

Business Management

Data Handling and Business Intelligence

ASSESSMENT 2: DATA ANALYSIS AND VISUALISATION

RESIT

MARINELA ANDREICA

Student ID:2010777

Module code:LCBB5000

Table of Contents

PART1 3

Determine the increase or decline in the average price of the hotel room 3

Pivot table- 3

PART 2 5

1- How many students are male and female?..... 5

2. How many student qualifies and does not qualify for reduced/free lunch? 7

3. What are the mean and median of students who qualifies and does not qualify for reduced/free lunch? 8

4. What are the mean and median differences of students pre-test and post-test scores? 9

5. Exemplification of clustering 13

Data Mining 14

Text Mining 14

7. Advantages and disadvantage of using SPSS and Excel by use of relevant practical examples 15

Conclussion.....**Error! Bookmark not defined.**

REFERENCES 17

PART1

Determine the increase or decline in the average price of the hotel room.

In this part I will explain how to use different tables in order to find out the results for the questions asked in assessment.

Firstly I will discuss about, **Pivot table-** In databases and spreadsheets, pivot tables are used as a data summary tool to automatically summaries and analyses big datasets. It is a flexible and powerful tool that gives users the ability to quickly turn complicated data sets into tables that are straightforward and simple to understand. (Dierenfeld and Merceron, 2012)

Simple data organized column-by-column is the starting point for a pivot table, which then summarizes the information by grouping the entries into categories and displaying it in a new table. When dealing with huge data sets, where it might be difficult to make sense of the data in its original form, pivot tables are extremely valuable because of the flexibility they provide. Users are able to drag and drop fields from the primary data set into the pivot table layout, which is essential to the usefulness of pivot tables. Users have the ability to select which fields should be displayed in rows and columns, as well as apply a variety of aggregation functions to the data, such as the total, the average, the minimum, and the maximum (Dierenfeld and Merceron, 2012) .

Users are able to rapidly spot trends and patterns, as well as filter and sort the data in a variety of different ways thanks to the pivot table that was produced as a consequence. This table provides a condensed perspective of the data. In the fields of business, finance, and accounting, pivot tables are frequently utilized to analyses sales statistics, budgeting, and financial performance. Additionally, pivot tables are utilized in the fields of research and data analysis to summaries and report on extensive volumes of data.

In general, pivot tables are an effective tool for conducting data analysis because they enable users to acquire insights into enormous datasets in an efficient and easy manner.

Use of the pivot table to determine the change in the average price

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			

(data showing in the excel table) (created by student)

PIVOT TABLE

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

(created by student)

The data related to hotel bookings is displayed in the provided pivot table. More precisely, the table displays the total number of arrivals each year, the total number of adults and children per booking, and the total average price per booking for both cancelled and not-cancelled bookings. The grand total column displays the total number of reservations contained in the dataset (Alexander and Jelen, 2001).

A pivot table is a helpful tool that can be used for the purpose of summarizing and analyzing huge datasets. In this instance, it makes it

possible for us to simply compare the total number of arrivals—adults and children—as well as the average price for cancelled and non-cancelled appointments. The overall number of people in the cancelled bookings is 22,695, and the total number of children is 1,478. This is a significantly lower amount than the total number of adults and children in the bookings that were not cancelled, which is 44,231 and 2,341, respectively. This shows that appointments that are cancelled are more likely to be made by individuals or by smaller groups.

When a reservation is cancelled, the overall average price per booking is \$1,314.36, but when a reservation is not cancelled, the total average price per booking is \$2,437.33. This indicates that the average price of a reservation that was not cancelled is nearly two times higher than the average price of a reservation that was cancelled. This may be the result of a variety of different circumstances, such as the time of year during which the reservation was made, the length of the stay, or the meal plan that was chosen.

Pivot table that shows the groups that has done the reservation in the hotel

	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

(data showing in the excel table)

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

(created by student)

According to the table, there are two groups of hotel reservations: those with one child and those with two children. The total number of children for reservations with just one child is displayed in the left column, while the total number of children for reservations with two children is displayed in the right column.

For example, the Grand Total row in the left column indicates that there were 235,989 reservations with just one child overall. There were a total of 235,989 reservations with 2 children, according to the Grand Total row in the right column.

The grand total for all bookings in the dataset is displayed in the last row, "Grand Total," which is labelled. In this situation, there are 471,978 kids in all.

It's important to note that the table only displays information for reservations with either 1 or 2 children. There could have been reservations for either more than two children or no children at all. The chart also doesn't say how many reservations included one child versus two children. This data would be valuable for researching family travel patterns or figuring out how many rooms should include additional beds.

LOOKUP

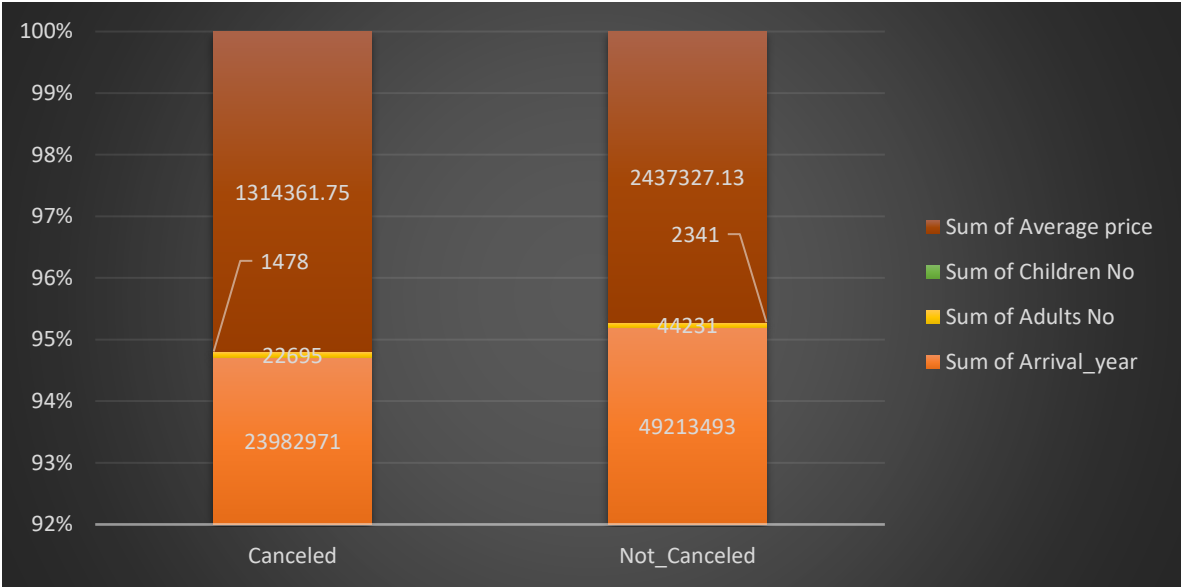
In this, booking ID will be used in the formula to find up the average price for that particular The formula looks for the booking ID within the given table range and then provides the average price that should be associated with that booking ID. In this situation, the typical cost of a reservation with the ID INN00001 is \$65.

Based on the information provided, it is derived that the bookings that are not cancelled contribute a significantly greater amount to the total revenue than bookings that are cancelled. It is also possible to see that the average price for a reservation that was not cancelled is greater than the average price for a reservation that was cancelled. This could be the result of a number of different things, including the duration of the stay, the type of room that was chosen, or the time of year that the reservation was made.

It is also important to observe that the average price for all bookings combined is slightly lower than the average price for a booking that was not cancelled. This finding raises an intriguing point. This could be owing to the fact that the vast majority of the reservations were kept, and the average price for these reservations is greater than the average price for reservations that were cancelled. It is also likely that there were some outliers in the data, which caused the average price for all bookings to be somewhat lower than the average price for bookings that were not cancelled. This is something that needs to be investigated further.

In conclusion, the data presented reveals insightful information regarding hotel bookings and the average pricing of their rooms. It demonstrates the significant impact that not-canceled bookings have made to the overall revenue as well as the higher average price that not-canceled bookings have in comparison to cancelled bookings. A helpful tool for analyzing individual bookings is provided by the formula, which is used to seek up the average price for a certain booking ID.

Chart showing the different segments such as year, adults children, price as well as booking status.



(chart showing the data) (created by student)

The change in the price of booking or an average cost of bookings can be shown in a column chart over time, however there is no time-related information in the provided data. As a result, we are unable to utilize a line chart to depict the provided data (Aitken, 2006) . To contrast the total number of visitors, adults, kids, and the overall average price for cancelled and uncanceled reservations, we can use a bar chart or stacked column chart. Four categories would be listed on the X-axis of the graph: "Arrivals," "Adults," "Children," and "Average Price." The total values for each category would be shown on the Y-axis.

The graph, for instance, would demonstrate that the total number of adults for reservations that were not cancelled was higher than for reservations that were. Additionally, it would demonstrate that non-cancelled bookings generally cost more overall than cancelled ones. As a result, we would be able to quickly compare the sum numbers for each category and obtain understanding of the variations between cancelled and uncanceled appointments. Overall, while a column chart cannot be used to depict the supplied data, a bar chart or stacked column chart can be used to show the differences between cancelled and uncanceled bookings.

PART 2

1- How many students are male and female?

Statistics		
Gender		
	Valid	2133
	Missing	0

		Frequency	Percent
	male	1077	50.5
	Female	1056	49.5
	Total	2133	100.0

(created by student)

INTREPRETATION

Above table gives information on gender distribution of the sample size of the 2133 students. By analyzing it has been interpreted that there are 1077 male students, representing 50.5% of sample size and about 1056 female students representing 49.5% of sample size.

2. How many student qualifies and does not qualify for reduced/free lunch?

Statistics

Lunch		
	Valid	2133
	Missing	0

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)

INTREPRETATION

As per the above table it has been found that sample size of study is 2133 and out of all these students about 1215 students who are 57% of sample size who do not qualify for the free lunch or reduced lunch, meanwhile about 918 students representing 43% of sample have qualified for free/reduced lunch.

3. What are the mean and median of students who qualifies and does not qualify for reduced/free lunch?

Statistics		
Lunch		
	Valid	2133
	Missing	0
Mean		1.4304
Median		1.0000
Mode		1.00

(created by student)

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)

INTREPRETATION

The above table summary that mean is 1.4304, which states that majority if the students from the sample size have not qualified for the free lunch/reduced lunch. Median is 1.0000, stating that many of students haven’t qualify for reduced or free lunch. Mode also supports the above statements that common category for qualification of lunch also do not qualify for the lunch.

4. What are the mean and median differences of students pre-test and post-test scores?

Statistics			
		Pretest	Posttest
	Valid	2133	2133
	Missing	0	0
Mean		54.9559	67.1022
Median		56.0000	68.0000
Mode		60.00	69.00

(created by student)

Table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	22.00	1	.0	.0	.0
	23.00	1	.0	.0	.1
	25.00	1	.0	.0	.1
	26.00	3	.1	.1	.3
	27.00	5	.2	.2	.5
	28.00	9	.4	.4	.9
	29.00	17	.8	.8	1.7
	30.00	27	1.3	1.3	3.0
	31.00	19	.9	.9	3.9
	32.00	23	1.1	1.1	5.0
	33.00	26	1.2	1.2	6.2
	34.00	24	1.1	1.1	7.3
	35.00	18	.8	.8	8.2
	36.00	36	1.7	1.7	9.8
	37.00	29	1.4	1.4	11.2
	38.00	38	1.8	1.8	13.0
	39.00	46	2.2	2.2	15.1
	40.00	45	2.1	2.1	17.3
	41.00	42	2.0	2.0	19.2
	42.00	42	2.0	2.0	21.2
	43.00	49	2.3	2.3	23.5
	44.00	43	2.0	2.0	25.5
	45.00	48	2.3	2.3	27.8
	46.00	37	1.7	1.7	29.5
	47.00	42	2.0	2.0	31.5
	48.00	44	2.1	2.1	33.5
	49.00	43	2.0	2.0	35.5
	50.00	44	2.1	2.1	37.6
	51.00	40	1.9	1.9	39.5
	52.00	54	2.5	2.5	42.0
	53.00	59	2.8	2.8	44.8
	54.00	52	2.4	2.4	47.2
	55.00	59	2.8	2.8	50.0
	56.00	56	2.6	2.6	52.6
	57.00	45	2.1	2.1	54.7
	58.00	63	3.0	3.0	57.7
	59.00	61	2.9	2.9	60.5
	60.00	73	3.4	3.4	63.9
	61.00	56	2.6	2.6	66.6
	62.00	60	2.8	2.8	69.4
	63.00	64	3.0	3.0	72.4
	64.00	48	2.3	2.3	74.6
	65.00	59	2.8	2.8	77.4
	66.00	51	2.4	2.4	79.8
	67.00	46	2.2	2.2	82.0
	68.00	47	2.2	2.2	84.2
	69.00	28	1.3	1.3	85.5
	70.00	30	1.4	1.4	86.9
	71.00	31	1.5	1.5	88.3
	72.00	34	1.6	1.6	89.9
	73.00	27	1.3	1.3	91.2
	74.00	28	1.3	1.3	92.5
	75.00	23	1.1	1.1	93.6
	76.00	18	.8	.8	94.4

	77.00	15	.7	.7	95.1
	78.00	15	.7	.7	95.8
	79.00	13	.6	.6	96.4
	80.00	7	.3	.3	96.8
	81.00	12	.6	.6	97.3
	82.00	11	.5	.5	97.8
	83.00	11	.5	.5	98.4
	84.00	10	.5	.5	98.8
	85.00	6	.3	.3	99.1
	86.00	8	.4	.4	99.5
	87.00	6	.3	.3	99.8
	88.00	1	.0	.0	99.8
	89.00	2	.1	.1	99.9
	91.00	1	.0	.0	100.0
	93.00	1	.0	.0	100.0
	Total	2133	100.0	100.0	

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	32.00	1	.0	.0	.0
	34.00	1	.0	.0	.1
	35.00	1	.0	.0	.1
	36.00	2	.1	.1	.2
	37.00	3	.1	.1	.4
	38.00	4	.2	.2	.6
	39.00	9	.4	.4	1.0
	40.00	9	.4	.4	1.4
	41.00	11	.5	.5	1.9
	42.00	15	.7	.7	2.6
	43.00	23	1.1	1.1	3.7
	44.00	19	.9	.9	4.6
	45.00	26	1.2	1.2	5.8
	46.00	21	1.0	1.0	6.8
	47.00	30	1.4	1.4	8.2
	48.00	39	1.8	1.8	10.0
	49.00	39	1.8	1.8	11.9
	50.00	35	1.6	1.6	13.5
	51.00	42	2.0	2.0	15.5
	52.00	38	1.8	1.8	17.3
	53.00	45	2.1	2.1	19.4
	54.00	45	2.1	2.1	21.5
	55.00	48	2.3	2.3	23.7
	56.00	46	2.2	2.2	25.9
	57.00	49	2.3	2.3	28.2
	58.00	48	2.3	2.3	30.4
	59.00	57	2.7	2.7	33.1
	60.00	44	2.1	2.1	35.2
	61.00	48	2.3	2.3	37.4
	62.00	53	2.5	2.5	39.9
	63.00	40	1.9	1.9	41.8
	64.00	39	1.8	1.8	43.6
	65.00	37	1.7	1.7	45.3
	66.00	37	1.7	1.7	47.1
	67.00	49	2.3	2.3	49.4
	68.00	58	2.7	2.7	52.1
	69.00	65	3.0	3.0	55.1
	70.00	62	2.9	2.9	58.0

	71.00	51	2.4	2.4	60.4
	72.00	60	2.8	2.8	63.2
	73.00	54	2.5	2.5	65.8
	74.00	57	2.7	2.7	68.4
	75.00	45	2.1	2.1	70.6
	76.00	50	2.3	2.3	72.9
	77.00	47	2.2	2.2	75.1
	78.00	54	2.5	2.5	77.6
	79.00	57	2.7	2.7	80.3
	80.00	43	2.0	2.0	82.3
	81.00	39	1.8	1.8	84.2
	82.00	39	1.8	1.8	86.0
	83.00	31	1.5	1.5	87.4
	84.00	21	1.0	1.0	88.4
	85.00	29	1.4	1.4	89.8
	86.00	16	.8	.8	90.5
	87.00	23	1.1	1.1	91.6
	88.00	19	.9	.9	92.5
	89.00	15	.7	.7	93.2
	90.00	23	1.1	1.1	94.3
	91.00	22	1.0	1.0	95.3
	92.00	15	.7	.7	96.0
	93.00	23	1.1	1.1	97.1
	94.00	19	.9	.9	98.0
	95.00	15	.7	.7	98.7
	96.00	8	.4	.4	99.1
	97.00	3	.1	.1	99.2
	98.00	6	.3	.3	99.5
	99.00	3	.1	.1	99.6
	100.00	8	.4	.4	100.0
	Total	2133	100.0	100.0	

(created by student)

INTREPRETATION

On the basis of given table and after interpreting it, it is clearly evident that’s pretest score mean is 54.96 & posttest score mean is 67.10, indicating increase in average score by about12.14 points from the pretest to posttest score. Whereas, pretest score median is 56 & posttest score median is 68, which indicates increase in 12 points. Mean and Median both are showing increase in scores of posttest & posttest, which shows overall improvement in performance of students.

5. Exemplification of clustering

Quick Cluster

Initial Cluster Centers		
	Cluster	
	1	2
P	23.00	93.00
	35.00	100.00

	Change in Cluster Centers	
	1	2
1	30.191	33.606
2	.885	.791
3	.611	.532
4	.212	.181
5	.227	.200
6	.150	.128
7	.000	.000
a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 7. The minimum distance between initial centers is 95.525.		

Final Cluster Centers

	Cluster	
	1	2
Pretest	43.05	65.32
Posttest	54.70	77.90

(created by student)

Number of Cases in each Cluster

Cluster	1	993.000
	2	1140.000
Valid	2133.000	
Missing	.000	

(created by student)

INTREPRETATION

Above table showcase outcome of cluster analysis on data set of 2133 students. First, cluster center was set at the pretest score of 23 & 93 for the 1 and 2 cluster & posttest score of 35 & 100 for the cluster 1 and 2. Analysis updated iteratively till achievement of convergence which happened at 7 iteration. Final cluster center pretest and posttest score were 43.95& 65.32 and 54.70 & 77.90 for the cluster 1 & 2. On the basis of above findings, it has been interpreted that 1 cluster have students with the high scores in comparison to 2 cluster.

6. With real world (practical) examples, explain minimum 5 most common data mining and text mining methods for each that can be used in the contemporary business.

Here are some examples of data and text mining methods which are used in settings of contemporary business-

Data Mining

Clustering Analysis-It is the technique which groups same data together on the basis of specific criteria. For instance- retailer may use this analysis for grouping consumers on the basis of their buying habits & then make customized marketing campaigns for each of the cluster (Antons and et. al., 2020)

Classification Analysis: This is the technique which categories data in pre- defined groups on basis of attributes. For example- Bank may use this for categorizing applications of loans as low or high risk on basis of factors such as income, score of credit & history of employment.

Regression Analysis: In this, first identification of relationship among dependent and 1 or more independent variables. For example- real estate firm may use this analysis for identifying factors which affects the prices of houses in specific area like square footage, location, bedrooms numbers.

Association Rule Mining: It’s a technique which identify relationship and patterns among items in the dataset. Retailers use this particular technique for identifying goods which are purchased collectively and then use the information for creating target promotions or the offers for product bundling (Papagiannopoulou and Tsoumakas, 2020).

Text Mining

Machine Learning: This is a text mining technique which uses natural language processing & machine learning for identification & extracting subjective information from the text. For instance- hotels can use this for analyzing the consumer service and identify the most common issues and complaints.

Prediction: In this technique, firstly identification and extraction from the topics from big sets of data of the text. For example- new organization may use it for identifying common discussed topics across various sources of new.

Data Visualization: It is used for extracting and identifying named entities like organizations, people & locations from the text. For example, law firms can use this for extraction of info about the companies and individuals mentioned in the legal documents.

Text Classification: This technique is used for categorizing text in pre-defined groups on the basis on its content. For instance- media platforms may use this for identifying and removing inappropriate content.

Information Extraction: In this technique, information is extracted from specific pieces like numbers, dates and locations. For example- medical research firm may use this for data extraction from clinical trial reports(Witten, 2011)

7. Advantages and disadvantage of using SPSS and Excel by use of relevant practical examples

Advantages and Disadvantages of using SPSS and Excel-

Criteria	SPSS	Excel
Data management & organization	SPSS is specifically designed for the statistical analysis & offers the advanced features for the data management & organization which includes variables and value labels, handling of missing value & transformations of data. SPSS might handle the large datasets & supports vast range of data formats (Pejić Bach, and et. al., 2019).	Excel is general-purpose spreadsheet program which lacks advanced data management features of the SPSS. Excel may be hard for managing the large datasets & can need transformations of manual data.(Java,2021)
Data analysis	SPSS provides vast range of the statistical analyses consisting of inferential statistics, descriptive statistics, factor analysis & regression analysis. SPSS have the user-friendly interfaces for every analysis &s generates the detailed reports which are easy in interpreting.	Excel gives the basic statistical functions like mean, median & standard deviation, but lacks advanced statistical features of the SPSS. Excel needs more manual input & ca not give same level of the output as the SPSS.
Graphical representation of the data	SPSS offers various chart types & customization options for the visualization of data including histograms, scatterplots and bar charts, SPSS graphs are ready for publication & may be easily exported to the other software.	Excel have limited selection of chart kinds & customization options & can need more effort for creating publication-ready graphs. Excel graphs might not be as the visually appealing as the SPSS graphs (Rahman and Muktadir, 2021).
User interface & ease of use	SPSS have user-friendly interface with intuitive menus & dialog boxes which guide users through each of the analysis. SPSS gives helpful documentation & online resources for the learning of how to use software.	Excel have steeper learning curve for the statistical analysis & can need more technical knowledge for using effectively. Excel may not be as the user-friendly for statistical analysis as the SPSS.
Cost	SPSS requires license fee, but can be available for free by academic institutions. SPSS offers free trial period for the new users.	Excel is included often in the Microsoft Office suites & can already be available on various computers. Excel is less expensive than the SPSS but might need purchase of the add-ons or plugins for the advanced statistical analysis.

Practical Examples

Data management and organization- If the researcher has big dataset with various variables & values, SPSS can be better choice for the data management & organization. For instance, study on the employee job satisfaction might have different survey questions which needs to be labeled & organized in particular way. SPSS may handle all these tasks efficiently & give clean, organized dataset for the analysis. On other hand, if researcher have small dataset with less variables. Excel can be sufficient for the basic data management tasks.

Data analysis- If researcher requires to conduct the complex statistical analysis, like multiple regression or factor analysis, SPSS would be better choice. For instance, study on the factors which influence the college graduation rates can require regression analysis with the multiple predictor variables. SPSS may perform this analysis with quite ease & generate comprehensive report which explains results. Excel, on other hand, might not have essential functions or can need manual input for conducting same analysis.

Graphical representation of data- If researcher wants for creating graphs ready for publication or visualizations, SPSS will be better choice. For instance, study on relationship among age & income might need scatterplot with the regression line. SPSS may create this graph with the custom labels & formatting options which make it ready. (Witten, 2011)

Conclusion

Based on information that I collected SPSS system is very useful as has various statistics used in education, researches, planning, collecting and analyzing lots of industries, also very helpful in quick calculations.

REFERENCES

Books and Journals

Aitken P.G. (2006) *Excel pivot tables and charts*. John Wiley & Sons.

Alexander M. and Jelen B. (2001). Pivot table data crunching.

Antons D. and et. al. (2020) The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, 50(3), pp.329-351.

Dierenfeld H. and Merceron A. (2012) Learning analytics with excel pivot tables.

Java (2021). *Advantages and Disadvantages of MS Excel - Javatpoint*. [online] Available at: <https://www.javatpoint.com/advantages-and-disadvantages-of-ms-excel> [Accessed 10 May 2023].

Papagiannopoulou E. and Tsoumakas G. (2020). A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2), p.e1339.

Pejić Bach and et. al. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability*, 11(5), p.1277.

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.

Witten,I. H.(2011) Data mining practical machine learning tools and techniques .3rd ed.Amsterdam:Elsevier /Morgan Kaufmann.