**Capstone Project 2: Milestone Report 1**

**Problem:**

As people travel from one city to another, they want to be able to experience aspects, cultures and peculiarities of that city. One of such distinctive features is the gastronomy, restaurants or eateries of that city. Being new to the city the average traveler will not know where to eat. As such his or her go-to will be Google or Yelp. Yelp would provide a list of restaurants closest to the where the user of the app is based on his GPS coordinates and not necessarily a list of restaurants that are popular in the city and that others just like the user would recommend. The objective here is to build a recommender system that takes eateries and restaurants within a city and recommends them based on how popular they are and also how likely they would be recommended to a user by other users of similar taste.

**Client:**

Business travelers, vacationers and yelp users who are foodies and like to travel. Restaurants will benefit from this as they get to be recommended not based on their distance from a user but rather their popularity and the likelihood that some other user similar in taste to the visiting user would recommend them.

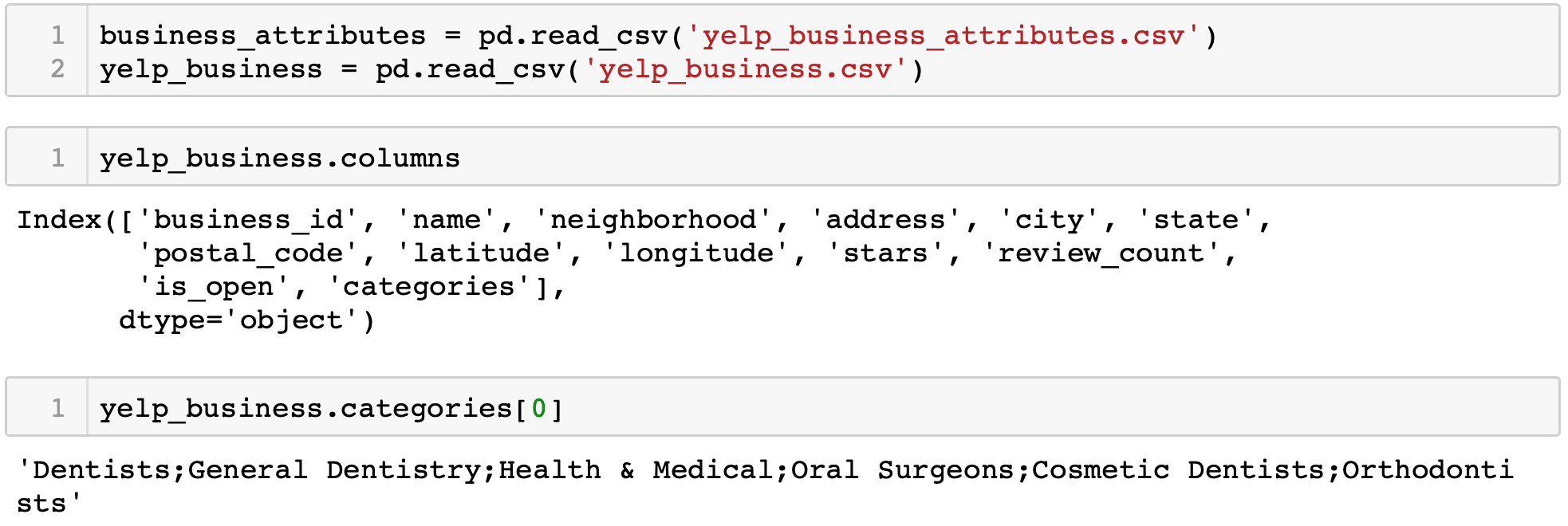
**Data Set:**

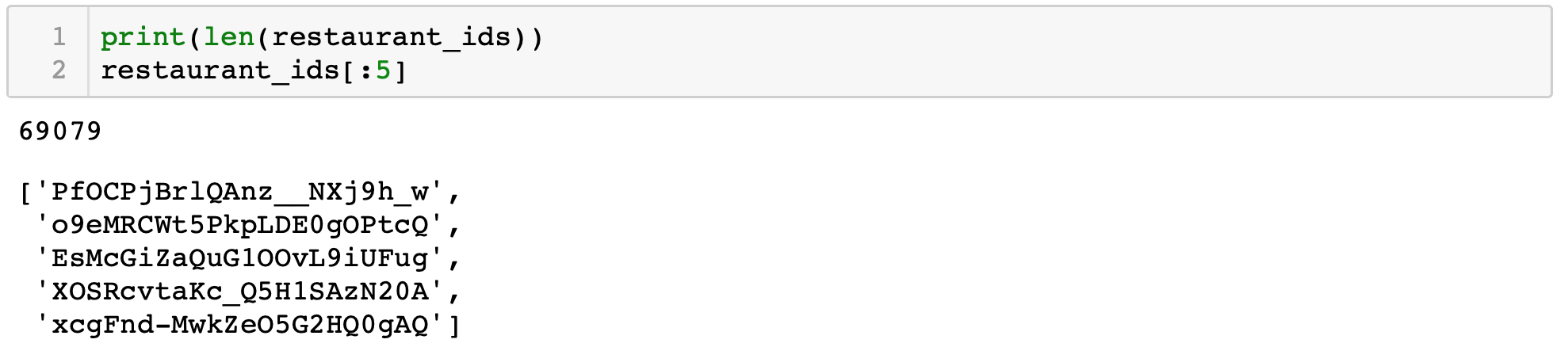
Yelp data set which has 1.637M users with 6.685M reviews for 192,609 businesses of all types in 10 metropolitan areas.

**Approach:**

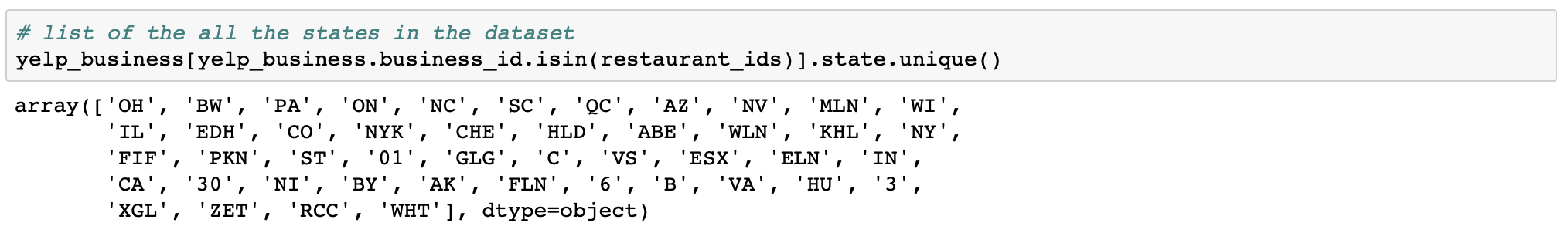
The dataset is made up of 4 different files yelp\_business.csv, yelp\_business\_attributes.csv, yelp\_user.csv and yelp\_review.csv. Used pandas to load the data from these files with its chunk feature coming handy for the very large files. Used the nltk library to determine from a string in the categories column of the yelp data set, which businesses were restaurants or eateries. And combined all the relevant data into a single table. Wrangled the data to eliminate missing values and also did some EDA on the ratings distribution of the users. The recommender is collaborative and the data set was narrowed down to restaurants in cities of Illinois found in the data set. Please note the reference to restaurants also refers to eateries.

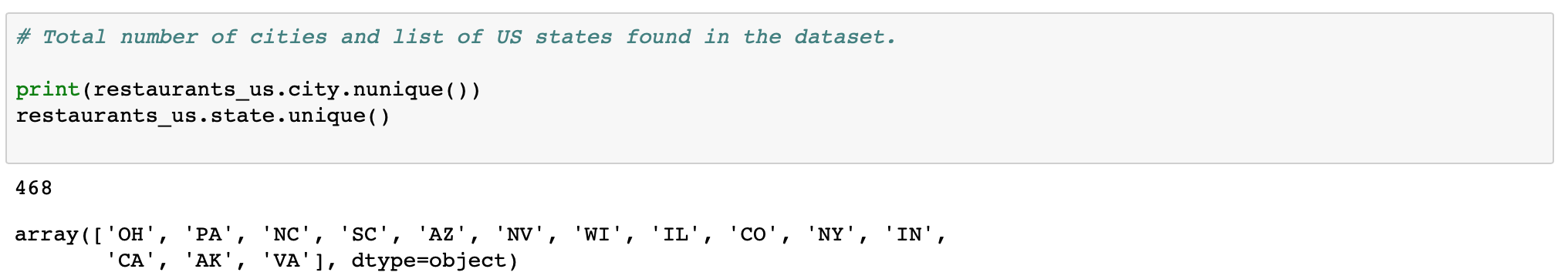
**Data Wrangling**

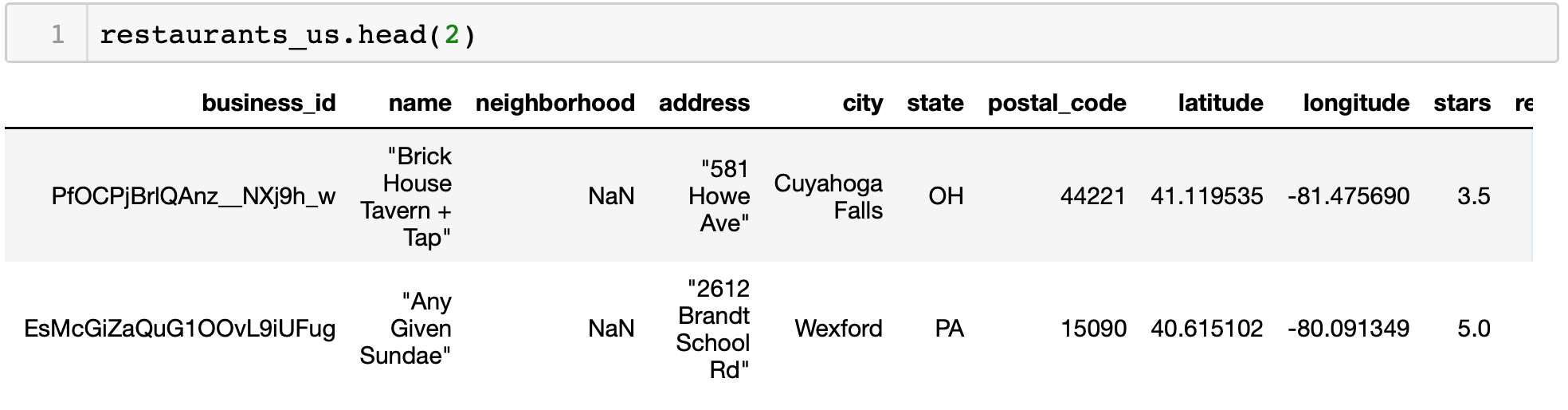
* Loading files in Pandas**:** 
  + yelp\_business.csv and yelp\_business\_attributes.csv files were loaded and the nltk library used on the “categories” column of the yelp\_business file to determine which businesses were restaurants or eateries and combined all the restaurant business ids in a list.

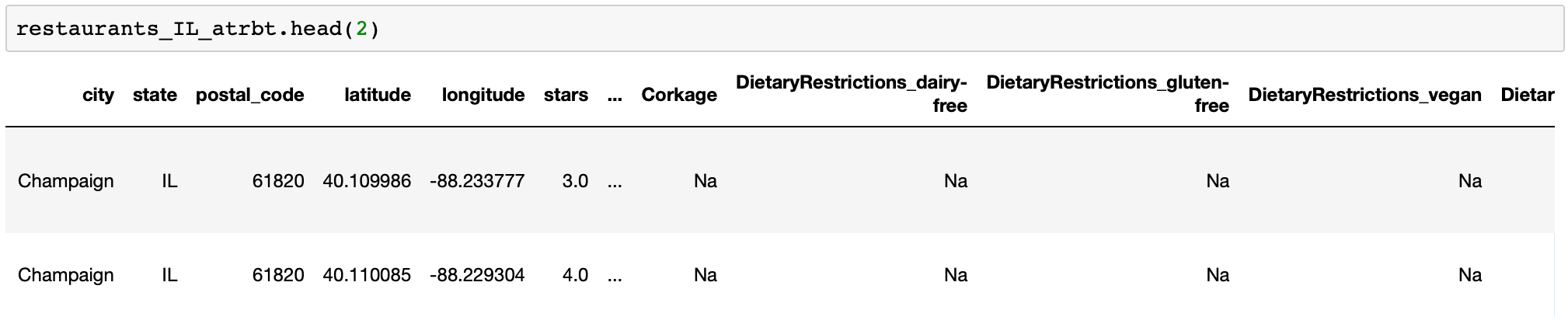


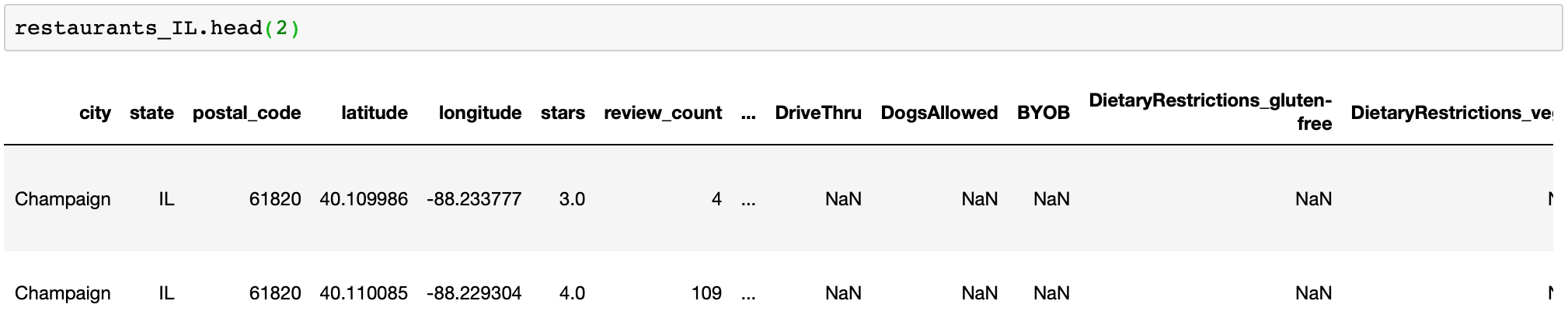
* + Examining the states found in the dataset



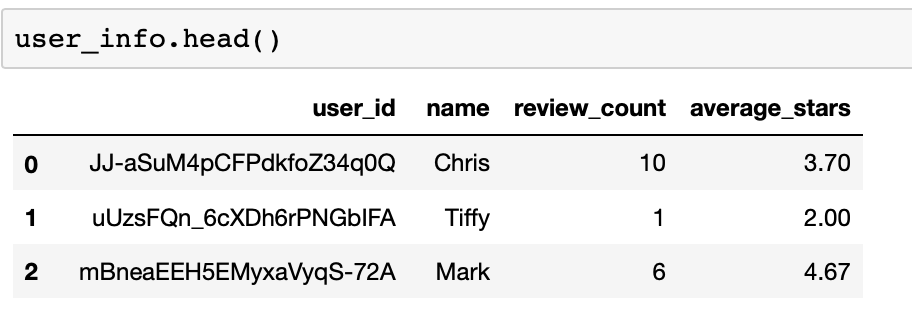
* + Loaded a txt file with US states name codes and combined with the list of restaurants ids to pull data from the yelp\_business data set of only restaurants in the US.

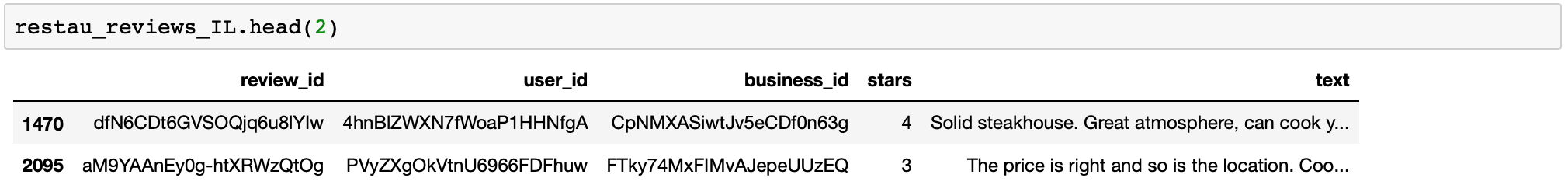


* + Pull from the dataset only restaurants found in Illinois and merge with the attributes file to add attributes to restaurants in Illinois. Change all Na entries to NaN entries.

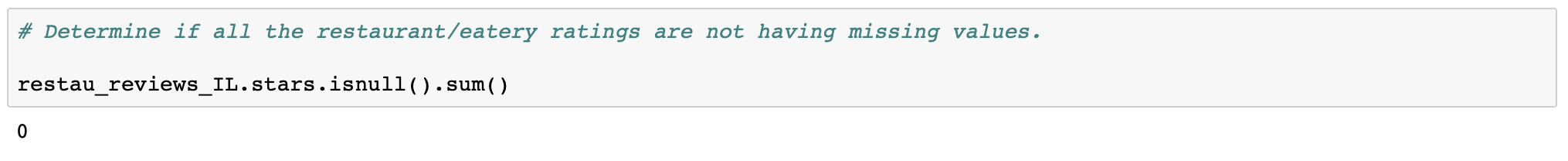


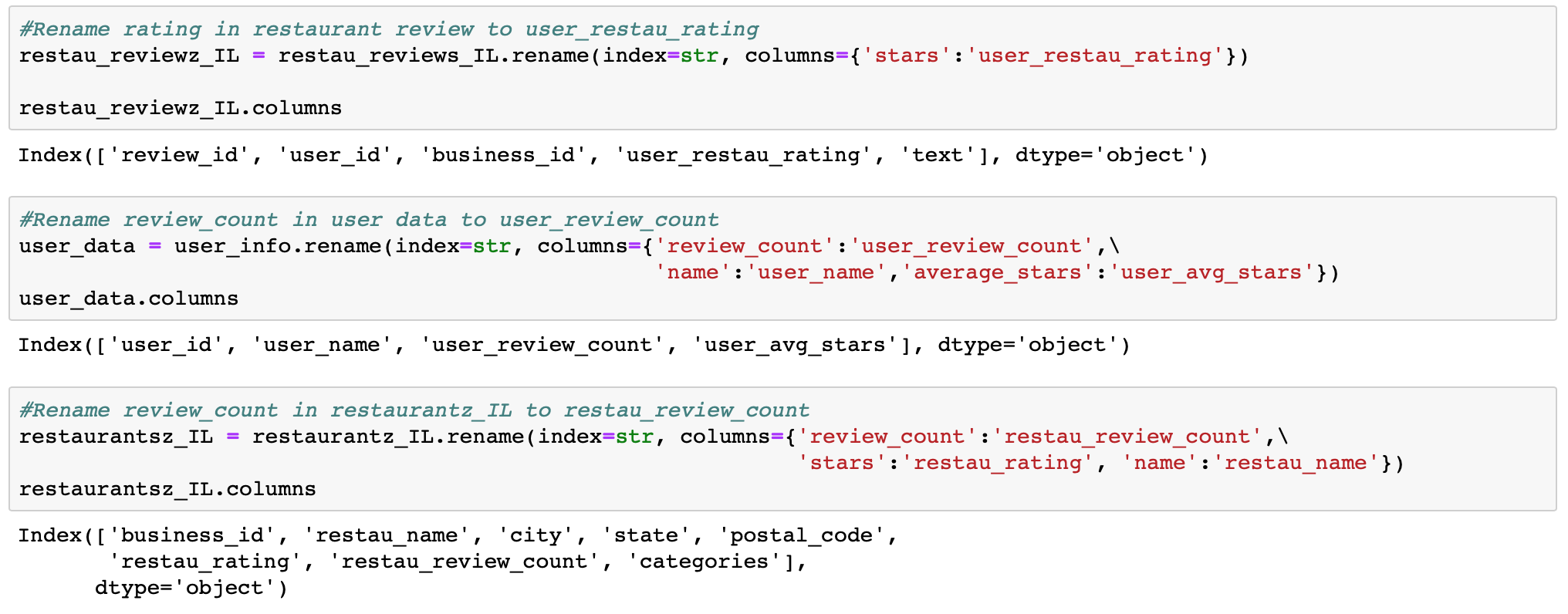
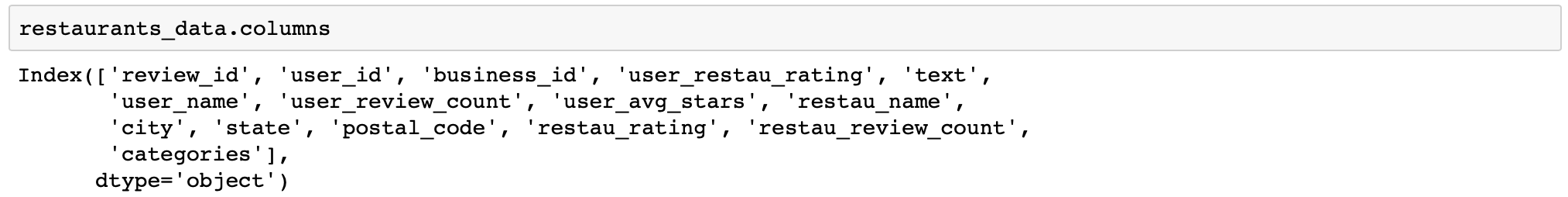
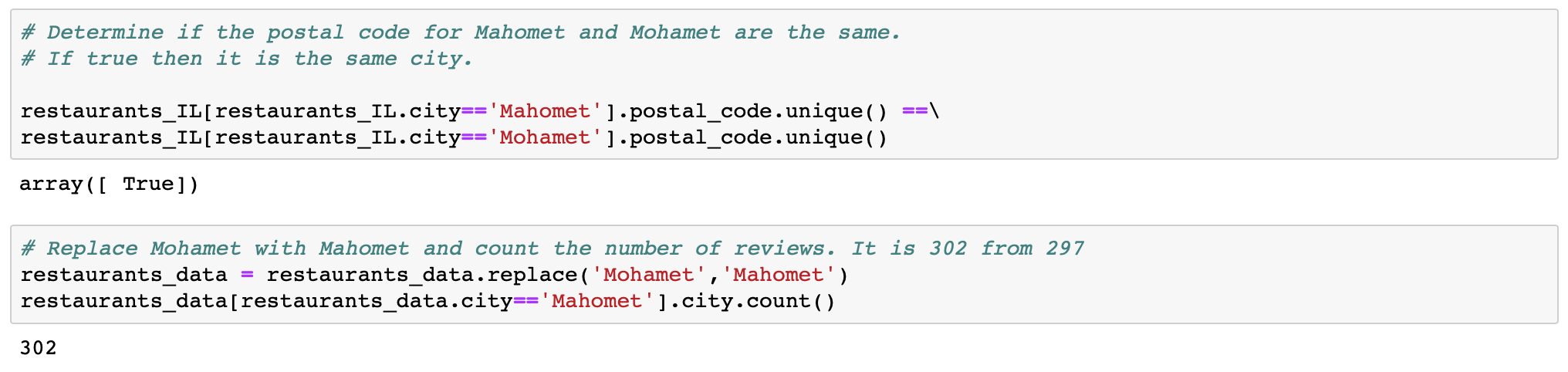
* + Load relevant user information from the 1.6GB yelp\_user.csv file in chunks of 10000 rows and also load from the 3.79GB yelp\_reviews.csv file reviews of users for restaurants in IL only.



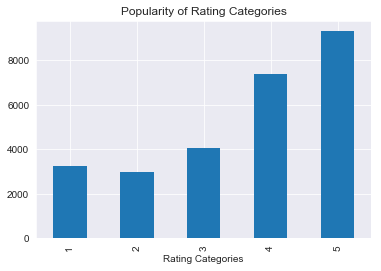


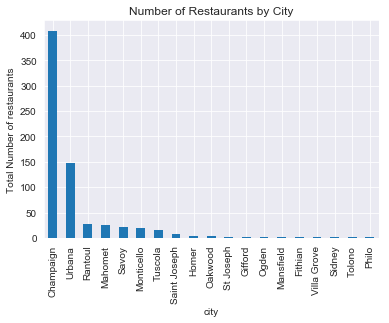
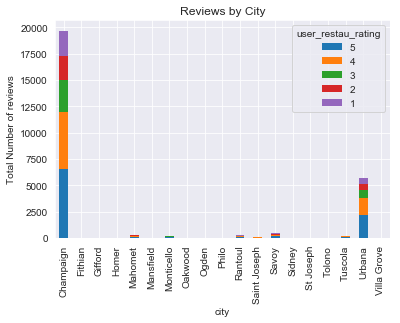
* + Check for any missing ratings.

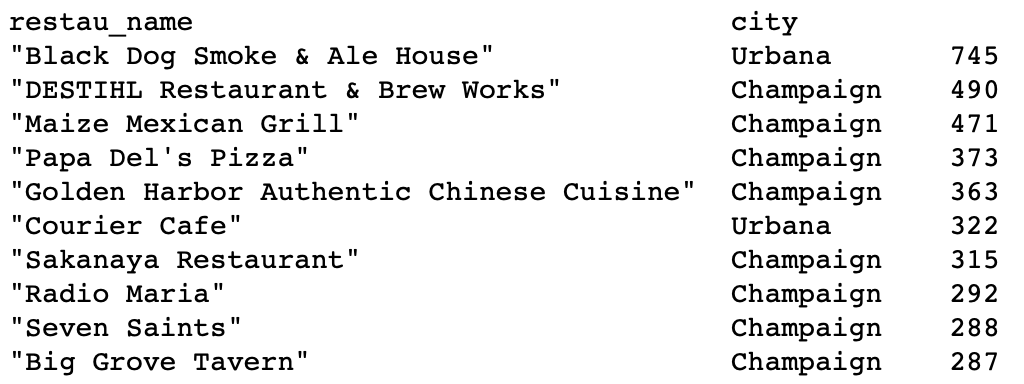


* + Before merging the files, rename columns having the same name between the files
  + Merge the three files into a single data frame restaurants\_data with column names thus:
  + Check City column and correct some misspelled city entries from Mohamet to Mahomet. Fill up missing postal codes and save data frame to restaurants\_all\_info\_IL.csv

**Data Story Telling:**

* + The most popular of the ratings given by the users for the different restaurants is 5 while the least popular is 2.
  + The highest number of the reviews were given for restaurants in the city of Champaign followed by the city of Urbana. This is primarily because most of the restaurants in the data set are found in Champaign and Urbana with 408 and 148 restaurants respectively.

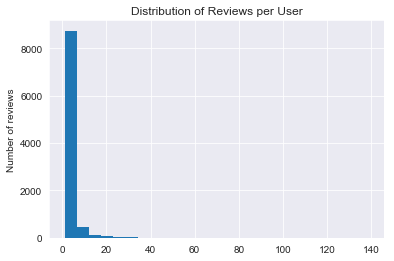
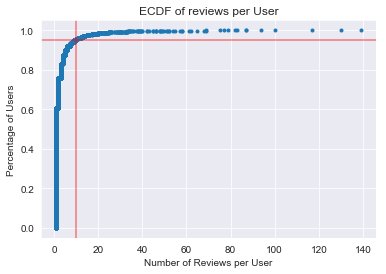


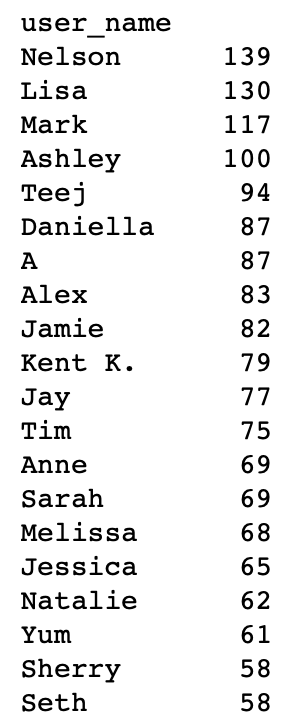
* + The top 10 restaurants by number of reviews are found in Champaign and Urbana as shown below:

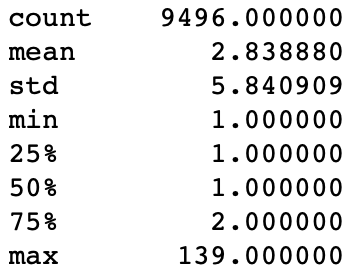
**Exploratory Data Analysis :**

The dataset used is made up of ratings from eateries in 21 cities IL extracted from the yelp dataset. The former being a combination of user information, business attributes, business reviews and user reviews from 4 different files. The objective is to recommend to users, restaurants in the city that other users similar to them like and would recommend.

The main variables considered in the recommendation are the ratings given by the users.

The number of reviews by each user varies by a huge margin with about 95% of the users having about 10 or less reviews while 5% has between 10 and 140 reviews. This gives a right skewed distribution of number of reviews per user with a mean of approximate 3.0 per user and median of 1.0 as shown below:





A distribution of the ratings is left skewed with a majority of the ratings being 3 to 5 and a mean of 3.61 and median of 4.0.

