A STATISTICAL ANALYSIS ON GLOBAL SUPERSTORE DATA

CWID	First Name	Last Name	IIT Email
A20460285	Priyanka	Hosur Mahadevu	pmahadevu@hawk.iit.edu
A20457612	Padmashri	Adgonda Malgonnavar	padgondamalgonnavar@hawk.iit.edu

Introduction

The Superstores industry comprises companies that operate by having large size spaces that store and supply large amounts of goods. The superstore industry is comprised of extensive stores that sell a typical product line of grocery items and merchandise products, such as food, pharmaceuticals, cosmetics and personal care items, health products, games and toys, furniture, and appliances. Superstore provides a membership fee for consumers to shop within the store and once a member, the superstore provides consumers a broad array of products for discounted costs.

The superstore industry is part of the retail trade market. Most of the products bought at superstores are used by other wholesalers and smaller retail businesses for their own companies. There is constant competition between the superstores and supercenters with many merchandisers, department stores, wholesalers, and grocery stores. Large superstores and superstore chains are predominant in this market because of their economics of scale in financing, purchasing, and distributing.

To analyze such an industry is of great importance and induced us as it gives insights into the sales and profits of various products. Our analysis is based on global superstore data where the products are ordered between the years 2011-2015. Here, in this superstore data we analyze and discover various aspects that determine the profit of superstore based on some parameters like discount, products, and sales.

Data Sets

In this project, we have chosen the dataset from Kaggle and retrieved from the link: https://www.kaggle.com/jr2ngb/superstore-data.

Our analysis is based on a retail dataset of a global superstore from the year 2011 – 2015 (4 years) and the dataset belongs to the retail domain with **51290 observations**. We are exploring the relationship between sales against different market groups and sales against different product category groups where we are trying to predict the profit (dependent variable) with the help of the information contained in the other variable with a **95% confidence level**. The final model would be useful for the superstore manager to predict the store profit with respect to qualified independent variables such as store market region, category of product, discounts, etc.

Research Problems

- Do superstore have different Sales with respect to different groups of Market regions.
- Whether the group of product categories in the superstore have the same sales or not.
- Build multiple linear regression models to predict profit of superstore with respect to region-wise, sales-wise, product wise sales etc.

Potential Solutions

- As we observed that the Market has a group within groups, we are implementing ANOVA to solve the problem by building the ANOVA regression model. And to check model assumptions will perform residual analysis, then perform F-test and compare p-value with significance level to accept or reject the Null Hypothesis. At last, we check the p-value of each slope in the t-test to know which groups the same and which groups are not if the mean of all market groups is not the same.
  Categorical variable: Market and Y variable: Sales
  Null Hypothesis: Mean of all market groups are the same.
  Alternate Hypothesis: Not all the market groups of means are the same.

- As we observed that the product category has groups within groups, we are implementing ANOVA to solve the problem by building the ANOVA regression model. We perform F-test and compare the p-value with the significance level to accept or reject the Null Hypothesis. At last, will look at the p-value of each slope in the t-test to know which groups the same and which groups are not if the mean of all product categories is not the same.
  Categorical variable: Category and Y variable: Sales
  Null Hypothesis: Mean of all product category groups are the same.
  Alternate Hypothesis: Not all the product category groups of means are the same.

- According to the size of the dataset, we choose the **hold-out evaluation** method to split the data by considering **70%** of the data as training data and **30%** of the data as testing data. And asses the multicollinearity issue with the VIF method. By using feature selection will build different models by parameter estimates and check the goodness of fit. And perform residual analysis to validate the model is qualified or not and rebuilding the model until the model is qualified. Look for influential points if exists remove and rebuild the model.

Evaluations

Once the fitted final multiple regression model is built with training data set by considering the least RMSE value will produce confidence level for prediction to check how accurate, the

prediction is by evaluating model with the test dataset. If the dependent involved with any transformations will be reverted to the original Y variable before producing predictions.

Expected Outcomes

- We expect that there would be a significant difference in Sales with respect to different Market regions.
- We expect that there would be a significant amount of difference in sales of products with respect to different product categories based on demand or discount.
- Fit multiple linear regression model would be useful for superstore manager to predict profit with respect to useful predictor variables.