Module 1 Final Project

SAXA 3 - Mike Johnson | Kesh Kamani | Ryan Mathis | Khushi Patel | Andrew Singh

Summary

In the Washington DC metro area, we found there to be significant potential for investment in the Georgetown and Capitol Hill neighborhoods based on the popularity of the neighborhood, listing price on a per-person basis, and variability of prices within those neighborhoods. Capitol Hill is the second most populated neighborhood in terms of total listings, while Georgetown ranks fourth out of the six neighborhoods. However, they attract the highest rate of visitors, with almost 6 more stays per listing than units in the next most popular neighborhood. Average listing prices for entire homes/apartments in Georgetown and Capitol Hill are $253 and $198, and average prices for private rooms are $144 and $83.50, respectively. On a per-person basis, these numbers are adjusted to $58.70 and $49.60 for entire homes/apartments and $63.90 and $27.80 for private rooms. Prices in these neighborhoods are also quite variable, with an average IQR of $115 for entire homes/apartments and an average of $29 for private rooms. This allows hosts more flexibility in setting the prices which can result in returning a higher rate of stays in the units. Our recommendation based on the data is that an investment into an entire home or apartment in either the Capitol Hill or Georgetown neighborhoods will be the most lucrative for a long term investment.

Question 1: Exploratory Data Analysis

In our analysis, we focused on providing a concise summary of key statistics to inform investors about the various neighborhoods in the Washington D.C. metropolitan area. Our insights provide an overview of the total number of reviews, number of listings, average price per person per commendation, and the number of stays per listing. This provides a series of quick insights when trying to get a holistic view of the AirBnB market in various Washington D.C. neighborhoods.

Table 1: Analysis of key market factors within the entire dataset

Neighborhood	Stay Count	Price per Person	Listing Count	Stays per Listing
Capitol Hill	36,398	$48.10	412	88
Georgetown	12,021	$59.00	138	87
Dupont Circle	39,344	$48.10	487	81
Union Station	21,721	$53.70	280	78
Shaw	21,465	$59.30	305	70
Foggy Bottom	6,267	$58.70	96	65

Module 1 Final Project
SAXA 3 - Mike Johnson | Kesh Kamani | Ryan Mathis | Khushi Patel | Andrew Singh

Table 2 & 3: Analysis of the key market factors

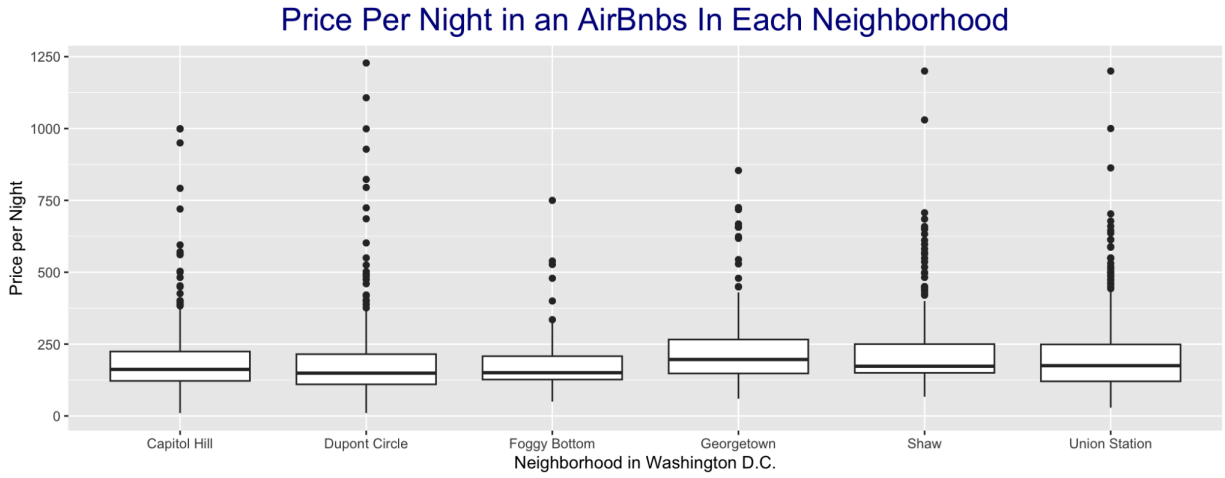
Entire Home/Apartment

Neighborhood	Stay Count	Price per Person	Listing Count	Stay per Listing
Georgetown	11,929	$58.70	126	95
Capitol Hill	35,978	$49.60	397	91
Shaw	20,043	$55.40	231	87
Union Station	38,644	$47.70	475	81
Dupont Circle	19,138	$57.20	288	67
Foggy Bottom	6,034	$56.50	91	66

Private Room

Neighborhood	Stay Count	Price per Person	Listing Count	Stay per Listing
Dupont Circle	2,237	$112.95	17	137
Union Station	700	$70.62	12	58
Foggy Bottom	233	$92.78	5	47
Shaw	1,301	$42.80	41	32
Capitol Hill	34	$27.83	2	17
Georgetown	92	63.85	12	8

Table 4: Distribution of price Night in an AirBnB per Neighborhood: The median price distribution per a one night stay at an Airbnb across different neighborhoods is around a similar range, where Dupont Circle is the most affordable option at $149, relative to Georgetown being the most expensive at $196.



NA Handling

The data analyzed had NA's in four variables. Below are notes on how these NA's were addressed.

- **Host Acceptance Rate:** NA's were replaced with the mean host acceptance rate.
- **Room Type:** NA's were replaced with the most common room type in the data. The most common room type in the data was 'Entire home/apt'.
- **Bedrooms**: The number of beds was identified as a predictor of the number of bedrooms with an r-squared value 0.60. A linear regression model was used to predict the number of bedrooms based on the number of beds for NA's.
- **Average Rating:** NA's were replaced with the mean average rating.

Question 2

1. **Which combination of neighborhood and room type has the highest average price?**
   The combination of neighborhood and room type that has the highest average price is the private room in Foggy Bottom with an average price of $334 per night per unit.

2. **Which one has the lowest?**
   The combination of neighborhood and room type that has the lowest average price is the shared room in Capitol Hill with the average price being $35.08 per night per unit.

3. **Which combination has the highest variability?**
   The combination that has the highest variability in price of entire homes or apartments in Georgetown where there was a $129.25 difference in price between the 25th and 75th percentiles.

4. **Which combination has the lowest?**
   The combination that has the lowest variability in price in the entire dataset were shared rooms in Shaw, where there was a $0 difference between the 25th and 75th percentiles. However, when viewing the dataset, 38 of the 41 shared room units in this neighborhood were listed at the exact same price by what appeared to be the same Host, given that the host in question had 41 listings. Based on that finding, it would be more valuable to look at the next combination in question, which is private rooms in Foggy Bottom, with a $20 range between the 25th and 75th percentiles.

Question 3: Provide an interpretation of the computed confidence interval in the context of the Airbnb listings.

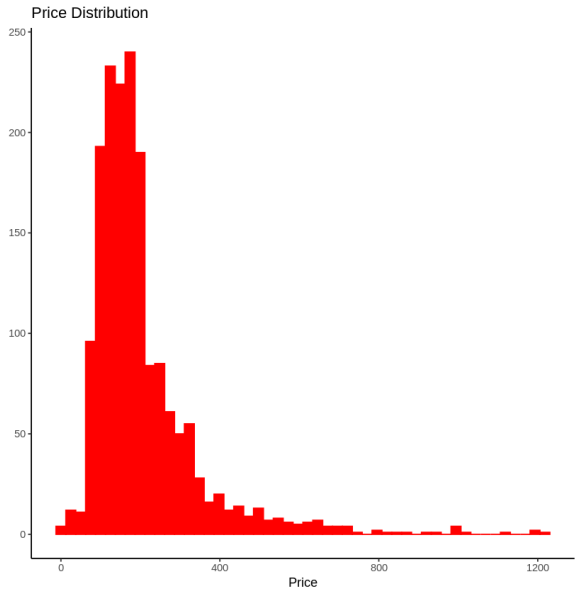
The confidence interval was computed using a custom R function designed to return the lower and upper bounds for the specified variable — in this case, the price. The calculated 95% confidence interval for the average price is between 198.35 and 211.34 dollars. This interval suggests that we can be 95% confident that the true average price of Airbnb listings in Washington, DC falls within this range.

Question 4: Test whether the average price of all listings in the population is more than $200 (at 95% level of confidence). What is the statistical conclusion based on this result? Is this conclusion in line with the estimated confidence interval reported in question 3?

The null hypothesis (H0) posits that the average price is $200, while the alternative hypothesis (H1) suggests that it is greater than $200.

A one-sample t-test was conducted, comparing the sample mean to the hypothesized value of $200. The results yielded a **t-statistic of 1.46** and a **p-value of 0.07**. Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis. Therefore, based on the data, there is insufficient evidence to conclude that the average price is significantly greater than $200. This result is consistent with the confidence interval reported earlier, which includes values both above and below $200.

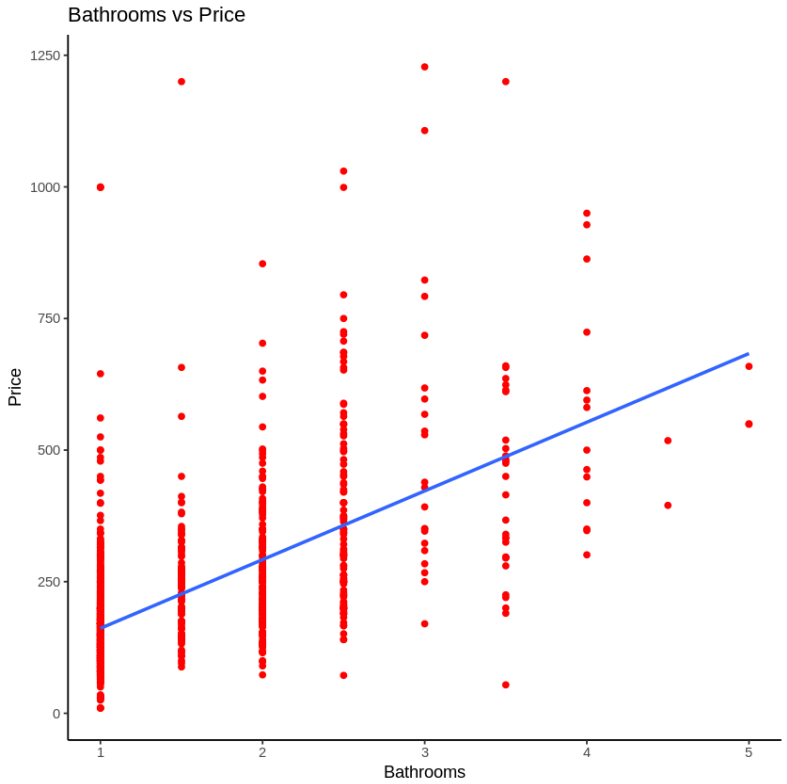
Question 5: Visualize price to test for normality and comment on the results



When visualizing the normality of the price, we found the price is skewed to the right, with the majority of the distribution of prices falling between approximately $0 to $400 per night. The median price for an AirBnB is $171 per night. We found that there are only 7% of AirBnB's with prices greater than $400 per night.

Question 6: What's the best simple linear regression model for "price" of the listings based on R-squared and residual standard error?

The best simple regression model for price is bathrooms. This is because the model has the highest R-squared value at 0.38 and the lowest residual standard error. A higher R-squared means that the model does a better job at explaining variability in price compared to the other models tested. A lower residual standard error means that the model makes predictions that are closer to actual values. A visual has been included below for reference.



Summary of R-squared and RSE

The table below summarizes the R-squared and residual standard errors for each variable tested against price. Interestingly, the variables with the highest R-squared all just happen to be related to the size of the listing.

Model	R-Squared	RSE
Bathrooms	0.38	107.99
Accommodates	0.34	111.15
Bedrooms	0.34	111.71

Beds	0.27	117.36
Neighborhood	0.01	136.49
Superhost	0.01	136.60
Host Acceptance Rate	0.01	136.89
Room Type	0.01	136.39
Min Nights	0.01	136.69
Total Reviews	0.01	136.71
Host Since	0.00	137.27
Host Total Listings	0.00	137.24
Avg Rating	0.00	137.23

Question 7: Implement a multiple linear regression model for the "price" of the listings. Report the regression coefficients and measures of fit, and write an interpretation of the regression coefficients in the context of this model. Are there any violations of model assumptions?

A multiple regression model was created using the top three variables based on their R-squared. A summary of the model included below.

Regression Coefficients

- **Intercept:** 6.205 - Not significant with a high p-value (0.31)
- **Bathrooms:** Each additional bathroom is associated with a $75.81 increase in price.
- **Accommodates:** Each additional guest accommodated is associated with a $17.42 increase in price.
- **Bedrooms:** Each additional bedroom is associated with a $18.28 increase in price.

Measures of Fit

- **Adjusted R-Squared:** 43.28% of the variance in price is explained by the model
- **Residual Standard Error:** 103.4. This is a measure of the average distance that the observed values fall from the regression line.
- **F-statistic:** 437.6, $p << 2.2e-16$. This indicates that at least one of the variables is significantly related to the response variable.

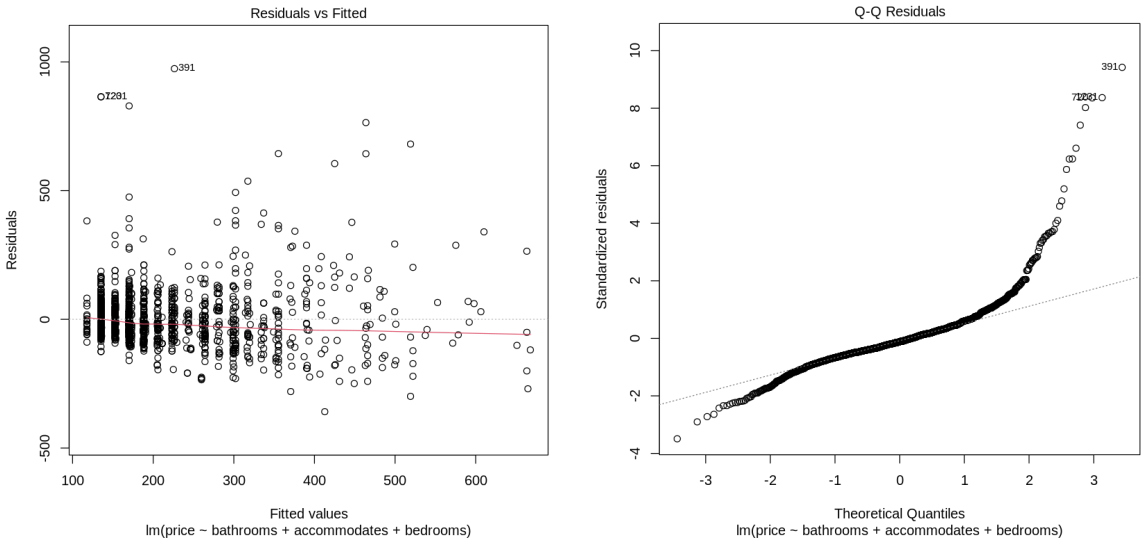
Interpretation

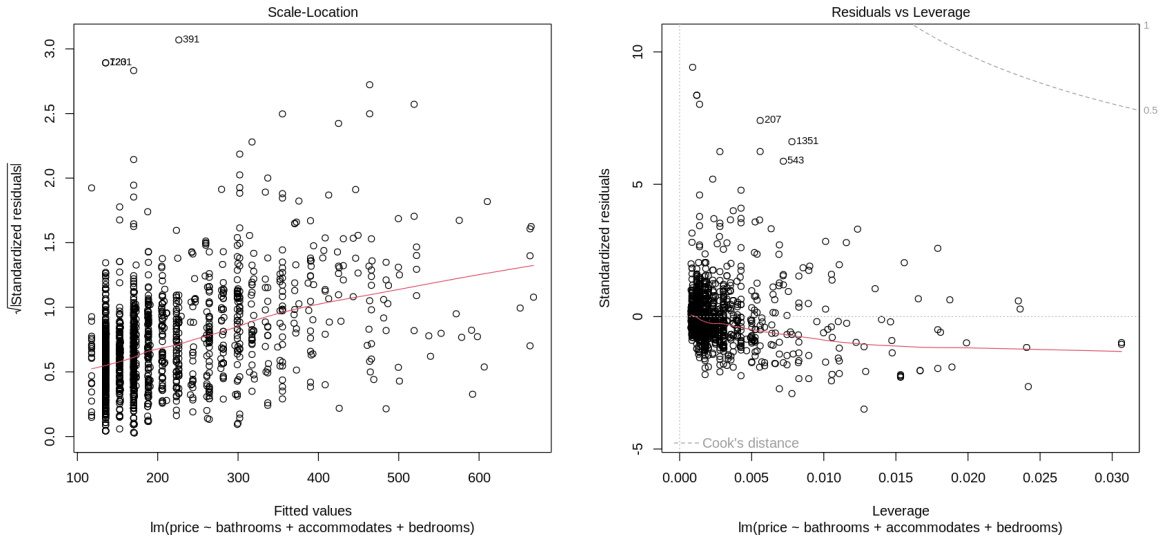
Each additional bathroom, person accommodated, and bedroom increases the price. The small p-values suggest that the relationship to price is not due to random chance.

Diagnostic Review

The diagnostic plots suggest a violation of model assumptions with a non-constant variance.

- **Residuals vs Fitted** shows a "funneling", suggesting an increasing variance.
- **Q-Q Residuals** shows deviations from the normal line. The residuals should remain close to the normal line.
- **Scale-Location** shows an upward line - meaning a non-constant variance.





Question 8: Are there any multicollinearity concerns among the independent variables selected for the multiple regression model of question 7?

There is multicollinearity among the independent variables, with correlation coefficients ranging from moderate to strong positive. This means a change in one independent variable is associated with changes in another, leading to less reliable inferences in the model.

The correlation matrix below summarizes the correlation between each variable.

	Price	**Bathrooms**	**Accommodates**	**Bedrooms**
Price	1.00	0.62	0.59	0.58
Bathrooms	0.62	1.00	0.69	0.76
Accommodates	0.59	0.69	1.00	0.79
Bedrooms	0.58	0.76	0.79	1.00

Due to the multicollinearity, it is advised to select just one of the variables to use in a regression model related to price.