Exploratory Data Analysis- Superstore Sales

Overview

Exploratory Data Analysis provides a sense of the data and often sets the initial roadmap for our data analysis. Often Data scientists spend maximum time in Exploratory Data analysis so it would be worthwhile to automate some tasks for performing EDA.

This paper presents an analysis of sales data from a hypothetical SuperStore. The objective is to uncover insights that can drive business decisions, focusing on various aspects of the data, including univariate, bivariate, and multivariate analysis.

Data Cleaning

Before delving into the analysis, it was necessary to ensure the data was clean and in the correct format. One of the primary steps was to change the datatype of the date column to DateTime, which allows for more efficient and accurate time-series analysis.

Feature Engineering

Two features were calculated using the ‘Date’ and ‘Time’ features in the dataset.

- Day
- TimeoftheDay

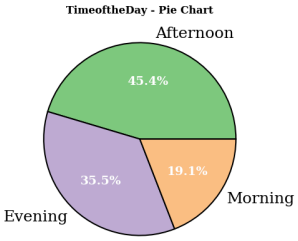
Data Visualization

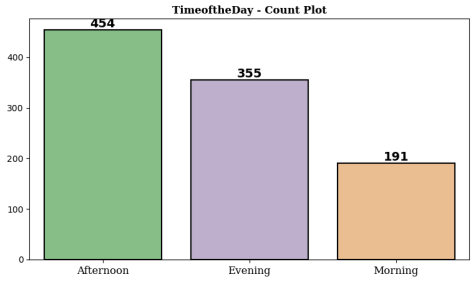
Univariate Analysis

Categorical Variables

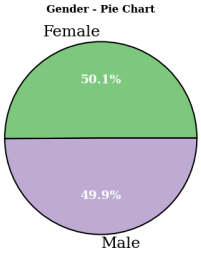
Pie charts and count plots were plotted for all the categorical variables. Following observations were made.

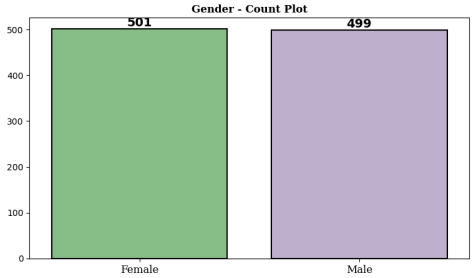
- Around 81% purchases are made in the afternoon and evening hours as shown below.





- The SuperStore has an almost equal number of male and Female customers.

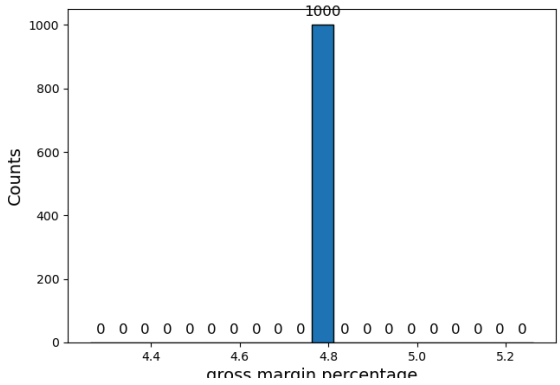


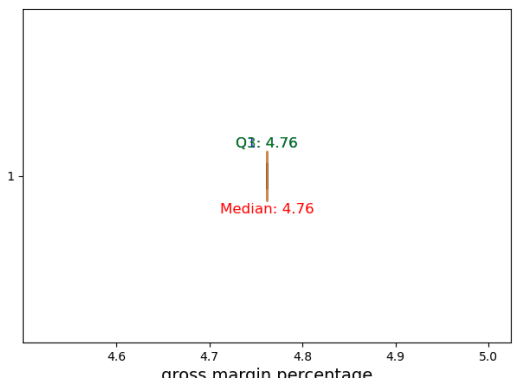


Continuous Variables

Histograms and Box plots were used to visualize distribution of the numerical variables of the dataset. Following points may be noted.

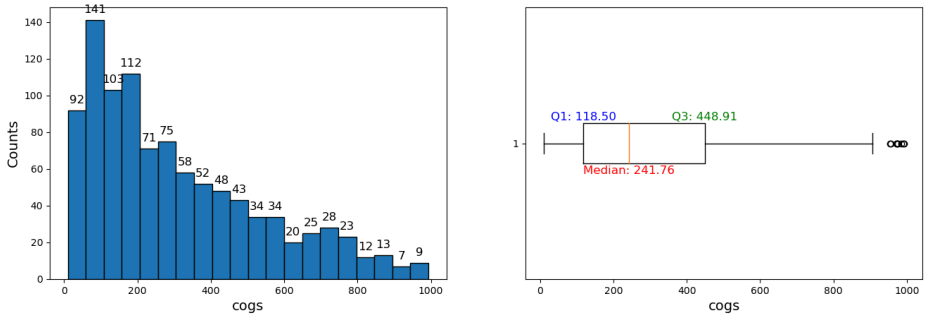
- ‘Gross margin percentage’ is a constant variable.



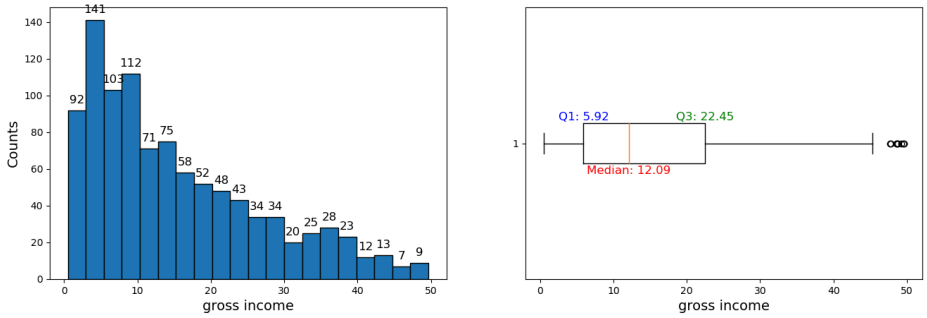


- Variables ‘cogs’, ‘gross income’, ‘Tax 5%’ and ‘Total’ have similar distribution and must be perfectly correlated.

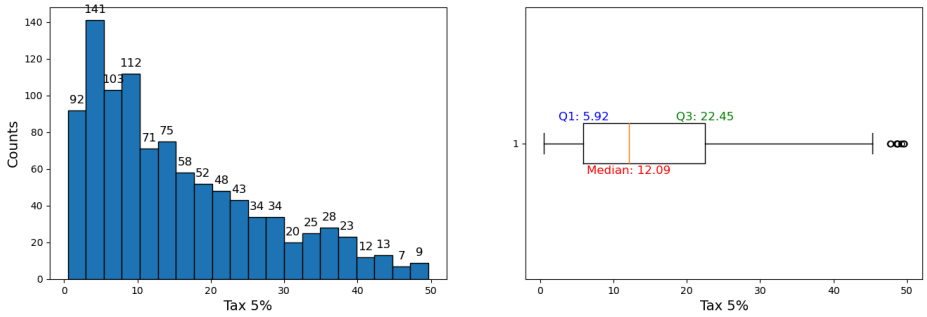
Distribution of Continuous Variables



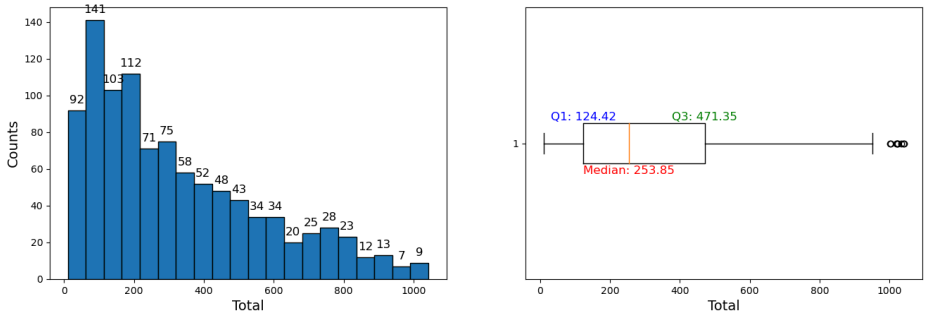
Distribution of Continuous Variables



Distribution of Continuous Variables



Distribution of Continuous Variables

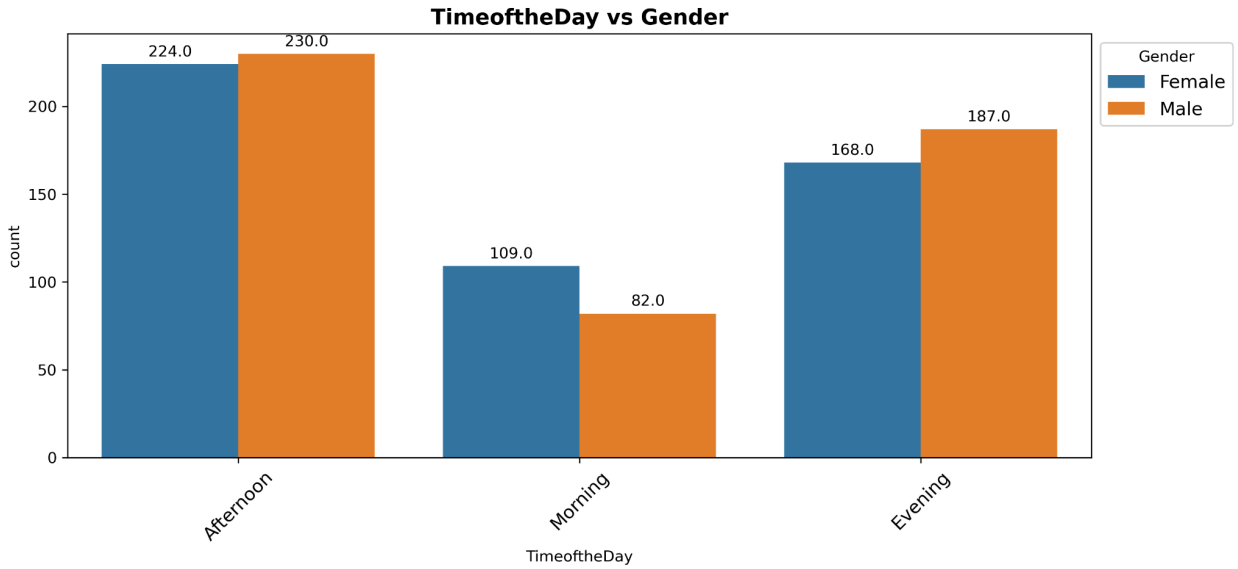


Bivariate analysis

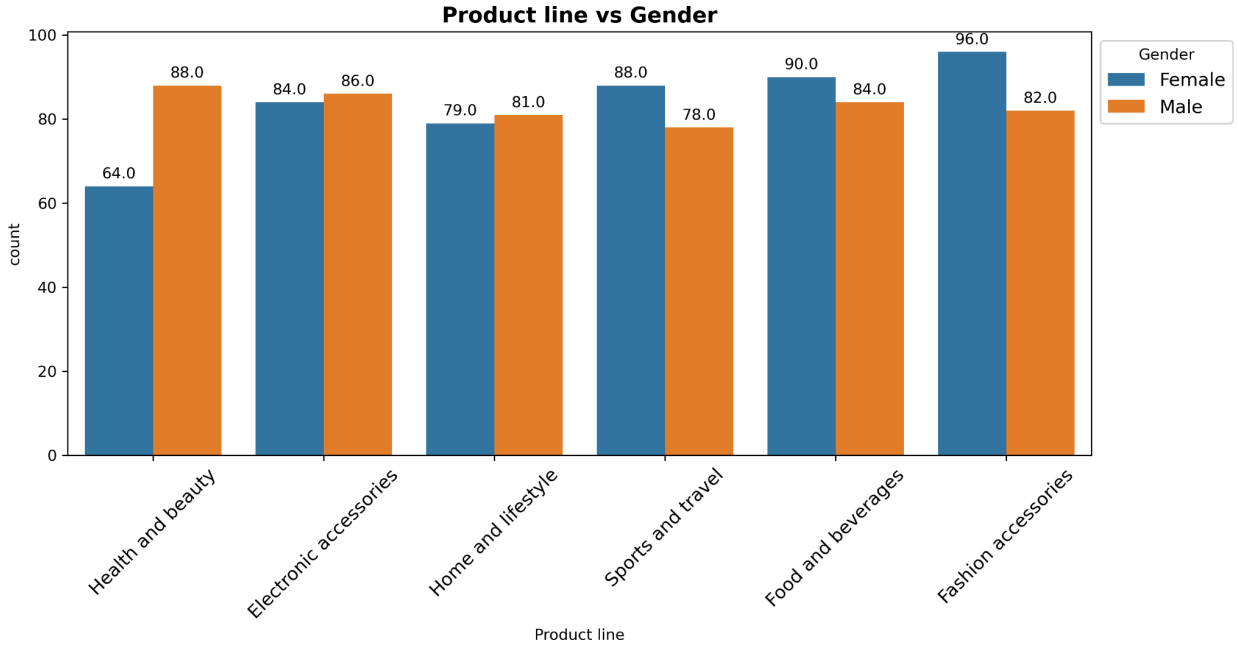
Categorical Features vs. Categorical Features

Count Plots for different combinations of the continuous variables were displayed.

- Afternoon is the popular shopping time among both genders.



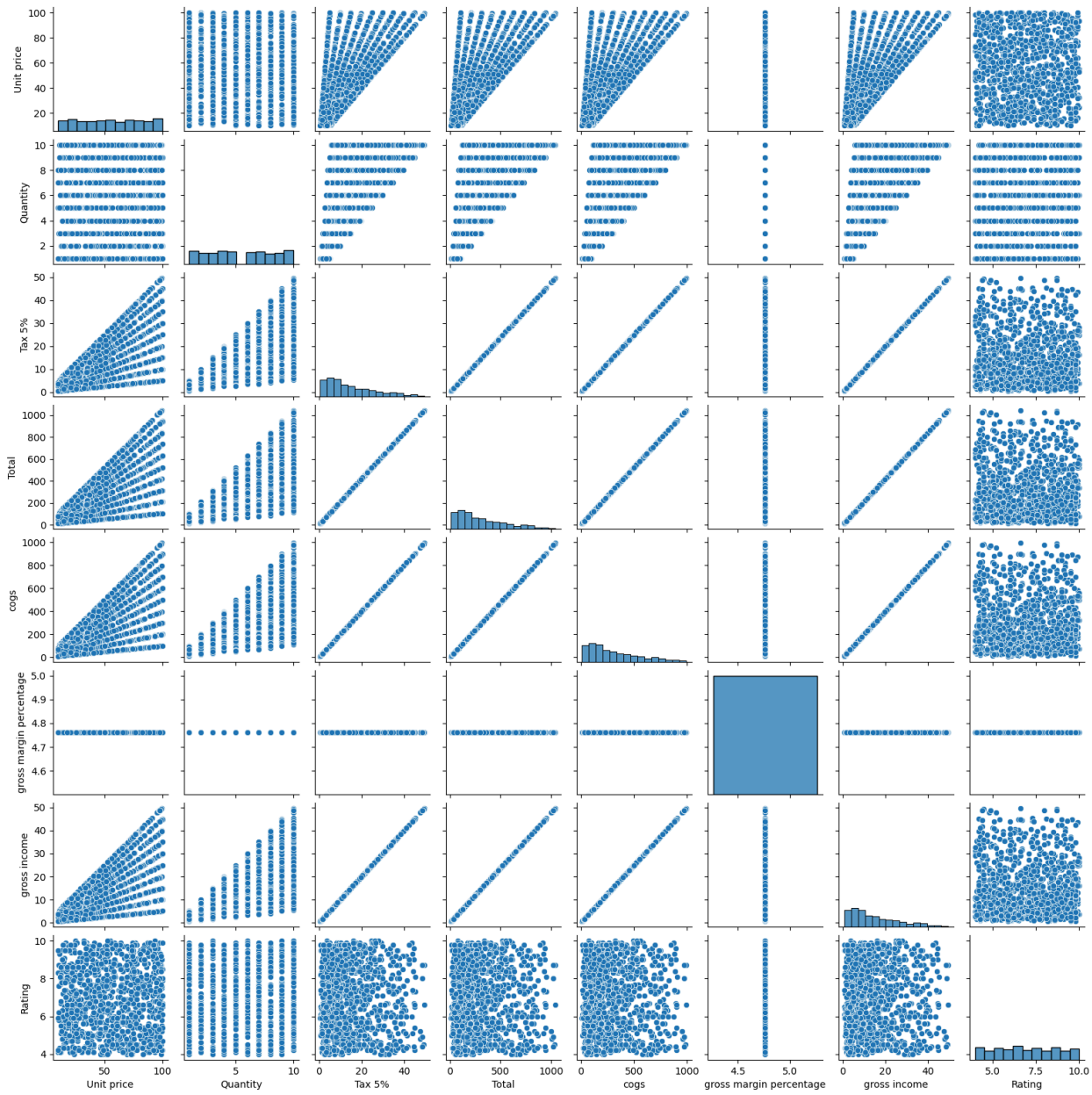
- 'Health and beauty' segment is more popular among males whereas 'Fashion accessories' is slightly more popular among females.



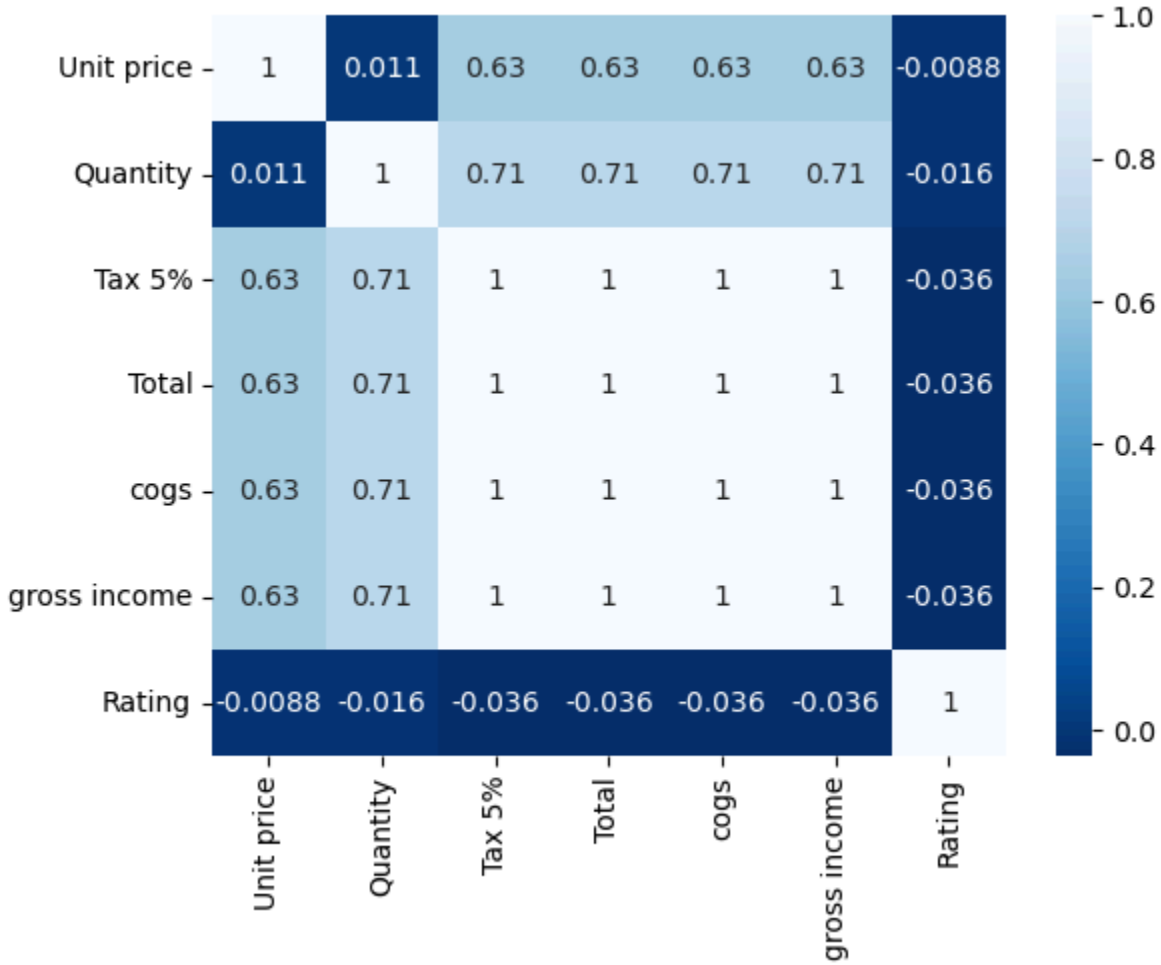
Continuous Features vs. Continuous Features

Pair plots and correlation matrix were used to visualize relationships among different combinations of the numerical variables of the dataset.

- We can see that some variables are perfectly correlated and ‘gross margin percentage’ is a constant variable.



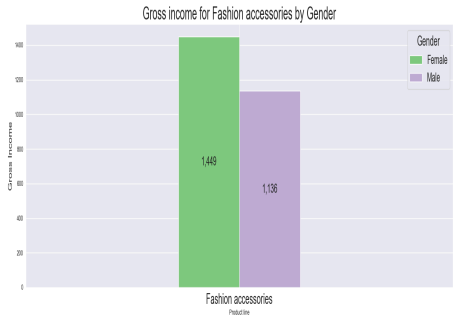
- Correlation matrix shows that ‘gross income’, ‘cogs’, ‘Total’ and ‘Tax 5%’ are perfectly positively correlated with each other. We can use ‘gross income’ for further analysis and ignore other variables.

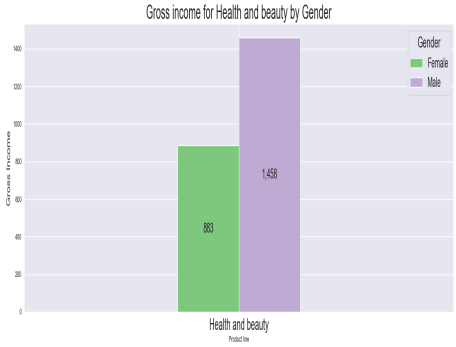


Multivariate analysis

Categorical Features vs Numerical Features

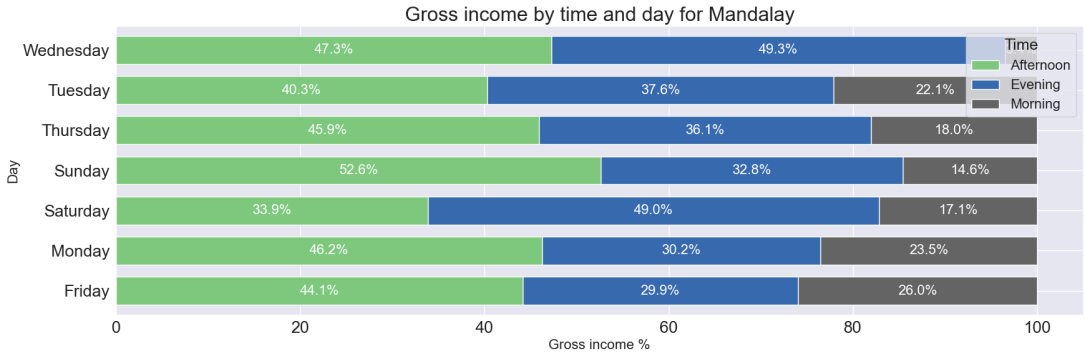
- Earlier we observed that the ‘Health and Beauty’ segment is more popular among males and the ‘Fashion accessories’ segment is slightly more popular among females. We can further check it in terms of the gross income that the store is earning for these two segments based on gender. As shown below, males are bringing more revenue to the ‘Health and Beauty’ segment compared to female customers of the store.

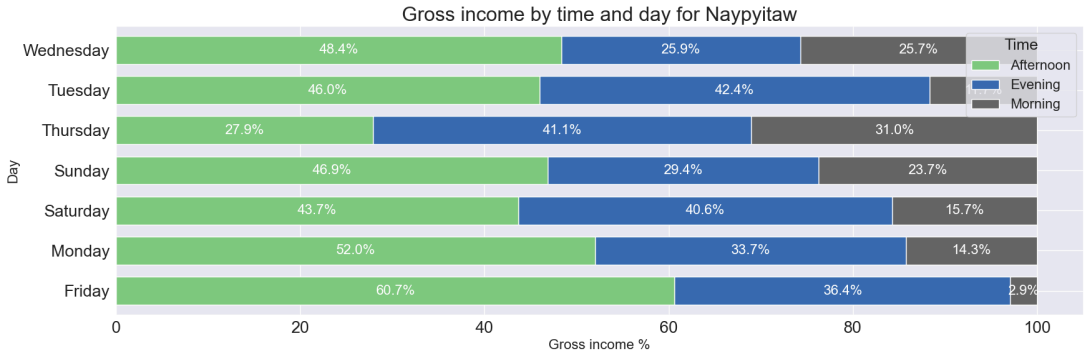


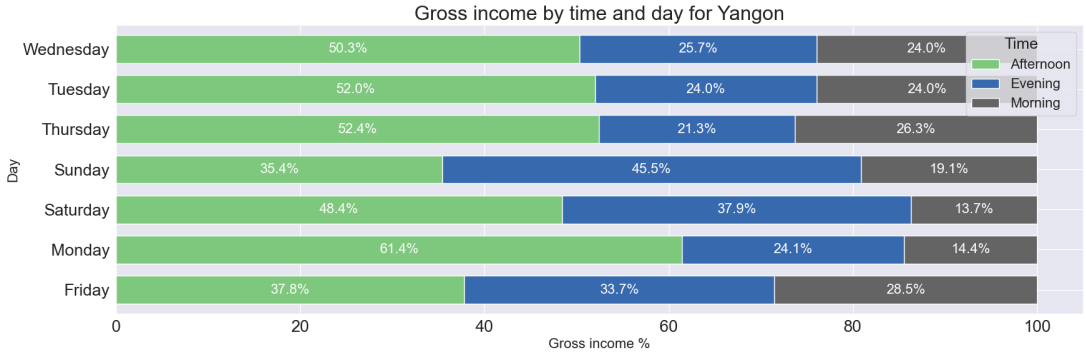


Product line	Electronic accessories	Fashion accessories	Food and beverages	Health and beauty	Home and lifestyle	Sports and travel
Female	1,290.57	1,449.40	1,579.57	883.86	1,430.33	1,360.70
Male	1,296.93	1,136.60	1,094.00	1,458.70	1,134.53	1,264.20
Total	2,587.50	2,585.99	2,673.56	2,342.56	2,564.85	2,624.90

- We further evaluated popular shopping time among the customers and mostly customers like to shop in afternoon and evening hours compared to morning hours. It can be seen that afternoon is the most popular time of shopping for all the cities and for all the days of the week.







TimeoftheDay	Afternoon	Evening	Morning
Friday	47.67%	33.39%	18.94%
Monday	53.62%	28.94%	17.44%
Saturday	41.39%	42.96%	15.65%
Sunday	43.83%	36.36%	19.80%
Thursday	42.50%	32.61%	24.89%
Tuesday	45.66%	35.29%	19.05%
Wednesday	48.64%	32.65%	18.71%

Conclusion

Following data insights were inferred from the above data visualization.

- Store has almost an equal number of male and female customers.
- Products from the ‘Health and Beauty’ segment are more popular among the males.
- Afternoon hours are the busiest time for the store.