Predicting AirBnb Prices in NYC

Nadya, Gustavo & Maria

In the heart of the bustling metropolis that is New York City, a dynamic marketplace of AirBnb listings unfolds, offering a plethora of accommodations to travelers from all corners of the globe. From snug apartments in the heart of Manhattan to expansive lofts in Brooklyn, these listings not only span the spectrum in terms of location but also come with a rich array of features and amenities. Today, we embark on a statistical exploration to uncover the underlying factors that dictate the price of these listings. Our investigative journey will revolve around two pivotal questions:

- **Predicting Airbnb Listing Prices:** Can we forecast the price of an AirBnb listing based on its features?
- **The Anatomy of Price Differences:** What are the discerning characteristics that elevate some listings to a higher price bracket than others?

Data

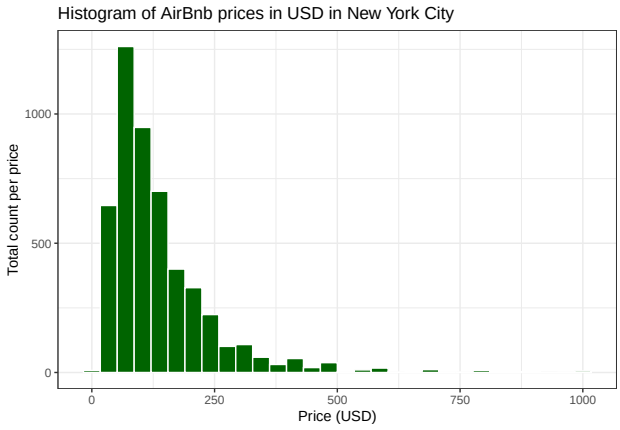
We rely on data from AirBnb, which we joined with borough data for NYC. Each entry represents a different AirBnb listing. The joint data set has more than 40000 entries, and we randomly sampled 5000 for this project. We limited prices to $1000 per night to limit extreme outliers. We selected 24 variables that we believe are good predictors for price, like number of bathrooms, reviews per month, and the property type. Variables include logical, numerical, and strings.

Our outcome variable is price, a numerical variable measured in USD, ranging from $0 to $1000. We constructed a histogram to better understand the distribution of AirBnb prices, and we see a left-skewed histogram around the median price of $100.

Other numerical variables include *latitude and longitude*, *accommodates* (people an AirBnb can host), *maximum_night*s (that an AirBnb can be rented for), and *reviews_scores_rating* (user's rating of an AirBnb), amongst others. Categorical variables include *property_type*, *room_type,* and *neighborhood_group*. We self-coded logical variables using the **stringr** package to detect strings in the *amenities* variable. This variable has 5000 unique values (one for each listing) because it just lists amenities. We selected amenities that we thought would demonstrate a higher price (like having a balcony, a gym, or a doorman).

```
set.seed(253)
airbnb_sub <- airbnb %>%
  left_join(nyc, by=c("neighbourhood_cleansed"="neighbourhood")) %>%
  filter(price >= 0, price <= 1000) %>%
  sample_n(5000) %>%
  select(!id & !require_guest_profile_picture & !host_has_profile_pic & !is_location_exact &
           !calendar_updated & !instant_bookable & !is_business_travel_ready &!bed_type &
           !neighbourhood_cleansed & !square_feet & !host_response_rate & !host_response_time &
           !property_type) %>%
  mutate(has_TV=as.factor(str_detect(amenities, "TV"))) %>%
  mutate(has_Gym=as.factor(str_detect(amenities, "Gym"))) %>%
```

```
  mutate(has_Pets=as.factor(str_detect(amenities, "Pets"))) %>%
  mutate(has_Doorman=as.factor(str_detect(amenities, "Doorman"))) %>%
  select(!amenities)
```

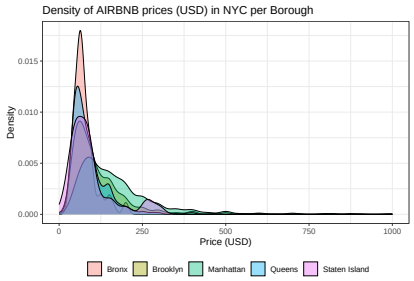


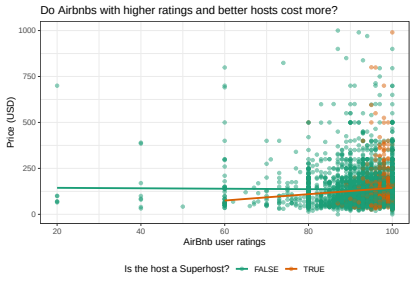
Exploratory Data Analysis

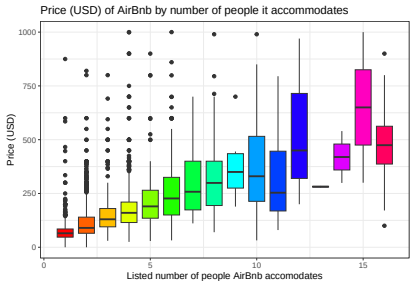
We started our exploratory data analysis through bivariate and multivariate visualizations. This data set includes many variables and we thought best to find possible relationships through data visualizations. Our instincts pointer towards *neighborhood_group* as a strong predictor of price. However, the density plot below shows a similar pattern to our histogram above: a left-skewed plot where all boroughs exhibit similar prices.

We then thought about how a higher user rating might prompt hosts to increase prices, as well as how the quality of a host might impact price. The scatter plot below showed us that there might be some relationship between having a superhost, having a high rating, and having a higher price. However, we had not found a strong relationship yet.

Lastly, we thought about the size of an Airbnb and its price. We removed the variable *square_feet* from our data set because it had virtually no entry points. We took instead our *accommodates* variable for this purpose, assuming that a higher number of guests means a bigger Airbnb. The box plot bellow actually shows a strong positive relationship, and, to our surprise, a rather linear relationship. Using these visualizations, we moved onto creating a Least Squares regression that models price by our 23 predictors.





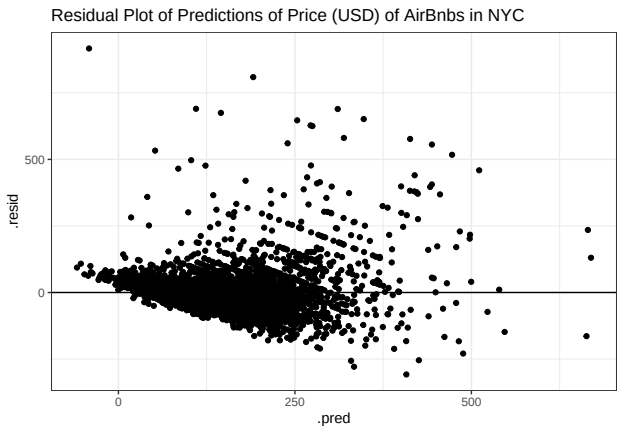


Methods and Models

Least Squares Model

We constructed a Linear Regression model using all predictors in our data set. We decided to include all of them to explore the relationships, coefficients, and p-values that our model presented. We collected the

10-fold Cross-Validated MAE for our model to see how accurate this LS model is. We also created a residual plot (seen below) to see if our predictions were wrong.



LASSO Model

Although our linear model offered appropriate predictions (CV MAE of $42), we feared overfitting to our data. Plus, although 42 is not a big difference in this data set, who wants to be surprised by 42 more dollars per night when staying at an AirBnb?

This prompted us to create a LASSO model to penalize predictors that did not contribute to our model. We created 50 different models with normalized predictors with a tested range of λ values (from 0.01 to 1000). We selected the most parsimonious penalty to ensure a simpler model with the lowest errors possible.

Results

In order to understand the factors that contribute to variations in listing prices, we used the LASSO algorithm. LASSO is a regularization technique that allows us to identify the most influential predictors while minimizing overfitting. The objective was to discern the key features that make certain listings more expensive than others in the Airbnb marketplace.

```
# Filter for the desired predictors
desired_predictors <- c("longitude", "accommodates", "bathrooms", "bedrooms",
                        "availability_30", "review_scores_rating", "reviews_per_month",
                        "property_type_Condominium", "property_type_Loft",
                        "room_type_Private.room", "room_type_Shared.room",
                        "neighbourhood_group_Manhattan", "neighbourhood_group_Staten.Island",
                        "has_TV_TRUE.", "gym_TRUE.", "doorman_TRUE.")

coefficient_data <- tidy(final_airbnb) %>%
  filter(term %in% desired_predictors)

coefficient_data
```

```
## # A tibble: 12 x 3
##    term                              estimate penalty
##    <chr>                                <dbl>   <dbl>
##  1 longitude                           -11.2     5.69
```

```
##  2 accommodates                          32.8     5.69
##  3 bathrooms                              8.95    5.69
##  4 bedrooms                              12.7     5.69
##  5 availability_30                        6.31    5.69
##  6 review_scores_rating                   0       5.69
##  7 reviews_per_month                     -1.05    5.69
##  8 room_type_Private.room               -25.5     5.69
##  9 room_type_Shared.room                 -7.05    5.69
## 10 neighbourhood_group_Manhattan         18.8     5.69
## 11 neighbourhood_group_Staten.Island     -1.32    5.69
## 12 has_TV_TRUE.                           4.31    5.69
```

Interpreting Coefficients

To understand what influences the pricing of Airbnb listings, the coefficients of our selected predictors provide valuable insights. Among these, several predictors stand out as the most influential factors shaping pricing dynamics. Notably, "Accommodates" emerges as a pivotal factor, with a substantial positive coefficient of 28.984. This indicates that a listing's capacity to host more guests directly translates to higher prices, emphasizing the significance of space. Additionally, "Bedrooms" and "Bathrooms" play notable roles, with coefficients of 15.602 and 10.877, respectively. Listings offering additional bedrooms and bathrooms tend to command higher prices, underscoring the importance of comfort and convenience. Moreover, the positive coefficient of 7.785 associated with "Availability in 30 days" signifies that listings with increased short-term availability are linked to higher prices, reflecting the dynamics of supply and demand in the Airbnb marketplace. Geographic location, represented by "Longitude," plays a notable role, with a negative coefficient of -12.245. Moving east or west within the city corresponds to a decrease in listing prices, suggesting that specific areas hold a premium.

Furthermore, the influence of categorical predictors cannot be overlooked. "Neighbourhood Group (Manhattan)" stands out with a substantial coefficient of 18.224, indicating that listings in Manhattan are associated with a premium. Additionally, amenities such as having a gym, a private room or shared room, or a doorman exhibit coefficients that affect pricing. "Room Type (Private Room)" exhibits a significant negative coefficient of -25.203, suggesting that offering a private room is associated with lower prices compared to other room types. Similarly, "Room Type (Shared Room)" has a negative coefficient of -8.347, indicating that listings offering shared rooms tend to be priced lower. While these predictors hold significant influence, it's worth noting that some predictors with non-significant coefficients still offer valuable insights. A positive coefficient for non-significant predictors implies that the variable positively influences pricing, albeit not significantly, while a negative coefficient suggests a negative but non-significant influence.

Model Evaluation

We evaluated our predictive model for Airbnb listing prices to understand its accuracy and the factors impacting listing prices.

Does the Model Produce Accurate Predictions?

To evaluate the accuracy of our model, we used the key metric Mean Absolute Error(MAE).

MAE: The MAE was calculated by comparing predicted prices to actual listing prices. The MAE score, with a value of 46.09506, reveals the extent of prediction accuracy. A lower MAE indicates that the model is more adept at producing accurate predictions.

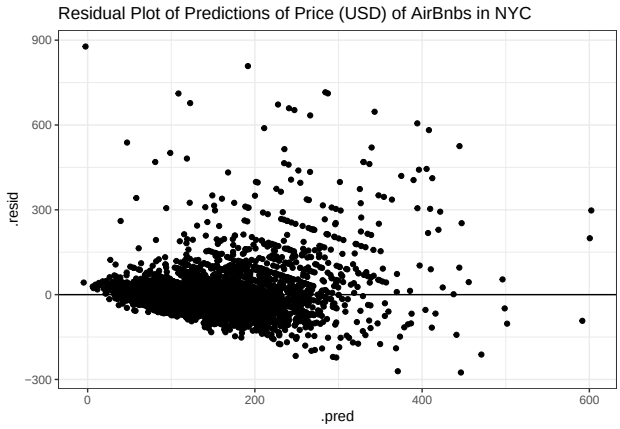
```r
# Does it produce accurate predictions? (MAE)
final_airbnb %>%
  augment(new_data = airbnb_sub) %>%
  mae(truth = price, estimate = .pred)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 mae     standard        46.0
```

Is the Models Residual Plot Indicative of a Good Fit?

We created and analyzed the residual plot to understand the model's data capture capabilities.

Residual Pattern: The residual plot analysis revealed that the residuals exhibit a clustered pattern around the line, and there was not really a balance of points above and below zero across the span of the model. This pattern suggests the presence of some systematic bias in the model's predictions rather than a random, evenly distributed spread.



Based on a comprehensive analysis of our model's Mean Absolute Error (MAE), and residual plots, it becomes evident that our model currently falls short in generating accurate predictions. In fact, it performs poorly, as indicated by the high MAE, which signifies a considerable gap between the predicted prices and the actual listing prices.

To respond effectively to our first research question, "How accurately can we predict a listing's price by its features?" we must acknowledge that our current model is not yet strong in this regard. Nevertheless, recognizing the need for model refinement and the inclusion of additional relevant features, along with addressing systematic biases, can pave the way for enhanced predictive performance and more accurate listing price predictions in the future.

Limitations

Despite the valuable insights garnered from our analysis of Airbnb listing prices based on property features in New York City, it is essential to recognize the inherent limitations to the scope and accuracy of our model. For the purpose of this study we used a relatively small sample size of 5,000 entries from the original dataset, which contained over 40,000 entries. This limited sample size might not fully capture the diversity of Airbnb

listings in New York City, potentially leading to a biased analysis. In terms of accuracy, the Lasso model used in the analysis had a Mean Absolute Error (MAE) of 46.1, meaning that our listing price predictions might be off by $46.1. Given the context of our data, this is a significant difference in price, which indicates that the model's predictive accuracy has room for improvement.

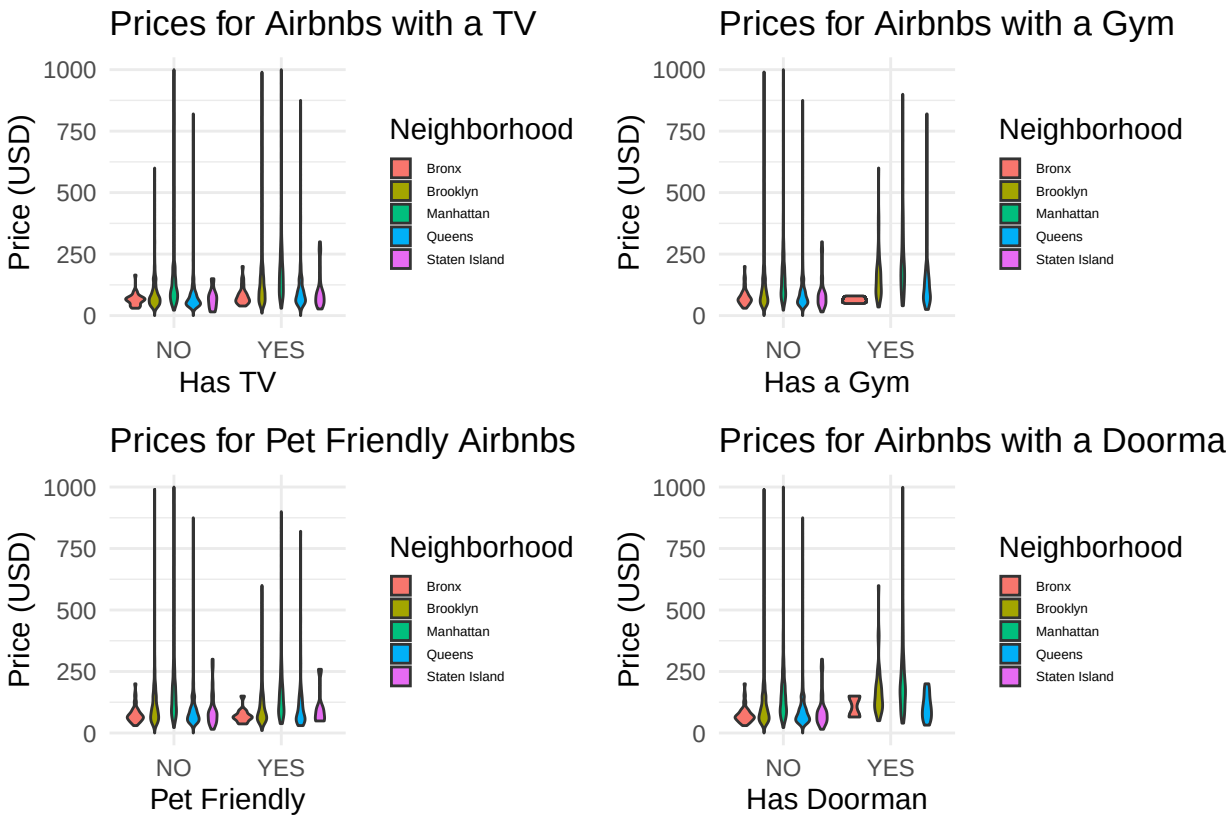
Further model refinement is needed to achieve more accurate price predictions. While this study considered various predictors such as accommodates, maximum_nights, and reviews_scores_rating, property_type, amenities, among others, there might be other relevant variables not included in the analysis that could significantly impact listing prices. On the other hand, we might have to consider a different model type to more accurately describe the relationship between the predictor variables and the listings prices. Our Lasso model, as a parametric model assumes linear relationships between our predictors and the price of the Airbnbs. However, the impact of these variables on price might be more complex and nonlinear, which our model won't be able to capture. It is also important to underscore the presence of systematic bias in our model's prediction evidenced by the clustered patterns in our residuals. Addressing this bias will be crucial in future research for improving predictive performance.

Lastly, based on the data that we are using to train our model, we are unable to identify price's temporal dynamics. Thus, this study does not consider potential changes in Airbnb listing prices over time. Prices may vary seasonally or due to other temporal factors, which are not accounted for in this analysis.

Key Takeaways

With respect to our first research question, this study demonstrates that it is possible to forecast Airbnb listing prices based on property features. While the model's initial accuracy can be improved, it provides a valuable starting point for understanding the price dynamics of Airbnb listings in New York City. We must acknowledge the need for model refinement and the inclusion of additional relevant features. Our model evaluation emphasizes the importance of addressing systematic bias in predictions to enhance accuracy. It also suggests that future research could delve deeper into Airbnb pricing dynamics, considering a broader range of variables and temporal factors to improve predictive accuracy. We also see a potential improvement in developing an nonparametric model that may better capture the relationship between predictor variables and price, and hold lower bias. In addition, if the resources are available, using a larger sample size to train our model could potentially increase prediction accuracy.

To respond to our question on the anatomy of price difference, we observed that several property features have been identified as influential factors that affect listing prices. These include the capacity to accommodate guests, the number of bedrooms and bathrooms, short-term availability, and geographic location. The location of an Airbnb property, as represented by the longitude, plays a significant role in pricing. Specific areas within New York City, such as Manhattan, command premium prices. Room types and the presence of amenities like gyms, private rooms, and doormen also have an impact on pricing although to different extents. In the violin plots below we can observe how different amenities may —or may not— impact a listing's price:



One of the main features we noticed is that different amenities have different impacts in price depending on which neighborhood they are in. For example, having a doorman in a Brooklyn or Manhattan listing will drive the prices up, however it won't have as much of an impact in the Bronx listings. Understanding the importance of these categorical predictors may be of interest for property owners and potential guests.