Statistical Analysis of Airbnb Listings in Seattle

Sung Keum

Golden Gate University

Table of Contents

Introduction

Following the COVID-19 pandemic, the travel industry has been a boon for the US economy. The pent-up demand for travel following two years of different variations of travel restrictions has resulted in a rise in lodging bookings, boosting the economy. Per new figures released by the U.S. Travel Association, spending on travel in the United States exceeded $1.2 trillion in 2022, on par with pre-pandemic levels (Taylor, 2023). This expenditure has supported jobs in the travel industry as well as other tourism-related industries including restaurants, lodging, and retail. One of the beneficiaries of an increase in travel appetite is Airbnb. The company is an online platform that allows people to book places to stay. While it is famous for providing vacation homes, users of Airbnb can also book private rooms, shared rooms, and even hotel rooms. The properties available on Airbnb are mixed: some are owned by private individuals, while others are led by property management companies (Rawson, 2023). In 2022, there were 393.7 million nights stayed in Airbnb, an increase of 31% from the prior year (Woodward, 2023). The company has generated $8.3 billion in revenue in the same year, an increase of 40% from the prior year (Curry, 2023), and has a revenue CAGR of 23.10% in the last 5 years (Finance Charts, n.d.). Whether this trend will continue is a question that's been asked by many market pundits. The company has certainly been the beneficiary of the travel boom and the change in the working environment to a remote work. But with the rise in fuel costs, airfares, and lodging prices, as well as the overall costs of living, there is skepticism surrounding the sustainability of Airbnb's growth, let alone the growth of travel.

The recent, remarkable turnaround in the travel industry since the pandemic has inspired me to perform a statistical analysis into Airbnb listings, specifically in the Seattle area. The city is vibrant and is known for its thriving economy. It is a home to Starbucks, Microsoft, Amazon, and Costco, to name a few, and due to beautiful natural scenery, diverse culture, and excellent food and drink, it's one of the top places to visit.

Data Origination and Description

The data for this paper is gathered from insideairbnb.com. The "company" provides data on Airbnb listings in various cities around the world. The data is sourced from publicly available information from the Airbnb site and has already been cleansed and aggregated to facilitate public discussion. The dataset used for this study provides most recent listings in Seattle in June of 2023 and consists of 6636 records (listings) and 18 attributes, describing price per night, host name, number of reviews, room type, minimum nights required, and neighborhood of the listing, to name a few. Below is a full description of the dataset.

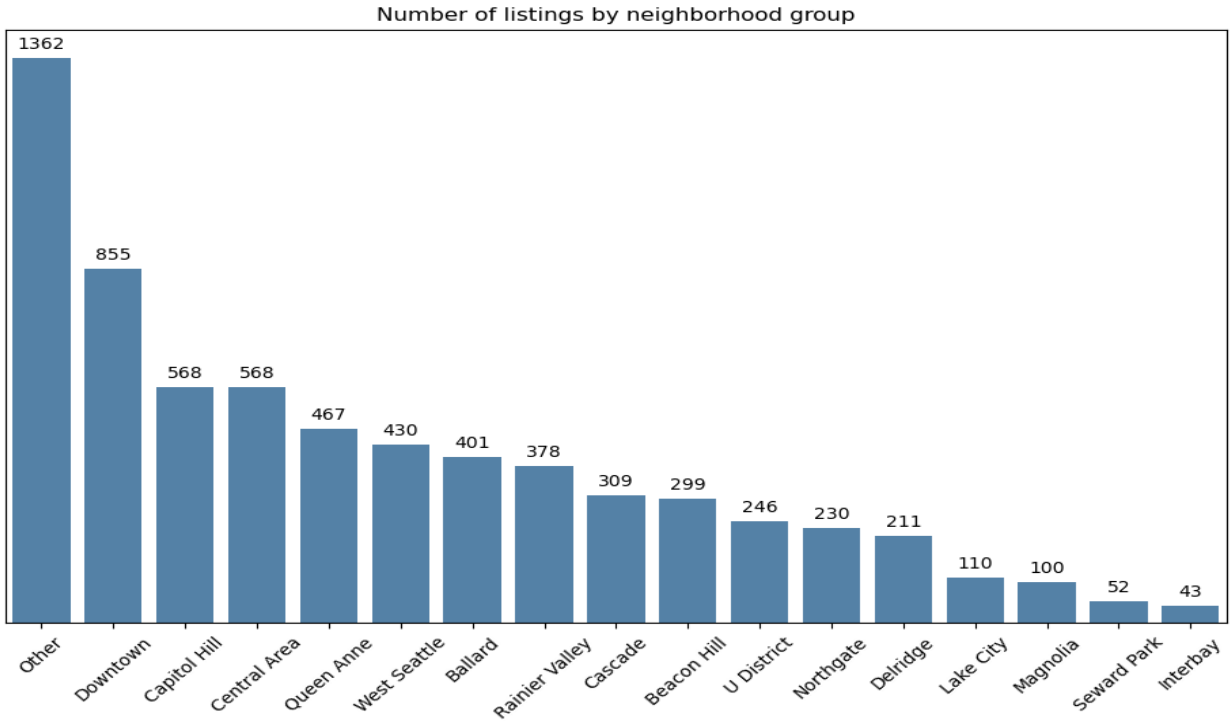
Table 1: Dataset variables

Name	Description	Name	Description
id	*Airbnb's unique identifier for listing*	price	*Daily price*
name	*Name of the listing*	minimum_nights	*Minimum number of night stay*
host_id	*Airbnb's unique identifier for the host*	number_of_reviews	*Number of reviews the listing has*
host_name	*Name of the host*	last_review	*The date of the last/newest review*
neighbourhood_group		reviews_per_month	
neighbourhood		calculated_host_listings_count	*The number of listings the host has in city*
latitude		availability_365	
longitude		number_of_reviews_ltm	*Number of reviews in the last twelve month*
room_type	*Entire place/Private room/Shared room/Hotel*	license	*License number*

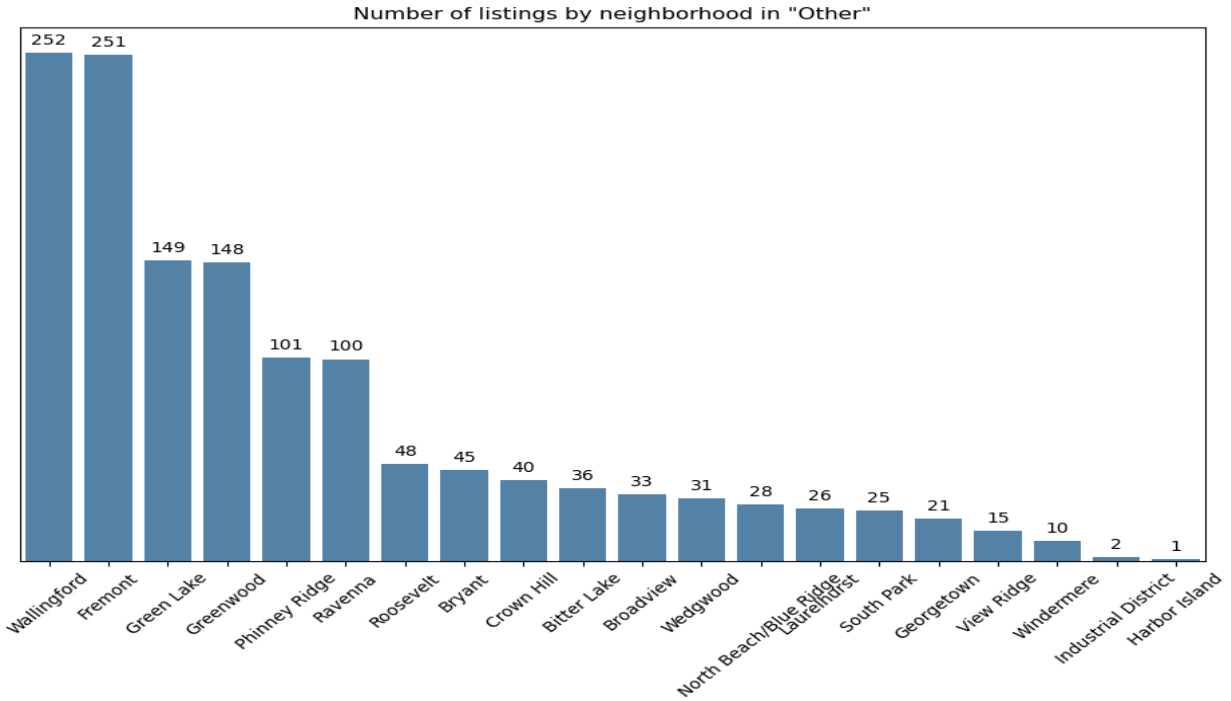
While I was exploring the dataset, I identified listings with exorbitant prices of $10,000 per night under the host name, Jen. While I believe that $10,000 per night Airbnb could exist, I had a doubt, so I went directly to an Airbnb website and searched up the host name Jen in Seattle. Upon investigation, I learned that the price for her listings was indeed $10,000 on display, but it changed as soon as I selected dates. The price varied depending on the day of the week the reservation was for, so it was impossible to replace the old value with a new price. Thus, I decided to delete all the listings under Jen from the dataset: there were a total of 6. Furthermore, a listing with a price of 0 was also deleted. After cleaning the data, I was left with 6629 observations and 18 attributes. A new variable named "seattle_region" was created in order to gain insight on price at a regional level. This variable contains four values—west, east, north, and central—and the value was assigned to each listing based on regional location.

Data Analysis

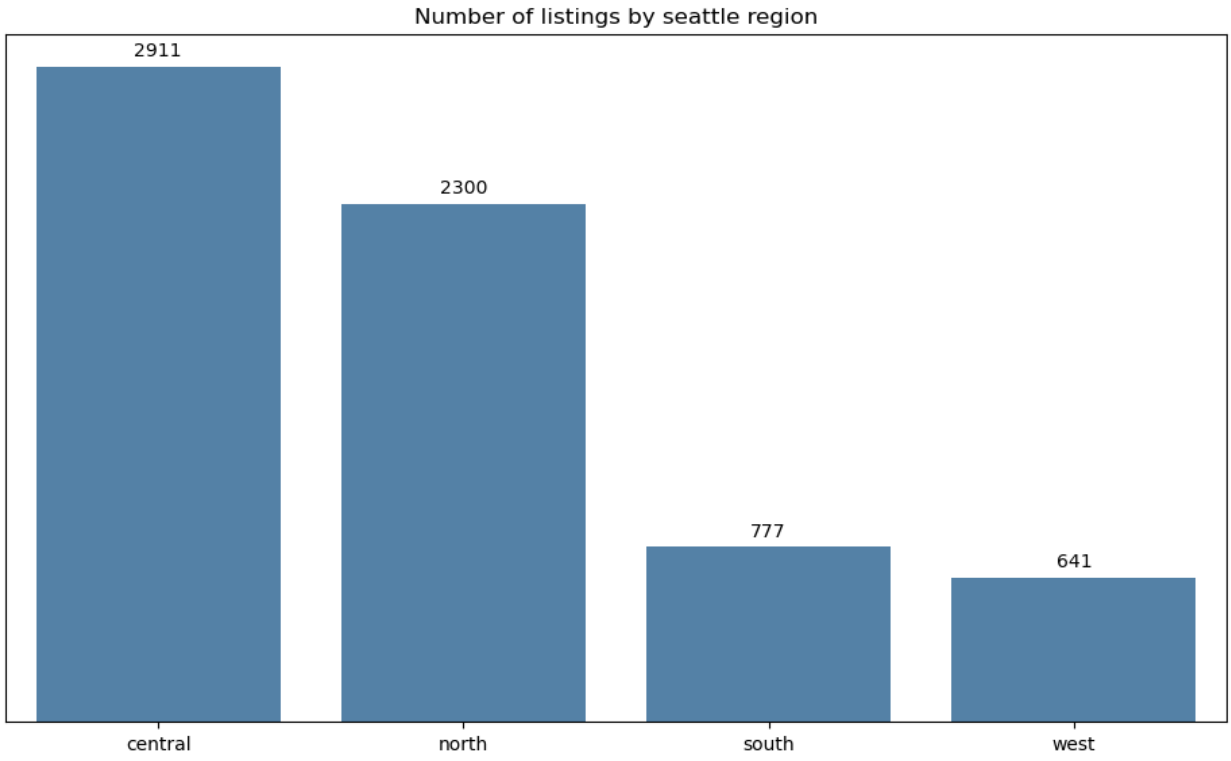
The main objective of this paper is to find factors impacting the price of Airbnb listings in Seattle and how much that impact is for each of those factors. In addition to regression analysis, I conduct hypothesis tests, such as t-test and ANOVA, on whether the difference in average prices (if they differ) by region, neighborhood, minimum nights required, room type, and host type (individual vs. company) is statistically significant. I also perform descriptive statistics to gain insights into the variables of interest.



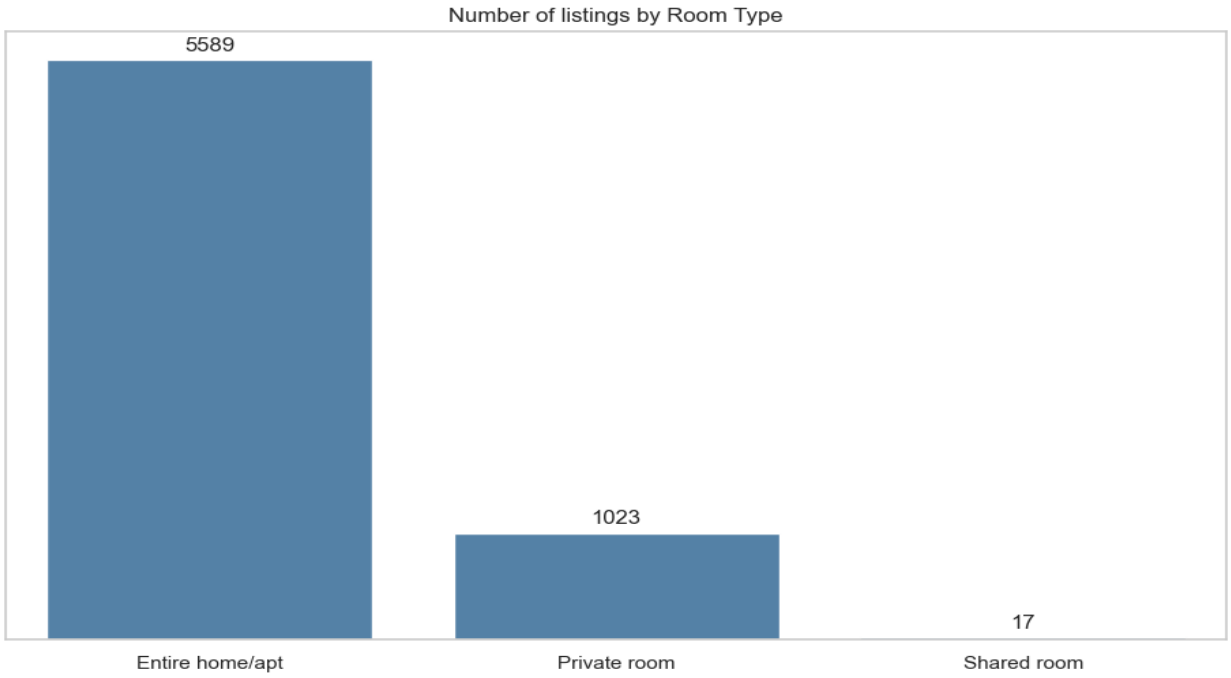
The listings in Seattle are divided into 17 neighborhood groups. The “Other” neighborhood group has the most listings at 1362, followed by Downtown (855), Capitol Hill and Central Area at 568, and so on. The chart below shows a result of drilling down “Other” neighborhood group



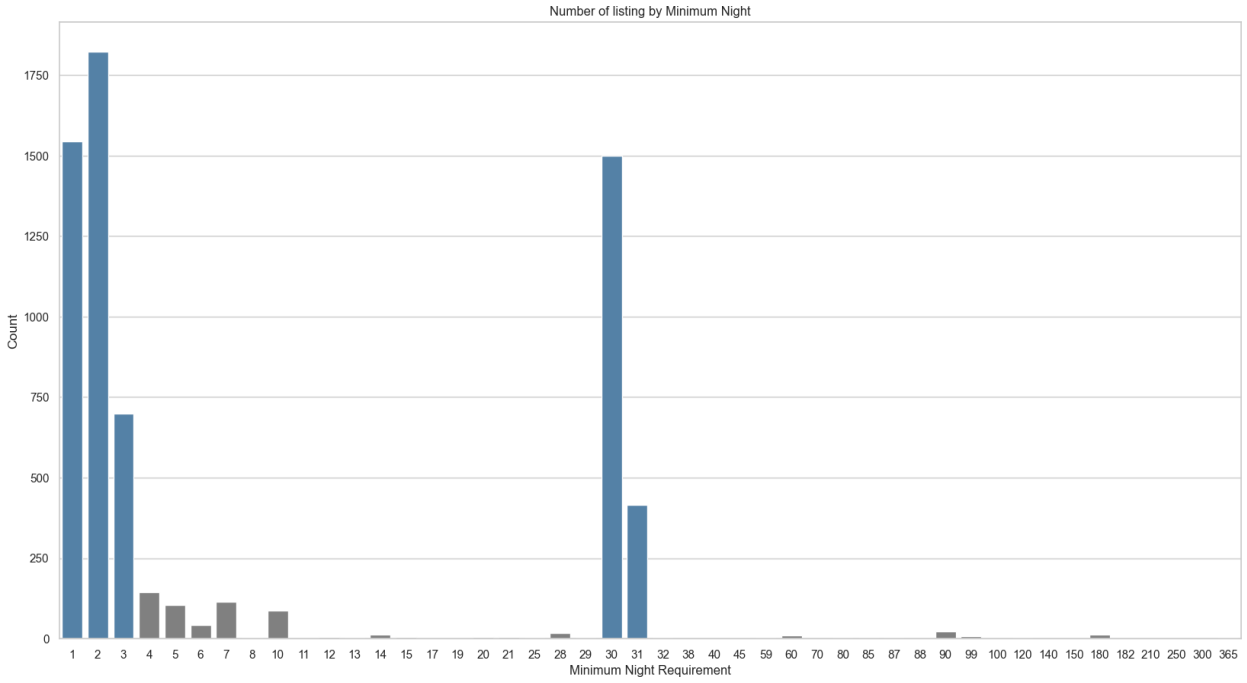
into the neighborhood. The top 6 neighborhoods dominate the listings. Wallingford and Fremont have 253 and 251 listings, respectively, followed by Green Lake (149), Greenwood (148), Phinney Ridge (101), Ravenna (100), and so on. The chart below shows the number of listings by region. The listings are concentrated in Central and North Seattle, followed by South and West Seattle.



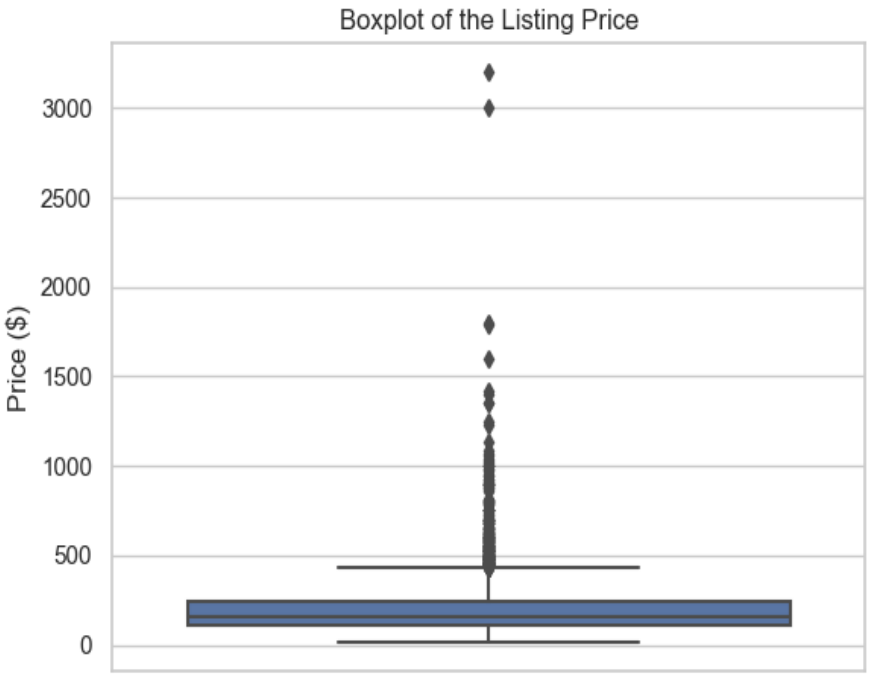
The type of room is heavily concentrated in Entire home/apartment as people tend to favor seclusion over sharing of space. This is not to say listings of Private room and Shared room are null. There are 1023 and 17 listings of Private Room and Shared Room, respectively.

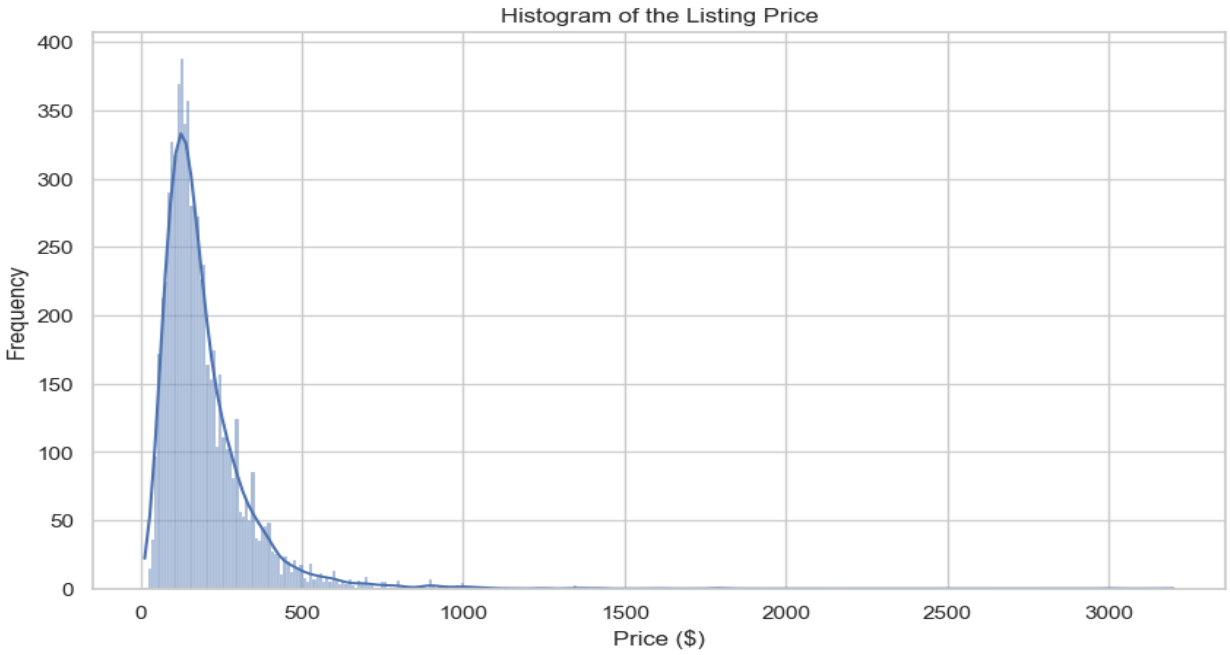


Per chart below, listings with minimum night requirement of 1, 2, 3, 30, and 31 are predominant in our dataset.



Below are boxplot, histogram, and summary statistics of the price of Airbnb listings in Seattle.





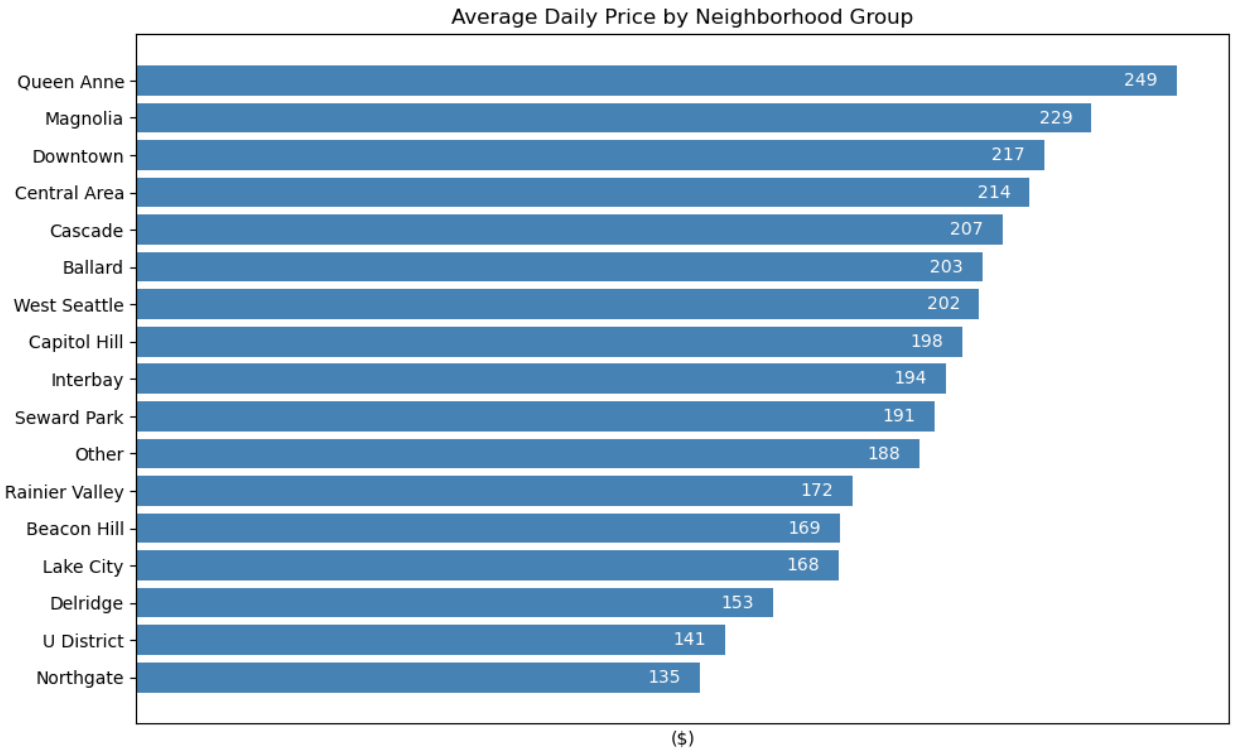
Summary Statistics of Price

Mean	Std	Min	25%	50%	75%	Max
195.64	146.1	13	111	158	238	3200

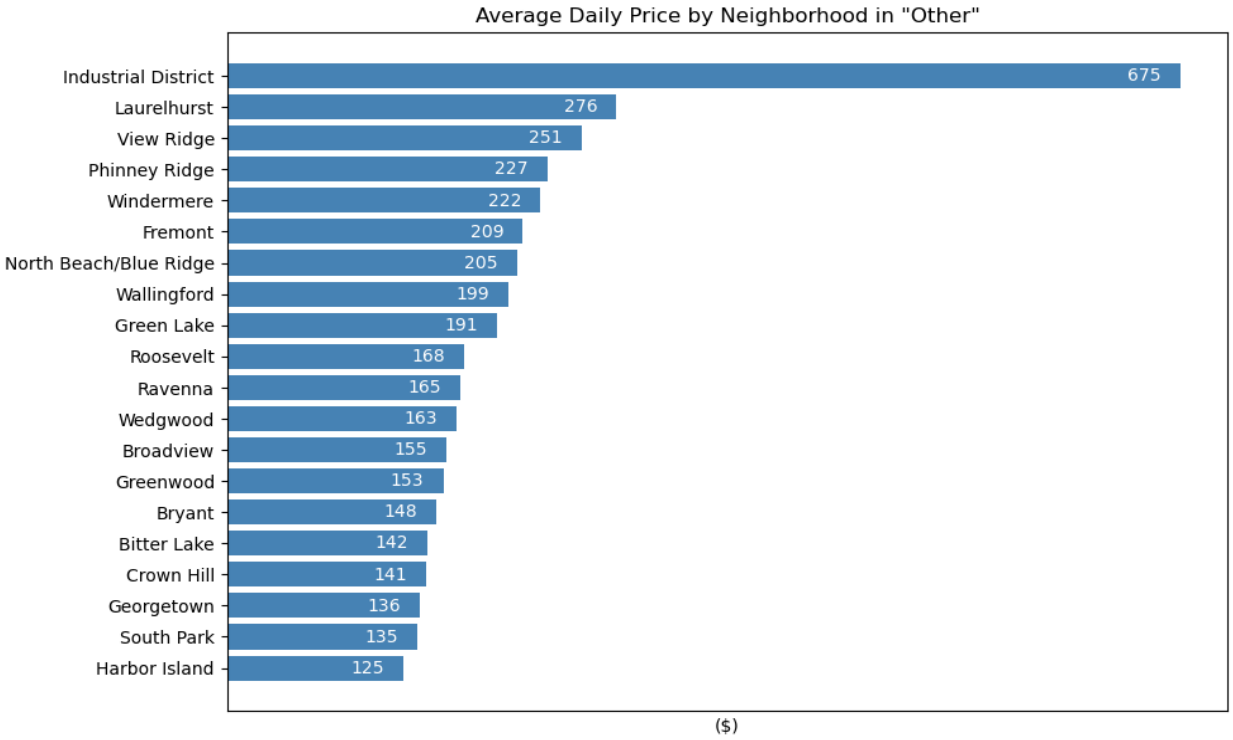
From the charts and a summary statistics table, the following information can be retrieved:

- Mean price of all Airbnb listings in Seattle is $195.64.
- Standard deviation of all Airbnb listing in Seattle is $146.1
- Median price of all Airbnb listings in Seattle is $158
- Maximum price of all Airbnb listings in Seattle is $3200
- **Data is highly right-skewed.**

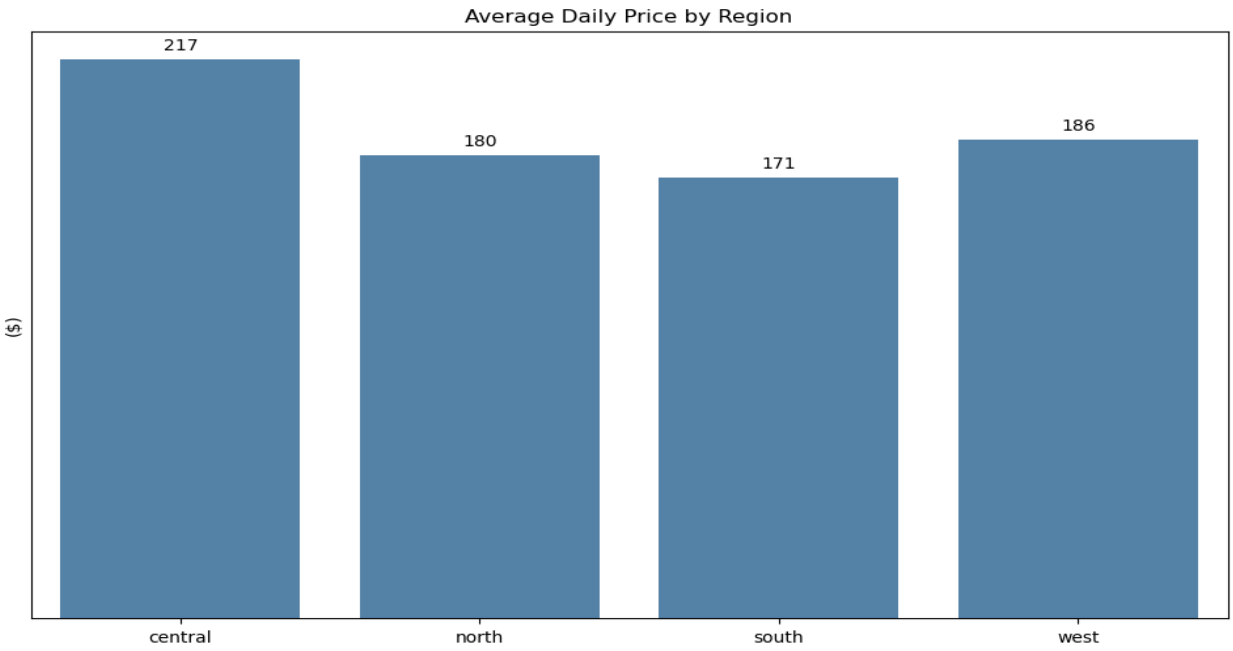
Next, the graphs displaying price distribution (boxplot) as well as the price comparison (bar plot) by room type, neighborhood group, region, and minimum nights required are provided.



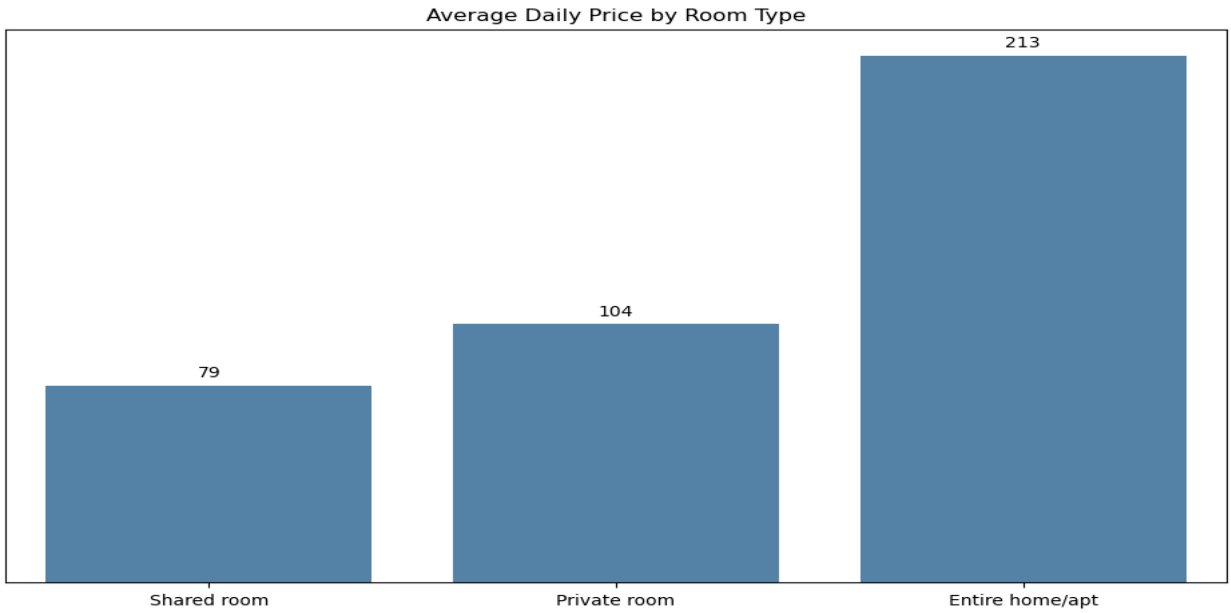
Queen Anne ($249), Magnolia ($229), and Downtown ($217) locations have the top 3 highest average daily price of Airbnb listings in Seattle. Just like in the number of listings, I drill down average price in the “Other” neighborhood group.



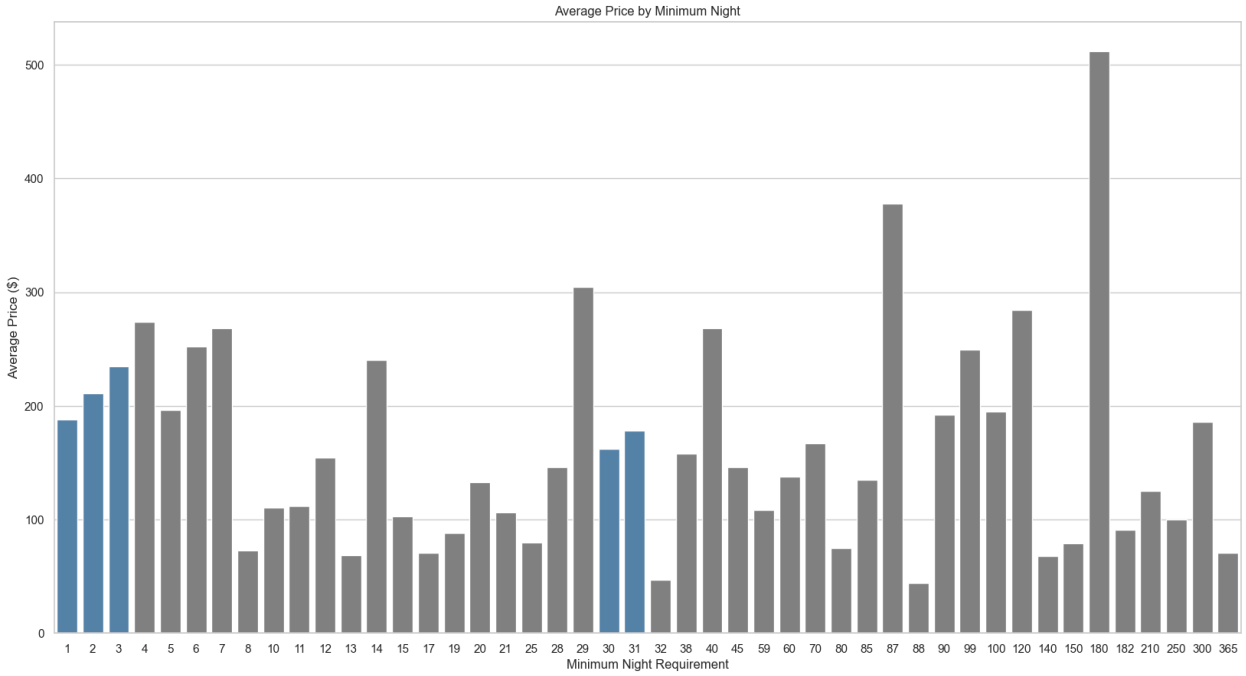
Industrial District holds the top spot in the "Other" neighborhood group with an average daily price of $675, followed by Laurelhurst ($276) and View Ridge ($251). One thing to note is that the area has a measly two listings, so comparing price with other areas may not be appropriate.



On a regional level, listings in Central Seattle have an average daily price of $217, followed by West ($186), North ($180), and South ($171).

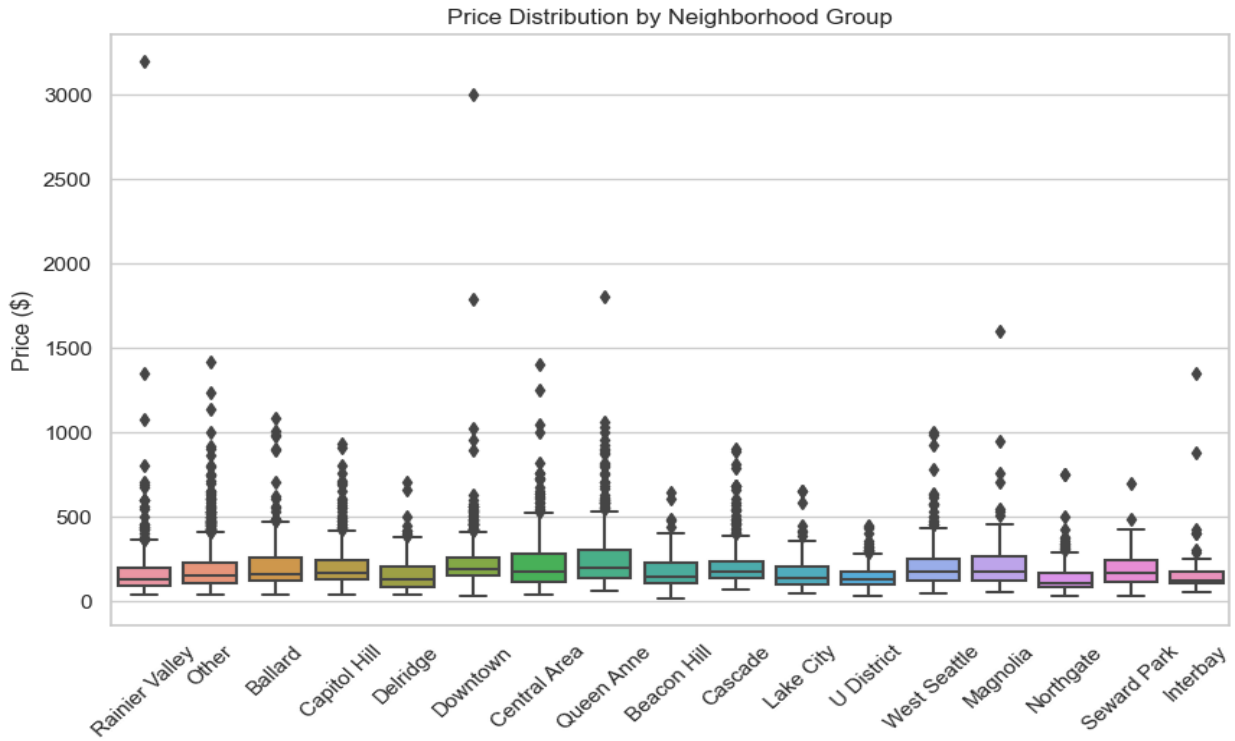


As one can anticipate, average price of Entire home/apt ($213) is higher than cheaper alternatives of Private room ($104) and Shared room ($79).

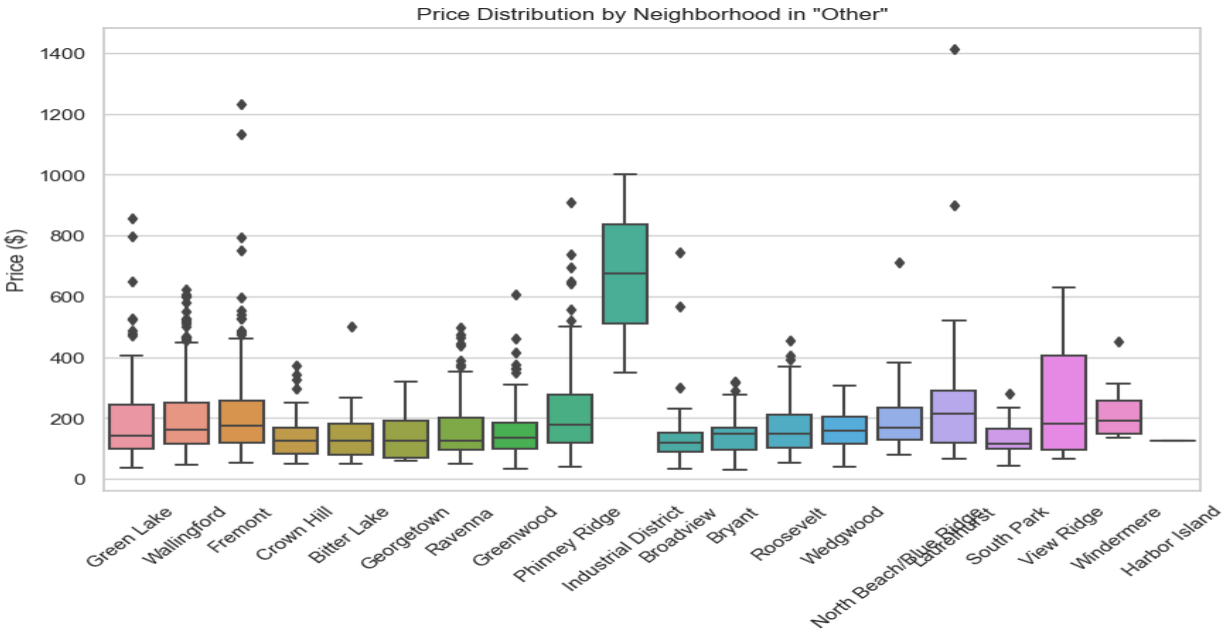


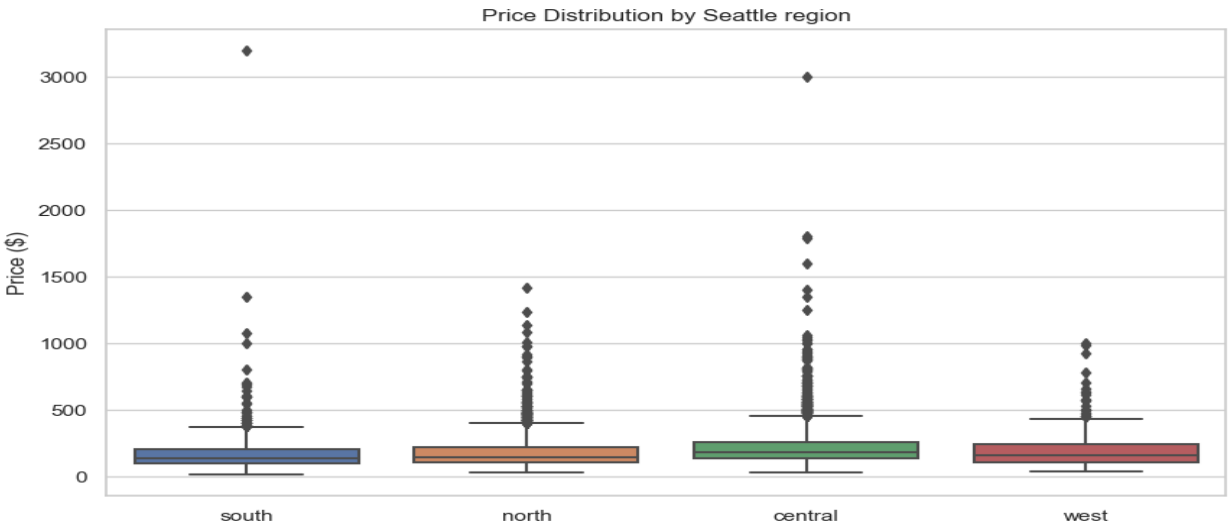
A chart comparing the number of listings by minimum nights showed that the listings with a night requirement of 1, 2, 3, 30, and 31 are predominant. Next, I explore the average price by night requirement to see if there is substantial price difference. The chart above shows that the listings with night requirement of 3 has the highest average price ($234), followed by night requirement of 2 ($210), night requirement of 1 ($188), night requirement of 31 ($178), and night requirement of 30 ($162). Although, comparing the prices between, for example, minimum night of 2 and minimum night of 30 may not be appropriate due to a difference in the intention (a weekend booking vs a month booking for short-term stay), comparing the prices between night requirements close to each other may yield insights to future guests and hosts.

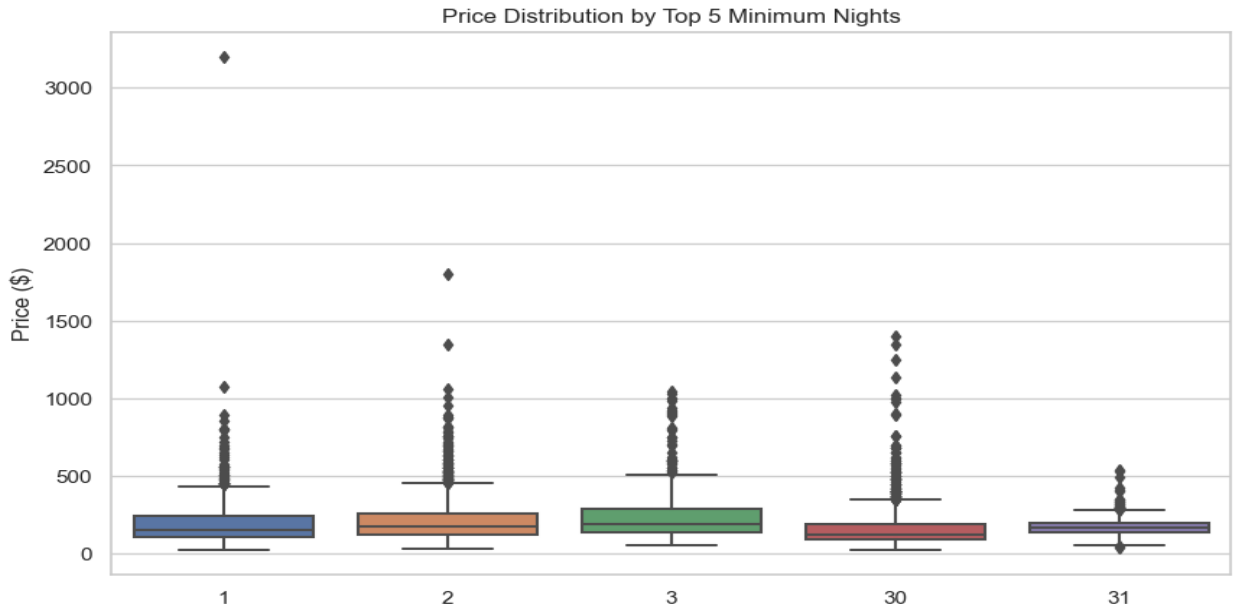
Now, let's look at the price distribution by neighborhood group, areas in the "other" neighborhood, region, and listings with minimum night requirement of 1, 2, 3, 30, and 31.



From the bar chart showing the average price by neighborhood group, we know that Queen Anne, Magnolia, and Downtown are the top 3 places with the highest average price. The boxplot is used to show the price variation for each group, indicating that future guests can find good deals in a neighborhood group with a higher average price. As anticipated, the data is highly right skewed. Next, we explore price distributions by areas in the “Other” neighborhood group, by region, and by listings with popular minimum night requirements.







Analysis of Price Differences in Minimum Nights, Region, Host Type

The One-Way Analysis of Variance, also known as ANOVA, will be used in this study to test for the statistical significant difference of the average prices between listings with different attributes for 3 or more groups. Similarly, the two-sample t-test will be used to test for statistical significance for 2 groups. I will be using these two tests to find out:

- If there exists a statistically significant difference between average prices for listings with minimum night requirement of 1, 2, and 3 (*Table 2*).
- If there exists a statistically significant difference between average prices for listings with minimum night requirement of 30 and 31 (*Table 3*).
- If there exists a statistically significant difference between average prices of listings in different regions (*Table 4*).
- If there exists a statistically significant difference between average prices of listings posted by an individual and a company (*Table 5*).

Table 2: Price difference in Minimum Night Requirement 1, 2, 3.

F statistic	p-value
28.29799291	6.23945E-13

The extremely small p-value indicates that we can reject the null hypothesis and conclude that there is indeed a statistical significant difference in average prices for listings with minimum night of 1, 2, and 3.

Table 3: Price difference in Minimum Night Requirement 30 and 31.

T statistic	p-value
-3.516012739	0.000451589

The extremely small p-value indicates that we can reject the null hypothesis and conclude that there is indeed a statistical significant difference in average prices for listings with minimum night of 30 and 31.

Table 4: Price difference by Region

F statistic	p-value
38.80115494	7.54526E-25

The extremely small p-value indicates that we can reject the null hypothesis and conclude that there is indeed a statistical significant difference in average prices of listings by region.

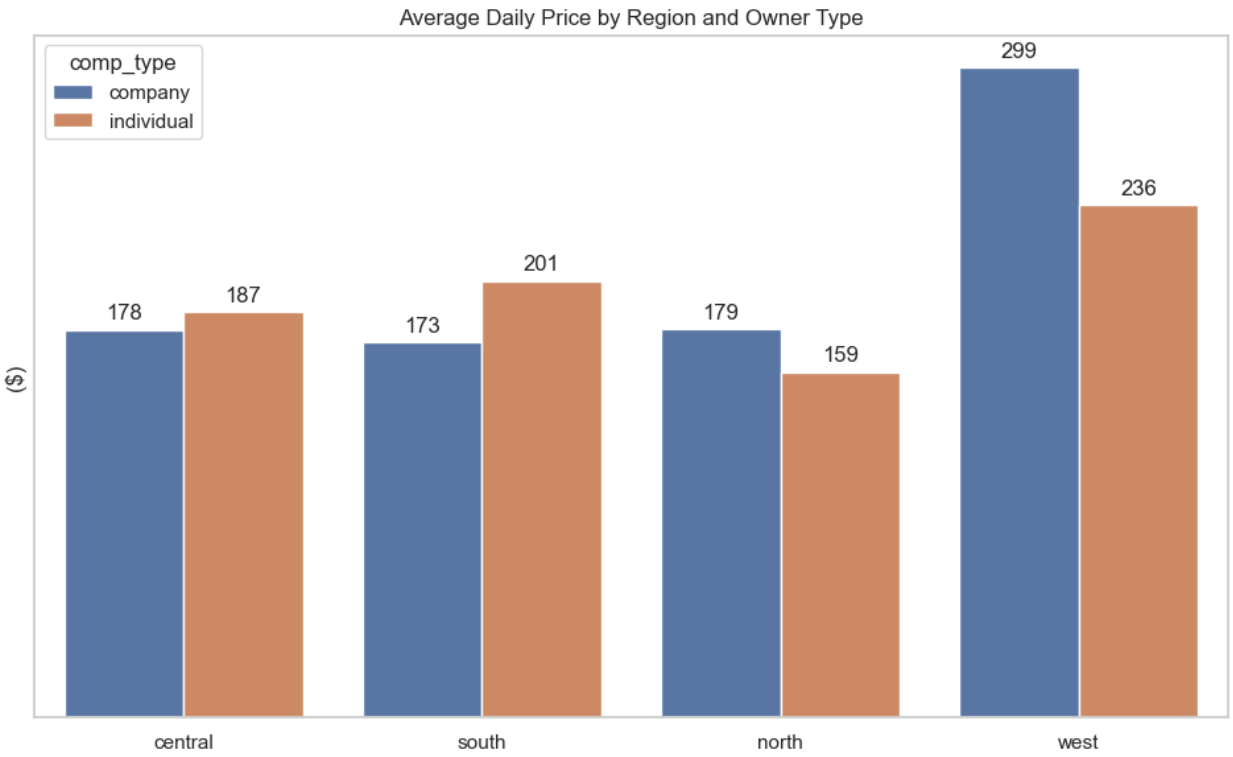
Table 5: Price difference by Host Type (Individuals vs Company)

T statistic	p-value
1.130509929	0.258623856

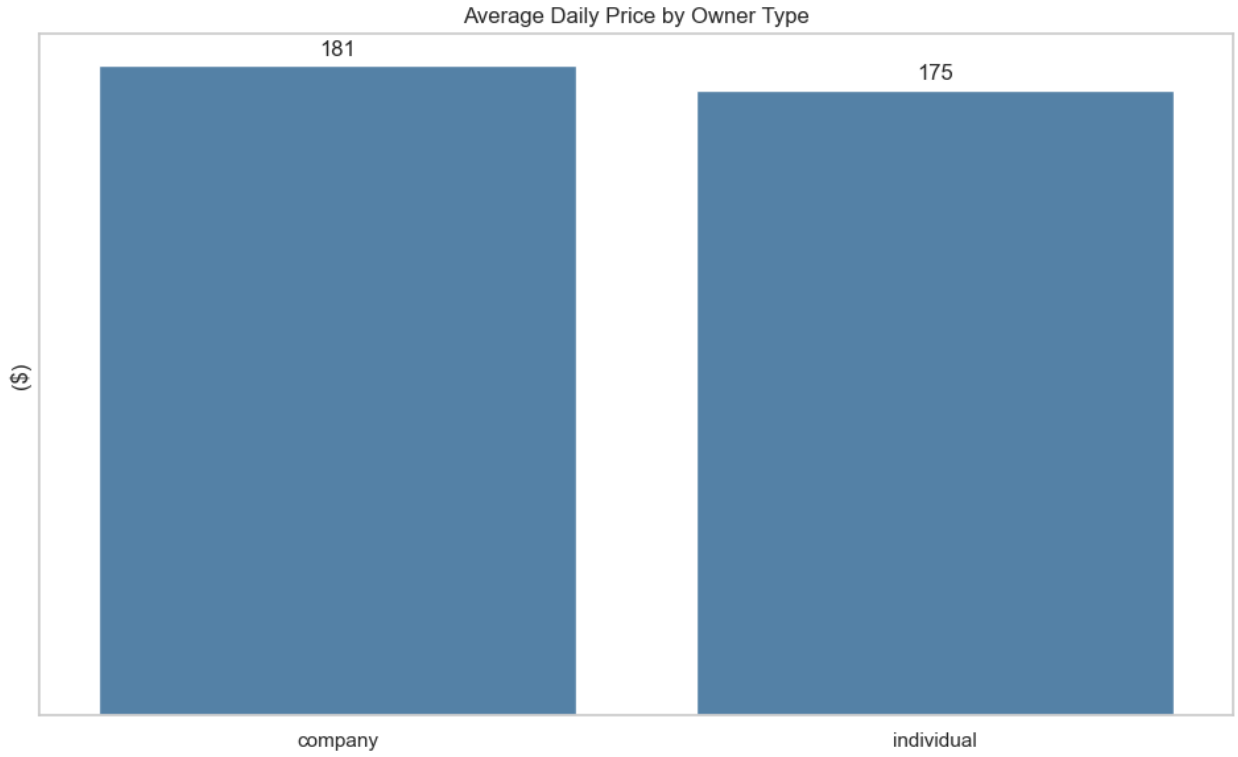
The small T statistic and a large p-value of 0.2586 fail to reject the null hypothesis and conclude that there is no statistical significant difference in average prices of listings posted by individuals and companies.

Analysis by Host Type and an Assumption

For this section of the study, the listings were aggregated by host name and sorted by the number of listings. Only the Top 10 hosts were considered for this analysis due to predominance in the numbers. Each host was put into one of two categories, company and individual. If the host name is a personal name, it is given an “individual” status. If it is not, it is given a “company” status. The objective of this analysis is to explore if the listings by companies and by individuals differ in price. An assumption is made to move forward with this analysis, which is that a host name with a personal name is an individual, not a company. For example, a host named “Blueground” was put into a company category, whereas “David” was put into an individual category (Note: This assumption may be false).



The chart above shows that the average price of listings in the West by host type differs quite a bit ($299 vs $236). In the North region, the average price of listings under companies is higher by $20 compared to that under individuals. In contrast, the average prices of listings under individuals in Central and South regions are higher than those under companies by $9 and $28, respectively. Next, the average price of listings by ownership type is shown at an aggregate level. The average price of listings by companies is higher by $6 than the average price of listings by individuals. In table 5 above, this difference in price is tested for statistical significance. The result is that there is no statistical significant difference in the average prices by host type.



Regression Analysis

Multivariate regression technique is used to estimate the influence of the variables on price. To begin, **Model_1** includes *Region* (seattle_region) and *Room Type* (room_type) as predictor variables. Each model subsequently adds complexity with the addition of predictor variables. **Model_2** adds *Number of Reviews* (number_of_reviews). Model_3 adds *Minimum Nights of Stay* (minimum_nights). (Note: Many financial variables including price typically have a log normal distribution. Thus, log transformation is used on price to restore symmetry).

Model_1: log(price) = Region + Room Type

OLS Regression Results

Dep. Variable:	log10_price	R-squared:	0.234
Model:	OLS	Adj. R-squared:	0.234
Method:	Least Squares	F-statistic:	405.0
Date:	Tue, 08 Aug 2023	Prob (F-statistic):	0.00
Time:	09:32:01	Log-Likelihood:	547.86
No. Observations:	6629	AIC:	-1084.
Df Residuals:	6623	BIC:	-1043.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.3027	0.004	543.874	0.000	2.294	2.311
seattle_region[T.north]	-0.0683	0.006	-10.951	0.000	-0.081	-0.056
seattle_region[T.south]	-0.0905	0.009	-10.015	0.000	-0.108	-0.073
seattle_region[T.west]	-0.0732	0.010	-7.525	0.000	-0.092	-0.054
room_type[T.Private room]	-0.3118	0.008	-40.875	0.000	-0.327	-0.297
room_type[T.Shared room]	-0.4306	0.054	-7.951	0.000	-0.537	-0.324

Omnibus:	542.252	Durbin-Watson:	1.933
Prob(Omnibus):	0.000	Jarque-Bera (JB):	745.890
Skew:	0.684	Prob(JB):	1.08e-162
Kurtosis:	3.911	Cond. No.	21.6

With the F-statistic of 405 and p-value essentially 0, the model is statistically significant. All the predictor variables (region and room type) have high t-statistic values as well. The Adjusted R-squared (0.234) indicates that about 23% of the variation in price is explained by the model.

Model_2: log(price) = Region + Room Type + Number of Reviews

OLS Regression Results

Dep. Variable:	log10_price	R-squared:	0.245
Model:	OLS	Adj. R-squared:	0.244
Method:	Least Squares	F-statistic:	358.0
Date:	Tue, 08 Aug 2023	Prob (F-statistic):	0.00
Time:	09:48:12	Log-Likelihood:	594.68
No. Observations:	6629	AIC:	-1175.
Df Residuals:	6622	BIC:	-1128.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.3197	0.005	509.396	0.000	2.311	2.329
seattle_region[T.north]	-0.0685	0.006	-11.052	0.000	-0.081	-0.056
seattle_region[T.south]	-0.0914	0.009	-10.187	0.000	-0.109	-0.074
seattle_region[T.west]	-0.0724	0.010	-7.500	0.000	-0.091	-0.054
room_type[T.Private room]	-0.3152	0.008	-41.571	0.000	-0.330	-0.300
room_type[T.Shared room]	-0.4155	0.054	-7.723	0.000	-0.521	-0.310
number_of_reviews	-0.0003	2.64e-05	-9.707	0.000	-0.000	-0.000

Omnibus:	488.318	Durbin-Watson:	1.938
Prob(Omnibus):	0.000	Jarque-Bera (JB):	662.971
Skew:	0.637	Prob(JB):	1.09e-144
Kurtosis:	3.882	Cond. No.	2.41e+03

With the addition of Number of Reviews, the F-statistic drops to 358, but the model is still highly significant with a p-value of 0. The Region and Room Type variables are still statistically significant as well as the Number of Reviews variable. The Adjusted R-squared (0.244) increased marginally, indicating that about 24% of the variation in price is explained by this model.

Model_3: log(price) = Region + Room Type + Number of Reviews + Minimum Nights

OLS Regression Results

Dep. Variable:	log10_price	R-squared:	0.275
Model:	OLS	Adj. R-squared:	0.274
Method:	Least Squares	F-statistic:	358.3
Date:	Tue, 08 Aug 2023	Prob (F-statistic):	0.00
Time:	10:02:56	Log-Likelihood:	728.23
No. Observations:	6629	AIC:	-1440.
Df Residuals:	6621	BIC:	-1386.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.3582	0.005	468.196	0.000	2.348	2.368
seattle_region[T.north]	-0.0746	0.006	-12.252	0.000	-0.086	-0.063
seattle_region[T.south]	-0.1031	0.009	-11.688	0.000	-0.120	-0.086
seattle_region[T.west]	-0.0845	0.009	-8.899	0.000	-0.103	-0.066
room_type[T.Private room]	-0.3154	0.007	-42.438	0.000	-0.330	-0.301
room_type[T.Shared room]	-0.3861	0.053	-7.317	0.000	-0.489	-0.283
number_of_reviews	-0.0003	2.64e-05	-12.989	0.000	-0.000	-0.000
minimum_nights	-0.0024	0.000	-16.499	0.000	-0.003	-0.002

Omnibus:	746.169	Durbin-Watson:	1.940
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1277.685
Skew:	0.774	Prob(JB):	3.58e-278
Kurtosis:	4.492	Cond. No.	2.41e+03

With the addition of Minimum Nights, the model is statistically robust with the F-statistic of 358.3 and the p-value of 0. All the predictor variables continue to be statistically significant, w ith extreme t-statistics for some variables. The Adjusted R-squared of 0.274, which is an increase from the previous models, indicates that about 27% of the variation in price is explained by this model.

Model_1 vs. Model_2

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	6623.0	328.994605	0.0	NaN	NaN	NaN
1	6622.0	324.379351	1.0	4.615254	94.217499	3.967740e-22

With the F-statistic of 94.22 and an extremely low p-value, it appears that Model_2 (inclusion of Number of Reviews) is an improvement over the Model_1.

Model_2 vs. Model_3

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	6622.0	324.379351	0.0	NaN	NaN	NaN
1	6621.0	311.569397	1.0	12.809954	272.217702	5.796404e-60

With the F-statistic of 272.22 and an extremely low p-value, it appears that Model_3 (inclusion of Minimum Nights) is an improvement over the Model_2.

Conclusions

This paper delved into the most recent listings of Airbnb in Seattle. The statistical analysis was conducted on the attributes of the listings, with a price variable as an anchor. The number and the average price of the listings were different across the regions, neighborhood groups, by room type, and by minimum night of stay requirement. The listings with the minimum night of stay requirement of 1, 2, 3, 30, and 31 days were predominant. At a regional level, Central Seattle had the most listings and the highest average price per night. At a group level, Downtown, Capitol Hill, and Central Area had the most listings. The higher number of listings did not exactly translate to a higher average price per night. Queen Anne and Magnolia had the listings with the highest average price per night. As one could have anticipated, the listings under Entire home/apt were most expensive, followed by Private room and Shared room. The price varied greatly for each neighborhood group and region, which goes to show that one can find a listing with a “good” price at a neighborhood group or a region with higher priced listings. The difference in prices by region and by listings with different minimum night requirements were statistically significant. An additional analysis was conducted to find if price was different for listings posted by companies and individuals. The difference in price was

significant in the West region, but miniscule in other regions. A statistical test has shown that the difference by host type is not statistically significant.

As shown by the final regression model, the Region, the Room Type, the Minimum Nights of stay, and the Number of Reviews on the whole, explained approximately 27% of the variation in price: they were all statistically significant factors in prediction of price.

- On average, the listings in the North region, relative to those of Central, are 7.18% lower in price.
- On average, the listings in the South region, relative to those of Central, are 9.8% lower in price.
- On average, the listings in the West region, relative to those of Central, are 8.1% lower in price.
- On average, the listings under Private room type, relative to those of Entire room/apt, are 27.05% lower in price.
- On average, the listings under Shared room type, relative to those of Entire room/apt, are 32.03% lower in price.
- A one-unit increase in the minimum nights of stay requirement results in a measly 0.24% decrease in price.
- A one-unit increase in the number of reviews results in a measly 0.03% decrease in price.

References

Curry, D., (May 24, 2023). *Airbnb Revenue and Usage Statistics (2023).* BusinessofApps. https://www.businessofapps.com/data/airbnb-statistics/

FinanceCharts, (n.d.). *Airbnb (ABNB) Revenue CAGR*. https://www.financecharts.com/stocks/ABNB/income-statement/revenue-cagr

InsideAirbnb, (n.d.). http://insideairbnb.com/

Rawson, C., (July 11, 2023). *How Does Airbnb Work?* NerdWallet. https://www.nerdwallet.com/article/travel/how-does-airbnb-work

Taylor, M., (April 19, 2023). *US Travel Spending of $1.2 Trillion on Par with Pre-Pandemic Levels.* TravelPulse. https://www.travelpulse.com/news/destinations/us-travel-spending-of-1-2-trillion-on-par-with-pre-pandemic-levels

Woodward, M., (July 18, 2023). *Airbnb Statistics [2023]: User & Market Growth Data*. Search Logistics. https://www.searchlogistics.com/learn/statistics/airbnb-statistics/