**Data Analysis and Visualization**

**Ecommerce Data Analysis**

Acknowledgment

I would want to express my appreciation to everyone who helped me complete this report. Mr. Saurav Poudel, our instructor, deserves special recognition for his energizing suggestions and support, which allowed me to complete my project on time.

In addition, I would like to thank Lecturer Mr. Saurav Poudel, whose lectures supported me in completing my research. Mr. Rohit Panday, the department head, deserves many thanks for his tremendous efforts in guiding me and providing me with all the essential materials to complete the job. I must recognize the contribution of other instructors and friends to the improvement of my research and report.

Abstract

The purpose of this document is to demonstrate the knowledge gained from the "Data Analysis and Visualization" module and to practice applying acquired programming knowledge and abilities to data analysis and data visualization utilizing a variety of tools and methodologies. In addition, showcasing the procedures involved in a Data Science pipeline, from data gathering and Exploratory Data Analysis to Machine Learning / Modeling algorithms and Data Visualization approaches, along with problem-solving and critical thinking or evaluation abilities. Analysis and visualizations are carried out using R along with its popular libraries like dplyr, skimr, tidyr and ggplot2. In addition to data visualizations, data comprehension, data preparation, data analysis, predictive modeling, and hypothesis testing were performed consecutively. In data comprehension, each variable's attributes were examined. In the data preparation process, all features were examined and eliminated in accordance with the needs of the study. In the process of data analysis, correlation was computed using descriptive statistics. In data exploration, data were also presented as histograms, bar charts, stacked bars, pie charts, scatter points and heat maps.

Table of Contents

[1. Introduction 5](#_Toc104558969)

[1.1 Objectives 5](#_Toc104558970)

[2. Data Understanding 6](#_Toc104558971)

[3. Data preparation and Manipulation 7](#_Toc104558972)

[4. Exploratory Data Analysis 14](#_Toc104558973)

[4.1 Data Visualization 14](#_Toc104558974)

[5. Statistical Testing 25](#_Toc104558975)

[5.1 Hypothesis Testing 25](#_Toc104558976)

[5.2 Correlation Analysis 25](#_Toc104558977)

[6. predictive Data Analysis 27](#_Toc104558978)

[6.1 Linear Regression 27](#_Toc104558979)

# Introduction

The dataset used for this coursework is obtained from tableau Community. The data set Global Super Store has around 51000 values. It is a customer-centric data set that includes information on all orders placed with various vendors and markets between 2011 and 2014. It has features such as order id, customer name, segment, state, country, region, market, postal code, category, subcategory, sales, Order Date, Ship Date and has 51290 rows collected for Global Super stores record. This data can help in reducing the shipping cost and optimize the logistics by analyzing the data. Order priority defines any product with its delivery time in that must be shipped in delivery time. Link for data source https:// <https://data.world/asepetruk/global-superstore>.

## Objectives

The main objective of this coursework is data analysis and visualization by using different statistical approaches and machine learning algorithms to extract some valuable outcome. Also, we can get insights from following data:

• We can get insights about the average shipping cost according to priorities of orders.

• We can quantity distribution as per the categories.

• To find Total sales as per Category and their percentages.

• We can know about the shipping cost according to market areas.

• We can Find correlation among sales, shipping time, quantity, profit, and discounts.

• Which subcategory has the best profit as per the market areas.

• Distribution of Profit and sales with respect to category and shipping modes.

• Which category and subcategory has highest sales and profit?

• Which city is most likely to get faster or slower delivery time for a product.

• Which type of segment has the highest number of sales.

# 2. Data Understanding

**Summary of Dataset:** This dataset provides summary information of the global stores sales data which is in csv format. This dataset is provided in csv format which consists of 24 columns with 51k rows of data. There are some missing values in the data set in Postal Code column. Most of the data inside the files are in Categorical type and continuous data types.

|  |  |  |  |
| --- | --- | --- | --- |
| **SN.** | **Features** | **Data Types** | **Variable Types** |
| 1 | Order ID | Categorical | Categorical |
| 2 | Order Date | Date | Continuous |
| 3 | Ship Date | Date | Continuous |
| 4 | Ship Mode | Categorical | Categorical |
| 5 | Customer ID | Categorical | Categorical |
| 6 | Segment | Categorical | Categorical |
| 7 | Postal Code | int | Discrete |
| 8 | City | Categorical | Categorical |
| 9 | State | Categorical | Categorical |
| 10 | Country | Categorical | Categorical |
| 11 | Region | Categorical | Categorical |
| 12 | Market | Categorical | Categorical |
| 13 | Product ID | Categorical | Categorical |
| 14 | Category | Categorical | Categorical |
| 15 | Subcategory | Categorical | Categorical |
| 16 | Product Name | Categorical | Categorical |
| 17 | Sales | float | Continuous |
| 18 | Quantity | int | Discrete |
| 19 | Discount | float | Continuous |
| 20 | Profit | float | Continuous |
| 21 | Shipping Cost | float | Continuous |
| 22 | Order Priority | Categorical | Categorical |

# 3. Data preparation and Manipulation

Cleaning and converting raw data before processing and analysis is known as data preparation. It's a crucial stage before processing that often include reformatting data, making data changes, and integrating data sets to improve data.

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Text

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

A picture containing text

Description automatically generated

Application

Description automatically generated with medium confidence

Table

Description automatically generated

A picture containing graphical user interface

Description automatically generated

# 4. Exploratory Data Analysis

Exploratory Data Analysis refers to the fundamental way of conducting preliminary data investigations with the use of summary statistics and graphical representations to identify trends, identify anomalies, test hypotheses, and verify conclusions.

## 4.1 Data Visualization

Chart, bar chart

Description automatically generated

The above bar chart displays the total number of orders by market areas which clearly shows Africa has the minimum order counts (less than 5000). The Asia pacific has maximum orders placed while Europe is following it and remaining LATAM and USCA are almost identical at close to 10k.

Chart, histogram

Description automatically generated

The above histogram visualization shows the frequency distribution of the quantity ordered. The maximum ordered quantity is from 0 to 2 which is greater than 20000. It is then followed by 2 to 3, the frequency for which is close to 10000. This plot clearly shows as the quantity gets increased the frequency count gets decreased. The 10 to 14 number of quantities has the least frequency. There is a huge gap between the 0 to 2 and remaining other quantities.

Chart, pie chart

Description automatically generated

The above pie chart shows the percentage sales by category that includes Technology, Furniture and Office Supplies. Product category "Technology" has contributed maximum sales which is 37%. It is then followed by "Furniture” which is 33%. "Office Supplies" has contributed the least which is 30%.

Graphical user interface, application

Description automatically generated with medium confidence

A picture containing table

Description automatically generated

The above co-relation matrix chart provides the co-relationship information among different variables. The color gradient from Red to Blue describes the extent of co relationship among Sales, Quantity, Discount, Profit and Shipping Cost red being the negative co relationship and blue being the positive co-relationship. It can be seen that "Sales" and "Shipping Costs" are positively related. "Profit" and "Discount" are very weakly related. "Profit" and "Discount" are negatively related.

Graphical user interface, text, application, email

Description automatically generated

Chart, bar chart

Description automatically generated

In above bar plot we have grouped order by Priorities and calculated their mean shipping cost. We can clearly describe the graph as critical priority has the maximum mean value while medium has the lowest in mean range of values.

Chart, scatter chart

Description automatically generated

From above scatter points we can see more profits and loss have been availed from the standard class while there are small range of profits seen in increment of sales feature. First class shipping mode has some better positive in profit generation in compared to other modes.

Chart, scatter chart

Description automatically generated

From above plot we can explain that the Sales to Profit ratio is same in every category, no matter how they are clubbed. Furniture is neutral in comparison with others while mostly Office Supplies and Technology are in positive ranges.

Chart, bar chart

Description automatically generated

From above bar plot we can say that more sales have been incurred by the technology category and least sales are of Office Supplies. Most preferred order priority is of medium while lowest and critical priority are least used.

Chart, bar chart

Description automatically generated

From the above plots we can clearly say consumer have huge number of orders while home Office segment has a smaller number of orders in the store data. Asia pacific Market place has high number of order while Africa has contributed very low in numbers.

Chart

Description automatically generated

From the above facet wrap plots we can say Office supplies has more orders than other two categories. Comparing the order priorities low and critical priorities have very a smaller number of orders as compared to high and medium priorities.

Chart

Description automatically generated

From the above bar plot, it can be concluded that Losses have been incurred by the tables industry mainly in the Asia Pacific region while most profit generating markets are also in Asia pacific. Copiers, Phones, Chairs and Bookcases are almost identical in profit. Africa’s Market area have been very less active in profit contribution in all the subcategories.

# 5. Statistical Testing

Statistical testing examines if the given data adequately supports a specific hypothesis by testing the hypothesis, concluding the population using sample data, and evaluating the hypothesis. The T-test, Z-test, chi-square, ANOVA, binomial, and one-sample median test are examples of statistical tests. Statistical tests are chosen based on the data's structure, distribution, and variable type. The p-value of your results is computed using the test statistic, which aids in deciding whether to reject the null hypothesis.

The t-test compares the mean of two groups statistically. It is commonly used in hypothesis testing to see if a method or treatment influences the population of interest, or if two groups are distinct.

## 5.1 Hypothesis Testing

Hypothesis testing is a statistical procedure in which an analyst tests a hypothesis about a population parameter. The analyst's approach is determined by the type of the data and the purpose of the study. It is a technique for determining the outcome of a hypothesis based on a sample of data from a wider population.

Steps for Hypothesis Testing

The analyst should first state the two hypotheses so that only one could be correct.

The next step is to create an analysis plan that explains how the data will be analyzed.

The third stage is to set the strategy into action and analyze the sample data physically.

The fourth and last step is to analyze the data and decide whether the null hypothesis should be accepted or rejected.

## 5.2 Correlation Analysis

Correlation analysis is a statistical method for determining the strength of a relationship between two numerically measured continuous variables (for example, height and weight). When a researcher wants to see if there are any relevant links between variables, this analysis is helpful. Correlation analysis is frequently mistaken as determining cause and effect; however, this is not the case because other variables not included in the study could have influenced the outcomes.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

# 6. predictive Data Analysis

## 6.1 Linear Regression

Linear regression analysis is a statistical technique for predicting the value of one variable based on the value of another. The dependent variable is the variable you want to forecast. The independent variable is the one you are using to forecast the value of the other variable. It is one of the essential and most basic regression techniques which is extensively used.

This type of analysis involves one or more independent variables that best predict the value of the dependent variable to estimate the coefficients of the linear equation. It creates a straight line or surface that reduces the difference between expected and actual output values. Simple linear regression calculators that employ the "least squares" method to get the best-fit line for a set of paired data are available and uses Y (independent variable) to calculate the value of X (dependent variable).

Chart, box and whisker chart

Description automatically generated

A new data frame df\_standardize is created for standardization as the distribution of the features are not in gaussian or normal distributed curve. Then we sorted the Sales data in descending order as it has outliers in the boxplot above. So, we removed the last 20 rows and sorted them.

Chart, box and whisker chart

Description automatically generated

In the above figure, we sorted the profit column in descending order as the profits are in negative index so that we can remove those outlier data after the standardization. Top 15 index are removed from the data and sorted it and assigned them to the original data frame for easy assigning of variable.

Chart, box and whisker chart

Description automatically generated

In the above figure, we sorted the profit column in ascending first so that we can remove those outlier data after the standardization. Top 15 index are removed from the data and sorted it and assigned them to the original data frame for easy assigning of variable.

Graphical user interface, text, application, email

Description automatically generated

Applying the four independent features from the data frame and using linear regression to create a model using Sales as the dependent feature.

Chart, scatter chart

Description automatically generated

Above figure demonstrates us about the predicted and model created visualizations and assigned it in a variable.

Graphical user interface, text, application, email

Description automatically generated

In this figure the summary of the model created is shown.

Graphical user interface, text, application, email

Description automatically generated

We now split the data into two parts test and train which are at a ratio of 20 and 80 respectively.

Chart, scatter chart

Description automatically generated

A predicted visualization is shown for trained data from the model created where x axis represents the predicted value from model and y is the Sales values.



Chart, scatter chart

Description automatically generated

Here the graph of test model is shown where x axis represents the predicted on test model and Sales data on y axis.

Graphical user interface, application

Description automatically generated

In above figure the Root mean square is shown which is 72 percentage so that we can confirm this model as an average model.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

The above figure shows the residuals to calculate the RMSE and Standard Deviation on test models. We can say it average model as the RMSE is less in compared to Standard Deviation.