ML Group project:

Members:

- Mark Ditchburn
- Andrea Butera
- Kuno de Leeuw-Kent

Dataset:

The dataset selected is data on all listing activity and metrics in NYC, NY for Airbnb captured for the year 2019.

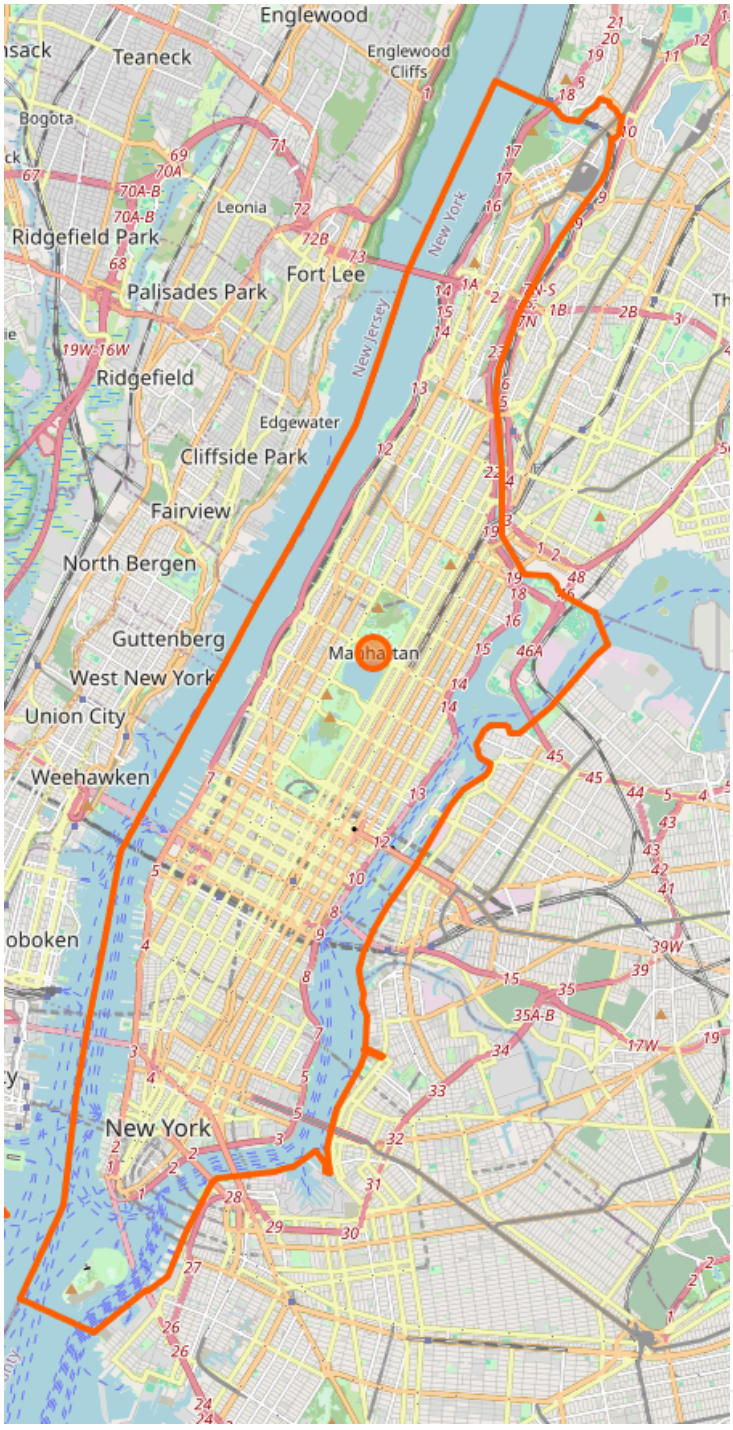
The dataset can be found here, with the provider of the data here.

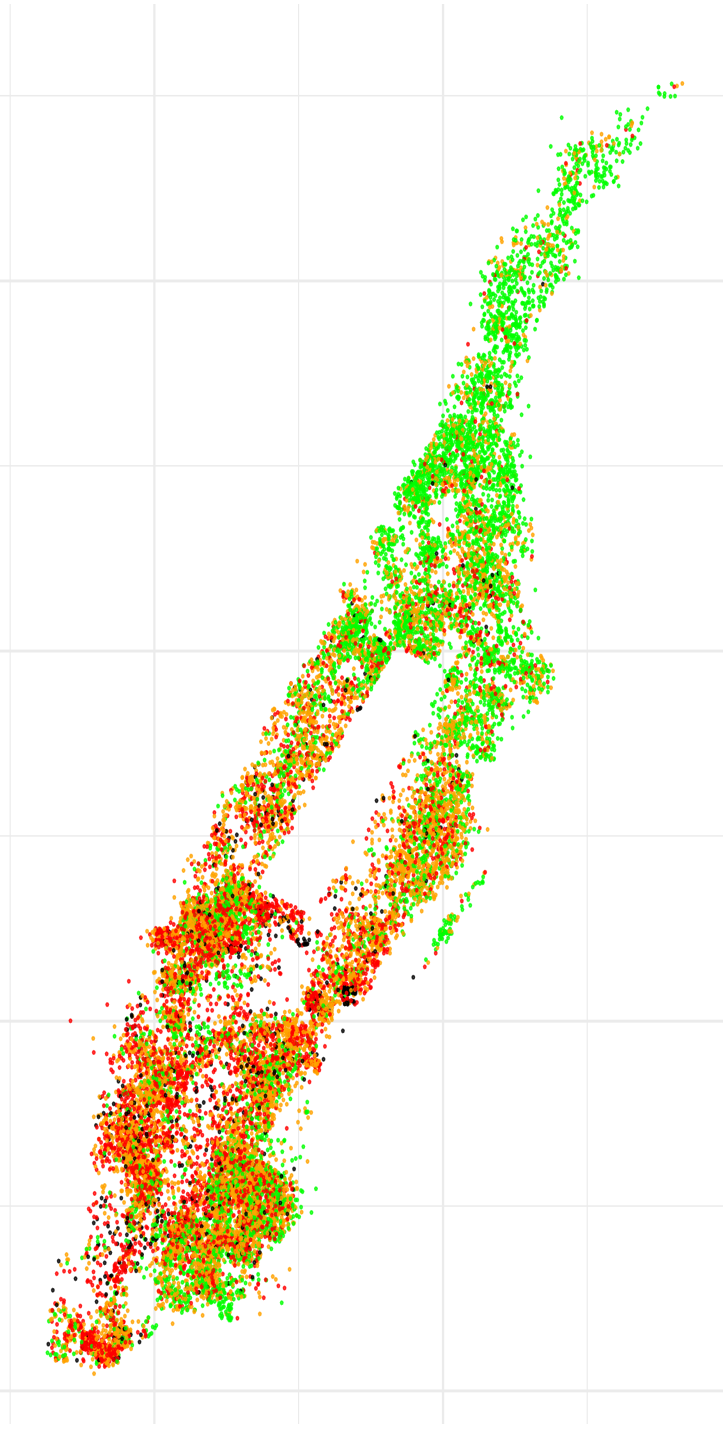
Data Summary:

```
'data.frame':   48895 obs. of  16 variables:
 $ id                            : int  2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...
 $ name                          : chr  "Clean & quiet apt home by the park" "Skylit Midtown Cas
 $ host_id                       : int  2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...
 $ host_name                     : chr  "John" "Jennifer" "Elisabeth" "LisaRoxanne" ...
 $ neighbourhood_group           : chr  "Brooklyn" "Manhattan" "Manhattan" "Brooklyn" ...
 $ neighbourhood                 : chr  "Kensington" "Midtown" "Harlem" "Clinton Hill" ...
 $ latitude                      : num  40.6 40.8 40.8 40.7 40.8 ...
 $ longitude                     : num  -74 -74 -73.9 -74 -73.9 ...
 $ room_type                     : chr  "Private room" "Entire home/apt" "Private room" "Entire
 $ price                         : int  149 225 150 89 80 200 60 79 79 150 ...
 $ minimum_nights                : int  1 1 3 1 10 3 45 2 2 1 ...
 $ number_of_reviews             : int  9 45 0 270 9 74 49 430 118 160 ...
 $ last_review                   : chr  "2018-10-19" "2019-05-21" "" "2019-07-05" ...
 $ reviews_per_month             : num  0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...
 $ calculated_host_listings_count: int  6 2 1 1 1 1 1 1 1 4 ...
 $ availability_365              : int  365 355 365 194 0 129 0 220 0 188 ...
```

Data Visualisation:

Here is a sample of the data filtered by Manhattan plotted base on location and coloured by price. As you can see the data lines up with the map correctly and is very dense. This is most likely due to the dense population and the high amount of tourism in Manhattan and new york as a whole





Model Aims:

For our group project we want to create a prediction model that will predict the price of a give property. These will be regression models and we will explore 3 different approaches and evaluate the accuracy of each.

To ensure each model is comparable we will be using the same seed for separating our data into training and testing subsets with 70% our data used for training and the remaining 30% for evaluating the model.

Here is the script we use to load and separate our data, we use random separation to protect against there being changes in the data overtime:

```
# Import CSV file
AirbnbData <- read.csv("DataSet/AB_NYC_2019.csv")

# Create Train and Test subsets
set.seed(25)
train_indices <- sample(seq_len(nrow(AirbnbData)), size = 0.7 * nrow(AirbnbData))

training <- AirbnbData[train_indices, ]
testing <- AirbnbData[-train_indices, ]
```