Preparing data

In BuddyMove Dataset Data Set was populated from destination reviews published by 249 reviewers of holidayiq.com till October 2014. Reviews falling in 6 categories among destinations across South India were considered and the count of reviews in each category for every reviewer (traveler) is captured.

We dont use `Attribute 1 : Unique user id` and we cluster based on other features

Attribute Information

- Attribute 1 : Unique user id
- Attribute 2 : Number of reviews on stadiums, sports complex, etc.
- Attribute 3 : Number of reviews on religious institutions
- Attribute 4 : Number of reviews on beach, lake, river, etc.
- Attribute 5 : Number of reviews on theatres, exhibitions, etc.
- Attribute 6 : Number of reviews on malls, shopping places, etc.
- Attribute 7 : Number of reviews on parks, picnic spots, etc.

Analysis

Silhouette (clustering)

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster.

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

Kmeans

In this way we want to divide the data into k cluster and k as the input. Because we use classification so k equal to the number of classes in the classification.) k=6) Given the silhouette value obtained (0.2), we can conclude that the clustering is almost done well.

```
n_clusters : 6
silhouette_score : 0.2978550526393102
```

Hierarchical clustering

In this method, we want to divide the data into k-class, but in this way we do not have the number of classes as inputs. In this way, we use different linkage for clustring, and in each linkage, clustring is done for different k (1> k >10), and the best ones are selected.) compared based on the silhouette.)

linkage : ward

```
n_clusters : 2
silhouette_score : 0.24418093101151334

n_clusters : 3
silhouette_score : 0.28973852605254324

n_clusters : 4
silhouette_score : 0.24988027361895215

n_clusters : 5
silhouette_score : 0.24548845594378832

n_clusters : 6
silhouette_score : 0.2537695943060857

n_clusters : 7
silhouette_score : 0.26526788561620696

n_clusters : 8
silhouette_score : 0.2894292674564693

n_clusters : 9
silhouette_score : 0.2836963765775202
```

`K=3 > k=8 > k=9 > k=7 > k=6 > k=4 > k=5 > k=2`

In this case, the silhouette for k=3 has the highest value, so we can say that the k=3 is better.

linkage : average

```
n_clusters : 2
silhouette_score : 0.4763807066815255

n_clusters : 3
silhouette_score : 0.28849394043596843

n_clusters : 4
silhouette_score : 0.3247230508277392

n_clusters : 5
silhouette_score : 0.30359476855496775

n_clusters : 6
silhouette_score : 0.21029769932379913

n_clusters : 7
silhouette_score : 0.2569134280073912

n_clusters : 8
silhouette_score : 0.24288145683859752
```

```
n_clusters : 9
silhouette_score : 0.2451663065476037
```

`K=2 > k=4 > k=5 > k=3 > k=7 > k=9 > k=8 > k=6`

Based on the values, we can say that k=2 (s=0.4763) is better, because the silhouette has the highest value.

linkage : single

```
n_clusters : 2
silhouette_score : 0.4763807066815255

n_clusters : 3
silhouette_score : 0.24000100804992566

n_clusters : 4
silhouette_score : -0.028251777214029824

n_clusters : 5
silhouette_score : -0.10104153455087905

n_clusters : 6
silhouette_score : -0.11106803869427169

n_clusters : 7
silhouette_score : -0.2030951913529466

n_clusters : 8
silhouette_score : -0.08213424961087658

n_clusters : 9
silhouette_score : -0.07556247582712967
```

`K=2 > k=3 > k=4 > k=9 > k=8 > k=5 > k=6 > k=7`

Based on the values, we can say that k=2 (s=0.4763) is better. On the other hand, with regard to the values of the silhouette that are negated, we can say that they are not very good clustring.

linkage : complete

```
n_clusters : 2
silhouette_score : 0.4763807066815255

n_clusters : 3
silhouette_score : 0.28849394043596843

n_clusters : 4
silhouette_score : 0.24955186842539795

n_clusters : 5
silhouette_score : 0.23637582388852013
```

```
n_clusters : 6
silhouette_score : 0.24591972732074335

n_clusters : 7
silhouette_score : 0.24807474873485053

n_clusters : 8
silhouette_score : 0.23478197924553293

n_clusters : 9
silhouette_score : 0.26757421494826006
```

`K=2 > k=3 > k=9 > k=4 > k=7 > k=6 > k=5 > k=8`

Based on the silhouette value obtained for number of different classes, we can conclude that k=2 (s=0.4763) that has the highest amount of silhouette is a better clustring. On the other hand, since in all cases the value of the silhouette is positive, we can say that all clustring are pretty good.

Conclusion

According to the silhouette in two ways we can say that hierarchical method is better because it has largest value in the silhouette score.