Universidad Francisco de Vitoria
UFV Madrid

SMART CITIES COMPETING FOR TALENT:

NEURAL NETS AND CLUSTERING ALGORITHM FOR CITIES

Student: Marta Meneses Muñoz

Business Analytics & Computer Science

4th Year

DFP

Francisco de Vitoria University

I dedicate it to my parents, for their enormous effort in giving me the best possible education that allows me to fully develop myself. They are the ones who have always supported, helped, and encouraged me.

Acknowledgement

To my family and friends for allowing me the necessary time to carry out this research and analysis. To Juan, for being a fundamental support in these last four years. To the university, for giving me the opportunity to learn about the academic research, from which this project stems, during my internship last year at the Global Observatory of Attractive Cities. In addition, I am grateful for all the tools and knowledge learned there. Thanks to my tutors León Beleña, José María Peláez and, especially, Ana Lazcano, for guiding and advising me during this project. Finally, thanks to José Antonio Ondiviela, for awakening my curiosity and passion for cities and for his confidence in me, as well as for his support throughout this project. I would also like to mention Blanca Herreros de Tejada, my partner and friend with whom I was able to collect all the data I use in this project during my previous internships.

Abstract

The future of cities is based on the 3Ts: technology, talent and tolerance. Talent is a fundamental pillar for the development of cities and therefore we must know how to be an attractive city to attract it. Choosing the place where we can develop our potential and put it at the service of others is what every talented young person wants to do. In a globalized world, the information that we can evaluate to facilitate decision making is increasing. As a result, there is great competition and, just like companies, cities need to attract that talent that will lead them to advance and grow.

Through this study, we will group 175 cities considered Smart Cities, by similarity through an unsupervised model without any human bias or subjective opinion. In the same way, this study will help governors to know in which group of cities they are in the search for talent and to know what they should offer to improve their positioning. Thus, we will be able to foster the desire to make cities the place where the person is the center and attract technology companies that enrich cities, and therefore we will also be considering the first "T", technology.

Keywords

Smart Cities, Attractiveness, technology, clustering, PCA, Autoencoder

Universidad Francisco de Vitoria UFV Madrid

INDEX

Content Index

Annexes

Figures Index

Tables Index

0. INTRODUCTION

According to the UN, by 2050 more than 70% of society will live in cities, and cities will want to have the best human resources to develop, grow and offer citizens the best services. For this reason, all cities in the contemporary world seek to attract talent. According to the RAE, talent refers to "intelligent person or apt for a certain occupation" (RAE, 2014), but we can go a little further, valuing the whole person, knowledge and skills.

In today's companies there is great competition to attract talent and cities must also fight for it. We are going through the Fourth Industrial Revolution and the need to keep up with the times is increasingly competitive, which means having the right people for it.

Smart Cities are cities based on sustainable urban development that, through data exchange, interconnection and innovation, manage to offer better services in terms of governance, economy, mobility, environment, energy, health and safety. All this is possible thanks to technologies such as IoT, algorithms, cloud, big data, artificial intelligence and blockchain. It is important to take care of cities as they are not only organized streets, but also where we breathe, move, live, learn, everything happens in them (ESMARTCITY, 2022).

Attractiveness is something that governments need to consider, as it will be the most attractive cities that attract talent, and it is important that policymakers compete for it. Talent is not about favoring the city economically, but about enriching it in all its aspects. Similarly, we must not lose sight of the first "T", as technology plays a key role. Cities that prosper are those that take care of citizens and their environment, but for this to be possible, technology is necessary (Ondiviela, 2021). Therefore, the attraction of technological companies would also be a key element in the future of cities.

It is important for cities to know where they stand in this competition for talent. Similarly, it is important for companies to know where the talent is in order to bring their delegations to these cities. Therefore, we will also focus on the first "T", technology, because by attracting talent to cities, technology companies will also be attracted. Citizens will benefit from this competitiveness of the cities, as they will be willing to offer them the best opportunities.

1. PROBLEM DESCRIPTION

Previous Research

The focus is increasingly on smart cities. Life expectancy is longer, cities are growing faster and connected cities are needed to manage change efficiently, while trying to do as little damage as possible to the environment. Obviously, all this goes hand in hand with technological breakthroughs.

There are many models based on smart cities and different classifications according to their attractiveness. Evidently they do not usually coincide since it is not possible to establish a universal ranking according to the attractiveness of a city to attract talent. Each person has his or her own preferences and what may be attractive to one may not be attractive to another. In addition, each of these rankings will be biased by those who elaborate them, even by the experts who generate them.

Although the economic factor continues to be of great importance, services, environmental care, innovation and the culture of cities are increasingly valued (Kelly, 2020). As Dan Doctoroff said in an interview (Hong, 2019), cities are always going to be immensely complex human organisms to manage, similar to the challenge of trying to solve a Rubik's cube with 50 faces.

Competitiveness between cities is a reality and the criteria used to decide where to study, work or invest are becoming increasingly broad. The company JLL developed a model with 10 groups of cities according to their role in the world, focused on the opportunities they offer to real estate investors (JLL & The Business of Cities, 2018). These 10 groups are differentiated from each other into four larger groups that contain them: established world cities, new cities, emerging cities, and growing hybrids.

Figure 1: JLL (JLL & The Business of Cities, 2018)

Similarly, IESE Cities in Motion publishes an annual ranking of cities based on a number of indicators, from economic factors to citizen services, tolerance and equality. In the 2020 ranking, we were able to see the effects produced by COVID in these cities. This ranking uses 101 indicators divided into the following dimensions: Human capital, social cohesion, economy, governance, environment and transportation, urban plan, international projection and technology (IESE Cities in Motion, 2020).

In addition, cities are increasingly trying to take new initiatives to improve their services. A study conducted by MDPI that was published in several scientific journals, uses clustering and deep learning models, managing to group the characteristics of a city into six major factors: technology, energy, environment, transportation, government, human capital and quality of life (Parlina , Ramli & Murfi 2021). However, during this year I have been immersed in a research project: UFV's World Observatory of Attractive Cities, thanks to which I have been able to delve deeper into this topic and understand that there are many more factors that define a city.

Given the complexity of cities, this UFV project has tried to extract more than 200 indicators of the attractive city, trying to bring a very complex reality to a simpler model. These indicators can be grouped into magnetism and profitability. This research deals with the attractiveness to attract talent of 175 smart cities around the world, based on a doctoral thesis. This thesis was later published in the book Beyond Smart Cities: Creating the Most Attractive Cities for Talented Citizens by José Antonio Ondiviela, Western Europe Director for Smart Cities Solutions at the large company and technology benchmark Microsoft (Ondiviela, 2021).

Magnetism has the most emotional and subjective components, while profitability has the most rational and objective side of a city. Based on these two main indicators, a ranking of the attractiveness of cities was calculated, but in a personalized way, since each person has his or her priorities and preferences when it comes to attributing weights to the characteristics that define an attractive city for him or her. With this in mind, a mobile application was created (See Annex 7) in which the user could give his own weightings to the indicators of magnetism and profitability. Within magnetism we found the following indicators: identity, dynamism and future, that is, the city's strategy; while profitability was defined by: cost of living indicators, purchasing power and the services that the city offers its citizens, which also influence when describing how attractive a city is (Ondiviela, 2021). The sum of the weights of these indicators returns a personalized ranking for the user with the most attractive cities for him.

However, based on the observatory and as published in the aforementioned book, the model built also aimed to provide a ranking of the attractiveness of cities to attract talent according to the criteria of experts. This ranking was calculated with the sum of the two large groups of indicators mentioned: magnetism and profitability, which were given a weight of 50 and 50 respectively to calculate the attractiveness index. In the absence of knowing which might be more important, it was decided to give equal weight to both. Magnetism was then divided into the indicators of identity, dynamism and strategy, which in turn are made up of several sub-indicators (see Annex 1). The same goes for profitability, where it is divided into 10 city services and cost of living, which in turn have several sub-indicators (See Annex 2).

We could say that this model contains several biases. First, magnetism and profitability are assigned equal percentages according to the justified judgment of a subject matter expert. The weights of the indicators within magnetism and profitability were obtained through a survey of Smart Cities experts. A survey was launched to 21,334 attendees at the 2018 SmartCity Expo & WWW Congress in Barcelona (Spain). These people were considered people with high knowledge in cities. 1550 people responded, so it proved to be a reliable sample with 95% confidence and an error of less than 2%. Therefore, the results were used to calculate the weight of the indicators within magnetism and profitability, creating a model from which a ranking of the most attractive cities was extracted. The weights of the sub-indicators, which make up identity, dynamism, strategy, 10 city services and cost of living, were decided by the expert who created the model, i.e. another bias.

After applying these weights, a sum was made to obtain the ranking of attractiveness of the cities according to the opinion of the experts who participated in the survey. The results of this ranking can be visited on the UFV Observatory website (UFV, 2021) or at the Smart CityCongress. The cities in this ranking were divided into four large groups according to their position in the classification: Advanced (1-93), Challenging (94-116), Emerging (117-152) and Basic (152-175). The latter are those that do not have the basic services or characteristics of an attractive city, but nevertheless meet the minimum requirements to appear in the ranking. These groups bear some resemblance to those established in the JLL, but different criteria are applied.

As we can see, previous research conducted by the World Observatory of Attractive Cities focused on obtaining a ranking of city attractiveness based on a broad set of data. This attractiveness ranking, grouped into four clusters, was intended to help policymakers in the different cities participating in the study to know which group of cities they fall into and how they could influence city prosperity.

However, taking advantage of advances in machine learning models, we want to compare these groups with those obtained by an unsupervised model that applies its own logic, without any intervention in the model's weightings by a human. We intend to group these 175 cities by similarity of their characteristics. These characteristics will be the indicators and sub-indicators of the UFV Smart Cities Observatory project that make cities attractive to attract talent, but not through a human model, but through an unsupervised clustering model. Then, we will be able to check if there is a relationship between the clusters obtained by the unsupervised model and the human model and its divisions (advanced, challenging, emerging and basic).

This grouping of cities would be very useful for governors to know where they are in that search and competition for talent. Knowing where they are positioned and how to improve will also attract the attention of technology companies looking for that constant talent. In this way we will be covering two

of the 3 T's of the cities of the future: talent and technology. We can say that all parties would gain value, as governments get to understand how to attract talent to their cities, companies can settle in those cities where talent resides and enrich them, and citizens benefit from the opportunities this offers them. The Digital Age becomes an Urban Age in which cities are a fundamental factor. (Alcalde, 2017)

Deep Learning

Fundamental to the present project have been the enormous advances in Deep Learning in recent years. This refers to a branch of Machine Learning or automated learning based on modeling high-level abstractions of input data. This is achieved by building complex structures with multiple layers and nonlinear transformations of the data (Hao, Zhang, & Ma, 2016). The algorithm is trained and learns through data and experience to give better results. This is used for advances in facial or speech recognition, image classification, and making predictions, among many other applications.

Unsupervised Models

In our research we will use the following unsupervised models:

- To reduce the dimensionality of the data, we will compare two algorithms:
    - Principal Component Analysis (PCA).
    - Autoencoder.
- K means will be the clustering algorithm.

Like the present one, there are many researches using unsupervised models, such as the IEEE study for clustering urban areas in Seoul. This used K means for clustering and PCA for data reduction, considering that it was only useful for linear relationships (Han & Sohn, 2016).

An unsupervised model is a model in which there is a set of input data, but there is no target variable; rather, one seeks to find a pattern, a structure in the data. In unsupervised learning, we do not have a previously known label in the data, i.e., there is no known target variable to train the model. Therefore, the difficulty with unsupervised models is that there is no target variable to compare the results obtained by the model to.

There are many unsupervised models that could solve our case study. We will achieve a reliable simplification, reducing the data model from 175 cities after a previous cleaning of the almost 200 variables to 126 that we will finally reduce to a lower dimensionality to work with them. In addition, an unsupervised model will be used to segment the data into clusters.

Principles Components Analysis (PCA)

Principal component analysis, known as PCA (Amat, 2017), is a linear combination of the original variables in our data set. Thus, it returns a data set that explains approximately the same data as the initial set of values, but with a smaller dimension. It is an unsupervised model, since this model does not use any target variable, but rather seeks to extract information about the variables in question.

Each of the variables calculated by the linear combination of the initial data set has the name of principal component. The calculation of the principal components requires prior standardization of the data.

The way to calculate the optimal number of principal components is by means of the cumulative explained variance. This is the sum of the explained variance, for instance, the information that each principal component is able to capture. Normally, since we want to reduce the dimensionality of the data, we are left with the number of principal components beyond which the increase in the cumulative explained variance is no longer substantial, for example. the change is minimal and, therefore, not much additional information is collected.

Neural Nets

Neural networks are simple models inspired by the functioning of the nervous system. The nodes are called "neurons", which are the basic units of the model that are organized in layers. A neural network typically has three layers: the input layer, one or more hidden layers and the output layer. The input layer has the input data, these pass through the weights from one layer to another until they reach the output layer, where the result is obtained (IBM, 2021).

First, the process is virtually random, but the network learns from each output and compares it to the expected result. It is in the training phase where the process is repeated, gradually changing the weights until a valid error measure is obtained. At the end of the training phase, the validity of our neural network can be tested with new input data not used during the training phase to see if our network generalizes well with the new observations. For this purpose, the data are divided into a training set and a validation set.

Autoencoder

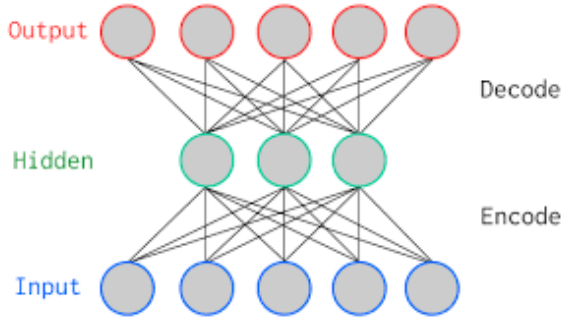


Figure 2: Autoencoder (Wkipedia Commons).

An autoencoder is an algorithm that we will use for feature extraction, reducing the dimensionality of the input data. It consists of three elements: encoder, bottleneck and decoder. With this algorithm, it will try to reconstruct the input from the bottleneck where the most important data are located. It wants to generate the same input data repeatedly to arrive at a lower dimension of that data that collects all the information in the dataset (Roy, 2020).

The encoder tries to compress the input data through the weights and calculations performed, reducing the dimensionality, while the decoder tries to recreate that more compressed version of the data in the initial dataset. After training the data, the decoder disappears, and the encoder remains. With the encoder, we get a reduced version of the dimensionality of the data (latent space) and this data is what we use for clustering using a new clustering algorithm. (Hubes, 2018). Then, we will detect the cities into clusters according to their similarity.

The autoencoder works as an MLP. They are connected layers where each neuron is connected to neurons in the next layer and has an associated weight (Charte, 2021). An activation function is applied to each neuron and the parameters are reset by backpropagation. We use the multilayer neural network model, which should be symmetric, using an intermediate layer to reproduce the initial data (Muaz, 2019).

The model learns by trying to achieve the minimum error. The error function tries to see how the input data can be reconstructed. In this case, the MSE (Mean Squared Error) between the input and output variables will be calculated to determine the efficiency of the network in reproducing the initial variables. The weights store the data information (Charte, 2021). The error function, MSE, is the one that corrects the adjustment of the weights during the backpropagation process, comparing the original value of the obtained result. The objective is to minimize the error.

There are several optimizers for parameter tuning, so we must find out which one is the best for our research. We will compare the results between Adam or SGD, which will be the optimizers under study.

As with PCA, the data from an Autoencoder must be standardized. It is important to standardize the training dataset data and then the validation data so that the validation data does not intervene in the training, as it can lead to problems of low generalization of the model.

As for the activation function, the sigmoid and hyperbolic do not give good results when applied to multilayer models. The SGD optimizer is the most used when more than two layers are used for the backpropagation process (Brownlee, 2020).

Regarding the layers and their dimensionality, we can say that the higher the number of layers, the more careful we must be to not overfit the model, although usually the algorithm learns better. In addition, since we will not only have linear relationships, but we will also use an activation function that learns from the nonlinear ones.

The activation function, which applies a nonlinear transformation to the neurons, RLU is the most commonly used to replace the sigmoid with linear neural networks and improves its performance in the field of Autoencoders. It has two linear parts instead of one (Goodfellow, Bengio, & Courville, 2016). The computation is faster and less expensive as it does not need to perform exponential functions (Glorot, Bordes & Bengio, 2011). Moreover, it overcomes the problem of gradient fading, since it returns the positive value instead of a 1 as in the sigmoid, and a 0 when it is a negative value. However, the problem of gradient blowup can occur (Charte, 2021).

The lines of work for which the autoencoder is used are several:

- Reduce the dimensionality of the data
- Identify anomalies
- Noise removal
- Imputation of missing data
- Data anonymization
- Semantic clustering
- Generation of new data

PCA vs Autoencoder Comparison

Principal components, or PCA, are used especially when linear relationships can be extracted from the data. The risk of overfitting is reduced by reducing the number of input variables. However, the Autoencoder can learn linear and nonlinear relationships from the data. The Autoencoder works much like principal components, if a single layer and the sigmoid is used as the activation function (Balodi, 2021). The sigmoid is also widely used with probability-related models, since the range of its values is between 0 and 1.

In addition, the autoencoder is more likely to overfit the model, as it learns from the model itself and may not generalize well to new data. In our case, we do not want it to generate new data, since we will not be using the model with a test set. Our goal is that the autoencoder learns to reduce the dimensionality of the data by collecting the most important variables from the data (Fawi, 2021).

On the other hand, we employ an unsupervised clustering model, K means. We could have explored other algorithms for data clustering, e.g., DBSCAN. With DBSCAN, no predetermined number of clusters is necessary, but it does not work very well when the clusters that are created have different densities (Cioffi, 2021). In addition, DBSCAN may serve better to avoid the problem of high dimensionality of the data. However, according to previous studies, K means is better in terms of time and speed (Charkraborty, Nagwani & Dey, 2011).

K means

It is a clustering algorithm that divides the dataset into similarity groups to find patterns in the data. Observations within the same cluster resemble each other and differ from observations in other clusters (Garbade, 2018). For this reason, the algorithm looks for a specific number of groups called clusters. A cluster would then be a smaller set of the initial data set aggregated by similarity.

For clustering, the following steps are needed: establishing a proximity measure that indicates similarity or not, an error function to evaluate the groups, and an algorithm for clustering (Amat, 2020). Some methods used to calculate the distance are the Euclidean distance and the Manhattan distance.

First, a number "k" of centroids is defined. The centroids are the centers of the clusters, which are first defined randomly. Next, the observations in the data set are assigned to the nearest cluster. Next, the sum of squares is calculated to find the dispersion within the cluster, a new nearest centroid is calculated, and the observations are re-clustered.

To calculate the validity of the centroids, the cohesion of the clusters is measured. It is usual to use the SSE and look for the minimum dispersion within the cluster, while the separation between clusters should be the maximum (Leon, 2021).

Why K-means? It is the most popular because it is very simple to use and time efficient. It does not reach the global optimum of the clusters, but the local minimum. A negative point is that the user must indicate the number of clusters he wants to create, and that the algorithm is very sensitive to outliers, so they will have to be prepared and the data will have to be normalized (Ullman, Poggio, Harari, Zysman & Seibert, 2014).

2. OBJECTIVES

General Objective

The general objective is to perform an unsupervised model contrast to reduce the dimensionality of the data and bring the complexity of the reality of the cities to a simpler model. We are looking for a model that does not have biases applied by humans. Starting from 175 smart cities and analyzing up to 200 indicators that influence the attractiveness of the city, it is intended to group the 175 cities under study into clusters based on the similarity of their characteristics that define an attractive city for talent. For these groups of cities, common characteristics will be indicated.

These groups will be compared with the four groups defined by the attractiveness to attract talent of the cities: advanced, challenging, emerging and basic. In this way, we will be able to know if there is a possible relationship between the four groups of similar cities in our model and the attractiveness groups of the UFV World Observatory of Attractive Cities.

This information will be very useful for decision-makers to understand where they stand when it comes to attracting talent. On the other hand, the positioning of the city clusters will attract the attention of those technology companies looking for talented people in the cities.

Specific Objectives

In addition, as we have already mentioned, we will contrast different unsupervised models. We will deepen our knowledge of unsupervised algorithms and the behavior of neural networks. Neural networks are increasingly talked about, but do you know exactly what they consist of? When is it more interesting to use any other model? In this case, we will explain the use of PCA, a neural network called autoencoder and the k-means algorithm. Finally, the model used will be justified.

We summarize the objectives in the following points:

1. to deepen the dimensionality reduction of data through unsupervised algorithms such as PCA and Autoencoder.
2. Grouping of cities by similarity using an unsupervised clustering algorithm such as K means. We will obtain a grouping of cities without setting the weights ourselves, i.e. without applying our opinion.
3. Interpretation of the clusters obtained by our model and extract a valid knowledge for the rulers of the characteristics of the group in which they are positioned. We will be able to understand which groups are more attractive when it comes to attracting talent based on the characteristics they share.
4. Comparison of the clusters obtained by our model with the attraction groups according to the ranking of experts of the study of the Observatory of Attractive Cities of the Francisco de Vitoria University (advanced, challenging, emerging and basic).
5. Extract value for technology companies to find which are the leading cities today in the extraction of talent and establish their delegations in them.
them.

3. METHODOLOGY

Methodology

For the clustering of cities, we have a very rich set of data that we can reduce thanks to algorithms that aim to reduce the dimensionality of the data. We will compare two algorithms that can be applied for this purpose. On the one hand, Principal Components can collect in a smaller number of variables the linear combination of others from a larger set. On the other hand, it will be possible to train a neural network that also allows to reduce the dimensionality of the data. Finally, an unsupervised clustering model is applied that allows us to group the 175 cities of the study into 5 clusters.

To do this, we wanted to test the potential of machine learning tools and algorithms for data analysis and clustering. In this way we obtained an algorithm capable of grouping the cities according to their similarity. Having the model, we can include more and more cities and the algorithm will keep learning from them. Therefore, instead of using the weights from the previous UFV research mentioned above, we will only use the indicators and sub-indicators that influence the attractiveness of a city, and we will try to reduce the dimensionality of the data using an algorithm to then cluster the cities. We will compare PCA or Autoencoder for this dimensionality reduction. These algorithms will be able to perform their own weighting and calculations to reduce the more than 200 indicators to a smaller number, which will allow us to represent a city more easily. We will choose the algorithm whose result is better than the

rest. We will be able to find that pattern that will make some cities look like others. In addition, we will be able to check if the clusters correspond to those mentioned in the previous study: advanced, challenging, emerging or basic.

Several previous studies have been done on the performance and efficiency between PCA or autoencoder, and then applying a clustering algorithm such as K-means, so we want to test and compare them on our dataset.

As for performance, we will add the time component to our code to monitor this indicator and check which algorithm spends more time on training, PCA or Autoencoder. To know the training time between one model and another, we add the following lines to the code:

```
import time
start_time = time.time()
....
print("--- %s seconds ---" % (time.time() - start_time))
```

4. SOLUTION DEVELOPMENT

Tools used

Python is a programming language used primarily for automated learning (Keepcoding, 2022). It is one of the most widely used programming languages by the data analyst and data scientist community, it is open source and very useful for data analysis and exploration, as well as for building automated learning models.

The environment in which I am going to work is Google Colab. It is an environment in which to write and run our Python code. It is a Google Research product available in the browser and is widely used in data analytics and machine learning. It is a cloud platform, like Jupiter Notebook and allows execution on GPU. Another advantage is that it has most of the libraries I will be using. It allows me to have the security that it is stored in the cloud and an easy upload of the data tables.
As for the libraries available in Python for machine learning, Tensorflow is mainly used by academic communities for deep learning among others such as matplotlib (matplotlib, 2022) for visualization, numpy (Numpy, 2022) and many more that facilitate programming through libraries already created by

the community that forms Python. As Python is an open source language, it is continuously nourished by improvements and packages.

Between Pytorch and Tensorflow we will finally use Tensorflow for its robustness, for being widely used and a library especially fed by the community and specialized in the creation of deep learning models (Dubovikov, 2017).

In the execution environment we will make use of the GPU. The CPU is the central processing unit centered on the Von Neumann architecture. The CPU stores programs and data in memory and has optimization cores for sequential processing. The ALUs of the CPU, are the arithmetic logic units that perform the multiplication and addition calculations, working sequentially and stored in memory, they have to access the memory every time there is a new operation, which slows down the processing time. However, the GPU is the graphics processing unit, it uses thousands of ALUs in a single processor, allowing multiple operations to be performed at the same time in parallel. The GPU accesses registers or shared memory to read and store the results of intermediate calculations, which is faster than sequential processing with numerous CPU memory accesses. For this reason, we will use the parallel computing platform and the CUDA programming application interface, which allows us to use the GPU, which will be faster in computations (NVIDIA, 2020).

```
device = tf.config.list_physical_devices('GPU')
```

Finally, both Python and Tableau will be used for data visualization. Regarding the exploration of data and analysis, Python will be the main tool and Tableau for the interpretation of results and conclusions.

Data

The data come from a study conducted by the Observatory of Attractive Cities of the Francisco de Vitoria University.

There is a wide variety of variables to be studied. The number of variables is finally 126 after a previous filtering of the almost 200 collected. However, the available observations are 175 since we are going to study the comparison only among 175 cities. There are many variables, most of them possibly correlated, so we will first have to do some data cleaning. This is the most tedious and time-consuming part, but it is very relevant. If the set of variables increases a lot, the observation in the data set should also increase. Therefore, we will try to eliminate those variables that are not relevant, since we cannot increase the number of observations.

The 175 observations are cities that were chosen based on two of the best city rankings based on facts rather than surveys. These leading cities in terms of quality of life were chosen based on the IESE Cities in Motion research (IESE, 2020). In addition, all of these cities scored above 50 on the liveability index calculated by The Economist (Economist Intelligence, 2021). This last indicator is speaciallu relevant, as no one would live in a dangerous place (Ondiviela, 2021).

In addition to the indicators of magnetism and profitability, there are others that are included in a separate sheet called "attractiveness" with general data that can define the characteristics of attractive cities, such as population.

Transformation

The preparation step is key, since having good results or not depends not only on the trained algorithm, but also on the purity of the data. Therefore, it is relevant to analyze if there are outliers, unique or non-representative data and null values. It would also be interesting to explore the type of data and their distribution.

Variables that do not provide relevant information or that negatively affect the reliability of the study will be eliminated. We start from the assumption that it will not be necessary to eliminate observations, since complete data are obtained for a total of 175 cities, so no missing values will be found. There are columns calculated by means of others. Therefore, there will be redundant information and will give a high correlation coefficient. These, not adding relevant information, will be removed from the data set. Finally, there are spreadsheets within the three Excel files in question with summaries and graphs that should be eliminated as they do not contribute to the objective pursued in this study.

It is important to standardize the variables to avoid that those of higher rank are considered of greater importance than those of lower rank. To achieve this purpose, we will use the Python function StandardScaler() (scikit-learn.org, 2022).

We can check for correlations between the variables in the three initial tables using the function: corr(). Darker colors indicate strong positive correlations and lighter colors strong negative correlations.

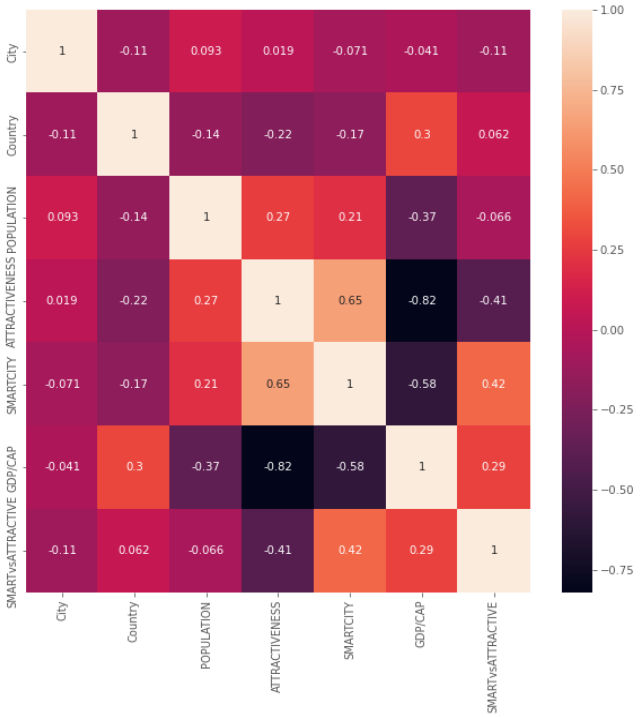


Figure 1: *Correlation Magnetism Matrix (Own work)*

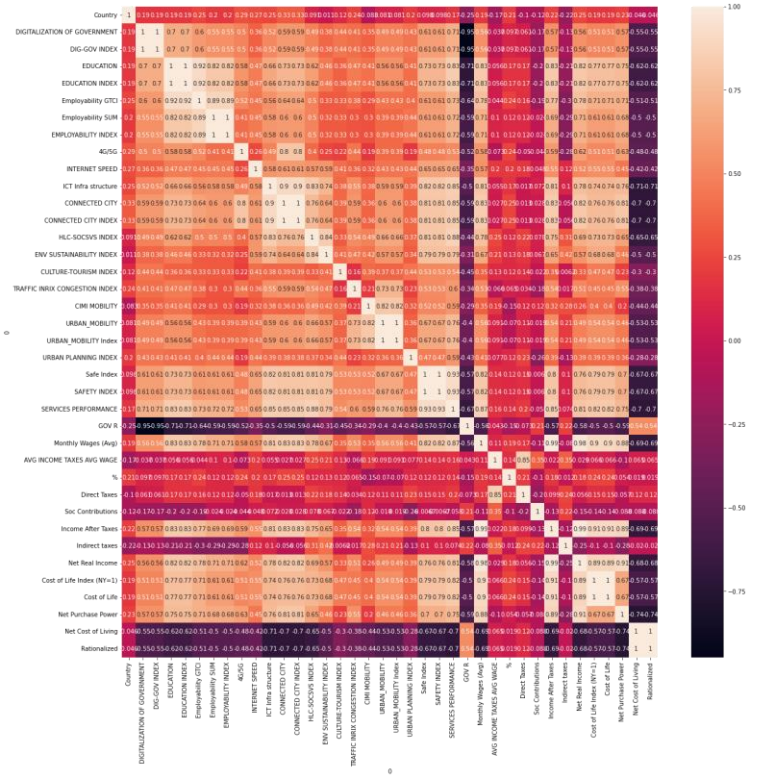


Figure 4 Correlation Profitability Matrix (Own work)

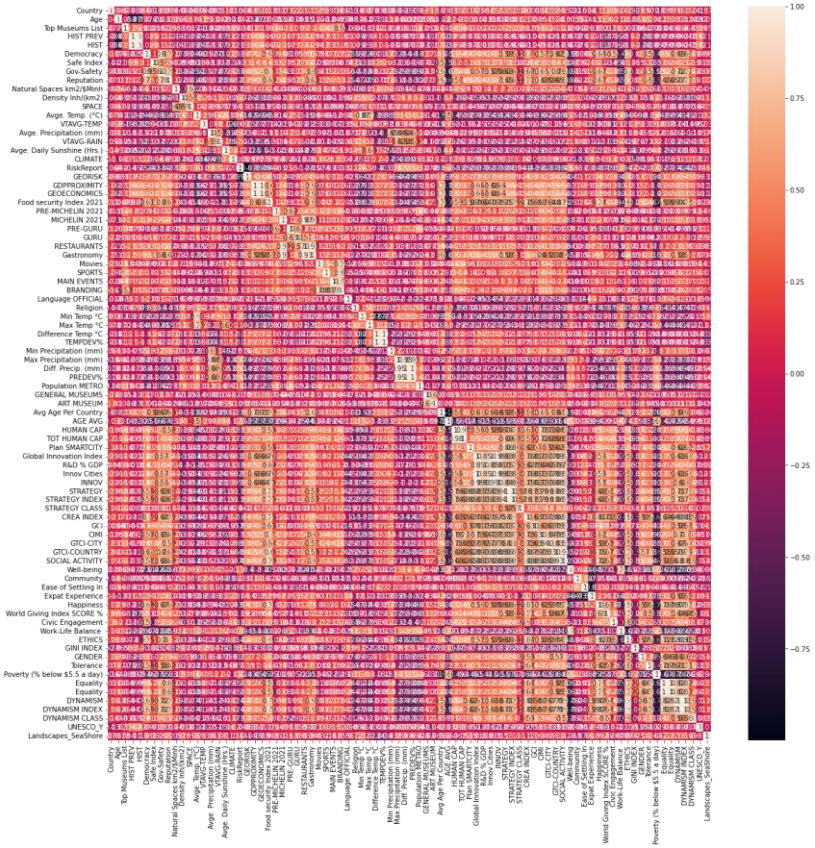


Figure 5 · Correlation Attractiveness Matrix (Own work)

Analysis

Our objective is to group the cities according to their similarity. We will compare the results by applying K means in three different scenarios: on standardized data without reducing dimensionality, reducing dimensionality with PCA and finally, reducing dimensionality with an Autoencoder.

To compare each of the three models created, we will use the K means inertia measure. This allows us to identify how well our algorithm has clustered the data. It measures the distance between each observation and its centroid, squares that distance, and sums these values over its entire cluster. A good model is one that has a low inertia value, indicating greater cohesion in the clusters.

1. We normalize the data and cluster them without reducing the dimensionality of the data.
2. We use K means

We apply K means with a previous standardization of the initial data with StandardScaler. Then, we explore according to the Elbow Rule the optimal number of clusters. The optimal number is found at the point on the line where the curve is made. We will say that the optimal number we will use is 5, since it is a point that is on the curve and that we can easily compare between them.

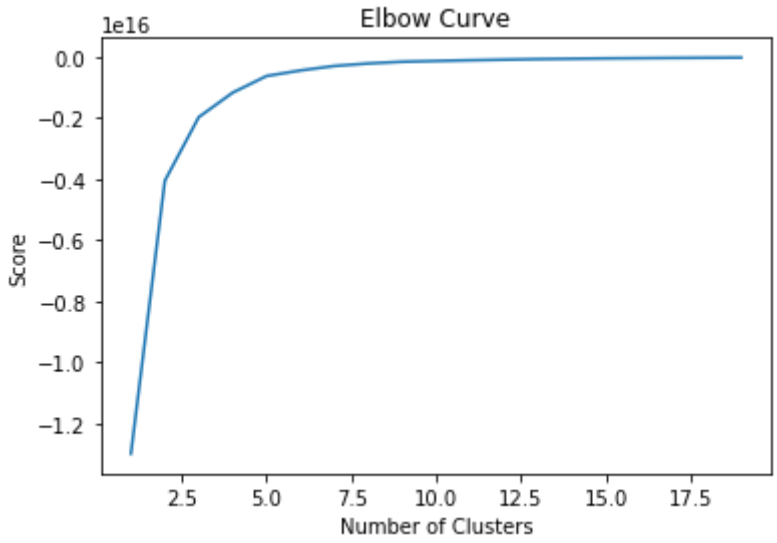


Figure 6 : Elbow Rule for first model (Own work)

After deciding the optimal number of clusters, we apply the K means algorithm to the standardized data set.

For this first model the measure for *kmeans* inertia_ = 12367.903610287718

Dimensionality reduction algorithms comparison

1. Reduction of the dimensionality with PCA and apply *K means*.

We have previously seen that there are strong linear correlations between the variables in our data set, so we will only keep those variables that are not correlated, nor computed. Each principal component will collect independent information from the data (see Annex 6). First, we will leave this work to the PCAs.

Having too few observations and too many variables can lead to over-fitting of the model or to very poor generalization. Therefore, we see it relevant to reduce such dimensionality as it will also improve the training time and the productivity of the algorithm.

Next, we calculate the percentage of explained variance accumulated, which will help us to decide the optimal number of Principal Components to use:

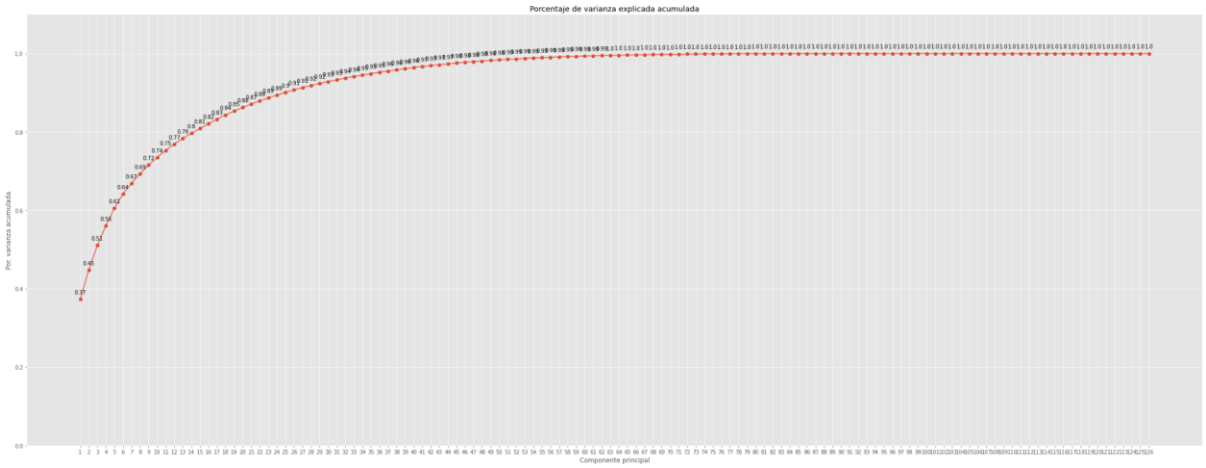


Figure 7. Accumulated Explicative Variance (Own work)

We see that it is with 63 principal components when the model does not provide additional information, so we will only work with 63 principal components. This is where the increase in cumulative explained variance starts to become insignificant, so we consider only the smallest number of components that explain the most information in the data set.

In addition, we will apply the K-means model, but to the data reduced by the PCA algorithm. We also start in the same way as before, looking for the optimal number of clusters to create according to the data set.

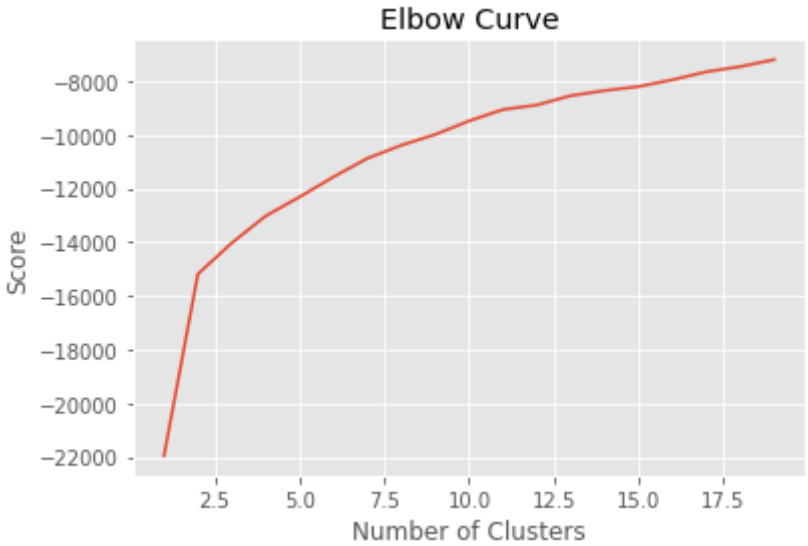


Figure 8 Elbow Curve for the second model (Own work)

The inertia measure for the cluster using the data reduced by the PCA is:

kmeans.inertia_ = 12268.082891594982

2. Finally, we apply the K means model to the data set reduced by the neural network: Autoencoder.

We have an encoder, the intermediate layer or bottleneck and the decoder that tries to represent the initial data.

There are people who use as activation function in Autoencoders: Selu. However, it generally does not give better results than RELU and the training time is usually longer. (Patel, 2017). Also, for the optimization process, we tested both Adam and SGD as optimizers.

- Autoencoder: we start with a learning rate of 0.1. (Can be decreased)
- Block size (batch): 32
- Number of iterations: 25
- Activation function: Relu
- Optimizer: Adam or SGD. We verified that SGD has less training time and better outcomes.

The training results for the model using SGD as optimizer are:

SGD: Epoch 25/25
0s 28ms/step - loss: 0.9964 - val_loss: 0.9971

Training time: --- 3.664219856262207 seconds ---

Next, we see that there is no evidence of overfitting, since the error decreases as the iterations progress, both with the training and the validation data.

9

Figure 6: Training and validation error SGD (Own work)

The training results for the model using Adam as optimizer are:

Adam: Epoch 25/25
- 0s 33ms/step - loss: 0.0585 - val_loss: 0.2785
Trainig time: --- 5.7519371509552 seconds ---

In the graph below, we see that the training error decreases with each iteration. However, with the validation dataset there is a point where the error remains constant. This means that the model does not improve with each iteration when applying new data because it fits the training data, so it does not generalize correctly.

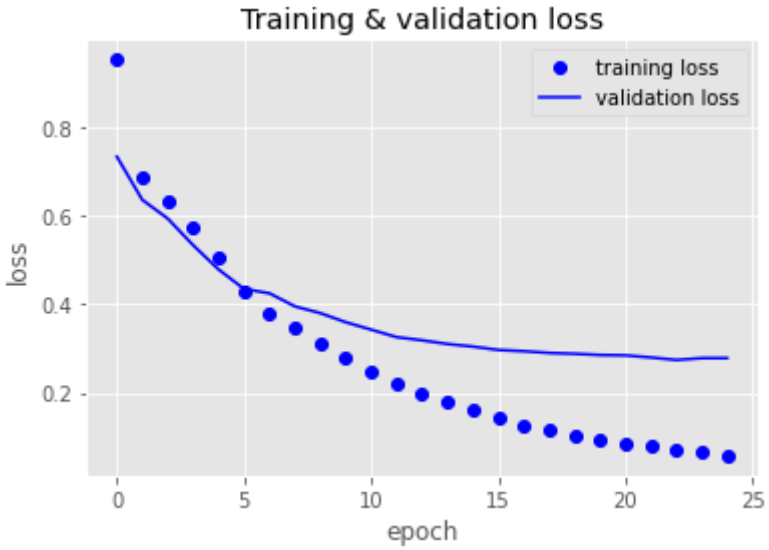


Figure 10: Training and validation eror Adam (Own work)

If we compare the results after applying each of the optimizers, we see that the training time is shorter with SGD. In addition, although the validation loss is lower for Adam, we can see in the graph of its error function that the model has a higher risk of overfitting. This is because there comes a time when the model remains static and does not improve. However, in the case of SGD, the validation error decreases proportionally with the training error. For this reason, we select SGD as the optimizer. Moreover, if we apply K averages to the data obtained by each of them, we find that we obtain better results with the data reduced by the Autoencoder when SGD is applied as the optimizer.

kmeans.inertia_
14173.1455078125 (ADAM)
89.43651580810547 (SGD)

As a last point of analysis, we explore the possibility of using a regularizer to avoid the risk of overfitting the model, through regularization l1 and l2. (Brownlee, 2020), but the results did not improve, so the proposal was rejected.

Based on the data returned, we select the model with the least error. Next, the errors of each of the different k means trained are represented. The first bar corresponds to the original standardized data, the second to the data reduced by PCA and, finally, the data reduced by the Autoencoder in a bar chart.

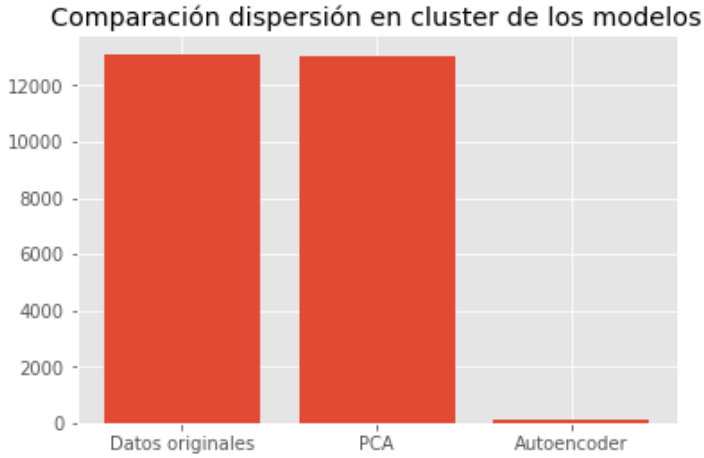


Figure 11 Error measure comparative (Own work)

All in all, we can observe KMEANS with the model that has the least error, that is, the Autoencoder. 5 clusters are shown below in the graph, each of them represented in a different color.

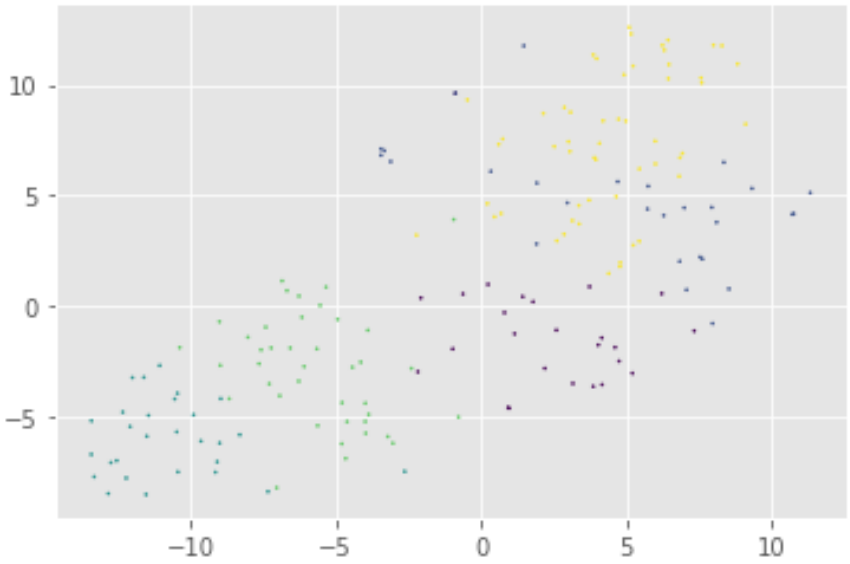


Figure 12 Clusters diagram (Own work)

5. RESULTS

First, as mentioned above, the neural network, specifically the Autoencoder, can reduce the dimensionality of the data in such a way that by means of linear and nonlinear combinations we can avoid the problem derived from the high dimensionality of the data and allow a more efficient and accurate training than the PCA. It has been possible to reduce the 126 of the initial cleaning to 20 variables (in the latent space or bottleneck) which facilitates the grouping and representation of the cities. This measure of inertia error does not correspond to the error with respect to a target variable, but to the dispersion of the data, since the objective is to have groups as far apart as possible, but as compact as possible. For this reason, it was decided to use the autoencoder for the application of k means.

Next, we can interpret the results obtained from the grouping and check what these clusters are due to.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Toronto	Vienna	La Paz	Buenos Aires	Sydney
Ottawa	Antwerp	Medellín	Córdoba	Melbourne
Montreal	Vancouver	Bogotá	Manama	Adelaide
Tel Aviv	Paris	San José	Minsk	Canberra
Yokohama	Lyon	Santo Domingo	Brasilia	Linz
Osaka	Berlin	Quito	São Paulo	Brussels
Nagoya	Hong Kong	Cairo	Rio de Janeiro	Prague
Santander	Milan	Accra	Sofia	Copenhagen
Dubai	Rome	Mumbai	Santiago	Aarhus
Abu Dhabi	Florence	Bangalore	Shanghai	Tallinn
Boston	Torino	New Delhi	Beijing	Helsinki

Washington, D.C.	Tokyo	Hyderabad	Guangzhou	Tampere
Chicago	Luxembourg	Jakarta	Shenzhen	Espoo
Seattle	Singapore	Kuwait City	Chengdu	Oulu
Los Angeles	Seoul	Kuala Lumpur	Chongqing	Marseille
Baltimore	Barcelona	Guadalajara	Shenyang	Nice
Philadelphia	Madrid	Casablanca	Wuhan	Bordeaux
Dallas	Bilbao	Rabat	Suzhou	Lille
Phoenix	Zaragoza	Panama City	Tianjin	Munich
Houston	Zurich	Asunción	Harbin	Dusseldorf
Atlanta	Geneva	Lima	Zagreb	Frankfurt
Miami	Bern	Manila	Tbilisi	Hamburg
Las Vegas	Basel	Cape Town	Athens	Stuttgart
Kansas City	Taipei	Durban	Budapest	Cologne
Montevideo	London	Johannesburg	Jerusalem	Dublín
	San Francisco	Bangkok	Riga	Amsterdam
	New York City	Ho Chi Minh City	Vilnius	Eindhoven
	Honolulu	Hanoi	Mexico City	Rotterdam
			Monterrey	Den Haag
			Doha	Auckland
			Bucharest	Wellington

Moscow	Oslo
St Petersburg	Bergen
Riyadh	Stavanger
Belgrade	Warsaw
Bratislava	Wroclaw
Tunis	Lisbon
Istanbul	Porto
Ankara	Ljubljana
Kiev	Málaga
	Valencia
	Seville
	Stockholm
	Gothenburg
	Malmo
	Edinburgh
	Birmingham
	Liverpool
	Manchester
	Belfast
	Bristol

Nottingham
Glasgow
Denver

Table 1: *Clusters (Own work)*

To facilitate the understanding of the results, we visualize the location of the cities of the different groups on a map to find out if there is any relationship with the country to which the city belongs to.

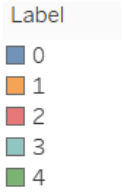


Figure 13: *Color leyend (Own work)*

As can be seen below, most cities in the same country belong to the same group, except for cities in countries such as France, Germany, Spain and Japan. However, all countries tend to cluster or fall almost equally between 1 and 4, which represents the difference between the first and second cities in each country.

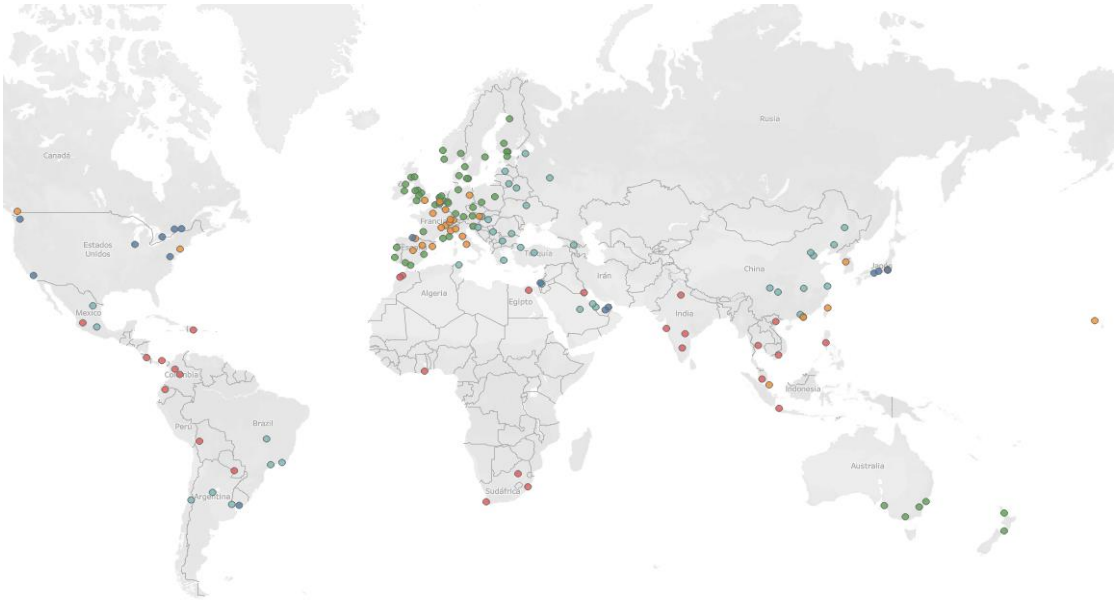


Figure 14: *Clusters map (Own work)*

Cluster 0 includes countries such as the United States, the United Arab Emirates, Japan, Canada, the capital of Israel and Santander in Spain. Cluster 1 includes the capitals of European countries such as

Vienna, Berlin, Paris and London, as well as Taiwan, part of the United States, Tokyo, Japan, Italy, Switzerland, Singapore, Seoul and the leading Spanish cities such as Madrid and Barcelona. In the second cluster, we find Latin America, India and Africa, while in the third, we find Brazil, Russia, China and Turkey. Finally, in the fourth cluster we find the Nordic countries, Sweden, Finland, Denmark, Norway, part of France and Belgium, Germany, the Netherlands, Poland and part of the United Kingdom and Spain.

This classification places the cities in groups 0 and 4 as those cities that we could consider advanced in attracting more talent and whose characteristics make them attractive to talent. Why does our model divide the cities in groups 0 and 4 into two different groups? The cities in group 0, as we have already mentioned, correspond to the cities of the United States, Japan, Canada, the United Arab Emirates and Israel. All of them major world leaders that stand out mainly in economic matters and, therefore, those that stand out in city profitability components such as city purchasing power. However, the cities in group 4 are those large cities that combine more emotional and rational factors, such as those in the Nordic countries, the Netherlands, Germany, Australia, part of Spain and part of the UK. All of them coincide in being leading and highly competitive cities with great potential, but they stand out to a great extent in the profitability of the services they offer, and in the magnetism in their dynamism, due to their high values in the social, environmental and innovation components. Those cities in which the person is put at the center is what is sought after today (Gehl, 2010).

In cluster 1 are the large capitals of the countries whose cities mostly fall into groups 0 and 4. This could be because the capitals usually offer a poorer quality of life and have the economic and environmental problems that usually occur in the big city, so they are positioned in a separate group.

All this we can see in an example, as would be the case of the United States, a country that does not stand out for the services it offers to citizens or for the care of the environment. Especially not when Trump was in power in 2021, the year in which the data was collected. However, it is one of the best in the economic aspect. We can see this in the data on which our study is based. If you want to check the above in the data, for example, in terms of healthcare as a service, the United States does not appear until the 16th position. (See Annex 3) while in the economic issue all American cities appear in the top positions (See Annex 8).

On the other hand, cluster 3 clearly represents those emerging cities that try to compete for talent, the so-called BRICs, but without India. They are those trying to outperform and improve, such as China, Russia, Brazil and Argentina. China would soon position itself in the advanced ones, due to its strong investments in innovation (See Annex 5), but despite being very powerful economically, it is not in terms of quality of life since its government is a dictatorship. It is worth mentioning the fact that this

cluster does not include India. India is not good in terms of social structure, services or emotional aspects of the city, which makes it fall in this last group.

Finally, we could place the basic ones within cluster 2. Cities in Africa, India, Vietnam and Latin America that need to improve in many aspects to position themselves and compete for talent.

We could then say that it is in the cities of groups 0 and 4 where there is currently more talent and greater prosperity to a large extent for technology companies. A clear example is Dublin, which is in group 4, and is the cradle of technology today.

Previous Study Comparison

According to our results, we can see that there is a relationship with the clusters created in the previous research on which our project is based. Recall that these clusters were: advanced, challenging, emerging and basic cities, and that the cities that fall into these clusters come from a ranking calculated according to the results of the surveys of smart city experts who attended the Smart Expo 2018 in Barcelona. (See Annex 9) Advanced cities are those that stand out above the rest and, therefore, those that attract talent at present. Challenging ones are those that pose a challenge or threat to the advanced ones, as they grow with agility and compete to attract talent. On the other hand, emerging companies are those that are beginning to stand out and position themselves, while basic companies are those that need to fix and resolve many basic issues in order to compete for talent. According to this definition, we will understand if there is a relationship with the groups created by our model.

We can say that our interpretation can be correct and consistent with the results of this study. According to the division of the ranking positions and the groups created by our model, we could say that the cities of groups 0 and 4 are positioned as the advanced ones, since 77% of the cities that fall in the first 93 positions of the ranking created by the UFV Observatory of Smart Cities, belong to group 0 and 4. From position 94 to 116 there are cities of group 1, and some of the groups 0 and 4. This leads us to think that there is no longer much differentiation between cities and that this "gap" is becoming smaller and smaller, reaching a more homogeneous world, to some extent, thanks to technological advances, investment in innovation and globalization. On the other hand, the emerging (117-151) are very clear, where 83% of those occupying these positions belong to group 3. Finally, those in group 2 would be those who must take action to avoid being left behind. There is a clearly defined group, since 92% of the cities that are positioned in the ranking in the place of the basic ones (152-175), are within group 2 of our study.

1-93	Advanced	Cluster 0	21	93	77%
		Cluster 4	51		
94-116	Challengers	Cluster 1,0,4	15	23	65%
117-151	Emerging	Cluster 3	29	35	83%
152-175	Starters	Cluster 2	22	24	92%

Table 2: *UFV results vs K means (Own work)*

6. CONCLUSIONS

These results are consistent and closely related to those obtained in the previous study conducted at the Smart Cities Observatory of the UFV on which this model was inspired.

We have managed to group the cities by similarity of their characteristics, in addition to placing these groups in advanced, challenging, emerging and basic cities, so that young people know in which group of cities they can develop their talent, and the governors know how to compete for talent. In addition, we have been able to verify that these groups correspond to those created by the model of the previous study at the UFV's Observatory of Attractive Cities.

Likewise, we have found certain differences and we have understood the bias of the model created by the subjectivity of human beings. This was one of the objectives pursued, to create a model that classifies by similarity, avoiding the influence of human decision in the weightings given to the indicators. In this case, the model chosen has turned out to be a neural network that learns from data and experience, managing through linear and non-linear combinations to find the way to combine these characteristics of attractive cities to then apply a model of grouping cities by similarity. In these models there is no emotional component, but a rationalization component. A clear conclusion is that the expert-based weights model used for the Attractive Cities Observatory was very correct, as the trained model has generated very similar clusters without bias or subjective weights. Our biggest bias could be said to be subjectivity when interpreting the clusters.

According to the clusters generated, they are not very separate groups, although the groups are clearly differentiated by color, and do not mix with each other. A solution to this could have been to create more clusters, as 175 cities in 5 clusters create large groups. However, this number was decided by the elbow rule. It should be noted that we are working with a model with very few observations with respect to the large number of variables used, which limits our scope for action and the ability to obtain reliable results. It is to be expected that with a larger number of observations on which to train the model, better results could have been obtained.

On the other hand, it is worth mentioning that neural networks are called black boxes because it is difficult to interpret the results and to know the influence of the variables on the output of the results. With SAS we can know some graphs for predictive models and know the influence of the variable on the results, but we do not contemplate it in unsupervised models in Python.

In the long term, one could continue to feed the model with more cities and indicators, so that it learns from the changes produced year after year. In addition, it would be very interesting to compare clusters from year to year and see how cities move from one cluster to another, depending on events that occur or improvements in terms of what their cities offer.

The first conclusion we can draw is that those talented young people may decide to move to those cities positioned in clusters 0 or 4, depending on whether they are looking more for the economy or social services, whether they focus on the person or the environment. These are the three main drivers of cities: economy, social services and sustainability. Therefore, cities located in cluster 1 excel in many aspects, but they must continue to innovate to catch up with the most advanced ones. This is done by challenging with strong innovation, better services, or reform of the tax system, like Belgium or through gastronomy and traditions like the Spanish ones. Finally, as for cluster 2 cities, their governors and legislators should take action and focus on challenging cities to be able to change clusters in the near future and move forward.

To conclude, companies are similarly advised to relocate to those cities that attract talent, i.e. the advanced ones. Depending on the objectives and purpose of the company, it is possible to opt for group 0, more focused on economic aspects, or group 4, more focused on people and the services offered. However, the challengers are those cities that need these technological companies to advance quickly, so they will also be appreciated.

7. BIBLIOGRAPHY

Alcalde, I. (2017) Ciudades con Alma en la Era Digital. CITIZEN http://thecitizen.es/cultura/ciudades-con-alma-en-la-era-digital

Amat, J. (2017, junio). *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*. cienciadedatos. https://www.cienciadedatos.net/documentos/35_principal_component_analysis#Proporci%C3%B3n_de_varianza_explicada

Amat, J. (2020, diciembre). *Clustering con Python*. cienciadedatos. https://www.cienciadedatos.net/documentos/py20-clustering-con-python.html

Balodi, T. (2021, 2 agosto). *3 Difference Between PCA and Autoencoder With Python Code | Analytics Steps*. Analyticssteps. https://www.analyticssteps.com/blogs/3-difference-between-pca-and-autoencoder-python-code

Bankinter (2022) *Precio de la vivienda en 2022: ¿En qué ciudades españolas es más caro comprar una vivienda'?* Bankinter. https://www.bankinter.com/blog/finanzas-personales/precio-vivienda-ciudades

Brownlee, J. (2020, 20 agosto). *A Gentle Introduction to the Rectified Linear Unit (ReLU)*. Machine Learning Mastery. https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/

Charkraborty, S, Nagwani, N.K & Dey, L (2011). *Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms.* International Journal of Computer Applications (0975-8887) Volumen 27. https://arxiv.org/ftp/arxiv/papers/1406/1406.4751.pdf

Charte, F. (2021, 27 enero). *Autoencoders ¿Qué son, para qué sirven y cómo funcionan?* Universidad de Jaén (UJA). http://cemixugrmadoc.ugr.es/pages/10-banners/siadef/sesion14_transparencias/!

Cioffi, N. (2020, 29 diciembre). *AI & ML: How combine KMeans and DBSCAN. A Methodological Approach for Segmentation*. Medium. https://medium.com/analytics-vidhya/ai-ml-how-combine-kmeans-and-dbscan-a-methodological-approach-for-segmentation-ad2735dbf42d

Dubovikov, K. (2017, 20 junio). *PyTorch vs TensorFlow — spotting the difference - Towards Data Science*. Towards Data Science. https://towardsdatascience.com/pytorch-vs-tensorflow-spotting-the-difference-25c75777377b

Economist Intelligence. (2021). *The Global Liveability Index 2021*. Economist Intelligence. https://www.eiu.com/n/campaigns/global-liveability-index-2021/

ESMARTCITY. (2022, 20 abril). *Ciudades Inteligentes*. ESMARTCITY https://www.esmartcity.es/ciudades-inteligentes

Fawi, M. (2021, 9 febrero). *Autoencoder with Manifold Learning for Clustering in Python*. Minimatech. https://minimatech.org/autoencoder-with-manifold-learning-for-clustering-in-python/

Garbade, M. J. (2018, 13 septiembre). *Understanding K-means Clustering in Machine Learning*. Towards Data Science. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

GEHL, J. (2010) *Cities for People*. Island Press, Washington DC

Glorot, X., Bordes, A., & Bengio, Y. (2011, 14 junio). *Deep Sparse Rectifier Neural Networks*. PMLR. http://proceedings.mlr.press/v15/glorot11a

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press. Page 175, 226

Han, G., & Sohn, K. (2016, March). Clustering the Seoul metropolitan area by travel patterns based on a deep belief network. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1-6). IEEE.

Hao, X., Zhang, G., & Ma, S. (2016). Deep learning. *International Journal of Semantic Computing*, *10*(03), 417-439.

Hubens, N. (2018, 8 mayo). *Introducción al autoencoder*. DeepLearningItalia. https://www.deeplearningitalia.com/introduzione-agli-autoencoder-2/

Hong, G. (2019) *Urban Innovation with Dan Doctoroff (CEO of Sidewalk Labs): Catalyzing Growth and Securing Data in Modern Cities.* businesstoday. https://journal.businesstoday.org/bt-online/2019/urban-innovation-with-dan-doctoroff-ceo-of-sidewalk-labs-catalyzing-growth-and-securing-data-in-modern-cities

IBM. (2021). *El modelo de redes neuronales*.IBM. https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model

IESE (2020). *IESE Cities in Motion Index 2020*. https://media.iese.edu/research/pdfs/ST-0542-E.pdf

IESE Cities in Motion (2020). *IESE Cities in Motion Index 2020.* IESE Business School. https://citiesinmotion.iese.edu/indicecim/index.eng.html?lang=en

JLL & The Business of Cities (2018). *World Cities: Mapping the Pathways to Success.* JLL and The Business of Cities. https://www.us.jll.com/en/trends-and-insights/research/world-cities-mapping-the-pathways-to-success

Keep Coding. (2022, 20 enero). *¿Qué es Machine Learning? ¿Cómo empezar a usarlo?* KeepCoding Tech School. https://keepcoding.io/blog/que-es-machine-learning/#:%7E:text=Python%20es%20sin%20duda%20el,%2C%20Shell%2C%20TypeScript%20y%20Scala.

Kelly, J. (2020) *This is the ever-changing state of the world's top cities.* Word Economic Forum. https://europeansting.com/2020/02/07/this-is-the-ever-changing-state-of-the-worlds-top-cities/

León, E. (2021). *Métricas para la validación de clustering.* Universidad Nacional de Colombia. https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf

matplotlib. (2022). *matplotlib.pyplot — Matplotlib 3.5.1 documentation*. Matplotlib https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html

Ministerio de Asuntos Económicos y Transformación Digital. (2014, 19 septiembre). *Informe: Open Data como herramienta para Smart Cities*.datos.org https://datos.gob.es/es/noticia/informe-open-data-como-herramienta-para-smart-cities

Ministerio de Asuntos Económicos y Transformación Digital. (2017, 5 junio). *Informe Datos abiertos y Ciudades inteligentes: una visión alternativa desde el Derecho*. Datos.org https://datos.gob.es/es/noticia/informe-datos-abiertos-y-ciudades-inteligentes-una-vision-alternativa-desde-el-derecho

Muaz, U. (2019, 25 julio). *Autoencoders vs PCA: when to use ? - Towards Data Science*. Towards Data Science. https://towardsdatascience.com/autoencoders-vs-pca-when-to-use-which-73de063f5d7

Numpy. (2022) *NumPy Documentation*. (2022). numpy.org. https://numpy.org/doc/

Nvidia (2020). *¿Qué es la computación acelerada por GPU?.* nvidia. https://www.nvidia.com/es-la/drivers/what-is-gpu-computing/#:~:text=Una%20CPU%20tiene%20unos%20cuantos,varias%20tareas%20al%20mismo%20tiempo.

Ondiviela, J.A. (2021) Beyond SmartCities: Creating Attractive Cities for talented citizens. Springer.https://doi.org/10.1007/978-3-030-83371-8

Parlina, A, Ramli, K. & Murfi, H. (2021). Exposing emerging trends in smart sustainable city research using deep autoencoders-based fuzzy C-means, *Sustainability*, 13(5), 2876

Patel, H. (2017, 25 julio). *SELU vs RELU activation in simple NLP models*. Hardikp. https://www.hardikp.com/2017/07/24/SELU-vs-RELU/

RAE. (2014). *talento | Diccionario de la lengua española*. «Diccionario de la lengua española» - Edición del Tricentenario. https://dle.rae.es/talento

Roy, A. (2020, 12 diciembre). *Introduction To Autoencoders*. Towards Data Science. https://towardsdatascience.com/introduction-to-autoencoders-7a47cf4ef14b

scikit-learn.org. (2022). *sklearn.preprocessing.StandardScaler*. Scikit-Learn.

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

UFV, 2021 *World Wide Observatory for Attractive Cities.* ufv.es http://ddfv.ufv.es/handle/10641/2619

Ullman, S., Poggio, T. , Harari, D. , Zysman, D. & Seibert, D. (2014) *Unsupervised Learning Clustering.* MIT.edu http://www.mit.edu/~9.54/fall14/slides/Class13.pdf

8. ANNEXES

Annex 1: Magnetism indicators

Area	W	Subarea	Class	Indicator
Magnetism	User Input	Identity	History. Culture	Age
				UNESCO
				Top Museums
			Government Basics	Democracy Index
				Safe City Index
			Reputation	Reputation
			Space. Density	% Natural Space
				Density (inh/km2)
			Climate	Avge. Temperature Desviation
				Avge. Precipitation Desviation
				Avge. Daily Sunshine
			Geo Risk	Natural Disaster Risk
			GeoEconomics	GDP Proximity
			Gastronomy	Food Security Index
				Guru Restaurant
				Michelin Guide and Guru

			Branding. External Image	Music
				Movies
				Sports
				Main Events
	User Input	Dynamism	Competitiveness	Creativity Index
				Global Competitivenes
				Cities In Motion
				Global Talent Competitiveness
			Expat Social Experience	Life Style - Quality
				People Around
				Relationship - Social Life
			Ethics. Well-being	Happiness
				World Giving Score
				Civic Engagement

				Work-Life Balance
			Equality	GINI Index
				Gender
				Tolerance
				Poverty
	User Input	Strategy	Human Capital	Population Age Average Per Country
				Ranking Human Capital
			Smart Cities Plan	Plan Smart Cities
			Innovation	R&D (% GDP)
				Global Innovation Index
				Innovation Cities

Annex *2*: Profitability indicators

Area	W	Subarea	Class	Indicator
Profitability	Services	User Input	Digital Government	Online Service Index
				eParticipation Index
				Digitalization of Government
		User Input	Education. LifeLong Training	Quality of Management Schools
				Prevalence of Training in firms
				Employee Development
		User Input	Employability	LinkedIn Talent Hiring Demand
				Employability
		User Input	Connected City	4G LTE
				5G LTE
				Internet Speed
				ICT Infraestructure
		User Input	Health/Social SVS	Social Expenditure (% GDP)
				Life Expectancy at age 60
				Physicians (per 1k)

				Public Health Expenditure (%GDP)
		User Input	Environmental Sustainability	Carbon Neutrality Plan
				Sustainable City Index
				Environment
		User Input	Culture-Tourism	Culture Creative Jobs %
				City Destination
		User Input	Urban Mobility	Smart Parking
				Car Sharing Services
				Traffic INRIX Congestion
				Mobility and Transportation
		User Input	Urban Planning	Urban Planning
		User Input	Safety	Safe Cities Index
				Cities In Motion
				Personal Safety
	Cost Of Living. Net Purchase Power	50	Net Real Income	Avg Wages/month
				Direct Tax + Social Contributions
				Indirect Tax
		50	Cost Of Life	Purchase Power Parity Plus Rent (NY=1)

Annex 3: Health Services Index Top 16

Country	Health services index
Norway	10.00
France	9.95
Austria	9.82
Sweden	9.72
Denmark	9.56
Germany	9.45
Belgium	9.27
Italy	9.17
Finland	9.10
Luxembourg	8.70
Spain	8.47
Greece	8.42
Japan	8.37
Portugal	8.03
Australia	7.82
United States	7.75

Annex 4: EXPAT Experience Index Top 15

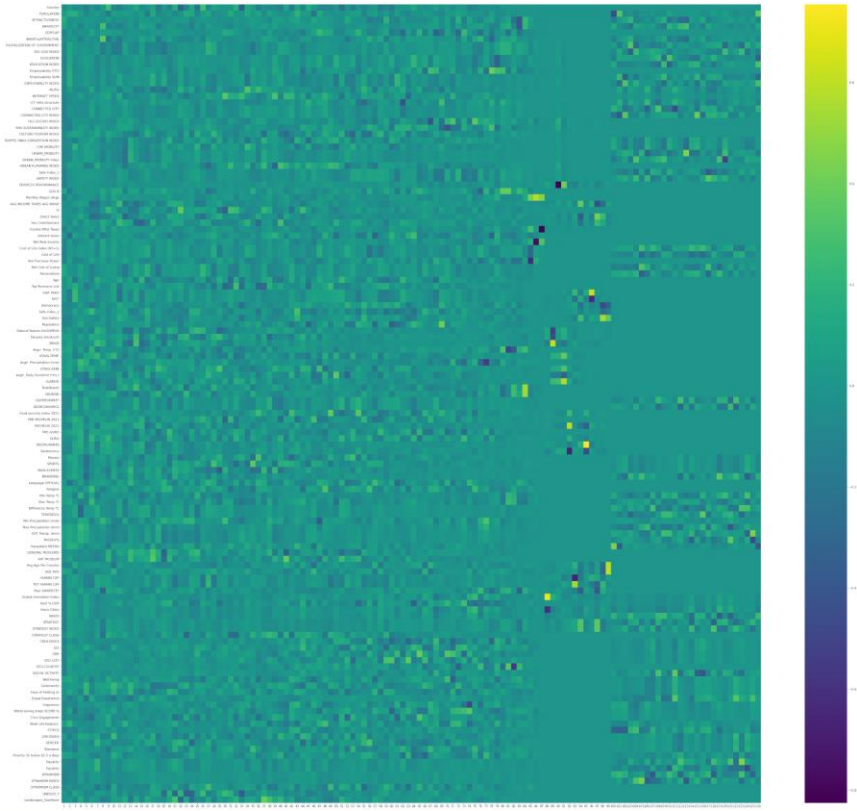
Country	Expat Experience NOR
Spain	10.00
Portugal	9.13
New Zealand	8.35
Bahrain	7.91
Singapore	7.74
Malaysia	7.74
Argentina	7.65
Uruguay	7.65
Argentina	7.65
United Arab Emirates	7.30
Greece	6.78
Croatia	6.78
Panama	6.78
Bulgaria	6.70
Australia	6.61

Anexo 5: Innovation Index Top 17

Country	Índice de Servicios de Innovación
Switzerland	10.00
Sweden	9.30
South Korea	8.86
United States	8.64
Israel	8.40
Denmark	8.26
Germany	8.04
Finland	7.96
Netherlands	7.83
United Kingdom	7.82
Singapore	7.56
Japan	7.42
Taiwan	7.10
Hong Kong	7.08
France	7.00
Austria	6.92

China	6.89

Annex 6: PCA information collected



Annex 7: QR Code for the app

Annex 8: Economic results

Country	Net Purchase Power
Kuwait	7.700
United States	4.824
United States	4.820
Bahrain	4.748
United States	4.729
Qatar	4.644
United States	4.529
United States	4.510
United States	4.368
United States	4.291
Australia	4.285
Switzerland	4.267
United States	4.201
Australia	4.180
United Kingdom	4.126
United States	4.099
Switzerland	4.087
Australia	4.073

Annex 9: Attractivity Ranking results

City	Country	MAGNETISM	IDENTITY	DYNAMISM	STRATEGY	PROFITABILITY	PERFORMANCE	NetPurchase Power	ATTRACTIVENESS
Zurich	Switzerland	12	12	31	45	9	22	21	1
Amsterdam	Netherlands	1	18	1	8	43	13	59	2
Kansas City	United States	42	60	77	13	4	42	8	3
London	United Kingdom	2	1	60	22	45	7	63	4
Berlin	Germany	14	24	47	17	19	11	41	5
Sydney	Australia	13	53	8	29	23	41	26	6
Melbourne	Australia	24	61	14	43	15	59	14	7
Copenhagen	Denmark	3	45	4	2	54	1	70	8
Den Haag	Netherlands	18	68	5	50	18	32	27	9
Bern	Switzerland	51	47	39	74	6	34	12	10
Dallas	United States	82	110	77	35	1	27	2	11
Vienna	Austria	16	11	50	47	22	6	47	12
Rotterdam	Netherlands	15	64	6	33	27	19	45	13
Washington, D.C.	United States	27	21	75	32	16	24	30	14
Basel	Switzerland	34	27	27	75	12	27	22	15
Toronto	Canada	10	43	11	38	37	45	34	16
Phoenix	United States	84	132	82	5	2	50	3	17
Manchester	United Kingdom	23	42	59	12	35	49	33	18
Atlanta	United States	61	100	83	9	10	58	10	19
New York City	United States	8	10	56	14	59	17	64	20
Adelaide	Australia	57	107	12	58	14	71	11	21
Chicago	United States	38	56	70	16	28	43	29	22
Aarhus	Denmark	9	51	9	26	57	12	66	23
Tokyo	Japan	60	35	96	46	13	8	35	24
Hamburg	Germany	37	44	48	53	30	18	46	25
Geneva	Switzerland	48	19	35	95	21	69	17	26
Oslo	Norway	25	37	30	55	42	14	56	27
Stockholm	Sweden	5	24	22	3	70	8	78	28
Eindhoven	Netherlands	17	72	7	37	53	39	52	29
Glasgow	United Kingdom	52	67	61	31	25	56	23	30
Edinburgh	United Kingdom	21	32	58	21	51	66	44	31
Helsinki	Finland	7	80	10	1	67	2	80	32
Houston	United States	90	113	79	51	5	69	5	33
Miami	United States	69	94	81	27	17	56	16	34
Cologne	Germany	32	37	51	44	44	53	42	35
Montreal	Canada	65	75	15	92	20	65	19	36
Denver	United States	88	129	85	15	7	26	13	37
Birmingham	United Kingdom	54	64	71	34	31	55	28	38
Los Angeles	United States	44	54	67	30	40	62	31	39
Malmo	Sweden	20	57	26	24	61	20	65	40
San Francisco	United States	33	58	68	6	52	35	54	41
Espoo	Finland	22	86	18	18	63	3	75	42
Stavanger	Norway	45	82	28	49	46	14	57	43
Ottawa	Canada	71	99	17	79	32	61	25	44
Belfast	United Kingdom	75	101	64	36	29	80	15	45
Bergen	Norway	66	83	28	72	39	14	53	46
Gothenburg	Sweden	19	75	32	10	69	27	69	47
Las Vegas	United States	101	127	88	68	3	36	7	48
Baltimore	United States	97	105	86	77	8	48	9	49
Munich	Germany	41	29	44	71	60	30	60	50
Frankfurt	Germany	70	48	54	91	38	23	48	51
Wellington	New Zealand	55	96	2	83	55	46	49	52
Tampere	Finland	43	113	18	18	64	4	73	53
Stuttgart	Germany	74	52	55	87	41	43	40	54
Yokohama	Japan	85	97	100	23	26	31	35	55
Bristol	United Kingdom	67	73	64	48	50	68	38	56
Luxembourg	Luxembourg	50	40	33	85	62	51	55	57
Oulu	Finland	39	116	18	11	68	4	79	58

City	Country	MAGNETISM	IDENTITY	DYNAMISM	STRATEGY	PROFITABILITY	PERFORMANCE	NetPurchase Power	ATTRACTIVENESS
Boston	United States	63	74	74	40	58	73	43	59
Liverpool	United Kingdom	76	62	64	67	47	83	24	60
Canberra	Australia	91	121	12	99	24	71	18	61
Philadelphia	United States	86	90	87	65	36	82	20	62
Dublin	Ireland	26	41	25	56	72	83	58	63
Nottingham	United Kingdom	79	68	63	69	49	78	32	64
Vancouver	Canada	92	128	16	86	33	38	37	65
Singapore	Singapore	56	135	21	7	66	8	76	66
Dusseldorf	Germany	94	68	53	115	34	32	39	67
Seattle	United States	87	108	76	42	48	36	50	68
Auckland	New Zealand	58	84	3	98	71	67	61	69
Paris	France	4	2	46	52	87	24	102	70
Barcelona	Spain	6	4	24	63	86	40	95	71
Linz	Austria	83	48	52	113	65	21	67	72
Valencia	Spain	31	17	34	79	77	62	82	73
Antwerp	Belgium	59	37	99	41	74	59	74	74
Madrid	Spain	11	5	23	81	89	62	94	75
Seville	Spain	28	8	37	90	80	76	87	76
Zaragoza	Spain	29	9	38	88	83	88	82	77
Lyon	France	35	20	57	61	81	51	91	78
Brussels	Belgium	73	35	94	70	75	74	72	79
Málaga	Spain	36	23	36	76	85	86	86	80
Marseille	France	64	16	69	93	78	75	84	81
Nice	France	46	14	62	78	88	85	90	82
Honolulu	United States	107	157	88	64	56	47	51	83
Seoul	South Korea	30	15	112	4	92	79	100	84
Lille	France	80	50	72	84	84	87	85	85
Bordeaux	France	68	26	72	89	90	89	89	86
Osaka	Japan	104	106	104	73	73	53	71	87
Santander	Spain	47	33	41	82	96	92	98	88
Bilbao	Spain	53	22	41	96	95	91	98	89
Milan	Italy	49	6	105	62	98	80	109	90
Nagoya	Japan	99	109	107	57	82	77	88	91
Florence	Italy	40	7	115	39	101	100	104	92
Hong Kong	Hong Kong	77	86	95	25	97	96	96	93
Tel Aviv	Israel	93	129	91	20	94	97	92	94
Rome	Italy	62	3	111	102	104	98	106	95
Jerusalem	Israel	78	59	92	54	99	114	81	96
Lisbon	Portugal	81	46	43	114	103	90	111	97
Doha	Qatar	119	161	97	94	76	119	6	98
Porto	Portugal	72	30	49	107	107	99	110	99
Torino	Italy	95	12	116	106	100	101	101	100
Dubai	United Arab Emirates	108	150	40	103	93	105	77	101
Ljubljana	Slovenia	100	64	103	100	102	101	105	102
Abu Dhabi	United Arab Emirates	117	166	45	124	91	109	62	103
Manama	Bahrain	131	154	84	149	79	129	4	104
Kuwait City	Kuwait	159	149	145	161	11	130	1	105
Tallinn	Estonia	96	93	108	59	109	93	116	106
Prague	Czech Republic	89	34	98	97	113	95	121	107
Wroclaw	Poland	106	62	129	101	106	104	107	108
Taipei	Taiwan	103	120	131	28	110	94	117	109
Warsaw	Poland	113	78	120	131	108	103	112	110
Athens	Greece	105	31	128	128	114	110	114	111
Santiago	Chile	122	134	114	116	105	107	93	112
Shanghai	China	98	27	150	60	125	113	145	113
Budapest	Hungary	110	55	136	110	115	114	115	114
Bratislava	Slovakia	123	88	142	138	112	106	113	115
Vilnius	Lithuania	109	85	122	105	117	111	123	116
Buenos Aires	Argentina	111	104	80	140	121	114	135	117
Riga	Latvia	116	103	109	133	119	111	126	118

City	Country	MAGNETISM	IDENTITY	DYNAMISM	STRATEGY	PROFITABILITY	PERFORMANCE	NetPurchase Power	ATTRACTIVENESS
Zagreb	Croatia	114	95	124	123	122	120	120	119
Moscow	Russia	102	112	101	66	141	134	144	120
Beijing	China	118	75	153	109	128	120	138	121
Istanbul	Turkey	125	71	147	143	124	147	108	122
Mexico City	Mexico	115	78	119	139	142	139	139	123
Minsk	Belarus	138	132	121	155	116	143	97	124
Sofia	Bulgaria	112	80	127	112	147	133	150	125
Córdoba	Argentina	121	123	90	159	136	132	135	126
Montevideo	Uruguay	120	138	93	130	138	127	148	127
Suzhou	China	130	91	163	129	130	124	128	128
Buchanest	Romania	128	111	133	142	133	138	118	129
St Petersburg	Russia	124	136	110	127	140	140	125	130
Kuala Lumpur	Malaysia	142	175	102	104	118	108	134	131
Chongqing	China	135	115	158	122	129	123	128	132
Shenyang	China	134	119	160	117	130	124	128	133
Guadalajara	Mexico	132	145	125	120	137	142	119	134
Tianjin	China	133	102	165	132	135	130	128	135
Guangzhou	China	145	147	157	110	120	117	122	136
Chengdu	China	141	116	156	141	126	122	128	137
Rio de Janeiro	Brazil	129	124	126	136	145	151	124	138
Shenzhen	China	146	152	155	108	127	118	143	139
Monterrey	Mexico	139	155	123	134	139	137	137	140
Wuhan	China	144	140	159	117	134	124	140	141
Riyadh	Saudi Arabia	166	172	132	167	111	146	68	142
Ankara	Turkey	152	125	154	166	123	152	103	143
Sao Paulo	Brazil	137	140	117	145	146	149	141	144
Kiev	Ukraine	127	118	118	146	152	145	163	145
Belgrade	Serbia	126	92	137	148	153	141	166	146
San José	Costa Rica	147	153	134	150	144	136	149	147
Bangkok	Thailand	136	139	106	147	151	135	165	148
Panama City	Panama	143	160	113	154	148	155	147	149
Brasilia	Brazil	153	168	130	158	143	148	127	150
Harbin	China	163	173	162	117	132	128	128	151
Cape Town	South Africa	140	151	138	121	163	153	173	152
Bogota	Colombia	151	159	148	135	154	157	162	153
Lima	Peru	161	143	164	151	150	162	146	154
Durban	South Africa	150	169	140	125	160	150	171	155
Johannesburg	South Africa	148	142	139	153	165	153	174	156
Tbilisi	Georgia	160	125	146	173	157	160	160	157
Quito	Ecuador	162	165	152	144	155	158	161	158
Tunis	Tunisia	155	88	167	175	164	168	164	159
Jakarta	Indonesia	149	146	161	126	174	161	175	160
Manila	Philippines	154	137	135	169	171	165	172	161
Hanoi	Vietnam	157	144	143	164	170	164	168	162
Casablanca	Morocco	156	122	168	160	172	172	158	163
Medellín	Colombia	170	174	151	152	149	144	159	164
Asuncion	Paraguay	164	148	149	168	168	166	167	165
Ho Chi Minh City	Vietnam	165	171	141	156	167	159	170	166
La Paz	Bolivia	167	161	166	162	166	171	154	167
Santo Domingo	Dominican Republic	168	163	144	171	162	156	169	168
New Delhi	India	171	167	175	137	156	163	152	169
Cairo	Egypt	158	97	171	165	175	175	156	170
Rabat	Morocco	169	131	169	172	173	173	155	171
Mumbai	India	172	158	172	157	159	167	157	172
Bangalore	India	173	156	173	163	161	170	153	173
Hyderabad	India	174	164	174	170	158	169	151	174
Accra	Ghana	175	170	170	174	169	174	142	175