Mini-Project III

By: Vamsi, Sunny

Demographics

Hypothesis

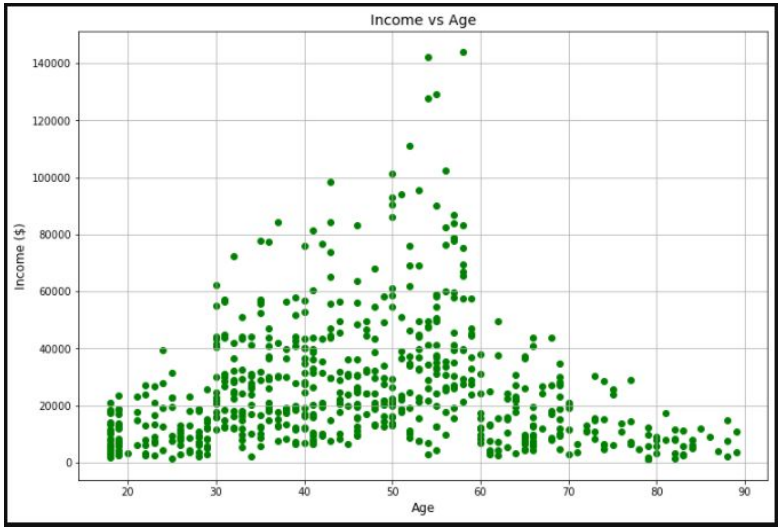
- Could we possibly target the wrong demographic using this data (bias)?
- Which age/gender group should we target with our credit cards so that we can avoid excess liability and ensure that credit card bills will be paid?

Data Preparation

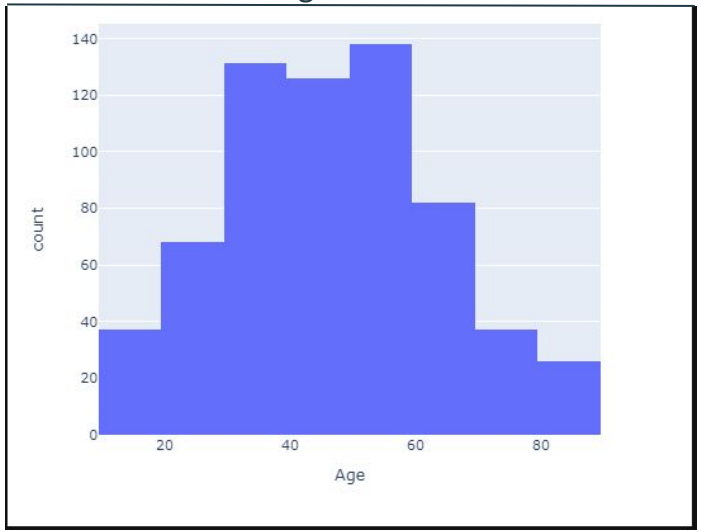
1. Remove unnecessary features and samples with $0 income as we wouldn’t want to target this demographic with bank accounts or products.
2. Transform the gender feature (One-Hot Encode).
3. Normalize the data by using StandardScaler.
4. Attempt to remove features by using Forward Selection.

Visualizations

Scatter (income vs age)

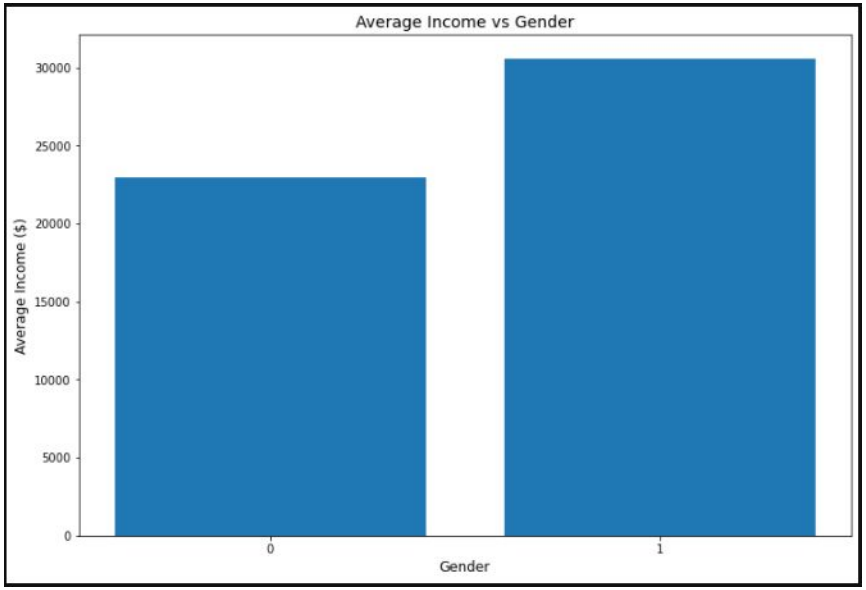


Age Distribution

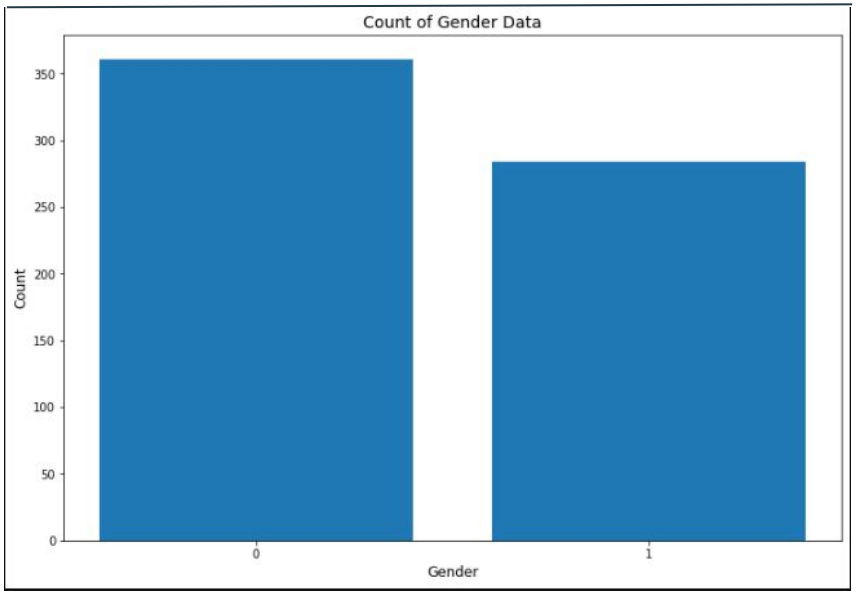


Visualizations

Gender Wage Gap ($8,000 difference in means)



Potential Bias (12% more females than males)



Unsupervised Learning Methods

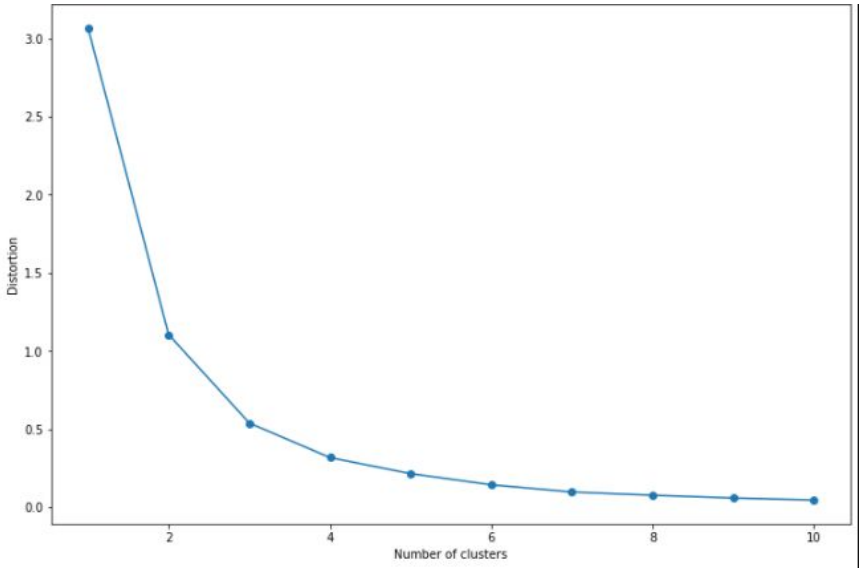
Clustering

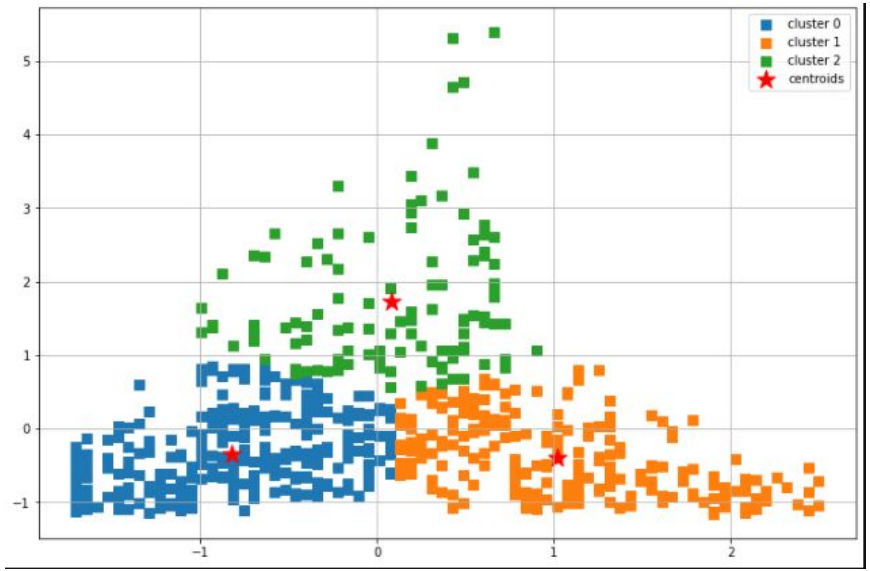
- Plot distortion and use the elbow rule to determine the appropriate number of clusters
- Apply K-Means clustering

PCA

- Clustering after PCA was applied

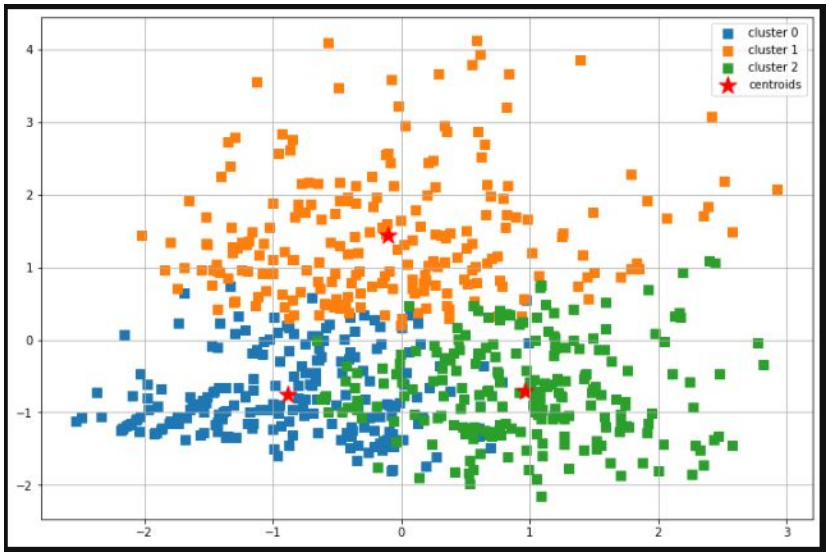
Clustering


Distortion
Number of clusters
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
8
10


cluster 0
cluster 1
cluster 2
centroids
5
4
3
2
1
0
−1
−1
0
1
2

Clustering Post-PCA

Confirms initial clusters



Patterns in banking behavior

Hypothesis

- Young people tend to be reckless in spending
- High income earners spend more
- People who spend more in credit card will have less savings

Clusters:

- More information on the users who are active credit card users?
- List of customers who are risky and safe in terms of their spending behavior

All the above questions can be answered with credit card and savings transactions and customer data. Each customer in the credit card transactions has been grouped together and summarized by average spending/saving per month

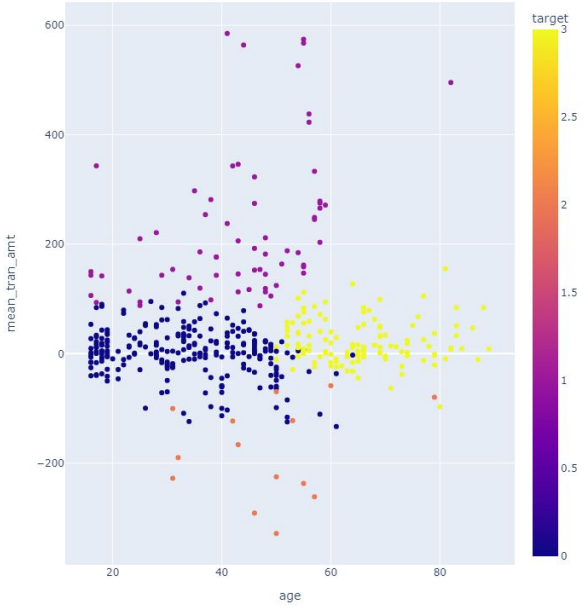
Data Preparation

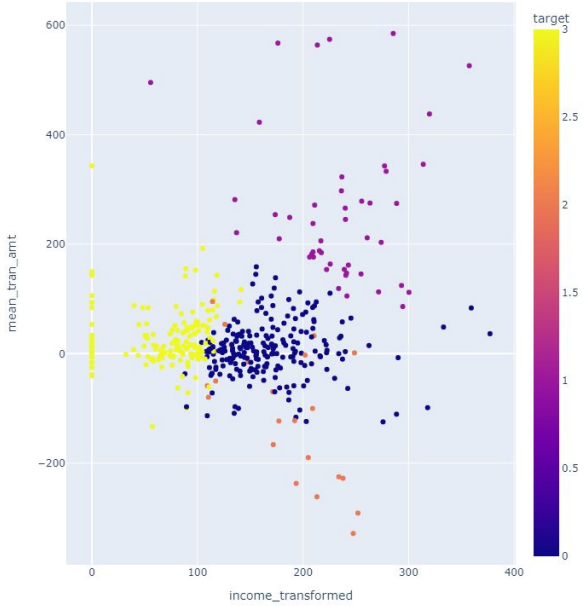
1. Dealing with missing values and outliers
    a. Dealing with nan - Filled end_date for accounts with today's date for active customers
    b. Dealing with outliers - Removed 10 customers (More than 3 standard deviations away)
2. Feature Selection
    a. Appended customer age, gender, income details to account/transactions table
    b. Removed transaction_code and top_channel data (feature description unknown)
3. Value Transformation
    a. One hot-encoded the gender
4. Before modelling
    a. Ensured normality of the features through q-q plot
    b. Scaled the features

Modelling

1. PCA
    a. Converted the features into PCA components
    b. Selected appropriate number of components through **Scree plot** that captures high variability in data
2. Cluster
    a. Selected number of clusters by plotting distortion levels per cluster (**elbow rule**)
    b. Clustered the data points using **KMeans** method

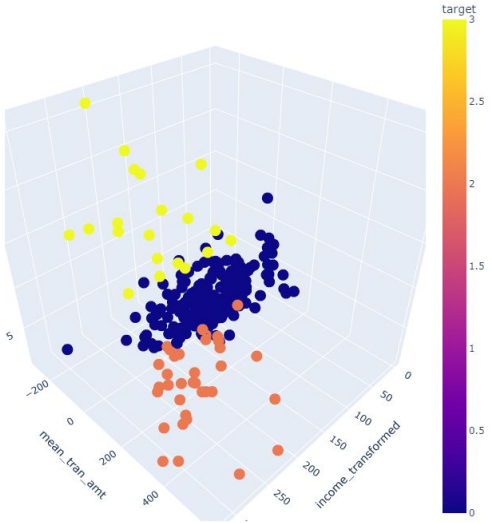
Hypothesis Testing

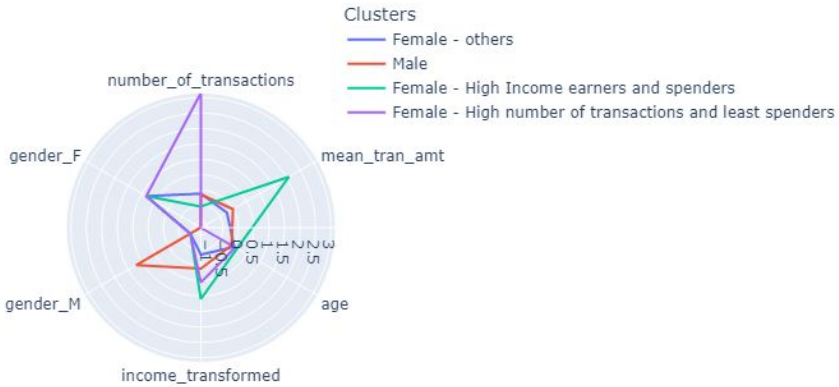




446 customers from credit card transactions

Demographic vs credit card behavior





The algorithm clustered Female into 3 categories and all Male into 1 category

Savings vs Spending