IST 707 Applied Machine Learning
By Prof. Kelvin King

Assignment 3

Visualization and Dimensionality Reduction

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

Date of Submission: 9/24/2023
Syracuse University – School of Information Studies

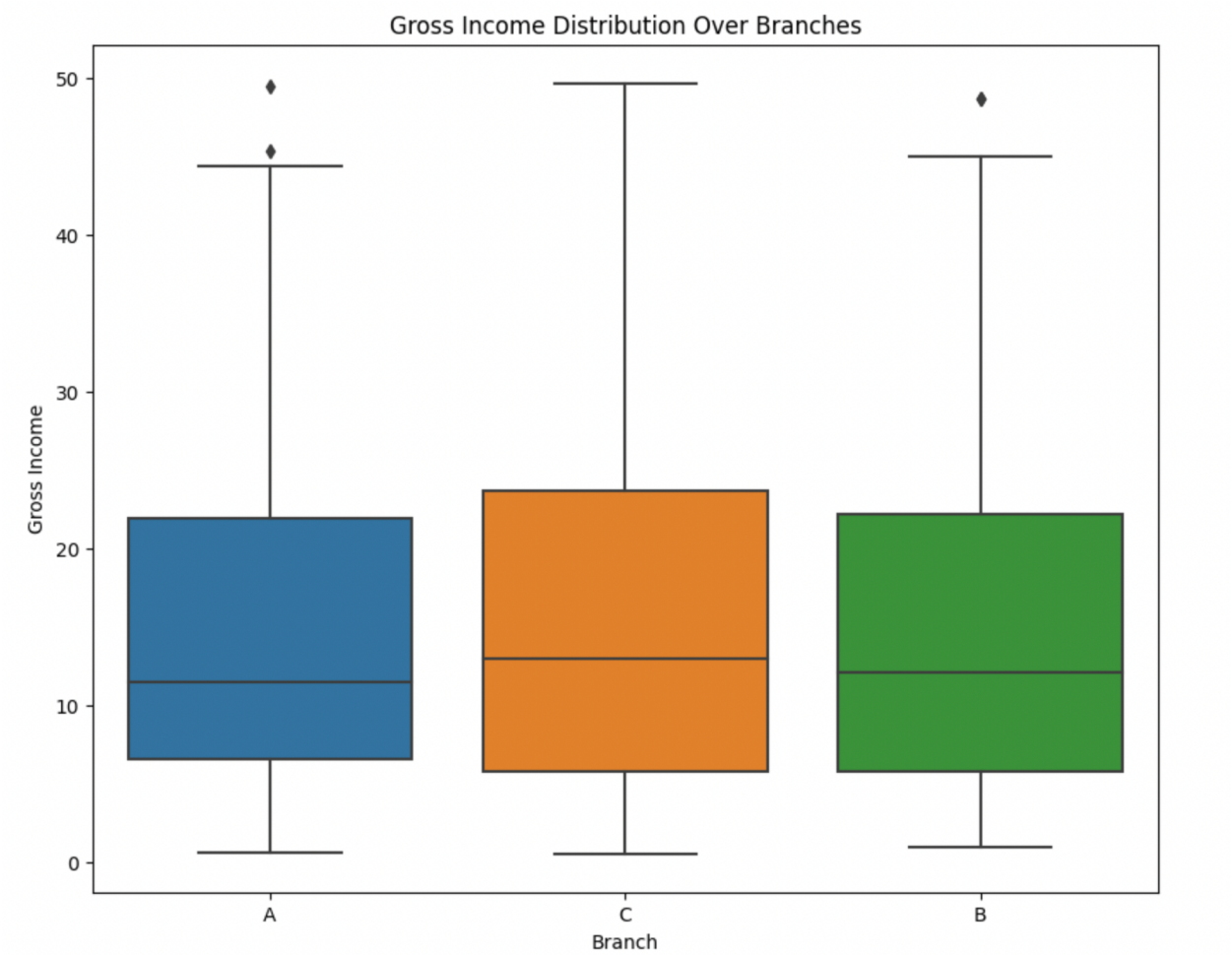
Table of Contents

1. Introduction – Part 1

In this report – part 1, we focus more on visualizations and dimensionality reduction. We explore more visualizations on the sales data of supermarket recorded in three different branches – A, B and C. Our main objective here is to give more insights on the sales data and learn more visualizations techniques like scatter plot, box plot and bar plot. We then perform Principal Component analysis on the MNIST dataset. We perform different data preprocessing techniques to transform the data in standard and normalized form for PCA. Our main goal is to understand the MNIST dataset using the PCA. In this report, we will try to provide the supermarket with recommendations based on the visualizations and analysis and extract valuable insights from the data for appropriate business decisions.

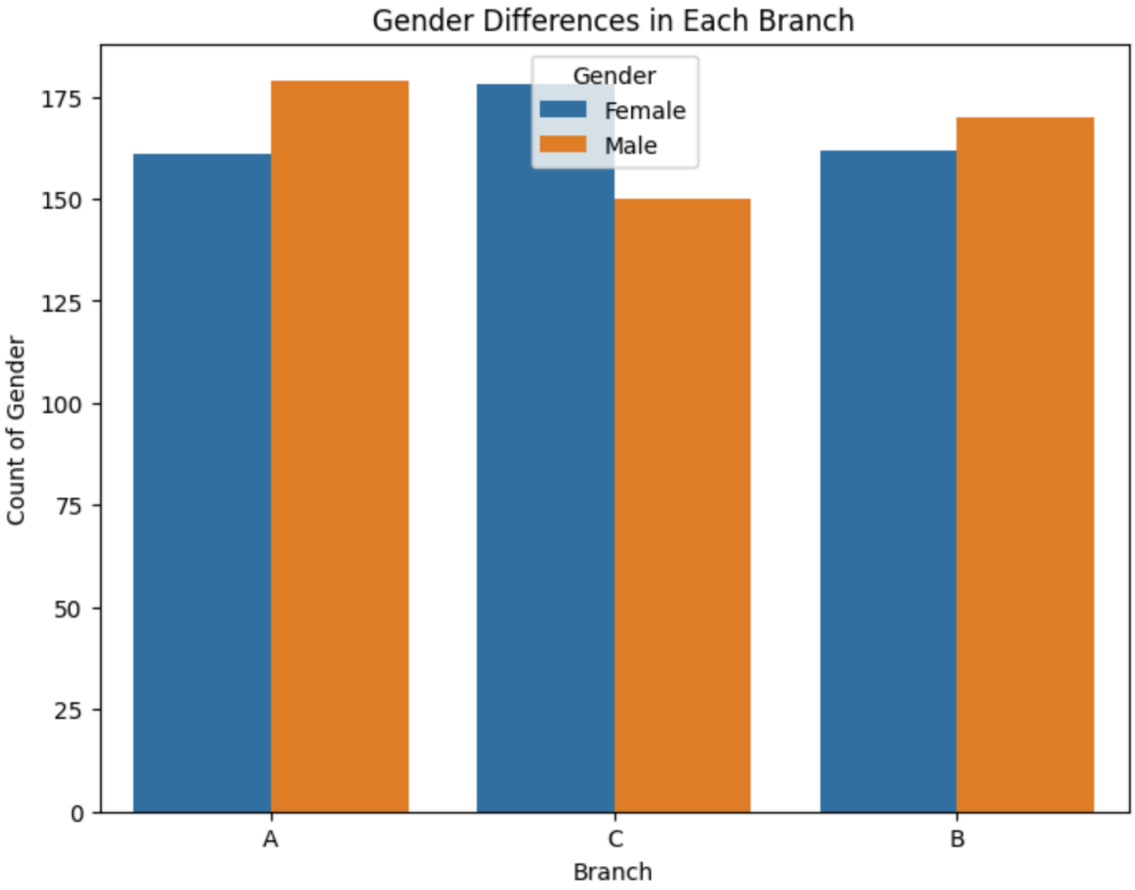
2. Visualizations of Sales Data

2.1 Gross Income Distribution over different branches



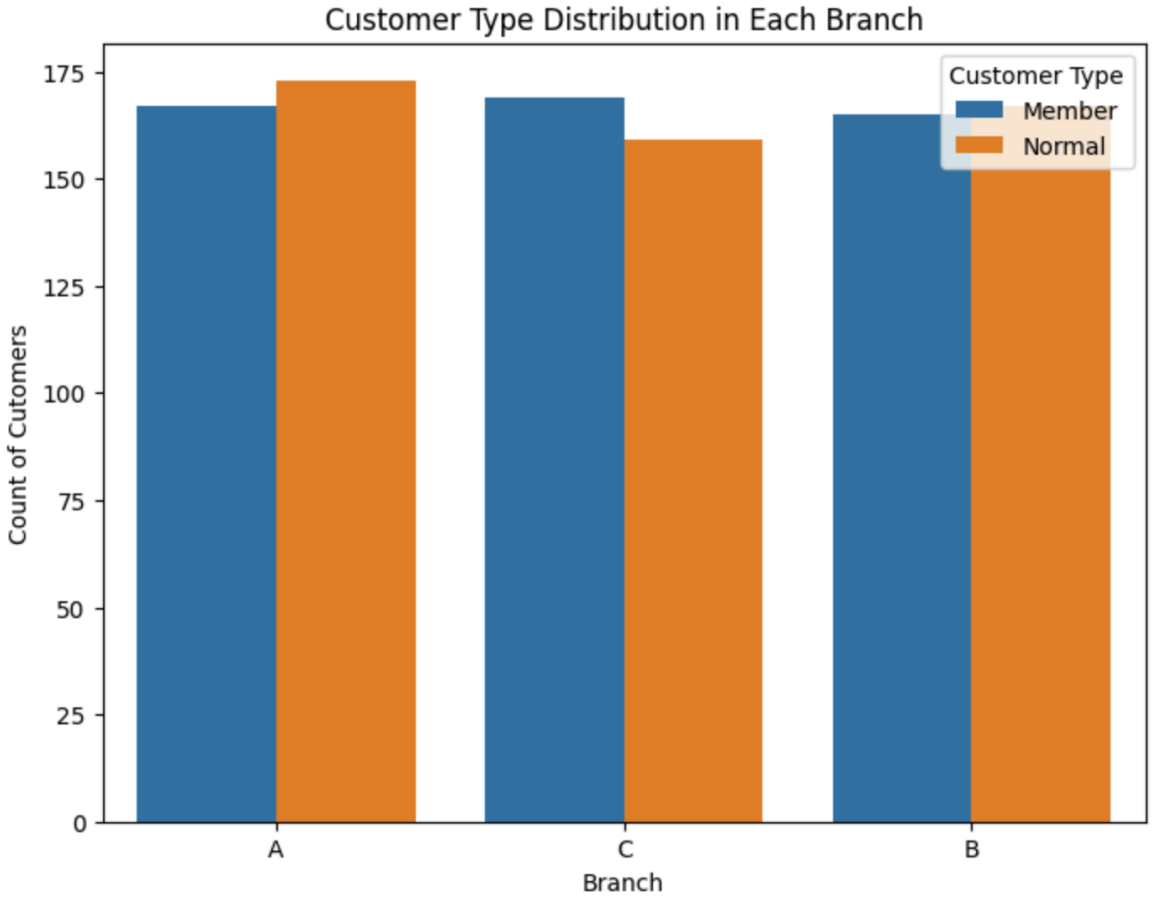
In the above box plot, we have tried to plot the gross income distribution over the three branches. From the above visualization, we can clearly say that branch C has the highest gross income which indicates that branch C is the most profitable branch followed by branch B and A.

2.1 Gender Differences in Each Branch



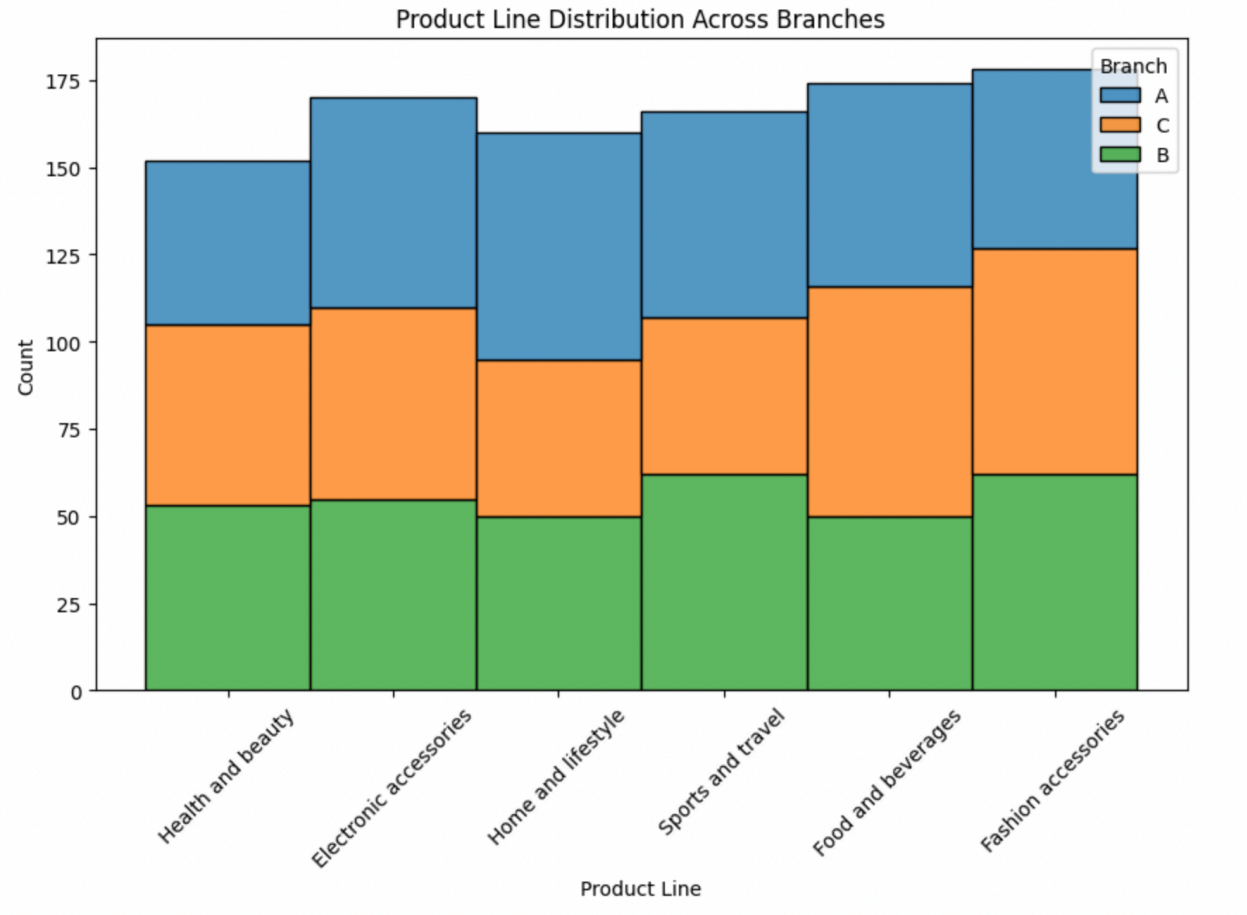
In the above Count plot, we have plotted the count of male and female customers in each branch. From the visualization, we can say that branch C has the highest number of Female customers followed by B and C. Branch C has a greater number of female customers than male while both branches A and B are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

2.2 Customer Type Distribution in Each Branch



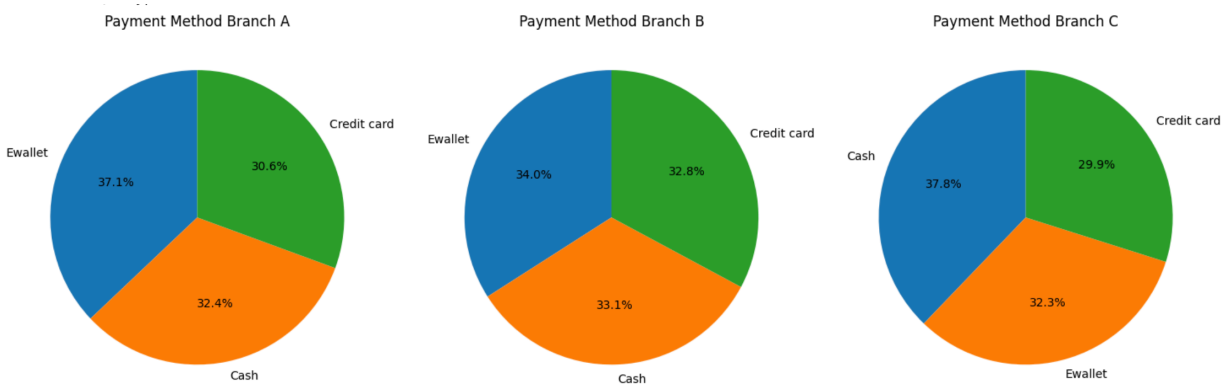
In the above graph, we plot the type of customer – Member or Normal in each branch using count plot. From the above graph, we can say that Branch A and B both have a balanced distribution of members and normal customers while we can see a difference in branch C. In branch C, there are more members than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

2.3 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.

2.4 Payment method preference in each branch

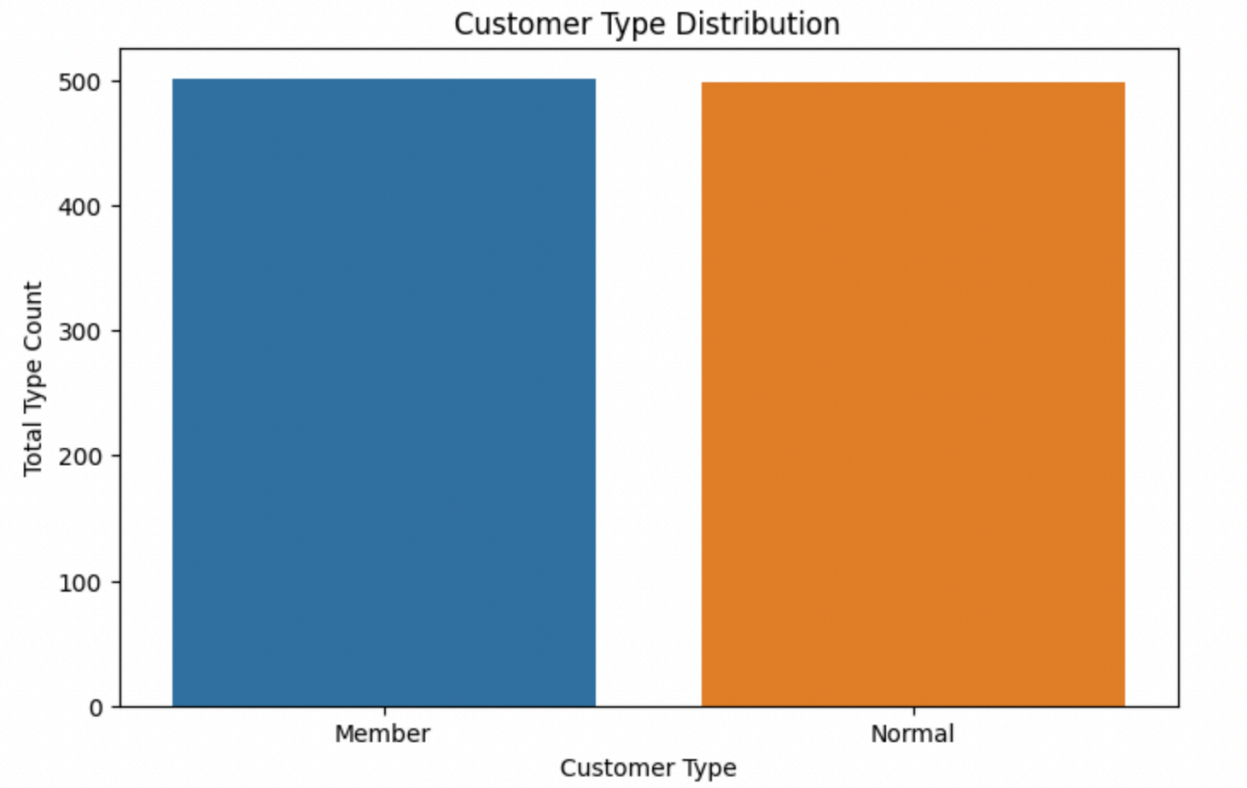


In the last pie chart, we plot the distribution of different payment methods across different branches. From the above pie chart, we can conclude that “E-wallet” and “Cash” are the most preferred payment method while “Credit card” is the least preferred payment method. Distribution seems balanced across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

2.5 Relationship between customer rating and total sales using scatter plot

In the above plot, we try to explore the customer ratings variable against total sales to see if ratings affect total sales. From the above scatter plot, we can see that data points are spread across the plot. There are wide range of data points across total sales and customer rating with no clear relationship in these two. There are high sales outliers for all range of ratings which might not affect total sales significantly. We can conclude that customer ratings alone is not sufficient to predict total sales.

2.6 Customer type distribution using bar plot



In the above bar plot, we try to plot two different customer type – member and normal against total count of customers in each type. From the above graph, we can see that member count is slightly more than normal customer type. We can see a balanced distribution overall indicating that the supermarket has high number of normal customers, and the plot gives a clear breakdown of both types. This can help the supermarket marketing team to work on member promotions cater to strategic marketing to targeted normal customers. The team can use this plot to monitor the total number of customers and members.

Conclusion and Recommendation:

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.

3. PCA on MNIST dataset

In this part of assignment, our objective is to perform PCA on MNIST dataset which is a collections of handwritten digit images. Main goal is to apply dimensionality reduction on the dataset trying to preserve maximum variance using PCA.

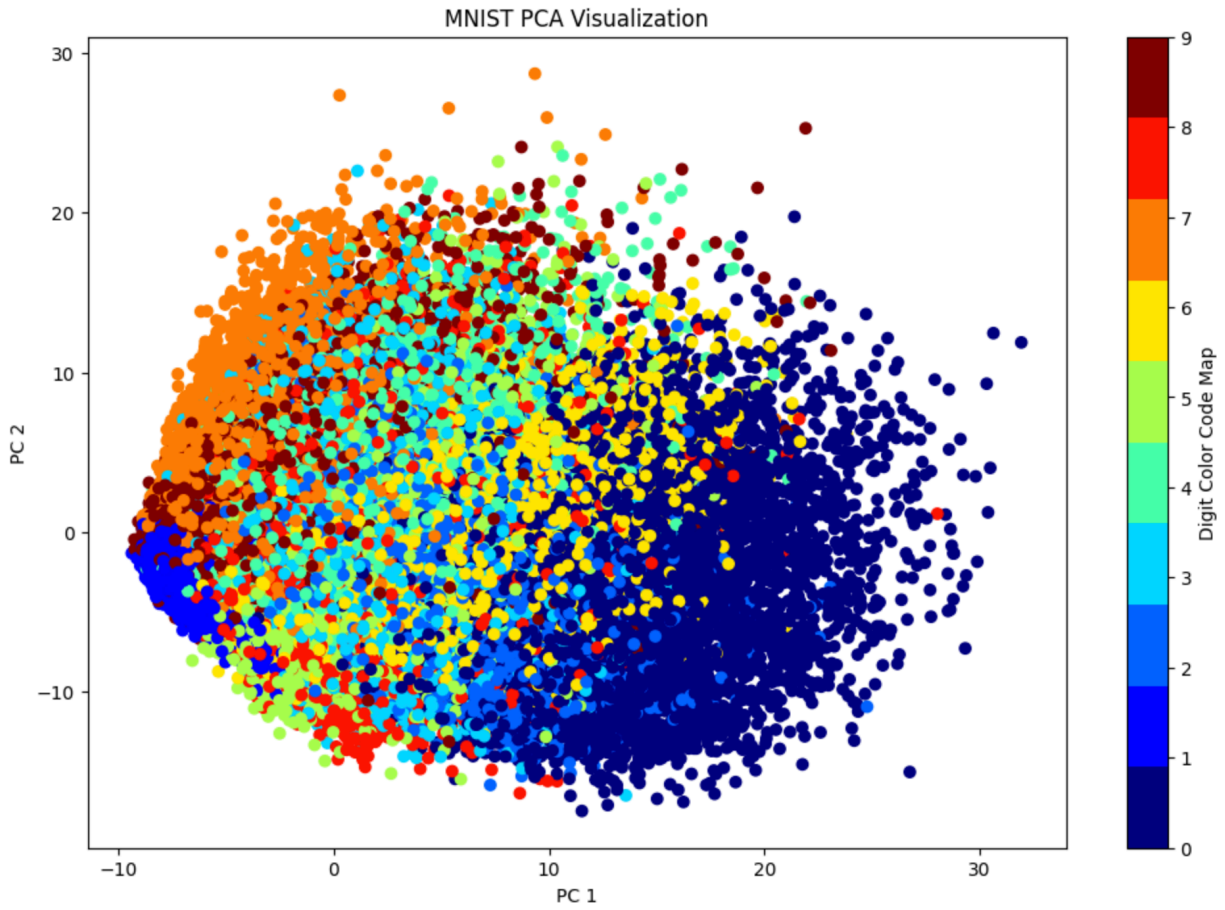
Dataset: The MNIST dataset contains many images of 28x28 pixel grayscale images of handwritten digits from 0 – 9. Dataset is loaded from a function called fetch_openml in scikit-learn datasets.

Preprocessing: Standardization was applied so that every attribute will have the same scale. We use a formula where subtract the mean and divide by standard deviation to each variable. This ensures that each variable contributes equally to the analysis.

We then used Simple imputer to impute the missing values with mean. This strategy will fill any missing values with the mean.

PCA: We then applied PCA on the normalized dataset to project the dataset to 2 dimensional space as we need to visualized the data in 2D space.

Visualization: We then visualize the data using scatter plot with two PCA's - A_pca[:, 0], A_pca[:, 1]. Color of each data point is identified by target variable and then we map it using color map 'jet' with 10 different colors.



In this 2D scatter plot, we can observe that there are clusters which represents a single digit. The digits are color coded on the map to the right of plot. Some clusters can be easily identified like 0 and 7 as they are more distinct visually while others like 3,4,5,6,8 are spread across and overlapping each other which means they have digits with similar visual representation. We can see lot of gaps in data points in clusters from 3-8 which also indicates that the digits in these clusters have common handwritten characteristics. The plot has few outliers which indicate that these are digits with poor handwriting which do not come in any cluster.

Conclusion and Recommendation:

In this analysis, we applied PCA on the MNIST dataset to reduce our data to 2D space. While the two principal components PC1 and PC2 covered very less variance of the total, we were able to make a few conclusions based on the scatter plot we did after.

The scatter plot showed that the data consisted of clusters of similar digits, overlapping of digits, outliers in the dataset and the spread across the dataset. This suggested that the handwriting in the dataset consists of variability, and this can be one starting point for further analysis.

Digit recognition and classification are one of the many machine learning techniques which can be used for further analysis of data.

4. Introduction – Part 2

In this part, we analyze a bank dataset which contains demographic and personal information of its customers like age, income, marital status. Our goal of this assignment is to determine whether the customers of the bank would be interest in a Personal Equity Plan (PEP) based on different attributes present in the dataset. We first preprocess the data and perform exploratory data analysis with the help of different visualization techniques like scatter plot, box plot, bar plot like part 1 which will help us understand the data better. PCA is then applied to this dataset to reduce the dimensionality of it and plot it in 2D space using scatter plot. In this report, we will try to provide the bank with recommendations based on the visualizations and analysis and extract valuable insights from the data for appropriate business decisions.

5. Data Preprocessing and Transformation

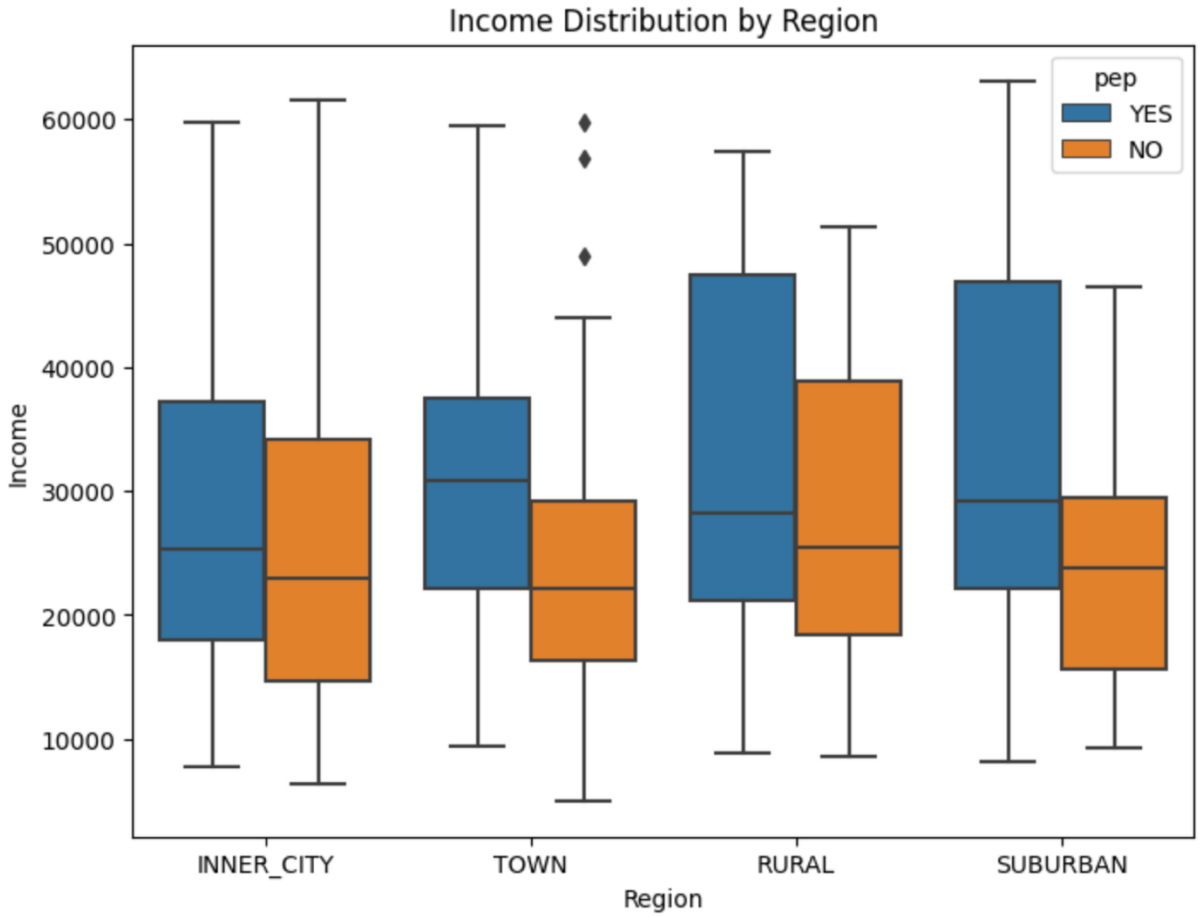
To ensure data quality and precision in our analysis, we first preprocess the data using different methods. We first check for any missing values, encode categorical data and standardize numerical features for consistency in data analysis.

6. Visualization of Bank Dataset

6.1 Scatter Plot - Age vs Income by PEP

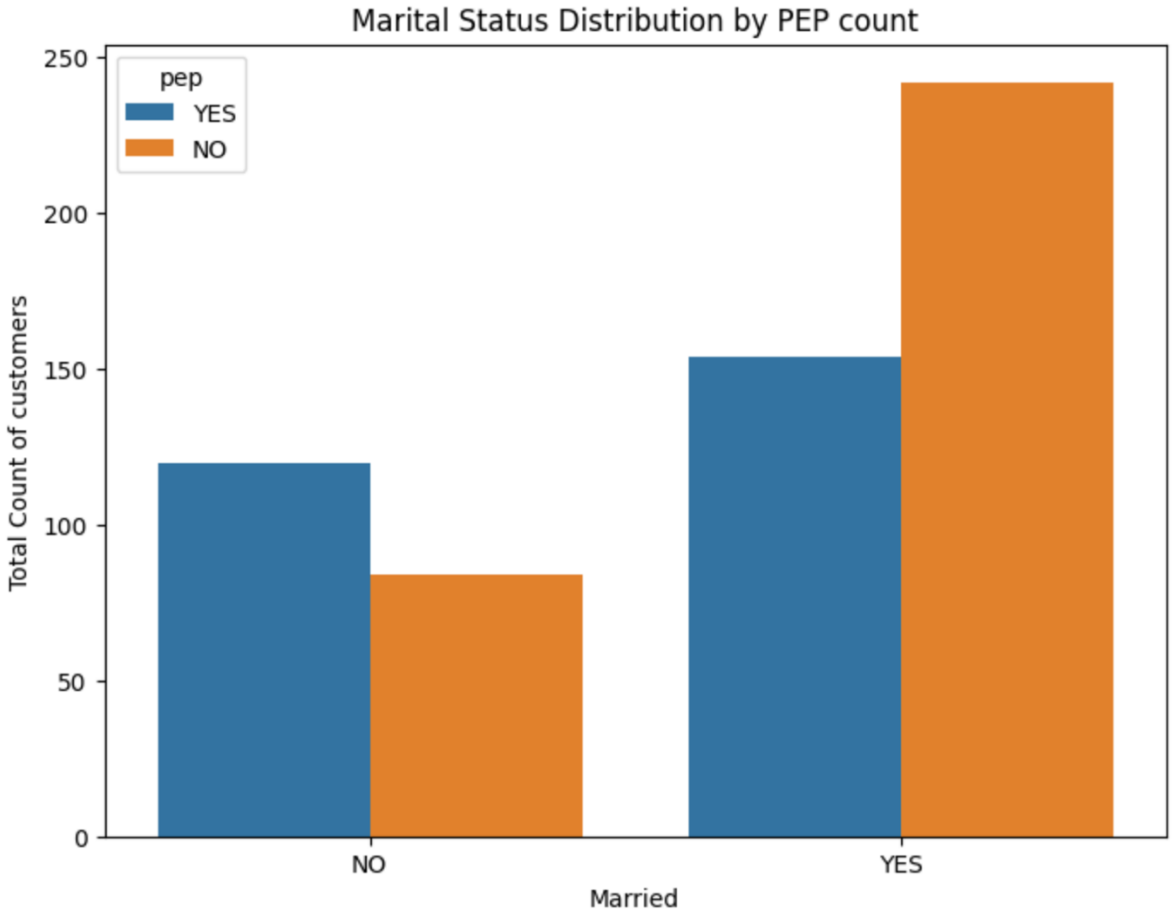
In the above scatter plot, we plot Age distribution of customers against their income with PEP to see whether we see a pattern of customers purchasing PEP based on age. From the above, graph we can say that the income of people increases consistently as their age increases. This says that income of a person is directly proportional to their age. While we see a pattern in income and age, it is hard to identify a pattern in PEP purchase. We can see orange and blue data points spread across all age groups with all income ranges. Orange indicates that a customer has not purchased a PEP while blue indicates vice versa. We can conclude that, alone age and income is not sufficient to determine whether a customer would purchase a PEP.

6.2 Box Plot - Region vs Income by PEP



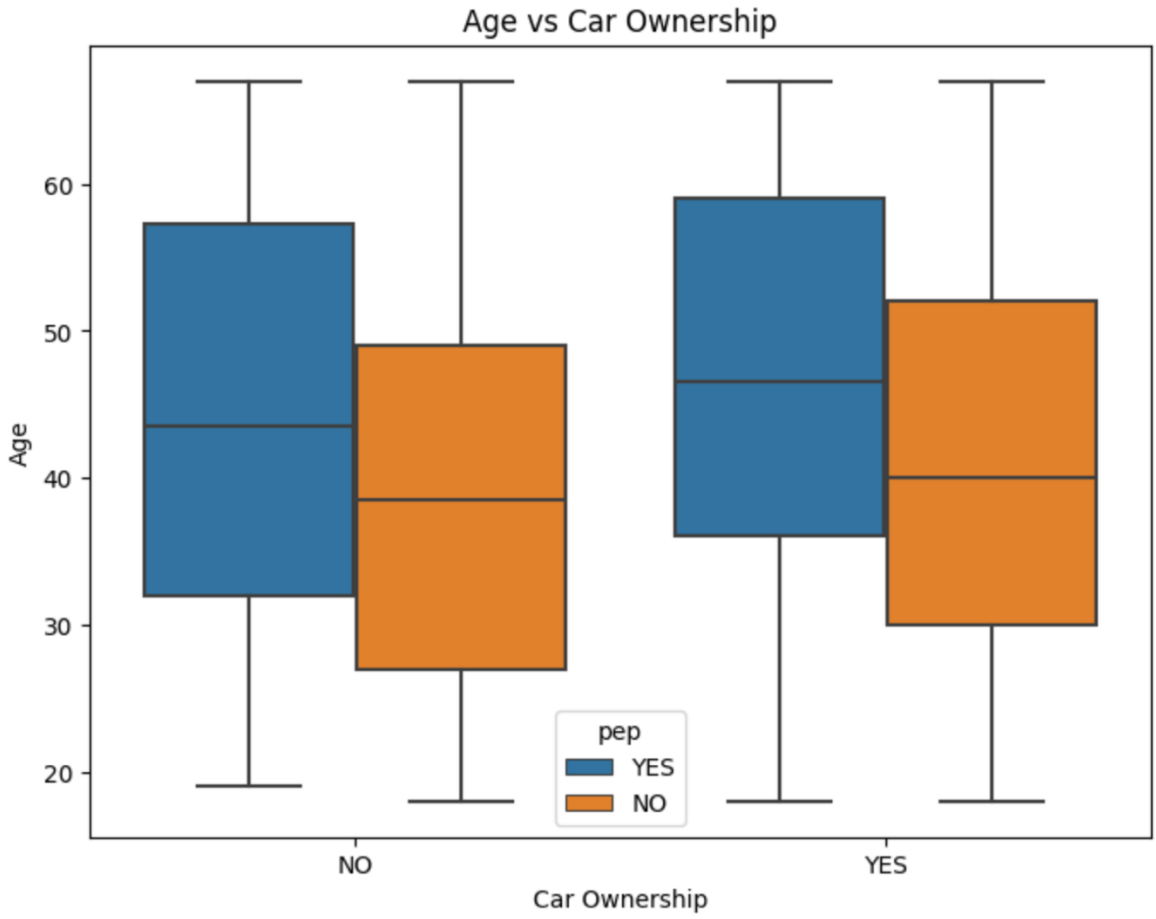
In this above box plot, we plot different regions across income and PEP purchase. From the above visualization, we can clearly see that people living in Suburban and Rural areas have higher income as compared to other regions and in turn have higher number of PEP purchase than other regions.

6.3 Bar Plot - Marital status and PEP



In the above bar plot, we plot Marital status distribution vs total count of customers and PEP. From the above graph, we can conclude that the dataset contains more people who are married. There is a balanced distribution in terms of PEP purchase in married and unmarried people with married being slightly higher than unmarried. There are more number of people who do not have a PEP in married segment while people who are unmarried have more PEP purchase as compared to ones who do not have in unmarried segment.

6.4 Box Plot - Age distribution by car ownership and PEP



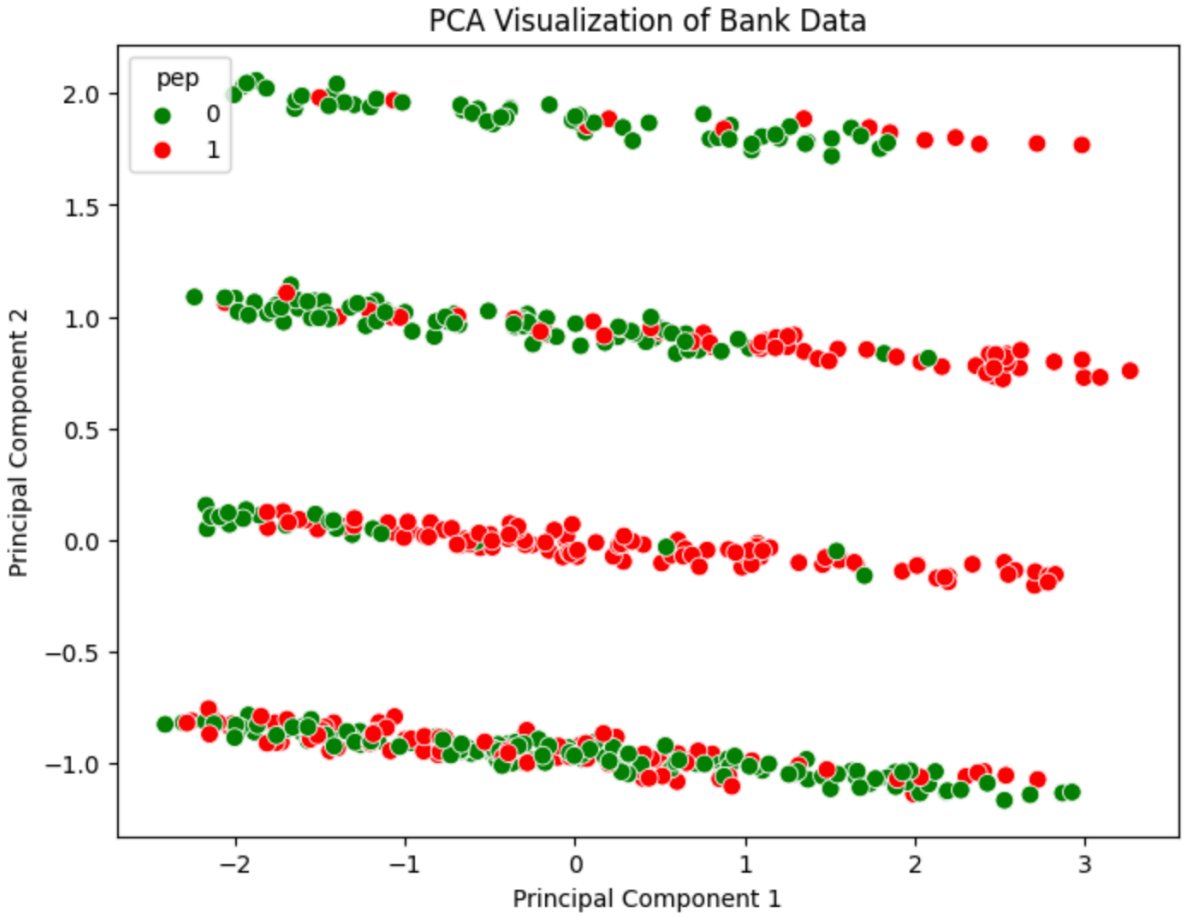
In this plot we try to plot a box plot to identify whether having a car affects PEP purchase or not. We plot Car ownership against Age. We can see that customers who are slightly old usually own a car but having a car ownership doesn't significantly increase the chances of having PEP.

6.5 Bar Plot - Customers with savings account and PEP

In this bar plot, we plot distribution of savings account and its relationship with PEP purchase. From the above graph, we can see that people with savings account have lot more PEP purchase than people without a savings account. We see a balanced distribution within both segments of savings account.

7. PCA on Bank Dataset

PCA is a technique used to reduce the dimension of data that identifies the most significant components in the data. In our analysis, we applied PCA to all components and then recognized the two most significant components by visualizing it in a 2D space.



The above visualizations indicate different clusters based on similar features. These clusters might have customers with similar financial situation or age or marital status – behavior and preferences.

Conclusion:
From the above visualizations, we can conclude that there is single most significant variable leading to PEP purchase. There are multiple factors at play and PEP purchase depends on multiple features, behavior, and preferences of customer.

8. Recommendations to Client

Recommendation 1: Based on above visualizations of numerical data features like age, income, and children, we can say that customers in a particular age group with equivalent average income and children are more likely to be inclined to by PEP. Client can do target marketing in these groups to increase PEP sales.

Recommendation 2: In above visualization with categorical features like marital status, car, savings account, we can understand that people who are married, own a car, and have a savings account are more likely to have a PEP. Again, client can conduct marketing campaigns focusing on these groups.

Recommendation 3: People with high income and a particular age group are more likely to buy a PEP. Client can focus on these groups as well.

9. References

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://conx.readthedocs.io/en/latest/MNIST.html
https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_openml.html
https://builtin.com/data-science/step-step-explanation-principal-component-analysis#:~:text=Step%201%3A%20Standardization&text=So%2C%20transforming%20the%20data%20to,transformed%20to%20the%20same%20scale.
https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.colorbar.html
https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
https://stackoverflow.com/questions/32857029/python-scikit-learn-pca-explained-variance-ratio-cutoff