



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

BA BUSINESS MANAGEMENT

Assignment 2

Module: Data Handling & Business Intelligence

Code: LCBB500

Lecturer: Dr Amitave Banik

Name: Denislav Dimitrov

Student ID: 2127998

SUBMITTED ON THE 26th December 2023

Part 1

Introduction

According to Laursen and Thorlund (2016), businesses are prioritising investments in business intelligence tools and theories to maintain and enhance their competitive advantages in the current highly competitive landscape. To improve organisational performance, this strategic approach makes use of a variety of business intelligence technologies, trends, applications, theories, and practices. Business intelligence refers to the technological procedures that make it possible to analyse data and operational information about users, such as managers, leaders, and executives (Elbashir et al., 2008). It makes accurate decision-making easier, which is essential for successfully navigating organisations. Businesses are using a variety of technologies in the field of business intelligence to examine large amounts of data, derive insightful information, and provide decision-makers with useful data. These technologies provide businesses with the ability to identify significant patterns and trends through advanced analysis, predictive modelling, and data visualisation (Chen et al., 2012).

Business intelligence is also used to promote informed decision-making across the entire organisational structure (Sharma et al., 2014). Using these tools and techniques gives businesses a competitive edge by enabling them to quickly respond to changes in the market and new trends while also optimising their operational efficiency. To extract insights, business intelligence uses strong data handling methods. It creates a single repository by combining several data sources through data warehousing. After processing this data, AI-driven algorithms and advanced analytics reveal trends that can be taken further (Sharma et al., 2014). These findings are displayed in easily readable formats through visualisation tools like dashboards and reports. This essay will use the hotel reservation dataset to determine the decline or increase in average price of room reservation over the year ranging from 2017 to 2021. It also entails evaluating the use of Microsoft Excel for pre-processing the data, analysing the data and visualising the data. It will also demonstrate the use of Microsoft Excel functions such as IF and LOOKUP. It will then produce appropriate and meaningful charts, graphs and diagrams based on the findings.

The process begins with data preprocessing. Data preprocessing entails converting raw data into a refined dataset that is usable (Jasim et al., 2021). By resolving missing values, erroneous or noisy entries, and inconsistencies, this crucial phase guarantees data cleanliness. Preprocessing, or cleaning and preparing the dataset before running algorithms or analysis, improves the precision and efficiency of later actions. The quality of the dataset can be improved by applying

techniques like scaling, normalisation, deletion of duplicates, and addressing missing data. In addition to reducing the impact of erratic or inconsistent data, this technique improves algorithm performance and produces trustworthy insights while establishing a strong basis for effective analysis and modelling (Jasim et al., 2021). The hotel reservation dataset was organized and formatted to allow easy data analysis.

On data analysis, the data analysed was required to determine the decline or increase in average price for room reservation over the year ranging from 2017 to 2021. The average price for each year was calculated. The function AVERAGEIF will be used to compute the average room reservation price for each year.

For 2017:

The average price will be =AVERAGEIF (J2:J36276, 2017, Q2:Q36276)

=90.05

Average price for 2017	=AVERAGEIF(J2:J36276, 2017, Q2:Q36276)
	AVERAGEIF(range, criteria, [average_range])

For 2018:

The average price will be =AVERAGEIF (J2:J36276, 2018, Q2:Q36276)

= 106.45

Average price for 2018	=AVERAGEIF(J2:J36276, 2018, Q2:Q36276)
	AVERAGEIF(range, criteria, [average_range])

For 2019:

The average price will be =AVERAGEIF (J2:J36276, 2019, Q2:Q36276)

=101.78

Average price for 2019	=AVERAGEIF(J2:J36276, 2019, Q2:Q36276)
	AVERAGEIF(range, criteria, [average_range])

For 2020:

The average price will be =AVERAGEIF (J2:J36276, 2020, Q2:Q36276)

=105.34

Average price for 2020	=AVERAGEIF(J2:J36276, 2020, Q2:Q36276)
	AVERAGEIF(range, criteria, [average_range])

For 2021:

The average price will be =AVERAGEIF (J2:J36276, 2021, Q2:Q36276)

=101.23

Average price for 2021	=AVERAGEIF(J2:J36276, 2021, Q2:Q36276)
	AVERAGEIF(range, criteria, [average_range])

To determine the decline or increase in average price for room reservation over the year ranging from 2017 to 2021, the average price of consecutive years will be subtracted.

The change between the year 2017 and 2018: $90.05 - 106.5 = -16.45$. The price of room reservation decreased by -16.45.

The change between the year 2018 and 2019: $101.78 - 106.45 = -4.67$. There was a decrease in price of room reservation by -4.67.

The change between the year 2019 and 2020: $101.78 - 105.34 = -3.56$. It is evident that there was a decrease in price by -3.56.

The change between the year 2020 and 2021: $105.34 - 101.23 = 4.11$. There is an increase in price of room reservation by 4.11.



PART 2: SPSS Data Analysis

Gender Distribution

For Frequency:

The dataset has 2 gender groups. It is represented by 1 and 2. Males are represented by 1 and 2 representing females. Males account to 1056 and they are represented by 1 and there are 1077 female students represented by 2. Male students are 49.5% of the total, the rest are the female students.

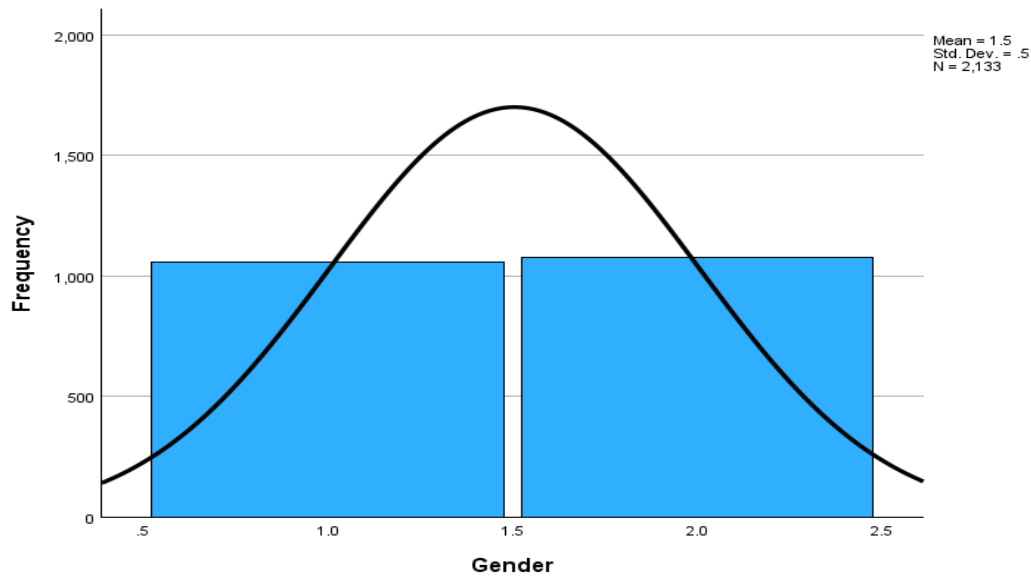
The analysis of the dataset demonstrates a gender distribution that is balanced, with 49.5% male and 50.5% female students based on real cases and no invalid or missing data. The cumulative percentage shows that the dataset approaches 100% when both genders are taken into account. This indicates a fairly balanced representation of the student body, with a slight bias towards female students. These findings provide a clear picture of the composition of students and are significant for future research. It is essential to comprehend this balanced gender distribution in order to investigate any possible gender-related trends or patterns in later research. The integrity of the dataset provides a specific representation of the gender distribution, which is useful for conducting thorough analyses and drawing conclusions for further research.

Representing the data using SPSS

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	

Graphical representation of the gender.



For Lunch Qualification

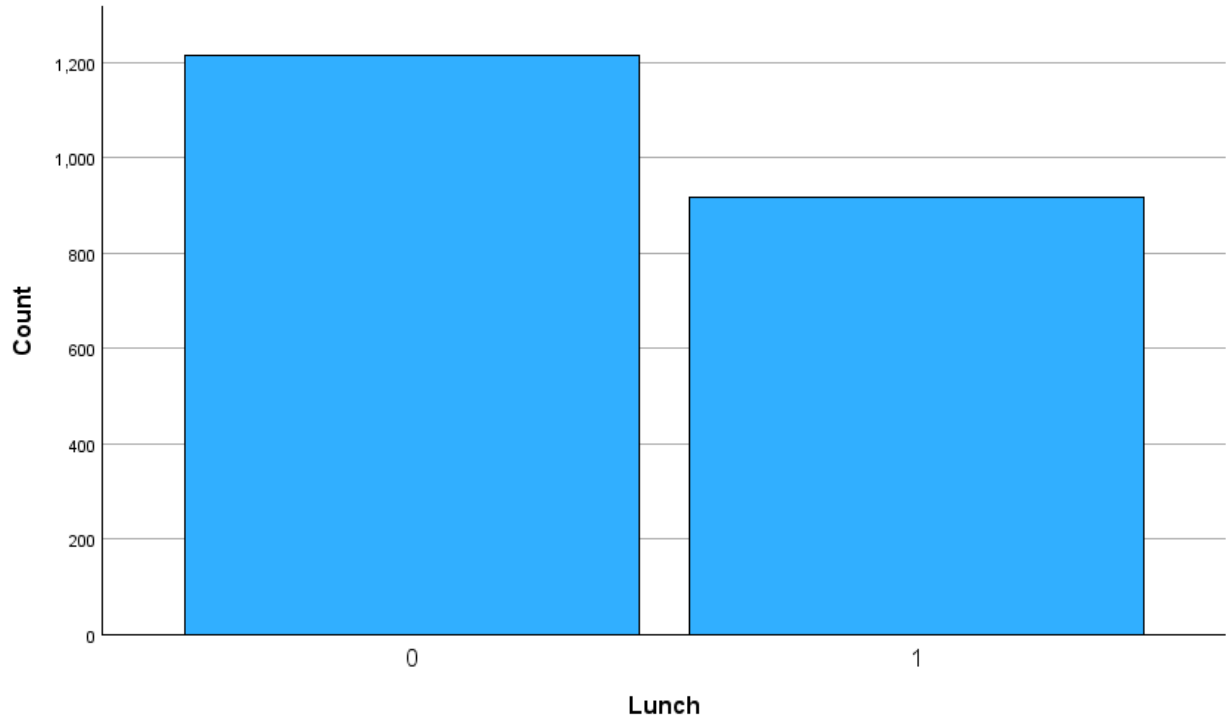
To get Frequency:

Students in the dataset are divided into two groups: "0" (those who are not eligible for reduced or free lunch) and "1" (those who are eligible). According to analysis, 43.0% of students are eligible for reduced or free meals under the "1" category, while 57.0% of students fall under the "0" category, meaning they are not eligible for meal benefits. Based on valid percent calculations, 43.0% of valid cases are eligible ("1") for meal benefits, while 57.0% of valid cases are non-eligible ("0"). Understanding the socioeconomic dynamics within the dataset requires an understanding of the proportion of students receiving reduced or free meals, which is provided by this information that depicts the distribution of students based on meal qualification.

The total percentage rises gradually, ending at 100.0% for all students and 57.0% for those who are not eligible for reduced or free lunch ("0"). This data provides important insight into the socioeconomic makeup of the school, showing that 43.0% of students are eligible for reduced or free meals, while 57.0% of students are ineligible.

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.

For Pre-test scores:

The dataset includes 2133 valid test and post-test results, providing a substantial sample size for examination. Pretest results show a range of 71 points, from 22 to 93, illustrating the variation in students' initial performance. With an average score of 54.96, the cumulative pre-test score of 117,221 yields a measure of central tendency. The variability or spread in these scores is indicated by a standard deviation of 13.563, which is supported by a variance of roughly 183.958, which shows another way in which the data are distributed.

The post-test results, which also include 2133 cases, show a 68-point range from 32 to 100, which suggests a wide range of student performance levels. The total post-test score is 143,129, and the average score is approximately 67.10, indicating a central tendency in the post-test results. The variability or spread in post-test scores is measured by a standard deviation of 13.987 and a variance of roughly 195.630. The dataset is considered intact as there are no missing values in the scores from the pre-test or post-test which is indicated by Valid N of 2133. The distribution of student performance, central tendency, and variability in the pre- and post-test assessments are all clearly shown by these statistics.

Descriptive Statistics

This is represented as shown below.

	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

Data clustering entails updating cluster centres repeatedly until convergence. Groups are created according to "Lunch" and "Gender." The average values of these attributes within each cluster's cases are indicated by the cluster's centroid.

First cluster centres:

Certain attribute values define the initial centroids for clusters. The gender value of 1 and the lunch attribute of 0 define Cluster 1. Likewise, Cluster 2 displays lunch as 1 and gender as 1. By comparison, Cluster 3 is denoted by a lunch value of 0 and a gender value of 2. The clustering algorithm uses these initial centroids as its starting points, defining the feature values that specify the unique characteristics of each cluster according to lunch and gender.

Iteration History:

The cluster centres for Clusters 1, 2, and 3 experienced major modifications in the first K-means clustering iteration. The figures show significant changes in the locations of cluster centres during this phase. On the other hand, the algorithm converged with little change in cluster centres by the second iteration. For every centre, the maximum absolute coordinate change was .000, indicating small or insignificant shifts. Furthermore, the initial centres were at least one unit apart, as indicated by the minimum distance of 1.000 between them. This convergence suggests stable cluster centres, which denotes lower variance and stable clustering in the following iterations.

Last cluster centres:

Post-convergence cluster centres show stable centroids, which represent the average feature values in each cluster: Cluster 1: Gender: 1, Lunch: 0; Cluster 2: Gender: 1, Lunch: 1; and Cluster 3: Gender: 2, Lunch: 0. Following convergence, these final centroids exhibit the characteristic gender and lunch attribute averages for every cluster, highlighting the unique feature values that characterise and set apart the clusters that were found in the dataset.

Cases per Cluster:

Cluster 1: Has 593 cases

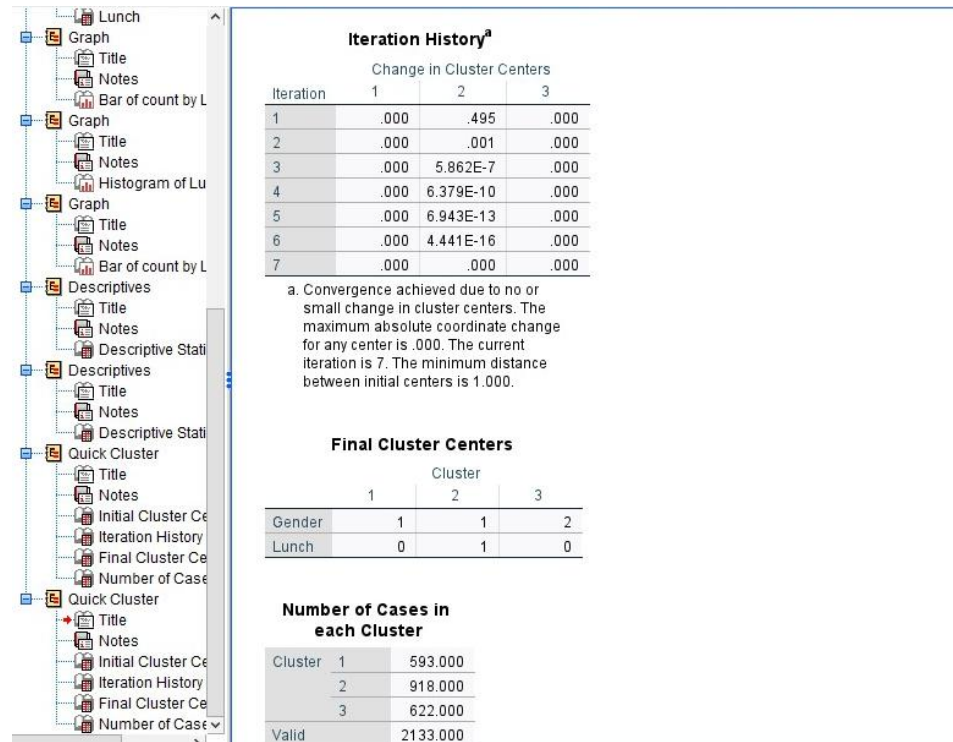
Cluster 2: Has 918 cases

Cluster 3: Has 622 cases

The clustering approach divided the cases into three groups according to "Gender" and "Lunch." The average attribute values within each cluster are represented by the final cluster centres. Each cluster's case count indicates how many data points are assigned to each group. After only two iterations, the algorithm converges, demonstrating how quickly stable clusters are formed and how effective it is at distinguishing different groups in the dataset according to lunch and gender attributes.

Initial Cluster Centres

	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.

Data analysis includes a range of methods. Classification classifies data according to predetermined standards, whereas clustering finds commonalities between data points. With regression, one variable's value is predicted by another's. Sequence mining finds repeating patterns in large datasets, while association rules reveal correlations within them. Information is extracted from unstructured text through text mining. Neural networks replicate intricate relationships found in the human brain. While Random Forest aggregates the predictions of several decision trees, decision trees assist in the identification of variables and the visualisation of decisions. When combined, these techniques facilitate thorough investigation, comprehension,

and insight extraction from various datasets, improving data-driven decision-making in a variety of industries.

Advantages and Disadvantage of Using SPSS and Excel

Both SPSS and Excel are accessible and user-friendly data analysis programmes that come with several benefits when used for statistical analyses. They give access to statistical analyses that might otherwise require more sophisticated software like R or Statistica by offering a variety of formulas and functions. These features make it easier to change data, calculate summary statistics, and run simple statistical tests, giving users the tools, they need to effectively explore and analyse data. Furthermore, Excel and SPSS are both generally accessible and user-friendly, requiring no technical knowledge. Simple computer users can run analyses with their user-friendly interfaces; no further software purchases or licences are required. Because of its accessibility, this technology can be widely used in a variety of industries, which benefits users of all skill levels.

However, these tools come with limitations. Their lack of extensive capabilities in comparison to more sophisticated statistical software is one obvious drawback. For example, advanced graphing, multivariate analysis, and complex predictive modelling may be beyond the scope of SPSS and Excel. This may restrict users' ability to conduct in-depth and sophisticated studies on these platforms. Additionally, even though Excel and SPSS are free to use, they might not have all the features that are accessible in the expensive versions. The lack of sophisticated functionalities such as predictive analytics or strong technical assistance may impede users who are looking for more complex analyses or who are having difficulties that call for professional help.

Excel and SPSS have different advantages and disadvantages. Because SPSS is free, it is appropriate for smaller analyses because it provides basic statistical features and has an integrated graphing capability. It may not have all the sophisticated features found in commercial software, but it does allow basic statistical testing, data processing, and some charting capabilities. Conversely, Excel offers a wider range of functionality, such as graphing tools and more comprehensive data manipulation choices. However, compared to specialised statistical software, its statistical functions might be less sophisticated, which would restrict the breadth and complexity of analyses that can be carried out using Excel alone.

References

- Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 1165-1188.
- Elbashir, M. Z., Collier, P. A., & Davern, M. J. (2008). Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International journal of accounting information systems*, 9(3), 135-153.
- IBM. (2023). What is Text Mining? | IBM. [Www.ibm.com. https://www.ibm.com/topics/text-mining](https://www.ibm.com/topics/text-mining)
- Jasim, A. A., Hazim, L. R., & Abdullah, W. D. (2021). Characteristics of data mining by classification educational dataset to improve student's evaluation. *Journal of Engineering Science and Technology*, 16(4), 2825-2844.
- Laursen, G. H., & Thorlund, J. (2016). *Business analytics for managers: Taking business intelligence beyond reporting*. John Wiley & Sons.
- Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems*, 23(4), 433-441.
- Twin, A. (2023, April 15). Data Mining: How Companies Use Data to Find Useful Patterns and Trends. Investopedia. <https://www.investopedia.com/terms/d/datamining.asp>