Spring 5-8-2022

Comparative Analysis of Imputation Methods in Real Estate Data

Connor Donlen

A COMPARATIVE ANALYSIS OF IMPUTATION METHODS IN REAL ESTATE DATA

by
Connor Jacob Donlen

A thesis submitted to the faculty of The University of Mississippi in partial fulfillment of the requirements of the Sally McDonnell Barksdale Honors College.

Oxford
May 2022

Approved by

Advisor: Dawn E. Wilkins

Reader: Yixin Chen

Reader: Philip Rhodes

Abstract

Comparative Analysis of Imputation Methods in Real Estate Data

by

Connor Donlen

University of Mississippi, 2022

This project involves comparing different methods of missing data imputation in the context of predicting real estate listing prices. These methods are compared against each other in both their ability to recreate the original data and their effects on a final predictive model. In order to evaluate their effectiveness, first, a predictive model is made using the complete dataset to use as a benchmark for the imputed datasets. Then, a complete dataset is split into 80% training and 20% testing datasets, and missing values are created in the training data using two different missing data mechanisms, missing completely at random (MCAR) and missing at random (MAR). These datasets are then imputed using several popular imputation methods and used as training data for the same model architecture as the benchmark.

The final predictive models show that multiple imputation using deterministic regression gives the best results for MCAR data, and multiple imputation using stochastic regression gives the best results for MAR data.

Since MAR data is encountered more frequently, this reaffirms the viewpoint that proper imputation requires more than just predicting the missing values as accurately as possible, and an analyst should also be concerned with preserving the variability of the data. However, the results were similar enough in some trials that, in some instances, using multiple imputation and stochastic methods over single imputation and deterministic methods may be a matter of best practice rather than one that gives definitive improved results.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Overview

The primary goal of this project was to find the imputation method that was able to minimize the error in predicting the price of real estate property listings in the presence of missing data. The potential real-life application considered for certain design choices was finding property listings that were undervalued and thus good for investing or flipping for profit. Because of this, the primary metric for both model building and imputation evaluation was the root mean square error when predicting the test dataset.

The secondary goal of the project was to learn more about imputation as a whole and the potential advantages and disadvantages of different imputation methods. Although the results of the research are only for real estate data and, more specifically, this dataset, this research will help to make more informed decisions on handling missing data.

1.1 Predictive Models

In machine learning, predictive models are built on one or more predictor variables with the goal of achieving the highest accuracy in predicting a target variable. Models are usually built on training data and evaluated on test data that was not seen during training so that an analyst can be more confident that a model will perform well in deployment. There are many algorithms for predictive analysis, and this project focuses on ones that predict continuous values like listing prices.

1.1.1 Linear Regression

Linear regression is one of the most used algorithms for continuous value prediction. In linear regression, the model attempts to define the relationship between two or more variables by fitting a linear equation through all the observed data. Ordinary least squares is often the default method of linear regression and finds the line that creates the smallest total error between the observed and predicted values. These models can help describe linear relationships between variables and determine which ones are significant in explaining how a target variable changes.

1.1.2 Random Forest

Random forest is another machine learning algorithm for prediction. Random forests are able to predict both continuous and categorical variables, making them flexible and increasingly popular. Random forests use ensemble learning with decision trees, meaning that they create many decision trees and combine their predictions to form a single random forest prediction. Decision trees form a branch-like structure by forming ‘splits’ based on values in a predictor variable until they reach a leaf node. An example split could be that all observations with a home size less than 1,000 square feet follow the left branch, and all observations with one above follow the right branch.

1.1.3 Root Mean Square Error

Root mean square error (RMSE) is a metric used to evaluate how well a predictive model performs. The root mean square error takes the square of each residual (the difference between the actual and the observed value), divides it by the number of observations, and takes the square root of this found average. By squaring the residuals, the metric is 1. Able to account for negative residuals that would decrease the error if summed normally and 2. Punishes the model more heavily for outlier errors, as the large residuals get even larger compared to small residuals. Finally, taking the square root of this metric allows the resulting number to be more intuitive for evaluating accuracy, as it will mostly follow the scale of the dependent variable. It is important to note that because the RMSE follows the scale of the dependent variable, it is unable to be used to compare models with different scales.

1.2 Missing Data Mechanisms

There are two main ways that missing data can cause problems. The first is that, when dropping observations with missing values, there is less information for a model to use and, thus, less statistical power. The second is that, depending on why and in what patterns the data is missing, bias can be introduced and result in misleading conclusions. Dropping cases with missing data may lead to underrepresented subgroups that skew sample parameters further from the population parameters. For this reason, it is important to understand the nature of the missing values when deciding how to deal with them.

An important figure in missing data and imputation research who is cited throughout this report is Donald B. Rubin. Rubin created the foundation of types of missing data, initially proposed multiple imputation, and his works are referenced in almost all related academic texts. Rubin defined missing data with three categories, or mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1976). These categories are determined by whether a data point's probability of being missing is dependent on other observed or unobserved data.

1.2.1 Missing Completely at Random (MCAR)

Data points in a dataset that are said to be MCAR have an equal probability of being missing that is completely independent of both observed and unobserved data. This means that there is no systematic difference between the observations with missing data and those without. Examples of MCAR data are an electronic error that caused random measurements to be lost or a survey that was lost in the mail. MCAR data is the most convenient of the missing data mechanisms in that no bias is introduced due to excluding missing subgroups, although statistical power is still decreased due to fewer observations (Rubin, 1987). Data that is MCAR can be considered a whole random sample of less observations compared to the possible dataset with no missing values. Missing completely at random is considered an unrealistic assumption in most real use cases (Buuren, 2021).

1.2.2 Missing at Random (MAR)

Data points in a dataset that are said to be MAR have an unequal probability of being missing depending on their relation to another observed variable. This means that observations with missing values are systematically related to the observed data but not the unobserved data (Rubin, 1987). An example of MAR from the dataset used in this research is listings in the city of San Francisco being more likely to have a missing home size. This means that the home size variable having a missing value is dependent on which city the property is located. Performing an analysis with MAR data by removing the observations with missing values is more likely to introduce bias in the results, as there will be an underrepresentation of subgroups that needs to be accounted for (Rubin, 2021).

1.2.3 Missing Not at Random (MNAR)

Data points in a dataset that are said to be MNAR have an unequal probability of being missing that is related to unobserved variables, rather than the observed ones. This means that observations with missing values are systematically related to the unobserved data but not the observed data (Rubin, 1987). An example of MNAR is people not participating in a drug test because they know they still have drugs in their system. This would create an unobserved group in the dataset that cannot be explained by the other observed variables. Like MAR data, MNAR data can also introduce bias into analysis results if unaccounted for (Buuren, 2021).

The only reliable way found to diagnose MNAR data was to measure the unobtained data. For example, one could follow up with non-respondents of their survey to determine why they did not participate and, potentially, find key differences between their results and the observed results.

1.3 Missing Data Handling

After determining which category or categories the missing data belongs to, one can then decide which method to use to handle it. The method depends on the category and how many of the observations contain missing values, and there is currently no one correct answer even when considering these factors. Research into imputation methods is still lacking, and many analyses are done without much consideration for how the missing values should be handled. In a 1994 collection of datasets used in statistical literature (Hand et al., 1994), only 13 of the 510 datasets had a code for how the missing values were handled, including how many there originally were (Buuren, 2021). This shows that missing data is often not reported and the importance of documenting it is overlooked.

1.3.1 Complete Case Analysis

In complete case analysis, also known as listwise deletion, all observations with missing values are deleted and further analysis is done on all complete cases in the dataset. Complete case analysis is treated as the default method of handling missing data and is often done without further research or reporting. Deleting the entire observations has a high chance of creating bias in parameter estimation if the data is not MCAR (Buuren, 2021).

The error in parameter estimation after complete case analysis also grows rapidly relative to the percentage of missing data, even if it is MCAR. In addition, depending on the field, it is not uncommon for large amounts of data to be missing. A study by King et. al. estimated that the number of incomplete observations in political science data is over

50% on average (2001), enough for a complete case analysis to lose much of its statistical power and lead to unreliable results.

1.3.2 Imputation

The other main way of handling missing data, along with complete case analysis, is imputing the missing values. In imputation, the missing values in a dataset are replaced by a number based on a chosen algorithm. This allows further analysis to keep the data in incomplete observations that would otherwise be lost if removed. Imputation hopes to avoid the problems of complete case analysis by keeping as much data as possible and preventing the deletion of entire subgroups. The downside of imputation is its complexity compared to simply dropping incomplete observations; there is no one correct way to impute missing values, and it can be difficult to understand the effects of a certain method on a dataset.

Single Imputation

Single imputation (SI) methods are every method that results in one imputed dataset. The two most common methods of SI are mean imputation and regression imputation. In mean imputation, the mean value for all non-missing values in a given variable is used to replace all missing values in that column, resulting in only one unique value being used for imputation in each variable. In regression imputation, a linear regression model is made using all complete observations and variables in order to predict each missing value, resulting in a unique imputed value for each missing one.

A possible misconception is that mean imputation is SI due to using one value for each variable, and regression imputation is MI due to using multiple imputed values. However, the distinction between SI and MI comes from the number of datasets made and not the number of values used, thus, both mean and regression imputation are SI methods.

Multiple Imputation

Multiple imputation (MI) was first introduced by Rubin as a way to deal with the inherent uncertainty of imputations, and it creates multiple datasets using a chosen SI method (Rubin, 1987). The steps in MI can be seen in Figure 1. The key part of MI is that the analysis is performed on each of the created datasets first, and the results of those analyses are pooled only after this is done. This is important because, as stated by Stef van Buuren, the creator of the multiple imputation by chained equations (MICE) package for R, averaging the datasets first and then analyzing the single dataset "ignores the between-imputation variability, and hence shares all the drawbacks of single imputation" (Buuren, 2021).

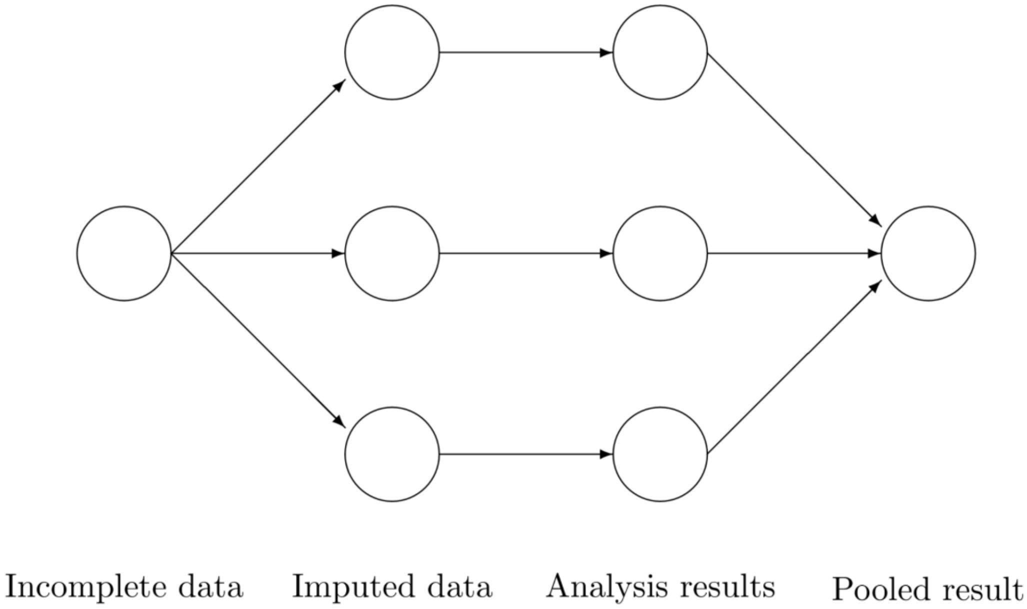

Figure 1: Multiple Imputation Workflow (Buuren, 2021)

Despite Rubin first proposing MI in the 1980's, it has only more recently, in the early 2000's, seen growth in use and is still not used very frequently (Sheuren, 2005). The primary drawbacks of MI lie in its difficulty of understanding and use, especially in the analysis step of the workflow. In addition, due to the lack of work on the topic, the types of analyses used after imputing the data are mainly limited to regression models that have easily poolable parameter coefficients.

1.4 Imputation Schools of Thought

Currently, two main schools of thought can be found in scholarly works dealing with imputation. The first is that imputation should seek to predict the original values as accurately as possible with the thought that if the imputations are accurate, then there is no missing information in the resulting dataset. The second school of thought is more conservative and says that imputation should seek to preserve the variance of the variables and the relationships between the variables even if it means less accurate replication of the data. This school also says that missing values will always be missing and that, since machine learning models assume the imputed values are real, trying to predict the best value undermines this uncertainty and can lead to invalid results.

Despite how it may seem these goals go together, many commonly used imputation methods are unable to accomplish both of these. One such example is regression imputation, which is able to accurately predict the missing values but tends to inflate variable correlation and underestimate their variance. The two overarching categories of imputation methods following these schools of thought are single imputation (SI) for the first and multiple imputation (MI) for the second. Members of the second school of thought could argue that MI methods are the most statistically valid imputations, as they are able to preserve variability and uncertainty in ways that SI methods typically can not.

2 Methods

2.1 Dataset Selection

This project considered two datasets of property listings located primarily in the San Francisco Bay Area. Dataset one contains single family home listings along with location data like commute time scraped by Michael Boles in his Towards Data Science project (Boles, 2019). In Table 1, a description of the dataset can be seen. This table only contains variables used in the model to make it fit.

Table 1: Description of Dataset

City	Price	Beds	Baths	Home Size	Lot Size	School Score	Commute Time
Belmont	1,595,000	4	2	2,220	3,999	77.9	33
Belmont	899,999	2	1	840	4,234	77.9	33
Belmont	1,588,000	3	2	1,860	5,210	77.9	33

This dataset had problems with inconsistent recording of half baths and was more skewed in its variables with large mansions influencing prices and home sizes.

Dataset two was initially much broader and larger, containing all types of properties like single family, condos, townhomes, and empty lots of land. Dataset two was sourced from Kaggle (Roehrich, 2022). Dataset two had parking, number garage spaces, number of stories, pool, and whether it was a new construction variables in addition to the variables in dataset one. However, it also had a problem where lot sizes in acres were not actually given, meaning that the majority of the dataset was rendered unusable. It was also difficult to wrangle the categorical variables due to the lack of structure (36 unique categories for ‘levels’ variable). The data was very noisy and showed similar trends as dataset one with much more variability despite having lower bounds on actual home prices.

Dataset one was chosen because it had a very similar number of observations compared to dataset two after cleaning but with much more focus and less noise. This resulted in a better specified model that could draw stronger statistical conclusions.

2.2 Model Building

Going along with the primary goal of finding the best imputation method for price prediction, it was first necessary to build a predictive model to be used as a benchmark. The first part of the project was evaluating different models on their RMSE in predicting test data.

The dataset was cleaned so that duplicate listings were removed and the 1st and 99th percentiles of each variable were filtered out due to large differences from the rest of data. Keeping these percentiles would have created a lack of observations in those ranges that would make less powerful predictions. Also, all cities with less than 10 listings in them were removed for similar reasons. The workflow for model creation and evaluation can be seen in Figure 2.

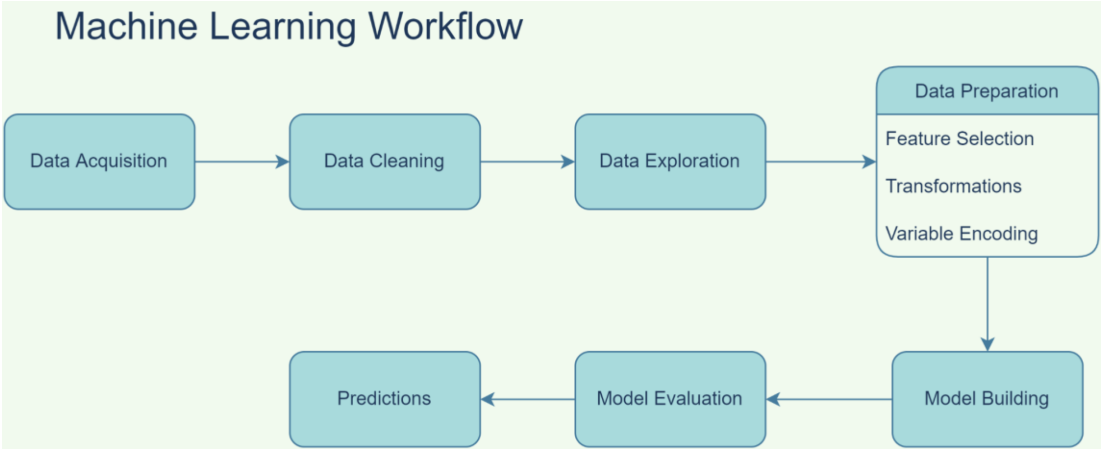

Figure 2: Machine Learning Workflow

The machine learning algorithms tried in this project were multiple linear regression, weighted linear regression, robust linear regression, support vector regression, keras sequential neural network (Allaire & Chollet, 2022), decision tree, random forest, and gradient boosted machine.

The two best performing models were the multiple linear regression and random forest, so time was put into the regression inputs and random forest hyperparameters for final model selection.

2.3 Model Selection

The best performing linear regression model was found using the R lm function in the stats package (R Core Team, 2021). In order to deal with the heteroscedastic and skewed data, the log transformation of price, home size, and lot size were taken. The model also dummy encoded the city variable and included second order interactions (squared terms) of home size and lot size. The average RMSE in an 80/20 train-test split after reversing the log transformation was 323,928.

The best performing random forest model was found using the R randomForest method from the randomForest package (Liaw & Wiener, 2002). This forest contained 500 trees, randomly selected three variables as candidates for each split, and did not encode the city variable. The average RMSE in an 80/20 train-test split was 305,448.

The random forest was selected as the benchmark model for imputation research due to its better performance and easier interpretability compared to a regression with a log transformed target variable.

2.4 Missing Data Creation

For this project, missing data points were created in the filtered dataset with the R package, missMethods (Rockel, 2022), in order to simulate MCAR and MAR data with different missing percentages. Before the missing data was created, the dataset had 20% of observations removed to be used as test data for the final models. A single test split was used for all imputation methods in each trial and no further processing was done on the testing data after the split. This allowed all the models to be evaluated on the same test data in order to compare each method's effectiveness for the primary goal of price prediction.

MCAR data, despite often being unrealistic in actual use, made for a good way of purely evaluating each method's ability to recreate missing data under different amounts of missing values. Because the missing values were not based on any other variables, they could be imputed without concern for any potential underlying relationships. This project stochastically created missing values in 10%, 20%, 30%, and 40% of the home size, lot size, number of baths, number of beds, commute time, and school score variables in multiple trials. The city and price variables were left out of the missing data creation because 1. A listing should never be missing the city because it is part of the address and 2. Price is the target variable, and imputation of the target variable sparks a separate debate that is outside the scope of this project. The completely random layout of missing values can be seen in Figure 3.

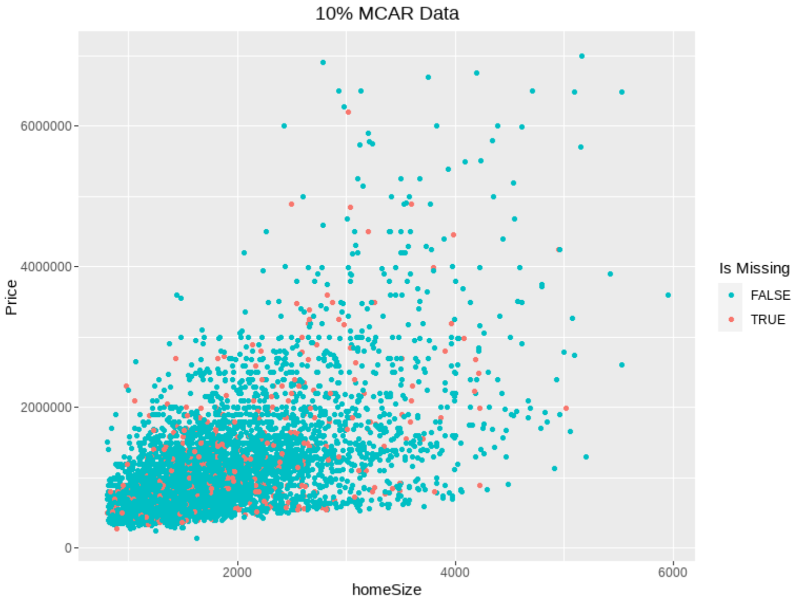

Figure 3: 10% MCAR Data Scatter Plot

MAR trials were also done to get a more real-life applicable understanding of how the different imputation methods performed. When observing the original, uncleaned dataset, it was found that the probability of an observation missing the home size variable was based on the city in which it was located. More specifically, a large amount of the homes in San Francisco had no value for the home size variable. In order to replicate and expand upon this real relationship in the training data, all properties located in San Francisco had their home size removed, and all properties located in Oakland had their lot size removed. This amounted to around 10% of the observations being given missing data. Note that some observations from San Francisco and Oakland were already moved to the test data before this removal process, meaning that the models still had to predict property prices from these cities, just like what would occur during real analysis. It can be seen in Figure 4 that there was a clear trend in the missing data as it made a much tighter line of missing values as price increased.

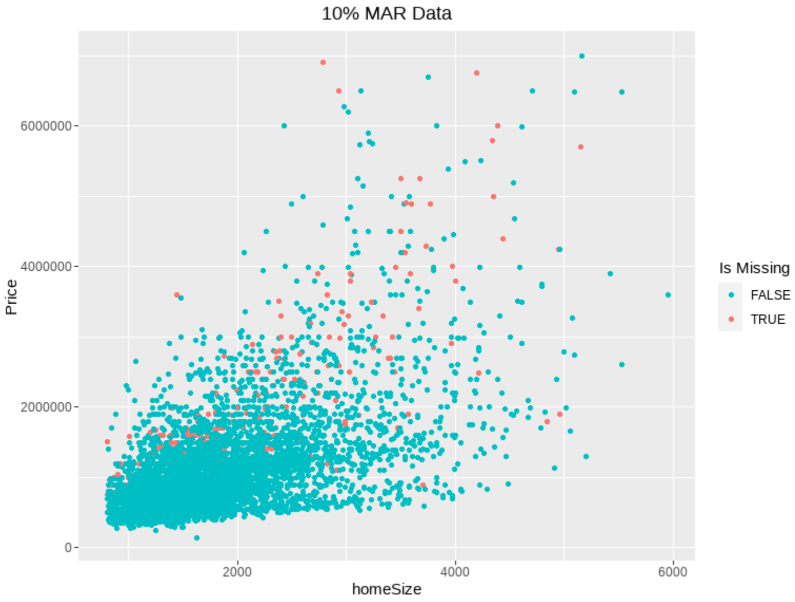

Figure 4: 10% MAR Data Scatterplot

2.5 Imputation Methods Used

The imputation methods tested were chosen for two main reasons. The first reason was their popularity in data analyses, meaning these were the most commonly found methods in other imputation researchers' works. The second reason was that the different methods made up both of the schools of thought on the goals and best practices of imputation.

2.5.1 Mean and Median Imputation

Mean imputation is one of the most commonly used imputation methods due to its simplicity to understand. In mean imputation, all missing values of a certain variable are replaced with the mean value of their respective variable. This means that there is only one unique imputed value for each variable. Median imputation is similar except it uses the median of the variable instead of the mean, resulting in a value that is not affected by outliers.

A possible drawback of mean and median imputation is that imputing the average value for all missing data can decrease standard deviations and variable correlations from their actual values.

2.5.2 Regression Imputation

In regression imputation, the missing values for each variable are predicted by a linear regression model made using the complete cases from all other variables. For a simple example, if a dataset has three variables, then missing values in variable 1 would be predicted using the complete cases in variables 2 and 3, while the missing values in variable 2 would be imputed using the complete cases in variables 1 and 3.

The possible drawbacks of regression imputation are that predicting each value too accurately can actually decrease variable standard deviations and increase the correlations between variables when data is not MCAR.

2.5.3 Stochastic Regression Imputation

Stochastic regression imputation follows the same premise as regression imputation but adds a random error to the predictions in order to minimize the drawbacks of deterministic regression imputation (stop inflation of variable correlations and preserve variance).

A possible drawback of stochastic regression imputation is that adding random errors can lead to implausible imputed values. Another potential problem is that when data is heteroscedastic, the random error changes throughout the distribution, which means that the random error should not be the same throughout the entire dataset.

2.5.4 Predictive Mean Matching Imputation

Predictive mean matching (PMM) is an imputation method added to the end of other methods like regression imputation. PMM takes the predicted value, finds a user-specified number of nearest neighbors to this value, and randomly selects one of them to impute.

Similarly to stochastic regression imputation, PMM attempts to solve the potential problems of regression imputation by randomly selecting neighbors to preserve variance. However, PMM also attempts to solve the potential drawbacks of stochastic regression by only using plausible imputed values from other observed values. Heteroscedasticity is also less of a problem due to using nearest neighbors for each individual value.

2.5.5 Multiple Imputation Methods

This project used the R multiple imputation by chained equations (MICE) implementation of Rubin’s proposed MI method (Buuren & Groothuis-Oudshoorn, 2011) using regression and stochastic regression to evaluate how SI compared with MI. The goal of using MICE was to determine if creating multiple datasets actually had an effect on predictive models and to see how much MICE benefits depended on the imputation method used.

One potential problem with MICE is that, according to the proper workflow, the pooled analysis results are supposed to come from combining the coefficients of the models made using each created dataset. This means that if three datasets are made from MI, then three predictive models are built, and a single model is developed by averaging the coefficients of each model. As mentioned earlier, it is advised not to do any alternative method, such as making an ‘average dataset’ or ‘stacked dataset’ to analyze after (Buuren, 2021).

The problem resulting from this is that MICE documentation does not consider models that do not use clear, interpretable coefficients that cannot be combined. In the context of this project, no right answer on how to perform analysis of MICE datasets on random forest models was found during this research.

Faced with this issue, two options were considered for how to perform MICE analysis on random forests while still following the proper workflow. The first option was to create a smaller forest for each dataset such that they could be combined to contain the same number of trees as the benchmark model. This combined forest is similar to combining model coefficients in that parts of the forest were trained on different imputed datasets and, thus, should preserve the between-imputation variability. The second option considered was to create one full-size forest for each of the datasets and average their predictions. This would still allow separate models to be created according to MICE workflow, but it is unclear how averaging a vote differs from averaging the datasets to begin with.

Because of this uncertainty, the first option was chosen for implementing MICE methods in this research, but there may still be better options that were not encountered while researching solutions.

2.6 Evaluation of Imputation

The primary metric used to evaluate the imputation methods was the root mean square error (RMSE) of the final predictive models trained on each imputed dataset. This is because even though other metrics may allow one to compare the imputed data to the complete data, they do not actually reveal how well these datasets are actually able to predict the unseen test data, which is the primary goal of predictive analysis. The secondary metrics used for imputation evaluation were the RMSE of the imputations themselves, the imputation bias, and how the variable means and standard deviations changed after imputation. The imputation research workflow can be seen in Figure 5.

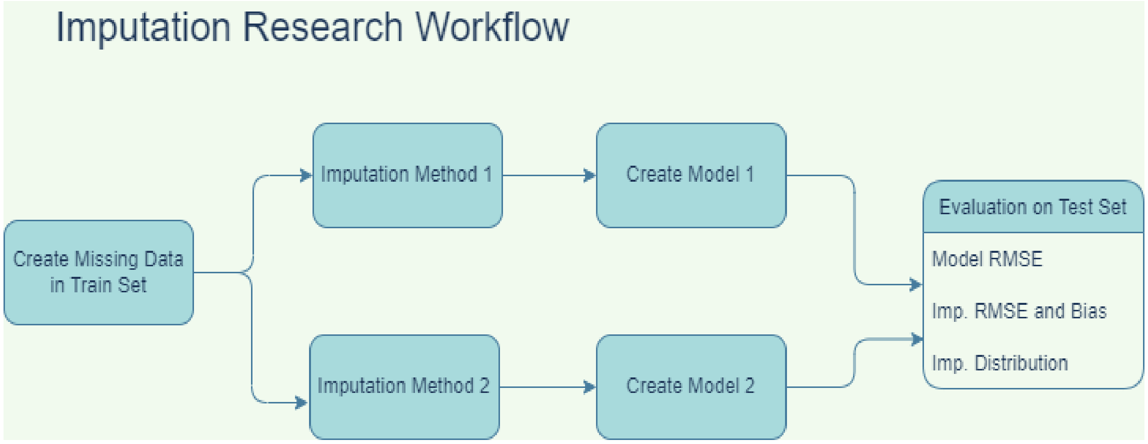

Figure 5: Imputation Research Workflow

One problem encountered was that the MICE methods were unable to be evaluated using the secondary metrics because they create multiple datasets and evaluating their averages defeats the purpose of MI.

3 Results

3.1 MCAR

The first results are for the MCAR tests. Table 2 and Figure 6 show the primary metric, the RMSE of the final model, when trained on the imputed datasets for each method and amount of missing data. The bolded numbers are the ones that performed the best and had the lowest model RMSE.

Table 2: MCAR Model RMSE Results

	10% Missing	20% Missing	30% Missing	40% Missing
Mean	324,905	336,057	352,779	371,417
Median	326,426	337,161	353,910	375,938
Regression	320,980	325,813	332,212	342,631
Stochastic Regression	323,219	330,905	337,374	348,382
PMM	322,299	325,712	335,442	346,382
MICE Regression	**320,024**	**325,216**	**330,270**	**342,460**
MICE Stoch. Regression	320,485	326,143	333,523	343,181
Drop Missing	324,868	347,550	388,718	441,402
Full Data	313,154			

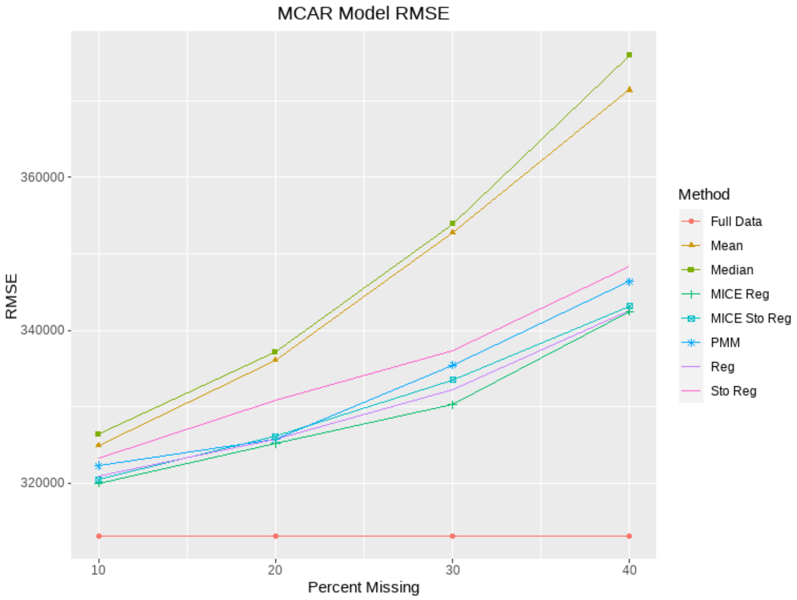

Figure 6: MCAR Model RMSE Results Visualization

The results show that, although other methods perform quite similarly, the MICE regression was able to create the best dataset for predicting the test data for every amount of missing data. It is worth noting that the errors for other regression methods were often less than 1,000 higher than the MICE regression and that all MICE imputations took significantly longer to perform due to building a separate random forest for each dataset created. In addition, a complete case analysis (Drop Missing) performed increasingly worse as the amount of missing data increased.

The secondary metrics, imputation RMSE and imputation bias, gave somewhat unexpected results when compared to model performance. The results can be seen in Tables 3 and 4 and visualized in Figures 7 and 8. The bolded numbers are the ones that performed the best and had the lowest RMSE or bias.

Table 3: MCAR Imputation Percent Bias Results

	10%	20%	30%	40%
Mean	.0166	.0329	.0483	.0663
Median	.0078	**.0156**	**.0227**	**.0311**
Regression	.008	.0167	.0258	.0371
Stoch. Reg.	.008	.0175	.0276	.0396
PMM	**.0077**	.0166	.0265	.0384

Table 4: MCAR Imputation RMSE Results

	10%	20%	30%	40%
Mean	350	489	606	702
Median	354	494	614	709
Regression	**301**	**428**	**540**	**632**
Stoch. Reg.	427	600	763	898
PMM	419	600	767	890

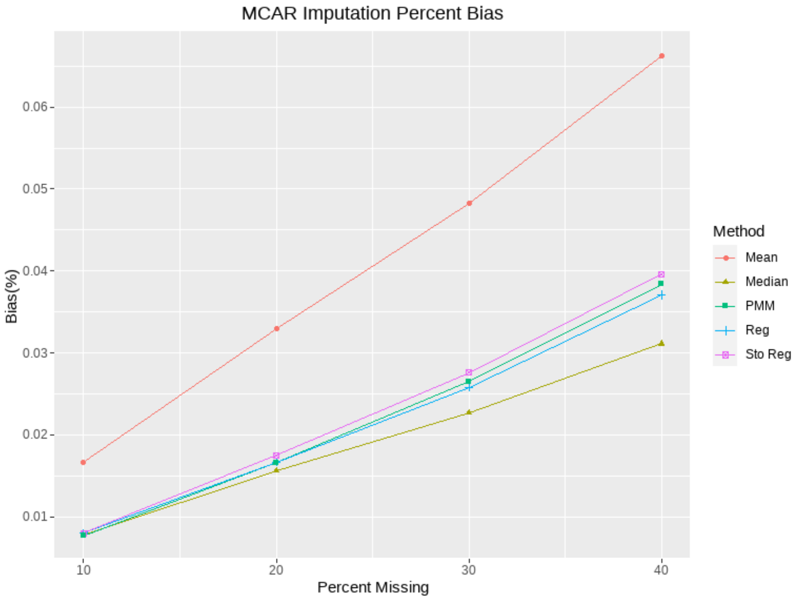

Figure 7: MCAR Imputation Percent Bias Visualization

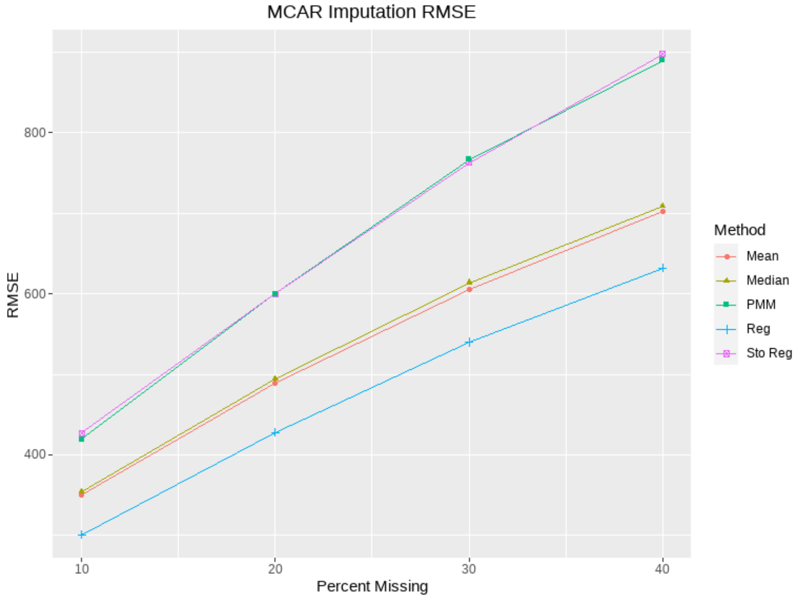

Figure 8: MCAR Imputation RMSE Visualization

The results show that imputation with the median was actually .0001% bias away from having the lowest bias under all percentages of missing data. This was unexpected for two reasons. The first is that median imputation had under half the bias of mean imputation for all percentages of missing data despite how similar the two imputation methods are. The second reason is that median imputation had lower bias than predictive methods like regression, which was very surprising since the dataset's high variance was expected to hurt mean and median imputation performance.

The results also show that regression and MICE regression generally had the best results for all three of these metrics, performing even better than stochastic regression and PMM. As previously mentioned, regression under MCAR is not as dangerous as under MAR and MNAR, but it was still expected that the methods that were made to improve deterministic regression did not outperform it for any of the percentages missing.

This research also compared the mean and standard deviations of each variable for all SI methods in order to compare how well they preserved the distribution of the data. In Figures 9-12 the density plots do well in visualizing the mean and standard deviation in each imputed dataset compared to the original data. The density curves changed increasingly as the percentage of missing data went up, but the overall trends and rankings across the methods stayed the same. Because of this, only 10% and 30% MCAR data is shown, and mean and median imputation were plotted separately from the rest to increase visibility because they extend the y-axis.

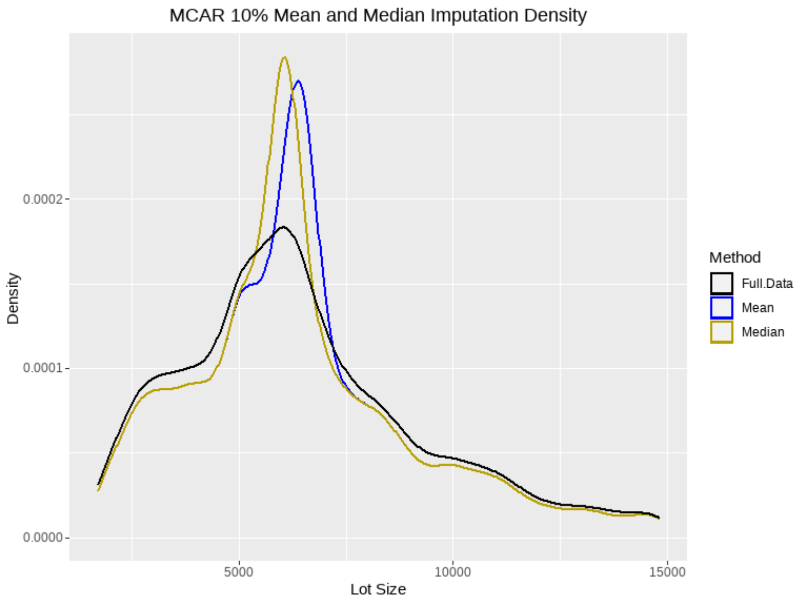

Figure 9: MCAR 10% Average Imputation Density Plot

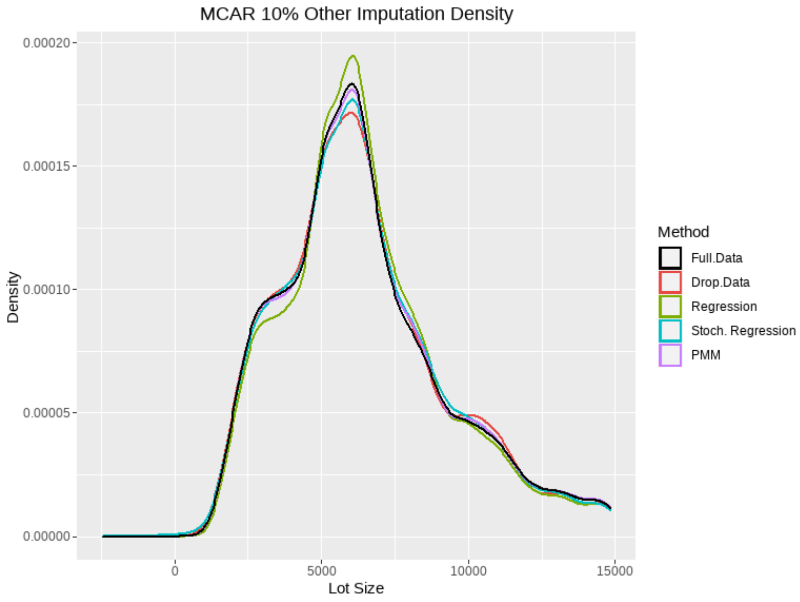

Figure 10: MCAR 10% Other Imputation Density Plot

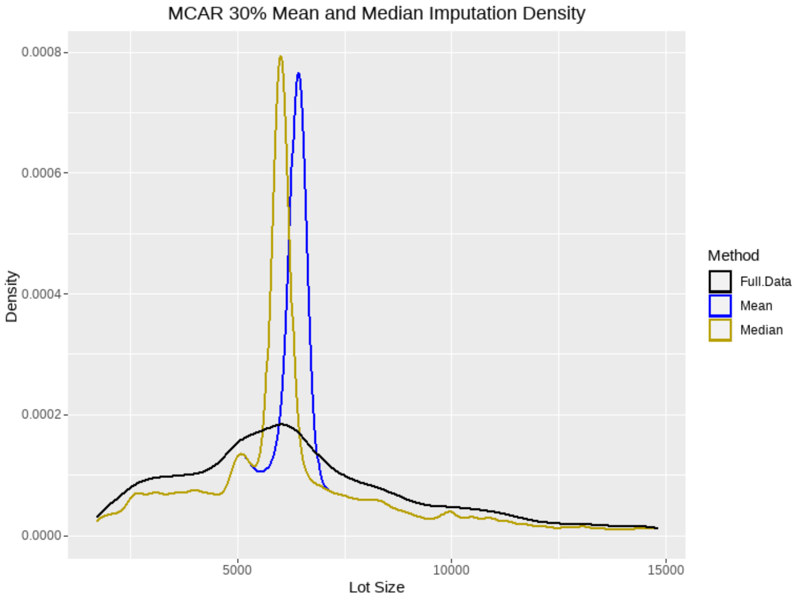

Figure 11: MCAR 30% Average Imputation Density Plot

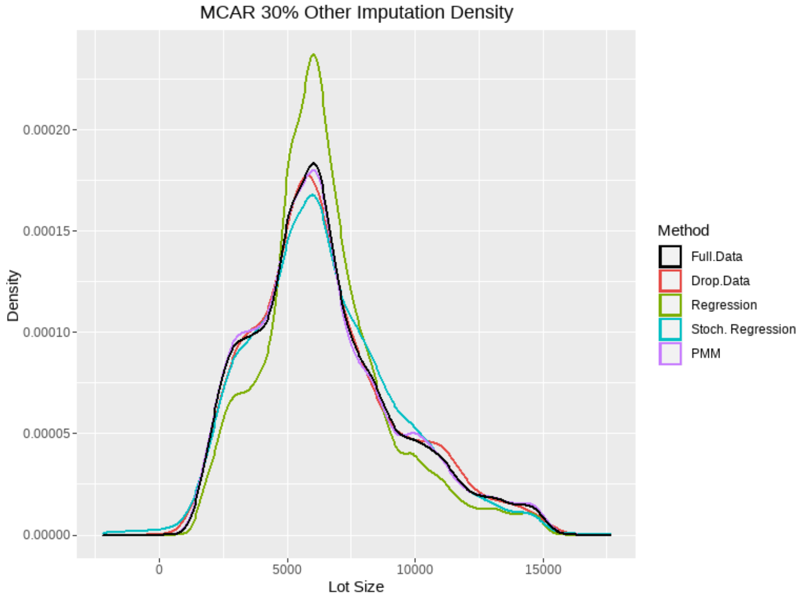

Figure 12: MCAR 30% Other Imputation Density Plot

The results show that complete case analysis resulted in minimal loss of the original data's distribution for all levels of percentage missing. This confirms that, in MCAR data, a complete case analysis results in a large loss of information but does not harm the overall structure of the observations. It can also be seen that, as the percentage of missing data increased, mean, median, and regression imputation increasingly underestimated the standard deviation of the original data. Stochastic regression and PMM accomplished their job of improving this issue by preserving the appropriate amount of variance throughout all amounts of missing data; however, they still ended up with worse model results than deterministic regression. The results also show no trends between mean and median imputed datasets that are able to explain why median imputation's bias was so much lower.

3.2 MAR

Next are the research results for MAR data. To reiterate, missing data points were created in home size or lot size if the listing was located in San Francisco or Oakland respectively, resulting in about 10% incomplete observations being created. This means that no additional trials were done with different percentages of missing data. Table 5 and Figure 13 show the primary metric, the RMSE of the final model, when trained on the imputed datasets for each method and amount of missing data. The bolded number is the one that had the lowest model RMSE.

Table 5: MAR Model RMSE Results

	RMSE
Mean	317,378
Median	317,564
Regression	314,525
Stochastic Regression	310,600
PMM	314,714
MICE Regression	314,471
Mice Stochastic Regression	**310,173**
Drop Missing	328,795
Full Data	311,177

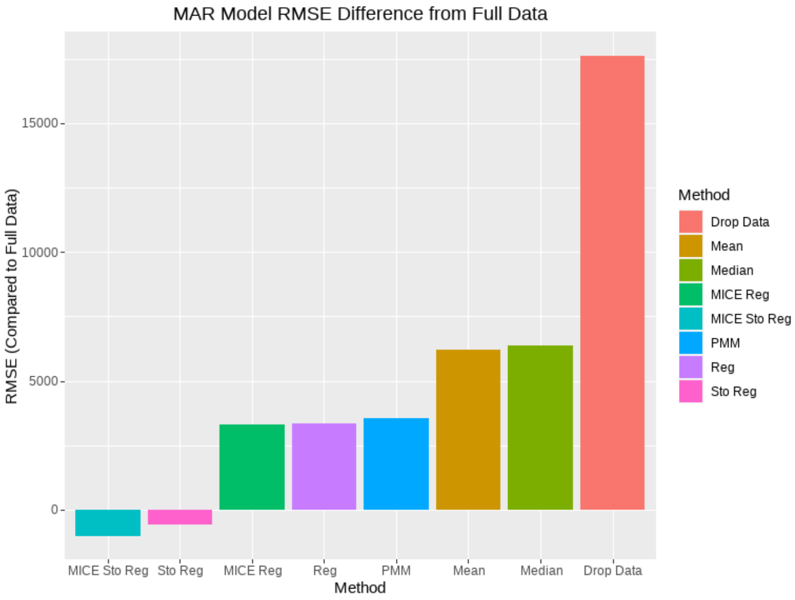

Figure 13: MAR Model RMSE Visualization

The first point to address is that it can also be seen that both SI and MI stochastic regression actually outperformed the full data model; it was not uncommon for most of the imputation methods to outperform the full data model in specific train/test splits. There is no good explanation for this other than possibly the specific split having less outliers either the train or test set that allowed the overall error to be below the full data's. This is not indicative of imputed data being more useful than original data, and the model RMSE is only being used for imputation cross-validation.

The results show that there is a clear difference in imputation method performance based on whether the data is MCAR or MAR. Stochastic regression, both single and multiple, outperformed definitive regression by a fair margin, although performance was closer in some specific train/test splits. This may imply that the downsides of regression imputation are negligible under MCAR but become more important under MAR. The results also show that the difference between SI and MI performance was much smaller under MAR than in MCAR. It is also worth noting that PMM, which was designed to improve upon stochastic regression, performed worse under MAR and worse than MICE stochastic regression under MCAR.

The secondary metrics, imputation RMSE and imputation bias, also gave different results with MAR data. They can be seen in Tables 6 and 7. The bolded numbers are the ones that performed the best and had the lowest RMSE or bias.

Table 6: MAR Imputation Bias Results

	Bias
Mean	.00832
Median	.00611
Regression	.00166
Stochastic Regression	**.00159**
PMM	.00198

Table 7: MAR Imputation RMSE Results

	RMSE
Mean	311
Median	284
Regression	**212**
Stochastic Regression	343
PMM	309

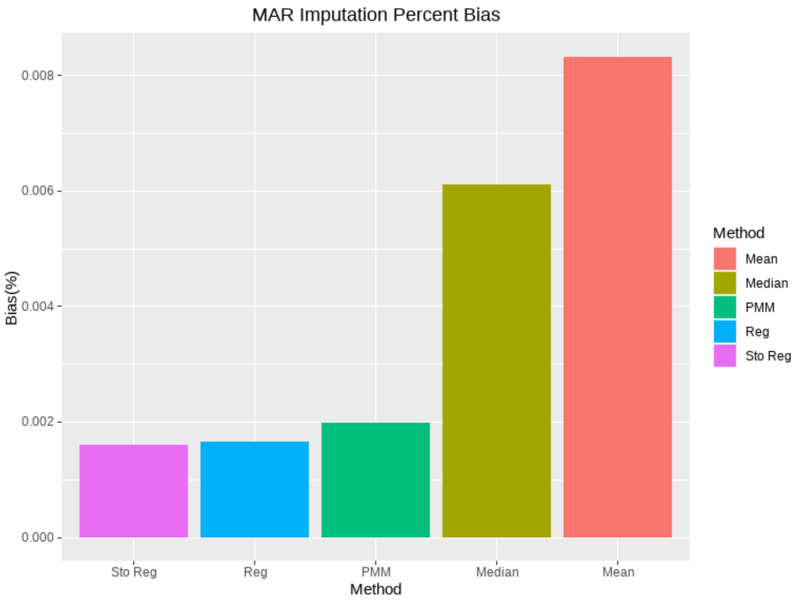

Figure 14: MAR Imputation Percent Bias Visualization

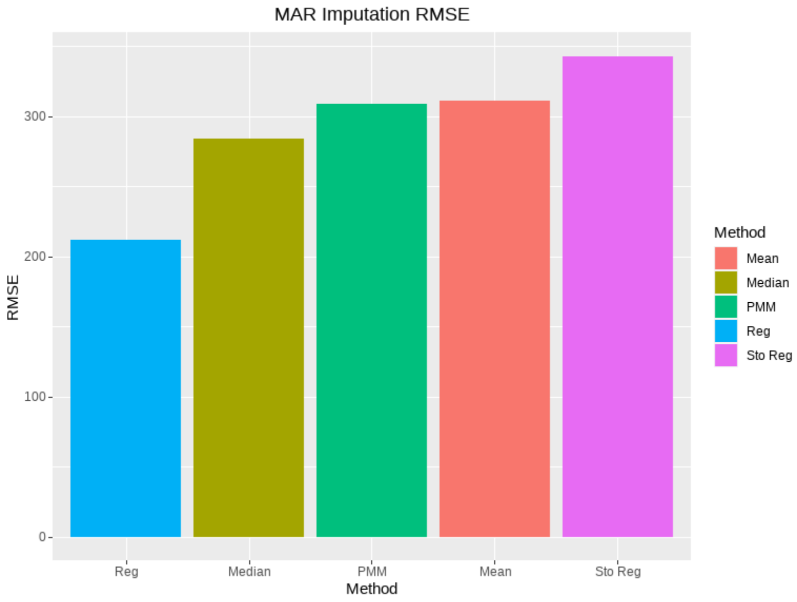

Figure 15: MAR Imputation RMSE Visualization

In the previous tests with MCAR data, stochastic regression had both the worst bias and the worst RMSE, but now with MAR data, stochastic regression has the best bias and still the worst RMSE. This supports Stef van Buuren's claim that RMSE is not a good metric for evaluating imputation results (Buuren, 2021), as stochastic regression had the best model performance and bias but the worst RMSE. Mean and median imputation also had very high bias relative to the other imputation methods. In MCAR data, their biases were close to the other methods with median even having the best, but in MAR data, their biases were about four times higher than the regression based methods.

The MAR research also compared the mean and standard deviations of each variable for all SI methods in order to compare how well they preserved the distribution of the data. In Figures 16 and 17, the density plots do well in visualizing the mean and standard deviation in each imputed dataset compared to the original data.

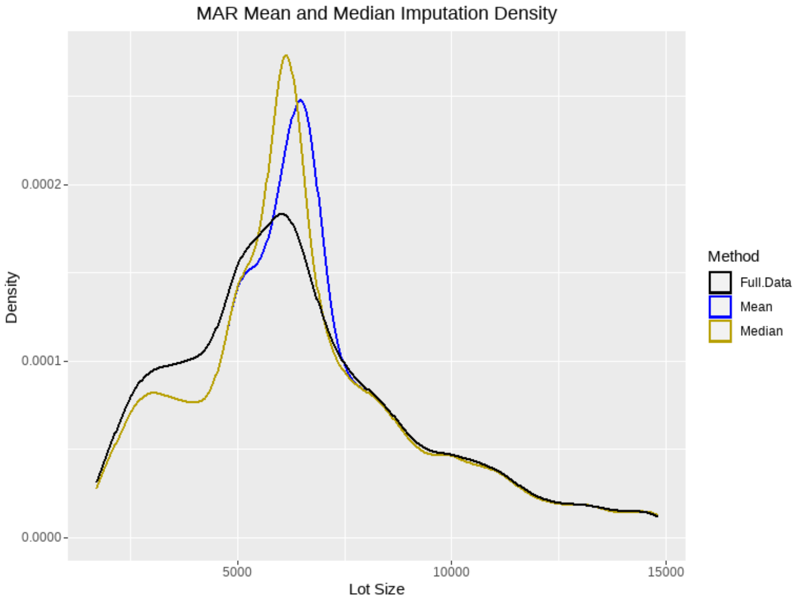

Figure 16: MAR Average Imputation Density Plot

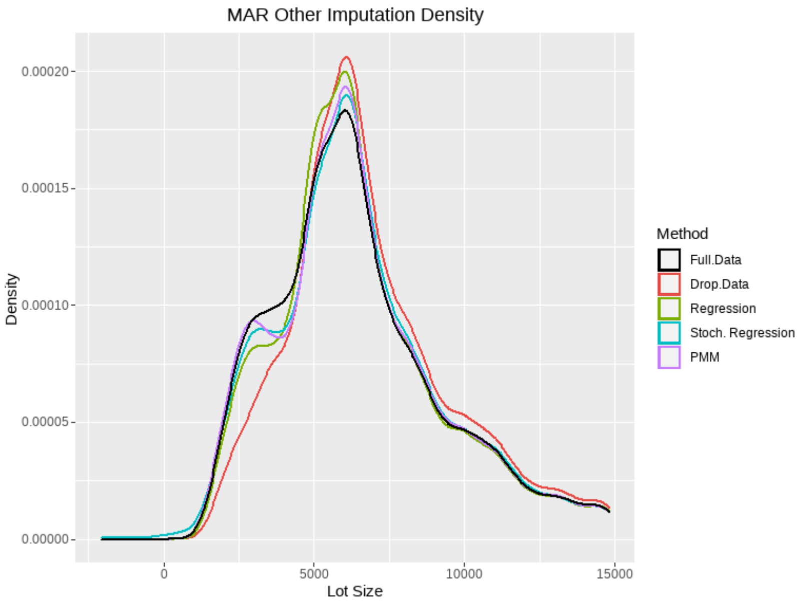

Figure 17: MAR Other Imputation Density Plot

The results show that PMM and stochastic regression best preserved the variance at different points of the data. Complete case analysis (Drop Data) actually preserved the data variance better than mean and median imputation, although it performed far worse in the model due to information loss. Overall, the mean and standard deviation results for MAR were very similar to 10% MCAR due to the similar amount of missing data.

4 Conclusions

This paper compares different single and multiple imputation methods on their ability to minimize error in predicting real estate house listing prices. With the results of this paper, it can be concluded that, out of the methods tried, MICE regression imputation led to the best predictive model results for MCAR data, and MICE stochastic regression led to the best predictive model results for MAR data. Since regression imputation outperformed stochastic regression in MCAR data, stochastic regression cannot be called a total improvement over deterministic regression even if MCAR data is often unrealistic.

The results, for this dataset specifically, support always imputing data over using complete case analysis unless the data is MCAR and has less than 10% missing data, as it performed similar to mean and median imputation at that percentage. However, the regression-based methods outperformed mean and median imputation in every test, meaning that all of complete case analysis and average imputation should generally be avoided in favor of more advanced methods.

The evaluation metrics used, despite being the most used imputation metrics, were unable to explain why some imputation methods performed better than others. There were situations where both schools of thought produced better results than the other, but no clear answer was found for which one should be followed. It was especially difficult to evaluate MI methods using random forest for analysis because there were no parameters to compare with other imputation methods. The model performance benefits of MI were small in both MCAR and MAR, and the additional complexity of analysis may make it unfit for some use cases.

Due to these reasons, this paper concludes that adequate evaluation metrics for imputation were not encountered during research and are not widely in use. However, the results display the importance of understanding the missing data mechanism present in the dataset. The results were noticeably different in the MAR data compared to the MCAR data, meaning that thorough investigation of the missing data should be done before an imputation method is chosen.

The primary goal of this project was to get the best predictive results. If an analyst's concern is exploratory analysis, and they are worried that a certain imputation method would harm data distribution, then the conclusions may differ. Results show that, in all situations tested, PMM imputation was best at preserving the data's distribution. Complete case analysis performed decently in MCAR but the second worst behind regression in MAR, and regression imputation actually preserved the distribution better in MAR than MCAR, despite expert claims of additional bias in MAR.

5 Future Work

With the goal of better prediction, one improvement in future research would be improving the benchmark predictive model. This would be done by finding more variables and new, potentially significant, information for the model. Another improvement would be optimizing the neural network architecture, although this would harm the imputation research process with largely increased model training time.

For improving the imputation research, future work would test more datasets to validate that the results are the same across all datasets. Future work could also have more focus on finding solutions to the MICE implementation issues and seek to find more conclusive results on the debate of SI and MI, potentially trying non-regression based MICE methods. Further MI research could also use regression models instead of random forests with the intent of putting less focus on the model performance and more focus on having the ability to properly evaluate the imputation results with multiple metrics.

Another concern with the results of the research is that the evaluation metrics used were unable to explain parts of the results. Future research would attempt to find new metrics that could better explain the results that this research was unable to explain.

BIBLIOGRAPHY

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Boles, M. (2019, August). *mboles01/Realestate: Contains a set of python scripts for real estate data analysis, visualization, and fitting.* GitHub. Retrieved April 13, 2022, from https://github.com/mboles01/Realestate

Buuren, S. van. (2021). *Flexible Imputation of Missing Data* (2nd ed.). Chapman & Hall/CRC.

Hand, D. J., Daly, F., McConway, K., Lunn, D., & Otrowski, E. (1994). *A Handbook of Small Data Sets*. Chapman & Hall.

JJ Allaire and François Chollet (2022). keras: R Interface to 'Keras'. R package version 2.8.0. https://CRAN.R-project.org/package=keras

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, *95*(1), 49–69. https://doi.org/10.1017/s0003055401000235

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roehrich, G. (2022, February 1). *Real Estate California*. Kaggle. Retrieved April 13, 2022, from https://www.kaggle.com/datasets/yellowj4acket/real-estate-california

Rubin, D. B. (1976). *Inference and Missing Data* (Vol. 63, Ser. Biometrika). Oxford University Press.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Scheuren, F. (2005). Multiple Imputation: How It Began and Continues. *The American Statistician*, *59*(4), 315–319. https://doi.org/10.1198/000313005x74016

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. DOI 10.18637/jss.v045.i03.

Tobias Rockel (2020). missMethods: Methods for Missing Data. R package version 0.2.0. https://CRAN.R-project.org/package=missMethods