Airbnb Data Report: Predicting Price of New York City Listings

Tyler Davis Grey Gergen Jack Macfadyen Jesus Rodriguez

Motivation

There are many considerations that must be made when finding temporary housing, either for vacations or business trips. Airbnb is a popular website that allows individuals or businesses to list potential places for users to choose. These users can choose where they wish to stay based on location, accommodations, how many beds they need, and price. We sought to create a model that predicts the price of an Airbnb listing based on various variables. This will be useful to both users who wish to find an expected price of an accommodation based on various desired characteristics as well as hosts who want a robust method of finding appropriate pricing for their listing.

Airbnb data for major cities across the world has been scraped by a team of contributors and gathered on a website here in October 2025. Their motivation is to show transparency in how spaces are being rented to tourists in their communities. For the purposes of this report, we took a sample of all Airbnb listings from New York City.

When coming across this dataset, we found several missing values in the price of different Airbnb listings. Of the 36111 listings we found 14783 missing prices. We believe that we can use the latitude, longitude, beds, bathrooms, and neighborhood location among other variables to reliably impute the missing pricing data for these listings.

Table 1: Summary Statistics of `Price`

price
Min. : 10.0
1st Qu.: 88.0
Median : 152.0
Mean : 234.5
3rd Qu.: 272.0
Max. :10000.0
NA’s :14783

Data Documentation

Data Cleaning

Variable Overview

The process for choosing important variables in our model relied on a combination of determining which information is pertinent and relevant to actual usage by both users and hosts. The variable selection used for each variable is explained in the following sections.

`Price`

The `price` variable describes the daily price for a listing in its local currency. Since all listings in this analysis are from New York City, all prices are in $USD. There are 21,110 valid observations of price where it not missing and has a numerical value.

Summary Statistics

The `price` variable has a mean of $234.53 and a median of $152.

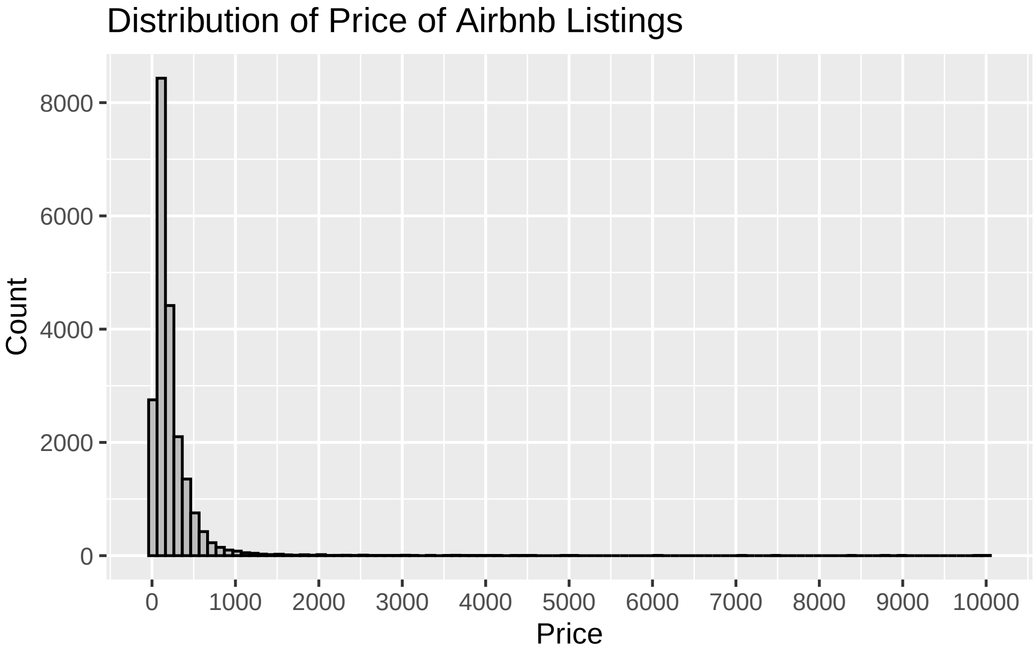
We decided to consider all listings at or above the price of $10,000 per night an outlier. There were a considerable group of listings above this number, going all the way to $50,000 per night. Many of the listings we examined manually had different, more realistic prices on the actual Airbnb website, indicating that either the listing was improper during the webscrape, or the webscrape itself had flaws. We chose $10,000 as the cutoff point as there was a listing at $9,999 per night that was verifiable. The smallest price in the dataset is $10 a night.

Table 2: Number of Listings in Each Price Range

price_range	n
<=1000	20755
>1000	355
N/A	14783

There are 14783 missing values in `price`, which means missing values account for 41.19% of the observations in the dataset.

Histogram and Price Ranges



A histogram of `price` shows that the distribution is very right-skewed, as most of the dataset has listings in the ≤ $1000 price range. In fact only, 0.99% of the dataset (including NA values) has Airbnb listings over the $1000 price range.

`Beds`

The `beds` variable describes the amount of beds an apartment/house/etc will have for guests. This differs from `bedrooms` as there could be more than one bed in a bedroom. There are 2.1425×10^4 observations in `beds` where the value is not missing.

Summary Statistics

Table 3: Summary Statistics of `Beds`

beds
Min. : 0.000
1st Qu.: 1.000
Median : 1.000
Mean : 1.634
3rd Qu.: 2.000
Max. :40.000
NA's :14468

The expected value of beds in a listing is 1.63 beds. Most listings in the dataset have 1 bed.

The highest number of beds a listing has in the data is 40, while the lowest number of beds is 0. **COMMENT: The value of 40 is very likely an outlier!**

Similarly to `price`, there are 14468 missing values in the `beds` variable, that is 40.31% of observations in the dataset.

This was supposed to be a ***`callout`*** *but it had problems rendering as a pdf* Taking a look at listings with 0 beds, and I found that there are some like this:

- www.airbnb.com/rooms/1111666966430724392
- https://www.airbnb.com/rooms/21456

where it is scraped as a 1 bedroom, 0 beds listings, however it does indeed have a bed.

When looking at listings with 0 beds AND 0 bedrooms:

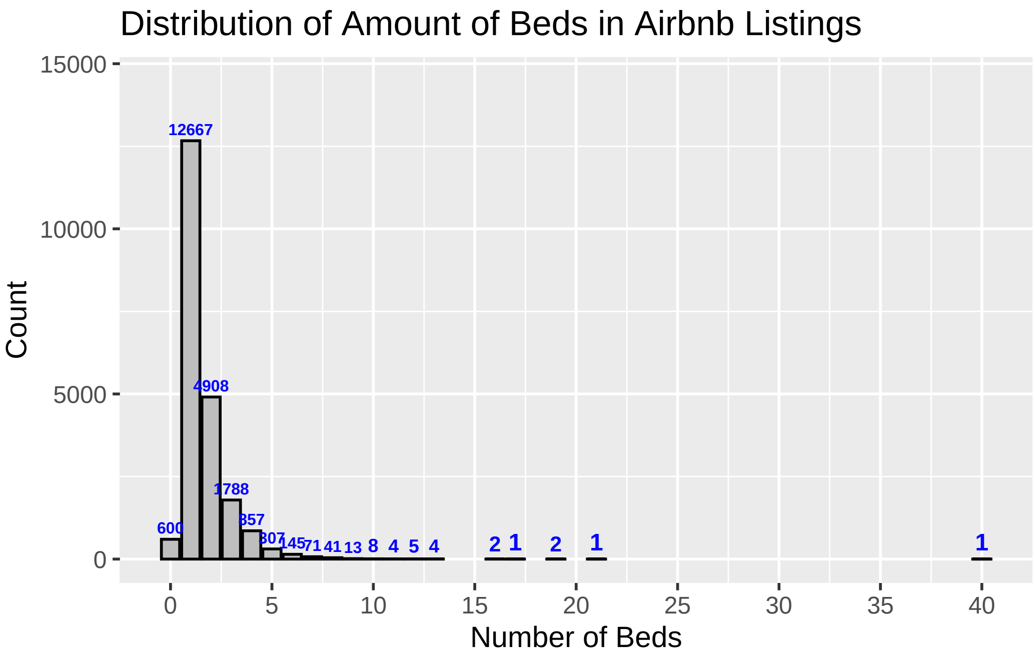
- https://www.airbnb.com/rooms/6359111
  - A loft meant for gatherings/parties… There doesn't seem to be any beds, but there are plenty of couches
- https://www.airbnb.com/rooms/1936633
  - A "Great 1BD waterfront City Island NY"
    - Literally has 1 bedroom in the name, and yes it does have a bed.

Vice versa (0 bedrooms, 1 bed) seems fine however, as there are mostly studio apartments like the following. https://www.airbnb.com/rooms/2595

Overall, the data is very messy, and the `source` of the scrape-`city scrape`-leaves a lot of listings that are very inaccurate to what is actually listed on the website.

End of callout, at this time, I think the model with imputation Tyler has done is good that we don't have to worry about these cases and innaccuracies, but it might be a good idea to include it in a discussion/limitations section

Histogram



COMMENT: I think these tables, especially the one the tail end of bed observations would be good to include. Maybe discuss how values above 20 should be considered outliers??

««««< HEAD

COMMENT 2: I took out the tables, I think the histogram now does the job, it still would be good to discuss outliers the same however. Feel free to delete the comments above if you don't want to use these two tables

Table 4: Summary Statistics of `Bedrooms`

bedrooms
Min. : 0.000
1st Qu.: 1.000
Median : 1.000
Mean : 1.385
3rd Qu.: 2.000
Max. :16.000
NA's :5949

`Bedrooms`

The `bedrooms` variable describes the amount of bedrooms an apartment/house/etc... will have for guests. This differs from `beds` as 1 bedroom could possibly contain 1 or more beds, or sometimes even no beds. There are 2.9944×10^4 observations in `bedrooms` where the value is not missing.

Summary Statistics

The expected value of bedrooms in a listing is 1.39 bedrooms. Most listings in the dataset have 1 bedroom.

The highest number of beds a listing has in the data is 16, while the lowest number of beds is 0.

There are 5949 missing values in the `bedrooms` variable, that is 16.57% of observations in the dataset.

`Bathrooms`

The `bathrooms` variable describes the amount of bedrooms an apartment/house/etc... will have for guests. This differs from `beds` as 1 bedroom could possibly contain 1 or more beds, or sometimes even no beds. There are 3.5849×10^4 observations in `bedrooms` where the value is not missing.

Table 5: Summary Statistics of `Bathrooms`

bedrooms
Min. : 0.000
1st Qu.: 1.000
Median : 1.000
Mean : 1.385
3rd Qu.: 2.000
Max. :16.000
NA's :5949

Summary Statistics

The mean amount of bathrooms in a listing is 1.17 bathrooms. Most listings in the dataset have 1 bathroom.

The highest number of beds a listing has in the data is 15.5, while the lowest number of beds is 0.

There are 44 missing values in the `bedrooms` variable, that is 0.12% of observations in the dataset.

COMMENT 2: I took out the tables, I think the histogram now does the job, it still would be good to discuss outliers the same however. »»»> 5c96e31971cf0db4e62f00571f595d400030916a

Variable Selection and Discussion

The variables of interest for modeling and analysis were selected based on their expected influence on `price`, the outcome variable. Only measurable characteristics of the list or host were included, rather than variables related to the web-scrape process itself (ie., `last_scraped`, `source`) or variables such as urls or images.

The variables were grouped into categories according to the aspect of the listing they represent:

- **Location**
    - `neighbourhood_group_cleansed`
    - `neighbourhood_cleansed`

  - `latitude`
  - `longitude`

- **Property characteristics**
  - `property_type`
  - `room_type`
  - `accommodates`
  - `bathrooms`
  - `bedrooms`
  - `beds`

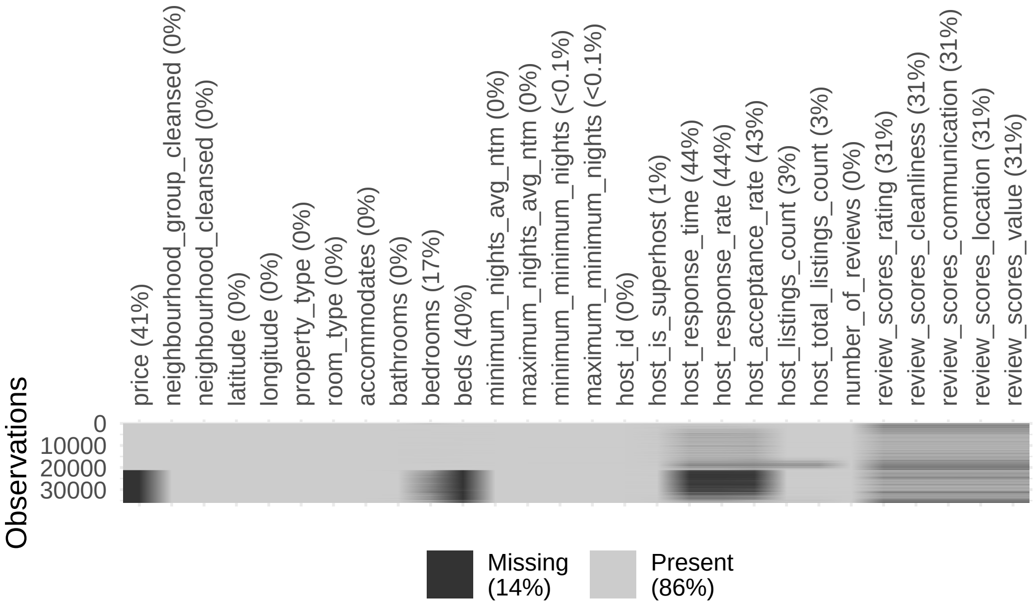
- **Booking rule summaries (night requirements)**
  - `minimum_nights`
  - `maximum_nights`
  - `minimum_nights_avg_ntm`
  - `maximum_nights_avg_ntm`
  - `minimum_minimum_nights`
  - `maximum_minimum_nights`

- **Host characteristics**
  - `host_id`
  - `host_is_superhost`
  - `host_response_time`
  - `host_response_rate`
  - `host_acceptance_rate`
  - `host_listings_count`
  - `host_total_listings_count`
  - `host_identity_verified`

- **Reviews**
  - `number_of_reviews`
  - `review_scores_rating`
  - `review_scores_cleanliness`
  - `review_scores_communication`
  - `review_scores_location`
  - `review_scores_value`

**START COMMENT: I'm not sure if it is my browser, but the vis_miss() plot seems blurry when it is rendered

To visually understand the completeness of the data, missingness patterns were examined using a `vis_miss()` plot. The plot was arranged by increasing price.



From this plot, missingness in `price` appeared related to variables such as `beds`, `host_response_time`, `host_response_rate`, and perhaps `host_acceptance_rate`. Another pattern can be seen with increasing `price` and all the variables related to reviews.

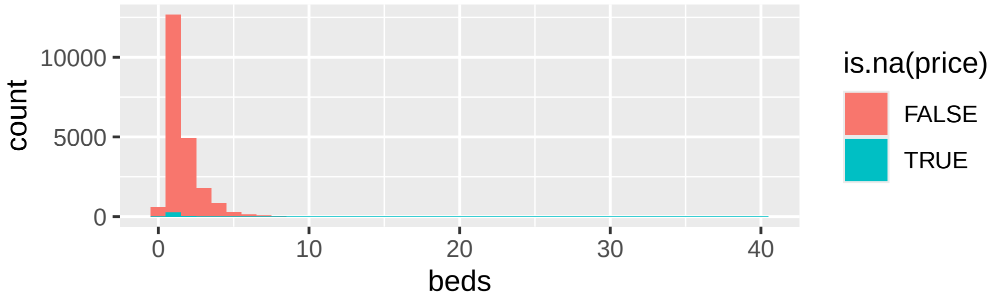
**END COMMENT: vis_dat() doesn't work either

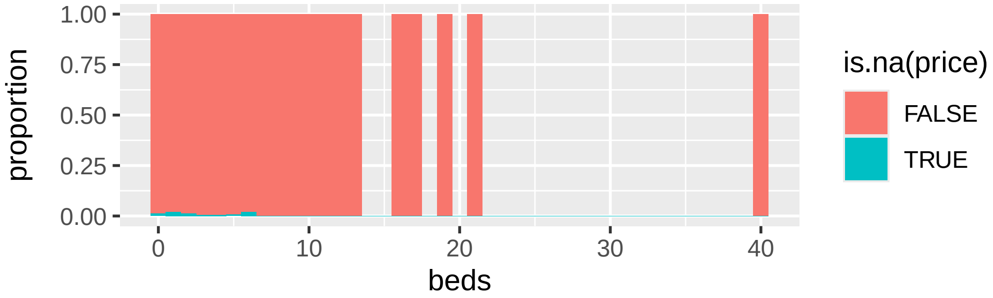
Prioritzing interpretability, only variables with simple measures were considered for the final model. These predictors were: `beds`, `bedrooms`, `bathrooms`, `accomodates`, `latitude`, `longitude`, and `host_is_superhost`.

Selected Predictors Relation to Price

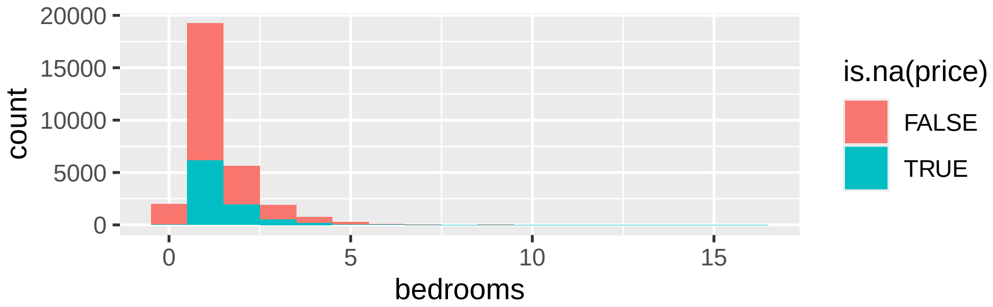
How do the predictors `beds`, `bedrooms`, `bathrooms`, `latitude`, `longitude`, and `host_is_superhost` relate to the outcome variable `price`?

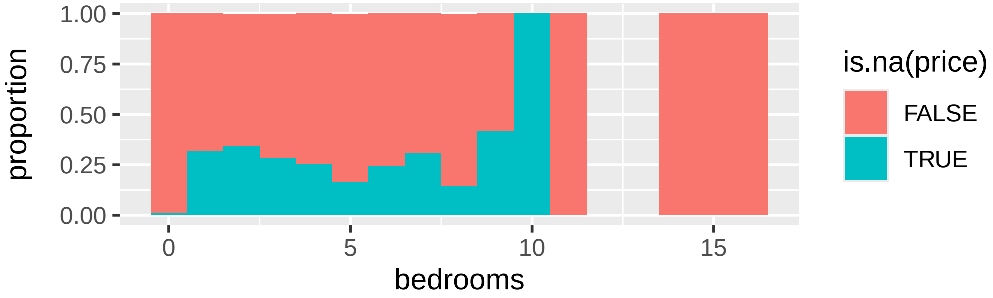
Beds



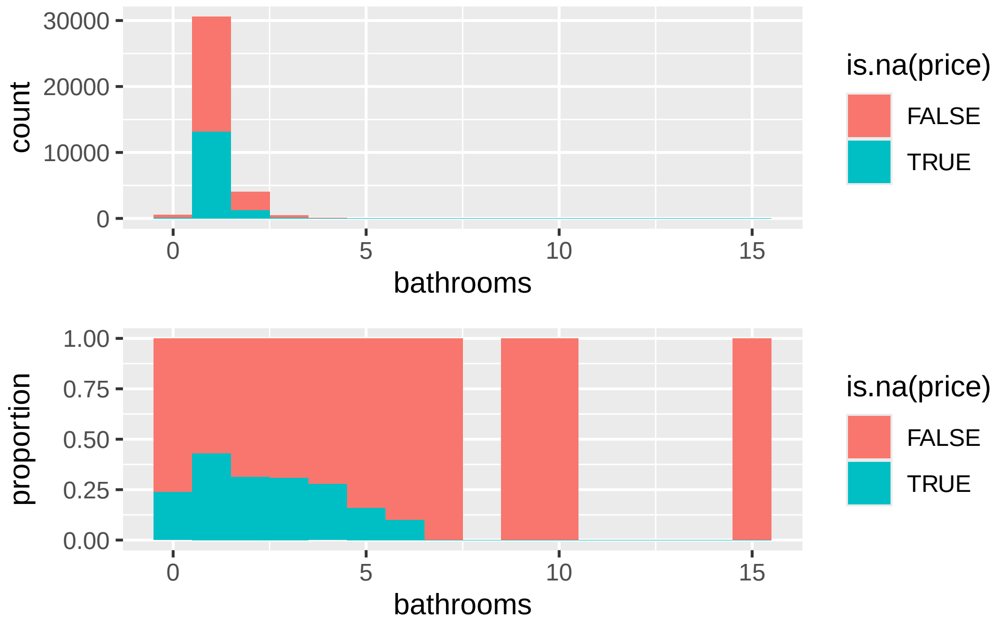


Bedrooms

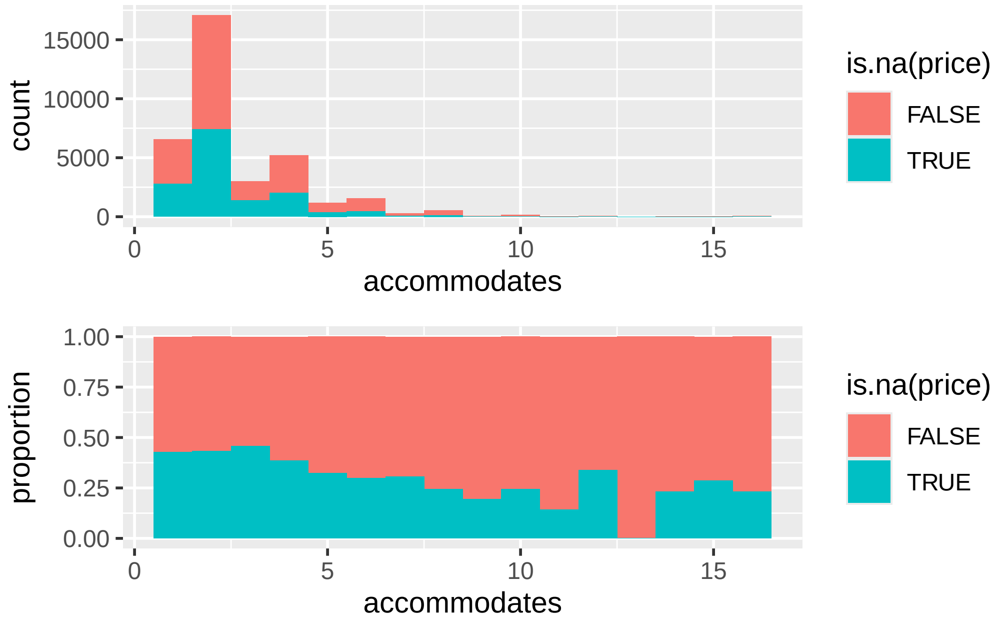




Bathrooms



Accomodates



Model Construction

The next step in the analysis will be to build a model with price as the response variable. The predictor variables will be values that are suspected to be indicators of price such as beds, geographic coordinates or UTM, accommodates, bathrooms, etc. This model will then be run and model accuracy will be analyzed. The price variable will be imputed first using geographic coordinates in the predictor set then UTM. The final model was developed from a variety of model building techniques, along with understanding our predictors of interest. This model will create a basis for comparing imputation methods with different coordinate points.

Statistical Model

$$\begin{aligned}\log(\text{Price}) &= \beta_0 + \beta_1 \cdot \text{bedrooms} + \beta_2 \cdot \text{beds} + \beta_3 \cdot \text{latitude} && (1)\\ &+ \beta_4 \cdot \text{longitude} + \beta_5 \cdot (\text{latitude} \times \text{longitude}) && (2)\\ &+ \beta_6 \cdot \text{accommodates} + \beta_7 \cdot \text{bedrooms} && (3)\end{aligned}$$

Imputation Analysis

The overall goal of this study is to test the imputive quality of different distance based metrics on the price variable from the AirBnb data set. This analysis will then provide information regrading the potential use of UTM more consistently over geographic coordinates. The response variable of our model, price, is the variable that is going to be imputed. The model accuracy of both the imputation with geographic coordinates and UTM will be compared using AIC and BIC. The RMSE of each imputation method will also be calculated and the difference will be statistically tested. The imputation analysis throughout this study will be using the mice package in R. The mice package contains a variety of methods for imputation, for this analysis predictive mean matching (pmm) will be used. The price variable and the entire predictor set are all quantitative variables for which pmm is the recommended method of imputation.

Missing Values of Price

We know there are 14783 missing values in `price`. There are 355 occurrences when `price` is the only missing value. The `price` variable is missing the most when `beds` is also missing. The second most time `price` is missing is when `beds` AND `bedrooms` are missing. Note that, `latitude`, `longitude`, and `accomodates` are the only variables of interest that do not have any missing values. The variable of most interest to us is the price variable and how imputation predictor sets impact potential results.

Comparing Model Accuracy Between Predictor Sets

Table 6: Comparing Model Fit Between Geographic vs. UTM

	Mean AIC	Mean BIC
Geographic	70677.62	70754.02
UTM	70911.21	70987.61

The imputation was run on the price variable using both geographic coordinates and UTM. The table above shows the mean AIC and BIC measurements for each imputation predictor set. The AIC and BIC measurements are lower for the geographic coordinates suggesting that they represent the relationship better than UTM. The model comparison gives the idea that geographic coordinates could be be a better fit for the model when compared to UTM.

Comparing Imputation Methods with RMSE

The predictor set containing geographic coordinates seems to be a slightly better fit in our model. Yet, AIC and BIC do not have anything to say about imputation quality.This difference can truly be tested by comparing RMSE for each imputation method. This will be done by masking 10 percent of the price values that are not missing, then imputing using each predictor set and calculating the RMSE. This process will then be simulated 10 times and the mean RMSE value will be calculated and compared.

Table 7: Impuation Analysis using RMSE

	Mean RMSE	SE	CI (95%)
UTM	456.015	17.995	(420.75 - 491.28)
Geographic	457.920	17.204	(424.2 - 491.64)

It can be seen in the table above the the predictor set containing UTM has the slightly lower RMSE signifying that it could have higher imputation accuracy. The predictor set with the geographic coordinates does have the lower SE which indicates slightly less variability. These results do not give any significant conclusions, and will need to be statistically tested. This can be done with an anova test, which will show if the the two predictors sets mean RMSE values are different from each other.

Anova Testing

The anova test above shows that we fail to reject the null hypothesis, (p - value > .05), which is that the predictor sets mean RMSE values are equal. This shows that there is no

Table 8: Anova Testing RMSE vs. Predictor Sets

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
predictors	1	18.148	18.148	0.006	0.94
Residuals	18	55781.197	3098.955	NA	NA

statistical difference in using the predictor set with geographic coordinates or UTM in regards to imputation quality. It can be said that neither the imputation or the model would directly benefit from using UTM over geographic coordinates.

Conclusion

Overall, the study that has been completed shows that geographic coordinates and UTM do not differ in imputation quality in the AirBnb data set for New York City. This conclusion is supported the the fact the RMSE values were shown to be not significantly different. The result from this study can be used in further analysis of this data to further build a more complex model to model price and test its predictive quality. This conclusion does not support the earlier hypothesis that UTM is a better distance based measure for regression and imputation, due to its more linear qualities when compared to geographic coordinates. This relationship could be impacted by the size of the analysis area (New York City), and might be impacted if a larger or smaller area was considered. This study leaves the door open for analyzing the imputive quality of more linear distance based measures such as UTM.