

# The Airbnb Market in NYC: An Exploration Data Analysis

Tai Chou-Kudu

```
library(ggribbles)
library(tidyverse)
library(ggplot2)
library(dplyr)
```

```
nycbnb = nycbnb |>
  select(
    id,
    price,
    neighborhood,
    borough,
    accommodates,
    bathrooms,
    bedrooms,
    beds,
    review_scores_rating,
    number_of_reviews,
    listing_url )
```

## Initial Exploration

Q: How many rows does the dataset have?

A: The dataset has 37765 rows.

Q: What does each row in the dataset represent?

A: Each row represents an Airbnb listing. Each column represents an aspect of the listing, such as number of bedrooms, neighborhood location, price of rental, or listing url.

## Data Preparation

```
# Filter data for Brooklyn
brooklyn_data <- nycbnb[nycbnb$borough == "Brooklyn", ]
# Remove null values
clean_data <- brooklyn_data[is.finite(brooklyn_data$price), ]
```

## Data Visualization

We'll create a faceted histogram of this cleaned Brooklyn Airbnb data. We have 48 neighborhoods to display, so a layout of 6 columns and 8 rows will allow us to see all of the facets/neighborhoods at once, while being able to scan for trends in pricing.

```
ggplot(clean_data, aes(x = price)) +
  geom_histogram(binwidth = 100, fill = "skyblue", color = "black") +
  facet_wrap(~ neighborhood, ncol = 6) +
  labs(title = "Faceted Histogram of Prices in Brooklyn by Neighborhood",
       x = "Price",
       y = "Count") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(size = 8),
    axis.text.y = element_text(size = 8),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    plot.title = element_text(size = 15, hjust = 0.5)
  )
```

Faceted Histogram of Prices in Brooklyn by Neighborhood



## Data Visualization Improvements

The spatial layout has room for improvement. Bed-Stuy shows an outlier of maximum price in its histogram chart, which skews data visualization and leads to scale distortion. A future idea is to normalize the data or transform to account for the outlier, so that we can see more of the data for each neighborhood, instead of the scale distortion creating a “vertically squished” effect.

## Further Data Exploration

Identify the neighborhoods city-wide with the top five median listing prices that have a minimum of 50 listings.

```
top_5_neighborhoods <- clean_data %>%
  group_by(neighborhood) %>%
  filter(n() >= 50) %>%
  summarize(median_price = median(price, na.rm = TRUE)) %>%
  arrange(desc(median_price)) %>%
  slice_head(n = 5) %>%
  pull(neighborhood)

top_5_neighborhoods
```

```
[1] "Boerum Hill"      "Greenpoint"      "Park Slope"      "Carroll Gardens"
[5] "Clinton Hill"
```

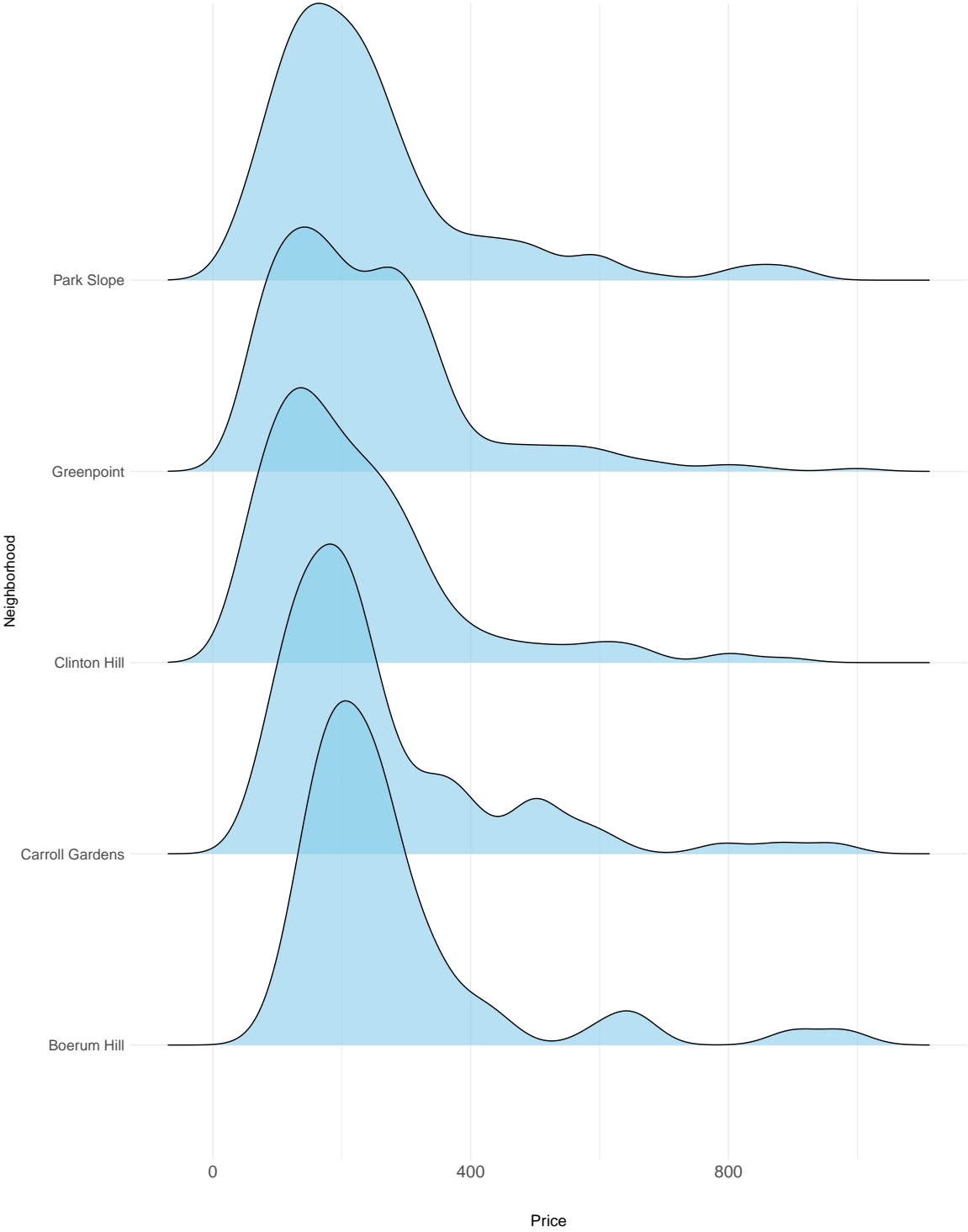
Filter the data for these five neighborhoods and make ridge plots of the distributions of listing prices in these five neighborhoods

```
filtered_data <- clean_data %>%
  filter(neighborhood %in% top_5_neighborhoods)

#|message: false
ggplot(filtered_data, aes(x = price, y = neighborhood)) +
  geom_density_ridges(fill = "skyblue", color = "black", alpha = 0.6) +
  labs(title = "Distribution of Listing Prices in Top 5 Neighborhoods",
       x = "Price",
       y = "Neighborhood") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(size = 15),
    axis.text.y = element_text(size = 13),
    axis.title.x = element_text(size = 14, margin = margin(t = 30)),
```

```
axis.title.y = element_text(size = 13),  
plot.title = element_text(size = 16, hjust=0.5, margin=(margin(b=40)))  
)
```

Distribution of Listing Prices in Top 5 Neighborhoods



Calculate the minimum, mean, median, standard deviation, IQR, and maximum listing price in each of these neighborhoods.

```
summary_stats <- filtered_data %>%
  group_by(neighborhood) %>%
  summarize(
    min_price = min(price, na.rm = TRUE),
    mean_price = mean(price, na.rm = TRUE),
    median_price = median(price, na.rm = TRUE),
    sd_price = sd(price, na.rm = TRUE),
    iqr_price = IQR(price, na.rm = TRUE),
    max_price = max(price, na.rm = TRUE)
  )

summary_stats
```

```
# A tibble: 5 x 7
  neighborhood min_price mean_price median_price sd_price iqr_price max_price
  <chr>         <dbl>     <dbl>         <dbl>     <dbl>     <dbl>     <dbl>
1 Boerum Hill    99       279.           231       173.       115       980
2 Carroll Gardens 75       261.           200       177.       163       963
3 Clinton Hill   48       222.           185       151.       154       890
4 Greenpoint     46       240.           209       151.       170.       998
5 Park Slope     44       244.           201       169.       146.       900
```

## Data Analysis

Boerum Hill and Carroll Gardens show the highest mean price and lowest minimum price of Airbnb rentals. These are more upscale neighborhoods, and that is reflected in these summary statistics. Rentals generally start at higher prices in Boerum Hill, as seen in the ridge distribution chart.

Clinton Hill, Greenpoint, and Park Slope share the lowest minimum prices of Airbnb rentals. Park Slope seems to have a greater amount of highly priced rentals than the other 2 neighborhoods, judging by the distribution charts. The maximum price, however is highest in Greenpoint, showing the upscale atmosphere of the neighborhood. Park Slope and Clinton Hill, perhaps, have various areas of the neighborhood, some that may be considered lower in market value. Clinton Hill has the lowest median price and mean price of the 5 neighborhoods, which isn't surprising to me as a New Yorker. I think this speaks to the rapid gentrification of Clinton Hill over the past decade, and the fact that some areas are still considered "shady". However, noise complaints may be a potential factor to consider, among other factors. Park Slope is considered a wealthy area, but some areas are super wealthy, while others may have

predominantly middle-class residents. A comparison of resident income/wealth data with these Airbnb rental datasets could provide additional insights.

Create a visualization that will help you compare the distribution of review scores (review\_scores\_rating) across neighborhoods.

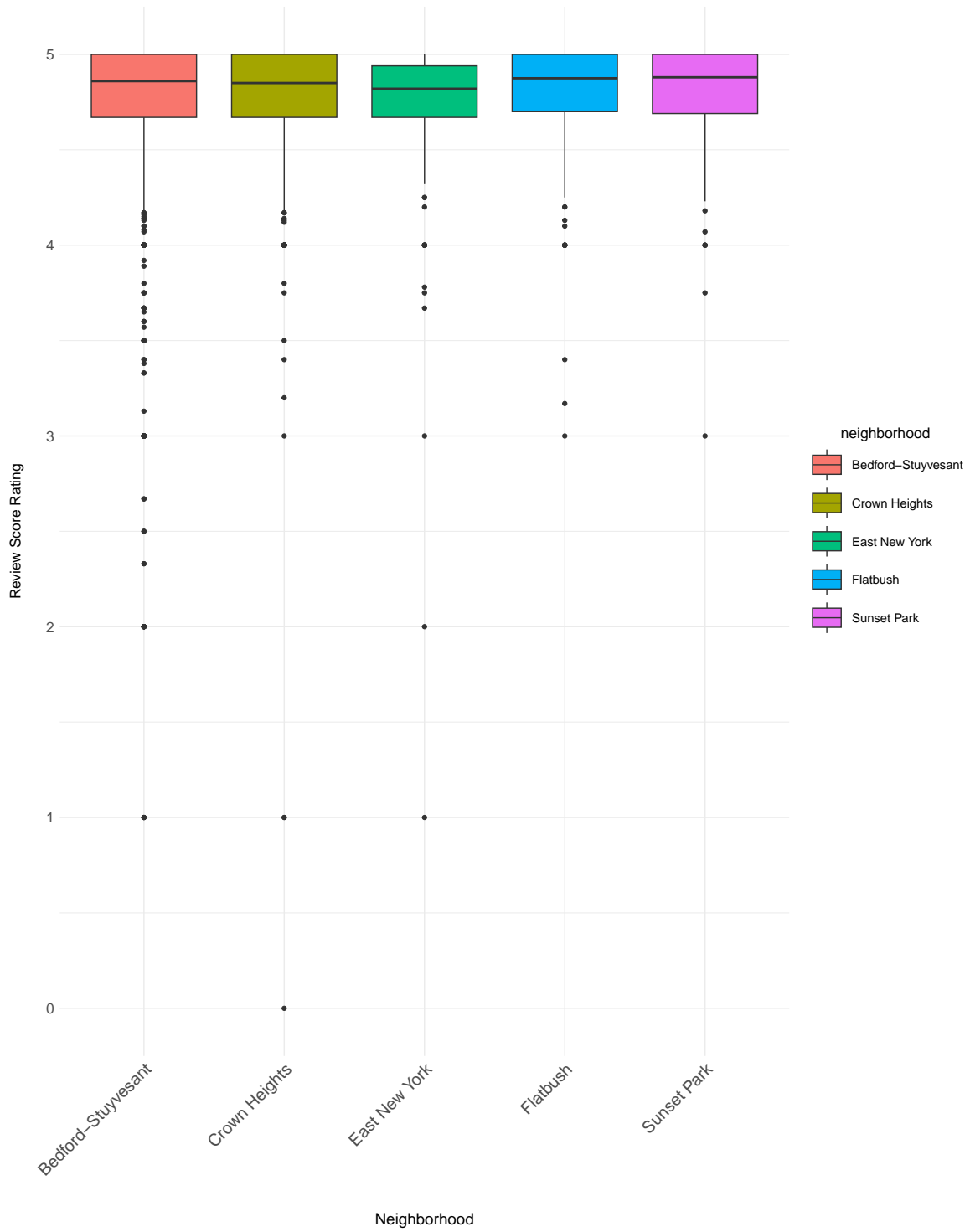
```
neighborhoods_of_interest <- c("Bedford-Stuyvesant",
                              "East New York", "Crown Heights",
                              "Flatbush", "Sunset Park")

filtered_data <- nycbnb %>%
  filter(!is.na(review_scores_rating)) %>%
  filter(neighborhood %in% neighborhoods_of_interest)

ggplot(filtered_data, aes(x = neighborhood,
                        y = review_scores_rating,
                        fill = neighborhood)) +
  geom_boxplot() +
  labs(title = "Distribution of Review Scores Across Neighborhoods",
       x = "Neighborhood",
       y = "Review Score Rating") +
  theme_minimal() +
  theme(
    legend.title = element_text(size = 13, hjust=0.5),
    legend.text = element_text(size = 11),
    legend.key.size = unit(1.2, "cm"),
    axis.text.x = element_text(size = 15, angle = 45, hjust = 1),
    axis.text.y = element_text(size = 13),
    axis.title.x = element_text(size = 14, margin=(margin(t=30))),
    axis.title.y = element_text(size = 13, margin=(margin(r=20))),
    plot.title = element_text(size = 16, hjust=0.5, margin=(margin(b=40)))
  )
```



Distribution of Review Scores Across Neighborhoods



It seems like Bed-Stuy has the lowest overall range of reviews, which may be due to the rapid recent gentrification— with too many resulting factors and social dynamic to explain briefly, that cause dissatisfaction for renters. I'm surprised East New York's ratings are so high because it's known as one of New York's dangerous neighborhoods. It's interesting to think about local New Yorkers' perceptions of neighborhoods—those of us who have seen the rapid change and gentrification of neighborhoods, versus those who are merely visiting from elsewhere, only seeing the neighborhood through these outside eyes or through their romanticizing of NYC.