**Homework #2**

**CS 6675, Spring 2017**

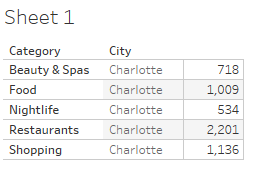
**Task 1: Data Download and Descriptive Analysis**

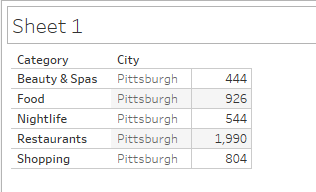
In this task, extract business IDs from yelp\_academic\_dataset\_business.json if a business’s city/location is either “Pittsburgh” or “Charlotte”. By using each business ID as a unique key, extract corresponding checkins and reviews from yelp\_academic\_dataset\_checkin.json and yelp\_academic\_dataset\_review.json. Then, answer the following questions:

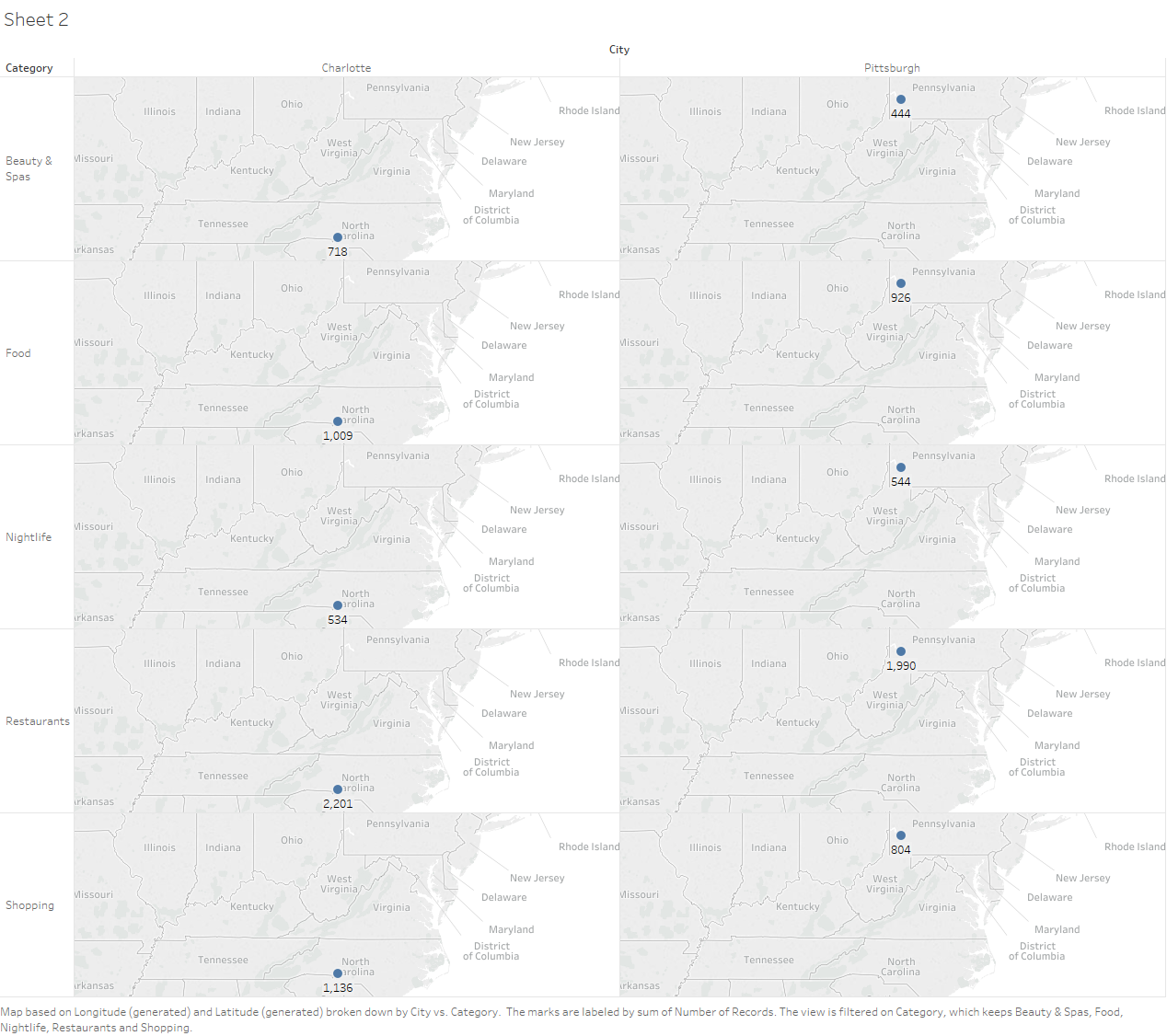
* Which types of restaurants are the most popular in each city? Define and describe what “popularity” means.
* In order to understand the popularity of each city, visualize distribution on the map using tools like Tableau, Google maps API, basemap, D3, etc. Report your findings including some figures like snapshots of the maps.
* Report one more interesting finding through descriptive analysis.

**Answer:** I extracted the data from JSON and saved it in the form of CSV using python script. First I read the business json and extracted the ‘business\_id’ for ‘Pittsburgh’ and ‘Charlotte’ and saved them in csv file. Then used each business\_id from csv to find checkin and review details from checkin and review json files and saved them in csv files respectively.

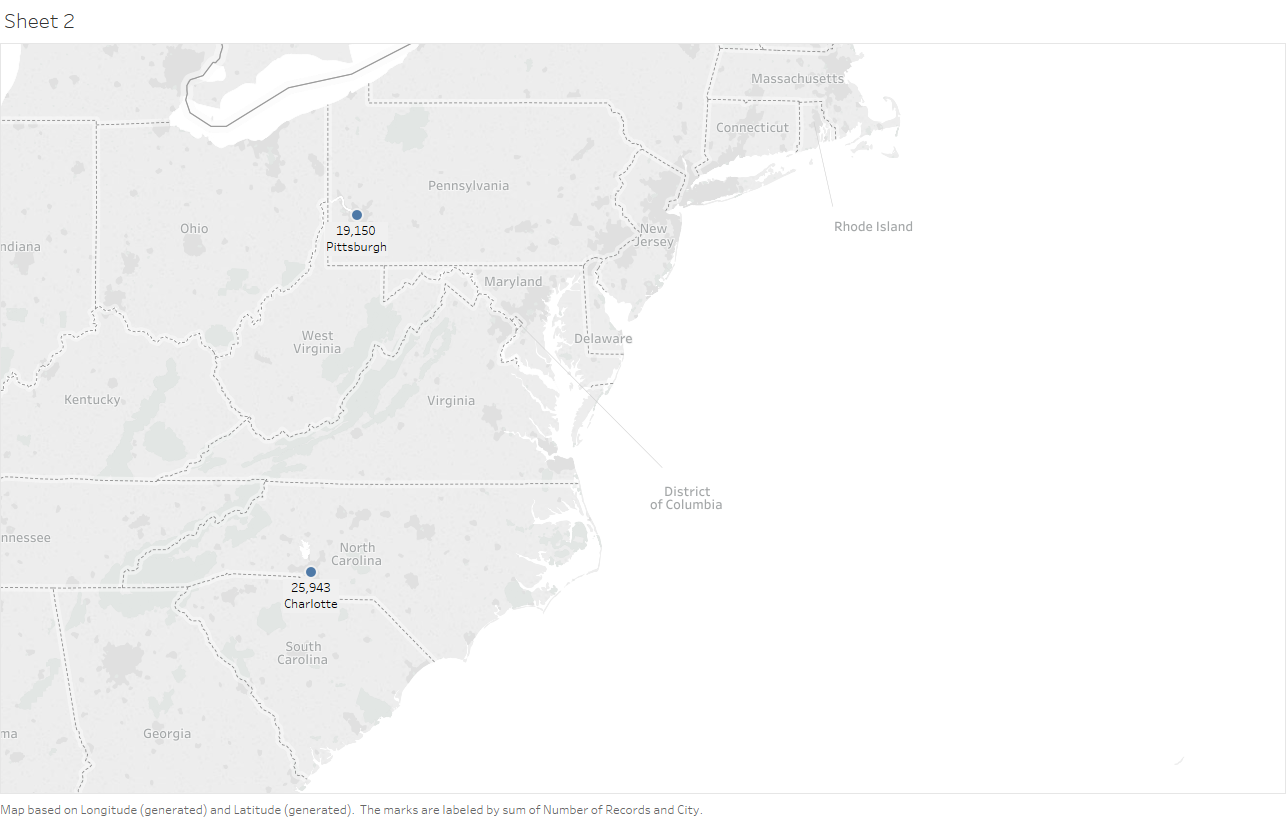
To find the popularity of each city, I read the csv file and separated each category and counted each category using Tableau filtered by city.





Popularity of each city 

From the above map, we can infer that Charlotte is more popular in most of the top 5 categories. While Pittsburgh is more popular in Nightlife.



As an interesting fact we can infer from the above image that Charlotte is more popular city in overall categories.

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

Text processing is a fundamental element of creation or manipulation of text. Latent Dirichlet

allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a

small number of topics and that each word's creation is attributable to one of the document's

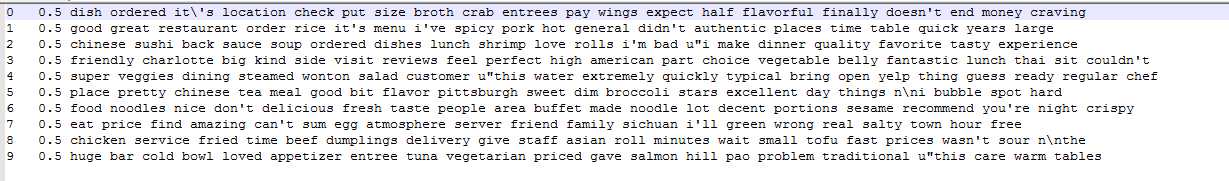
topics.

* What words are most frequently used to describe Chinese restaurants? To identify a restaurant’s type, refer to a business’s meta info in yelp\_academic\_dataset\_business.json
* What are the major themes/topics in the reviews of Chinese restaurants?

**Answer:** To find the Chinese restaurants I used the python script used in task 1 to get the business\_id of Chinese restaurants by searching in the categories list and saved all these business\_id in csv. Then I read this csv and extracted reviews for all these Chinese restaurants from review json and saved them in a csv file.

I have converted csv to text and then to mallet and passed to LDA for text processing. The command I used is:

bin\mallet import-dir --input (input file location) --output chinese\_review.mallet --keep-sequence --remove-stopwords



Then I created topic model.

Major theme topics are:

* Time
* Ordered
* Lunch
* Good
* Chinese