

Data Handling and Business Intelligence

Programme: BA Hons Business Management

Lecturer: Dr Anupam Mazumdar

Student Name: Yordan Stankov

Student ID: 2021805

Determine an increase or decrease within the average price of rooms

Pivot table: This word is mainly applied in creating summaries of data that help in analysing data in spreadsheets and databases. With the help of this tool, it has become easier to make complex data to be simple considering the flexibility and the different powerful tools it offers (Domino, 2021). With the help of the tools, it turns tables that are complicated to be easy to understand as they are made more straightforward.

First, data is arranged in columns beginning with the pivot table, after this, data is classified into different categories and presented in a different new table. When using the pivot tool, the user has the freedom of deciding which kind of row they will display data while still applying various aggregations such as average, maximum, total and minimum. While using the pivot table it helps the user for finding the patterns and trends. Furthermore pivoting helps in filtering and sorting data in different ways. Furthermore, it also plays a role in giving a detailed idea about the information. In the field of accounting and finance, pivot tables are used in creating sales trends, creating budgets and also measuring financial performances. Pivot tables are also applied in the field of research, especially in analysing data in circumstances where the data being analysed is in large quantities and has complexity (Jelen., 2020)

Using a pivot table to determine the change in average prices

PIVOT TABLE				
	Data			
Booking status ▾	Sum of Arrival_year	Sum of Adults No	Sum of Children No	Sum of Average price
Canceled	23982971	22695	1478	1314361.75
Not_Canceled	49213493	44231	2341	2437327.13
Grand Total	73196464	66926	3819	3751688.88
VLOOKUP Formula				
Booking ID ▾	Average Price ▾			
INN00001	65			

PIVOT TABLE				
	Data			
Booking status	Sum of Arrival_year	Sum of Adults No	Sum of Children No	Sum of Average price
Cancelled	23982971	22695	1478	1314361.75
Not Cancelled	49213493	44231	2341	2437327.13
Grand Total	73196464	66926	3819	3751688.88

The above tables show the data that is related to the hotel bookings that were made. The tables display the entire figure of arrivals yearly, the entire figure of children and adults booking showing the average price paid in cases of both bookings that were not cancelled and those that were cancelled. The last column displays the total number of reservations that were done in the set.

The table has distributed the data in a way that a data set that was complex can now be easily understood. The table has separated the total number of bookings for children and adults from the total number of bookings yearly. This table has also portrayed the number of bookings that were cancelled and the bookings that were not cancelled. According to the table, the number of adults in the bookings that were cancelled was **22, 695**, while in the case of children, they

were up to **1478**. Furthermore, the bookings that were cancelled were almost double the bookings that were not cancelled, this is compared to the bookings that were not counselled for adults were **44, 231** while for children they were **2341**.

Considering the reservations that were cancelled, the average general price was \$131.36 on every booking while as compared to the non-cancelled reservations with an average price of \$2437.33 on every booking. The existing two-time difference could be a result of the fluctuations in prices of every reservation considering the bookings that were cancelled and those that were not cancelled as a result of the different factors such as the time the reservations were made, the number of days they were staying and the kind of meal plans that they prefer

Pivot table showing the reservation for the year 2017

	A	B	C
1	Arrival_year	2017	
2			
3	Sum of Arrival_year	Adults No	
4	Children No	2	Grand Total
5	1	235989	235989
6	Grand Total	235989	235989

Children No	2	Grand Total
1	235989	235989
Grand Total	235989	235989

According to the table above, the hotel has two categories under reservations. The first column is with the children's slot, while the second one is the column with two children. The column for bookings with one child has been displayed on the left side, while for parents with

two children has been displayed on the right side of the table. The total number of bookings for those with one child and two children are 235,989 and 235989, as they appear respectively. This then brings the total number of bookings to 471, 987.

This table displays the reservation data for bookings with one and two children. However, it does show only those with two and one children but it does not have the data of a column in which the number of children was more than two or that which had no children at all. With this data, it is of high significance to focus on the guidelines on the research having the type of family that books the hotel, to ensure that the prices have been set considering all the regulatory

The formula for "LOOKUP" in Microsoft Excel:

VLOOKUP	
Formula	
Booking	Average Price
ID	
INN0000	
1	65

To find the average price on a particular booking, the INN00001 ID will be used within the VLOOKUP formula. Considering the associated given table range, the formula helps in searching the booking ID considering the price range which has been given in the table. In this case study the given case price of reservation that is related to the ID INN00001 is \$65

With the data above, it can be reported that the revenue from the non-cancelled booking is upper than the revenue that was recorded from the reservations that were cancelled. It also shows that the average price of reservations for the bookings that were not cancelled is higher than the average cost for the reservations having cancelled bookings (Jelen, 2020). This has been explained to be as a result of many of the bookings were not cancelled and the average price for the booking was high as to the average cost of the bookings that were cancelled.

With the above, it can be concluded that the above gives us data relating to bookings within a hotel and the total amount of average price that the hotel charged their clients (Murray, 2022).

The above also gives the main impacts of the bookings that were not cancelled having revenue and a more increased price as to the bookings that were cancelled. The above also offers a tool to find the average price that is charged for a single booking ID which is provided in the form of the VLOOKUP formula.

Chart displaying the segmentation of booking having Year, Adults, Children and Price

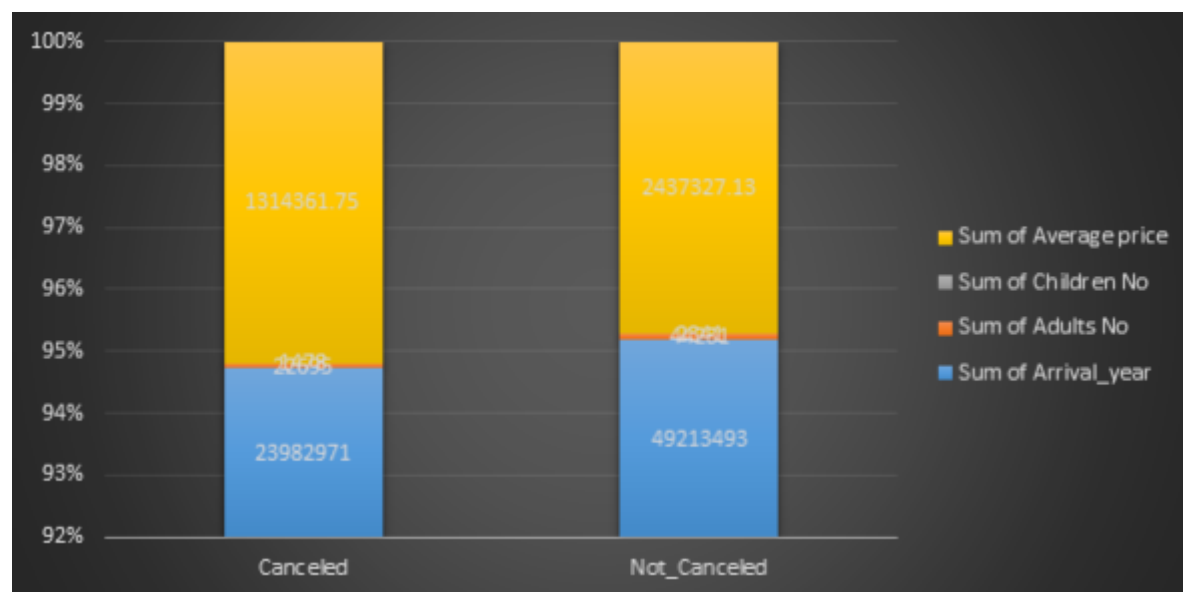


Chart displaying data

The above chart shows the average cost of bookings or the fluctuation in prices within a column for a particular time. No data has been given relating to the time of the data. This has enabled the creation of a line chart. To give the values of numbers that are related to the number of bookings having, kids, adults and the prices that were charged averagely for both bookings that were cancelled or not it is effective to use a stacked column bar chart.

In the chart, the X axis is used to show the arrivals, average price and the children. The Y axis is used to show the total value for the categories that were mentioned. The graph expounds on the total number of reservations that are cancelled were lower as to those which were not cancelled. It also shows that the average price for the cancelled bookings was lower than those that were not cancelled. With this type of data, it becomes hard to represent them in a column chart and it would not have been explained with such depth hence the use of bar graphs within this case is more impactful.

Part two

How many students are male and female?

Statistics

Gender

N	Missing	0
	Valid	2133

Gender

	Frequency	%	Valid %	Cumulative %
Male	1076	50.5	50.5	50.5
Valid Female	1055	49.5	49.5	100.0
Total	2133	100	100	

About the table above, it is showcased that there are a total of 2133 respondents that the study was conducted. From the table above, it shows that of the total 2133 respondents, 1056 were females 49.5% while 1077 were male (50.5%)

2) How many qualify and do not qualify?

Statistics

Lunch		
N	Valid	2133
	Missing	0

		Frequency	Percent	Valid percent	Cumulative Percent
Valid	Do not qualify	1215	57.0	57.0	57.0
	Qualify for free/ reduced lunch	918	43.0	43.0	100.0
	Total	2133	100.0	100.0	

According to the data from the tables, it is recorded that the research has a population sample of 2133 that participated in the study. The table above shows that, of the total participants 57% having 1215 did not qualify for lunch and 43% of the participants qualified for reduced/free lunch (918)

Mean and median of students who qualify and do not qualify for reduced/free lunch?

Statistics		
Lunch		
N	Valid	2133
	Missing	0
Mean		1.4304
Median		1.0000
Mode		1.00
Std. Deviation		.49525
Variance		.245
Range		1.00
Maximum		2.00

This table displays the median, mean as well as mode of the students who meet the qualifications the free or reduced lunch and those who do not meet. The table gives the mean value as 1.43 that states that many of the students do not meet the qualifications for the free or reduced lunch. Furthermore, the table shows that the mode and median value is 1.0. With this, we can also conclude that many of the participants do not meet the requirements for the free reduced lunch.

Mean and median differences of students' pre-test and post-test scores

Statistics			
		Pre-test	Post-test
N	Missing	0	0
	Valid	2133	2133
Mean		54.96	67.10
Mode		60.00	60.00
Median		56.000	67.000

According to the above data, it is clear that there are no values that are missing in the pre-test and post-test variables. After analysing the data above, it is clear that in the mean value of the variable that is pre-test is 54.96 and the mean value that is post-test is 67.10. This shows that after completing the test the average score can be improved by 12.1 points. On the other side, the pre-tests score considering the median value records 56 to the post-test its median value is 67. This shows that there is a gradual rise in the median value having 12 just after carrying out the experiment

Using the dataset student Score provided in conjunction with SPSS, provide analysis clustering (ex. K-Means Clustering). Show workings with screenshots and clearly explain the results.

Initial Cluster Centres		
	Cluster	
Post-test	35.00	100.00
Pre-test	23.00	93.00

Iteration History		Change in Cluster Centres	
Iteration		1	2
1		30.2	30.61
2		.884	.791
3		.61	.532
4		.210	.180
5		.226	.200
6		.150	.127

7	.000	.000

Final Cluster centres

	Cluster	
	1	2
Pre-test	43.05	65.32
Post-test	54.70	77.90

Distances between final

Cluster centre

Cluster	1	
1		32.12
2	32.12	

Number of cases in every cluster

Cluster	1	993.000
	2	1140.000
Valid		2133.000
		.000

The data above shows the cluster analysis just post-analysis of the data set for the 2133 people who participated in the study. According to the above data, it is clear that the first and second has a score of 23 and 93 while considering the post-tests set had 35 and 100. From the iteration history, it shows that the level of 7 and the label of iteration are achieved. Additionally, the final cluster table shows that the posters and pretest scores were 65.32 and 43.95 and 77.09 and 54.70 having both clusters 2 and 1 respectively. Hence according to the above data it can be concluded that the students existing within the first cluster have high scores and to those in cluster 2.

Explain a minimum of 5 most common data mining and text mining methods

Data Mining

Clustering Analysis:

This refers to a framework that puts similar data collectively focusing on individual criteria (Goel, 2020). For example, when retailers want to group clients considering their buying habits and develop a campaign that is personalised, they could use Clustering analysis

Classification analysis: This technique groups data within groups that are predefined considering the attributes. For example, banks use this tool in creating categories when applying for loans in terms of high and low risks, considering a variety of factors such as income, employment status and credit score.

Association Rule Mining: This is the technique that is used in the identification of patterns and relationships within items in a set of data. This tool is sometimes used by retailers during the identification of products that have been bought together. With this tool, they collect data

that is useful when creating personalised promotions or when the company wishes to offer promotions such as discounts for products bundling.

Anomaly detection: This tool is mainly used during the location of data points and patterns that are not usual which showcases cases of abnormality. For example, banks and financial institutions could use this to locate the transactions that were made fraudulently or any activities that are suspicious within the accounts of the clients

Regression analysis: With this tool, it helps in identifying relationships between the dependents and one or more independents which is initially done. For example, a real estate could use this analysis to locate factors that drive or lowers the prices of houses with a particular geographic location. These could include factors such as square footage, number of bathrooms, views and the number of rooms the house has.

Text Mining

Named entity recognition:

This tool is mainly used in extracting and identifying entities such as business companies, the names and the locations that a text has been sent from.

Text Classification: This technique is mainly used in categorising text within predefined groups focusing on a particular data. In an example, this could be used by the media platforms in identification as well as elimination of content that is inappropriate

Sentiment Analysis: This is a technique in text mining that uses natural language processing and machine learning to identify and extract the subject data within a text. For example,

hotels can use this technique to analyse client service as well as identify client service and identify the common difficulties and inadequacies that are being complained about

Topic Modelling: Using this technique performs the role of identifying and extracting from the key topic from the various significant. For example, the newly launched business could use it to locate the topics that have been discussed from various sources.

Discuss the advantages and disadvantages of using SPSS and Excel by use of relevant practical examples Advantages and Disadvantages of using SPSS

Graphical User Interface: Among the significant benefits of the software is its user-friendly interface. It becomes simpler for research people to carry out explorations and perform statistics studies as well as in a more effective way. The people who make use of this software could carry out operations that are complex and test statistics even if they do not have expertness in programming

Data manipulation and Handling: Various forms of managing data description have been provided by SPSS. It does not only give the user the ability to manage the missing values but it also plays a significant role in location of the variables that are independent and those that are dependent. Furthermore, it gives the consumer a chance to sort and update information and also carry out data importation from the various formats available. Furthermore using the SSP also provides the user with tools to record variables, dataset combination tools and data aggregation tools to ensure that effective presentation and preparation of data has been achieved.

Reporting and Results: An additional significance of using the SPSS is, it helps in developing reports that are detailed which give statistics analysis and reports (Rahman and

Muktadir, 2021). With these reports, it becomes simpler to write research reports and also create presentations using formats such as PDF and Word

Disadvantages

Cost: One of the biggest limitations of SPSS is that it needs a paid license. People who have limited and restricted budgets can find it challenging to use it is that there are other free alternatives such as Python and R. Furthermore with the cost factors this could be a disadvantage to companies as it increases the cost of production.

Programming capabilities are limited: Flexibility for statistics that are complex and handling of data using the SPSS could be limited. This is a result of the point-and-click interface. Despite that SPSS gives a syntax editor, it does not offer friendly services as compared to other languages for instance Python and R.

Advantages and Disadvantages of Excel

Advantages

Budgeting and planning of finances: Excel offers features that make it easier to budget and create financial plans. With this, it becomes easier for the user to calculate the totals as it also plays a role in developing projections as well as tracking expenses and incomes.

Project managing: When using Excel one can create timelines and track all the tasks and also oversee the progress while managing projects. With such, excel gives the user benefit as it offers a graphic presentation of the project schedules that plays a role in the identification of bottlenecks and managing all the dependents

Inventory management: Excel can be applied when managing inventory. With this, the tool offer feature helps follow the stock levels, work out recorder points, oversee the sales and also develop other reports concerning the inventory. This feature gives a well-organised method of analysing and organisation the inventory data (Steinke and Kopp, 2020).

Disadvantages

Limitations in scalability: The tool is not so effective in managing and handling huge datasets. When using the tool, the more the data, the slower it becomes. Hence considering processing data that is in large amounts excel is not an effective tool.

Data integrity issues: This tool mainly considers manual data entry that could lead to errors while inputting the data. With this, it stands a risk of losing the data or misinterpretations. Hence while using this tool; it is difficult to maintain the integrity of the data when many people are working on the same file at the same time

Practical example

For example, in a survey carried out by a researcher to carry out an investigation on the link between the number of hours a student takes and the type of grade that they will get. Data has been composed based on 50 participants that include the number of hours that they study and the exam grade. Below is an analysis of data with SPSS and Excel

- Create a column for the number of hours and another column for the Exam grade.
Input all the data regarding what was collected for the 50 students
- Select all the columns

- Choose a chart type that will be more suitable from the Insert section at the top part of the Excel
- Consider the type as adjustments such as the Y-axis label and positions adjustments are also being made
- The graph will then give an estimation of the amount of time that the students spent and their results

In SPSS

- Navigate through SPSS and create new data files
- Copy and paste the data from the Excel and begin testing the results using the SPSS
- After inputting the data, select analyse which is found at the top menu followed by selecting descriptive statistics followed by Frequencies
- Put the values for study time and the examination results within the right-hand box
- To have descriptive results make a report using measures such as Range, Mode, Mean and more
- Finalise by clicking OK

Conclusion

From the report above, it is clear that SPSS and Excel are tools of high significance when carrying out data analysis and arranging it in a particular form. This report analysed the use of various functions such as VLOOKUP, LOOK UP and Pivot table. Furthermore, research has been done and descriptive results alongside cluster analysis have also been carried out.

Furthermore, from this research, it has been concluded that SPSS is user-friendly and has capabilities of data visualisation and data handling. However, it has also been linked with limitations such as limitations in data handling. Furthermore, this report has also analysed the benefits and limitations of Excel

References

- Brzozowska, J. and Gola, A., 2021. Computer-aided assembly planning using MS Excel software—a case study. *Applied Computer Science*, 17(2), pp.70-89.
- Čaplová, Z. and Švábová, P., 2020. IBM SPSS statistics. In *Statistics and Probability in Forensic Anthropology* (pp. 343-352). Academic Press.
- Domino, M.A., Schrag, D., Webinger, M. and Troy, C., 2021. Linking data analytics to real-world business issues: The power of the pivot table. *Journal of Accounting Education*, 57, p.100744.
- Ferreira-Mello, R and et.al., 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6), p.e1332.
- Goel, N., 2020. A study of text mining techniques: Applications and Issues. *Pramana Research Journal*, 8(12), pp.307-316.
- Hassani, H and et.al., 2020. Text mining in big data analytics. *Big Data and Cognitive Computing*, 4(1), p.1.
- Jelen, B., 2020. Excel: A PIVOT TABLE WITH DATA FROM DIFFERENT WORKSHEETS. *Strategic Finance*, 101(8), pp.56-57.
- Jelen, B., 2022. Excel VLOOKUP FOR SPEED. *Strategic Finance*, 104(2), pp.60-61.
- Murray, A., 2022. The Infamous VLOOKUP Function. In *Advanced Excel Formulas: Unleashing Brilliance with Excel Formulas* (pp. 329-372). Berkeley, CA: Apress.

Rahman, A. and Muktadir, M.G., 2021. SPSS: an imperative quantitative data analysis tool for social science research. *Int. J. Res. Innov. Soc. Sci*, 5, pp.300-302.

Steinke, A. and Kopp, B., 2020. RELEX: An Excel-based software tool for sampling split-half reliability coefficients. *Methods in Psychology*, 2, p.100023.