Mini Project Written Answers by: Hristian Hristov

Team Name:Team 20

Team Members:

Hristian

Mickey

Mohammed

Nihil

Oluwatobiloba

Roles and Responsibilities:

In the second mini-project, as a group, we decided to follow a different approach from mini-project 1. We held meetings daily or every other day to share ideas and come up with the best possible solutions to be presented and presented.

One of the reasons we chose a different approach from the one used in the previous project is that mini-project 1 consisted of 3 worksheets, each with questions related to its specific worksheet.

This division facilitated the distribution of tasks within the team. However, in this project the responses are linked to a single dataset.

For this mini-project, we were tasked with answering question 1, question 2, and question 3 using the data as is.

Question 1:

The process of combining the Sales, Customers, and Marketing sheets into one Excel sheet resulted in about 36 columns. In this question I am asked to determine if it is possible to further combine certain columns due to their category similarity.

Upon inspection, I found a few columns that could be considered to belong to the same category.

This formula looks for the value 1 in the corresponding columns. If all columns contain the value null will return "Not Given".

By applying such formulas, we were able to reduce the number of columns from 36

to 16, achieving a more organized and simplified data set.

Question 2:

In this particular question, our focus was on identifying the most important columns and examining their minimum, maximum, mode, range, and frequency of the most common value. The key columns we considered in our approach were Income, Age, Children, Product,Premium and ordinary.

Turning to the first four parts of the question, we arrived at clear answers using formulas and simple calculations

Question 3:

Charts can be helpful in finding the minimum and maximum values and therefore the range. However, they may not be as effective at determining accurate central measures, such as the mean and mode. Nevertheless, they can be helpful in estimating the approximate location of these values on a chart.

Among the best charts for visual analysis are histograms and line charts. I think it's a replica.

The chart can show both minimum and maximum values, as well as show the concentration of data points, providing an idea of the possible range in which the average value may lie. However, as mentioned above, it cannot provide an exact value. However, by minimizing the basic units on the y-axis into smaller units, we can try to derive a more satisfactory estimate.

In summary, charts are useful for visualizing data trends, extremes, and general distributions, but precise measures such as mean and mode may require additional calculations and interpretation.

Question 4:

As required in this question, data cleansing involves removing records that do not represent realistic data sets. For example, there is a negative value (-64) in the Age column, which strongly affected the range value in Question 2. As a result, the minimum value becomes negative,resulting in an unrealistically high value for the range (max-min) in terms of human ages.

Question 5:

This question involves repeating the calculations from question 2 to check how the data is cleaned affects the data.

After analyzing the cleaned data, I noticed that it looks more realistic. For example, in the Age column the range is 56, while in the pre-clean data the range is 144. This shows a significant improvement in the accuracy of the data after cleaning.

Learning outcomes:

This project emphasizes the importance of data accuracy to obtain valuable and realistic measures. For example, it demonstrates how a negative value in an entire column consisting of 2225 lines resulted in an invalid range calculation, making it meaningless.