Case Study 1

BSDS Spring 2021

Instructions

- You must work in your assigned group on Monday, 2/22 in a breakout room during the live lecture (unless you have made prior arrangements).
- On Friday, 2/26 each group must present one or two of the open-ended problems in the assignment.
- Each group will submit a Jupyter Notebook file (.ipynb) and an associated .html file via Canvas. Each group member will submit the same two files, except for the "equal work pledge".
- The submitted notebook must start with a Markdown cell header indicating the Case Study number, dataset, group members, and any external sources. Everyone will additionally include an "equal work pledge" which states that they understand all code/answers in the assignment and that all group members contributing equally. If this is not the case please indicate.
- Each question must be annotated appropriately with Markdown cells. The Notebook file should written in a way that a third party with no knowledge of the questions can read it.
- All group members will receive the same grade unless the "equal work pledge" is violated.

Import the `nycAirBNB19.csv` dataset available on Canvas. Here is a description for each variable:

- `id`: The listing ID
- `name`: name of the listing
- `host_id`: The ID of the host
- `host_name`: The name of the host
- `neighbourhood_group`: Which of the five boroughs the listing is located in
- `neighbourhood`: Which neighbourhood the listing is located in
- `latitude`: The latitude of the listing
- `longitude`: The longitude of the listing
- `room_type`: The room type of the listing
- `price`: The price per night in dollars
- `minimum_nights`: The minimum number of nights one can book

- `number_of_reviews`: The number of reviews for this listing
- `last_review`: the date of the latest review
- `reviews_per_month`: The number of reviews per month
- `calculated_host_listings_count`: The number of listings associated with this host
- `availability_365`: Number of days when listing is available for booking

1. Create the following visualizations.

a. Make a scatter plot of `price` vs. `number_of_reviews`.

b. Create a histogram of `availability_365`.

2. Create a facet grid of scatter plots from 1a with the categorical variables `neighbourhood_group` and `room_type`. What do you observe?

3. Create a dataset of listings with `price` strictly less than $100 in the `neighbourhood_group` Brooklyn.

4. List the mean `price` for each `neighbourhood`.

5. Create a dataset of just private room listings. Add a column for total minimum booking cost given by the `price` times the `minimum_nights`. Select the observations where the minimum booking cost is less than $1000 and create a scatter plot of minimum booking cost vs. `price` for these observations.

Note: In the following questions (6. - 10.) include summary statistics and/or visualizations to support your claims!

6. Create a scatter plot of `latitude` vs. `longitude`. Use different aesthetics to visualize `neighbourhood_group` and `price` on the same plot.

7. Count the number of missing values for each variable. Why do you think these variables have missing values?

Open-ended Questions

8. Devise a metric for ranking `neighbourhoods` by which are the most lucrative to be an AirBnB host in (don't just use `price`, think about frequency of visits, competition, etc.). Based on this metric, which are the worst and best `neighbourhoods` to be a host? Also using this metric, rank a different categorical variable.

9. Compare/contrast the different `room_types` using both numeric and categorical variables.

10. Develop and answer you own hypothesis; get creative!