Data Analytics In Supermarkets

Student Name

Institution Name

8/5/2023

**Introduction**

Improving business performance is often a priority for companies to remain competitive and achieve long-term success. Data analysis has emerged as a crucial aspect of every business's operation as it helps companies to make informed decisions based on the insights and patterns retrieved from the data. In today's era of enormous technological uses, organizations have access to large amounts of data, and with proper analysis, they could yield patterns that are not easily recognized. Practical data analysis can increase organizations' profitability, increase competition power, and improve the efficiency of the business.

Supermarkets are large shops offering various food, beverages, and household products. They are more prominent and extensive than grocery stores and tend to have broad sections. The primary objectives of supermarkets across the globe are to sell products and generate high profits. Due to increased competition from retailers, shops, and mini-markets, supermarkets need to optimize their operations to generate large amounts of money. The supermarket data set contains 9994 observations of 13 variables. Shipping mode, segment, nation, city, state, Region, postal code, Region, category, subcategory, sales, quantity, discount, and profit are among the variables. The data set can be found in the Kaggle data set repository.

**Problem statement**

Like any other institution, supermarkets are exposed to various risks and face numerous challenges. Supermarkets must understand customer needs, wants, preferences, and behaviors to provide the best services. To offer better customer services, supermarkets must apply data analytics to maximize their returns. The goal of supermarkets is maximizing profits and minimizing costs. The dataset analysis will involve different aspects of data analysis, including, Exploratory data analysis, data cleaning, data visualization, and prescriptive and descriptive analysis. EDA will be used to understand the data better. The data will be cleaned to remove errors that may cause investigation inaccuracies. Descriptive statistics will be used to summarize the data, while correlation analysis will be used to show the relationship between the different variables. The analysis will provide the supermarket industry with valuable information that will help them improve their manufacturing process and the quality of their products, improving their overall efficiency.

**Data Cleaning**

Data cleaning is an essential process in the data analysis process. The process involves removing the duplicates, handling the missing values, and correcting the inconsistencies to ensure the data is accurate, reliable, and complete (Imtiaz & Shah, 2008). In the case of the Supermarket store data, the dataset contained inconsistencies and errors that needed to be addressed. R programming software was used to clean the supermarket store dataset. The first step involved ensuring all the columns were in the correct data types.

**Data Types**

The Ship mode, Region, and Category columns were in the character data type and must be transformed into the factor data type. A function in R software transforms the variables from character to factor data type to achieve this.

**Handling Missing Values**

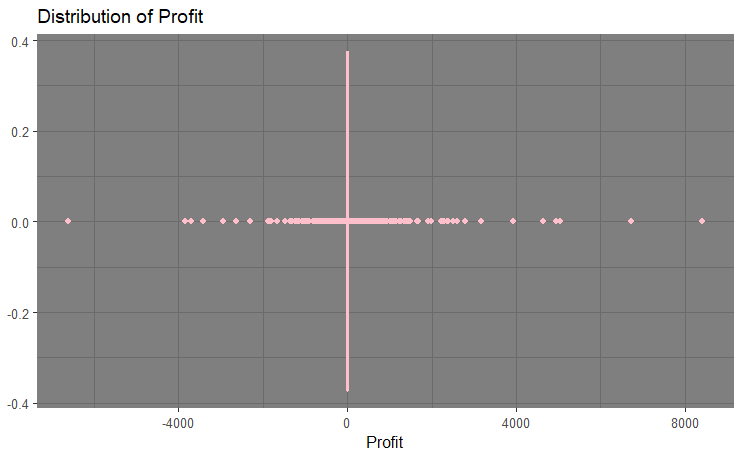
The second step in the data-cleaning process involved identifying missing values and choosing the best way to deal with them according to the data(Pellegrin & Zentner, 2021). The supermarket stores data set contained no missing observations.

**Duplicates**

The final steps in the data cleaning involved Identifying and removing duplicates in the dataset. The supermarket stores data set contained no duplicated values.

**Outlier Detection**

This step involved identifying outliers present in the supermarket store's data and omitting them. Outliers can significantly have an impact on statistical analysis and machine learning algorithms. The outliers were detected by generating a box plot of the data. This was done by first loading the ggplot2 package and using the geom\_boxplot function to generate the boxplot.



The boxplot above showed that the data contained a lot of outliers, and there was a need to omit them before any analysis was done since the process led to improved accuracy, better visualization, improved predictive modeling, and enhanced insights.

**Exploratory Data Analysis**

**Descriptive Statistics**

Exploratory data analysis is a crucial step in data analysis as it helps to understand the data, identify the patterns, and look at the relationships between the variables (Pearson, 2018). For this project, EDA was conducted by generating a summary of the data to gain insights and understand the data more. The first step was to explore the numerical and integer variables present in the data as follows;

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sales | Quantity | Discount | Profit |
| Minimum | 0.44 | 1.00 | 0.000 | -6599.98 |
| 1st Qu | 17.28 | 2.00 | 0.000 | 1.729 |
| Median | 54.49 | 3.00 | 0.200 | 8.666 |
| Mean | 229.858 | 3.79 | 0.156 | 28.657 |
| 3rd Qu | 209.94 | 5.00 | 0.200 | 29.364 |
| Max | 22638.48 | 14.00 | 0.800 | 8399.976 |

The summary above contained the data's minimum, maximum, median, mean, and quartiles. It was observed that the maximum discount, profit, quantity, and sales in the supermarket were 0.8, 8399.976, 14.0, and 22638.480, respectively. The mean of the discount is 0.1562, which is lower than its median of 0.2. This means that the discount is skewed to the left. Most discounts are concentrated at the lower end of the range. The means of the other numerical variables were higher than the medians, which meant that most sales and quantities were concentrated at the upper end of the range. In most sales, zero discount was offered because small quantities of goods were bought. This showed that supermarkets sold high quantities of goods to maximize profit while offering small discounts on sales. The descriptive statistics of the profit column revealed that the mean of the profits was 28.657, with a standard deviation of 234.2601. The median value of the profit column was 8.666. The mode was unavailable, but the variance was 54877.8, indicating how the profits were spread over a wide range. The total profits obtained in the supermarkets were 286397. The sample variance measures how much the profit values vary from the mean; in this case, it is 54877.8. A low variance shows that the data is clustered and grouped closely around the norm, while a significant variance indicates that the data is spread out. The kurtosis value is 399.98, which shows that the profit distribution differs from a normal distribution.

The second step involved checking on the factors present in the categorical variables;

There were 3 methods of shipping present in the supermarket. It is seen that the standard class shipping had the most customers.

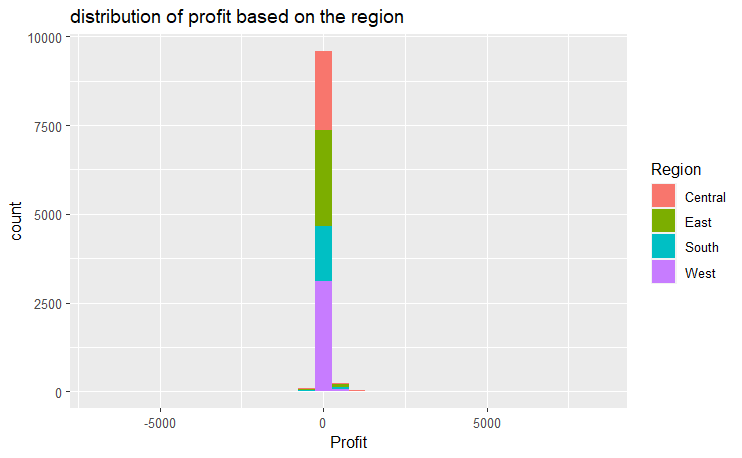
First Class Same Day Second Class Standard Class

1538 543 1945 5968

The supermarket served citizens in the United States across 4 regions: Central, East, South, and West. It was observed that most customers were from the Western Region while the least was from the South Region.

Central East South West

2323 2848 1620 3203



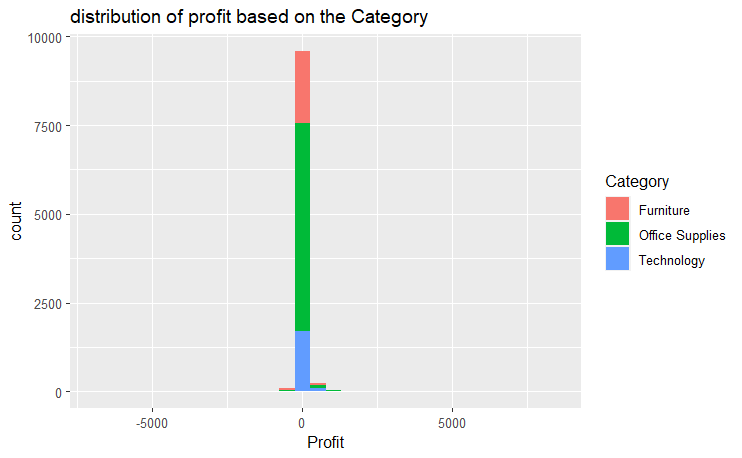
The histogram shows that the western Region contributes the most profit in the supermarket.

The supermarket offered three main categories of goods, as shown below. Office supplies goods were the most bought commodities in the supermarket store. This means that office supplies were the most and technological goods were in the least demanded.

Furniture Office Supplies Technology

2121 6026 1847

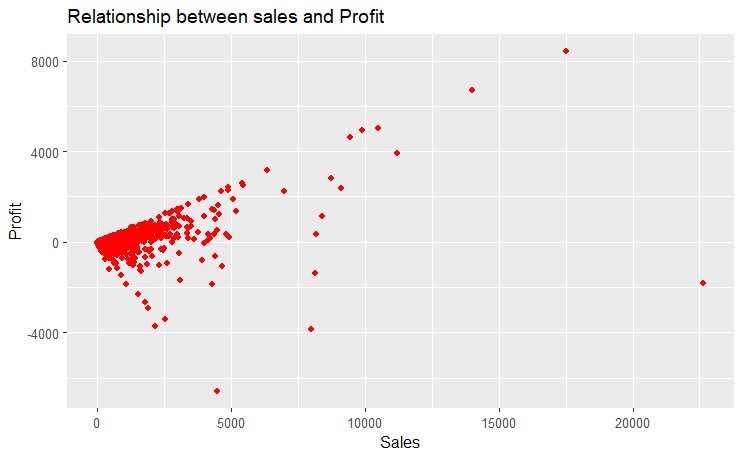
The histogram below shows that office supplies are the most sold while technology goods are the least. Office supplies generate the most profits for the supermarket. The supermarket should invest more in office supplies so as maximize their profit.



**Correlation analysis**

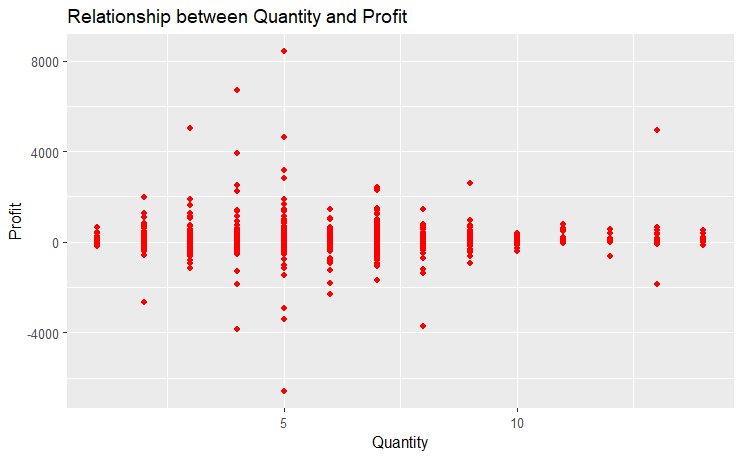
This step involved checking the data set's relationship between numerical and integer variables. The correlation coefficient ranges from -1 to 1; a correlation near 1 indicates a strong relationship, while a correlation near 0 indicates a weak connection. The relationship between profit and other variables was as follows;

**Relationship Between Profit and Sales**

****

There is a moderate positive relationship between sales and profit. This means that as sales increase, the profits increase too. The p-value 2.2e-16 is extremely low, meaning the relationship between profit and sales is statistically significant; hence it did not just occur by chance. The 95% confidence intervals between 0.4638148 and 0.4940301 mean that 95% of the confidence would lie between 46% and 49% when many samples were taken. The graph shows that as sales increase, profit also increases. For the supermarket to generate the most profits, it should make the maximum sales.

**Relationship Between profit and Quantity**

****

There is a weak relationship between profit and quantities of goods sold. The p-value 3.362e-11 is extremely low, meaning that the relationship between profit and sales is statistically significant; hence it did not just occur by chance. The 95% confidence intervals between 0.04670791 and 0.08574776 mean that 95% of the confidence would lie between 4% and 9% when many samples were taken. But it is also seen that the profit increases as the quantity of goods increases.

**Regression Analysis**

**Multiple linear regression to predict profit**

This step involved examining the relationship between the response and explanatory variables. The response variable was profit, while the explanatory variables were quantity discounts and sales. Multiple linear regression was built to predict the supermarket's profit to achieve this. Multiple regression was used since independent variables were more than one.

The regression model to predict profit in the supermarket was as follows: Profit = 3.497e+01 + 1.800e-01\*Sales - 2.962e+00\*Quantity - -2.335e+02\*Discount

Call:

lm(formula = Profit ~ Sales + Quantity + Discount, data = supermarket\_data)

Residuals:

Min 1Q Median 3Q Max

-7266.8 -23.8 -0.4 25.6 5229.6

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.497e+01 4.218e+00 8.291 < 2e-16 \*\*\*

Sales 1.800e-01 3.275e-03 54.961 < 2e-16 \*\*\*

Quantity -2.962e+00 9.171e-01 -3.230 0.00124 \*\*

Discount -2.335e+02 9.686e+00 -24.101 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199.8 on 9990 degrees of freedom

Multiple R-squared: 0.2727, Adjusted R-squared: 0.2725

F-statistic: 1249 on 3 and 9990 DF, p-value: < 2.2e-16

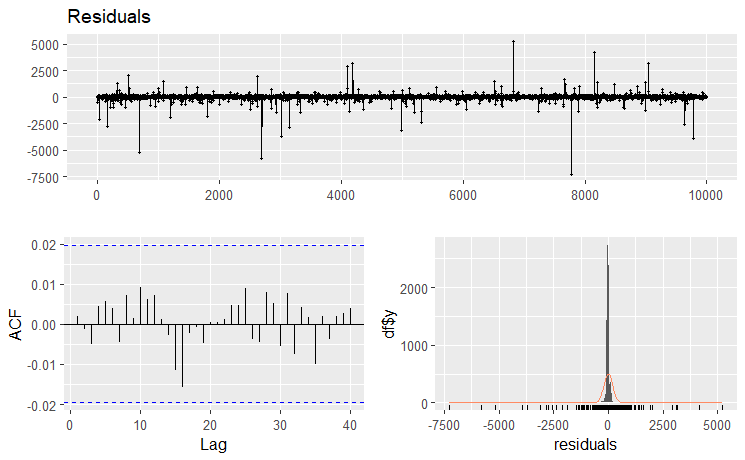
The coefficients of the model were both positive and negative. The coefficient of sales is positive, meaning that for every unit increase in sales in the supermarket, the profit increases by 1.800e-01. A coefficient of -2.335e+02 of discount means that for every unit increase in the discount, the profit decreases by -2.335e+02. The model intercept is 3.497e+01 meaning that when all the independent variables are zero, the supermarket's profit is 3.497e+01. The coefficient of determination represents the fraction of the variance in the dependent variable that can be explained by the explanatory factors.. A coefficient of determination R-squared(R) of 0.2727 indicates that the model explains 27.27% of the profit. A p-value less than 5% is significant, and an F-statistic of 1249 suggests that our model is statistically significant. The residual error of our model is 199.8 meaning the measure of residuals around the regression line.

Residuals indicate the difference between the actual and predicted values. They represent the portion of the dependent variable not explained by the model. A good regression model always has normally distributed residuals.  
 Breusch-Godfrey test for serial correlation of orders up to 10

data: Residuals

LM test = 2.609, df = 10, p-value = 0.9892

The Breusch-Godfrey test is used in testing autocorrelation between residuals. Autocorrelation of residuals occurs if the errors of a residual model are dependent. A p-value of 0.9 is greater than 5% significance, meaning we can’t reject the null hypothesis. This means there is no serial correlation between the residuals; hence they are independent.



Based on the above chart, the residuals are normally distributed, meaning our model is a good fit and statistically acceptable. From the histogram, it is bell-shaped, with a single pick at the center and a broad width.

**Recommendations**

Based on the insights drawn from the data, I recommend the following to supermarket stores to maximize profitability and returns.

* The supermarket store should minimize offering huge discounts to its customers. It is seen from the regression model that as the discount increases, the profit decreases.
* The supermarket should advertise its products in the south and central regions. This is because there is less demand for goods in the south and central regions than in the west and northern regions.
* The supermarket should understand its customer's needs and preferences. It is seen that office supplies goods are in high demand; hence it should always ensure it has ready items.
* Offering better and satisfactory services- The supermarket store should invest much in the standard class shipping mode since it has been used by most customers recently. By investing in standard shipping mode, the supermarket ensures its services are efficient, thus improving customer satisfaction.

**Conclusion**

Data analytics has been an important aspect of almost every business, including supermarkets, due to the gradual production of data. Over the recent years, supermarkets have produced large amounts of data which, analyzed properly, can help the manager draw hidden patterns and relationships between variables present. In this project, data cleaning, visualization, descriptive statistics, and prescriptive modeling have been covered. Exploratory data analysis helps to understand the data more. In a nutshell, data analytics is a valuable tool for supermarkets to maximize their profits. Supermarkets should always understand the customers' wants and preferences and meet their expectations to ensure success.

**References**

Imtiaz, S. A., & Shah, S. L. (2008). Treatment of missing values in process data analysis. The Canadian Journal of Chemical Engineering, 86(5), 838-858. <https://doi.org/10.1002/cjce.20099>

Pearson, R.K. (2018) “Exploratory Data Analysis: A first look,” *Exploratory Data Analysis Using R*, pp. 79–140. Available at: https://doi.org/10.1201/9781315382111-3.

Pellegrin, N. and Zentner, A. (2021) “Identifying missing values,” *Cleansing Data* [Preprint]. Available at: https://doi.org/10.4135/9781071868317.

**Appendix**

*supermarket\_data<-read.csv("~\\SampleSuperstore.csv")*

*# Structure of the data set*

*str(supermarket\_data)*

*# Changing the data types of variables*

*supermarket\_data$Ship.Mode<-as.factor(supermarket\_data$Ship.Mode)*

*supermarket\_data$Region<-as.factor(supermarket\_data$Region)*

*supermarket\_data$Category<-as.factor(supermarket\_data$Category)*

*# Checking missing values*

*colSums(is.na(supermarket\_data))*

*# Checking duplicates*

*duplicated(supermarket\_data)*

*# Checking the presence of outliers in profit*

*library(ggplot2)*

*ggplot(data=supermarket\_data,aes(Profit))+geom\_boxplot(col="pink")+ggtitle("Distribution of Profit")+theme\_dark()*

*# Ommiting outliers*

*Q1 <- quantile(supermarket\_data$Profit, 0.25, na.rm = TRUE)*

*Q3 <- quantile(supermarket\_data$Profit, 0.75, na.rm = TRUE)*

*IQR <- Q3 - Q1*

*lower <- Q1 - 1.5 \* IQR*

*upper <- Q3 + 1.5 \* IQR*

*outliers <-supermarket\_data$Profit < lower | supermarket\_data$Profit > upper*

*supermarket\_data$Profit[outliers] <- NA*

*supermarket\_data <- supermarket\_data[!outliers, ]*

*# Descriptive statistics of the numerical and integer variables.*

*library(tidyverse)*

*supermarket\_data %>% select(Sales,Quantity,Discount,Profit) %>% summary()*

*var(supermarket\_data$Profit)*

*var(supermarket\_data$Quantity)*

*var(supermarket\_data$Discount)*

*var(supermarket\_data$Sales)*

*# totals*

*sum(supermarket\_data$Profit)*

*sum(supermarket\_data$Quantity)*

*sum(supermarket\_data$Discount)*

*sum(supermarket\_data$Sales)*

*sum(supermarket\_data$Profit)*

*sd(supermarket\_data$Profit)*

*var(supermarket\_data$Profit)*

*library(moments)*

*kurtosis(supermarket\_data$Profit)*

*# Factors present in the data*

*table(supermarket\_data$Ship.Mode)*

*table(supermarket\_data$Region)*

*table(supermarket\_data$Category)*

*# Visualization*

*ggplot(data = supermarket\_data,aes(Profit,fill=Region))+geom\_histogram(bins=30)+ggtitle("distribution of profit based on the region")*

*ggplot(data = supermarket\_data,aes(Profit,fill=Category))+geom\_histogram(bins=30)+ggtitle("distribution of profit based on the Category")*

*ggplot(data = supermarket\_data,aes(Profit,fill=Ship.Mode))+geom\_histogram(bins=30)+ggtitle("distribution of profit based on the Ship.Mode")*

*cor.test(supermarket\_data$Quantity,supermarket\_data$Sales)*

*ggplot(data = supermarket\_data,aes(Sales,Profit))+geom\_point(col="red")+ggtitle("Relationship between sales and Profit")*

*cor.test(supermarket\_data$Profit,supermarket\_data$Sales)*

*ggplot(data = supermarket\_data,aes(Quantity,Profit))+geom\_point(col="red")+ggtitle("Relationship between Quantity and Profit")*

*cor.test(supermarket\_data$Profit,supermarket\_data$Quantity)*

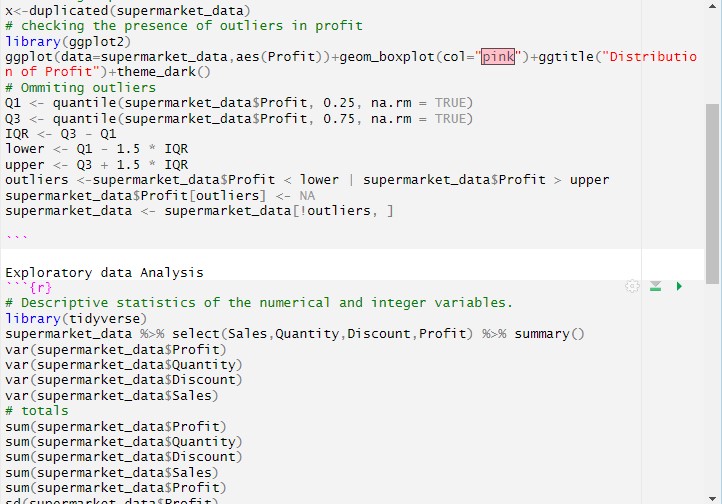
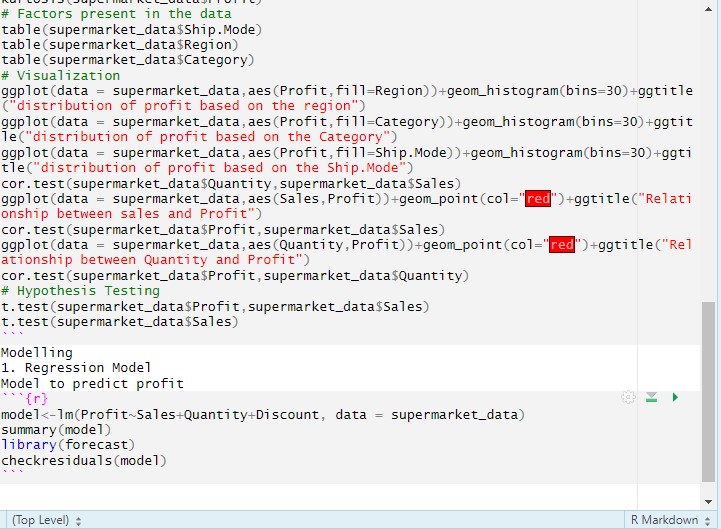
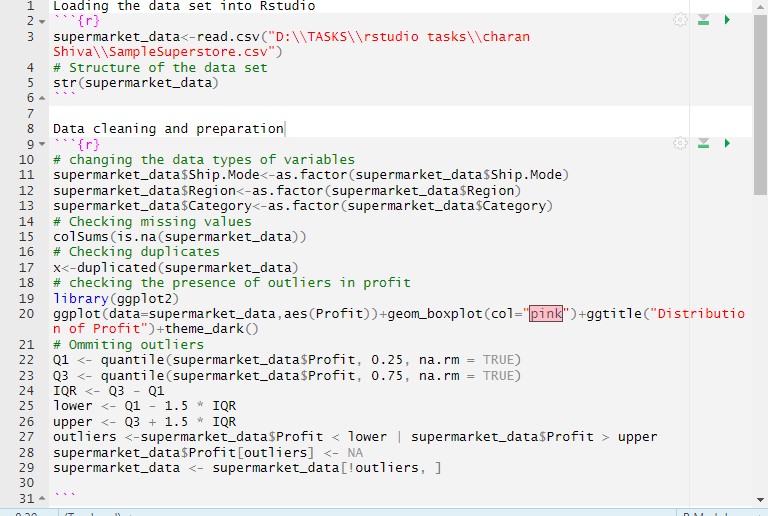
*model<-lm(Profit~Sales+Quantity+Discount, data = supermarket\_data)*

*summary(model)*

*library(forecast)*

*checkresiduals(model)*

**SCREEN CAPTURE OF THE CODE**

**