**Missing Value Exercise**

**Tasks to be performed**

1 How to handle missing values in the data.  
2. When we need to discard missing value.  
3. What are different missing value filling techniques.  
4. When to use which technique.

**1 Handling missing values**

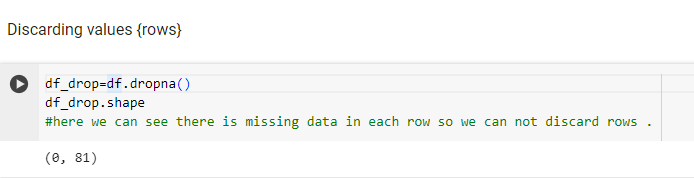
There are 3 ways to handle missing values

* 1. Removing Missing values
  2. Imputing Missing values
  3. Use an algorithm which can handle missing values

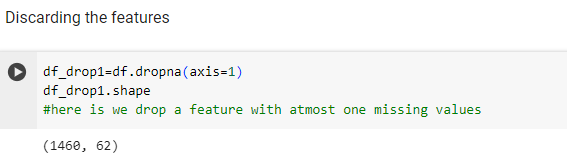
**2 When we need to discard values**

This method cause data loss

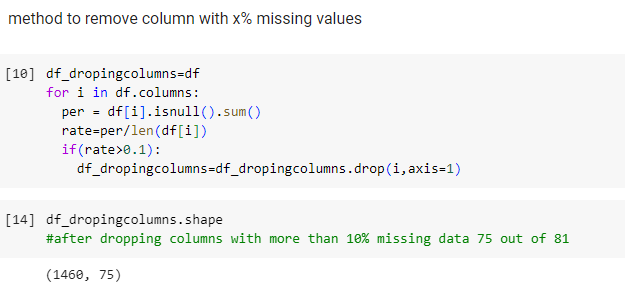
* 1. we can remove the rows when there is any missing value in rows



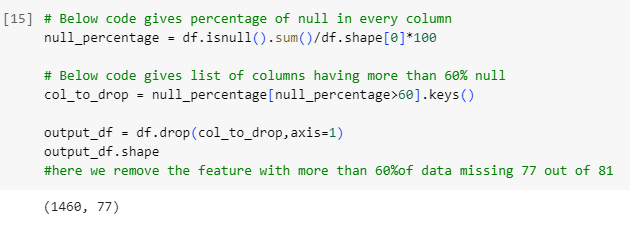
* 1. we can remove columns when there are many missing values in column



here removing values depend on data loss and dependency on result, when our feature is not correlated to any other feature in that case we remove the column{when there is more than 50% data missing in the column}.



We can discard missing values when we have large dataset and missing values are very few.



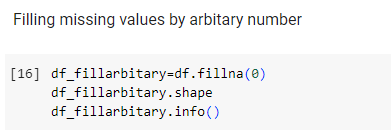
**3. Different missing value filling techniques**

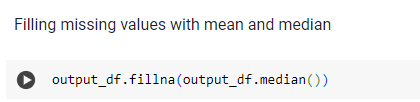
There are multiple missing values filling techniques …

1. Arbitrary/Mean/Median imputation
2. Mode imputation
3. Forward/backward imputation
4. Deep Learning – KNN,regression.

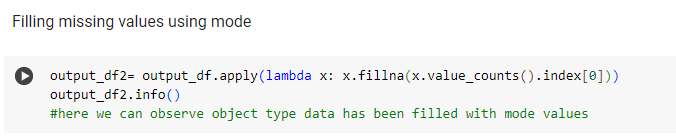
**4 When to use which technique**

**1 Arbitary/mean/median :** Use mean imputation for normally distributed data and median imputation for data with outliers. This method is simple and suitable for numerical data with missing values. It can not be used for categorical data.



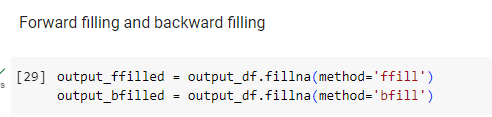


1. **Mode Imputation:** Use mode imputation for categorical data with missing values. This method replaces missing values with the most frequent value in the column.



1. **Forward Fill/Backward Fill:** Use forward fill or backward fill for time-series data or ordered

data where the missing values can be reasonably assumed to be close to the non-missing values. While using this method we need to make sure the data distribution is not hampered.



4. **K-Nearest Neighbors (KNN) Imputation:** Use KNN imputation when the missing values are related to the values of other features in the dataset. This method works well for datasets with complex relationships between features. Can be used for MCAR , MAR and MNAR.

**Sources used**

**Kaggle : dataset has been picked from kaggle.**

**Github / GFG : used for filling data**

**Colab : program is done on colab .**

Google Colab Notebook link: [click Here](https://colab.research.google.com/drive/1WjxF_dhQy6Gqj9aVr9JWr0NNHyUBHBd5?usp=sharing)