BACHELOR

Visualization For Troubleshooting CSV Files

Mei, Jingxian

Award date:
2022

Visualization For Troubleshooting CSV Files

Data Science - Bachelor End Project

Jingxian Mei

Supervisor:
Dr. Ir. Stef van den Elzen

1 August 2022

Abstract

The paper illustrates the visualization of missing values and data type errors in a CSV file. This report contains two types of visualizations, a table visualization and a pie chart. The purpose of a table visualization is to provide users with an interactive way to correct errors in their CSV files by providing them with guidance. The purpose of the pie chart is to determine which type of error is most commonly occurring, out of two different types.

Keywords: interactive table visualization, errors in CSV, C#, CSVHelper.

Table of Contents

Abstract **2**

Introduction **4**

Related Work **7**

Current Methods 7
Issues in CSV 10
Select Errors to Visualize 7

Visualization Methodology **15**

Data 15
Mockup 16
Design Decisions 18
Tool Support and Goals 20

Implementation **21**

Attempt 1: Python 21
Attempt 2: C# 23

Evaluation **33**

User Experiment Set Up 33
Result Analysis 33

Conclusion **34**

Future Work **35**

References **36**

Appendix **39**

A. Coding 39
B. User Experiment Questionnaire 39
C. User Experiment Result 41

1. Introduction

Data analytics tools have made significant improvements for analytics and data management (Kandel et al., 2011). Therefore, the file system plays an important role in the storage and transmission of information, the value of information today is immeasurable. One of the most popular file formats is the Character Separated Values (CSV) file, it has a simple structure to store data and it can be used in different programming languages because of the simple storing structures, processing, and transmitting of information, it can be used such as Python, EXCEL, SQL, and Powershell (Abba & Hassan, 2018).

But the problem of troubleshooting files remains. Due to the simple storing structure, CSV files can easily hide errors. When people load CSV files, they are not always correctly structured, such as incorrect data types, unrecognized symbols, misspellings, duplicates, missing data, and outliers. This will cause the loading process to fail and might lose valuable information. These mistakes would discourage analyzers use CSV files (Abba & Hassan, 2018). Therefore, repairing mistakes in the CSV files is vital in this case. Despite the fact that analysts can manually check or fix it when they find errors, it is time-consuming. T. Dasy and T. Johnson estimate that data cleaning accounts for approximately 80% of the time that data warehousing projects take because it requires special scripts for programming languages such as Python and Excel. Due to this inconvenient fact, some people avoid data analysis (Kandel et al., 2011).

The most common errors in CSV files are file size too large, matching(this could be cells don't match column names, the column name addresses but give email address), translation(format does not translate correctly when it is opened on other platforms), values(data length is too long or too short when importing CSV files, or unacceptable characters), missing data(some information in the columns are missing), non-digestible formats(formats are not matched with it should be) (Lahar, 2020) Incorrect CSV files bring unexpected problems when people explore and analyze data. For instance, missing commas lead to the instance of CSV file meaningfulness or unusable. Typing error gives wrong information and is not valuable. The worst scenario is in the bank or university when people need to transfer money to a wrong number or read students' records as missing (Abba & Hassan, 2018).

On the other hand, problems associated with reformatting and validating data may be difficult to correct manually. Analysts, for instance, do not have standard rules to evaluate errors that occur when splitting the address and date of birth in records, and so their transformation may not be fully accurate. Likewise, transforming coded values is a time-consuming process, which is exhausting for analysts. When a person needs to find the FIPS code associated with a state name in the United States, they need to combine data from several tables. It will be even more challenging to modify the data if the data size and complexity grow (Carvalho et al., 2015).

In order to solve such problems, it is beneficial to provide techniques and tools for error checking and correction. I propose a method for detecting and correcting errors in CSV files using extensible table technology. In order to demonstrate the implementation of the table technique, an application is developed.

Detecting errors in the CSV files can be accomplished in a number of methods. In the paper, my approach is to develop an interactive visualization tool that can detect and correct errors in the CSV files, which would be useful to improve data quality, allowing analysts to correct the errors. Here is an overview of the tool. The application comes with a GUI which allows users to visualize the errors in the CSV files. The tool is used as such:

- Upon uploading the files, users are directed to the first window, which displays a table visualization and allows users to check the error position and edit the errors in the specified cells, these errors are color-coded.

- The second window allows users to check general overview of the error distribution.

Hence, the research question of this paper is*:*

- What visualization provides a clear overview of CSV files in a way that errors can be identified and corrected effectively?

And the sub-questions are:

- What are the suitable visualizations to represent errors?

- How does the tool help users to correct errors indicated by the visualization?

- What are the limitations of the tool?

In this paper, the structure is as follows: relevant work is discussed first, followed by an overview of visualization methodology that includes an introduction to data, mockup design, , and design decisions. Followed by implementation and their results. The next section is an evaluation of the tool. An illustration of a conclusion will be presented. As a final point, future work will be discussed.

2. Related Work

2.1 Current Methods

There are various approaches to visualizing errors in CSV files. In this report, I would like to discuss 5 methods here, they are Wrangler, Table Lens, Parser, and Regular Expression. I will first talk about their definitions, and then make a summary to discuss if the method suitable for this paper.

2.1.1 Wrangler

A declarative transformation language underlies the Wrangler solution, which generates a mixed-initiative user interface. A list of basic transforms is suggested as a way to identify the transformations that can be interactively transformed based on the user's selections of the data. More specifically, the selection of data means that users can upload CSV files and they can select rows, columns, and cells. In order to provide a high level of relevance to the possible transformations, he develops a model that ranks the options based on their frequency, diversity, and difficulty. To achieve high-quality data transformations, Wrangler provides users with natural language descriptions - which they can customize with visual previews of interaction parameters and transformation results. In contrast to using Excel, Wrangler's method decreases the running time and encourages the use of transformations. In addition to being easy to use, his method can be easily navigated by users, allowing them to complete the data cleaning tasks in a timely manner. However, when users are unfamiliar with transforms, it takes a longer period of time (Kandel et al., 2011). The method is therefore most suitable for people who are familiar with transforms when it comes to cleaning data, although it cannot detect errors.

2.1.2 Table Lens

A CSV file of a small size is easy to understand. However, as the number of rows and columns increases, a CSV file of a large size can be difficult to comprehend. Table Lens utilises a *fisheye* method to visualize and understand large files, which is a solution for massive tables (Carvalho et al., 2015). It can be shown a maximum of 660 cells at once on a size 19 inches screen in the spreadsheet, while the Table Lens method can display 100 times of the spreadsheet, and handle 30 times of the spreadsheet. With *Fisheye* or *focus + content* (Ramana & Card, 1995), visual distortion concepts are applied, meaning that the eyes are zoomed in on the center area while the other areas are zoomed out (Carvalho et al., 2015). In other words, as a result of these techniques, it is possible to interact with large information structures by dynamically distorting the spatial layout of the structure in accordance with the different levels of interest of each part (Ramana & Card, 1995). This characteristic can provide more details in the analytics and focus on a part of a table (Carvalho et al., 2015). Other than that, Table Lens is a good approach to looking for outliers and patterns in multivariate tables, it is specialized for numerical and categorical data (Rao & Card, 1994). In conclusion, Table Lens is suitable for a large file and can focus on a specific part of the file in error detection.

2.1.3 Parser

The definition of a parser is that it is a program that recognizes valid syntax, and it can be used for all context-free grammar. Syntax specifications are used by the program parser as part of the grammar specification. Grammars are the rules that define valid syntax (Abba & Hassan, 2018). In simple terms, parsers use grammar to check if there are any mistakes in the CSV files. Nevertheless, it is not easy to define a suitable parser if the CSV files do not exist CFG. Therefore, if people want to use Parser to do error detection, they have to prove the CSV files have CFG (Abba & Hassan, 2018). This method is not suitable for doing error checking in this paper, because this method needs two works - prove CSV files have a CFG and find a suitable grammar to detect eros. Moreover, it is difficult to find suitable grammar for error detection.

2.1.4 Regular Expression

A common and powerful approach is Regular Expression. The regular expression can be used to correct data such as address and date or to correct grammatical errors. However, this approach can not detect hidden errors because of no standard writing formats. In the detail, there are many formats in the CSV files but they have the same pattern with slight differences in expression, regular expression can not detect the differences. The most common and popular format is the RFC4180 standard. (Abba & Hassan, 2018). Based on the RFC-4180 standard, the definition of CSV format can be as follows (Ge et al., 2019):

- File = [Header \n]Record \n ... \n Record Record=Field , ... , Field
- Field = Quoted | Unqoted
- Quoted = " Escaped...Escaped "
- Escaped = Char| " " | \n | , Unqoted = Char...Char

- Char = Any char except of " , \n , and ,

In summary, it is an option to detect simple pattern errors in this paper, but tested CSV files should have the same format. Therefore, if it is possible to test data type errors with it, it would be a good idea to do so. Data type errors are detected if there are any different data types in the cell compared to the whole column. For example, the data type of student number is an integer, but it detects one of the cells in the column exists a string data type. Here are some regular expressions that can be tested for wrong data types in the CSV files:

(1) Detect integer errors : ^\d+$
(2) Detect float errors: [+-]?([0-9]*[.])?[0-9]+
(3) Detect string errors: /^[A-Za-z]+$/

2.1.5 Pie Chat

Another name for a pie chart is the circle chart (Khan & Khan, 2011). Pie charts are used to make comparisons between parts and the entire picture (Sadiku et al., 2016). Pie chart circles contain a number of sectors, each of which reflects a percentage in a whole number. Pie chart controls are used to determine the size of data wedges as compared to other data wedges (Khan & Khan, 2011). It is one of the goals of this paper to show each error's distribution for a whole table, so that the users will be able to identify which error most frequently occurs. Due to this, a pie chart can be compared with each error, whereas there is no error within the entire CSV file, which is the characteristic of a pie chart. Additionally, Kiyono and his colleagues (Kiyono et. al., 2020) developed models for grammatical error correction in the datasets, and when they analyzed the error type distribution across different proficiency levels, they used the pie charts to illustrate it.

In conclusion, each method has its strengths and specializations. Wrangler is specialized for transforms, Table Lens is for large data sets, and it can especially analyze part of the data by applying the *Fisheye* method, regular expressions is for pattern errors with the same writing formats in the CSV files, Pie chart is for comparison between part and whole. In this paper, regular expressions for data type errors will be tested. Pie charts will be used for comparing the distribution of each error, no errors, within the whole table.

2.2 Issues In CSV

A discussion of common errors in CSV files is provided in this chapter. The studies conducted by Döhmen (Döhmen, 2017) were based on an analysis of 80 CSV files randomly chosen from data.gov.uk, which is one of the largest open data portals operated by the United Kingdom government. In his analysis, Döhmen identified 16 errors, which were the encoding issue, aggregated cell, visual table element, trailing whitespace, European thousand, separator, empty column, footnote, ragged row, multiple tables, title, multiple header rows, inconsistent data formatting, missing value qualifier, aggregated column, different variables in the same column, title, footnote, hierarchical column headers, spanning cells, and same variables in the multiple columns. It was stated by Döhmen that there are more errors occurring in the files, but these errors are most frequently occurring in the samples.

The following are definitions of the 16 errors.

- **Encoding Issue** The encoding issue refers to the file that can be encoded in a variety of versions, when people load it, it reads in different versions. This problem cause incorrectly loading and leads to wrong assumptions. The most common encoding is UTF-8, if all files are encoded as this, it will solve the problem. For example, as the picture is shown in Figure 2.2.1, the £ sign was read as the Polish letter Ł, then people would wrongly assume this is a Polish letter. This is because the file was ISO 8859-1 encoding, but read as encoding ISO 8859-16 (Döhmen, 2017). In addition, Mitlöhne and his partners (Mitlöhne et. al., 2016), concluded that in the Open Data portals, only 100k out of 200k CSV files can be correctly opened.

- **Visual Table Element** A visual table element error can be defined as a row that should not be in the table. This is because it disturbs users when they read the table and disturbs them to detect the type of data when parsing the table. As an illustration in Figure 2.2.2, an unnecessary white space row is in the first row (Döhmen, 2017).

- **Trailing Whitespace** This error refers to cells that have trailing whitespace to align the values for visual convenience, but this is not meaningful when processing data in the future (Döhmen, 2017).

- **European Thousand Separator** A thousand separators have different meanings in America and the European area. European use dots to separate thousand, while America uses dots to separate decimal numbers. In general, the American format is the most commonly used, but some packages apply the format based on the local of the system (Döhmen, 2017). It is possible for people to become confused by different systems.

- **Empty Column** A whole column is empty, but some empty columns have headers, which means that the information was not recorded correctly, the users should drop it or keep it (Döhmen, 2017).

- **Footnote** It is unnecessary to be included in the tables because it can disturb detecting data types (Döhmen, 2017).

- **Ragged Row** The rows have different numbers of delimiters (Döhmen, 2017). For example, a row only contains one cell.

- **Aggregated Cell** This is an issue that several columns in the table contain sums, and an additional row contains the sums of those columns. There is no need to include this cell in the table as it disrupts its rectangular shape. After parsing, it is easy to calculate totals or averages. An example is shown in Figure 2.2.2, 'T531,445' should not be in there, it destroys the rectangular shape (Döhmen,2017).

- **Multiple Tables** A table contains more than one table. As shown in FIGURE 2.2.3, the table has another table, it is most likely fail to detect data type (Döhmen, 2017).

- **Title** As shown in FIGURE 2.2.3, a title is a row, it is similar to footnote error, which disturbs data type detection (Döhmen, 2017).

- **Multiple Header Rows** The table's header is in ordered structure. As show in FIGURE 2.2.3, "Air", "Rail", "Taxi/Car" belong to "Travel" (Döhmen, 2017).

- **Inconsistent Data Formatting** In a currency row, some cells have the currency sign and some not, which is inconsistent formatting (Döhmen, 2017).

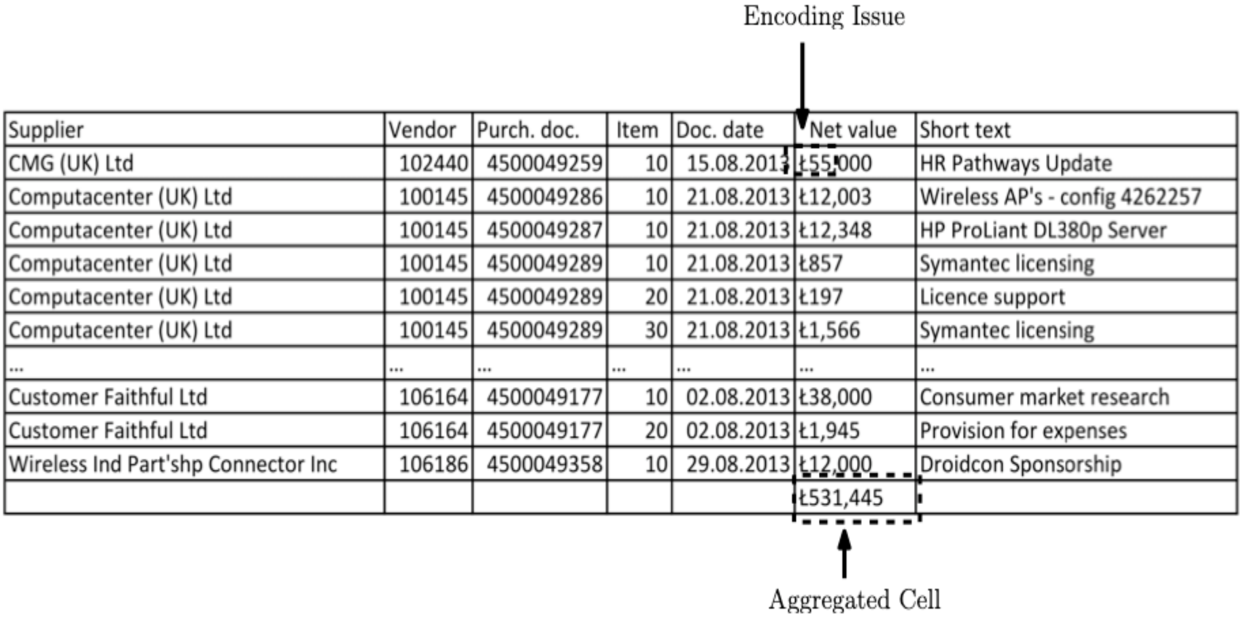
- **Missing Value Qualifier** Missing values in the cell, it can be denoted as “NA”, “NaN”, “-999”, ‘-1’, and “-”. The detection of data types may be impacted if NA values are not detected (Döhmen, 2017).

- **Aggregated Column** Cells have unnecessary information, which can be produced after loading (Döhmen, 2017).

- **Different Variables in the Same Column** In the FIGURE 2.2.4, a column contains different types of variables.

- **Spanning Cells** In most cases, spanning cells result from exports from formats that support spanning cells (Döhmen, 2017).

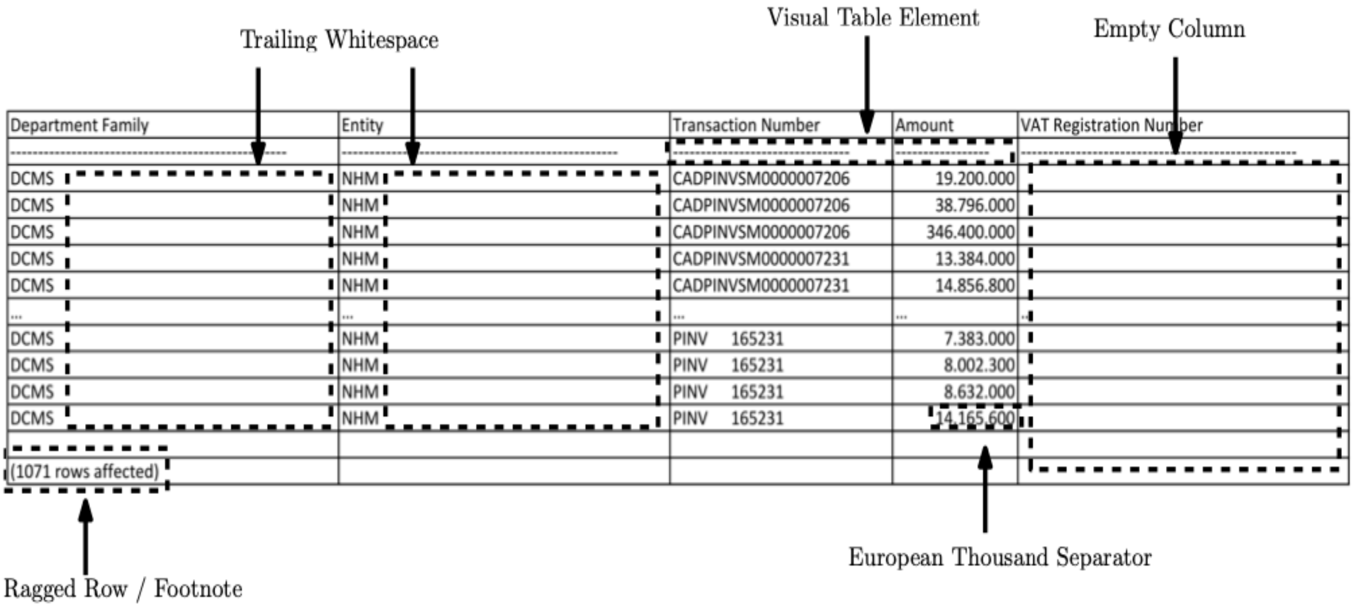
- **Same Variable in Multiple Columns** This is an issue that a table has more than one column that has the same content. For example, in Figure 2.2.4, the last row has 2 two repeated columns, which are “headcount” and “Full-time equivalent” (Döhmen, 2017).



Encoding Issue
Aggregated Cell

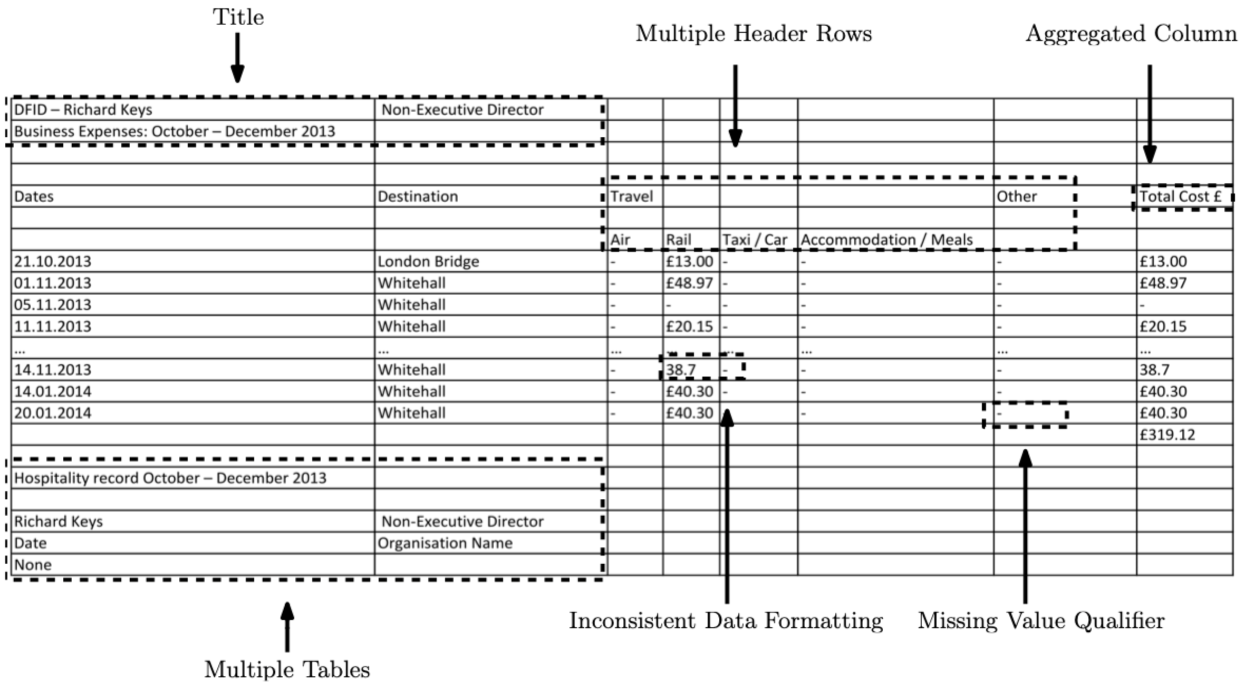
Supplier	Vendor	Purch. doc.	Item	Doc. date	Net value	Short text
CMG (UK) Ltd	102440	4500049259	10	15.08.2013	ł55,000	HR Pathways Update
Computacenter (UK) Ltd	100145	4500049286	10	21.08.2013	ł12,003	Wireless AP's - config 4262257
Computacenter (UK) Ltd	100145	4500049287	10	21.08.2013	ł12,348	HP ProLiant DL380p Server
Computacenter (UK) Ltd	100145	4500049289	10	21.08.2013	ł857	Symantec licensing
Computacenter (UK) Ltd	100145	4500049289	20	21.08.2013	ł197	Licence support
Computacenter (UK) Ltd	100145	4500049289	30	21.08.2013	ł1,566	Symantec licensing
...	...	...	...	...	...	...
Customer Faithful Ltd	106164	4500049177	10	02.08.2013	ł38,000	Consumer market research
Customer Faithful Ltd	106164	4500049177	20	02.08.2013	ł1,945	Provision for expenses
Wireless Ind Part'shp Connector Inc	106186	4500049358	10	29.08.2013	ł12,000	Droidcon Sponsorship
					ł531,445	

FIGURE 2.2.1 ENCODING ISSUE AND AGGREGATED CELL (DÖHMEN, 2017).



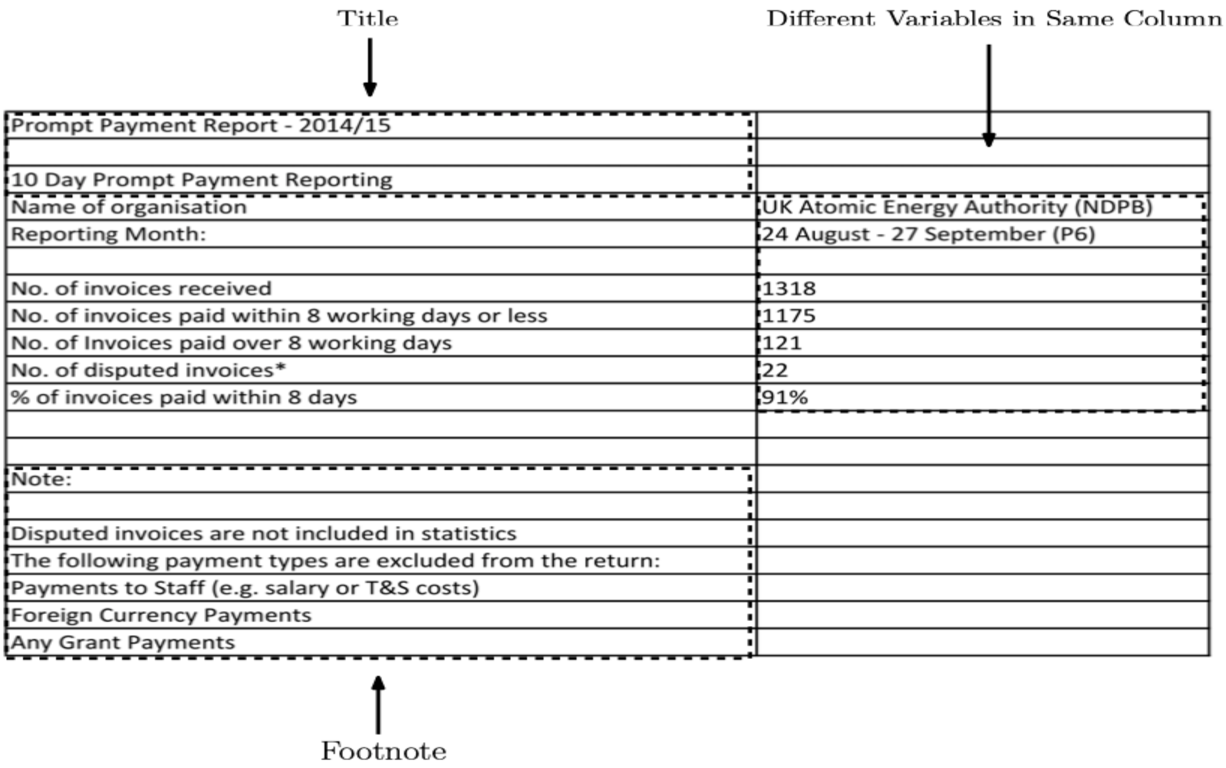
Department Family	Entity	Transaction Number	Amount	VAT Registration Number
--	--	--------------------------	----------------	--
DCMS	NHM	CADPINVSM0000007206	19.200.000	
DCMS	NHM	CADPINVSM0000007206	38.796.000	
DCMS	NHM	CADPINVSM0000007206	346.400.000	
DCMS	NHM	CADPINVSM0000007231	13.384.000	
DCMS	NHM	CADPINVSM0000007231	14.856.800	
...	...	...	...	...
DCMS	NHM	PINV 165231	7.383.000	
DCMS	NHM	PINV 165231	8.002.300	
DCMS	NHM	PINV 165231	8.632.000	
DCMS	NHM	PINV 165231	14.165.600	
(1071 rows affected)				

FIGURE 2.2.2 RAGGED ROW/FOOTNOTE, TRAILING SPACE, VISUAL TABLE ELEMENT, EUROPEAN THOUSAND SEPARATOR, AND EMPTY COLUMNS (DÖHMEN, 2017).



DFID – Richard Keys	Non-Executive Director						
Business Expenses: October – December 2013							
Dates	Destination	Travel				Other	Total Cost £
		Air	Rail	Taxi / Car	Accommodation / Meals		
21.10.2013	London Bridge	-	£13.00	-	-	-	£13.00
01.11.2013	Whitehall	-	£48.97	-	-	-	£48.97
05.11.2013	Whitehall	-	-	-	-	-	-
11.11.2013	Whitehall	-	£20.15	-	-	-	£20.15
...	...	...	...	...	...	...	...
14.11.2013	Whitehall	-	38.7	-	-	-	38.7
14.01.2014	Whitehall	-	£40.30	-	-	-	£40.30
20.01.2014	Whitehall	-	£40.30		-	-	£40.30
							£319.12
Hospitality record October – December 2013							
Richard Keys	Non-Executive Director						
Date	Organisation Name						
None							

FIGURE 2.2.3 TITLE, MULTIPLE TABLES, MULTIPLE HEADER ROWS, INCONSISTENT DATA FORMATTING, AGGREGATED COLUMNS, AND MISSING VALUE QUALIFIER (DÖHMEN, 2017).



Prompt Payment Report - 2014/15	
10 Day Prompt Payment Reporting	
Name of organisation	UK Atomic Energy Authority (NDPB)
Reporting Month:	24 August - 27 September (P6)
No. of invoices received	1318
No. of invoices paid within 8 working days or less	1175
No. of Invoices paid over 8 working days	121
No. of disputed invoices*	22
% of invoices paid within 8 days	91%
Note:	
Disputed invoices are not included in statistics	
The following payment types are excluded from the return:	
Payments to Staff (e.g. salary or T&S costs)	
Foreign Currency Payments	
Any Grant Payments	

Figure 2.2.4 TITLE, Footnote, and Different Variable In Same Column (Döhmen, 2017).

2.3 Select Errors to Visualize

There are lots of issues in the CSV files, here are selections of errors to visualize, they are missing value qualifiers, and different variables in the same column. This section will explain why these errors are chosen for this paper.

The reason for choosing the missing value qualifier is that, according to Döhmen, not properly detecting NaN values will cause failure to detect data type. Moreover, missing values should not be deleted just because this is a NaN cell (Döhmen, 2017). According to (Mitlöhne et. al., 2016) parsing CVS statistics, NA values takes 23% of 200k CSV files. Therefore, I think this is necessary to detect NA values.

The second error is different variables in the same column, based on Döhmen’s statement, if this error exists, it is hard to decide the data type for the column (Döhmen, 2017). In addition, according to (Mitlöhne et. al., 2016), a column consisting of multiple data types is not logical, and this mistake is very common in the Open Data portals.

3. Visualization Methodology

3.1 Data

A total of 9 CSV files have been randomly selected to be tested, which come from Open Data portal data.gov.uk, Kaggle, and stats.govt.nz. The majority of these CSV files exceed 10,000 rows in order to achieve the goal of handling large datasets. In terms of detect data type errors and empty cell errors, selected files should at least contain those errors. In terms of size, the largest CSV file is small business compete, which contains approximately 32,0000 rows with 18 columns. In terms of number of columns, rows and datatypes, which shows in TABLE 3.1.1. According to (Mitlöhne et. al., 2016), they parsed 200k CSV files and found that data types of distribution are numeric, date, and character strings. Therefore, numeric, date, and string are checked. Data types are detected in the tool.

Details about checking data type will be discussed in the implementation chapter. In the table, it shows the details of the CSV files are selected.

Most of file's names are too long to display in the table below, here are the file names for each file:

- File 1: Biotoxin+Results+2021+160322
- File 2: annual-enterprise-survey-2021-financial-year-provisional
- File 3: annual-enterprise-survey-2021-financial-year-provisional-size-bands
- File 4: bos2021ModC
- File 5: data
- File 6: error
- File 7: research-and-development-survey-2021
- File 8: time_series_covid19_confirmed_US
- File 9: time_series_covid19_deaths _US

	Number of rows	Number of columns	String	Numerical	Date
File 1	966	15	Yes	Yes	
File 2	47,175	10	Yes	Yes	
File 3	17,028	7	Yes	Yes	
File 4	13,489	6	Yes	Yes	
File 5	64,958	14	Yes	Yes	
File 6	323,418	18	Yes	Yes	Yes
File 7	27664	8	Yes	Yes	
File 8	3262	135	Yes	Yes	
File 9	3262	136	Yes	Yes	

TABLE 3.1.1 DETAILS ABOUT THE CSV FILES.

3.2 Mockup

This section presents a mockup of the GUI, which may not be the actual implementation. The actual design can be found in the chapter on Implementation.

The first step in the process is to design a graphical user interface (GUI), which allows users to navigate the tool. A mockup of the upload page is shown in FIGURE 3.2.1, to the left is a relevant visualization picture, as Galitz (Galitz, 2007) explained that "When all aspects of an image or object are relevant, present a picture or photograph of it." This is because pictures and photographs can draw attention to the subject matter. You will also find a button on the right side that allows you to upload a CSV file. As soon as the CSV file has been uploaded, the user will see FIGURE 3.2.2, which shows the new window with two buttons and a title of "Error CSV visualization". The first of button which is labelled 'Error Distribution', which presents the error distribution of the cells in the CSV file. The number of cells for each error is displayed in a pie chart, color-coded to indicate the most frequent errors out of selected errors to visualize. If the user clicks on the 'Table View' button in FIGURE 3.2.3, a table visualization appears, in which each color corresponds to a specific error. This can assist the user in identifying the error. Furthermore, users have the option to modify errors directly on the GUI by clicking the cell and entering the correct data. Once the errors have been modified, users have the option of downloading the data as a CSV file. When the number of rows and columns increases, users must scroll data to modify the data on the GUI, which can become very time consuming. The solution is to click the cell in order to acquire the row number and column number, so they may correct errors in EXCEL or other software applications. When row and column sizes increase, users will be able to view a scaled table view on the right side, known as the grid view, as shown in FIGURE 3.2.3. It represents the same meaning as the table view, but this is scaled smaller.

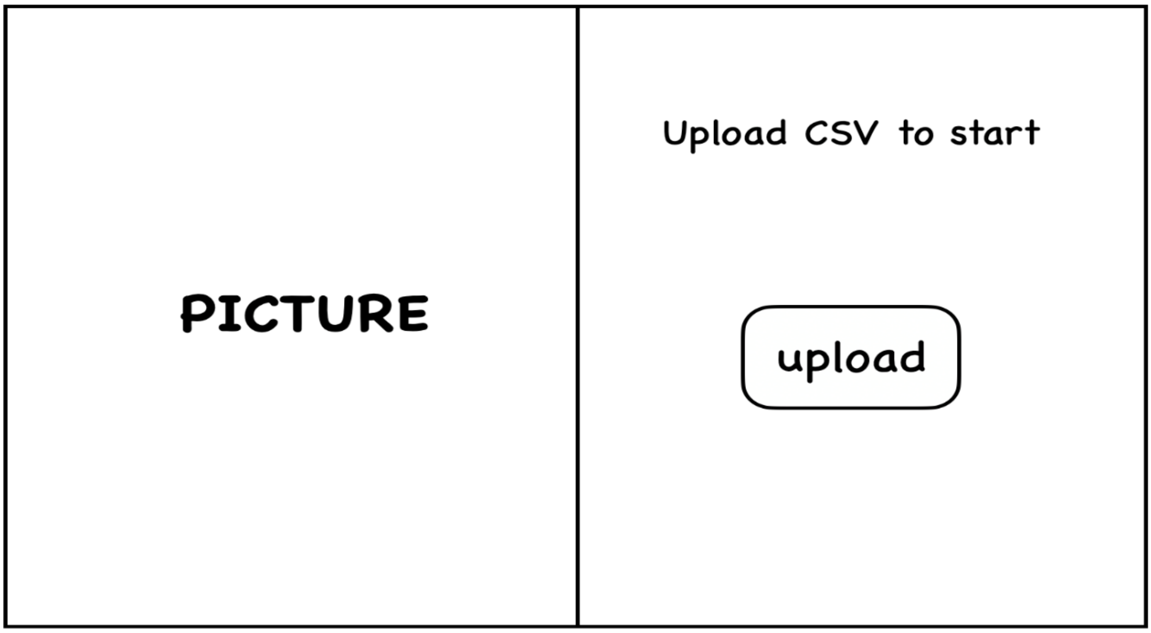


FIGURE 3.2.1. UPLOAD WINDOW

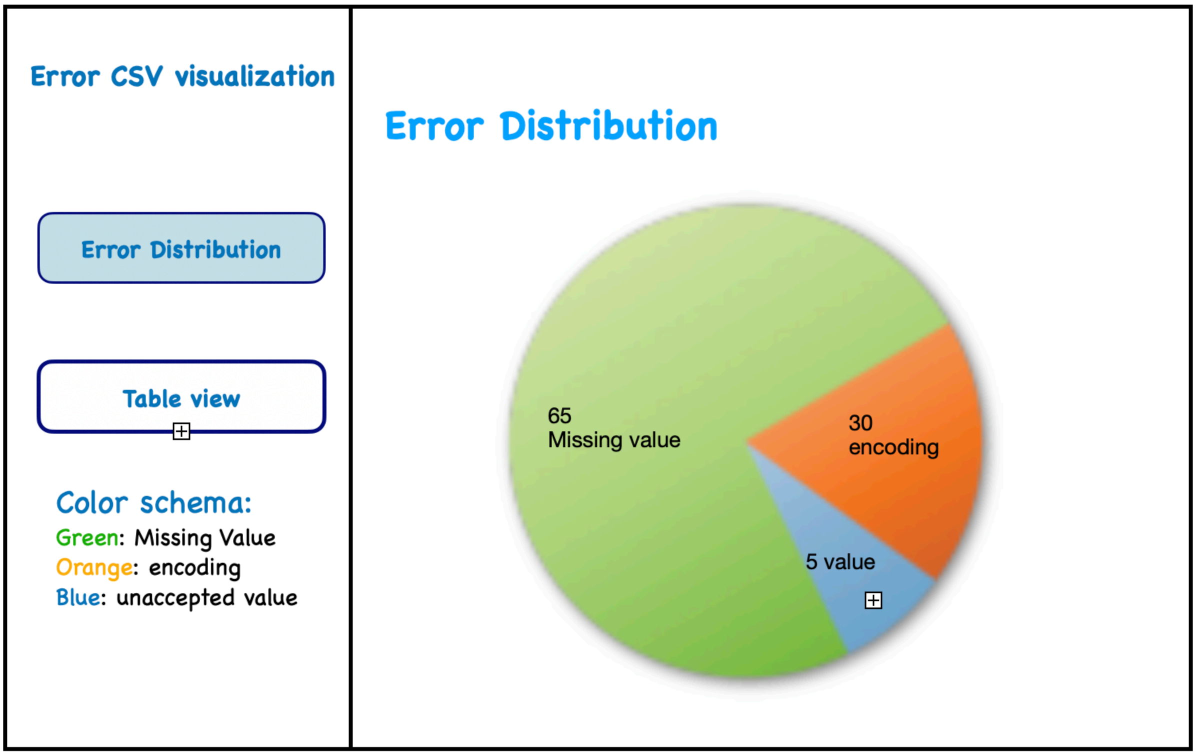


FIGURE 3.2.2 PIE CHART WINDOW

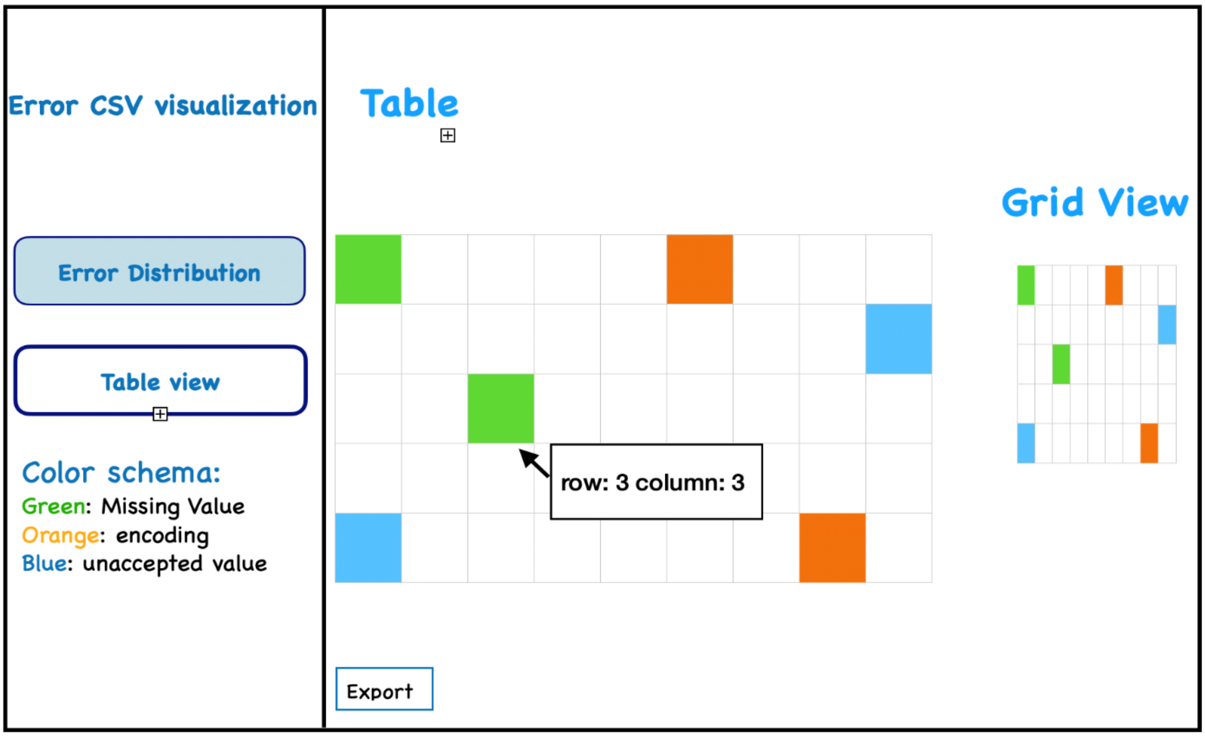


FIGURE 3.2.3. TABLE VIEW WINDOW

3.3 Design Decisions

There are lots of methods to display the visualizations. In this section, I will discuss why I designed a GUI, selected the scalable grid view, loaded data into the table, and the color scheme for each error.

3.3.1 GUI

GUI can enhance the tool's effectiveness, a good GUI design can reduce the 25% of processing time and decrease 25% of chance to make mistakes Galitz (Galitz, 2007). The GUI allows users to construct visualizations without coding and require minimal knowledge (North et. al., 2002), users just need to click buttons to upload file, visualize errors in the CSV files and edit cells. Visualizations and uploading of files can be integrated in a GUI, which complys the goal 7 in section 3.4.

The GUI is shown in FIGURE 4.2.1. Based on the design principle of Galitz (Galitz, 2007), GUI has a background color can increase screen appeal, and different colors can separate the screen components. The background colors should be harmony, which means the colors are not sharp contrast to put together. Therefore, all backgrounds' colors are light. Secondly, the buttons should place into one group, therefore, "table visualization" and "error distribution" buttons belong to visualization part, which are together above the visualization's component, "Upload File" and "Export File" belong to file level component, they are left next to the visualization component. Other buttons below to edit visualization part, which are together below the visualization's component, the main visualizations are in the middle of the windows form.

3.3.2 Table Visualization

According to Hoffman (Hoffman, 2000), table visualization refers to the presentation of data in the form of a table. In a table format, the information is arranged according to relationships between rows and columns, in which the row refers to independent variables in records or "Dimension", and the column refers to dependent variables or records or "Variates". In this structure, rows represent records, and columns represent variables, depending upon attributes and data. (Khan & Khan, 2011). Additionally, the table format can be used in the same manner as Excel, which requires little knowledge, and users are familiar with Excel, complies with goal 6 in section 3.4, and allows users to correct errors in the GUI.

3.3.3 Scaled Table

There is a challenge to check errors as the number of columns and rows increase, because scrolling down the CSV files in the GUI becomes time-consuming. Further, some users wish to view the distribution of errors in the tables, but they are only able to see a small portion of table. In addition, it might be overwhelming to look a large number of rows in a small GUI for users. In order to support this, as shown in FIGURE 4.2.2, we require a scaled table.

3.3.4 Error Color Schema

It has been stated by Galitz (Galitz, 2007) that the most appropriate colors to discriminate are red, yellow, green, blue and brown; the best colors to emphasize elements are white, yellow, green, blue, and red, which is from the least order. It is best to emphasize with bright colors in order to attract the viewer's attention. In this paper, error types are emphasized. Errors are therefore colored blue and yellow. Tables and pie charts indicate no mistakes by white and green, respectively. However, approximately 8 percent of males and 0.4 percent of females are color blind, most of whom are unable to recognize green, red, and blue colors (Galitz, 2007). In consideration of the fact that some people may have color-blindness, each color is labeled with the errors' name, which does not prevent them from using this tool.

3.4 Tool Support and Goals

Plotly and Tkinter are used in the interactive visualization developed for Python.

C# uses an external library called CsvHelper, which is used for reading and writing CSV files. This library is easy to use and can be used on most operating systems, it requires little memory to run, and it is compliant with RFC 4180 to ensure compatibility across systems (Close, 2009).

The general objective of this project is to develop an interactive tool in which users can fix errors in CSV files using these libraries. In addition, there are some sub-objectives for the visualization:

1. Developing user-friendly code (clean, simple)

2. The visualization tool is open source
3. GUI design is easy to use
4. Designing an interactive tool capable of handling large CSV files.
5. Creating visualizations that can aid the user in correcting errors
6. Visualizing information in a way that is easy to consume, allowing users to spend a few minutes or even obtain information with minimal knowledge.

7. Integrating upload file function and visualizations together

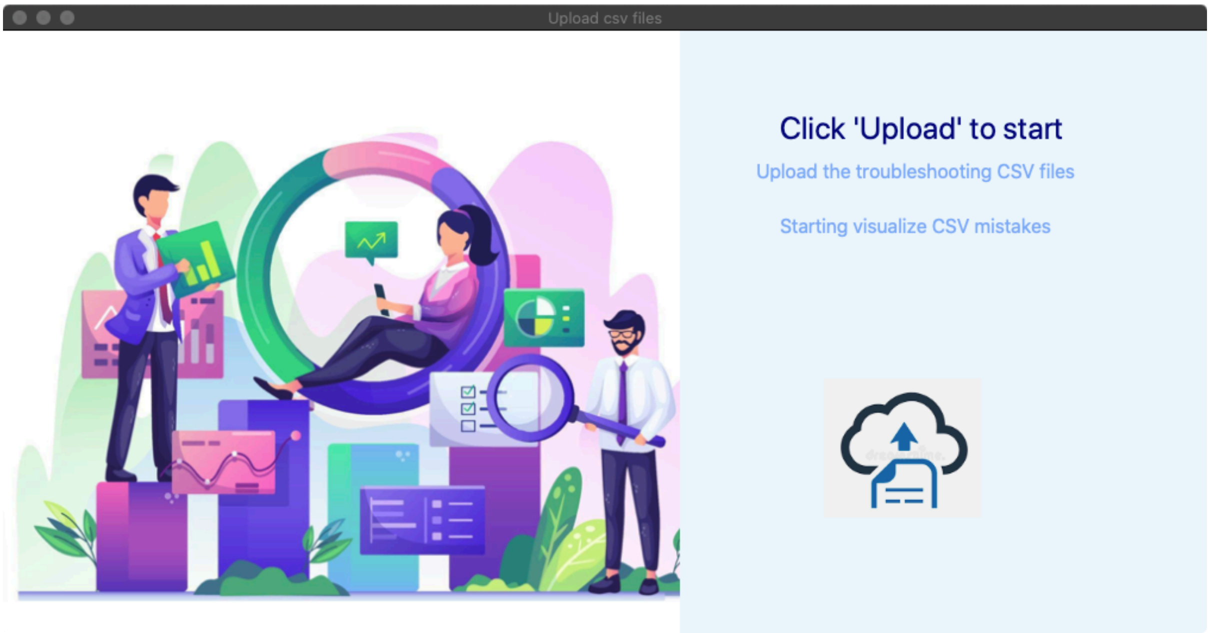
A variety of functions will be available in the visualization. First, a general overview is provided of the errors in the CSV files, with errors of wrong data type and missing value distribution. This uses a pie chart. A second function provides details of the errors. Each error is a different color, this can help users to check where the error is because each cell is clickable. For example, users can click on the cell, it shows column name, column number, and row number. This can help users to find the error position and modify them in their own platform if they want to. It is possible to determine which column or row has the error and fix it accordingly. In addition, users could edit the mistake values in the tables, and they are able to download it in CSV format.

4. Implementation

There are two attempts for this report, which are Python and C#. The first attempt Python failed because the Tkinter library and Plotly library cannot be connected, which fail to perform one of our goals – integrate upload files and visualizing it together. Therefore, only table visualization is finished in Python attempt, including show clicked column number and row number, edit cells, delete columns and rows, add columns and rows are implemented. The following sections will describe the two attempts. In the first attempt, available functions and GUI design are introduced. In the second attempt, firstly, I will talk about a summary of libraries used. Then, I will list the provided functions and describe how to use these functions. Finally, I will show the tool's results.

4.1 Attempt 1: Python

FIGURE 4.1.1 shows the Upload file window which is created by the Tkinter library. As the first step, we need to develop a GUI for users. It is possible to view the entire table on the GUI, as shown in FIGURE 4.1.2. Each color represents an error, color orange means empty cell error and color pink means data type error. Thus, users can clearly discern where errors are, as shown in FIGURE 4.1.3, they will be able to click on the cell to view the actual column number and row number. Furthermore, they will be able to modify their mistakes in the cell, accompanied by a guide. Following visualization, FIGURE 4.1.2, there is a "Export" button, users may download the modified file as CSV, which allows them to conduct further analysis, such as visualization, data cleaning, machine learning, etc. In addition, users can delete rows and columns if they click "x" button, they also could modify the column name.

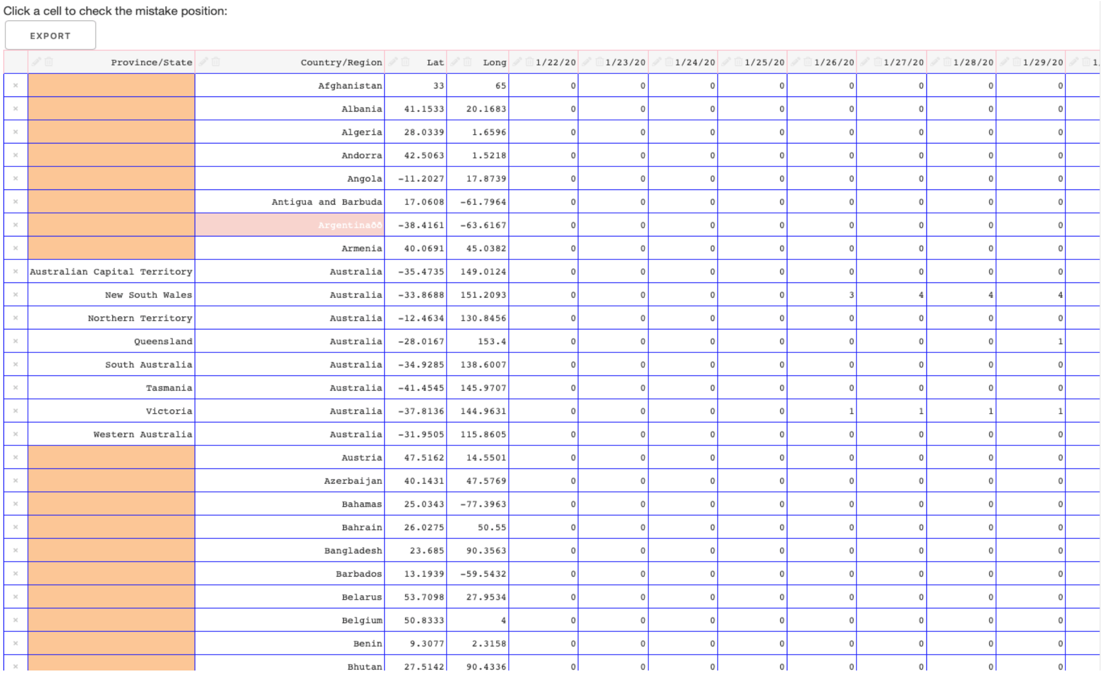


Click 'Upload' to start

Upload the troubleshooting CSV files

Starting visualize CSV mistakes

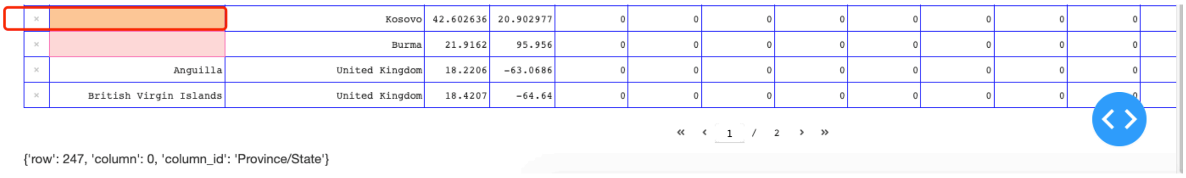
FIGURE 4.1.1. UPLOAD WINDOW



Click a cell to check the mistake position:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1
×		Afghanistan	33	65	0	0	0	0	0	0	0	0	
×		Albania	41.1533	20.1683	0	0	0	0	0	0	0	0	
×		Algeria	28.0339	1.6596	0	0	0	0	0	0	0	0	
×		Andorra	42.5063	1.5218	0	0	0	0	0	0	0	0	
×		Angola	-11.2027	17.8739	0	0	0	0	0	0	0	0	
×		Antigua and Barbuda	17.0608	-61.7964	0	0	0	0	0	0	0	0	
×		Argentinaðð	-38.4161	-63.6167	0	0	0	0	0	0	0	0	
×		Armenia	40.0691	45.0382	0	0	0	0	0	0	0	0	
×	Australian Capital Territory	Australia	-35.4735	149.0124	0	0	0	0	0	0	0	0	
×	New South Wales	Australia	-33.8688	151.2093	0	0	0	0	3	4	4	4	
×	Northern Territory	Australia	-12.4634	130.8456	0	0	0	0	0	0	0	0	
×	Queensland	Australia	-28.0167	153.4	0	0	0	0	0	0	0	1	
×	South Australia	Australia	-34.9285	138.6007	0	0	0	0	0	0	0	0	
×	Tasmania	Australia	-41.4545	145.9707	0	0	0	0	0	0	0	0	
×	Victoria	Australia	-37.8136	144.9631	0	0	0	0	1	1	1	1	
×	Western Australia	Australia	-31.9505	115.8605	0	0	0	0	0	0	0	0	
×		Austria	47.5162	14.5501	0	0	0	0	0	0	0	0	
×		Azerbaijan	40.1431	47.5769	0	0	0	0	0	0	0	0	
×		Bahamas	25.0343	-77.3963	0	0	0	0	0	0	0	0	
×		Bahrain	26.0275	50.55	0	0	0	0	0	0	0	0	
×		Bangladesh	23.685	90.3563	0	0	0	0	0	0	0	0	
×		Barbados	13.1939	-59.5432	0	0	0	0	0	0	0	0	
×		Belarus	53.7098	27.9534	0	0	0	0	0	0	0	0	
×		Belgium	50.8333	4	0	0	0	0	0	0	0	0	
×		Benin	9.3077	2.3158	0	0	0	0	0	0	0	0	
×		Bhutan	27.5142	90.4336	0	0	0	0	0	0	0	0	

FIGURE 4.1.2. TABLE VIEW WINDOW



×		Kosovo	42.602636	20.902977	0	0	0	0	0	0	0	0
×		Burma	21.9162	95.956	0	0	0	0	0	0	0	0
×	Anguilla	United Kingdom	18.2206	-63.0686	0	0	0	0	0	0	0	0
×	British Virgin Islands	United Kingdom	18.4207	-64.64	0	0	0	0	0	0	0	0

{'row': 247, 'column': 0, 'column_id': 'Province/State'}

FIGURE 4.1.3. SHOW SELECTED COLUMN AND ROW NUMBER

4.2 Attempt 2: C#

For C# attempt, the external library uses CsvHelper, which is a .Net library for reading and writing CSV files. The CsvHelper is for reading csv files in this paper, and the CSV file is loading into DataGridView windows form. To implement GUI part, no extra library is used, since C# has its built-in functions that allow users to drag certain tool (such as button, label) and put in the main windows form, which is very handy for APP design. And the background color, font can be set in the "property" menu. The built-in tool is called toolbox. For the rest of functions, only built-in functions are used.

Here is the main idea of detecting the type of data in a CSV file. Since we do not know the details of each column, we cannot detect it by simply using built-in functions, such as typeOf(). In short, we loop the table and count the data type of each cell on that column, and if the count is the highest, then that column is the data type. As an example, we set a variable named number of integers as 0 and a variable named number of strings as 0, then we loop the column of each cell to detect its data type and if one of the cells is integer, we set the number of integers plus one repeatedly, the same for the number of strings, and finally we compare the number of integers with the number of strings, and if the number of integers is greater than the number of strings, then we set the data type as int. To check each cell's data type, the built-in function is used. In addition, as mentioned in Chapter 2.1.4, 3 regular expression are tested for detecting data type, but this method fails. Moreover, the idea of detecting empty cell is using the built-in function. Different variable in the same column error simplify to "Data Type error" in this paper.

Here are available functions:

1. Load (large) CSV file.
2. Edit cell on the GUI
3. Add a row
4. Add a column
5. Delete a column
6. Delete a row
7. Show the selected row number and column number
8. Show table visualization
9. Show error distribution as a pie chart
10. Export corrected file as CSV format

Function 2, 3, 4, 5, 6 and 7 can help users to repair the CSV file. More specifically, 2, 3, 4,5,6 allow users to correct the mistakes for a small CSV file, such as 200 rows. While for large data, they can check function and modify the errors in the EXCEL, which takes less time.

The tool can used as such:

- The first step is to upload the CSV file (Figure 4.2.1), if it uploads successfully, users can see a table view and a grid view with the error color schema on the right side (FIGURE 4.2.2).

- The second step is checking the error distribution, this can be done with clicking “Error Cell Distribution” button. If users want to back the table visualization, they can click “Table Visualization”.

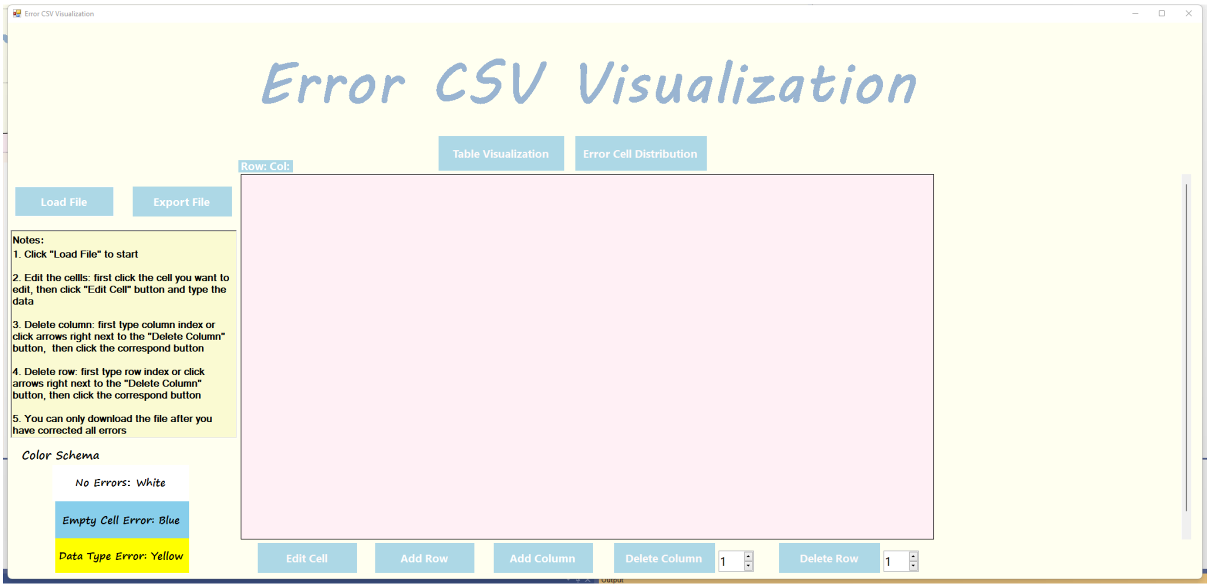
- The third step is to edit the error cells, users need to first click the cell then click the “Edit Cell” button, a message box will pop up and allow users to type the data. After users modify the data, error types are checked, and this will be updated in the table visualization and error cell distribution.

- The fourth step is to add column or a row, users can click button “Add Column” and “Add Row” to do so, if it is “Add Colum”, then users will be requested to fill in a non-empty column name. The users are allowed to delete rows and columns by their index, they can click the box with up or down to decide the row or column index, or they could input the index on that box. To clarify the index for delete row and column, (the index box is called NumericUpDown in C#), the NumericUpDown right next to “Delete Column” is for delete column and the NumericUpDown right next to “Delete Row” is for delete row.

- The last step is export file, files can be exported only if all errors are corrected, otherwise users will be informed by a message box.

Further, every time a user edits a table, including editing cells, deleting/adding columns, deleting/adding rows, the inputs will be checked to ensure that they are correct, which can be observed by looking at the table visualization and error cell distribution. If a user adds a column, it is initially empty. By default, this column is colored as bule, which represents an empty cell error. In the chart view, the number of empty cells will be shown.

The next section are the outputs of the tool, the order is the same as described as in the above steps:



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

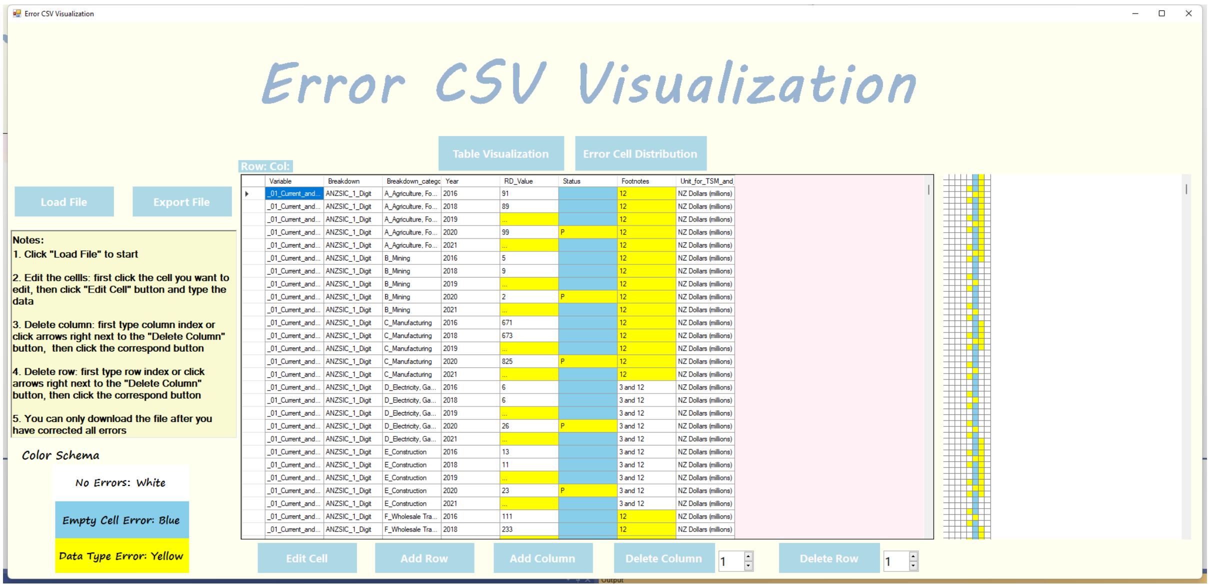
Color Schema

No Errors: White

Empty Cell Error: Blue

Data Type Error: Yellow

FIGURE 4.2.1. GUI



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Color Schema

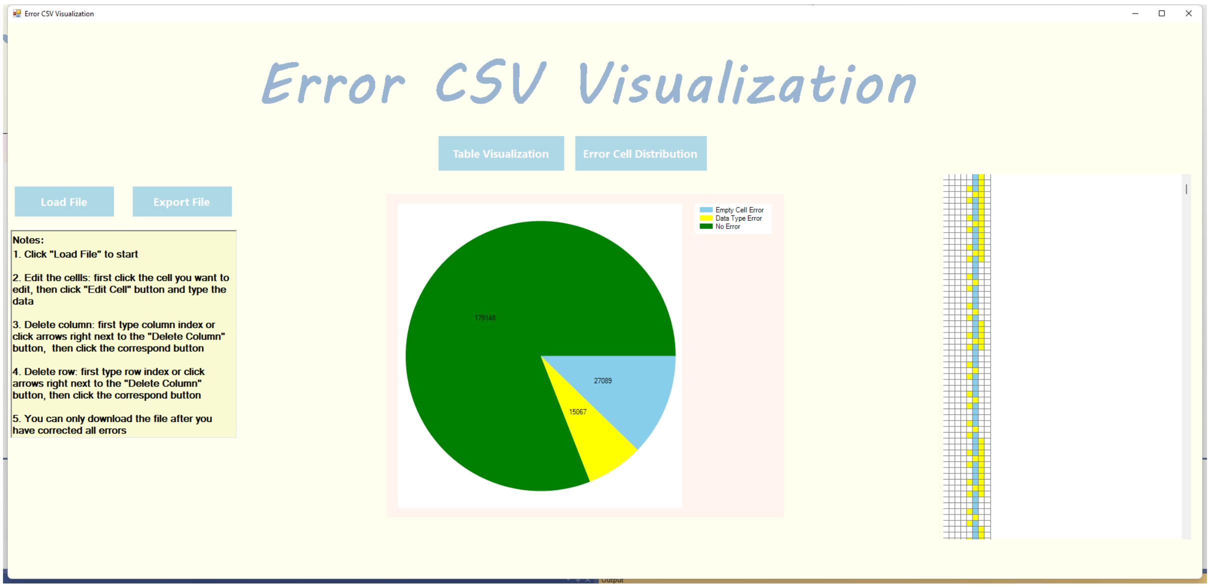
No Errors: White

Empty Cell Error: Blue

Data Type Error: Yellow

Variable	Breakdown	Breakdown_catego	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2018	89		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2019	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2020	99	P	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2021	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2016	5		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2018	9		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2019	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2020	2	P	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2021	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2016	671		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2018	673		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2019	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2021	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2016	13		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2018	11		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)

FIGURE 4.2.2. LOADING CSV AND SHOWING TABLE VISUALIZATION: RESEARCH-AND-DEVELOPMENT-SURVEY-2021-CSV



Notes:
1. Click "Load File" to start

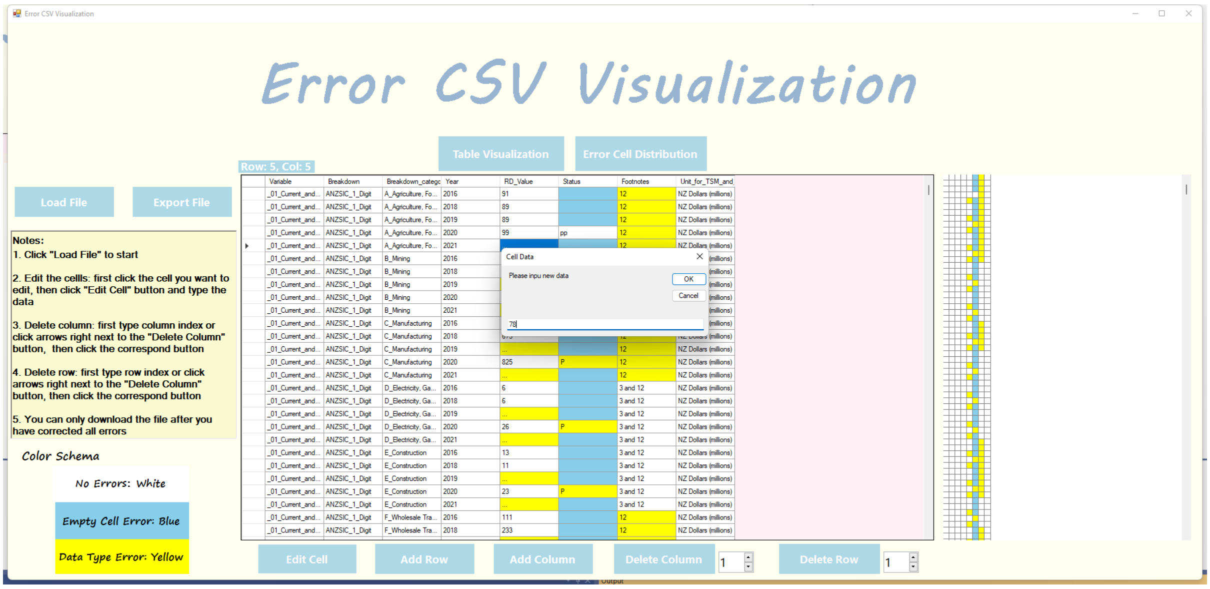
2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

FIGURE 4.2.3. DISPLAYING ERROR CELL DISTRIBUTION



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Row: 5, Col: 5

Variable	Breakdown	Breakdown_categc	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2018	89		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2019	89		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2020	99	pp	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2021			12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2016				(millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2018				(millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2019				(millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2020				(millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2021				(millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2016				(millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2018	675		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2019	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2021	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2016	13		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2018	11		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)

Cell Data

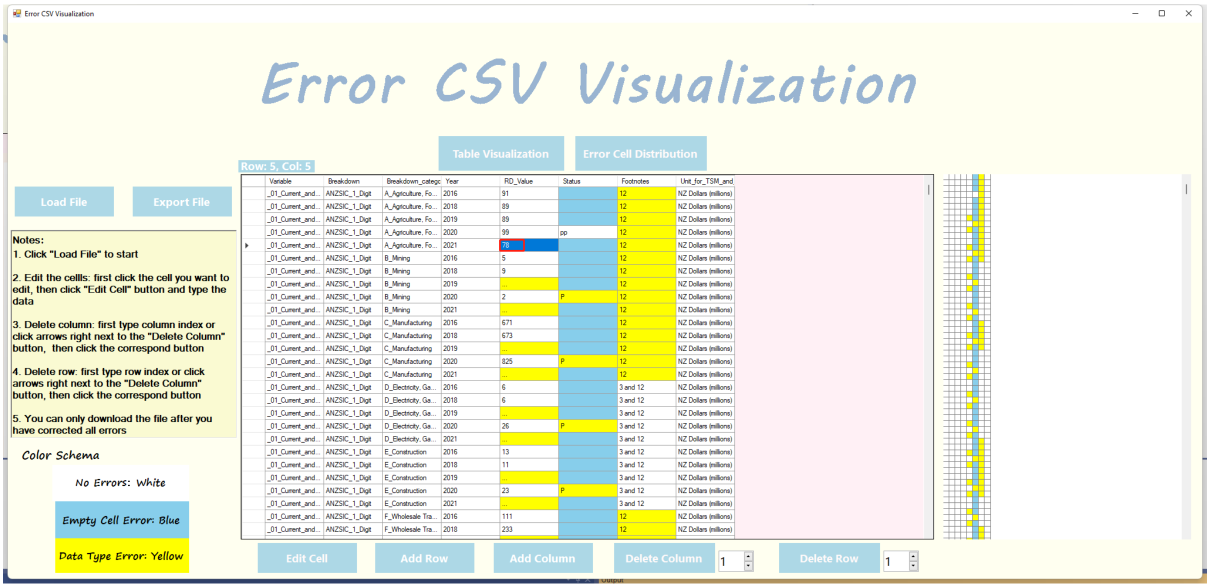
Please inpu new data

OK

Cancel

78

FIGURE 4.2.4.1. EDITING ERROR CELL, CHANGE VALUE TO 78



Notes:

1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Color Schema

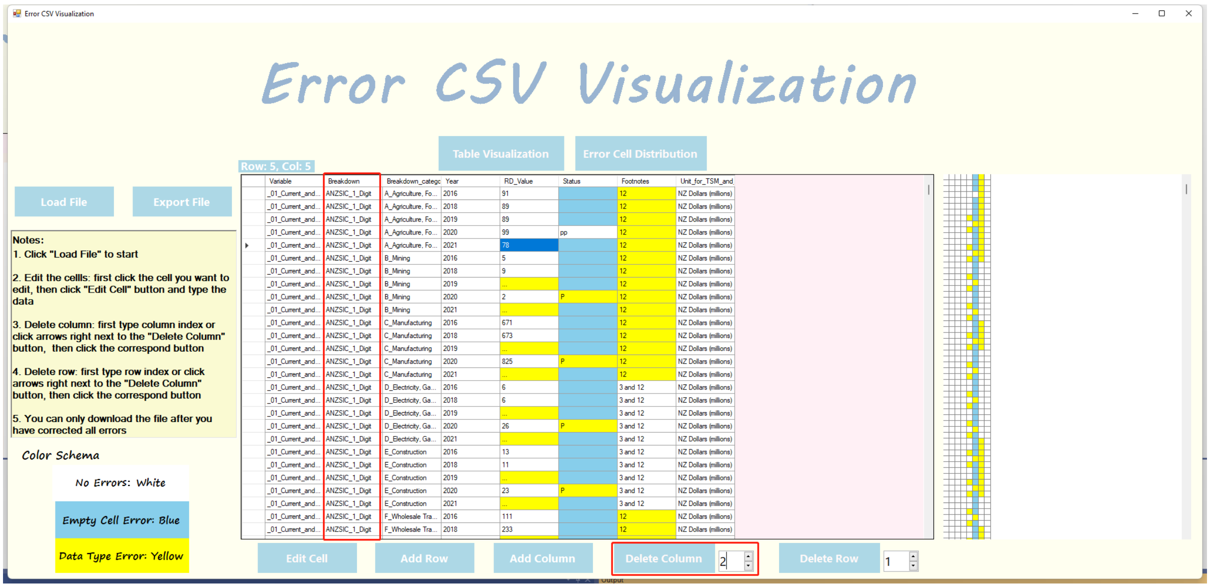
No Errors: White

Empty Cell Error: Blue

Data Type Error: Yellow

Variable	Breakdown	Breakdown_catego	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2018	89		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2019	89		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2020	99	pp	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2021	78		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2016	5		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2018	9		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2019	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2020	2	P	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2021	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2016	671		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2018	673		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2019	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2021	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2016	13		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2018	11		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)

FIGURE 4.2.4.2. EDITING ERROR CELL, SHOWING RESULT OF CHANGING VALUE TO 78



Notes:

1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Color Schema

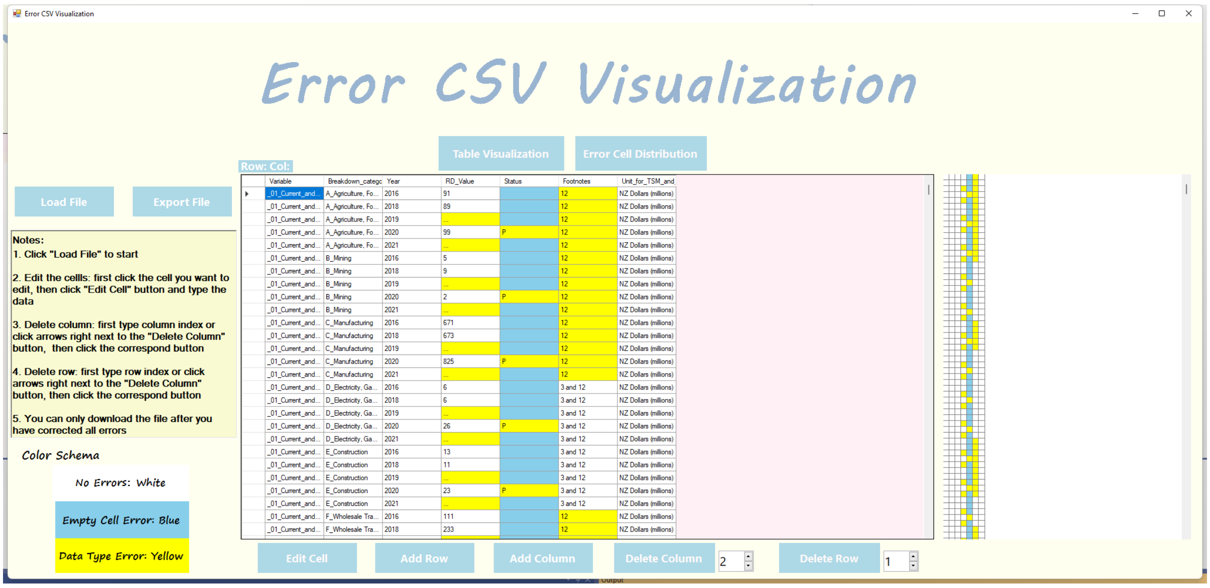
No Errors: White

Empty Cell Error: Blue

Data Type Error: Yellow

Variable	Breakdown	Breakdown_catego	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2018	89		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2019	89		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2020	99	pp	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	A_Agriculture, Fo...	2021	78		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2016	5		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2018	9		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2019	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2020	2	P	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	B_Mining	2021	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2016	671		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2018	673		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2019	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	C_Manufacturing	2021	...		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2016	13		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2018	11		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	E_Construction	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)
_01_Current_and...	ANZSIC_1_Digit	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)

FIGURE 4.2.5.1. DELETE COLUMN OF INDEX 2



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

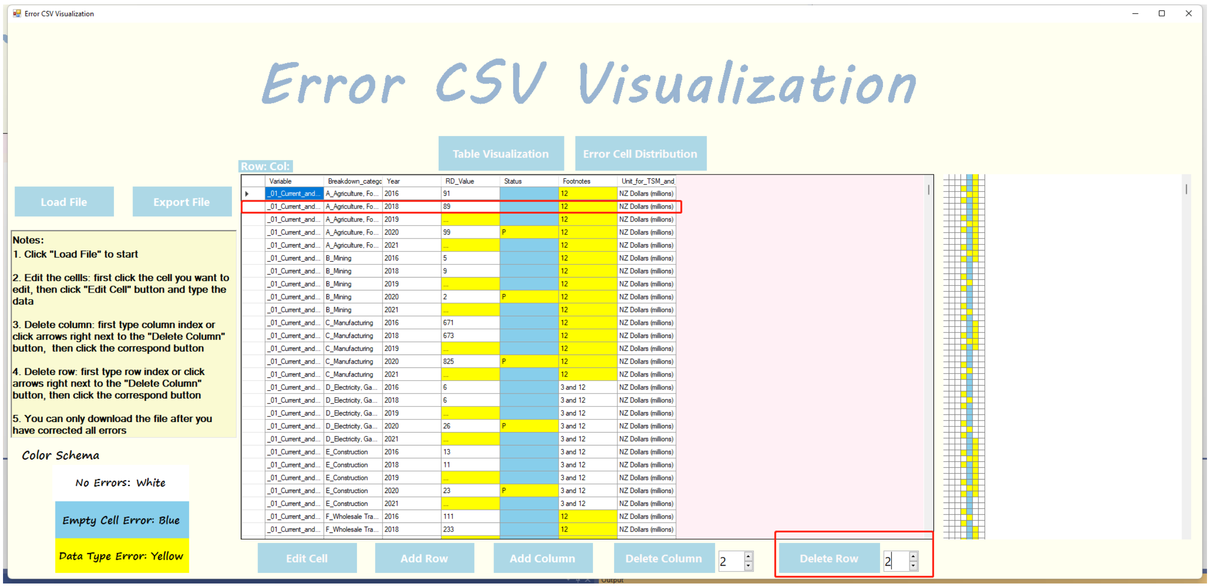
3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Variable	Breakdown_catego	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_01_Current_and...	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2018	89		12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2019	...		12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2020	99	P	12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2021	...		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2016	5		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2018	9		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2019	...		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2020	2	P	12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2021	...		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2016	671		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2018	673		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2019	...		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2021	...		12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2016	13		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2018	11		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)
_01_Current_and...	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)

FIGURE 4.2.5.2. DISPLAYING THE RESULT: DELETE COLUMN OF INDEX 2



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

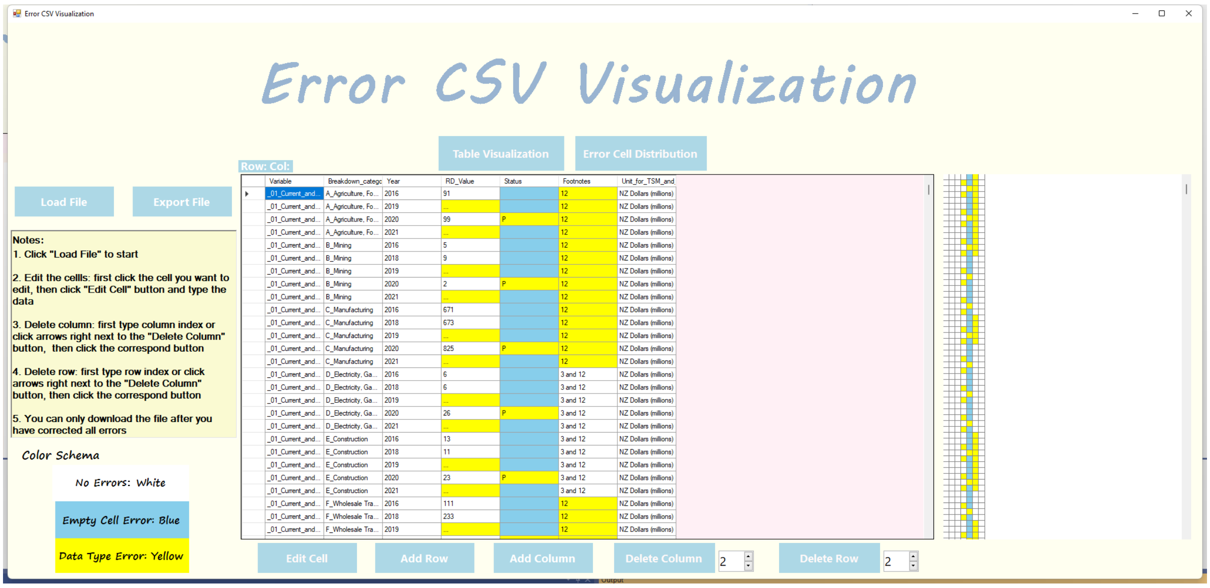
3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Variable	Breakdown_catego	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_01_Current_and...	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2018	89		12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2019	...		12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2020	99	P	12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2021	...		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2016	5		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2018	9		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2019	...		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2020	2	P	12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2021	...		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2016	671		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2018	673		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2019	...		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2021	...		12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2016	13		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2018	11		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)
_01_Current_and...	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)

FIGURE 4.2.6.1. DELETE ROW OF INDEX 2



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

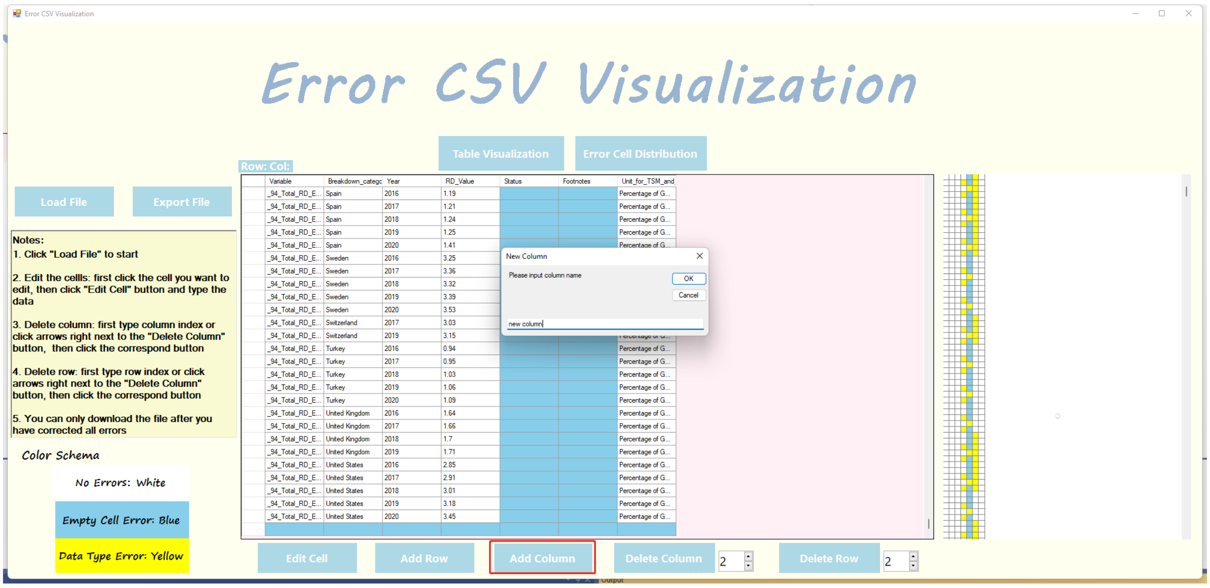
3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Variable	Breakdown_catego	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_01_Current_and...	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2019	...		12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2020	99	P	12	NZ Dollars (millions)
_01_Current_and...	A_Agriculture, Fo...	2021	...		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2016	5		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2018	9		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2019	...		12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2020	2	P	12	NZ Dollars (millions)
_01_Current_and...	B_Mining	2021	...		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2016	671		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2018	673		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2019	...		12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)
_01_Current_and...	C_Manufacturing	2021	...		12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2016	13		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2018	11		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2019	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)
_01_Current_and...	E_Construction	2021	...		3 and 12	NZ Dollars (millions)
_01_Current_and...	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)
_01_Current_and...	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)
_01_Current_and...	F_Wholesale Tra...	2019	...		12	NZ Dollars (millions)

FIGURE 4.2.6.2. DISPLAYING THE RESULT: DELETE ROW OF INDEX 2



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

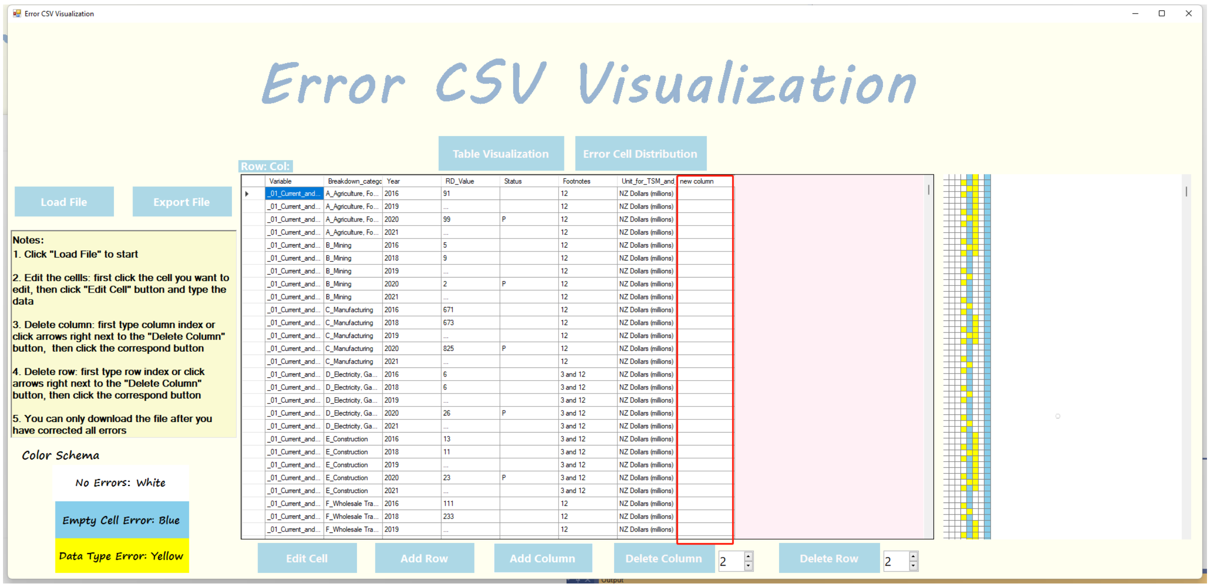
Variable	Breakdown_catego	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_94_Total_RD_E...	Spain	2016	1.19			Percentage of G...
_94_Total_RD_E...	Spain	2017	1.21			Percentage of G...
_94_Total_RD_E...	Spain	2018	1.24			Percentage of G...
_94_Total_RD_E...	Spain	2019	1.25			Percentage of G...
_94_Total_RD_E...	Spain	2020	1.41			Percentage of G...
_94_Total_RD_E...	Sweden	2016	3.25			
_94_Total_RD_E...	Sweden	2017	3.36			
_94_Total_RD_E...	Sweden	2018	3.32			
_94_Total_RD_E...	Sweden	2019	3.39			
_94_Total_RD_E...	Sweden	2020	3.53			
_94_Total_RD_E...	Switzerland	2017	3.03			
_94_Total_RD_E...	Switzerland	2019	3.15			Percentage of G...
_94_Total_RD_E...	Turkey	2016	0.94			Percentage of G...
_94_Total_RD_E...	Turkey	2017	0.95			Percentage of G...
_94_Total_RD_E...	Turkey	2018	1.03			Percentage of G...
_94_Total_RD_E...	Turkey	2019	1.06			Percentage of G...
_94_Total_RD_E...	Turkey	2020	1.09			Percentage of G...
_94_Total_RD_E...	United Kingdom	2016	1.64			Percentage of G...
_94_Total_RD_E...	United Kingdom	2017	1.66			Percentage of G...
_94_Total_RD_E...	United Kingdom	2018	1.7			Percentage of G...
_94_Total_RD_E...	United Kingdom	2019	1.71			Percentage of G...
_94_Total_RD_E...	United States	2016	2.85			Percentage of G...
_94_Total_RD_E...	United States	2017	2.91			Percentage of G...
_94_Total_RD_E...	United States	2018	3.01			Percentage of G...
_94_Total_RD_E...	United States	2019	3.18			Percentage of G...
_94_Total_RD_E...	United States	2020	3.45			Percentage of G...

New Column

Please input column name

new column

FIGURE 4.2.7. 1 ADD A COLUMN: NEW COLUMN



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

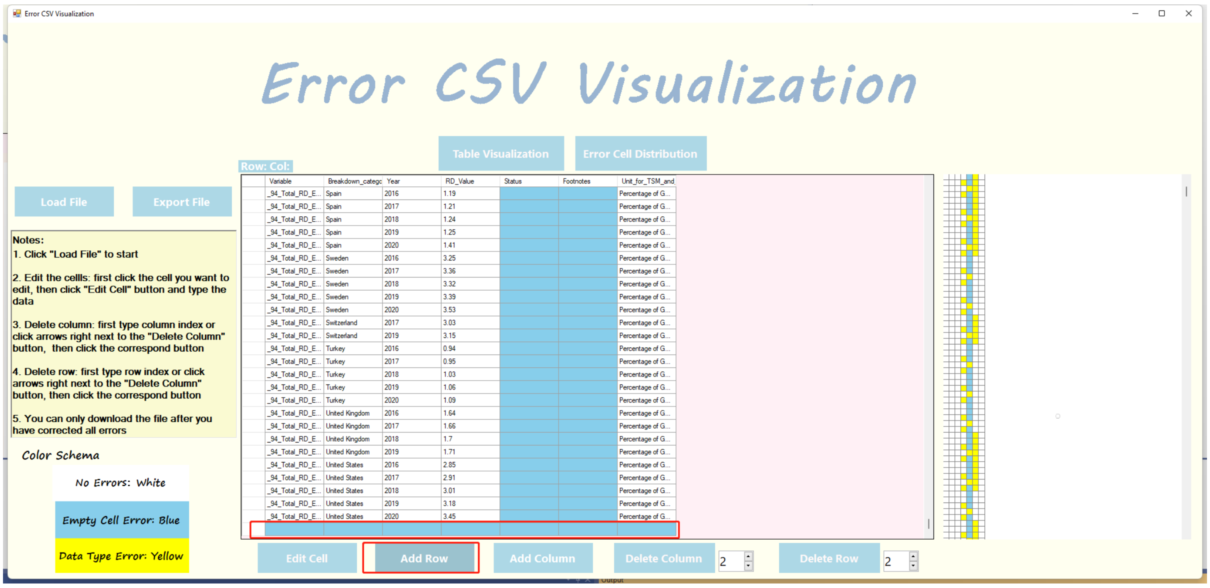
3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Variable	Breakdown_categc	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and	new column
_01_Current_and...	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)	
_01_Current_and...	A_Agriculture, Fo...	2019	...		12	NZ Dollars (millions)	
_01_Current_and...	A_Agriculture, Fo...	2020	99	P	12	NZ Dollars (millions)	
_01_Current_and...	A_Agriculture, Fo...	2021	...		12	NZ Dollars (millions)	
_01_Current_and...	B_Mining	2016	5		12	NZ Dollars (millions)	
_01_Current_and...	B_Mining	2018	9		12	NZ Dollars (millions)	
_01_Current_and...	B_Mining	2019	...		12	NZ Dollars (millions)	
_01_Current_and...	B_Mining	2020	2	P	12	NZ Dollars (millions)	
_01_Current_and...	B_Mining	2021	...		12	NZ Dollars (millions)	
_01_Current_and...	C_Manufacturing	2016	671		12	NZ Dollars (millions)	
_01_Current_and...	C_Manufacturing	2018	673		12	NZ Dollars (millions)	
_01_Current_and...	C_Manufacturing	2019	...		12	NZ Dollars (millions)	
_01_Current_and...	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)	
_01_Current_and...	C_Manufacturing	2021	...		12	NZ Dollars (millions)	
_01_Current_and...	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)	
_01_Current_and...	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)	
_01_Current_and...	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)	
_01_Current_and...	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)	
_01_Current_and...	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)	
_01_Current_and...	E_Construction	2016	13		3 and 12	NZ Dollars (millions)	
_01_Current_and...	E_Construction	2018	11		3 and 12	NZ Dollars (millions)	
_01_Current_and...	E_Construction	2019	...		3 and 12	NZ Dollars (millions)	
_01_Current_and...	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)	
_01_Current_and...	E_Construction	2021	...		3 and 12	NZ Dollars (millions)	
_01_Current_and...	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)	
_01_Current_and...	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)	
_01_Current_and...	F_Wholesale Tra...	2019	...		12	NZ Dollars (millions)	

FIGURE 4.2.7.2 DISPLAYING THE RESULT: ADD A COLUMN NAMED NEW COLUMN



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

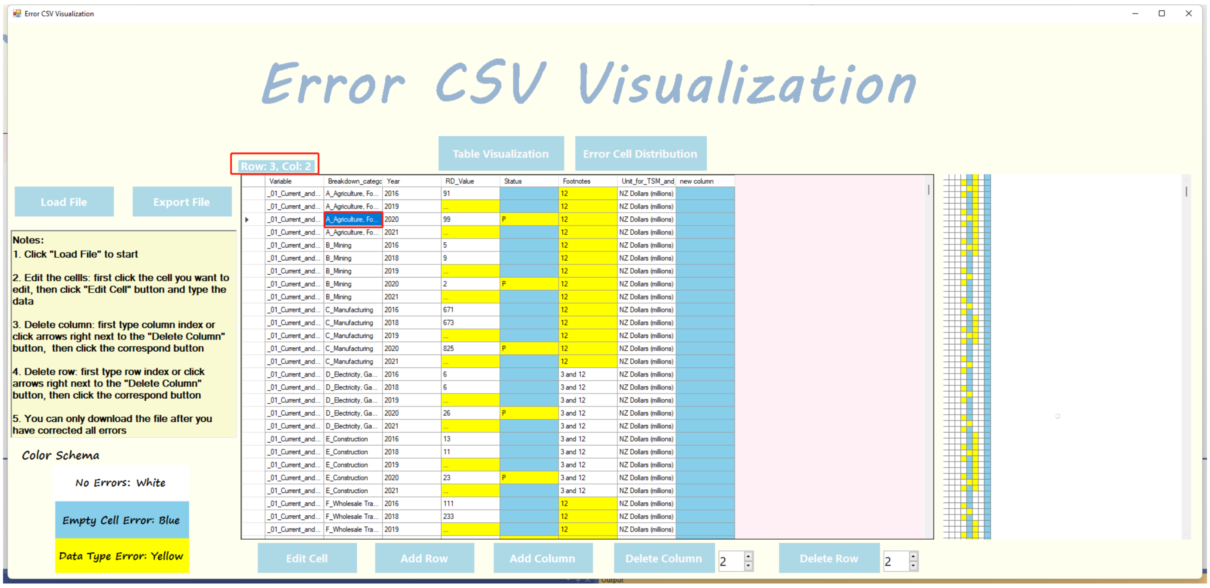
3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

Variable	Breakdown_categc	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and
_94_Total_RD_E...	Spain	2016	1.19			Percentage of G...
_94_Total_RD_E...	Spain	2017	1.21			Percentage of G...
_94_Total_RD_E...	Spain	2018	1.24			Percentage of G...
_94_Total_RD_E...	Spain	2019	1.25			Percentage of G...
_94_Total_RD_E...	Spain	2020	1.41			Percentage of G...
_94_Total_RD_E...	Sweden	2016	3.25			Percentage of G...
_94_Total_RD_E...	Sweden	2017	3.36			Percentage of G...
_94_Total_RD_E...	Sweden	2018	3.32			Percentage of G...
_94_Total_RD_E...	Sweden	2019	3.39			Percentage of G...
_94_Total_RD_E...	Sweden	2020	3.53			Percentage of G...
_94_Total_RD_E...	Switzerland	2017	3.03			Percentage of G...
_94_Total_RD_E...	Switzerland	2019	3.15			Percentage of G...
_94_Total_RD_E...	Turkey	2016	0.94			Percentage of G...
_94_Total_RD_E...	Turkey	2017	0.95			Percentage of G...
_94_Total_RD_E...	Turkey	2018	1.03			Percentage of G...
_94_Total_RD_E...	Turkey	2019	1.06			Percentage of G...
_94_Total_RD_E...	Turkey	2020	1.09			Percentage of G...
_94_Total_RD_E...	United Kingdom	2016	1.64			Percentage of G...
_94_Total_RD_E...	United Kingdom	2017	1.66			Percentage of G...
_94_Total_RD_E...	United Kingdom	2018	1.7			Percentage of G...
_94_Total_RD_E...	United Kingdom	2019	1.71			Percentage of G...
_94_Total_RD_E...	United States	2016	2.85			Percentage of G...
_94_Total_RD_E...	United States	2017	2.91			Percentage of G...
_94_Total_RD_E...	United States	2018	3.01			Percentage of G...
_94_Total_RD_E...	United States	2019	3.18			Percentage of G...
_94_Total_RD_E...	United States	2020	3.45			Percentage of G...

FIGURE 4.2.8 ADDING A NEW ROW



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

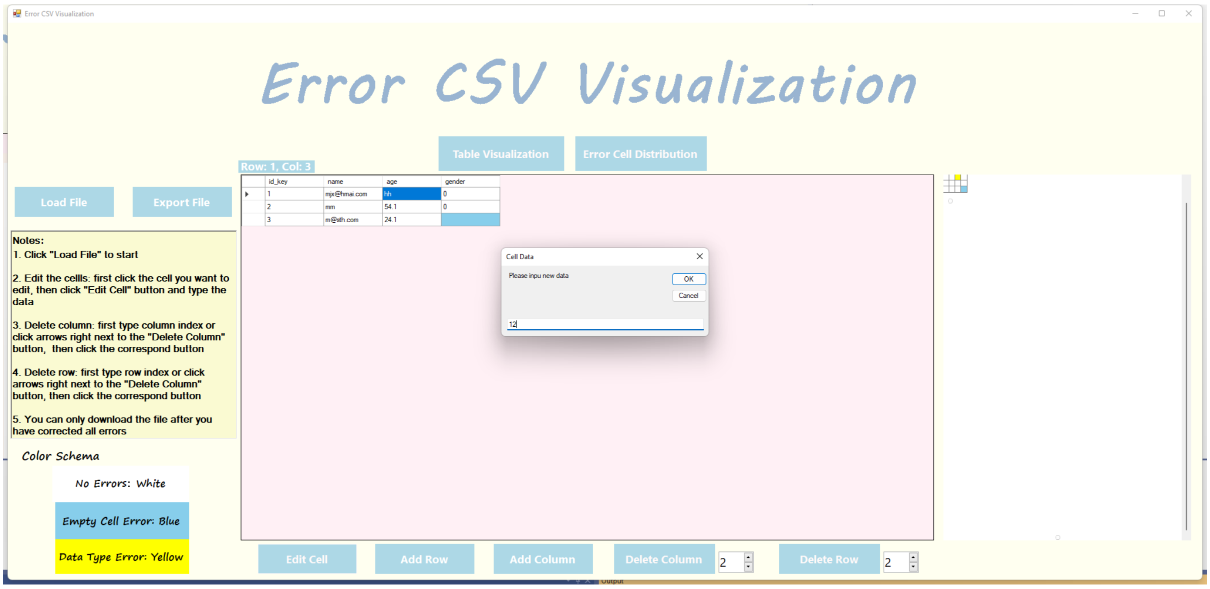
3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

5. You can only download the file after you have corrected all errors

	Variable	Breakdown_catego	Year	RD_Value	Status	Footnotes	Unit_for_TSM_and	new column
	_01_Current_and...	A_Agriculture, Fo...	2016	91		12	NZ Dollars (millions)	
	_01_Current_and...	A_Agriculture, Fo...	2019	...		12	NZ Dollars (millions)	
▸	_01_Current_and...	A_Agriculture, Fo...	2020	99	P	12	NZ Dollars (millions)	
	_01_Current_and...	A_Agriculture, Fo...	2021	...		12	NZ Dollars (millions)	
	_01_Current_and...	B_Mining	2016	5		12	NZ Dollars (millions)	
	_01_Current_and...	B_Mining	2018	9		12	NZ Dollars (millions)	
	_01_Current_and...	B_Mining	2019	...		12	NZ Dollars (millions)	
	_01_Current_and...	B_Mining	2020	2	P	12	NZ Dollars (millions)	
	_01_Current_and...	B_Mining	2021	...		12	NZ Dollars (millions)	
	_01_Current_and...	C_Manufacturing	2016	671		12	NZ Dollars (millions)	
	_01_Current_and...	C_Manufacturing	2018	673		12	NZ Dollars (millions)	
	_01_Current_and...	C_Manufacturing	2019	...		12	NZ Dollars (millions)	
	_01_Current_and...	C_Manufacturing	2020	825	P	12	NZ Dollars (millions)	
	_01_Current_and...	C_Manufacturing	2021	...		12	NZ Dollars (millions)	
	_01_Current_and...	D_Electricity, Ga...	2016	6		3 and 12	NZ Dollars (millions)	
	_01_Current_and...	D_Electricity, Ga...	2018	6		3 and 12	NZ Dollars (millions)	
	_01_Current_and...	D_Electricity, Ga...	2019	...		3 and 12	NZ Dollars (millions)	
	_01_Current_and...	D_Electricity, Ga...	2020	26	P	3 and 12	NZ Dollars (millions)	
	_01_Current_and...	D_Electricity, Ga...	2021	...		3 and 12	NZ Dollars (millions)	
	_01_Current_and...	E_Construction	2016	13		3 and 12	NZ Dollars (millions)	
	_01_Current_and...	E_Construction	2018	11		3 and 12	NZ Dollars (millions)	
	_01_Current_and...	E_Construction	2019	...		3 and 12	NZ Dollars (millions)	
	_01_Current_and...	E_Construction	2020	23	P	3 and 12	NZ Dollars (millions)	
	_01_Current_and...	E_Construction	2021	...		3 and 12	NZ Dollars (millions)	
	_01_Current_and...	F_Wholesale Tra...	2016	111		12	NZ Dollars (millions)	
	_01_Current_and...	F_Wholesale Tra...	2018	233		12	NZ Dollars (millions)	
	_01_Current_and...	F_Wholesale Tra...	2019	...		12	NZ Dollars (millions)	

FIGURE 4.2.9 DISPLAYING ROW NUMBER OF COLUMN NUMBER OF CLICKED CELL



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

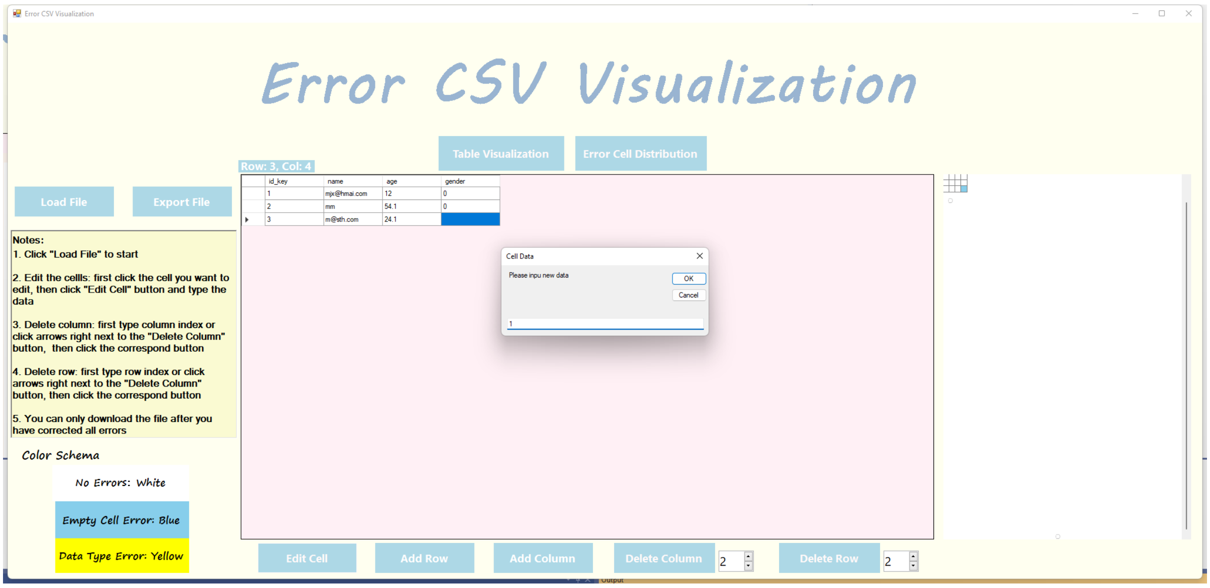
5. You can only download the file after you have corrected all errors

	id_key	name	age	gender
▸	1	mjx@hmai.com	hh	0
	2	mm	54.1	0
	3	m@sth.com	24.1	

Please inpu new data

12

FIGURE 4.2.10.1 EXPORT FILE: CHANGE THE AGE TO 12



Notes:
1. Click "Load File" to start

2. Edit the cellls: first click the cell you want to edit, then click "Edit Cell" button and type the data

3. Delete column: first type column index or click arrows right next to the "Delete Column" button, then click the correspond button

4. Delete row: first type row index or click arrows right next to the "Delete Column" button, then click the correspond button

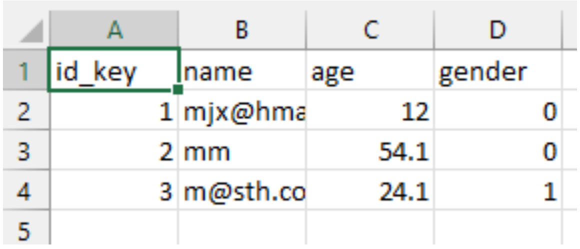
5. You can only download the file after you have corrected all errors

	id_key	name	age	gender
	1	mjx@hmai.com	12	0
	2	mm	54.1	0
▸	3	m@sth.com	24.1	

Please inpu new data

1

FIGURE 4.2.10.1 EXPORT FILE: CHANGE THE GENDER TO 1



	A	B	C	D
1	id_key	name	age	gender
2	1	mjx@hma	12	0
3	2	mm	54.1	0
4	3	m@sth.co	24.1	1
5				

FIGURE 4.2.10.3 EXPORT FILE: DISPLAYING THE RESULT

4.2.1 Runtime

For the purpose of testing performance, here are the running time for each file. The OS of testing computer is Windows 11 Pro, the processor is 11th Gen Intel(R) Core(TM) with i7 system and 16 logical processors, the RAM is 32.0 GB. The TABLE 4.3.1shows the run time of each file. “s” means seconds, “min” means minutes. From the table, we can see that for most of files only take few seconds to run, only the file with around 320,000 rows need to take 2 minutes.

	Number of rows	Number of columns	Runtime
File 1	966	15	5
File 2	47,175	10	13s
File 3	17,028	7	10s
File 4	13,489	6	20s
File 5	64,958	14	2 min
File 6	323,418	18	13s
File 7	27664	8	15s
File 8	3262	135	13s
File 9	3262	136	13s

TABLE 4.3.1 RUNTIME

5. Evaluation

This chapter will discuss the evaluation of the tool using a qualitative approach. It will begin by discussing the user experiment setup, followed by an analysis of the results.

5.1 User Experiment Set Up

The questions asked are the most important points for a questionnaire (Krosnick, 2018). To evaluate if the tool satisfies the goals, the questions on the survey are generated based on the goals in section 3.4, and most of them are Likert-scale. This is because Likert-scale is an efficient approach to give feedback on assessment (Barua, 2013). To improve the tool for future work, the suggestions from users will be asked as well. Results will be analyzed based on these goals. In this survey, 3 people are invited to take it, 3 of them are data science students and they have experienced with processing CSV files. The questionnaire is generated by Google Form, and the questions can be found in the Appendix B. The survey results can be found in Appendix C.

5.2 Result Analysis

According to the survey results, two out of three users are very happy with the GUI design, and one person thinks the design is excellent. They all agree that the application is informative, easy to use, and understandable, which meets goal 3. According to user feedback, the application fulfills the goal of helping users correct errors through visualizations, so goal 5 and 6 are achieved. Additionally, the code is open source, so users can modify it themselves, and it is easy to understand, thus fulfilling goal 2. It appears to be capable of loading large CSV files and displaying them, although it is slow to do so. In this regard, goal 4 has been met. When users use the tool, they have not encountered any bugs, according to user feedback.

There are three suggestions from the survey. As for the first suggestion, a person suggested that the tool should be able to visualize more types of errors. The second suggestion is, another people think that it should be able to visualize them in full-screen mode, the visualizations are too small. The last suggestion is that there is a slowdown in the display of large files, which creates a leggy experience for users. In the opinion of another person, the app should make use of the resources available.

Overall, all goals can be achieved, but the tool has some limitations, which can be improved in the future.

6. Conclusion

The purpose of the application is to assist users in visualizing errors in CSV files. In addition to being easy to use, the visualization tool should have an informative GUI. The users are able to interact with the table visualizations and correct the errors in the GUI or EXCEL. Once the errors have been corrected, they are able to download the corrected CSV file. It takes longer for the GUI to respond when editing the table for a large CSV file, e.g., 10k rows, which is a limitation of the application. It has been tested to handle files with up to 320,000 rows of data (the largest file tested was 320,000 rows). Despite the fact that the APP is only capable of visualizing two types of errors, which are empty cell errors and different variable errors in the same column, the paper selects them because empty cell errors can affect the data type detection for other columns, and for different variable errors in the same column, this is due to the fact that the error is quite common in CSV files.

The table visualization can provide information regarding errors within the cells. As a result, users will be able to identify which error they are experiencing. As a means of guiding users in repairing errors, each color represents an error. If users wish to correct errors in the EXCEL, they may click on the error cell and check the column number and row number. By examining the error cell distribution, they can identify which errors are most frequently occurring, which allows them to modify the algorithm used to generate the data. A pie chart represents the error cell distribution, which is an effective method for displaying the relationship between different components.

There are several limitations to the visualization tool. First, only two types of CSV errors can be detected, while as noted in chapter 2 related work, there are a number of different types of CSV errors. The survey's suggestion also suggests to provide more errors. When loading a large CSV file, the response time is prolonged. It responds very slowly when you edit the cell, add a column, add a row, delete a column, delete a row from a table, and scroll down data in scaled grid table. This is because every time when user edit the table, it needs to check the error and generating the grid table again. In addition, according to the survey feedback, the visualization should be in full-screen mode.

7. Future Work

Some improvements are still required for the tools. In this chapter, the future work will be discussed.

To begin with, it will be able to visualize more types of errors in the future, such as aggregated column or row errors and inconsistent data formatting errors. Depending on the results of a survey, which is a survey asking the respondents what errors you expect to visualize, the error selections can be made based on those results. People who are often involved in the processing of CSV files will receive the survey.

There is a possibility to reducing the size of columns and rows in the GUI when users visualize them. By the principle of the *Fisheye* method, only rows containing the error cells are chosen, but original row numbers will not be changed. So, when users work with large file, such as 10k rows, they take less time to scroll the file in the GUI since only error rows are displayed, they can check the selected row number and column number, and correct errors in the CSV files on EXCEL. For example, they are 10 rows, row indices 3 and 5 have errors, so only the row 3 and row 5 are selected to be display, but they are still identified as row 3 and row 5, not row 1 and row 2.

It will be possible for users to correct the errors more efficiently by using a machine learning algorithm that will correct them by simply clicking the button. The algorithm can also be interactively used by the users, for example, they can select which errors they want to see visually, and then click on a button to fix those errors.

Additionally, the loading time will be reduced when a large CSV file is loaded, such as a CSV file containing 32k rows. The results of the tests have shown that it takes around two minutes for the program to generate a table visualization for a file with 32k rows, and take longer for generating scaled table, which are not efficient. Therefore, it is anticipated that a new approach will be taken to solve the issue of slow loading times in the future.

8. References

Abba, A. H., & Hassan, M. (2018). Design and Implementation of a CSV Validation System. *In Proceedings of the 3rd International Conference on Applications in Information Technology*, pp. 111-116.

Carvalho, P., Hitzelberger, P., Bouali, F., & Venturini, G. (2015). A Visual Technique to Assess the Quality of Datasets-Understanding the Structure and Detecting Errors and Missing Values in Open Data CSV Files. *In International Conference on Data Management Technologies and Applications*, *2*, pp. 134-141.

Cota, M. P., Rodríguez, M. D., González-Castro, M. R., & Gonçalves, R. M. M. (2017). Massive data visualization analysis of current visualization techniques and main challenges for the future. *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*, pp. 1-6. 10.23919/CISTI.2017.7975704

Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). Wrangler: Interactive Visual Specification of Data Transformation Scripts. *In Proceedings of the sigchi conference on human factors in computing systems*, pp. 3363-3372. 10.1145/1978942.1979444

Lahar, S. (2020, November 5). *6 Common CSV Import Errors and How to Fix Them*. Flatfile. Retrieved June 5, 2022, from https://flatfile.com/blog/top-6-csv-import-errors-and-how-to-fix-them/

Rao, R., & Card, S. K. (1994, April). The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 318-322.

Sadiku, M. N.O., Shadare, A. E., Musa, S. M., & Akujuobi, C. M. (2016). Data visualization. *International Journal of Engineering Research And Advanced Technology (IJERAT)*, *02*(12), pp. 11-16.

Khan, M., & Khan, S. S. (2011). Data and information visualization methods, and interactive mechanisms: A survey. *International Journal of Computer Applications*, *34*(1), pp. 1-14.

Ge, C., Li, Y., Eilebrecht, E., Chandramouli, B., & Kossmann, D. (2019, June). Speculative distributed CSV data parsing for big data analytics. In *Proceedings of the 2019 International Conference on Management of Data* (pp. 883-899).

Döhmen, T. (2016). *Multi-Hypothesis Parsing of Tabular Data in Comma-Separated Values (CSV) Files* (Doctoral dissertation, Master's thesis, Vrije Universiteit Amsterdam, www.cwi.nl/boncz/msc/2016-Doehmen. pdf).

Mitlöhner, J., Neumaier, S., Umbrich, J., & Polleres, A. (2016, August). Characteristics of open data CSV files. In *2016 2nd International Conference on Open and Big Data (OBD)* (pp. 72-79). IEEE.

Galitz, W. O. (2007). *The essential guide to user interface design: an introduction to GUI design principles and techniques*. John Wiley & Sons.

Close, J., (2009) *CsvHelper* https://joshclose.github.io/CsvHelper/

Kiyono, S., Suzuki, J., Mizumoto, T., & Inui, K. (2020). Massive exploration of pseudo data for grammatical error correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2134-2145.

Hoffman, P. E. (2000). *Table visualizations: a formal model and its applications*. University of Massachusetts Lowell.

North, C., Conklin, N., Indukuri, K., & Saini, V. (2002). Visualization schemas and a web-based architecture for custom multiple-view visualization of multiple-table databases. Information Visualization, 1(3-4), 211-228.

Krosnick, J. A. (2018). Questionnaire design. In The Palgrave handbook of survey research (pp. 439-455). Palgrave Macmillan, Cham.

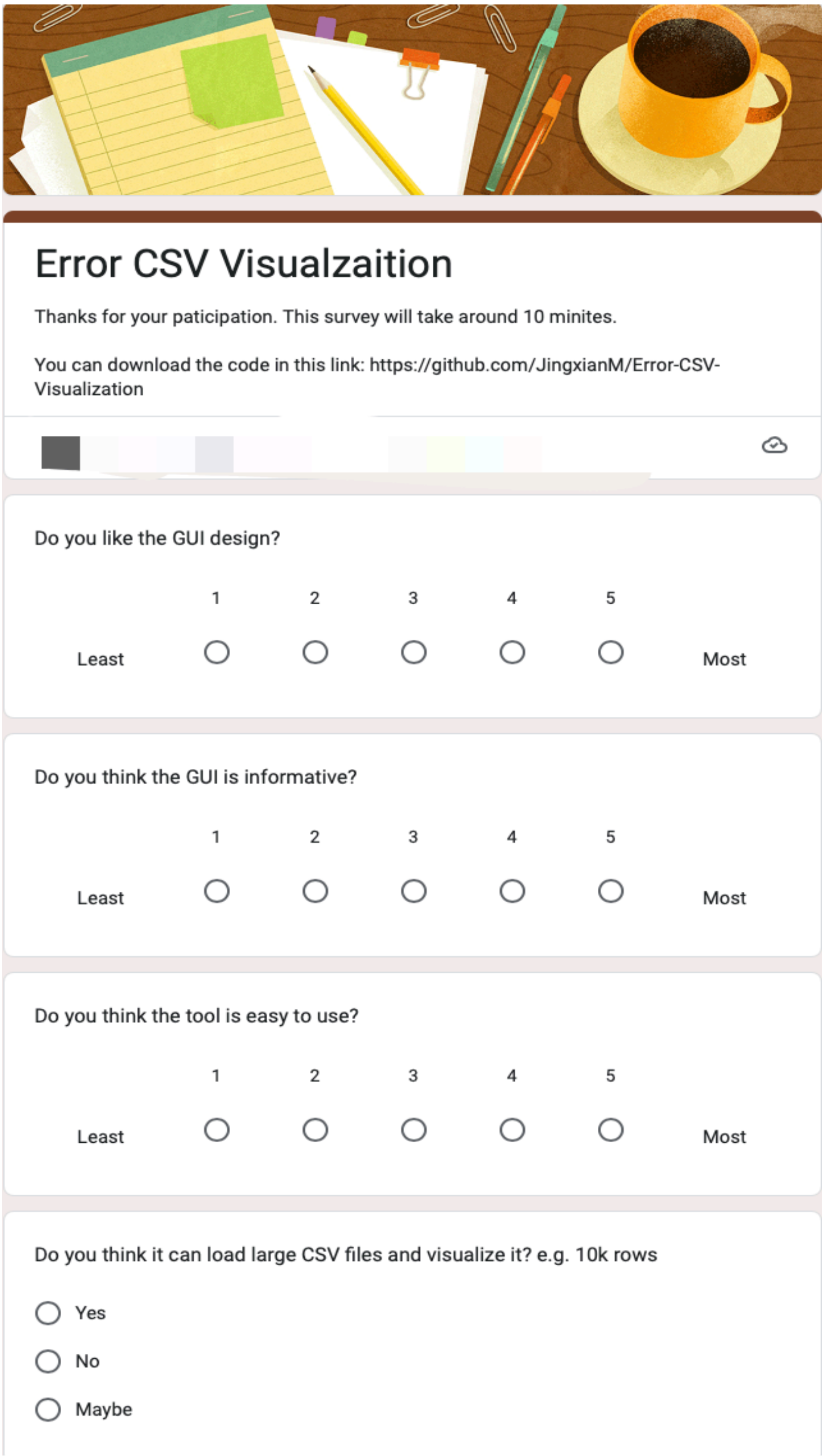
Barua, A. (2013). Methods for decision-making in survey questionnaires based on Likert scale. Journal of Asian Scientific Research, 3(1), 35-38.

9. Appendix

A. Coding

https://github.com/JingxianM/Error-CSV-Visualization

B. User Experiment Questionnaire



Error CSV Visualzaition

Thanks for your paticipation. This survey will take around 10 minites.

You can download the code in this link: https://github.com/JingxianM/Error-CSV-Visualization

Do you like the GUI design?

	1	2	3	4	5	
Least	○	○	○	○	○	Most

Do you think the GUI is informative?

	1	2	3	4	5	
Least	○	○	○	○	○	Most

Do you think the tool is easy to use?

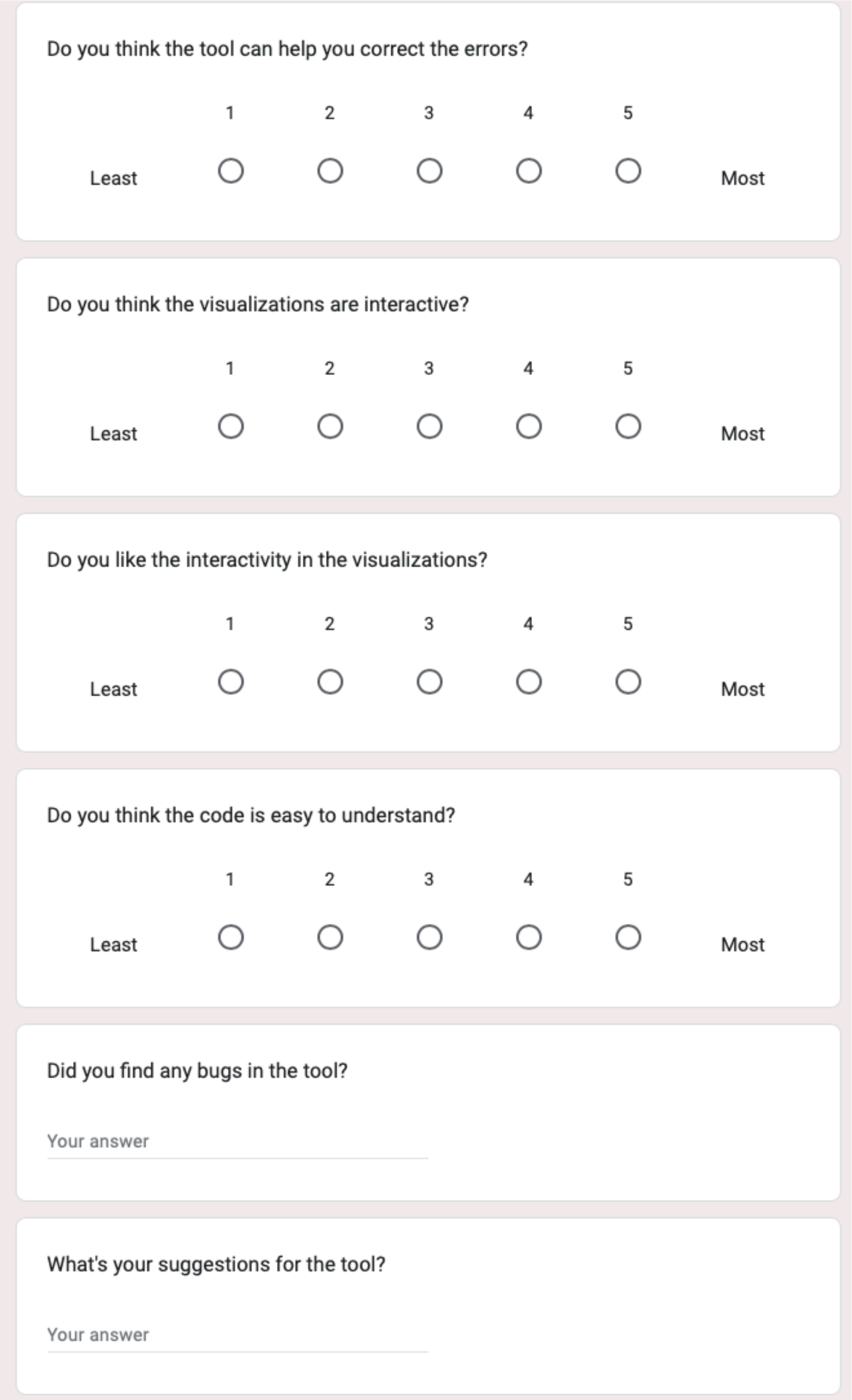
	1	2	3	4	5	
Least	○	○	○	○	○	Most

Do you think it can load large CSV files and visualize it? e.g. 10k rows

○ Yes

○ No

○ Maybe



Do you think the tool can help you correct the errors?

	1	2	3	4	5	
Least	○	○	○	○	○	Most

Do you think the visualizations are interactive?

	1	2	3	4	5	
Least	○	○	○	○	○	Most

Do you like the interactivity in the visualizations?

	1	2	3	4	5	
Least	○	○	○	○	○	Most

Do you think the code is easy to understand?

	1	2	3	4	5	
Least	○	○	○	○	○	Most

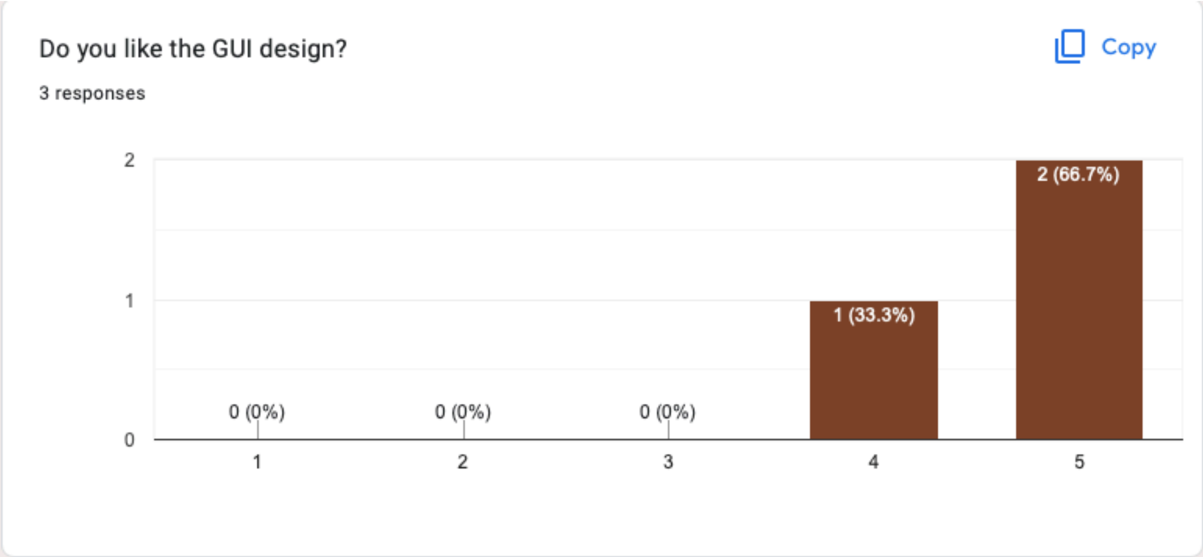
Did you find any bugs in the tool?

Your answer

What's your suggestions for the tool?

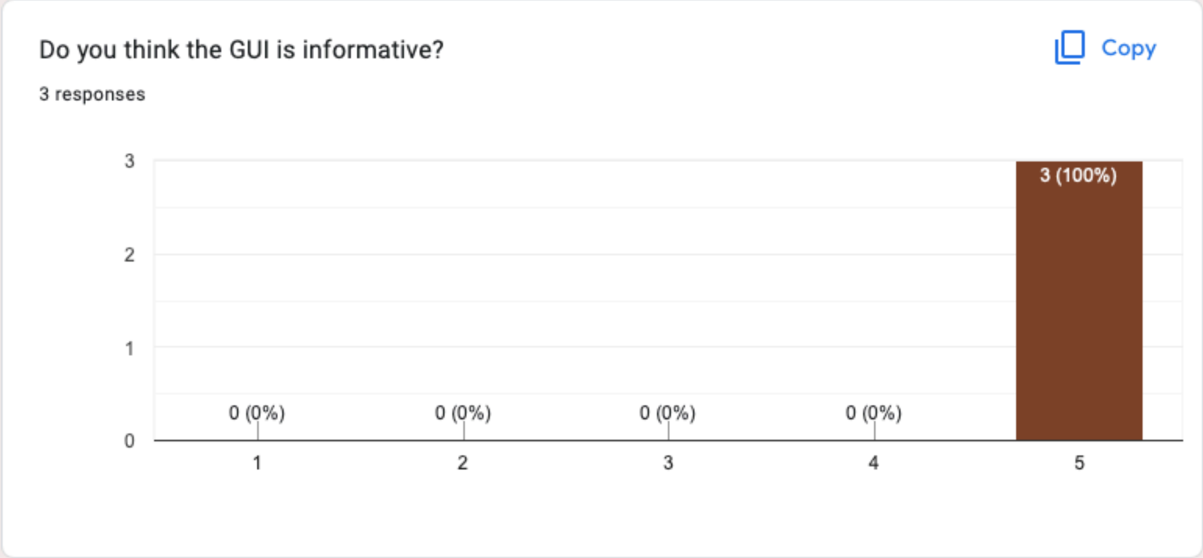
Your answer

D. User Experiment Results



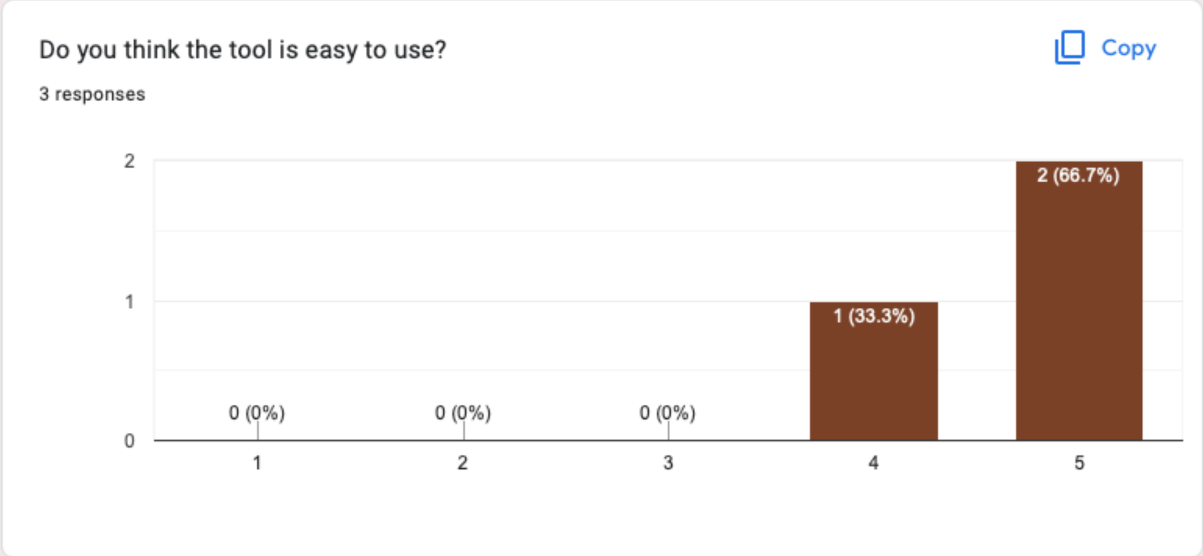
Do you like the GUI design?

3 responses



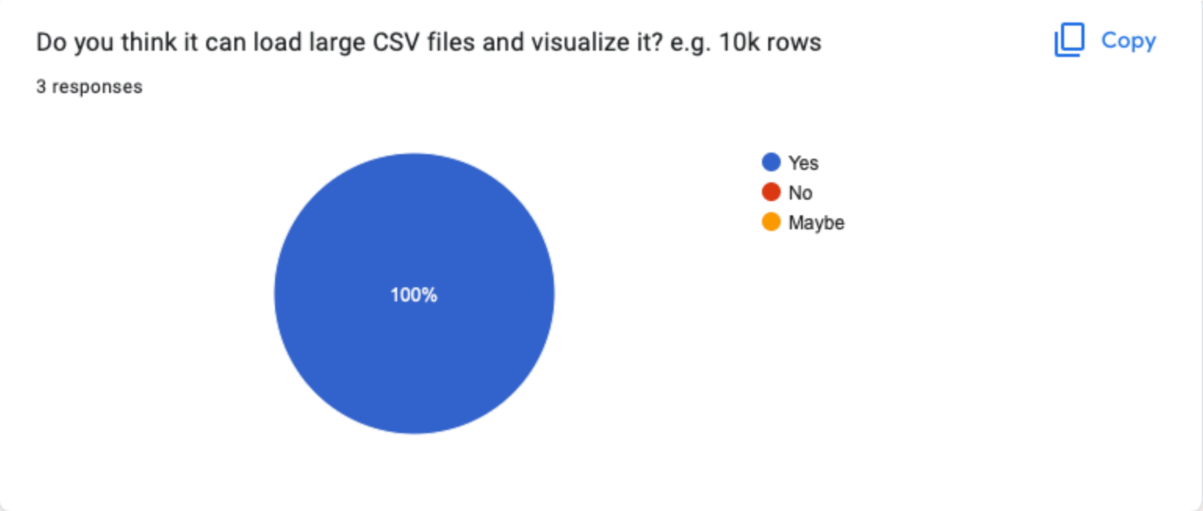
Do you think the GUI is informative?

3 responses



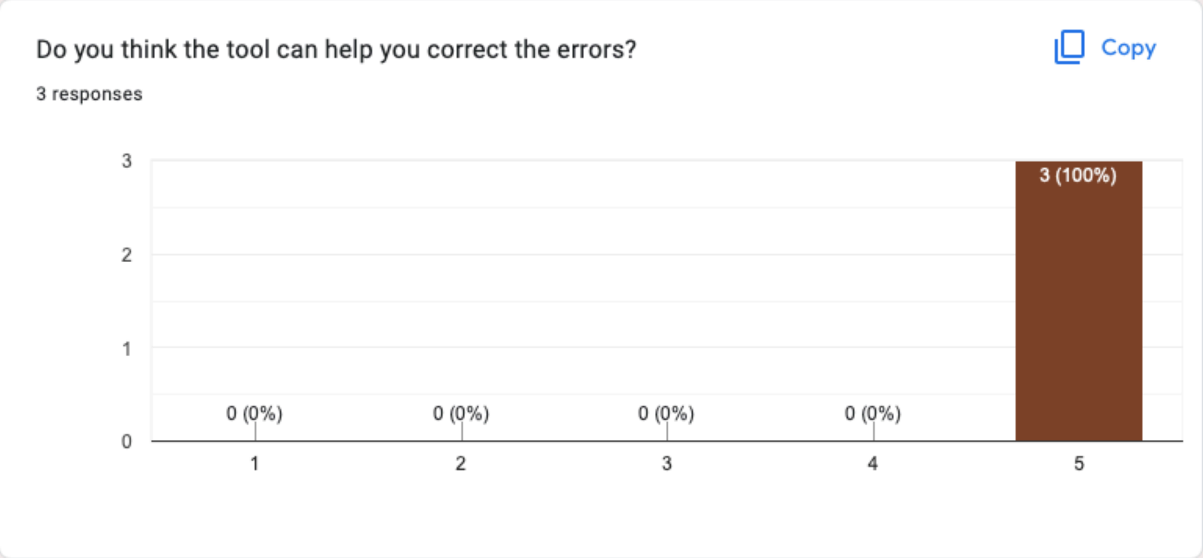
Do you think the tool is easy to use?

3 responses



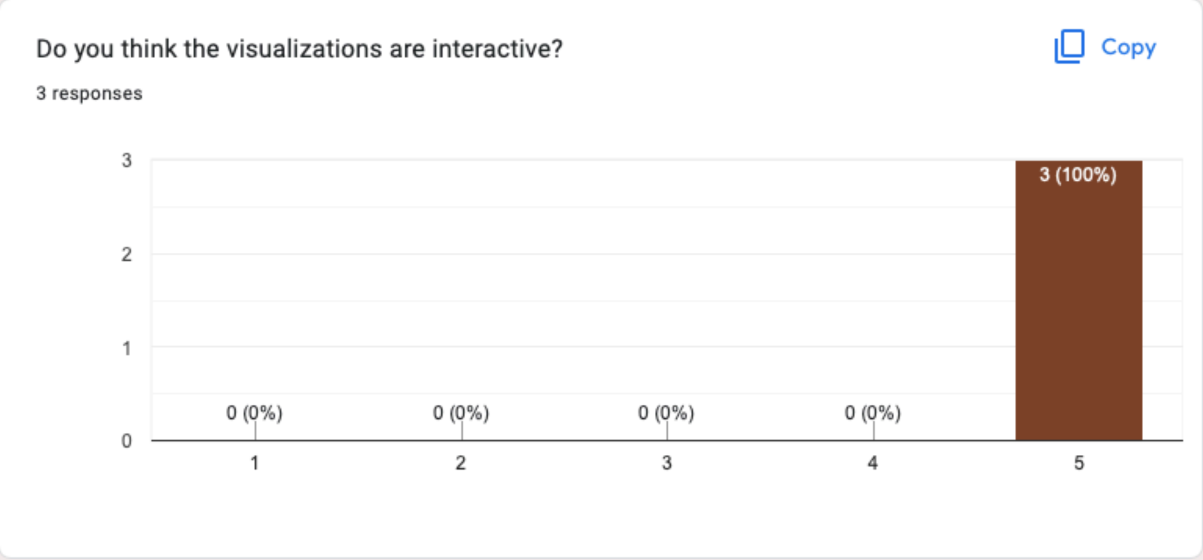
Do you think it can load large CSV files and visualize it? e.g. 10k rows

3 responses



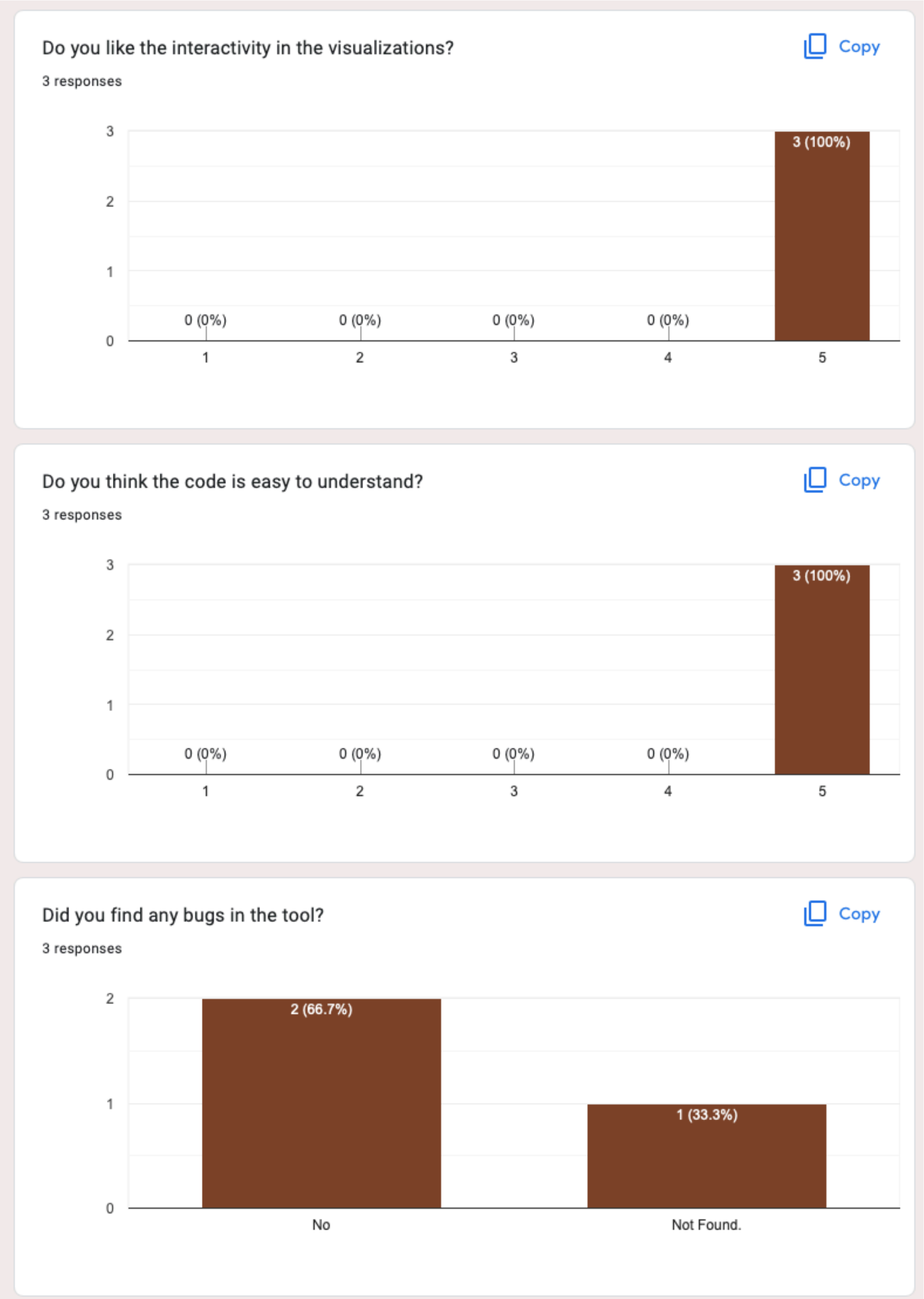
Do you think the tool can help you correct the errors?

3 responses



Do you think the visualizations are interactive?

3 responses



Do you like the interactivity in the visualizations?

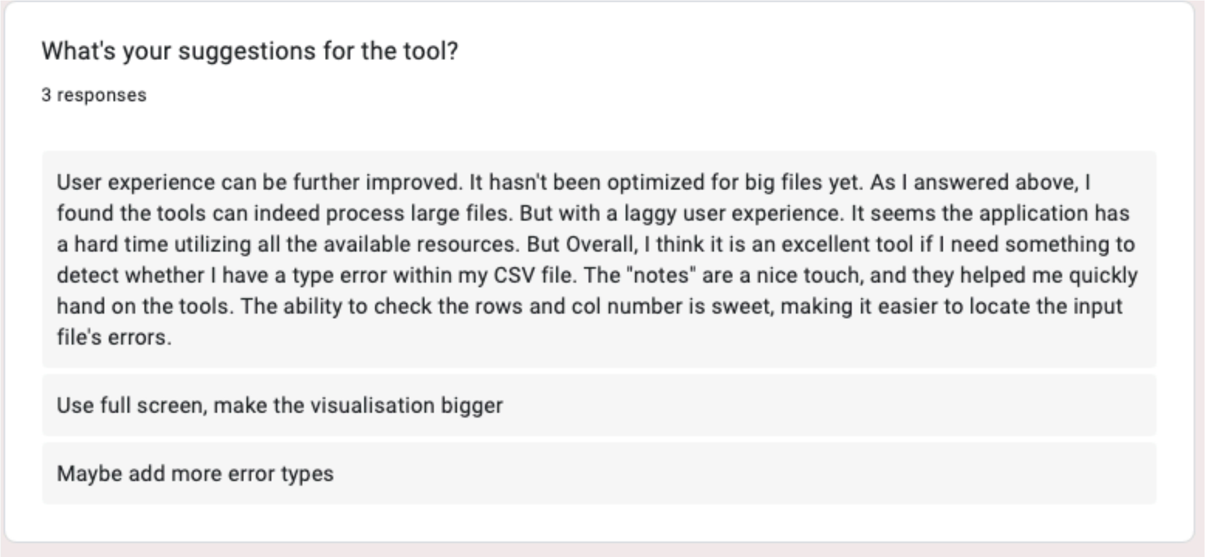
3 responses

Do you think the code is easy to understand?

3 responses

Did you find any bugs in the tool?

3 responses



What's your suggestions for the tool?

3 responses

User experience can be further improved. It hasn't been optimized for big files yet. As I answered above, I found the tools can indeed process large files. But with a laggy user experience. It seems the application has a hard time utilizing all the available resources. But Overall, I think it is an excellent tool if I need something to detect whether I have a type error within my CSV file. The "notes" are a nice touch, and they helped me quickly hand on the tools. The ability to check the rows and col number is sweet, making it easier to locate the input file's errors.

Use full screen, make the visualisation bigger

Maybe add more error types