DSAN 5100

Airbnb: What Factors Influence Rental Prices?

Probabilistic Modeling & Statistical Computing

Car Barn Room 204 - 12/04/2024

Team Members

Adam Stein

Jessica Joy

Hung Tran

Soong-Ping Hill

David Corcoran

Table of Contents

1. Data Science Questions
2. Data Collection and Cleaning
3. Exploratory Data Analysis
4. Statistical Testing
5. Conclusion

Data Science Questions

1. What factors most strongly influence the price of an Airbnb listing?

2. Are rental prices of Superhosts greater than those of regular hosts?

3. How do Airbnb prices differ between major cities?

4. How does the crime rate of the neighborhood influence Airbnb prices?

5. To what extent does distance from the city center play a role in Airbnb pricing?

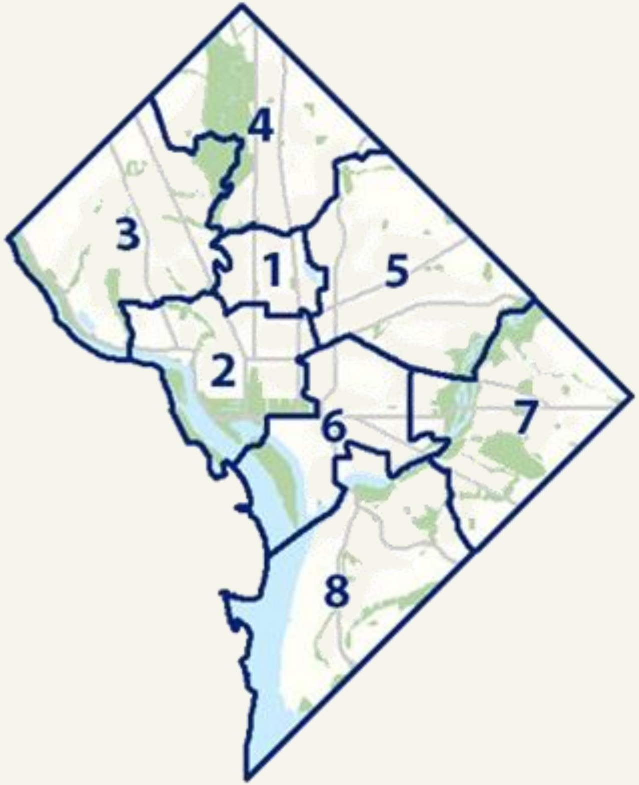
Data Collection

3 Datasets with 75 attributes from “Inside Airbnb” for 2024

1 Dataset on crime sourced from opendata.dc.gov for 2023

1 Original dataset created to match DC neighborhoods to wards

distance_to_city_center column calculated using existing longitude and latitude values



DC.Gov, 2024

Data Cleaning

Step 1: Merge Datasets

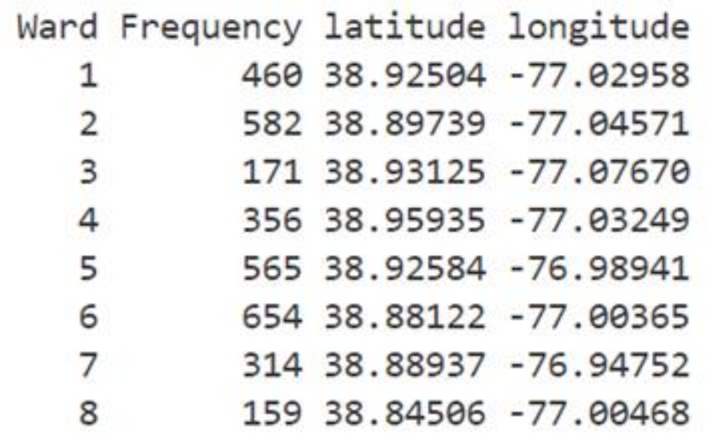
- listings.csv.gz and wards.csv

Step 2: Drop Unnecessary Columns

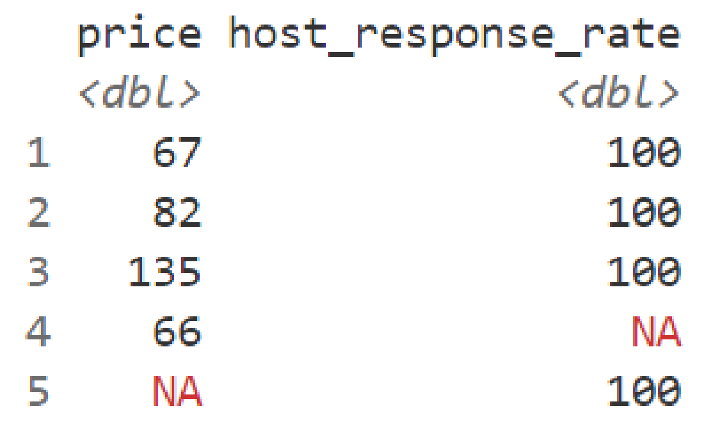
- 75 columns -> 20 columns

Step 3: Alter Column Data Types

- **price:** Remove “$” & convert to numeric
- **host_response_rate:** Remove “%” & convert to numeric



```
Ward Frequency latitude longitude
   1       460 38.92504 -77.02958
   2       582 38.89739 -77.04571
   3       171 38.93125 -77.07670
   4       356 38.95935 -77.03249
   5       565 38.92584 -76.98941
   6       654 38.88122 -77.00365
   7       314 38.88937 -76.94752
   8       159 38.84506 -77.00468
```

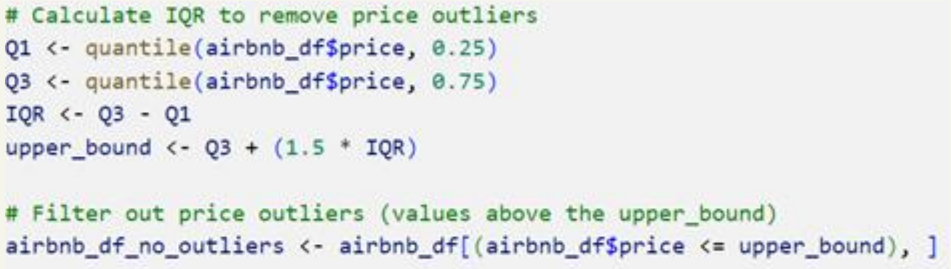


```
  price host_response_rate
  <dbl>              <dbl>
1    67                100
2    82                100
3   135                100
4    66                 NA
5    NA                100
```

Data Cleaning

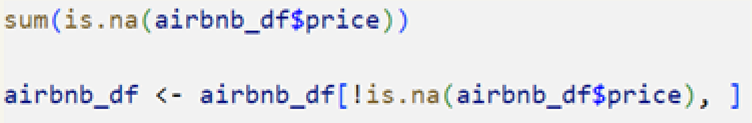
Step 4: Remove Price Outliers

- Calculate IQR and the upper bound
- Filter out values greater than the upper bound



```
# Calculate IQR to remove price outliers
Q1 <- quantile(airbnb_df$price, 0.25)
Q3 <- quantile(airbnb_df$price, 0.75)
IQR <- Q3 - Q1
upper_bound <- Q3 + (1.5 * IQR)

# Filter out price outliers (values above the upper_bound)
airbnb_df_no_outliers <- airbnb_df[(airbnb_df$price <= upper_bound), ]
```

Step 5: Remove Null Price Values



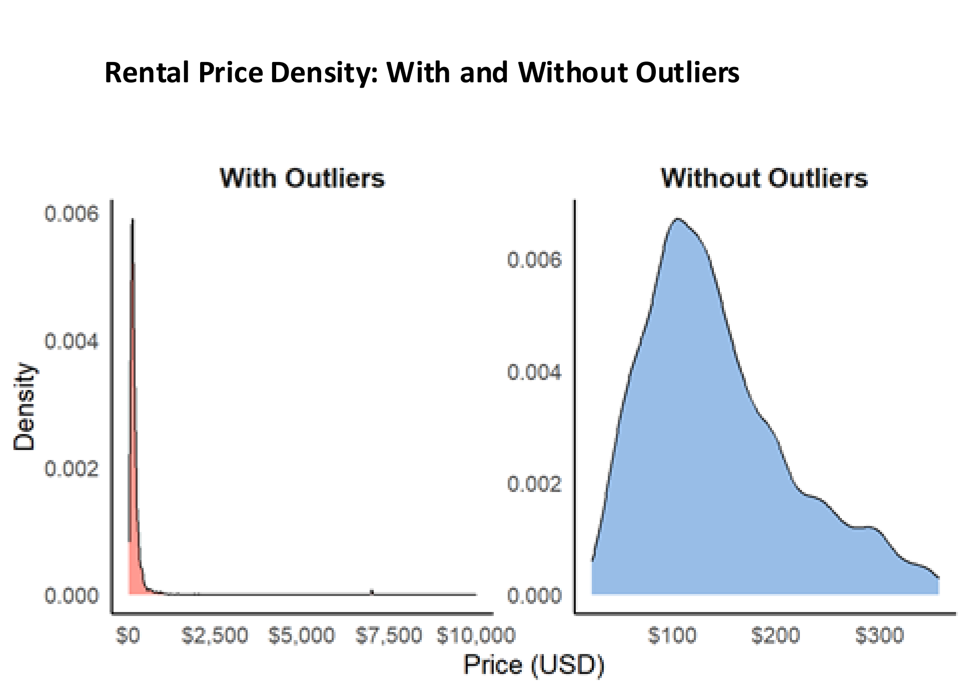
```
sum(is.na(airbnb_df$price))

airbnb_df <- airbnb_df[!is.na(airbnb_df$price), ]
```

Row Count	D.C.	Boston	Chicago
Initial	4928	4325	7952
Final	3887	3302	7112

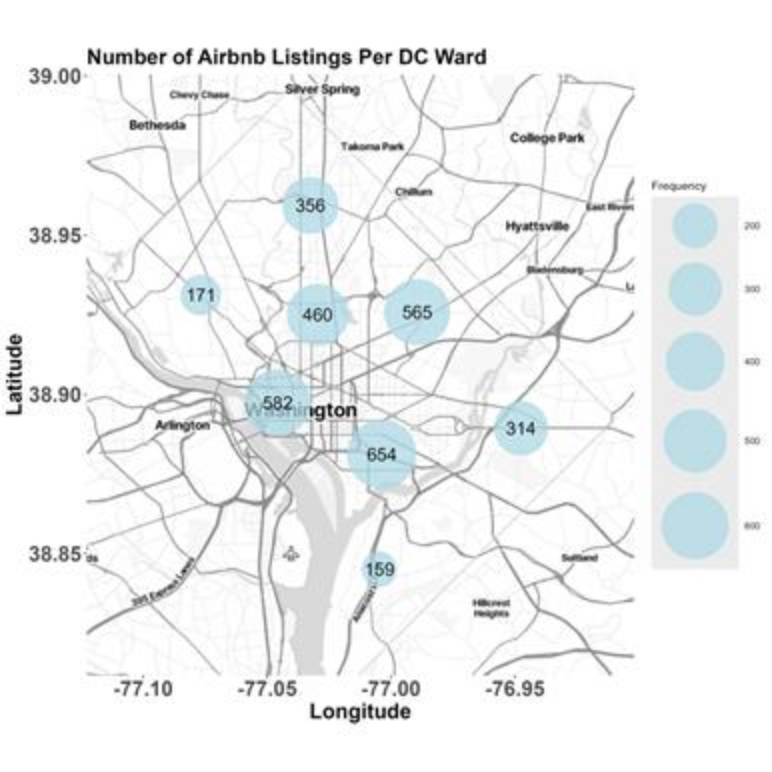
Exploratory Data Analysis

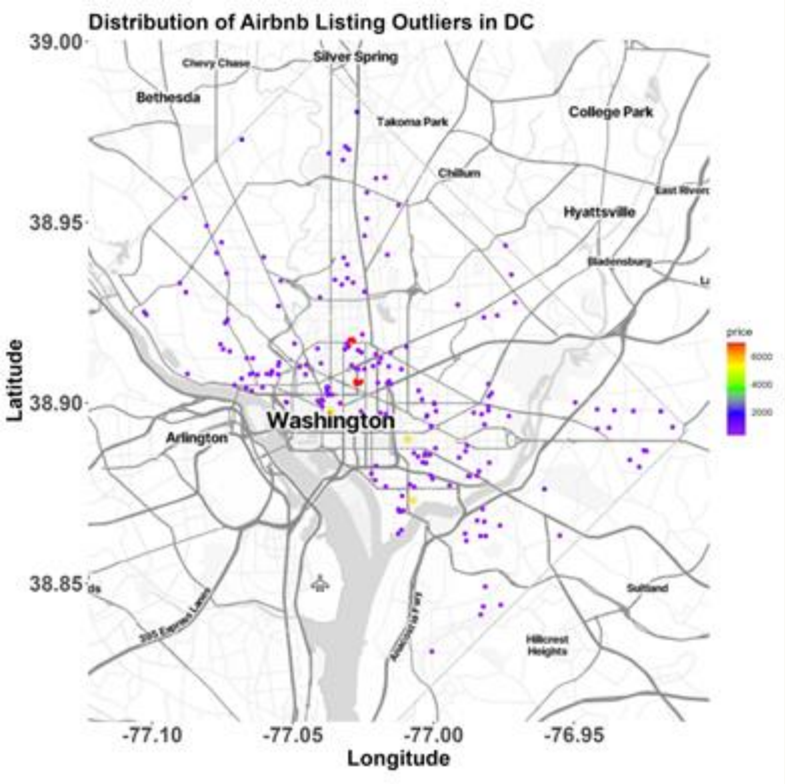
To explore the data across DC, Boston, and Chicago, we created numerous visualizations to get a good understanding of how rental prices are distributed. The following visuals were created for all three cities:

- **Price density with and without outliers**
- Average price per neighborhood
- Boxplot of price by room type
- Geospatial plots of rental prices
- Boxplot of superhost vs not superhost prices
- Price by superhost ratings
- Price by property type
- Price by accommodates, bathrooms, bedrooms, and beds


Rental Price Density: With and Without Outliers

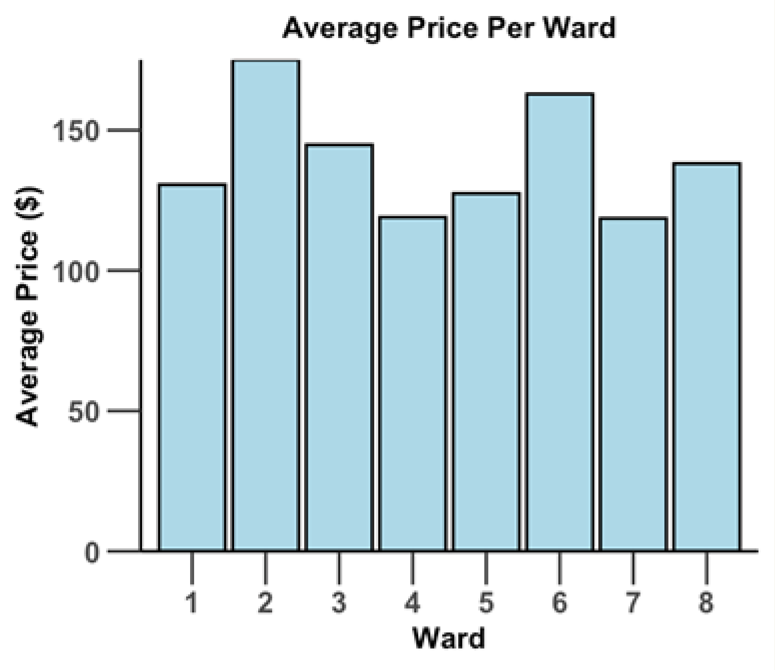
Exploratory Data Analysis



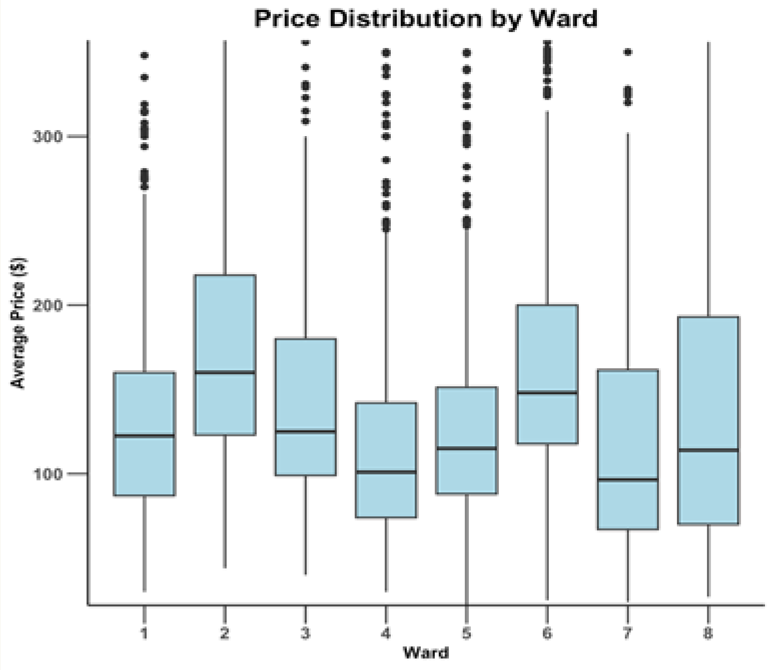


Exploratory Data Analysis

Average Price Per Ward

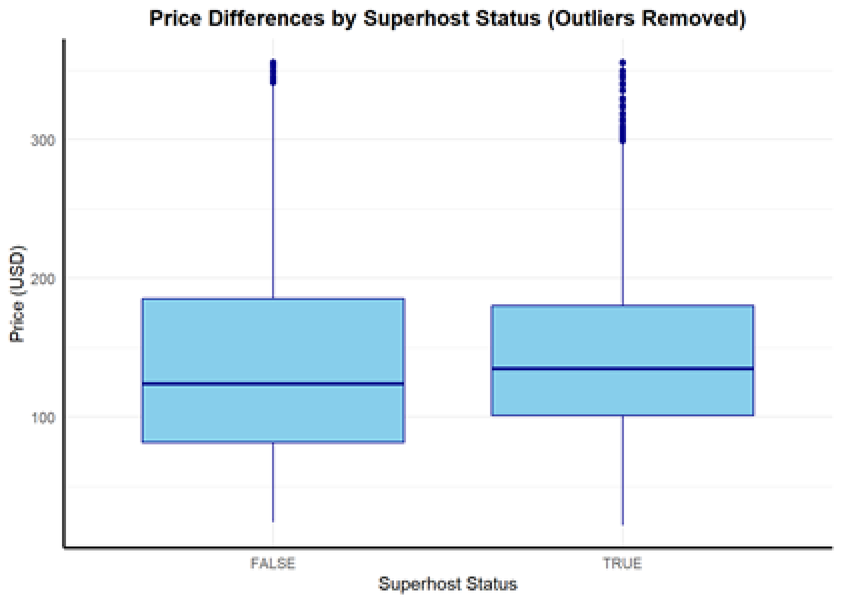


Price Distribution by Ward

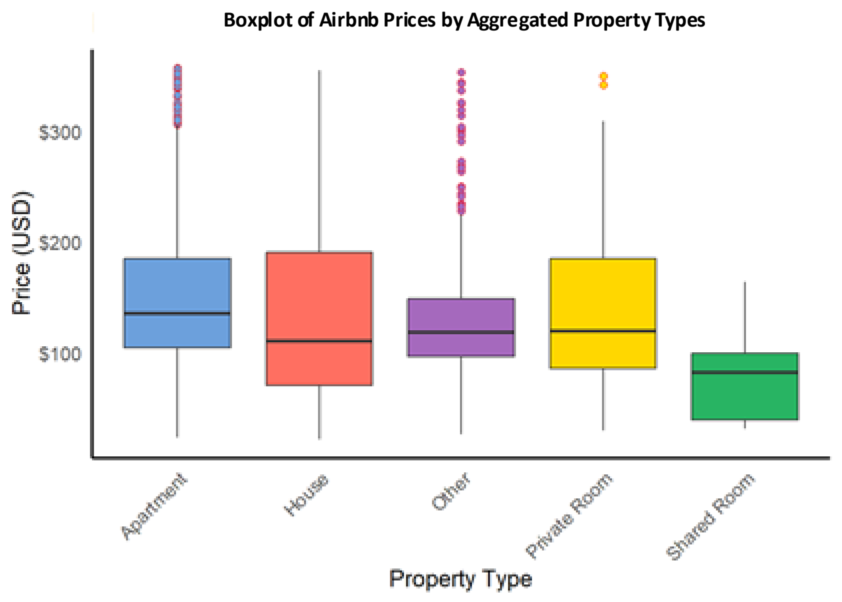


Exploratory Data Analysis

Price by Superhost Status

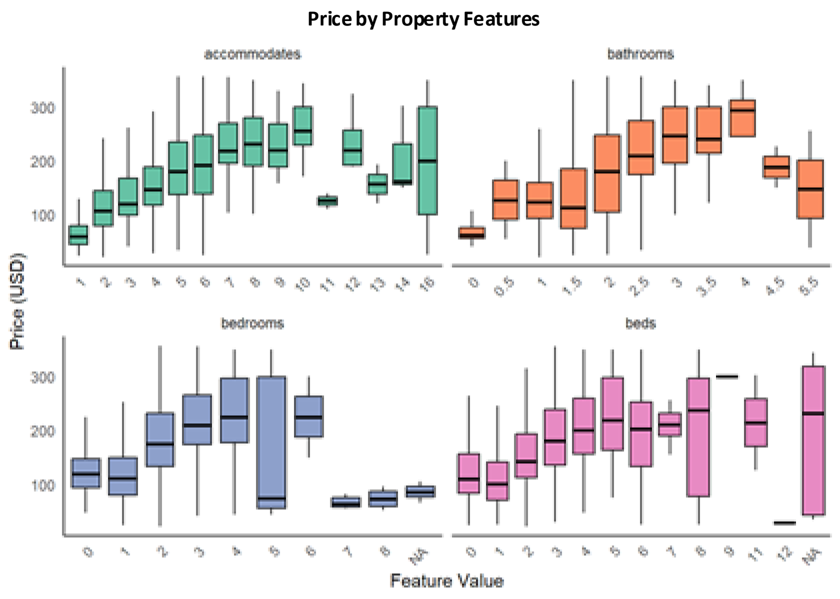


Price by Property Type

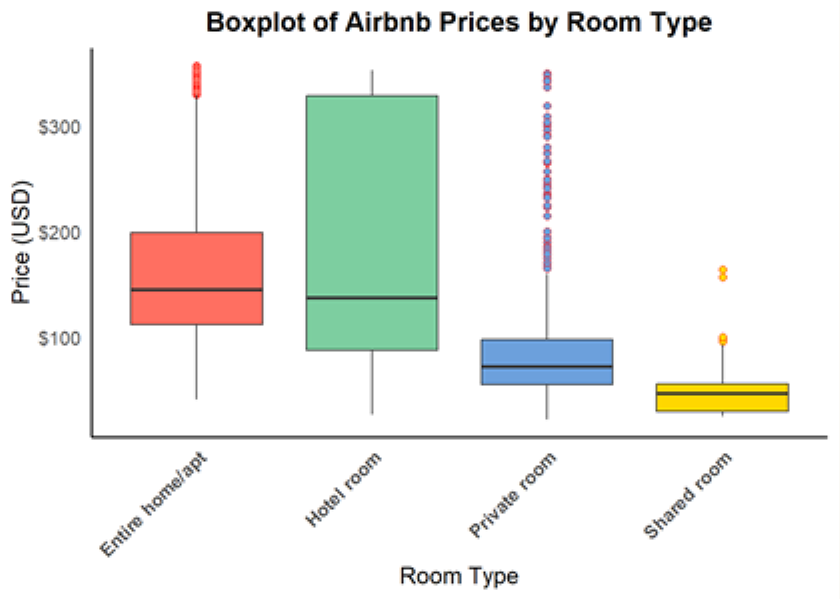


Exploratory Data Analysis

Price by Property Features

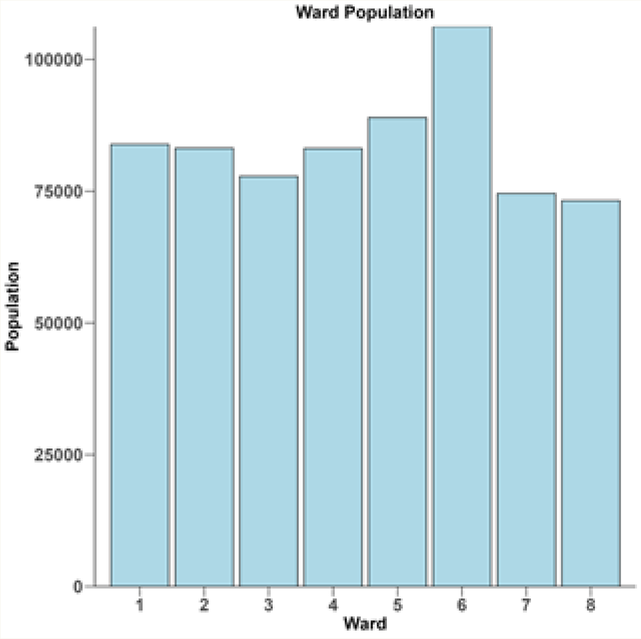


Price by Room Type

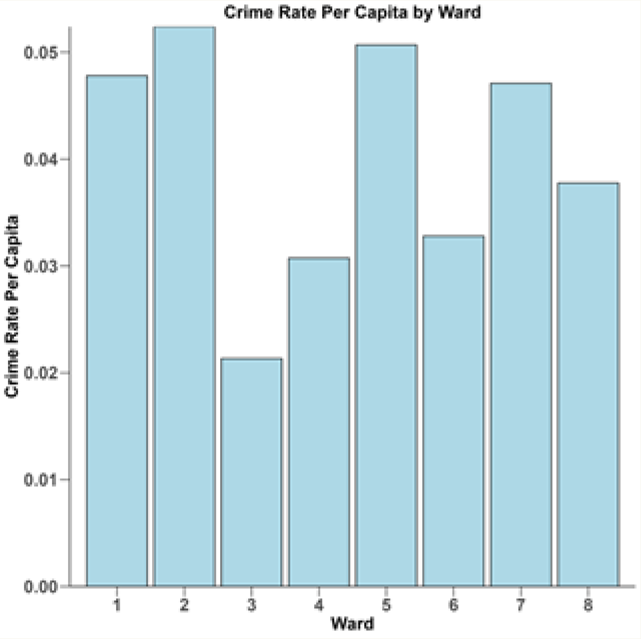


Exploratory Data Analysis

Ward Population



Crime Rates by Ward



Statistical Testing

Price vs Ward

Analysis of Variance Test (ANOVA)

Null Hypothesis (H_0): There is no significant difference in mean Airbnb listing prices across DC Wards

Alternate Hypothesis (H_A): There is at least one DC Ward that significantly differs in mean Airbnb listing price

Results

F-Value = 54.62, p-value = 2e-16

P-value < 0.05, Reject null hypothesis

The ANOVA test suggests that **there is at least one DC Ward** that significantly differs in mean Airbnb listing price

Price vs Ward Crime Rate

Analysis of Variance Test (ANOVA)

Null Hypothesis (H_0): There is no significant difference in mean Airbnb listing prices across crime rate categories (low, medium, high)

Alternate Hypothesis (H_A): There is at least one crime rate category (low, medium, high) in which the mean Airbnb listing price significantly differs from the rest

Results

F-Value = 22.18, p-value = 2.64e-10

P-value < 0.05, Reject null hypothesis

The ANOVA test suggests that there is **at least one crime rate category** in which the mean Airbnb price differs from the rest

Price vs Superhost Status

Two Sample T-Test

Null Hypothesis (H_0): There is no significant difference in mean Airbnb listing prices between superhost and non-superhost listings

Alternate Hypothesis (H_A): The mean Airbnb listing price is greater for superhost listings than non-superhost listings

Results

t = -2.59, df = 3549.8, p-value=0.0049
95% CI: - Infinity to -2.17

p-value < 0.05, Reject null hypothesis

The two sample t-test suggests that the **mean listing prices by non-superhosts are *less* than those of superhosts** on Airbnb

Superhost Status vs Room Type

Chi Squared Test of Independence

Null Hypothesis (H_0): There is no association between Airbnb Superhost Status and Room Type

Alternate Hypothesis (H_A): There is an association between Airbnb Superhost Status and Room Type

Results

X-squared = 121.27, df = 2, p-value < 2.2e-16

p-value < 0.05, Reject null hypothesis

The chi squared test of independence test suggests that **there is an association between Superhost status and the room type** of the Airbnb listing.

Price vs Distance to City Center

Analysis of Variance Test (ANOVA)

Null Hypothesis (H_0): There is no significant difference in mean Airbnb listing prices across distance to city center (Near, Medium, Far)

Alternate Hypothesis (H_A): There is at least one distance category (Near, Medium, Far) in which the mean Airbnb listing price significantly differs from the rest

ANOVA Results

F-Value = 71.33 p-value = <2.2e-16
P-value < 0.05, Reject null hypothesis

The ANOVA test suggests that **there is evidence of a significant difference** in average Airbnb price among distance to city center (Near, Medium, Far)

Tukey's HSD Test Results

Distance Bins	Difference	Lower	Upper	P-value
Medium-Near	-23.00	-28.60	-17.40	0.00
Far-Near	-37.04	-46.51	-27.56	0.00
Far-Medium	-14.04	-23.82	-4.24	0.002

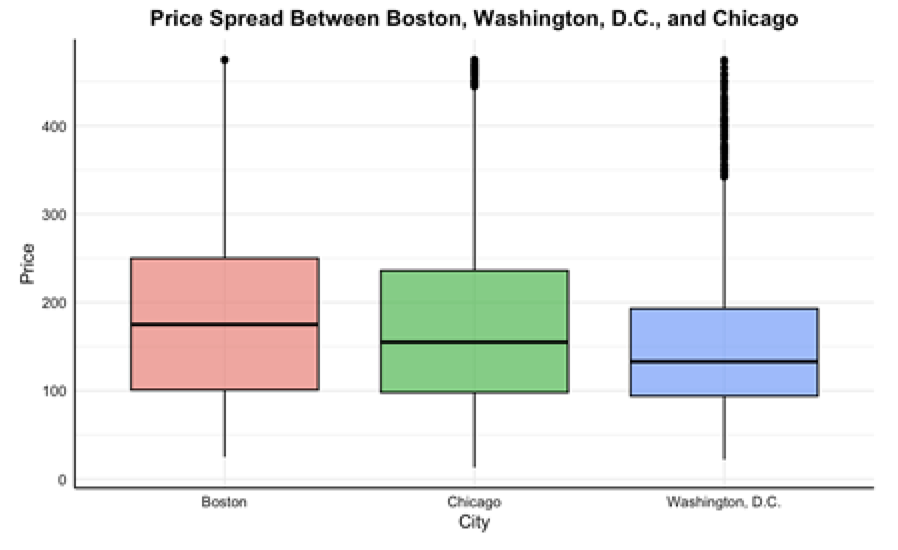
Price Comparisons Across Three Cities

Analysis of Variance Test (ANOVA)

Null Hypothesis (H_0): There is no difference between mean Airbnb listing price and city

Alternate Hypothesis (H_A): There is a difference between mean Airbnb listing price and city

Average Price Per City



Price Comparisons Across Three Cities

Analysis of Variance Test (ANOVA)

Null Hypothesis (H_0): There is no difference between mean Airbnb listing price and city

Alternate Hypothesis (H_A): There is a difference between mean Airbnb listing price and city

Results

F Value = 113.9 df = 2, p-value = 2e-16

P-value < 0.05, reject null hypothesis

The ANOVA test suggests that **there is evidence of difference** in average Airbnb price among different cities (Boston, DC, Chicago)

Conclusion

1. What factors most strongly influence the price of an Airbnb listing?

- Neighborhood
- Distance to city center
- Superhost status
- Property features

2. Are rental prices of Superhosts greater than those of regular hosts?

Yes, but moderately so

3. How do Airbnb prices differ between major cities?

Prices vary moderately between cities with DC having the largest spread

4. How does the crime rate of the neighborhood influence Airbnb prices?

There is a significant difference between the mean prices of properties and ward crime rates

5. To what extent does distance from the city center play a role in Airbnb pricing?

There is a significant difference between the prices of properties near, medium, and far distance away

References

“Crime Incidents in 2024.” *Open Data DC*, opendata.dc.gov/datasets/c5a9f33ffca546babbd91de1969e742d_6/explore?location=38.904150%2C-77.011950%2C11.31&showTable=true&uiVersion=content-views. Accessed 2 Dec. 2024.

“Get the Data.” *Inside Airbnb*, insideairbnb.com/get-the-data/. Accessed 2 Dec. 2024.

“III.B. Overview of the State - District of Columbia - 2023.” *Health Resources and Services Administration*, mchb.tvisdata.hrsa.gov/Narratives/Overview/5ff83faa-6561-405a-bd59-a6c00d8c7cef. Accessed 2 Dec. 2024.

“What’s My Ward?” *DC.Gov*, planning.dc.gov/whatsmyward. Accessed 2 Dec. 2024.

Questions?

Appendix

Price vs Host Response Rate

Correlation Test

Null Hypothesis (H_0): There is no correlation between mean Airbnb listing price and host response rate

Alternate Hypothesis (H_A): There is a correlation between mean Airbnb listing price and response rate

Results

t=0.38, df = 3715, p-value = 0.70
95% CI: -0.026 to 0.038
R = 0.0063

P-value > 0.05, Fail to reject null hypothesis

The correlation test suggests **there is not enough evidence** to conclude a significant linear relationship between the mean listing price and host response rate.