MSA 8050 SCALABLE DATA ANALYTICS

Final Project - Report

❖ **TEAM MEMBERS:**

Full Name	Email ID	Panther ID
Pranshu Srivastav	psrivastav1@student.gsu.edu	002602585
Shruti Shivaji Lanke	slanke1@student.gsu.edu	002583158
Suvaleena Paul	spaul25@student.gsu.edu	002638763
Ankita Mitra	amitra2@student.gsu.edu	002591022
Tania Haldar	thaldar1@student.gsu.edu	002253341

Industry Analysis Using Yelp Review

❖ **PROBLEM STATEMENT:**

For our project, we have covered three main csv files: **Business, User, and Review**. We have extracted some insights using these files individually or by combining them. Our primary focus is on building a model that can take any comment as input and determine whether that comment is made by an **"Elite"** Yelp user or by a **"Non-Elite"** Yelp user.

❖ **INTRODUCTION:**

Yelp is used by the general population frequently as visiting Yelp's website to see reviews has become a common step while looking for places to visit. While anyone can create an account and review, certain experienced users are given the title of "Elite". These Elite yelp users, according to Yelp, have a high impact on their community. They usually have detailed posts about their experience with good quality images to support their review. Their reviews and ratings can create certain hotspots around town for events and restaurants and can help small business owners grow. They can even start trends from their reviews based on the experience of their visit. As they are a critical group of users, we want to be able to identify who they are and what makes them different.

❖ DATA DESCRIPTION:

We have used three of the csv files provided in the Yelp dataset: **Business, User, and Review**. We have removed several columns that we did not use for our analysis. Below are the columns that we have used from each of the csv files to find generalized insights.

BUSINESS

Column	Description
Business_id	Unique ID of the Business
Name	Name of the Business
City	City of the Business
State	State of the Business
Stars	Star rating of the Business
Review_count	Total number of reviews for the business
Is_open	It's a binary value, 0 for closed and 1 for open
Categories	Name of the categories the Business is associated with.

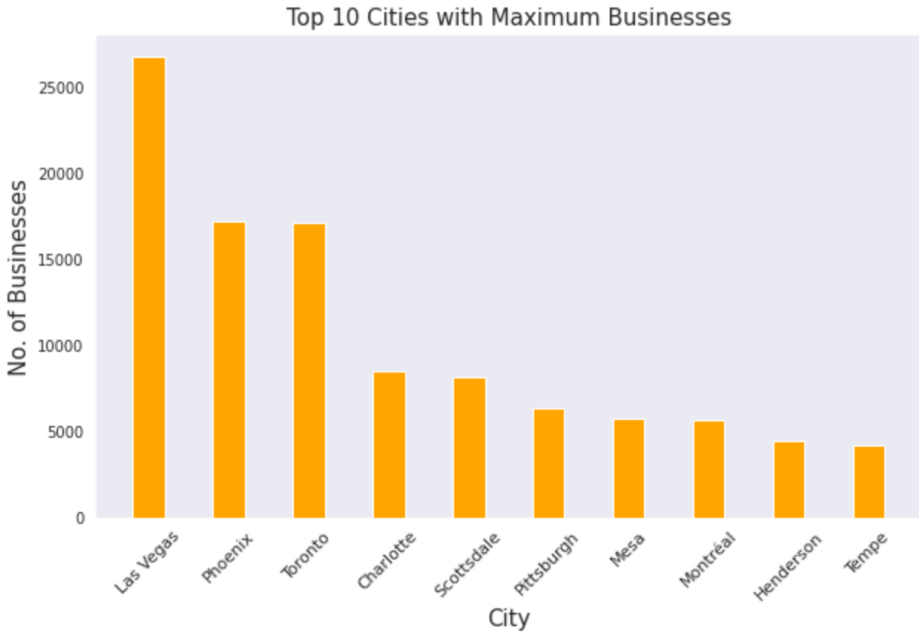
USER

Column	Description
User_id	Unique user id
Yelping_since	The date user joined Yelp
Useful	Number of useful votes sent by the user
Funny	Number of funny votes sent by the user
Cool	Number of cool votes sent by the user
Elite	The years the user was elite
Average_stars	Average rating of all reviews

REVIEW

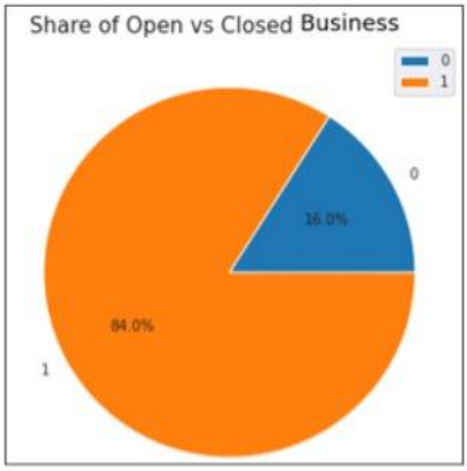
Column	Description
Review_id	Unique review id
User_id	Unique user id which maps to the user in yelp_user
Business_id	Unique business id which maps to the business in yelp_business
Stars	Star rating of the review
Text	The review itself

❖ DATA VISUALIZATIONS AND ANALYSIS:

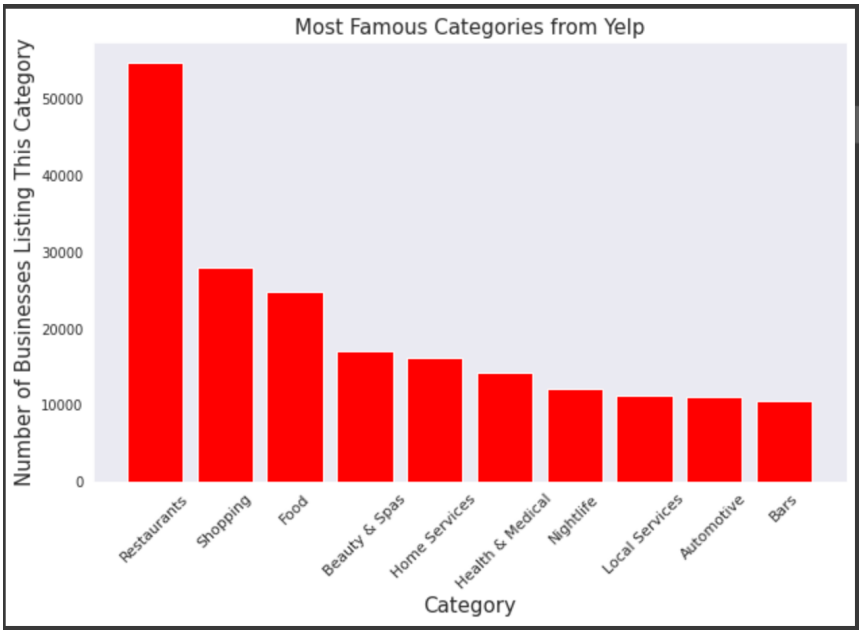


-

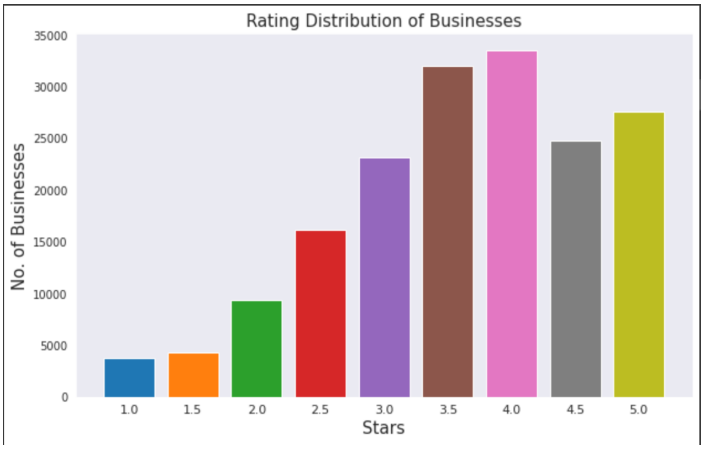
These cities have the most businesses in order and it is capped at 10 cities. Also, we notice that the dataset includes cities outside of the United States as well.



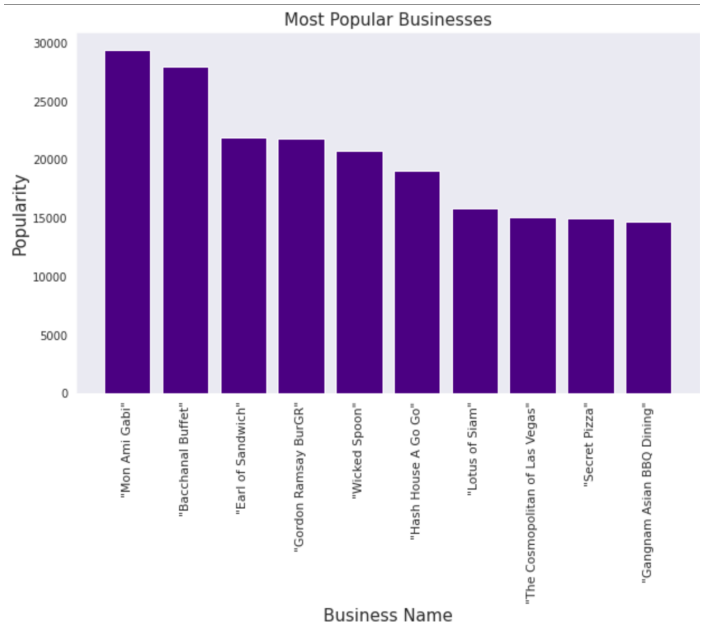
This pie chart shows a distribution of how many businesses remain open and how many have shut down.



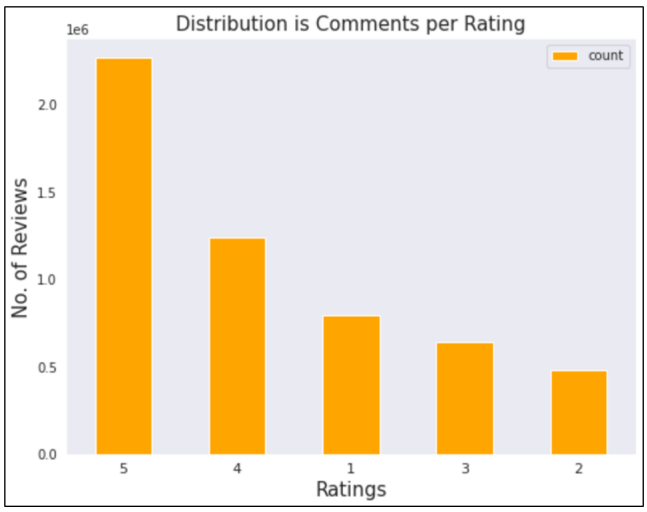
These are the top 10 categories used by businesses in the given Yelp dataset. Most businesses use the restaurant tag.



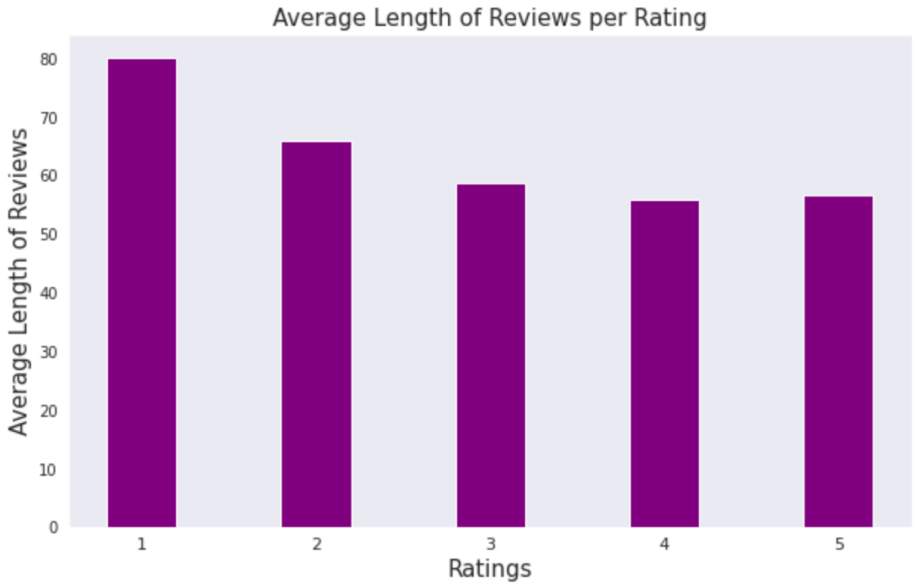
This bar graph shows us the distribution of how many of each rating is spread out across the businesses. An important note here is that the average rating of the business has been rounded. We are taking the average of the star ratings, then plotting the number of businesses against each star. 4.0 stars is the most common rating for businesses.



This graph shows the top 10 popular businesses within our dataset; most popular being Mon Ami Gabi. Popularity is determined by the number of reviews multiplied by the average star rating.



We have calculated the number of reviews per category. The number of reviews for 5-star rating is the highest which is an interesting point as next we will see that the length of comments for 5 star rating is one of the shortest; whereas, the length of 1 star rating is the highest.

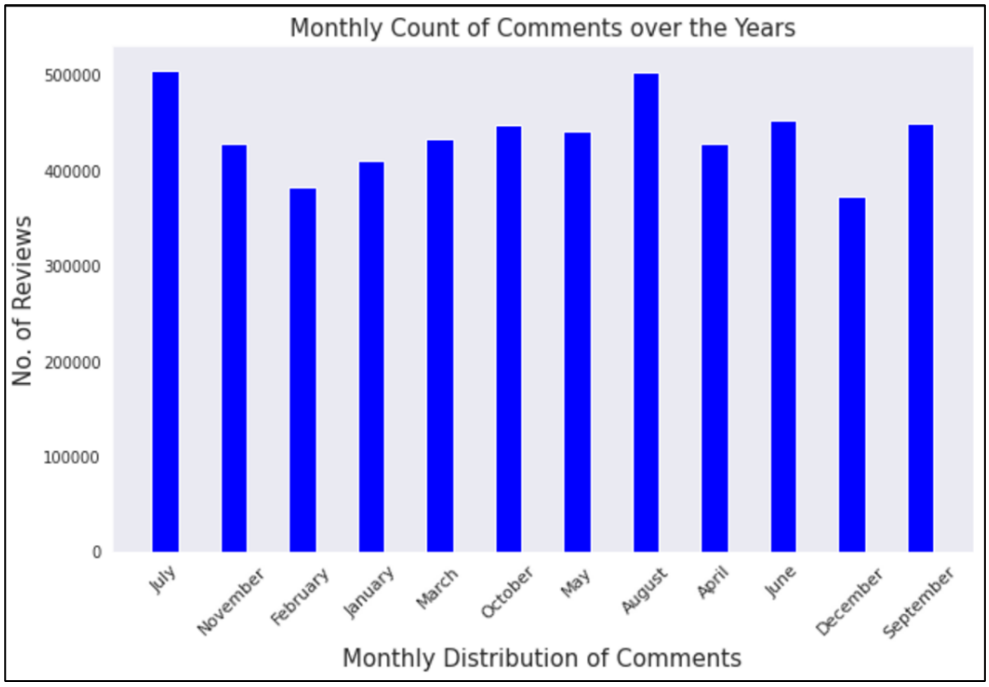


Here it is shown what the average length of reviews is in terms of words for each star rating. This could indicate that the person giving a one-star rating may be complaining and explaining the situation in the comments; whereas, the better reviews may just be short and simple. We have used **RDDs** to calculate the length of comments per star rating.

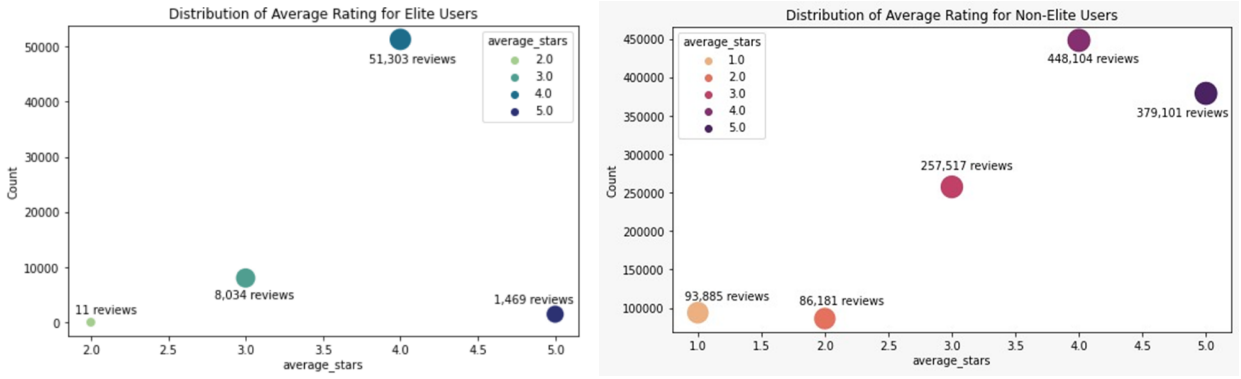
```
print("1 star rating: ",sort_1_words.collect())
print("2 star rating: ",sort_2_words.collect())
print("3 star rating: ",sort_3_words.collect())
print("4 star rating: ",sort_4_words.collect())
print("5 star rating: ",sort_5_words.collect())
```

```
1 star rating:  ['would', 'place', 'service', 'food', 'get', 'time', 'back', 'one', 'like', 'never']
2 star rating:  ['food', 'place', 'good', 'like', 'service', 'time', 'would', 'get', 'one', 'back']
3 star rating:  ['good', 'food', 'place', 'like', 'service', 'would', 'great', 'time', 'get', 'really']
4 star rating:  ['good', 'place', 'food', 'great', 'service', 'like', 'time', 'really', 'one', 'get']
5 star rating:  ['great', 'place', 'food', 'service', 'good', 'time', 'best', 'love', 'back', 'amazing']
```

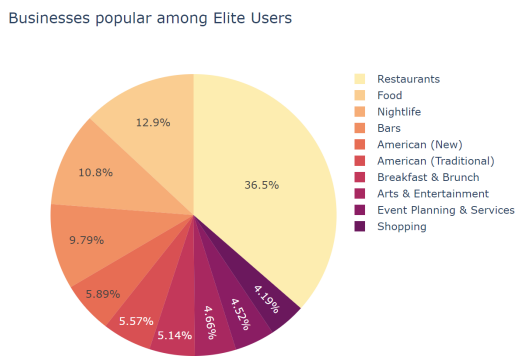
We have also used **RDDs** to calculate the top 10 words per star rating. For ratings 3 and 4, the top word used is “good”; whereas, for a five star rating, it is “great.”

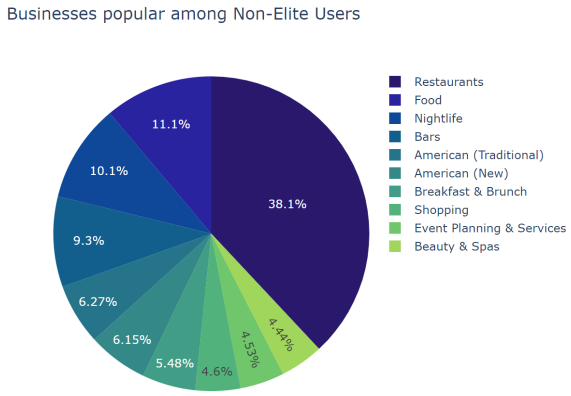


Here we have plotted how many reviews are left on Yelp's website on a monthly basis. We can infer that the months of July and August are busier than the other months.

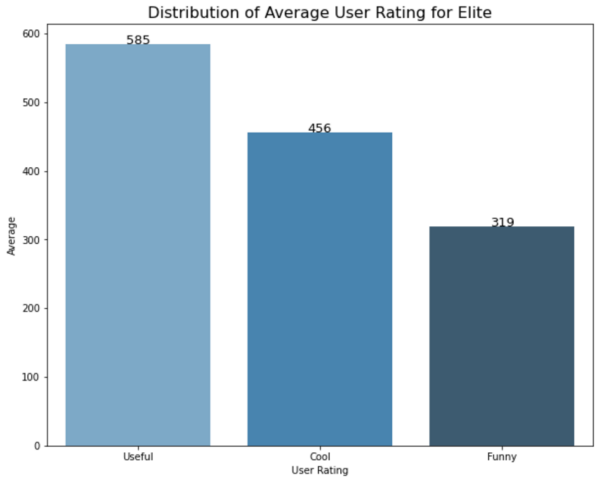


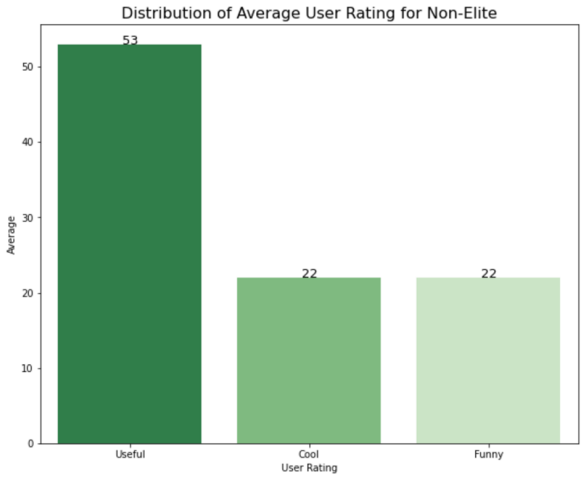
The scatter plot on the left shows the star rating distribution for Elite users and the one on the right shows it for the Non-Elite users. Since, there are more Non-Elite users in general, naturally they have a much higher average count. However, from both plots we can see that both Elite and Non-Elite users tend to give 4 star ratings most.



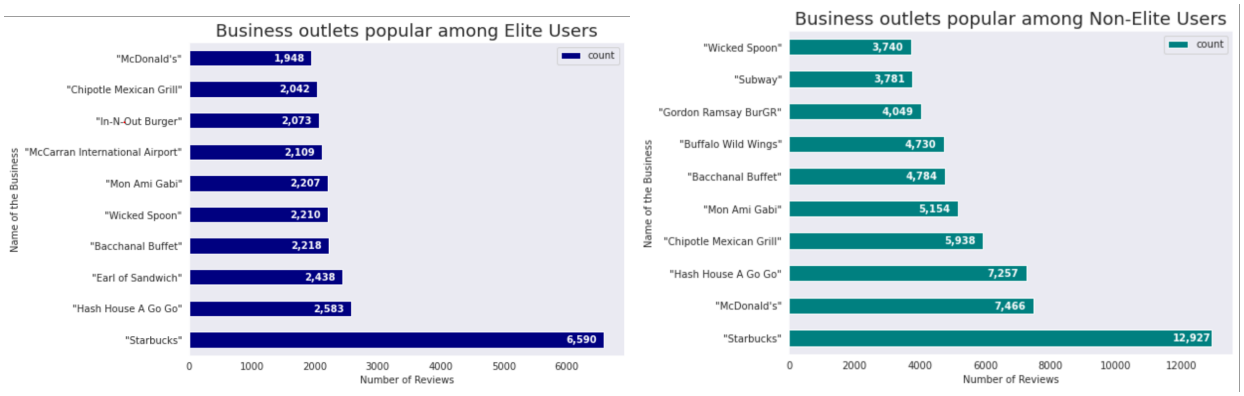


These two pie charts use **all 3 csv files** to show that both Elite users (left) and Non-Elite users (right) generally review the same types of businesses the most. In this case, restaurants are the most common types of businesses to review for both, followed by places of food, nightlife, bars, and so on.





These bar graphs show on average how much of the “useful,” “cool,” and “funny” tags Elite (left) and Non-Elite (right) users tend to get. Elite users have higher of all 3 tags compared to Non-Elites.



These horizontal bar plots above also combine data from all 3 csv files to show the most reviewed businesses amongst Elite users (left) and Non-Elite users (right). We can see that Starbucks is popular for both classes of users.

---

word	count
	1873049
place	434095
good	349986
food	323055
great	271609
like	248811
time	210253
really	201687
service	180477
nice	134555
love	134245
restaurant	115930
little	115204
pretty	114237
well	114172
always	113145
first	108588
best	107897
try	104575
much	97900

only showing top 20 rows

word	count
	7013420
food	1356110
place	1348042
great	1319842
good	1176111
service	1052788
time	880065
like	719647
really	554087
best	522560
staff	466179
love	460487
always	452751
nice	440401
friendly	427853
first	399293
amazing	398822
never	398014
well	386055
experience	343646

only showing top 20 rows

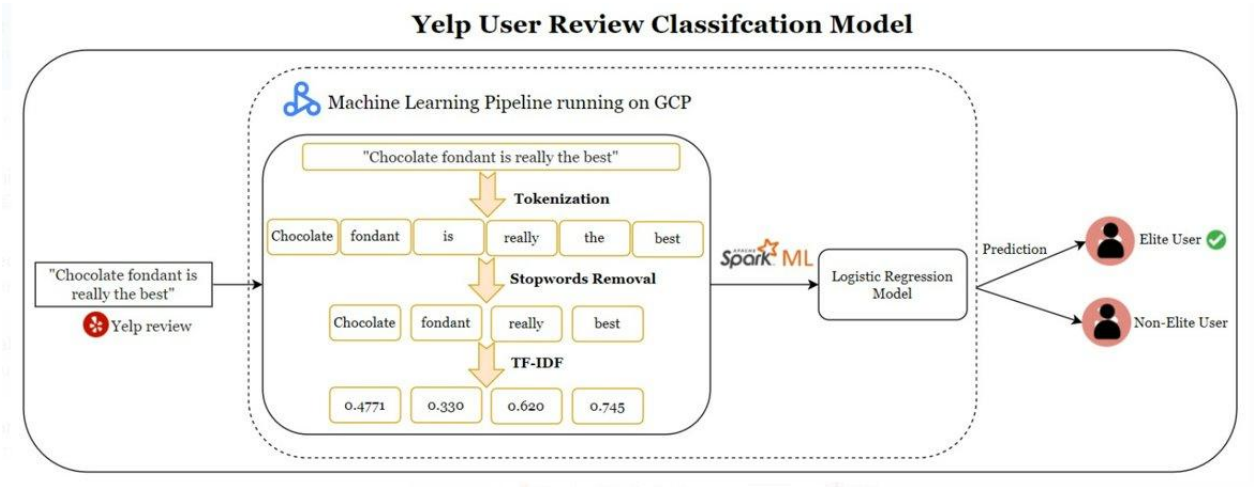
Elite **Non-Elite**

We also extracted the top words used in reviews for Elite and Non-Elite users and found that Elite users (left) mostly use “really,” “love,” “restaurants,” “place,” and “good,” whereas, Non-Elite users (right) mostly use “staff,” “service,” and “friendliness” in their reviews.

*All the aggregations and computations are performed using **PySpark dataframes** and **RDDs**, and then the aggregated tables are converted to Pandas dataframes only for visualizations.*

❖ MACHINE LEARNING PIPELINE & STREAMING:

We are working with a **classification problem** where the label Is_Elite is predicting whether a particular review is from an Elite user or not. We used **Tokenizer** to tokenize the review where input was “Text” and output was “Words”. We removed all stop words using **StopWordsRemover** where input was “Words” and output was “Filtered”. We calculated both the **Term Frequency** using **HashingTF** and **Inverse Document Frequency**. We then used **Logistic Regression** with **max iteration as 10** and **parameter as 0.001**. We then split the data in the ratio of **90% for train and 10% for test** and ran the model pipeline on **GCP**. On testing this model for the test data, we get an **accuracy of 77.55%**. The model is then saved and downloaded to local using scp commands for streaming computation.



In our **streaming demo**, we used a Non-Elite user’s comment as input (top), to which our model correctly predicted that it came from a Non-Elite user, hence the 0 for prediction (left). We also tested our model with an Elite user’s comment as input (bottom), to which our model correctly predicted that it came from an Elite user, hence the 1 for prediction (right).

```
Hanging out in Vegas this weekend and staying at Paris so decided to check this out at Planet Hollywood since its next door. Have always wanted to check out one of Gordon Ramsey's restaurants. Had the tur
key burger (super good!) and the sweet potato fries. They were a little sweet for my taste but still good. You also get several different sauces to dip them in (house ketchup, spicy ketchup, curry, etc) s
o those were fun to try. If your looking for a good burger, should definitely check it out.
```

```
This is definitely a good example of a brunch place where what you order can have a significant impact on whether you love or hate the food. I ordered the California crepe, which was not made well and tas
ted fairly bland, despite how promising the combination of ingredients sounded. My friend, however, ordered the chicken fried steak, which was great. This brunch place has a HUGE menu so just make sure yo
u order something that you will definitely like, and you're in the clear. Additionally, the free banana muffin is delicious.
```

```
+--------------------+----------+
|                text|prediction|
+--------------------+----------+
|Hanging out in Ve...|       0.0|
+--------------------+----------+
```

```
+--------------------+----------+
|                text|prediction|
+--------------------+----------+
|This is definitel...|       1.0|
+--------------------+----------+
```