BACHELOR

An analysis of commuter’s outflow from train platforms via stairs

Katiliūtė, Veja

Award date:
2023

Department of Mathematics and Computer Science
Data Science

An analysis of commuter's outflow from train platforms via stairs

Bachelor's Thesis

Vėja Katiliūtė

Supervisor:
Marko Boon

1.2 version

Eindhoven, 23 January 2023

Abstract

With the upsurge in crowding, there is an increasing risk of accidents happening. The problems call for crowd management solutions to simplify the flow. However, if we want to nudge people to take certain routes, it is important to understand the underlying psychology and physics. Using data given by the challenge owner Prorail company, we investigate the outflow at Eindhoven train station and try to measure what influenced people's choice to choose stairs or escalators. The purpose of this project is to verify some factors based on research found but also look into some factors that do not have much explanation discovered yet. We first define the processing procedures that should increase the quality of the given data by reducing errors and outliers. And secondly, we apply different statistical methods to verify previous assumptions and identify new potential directions. These statistical analyses include Jaccard's coefficient, Pearson's correlations, Mann-Whitney U test, Pearson's chi-square, and logistic regression. Using these methods we find that larger crowd size, speed, and path length have positive effects toward people taking stairs. We also find that longer time spent on the platform is associated with escalator usage. Traditionally, we select the best features and a model that we would use for decision prediction. We achieve up to 66% prediction accuracy using the engineered variables and logistic regression.

Contents

1 Introduction

Public transportation usage has grown significantly in the Netherlands over the past decades. As a result, the Netherlands is experiencing more problems with an incrase in crowding. Higher crowdedness causes more accidents and service disturbance, which calls for a long-term solution. Primarily for safety, convenience, and customer satisfaction, it is essential to implement crowd management techniques. One of the most straightforward solutions lies in the design, which is the physical layout of the environment where crowds form. Even though physical transformations are the key to dictating the movement, design changes are usually costly and highly time-consuming. If not the physical layout, another concept we could explore is nudging. Nudging is an applied theory in behavioural economics and decision-making proposing an adaptive design of the decision environment as a way to influence people's behaviour. For example, in a train station setting, blocked rail trails or signs of arrows showing the exit or other directions could be used as nudging techniques. Train stations in the Netherlands already implement some methods, such as directional arrows and optimally regulating train schedules. However, to prevent raising crowd issues in the future, it is essential to look into the critical points where crowding occurs the most. In the case of the train stations, the critical points would be near the exit, when people's goal is to leave as fast as possible. At the exit points, we would aim to reduce the crowdedness and increase the flow to ensure a higher capacity for more people. This problem requires an efficient plan for utilizing nudging techniques or evidence of a need for altered station layouts.

Nonetheless, before suggesting any potential improvements, we must understand what factors play a role in the movement process at the critical points in the train stations. In collaboration with Prorail, the company responsible for the track infrastructure in the Netherlands, we use the resources at hand and look closer into the flow at the exit points of the platforms in a train station. Particularly at the exit points of the train platforms, we observe an exiting layout that offers travellers to exit through stairs or escalators. When people choose to leave through stairs or escalators, we observe a decision outcome. When we look into the person's path, we can distinguish the individual speed, distance and time spent on the platform. These important observations can be used to research the difference in some factors based on whether people took stairs or escalators. For these analyses, we will use Jaccard's coefficient, Pearson's correlation, Mann-Whitney U test and Pearson's chi-square. To draw the main conclusions we use univariate and multivariate logistic regression analysis. These methods will be elaborately explained in the report. As a result, we can find out whether some factors are connected to people's choices every day when exiting the trains.

1.1 Literature review

In recent years, the crowd management topic has become increasingly prevalent. Since many industrial structures have escalators and stairs constructed, there is already some research into decision-making concerning stairs and escalators.

One of the factors that have been probed is the crowd size. Firstly, we found evidence that crowdedness, in general, influences decision-making. The research showed that when making choices, social crowding level was significantly influential (Maeng et al. [2013]). Specifically, discussing decision-making between stairs or escalators, crowdedness also has a significant level of prediction power. Using simulation techniques, researchers looked into different crowdedness measurements and found that using such measurements alone could reach up to 90% of the predictability of people's choices (Srikukenthirana et al. [2014]).

Besides crowdedness, some discussions on height suggested that a higher level of the pedestrian route would lead to higher usage of escalators despite the path direction. Using the height variable, it was possible to reach up to 85% accuracy in bidirectional prediction (Li et al. [2014]) when looking into environments of different height levels.

Another factor that research has touched upon is the moment of the day, meaning peak and non-peak hours. There was evidence that people prefer to take escalators during off-peak hours, as opposed to higher stairs usage findings during peak hours (Lazi and Mustafa [2015]). The main criticism of such observation is that peak hours are closely related to crowd size because people travel more during peak hours. Since crowd size may depend on hours, the factor of peak hours may have an overlapping effect compared to crowd size.

Lastly, some findings suggested that speed can also be significantly linked to taking stairs. The faster people speak, the higher the likelihood of choosing stairs. For example, it was found that using Artificial Neural Network, the accuracy reaches up to 85% when using a measurement of velocity (Yuen et al. [2012]). It indicates that speed may affect people's decision-making.

In a nutshell, some research already indicates people take stairs when crowd size increases. Peak hours and higher individual speed are associated with stairs, while the height of the pedestrian path leads to higher escalators usage. These variables have shown predictability and explicability for the decision-making system between stairs and escalators. We will use this information to verify our research.

1.2 Research question

For this report, we combine the advantage of collaboration with Prorail and literature review. Prorail company contributed to this project by offering their collected data, which is the base for this data science project. In this report, we will investigate features that can be extracted with high precision from the given data. The information to which Prorail has access offers a view of the transit area between the platform and the train station. It allows us to investigate people's movement around the critical exit point. Based on discussed research, we will explore some aforementioned features and some other features that can be conveniently extracted from the given data. We will look into the crowdedness levels, individual speed and peak hours. Besides, we will also see the effects of path length and time spent on the platform. To add another level to our research, we will analyse the measurements of individuals and the average crowd speed, path length and time spent on the platform. The performed analysis will help us answer what factors influence people to take more stairs. These results will hopefully contribute towards the research that connects crowd management and decision-making by adding another layer of evidence. In this report, we will answer three questions relating to decision-making, stated as:

- Can we verify the previous research that crowdedness, peak hours and higher individual speed are linked to higher stairs usage?
- Do factors such as path length and time spent on the platform affect stairs usage when exiting the platform? If yes, to what extent are they influential?
- Can average crowd speed, crowd distance, and time the crowd spent on the platform indicate an individual's decision-making when exiting?

These questions will help us conclude our main question:

- What factors can we associate with people taking stairs when exiting from the train's platform?

This project will inspect the Prorail data and suggest a processing technique for our case, which is essential for the analysis. Furthermore, we will use statistical analysis to better understand the chosen factors in the context of our data. Lastly, we will measure the effects of selected factors on the decision outcome, whether a person decides to take stairs or escalators.

2 Exploratory analysis

After discussing the subject of this research project, we will now look closer into the resources we are provided with, which is the Prorail data. This chapter will offer a close look into the information and errors that we encounter in the data. It will help us decide on the data processing framework, which is key for selecting the correct information for analysis.

2.1 Raw data

In order to tackle these questions, Prorail representatives gave us access to their server. The database contained information on five different stations, including Eindhoven Central Station, which will be our target station. This data was extracted from the sensors that Prorail installed for crowd observation and management research. Some of these sensors date back to 2017, and some were installed a year ago. Each station has a different volume of data; however, even the latest ones provide enough information due to the reporting frequency. These sensors continuously create shots of people's positions ten times a second, giving us detailed information. In the database, we get access to hourly blocks with an accuracy of a tenth of a second. One example of data that we will be using in the project is seen in Table 1. This shows a few datapoints that were captured in Eindhoven Station on the platform between the first and the second railway. The database contains a datetime which shows the date and time information, the object's coordinates encoded in x_pos and y_pos parameters and an anonymised tracked_object id, which is unique per detected person.

Example of data				
date_time_utc	tracked_object	x_pos	y_pos	datetime
1.636715e+12	500270	-2719.0	63889.0	2021-11-12 12:00:00+01:00
1.636715e+12	500577	-414.0	24695.0	2021-11-12 12:00:00+01:00
1.636715e+12	500647	1657.0	29373.0	2021-11-12 12:00:00+01:00
1.636715e+12	500674	-1619.0	17582.0	2021-11-12 12:00:00+01:00
1.636715e+12	500680	-2249.0	62697.0	2021-11-12 12:00:00+01:00

Table 1: Example of Prorail dataset from 2021-11-12 12PM

The size of hourly data frames fluctuates significantly. Less busy hours could produce a few MB of files, while hectic hours can create 30MB or larger files. This means that processing one week alone can take multiple GB of data, which can be very expensive to compute, depending on applied methods. Due to time and computational power limitations, we will use data samples instead of all data for our analysis. It is essential to consider that throughout the project because it can be costly concerning time. Besides, our produced samples are arbitrarily large for analysis. For this project, we will use samples of collected data in Eindhoven Central Station on the platform between the first two railways in 2021.

2.2 Data content

The given data allows us to understand where and when people are located. Observing people's position at a given time can show a momentary object distribution on the platform, as indicated in Figure 1.

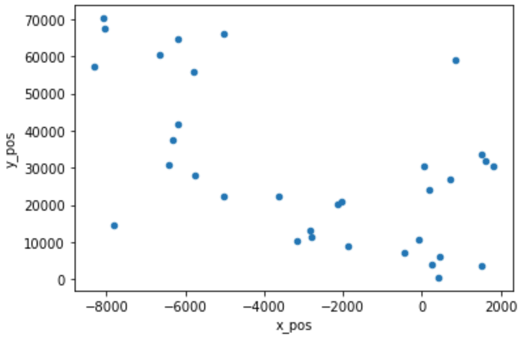


Figure 1: Object positions at 2021-11-12 12PM

We can also follow objects' movement in terms of x and y positions, which suggests another utility for the dataset. We can track the movement by looking at unique trajectories and plot them accordingly, as pictured in Figure 2. Each coloured line here represents a different tracked object path, and the grey background image allows us to better visualize the action on the platform. The grey imaging will be used for many following visualizations. We will later extract features from these trajectories that we use for our analysis.

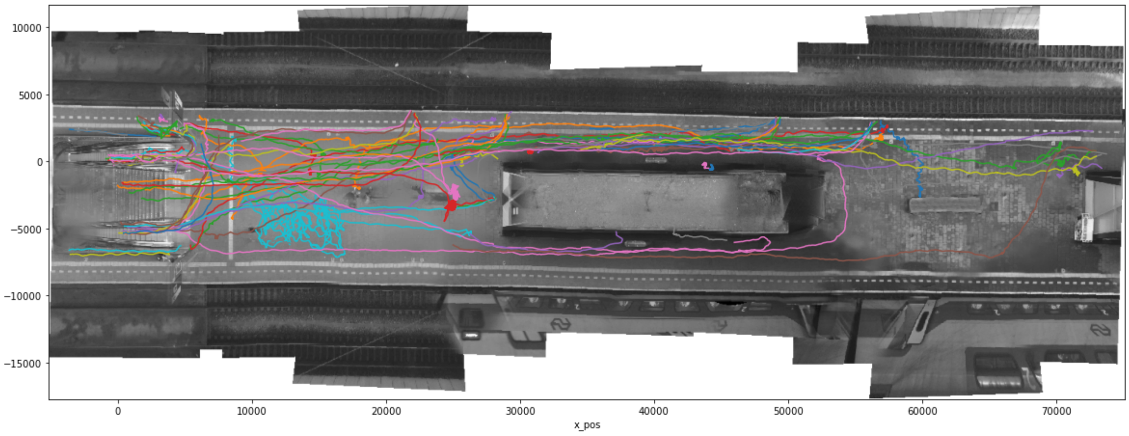


Figure 2: Trajectories at 2021-11-12 12PM

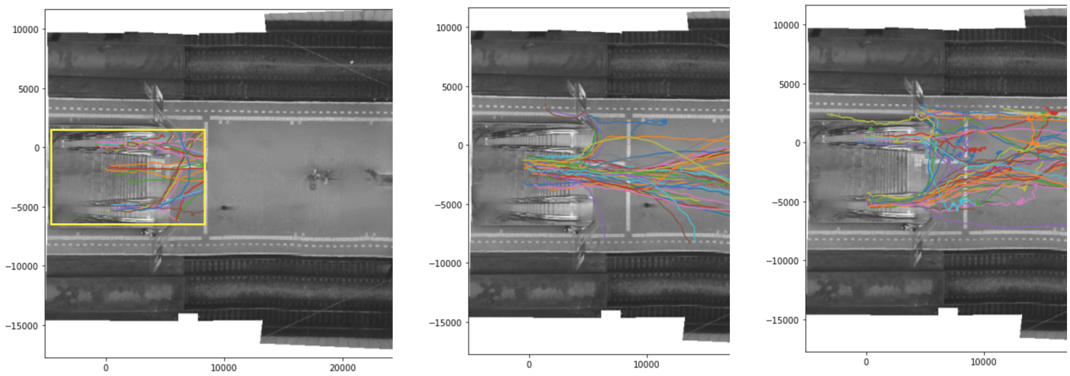
2.3 Area definition

Using our dataset, we can also identify the area on which we will focus in the report. It is the exit and entrance area connecting a platform and the rest of the station. We will briefly discuss the area and direction definitions that will be important whenever we mention stairs or escalators area.

2.3.1 Stairs and escalators

For this project, we manually found the exit area. Looking at the data, we configured the target area to be within $x \in$ (-5134, 8300) and $y \in$ (- 6185, 1275).

In Figure 3a, we can see a yellow outlined area which uses the mentioned x and y coordinates intervals. This is the definition for the stairs and escalators.



(a) Exit area (b) Stairs paths (c) Escalators paths

Figure 3: Selected trajectories at 2021-11-12 12PM

Going a little further, we also split the target area into specific parts: stairs and top and bottom escalators. We defined the stairs and escalators' common area quite liberally, which means the area includes a further margin than where the actual stairs end. However, for the exact space of stairs alone, we redefined the x to have a lower margin so we do not have ambiguous values. So for stairs, we manually configured $x \in$ (-5134, 4500) and $y \in$ (-4000, -500).

These configurations allow us to select specific paths leading through the stairs, which is visualised in Figure 3b. For escalators, we used the same x coordinates as for stairs. The y for the bottom escalator, which goes in the direction towards the platform, was $y \in$ (-6185, -4000) and the upper escalator, going in the opposite direction, was described as $y \in$ (-500, 1275). The escalator's definition allows for escalator path selection, as shown in

Figure 3c. In technical parts, when coding and analysing the target area, we will use the full prefix ”stairs” to describe stairs, and for escalators, we use the abbreviation ”esc” for convenience. The following sections will explain the technical terminology and variable abbreviations in more detail so the reader can follow the report.

2.3.2 Direction

One of the features that we introduce in data is a bidirectional division. We introduce a factor that tells whether the traveller is entering or exiting the train platform. In terms of data marking, we will use a suffix or prefix “to”, for example, “to_platform” and “count_to”, when a person is going to the platform. If the person is going from the platform, we will use a suffix “from”, such as “from_platform” as well as “count_from”. To find a walking direction, the object position at the x coordinate was used as defining criterion. When we observe the starting position x smaller than the ending x positions, we assume that the person was going to the platform. If the x was larger at the start than the x at the end, we assume the opposite: the person was going from the platform to the station.

These definitions help us generate the directions that can be visualised in Figure 4, where red colour represents people walking to the platform and blue sorting out of the platform. This division will be further used to select only the paths that lead to exiting.

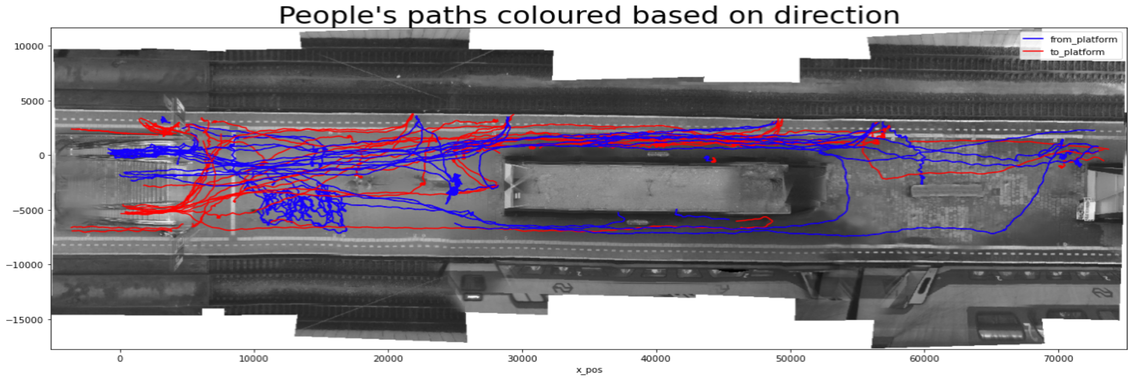


Figure 4: Directional trajectories at 2021-11-12 12PM

2.4 Data files selection

Moreover, raw data includes errors that we will briefly discuss in this section. Some parts of the data are inaccurate due to sensor errors, or other possible reasons. We use x and y coordinates to evaluate each hourly block in the database and create x and y filters. We will automatically apply these filters to select the files that are not erroneous and considered suitable for further processing.

2.4.1 Filter x

When looking for inaccuracies in hourly files, we manually found erroneous patterns we want to remove before proceeding with the process. We found a desired case using density plot detection, which we can see in Figure 5.

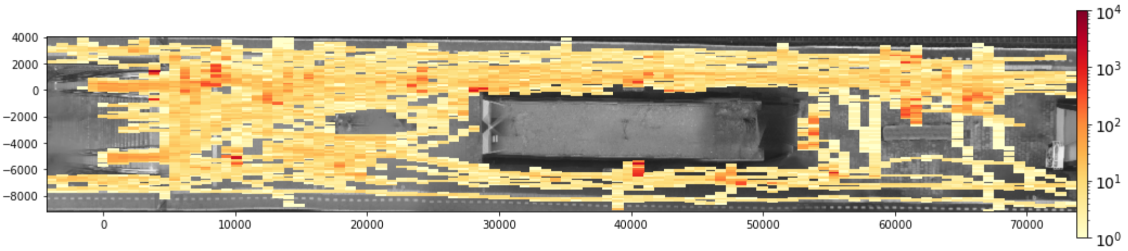
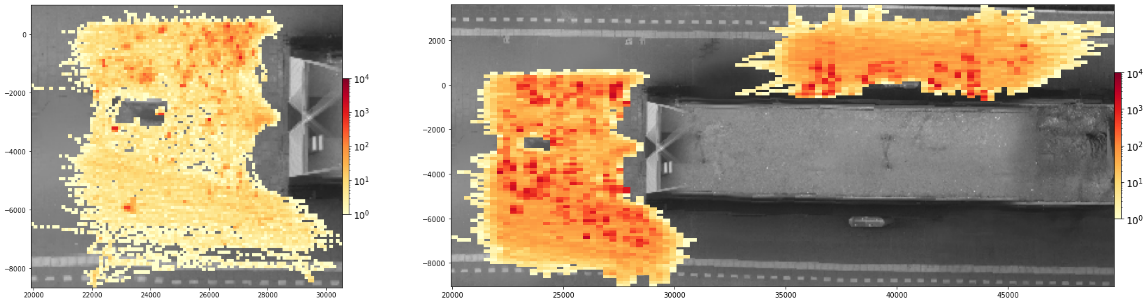


Figure 5: Density plot 2021-11-11 10AM

It shows where and how many people were present during the given hour. It represents the crowd walking and spreading out on the platform, which is the expectation of crowd behaviour. On the contrary, we found some hours that showed strange density concentrations only in particular areas. Such blocks of data we consider erroneous because the crowd is not spread out everywhere but only in an area not connected to an exit or entrance.



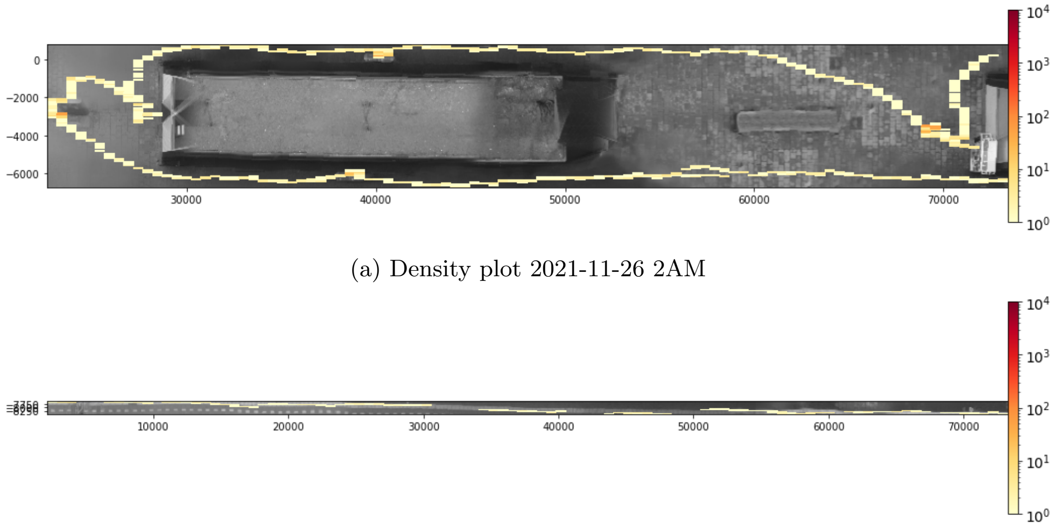
(a) Density plot 2021-11-06 12PM

(b) Density plot 2021-11-08 12PM

Figure 6: Density plots of erroneous data

For instance, Figure 6a visualizes an hourly file density where the movements are condensed only near the waiting area. In Figure 6b, we see the movement concentrated in two specific areas. Such hourly data might have occurred because of sensor misconfiguration, or perhaps some sensors stopped reporting data during some time. There could be other technical issues, or it could be other reasons. One explanation might suggest that during certain hours there were no travellers. Perhaps a very concentrated pattern was created by maintenance workers during those hours, and the sensor strangely captured the movement. Another reason can be singular homeless people sleeping, which could be detected using the benches as a reference. Despite the reason, these figures depict cases we would like to exclude from our analysis because they do not offer data around the exit area.

Nevertheless, many frames were rather ambiguous, as seen in Figure 7. These images show the density plots that extend through a substantial part of the platform; however, they do not lead to an exit. They might look correct, but amid our research, they would not offer much additional information. As a result, losing these paths from such files is not considered a substantial loss.



(a) Density plot 2021-11-26 2AM

(b) Density plot 2021-11-21 4AM

Figure 7: Strange and ambiguous density distributions

Besides individual examples, we also looked into the statistical differences of hourly data files, which serve a complimentary purpose for data selection criteria definition. To compare the files, we took a sample of one month and looked into the statistics of each file. We used the minimal values of x position (x_min), and standard deviation of x (x_std) and y (y_std) positions. We assumed that the minimal x coordinate would explain how close to the stairs our hourly observations would be. If minimal x reaches a low enough value, we would observe at least one position in the exit area, meaning that there are paths that exit or enter which is in our interest. The standard deviation measurement for both positions, x and y would show the variation in observations. If the variation of x is high enough, people spread out around the platform, which is a desirable frame. We expected that the standard deviation of y coordinate could also show some variation. We looked into the exact values of each hourly file displayed in Figure 8. We can see a high fluctuation in x_min and x_std. Indeed where we had a desirable density distribution on the platform, we observed a low x_min and a high x_std. It suggested that our assumption for x coordinate is reasonable. For y_std we can only say that it has less fluctuation.

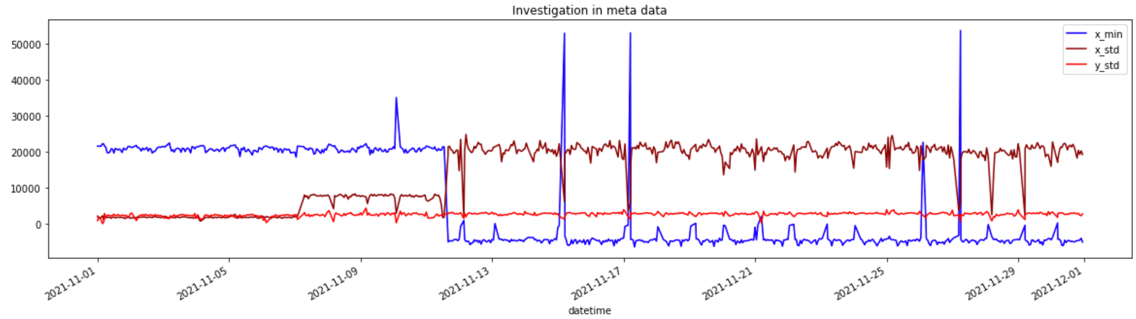


Figure 8: Statistics of unfiltered data

To quantify the criterion for x coordinate, we looked into the statistics of x_min and x_std. In Table 2, from the mean and median(aka 50th percentile), we see that most of x_min values are below 5000. For x_std, we see that most of the values are above 14000. Thus, we took advantage of these values and used them to create criteria for file selection. This criterion is defined by minimal x being less than 5000 and standard deviation of x more than 14000. We used sufficient but liberal criteria to select as many files as possible because we will apply other filtering techniques. After having applied the criteria, the statistics on metadata were as expected. In Figure 9 we notice the metadata of selected files using the x filters. All files fit in the pattern of small x position and high x standard deviation. These files are desirable because they have a naturally occurring distribution of people on the platform.

Statistical description		
	x_min	x_std
count	661	661
mean	4709.31	14547.46
std	12532.25	8144.60
min	-6380	765.19
25%	-4648	7501.25
50%	-4308	19616.91
75%	20468	20965.93
max	53651	24815.37

Table 2: Statistical description on data used for filtering

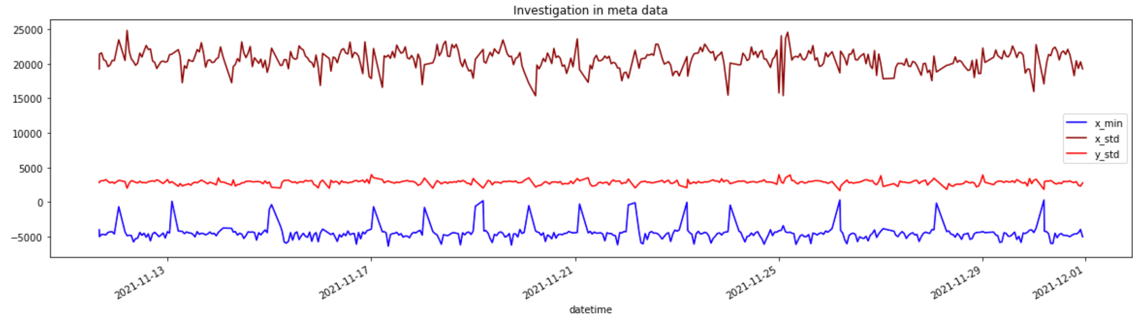


Figure 9: Statistics of filtered data

After we defined our criterion for x coordinate, the question arose on whether we should keep all the hours that show reasonable distributions or the full days with only non-erroneous information. Both applications have advantages and disadvantages. If we apply the filters on full days, we might lose some days that were meaningful because one hour poses errors. So, for instance, we might not be able to compare Mondays of one month if one Monday had any erroneous data, which has been eliminated. However, if we apply data filtering on hours, we might have comparable days but not necessarily the moments. We also cannot explicitly say whether the data was erroneous or there were no people on the platform. Because of that, it can be impossible to compare two events that happen yearly, like Dutch Design week or Glow. There are some arguments for both cases shown in Table 3; however, applying processing on hourly data is entirely sufficient and more efficient. For this project, we investigate different moments and people's behaviour. Because we do not compare specific events, it is better to keep even a few hours of the day if they do not have errors.

Arguments	
Filter by day	Filter by hour
Against: • loss of important data because we delete full days that had only one erroneous hour; • conservative selection; • not clear how to interpret the empty hours when there were no people on the platform;	Against: • in some cases we can not compare the different days; • we might not be able to investigate specific events;
Pro: • all days are non erroneous; • we still have loads of data;	Pro: • we have more data; • faster and simpler to implement;

Table 3: Arguments for filter application on days vs hours

2.4.2 Filter y

Previously, we looked into the rough outlines of the errors. This resulted in a clear definition of the x coordinate. We saw that the standard deviation of y coordinate did not show drastic change, which meant that we needed to approach y differently. For this, we reasoned to apply the knowledge of stairs and escalators area. If we observe any paths within the exit area in terms of y coordinate, the data should be considered non-erroneous and valuable.

Together with x filter, this selection procedure is to verify whether there are any people in the stairs and escalators area and whether the distribution is as expected. We apply criteria only to select the useful hourly data files and other files that are skipped and unused.

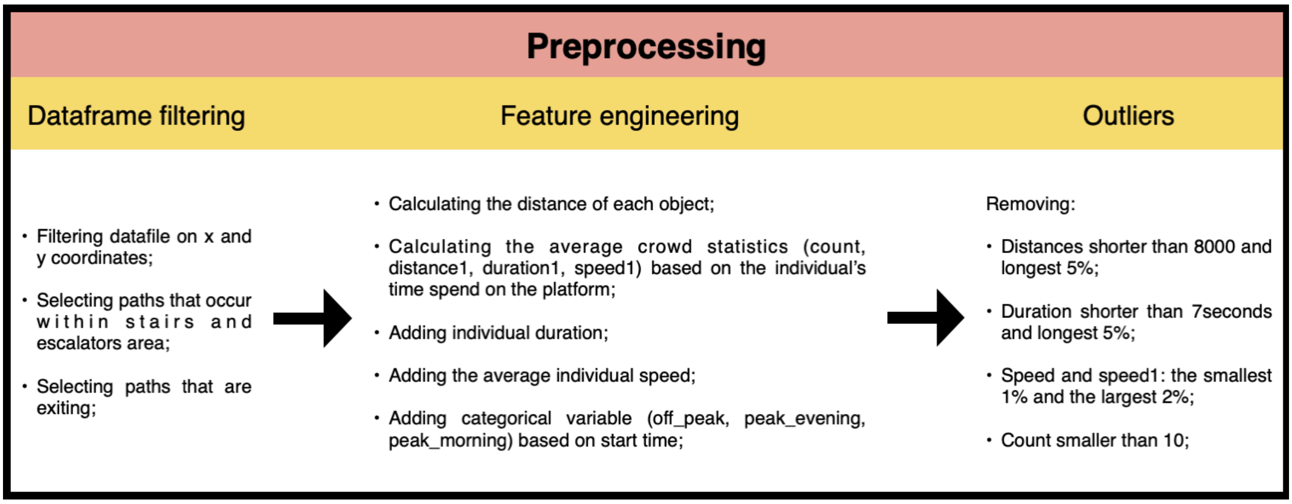
We now work to reduce the data even further to ensure the quality and reliability of this research. This will be explained in the following section.

3 Preprocessing

Earlier, we looked into the datapoint distribution on the platform, we created a dataframe filtering method for our particular data. This is a part of the first stage of filtering. This project includes a few steps of data selection in different parts of the process to ensure the usability of the data and result accuracy.

3.1 Path selection

This research aims to look into people's decision-making patterns and investigate whether some chosen factors are causal for decision outcomes. Thus, after applying x and y filters, we filter out the data files that are considered accurate in terms of data quality. Then heeding our research purpose, we pick out only the relevant paths. These are the exiting paths. In practice, it meant selecting the paths that pass through the stairs and escalators area. We then added a directional variable and performed the last step of excluding the travellers entering the platform. The procedure is briefly framed on the left side of Figure 10, named "Dataframe filtering". As a result, we are left with a collection of paths defined by people who are exiting the platform using stairs or escalators. This selection was to minimize technical errors and irrelevant data.



Preprocessing

Dataframe filtering

- Filtering datafile on x and y coordinates;
- Selecting paths that occur within stairs and escalators area;
- Selecting paths that are exiting;

Feature engineering

- Calculating the distance of each object;
- Calculating the average crowd statistics (count, distance1, duration1, speed1) based on the individual's time spend on the platform;
- Adding individual duration;
- Adding the average individual speed;
- Adding categorical variable (off_peak, peak_evening, peak_morning) based on start time;

Outliers

Removing:

- Distances shorter than 8000 and longest 5%;
- Duration shorter than 7seconds and longest 5%;
- Speed and speed1: the smallest 1% and the largest 2%;
- Count smaller than 10;

Figure 10: Preprocessing steps

3.2 Engineered variables

The next step of the processing before analysis is the obtaining of the measurements defined in the research question. From the raw data, we would like to compute the crowd size, peak hours and numeric variables as previously described. We also want to find the individual's path, time spent and the speed. Afterwards, we create some parameters to store the crowd statistics, such as average distance, time spent on the platform and the average crowd speed. Lastly, we engineer a binary target variable to encode the decision outcome, whether we observed the person taking stairs or escalators. The middle section named "feature engineering" in the Figure 10 summarizes the processing techniques used to obtain the extracted variables that we use for analysis. The overview and more detailed explanation of engineered features can be seen in the Table 4.

Our engineered variables can also be classified into different types. We have categorical or qualitative data such as stairs, off_peak, peak_morning, and peak_evening. They are defined as nominal variables, meaning the values are qualitatively different and can not be interpreted numerically. The other variables, such as duration, distance and speed, are numeric interval continuous because they are measurements and can take an unspecified number of values. The variable count is also numeric, but it is countable and can take any value in a set of non-negative natural numbers. This variable also has natural order or rank based on a scale of low to high. It makes count a ratio variable, which can be seen as a discrete numeric and ordinal variable. We will use this ambiguity to our advantage, based on the purpose of analysis. It is important to consider the data types of our information, so we can determine the right methods for statistical analysis.

Engineered variables description	
Variable	Description
count	Count of people that are exiting the platform
distance	The path length of the object (in terms of coordinates distance)
duration	The time that an object spends on the platform (in seconds)
speed	The average speed of the object (coordinates distance divided by duration in seconds)
distance1	The average path length of all objects on the platform that are exiting during the time when the unique object is on the platform (in terms of coordinates distance)
duration1	The average time that people spend on the platform during the given time (in seconds)
speed1	The average speed of all objects on the platform that are exiting during the time when the unique object is on the platform (coordinates distance divided by duration in seconds)
stairs	Categorical binary variable, describing the decision outcome. 0 ← escalators 1 ← stairs
off_peak	Categorical binary variable, describing the travel time. 0 ← peak hours 1 ← off peak hours
peak_morning	Categorical binary variable, describing the travel time. 0 ← off peak hours 1 ← morning peak hours (6:30 - 9:00AM on workdays)
peak_evening	Categorical binary variable, describing the travel time. 0 ← off peak hours 1 ← evening peak hours (4:00 - 6:30PM on workdays)

Table 4: Engineered variables used for the analysis

3.3 Data transformation

Another step of processing the features is data transformation. Usually we do not have a normal distribution in our observations, which is needed for many classical statistical methods. Often our data is skewed and do not have a clean normal pattern. One of the methods that are widely used in dealing with skewed data is logarithmic transformation. This transformation changes the variable x with $log(x)$, which now reflects the percentage difference instead of the value difference. This simple method faced some criticism about its adequacy. According to Feng et al. [2019], there is evidence that standard statistical tests performed on log-transformed data are often not relevant for the non-transformed original data. Because logarithmic transformation often reverts data to a normal distribution and decreases data variability, it is easy to apply traditional methods for analysis. However, the paper suggests not applying the transformations and using newer statistical methods that do not assume a normal distribution.

There are more research on how to transform the variables into approximately normal distribution. However it highly differs per data case and often finding the right equation of transformation requires a separate in depth analysis which is outside of the scope of this research. Consequently, we will refrain from using any transformations in order to gain more reliability in our results.

3.4 Outliers

Besides transformations, there is another good processing practice used for data preparation. This includes removing outliers from our data. We remove the outliers from the engineered variables such that we leave some noise out of the analysis. The steps are depicted again in the Figure 10 on the right side under "Outliers". The outlier selection was obtained through trial of manual configurations and statistical summaries to remove the datapoints that add little volume but introduce high variability in the data. In this project we use a Logistic Regression, which is highly sensitive to outliers and can show imprecise results. Hence, removing outliers ensures the reduction of measurement error and lower variance therefore more consistent estimation. As pictured in the Figure 10, all three main steps of preprocessing or data preparation were to ensure that we have measurable subjects, adequate data quality and reliable results from our analysis. These steps are highly important and must have been looked closer into. The next section will explicate the observed and engineered variables and statistical analysis.

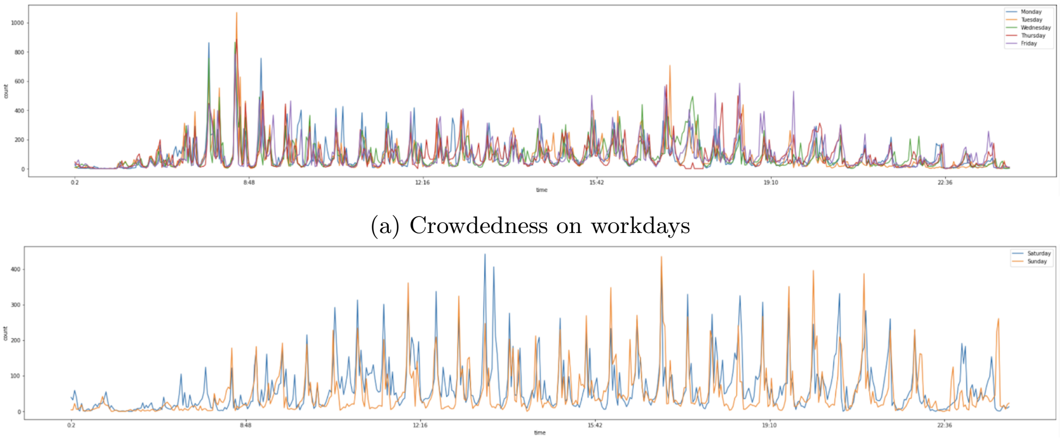
4 Statistical analysis

After discussing the processing steps, we are aware of modifications performed on Prorail data. This chapter will be dedicated to statistical analysis of the target variables that we obtain from feature engineering steps. Besides count, we inspect path length, time spent on the platform and speed. For categorical variables, we look into the peak hours.

4.1 Count

One of the measurements that we obtain is the crowd size. We measure it by computing the unique people count during a particular time of the day. This simple calculation allows us to plot a general curve of the count during the day and compare it between the different hours or days. The general crowdedness can give an idea of crowd distribution during the weekdays and indicate when the crowd is larger and when there are fewer people on the platform.

Figure 11 shows the total crowd size on the whole platform on all weekdays, which is drawn from a one-week sample. Figure 11a indicates the general trends for work days, which is enough to show that we have a relatively similar distribution on every work day. Generally, we notice a high fluctuation of a crowd during the whole day. These fluctuations are due to regular train schedules that are the same on workdays and weekends. Nevertheless, the crowd size differs throughout the day, and we can observe some patterns when the crowd size increase. For instance, on weekdays, the highest peaks are during morning peak hours, which can be explained by the fact that people are going to work or school. During the evening peak hours, there is also an increase in travellers; however, it is less noticeable and lasts till later after the peak hours end. This could be explained by the different hours at which people finish work or school or the price incentives to travel off-peak hours. Generally, peak hours seem to be closely connected to the crowd size, which can be attributed to fixed work hours.



(a) Crowdedness on workdays

(b) Crowdedness on weekends

Figure 11: Count of people every two minutes on different weekdays

On the weekend, we observe another distinctive pattern for Saturdays and Sundays (Figure 11b). Compared to the curve of workdays, we see a more uniform spread in peaks, suggesting that there are no particular hours that are more appealing to travellers. Except for morning hours, on weekends, we could expect a somewhat uniform crowd distribution. Generally, the busiest hours are in the mornings of the workdays.

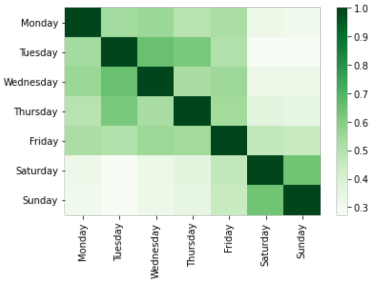


Figure 12: Correlations between crowds on different weekdays

To verify the similarity between the weekdays, we created a correlation matrix for statistical comparison. Pearson's correlations measure the inter-dependencies of given variables, which will be used further in the report and explained in detail in the chapter on causal analysis. From the correlation scores depicted in Figure 12, we can identify two distinct clusters of similar patterns. One collection is formed on workdays such as Monday, Tuesday, Wednesday, Thursday and Friday, while the weekend, Saturday and Sunday would create another group similar in crowd patterns. To add, Friday also has some significant similarities with the weekend patterns. It could be explained by people's tendency to consider Friday evening as the beginning of the weekend and thus explains more activity in the later hours of the day. As mentioned before, these trends can mainly be attributed to trains' schedules, which remain fixed on work days and weekends.

Besides comparing the weekdays, we can also compare one weekday to understand the similarities. As an example, we compare two Saturdays, one with an event and one without. Saturday, November 13th was the last day of Glow 2021, which is a big event in Eindhoven, when many people come in the evening to visit the infamous light festival. November 20th was a casual Saturday when no specific event took place. Comparing these two days, we again see a similar pattern except for late hours shown in Figure 13. On the day of Glow, we see large crowds late Friday night and higher numbers on Saturday night. We see that these Saturdays look very similar. Crowds at night, however, differed a lot, which is clearly due to Glow. It suggests that we could expect every weekend to look alike and most likely the workdays as well, but in case of an event, we will see increased numbers of people.

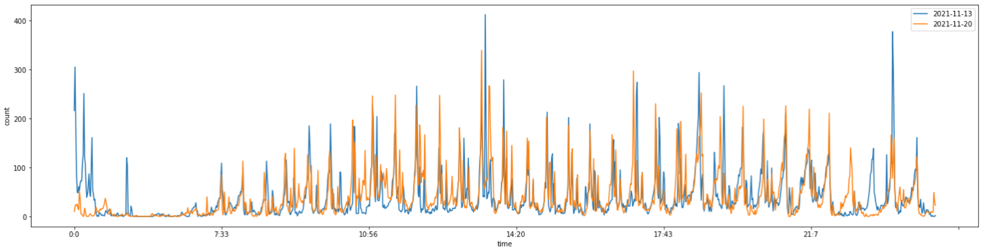


Figure 13: Crowdedness on two different Saturdays

Besides looking into the general count, we can also investigate the count numbers further when zooming into peaks. For that, we take one of the busiest hours mentioned above, which is on Tuesday morning between 8-10 am.

To begin with, we may look into the conditional count based on the direction that has been defined. For example, in Figure 14 we can see the count computed with respect to different directions when coming and leaving the platform. Interestingly, we see a clear difference between the directional count. This is because people walking from the platform create a denser crowd as compared to those walking to the platform. It can be explained by the fact that once a train arrives, a momental crowd passes from the platform to the station and when people are coming to the station to take the train, that crowd may be spread out over a more extended period of time.

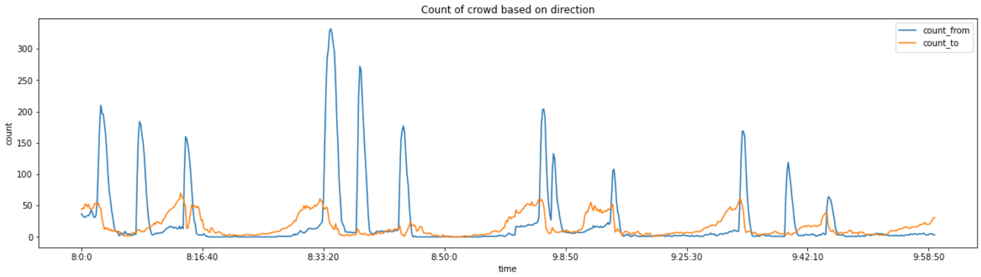
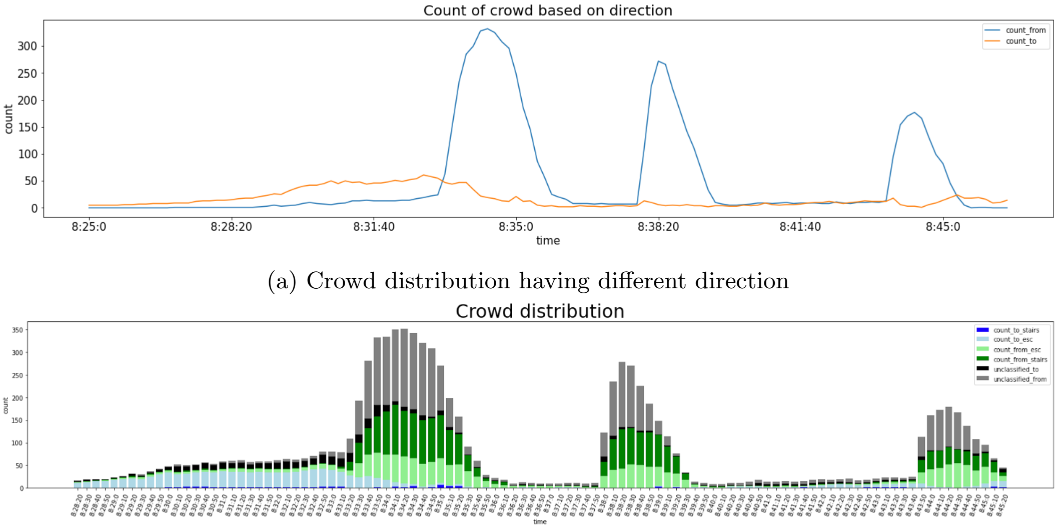


Figure 14: Crowdedness based on direction that people are taking

In order to analyze the crowd on the peak, we are zooming into the highest peak of the day, which occurred roughly between 8:25 and 8:45 am. Figure 15a depicts the zoomed bidirectional count for convenience. The plot that gives us important information is in Figure 15b. This stacked bar chart was generated using a nested conditional count, which has a division in direction and escalators or stairs. The blue colours show the movement towards the platform or, more precisely, people taking stairs and escalators to reach the platform. In contrast, the green colour shows the movement from the platform to sort out through stairs and escalators. If we look at the darker green colour, we shall notice that the count of people taking stairs largely increases with the total count. The light blue and green colours symbolizing the count of people taking escalators is rather stable and increases less than the crowd on the stairs. Overall, people prefer escalators, so when the crowd is relatively small (less than 100 people on the platform), most people take the escalators. But if the crowd significantly increases in a short period of time, the crowd naturally spreads out on stairs and escalators, hence nudging a majority towards the stairs. This happens because stairs have a larger capacity for movement; however, escalators are more convenient.



(a) Crowd distribution having different direction

(b) Crowd distribution having division between stairs and escalators

Figure 15: Conditional crowd distribution

4.1.1 Error margin

We also must address the grey and black area in Figure 15b since it involves almost half of the data points during busy moments. The grey and black colours depict data points classified by direction but not found within the stairs and escalators area. One of the reasons could be cross-platform transit or people opting for an elevator. However, common intuition suggests that it can not be half of the crowd transiting or coming/exiting through the elevator. In Figure 16, we see an investigation into those trajectories that never reach stairs or escalators. Some trajectories within the yellow circle area seem to be cut at a certain point, which may be attributed to the sensor error. Suppose there are too many people in one place moving hectically. In that case, it may trigger inaccuracies in sensors, and they may attribute new trajectories to the same objects or objects losing trajectories along the way.

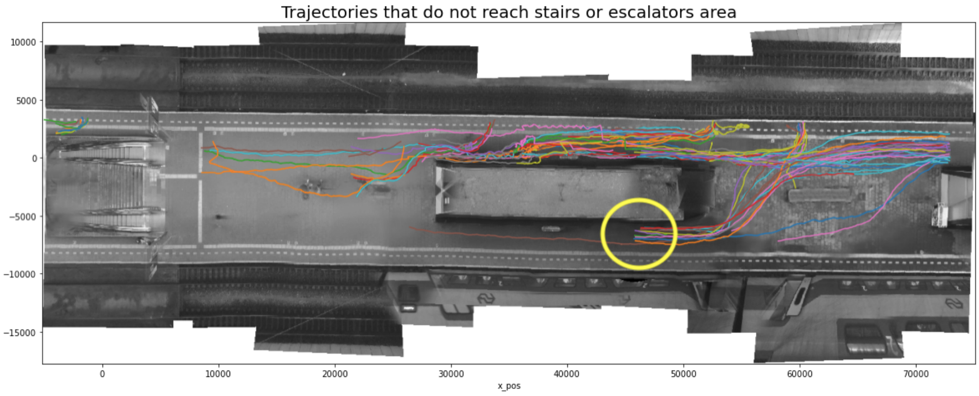


Figure 16: Trajectories that follow towards the exit but do not reach it

It is impossible to calculate the errors precisely because we do not know exactly why these inaccuracies appear. However, these errors create much noise that is excluded from the later analyses. Using Figure 15b, we can estimate that up to 50% of information at very busy moments is lost due to sensor errors. It challenges the processing of crowded moments and the extraction of quality information. When drawing conclusions, we must remember that we use partially incomplete data and that many paths might be lost, which can introduce bias. Together with the previously described data selection, we present some disproportionality in the data, which should be regarded.

Despite the error, we see that stairs usage is connected to crowd size. So now we shall look into other features.

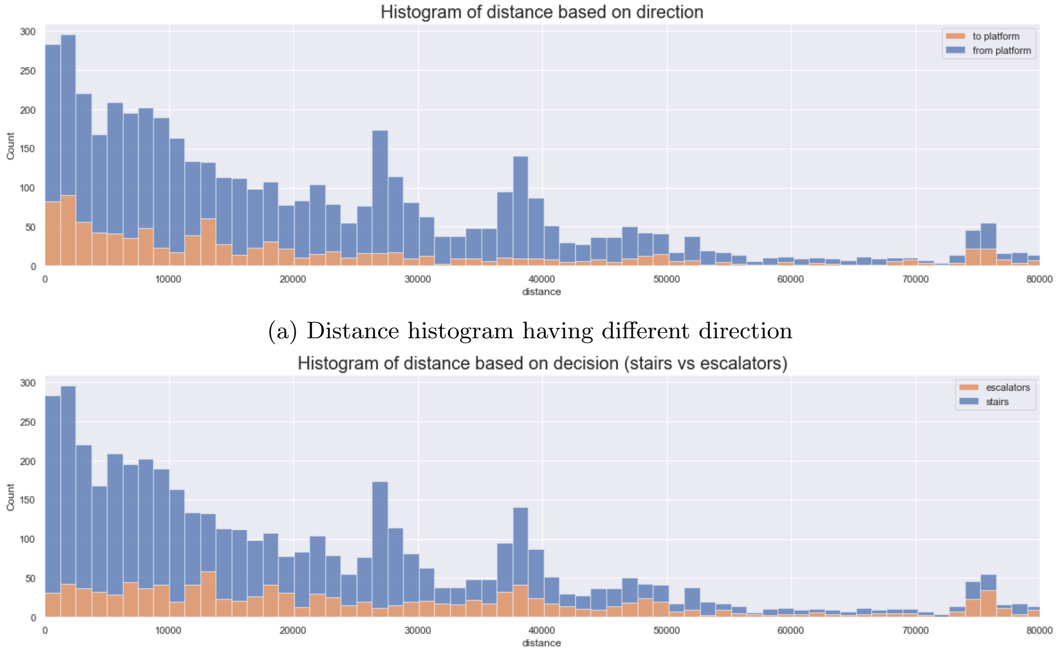
4.2 Distance

Furthermore, we looked into the measurement of distance, which is defined by the length of the path. The length of the path was calculated with respect to x and y coordinates. Therefore, the scale corresponds to the distance of the coordinates calculated by:

$$distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

We looked into a sample of individual distances to examine whether we distinguish any patterns in conditional path length. For example, it is possible that distance would affect decision-making in a situation of large crowds going from the platform. Intuitively, large groups result in people being forced to walk around and therefore have a smaller chance to get to escalators, thus needing to take stairs. In Figure 17a, we see a distribution of individual distances based on direction and in Figure 17b, we see the conditional distribution based on escalators and stairs usage.

From the total distribution, most people take relatively shorter paths, but we also observe a non-linear decrease. We see that the smallest distances are the most frequent, but there are significantly many paths that are about 28000 and 39000 points in length. This suggests that there could be some pattern in what paths people take. Perhaps there are types of trajectories that are more frequent. Those trajectories lead to a certain distance that travellers appear to take. In terms of conditional distribution on direction (Figure 17a) or on decision outcome (Figure 17b), we see no apparent difference. Stairs usage is relatively homogeneous and does not necessarily depend on distance.



(a) Distance histogram having different direction

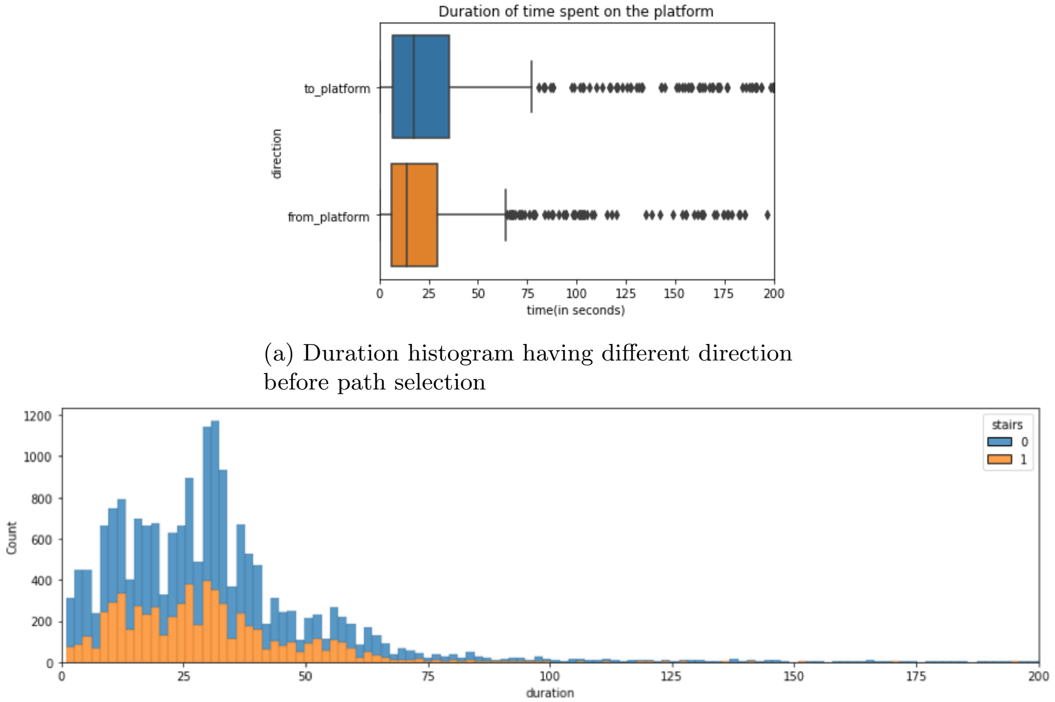
(b) Distance histogram having division between stairs and escalators

Figure 17: Histograms of distance variable

4.3 Duration

Another variable we look into is the time people spend on the platform or duration. We look into the measurement of duration, which is calculated as the difference between the start and end time in terms of seconds. It is possible that the time that one spends on the platform would differ when coming to the platform or exiting. It can happen because people that come to the platform are less rushed and might come earlier and spend significantly more time than people who are exiting and are directed towards faster movement. In Figure 18a, we observe boxplots of distributions when people come from and to the platform. For illustration purposes, we reduce the outliers, and it is still visible that the values seem quite similar on average, despite the direction. One thing that can be said is that the distribution of the crowd exiting the platform is more condensed and consistent compared to the other law. The absolute majority take less than one minute on the platform, and direction does not seem to affect the duration as much. However, this analysis was performed before path selection, which suggests that there could be a substantial error.

We also compare the duration based on the decision outcome. The following visualisation is already performed after the path selection when noise is removed. In Figure 18b, we look at a histogram that shows people's time spent on the platform with respect to decision outcomes. The figure depicts that duration does not differ very much based on the decision result, which does not suggest any clear patterns. Based on these analyses, we can conclude that there is no evidence that duration depends on the direction or the decision outcome.



(a) Duration histogram having different direction before path selection

(b) Duration histogram of exiting paths given stairs and escalators. 1 refers to stairs, 0 refers to escalators. This is drawn from selected data

Figure 18: Histograms of duration variable in seconds

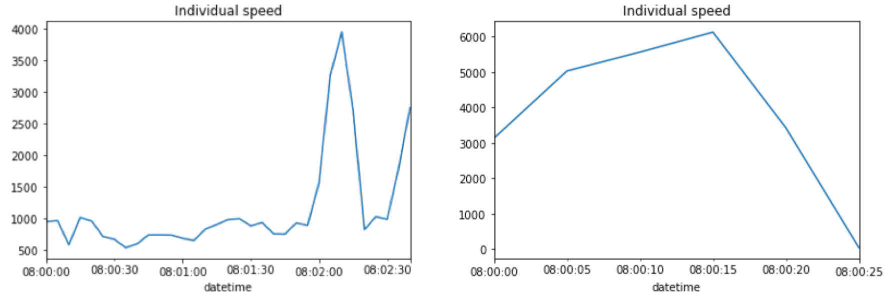
4.4 Speed

Another variable to look into is speed. For speed, we use the formula of distance and time division. We use a relative speed unit, which is the coordinate distance divided by seconds.

Speed is an interesting measurement because it can be expressed as an individual function. Figure 19 shows two examples of individual speed functions. It depicts how different each speed function can be. However, for our measurements, we use the average individual's speed because it is much easier to extract, and modelling individual speed is out of the scope of this report. It can be considered an idea for later research.

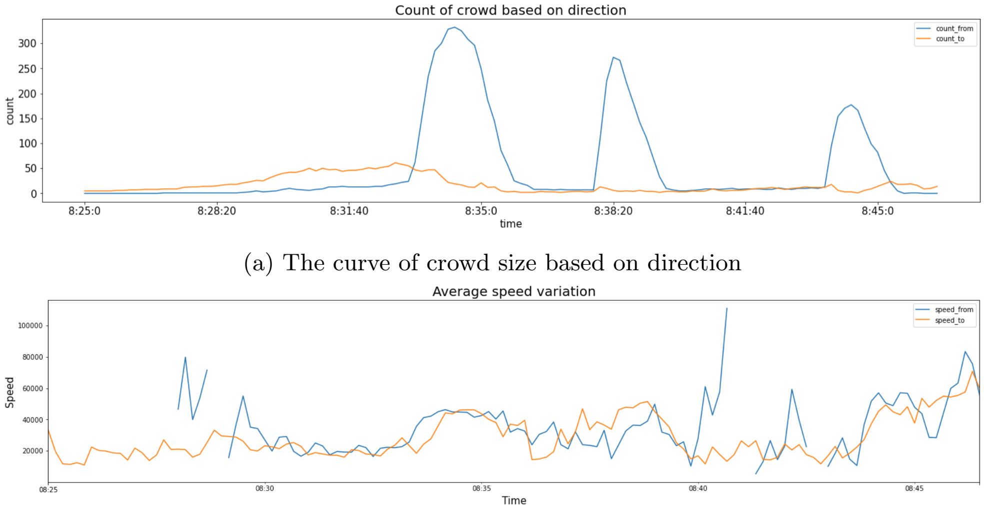
Furthermore, we look into the average speed variations depending on the direction and the decision outcome. From Figure 20, we see how the average crowd speed (Figure 20b) fluctuates compared to different count values (Figure 20a), which does not yield clear patterns. We only notice that when the crowd size increase drastically, the speed seems to stabilise and have less variance. It means that at highly crowded moments, the speed of separate travellers becomes more similar. It can be explained by the crowd effect. It occurs due to a particular crowd speed, and one person can hardly move much faster or much slower; hence, the speed differs less. On the other hand, when there is little to no crowd, the speed has more variation because individuals move the way they want and therefore crowd has no effect.

In the following Figure 21, we look at the average speed variation in function of time based on taking stairs and escalators. Compared to count (Figure 21a), we can identify some patterns for speed variables. Firstly, we



(a) Speed function of object 628314 (b) Speed function of object 628207

Figure 19: Speed functions of two unique objects



(a) The curve of crowd size based on direction

(b) The curve of crowd speed based on direction

Figure 20: Curves of count and speed depending on direction

see that when the crowd increases, the speed also increases. This becomes visible when comparing the peaks of the crowd curve with the speed curve (Figure 21b).

It can be said that people that take escalators have a more consistent average speed while the stairs crowd has a more fluctuating speed. It could mean that looking into individual speed functions can be interesting, and it may result in some connection between people's movement function and decision outcomes.

The previous analysis was performed before path selection, which includes noise. In Figure 22, we also showcase a speed histogram computed after the path selection. It has little to no noise and errors. The histogram shows a general distribution of speed in selected paths. We see that exiting people have a similar speed when choosing stairs or escalators. Only people with low speed tend to rarely select the stairs. It must include older people or people who have problems moving, leading them to choose a more convenient descent which is escalators.

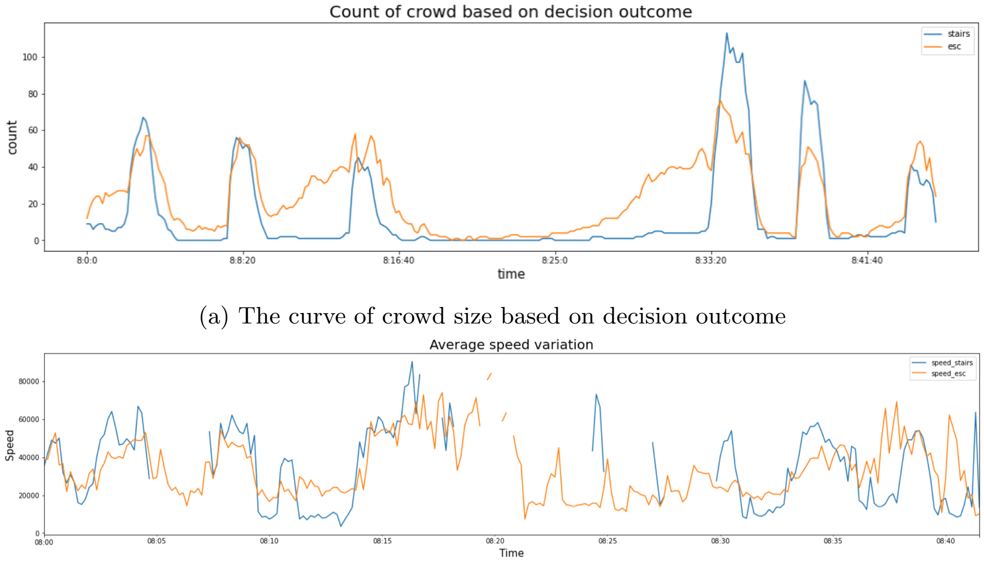
These analyses suggest some patterns between the speed and direction or decision outcome, but further investigations will be needed to assess the relationships closer.

We discussed the variables of all observations. Some visualisations were already drawn from data that was filtered using all three processing steps that we discussed in the previous preprocessing section. Now we will assess the fully processed data and explain its relevancy for the causal analysis.

4.5 Distributions after preprocessing

We have already considered some important remarks on the subject variables. This section will explain the distribution of the fully processed information and important aspects of causal analysis.

Firstly, we look closer into the distributions of the engineered variables that were filtered, selected and cleaned from outliers. Then, the data is prepared to be assessed for further statistical analysis. After data cleaning, we are working with a sample of around 28000 data points, where around 56% of entries have decision outcomes



(a) The curve of crowd size based on decision outcome

(b) The curve of crowd speed based on decision outcome

Figure 21: Curves of count and speed depending on decision outcome

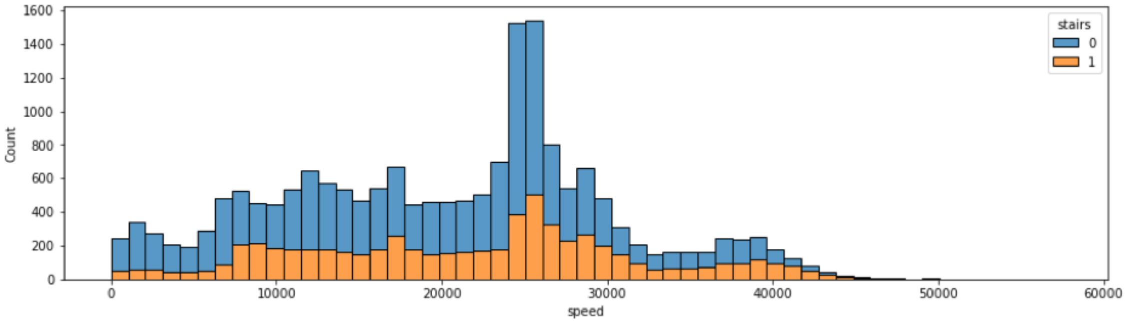


Figure 22: Histogram of speed variable conditioned on decision outcome. This histogram is made after the path selection and does not contain substantial noise.

equal to stairs.

To begin with, we first look into the categorical parameters and their distribution. We have many data points in each categorical variable to look at these features and compare their distributions. For categorical data, we can draw a boxplot to visualize the difference in distribution; see Figure 23. From the boxplots, the distribution of the population taking stairs has a higher median compared to the population taking escalators in each variable. For morning peak hours, we see the highest difference: there are many more people taking stairs than escalators. Because all variables' distributions differ based on stairs and escalators, this advises that these peak variables might affect people's decision outcomes. However, we will need further statistical analysis to examine it more closely.

We also visualize the conditional distributions of numeric interval variables, see Figure 24. Firstly, we notice that all parameters have a non-normal distribution; only speed and speed1 show a closer-to-normal distribution. Some significant peaks are fascinating in an individual's measurements, such as distance and duration. In the distribution of distance variable, we see a highly significant peak at around 40000 of the distance measurement. It could occur due to the fixed physical distance leading from the train. Perhaps some paths are more frequent than others and have a particular distance that measures around 40000 units of length. We also observe a couple of substantial peaks at stamps of 20 and 30 seconds for individual duration measurement. Such peaks can occur due to the crowd effect. For instance, if people are exiting the same train and moving along the same path, it is likely that they will spend the same time on the platform and perhaps then distance. Nevertheless, it is beneficial to investigate these points closer because it might show an undiscovered type of error that occurs in the data or sensor instability. These random-looking peaks suggest anomalies or an interesting indication of patterns in people's behaviour.

Moreover, looking at the conditional distributions for each parameter, a similar variability becomes highly

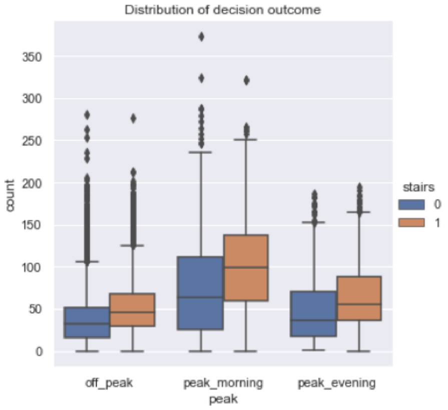


Figure 23: Distribution of categorical variables

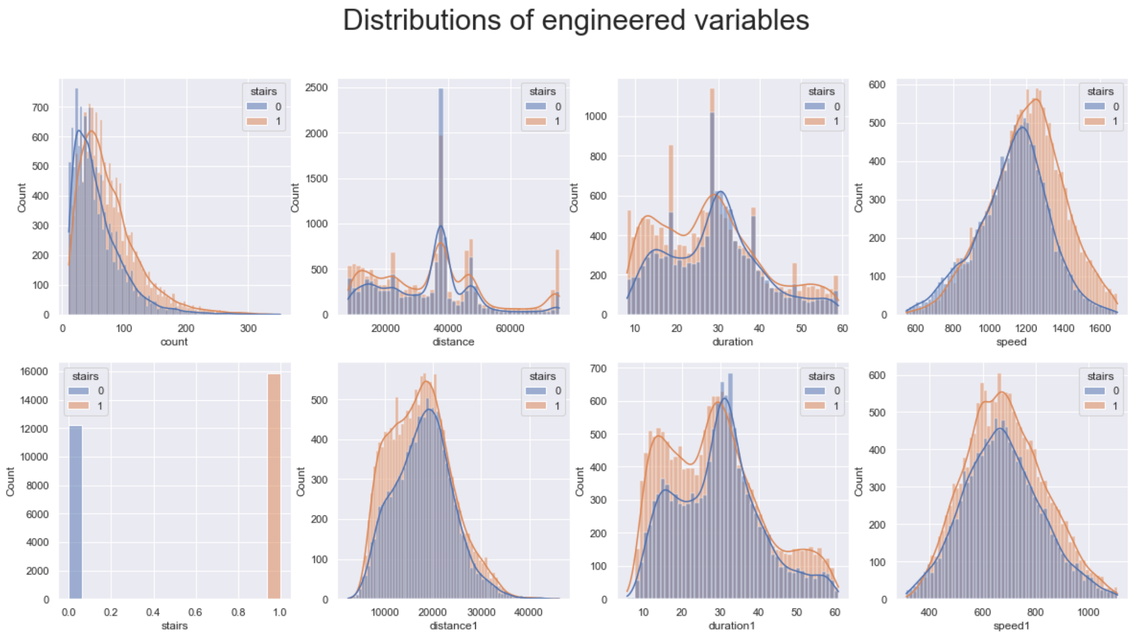


Figure 24: Distributions of numeric variables

noticeable. For example, we see that distributions of the population taking stairs and taking escalators are similarly shaped. The shape of the distribution determines the hypothesis testing methods, which are usually better performed on normal distributions. Because we do not have normal distributions, we will look into tests that do not require it. Generally, tests are performed to validate hypothesis and is considered essential in causal analysis and argument validity.

4.6 Interactions

Besides each distribution of each parameter, it is important to look into the interactions between the variables. For the categorical variables, we use Jaccard's coefficient, and for numeric variables, we use Pearson's Correlation coefficient. This section will overview the interactions between the subject targets.

4.6.1 Jaccard's coefficient

In the research, we use a few statistical methods. Firstly, we use Jaccard's coefficient to determine the interaction between the engineered categorical variables. The Jaccard coefficient measures the similarity between two binary data sets to see which members are shared and distinct. The Jaccard similarity is calculated by dividing the number of observations into both sets by the number of comments. In other words, the Jaccard similarity can be computed as the size of the intersection divided by the size of the union of two sets A and B (Karabiber), mathematically expressed as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

In Table 5, we see Jaccard's coefficients calculated between the categorical variables. The highest similarity with the stairs variable is the off_peak (the table cells are in blue). Off_peak has the most correlation with stairs and thus could explain more variance within stairs and be a better explanatory variable and predictor. We also see some similarity with other peak variables like peak_evening and peak_morning, but it is less significant.

Jaccard's coefficient				
	stairs	off_peak	peak_morning	peak_evening
stairs	1.00	0.41	0.22	0.11
off_peak	0.41	1.00	0.0	0.0
peak_morning	0.22	0.0	1.00	0.0
peak_evening	0.11	0.0	0.0	1.00

Table 5: Jaccard's similarity coefficient between categorical variables

Opposing the degree of similarity, we also must take dissimilarity into account. In Table 5, some interactions are computed as 0, such as peak_evening - off_peak and peak_morning - off_peak. It suggests that they are entirely dissimilar. However, it could indicate that these pairs are opposite versions of each other. It would mean that they are dissimilar, but in reality, they could be just the opposite variable, which ultimately expresses the same information. For this, we can also use Pearson's correlation, which we will discuss in the following part.

4.6.2 Pearson's correlation

To determine the relationship between two continuous variables, we will use Pearson Correlation coefficients. It is based on the covariance method and gives information about the magnitude of the association, or correlation, and the direction of the relationship (Turney [2022]). The formula uses observed values (x, y) and their means ($\bar{x}$, $\bar{y}$) as follows:

$$correlation(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2(y_i - \bar{y})^2}}.$$

Correlation scores can also indicate where multicollinearity may occur. Multicollinearity is a statistical concept in which several independent model variables are correlated. A high correlation between independent variables can result in less reliable statistical results. Hence, we would like to avoid highly overlapping interactions when choosing the variables.

Consequently, we compute and display the correlations in Figure 25. The interactions between the engineered variables indicate some high positive correlations and some high negative correlations that should be discussed. We can notice six high positive correlations between duration and distance variables. Such pairs include duration-duration1, distance-distance1 and interactions of distance and duration variables. The longer the duration, we can expect the longer the distance for both individual and crowd measurements. When using these variables, we would like to ensure reduced multicollinearity. Instead of all four, we would include only a couple of variables. Standard practice would be to pick two or one out of four distance or duration variables.

In addition, there are high negative interactions considering crowd speed and off-peak variables. The crowd speed negatively correlates with individual distance and individual and crowd duration. It suggests that the higher the distance and longer the time, the smaller the average crowd speed. It could suggest that people waiting longer on the platform would be less in a rush and stand still, which would reduce the average crowd speed. From the correlations figure, we can also draw that larger crowds lead to lower average speed. It can be attributed to the crowd effect since having more people on the platform would encourage the crowd to slow down and move uniformly. Our data also shows that crowd speed has more substantial correlations with distance and duration than individual speed. It means that crowd speed is more connected and predictable than individual speed. Individual speed is more independent of other variables.

The same Figure 25 is also valid for looking into the relationship between continuous and categorical variables. We use the point biserial correlation coefficient for such relationships, which is mathematically equivalent to Pearson's correlation. Hence, in the application, we use the same Pearson's correlations between categorical variables. Highly significant negative interaction could be seen in the relationship between off-peak and both peak-morning and peak-evening variables. According to Jaccard's coefficients' results, off-peak variables showed complete dissimilarity with the peak hours variables. Connecting to very high negative correlations, it is now

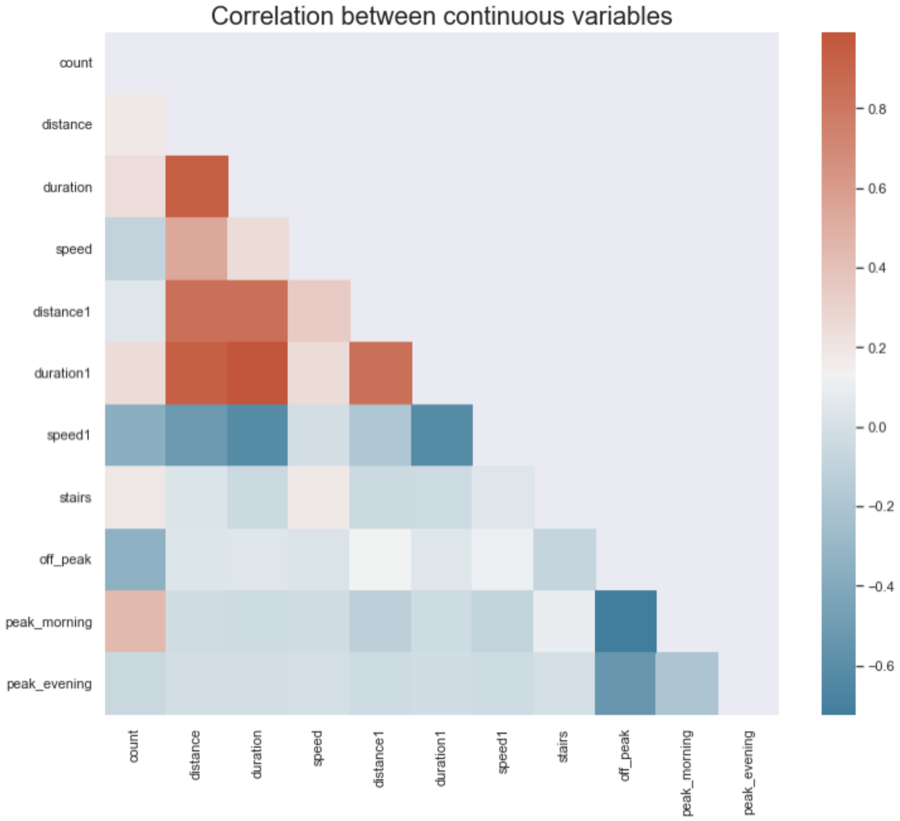


Figure 25: Correlations between engineered variables

evident that off-peak is an opposite variable to peak morning and evening hours. Hence, the off-peak variable can be seen as a negative version of peak variables, which should be taken into account to avoid multicollinearity. It is better to avoid using off-peak and peak morning variables together in a statistical model.

Formerly we wondered how much peak variables are associated with the crowd size. From the correlations, we also could see that count is negatively related to off-peak hours and slightly positively related to morning peak hours. This information shows some evidence that people travel more during morning peak hours and less during off-peak hours. Therefore, if crowd size significantly impacts decision-making, peak variables may reflect the same information. Consequently, morning peak hours and off-peak hours might explain some people's decision outcome only because it reflects the crowd size. However, this requires more evidence.

4.6.3 Distributional interactions

For peak variables, Jaccard's coefficient and the correlation coefficient are enough to understand the relationship because the values are binary. Continuous variables offer distributional information. We can also investigate the correlations closer and look for patterns conditioned on decision outcomes for continuous numeric variables.

In Figure 26, we see each continuous variable relationship distributed based on decision outcome, which corresponds to correlations calculated above. In relationships where we notice high correlations, we observe relatively linear relationships, such as duration1 - duration and duration - distance. Where we found a negative correlation, we see an L-shaped distribution that defines a highly negative correlation, such as duration1 - speed1. For most variables, we see non-linear relations. It suggests that some variables could profit from statistical transformations when using the data for prediction or causal investigation; however, this is outside this research's scope.

Nonetheless, these relational scatter plots also show conditional relationships with the stairs variable used as a hue. However, we can not indicate any differences when looking into relations between variables having people taking stairs or escalators as an outcome. It suggests that most connections are more complicated than linear. Moreover, the information depicted in Table 5 and Figure 26 shows highly close relationships between some factors that we may want to use as regressors for the logarithmic regression on stairs as a target variable. Thus, it can help us filter out some parameters, so we avoid multicollinearity in our model.

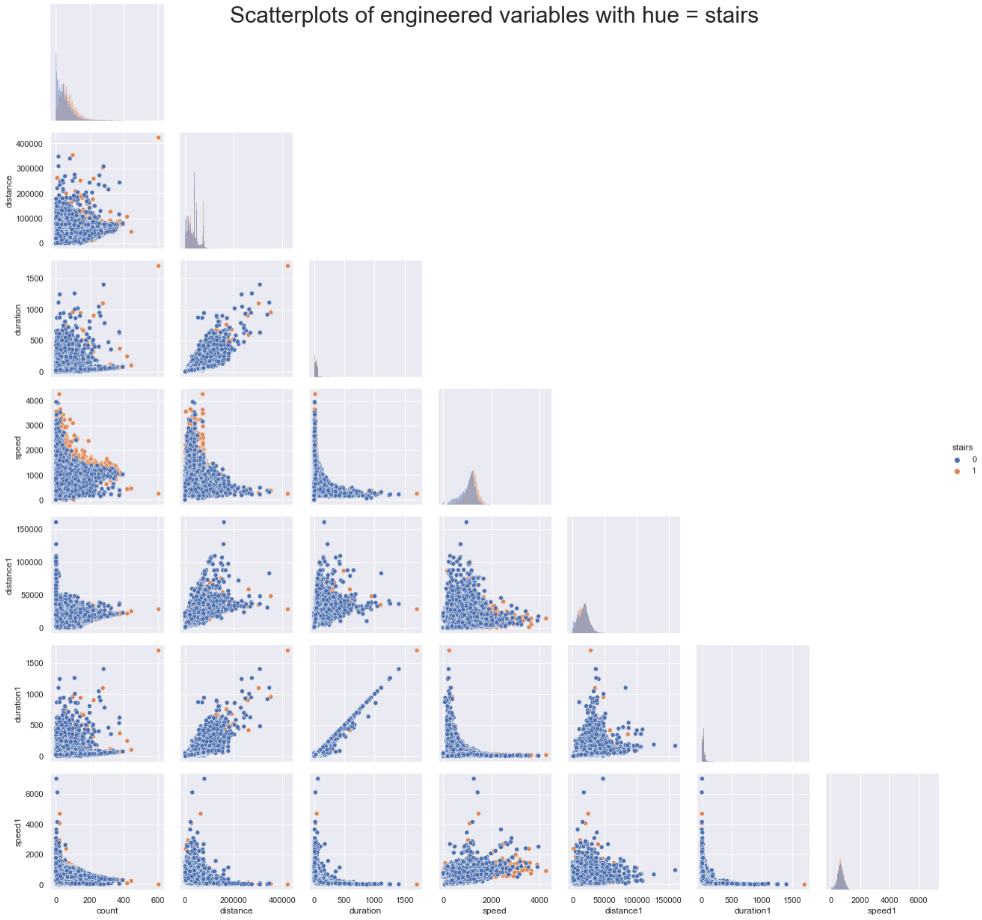


Figure 26: Interactions conditioned on decision outcome

4.7 Dispersion

Another useful metric that can be handy is the dispersion coefficient. It firstly measures the variation of the items among themselves, and second, it measures the variation around the average. If the difference between the value and average is high, then dispersion will be high. Higher dispersion means lower accuracy. Measuring variation can help us understand how reliable the regressors may be. For dispersion calculation, it is common to use the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean. The higher the coefficient of variation, the greater the level of dispersion around the mean.

Coefficient of variation		
Variable	CV when $y = 0$	CV when $y = 1$
count	69%	63%
distance	44%	51%
duration	38%	45%
speed	17%	17%
distance1	33%	37%
duration1	38%	44%
speed1	21%	21%

Table 6: Coefficient of variation for numeric variables

Table 6 shows the coefficient of variation in percentage in the engineered data. We first see that all coefficients are less than 100%, which means that all variables are quite consistent. We also see that the samples' conditional variation coefficient is similar when the decision outcome is stairs ($y = 1$) versus escalators ($y = 0$). The coefficient of variation between these samples differs by 7% at most. Generally, for the sample where the decision outcome was to take escalators, we see smaller coefficients, suggesting a little higher measurement accuracy for this sample.

To sum up, the variable count is the least consistent around the mean, while speed variables show less variation around the mean. It indicates that some variables are more consistent explanatory and predictive variables. Therefore, we should take into consideration these variation coefficients when concluding descriptive and predictive power so we have a better idea of our results' accuracy and reliability. But we have to first understand the relevance of those variables for the outcome, for which we make causal analyses that are explained in the next section of this report.

5 Causal analysis

After analysing the separate variables, we will now look at the causal analysis. One of the statistical methods is the difference in sample estimation, which concludes whether the samples that differ in decision outcome are statistically different. We used Mann-Whitney U and Pearson's chi-squared tests. These tests aim to estimate whether two samples are derived from the same population or, in other words, whether both samples have the same shape. For the main method to estimate factor effect for decision outcome, we use logistic regression analysis. These statistical analyses will help answer our research question.

5.1 Differences-in-samples estimation

The mean difference, or difference in samples, measures the absolute difference between the mean values in two groups. It gives an idea of how much difference there is between the two groups. To test the statistical difference between two samples, we use hypothesis testing for differences between the samples (Glen). We saw in the previous section that our samples are not normally distributed. In such cases, a classical method suggests applying transformations to our data; however, as discussed in the Preprocessing chapter, we will not use it. As a result, we select a statistical test that is allowed for non-normal distributions. For continuous variables, we will use the Mann-Whitney test. For categorical variables, we will use Pearson's chi-square test.

5.1.1 Mann-Whitney U test

Mann-Whitney's test is considered a non-parametric equivalent to the traditional two-sample independent t-test. In statistics, non-parametric (or distribution-free) tests are techniques of statistical analysis that do not require a particular distribution. It is applied when the distributions are not transformed to be normal. It calculates the test statistic based on sample size (n) and the rank sum according to Fay and Proschan [2010], as shown here:

$$U_{stat} = ranksum - \frac{n(n-1)}{2}.$$

However, before applying the tests, we have to check for assumptions for the Mann-Whitney U Test, which are more liberal than standard t-test assumptions. The first assumption suggests that the analysed variable must be ordinal or continuous. The other assumption states sample independence, and the last one suggests that the shapes of the distributions of the two samples are roughly the same. In our case, first of all, we have both numeric and categorical variables. Therefore, we will choose to apply this test to our continuous variables, which are speed, distance, duration, and crowd size, considering it as an ordinal variable. Secondly, the sample independence assumption holds by the definition of our data. Sample shape similarity assumption holds as well because it was shown in the discussion about sample distributions. Clearly, the Mann-Whitney U test is suitable for our variables (Fay and Proschan [2010]).

In our case, a Mann-Whitney U test is performed as a two-sided test, with the hypothesis given:

- Null hypothesis: the mean of the first sample is equal to the second one.
- Alternative hypothesis: the mean of the first sample is unequal to the second one.

To determine significance, we will use a classical method of the p-value. This means that in case our p-value is less than 0.05, we will reject the null hypothesis.

Mann-Whitney U test			
X	E(X|class = 0)	E(X|class = 1)	Significance
count	57.964	75.219	***
distance	33297.06	34299.61	
duration	29.136	27.932	***
speed	1125.815	1201.499	***
distance1	18035.31	17397.24	***
duration1	29.433	28.406	***
speed1	675.181	687.075	***

Table 7: Mann-Whitney U test results for numeric variables

We performed the Mann-Whitney U test for each variable, and in Table 7, we provide a conditional means and significance level for each variable. From the table, we see that we do not reject the hypothesis only for the distance variable. For the rest of the variables, we reject the null hypothesis. For most variables, we collect evidence that samples are statistically different from each other. Furthermore, it shows that these variables have different values when comparing people who took stairs and escalators. This may suggest that these variables are significant in determining whether people took stairs or escalators and can be seen causally connected to the target.

5.1.2 Pearson's chi squared test

For categorical variables, we use Pearson's chi-square test that calculates the scores using the observed and expected frequencies(Swinscow [1997]). It can be calculated by:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

For this test, we have some assumptions that need to be validated. Firstly, we need to validate that our variables are categorical, evident from the binary definition. Another assumption is about the observation independence that holds the same way as we discussed for the Mann-Whitney U test. The third assumption is that the cells in the contingency table are mutually exclusive, suggesting that an individual cannot belong to more than one cell. The person can travel either during peak hours in the morning, in the evening or outside those hours that would be off-peak hours. Thus, there is mutual exclusivity. The last assumption suggests that the sample should be large enough and most values should be greater than 5. Our sample is significantly large, and Table 8 illustrates that.

Contigency table			
stairs	off_peak	peak_evening	peak_morning
0	16442	4337	4915
1	8206	2193	4415

Table 8: Contigency table on categorical variables

In our case, we performed Pearson's chi-squared test using the defined hypothesis:

- Null hypothesis: no relationship exists between the categorical variables; they are independent.
- Alternative hypothesis: there is a statistically significant relationship between categorical variables.

To determine significance, again, we will use a classical method of the p-value with a 0.05 rejection threshold.

We perform the chi-square test between the stairs and peak variables and get the test statistic and p-values in Table 9. We see that the peak_evening variable has a p-value larger than our defined significance level, thus confirming the null hypothesis. It means that the peak_evening variable does not have a significant difference between the two samples. The other variables, off_peak and peak_morning, have very small p-values leading to null hypothesis rejection. It suggests a significant relationship between the stairs and the two variables, suggesting that peak hours in the morning and off-peak hours may affect people's decision whether they take stairs or escalators.

Pearson's chi-squared test			
X	Chi-squared	p-value	Significance
off_peak	178.39	2.2e-16	***
peak_evening	1.4052	0.23597	
peak_morning	270.87	2.2e-16	***

Table 9: Pearson's chi-squared test results for categorical variables

Differences-in-sample estimations offer us statistical insights into relationships between the stairs variable and other engineered factors. We performed the Mann-Whitney U test on numeric data and Pearson's chi-squared on categorical data. It was found that most variables show statistical differences when comparing populations with binary decision outcomes. Only individual distance and peak_evening showed that the samples are not different. This indicates factor difference when comparing people who took stairs versus those who took escalators. It means that besides individual distance and evening peak hours, every variable significantly differs when comparing two populations. Yet it is not enough to argue for causal relationships. Hence we use logistic regression for the following discussion aspect.

5.2 Logistic regression analysis

As one of the statistical analysis methods, we use logistic regression, which is also often used for classification and predictive analytics. Logistic regression usually examines the relationship of a binary (or dichotomous) outcome with one or more predictors, which may be either categorical or continuous. In our case, the dependent or the outcome variable is stairs which reference whether a person took stairs or escalators. The other variables that we investigated before are the independent variables, which are named regressors or predictors. We will start this section by discussing the theory behind logistic regression and the assumptions required for this statistical method. The following step will be the application of logistic regression to identify the effect size of each engineered variable.

5.2.1 Introduction to logistic regression

Logistic regression estimates the probability of an event occurring based on a given dataset of independent variables (IBM). This logistic function is also called log odds and is represented by the formula in Figure 27, where z represents the regression equation. This regression equation includes an intercept, independent variables and weights that are computed for each independent variable. The graph in Figure 27 depicts logistic regression as a separation function between two classes of points. We will use logistic regression to identify the effect size of each variable and will make a predictive model with our engineered variables.

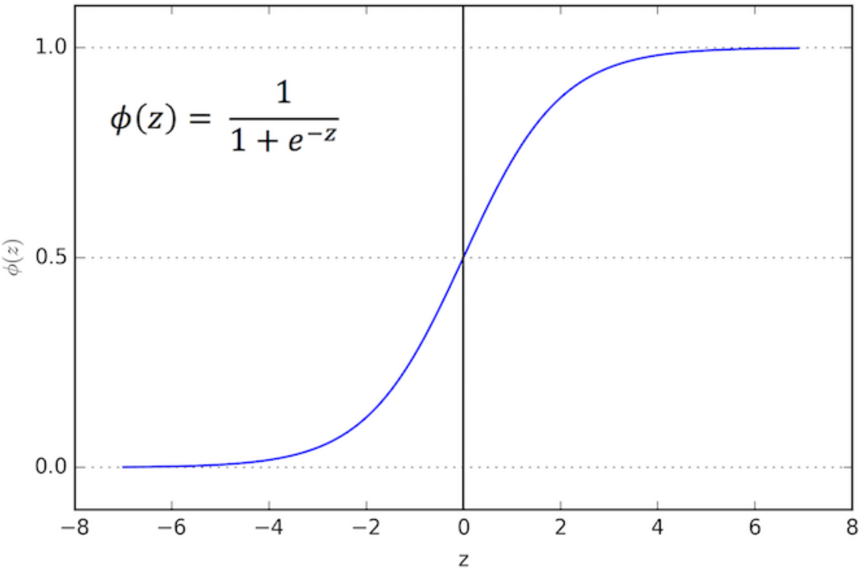


Figure 27: Logistic regression

The logistic regression algorithm works by first calculating the baseline odds of having the outcome versus not having the outcome without using any predictor. This gives us the constant, also known as the intercept. Then, the chosen independent predictor variables are entered into the model, and we calculate the regression coefficients, showing the effect size (LaValley [2008]). Our goal is to select the significant variables and use them for prediction. The conventional technique is first to run the univariate analyses, so each predictor is regressed on the stairs variable at the time to see whether there is a significant effect. This technique has a downside that it does not take interactions between the variables into account. The alternative technique is to load all the variables simultaneously and compute coefficients, performing multivariate analysis. Then we can see which variables are significant given the whole set of variables. Some variables can show significance only in the context of other variables, which is harnessed using this technique.

Logistic regression uses hypothesis testing, with:

- Null hypothesis: the given regressor(s) has(have) no significant effect on the target variable.
- Alternative hypothesis: there variable(s) have a significant effect on the outcome.

To determine significance, we will use a threshold of 0.05 for the coefficient. We will reject the null hypothesis when the coefficient is higher than the significance level.

The key to a successful logistic regression model is to choose the correct variables to enter into the model. Too many variables could lead to larger standard errors and imprecise confidence intervals, which makes our results less reliable. Consequently, our goal is to include fewer variables to obtain significant results (Ranganathan et al. [2017]). After having discussed how logistic regression works, it is essential to validate the assumptions required for this model.

5.2.2 Logistic regression assumptions

Now we will explain all six logistic regression assumptions and argue for the model's validity.

The first assumption is the main one, stating that the outcome variable is binary or multinomial. In our case, it is easy to argue since our target variable is binary, having a value of 0 for escalators and 1 for stairs. The following assumption is meant to verify the independence of observations. Randomisation is one of the critical assumptions for most statistical methods because it removes interdependence and influence between variables, making it easier to compare them. Our data, by nature, is random because we have a unique situation for each data point, and they do not influence each other.

The third assumption is about the number of observations for each independent variable in the dataset to avoid creating an overfit model. Our dataset contains around 28000 uniformly engineered data points, which confirms a large sample size.

Another assumption states no multicollinearity, which corresponds to a situation where the data contain highly correlated independent variables. It is a problem because it reduces the precision of the estimated coefficients, which weakens the statistical power of the logistic regression model. We already looked into the features that caused the most multicollinearity and discussed what features we should exclude. It would help remove the perfect correlation.

Box-Tidwell test				
	coef	std err	z	p-value
duration	0.1569	0.292	0.538	0.591
duration1	-0.1238	0.230	-0.538	0.591
distance	-0.0011	0.000	-2.618	0.009
distance1	0.0008	0.001	1.432	0.152
speed	0.0023	0.010	0.228	0.819
speed1	0.0233	0.014	1.654	0.098
duration:Log_duration	-0.0414	0.060	-0.690	0.490
duration1:Log_duration1	0.0445	0.049	0.907	0.364
distance:Log_distance	9.463e-05	3.47e-05	2.727	0.006
distance1:Log_distance1	-7.738e-05	4.78e-05	-1.619	0.105
speed:Log_speed	0.0002	0.001	0.158	0.874
speed1:Log_speed1	-0.0031	0.002	-1.764	0.078
const	-3.8867	2.900	-1.340	0.180

Table 10: Box-Tidwell test result for continuous variables

Furthermore, we have another assumption that is more complicated than previous ones. It talks about the relationship between the logit (aka log-odds) of the outcome, and each continuous independent variable is linear. The logit is the logarithm of the odds ratio, where p = probability of a positive outcome. To check this, we use a box-Tidwell test that verifies for a linear relationship between the log odds and continuous independent outcome. If the p-value is very small, we have to reject the null hypothesis and conclude that the relationship is non-linear. It would encourage changing the model, transforming the variables, or using another model. As an example, we made a model with all numeric variables. Table 10 shows the results of the Box-Tidwell test. We should look at the p-values of rows that show an interaction between the variable and its logarithmic transformation. Having a p-value of 0.05, we see that most factors have a higher p-value. Therefore, we accept the null hypothesis, stating that duration, duration1, distance1, speed and speed1 have a linear relationship with the logit of the outcome. For distance1, we have a low p-value which shows a non-linear relationship. This means that we should either remove this variable or transform it.

Lastly, logistic regression assumes that there are no highly influential outlier data points, as they distort the outcome and accuracy of the model. We can use Cook's Distance to determine the influence of a data point. It is calculated based on residual and leverage. It summarizes the changes in the regression model when that particular observation is removed. We use standardized residuals to determine whether a data point is an outlier. We can identify the strongly influential outlier data points by finding the top observations based on thresholds and standardized residuals. Having our data, we looked into the influence of the data points and found some visible outliers, constituting around 4% of the dataset (Figure 28). It is important to keep in mind that we already filtered out many data points and finally reduced the data that we used. We could remove more outliers or transform them, reducing the variance and increasing consistency. However, by removing these outliers, we are also introducing higher bias, which leads to results relevant only to some parts of the population. Since the outliers take up less than 5% we can assume that there are not many outliers, and these points do not change the model results as much compared to the other points.

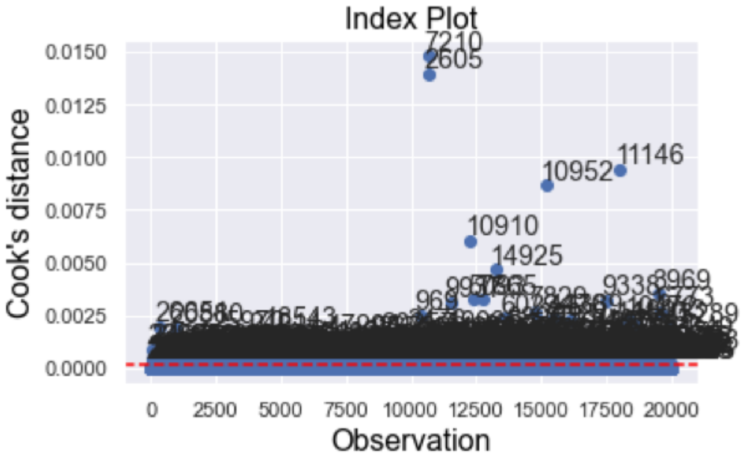


Figure 28: Influencial outliers based on Cook's distance

We looked through the assumptions, discussed them briefly and witnessed that using logistic regression is valid for our data. Before applying Logistic regression on different sets of variables, it was important to assess what requirements should be considered.

5.2.3 Standardisation

Before performing the logistic regression and comparing the results between the variables, we must apply standardization. It is preferred when we want to compare values with different measurement units. In our case, we want to compare the count of people, duration in seconds, distance in custom data units and speed. Unlike transformations, standardization does not change the distribution and maintains the same information in data. It only makes the units comparable. We apply a standardization technique of removing the mean and scaling to unit variance to all numeric features, such as:

$$z = \frac{(x - \mu)}{\sigma}$$

After this process, every factor has an unchanged distribution but a standard and comparable scale, where the mean equals 0 and the standard deviation is 1. Consequently, we can compare the results and use the features in the regression, avoiding difficult interpretation.

5.2.4 Logistic regression results

Moreover, we explained the logistic regression and prepared the variables ultimately for the logistic regression application. Based on regression methods, we performed univariate logistic regression analysis and multivariate analysis.

Table 11 shows the results of univariate regression. We see that only the evening peak hours are in accordance with the null hypothesis, suggesting that it does not influence the target variable. We find these results in agreement with Pearson's chi-square test, confirming the insignificance of this variable. All of the other features are significant in explaining variable stairs. In Table 12, we see the results of multivariate regression made with all the variables simultaneously. We see that, unlike univariate analysis, it shows insignificance for morning peak hours. After removing morning peak hours, the rest of the variables stayed significant. Besides, logistic regression gives us insights into effect size and direction. This can help us indicate what factors influence people to take stairs and what factors are related to higher stairs usage.

From Tables 11 and 12, we can also distinguish the effect size and direction in each independent variable. The effect size is defined by the coefficient that is associated with the variable.

Univariate logistic regression analysis			
Variable X	Intercept	Coefficient	significance
count	0.282	0.427	***
distance	0.263	0.061	***
duration	0.263	-0.099	***
speed	0.271	0.382	***
distance1	0.263	-0.101	***
duration1	0.263	-0.085	***
speed1	0.263	0.080	***
peak_evening	0.268	-0.043	
peak_morning	0.161	0.501	***
off_peak	0.493	-0.344	***

Table 11: Univariate logistic regression test results

Multivariate logistic regression analysis		
Variable X	Coefficient	Significance
count	0.620	***
distance	1.129	***
duration	-1.602	***
speed	0.235	***
distance1	-0.571	***
duration1	0.975	***
speed1	0.399	***
peak_evening	0.126	***
peak_morning	-0.007	
off_peak	0.091	***

Table 12: Multivariate logistic regression test results

To begin with, both regressions show negative effects for duration and distance1. They are suggesting that these variables, in general, are associated with escalators. It means the higher the individual duration and crowd distance, the more likely people would take escalators. If people spend a long time on the platform, it is more likely that they will take the escalator. Alternatively, individuals in crowds that take relatively longer paths could also be an indication against stairs usage. These results are consistent, and they will lead towards the choice of comfort, which is taking the escalators. From univariate regression, crowd duration and off-peak hours would encourage people to take escalators. Crowd duration is very similar to individual duration. Hence, it is easy to argue that both variables affect the decision in the same way. According to correlations, we saw that off-peak hours are negatively correlated with the stairs variable and with the crowd size. It is credible that people avoid stairs when less crowding is present, and a more convenient means of descent is available. From univariate regression, we see that more variables are associated with escalators compared to multivariate regression. Multivariate regression, however, does not indicate any other significant variables that would be associated with the escalators' usage. This means that in the context of other variables, only crowd distance and individual duration have a positive association with escalators.

Besides negative effects, we also see positively influential factors. Firstly, we see that count, individual distance and speed, as well as crowd speed, are associated with stairs usage in both regressions. If there are more people exiting the platform at the same moment, people will be more willing to take the stairs. Likewise, with speed, people that are walking faster are more likely to take stairs. The longer the individual distance, the more it is associated with the willingness to choose to exit through stairs. Uniquely for univariate logistic regression, we see significant positive effects of peak morning hours. It is an inverse result from off-peak hours, which is as expected given they are negatively correlated (as discussed in Figure 25). When we look at multivariate logistic regression, we see more variables that have a positive effect on people taking stairs. Such variables are average crowd duration, evening peak and off-peak hours. In a previous chapter, we saw that individual duration has a perfect correlation with crowd duration (due to technical computations). Therefore, it is not credible that average crowd duration would have a positive effect towards stairs while individual duration has a negative effect. For evening and off-peak hours, we can not find convincing arguments. Mainly positive effects are due to interactions with other variables.

In the context of many variables, we see that peak variables and crowd duration are inconsistent and depend on what other variables are used together in a model. This can happen due to multicollinearity because some variables are almost linear transformations of others; the model tries to balance out the effects. If one variable has a small effect in univariate analysis, it might change the direction of the effect in the context of other stronger predictors.

From the logistic regression analysis, we can say that individual duration and crowd distance have a positive effect on people taking escalators. Comparing these variables, individual duration has a much larger negative effect, estimated to be around -1.6, while crowd distance is around -0.6 (Table 12). Variables that encourage people to choose stairs are crowd size, individual distance and individual and crowd speed. Comparing these four factors, we see that distance has the largest effect size, around 1.1. The smallest effect size is shown by individual speed, around 0.2.

5.3 Causal analysis conclusion

In a nutshell, to determine causal effects, we used sample difference estimation and logistic regression analysis. The difference in samples showed that distance and evening peak hours were the two variables that did not show a difference when comparing two populations with different decision outcomes. However, in logistic regression, we saw different results. Logistic regression is a powerful tool that was used to determine the engineered variables' effect size on the decision outcome. We found that individual duration and average crowd distance when an individual is on the platform showed the highest associations with escalator usage. On the other hand, we found that larger crowd size influenced more people towards choosing stairs. Longer individual distances also lead to takings stairs, which might occur because of huge crowds concentrated around the exit and blocking the route to the escalators. We also consistently observed that higher individual speed was associated with the choice to take stairs, which is interesting because, generally higher crowd lead to a slower crowd. These variables take into account the interactions with other variables, but some of them show inconsistent results, leading to a less clear dependency on decision-making.

After causal analysis, we found features that have the most validity based on logistic regression. In the following section, we will try different feature selection methods and compare them to our causal analysis. The sets of selected features will be used for prediction and best model selection. Further on, we will come back to logistic regression as a predictive model in the later chapter.

6 Feature selection

Feature selection is important when making a prediction model because we want to select the set of variables with the highest prediction score. Logistic regression is one of the methods for feature selection. It helps us find the most indicative variables based on their effect size on the dependent outcome variable. Logistic regression looks at causal effects and helps us argue which variables are the most influential for the outcome. Besides logistic regression, there are other methods that are focused on variable predictive performance rather than on effect size. Depending on the data, variables with smaller effect sizes could potentially predict the outcome better, or a certain group of variables achieve the best results. This chapter will look into other feature selection techniques such as correlations, Fisher's test and knowledge gain.

6.1 Correlation

As previously discussed, correlations have a high application power for variable interaction analysis. Correlation analysis is a common way to select the best features. According to this method, the higher correlation, the higher the explanatory variance we observe.

We measure our independent variables in comparison with the decision outcome variable and rank them in Figure 29. Considering the threshold at 0.05, we select significant variables in descending order, such as count, speed, peak_morning, off_peak, crowd distance and duration. The other variables are regarded as inconsequential. The correlation method is easy to apply but does not account for effect size and multicollinearity. Besides the primary correlation method, we look into other techniques to calculate meaningful variables.

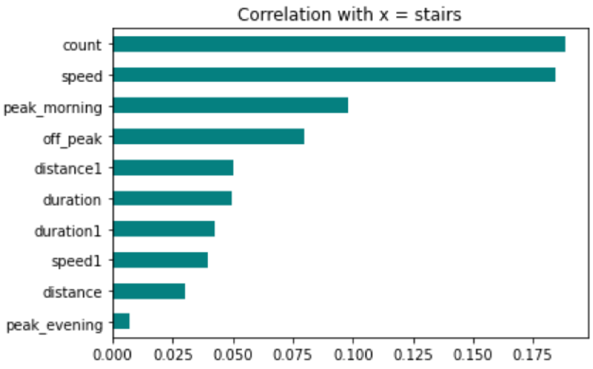


Figure 29: Correlations between the outcome and each variable

6.2 Fisher's test

We use Fisher's score as one of the most widely used supervised selection methods(Gu et al. [2011]). Supervised feature selection refers to the method which uses the output label class for feature selection. For Fisher's score, the algorithm computes the Fisher score for each feature. It is computed using a ratio of between-class and within-class variance. The algorithm helps select and rank the variables with the largest Fisher scores.

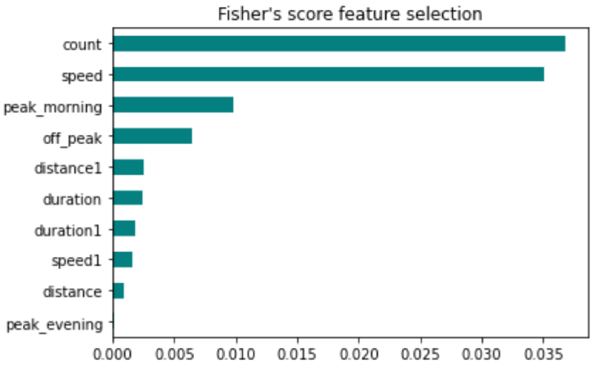


Figure 30: Feature selection using Fisher's score

In our case, we calculated the scores for each variable and displayed the results in Figure 30. Using a significance level of 0.005, we can select count, speed, morning peak and off-peak hours. These variables are very similar to correlation selection.

However, this method also has disadvantages. Because the score of each feature is computed independently, the selected features are suboptimal and can be used only as a feature-filtering technique. More importantly, the heuristic (or nearly optimal) algorithm fails to select features with relatively low individual scores but very high scores when combined. In addition, it cannot handle redundant features because it does account for multicollinearity.

6.3 Knowledge gain

The last technique we used is knowledge gain or mutual information selection. Mutual information comes from the field of information theory and is widely used for feature selection. It is also, by default, applied to select the feature in decision tree making.

Mutual information is calculated using probabilities between two variables. It measures the reduction in uncertainty for one variable given a known value of the other variable. Given the stochastic nature of this evaluation procedure, we have different results every time we calculate the mutual information, which leaves room for more inaccuracies. This main disadvantage makes this algorithm less reliable and harder to interpret(Huijskens [2017]). To minimize inaccuracies in the measurement of knowledge gain, we measured 200 random samples and found average measurements for each factor, displayed in Figure 31. We can use a significance of 0.01, which is used to describe a weak connection to the outcome variable. According to the calculations, mutual information selection extracts speed, distance, count, speed1, distance1 and duration1 as important features.

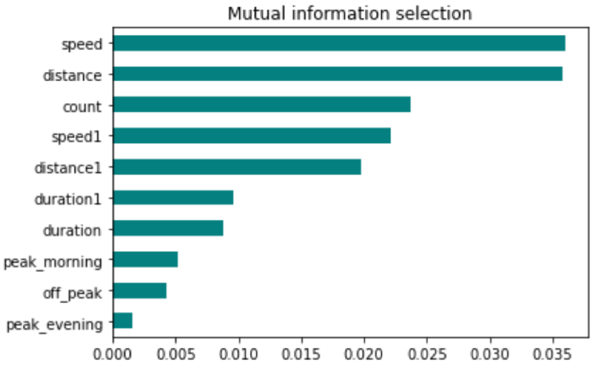


Figure 31: Feature selection using mutual information

6.4 Multicollinearity

We have selected the features using a few different techniques. However, these methods do not consider multicollinearity, which is important to reduce for result reliability. We will use logistic regression for prediction modelling, which requires feature independence. We briefly discussed multicollinearity in the correlations chapter. For this section, we look into the highest correlation to decide how to increase independence among features.

Table 13 shows the ten highest correlations, including feature names and Pearson's correlation score. We see that variables that appear among the highest correlations are duration, duration1 and distance with distance1. The colours map these four distinct features, and easier indicate their appearance. To remove multicollinearity, we need to remove at least two of these four features if we want to remove the highest correlations that are above 0.7. There is no one answer, but a common practice would be to take one distance and one duration variable. Besides duration and distance, we also see that peak_morning and off_peak hours also have a high correlation among themselves, and thus better to take one measure from them if we want to build a reliable model.

6.5 Selected feature overview

Using the feature selection methods, we indicated sets of variables that each algorithm ranked the highest. We also apply multicollinearity reduction, so the features can be used for the model. Eventually, we have eight feature sets, described in Table 14. In addition, some methods filtered variables that are highly correlated together. In such cases, we split the set into two subsets to separate the highly correlated variable. These feature sets will be used in the following chapter to test predictability.

Highest correlations		
Variable 1	Variable 2	Correlation score
duration1	duration	0.991
duration	distance	0.936
duration1	distance	0.931
duration1	distance1	0.849
duration	distance1	0.846
distance	distance1	0.838
peak_morning	off_peak	0.725
duration1	speed1	0.623
duration	speed1	0.614
distance	speed	0.538

Table 13: Top 10 Highest correlations between the engineered variables

Features selection	
Feature set	Description
’count’, ’speed’, ’distance1’, ’duration1’, ’speed1’, ’peak_evening’, ’off_peak’	Feature selection based on multivariate logistic regression using crowd distance and duration
’count’, ’distance’, ’duration’, ’speed’, ’speed1’, ’peak_evening’, ’off_peak’	Feature selection based on multivariate logistic regression using individual distance and duration
’duration’, ’count’, ’distance’, ’speed’, ’speed1’	Feature selection based on both regressions using individual distance
’duration’, ’distance1’, ’count’, ’speed’, ’speed1’	Feature selection based on both regressions using crowd distance
’peak_morning’, ’distance1’, ’duration’, ’speed’, ’count’	Feature selection based on correlation
’count’, ’speed’, ’morning_peak’ and ’off_peak’	Feature selection based on Fisher’s score
’speed’, ’distance’, ’count’, ’speed1’, ’duration1’	Feature selection based on information gain, only individual distance
’speed’, ’distance1’, ’count’, ‘speed1’, ’duration1’	Feature selection based on information gain, only crowd distance

Table 14: Features selection removing multicollinearity

7 Predictive modelling

We have concluded the causal analysis and feature selection. In addition, we test and evaluate how well each set can perform in predicting the outcome. Predictive modelling is a commonly used statistical technique to predict future behaviour. It works by analysing the previous behaviour by fitting a function using a training set and performing on the testing set. We compared the observed values and predicted values to conclude predictive power. This chapter will explain in detail the data split for modelling and evaluation metrics and will compare the models using our selected feature sets.

7.1 Train and test split

For predictive modelling, we use a train-test split in our data. We use 70% of the data for training and the rest 30% for testing, which results in about 20000 training points and 8000 testing points. We first randomise the datapoint order in our data and then perform the splitting to make sure both sets capture similar variance. We then define the prediction target as stairs and the independent variables as predictors. For each model, we use a different feature set and evaluate the performance.

7.2 Evaluation metric

In a classification problem, we predict a categorical outcome, for which we can draw a confusion matrix. The confusion matrix defines how many cases were predicted correctly and how many cases were false positives or false negatives and can be drawn like in Table 15. The blue cells symbolise the well-predicted values, which we aim to maximise.

Confusion matrix		
	STAIRS	ESCALATORS
PREDICTED STAIRS	True positive	False positive
PREDICTED ESCALATORS	False negative	True negative

Table 15: Confusion matrix

We use evaluation based on accuracy and F1 score for this classification problem. Accuracy is a classification metric calculated by adding true positives and negatives and dividing by the sum of all cases, such as:

$$accuracy = \frac{TrueNegative + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}.$$

It shows the proportion of true results among the total number of cases. The F1 score is a number between 0 and 1 and is the mean of precision and recall measurements that describe how well the algorithm classified the true positives, formulated as:

$$F1 = \frac{TruePositive}{TruePositive + \frac{1}{2}(FalsePositive + FalseNegative)}.$$

F1 score and accuracy are evaluation techniques widely used in practice and suitable for our prediction evaluation.

7.3 Predictive power

Logistic regression with features			
Names of features	Method	Accuracy	F1
’count’, ’speed’, ’distance1’, ’duration1’, ’speed1’, ’peak_evening’, ’off_peak’	Multivariate regression	0.659	0.717
’count’, ’distance’, ’duration’, ’speed’, ’speed1’, ’peak_evening’, ’off_peak’	Multivariate regression	0.657	0.721
’duration’, ’count’, ’distance’, ’speed’, ’speed1’	Both regressions	0.659	0.723
’duration’, ’count’, ’distance1’, ’speed’, ’speed1’	Both regressions	0.656	0.715
’duration’, ’count’, ’distance1’, ’speed’, ’peak_morning’	Correlations	0.652	0.713
’count’, ’off_peak’, ’speed’, ’peak_morning’	Fisher’s score	0.620	0.694
’duration1’, ’count’, ’distance’, ’speed’, ’speed1’	Information gain	0.652	0.715
’duration1’, ’count’, ’distance1’, ’speed’, ’speed1’	Information gain	0.661	0.718

Table 16: Logistic regression with selected feature sets

We looked into the engineered variables and their selection based on the previously explained methods. We used logistic regression to predict for each feature set. Table 16 is our overview of the features evaluated by accuracy and F1 score.

Generally, we notice that all feature sets perform similarly. The set with the lowest number of variables (four features) was selected by Fisher's score, which showed lower accuracy and F1 score. Sets with five features (Information gain) performed the same as seven features (Multivariate regression method), with the highest accuracy of 66% and F1 score of around 72%. Finally, we choose the best set of features based on our project. In the table, the feature set is marked in a blue cell and includes count, individual speed and crowd distance, duration and speed. These features perform the best out of the proposed ones.

Since we selected the feature set, we can also compare the predictive techniques. We compare logistic regression performance with well-known machine learning techniques such as Random Forest Classifier and K-nearest neighbour. Random Forest Classifier is a powerful algorithm that works based on decision trees, while the K-Nearest Neighbour classifier finds neighbouring data points based on the smallest difference and tries to cluster the data. We will not delve deep into these methods, but the main message is that these two algorithms are more powerful and more complex than logistic regression. Nevertheless, they are often used in practice because they perform well. Table 17 depicts the results of all three algorithms, which are very similar. Random Forest Classifier yields better results by 2%, but it is clear that the Logistic Regression is a good fit for this data and the chosen variables. More powerful algorithms do not show higher accuracy, which suggests that the predictive power of the chosen feature set is limited. If we want to improve our predictions, we could look at more advanced statistical methods or add more information to the model.

Prediction performance		
Algorithms	Accuracy	F1
LogisticRegression()	0.661	0.718
RandomForestClassifier()	0.688	0.730
KNeighborsClassifier(n_neighbors=5)	0.648	0.696

Table 17: Score of different algorithms with chosen features

7.4 Variance and Bias

Statistical models inherent errors that are reducible and irreducible. Irreducible errors occur due to natural variability in the data. The reducible errors can be controllable and minimized for better accuracy. The reducible error includes bias and variance. Variance expresses how scattered and variant the data is, while bias describes how relevant the results are. If we simplify the data, we remove the variations, but we introduce bias because, for certain variations, the results will be simplified and do not reflect reality.

In our case, we introduced bias when processing the data. It also led to reduced variance, on the other hand. When processing the data, we lost up to 20% of entries when filtering the hourly data files, and roughly 30% of paths that we assumed were irrelevant. The last step refined the data even more by removing 20% of variability. We are left with more focused data, which is simplified and lost some relevant variability. Based on our data, it was an essential part to have meaningful results; however, for later research, it is possible to maintain more variability by modifying the processing steps and using a more complex statistical model.

8 Conclusion

In conclusion, we found some significant evidence that some variables explain people's choice to take stairs, and some variables are associated with escalators. We summarised the findings by estimating the evidence and the credibility and drew a plot seen in Figure 32. These findings will help answer the research questions that we described in the beginning.

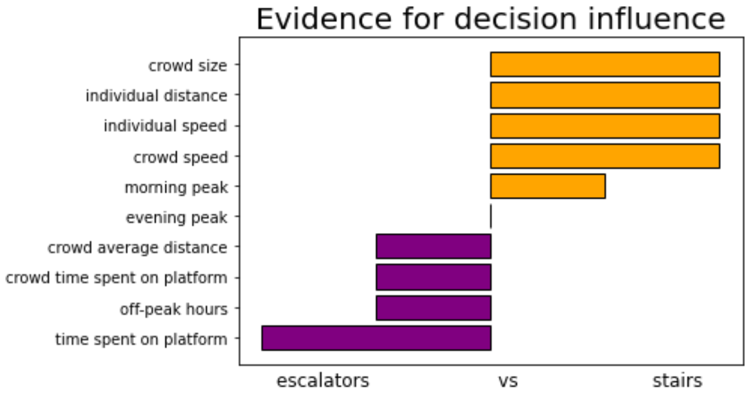


Figure 32: Findings for variable significance

- Can we verify the previous research that crowdedness, peak hours and higher individual speed are linked to higher stairs usage?

We found that larger crowd size and higher individual speed are associated with a higher probability of taking stairs. This verifies the previous research. For morning peak hours, univariate logistic regression results suggested some positive effects, while evening peak hours were found to be insignificant as a sole factor. Multivariate regression showed that in the context of all discussed variables, evening peak hours but also off-peak hours showed a positive effect. Because our analysis showed that morning peak and off-peak variables could be considered the inverse of one another, the inconsistent results suggest that more investigation is needed. There is not enough evidence to conclude that during peak hours, people tend to take the stairs more often. These features did not show high predictive power compared to other features. Morning peak hours yield significant results because the count is always higher, and evening peak hours did not show much crowd difference compared to some off-peak hours, which renders results inconsistent. Thus, it is very likely that peak hours show significance because of crowd size. This could be the next step for peak feature investigation.

- Do factors such as path length and time spent on the platform affect stairs usage when exiting the platform? If yes, to what extent are they influential?

For the individual distance, we found substantial evidence for stairs usage. The longer the distance, the higher the probability of taking stairs. However, we found that longer time spent on the platform increases the likelihood of taking escalators. These two features have a similar effect but in opposite directions. However, these two features have a very high correlation, so the opposite direction evidence could occur due to the fact that these variables balance each other out. However, they offer little explanation for the decision outcome. An idea for further analysis would be to look into these factors separately but more precisely. For instance, measure a more specified group of travellers and narrow down the research variables, like measuring the time spent and distance within the area of the exit.

- Can average crowd speed, crowd distance, and time the crowd spent on the platform indicate an individual's decision-making when exiting?

From the analysis, we found that crowd speed is also related to people taking stairs. It can be associated with the crowd effect. In the presence of large crowds, people tend to adapt to the crowd speed, which makes the crowd speed a relevant variable for decision-making. For crowd distance, we found negative effects, which means that the longer the crowd distance, the higher likelihood for an individual to take escalators. However, crowd distance is highly correlated to individual distance, which is associated with stairs. This ambiguity suggests that there can be some effect, but the crowd distance variable is too correlated with individual measurements to draw strong opposite and independent conclusions. The time that the crowd spent on the platform is also highly correlated with individual duration. However, for both measurements, we saw an influence leaning towards escalator usage. There is not enough evidence to state clear conclusions for crowd duration. We saw that crowd speed has more credible arguments and more independence from other variables; therefore, results are more reliable than for crowd distance and duration.

- What factors can we associate with people taking stairs when exiting from the train's platform?

Generally, looking at Figure 32, we see that factors that influence people to choose stairs are larger crowds, higher individual and surrounding speed and longer individual distance. People that waited longer on the platform seemed to take escalators more. It could be explained by the fact that people choose to wait for the crowd to decrease and sort out when more space is available. Thus, leading to less crowding around the exit and people choosing escalators instead of stairs.

This information that we discovered allowed us to achieve some prediction power. From our data, we saw that using five predictor variables, we can achieve 65% accuracy or 70% prediction score for a population that is taking stairs. These predictions are significant and only create a baseline for a predictive model that contains crowd size, speed, distance, duration and encoded peak hours information. In the future, more analysis could indicate a better specification of these features for the decision-making problem between stairs and escalators.

Our research confirms some of the already researched variables and introduces explanations and measurements for some new but primarily important features. Even more than that, this project posed many new questions and directions for further research. Later steps would involve looking closer into each discussed feature, comparing different stations or other environments and allow to verify even more assumptions.

References

M.P. Fay and M.A. Proschan. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Stat Surv., Author manuscript*, 4:1–39, 2010. doi: 10.1214/09-SS051.

C. Feng, H. Wang, N. Lu, T. Chen, H. He, Y. Lu, and X.M. Tu. Correction: Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry*, 26(2):105–109, 2019. doi: 10.1136/gpsych-2019-100146corr1.

S. Glen. Mean difference / difference in means(md). `https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/mean-difference`.

Q. Gu, Zh. Li, and J. Han. Generalized fisher score for feature selection. *UAI'11: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, page 266–273, 2011.

Th. Huijskens. Mutual information-based feature selection. `https://thuijskens.github.io/2017/10/07/feature-selection/`, 2017.

IBM. What is logistic regression? `https://www.ibm.com/topics/logistic-regression`.

F. Karabiber. Jaccard similarity. `https://www.learndatasci.com/glossary/jaccard-similarity/`.

M.P. LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008. doi: 10.1161/CIRCULATIONAHA.106.682658.

M.Kh.A.M. Lazi and M. Mustafa. Pedestrian route choices between escalator and staircase during descending at masjid jamek terminal. *Malaysian University Transportation Research Conference 2015*, 2015.

Q. Li, Ch. Ji, L. Jia, and Y. Qin. Effect of height on pedestrian route choice between stairs and escalator. *Green Intelligent Transport System*, 2014(965305), 2014. doi: 10.1155/2014/965305.

A. Maeng, R.J. Tanner, and D. Soman. Conservative when crowded: Social crowding and consumer choice. *Journal of Marketing Research*, 50(6), 2013. doi: 10.1509/jmr.12.0118.

P. Ranganathan, C.S. Pramesh, and R. Aggarwal. Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research*, 8(3):148–151, 2017. doi: 10.4103/picr.PICR_87_17.

S. Srikukenthirana, A. Shalabya, and E. Morrowb. Mixed logit model of vertical transport choice in toronto subway stations and application within pedestrian simulation. *Transportation Research Procedia*, 2:624–629, 2014. doi: 10.1016/j.trpro.2014.09.104.

T.D.V. Swinscow. *Statistics at Square One: 8-chi-squared-tests*. BMJ Publishing Group, ninth edition edition, 1997.

S. Turney. Pearson correlation coefficient (r) | guide examples. `https://www.scribbr.com/statistics/pearson-correlation-coefficient/`, 2022.

J.K.K. Yuen, E.W.M. Lee, and W.W.H. Lam. An intelligence-based route choice model for pedestrian flow in a transportation station. *Applied Soft Computing*, 24:31–39, 2012. doi: 10.1016/j.asoc.2014.05.031.