Where Should New York Renters Look?

Housing Data Exercise

Colin Adams

09/25/22

1. Data Introduction

The data I have chosen follows listing prices of units in different New York City neighborhoods along with many features of the unit. The data falls under in the Housing market.

- Originally obtained the data from https://github.com/Codecademy/datasets/tree/master/streeteasy.
- Link to original dataset: https://drive.google.com/file/d/1JpquBHuVTaBsCM53XSGip51hSnvzQmSj/view
- Link to cleaned dataset: https://drive.google.com/file/d/1Z3zeL9CWym07VOg59Hue8iYX4BoHFBv_/view

This data is important for those who are looking to rent/purchase units in New York, to investigate what they should be looking for. In addition, it is important for those moving to New York to find out which area suits them the best.

Variable Dictionary:

- **rental_id**: Rental ID
- **building_id**: Building ID
- **rent**: Cost of rent (in USD)
- **bedrooms**: Number of bedrooms
- **size_sqft**: Size of the rental listing in square-footage
- **min_to_subway**: Time it takes to get to the subway (in minutes)
- **floor**: The number of floors
- **building_age_yrs**: Age of the listing's building (in years)
- **no_fee**: Does it have a broker fee? ("1" = yes , "0" = no)
- **has_roofdeck**: Does it have a roof deck? ("1" = yes , "0" = no)
- **has_washer_dryer**: Does it have a washer/dryer in the unit? ("1" = yes , "0" = no)
- **has_doorman**: Does the building have a doorman? ("1" = yes , "0" = no)
- **has_elevator**: Does the building have an elevator? ("1" = yes , "0" = no)
- **has_dishwasher**: Does the listing come with a dishwasher? ("1" = yes , "0" = no)
- **has_patio**: Does the unit have a patio? ("1" = yes , "0" = no)
- **has_gym**: Does the building have a gym? ("1" = yes , "0" = no)
- **neighborhood**: The neighborhood where the unit is located.
- **submarket**: The submarket where the unit is located.
- **borough**: The borough where the unit is located.

Glancing at the dataset.

```
# A tibble: 6 x 20
  rental_id building_id  rent bedrooms bathrooms size_sqft min_to_subway floor
      <dbl>       <dbl> <dbl>    <dbl>     <dbl>     <dbl>         <dbl> <dbl>
1      1545    44518357  2550        0         1       480             9     2
2      2472    94441623 11500        2         2      2000             4     1
3     10234    87632265  3000        3         1      1000             4     1
4      2919    76909719  4500        1         1       916             2    51
5      2790    92953520  4795        1         1       975             3     8
6      2869     8967298  3600        3         2       900             4     1
# ... with 12 more variables: building_age_yrs <dbl>, no_fee <dbl>,
#   has_roofdeck <dbl>, has_washer_dryer <dbl>, has_doorman <dbl>,
#   has_elevator <dbl>, has_dishwasher <dbl>, has_patio <dbl>, has_gym <dbl>,
#   neighborhood <chr>, submarket <chr>, borough <chr>
```

2. Analysis & Discussion

According to Manhattan Miami Real Estate, location is most important when purchasing an apartment in NYC, and it tops amenities. While a great location is great, not many people are able to afford living in such an expensive city, and many people live in very tiny, closet-sized studio apartments.

Where should movers look for a roomy apartment?

2.1 Summary Statistics

Summary of each variable in the dataset.

3.1.1 Figure 1:

Table 1: Table continues below

rental_id	building_id	rent	bedrooms
Min. : 1	Min. : 7107	Min. : 1250	Min. :0.000
1st Qu.: 2700	1st Qu.:26998106	1st Qu.: 2750	1st Qu.:1.000
Median : 5456	Median :50698935	Median : 3600	Median :1.000
Mean : 5527	Mean :51220069	Mean : 4537	Mean :1.396
3rd Qu.: 8306	3rd Qu.:75720641	3rd Qu.: 5200	3rd Qu.:2.000
Max. :11349	Max. :99987207	Max. :20000	Max. :5.000

Table 2: Table continues below

bathrooms	size_sqft	min_to_subway	floor
Min. :0.000	Min. : 250.0	Min. : 0.000	Min. : 0.00
1st Qu.:1.000	1st Qu.: 633.0	1st Qu.: 2.000	1st Qu.: 3.00
Median :1.000	Median : 800.0	Median : 4.000	Median : 6.00
Mean :1.322	Mean : 920.1	Mean : 5.079	Mean :10.19
3rd Qu.:2.000	3rd Qu.:1094.0	3rd Qu.: 6.000	3rd Qu.:14.00
Max. :5.000	Max. :4800.0	Max. :51.000	Max. :83.00

Table 3: Table continues below

building_age_yrs	no_fee	has_roofdeck	has_washer_dryer
Min. : 0.00	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.: 12.00	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median : 44.00	Median :0.0000	Median :0.0000	Median :0.0000
Mean : 52.09	Mean :0.4296	Mean :0.1286	Mean :0.1338
3rd Qu.: 89.00	3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:0.0000
Max. :180.00	Max. :1.0000	Max. :1.0000	Max. :1.0000

Table 4: Table continues below

has_doorman	has_elevator	has_dishwasher	has_patio
Min. :0.000	Min. :0.00	Min. :0.0000	Min. :0.0000
1st Qu.:0.000	1st Qu.:0.00	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.000	Median :0.00	Median :0.0000	Median :0.0000
Mean :0.228	Mean :0.24	Mean :0.1556	Mean :0.0456
3rd Qu.:0.000	3rd Qu.:0.00	3rd Qu.:0.0000	3rd Qu.:0.0000
Max. :1.000	Max. :1.00	Max. :1.0000	Max. :1.0000

has_gym	neighborhood	submarket	borough
Min. :0.0000	Length:5000	Length:5000	Length:5000
1st Qu.:0.0000	Class :character	Class :character	Class :character
Median :0.0000	Mode :character	Mode :character	Mode :character
Mean :0.1438	NA	NA	NA
3rd Qu.:0.0000	NA	NA	NA
Max. :1.0000	NA	NA	NA

To start, let's create a new variable to find out which apartments utilize space the best for the price. A new variable will be created to determine the square-feet per dollar of each unit.

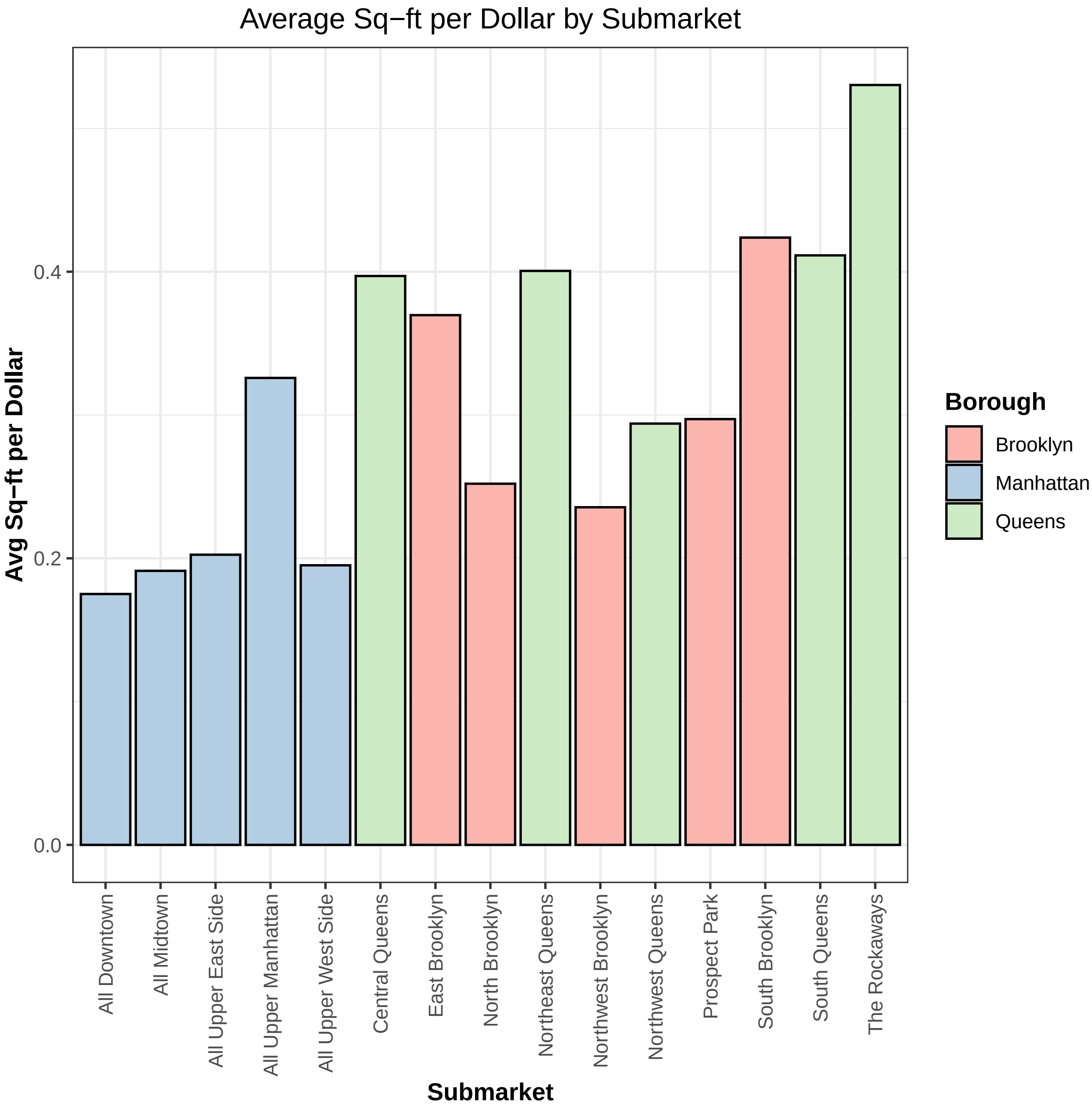
2.1.2 Figure 2 - Summary Statistics on Square-Feet per Dollar:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.04499	0.1714	0.203	0.2278	0.2574	0.8929

The table above shows us the statistics of the sqft per dollar of each listing unit. We can place these values into three different categories based off their values: "Expensive" (0) , "Average" (1) , "Bargain" (2). "Expensive" denotes a sqft per dollar value that is below the first quartile. "Average" denotes a sqft per dollar value between the first and third quartiles. "Bargain" denotes a sqft per dollar value that is greater than the third quartile, meaning that you are getting a lot of space for the price you are paying.

Now let's look at the average square-feet per dollar value of each location.

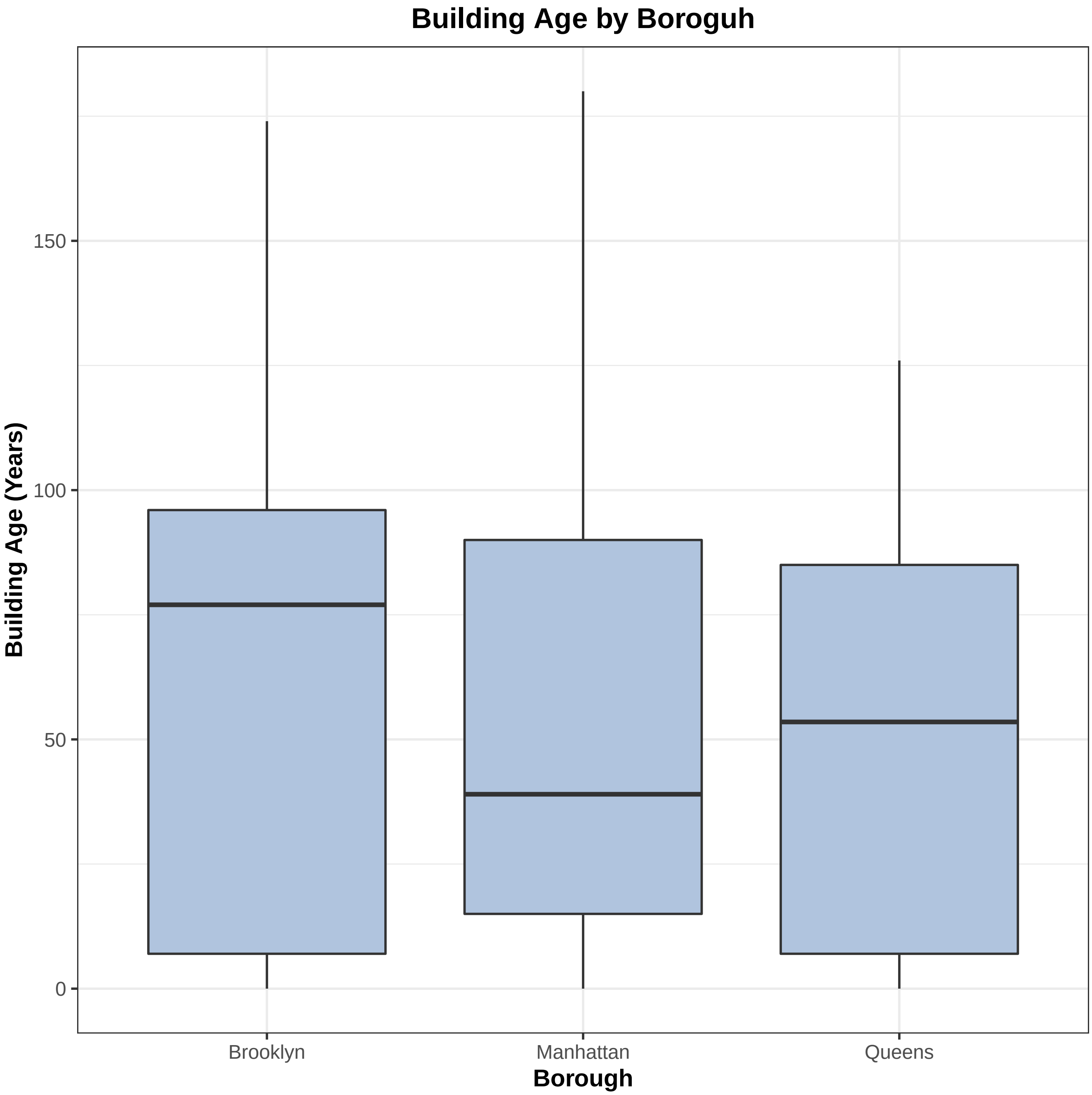
2.1.3 Figure 3:



From the plot above, we see that Manhattan, on average, has smaller units for the price in comparison to Brooklyn and Queens. The submarkets of Brooklyn and Queens are generally close, however Queens submarkets tend to give a slightly roomier unit for the price.

While roomier apartments are beneficial for living conditions, outdated apartments may be a deal breaker for movers. We can now inspect the age of the building that each apartment is in based on it's location.

2.1.4 Figure 4:



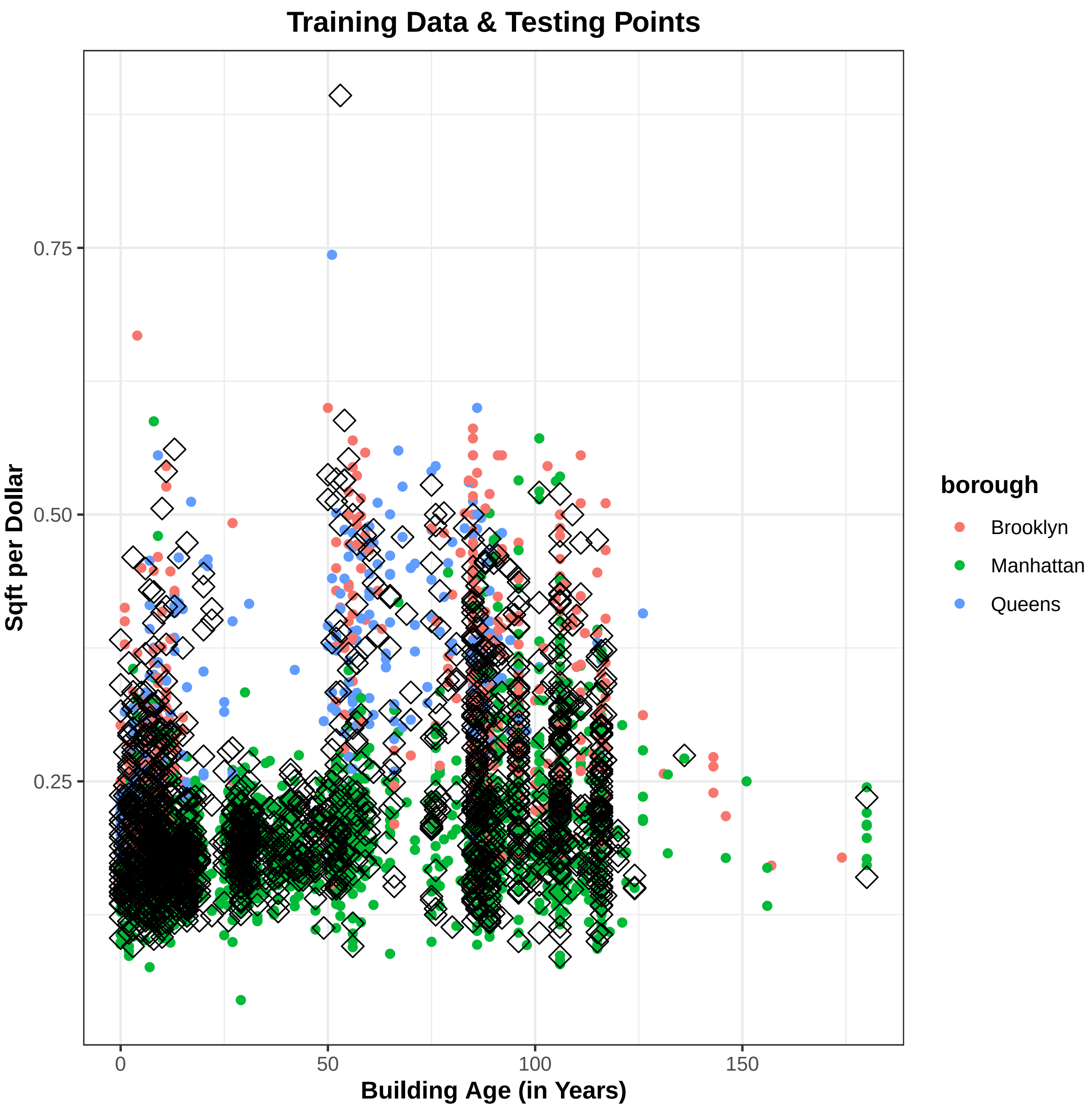
From the boxplot above, we can see that on average, Brooklyn has the oldest buildings but has a lot of variability. Manhattan, on average, has the newest buildings. When renting a place to live in New York, size does matter, however, some of the units with a large square footage may be outdated.

2.2 KNN Classification

Can the location of a unit be predicted based on the square-footage and age of the unit?

To to create this model, KNN will be performed with the variables "building_age_yrs" and "sqft_per_dollar" being used to predict which borough the unit is located in.

2.2.1 Figure 5



A value of k = 23 will be used (square-root of 5000 rows of data).

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Col Total |
|-------------------------|
```

```
Total Observations in Table:  1500


                   | test_classes
street_knn_classes |  Brooklyn | Manhattan |    Queens | Row Total |
-------------------|-----------|-----------|-----------|-----------|
          Brooklyn |       133 |        73 |        57 |       263 |
                   |     0.463 |     0.068 |     0.388 |           |
-------------------|-----------|-----------|-----------|-----------|
         Manhattan |       151 |       985 |        82 |      1218 |
                   |     0.526 |     0.924 |     0.558 |           |
-------------------|-----------|-----------|-----------|-----------|
            Queens |         3 |         8 |         8 |        19 |
                   |     0.010 |     0.008 |     0.054 |           |
-------------------|-----------|-----------|-----------|-----------|
      Column Total |       287 |      1066 |       147 |      1500 |
                   |     0.191 |     0.711 |     0.098 |           |
-------------------|-----------|-----------|-----------|-----------|


Confusion Matrix and Statistics

           Reference
Prediction  Brooklyn Manhattan Queens
  Brooklyn       133        73     57
  Manhattan      151       985     82
  Queens           3         8      8

Overall Statistics

               Accuracy : 0.7507
                 95% CI : (0.728, 0.7724)
    No Information Rate : 0.7107
    P-Value [Acc > NIR] : 0.0002978

                  Kappa : 0.3576

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Brooklyn Class: Manhattan Class: Queens
Sensitivity                  0.46341           0.9240      0.054422
Specificity                  0.89283           0.4631      0.991870
Pos Pred Value               0.50570           0.8087      0.421053
Neg Pred Value               0.87551           0.7128      0.906144
Prevalence                   0.19133           0.7107      0.098000
Detection Rate               0.08867           0.6567      0.005333
Detection Prevalence         0.17533           0.8120      0.012667
Balanced Accuracy            0.67812           0.6936      0.523146
```

Based on the results of the model, we can accurately predict the borough of a NYC listing unit with **79.47% accuracy** based on its square-feet per dollar and the age of its building. Manhattan and Brooklyn were the most frequently mistaken boroughs for this model. This model tells us that the best "bang for your buck" will likely depend on which borough you decide to rent from. While Queens typically has a larger amount of space for the price, the age of the building is generally not as old as those in Brooklyn. Manhattan has the worst average sqft per dollar, however the buildings are newer on average. Brooklyn has a similar avg sqft per dollar to Queens, however the building age has a lot of variability with a very large average value.

While there are numerous factors to consider when living in New York City, it appears that Queens would likely be the best place to rent a home based solely on the age of the building and the space you are able to get for the price.

3. Citations

- ManhattanMiami Real Estate: https://www.manhattanmiami.com/resources/buying-an-apartment-in-nyc
- https://github.com/Codecademy/datasets/tree/master/streeteasy
- http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/

4. Appendix

DATA IMPORT:

```
# setting object to data url
data_url = "https://raw.githubusercontent.com/Codecademy/datasets/master/streeteasy/streeteasy.csv"

# importing the dataset
street_easy <- read_csv(url(data_url))

# viewing data
street_easy %>% head()

# removing the na values for neighborhood, burrow, etc.
street_easy <- subset(street_easy, !is.na(borough) & !is.na(submarket) & !is.na(neighborhood))
```

FIGURE 1:

```
street_easy %>% summary() %>% pander()
```

FIGURE 2:

```
# new var: sqft per dollar
street_easy$sqft_per_dollar <- (street_easy$size_sqft / street_easy$rent)

# summary of new variable
street_easy$sqft_per_dollar %>% summary() %>% pander()
```

FIGURE 3:

```
# aggregating the mean for each submarket
agg_submarket <- aggregate(street_easy,
                           by = list(street_easy$submarket),
                           FUN = mean)
agg_submarket$Borough <- c('Manhattan', 'Manhattan', 'Manhattan', 'Manhattan', 'Manhattan', 'Queens', 'Brookly
                           'Brooklyn', 'Queens', 'Brooklyn', 'Queens', 'Brooklyn', 'Brooklyn', 'Queens', 'Quee

# plotting avg sqft per dollar vs. submarket
ggplot(data = agg_submarket,
       aes(x = Group.1, y = sqft_per_dollar, fill = Borough)) +
  geom_bar(stat = "identity", color = 'black') +
  labs(title = 'Average Sq-ft per Dollar by Submarket', x = 'Submarket', y = 'Avg Sq-ft per Dollar') +
  theme_bw() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
                     legend.position = 'right', plot.title = element_text(hjust = 0.5),
                     axis.title = element_text(face="bold"),
                     legend.title = element_text(face="bold")) +
  guides(fill = guide_legend(override.aes = list(colour = "black"))) +
  scale_fill_brewer(palette="Pastel1")
```

NEW VARIABLE:

```
# Creating a new categorical variable based off of the sqft per dollar
# This will categorize the data based on the size of the unit you are getting for the price
# 0 = expensive , 1 = average , 2 = bargain
street_easy <- street_easy %>% mutate(
  bargain = if_else(street_easy$sqft_per_dollar < 0.1714, 0,
                    if_else(street_easy$sqft_per_dollar >= 0.1714 & street_easy$sqft_per_dollar <= 0.2574,
                            1, 2)))
```

FIGURE 4:

```
ggplot(data = street_easy, aes(x = borough, y = building_age_yrs)) + geom_boxplot(fill = 'lightsteelblue', ) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", color = 'black'),
        axis.title = element_text(face="bold", color = 'black'),
        legend.title = element_text(face='bold')) + labs(x = 'Borough', y = 'Building Age (Years)', title = 'B
```

KNN SETUP:

```
# subsetting the data into an approximately 70%/30% training/testing split
index <- sample(1:nrow(street_easy), round(nrow(street_easy) * 0.7))
training_df <- street_easy[index, ]
testing_df <- street_easy[-index, ]

# Storing the training/testing data features
train_features <- training_df[, c(9,22)]
test_features <- testing_df[, c(9,22)]

# Storing the actual labels
train_classes <- training_df$borough
test_classes <- testing_df$borough
```

FIGURE 5:

```
p1 <- ggplot(training_df, aes(x = building_age_yrs, y = sqft_per_dollar, color = borough)) + geom_point() + th
geom_point(data = testing_df, aes(x = building_age_yrs, y = sqft_per_dollar), color = "black", pch = 5, size =
  labs(x = 'Building Age (in Years)', y = 'Sqft per Dollar', title = 'Training Data & Testing Points') +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", color = 'black'),
        axis.title = element_text(face="bold", color = 'black'),
        legend.title = element_text(face='bold'))

p1
```

KNN RESULTS:

```
# knn with k = 23 (square root of 5000 rounded)
street_knn_classes <- knn(train = train_features, test = test_features,
cl = train_classes, k = 23)


# Show the confusion matrix
CrossTable(x = street_knn_classes, y = test_classes, prop.chisq = FALSE, prop.t = F, prop.r = F)

# confusion matrix
confusionMatrix(data = street_knn_classes, reference = as.factor(test_classes))
```