**Data Description:**

There are two csv files. These files contain Yelp and Airbnb data from zip codes in NYC. Please find a description of each file below:

**panel.csv:**

This csv file has three columns. The first column has the zip code, the second column has the year, and the third column has a binary (1/0) variable called airbnb. The value of the airbnb variable is equal to 1 if there is significant Airbnb activity in that zip code during the year of the row. Otherwise it is equal to 0.

**yelp.csv**

This csv file has a collection of Yelp restaurant reviews written for restaurants in the zip codes from the panel.csv file. Here is the description of the columns:

|  |  |
| --- | --- |
| review\_date | The date that a restaurant review was written. |
| address\_zipcode | The zip code that the review was written for. |
| review\_text | The text of the review. |
| review\_rating | The rating given to the restaurant by the reviewer. |
| price | The relative price of the restaurant. The possible values in this column are: $ , $$ , $$$ , $$$$, and “UNCLAIMED RESTAURANT”. The dollar signs indicate how expensive the restaurant is with more dollar sign indicating more expensive restaurants. The “UNCLAIMED RESTAURANT” value implies that the restaurant was not claimed and therefore the price point is not verified. |
| NYC\_reviewer | This is a binary variable (1/0) that indicates whether the reviewer that wrote the restaurant review is from New York City. 1 means that review is written by a resident of NYC and 0 means that it was written by a visitor to NYC. |

* 1. Added a column to the data in panel.csv. This column contains the number of Yelp restaurant reviews that occurred in a zip code during the year in question. Named this column **yelp\_count**.
  2. Added a column to the data in panel.csv that contains the average rating of Yelp restaurant reviews that occurred in a zip code during the year. Namede this column **yelp\_average**.
  3. Added a column to the data in panel.csv. This column contains the number of Yelp restaurant reviews **written by NYC residents** that occurred in a zip code during the year in question. Named this column **yelp\_NYC\_count**.
  4. Added a column to the data in panel.csv that contains the average rating of Yelp restaurant reviews **written by NYC residents** that occurred in a zip code during the year. Named this column **yelp\_NYC\_average**.
  5. Wrote the new Panel data, with all the added columns and the original columns, to a csv file called panel\_output.csv.

1. Difference in ratings by visitors and NYC residents.
   1. Determined whether there is a statistically significant difference between the average ratings of the reviews written by residents of NYC and those written by visitors to NYC. In this part, ignored other factors that may impact average rating, only whether the difference between the visitors and residents is statistically different.
   2. Designed a statistical model that will estimate the change in average rating if a reviewer is a visitor vs. if the reviewer is a resident of NYC. In this model, controlled for other factors that could impact the rating. Answered: Are reviewers giving higher/lower ratings in one year relative to another? Do reviewers give higher/lower ratings to more/less expensive restaurants? Provided succinct rational for the controls included in the statistical model.
2. Using review text to understand reviewer location: 
   1. In this question, the goal is to determine whether there is a difference in the characteristics of restaurants that the NYC locals are interested in and the characteristics that the visitors to NYC are interest in. Here are the steps:
      1. Used the Microsoft Azure key phrase extraction API to identify the key phrases associated with each review in the review text in the yelp.csv file.
      2. Compared the key phrases associated with the reviews written by the visitors and the key phrases associated with the reviews written by the NYC locals. Based on these, determined the phrases that are high frequency phrases and examine the distribution of the phrases with a focus on the cases where the words are very high frequencies for visitors or for locals but not for both. . Clearly stated rational for the approach adopted.
3. Airbnb and tourist destination.
   1. Identified 5 tourist destinations in NYC and used the google distance API to determine the distance between these locations and the 10 zip codes in the panel data. Analyzed whether distance to these tourist locations is a significant factor in determining whether a zip code eventually contains significant Airbnb activity.
   2. Used a secondary data source to provide further evidence and details regarding the reasons why Airbnb becomes more popular in one neighborhood and not another neighborhood.