Optimal Kmeans Clustering for Market Segmentation on Airline Customers

Student ID: s0920609 Name (English): Jaeson Thendian Name (Chinese): 戴嘉盛

GitHub Link for Full Source Code & Explanation of Project:

https://github.com/JaesonT/Final-Project-ML/blob/main/Jaeson%20Final%20Project.ipynb

1. Introduction

In today's competitive business landscape, understanding customer behavior is essential for fostering loyalty and tailoring services effectively. Traditional marketing approaches often rely on a one-size-fits-all strategy, overlooking the diverse needs and values of different customer groups. With the rise of artificial intelligence (AI), particularly unsupervised learning techniques like clustering, businesses can now uncover hidden patterns in customer data to drive smarter, more strategic decisions.

For this project, we will be exploring customer segmentation using airline customer data to better understand how artificial intelligence can support modern marketing strategies. By leveraging AI techniques to group customers based on their behavior and characteristics, businesses can move away from one-size-fits-all approaches and instead create more personalized, data-driven campaigns. This not only improves customer engagement but also helps companies build stronger relationships and increase overall brand loyalty.

2. Problems

In the airline industry, customer retention and personalized marketing have become increasingly important due to rising competition and evolving consumer expectations. Airlines cater to a diverse customer base including frequent business travelers, premium passengers, occasional tourists, and budget-conscious flyers. Each of these segments has distinct preferences, behaviors, and value to the company. Yet, many airlines continue to rely on broad, undifferentiated marketing strategies that overlook these important differences.

This “one-size-fits-all” approach often leads to inefficient marketing efforts, reduced customer satisfaction, and missed opportunities to build long-term loyalty. For instance, high-value customers may receive promotions that do not reflect their needs, while less frequent travelers may feel disconnected due to irrelevant or poorly timed offers.

To address this challenge, airlines need to adopt more advanced methods of customer segmentation that go beyond traditional, rule-based approaches. The integration of artificial intelligence into marketing strategies enables companies to analyze large volumes of customer data and uncover hidden patterns. By doing so, airlines can better understand their audience, personalize communications, and make more informed, data-driven decisions.

3. Data Requirements:

My study is going to be based on the RFM model (Recency, Frequency, Monetary). Using this model, I will need datasets that collects transactional level information that captures customer behaviour over time. But aside from looking at recenct, frequency and monetary, to make the segmentation of customers clearer, we can also take measurements like tenure and the flight industry specific features. From data set we can use the purchase history of customers, how many flight the customer have taken, and amount of money customer spent, in order to identify the measurements of the RFM model.

4. Data Acquisition:

The dataset will be taken from Kaggle and it is titled “flight.csv”. This dataset can be download as a CSV file from the link written below:

https://www.kaggle.com/datasets/edisugiarto/unclean-airline-customer-dataset

5. Dataset Description (EDA)

This Dataset has 62,988 lines and 23 columns, in which is divided into 4 data types. The divided data are, 4 datetime, 10 numeric integers, 5 numeric float, 4 categorical, totaling to 23 columns. The dataset contains missing values in several columns, including GENDER (3), WORK_CITY (2,936), WORK_PROVINCE (3,248), WORK_COUNTRY (26), AGE (420), SUM_YR_1 (551), SUM_YR_2 (138), and LAST_FLIGHT_DATE (421). Descriptive Statistics shows that the average customer is 42 years old, takes around 12 flights, and travels about 17,000 km. Income and points are highly skewed, with a few customers earning exceptionally high amounts and accumulating large point balances. Most customers are male and based in China.

Data Distribution

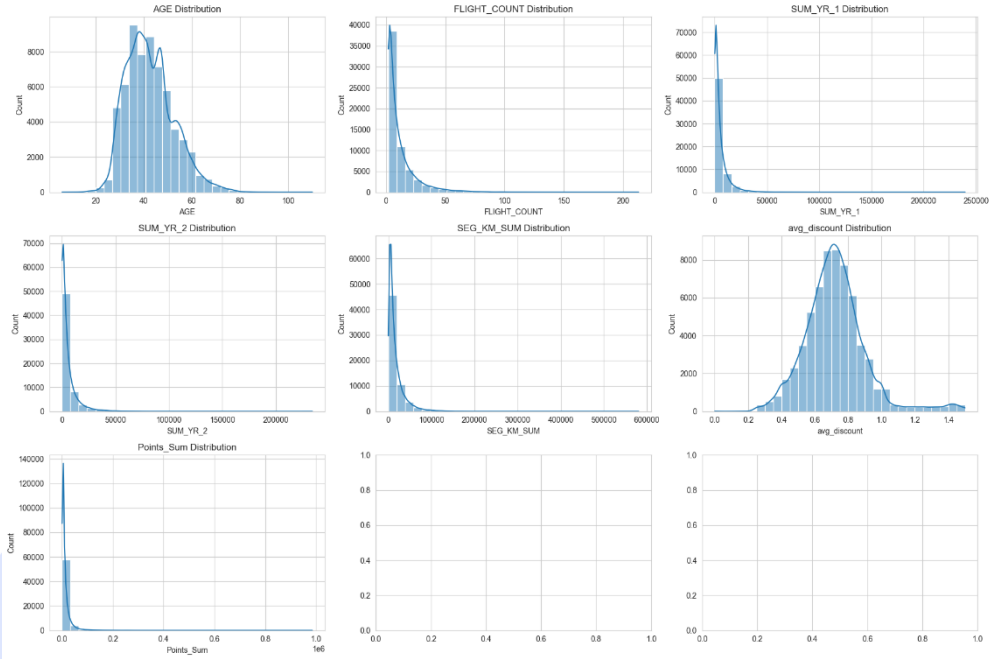

Figure 1. Distribution of numerical variables including AGE, FLIGHT_COUNT, SUM_YR_1, SUM_YR_2, SEG_KM_SUM, avg_discount, and Points_Sum.

The dataset shows a mix of normal and skewed distributions. Age is roughly normal but includes outliers. Flight count, distance traveled, income, and points are right-skewed, with most values concentrated at the lower end and a few extreme outliers. Average discount centers around 70%, but some values exceed 100%, indicating anomalies.

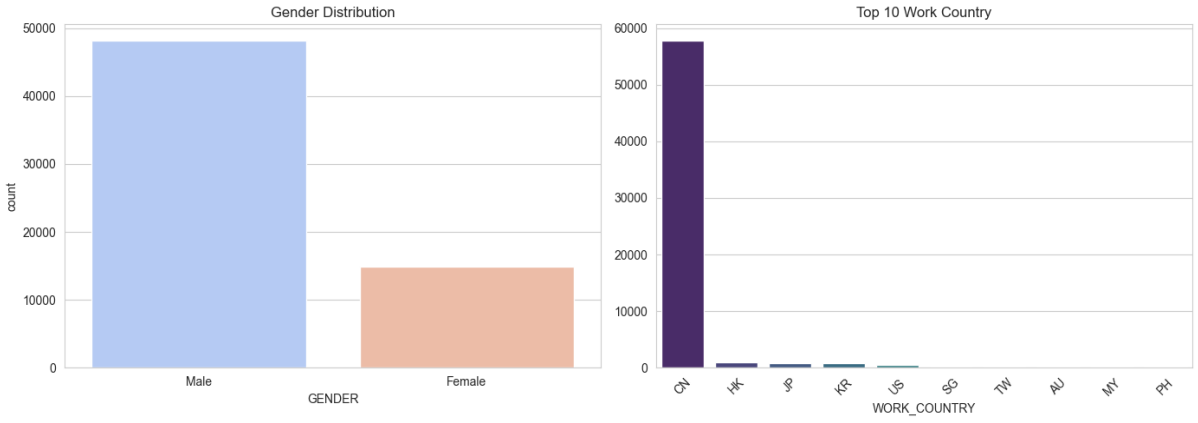

Figure 2. Distribution of categorical variables including GENDER and the top 10 WORK_COUNTRY values.

From the figure shown above it shows that the Majority of gender is male (~76%) and most customers are from china, followed by hongkong, Japan, South Korea and the US.

Corelation Analysis

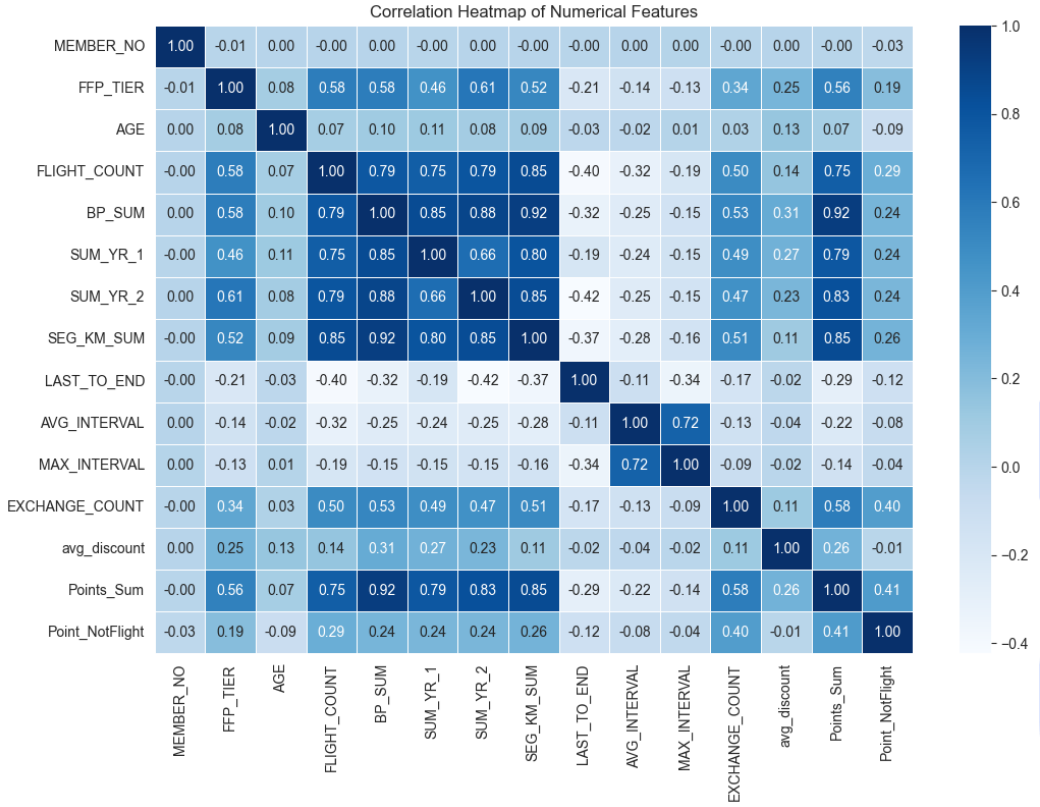

Figure 3. Corelation Heatmap of Numerical Features

Some Interesting revelations from the heatmap, shows that SUM_YR_1 and SUM_YR_2 are highly correlated, while SEG_KM_SUM and FLIGHT_COUNT also share a strong positive relationship. Points_Sum correlates with both flight frequency and distance, indicating that frequent flyers earn more points. Meanwhile, AGE has little to no correlation with other features, suggesting minimal influence on flight behavior.

EDA Conclution

In conclution, this dataset contains some missing values. There is also outliers in this data which are AGE, SUM_YR_1, SUM_YR_2, and SEG_KM_SUM and need to be addressed. Demographically, most customers are from china and and there are significantly more male customers. Finally, Customers with high flight counts and long travel distances tend to have higher income and point balances.

6. Data Preprocessing and Feature Engineering.

Amount of data points at the beginning (62988, 23).

AGE

There is 420 data row that has missing values in the column 'AGE'. Because the amount is very small compared to the total data, we can discard AGE data with missing value.

Tenure (Additional Feature from RFM)

For tenure, we calculate the number of days between each customer's FFP_DATE (membership start date) and LOAD_TIME (data extraction date), storing the result as a new column to represent customer loyalty duration.

Recency

For Recency, we can get from the 'LAST_TO_END' feature

Frequency

For Frequency, we can take from 'FLIGHT_COUNT' feature

Monetary

For Monetary, I take from the total revenue from each customer from the feature 'SUM_YR_1' and 'SUM_YR_2'

Handling Work Location Data

```
df['WORK_COUNTRY'].unique()
0.0s
array(['CN', 'US', 'FR', 'JP', 'HK', 'MY', 'AU', 'NL', 'MX', 'CA', 'KR',
       'PK', 'TH', 'SG', 'BR', 'PE', 'AE', 'HN', 'PH', 'ID', 'MO', 'GN',
       'FL', 'KH', 'TW', 'VN', 'DE', 'IL', 'CH', 'PA', 'VT', 'EC', 'NG',
       'BI', 'GT', 'BE', 'GL', 'NZ', 'FA', 'NI', 'KO', 'AN', 'CC', 'UV',
       'MN', 'UK', 'AS', 'IT', 'OM', 'SE', 'BZ', 'IN', 'KZ', 'PT', 'NO',
       'MS', 'FI', 'UZ', 'NE', 'AT', 'AA', 'HU', 'CY', 'EG', 'MA', '中',
       'ES', 'MZ', 'IR', 'SW', 'RE', 'PS', 'KE', 'SN', 'BB', 'AZ', 'UA',
       'LB', 'YE', 'GE', 'SI', 'KA', 'MV', 'RO', 'ZA', 'DK', 'AB', 'CD',
       'UY', 'CM', 'TK', 'AR', 'BN', 'PN', 'BD', '北', 'SU', 'RU', '沈',
       'GU', 'GB', 'ZW', 'PP', nan, '芬', 'FO', 'IE', 'PR', 'BS', 'GH',
       'SK', 'cn', 'TR', 'LA', 'BG', 'CL', 'CR', 'SA'], dtype=object)
```

There are some problematic country codes which are: cn, FL, FA, KO, UV, MN, AS, RE, BB, AB, TK, PN, SU, GU, PP, FO, IA, 沈, 北, 芬, 中, nan. Therefore, these need to be handled one by one. To handle this I switch the chinese letter to english letters and standardized capital and lowercase characters.

Choosing the Features

```
df.shape # Checking total data after cleaning
[36]  0.0s
...   (61860, 28)
```

Post Cleaning the data, we have 61860 Rows & 28 columns

For the features, I am selecting are, AVG_INTERVAL, tenure, recency, monetary, frequency, and SEG_KM_SUM because they represent key aspects of customer behavior, such as flight regularity, membership duration, recency of activity, spending, engagement frequency, and travel distance, making them ideal for clustering and segmentation.

Outlier Treatment

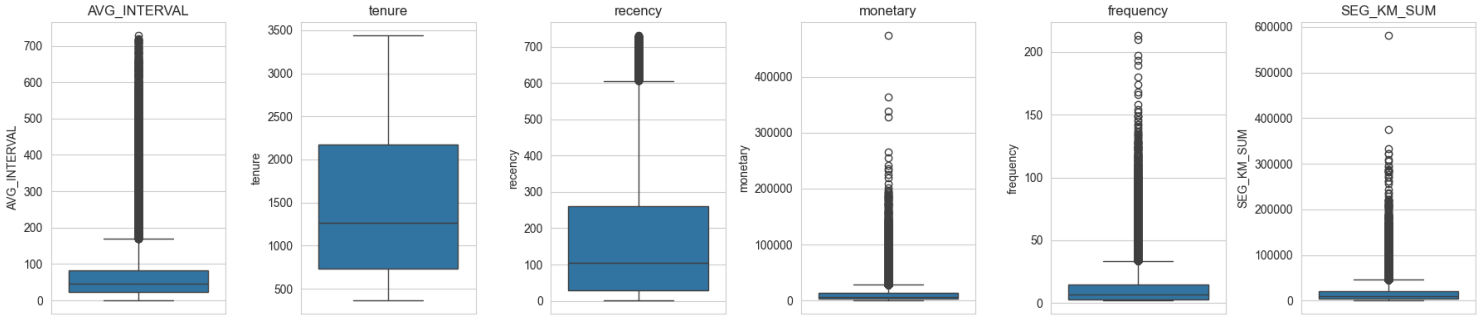

Figure 4. Boxplot of chosen feature pretreatment

Boxplot of the selected feature shown in figure 4, has visible extreme values, expecially in monetary, frequency and SEG_KM_SUM. These outliers represent high acitivity customers that could distort clustering if left unprocessed.

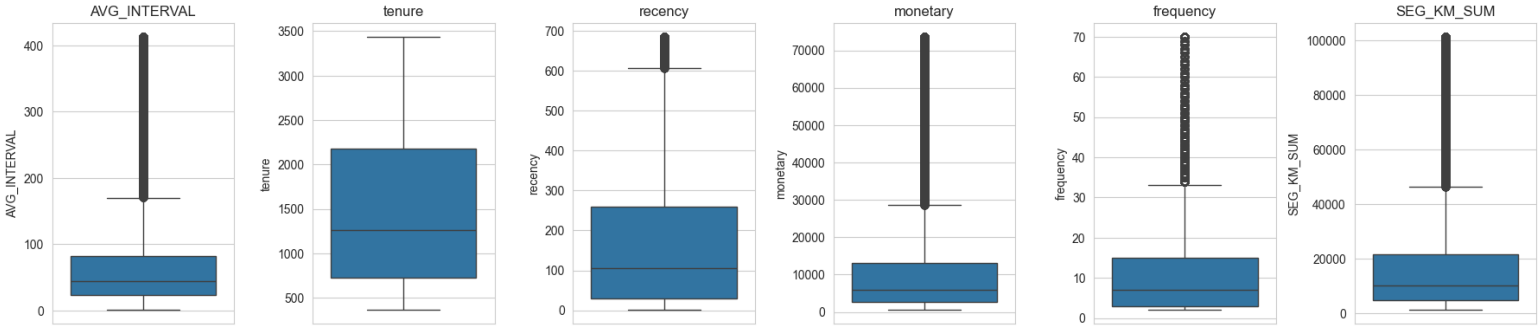

Figure 5. Boxplot of chosen feature after treatment

For the boxplot after treatment, it shows the values that has been compressed instead on removed. This reduces distortion while preserving data patterns. Hence, it will allow the model to better distinguish segements without removing important data from outliers.

Standardization

Since in this dataset the outliers represent important customer groups (e.g., long travel distance, high interval, high frequency, and high revenue), and we want these outliers to form their own clusters, hence MinMaxScaler is used instead of StandardScaler.

7. Apply Machine Learning

I have chosen to use the Kmeans clustering method for this project unsupervised machine learning. In order to find out the best optimal clusters for our data. I will be using two methods to determine, which is the elbow method and silhouette score.

Elbow Method

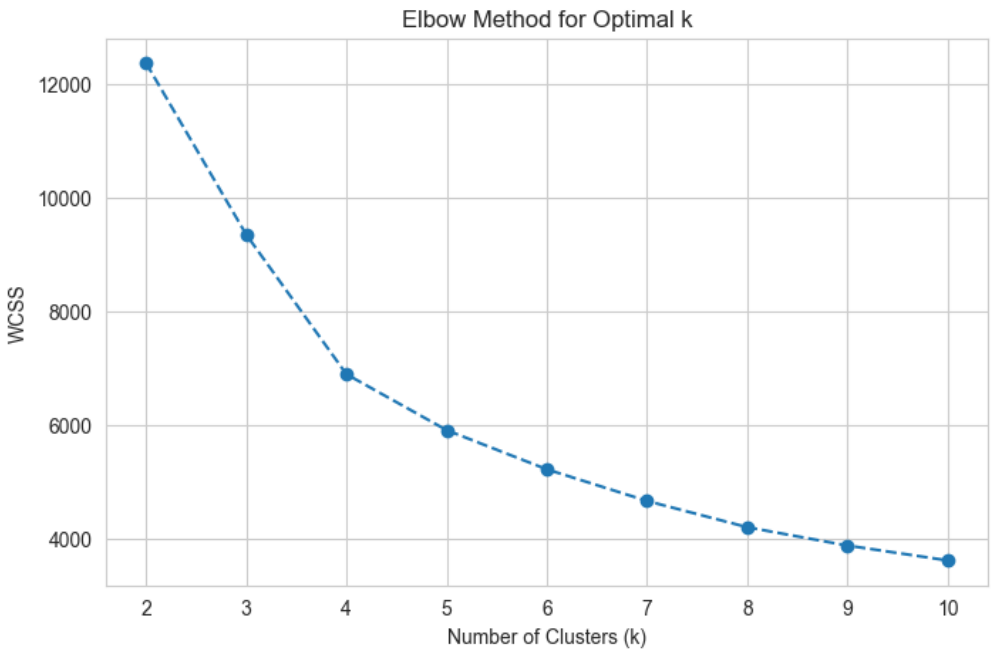

Figure 6. Graph of Elbow Method

Using the elbow method, we look for the hard bend representing the elbow. Anf from the above figure we can see the hard bend on 4 clusters. Hence, the elbow method has determined that 4 clusters is optimal for this data.

Silhouette Score

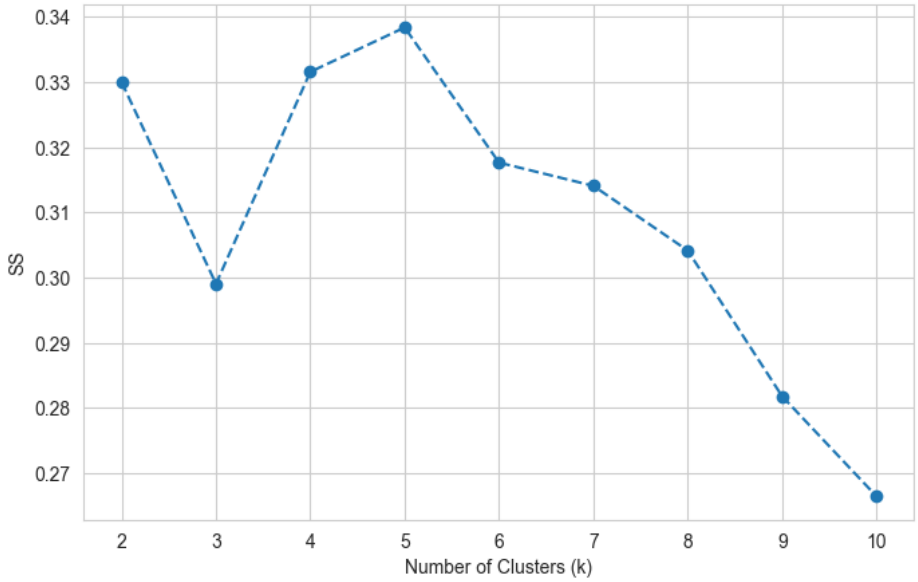

Figure 7. Graph of Silhouette Score

From the graph, the highest silhouette score occurs at k = 5, which suggests that 5 is the optimal number of clusters for this dataset.

Since the elbow method and Silhouette Score gives different results (4,5 optimal clusters). I will check for both 4,5 clusters in the next section.

PCA Analysis

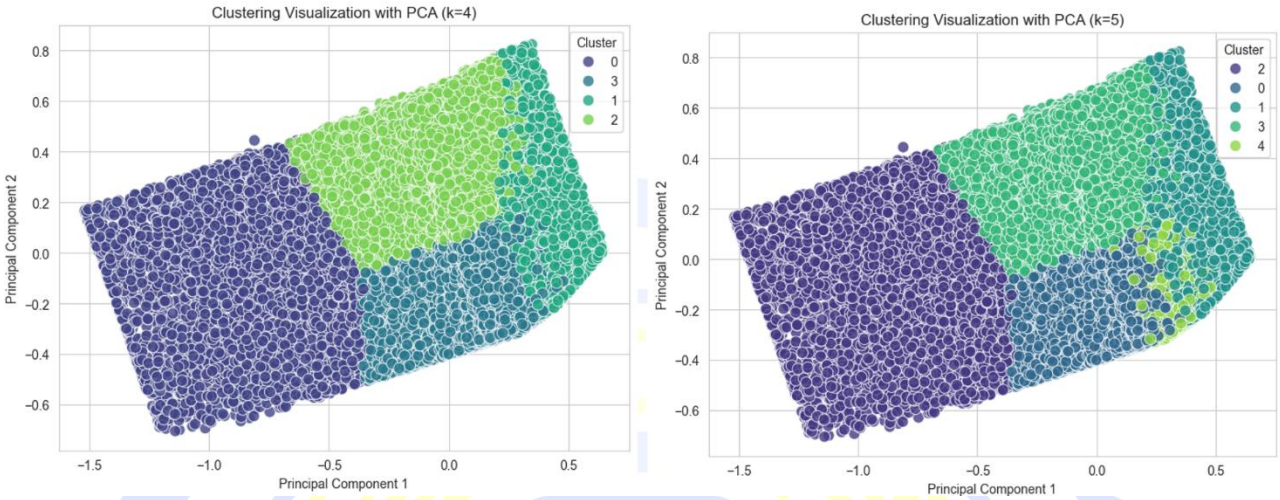

Figure 8. PCA analysis for K=4,5 Kmeans Clustering

The PCA visualizations for k = 4 and k = 5 show clear cluster separation in both cases, indicating that KMeans has effectively segmented the data. However, the plot for k = 5 reveals an additional distinct cluster (Cluster 4 in green), suggesting a more granular segmentation of customer behavior. This fifth cluster captures a small but meaningful group that is not distinguishable in the k = 4 configuration. Combined with the higher silhouette score at k = 5, this supports the conclusion that 5 clusters offer a better balance between segmentation quality and interpretability.

8. Analyze & Interpret Results

Results & Interpretation

	tenure				recency				frequency				monetary			
	mean	median	std	count	mean	median	std	count	mean	median	std	count	mean	median	std	count
K_Cluster																
0	893.748902	842.0	372.272011	22091	97.423747	76.0	81.705985	22091	10.366575	9.0	6.685950	22091	8786.057037	7200.0	6309.368940	22091
1	1193.967326	991.5	670.141982	12242	480.620732	473.0	117.818885	12242	4.003839	3.0	3.063775	12242	3704.414253	2640.0	3498.134637	12242
2	1875.342762	1851.0	833.956030	6611	31.262895	14.0	46.680614	6611	42.962789	38.0	20.581107	6611	43692.151112	35961.0	26460.279701	6611
3	2465.498442	2426.0	453.807687	16686	113.718207	84.0	104.227537	16686	9.910524	8.0	7.030212	16686	8576.208139	6816.0	6484.535957	16686
4	913.426005	848.0	375.857309	4230	127.006619	115.0	89.273553	4230	3.001418	3.0	0.940408	4230	2708.251300	2302.5	1941.958033	4230

1. Cluster 0 (Active-moderate Spenders)

- In this cluster, Customers have short tenure, high recency, high frequency and moderate spending. Hence, this group is relatively new but enganged to the company. They are valuable growing customers that needs nutturing to be potential loyal customers.

2. Cluster 1 (Lapsed Low-Spenders)

- This Cluster includes customers that are moderately tenure, very high recency, low frequency, low monetary. These are diengaged customers with their last flight being very old. They havent intracted recently and spend less money. They are likely to be dormant and not worth to rentent unless strategically targeted.

3. Cluster 2 (High Value Flyers)

- In this cluster, it shows the best customers with high tenure, low recency, very high frequency and very high monetary. These customers are your champions; they are the most valuable client the company has. They are active loyal and spend significantly. These groups should be your VIP and be a priority.

4. Cluster 3 (Loyal Customers)

- In this cluster, it shows loyal customers with longest tenure, recent flyers, moderate frequency and spending. This cluster include matured and steady customers; they have been purchasing tickets for a long time and a stable rate. They are regulars that respond well to regular engagement.

5. Cluster 4 (At risk, Low spenders)

- This cluster shows sign of early disengagement, they have moderate tenure, high recency and low frequency and low monetary. With relatively short tenure and low value, they may be at risk of churn. Targeted retention offers or automated reactivation campaigns could help re-engage them.

Business Reccomendation

Based on the Kmeans clustering analysis these are the business recommendations to increase retention and maximizing value for better customer lifecycle.

1. **Nuture Cluster 0 (Active Moderate Spenders)**
   - Profile: Newer customers with high frequency and moderate spending.
   - Reccomended actions: Loyalty incetives, up sell promotions.
2. **Minimalize on Cluster 1 (Low Spenders)**
   - Profile: Customers with low frequency and spend, and very high recency. mostly inactive.
   - Reccomended actions: Use low-cost automated promotion techniques like email to try to win back with low risk.
3. **Prioritize Cluster 2 (High Spending Customers)**
   - Profile: Recent, frequent, and high-spending. your most profitable customers.
   - Reccomended actions: Give VIP treatment such as faster lanes or lounges.
4. **Be Steady of Cluster 3 (Loyal Customers)**
   - Profile: Reliable and stable customers
   - Reccomended actions: Maintain with regular promotions
5. **Last effort for Cluster 4 (At Risk Low Spending Customers)**
   - Profile: Low spend, low frequency, and inactive. Likely to Churn
   - Reccomended Actions: Last chance win back campaign.

In conclution, by doing machine learning for customer market segmentation we can personalize each customers needs and market promotion to them correctly. This will maximize engagement and profits as it will make marketing more efficient. Hopefully, for future projects, I could use different clustering methods such as DBSCAN, BIRCH, ETC. With different clustering method, I could compare each of them ans check for the most efficient and clear segmentation of customers. However, this project have sucseeded in finding the optimal cluster for the Kmeans Clustering technique for customer segmentation in the airline industry.

Self Study Week Project; Comparing Kmeans & Agglomerative Clustering Methods

This section is a continuation of the above project. As shown previously, I have done customer segmentation using Kmeans which have shown results and feasibility to be used on field. However, in the world of machine learning; there are a lot of other clustering techniques to be explored. Hence, to improve on this project, this section is dedicated to comparing between another clustering technique which is the agglomerative clustering. Here, I compare agglomerative and the previous Kmeans clustering's performance and suitability for the flight's dataset. By analyzing and comparing more clustering method, I hope to find the optimal clustering method for this specific dataset.

Finding Best Clusters VIA Silhouette Score

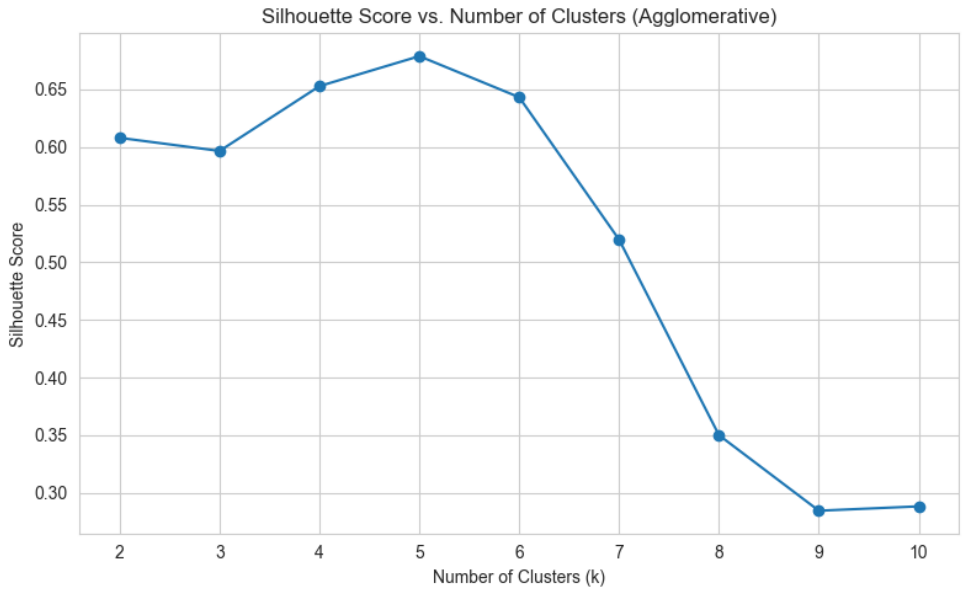

Figure 9. Silhouette Score for Agglomerative Clustering

The silhouette score above shows the best amount of cluster for agglomerative is 5 clusters, which is good because it shows simmilar optimal amount of clusters for both Kmeans and agglomerative indicating that 5 clusters is more likely to be the best amount.

PCA Results

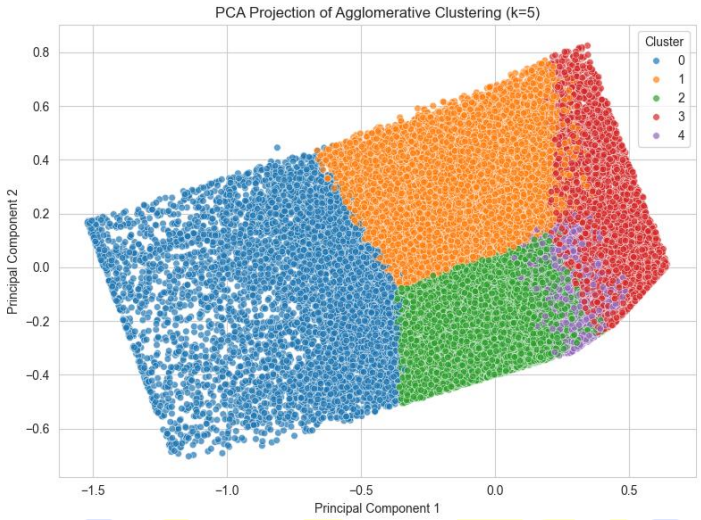

Figure 10. PCA analysis for Agglomerative Clustering

The PCA results for K=5 in agglomerative clustering shows clear cluster seperation, this indicates that 5 different group of customers have been sucsessfully and clearly segregated with the agglomerative clustering technique

Comparing PCA Results

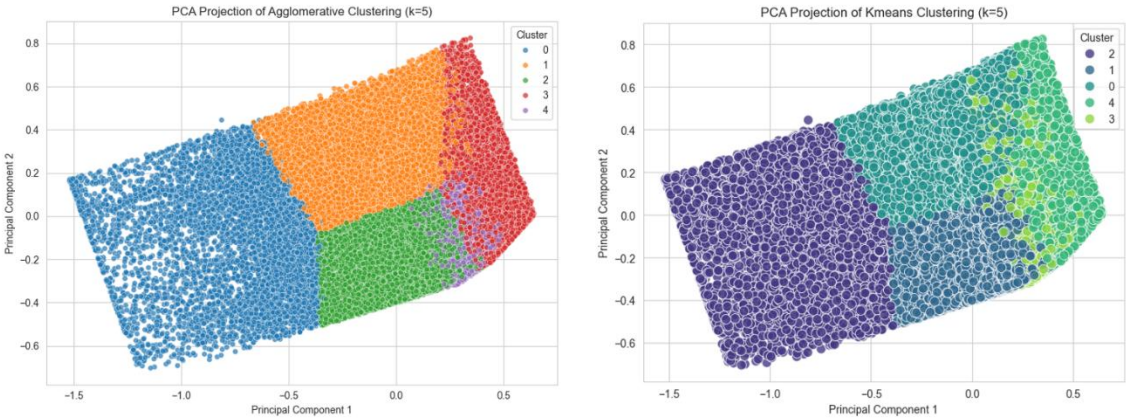

Figure 10. Side by Side PCA for Kmeans & Agglomerative K=5

The PCA visualizations comparing KMeans and Agglomerative Clustering (both with $k = 5$) reveal distinct differences in how each algorithm partitions the data. Agglomerative Clustering produces well-defined and evenly separated clusters, suggesting that its hierarchical approach effectively captures the underlying structure and dissimilarities within the dataset. The boundaries between clusters appear sharp and linear. In contrast, KMeans results in more gradual transitions between clusters, with some groups appearing less compact and more dispersed across the principal component space. This reflects KMeans' reliance on minimizing

Euclidean distances to centroids, which can struggle with non-spherical or uneven cluster shapes. Overall, Agglomerative Clustering demonstrates better performance in identifying natural groupings in this dataset, while KMeans, although computationally faster, appears less precise in capturing complex structures.

Results and Interpretation for Agglomerative Clustering

	tenure				recency				frequency				monetary			
	mean	median	std	count	mean	median	std	count	mean	median	std	count	mean	median	std	count
Cluster																
0	2490.516323	2455.0	447.547645	15653	111.635469	81.0	103.844865	15653	10.578292	9.0	7.271308	15653	9162.581422	7372.0	6754.401838	15653
1	901.186918	847.0	378.593718	23315	96.189964	75.0	81.160342	23315	10.584516	9.0	7.070977	23315	9010.697234	7253.0	6678.386840	23315
2	1907.653412	1907.0	837.338331	5935	29.589217	14.0	45.230810	5935	44.771693	40.0	20.869810	5935	45844.614153	37917.0	27015.541529	5935
3	1234.770740	1083.0	651.384845	4894	143.016551	126.0	106.081143	4894	2.803024	3.0	0.851792	4894	2568.574377	2176.0	1848.146496	4894
4	1193.834784	986.0	675.510922	12063	481.781646	477.0	118.410199	12063	4.005223	3.0	2.976343	12063	3704.291411	2664.0	3485.279192	12063

1. Cluster 0 (Loyal Customers)

- In this cluster, customers have the highest tenure with an average of 2490 days, which indicates loyalty. The recency is fairly low on average of 112 days, indicating they have flown somewhat recently, they also maintain moderate frequency of flights on average of 10.6. Finally, they contribute high monetary value of average 9163. This indicates that they are loyal and stable segment of customers.

2. Cluster 1 (Engaged Medium-Term Customers)

- This Cluster includes customers that are modetrantly tenured but still highly engaged. With an average tenure of about 901 days, they haven't been with the airline as long as Cluster 0, but their recent activity is comparable, with a recency average of 96 days. They also fly moderately frequently on average of 10.6 times and have a high average monetary contribution of 9,001. This group shows potential for loyalty.

3. Cluster 2 (High Value Frequest Flyers)

- In this cluster, it shows the most valuable customers the airline has. As they have high tenure avreage of 1907, low recency average of 30 days, high average frequency of 44.8. And given the highest monetary value of 45845. This makes the the top priority customers aka VIPS.

4. Cluster 3 (Low-Value Inactive Customers)

- In this cluster, it shows low value customers, with short tenure of 1235, high recency of 143 days and very low frequncy of 3. Additonally their monetary value is at the lowest at 2569. This indicates this segment of customers are slowly slipping away and might need re-engagement to put incentive on them.

5. Cluster 4 (Moderate, Low-Engagement Customers)

- This cluster has disengaged customers moderate tenure at 1194, but very high average recency of 482. Addtionally thier frequency flight is low at 4 and low monetary average of 3704. This segment of customers shows that they havent flown for a very long time and need a lot of efforts to re-engage them.

Business Recommendation

In order to maximize the customer segmentation, each segments need their own type of promotion to maximize engagement and profits. Here are some reccomended actions the company should take for each segement of customer.

6. **Nuture Cluster 0-1 (Stable Customers)**
   - Profile: Long-tenured, engaged customers with solid spending.
   - Reccomended actions: Maintain their engagement with seasonal offers.
7. **Focus on Cluster 2 (High spending Customers)**
   - Profile: Most active and profitable customers
   - Reccomended actions: Offer exclusive VIP rewards and elite benefits.
8. **Reactivate Cluster 4 (Dormant Customers)**
   - Profile: Inactive customers
   - Reccomended actions: Win-Back campaigns such as “miss you” promos
9. **Let go of Cluster 3 (Low Value Customers)**
   - Profile: Low spenders
   - Reccomended actions: Low-cost emails to try to re-engage or let go.