**Name: Daniyal Mufti**

**FE 541 Project Proposal**

**New York City Airbnb Open Data**

**Part 1:**

As part of my semester project, I will analyze New York City's Airbnb calendar year 2019 data and try to draw out some interesting statistical inferences and insights using the techniques learnt in this course. I pulled this project and data from the Kaggle competition website; After the completion of this semester, I plan on publishing the project on the website so it should be an exciting project to work on.

The three main ideas I will focus on will be: -

- Forming some hypothesizes in the data and then draw some statistical inferences from the data.
- Drawing out some meaningful statistics from the data which gives us some insights into the nature of the Airbnb business and its operation with NYC.
- Visualizing the data for insightful data storytelling.

The questions I might ask would be things like: -

- What can we learn about different hosts and areas?
    - Pricing differences by areas
    - Differences in pricing and housing availability based on host gender.
- What can we learn from different predictions?
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

These are just some examples of the questions I will look to answer. As I analyze the data in more depth, I should be able to draw out more interesting questions which will hopefully give me more interesting insights.

**Part 2:**

The data source is: -

New York City Airbnb Open Data | Kaggle

The data comes in a csv formatted file with 48,895 rows and 16 columns of data.

First 5 rows of data snipped from the R console: -

```
    id                                              name host_id   host_name neighbourhood_group neighbourhood latitude
1 2539                Clean & quiet apt home by the park    2787        John            Brooklyn    Kensington 40.64749
2 2595                             Skylit Midtown Castle    2845    Jennifer           Manhattan       Midtown 40.75362
3 3647               THE VILLAGE OF HARLEM....NEW YORK !    4632   Elisabeth           Manhattan        Harlem 40.80902
4 3831                    Cozy Entire Floor of Brownstone    4869 LisaRoxanne            Brooklyn  Clinton Hill 40.68514
5 5022 Entire Apt: Spacious Studio/Loft by central park    7192       Laura           Manhattan   East Harlem 40.79851
  longitude       room_type price minimum_nights number_of_reviews last_review reviews_per_month
1 -73.97237    Private room   149              1                 9  2018-10-19              0.21
2 -73.98377 Entire home/apt   225              1                45  2019-05-21              0.38
3 -73.94190    Private room   150              3                 0                          NA
4 -73.95976 Entire home/apt    89              1               270  2019-07-05              4.64
5 -73.94399 Entire home/apt    80             10                 9  2018-11-19              0.10
  calculated_host_listings_count availability_365
1                              6              365
2                              2              355
3                              1              365
4                              1              194
5                              1                0
```

The descriptions of each of the labels are as follows: -

1. id: This is the listing ID.
2. name: This is the name/description of the listing.
3. host_id: This is the host ID.
4. host_name: This is the name of the host.
5. neighbourhood_group: Borough location withing New York City.
6. neighbourhood: Neighborhood location within New York City.
7. latitude: Latitude coordinates.
8. longitude: Longitude coordinates.
9. room_type: Listing space type.
10. price: The price in dollars of listing.
11. minimum_nights: The minimum amounts of nights that need to be stayed for the listing.
12. number_of_reviews: The number of reviews for the host.
13. last_review: The latest review date for the host.
14. reviews_per_month: The number of reviews per month for the host.
15. calculated_host_listings_count: Number of listings per host.
16. availability_365: Number of days when the listing is available for booking.

**Part 3:**

*Step 1: -*

I will begin by cleaning and imputing the data as needed to makes sure there are no data quality issues that would hamper the analysis.

*Step 2: -*

I will create some basic summary statistics to help get some intermediate insights into the data which should help me get more clarity into what kind of different questions I can ask of the data. I will then see if I can create additional columns of information that would help aid the overall analysis.

*Step 3: -*

As part of my analysis, I will try to visualize the data to help create a more visually pleasing data story for the audience. There is good geographic data available to use in the dataset so I will use that to visualize the data.

*Step 4: -*

After visualizing the data, I plan to create the relevant statistics and metrics to conduct statistical inference testing.

**Conclusion:**

All of the above steps are subject to amendment as I learn more during this course and understand the data better. Hopefully by the end of the analysis, I would have been able to answer meaningful questions about the Airbnb business and operation trends within New York City.