Project Proposal

Linearity Legends - Barron Brothers, Jack Steel, Mert Bildirici, Kethan Poduri

```
listings <- read.csv('data/listings.csv')
```

Introduction

Airbnbs have historically been known as a cheaper (and better) alternative to hotels. As the Airbnb market has evolved, Airbnbs have typically been getting more expensive for solo travelers and couples. If many people rent an Airbnb for long periods of time, though, it could be very economical.

When people are travelling, it is important for them to get the best price and best experience possible. These two factors can be difficult to balance, though. As an Airbnb is more highly rated, the owner could decide to raise the price due to higher demand. People can also rate on multiple factors, such as the availability, the amount of the room rented, and/or the location.

The research question we will focus on is as follows: For Airbnbs in Austin, TX, **What is the relationship between price and customer ratings with respect to availability, room type, and the number of reviews a listing has?**

We predict that as availability decreases, the number of reviews a listing has increases, and as room type becomes larger and more private, and customer ratings become more positive, price will increase.

Data description

The source of the dataset is here. While the data itself is displayed on Kaggle, the data is derived from a clear external source: InsideAirbnb. Our team has cross-validated the data to make sure that it is accurate. The data set itself is updated frequently, with the last update taking place on 07 Oct. 2023. The data have been collected since 2020, but prior dates are included.

There are 14,681 observations in the dataset, 4,381 of which have `number_of_reviews` ≥ 30. Generally, many of the useful variables have to do with the ratings of an observation (one

Airbnb), the location of the observation, or describing the observation itself (i.e. `price`, `name`, ... etc.)

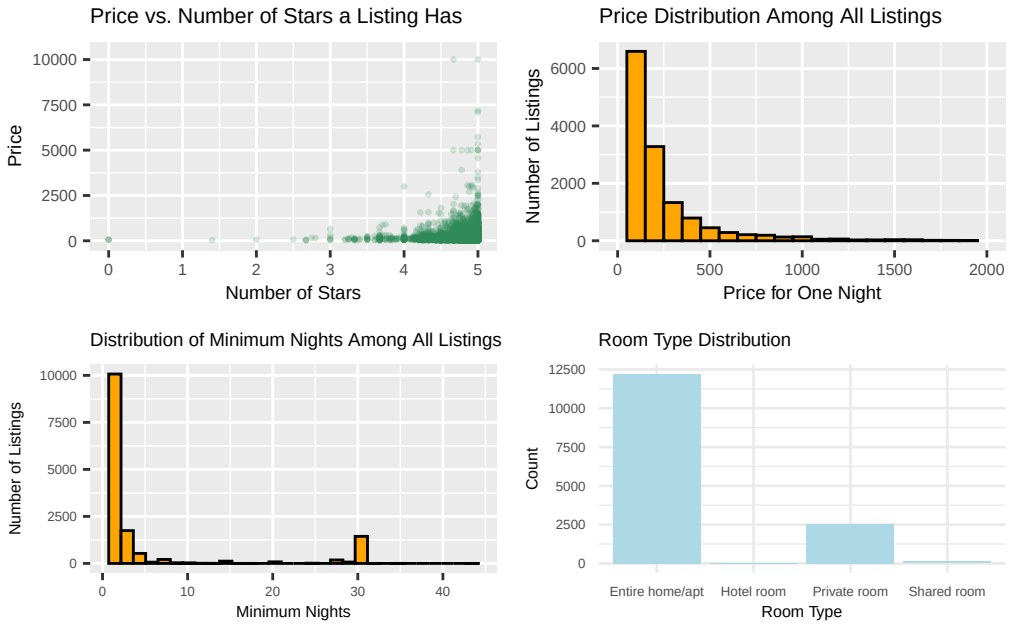
Initial exploratory data analysis

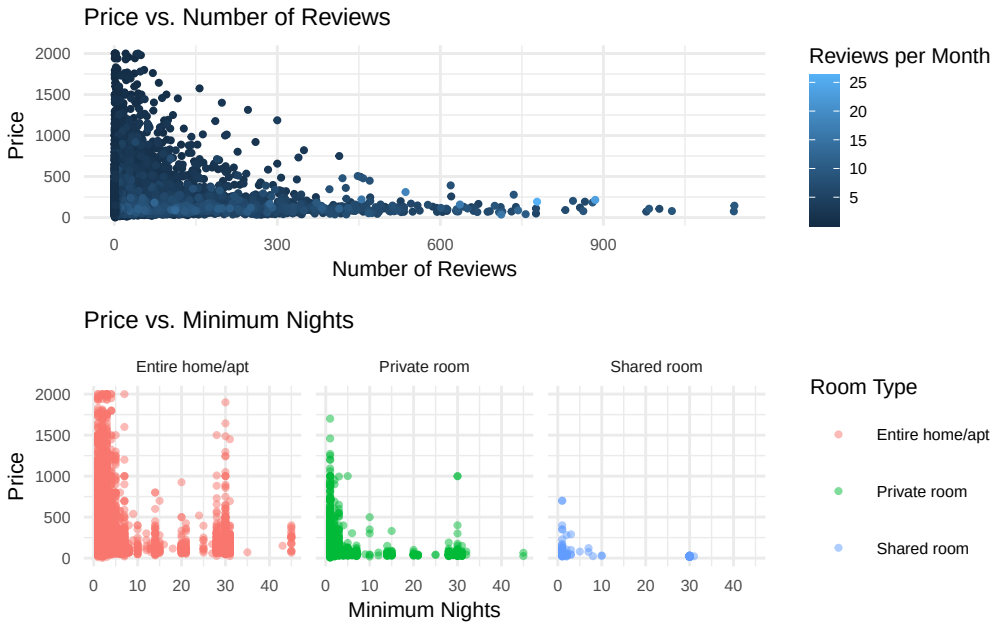
Used stack overflow to extract the rating from the `names` column, in which there are multiple variables presented.

Created a new column for the number of stars a listing has:

```
listings$stars <- as.numeric(gsub(".* (\\d+\\.\\d+).*", "\\1", listings$name))
```

Selected Visualizations





We are interested in exploring the interaction effect between minimum nights and availability. Although both are predictor variables, I expect there to be some interaction effect since the number of minimum nights might affect the availability.

NB: To make the visualizations clearer, we eliminated some outliers with some of our variables. For most of the plots, we limited `price` to be under $2000/night and `minimum_nights` to under 45. This is for visualization purposes only.

Brief Observations From Selected Graphs

Graph 1 (*Price vs. Number of Stars a Listing Has, top left*): There is a moderately strong, positive, non-linear relationship between the price of an Airbnb in Austin, TX and the number of stars the observation is rated. There seem to be possible outliers from about $5,000-$10,000/night and a small cluster around $2,000-$3,000 and 4.8-4.9 stars.

Graph 3 (*Distribution of Minimum Nights Among All Listings, middle-left*): The distribution of the minimum nights required for a rental for Airbnbs in Austin, TX is right-skewed with a median of 2 nights, an IQR of 2 nights (Q1: 1 night; Q3: 3 nights), a minimum of 1 night, and a maximum of 1,124 nights. There are potential outliers above 45 nights (that are cut off for clarity), but there is also a significant spike in `minimum_nights` around 30 nights. Some of these outliers might get filtered out when we consider the `number_of_reviews` and filter based on that. (This applies to any outliers mentioned below.)

Graph 5 (*Number of Reviews, second-to-bottom*): The relationship between the price and the number of reviews for Airbnb rentals in Austin, TX is relatively weak, negative, and non-linear. There are potential outliers above 900 reviews and in the cluster between ~150-450 reviews that has a higher price than similarly-rated Airbnbs. In addition, the vast majority of Airbnbs have a low number of ratings, but the lower-priced Airbnbs tend to have a higher number of reviews per month.

Graph 6 (*Price vs. Minimum Nights, bottom*): There is a relatively weak, negative, non-linear relationship between the minimum nights required and the price for an Airbnb rental in Austin, TX. This relationship is fairly consistent across the different room types, despite different observations. In entire home and private room rentals, there is a spike in both frequency and price variance around 30 nights.

Summary Statistics

Table 1: Price EDA

minimum	q1	median	mean	q3	maximum
1	95	150	269.3	266	19286

Table 2: Minimum Nights EDA

minimum	q1	median	mean	q3	maximum
1	1	2	7.6	3	1124

The prices range from 1 to 19286 dollars a night, which is a wide range. This is right skewed since the mean and the median is around the 100-300 range. There are some houses that are very expensive, which skew the mean to the right and makes the plot appear right-skewed.

Analysis approach

The response variable we are looking to predict is the price of the Airbnb per night for Airbnbs in Austin. We will choose to filter for rentals in Austin due to the extensive amount of Austin rentals within the data set. This will ensure that the rentals will have the price listed in the same currency and allow us to more confidently generalize our results to the larger population.

One of the predictor variables that we are looking to implement in our model is neighborhood, which contains the zip codes for each Airbnb rental in Austin. We plan to use this variable to identify the relationship between the location in Austin and the price of the Airbnb rental.

Another predictor variable that we will use to predict price is the room type, which contains the values entire home/apt (apartment), private room, hotel room, and shared room. Through our data analysis we intend to see the disparities in price between different types of Airbnb rentals.

The next predictor variable we'd consider implementing in our model to predict price is number of reviews, which contains the number of individuals that left a review on the Airbnb rental. With this value we can determine whether rentals with a high number of reviews typically have a high price per night when compared with rentals with a lower number of reviews.

The number of reviews per month is another predictor variable that we were considering incorporating into our model, however we may find that it is correlated with the number of reviews the Airbnb rental has and could lead to multicollinearity. This variable indicates the average number of times a review was left on the Airbnb rental each month since the rental was first listed.

The fifth predictor variable that we look to integrate into our statistical analysis is the calculated number of host listings, which indicates the number of rentals that the host of the given house has ever listed. This is another variable that has complexities to it, as several listings will have the same number of calculated host listings as the host is the same individual with various rentals in the same region or city.

Lastly, we are looking to include the availability out of the 365 days of the year into our model. This value displays how many days of the year an individual can rent the given Airbnb. Incorporating this availability into our model will allow us to determine whether houses with greater availability or less availability are more costly.

Our group intends on doing a multiple linear regression incorporating all or some combination of the variables listed above. We intend to incorporate interaction terms, as certain predictor variables, such as neighborhood, may have distinct values for price depending on the outcome of other predictor variables, such as room type. Therefore we may incorporate several interaction terms using a combination of the six predictor variables list above to create a multiple linear regression predicting price.

Data dictionary

The data dictionary can be found here