

# Handling missing values

Itai Mondshine

March 12, 2022

## 1 Abstract

Missing data is an everyday problem that analysts and data scientists need to deal with. Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models. To my knowledge, there is no automated pipeline that deals with this problem. My project tries to propose a pipeline to. This includes: investigating the data and understanding the type of missing data and then trying to fit the best solution in order to fix it.

## 2 Problem description

Missing data are a common challenge encountered by researchers while taking a research. it occurs across all types of studies. The optimum approach to missing data is to ensure that strategies are devised to ensure that the amount of missing data in a study is as small as possible. Dealing with missing data may be low on the list of priorities for a researcher when undertaking a study but it is a vital step in data analysis as inappropriate handling of missing data can lead to a variety of problems. These included a loss of statistical power, loss of representation of key subgroups of the cohort, biased or inaccurate estimates of treatment effects and increased complexity of statistical analysis. It's an important step in the data preparation step.

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. Below is a sample of the missing data from the Titanic data set. You can see the columns 'Age' and 'Cabin' have some missing values.

We all know, that data preparation is one of the most time-consuming stages in the data analysis process. We need to acquire missing values, check their distribution, figure out the patterns, and make a decision on how to fill the spaces.

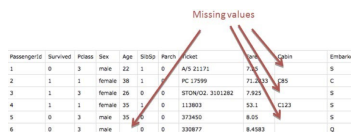
### 2.1 How is Missing Value Represented In The Dataset?

In the data set, blank rows show the missing values. In Pandas, usually, missing values are represented by NaN. It stands for Not a Number.

Missingness is broadly categorized into 3 categories:

#### 2.1.1 Missing Completely at Random (MCAR)

When we say data are missing completely at random, we mean that the missingness has nothing to do with the observation being studied (Completely Observed Variable (X) and Partly Missing Variable (Y)). For example, a weighing scale that ran out of batteries, a questionnaire might be lost in the post,



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	53.1	F85	C
3	1	3	female	26	0	0	STON/OJ: 2101282	7.92		S
4	1	1	female	35	0	0	113803	53.1	F123	S
5	0	3	male	26	0	0	374602	8.51		S
6	0	3	male	0	0	0	208077	8.45		Q

Figure 1: an example of missing values from the Titanic dataset.

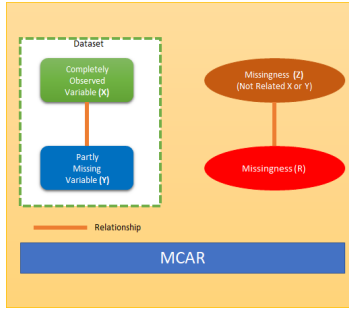


Figure 2: an example of missing data of type MCAR.

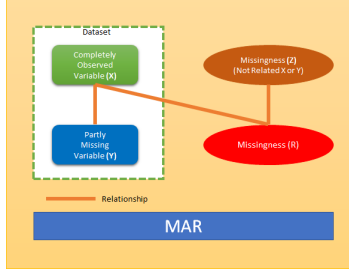


Figure 3: an example of missing data of type MAR.

or a blood sample might be damaged in the lab. MCAR is an ideal but unreasonable assumption. Generally, data are regarded as being MCAR when data are missing by design, because of an equipment failure or because the samples are lost in transit or technically unsatisfactory. The statistical advantage of data that are MCAR is that the analysis remains unbiased. A pictorial view of MCAR is below where missingness has no relation to dataset variables  $X$  or  $Y$ . Missingness is not related to  $X$  or  $Y$  but some other reason  $Z$ .

### 2.1.2 Missing at Random (MAR)

When we say data are missing at random, we mean that missing data on a partly missing variable ( $Y$ ) is related to some other completely observed variables( $X$ ) in the analysis model but not to the values of  $Y$  itself. A pictorial view of MAR as below where missingness relates to dataset variable  $X$  but not with  $Y$ . It can have other relationships ( $Z$ ). It is not specifically related to the missing information. For example, if a child does not attend an examination because the child is ill, this might be predictable from other data about the child's health, but it would not be related to what we would have examined had the child not been ill. Some may think that MAR does not present a problem. However, MAR does not mean that the missing data can be ignored.

### 2.1.3 Missing Completely at Random (MNAR)

If the data characters do not meet those of MCAR or MAR, they fall into the category of missing not at random (MNAR). When data are missing, not at random, the missingness is specifically related to what is missing, e.g. a person does not attend a drug test because the person took drugs the night before. A person did not take an English proficiency test due to his poor English language skill. The cases of MNAR data are problematic. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data, but that requires proper understanding and domain knowledge of the missing variable. The model may then be incorporated into a more complex one for estimating the missing values. A pictorial view of MNAR is below where missingness directly relates to variable  $Y$ . It can have other relationships ( $X$   $Z$ ).

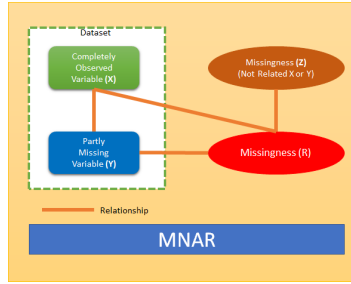


Figure 4: an example of missing data from of type MNAR.

### 3 Solution overview

As we will show later in the related work subsection there are a lot of solutions uniquely for each type of missing values. The problem is we can't know for sure which solution fits more.

Today, to my knowledge, there is no an automatic solution or an attitude which helps us to detect the exact type of missingness. This a great problem. Actually, the best way to detect the exact type of missingness is knowing and get familiar with the domain of data. Without it, it's very difficult to point the type.

In our solution we aim to show couple of things. Firstly, and most important, we show an automatic pipeline for detecting the type of missingness and in the second part, we programmed a automatic tool that finds the best handling method.

The logic behind our solution is that each type of missingness has a unique pattern of the missing values. Thus, because each type indicates to different phenomenon in the data. Using this understanding we can program a tool that tries to find patterns in the missing values and then tries to offer the best fit.

In order to do this, we used the library called missingno in Python library which helps us to identify and visualise missing Data. Missingno has two great functions: a matrix and a heatmap (you can see in the JUPYTER notebook an example of using this library)

In our solution we are focusing in analyzing the type of missingness visually by finding the patterns of missing data. In general, there are two types of missing data: MCAR and MNAR. In MNAR we need to find the dependencies between missing features, and start the data gathering process. In MCAR it means that there are no deep patterns in missing values. So, we can work with that and decide if some rows/features may be removed or imputed.

Our solution contains two steps: 1. Detecting the type of missing data 2. applying the best method to solve the missing values.

we will show a method based on visually investigating the data.

#### 3.1 Detecting the type of missingness

##### a. Missing Completely At Random, MCAR

In this type, such values have no relationship with existing values. They often occur because of technical of human errors during data entry. For MCAR we checked for two things: a low correlation and a small percentage of missing values among others.

we will have a look at the heatmap chat (fig 5):

in columns we see a low correlation and a number values less that 20 percent we will sign them as MCAR.

##### b. Missing At Random, MAR

As we have explained earlier MAR is a a broader class of MCAR. the random loss of information **is only related to a certain group of data**. in MAR we actually need to find a relation between our feature and another feature. It means, we look for a relation between the randomness of our columns and the other feature.

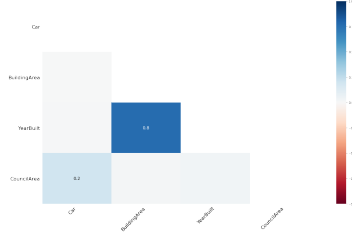


Figure 5: an example of missing data from of type MCAR.

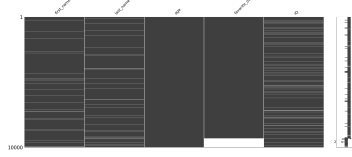


Figure 6: when we sort by the feature 'Age' , we can see that all the null values of the feature 'favorite\_s\_comeclose'.

in order to see this we sort the values of all the other features and for each sort we look at our feature distribution of missing values. If all the missing values will come close together - it means that there is a relation between the two features.

for example, let's have a look:

Visually, it's easy to see, but in code we encountered with a problem to program this. Finally, we decided to measure the difference between two continuous series of null values. We used a function that find the longest sub array of null values and compared the lengthens before and after sorting the arrays. If we show that there is high increase in the length of the sub array - it indicates that one feature affects the other one. Also, we checked that there is a correlation between the checked feature and other features.

### c. Missing Not At Random, MNAR

This is the most challenging type of missingness. Still, it is randomly scattered but there is a unobserved factor that affects the missingness. Usually, one can consider it when get familiar with the domain of data.

we can see the values are still random but we can't find any relation to any other feature.

To sum up, our method suggests a new approach for detecting the type of missingness. Today, there is no test for detecting the type of MAR and MNAR. The only test commenly used is chi2 test for detecting the hypothethis that the data is MCAR. Our method is unique because it gives an innovative and creative way to deal with missing values.

## 3.2 Applying a soulution to fix the missing values

### . Automatic imputing the missing values

In this part we trying to build an automate script that given a feature finds the best method to fill the missing data. At first, I have tried to find a solution for each type of missingness (mcar, mar, mnar), but i found it very difficult. In addition, to my opinion, it's more powerfull to use a generative method

### . Evaluation

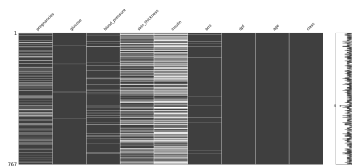


Figure 7: an example of missing data from of type MNAR.

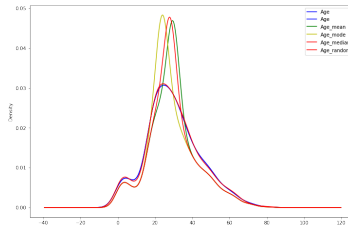


Figure 8: Comparing the different methods.

Data skewness represents the extent to which data values are not symmetrical around the mean. One way of measuring skewness is by studying the mean and median of the distribution.

In our automatic pipeline we search for the method which gives the closest skewness to the original skewness of the data without the missing values.

In this solution we compared between 4 methods recommended for handling missing values : mean, mode, median and random.

## 4 Experimental evaluation

Missing data is an everyday problem that analysts need to deal with. Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models. To my knowledge, there is no automated pipeline that deals with this problem. My project tries to propose a pipeline to. This includes: investigating the data and understanding the type of missing data and then trying to fit the best solution.

## 5 Related work

<https://www.scirp.org/journal/paperinformation.aspx?paperid=111220>

Missing data is an everyday problem that analysts need to deal with. Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models. To my knowledge, there is no automated pipeline that deals with this problem. My project tries to propose a pipeline to. This includes: investigating the data and understanding the type of missing data and then trying to fit the best solution.

Today, there is a statistical test called :

### 5.1 Little's MCAR test

Tests the null hypothesis that the missing data is Missing Completely At Random (MCAR). A p.value of less than 0.05 is usually interpreted as being that the missing data is not MCAR (MAR or MNAR). The problem is it's not absolute. Another method of attempting to identify the classification is creating a dummy variables, missing data will be 1 and observed 0. Then, will be performed t-tests, or chi-squared test.

### 5.2 MAR or MNAR

The real problem is performing a test of MAR against MNAR because it requires unavailable data.

making it easy to misclassify one mechanism for the center. One method is to follow up with a survey or questionnaire.

### 5.3 Handling missing values

There are a lot of ways to deal with missing data. Some of them give good results and some don't. The most useful methods are:

1. Delete Rows with Missing Values - simply delete rows or columns having null values. If columns have more than half of the rows as null then the entire column can be dropped.

2. imputing values with mean / median
3. Using Algorithms that support missing values - A great method we haven't used. in this method we use ML algorithms for handling missing values. Some of them are robust to missing values, like : K-NN. Navie bayes can also support missing values when making a prediction.
4. prediction of missing values - using regression or classification model for prediction of missing values.

## 6 Conclusion

Missing data is an everyday problem that analysts need to deal with. Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models. To my knowledge, **there is no automated pipeline that deals with this problem**. My project tries to propose a pipeline to. This includes: investigating the data and understanding the type of missing data and then trying to fit the best solution.

I have learned from the project that there is not enough research on the subject of dealing with missing data.