Part I:

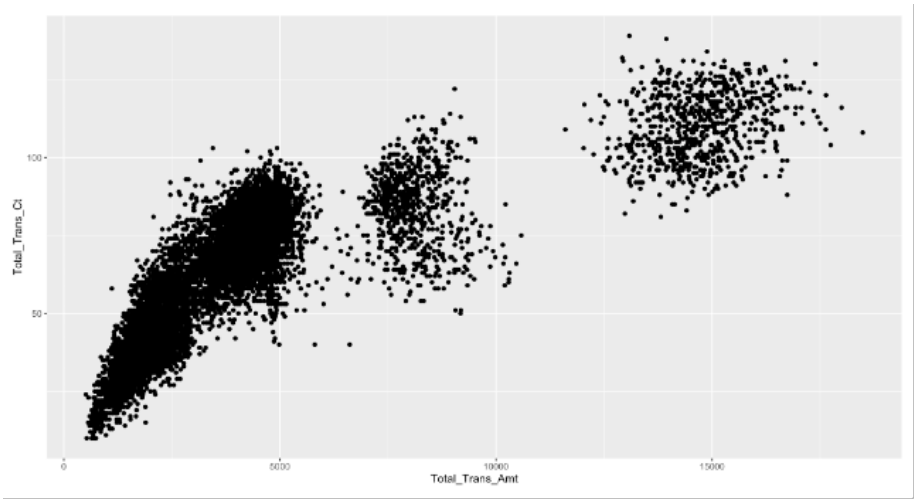
Hypothesis:

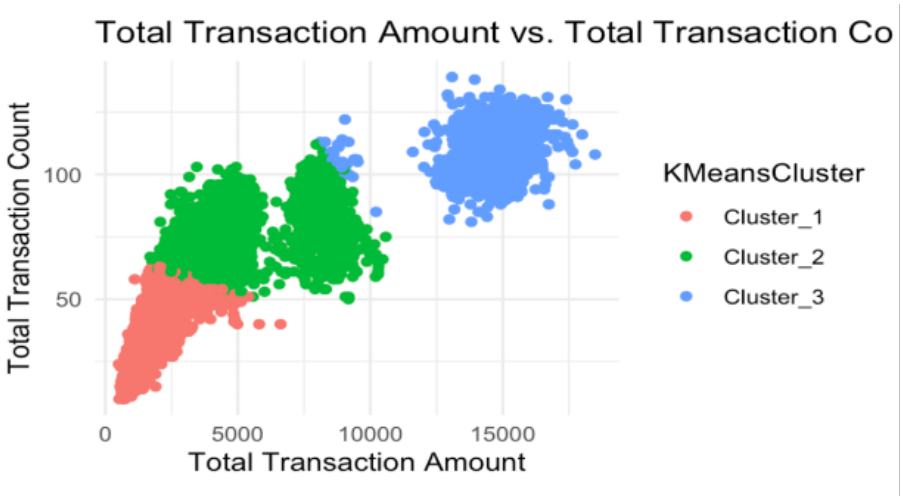
Higher Income and Higher Education Level Customers will exhibit higher Total Transaction Count and Amount

1. Columns Included In Cluster:

a. Total Transaction Count
b. Total Transaction Amount
c. Education Level
d. Income Level

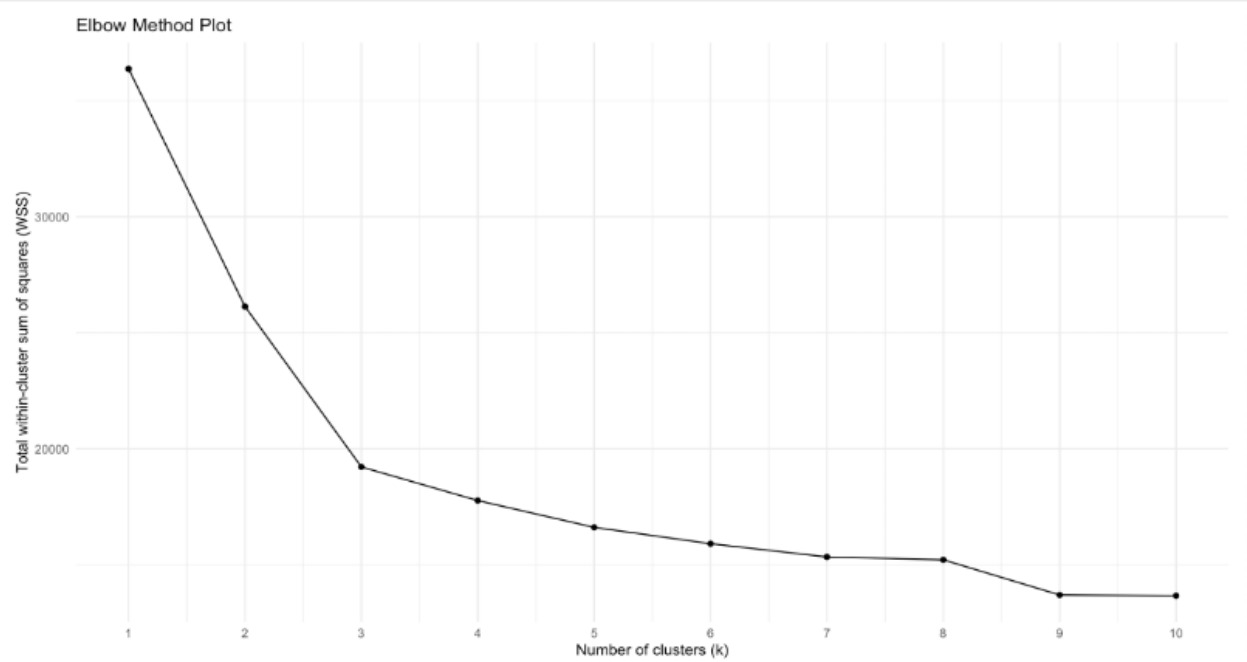
By plotting total transaction amount by total transaction count, there appear to be 3 clusters from the plot, we create the model using 3 clusters.



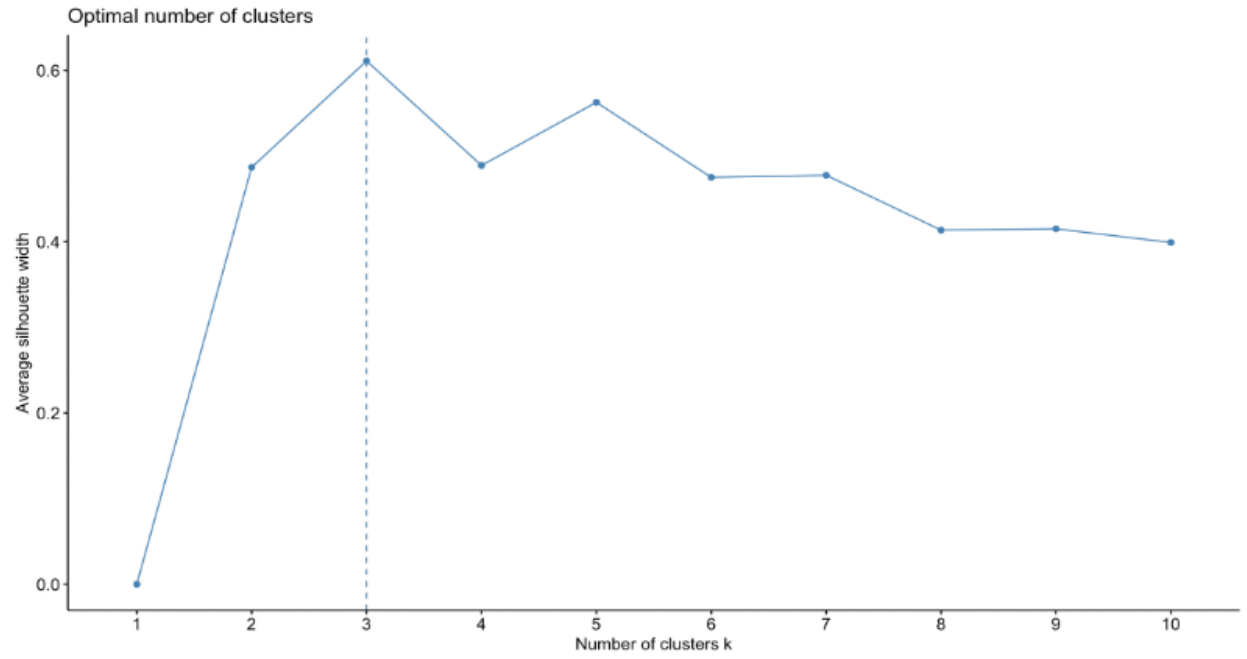


2. Optimal Cluster Size

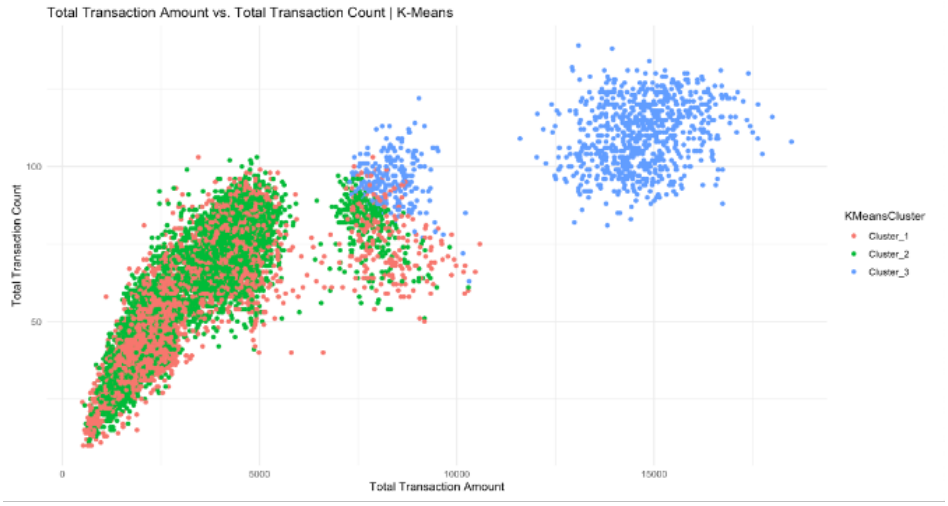
a. Elbow Method Graph



b. Silhouette Method Graph



Both methods confirm that the optimal number of clusters is three. Total Revolving balance was also added as a variable to test the hypothesis that higher income and higher education level customers had lower total revolving balance, but was ultimately removed as there was no significant difference in total revolving balance between the clusters. Demonstration shown below.



3. Clusters Analysis

Summary Statistics By Cluster

KMeansCluster <fctr>	Total_Trans_Ct <dbl>	Total_Trans_Amt <dbl>	Total_Revolving_Bal <dbl>	count <dbl>
Cluster_1	40.20073	1949.276	1136.963	3821
Cluster_2	75.54041	4686.532	1151.287	5531
Cluster_3	110.19742	14491.338	1372.533	775

3 rows

K Means Summary

Total_Trans_Amt <dbl>	Total_Trans_Ct <dbl>	Income_Category_1 <dbl>	Income_Category_2 <dbl>	Income_Category_3 <dbl>	Income_Category_4 <dbl>	Income_Category_5 <dbl>
-0.72261305	-1.0505011	-0.05915146	-0.10045676	0.06815687	-0.1176131	0.03122502
0.08314254	0.4550722	-0.13694022	-0.08520006	0.13116163	-0.1461358	0.04099000
2.96934588	1.9315620	-0.02513836	-0.11713373	0.05597383	-0.0999777	0.03023716

Income_Category_5 <dbl>	Education_Level_1 <dbl>	Education_Level_2 <dbl>	Education_Level_3 <dbl>	Education_Level_4 <dbl>	Education_Level_5 <dbl>	Education_Level_6 <dbl>
0.03122502	-0.07265504	-0.1046544	-0.04775896	0.02684671	0.08849218	0.1551206
0.04099000	-0.07636509	-0.1034671	-0.04886284	0.02523387	0.09163100	0.1473938
0.03023716	-0.07973831	-0.1006618	-0.05425751	0.01070965	0.09108837	0.1701760

Summary Statistics By Cluster

KMeansCluster <fctr>	count <dbl>	Customer_Age <dbl>	Dependent_count <dbl>	Months_on_book <dbl>	Total_Relationship_Count <dbl>	Months_Inactive_12_mon <dbl>	Contacts_Count_12_mon <dbl>
Cluster_1	3821	46.61921	2.225857	36.17980	4.063072	2.399895	2.681497
Cluster_2	5531	46.29579	2.438076	35.88248	3.847225	2.318387	2.333032
Cluster_3	775	45.09548	2.283871	35.01677	2.330323	2.214194	2.212903

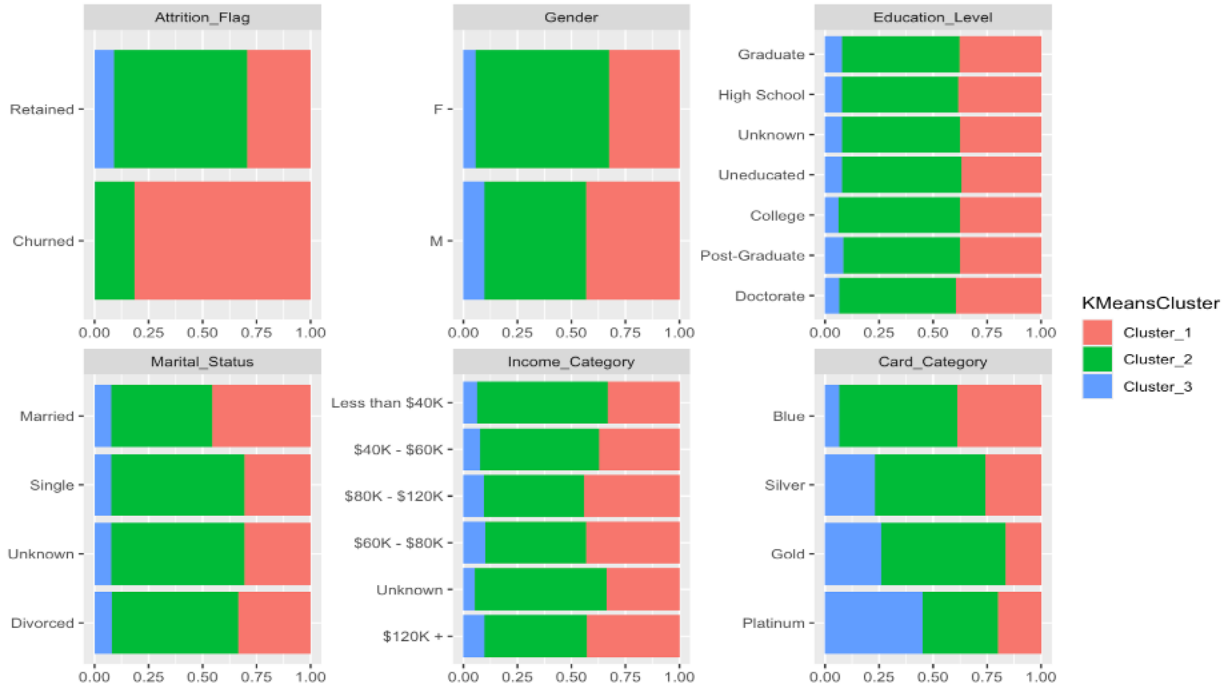
Credit_Limit <dbl>	Total_Revolving_Bal <dbl>	Avg_Open_To_Buy <dbl>	Total_Amt_Chng_Q4_Q1 <dbl>	Total_Trans_Amt <dbl>	Total_Trans_Ct <dbl>	Total_Ct_Chng_Q4_Q1 <dbl>
8428.854	1136.963	7291.891	0.7607404	1949.276	40.20073	0.6683805
7987.726	1151.287	6836.438	0.7571103	4686.532	75.54041	0.7373578
14231.009	1372.533	12858.476	0.7761974	14491.338	110.19742	0.7489910

3 rows | 9-15 of 16 columns

Total_Revolving_Bal <dbl>	Avg_Open_To_Buy <dbl>	Total_Amt_Chng_Q4_Q1 <dbl>	Total_Trans_Amt <dbl>	Total_Trans_Ct <dbl>	Total_Ct_Chng_Q4_Q1 <dbl>	Avg_Utilization_Ratio <dbl>
1136.963	7291.891	0.7607404	1949.276	40.20073	0.6683805	0.2597925
1151.287	6836.438	0.7571103	4686.532	75.54041	0.7373578	0.2986653
1372.533	12858.476	0.7761974	14491.338	110.19742	0.7489910	0.1796929

3 rows | 10-16 of 16 columns

Distribution of Clusters by Categorical Variables



Cluster 1:

- **Average Total Transaction Count**: Relatively low (approx. 40.20), suggesting these customers use their credit cards less frequently than the other clusters.
- **Average Total Transaction Amount**: Also on the lower side (approx. $1949.28), indicating these transactions are not of high value compared to the other clusters.
- **Average Total Revolving Balance**: Moderate (approx. $1136.96), which could suggest that these customers may sometimes carry a balance from month to month, but the lowest amongst the three clusters.
- **Average Total Relationship Count:** Ranks highest of the three clusters with 4
- **Average Credit Limit:** Ranked 2nd of the three clusters with a $8,428 approximate average
- **Average Customer Age:** Ranks highest of the three clusters with 46.6

- **Count**: This cluster has the second highest number of customers (3821), making it the second largest segment.
- **Attrition_Flag**: This cluster has a higher proportion of churned customers compared to retained ones.
- **Gender**: Appears to have more male than female customers.
- **Marital_Status**: Married customers seem to be a significant proportion, followed by divorced and then single.
- **Income_Category**: Mostly concentrated in the $80K-$120K+ income categories.
- **Education_Level**: A good mix across different education levels.
- **Card_Category**: Primarily holders of blue cards, which might indicate a standard level of credit card service.

This cluster may represent more conservative spenders. This group is mostly made up of middle aged married people, mostly men. They mostly earn upwards of $80K and are spread across different education levels.

Cluster 2:

- **Average Total Transaction Count:** Medium (approx. 75.54), indicating these customers use their credit cards fairly frequently.
- **Average Total Transaction Amount:** Higher (approx. $4686.53), showing that not only do they use their cards often, but also spend more per transaction.
- **Average Total Revolving Balance:** Similar to Cluster 1 (approx. $1151.29), which is quite interesting given their higher transaction count and amount.
- **Average Total Relationship Count:** Ranks middle of the three clusters with 3.8
- **Average Credit Limit:** Ranked lowest of the three clusters with a $7,987 approximate average
- **Average Customer Age:** Ranks highest of the three clusters with 46.2
- **Count:** This is the largest cluster in terms of customers (5531).
- **Attrition_Flag**: There's a higher proportion of retained customers, but also a noticeable amount of churned customers.
- **Gender**: Appears to have more female than male customers.
- **Marital_Status**: Single status is dominant in this cluster, with divorced status and unknown following. There is a slightly higher proportion of divorced status here compared to Cluster 1.
- **Income_Category**: More diverse in terms of income, with significant portions in the ranges below $60K and Unknown categories.

- **Education_Level**: A good mix across different education levels.
- **Card_Category**: This cluster has a broader distribution of card categories, with a significant number of silver cardholders and a few gold and platinum compared to Cluster 1.

This segment seems to consist of middle-ground customers, who are using their credit cards frequently and also spending significantly. This group is mostly made up of middle aged single and divorced people, mostly women. They mostly earn below $60K or have unreported income and are spread across different education levels.

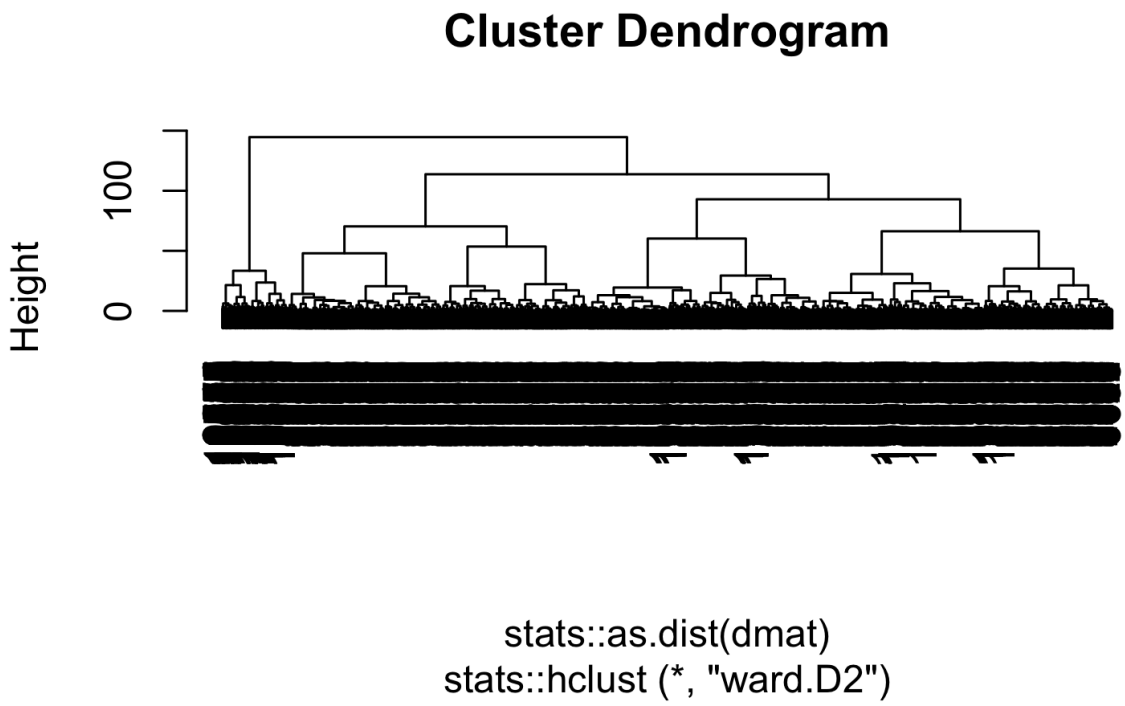
Cluster 3:

- **Average Total Transaction Count**: The highest (approx. 110.19), showing these customers are the most active in using their credit cards.
- **Average Total Transaction Amount**: Significantly the highest (approx. $14491.34), suggesting these are premium customers with high spending habits.
- **Average Total Revolving Balance**: The highest average revolving balance (approx. $1372.53), which could imply higher credit limits and possibly more comfort with maintaining a credit card balance.
- **Average Total Relationship Count:** Ranks lowest of the three clusters with 2.33
- **Average Credit Limit:** Ranked highest of the three clusters with a $14,231 approximate average
- **Count**: The smallest cluster (775 customers), which is not uncommon for a premium segment.
- **Attrition_Flag**: The proportion of retained customers is the highest among all three clusters, with very few churned customers.
- **Gender**: This cluster has a higher proportion of male customers compared to the other two clusters.
- **Marital_Status**: A good mix across different marital statuses.
- **Income_Category**: This cluster has the highest proportion of customers in the higher income categories ($80K-$120K and $120K+), indicating higher affluence.
- **Education_Level**: Contains the highest proportion of post-graduate and doctorate-educated individuals, but generally a good mix across different education levels..
- **Card_Category**: The distribution includes a higher proportion of gold and platinum cardholders, which are typically associated with higher spending and more exclusive benefits.

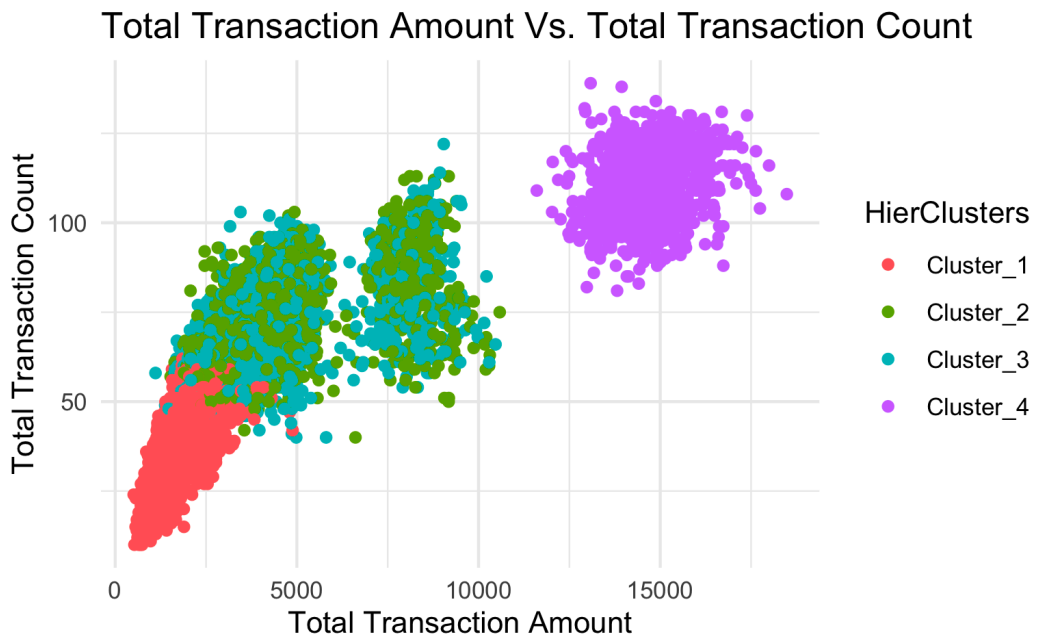
Cluster 3 represents the high-value segment, with customers who are frequent users and spend large amounts. This group may also be less price-sensitive and more service-oriented.

Part II:

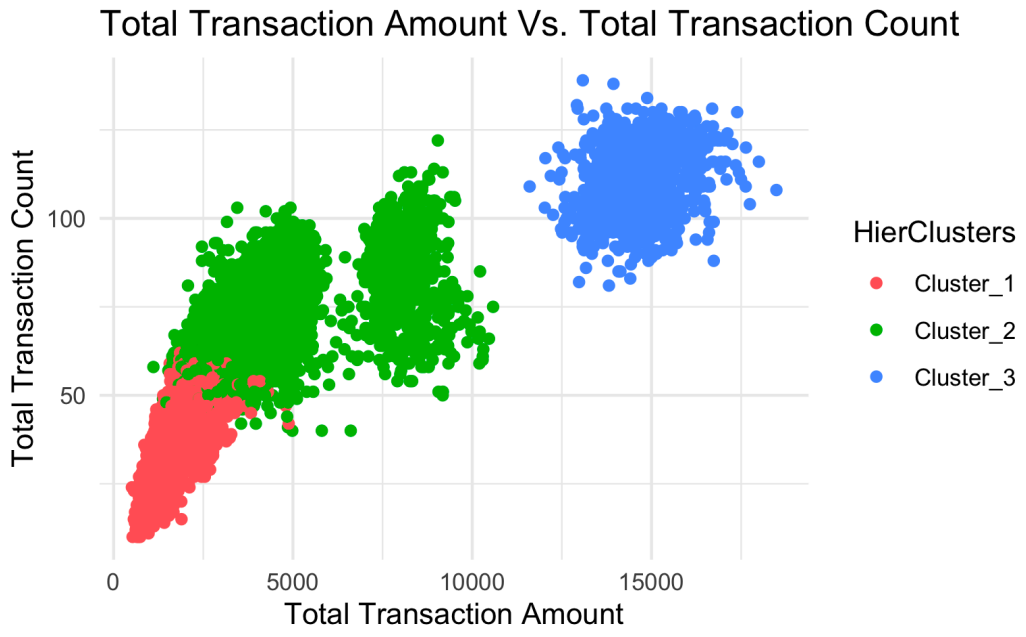
Hierarchical Clustering Dendrogram:



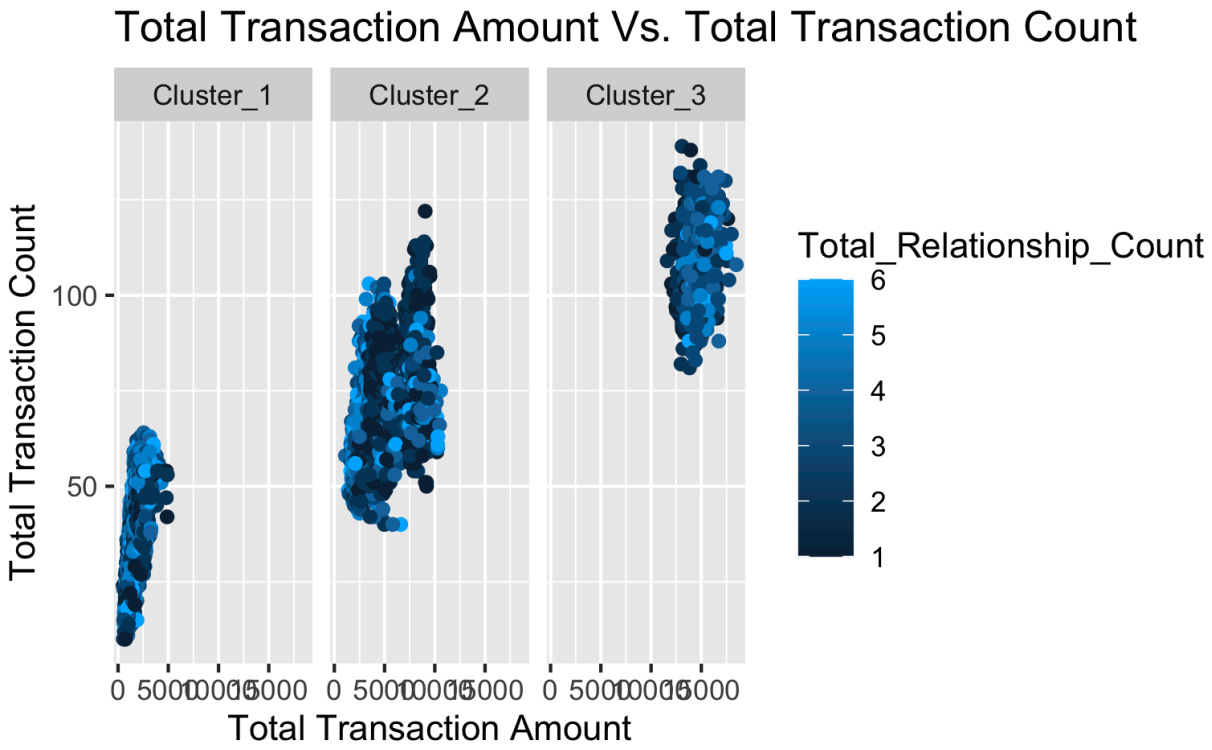
When using hierarchical clustering with 4 clusters, clusters 2 and 3 converge. This is problematic because relative to the configuration of our kmeans analysis comparing total transaction amount vs. total transaction count, it is difficult to extract any unique characteristics or insights from these clusters.



With 3 clusters, we obtain much different results than our kmeans analysis, with a better-separated cluster 3, and non-converging clusters 1 and 2.

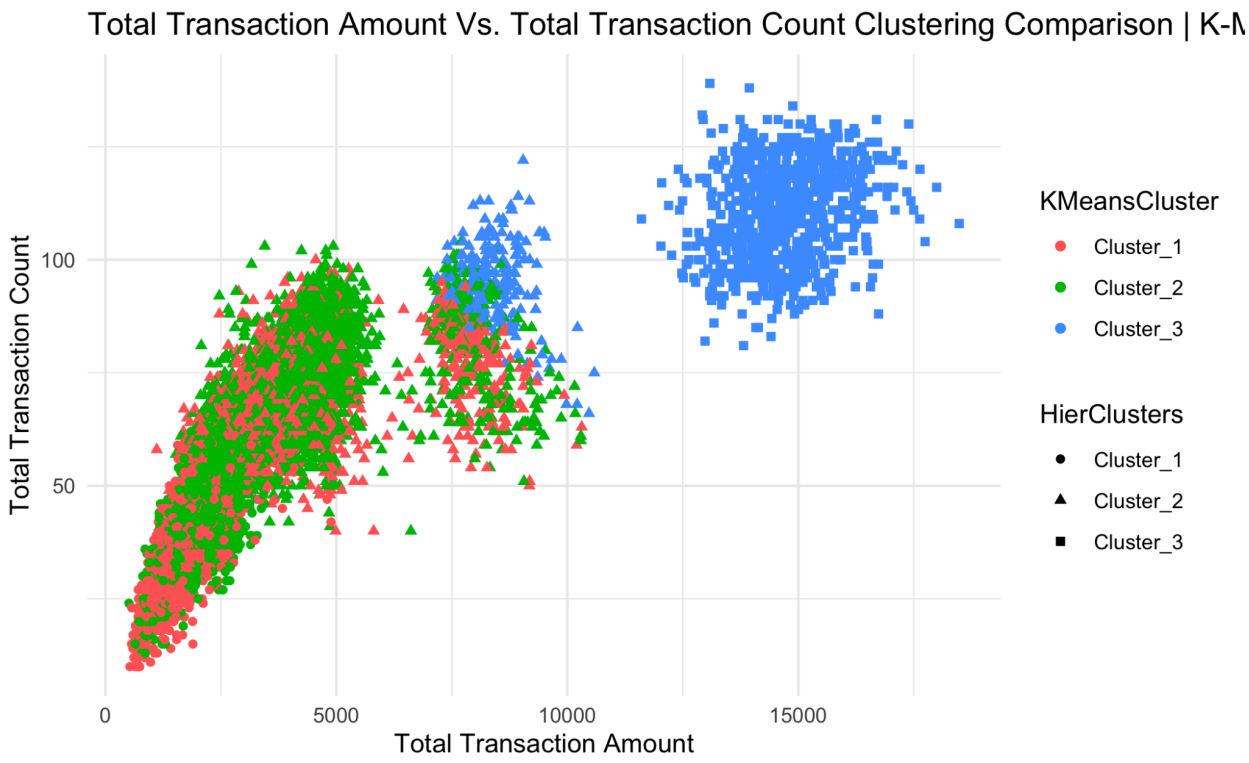


Of these clusters, we find that in cluster 3 of high transaction count and high transaction amount, these individuals tend to have had more relationships.



Based on a confucian matrix, we find that kmeans and hierarchical clustering differ the most when it comes to cluster 1, followed by cluster 2.

	Truth		
Prediction	Cluster_1	Cluster_2	Cluster_3
Cluster_1	1800	1613	0
Cluster_2	1804	3955	208
Cluster_3	0	0	747



Part III

A credit card company has data about its various customers. Here we use various facts such as their transaction count, transaction amount, education level, income level and revolving balance etc. The goal is to create clusters that are similar within itself, and dissimilar from others.

K-Means clustering is a type of flat clustering, chooses k centroids, groups the data points and then repositions the centroids until convergence is achieved. Banks can use K-Means clustering to divide customers into several clusters of different risk levels for risk management, then tailor their marketing strategies, product development, interest rates, and credit limits for each cluster. The method is simple and easy to implement, but it does not handle outliers well, and choosing the number of clusters in advance makes it not very flexible. It is also sensitive to initial centroid positions.

Unlike K-means clustering, for which the number of clusters needs to be specified beforehand, hierarchical clustering groups data points into a hierarchy of clusters based on their similarity or distance, and determines the optimal number of clusters by using a dendrogram. The banks can use the dendrogram to show the hierarchical relationship between the clusters and the sub-clusters, help the bank to identify the most risky and delinquent customers, as well as the most profitable and loyal ones. It apparently is not very suitable for larger datasets, though it can be combined with other clustering techniques like K-means.

DBSCAN is a clustering algorithm that uses density-based methods rather than distance-based clustering in K-Means and Hierarchical clustering; it works on the assumption that clusters are dense regions in space separated by regions of lower density. The clusters have a more non-linear shape with the given data and this method can be more used when the data is not linearly separable. DBSCAN can find clusters of arbitrary shapes and sizes, and is robust to outliers. unlike K-means assumes spherical clusters. However, it can be sensitive to the choice of distance metric and parameters, though it is less sensitive to initialization and does not require the number of clusters to be specified beforehand either.

In our experiment with the given data of credit card customers, it looks like K-means is the better option.

To increase credit card sales and revenue, personalization is necessary to deliver customized content and offers for customers based on their preferences, motivations, behavior and needs. Retention and satisfaction can be increased by a more engaging relationship with each customer. For example, the customers in the ‘full players’ cluster are active users who use credit cards most frequently for transactions or installment payments, normally with a good credit score, so banks can focus their marketing efforts on this cluster and offer more rewards. There are also customers who only use credit cards for installment purposes, and banks can offer zero interest installment programs to cater the product to their needs. For the students who just started to use credit cards, banks can design lower barrier products which help them to build their credit with waived fees, and devise marketing strategies to reduce their churn.