



# Yelp Data Analysis

**Vinay Kota,  
Utkarshika Baidya,  
Hemanth Doddala  
MPS Data Science**

## ☞ Yelp Data Analysis

---

The yelp data set is a trove of reviews, businesses, users, tips and check in data. In this project we are analyzing restaurant out of the other businesses found on yelp such as medical services, hotels etc. This notebook comprises of data for years(Feb 2005-Jan 2022). We have used yelp data to visualize the ratings and reviews for businesses in a particular area or industry. This could be a graphical representation that shows the average rating for each business, as well as the distribution of ratings across all businesses in the area. We also included other details, such as the number of reviews each business has received, and any significant trends or patterns that emerge from the data. This type of visualization could be useful for both consumers and businesses, as it provides a quick and easy way to understand the overall sentiment surrounding a particular area or industry.

# Dataset description

---

Extraction- <https://www.yelp.com/dataset>

Our dataset is from Yelp.com. The dataset contains 5 JavaScript Object Notation(JSON) files. The yelp data set is a trove of reviews, businesses, users, tips and check in data all of which can be used for private, academic and educational purposes.

1. Business Dataset: Contains information about businesses, such as geographic information, attributes, rating, and categories. This dataset consists of 150346 rows and 14 columns (5.34 GB)
2. Review Dataset: Contains complete review text information, such as the user who authored the review and the business for which it was written. This dataset consists of 6.9 million rows and 9 columns (5.34 GB)
3. User Dataset: User dataset contains information about user, friend mapping and all the metadata associated with user. This dataset consists of 2 million rows and 22 columns (3.36 GB)
4. CheckIn Dataset: Information about user checking at the business locations. This dataset consists of 131930 rows and 2 columns
5. Tip Dataset: user-written feedback on a company. Reviews are longer than tips, and tips typically contain brief recommendations. This dataset consists of 908915 rows and 5 columns.

# Objective

---

## Initial questions

We started with a vision to use the yelp restaurant dataset to build a recommender. Analyzing some of the questions ad digging into further investigations about the business and come up with recommendations that would help restaurant owners to better decide where to open business as well as ways to improve their businesses. During this process we discover the Yelp Dataset, dig deep and find interesting connections. We realized that our data only allowed us to take two tracks: analyze business or analyze user. Hence, we dug deeper into both kinds of data and came up with number of question before we decided to evaluate user data in greater detail. Narrow our investigation to Fast Food chains and build a recommender for those based on location.

---



## **Step-1 Cleaning data (handling null values, dropping columns, changing data types)**

```
df_business = df_business.drop('attributes','address', 'hours')
df_business.printSchema()
```

```
root
|-- business_id: string (nullable = true)
|-- categories: string (nullable = true)
|-- city: string (nullable = true)
|-- is_open: long (nullable = true)
|-- latitude: double (nullable = true)
|-- longitude: double (nullable = true)
|-- name: string (nullable = true)
|-- postal_code: string (nullable = true)
|-- review_count: long (nullable = true)
|-- stars: double (nullable = true)
|-- state: string (nullable = true)
```

```
[ ] df business.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in df business.columns]).show()
```

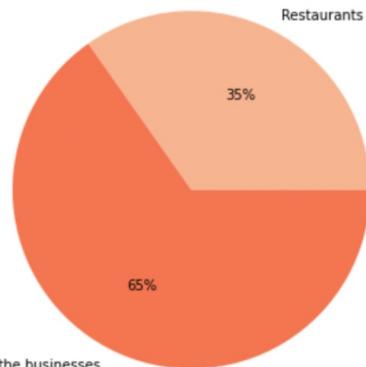
```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|business_id|categories|city|is_open|latitude|longitude|name|postal_code|review_count|stars|state|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|          0|     103|    0|      0|      0|      0|    0|        0|         0|    0|    0|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

```
[ ] # Filtering restaurants from the business dataset  
restaurants_df = df_business.filter(df_business['categories'].contains("Restaurant"))
```

```
▶ business_count=[restaurants_df.count(), df_business.count()-restaurants_df.count()]  
labels=["Restaurants", "Rest of the businesses"]  
colors = sns.color_palette('rocket_r')[0:2]  
plt.figure(figsize=(6, 6))  
plt.pie(business_count, labels = labels, colors=colors, autopct='%.0f%%')  
plt.title("Restaurants vs Other Business Percentage")  
plt.show()
```



Restaurants vs Other Business Percentage



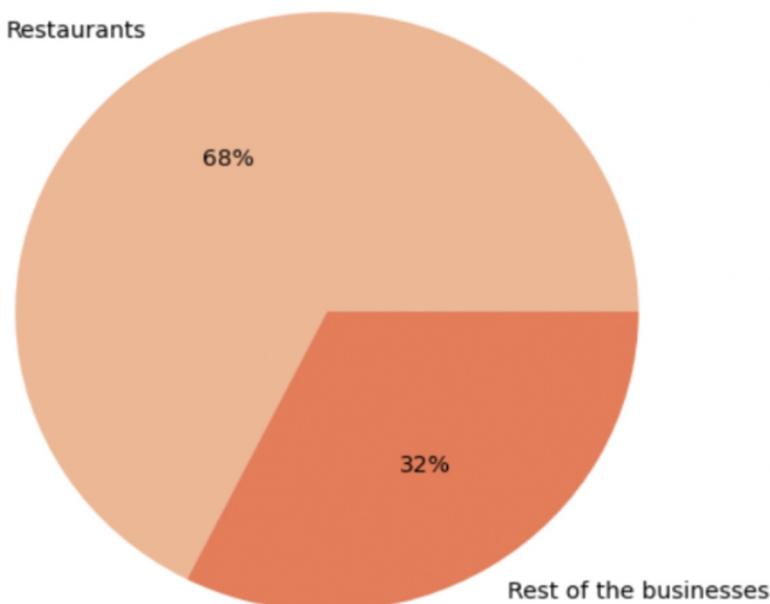
From this our observation was that we are working with 35% of the businesses which are restaurants and the rest which come up to be 65% are hotels, hospitals etc.

### Step- 3 (Analysis)

After understanding the datasets and the components we wanted to analyse the combination of the datasets and bring out meaningful information from it. Hence, we started to join the datasets. We have layers to our analysis which are shown below.

1. Reviews distribution: Our first finding was that out of total about 68% of the reviews that are on Yelp comprise of reviews for the restaurants and only 32% for other businesses, i.e 2/3 of the reviews.

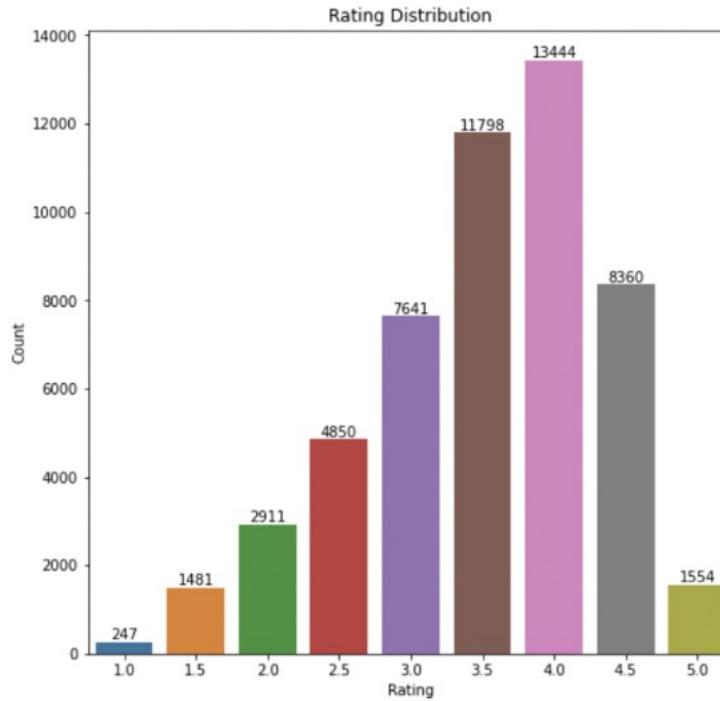
Restaurants vs Other Business Review Percentage



## Findings

Rating distribution: From our analysis we were able to see the overall rating or the distribution of ratings across the restaurants, where high number of restaurants saw a rating between 3.5 to 4.5

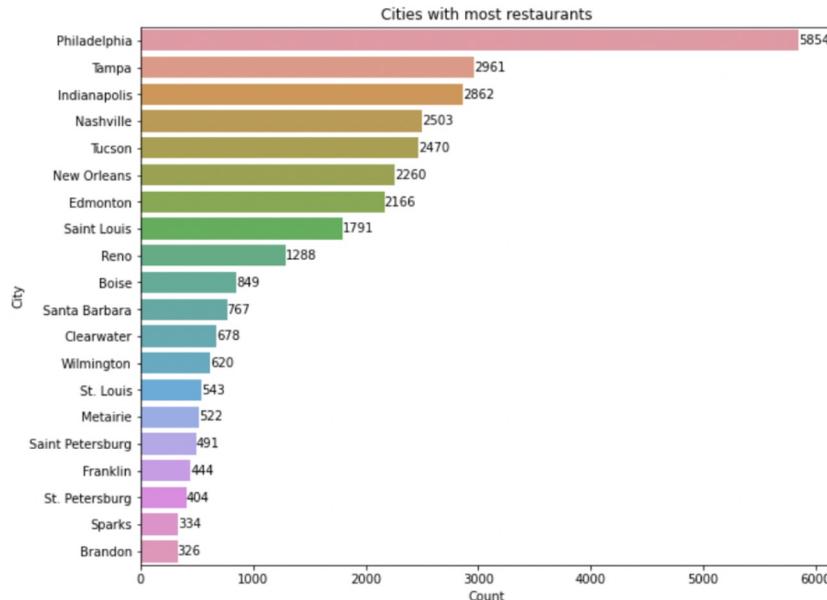
Trend for checkin and review to the star ratings: Checkin\_count and the review count for the restaurants which shows that because of the rating of the restaurants the checkins have been affected.



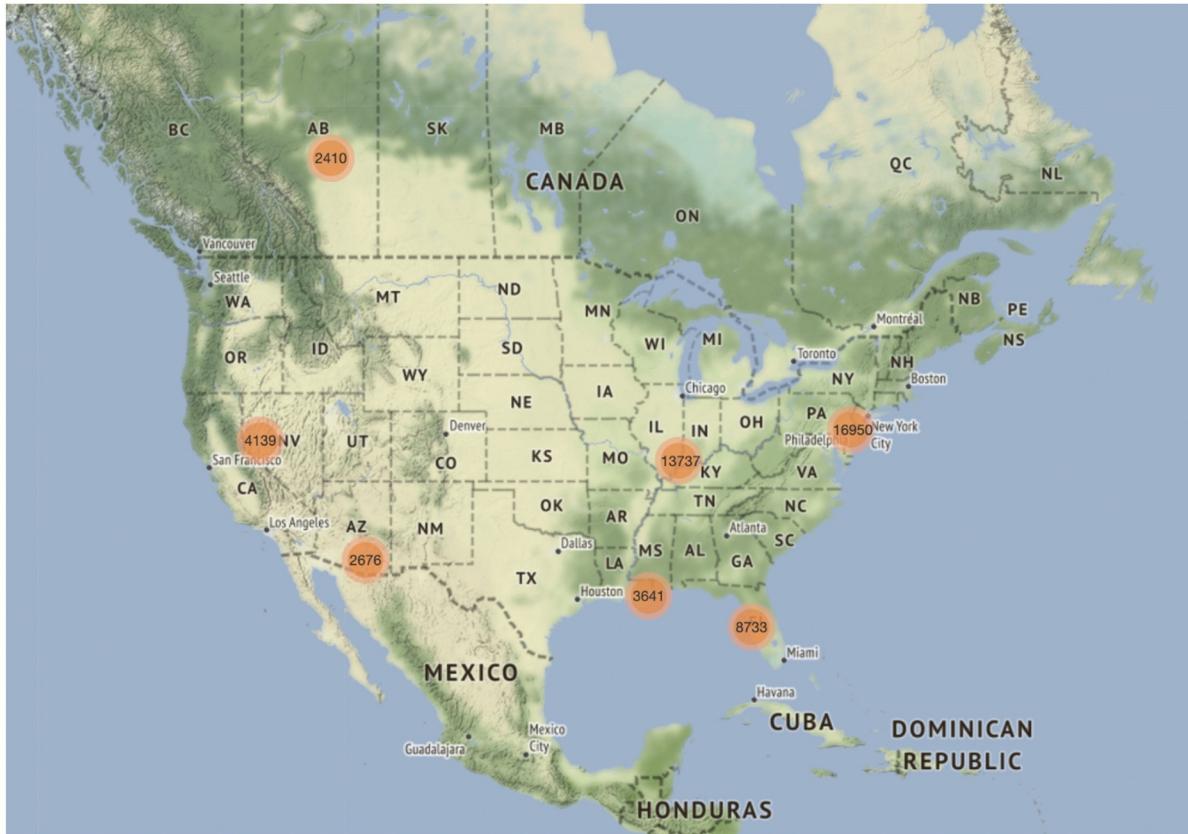
## Step -3.1 (Location based Analysis)

We wanted to get location based data so, We found

1. where the restaurants were located
2. how many in a particular city (most and least number of restaurants in the cities)



4. We also have a interactive heatmap to show the states and the restaurants in that area.



## Review Words

For 5 star and 1 star Ratings

Findings : Cuisine Names,  
Adjectives, for food and ambiance

We wanted to analyse the pattern of words that are used in the top rated(5 star) and worst rated( 1 star) using WordCloud. We were able to find out that both positive as well as negative meaning words are used in both the cases

Most used words in the 5 star reviews



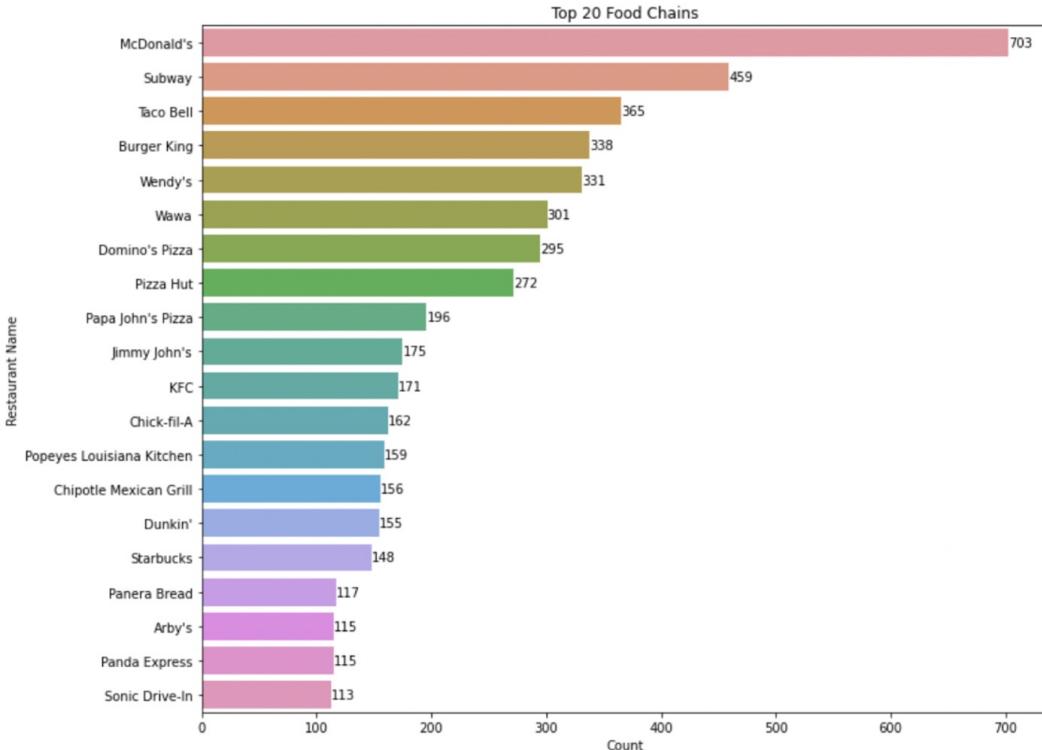
Most used words in the 1 star reviews



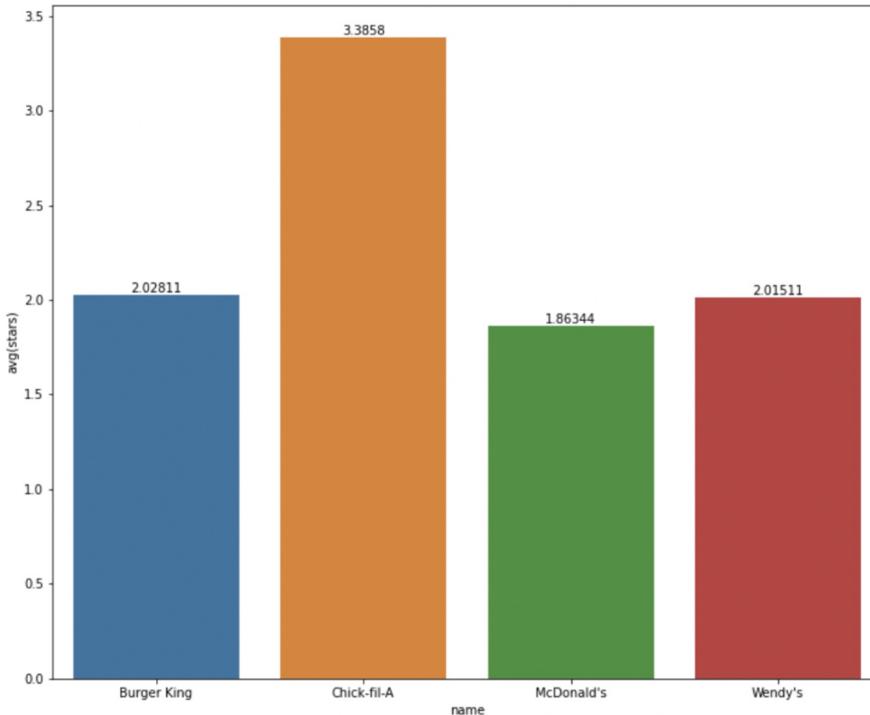
## Fast Food Chains:

Here we see that McDonald's is the number 1 fast food joint followed by Subway, Taco bell and so on. We then just took the most popular burger joints namely McDonald's, Burger King, Wendy's and Chick-fil-A for further analysis.

1. Top 20 fast food chains in US



2. Ratings of the food chain business : To our surprise thought the food joints were the top the list the star ratings told a different story. The ratings for these were 1.5 to 2 stars. Among the 4 chosen burger joints Chick-fil-A was the most promising one with around 3.5 star rating.





3. We analysed chick- fil-A which has the highest and the lowest rating and compared them .

	business_id	categories	city	is_open	latitude	longitude	name	postal_code	review_count	rating	state
0	I920q9aMfPWJ0Jq11rkfkQ	Fast Food, Event Planning & Services, Caterers...	Barrington	1	39.87259	-75.045165	Chick-fil-A	08007	5	5.0	NJ
1	dqhYXtw1P--1MpxBKQMkg	Restaurants, Fast Food	Franklin	0	35.89534	-86.877163	Chick-fil-A	37064	14	5.0	TN

	business_id	categories	city	is_open	latitude	longitude	name	postal_code	review_count	rating	state
0	EQoeJ4o6re7EC4_EONXPlg	Chicken Shop, Restaurants, Fast Food	Riverview	1	27.797713	-82.333970	Chick-fil-A	33579	87	1.5	FL
1	8Tg3imlkSI8ugCDM4bpSeg	Caterers, Chicken Shop, Event Planning & Servi...	Nashville	1	36.132446	-86.795202	Chick-fil-A	37212	6	1.5	TN

## Reviews for chick- fil -A:

In the picture we can see reviews for chick-fil-A with 1 star and the 5 star rating in the same area.

- 0      The people work there are very nice and friend...
- 1      Excellent Chick-fil-a restaurant experience. ...
- 2      Had a moment before heading over to the car de...
- 3      Wow! My boyfriend and I came right before clos...
- 4      Some of the best service I've ever received at...
- 5      The Chick- fili- A chicken was excellent and th...
- 6      This location isn't closed, not sure why it sa...
- 7      Yum! Always friend, friendly and delicious. \n...
- 8      As with any Chick Fil A the service was amazin...
- 9      Consistently friendly, clean and hot fresh foo...
- 10     How can you go wrong with CFA? The food is gre...
- 11     Very solid Chick Fil A. Food is what you expec...
- 12     Another excellent FSU (free standing unit)\n\n...
- 13     I did the drive thru during the lunch rush and...
- 14     It's chick fil a on the pike (and not the blac...
- 15     Impressive and clean shiny and new Chick Fil A...
- 16     Brand-new CFA serving up all your favorites wi...
- 17     Stopped here on the way home from camping in C...
- 18     I love chick-fil-a I eat chicken buggies and f...

3.1 Review: For Good Ratings

Name: text, dtype: object

- 0      At 5 pm you would think there was enough time ...
- 1      This knock off chick fil a is the worst! The f...
- 2      Awesomeness place friendly staff and made \nTo...
- 3      Even though I was on a church camp and used th...
- 4      The food was extremely over priced (even for C...

...

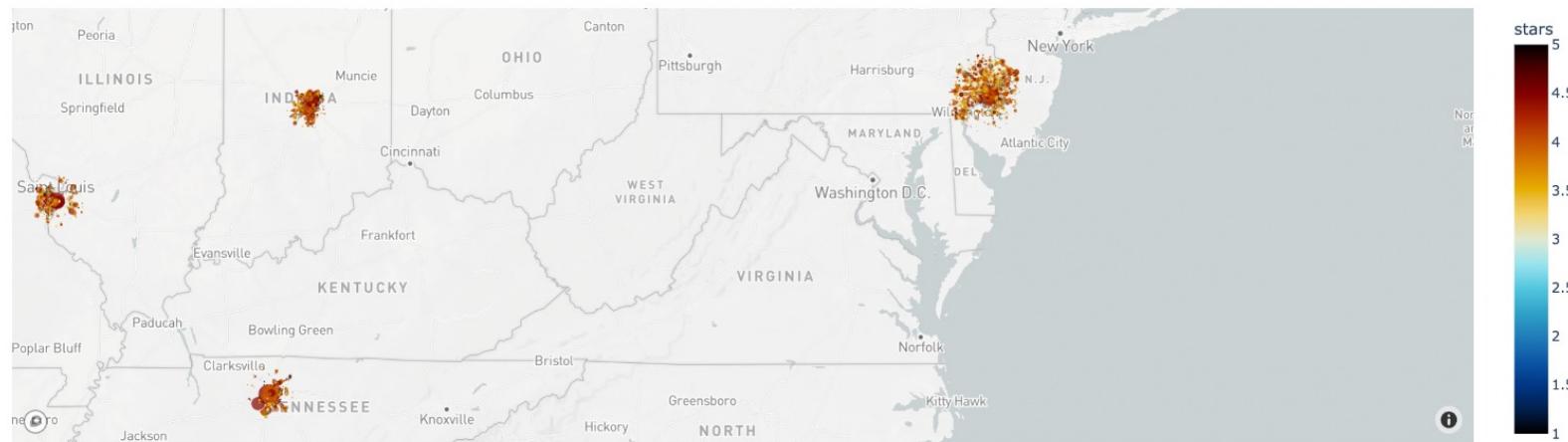
- 96     This location in Riverview is a nightmare to a...
- 97     Would give 0 starts if possible. The SLOWEST l...
- 98     It's guaranteed they forget utensils and straw...
- 99     The only Chick-fil-A in America that limits yo...
- 100    Absolutely the worst chic fila experience and ...

For Bad Ratings

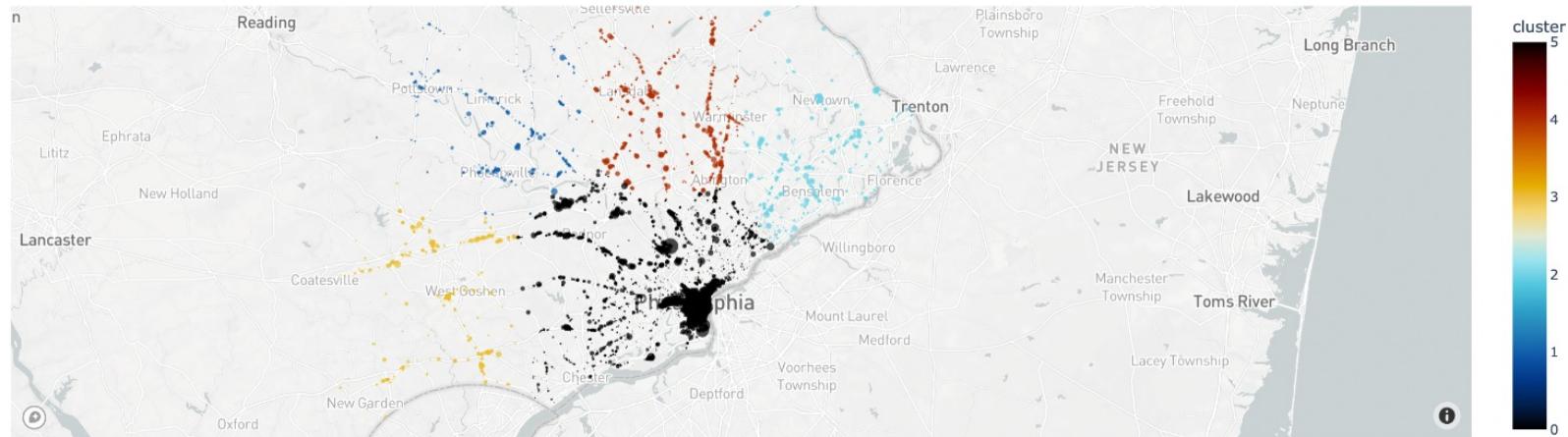
Name: text, Length: 101, dtype: object

## Step -5 (Recommendation system based on location)

Scatter plot that shows the Restaurants based on the ratings



- Since, Philadelphia consists of most number of high rated restaurants, the restaurant distribution has been divided into clusters using kmeans algorithm. Using silhouette score, the best number of clusters is chosen as 5. A function then helps in recommending the best restaurants given a certain co-ordinates.



Top 5 recommendations based on latitude and longitude (users location)

	city	name	rating
5839	Philadelphia	Tortilleria San Roman	5.0
9977	Philadelphia	Hikari Sushi	5.0
11897	Philadelphia	Christie's Deli	5.0
7180	Philadelphia	Miss Rachel's Pantry	5.0
6649	Philadelphia	El Rancho Viejo	5.0

- **Suggestions for the restaurant business based on the analysis finding**
  1. Service is the most important factor, if you want to have better ratings or customer attraction you must improve on the service.
  2. Hygiene is another factor that determines the likeness for the customers.
  3. Businesses should look at their reviews and try to improve based on that given reviews as they say a lot
- **Challenges**
  1. Reading data from Hadoop after inserting data into HDFS.
  2. While cleaning data for business- the same restaurant has multiple spellings listed(unique entries but is the same restaurant).
  3. Not able to Use MLlib for clustering.

# Conclusion

1. Among all the business that are found on yelp Restaurant business comprises of 35% of the total.
2. About 2/3rd of the reviews that can be found on yelp are comming from restaurants i.e 68%
3. Tips which are short reviews make up 71% from the restaurant business on yelp, which started from 2009
4. Majority of restaurants fall under 3.5 to 4.5 star ratings.
5. **Star ratings have an impact on the checkins from customers.**
6. Philadelphia, Tampa and indianapolis are the top 3 cities with highest number of restaurants
7. There are cities which have just 1 restaurant listed on yelp
8. Review count of Philadelphia is directly proportionate to it reviews and has the highest reviews for the restaurants.
9. **To our surprise the top fast food/ burger joints have the worst rating. out of which Chick - fil- A is the best as of the average rating with 3.3.**
10. **Chick- fil- A though was the best among the others they also had ratings that were 1.5 and 5 in certain areas. Since we found 1.5 and 5 both are in the same area , location really doesnt play a significant role but it is the service that determines whether the business is good or bad.**
11. **We were able to come up with a model that gave us a location based recommendation where user gives the latitude and longitude.**

Note: The orange bold bullets are key findings out of the many

Thank You