CS57300: Homework 2

Due date: Wednesday February 18, midnight (submit pdf to Blackboard)

Submit both your answers to the questions and the code that you used for analysis. Your homework must be typed. Use of Latex is recommended, but not required.

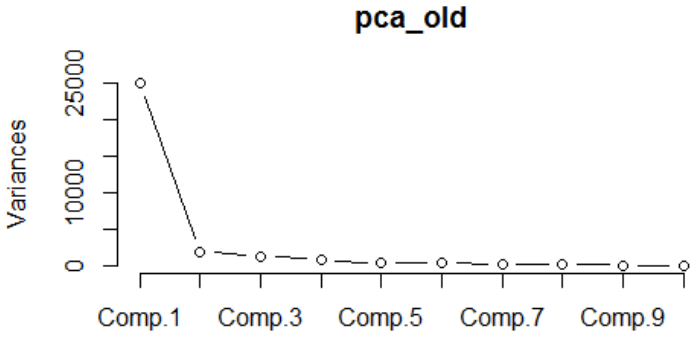
In this assignment, you will use R (and optionally python) to explore, transform, and analyze the Yelp data you started to use in HW1. Based on your analysis you will formulate hypotheses about the data.

1 Principal Component Analysis (6 pts)

Consider the subset of the Yelp data comprised of the 35 numeric attributes.

(a) Run principal component analysis on the data.

(b) Plot the scree plot. Identify what number of components are needed to explain more than 95% of the variance in the data.




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

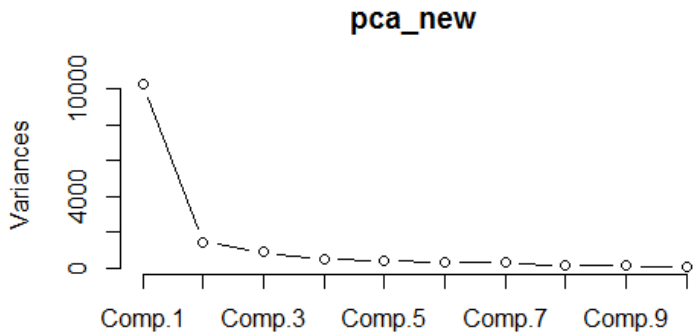
We can conclude that 6 components are needed to explain 95% of the variance.

(c) Inspect the weights for the first principal component and identify how many of the 35 attributes have a significant weight in this component.

```
> pca_old$loadings[,1]
         stars   review_count      longitude       latitude     sun_mid_6
-0.0006576914 -0.7974495627  0.0231482226  0.0038839967 -0.0132052832
    sun_6_noon    sun_noon_6     sun_6_mid     mon_mid_6    mon_6_noon
-0.0463759890 -0.1091093423 -0.0965695341 -0.0089152353 -0.0358196247
   mon_noon_6     mon_6_mid     tue_mid_6     tue_6_noon    tue_noon_6
-0.0879996510 -0.0994463461 -0.0074330786 -0.0356203098 -0.0868282128
    tue_6_mid     wed_mid_6    wed_6_noon     wed_noon_6     wed_6_mid
-0.1009176419 -0.0093340198 -0.0382205510 -0.0957347170 -0.1181184198
    thu_mid_6    thu_6_noon    thu_noon_6     thu_6_mid     fri_mid_6
-0.0149806440 -0.0531068490 -0.1471675898 -0.1877125929 -0.0367921921
   fri_6_noon    fri_noon_6     fri_6_mid     sat_mid_6    sat_6_noon
-0.0860707663 -0.2297316476 -0.2023235253 -0.0363321748 -0.0925543787
   sat_noon_6     sat_6_mid     tip_count liked_tip_count         likes
-0.2035730581 -0.1219815318 -0.2205086570 -0.0025775279 -0.0027697922
```

if we define significant weights is greater than 0.1, there are 11 attributes have significant weights.

(d) Transform the data by removing the original column for *review_count* and replace it with a new column containing log-transformed values of *review_count*. Repeat the above analysis and discuss what if any changes you see in the results.



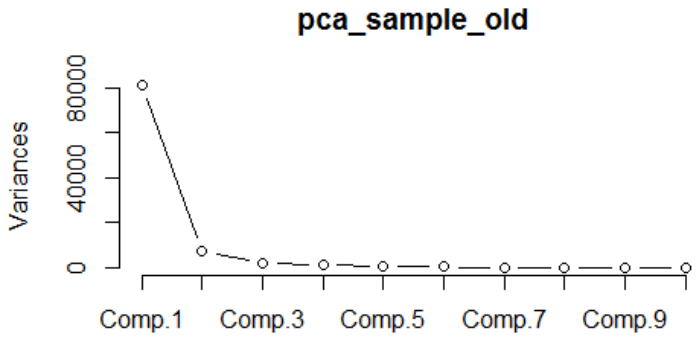

```
> summary(pca_new)
Importance of components:
                          Comp.1      Comp.2      Comp.3      Comp.4      Comp.5
Standard deviation     101.3239451 38.02957360 29.43999525 21.95453349 20.49728662
Proportion of Variance   0.6982976  0.09836923  0.05895109  0.03278422  0.02857651
Cumulative Proportion    0.6982976  0.79666684  0.85561793  0.88840214  0.91697865
                           Comp.6     Comp.7      Comp.8      Comp.9     Comp.10
Standard deviation     17.38528387 16.7095013 11.471300894 10.066793963 8.676868470
Proportion of Variance  0.02055796  0.0189908  0.008950385  0.006892848 0.005120854
Cumulative Proportion   0.93753661  0.9565274  0.965477795  0.972370644 0.977491498
                          Comp.11     Comp.12     Comp.13     Comp.14    Comp.15
Standard deviation     7.197203665 6.769874231 5.842470924 5.707321270 5.33042364
Proportion of Variance 0.003523254 0.003117293 0.002321718 0.002215547 0.00193259
Cumulative Proportion  0.981014752 0.984132045 0.986453763 0.988669310 0.99060190
                          Comp.16     Comp.17     Comp.18     Comp.19     Comp.20
Standard deviation     4.71408460 4.472099287 4.126835215 3.7079786749 3.6402633519
Proportion of Variance 0.00151151 0.001360314 0.001158379 0.0009351706 0.0009013262
Cumulative Proportion  0.99211341 0.993473725 0.994632104 0.9955672744 0.9964686005
                          Comp.21      Comp.22      Comp.23      Comp.24      Comp.25
Standard deviation     3.0376713062 3.0285005439 2.5877338923 2.5197544708 2.3096721156
Proportion of Variance 0.0006276217 0.0006238378 0.0004554656 0.0004318499 0.0003628416
Cumulative Proportion  0.9970962222 0.9977200600 0.9981755256 0.9986073755 0.9989702171
                          Comp.26      Comp.27      Comp.28      Comp.29      Comp.30
Standard deviation     1.927057884 1.7144886840 1.4994986022 1.2978393309 1.167236e+00
Proportion of Variance 0.000252584 0.0001999335 0.0001529356 0.0001145667 9.266878e-05
Cumulative Proportion  0.999222801 0.9994227346 0.9995756702 0.9996902369 9.997829e-01
```

```
> pca_new$loadings[,1]
         stars    review_count       longitude        latitude      sun_mid_6
-0.0009674105 -0.0078097777  0.0425589619  0.0076541379 -0.0207761585
    sun_6_noon     sun_noon_6      sun_6_mid      mon_mid_6     mon_6_noon
-0.0723556246 -0.1823487949 -0.1597903188 -0.0144290950 -0.0585936882
    mon_noon_6      mon_6_mid      tue_mid_6     tue_6_noon     tue_noon_6
-0.1533033126 -0.1680012177 -0.0123201054 -0.0594824406 -0.1538128742
     tue_6_mid      wed_mid_6     wed_6_noon     wed_noon_6      wed_6_mid
-0.1729564831 -0.0152939575 -0.0632078126 -0.1689259463 -0.2015364946
     thu_mid_6     thu_6_noon     thu_noon_6      thu_6_mid      fri_mid_6
-0.0239049456 -0.0858006049 -0.2611471902 -0.3237403023 -0.0589688973
    fri_6_noon     fri_noon_6      fri_6_mid      sat_mid_6     sat_6_noon
-0.1307646848 -0.3749014825 -0.3288140224 -0.0589312826 -0.1427036971
    sat_noon_6      sat_6_mid      tip_count liked_tip_count          likes
-0.3364539715 -0.1956675587 -0.3493698989 -0.0040907475 -0.0044044651
```

At this time, there are 7 components to explain 95% of the variance and 17 attributes have significant weights. The change is that the *review_count* doesn't dominant variance after using log value. The weight is change from 0.79 to 0.07.

(e) Sample a random set of 100 examples from the original data. Repeat the above analysis and discuss what if any changes you see in the results.

(a) 1 and 2




```
> summary(pca_sample_old)
Importance of components:
                          Comp.1      Comp.2      Comp.3     Comp.4       Comp.5
Standard deviation     284.6263644 86.37326852 45.70639490 33.3203433 19.302719955
Proportion of Variance   0.8731545  0.08040806  0.02251618  0.0119663  0.004015854
Cumulative Proportion    0.8731545  0.95356260  0.97607879  0.9880451  0.992060940
                           Comp.6      Comp.7      Comp.8      Comp.9      Comp.10
Standard deviation     15.395116482 13.948723654 9.2622740329 8.0327142186 6.220990460
Proportion of Variance  0.002554505  0.002097055 0.0009246473 0.0006954493 0.000417119
Cumulative Proportion   0.994615445  0.996712500 0.9976371473 0.9983325966 0.998749716
                           Comp.11      Comp.12      Comp.13      Comp.14      Comp.15
Standard deviation     5.6056496215 4.3483195960 3.9903946931 3.031736e+00 2.647807e+00
Proportion of Variance 0.0003386825 0.0002037904 0.0001716218 9.906576e-05 7.556375e-05
Cumulative Proportion  0.9990883981 0.9992921885 0.9994638104 9.995629e-01 9.996384e-01
                           Comp.16      Comp.17      Comp.18      Comp.19      Comp.20
Standard deviation     2.429867e+00 2.244382e+00 1.933649e+00 1.847583e+00 1.730489e+00
Proportion of Variance 6.363643e-05 5.429184e-05 4.029915e-05 3.679162e-05 3.227591e-05
Cumulative Proportion  9.997021e-01 9.997564e-01 9.997967e-01 9.998335e-01 9.998657e-01
                           Comp.21      Comp.22      Comp.23      Comp.24      Comp.25
Standard deviation     1.608412e+00 1.516112e+00 1.466197e+00 1.1555977332 1.034928e+00
Proportion of Variance 2.788274e-05 2.477441e-05 2.316999e-05 0.0000143931 1.154413e-05
Cumulative Proportion  9.998936e-01 9.999184e-01 9.999416e-01 0.9999559551 9.999675e-01
```

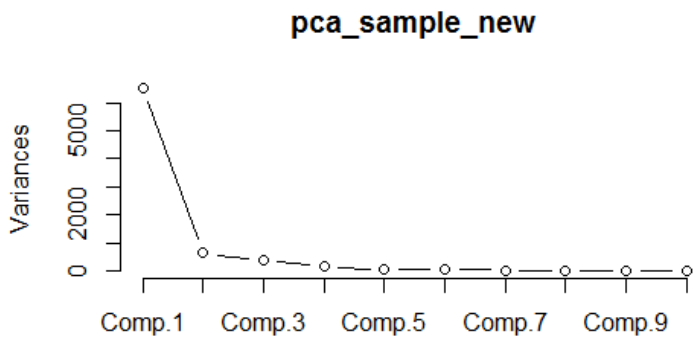
We can conclude that 2 components are needed to explain 95% of the variance.

(b) 3

```
> pca_sample_old$loadings[,1]
        stars  review_count     longitude      latitude    sun_mid_6
-0.0003893774 -0.8849626281  0.0122888210  0.0015867356 -0.0027656102
   sun_6_noon    sun_noon_6     sun_6_mid     mon_mid_6    mon_6_noon
-0.0775924531 -0.0838094947 -0.0529273409 -0.0019125054 -0.0516884458
   mon_noon_6     mon_6_mid     tue_mid_6     tue_6_noon    tue_noon_6
-0.0588093255 -0.0481380621 -0.0021781674 -0.0482673213 -0.0491387557
    tue_6_mid     wed_mid_6     wed_6_noon     wed_noon_6     wed_6_mid
-0.0532412616 -0.0023199700 -0.0535145643 -0.0521689541 -0.0520767599
    thu_mid_6     thu_6_noon     thu_noon_6     thu_6_mid     fri_mid_6
-0.0040786672 -0.0850317229 -0.0920574255 -0.1018033503 -0.0082213599
   fri_6_noon    fri_noon_6     fri_6_mid     sat_mid_6    sat_6_noon
-0.1575767702 -0.1486152166 -0.1207471941 -0.0095389373 -0.1397439621
   sat_noon_6     sat_6_mid     tip_count liked_tip_count         likes
-0.1707780194 -0.0615084042 -0.1904661343 -0.0007820739 -0.0007482548
```

There are 7 attributes have significant weight.

(c) 4




```
> pca_sample_new$loadings[,1]
        stars  review_count     longitude      latitude    sun_mid_6
-0.0009149156 -0.0104438522  0.0448168658  0.0062953562 -0.0159052031
   sun_6_noon    sun_noon_6     sun_6_mid     mon_mid_6    mon_6_noon
-0.0354477376 -0.1843199035 -0.1790395301 -0.0123246436 -0.0355811672
   mon_noon_6     mon_6_mid     tue_mid_6     tue_6_noon    tue_noon_6
-0.1563758081 -0.2051985884 -0.0153112348 -0.0217037158 -0.1391529756
    tue_6_mid     wed_mid_6     wed_6_noon     wed_noon_6     wed_6_mid
-0.2104683084 -0.0140466930 -0.0288925430 -0.1423851660 -0.1866404694
    thu_mid_6     thu_6_noon     thu_noon_6     thu_6_mid     fri_mid_6
-0.0233644865 -0.0256997056 -0.1569776186 -0.4130851318 -0.0416087402
   fri_6_noon    fri_noon_6     fri_6_mid     sat_mid_6    sat_6_noon
-0.0380103385 -0.2403786844 -0.4801310609 -0.0457915247 -0.0522541424
   sat_noon_6     sat_6_mid     tip_count liked_tip_count         likes
-0.2371612620 -0.2682245711 -0.3475604832 -0.0039679860 -0.0040782800
> summary(pca_sample_new)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     80.7154142 25.0983736 19.92578795 11.95271672 7.56184646
Proportion of Variance  0.8215752  0.0794375  0.05006859  0.01801638 0.00721091
Cumulative Proportion   0.8215752  0.9010127  0.95108133  0.96909771 0.97630862
                           Comp.6     Comp.7     Comp.8     Comp.9     Comp.10
Standard deviation     6.327219830 5.620188337 4.654751299 4.344325493 3.997188561
Proportion of Variance 0.005048475 0.003983237 0.002732293 0.002380012 0.002014854
Cumulative Proportion  0.981357095 0.985340332 0.988072625 0.990452637 0.992467491
                          Comp.11     Comp.12     Comp.13     Comp.14     Comp.15
Standard deviation     3.013973911 2.926614819 2.5904915218 2.4213638671 2.2118046071
Proportion of Variance 0.001145548 0.001080104 0.0008462501 0.0007393575 0.0006169186
Cumulative Proportion  0.993613039 0.994693143 0.9955393933 0.9962787508 0.9968956695
                          Comp.16     Comp.17     Comp.18     Comp.19     Comp.20
Standard deviation     2.0831334498 1.8222287259 1.7860986313 1.5228825185 1.4590566931
Proportion of Variance 0.0005472283 0.0004187359 0.0004022956 0.0002924605 0.0002684595
Cumulative Proportion  0.9974428978 0.9978616337 0.9982639292 0.9985563897 0.9988248492
                          Comp.21     Comp.22     Comp.23     Comp.24     Comp.25
Standard deviation     1.4402042891 1.3269419105 1.0559405377 0.9841261427 0.9464797904
Proportion of Variance 0.0002615668 0.0002220436 0.0001406091 0.0001221338 0.0001129684
Cumulative Proportion  0.9990864159 0.9993084595 0.9994490686 0.9995712024 0.9996841708
                          Comp.26     Comp.27     Comp.28     Comp.29     Comp.30
```

At this time, there are 4 components to explain 95% of the variance and 15 attributes have significant weights. The change is that the *review_count* doesn't dominant variance after using log value. The weight is change from 0.88 to 0.01. The sample result will take less components to explain most of difference.

2 Scoring and search (12 pts)

Consider the subset of the Yelp data with only the *review_count* and *tip_count* attributes.

(a) Run principal component analysis on the data. Report the eigenvector values (i.e., component weights) in the solution returned by R.

```
> princomp(newdata)
Call:
princomp(x = newdata)

Standard deviations:
   Comp.1    Comp.2
132.95500  18.14155

 2  variables and  14191 observations.
> loadings(pca)

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

               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
```

The eigen vactor values will be (-0.96,-0.251)

(b) Develop your own algorithm to search over possible eigenvector solutions.
Recall that solutions must be orthogonal vectors of norm 1. Since the p^{th} dimension is constrained by the solutions for the $[1, p-1]$ principal components, and your data for this question is 2-dimensional, you will only need to search for the values in first eigenvector. Moreover, since the eigenvector must have a norm of 1, you will only need to search over the first value for the eigenvector.

- Mean center your data.
- Consider a grid search over [-0.95,+0.95] with a step-size of 0.05 for the first eigenvector value (i.e., for *review_count*, let's call this v_1).
- For each possible value of v_1, calculate a positive value for v_2 (i.e., for *tip_count*) that constrains the vector $[v_1, v_2]$ to have a norm of 1. (Note that searching over positive and negative values for v_1 and only positive values for v_2 will cover all directions.)
- For each choice of $[v_1, v_2]$, project the mean-centered data onto the vector and calculate the PCA score function (i.e., the variance of the projected data).
- Plot the score as a function of v_1 and identify the solution with the best score. Compare it to the solution returned by R and discuss any differences.

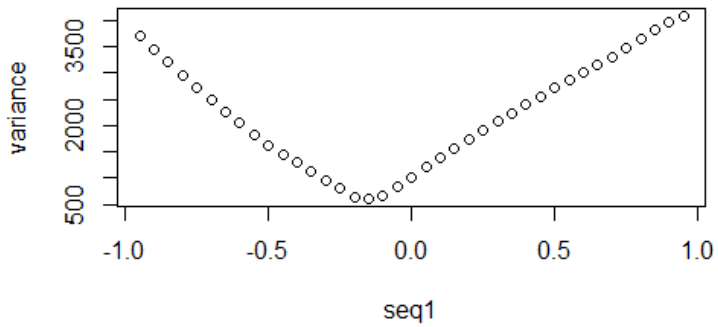
```
> princomp(newdata)
Call:
princomp(x = newdata)

Standard deviations:
   Comp.1    Comp.2
132.95500  18.14155

 2  variables and  14191 observations.
> loadings(pca)

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

               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
```



According the graph, we can find that find the best score is at v1=0.95, which means the basis vector is at b=(0.95,0.312), close to result from pca by R. The difference is only the direction, but it doesn't matter.

3 Transformations and associations (16 pts)

Consider the binary feature construction that you did in HW1 (e.g., Nightlife vs. not-Nightlife). In this question, you will construct binary features for values in the *category* and *city* attributes.

(a) Extract all the unique values in the *category* attribute by parsing the comma-separated

lists (e.g., "`Mexican, Restaurants`" → two values, one for `Mexican` and one for `Restaurants`). Sort the list of values and choose the top 30. Construct binary features for each of these 30. (Note: you should figure out how to do this in a loop or a function, do not do it manually!)

(b) Repeat the same process of binary feature construction for the *city* attribute, but this time use the top 30 most frequent cities in the data (i.e., reverse sort by number of examples in the city). Note: you do not need to parse this attribute.

(c) For each pair of binary features (*category* vs. *city*; 30×30 pairs), determine whether there is any association by calculating χ^2 scores (using `chisq.test`) from a contingency table of counts, e.g.:

	City i	
Category j	0	1
0	N_{00}	N_{01}
1	N_{10}	N_{11}

Report the top five features combinations with the largest χ^2 scores, along with assessments of significance (i.e., p values), and discuss whether the correlations are interesting or expected, given your domain knowledge.

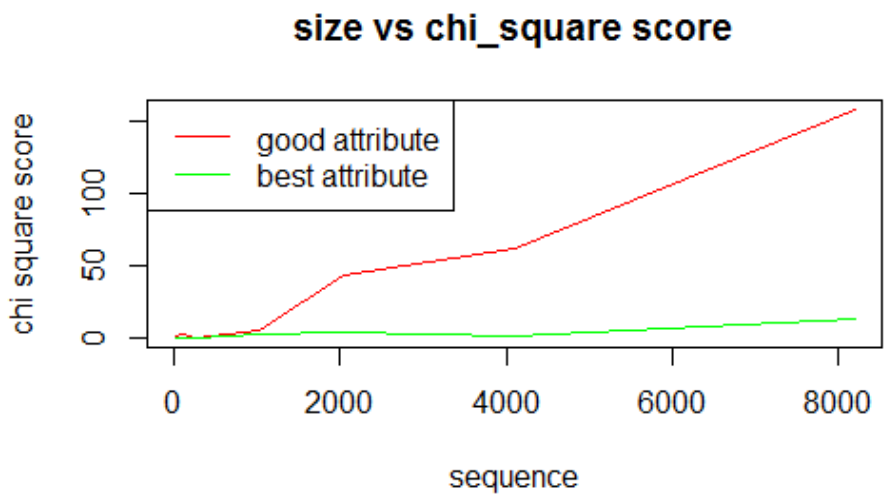
```
> source('~/top5.R')
city:  Edinburgh
category:  Indian
chi_square:  286.4583
p:  2.939462e-64
city:  Edinburgh
category:  Mexican
chi_square:  117.9061
p:  1.817955e-27
city:  Edinburgh
category:  Turkish
chi_square:  65.36709
p:  6.216837e-16
city:  Edinburgh
category:  Pizza
chi_square:  51.95515
p:  5.678296e-13
city:  Las Vegas
category:  Korean
chi_square:  51.34228
p:  7.758637e-13
```

According the result, we can conclude that Edinburgh has many food restaurant. Among the top 5 pairs, 4 of them is city Edinburgh.

(d) Consider the feature pair with largest χ^2 score (let's call this pair A^{max}) and another feature pair with a score that is barely significant (i.e., A^{good} with p-value ≈ 0.05). Investigate the effect of sampling on the scores of these feature pairs.

- Repeat ten times:
  - Create ten random samples of the following sizes: $[16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192]$.

  - Calculate the χ^2 scores for A^{max} and A^{good} on each sample.
- Calculate the mean and standard deviation of the scores for each feature pair, for each sample size.
- Plot the χ^2 scores as a function of sample size. Your plot should include one curve for A^{max} and one curve for A^{good} and include error bars to show the standard deviation.
- Discuss the results. What effect does sample size have on significance? Does the effect vary across the two attributes?



The chi square is from 0 to 9.4 for good attribute and form 0 to 149 for best attribute pair. According chart, we can conclude that the chi square will be more significant as the size of sample increased. The effect is similar across two attribute.

4 Identifying hypotheses (6 pts)

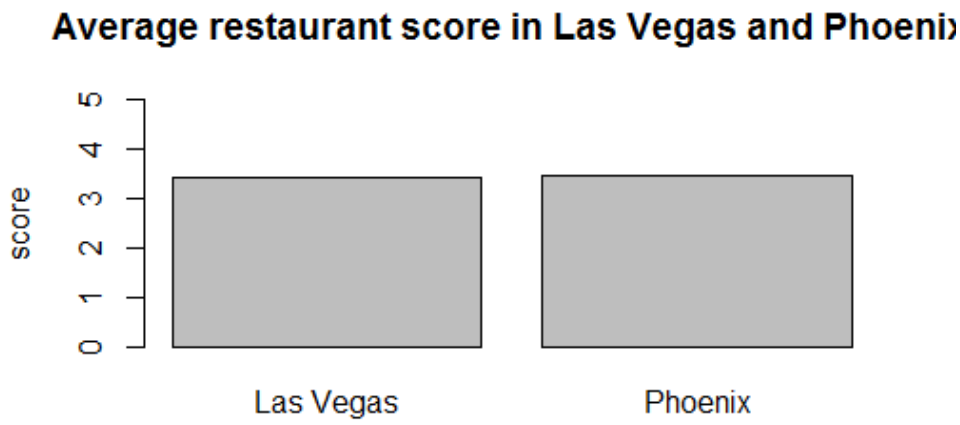
The *stars* attribute corresponds to a rating for the business. The *review count* attribute records the number of reviews/ratings that the business received. Investigate how the binary features you created for the *city* and *categories* attributes, as well as the *latitude*, and *longitude* attributes relate to these two *stars* and *review count* attributes. Identify two hypotheses about the relationships between the features (one for *stars* and one for *review count*). For each of your hypotheses:

(a) Identify the type of hypothesis (descriptive vs. relational vs. causal; direction vs. non-directional).

(b) State the hypothesis and discuss how your analysis of the data led you to the conjecture.

(c) Include a plot to support your hypothesis.

(a) Type:Directional relational
Hypothesis: The average stars of Las Vegas restaurant is similar to the average stars of Phoenix restaurant
How: Getting all Las Vegas and Phoenix restaurants' stars and take average
Chart:



(b) Type:Directional relational
Hypothesis: The average review count of Indian food is higher than Chinese food
How: Getting all Indian and Chinese restaurants' review count and take average
Chart: