Data Mining Assignment – 1

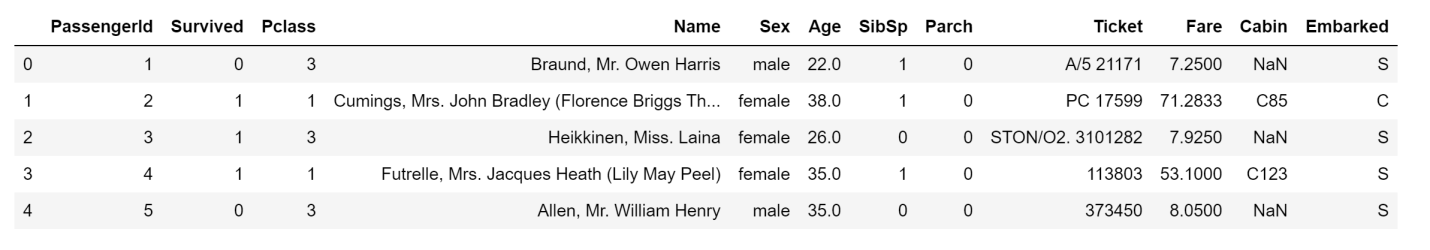
Data Imputation Analysis Report

**Name: H Subhashini**

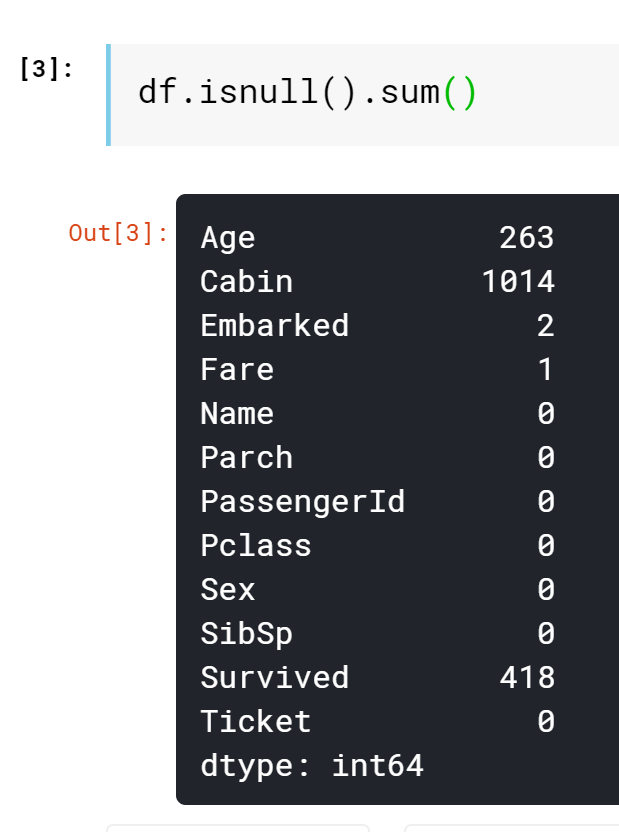
**Roll num: 16PT36**

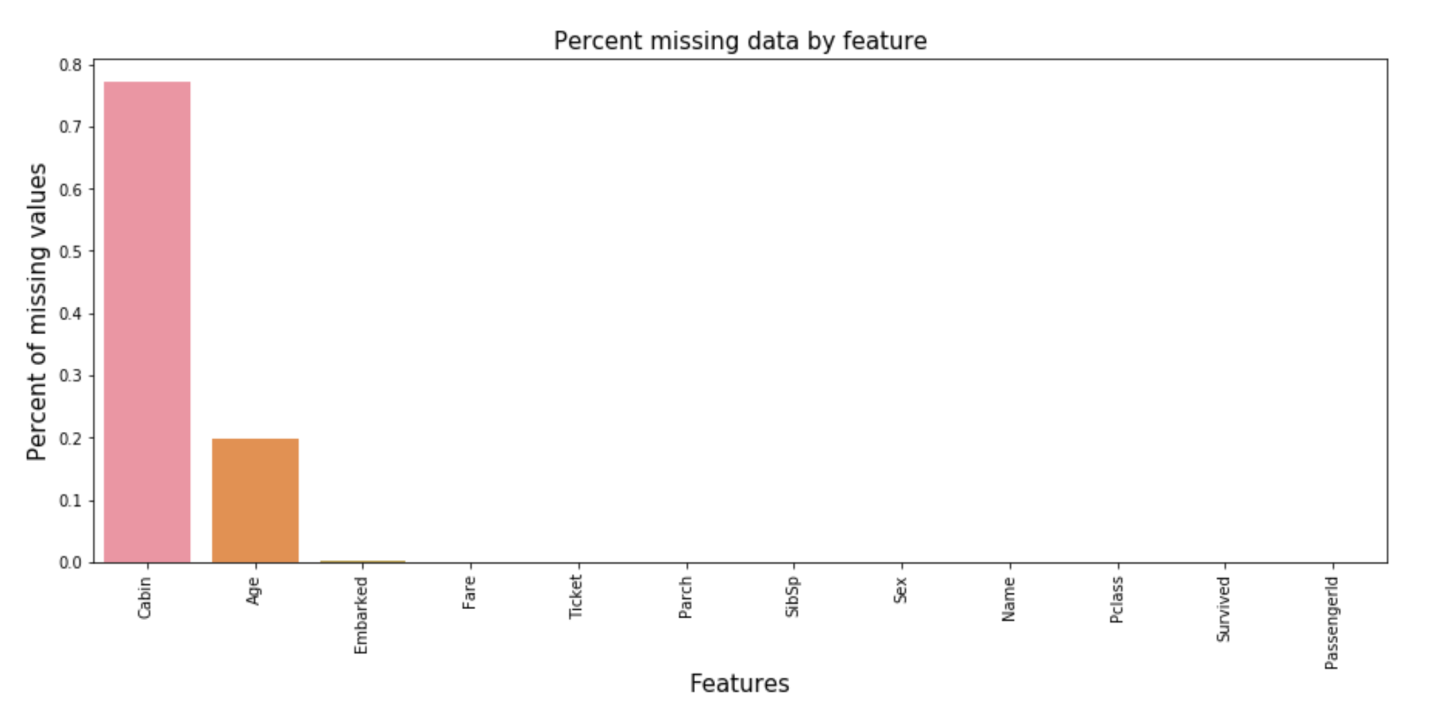
Dataset considered is **Titanic.**

Sample data



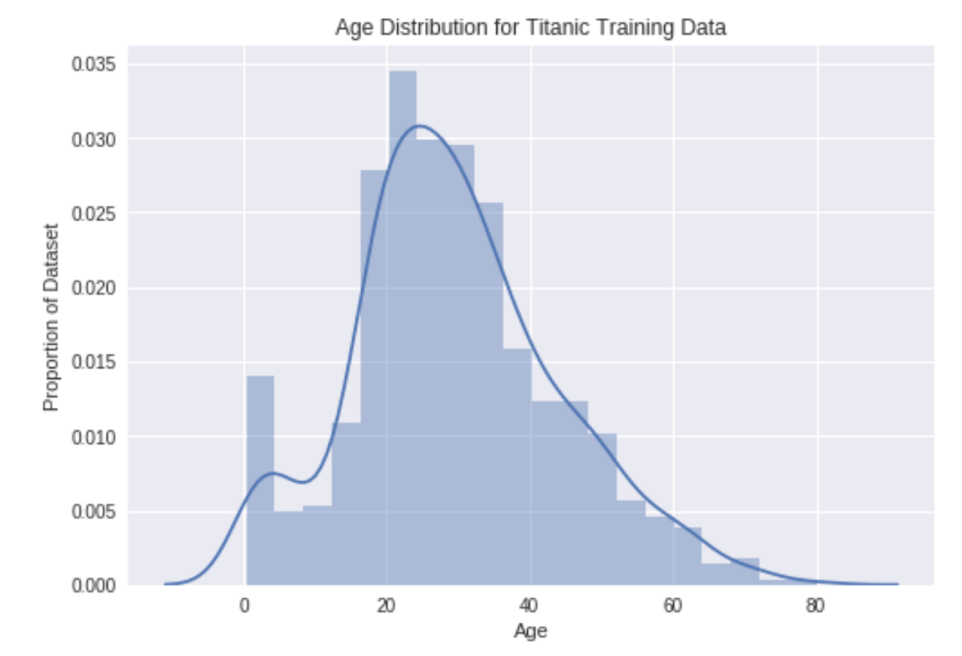
From the below figure, it is clear that the fields **Age (numerical)**, **Cabin (categorical)** and **embarked (categorical)** has missing values.





The above figure shows the missing value in percentage. Around 75% of “Cabin” is missing, 20% of “Age” is missing and less than 1% of “Embarked” is missing.

**Distribution of Age:**

****

We see a somewhat bimodal distribution with one peak for children and one peak for adults.

**Variance of Age** before imputation is **207.55**.

**Impute missing age values:**

**Mean/Median Imputation:**

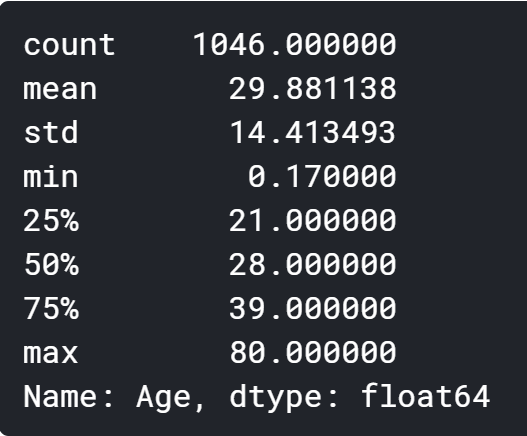
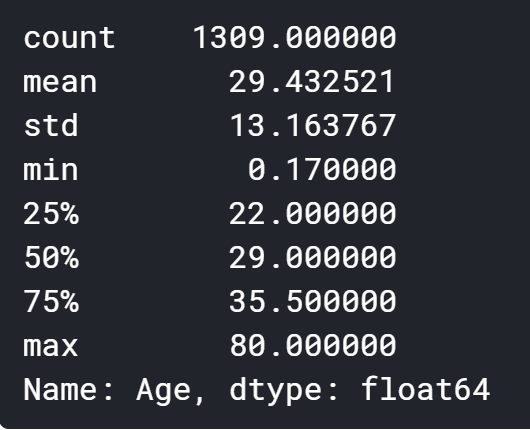
Age had 263 missing values.

From the name we have extracted the title of every individual, and when there is a missing value, **median** of the persons with the same title is replaced in place of nan. It is clear that the initial distribution and distribution after imputation is same.

**Variance of Age** after median imputation is **173.94**.

Here, since Age is normally distributed, both mean and median are approximately the same.

If **data is skewed**, **median** is a better representation.

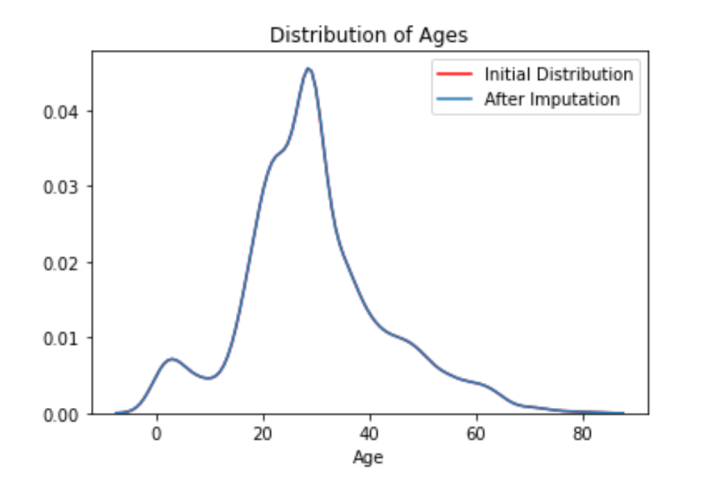
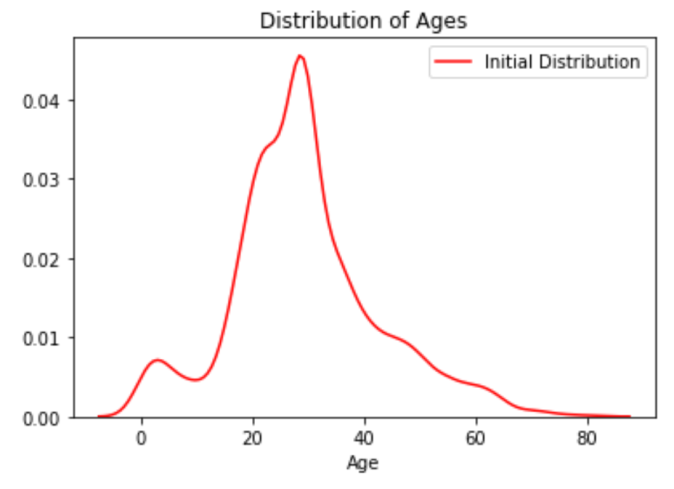
After imputation

Before imputation

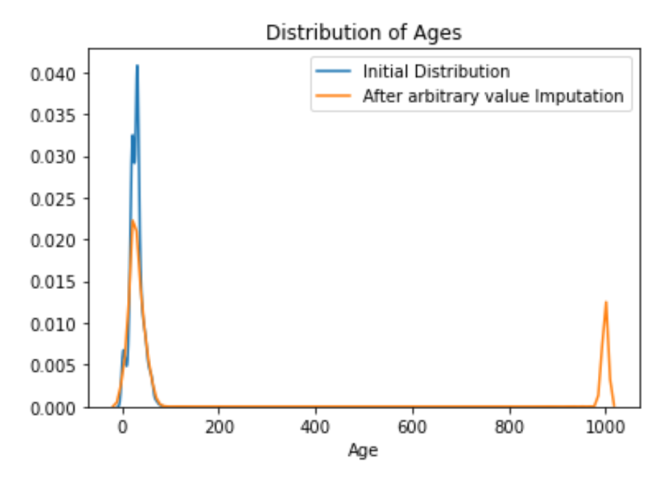
MEAN: Suitable for continuous data without outliers

MEDIAN : Suitable for continuous data with outliers

This mean/median imputation performs good for age feature since there is not significant difference in distribution.



**Arbitrary value Imputation:**

****

Arbitrary value imputation consists of replacing all occurrences of missing values (NA) within a variable with an arbitrary value. The arbitrary value is different from the mean or median and not within the normal values of the variable. **Variance of Age** after arbitrary imputation is 150952.27.

We can use arbitrary values such as 0, 999, -999 (or other combinations of 9s) or -1 (if the distribution is positive). Assumption is data is not missing at random.

This method is suitable for numerical and categorical variables.

We observe that the distribution is **not** the same before and after imputation.

**End of tail Imputation:**

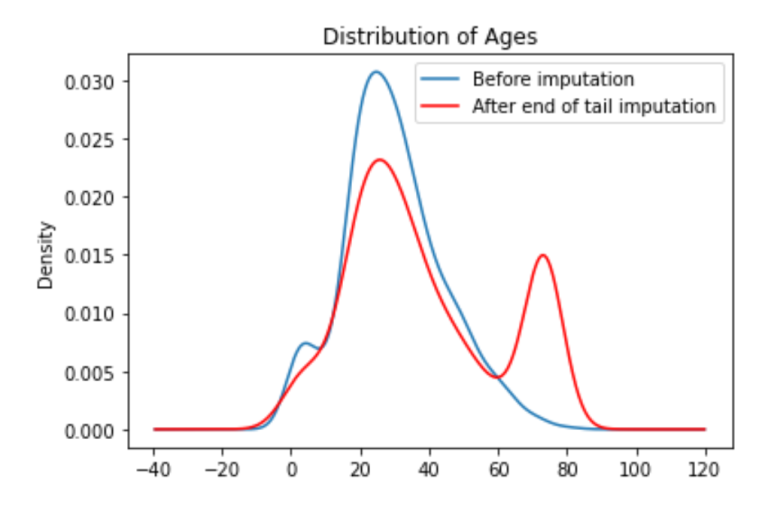
End of tail imputation is roughly equivalent to arbitrary value imputation, but it automatically selects the arbitrary values at the end of the variable distributions.

* If the variable follows a normal distribution, we can use the mean plus or minus 3 times the standard deviation.
* If the variable is skewed, we can use the IQR proximity rule.

Approximately 20% of data was missing in Age feature.

**Variance before imputation: 207.55**

**Variance after end of tail imputation: 471.19**

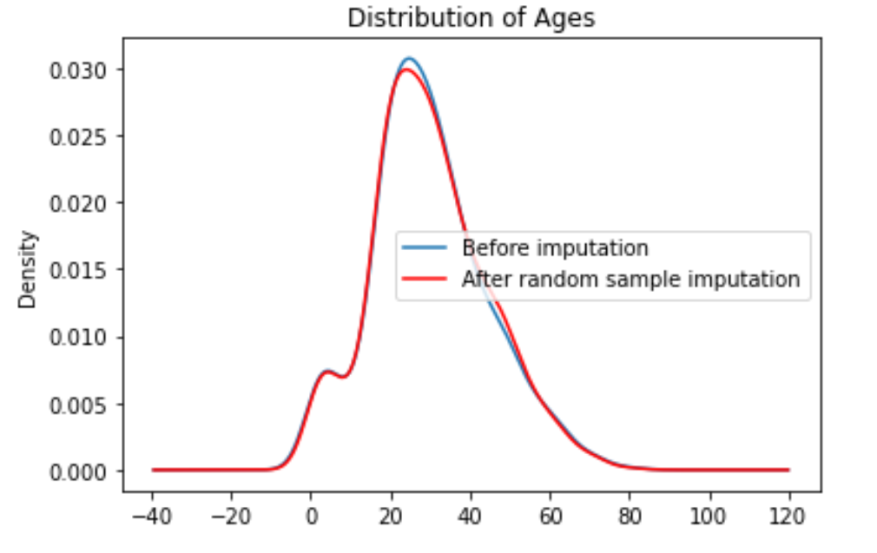


**Random Sample Imputation:**

Random sampling consists of taking a random observation from the pool of available observations of the variable and using those randomly selected values to fill in the missing ones.

This method is suitable for numerical and categorical variables.

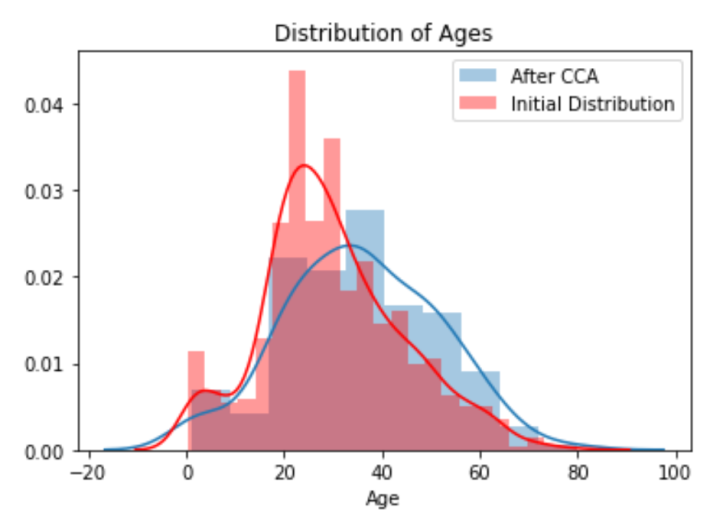
**Variance after Random Sample Imputation:** 209.67



We observe that the distribution is **almost the same** before and after imputation.

**Complete Case Analysis:**

Complete case analysis (CCA) is a technique that consists of discarding observations where values in any of the variables are missing.

****

We observe that the distribution is not the same before and after imputation.

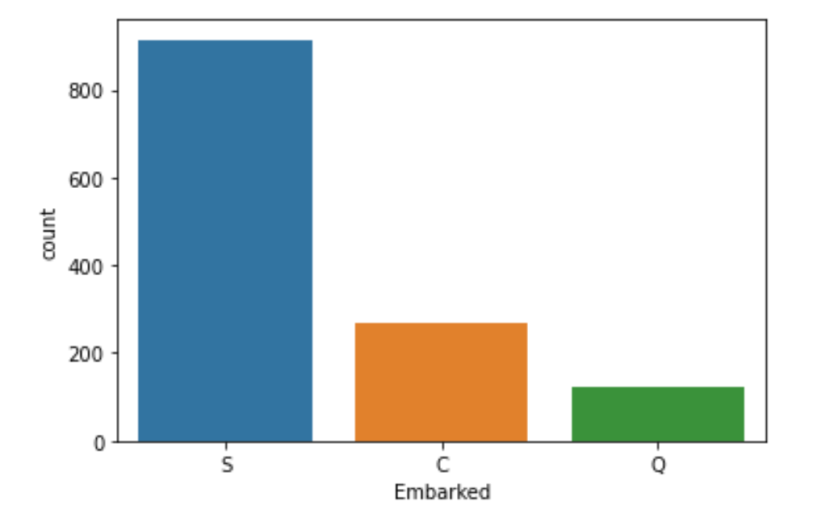
**Variance after Complete Case Analysis: 243.39**

**Impute missing embarked values:**

In this case, the change after imputation is not big, and it wouldn’t affect the data much since the missing entries were less than 1%.

**Frequent Category Imputation:**

Frequent category imputation—or **mode imputation**—consists of replacing all occurrences of missing values (NA) within a variable with the mode, or the most frequent value.



This method is suitable for numerical and categorical variables, but in practice, we use this technique with categorical variables. As we saw that **maximum passengers boarded from Port S**, we replace NaN with **S**.

**Missing Category Imputation:**

This method is the most widely used method of missing data imputation for categorical variables.

we replace NaN with **Missing**.

**Observation and Inference:**

For the **feature ‘Age’** considered, **mean/median imputation and random sample imputation** works better than the other imputations considered.

For the feature ‘Embarked’ considered, there are only 2 missing values. Generally For the Categorical variable, imputing with the mode is generally used (**Frequent Category Imputation**). But if there is a class imbalance in the variable, imputing with mode will increase the problem further. So in such cases, we can use a classification model to predict the missing classes.