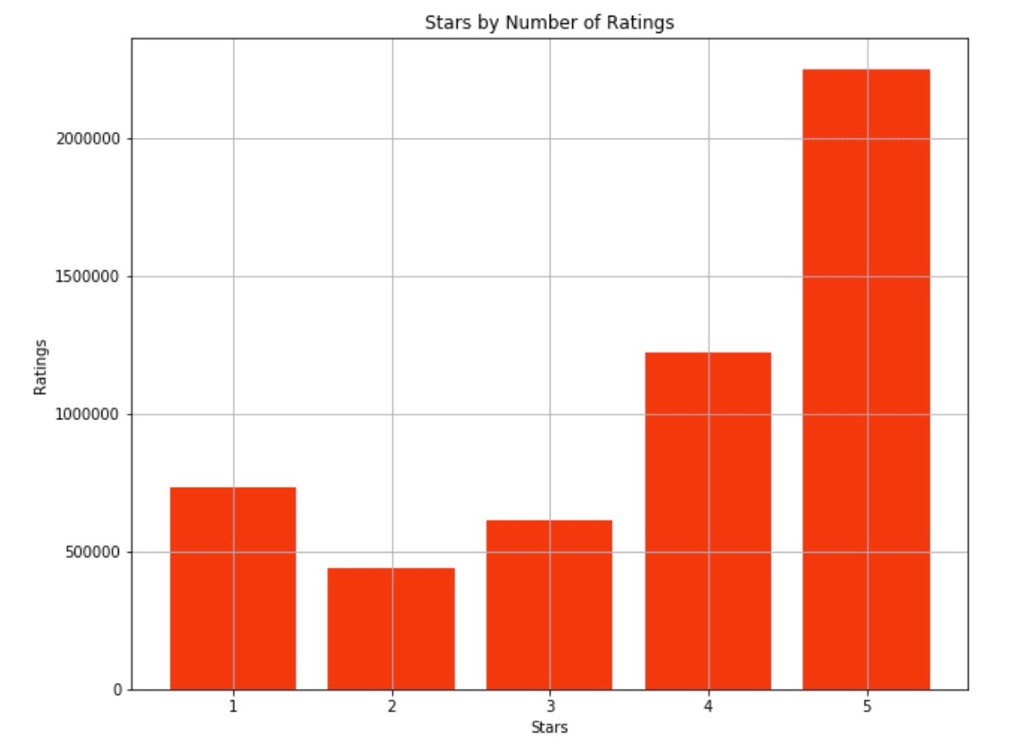
Using machine learning models, what can we learn from reviews on Yelp? From a database of over 3 million reviews of businesses on Yelp, we take a look at how our machine learning models can predict not only the positivity/negativity of reviews, but also how language varies based on location and type of business.

**Modelling Yelp Reviews**

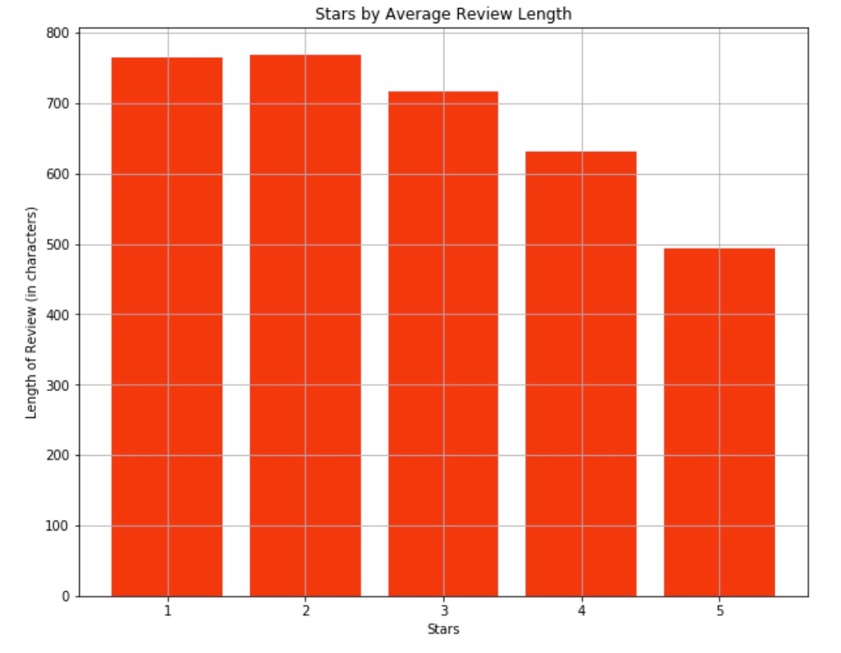
Yelp releases a subset of their review and user data to the public for database and machine learning purposes. Given the nature of these reviews, that each review is mapped to a star rating for a given business, this is a fantastic opportunity for doing NLP and other kinds of machine learning. The goal of my project is to use these reviews to form predictions about first and foremost, the star ratings, but also the supplementary data (time of day, geographical location, etc.)

**The Data**

The dataset consists of just over 3 million reviews, each accompanied by some user data and a star rating. The star ratings are distributed as follows:



The classes are fairly uneven. The plurality of reviews are 5-star and follow in descending order, with the exception of the 1-star reviews. On the other hand, reviews have fairly steady character counts across star-rating categories.



Even at 500 characters in the 5-star rating group, this is approximately 100 words. This gives us a sizeable amount of longer-utterances with which to do our analysis.

**The Analysis**

The size of the dataset will require us to further subset the data to train simpler models local before attempting to account for the wider dataset using an online computing platform (AWS). The first goal will be to train a model that accurately predicts whether a review is positive or negative. From there, we can further subdivide those two categories into the ones that are found in the data (i.e. star ratings).

We will begin with simpler models such as logistic regression models and machine learning models like K Nearest Neighbors, and then move to more complex ensemble methods as the data requires. Given the number of dimensions in text data, the more complex methods will be what requires solutions such as cloud computing.