Capstone Project 2: Milestone Report_1

Questions:

Define the problem:

- Sentiment analysis of the Yelp customer review
- Prediction of the star rating based on the review
- Recurring Themes
- Try to Tag/Entities with sentiment

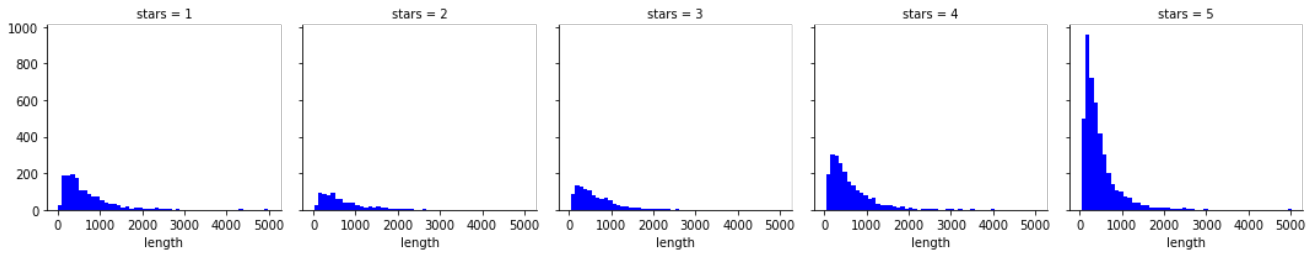
Identify your client:
Client could be all the e-commerce companies starting with Yelp

Describe your data set, and how you cleaned/wrangled it:

- The dataset was directly downloaded from https://www.yelp.com/dataset
- Several columns were renamed to make them readable
- column “review_id” was made the index
- Missing data was searched but their were no missing data
- A separate new column ‘length’ was made which counts and stores the length of the reviews to be used for further analysis

Explain your initial findings:

- Graph was plotted with number of review vs length for all the 5 stars rating separately.



As per the above graph we can see that the length of the reviews were more for the 1,4 and 5 stars. We can also deduce that for stronger feeling the length was more.

- We also observed that there is negative correlation between:

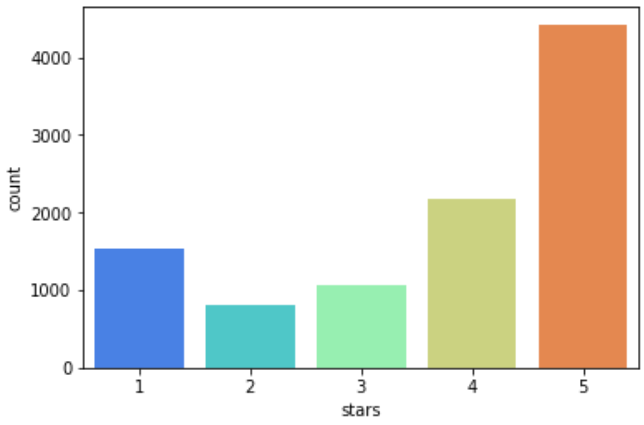
Cool and Useful
Cool and Funny
Cool and Length

- Thus, we can say that the reviews marked cool tend to be curt, not very useful to others and short. Whereas, there is a positive correlation between:

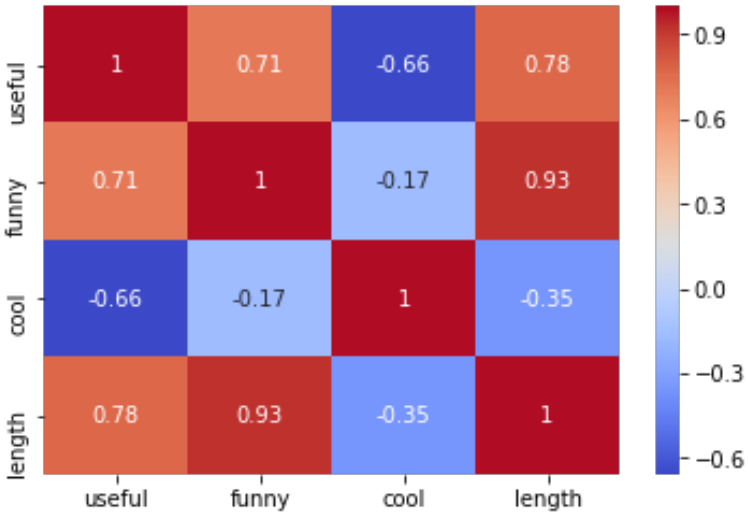
Funny and Useful
Funny and Length
Useful and Length

- Thus, we can say that longer reviews tend to be funny and more useful.

- We plotted graph between number of review v/s stars and observed that our dataset is dominated by 5 star reviews.

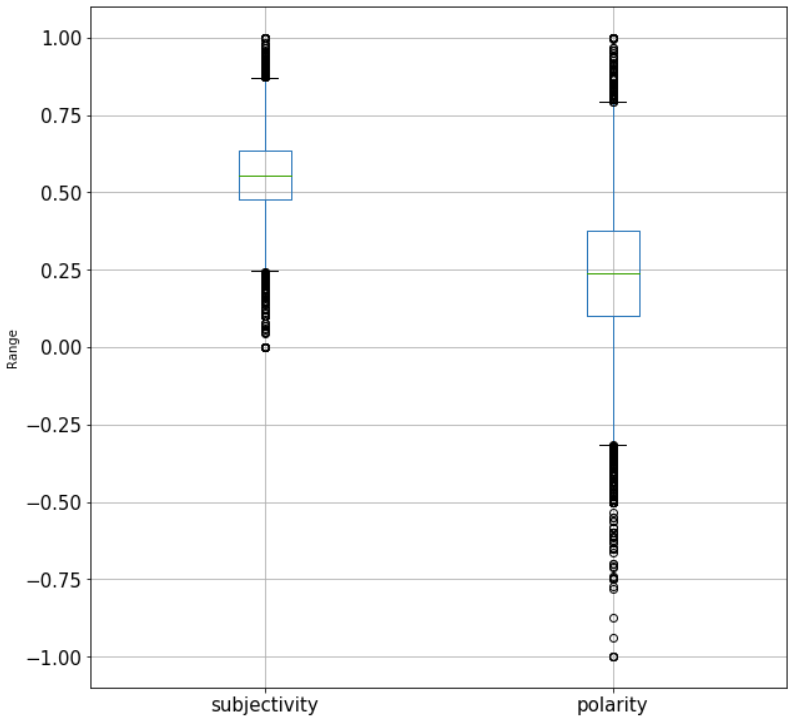


- We plotted heat map between the steal, and saw the following observation:



- Their are total of 4618 unique businesses reviewed in the above dataset of 10,000.

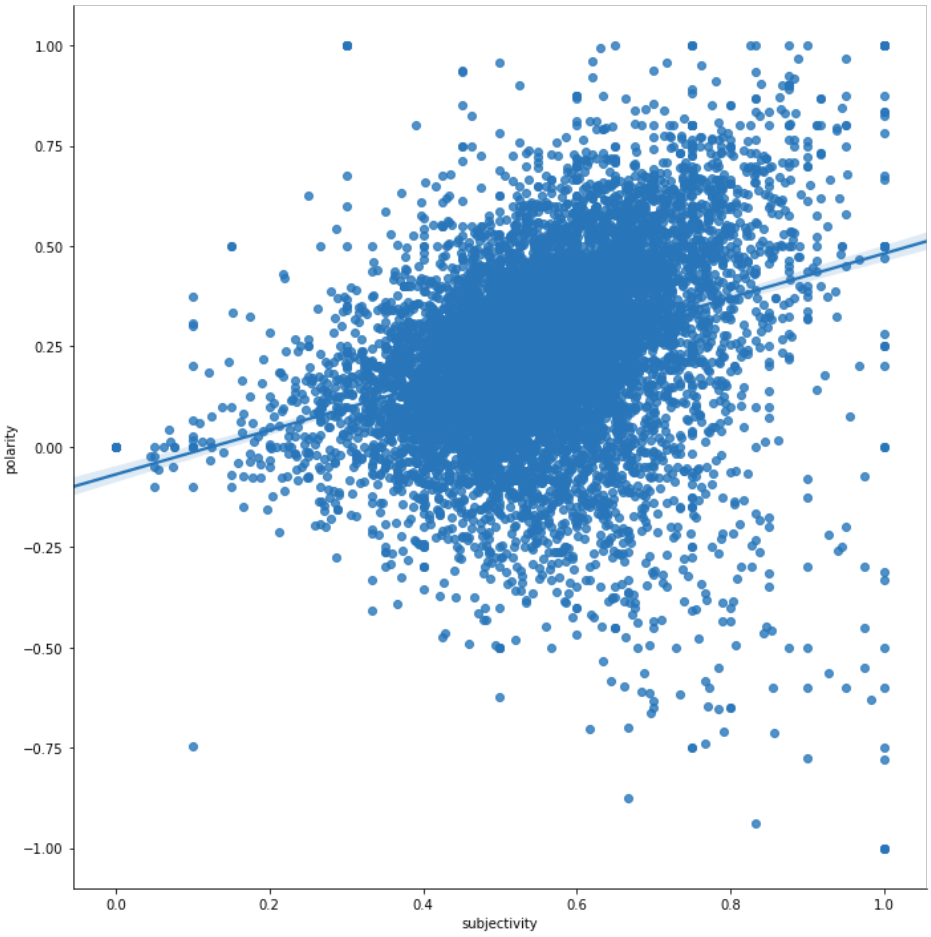
- We performed first type of Sentiment analysis using **textblob**. We checked for the polarity and subjectivity of the text reviews. Polarity — It simply means emotions expressed in a sentence, across a range of negative, to positive.
- Subjectivity — Subjective sentence expresses some personal feelings, views, or beliefs.
- So my program has confirmed to me that all the 10000 records are there and gave me a mean polarity of 0.24, which is good that means as an average, most people are in between neutral to positive with the services. And as you can see the 50% Value which means the median is above zero i.e., 0.24.



- The covariance between the two variables is 0.0287693. We can see that it is positive, suggesting the variables change in the same direction as we expect.

- We can see that the two variables(polarity and subjectivity) are positively correlated and that the correlation is 0.69351. This suggests a high level of correlation, e.g. a value above 0.5 and close to 1.0.

- The following plot shows a positive correlation between Subjectivity and Polarity. Meaning, as subjectivity increase, the polarity in the response increase too, Or in other words, the more strong feelings are expressed, the more the overall comment is subjective.


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
polarity
0.0
0.2
0.4
0.6
0.8
1.0
subjectivity