CLUSTERING FOR RETAIL CUSTOMER SEGMENTATION

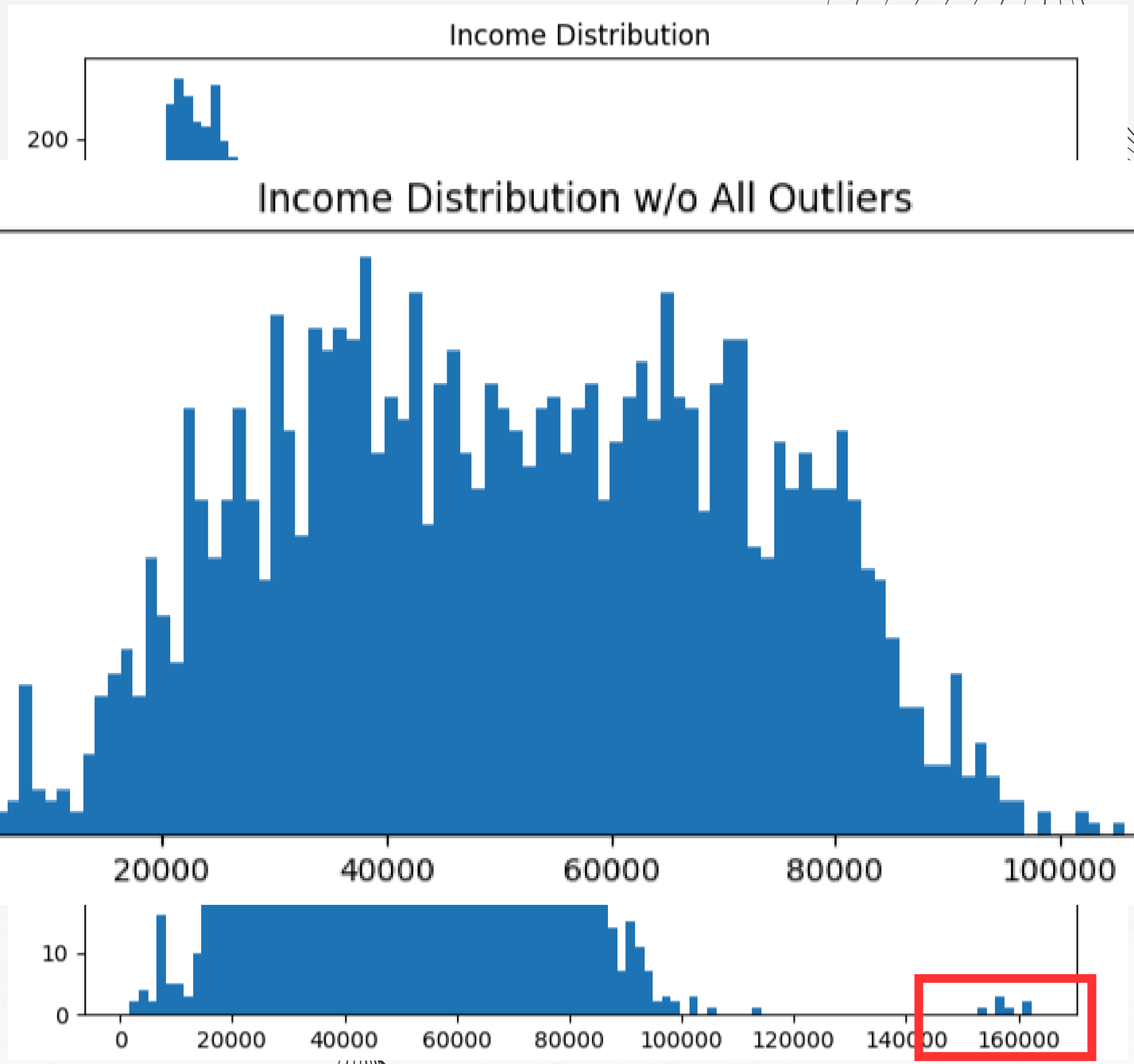
INTRODUCTION

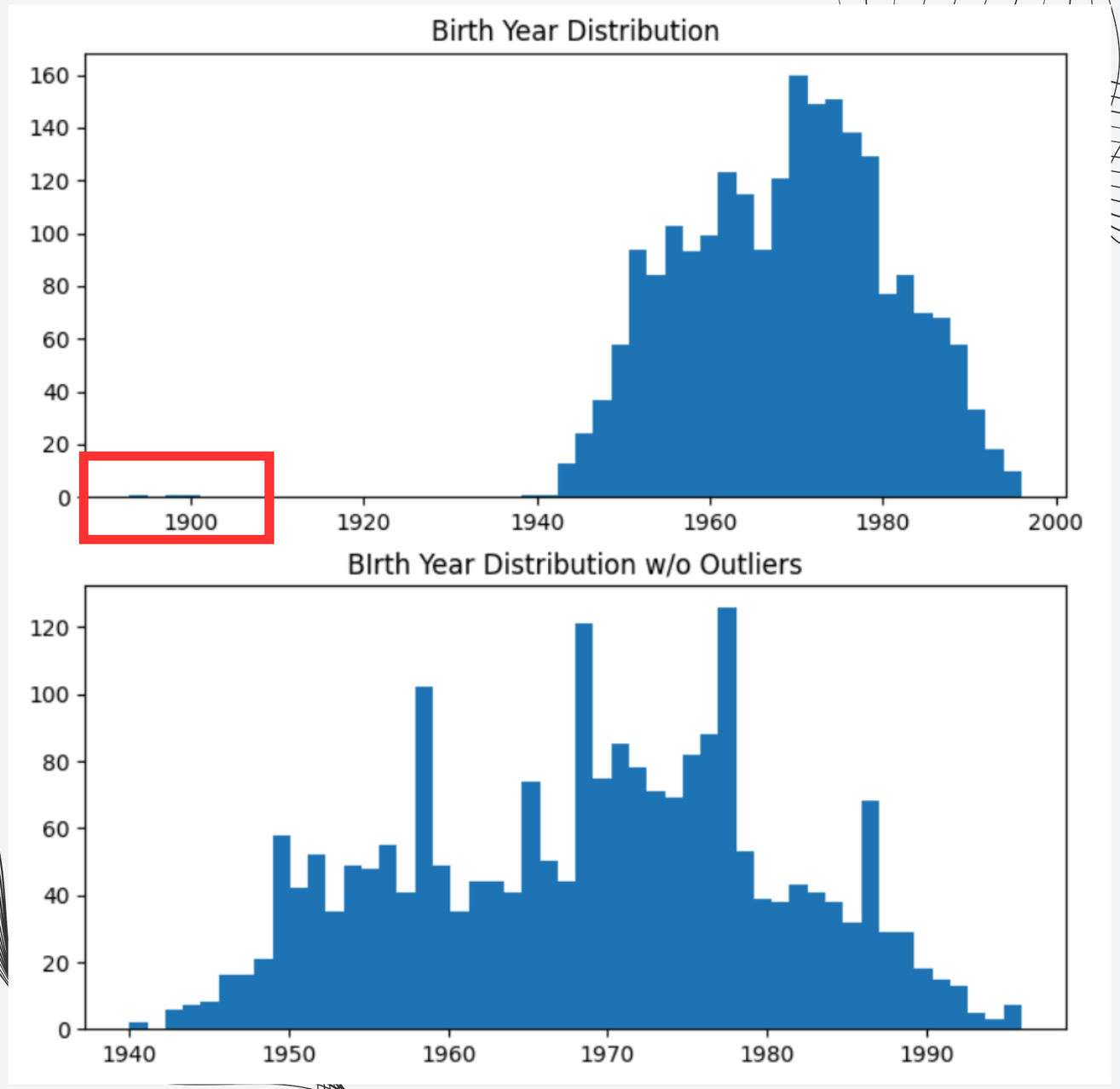
- **Problem:** Retail stores need insights into how their customers behave in order to provide relevant marketing, sales, and experiences

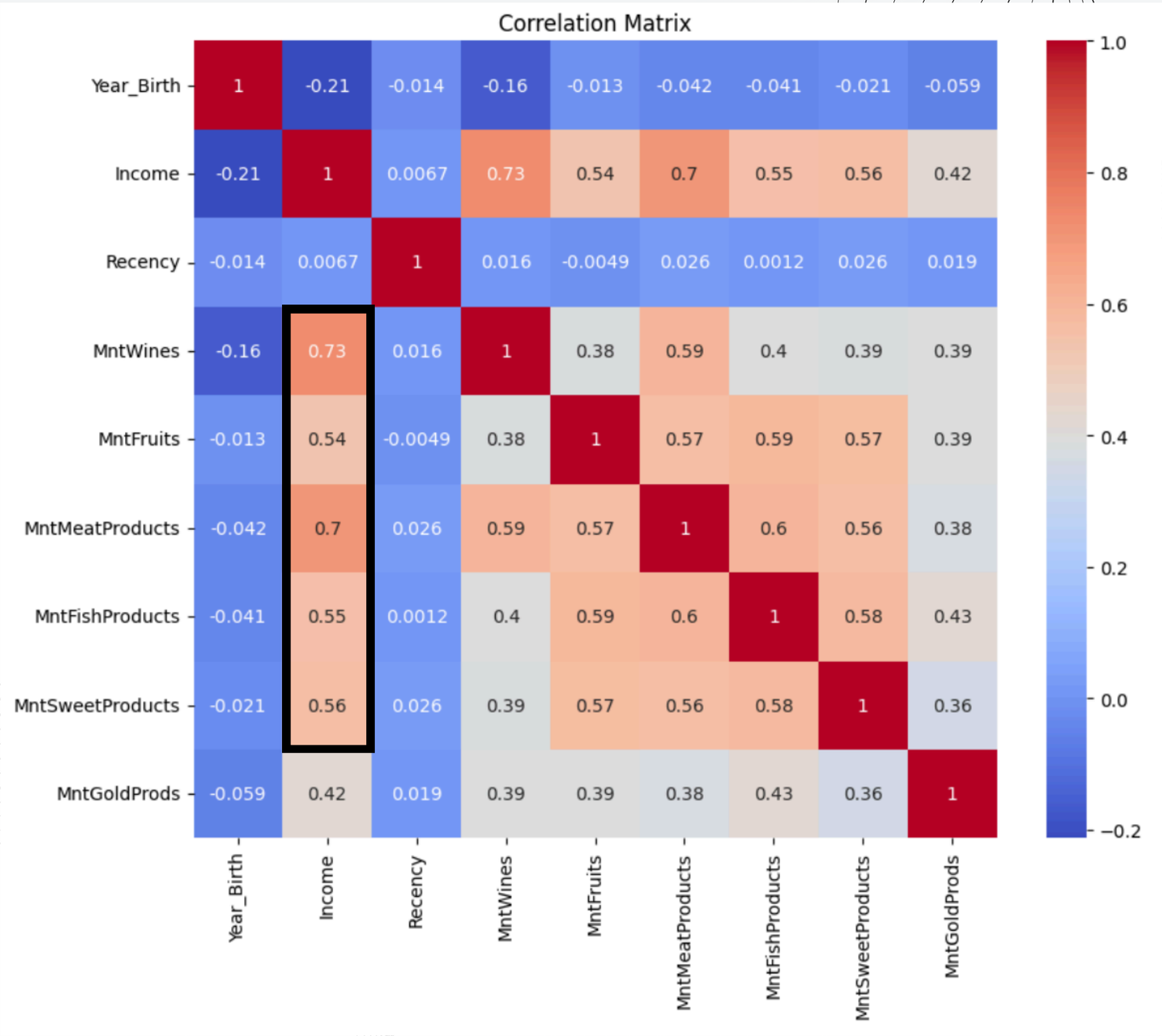
- **Project goal:** Determine how customers can be split into discrete clusters based on their demographics and spending habits

DATA CLEANING/EDA

- **Dataset:** Kaggle dataset with 2,240 observations
- **Key variables:** Age, income, recency, amount spent
- **Data cleaning:** drop irrelevant columns, reformat dates, remove outliers from income and age


Income Distribution
200
Income Distribution w/o All Outliers
50
40
30
20
10
0
0
20000
40000
60000
80000
100000
10
0
0
20000
40000
60000
80000
100000
120000
160000

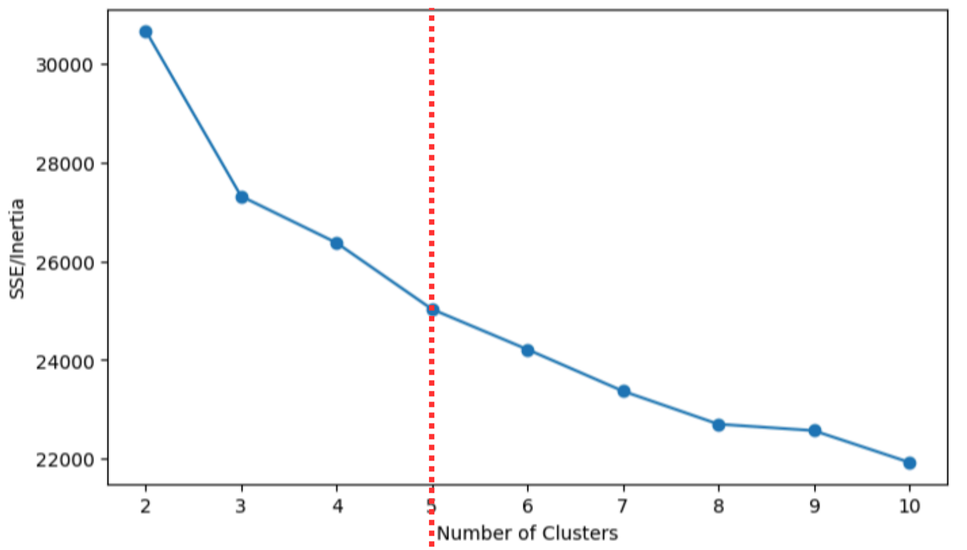

Birth Year Distribution
160
140
120
100
80
60
40
20
0
1900
1920
1940
1960
1980
2000
BIrth Year Distribution w/o Outliers
120
100
80
60
40
20
0
1940
1950
1960
1970
1980
1990

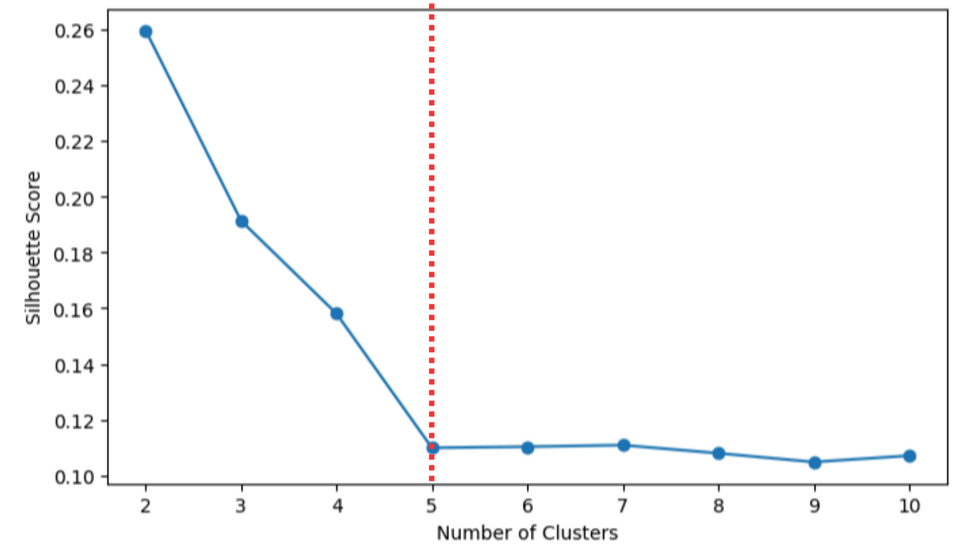

Correlation Matrix
Year_Birth 1 -0.21 -0.014 -0.16 -0.013 -0.042 -0.041 -0.021 -0.059
Income -0.21 1 0.0067 0.73 0.54 0.7 0.55 0.56 0.42
Recency -0.014 0.0067 1 0.016 -0.0049 0.026 0.0012 0.026 0.019
MntWines -0.16 0.73 0.016 1 0.38 0.59 0.4 0.39 0.39
MntFruits -0.013 0.54 -0.0049 0.38 1 0.57 0.59 0.57 0.39
MntMeatProducts -0.042 0.7 0.026 0.59 0.57 1 0.6 0.56 0.38
MntFishProducts -0.041 0.55 0.0012 0.4 0.59 0.6 1 0.58 0.43
MntSweetProducts -0.021 0.56 0.026 0.39 0.57 0.56 0.58 1 0.36
MntGoldProds -0.059 0.42 0.019 0.39 0.39 0.38 0.43 0.36 1
Year_Birth
Income
Recency
MntWines
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
MntGoldProds
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

MODELING AND ANALYSIS

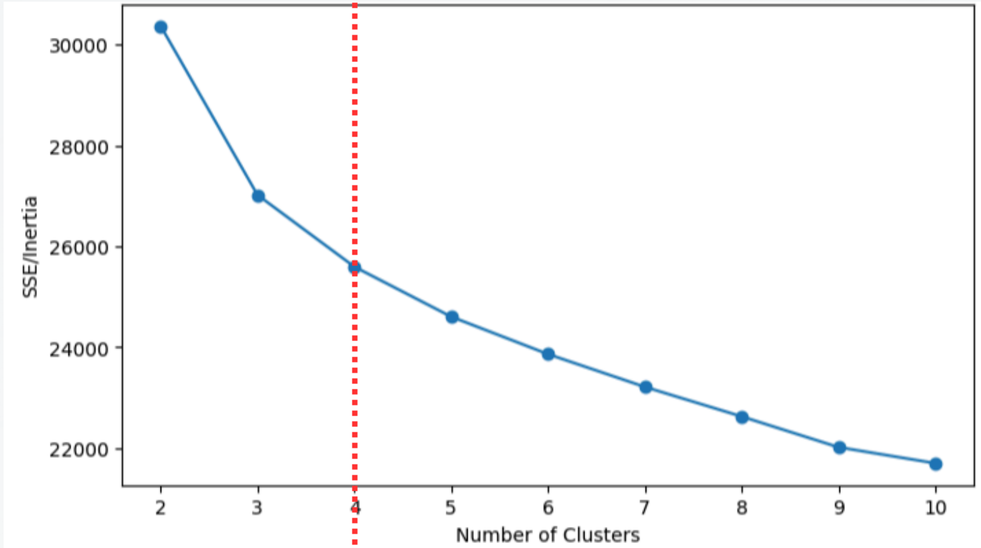
- **ML approach:** K-means, agglomerative clustering
- Evaluated models and chose number of clusters with sum of squared errors and silhouette scores
- Used principle component analysis to confirm results and improve performance/efficiency

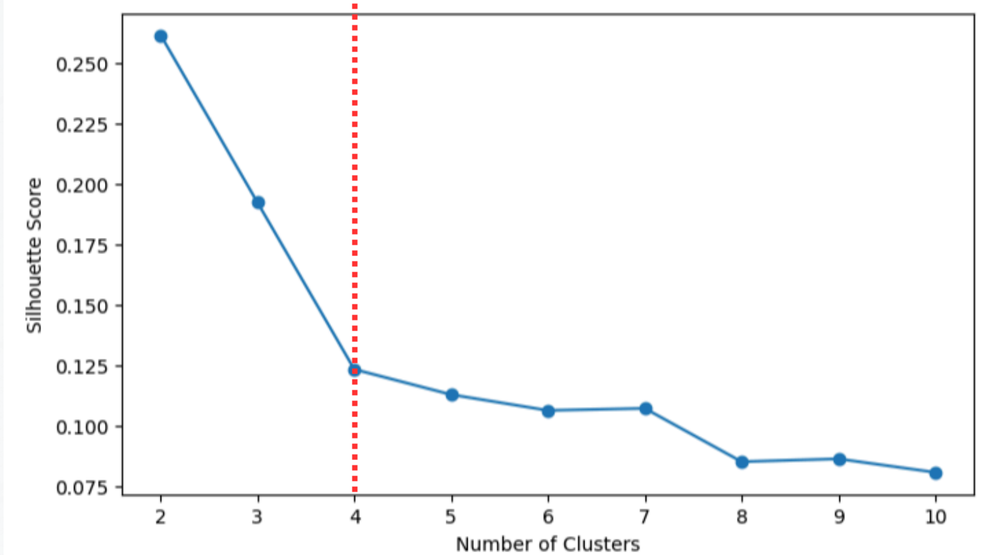
WITHOUT PCA



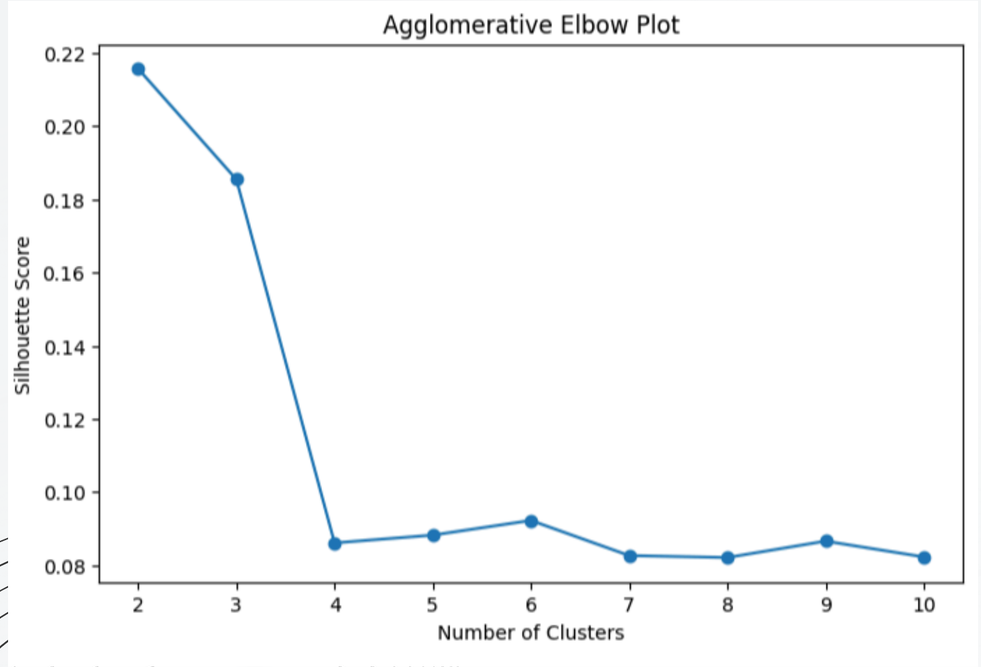


WITH PCA

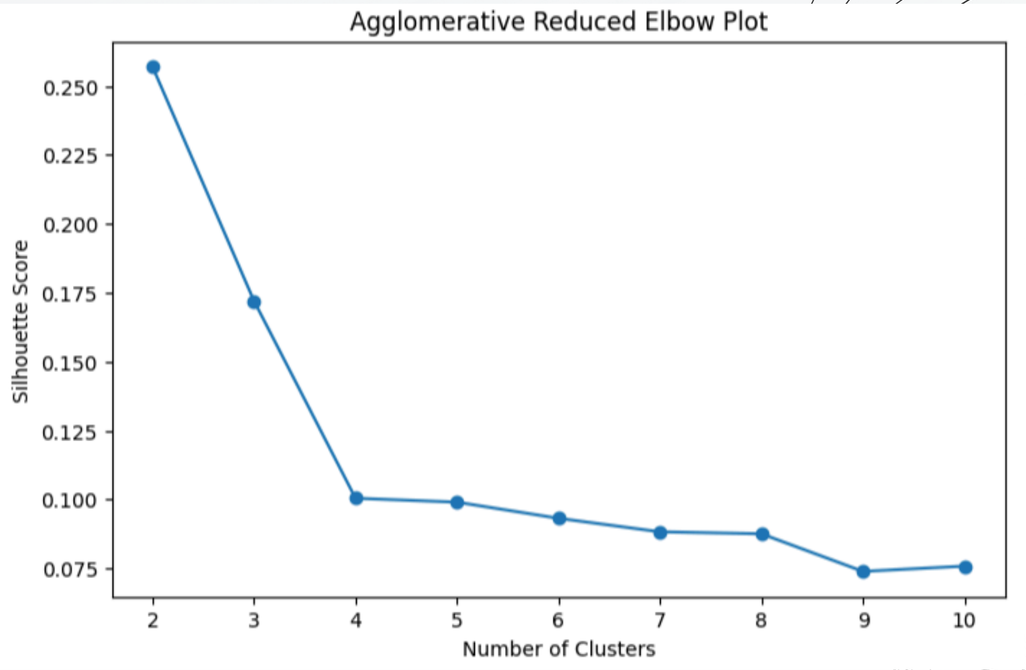




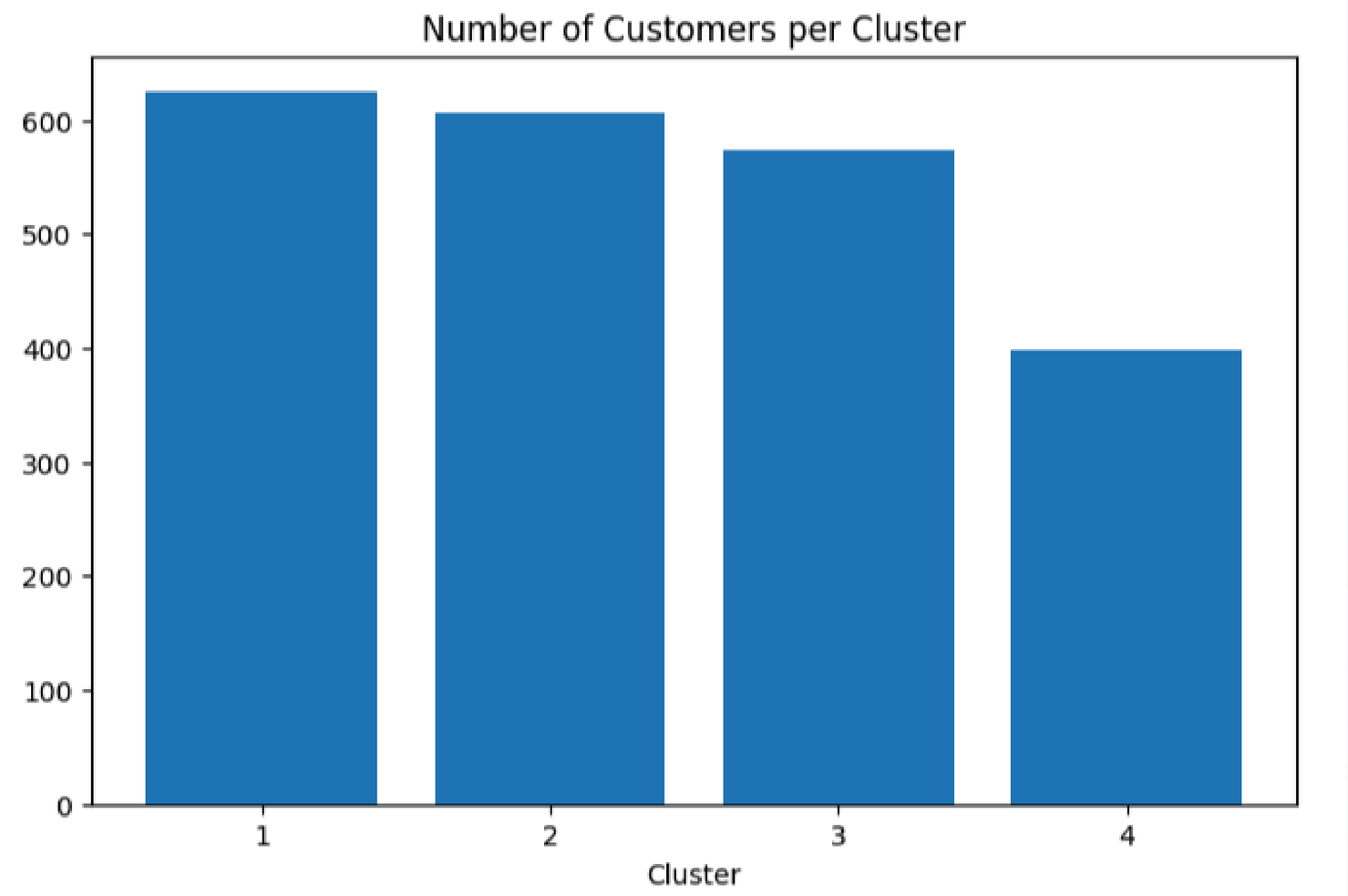
WITHOUT PCA

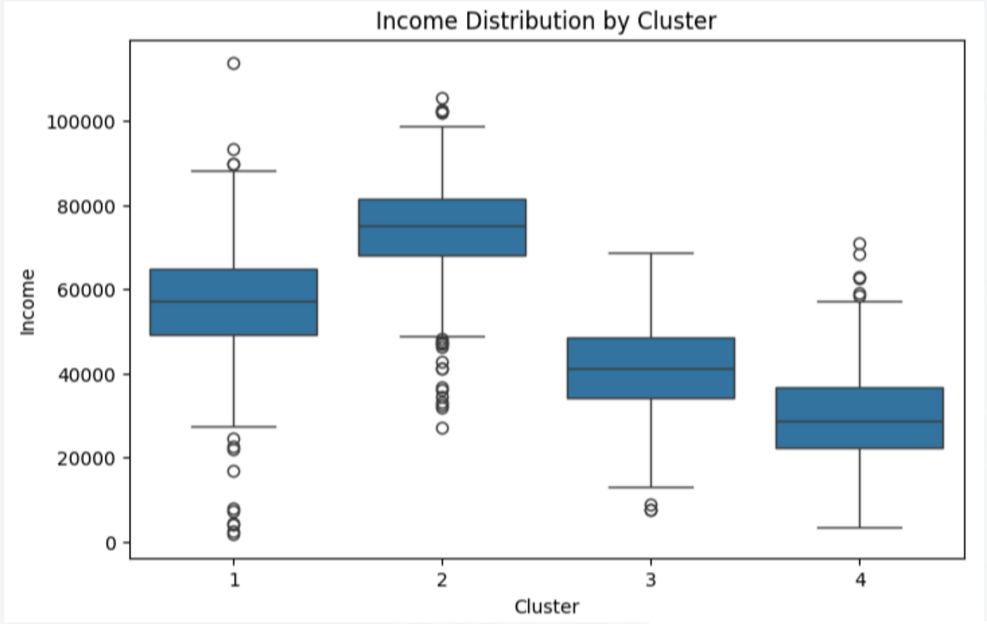


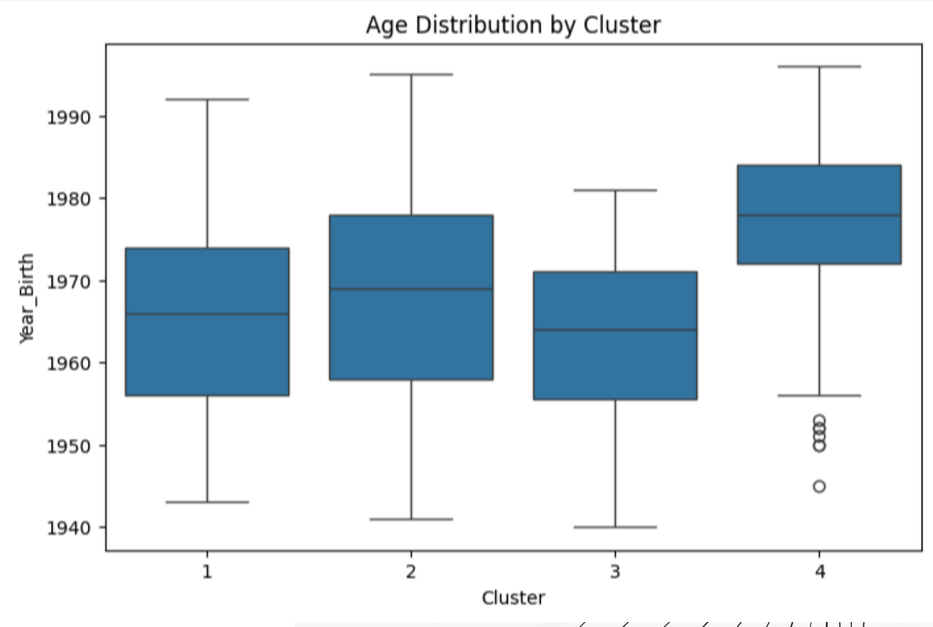
WITH PCA

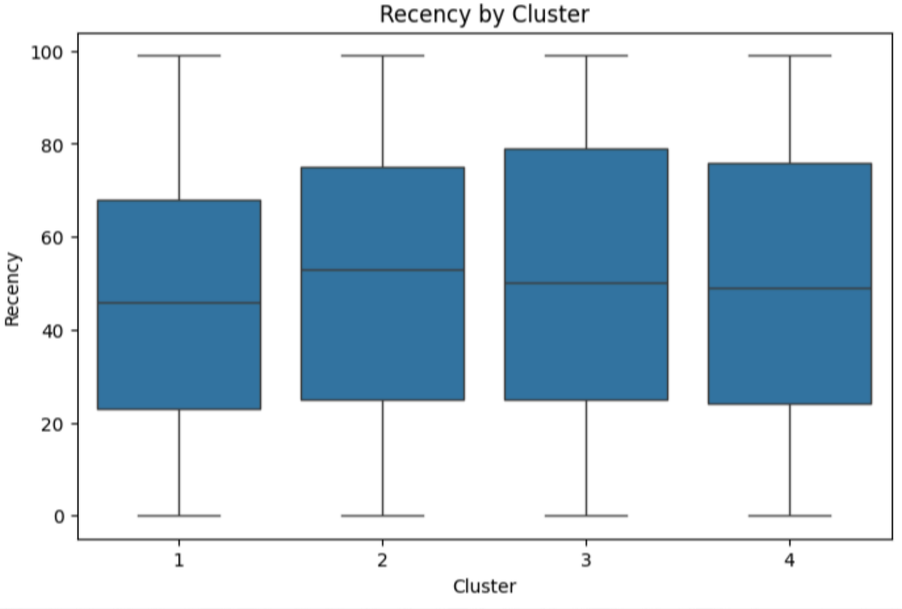


AGGLOMERATIVE WITH PCA




Income Distribution by Cluster
Income
0
20000
40000
60000
80000
100000
1
2
3
4
Cluster


Age Distribution by Cluster
Year_Birth
1940
1950
1960
1970
1980
1990
1
2
3
4
Cluster


Recency by Cluster
Recency
0
20
40
60
80
100
1
2
3
4
Cluster

CONCLUSION

- While the clusters do have some overlap, there are distinct characteristics for each, especially in income, frequency, and age

- **Future work:** Combining this analysis with qualitative research to create formal customer assessments/profiles