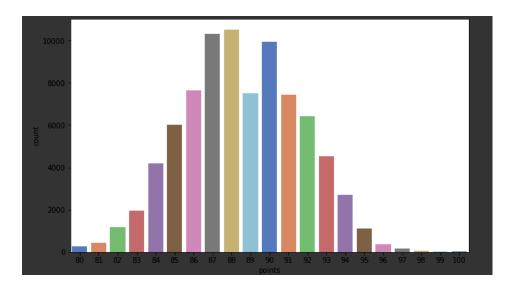
ML Challenge:

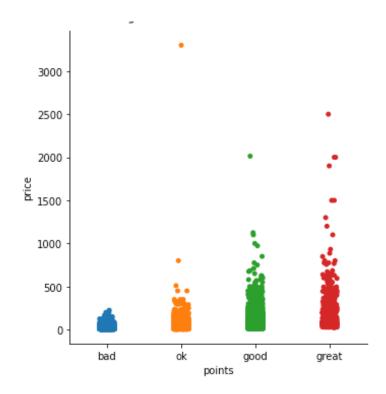
Github: https://github.com/parthvishah/Projects

Summary of the steps I have done:

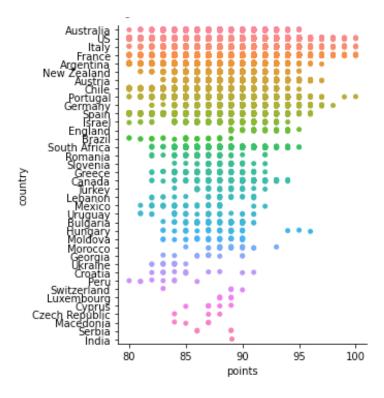
- 1 Import the train/test files.
- 2- Find correlation between various variables



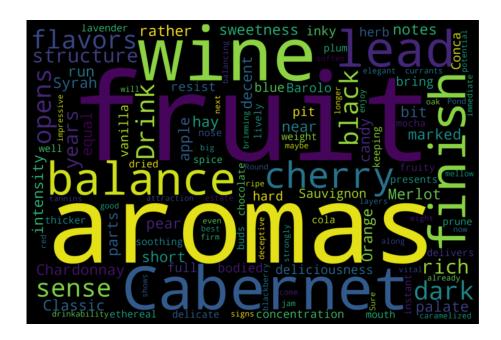
Distribution of points.



Correlation between price and points after I segregated points into bad, ok, good and great.

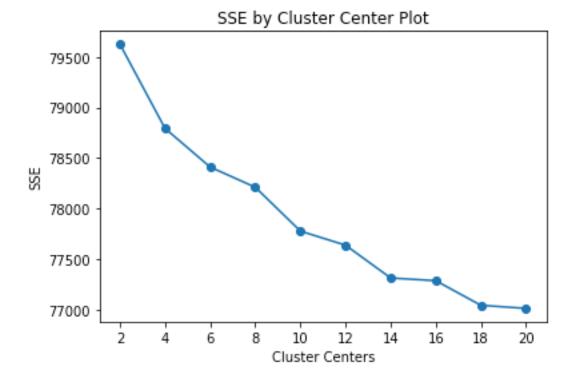


Distribution of country with respect to points.



Frequency of words appeared in the review.

- 3 Preprocessing of texts
- Removal of Stopwords
- Removal of html scripts
- · Removal of emojis
- Tokenization
- Stemming
- 4 TFIDF of the preprocessed_review.
- 5 Fit test data to the tfidf trained to training set.
- 6 Achieved an F1 score of 70.



My second approach was replace each review with a cluster value after applying K-Means to the preprocessed text.

This gave a very low accuracy

I have used SGD and RandomForest to classify the data.

I believe I could have achieved a better metric value if I could have performed GridSearchCV But the RAM was constantly crashing.