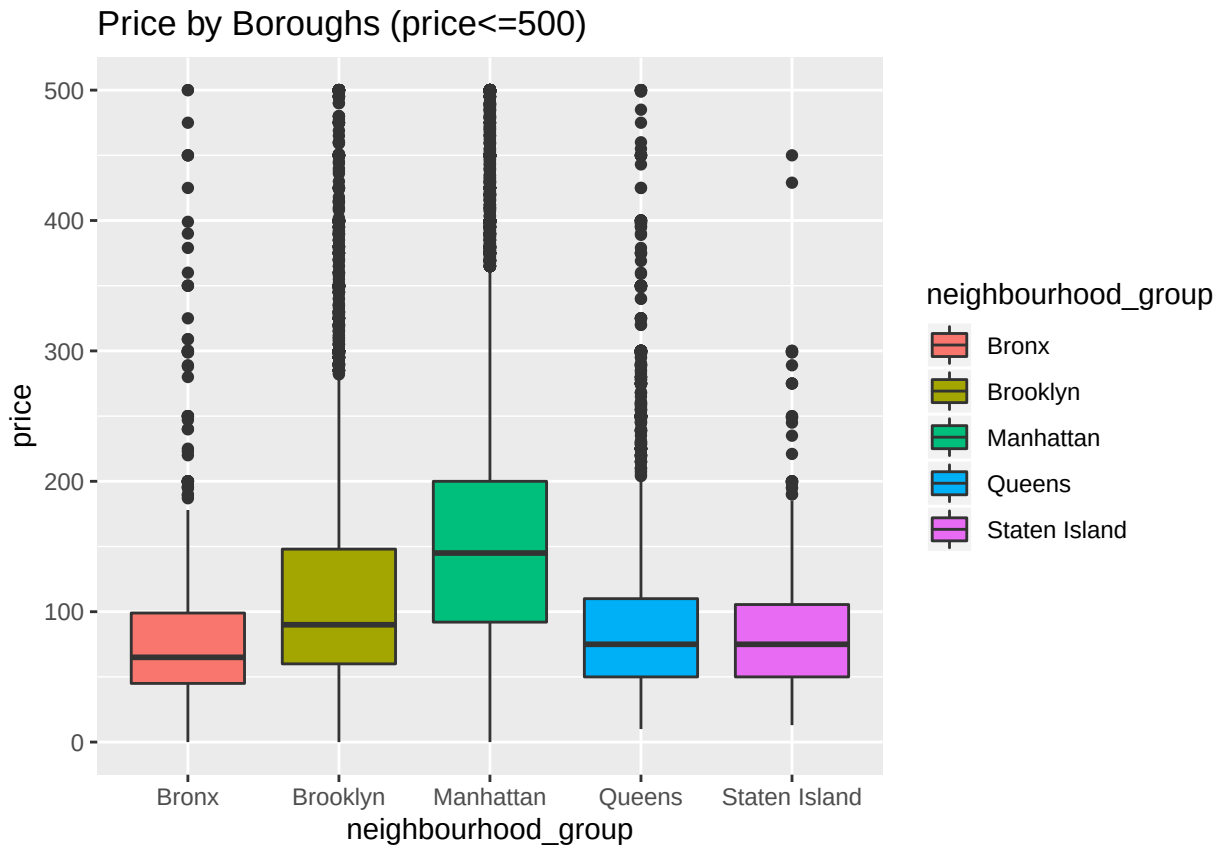
Exploratory Analysis of Data for Airbnb Listings in NYC

Youngsoo Baek, Irene Yi Ji, Phuc Nguyen

Exploratory Analysis

```
## # A tibble: 5 x 5
##   neighbourhood_group median    Q1    Q3   max
##   <fct>                <dbl> <dbl> <dbl> <dbl>
## 1 Bronx                   65    45    99  2500
## 2 Brooklyn                90    60   150 10000
## 3 Manhattan              150    95   220 10000
## 4 Queens                  75    50   110 10000
## 5 Staten Island           75    50   110  5000
```

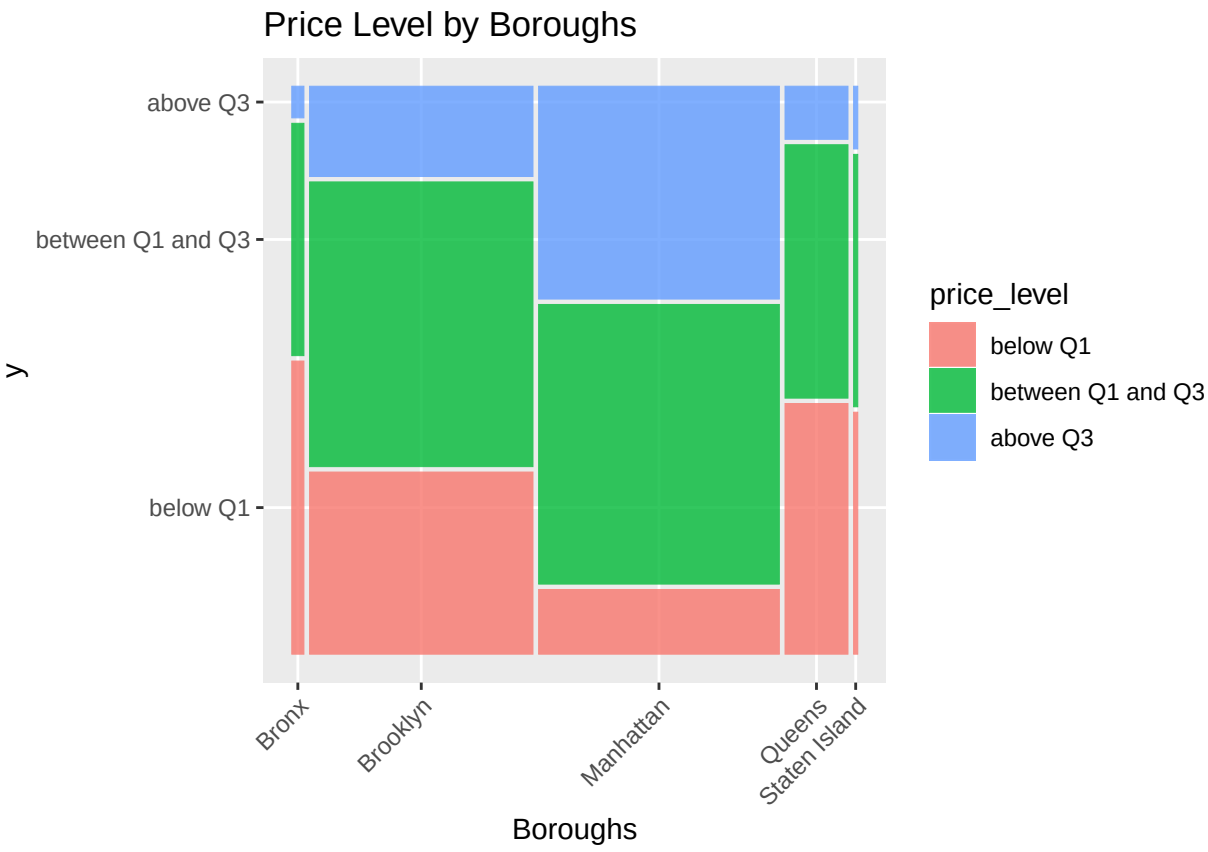
```
##   median Q1  Q3   max
## 1    106 69 175 10000
```




```
##
##                      Bronx Brooklyn Manhattan Queens Staten Island
##   below Q1             574     6572      2514   2549           162
##   between Q1 and Q3    455    10274     10872   2577           169
```

```
##   above Q3             62      3258       8275     540             42
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_price_borough
## X-squared = 6198.5, df = 8, p-value < 2.2e-16
```



A p-value less than the signficiance level of 0.05 tells us that these groups show statistically significant differences in their distributions between the 5 categories. The samples are heterogeneous.

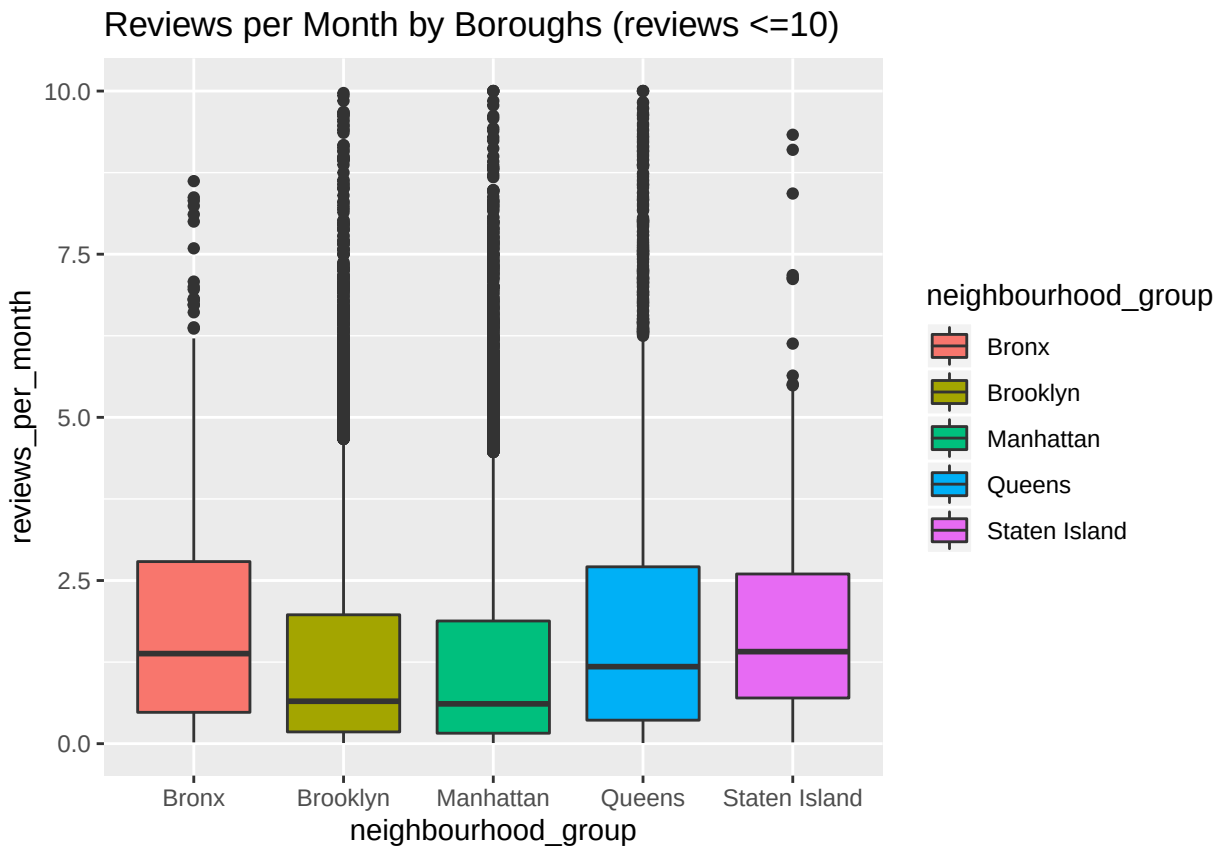
As shown in the boxplot, Manhattan has the highest price.

```
## [1] 0
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_price_ngh
## X-squared = 14322, df = 440, p-value < 2.2e-16
```

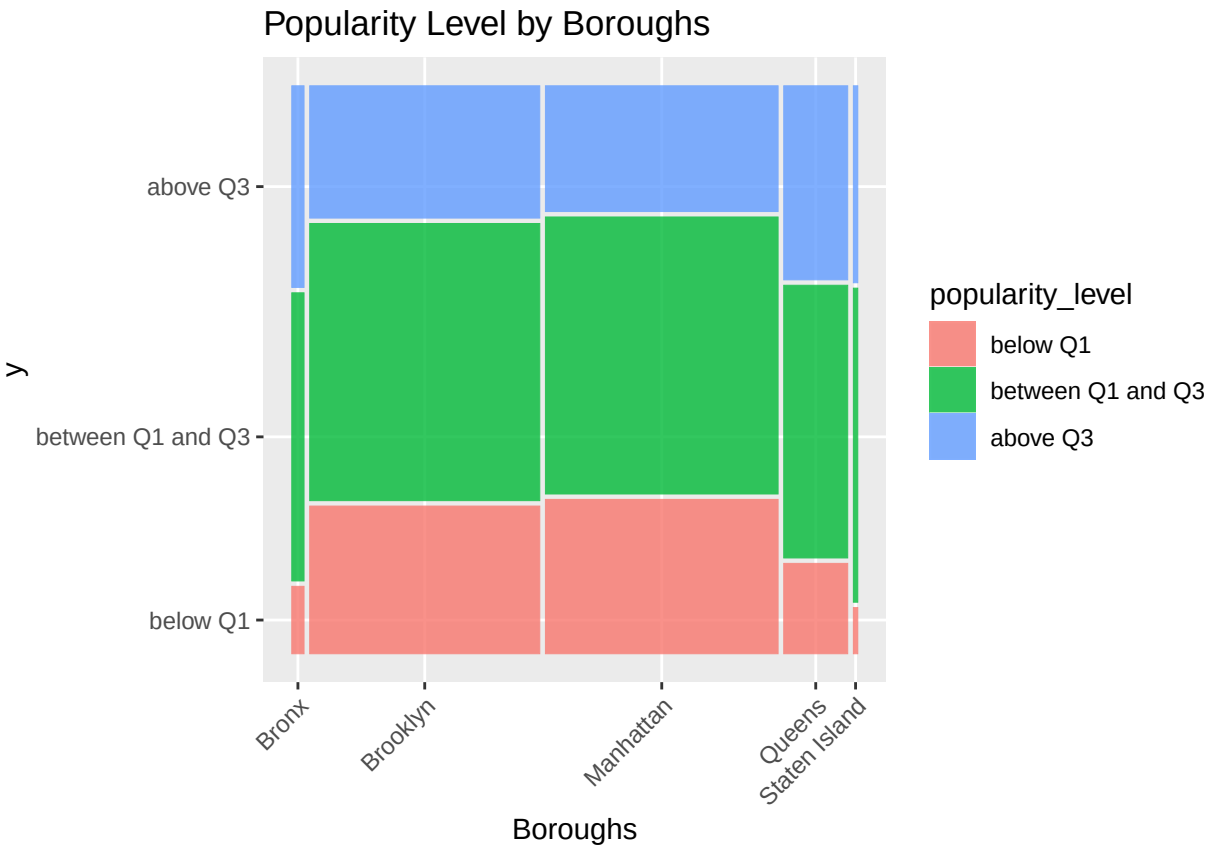
```
## # A tibble: 5 x 5
##   neighbourhood_group median    Q1    Q3   max
##   <fct>                <dbl> <dbl> <dbl> <dbl>
## 1 Bronx                    1     0     3    10
## 2 Brooklyn                 1     0     2    14
## 3 Manhattan                1     0     2    58
## 4 Queens                   1     0     3    21
## 5 Staten Island            1     1     3    10
```

```
##   median Q1 Q3 max
## 1      1  0  2  58
```




```
##
##                     Bronx Brooklyn Manhattan Queens Staten Island
##   below Q1            106     4360      4608    742            26
##   between Q1 and Q3   453     8182      8270   2239           177
##   above Q3            317     3905      3754   1593           111
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_pop_borough
## X-squared = 580.84, df = 8, p-value < 2.2e-16
```



According to the chi-square test, there is still heterogeity among boroughs regarding popularity of listings. However, as shown in the boxplot and mosaic plot, the difference is not as significant as price. Manhattan has the lowest popularity, in terms of monthly review rate.

```
## [1] 0
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_pop_ngh
## X-squared = NaN, df = 440, p-value = NA
```

Response of Interest: Price and Popularity

Choosing a Meaningful Measure of Popularity

Heterogeneity across Neighbourhoods/Boroughs

Spatial Correlation

Predictors of Interest

Possibly Unreliable Predictors

Modeling

Price and Popularity: Bivariate Mixed Effects Regression

Did We Miss Spatial Correlation Within Neighbourhoods?

Text Analysis for Listing Names

Further Work