



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

BA Business Management
Data Handling and Business Intelligence
LCBB5000

Assessment 2
Data Analysis and Visualisation

Student Name: Miroslava Velkova

Student Number: 2125655

Lecturer: Amitave Banik

Contents

Introduction	3	
Part 1	4	
Graph	6	
Pivot Table	7	
Sum if	8	
Part 2	9	
SPSS.....	9	
Number of Male & Female Students	10	
Number of Students Qualified for reduced/free lunch.....	10	
Mean and median of students who qualifies and does not qualify for reduced/free lunch	11	11
The mean and median differences of students pre-test and post-test scores	11	
Clustering	12	
5 Common Data mining and text mining methods for contemporary business.....	12	
Advantages and Disadvantage of using SPSS and Excel	12	
Conclusion	17	
Reference	18	

Introduction

In the contemporary business scene data handling as well as business intelligence are essential for the success of modern businesses; As it leads to better decision making, improved operational efficiency and provides the firm with a competitive advantage. In this report, data analysis and its subsequent visualization will be carried out. The report is divided into two parts part one consists of analyzing hotel reservation data set as well as finding out the decline or increase in average price over the year from 2017 to 2021; as well as finding out that information by utilizing 2 excel functions some if and pivot table. A graph will be provided for better understanding of the data.

Part 2 will consist of analyzing your local school on various aspects using spaces such as the number of male and female students;

In this report five most common data mining and text mining methods will be provided that can be used for contemporary businesses and advantages and disadvantages of using SPSS and Excel will be discussed with practical examples.

Part 1

The first step in data analysis in excel is the preprocessing of data in excel which are steps undertaking to clean and transform raw and unfiltered data into more suitable and further analysis (García et al, 2016). This involves removing duplicates or correcting data entry errors or converting data types or rearranging data to make it easier for analyzing or filtering out the data as well as sorting out the data (Katal et al, 2013).

The first requirement would be to find the decline and increase of average price or otherwise known as room reservation price from the hotel reservation data set From 2017 to 2021; In order to determine that data out of the data set 3 columns must be extrapolated they are arrival year to know more about in which year the average price increased or decreased; Not to mention the average price of the year and lastly the most important data is booking status as many reservations were cancelled thus making the prize being reduced as there are; total of 36,275 entries an out of them 24,390 are not cancelled and 11,885 are cancelled; this was done by utilizing the count if function.

Total	36275
Not Canceled	24390
Canceled	11885

(Created by Student)

In order for better data handling the three mentioned columns of data arrival year, average price and booking status will be extrapolated from the main sheet to another sheet that will be renamed to data set. Due to the large number of data only the first 22 data were shown as evidence in the following table.

A	B	C
Arrival_year ▾	Average price ▾	Booking status ▾
2017	65	Not_Canceled
2018	106.68	Not_Canceled
2018	60	Canceled
2018	100	Canceled
2018	94.5	Canceled
2018	115	Canceled
2017	107.55	Not_Canceled
2018	105.61	Not_Canceled
2018	96.9	Not_Canceled
2018	133.44	Not_Canceled
2018	85.03	Not_Canceled
2018	140.4	Not_Canceled
2018	88	Canceled
2018	90	Canceled
2019	94.5	Not_Canceled
2018	115	Canceled
2017	96	Not_Canceled
2017	96	Not_Canceled
2017	65	Canceled
2019	72	Not_Canceled
2017	65	Canceled
2019	67	Not_Canceled
2018	85	Not_Canceled
2018	105	Not_Canceled
2018	94.5	Not_Canceled
2018	114.3	Not_Canceled
2017	65	Not_Canceled

(Created by Student)

The first step on processing and analyzing the data would be to use filtering on the three columns of arrival year, average price and booking status for father ease of management and finding out the cancelled and not cancelled status of booking then by adding 2017, 2018, 2019, 2020 and 2021 in a different table with Their own columns and adding the average price of those corresponding years we have found the following table that details the average price of that year of the hotel reservation that weren't cancelled;

Arrival Year	Total
2017	466159.06
2018	1829216.13
2019	11374.19
2020	13334.55
2021	14023.05
2022	103220.15
Grand Total	2437327.13

(Created by Student)

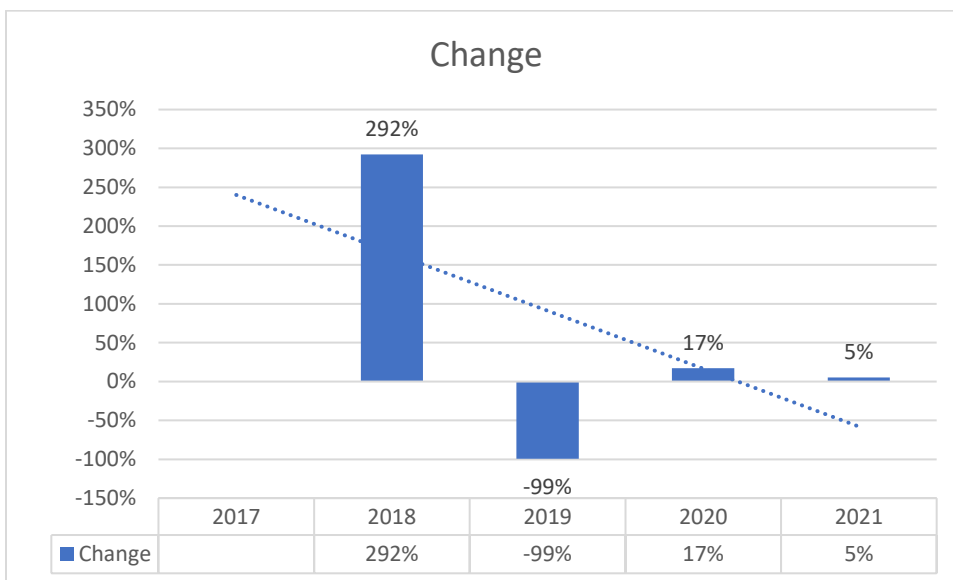
In 2017 the total average price was 466159.06; 2018 it was 1829216.13 a major increase from 2017 onwards but in 2019, 2020 and 2021; the table shows that there has been a drop in average pricing and the business is on a downhill trajectory, but the business has picked up in 2022.

Arrival Year	Total	Change
2017	466159.06	
2018	1829216.13	292%
2019	11374.19	-99%
2020	13334.55	17%
2021	14023.05	5%

(Created by Student)

Graph

In the above table we can see that in 2018 the business saw a meteoric rise of 292% in their operations but in 2019 they saw a fall off 99% this can be due to the covid-19 and the table shows that the business picked up average pricing in 2020; a 17% increase and in 2021 that increase has been slowed down to 5%.



(Created by Student)

Pivot Table

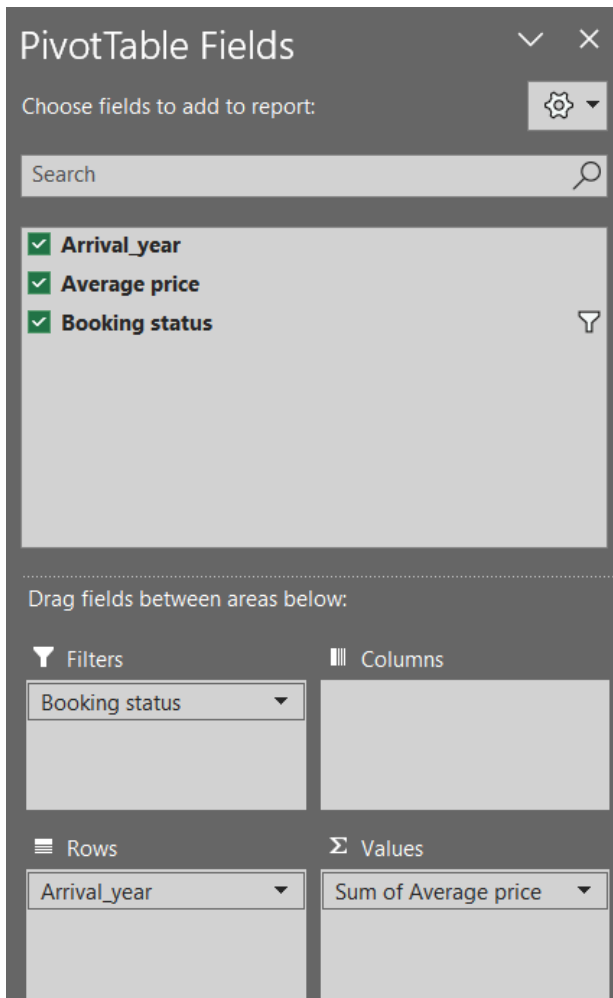
The importance of the people table cannot be understated as it is a powerful tool for data analysis that allows for the summarization and organization of large amounts of data in a compact, flexible and customizable way. pivot table allows for the simplification of data analysis. With the help of pivot table, it is easier to summarize and visualize substantial amounts of data (Wegner, 2010). By utilizing the pivot table users can transform raw data into meaningful insights or make informed decisions upon those data. It allows for users to quickly and efficiently export data and identify patterns trends or any other anomalies in the data in; one of the key strengths of the pivot table is its flexibility as the pivot table is highly customizable and can show different variables to display and how they want the data to be summarized or interconnected (Wegner, 2010). But the major importance of the pivot table is that it's time saving as it eliminates the need for manual sorting and filtration of the data and without pivot tables users cannot easily analyze data and make decisions in a fraction of the time that would take to manually analyze it (Wegner, 2010).

In this report it is easier to identify the relation between the increase and decrease of average price by the use of a pivot table; as shown in the table below:

Booking status	Not Canceled/ Canceled
Sum of Average price	
Arrival Year	Total
2017	466159.06
2018	1829216.13
2019	11374.19
2020	13334.55
2021	14023.05
2022	103220.15
Grand Total	2437327.13

(Created by Student)

This is the pivot table of the report. It was based on the earlier mentioned extrapolated data of Arrival Year, Avg Price and Booking status. By using the built-in pivot table, it was easy to find and summarize the data. These are the PivotTable Fields; the rows the arrival year and the values are the sum of average price of the year.



(Created by Student)

Sum if

The Sum IF function of excel is a particularly important and efficient tool for summarizing the values in order to get a more comprehensive and easier to manage calculation as well as making the calculations more simplified. This function simplifies calculations by allowing the summation of values that meet specific criteria without having to manually add each value this automize the function (Wegner, 2010). This allows for greater flexibility and automates the reporting process as it combined with other functions such as the if statements the summit function can automate the reporting process and allowing for quicker generation of summaries and insights in large amounts of data (Wegner, 2010).

Year	Total Avg Price	Cancelled Avg Price	Final Avg Price
2017	549304.11	83145.05	466159.06
2018	2984721.57	1155505.44	1829216.13
2019	18015.29	6641.10	11374.19
2020	20857.3	7522.75	13334.55
2021	19436.8	5413.75	14023.05
2022	159353.81	56133.66	103220.15

(Created by Student)

In this report to find the total average price of each year; First a table was constructed that consisted of the year, total average price, cancelled average price then the final average price. Then pick column was consistent in the year section from 2017 to 2022. In the total average price column of 2017; the first function of sum ifs was used; the first element of the formula was the some range that was the average price column update excel then it was the criteria range which was the years in the excel and lastly the criteria one which was the 2017 of the above table;

The function is; =SUMIFS(B2:B36275,A2:A36275,E28)

Thus the total average price was formulated and applied to the rest of the years. As we add the final average price that excluded the cancelled average price from the people table it was placed on the table then subtracted from the total average price of the subsequent row and the result was place on the canceled the average price as the final average price was calculated without the values from the cancelled average price. This was carried out for the rest of the years too; ergo we can see the total price increase and decrease over the years.

Part 2

SPSS

SPSS or statistical package for social sciences is a tool that is used for statistical analysis and is widely considered by many researchers as one of the prominent tools for analysis of data; SPSS allows for more efficient data management as it allows the user to manage and organize their data more easily and it has various tools that enable that such as data entry, cleaning and manipulation (Rovai et al, 2013). SPSS has various tools for greater visualization of data such as using charts, graphs and histograms and it allows for more streamlined statistical analysis as it is specifically designed for that function and has a wide range of statistical procedures that can be utilized to analyze any form of data no matter the size (Rovai et al, 2013); It uses various to such as descriptive statistics, correlation

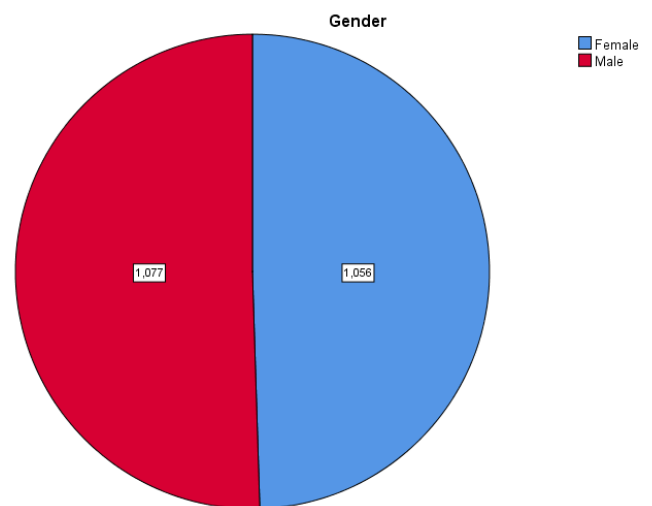
analysis, regression analysis etc. to conclude spaces is a great tool for statistical analysis as it allows the user to manage, analyze and visualize the data more efficiently (Rovai et al, 2013).

Number of Male & Female Students

		Gender			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	1056	49.5	49.5	49.5
	Male	1077	50.5	50.5	100.0
	Total	2133	100.0	100.0	

(Created by Student)

As shown in the above table the total number of students is 2133 out of those 2133 students 1056 or 49.5% are female and 1077 or 50.5% are male.



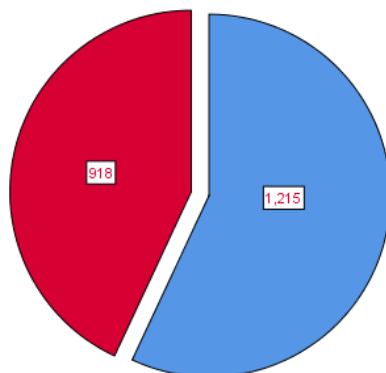
Number of Students Qualified for reduced/free lunch.

		Lunch			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Does not qualify	1215	57.0	57.0	57.0
	Qualifies for reduced/free lunch	918	43.0	43.0	100.0
	Total	2133	100.0	100.0	

(Created by Student)

■ Does not qualify
■ Qualifies for reduced/free lunch

The development table shows that out of 2133 students 1215 or 57 % do not qualify for the reduced/free lunch program and 918 or 43% are qualified for the reduced/ free lunch.



(Created by Student)

Mean and median of students who qualifies and does not qualify for reduced/free lunch.

Statistics

		Pretest	Posttest
N	Valid	2133	2133
	Missing	0	0
Mean		54.96	67.10
Median		56.00	68.00

(Created by Student)

The mean and median differences of student's pre-test and post-test scores

Case Processing Summary

	Included		Cases Excluded		Total	
	N	Percent	N	Percent	N	Percent
Pretest	2133	100.0%	0	0.0%	2133	100.0%
Posttest	2133	100.0%	0	0.0%	2133	100.0%

(Created by Student)

Report

	Pretest	Posttest
Mean	54.96	67.10
Median	56.00	68.00

(Created by Student)

Clustering

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Student No	2133	14	31	22.80	4.229
Pretest	2133	22	93	54.96	13.563
Posttest	2133	32	100	67.10	13.987
Valid N (listwise)	2133				

(Created by Student)

Initial Cluster Centers

	Cluster	
	1	2
Zscore (Student No)	1.23046	-1.37070
Zscore (Pretest)	-2.35609	2.80497
Zscore (Posttest)	-2.29518	2.35206

(Created by Student)

Final Cluster Centers

	Cluster	
	1	2
Z score (Student No)	.65914	-.62249
Zscore (Pretest)	-.81788	.77240
Zscore (Posttest)	-.82717	.78117

(Created by Student)

Iteration History^a

	Change in Cluster Centers	
Iteration	1	2
1	2.307	2.569
2	.061	.054
3	.024	.023
4	.015	.014
5	.011	.011
6	.003	.003
7	.003	.002
8	.004	.004
9	.001	.001
10	.001	.001

(Created by Student)

Number of Cases in each Cluster

Cluster	1	2
1	1036.000	1097.000
2	1097.000	1036.000
Valid	2133.000	
Missing		.000

(Created by Student)

Five Common Data mining and text mining methods for contemporary business

Data mining and text mining are both critical tools that are utilized for the success of contemporary businesses as they have the ability to give inside from large volumes of data and text which leads to a more informed decision-making process and they as leading to more operational efficiency and

increased competitive advantage by being more streamlined (Sumathy et al, 2013). These tools help the business be more data-driven on their decision-making process as data mining allows business to analyze and extract patterns, trends and relationships from large data. By utilizing the tools businesses can identify potential opportunities and streamline their processes as well as uncovering hidden insights. Data mining and text mining lead businesses to be more efficient in their operations by finding out various inefficiencies, redundancy and bottlenecks that are hampering the business process (Sumathy et al, 2013).. This will give the business a more competitive edge over their competition as it allows to them be more prepared and innovative during change. Here are five methods that are used in data mining and text mining;

1. **Clustering:** This is a method that is used in data mining to group similar data points together on the basis of their similarities or even sometimes dissimilarities. In today's business world clustering can be used for important business activities such as customer segmentation, market segmentation, product differentiation, fraud detection etc. To identify the various patterns or trends in data and improve the decision making as well as streamlining the strategic planning clustering is highly important in contemporary businesses (Giordani et al, 2020).
2. **Classification:** this method is used in data mining as well as text mining 2 categorize various data points in different categories or classes. In contemporary businesses classification place various rules and use utilized for different analysis such as sentiment analysis, spam detection, credit risk assessment and product categorization (Bhardwaj and Pal, 2012); This tool enables businesses to be more automated and have a more streamlined operational efficiency. This tool can be used to automate the operations by categorizing and tagging large volumes of text data; making it more easier to manage and operate as well as other functions such as organize, search and analyze with the combination of tools this will lead to a more streamlined efficient business process (Bhardwaj and Pal, 2012).
3. **Association Rule Mining:** Another tool that can be utilized for data mining is association rule mining that helps to discover various patterns or associations between different variables in a data set. In modern businesses there are various variables and association rule mining can help on finding out various patterns and help on analysis by utilizing various methods such as Market Basket analysis, recommendation system and cross selling as well as upselling.
4. **Sentiment Analysis:** this method is also known as opinion mining is a text mining method that involves the process of systematically determining the sentiment or emotional tone of a piece of task this is a very important element on business analysis as it helps on finding out the emotional

tone of a piece of text such as customer reviews, social media post or customer feedback which can help on marketing and giving the firm a competitive edge over their competition (Sukanya and Biruntha, 2012).

5. **Text Classification:** this helps to automatically categorize or tag text documents into a more predefined class in large datasets. This can be used for document categorization, content recommendation and even to an extent of spam filtering; This can be used in tangent with sentiment analysis for father impact on business operation as it helps on businesses efficiently organize and manage large volumes of text data (Sukanya and Biruntha, 2012).

Advantages and Disadvantage of using SPSS and Excel

Here are some advantages of using SPSS and Excel

SL	Excel	SPSS
1	one of the major advantages of excel is familiarity and ease of use; excel is one of the most widely used spreadsheet software and is more of a common to and familiar to many users; with its ease of use and user friendly UI it may the process of learning aid vary easier compared to other models. As many other spreadsheets integrate various excel systems and commands as well as its UI to their spreadsheet; It makes the transaction period from excel to other spreadsheets easier. It is the one stop software for all of data management excluding specific data analysis.	SPSS is a more specialized software for statistical analysis. SPSS offers a diverse range of complex statistical techniques that includes techniques such as descriptive statistics, regression analysis, factor analysis, anova in many others. It has a wide range of tool set that is used for conducting advanced data analysis and more specific analysis making it a widely regarded software by many researchers etc. this software offers a more advanced statistical techniques such as the structural equation modelling, cluster analysis and multi-level modelling that are not usually found in excel but is very needed for advanced analysis.
2	Excel offers greater data visualization as it offers a variety of chart types that can be easily modified and has charts and graphs for any kind of data visualization. Because of its familiarity and ease of use any user can just simply visualize data with the Press of one button. For example a sales manager can use excel to create a bar chart to compare sales performance.	SPSS offers a more streamlined and complex data management capabilities that allows the users to easily clean out redundancies and transform data; it allows the users to filter, transform and even merge data and it has a user-friendly interface for data preparation tasks. For example and analyzed working with a large number of customer survey data may use SPSS to filter out incomplete response or redundancies.
3	Due to the most streamlined approach of excel it has integration with other	just like excel SPSS also provides a high quality output in the form of tables, charts

	Microsoft Office products that allows greater ease of data exchange and presentation; For example a business analyst can use excel to analyze data and create charts as well as import them to PowerPoint presentation for a meaningful presentation.	and graphs that can be easily exported to other software for further analysis.
--	---	--

Here are some disadvantages of SPSS and Excel.

Excel	SPSS
one of the major disadvantages of excel would be its limited statistical analysis capabilities as excel leans towards more a casual and user friendly experience although it has built-in functions for basic statistical analysis but it may not be suitable or be appropriate for various advanced statistical techniques that are commonly used in data analysis. Such as excel does not offer robust tools for advanced regression analysis, multivariate analysis or other data manipulation tools.	Although SPSS offers a more specialized data management tool set but it has a steeper learning curve compared to excel are other spreadsheet software's so we go on especially for users who are not familiar with statistical concepts or techniques it may be hard for them to be adjusted. It also provides a limited flexibility and customization as it's a more specialized tool designed for statistical analysis and functionality may be limited in terms of customization.
Excel also suffers from the same disadvantage of having a high cost but it's usually bundled with various prebuilt PC's and laptops; as it only doesn't offer one tool but also the whole range of Microsoft Office so the cost of individual tools such as excel goes down. Excel has the potential for errors as it's prone to human	The cost of a commercial license for SPSS is high and can deter individual users of small researchers for simple research projects with limited budgets it can prohibit them. For example a student conducting a research on a limiting budget can may find difficult to afford the high cost of SPSS license.

error such as manual data entry or formal errors etc.	
---	--

Conclusion

To conclude this report showcased the use of excel and spss in data management. On the first part of the report, it showcased how excel can be used on fighting out a specific and simple data by utilizing various tools such as the pivot table or sum if or by just manually filtering out the information and finding out but that is prone to error. How the data can be easily visualized by utilizing the various built in graphs ().

The second part of the report consisted of the use of css on finding out the number of male and female students as well as the number of students qualified for reduced launch or the number of students who don't qualify for it as well as the mean and median differences it also showcases the use of clustering. Five common data mining and text mining methods for contemporary businesses were discussed as well as the advantages and disadvantages of using spss and excel in contemporary business using relevant examples.

Reference

- Katal, A., Wazid, M. and Goudar, R.H., 2013, August. Big data: issues, challenges, tools and good practices. In 2013 Sixth international conference on contemporary computing (IC3) (pp. 404-409). IEEE.
- García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F., 2016. Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1), pp.1-22.
- Wegner, T., 2010. *Applied business statistics: Methods and Excel-based applications*. Juta and Company Ltd.
- Rovai, A.P., Baker, J.D. and Ponton, M.K., 2013. *Social science research design and statistics: A practitioner's guide to research methods and IBM SPSS*. Watertree Press LLC.
- Sumathy, K.L. and Chidambaram, M., 2013. Text mining: concepts, applications, tools and issues-an overview. *International Journal of Computer Applications*, 80(4).
- Sukanya, M. and Biruntha, S., 2012, August. Techniques on text mining. In 2012 IEEE international conference on advanced communication control and computing technologies (icaccct) (pp. 269-271). IEEE.
- Bhardwaj, B.K. and Pal, S., 2012. Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.
- Giordani, P., Ferraro, M.B., Martella, F., Giordani, P., Ferraro, M.B. and Martella, F., 2020. *Introduction to Clustering* (pp. 3-5). Springer Singapore.