MGMT 590

Web Data Analytics

Assignment 2

Total Points: 60

**Ground Rules:**

*For all questions, you must submit the code that you used and the code must be clearly commented. If the code is confusing, specifically with regard to which question it is referring to, the code will not be read and you will not be awarded the points. If the code does not run, for whatever reason, you will be deducted points. The same applies to statistical analysis. Be clear as to which question you are answering in your analysis.*

*You may work with* ***ONE*** *partner on this assignment. Only one member should submit the assignment but should clearly mention who they are working with.*

**Data Description:**

There are two csv files that you have access to. 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. |

**Instructions:**

1. Organize the Panel Data **(5 points)**
   1. Add a column to the data in panel.csv. This column will contain the number of Yelp restaurant reviews that occurred in a zip code during the year in question. Name this column **yelp\_count**.
   2. Add 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. Name this column **yelp\_average**.
   3. Add a column to the data in panel.csv. This column will contain the number of Yelp restaurant reviews **written by NYC residents** that occurred in a zip code during the year in question. Name this column **yelp\_NYC\_count**.
   4. Add 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. Name this column **yelp\_NYC\_average**.
   5. Write the new Panel data, with all the added columns and the original columns, to a csv file called panel\_output.csv.
2. Difference in ratings by visitors and NYC residents. **(12.5 points)**
   1. Determine 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, you don’t need to consider other factors that may impact average rating, only whether the difference between the visitors and residents is statistically different.
   2. Design 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, you will need to control for other factors that could impact the rating. You must think about the following items specifically: Are reviewers giving higher/lower ratings in one year relative to another? Do reviewers give higher/lower ratings to more/less expensive restaurants? Provide succinct rational for the controls included in the statistical model.
3. Using review text to understand reviewer location:  **(17.5 points)**
   1. In this question, your 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. Use 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. **Make sure to go through the Azure text analytics documentation to understand how best to go about sending requests so as not to deplete your free funding.** If you had chosen westus as your location, you would use the following endpoint: https://westus.api.cognitive.microsoft.com/text/analytics/v2.0/keyPhrases
      2. Compare 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, determine 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. You will be graded based upon the rigor and methodology with which you use to determine the characteristics, based on the reviews that are interesting to visitors vs. those that are interesting to NYC locals. Clearly state rational for the approach(es) you adopt.
4. Airbnb and tourist destination. **(25 points)**
   1. Identify 5 tourist destinations in NYC and use the google distance API to determine the distance between these locations and the 10 zip codes in the panel data. Analyze whether distance to these tourist locations is a significant factor in determining whether a zip code eventually contains significant Airbnb activity. This does not have to be done through statistical methods. However, you should use visuals to argue for or against the importance of distance to these tourist attractions in driving Airbnb demand in a zip code.
   2. Use a secondary data source of your choosing to provide further evidence and details regarding the reasons why Airbnb becomes more popular in one neighborhood and not another neighborhood. You will be graded on you rigor and analysis.