DIC Project Report

Phase 1

Restaurant Prediction using Yelp Dataset

(Catering United States)

Prashant Godhwani

Akhilesh Goriparthi

I. Problem Statement

a. Form a title and problem statement that clearly state the problem and questions you are trying to answer

With more and more places to eat being opened every day, people want to make informed decisions and have a good experience when they do choose a restaurant. Yelp, an open crowd-source platform, contains information about restaurants including but not limited to reviews, categories, locations and other relevant features that can be leveraged to recommend users places to eat based on their preferences. The challenge is how to use this big data to make reasonable inferences and model it to build a recommendation system that can save people the hassle and energy while ensuring a better experience.

What are trying to answer?

1) How can we effectively find patterns in the data spread across multiple sources to extract valuable information for a restaurant?
2) Is it possible to make accurate recommendations using the data available to us?
3) How can we deal with the inherent bias introduced in the system during prior data collection stages?
4) How often should the model be retrained to incorporate new restaurants or food chains and updated reviews?
5) What input we would need from the user to suggest them restaurants.
6) If the model made an inaccurate prediction, could the user provide feedback to the system? If so, how will the model take those into consideration and recommend better?

b. Discuss the background of the problem leading to your objectives. Why is it a significant problem?

In current times, we see huge number of dining options around us. With the advent of internet, people are more aware than before. This leads to them scouting for the perfect place they can dine in, based on their preferences. The current options in the market are

not crowdsourced and their recommendations can be influenced by partnerships with certain restaurants. This makes it difficult for customers to find trusted restaurants for their needs. On the other hand, the restaurant owners of small, independent businesses also face issues in gaining visibility.
This recommendation system utilizes the power of crowdsourced reviews to suggest users based on their preferences some of the best places they can dine at. Moreover, it also helps to level the field for these smaller, independent restaurants.

c. Explain the potential of your project to contribute to your problem domain. Discuss why this contribution is crucial?

The contribution is crucial as the food sector is one of the biggest economic drivers of modern times with millions of restaurants opening in the world every day. In addition to helping users save time by not scrolling through plethora of places to eat, money and ensure a better dining experience, this recommendation system can also help the restaurants lead customers towards their quality restaurants. This platform also has use case to recommend users' who are travelling, places to dine based on their dietary, monetary, location and cuisine preferences.

II. Data Sources

a. Introduction

We have taken Yelp's dataset from Yelp's official website and Kaggle. The data contains 5 files with json extension.

i. You must cite and link your data sources in the report.

The data has been taken from Yelp's official website and Kaggle-
https://www.yelp.com/dataset
https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset

b. About Data

Business Data File -

Column Name	Sample Value	Comment
business_id	"tnhfDv5Il8EaGSXZGiuQGg"	22-character unique string business id
name	"Garaje"	The business's name

address	"475 3rd St"	The full address of the business
city	"San Francisco"	The city
state	"CA"	2-character state code, if applicable
postal code	"94107"	The postal code
latitude	37.7817529521	Latitude
longitude	-122.39612197	Longitude
stars	4.5	Star rating, rounded to half-stars
review_count	1198	Number of reviews
is_open	1	0 or 1 for closed or open permanently, respectively

categories	["Mexican", "Burgers", "Gastropubs"]	An array of strings of business categories

Review Data File

Column Name	Sample Value	Comment
review_id	"zdSx_SD6obEhz9VrW9uAWA"	22-character unique review id
user_id	"Ha3iJu77CxlrFm-vQRs_8g"	22 character unique user id, maps to the user in user.json
business_id	"tnhfDv5Il8EaGSXZGiuQGg"	22 character business id, maps to business in business.json
stars	4	Star rating
date	"2016-03-09"	Date formatted YYYY-MM-DD

text	"Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks."	The review itself
useful	0	Number of useful votes received
funny	0	Number of funny votes received
cool	0	Number of cool votes received

Users Data File

column_name	sample value	comment
user_id	"Ha3iJu77CxlrFm-vQRs_8g"	string, 22 character unique user id, maps to the user in user.json
name	"Sebastien"	string, the user's first name

review_count	56	integer, the number of reviews they've written
yelping_since	"2011-01-01"	string, when the user joined Yelp, formatted like YYYY-MM-DD
friends	["wqoXYLWmpkEH0YvTmHBsJQ", "KUXLLiJGrjtSsapmxmpvTA","6e9rJKQC3n0RSKyHLViL-Q"]	array of strings, an array of the user's friend as user_ids
useful	21	integer, number of useful votes sent by the user
funny	88	integer, number of funny votes sent by the user
cool	15	integer, number of cool votes sent by the user
fans	1032	integer, number of fans the user has

elite	[2012,2013]	array of integers, the years the user was elite
average_stars	4.31	float, average rating of all reviews
compliment_hot	339	integer, number of hot compliments received by the user
compliment_more	668	integer, number of more compliments received by the user
compliment_profile	42	integer, number of profile compliments received by the user
compliment_cute	62	integer, number of cute compliments received by the user
compliment_list	37	integer, number of list compliments received by the user

compliment_note	356	integer, number of note compliments received by the user
compliment_plain	68	integer, number of plain compliments received by the user
compliment_cool	91	integer, number of cool compliments received by the user
compliment_funny	99	integer, number of funny compliments received by the user
compliment_writer	95	integer, number of writer compliments received by the user
compliment_photos	50	integer, number of photo compliments received by the use

Checkin Data File

column_name	sample value	comment
business_id	"tnhfDv5Il8EaGSXZGiuQGg"	string, 22 character business id, maps to business in business.json
date	"2016-04-26 19:49:16, 2016-08-30 18:36:57, 2016-10-15 02:45:18, 2016-11-18 01:54:50, 2017-04-20 18:39:06, 2017-05-03 17:58:02"	string which is a comma-separated list of timestamps for each checkin, each with format YYYY-MM-DD HH:MM:SS

III. Data Cleaning/Processing

Since the complete dataset for this project is split into 4 different source json files. We performed some basic Data cleaning steps on all of them and then some domain specific steps on each to improve the data quality for modeling further –

1) **Handling Missing Values:** We started by identifying if there was any missing/null data in the dataset and performed data cleaning by removing the entries with missing, null, empty data.

   Why we chose to delete the data instead of Imputation?
   Since compared to the total population, the data that was partially available was less, and didn't cause loss of valuable information. We decided to go ahead with deleting it.

   For example – In the business dataset – we could identify missing values for neighborhood, city, state, postal code, latitude and longitude.

   Ref – Data Cleaning > Business > 1

2) **Handling Data with incorrect state:** Since our business use case is only based on building a recommendation system for the United States, we decided to get rid of the data that didn't lie in the population dedicated for the problem.

   We solved this problem by removing all the data in the data frame that didn't belong to the 50 US states. This narrowed down the data to the population we intend to cater.

This data cleaning step is only valid for the Business Dataset.

Ref – Data Cleaning > Business > 2

3) **Handling Incorrect Latitude and Longitude values:** Since the population we are catering resides in the US, and locations play a big role in recommending places to eat, we decided to find out that even after performing Data Cleaning (2), is there any data that has incorrect location co-ordinates?

   After finding out traces of incorrect data, we removed the data with incorrect latitude and longitude values.

   This data cleaning step is only valid for the Business Dataset.

   Ref – Data Cleaning > Business > 3

4) **Removing Businesses that are closed:** Since the recommendation system should recommend dining options where people can go and dine-in. We removed all the businesses that were closed.

   We achieved this simply by filtering out the rows where is_open was marked as 0.

   This data cleaning step is only valid for the Business Dataset.

   Ref – Data Cleaning > Business > 4

5) **Filtering out Restaurants:** The Yelp dataset contains crowd sourced data of all the businesses. We however, intend to build a restaurant recommendation system. Hence, we filtered out our population by removing all the businesses that were not serving food.

   This ensures that the population being considered by our recommendation model contains the correct businesses to recommend to its users.

   We achieved this by filtering out on the categories assigned to each business.

   This data cleaning step is only valid for the Business Dataset.

Ref – Data Cleaning > Business > 5

6) **Sampling and Keeping Reviews only for restaurants in the US:** Since the review dataset was large and could not be loaded into the main memory of multiple options we tried, we decided to sample the data.

   We picked out 14,000 data points from 2,448,503 data points by random sampling. We further obtained these points by keeping only the reviews that relate to businesses we cater to from Data Cleaning(5).

   For Phase 2 of the project and as the class progresses, we intend to use some of the tools taught in class to handle such huge amount of data without introducing bias.

   This data cleaning step is only valid for the Business Dataset.

   Ref – Data Cleaning > Business > 6

7) **Parsing fields to their correct Format –** Across all the datasets, we have formatted the fields to their correct format to ensure that any operations involving them go without hiccups.

   For example – we change the date in the reviews from object to datetime. Further during our EDA, this helps us track sentiment across timeline for a restaurant.

   Ref – Data Cleaning > Reviews > 7

8) **Identifying and Removing reviews with encoded HTML –** To ensure correct sentiment analysis, we remove reviews with encoded HTML.

   This helps us correctly extract menaningful insights from the data, as HTML is used to structure and format text for web pages, and isn't as useful to actual intent and content of the text.

   Ref – Data Cleaning > Reviews > 8

9) **Removing Stop words, punctuation for Sentiment Analysis –** To reduce noise in the text and improve computational efficiency, next we remove all the stop words like a, an, and, as, at, be, by from the actual review. This helps us capture the intent of the review efficiently and reduce resources to compute the same.

   We achieved this using NLTK library to download and use the list of stop words and remove them before proceeding with the next stage.

   Ref – Data Cleaning > Reviews > 9

10) **Checking Duplicate users –** To ensure that we maintain the sanity of the data, we remove duplicate users from the users dataset. This helps us reduce the bias that may occur due to these duplicate users having same preferences and taste.

    This also improves the data quality and save computation resources.

    Ref – Data Cleaning > Reviews > 10

There are some other Data cleaning steps that were performed, but were not as significant enough to be highlighted. They are mentioned below –

11) Removing columns from each dataset that are not useful.
12) Merging data frames to reduce clutter. E.g. - We merged the business hours datasets with our filtered restaurant list to filter out businesses that are not participating or have been filtered out before. (11)
13) Renaming labels in datasets to more meaningful ones. (12)
14) Making categories consistent by making them lowercase

IV. Exploratory Data Analysis (EDA)

We tried to do EDA to discover patterns that can help our recommendation system make useful recommendations whilst also helping business owners make smart decisions based on the statistics.

1) Peak hours for each restaurant across all business days

- **Objective -** The main objective of this EDA is to identify the peak hours of each restaurant for each day.

- **How -** For this we took the checkins dataset and took the business_id, hours, weekday grouped them and summed the number of checkins.

```
result
```

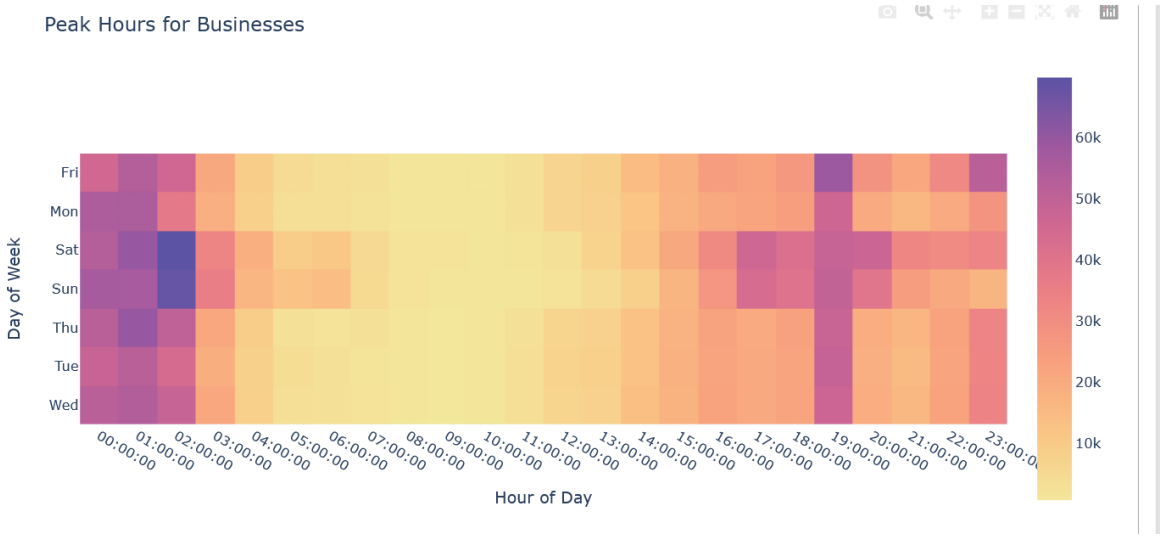
[60]:

	business_id	weekday	hour	sum
0	--6MefnULPED_I942VcFNA	Fri	23:00:00	5
1	--6MefnULPED_I942VcFNA	Mon	00:00:00	6
2	--6MefnULPED_I942VcFNA	Sat	23:00:00	12
3	--6MefnULPED_I942VcFNA	Sun	01:00:00	6
4	--6MefnULPED_I942VcFNA	Thu	23:00:00	2
...	...	...	...	...
733571	zzzaIBwimxVej4tY6qFOUQ	Sat	01:00:00	3
733572	zzzaIBwimxVej4tY6qFOUQ	Sun	19:00:00	5
733573	zzzaIBwimxVej4tY6qFOUQ	Thu	01:00:00	2
733574	zzzaIBwimxVej4tY6qFOUQ	Tue	22:00:00	3
733575	zzzaIBwimxVej4tY6qFOUQ	Wed	21:00:00	2

ick to scroll output; double click to hide

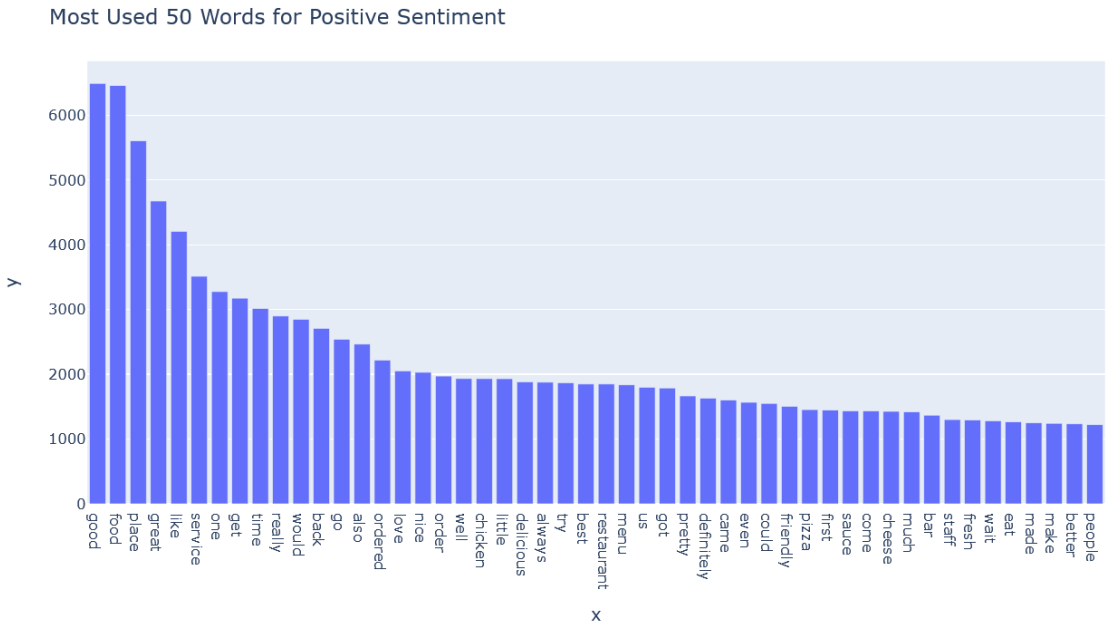
733576 rows × 4 columns

- The above is the result and its show for each business and week day which is the peak hours.
- **Inference -** By plotting the average peak hours of all businesses and we see that most businesses have receive customers on weekends between 7:00 p.m. to 2:00 a.m. The result can be shown in the below plot.

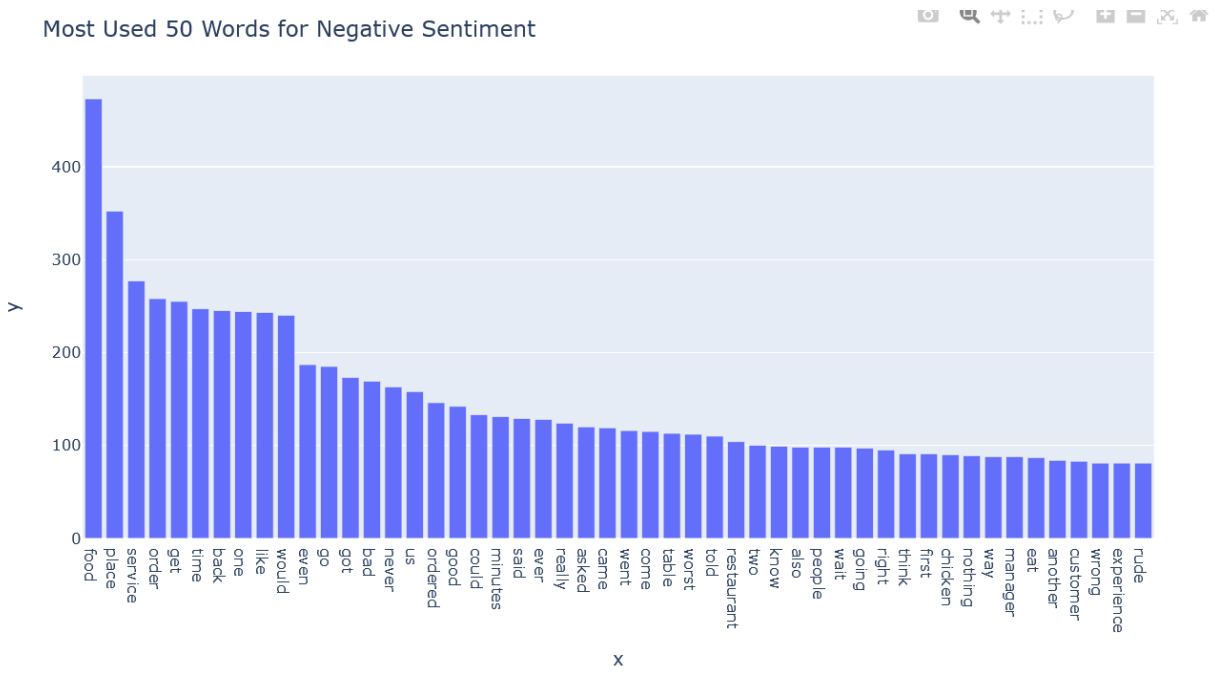


2) Sentiment Analysis of Reviews

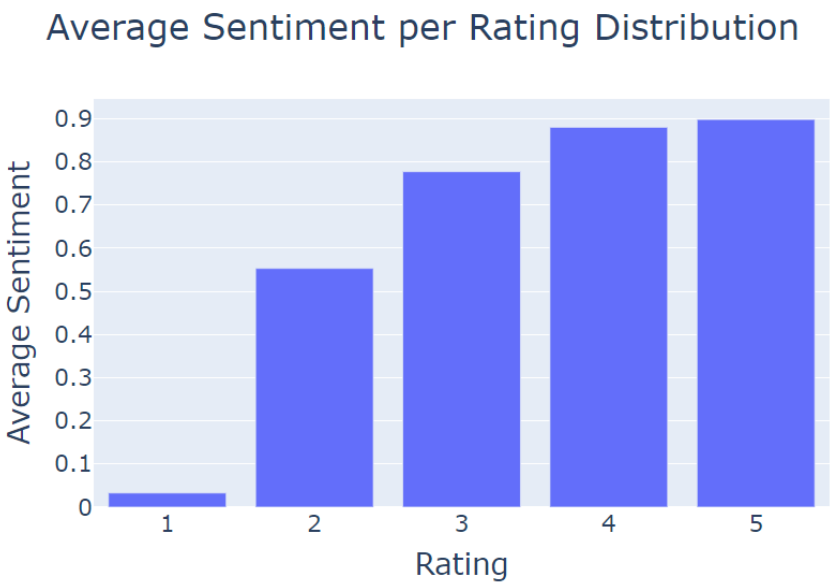
- Using Vader we analyzed the text and given each review a score and if the score is above 0 it is considered as positive, below zero is negative otherwise neutral.
- After defining the class for each review we calculated the frequency of wors in each class.
- And plotted them in the bar graph which can be seen below.



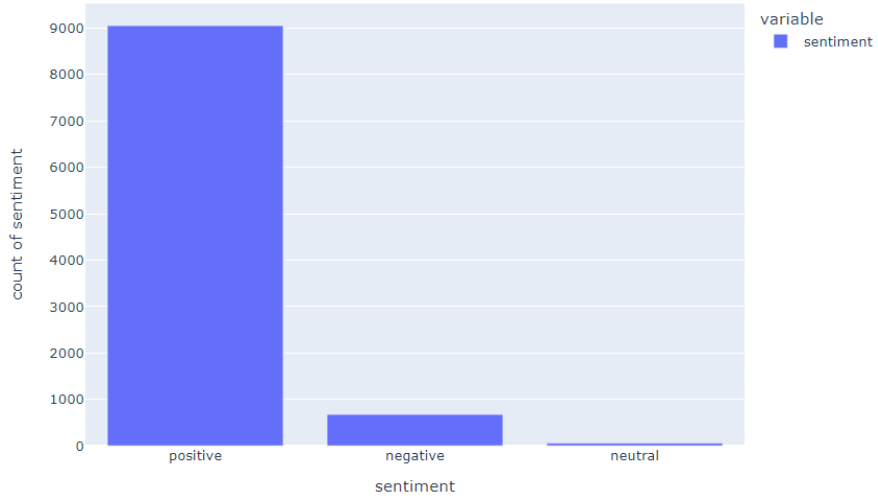
- Here the x axis is the top 50 words and y- axis the count of their word in all the reviews of positive class.
- if a review has good, better, delicious, great then it most likely to be a positive review.



- Here the x axis is the top 50 words and y- axis is the count of their occurrences in all the reviews of negative class.
- if a review has bad, worst, never, rude or another then it most likely to be a negative review.
- Based on the Average Sentiment per rating distribution (below), we found out that reviews with lesser ratings were negative, and ones with higher ratings were positive. There was not much difference between reviews rated 4 and 5 stars signifying that users were similarly satisfied with the service. Here x-axis represents the rating and y-axis represents that sentiment_score.

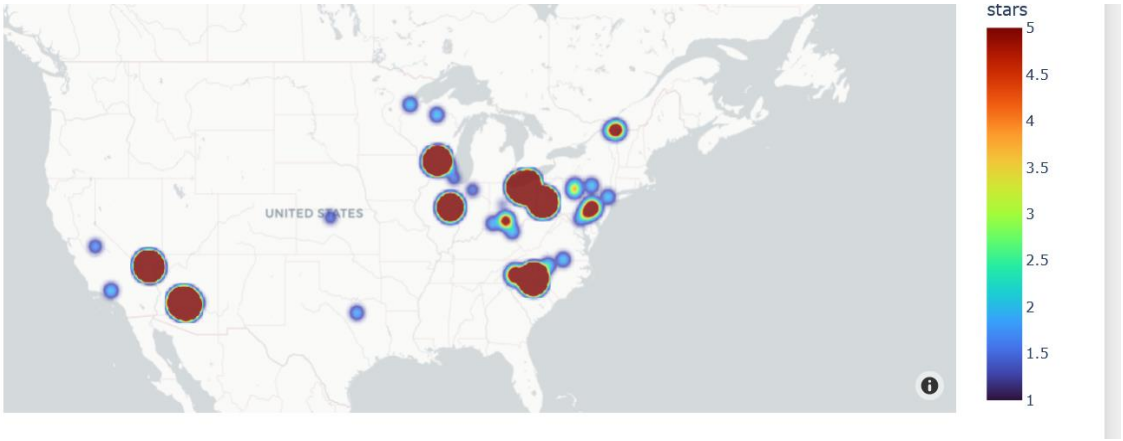


- Observing below, we can infer that people tend to leave reviews when they are either satisfied or dis-satisfied with the service. Diners with neutral sentiment don't tend to leave a lot of reviews or they tend to form a strong positive/negative opinion of their experience and hence we can see such a low amount of reviews with neutral sentiment. Here x-axis has the sentiment and y-axis represents the number of reviews for each sentiment.



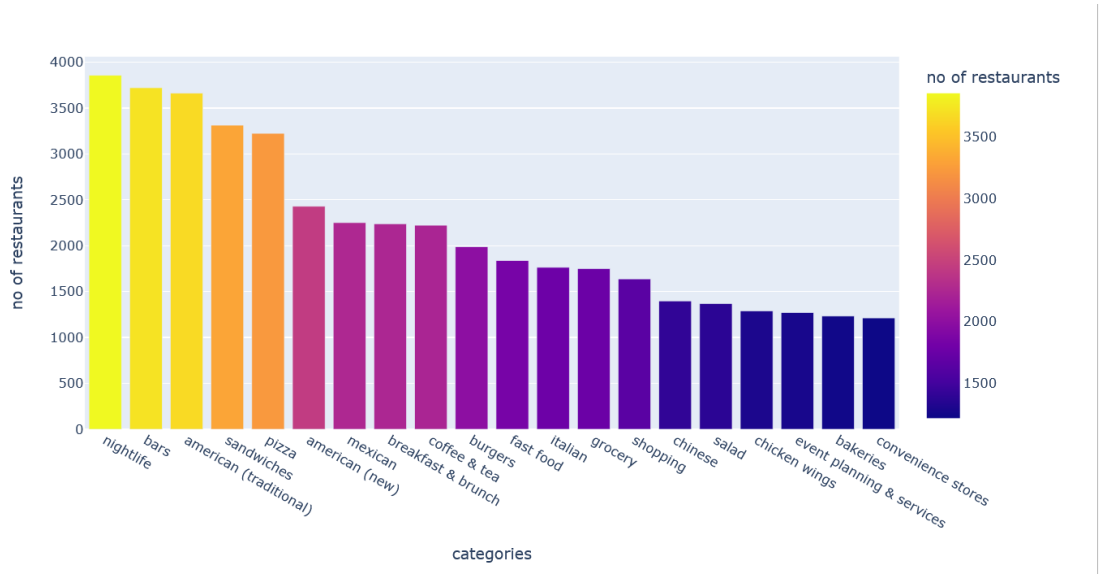
3) Identifying areas with highly rated Restaurants

- **How -** We took the business_final data-frame and grouped it with latitude and longitude and took the average of the stars and plotted them geographically.
- **Inference -** Looking at the heatmap we can figure out the areas of highly rated restaurants.
- Further observing, we notice that places with larger cluster of restaurants have higher average rating, compared to regions with lesser number of places to dine. This maybe because of the better dining experience offered by them owing to the competition that they face from the neighboring restaurants.
  This is not the case with the latter as the customers there might not have a lot of options which in turn adversely affects the quality of service offered.



4) No of restaurants per food category

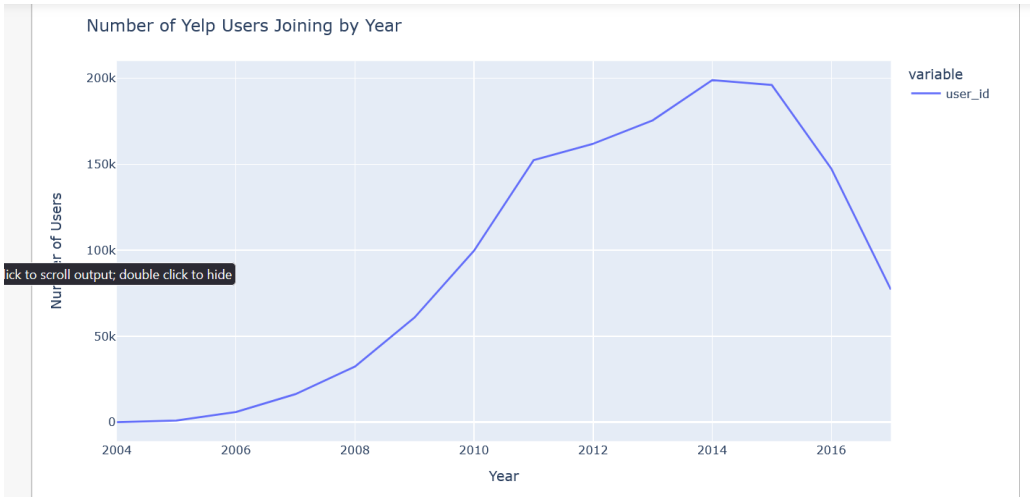
- **How -** taking only the columns categories and business_id created a data frame called cat.
- Replaced the categories restaurant with empty because we doing EDA on the restaurant category so we need to find the categories within restaurants.
- Using str.split function split the categories which are combined with ; and using value counts method in pandas counted the number of restaurant with particular categories and took the top 20 categories and plotted them on a bar graph. Below is the result.



- In the above graph the x-axis is the categories and the y-axis is the number of restaurants.
- We performed this to know the distribution of restaurants for a specific category.
- **Inference –** It can be observed there are a lot of options to cater people looking for nightlife and bars owing to the demand. Some of other items that have a higher demand are pizzas and sandwiches due to which there seems to be a large number of restaurants selling these.

5) Plot of growth of Yelp over the years

- **How -** Using yelp_user dataset finds the growth of yelp over the years.
- Took the "yelping_since " column in yelp_user dataset and changed it into datetime and counted the number of user for each year and plotted a line showing the trend of it.
- We did this EDA to check whether the yelp is still growing or not for us to use for the future recommendation. Below is the line graph .

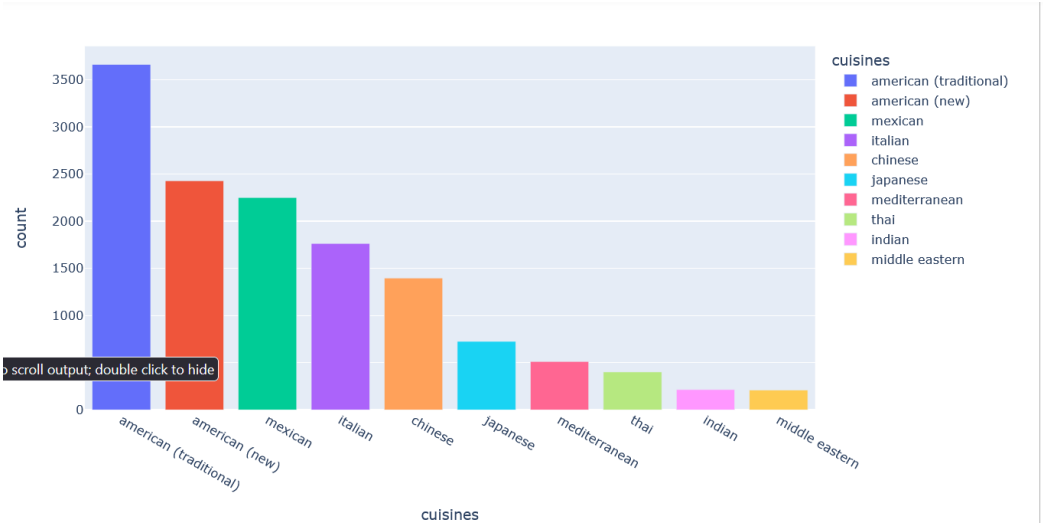


- The x-axis is the years and the y- axis is the count of user.

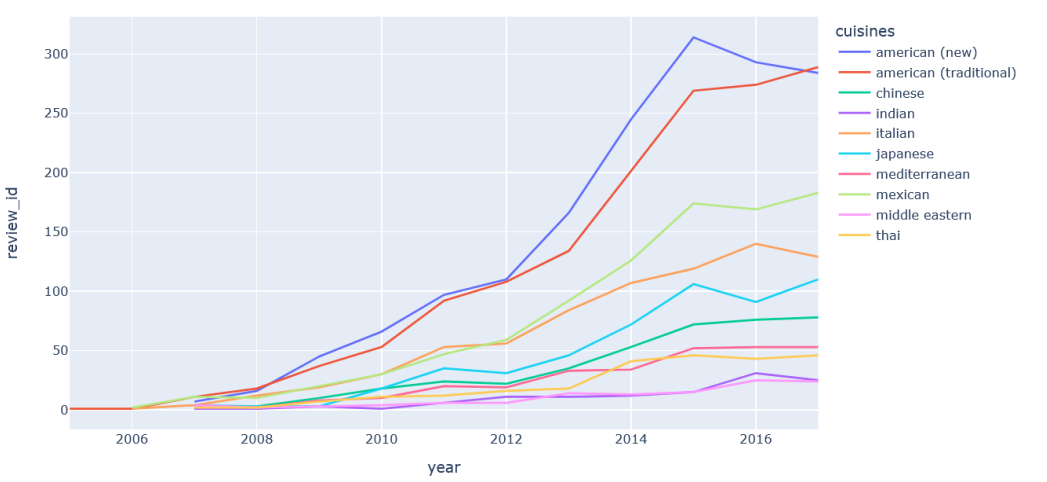
- **Inference –** There has been a steep and constant rise in the number of Yelp users joining every year till 2014. The number of new users joining the platform declining after 2014 can be owed to the assumption that the platform onboarded most of its users by that time.

6) Demand and Supply of Different Cuisines across time

- This EDA is done because to check the demand for the different cuisines and what users are recommending more by counting the number of reviews for each cuisine.
- First we created a list which has the names of popular cuisines and using for loop for business_final dataframe we counted the number of restaurants for each cuisine and plotted a bar graph.



- For the above graph the x-axis represents the cuisines and y-axis represents the count(no. of restaurants). As we can see the cuisine American (new) has the most number of restaurants.
- Next to check the trend of each cuisine we combined the cuisine with our reviews table and counted the number of reviews for each cuisine.

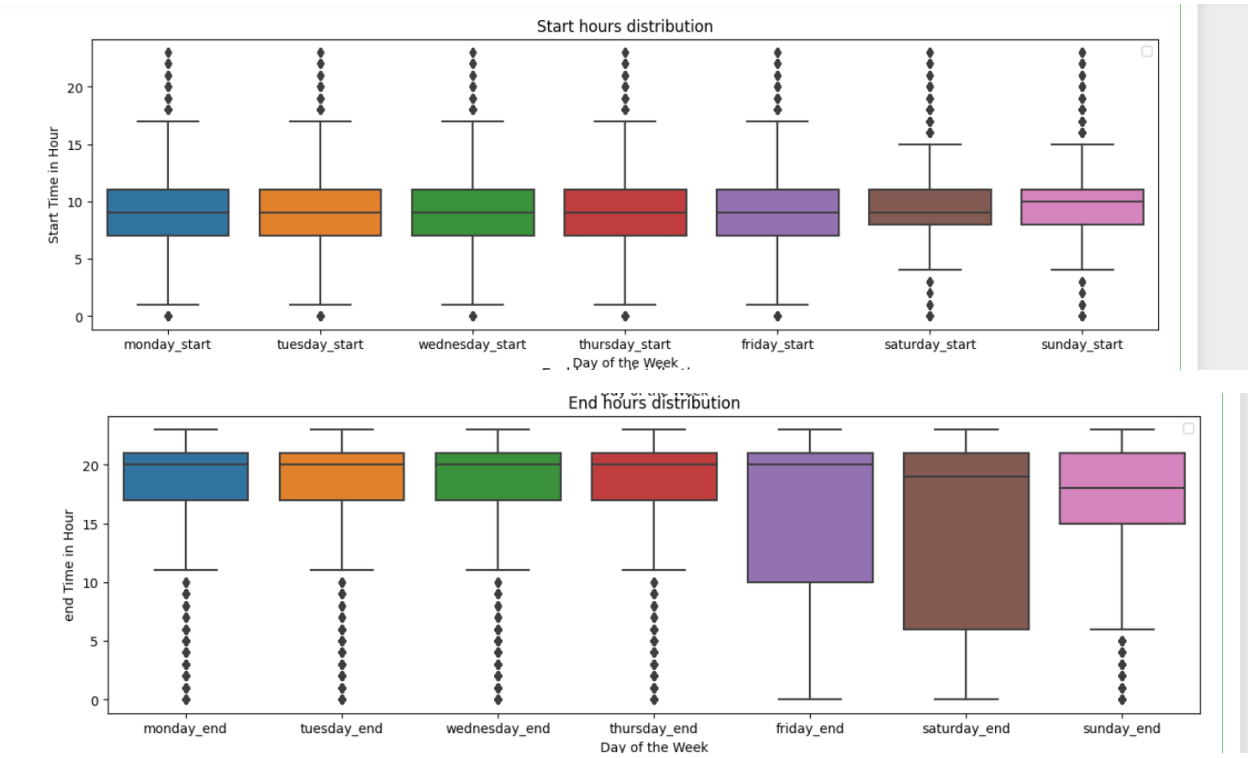


- For the above graph the x-axis is the years and y-axis the number of reviews and each line represents a different cuisine and type of cuisine can said by the color.

- **Inference –** It can be observed that the demand for American (new) and American (traditional) superseded the demand of other cuisines in the United States from early 2008. The rate of growth for these cuisines was very high. Next In list were Mexican and Italian followed by various Asian cuisines.

7) Average start and end time for restaurants

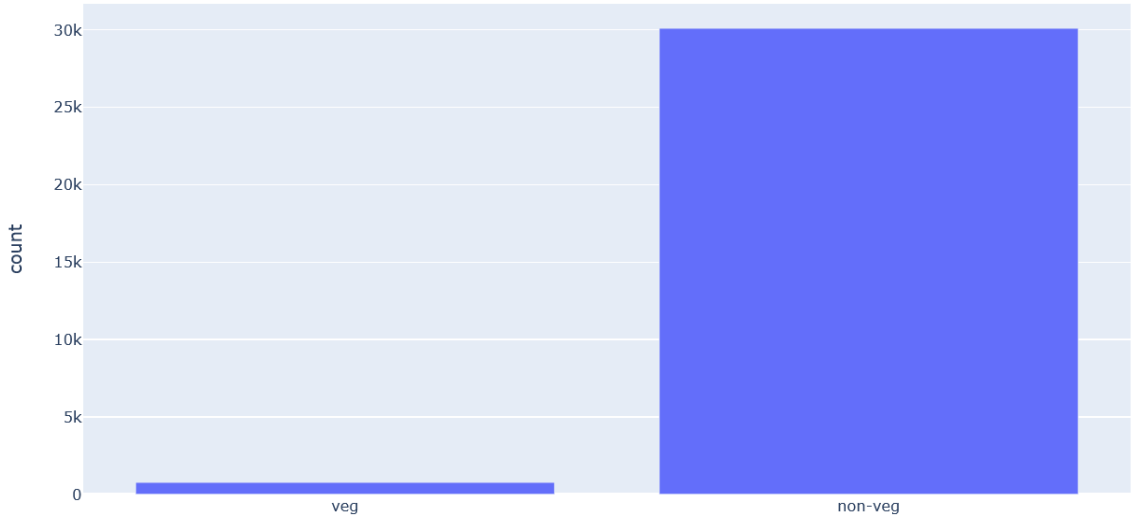
- Previously in the data cleaning we combined the business_hours with the business table and from that we extracted the opening and closing hours for each business.
- After the converted the hours for each day to datetime format for our analyzation.
- As the opening and closing hours are combined we divided using the str.split function.
- Plotted them in the box plot to get the high low and average start and end hours for all the restarants.



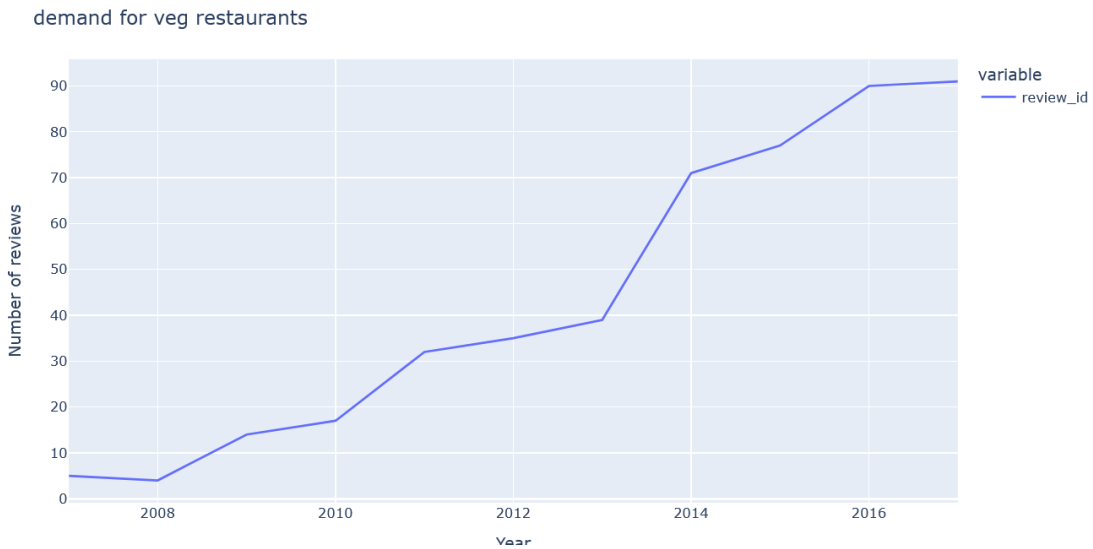
- In the above two plots the x-axis is day of the week and the y-axis the start in plot 1 and end in plot 2 in hours.
- **Inference -** On Weekends some restaurants close early, this can be owed to more demand or increasing awareness about work-life balance from the point of view of the restaurant staff.
- From the box plots we see that the average start time from monday- saturday is 9:00 a.m but on sunday its 10:00 a.m.
- the average end time on weekdays is 8:00 p.m but on weekends for satuday it is between 6:00 to 7:00 p.m

8) Analyzing the Supply and Demand for Vegetarian Options across US

- As the demand for vegetarian restaurants so we counted the number of restaurants and plotted the difference.
- And also taking the number of reviews for vegetarian restaurants for each year plotted the trend.



- For the above bar graph x- axis is the type of restaurant and y-axis is the count of the number of restaurants. As u can see the difference is huge.

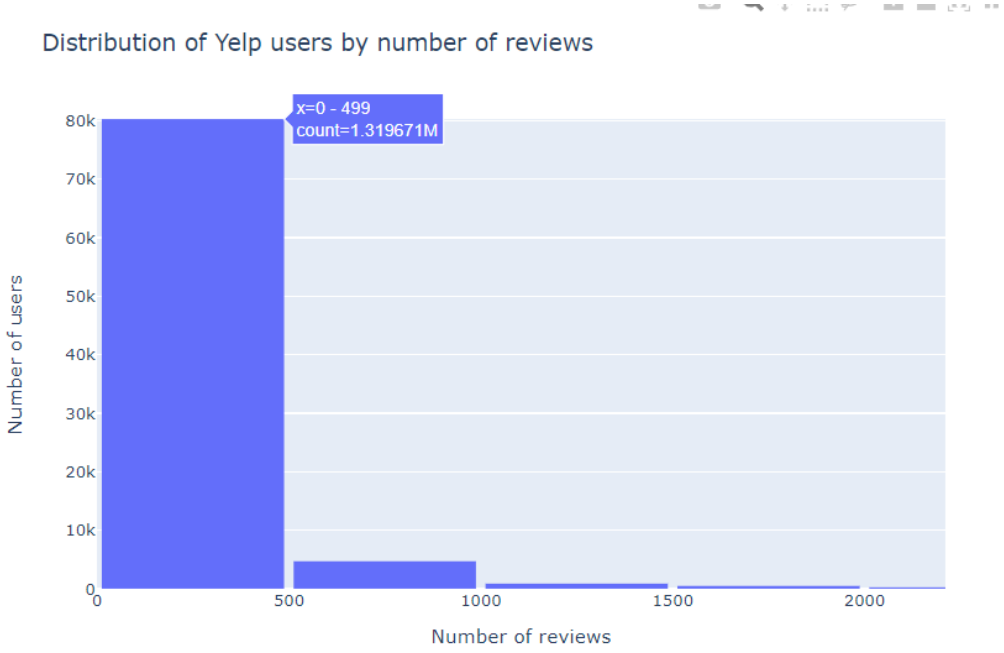


- The above graph is the trend of vegetarian restaurants. The x-axis represents years and y-axis represents the number of reviews. The trend is going up showing us the demand for vegetarian restaurants is going up.
- **Inference – We** can see that there are very less vegetarians' options in the United States but the demand for vegetarian restaurants is increasing over time.

- Hence starting a vegetarian restaurant would help cater this increasing demand, providing economic benefits for restaurant owners.

9) Analysis of Active Contributors on Yelp

- Getting the count of reviews for each user and calculating the users with non-zero reviews helps us figure out the number of users on the platform who contribute.
- This can help us analyze and improve engagement on the platform. Further it can prove beneficial to businesses who can track the users with more influence and reach to help promote their brands.



- **Inference –** We can see that most of the users on the platform have less than 500 reviews and do not actively contribute to the platform. Moreover, there are only 28 users who have contributed the most by writing more than 2500 reviews. This further enables them to be classified as elite contributors of the platform.
- We observe the trend that although the number of new users is increasing on the platform most of them didn't review, which suggests that users use the platform to only read the reviews and not give them, and the ones who have more reviews are likely to be influencers or food bloggers.

10) Mapping Distribution of Restaurants w.r.t Ratings

- We calculated the sum of restaurants per rating numeric by grouping the business_ids based on the stars.

- **Inference –** We can observe that diners rarely prefer restaurants rated 1 or 2 stars. It can also be inferred from the graph above that most of the businesses are rated between 3 to 4 stars and as we move outside the number of businesses gradually decrease, which loosely resembles gaussian distribution of the businesses across ratings.

V. References

Data sets -

1) https://www.yelp.com/dataset

2) https://www.kaggle.com/yelp-dataset/yelp-dataset

Other –

1) **Dr Eric Mikida & Dr. Shamsad Parvin – Class Lecture Slides and Lecture videos**
2) https://www.analyticsvidhya.com/blog/2020/04/beginners-guide-exploratory-data-analysis-text-data/
3) https://towardsdatascience.com/what-is-exploratory-spatial-data-analysis-esda-335da79026ee
4) https://www.kaggle.com/code/ponybiam/03-enth-geospatial-eda
5) https://www.kaggle.com/code/ekami66/detailed-exploratory-data-analysis-with-python/notebook
6) https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/