Yelp Data Analysis

By rf1999, rrc370, wb630, spm443

About Yelp

- Yelp is an online platform which connects users to local businesses.
- Users can in turn post reviews, give ratings etc about the businesses on basis of the service they experienced.
- There are also events, event reminders and special offers from businesses and you can even make friends with other yelpers, just like Facebook or MySpace. This allows you to send messages to other users or even "follow" them, which allows you to see a specific reviewer's posts before any others when you're looking at a business.
- Yelp appears to be a place where users go to give a balanced opinion about their personal experiences.
- The vast amount of data on yelp provided by millions of users can be analysed for deeper insights and visualisations to output the underlying information.

Dataset

- https://www.kaggle.com/yelp-dataset/yelp-dataset\

  Yelp business json has entries such as business name, category, location, rating stars, no. of reviews, open time and 39 detailed features to rate the business, like TV, WiFi, environment-friendly etc.

  Yelp tip data includes ID, review text and date.

  For Review json, Yelp gives data like ID, rating stars, review data, review text, and number of reviews marked by 'cool', 'useful' and 'funny'.

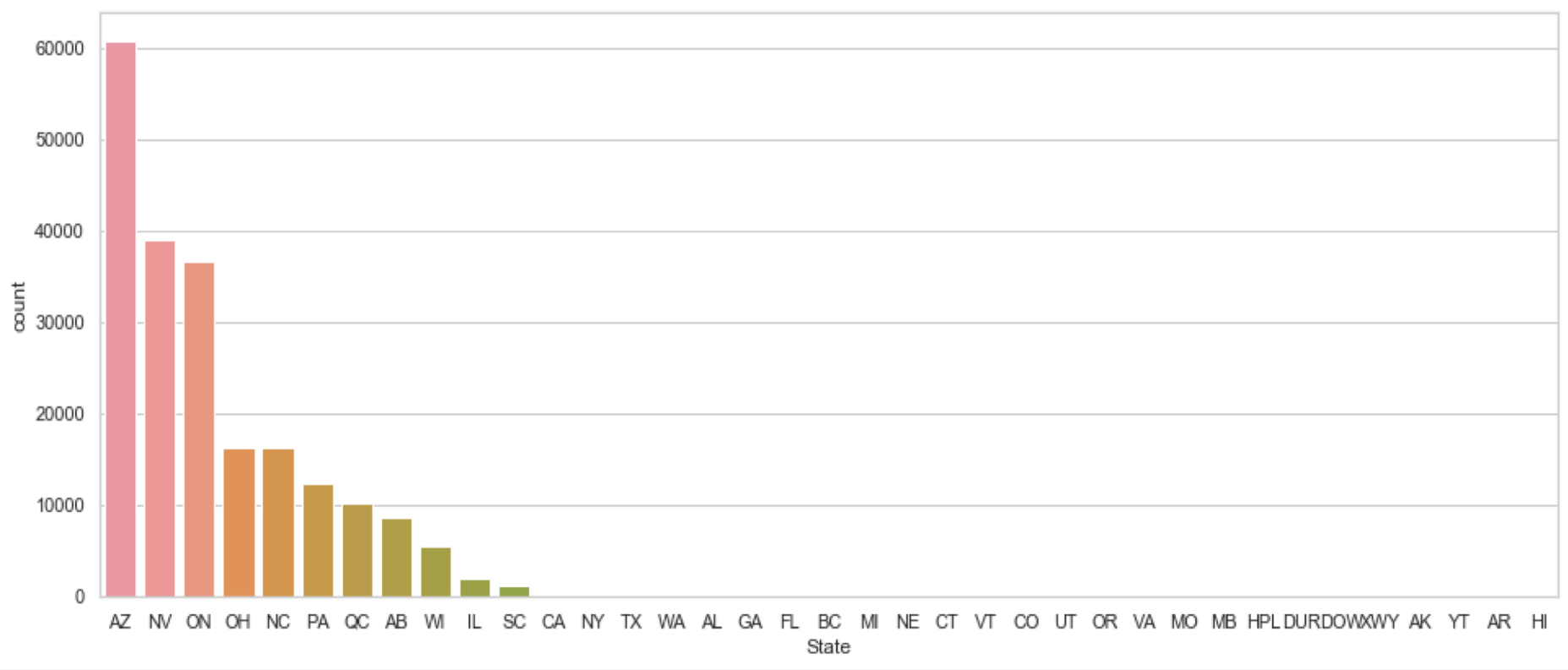
  Yelp user data has information on consumers, for example, ID, their average rating stars, count for reviews marked as 'cool', 'useful', 'funny' and the Yelp account information

Tech Stack

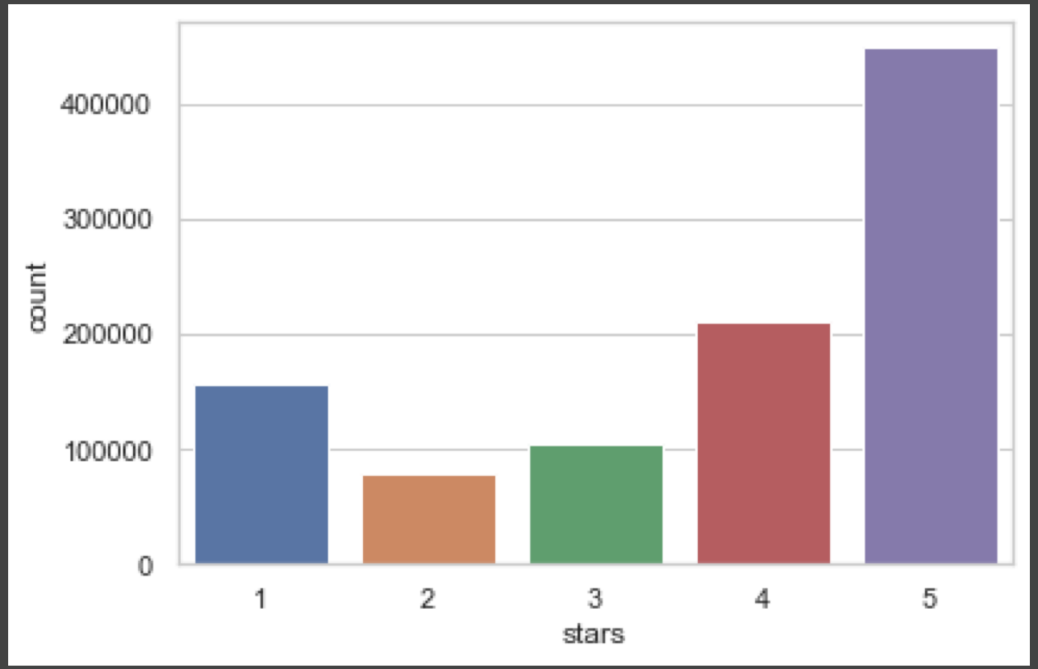
Dataset Handling and parsing (Uniqueness)

- Unlike most of the projects out there which are based on MongoDB and which only work on a subset of the dataset, we tried to work on the whole dataset by converting our JSON files to CSV format using R, which efficiently converted the huge 2-3 GB files to <700 Mb
- R provides a package jsonlite which efficiently transfers file format with just one line.
- In our case, although the dataset was about 4GB in size, using R we converted the largest file to 2.6GB in size.
- We could've used subsets of the whole dataset but that would not allow us to extract the underlying insights which we achieved with the whole dataset.

Analysis of number of reviews by state



Number of reviews by star ratings



Star Rating	Count
1	156690
2	78883
3	104973
4	210363
5	449091

Interesting Observations

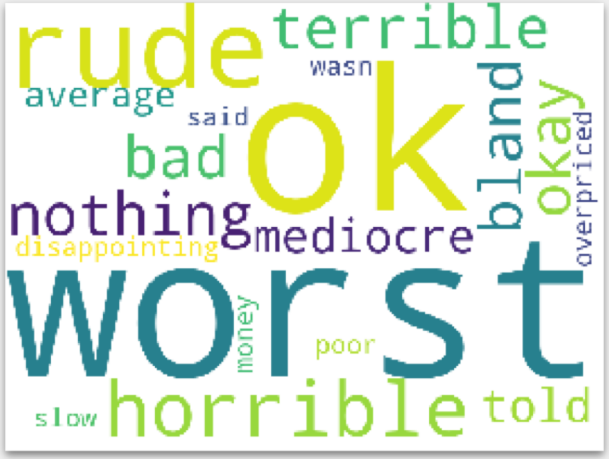
- We can observe that the main cluster of the reviews are on the positive areas, meaning people tend to leave reviews when they get a good experience.
- Also it can be observed that the number of bad or (extreme) negative reviews (1 star) are much more than two or three star (neutral reviews) also showing a correlation between when people post reviews based on their experience.
- ***So, people tend to post reviews when their experience was positive i.e. 4 or 5 start or when their experience was extremely negative i.e. 1 star.***

Sentiment Analysis

- We wanted to predict the sentiment of user (positive or negative) from the star ratings, features or review posted by the user for particular business. We made a SVM model in which we relabeled the review star rating 0-3 as 0 and 4-5 as 1 and based on the review text we tried to predict the positve and negative reviews.
- We also built a Logistic regression model in which we regress on the provided star ratings (labelling 0-3 as 0 and 4-5 as 1) and multiple features such as business location down to city, state, latitude, longitude, review count, reviews tagged as ‘funny’, ‘cool’ or ‘useful’, average review count etc to predict the start ratings for the businesses.

Conclusion (Sentiment Analysis)

- We found out that LR model can predict efficiently whether a business will receive high or low rating stars with an accuracy of 82.6%. Also that SVM model can predict the whether a review is positive or negative in nature with an accuracy of 88.632% F1 Score.
- We extracted the cloud of most positive and negative words from the svm weights visualised below.



Correlation of features with business ratings

- In the business dataset, we saw that there are many additional features available which might have an impact on the user rating.
- In order to analyse that we created multiple computational models to extract the underlying insight.
- We analysed features like WiFi_No, environment-friendly, TV etc to understand their impact on reviews and ratings of restaurants particularly and to do so efficiently we removed all non-restaurant data.
- We built a custom one-hot encoder to preprocess the data and relabelled star rating again (0-4 as 0 and 5 as 1)

Conclusion (Correlation of features with business ratings)

- The results of this analysis were not same as the previous models, despite significant feature, model experimentation and training time, a prediction engine for extracting correlation between features and business did not come out strongly.
- Within the context of restaurant data, these business attributes are not particularly significantly correlated with the number of stars in a review and thus doesn't affect the review sentiment in any significant manner.

Helping Businesses

- Our research could serve as a valuable tool for businesses in need of algorithmic sentiment prediction.
- In lieu of simple rating systems, businesses could use the models we've defined to parse text for sentiment and respond accordingly, either directly or in their business decisions.
- The models we've developed are built on a fully scalable foundation, and could be adapted to a wide variety of business applications.

Possible Extensions

- Use of AWS EMR for smoother integration with S3 bucket and more computing power
- Analysis on a subset of the total dataset on basis of geolocation using the same models to find more cohesive correlations.
- Creating a end to end toolkit to provide stakeholders with meaningful insights from the dataset for business purposes

Thank You