# CS57300: Homework 2

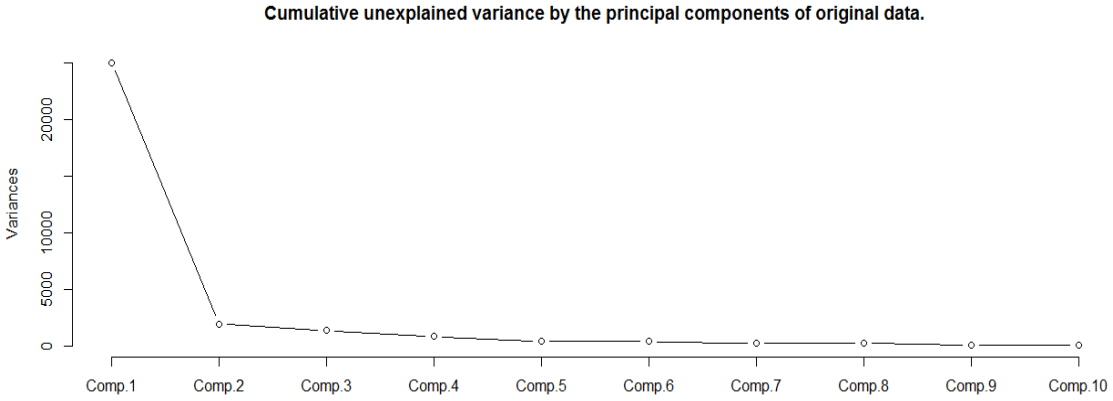
Hasini, Urala Liyanage Dona Gunasinghe

## 1 Counting

(a) Following is the summary of the principal component analysis on the data with 35 numeric attributes.

```
Importance of components:
                          Comp.1     Comp.2      Comp.3      Comp.4      Comp.5      Comp.6      Comp.7      Comp.8      Comp.9     Comp.10     Comp.11
Standard deviation     158.3208666 44.2440701 37.26238490 29.39718627 20.50120241 20.32685616 17.120982866 16.380360858 11.428692855 9.326342495 8.671547361
Proportion of Variance   0.8012007  0.0625713  0.04438194  0.02762336  0.01343457  0.01320704  0.009369629  0.008576537  0.004175016 0.002780275 0.002403578
Cumulative Proportion    0.8012007  0.8637720  0.90815398  0.93577735  0.94921192  0.96241895  0.971788583  0.980365120  0.984540136 0.987320411 0.989723989
                          Comp.12     Comp.13     Comp.14     Comp.15     Comp.16     Comp.17     Comp.18     Comp.19     Comp.20     Comp.21     Comp.22
Standard deviation     6.780620300 6.638590351 5.811934150 5.701641177 5.320085275 4.7046892950 4.4671920018 4.1061028004 3.7079725711 3.640107370 3.0359178765
Proportion of Variance 0.001469616 0.001408694 0.001079708 0.001039118 0.000904695 0.0007075008 0.0006378731 0.0005389204 0.0004394789 0.000423539 0.0002946084
Cumulative Proportion  0.991193605 0.992602300 0.993682008 0.994721126 0.995625821 0.9963333215 0.9969711946 0.9975101150 0.9979495940 0.998373133 0.9986677413
                          Comp.23      Comp.24      Comp.25      Comp.26      Comp.27      Comp.28      Comp.29      Comp.30      Comp.31      Comp.32
Standard deviation     3.025459181 2.5723657386 2.5197448872 2.3095050223 1.9263128189 1.714360e+00 1.4986860019 1.295265e+00 1.167128e+00 1.068947e+00
Proportion of Variance 0.000292582 0.0002115098 0.0002029449 0.0001704915 0.0001186093 9.394401e-05 0.0000717937 5.362683e-05 4.354138e-05 3.652394e-05
Cumulative Proportion  0.998960323 0.9991718331 0.9993747780 0.9995452696 0.9996638788 9.997578e-01 0.9998296166 9.998832e-01 9.999268e-01 9.999633e-01
                          Comp.33      Comp.34      Comp.35
Standard deviation     8.172108e-01 6.838136e-01 1.115767e-01
Proportion of Variance 2.134682e-05 1.494653e-05 3.979352e-07
Cumulative Proportion  9.999847e-01 9.999996e-01 1.000000e+00
```

(b) Following is the scree plot:

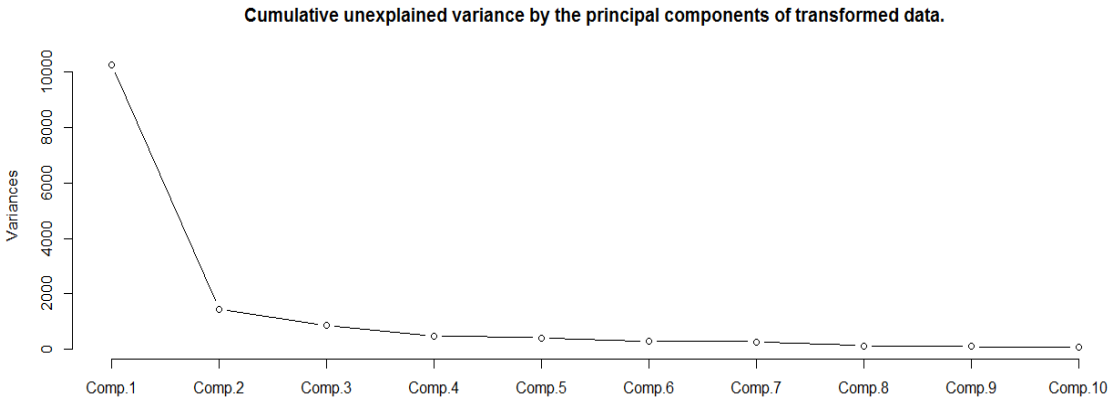



Accordingly, minimum six components are needed to explain more than 95% of data as the cumulative proportion of variance explained by the 6h component is 96.24%.

(c) By inspecting the loadings of the principal component matrix in R, only 11 attributes are shown with significant weights in the first principal component. The are: review_count, sun_noon_6, tue_6_mid, wed_6_mid, thu_noon_6, thu_6_mid, fri_noon_6, fri_6_mid, sat_noon_6, sat_6_mid, tip_count. Loadings matrix is shown below:

```
                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
stars
review_count    -0.797  0.548  0.184         -0.121
longitude                      -0.974
latitude                       -0.193
sun_mid_6                             0.148 -0.169
sun_6_noon                -0.205        -0.134                -0.219
sun_noon_6      -0.109 -0.132 -0.127         0.158 -0.154 -0.135 -0.141 -0.138                 0.309  -0.328   0.347
sun_6_mid              -0.142                                      -0.318 -0.111  0.258  -0.165  -0.144  -0.181   0.259
mon_mid_6                                    0.103 -0.117
mon_6_noon                -0.175        -0.158 -0.115 -0.110        -0.224 -0.112
mon_noon_6             -0.157                     -0.240 -0.141 -0.102                 -0.135   0.346          -0.221
mon_6_mid              -0.175  0.110                                -0.277         0.237  -0.268  -0.176   0.277  -0.532
tue_mid_6
tue_6_noon                -0.196        -0.190 -0.144 -0.127        -0.268 -0.115          -0.132
tue_noon_6             -0.175                     -0.257 -0.133                                0.337
tue_6_mid       -0.101 -0.196  0.112         -0.129                -0.184         0.149  -0.157           0.335   0.468
wed_mid_6                                    0.103 -0.110
wed_6_noon                -0.201        -0.192 -0.138 -0.122        -0.250 -0.113          -0.155
wed_noon_6             -0.188 -0.103               -0.255 -0.155                                0.244
wed_6_mid       -0.118 -0.224  0.135         -0.142                -0.102         0.144  -0.151  -0.131   0.174
thu_mid_6                                    0.169 -0.189
thu_6_noon                -0.252        -0.223 -0.121 -0.110        -0.238 -0.103          -0.235
thu_noon_6      -0.147 -0.307 -0.125               -0.258 -0.218  0.414  0.323  0.326   0.228  -0.284  -0.213  -0.153
thu_6_mid       -0.188 -0.390  0.244         -0.309         0.119  0.323  0.133          0.215           0.141   0.317
fri_mid_6                                    0.403 -0.460  0.119
fri_6_noon                -0.413        -0.322  0.187  0.203         0.159  0.143  -0.213          -0.202
fri_noon_6      -0.230 -0.201 -0.250         0.485  0.165  0.129  0.208         -0.237  -0.515  -0.331
fri_6_mid       -0.202 -0.271  0.267         -0.235  0.120  0.236  0.196 -0.300 -0.365           0.283  -0.271  -0.291
sat_mid_6                                    0.394 -0.446  0.113                                0.116
sat_6_noon                -0.417        -0.303  0.253  0.263         0.187  0.239           0.273
sat_noon_6      -0.204 -0.191 -0.274         0.407  0.212  0.266        -0.217          0.489   0.162   0.394
sat_6_mid       -0.122 -0.132  0.121               0.140         -0.267 -0.297  0.312   0.289  -0.135  -0.492
tip_count       -0.221 -0.141                     0.166 -0.160 -0.517  0.495 -0.516   0.218  -0.136  -0.109
liked_tip_count
likes
                Comp.15 Comp.16 Comp.17 Comp.18 Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24 Comp.25 Comp.26
stars
review_count
```

(d) Following is the scree plot for the transformed data:

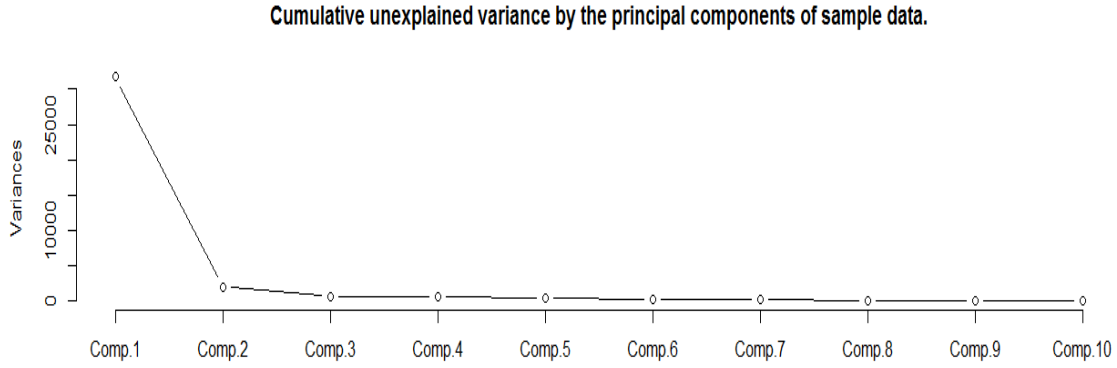



Now it takes minimum seven components to explain 95% of the data. And now there are 17 attributes shown with significant weights in the first principal component of the new loadings matrix. However, log of review count is not among them where as in original loadings matrix, review count was shown with a significant weight in the first principal component. New loadings matrix is shown below.

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 | Comp.13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| log_r_count | | | | | | | | | | | | | |
| stars | | | | | | | | | | | | | |
| longitude | | | -0.975 | | | | | | | | | | |
| latitude | | | -0.194 | | | | | | | | | | |
| sun_mid_6 | | | | -0.114 | | -0.184 | | | | | | | |
| sun_6_noon | | 0.201 | | | 0.140 | | | -0.228 | | | | | 0.109 |
| sun_noon_6 | -0.182 | | | 0.189 | -0.128 | | -0.203 | -0.140 | | | -0.164 | -0.392 | 0.406 |
| sun_6_mid | -0.160 | -0.131 | | | | 0.137 | -0.257 | | 0.265 | 0.183 | -0.196 | | 0.304 |
| mon_mid_6 | | | | | | -0.122 | | | | | | | |
| mon_6_noon | | 0.161 | | | 0.176 | | | -0.233 | | | | 0.112 | |
| mon_noon_6 | -0.153 | | | 0.275 | | | -0.171 | | | 0.131 | 0.155 | -0.311 | -0.219 |
| mon_6_mid | -0.168 | -0.158 | | | | 0.121 | -0.218 | | 0.265 | 0.279 | 0.115 | 0.246 | -0.526 |
| tue_mid_6 | | | | | | | | | | | | | |
| tue_6_noon | | 0.178 | | 0.127 | 0.212 | | | -0.274 | | | | 0.155 | |
| tue_noon_6 | -0.154 | | | 0.304 | | | -0.139 | | | | 0.153 | -0.305 | |
| tue_6_mid | -0.173 | -0.168 | | | 0.130 | | -0.121 | | 0.222 | 0.159 | 0.281 | 0.130 | 0.354 |
| wed_mid_6 | | | | | | -0.119 | | | | | | | |
| wed_6_noon | | 0.184 | | 0.121 | 0.214 | | | -0.258 | | | | 0.180 | |
| wed_noon_6 | -0.169 | | | 0.318 | | -0.111 | | | | | | -0.228 | -0.117 |
| wed_6_mid | -0.202 | -0.198 | | | 0.140 | | | | 0.186 | 0.156 | 0.105 | 0.178 | |
| thu_mid_6 | | | | -0.128 | | -0.212 | | | | | | | |
| thu_6_noon | | 0.237 | | 0.104 | 0.241 | | | -0.250 | | -0.104 | | 0.236 | |
| thu_noon_6 | -0.261 | | | 0.431 | | -0.331 | 0.266 | 0.318 | 0.161 | -0.193 | -0.436 | 0.190 | -0.153 |
| thu_6_mid | -0.324 | -0.357 | | | 0.297 | | 0.372 | 0.137 | | -0.227 | 0.209 | | 0.352 |
| fri_mid_6 | | | | -0.292 | | -0.545 | | | | | | | |
| fri_6_noon | -0.131 | 0.425 | | -0.213 | 0.285 | 0.105 | 0.161 | 0.175 | 0.109 | 0.224 | -0.182 | | |
| fri_noon_6 | -0.375 | 0.182 | | | -0.506 | | 0.249 | | -0.258 | 0.499 | | 0.303 | |
| fri_6_mid | -0.329 | -0.322 | | -0.182 | 0.209 | 0.132 | 0.270 | -0.341 | -0.375 | | -0.113 | -0.353 | -0.278 |
| sat_mid_6 | | | | -0.285 | | -0.531 | | | | | | | |
| sat_6_noon | -0.143 | 0.425 | | -0.260 | 0.254 | 0.142 | 0.183 | 0.224 | 0.259 | | | -0.268 | |
| sat_noon_6 | -0.336 | 0.205 | | | -0.437 | 0.189 | 0.109 | -0.169 | 0.277 | -0.494 | 0.395 | | -0.106 |
| sat_6_mid | -0.196 | -0.143 | | -0.173 | | 0.108 | -0.210 | -0.271 | 0.280 | -0.258 | -0.550 | | |
| tip_count | -0.349 | | | -0.237 | | | -0.521 | 0.462 | -0.500 | -0.251 | | 0.103 | |
| liked_tip_count | | | | | | | | | | | | | |
| likes | | | | | | | | | | | | | |

| | Comp.14 | Comp.15 | Comp.16 | Comp.17 | Comp.18 | Comp.19 | Comp.20 | Comp.21 | Comp.22 | Comp.23 | Comp.24 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| log_r_count | | | | | | | | | | | |

(e) Analysis on a sample of size 100:
The minimum number of components to explain 95% of data is: 3.
The scree plot is shown below:




Eight attributes have significant weights in the first principal component as shown below. Accordingly, the set of attributes which had significant weights in the first principal component w.r.t the sample dataset is a subset of that of the whole dataset, which is interesting.

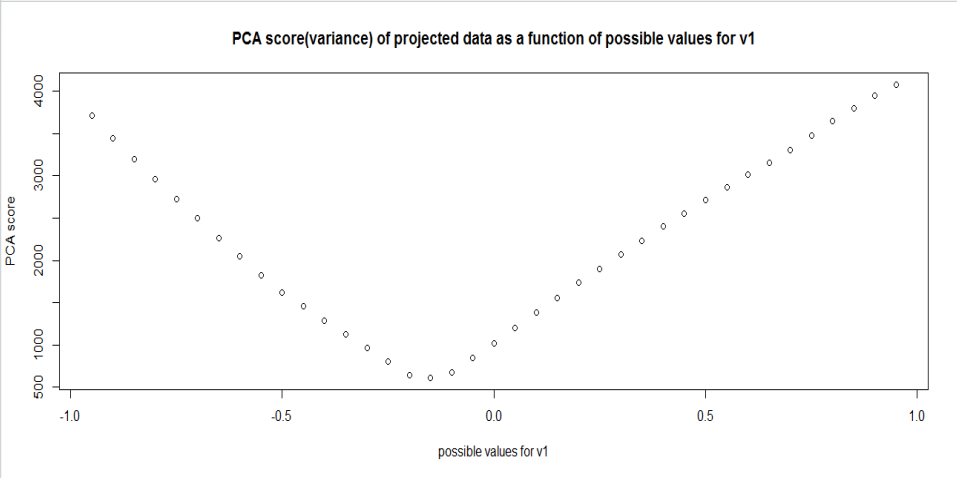
```
Loadings:
                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
stars
review_count    -0.883 -0.306 -0.250
longitude                             -0.974
latitude                              -0.192
sun_mid_6
sun_6_noon                    -0.185         -0.125
sun_noon_6              0.147  0.117         -0.272
sun_6_mid                      0.258
mon_mid_6
mon_6_noon              0.123 -0.222
mon_noon_6              0.260
mon_6_mid                      0.215
tue_mid_6
tue_6_noon              0.135 -0.234         -0.186
tue_noon_6              0.234
tue_6_mid                      0.198
wed_mid_6
wed_6_noon              0.123 -0.239         -0.122
wed_noon_6              0.223 -0.111         -0.300
wed_6_mid                      0.142
thu_mid_6
thu_6_noon              0.157 -0.282         -0.189
thu_noon_6      -0.120  0.359               -0.455
thu_6_mid       -0.142         0.298         -0.142
fri_mid_6
fri_6_noon              0.169 -0.334          0.350
fri_noon_6      -0.153  0.383                 0.238
fri_6_mid       -0.181 -0.203  0.138         -0.189
sat_mid_6
sat_6_noon              0.108 -0.181          0.335
sat_noon_6      -0.148  0.367  0.326          0.169
sat_6_mid       -0.108         0.253
tip_count       -0.205  0.369  0.158          0.332
liked_tip_count
likes
```

## 2 Scoring and Search

(a) Eigen vector values (component weights) in the solution returned by R is shown below:

```
Loadings:
             Comp.1 Comp.2
review_count -0.968  0.251
tip_count    -0.251 -0.968
```

(b) Plot of the PCA score (variance) of the projected data as a function of possible values for v1:




The solution with the best score (i.e:maximum variance of 4073.6674) is: v1=0.95 and v2=0.3122499.
Comparing it with the solution returned by R, we can observe that the **magnitudes** of the weights in principal component 1 are approximately equal in both the solutions. However, they have **opposite signs**.
One reason is that we searched only in the interval of [-0.95, +0.95], for v1 whereas the weight found by R for v1 is out of that interval. I tried making the interval [-1.0, +0.95] in which case I got -1 as the best weight for v1.
Likewise, if we extended our search space, we could have got values which are close to the values returned by R. These observations signals that the this algorithm is able to find the optimal solution for principal component vectors.

## 3 Transformations and Associations

(a) Top 30 categories:

```
 [1] "Restaurants"          "Mexican"              "American (Traditional)" "Fast Food"          "Pizza"          "Sandwiches"
 [7] "Nightlife"            "Bars"                 "Food"                   "American (New)"     "Italian"        "Chinese"
[13] "Burgers"              "Breakfast & Brunch"   "Japanese"               "Delis"              "Sushi Bars"     "Steakhouses"
[19] "Seafood"              "Sports Bars"          "Cafes"                  "Buffets"            "Barbeque"       "Thai"
[25] "Coffee & Tea"         "Mediterranean"        "Chicken Wings"          "Asian Fusion"       "Pubs"           "Indian"
```

(b) Top 30 cities: (I used each unique string as a distinct city. i.e: Edinburgh and City of Edinburgh are two different cities.)

```
 [1] "Las Vegas"      "Phoenix"        "Edinburgh"        "Scottsdale"       "Mesa"
 [6] "Madison"        "Tempe"          "Henderson"        "Chandler"         "Glendale"
[11] "Gilbert"        "Peoria"         "North Las Vegas"  "Surprise"         "Goodyear"
[16] "Waterloo"       "Avondale"       "Kitchener"        "Queen Creek"      "Middleton"
[21] "Cave Creek"     "Casa Grande"    "Fountain Hills"   "Apache Junction"  "Buckeye"
[26] "Sun Prairie"    "Fitchburg"      "Maricopa"         "Monona"           "Sun City"
> |
```

(c) Top five feature combinations with largest $\chi^2$ scores along with the p-values:

| | attr | score | pval |
|---|---|---|---|
| Coffee & Tea_Edinburgh | | 5.344245e+02 | 3.078803e-118 |
| Indian_Edinburgh | | 2.864583e+02 | 2.939462e-64 |
| Food_Edinburgh | | 1.273845e+02 | 1.530541e-29 |
| American (New)_Scottsdale | | 1.207110e+02 | 4.420669e-28 |
| Mexican_Edinburgh | | 1.179061e+02 | 1.817955e-27 |

I think the associations among the above feature combinations can be expected.
All the above feature pairs with the highest $\chi^2$ scores, seem to have the lowest p-values which are below the significance level (0.05). That means the null hypothesis of independence between the two features in each feature pair can be rejected.
Since Edinburgh is the capital of Scotland and the second most populous city in Scotland, it can be expected that there is an association between the city and the different types of restaurants in there.
However, looking at the following contingency tables related to each of the above feature pairs, it is hard to infer a reason behind the order of the $\chi^2$ scores.

| | Edinburgh | |
|---|---|---|
| Coffee & Tea | 0 | 1 |
| 0 | 12982 | 900 |
| 1 | 182 | 127 |

| Indian | Edinburgh 0 | 1 |
|---|---|---|
| 0 | 13039 | 950 |
| 1 | 125 | 77 |

| Food | Edinburgh 0 | 1 |
|---|---|---|
| 0 | 12183 | 847 |
| 1 | 981 | 180 |

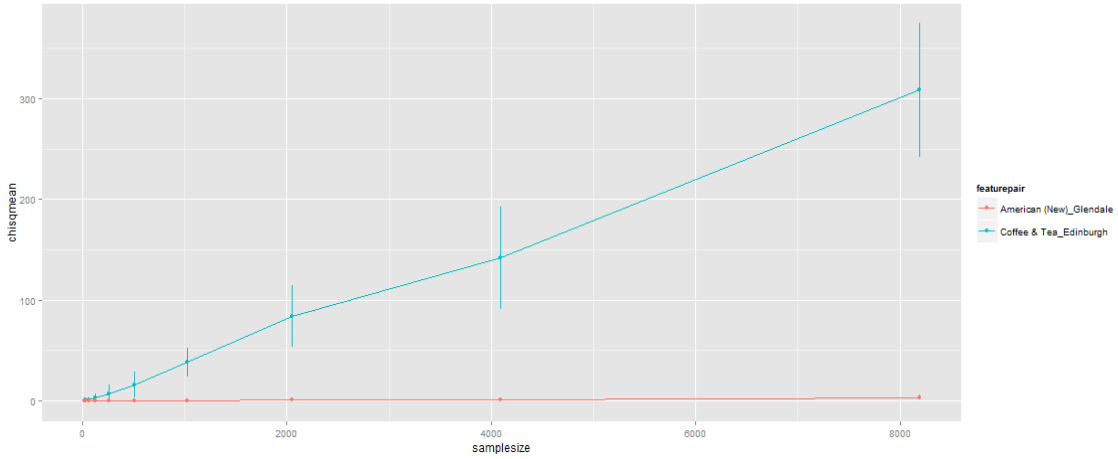
| American (New) | Scottsdale 0 | 1 |
|---|---|---|
| 0 | 12297 | 851 |
| 1 | 880 | 163 |

| Mexican | Edinburgh 0 | 1 |
|---|---|---|
| 0 | 11450 | 1012 |
| 1 | 1714 | 15 |

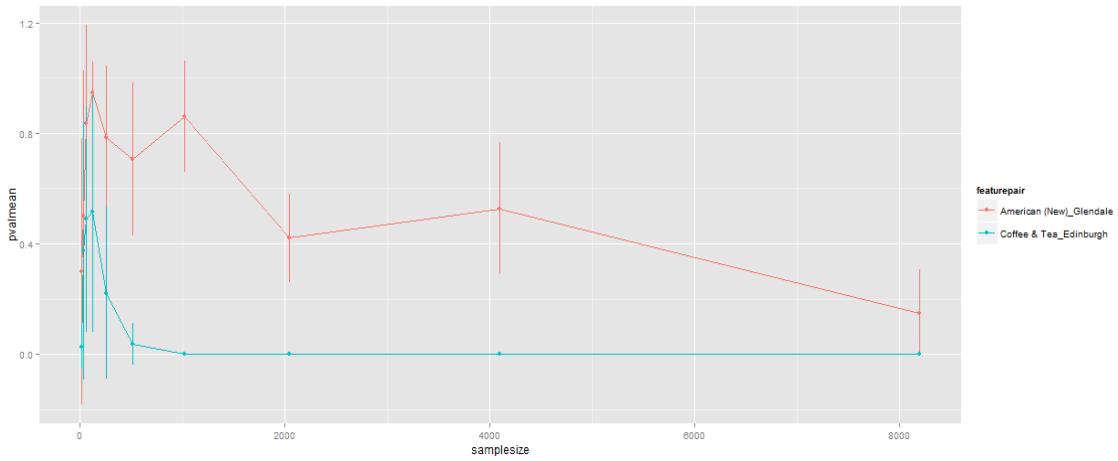
(d) Feature pair with the largest $\chi^2$ score: **Coffee & Tea, Edinburgh**
Feature pair with a score that is barely significant (selected from the following section of the $\chi^2$ score table): **American (New), Glendale**

| | | |
|---|---|---|
| Mexican_Casa Grande | 3.952205e+00 | 4.680998e-02 |
| Italian_Cave Creek | 3.951397e+00 | 4.683246e-02 |
| Coffee & Tea_Henderson | 3.910551e+00 | 4.798387e-02 |
| American (Traditional)_Cave Creek | 3.904881e+00 | 4.814604e-02 |
| Bars_Mesa | 3.898307e+00 | 4.833478e-02 |
| American (New)_Glendale | 3.827959e+00 | 5.040427e-02 |
| Coffee & Tea_Maricopa | 3.652188e+00 | 5.599562e-02 |
| Burgers_Chandler | 3.650391e+00 | 5.605604e-02 |
| Indian_Scottsdale | 3.638907e+00 | 5.644398e-02 |
| Chinese_Chandler | 3.634144e+00 | 5.660569e-02 |
| Pubs_Mesa | 3.513621e+00 | 6.086628e-02 |

Plot of the mean $\chi^2$ scores as a function of the sample size with error bars including standard deviation, for the aforementioned two feature pairs:




Plot of the mean p values as a function of the sample size with error bars including standard deviation, for the aforementioned two feature pairs:




$\chi^2$ score of (Coffee & Tea, Edinburgh) feature pair which has the largest score w.r.t the whole dataset, grows almost linearly with the increase of the sample size where as the $\chi^2$ score of (American (New), Glendale) feature pair which has a barely significant p-value w.r.t the whole dataset, stays almost constant with the increase of the sample size. Therefore, the effect of sample size on the $\chi^2$ score of the two feature pairs are different.

Looking at the variation of the p-values of $\chi^2$ statistic of the two feature pairs with the sample size, we can conclude that the p values of the $\chi^2$ statistic of both feature pairs vary with the sample size.

However, the sample size directly impacts the decision of rejecting or not rejecting the null hypothesis of independence between the two features only in the case of first feature pair: (Coffee & Tea, Edinburgh).
For example, if we analyzed only a sample of size below 1000, we would not have rejected the null hypothesis that those two features are independent, because the p-values of the $\chi^2$ statistic is higher than 0.05 in such samples.

Another difference is that the variation of the p-value of the $\chi^2$ statistic of the feature pair: (Coffee & Tea, Edinburgh) gets lower than that of the feature pair: (American (New), Glendale), as the increase of the sample size.

# 4 Identifying hypothesis

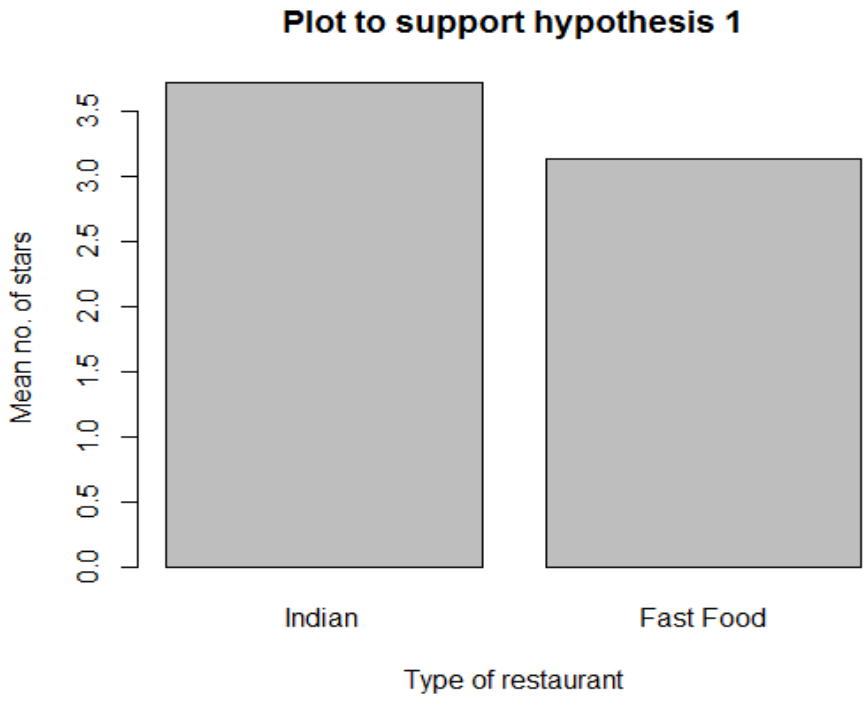
(a) Hypothesis 1: Indian restaurants gets more stars than fast food restaurants as Indian food is of more quality, authentic and healthy.
This is a directional, relational and causal hypothesis. However, we have data only to analyze the direction and the relationship, but not the cause. Cause was mentioned from the domain knowledge.

Hypothesis 2: Restaurants in Edinburgh gets more reviews than the restaurants in Waterloo because Edinburgh is a capital of a country with more population than Waterloo which is a suburb in Pheonix in Arizona state.
This is a directional relational causal hypothesis. In this case also, the dataset provides data only to analyze the direction and the relationship, but not the cause. Cause was mentioned from the domain knowledge.

(b) How the analysis of data led to the conjecture:
From the analysis of HW1, we observed that there is no interesting relationship between longitude/latitude and stars. And from the box plots drawn for part (g) of the last question of HW1, we observed there is a relationship between the no. of stars and the type of the restaurant. Also, from the analysis in the question 2 of HW2, we observed that there can be an association between the city and the categories of the restaurants. From those observations and the reasoning I gave to those previous observations, I came to form these two hypothesis, which I would like to investigate further with proper hypothesis testing.

(c) Following are two basic plots to support the two hypothesis. In order to come to a conclusion (whether to reject or not reject the hypothesis), we need to do a proper hypothesis testing.

Plot for hypothesis 1:



Plot for hypothesis 2: