Capstone1: Write up for Inferential Statistics

The purpose of this document is to explain the steps taken to validate the statistical significance of the parameters that were dropped as well as the variables that are particularly significant in terms of explaining the answer to the project question.

Conclusions

Consistent with the objective of predicting the star rating a review will receive, this section of inferential statistics supports the decision to focus on using the review text to build our predictive model. It also supports the decision to drop the remaining columns (for the purpose of predicting star ratings for the review).

However, it should also be noted that the analysis indicates strong correlations between the columns 'cool', 'funny' and 'useful'. This might prove useful in a future analysis of understanding what makes a good review.

Sections

1. Set up jupyter notebook - `inferential_statistics.ipynb`
2. Load in data for reviews and businesses, merge and pre-process data
3. Validate decision to drop columns other than review text and review stars
4. Verify that reviews text contain words that may be statistically significant

1. Set up jupyter notebook

The inferential statistics stage can be found a Jupyter notebook called `inferential_statistics.ipynb `. It will contain the key steps laid out in this document.

As part of the set up, the library dependencies are loaded.

2. Load in data for reviews and businesses, merge and pre-process data

The dataset used is consistent with prior sections - `review.json` and `business.json` files from Yelp.

As with the data wrangling section, the review and business data sets were merged. Only reviews of restaurant businesses in the USA have been retained. For this analysis, no columns were dropped.

3. Validate decision to drop columns other than review text and review stars

Firstly, a correlation matrix was generated between the numerical data fields.

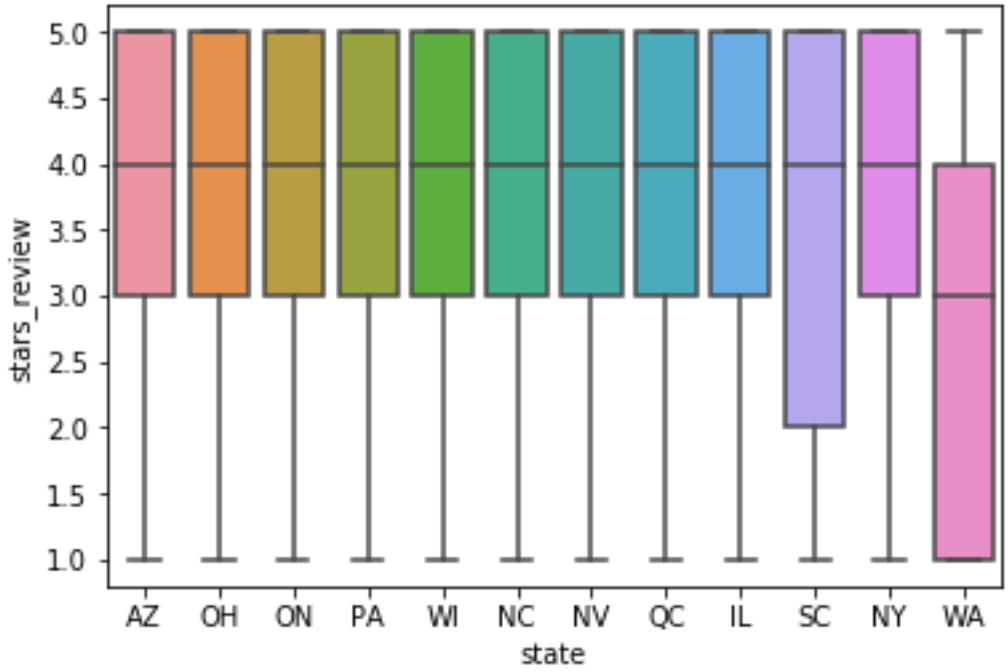
	cool	funny	stars_review	useful	is_open
cool	1.000000	0.846138	0.042902	0.854624	-0.015011
funny	0.846138	1.000000	-0.041543	0.823759	-0.020031
stars_review	0.042902	-0.041543	1.000000	-0.040162	0.051385
useful	0.854624	0.823759	-0.040162	1.000000	-0.028944
is_open	-0.015011	-0.020031	0.051385	-0.028944	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

It can be noted that there are high correlations between 'cool', 'funny' and 'useful' reviews. This might be helpful for a future analysis of what makes a good review. However, the target of our current analysis, 'stars_review', does not have notable correlations to anything other than the 'stars_business'. This makes sense since 'stars_business' is the average of 'stars_review' for a restaurant and there are a large number of restaurants with a low number of reviews (see data story telling), however this isn't very helpful.

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'

To validate if location significantly affected star ratings, the distribution of review ratings were plotted by state:



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'.

The point of a review is to rate a business, to assign the rating based on the business ID or name would reinforce the bias, therefore we will remove 'name', 'business_id'. Since 'review_id' is an idex, we'll remove that too.

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

- 'user_id' : in the previous analysis we noted that majority of reviews have no previews reviews
- 'attributes', 'categories', 'hours': are descriptors of the business and not the review

4. Verify that review text contain words that may be statistically significant

The pre-processed Unigram data from the data wrangling section loaded into a data frame and vectorized into a dense matrix (and coverted into a dataframe).

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.

Reviews were assumed to be independent events, the Central Limit Theorem (CLT) was therefore applicable to the difference in the rate of appearence of a frequently used word such as 'bad' or 'best'. In addition, the difference in frequency of occurences 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.