Missing Values and replacement policies

- Most of the effort with the project connected with data is spent on data preparation, sometimes it can take up to 90 percent of the overall time spent on the project.
- Dealing with missing data is one of the most difficult parts in the data preparation phase.
- One of the reasons that it is considered difficult is that there is no best way to deal with missing values.
- In order to understand what to do with missing values found in your dataset, firstly, you need to understand what type of missing values you have. When I first faced the problem of missing data, it was difficult to understand the meaning of their types, that's what I will try to explain in this article with simple and clear examples.

Three kinds of missing data:

- Missing at Random (MAR)
- Missing Completely at Random (MCAR)
- Missing Not at Random (MNAR)

Let's imagine that we are trying to predict the price of the car that is being sold, for example in eBay, the data may look like this:

Model	Year	Color	Mileage	Price
Chevrolet	2014	NaN	10000	50000
Ford	2001	White	NaN	20000
Toyota	2005	Red	NaN	30000
Crysler	2019	Black	0	100000

Let us understand the different types of missing values with Example:

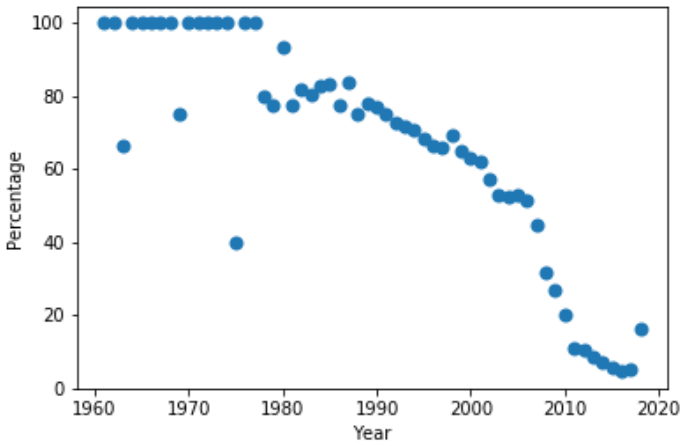
Missing at Random (MAR)

MAR data — means there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data.

What it means, is that the missingness of data can be predicted by other features in the dataset.

In the table above table

- the mileage has few missing values.
- the manufacturing year of cars with missing values is lower than in other examples
-



Graph for Percentage of missing values in the mileage column depending on the year of the car

- The correlation between the percentage of missing values in the mileage column and the manufacturing year of the car is clearly seen
- It is clearly seen that older the car, more the probability that the mileage will not be provided by the seller of the car.
- we can predict the missingness of the mileage of the car, from its manufacturing year (may be with help of some expert)

Missing Completely at Random (MCAR)

- MCAR means there is no relationship between the missingness of the data and any values, observed or missing.
- This kind of missing values is the easiest to understand.
- The fact that the data is missing has nothing to do neither with observed data nor with non-observed data, it's just missing.
- There is no logic in it.
- In the above given table there is missing value in the color column… random and non systematic one. (May be Someone just forgot to mention the color)

Missing Not at Random (MNAR)

MNAR data is the most complicated one both in terms of finding it and dealing with it. The fact that the data is missing is related to the unobserved data, i.e. the data that we don't have, the missingness is related to factors that we didn't account for.

Another Example

MAR :

- The missing data here is affected only by the complete (observed) variables and not by the characteristics of the missing data itself.
- in other words , for a data point , to be missing is not related to the missing data, but it is related to some of (or all) the observed data , the following example will depicts the situation and make it more clear :

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	
29	
30	
30	
31	
44	118
46	93
48	141
51	104
51	116
54	97

- We could easily notice that IQ score is missing for youngsters (age < 44 yo) , and thus the missing data depends on the observed data , however there is no dependency with the values of the missing column itself .

Missing Completely at Random (MCAR)

- There's no relationship between whether a data point is missing and any values in the data set (missing or observed) .
- The missing data are just a random subset of the data .
- The missingness is nothing to do with any other variable . By the way , data are rarely MCAR.

- following example will depicts this kind of problem :

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

- It is relatively easy to check the assumption that in our example data is missing completely at random. If you can predict any reason for missing data (e.g., using common sense, regression, or some other method) whether based on the complete variable Age or the Missing variable IQ score , then the data is not MCAR !

Missing Not at Random (MNAR)

- The data will be missing based on the missing column itself , for instance the following example points out the fact that data are missing on IQ score with only the people having a low score .

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

As you can see , it is impossible to detect MNAR cases without knowing the missing values !

Coping And Dealing With Missing Data Problems (Missing Value Management)

Method 1: Deletion

it falls under two different techniques :

- **Listwise Deletion** : In this method, an entire record is excluded from analysis if any single value is missing , and therefore we have the same N (number of records) for all analysis .
- **Pairwise Deletion** : during our analysis the number of records taken into consideration denoted "N" will vary according to the studied variable (column) , and for instance we could compute the mean for 2 features (Complete VS missing) and while dividing by the number of samples , we end up dividing by different N , one is the total number of rows and the other is the total number on complete values on the missing feature .

Example:

```
X =
   {. 3 2,
    8 . 2,
    1 5 8,
    1 3 5,
    2 4 3,
    4 5 3 };
```

Listwise deletion is the operation used by regression procedures to deal with missing values.

During listwise deletion, an observation that contains a missing value in *any* variable is discarded;
no portion of that observation is used when building "cross product" matrices such as the covariance or correlation matrix.
For our example, listwise deletion means that the correlation matrix is formed by using rows 3–6, as follows:

```
/* Listwise deletion matrix:
    1 5 8,
    1 3 5,
    2 4 3,
    4 5 3     */
```

ListCorr		
1.000	0.492	-0.698
0.492	1.000	0.184
-0.698	0.184	1.000

What happens if you form the matrix that consists of pairwise correlations? That is, form the array C such that C[*i*,*j*] is the correlation between the *i*th and the *j*th columns of X. The missing values *for each pair* of variables are deleted based on whether either variable contains a missing value.
Under this pairwise-deletion scheme, each element of C is computed by using different observations:

- The element C[1,2] is computed by using observations 3–6 because the first observation has a missing value for X1 and the second observation has a missing value for X2.
- The element C[1,3] is computed by using observations 2–6 because the first observation is missing for X1.
- The element C[2,3] is computed by using observations 1 and 3–6 because the second observation is missing for X2.

The following SAS/IML statements compute the array of pairwise correlations:

```
/* vectors used for pairwise correlation
  R12    R13    R23
              3 2
         8 2
  1 5    1 8    5 8
  1 3    1 5    3 5
  2 4    2 3    4 3
  4 5    4 3    5 3   */
PairCorr = corr(X, "Pearson", "pairwise");
Eigenval = eigval(PairCorr);
print PairCorr[format=6.3], Eigenval;
```

PairCorr		
1.000	0.492	-0.717
0.492	1.000	0.419
-0.717	0.419	1.000

For the matrix of pairwise correlations, one eigenvalue is negative. This indicates that the matrix is not a valid correlation matrix. *There is no multivariate distribution for which this matrix represents the correlation between variables!*

Method 2: Single Imputation Methods

- **Single value imputation :** replacing the missing value with a single value utilizing one strategy such as : Mean , Median , Most Frequent , Mean Person , ... of the corresponding feature .

If there is a dataset that have great outliers, I'll prefer median. E.x.: 99% of household income is below 100, and 1% is above 500.

On the other hand, if we work with wear of clothes that customers give to dry-cleaner (assuming that dry-cleaners' operators fill this field intuitively), I'll fill missings with mean value of wear.

Mean/Median/Mode Imputation: For all observations that are non-missing, calculate the mean, median or mode of the observed values for that variable, and fill in the missing values with it. Context & spread of data are necessary pieces of information to determine which descriptor to use.

- Ok to use if missing data is less than 3%, otherwise introduces too much bias and artificially lowers variability of data

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

3. **Hot or Cold Deck Imputation**

"Hot Deck Imputation" : Find all the sample subjects who are similar on other variables, then *randomly* choose one of their values to fill in.

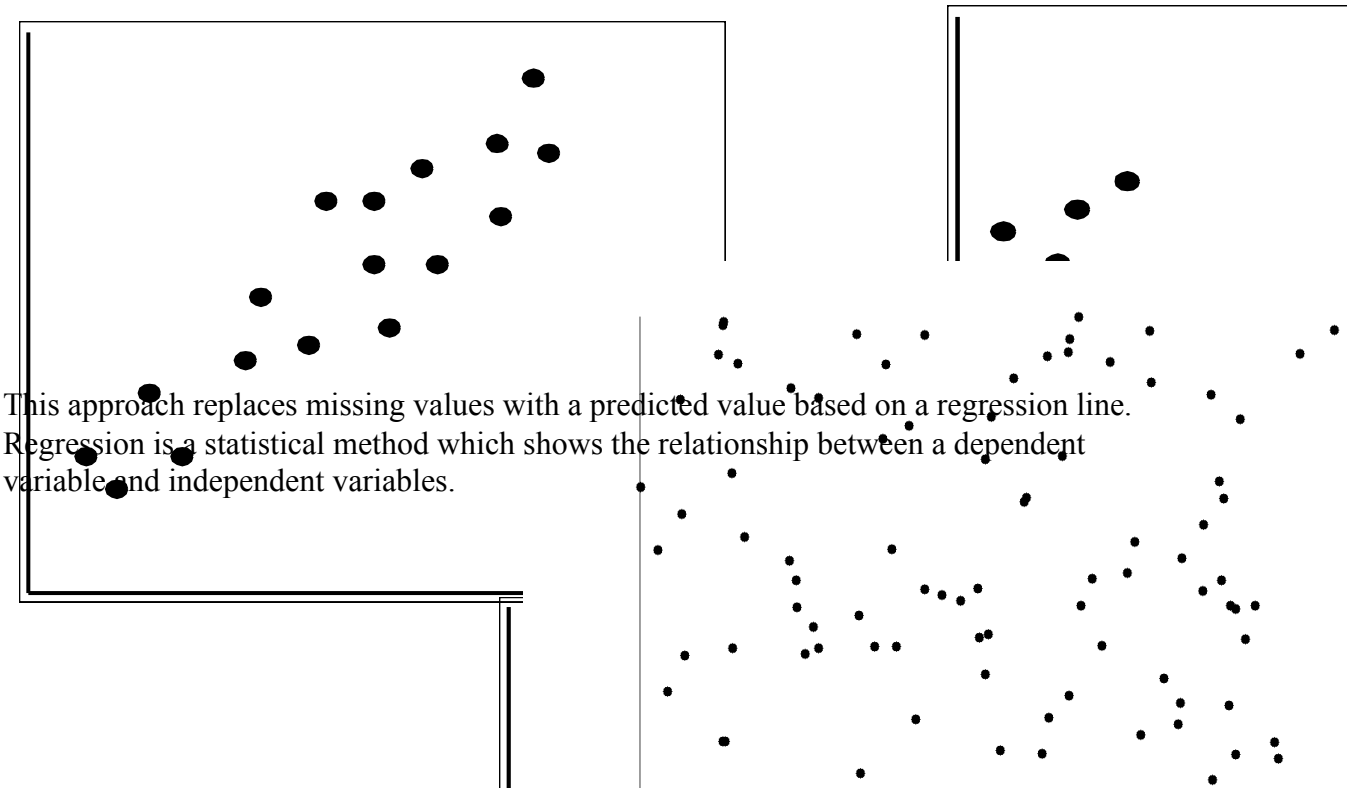
- Good because constrained by pre-existing values, but the randomness introduces hidden variability and is computationally expensive

"Cold Deck Imputation" : Systematically choose the value from an individual who has similar values on other variables (e.g. the third item of each collection). This option removes randomness of hot deck imputation.

- Positively constrained by pre-existing values, but the randomness introduces hidden variability and is computationally expensive
- **Similarity** : trying to find the closest (top-N closer) row(s) to the row containing our missing value , and fix a strategy among them to assign a value to our missing value .

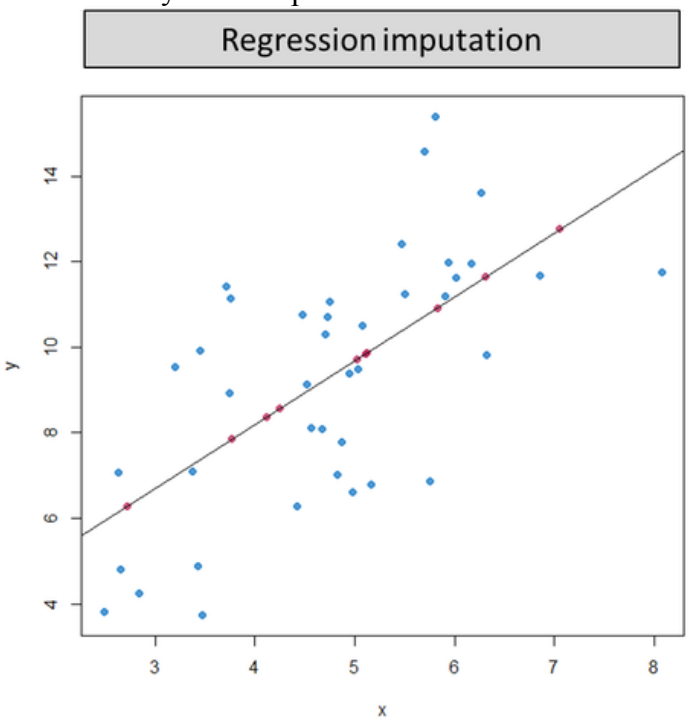
4. Regression Imputation :

- In single regression imputation the imputed value is predicted from a regression equation , we assume that the missing values are in a regression line with a nonzero slope with one of the complete features (predictors)
- Fill in with the predicted value obtained by regressing the missing variable on other variables; instead of just taking the mean, you're taking the *predicted value*, based on other variables.



- This approach replaces missing values with a predicted value based on a regression line.
- Regression is a statistical method which shows the relationship between a dependent variable and independent variables.

- It's expressed as $y = mx + b$ where m is the slope, b is a constant, x is the independent variable and y is the dependent variable.



- This method assumes that the imputed values fall directly on a regression line with a non-zero slope. As you can see, it is easy to comprehend and seems logical at the same time but it can affect the variability and the distribution of the data to some extent.

"Stochastic regression imputation" : The predicted value from a regression, *plus* a random residual value.

- This has all the advantages of regression imputation but adds in the advantages of the random component.

Challenges of Single value mutation:

- values found in single imputation might be biased by the specific values in the current data set, and
- not represent the total values of the full population.

Sol : Multiple Mutation

Multiple Imputation

Multiple imputation was a huge breakthrough in statistics about 20 years ago because it solved a lot of these problems with missing data (though, unfortunately not all). If done well, it leads to unbiased parameter estimates and accurate standard errors.

While single imputation gives us a *single* value for the missing observation's variable, multiple imputation gives us (you guessed it) *multiple* values for the missing observation's variable and then averages them for the final value.

To get each of these averages, a multiple imputation method would run analyses with 5–10 unique samples of the dataset and run the same predictive analysis on each**. The predicted value at that point would serve as the value for that run; the data signature of these samples change each time, which causes the prediction to be a bit different. The more times you do this, the less biased the outcome will be.

Once you take the mean of these values, it is important to analyze their spread. If they're clustering, they have a low standard deviation. If they're not, variability is high and may be a sign that the value prediction may be less reliable.

While this method is much more unbiased, it is also more complicated and requires more computational time and energy.

Numeric Data Imputation

Mean/Median

Imputing with mean/median is one of the most intuitive methods, and in some situations, it may also be the most effective. We basically take the average of the data, or we take the median of the data and replace all missing values with that value.

- Assumptions: Data is missing at random; Missing observations look like the majority of observations
- Advantages: Easy and quick to implement; Preserves the loss of data
- Disadvantages: Co-variance and variance may change; More the missing data, higher the distortion

Arbitrary Value Method

Here, the purpose is to flag missing values in the data set. You would impute the missing data with a fixed arbitrary value (a random value).

It is mostly used for categorical variables, but can also be used for numeric variables with arbitrary values such as 0, 999 or other similar combinations of numbers.

- Assumptions: Data is not missing at random
- Advantages: Quick and easy to implement; bring out underlying importance of missing values
- Disadvantages: Changes co-variance/variance; may create outliers

End of Tail Method

This method is similar to the arbitrary value method, however, the arbitrary value here is chosen at the tail-end of the underlying distribution of the variable.

If normally distributed, we use the mean +/- 3 times Standard Deviation.

If the distribution is skewed, use the IQR proximity rule. [Refer the Document of Data Distribution]

- Assumptions: Data is not missing at random; Data is skewed at the tail-end
- Advantages: Can bring out the importance of missing values;
- Disadvantages: Changes Co-variance/variance; may create biased data.

Categorical Data Imputation

Mode

As the name suggests, you impute missing data with the most frequently occurring value. This method would be best suited for categorical data, as missing values have the highest probability of being the most frequently occurring value.

- Assumptions: Data is missing at random; missing values look like majority
- Advantages: Quick and easy to implement; Suitable for categorical data
- Disadvantages: May create a biased data-set, favoring most frequent value

Add a Category for Missing Data

This next method is quite straightforward and only works for categorical data. You would create a separate label for missing values — 'missing' or it could be anything relevant. The idea is to flag missing values and understand the importance of being missing.

- Assumptions: No assumption
- Advantages: Quick and easy to implement; Helps understand importance of missing data
- Disadvantage: Potentially misunderstood data; Number of missing data should be large enough