

4KG3 Data Mining Assignment 3 – Clustering

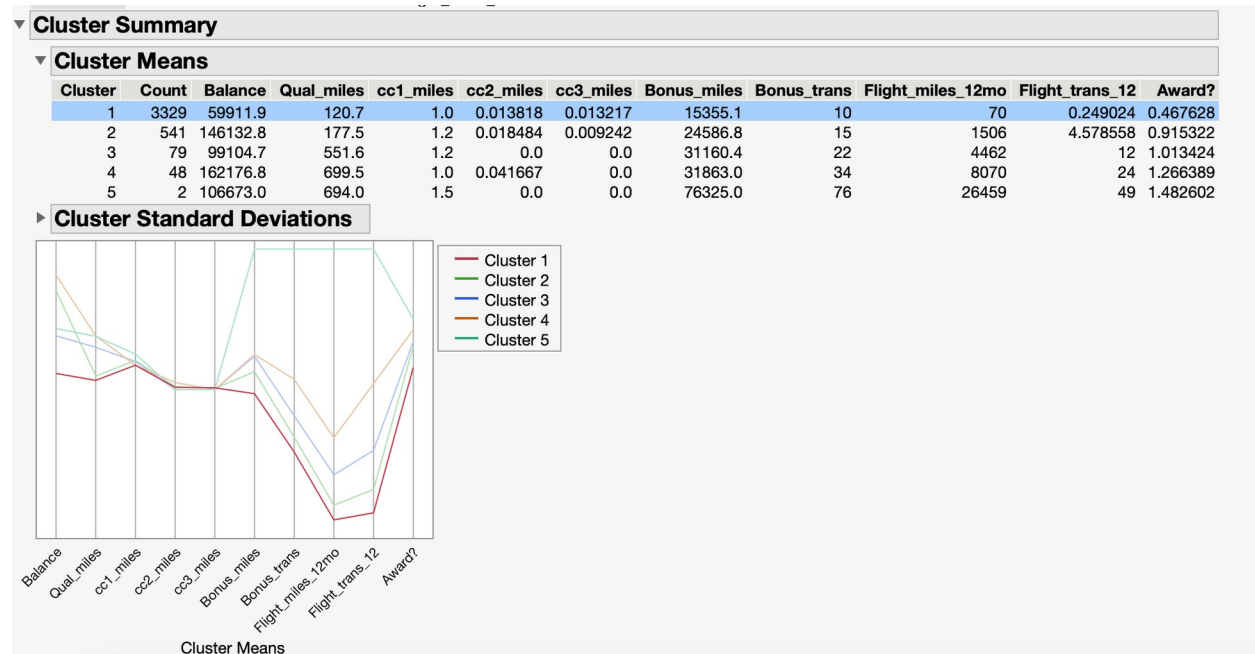
Student Name: Jieyu Zhang

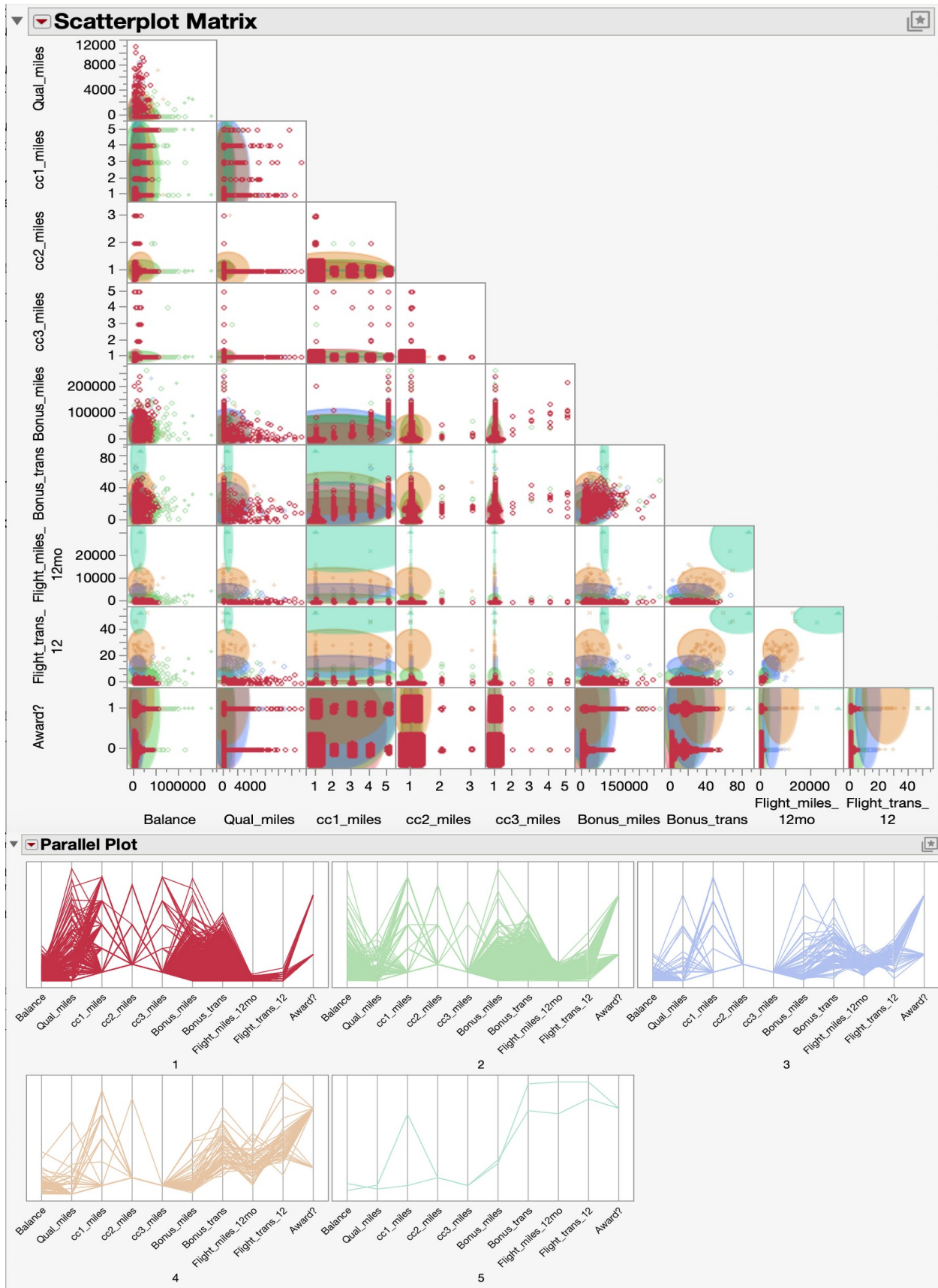
Student Number: 400368960

Tasks:

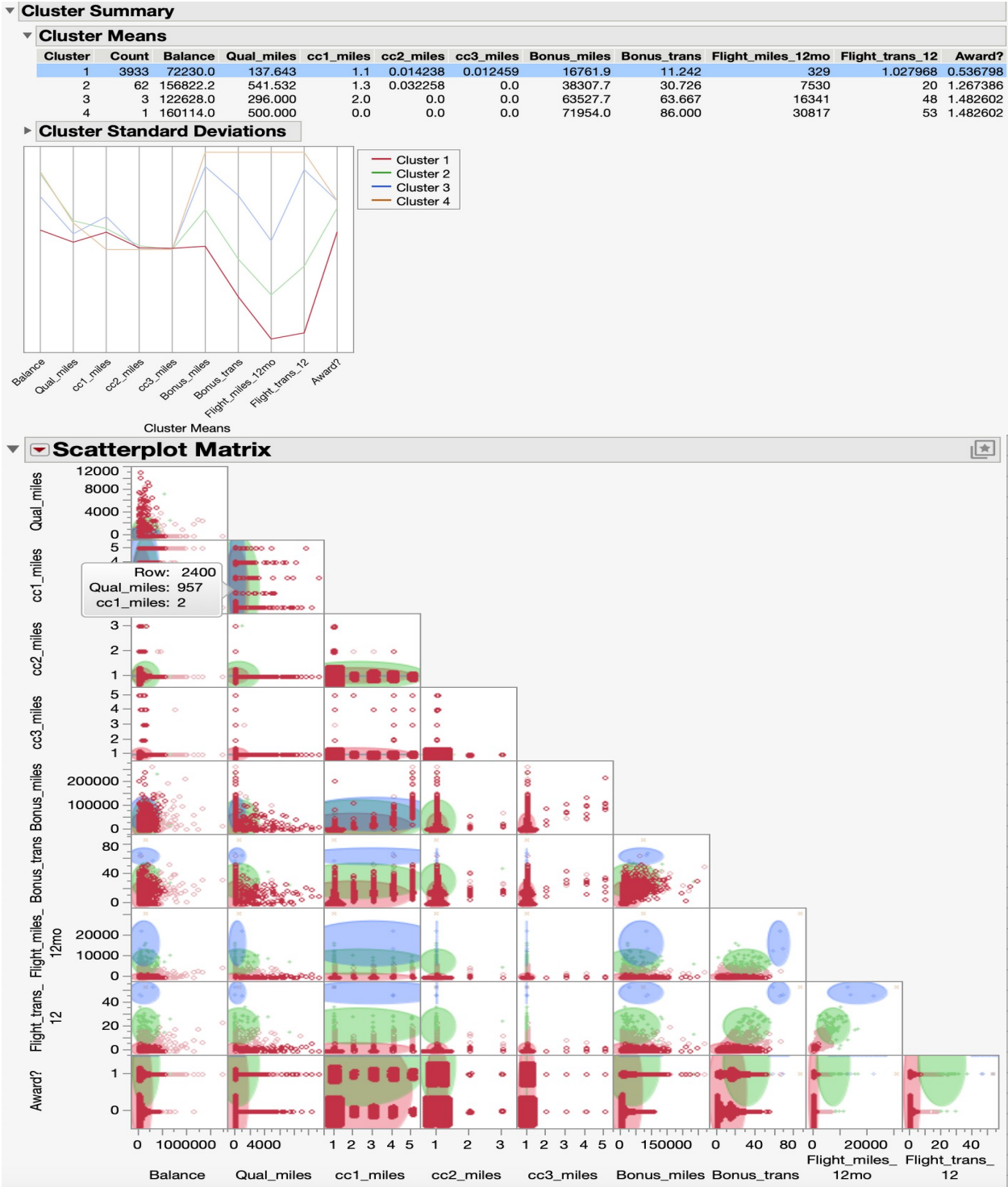
1. Apply **hierarchical clustering** to the EastWest Airline customers. Use the **Dendrogram** along with practical considerations to explore and identify the reasonable number of clusters. Compare different distance measurement methods (such as Ward, Centroid, and Average) used in hierarchical clustering. Report what **number of clusters** you have selected in the Dendrogram and the corresponding output of the selected clusters including **Cluster Summary**, **Scatterplot Matrix** and **Parallel Coord Plots**.

Ward method: I believe using **5 clusters** for this Dendrogram is a good choice. The structure suggests that 3 to 5 clusters would work well, and selecting 5 allows me to capture more detailed segmentation without losing meaningful groupings. It also helps identify subtler patterns in the data.

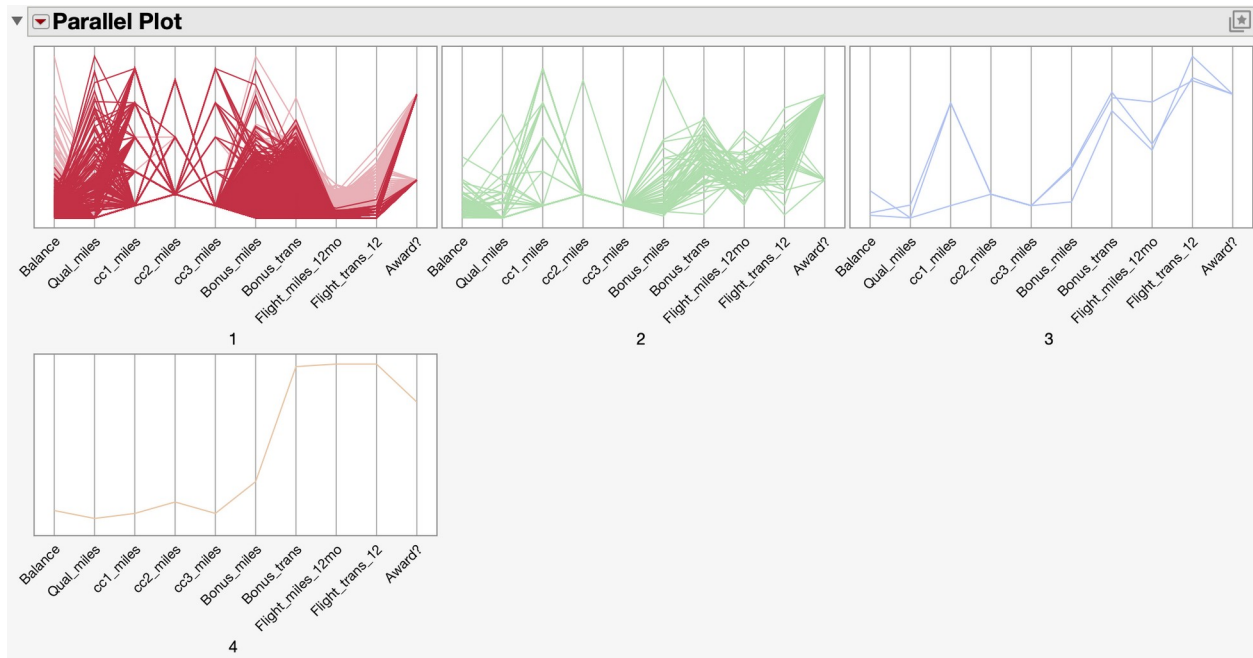




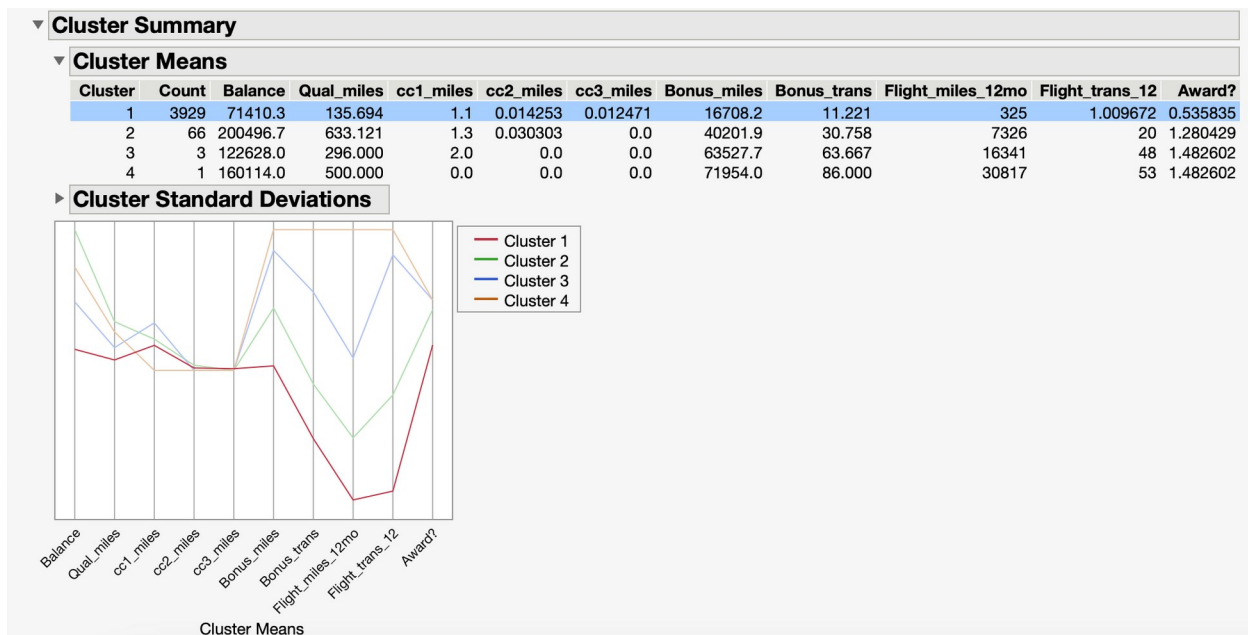
Average method: I believe using **4 clusters** for this Dendrogram makes the most sense because the structure shows a few major separations, with most of the data densely packed on the left and only a few large vertical splits on the right. Choosing 3 clusters would give a broader view, while 4 clusters would capture slightly more detail without over-segmenting. If I use 5 clusters, I might be forcing unnecessary splits where the data naturally doesn't separate as much. However, if I need finer segmentation, I could still consider it.

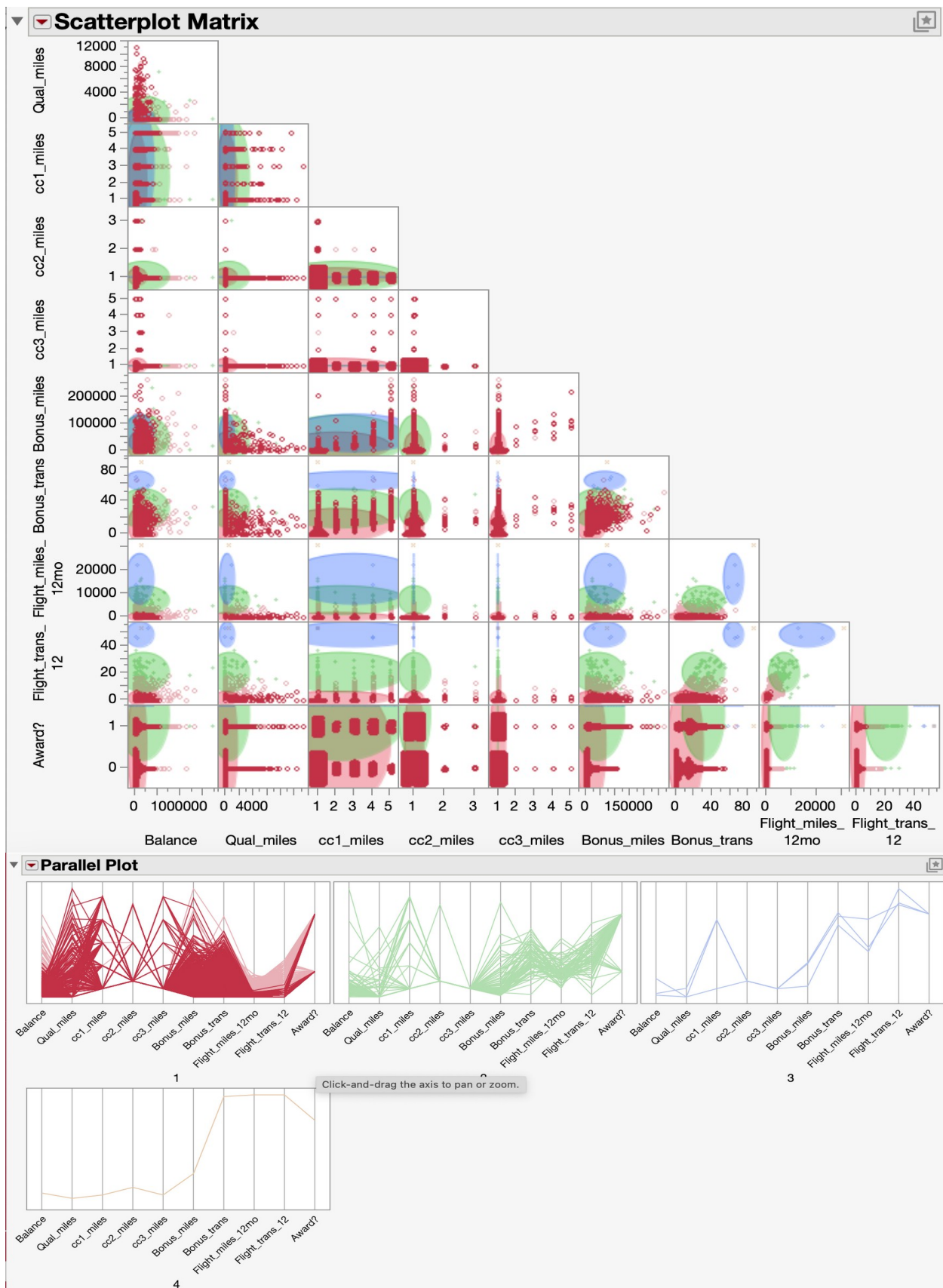


Scatterplot Matrix



Centroid method: I believe using **4 clusters** is the best choice because the Dendrogram shows a few major separations, making it a natural fit. This allows me to capture meaningful groupings without over-segmenting the data. If I used 5 clusters, I might be forcing unnecessary splits, which could reduce clarity. With 4 clusters, I can balance **detail and simplicity**, ensuring that each group remains distinct and interpretable.





2. Explore the clusters to characterize them. Compare the cluster centroids and try to give each cluster a meaningful label.

Cluster 1 – High Bonus Miles, Low Flight Activity (Credit Card Point Collectors)

This group gets the majority of their miles through bonus transactions (credit cards, promotions) and flies extremely seldom. Their balance is moderate, implying that they earn and redeem frequently. With low flight miles but high bonus miles, individuals participate in mileage programs but are not frequent travelers. They also have excellent award redemption rates, indicating that they make good use of their points.

Cluster 2 – Frequent Flyers with High Engagement (Elite Travelers)

This cluster flies the most frequently, earning the most Flight Miles and Qualifying Miles, showing elite status. They also participate in bonus schemes, however not as vigorously as Cluster 1. Their balance is high, implying that they earn and spend miles regularly. They are more likely to redeem rewards and benefit from frequent travel.

Cluster 3 – Large Balance Hoarders, Infrequent Travelers (Miles Savers)

This group has the highest Balance but flies infrequently. They have accrued miles throughout time but do not currently earn much. Their Bonus Miles and Transactions are moderate, which means they participate in some programs but do not aggressively accrue new miles. They rarely redeem, maintaining big balances with little activity.

Cluster 4 – Low Activity, New or Occasional Members (Light/Inactive Users)

This category has the lowest balance and kilometers earned, indicating minimal activity. They rarely fly and participate in bonus programs, resulting in poor award redemption rates. Many are new members or infrequent users who have not established allegiance.

3. Use k-means clustering with the number of clusters that you found above. Use the built-in graphical tools to characterize the clusters. How do you compare these results with the results from hierarchical clustering? Are they similar or different?

The segmentation uses K-means clustering with six clusters to provide a more precise analysis of client segments. The groupings include frequent elite travelers, credit card maximizers, high-value travelers, low-engagement members, mileage collectors, and a recently identified moderate-engagement category. K-means can efficiently group clients based on balance, qualifying miles, bonus miles, flight activity, and award redemption. The addition of a sixth cluster refines segmentation even more, distinguishing intermediate flyers from low-engagement individuals, who were previously merged in the five-cluster model.

When these results are compared to hierarchical clustering, both methods identify frequent flyers, bonus mile collectors, and low-engagement individuals, but hierarchical clustering offers a more organized view of how groups merge at different levels. K-means optimizes cluster assignments with centroids, which may force data into predefined groups, whereas hierarchical clustering preserves a natural tree-like development, displaying how smaller clusters interact to build larger groupings. While hierarchical clustering improves visualization and interpretability, K-means is still more versatile and scalable, particularly for large datasets.

The overall segmentation is consistent across the two techniques, but K-means with six clusters allows for finer distinctions, particularly when isolating a moderate-engagement group. Hierarchical clustering provides a more visual portrayal of how segments evolve, but K-means efficiently differentiates client kinds using numerical optimization. Hierarchical clustering is preferred when trying to understand the links between segments. If scalability and efficiency are important considerations, K-means is a superior option.

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC Best
K Means Cluster	3	-11.41
K Means Cluster	4	13.2796
K Means Cluster	5	28.5579
K Means Cluster	6	63.2427 Optimal CCC

Columns Scaled Individually

Control Panel

K Means NCluster=3

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	15	26	0
2	2673		
3	1311		

Cluster Means

Cluster	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Award?
1	138061.4	78.8	3.46666667		1 4.06666667	93927.8667	28.0666667	506.666667	1.6	0.53333333
2	45468.0296	91.0074822	1.29405163	1.01758324	1.00037411	5245.40142	7.26861205	213.594463	0.6221474	0.2020202
3	130224.825	253.141876	3.60411899	1.00839054	1.00152555	40528.119	20.2486651	962.032799	2.90312738	0.71167048

Cluster Standard Deviations

K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	1223	17	0
2	158		
3	2603		
4	15		

Cluster Means

Cluster	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Award?
1	117668.99	147.918234	3.77105478	1.00245298	1.00163532	40497.4064	18.9901881	345.991823	1.02616517	0.65331153
2	192704.032	798.196203	2.16455696	1.03797468		32008.0063	27.4113924	5511.92405	16.0949367	0.8164557
3	45295.5728	103.001537	1.24087591	1.01882443	1.00038417	4828.17096	7.07606608	206.734921	0.64195159	0.2093738
4	138061.4	78.8	3.46666667		1 4.06666667	93927.8667	28.0666667	506.666667	1.6	0.53333333

K Means NCluster=5

Cluster Summary

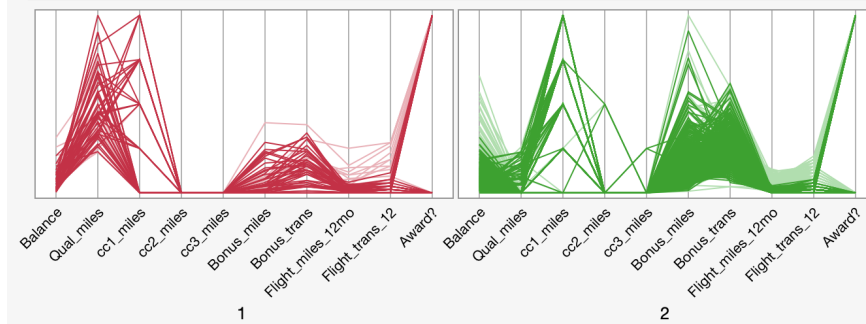
Cluster	Count	Step	Criterion
1	61	26	0
2	1210		
3	141		
4	2572		
5	15		

Cluster Means

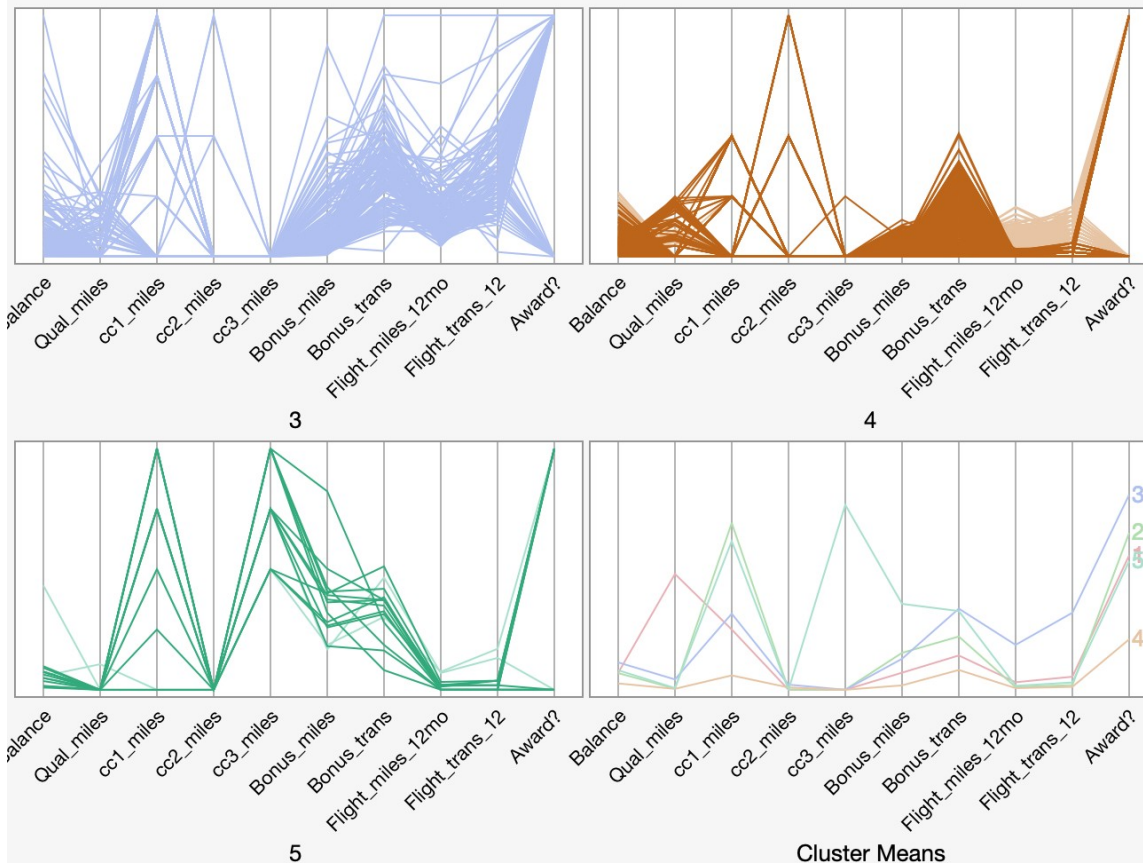
Cluster	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Award?
1	119660.492	5351.06557	2	1	1	18683.4754	12.2295082	963.639344	2.86885246	0.55737705
2	117723.49	56.3975207	3.76694215	1.00247934	1.00165289	40393.2438	18.9958678	347.080165	1.03057851	0.65041322
3	193487.085	478.255319	2.26241135	1.04255319	1	33887.4752	28.9361702	5736.56738	16.929078	0.80851064
4	44803.4199	43.9510109	1.23833593	1.01905132	1.0003888	4805.47084	7.0622084	211.725505	0.64541213	0.20917574
5	138061.4	78.8	3.46666667	1	4.06666667	93927.8667	28.0666667	506.666667	1.6	0.53333333

Cluster Standard Deviations

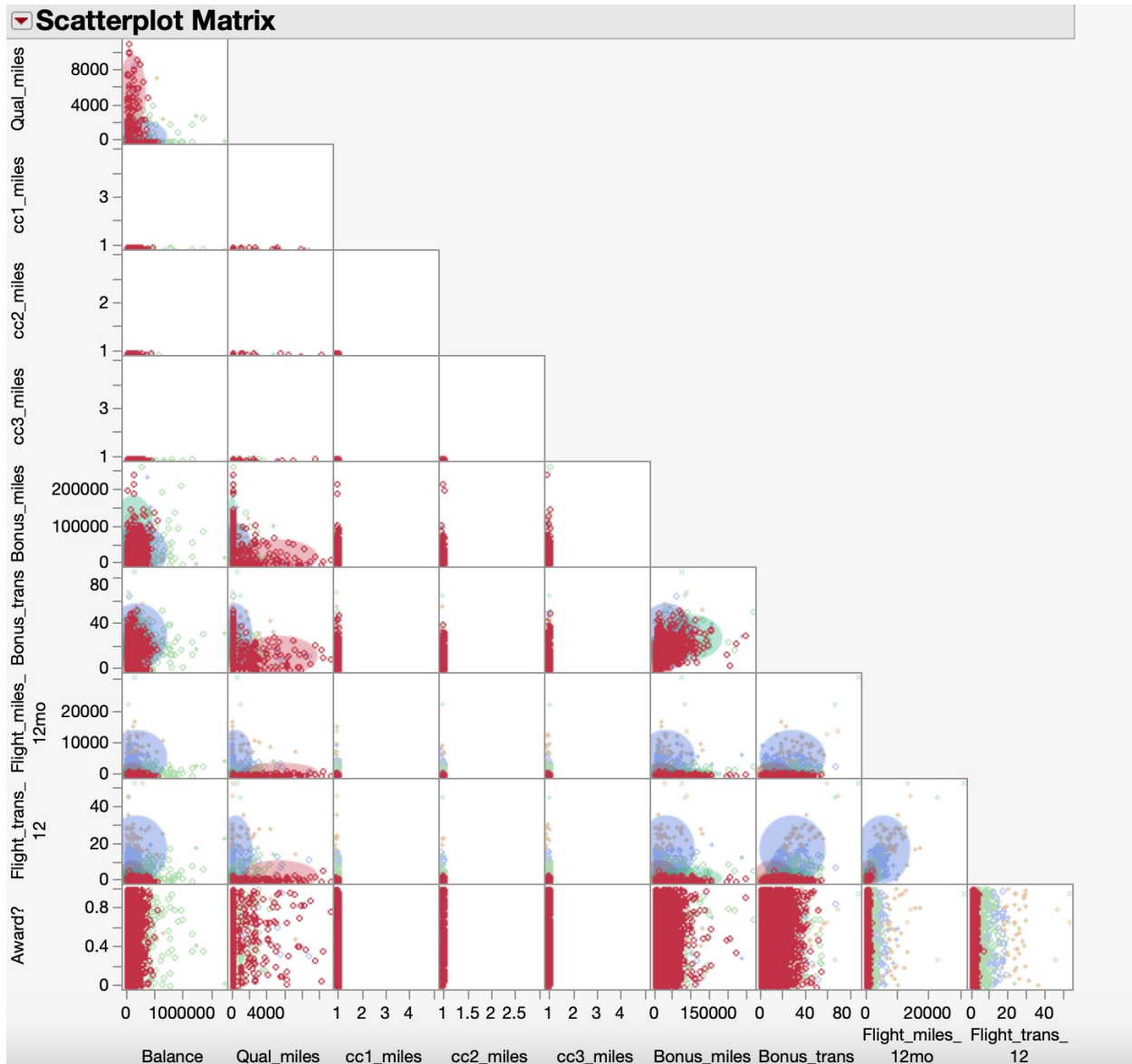
Parallel Coordinate Plot



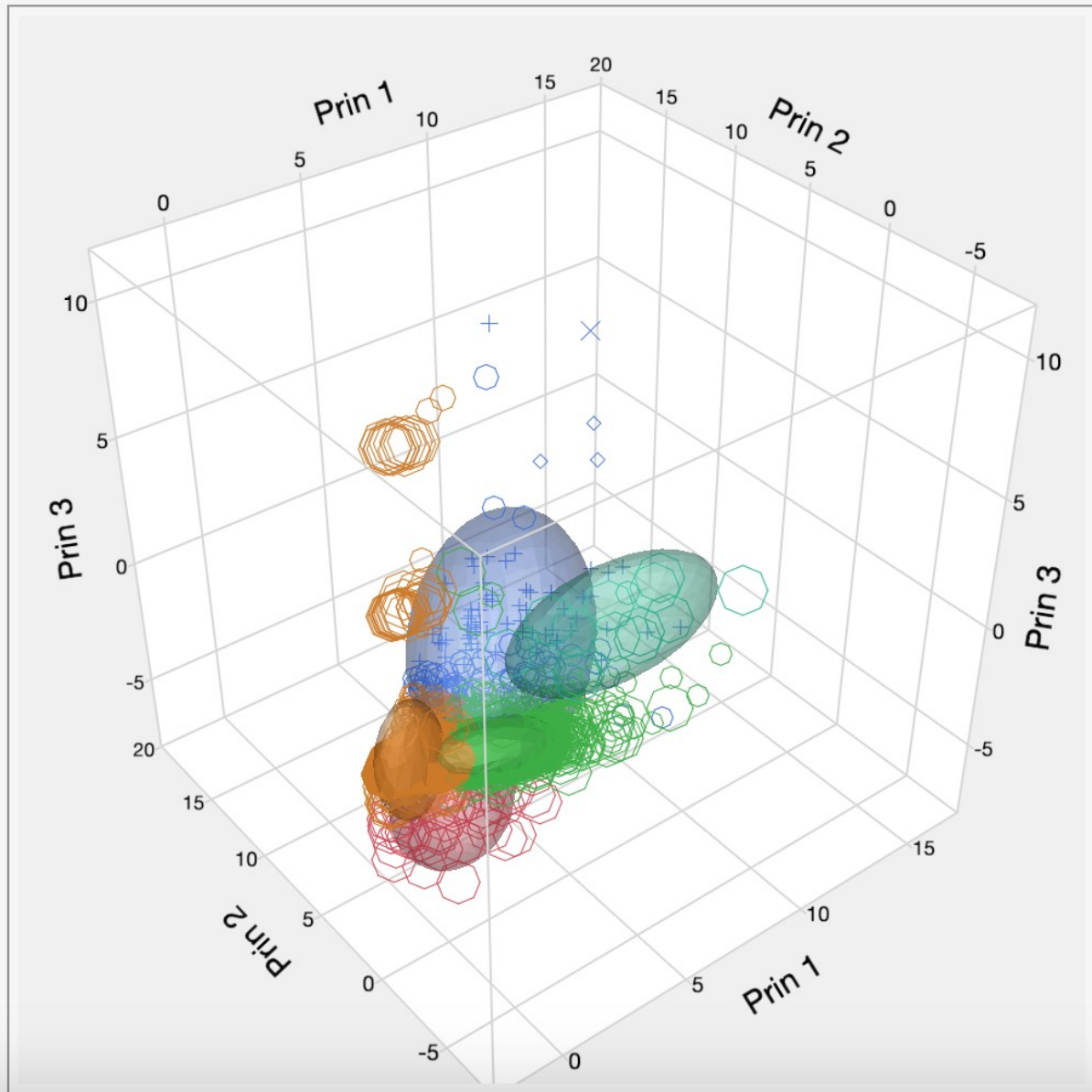
Parallel Coordinate Plot



Cluster Means



▼ Biplot 3D



K Means NCluster=6

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	61	19	0
2	43		
3	15		
4	2539		
5	1204		
6	137		

Cluster Means

Cluster	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Award?
1	119660.492	5351.06557	2	1	1	18683.4754	12.2295082	963.639344	2.86885246	0.55737705
2	68876.5814	23.255814	1.13953488	2.34883721	1	14689.8372	17.5348837	582.627907	2.20930233	0.39534884
3	138061.4	78.8	3.46666667	1	4.06666667	93927.8667	28.0666667	506.666667	1.6	0.53333333
4	44943.6625	45.3194171	1.2426152	1	1.00039386	4720.8657	6.9661284	211.523434	0.64552974	0.20834974
5	117663.575	54.9975083	3.77491694	1	1.00166113	40478.9535	18.9842193	346.722591	1.02574751	0.65033223
6	191393.445	484.919708	2.28467153	1	1	34007.5109	28.6934307	5794.27737	16.9708029	0.80291971

Cluster Standard Deviations

Cluster	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Award?
1	107992.375	2051.54679	1.35501497	0	0	21750.5092	8.12886668	1404.78043	3.84313486	0.49669696
2	97836.6561	150.7149	0.55327336	0.47660236	0	12308.9466	8.79309676	1227.40407	4.78640584	0.48892549
3	166325.988	294.842602	1.49962958	0	0.77172246	41430.0621	9.46197067	741.814592	2.65329983	0.49888765
4	51170.81	286.260284	0.574689	0	0.01984189	5978.69346	6.61735003	596.545474	1.56655885	0.40612822
5	120691.86	305.070635	0.87223816	0	0.04072309	26280.7443	6.93675153	661.021054	1.88017322	0.47686499
6	247055.977	892.353102	1.61176743	0	0	35556.2943	13.5434214	3941.86603	8.56826738	0.39779348

- What are the advantages and disadvantages of hierarchical vs. k-means clustering methods?

Hierarchical clustering is better suited for small datasets and gives easy interpretability via dendrograms, allowing users to visualize associations without limiting the number of clusters. However, it is computationally intensive and susceptible to outliers. In comparison, K-means clustering is faster and more scalable, making it perfect for large datasets. However, it needs predefining K, is sensitive to initialization, and assumes spherical clusters. When it comes to comprehending relationships, hierarchical clustering is preferred, whereas K-means is better for efficiency and large-scale applications.

- Search web to understand and describe the objectives and challenges of an airline frequent-flyer program. If you were the manager of an airline frequent-flyer program, which clusters would you target for incentive offers, and what types of offers would you provide to customers in that cluster?

Objectives and Challenges of Airline Frequent-Flyer Programs

Airlines' frequent-flyer programs (FFPs) seek to improve customer loyalty, generate revenue, and provide useful data about passenger behavior. These programs encourage repeat business by providing miles or points that can be redeemed for flights, upgrades, and other exclusive privileges. However, maintaining an FFP presents major issues, including balancing profitability

and customer happiness. If airlines make rewards more difficult to earn or redeem, they risk alienating loyal consumers. For example, Qantas recently raised the points required for premium cabin incentives by 20%, causing unhappiness among frequent fliers.

Another significant difficulty is reward availability. Customers may feel unrewarded if there aren't enough seats available for redemption. To remedy this, Qantas introduced Classic Plus, which adds 20 million more reward seats every year to improve accessibility. Airlines must also adjust to shifting market conditions. British Airways, for example, modified its loyalty program to base status points on spending rather than distance flown, bringing awards more in line with customer profitability. These continuing changes underscore the need for continuous innovation to make frequent flyer programs appealing and competitive.

Targeting Clusters for Incentive Offers

As an FFP manager, I would prioritize three main customer clusters to maximize engagement and revenue. First, Elite Frequent Flyers (High-Value Travelers) are the airline's most loyal clients. They travel frequently, earn a lot of money, and redeem rewards on a regular basis. To retain their loyalty, I would provide complimentary upgrades, priority boarding, and exclusive lounge access, guaranteeing that they continue to choose the airline over competitors.

Next, Credit Card Maximizers (High Bonus Earners, Moderate Flyers) earn miles primarily from credit card purchases and partner promotions, rather than flights. Because they are actively involved with the program but not frequent travelers, I would give bonus miles on select purchases, discounted award flights, and exclusive credit card advantages to encourage them to fly more frequently.

Finally, Low-Engagement Members (frequent flyers) are a category with unrealized potential. Because they fly seldom, personalized incentives such as bonus miles for first bookings, limited-time discounted flights, and promotional emails with tailored offers may reengage them. The idea is to turn occasional consumers into frequent fliers by providing tailored discounts that make participating in the program more rewarding.

Understanding the habits and preferences of various client segments allows me to build an incentive strategy that maximizes loyalty, satisfaction, and revenue growth, ensuring that the airline's frequent-flyer program remains competitive and useful to both the company and its consumers.

References

British Airways. (2024). *British Airways tweaks frequent-flyer scheme*. *The Times UK*. Retrieved from <https://www.thetimes.co.uk/article/british-airways-frequent-flyer-scheme-tweak-97m5gx0pb>

Qantas. (2024). *You will need 20% more points for Qantas reward seats in premium cabins under changes*. *The Australian Business*. Retrieved from <https://www.theaustralian.com.au/business/aviation/you-will-need-20-per-cent-more-points-for-qantas-reward-seats-in-premium-cabins-under-changes/news-story/5e277187fc2190a2daa2d2500b82845f>

Qantas. (2024). *Scoring Qantas rewards seats might be about to get easier*. *The Australian Business*. Retrieved from <https://www.theaustralian.com.au/business/aviation/scoring-qantas-rewards-seats-might-be-about-to-get-easier/news-story/9cb7539ca60413bd5f8c3034b374a19a>

Wikipedia contributors. (2024). *Frequent-flyer program*. *Wikipedia, The Free Encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Frequent-flyer_program

6. Report how much time you have spent on this assignment and what you have learned.

I worked on this assignment for around 5 hours, which included data analysis, cluster interpretation, and comparing several clustering approaches. The procedure included investigating hierarchical and K-means clustering, assessing cluster centroids, and finding relevant consumer groups.

This assignment helped me have a better grasp of client segmentation in airline frequent-flyer programs. I discovered how clustering algorithms can assist spot trends in client behavior, distinguish between frequent and infrequent users, and customize marketing campaigns based on data-driven insights. Furthermore, I investigated the benefits and drawbacks of hierarchical and K-means clustering, determining when to employ each method based on dataset size and clustering aims.

This task also helped me strengthen my data interpretation skills, namely examining numerical cluster summaries and generating strategic recommendations. By applying clustering techniques to real-world circumstances, I gained a deeper understanding of how firms employ machine learning for consumer segmentation and targeted marketing.