**Final Project | Statistical Computing BANA 6043**

**Anukriti Sharma, Sukanya Wadawadagi, Laasya Akunuru**

**October 2021**

**Project Contents**

1. Introduction
   1. Dataset
2. Analysis
   1. EDA
   2. Our Findings
3. New Functions
4. Overview
5. **Introduction**

[Retail sales](https://www.thebalance.com/u-s-retail-sales-statistics-and-trends-3305717) are a very important economic indicator because when consumers open their pocketbooks, the economy tends to hum along. Think of all the people and organizations involved in producing, distributing, and selling the goods you use daily like food, clothes, fuel, and so on.

One such large organization interacting with consumers providing many good for purchase is Easy Shop, Easy Shop is a superstore which provides a wide range of products for sale, including, office supplies, furniture, and tech products. Analyzing the stores data over a significant amount of time from last few years, gives us a chance to predict the sales in the coming years, and what demographic the marketing team to cater to.

Easy Shop takes orders from customers, and ships out multiple products across different states and cities in the United States based on the requirement.

* 1. **Dataset**

The dataset contains retail data of a global superstore for 4 years. It contains 17 variables where the first one is Row ID – which is the numbering of each row of data. The other 16 are Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Country, City, State, Postal Code, Region, Product ID, Category, Sub-Category, Product Name and then the most important one, Sales. This dataset was sourced from Kaggle.

The data spans over 4 years and includes one data point (one row of data) per week. This dataset is a CSV file, and we will use R and SAAS for our analysis. We will use time series analysis as well to observe the trends as well.

1. **Analysis**
   1. **EDA – Exploratory Data Analysis**

We first conducted basic Exploratory Data Analysis on the dataset in R. It helps identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables. These are the following steps we went through to complete EDA.

* **Install and Load the libraries by using: library(DataExplorer)**

str(superstore)

Text

Description automatically generated

*After importing the dataset, we noticed that Row ID should be numeric, since there is % attached to it. The value here is character. We then proceeded to clean this data to the appropriate data type for a better analysis.*

* **Data Cleaning**

superstore$Row.ID = as.numeric(gsub('%','',superstore$Row.ID))

Text, letter

Description automatically generated

*After cleaning, the Row.ID values get converted to numeric values.*

* **We checked the dimension of the input dataset and the time of variables.**

plot\_str(superstore)

A picture containing chart

Description automatically generated

*With that, we saw we have some Continuous variables and some Categorical variables. We can see an overview of each variable with its datatype.*

* **Check for Missing values before diving deep in the analysis.**

plot\_missing(superstore)

Chart

Description automatically generated

*In our dataset, there are no missing values.*

* **Histogram is analyst’s best friend to analyze/represent Continuous Variables.**

Graphical user interface

Description automatically generated

plot\_density(superstore)

Chart, line chart

Description automatically generated*That marks the end of univariate analysis and the beginning of bivariate/multivariate analysis, starting with Correlation analysis.*

* **Bivariate Analysis**

plot\_correlation(superstore, type = 'continuous','Review.Date')

Chart, bar chart

Description automatically generated

*Row ID here is essentially a Customer ID. This creates a heat map for all discrete categories, which is Sales, Postal Code and Row ID. This function first dummifies all categories, then calculates the correlation matrix. The darkest color is perfect correlation. Postal code and Row ID have a stronger relationship. There exists a relationship between Sales & Postal code, Sales & Row ID.*

* **Categorical Variables — Barplots**

plot\_bar(superstore)

Graphical user interface, text

Description automatically generated

Graphical user interface

Description automatically generated

We then used the function create\_report() that gives a sharable rendered markdown in html.

* **create\_report(superstore)**

A picture containing table

Description automatically generated

**The downloaded Data Profiling Report can be accessed by double clicking on this html link:**



* 1. **Our Findings**

1. We first used R to observe the Statewide Sales.

Chart, scatter chart

Description automatically generated

*It is observed that North Carolina reaches the highest number of Sales which is around 8000 followed by Georgia just crossing 6000. More consistently, based on the frequency of the plots, there is a higher sales record for Texas and California. The least sales over 4 years are from Nevada and Maryland.*

1. It was imperative to know about shipment method that an individual prefers in different cities, so that from a store point of view, you can invest to cater the needs of the customers with respect to their preferences.

Chart, bar chart

Description automatically generated

Chart, bar chart, treemap chart

Description automatically generated

*From the above histograms, we can conclude that a customer generally does not change his shipping preference method. Once a person chooses either First Class, Second Class, Standard Class or Same Day shipment method, it is highly unlikely that the customer would change that. This holds for other cities as well. Easy Shop must prioritize the customer to those shipment methods to continue getting their sales, maybe even offer discounts to retain the trust and confidence.*

1. For our next analysis, we wanted to explore the sales of different categories sold in the superstore.

Chart, pie chart

Description automatically generated

*The pie chart shows out of the main 3 categories offered by Easy Shop, Technology products was sold most. Followed by Furniture. Office Supplies had least sales. Based on this breakdown, the stock of inventory, the marketing of sales targeted can be specific.*

1. The next main criteria is of Shipment Mode with respect to Average Sales. We used proc means function in SAAS to get an output of this table.

Table

Description automatically generated

*The average sales are the most for Same Day shipment, followed by Second Class. Time is important factor for delivery, and hence more resources need to be allocated for Same Day since that is the preference of majority of customers.*

1. Our last analysis is the number of sales of different categories with respect to the shipment mode.

A picture containing timeline

Description automatically generated

*In this plot, we can identify which category has the highest sales for which shipment mode. Technology products being delivered on the same day have the highest sales of 8000. For Office Supplies, Standard Delivery has highest sales of above 6000.*

1. **New Functions**
2. **create\_report()**

This function in R is to generate and render element in a report object in EDA. We chose this function because this creates a clear and concise html report holding all the details of the exploratory analysis that was done.

1. **mapdata**

We have used mapdata function in R to show the sales based on region in our superstore data. mapdata package was first installed, and then longitude, latitude, group and region were taken into consideration. We chose mapdata to see the sales interacting with the regions, and the density of lines in an area shows the popularity of the store in that region. We can use this visualization to further increase sales catering to specific regions and encourage the store to expand in the US and globally.

1. **Left Inner Join**

Using left inner join in the previous mapdata function, we have mapped the superstore data based on the region variable. The sales data is then matched. Based on those sales the region is mapped. We used this to get the sales data to the mapped data frame.

Engineering drawing

Description automatically generated with low confidence

1. **Overview**

By conducting past sales analysis on a global superstore like this, you can predict the likelihood of a prospect converting into a customer and make personalized offers to leads that are ready to buy. You can also increase the lifetime value of existing customers by identifying upselling and cross-selling opportunities in customer behavior. Analyzing the historic data from the past 4 years sheds some light on sales trends, how likely are customers to buy products from different categories over time, what are they willing to buy, how important is the delivery of these products to them, and so on. Association Rules here are widely used to analyze sales and transaction data which are intended to identify strong rules discovered in this process.