19 November 2018
16:51

Missing data can have a severe impact on building predictive models because the missing values might be contain some vital information which could help in making better predictions. So, it becomes imperative to carry out missing data imputation. There are different methods to treat missing values based on the problem and the data. Some of the common techniques are as follows:

1. **Deletion of rows**: In train dataset, observations having missing values in any variable are deleted. The downside of this method is the loss of information and drop in prediction power of model.
2. **Mean/Median/Mode Imputation**: In case of continuous variable, missing values can be replaced with mean or median of all known values of that variable. For categorical variables, we can use mode of the given values to replace the missing values.

3. **Building Prediction Model**: We can even make a predictive model to impute missing data in a variable. Here we will treat the variable having missing data as the target variable and the other variables as predictors. We will divide our data into 2 datasets

   one without any missing value for that variable and the other with missing values for that variable. The former set would be used as training set to build the predictive model and it would then be applied to the latter set to predict the missing values.

You can try the following code to quickly find missing values in a variable.

```
sum(is.na(combi$Item_Weight))
```
[1] 2439

Imputing Missing Value

As you can see above, we have missing values in Item_Weight and Item_Outlet_Sales. Missing data in Item_Outlet_Sales can be ignored since they belong to the test dataset. We'll now impute *Item_Weight* with mean weight based on the *Item_Identifier*variable.

```
missing_index = which(is.na(combi$Item_Weight))
for(i in missing_index){

 item = combi$Item_Identifier[i]
 combi$Item_Weight[i] = mean(combi$Item_Weight[combi$Item_Identifier == item], na.rm = T)
}
```
Now let's see if there is still any missing data in *Item_Weight*

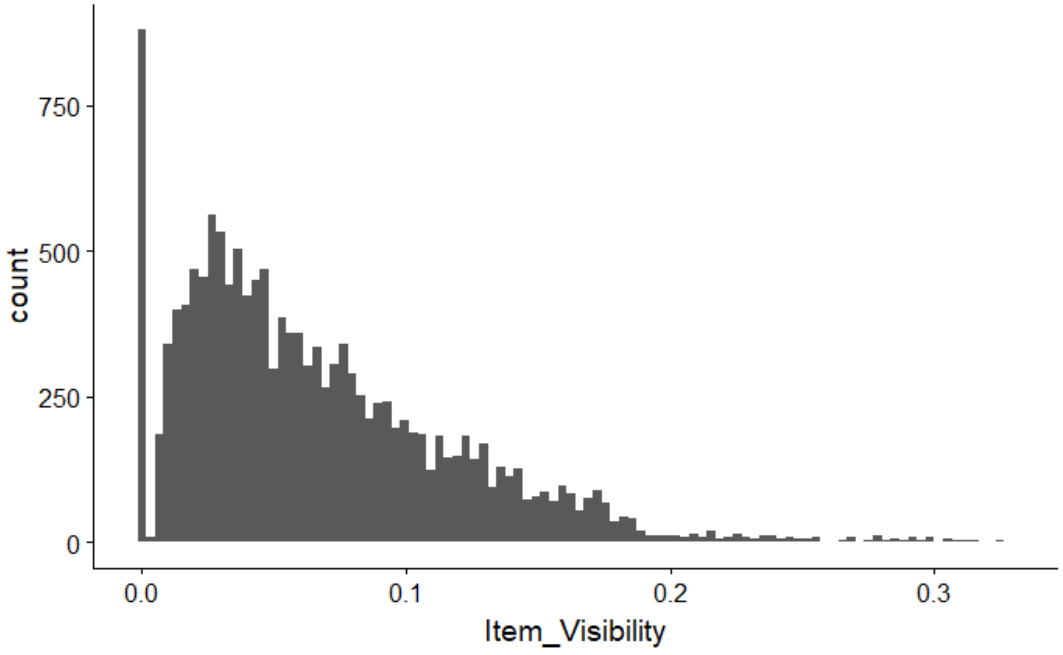
```
sum(is.na(combi$Item_Weight))
```
[1] 0

0 missing values! It means we have successfully imputed the missing data in the feature.

Replacing 0's in Item_Visibility variable

Similarly, zeroes in Item_Visibility variable can be replaced with *Item_Identifier* wise mean values of *Item_Visibility*. It can be visualized in the plot below.

```
ggplot(combi) + geom_histogram(aes(Item_Visibility), bins = 100)
```



Let's replace the zeroes.

```
zero_index = which(combi$Item_Visibility == 0)
for(i in zero_index){

 item = combi$Item_Identifier[i]
 combi$Item_Visibility[i] = mean(combi$Item_Visibility[combi$Item_Identifier == item], na.rm = T)

}
```

After the replacement of zeroes, We'll plot the histogram of *Item_Visibility* again. In the histogram, we can see that the issue of zero item visibility has been resolved.

```
ggplot(combi) + geom_histogram(aes(Item_Visibility), bins = 100)
```