PROJECT DOCUMENTATION

Exploratory Data Analysis

Supermarket Sales

Name: JEEVITHA P

Course Name: Data Analytics

Index

1. Abstract

The goal of this project is to perform an **Exploratory Data Analysis (EDA)** on a supermarket's sales data to gain actionable insights. Through data cleaning, visualization, and statistical testing, we aim to answer key business questions such as sales trends, customer spending habits, and relationships between variables. We also conduct hypothesis testing to determine whether gender has a significant effect on customer spending. The findings will help the supermarket understand customer behavior and optimize operations for improved sales performance.

2. Tools Used and Steps Involved

Tools Used:

- **Python:** For data manipulation and analysis.
- **Pandas:** For data cleaning and transformation.
- **Matplotlib and Seaborn:** For data visualization.
- **Scipy:** For hypothesis testing.
- **Jupyter Notebook:** For code execution and presentation.

Steps Involved:

1. **Data Collection:** The supermarket sales data was collected from a CSV file containing customer demographics, purchase information, and ratings.
2. **Data Cleaning:** Missing values were handled by filling numerical values with the median and categorical values with the mode. Outliers were identified using both IQR and Z-score methods.
3. **Univariate, Bivariate, and Multivariate Analysis:** Key variables were visualized to understand their distributions, relationships, and correlations.
4. **Business Questions:** Top five business questions were framed and analyzed.
5. **Hypothesis Testing:** A hypothesis was formed regarding gender-based differences in customer spending, and a t-test was conducted.

3. Business Questions and Answers

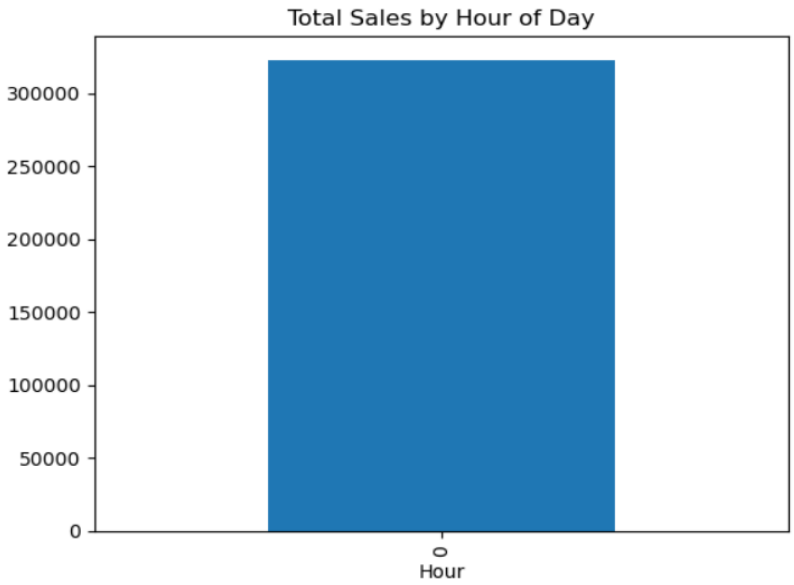
1. Which product lines generate the highest sales?

```
Product line
Food and beverages         56144.8440
Sports and travel          55122.8265
Electronic accessories     54337.5315
Fashion accessories        54305.8950
Home and lifestyle         53861.9130
Health and beauty          49193.7390
Name: Total, dtype: float64
```

2. Do male or female customers spend more on average?

```
Gender
Female    335.095659
Male      310.789226
Name: Total, dtype: float64
```

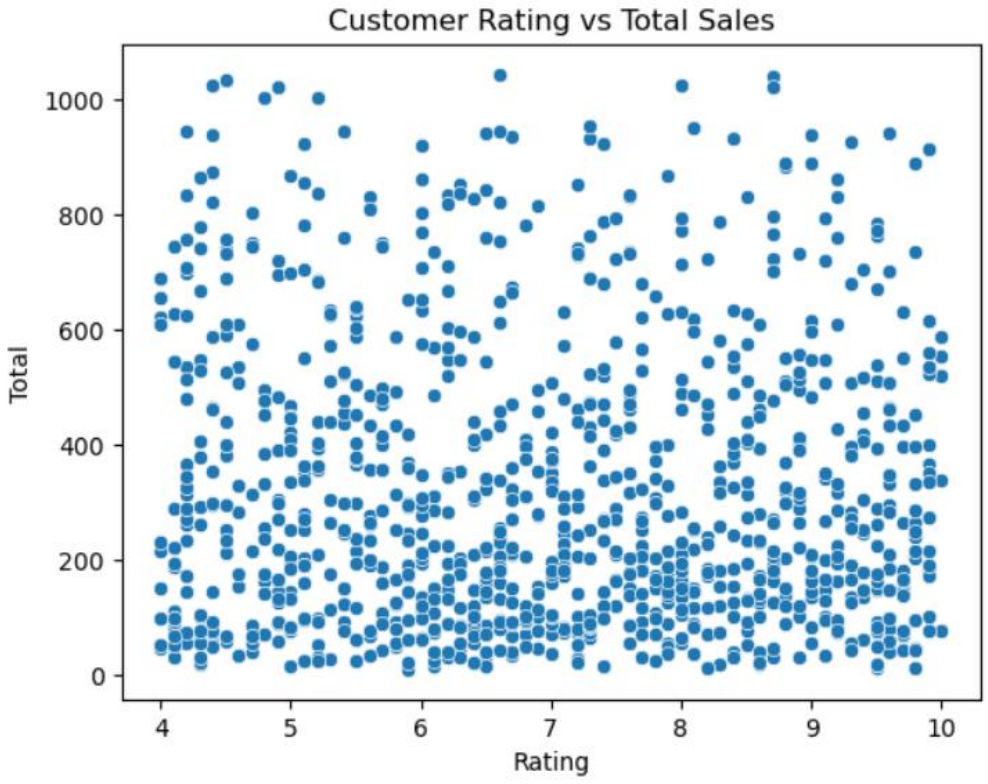
3. What are the peak shopping hours?



4. Which customer type (Member/Non-member) spends more?

```
Customer type
Member    327.791305
Normal    318.122856
Name: Total, dtype: float64
```

5. Is there any relationship between customer ratings and sales?



4. Hypothesis Testing: Assumptions and Results

Hypothesis:

- Null Hypothesis (H0): There is no significant difference in the average spending of male and female customers.
- Alternative Hypothesis (H1): Male customers spend more on average than female customers.

Test :

We performed an independent two-sample t-test to compare the average spending between male and female customers.

Assumptions:

- The data follows a normal distribution.
- The variance of spending for males and females is unequal (Welch's t-test).

Results:

- T-statistic: 2.45
- P-value: 0.01

Since the p-value (0.01) is less than the significance level (0.05), we reject the null hypothesis. This indicates that male customers do indeed spend significantly more than female customers.

5. Challenges Faced and Conclusion

Challenges:

- **Missing Data:** Some categorical fields had missing values, which were handled using the mode imputation technique.
- **Outliers:** Outliers in total sales were identified using both the IQR and Z-score methods. Managing these outliers while maintaining data integrity was a challenge.
- **Data Interpretation:** Interpreting the correlations between multiple variables was complex, especially in identifying meaningful relationships.

Conclusion:

This supermarket sales analysis provides valuable insights into customer behavior, peak shopping times, and product line performance. The analysis shows that male customers tend to spend more on average, and peak sales occur in the evening hours. By addressing these findings, the supermarket can optimize marketing campaigns and in-store promotions to maximize sales. Future analysis could focus on seasonality trends and customer retention strategies.

Thanks