

IST 707 Applied Machine Learning
By Prof. Kelvin King

Assignment 2

Supermarket Sales Data Analysis

Submitted by:
Samarth Sandesh Mengji
SUID: 718473878
NetID: smengji@syr.edu

Date of Submission: 9/17/2023
Syracuse University – School of Information Studies

Table of Contents

- I. Introduction
- II. Data Exploration and Transformation
- III. Insights into Sales Data
 - A. Gross Income Distribution Over Different Branches
 - B. Gender Differences in Each Branch
 - C. Customer Type Distribution in Each Branch
 - D. Product Line Distribution Across Branches
 - E. Payment Method Preference in Each Branch
- IV. Conclusion of Insights
- V. Part 2
- VI. References

I. Introduction

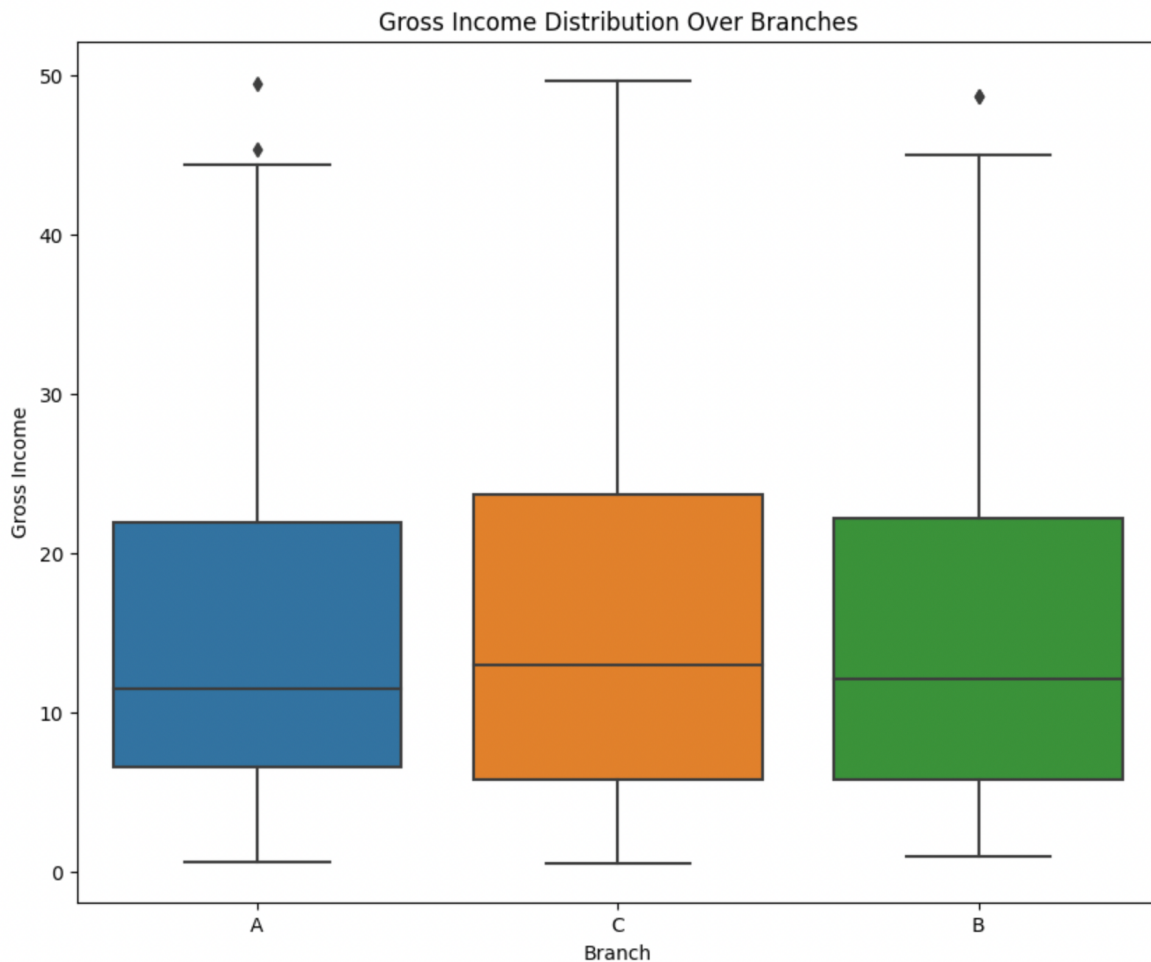
In this report, we analyze the sales data of a supermarket recorded in three different branches – A, B and C. We perform some data exploration and check for missing data before we start visualizing data based on specific branches. Our main goal of this assignment is to provide insights into sales data for each branch of the supermarket. We have performed different data exploration and transformation techniques on data to visualize and find meaningful insights on it.

II. Data Exploration and Transformation

Before we are going into analysis, we perform some data exploration and transformation. We checked for missing/null values and used functions like `head()` to better understand the attributes in the dataset. We found that data had no missing values and we can go ahead with visualizing the data using different attributes.

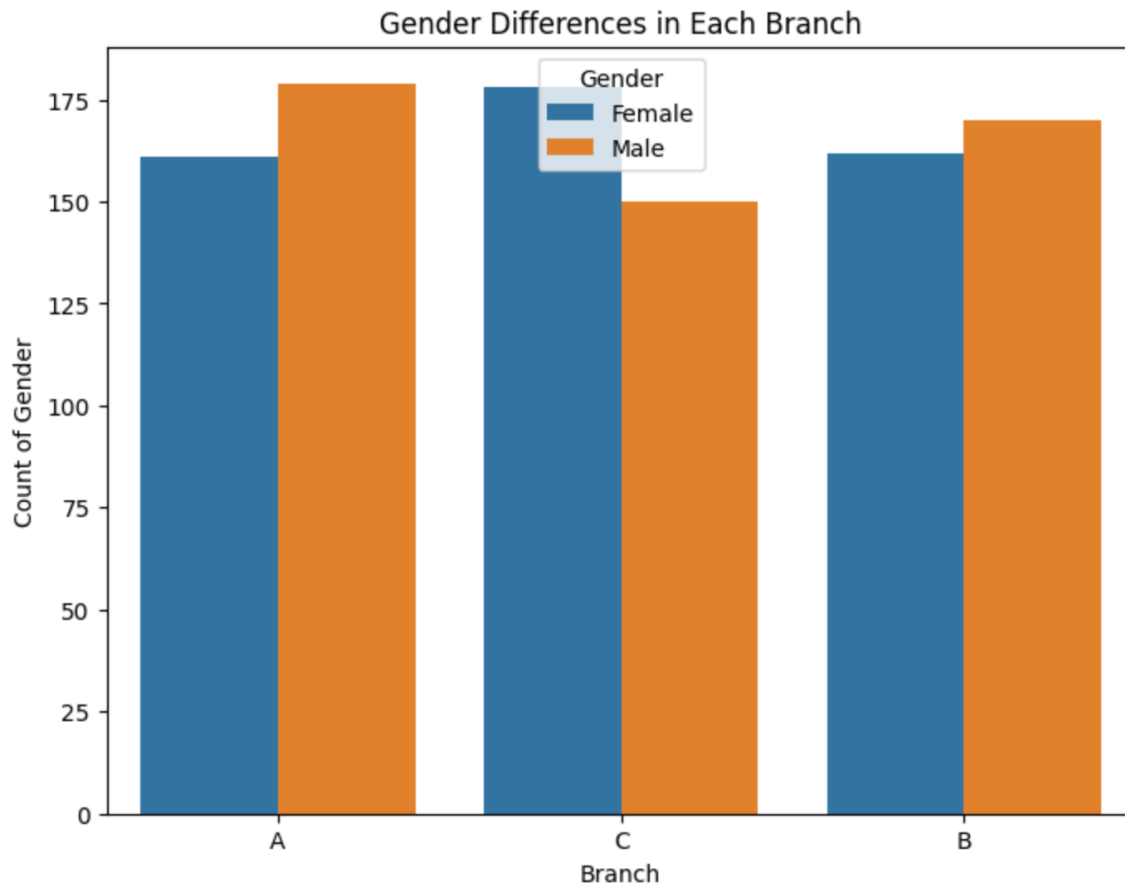
III. Insights into Sales Data

A. Gross Income Distribution over different branches



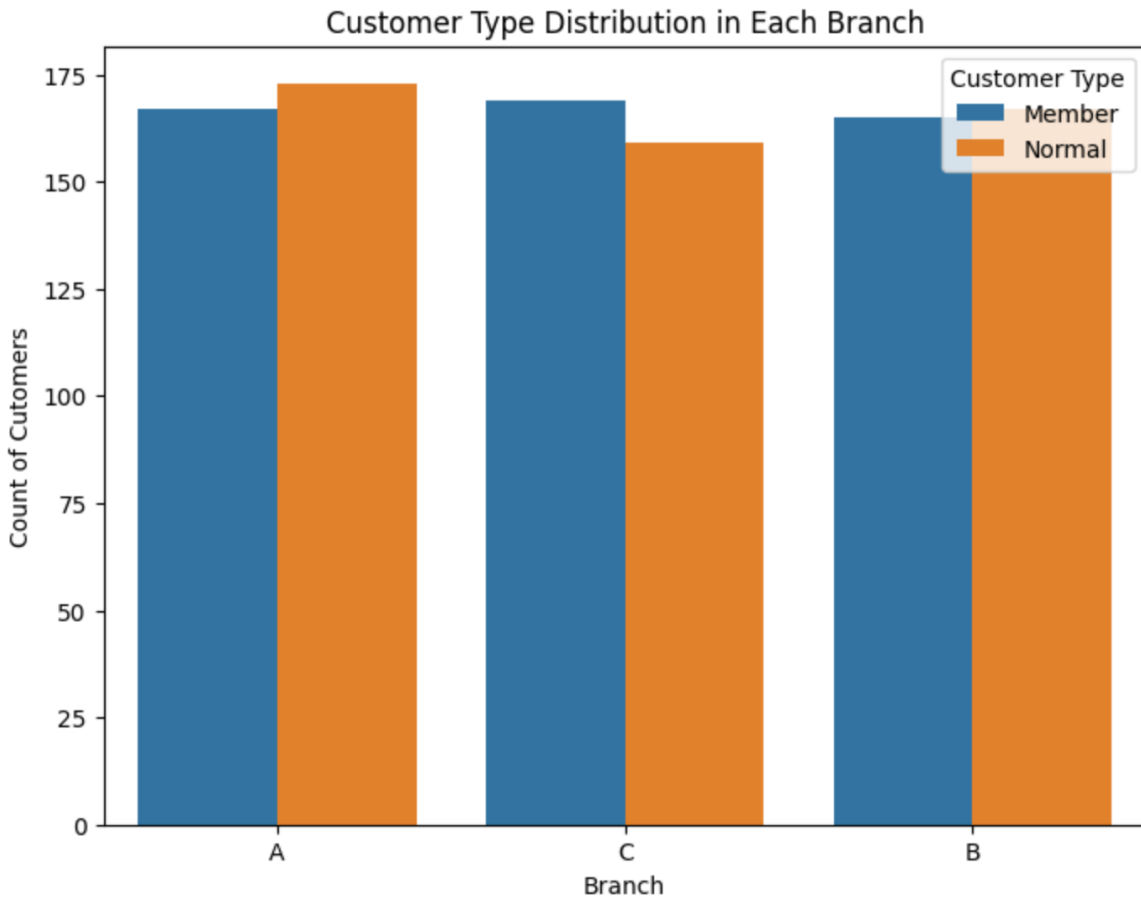
In the above box plot, we have tried to plot the gross income distribution over the three branches. From the above visualization, we can clearly say that branch C has the highest gross income which indicates that branch C is the most profitable branch followed by branch B and A.

B. Gender Differences in Each Branch



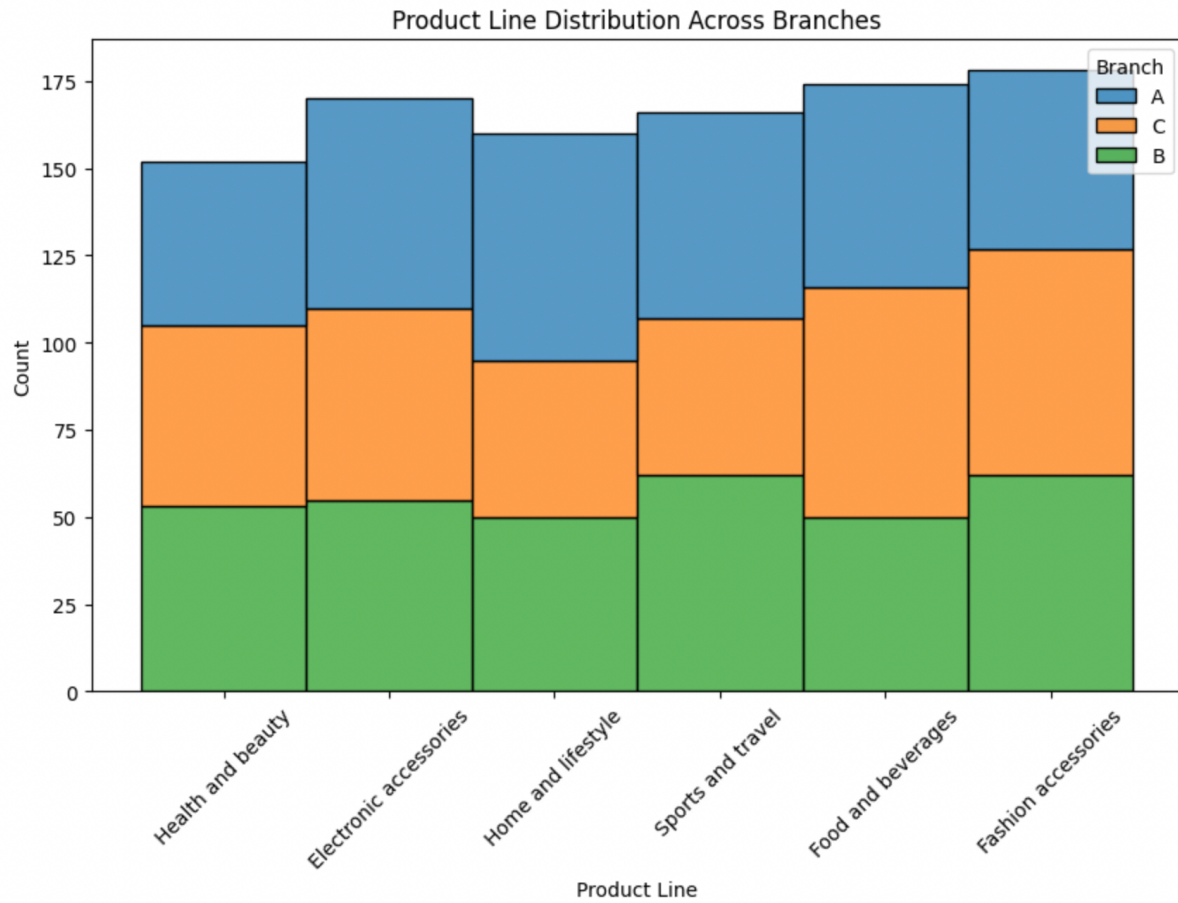
In the above Count plot, we have plotted the count of male and female customers in each branch. From the visualization, we can say that branch C has the highest number of Female customers followed by B and C. Branch C has a greater number of female customers than male while both branches A and B are more male dominant than female. In branch A and B, we see a slightly similar balanced distribution of male and female customers.

C. Customer Type Distribution in Each Branch



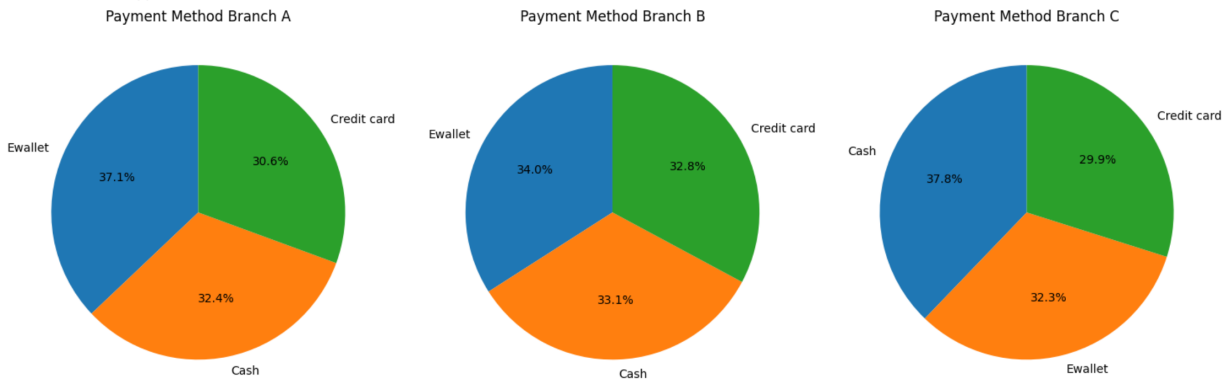
In the above graph, we plot the type of customer – Member or Normal in each branch using count plot. From the above graph, we can say that Branch A and B both have a balanced distribution of members and normal customers while we can see a difference in branch C. In branch C, there are more members than normal customers, also having the highest number of members in all three branches.

D. Product Line Distribution across branches



In the above multi-stacked histogram, we have tried to plot the sale of different product lines in each branch by representing it by three different colors. From the above graph, we can conclude that product line “Fashion accessories” and “Food and beverages” have the highest sales across all branches while “Health and Beauty” products have the lowest sales. There is less variation in product lines across branches as many have similar distribution.

E. Payment method preference in each branch



In the last pie chart, we plot the distribution of different payment methods across different branches. From the above pie chart, we can conclude that “E-wallet” and “Cash” are the most preferred payment method while “Credit card” is the least preferred payment method. Distribution seems balanced across all three branches.

IV. Conclusion from above graphs.

In conclusion, we can say that branch C is the most profitable branch due to its higher gross income. Branch C has most members, highest female customers with “Fashion Accessories” being the highest sales across product line. We can conclude that due to high female customers and member customers in Branch C, it has high fashion sales and in turn higher gross income. While E-wallet and Cash are the most preferred payment, product line had a balanced distribution across branches.

Based on these findings, the supermarket chain can work on targeted marketing and focus on targeted inventory management across branches.

V. Part 2

Main Idea (M):

Preamble A: Kelvin is depicted as a shy, helpful, detail-oriented with little interest in people or world of reality. He has a need for order and structure.

Preamble B: Lucy is depicted as a single, outspoken, and bright with a major in philosophy. She actively takes part in anti-nuclear demonstrations and is deeply concerned with social injustices in the society.

Evidence (E):

Preamble A: The characteristics defined for kelvin resemble more towards him being a librarian rather than a farmer. Kelvin has attention to detail and needs orientation which are qualities of a librarian.

Preamble B: Lucy's characteristics of being active in social justice issues and a degree towards philosophy makes her more suited for a role like bank teller with being active in feminist movement. Being bank teller requires good communication and analysis and involvement in social injustice issues provides evidence towards her being bank teller with active participation in feminist movement.

Analysis (A):

Preamble A: Kelvin's passion towards being detail oriented and being meek and tidy soul aligns with the characteristics of a librarian. A farmer needs characteristics like physical labor and agriculture background. These characteristics do not align with kelvin's characteristics.

Preamble B: Lucy has a degree in philosophy, and she actively takes parts in social injustice like anti-nuclear demonstrations which demonstrates her passion towards society and demonstrates her skills like communication and analysis.

Link (L):

Preamble A: From the above evidence, we can say that Kelvin is a librarian because of his detail oriented and organized characteristics.

Preamble B: From the above evidence, we can say that Lucy is a bank teller and is active in feminist movement due to her educational background and her active participation in social injustice issues.

I certify that this assignment represents my work. I have not used any unauthorized or unacknowledged assistance or sources in completing it, including free or commercial systems or services offered on the internet.

VI. Reference:

Website documentation used for script:

<https://seaborn.pydata.org/generated/seaborn.boxplot.html>

<https://seaborn.pydata.org/generated/seaborn.countplot.html>

<https://seaborn.pydata.org/generated/seaborn.histplot.html>

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.pie.html

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.subplots.html

https://pandas.pydata.org/docs/getting_started/intro_tutorials/03_subset_data.html