Introduction to the Dataset and Clustering Methods

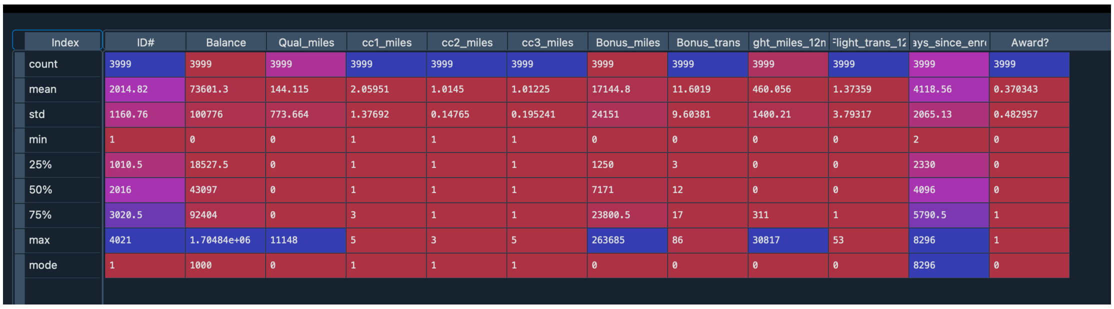
The dataset used in this analysis is from EastWest Airlines, containing customer-related features such as Balance, Qual_miles, Bonus_miles, and Flight_miles_12mo.

The goal is to segment customers into distinct groups using clustering techniques. Two clustering methods were applied: K-Means and Hierarchical Clustering.

Part 1.

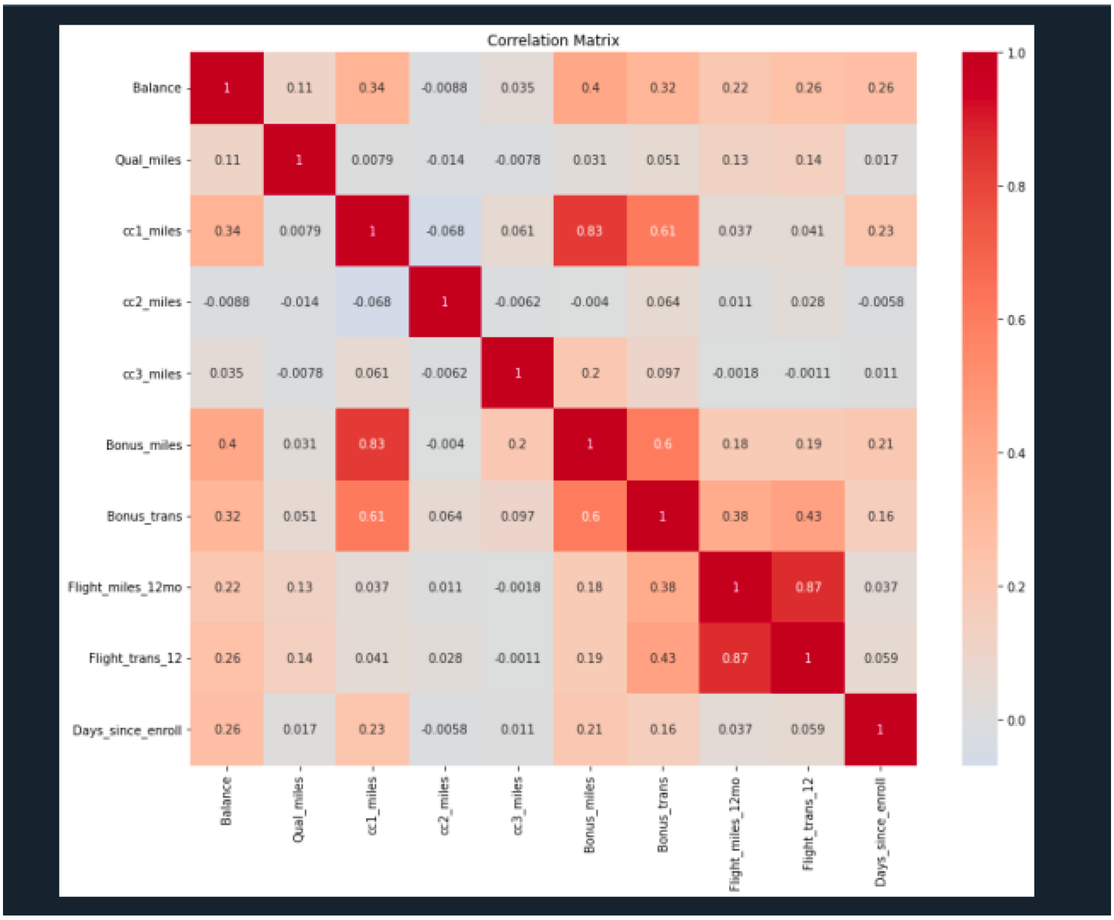
Data Processing

Descriptive statistics:



Index	ID#	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	ght_miles_12n	Flight_trans_12	ays_since_enr	Award?
count	3999	3999	3999	3999	3999	3999	3999	3999	3999	3999	3999	3999
mean	2014.82	73601.3	144.115	2.05951	1.0145	1.01225	17144.8	11.6019	460.056	1.37359	4118.56	0.370343
std	1160.76	100776	773.664	1.37692	0.14765	0.195241	24151	9.60381	1400.21	3.79317	2065.13	0.482957
min	1	0	0	1	1	1	0	0	0	0	2	0
25%	1010.5	18527.5	0	1	1	1	1250	3	0	0	2330	0
50%	2016	43097	0	1	1	1	7171	12	0	0	4096	0
75%	3020.5	92404	0	3	1	1	23800.5	17	311	1	5790.5	1
max	4021	1.70484e+06	11148	5	3	5	263685	86	30817	53	8296	1
mode	1	1000	0	1	1	1	0	0	0	0	8296	0

Correlation Matrix:

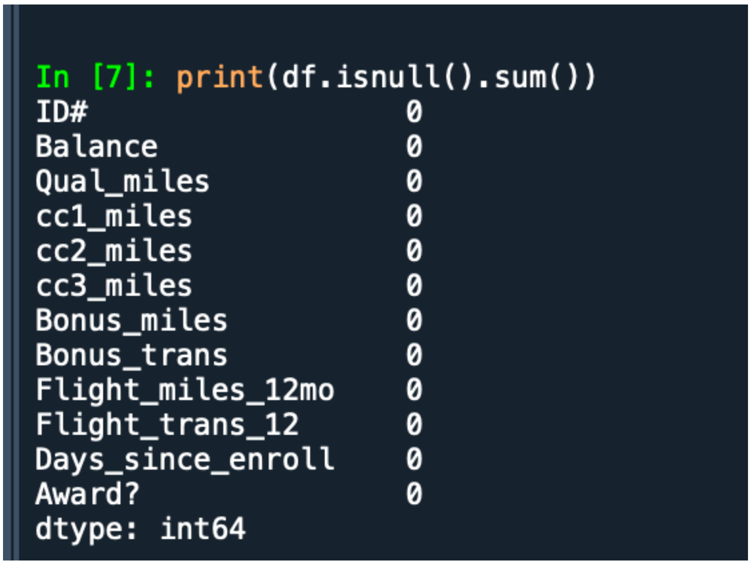


Correlation Matrix

	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll
Balance	1	0.11	0.34	-0.0088	0.035	0.4	0.32	0.22	0.26	0.26
Qual_miles	0.11	1	0.0079	-0.014	-0.0078	0.031	0.051	0.13	0.14	0.017
cc1_miles	0.34	0.0079	1	-0.068	0.061	0.83	0.61	0.037	0.041	0.23
cc2_miles	-0.0088	-0.014	-0.068	1	-0.0062	-0.004	0.064	0.011	0.028	-0.0058
cc3_miles	0.035	-0.0078	0.061	-0.0062	1	0.2	0.097	-0.0018	-0.0011	0.011
Bonus_miles	0.4	0.031	0.83	-0.004	0.2	1	0.6	0.18	0.19	0.21
Bonus_trans	0.32	0.051	0.61	0.064	0.097	0.6	1	0.38	0.43	0.16
Flight_miles_12mo	0.22	0.13	0.037	0.011	-0.0018	0.18	0.38	1	0.87	0.037
Flight_trans_12	0.26	0.14	0.041	0.028	-0.0011	0.19	0.43	0.87	1	0.059
Days_since_enroll	0.26	0.017	0.23	-0.0058	0.011	0.21	0.16	0.037	0.059	1

Missing Value:

There is no missing Value

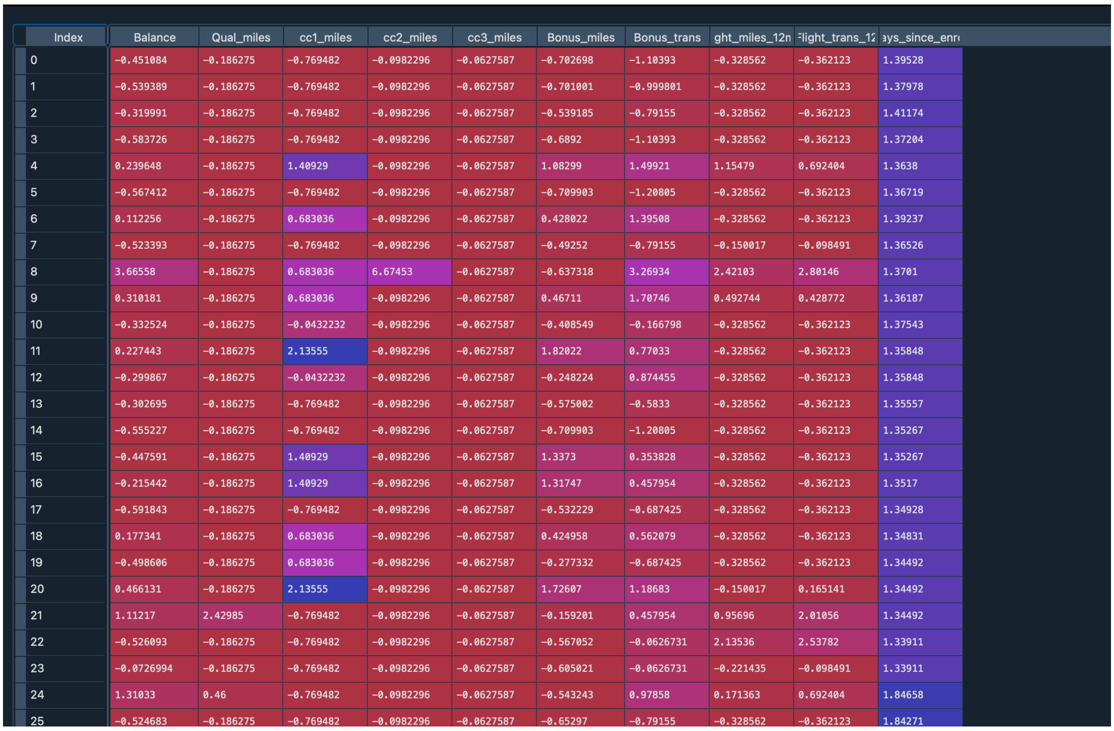


```
In [7]: print(df.isnull().sum())
ID#                  0
Balance              0
Qual_miles           0
cc1_miles            0
cc2_miles            0
cc3_miles            0
Bonus_miles          0
Bonus_trans          0
Flight_miles_12mo    0
Flight_trans_12      0
Days_since_enroll    0
Award?               0
dtype: int64
```

Convert Categorical Value to numerical:

There is no categorical variables

Normalize data:



Index	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	ght_miles_12n	light_trans_12	ays_since_enr
0	−0.451084	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.702698	−1.10393	−0.328562	−0.362123	1.39528
1	−0.539389	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.701001	−0.999801	−0.328562	−0.362123	1.37978
2	−0.319991	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.539185	−0.79155	−0.328562	−0.362123	1.41174
3	−0.583726	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.6892	−1.10393	−0.328562	−0.362123	1.37204
4	0.239648	−0.186275	1.40929	−0.0982296	−0.0627587	1.08299	1.49921	1.15479	0.692404	1.3638
5	−0.567412	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.709903	−1.20805	−0.328562	−0.362123	1.36719
6	0.112256	−0.186275	0.683036	−0.0982296	−0.0627587	0.428022	1.39508	−0.328562	−0.362123	1.39237
7	−0.523393	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.49252	−0.79155	−0.150017	−0.098491	1.36526
8	3.66558	−0.186275	0.683036	6.67453	−0.0627587	−0.637318	3.26934	2.42103	2.80146	1.3701
9	0.310181	−0.186275	0.683036	−0.0982296	−0.0627587	0.46711	1.70746	0.492744	0.428772	1.36187
10	−0.332524	−0.186275	−0.0432232	−0.0982296	−0.0627587	−0.408549	−0.166798	−0.328562	−0.362123	1.37543
11	0.227443	−0.186275	2.13555	−0.0982296	−0.0627587	1.82022	0.77033	−0.328562	−0.362123	1.35848
12	−0.299867	−0.186275	−0.0432232	−0.0982296	−0.0627587	−0.248224	0.874455	−0.328562	−0.362123	1.35848
13	−0.302695	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.575002	−0.5833	−0.328562	−0.362123	1.35557
14	−0.555227	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.709903	−1.20805	−0.328562	−0.362123	1.35267
15	−0.447591	−0.186275	1.40929	−0.0982296	−0.0627587	1.3373	0.353828	−0.328562	−0.362123	1.35267
16	−0.215442	−0.186275	1.40929	−0.0982296	−0.0627587	1.31747	0.457954	−0.328562	−0.362123	1.3517
17	−0.591843	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.532229	−0.687425	−0.328562	−0.362123	1.34928
18	0.177341	−0.186275	0.683036	−0.0982296	−0.0627587	0.424958	0.562079	−0.328562	−0.362123	1.34831
19	−0.498606	−0.186275	0.683036	−0.0982296	−0.0627587	−0.277332	−0.687425	−0.328562	−0.362123	1.34492
20	0.466131	−0.186275	2.13555	−0.0982296	−0.0627587	1.72607	1.18683	−0.150017	0.165141	1.34492
21	1.11217	2.42985	−0.769482	−0.0982296	−0.0627587	−0.159201	0.457954	0.95696	2.01056	1.34492
22	−0.526093	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.567052	−0.0626731	2.13536	2.53782	1.33911
23	−0.0726994	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.605021	−0.0626731	−0.221435	−0.098491	1.33911
24	1.31033	0.46	−0.769482	−0.0982296	−0.0627587	−0.543243	0.97858	0.171363	0.692404	1.84658
25	−0.524683	−0.186275	−0.769482	−0.0982296	−0.0627587	−0.65297	−0.79155	−0.328562	−0.362123	1.84271

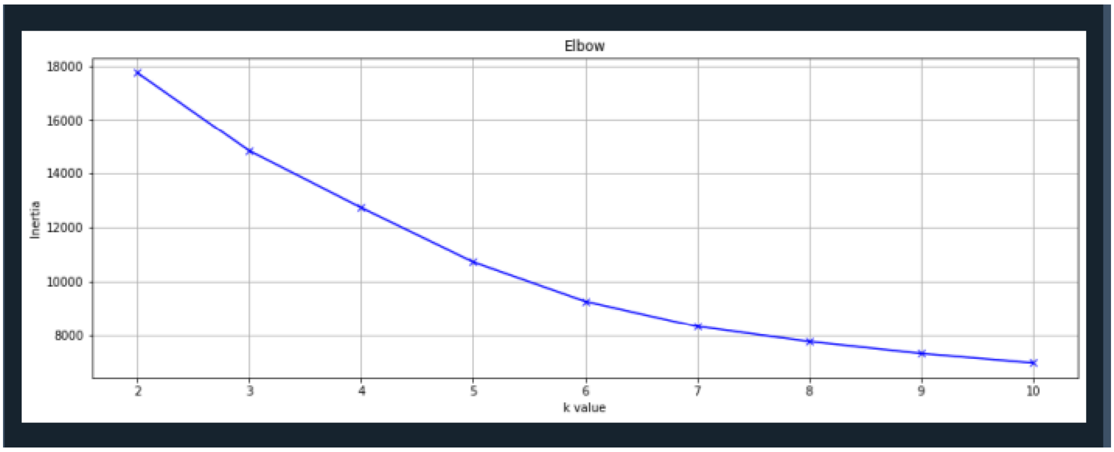
Part 2.

Clustering model building:

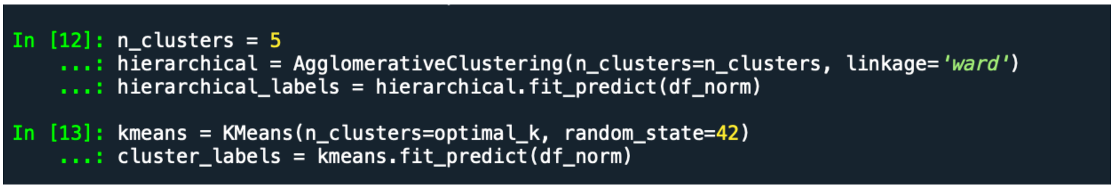
Selected feature:

Here I drop ID and Award at first, as ID doesn’t have any meanings award is kind of status which is not what we want in unsupervised learning. Flight_trans_12: Highly correlated (0.87) with Flight_miles_12mo so we drop it. Bonus_trans: Moderately correlated (0.6) with Bonus_miles, both of them seems talking about the same thing so we drop it. cc2_miles & cc3_miles: Very low correlation with all other variables (near 0) so we drop it.

Elbow graph to determine number of cluster:



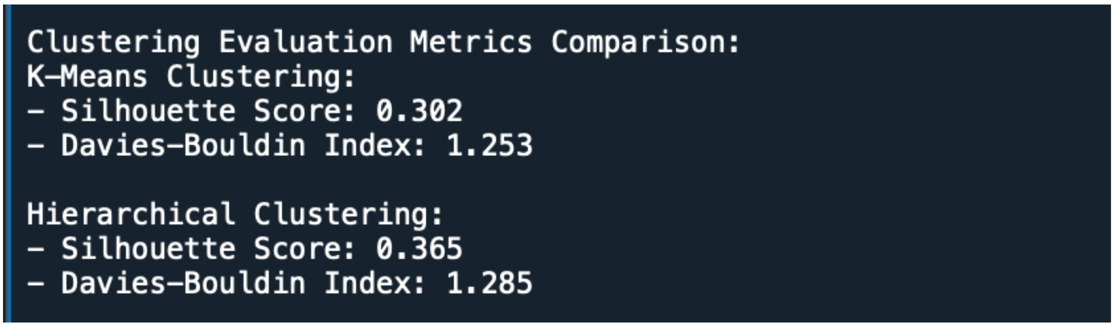
When K value around 5, the slope becomes smooth. So we choose to have 5 clusters.



```
In [12]: n_clusters = 5
    ...: hierarchical = AgglomerativeClustering(n_clusters=n_clusters, linkage='ward')
    ...: hierarchical_labels = hierarchical.fit_predict(df_norm)

In [13]: kmeans = KMeans(n_clusters=optimal_k, random_state=42)
    ...: cluster_labels = kmeans.fit_predict(df_norm)
```

Models are built as above.

Internal Evaluation:



```
Clustering Evaluation Metrics Comparison:
K-Means Clustering:
- Silhouette Score: 0.302
- Davies-Bouldin Index: 1.253

Hierarchical Clustering:
- Silhouette Score: 0.365
- Davies-Bouldin Index: 1.285
```

Silhouette Score:

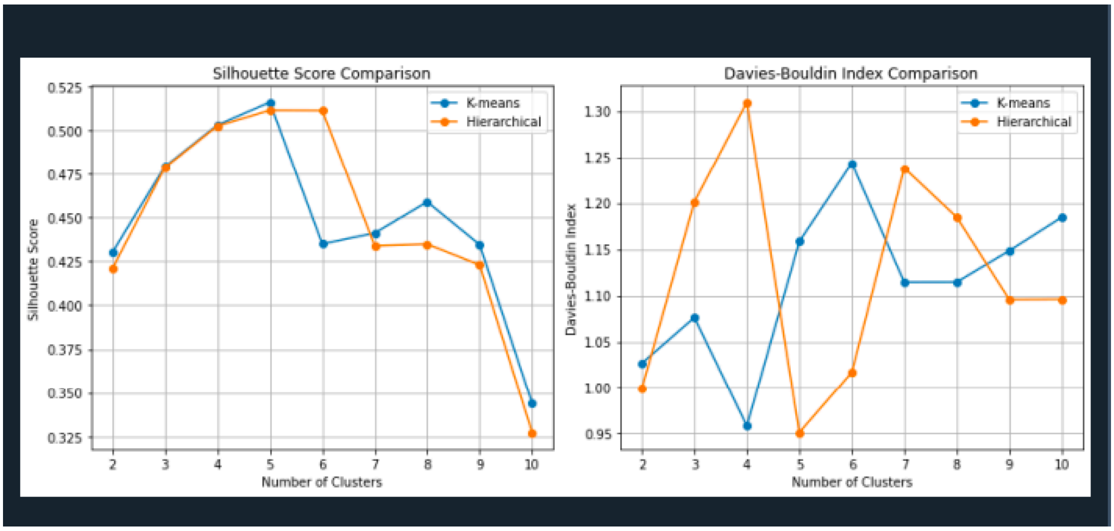
- K-Means: 0.302
- Hierarchical Clustering: 0.365
- Hierarchical Clustering has a slightly higher silhouette score, indicating better internal cluster cohesion.

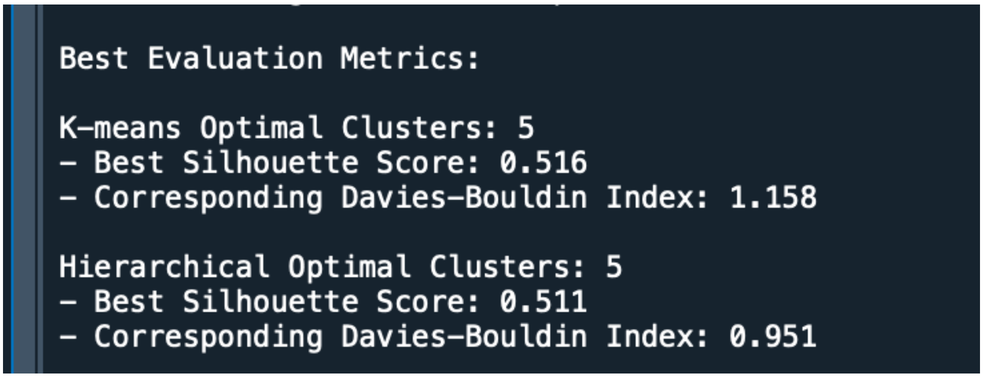
Davies-Bouldin Index:

- K-Means: 1.253
- Hierarchical Clustering: 1.285
- K-Means has a slightly lower Davies-Bouldin Index, indicating better separation between clusters.

Both methods doesn’t show an ideal model of this clustering as their silhouette scores are relatively low.

I will do one more evaluation to calculate the silhouette score and daviers-bouldin index.



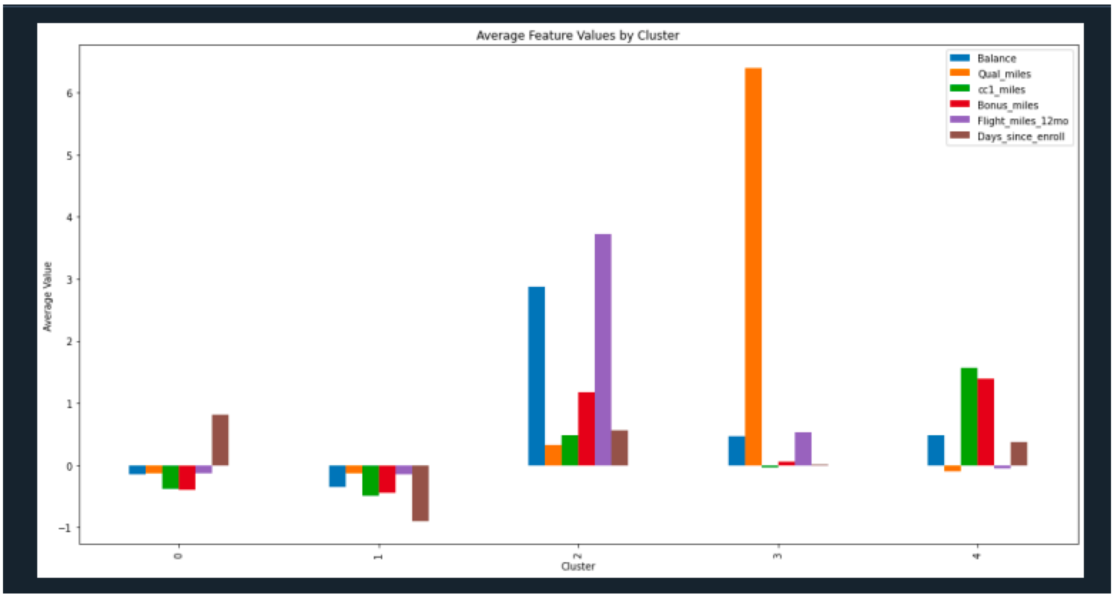


```
Best Evaluation Metrics:

K-means Optimal Clusters: 5
- Best Silhouette Score: 0.516
- Corresponding Davies-Bouldin Index: 1.158

Hierarchical Optimal Clusters: 5
- Best Silhouette Score: 0.511
- Corresponding Davies-Bouldin Index: 0.951
```

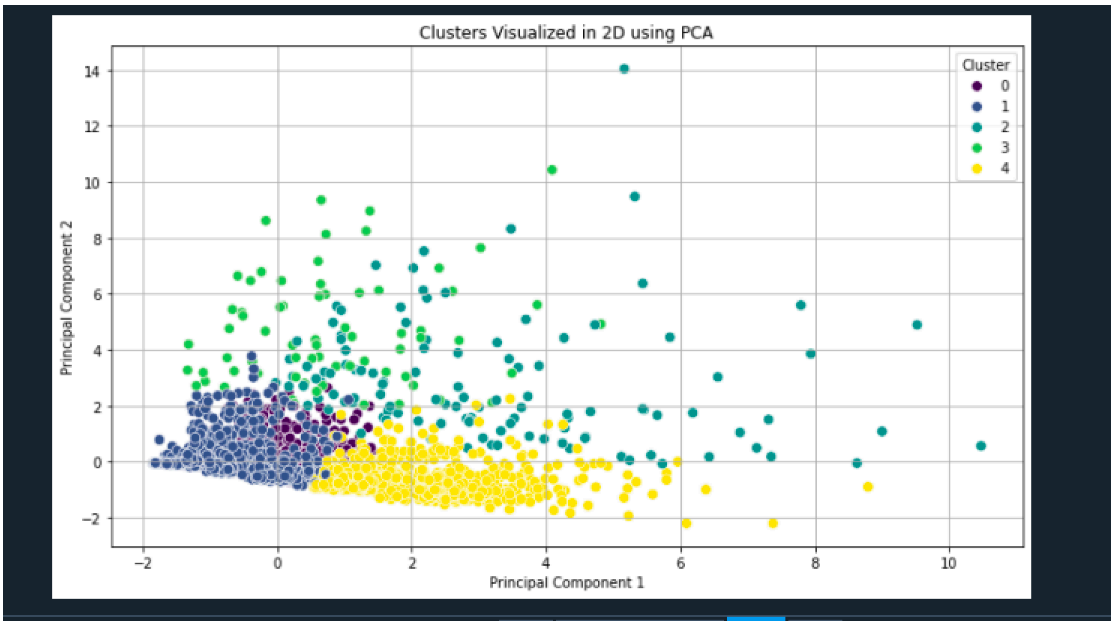
From the graph we can see that when k = 5, we have the highest silhouette score and davies-Bouldin index are close to 1. So here we will use 5 clusters to lable all the data and use K-means to build the final model as it has a larger silhouette score and less computation work.

Part3:

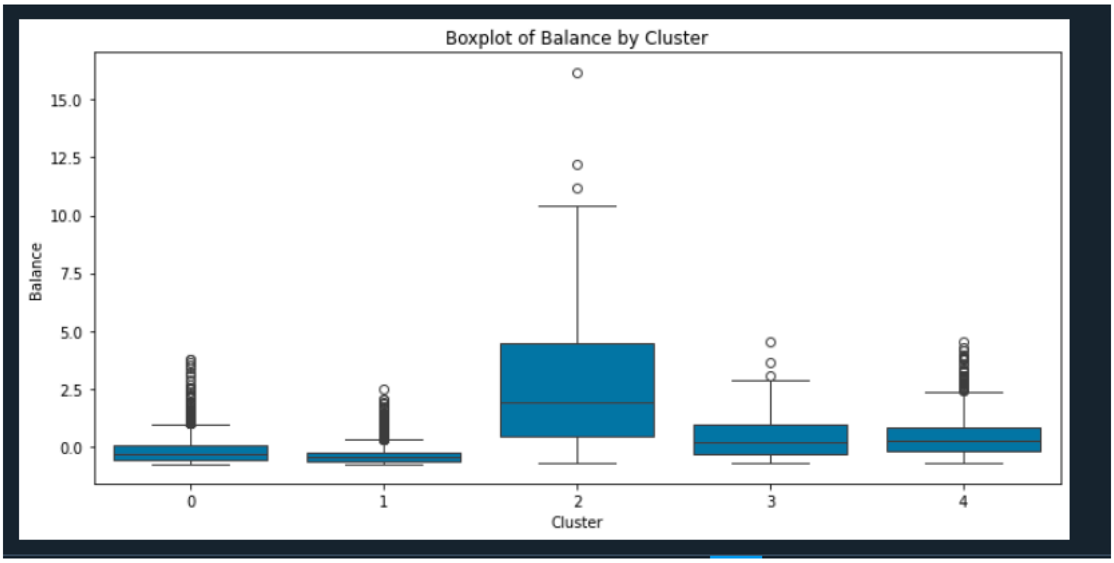
Model Diagnostics and Interpretation

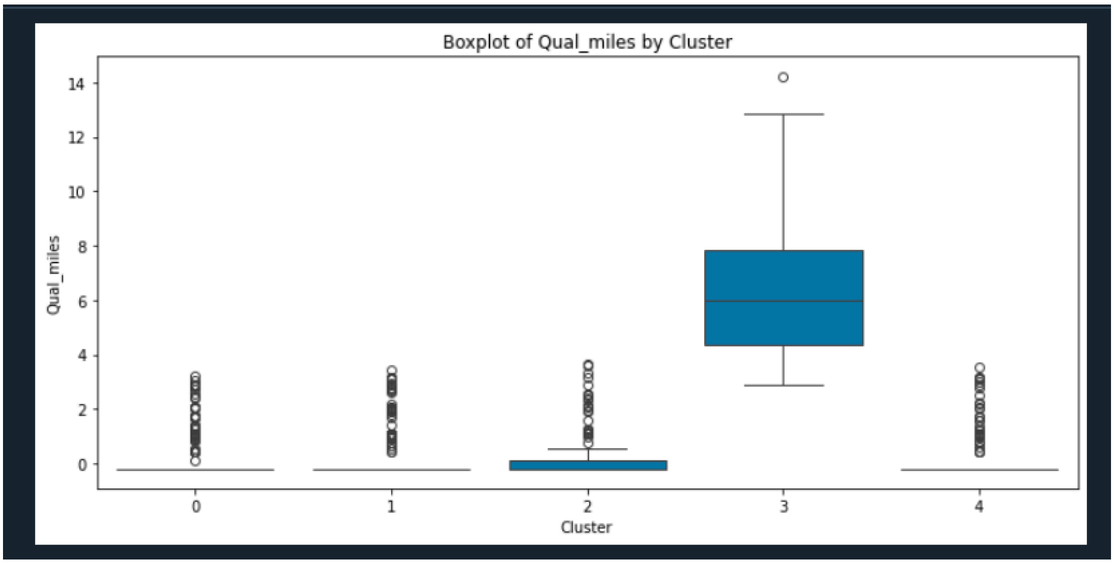


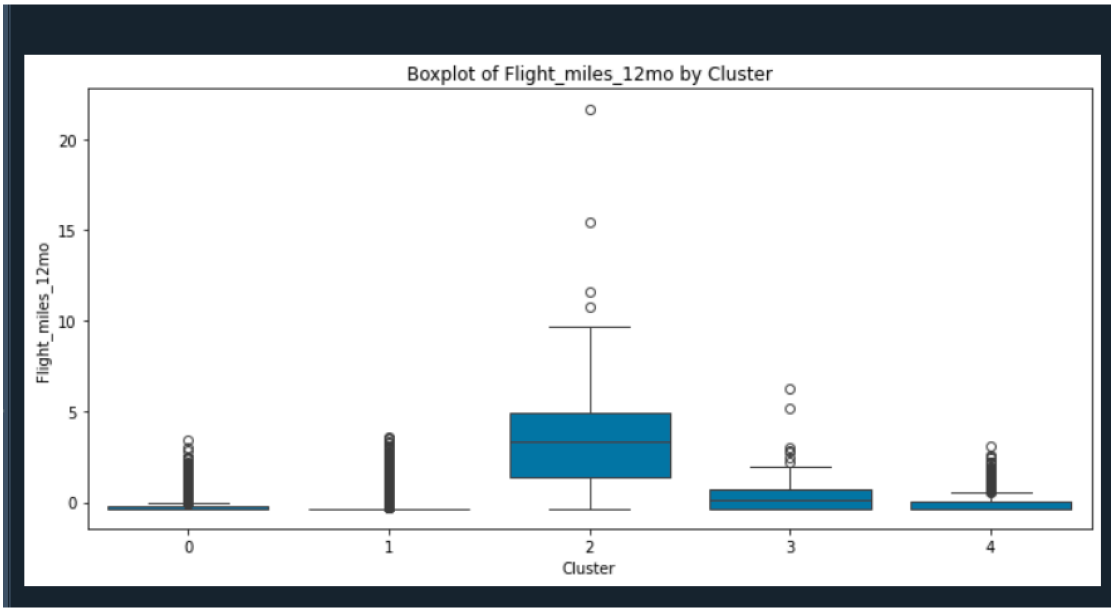
According to the average value per cluster plot we can tell that Cluster 2 has high values for most features, indicating highly engaged customers. Cluster 3 shows high values in 'Qual_miles', which may indicate frequent flyers. Cluster 0 and Cluster 1 have lower values overall, possibly representing less active customers.



The PCA scatter plot shows that Cluster 2 and Cluster 4 are more distinct, while Cluster 0 and Cluster 1 overlap significantly, indicating they have similar features.







Boxplots show differences in specific features across clusters. For example, Cluster 3 has high 'Qual_miles', and Cluster 2 has a higher 'Balance'. **Cluster 2** shows high values in Balance, Bonus_miles, and Flight_miles_12mo, indicating these customers are high-value, frequent travelers. The features Balance and Qual_miles are particularly effective at distinguishing Cluster 2 and Cluster 3 from other clusters, as these clusters show distinct distributions for these features.

Discussion of Implications and Future Improvements

The clustering results indicate distinct customer segments, such as high-value frequent travelers and less active customers. These insights can help the airline develop targeted marketing strategies.

However, some clusters, such as Cluster 0 and Cluster 1, show significant overlap, suggesting a need for further feature engineering or additional data to improve cluster separability.

Future analysis could explore incorporating more features, such as customer demographics or behavioral data, to better differentiate between clusters.