**FINAL PRESENTATION**

**E-COMMERCE SALES DATA**

**ANAMIKA SINGH**

**BANA 6760 E01: DATA VISUALIZATION**

**DATE: MAY 8, 2021**

**UNIVERSITY OF COLORADO, DENVER**

**Executive Summary**

Exploratory Data analysis is a crucial step in data analysis process therefore EDA is performed on the dataset before performing data modeling. EDA will be performed on **E-commerce Sales** dataset using **R** to understand the structure and pattern of the dataset. EDA will help in analyzing the relationship between variables and helps to identify errors, missing values, outliers, and erroneous data. Data visualization is very important step that helps to get clear insights from data to make better business decisions. Dataset will be examined, and visualizations will be created in **Tableau** to answer some of the business questions to improve the sales and profit. The purpose of the project is to create interactive worksheets and dashboard with Tableau using **E-commerce Sales** dataset.

**The Dataset**

The following dataset is used for the project:

E-commerce dataset from Kaggle. The dataset contains sales details from Indian E-commerce market. The dataset includes two excel files.

1. **Orders**- This file contains information on Order ID, Order Date, Customer Name, City and State. I selected this dataset as it contains geographical locations that will be plotted on map.
2. **Order Details**- This file contains information on Order ID, Amount, Profit, Quantity, Category and Sub-Category.

Reference for dataset:

R.Ben. (2020) E-commerce Dataset. *Kaggle*

<https://www.kaggle.com/benroshan/ecommerce-data?select=Sales+target.csv>

These two datasets will be joined with the help of a common column. I will join Orders to Order details using order ID column. After joining the dataset, data cleaning will be done on R and Excel. Dataset contains some null values that needs to be addressed before performing visualizations. EDA will be performed on the dataset to identify outliers and erroneous values and to get better understanding of the dataset.

**Data Analysis Tools**

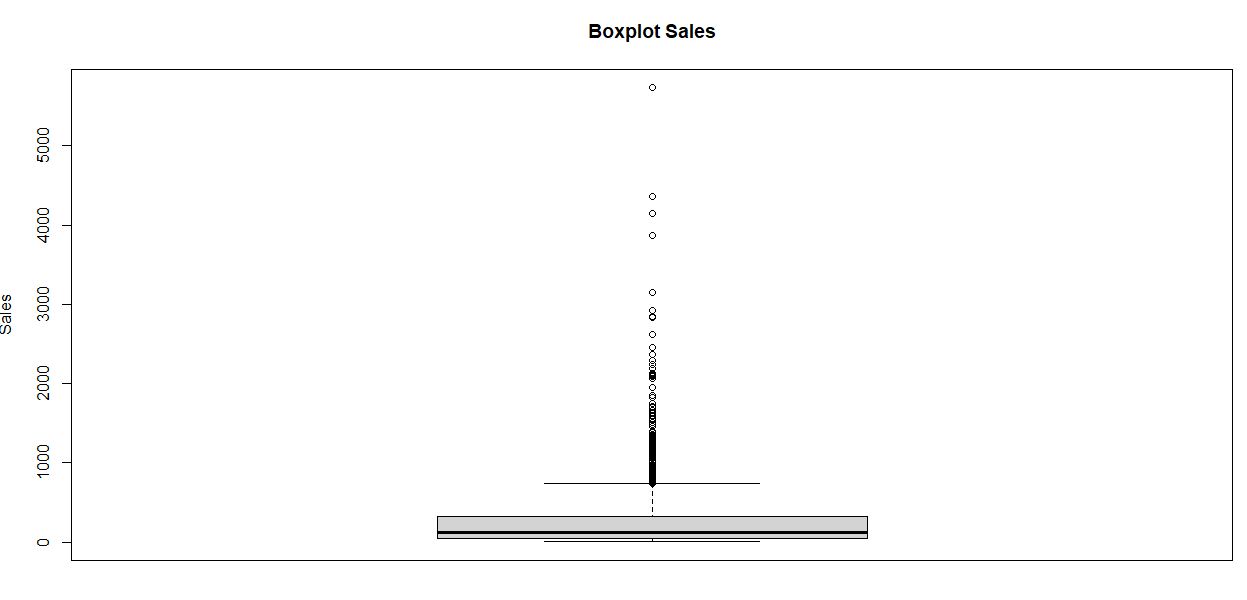
Tools I am planning to use are:

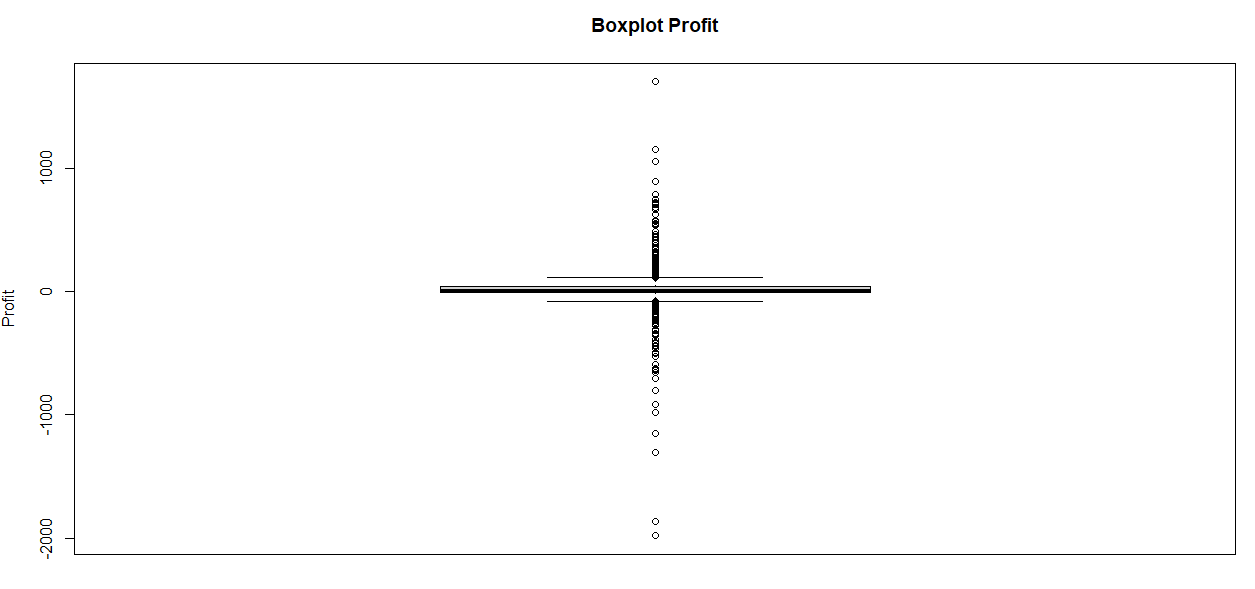
1. R and Excel- I like working with R and Excel, that is why I will use both the tools to understand the dataset. I will use Excel to get the basic structure of the dataset. R will be used to perform EDA on the dataset. That will help to understand the overall pattern of the dataset. It will help to find missing values, null values if present in the dataset.
2. Tableau- Tableau helps to build complex visualizations as it is more powerful than Power BI and deals with high volume of dataset, I will use Tableau to build visualizations, mapping dataset, preparing worksheets, Dashboard, and stories.
3. Power point- PowerPoint will be used to prepare presentation by adding write up slides and dashboard to it. It will also be used to record the presentation at the end.

**Exploratory Data Analysis**

EDA helps clean up the dataset and identify important variables in the dataset. It is used by data scientists to analyze and get familiar with the dataset. It helps to clean up the dataset. The E-commerce sales dataset contains information about order ID, order date, product category, profit, and sales.

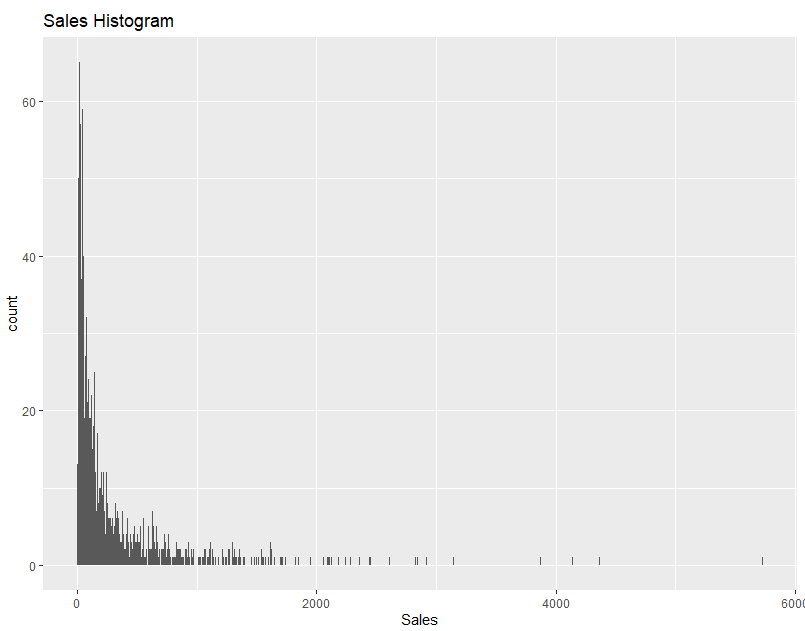
* **Missing Values**- By using Summary function in R, it is disclosed that no variables have missing values in the dataset.
* **Erroneous data** – Order date column has mixed date format. Changed it into standard single format using Lubridate package.
* **Rename column**- Renamed amount column with sales.
* **Outliers**- There are so many outliers present in the sales and profit dataset. It was determined by using Boxplot in R.

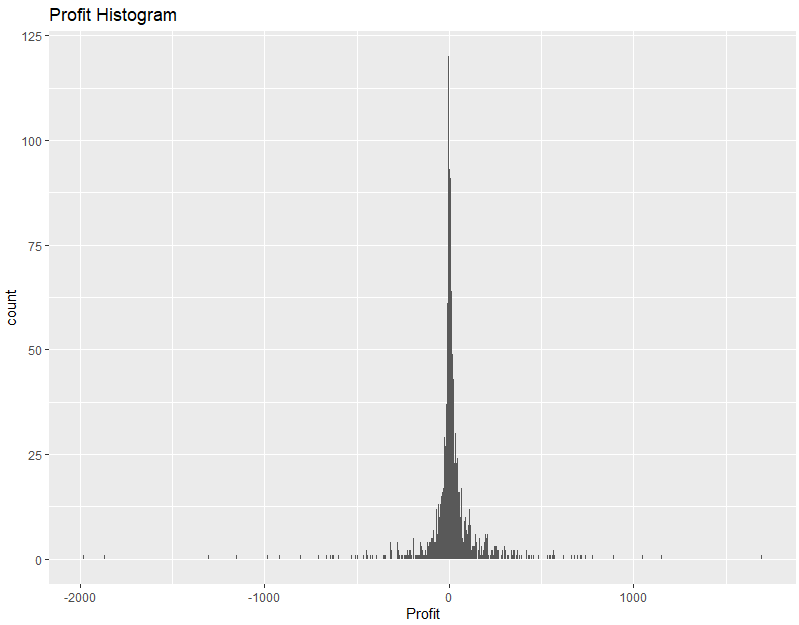




* **Distribution**- Using moments package in R, skewness and kurtosis of the sales and profit variables are determined. Shape of the sales and profit variables are also determined by plotting histogram and box plot with **five number summary**. Mean is greater than the median for Sales, meaning the variable is **right skewed**. Profit has negative skewness, meaning profit is left skewed. Both the variables also have high positive value of Kurtosis, which means the distribution is peaked with thick tails. Standard deviation is high for sales data, meaning data is more spread out.

The below two graphs shows the distribution of the sales and profit data. The histogram for sales data is highly right skewed, meaning mean is greater than the median. The histogram of the profit variable is left skewed. Some of the values in profit are in negative.





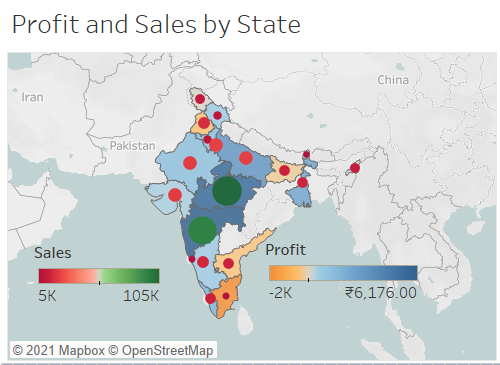
* **Correlation**- By visualizing scatter plot between sales and profit using cor function in R, we saw that profit is not related to sales as the correlation is only 0.24.
* **Merged Data**- Datasets from different files were merged into a single file to perform EDA and building visualization.

**Business Questions and Visualizations**

After examining and visualizing the dataset with the help of Bar plot, line graph, scatter plot, and maps in Tableau, business questions were answered to improve the profit and sales. **Dashboard** answers the following questions:

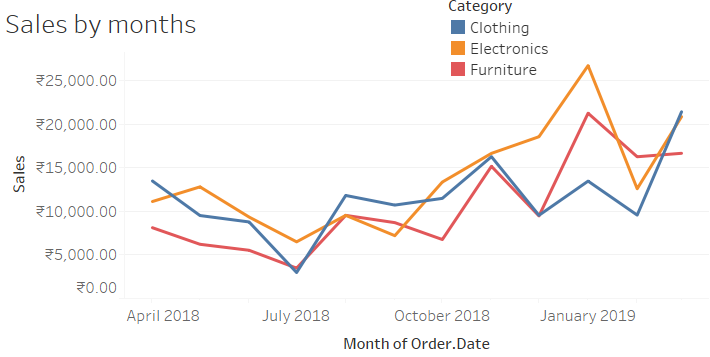
What are the sales and profit by each state?

Map visualization is created using latitude and longitude automatically generated by Tableau. Below map represents Sales and Profit by state. Color gradient is used to represent the profit and circle size represents sales. We can observe that Maharashtra and Madya Pradesh have high sales and profits.



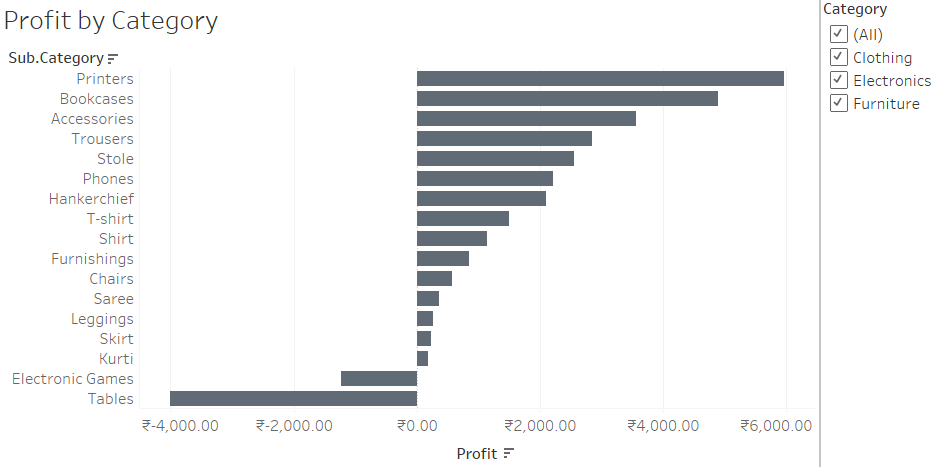
What are the sales by each category over months?

Line graph represents Sales over months by category. We can observe that sales of all the three products are increasing over months. In January 2019 sales are the highest for the E-commerce company. In July we can see a dip in the company sales for Clothing and Furniture.



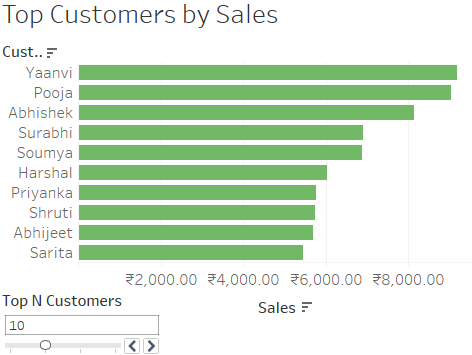
Which Product brings the highest profit?

The bar graph below depicts profit by product category. From the graph we can observe that Printers brings highest profit for the E-commerce company. If we scroll down on the dashboard, we can observe that Table brings loss to the company.



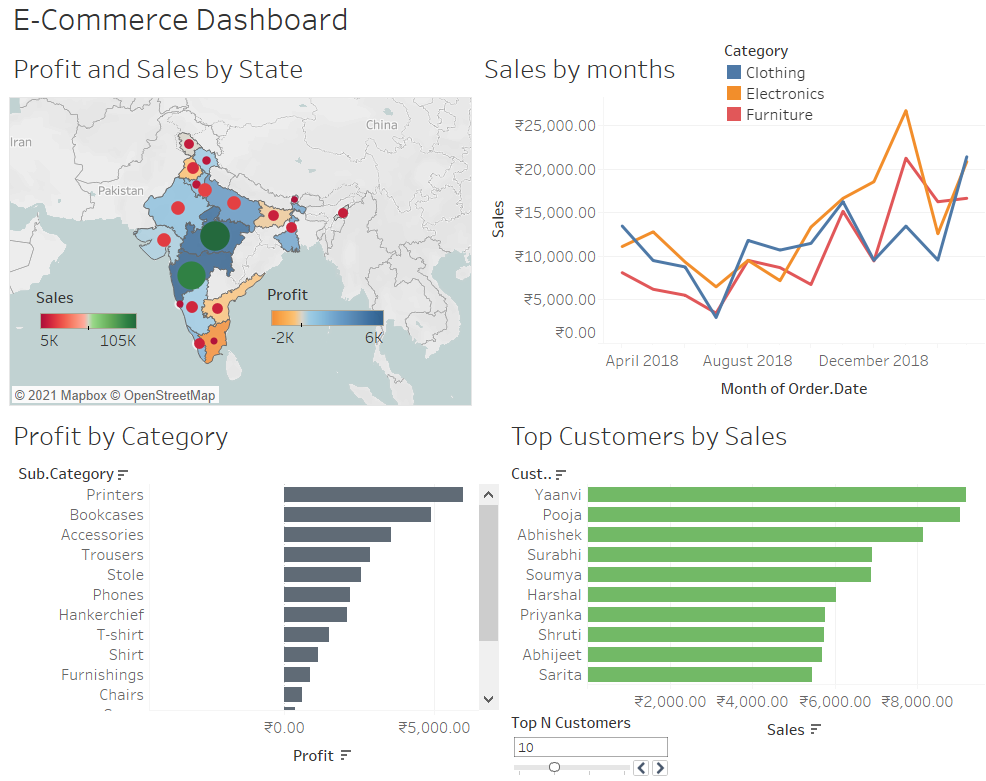
Show the name of top customers by sales.

The bar graph below shows the name of the customers making highest sales. Y axis of the horizontal bar graph represents customers name and x axis represents sales in Rupees. Using sets and parameters top 10 customers are shown. We can also change the view from top 10 to top 20 customers using parameter controls. We can always adjust the slicer to vary the range of number of customers. We can observe that Yaanvi is the top customer by sales.



**E-Commerce Dashboard**

E-Commerce dashboard is created using Tableau. The dashboard below contains all the graphs discussed above to answer all the business questions. Dashboard is made interactive by using “use as filter” option. Now if a user selects a graph, it will act as a filter for all the other graphs in the dashboard.

**Conclusion**

Worksheets and dashboard are created in Tableau by importing E-commerce csv file. It was easy to create interactive graphs and maps in Tableau to draw insights from the dataset. Map visualization showed that company receives highest revenue from Madhya Pradesh and Maharashtra. Line graph displayed that E-commerce company gets highest sales in January. We also observed that printers bring the highest profit for the company and Yaanvi is our top customer by sales.

We also used filter, sets and parameters to make the dashboard interactive. Use as filter option gives an option to use visualization as filter. When a bar is selected in the dashboard, it will filter all the graphs according to that bar.

REFERENCE:

Chang, W. (December 2012) R Graphics Cookbook.*O’Reilly*

Advanced Tableau (Professor Marcus Ellis, March 02, 2021)