

University of Wales Trinity Saint David

Analysing different kinds of Data

**Data Handling and Business Intelligence
Module: LCBB5000**

Lecture: Xavier Matieni

Student name: Daniela Georgieva

Student ID: 1915925

Table of Contents

<i>Introduction.....</i>	<i>2</i>
<i>Discussion.....</i>	<i>3</i>
<i>Part A</i>	<i>3</i>
<i>Part B</i>	<i>7</i>
<i>Conclusion</i>	<i>13</i>
<i>References</i>	<i>14</i>

Introduction

The systematic collecting, processing, and analysis of data to produce insightful information for well-informed decision-making is the foundation of data handling and business intelligence. Data handling is the process of organising, storing, and guaranteeing the correctness and accessibility of data. Business intelligence gives companies a competitive edge by converting raw data into intelligence that can be used through the use of sophisticated analytics and reporting technologies. Business intelligence fosters data-driven tactics for successful execution and execution by helping organisations spot patterns, developments, and prospects through intricate algorithms and visualisations. Hence, this essay will shed light on the appropriate use of Excel functions to use the data for certain interpretations. Other than this, the concept of SPSS, along with the pros and cons, the concept of data mining and text mining will also be covered in the essay. All the calculations and graphical diagrams will be represented and covered in the essay respectively.

Discussion

Part A



Figure: Column graph

Source: Self-created

The hotel prices for the years 2017, 2019, 2020, and 2021 are represented in the dataset. It sheds light on the variations and patterns in hotel rates over this time. Prices in 2017 exhibit a great deal of volatility (Bothell, 2023). Values like 0 and 6 indicate situations where the price may not have been recorded or was very low, while values like 159.75 and 200.75 indicate comparatively high prices. As we go into 2019, the price fluctuations persist, but with increasingly discernible patterns. A range of pricing is indicated by notable figures 155, 168.3, and 186.22. Even yet, some values remain inconsistent, like 0, which could indicate missing or unrecorded data. 2019 has a mixed bag of high and low costs, which could indicate that certain events or seasons affect hotel rates. Values such as 1 and 0 indicate a clear decline in prices in 2020, the year of the global COVID-19 epidemic (Bothell, 2023). The tourism and hospitality sectors were severely disrupted by the pandemic, which resulted in decreased demand and ultimately cheaper hotel rates. The data for

2020 shows this declining trend, with a few outliers (such as 215.6) that may not be impacted by the overall industry slump. With figures like 224.67 and 205, the prices in 2021 show indications of recovery when compared to 2020. The price increase over time points to a recovery in the hotel business as demand started to pick up and travel restrictions relaxed (Bothell, 2023). Still, some values, such as 0 and 6, are present, highlighting the significance of closely examining outliers and possible data abnormalities.

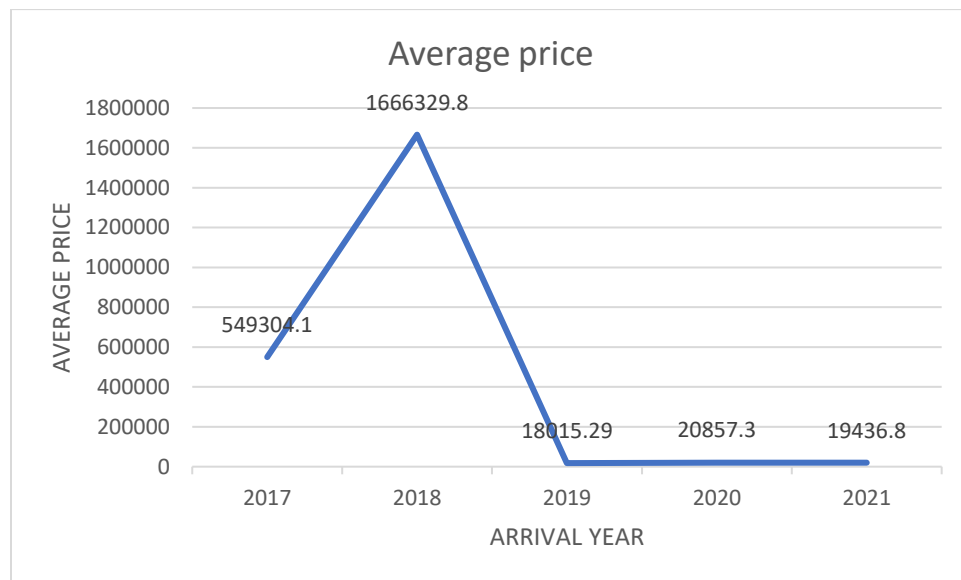


Figure: Showing the line graph

Source: Self-created

Moving ahead the data that has been provided seems to consist of a set of numerical numbers that correspond to several years, namely 2019, 2020, and 2021. Each row has a year and a corresponding numerical value, suggesting that it represents a range of measures or metrics. Three years are covered by the dataset: 2019, 2020, and 2021. There are numerical values connected to each year, and it is essential to comprehend what these values mean (Bowler-Hill and Hernandez, 2023). The data has a wide range, ranging from zero to numbers greater than 200. This means that different metrics, like amounts, percentages, or scores, could be

represented by the measurements. There is a range of values in the data for 2019 from zero to 228.33. A few entries appear more than once, suggesting that there may be mistakes in the dataset or that there are different measurements for the same statistic. The maximum value is 228.33, however, there are many zeros, which could indicate that certain data are missing or not defined. With values ranging from zero to 215.6, the 2020 data provides new metrics (Bowler-Hill and Hernandez, 2023). As in 2019, there are duplicate entries and instances of zero, suggesting that the dataset may have gaps. In contrast to 2019, the year 2020 appears to have a wider range of measures and a more diverse range of values. With numbers ranging from zero to 205, the 2021 data maintains the trend of varied measurements. There are also recurrent entries and zero incidents this year. Interestingly, certain measurements, such as "0" and "75," show up every year for all three, pointing to a possible shared statistic or signal.

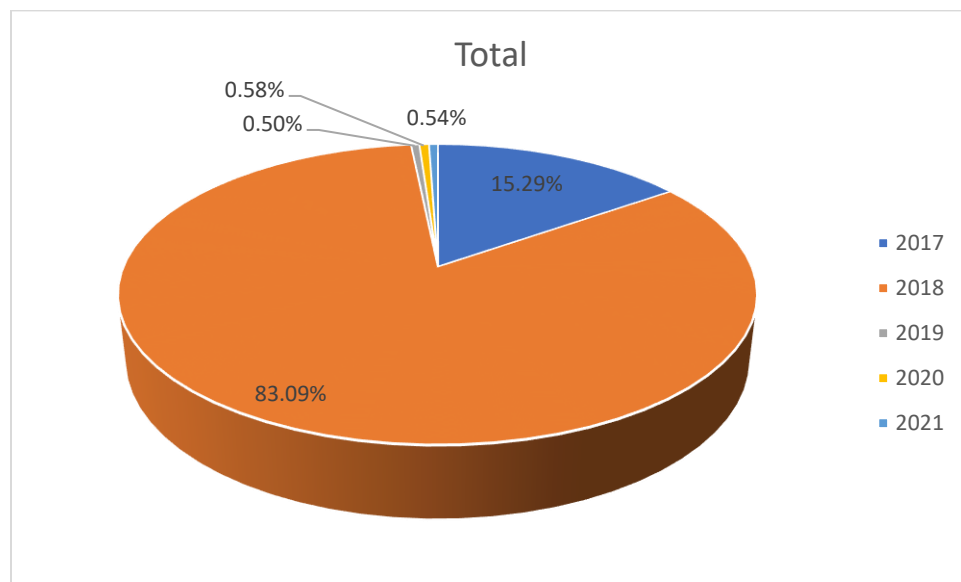


Figure: Showing the pie chart

Source: Self-created

One data processing tool used in spreadsheet programs such as Microsoft Excel is the pivot table. It enables users to condense, examine, and display huge datasets in an organised manner. Users are able to obtain insightful information quickly by reorganising and classifying data. With the use of pivot tables, the analysis may be dynamic and interactive, making it simple to identify patterns and trends in the data (Gallahar, 2023). To arrange data according to many criteria, including counts, percentages, averages, sums, and counts, users can drag and drop fields. Due to this feature, pivot tables are an effective tool for producing insightful reports that help with decision-making and for decomposing complicated data sets into simpler forms. The distribution of arrivals across several years is seen in the data supplied, expressed as percentages. By examining these percentages, one can gain an understanding of the patterns and variations in hotel occupancy during the designated years. With arrivals making up 15.29% of the total in 2017, there may have been a smaller surge of visitors (Gallahar, 2023). The corresponding hotel rates, as covered in the prior dataset, showed a good deal of variation without a clear pattern, suggesting a variety of pricing factors or possible seasonality. There was a notable increase in arrivals in 2018, accounting for an astounding 83.09% of the total. There could be several reasons for this surge in demand, including more travel, business, or events. Consequently, it would be fascinating to investigate how hotel rates changed in response to this spike in demand. Hotel prices typically increase when supply cannot keep up with demand, which is consistent with the economic theory of supply and demand (Hertens *et al.*, 2023). The percentages of total arrivals in the following years, 2019, 2020, and 2021, range from 0.50% to 0.58%, which are significantly lower. The COVID-19 epidemic and other external variables that may be influencing travel patterns could be the cause of these low percentages. Hotel prices, as covered in the preceding dataset, had a considerable decline in 2020 when the pandemic severely affected tourism worldwide (Hertens *et al.*, 2023). The low arrival rate in that year is consistent with the general pattern of lower travel during the pandemic. Despite a little uptick from 2020, the percentage of arrivals in 2021 stayed quite low at 0.54%. The steady recovery may have had an impact on hotel rates, indicating a

possible rise, but a thorough understanding of pricing dynamics cannot be obtained solely from percentages. Significant fluctuations in demand over time are highlighted by the arrival percentage statistics. 2018's spike in arrivals probably had a favourable impact on hotel rates; but, smaller percentages in the following years may have contributed to more steady or even declining pricing trends, particularly during the pandemic-induced recession in 2020 (Hertens *et al.*, 2023). A thorough grasp of the intricate relationship between demand and pricing in the hospitality sector throughout the chosen years can be obtained by examining both arrival percentages and hotel rates.

Part B

Interpreting the number of male and female students in the school with interpretation

Formula used for determining the male =COUNTIF (B: B, "Male")

Formula used for determining the female =COUNTIF (B: B, "Female")

Within the provided list, there are 1077 instances of men and 1056 instances of women. With a total of 2133 entries, the data appears to represent a series of gender labels. With a slightly higher number of men than women, the gender distribution seems to be fairly balanced (Iannone and Thoma, 2023). This harmony raises the possibility that the dataset represents a varied or haphazard selection of people. Numerous disciplines, including sociology, marketing, and demographics, can benefit from the analysis of gender distributions. Gaining an understanding of how gender is represented in a dataset can assist in spotting biases or trends or provide guidance for decision-making processes that take gender into account.

Interpreting how many students qualifies and do not qualify for reduced/free lunch.

Formula used for determining the qualify = =COUNTIF (D:D, "Qualifies for reduced/free lunch")

Formula used for determining the does not qualify= =COUNTIF (D:D, "does not qualify")

Based on data analysis, the school ANKYI has 918 students who are eligible for reduced or free meals out of the total student population, while 1215 students are not (Iannone and Thoma, 2023). The fact that 918 pupils are eligible for free or reduced lunches suggests that a sizeable percentage of kids are having financial difficulties. In order to address any potential concerns regarding financial limitations that children and their families might be facing, the school needs access to this information. To guarantee that students have access to a healthy diet and a supportive learning environment, it advocates for focused interventions and support services, such as extra resources for those who qualify. However, the number of 1215 pupils who are not eligible for free or reduced lunches indicates that they are a somewhat better-off population within the school (Iannone and Thoma, 2023). The distribution of resources and the customization of educational programmes to accommodate the varied demands of the student body can be aided by this data. Comprehending the economic heterogeneity among the student body enables the educational institution to execute tactics that cultivate a welcoming atmosphere and tackle the particular obstacles encountered by distinct student demographics. An essential tool in this research is Excel's Countif function. With its ability to count cells accurately and efficiently according to predetermined criteria, it is a very useful tool for working with enormous datasets. Countif makes it easier to extract useful information from data, enabling administrators and educators to spot trends and patterns in the data quickly (Jelen, 2023). In this instance, it makes it easier to determine precisely how many pupils are eligible for reduced or free lunches and how many are not.

Determining the mean and median of students who qualify and do not qualify for reduced/free lunch.

Formula used for calculating mean =AVERAGE (E3:E4)

Formula used for calculating median =MEDIAN (E3:E4)

The number of students who do and do not qualify for free or reduced lunch is 1066.5, according to the mean and median estimates (Jelen, 2023). This statistical data offers insightful information about the distribution and central patterns of the ANKYI school's student body. Subtracting the entire quantity of values from the total value of all values yields the mean, commonly referred to as the average. According to the given circumstances, a mean of 1066.5 indicates that, on average, 1066.5 students fall into each category those who are eligible for free or reduced lunch and those who are not (Jelen, 2023). This metric is affected by values that are extreme and gives a distribution with a balance point. Conversely, the centre value in a dataset arranged from least to largest is represented by the median. A median score of 1066.5 signifies that the proportion of students in each category who fall above and below this threshold is equal. As it is less susceptible to extreme values, this statistic provides a more reliable indication of central tendency. A somewhat symmetric distribution is implied by the two groups' mean and median values being equal. A fair even ratio across the requirements for being eligible for reduced/free lunch is highlighted by the uniformity, which shows a balanced proportion of kids in both categories (Joo *et al.*, 2023). It is easier to evaluate the general socioeconomic status of the student body when one is aware of the mean and median. Based on the convergence of these indicators at 1066.5, it may be inferred that a considerable proportion of students are eligible for free or reduced lunch on average, while an equal number are not.

The utilisation of data mining and text mining techniques is widely acknowledged as a crucial component of modern corporate strategy, as they offer valuable insights, facilitate decision-making, and boost overall operational effectiveness in a range of sectors (Joo *et al.*, 2023). Data mining has five different parts, the first includes association rule mining. Association rule mining can be used in the retail industry to find trends in consumer purchase behaviour. By identifying which products customers most often purchase together, retailers may strategically arrange comparable products in-store or online to boost sales. The next comes credit scoring. Based on

past data, categorising engines in credit scoring assist in determining if a consumer is likely to fail on a loan. Informed loan decisions and risk assessments are facilitated by these particular lenders (Joo *et al.*, 2023). Social networking sites group people based on shared interests or habits through a process called clustering. Platforms can increase user engagement by delivering recommendations, ads, or tailored content to groups of users. Regression analysis is a tool that manufacturers use to forecast equipment failure based on a variety of factors, including usage, temperature, and pressure. By scheduling maintenance ahead of time, businesses may reduce downtime and avoid crucial breakdowns. Anomaly detection algorithms in cybersecurity are able to recognise peculiar network traffic patterns that can point to a cyberattack. Organisations may quickly identify and address security threats with the aid of this technique.

Now, text mining also constitutes the five major elements that contribute to the overall business environment and those are elaborated in the next section (Sena and Ariyachandra, 2023). Sentiment analysis is a tool used by social media platforms to learn about people's opinions of goods, companies, and occasions. Businesses can use this data to measure public sentiment and adjust their marketing tactics accordingly. Medical record data, including patient names, conditions, and treatment regimens, are extracted and categorised using NER in the healthcare industry. Data accessibility and organisation are enhanced by this. Topic modelling is a technique used by news agencies to automatically group news articles into categories like finance, sports, and politics. This makes it easier to organise and retrieve content efficiently (Sena and Ariyachandra, 2023). Based on the type of client inquiry, customer support teams automatically classify and prioritise help requests using text classification. This expedites response times and raises client contentment. By taking into account variables like user behaviour, content popularity, and keyword relevancy, search engines employ information retrieval techniques to present relevant search results. This guarantees that consumers get the most relevant information fast.

SPSS	
<i>Advantages</i>	<i>Disadvantages</i>
1. Advanced statistical analysis is the focus of SPSS's design. It may be used for sophisticated studies because it provides a broad range of statistical processes. For instance, SPSS gives researchers the tools and results they need to perform a multivariate analysis of variance (MANOVA) and determine how several independent factors affect several dependent variables. 2. Strong tools are available in SPSS for reporting and data visualisation (Sylvia <i>et al.</i>, 2023). Within the software, researchers can easily and effectively generate tables, graphs, and charts. When presenting findings to stakeholders, this feature is helpful. In a market research study, for example, SPSS may produce aesthetically pleasing graphs that show customer patterns and preferences. 3. With its capabilities for handling, processing, and cleaning up big information, SPSS is an excellent data management tool. It is simple for researchers to compute new variables, select cases, and recode variables.	1. Particularly for those who are not experienced with statistical software, SPSS can have a challenging learning curve (Sena and Ariyachandra, 2023). It could be difficult for beginners to understand the syntax needed for some analysis and to traverse the many options. This learning curve can present challenges, especially for individuals with little experience with statistics. 2. Being a commercial software program, SPSS can be expensive to licence, particularly for lone researchers or small businesses with tight budgets. Those who cannot afford the licensing costs may not be able to access the programme due to this pricing aspect. 3. Although SPSS is more effective for statistical analysis, it is not as flexible as Excel when it comes to ordinary spreadsheet features (Sylvia <i>et al.</i>, 2023). While SPSS is mostly focused on statistical analysis, Excel is capable of performing a wide range of functions, such as project management, financial computations, and data storage.

SPSS can effectively handle and manipulate data to analyse patterns over time in a longitudinal study tracking academic performance (Sylvia <i>et al.</i> , 2023).	
--	--

<i>Excel</i>	
<i>Advantages</i>	<i>Disadvantages</i>
1. Excel is generally accessible and well-known to most users. Since it is pre-installed on the majority of computers, both individuals and organisations can simply access it. This ubiquity makes data entry and manipulation simple and fast. 2. Beyond statistical analysis, Excel is a flexible tool. It can be used for many different things, including basic data analysis, project management, and budgeting. Financial modelling in a company setting can be done with Excel, enabling users to produce dynamic financial forecasts (Ashtiani and Raahmei, 2023). 3. Excel is well-known for having an intuitive user interface that makes it accessible to users of all skill levels. Many users are already	1. Excel does not have the sophisticated statistical analysis features of specialised software such as SPSS, even if it does include fundamental statistical operations (Ashtiani and Raahmei, 2023). It may be difficult or impossible to carry out complex studies in Excel, such as hierarchical linear modelling or structural equation modelling. 2. Excel can make it difficult to maintain data integrity, particularly when dealing with big datasets. When compared to specialised statistical software, the likelihood of mistakes, unintentional data modifications, or misinterpretation is increased. Errors in data entry or formulas can have major repercussions in important applications such as medical research.

accustomed to fundamental features like sorting, filtering, and chart creation because of its user-friendly spreadsheet format.	3. The scientific process relies heavily on exact replication, which can be difficult with Excel worksheets. Ensuring that analyses can be replicated by others is crucial, especially in collaborative research or when sharing findings (Ashtiani and Raahmei, 2023). The lack of transparency in Excel's data conversions and algorithms can make replication more difficult.
---	--

Conclusion

Through complex algorithms and visualisations, business intelligence assists businesses in identifying patterns, developments, and opportunities, hence promoting data-driven approaches for successful implementation. As a result, this essay has clarified how to use Excel functions appropriately to use the data for specific interpretations. Apart from this, the essay has also discussed the idea of SPSS, its advantages and disadvantages, data mining, and text mining. Every computation and graphical diagram has been handled and illustrated in the essay, correspondingly.

References

- Ashtiani, M.N. and Raahmei, B., (2023). News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Systems with Applications*, p.119509.
- Bothell, L.J., (2023). Summary Tables. *MS Excel Bootcamp*.
- Bowler-Hill, S. and Hernandez, G.E., (2023). 24/7 Library: Using OpenRefine and Excel to Visualize Large Data Sets.
- Gallahar, T., (2023). Business Information: Excel Dashboard.
- Hertens, L., Van Cauwenbergh, C., Van den Broeck, F., Sambaer, J., Van Haute, M., Hertens, P., De Baere, E., Accou, G., Nerinckx, F. and Leroy, B.P., 2023. Results of Belgian patients with RPE65-related inherited retinal dystrophy 6 months after treatment with voretigene neparvovec. *Investigative Ophthalmology & Visual Science*, 64(8), pp.762-762.
- Iannone, P. and Thoma, A., (2023). Interactive theorem provers for university mathematics: an exploratory study of students' perceptions. *International Journal of Mathematical Education in Science and Technology*, pp.1-23.
- Jelen, B., (2023). MAX MINUS MIN IN A PIVOT TABLE. *Strategic Finance*, 104(8), pp.54-55.
- Joo, S.R., Kim, J.C. and Kim, S.J., (2023). Establishment of Data Mining-Based Public Education Administrative Work Automation System and Student Activity Analysis. *Journal of Autonomous Intelligence*, 6(3).
- Sena, M.P. and Ariyachandra, T., (2023). An Examination of Tableau as a Supplement to Excel to Enhance Data Literacy Skills. *Information Systems Education Journal*, 21(4).
- Sylvia, M.L. and Terhaar, M.F. eds., (2023). *Clinical analytics and data management for the DNP*. Springer Publishing Company.