Missing Data Analysis

Gupta, Moro, Kannan

1. Business Understanding

1.1 Introduction

The idea of this case study is to understand how a simple linear regression model behaves in the presence of missing data and different imputation techniques. The questions we would like to answer are (1) how the performance of the model is impacted by an increasing number of imputed missing datapoints, (2) what the impact of various types of missingness is on the model performance, and (3) what imputation techniques work best in what scenarios. Very often in informal literature a "Median" imputation is chosen as a de facto imputation technique without regard to the reason for missingness and this can have severe consequences on the model's performance and generalization capability. We hope the outcome of this study would help current and future data scientists understand the importance of determining the reason for missingness and take appropriate actions to overcome this in the model development process.

The study is based on the California housing dataset derived from the 1990 U.S. census, available from the Sklearn python library version 0.23.2, as of November 2020. The dataset contains one observation per census blog group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people) [1]. The model used to answer this case study question is focused on predicting the median income of the residents in each block based on attributes of the houses in the block.

1.2 Data Dictionary

The California housing dataset contains 20,640 entries. Each entry is relative to a specific block group and has a total of eight attributes explained in Table 1.

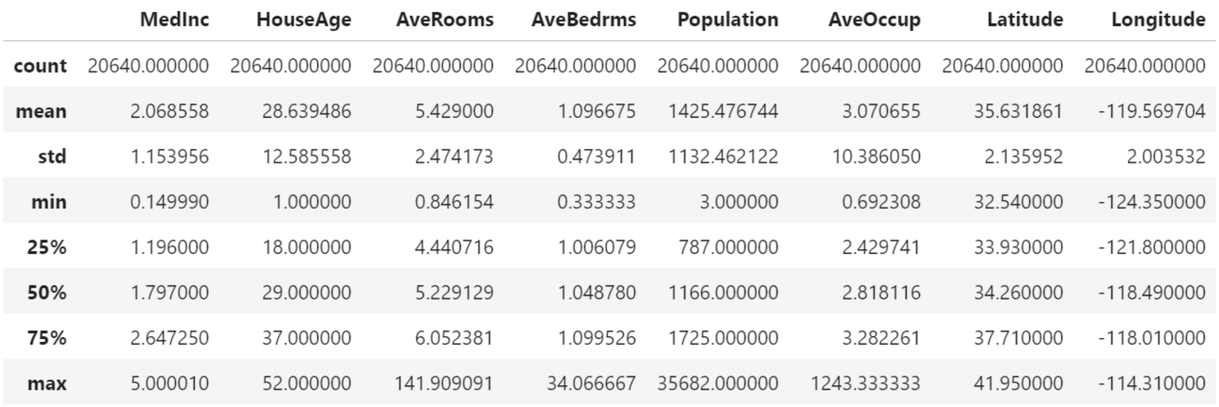
Table 1 Fields Description

Column Name	Type	Description
`MedInc`	Numeric	Median income in the block, measured in tens of thousands of US Dollars
`HouseAge`	Numeric	Median house have in the block
`AveRooms`	Numeric	Average number of rooms per house in the bloce
`AbeBedrms`	Numeric	Average number of bedrooms per house in the block
`Population`	Numeric	Total population in the block
`AveOccup`	Numeric	Average house occupancy in the block
`Latitude`	Numeric	Latitude of the house block
`Longitude`	Numeric	Longitude of the house block.

1.3 Exploratory Analysis

Feature Ranges: An initial analysis of the original dataset (Table 2) shows no missing data point. We notice a large difference in the ranges of the values across columns. For example, `Population` ranges from 3 to 35,682, while `MedInc` ranges from 0.15 to 5. We will need to standardize the range of these values in these features to aid in convergence during the model training process. This will also ensure that the model coefficients are on comparable scales for further evaluation in this case study.

Table 2 Basic statistics on dataset fields



	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	2.068558	28.639486	5.429000	1.096675	1425.476744	3.070655	35.631861	-119.569704
std	1.153956	12.585558	2.474173	0.473911	1132.462122	10.386050	2.135952	2.003532
min	0.149990	1.000000	0.846154	0.333333	3.000000	0.692308	32.540000	-124.350000
25%	1.196000	18.000000	4.440716	1.006079	787.000000	2.429741	33.930000	-121.800000
50%	1.797000	29.000000	5.229129	1.048780	1166.000000	2.818116	34.260000	-118.490000
75%	2.647250	37.000000	6.052381	1.099526	1725.000000	3.282261	37.710000	-118.010000
max	5.000010	52.000000	141.909091	34.066667	35682.000000	1243.333333	41.950000	-114.310000

Feature Distribution: As visible in Fig. 1, all the variables have a non-normal distribution which is expected with real-life data. Furthermore, some variables which are indicative of wealth (e.g. average rooms, average bedrooms) show a right skewed distribution. As the primary objective of this case study was to analyze the impact of missing data on a standard linear regression model, we choose not to transform the non-normal distributed features nor did we analyze the residuals of the models to check for model assumption violations. Since the same non normalized features were used for all model comparisons, it was still a fair comparison. However, for best model metrics, it is recommended that the model assumptions be checked, and appropriate feature transformation be performed to overcome any violations. We will leave this as a post evaluation exercise.

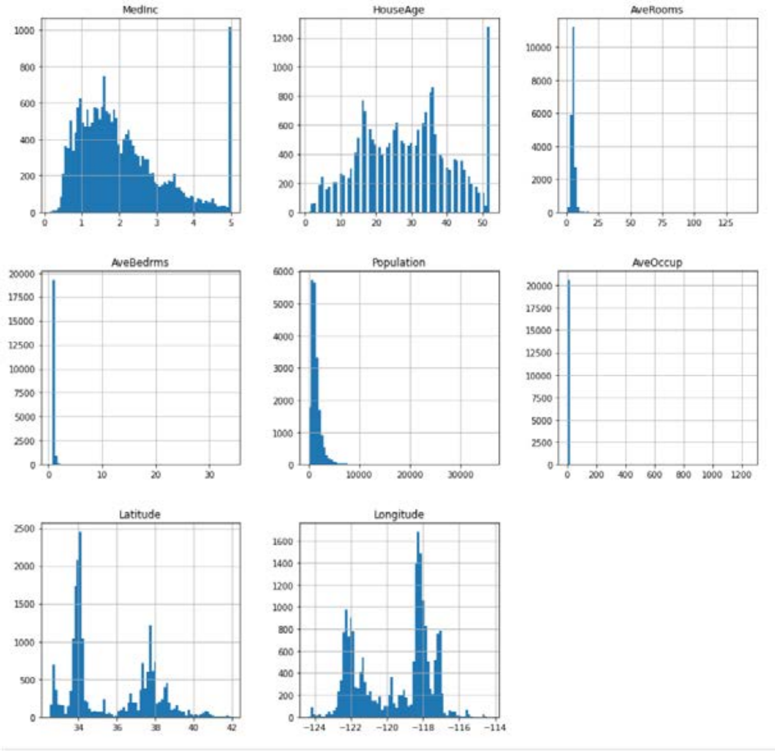


Fig. 1 Distribution of values per each column in the dataset

Target Variable: Fig. 1 also shows that the `MedInc` attribute shows a spike at value 5.0 which was quite unusual. One reason for this could be the grouping of all the values greater than 5.0 into a single value (5.0). Even if this could be the most likely scenario, we didn't find any source confirming it.

Feature Correlation: Fig. 2 shows the scatter plots between pairs of variables, and we notice that some variable shows a strong correlation, such as `AveRooms` and `AveBedrms`. Other variables show a less pronounced linear relation. The correlation matrix in Fig. 3 shows that `MedInc` is slightly positively correlated with `AveRooms` and `HouseAge` columns, and slightly negative correlated with `Latitude` column. This information is important for depicting the best strategy to impute and select missing values as we will see in the subsequent sections.

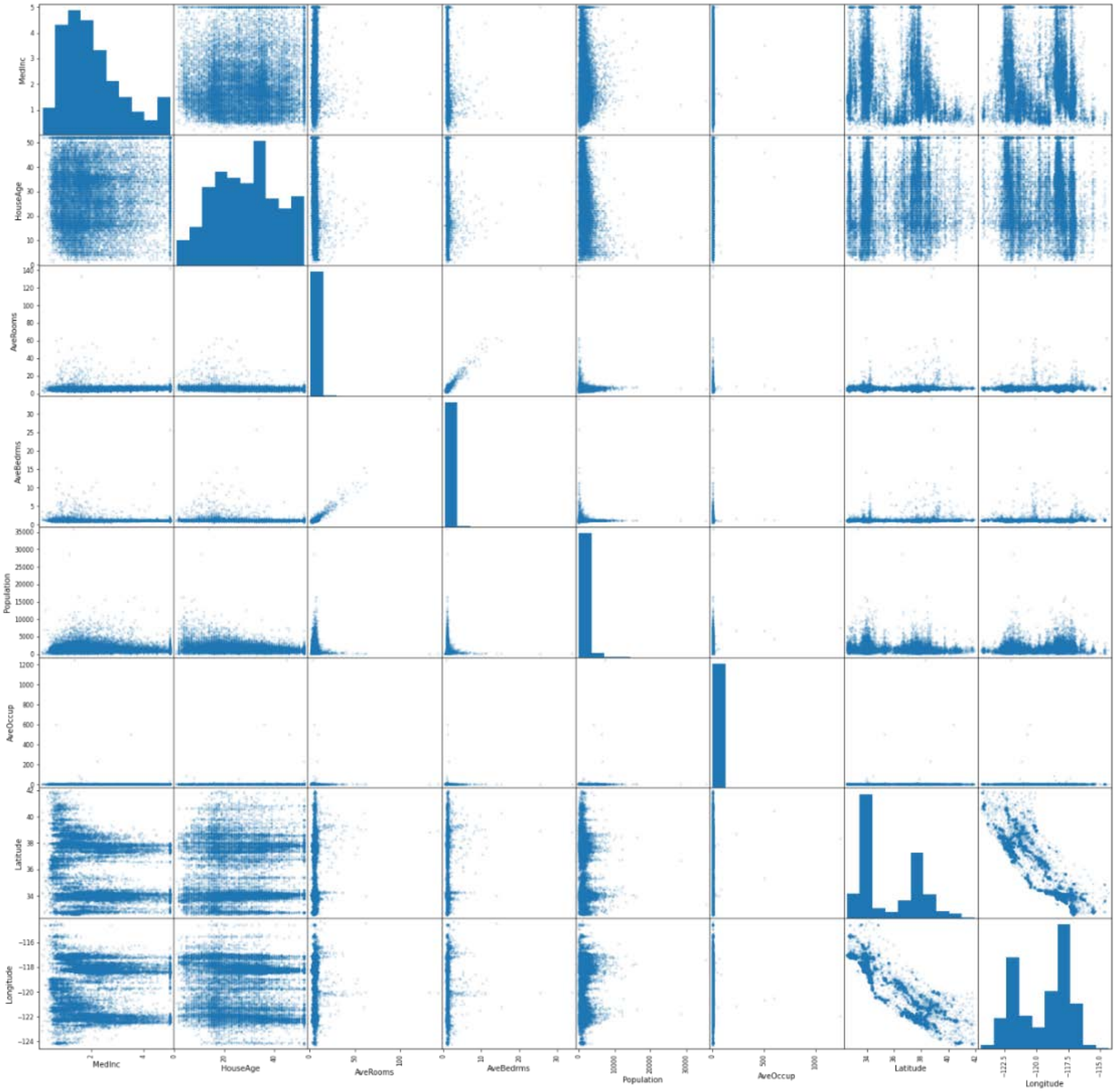


Fig. 2 Scatter Plots Matrix

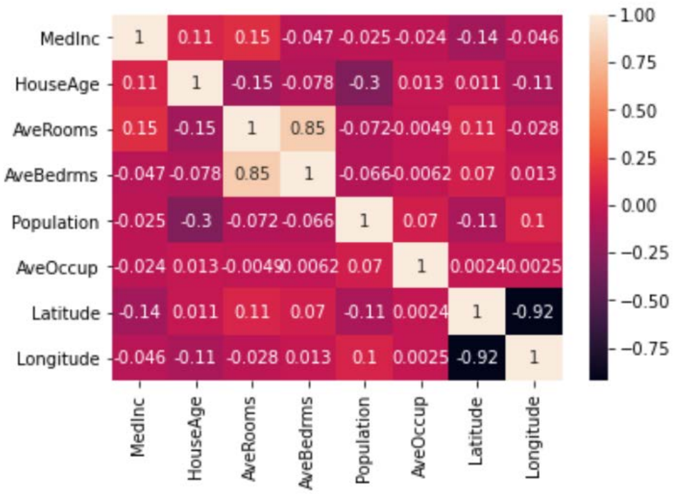


Fig. 3 Correlation Matrix

2. Baseline Model Development

The baseline model is as simple regression model with `MedInc` as the target variable and all other variables used as predictors. As mentioned in 1.3, we performed a standardization of all predictors but no feature normalization. The standardization process subtracts the mean from the individual values and divides the result by its standard deviation. The values obtained after standardization have a mean of 0 and a standard deviation of 1. This helps in variable coefficients comparison since the range of all variables (and hence their coefficients) will be on the same scale.

The model has been trained based of a random selection of 70% of the original dataset and has been measured against the remaining 30% of the dataset (test). This split of train and test dataset has not been altered between the different models to provide a fair assessment of the impact of the missing values and their imputation. The baseline model resulted on the following model:

$$\begin{aligned}\boldsymbol{medInc} = {} & -59.015 + 0.005 * HouseAge + 0.364 * AverRooms - 1.381 * AveBedrms \\ & -0.0001 * Population - 0.001 * AveOccup - 0.731 * Latitude - 0.724 * Longitude\end{aligned}$$

The model developed had a MAE (Mean Absolute Error) of 0.678, an MSE (Mean Squared Error) of 0.808, a RMSE (Root Mean Squared Error) of 0.899 and an R^2 of 0.390 (Table 4). This model can explain 39% of the variance of the target variable (`medInc`), and all the other measures indicates we hare a large range of errors in the model prediction. For the purpose of this case study we will consider all these metrics when comparing various models developed with missing data.

3. Missing Data

After determining the performance of the baseline model (which included no missing values), we analyzed the impact of missing values in one of the variables. We choose the `AvgBedrms` variable to study the impact of missing data as it has the largest magnitude of the coefficient in the baseline model (after standardization) and hence the largest impact on the prediction. In all the imputation techniques that we studied, once a certain number of values were nullified from this selected variable, we replaced them with the **median** of the remaining values. We selected this method as this variable had a non-normal distribution, and the median value is the best representation of the central tendency of the data in this case (as the mean can be skewed due to the non-normal distribution). In addition, we wanted to also study the impact of this often de facto imputation technique on the performance of the model for various types of missingness. Once the missing data was imputed, a new model was trained and compared to the baseline model across the four metrics using the same test dataset used for the baseline model.

3.1 Missing Completely at Random (MCAR)

In this section, we study the behavior of the model when a random selection of values in `AvgBedrms` were replaced with the median of the remaining values. The random selection of the values was based on uniform sampling from the full dataset. Hence all the observations had the same probability of being selected to be converted into a missing entry. We analyzed six different scenarios ranging from 1% to 50% missing values for the `AvgBedrms` column. Table 3 shows the parameters used for all the imputations scenarios. Table 4 and Table 5 show the model performance and coefficient comparison to the baseline respectively.

Table 3 Imputations parameters

Missing %	# rows with missing data	# rows without missing data	Median of rows without missing data (imputed value)
Baseline	0	20,640	
1%	206	20,434	1.04878
5%	1,032	19,608	1.04884
10%	2,064	18,576	1.04925
20%	4,128	16,512	1.04918
33%	6,811	13,829	1.04979
50%	10,320	10,320	1.04939

Table 4 Baseline Model Comparison to MCAR Imputed Dataset

Missing %	Imputation Method	MAE	MSE	RMSE	R^2	MAE diff	MSE diff	RMSE diff	R^2 diff
Baseline	None	0.678	0.808	0.899	0.390				
1%	Median	0.679	0.804	0.897	0.393	0.001	-0.004	-0.002	0.003
5%	Median	0.683	0.810	0.900	0.389	0.005	0.001	0.001	-0.001
10%	Median	0.685	0.812	0.901	0.387	0.007	0.002	0.002	-0.003
20%	Median	0.690	0.824	0.908	0.378	0.012	0.009	0.009	-0.012
33%	Median	0.695	0.833	0.912	0.372	0.017	0.013	0.013	-0.018
50%	Median	0.697	0.835	0.914	0.370	0.019	0.015	0.015	-0.020

Table 5 Coefficients of the Missing Completely at Random Models

Missing %	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	intercept
Baseline	0.005	0.364	-1.381	-0.00001	-0.001	-0.731	-0.724	-59.015
1%	0.004	0.318	-1.179	-0.00001	-0.001	-0.737	-0.732	-59.686
5%	0.004	0.268	-0.959	-0.00001	-0.001	-0.747	-0.745	-60.938
10%	0.004	0.258	-0.914	-0.00001	-0.001	-0.751	-0.750	-61.310
20%	0.003	0.226	-0.742	-0.00001	-0.001	-0.759	-0.759	-62.155
33%	0.003	0.203	-0.655	-0.00001	-0.001	-0.758	-0.770	-63.081
50%	0.003	0.191	-0.664	-0.00001	-0.001	-0.775	-0.777	-63.666

From Table 3, we observe that since the data was nullified (converted to missing) at random, the imputed value remains consistent for all datasets (at around 1.049). From Table 4 we can notice that the model performance degraded slightly as we increased the quantity of missing values, except for the 1% imputation case. For the data with 1% missing values, the R^2 improved by 0.003, MAE improved by 0.001, while MSE, and RMSE had a very small negative impact. Starting from 5% of imputation level, the model started to have a slightly lower performance compared to the baseline. However, it is important to note that this degradation in performance is miniscule compared to the absolute value of the baseline metrics and may not be statistically significant if we were to repeat this process using several different train/test splits or using a cross validation technique. This is expected behavior since the data was missing completely at random and we used a median imputation technique. The median estimate is a valid estimate for a large sample that is randomly picked from an underlying distribution. From this analysis, we can conclude that the median imputation works for data that is missing completely at random.

From Table 5, we also notice that as we increase the percent of missing values, the coefficient for `AvgBedrms` reduces in magnitude from 1.381 in the baseline model (no missing data) to 0.664 for the 50% missing data model. This is accompanied by a slight increase in the magnitude of the coefficients for the `Latitude` and `Longitude` columns which are the columns with the next highest magnitude of coefficients in the original model (after `AvgBedrms`). In essence, the model is transferring importance from one variable (`AvgBedrms`) to others (`Latitude` and `Longitude`) since the amount of "signal" `AvgBedrms` is reducing due to the imputation of the missing values.

3.2 Missing at Random (MAR)

In this section, we study another common type of missing data which is "Missing at Random". In this type of data, the missing data (x) does not depend on the value of x after controlling for another variable z. This can be illustrated using the data in our case study. For example, if people were asked to fill out a survey to collect the housing data, there are chances that some people might not fill the answer to the "number of bedrooms" question if they have already provided the information on the "number of rooms" in their house. This could also be stated as: presence or absence of data in `AveBedrms` is conditionally dependent on the data provided in `AvgRooms.` For this case study, we try to simulate this scenario.

In order to understand the effect of data that is "Missing at Random", we imputed the value of `AvgBedrms` in some of the observations where the value of `AvgRooms` was more than three (3). The value "3" is slightly less than the 25th percentile for `AvgRooms` (Table 2). Hence, after controlling for `AvgRooms` greater than 3, we consider more than 75% of the data in `AvgBedrms` as potential candidates for imputation. This is consistent with the description of "Missing at Random" as explained above. In all, we considered three different scenarios for imputation in this case - with 10%, 20%, and 30% missing data. Like the previous case, we used the median value of the `AvgBedrms` column (for only those entries that were not missing data) for imputation. The imputation statistics are shown in Table 6. Although the data is not missing completely at random, it is still missing at random (after controlling for another variable). Hence the median imputed value remains consistent (at around 1.049) across the various datasets.

Table 6 Missingness statistics for MAR datasets and Imputed Values

Missing %	# rows with missing data	# rows without missing data	Median of rows without missing data (imputed value)
Baseline	0	20,640	
10%	2,018	18,622	1.04870
20%	4,037	16,603	1.04895
30%	6,056	14,584	1.04857

Table 7 shows the model evaluation results for the MAR imputed datasets. Like the previous case when data was missing MCAR, there is a slight degradation in the metrics. We notice that degradation is slightly more than the degradation observed in the MCAR case. For example, in the 10% and 20% cases, the degradation in R^2 was only 0.003 and 0.012 in the case of MCAR. However, in this case (MAR), it is 0.031 and 0.043 which is substantially worse than the MCAR case. This was expected since we have not considered the values of the controlling variable (`AvgRooms`) during the imputation step. However, since the missingness still remains random, the impact of median imputation is still relatively small compared to the absolute value of the baseline metrics.

Table 7 Baseline Model Comparison to MAR Imputed Dataset

Missing %	Imputation Method	MAE	MSE	RMSE	R^2	MAE Diff	MSE Diff	RMSE Diff	R^2 Diff
Baseline	None	0.678	0.808	0.899	0.390				
10%	Median	0.684	0.85	0.922	0.359	0.006	0.042	0.023	-0.031
20%	Median	0.687	0.866	0.93	0.347	0.009	0.058	0.031	-0.043
30%	Median	0.693	0.876	0.936	0.339	0.015	0.068	0.037	-0.051

Table 8 shows the model coefficients and the intercept for each dataset. Like the MCAR case, we see that the coefficient for `AveBedRms` reduces in magnitude as we increase the amount of missing data and this is accompanied with a slight increase in the magnitude of the coefficients for `Latitude` and `Longitude`.

Table 8 Coefficients of the Missing at Random Models

Missing %	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	Intercept
Baseline	0.005	0.364	-1.381	-0.00001	-0.001	-0.731	-0.724	-59.015
10%	0.005	0.314	-1.147	-0.00001	-0.001	-0.745	-0.741	-60.501
20%	0.005	0.309	-1.125	-0.00001	-0. 001	--0.750	-0.746	-60.948
30%	0.004	0.275	-0.967	-0.00001	-0. 001	-0.755	-0.753	-61.585

3.3 Missing “Not At” Random (MNAR)

Next, we evaluate another type of missingness called “Missing Not at Random”. As the name suggests, in this case, the data is not missing at random, i.e. the missing values of a variable depend on what the variable value would have been had it not been missing. For example, in our case study, it may be possible that people on the higher spectrum of the income bracket might not report the average number of bedrooms in their house. If this were indeed the case, then the missing data in the `AvgBedrms` variable would not be missing at random and would be correlated to the value of the variable `AvgBedrms` itself. We will simulate this scenario and study the impact on of imputing this data on the model’s performance.

We saw previously from Table 2 that the complete dataset had a median number of $\mathtt{AvgBedrms} = 1.04878$ with a 75% percentile as 1.099526. Hence the top 25% of blocks have an average number of bedrooms that are greater than 1.099526. We will consider this subset of the dataset (`in_sample`) for imputation in this exercise. The rest of the data (`out_sample`) is left “as is” and used for imputing the missing values.

Table 9 Missingness statistics for MNAR datasets and Imputed Values

Missing %	# rows with missing data	# rows without missing data	Median of rows without missing data (imputed value)
25%	5,162	15,478	1.02817

The statistics of the `out_sample` dataset (Table 9) indicate that it has a median `AvgBedrms` value of 1.02817 which is lower than the full dataset. This is expected since we choose to intentionally nullify the highest 25% of values for this variable. Performing a median imputation in this case would, by definition, introduce a bias in the data since median imputation is more suitable for cases when the data is missing at random. However, to remain consistent with the previous evaluations, we will perform a median imputation. We will follow this up with other imputations to check out the impact of various imputation techniques when data is MNAR.

Table 10 shows the comparison of the original model with the 25% MNAR dataset imputed using the Median value of the `AvgBedrms` column. As we can see, the median imputation is not a very good imputation method for the MNAR case as the data is not missing at random. We have introduced a bias in the analysis by using median imputation and this reflects in the model statistics. The MAE, MSE and RMSE all increase compared to the baseline model and the R^2 shows a significant reduction of 0.137 (> 35% reduction from baseline). This degradation is performance is significantly more than what was observed for the MAR and MCAR cases described above.

Table 10 Baseline Model Comparison to MNAR Imputed Dataset

Missing %	Imputation	MAE	MSE	RMSE	R^2	MAE Diff	MSE Diff	RMSE Diff	R^2 Diff
Baseline	None	0.678	0.808	0.899	0.390				
25%	Median	0.717	0.989	0.995	0.253	0.039	0.181	0.096	-0.137

We think that a better way to impute the MNAR data would be to use the nearest neighbor's (KNN) approach. In this approach, we don't impute using the mean or median of the whole dataset, but rather look for "K" observations (neighbors) that are close to the one that has missing data and then use the mean/median of only those neighbors to impute the data. This can give a more context specific imputation which can be helpful in cases where the data is not missing at random. For example, in this case, we saw that `AvgRooms` was highly correlated to `AvgBedrms` and that observations in `AvgRooms` did not have missing values. Hence, for the observations that had missing values in `AvgBedrms`, we could use `AvgRooms` values to find the nearest neighbors and then use the mean value of `AvgBedrms` for those neighbors to impute the missing data. In this case, each observation with missing data will have its own set of "K" closest neighbors and will potentially get a different imputed value compared to other observations with missing data. This is what we refer to context specific imputation.

In addition to Median and KNN imputation, we tested a few other imputation techniques namely, "Mean Imputation" and "Zero Imputation" (Missing values are imputed with 0 values). The methodology to perform the next set of tests were slightly different compared to the previous approach as outlined in [2]. We used 5-fold cross validation, but the same folds were used to test out each imputation technique, so the comparison was still deemed a fair one.

Table 11. MNAR Imputation Metrics across 5 fold Cross Validation

Imputation Method	Mean "-MAE"	Mean "-MSE"	Mean "-RMSE"	Mean R^2
Baseline (Full Data)	-0.682	-0.817	-0.904	0.386
Zero Imputation	-0.723	-0.929	-0.964	0.302
KNN Imputation	-0.722	-0.929	-0.964	0.303
Mean Imputation	-0.723	-0.929	-0.964	0.302
Median Imputation	-0.723	-0.929	-0.964	0.302

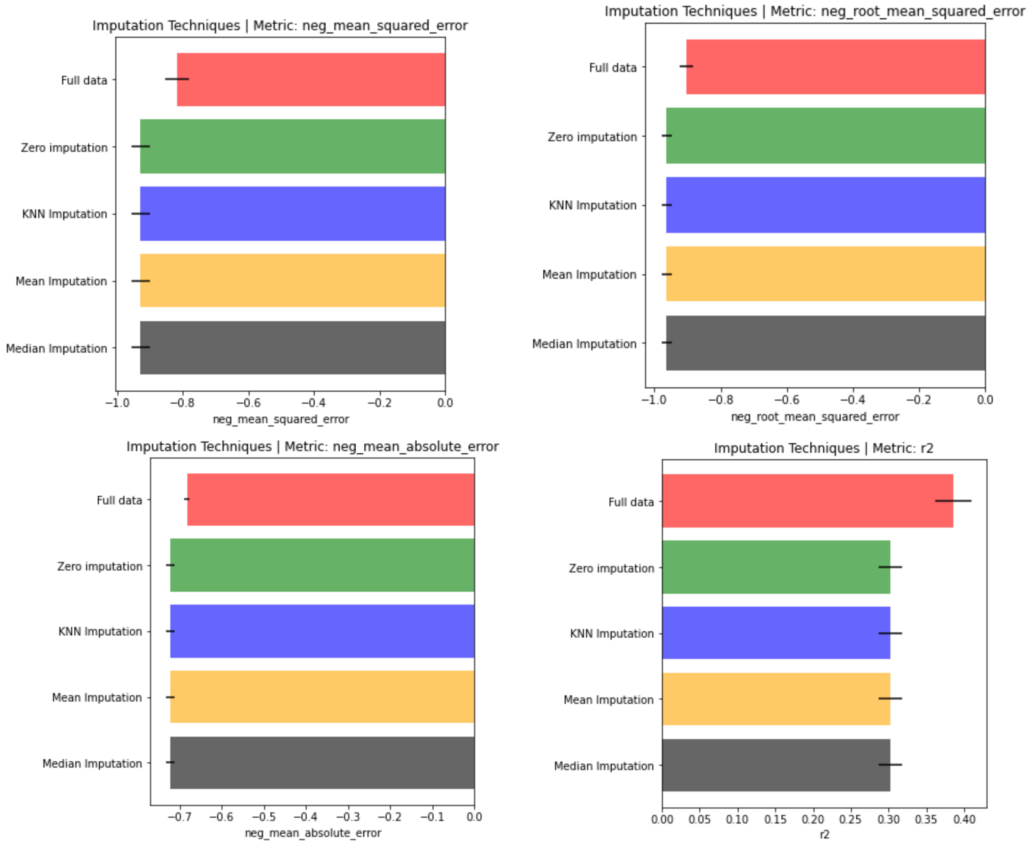


Fig. 4. Impact of Various Imputation Techniques on MNAR Dataset

The results from the test are shown in Fig. 4 and Table 11 which show how the mean "metric" changes with the various imputation techniques used. The same metrics that were used earlier (MAE, MSE, RMSE, R^2) albeit in a slightly different form. For MAE (bottom left), MSE (top left) and RMSE (top right), the values are captured in negated format. The tick marks on the bar plots represents the Standard Deviation of the metric across the 5 folds used. As expected, all imputation methods show a significant deterioration of the model metrics. Also, the standard deviation bars show no overlap between the baseline model and any of the MNAR data sets with various imputation methods. Hence these results are statistically significant as well.

However, we did observe some unexpected results. Based on the theory outlined above, we would have expected the KNN imputation to perform better than the Mean, Median and Zero Imputation techniques, but we did not observe this in the end results. We surmise that this could be because although data that was imputed represented the top 25% of the `AvgBedrms` feature, it was not very different from the bottom 75% (baseline median for `AvgBedrms` was 1.04878 compared to 1.02817 for the MNAR dataset). Hence even though we imputed the top 25% with the nearest neighbor's approach, the imputed values may not have been very different from what would have been imputed using the median imputation or other approaches. There may be other reasons as well and this may warrant further investigation in the future, but we still believe that the KNN imputation will be a better imputation strategy in this case in general.

4. Conclusion

Through this study, we set out to answer three questions - (1) how the performance of the model is impacted by an increasing number of imputed missing datapoints, (2) what the impact of various types of missingness is on the model performance, and (3) what imputation techniques work best in what scenarios.

We considered three different missing data types – "Missing Completely at Random" (MCAR), "Missing at Random" (MAR) and "Missing Not at Random" (MNAR). In all three cases, we evaluated the "median" imputation technique that is often the de facto method used in analysis. We found that in all cases where data is missing, imputation led to a degradation in performance. The degradation was minimal in case of MCAR and MAR with MCAR imputation performing better than MAR. This was expected since although both these data types had data missing at random, in case of MAR, it was missing at random after controlling for another variable. We did not consider this controlling variable in the imputation process. Hence, we expected a slight degradation in MAR performance compared to MCAR through it was not substantial in the grand scheme of things.

The most significant finding was the fact that the median imputation technique does not work well when data was MNAR. Median imputation introduced a bias in the data which led to a significant degradation in performance. In case of MNAR data, other imputation techniques such as KNN imputation may work better. Although we were not able to prove this conclusively due to a lack of variability in our dataset, we still believe that KNN imputation will give better model performance in general for MNAR data since the imputation is context specific and can capture the non-randomness in the missing values better. Proving this conclusively will be left as an exercise for the future.

In conclusion, we hope this case study can serve as a guide for data scientists when dealing with missing data. Very often, little attention is given to the reason for missing data and median imputation technique is applied blindly. However, this case study shows that this may not be the best strategy to obtain the most generalized model and that getting to the root cause of the missing data should be an important first step before any models are developed. Only once we know the root cause of the missing data can appropriate actions be taken to mitigate their impact with minimal loss of performance.

5. References

[1] schikit-learn developers, "California Housing dataset," schikit-learn , 2020. [Online]. Available: https://scikit-learn.org/stable/datasets/index.html#california-housing-dataset. [Accessed 01 11 2020].

[2] "Scikit Imputation," [Online]. Available: https://scikit-learn.org/stable/auto_examples/impute/plot_missing_values.html#sphx-glr-auto-examples-impute-plot-missing-values-py. [Accessed 30 10 2020].

6. Appendix

The entire code used for this case study is available in the following GitHub repository: Repository Link