Gmail by Google

Sung Moon <monspo2@gmail.com>

Imputing Missing data

Sung Moon <monspo2@gmail.com>
To: John Moon <monspo2@gmail.com>

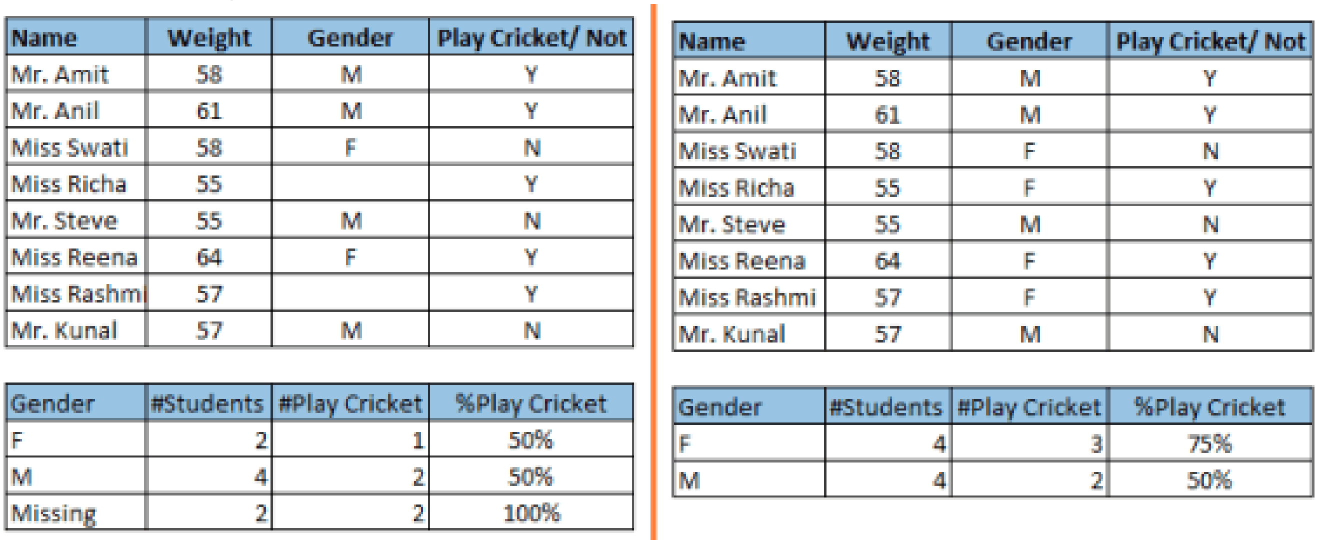
Wed, Mar 9, 2016 at 1:19 PM

7 Steps of Data Exploration & Preparation – Part 2

We will look at the methods of Missing values treatment. More importantly, we will also look at why missing values occur in our data and why treating them is necessary?

Why missing values treatment is required?

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analysed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.



Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Notice the missing values in the image shown above: In the left scenario, we have not treated missing values. The inference from this data set is that the chances of playing cricket by males is higher than females. On the other hand, if you look at the second table, which shows data after treatment of missing values (based on gender), we can see that females have higher chances of playing cricket compared to males.

Why my data has missing values?

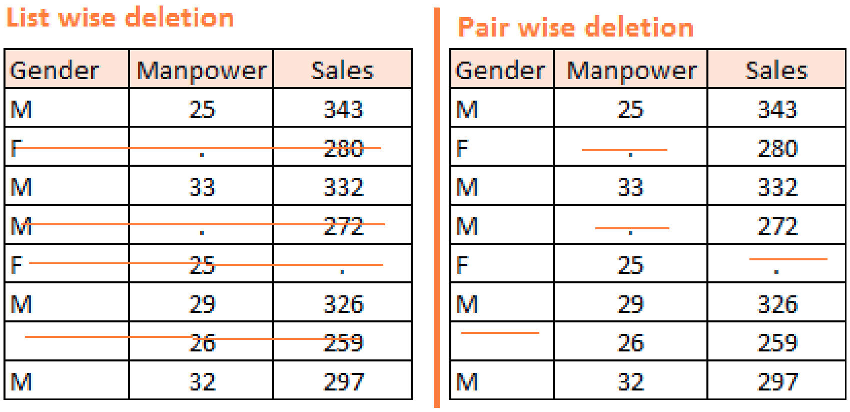
We looked at the importance of treatment of missing values in a dataset. Now, let's identify the reasons for occurrence of these missing values. They may occur at two stages:

1. **Data Extraction**: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are

typically easy to find and can be corrected easily as well.

2. **Data collection**: These errors occur at time of data collection and are harder to correct. They can be categorized in four types:
    - **Missing completely at random:** This is a case when the probability of missing variable is same for all observations. For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin. If an head occurs, respondent declares his / her earnings & vice versa. Here each observation has equal chance of missing value.
    - **Missing at random:** This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.
    - **Missing that depends on unobserved predictors:** This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop out from the study. This missing value is not at random unless we have included "discomfort" as an input variable for all patients.
    - **Missing that depends on the missing value itself:** This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

Methods to treat Missing Values

1. **Deletion:** It is of two types: List Wise Deletion and Pair Wise Deletion.
    - In list wise deletion, we delete observations where any of the variable is missing. Simplicity is one of the major advantage of this method, but this method reduces the power of model because it reduces the sample size.
    - In pair wise deletion, we perform analysis with all cases in which the variables of interest are present. Advantage of this method is, it keeps as many cases available for analysis. One of the disadvantage of this method, it uses different sample size for different variables.
    - 

List wise deletion

Gender	Manpower	Sales
M	25	343
~~F~~	~~.~~	~~280~~
M	33	332
~~M~~	~~.~~	~~272~~
~~F~~	~~25~~	~~.~~
M	29	326
	~~26~~	~~259~~
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	~~.~~	280
M	33	332
M	~~.~~	272
F	25	~~.~~
M	29	326
	26	259
M	32	297

    - Deletion methods are used when the nature of missing data is "**Missing completely at random**" else non random missing values can bias the model output.
2. **Mean/ Mode/ Median Imputation:** Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing

data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable. It can be of two types:-

- **Generalized Imputation:** In this case, we calculate the mean or median for all non missing values of that variable then replace missing value with mean or median. Like in above table, variable "**Manpower"** is missing so we take average of all non missing values of "**Manpower"** (**28.33**) and then replace missing value with it.
- **Similar case Imputation:** In this case, we calculate average for gender "**Male"** (29.75) and "**Female**" (25) individually of non missing values then replace the missing value based on gender. For "**Male**", we will replace missing values of manpower with 29.75 and for "**Female**" with 25.

3. **Prediction Model**: Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and variable with missing values is treated as target variable. Next, we create a model to predict target variable based on other attributes of the training data set and populate missing values of test data set.We can use regression, ANOVA, Logistic regression and various modeling technique to perform this. There are 2 drawbacks for this approach:
    - The model estimated values are usually more well-behaved than the true values
    - If there are no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

4. **KNN Imputation:** In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function. It is also known to have certain advantage & disadvantages.
    - **Advantages:**
        - k-nearest neighbour can predict both qualitative & quantitative attributes
        - Creation of predictive model for each attribute with missing data is not required
        - Attributes with multiple missing values can be easily treated
        - Correlation structure of the data is taken into consideration

- **Disadvantage:**
    - KNN algorithm is very time-consuming in analyzing large database. It searches through all the dataset looking for the most similar instances.
    - Choice of k-value is very critical. Higher value of k would include attributes which are significantly different from what we need whereas lower value of k implies missing out of significant attributes.

End note

In this article, we have discussed the importance of treating missing values in a dataset, followed by the reason which causes the missing values to prevail in a dataset. We have also looked at the basics of treating missing values. The treatment method can be hybrid approach based on above methods or more complex. In next series of data exploration, we will look at the outlier treatment.

Have you ever come across dealing with similar situation? We are excited to hear your experience & learning talks. Talk with us in the comments section below.