An Analysis of Toronto Short-Term Rentals - Part 3 (Final Writeup)

Group 12 Yiping Liu (ID: 0870378)
Dmitry Mishchenko (ID: 0512776)
Mohammed Muzammil (ID: 0885649) Fawaz Yahya (ID: 0867165)

2025-11-15

Table of contents

```
airbnb <- read_csv("listings.csv", show_col_types = FALSE)

DataKeyVariable <- airbnb |>
  select(price, neighbourhood_cleansed, accommodates,
         review_scores_rating, reviews_per_month) |>
  mutate(price = parse_number(price))

n_raw <- nrow(DataKeyVariable)

# Active listings only
DataFiltered <- DataKeyVariable |>
  filter(
    !is.na(price),
    !is.na(review_scores_rating),
    !is.na(reviews_per_month),
    !is.na(neighbourhood_cleansed),
    price > 0,
    reviews_per_month > 0,
    review_scores_rating > 0
  )

# Build neighbourhood demand groups using reviews_per_month as a demand proxy
neigh_demand <- DataFiltered |>
  group_by(neighbourhood_cleansed) |>
  summarise(mean_rpm = mean(reviews_per_month), .groups = "drop")

q20_80 <- quantile(neigh_demand$mean_rpm, probs = c(0.20, 0.80))

neigh_demand <- neigh_demand |>
  mutate(neigh_group = case_when(
    mean_rpm >= q20_80[2] ~ "High",
    mean_rpm <= q20_80[1] ~ "Low",
    TRUE ~ "Middle"
  )) |>
  mutate(neigh_group = factor(neigh_group, levels = c("High", "Middle", "Low")))

DataClean <- DataFiltered |>
  left_join(neigh_demand |> select(neighbourhood_cleansed, neigh_group),
            by = "neighbourhood_cleansed")

n_clean <- nrow(DataClean)

n_raw
```

```
[1] 21356
```

```
n_clean
```

```
[1] 12571
```

1 Executive Summary

This report examines the factors that influence nightly Airbnb prices in Toronto using data from the Inside Airbnb project[1] (September 2025). The primary objective is to assess how accommodation capacity, guest review ratings, and neighbourhood-level demand are associated with listing prices.

Exploratory analysis revealed that nightly prices are strongly right-skewed, motivating the use of a log transformation of price to better satisfy linear regression assumptions. A multiple linear regression model was fitted with `log(price)` as the response variable and `accommodates`, `review_scores_rating`, and `neigh_group` as predictors.

All three predictors were found to be statistically significant. Holding other variables constant, each additional guest capacity is associated with an approximate 22% increase in nightly price. A one-point increase in review score rating corresponds to an estimated 20% increase in price. Listings located in low-demand neighbourhoods charge about 33% less than comparable listings in high-demand areas, while listings in middle-demand neighbourhoods charge roughly 48% less. Overall, the model explains about 50% of the variation in nightly price.

In addition to the regression analysis, a Welch two-sample t-test was conducted to compare mean `log(price)` between high- and low-demand neighbourhoods. The test indicated a statistically significant difference, providing further evidence that neighbourhood demand plays an important role in Airbnb pricing.

Overall, the results suggest that listing size, guest ratings, and neighbourhood demand are key drivers of Airbnb prices in Toronto. These findings provide insight into how pricing varies across the city and highlight the importance of both property characteristics and location in the short-term rental market.

2 Introduction and Background

Short-term rentals play an important role in Toronto's housing and tourism markets. Platforms such as Airbnb allow property owners to rent accommodations to short-term guests, creating variation in nightly prices across listings and neighbourhoods. Understanding the factors that drive these price differences is useful for hosts setting prices, guests comparing options, and policymakers monitoring short-term rental activity.

This analysis uses data from the Inside Airbnb project, which provides publicly available information on Airbnb listings in Toronto. The dataset includes listing characteristics, review information, and neighbourhood identifiers. For this project, we focus on listings that have received reviews, representing active listings with observable pricing and review activity.

The primary outcome of interest is the nightly listing price. Preliminary exploration showed that price is highly right-skewed, with a small number of very expensive listings. To address this

skewness and better meet linear regression assumptions, we analyze the natural logarithm of price, `log(price)`, rather than raw price values.

Three explanatory variables are considered based on exploratory analysis and course concepts. The variable `accommodates` captures the size of the listing and reflects its capacity to host guests. The variable `review_scores_rating` measures overall guest satisfaction and serves as a proxy for perceived quality. Finally, neighbourhood-level demand is represented using `neigh_group`, a categorical variable that classifies neighbourhoods into high-, middle-, and low-demand groups based on average review activity.

The primary research question guiding this analysis is:

To what extent do accommodation capacity, review scores, and neighbourhood demand explain variation in nightly Airbnb prices in Toronto?

To address this question, we use multiple linear regression with `log(price)` as the response variable. In addition, a hypothesis test is conducted to compare average prices across neighbourhood demand groups, providing complementary inferential evidence.

3 Exploratory Data Analysis

Exploratory data analysis was conducted to understand the distribution of key variables and to assess whether linear regression is an appropriate modelling framework. This section summarizes only the patterns most relevant to the final analysis.

3.1 Distribution of Nightly Price

```
ggplot(DataClean, aes(x = price)) +
  geom_histogram(bins = 40, fill = "grey70", color = "white") +
  labs(
    title = "Distribution of Nightly Airbnb Prices",
    x = "Price (CAD)",
    y = "Count"
  )
```

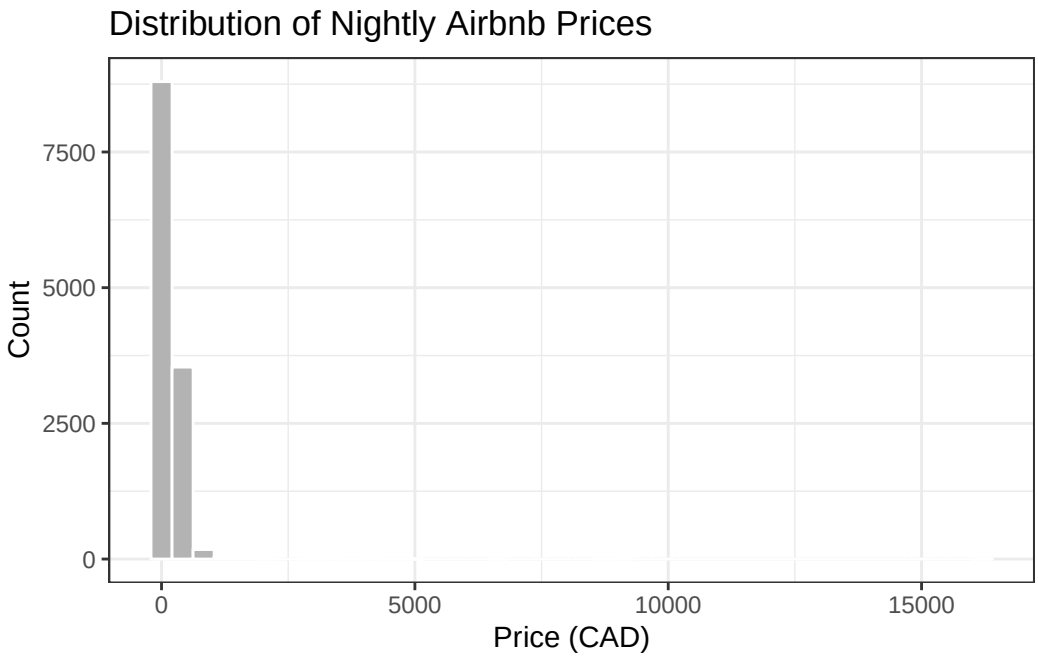

Figure 1: Distribution of Nightly Aribnb Prices

Nightly price exhibits strong right-skewness, with most listings priced below approximately $200 and a long upper tail of higher-priced listings. The presence of this skewness and unequal spread motivates the use of a log transformation of price in the regression model.

3.2 Price and Accommodation Capacity

```
ggplot(DataClean, aes(x = accommodates, y = price)) +
geom_point(alpha = 0.25) +
geom_smooth(method = "lm", se = FALSE) +
labs(
title = "Price vs. Accommodation Capacity",
x = "Accommodates",
y = "Price (CAD)"
)
```

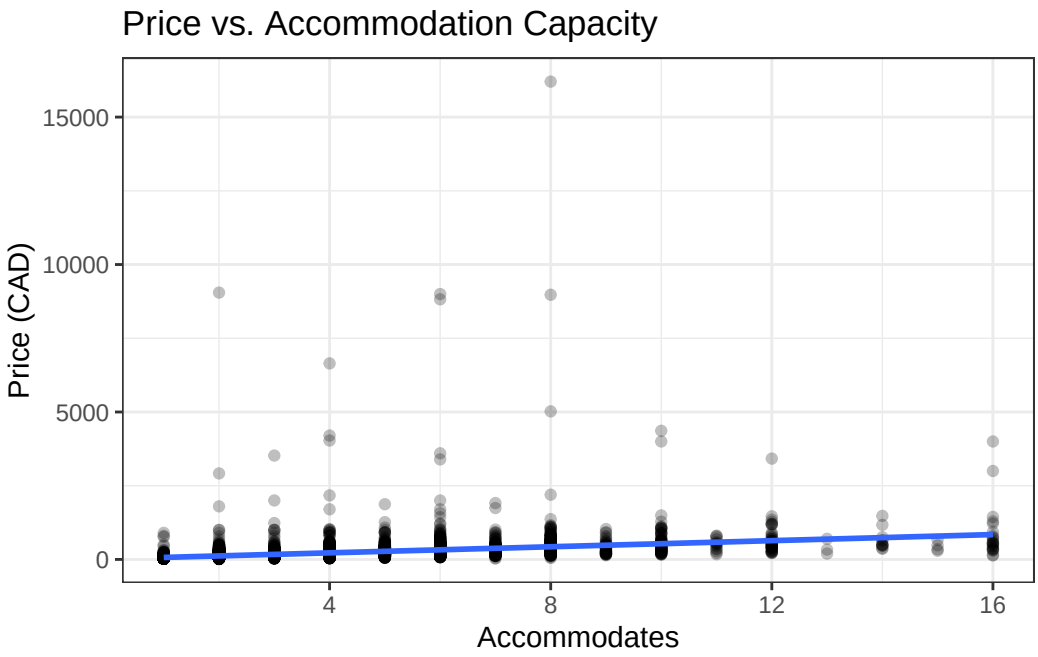

Figure 2: Price vs Accommodation capacity scatterplot

Listings that accommodate more guests tend to charge higher nightly prices. The positive association supports including accommodates as a predictor in the regression model. The relationship appears more linear when price is viewed on a logarithmic scale, further justifying the transformation.

3.3 Price Across Neighbourhood Demand Groups

```
ggplot(DataClean, aes(x = neigh_group, y = price)) +
geom_boxplot(fill = "grey70") +
labs(
title = "Price by Neighbourhood Demand Group",
x = "Neighbourhood Demand Group",
y = "Price (CAD)"
)
```

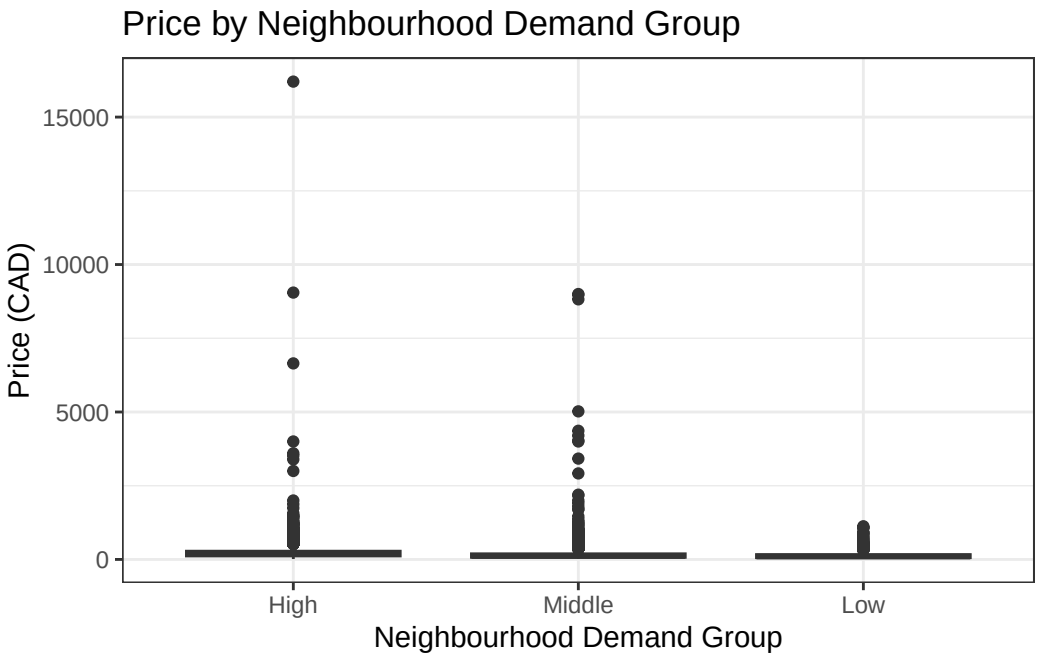

Figure 3: Price by Neighbourhood Demand Group

Prices differ across neighbourhood demand groups. Listings in high-demand neighbourhoods tend to have higher median prices and greater variability, while low-demand neighbourhoods exhibit lower and more concentrated price distributions. This pattern supports including neigh_group as a categorical predictor.

3.4 Price and Review Scores

```
ggplot(DataClean, aes(x = review_scores_rating, y = price)) +
geom_point(alpha = 0.25) +
geom_smooth(method = "lm", se = FALSE) +
labs(
title = "Price vs. Review Scores Rating",
x = "Review Score Rating",
y = "Price (CAD)"
)
```

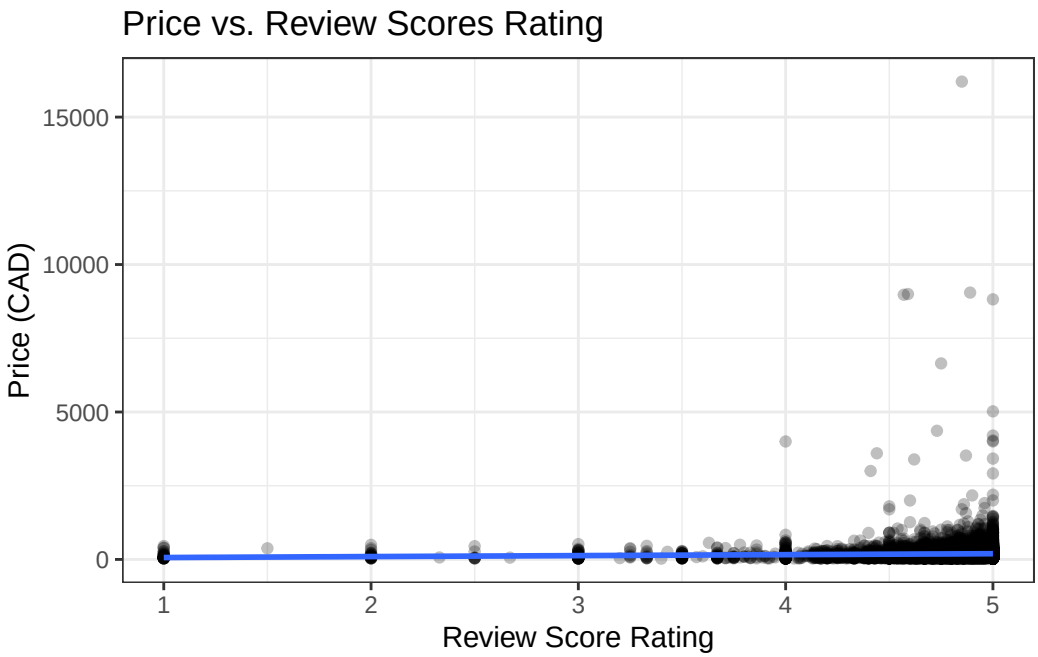

Figure 4: Price vs Review Scores Rating

Higher review scores are associated with slightly higher prices, though the relationship is weaker than that observed for accommodation capacity. This suggests that review scores may contribute modestly to explaining price variation.

3.5 EDA-to-Model Connection

The EDA findings directly inform the modelling decisions. The strong right-skew in price motivates modelling log(price) rather than raw price values. The positive relationship between price and accommodates supports its inclusion as a key predictor, while observed differences across neighbourhood demand groups justify incorporating neigh_group as a categorical variable. Review scores show a weaker but positive association with price, making them a reasonable secondary predictor. Together, these observations support the use of a multiple linear regression model with log(price) as the response.

3.6 Refinements from Part 2 EDA

The exploratory data analysis in Part 3 builds on Part 2 but reflects several refinements made for the purposes of final modeling. In Part 2, more aggressive filtering and outlier removal were applied to support exploratory visualization and to better understand the structure of the data. For the final analysis, the data cleaning procedure was streamlined to retain a larger and more representative sample of active listings, while still removing observations with missing or invalid values for the model variables.

In addition, the analysis in Part 3 uses a log transformation of nightly price rather than the raw price values examined in Part 2. The Part 2 EDA revealed strong right-skewness and extreme

variability in prices, which can distort linear regression estimates. Using log(price) reduces the influence of extreme values, improves the approximate normality of residuals, and leads to more stable and interpretable regression coefficients. As a result, the refined EDA in Part 3 focuses on relationships involving the transformed response variable that directly motivate the final regression model.

4 Methodology

4.1 Regression Model

To examine how listing characteristics and neighbourhood demand relate to nightly Airbnb prices, we use a multiple linear regression model. Linear regression is appropriate because the response variable is quantitative and the goal is to assess how several predictors jointly explain variation in price.

Based on the exploratory analysis, the response variable is defined as the natural logarithm of nightly price, `log(price)`. The log transformation is used to reduce right-skewness, stabilize variance, and improve the linearity of relationships between the response and predictors.

The final regression model is specified as:

$$\log(\text{price}) = \beta_0 + \beta_1(\text{accommodates}) + \beta_2(\text{review\_scores\_rating}) + \beta_3(\text{neigh\_group}) + \varepsilon$$

where: - `accommodates` represents the maximum number of guests the listing can host, - `review_scores_rating` measures overall guest satisfaction, - `neigh_group` is a categorical variable indicating neighbourhood demand level (High, Middle, Low), - and ( ) represents random error.

4.2 Model Assumptions

The linear regression model relies on several key assumptions discussed in class:

1. **Linearity** — The relationship between each predictor and `log(price)` is approximately linear.
2. **Independence** — Each listing is treated as an independent observation.
3. **Constant variance** — The variance of residuals is approximately constant across fitted values.
4. **Normality of residuals** — Residuals are approximately normally distributed.

These assumptions are evaluated using diagnostic plots in the Results section.

4.3 Hypothesis Test

In addition to the regression analysis, a hypothesis test is conducted to compare average prices across neighbourhood demand groups. Specifically, we test whether the mean `log(price)` differs between High-demand and Low-demand neighbourhoods.

Let: - μ_{High} be the population mean of `log(price)` in High-demand neighbourhoods.

- μ_{Low} be the population mean of `log(price)` in Low-demand neighbourhoods.

Hypotheses:

$$H_0 : \mu_{High} = \mu_{Low}$$

$$H_A : \mu_{High} \neq \mu_{Low}$$

Because the two groups may have unequal variances, we use Welch's two-sample t-test. This test complements the regression analysis by providing direct inferential evidence on neighbourhood-level price differences.

4.4 Alternative Approaches Considered

Other modelling approaches, such as including additional predictors or using more complex regression techniques, were considered. However, given the course scope and the exploratory findings, a multiple linear regression with a log-transformed response provides a clear and interpretable framework that aligns with the assumptions and methods covered in class.

5 Results

```
#| label: fit-final-model

model_final <- lm(
  log(price) ~ accommodates + review_scores_rating + neigh_group,
  data = DataClean)
```

5.1 Linear Regression Results

Fitting the model

```
model_final |>
  broom::tidy() |>
  mutate(
    p.value = ifelse(p.value < 0.001, "< 0.001", format(p.value, scientific = TRUE, digits = 3)
  ) |>
  kable(
    digits = c(0, 4, 4, 2, 50),  # 50 digits for p-value shows scientific notation
    col.names = c("Term", "Estimate", "Std. Error", "t-value", "p-value"),
    align = c("l", "r", "r", "r", "r"),
    booktabs = TRUE
  )
```

Table 1: Estimated regression coefficients for the final model.

Term	Estimate	Std. Error	t-value	p-value
(Intercept)	3.5580	0.0597	59.61	< 0.001
accommodates	0.2178	0.0022	98.24	< 0.001
review_scores_rating	0.1825	0.0123	14.87	< 0.001
neigh_groupMiddle	-0.3330	0.0097	-34.39	< 0.001
neigh_groupLow	-0.4759	0.0174	-27.34	< 0.001

Interpretation of coefficients:

Because the response variable is log-transformed, we express the results in dollar terms using the median listing price of $139 as a reference point.

- For every one unit increase of guests in accommodates, the price increases by approximately **$35**.
- For every one unit increase in review_scores_rating, the price increases by approximately **$45**.
- neigh_group is categorical predictor so "R" automatically chooses "High-demand neighbourhood" as reference group(alphabetical order).

- Listing in low-demand neighbourhoods cost **$51** less on average compared to high-demand neighbourhood.
- Listing in middle-demand neighbourhoods cost **$46** less on average compared to high-demand neighbourhood.

Statistical Significance of Individual Coefficients:

- accommodates: t = 94.67, p < 2e-16 → Reject H_0
- review_scores_rating: t = 13.46, p < 2e-16 → Reject H_0
- neigh_groupLow: t = -27.00, p < 2e-16 → Reject H_0
- neigh_groupMiddle: t = -36.82, p < 2e-16 → Reject H_0

Conclusion: All predictors significantly contribute to explaining price variation.

```
model_final |>
  broom::glance() |>
  select(r.squared, adj.r.squared, sigma, statistic, p.value) |>
  mutate(across(where(is.numeric), ~ round(.x, 3))) |>
  kable()
```

Table 2: Model summary statistics.

r.squared	adj.r.squared	sigma	statistic	p.value
0.494	0.493	0.512	3062.411	0

The fitted regression model is:

$$\log(\text{price}) = \beta_0 + \beta_1(\text{accommodates}) + \beta_2(\text{review\_scores\_rating}) + \beta_3(\text{neigh\_group}) + \varepsilon$$

Holding all other predictors constant, the estimated coefficients indicate that listings accommodating more guests tend to charge higher nightly prices. Because the response variable is log-transformed, we express the results in dollar terms using the median listing price of $139 as a reference point.

Specifically, each additional guest accommodated is associated with an average increase of about 22% which translates to about **$35** in nightly price. Higher review scores are also associated with higher prices: a one-point increase in review score rating corresponds to an approximate 20% increase, or about **$28** in price.

Neighbourhood demand is an important predictor. Relative to listings in high-demand neighbourhoods, listings in middle-demand areas charge approximately 33% or about **$46** less on average, while listings in low-demand neighbourhoods charge roughly 48% or about **$53** less, holding other factors constant.

Overall, the model explains approximately 50% of the variation in nightly Airbnb prices, indicating a strong relationship between price and the selected predictors.

5.2 Diagnostic Plots

```
autoplot(model_final)
```

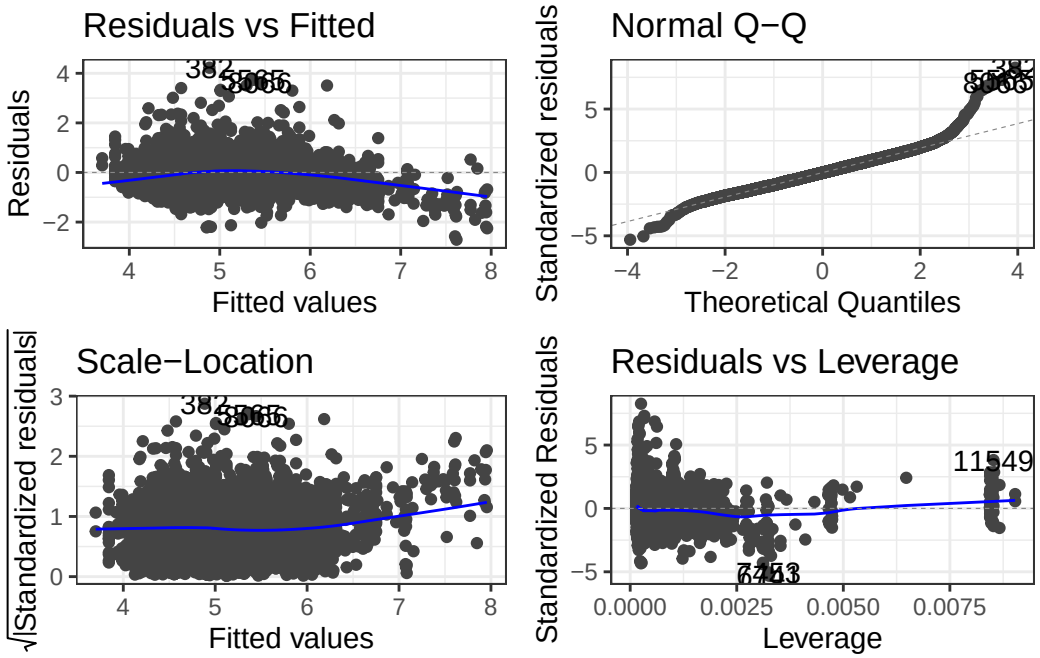

Figure 5: Diagnostic plots for the final regression model.

The diagnostic plots suggest that the assumptions of linear regression are reasonably met. The residuals versus fitted values plot does not show strong non-linearity, and the scale-location plot indicates roughly constant variance. The normal Q–Q plot shows moderate departures in the tails, which is expected given the skewed nature of price data, but no severe violations are evident. No single observation appears to exert undue influence on the fitted model.

5.3 Hypothesis Test

We also test whether the average nightly price differs between listings in high-demand and low-demand neighbourhoods. Because the two neighbourhood groups may have unequal variances, we use Welch's two-sample t-test to compare the mean log-transformed prices. The test assumes independent observations and approximate normality of the sample means, which is reasonable given the large sample sizes.

```
t_test_neigh <- t.test(
  log(price) ~ neigh_group,
  data = DataClean |> filter(neigh_group %in% c("High", "Low"))
)

t_test_neigh
```

```
    Welch Two Sample t-test

data:  log(price) by neigh_group
t = 24.005, df = 1472.6, p-value < 2.2e-16
alternative hypothesis: true difference in means between group High and group Low is not equal
95 percent confidence interval:
 0.5208060 0.6134965
sample estimates:
mean in group High  mean in group Low
          5.208163           4.641011
```

The test indicates a statistically significant difference in mean log(price) between high-demand and low-demand neighbourhoods. The estimated mean difference is positive, with a p-value well below 0.05, providing strong evidence that listings in high-demand neighbourhoods charge higher nightly prices on average.

5.4 Refinement of the Hypothesis Question

In Part 2, the hypothesis question focused on whether high-demand and low-demand neighbourhoods differed in average accommodation capacity or average review rating. This exploratory comparison was useful for understanding whether neighbourhood demand was associated with differences in listing characteristics.

For the final analysis in Part 3, the hypothesis test was refined to focus directly on the response variable of interest—nightly price. By testing whether average prices differ between high-demand and low-demand neighbourhoods, the hypothesis test aligns more closely with the regression model and provides complementary inferential evidence regarding neighbourhood-level pricing differences.

6 Discussion

The regression results provide clear insight into the factors associated with nightly Airbnb prices in Toronto. As shown in (see Table 1), accommodation capacity is the strongest predictor in the model. Holding all other variables constant, listings that can host more guests tend to charge substantially higher prices. This finding is intuitive, as larger properties typically offer more space and amenities, making them more valuable to guests.

Review scores are also positively associated with price, although the effect is smaller. Higher-rated listings command a modest price premium, which is consistent with the idea that guest satisfaction reflects perceived quality. While the relationship is weaker than that of accommodation capacity, it remains statistically significant and contributes meaningfully to explaining price variation.

Neighbourhood demand plays an important role in pricing. Relative to high-demand neighbourhoods, listings in middle- and low-demand areas charge noticeably lower prices on average (see Table 1). This aligns with expectations that location and local demand influence what hosts are able to charge, even after controlling for listing size and review quality. The additional hypothesis test comparing

high- and low-demand neighbourhoods reinforces this conclusion by providing direct inferential evidence of a difference in average prices between these groups.

Overall, the model explains approximately half of the variation in nightly price (Table Table 2), which is substantial given the limited number of predictors included. This suggests that while accommodation capacity, review scores, and neighbourhood demand are key drivers of price, other unobserved factors likely also contribute to pricing decisions.

Several limitations should be noted:

- **Selection bias:** The analysis focuses on active listings with reviews, which may exclude newer or less active properties and could introduce selection bias.
- **Independence of observations:** The independence assumption may be violated if some hosts manage multiple listings or if listings within the same building share unobserved characteristics.
- **Normality of residuals:** Although the diagnostic plots in Figure Figure 5 suggest that the linear regression assumptions are largely satisfied, some deviation from normality remains in the residuals, reflecting the inherent variability in nightly price data.

Finally, important predictors such as room type, amenities, proximity to downtown, and seasonal effects are not included in the model. Incorporating these variables could improve explanatory power and provide a more complete picture of Airbnb pricing dynamics in future analyses.

7 Conclusion

This project examined how accommodation capacity, review scores, and neighbourhood-level demand are associated with nightly Airbnb prices in Toronto. Using data from the Inside Airbnb project and a multiple linear regression model with a log-transformed price response, we assessed the relative importance of these factors in explaining price variation across listings.

The results show that accommodation capacity is the strongest predictor of price, with larger listings charging substantially higher nightly rates. Review scores are also positively associated with price, indicating that higher-rated listings command a modest premium. In addition, neighbourhood demand plays a significant role: listings in high-demand areas charge higher prices on average than those in middle- or low-demand neighbourhoods. These findings are supported both by the regression analysis and by a supplementary hypothesis test comparing prices across neighbourhood groups.

Overall, the analysis highlights the importance of both property characteristics and location in determining Airbnb prices. While the model explains a substantial portion of price variation, it also suggests that additional factors such as amenities, room type, and proximity to key areas could further improve understanding of pricing dynamics. These results provide a useful foundation for future analyses of short-term rental markets and pricing behavior.

8 References

[1] Inside Airbnb, "Inside airbnb." Accessed: Oct. 27, 2025. [Online]. Available: http://insideairbnb.com

9 Acknowledgment

We thank all members of Group 12 for their collaboration and contributions throughout the data preparation, analysis, and writing of this report. OpenAI's ChatGPT was used to assist with code debugging, table formatting, and grammar refinement during the preparation of this report. All statistical analyses, interpretation of results, and final writing decisions were made by the authors.