

Assessment 2
Data Analysis and Visualisation
By Ibrahim Gazi

BABM – LCBB5000

Lecturer: Amitave Banik

Student ID: 2018574

UWTSD - London

September 2023

Table of Contents

Part 1: Hotel Reservation Data Analysis and Visualisation	3
Introduction	3
Excel: Data Pre-Processing.....	4
First Step.....	4
Second Step.....	5
Third Step.....	5
Fourth Step	6
Excel: Data Analysis.....	7
IF Formula.....	7
PIVOT Table	8
First step	9
Second step	9
Third Step.....	9
Fourth Step	10
Excel: Visualisation and Interpretation of the Data.....	12
Part 2: Student Scores Data Analysis and Visualisation	13
Introduction	13
Distribution of Gender.....	14
Qualified and Disqualified Students for Reduced or Free Lunch.....	15
Mean and Median Value for Qualified and Disqualified Students.....	15
Mean and Median Differences of Pre-Test and Post-Test Scores	16
Clustering	17
Data and Text Mining Methods.....	18
Using SPSS and Excel.....	19
Conclusion	21
References.....	22

Part 1: Hotel Reservation Data Analysis and Visualisation

Introduction

The Hotel Reservation dataset is downloaded and available in spreadsheet format as Excel file. The Dataset consists of the booking details of the arriving guests at the hotel from 2017 to 2021 – 5 years. According to the reservation sheet, different variables present in the spreadsheet include booking status, booking ID, average price, special requests, demography, lead time, arrival, meal type, parking required, room type, repetition, and market type. Derrfuss (2023) asserts that Microsoft Excel is significant for the hotel reservation dataset to pre-process, review and visualise the data. This report, thus, explores the application of formulas and functions in considering the use of IF and PIVOT tables to analyse the increase or decline in average price over five years, i.e., 2017 to 2021; to solve and visualise the Hotel Reservation data analysis.

Excel: Data Pre-Processing

According to Derrfuss (2023), data pre-processing is a technique that utilises raw data to ensure its acceptability for review. Some of the steps for Data pre-processing include Data Cleaning, Data Integration, Data Transformation, Data Reduction, Data Discretisation and Data Normalisation (Han, Zhang and Qiao, 2019). This way, data pre-processing is essential in the Hotel reservation data analysis for any analysis – since the raw data must be prepared for analysis with no faults.

First Step

The first step is establishing the original Dataset – where data ready for processing is gathered and located. See **Figure 1** below.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Booking ID	Adults No	Children No	Weekend Nights	Weekday nights	Meal_plan	Required_car_parking_space	Room_Type_reserved	Lead_time	Arrival_year	Arrival_month	Arrival_date	Market_type	Repeated_guest	Previous Cancellation	Previous bookings	Average price	Special requests	Booking status
2	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	10	2	Offline	0	0	0	65	0	Not_Canceled
3	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	2018	11	6	Online	0	0	0	106.68	1	Not_Canceled
4	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018	2	28	Online	0	0	0	60	0	Canceled
5	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018	5	20	Online	0	0	0	100	0	Canceled
6	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	2018	4	11	Online	0	0	0	94.5	0	Canceled
7	INN00006	2	0	0	2	Meal Plan 2	0	Room_Type 1	346	2018	9	13	Online	0	0	0	115	1	Canceled
8	INN00007	2	0	1	3	Meal Plan 1	0	Room_Type 1	34	2017	10	15	Online	0	0	0	107.55	1	Not_Canceled
9	INN00008	2	0	1	3	Meal Plan 1	0	Room_Type 4	83	2018	12	26	Online	0	0	0	105.61	1	Not_Canceled
10	INN00009	3	0	0	4	Meal Plan 1	0	Room_Type 1	121	2018	7	6	Offline	0	0	0	96.9	1	Not_Canceled
11	INN00010	2	0	0	5	Meal Plan 1	0	Room_Type 4	44	2018	10	18	Online	0	0	0	133.44	3	Not_Canceled
12	INN00011	1	0	1	0	Not Selected	0	Room_Type 1	0	2018	9	11	Online	0	0	0	85.03	0	Not_Canceled
13	INN00012	1	0	2	1	Meal Plan 1	0	Room_Type 4	35	2018	4	30	Online	0	0	0	140.4	1	Not_Canceled
14	INN00013	2	0	2	1	Not Selected	0	Room_Type 1	30	2018	11	26	Online	0	0	0	88	0	Canceled
15	INN00014	1	0	2	0	Meal Plan 1	0	Room_Type 1	95	2018	11	20	Online	0	0	0	90	2	Canceled
16	INN00015	2	0	0	2	Meal Plan 1	0	Room_Type 1	47	2019	10	20	Online	0	0	0	94.5	2	Not_Canceled
17	INN00016	2	0	0	2	Meal Plan 2	0	Room_Type 1	256	2018	6	15	Online	0	0	0	115	1	Canceled
18	INN00017	1	0	1	0	Meal Plan 1	0	Room_Type 1	0	2017	10	5	Offline	0	0	0	96	0	Not_Canceled
19	INN00018	2	0	1	3	Not Selected	0	Room_Type 1	1	2017	8	10	Online	0	0	0	96	1	Not_Canceled
20	INN00019	2	0	2	2	Meal Plan 1	0	Room_Type 1	99	2017	10	30	Online	0	0	0	65	0	Canceled
21	INN00020	2	0	1	0	Meal Plan 1	0	Room_Type 1	12	2019	10	4	Offline	0	0	0	72	0	Not_Canceled
22	INN00021	2	0	2	2	Meal Plan 1	0	Room_Type 1	99	2017	10	30	Online	0	0	0	65	0	Canceled
23	INN00022	1	0	0	1	Meal Plan 1	0	Room_Type 1	122	2019	11	25	Corporate	0	0	0	67	0	Not_Canceled
24	INN00023	2	0	2	4	Meal Plan 1	0	Room_Type 1	2	2018	3	20	Offline	0	0	0	85	0	Not_Canceled
25	INN00024	2	0	0	3	Meal Plan 1	0	Room_Type 1	37	2018	10	13	Offline	0	0	0	105	0	Not_Canceled
26	INN00025	2	0	2	1	Not Selected	0	Room_Type 1	130	2018	5	22	Online	0	0	0	94.5	1	Not_Canceled

Figure 1: Hotel Reservation Dataset

Second Step

The next step would be to remove cell formatting models that may need to be more helpful. The reason for eliminating unnecessary cell formatting is to select the entire Dataset – this way, we click the AutoFit Row Height and AutoFit Column Width. This option is present in the Format section, under Home Tab. Consequently, specific columns and rows extract into isolated sheets for further analysis. See **Figure 2** below.

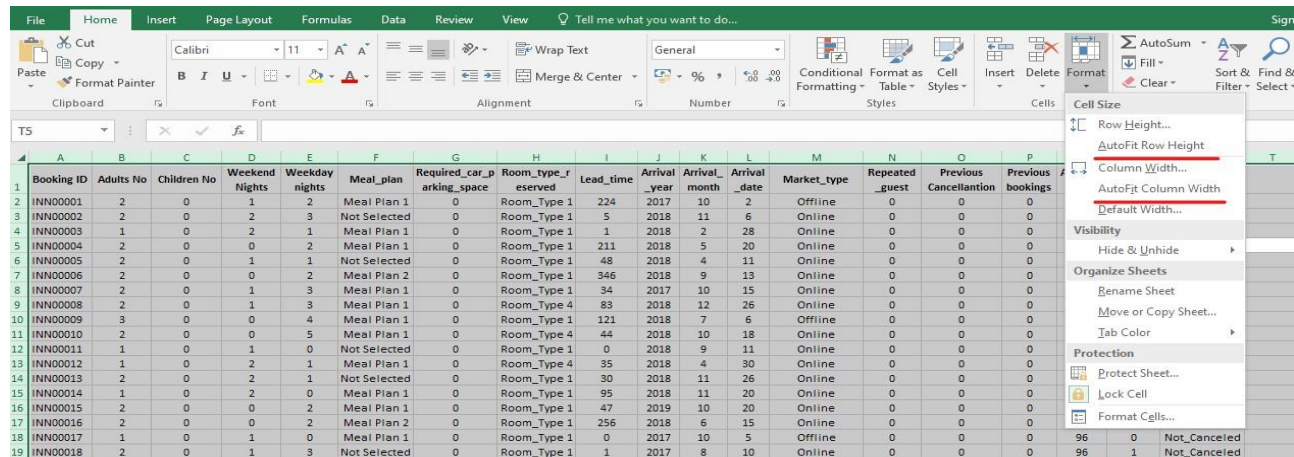


Figure 2: Format Cells

Third Step

By selecting the header row, we then clicked Sort and Filter option available in the Home Tab – to ensure recognition of data inconsistencies with each of the specific columns. Sample columns for analysis purposes according to **Figure 3** below include Average Year and Average Price.

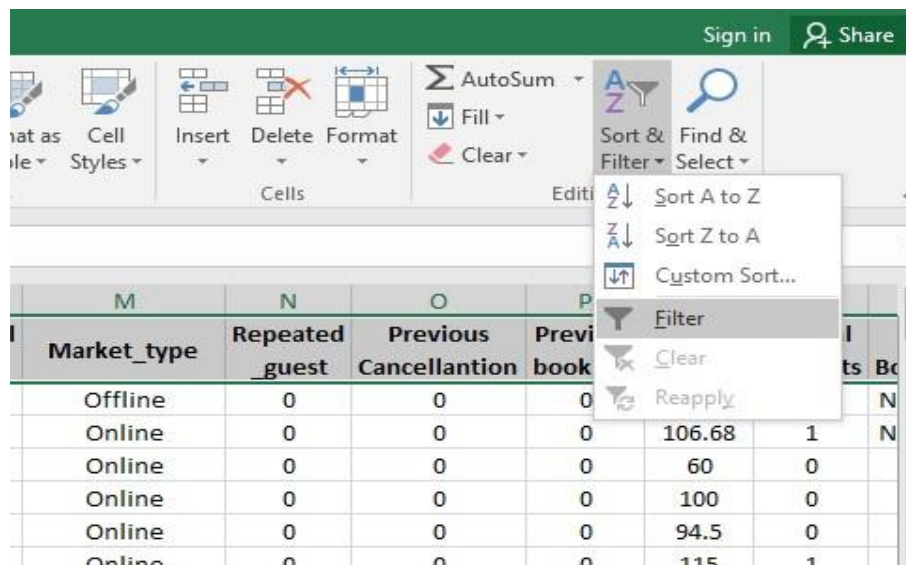


Figure 3: Sort and Filter

See **Figure 4** and **5 below**, which shows Filtering of the Average Price and Arrival Year columns respectively.

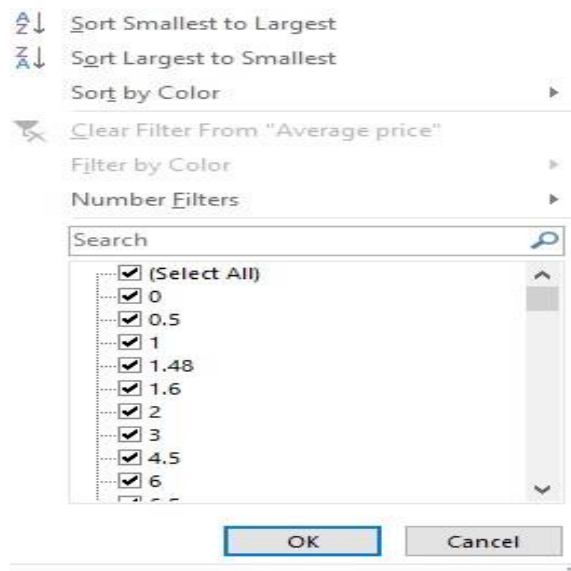


Figure 4: Filtering the Column for Average Price

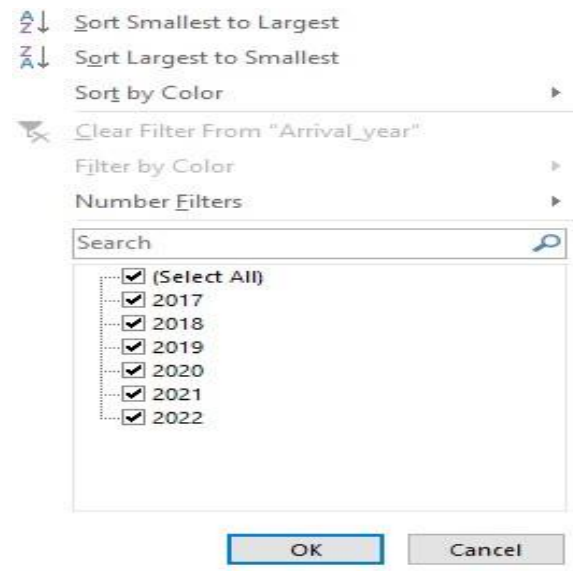


Figure 5: Filtering Arrival Year Column

Fourth Step

The spreadsheet confirms that missing values exist in the Average Price; the missing values need to be eliminated. To do this, as confirmed in **Figure 6** below, we selected rows for null values only i.e., from the table — in this case, only the rows in which the price is zero and collectively eliminating those rows once. As a result, **Figure 7** shows the values deleted for the average price that consists of null values.

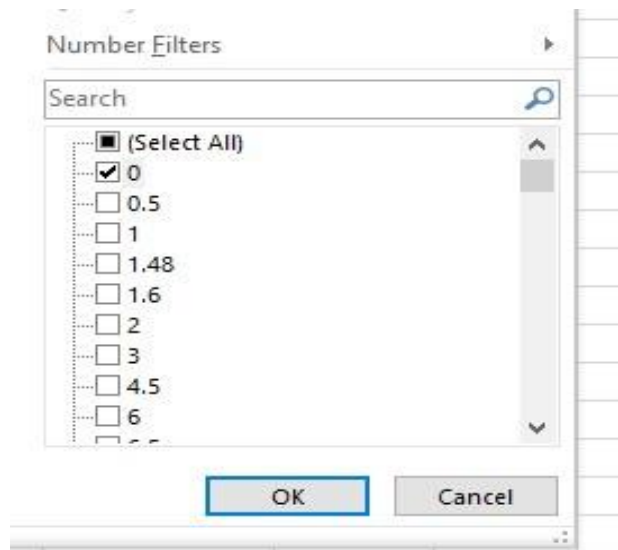


Figure 6: Application of the Drop-Down List for the Column on Price in Selecting Null Values

	L	M	N	O	P	Q
	Arrival date	Market type	Repeated guests	Previous Cancellatio	Previous booking	Average price
10	10	omplementar	0	0	0	0
1	1	omplementar	1	3	5	0
27	27	omplementar	0	0	0	0
1	1				1	0
2	2				0	0
7	7				3	0
2	2				4	0
2	2				0	0
1	1				0	0
19	19	omplementar	0	0	0	0
23	23	omplementar	0	0	0	0
19	19	omplementar	1	0	1	0
15	15	omplementar	1	0	1	0
1	1	omplementar	1	0	1	0
6	6	omplementar	1	2	7	0
3	3	omplementar	0	0	0	0
13	13	omplementar	0	0	0	0
20	20	omplementar	1	0	2	0
10	10	Online	0	0	0	0
15	15	omplementar	0	0	0	0
21	21	omplementar	0	0	0	0

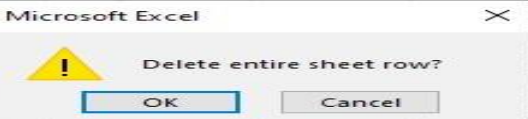


Figure 7: Deleting Rows for Average Price Consisting of Null Values

Excel: Data Analysis

IF Formula

For purposes of analysis, the needed data points for assessment are extracted and inserted in a structured method – this way, data points on the average year and arrival year are extracted from the whole table and inserted in a structured method in the table. Then, the conditional Formula is used to extract each year's overall value. In preparing the new Dataset, **Figure 8** below shows the application of the condition formula i.e., SUMIF (). This way, the syntaxes “A: A” and “B: B”; “lookup range” and “return range”, respectively, show the application of the conditional Formula IF. See **Figure 8** below, which depicts SUMIF (A: A, "2017", B: B).

	A	B	C	D	E	F	G	H
	Arrival _year	Average price			Arrival_year	Sum of Average price		
1	2017	65			2017	=SUMIF(A:A, "2017", B:B)		
2	2018	106.68			2018			
3	2018	60			2019			
4	2018	100			2020			
5	2018	94.5			2021			
6	2018	115						
7	2017	107.55						
8	2018	105.61						

Figure 8: Extracting Necessary Fields

Figure 9 below shows the complete Dataset that has been worked on – for further evaluation: the Dataset comprises the Arrival year and Sum of Average price.

Arrival_year	Sum of Average price
<u>2017</u>	549304.11
<u>2018</u>	2984721.57
<u>2019</u>	18015.29
<u>2020</u>	20857.3
<u>2021</u>	19436.8

Figure 9: Conditional Formula in Dataset Analysis

PIVOT Table

According to Excel chat (2019), a PIVOT table condenses data automatically – thus makes it useful in analysing massive data in worksheets and databases. Therefore, the PIVOT table simplifies complex data because of its feature, such as distinct powerful elements and flexibility rate (Večeřa, 2020). The powerful element's convert's complex data into tables, making data analysis and interpretation easy. Firstly, it is important to organise data in columns at the start of the PIVOT Table – then, group the information into distinct categories to display a new table. This way, PIVOT tables play important role in scenarios where the Dataset is huge – and is difficult to interpret in the raw format. With the help of different elements and their flexibility rate, it is possible to transform the hotel reservation dataset into a simpler format.

This way, the raw data can easily be displayed in the PIVOT table framework. Working Data (2020) asserts that the PIVOT table enables users to gain freedom on the field they need to display in columns and rows — thus realising the application of various aggregations such as minimum, maximum, average, and total. PIVOT tables in Excel help analysts report massive data in the Dataset in a simplified, summarised, and condensed manner – considering the required data and their respective fields (Assessment and Curriculum Support Centre, 2022). In the case scenario, the Hotel reservation dataset consists of many records that must be properly structured for analysis. This way, PIVOT characteristics would help achieve analysis objective as pre-defined in the steps below:

First step

We select the entire Dataset that must be simplified, condensed, and summarised for further analysis. See **Figure 10** below.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Booking ID	Adults No	Children No	Weekend Nights	Weekday nights	Meal_plan	Required_car parking_space	Room_type_reserved	Lead_time	Arrival_year	Arrival_month	Arrival_date	Market_type	Repeated_guest	Previous Cancellation	b
2	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	10	2	Offline	0	0	
3	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	2018	11	6	Online	0	0	
4	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018	2	28	Online	0	0	
5	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018	5	20	Online	0	0	
6	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	2018	4	11	Online	0	0	
7	INN00006	2	0	0	2	Meal Plan 2	0	Room_Type 1	346	2018	9	13	Online	0	0	
8	INN00007	2	0	1	3	Meal Plan 1	0	Room_Type 1	34	2017	10	15	Online	0	0	
9	INN00008	2	0	1	3	Meal Plan 1	0	Room_Type 4	83	2018	12	26	Online	0	0	
10	INN00009	3	0	0	4	Meal Plan 1	0	Room_Type 1	121	2018	7	6	Offline	0	0	
11	INN00010	2	0	0	5	Meal Plan 1	0	Room_Type 4	44	2018	10	18	Online	0	0	
12	INN00011	1	0	1	0	Not Selected	0	Room_Type 1	0	2018	9	11	Online	0	0	
13	INN00012	1	0	2	1	Meal Plan 1	0	Room_Type 4	35	2018	4	30	Online	0	0	
14	INN00013	2	0	2	1	Not Selected	0	Room_Type 1	30	2018	11	26	Online	0	0	
15	INN00014	1	0	2	0	Meal Plan 1	0	Room_Type 1	95	2018	11	20	Online	0	0	
16	INN00015	2	0	0	2	Meal Plan 1	0	Room_Type 1	47	2019	10	20	Online	0	0	
17	INN00016	2	0	0	2	Meal Plan 2	0	Room_Type 1	256	2018	6	15	Online	0	0	
18	INN00017	1	0	1	0	Meal Plan 1	0	Room_Type 1	0	2017	10	5	Offline	0	0	
19	INN00018	2	0	1	3	Not Selected	0	Room_Type 1	1	2017	8	10	Online	0	0	
20	INN00019	2	0	2	2	Meal Plan 1	0	Room_Type 1	99	2017	10	30	Online	0	0	

Figure 10: Selection of the Necessary Columns and Rows within the Dataset

Second step

Selection of the Pivot table option – available under the Insert Tab (first option on the far left) of the Excel architecture. See **Figure 11** below, which shows how to insert a Pivot table.

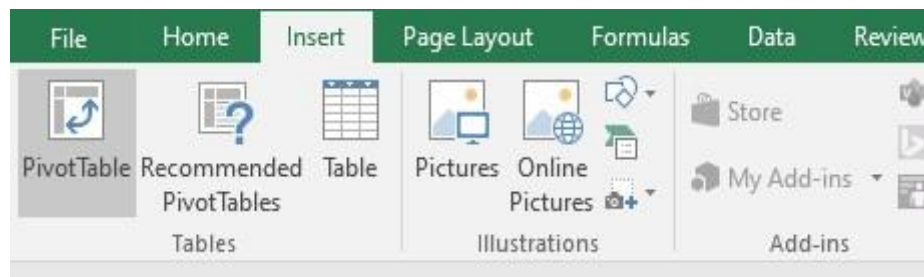


Figure 11: Pivot Table under Insert Table

Third Step

After utilising the Pivot table option under the Insert section, a new window labelled “Create Pivot table” emerges. Thus, the window needs details concerning the table range and whether the PIVOT table would be inserted in a new Excel worksheet. See **Figure 12** below on the selected parameters that enabled the creation of the PIVOT Table.

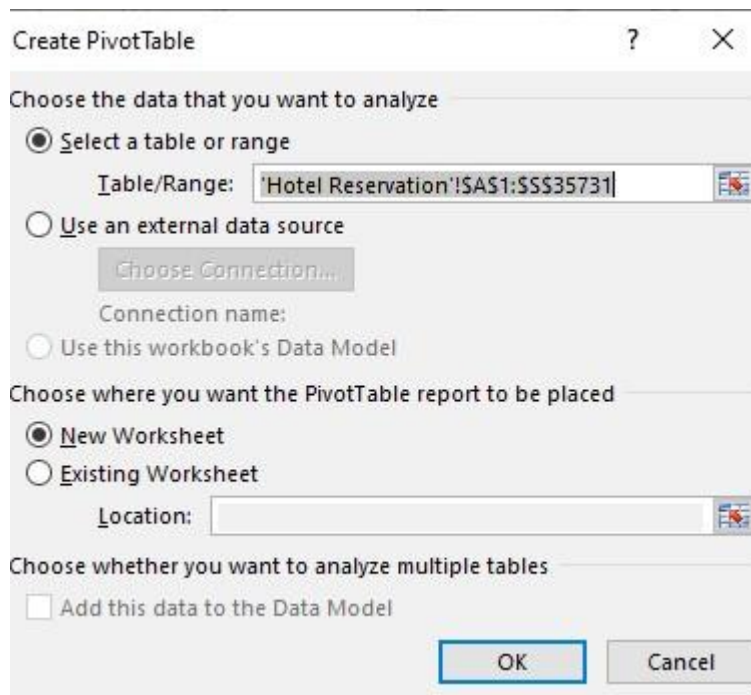


Figure 12: Creating a Pivot Table

Fourth Step

The new worksheet will emerge where the PIVOT table pane was extracted. The University of Michigan Library (2023) — confirm that with the help of the checklist provided beside each column name, the step resumes selecting all the necessary fields significant for analysis. Loffing (2022) adds that there is a reason to select the required conditions. For this specific scenario, the required elements include Average Price and Arrival Year columns. By extracting the needed values from the pivot pane, it is possible to create a structured table similar to the one in Figure 9 (as previously mentioned). It is also important to note that the Arrival year must be dragged into the filter column to measure individual results. **Figures 13, 14 and 15 (a) and (b)** shown below on inserting arrival year, inserting average price field, and dragging values in the needed panels, respectively, depict the complete procedure. **Figure 15 (b)** shows the use of PIVOT Tables to extract values.

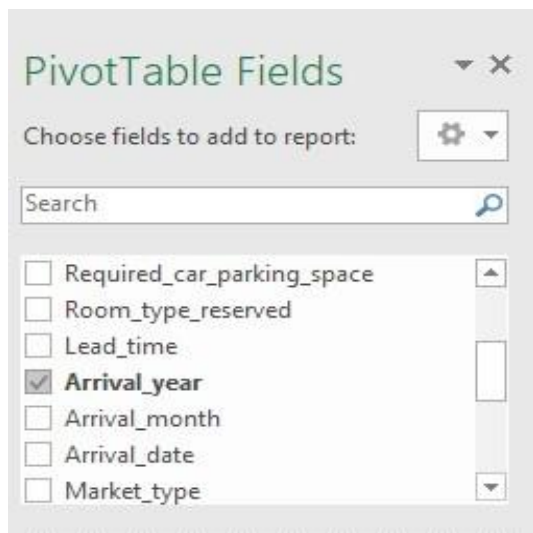


Figure 13: Inserting Arrival Year

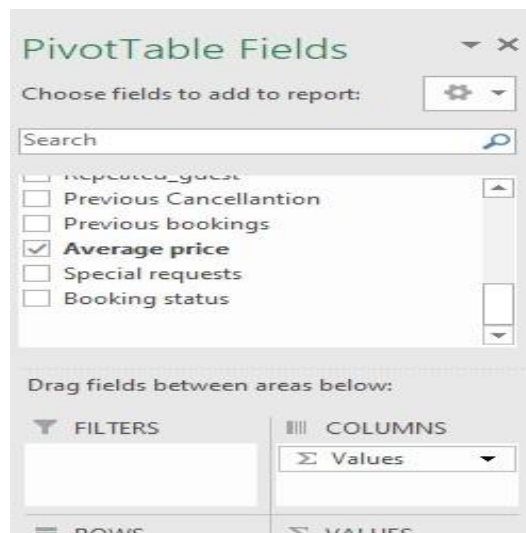


Figure 14: Inserting Average Price Field

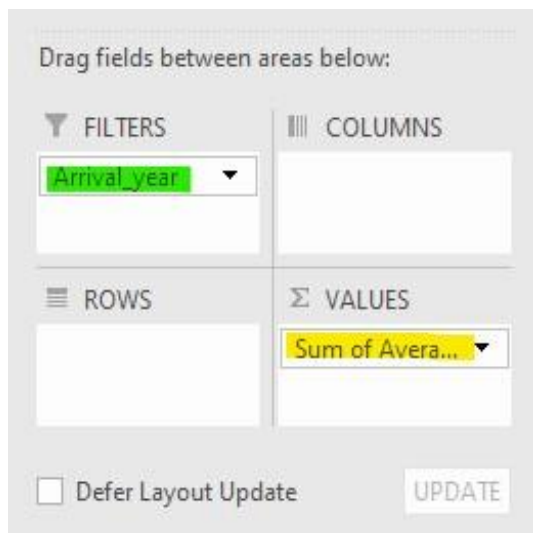


Figure 15 (A): Dragging Values in the Needed Pages

E	F
Arrival_year	Sum of Average Price
2017	549304.11
2018	2984721.57
2019	18015.29
2020	20857.3
2021	19436.8

Figure 15 (B): Use of Pivot Tables to Extract Values

Excel: Visualisation and Interpretation of the Data

Estimated data in the prior section of the analysis i.e., data analysis using the IF formula and PIVOT, as Bordelon (2021) asserts, is depicted as shown in **Chart 1** below, consisting of a line chart for purposes of data visualisation. The major solution that visualisation needs to address is whether there has been a deprecation or escalation of the total average price from 2017 to 2021.

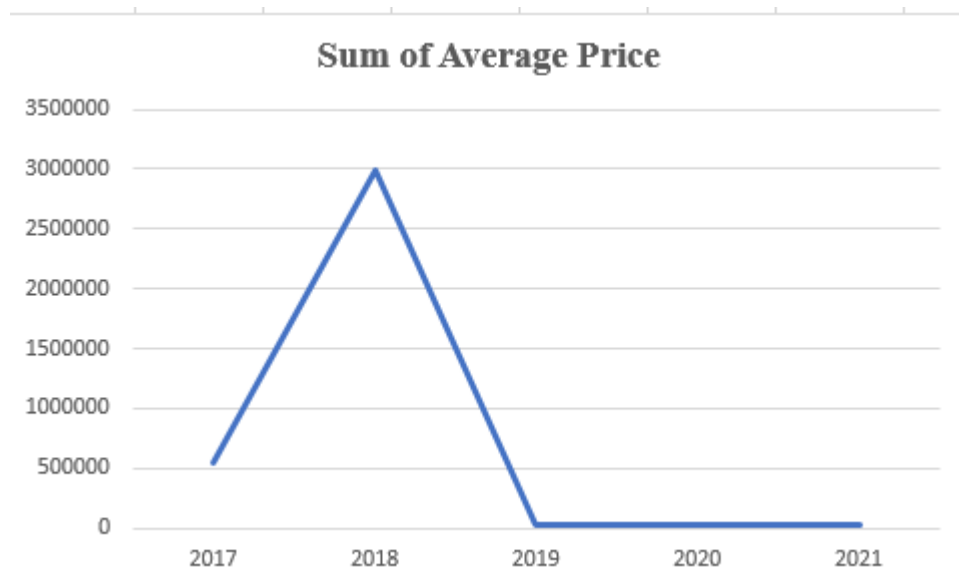


Chart 1: Line Depiction of the Average Price Pattern from 2017 To 2021

From Chart 1 above – considering the total average price, the chart establishes an original upward trend from 2017 to 2018. This way, the general growth realised during the 2017 to 2018 period is four times that of the original value. From 2018 to 2021, the total average price realised a downward pattern. The most severe downfall was recorded from 2018 to 2019 – the maximum downfall within the specified period realised a 99.3% drop. However, there was not much growth realised from 2019 to 2021. There was a minimum record of £18k in 2019 for the average price of the hotel bookings – whereas the year 2018 realised the maximum average hotel bookings price of £ 2984k. It can be concluded that the hotel reservation experienced a decline in the total average price from 2017 to 2021 by 96%.

Part 2: Student Scores Data Analysis and Visualisation

Introduction

The database extracted from the source – the “Student_Score.csv file” – comprises 2133 records, translating to 14 fields. See **Figure 1** below, which demonstrates the “Student_Score.csv” dataset in the Excel framework. This study explores the application of SPSS that aid in analysing statistical data of the student score for various scenarios attributed to reduced and free lunch. The report also explores some of the crucial data and text mining methods, advantages and disadvantages of using SPSS and Excel.

	School	School_setting	School_type	Classroom	Teaching_method	Student_No	Student_ID	Gender	Lunch	Pretest	Posttest	V12	V13	V14
1	ANKYI	Urban	Non-public	6OL	Standard	20	2FHT3	Female	Does	not	qualify	62	72	
2	ANKYI	Urban	Non-public	6OL	Standard	20	3JIVH	Female	Does	not	qualify	66	79	
3	ANKYI	Urban	Non-public	6OL	Standard	20	3XOWE	Male	Does	not	qualify	64	76	
4	ANKYI	Urban	Non-public	6OL	Standard	20	55600	Female	Does	not	qualify	61	77	
5	ANKYI	Urban	Non-public	6OL	Standard	20	74LOE	Male	Does	not	qualify	64	76	
6	ANKYI	Urban	Non-public	6OL	Standard	20	7YZO8	Female	Does	not	qualify	66	74	
7	ANKYI	Urban	Non-public	6OL	Standard	20	9KMZD	Male	Does	not	qualify	63	75	
8	ANKYI	Urban	Non-public	6OL	Standard	20	9USQK	Female	Does	not	qualify	63	72	
9	ANKYI	Urban	Non-public	6OL	Standard	20	CS5QP	Male	Does	not	qualify	64	77	
10	ANKYI	Urban	Non-public	6OL	Standard	20	D6HT8	Female	Does	not	qualify	61	72	
11	ANKYI	Urban	Non-public	6OL	Standard	20	DZMKU	Male	Does	not	qualify	61	73	
12	ANKYI	Urban	Non-public	6OL	Standard	20	FH7B9	Male	Does	not	qualify	64	74	
13	ANKYI	Urban	Non-public	6OL	Standard	20	J19VG	Male	Does	not	qualify	66	78	
14	ANKYI	Urban	Non-public	6OL	Standard	20	JQM2W	Female	Does	not	qualify	60	71	
15	ANKYI	Urban	Non-public	6OL	Standard	20	MEUC4	Female	Does	not	qualify	64	77	
16	ANKYI	Urban	Non-public	6OL	Standard	20	R4U8H	Male	Does	not	qualify	64	73	
17	ANKYI	Urban	Non-public	6OL	Standard	20	TH7KI	Male	Does	not	qualify	63	70	
18	ANKYI	Urban	Non-public	6OL	Standard	20	U1FV7	Male	Does	not	qualify	67	73	
19	ANKYI	Urban	Non-public	6OL	Standard	20	WC5I6	Female	Does	not	qualify	63	71	
20	ANKYI	Urban	Non-public	6OL	Standard	20	ZBQ4T	Male	Does	not	qualify	64	73	
21	ANKYI	Urban	Non-public	ZNS	Standard	21	0CRO6	Male	Does	not	qualify	60	68	
22	ANKYI	Urban	Non-public	ZNS	Standard	21	1QMDI	Male	Does	not	qualify	60	70	

Figure 1. Overall view of the “Student_Score.csv” data set layout

Distribution of Gender

See **Table 1** below that shows the output for gender distribution, i.e., male, and female students, regarding frequency, percent, valid percent, and cumulative percent.

Table 1: Students' Gender – Frequency Distribution Table

Statistics					
Gender					
N	Valid	2133			
	Missing	0			

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	

Consequently, the chart in **Figure 2** below shows that – out of 2133 students participating in the study, gender distribution of each gender contributes to half of the total percentage. Specifically, the rate of female students is 50.5%; and that of male students is 49.5% - thus, the female students are slightly fewer than the male students.

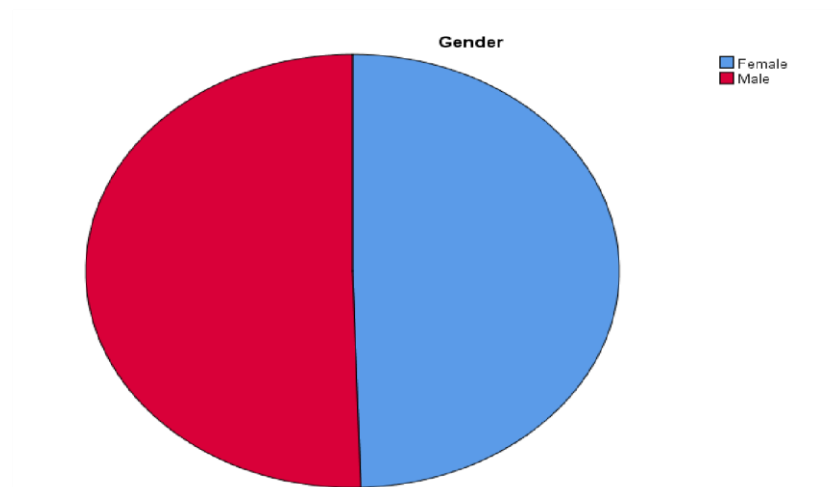


Figure 2. Gender distribution

Qualified and Disqualified Students for Reduced or Free Lunch

Table 2 below shows the output, which depicts that out of 2133 students who participated in the study, only 918 students qualified for the reduced or free lunch. The 918 qualified students represent 43% of the total cumulative per cent. On the other hand, 1215 students failed so they did not qualify for the reduced or free lunch, representing 57% of the total cumulative per cent.

Table 2: Qualification Statistics for Reduced or Free Lunch

Statistics					
Lunch					
N	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	

Mean and Median Value for Qualified and Disqualified Students

The mean value for the total number of students who qualified and failed for the reduced or free lunch reward is represented in the equation below:

$$(918 + 1215) / 2 = 1066.5$$

The median value for the total number of students who qualified and failed for the reduced or free lunch reward is represented in the equation below:

$$\frac{1}{2} (1 \text{ observation} + 2 \text{ observation}) = (918 + 1215) / 2 = 1066.5$$

Mean and Median Differences of Pre-Test and Post-Test Scores

Table 3, shown below, shows the output of the total record tests for 2133 students, depicting the differences in the mean and median values for the pre-test and post-test scores. The average pre-test score (mean) of 54.96 for the total number of students represented a median score of 56.00, translating to achievement on either side of the value on half of students' marks. On the other hand, the average post-test score (mean) of 67.10 for the total number of students represented a median score of 68.00, translating to students' achieving a higher post-test score than the previous test (pre-test).

Frequencies



Statistics			
		Pretest	Posttest
N	Valid	2133	2133
	Missing	0	0
Mean		54.96	67.10
Std. Error of Mean		.294	.303
Median		56.00	68.00

Table 3: Differences in Mean and Median of Students' Pre-Test and Post-Test Scores

Clustering

K-Means Clustering

In the case scenario, there are two categories for “k” – or the clusters selection, i.e., Z score for Pre-test and Z score for Post-test. This way, two clusters for pre-test and post-test have been discovered. **Figure 3** below demonstrates the segmentation of groups in the case scenario, with 993 students in Cluster 1 and 1140 in Cluster 2.

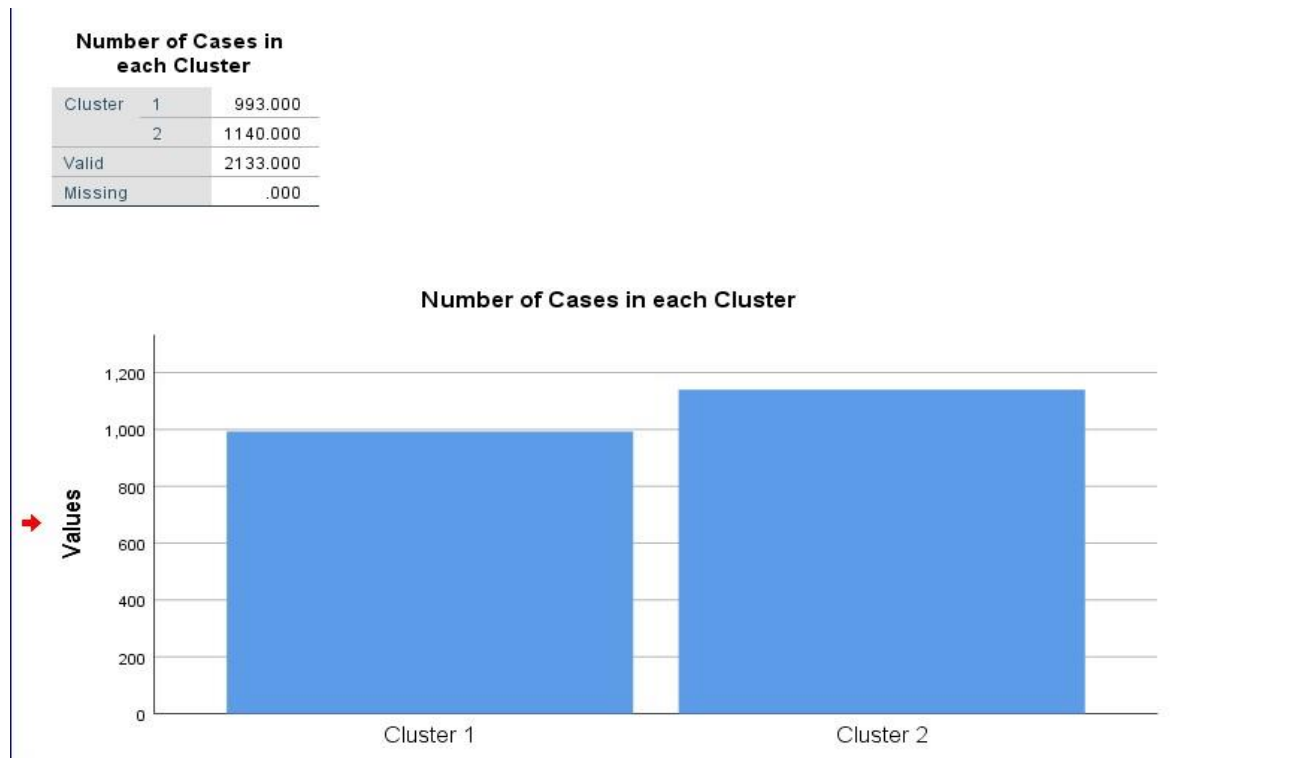


Figure 3. Cluster segmentation

Cluster centres are the points that indicate the average score of the points located at each of the clusters. **Table 4** below shows the original cluster centres for pre-test and post-test scores. Similarly, Table 5 shows the final centres for pre-test and post-test scores. For the pre-test scores, Table 5 indicates that the average score for cluster 1 is 0.87 and that of cluster 2 is 0.76. For the post-test scores, **Table 5** shows that the average score for cluster 1 is 0.88 and that of cluster 2 is 0.77.

	Cluster	
	1	2
Zscore(Pretest)	-2.35609	2.80497
Zscore(Posttest)	-2.29518	2.35206

Table 4: Original Cluster Centres

	Cluster	
	1	2
Zscore(Pretest)	-.87764	.76447
Zscore(Posttest)	-.88650	.77219

Table 5: Final Cluster Centres

Further, **Table 6** shows the statistical significance illustrated in the form of the ANOVA table between the cluster centres for each of the variables in the study. **Table 6** shows a statistical significance representing mean values significantly distinct. This way, cluster 1 contains the most dependent pre-test and post-test scores than the cluster 2 counterpart.

ANOVA						
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(Pretest)	1431.104	1	.329	2131	4351.122	.000
Zscore(Posttest)	1460.133	1	.315	2131	4631.188	.000

Table 6: Statistical Significance and the ANOVA Table

Data and Text Mining Methods

Neural Networks

According to Kaur et al. (2021) and Pedamkar (2023), neural network is a data mining method that employs 3-pillar elements, namely model, learning algorithm, and activation function to model the relationship between inputs and their corresponding outputs. According to Yates and Islam (2022), such businesses as IBM, Microsoft, and Qualcomm use neural networks to harvest user information from datasets with an objective to find data patterns.

Association

In association data mining, critical operations involve data analysts searching for connections, relationships, and correlations between data points (University of Kansas, 2020). This is important primarily when hidden suggestions are derived to display essential information. Medical facilities, such as Mayo Clinic (2022), use association rules in medical diagnoses to determine the likeliness of illness based on pre-defined symptoms and the patient's specifics on the "if-then" basis.

Categorisation

According to Benício et al. (2022), Natural Processing Language or Categorisation is a text mining method that extracts relevant abbreviations and synonyms from textual data. For Instance, Ball (2018) reports that categorisation serves as an automated process that businesses like FedEx use in personalised commercials deliveries, categorising customer web pages, and spam filtering.

Outlier Analysis

University of Baltimore (2020) asserts that outlier data mining identifies data that fails to comply with end behaviour or pattern; thus, the unexpected data items translate to noise or definition of outlier. According to Chatterjee et al. (2021), Amazon Inc. uses anomaly or outlier analysis as a security measure to detect intrusion, faults, or credit card fraud.

Clustering

Aggarwal and Reddy (2014) assert that clustering involves segmenting distinct data points with relevancy in their features. Categorising data into different subsets is a popular method to target specific users; thus, most businesses use clustering as a method in applications related to data mining. According to Big Data Tips (2016), Coca-Cola Company uses clustering analysis to segment goods based on their general popularity, local customer preferences, and flavour.

Using SPSS and Excel

According to Rahman and Muktadir (2021), SPSS is widely used in many fields, such as healthcare, social sciences, education, and economics, to easily apply statistical methods in documenting and managing data efficiently. Its main advantage is that SPSS incorporates a comprehensive system of extensive data integration, machine learning procedures, and text analysis in its sub-sections to merge files, e.g., the same subjects and different variables or different subjects and the same variables (Team TMT, 2021). Despite having useful features that help analysts to conduct statistical operations efficiently and effectively, it also has limitations, such as being expensive to purchase (Rahman and Muktadir, 2021). Nonetheless, regarding real-world application, SPSS is the most common statistical tool used (52.1%) in health sciences research as compared to SAS (12.9%) and Stata (12.6%) (Masuadi et al., 2021).

Peck, Olsen, and Devore (2008) state that Excel as a spreadsheet framework efficiently performs data analysis, data computing, and storage operations. Excel can also clearly display analysis findings in real-time graphs, tables, and charts. However, unlike SPSS, Excel is prone to many errors, especially when inserting formulas into spreadsheets. Excel leaves room for human errors, lost files, and faulty procedures (Peck, Olsen, and Devore, 2008). Despite that, Excel is used by more than 731 000 companies in the US alone, among them being PayPal, JP Morgan

Chase, and Intel for data management and visualisation (Mangindin, 2022). Advantages and disadvantages of SPSS and Excel are summarised in **Table 7** below.

Table 7: Advantages and Disadvantages of SPSS and Excel

Software	Advantages	Disadvantages
Excel	It is easy to use; most learners can learn to use it by themselves.	It is prone to human error.
	It provides a wide assortment of free templates to download and use.	It is difficult to troubleshoot in case of problems (Java T Point, 2023).
	It allows for using large volumes of data via Power Pivot (Excel Champs, 2023).	It is not designed for real-time collaborations (Java T Point, 2023).
SPSS	It incorporates advanced technology and more efficient algorithms, helping firms to perform better in the long run.	It requires specific knowledge to apply all of its available functions, making it harder to learn.
	It facilitates more exact calculations that do not include approximations (Team TMT, 2021).	The cost of SPSS can be prohibitively expensive, particularly for individual users, hence potentially impeding its accessibility.
	It can merge files and transfer data from one programme to another via its compatibility with other data management programmes.	It is less flexible in terms of performing specific multi-level analyses or structural equation modelling (Rippey, 2021).

Conclusion

There are notable distinctions between SPSS and Excel. There is no basis for comparison between these two programmes. Excel is a software application designed for the creation and manipulation of spreadsheets, while SPSS is a software programme specifically developed for statistical analysis purposes.

Excel is a versatile tool that facilitates the execution of many statistical analyses. In contrast, SPSS possesses a multitude of features that can be regarded as comprehensive and robust tools for conducting statistical analyses. The Statistical Package for the Social Sciences (SPSS) has a range of inherent tools for the manipulation of data, facilitating the processes of recording and transforming variables.

In contrast, Excel does not include built-in functions. Significant effort is required to complete the assigned task. In SPSS, each column represents a distinct variable. However, columns and rows are not treated in the same manner.

Ultimately, it can be concluded that both Excel and SPSS effectively fulfil their respective roles within their domains. Both statistical approaches are highly effective and widely used within their respective fields.

After analysing and visualising different datasets, such as the student score dataset in terms of the qualifications for reduced or free lunch, it would be recommended to use SPSS over Excel, as it is less likely to present inaccurate results if applied correctly. Although SPSS has some disadvantages, its pros outweigh the cons in the proposed case study on student lunch.

References

- Aggarwal, C.C. and Reddy, C.K., 2014. *Data clustering: algorithms and applications*. London: Chapman & Hall/CRC. [Link](#)
- Assessment and Curriculum Support Centre, 2022. *Using Excel's Pivot Table to Analyse Learning-Assessment Data*. The University of Hawai'i at Mānoa. [Online] Available at: [Link](#) [Accessed 26 July 2023].
- Ball, T., 2018. *FedEx AWS silo leaks 120,000 sets of customer information*. [Online] Available at: [Link](#) [Accessed 2/8/2023].
- Benício, D.H.P., Xavier Junior, J.C., de Paiva, K.R.S. and de Camargo, J.D., 2022. Applying text mining and natural language processing to electronic medical records for extracting and transforming texts into structured data. *Research, Society and Development*, **11**(6), p. e37711629184. [Link](#)
- Big Data Tips. (2016). Coca Cola | Big Companies Data Mining & Machine Learning. [online] Available at: [Link](#) [Accessed 31 July 2023].
- Bordelon, B., 2021. *Getting Started in Data Analysis: Stata, R, SPSS, Excel: Home*. Princeton University Library. [Online] Available at: [Link](#) [Accessed 26 July 2023].
- Chatterjee, I., Zhou, M., Abusorrah, A., Sedraoui, K. and Alabdulwahab, A., 2021. Statistics-based outlier detection and correction method for Amazon customer reviews. *Entropy*, **23**(12), p. 1645. [Link](#)
- Derrfuss, J., 2023. *Chapter 25 Introduction to data pre-processing*. School of Psychology, The University of Nottingham. [Online] Available at: [Link](#) [Accessed 26 July 2023].
- Derrfuss, J., 2023. *Chapter 26 Data pre-processing with Excel*. School of Psychology, The University of Nottingham. [Online] Available at: [Link](#) [Accessed 26 July 2023].
- Excel Champs, 2023. *Top 10 benefits of Microsoft Excel – why you should learn it in 2023*. [Online] Available at: [Link](#) [Accessed 01 August 2023].
- Excel chat, 2019. *Using a Pivot Table to Analyse Worksheet Data*. [Online] Available at: [Link](#) [Accessed 26 July 2023].
- Han, X., Zhang, Y. and Qiao, J., 2019. Pre-treatment of Cotton Processing Data Based on SPSS. In Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference (HPCCT '19). [Online]. New York: Association for Computing Machinery, pp. 134-138. [Accessed 26 July 2023]. Available at: [Link](#)
- Java T Point, 2023. *10 disadvantages of Microsoft Excel*. [Online] Available at: [Link](#) [Accessed 01 August 2023].

- Kaur, H., Ahsaan, S.U., Alankar, B. and Chang, V., 2021. A proposed sentiment analysis deep learning algorithm for analysing COVID-19 tweets. *Information Systems Frontiers*. [Link](#)
- Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E.A. and Ceder, G., 2021. Opportunities and challenges of text mining in materials research. *iScience*, **24**(3), p. 102155. [Link](#)
- Loffing, F., 2022. Raw Data Visualization for Common Factorial Designs Using SPSS: A Syntax Collection and Tutorial. *Frontiers in Psychology*, **13**, p. 808469. [Link](#)
- Maheshwari, A.K., 2015. Business intelligence and data mining. New York: Business Expert Press, LLC.
- Mangindin, G., 2022. *Who uses Excel? Companies that use Excel and what Excel is used for*. [Online] Available at: [Link](#) [Accessed 2/8/2023].
- Masuadi, E., Mohamud, M., Almutairi, M., Alsunaidi, A., Alswayed, A. and Aldhafeeri, O., 2021. Trends in the usage of statistical software and their associated study designs in health sciences research: a bibliometric analysis. *Cureus*, **13**(1), p. e12639. [Link](#)
- Mayo Clinic, 2022. *Telehealth: technology meets health care*. [Online] Available at: [Link](#) [Accessed 31 July 2023].
- Peck, R., Olsen, C. and Devore, J., 2008. *Introduction to statistics and data analysis*. Belmont: Thomson Brooks/Cole.
- Pedamkar, P., 2023. *Data mining methods*. [Online] Available at: [Link](#) [Accessed 02 August 2023].
- Rahman, A. (2021). (PDF) SPSS: An Imperative Quantitative Data Analysis Tool for Social Science Research. [online] ResearchGate. Available at: [Link](#)
- Rahman, A. and Muktadir, M.G., 2021. SPSS: An Imperative Quantitative Data Analysis Tool for Social Science Research. *International Journal of Research and Innovation in Social Science*, **05**(10), pp. 300-302. [Link](#)
- Rippey, P., 2021. *Pros & cons of SPSS*. [Online] Available at: [Link](#) [Accessed 01 August 2023].
- Team TMT, 2021. *What Is SPSS? Definition, features, and advantages*. [Online] Available at: [Link](#) [Accessed 01 August 2023].
- The University of Michigan Library, 2023. *Using Microsoft Excel: Filtering and Sorting Data*. [Online] Available at: [Link](#) [Accessed 26 July 2023].
- University of Baltimore, 2020. *Time series analysis for business forecasting*. [Online] Available at: [Link](#) [Accessed 31 July 2023].
- University of Kansas, 2020. *Section 5. Collecting and analysing data*. [Online] Available at: [Link](#) [Accessed 31 July 2023].

Večeřa, M., 2020. *What is a Pivot Table and how to create it: a complete guide for 2020 (from beginners to advance with real-world examples)*. Lumeer. [Online] Available at: [Link](#) [Accessed 26 July 2023].

Working Data, 2020. *How to analyse sales using pivot tables*. Working Data. [Online] Available at: [Link](#) [Accessed 01 August 2023].

Yates, D. and Islam, M.Z., 2022. Data mining on smartphones: an introduction and survey. *ACM Computing Surveys*, **55**(5). [Link](#)