Buddy Move Data

Unsupervised learning

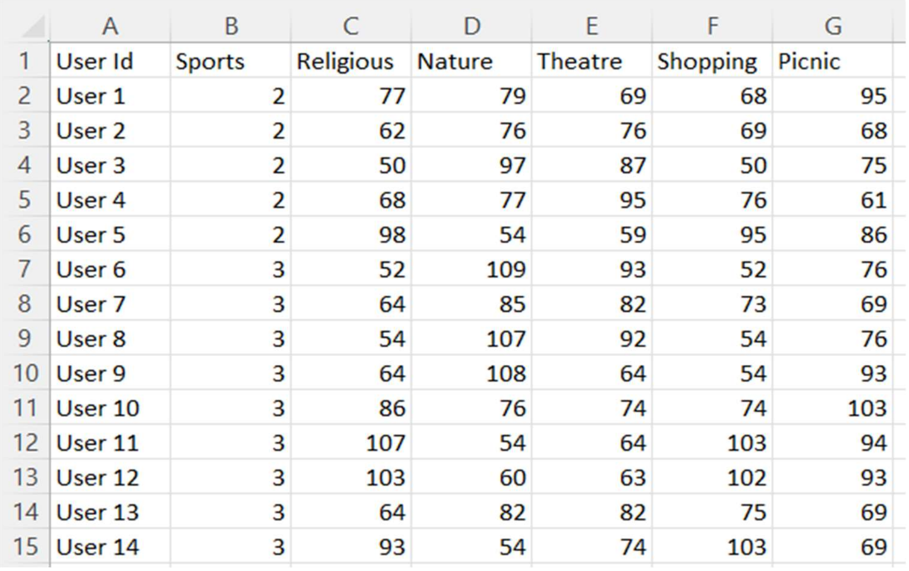
Nividha Jadeja (340)

1. **Introduction**:

The dataset is a review of Holiday destinations published by reviewers of a company till 2014. The destinations are places in South India. This data can be used to club customers into groups and provide them specialised deals.

2. **Dataset**:

The data is in a structured format in a CSV file. The customer reviews are historical data and so it is static data. The dataset contains total 7 attributes: UserID, number of reviews on sports, religious institutions, natural places like beach, lakes, and places like theatres, malls and parks.



	A	B	C	D	E	F	G
1	User Id	Sports	Religious	Nature	Theatre	Shopping	Picnic
2	User 1	2	77	79	69	68	95
3	User 2	2	62	76	76	69	68
4	User 3	2	50	97	87	50	75
5	User 4	2	68	77	95	76	61
6	User 5	2	98	54	59	95	86
7	User 6	3	52	109	93	52	76
8	User 7	3	64	85	82	73	69
9	User 8	3	54	107	92	54	76
10	User 9	3	64	108	64	54	93
11	User 10	3	86	76	74	74	103
12	User 11	3	107	54	64	103	94
13	User 12	3	103	60	63	102	93
14	User 13	3	64	82	82	75	69
15	User 14	3	93	54	74	103	69

3. **Data Preparation**:

The following steps have been ensured for data preparation:

- Querying the data using Pandas to check if all the data has been imported for processing.
- The dataset has been cleaned to drop the null values and remove outlier values for processing- using isnull().sum()
- Formatting of data- this wasn’t required as the dataset columns were in the required form

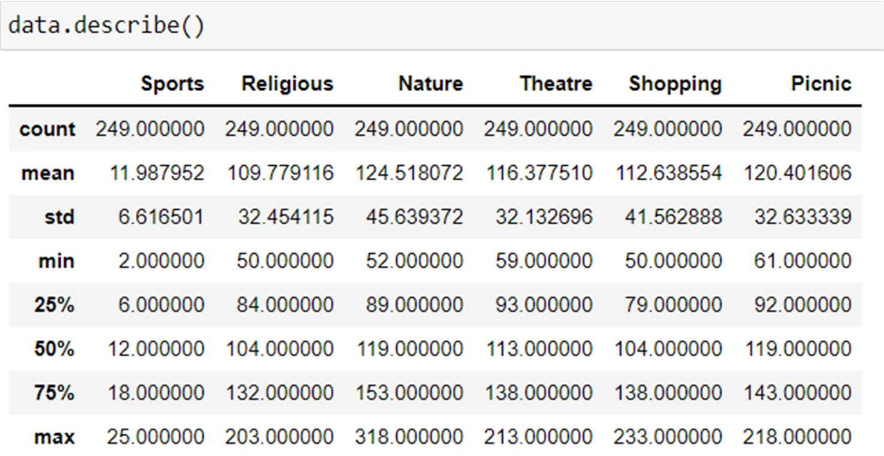
4. **Feature Engineering**:

The main types of features are- Categorical, Continuous and Derived features.

- This dataset did not have any value that can be categorised in Yes/no, Sex- M/F, or any Boolean value.
- Majority of the columns had continuous features with numerical values of their reviews.
- As with the categorical feature, derived features were not necessary as there were no time stamp to form a weekday- "yes" (is a weekday) or "no" (is not a weekday)

5. **Data Modelling**:

Data.describe(), gives us descriptive statistics for the dataset. We can discern what most clients prefer to visit as their holiday destination.

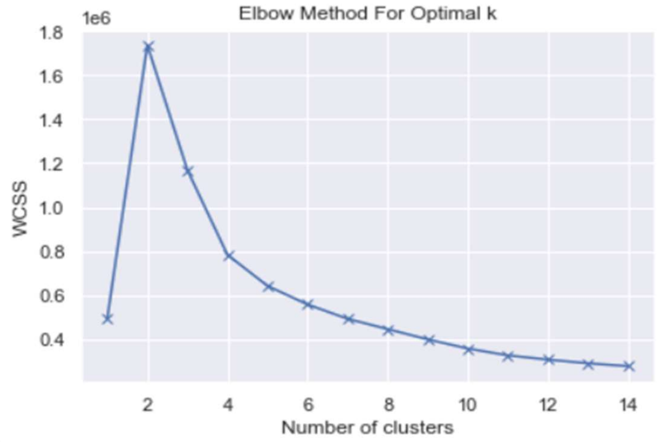


```
data.describe()
```

	Sports	Religious	Nature	Theatre	Shopping	Picnic
count	249.000000	249.000000	249.000000	249.000000	249.000000	249.000000
mean	11.987952	109.779116	124.518072	116.377510	112.638554	120.401606
std	6.616501	32.454115	45.639372	32.132696	41.562888	32.633339
min	2.000000	50.000000	52.000000	59.000000	50.000000	61.000000
25%	6.000000	84.000000	89.000000	93.000000	79.000000	92.000000
50%	12.000000	104.000000	119.000000	113.000000	104.000000	119.000000
75%	18.000000	132.000000	153.000000	138.000000	138.000000	143.000000
max	25.000000	203.000000	318.000000	213.000000	233.000000	218.000000

Hence the clustering technique for unsupervised model- K means algorithm is used here.

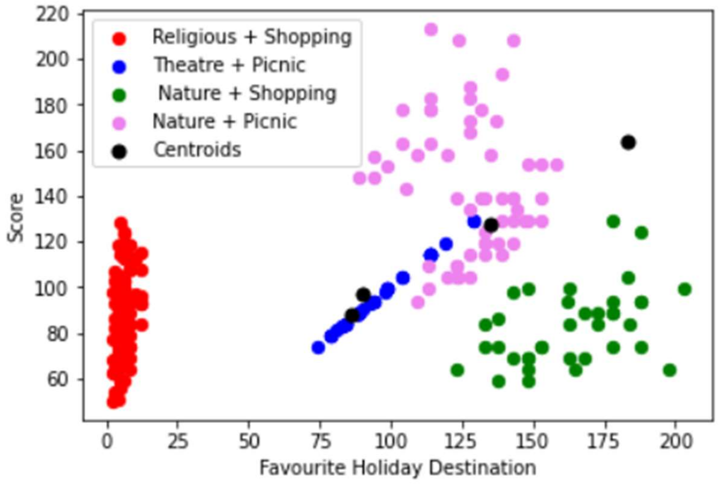
- To find the optimal number of clusters, Elbow method is used. The angle at which the line goes parallel with the X-axis is the number of clusters for our model.



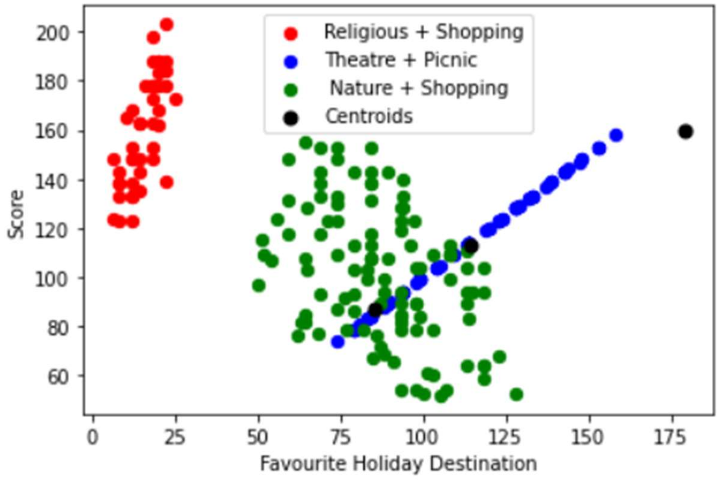
6. **Clusters / Observation**:

By using various combinations and offering a combined holiday package of nature and picnic places will be the most popular choice.

A religious pilgrim also has a mythological story attached with the place, can be of interest to the visitors. Thus, by using various combinations, we can provide specialised deals to the targeted groups.



By changing the values, we can get different clusters/grouping:



7. **References**:

- https://archive.ics.uci.edu/ml/datasets/BuddyMove+Data+Set
- https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f