

# Comparing Imputation Quality of Global Coordinate System and Universal Transverse Mercator for Statistical Modeling

Tyler Davis      Grey Gergen      Jack Macfadyen      Jesus Rodriguez

## Motivation

When it comes to creating a model that includes location or coordinate data, there are various methods one can use for implementation. The Geographic Coordinate System (GCS) is a non-cartesian, global standard that communicates position using latitude and longitude. The Universal Transverse Mercator (UTM) is a projectile-based mapping of Earth that creates 60 zones. These coordinates are more easily interpreted on a linear scale, which allows for better quantification of distance. We wish to see whether this different mapping of coordinate data has an effect on imputation and creating a model. By conducting imputation using the commonly-used latitude and longitude and then again with UTM, we can compare and contrast the different coordinate systems to examine the potential use of UTM over GCS for imputation before modeling. For this study, we will be examining web-scraped Airbnb listings from New York City and create a model predicting the price of a listing based on location and various other metrics.

## Data Description

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 to show how spaces are being rented to non-locals. We took a sample of all Airbnb listings from New York City to use for our models. Data documentation can be found in @appendix-A

When examining the dataset, there were a large proportion of missing values for the price of Airbnb listings. Of the 36111 listings in the data, 14783 (41%) had missing price values, giving us an opportunity to use various variables, including location with our two coordinate systems, to reliably impute the missing pricing data for these listings.

## Variable Overview

Of the various variables that were included in the dataset, there were a few that we considered strong, interpretable predictors of price for our model: the number of bedrooms, beds, and bathrooms; the total amount of people who could be accommodated; and the location (latitude and longitude). A list of all variables in the dataset, their selection process, and missingness patterns can be found in @appendix-B.

## Model Construction

The next step in the analysis will be to build a model with price as the response variable as well as the predictor variables we assume to be strong indicators. The price variable in the data will be imputed first using geographic coordinates and then once more with UTM as a separate dataset. The model will then be run with each dataset and model accuracy will be analyzed. 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

$$\log(\text{price}) = \beta_0 + \beta_1(\text{bedrooms}) + \beta_2(\text{beds}) + \beta_3(\text{latitude}) + \beta_4(\text{longitude}) + \beta_5(\text{latitude} \times \text{longitude}) + \beta_6(\text{accommodates})$$

## Imputation Analysis

Of the 14783 missing price values, there are only 355 occurrences when price is the only missing variable. Mostly commonly, when price is missing, beds is also missing. There were no observations where latitude, longitude, or accommodates were missing.

The response variable of our model, price, is the variable that is going to be imputed using predictive mean matching from the `mice` package in R, first with geographic coordinates and then again with UTM. The model accuracy of both the imputation with geographic coordinates and UTM will be compared using AIC and BIC. The Root Mean Square Error (RMSE) of each imputation method will also be calculated, and the difference will be statistically tested.

## Comparing Model Accuracy Between Predictor Sets

@tbl\_model\_comp shows the mean AIC and BIC measurements for each imputation predictor set. For both metrics, the model with GCS-imputed data has a lower value, suggesting that the model is a better fit and represents the relationship to price better than the dataset imputed with UTM.

Table 1: Comparing Model Fit Between GCS vs UTM

	Mean AIC	Mean BIC
GCS	70677.62	70754.02
UTM	70911.21	70987.61

## Comparing Imputation Methods with RMSE

The predictor set containing geographic coordinates seems to be a slightly better fit with our model. Yet, AIC and BIC do not have anything to say about imputation *quality*. This difference can 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 values will be calculated and compared.

Table 2: Impuation Analysis using RMSE

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

@tbl\_rmse\_table shows the UTM predictor set 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 ANOVA, which will show if the mean RMSE of the two predictor sets are different from each other.

Table 3: 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

The ANOVA shows that we fail to reject the null hypothesis, ( $p\text{-value} = 0.94$ ), which is that the predictor sets mean RMSE values are equal (@tbl\_rmse\_comparison). This shows that there is no significant 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, this study shows that geographic coordinates and UTM do not differ in imputation quality in the case of Airbnb data for New York City. This conclusion is supported by the fact the RMSE values were shown to be not significantly different. As a result, we can't claim that switching to UTM as a distance-based coordinate system leads to better imputation for modeling than GCS, an angular coordinate system. This outcome could have been influenced by the size of the analysis area (New York City) and have different results if a differently-sized area was considered. The procedure from this study can potentially be used for analyzing the imputation quality of other coordinate systems, as well as further analysis of this data to build a more complex model and test its predictive quality.

## Appendix A: Data Documentation

### Data Documentation for Model

Field	Type	Description
host_is_superhost	boolean [t=true; f=false]	NA
latitude	numeric	Uses the World Geodetic System (WGS84) projection for lat
longitude	numeric	Uses the World Geodetic System (WGS84) projection for lat
accommodates	integer	The maximum capacity of the listing
bathrooms	numeric	The number of bathrooms in the listing
bedrooms	integer	The number of bedrooms
beds	integer	The number of bed(s)
price	currency	daily price in local currency. NOTE: the \$ sign is a technical

### 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:

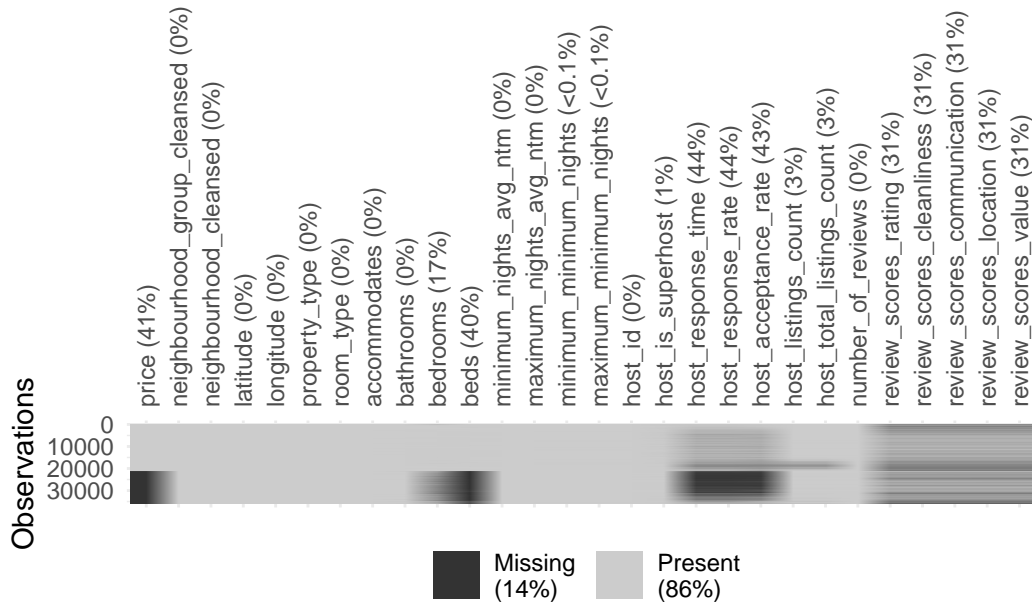
- **Location**
  - neighbourhood\_group\_cleansed, neighbourhood\_cleansed, latitude, longitude
- **Property characteristics**
  - property\_type, room\_type, accommodates, bathrooms, bedrooms, beds
- **Booking rule summaries (night requirements)**
  - minimum\_nights, maximum\_nights, minimum\_nights\_avg\_ntm, 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

```

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.

## Appendix B: Variables and Variable Selection

### 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.

Table 5: 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.

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.

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



Table 6: 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, with less than 1% of listing being over \$1000 per night.

## 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  $2.1425 \times 10^4$  observations in `beds` where the value is not missing.

## Summary Statistics

Table 7: 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](http://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:

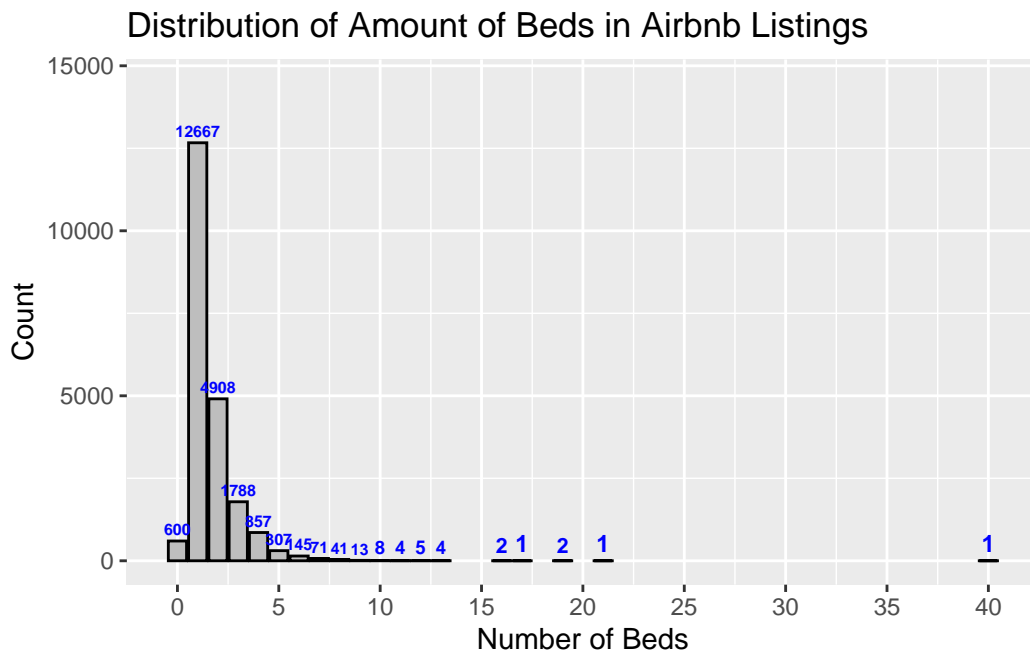
- <https://www.airbnb.com/rooms/6359111>
  - A loft meant for gatherings/parties... There doesn't seem to be any beds, but there are plenty of couches
- <https://www.airbnb.com/rooms/1936633>
  - A “Great 1BD waterfront City Island NY”
    - \* Literally has 1 bedroom in the name, and yes it does have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

## 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  $2.9944 \times 10^4$  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.

Table 8: 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

Table 9: 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 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  $3.5849 \times 10^4$  observations in **bedrooms** where the value is not missing.

## 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.

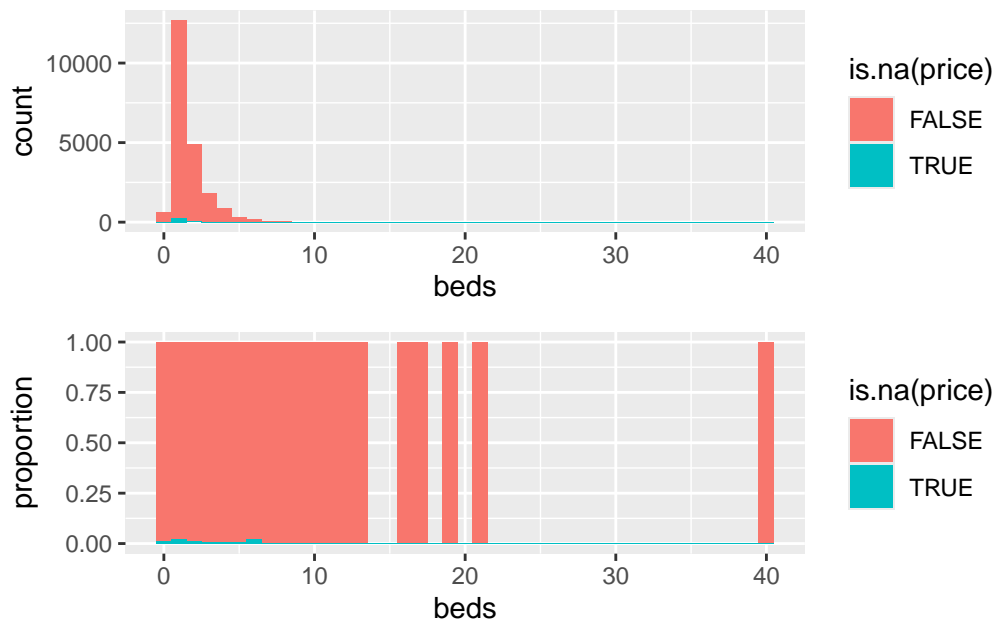
There are 44 missing values in the `bedrooms` variable, that is 0.12% of observations in the dataset.

## Appendix B

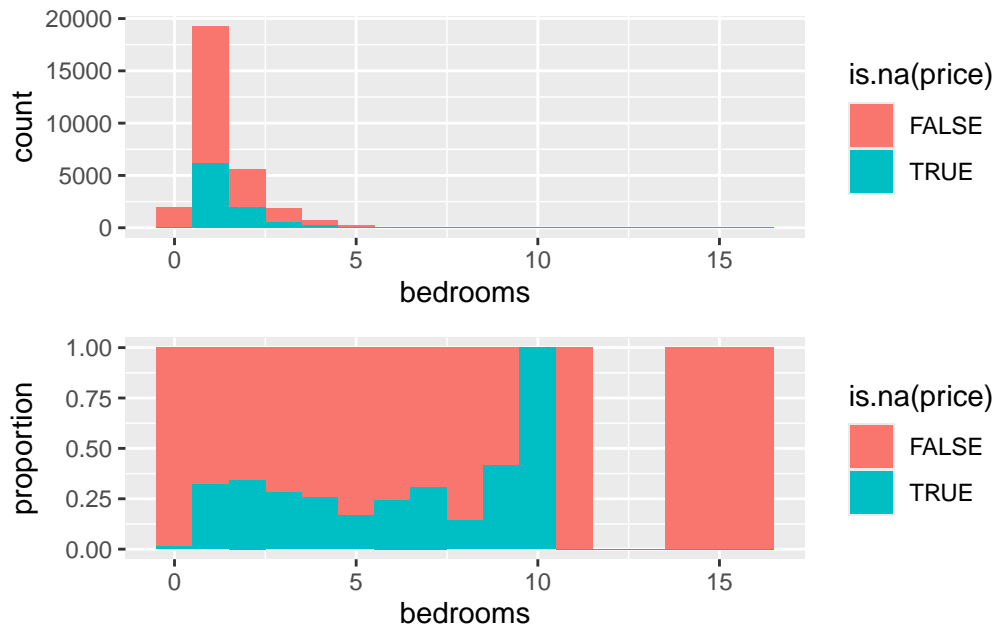
### Selected Predictors Relation to Price

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

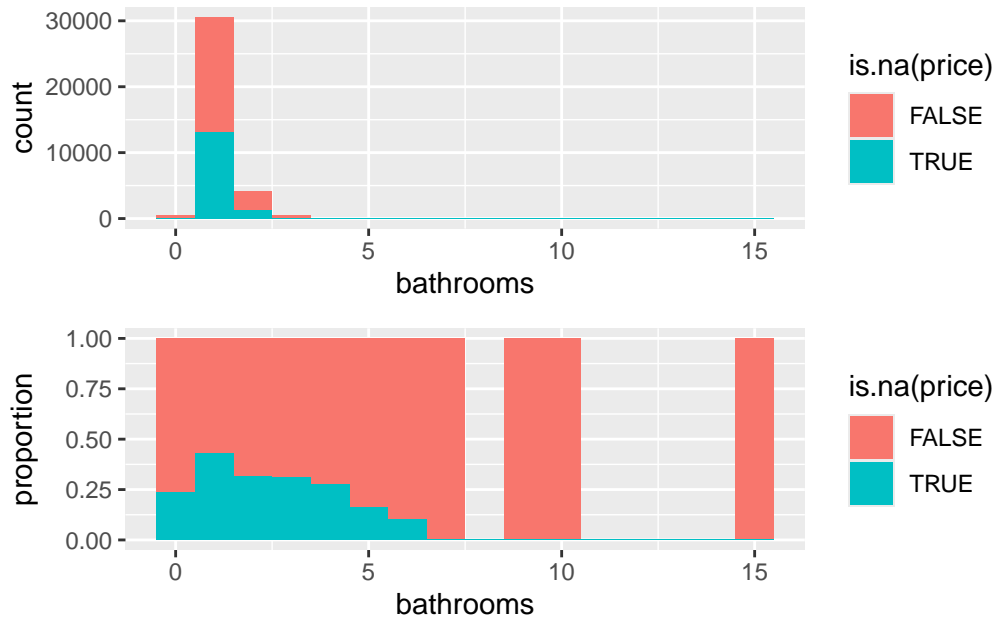
#### Beds



### Bedrooms



### Bathrooms



## Accommodates

