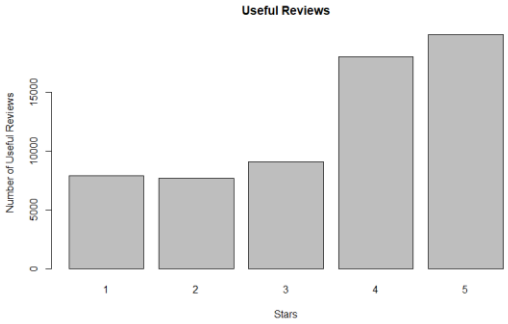
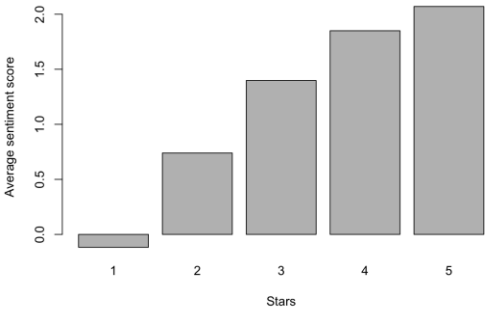
Stat 333 Group 6

Project 2

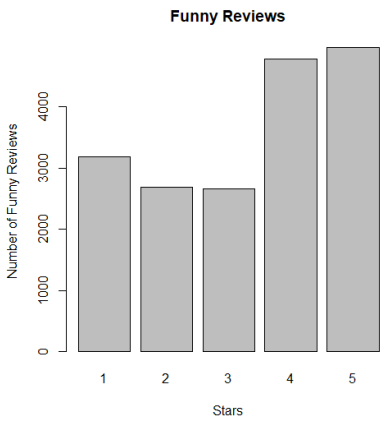
Yelp is a site where users can write online reviews for places they visit. We are given data that includes each review's text, as well as several other characteristics of each review. We are aiming to use a sample of Yelp restaurant reviews to produce a model that will allow us to predict the star ratings of other Yelp restaurant reviews.

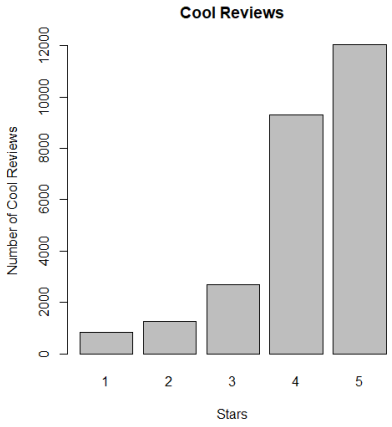




From the histogram of useful reviews, we decided to use it because there seems to be a trend in how useful a review is and its star rating. It seems like higher star reviews are more likely to be marked as useful.

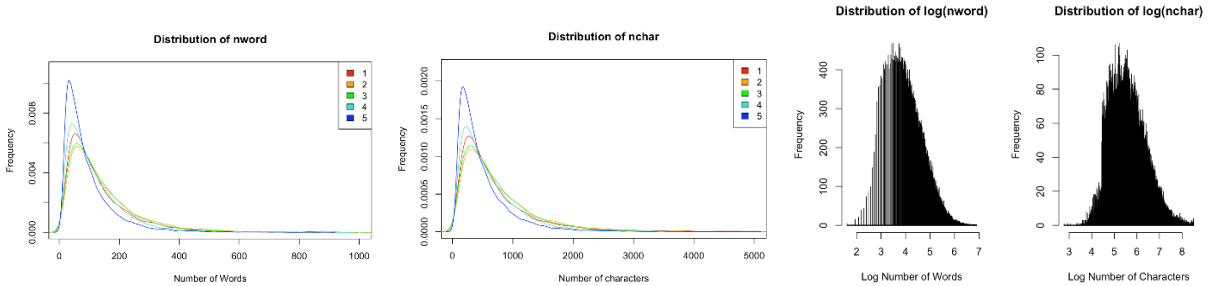
There seems to be a clear relationship between the average sentiment score for a review and its star rating. Lower star reviews seem to have a lower average sentiment score.



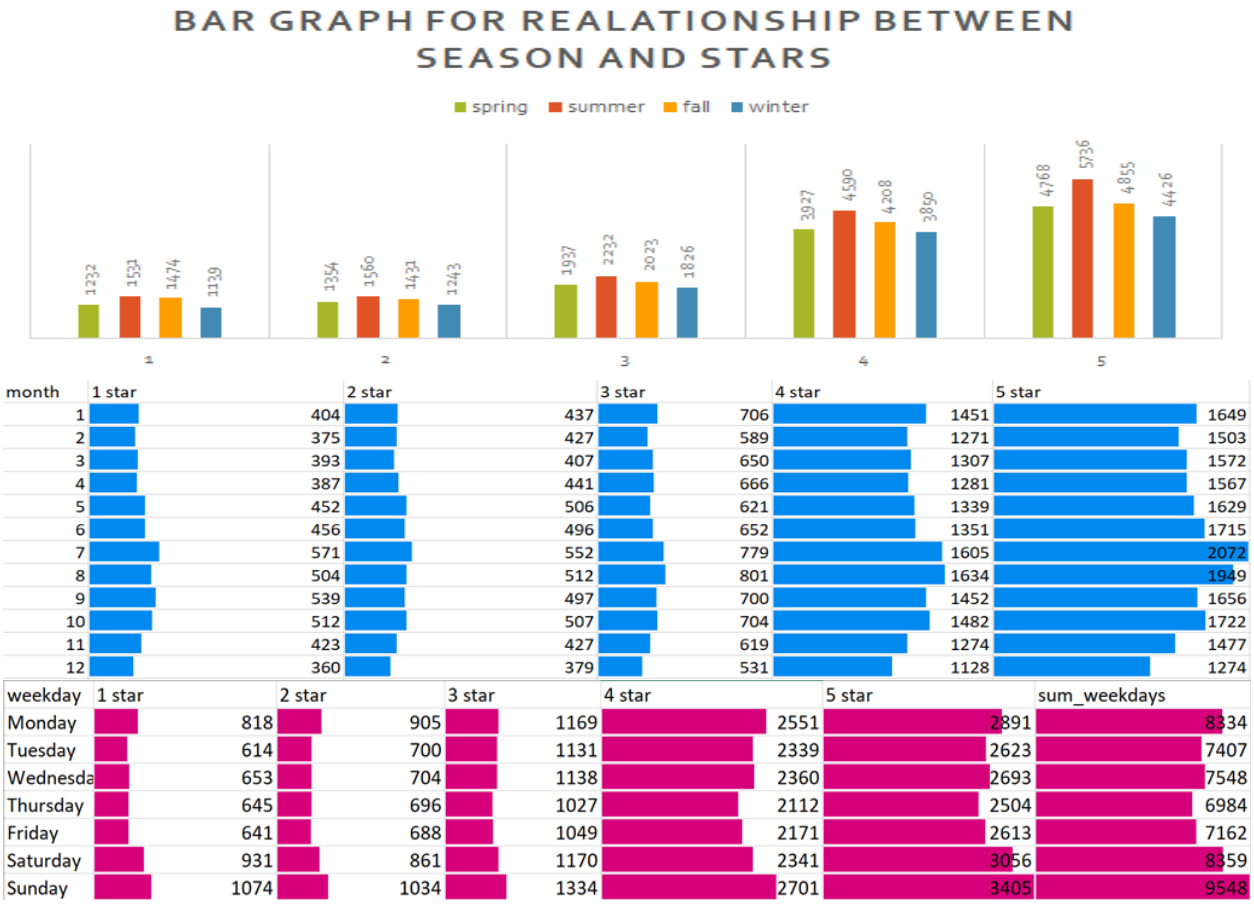


It seems like there is also a clear relationship between if a review is marked cool and its star rating. A review seems to be more likely to be marked cool if it is a higher star review.

There also seems to be a distinction between if a review is marked as funny and what its star rating might be. It seems like higher starred reviews are more likely to be marked as funny. This may be because higher starred reviews are more positive while lower starred reviews are more negative (it may be harder to make a funny review that also effectively criticizes an establishment).



From the distribution of nchar and nword, we can see there is need for a log transformation if we decide to use them.



month	1 star	2 star	3 star	4 star	5 star
1	404	437	706	1451	1649
2	375	427	589	1271	1503
3	393	407	650	1307	1572
4	387	441	666	1281	1567
5	452	506	621	1339	1629
6	456	496	652	1351	1715
7	571	552	779	1605	2072
8	504	512	801	1634	1949
9	539	497	700	1452	1656
10	512	507	704	1482	1722
11	423	427	619	1274	1477
12	360	379	531	1128	1274

weekday	1 star	2 star	3 star	4 star	5 star	sum_weekdays
Monday	818	905	1169	2551	2891	8334
Tuesday	614	700	1131	2339	2623	7407
Wednesda	653	704	1138	2360	2693	7548
Thursday	645	696	1027	2112	2504	6984
Friday	641	688	1049	2171	2613	7162
Saturday	931	861	1170	2341	3056	8359
Sunday	1074	1034	1334	2701	3405	9548

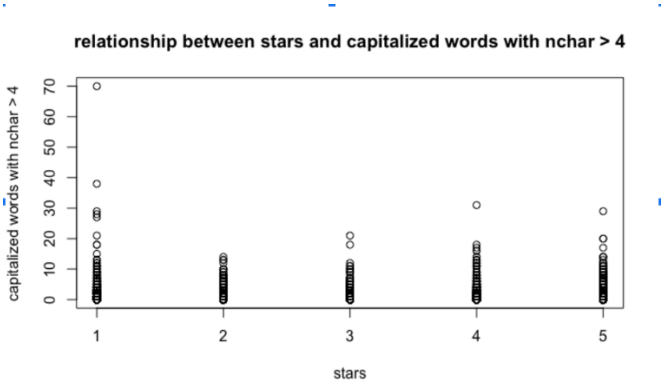
Date was split into season (quarter), day of the week, and month to capture the effects of when a review is posted on the star rating of the review.

The categories variable was also used to produce another variable, country. This variable tells us the type of cuisine a restaurant serves. This aims to capture the effect of what types of cuisine are commonly rated high by people around the Madison area.

City was also used as a predictor in our model. This is to capture the effect that different cities may have different population distributions (ethnicity/cultural background). This would result in

different tastes in cuisine from city to city as different ethnicities have differing preferences for cuisine.

We used many words/phrases as predictors for the star rating of a review. We included the words that were given to us, as well as words and phrases we found were common in reviews. We included phrases like, “not worth”, “long wait”, “definitely not”, “never go”, etc. After collecting our list of word/phrase predictors, we used cross-validation to determine which words/phrases were useful for prediction.



We figured that reviews with a high number of ALLCAPS words would be more likely to be a lower star review. Unfortunately, according to the graph, this does not seem to be the case.

Due to the complexity of this review data, we have decided not to identify/exclude outlier reviews from our data. As each review is written by a person, we assume they have a good reason for sharing their experience.

We attempted to use Lasso for our regression, but inspection of the lambda – squared error plot showed that as lambda increased, our squared error also increased. We also saw that for lower lambda values (below -5 or so), our squared error was not changed. This told us that not all our predictors were as useful as we would want. This led us to using cross validation to cut back the predictors which were not very useful.

We use a multiple lm(y~…) to produce a linear regression model. Using the multiple linear regression model allows us to have p-values for each of the predictors we add. To make this process highly efficient, we introduce the 10-fold cross-validation of our yelp data. Cross-Validation is a resampling procedure used to evaluate our model on a limited data sample. We shuffle the dataset randomly and split our data to 10 folds and select 9 folds as training folds and the left one as validation fold. Each time, we fit our linear regression model on the training folds and evaluate it on the test fold. Using this method, we are better able to avoid overfitting our model. For any given run of cross-validation, we are not using the full set of training data to train our model, making it impossible to overfit the model with respect to the fold of data that is saved for testing. With this method, we can test our model using limited sample data and predict the performance on other hidden data. K-fold cross-validation generally results in a less overfit model.

Our final model includes the following predictors: useful, funny, cool, sentiment, city, nchar, nword, country (“Italian”, “Chinese”, “American”, “Thai”, “Other”, etc.), season/quarter, month, day of the week, many of the word predictors provided in “Yelp_train..csv”, and several word/phrase predictors that we came up with by looking for commonly used words/phrases in the reviews.

We all were able to contribute to this project by helping each other with writing R code, making the presentation, and the summary write up. We each focused more on a particular part of the project (Ching-Wen worked more on the R code, Yuxin worked more on the presentation, and Roshan worked more on the summary write up). To be clear, each of us worked on all parts of this project, while also focusing on one of the three big deliverables.

Peer Review (Roshan):

Ching-Wen – 5

Yuxin – 5

Peer Review (Ching-Wen):

Roshan – 5

Yuxin – 5

Peer Review (Yuxin):

Roshan – 5

Ching-Wen – 5