Capstone1: Milestone Report

The purpose of this document is to review the objectives of the project and to consolidate the findings thus far.

Sections

1. Define the problem
2. Identify the client
3. Describe the data set, and how it was cleaned/wrangled
4. List other potential data sets that could be used
5. Explain the initial findings
6. Share Capstone Project 1 code and milestone report

1. Define the problem

With the event of social media and online directories - such as Yelp, TripAdvisor and Zomato - online marketing is becoming an increasing important aspect of promoting a business.

In particular, reviews and ratings of a restaurant could make or break a new establishment. It is now a common practice for customers to checkout a restaurant (or establishment) using online reviews and ratings before giving it a try.

This project will predict Yelp's 5-star rating based on the review written.

The prediction could be used as part of a new feature to change the default rating from a static value (e.g. default of 3 stars or 0 stars) into a dynamic default value based on the review. This default rating would still need to be accepted by the reviewer. Alternatively, this could be used to trigger is a verification for the user to confirm the rating if the predicted rating is significantly different from the actual rating. Regardless of implementation method, the intent of the prediction is to help make review ratings more accurately reflect the reviewer's opinion.

2. Identify the client

The client for this project would be Yelp.

An easier review experience with more reflective ratings, would add value to all to their major stakeholders:
- a dynamic rating could help reduce the cognitive load on the reviewers, making reviews easier
- better rating accuracy helps new customers with their online research
- more reflective ratings reduce frustration / confusion of establishment proprietors arising from an unconsidered rating
- the platform's (e.g. Yelp) quality of reviews is improved - providing better, cleaner data

3. Describe the data set, and how it was cleaned/wrangled

For more details, please refer to the document titled:
`Capstone1_writeup_datawrangling ` and its accompanying jupyter notebook.

The dataset was obtained as part the '[Yelp dataSet Challenge - Round 10](https://www.yelp.com.sg/dataset)' (September 1, 2017 to December 31, 2017). In particular, the `review.json` and `business.json` files from the JSON dataset.

These are currently stored in the folder labeled ` 01_raw_data`. In accordance with the terms and conditions of the dataset usage, the data shall not be shared in the Github repo.

Exploring the dataset

After the dataset was loaded in, 3 functions were used: `.shape`, `.info`, `.describe`. A null check was performed, and a head of 50 was used to eyeball the data set.

The following observations were made about the `review` data:
1. There were no null values
2. Except for the data, the data types from the column were correct
3. The full data set had ~ 4.7 million observations
4. Non-restaurant businesses were also reviewed (e.g. accommodation)
5. Reviews were not all in English

For the `business` data:
1. There were only 2 null values (1 Lat, 1 Long)
2. The data types were not an issue but there were a number of columns related to the location of the business
3. The full data set had ~ 150 k observations
4. Non-restaurants were included
5. A number of locations were outside the USA (e.g. state: ON, postal_code: M4K 1N7 refers to Toronto). Unique state values and the range of lat long values were further analyzed and revealed that a number of businesses were outside the USA.

Cleaning the dataset

To clean the review data, very little clean up was necessary. The date was converted to datetime but the stars were kept as integers vs. categories.

To clean the business data:
- non-restaurant businesses were removed. Leaving ~ 50k observations (almost 1/3 the original number of businesses although the impact on the number of reviews is less severe).
- next, restaurants outside the USA were removed using a bounding box of min-max latitude and longitudinal values. The bounding box for the US is (49.3457868 # north lat) (24.7433195 # south lat) (-124.7844079 # west long) (-66.9513812 # east long). This reduced the dataset to ~48k observations.

The restaurant data frame was merged with the review dataframe using an inner join on the 'business_id'. For the merged data, the following observations were made:
1. There were no null values
2. Inner join reduced the total number of observations (i.e. reviews) from ~4.74 million from all businesses to ~ 2.92 million from restaurants alone to the ~2.88 from restaurants in the USA bounding box.

A bar chart to showed a bias towards positive 5 star ratings but uncovered no outliers. Data exploration will be discussed in the next section.

Pre-processing text

The pre-processing of the review text involved 5 steps. These were defined in function named `pre_process_review`:
1. Removed punctuation
2. Convert characters to lower case
3. Removal of stop words
4. Lemmatization of words
5. Conversion for review text string to a list of words (tokenization)

Note that the additional step of expanding contractions was not used, but it would come as step 1 if it were needed.

4. List other potential data sets that could be used

Other potential data set that could be used to explore the current data set is: `user.json` file from the '[Yelp dataSet Challenge - Round 10](https://www.yelp.com.sg/dataset)' dataset. This could be joined onto the column in the review dataframe titled: user_id.
This data would be especially interesting if one were to do an analysis regarding the usefulness of a review. However, it is not expected to impact the current analysis on prediction of the star rating for a given review, since majority of reviews come from reviewers with few reviews (as seen in the next section).

5. Explain the initial findings

For more details, please refer to the documents titled: `Capstone1_datastorytelling` , `Capstone1_writeup_inferentialStatistics` and their accompanying Jupyter notebooks.

An initial analysis of the data helped support the initial hypothesis of the project outlined at the start of this document:
- Yelp reviews are relevant in today's context
- Review ratings could have significant impact on a restaurant
- Review text can be used as a predictor of star ratings

Hypothesis: Yelp reviews are relevant in today's context

A plot of the number of Yelp reviews over time (figure 1) suggests that Yelp continues to be relevant in 2017. It can be seen that the number of reviews grows rapidly till 2015 when they begin to a phase of slower but positive growth.

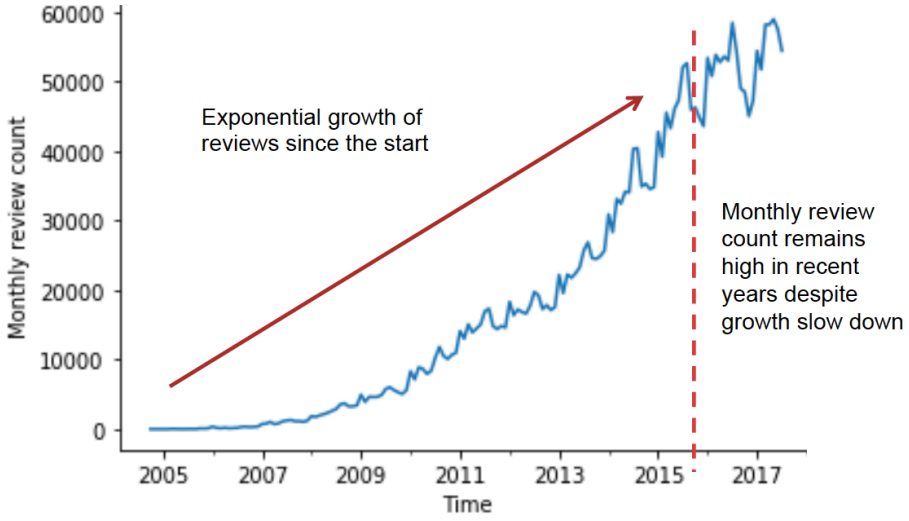

Figure 1: Monthly restaurant review counts

Hypothesis: Review ratings could have significant impact on a restaurant

It was observed that the majority of restaurants have less than 20 reviews (see figure 2). For these restaurants, a single review rating has a significant impact on the average star rating for the restaurant. This would impact the restaurant's standing within the Yelp community, which might significantly impact those restaurants.

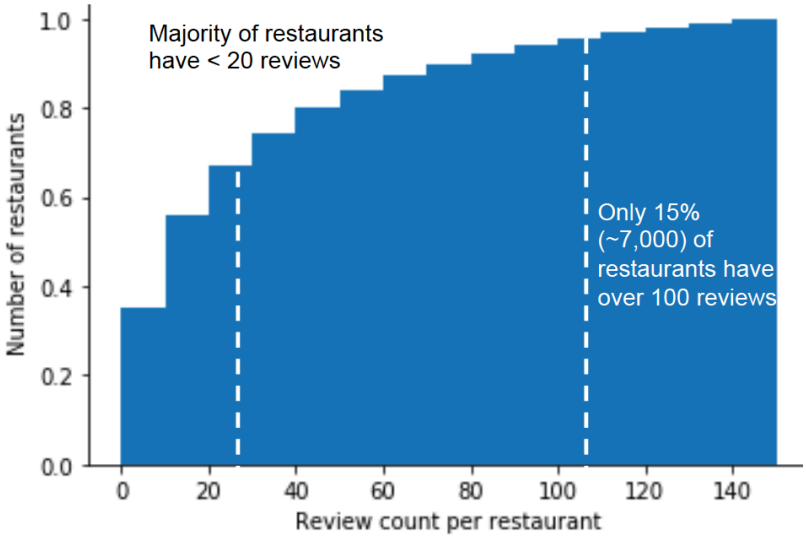

Figure 2: Cumulative distribution of review counts

Hypothesis: Review text can be used as a predictor of star ratings

The hypothesis that review text is a good predictor of star ratings is explored by reviewing other available data (e.g. user data, restaurant data, location, text length etc).

An analysis of users writing reviews showed that the top two reviewers wrote thousands of reviews and the top users all remain prolific writers (see figure 3). However, further analysis of the distribution of restaurant reviews per user (figure 4), revealed that the majority of reviews were written by first time reviewers. This implies that analyzing a particular user's behavior might be helpful for prolific writers, but there is limited impact of such an analysis on review ratings as a whole.

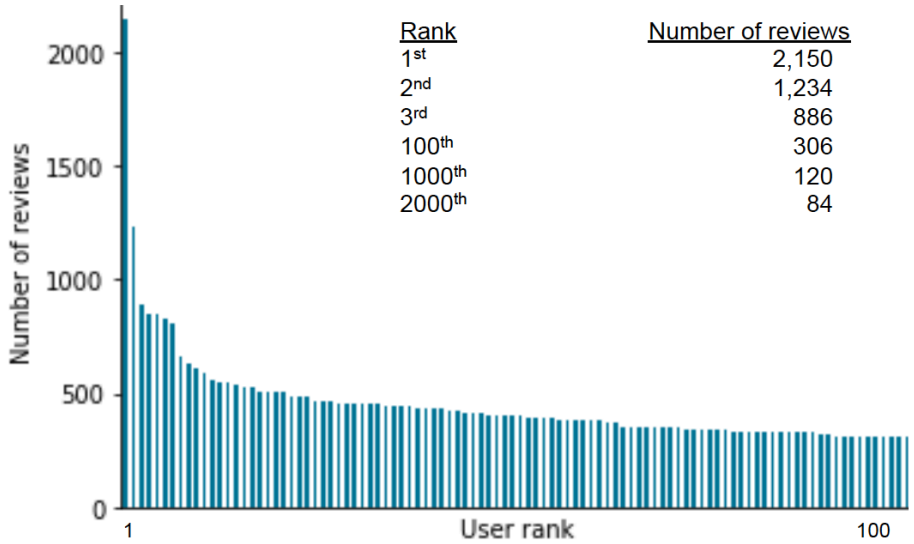

Figure 3: Number of reviews written by the top reviewers (each)

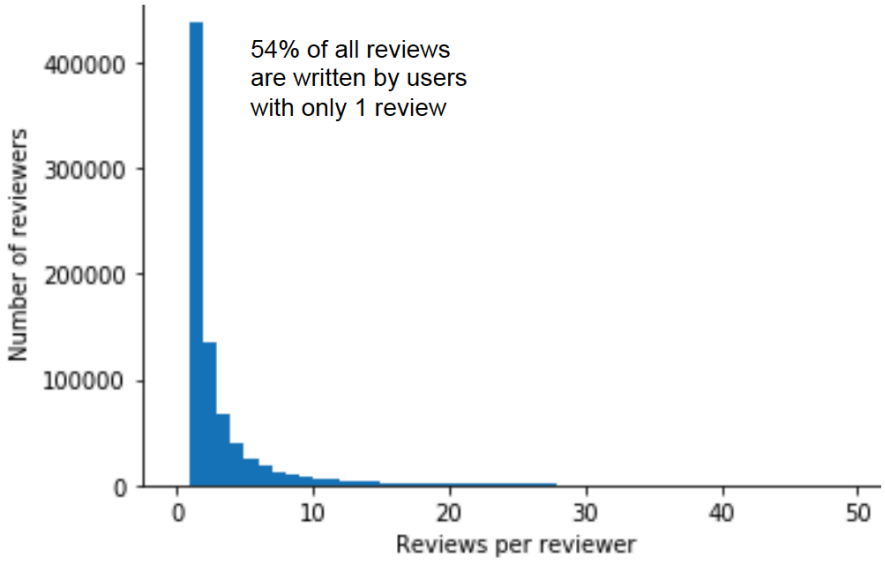

Figure 4: Distribution of restaurant reviews per user

Next, the restaurants were analyzed to understand what ratings they got and the number of reviews they received. It was found that on average, restaurants were rated fair to good (3-4 stars) (see figure 5). However, restaurants with better ratings tend to get more reviews (see figure 6).

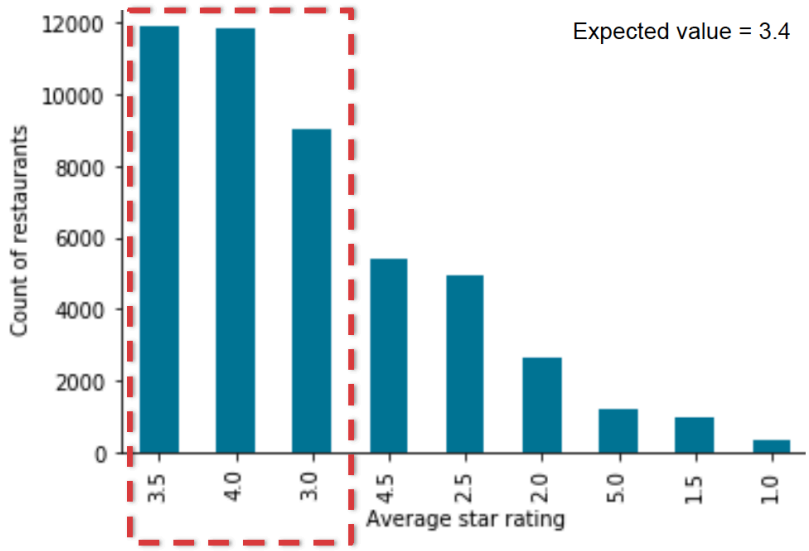

Figure 5: Distribution of average star ratings for restaurants

Figure 6: Distribution of reviews by restaurant's average star rating

If we were to use the expected value of star ratings based on the average star ratings received by restaurants, this might result in a popular restaurants (who get higher ratings) receiving a lower rating. Assigning a value based on the most common review rating or on the expected value for a review rating might result in higher than justified rating. The analysis of the distribution of reviews by star ratings (figure 7) provides the mean and mode values for review ratings.

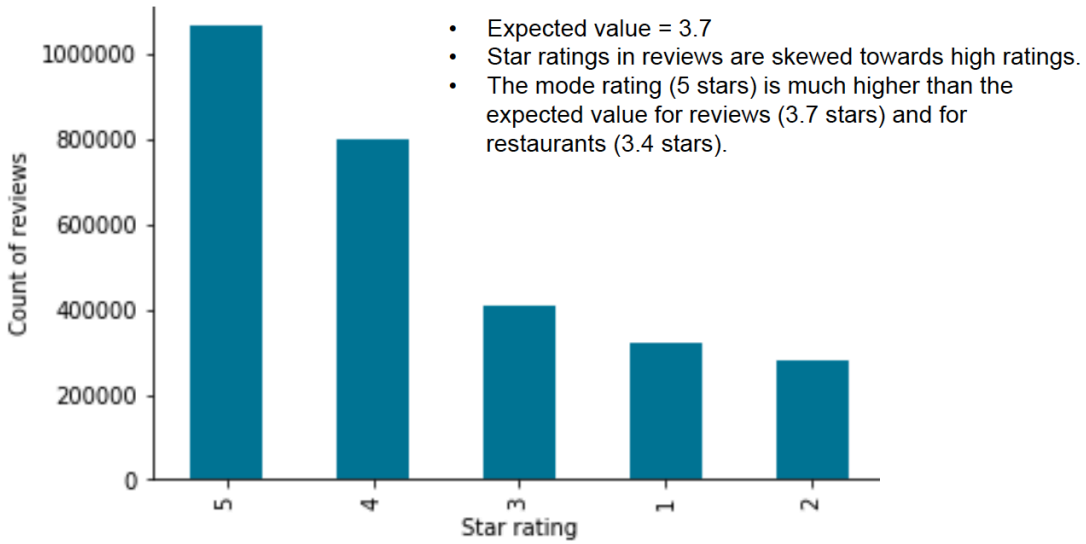

Figure 7: Distribution reviews by star ratings

An analysis of the review text was conducted to explore the link between review text and ratings. First, a simple analysis of the distribution of review lengths by stars (figure 8) indicated that the length of a review appears to be a poor proxy indicator of star ratings. Therefore, the content (and not just the length) of the text would need to be analyzed.

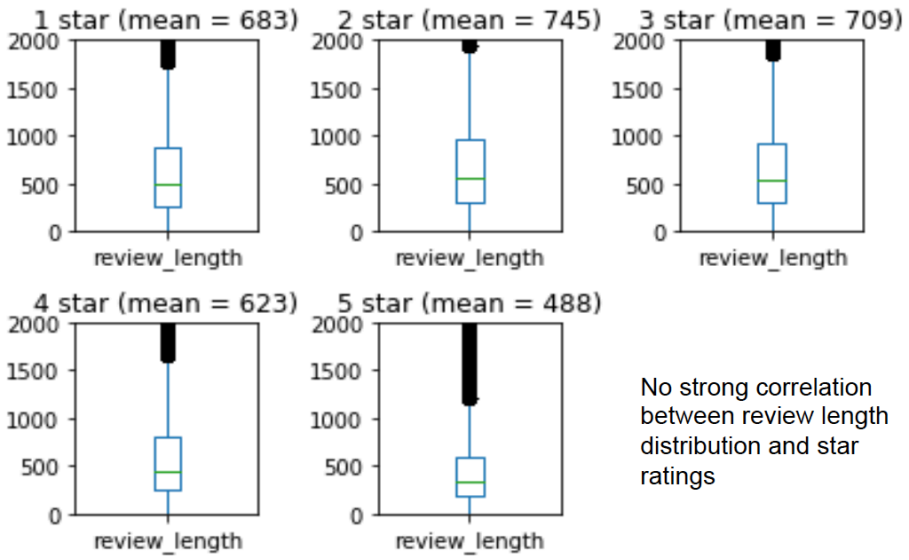

Figure 8: Distribution of number of characters in a review by stars

In addition, inferential statistics was used to validate the decision to drop all other columns other than the review text and review stars. A correlation matrix (figure 9) was generated between the numerical data fields.

	cool	funny	stars_review	useful	is_open
cool	1.000000	0.846138	0.042902	0.854624	-0.015011
funny	0.846138	1.000000	-0.041543	0.823759	-0.020031
stars_review	0.042902	-0.041543	1.000000	-0.040162	0.051385
useful	0.854624	0.823759	-0.040162	1.000000	-0.028944
is_open	-0.015011	-0.020031	0.051385	-0.028944	1.000000
latitude	-0.029170	-0.037964	-0.031977	-0.026753	0.011822
longitude	-0.047311	-0.051008	-0.033299	-0.035423	0.015738
review_count	0.022783	0.022769	0.066223	0.007781	0.084150
stars_business	0.043261	-0.000225	0.413689	0.012131	0.115686

	latitude	longitude	review_count	stars_business
cool	-0.029170	-0.047311	0.022783	0.043261
funny	-0.037964	-0.051008	0.022769	-0.000225
stars_review	-0.031977	-0.033299	0.066223	0.413689
useful	-0.026753	-0.035423	0.007781	0.012131
is_open	0.011822	0.015738	0.084150	0.115686
latitude	1.000000	0.810437	-0.159074	-0.073632
longitude	0.810437	1.000000	-0.291484	-0.078719
review_count	-0.159074	-0.291484	1.000000	0.157608
stars_business	-0.073632	-0.078719	0.157608	1.000000

Figure 9: Correlation matrix of numerical columns in the dataset

The target of our current analysis, 'stars_review', did not have notable correlations to any numerical data field, other than the 'stars_business' (the average of 'stars_review' for a restaurant). Since this field is an aggregated function of the review rating (stars_review), it shan't be used to predict the star rating for the current review as such a practice would promote the enforcement of existing opinions and negatively impacts independent judgment.

Similarly, to encourage independent judement of reviews to rate a business, we will remove 'name', 'business_id' as assignment of the rating based on the business ID or name would reinforce the bias. Since 'review_id' is an idex, we'll remove that too.

Therefore, we can conclude that for our analysis of 'stars_review', we can drop the columns: 'cool', 'funny', 'useful', 'is_open', 'latitude', 'longitude', 'review_count', 'stars_business', 'name', 'business_id', 'review_id'

To validate if location significantly affected star ratings, the distribution of review ratings were plotted by state (figure 10).

The distribution of stars awarded by reviews are consistent with the exception of WA and SC. We see that WA has a very low number of reviews do not expect it to impact the analysis significantly.

SC has a greater proportion of low star ratings but not a very high number of reviews. We suspect this will not significantly impact the analysis either.
Since there is little correlation between 'stars_review', 'latitude' and 'longitude' as well as 'state', we will infer that location associated columns will not help us to determine the 'stars_review'.

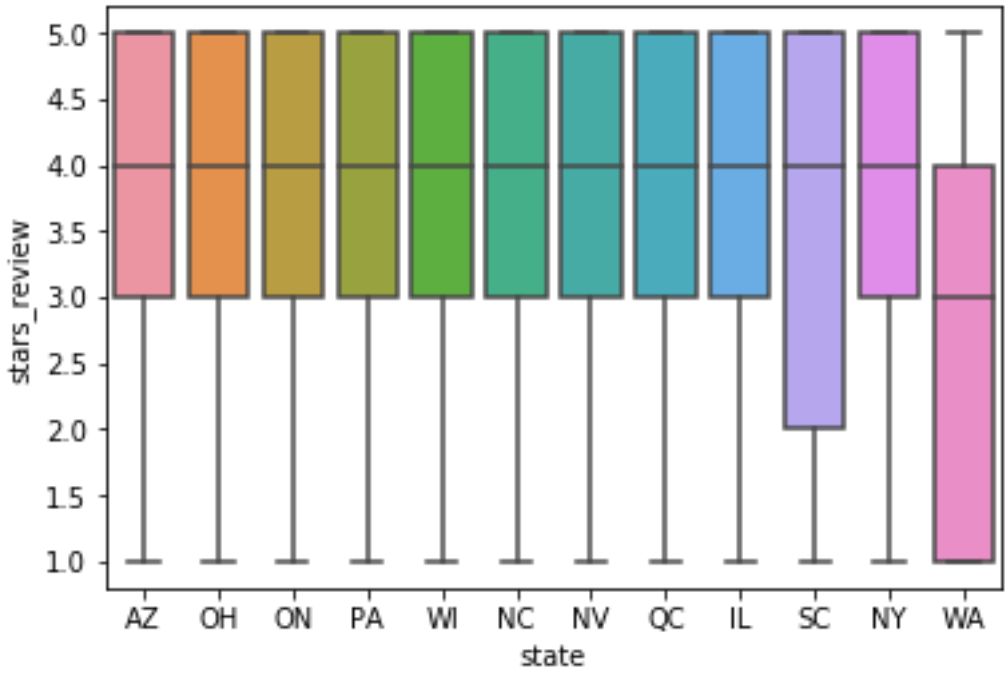

Figure 10: Distribution of review ratings by state

The remaining columns: 'user_id', 'attributes', 'categories', 'hours'

- 'user_id' : in the previous analysis we noted that majority of reviews have no previews reviews and should not be used as the basis of a predictor
- 'attributes', 'categories', 'hours': are descriptors of the business and not the review

After ruling out other columns in the data set as suitable predictors for the review star rating, sn initial look at the unigrams, bigrams and trigrams through word clouds (figure 11) suggests that the review text content might be an appropriate way of predicting the star ratings.

Figure 11: N-gram word clouds

Basic analysis of top words used - exploring the top 10 uni-grams based on star ratings (figure 12) and the rank of the word in terms of frequency of use (figure 13) - suggest differences based on star ratings.

TOP WORDS IN REVIEWS GROUPED BY STAR RATINGS

Rank of word in terms of frequency of use (in text by star rating):

1	time	good	good	good	great
.	like	like	like	great	good
.	one	time	time	like	time
.	back	one	would	time	best
5	get	get	one	one	one
.	would	would	get	really	like
.	go	back	really	get	delicious
.	good	really	great	go	go
	even	go	go	would	back
10	never	table	back	back	love

Figure 12: Top 10 unigrams by star rating

TOP WORDS IN REVIEWS GROUPED BY STAR RATINGS

Rank of word in terms of frequency of use (in text by star rating):

best	153	76	56	31	4
great	60	16	8	2	1
good	8	1	1	1	2
ok	118	50	35	251	540
bad	22	30	48	148	221
never	10	48	110	97	55

Top 10

Figure 13: Rank of words in terms of frequency of use (in text by star rating)

Since the exploratory data analysis section suggested that particular words such as 'bad' and 'best' might be good indicators of the star rating of the review. The hypothesis that frequency of occurrence for the two words are statistically significant were tested using permutation tests with a null hypothesis of identical distributions.

The Central Limit Theorem (CLT) is applicable to the difference in the rate of appearance of frequently used words such as 'bad' or 'best' since reviews are assumed to be independent events. In addition, the difference in frequency of occurrences is expected to be normally distributed.

Using the test statistic - Frequency (rate) of word (bad / best) appearing – the following null hypotheses were tested:
1. The frequency of appearance of the word 'bad' is identical for 1 star rated reviews and the other reviews
2. The frequency of appearance of the word 'bad' is identical for 5 star rated reviews and the other reviews
3. The frequency of appearance of the word 'best' is identical for 1 star rated reviews and the other reviews
4. The frequency of appearance of the word 'best' is identical for 5 star rated reviews and the other reviews

In all 4 cases, the p-value for the hypotheses tests suggested that the Null Hypotheses can be rejected. The alternative hypothesis being that the words 'bad' and 'best' are statistically significant to the star rating that the review receives.

6. Share Capstone Project 1 code and milestone report

It should be noted that due to the previous findings and what was learnt, a new file was created for feature engineering. This working was done in a Jupyter notebook called `milestone_1.ipynb`.

This file repeats much of the work found in the previous notebooks and leverages the findings made so far. In particular:
- Inferential statistics analysis supports the notion that all columns other than the preprocessed text may be removed.
- Exploratory data analysis suggested the review text might be analyzed as a uni-gram, a bi-gram or a tri-gram. The current pre-processing in the data wrangling stage removes the stop words.

In addition to dropping all columns other than the pre-processed text and the star ratings, the following pre-processing steps will be used:
1. Removed punctuation
2. Convert characters to lower case
3. Lemmatization of words

The steps:
1. Removal of stop words
2. Conversion for review text string to a list of words (tokenization)

Will be handled by the CountVectorizer and the removal of n-grams solely comprising of stop words.

In view of this, a new CSV export has been prepared.