DATA583 Project Proposal

Madison Greenough & Jonah Edmundson

1 Statistical Description

The dataset contains 800 real estate properties for sale in New York City. There are 10 variables for each listing, including the price of the listing (integer in $USD), number of bedrooms (numeric, int), number of bathrooms (numeric), square feet (numeric), address (string), flag of whether it is new or not (binary 0 or 1), the listing company name (string), latitude (numeric), longitude (numeric), and distance to Central Park (numeric).

There are missing data points throughout the dataset, given the nature of how the data comes from many different listing agencies which have different protocols. Very few, <5% of listings were missing bedroom or bathroom details. Approximately 25% of our dataset has missing latitude or longitude values, which in turn results in 25% of distance measures, as it is calculated from those values. Also, square footage is missing in approximately 35-30% of the dataset.

Below, we can see a list of all of the variables, as well as a descriptive summary for each.

```
> df = read.csv('../data/NY_realestate2023-02-17-cleaned.csv')
> df$price = as.numeric(df$price)
> df$bed = as.numeric(df$bed)
> df$bath = as.numeric(df$bath)
> names(df)

 [1] "price"     "bed"       "bath"      "feet"      "address"   "new_flag"
 [7] "company"   "latitude"  "longitude" "distance"

> summary(df[c(1,2,3,4,6,8,9,10)])

     price               bed              bath             feet
 Min.   :  132500   Min.   : 0.000   Min.   : 1.000   Min.   :     1.0
 1st Qu.:  561250   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:   863.8
 Median :  875000   Median : 2.000   Median : 2.000   Median :  1285.0
 Mean   : 1606935   Mean   : 2.912   Mean   : 2.399   Mean   :  3087.6
 3rd Qu.: 1584750   3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.:  2368.5
 Max.   :46395000   Max.   :24.000   Max.   :16.000   Max.   :325000.0
                    NA's   :15       NA's   :27       NA's   :248
  new_flag       latitude       longitude         distance
 False: 71   Min.   :40.51   Min.   :-77.77   Min.   :  0.2015
 True :751   1st Qu.:40.65   1st Qu.:-73.99   1st Qu.:  4.2443
             Median :40.74   Median :-73.96   Median : 11.0958
             Mean   :40.73   Mean   :-73.95   Mean   : 13.1292
             3rd Qu.:40.77   3rd Qu.:-73.91   3rd Qu.: 18.2800
             Max.   :42.64   Max.   :-73.59   Max.   :346.8985
             NA's   :206     NA's   :206      NA's   :206

> nrow(df)

[1] 822
```

2 Data Collection

The data was collected by performing a web scrape on Trulia, which is a site listing all real estate properties for sale in New York City. The site lists a maximum of 20 pages, which is a total of 800 listings. The web scrape was performed on 02/17/2023, so it includes all listings posted on or before that date, up to the maximum of 800. When the time of listing is mentioned, in reference to whether it is flagged as new or not, this is in reference to the listing being posted within 7 days of the 02/17/2023 web scrape. When cleaning the scraped data, the columns were converted to numeric values, and new columns were created to indicate the latitude and longitude coordinates of the listing, based on the address location provided. Afterwards, another new column was added to calculate the distance to Central Park using the coordinates of the listing location. Central Park was chosen as the reference point as it is not only a desirable location to reside in, but it is also located centrally in NYC near expensive neighbourhoods, so it may be a factor to help indicate price of a listing. The web scrape is completed and the data is ready in advance of this proposal submission.

3 Scientific Questions

We will try to model price as a response variable using the given dependent variables we have in the dataset to hopefully create a model that can help predict prices of listings. This model can be a beneficial tool to listing companies when appraising a home to help select a list price, but it can also be used on the other side by buyers when determining whether a home is listed at a fair price or when determining an offer price. One potential modelling method is through clustering. We will look to find various combinations of latitude and longitude that give unusually high or low prices, relative to the dataset, and then we will map these locations onto different neighbourhoods in New York City.