

Data Analysis and Visualisation



Florin - Cristi Movilianu – 2112865

Data Handling & Business Intelligence – LCBB5000

Lecturer – Gasso Mwaluseke

Table of contents

1. INTRODUCTION	3
2. DATA ANALYSIS OF HOTEL RESERVATION	3
2.1 EXCEL	3
2.2 DESCRIPTION OF HOTEL RESERVATION DATASET	5
2.3 AVERAGE PRICE PER ROOM.....	6
2.4 SALES CHANNELS	7
3. DATA ANALYSIS OF THE STUDENT SURVEY	8
3.1 GENDER ANALYSIS	8
3.2 LUNCH ANALYSIS	9
3.3 MEAN AND MEDIAN OF STUDENTS OVER LUNCH QUALIFICATION	9
3.4 MEAN AND MEDIAN DIFFERENCES IN STUDENT'S TEST SCORES	10
3.5 CLUSTERING	11
3.6 DATA MINING	14
3.7 TEXT MINING	14
3.8 ADVANTAGES AND DISADVANTAGES OF USING SPSS AND EXCEL.....	15
4. CONCLUSION	16
REFERENCES.....	17

1. Introduction

Because of its accessibility, adaptability, and user-friendly interface, Microsoft Excel has been used extensively in the field of data analysis for a considerable amount of time. As a data analyst, the aim of this project is to evaluate how well Excel can be used for the pre-processing, analysis, and visualisation of data. In addition, I will analyse the data using the capabilities that are available in Excel to build useful charts and graphs that can be used to graphically portray the findings.

Furthermore, I will discuss the school's recent interest in the SPSS software for data mining as an alternative to Excel for data analysis, and the primary focus of our discussion will be on examining the relationships between students' eligibility for reduced or free lunches, gender, and their pre-test and post-test scores. The findings will be presented using graphical representations of data that have significant meaning.

2. Data Analysis of Hotel Reservation

2.1 Excel

Microsoft Excel is a spreadsheet programme that can be used to organise numerical and tabular data using various predefined formulas and functions. Financial analysis in Excel is widely utilised by organisations of all sizes throughout the world. The financial industry frequently employs Excel for data management and analysis. It is utilised in every department of businesses, both big and small. Excel's primary function is to facilitate the organisation of virtually anything that requires data entry, accounting, visual analysis, data management, or time management (CFI, 2022).

Data Pre-processing

Data pre-processing consists of preparing and cleaning data before conducting analysis. Excel offers several advantages for data pre-processing tasks:

- Familiarity: Excel's widespread usage and user-friendly interface make it accessible to a wide range of users (Poutintsev and Aditya, 2020).
- Simplicity: Worksheets in Microsoft Excel follow a standard tabular spreadsheet layout, with rows and columns. When information is well-structured and straightforward, user engagement improves (Jacob, 2021).
- Collaboration: Multiple users can enter data at once in Excel (Poutintsev and Aditya, 2020).

However, Excel has certain limitations for data pre-processing:

➤ **Inefficiency:** New information must be added to the Excel spreadsheet by hand if a user-friendly form option is not available. This might be a very laborious process if there is a lot of data to enter (Jacob, 2021).

➤ **Lack of relations:** Excel is using a basic spreadsheet format to store and organise data, but it does not allow for straightforward relational connections between individual sheets (Jacob, 2021).

➤ **Not secured:** Unfortunately, there is no built-in encryption in Excel to keep private information safe. Spreadsheets are naturally insecure, so even if you protect them with a password, hackers will still be able to access your data (Caspio, 2020)

➤ **Prone to errors:** Having many persons enter data into the same file increases the likelihood of mistakes being made. These errors are not only costly, but they are also difficult to detect later in the process (Poutintsev and Aditya, 2020).

➤ **Limited data entry:** Worksheets can have a maximum of 1,048,576 rows and 16,384 columns (Microsoft, 2023).

Data Analysis:

Excel provides basic analytical capabilities that allow analysts to perform calculations, statistical analysis, and exploratory data analysis. Some advantages of using Excel for data analysis include:

➤ **Versatility:** Cleaning, transforming, and manipulating data is made easier by Excel's in-built functions and formulas (JavaTpoint, 2021). With these tools, analysts may execute operations like eliminating duplicates, standardising data, and combining data sets.

➤ **Formula-Based Calculations:** Excel facilitates instantaneous, calculator-free computation of big quantities. Numerous instantaneous calculations of difficult mathematical equations are possible with the use of simple formulas (Jacob, 2021). Analysts can utilize formulas for basic statistical measures, such as mean, median, and standard deviation.

However, Excel has limitations when it comes to advanced data analysis:

➤ **Limited Statistical Capabilities:** Excel's built-in statistical capabilities are not nearly as sophisticated as those found in specialised statistical software such as SPSS; nonetheless, due to Excel's user-friendly interface, the programme is still commonly utilised in the research community (Gray et al., 2022).

Data Visualization:

Excel offers visualization tools that allow analysts to create basic charts, graphs, and dashboards. The advantages of using Excel for data visualization include:

➤ **Charting Options:** Producing charts from pre-existing data sets is a complex task, but Excel simplifies this operation. Existing, up-to-date data in a spreadsheet may be quickly and

easily converted into a chart by using the chart function (Smith, 2019). Analysts can choose from bar charts, line graphs, scatter plots, and many others.

➤ **Conditional Formatting:** Users can benefit from conditional formatting because it facilitates the identification of patterns and outliers in their data, facilitates the rapid identification of relevant information, gives a visual representation of their data, and facilitates the communication of their results to others (Aydin, 2023).

➤ **Tables:** Data sets might be overwhelming, but tables, pivot tables, and other visualisation methods paired with calculations can help the user identify the facts they need quickly (Poutintsev and Aditya, 2020).

However, Excel has limitations in terms of advanced data visualization:

➤ **Design Limitations:** Excel's visualizations may lack advanced design elements and may not be as visually appealing or interactive compared to Power BI.

➤ **Limited interactive capabilities:** While Excel does allow for the creation of dashboards, they are static, provide just an overview of patterns, and act more as a final product than an introduction for further investigation (Sigma, 2020). Interactive features commonly found in dedicated data visualization tools, such as filtering, zooming, or interactive tooltips, may be lacking in Excel.

2.2 Description of Hotel Reservation Dataset

A dataset, as defined by Ella Simon (2023), is a compilation of records related to a topic, subject, or industry that may include numerical data, textual information, visuals, video recordings and sounds.

The Hotel Reservation dataset contains records for every reservation made between 2017 and 2022. Each reservation has its records, which include the number of guests, the part of the week they needed, the meal plan they had, the vehicle parking needs, the room they requested, how they made the reservation and the year, the average price per room, and many other details. In total, 18 details are kept for each reservation.

Using the Pivot Table function, we can generate a table that summarises the dataset and can be filtered to display only the information of interest and also, generate a graphic representation of the data.

Sum of Average price	
Arrival_year	Total
2017	£ 549,304
2018	£ 2,984,722
2019	£ 18,015
2020	£ 20,857
2021	£ 19,437
2022	£ 159,354
Grand Total	£ 3,751,689

Table 1 *Turnover per year from 2017 to 2022*

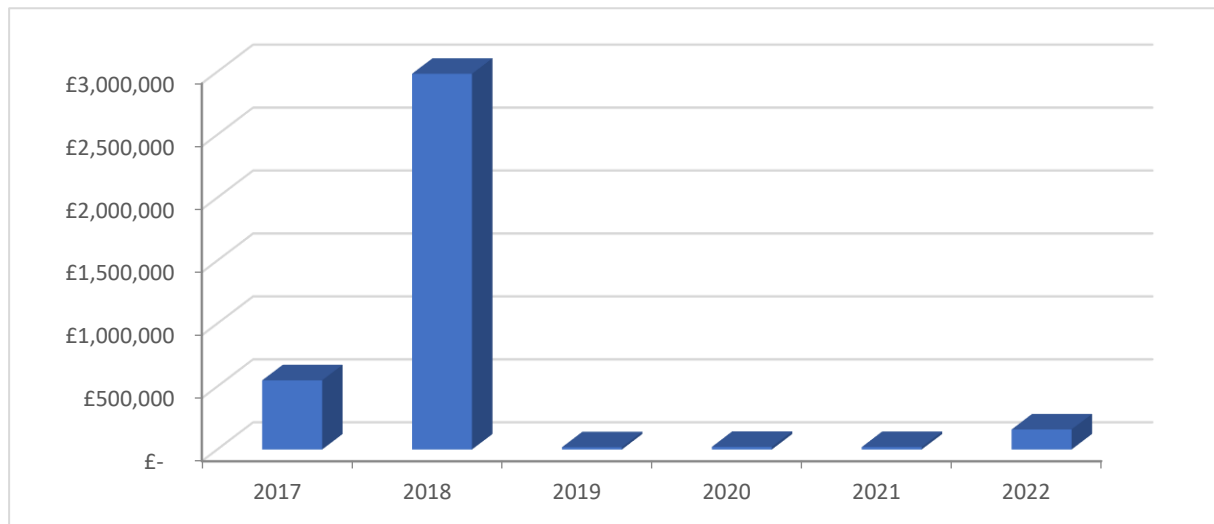


Figure 1 *Turnover per year from 2017 to 2022*

Based on the facts collected from the pivot table, we can observe that the best year for this hotel was 2018 with £2,984,722 in revenue, however, the revenue dropped significantly as a result of the government's Covid-19 limits and the drop in demand. As the Covid-19 restrictions were eased, we are looking at a recovery in turnover which started in 2022 (ONS, 2021).

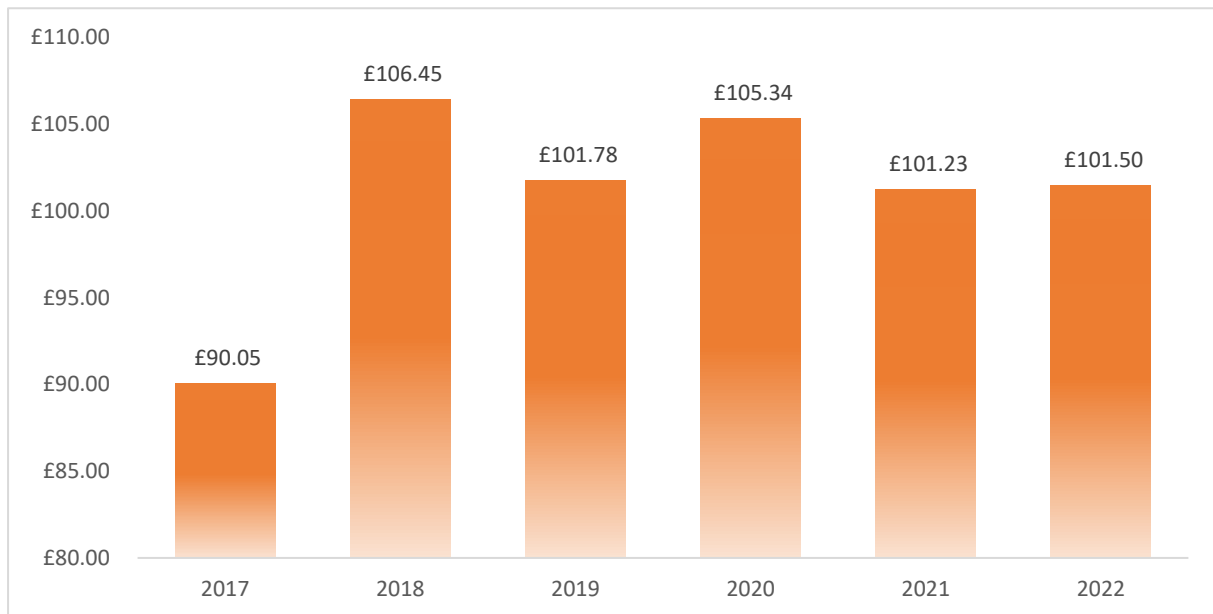
2.3 Average price per room

We use one of Excel's many functions called AVERAGEIF to discover, for each year, the average price per room by using the average price from the dataset provided. This allows us to take advantage of Excel's extensive functionality.

Year	Average price per room	
2017	£	90.05
2018	£	106.45
2019	£	101.78
2020	£	105.34
2021	£	101.23
2022	£	101.50
Total	£	101.06

Table 2 *Average price per room from 2017 to 2022*

The data shows that the hotel's average room rate stays relatively stable between 2018 and 2022, at around a median value of £101.06, except 2017, when it was £90.05 on average.



Despite the rapid expansion and a bigger than predicted return of visitor numbers in 2022, the recovery cycle in the UK hotel industry is likely to slow down in 2023 due to continued uncertainty of economic situations and increased operational expenditures (PwC, 2023).

2.4 Sales channels

Hotels may benefit from a well-thought-out distribution plan in several ways: they can reach more potential guests, increase their number of reservations, decrease their reliance on only one or two channels and learn which ones bring them the most revenue (Cloudbeds, 2022).

To compare reservation methods over the whole data set, we are utilising Excel's built-in COUNTIF function. For each year specified in *Market_type*, this method will return the total number of reservations for that year.

Reservations channels	Total	2017	2018	2019	2020	2021	2022
Online	23,214	2,730	18,992	119	156	118	1,099
Offline	10,528	2,766	7,262	45	34	58	363
Aviation	125	0	121	0	0	0	4
Complementary	391	128	228	4	3	4	24
Corporate	2,017	476	1,435	9	5	12	80
Total	36,275	6,100	28,038	177	198	192	1,570

Table 3 Sales channels per year from 2017 to 2022

When split down by year and channel, online booking consistently beat all others, with the exception lockdown period due to Covid-19. In all periods, the offline channel performed second best, in front of the corporate channel.

There is evidence to show that the company has been effective in producing revenue across several channels, with the best results coming from its online and offline activities. The reservations reduction from 2020 and 2021 implies that the company was affected by issues

outside of its control, such as Covid-19 restrictions. However, the significant rise in sales in 2022 indicates that the company has recovered and is growing its revenue through its existing sales channels for the long term.

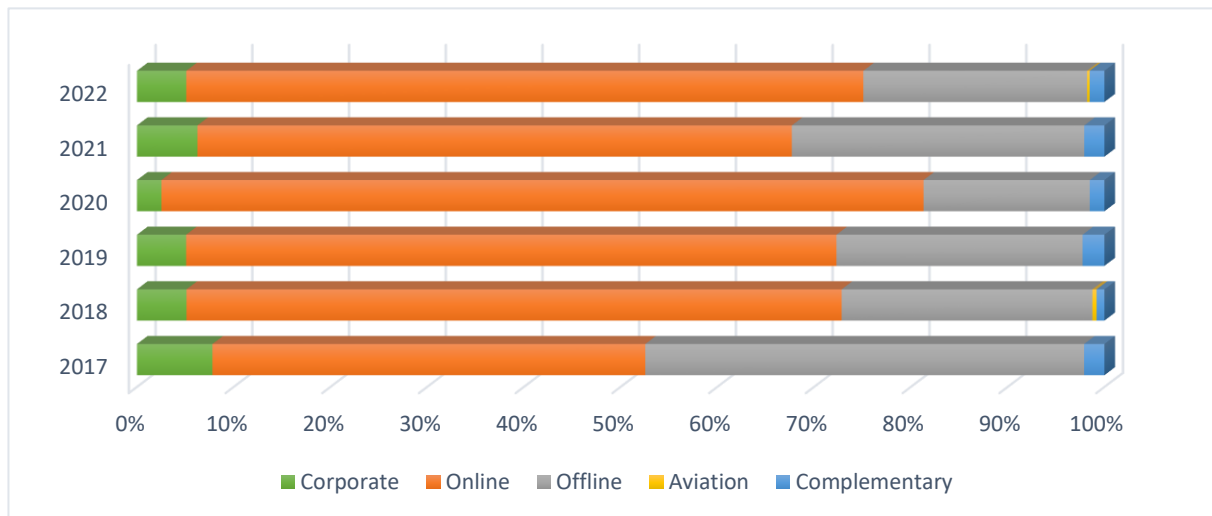


Figure 3 Sales channels per year from 2017 to 2022

3. Data Analysis of the student survey

Descriptive statistics assist with making sense of your data by providing a high-level overview of the connections between different categories. It's the first thing you should do before diving into any data analysis (Kaur, Stoltzfus and Yellapu, 2018).

3.1 Gender analysis

	Frequency	Percent
Female	1056	49.51%
Male	1077	50.49%
Total	2133	100.00%

Table 4 Gender analysis

Out of a total of 2133 student entries, 1056 are female, or 49.51 percent of the total, while 1077 are male, or 50.41 percent. This information was obtained from the output data gathered during the gender analysis. This indicates that

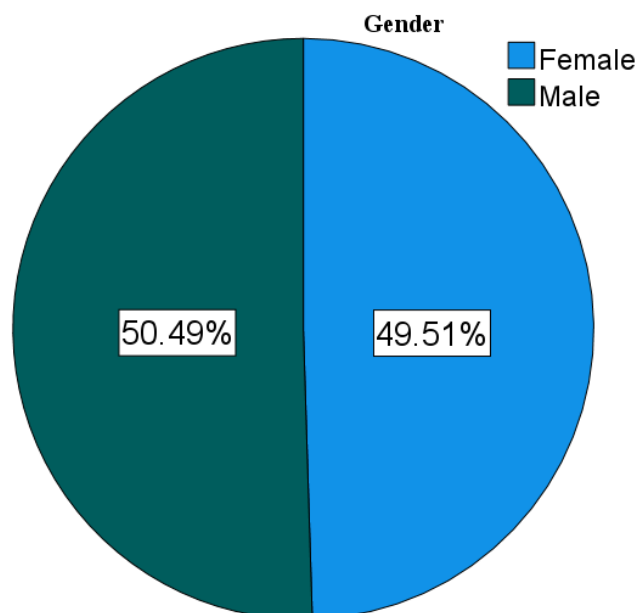


Figure 4 Gender visual description

gender is almost equally represented in the student survey.

3.2 Lunch analysis

	Frequency	Percent
Does not qualify	1215	56.96%
Qualifies for reduced/free lunch	918	43.04%
Total	2133	100%

Table 5 SPSS Launch Analysis

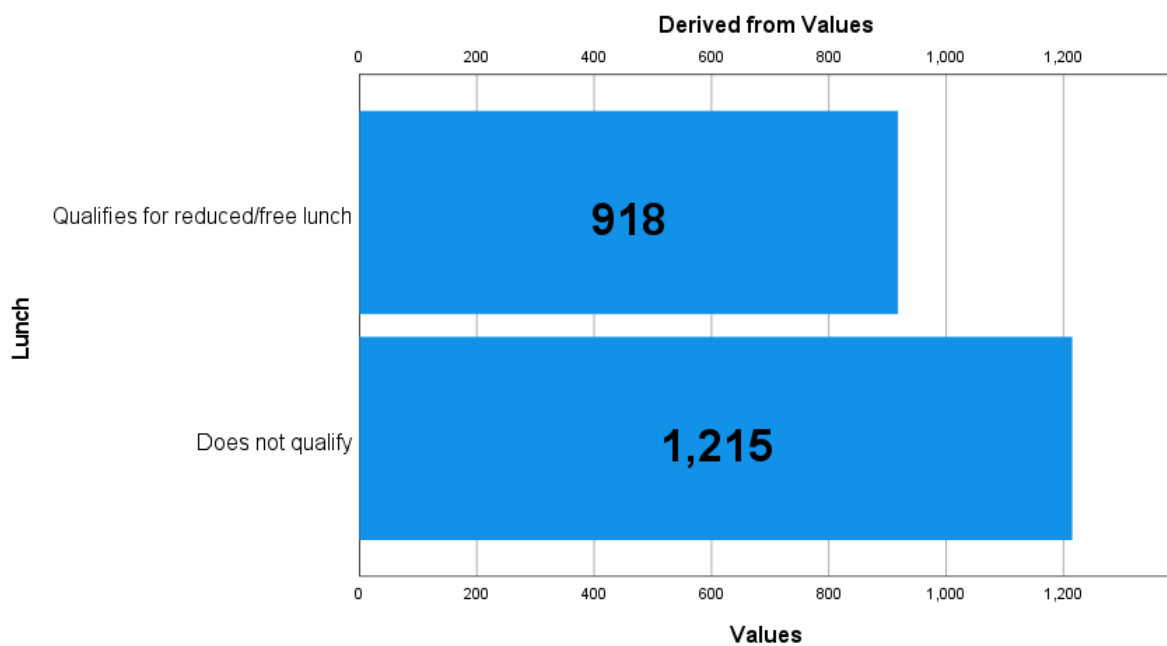


Figure 5 Lunch analysis visual description

In our data analysis to determine which students qualify for free or discounted lunches, we found that out of a total of 2133 students, 1215 students (or 56.96%) do not qualify for lunch, while 918 students (or 43.04%) do. This means that there are a lot of students who aren't able to get reduced or free lunches.

3.3 Mean and median of students over lunch qualification

N	Valid	2133
	Missing	0
Mean		1.43
Median		1.00

Table 6 Mean and median for launch qualification

This report has a mean of 1.43, which is closer to 1, indicating that 43% of the pupils in the study do not meet the requirements for free lunch. More students will be eligible for free or discounted lunches if this value gets closer to 2. In addition, the median value is 1, which indicates that the majority of entries contain information indicating that the students do not qualify for reduced or free lunch.

3.4 Mean and median differences in student's test scores

This report provides a breakdown of the pre-test and post-test scores based on whether or not students qualify for reduced/free lunch.

Lunch		Pre-test	Post-test
Does not qualify	N	1215	1215
	Mean	62.29	74.38
	Median	62.00	74.00
Qualifies for reduced/free lunch	N	918	918
	Mean	45.25	57.48
	Median	44.00	55.00
Total	N	2133	2133
	Mean	54.96	67.10
	Median	56.00	68.00

Table 7 Launch report in correlation with test results

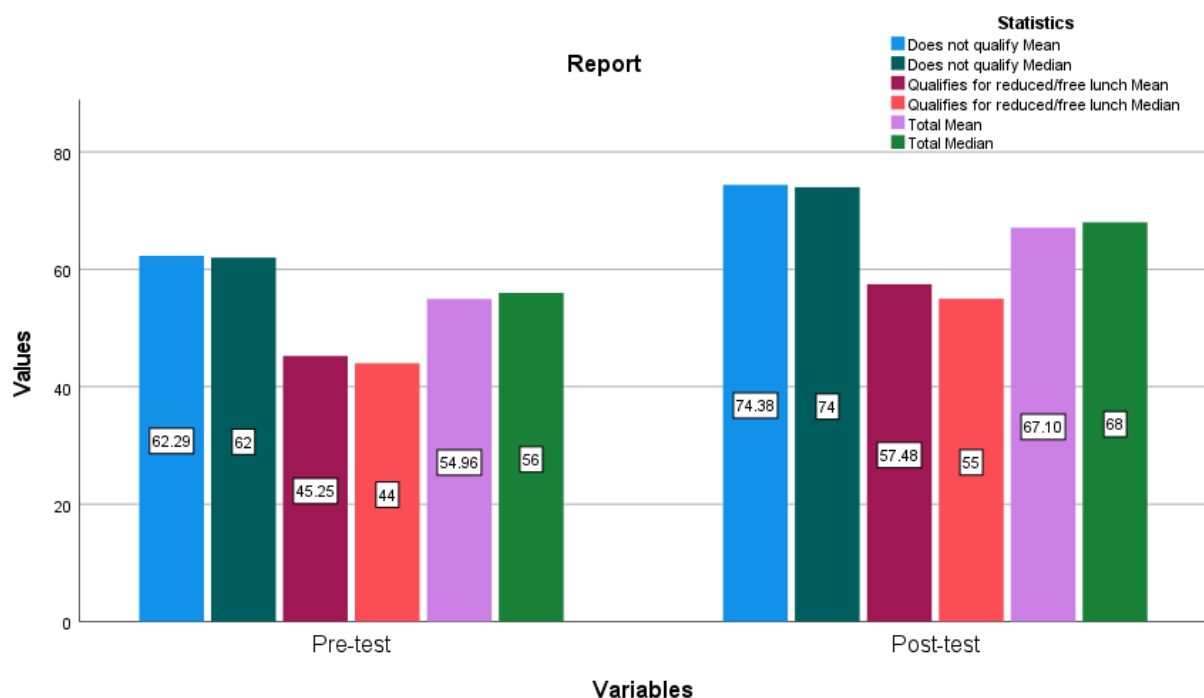


Figure 6 Visual description of launch report over test results

This analysis includes 1215 students who do not qualify for free or reduced lunch; their average pre-test score is 62.29 and their average post-test score is 74.38. Before and after testing, the median scores are 62 and 74, respectively.

There were 918 kids included in the analysis who were eligible for reduced or free lunch; their average pre-test score was 45.25 and their average post-test score was 57.48. The median pre-test and post-test scores were 44 and 55, respectively.

The average difference between students' pre-test and post-test scores was 12.14 points or 54.96 to 67.10. There was an increase of 12 points from the median initial score of 56 to the final score of 68.

The results of this study indicate that there is a statistically significant gap between the academic performance of children who get free or reduced-price lunches and those who do not.

3.5 Clustering

Once the data is ready to be analysed, we go for, as a clustering method, the TwoStep Cluster Analysis:

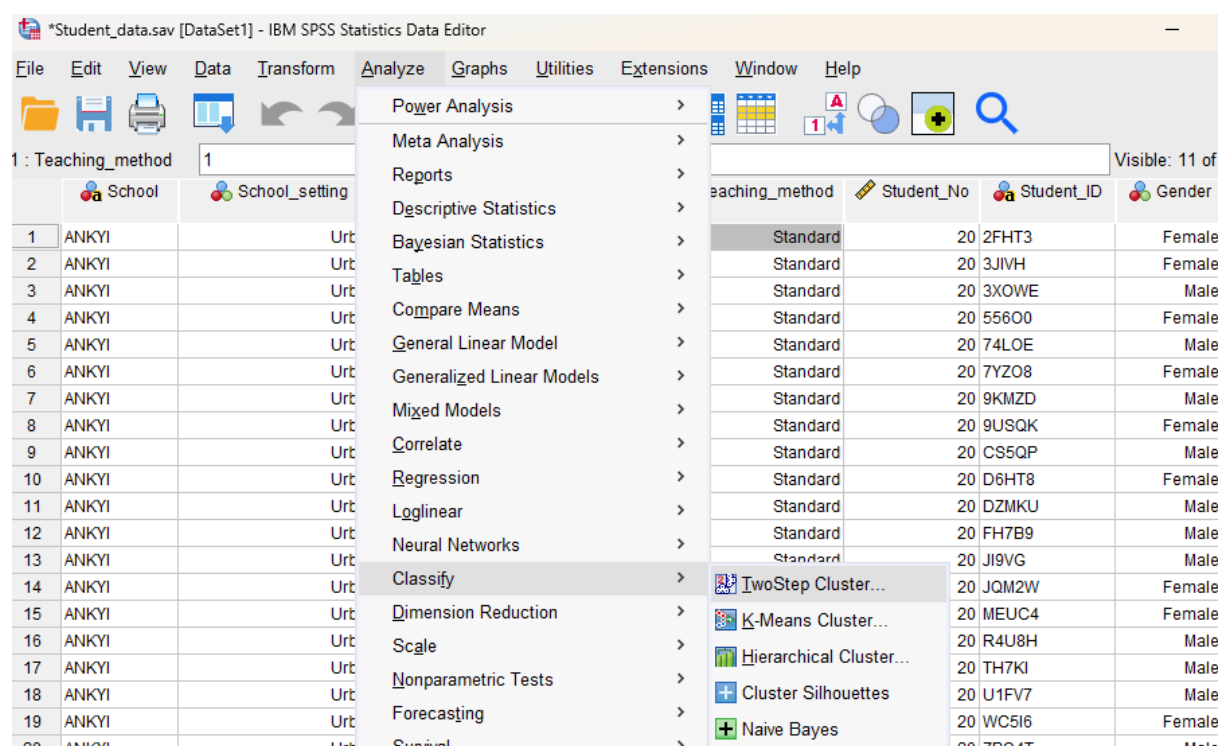


Figure 7 SPSS data base with Student_data

The investigation will examine the impact of teaching practises on students' performance and for this we will use the variable *Teaching method* and *Post-test*.

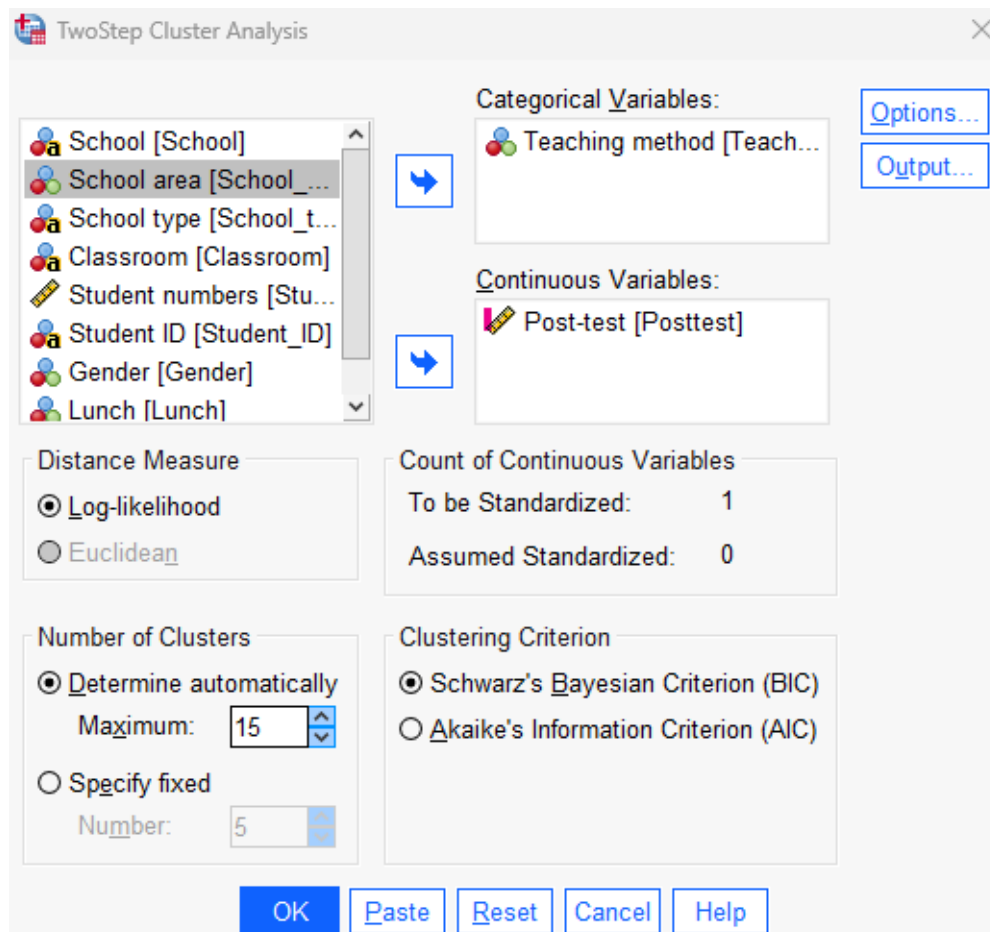


Figure 8 SPSS TwoStep Cluster Analysis choosing

Once the above step is completed, SPSS will provide a summary and a quality report.

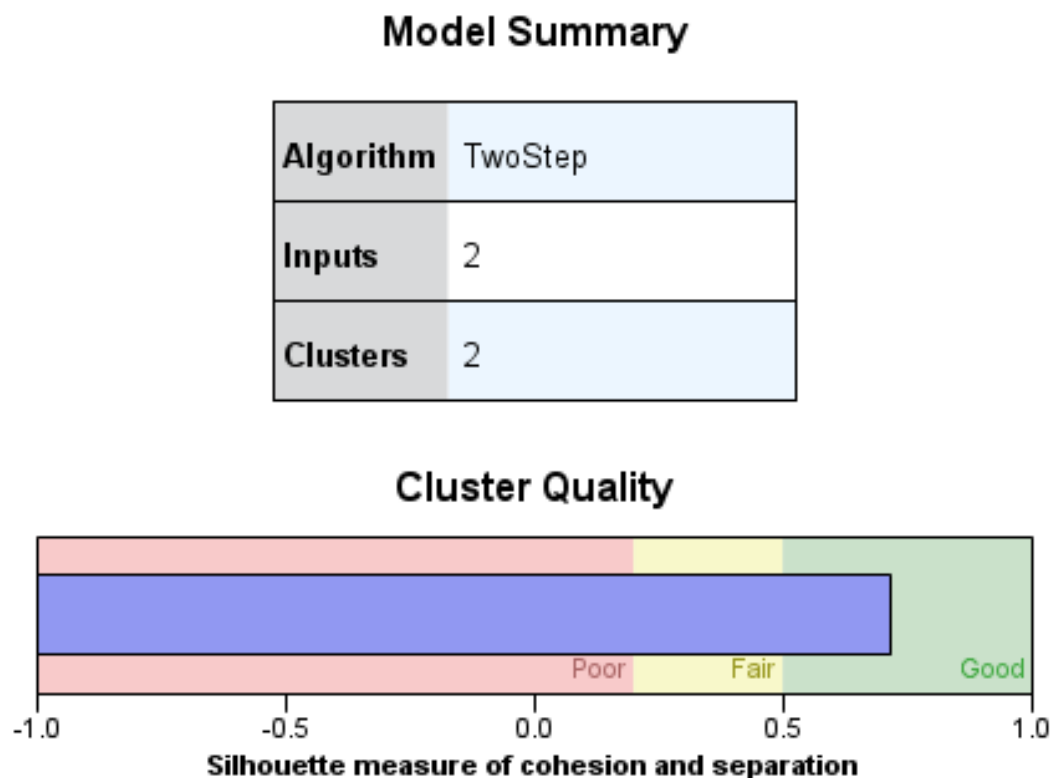


Figure 9 SPSS report on TwoStep Cluster Analysis

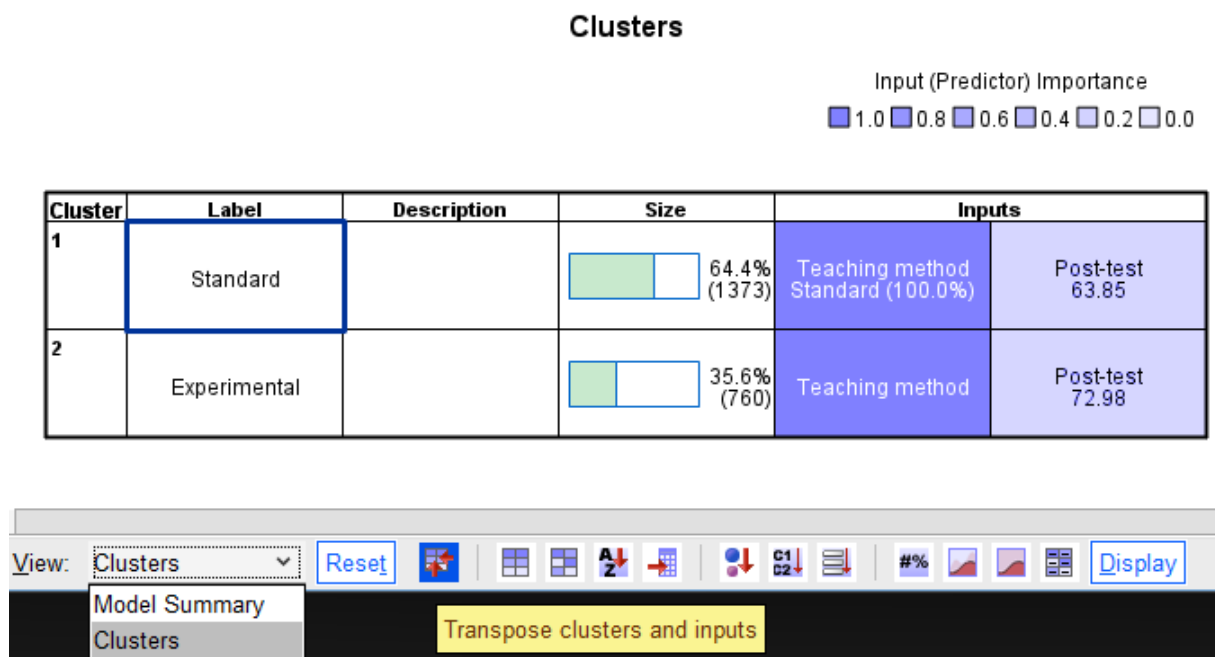


Figure 10 *Editing the report on TwoStep Cluster Analysis*

Once we begin editing the SPSS provided summary by changing the view from *Model Summary* to *Clusters* and adding labels to both clusters, 1 for Standard and 2 for Experimental, we'll have access to more interesting details about the analysis. Over 64% of students are being taught using the conventional approach; these students score an average of 63.85 on final exams, while only about 35.6% are being taught using the experimental approach, scoring an average of 72.98.

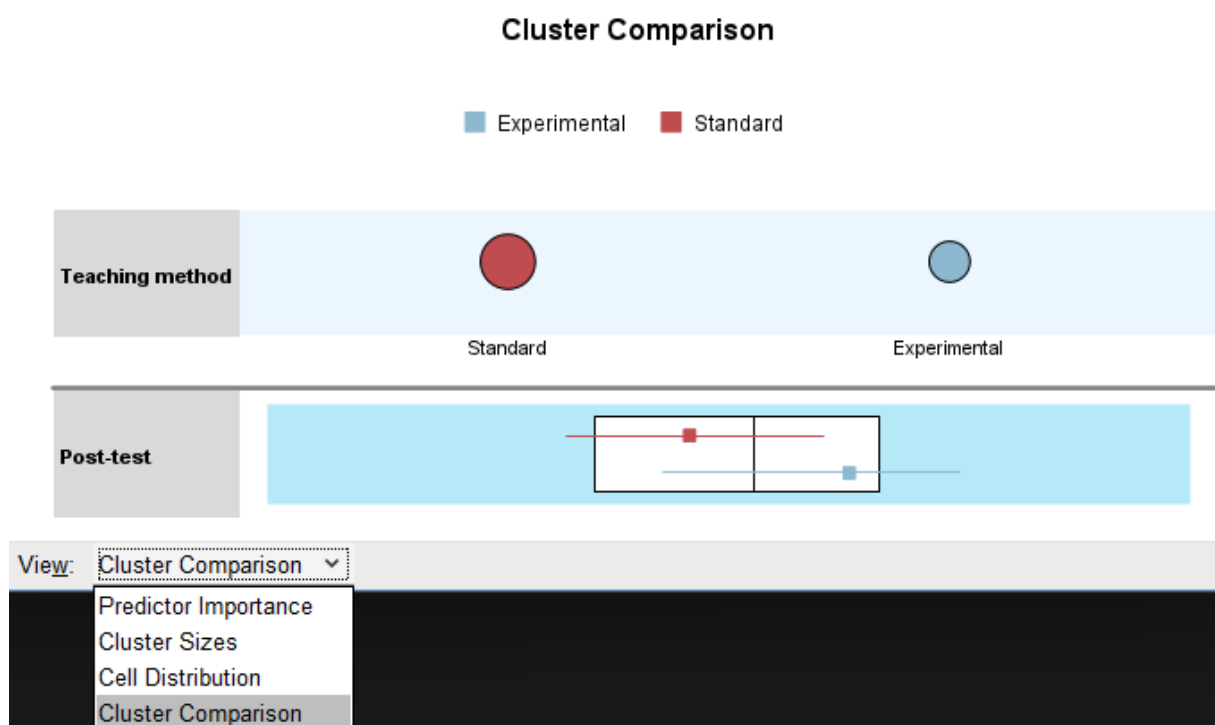


Figure 11 *Cluster Comparison*

While selecting both clusters, we can compare them in the other window, where we can see that even though the experimental teaching method cluster is smaller, the results are closer to the highest point, which indicates that the students have better results with this method.

3.6 Data mining

To find useful, complicated information hidden in enormous datasets, data miners use computing methods. As stated by Kallio and Tuimala (2013), data mining is not only utilised for discovering interesting patterns in the data, but also for studying big data sets, constructing models that identify particular elements of data, and generating forecasts based on the data.

- Cluster analysis is a method of statistical analysis that is used to learn more about the patterns and extreme cases that may be found in any given data collection. Cluster analysis is an unsupervised learning technique since it does not need the data points to be sorted into any predetermined groups (TIBCO, 2022).

Cluster analysis is being used by the marketing departments of retail businesses to classify clients into distinct groups according to their preferences and behaviours. Companies may often adapt their advertising campaigns to reach specific groups of their audience.

- One way to investigate and quantify this association is using regression analysis. Using a sample's worth of data as a starting point, regression analysis attempts to identify a mathematical connection between two or more variables (Dodge, 2008).

In business, regression is typically employed for foreseeing future outcomes. Regression analysis is used extensively by insurance providers to forecast the likelihood that customers will experience undesirable occurrences such as accidents or burglaries.

- Association rule mining is a method for exploring a huge dataset for patterns of association between previously unnoticed variables (Goh and Ang, 2007).

This method helps businesses in developing sales strategies, for example, due to the high salt level of fast-food customers often end up buying a soft drink because of their thirst. They capitalised on this by including the soft drink as part of combination meals that included food that would make you thirsty.

3.7 Text mining

Zanini and Dhawan (2015) to derive useful numerical indices from text, a technique known as text mining is employed. Its primary goal is to transform the text into a structured format that may be used by a variety of data mining statistical methods.

- The purpose of using sentiment analysis is to evaluate whether the emotion expressed in a piece of writing is positive, negative, or neutral. To do this, points are assigned to each sentence depending on the number of positive and negative words and adjectives that appear inside it (Follestad, 2022).

Sentiment analysis is a powerful method to monitor brand perception, identify areas of improvement for your company, and get insight into your target audience. You'll be able to narrow in on the exact features of your goods or service that clients love or hate.

- Using an approach called topic modelling, you may find groupings of terms that frequently appear together in the same documents throughout all the content and this is done by using unsupervised machine learning (Sheridan, 2022).

Most modern organisations use a topic modelling algorithm that can tell from the text of documents that arrive whether they are contracts, bills, feedback, or something else entirely. This method helps businesses to be more effective and quicker to react in their daily duties.

3.8 Advantages and disadvantages of using SPSS and Excel

	Advantages	Disadvantages
SPSS (Pallant, 2020)	<i>Advanced statistical analysis:</i> SPSS is well-suited to in-depth data analysis projects because of its broad range of sophisticated statistical methods, including regression and factor analysis.	<i>Cost:</i> Because SPSS is commercial software that requires a licence, it can be too costly for individuals and small businesses.
	<i>Easy data management:</i> users may easily manage massive datasets thanks to their data cleaning, manipulation, and transformation functions.	<i>Steep learning curve:</i> SPSS is requiring training to use it at its full potential compared to Excel that it's easier
	<i>Graphical capabilities:</i> When it comes to exploring and presenting data, SPSS's extensive visualisation tools let users generate a wide range of charts, graphs, and plots.	<i>Resource consumption:</i> SPSS may be demanding on system resources, especially when processing huge datasets or performing complicated examinations.
	<i>Integration with other software:</i> SPSS can communicate with other statistical programmes and databases, making it easier to share and use information.	<i>Limited customizability:</i> When it comes to modifying analyses or adding new features, SPSS has certain restrictions.
Excel (Shamshi, 2023)	<i>Familiarity and accessibility:</i> Excel is popular and well-known among many users, making it accessible and simple to	<i>Learning curve:</i> Gaining proficiency in Excel will open up several career doors for you. However, Excel has a steep

	Advantages	Disadvantages
	use for basic data analysis and visualisation activities.	learning curve and may take a lot of time to master.
	<i>Flexibility and versatility:</i> Excel have an extensive number of functions and formulas that enable users to conduct different computations and data conversions.	<i>Resource consumption:</i> Excel has limitations on handling large datasets, and performance can be affected when dealing with extensive computations or complex data manipulation.
	<i>Integration:</i> Excel's compatibility with other programmes makes it easy to exchange and collaborate on data.	<i>Data size limitations:</i> Excel's incapacity to handle large data sets could limit effective decision-making.
	<i>Easy learning curve:</i> training can be accessed both digitally, through sites and applications, and physically, through educational institutions or private courses.	<i>Prone to errors:</i> Because human error is unavoidable with Excel, this means most sheets are likely to include inaccuracies that will go unnoticed.

4. Conclusion

Excel's various features and functionalities make it a powerful data analysis tool, allowing for quick and easy pre-processing, analysis, and visualisation. We can efficiently manage and modify the Hotel Reservation dataset by employing Excel features like IF statements, LOOKUP functions, and PIVOT tables.

Because of its intuitive design and all-around availability, Excel may be used by a wide variety of people, including those without any programming experience. Data visualisation in the form of informative charts, graphs, and tables facilitates learning and transmission of findings.

Nevertheless, it is essential to take into consideration SPSS, an alternative data mining tool that was only lately obtained by the educational institution. The statistical analysis methods offered by SPSS are among the most sophisticated available, and the software was developed with data mining activities in mind.

References

- Aydin, C. (2023) Benefits of Using Conditional Formatting in Pivot Tables. Available at: <https://projectcubicle.com/benefits-of-using-conditional-formatting-in-pivot-tables/> (Accessed: 16 April 2023).
- Caspio (2020) The Risks of Using Excel for Data Management. Available at: <https://caspio.com/blog/excel-data-management-problems/> (Accessed: 14 April 2023).
- CFI (2022) Excel Definition. Available at: <https://corporatefinanceinstitute.com/resources/excel/excel-definition-overview/> (Accessed: 12 April 2023).
- Cloudbeds (2022) How to design a profitable hotel distribution strategy in 6 steps. Available at: <https://cloudbeds.com/articles/creating-a-profitable-hotel-distribution-strategy/> (Accessed: 17 April 2022).
- Derrfuss, J. (2022) Data preprocessing with Excel, psychology.nottingham.ac.uk. Available at: <https://psychology.nottingham.ac.uk/staff/lpzjd/psgy1001-21/preprocessing-excel.html> (Accessed: 14 April 2023).
- Dodge, Y. (2008) 'Regression Analysis', The Concise Encyclopedia of Statistics, pp. 450–452. Available at: https://doi.org/10.1007/978-0-387-32833-1_348.
- Follestad, M. (2022) Text and data mining methods. Available at: https://libguides.library.usyd.edu.au/text_data_mining/methods (Accessed: 6 May 2023).
- Goh, D.H. and Ang, R.P. (2007) 'An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents', Behavior Research Methods, 39(2), pp. 259–266. Available at: <https://doi.org/10.3758/bf03193156>.
- Gray, A. et al. (2022) 'What statistical software should I use? And does it actually matter?', New Zealand Medical Student Journal, 0(34), pp. 38–40. Available at: <https://doi.org/10.57129/gadb4864>.
- Jacob, J. (2021) Microsoft Excel vs Access - Pros and Cons. Available at: <https://hyperoffice.com/blog/2021/09/03/blog-microsoft-excel-vs-access-pros-and-cons/> (Accessed: 14 April 2023).
- JavaTpoint (2021) Advantages and Disadvantages of MS Excel - Javatpoint. Available at: <https://javatpoint.com/advantages-and-disadvantages-of-ms-excel> (Accessed: 14 April 2023).
- Kallio, A. and Tuimala, J. (2013) 'Data Mining', Encyclopedia of Systems Biology, pp. 525–528. Available at: https://doi.org/10.1007/978-1-4419-9863-7_599.
- Kaur, P., Stoltzfus, J. and Yellapu, V. (2018) 'Descriptive statistics', International Journal of Academic Medicine, 4(1), pp. 60–63. Available at: https://doi.org/10.4103/ijam.ijam_7_18.

- Microsoft (2023) Excel specifications and limits. Available at: <https://support.microsoft.com/en-us/office/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3> (Accessed: 14 April 2023).
- ONS (2021) Coronavirus and its impact on UK hospitality - Office for National Statistics. Available at: <https://ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/articles/coronavirusanditsimpactonukhospitality/january2020tojune2021> (Accessed: 15 April 2023).
- Pallant, J. (2020) SPSS Survival Manual. 7th ed. Berkshire, England: Open University Press. Available at: <https://lms.su.edu.pk/download?filename=1588697869-julie-pallant-spss-survival-manual-mcgraw-hill-house-2016-1.pdf&lesson=17247> (Accessed: 18 April 2023).
- Poutintsev, F. and Aditya, J. (2020) Pros and Cons of Using MS Excel. Available at: <https://honestproscons.com/pros-and-cons-of-using-ms-excel/> (Accessed: 14 April 2023).
- PwC (2023) Hotels Forecast 2022 - 2023. Available at: <https://pwc.co.uk/industries/hospitality-leisure/insights/uk-hotels-forecast.html> (Accessed: 17 April 2023).
- Shamshi, S. (2023) Advantages and Disadvantages of Excel: Top 10. Available at: <https://techquintal.com/advantages-and-disadvantages-of-excel/> (Accessed: 17 April 2023).
- Sheridan, S. (2022) What Is Topic Modeling? A Beginner's Guide. Available at: <https://levity.ai/blog/what-is-topic-modeling> (Accessed: 18 April 2023).
- Sigma (2020) 5 Disadvantages of Excel for Data Visualization. Available at: <https://sigmacomputing.com/blog/5-excel-disadvantages-for-data-visualization> (Accessed: 16 April 2023).
- Siman, E. (2023) What Is a Dataset? Definitive Guide. Available at: <https://brightdata.com/blog/web-data/what-is-a-dataset> (Accessed: 14 April 2023).
- Smith, S. (2019) In What Ways Is MS Excel Used to Interpret Data? Available at: <https://smallbusiness.chron.com/ways-ms-excel-used-interpret-data-72984.html>.
- TIBCO (2022) What is Cluster Analysis? Available at: <https://tibco.com/reference-center/what-is-cluster-analysis> (Accessed: 18 April 2023).
- Zanini, N. and Dhawan, V. (2015) 'Text Mining: An introduction to theory and some applications', Research Matters, pp. 38–44. Available at: <https://cambridgeassessment.org.uk/Images/466185-text-mining-an-introduction-to-theory-and-some-applications-.pdf> (Accessed: 18 April 2023).