AIT 664
Final Project Report
Yelp's Dataset
Meenakshi Meenakshi

Introduction

Yelp helps people to connect to great local businesses. It is one of the biggest local search engines with more than 70 million desktop users and many more on the mobile("Factsheet" n.d.). There are over 177 million reviews on the website("80 Amazing Yelp Statistics" 2014) with over 19% reviewed only for Restaurants. In this paper, the aim is to find the attributes or features that affect the rating of a restaurant. To predict the rating for a restaurant operating on the historical data, such as the review text, restaurant's statistic, user check-ins etc. by examining the data set provided by Kaggle.com("Yelp Dataset" n.d.).Also, to discover the major cities where the businesses flourish, the keywords used in the reviews and the major categories that exist in the business. Three machine learning algorithms are used, random forest, KNN (K nearest neighbor), sentiment analysis to scrutinize the dataset and uncover insights.

Initial Requirements:

Tools and Techniques:

- Downloaded and Installed Tableau and tableau Prep. ("Tableau for Students" n.d.)
- Performed initial Training "Getting Started with Tableau". Also looking at the training videos in order to get familiar with the tool.
- R proficiency -Intermediate
    - For Data preprocessing, cleaning and analysis
    - Libraries used include randomForest, caret, mice, ggplot2, VIM etc.
- .Net - Beginner
- Python proficiency- Intermediate
    - Libraries like panda, NumPy, seaborn, and nltk, Text Blob for sentiment analysis of reviews.

DATA

Data Collection: Data was collected from Kaggle("Yelp Dataset" n.d.)

- The data set is a subset of Yelp's actual business data and consist of information about businesses in 11 metropolitans from 4 countries.
- There are files which are Business, Reviews, Check-in, Attributes and all of them are in both csv format.
- After milestone 2 also got the data from yelp in Json format and used the Business_Academic.Json for the business and attributes dataset.("Yelp Dataset" n.d.)
- The files contain approximately **5,200,000** user reviews
- Looking at the initial data, it needs cleaning as:
    - the Business file contains information about more than just restaurants.
    - There are missing/null values in columns (see appendix, figure 1)

DATA CLEANING AND PREPROCESSING

Data Preparation

The initial analysis which was done on the Business file (see appendix, figure 1) revealed that the data needed preprocessing and there are over 50 attributes in Business file and several others in other files, thus there has to be feature selection. The columns of all the files are present in the Appendix.

Merging the files to create a single dataset for Business attributes

- Merging Business, attributes, and check-in files
- Business academic Json file was converted to csv using .Net console application. The application deserialized to convert the Json object schemas into classes which were then written into a CSV file. The file had three schemas Business, Attributes and Hours. Business and Attributes schemas were used for the further analysis and converted to a single file.
- Filtered the data for only Restaurants
- Merged the Check-in Information from the Check-in file using the index formula in excel matching the business_id from both the files.
- Cleaned the Csv file using excel and removed the columns with less than 4% data
- Converted all the columns with binary predictors, TRUE and FALSE to a 0 for False and 1 for True.

Reviews Dataset Preprocessing

- Filtered the reviews for only restaurants.
- **Stop Word Removal**

  The preprocessing steps assist in removing ineffective words such as: the, a, in, an, are, at, for, I, then, etc. Some special characters included in the data file containing reviews were mostly punctuation marks dashes and spacing. Removal of these words and characters can result in cleaner data for any model. The stop words were removed using a predefined set in the NLTK package in Python. Other special characters were removed using the regular expressions and functions in python.

```
In [19]: ##Removing the stopwords from reviews 1 star
         from nltk.corpus import stopwords
         stop=stopwords.words('english')
         df_1_star['text']=df_1_star['text'].apply(lambda x:" ".join(x.lower() for  x in str(x).split() if x not in stop))
         df_1_star['text'].head()

         /Users/meenakshimadan/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:4: SettingWithCopyWarning:
         A value is trying to be set on a copy of a slice from a DataFrame.
         Try using .loc[row_indexer,col_indexer] = value instead

         See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
           after removing the cwd from sys.path.

Out[19]: 6     this place disgusting. the components hotdog h...
         40    i'm giving 1 can't give 0 yelp. nasty fried bo...
         43    j'avoue n'avoir jamais compris cet intérêt aut...
         52    the wilensky special tastes like bologna, whit...
         59    if came jail 35 years...come here! you've lost...
         Name: text, dtype: object
```

Figure 1. Stop word removal

- Converted all the data to lowercase
- Removed punctuations and non-useful characters

```
In [24]: #Removing unuseful characters from 1 star
         df_1_star['text']=df_1_star['text'].str.replace('[^\w\s]','')
         df_1_star['text'].head()

         /Users/meenakshimadan/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
         A value is trying to be set on a copy of a slice from a DataFrame.
         Try using .loc[row_indexer,col_indexer] = value instead

         See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy

Out[24]: 6     this place disgusting the components hotdog ha...
         40    im giving 1 cant give 0 yelp nasty fried bolog...
         43    javoue navoir jamais compris cet intérêt autou...
         52    the wilensky special tastes like bologna white...
         59    if came jail 35 yearscome here youve lost hair...
         Name: text, dtype: object
```

Figure 2. Removal of non-useful characters

Missing Data

The Completeness of data was an issue**.** The original dataset contained more than 60 binary predictors. The granularity of the dataset was pretty good. One of the main challenges was dealing with the missing data. For a lot of predictors like WIFI, Open24Hours, smoking etc. had almost 80-90% missing data, and thus would not be a good fit for any model. Furthermore, just neglecting the missing values was not an option as it would remove most of the observations. Data with no missing values was only 3% of the total dataset. Thus, preprocessing the missing data was irrefutably important.

	Column_Names	percentage_null
1	Business_id	0
2	Name	0
3	Categories	0
4	stars	0
5	review_count	0
6	is_open	0
7	longitude	0
8	Latitude	0
9	city	0
10	state	0
11	postal_code	0
12	checkins	13.3556439633241
13	GoodForKids	10.3420014684037
14	RestaurantsReservations	9.69146447658238
15	GoodForDessert	28.1438352655933
16	GoodForLatenight	28.1438352655933
17	GoodForLunch	28.1438352655933
18	GoodForDinner	28.1438352655933
19	GoodForBrunch	28.1438352655933
20	GoodForBreakfast	28.1438352655933
21	GarageParking	15.6521590656855
22	StreetParking	15.6572814042037
23	Validated	15.6572814042037
24	LotParking	15.6572814042037
25	ValetParking	15.6589888503765
26	Caters	26.6310379565284
27	NoiseLevel	0
28	RestaurantsTableService	47.4960301876483
29	RestaurantsTakeOut	7.63569928458005
30	RestaurantsPriceRange2	9.4421773353595
31	OutdoorSeating	14.0010586166271
32	BikeParking	18.5975037136954

32	BikeParking	18.5975037136954
33	HasTV	15.8690047296259
34	WiFi	0
35	Alcohol	0
36	RestaurantsAttire	0
37	RestaurantsGoodForGroups	10.1900387590281
38	RestaurantsDelivery	10.2532142674202
39	BusinessAcceptsCreditCards	30.4915737531374
40	BusinessAcceptsBitcoin	66.5869858452712
41	ByAppointmentOnly	70.215308962385
42	AcceptsInsurance	71.0843990643195
43	GoodForDancing	66.1447572865265
44	CoatCheck	67.4936397630065
45	HappyHour	65.6274010961804
46	BestNights	0
47	WheelchairAccessible	60.2284562979152
48	DogsAllowed	65.4293373401403
49	BYOBCorkage	0
50	DriveThru	66.354773165776

Figure 3. Missing Data Percentage

Missing Data Imputation

The predictors with less than 50% missing data were chosen for imputation. Used the mice function from the package mice (Stef van Buuren, Karin Groothuis-Oudshoorn (2011)) in R to impute the missing data. Method chosen was PMM (predictive mean matching) and the iterations done were 5. MICE PMM produces imputed values which are very close to the real values and is a better approach than other mice methods.

Since the dataset had around 59000 rows and 51 predictors, it had to be broken down into 4 parts for mice to be able to impute. Then all these parts were combined after imputation.

```
> #####imputation of 3rd part and then picking the final result
> tempData3 <- mice(data_part3,m=3,maxit=5,meth='pmm',seed=100)

 iter imp variable
  1   1  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  1   2  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  1   3  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  2   1  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  2   2  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  2   3  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  3   1  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  3   2  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  3   3  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  4   1  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  4   2  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  4   3  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  5   1  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  5   2  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
tCards
  5   3  BikeParking  HasTV  RestaurantsGoodForGroups  RestaurantsDelivery  BusinessAcceptsCredi
```

Figure 4. Missing data imputation using mice package R

The final predictors: Review_count, GoodForKids, GoodForDessert, GoodForLatenight, GoodForLunch, GoodForDinner, GoodForBrunch, GoodForBreakfast, GarageParking, StreetParking, Validated, Lot Parking, ValetParking, Caters, RestaurantsTableService, Outdoor Seating, Bike Parking, HasTV, RestaurantsGoodForGroups, RestaurantsDelivery, BusinessAcceptsCreditCards

EXPLORATORY DATA ANALYSIS

Once the data was cleaned the first step is to have some graphs to understand the underlying relationships and patterns in the dataset.

The package pandas-profiling was used for initial exploration of the dataset. A html version of the whole data profile was saved. It gives the total number of variables, the type of variables, % of missing data in the whole dataset and the variable types.

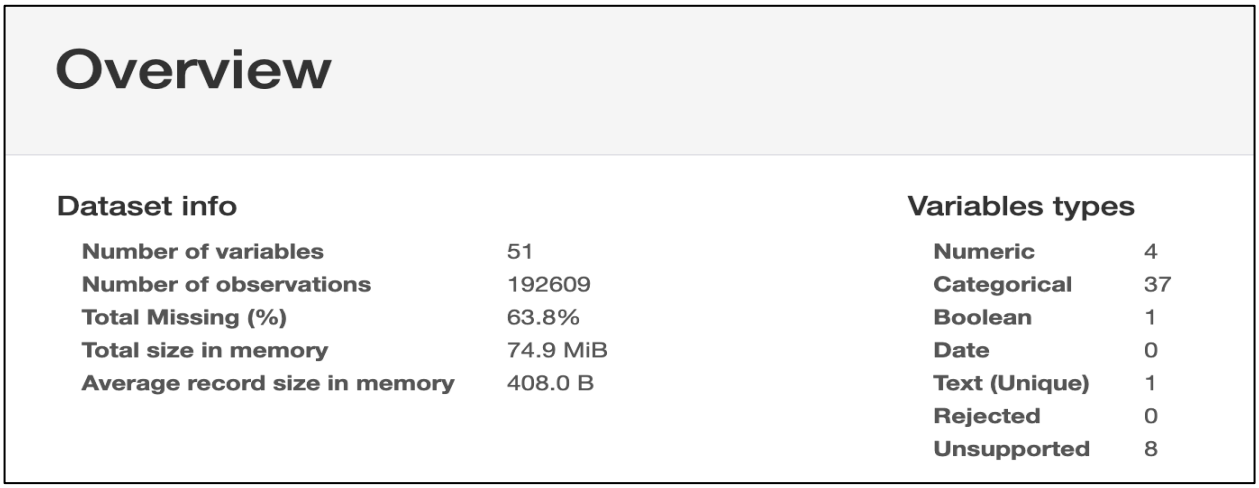


Figure5. Data profile for the Business attribute dataset

Variables

The variable details are given in the Figure 6, with the name of variable, and the number of unique values, % of missing values, distinct values etc.

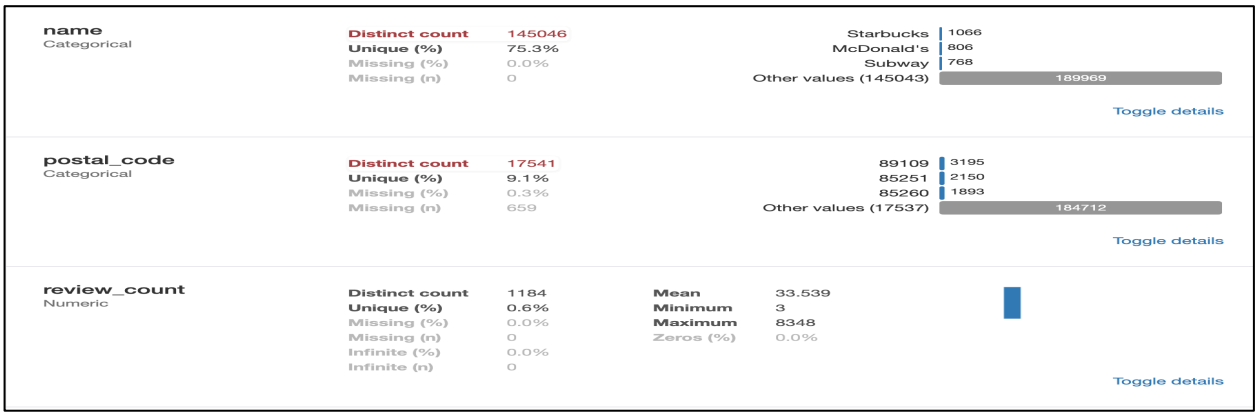


Figure 6. Data profile for the Variables in Business Attribute dataset

Business Attributes

For the business and its attributes few graphs were built to understand what the major categories in business are. The package seaborn from python was used for the graph below describes the top business categories with most reviews. Looking at the graph it is clear that the highest reviews are for the category "Restaurants", and second being "Shopping", and the third "Food". Restaurants are almost 31% of the total dataset, the total dataset is 192610 and out of these almost 59000 are rows where the category is restaurant.

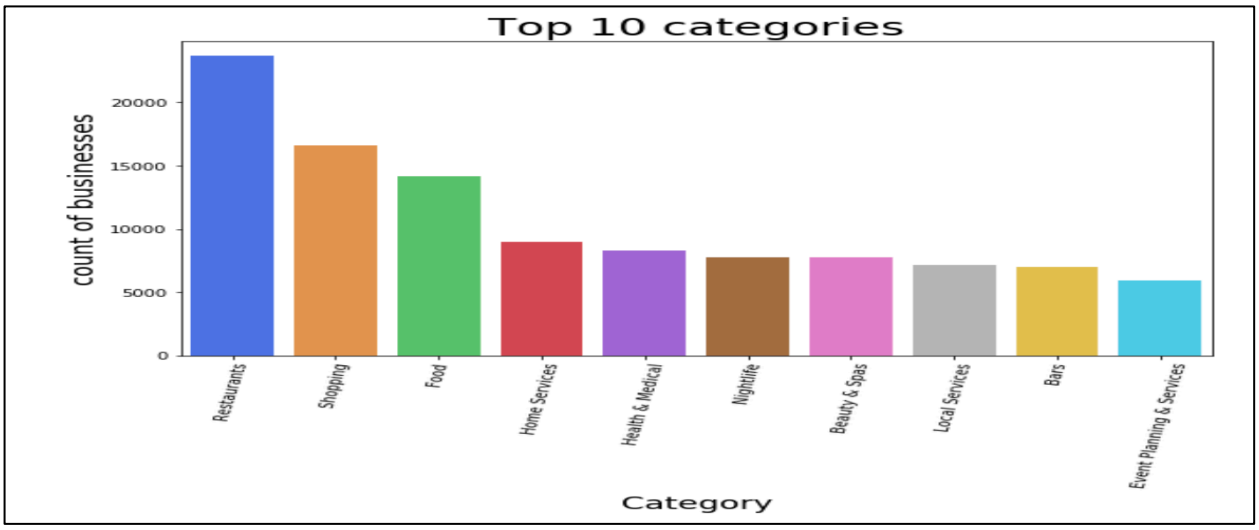


Figure 7. Top 10 categories of Business

Business Reviews Vs Star Ratings

To understand the relationship of the variable Star ratings with the number of reviews, a bar plot was plotted using the seaborn library of python. Maximum number of stars given are a "4" or "3.5", and the others are less.

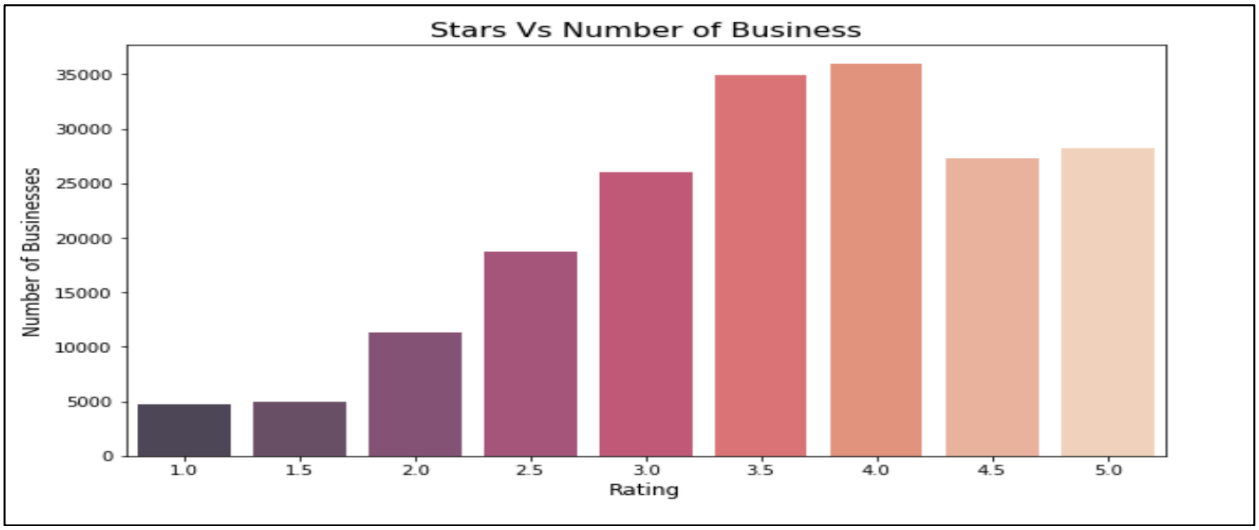


Figure 8. Stars vs Number of business reviews

- Which **cities have the highest number of reviews**?
  Las Vegas and Toronto have the highest number of reviews. Which could be because Vegas is a tourist destination and the number of people eating outside are far more than other cities. The bar chart was plotted using the seaborn library of python.

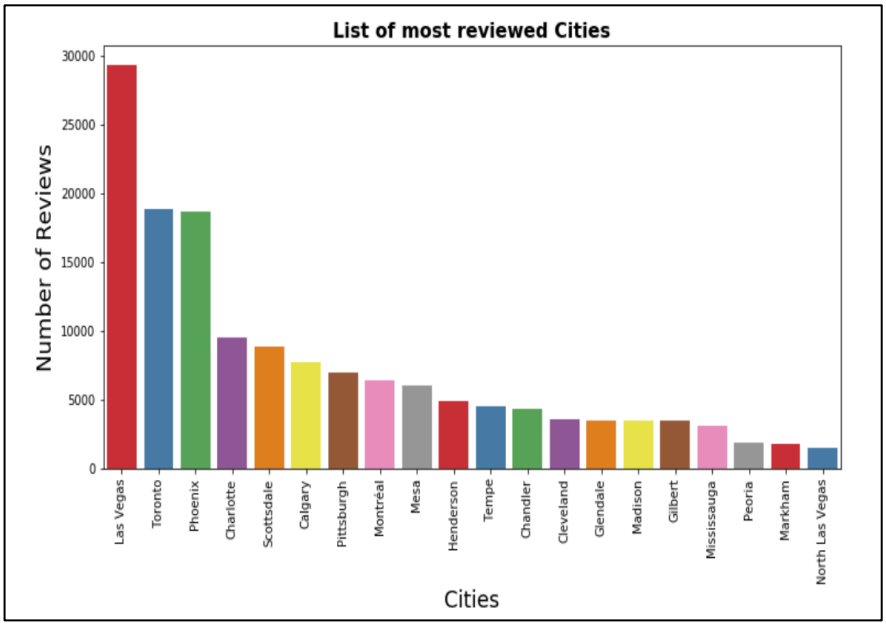


Figure 8. List of most reviewed cities

- Which **States have the highest Number of reviews in US?**
  Build a heat map in Tableau for the US states with the number of reviews, the higher the number the darker the color, the lower number the number has less color.

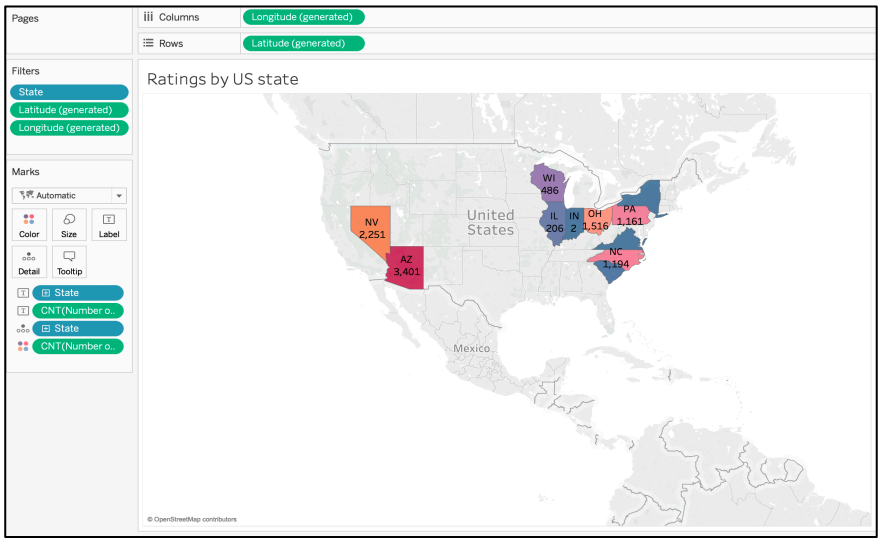


Figure 9. Map for the highest rated states

The bar chart in Figure 10 was generated using the seaborn package from python; the reviews by states not just in US but all the countries that have are present in the data, Nevada has the 2nd highest number of reviews and Illinois the lowest of these.

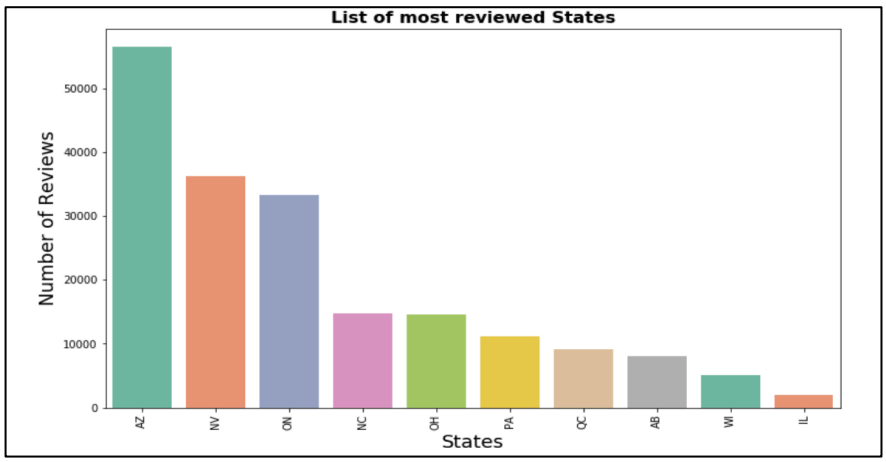


Figure 10. List of most reviewed states

States with most checkins

The number of checkins for the states are highest for AZ, the same where the number of reviews counts are higher. Tableau was used for this histogram. The x axis represents the states and the y axis has the number of checkins, and the numbers on the histograms are the count of check-in in that state.

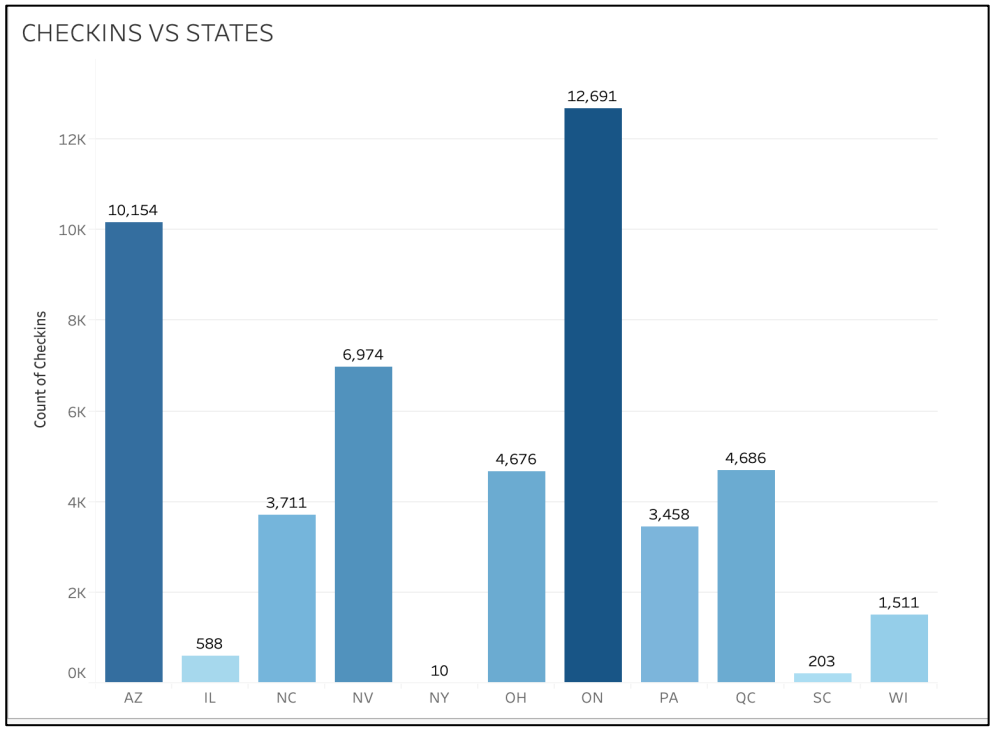


Figure 11. State Vs Checkin

Review Dataset

The second part of the dataset is about the reviews written for a particular restaurant. Since the data set was huge, divided the whole dataset into the 5 parts as per the star ratings for the analysis.
1 star → [=1 and <2]
2 stars→ [=2 and <3]
3 stars→ [=3 and <4]
4 stars→[=4 and <5]
5 stars→[=5]

Word cloud

To gain a better understanding of the huge textual data the first thing was to have a look at the keywords and their frequency. Word clouds or tag cloud work in a simple but efficient manner of displaying the most frequently occurring text in the data along with the frequency by the use of the different colors and the text sizes. The keywords used are highlighted by the different colors. The word cloud library from python was used to produce the images in figure 12 and 13.

The keywords used mainly in a review of 5 star can be seen in the word cloud as being highlighted in the big text size and different color than rest of the words. The figure 12 demonstrates the difference in the use of words for different star rating. For the 1and 2-star ratings the review contains the keywords like “time”, “order”, “never”. However, in 3star reviews words like “great” appear.

Figure 12. Word cloud for 1,2- and 3-star ratings

In the Figure12. the 4- and 5-star reviews have the keywords “great”, “really”, “menu”, “love” and “delicious” which cannot be seen in the reviews for the 1 to 3-star ratings, thus people do use different words frequently for different ratings.

Figure 13. Word cloud for 4- and 5-star ratings

Correlation Between review length and Star rating

Correlation between the length of review and star rating is done using the histograms. In Figure.14 the seaborn package was used from python along with Facet to develop five graphs side by side. The overall length of text doesn't seem to be a good indicator of the star rating, the high spike in star =5 graph is only because the number of reviews is more for rating 5.

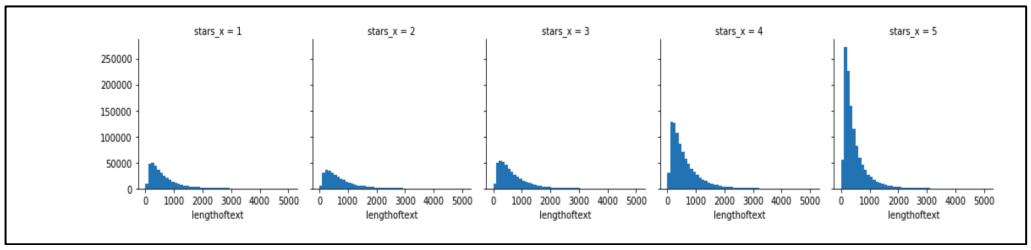


Figure 14. Histogram for the length of review Vs number of reviews for each Rating

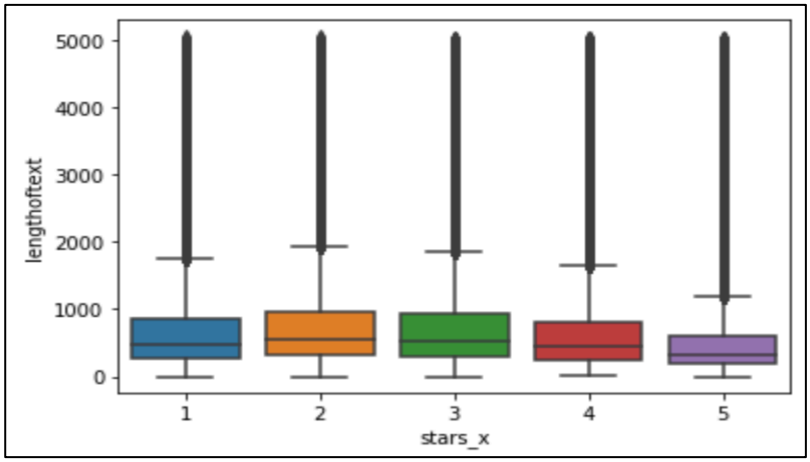


Figure 15. Boxplot for length of reviews and stars rating

The box plot in the Figure.15 suggest that the length of 2- and 3-star rating are more than 4- and 5-star ratings. It looks like as if there lower the length of reviews are either for the very low rating which is 1, or very high ratings i.e. 5. The seaborn package was used from python.

Top 10 Restaurants with maximum views

The maximum number of reviews were for “Hash House A Go Go” and “McDonald’s” and “Chipotle”. With the high-end restaurant’s like “Gordon Ramsay BurGR” and “Mon Ami Gabi” later in the list suggest people like to go to affordable places a lot more than these.

	Name
Hash House A Go Go	9322
McDonald's	9161
Chipotle Mexican Grill	7569
Mon Ami Gabi	7362
Bacchanal Buffet	7006
Wicked Spoon	5951
Buffalo Wild Wings	5933
Gordon Ramsay BurGR	5448
Earl of Sandwich	5442
In-N-Out Burger	5374

Figure 16. List of top 10 highly reviewed restaurant’s

Sentiment Analysis

Sentiment analysis is a way to find the underlying context or tone of the sentence. Used the Text Blob package to find the sentiments of the reviews. The number of positive sentiment reviews are much more than the negative reviews which is visible in Figure 17.

Figure 17. Sentiment of reviews

TECHNIQUES AND ALGORITHMS

According to the data analysis from exploratory data analysis the star ratings cannot be predicted from the length of a review , or sentiment of a review , now exploring the business attributes for prediction .The star rating is divided into a binary classification as Good(restaurant having 3 or more stars) and Bad (restaurant

having less than 3 stars) and is saved in a new column called “Restaurant Type”. The business, attributes, check-in datasets combined are used. The dataset is divided into two parts the training (80%) and test set (20%). Two machine learning algorithms were used for the prediction of star ratings.

All the categorical variables were converted as factors to be used by the algorithms. All the numeric values were kept as is for the analysis. Variables like state, city, postal_code, Name etc. were not used in the algorithm as they are textual data it would not be contributing to the model performance .Business_id was also removed from the analysis as it is just a unique identifier and would not be helpful in analysis. The data for restaurant’s that are open are selected.

Random Forest

It is a supervised learning algorithm. These are ensemble learning methods and operate by building multiple decision trees and then merge them to get the most precise answer for a class. The algorithm makes it easier to understand the results and the importance of the predictors on the final results.

The results on the test dataset is as below, on the left is the accuracy with imputation which is 80% and on the right is the accuracy without imputation and with no missing data after removing all the null rows in the data, with accuracy of 41%.

Random Forest – Imputed data

```
               Accuracy : 0.8097
                 95% CI : (0.8011, 0.8181)
    No Information Rate : 0.8097
    P-Value [Acc > NIR] : 0.5067

                  Kappa : 0.0108

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.008833
            Specificity : 0.997924
         Pos Pred Value : 0.500000
         Neg Pred Value : 0.810768
             Prevalence : 0.190276
         Detection Rate : 0.001681
   Detection Prevalence : 0.003361
      Balanced Accuracy : 0.503379
```

Random Forest – no missing data

```
               Accuracy : 0.4198
                 95% CI : (0.3626, 0.4786)
    No Information Rate : 0.372
    P-Value [Acc > NIR] : 0.05218

                  Kappa : 0.1416

 Mcnemar's Test P-Value : NA
```

Figure 18. Random Forest results

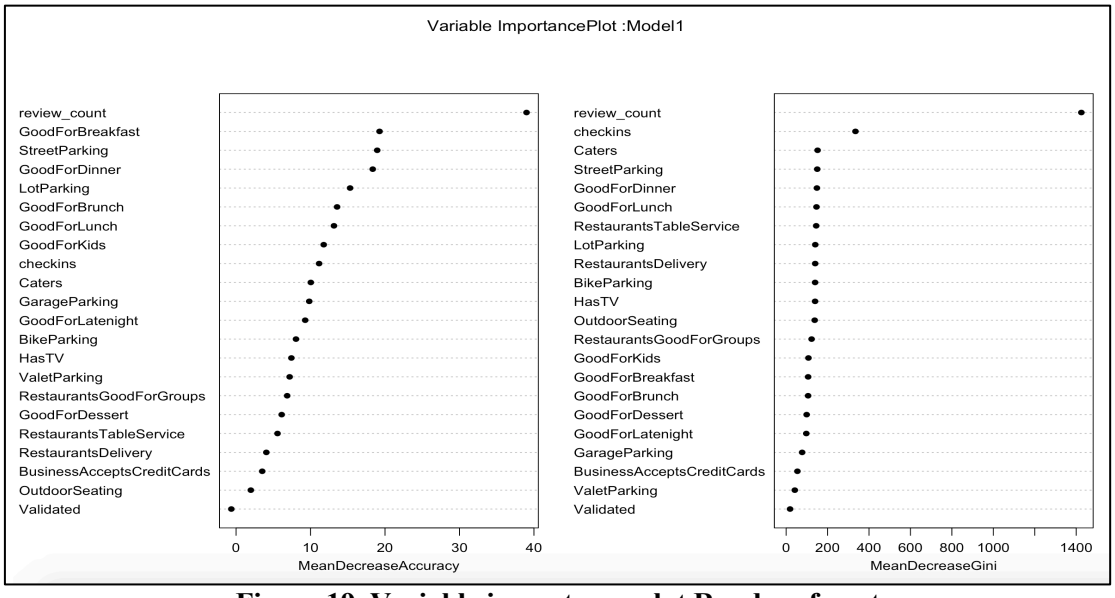


Figure 19. Variable importance plot Random forest

The variable importance plot above suggests that the most important variables for prediction of the stars are “Review count”, “checkins” and “parking”, “Caters”, “Table service”.

KNN (K- Nearest Neighbor)

The KNN for classification predicts a new sample using the K closest samples from the training set which was supplied. “*k*-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally, and all computation is deferred until classification.”(“*K*-Nearest Neighbors Algorithm” 2019).

For Better accuracy and allowing all the independent variables to participate in the distance calculation the 5 folds cross validation was done. On the left is the summary of KNN with the prediction on the unseen test data being 80%, on the right is the k-5 with the highest accuracy value.

```
> knn.summary
  Accuracy      Kappa
0.80727666 0.01848155
> confusionMatrix(knn.prediction,testing.y, positive = "Good")
Confusion Matrix and Statistics

          Reference
Prediction  Bad Good
      Bad    27   35
      Good 1570 6696

               Accuracy : 0.8073
                 95% CI : (0.7986, 0.8157)
    No Information Rate : 0.8082
    P-Value [Acc > NIR] : 0.5946

                  Kappa : 0.0185

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99480
            Specificity : 0.01691
         Pos Pred Value : 0.81007
         Neg Pred Value : 0.43548
             Prevalence : 0.80824
         Detection Rate : 0.80403
   Detection Prevalence : 0.99256
      Balanced Accuracy : 0.50585

       'Positive' Class : Good

>
```

k	Accuracy	Kappa
5	0.7820162	0.10218556
7	0.7897796	0.09265836
9	0.7940713	0.08311435
11	0.7967725	0.07183867
13	0.8001140	0.06658759
15	0.8012245	0.05909090
17	0.8023449	0.05637699
19	0.8031953	0.05090725
21	0.8035956	0.04647468
23	0.8042058	0.04408754
25	0.8044360	0.03845967
27	0.8047961	0.03549326
29	0.8051262	0.03231883
31	0.8058565	0.03092457
33	0.8061466	0.02923000
35	0.8065968	0.02971723
37	0.8064468	0.02749373
39	0.8067469	0.02718723
41	0.8067869	0.02593972
43	0.8071871	0.02607269

Figure 20. KNN results

The variable importance plot for KNN is given in the Figure 19

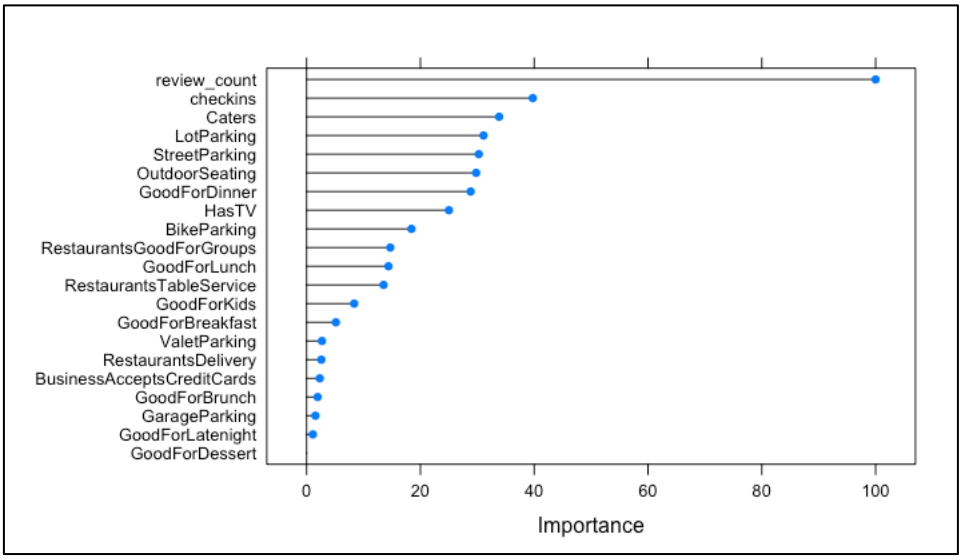


Figure 21. Variable importance plot KNN

From the Figure 21. variable importance plot the predictors which come out to be the best predictors of star rating are “review_count”, “checkins”, “caters”, “parking”, “has TV”, “Outdoor seating”. Some of these were also produced by the Random forest model.

Challenges

The major challenge for this project was data cleaning and data quality. The Business Json file contained three schemas in it and even after converting the file from Json to csv via a program considerable amount of time was spent cleaning it and merging it with other files to construct a dataset that can be used for analysis.

No doubt the number of variables for the analysis are a lot but most of these didn’t have enough data to contribute to the analysis. The quality of data is not good with missing data for some predictors in the range of 90%. Imputation of data took a lot a time as the mice package from R cannot work with a dataset with 56 columns because of the memory issues.

Conclusions

The results presented in this paper demonstrate that the reviews of the restaurants do not contribute a lot to the prediction of the rating. Of the two models performed Random forest was better in terms of accuracy. With the reviews length and sentiment not being a good predictor for Restaurant Rating; the six factors which came up as highly weighted by both the models were “Parking”, “Good For kids”, “Number of check-ins”, “Table service”, “Caters” and “outdoor seating”. There were total 20 features which have proved to be most important features in determining the rating. Thus, according to the analysis that I performed people are more concerned about these attributes than food, which means that the whole dining experience matters and not just food. This analysis can be useful for restaurant owners to understand their customers and focus on factors other than food as well to have a better rating. The results of the analysis can be more accurate if there was good quality and complete dataset present. Even with imputation the accuracy was 80% which was an unsolvable challenge in this project.

References

"80 Amazing Yelp Statistics." 2014. DMR. May 5, 2014. https://expandedramblings.com/index.php/yelp-statistics/.

Borole, Kaustubh. 2018. "Web Scraping Yelp, Text Mining and Sentiment Analysis for Restaurant Reviews." *Medium* (blog). October 9, 2018. https://medium.com/@kborole7/web-scraping-yelp-text-mining-and-sentiment-analysis-for-restaurant-reviews-ea500e1ef84d.

"Build a Simple Map - Tableau." n.d. Accessed April 11, 2019. https://onlinehelp.tableau.com/current/pro/desktop/en-us/maps_howto_simple.htm.

"Create CSV from JSON in C#." n.d. Accessed May 7, 2019. /Tips/565920/Create-CSV-from-JSON-in-Csharp.

Donges, Niklas. 2018. "The Random Forest Algorithm." Towards Data Science. February 22, 2018. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd.

"Factsheet." n.d. Yelp. Accessed May 7, 2019. https://www.yelp.com/factsheet.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

"Instance-Based Learning." 2019. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Instance-based_learning&oldid=886300144.

"Introduction to KNN, K-Nearest Neighbors: Simplified." n.d. Accessed May 7, 2019a. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.

"Introduction to KNN, K-Nearest Neighbors: Simplified." ———. n.d. Accessed May 7, 2019b. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.

"*K*-Nearest Neighbors Algorithm." 2019. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=893452483.

"Merge, Join, and Concatenate — Pandas 0.24.2 Documentation." n.d. Accessed April 11, 2019. https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html.

"Mice Citation Info." n.d. Accessed May 7, 2019. https://cran.r-project.org/web/packages/mice/citation.html.

"Natural Language Toolkit — NLTK 3.4.1 Documentation." n.d. Accessed May 7, 2019. https://www.nltk.org/.

Oliphant, Travis E. 2007. "Python for Scientific Computing." *Computing in Science & Engineering* 9 (3): 10–20. https://doi.org/10.1109/MCSE.2007.58.

"Pandas.Series.Str.Split — Pandas 0.24.2 Documentation." n.d. Accessed April 11, 2019. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.str.split.html.

Perez, Fernando, and Brian E. Granger. 2007. "IPython: A System for Interactive Scientific Computing." *Computing in Science & Engineering* 9 (3): 21–29. https://doi.org/10.1109/MCSE.2007.53.

"Python Data Analysis Library — Pandas: Python Data Analysis Library." n.d. Accessed May 7, 2019. https://pandas.pydata.org/.

"Tableau for Students." n.d. Tableau Software. Accessed February 14, 2019. https://www.tableau.com/academic/students.

"Yelp Dataset." n.d. Accessed February 14, 2019a. https://kaggle.com/yelp-dataset/yelp-dataset.

"Yelp Dataset." ———. n.d. Accessed May 7, 2019b. https://www.yelp.com/dataset/challenge.

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

Alexander Kowarik, Matthias Templ(2016). Imputation with the R Package VIM. Journal of Statistical Software, 74(7),1-16. doi:10.18637/jss.v074.i07

Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr

"Seaborn: Statistical Data Visualization — Seaborn 0.9.0 Documentation." n.d. Accessed May 7, 2019. https://seaborn.pydata.org/.

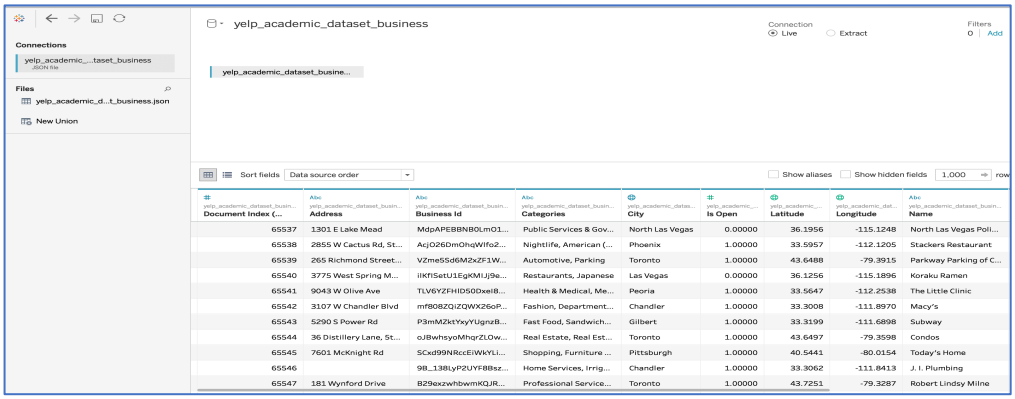
Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of StatisticalSoftware, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

Thibault Helleputte (2017). LiblineaR: Linear Predictive Models Based On The Liblinear C/C++ Library.R package version 2.10-8.

Appendix



Document Index (...	Address	Business Id	Categories	City	Is Open	Latitude	Longitude	Name
65537	1301 E Lake Mead	MdpAPEBBNB0LmO1...	Public Services & Gov...	North Las Vegas	0.00000	36.1956	-115.1248	North Las Vegas Poli...
65538	2855 W Cactus Rd, St...	AcjO26DmOhqWIfo2...	Nightlife, American (...	Phoenix	1.00000	33.5957	-112.1205	Stackers Restaurant
65539	265 Richmond Street...	VZmeS5d6M2xZF1W...	Automotive, Parking	Toronto	1.00000	43.6488	-79.3915	Parkway Parking of C...
65540	3775 West Spring M...	IlKfISetU1EgKMIJj9e...	Restaurants, Japanese	Las Vegas	0.00000	36.1256	-115.1896	Koraku Ramen
65541	9043 W Olive Ave	TLV6YZFHID50Dxel8...	Health & Medical, Me...	Peoria	1.00000	33.5647	-112.2538	The Little Clinic
65542	3107 W Chandler Blvd	mf808ZQiZQWX26oP...	Fashion, Department...	Chandler	1.00000	33.3008	-111.8970	Macy's
65543	5290 S Power Rd	P3mMZktYxyYUgnzB...	Fast Food, Sandwich...	Gilbert	1.00000	33.3199	-111.6898	Subway
65544	36 Distillery Lane, St...	oJBwhsyoMhqrZLOw...	Real Estate, Real Est...	Toronto	1.00000	43.6497	-79.3598	Condos
65545	7601 McKnight Rd	SCxd99NRccEiWkYLi...	Shopping, Furniture ...	Pittsburgh	1.00000	40.5441	-80.0154	Today's Home
65546		9B_138LyP2UYF8Bsz...	Home Services, Irrig...	Chandler	1.00000	33.3062	-111.8413	J. I. Plumbing
65547	181 Wynford Drive	B29exzwhbwmKQJR...	Professional Service...	Toronto	1.00000	43.7251	-79.3287	Robert Lindsy Milne

Figure 1. Data with a sample of 1000 rows showing Null values

```
In [2]: ##Reading in the data for Yelp Reviews
data_reviews=pd.read_csv("yelp_review.csv")
data_reviews.head()
```

Out[2]:

	review_id	user_id	business_id	stars	date	text	useful	funny	cool
0	vkVSCC7xljjrAI4UGfnKEQ	bv2nCi5Qv5vroFiqKGopiw	AEx2SYEUJmTxVVB18LlCwA	5	2016-05-28	Super simple place but amazing nonetheless. It...	0	0	0
1	n6QzIUObkYshz4dz2QRJTw	bv2nCi5Qv5vroFiqKGopiw	VR6GpWIda3SfvPC-lg9H3w	5	2016-05-28	Small unassuming place that changes their menu...	0	0	0
2	MV3CcKScW05u5LVfF6ok0g	bv2nCi5Qv5vroFiqKGopiw	CKC0-MOWMqoeWf6s-szl8g	5	2016-05-28	Lester's is located in a beautiful neighborhoo...	0	0	0
3	IXvOzsEMYtiJI0CARmj77Q	bv2nCi5Qv5vroFiqKGopiw	ACFtxLv8pGrrxMm6EgjreA	4	2016-05-28	Love coming here. Yes the place always needs t...	0	0	0
4	L_9BTb55X0GDtThi6GlZ6w	bv2nCi5Qv5vroFiqKGopiw	s2I_Ni76bjJNK9yG60iD-Q	4	2016-05-28	Had their chocolate almond croissant and it wa...	0	0	0

Figure 2. Review dataset

```
In [6]: ##Reading in the data for the Business from the combined file
data_business=pd.read_csv("combined_yelp_new.csv")
data_business.head()
```

Out[6]:

	Business_id	Name	Categories	stars	review_count	is_open	longitude	Latitude	city	state	...	WheelchairAccessible	Dogs/
0	QXAEGFB4oINsVuTFxEYKFQ	Emerald Chinese Restaurant	Specialty Food; Restaurants; Dim Sum; Imported...	2.5	128	1	-79.65229	43.60550	Mississauga	ON	...	NaN	
1	gnKjwL_1w79qoiV3IC_xQQ	Musashi Japanese Restaurant	Sushi Bars; Restaurants; Japanese	4.0	170	1	-80.85913	35.09256	Charlotte	NC	...	NaN	
2	1Dfx3zM-rW4n-31KeC8sJg	Taco Bell	Restaurants; Breakfast & Brunch; Mexican; Taco...	3.0	18	1	-112.02860	33.49519	Phoenix	AZ	...	NaN	
3	fweCYi8FmbJXHCqLnwuk8w	Marco's Pizza	Italian; Restaurants; Pizza; Chicken Wings	4.0	16	1	-81.35956	41.70852	Mentor-on-the-Lake	OH	...	NaN	
4	PZ-LZzSlhSe9utkQYU8pFg	Carluccio's Tivoli Gardens	Restaurants; Italian	4.0	40	0	-115.12850	36.10002	Las Vegas	NV	...	NaN	

5 rows × 56 columns

Figure 3. Business, attributes, Checkin dataset combined

Business
Business_id
name
address
city
state
postal_code
longitude
latitude
stars
review_count
is_open
categories

Attributes
GoodForKids
RestaurantsReservations
GoodForMeal
BusinessParking
Caters
NoiseLevel
RestaurantsTableService
RestaurantsTakeOut
RestaurantsPriceRange2
OutdoorSeating
BikeParking
Ambience
HasTV
WiFi
Alcohol
RestaurantsAttire
RestaurantsGoodForGroups
RestaurantsDelivery
BusinessAcceptsCreditCards
BusinessAcceptsBitcoin
ByAppointmentOnly
AcceptsInsurance
Music
GoodForDancing
CoatCheck
HappyHour
BestNights
WheelchairAccessible
DogsAllowed
BYOBCorkage
DriveThru
Smoking
AgesAllowed
HairSpecializesIn
Corkage
BYOB
DietaryRestrictions
Open24Hours
RestaurantsCounterService

Checkin
business_id
weekday
hour
checkins

Reviews
review_id
user_id
business_id
stars
date
text
useful
funny
cool

```
Confusion Matrix and Statistics

          Reference
Prediction  Bad Good
      Bad    14   14
      Good 1571 6731
```

Figure 4. Random forest confusion matrix with imputation

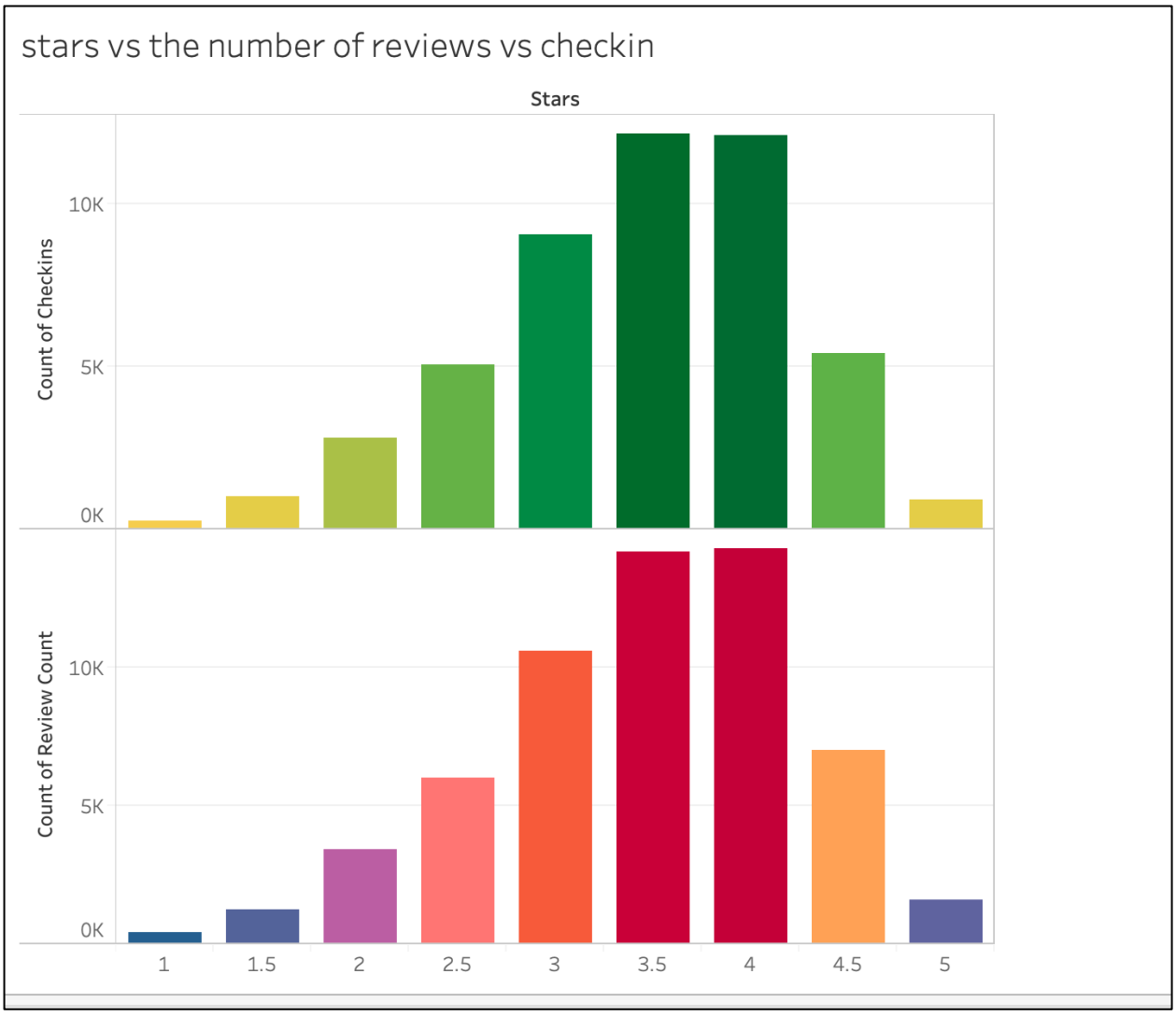


Figure 5. Stars vs Number of Reviews Vs Checkin using Tableau

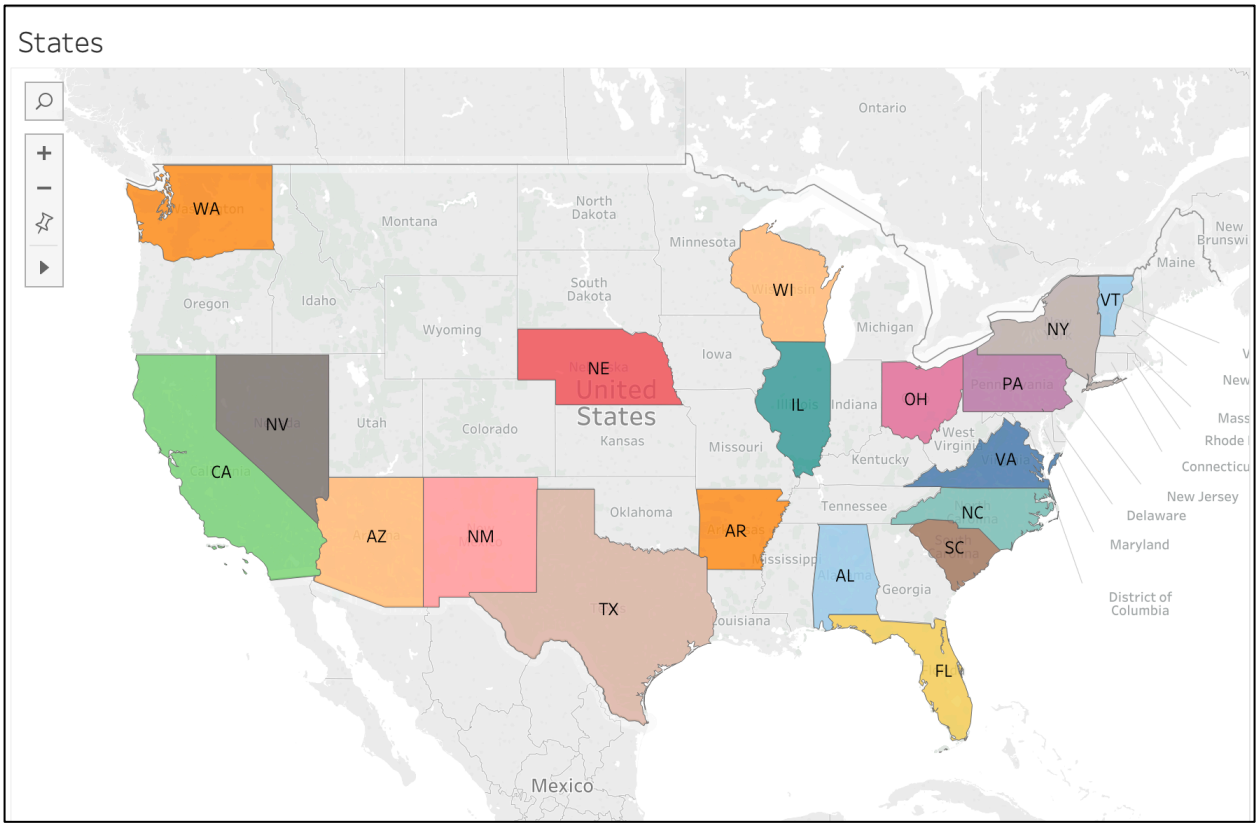


Figure 6. States for which data is present in US with restaurant's using Tableau

Figure 7. Attributes affecting star rating using Tableau