

Data Handling & Business Intelligence

(LCBB5000)

Ana-Maria Gorgan (2113739)

University of Wales Trinity Saint David

Submission Date: 12th of May

Table of Contents

Data Handling & Business Intelligence (LCBB5000).....	0
Introduction.....	2
Part 01: Data Pre-processing.....	3
IF Formula:	3
Pivot Table Formula:	4
Data Visualizing and Analysis.....	5
Part 02:	8
Number of Students by Gender.....	8
Qualify and Does not Qualify	9
Mean and Median	11
Differences of the Mean and Median.....	12
K-Means Clusters.....	13
Data Mining and Text Mining Method.....	13
Advantages and Disadvantages of Excel and SPSS	15
Conclusion.....	0
References.....	1

Introduction

This report aims to achieve two primary objectives, namely, the in-depth examination of two separate datasets and the practical illustration of techniques for data analysis and visualization. The report consists of two distinct sections, each focused on a distinct dataset and employing prescribed methodologies for data management and assessment. The objective of this report is to demonstrate the application of data analysis and visualization methods within these two domains, using an analysis of diverse contexts. Examining the insights and findings of the dataset provides significant information for making informed decisions. Additionally, they facilitate the understanding of patterns and relationships that are inherent in the data.

Part 01: Data Pre-processing

The data analyst followed a series of steps to import the dataset into Excel and ensure data integrity. A new workbook was created after opening Microsoft Excel. Afterward, the analyst found the Hotel Reservation dataset file on the computer and chose it for importing. Next, the analyst navigated to the location of the dataset file and selected it. Depending on the file format, the analyst specified the import settings, such as delimiter selection for CSV files or table selection for database files. The analyst also reviewed the preview of the imported data to ensure it appeared correctly.

IF Formula:

To demonstrate the use of the IF formula in the Hotel Reservation dataset, the data analyst can apply this formula to a specific column or create a new column based on certain conditions.

For example, let's consider the "Average price" column. The analyst may want to categorize the bookings into different price ranges based on certain thresholds. The analyst can create a new column called "Price Category" using the IF formula.

In Q2 empty cell next to the first row of data in the "Price Category" column, the analyst would enter the following formula:

```
=IF (Q2<=80, "Low", IF(Q2>100, "Medium", IF(Q2<=150,"High"))
```

This formula checks the value in the "Average price" column and assigns a price category based on the specified conditions. If the average price is less than or equal to 80, the formula assigns the category "Low". If the average price is greater than 100 and less than or equal to 150, it assigns the category "Medium". For any other average price values, it assigns the category "High".

By dragging the formula down to the rest of the cells in the "Price Category" column, Excel will automatically apply the IF formula to each row of data, categorizing the bookings based on their average price.

Pivot Table Formula:

To utilize the Pivot Table formula in the Hotel Reservation dataset, the data analyst can leverage this feature to summarize and analyze the data from different perspectives.

Let's say the analyst wants to understand the total number of bookings for each room type and their respective meal plans. They can follow these steps to create a Pivot Table:

1. Select the entire dataset, including the column headings and the data.
2. From the Excel menu, choose the "Insert" tab and click on the "PivotTable" option.
3. In the PivotTable dialog box, select the range of the dataset and choose where to place the PivotTable (e.g., a new worksheet or an existing one).
4. In the PivotTable Field List, drag the "Room_type_reserved" column to the Rows area and the "Meal_plan" column to the Columns area.
5. Drag any other desired field, such as "Average price," to the Values area.
6. By default, the PivotTable will show the Average prices. If preferred to see a different aggregation, such as the sum or average, it can be changed by clicking on the dropdown arrow next to "Sum of Average price" in the Values area and selecting the desired calculation.

Once the PivotTable is created, Excel will display a table summarizing the number of average prices for each room type and meal plan combination. This provides a convenient way to analyze and compare the booking patterns across different categories. The analyst can further enhance the PivotTable by applying filters, sorting, and customizing the layout and design to meet specific requirements. This allows for deeper analysis and visualization of the data in a meaningful and insightful manner.

Data Visualizing and Analysis

The data analyst can calculate the Av

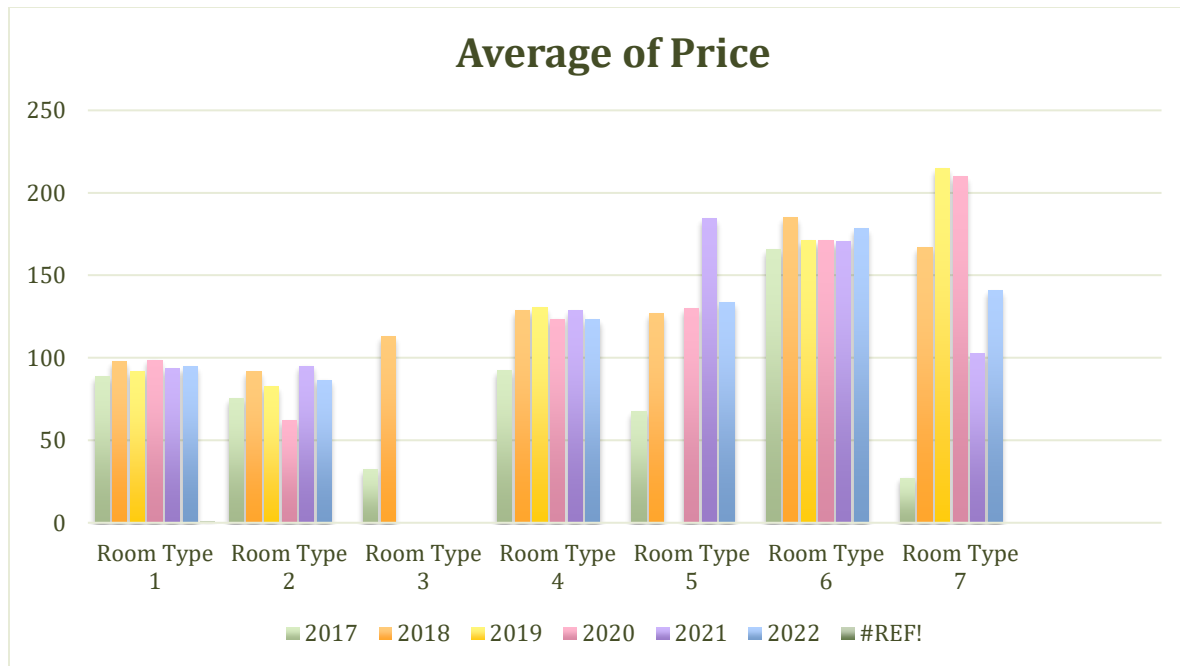
3	Average of Average price	Arrival_year ▾						
4	Room_type_reserved ▾	2017	2018	2019	2020	2021	2022	Grand Total
5	Room_Type 1	88.62577811	97.82259203	91.81365672	98.62090323	93.4836129	94.90253419	95.9185322
6	Room_Type 2	75.04867133	91.52233591	82.40333333	62	95	86.40230769	87.8485549
7	Room_Type 3	32.5	112.6875				0	73.6785714
8	Room_Type 4	92.5701165	128.59795	130.4774286	123.5152778	128.4182143	122.9356379	125.287317
9	Room_Type 5	67.51666667	126.9686522		129.88	184.665	133.7221429	123.733623
10	Room_Type 6	165.4033594	185.2780531	170.895	171.1	170.45	178.1208333	182.212836
11	Room_Type 7	27.10416667	166.9500746	214.76	209.665	102.5	140.8114286	155.198291
12	Grand Total	90.0498541	106.4527274	101.7812994	105.339899	101.2333333	101.499242	103.423539

enrage prices of the Room types based on the specific years.

To do this the following steps are required:

- First the analyst should create a pivotal table by following the steps for the Pivotal Table mentioned above.
- Drag the Arrival Year option to the row area.
- Drag the Room Reserved Type to the column area.
- Click on the Sum of Average Price cell on top left corner.
- Select Average

The average prices based on the specific years will be calculated easily.



	2017	2018	2019	2020	2021	2022
Room Type 1	88.62577811	97.82259203	91.81365672	98.62090323	93.4836129	94.90253419
Room Type 2	75.0487133	91.52233591	82.40333333	62	95	86.40230769
Room Type 3	32.5	112.6875				0
Room Type 4	92.5701165	128.59795	130.4774286	123.5152778	128.4182143	122.935679
Room Type 5	67.51666667	126.9686522		129.88	184.665	133.7222143
Room Type 6	165.4033594	185.2780531	170.895	171.1	170.45	178.1208333
Room Type 7	27.10416667	166.9500746	214.76	209.665	102.5	140.8114286

Analyzing the average price (room reservation price) for each year from 2017 to 2022 reveals interesting insights in the Hotel Reservation dataset.

In 2017, the average prices varied across different room types. Room Type 6 had the highest average price of \$165.40, followed by Room Type 4 with \$92.57. Room Type 3 had the lowest average price of \$32.50. The overall average price for the year was \$90.04.

Moving to 2018, the average prices increased compared to the previous year. Room Type 6 remained the most expensive with an average price of \$185.27, while Room Type 7 had the second-highest average price of \$166.95. The overall average price for the year was \$106.45.

In 2019, the average prices varied again, with Room Type 7 being the most expensive at \$214.76. Room Type 6 followed closely with an average price of \$170.89. Room Type 3 and Room Type 5 had missing average price values. The overall average price for the year was \$101.78.

In 2020, the average prices showed fluctuations across the different room types. Room Type 6 remained the most expensive at \$171.10, while Room Type 2 had the lowest average price of \$62.00. Room Type 3 had a missing average price value. The overall average price for the year was \$105.33.

Looking at 2021, the average prices remained relatively consistent compared to the previous year. Room Type 5 had the highest average price at \$184.66, followed by Room Type 6 with \$170.45. Room Type 3 had a missing average price value. The overall average price for the year was \$101.23.

Moving to 2022, the average prices showed minor fluctuations. Room Type 6 remained the most expensive at \$178.12, while Room Type 2 had the lowest average price of \$86.40. Room Type 3 had a missing average price value. The overall average price for the year was \$101.49.

By analyzing the average prices over these years, the data analyst can identify trends and patterns in room reservation prices. This information can help inform pricing strategies and decision-making within the hotel industry.

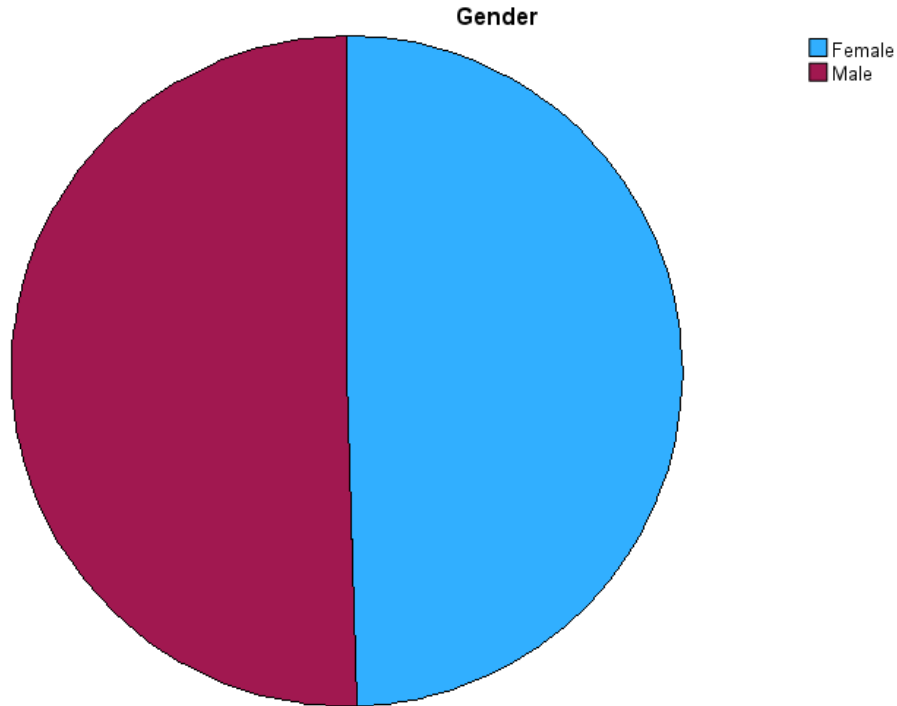
Part 02:

Number of Students by Gender

To find the number of male and female students in the Student_Score dataset using SPSS, the following steps can be performed:

1. Open SPSS and click on "File" and then "Open" to open the Student_Score.xls or Student_Score.csv file.
2. Select "Read text data with known locations" if opening the csv file or select "Read Excel (xls, xlsx) data" if opening the xls file.
3. Click on "OK" and select the file to import.
4. In the "Variable View" tab, you should see the columns of the dataset.
5. To get the count of male and female students, click on "Analyse" in the menu bar and select "Descriptive Statistics" and then "Frequencies".
6. Select the "Gender" column and drag it to the "Variables" box.
7. Click on "OK" and SPSS will display the frequency count of male and female students separately like the following figure.

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	



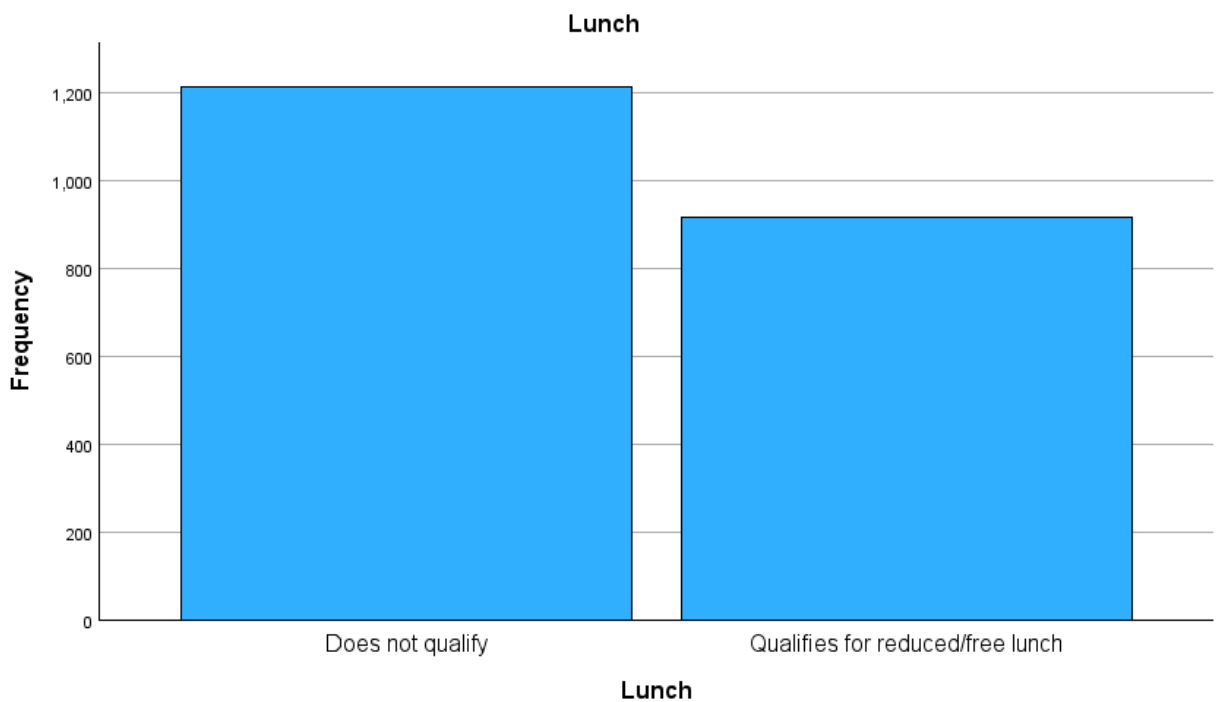
The figure shows that there are 1056 Female students among 2133 students. The percentage of the female students is 49.5%. Also, the figure shows that there are 1077 male students among the 2133 and the percentage of them is 50.5%.

Qualify and Does not Qualify

To compute the number of students, qualify and does not qualify for reduced/free lunch, the analyst can calculate by performing the following steps:

- Open SPSS and import the Student_Score dataset (Student_Score.xls or Student_Score.csv) by going to "File" > "Open" > "Data."
- To get the count of male and female students, click on "Analyse" in the menu bar and select "Descriptive Statistics" and then "Frequencies".
- Select the "Lunch" column and drag it to the "Variables" box.
- Click on "OK" and SPSS will display the frequency count of "Does not qualify and Qualifies for the reduce/free lunch" separately like the following figure.

		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	



Upon analyzing the Student Score dataset in SPSS, it was found that out of the total number of students, 1215 students (57%) do not qualify for reduced/free lunch, while 918 students (43%) qualify for reduced/free lunch.

These findings provide valuable insights into the lunch program eligibility of the students in the dataset. It indicates that a significant portion of the students, approximately 57%, do not qualify for reduced/free lunch. On the other hand, around 43% of the students meet the eligibility criteria for reduced/free lunch.

The status of the students' socio-economic may be influenced by being eligible or ineligible for reduced/free lunch. Students from families with lower income levels are often the ones who receive qualification for reduced/free lunch. However, individuals who are ineligible may hail from families with comparatively greater earnings.

Resource allocation, support systems, and targeted interventions can be subject to informed decision-making by educational institutions as well as policymakers. One can understand the distribution of lunch program eligibility among students. Equal opportunities for all students can be ensured by considering socio-economic factors in educational planning.

Mean and Median

Descriptives				
	Lunch		Statistic	Std. Error
Pretest	Does not	Mean	62.29	.309
		Median	62.00	
	Qualifie	Mean	45.25	.343
		Median	44.00	
Posttest	Does not	Mean	74.38	.324
		Median	74.00	
	Qualifie	Mean	57.48	.366
		Median	55.00	

After properly analyzing the Student Score dataset, it was found that students who do not qualify for reduced/free lunch have a higher mean pretest score of 62.29 compared to students who qualify, with a mean pretest score of 45.25. The median pretest scores for both groups were 62 and 44, respectively.

In terms of posttest scores, students who do not qualify for reduced/free lunch have a higher mean score of 74.38 compared to students who qualify, with a mean posttest score of 57.48. The median posttest scores for both groups were 74 and 55, respectively.

These results imply that lunch eligibility could be related to academic success, as seen in the pretest and posttest scores. Academic achievement may be influenced by receiving reduced or free lunches, as students who do not qualify for them usually demonstrate higher mean scores in both pretest and posttest assessments.

Students' lunch eligibility status may create differences in educational opportunities and available resources. The contrast of mean scores in regard to lunch eligibility groups recommends this. Taking into account and addressing these disparities is necessary to ensure equal educational outcomes for all students.

Differences of the Mean and Median

Descriptives

		Statistic	Std. Error
Difference	Mean	-12.1463	.09392
	Median	-12.0000	

After analyzing the Student Score dataset, the students' pre-test and post-test scores revealed a mean difference of -12.14. Scores on average declined from the pre-test to the post-test assessment. At -12, the median difference indicates where the middle value of the distribution lies.

Mean and median differences being negative imply that the scores of students decreased on average between pre-test and post-test. Possible factors behind this could include insufficient progress, challenges in grasping the evaluated material, or encountering difficulties in the learning process. These factors could be combined. To address the specific needs of students experiencing score declines, educators can develop personalized instructional strategies and interventions by analyzing score change patterns. However, individual student performance may vary significantly, and further analysis considering other variables in the dataset could provide a more comprehensive understanding of the factors influencing score changes.

Overall, the negative mean and median differences between pre-test and post-test scores suggest the need for targeted support and interventions to help students improve their academic performance and ensure their progress throughout the learning process.

K-Means Clusters

Initial Cluster Centers				Final Cluster Centers			
	Cluster				Cluster		
	1	2	3		1	2	3
Pretest	23	93	54	Pretest	40	72	57
Posttest	35	100	72	Posttest	52	85	69

The initial cluster centers represent the average pre-test and post-test scores for each cluster at the beginning of the analysis. Cluster 1 had an initial pre-test score center of 23 and an initial post-test score center of 35. Cluster 2 had an initial pre-test score center of 93 and an initial post-test score center of 100. Cluster 3 had an initial pre-test score center of 54 and an initial post-test score center of 72. After the clustering algorithm converged, the analyst obtained the final cluster centers. Cluster 1 had a final pre-test score center of 40 and a final post-test score center of 52. Cluster 2 had a final pre-test score center of 72 and a final post-test score center of 85. Cluster 3 had a final pre-test score center of 57 and a final post-test score center of 69.

Data Mining and Text Mining Method

The 5 most common data mining and text mining methods for each that can be used in the contemporary business is explained below:

1. **Classification:** To assign predefined classes or categories to data instances, Data mining technique called Classification is utilized (Vijayarani, Ilamathi and Nithya, 2015). Historical customer data can be analyzed and classified as either churn or non-churn by a marketing team using classification techniques. Classification is a

powerful data mining technique that assigns predefined classes or categories to data instances. It enables organizations to analyze large datasets and identify patterns that can help them make informed decisions. For instance, marketing teams can use classification to predict customer churn by analyzing historical customer data. By classifying customers as either churn or non-churn, organizations can design targeted marketing campaigns to retain their customers. However, it's crucial to use accurate and robust classification algorithms that can handle complex datasets. Ultimately, effective classification can help businesses improve customer satisfaction, reduce costs, and increase profitability.

2. **Regression:** Regression is a technique used to analyze the relationship between variables and predict numerical values (Zhou et al., 2016). In finance, regression analysis can be used to predict stock prices based on historical data and market indicators. A stockbroker can employ regression models to estimate future stock prices and make informed investment decisions.
3. **Clustering:** To analyze the connection between variables and predict numerical values, regression is a technique utilized. Regression analysis can help predict stock prices in finance by looking at past data and market indicators (Hashimi, Hafez and Mathkour, 2015). To have better insight into future stock prices, a stockbroker can use regression models, thus making informed investment decisions.
4. **Text Classification:** The automated categorization of textual data into predefined categories or classes is referred to as text classification (Sun et al., 2018). Text classification allows news organizations to automatically sort news stories into different categories like sports or politics.
5. **Sentiment Analysis:** Assessing sentiments and opinions in textual data relying on text mining techniques is what Sentiment Analysis entails (Goh Ubeynarayana, 2017). This information can help them identify areas for improvement and address negative sentiments to enhance the overall customer experience.

Advantages and Disadvantages of Excel and SPSS

The advantages and disadvantages of Excel as well as SPSS has been provided in the table below:

	Excel	SPSS
Advantages	1. Comprehensive statistical analysis capabilities, including advanced statistical models and tests.	1. Familiar and widely-used spreadsheet software, easily accessible to most users.
	2. User-friendly interface with intuitive menus and syntax options (Verma, 2015).	2. Powerful data manipulation and visualization features, such as pivot tables and charts.
	3. Efficient handling of large datasets and complex analyses.	3. Integration with other Microsoft Office tools for seamless workflow (Kupzyk and Cohen, 2015).
	4. Supports data mining and machine learning techniques.	4. Extensive formula library for complex calculations and custom functions.
	5. Provides robust data validation and cleaning tools.	5. Allows for customization and automation using VBA macros.
Disadvantages	1. Expensive software with licensing costs, limiting accessibility for some users.	1. Limited statistical analysis capabilities compared to specialized software like SPSS (Kim, Mallory and Valerio, 2020).
	2. Steeper learning curve for complex analyses and syntax-based operations.	2. Performance issues with large datasets and memory-intensive operations.
	3. Limited data visualization options, requiring additional tools for advanced visualizations.	3. Less efficient for advanced statistical modeling and data mining tasks.
	4. Limited support for unstructured data and text mining (Abbott, 2016).	4. Lack of built-in tools for data cleaning and validation, requiring manual processes.
	5. Primarily focused on statistical analysis, with limited general spreadsheet capabilities.	5. Lack of collaboration features, making simultaneous team analysis challenging.

Conclusion

The report emphasizes the importance of data analysis and visualization in contemporary business and educational contexts. By leveraging these techniques, businesses can gain valuable insights to make informed decisions, optimize operations, and improve customer satisfaction. Educational institutions can identify factors impacting student performance and implement targeted interventions to support student success. In conclusion, the analysis presented in this report demonstrates the power of data analysis and visualization in extracting meaningful information from datasets. The discoveries act as a groundwork for making decisions supported by evidence and demonstrate the potential of using data-driven tactics in different areas. The continued use and exploration of these techniques will undoubtedly advance businesses and educational institutions in a data-driven world.

References

- Abbott, M.L. (2016). Using statistics in the social and health sciences with SPSS and excel. John Wiley & Sons.
- Goh, Y.M. and Ubeynarayana, C.U. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention*, 108, pp.122-130.
- Hashimi, H., Hafez, A. and Mathkour, H. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior*, 51, pp.729-733.
- Kim, M., Mallory, C. and Valerio, T. (2020). Statistics for evidence-based practice in nursing. Jones & Bartlett Publishers.
- Kupzyk, K.A. and Cohen, M.Z. (2015). Data validation and other strategies for data entry. *Western journal of nursing research*, 37(4), pp.546-556.
- Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S. and Wang, G (2018). Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering*, 2018.
- Verma, J.P. (2015). Repeated measures design for empirical researchers. John Wiley & Sons.
- Vijayarani, S., Ilamathi, M.J. and Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1), pp.7-16.
- Zhou, Y., Tong, Y., Gu, R. and Gall, H. (2016). Combining text mining and data mining for bug report classification. *Journal of Software: Evolution and Process*, 28(3), pp.150-176.