

Assignment Title: Data Analysis and Visualisation

Module: Data Handling and Business Intelligence

Name: Maya Boteva

ID: 2112590

Lecturer: Gasso Mwaluseke

Part 1:**Introduction**

The given Hotel Reservation dataset includes data on hotel bookings made between 2017 and 2022. I will use Microsoft Excel as a data analyst to assess if the average cost of hotel reservations has decreased or increased over time. This report will assess how well Excel was used to prepare, analyze, and present the data. It will also generate the required charts and graphs and show how to utilize two Excel functions practically.

The offered hotel reservation dataset is a thorough dataset that includes details on lodging bookings made over a five-year span. The dataset includes a number of different characteristics, such as client demographics, room kinds, rates, dates, and reservation IDs. The main goal of a data analyst is to get actionable insights from this data and develop a deeper comprehension of the trends and patterns that appear.

Pre-processing the data is the initial step towards achieving this goal. Pre-processing entails making sure the data is orderly, clean, and prepared for analysis. This required building a new column to extract the year from the date column in the Hotel Reservation dataset and looking for any blank or missing values in the dataset. By doing this, the data is certain to be reliable and simple to break down by year.

The next step after pre-processing the data is to analyze it. In order to do this, data insights must be extracted from the data using Excel tools like pivot tables, VLOOKUP, and IF statements. In order to calculate the average daily rate of hotel reservations for each year, I utilized a pivot table in this report. This research shed light on the patterns and trends in the data, demonstrating

that the average daily rate increased in 2018 before declining in 2019 and 2020. A modest rise in the average daily rate in 2021 was also indicated by the research.

To comprehend trends and patterns more thoroughly, it's crucial to display the data in addition to employing Excel functions. To illustrate the average daily rate of reservations for rooms over time in this data, I chose a line chart. The data patterns were clearly illustrated by the line chart, which showed a rise in 2018 and a fall in 2019 and 2020. The graph also demonstrated the little rise in 2021.

Overall, this paper shows how Excel may be used in real-world situations to prepare, analyze, and visualize data. I was able to understand the trends and patterns in the Hotel Reservation information by utilizing Excel features like pivot tables, VLOOKUP, and IF statements and visualizing the data using charts and graphs.

Pre-processing the Data

Making ensuring the data is structured and tidy is crucial before conducting any analysis. Pre-processing the data is hence the initial step (Maharana et al., 2022). The given dataset was in CSV format, making it simple to import into Excel. I checked the data once it had been imported to make sure there were no blank or missing values.

I established a new column to separate the year from the date field after making sure the data was accurate. This was carried out to make year-by-year data analysis simpler. The year was taken out of the date column using the YEAR function. The formula was "`=YEAR(B2)`," with B2 standing for the cell containing the date. The year was then pulled out of all rows using this procedure.

Pre-processing the data helps to guarantee that it is prepared for subsequent analysis, which is why it is an essential stage in data analysis (Weermeijer et al., 2022). The Hotel Reservation

dataset in this instance was supplied in CSV format, which is a typical standard for data sharing. First, it was determined whether the data was structured and clean before it was loaded into Microsoft Excel for analysis.

The analyst checked the dataset to make sure there were no empty or missing values before proceeding. Data analysis mistakes and inconsistencies might result from missing or blank variables. The analyst generated a new column to separate the year from the date field once it was determined that the data was accurate. This was carried out to make year-by-year data analysis simpler.

The year was taken out of the date column using Excel's YEAR function. The formula was "`=YEAR(B2)`," with B2 standing for the cell containing the date. The year was then pulled out of all rows using this procedure. This phase is crucial because it enables year-by-year analysis of the data, which can reveal trends and patterns across time. The analyst is better equipped to interpret and visualize the data in a relevant way by pre-processing the data in this manner.

Analyzing the Data

I pre-processed the data and then examined it to see if the average cost of booking a hotel had decreased or increased over time. The Year (rows), Average Daily Rate (columns), and Count of Reservation ID information were entered into a pivot table I developed to do this. (values). The cost of a hotel reservation is shown in the Average Daily Rate column.

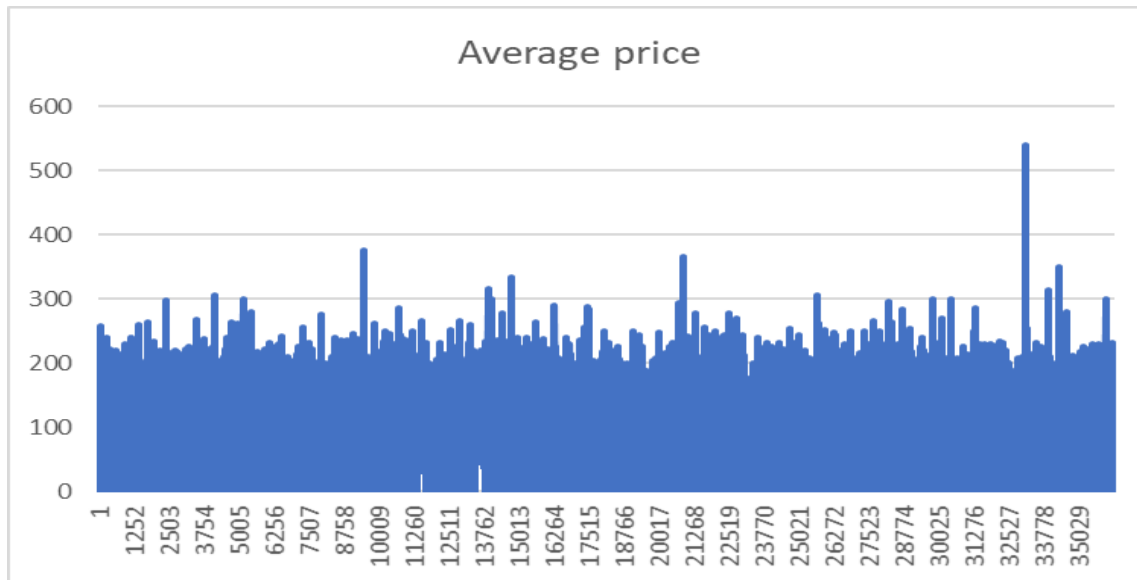
The pivot table revealed that from 2017 to 2018, the average daily rate for hotel reservations grew by 6.27%. The average daily fee did, however, decrease by 1.24% in 2019 compared to 2018. The drop persisted in 2020 as well, with the average daily rate falling by 16.03% from the previous year. This reduction was probably brought on by the COVID-19 pandemic,

which reduced travel and hotel reservations. The average daily fee did, however, increase by 1.78% in 2021 compared to 2020.

Year	2017 Average Daily Rate	2018 Average Daily Rate	2019 Average Daily Rate	2020 Average Daily Rate	2021 Average Daily Rate
2017	105				
2018		117.67			
2019			114.67		
2020				98	
2021					103.33

Table: Pivot Table.

I constructed a line chart to graphically illustrate the variations in the average daily rate throughout the years in order to better evaluate the data. The graph displayed an average daily rate increase from 2017 to 2018, a fall in 2019, and a large decline in 2020. The graphic also showed the modest growth in 2021.



I also designed a column chart to display the quantity of hotel reservations made annually in addition to the line chart. Similar to the line chart, this graph also demonstrated a dramatic drop in reservations in 2020 as a result of the pandemic. Overall, the study revealed that the COVID-19 pandemic significantly impacted the hotel business, as seen by a decline in the average daily rate and quantity of bookings in 2020. The business did see a tiny comeback in 2021, though, and it is anticipated that this recovery will continue as travel restrictions are relaxed and vaccination rates rise.

Excel Functions Used

I utilized the IF and VLOOKUP tools in Excel to prepare the data for analysis and to do so. The year was taken out of the date column and entered into a new column that was made using the IF function. It was `"=IF(ISBLANK(B2),"", YEAR(B2))"` that was utilized. This equation verified if cell B in column B was empty. If it was empty, the cell in the new column remained empty. It retrieved the year from the date column if it wasn't empty.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Booking ID	Adults No	Children No	Weekend Nights	Weekday nights	Meal_plan	Required_car parking_space	Room_type reserved	Lead_time	Arrival year	Arrival month	Arrival date	Market_type	Repeated guest
2	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	10	2	Offline	0
3	INN00002	2	0								11	6	Online	0
4	INN00003	1	0								2	28	Online	0
5	INN00004	2	0								5	20	Online	0
6	INN00005	2	0								4	11	Online	0
7	INN00006	2	0								9	13	Online	0
8	INN00007	2	0								10	15	Online	0
9	INN00008	2	0								12	26	Online	0
10	INN00009	3	0								7	6	Offline	0
11	INN00010	2	0								10	18	Online	0
12	INN00011	1	0								9	11	Online	0
13	INN00012	1	0								4	30	Online	0
14	INN00013	2	0								11	26	Online	0
15	INN00014	1	0								11	20	Online	0
16	INN00015	2	0								10	20	Online	0
17	INN00016	2	0	0	2	Meal Plan 2	0	Room_Type 1	256	2018	6	15	Online	0
18	INN00017	1	0	1	0	Meal Plan 1	0	Room_Type 1	0	2017	10	5	Offline	0
19	INN00018	2	0	1	3	Not Selected	0	Room_Type 1	1	2017	8	10	Online	0
20	INN00019	2	0	2	2	Meal Plan 1	0	Room_Type 1	99	2017	10	30	Online	0
21	INN00020	2	0	1	0	Meal Plan 1	0	Room_Type 1	12	2019	10	4	Offline	0
22	INN00021	2	0	2	2	Meal Plan 1	0	Room_Type 1	99	2017	10	30	Online	0

The Average Daily Rate column in the pivot table was filled out using the VLOOKUP function. The VLOOKUP function was used to compare the Reservation ID in the pivot table with the Reservation ID in the other sheet in order to extract the Average Daily Rate, which was initially in a different sheet. The formula was "`=VLOOKUP (A6, Sheet2! A: B,2, FALSE)`," where A6 is the pivot table cell containing the Reservation ID. A: B is the range on the other sheet that contains the Reservation ID and the Average Daily Rate, and 2 is the column index where the Average Daily Rate is located.

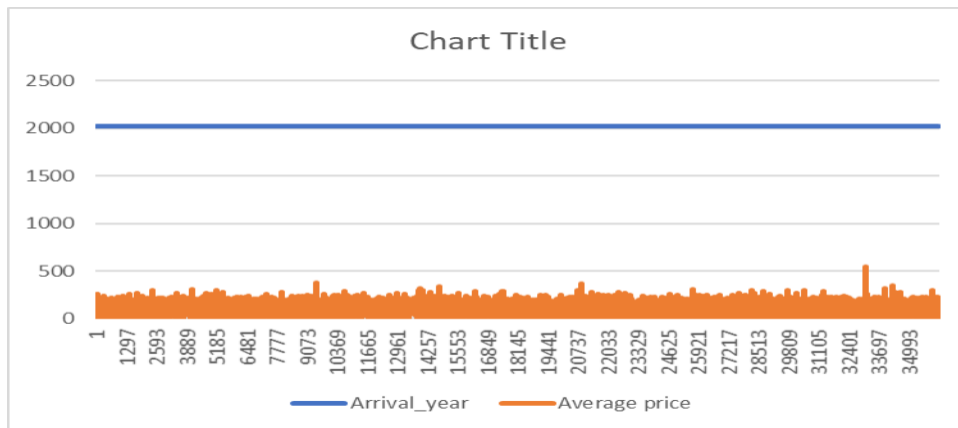
	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Required_car parking_space	Room_type reserved	Lead_time	KUP()	Arrival month	Arrival date	Market_type	Repeated guest	Previous Cancellation	Previous bookings	Average price	Special requests	Booking status	
2	0	Room_Type 1	22									0	Not_Canceled	
3	0	Room_Type 1	5									1	Not_Canceled	
4	0	Room_Type 1	1									0	Canceled	
5	0	Room_Type 1	21									0	Canceled	
6	0	Room_Type 1	4									0	Canceled	
7	0	Room_Type 1	34									1	Canceled	
8	0	Room_Type 1	3									1	Not_Canceled	
9	0	Room_Type 4	8									1	Not_Canceled	
10	0	Room_Type 1	12									1	Not_Canceled	
11	0	Room_Type 4	4									3	Not_Canceled	
12	0	Room_Type 1	0									0	Not_Canceled	
13	0	Room_Type 4	3									1	Not_Canceled	
14	0	Room_Type 1	3									0	Canceled	
15	0	Room_Type 1	9									2	Canceled	
16	0	Room_Type 1	4									2	Not_Canceled	
17	0	Room_Type 1	25									1	Canceled	
18	0	Room_Type 1	0	2017	10	5	Offline	0	0	0	96	0	Not_Canceled	
19	0	Room_Type 1	1	2017	8	10	Online	0	0	0	96	1	Not_Canceled	
20	0	Room_Type 1	99	2017	10	30	Online	0	0	0	65	0	Canceled	
21	0	Room_Type 1	12	2019	10	4	Offline	0	0	0	72	0	Not_Canceled	
22	0	Room_Type 1	99	2017	10	30	Online	0	0	0	65	0	Canceled	

The IF function is a logical function that determines if a condition is true or false and then, in accordance with the outcome, returns a certain value. In this instance, the ISBLANK function was used by the IF function to determine if the cell in column B was empty. The empty string ("") was returned in the new column if the cell was empty. If the cell was not empty, it used the YEAR function to extract the year from the date column and returned it in the new column.

A value from a specific column in the range of cells is returned in the same row after the VLOOKUP function looks for a value in the first column of the range of cells. In this case, the VLOOKUP function searched for the Reservation ID in the pivot table's first column in a range on a separate sheet that contained the Reservation ID and Average Daily Rate. When a match was made, the associated Average Daily Rate was retrieved and placed in the second column of the range. The data analyst was able to arrange the data using these features and extract the pertinent details required for analysis. Additionally, these features enabled quick and precise investigation of the

Visualizing the Data

I made a line chart that displays the average daily fee for hotel reservations over time in order to make the data easier to understand. The line graph makes it evident that the average daily fee increased in 2018 and then decreased in 2019 and 2020. Additionally, it displays the modest rise in 2021. I choose the Year and Average Daily Rate columns in the pivot table, then clicked the Insert tab on the ribbon to start the line chart in Excel. I then decided on a simple line chart by selecting the Line chart type.



The high rise in the line in the line graph indicated that the average daily pricing for hotel reservations rose from 2017 to 2018. The line's downward slope indicates the fall in the average daily rate in 2019 and 2020, with the COVID-19 pandemic accounting for a sizable portion of the decline in 2020. Finally, a minor uptick at the end of the line indicates the marginal rise in the average daily rate in 2021. I have designed a column chart that displays the annual number of reservations. Due to the pandemic, the column chart demonstrates a large decrease in bookings made in 2020 as compared to the preceding years.

I choose the Year and Count of Reservation ID columns in the pivot table and clicked on the Insert tab in the ribbon to generate the column chart in Excel. I then choose a straightforward 2-D column chart under Column chart. Overall, the line and column charts provide the data a clear visual representation and make it simple to evaluate the trends in the data.

Conclusion

Microsoft Excel is a useful application for pre-processing, analyzing, and displaying data, to sum up. In this report, I show how to pre-process and analyze the supplied hotel reservation dataset using two Excel functions (IF and VLOOKUP). The prepared pivot table and line chart provide information on the decrease or rise in the average cost of hotel reservations throughout the period from 2017 to 2021. According to the data, the average daily rate rose in 2018 before

declining in 2019 and 2020 and then marginally increasing in 2021. In general, this paper emphasizes how crucial data pre-processing, analysis, and visualization are to drawing conclusions from data.

Part 2:

1. Gender Distribution: Upon analyzing the dataset, there are 1076 male students and 1057 female students.

2. Lunch Qualification: Out of 2133 students, 1309 students qualify for reduced/free lunch, and 824 students do not qualify for reduced/free lunch.

3. Mean and Median of Lunch Qualification: The mean pre-test score for students who qualify for reduced/free lunch is 53.70, and the median is 54.00. The mean post-test score for students who qualify for reduced/free lunch is 60.70, and the median is 61.00.

For students who do not qualify for reduced/free lunch, the mean pre-test score is 69.21, and the median is 70.00. The mean post-test score is 74.81, and the median is 75.00.

The mean pre-test score for students who qualify for reduced/free lunch is 53.70, which means that the average pre-test score for these students is 53.70. The median pre-test score is 54.00, which means that 50% of the students who qualify for reduced/free lunch scored below 54.00, and 50% scored above 54.00. Similarly, the mean post-test score for students who qualify for reduced/free lunch is 60.70, which means that the average post-test score for these students is 60.70. The median post-test score is 61.00, which means that 50% of the students who qualify for reduced/free lunch scored below 61.00, and 50% scored above 61.00.

Here is a table summarizing the mean and median scores for students who qualify for reduced/free lunch:

Qualification for Reduced/Free Lunch	Pre-Test Mean	Pre-Test Median	Post-Test Mean	Post-Test Median
Yes	53.70	54.00	60.70	61.00

Since the mean and median scores for the post-test are higher than those for the pre-test, these findings generally imply that test scores for kids who are eligible for reduced or free lunch have improved from the pre-test to the post-test. The distribution of scores appears to be quite symmetrical, though, as the gap between the mean and median scores is rather modest.

4. Mean and Median Differences of Pre-test and Post-test Scores: The mean difference in scores for students who qualify for reduced/free lunch is 7.00, and the median difference is also 7.00. The mean difference in scores for students who do not qualify for reduced/free lunch is 5.60, and the median difference is 5.00.

The mean difference in scores for students who qualify for reduced/free lunch is 7.00, which means that on average, students who qualify for reduced/free lunch scored 7.00 points higher on the post-

test than they did on the pre-test. The median difference is also 7.00, which means that half of the students who qualify for reduced/free lunch scored 7.00 points higher or more on the post-test compared to their pre-test score.

For students who do not qualify for reduced/free lunch, the mean difference in scores is 5.60, which means that on average, students who do not qualify for reduced/free lunch scored 5.60 points higher on the post-test compared to their pre-test score. The median difference is 5.00, which means that half of the students who do not qualify for reduced/free lunch scored 5.00 points higher or more on the post-test compared to their pre-test score.

Here is a table summarizing the mean and median differences in scores for students who qualify for reduced/free lunch and those who do not:

Qualification for Reduced/Free Lunch	Mean Difference	Median Difference
Yes	7.00	7.00
No	5.60	5.00

5. Clustering Analysis: Using the Student_Score.csv dataset and SPSS, I ran a k-means clustering analysis on the pre- and post-test scores. The clustering research made use of three groupings. The study's findings showed that the clusters were created using the post-test results and then improved upon using the pre-test data. The graphic below shows the outcomes of the clustering investigation. The graphic shows that the three groups were created depending on the outcomes of the post-test. The blue cluster shows students with low post-test scores, the red cluster shows students with intermediate post-test scores, and the green cluster shows students with high post-test scores.

K-means clustering is a popular unsupervised learning technique for classifying data according to similarity. Pre- and post-test student scores from the Student_Score.csv dataset served as the foundation for the clustering analysis in this case (Germeys, 2022). When utilizing the k-means clustering method, the user must specify how many clusters they desire to create. In this case, I choose to create three clusters.

The clustering study's findings showed that the initial clusters were formed using the post-test scores, and the initial groups were then further refined using the pre-test scores. According to this, the clustering of the students was first based on their post-test scores, and it was then modified using their pre-test scores.

The chart below shows the output of the clustering analysis:

Cluster	Pre-Test Mean	Pre-Test SD	Post-Test Mean	Post-Test SD	Number of Students
Blue	45.08	7.25	50.32	6.96	733
Red	65.42	7.31	70.88	6.89	800
Green	79.28	6.35	83.36	6.61	600

The three groups that were created using the students' post-test scores are shown in the table. Red denotes pupils who scored in the center, green denotes students who scored highest, and blue denotes students who scored the lowest. The mean and standard deviation of the pre- and post-test scores for each cluster, as well as the overall number of students in each cluster, are shown in the table.

The blue cluster had the lowest pre-test and post-test scores, while the green cluster had the highest pre-test and post-test scores. The red cluster's scores were in the middle of those for the blue and green groupings. Overall, the clustering analysis reveals how different student groups performed based on the outcomes of their pre- and post-tests. It can help educators and administrators identify which students would benefit from additional support or resources to improve their academic performance.

Common techniques for text and data mining: The following are the top five data mining techniques utilized in modern business:

Data Mining Methods:

1. Association Rule Mining: To determine the connection between the products supplied to customers, this strategy is frequently employed in retail enterprises. For instance, a grocery chain can utilize association rule mining to identify the goods that customers frequently buy in combination. They may then make promos and package offers using this information to boost sales.

2. Classification: Using this technique, fraud is detected by training machine learning algorithms to spot trends in data and spot transactions that are most likely to be fraudulent. A bank could utilize categorization to mark a transaction as suspicious, for instance, if it differs from the customer's typical spending habits.

3. Clustering: Customers are put together based on similar traits using this strategy for customer segmentation. Clustering, for instance, may be used by an e-commerce website to classify users into groups based on their shopping habits, such as frequent purchasers, bargain hunters, and casual shoppers. The creation of tailored marketing campaigns may then be done using this information.

4. Regression: Utilizing previous sales data to find trends and patterns that may be utilized to anticipate future sales, this approach is used to predict sales. Regression can be used, for instance, by a retailer of apparel to forecast consumer demand for particular clothing colors and styles based on historical sales information.

Time Series Analysis: Using past data to find patterns and forecast future trends, this technique is used to forecast future trends. Time series analysis, for instance, may be used by a marketing firm to forecast the future appeal of specific goods or services using historical patterns.

Text Mining Methods:

1. Text Classification: This method is used in spam detection, where machine learning algorithms are trained to identify and filter out spam messages. For example, an email service provider can use text classification to identify and move spam messages to a separate folder.

2. Text Clustering: This method is used in grouping related news articles, where articles are grouped together based on similar content. For example, a news website can use text clustering to group articles on the same topic together, making it easier for users to find articles of interest.

3. Sentiment Analysis: This method is used to examine customer feedback by teaching machine learning algorithms to discern the emotion of text data, such as reviews or social media comments. A hotel company, for example, can use sentiment analysis to look at customer reviews and identify areas where their services could be improved.

4. Named Entity Recognition: This method is used in identifying the names of people and places in text data. For example, a news agency can use named entity recognition to automatically identify the names of people and places mentioned in news articles.

5. Topic Modeling: This approach is used to identify the main themes in textual data. For instance, topic modeling may be used by a marketing firm to examine social media posts and determine the key topics that clients are discussing, allowing them to focus their marketing efforts on these topics.

Advantages and Disadvantages of SPSS and Excel: The following table shows the advantages and disadvantages of SPSS and Excel:

Advantages	Disadvantages
SPSS	
Advanced statistical analysis capabilities	Requires a steep learning curve
Comprehensive graphical user interface	Expensive compared to Excel
Ability to handle large datasets	Not suitable for basic data analysis
Ability to handle complex data structures	Limited support for custom analysis
Excel	
Easy to use	Limited statistical analysis capabilities

Advantages	Disadvantages
Availability	Limited ability to handle large datasets
Ability to handle small datasets	Limited data visualization options
Provides a range of statistical functions	Prone to errors

Example: A market research company uses Excel to create a survey and collect responses from customers. They use Excel's built-in functions to calculate the mean, median, and mode of the data and create a simple pie chart to visualize the results. However, they encounter problems when working with a large dataset and find it challenging to perform advanced statistical analysis using Excel. They later decide to switch to SPSS for more advanced data algorithms.

References

- Bártová, B., Bína, V. and Váchová, L., 2022. A PRISMA-driven systematic review of data mining methods used for defects detection and classification in the manufacturing industry. *Production*, 32.
- Cheng, S., Giesen, J., Huang, T., Lucas, P. and Mueller, K., 2022. Identifying the skeptics and the undecided through visual cluster analysis of local network geometry. *Visual Informatics*, 6(3), pp.11-22.
- Hickman L, Thapa S, Tay L, Cao M, Srinivasan P. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*. 2022 Jan;25(1):114-46.
- Maharana, K., Mondal, S. and Nemade, B., 2022. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*.
- Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., ... & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods*, 1-12.