Statistical Analysis Report

Author: Aurora

Background

There are many reviews on the website of TripAdvisor everyday. The data includes 249 different high volume reviewers and the some information of their reviews. This analysis will use k-means clustering to segment these reviewers into distinct clusters to help explain the common and the different between reviewers.

Data Source

Data was obtained from TripAdvisor. There are ten variables, User ID(Unique Identifier), Sports(Percentage of reviews related to sporting locations), Religious(Percentage of reviews related to religious locations), Nature(Percentage of reviews related to nature locations), Theatre(Percentage of reviews related to theatre), Shopping(Percentage of reviews related to shopping locations), Picnic(Percentage of reviews related to picnic locations), Age(Self-reported age), Income(Income inferred from geographic tax records), Nbr(Total Number of Reviews)

Data Transformation and Cleaning (Description)

Normalizing data with N(0,1) function

	ï..User.Id	Sports	Religious	Nature	Theatre	Shopping	Picnic	Age	Income	Nbr
1	User 1	0.005128205	0.1974359	0.2025641	0.1769231	0.1743590	0.2435897	44	53441.2	390
2	User 2	0.005665722	0.1756374	0.2152975	0.2152975	0.1954674	0.1926346	35	61412.0	353
3	User 3	0.005540166	0.1385042	0.2686981	0.2409972	0.1385042	0.2077562	18	66359.1	361
4	User 4	0.005277045	0.1794195	0.2031662	0.2506596	0.2005277	0.1609499	24	61344.3	379
5	User 5	0.005076142	0.2487310	0.1370558	0.1497462	0.2411168	0.2182741	52	53163.7	394
6	User 6	0.007792208	0.1350649	0.2831169	0.2415584	0.1350649	0.1974026	51	49523.6	385

	SportsNorm_RW	ReligiousNorm_RW	NatureNorm_RW	TheatreNorm_RW	ShoppingNorm_RW	PicnicNorm_RW
1	-1.880073	0.3522515	-0.12012547	-0.4723631	-0.2730941	1.4201734
2	-1.805406	-0.2400670	0.08743659	0.4113481	0.1551331	-0.3063525
3	-1.822847	-1.2490676	0.95790223	1.0031802	-1.0004809	0.2060185
4	-1.859398	-0.1372973	-0.11031045	1.2256926	0.2577910	-1.3799321
5	-1.887305	1.7460634	-1.18795396	-1.0982110	1.0812211	0.5623982
6	-1.510017	-1.3425196	1.19293867	1.0161042	-1.0702523	-0.1447961

	AgeNorm_RW	IncomeNorm_RW	NbrNorm_RW
1	0.5875722	0.21664671	-1.604465
2	-0.2090094	0.50408327	-1.893061
3	-1.7136634	0.68248159	-1.830662
4	-1.1826090	0.50164192	-1.690264
5	1.2956447	0.20663973	-1.573265
6	1.2071356	0.07537338	-1.643464

Descriptive Data Analysis

Summary data

```
   ï..User.Id      Sports           Religious          Nature           Theatre          Shopping          Picnic            Age
User 1  :  1   Min.   :0.005076   Min.   :0.1087   Min.   :0.08828   Min.   :0.1115   Min.   :0.1056   Min.   :0.1438   Min.   :18.00
User 10 :  1   1st Qu.:0.011905   1st Qu.:0.1556   1st Qu.:0.16584   1st Qu.:0.1630   1st Qu.:0.1456   1st Qu.:0.1801   1st Qu.:27.00
User 100:  1   Median :0.019200   Median :0.1788   Median :0.20849   Median :0.1869   Median :0.1834   Median :0.1969   Median :38.00
User 101:  1   Mean   :0.018663   Mean   :0.1845   Mean   :0.20993   Mean   :0.1974   Mean   :0.1878   Mean   :0.2017   Mean   :37.36
User 102:  1   3rd Qu.:0.024828   3rd Qu.:0.2106   3rd Qu.:0.26555   3rd Qu.:0.2337   3rd Qu.:0.2163   3rd Qu.:0.2254   3rd Qu.:48.00
User 103:  1   Max.   :0.032342   Max.   :0.2736   Max.   :0.37722   Max.   :0.3027   Max.   :0.3190   Max.   :0.2686   Max.   :55.00
(Other) :243
     Income             Nbr          SportsNorm_RW      ReligiousNorm_RW  NatureNorm_RW      TheatreNorm_RW    ShoppingNorm_RW
Min.   :  962.9   Min.   :353.0   Min.   :-1.88730   Min.   :-2.0598   Min.   :-1.98310   Min.   :-1.9788   Min.   :-1.66851
1st Qu.:23789.5   1st Qu.:494.0   1st Qu.:-0.93874   1st Qu.:-0.7837   1st Qu.:-0.71873   1st Qu.:-0.7929   1st Qu.:-0.85714
Median :47985.8   Median :595.0   Median : 0.07464   Median :-0.1542   Median :-0.02346   Median :-0.2429   Median :-0.08903
Mean   :47433.4   Mean   :595.7   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000
3rd Qu.:67165.4   3rd Qu.:710.0   3rd Qu.: 0.85636   3rd Qu.: 0.7109   3rd Qu.: 0.90653   3rd Qu.: 0.8342   3rd Qu.: 0.57696
Max.   :99949.1   Max.   :843.0   Max.   : 1.90012   Max.   : 2.4227   Max.   : 2.72695   Max.   : 2.4232   Max.   : 2.66161

 PicnicNorm_RW       AgeNorm_RW       IncomeNorm_RW        NbrNorm_RW
Min.   :-1.9617   Min.   :-1.71366   Min.   :-1.67578   Min.   :-1.893061
1st Qu.:-0.7298   1st Qu.:-0.91708   1st Qu.:-0.85263   1st Qu.:-0.793273
Median :-0.1621   Median : 0.05652   Median : 0.01992   Median :-0.005482
Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.000000
3rd Qu.: 0.8036   3rd Qu.: 0.94161   3rd Qu.: 0.71156   3rd Qu.: 0.891508
Max.   : 2.2677   Max.   : 1.56117   Max.   : 1.89378   Max.   : 1.928897
```

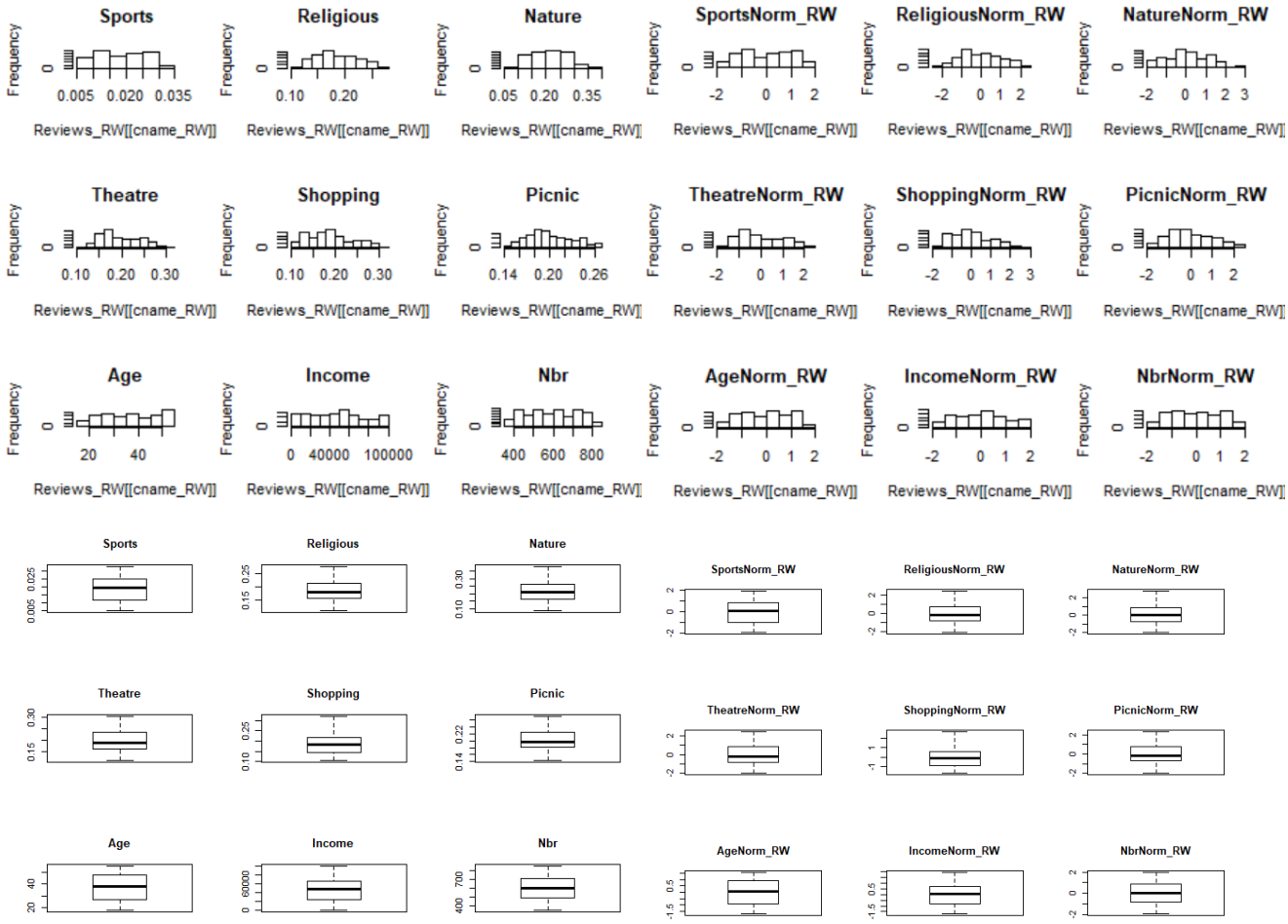
Descriptive data

	ï..User.Id	Sports	Religious	Nature	Theatre	Shopping	Picnic	Age
nbr.val	NA	249.000000000000	249.000000000	249.000000000	249.000000000	249.000000000	249.0000000000	249.0000000
nbr.null	NA	0.000000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.0000000000	0.0000000
nbr.na	NA	0.000000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.0000000000	0.0000000
min	NA	0.00507614200	0.108667529	0.088275862	0.111506524	0.105575326	0.1437802910	18.0000000
max	NA	0.03234152700	0.273631841	0.377224199	0.302658487	0.319018405	0.2686025410	55.0000000
range	NA	0.02726538500	0.164964312	0.288948337	0.191151963	0.213443079	0.1248222500	37.0000000
sum	NA	4.64700966200	45.933612544	52.273431945	49.161321069	46.767305194	50.2173195950	9303.0000000
median	NA	0.01920000000	0.178797468	0.208494208	0.186885246	0.183431953	0.1968911920	38.0000000
mean	NA	0.01866268941	0.184472340	0.209933462	0.197435024	0.187820503	0.2016759823	37.3614458
SE.mean	NA	0.00045621270	0.002332232	0.003887719	0.002751891	0.003123794	0.0018703186	0.7159993
CI.mean	NA	0.00089854542	0.004593507	0.007657158	0.005420057	0.006152549	0.0036837339	1.4102148
var	NA	0.00005182438	0.001354387	0.003763476	0.001885653	0.002429765	0.0008710248	127.6510882
std.dev	NA	0.00719891492	0.036801999	0.061347179	0.043424107	0.049292645	0.0295131292	11.2982781
coef.var	NA	0.38573834460	0.199498737	0.292222012	0.219941255	0.262445494	0.1463393354	0.3024047

	Income	Nbr	SportsNorm_RW	ReligiousNorm_RW	NatureNorm_RW	TheatreNorm_RW	ShoppingNorm_RW
nbr.val	249.0000000	249.0000000	2.490000e+02	2.490000e+02	2.490000e+02	2.490000e+02	2.490000e+02
nbr.null	0.0000000	0.0000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
nbr.na	0.0000000	0.0000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	962.9000000	353.0000000	-1.887305e+00	-2.059801e+00	-1.983100e+00	-1.978820e+00	-1.668508e+00
max	99949.1000000	843.0000000	1.900125e+00	2.422681e+00	2.726951e+00	2.423158e+00	2.661612e+00
range	98986.2000000	490.0000000	3.787430e+00	4.482482e+00	4.710051e+00	4.401978e+00	4.330120e+00
sum	11810928.5000000	148330.0000000	-1.859624e-15	-5.551809e-14	-2.125036e-14	1.888767e-14	-7.267624e-15
median	47985.8000000	595.0000000	7.463772e-02	-1.542001e-01	-2.346080e-02	-2.429475e-01	-8.903052e-02
mean	47433.4477912	595.7028112	-7.904226e-18	-2.229982e-16	-8.533961e-17	7.585616e-17	-2.921914e-17
SE.mean	1757.3579833	8.1247580	6.337243e-02	6.337243e-02	6.337243e-02	6.337243e-02	6.337243e-02
CI.mean	3461.2494985	16.0023255	1.248168e-01	1.248168e-01	1.248168e-01	1.248168e-01	1.248168e-01
var	768988463.2535536	16436.9113227	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
std.dev	27730.6412341	128.2065183	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
coef.var	0.5846221	0.2152189	-1.265146e+17	-4.484342e+15	-1.171789e+16	1.318284e+16	-3.422414e+16

	PicnicNorm_RW	AgeNorm_RW	IncomeNorm_RW	NbrNorm_RW
nbr.val	2.490000e+02	2.490000e+02	2.490000e+02	2.490000e+02
nbr.null	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	-1.961693e+00	-1.713663e+00	-1.675783e+00	-1.893061e+00
max	2.267688e+00	1.561172e+00	1.893777e+00	1.928897e+00
range	4.229380e+00	3.274835e+00	3.569560e+00	3.821959e+00
sum	3.165870e-14	-2.039341e-14	-1.874716e-14	4.257011e-14
median	-1.621241e-01	5.651783e-02	1.991848e-02	-5.481868e-03
mean	1.271500e-16	-8.188753e-17	-7.522796e-17	1.697025e-16
SE.mean	6.337243e-02	6.337243e-02	6.337243e-02	6.337243e-02
CI.mean	1.248168e-01	1.248168e-01	1.248168e-01	1.248168e-01
var	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
std.dev	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
coef.var	7.864725e+15	-1.221187e+16	-1.329293e+16	5.892667e+15

Graphical summary



From the numeric and graphical summaries, we conclude that the normalization worked properly. Also, all of the data looks reasonable. That means there are no values that seem like they are necessarily wrong. And it seems that there are no outliners.

Clustering

Model 1: 2 clusters (k=2)

When clustering 2 clusters, the sizes of them is 133,116. Ratio of variance is 67.5%. Total within cluster sum of square is161.3.

```
K-means clustering with 2 clusters of sizes 133, 116

Cluster means:
  ReligiousNorm_RW NatureNorm_RW
1       -0.7866884     0.7439186
2        0.9019789    -0.8529412
```

```
Within cluster sum of squares by cluster:
[1] 81.52903 79.79127
 (between_SS / total_SS =  67.5 %)

ClstrRev_RW$tot.withinss
[1] 161.3203
```

Model 2: 3 clusters (k=3)

When clustering 3 clusters, the sizes of them is 115, 56, 78. Ratio of variance is 84.2%. Total within cluster sum of square is 78.4.

```
K-means clustering with 3 clusters of sizes 115, 56, 78

Cluster means:
  ReligiousNorm_RW NatureNorm_RW
1       0.02834853    -0.1315929
2       1.41361248    -1.3811712
3      -1.05669717     1.1856252
Within cluster sum of squares by cluster:
[1] 39.92764 14.26429 24.25126
 (between_SS / total_SS =  84.2 %)

ClstrRev_RW$tot.withinss
[1] 78.44319
```

Model 3: 4 clusters (k=4)

When clustering 4 clusters, the sizes of them is 71, 53, 66, 59. Ratio of variance is 90.2%. Total within cluster sum of square is 48.5.

```
K-means clustering with 4 clusters of sizes 71, 53, 66, 59

Cluster means:
  ReligiousNorm_RW NatureNorm_RW
1       -1.0924507     1.2483583
2        1.4515787    -1.4092726
3        0.4274348    -0.3613061
4       -0.4674638     0.1678680
Within cluster sum of squares by cluster:
[1] 19.682612 11.888741  9.290550  7.591638
 (between_SS / total_SS =  90.2 %)

ClstrRev_RW$tot.withinss
[1] 48.45354
```

Model 4: 5 clusters (k=5)

When clustering 5 clusters, the sizes of them is 36, 57, 53, 37, 66. Ratio of variance is 91.9%. Total within cluster sum of square is 40.3.

```
K-means clustering with 5 clusters of sizes 36, 57, 53, 37, 66

Cluster means:
  ReligiousNorm_RW NatureNorm_RW
1       -0.7285288     1.2561847
2       -0.4667167     0.1467334
3        1.4515787    -1.4092726
4       -1.4139048     1.2148972
```

```
5        0.4274348    -0.3613061
Within cluster sum of squares by cluster:
[1]  7.113628  6.760279 11.888741  5.199727  9.290550
 (between_SS / total_SS =  91.9 %)

ClstrRev_RW$tot.withinss
[1] 40.25293
```

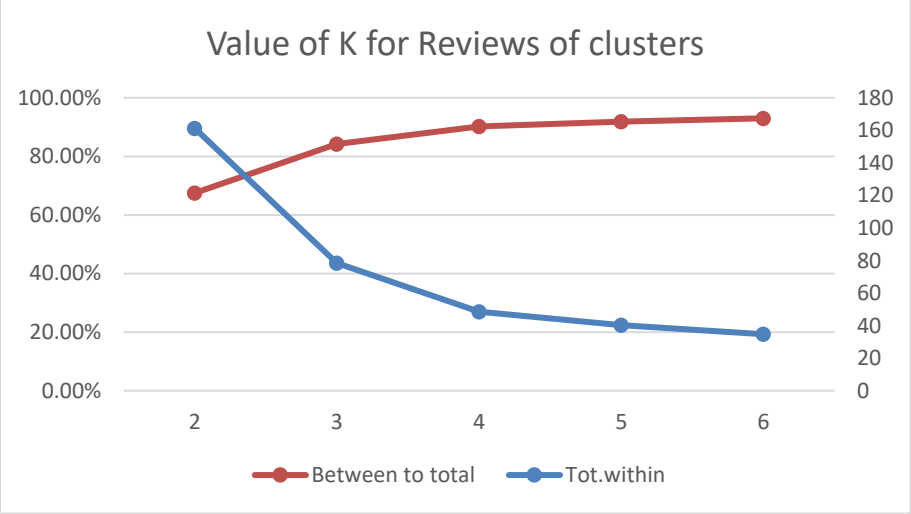
Model 5: 6 clusters (k=6)

When clustering 6 clusters, the sizes of them is 37, 46, 54, 29, 36, 47. Ratio of variance is 93.0%. Total within cluster sum of square is 35.7.

```
K-means clustering with 6 clusters of sizes 37, 46, 54, 29, 36, 47

Cluster means:
  ReligiousNorm_RW NatureNorm_RW
1       -1.4139048     1.2148972
2        0.2776017    -0.2138520
3       -0.4884886     0.1575676
4        0.7489530    -0.6797258
5       -0.7285288     1.2561847
6        1.4985230    -1.4709204
Within cluster sum of squares by cluster:
[1] 5.199727 4.016297 5.970179 3.615845 7.113628 8.818432
 (between_SS / total_SS =  93.0 %)

ClstrRev_RW$tot.withinss
[1] 34.73411
```

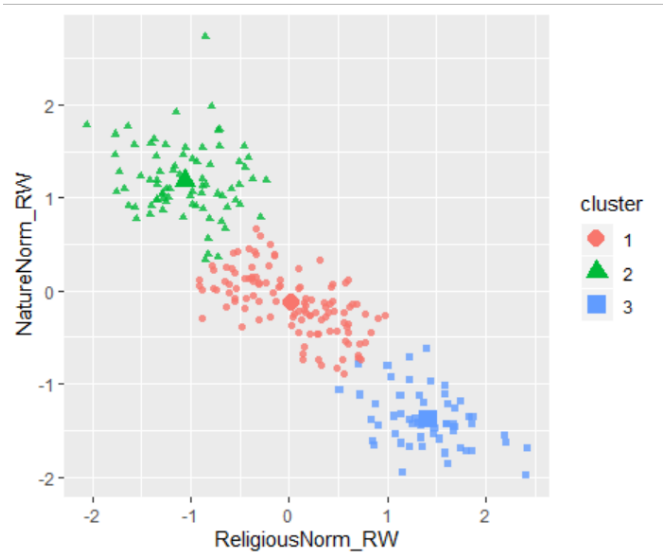


According to the value of between to total and the value of total within, we could find that between to total is increasing and total within is decreasing with the K value increasing. Based on the definition of 'elbow', the best model is the one that K is 3 which means the number of clusters is 3.

Model Evaluation

We choose model 2 which K is 3 and evaluate this model.

Scatter plot



Summary table

cluster	Sports	Religious	Nature	Theatre	Shopping	Picnic	Age	Income	Nbr	N
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
1 1	0.0178	0.186	0.202	0.209	0.188	0.198	37.1	48863.	587.	115
2 2	0.0201	0.146	0.283	0.202	0.138	0.211	36.9	45520.	592.	78
3 3	0.0184	0.236	0.125	0.168	0.257	0.195	38.4	47162.	618.	56

Descriptive name

1 Wealthy people who neither like nor dislike both religious and nature.

2 Lower income people who like nature more than religious.

3 Medium income people who like religious more than nature.

Suggest use

This model could be used for improving this website. This model indicates that what kind of person prefer to review the religious topic and what kind of person prefer to review the nature topic. According to the difference between these three groups, the website could update the information respectively.