Marketing Analytics – Predictive Modeling

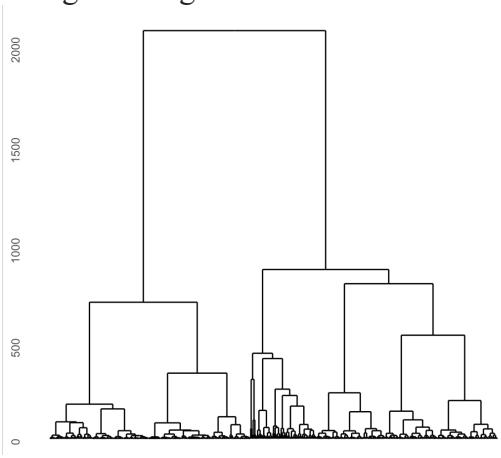
(Airline Frequent Flyer Customer Segmentation Using Clustering Analysis)

Project summary:

This project analyzed customer behavior data from the ***East West Airlines Frequent Flyer Program***, comprising 3,999 passengers and 13 behavioral variables related to mileage accrual, spending patterns, and bonus usage. The objective was to identify distinct customer segments with similar characteristics to support targeted mileage offers, personalized marketing strategies, and improved customer engagement.

1. **Explore, prepare, and transform the data to facilitate predictive modeling. Describe your process. (5)**
   - Reviewed data structure and identified key features, the dataset contains 13 columns and no missing values. Descriptive statistics revealed that data has wide variation in feature like “Balance” and “Bonus miles.”
   - Removed irrelevant column (column 1, ID), considered feature scaling for numerical data to handled high variability, and transformed categorical variables for predictive modeling.
   - Get the structure and summary of the data, as per that scale the data for further process.

2. **Apply hierarchical clustering with Euclidean distance and Ward's method. Make sure to normalize the data first. How many clusters would you pick and why? (5)**
   - Used Euclidean distance DIST ()_ to measure similarity between data points and ward’s method to minimize variance within cluster as they merge.
   - Using dendrogram to decide on the best k (no of cluster), and using cutree () functiree



   - According to the dendrogram diagram k=4 would be the optimum no of clusters.
3. **Compare the cluster centroid to characterize the different clusters and try to give each cluster a label. (5)**
   - After analyzing the model we calculated the centroids for each model from the characteristic analysis. in which,
   - **Cluster1**: Having high average balance and bonus miles, suggesting these customers are frequent and loyal flyers.

- **Cluster2:** shows the low balance and bonus miles, indicating budget conscious or infrequent flyers.
- **Cluster3:** It might exhibit high flight miles but bonus transaction, suggesting frequent travelers who don't often take advantage of bonus awards.
- **Cluster4:** Possible reflecting mix with moderate values across all features.
- Cluster characterization for k=4

	cluster	BM	BO	BT	FMile	Ftran	EnrD
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	39868.	3678.	6.40	120.	0.386	3499.
2	2	161110.	34913.	20.6	2279.	6.61	4462.
3	3	96738.	35506.	16.8	129.	0.407	5126.
4	4	50783.	9547.	9.68	255.	0.825	3956.

4. **Use k-means clustering with the number of clusters that you found above. Does the same picture emerge? (5)**
   - For choosing k=5, the resulting cluster differed in size and centroid identified customer groups with unique characteristics suing ward's method.
   - K means cluster were more even in sizes, while the hierarchal cluster showed greater size variation. this shows the clustering structure is not fully consistence between the 2 methods. Due to different approach hierarchal clustering build nested clusters, whereas k-means is a centroid based method.

	`km$cluster`	BM	BO	BT	FMile	Ftran	EnrD
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	137911.	49001.	19.9	369.	1.13	5026.
2	2	57564.	10753.	10.6	403.	1.21	4248.
3	3	43621.	4907.	7.25	152.	0.459	3635.
4	4	195903.	32926.	28.1	5596.	16.4	4729.

for hierarchal model

	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award.
1	0.4798738	0.009482464	1.32335079	-0.07987524	0.15932825	1.1003005	0.8123748	-0.06695137	-0.07562866	0.3168652	0.6584529
2	-0.3670849	-0.060376061	-0.57318281	-0.05280531	-0.06275873	-0.5082269	-0.5020806	-0.20883348	-0.22827342	-0.9144714	-0.4682490
3	1.1876857	0.841891007	0.08713096	0.16226123	-0.06275873	0.6271981	1.6687444	3.62101257	3.91175076	0.2757743	0.9321128
4	-0.1358841	-0.041615565	-0.50114553	0.11392920	-0.05864477	-0.4482805	-0.3301372	-0.14409002	-0.14952331	0.7788645	-0.1414921

#k-means model

	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award.
1	-0.1371691	-0.04091503	-0.5092247	0.11495660	-0.05862485	-0.4509295	-0.3378195	-0.13878826	-0.14636600	0.7768307	-0.1451485
2	-0.3668881	-0.05978220	-0.5791084	-0.05259105	-0.06275873	-0.5110696	-0.5042936	-0.20641590	-0.22639851	-0.9155379	-0.4724217
3	1.1934222	0.78586921	0.1858942	0.14365475	1.21771151	0.8963611	1.7031191	3.32934369	3.64256651	0.2771122	0.8970343
4	0.4638263	0.00666469	1.3168737	-0.07989180	-0.03502291	1.0528341	0.7949513	-0.07314971	-0.08540461	0.3155327	0.6590353

5. **To check the stability of the clusters, remove a random 5% of the data, and repeat the analysis. Does the same picture emerge? (5)**
   - To check the stability of the cluster, removed 5%of the data from the original data and perform the same model and it emerges the same result in both models, which is nearly same for both models.

#hierachal method with 95% data

	`km_sample$cluster`	BM	BO	BT	FMile	Ftran	EnrD
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	197319.	40485.	28.4	5264.	15.5	4768.
2	2	119107.	42184.	19.2	364.	1.07	4732.
3	3	59907.	6311.	8.43	273.	0.840	5726.
4	4	36994.	4838.	6.74	185.	0.547	2238.

#k- means methos with 95% data

	`km$cluster`	BM	BO	BT	FMile	Ftran	EnrD
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	59778.	6254.	8.36	266.	0.818	5723.
2	2	36628.	4802.	6.76	171.	0.515	2228.
3	3	193869.	38793.	28.0	5122.	15.2	4691.
4	4	120344.	42572.	19.2	358.	1.05	4770.

   - This dendrogram shows that the cluster structure largely consistent between the original and 95%datasets. Although there were minor difference in cluster sizes and centroid values an expected outcome when part of the data removed, the overall cluster structure was preserved, this stability suggest that the clustering approach is robust, with the cluster retaining their meaning and reliability even with a slight reduction in data.

6. **Which clusters would you target for offers, and what types of offers would you target to customers in that cluster? (5)**

   - This analysis reveals distinct customer segments, each with unique behaviors and engagement levels, allowing for targeted marketing strategies to improve engagement and satisfaction. Key findings include:
   - **Clusters 2, 4:** Represent high-value, engaged customers, ideal for premium and loyalty-focused offers**.(will be our target customers**)
   - Clusters 1 and 3: Include lower or moderate engagement customers who could benefit from incentives aimed at increasing loyalty and transaction frequency.
   - This segmentation, supported by both hierarchical and k-means clustering insights, provides a strong foundation for customized marketing approaches for each group

Conclusion:

Based on the analysis, high-value and moderately engaged clusters were identified as primary targets for premium rewards, loyalty incentives, and personalized mileage offers, while lower-engagement segments were recommended for activation and retention-focused campaigns.

This project demonstrates the effective use of unsupervised machine learning to translate customer behavior data into actionable marketing strategies for airline loyalty programs.

Codes

```
#Loading libraries
library(ggdendro)
library(tidyverse)
library(dplyr)
library(conflicted)
setwd("/Volumes/WORK/Fall 2024/R")
# Laod data and normalize it
set.seed(123)
ew.df <- read.csv("EastWestAirlines.csv")
head(ew.df)
str(ew.ef)
summary(ew.df)
eair <- ew.df[, -c(1,2)]
eair
d <- dist(eair, method = "euclidean")
eair.norm <- scale(eair)
eair.norm
d.norm <- dist(eair.norm, method="euclidean")
hc1 <- hclust(d.norm, method="ward.D")
ggdendrogram(hc1)
cluster <- cutree(hc1, k = 4)
cluster
##Characterizing clusters
```

```
cluster.means <- data.frame(cluster, eair) %>%
  group_by(cluster) %>%
  summarize(BM=mean(Balance),
            BO=mean(Bonus_miles),
            BT = mean(Bonus_trans),
            FMile = mean(Flight_miles_12mo),
            Ftran  = mean(Flight_trans_12),
            EnrD = mean(Days_since_enroll))
cluster.means
```

```
### Non-hierarchical Clustering:kmeans algorithm
km <- kmeans(eair.norm, 4)

# Investigate clusters formed
sort(km$cluster)    # show cluster membership
km$centers          # centroids
km$withinss         # within-cluster sum of squares
km$size             # cluster size

##Characterizing clusters
cluster.means <- data.frame(km$cluster, eair) %>%
  group_by(km$cluster) %>%
  summarize(BM=mean(Balance),
            BO=mean(Bonus_miles),
            BT = mean(Bonus_trans),
            FMile = mean(Flight_miles_12mo),
            Ftran  = mean(Flight_trans_12),
            EnrD = mean(Days_since_enroll))
cluster.means
```

```
##stability analysis with random 95% data
idx <- sample(1:nrow(eair.norm), size=0.95*nrow(eair.norm))
eair_sample <- eair.norm[idx,]
km_sample <- kmeans(eair_sample, 4)
sort(km_sample$cluster)    # show cluster membership
km$centers          # centroids
km$withinss         # within-cluster sum of squares
km$size
cluster_sample <- data.frame(km_sample$cluster, eair[idx, ]) %>%
  group_by(km_sample$cluster) %>%
  summarize(BM=mean(Balance),
            BO=mean(Bonus_miles),
            BT = mean(Bonus_trans),
            FMile = mean(Flight_miles_12mo),
            Ftran  = mean(Flight_trans_12),
            EnrD = mean(Days_since_enroll))
cluster_sample
```