**Week 8 Deliverables**

Lisum20- Berna Gündoğdu Özbaşaran

Email: [gundogduberna66@gmail.com](mailto:gundogduberna66@gmail.com)

Middle East Technical University- Sociology- Data Science

Problem Description

As our Australian company to make a forcast of the future data, what needs to be done is to firstly understand the past data. Hence, I have made some hypothesis that could bring light upon the questions at hand. By referencing the data and the insights it might provide, I will eliminate the top hypothesis that are most crucial. The hypothesis model requires there being a null hypothesis that must be rejected with a significant p-value hence we aim to arrive at the alternative hypothesis.

H0.1: The fluctuating values of ‘Google Mobility’ are the same in each year

Ha.1: The fluctuating values of ‘Google Mobility’ is due to Covid pandemic which has occurred in 2020

H0.2: All 6 products get the same discount throughout years

Ha.2: There are some products that gets more discounts

H0.3: Price discounts have no effect in sales during Covid

Ha.3: Promotions (three type) have caused more gross income

H0.4: There is no statistically significant different in the success of products in terms of sales (sku1 to sku6)

Ha.4: Some products are more successful than others

H0.5: There is no statistically significant difference in success throughout years (2017 to 202)

Ha.5: The firm was more successful in some years are more successful than others

H0.6: There is no statistically significant difference between holiday promotions success (V-Day, Easter, Christmas)

Ha.6: Some holiday sales were more successful than other

H0.7: There is no statistically significant difference between the success between Instore Catalogue and Store end promotion

Ha.7: Some promotions are more successful than others.

The hypothesis I have decided to utilize are hypothesis 2,3, 4 and 7 seems the others seem quite common sensical.

Data Understanding

To analyze the descriptive statistics, one can use the *.describe()* method. In addition to this, one should firstly examine what the variables are which will provide insight for the analysis. The product variable is a nominal value where every product has a distinct sku number. The date will be important in this anlaysis for the trend and pattern analysis. This variable would be the independent variable in our study since a difference in this value would effect the dependent variables like sales. The sales variable is a discrete variable and also a dependent variable. The price discount would also be an independent variable but we cannot be sure about it just now. The rest of the variables are possibly independent hence affects the sales rate, which is why we are doing the analysis.

What Kind Of Data Is Desired

The optimal data in hand would be the one that is bell-shaped. There might be a slight skew but there is no need to worry since statistical methods of pandas could take care of that. Missing values also pose a concern to the accuracy and quality of data.

What Kind of Problems and Ways to Tackle them

According to the pandas analysis *df.isnull().sum()* there are no outliers in the dataset. However, if there were we could use multiple methods. These methods could be to drop, interpolate( the average of the forward and backward value), copying the forward one with the pad method and the backward value with the bfill method.