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

BA (Hons) Business Management

Data Handling and Business Intelligence- LCBB5000

Data Analysis and Visualisation

Session: February 2023 Term

by Constantin Gabriel Tacea (2122369)

Module Coordinator: Amitave Banik

Contents

INTRODUCTION **3**

PART 1 ..... **3**

PIVOT TABLE 3

IF AND SUMIF 8

PART 2 ..... **11**

1. Male and Female students 11
2. Qualify/not qualify for reduced/free lunch 12
3. Mean and Median 13
4. Pretest and Posttest scores 15
5. K-MEANS CLUSTERING 15
6. Data mining methods and text mining methods 18
7. Advantages and disadvantage of using SPSS and Excel 19

CONCLUSION: **20**

REFERENCES **21**

INTRODUCTION

The economic-financial analysis has a permanent character and is a means to achieve the business objectives, so the analysis must offer solutions to be able to make decisions. The economic analysis contributes to achieving and exceeding profitability and ensuring financial balance throughout the year.

"There are simple end-user data mining tools, such as MS Excel, and there are more sophisticated tools, such as IBM SPSS Modeler… MS Excel is a relatively simple and easy data mining tool. It can get quite versatile once analyst pack and some other add-on products are installed on it. IBM's SPSS Modeler is an industry-leading data mining platform. It offers a powerful set of tools and algorithms for most popular data mining capabilities. It has colorful GUI format with drag-and-drop capabilities. It can accept data in multiple formats, including reading Excel files directly. " (Maheshwari, 2015, p.54)

The study analyzes with the help of Microsoft Excel functions the fluctuation of hotel booking prices and determines one of the possible causes of the decrease in the average price of hotel room bookings, which could be easily combated.

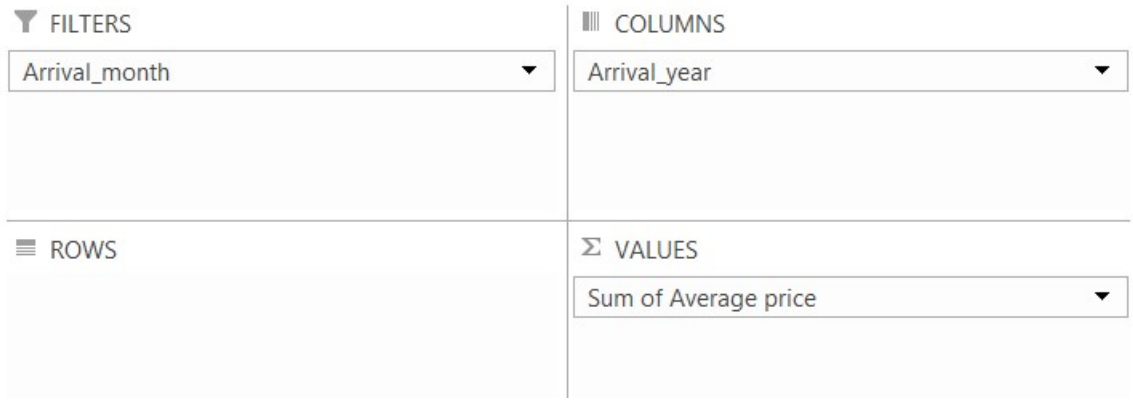
In the second part, SPSS will analyze a survey made on a sample of 2133 students. The database variables will be defined and the data will be verified and corrected. At the same time, the importance of data mining and text mining methods in a practical way will be discussed and the pros and cons of using Microsoft Excel and SPSS will be discussed.

PART 1

As confirmed by Anderson (2020), Microsoft Excel offers logical functions for data processing and statistics. It is important to know the basic functions you need in the analysis and how they can be used. At the same time, in the Microsoft Excel Support online program, each function and their use are explained in detail.

PIVOT TABLE

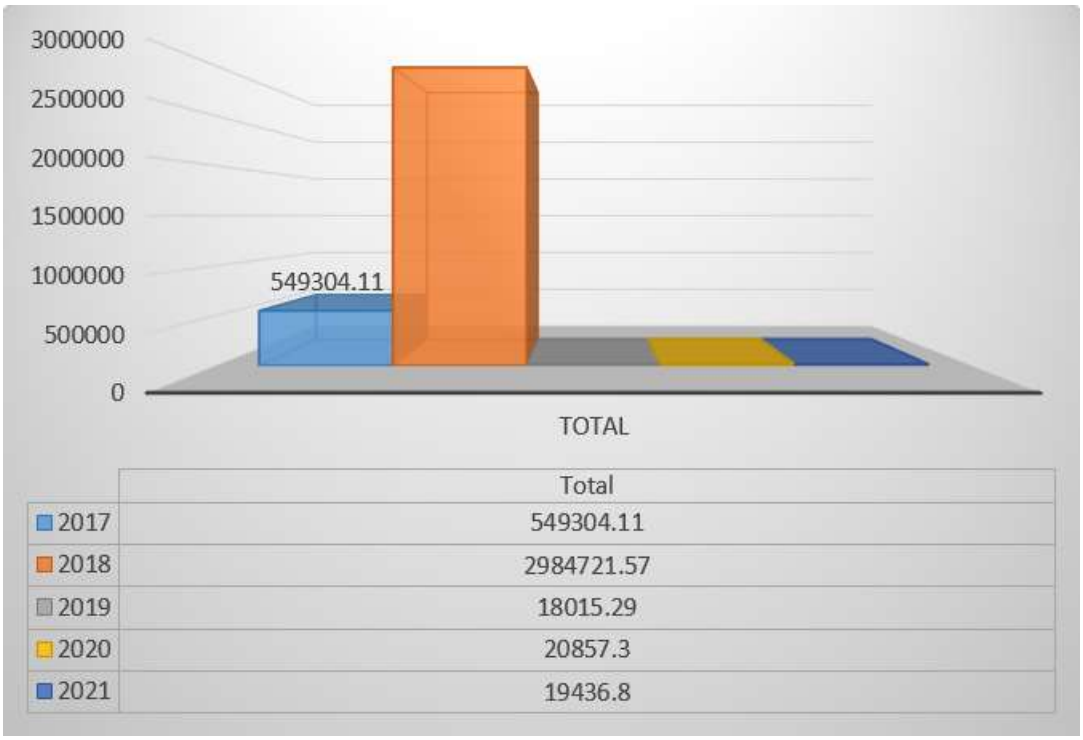
Microsoft Excel offers the chance to structure, organize and operate a large series of data through mathematical functions, the analyst being able to analyze and statistically calculate economic and financial phenomena, based on which decisions can be made. Graphs and diagrams provide an overview of the collected data. Pivot Table extracts and calculates large databases, providing a summary of numerous queries. “The analytical features include basic statistical and financial functions. Pivot tables help do sophisticated what-if analysis.” (Maheshwari, 2015, p.25)



Arrival_month	(All)						
Sum of Average price	Arrival_year						
	2017	2018	2019	2020	2021	2022	Grand Total
Total	549304.11	2984721.57	18015.29	20857.3	19436.8	159353.81	3751688.88

TABLE 1: The evolution of average prices from 2017 to 2021

Source: Self-report, 2023



	Total
2017	549304.11
2018	2984721.57
2019	18015.29
2020	20857.3
2021	19436.8

Figure 1: The evolution of average prices from 2017 to 2021

Source: Self-report, 2023

Major differences at room reservation price. What happened?

As can be seen in PIVOT TABLE and Chart, the room reservation price fluctuates. There can be several reasons for it fenomen.

As Pivot Table 1 shows, in 2018 the amount of the average price was approximately 3 million Lire, and in 2017 it was approximately 0.55 million Lire, but in 2019 there is a significant price collapse, and in the following years the decrease continued, reaching 19437 Lire in 2021 (in 2019 the amount of the average price was 18015 Lire and in 2020 20857 Lire). The graph shows a representative visual scheme and the shock of the decrease is noticeable.

We can easily notice according to the figures that the Hotel had some popularity in the years 2017-2018, which collapsed starting with 2019. We used the Pivot Table to notice starting with which month the prices dropped and the surprising thing is that the beginning of 2019 was devastating (an average price of 486 Lire was recorded, although December 2018 was not a bad month, an average price of approximately 188 thousand Lire is noted.

Sum of Average price	Arrival_year	
Arrival_month	2019	Grand Total
1	486.67	486.67
2	855.21	855.21
3	1106.86	1106.86
4	1449.36	1449.36
5	1765.03	1765.03
6	2325.76	2325.76
7	1250.56	1250.56
8	1827.1	1827.1
9	2865.51	2865.51
10	1926.66	1926.66
11	998.37	998.37
12	1158.2	1158.2
Grand Total	18015.29	18015.29

TABLE 2: The evolution of average prices from 2019

Source: Self-report, 2023

Sum of Average price	Arrival_year	
Arrival_month	2018	Grand Total
1	71942.71	71942.71
2	128357.19	128357.19
3	204108.2	204108.2
4	260986.26	260986.26
5	276262.75	276262.75
6	337397.81	337397.81
7	279893.43	279893.43
8	320068.43	320068.43
9	344978	344978
10	361613.51	361613.51
11	211629.25	211629.25
12	187484.03	187484.03
Grand Total	2984721.57	2984721.57

TABLE 3: The evolution of average prices from 2019

Source: Self-report, 2023

At the same time, according to the collected data, the Hotel's activity started in July 2017 (see Table 3) and even so the average total income is much higher than in the years preceding 2018.

The causes can be multiple, including: attention to marketing campaigns, the market, the employees and the management of the hotel and/or the quality services.

Sum of Average price	Arrival_year					
Arrival_month	2017	2018	2019	2020	2021	Grand Total
1		71942.71	486.67	230.5	151	72810.88
2		128357.19	855.21	975.65	641.57	130829.62
3		204108.2	1106.86	1164.18	998.79	207378.03
4		260986.26	1449.36	1477.97	2047.77	265961.36
5		276262.75	1765.03	1967.57	1408.38	281403.73
6		337397.81	2325.76	1990.87	1406.67	343121.11
7	27460.43	279893.43	1250.56	1749.08	2010.29	312363.79
8	87335.51	320068.43	1827.1	2140.16	2409.84	413781.04
9	157860.3	344978	2865.51	3257.64	3109.89	512071.34
10	166510.21	361613.51	1926.66	2795.81	2588.02	535434.21
11	44055.68	211629.25	998.37	1644.04	1343.13	259670.47
12	66081.98	187484.03	1158.2	1463.83	1321.45	257509.49
Grand Total	549304.11	2984721.57	18015.29	20857.3	19436.8	3592335.07

TABLE 4: The evolution of average prices in the period 2017-2019 and the months of the year

Source: Self-report, 2023

The increase in popularity, along with the increase in the average price, as well as the increase in income are the reasons why the hotel should have made a profit and an improvement in service, followed by a greater increase in popularity and prices, but 2019 brings repeated decreases. If the Hotel has kept the same strategy of 2018, among the causes of failure may be the irrelevant decrease or increase of the price, the competition and change of customer preferences and at the same time we can note the carelessness of the management who did not foresee the change in the market and failed to make the right decisions. One of the external causes that could have caused the price drop in 2020 could be the Pandemic with COVID-19. At the same time, another reason for the drop in prices can be caused by the methods of booking the rooms and the attention that the management and marketing specialists paid to the customers' requests, which we will analyze below with the help of the IF functions.

The valuable information about the moments of price increase, the good months, the glory years, as well as the number of guests and their incomes, which are highlighted with the help of the Microsoft Excel-Pivot Table function, are sources that can make the difference and help in making good decisions for increasing the performance and increasing the popularity of the hotel.

IF AND SUMIF

According to Microsoft Support, the IF function offers the opportunity to logically and effectively compare a value with a condition, which can be true or false. IF is one of the most important functions and can easily connect with other Microsoft Excel functions.

Considering the collapse of the average price of booked rooms starting with 2019, the research continues by studying the hotel's adaptability regarding the booking methods chosen by customers.

In the first phase, we considered that only one adult made the decision to make the booking for the hotel room and we used the IF function to determine 1 adult, regardless of how many adults were accommodated.

=IF(B2=1, 1,1)

	B	C	D	E
ıg ID	Adults No	an adult reserved the room	Children No	Weeke Night
01	2	1	0	1
02	2	1	0	2
03	1	1	0	2
04	2	1	0	0
05	2	1	0	1
06	2	1	0	0
07	2	1	0	1
08	2	1	0	1
09	3	1	0	0
10	2	1	0	0
11	1	1	0	1
12	1	1	0	2
13	2	1	0	2
14	1	1	0	2
15	2	1	0	0
16	2	1	0	0
17	1	1	0	1
18	2	1	0	1

TABLE 5: IF function to determine an adult who reserved the room

Source: Self-report, 2023

Through the SUM function, we determined that 36275 bookings were made during the studied years 2017-2021.

=SUM(C2:C36276)

B	C	D
2	1	0
2	1	0
2	1	0
2	1	0
TOTAL	36275	

TABLE 6: SUM function to determine how many adults reserved the room

Source: Self-report, 2023

In order to determine the type of booking chosen in the years that we are analyzing, namely 2017-2021, we used the SUMIFS function, thus we determined that the most practiced booking method is the online one, 23214 customers chose this method, the offline booking being half of that, i.e. 10528, and 2533 bookings were made through other methods (Corporate/ Complementary Methods).

=SUMIFS(C2:C36276,N2:N36276,"Online")

B	C	N	X	Y
Adults No	**an adult reserved the room**	**Market_type**	**Market type**	**The number of reservations**
2	1	Offline	ONLINE	23214
2	1	Online	OFFLINE	10528
1	1	Online	TOTAL ON+OFF	33742
2	1	Online	OTHER METHODS	2533
2	1	Online		
2	1	Online		
2	1	Online		
2	1	Online		
3	1	Offline		
2	1	Online		
1	1	Online		
1	1	Online		
2	1	Online		

TABLE 7: SUMIFS function

Source: Self-report, 2023

Considering the sharp increase in online bookings, it can be seen that not enough marketing has been done and no effective methods have been found to attract customers to make bookings. If the hotel management had focused on advertising, collaboration with applications and companies that deal with online bookings and had provided customers with an easy-to-use system on the website to be able to make bookings directly, the average price would not take a downward direction after 2018. The implementation of effective technological methods can help increase the average price of the booking, by accommodating more customers, i.e. by increasing the number of bookings,

The financial performance of the hotel can be marked by a new vision towards the online environment and thus the barriers can be overcome.

The application of the two Microsoft Excel, IF and Pivot Table functions helps the analyst to identify the price variation in the hotel booking process during 2017-2021. It identifies the sudden price drop in 2019, followed by the same weak course in the following years and detects one of the causes that can be fixed, the room booking method.

PART 2

Hinton (2004) say "SPSS is a wonderful statistical analysis package...SPSS allows us to undertake a wide range of statistical analyses relatively easily. While this is extremely useful it does mean that we do need to know what analysis is appropriate to the data we have. So a certain amount of basic statistical knowledge is required before using SPSS." (Hinton, p.1)

The Microsoft Excel Student_Score data were imported into the SPSS program.

1. Male and Female students

It was determined that 1056 of the students are women and 1077 of the students are men.

The following SPSS Dialog tools were used: from Descriptive Analysis and then Frequency and then Gender was chosen and therefore we managed to determine how many students are men and how many students are women

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	

TABLE 7: How many students are male and female?

Source: Self-report, 2023

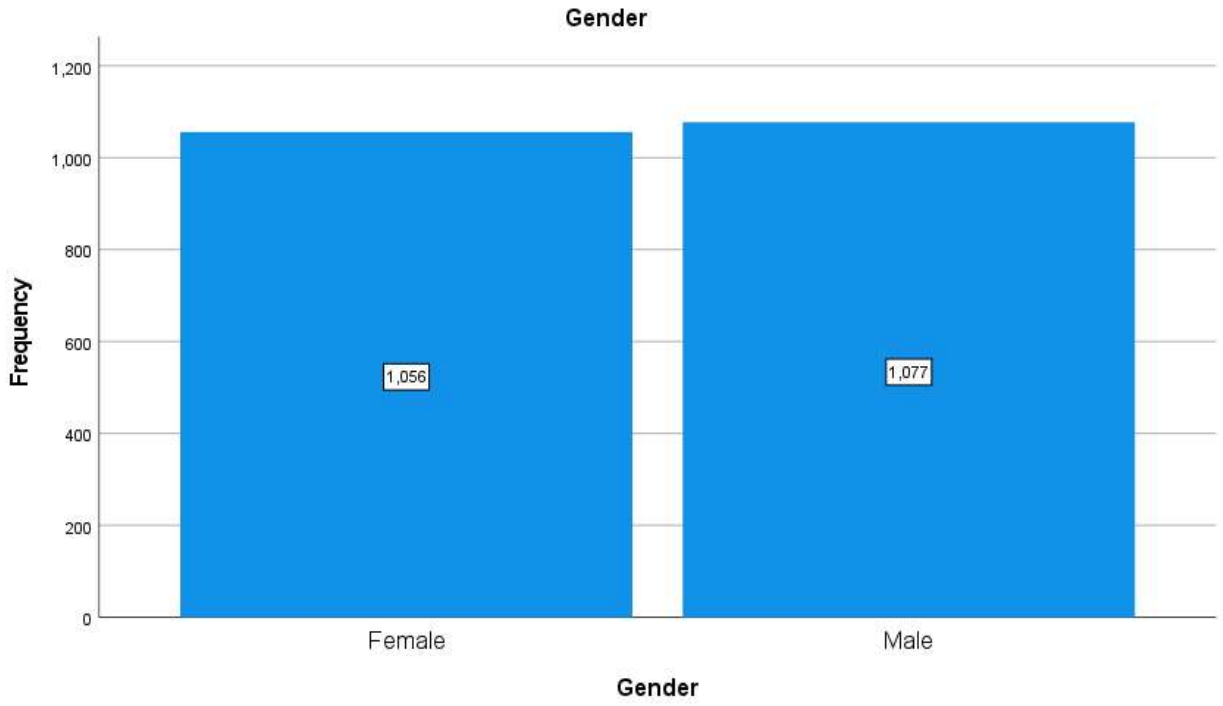


FIGURE 2: How many students are male and female?

Source: Self-report, 2023

2. Qualify/not qualify for reduced/free lunch

With the help of the same dialogue tools, it was possible to determine that 918 students qualify for the reduced/free lunch and 1215 students do not qualify. You can see the details in table no Y and in the diagram.

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	

TABLE 8: How many student qualifies and does not qualify for reduced/free lunch?

Source: Self-report, 2023

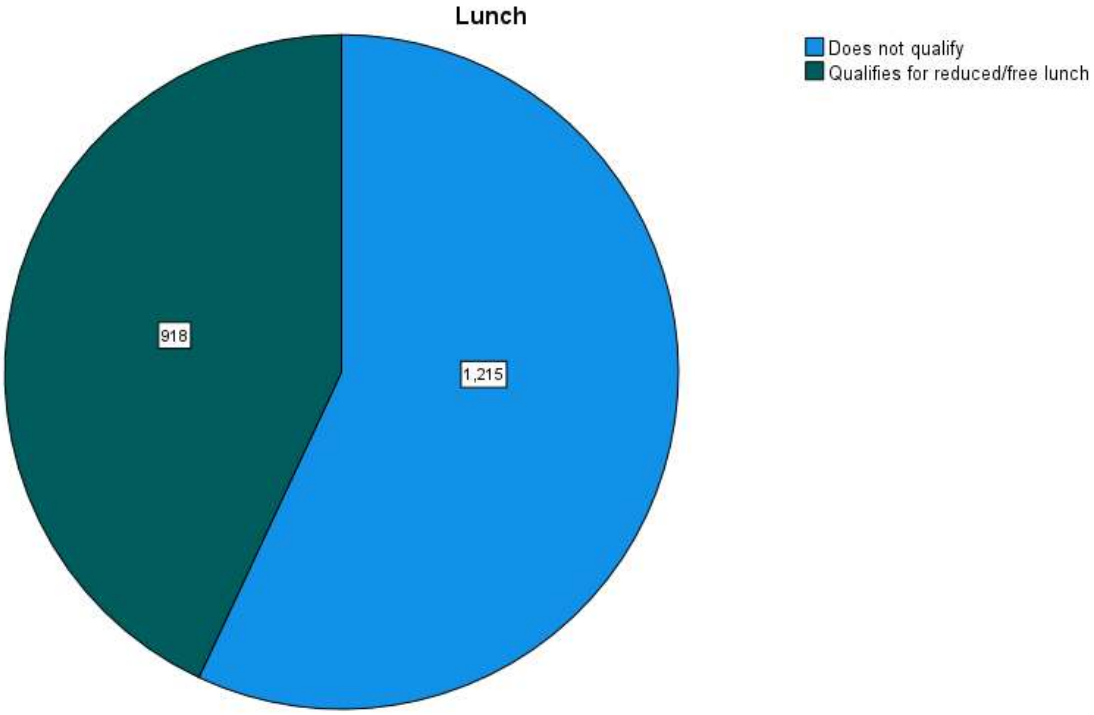
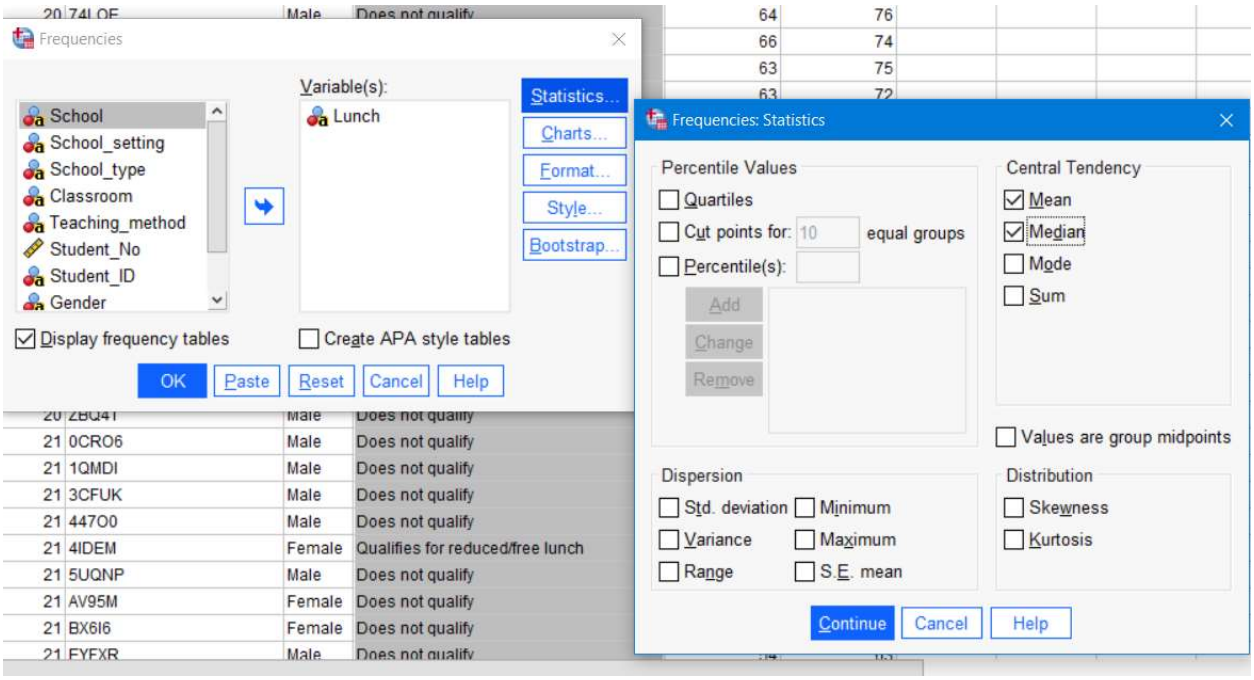


FIGURE 3: How many student qualifies and does not qualify for reduced/free lunch?

Source: Self-report, 2023

3. Mean and Median

Following the results obtained, the mean and median of the students who qualify and who do not qualify for the reduced/free lunch were calculated.



PICTURE 1: Ferestrele de Dialog Frenquencies for Mean and Median

Source: Self-report, 2023

Statistics

N	Valid	2133
	Missing	0
Mean		1.43
Median		1.00
Mode		1
Std. Deviation		.495
Sum		3051

TABLE 9: Mean and median of students who qualifies and does not qualify for reduced/free lunch

Source: Self-report, 2023

In table 9 it can be seen that the average is 1.43, and the median is 1.

4. Pretest and Posttest scores

Following the analysis, the following Pre-test and Post-test differences of the students were found:

In the case of the Pretest the average is 54.96 and in the Posttest is 67.10, and the Pretest median was 56 and the Posttest one was 68. The higher Posttest scores are obvious. Data can be seen in the table below:

DIFFERENCES

	Pretest	Posttest
Mean	54.96	67.10
N	2133	2133
Std. Deviation	13.563	13.987
Median	56.00	68.00

TABLE 10: Mean and median differences of students pre-test and post-test scores

Source: Self-report, 2023

5. K-MEANS CLUSTERING

Quick Cluster

Initial Cluster Centers

	Cluster	
	1	2
Pretest	23	93
Posttest	35	100

Cluster Membership

Case Number	Lunch	Cluster	Distance
1	Does not qualify	2	41.773
2	Does not qualify	2	34.205
3	Does not qualify	2	37.643
4	Does not qualify	2	39.408
5	Does not qualify	2	37.643
6	Does not qualify	2	37.483
7	Does not qualify	2	39.051
8	Does not qualify	2	41.037
2131	Qualifies for reduced/free lunch	1	27.203
2132	Qualifies for reduced/free lunch	1	29.206
2133	Qualifies for reduced/free lunch	1	22.204

Final Cluster Centers

	Cluster	
	1	2
Pretest	45	67
Posttest	56	79

Distances between Final Cluster Centers

Cluster	1	2
1		31.909
2	31.909	

Number of Cases in each Cluster

Cluster	1	1123.000
	2	1010.000
Valid		2133.000
Missing		.000

TABLE 11: K-MEANS CLUSTERING

Source: Self-report, 2023

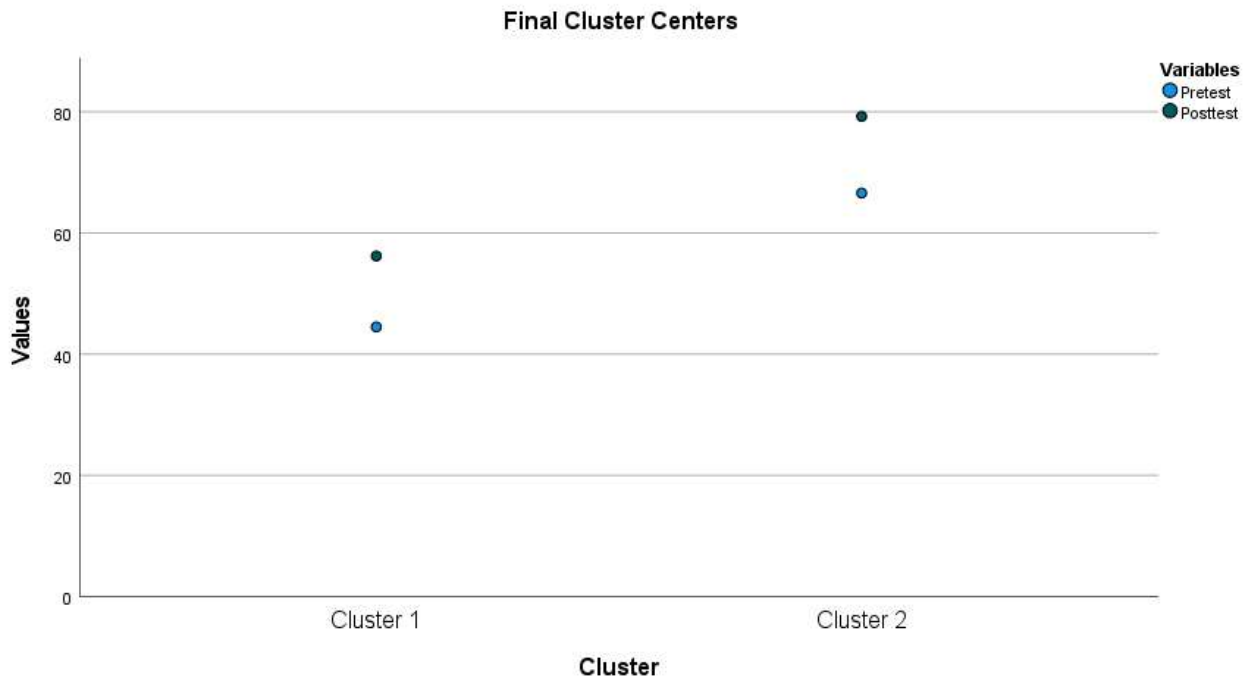


FIGURE 4: K-MEANS CLUSTERING

Source: Self-report, 2023

The K-means Clustering model can also be called Lloyd's Algorithm, because the first proposal was given by Stuart Lloyd in 1957, and then in 1967 James McQueen gave the name "k-means" (Christopher D., 2009).

K-means Clustering generates database efficiency between information and characteristics. For data analysts, K-means Clustering is attractive due to the fast operating data set.

In study of Zˇalik (2007) is describe "each cluster is represented by an adaptively-changing centroid (also called cluster centre), starting from some initial values named seed-points. k-Means computes the squared distances between the inputs (also called input data points) and centroids, and assigns inputs to the nearest centroid." (Zˇalik , p. 1)

Table W shows the data analysis of the pretest and posttest groupings of the students and aims to group the basic elements through the similarities between the data sets. The test was carried out on 2133 students and on their pretest and posttest scores in order to decide whether or not students receive reduced/free meals. Figure X identifies the two groups of clusters, those who receive reduced/free meals and those who do not. The model was efficient and easy to implement with the help of SPSS.

The test did not continue, because the difference is 0 and cannot be further analyzed. Unfortunately, the test is not representative and cannot be used in decision-making.

6. Data mining methods and text mining methods

Considering the earlier experiment, I would like to present “Clustering analysis”. According to Michael C.T. (2018), clustering analysis take into account a data set with n variables and the objective is to group the points into logical groups and this is classified into categories.

In practice, it can represent a help in the field of medicine, because certain states or diseases have states and offer similar sensations, and their grouping can be useful for discovering the subcategories of diseases, for example in the case of depression. At the same time, the method can be useful in the business environment. Customer groups are classified according to preferences/budget/age, etc., and marketing analyzes can be made based on customer groups and categories.

The classification is another essential method, because it can extract data effectively to be able to collect complex and competent information, representing a grouping of information. The method can be used in schools and universities, because a selection of candidates can be made based on the admission tests.

Regression analysis is used to connect at least two variables.

In real life, according to Maheshwari (2014), it can be used in energy companies, being an aid in modeling and predicting electricity consumption, depending on temperature, because regression has as its tool the drawing of a non-linear curve, and the energy consumption can be determined by this non-linear function.

Time series analysis studies chronological series and analyzes data over time. Practically, the method can be used in hotels or restaurants, because over time the demand and preference of consumers can be observed. The information can be used in making layout decisions or for the optimal diversification of menus.

Sentiment analysis: it is based on feelings and negativity or positivity is expressed according to Azevedo (2021). It is most often found in Social Media through reviews, surveys and forums, which are both the basis of practical use and the basis of method training. Any field of activity that receives a review resorts to Sentiment analysis.

Topic modeling: this method identifies topics and different information in several typologies/documents/websites through text. It is a method that is perfectly modeled with the

trend of the field of recruitment. As specified in David J. L (2021) in the published article, it is a useful strategy to observe what the tendency of the staff is regarding the additional hours through keyword analysis, regardless of the field.

Association rule mining: used to identify the relationships of a group of data, it can be a method put into practice by shopping centers, determining which products are bought together by customers (example: shampoo + hair conditioner) - through this specific information a strategy for placing goods according to the products purchased most often together can be thought of.

Text summarization is a useful method to emphasize the main ideas and to provide a shorter presentation by marking the essential ideas of a long text. This practice is most often encountered by television channels, especially news stations, when they have to briefly present the new events of the day or when communicating with the public online.

7. Advantages and disadvantage of using SPSS and Excel

Microsoft Excel and SPSS are data operating systems that offer the possibility of managing and operating information and have the necessary tools for data analysis and statistical analysis.

Microsoft Excel through collaborations with Google Drive and Google Sheet can create databases and management systems, maintenance, programming and organization through useful tools such as: vlookup, importrange, filter - these being just some of the ways to maintain a system. At the same time, data can be calculated and analyzed through Microsoft Excel.

Instead, SPSS is intended more for solid companies that want to operate a large stock of data in a short time. It is intended more for statistical analysis, not for maintaining data stocks.

	Advantages	Disadvantage
Microsoft Excel	- The collected data are easy to enter and manage - Microsoft Excel through Google sheet helps maintain databases, create files, express data, calculate variables, import information - you can create an	- Difficult operation in OneDrive - Limited control - When managing large stocks, Microsoft Excel handles the data with difficulty. - Lack of security - Loss of information - Difficulties in saving files

	entire operating and management site - Storing and analyzing large data stocks - Specific and logical formulas - Access to quick buttons (direct printing) - Increased popularity and an advantage in front of employers	- Useful for storing small business information, for large companies it is not useful, since CRM is practiced
SPSS	- Analyzes large data stocks - Data calculation speed - Performance in the data extraction and calculation process - A statistical system that offers direct analysis on the chosen sample according to the variables - It contains calculations and manipulates data with ease - SPSS is more of a system for analysis, not for data management like Microsoft Excel - Reduced analysis times - Intuitive system - Knowledge of SPSS operation is an advantage	- Requires security and confidentiality measures - Expensive licenses and guides purchased from IBM are required - Data import in SPSS is more difficult than in Excel - It is not intended for data storage, but is an operating system and data analysis. - It is not as well-known and used - There are not as many SPSS experts as Excel connoisseurs

CONCLUSION:

The presented report follows several directions focusing on Data Handling and Business Intelligence.

In the first stage, the importance of Microsoft Excel was presented and the value of financial economic analyzes was demonstrated by using excel functions, and the basis of the study was the situation of a hotel regarding room bookings. Worrying fluctuations were noted and it was found what were the reasons for the increase/decrease in prices when booking hotel rooms.

Pivot Table was used to determine the fluctuation, and IF functions were used to combat the situation.

In the second part of the study, the number of students of a University according to gender and their right to obtain reduced/free meals was quickly determined with the help of SPSS query windows.

At the same time, the importance and relevance of K-Means Clustering was discussed, the study being based on the same information provided by the University report. Next, it was brought forward through factual and real examples 5 most common data mining and text mining methods for each that can be used in the contemporary business and advantages and disadvantage of using SPSS and Microsoft Excel .

REFERENCES

Ana Azevedo, 2021, Integration Challenges for Analytics, Business Intelligence, and Data

Mining, Available at: https://www-igi-global-com.ezproxy.uwtsd.ac.uk/gateway/book/253124

[Accessed: 28 April 2022]

Anderson, D.R., Sweeney, D.J., Williams, T.A., Camm, J.D. and Cochran, J.J., (2020). Modern business statistics with Microsoft Excel. Cengage Learning, Available at: https://books.google.ro/books?hl=en&lr=&id=LwfFDwAAQBAJ&oi=fnd&pg=PP1&dq=Anderson,+D.R.,+Sweeney,+D.J.,+Williams,+T.A.,+Camm,+J.D.+and+Cochran,+J.J.,+(2020).+Modern+business+statistics+with+Microsoft+Excel.+Cengage+Learning&ots=S5CR4vl6-Q&sig=BpZ9JyhQR9uj2Ni7ZpXRx1y6oiM&redir_esc=y#v=onepage&q&f=false [Accessed: 29 April 2022]

Anil K. Maheshwari, 2015, Business Intelligence and Data Mining, Available at: https://web-s-ebscohostcom.ezproxy.uwtsd.ac.uk/ehost/detail/detail?vid=0&sid=e25658ec-73b1-4876-af31-

0ad43e28b2f1%40redis&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#AN=935070&db=nlebk

[Accessed: 28 April 2022]

Christopher D. Manning, Prabhakar Raghavan, Hinrich Sch¨utze, 2009, An Introduction to Information Retrieval

DataFlair, 2021, What is Text Mining in Data Mining – Process & Applications, Available at: https://data-flair.training/blogs/text-mining/ [Accessed: 01 May 2022]

Hinton, Perry R., 2004, SPSS explained, Available at: https://www-vlebooks-com.ezproxy.uwtsd.ac.uk/Product/Index/136475?page=0 [Accessed: 29 April 2022]

Michael Christoph Thrun . 2018, Projection-Based Clustering through Self-Organization and Swarm Intelligence Combining Cluster Analysis with the Visualization of High-Dimensional Data, [Accessed: 31 April 2022]

ScienceDirect, An efficient k'-means clustering algorithm, Available at: https://www-sciencedirect-com.ezproxy.uwtsd.ac.uk/science/article/pii/S0167865508000706 [Accessed: 2 May 2022]

ScienceDirect, Comparison of learning analytics and educational data mining: A topic modeling approach, Available at: https://www.sciencedirect.com/science/article/pii/S2666920X21000102 [Accessed: 1 May 2022]