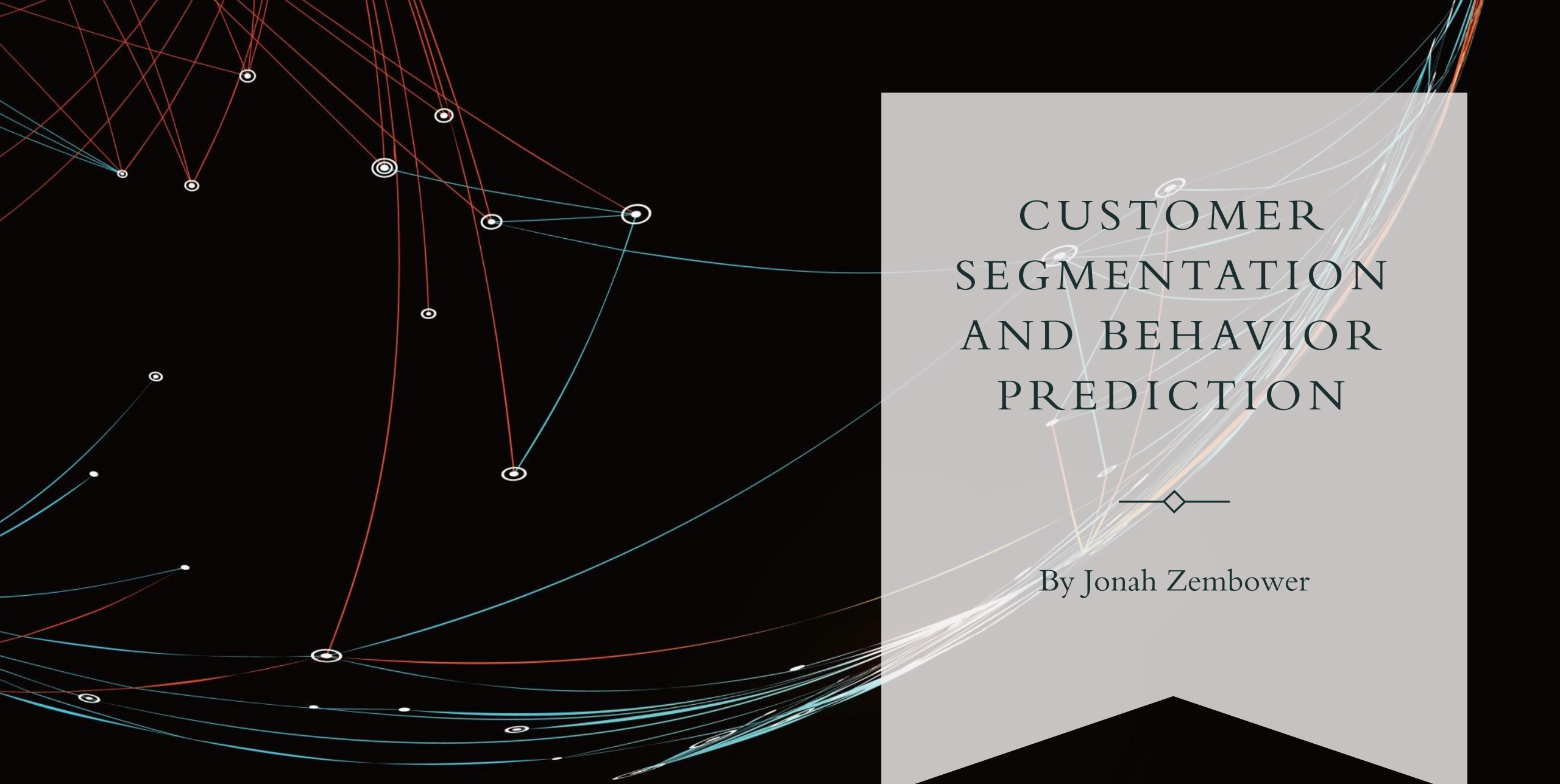


CUSTOMER SEGMENTATION AND BEHAVIOR PREDICTION

By Jonah Zembower

Data Preprocessing

-Customer_ID: Unique identifier for the customer

- Age: Customer's age

- Gender: Customer's gender

- Annual_Income: Annual income of the customer

- Total_Purchases: Total number of purchases made by the customer

- Average_Purchase_Value: Average value of purchases

- Product_Category_Most_Purchased: Category of the most purchased products

- Website_Visits_Last_Month: Number of times the customer visited the website in the last month

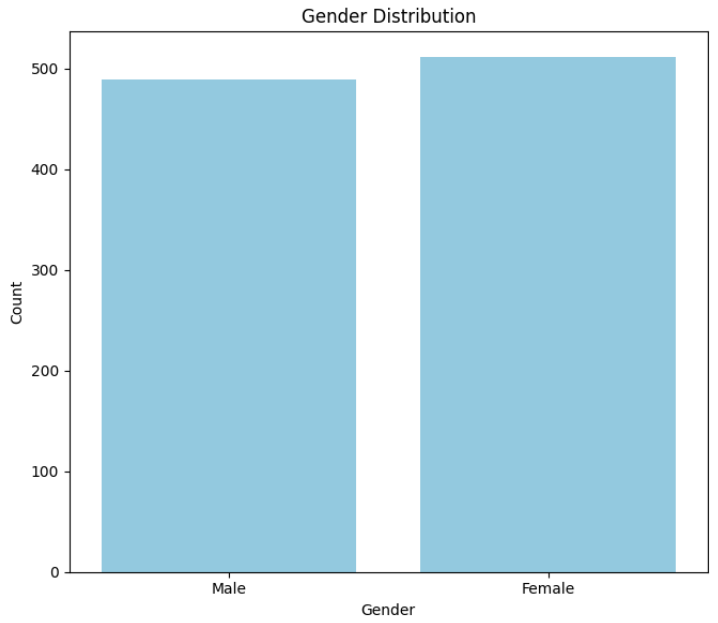
- Marketing_Emails_Opened: Number of marketing emails opened by the customer

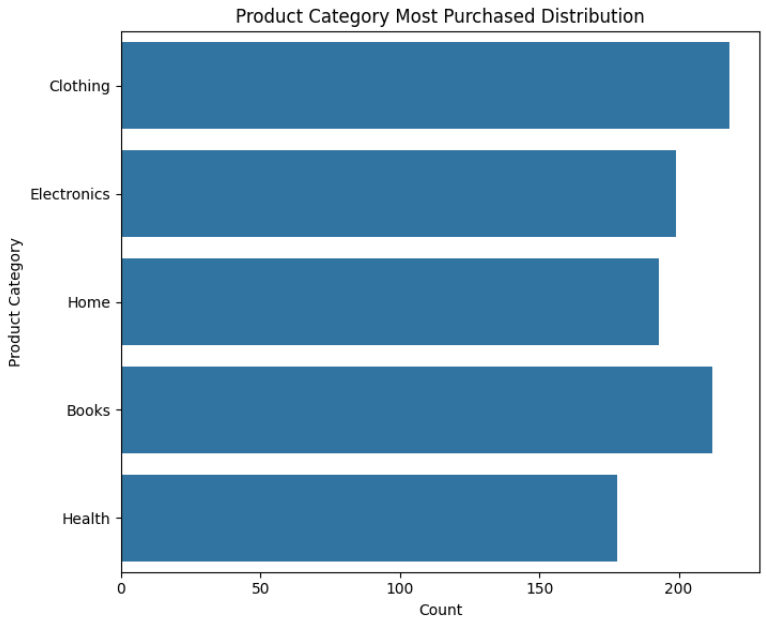
- Hours_Spent_on_Support_Calls: Total hours spent by the customer on support calls

- churn: 1 if they are leaving as a customer, and 0 if they stay

Categorical Features Distributions

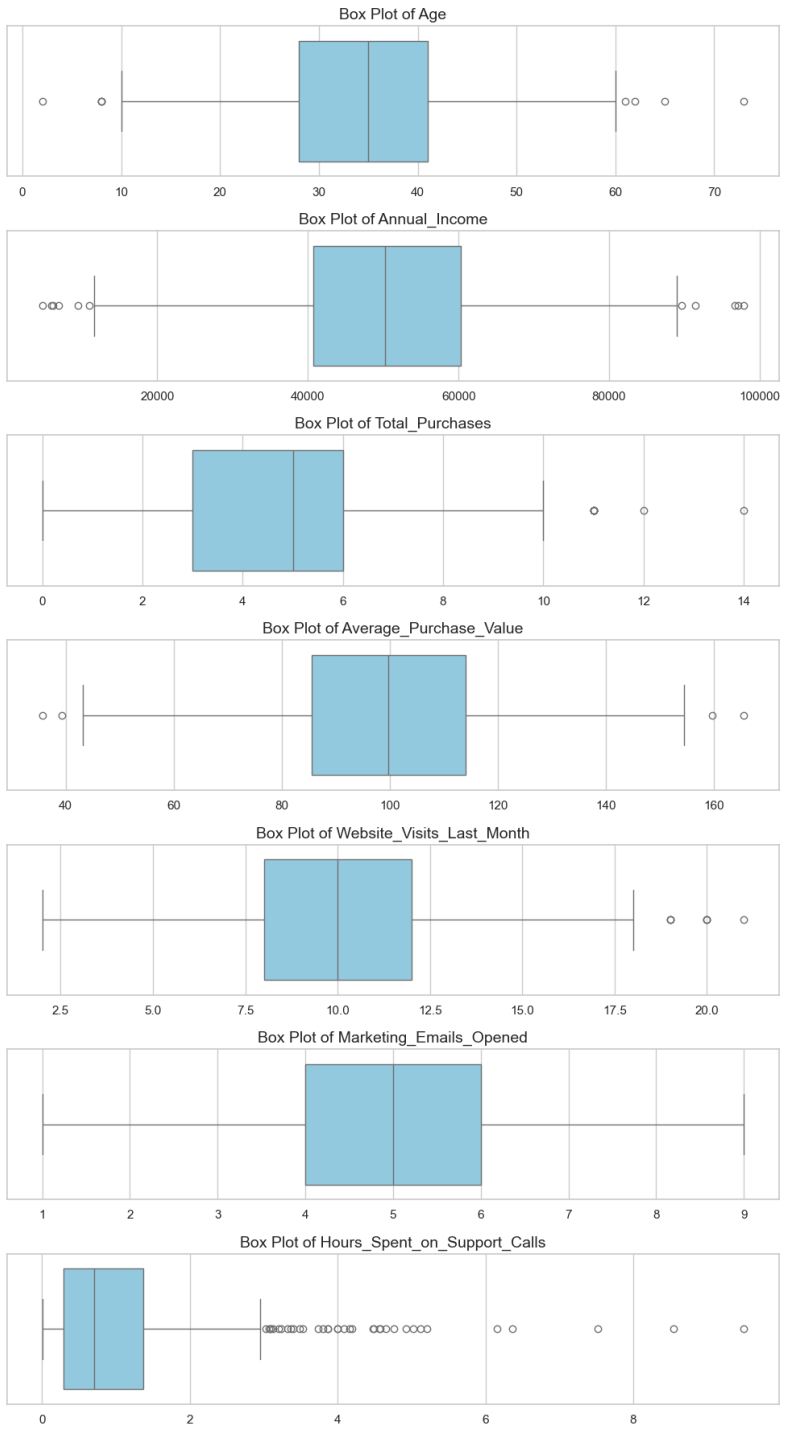
- Females > Males
- Clothing items most purchased
- Health items least purchased





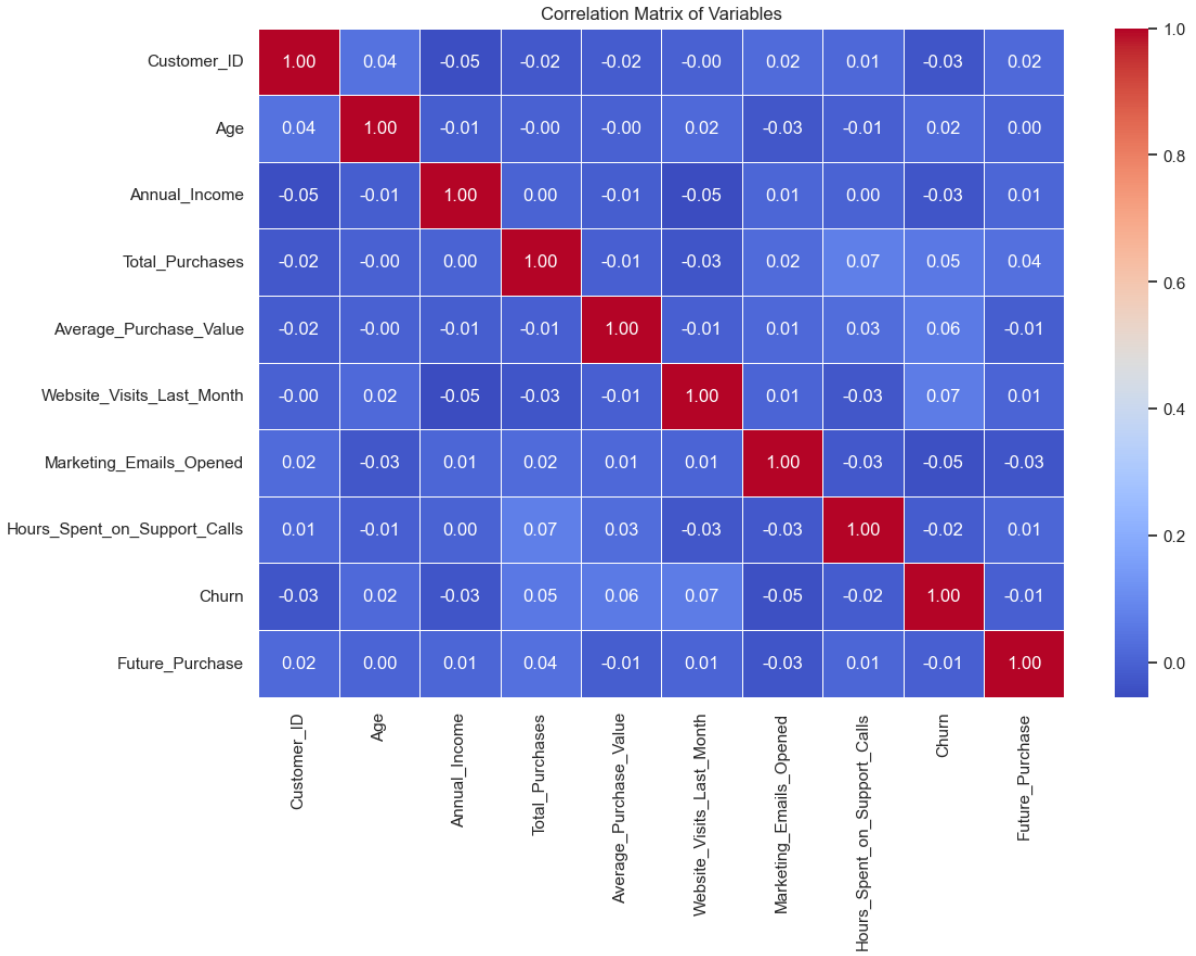
Numerical Features Distributions

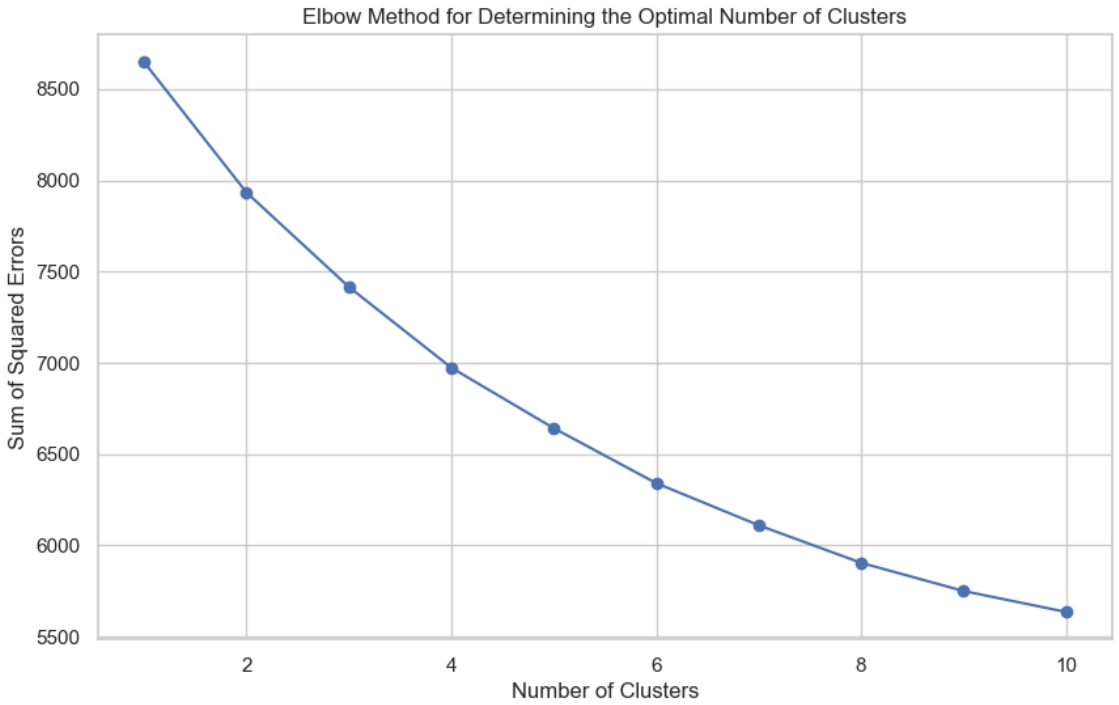
- Data is mostly evenly spread.
- Total purchases and hours spent on support calls are uneven.



Correlation Matrix of Features

- No correlating variables present.





Determining Clusters

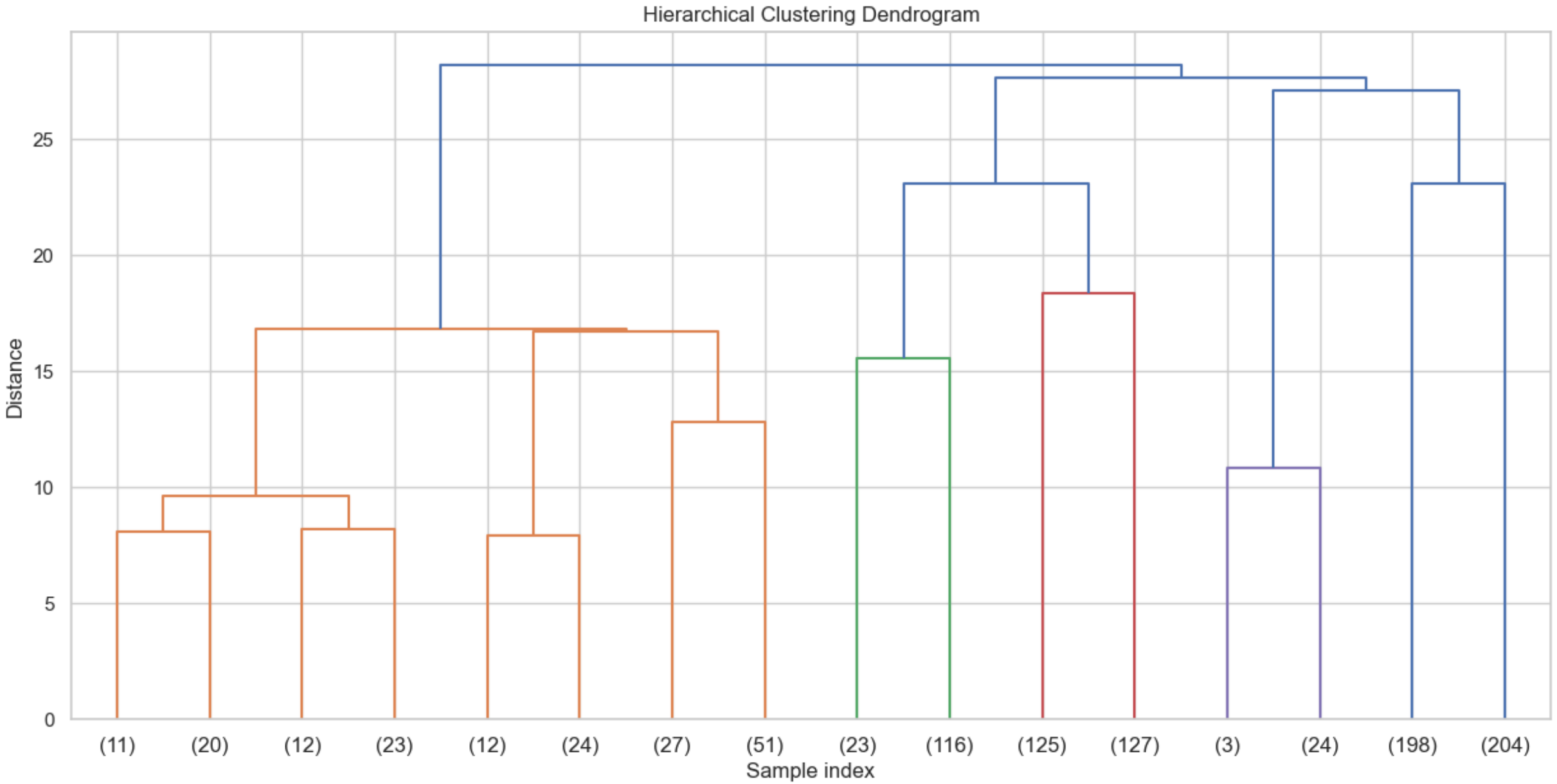
Most likely k = 3 or k=4

Silhouette Score

- SCORE_3 = 0.0834101672974745

- SCORE_4 = 0.0788651186252828 9

Hierarchical Clustering Dendrogram



Silhouette Score for Agglomerative

SCORE_3 =
0.0406311883047711

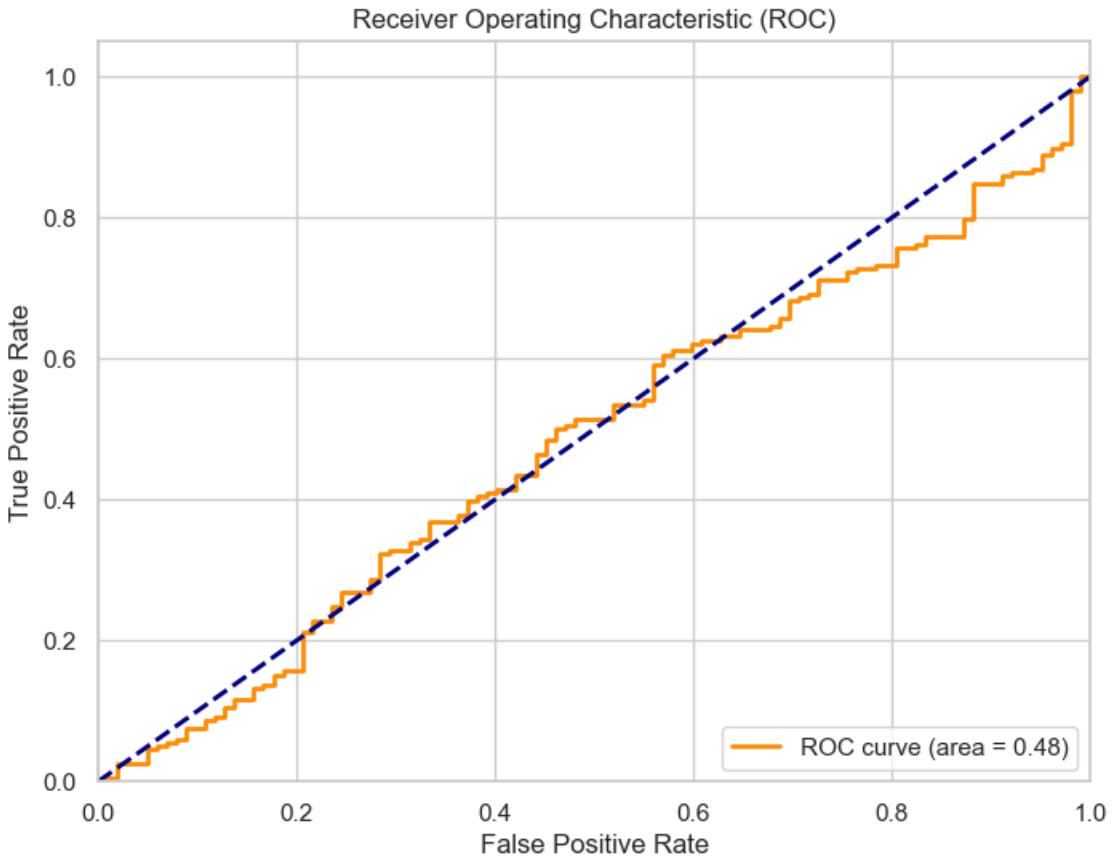
SCORE_4 =
0.0471788385838758 65

Customer Personas for k=3

	Age	Annual_Income	Total_Purchases	Average_Purchase_Value	Website_Visits_Last_Month	Marketing_Emails_Opened	Hours_Spent_on_Support_Calls	Gender_Female	Gender_Male
0	33.887498	47865.824948	5.649999	104.108542	10.243750	4.643750	2.693023	0.581250	0.418750
1	36.583956	54270.889856	5.270676	111.967372	9.022556	5.877194	0.682716	0.531328	0.468672
2	33.231289	47920.383852	4.317460	87.178206	10.616779	4.433107	0.647564	0.467120	0.532880

Product_Category_Most_Purchased_Books	Product_Category_Most_Purchased_Clothing	Product_Category_Most_Purchased_Electronics	Product_Category_Most_Purchased_Health
0.193750	0.243750	0.193750	0.225000
0.180451	0.258145	0.172932	0.182957
0.247166	0.172336	0.224490	0.156463

Product_Category_Most_Purchased_Home	Churn	Future_Purchase
0.143750	0.156250	0.700000
0.205514	0.167920	0.676692
0.199546	0.145125	0.678000



ROC Curve and AUC Score

- AUC score = 0.48

Neural Network MLP vs. Logistic Regression

The Multi-layer Perceptron (MLP) model's performance metrics are as follows:
Accuracy: 56 %
Precision: 64.73 %
Recall: 73.23 %

The logistic regression model's performance metrics are:
Accuracy: 66 %
Precision: 66 %
Recall: 100 %

Finding Best Configuration using Cross Validation

- Configuration (32, 16): 56.57 % Accuracy

- Configuration (128, 64): 55.71 % Accuracy

- Configuration (64, 32, 16, 8): 58.29 % Accuracy

Other Neural Networks

Decision Tree:

Accuracy: 56.7 %
Precision: 67 %
Recall: 67.7 %

SVM

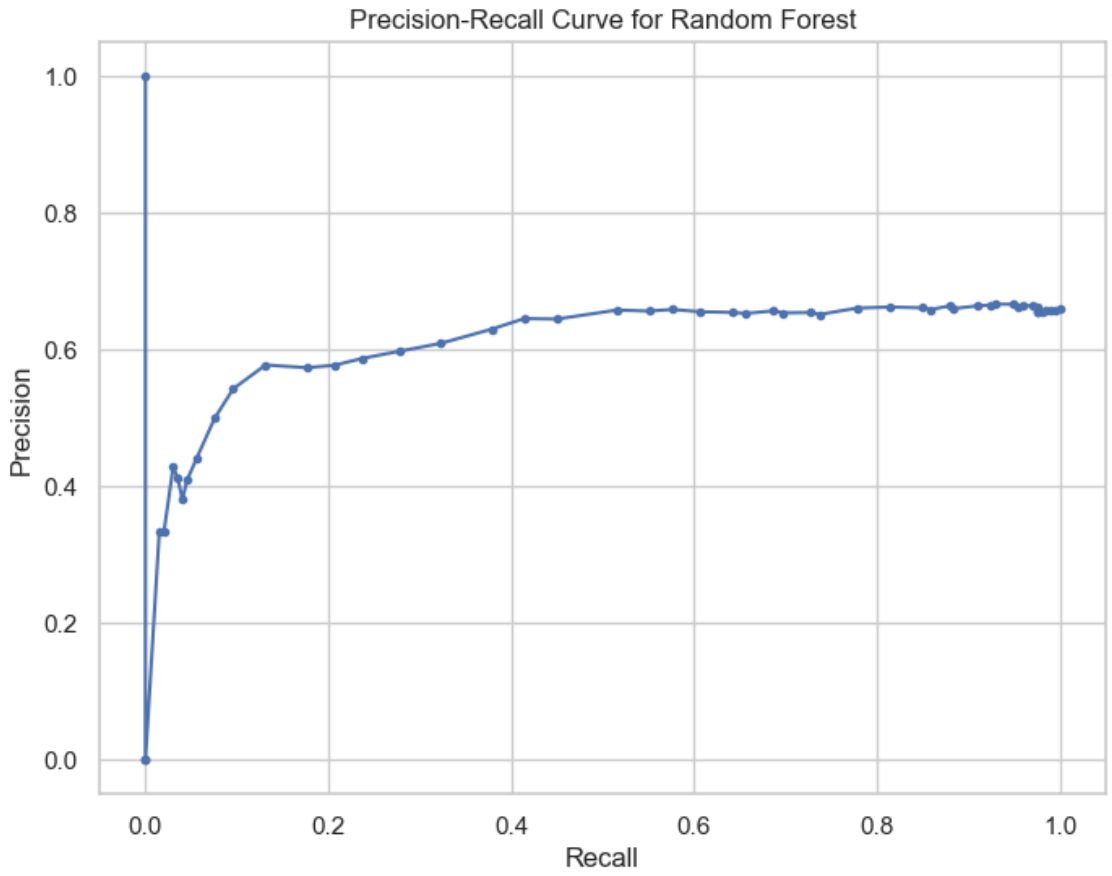
Accuracy: 66 %
Precision: 66 %
Recall: 100 %

Random Forest

Accuracy: 64.7 %
Precision: 66.7 %
Recall: 92.9 %

Gradient Boosting

Accuracy: 62 %
Precision: 65.6 %
Recall: 89.4 %



Precision-Recall Curve for Random Forest

- Threshold of 0.4 to 0.5 potentially
- As recall increases, precision decreases in incline.

Testing Threshold 0.4

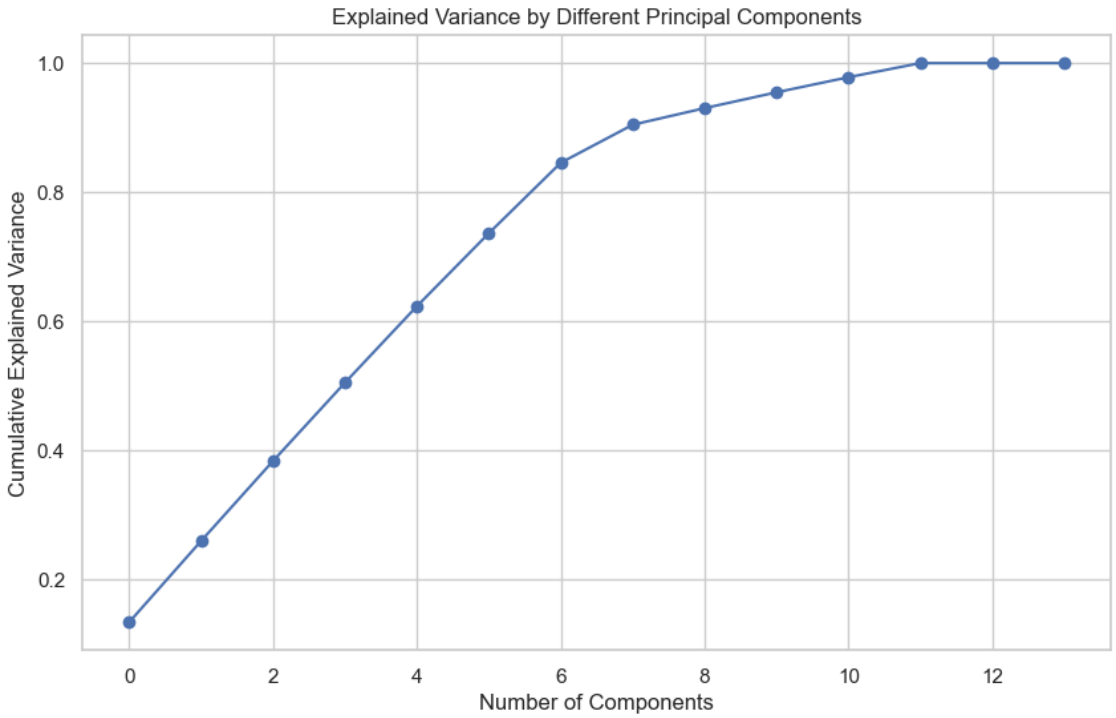
ACCURACY: 64.33 %

PRECISION: 65.42 %

RECALL: 97.47 %

Random Forest Thresholds Testing

- Tested between 0.3 and 0.5:
  - Found 0.3 is the best
    - F1 score: 0.795
  - New Predictions:
    - Accuracy: 66 %
    - Precision: 66 %
    - Recall: 100 %



PCA Analysis

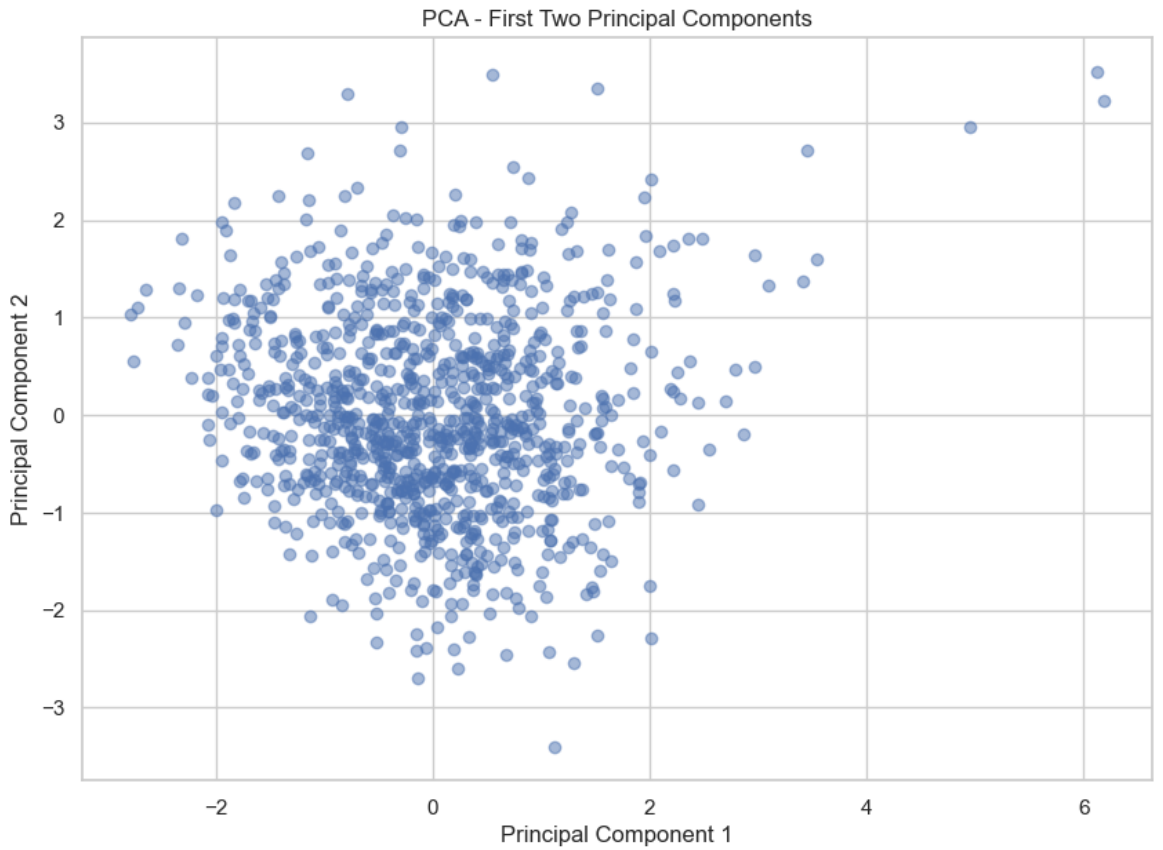
- About 73.6% variance by the 6th component
- After 8th component the variance tapers off

Loading Scores

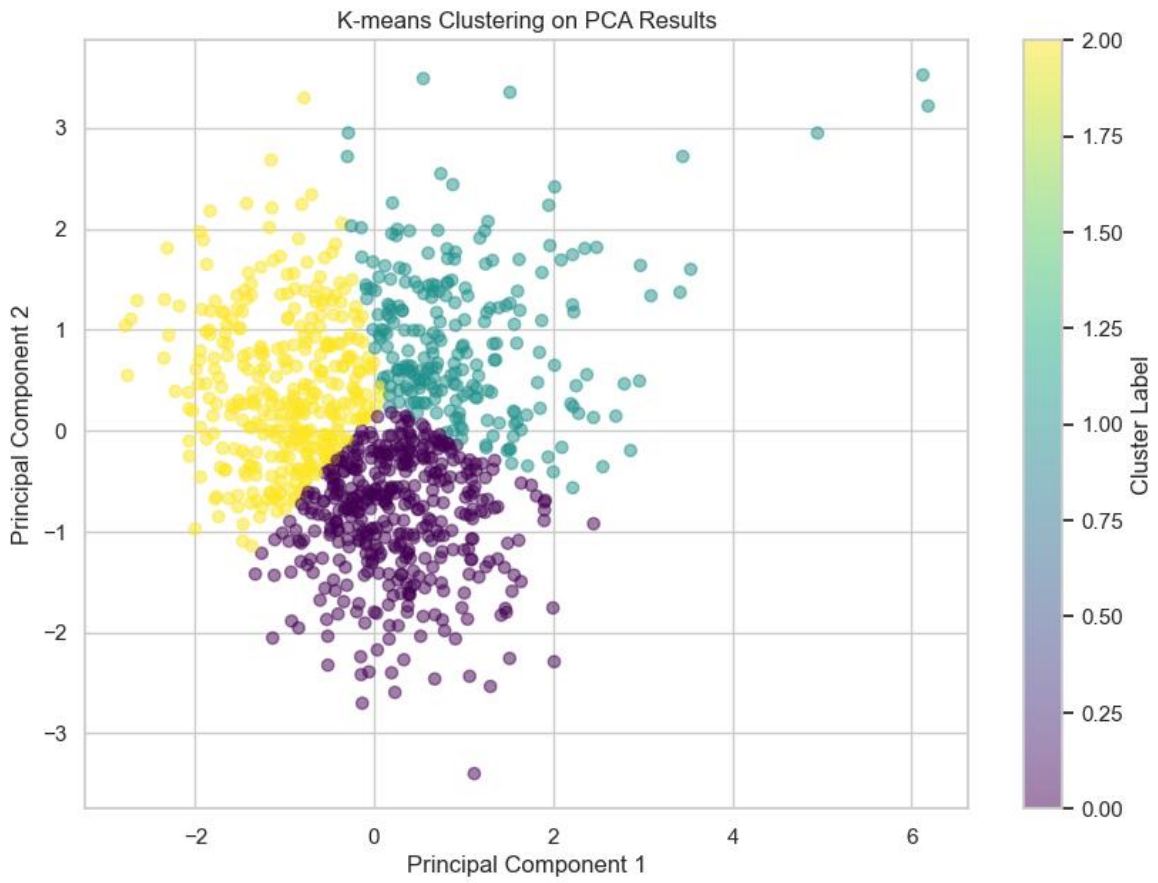
	0	1	2	3	4	5	6	7
Age	-0.201689	0.382449	0.419304	0.124943	0.771158	-0.095048	0.093643	-0.093012
Annual_Income	0.286248	-0.558823	0.362921	-0.120747	0.270024	0.596551	-0.171777	0.017354
Total_Purchases	0.543362	0.230120	-0.106401	0.512654	0.073039	-0.084194	-0.600776	0.063508
Average_Purchase_Value	0.138048	0.180178	-0.498746	-0.694789	0.374824	0.017947	-0.257354	0.093755
Website_Visits_Last_Month	-0.503243	0.293250	-0.273718	0.233960	-0.055724	0.698482	-0.199569	-0.002641
Marketing_Emails_Opened	0.032095	-0.409442	-0.593408	0.404782	0.413018	-0.034039	0.373051	-0.050846
Hours_Spent_on_Support_Calls	0.549633	0.449424	-0.011988	-0.022850	-0.092687	0.371382	0.589791	0.014010
Gender_Female	0.056059	0.008900	-0.037898	-0.048953	-0.034939	0.014384	-0.060647	-0.696548
Gender_Male	-0.056059	-0.008900	0.037898	0.048953	0.034939	-0.014384	0.060647	0.696548
Product_Category_Most_Purchased_Books	-0.005997	0.001645	0.042185	-0.006990	-0.006201	-0.000621	0.002704	0.051408
Product_Category_Most_Purchased_Clothing	0.007209	-0.009874	-0.013566	-0.012729	0.018623	0.012548	0.001420	-0.025527
Product_Category_Most_Purchased_Electronics	-0.005320	0.011837	-0.005689	0.010368	-0.034825	0.000965	0.004106	-0.039091
Product_Category_Most_Purchased_Health	0.014156	-0.000224	-0.017529	0.004709	0.009928	-0.000719	0.008053	-0.002577
Product_Category_Most_Purchased_Home	-0.010048	-0.003384	-0.005402	0.004642	0.012475	-0.012173	-0.016283	0.015787

FIRST TWO PRINCIPLE COMPONENTS

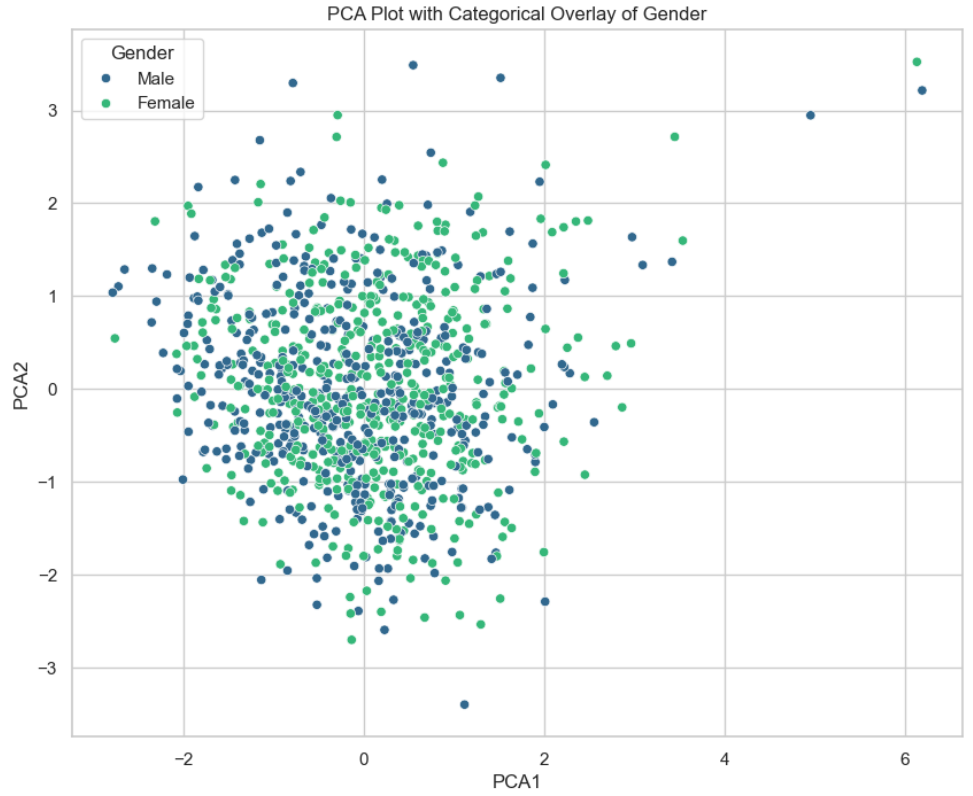
Dense cluster around a similar event



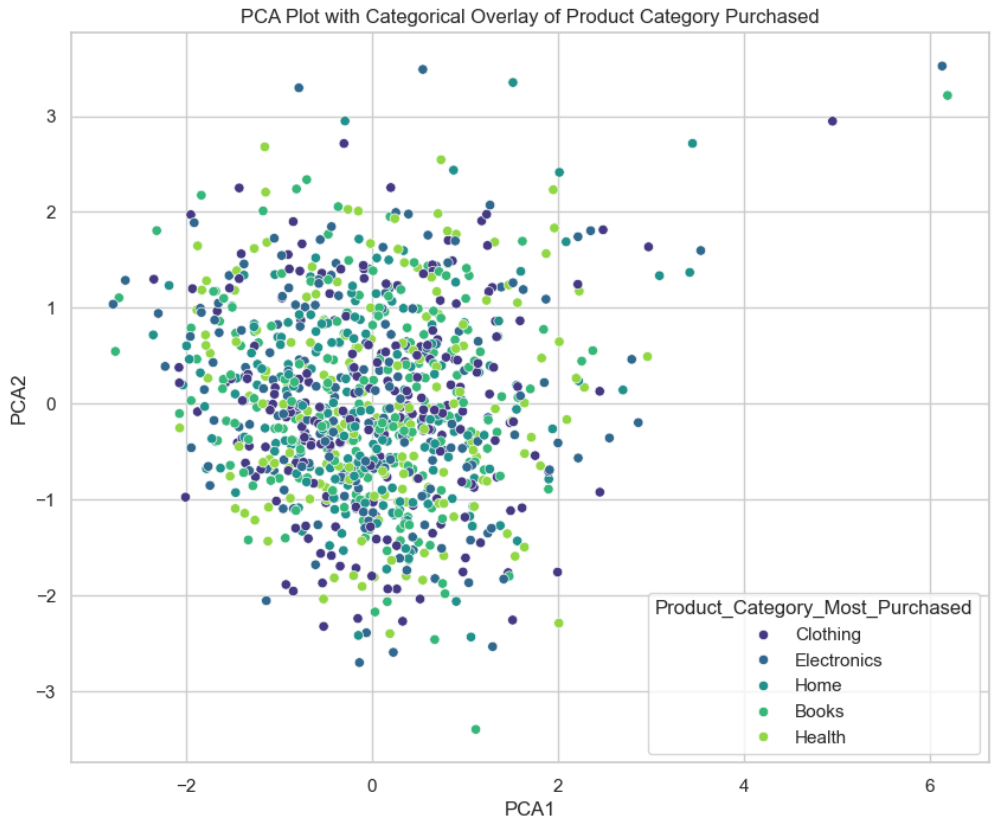
K MEANS CLUSTERING ON PCA



PCA PLOT WITH GENDER



PCA Plot with Product Category Purchased



Conclusion

Variables were discernibly distinct from each other in correlation.

The variables did show a lot of overlap though in principal component analysis.

Gender	Product Category Most Purchased

Overall, focus on annual income, total purchases, and website visits for future analysis.

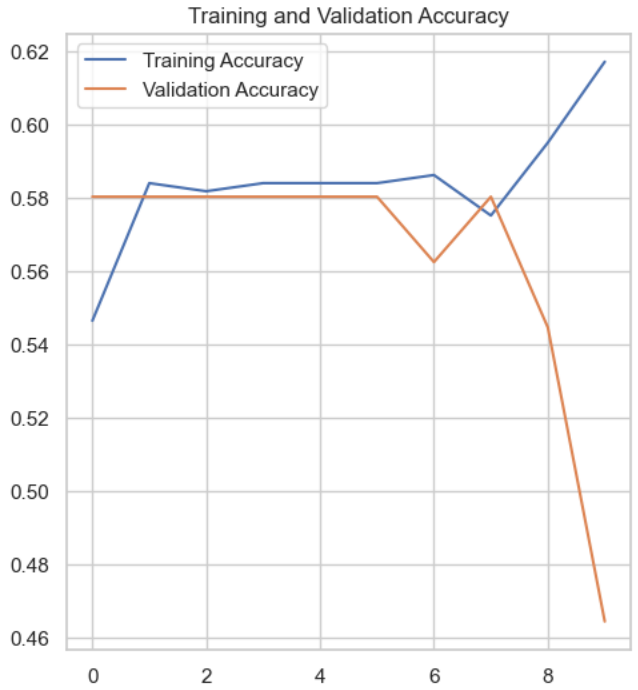
Image Analysis of Avengers

Given images of various Marvel Avengers

- Captain America
- Thor
- Hawkeye
- Iron Man
- Black Widow

Goal is to train a neural network

Initial Training Set Results



Negative

Negative

Positive

Negative

Positive

Positive

Positive

Negative

Negative

TEST SET RESULTS