PGPBDML – Chennai Capstone Project – Group 8

Vignesh Athavan , Shobin Joyakin, Sonia Mawandia,
Ganapathi Nagarajan , Nishita Ravindra

Friday, July 06, 2018

GREAT LAKES INSTITUTE OF MANAGEMENT
Global Mindset - Indian Roots

INTERIM REPORT

PROBLEM STATEMENT

Apply data science tools and techniques on the Yelp Dataset that is available

a) For restaurant/ business owners to improve their services and business and

b) For users to choose a best restaurant from the available choices.

Solution:

a) Help restaurants target potential customers who are not yet their customers but likely to enjoy the service

b) Recommend restaurants to customers based on their eating preferences and other information such as previous ratings and feedback for restaurants.

DATASET

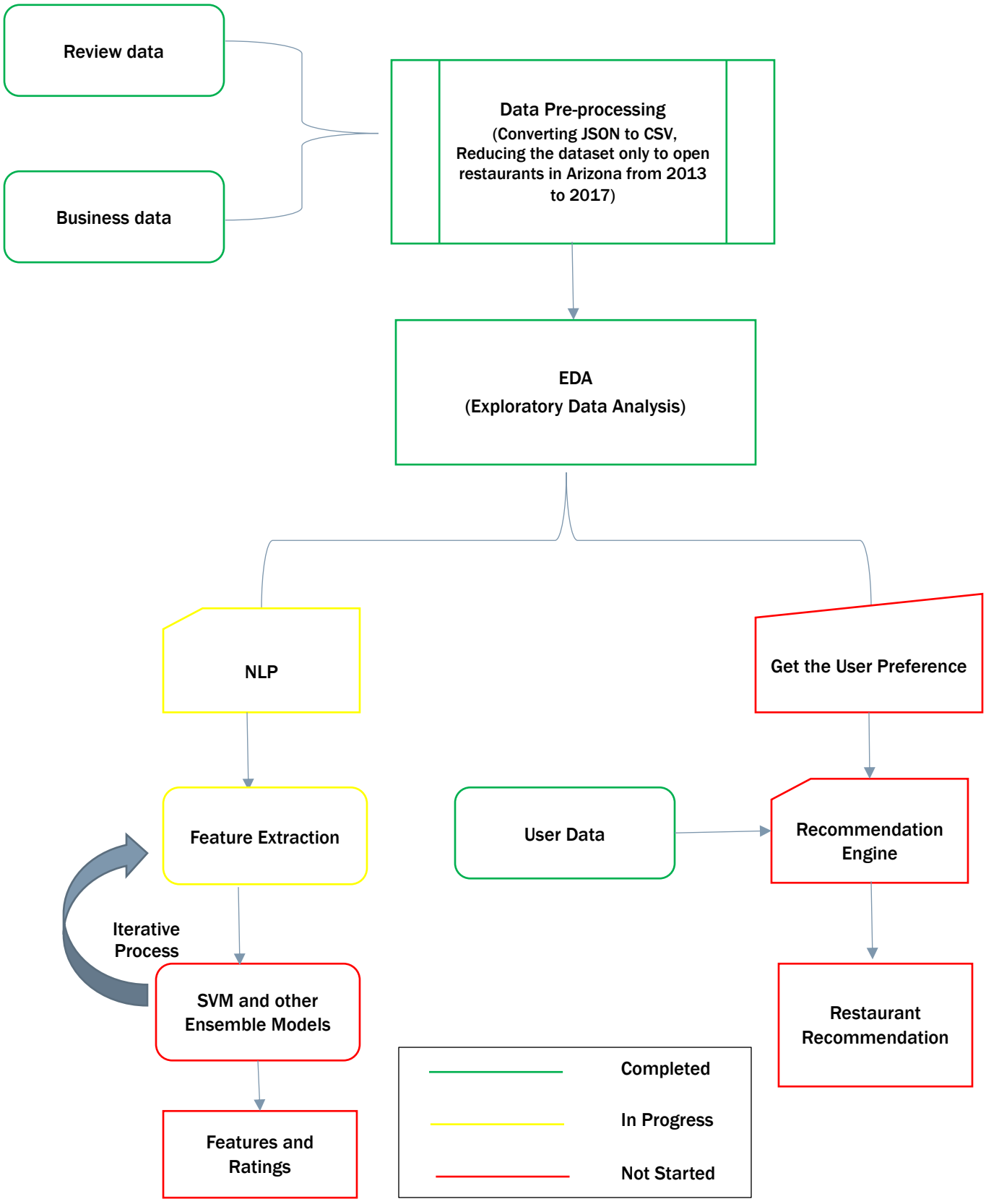
The dataset provided for the Capstone Project is part of the Yelp Dataset Challenge and the specific dataset used in this capstone corresponds to Round 11 of their challenge and can be accessed from the following link for download: http://www.yelp.com/dataset_challenge

The dataset is stored in 5 files of JSON format, where each file is composed of a single object type (a one-json-object per line). The respective data files provide information about:

1. Businesses and their attributes (business.json)
2. Check-in times of customers at given businesses store (checkin.json)
3. Reviews submitted by customers about the businesses (reviews.json)
4. Tips on the businesses (tips.json)
5. Users of the businesses (users.json)

For our Project purpose we are using business, reviews and users json.

PROPOSED SOLUTION



DATA PRE-PREPARATION / EXPLORATORY DATA ANALYSIS AND SUMMARY OF INITIAL FINDINGS

The Current JSON data available from the YELP website is very huge up to 7 GB in the form of 5 different JSON files. We will be creating a recommendation model using a sample of all open restaurants in the State of Arizona in USA that where reviewed between 2013 and 2017. By this we can restrict the data and create much lesser sparse dataset for our Analysis. This is achieved by filtering the JSON data in MongoDB. Since the User dataset is huge, we are analyzing only those users who reviewed the open restaurants in Arizona between 2013 and 2017. The created model can be scaled for other restaurants in USA which use similar features upon reengineering of model and features.

For our EDA, we loaded all the JSON files into Python and transformed the JSON objects to Pandas Dataframe by flattening the data. During our EDA using python, we found that almost all columns contained NaN (missing) values and it was required to handle these values. For our project we first cleaned the columns having attributed specific column values along with NaN values and columns having just True, False and NaN as their values.

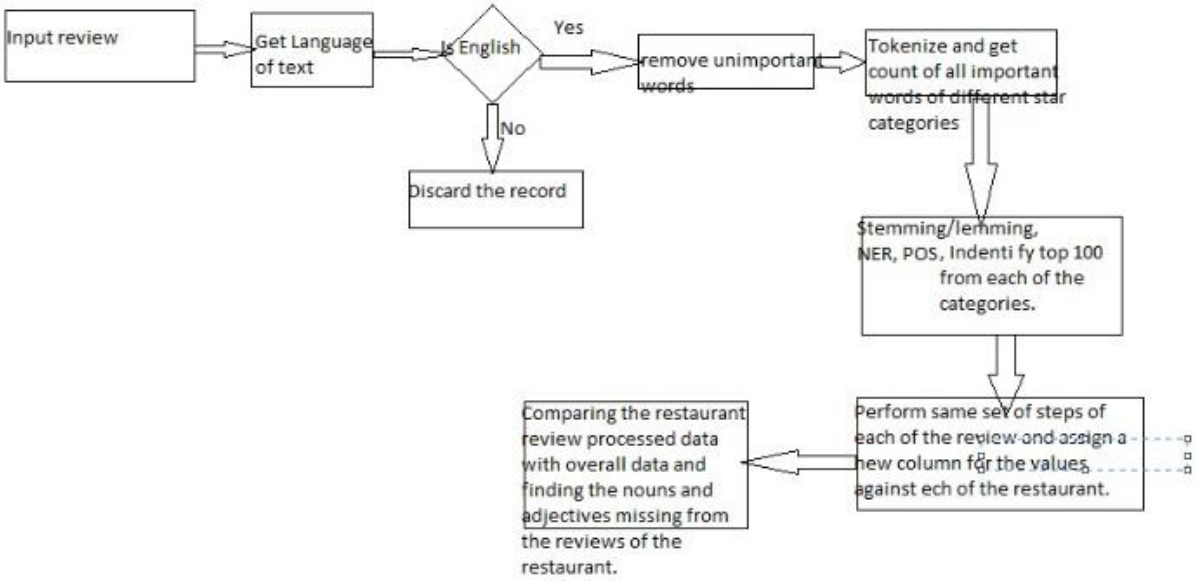
We handled the former list of columns followed by the later. In the later list of columns, we converted all False and NaN values to 0 and True values to 1 based on business reasoning.

The handling of the former list of columns is tabled below for easier understanding.

Column Name	Column Values Before Handling NaN	Column Values After Handling NaN	Final Values	Description
AgesAllowed	[nan 'allages' '21plus']	['allages' '21plus']	[0 1]	21plus : 1, allages : 0
Alcohol	['none' 'full_bar' nan 'beer_and_wine']	['none' 'full_bar' 'beer_and_wine']	[0 2 1]	none: 0, beer and wine: 1, Full bar: 2
Music_no_music	[nan False]	[nan False]	[0 1]	False: 1, True: 0
NoiseLevel	['loud' 'average' nan 'quiet' 'very_loud']	['loud' 'average' 'quiet' 'very_loud']	[2 1 0 3]	quiet: 0, average: 1, loud: 2, very_loud:3
RestaurantsAttire	['casual' nan 'dressy' 'formal']	['casual' 'dressy' 'formal']	[0 1 2]	casual:0, dressy:1, formal: 2
RestaurantsPriceRange2	[1. 2. nan 3. 4.]	[1. 2. 0 3. 4.]	[1 2 0 3 4]	0, 1, 2, 3, 4
Smoking	[nan 'no' 'outdoor' 'yes']	['no' 'outdoor' 'yes']	[0 1 2]	no:0, outdoor:1, yes:2
WiFi	['free' 'paid' nan 'no']	['free' 'paid' 'no']	[1 2 0]	no:0, free:1, paid:2

We have converted all columns other than business_id, name, address, city, state, postal_code, latitude, longitude, stars and review_count to their numerical values as it enables us to perform 'AND' operation during our recommendation.

NLP Process to be followed:



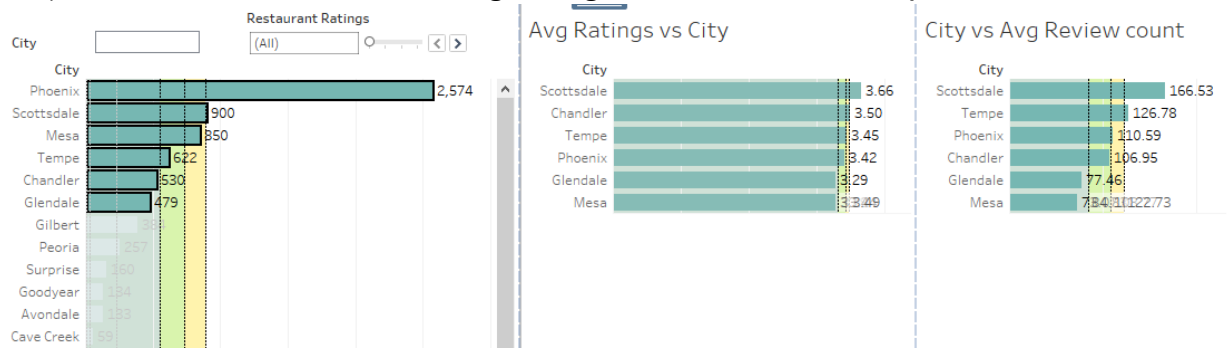
EXPLORATORY DATA VISUALAIZATION

1) The Restaurant spread across cities in Arizona

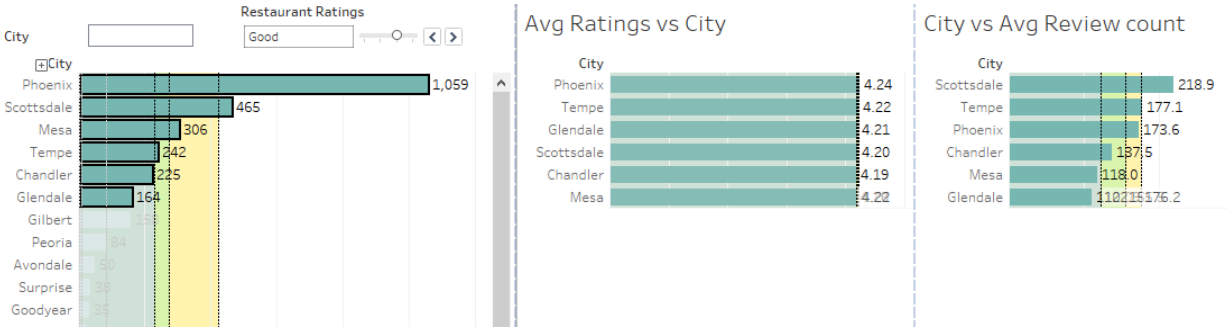
Here the maximum number of restaurant is from following cities.

- Phoenix
- Scottsdale
- Mesa
- Tempe
- Chandler
- Glendale

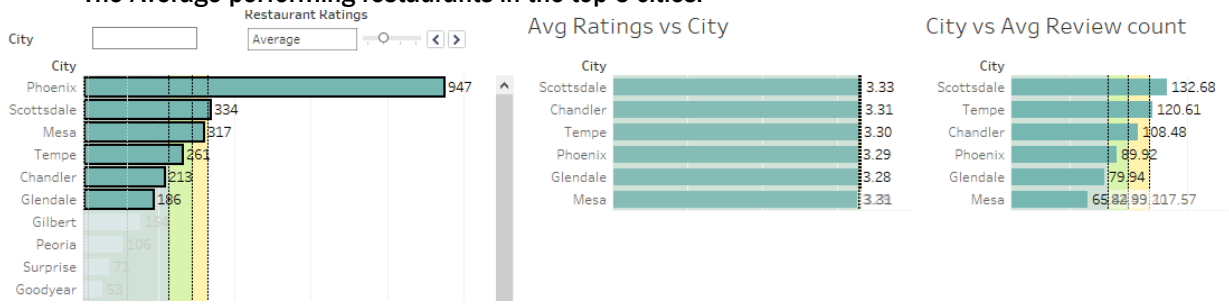
2) The Overall restaurant count, Average Rating and review count for the top 6 cities are below



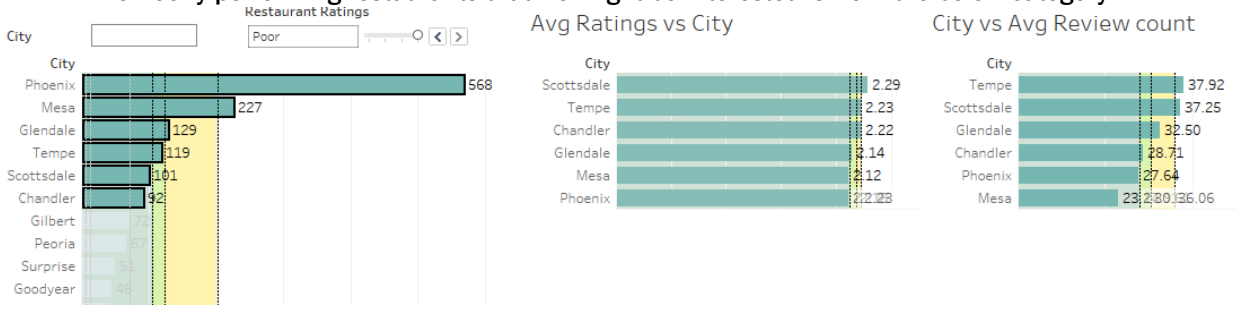
Below finding shows how Good restaurants perform on the top 6 cities



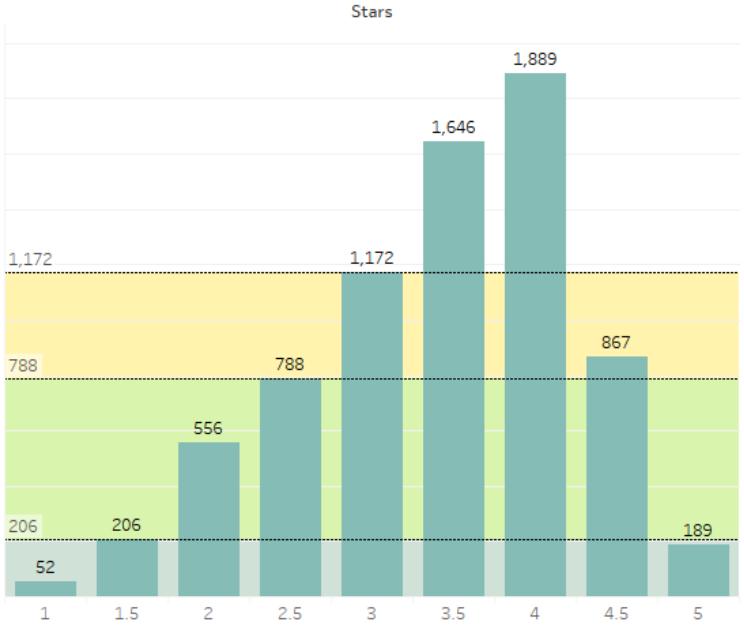
The Average performing restaurants in the top 6 cities:



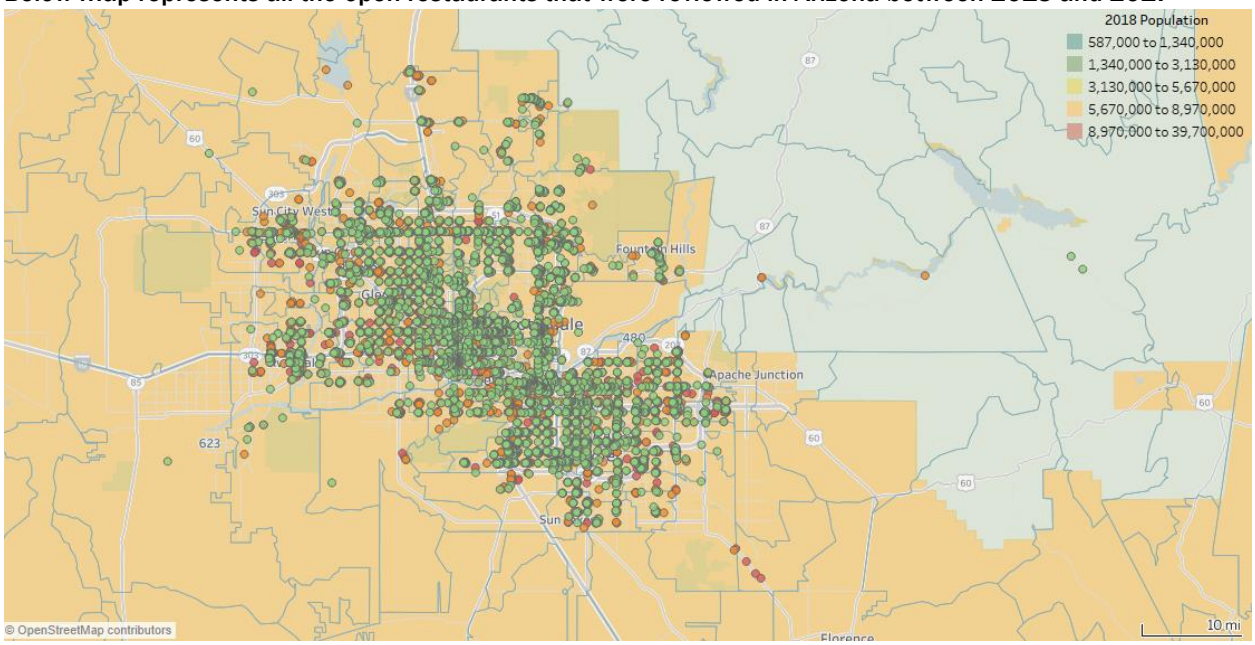
The Poorly performing restaurants that we might be interested is from the below category



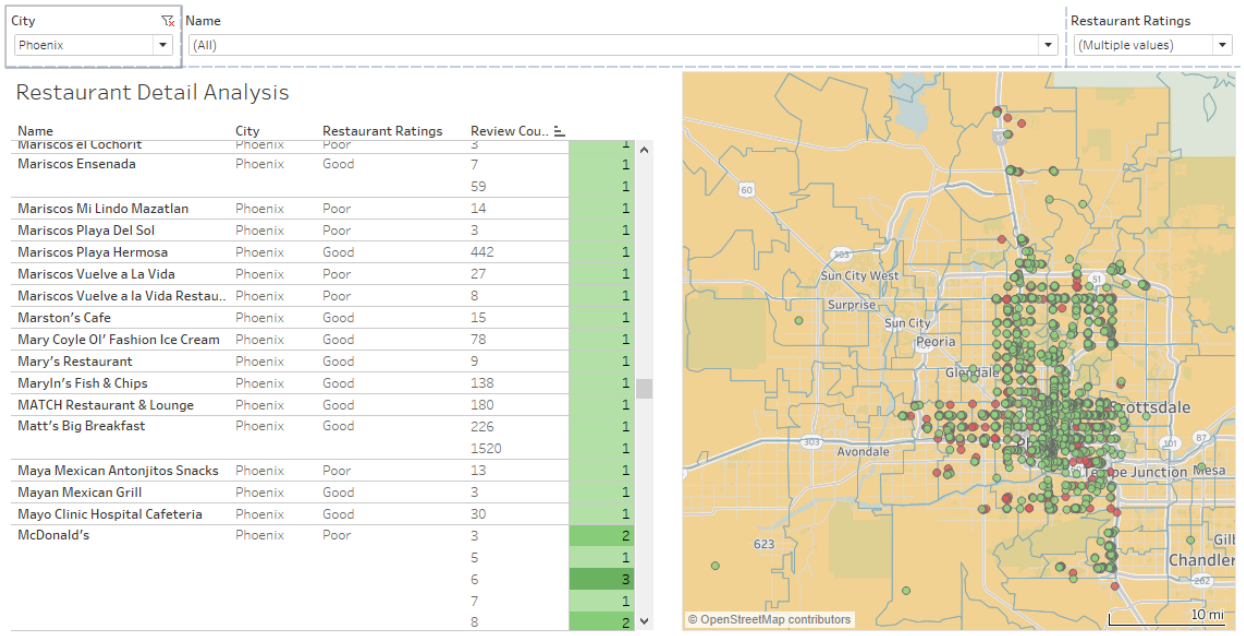
3) Analysis on the Restaurant Ratings
The data is pretty much normal with long tail in the left.



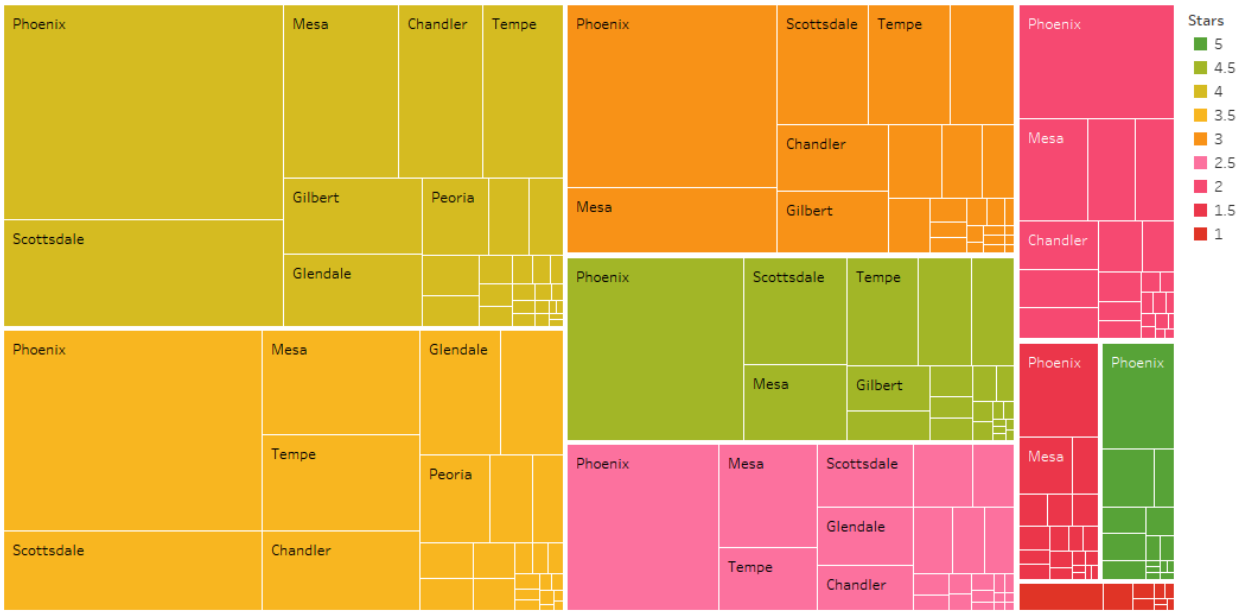
4) Below map represents all the open restaurants that were reviewed in Arizona between 2013 and 2017



5) Below map represents the list of Restaurants with Good and poor performance in the city of Phoenix along with review counts and the number of restaurants that were reviewed



6) A Treemap view of Restaurants and ratings are shown below by the count of restaurants



CHALLENGES

Data Size

5,200,000 reviews , 174,000 businesses possess huge challenge in terms of storage and processing. So we have decided to handle only Arizona state data for our analysis which has X reviews and Y business. Also we decided to take reviews from year 2013 to 2018 to further reduce the data size

Data Structure

JSON files even though was very useful in terms of viewing the data became difficult to handle in this data science project since **it has lot of Nested structure**. So we flattened the json and converted it into CSV for ease of handling.

EVALUATION METRICS

In order to evaluate our methods and models used, we need to agree on a set of success measures. For our project, we should identify common set of features on which the restaurant will be evaluated The ratings for these restaurants will be a function of these features, a high rating on these features should result on high rating on the restaurant and vice versa. We are planning to create a text regression model, utilizing bag of words and reviewers' RFM dimensions to predict usefulness of reviews and percentage error method which we have explained below

x: avg of ratings of reviews from dataset

y: avg of ratings on sentiment of reviews

Percentage Error: $(|y-x|/x) * 100$

Error Greatly impacts our analysis and recommendations

NEXT STEPS

1. Extract Features through NLP
2. Sentiment Analysis for different ratings Grouped as Poor (<3 rating) and Good (>4 Rating)
3. Build SVM model to predict ratings with Reviews and Features
4. Matrix Factorization Recommendation Engine
5. Build UI for user input (Category, Attributes, Facilities)
6. Pass Parameters from UI to Recommendation Engine, get the ordered list from the Algorithm and display the results