Mini Project Part 2

Your name

Department name, Institution name

Course number: Course name

Instructor’s name

Due date

**Q1.**

There is very little need for merging in the document. The complete data is categorized and divided into three sheets. The data from three sheets were combined into one to conduct this data cleaning and data wrangling. First, the data were sorted based on the ID to ensure the data through all the sheets was combined based on their respective data provider or ID. Next, the repeated columns were deleted. The combined column education was created. Instead of writing 0 and one for the respected column, the combined column uses if statements to show the education level of the users (Excel Chat, 2022). This can reduce the data entered into the Excel sheet and help use the sheet for analysis in analysis software such as Tableau and Power BI. This combined column also reduced the analysis time as it required excel to analyze and go through fewer columns.

**Q2.**

For summary analysis, four columns were chosen based on the provided data. First is the visit data. This data provides insights into the visits by each user. This data can provide useful information for the marketing and sales team and can be used to analyze those topics. In the visits data, the minimum number of visits was 0, and the maximum was 20. This shows that some people did not visit, and some did multiple visits. The mode of the data was 7. This shows the middle of the data. The range of the data was 20, while the mean was 5.34. Based on the analysis, the data was a little left distributed (Frost, 2020).

The second column was Premium. The premium column had a minimum of 0 which shows that some people don’t buy premium products (Extend Office, 2021). The max on the premium was 7570, which looks like wrong data. This data skews the complete results of the analysis and would make the analysis completely invalid. The mode of the column is 0, which means more than half the users don't go for premium products (Causal, 2023). The frequency of zero is 540. The range of the column is skewed because of the skewed max value. The data is very left distributed because of the high zero entries.

The third important column is Regular. This data is very similar to premium data and shows similar situations in min and max values. The min value is -6674, which is wrong as it conveys some people might regular products -6674 times and only go for premium ones. But the max data shows a wrong value. The mode of the data is 38, while the frequency of the mode is 62 (Waingankar, 2021). The range and mean of the data are heavily affected by the high max value within the data. The data is evenly distributed in this case.

The fourth and final important column is the age group. The min value in age data was -64, which is completely wrong. The max value is also 80, which is valid. The middle value is 44, which is valid. The age 44 was recorded 691 times.

**Q3.**

For uncleaned data, the charts cannot convey much information as they contain wrong values and values, which can skey the complete analysis. For the important columns discussed above, the age chart can convey some information, such as the common age groups of the customers or users. For premium and regular purchases, the scatter plots provide insights into the data. The data shows the entries are more focused towards the lower values between 0 to 400. This shows that most of the regular products sold are between 0 to 400. For premium, the values change as the max values in this range are between 0 to 200 (Thakur, 2020). Another similar chart can be created for visits which shows that most visits are between zero to 8 with some occasional spikes going up to 20(Microsoft, 2019). This can be a data consistency error, or it can be caused because some people visit more often than others.

**Q4.**

In the data, any blank spaces and negative values were removed. In this case, there cannot be any negative values such as a person visited negative times or a person has negative age. Therefore, any negative or zero values were removed from the dataset to provide accurate results for the analysis.

**Q5.**

Upon re-analysis after cleaning the data, there were some differences. For the first important column, the mean and frequency changed. This is due to the removal of unwanted and wrong values within the dataset. For the premium column, the maximum values changed from 7 thousand to 8 hundred. The new value is more valid as it provides a more practical number for the real-world analysis (Correlation One, 2023). The minimum value in the regular column changed from -6000 to 0. This is also a valid change, as no person can make negative purchases of a regular item. The zero here means there were no purchases which provide a clear definition for the analysis. The other values within the column were not significantly affected. After data cleaning, the age data was significantly affected. The min value changed from -64 to 24. The negative 64 was an invalid value which was replaced with a more analysis-friendly value of 24. The max value stayed the same. The range was affected significantly. Before data cleaning, the range was 144, which was changed to 56 after data cleaning.

In conclusion, the data was affected after data cleaning. The minimum and maximum values changed, which had effects on all the other formulas and variables, causing them to change.

**References**

Causal. (2023). *How to Find the Mode in Excel*. https://www.causal.app/how-to/how-to-find-the-mode-in-excel#:~:text=The%20MODE%20Function%20with%20the%20FREQUENCY%20Function,-Another%20way%20to&text=The%20FREQUENCY%20function%20is%20a,where%20you%20want%20the%20results.

Correlation One (Director). (2023). *Data Analytics 8—Mini-Project 2: Data Curation / Wrangling Walkthrough—Saturday, July 8th*. https://www.youtube.com/watch?v=GwAfw9NDYds

Excel Chat. (2022). *Syntax of the IF Formula*. https://www.got-it.ai/solutions/excel-chat/excel-tutorial/if/if-else

Extend Office. (2021). *How to find the maximum / minimum values in Excel?* https://www.extendoffice.com/documents/excel/1274-excel-find-max-min-absolute-value.html#:~:text=Then%20press%20Ctrl%2BShift%2BEnter,to%20get%20the%20correct%20result.

Frost, J. (2020). How to Identify the Distribution of Your Data. *Statistics By Jim*. https://statisticsbyjim.com/hypothesis-testing/identify-distribution-data/

Microsoft. (2019). *Create a chart from start to finish*. https://support.microsoft.com/en-us/office/create-a-chart-from-start-to-finish-0baf399e-dd61-4e18-8a73-b3fd5d5680c2

Thakur, M. (2020). Range in Excel. *EDUCBA*. https://www.educba.com/range-in-excel/

Waingankar, R. (2021). *FREQUENCY in Excel*. https://www.wallstreetmojo.com/frequency-function-excel/