AirBnB Data Report

Team Information

Team name: Team Team

Members: Grey Gergen, Tyler Davis, Jesus Rodriguez, Jack Macfadyen

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.

Airbnb data has been scraped by a team of contributors and gathered on a website here. Their motivation is to show transparency in how spaces are being rented to tourists in their communities.

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 impute the missing pricing data for these listings.

Data Documentation

Data Cleaning

Varaible Overview

The variables selected for analysis were chosen via variable selection explained further in this report.

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 observations of price where it has a numerical value and it is not missing from the data.

Summary Statistics

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

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

Due to cleaning, the max amount of price of limited to $10000 a night. The smallest price in the dataset is $10 a night.

There are 14783 missing values in `price`. Which means missing values make up around 41.19% of the observations in the dataset.
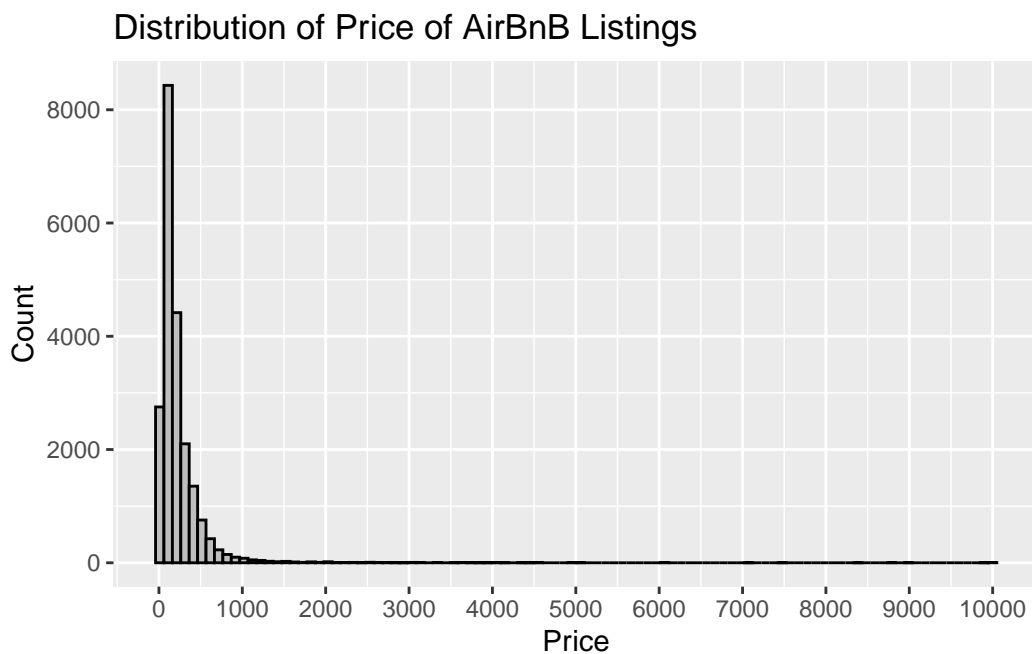
Histogram and Price Ranges

Distribution of Price of AirBnB Listings

Table 2: Number of Listings in Each Price Range

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

A histogram of `price` shows that the distribution is very right-skewed, as most of the dataset has listings in the $\leq$ \$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 listing, however it does indeed have a bed.
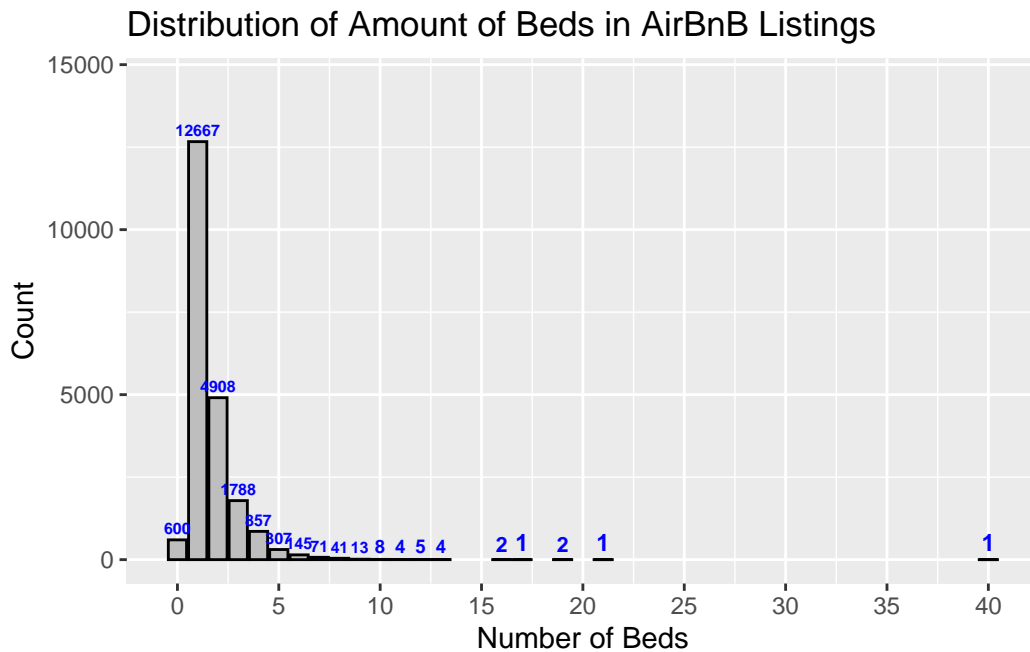
When looking at listings with 0 beds AND 0 bedrooms:

- 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

Distribution of Amount of Beds in AirBnB Listings

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??

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

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

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

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.

Summary Statistics

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

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.

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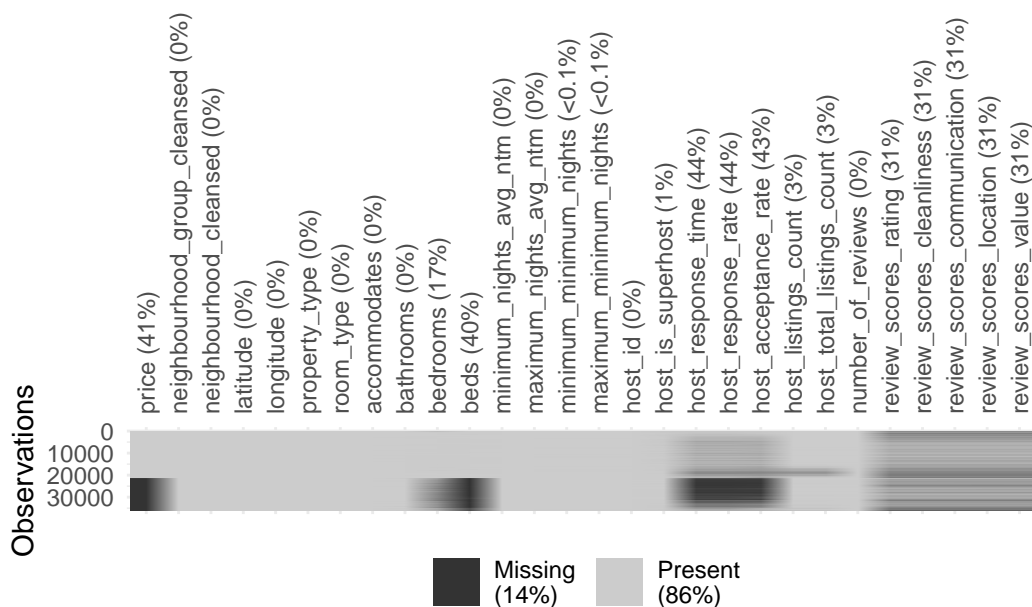
The variables were grouped into categories according to the aspect of the listing they represent:

- **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`

**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`.
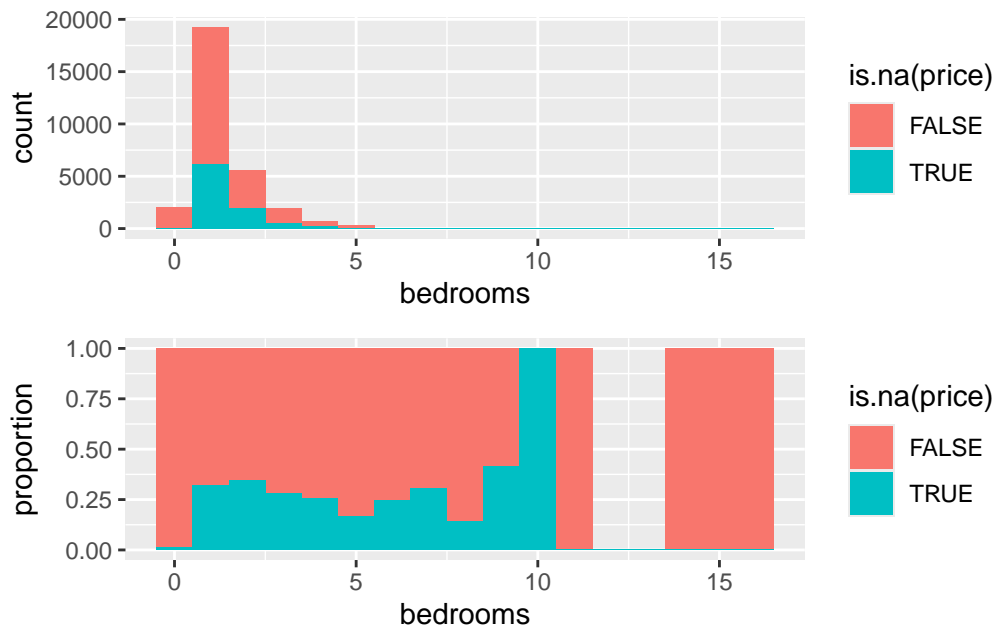
Selected Predictors Relation to Price

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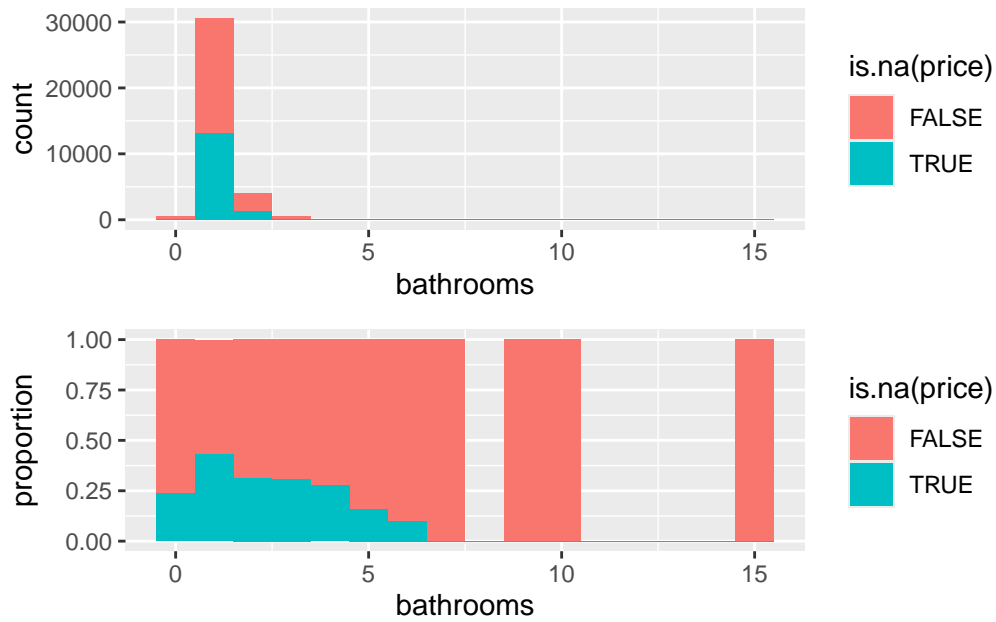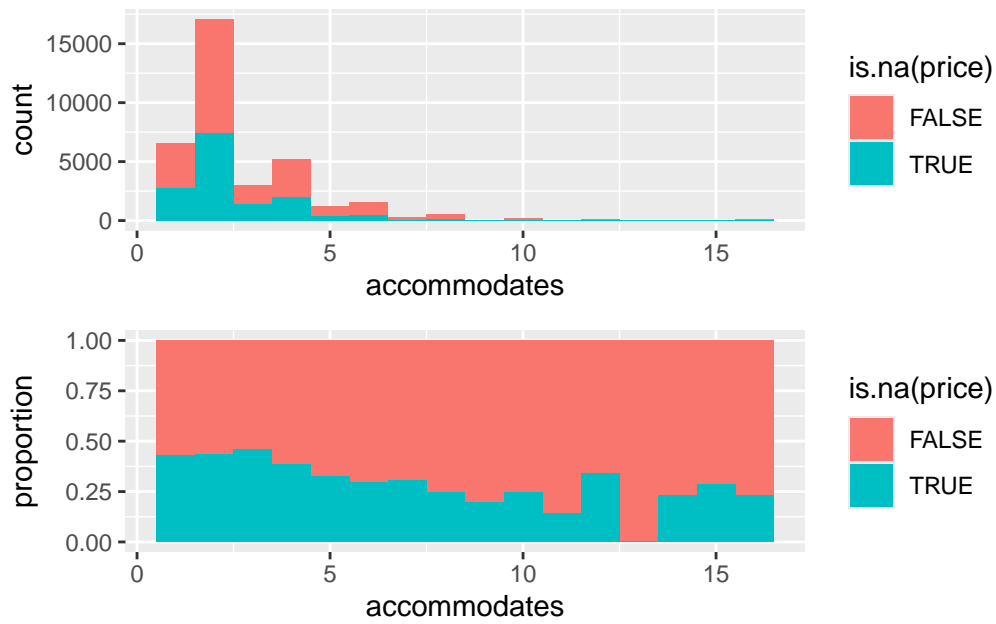
Beds

Bedrooms

Bathrooms

Accomodates

11

Model Construction

The next step in the analysis will be to build a model with price as the response variable. The predictors will be values that are suspected to be indicators of price such as beds, lat and long, accommodates, bathrooms, etc. This model will then be run and coefficients values and significance will be documented. This will all be done without imputation.

Matrix Plot to Assess Variable Correlation

Model Building

The airbnb model that will be used for this study will be attempting to predict the price variable. The predictors are beds, bathrooms, latitude, longitude, accommodates, and bedrooms. 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.

$$\log(\text{Price}) = \beta_0 + \beta_1 \cdot \text{bedrooms} + \beta_2 \cdot \text{beds} + \beta_3 \cdot \text{latitude} + \beta_4 \cdot \text{longitude} + \beta_5 \cdot \text{latitude*longitude} + \beta_6 \cdot \text{accomodates} + \beta_7 \cdot \text{b}$$
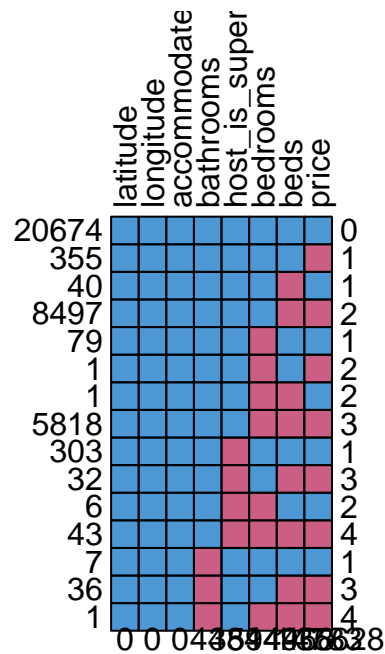
Coefficeint Estimates for Airbnb Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.300198e+05	3167.1630953	41.052448	0
bedrooms	-8.017570e-02	0.0057295	-13.993592	0
beds	4.220580e-02	0.0054763	7.706955	0
latitude	-3.202096e+03	77.8232893	-41.145733	0
longitude	1.759253e+03	42.8374615	41.068097	0
accommodates	1.993121e-01	0.0031530	63.214307	0
bathrooms	5.594000e-02	0.0090677	6.169126	0
latitude:longitude	-4.332788e+01	1.0525988	-41.162769	0

The coefficient estimate table above shows that all variables are significant predictors of price. This will then allow us to test the differences in imputation when using geographic coordinates versus Universal Transverse Mercator (UTM).

The overall goal of this study is to test the imputive quality of different distance based metrics on the price variable from the airbnb dataset. This analysis will then provide information regrading the potential use of UTM more consistently over geographic coordinates.

Geographic coordinates are based on a spherical Earth model and uses degrees to measure distance. This can create issues with interpretation as degrees are not uniform in distance and can add necessary complications to a model. Geographic coordinates are usually measured in latitude and longitude. UTM is a projected coordinate system which is based on 60 zones across the Earth. These coordinates are more easily interpreted on a linear scale. UTM is usually measure on a x and y scale. The hypothesis on this study is that if UTM is in the predictor set the imputation will have more accuracy than if geographic coordinates are in the predictor set. The first step is to analyze the missing value patterns of our price variable.

Missing Values of Price

	latitude	longitude	accommodates	bathrooms	host_is_superhost	bedrooms
20674	1	1	1	1	1	1
355	1	1	1	1	1	1
40	1	1	1	1	1	1
8497	1	1	1	1	1	1
79	1	1	1	1	1	0
1	1	1	1	1	1	0
1	1	1	1	1	1	0
5818	1	1	1	1	1	0
303	1	1	1	1	0	1
32	1	1	1	1	0	1

6	1	1	1	1	0	0
43	1	1	1	1	0	0
7	1	1	1	0	1	1
36	1	1	1	0	1	1
1	1	1	1	0	1	0
	0	0	0	44	384	5949

	beds	price	
20674	1	1	0
355	1	0	1
40	0	1	1
8497	0	0	2
79	1	1	1
1	1	0	2
1	0	1	2
5818	0	0	3
303	1	1	1
32	0	0	3
6	1	1	2
43	0	0	4
7	1	1	1
36	0	0	3
1	0	0	4
	14468	14783	35628

We know there are 14783 missing values in `price`. There are 355 occurrences when `price` is the only missing value. It looks like `price` 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.

Imputing Missing Values with Geographic Coordinates

The imputation process 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.

The table above shows the estimates for the airbnb model with the price variable imputed. The coefficient estimates for the model are slightly different from the original model showing that imputation does have an impact in this data set and model. The next step of the analysis to replace geographic coordinates (latitude and longitude) with UTM based distance measurements.

Imputing Missing Values with Universal Transverse Mercator

term	estimate	std.error	statistic	df	p.value
(Intercept)	4.4337415	0.0102707	431.690343	84.276261	0.0000000
bedrooms	-0.0664399	0.0064109	-10.363582	23.811081	0.0000000
bathrooms	0.1292279	0.0083531	15.470649	354.923050	0.0000000
beds	0.0137633	0.0072911	1.887691	10.916969	0.0859201
accommodates	0.1977896	0.0039064	50.632007	14.007389	0.0000000
X	-0.2750749	0.0044886	-61.282494	40.484154	0.0000000
Y	0.0993012	0.0068342	14.530062	7.307948	0.0000012
X:Y	-0.1112607	0.0038699	-28.750012	23.889664	0.0000000

The coefficient estimates above show the imputed model for price with UTM distance variables. This model shows a drastic difference from the previous two models. This shows that not only imputation but which distance based metric used impacts coefficient estimates. The only way to truly see model performance of each model is to compare AIC and BIC measurements for each of the models constructed.

Comparing Model Accuracy Between Predictor Sets

	mean_AIC	mean_BIC
Geographic	70677.62	70754.02
UTM	70911.21	70987.61

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 give the idea that geographic coordinates could be be a better fit in the model when compared to UTM. This difference can truly be tested by comparing the imputation quality of each method. This will be done by calculating the RMSE for each imputation method.

Comparing Imputation Methods with RMSE

	Predictor_Set	Mean_RMSE	SD_RMSE	Min_RMSE	Max_RMSE	SE	CI_lower	CI_upper
UTM	UTM	456.0150	56.90380	356.7984	547.1487	17.99456	420.7457	491.2844
Geographic	Geographic	457.9202	54.40467	361.9025	543.0629	17.20427	424.1998	491.6405

It can be seen in the table above the the predictor set containing UTM as the slightly lower RMSE signifying that it could have higher imputation accuracy. The predictor set with the geographic coordinates does have the lower SE. These results do not give any conclusions. The results in the table above will need to be statistically tested. This can be done with in a anova model, with RMSE as the response variable and the "predictor sets" (UTM vs. Geographic) as the predictor variable.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
predictors	1	18.14782	18.14782	0.0058561	0.9398454
Residuals	18	55781.19654	3098.95536	NA	NA

The anova summary above shows that we fail to reject the null hypothesis that the group means are equal. This means that there is no statistical difference in using the predictor set with geographic coordinates or UTM in regards to imputation quality. This shows that overall for the airbnb data set neither the imputation or the model would directly benefit from using UTM over geographic coordinates.