



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

BA BUSINESS MANAGEMENT

Assignment 2

Module: Data Handling & Business Intelligence

Code: LCBB5000

Lecturer: Dr Xavier Matieni

Name: Slavyana Nedyalkova

Student ID: 2129808

SUBMITTED ON THE 26th December 2023

Table of contents

Part 1: A Summary of The Dataset.....	3
IF Function.....	6
VLOOKUP FUNCTION	7
ANNUALY AVERAGE PRICES.....	8
PART 2: Data Analysis using SPSS.....	11
Descriptive Lunch Qualification Statistics.....	15
References	21

Part 1: A Summary of The Dataset

There are 36,276 columns and a total of 19 rows.

These are all the characteristics of the column, along with their respective meanings.

- A unique identification called the Booking_ID is given to every single booking.
 - Use: This is a crucial field that is used to uniquely identify each booking record inside the dataset.
- The term "no_of_adults" refers to the total number of adults that are part of the reservation.
 - Use: This function gives information about how many people are expected to attend each reservation.
- "no_of_children" refers to the total number of children that are covered by the reservation.
 - Similar to how the 'no_of_adults' element functions, this feature provides information about how many children are in the guest party.
- no_of_weekend_nights: • Definition: The quantity of Saturday or Sunday weekend nights the guest actually spent at the hotel or was originally scheduled to spend there.
 - Indicating how much time was spent there throughout the course of the weekend.
- The term "no of week nights" refers to the number of weeknights (Monday through Friday) that the guest either made a reservation for or actually stayed at the hotel.
 - When used on weekdays, it indicates how long the visitor will be staying.
- The sort of meal plan that the customer has set aside for themselves is indicated by the word "type_of_meal_plan".
 - Application: Here, details on dietary requirements and whether or not meals are included in the reservation are provided.
- required_car_parking_space: • Definition: This field indicates whether or not the customer needs a place to park their car (0 - No, 1 - Yes).
 - Application: This feature provides details on whether parking facilities are necessary for visitors arriving by car.
- The type of room that the customer has reserved (encoded) is the definition of the room_type_reserved variable.
 - Use: This denotes the specific kind of accommodation that the visitor has booked for their visit.
- The days that have elapsed between the reservation date and the arrival date constitute the "lead time."

- Application: This shows how much lead time or notification the visitor gave ahead of time for their arrival.
- The definition of "arrival_year" is the year of the arrival date.
 - Application: Indicates the year that visitors arrive at the venue.
- Arrival_month:
 - Arrival_month: • Definition: The month of arrival.
- Indication: This denotes the month of the visitors' arrival. The term "arrival_date" describes the month on which the arrival is planned to occur.
 - The use of this term indicates the day of the month that the guests come.
- market category
 - This phrase is defined as market segment designation.
- Application: Determines which market category the reservation falls into, providing information about the reservation's type or place of origin.
- repeated_guest:
 - Definition: Indicates if the client is a return visitor (0 for no, 1 for yes).
 - If the visitor has stayed at the hotel before, it is shown with a flag.
- The number of prior cancellations made by the customer: This is defined as the number of bookings that the customer canceled before the order that is currently being processed.
 - The guest's cancellation history is reflected in how this function is used.
- The number of former appointments that were not canceled: This word refers to the quantity of earlier appointments that the client did not cancel before the current booking was scheduled.
 - Application: This feature provides a history of the guest's confirmed reservations without a cancellation.
- Average price per room.
- This is the average cost, stated in euros, that is incurred for each day of the reservation.
 - Relates to the reservation and represents the average daily rate for a room.
- The no_of_special_requests variable is defined as the total number of special requests that the customer has made.
 - Any special requests or preferences that the guest may have had are included in this section.
- The booking_status is as follows:
 - Described as a signal indicating if the reservation was initially canceled or not.

- Application: Whether the reservation was fully canceled is indicated by the binary flag.

It was found that there were no missing values in any of the 36,276 columns when it came to the missing data.

The dataset is incredibly complete, with non-null entries for every variable in all 19 observations. There are several ramifications for this.

- The absence of missing values ensures that every column contains all of the information that would be expected of it and simplifies the procedures of modeling and data analysis.

Assurance of data quality: The dataset exhibits a high degree of data integrity, making it more suitable for reliable analyses and well-informed decision-making. • Researchers and analysts can move forward with confidence, as there are no missing values in the dataset that could potentially lead to uncertainties.

It is highly advised that users utilise the whole range of information available in the dataset due to its comprehensiveness in order to obtain robust modeling and pertinent insights.

IF FUNCTION

The goal of this exercise is to determine the "Book Status" based on the number of nights reserved in column E.

Getting Information from the Data:

This is Column E, which will show the "Book Nights."

The criterion that needs to be verified is if the number of nights reserved (E2) equals zero.

The "IF" function is a suitable option in this case since it allows us to specify several outcomes based on a certain condition.

Overview of the Structure of the "IF" Function:

The definition of the "IF" functionality's structure is =IF(condition, value_if_true, value_if_false).

In this specific case:

The requirement is that E2 must equal 0.

"BOOK" is the value when it is true; "CANCEL" is the value when it is false.

Formula explanation in accordance with:

Should the quantity of nights reserved in E2 be zero, the formula will get the outcome "BOOK."

If there are more nights reserved than zero, the formula will yield the output "CANCEL."

VLOOKUP FUNCTION

The Excel VLOOKUP function is a very helpful tool for browsing through and extracting data from a table.

This is the formula that was applied: `=VLOOKUP (10, FALSE, A2, A:S)` `=VLOOKUP(` The VLOOKUP function has only recently started at this stage. Excel is notified by this message that a vertical lookup analysis is about to be performed.

The letter A2 indicates one of the values you want to check in the table's first column. In this instance, a Booking ID is involved. A:S, This part of the code specifies the table range in which Excel should look for the value. The range A:S indicates, in the context of my case, that the table spans the full dataset, from column A to column S. Excel will search for the value represented by the letter A in the first column, which is indicated in A2.

10: The column index from which the data can be obtained is represented by this number. Since the value in your calculation is 10, Excel will pull the number from the tenth column in the database, which in this case is the Arrival_Year column.

NOT TRUE): Specifically, the last part of the function deals with the "range_lookup" option. When it is set to FALSE, it indicates that an exact match is required. This means that the VLOOKUP function won't return a result unless it finds a value in the table's first column that precisely matches the value given in option A2.

The formula effectively states, "Look up the value in A2 in the first column of the range A:S, and if you find it, give me the corresponding value from the tenth column." This provides a clear description of the goal of the formula.

ANUALLY AVARAGE PRICES

You can dynamically determine the average price for each year by using the formula `AVERAGEIF(J:J, J27, Q:Q)`. This makes it a useful tool for data analysis and summarization based on certain standards. The presented data sets represent the expected average prices for the years 2017 through 2021.

Column J lists the year of arrival, while Column Q lists the price.

It was found that the average price for 2017 was 90.0498541. In 2018, the average price was 106.4527274. In 2019, the average price was 106.4527274 as well. In 2020, the average price was 105.339899. In 2021, the average price was 101.2333333.

An analysis is presented using a histogram.

In order to determine whether the average cost of a room reservation has increased or decreased between 2017 and 2021, you can assess and compare the average expenses for each year.



Considering what we looked into, these are the conclusions:

Regarding the year 2018:

The significant increase in the average accommodation reservation charge from \$90.05 in 2017 to \$106.45 in 2018 indicates a significant upward shift. There are several possible reasons for this, such as:

There might have been a rise in the number of guests staying at the hotel or attending special events all year long, which allowed the hotel to hike prices.

Improvements to Services a higher price point than the preceding one could be explained by the hotel offering additional services or amenities.

From 2018 to 2019: The pricing strategy appears to be consistent, as seen by the lack of any noticeable change in average prices between 2018 and 2019. Here are a few theories that could be true:

It's possible that the hotel has identified the ideal pricing point at which supply and demand in the market are balanced.

Reliable Service: The hotel might not have felt the need to significantly alter the rates for its rooms if it had continued to provide consistent levels of service.

From 2019 to 2020: The slight decrease in average expenses, from \$106.45 in 2019 to \$105.34 in 2020, can be ascribed to several factors, such as the following:

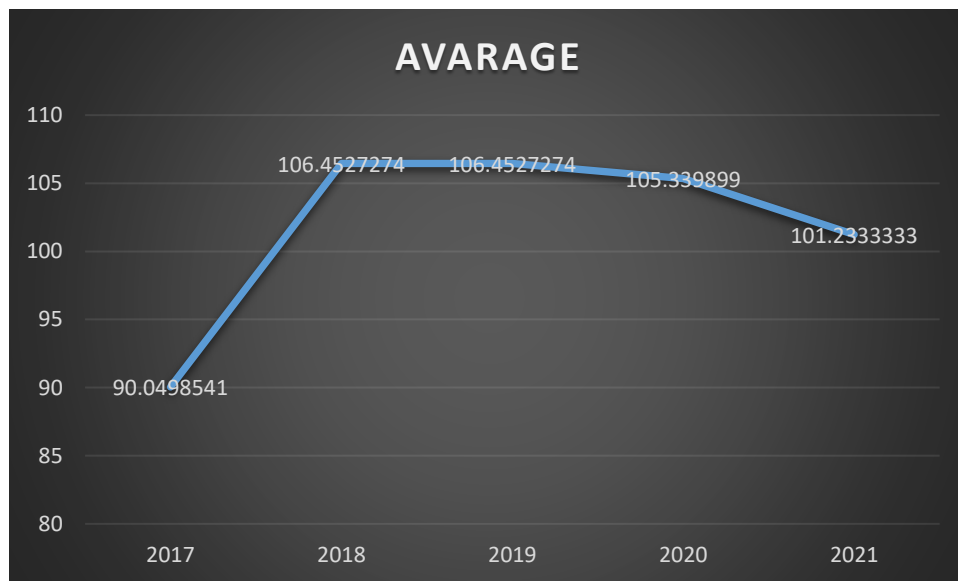
The state of the economy 2020 may have seen the use of a more cautious pricing strategy due to economic uncertainty or downturns.

It is plausible that competitive forces, such as heightened competition or the need to draw in cost-conscious customers, resulted in a modest price reduction.

Between 2020 and 2021: To further explain: The average expenses decreased further in 2021, from \$105.34 in 2020 to \$101.23, suggesting that adjustments are still being made, most likely due to the following factors:

Global Events (including pandemics): It is plausible that unforeseen events, like the COVID-19 epidemic, impacted travel patterns and led to deliberate modifications in prices.

LINE CHART TO REPRESENT THE ANALYSIS.



PART 2: Data Analysis using SPSS.

Gender Distribution

Frequency:

There are two gender groups in the dataset: "1" pertains to males and "2" pertains to females.

There are 1056 male pupils who are a "1" and 1077 female students who are a "2"

To a point:

Within the dataset, male students constitute 49.5% of the total (category "1").

Around fifty-five percent of the students are female (category "2").

Accurate Percentage:

This proportion is calculated based on the total number of legitimate instances, omitting any instances that are either missing or invalid.

There are 49.5% male pupils and 50.5% female students, according to the valid percent.

Overall Percentage:

As you scroll down the table, the cumulative percent provides a summary of the valid portions of the percentage.

After the "Male" classification, there are 49.5% of cases that are classified as genuine. When "Female" is taken into consideration, the cumulative proportion gets closer to 100 percent.

In summary, the dataset exhibits a gender distribution that is balanced, with a slightly higher proportion of female students than male students. These data provide information about the composition of the student population and may assist in explaining gender-related trends or patterns in research that will be conducted in the future.

Results from SPSS analysis follow.

Statistics

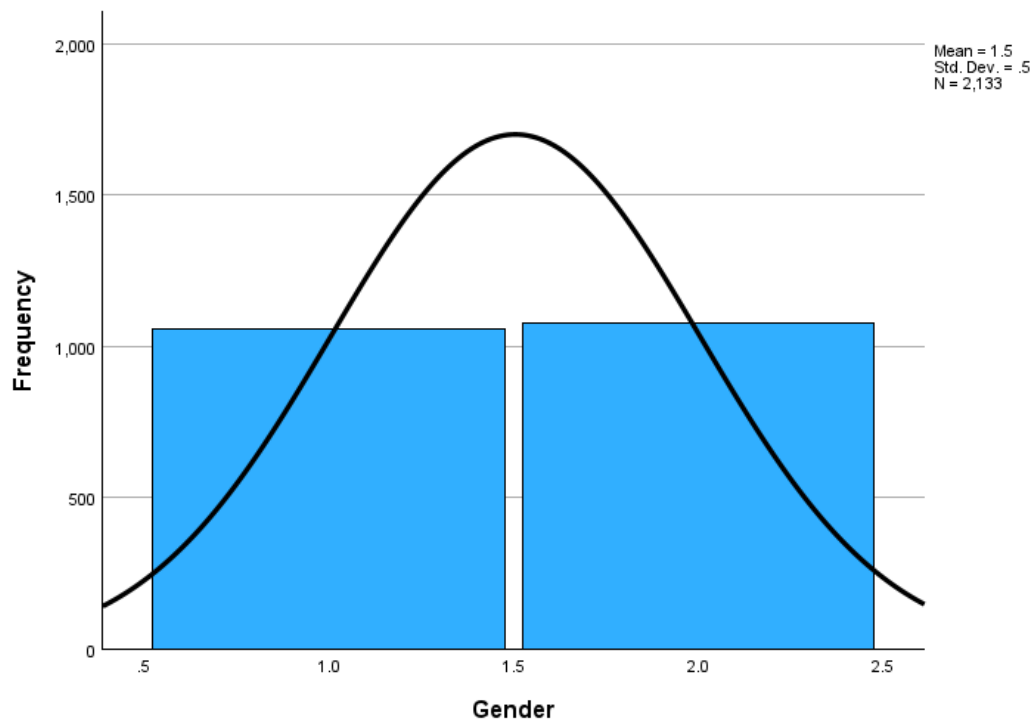
Gender

N	Valid	2133
	Missing	0

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1056	49.5	49.5	49.5
	2	1077	50.5	50.5	100.0
	Total	2133	100.0	100.0	

The graphical chart below represents the gender distribution.



Lunch Qualification

Frequency:

- The dataset is divided into two groups: • "0" (indicating pupils not eligible for reduced/free lunch).
- "1" (likely for reduced/free lunch pupils).

Approximately 57.0% of kids fall into the "0" category, meaning they do not qualify for reduced/free meals.

43.0% of students qualify for reduced/free meals under the "1" category.

Valid Percent:

Calculated by counting all valid cases, omitting missing or invalid data.

57.0% of genuine cases do not qualify for reduced/free lunch ("0"), while 43.0% do ("1").

Cumulative Percent:

This provides a running total as you progress down the table.

- 57.0% of genuine cases were covered after the "0" category.

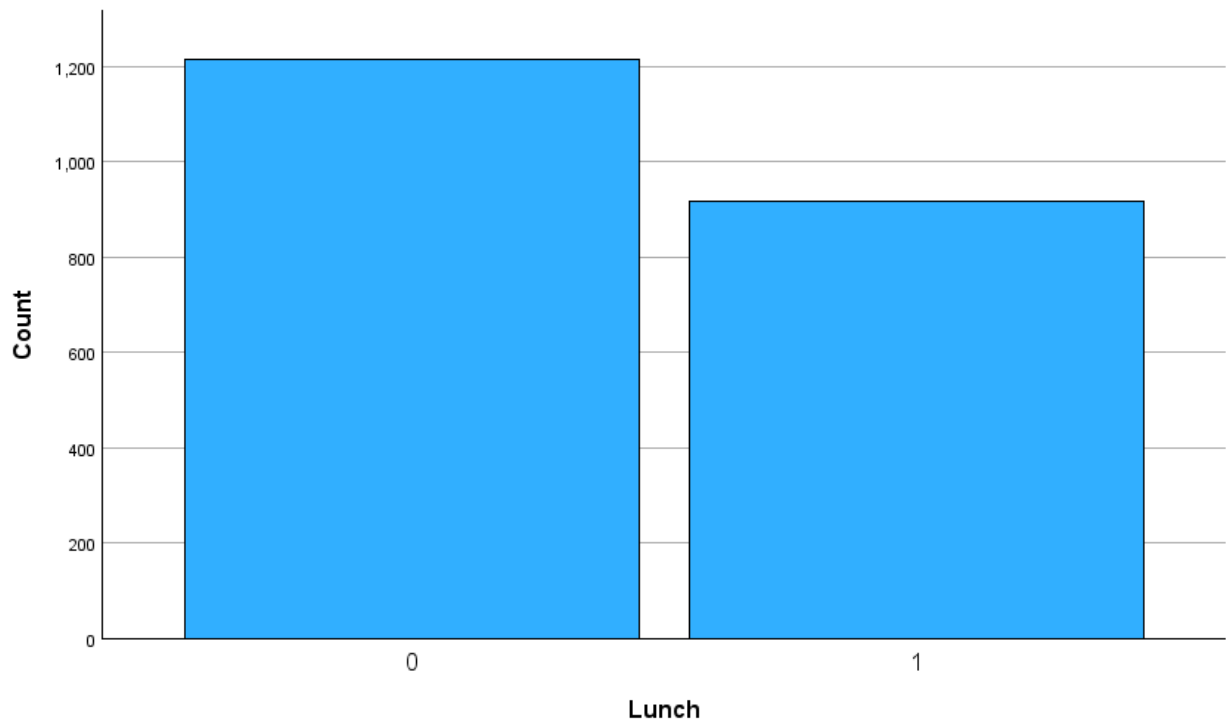
The cumulative percent gets to 100.0% after "1".

Interpretation: The statistics shows that 57.0% of pupils do not qualify for reduced/free lunch, whereas 43.0% do. This data on student lunch qualification distribution can help explain the school's socioeconomic context.

Below is the graphical representation of the findings.

Lunch

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1215	57.0	57.0	57.0
	1	918	43.0	43.0	100.0
	Total	2133	100.0	100.0	



Descriptive Lunch Qualification Statistics

The pre- and post-test scores:

Pre-test scores:

- N (Number of Cases): The dataset has 2133 valid pre-test scores, providing a sufficient sample size for analysis.
- The pre-test score range is 71, represented by the difference between the highest and lowest results. This range illustrates student pre-test performance variability.
- Pre-test scores range from 22 to 93. These values reveal the dataset's full pre-test scores.
- The total pre-test scores are 117,221, indicating pupils' cumulative performance.
- The average pre-test score is 54.96. This statistic measures central tendency, reflecting student performance averages.
- Standard Deviation: The 13.563 standard deviation indicates pre-test score variability or dispersion. Higher standard deviations indicate variability.
- Variance: The square of the standard deviation, approximately 183.958, measures data spread another way.
- Post-test results:
- N (Number of Cases): 2133 valid post-test scores establish a consistent dataset for analysis, like pre-test scores.
- Range: The post-test score range is 68, suggesting diversity in student performance.
- The least post-test score is 32, while the maximum is 100. These values show the dataset's post-test range.
- Sum: The post-test scores total 143,129, indicating student performance.
- Mean: The average post-test score is around 67.10, indicating a central tendency in performance.
- Standard Deviation: The 13.987 standard deviation evaluates post-test score variability, providing insights on performance spread.
- The variance, at 195.630, complements the standard deviation in measuring data dispersion.
- Valid listwise N: 2133
- The Valid N (listwise) of 2133 indicates no missing values in pre- or post-test scores, assuring dataset integrity.

The descriptive statistics for pre-test and post-test scores provide a thorough understanding of the dataset. The range, lowest, and maximum numbers show score distribution, while the mean

and standard deviation measure central tendency and variability. No missing values ensure dataset trustworthiness for further analysis.

These statistics provide a foundation for understanding student performance and examining correlations, trends, and probable factors affecting pre-test and post-test scores.

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
Pretest	2133	71	22	93	117221	54.96	13.563	183.958
Posttest	2133	68	32	100	143129	67.10	13.987	195.630
Valid (listwise)	N2133							

K-Means Clustering:

This unsupervised machine learning approach divides a dataset into a set number of clusters. Minimum variation among clusters, maximum variance between clusters. Data points are clustered and cluster centers updated until convergence is obtained by the method.

In the results, groups are generated based on "Gender" and "Lunch," and each cluster's centroid represents the average values of both attributes for its cases.

First cluster centers:

Each cluster begins at its initial center. The center of each cluster is a set of feature values. These aspects are "Gender" and "Lunch." Initial centroids for each cluster are shown:

In Cluster 1, gender is 1 and lunch is 0.

Cluster 2: Gender is 1 and Lunch is 1

In Cluster 3, gender is 2 and lunch is 0.

Iteration History: K-means clustering creates stable cluster centers through iteration. Iteration history shows cluster center changes:

In Iteration 1, Cluster Centers for Clusters 1, 2, and 3 change.

These figures show how much the cluster centers shifted in the first iteration.

Iteration 2: The method converged with minimal cluster center changes.

The highest absolute coordinate change for every center is .000, suggesting little or no change.

Minimum distance between starting centers is 1.000, indicating they were at least 1 unit apart.

Last cluster centers:

Final cluster centers are stable centroids following convergence. These centers represent each cluster's average feature values:

Cluster 1:

Gender: 1 Lunch: 0 Cluster 2:

Gender: 1 Lunch: 1 Cluster 3:

Gender: 2 Lunch: 0

Cases per Cluster:

"Number of Cases in each Cluster" lists the number of data points in each cluster:

Cluster 1: 593 cases

Cluster 2: 918 cases

Cluster 3: 622 cases

The clustering technique divided cases into three groups based on "Gender" and "Lunch."

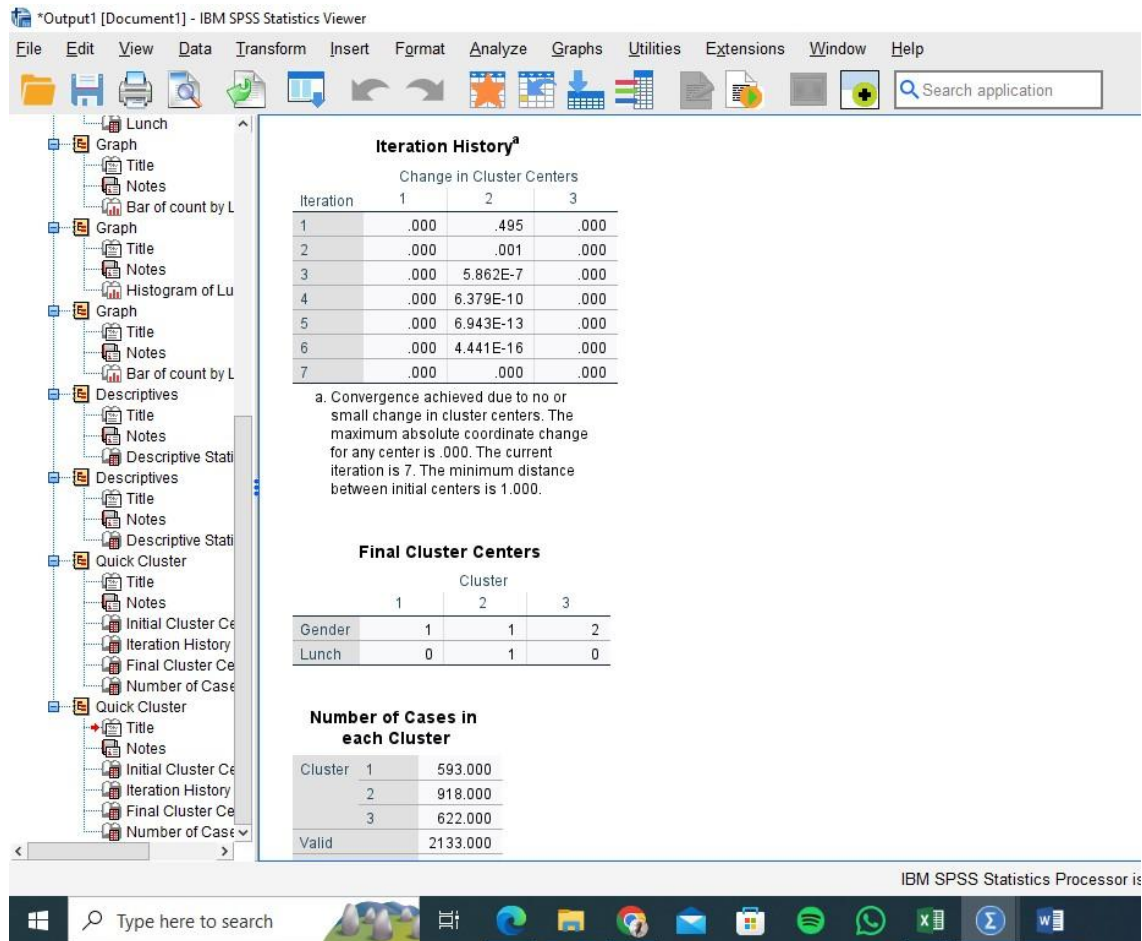
Each cluster has particular values of these attributes, and the final cluster centers are the average values.

Each cluster's case count shows how many data points belong to that group.

Convergence in two iterations shows the algorithm found stable clusters soon.

Initial Cluster Centers

	Cluster		
	1	2	3
Gender	1	1	2
Lunch	0	1	0



Common data mining and text mining methods for each that can be used in contemporary business.

1. Classification is the process of grouping data according to certain criteria.
2. Clustering is a technique that groups data pieces that are comparable based on their similarities.
3. The value of one variable can be predicted using regression analysis depending on the value of another variable.
4. Association Rules: This technique uses massive datasets to find correlations between different variables.
5. The process of identifying recurring sequences and patterns in data is known as sequence mining.
6. Identifying patterns and information in unstructured text data is the sixth step in the text mining process.

7. Neural networks, which are modeled after the human brain, are able to simulate complex interactions between several variables.
8. Identifying essential variables and visualizing decision-making can be accomplished using decision trees.
9. The Random Forest approach is an ensemble technique that combines the predictions of several different decision trees.

Advantages And Disadvantage of Using SPSS And Excel

Advantages of SPSS	Advantages of EXCEL
SPSS excels at fast statistical calculations and data manipulation.	Maintaining and handling huge amounts of client data reduces memory system data redundancy
Supercomputers and other fast machines employ SPSS, a statistical tool.	Excel's easy-to-use interface lets owners and staff insert data and calculate in any size firm.
Large-scale consumer data management in real time.	It displays data efficiently in tables, graphs, and charts.
SPSS excels at cluster, factor, logistic regression, and other sophisticated analytics.	Provides a more friendly approach since you can easily insert data in cells organized in columns and rows.
Makes it easy to use transforming variables to highlight data and show descriptive statistics.	Provides detailed insight into the usage of computing storage and links and relationships between data.
This software works for all technical and large enterprises.	Excel handles rows and columns like a relational database. So, data integration and storage are simpler.
SPSS complements a growing science and technology field.	Excel stores data in numerous formats.

(Natter, 2019; William, 2022)

Disadvantages of SPSS	Disadvantages of Excel
SPSS might be costly for students.	Excel makes statistical computations difficult. Complex analysis and advanced concepts are needed to perform the statistical process

For volume and rows, SPSS is like Access - not a fantastic spreadsheet style.	Complex analytics might take a long time to learn
	Formulas in Excel spreadsheets are prone to errors.
	It has a reduction in data redundancy of the memory system by maintaining and handling large amounts of customer data.

References

- Natter, E. (2019). The Advantages & Disadvantages of Spreadsheets. [online] Chron.com.
Available at: <https://smallbusiness.chron.com/advantages-disadvantages-spreadsheets-26551.html>.
- William, Kate (2022). What is SPSS? Definition, Features, Types, and Use Cases. [online] SurveySparrow. Available at: <https://surveysparrow.com/blog/what-is-spss/>.