Beds, Bedrooms, and Beyond: A Study of Airbnb Rental Prices

David Corcoran, Soong-Ping Hill, Jessica Joy, Adam Stein, Hung Tran

DSAN 5100 - Probabilistic Modeling and Statistical Computing

Georgetown University

December 11, 2024

Table of Contents

Section 1: Introduction..... 1

Section 2: Analysis..... 2

2.1 Data Collection..... 2

2.2 Data Cleaning..... 2

2.3 Exploratory Data Analysis (EDA)..... 4

2.4 Statistical Methods..... 11

Section 3: Results..... 13

3.1 Price vs Ward/Neighborhood..... 13

3.2 Price vs Ward Crime Rate..... 13

3.3 Price vs Superhost Status..... 14

3.4 Price vs Host Response Rate..... 15

3.5 Price vs Distance to City Center..... 17

3.6 Price Comparisons Across Major Cities..... 20

Section 4: Conclusion..... 21

Appendix..... 24

Section 1: Introduction

In a world shocked by inflation, consumers increasingly scrutinize costs when planning vacations. Many have turned to Airbnb, who have revolutionized the travel industry by often offering accommodations far more affordable than hotels, as a unique way to save or even earn additional income. The platform, founded in 2007, now represents a critical part of the vacation economy by serving over five million hosts who have welcomed over two billion guests into their homes (Airbnb Newsroom, *About Us*). However, Airbnb pricing varies widely and is influenced by factors such as amenities, home features, and location. Our goal is to understand how home and interior features drive Airbnb prices and to uncover whether neighborhood safety, measured through crime rates, can impact these values.

This analysis will focus on three major cities with millions of tourist foot traffic each year: Washington D.C., Boston, and Chicago. Each city offers distinct attractions, for example, Washington D.C. is a city of monuments and politics, Boston is home to an academic hub, and Chicago boasts rich architecture and culture. The combination of diverse Airbnb listings in these locations makes them ideal for comparative analysis. By uncovering patterns across these three cities, we aim to help potential travelers weigh the trade-offs between safety, comfort, and cost.

Additionally, Airbnb has become such a lucrative business that many people have invested in properties for the sole purpose of listing them on the site. Hosts have the goal of reaching "Superhost" status, a title given to Airbnb's most elite hosts who meet certain criteria. These include requirements for the number of reservations, response rate, overall rating, etc. (*What's required to be a Superhost - Airbnb Help Center,* Airbnb). Hosts are increasingly interested in what they can do to attract the most customers and ensure that they enjoy their stay.

Ultimately, this research aims to guide both Airbnb hosts and guests. Guests can make data-driven decisions on the best neighborhoods and stays based on their preferences of affordability, amenities, or safety, while hosts can understand what features of their properties will get them a higher return on their investment. Whether planning a vacation or looking to make some extra cash, this project offers valuable insights into how various factors shape the Airbnb experience.

Our research aims to answer the following five general 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 a listing's neighborhood influence Airbnb prices?*
5. *To what extent does distance from a city center play a role in Airbnb pricing?*

Section 2: Analysis

2.1 Data Collection

This project made use of three datasets sourced from Inside Airbnb, a website that hosts a plethora of data on Airbnb listings in over 120 cities worldwide. From this website, we pulled listings.csv.gz files for the three cities in question: Washington D.C., Boston, and Chicago. Each dataset had 75 columns that represent different features for each Airbnb listing, and an initial row count of 4928, 4325, and 7952, respectively. Additionally, one crime dataset for Washington D.C. was sourced from Open Data DC. We decided to explore the relationship between crime rate and Airbnb listings within just Washington D.C. because the city has fewer neighborhoods and its wards provide a natural framework for categorization. On the other hand, Boston and Chicago do not have a structure in which listings could be categorized into smaller groups. The Washington D.C. crime dataset was combined with city ward population counts to calculate the per capita crime rate per ward. For our analysis of Airbnb listings in Washington D.C., we created an original dataset to separate each neighborhood into the ward they are assigned to. Sorting the Washington D.C. data by ward allowed us to categorize each listing for the city into eight groups rather than 39.

2.2 Data Cleaning

Data cleaning is a vital step in data analysis, as raw data often contains errors, wrong data types and formats, missing data, duplicate entries, or outliers. Addressing these issues ensures that the data is accurate, consistent, and usable for meaningful analysis. Correcting these problems promotes data integrity by confirming that the data being used accurately reflects the population it's meant to represent. Without proper data cleaning, analyses can produce misleading results, leading to faulty insights and poor decision-making.

The first step in our data cleaning process involved merging datasets and creating additional data features. We combined the listings and ward datasets on their mutual neighborhood name column, ensuring that all Airbnb listings had their respective wards identified within our data. Next, we created four new data features: distance_to_city_center, distance_category, crime_rate_per_capita, and crime_rate_category. The distance_to_city_center column represents the Euclidean distance from a set longitude and latitude position; distance_category buckets each listing into a near, medium, or far category; crime_rate_per_capita was calculated by dividing total ward crime counts by populations; and crime_rate_category bins each listing into a crime rate categories of low, medium, or high.

After consolidating datasets, we removed data that was unrelated to our testing, starting with columns. This procedure dropped our total column count down from 75 to 19, guaranteeing that only relevant data like price, host status, location, and property details were retained. We then addressed missing data by excluding rows with NA values in all columns, again ensuring the dataset's integrity for analysis. The last part of the data removal step involved removing extreme price outliers, so that the dataset could more accurately reflect typical property prices and minimize the influence of anomalies. An interquartile range (IQR) method was used to calculate the difference between the 25th (Q1) and 75th (Q3) percentiles of the pricing data. Using this range, we established an upper bound of $Q3 + 1.5 \times IQR$ and identified outliers as price values exceeding this threshold. We did not remove rows in which the listing price was below the lower bound, as the price of an Airbnb listing cannot be negative. Upon filtering, rows with price values above the upper bound were removed from the dataset, confirming that extreme values will not skew the analysis.

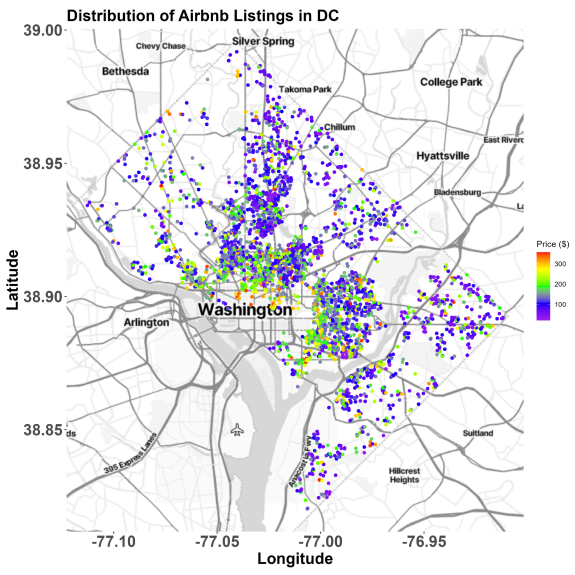
The final step in our data cleaning process was to standardize columns. To standardize, dollar signs and commas were removed from the price column, percentage signs were removed from the host_repsonse_rate column, and all values in both columns were converted to a numeric data type. As a result of our data cleaning, the Washington D.C., Boston, and Chicago datasets were reduced to 19 columns and 3261, 2506, and 5547 rows, respectively. Details of each feature in our datasets are listed in Appendix Table A.1.

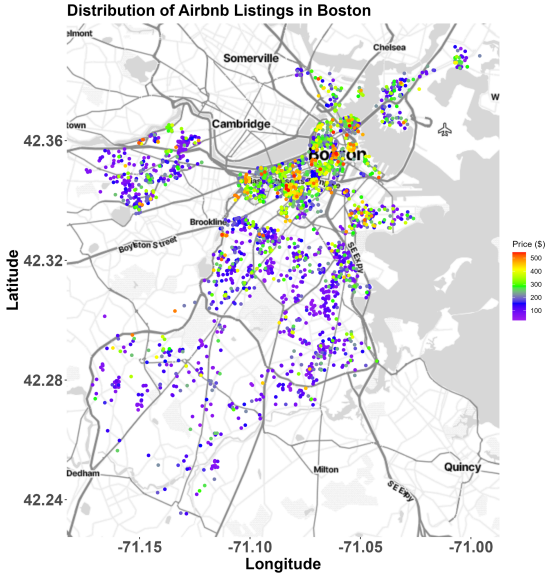
2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step prior to any statistical testing in understanding the factors influencing Airbnb pricing. By systematically exploring the dataset, we can uncover patterns, trends, and relationships that provide valuable insights into rental prices.

Geospatial plots were created to visualize the geographical distribution of Airbnb listings in our selected cities. Initially, blank maps were generated using Stamen Map tiles in combination with the Stadia Maps API. Stamen Map tiles defined the map style, while the Stadia Maps API allowed us to generate maps centered on specific locations using longitude and latitude values. To make use of Stadia Maps and its Stamen Map tiles, an API key must first be obtained. Once the blank maps were created, Airbnb listing data points were overlaid onto them using the ggmap R package, which integrates base maps with geospatial data visualizations.

A series of geospatial plots showcases the geographical distribution of Airbnb listings in Washington D.C., Boston, and Chicago, and is illustrated in Figure 1. These plots display the geographical distribution of non-outlier listings by visualizing the locations of Airbnb listings, with color gradients representing the range of prices for the listings in each city. The color scale indicates varying price levels, where darker shades correspond to lower prices and brighter hues represent higher-priced listings. Regardless of city, these geospatial plots reveal that higher-priced Airbnb listings are typically clustered closer to a city's center. This pattern suggests that listings in central locations tend to demand higher nightly rental prices compared to those in more distant neighborhoods, likely due to proximity to major attractions or commercial districts.





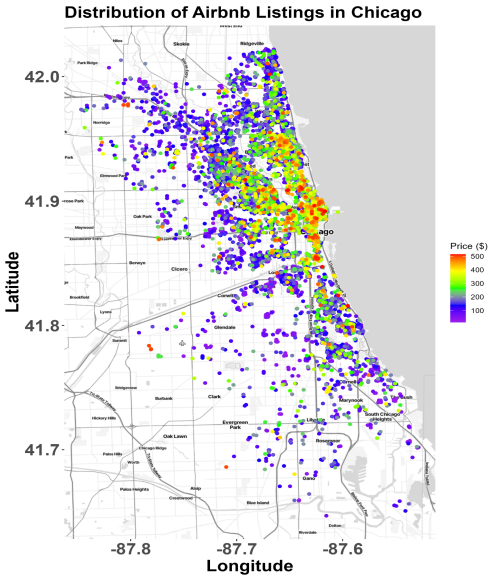
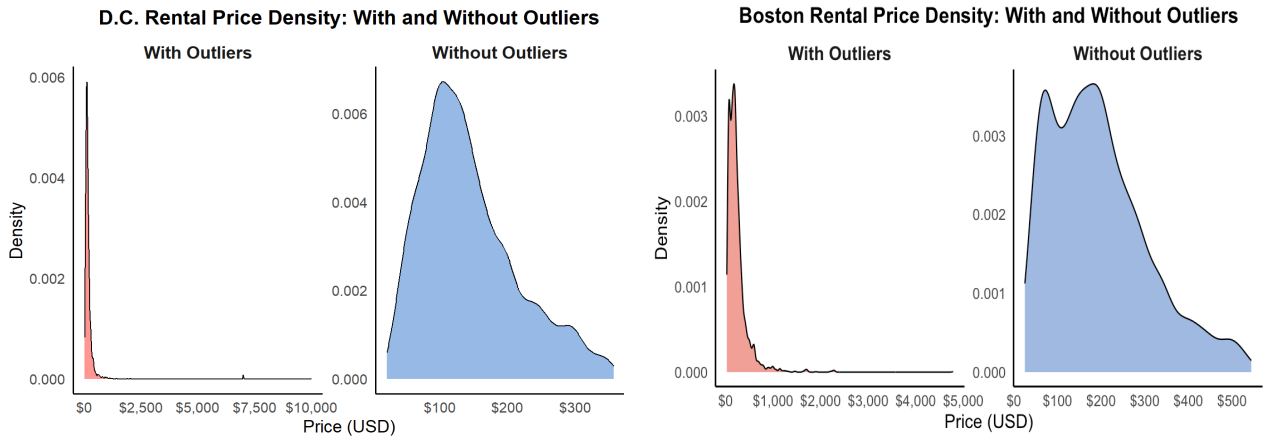


Figure 1: Geospatial distribution of Airbnb listings per city

Price Distribution Per City

To understand the distribution of prices across rental listings, we looked at the density of rental prices for all rentals across the three cities. Figure 2 shows density plots of rental prices for Washington D.C., Boston, and Chicago with and without outliers. The majority of rentals are priced between $50 and $350, $50 and $300, and $50 and $300 a night for D.C., Boston, and Chicago, respectively. All three cities have significant outliers, with D.C. showing the maximum rental price of $10,000 a night. The presence of large outliers in all three plots would skew our testing results, so we excluded outliers during this process (see section 2.4).



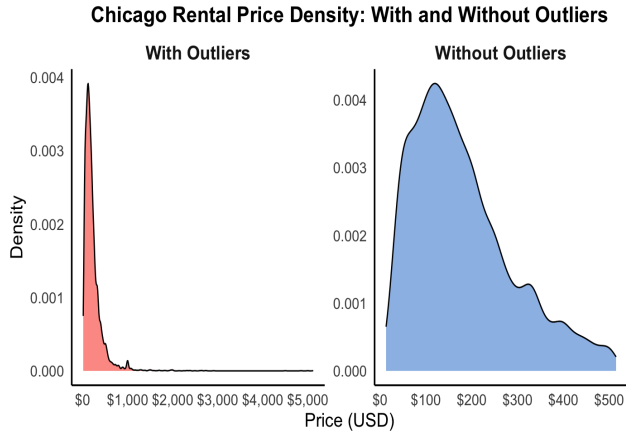
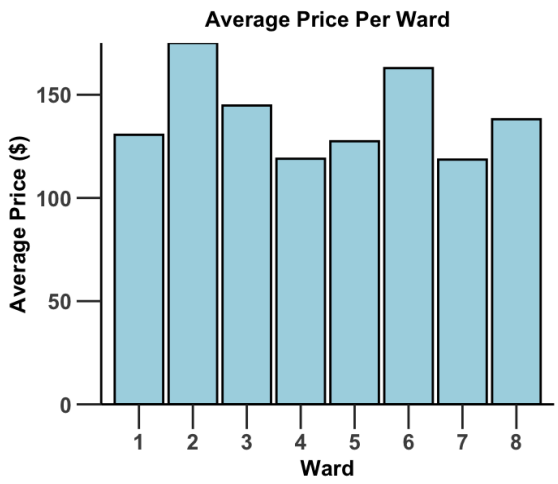
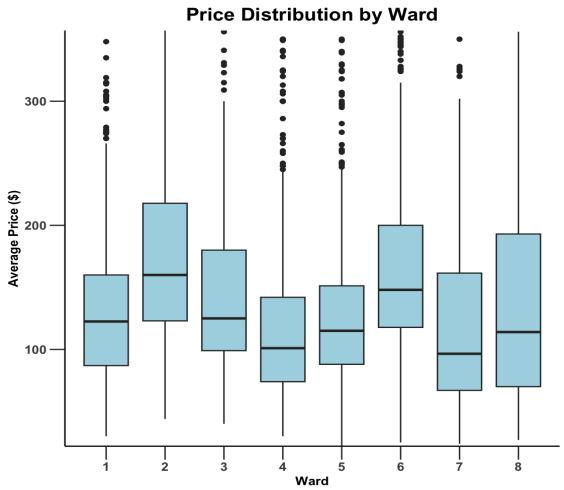


Figure 2: Rental price density with and without outliers per city

Feature Pairings: Variables vs Price

Next, we looked at how price varies with different features of our dataset. Specifically, we wanted to visualize how price is distributed by neighborhood/ward, superhost status, property type, property features, and room type. In D.C., the barplot of price by ward showed the highest average price in Ward 2 and the lowest in Ward 4. In Boston, the highest priced neighborhood was Longwood Medical Area (just south of Fenway Park) with an average price of over $400, and the lowest was Mattapan with an average price of a little over $100. In Chicago, average prices are highest near the North Side and lowest in the East Side. For D.C., we also include a boxplot that visualizes the distribution of average Airbnb prices across different wards. Wards 2 and 8 exhibit higher median prices compared to others, indicating a potential concentration of more expensive listings in these areas. Wards 1, 2, and 8 show wider interquartile ranges (IQR), reflecting greater price variability within these wards compared to others like Ward 5, which has a narrower IQR.





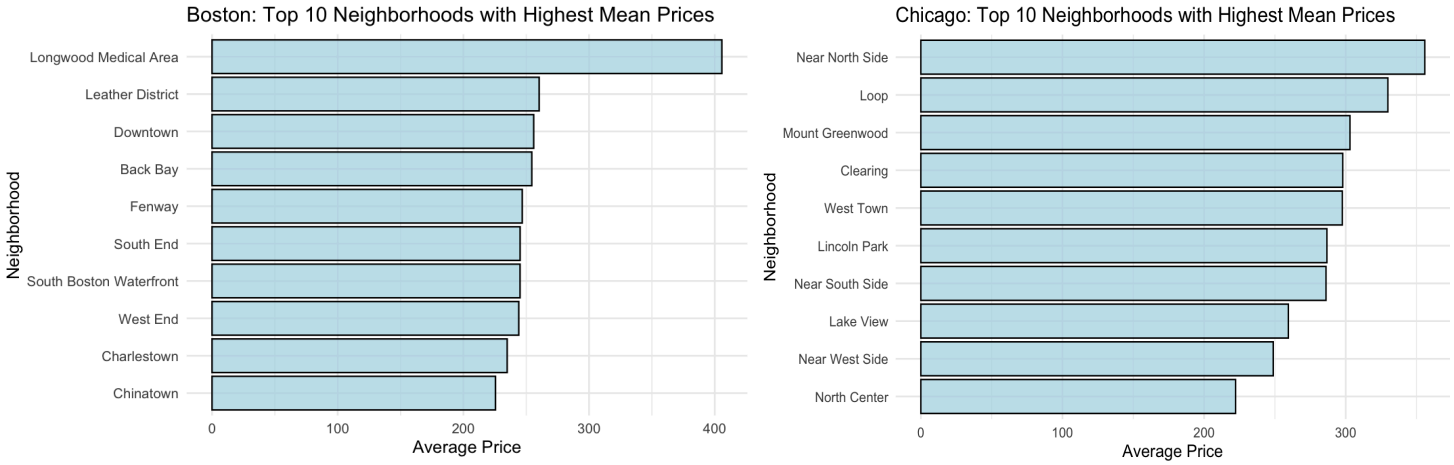
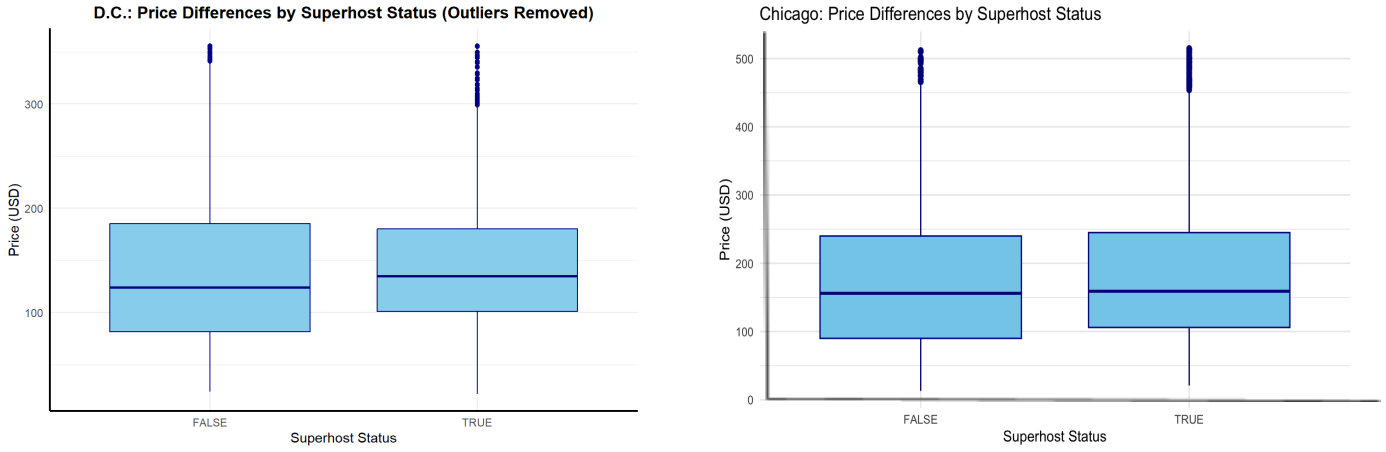


Figure 3: Average prices by neighborhood/ward

Comparing prices by superhost and non-superhost statuses in Figure 4, we see that in D.C. superhosts tend to charge slightly higher prices than regular hosts, suggesting that their higher status may allow for modest premium pricing. Listings by superhosts exhibit more outliers, indicating a broader range of high-end listings. Chicago is similar to D.C., where superhost median price is $159 per night whereas non-superhost is $156 per night. However, our findings differed in Boston where superhosts had a lower median price of $162 per night compared to the non-superhost average rate of $184 per night.



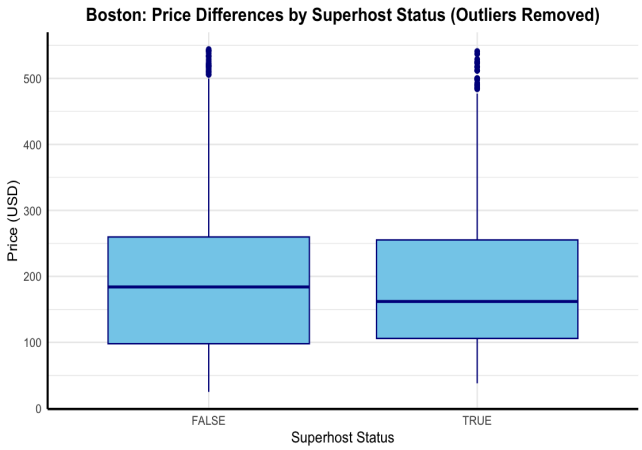


Figure 4: Distribution of prices by superhost and non-superhost status for Washington D.C., Boston, and Chicago. “FALSE” indicates non-superhost status whereas “TRUE” indicates that host is a superhost.

To analyze how prices differ by property type, we aggregated similar property types into broader categories since there are more than 20 different types of properties. The categories are Apartment, House, Private Room, Shared Room, and Other. The first four categories are the most frequent property types while the Other category includes less common properties like boats, bungalows, and other properties that do not fit neatly into the main four categories. Figure 5 depicts the distribution of prices by property type for Washington D.C., Boston, and Chicago. In D.C., the boxplot of prices by property types reveals that median prices are similar across properties, with Apartments having the highest median price and Shared Rooms having the lowest. This is expected since having your own room is typically pricier than sharing a room. In Boston, the Other category has the highest median price, followed by Apartments. Shared Rooms had the lowest median price, however, this is based on only one listing matching so this was not a good representative and sample size of the category. Similarly, in Chicago, the Other category possesses the highest median price while Shared Rooms have the lowest. It is also important to note that in D.C. and Chicago, houses have the largest IQR suggesting more variability in average price.

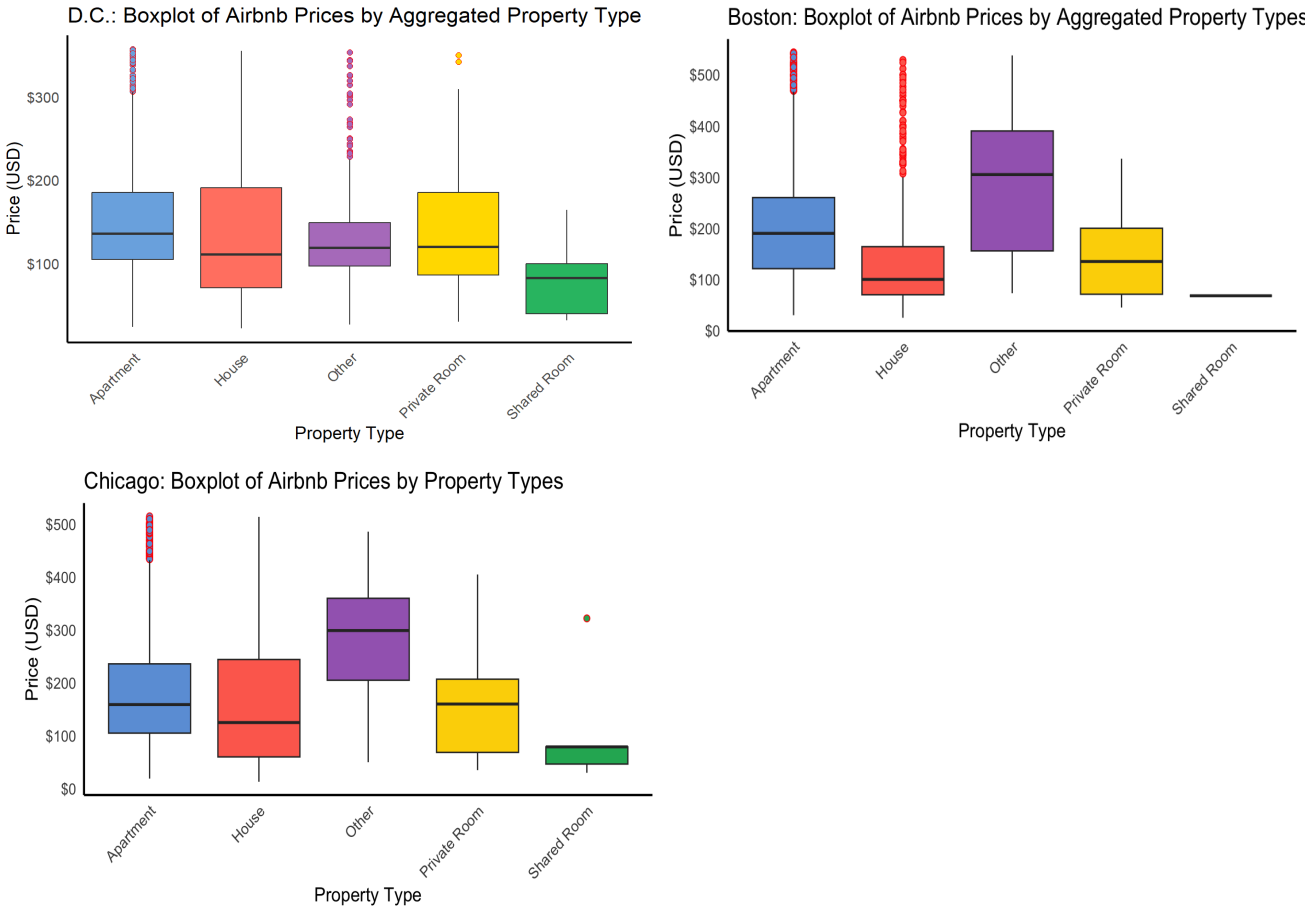


Figure 5: Distribution of prices by property type for Washington D.C., Boston, and Chicago

Finally, we analyzed the distribution of prices by room type. The four room types are an entire home or apartment, a hotel room, a private room, and a shared room. We expect that the entire home or apartment would have the highest price, followed by a hotel room, private room, and shared room. The distribution of prices by room type for Washington D.C., Boston, and Chicago is shown in Figure 6. The boxplot of prices by room type for D.C. confirms this expectation but differs for Boston and Chicago. For Boston, the most expensive room type was hotel rooms, followed by the entire home, while the lowest priced room type was shared room. Private rooms had many outliers compared to other room types. Similarly in Chicago, hotel rooms are the priciest. Private rooms also had the most outliers which could be caused by the type of home or apartment the private rooms are in, especially those located in more central and desirable areas. A private room in a mansion would presumably cost more than a private room in an apartment.

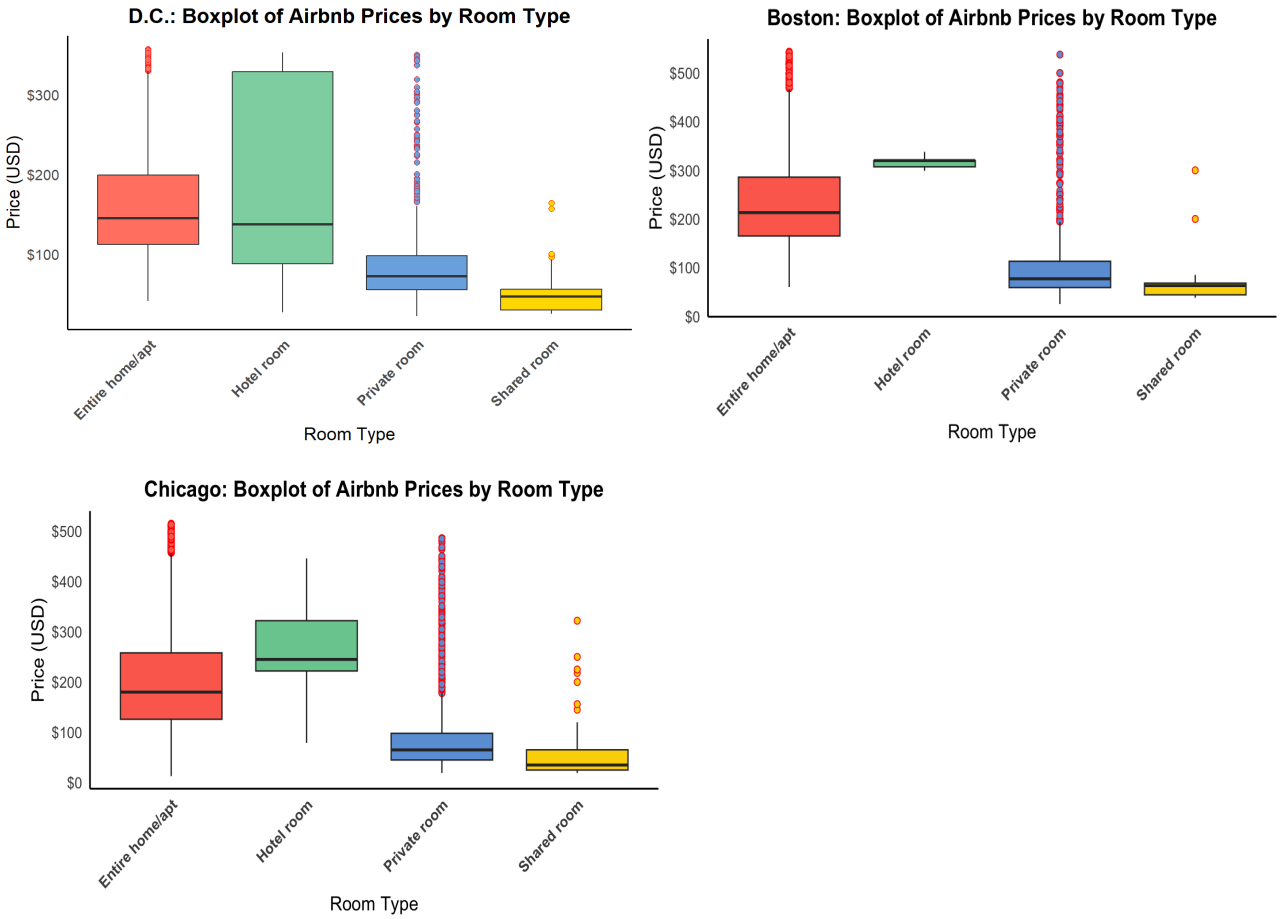


Figure 6: Distribution of prices by room type for Washington D.C., Boston, and Chicago

Crime by Neighborhood:

To understand how price and ward crime rates are related, we needed to calculate crime rates. Since the data for D.C. was the total number of crimes, to get a more accurate comparison of crime, we calculated a per capita crime rate for each ward based on population. Figure 7 shows the population and per capita crime rate for each ward. Wards 1, 2, and 7 have the highest crime rates, whereas Ward 3 has by far the lowest crime rate.

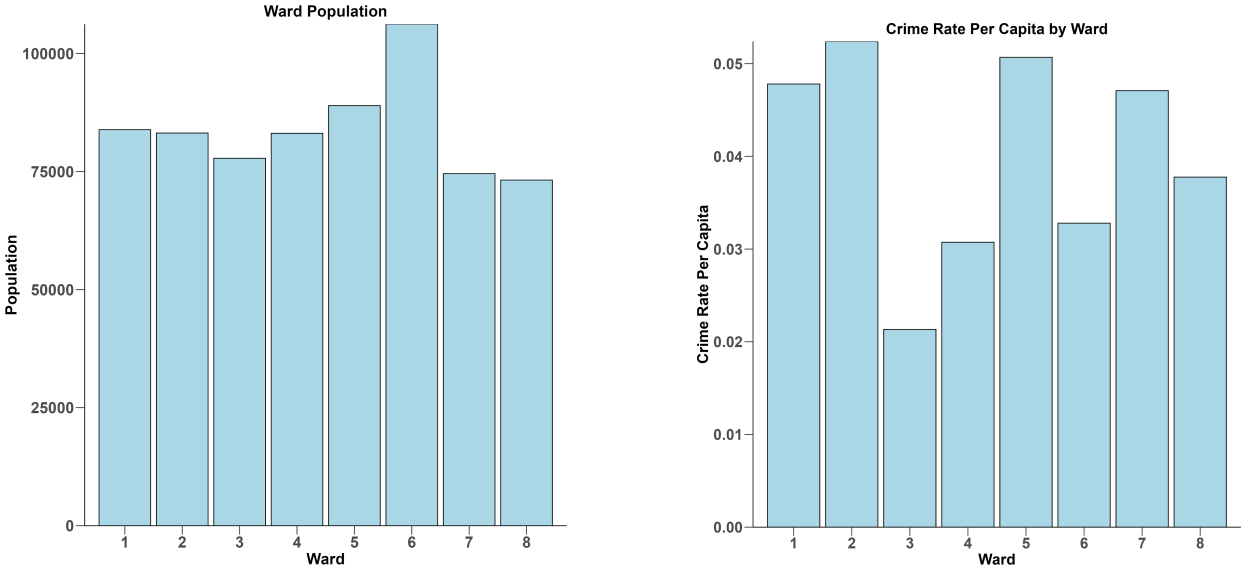


Figure 7: Washington D.C ward population and crime rate per capita

2.4 Statistical Methods

To determine the significance of the relationships, associations, or correlations between features in our project, we opted to perform a variety of hypothesis tests. In general, hypothesis tests use null and alternative hypotheses to affirm whether relationships in a population can be affirmed by sample data, or if it is more likely that they exist by chance. The five types of hypothesis tests that were conducted to answer our data science questions regarding Airbnb listings are two sample t-tests, analysis of variance tests (ANOVA), Tukey's Honest Significant Difference (HSD) Tests, correlation tests, and chi-squared tests of independence.

Two sample t-tests determine whether the means of two independent groups statistically differ from one another. T-tests produce a p-value that, when lower than 0.5, indicates the null hypothesis should be rejected. In our project, two sample t-tests were used to determine whether superhost status has a significant impact on mean listing price. In other words, they tested whether the mean listing prices of superhost listings are statistically different from the mean listing prices of non-superhost listings.

An ANOVA is a statistical method used to compare the means of more than two groups to determine if at least one group's mean differs from the rest. The test examines whether sample means provide evidence of differences in population means. ANOVA tests produce an F value, which measures the ratio of variability between groups to the variability within groups, and a

p-value, which measures the probability that the calculated F-value would occur under a null hypothesis. F value scores are computed with the formula $F = \frac{Mean\ Square\ Between\ Groups\ (MSB)}{Mean\ Square\ Within\ Groups\ (MSW)}$. MSB is calculated with the formula $MSB = \frac{Sum\ of\ Squares\ Between\ Groups\ (SSB)}{Degrees\ of\ Freedom\ Between\ Groups\ (dfB)}$, with $dfB = k - 1$, where k is the number of groups. MSW is calculated with $MSW = \frac{Sum\ of\ Squares\ Within\ Groups\ (SSW)}{Degrees\ of\ Freedom\ WIthin\ Groups\ (dfW)}$, with $dfW = n - k$, where n is the total number of observations and k is the number of groups. A high F value indicates that the variation between groups is significantly larger than the variation within groups. A p-value lower than 0.5 indicates that the null hypothesis should be rejected.

Tukey's HSD test is a post-hoc analysis that provides comparisons between different levels of a factor after conducting an ANOVA test. It determines specifically which groups have significant differences in means. In our project, we conducted ANOVA tests, followed by Tukey's HSD post hoc analyses, to explore the relationships between listing price and neighborhood/ward, listing price and neighborhood/ward crime rates, listing price and distances to city centers, and listing prices among major cities.

Correlation tests assess the relationship between two or more variables by producing a correlation coefficient ranging from -1 to 1. A positive coefficient indicates a positive correlation, while a negative coefficient signifies a negative correlation. The strength of the correlation is considered strong if the coefficient falls between -1 and -0.5 or 0.5 and 1, and weak if it lies between -0.5 and 0.5. In this project, correlation tests were conducted to examine the relationship between an Airbnb host's response rate and their rental listing price.

Chi-squared tests of independence evaluate whether two categorical variables are independent or associated with each other. The test uses observed and expected frequencies to calculate a chi-squared statistic, which is then used to determine a p-value. A p-value lower than 0.5 indicates that the null hypothesis should be rejected. In our project, chi-squared tests of independence were conducted to determine whether there is an association between superhost status and room type.

Section 3: Results

3.1 Price vs Ward/Neighborhood

To determine if there is a statistically significant difference between price and neighborhood/ward, we conducted an ANOVA test. The null and alternative hypotheses are listed below:

H_0: There is no significant difference in mean Airbnb listing prices across wards/neighborhoods

H_A: There is at least one ward/neighborhood that significantly differs in mean Airbnb listing price

Table 1: ANOVA test results for evaluating mean listing price and city neighborhood/ward

City	F value	p-value
Washington D.C	54.62	2.00×10^{-16}
Boston	33.38	2.00×10^{-16}
Chicago	25.91	2.00×10^{-16}

The results of each test are shown in Table 1. The high F values for all three cities indicate that the variation in the mean Airbnb listing prices between neighborhoods/wards is much larger than the variation within the neighborhoods/wards. Similarly, all three p-values are significantly smaller than the threshold of 0.05, indicating that the null hypothesis should be rejected and that at least one neighborhood/ward has a mean Airbnb listing price that is statistically significantly different from the others. Overall, this ANOVA test concludes that there is a meaningful relationship between Airbnb listing price and the neighborhood/ward in which the rental is located.

3.2 Price vs Ward Crime Rate

To evaluate whether crime rate has a significant effect on Airbnb listing price, the crime rate was calculated for each Washington D.C. ward by dividing the total count of crime in each

ward by the respective ward's population. This crime rate per capita value was then used to map each listing to a crime rate category of "low," "medium," or "high," where "low" wards corresponded to the bottom third of crime rate values, "medium" to the middle third (33rd to 66th percentile), and "high" to the top third of values. An ANOVA test was conducted to analyze the difference in mean rental prices among the categories. The null and alternative hypotheses are listed below:

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

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

Table 2: ANOVA test results for evaluating mean listing price and crime rate categories

	F value	**p-value**
Crime Rate Category	22.18	2.64×10^{-10}

The results of the ANOVA test are displayed above in Table 2. The ANOVA produced an F value of 22.18 and a p-value of 2.64×10^{-10}. The high F value of 22.18 indicates that the variation in the mean Airbnb listing prices between the crime rate categories is much larger than the variation within the crime rate categories. Similarly, the p-value of 2.64×10^{-10} is significantly smaller than the threshold of 0.05, indicating that the null hypothesis should be rejected and that at least one of the crime rate categories has a mean Airbnb listing price that is statistically different from the others. Overall, this ANOVA test concluded that there is a meaningful relationship between Airbnb listing price and the crime rate for the area in which an Airbnb rental property is located.

3.3 Price vs Superhost Status

To evaluate whether superhost status influenced Airbnb listing prices, data was filtered to separate superhosts from non-superhosts. To test if there was a statistically significant difference between superhost status and non-superhost status, a two sample t-test was conducted.

Comparing those who are superhost and non-superhost is important as it can provide insights into how verification of trust from the host can affect prices and attract potential guests. The null and alternative hypotheses for all three cities are stated below:

H_0: There is no significant difference in mean Airbnb listing prices between superhost and non-superhost listings.

H_A: The mean Airbnb listing price is greater for superhost listings than non-superhost listings.

Table 3: One-tailed t-test results of price and superhost status in three cities

City	t-value	p-value	95% Confidence Interval	Mean Difference
Washington D.C	-2.59	0.0049	(-Inf, -2.17]	-5.98
Boston	1.01	0.31	[-3.96, 12.44]	4.24
Chicago	-4.55	2.68×10^{-6}	[-Inf, -7.68]	12.02

The results of the T-tests are displayed above in Table 3. For both Washington, D.C. and Chicago, we reject the null hypothesis and conclude that the mean listing prices by non-superhosts are less than those of superhosts on Airbnb. From our confidence interval, we are 95% confident that the true mean difference in rental prices between superhosts and non-superhosts is less than -2.17 (i.e., true mean difference in rental prices between regular hosts and superhosts is at least $2.17). In Chicago, the difference is larger as we are 95% confident that the true difference in mean prices is at least $7.68.

For Boston, we fail to reject the null hypothesis with a computed p-value of 0.31. There is not enough evidence to suggest the average listing prices of non-superhosts is less than superhtst. The calculated 95% confidence interval of [-3.96, 12.44] contains zero, supporting the null hypothesis of there being no significant difference in mean Airbnb listing prices between superhost and non-superhost listings.

3.4 Price vs Host Response Rate

To assess if the host response rate is correlated with the price of listing, a correlation test was conducted. This can provide key insights into how the behavior of the host may influence their listing price. Host Response Rate is an indicator of trust and respect for the customer, and the customer may be willing to spend more for listings with a reliable host. These insights can be utilized by both Airbnb and the host to promote higher responsiveness, understand the priorities of the guest, and give hosts a competitive advantage. The null and alternative hypotheses for the correlation test for all three cities are stated below:

H_0: There is no correlation between Airbnb listing prices and Host Response Rate

H_A: There is a correlation between Airbnb listing prices and Host Response Rate

Table 4: Correlation test results of price and host response rate in three cities

City	t-value	p-value	95% Confidence Interval	Correlation Coefficient (r)
Washington D.C.	0.38	0.70	[-0.026, 0.038]	0.0063
Boston	2.22	0.027	[0.0047, 0.076]	0.04
Chicago	8.75	2.2×10^{-16}	[0.082, 0.13]	0.11

The results of the correlation tests are displayed above in Table 4. For Washington D.C., since the p-value is greater than 0.05, we fail to reject the null hypothesis in favor of the alternative hypothesis. From our confidence interval, we are 95% confident that the true correlation between prices and host response rate is between -.026 and 0.038. Zero lies within our 95% confidence interval, providing further evidence that we should not reject the null hypothesis. Thus, it is unlikely there is a correlation between host response rate and rental price.

For Boston and Chicago, the p-values were less than 0.05, so we reject the null hypothesis and conclude that there is a statistically significant correlation between price and host response rate. For Boston, we are 95% confident that the true correlation coefficient lies 0.0047 between 0.0760, and for Chicago, between 0.0819 and 0.1288. The correlation coefficients for

both cities exhibit a very weak, positive relationship, suggesting that there is not much significance.

3.5 Price vs Distance to City Center

To assess how the location of Airbnb listings influences listing price, we evaluated the listing's distance to the city center. A city's center is often an area with many attractions and a sought-after spot for tourists. To create this metric, we utilized the data's longitude and latitude columns, calculating an Euclidean distance approximation from each listing to the longitude and latitude of each city's center. The formula is as follows:

$$Distance = \sqrt{(CityCenterLatitude - ListingLatitude)^2 + (CityCenterLongitude - ListingLongitude)^2}$$

Since this metric does not factor in the curvature of the earth, we acknowledge that it is an approximation of the distance. The center of the city was selected as the Ellipse in D.C., the Loop in Chicago, and City Hall in Boston. The distribution of distances was grouped into three bins, "Near", "Medium" and "Far", allowing us to perform an ANOVA test, and analyze the difference in mean listing prices among the proximity groupings. The null and alternative hypotheses are listed below:

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

H_A: There is at least one proximity to the city center category (Near, Medium, Far) in which the mean Airbnb listing price significantly differs from the rest

Table 5: ANOVA test results for the mean listing price and distance to city center categories

City	F value	p-value
Washington D.C.	71.33	$< 2 \times 10^{-16}$
Boston	32.16	$< 1.48 \times 10^{-14}$
Chicago	305.3	$< 2 \times 10^{-16}$

The results of the ANOVA test are displayed above in Table 5. The ANOVA of all three cities produced large F values and p-values of less than 1.48×10^{-14}. These high F values indicate

that the variation in the mean Airbnb listing prices between the distance categories is much larger than the variation within the categories. Likewise, the small p-values are less than our threshold of 0.05, indicating that we can reject the null hypothesis and conclude that at least one of the distance categories (Near, Medium, Far) has a mean Airbnb listing price that is statistically different from the others. This ANOVA test suggests a significant relationship between Airbnb listing price and the proximity of the listing to the city center in Washington D.C., Boston, and Chicago.

To dive deeper into the difference between these distance categories, we performed Tukey's Honestly Significant Difference post hoc analyses.

3.5.1 Washington D.C.

Table 6: Tukey's HSD for distance to city center categories for Washington D.C.

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

All pairwise comparisons in Table 6 have p-values below 0.05, indicating statistically significant differences between them. In addition, since none of the confidence intervals (provided by the Lower and Upper columns) include 0, we can again conclude that the difference between these distance bins is statistically significant. On average, medium distance listings are $23 less than near listings, far listings are $37 cheaper than near listings, and far listings are $14 cheaper than medium listings. It is clear that in Washington D.C., prices decrease as listings get further from the city center. This information is valuable for a tourist who may be willing to trade longer travel time to attractions for a more affordable Airbnb.

3.5.2 Boston

Table 7: Tukey's HSD for distance to city center categories for Boston

Distance Bins	Difference	Lower	Upper	p-value
Medium-Near	-1.14	-12.24	9.95	0.96
Far-Near	-33.10	-44.11	-22.08	0.00
Far-Medium	-31.96	-42.97	-20.94	0.00

Analyzing the price comparisons between distances for Boston in Table 7, we see that the difference between Medium and Near listings has a p-value of 0.96, so we are unable to reject the null hypothesis. The confidence interval for Medium - Near also includes 0, providing more evidence to keep the null hypothesis that there is not a statistically significant difference in the means. The difference between Far-Near bins and Far-Medium bins has a p-value of 0, allowing us to reject the null hypothesis, indicating there are statistically significant differences. According to the difference column, on average, Far listings are $33 cheaper than Near listings, and Far listings are $31 cheaper than Medium listings. It is clear that when comparing listings Near - Far and Medium - Far, the prices are statistically different, but Medium - Near does not have this relationship. This is valuable for tourists trying to find the optimal place to stay while balancing price and distance to the city center. Travelers should take into account that there may not be a significant reduction in their cost by choosing a property of medium proximity to the center, they can just choose a closer option.

3.5.3 Chicago

Table 8: Tukey's HSD for distance to city center categories for Chicago

Distance Bins	Difference	Lower	Upper	p-value
Medium-Near	-63.54	-69.81	-57.28	0.000
Far-Near	-65.72	-81.12	-50.31	0.000
Far-Medium	-2.17	-17.96	13.61	0.94

Two of the three pairwise comparisons in Table 8 show p-values below 0.05, indicating statistically significant differences between them. The confidence intervals for Far-Medium include 0, so we fail to conclude that the difference between these distance bins is statistically

significant. However, for Medium-Near and Far-Near relationships, there are statistically significant differences, so we can reject the null hypothesis. The Difference column shows the mean price difference for the three distance categories. On average, medium distance listings are $64 less than near listings, and far listings are $66 cheaper than near listings. We can say that prices decrease as listings get further from the city center, but there seems to be less of a distinction between Medium and Far listings. Perhaps, for visitors who would like to stay closer to the center, Medium listings provide a better value proposition as they do not seem to cost much more than Far listings.

3.6 Price Comparisons Across Major Cities

To compare how Airbnb prices differed between cities, a comparative analysis of mean listing prices was done between Washington D.C., Boston, and Chicago. An ANOVA test was conducted to analyze any significant difference between mean listing prices among these three cities. The null and alternative hypotheses are below:

H_0: There is no significant difference between mean Airbnb listing price and city

H_A: There is a significant difference between mean Airbnb listing price and city

Table 9: ANOVA test results for mean listing price across Washington D.C., Chicago, Boston

	F value	**p-value**
Price Between Cities	113.9	$< 2 \times 10^{-16}$

The results of the ANOVA test are displayed above in Table 9. The ANOVA produced an F value of 113.9 and a p-value of less than 2×10^{-16}. This very high F value indicates that the variation in rental prices between cities is significantly larger than the variation within cities. Furthermore, the p-value of less than 0.05 indicates that we can reject the null hypothesis, concluding that at least one of the three cities has a mean Airbnb listing price statistically different from the others. This ANOVA test suggests that there is evidence of a difference in average Airbnb prices among the three different cities (Boston, Washington D.C., Chicago).

Table 10: Tukey's HSD Test of Average Price Differences Between Cities

Distance Bins	Difference	Lower	Upper	p-value
Chicago vs Boston	-11.18	-16.08	-6.27	3×10^{-7}
Washington D.C. vs. Boston	-33.46	-38.91	-28.02	0.00
Washington D.C. vs Chicago	-22.29	-26.84	-17.73	0.00

To evaluate the inter-city differences, we performed Tukey's Honestly Significant Difference post hoc analysis. All pairwise comparisons in Table 10 show p-values below 0.05, indicating statistically significant differences between them. In addition, none of the confidence intervals include 0, indicating that the difference between these cities is statistically significant. On average, Chicago listings are $11 cheaper than Boston listings, Washington D.C. listings are $33 cheaper than Boston listings, and Washington D.C. listings are $22 cheaper than Chicago. This shows that prices differ significantly among these three cities, with Boston being the most expensive and Washington D.C. being the cheapest. This information is valuable as it provides insights into the significant price differences between these three cities. This provides travelers with data on which cities would suit their needs and budget.

Section 4: Conclusion

Overall, our statistical analysis allowed us to answer the following questions:

- What factors most strongly influence the price of an Airbnb listing?
- Are rental prices of Superhosts greater than those of regular hosts?
- How do Airbnb prices differ between major cities?
- How does the crime rate of a listing's neighborhood influence Airbnb prices?
- To what extent does distance from a city center play a role in Airbnb pricing?

We rejected the null hypothesis in favor of the alternative hypothesis when looking at neighborhood/ward and distance to the city center. These factors most strongly influence the price of an Airbnb. Superhost status was moderately significant for D.C., but not for Boston or Chicago. Thus, more analysis of other cities would be needed to see if superhost status is a key factor in rental price. We found that there are statistically significant differences in mean rental

prices across the three crime categories, but further research is needed to conclude if the differences in ward crime affect rental prices. We did not find the host response rate to be a significant factor. These results were found for all three cities making us more confident in our findings.

When comparing Washington D.C., Boston, and Chicago, we found Washington D.C. to have the lowest median price but also the lowest spread. Chicago had the highest median price and the highest spread. While the differences in prices were small, we did find the differences were statistically significant in all pairwise comparisons between the cities. The true mean differences are likely no larger than $20 which means that the cities are comparable in prices, with D.C. having the most outliers and the highest maximum price. If a vacationer is looking for the most expensive and fanciest rental, Washington D.C. would be the best choice.

In light of rising inflation and heightened consumer attention to travel costs, our analysis provides key insights into the complex factors that shape Airbnb pricing. By examining home and interior features, neighborhood safety, and host status across three diverse cities – Washington D.C., Boston, and Chicago – we were able to cross-validate our findings and uncover actionable patterns for both guests and hosts.

For guests, our findings offer a data-driven guide to balancing affordability, amenities, and safety when selecting a listing, helping them make informed decisions that align with their travel priorities. For hosts, the analysis highlights how features like amenities and achieving Superhost status can significantly impact pricing. The premium pricing and consistency associated with Superhost listings, for example, underscore the value of achieving this distinction to maximize income.

As Airbnb continues to empower the vacation economy, understanding the interplay of these factors enables both travelers and hosts to navigate this dynamic marketplace effectively, ensuring better experiences and outcomes for all.

Citations

"About Us." *Airbnb Newsroom*, 22 Oct. 2024, news.airbnb.com/about-us/.

"What's Required to Be a Superhost - Airbnb Help Center." *Airbnb*, www.airbnb.com/help/article/829.

"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.

Stadia Maps, Stadia Maps, Inc., 2024, stadiamaps.com/.

"What's My Ward?" DC.Gov, planning.dc.gov/whatsmyward. Accessed 2 Dec. 2024.

Appendix

Table A.1: Dataset Attributes and Descriptions

Attribute	Data Type	Description
id	num	Unique ID for the listing
host_id	num	Unique ID for the host
host_response_rate	num	Percentage of inquiries responded to within 24 hours in the past 30 days.
host_is_superhost	chr	String value for if the host is a superhost (f or t)
neighbourhood_cleansed	chr	Name of the neighborhood the listing is in
property_type	chr	Property type of the listing
room_type	chr	Room type of the listing
accommodates	int	Maximum number of people the listing can hold
bathrooms	num	Number of bathrooms
bedrooms	int	Number of bedrooms
beds	int	Number of beds
amenities	chr	List of available amenities
price	num	Price of the listing
longitude	num	Longitude of the listing
latitude	num	Latitude of the listing
ward	int	Ward that the listing is in
distance_to_city_center	num	Distance (degrees) from the listing to the city center
distance_category	chr	Category of distance (near, medium, far) the listing is away from the city center
crime_rate_category	chr	Level of crime in which the listing is located (low, medium, high)

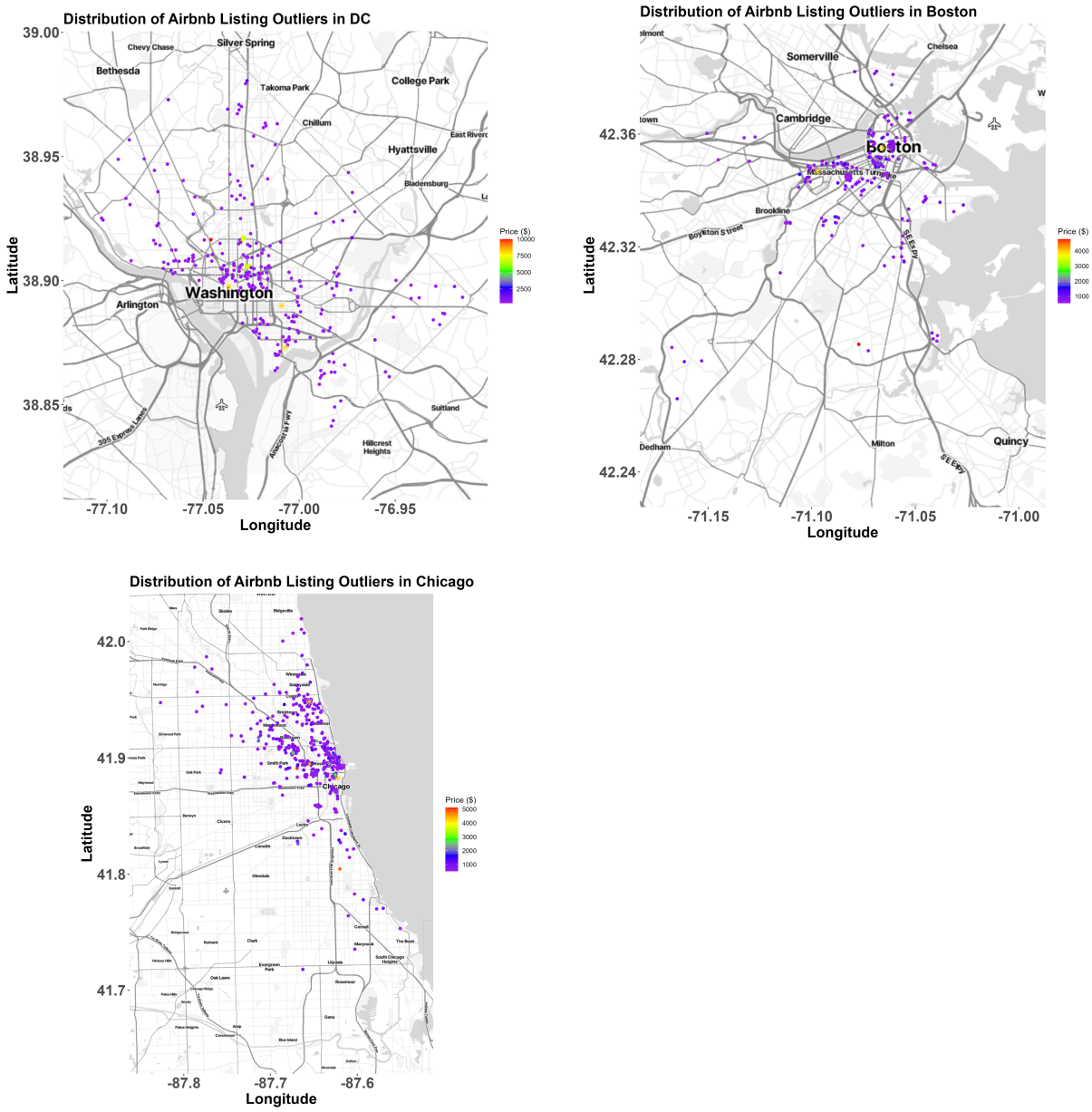


Figure A.1: Geographical distribution of Airbnb listing outliers per city

Price vs Property Features: Our analysis then focused on how price differs by various property features. This includes the number of people that the rental can hold (accommodates), bedrooms, beds, and bathrooms. We expected that as the number of property features increases (i.e., the more beds, bedrooms, bathrooms, etc.) a rental has, the more the rental should cost. We see this expectation mostly hold, except for a few interesting cases. In D.C., the prices for accommodates

increase up to 10 people, and then decrease with 11 people. Prices increase from 0 to 4 bathrooms and then decrease for 4.5 and 5 bathrooms. 5, 7, and 8 bedroom properties have lower median prices than even 0 bedrooms. For beds, the prices start to decrease after 9 beds. These results are surprising, but are likely due to there being few properties having such large numbers of these features (For example, not many properties can accommodate more than 10 people, have more than 4 bedrooms, or more than 4 bathrooms). Few rentals in these categories would greatly skew the results if there is an outlier. In Boston, the prices for accommodates also increase up to 10 people, decrease with 11 people, then fluctuate from 12 - 15. Prices also increase from 0 to 4 bathrooms and then decrease for 4.5 and 5 bathrooms. Properties with 5 and 6 bedrooms properties have lower median prices than even 0 bedrooms, but 7 bedrooms have the highest price. The prices start to remain stagnant after 5 beds, except at 9 beds, where it is the maximum. In Chicago, the price increases as each of the variables, accommodates, bathrooms, bedrooms and beds increase.

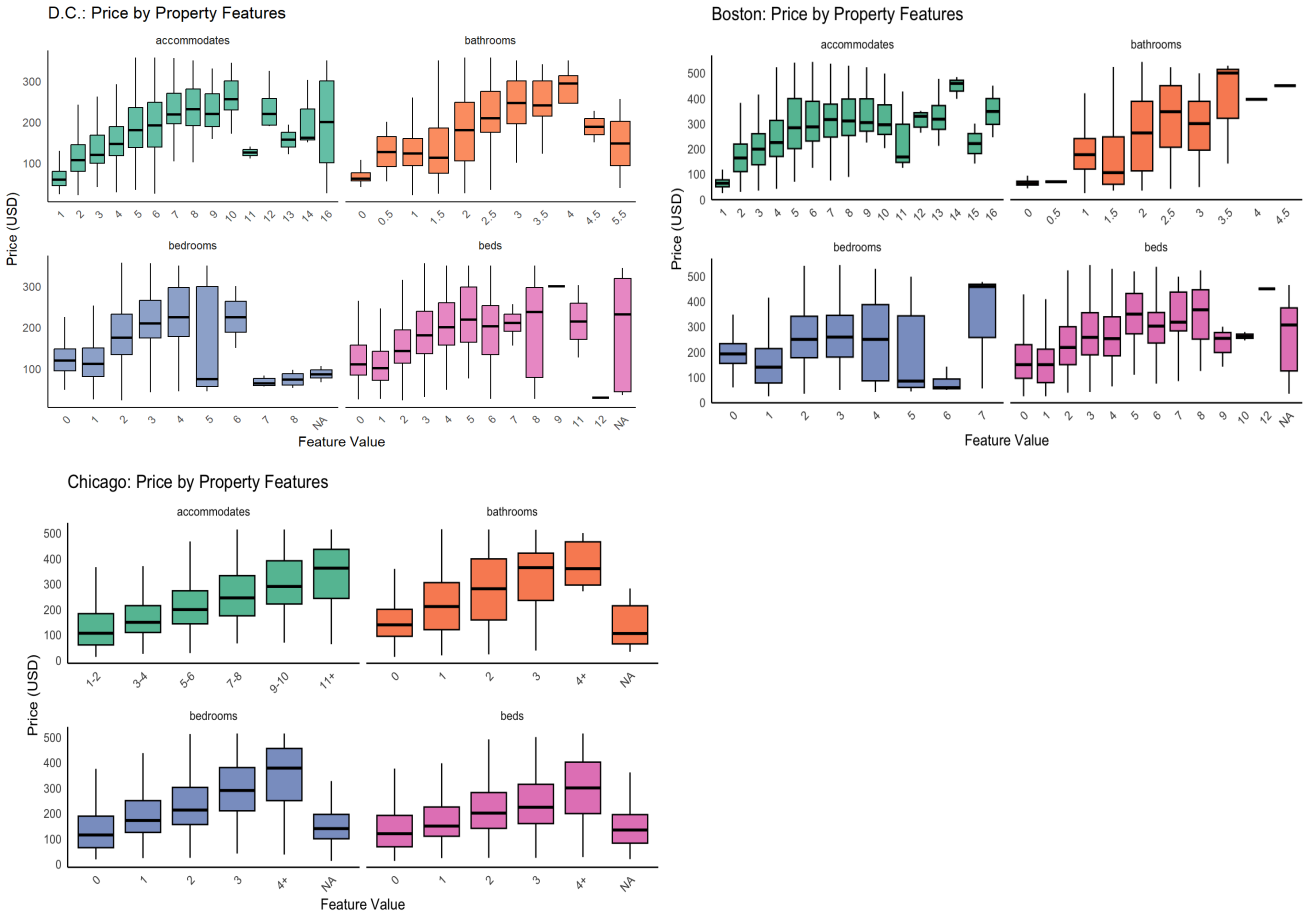


Figure A.2: Price by property features per city