Phase-End Project: Marketing Campaigns

SUBMISSION BY: COL RAKESH PEDRAM: Nov23 Cohort

Approach

- After reading the problem statement. Did a background study on how campaigns are run. How surveys are designed and the goals.
- I did the proj on google colab in Jupyter note books. I didn't run the code in single code block as the cleaning and testing the data was iterative process.
- Revised by Statistical study after reading the problem statement.
- Examined my files in excel to get the broad idea of the statistics and the effort required in data cleaning.
- Preliminary choice of what tests to apply to each hypothesis as when you are not familiar with python it takes a long time to do analysis and change the tool.
- Got my cheat sheets ready for referring code syntax.

Execution

Cleaning the data

I transformed the csv in excel using power query, just like the sql project. However I wrote code snippets in pyton for the learning value. Especially currency conversion to float.
The imported data was tested for data type. String integer and floats were checked. Date changed to datatepe64.

Missing values

The null values were checked and categorical data uniques to ensure categories were correctly classified.
I have reduced the dimensionality of my data to make model simpler for my first project to ensure a fit and avoid multinominal regressions.
Marital status has been clubbed into just two categories: Together and Single. Only these two have practical impact on shopping characters.
Education qualification were indainised and those were also clubbed as one category had just 2% of the whole data. SSC-Plus, Graduation, Masters. PhD was clubbed in Masters and 2nd was into SSC_Plus.

Addl Variables and Outliers

Addl variables were created for Age and total Spending. Outliers were iden with percentile and box plot and clipping(1.5IQR) on all numeric data done to remove outliers. Income field had one outlier. However initially I iterated the outlier treatment loop on all numeric data due to which my

binary data was all set to zero. Rectified by excluding promotion binary fiels from loop. Cleaned and normalised data was then rechecked with .describe()

Ordinal Encoding

My Education data I encoded by rank. Rest I used one-hot encoding. This creates more columns and is difficult to handle.
Ran the correlation algo on all the numeric data and used sns.heatmap() for visualisation.
The columns were then selected and corelation run only on select column. This gave a readable heat map.

Hypothesis testing

My Education data I encoded by rank. Rest I used one-hot encoding. This creates more columns and is difficult to handle.

Data Visualisation

This was simple mathplot functions to male scatter, pit and line. The learning value was what type of visualisation is suitable for what dats. I tried multiple visualisation. The coding was simple. One requirement in the problem statement was wrong as our data has not product wise sale details. Hence a summary of category sales was made.

Conclusion

Very challenging project and excellent learning. The next project will be much easier as I get familiar with the syntax.