Hubway Clustering Analysis_RituparnaDas

Part A: Importance of normalizing data in clustering

i. It’s a good idea to perform some type of scaling on our dataset (whether that is normalization or standardization) before running a clustering algorithm. Data scaling ensures that each feature/attribute in our data is being weighted equally by the clustering algorithm. Otherwise features with a much larger range of values compared to other features will influence the clustering output.

ii. After normalizing all the variables, we find the following results in R for all the variables

```
> mean(hubwaytripNorm$Duration)
[1] 3.819456e-18
> sd(hubwaytripNorm$Duration)
[1] 1
>
```

```
> mean(hubwaytripNorm$Morning)
[1] -2.017941e-17
> sd(hubwaytripNorm$Morning)
[1] 1
>
```

```
> mean(hubwaytripNorm$Afternoon)
[1] -1.310541e-17
> sd(hubwaytripNorm$Afternoon)
[1] 1
>
```

```
> mean(hubwaytripNorm$Evening)
[1] -6.87885e-17
> sd(hubwaytripNorm$Evening)
[1] 1
>
```

```
> mean(hubwaytripNorm$Night)
[1] -5.513973e-17
> sd(hubwaytripNorm$Night)
[1] 1
>
```

```
> mean(hubwaytripNorm$Weekday)
[1] -9.667016e-17
> sd(hubwaytripNorm$Weekday)
[1] 1
>
```

```
> mean(hubwaytripNorm$Weekend)
[1] 9.667016e-17
> sd(hubwaytripNorm$Weekend)
[1] 1
>
```

```
> mean(hubwaytripNorm$Male)
[1] 2.92973e-16
> sd(hubwaytripNorm$Male)
[1] 1
>
```

```
> mean(hubwaytripNorm$Age)
[1] 2.813452e-16
> sd(hubwaytripNorm$Age)
[1] 1
>
```

Part B: K means clustering with 10 clusters

i. The number of trips in each cluster is shown in the table below:

```
> table(KmeansClustering$cluster)

    1     2     3     4     5     6     7     8     9    10 
16287 31309  9893 15638 18632 30299 26187  4827 27482 13748 
>
```

ii. Examining the centroid of the 10 clusters in the unnormalized data(normalized data would render all means as 0 and non-informative), we come across some qualitative insights which we can present to the marketing team.

```
> tapply(hubwaytrip$Duration, KmeansClustering$cluster, mean)
        1         2         3         4         5         6         7         8         9        10 
 616.0338  795.6251 1388.7501  756.8779  655.0775  680.4691  581.8820  749.0445  625.8903  716.4264 
> tapply(hubwaytrip$Morning, KmeansClustering$cluster, mean)
           1            2            3            4            5            6            7            8            9           10 
1.0000000000 0.2273467693 0.0004043263 0.0000000000 0.0000000000 0.0000000000 1.0000000000 0.0000000000 0.0000000000 1.0000000000 
> tapply(hubwaytrip$Afternoon, KmeansClustering$cluster, mean)
          1           2           3           4           5           6           7           8           9          10 
0.000000000 0.501517136 0.000707571 1.000000000 1.000000000 0.000000000 0.000000000 0.000000000 1.000000000 0.000000000 
> tapply(hubwaytrip$Evening, KmeansClustering$cluster, mean)
        1         2         3         4         5         6         7         8         9        10 
0.0000000 0.2711361 0.9988881 0.0000000 0.0000000 1.0000000 0.0000000 0.0000000 0.0000000 0.0000000 
> tapply(hubwaytrip$Night, KmeansClustering$cluster, mean)
 1  2  3  4  5  6  7  8  9 10 
 0  0  0  0  0  0  0  1  0  0 
> tapply(hubwaytrip$Weekday, KmeansClustering$cluster, mean)
        1         2         3         4         5         6         7         8         9        10 
1.0000000 0.0000000 0.9995957 1.0000000 1.0000000 1.0000000 1.0000000 0.5815206 1.0000000 1.0000000 
> tapply(hubwaytrip$Weekend, KmeansClustering$cluster, mean)
           1            2            3            4            5            6            7            8            9           10 
0.0000000000 1.0000000000 0.0004043263 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.4184793868 0.0000000000 0.0000000000 
> tapply(hubwaytrip$Male,KmeansClustering$cluster, mean)
        1         2         3         4         5         6         7         8         9        10 
0.9981580 0.6929637 0.8321035 0.0000000 0.9977995 0.7224331 1.0000000 0.7984255 1.0000000 0.0000000 
> tapply(hubwaytrip$Age, KmeansClustering$cluster, mean)
       1        2        3        4        5        6        7        8        9       10 
49.68153 33.01310 46.49965 35.25233 49.90661 28.22469 29.87467 29.49492 29.44247 35.37962 
> 
```

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Duration	616	795.6	1388	756.8	655	680	581	749	625	716
Morning	1	0.22	0	0	0	0	1	0	0	1
Afternoon	0	0.5	0	1	1	0	0	0	1	0
Evening	0	0.27	1	0	0	1	0	0	1	0
Night	0	0	0	0	0	0	0	1	1	0
Weekday	1	0	1	1	1	1	1	0.58	1	1
Weekend	0	1	0	0	0	0	0	0.42	0	0
Male	0.99	0.69	0.83	0	0.99	0.722	1	0.79	1	0
Age	49.68	33.01	46.5	35.25	49.9	28.22	29.87	29.49	29.44	35.37

Cluster 10 has weekday morning data of all non-males, aged 35

Cluster 7 has weekday morning data of all males aged 29

Cluster 1 has weekday morning data of a different mean age of males, aged 49, lowest duration

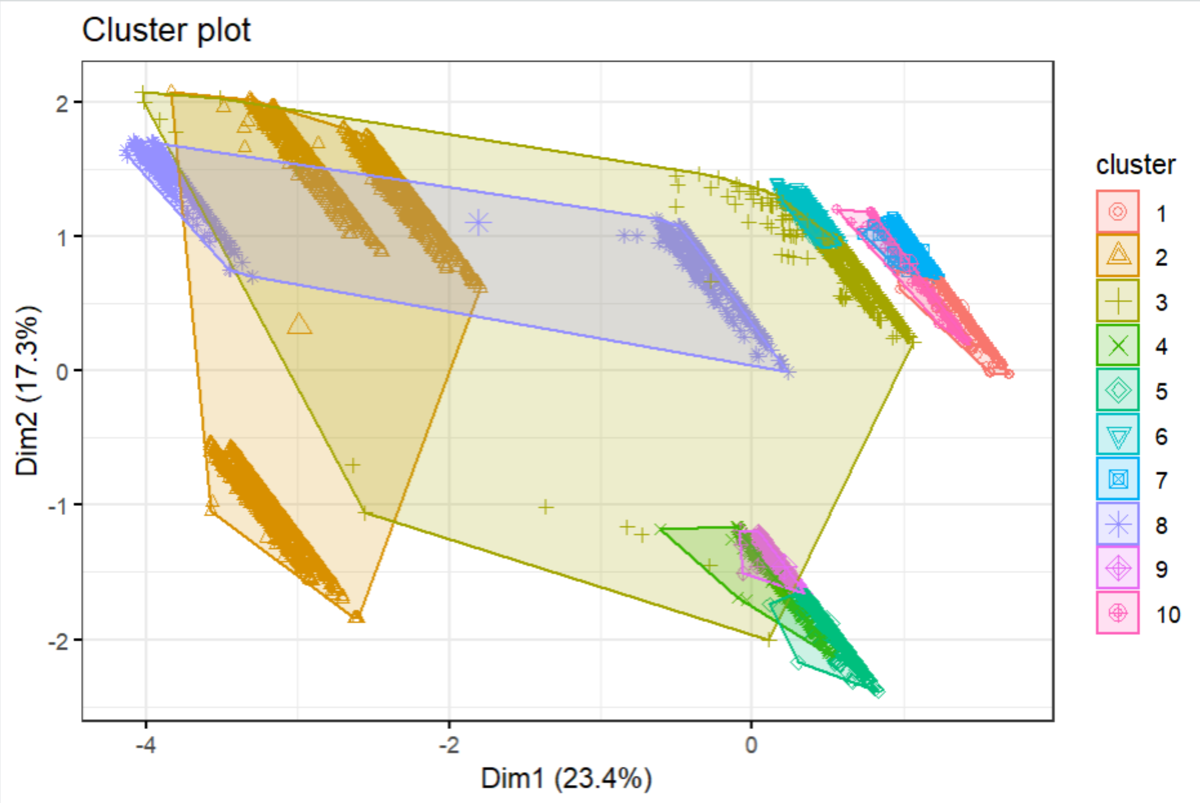
Cluster 1,5 and 6 morning, afternoon and evenings of weekdays of males of different ages

Cluster 4 gives weekday afternoons for non-males aged 35

Cluster 2 is the only cluster with weekend data

Looking at Cluster 3 and Cluster 6, both are weekday evenings, but one has less than half the duration of the other and 1.5 times the mean age.

Cluster 3 very high average duration, weekday evenings male



iii. Some of the clusters are more interesting than others like Clusters 1, 2,3,5,6,7,10.

iv. Looking at the graph above, I feel there are some data points close enough to be clustered together.

PART C: Repeating with 7 clusters

i. The number of trips in each cluster is shown in the table below:

```
> table(KmeansClustering$cluster)

    1     2     3     4     5     6     7 
22344 33329  9896 16298 46072 30301 36062 
> 
```

ii. We see some different clusters here

```
> tapply(hubwaytrip$Duration, KmeansClustering$cluster, mean)
        1         2         3         4         5         6         7 
 636.6340  792.8854 1414.7429  754.6363  635.8763  680.4317  620.6188 
> tapply(hubwaytrip$Morning, KmeansClustering$cluster, mean)
           1            2            3            4            5            6            7 
0.9944504117 0.2135677638 0.0004042037 0.0000000000 0.0000000000 0.0000000000 0.9428761577 
> tapply(hubwaytrip$Afternoon, KmeansClustering$cluster, mean)
           1            2            3            4            5            6            7 
0.0000000000 0.4711212458 0.0008084074 0.9620198797 1.0000000000 0.0000000000 0.0000000000 
> tapply(hubwaytrip$Evening, KmeansClustering$cluster, mean)
        1         2         3         4         5         6         7 
0.0000000 0.2547031 0.9983832 0.0000000 0.0000000 1.0000000 0.0000000 
> tapply(hubwaytrip$Night, KmeansClustering$cluster, mean)
           1            2            3            4            5            6            7 
0.0055495883 0.0606078790 0.0004042037 0.0379801203 0.0000000000 0.0000000000 0.0571238423 
> tapply(hubwaytrip$Weekday, KmeansClustering$cluster, mean)
        1         2         3         4         5         6         7 
1.0000000 0.0000000 0.9995958 1.0000000 1.0000000 1.0000000 1.0000000 
> tapply(hubwaytrip$Weekend, KmeansClustering$cluster, mean)
           1            2            3            4            5            6            7 
0.0000000000 1.0000000000 0.0004042037 0.0000000000 0.0000000000 0.0000000000 0.0000000000 
> tapply(hubwaytrip$Male,KmeansClustering$cluster, mean)
        1         2         3         4         5         6         7 
0.8280075 0.7009811 0.8321544 0.0000000 1.0000000 0.7224514 0.7244745 
> tapply(hubwaytrip$Age, KmeansClustering$cluster, mean)
       1        2        3        4        5        6        7 
48.73031 32.74632 46.49919 35.14210 37.68402 28.22527 29.24902 
> 
```

Cluster 1: Weekday mornings for males aged 48
Cluster 2: Weekends
Cluster 3: Very high duration evening weekday Male aged 46years
Cluster 4: Afternoon weekdays for non-males aged 35
Cluster 5: Weekday afternoon males mostly, average 37-year-olds
Cluster 6: Weekday evenings, mostly males aged 28
Cluster 7: Weekday mornings, mostly males aged 29

iii. Looking at the new numbers, I believe 10 clusters were better than 7. Because more clusters having non-males gave better perspective and inclusion. If anything without compromising the interpretability of the results, we can look at increasing the number of clusters.