CS 6675
Homework2

Task 1: Data Download and Descriptive Analysis

I have used to Python to extract the data from the json files and analyze the data.

Which types of restaurants are the most popular in each city? Define and describe what "popularity" means.

I have taken the business ID's of restaurants in each of the city and extracted the corresponding no of check-ins for each of the business ID from the check-in data.

I defined the **restaurant as popular** if it is among

- the top 50 restaurants of the 100 restaurants (Obtained by taking the top 100 restaurants with higher no of check-ins) which have higher average review rating of the restaurants.

Taking into consideration the top 50 restaurants in each city, I calculated the no. of restaurants in each of the different type of restaurants. The type of restaurant which has more restaurants in the city is the popular type of restaurants in the city.

For **Pittsburgh**:

The most popular type of restaurants are restaurants with **Nightlife, Bars and American cuisine**.

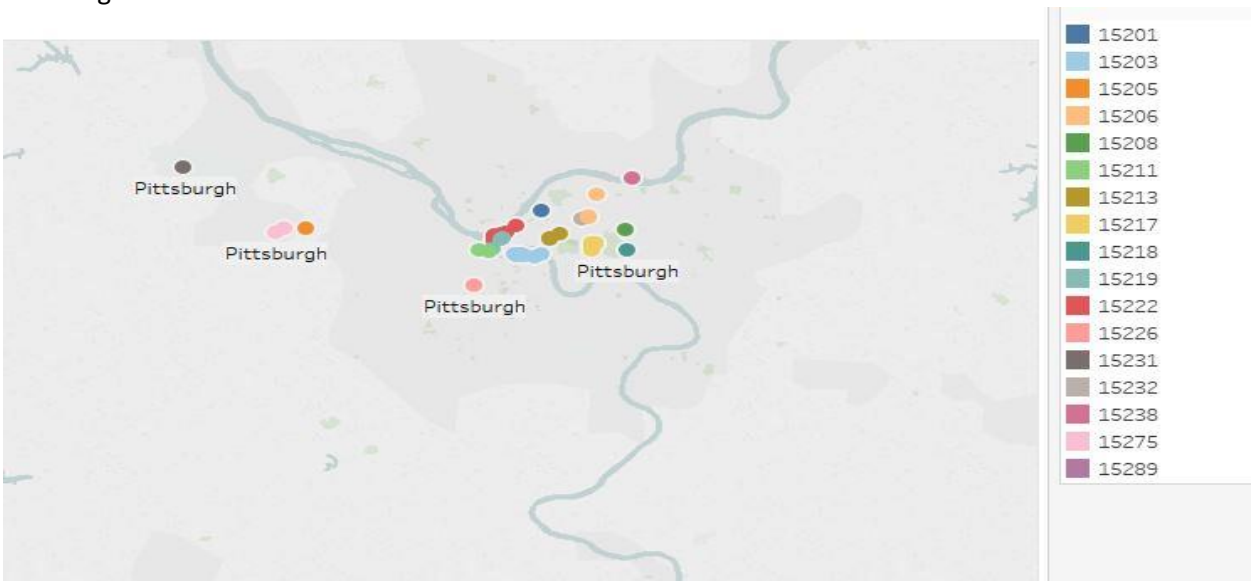
For **Charlotte**:

The most popular type of restaurants are restaurants with **Nightlife, Bars and restaurants serving Breakfast and Brunch.**

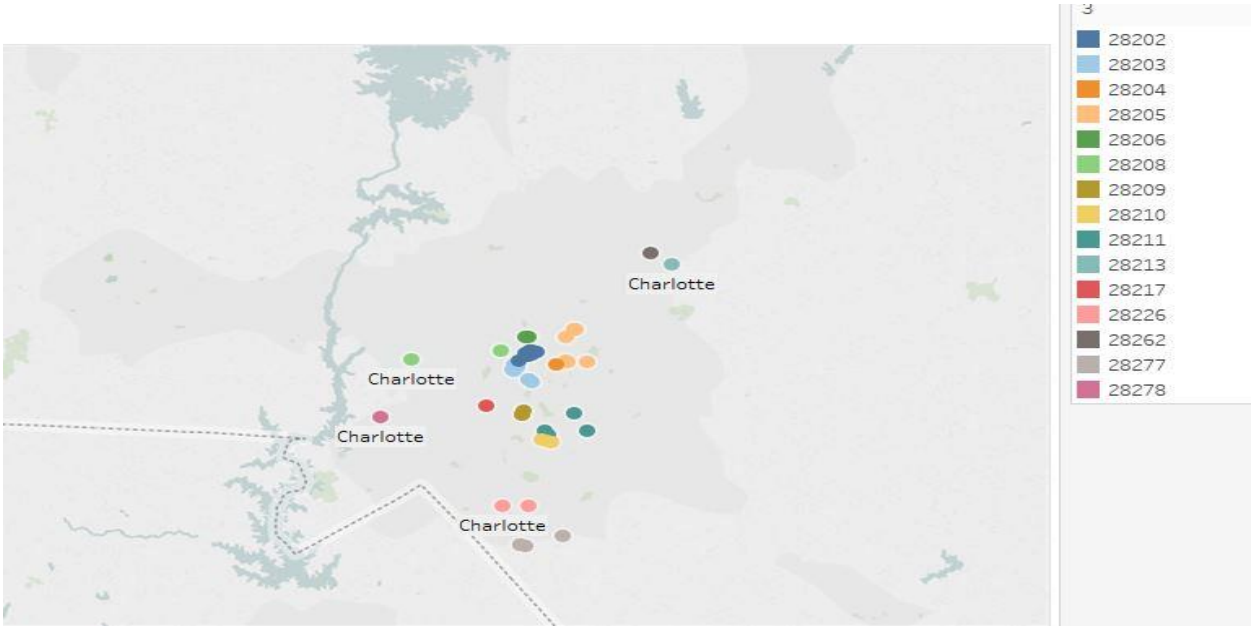
Distribution on the map.

The below are the locations of the popular restaurants in each of the cities classified by their Zip codes.

Pittsburg

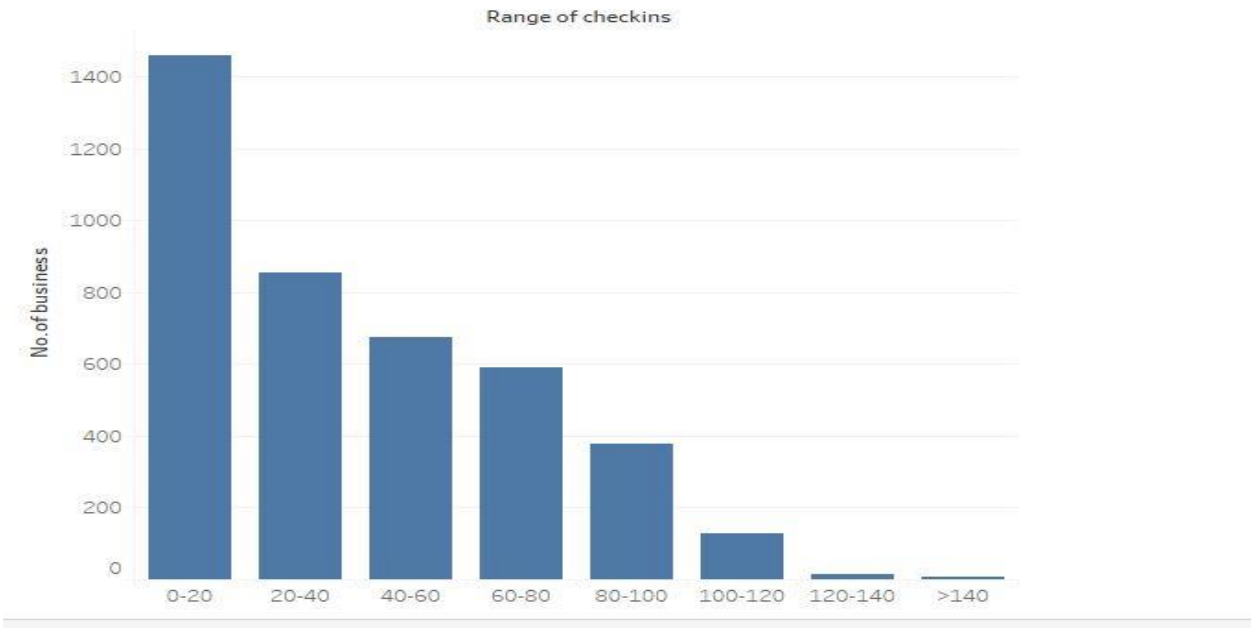


Charlotte



Descriptive analysis

The below figure shows the no. of business in each of the no. of check-ins range



A lot of business have less no. of check-ins and very few business have high no. of check-ins

Task 2: Text Processing & LDA (Latent Dirichlet Allocation)

Extracted business ID's of restaurants which have the terms "Chinese" and "Restaurants" in categories and then using the business ID's extracted all the reviews of the Chinese restaurants from the reviews data.

The following operations are applied on the each of the review extracted:

- Tokenization
- Removal of stop words in the review
- Stemming the words in the review

To find which words are most frequently used in the describing Chinese restaurant, I have counted the occurrences of all the words in the reviews.

The words that are frequently used to describe Chinese restaurants are **food, place, order, chicken, service, dish, noodles etc.**

Major themes/topics in reviews of Chinese restaurants

To find the major topics in reviews of Chinese restaurants, I have applied LDA on the review after applying the operations mentioned above.

The result of LDA is

1. 0.024*"chicken" + 0.016*"noodl" + 0.016*"rice" + 0.015*"fri" + 0.015*"soup"
2. 0.034*"food" + 0.025*"place" + 0.020*"good" + 0.018*"chines" + 0.018*"s"
3. 0.022*"t" + 0.016*"food" + 0.015*"order" + 0.010*"time" + 0.009*"us"
4. 0.039*"sushi" + 0.031*"buffet" + 0.018*"drink" + 0.018*"bar" + 0.014*"ice"
5. 0.029*"die" + 0.027*"und" + 0.026*"da" + 0.016*"nicht" + 0.015*"ist"

Hence the major topics are about

1. Dishes like Chicken noodles, fried rice and soup
2. How good is the place and food?
3. The service of the restaurant (Words like order and time)
4. Drinks available in the restaurant
5. No topic can be inferred from this.