Information and Airbnb, An Analysis on the Impact of Reviews and Ratings on Airbnb Prices in Canada*

Alaina Hu and Ahad Qureshi

February 14, 2024

Table of contents

1 Introduction

In the changing landscape of modern travel, Airbnb has emerged as a transformative force, reshaping traditional notions of accommodation and hospitality. No longer are travelers confined to hotels; people have the choice to choose between various forms of accommodation and find the option most suitable for their needs. With Airbnb's vacation rentals, travelers have the option to gain access to more space, kitchens, home amenities, and lower cost (Guttentag, 2016). Central to Airbnb's allure are the wealth of user-generated reviews and ratings, which serve as vital sources of information for prospective guests navigating a vast array of listings. These reviews not only offer insights into the quality and character of accommodations but also play a pivotal role in shaping consumer decisions. However, in this growing world of shared experiences, a fundamental question persists: What effect do these reviews and ratings have on Airbnb pricing strategies? This question lies at the heart of our research, as we delve

*Code and data are available at: https://github.com/alainahu/airbnb_analysis

into the relationship between information, consumer behavior, and pricing dynamics within the Airbnb marketplace in Canada. Through statistical analysis and data visualization, this paper endeavors to continue to explore the relationship between reviews, ratings, and Airbnb prices.

In an econometric analysis on Airbnb reviews and price in Boston, Lawani et al. find that reviews serve as a good proxy for rooms' quality and reviews affect both the host price and neighboring host price. Peng et al. (2020) build a machine learning model and use natural language processing to predict Airbnb prices in nine cities across the United States, Australia, and Europe. Their research concludes that customer reviews, house features, and geographical data are all predictive factors for Airbnb rental prices. While there is an extensive amount of statistical and machine learning research focused on the relationship between Airbnb reviews and prices, this paper fills in the gap in literature by focusing on the information-price relationship within Canada, specifically the cities of Toronto and Vancouver. Existing research often focuses on cities in the United States or Europe. We are interested in seeing if the same relationship between reviews and price can be found when we zoom into Toronto and Vancouver. Regional differences should be examined: Ghosh et al. (2023) find that predictability of Airbnb prices varies significantly across cities. For example, listing prices are the most predictable in Boston and least predictable in Chicago.

Our paper will follow a replication of Laouénan and Rathelot's research on the effects of information on ethnic discrimination in the Airbnb market. We follow their paper to replicate the following claims (1) the platform is effective in supplying useful information, so reviews and ratings impact the expectations of consumers and (2) there is a relationship between number of reviews and listed price. Data analysis for this reproduction is performed in `R` [@citeR], and additional help is provided by libraries such as `dplyr` [@dplyr], `ggplot2` [@ggplot], `ggrepel` [@ggrepel], `tidyverse` [@thereferencecanbewhatever], `kableExtra` [@kableextra], `knitr` [@knitr], `haven` [citehaven], 'readr',

2 Data

2.1 Source

The paper titled "Can Information Reduce Ethnic Discrimination? Evidence from Airbnb" [@citepaper] published by the American Economic Association [@citewebsite] is the focus of our paper. We reproduce and investigate several areas of the original paper, using the same datasets for our analysis.

2.2 Methodology

The dataset was compiled from the publicly accessible listings on Airbnb's platform, capturing details visible on the initial listing page. [@citepaper]. The collection process was repeated

every two to three weeks between June 2014 and June 2015 with an additional collection in November 2017, resulting in 21 total data collection waves. [@citepaper]. Following Laouénan and Rathelot's methodology, our analysis focuses on listings that received at least one review during the observation period, underlining the inclusion of only actively engaged listings. [@citepaper] This criterion narrowed the original dataset from 663,090 to 220,939 listings. For our specific interest in the Canadian market, we further refined the dataset to include listings solely from Toronto and Vancouver, as detailed in Table 1.

We employ a methodological approach akin to Altonji and Pierret (2001), observing both the quantity and quality of information available about a property to potential guests, which could influence their decision-making process.[@cite]

2.3 Features

The original research spanned 19 cities across North America and Europe, characterized by high listing volumes. Our study narrows this focus to Canadian cities, specifically Toronto and Vancouver, offering a localized perspective on the data while broadening the research question.

```
Rows: 87059 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): city
dbl (8): review, guest_satisfaction_overall, accuracy_rating, cleanliness_ra...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Figure 1 illustrates the broad spectrum of daily rental prices, highlighting significant price variability across listings. To mitigate outlier effects, we excluded the top and bottom 1% of the price range. The price distribution reveals a first quartile at $69, a median of $94, and a third quartile at $126 per night, indicating a skewed distribution with a mean price of $108.

```
Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.
```

```
Warning: Removed 102 rows containing non-finite values (`stat_bin()`).
```

```
Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

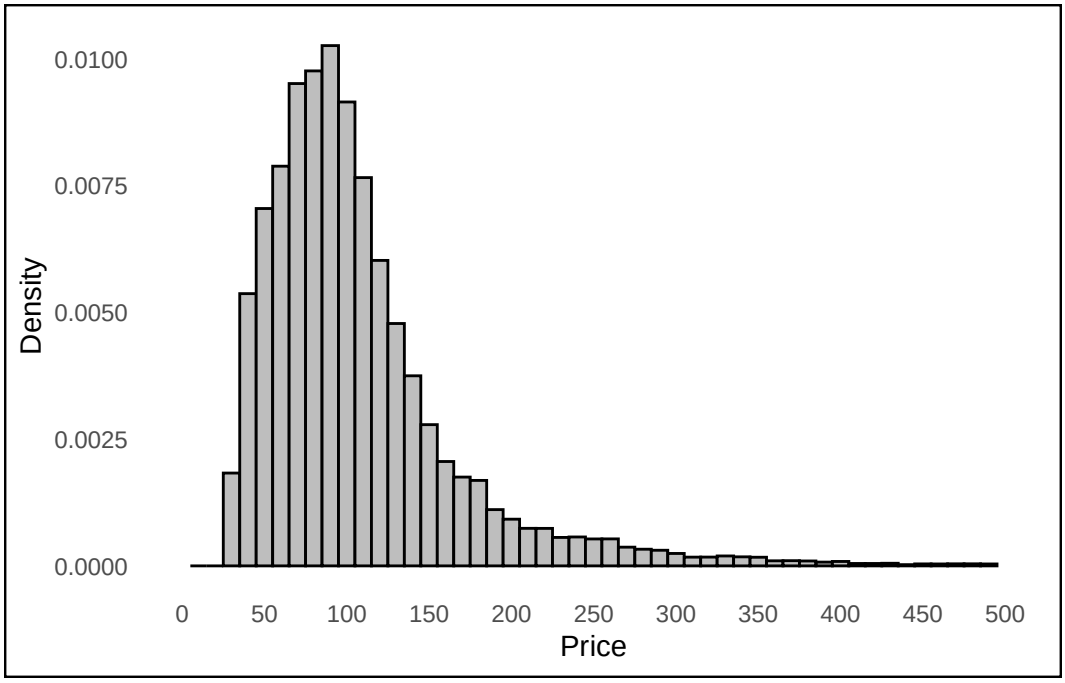


Figure 1: Distribution of Airbnb prices per night (including cleaning fees)

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
26.00   68.67   93.83  107.66  126.33  634.33
```

Our analysis emphasizes the importance of review quantity and quality in correlating information availability with listing prices. Utilizing the most recent rating for each property, which aggregates all received ratings over its Airbnb tenure, we noted a skewness towards higher ratings, a finding consistent with previous research by Fradkin, Grewal, and Holtz (2018). [@cite]

From the original dataset's 183 variables, we focused on 8 key variables deemed most relevant to our research question regarding information's impact on pricing:

1. Review: Represents the number of reviews received by the listing.It is the feedback provided by guests after their stay or experience.
2. Guest Satisfaction: A numerical rating indicating the overall satisfaction level of guests.
3. Accuracy Rating: Measured how well the description and images of a listing match the actual experience. Higher ratings suggest that guests found the property as advertised.
4. Cleanliness Rating: Assesses the cleanliness of the accommodation or service provided. It reflects guests' perceptions of the hygiene standards maintained at the property.

5. Location Rating: Rates the convenience, desirability, or attractiveness of the property's location, considering factors like proximity to tourist attractions, amenities, transport links, and the overall neighborhood.
6. Value Rating: Evaluates guests' perceptions of the worthiness of the service or accommodation relative to the price paid. It considers whether guests feel they received good value for their money.
7. New Price: Refers to the updated price of the listings after information and consequently due to the law of demand and supply.
8. Log Price: Relative change in prices due to availability of information

```
summary_df <- airbnb |>
  summarise(
    Mean_New_Price = mean(new_price, na.rm = TRUE),
    Median_New_Price = median(new_price, na.rm = TRUE),
    SD_New_Price = sd(new_price, na.rm = TRUE))
```

Table 1: Summary Statistics for the Reviews and Ratings

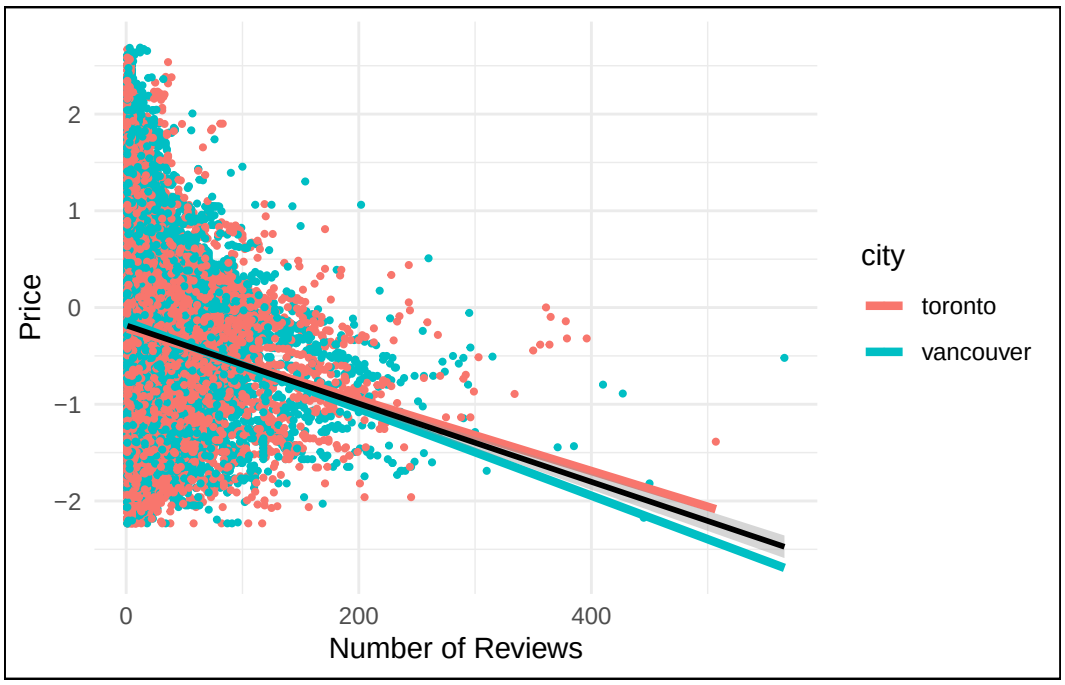
Variable	Mean	Median	Standard Deviation	Min	Max
Accuracy Rating	9.47	10	0.98	0	10
Cleanliness Rating	9.28	10	1.14	0	10
Guest Overall Satisfaction	93.47	95	7.78	20	100
Location Rating	9.48	10	0.95	0	10
Number of Reviews	14.90	7	23.86	1	566
Value Rating	9.29	9	1.00	0	10

«««< HEAD These variables were chosen for their direct relevance to our investigation into how informational transparency can affect Airbnb pricing strategies, with all variables being quantitative to facilitate our analysis.

======= »»»> b92cfdb (Started writing introductio)

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

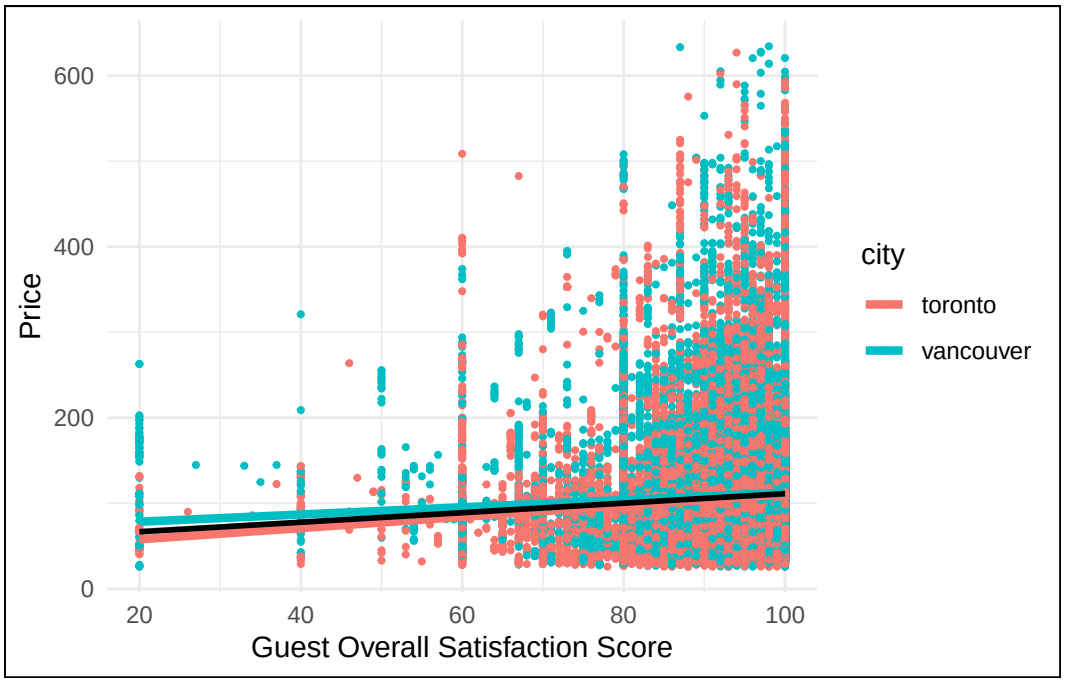
```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```




```
<ggproto object: Class ScaleDiscrete, Scale, gg>
    aesthetics: colour
    axis_order: function
    break_info: function
    break_positions: function
    breaks: waiver
    call: call
    clone: function
    dimension: function
    drop: TRUE
    expand: waiver
    get_breaks: function
    get_breaks_minor: function
    get_labels: function
    get_limits: function
    guide: legend
    is_discrete: function
    is_empty: function
    labels: waiver
    limits: NULL
    make_sec_title: function
    make_title: function
    map: function
```

```
    map_df: function
    n.breaks.cache: NULL
    na.translate: TRUE
    na.value: NA
    name: waiver
    palette: function
    palette.cache: NULL
    position: left
    range: environment
    rescale: function
    reset: function
    scale_name: brewer
    train: function
    train_df: function
    transform: function
    transform_df: function
    super:  <ggproto object: Class ScaleDiscrete, Scale, gg>
```

```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```

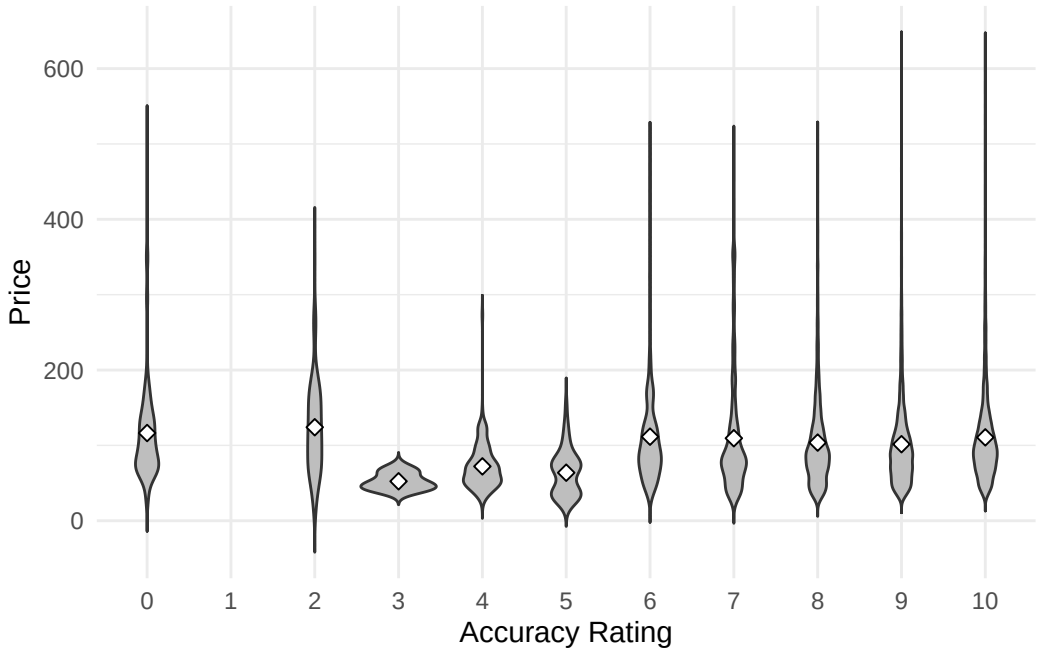


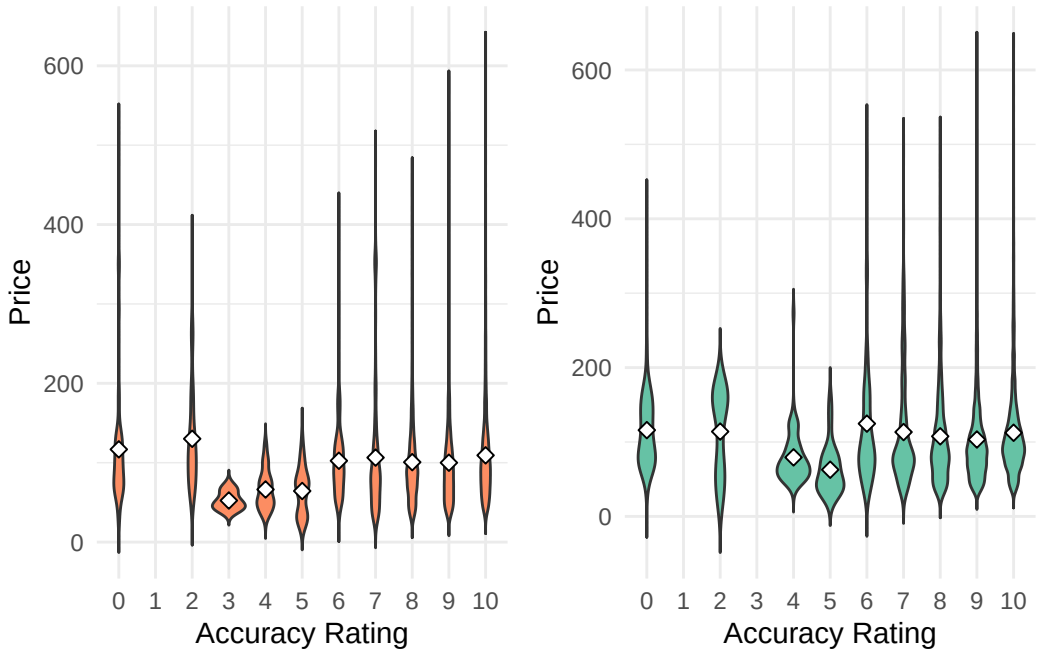

```
<ggproto object: Class ScaleDiscrete, Scale, gg>
```

```
aesthetics: colour
axis_order: function
break_info: function
break_positions: function
breaks: waiver
call: call
clone: function
dimension: function
drop: TRUE
expand: waiver
get_breaks: function
get_breaks_minor: function
get_labels: function
get_limits: function
guide: legend
is_discrete: function
is_empty: function
labels: waiver
limits: function
make_sec_title: function
make_title: function
map: function
map_df: function
n.breaks.cache: NULL
na.translate: TRUE
na.value: grey50
name: waiver
palette: function
palette.cache: NULL
position: left
range: environment
rescale: function
reset: function
scale_name: manual
train: function
train_df: function
transform: function
transform_df: function
super:  <ggproto object: Class ScaleDiscrete, Scale, gg>
```

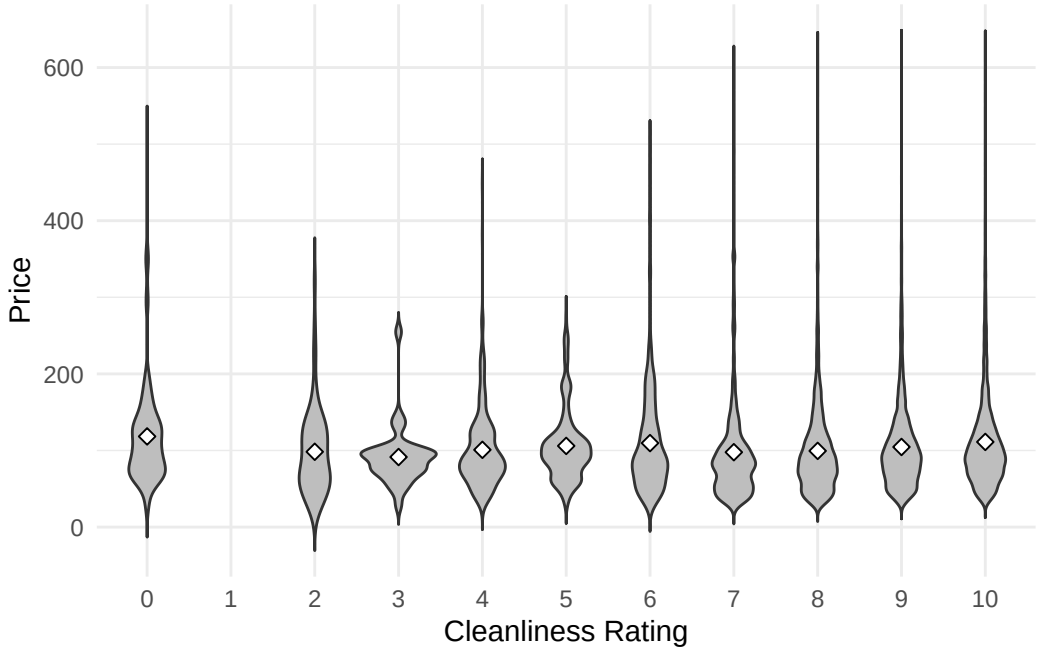
```r
unique_values <- unique(airbnb$accuracy_rating)

# Print the unique values
print(unique_values)
```

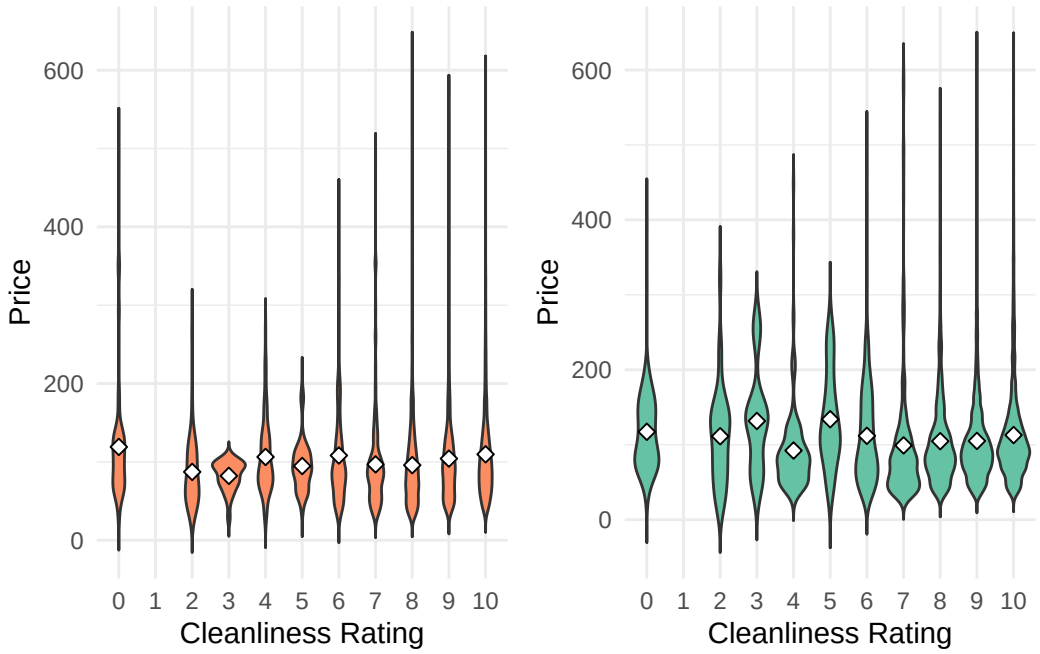
```
[1] 10  9  8  7  0  6  2  5  4  3
```



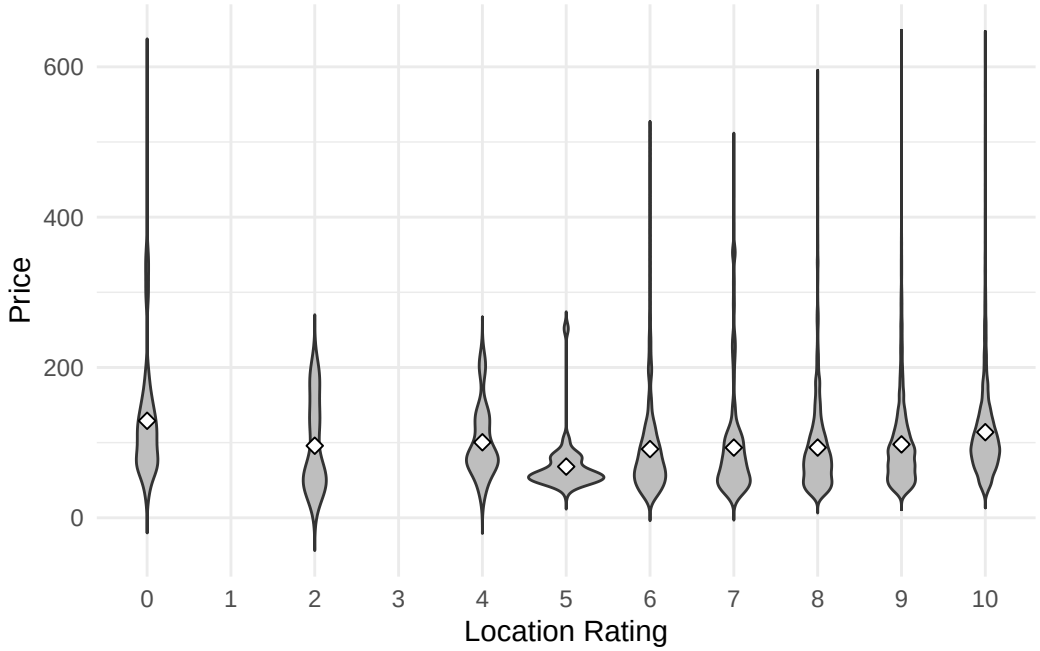

Price
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10
Accuracy Rating

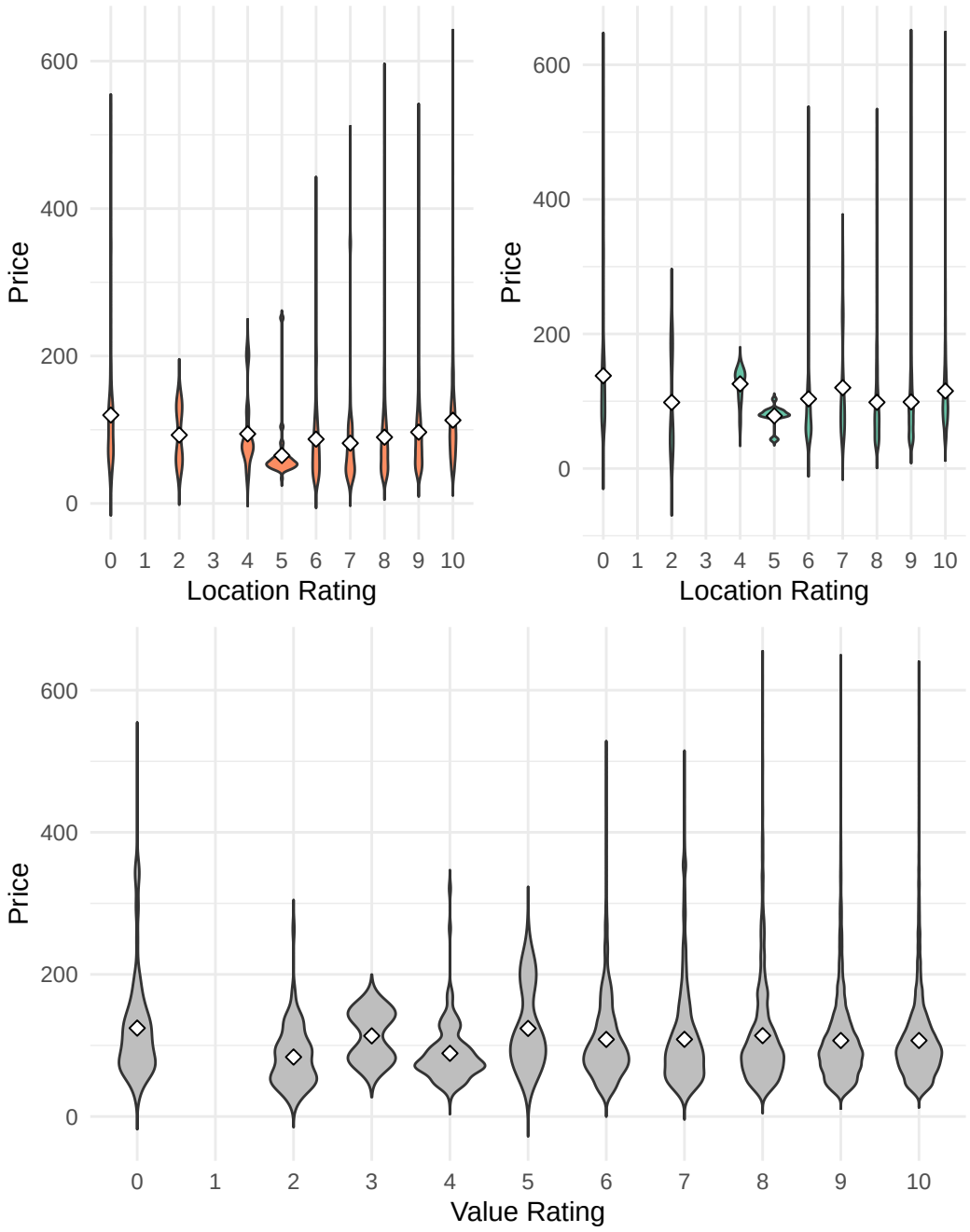
Price
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10
Accuracy Rating


Price
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10
Cleanliness Rating

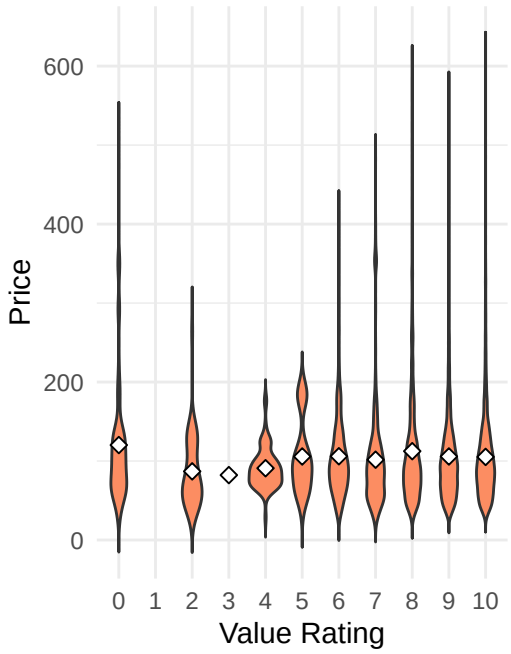

Price
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10
Cleanliness Rating

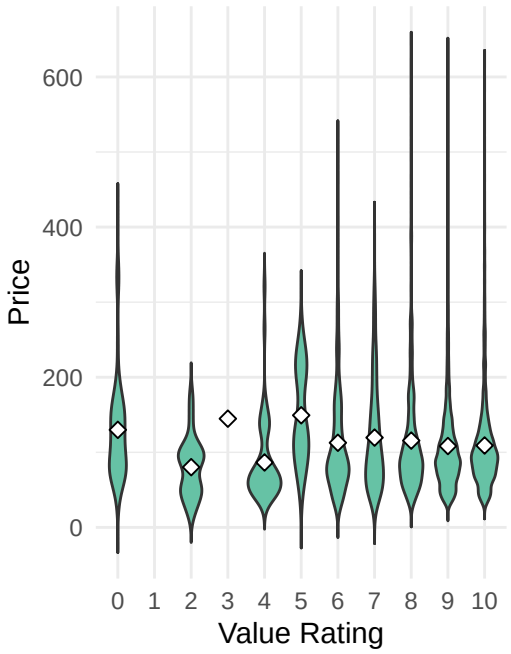
Price
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10
Cleanliness Rating


Price
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10
Location Rating




```
Warning: Groups with fewer than two data points have been dropped.
Groups with fewer than two data points have been dropped.
```


Price
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10
Value Rating


Price
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10
Value Rating