CS57300: Homework 2

Long Zhen, lzhen@purdue.edu

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.

See the code file.

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

I use both scree plot and the summary to find the answer. Both are provided in the following. Component 1 to Component 6 are needed to explain more than 95% of the variance in the data.

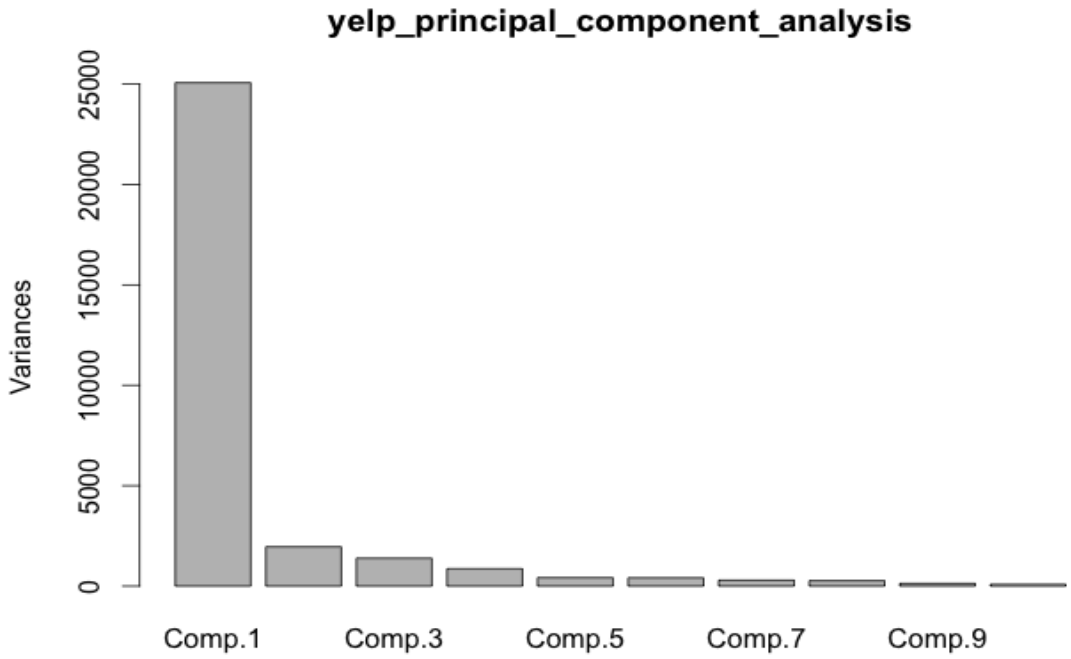

Figure 1: Q1(b) scree plot of principal component analysis

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

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	158.3208666	44.2440701	37.26238490	29.39718627
Proportion of Variance	0.8012007	0.0625713	0.04438194	0.02762336
Cumulative Proportion	0.8012007	0.8637720	0.90815398	0.93577735
	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	20.50120241	20.32685616	17.120982866	16.380360858
Proportion of Variance	0.01343457	0.01320704	0.009369629	0.008576537
Cumulative Proportion	0.94921192	0.96241895	0.971788583	0.980365120
	Comp.9	Comp.10	Comp.11	Comp.12
Standard deviation	11.428692855	9.326342495	8.671547361	6.780620300
Proportion of Variance	0.004175016	0.002780275	0.002403578	0.001469616
Cumulative Proportion	0.984540136	0.987320411	0.989723989	0.991193605

Table 1: Q1(b) Importance of components

From ***Table 2*** we can see that there are 11 attributes having a significant weight in this component. They are all listed in the table. All attributes without a significant weight are empty set by ***R***.

(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.

Similar to the previous questions, from ***Table 3*** summary and ***Figure 2*** scree plot, Comp.1 to Comp.7 are needed to achieve 95% requirement. There are 17 attributes having a significant weight in Comp.1, which are listed in ***Table 4***. It seems that the log does not bring in too many differences. However, the first component do drop a lot in proportion of variance ($0.80 \rightarrow 0.69$) and we do need one more component to achieve the 95% goal, which may not be a good thing. Although the overall proportion of variance for Comp.1 drops. There are more attributes having a significant weight on it ($11 \rightarrow 17$).

(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.

Comparing to the original data set, the random set with only 100 examples are much smaller. As I expected, we need less components to achieve the 95% goal for both log(reviewCount) and reviewCount. In both cases, only 4 components are needed. As the sample space becoming small, the number of attributes having a significant weight on Comp.1 are becoming similar (17 for reviewCount, 16 for log(reviewCount)). The gap between the proportion of variance for both cases is also smaller($0.78 \rightarrow 0.69$). Therefore, if the sample size is smaller, there is less information in it. We certainly need less components to achieve the same goal. Also, the log effect is smaller.

Attributes	Comp.1
stars	
review_count	-0.797
longitude	
latitude	
sun_mid_6	
sun_6_noon	
sun_noon_6	-0.109
sun_6_mid	
mon_mid_6	
mon_6_noon	
mon_noon_6	
mon_6_mid	
tue_mid_6	
tue_6_noon	
tue_noon_6	
tue_6_mid	-0.101
wed_mid_6	
wed_6_noon	
wed_noon_6	
wed_6_mid	-0.118
thu_mid_6	
thu_6_noon	
thu_noon_6	-0.147
thu_6_mid	-0.188
fri_mid_6	
fri_6_noon	
fri_noon_6	-0.230
fri_6_mid	-0.202
sat_mid_6	
sat_6_noon	
sat_noon_6	-0.204
sat_6_mid	-0.122
tip_count	-0.221
liked_tip_count	
likes	

Table 2: Q1(c) the weights for the first principal component

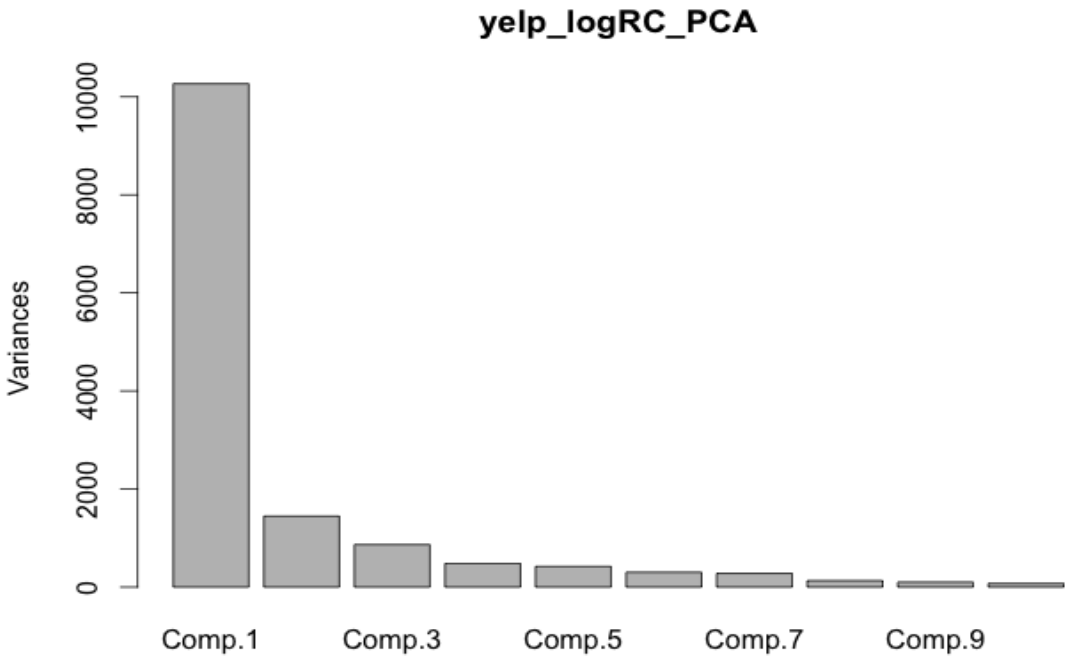

Figure 2: Q1(d) scree plot of principal component analysis with review log value

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	101.3239451	38.02957360	29.43999525	21.95453349
Proportion of Variance	0.6982976	0.09836923	0.05895109	0.03278422
Cumulative Proportion	0.6982976	0.79666684	0.85561793	0.88840214
	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	20.49728662	17.38528387	16.7095013	11.471300894
Proportion of Variance	0.02857651	0.02055796	0.0189908	0.008950385
Cumulative Proportion	0.91697865	0.93753661	0.9565274	0.965477795
	Comp.9	Comp.10	Comp.11	Comp.12
Standard deviation	10.066793963	8.676868470	7.197203665	6.769874231
Proportion of Variance	0.006892848	0.005120854	0.003523254	0.003117293
Cumulative Proportion	0.972370644	0.977491498	0.981014752	0.984132045

Table 3: Q1(d) Importance of components of review log value

Attributes	Comp.1
stars	
review_count	
longitude	
latitude	
sun_mid_6	
sun_6_noon	
sun_noon_6	-0.182
sun_6_mid	-0.160
mon_mid_6	
mon_6_noon	
mon_noon_6	-0.153
mon_6_mid	-0.168
tue_mid_6	
tue_6_noon	
tue_noon_6	-0.154
tue_6_mid	-0.173
wed_mid_6	
wed_6_noon	
wed_noon_6	-0.169
wed_6_mid	-0.202
thu_mid_6	
thu_6_noon	
thu_noon_6	-0.261
thu_6_mid	-0.324
fri_mid_6	
fri_6_noon	-0.131
fri_noon_6	-0.375
fri_6_mid	-0.329
sat_mid_6	
sat_6_noon	-0.143
sat_noon_6	-0.336
sat_6_mid	-0.196
tip_count	-0.349
liked_tip_count	
likes	

Table 4: Q1(d) the weights for the first principal component with review log value

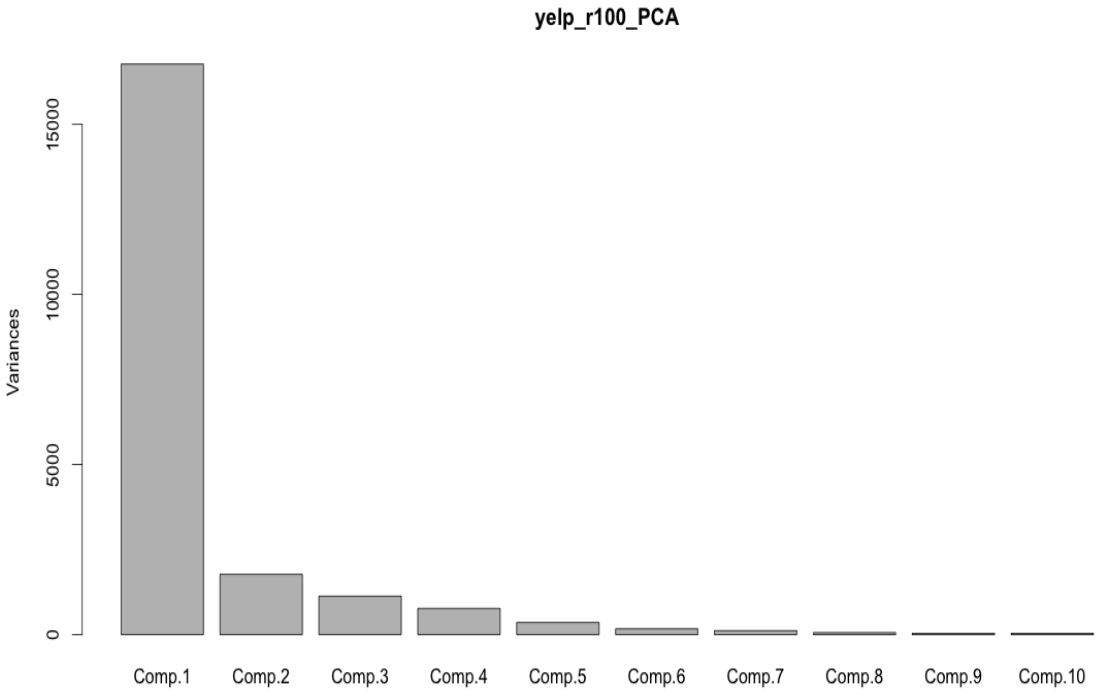

Figure 3: Q1(e) scree plot of principal component analysis with 100 random set

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	129.5207262	42.08150341	33.58919880	27.73625957
Proportion of Variance	0.7868799	0.08306392	0.05292114	0.03608492
Cumulative Proportion	0.7868799	0.86994382	0.92286496	0.95894988
	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	18.83727593	13.005522221	10.732811064	7.570561992
Proportion of Variance	0.01664432	0.007933878	0.005403273	0.002688352
Cumulative Proportion	0.97559420	0.983528076	0.988931349	0.991619701

Table 5: Q1(e) Importance of components of 100 random set

Attributes	Comp.1
stars	
review_count	-0.703
longitude	
latitude	
sun_mid_6	
sun_6_noon	
sun_noon_6	-0.118
sun_6_mid	
mon_mid_6	
mon_6_noon	
mon_noon_6	
mon_6_mid	-0.127
tue_mid_6	
tue_6_noon	
tue_noon_6	-0.125
tue_6_mid	-0.122
wed_mid_6	
wed_6_noon	
wed_noon_6	-0.121
wed_6_mid	-0.150
thu_mid_6	
thu_6_noon	
thu_noon_6	-0.161
thu_6_mid	-0.185
fri_mid_6	
fri_6_noon	-0.174
fri_noon_6	-0.229
fri_6_mid	-0.203
sat_mid_6	
sat_6_noon	-0.234
sat_noon_6	-0.201
sat_6_mid	-0.103
tip_count	-0.287
liked_tip_count	
likes	

Table 6: Q1(e) the weights for the first principal component of 100 random set

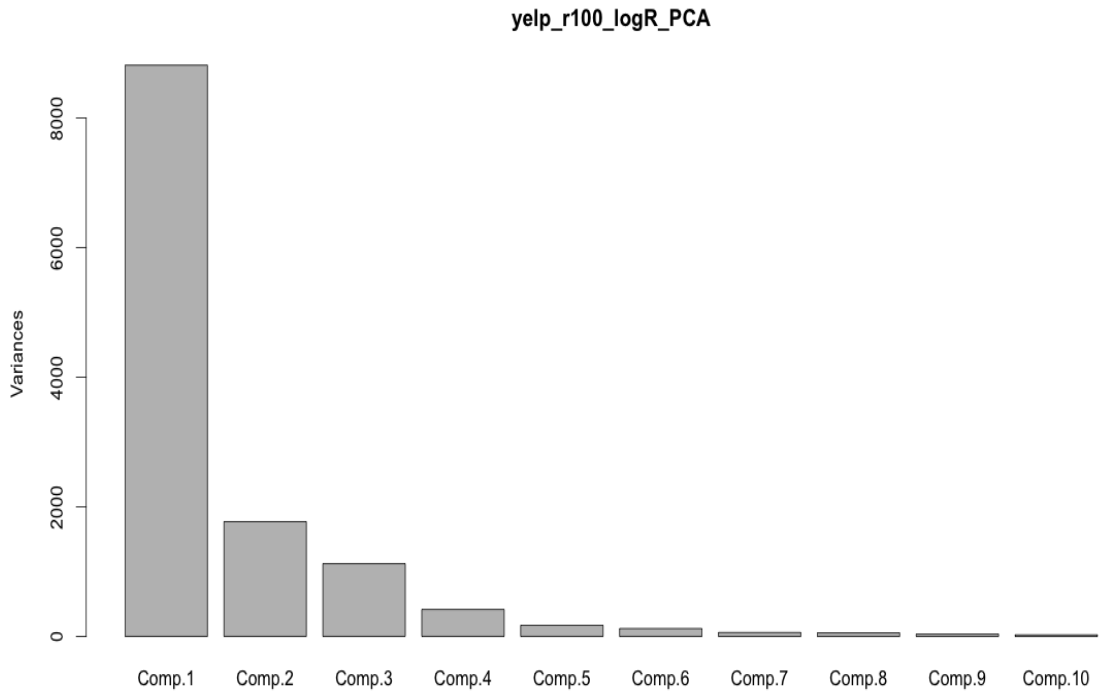

Figure 4: Q1(e) scree plot of principal component analysis with 100 random set with log review count

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	93.8946724	42.0832765	33.50803280	20.39720917
Proportion of Variance	0.6951799	0.1396479	0.08853463	0.03280627
Cumulative Proportion	0.6951799	0.8348278	0.92336247	0.95616873
	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	13.10222828	11.016541213	7.791275072	7.399853522
Proportion of Variance	0.01353648	0.009569866	0.004786658	0.004317791
Cumulative Proportion	0.96970521	0.979275076	0.984061734	0.988379525

Table 7: Q1(e) Importance of components of 100 random set with log review count

Attributes	Comp.1
stars	
review_count	
longitude	
latitude	
sun_mid_6	
sun_6_noon	
sun_noon_6	-0.166
sun_6_mid	-0.134
mon_mid_6	
mon_6_noon	
mon_noon_6	-0.136
mon_6_mid	-0.177
tue_mid_6	
tue_6_noon	
tue_noon_6	-0.174
tue_6_mid	-0.175
wed_mid_6	
wed_6_noon	
wed_noon_6	-0.168
wed_6_mid	-0.216
thu_mid_6	
thu_6_noon	
thu_noon_6	-0.222
thu_6_mid	-0.267
fri_mid_6	
fri_6_noon	-0.247
fri_noon_6	-0.316
fri_6_mid	-0.290
sat_mid_6	
sat_6_noon	-0.333
sat_noon_6	-0.278
sat_6_mid	-0.148
tip_count	-0.397
liked_tip_count	
likes	

Table 8: Q1(e) the weights for the first principal component of 100 random set with log review count

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.

See *Table 9* and *Table 10* for the eigenvector values, variances and other informations.

	Comp.1	Comp.2
Standard deviation	132.954997	18.14155395
Proportion of Variance	0.981722	0.01827798
Cumulative Proportion	0.981722	1.00000000

Table 9: Q2(a) Importance of components

	Comp.1	Comp.2
review_count	-0.968	0.251
tip_count	-0.251	-0.968
SS loadings	1.0	1.0
Proportion Var	0.5	0.5
Cumulative Var	0.5	1.0

Table 10: Q2(a) component weights information

(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.

Figure 5 is the plot for relation between v_1 and the score. **Table 11** also provides the detailed values. Therefore, we can easily tell that the best score is $17,607.580$ when $v_1 = 0.95$. Comparing to (a), $132.954997^2 = 17677.0312273$. They are very close but still have a gap. It is because this algorithm is limited by the step-size. As the step-size becoming small, we should have better score.

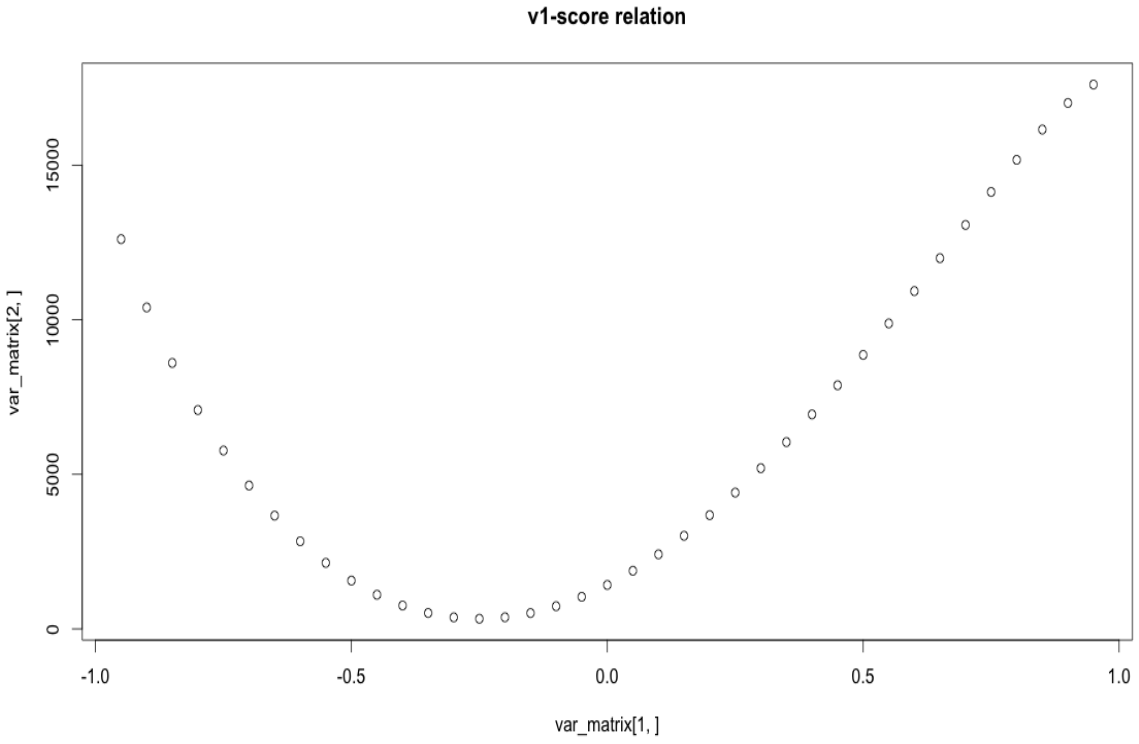

Figure 5: Q2(b) Plot the score as a function of v_1

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!)

See code file and **Table 12** for details.

v_1	-0.950	-0.900	-0.850	-0.800	-0.750	-0.700	-0.650
variance	12,606.600	10,397.570	8,603.217	7,080.493	5,769.776	4,638.195	3,664.741
v_1	-0.600	-0.550	-0.500	-0.450	-0.400	-0.350	-0.300
variance	2,834.663	2,136.899	1,562.735	1,105.036	757.774	515.724	374.248
v_1	-0.250	-0.200	-0.150	-0.100	-0.050	0	0.050
variance	329.156	376.590	512.941	734.787	1,038.833	1,421.869	1,880.723
v_1	0.100	0.150	0.200	0.250	0.300	0.350	0.400
variance	2,412.224	3,013.162	3,680.242	4,410.041	5,198.951	6,043.115	6,938.340
v_1	0.450	0.500	0.550	0.600	0.650	0.700	0.750
variance	7,879.992	8,862.842	9,880.864	10,926.920	11,992.310	13,065.950	14,133.10
v_1	0.750	0.800	0.850	0.900	0.950		
variance	14,133.100	15,172.750	16,152.040	17,011.330	17,607.580		

Table 11: Q2(a) Plot values

Category	Restaurants	Mexican	American (Traditional)	Fast Food
Count	14191	1729	1505	1486
Category	Pizza	Sandwiches	Nightlife	Bars
Count	1449	1329	1280	1207
Category	Food	American (New)	Italian	Chinese
Count	1161	1043	1005	1002
Category	Burgers	Breakfast & Brunch	Japanese	Delis
Count	864	675	507	429
Category	Sushi Bars	Steakhouses	Seafood	Sports Bars
Count	400	385	367	344
Category	Cafes	Buffets	Barbeque	Thai
Count	343	336	320	310
Category	Coffee & Tea	Mediterranean	Chicken Wings	Asian Fusion
Count	309	303	280	267
Category	Pubs	Greek		
Count	241	202		

Table 12: Q3(a) Top 30 Categories

(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.

See code file and **Table 13** for details.

City	Las Vegas	Phoenix	Edinburgh	Scottsdale	Mesa	Madison
Count	3814	2481	1027	1014	689	677
City	Tempe	Henderson	Chandler	Glendale	Gilbert	Peoria
Count	667	560	543	419	316	221
City	North Las Vegas	Surprise	Goodyear	Waterloo	Avondale	Kitchener
Count	197	144	118	117	100	96
City	Queen Creek	Middleton	Cave Creek	Casa Grande	Fountain Hills	Apache Junctio
Count	82	66	63	61	47	44
City	Buckeye	Sun Prairie	Fitchburg	Maricopa	Monona	Sun City
Count	42	39	38	37	32	31

Table 13: Q3(b) Top 30 Cities

(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.: 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.

See Table 14 for top 5 features combinations
The correlations are very interesting. It seems that the city ***Edinburgh*** is correlated with many categories. If we look at the full features combinations score table, we can see that among the top 10 combinations there are 8 combinations including ***Edinburgh***. It may be because ***Edinburgh*** is small town in indiana comparing to other cities in the data set, such as Las Vegas. So there are only small number of restaurants with limited categories. And if we extract the contingency tables of counts for further research, we can see for top 5 features combinations with ***Edinburgh***, the upper left corner number is always largest and much larger than others.

χ^2 scores	p values	City	Category
533.721930877579	4.37751738199327e-118	Edinburgh	Coffee & Tea
120.39976099922	5.17149646061583e-28	Scottsdale	American (New)
118.062323658308	1.68024595470501e-27	Edinburgh	Mexican
116.38110490353	3.92200635363141e-27	Edinburgh	Pubs
86.5258994548108	1.37919219884875e-20	Edinburgh	Cafes

Table 14: Q3(c) top five features combinations with the largest χ^2 scores

(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 sample size surely has effects. As the sample space size becoming larger, the score will become more accurate because the sampe space can more accurately reflect the characteristics of the real space. From the plot, we can see when the sample space is very small the score is not accurate at all. However, by comparing A^{max} and A^{good}, we can also learn that when the real score is small, the sample score will not change alot. In other words, the sample size may not have big effects.

A^{max}		A^{good}	
Mean	Standard Deviation	Mean	Standard Deviation
0.137	0.288	0	0
0.030	0.096	0	0
1.883	4.462	0.109	0.343
2.687	4.662	0.049	0.059
7.598	5.551	0.222	0.376
15.173	11.152	0.247	0.456
38.474	19.779	0.423	0.559
81.561	19.811	0.845	1.038
170.079	48.384	2.259	2.105
289.772	60.495	3.180	1.775

Table 15: Q3(d) Plot data for Figure 6 with mean and standard deviation

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:

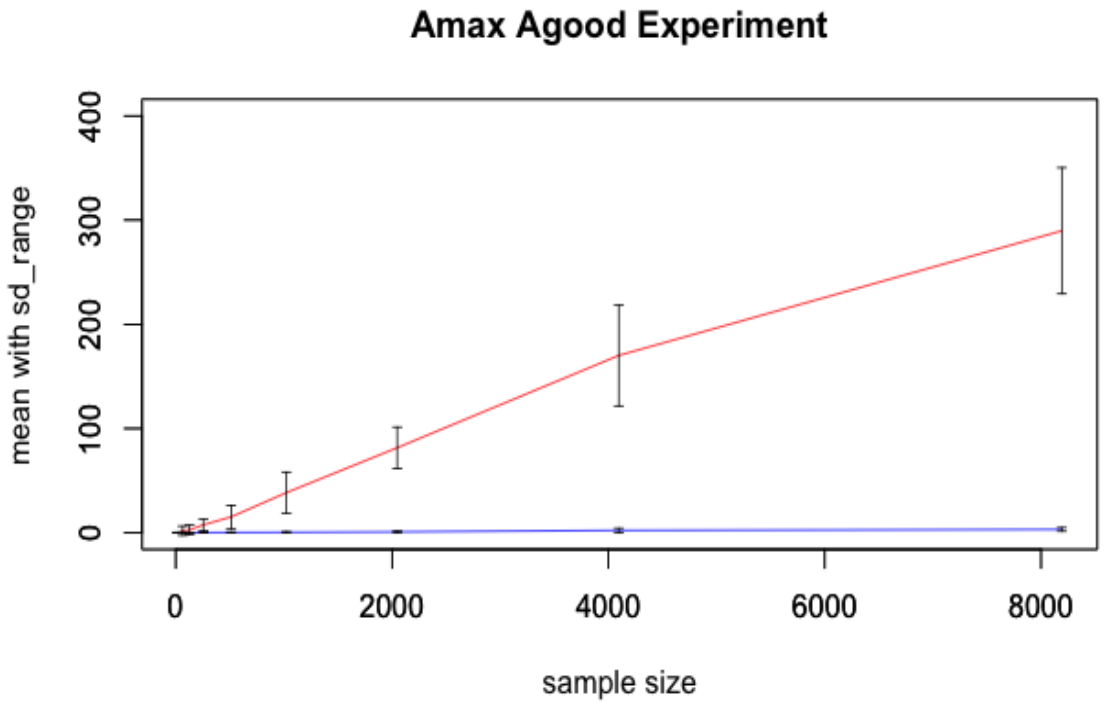

Figure 6: Q3(d) Plot of the χ^2 scores as a function of sample size, Red is A^{max}, Blue is A^{good}

(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.

Directional-relational The fast food restaurant has very poor reviews or less stars. **See Figure 7**. The reason why I made such a hypothesis is people usually go to fast food when they have no other choice or want some cheap and quick-served food. There is no quality guarantee. Therefore, I do not think fast food can receive very good reviews. I divide the yelp data set into two groups. One is all fast food restaurants. The other is non fast food restaurants. The result perfectly proved my hypothesis. The fast food median is very similar to the non fast food first quartile which is round 3. In other words, half of the fast food receive less than 3 stars. On the contrary, only one forth of non fast food receive less than 3 stars. This led to the conjecture.

Directional-relational Restaurants in Las Vegas(big popular city) have more reviews. **See Figure 8**. I come up with this hypothesis because Las Vegas seems like the biggest and most popular city in the data set by checking the state list. Also, since Las Vegas is tourism city with many famous hotels, restaurants and night clubs, people should be more willing to give more reviews. I used the similar approach. From the boxplot, you can tell the third quartile bar for Las Vegas is higher than other cities. Moreover, some restaurants have incredible large review numbers. Therefore, it proves my hypothesis.

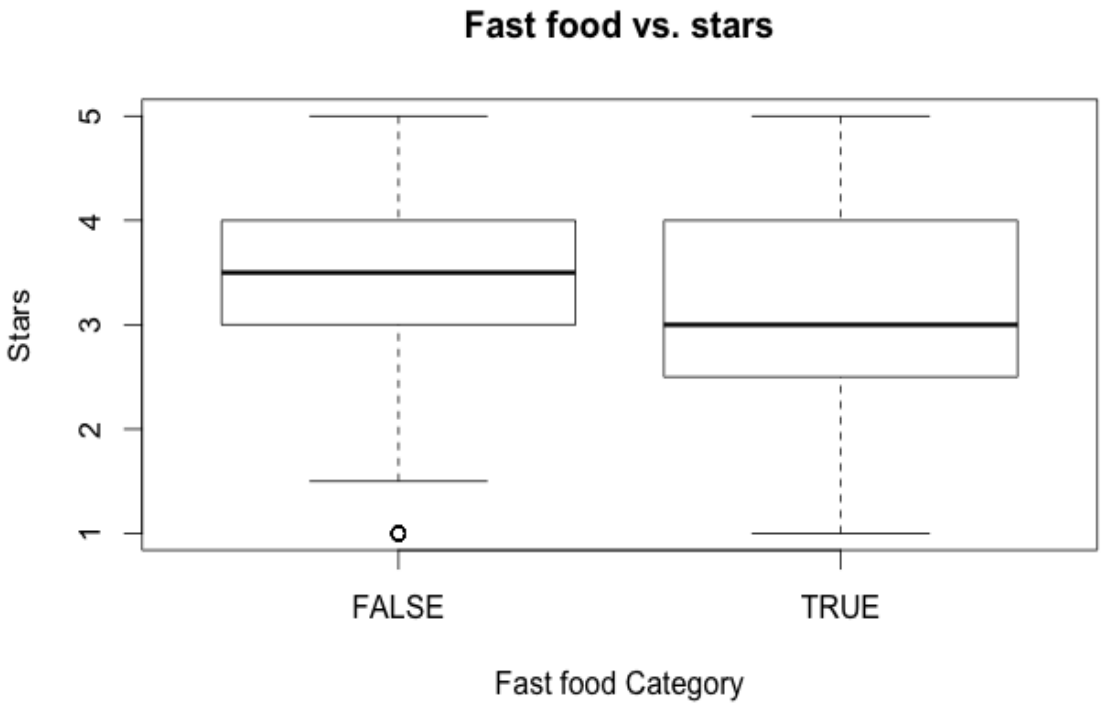

Figure 7: Q4 Fast Food vs. Stars

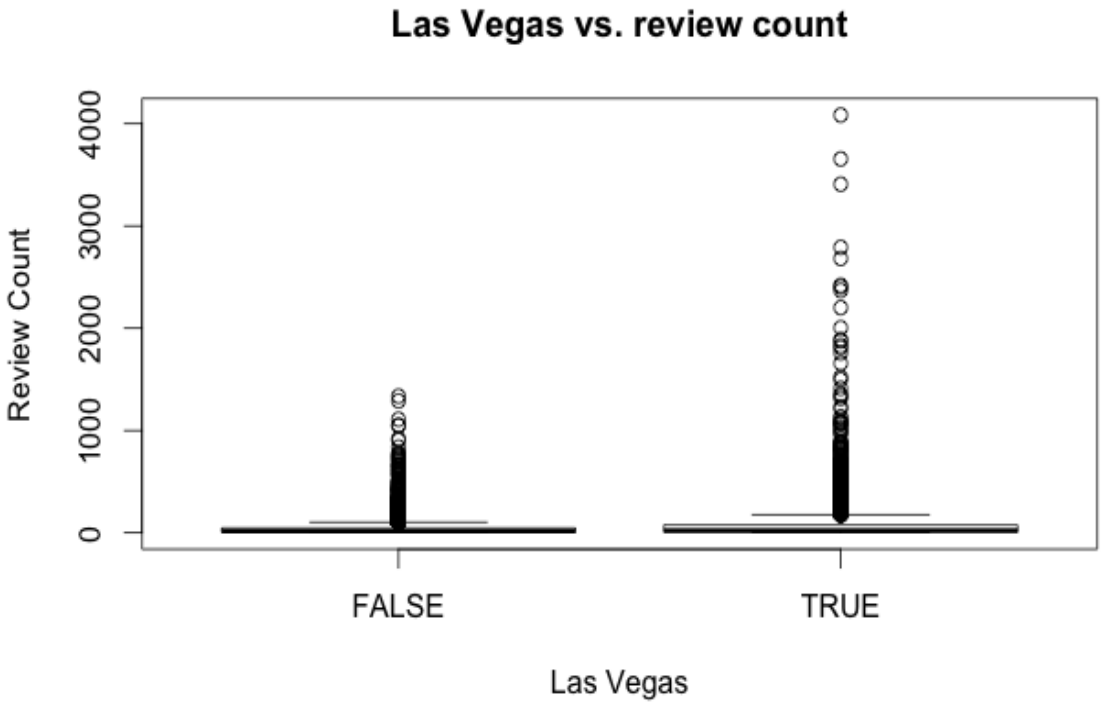

Figure 8: Q4 Las Vegas(City) vs. Review Count