Summary work No. 4

Topics:

1. Retrieval protocol

2. Data Examination (EDA)

3. Clear the data

Note: You can work in Python or R. It is recommended to work in jupyter notebook. Anyone who chooses to work in RStudio, in addition to the code will have to write the insights and copy the graphs in a word document.

Part 1 - Removal Protocol

1. Download the retrieval protocol from the link and complete all the columns. The SQL code to generate the VIEW with the flat file is at the link.

Part 2 - Exploratory data analysis

1. Describe the data with descriptive statistics (use concentration and dispersion indices)

2. Graphs depicting the behavior of each variable.

3. Create a correlation matrix and present it in a graph

4. Describe the revenue variable (revenue) - how is it distributed? Are there categories that show a large difference in the distribution of the target variable?

5. Create graphs that can help you check for extreme data. Describe them.

6. Describe the missing data: Which variables have missing data? Some?

7. Create a matrix of gaps (create a dataframe with the same dimensions from the source table and cells that are missing put a value of one and those that have data have a zero value). Show the matrix in the heatmap graph.

Part 3 - Clearing the Data

3.1 Extreme data

1. In variables where extreme values ​​are found, check the distribution of the variable with and without the extreme values. Does the distribution change?

2. Create a scatter graph with the variables in the previous question in X and the target variable in Y. Do extreme values ​​in variable X affect the behavior of variable Y? Do you see differences with or without extreme values?

3. In what variables would you delete the extreme values? How would you delete them? Necrosis.

4. Act on the data according to what you defined in the previous question.

3.2 Missing data

For each variable with missing data, you will see the distribution of other variables with or without missing. Use the missing matrix you created in part 2, question 7 for the missing / no indication. For the distribution, use a histogram or density graph with category / color as indicated.

2. Make a table of the variables in which there are gaps and describe what the mechanism of creation of the gaps is (based on the results of the previous question).

3. Which imputation technique is suitable for each variable? Use the chosen technique to replace the missing ones.

Part 4 - After clearing the data, repeat part 2 in full.