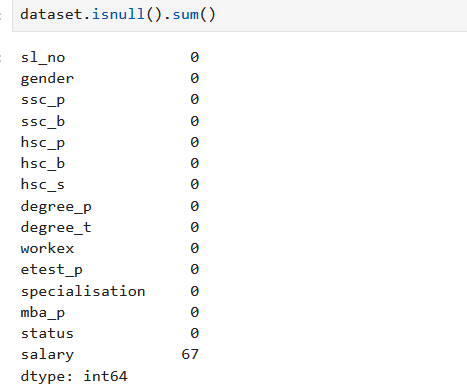
# Data Science

## Univariate Analysis

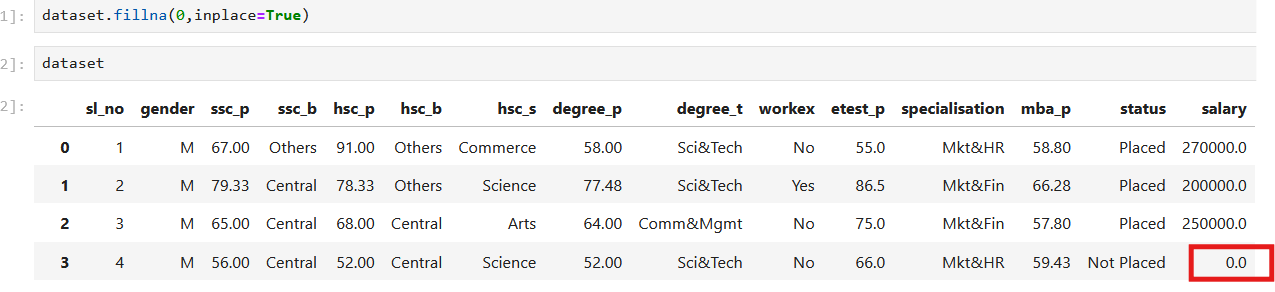
### Observation and Explanation

1. Finding if there is null values (empty cells) in the dataset



* .isnull() function of pandas shows Nan(null values) of the dataset
* Here the salary column has 67 null values

1. Replacing the Nan with zero or mean/median/mode respective to the dataset and removing data.

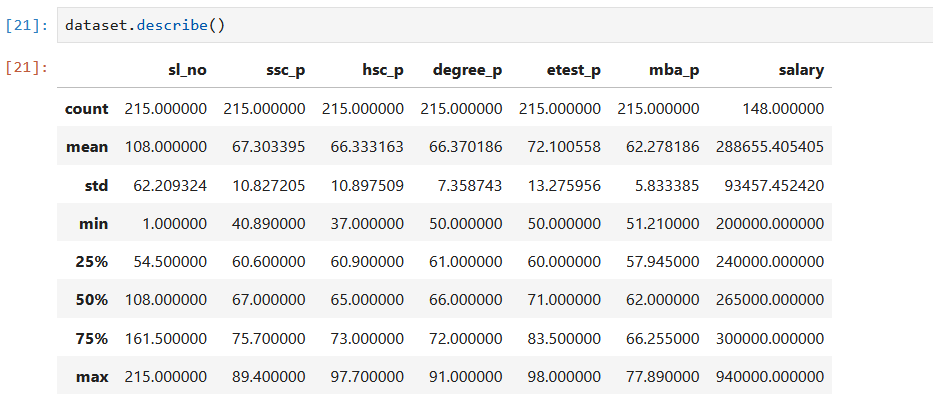


* .fillna() function helps replacing the NaN cells with 0, it can also be replaced with mean/median/mode.

Python: dataset([“column name here”].fillna([“column name here”].mean(),inplace=true)

* But this function can replace values at only one column at a time
* .drop(0,inplace=True), this function deletes row or column with null values where 0 refers to row and 1 refers to column

# Measure of Central Tendency:



* .describe() a function under pandas that gives mean (MCT) , count, Standard Deviation(MOS),percentile(Q1,Q2,Q3,Q4)(MLD), IQR, Outliers, min and max value of the data
* Here the average (mean) “of ssc\_p” is 67.3
* But we need mean, median,mode of the dataset , so we create a dataframe using pandas and for loop to calculate these values from the dataset and bring it under the dataframe



* Formulas:

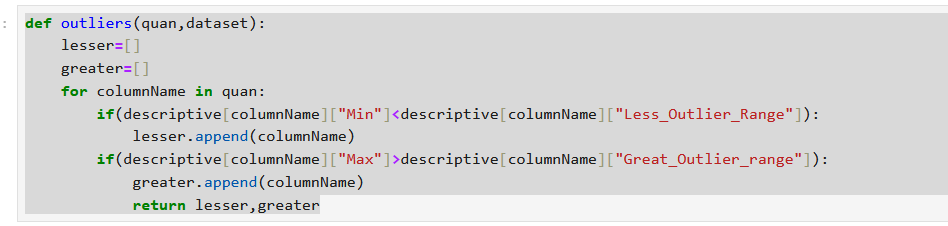
IQR=Q3-Q1

1.5 Rule= 1.5\*IQR

Less Outlier Range=Q1-1.5 Rule

Great Outlier Range=Q3+1.5 Rule

# Out Lier Detection



Less outlier exist if:

* Minimun value of a column is less than the lesser outlier range
* Means that a certain data point is far away from the mean in leftside

Greater outlier exist if:

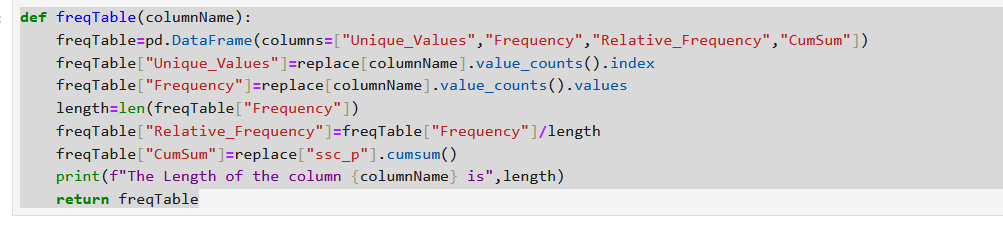
* Maximun value of a column id greater than greater outlier range
* Means that a certain data point is far away from the mean in rightside.

# Outlier Replacement

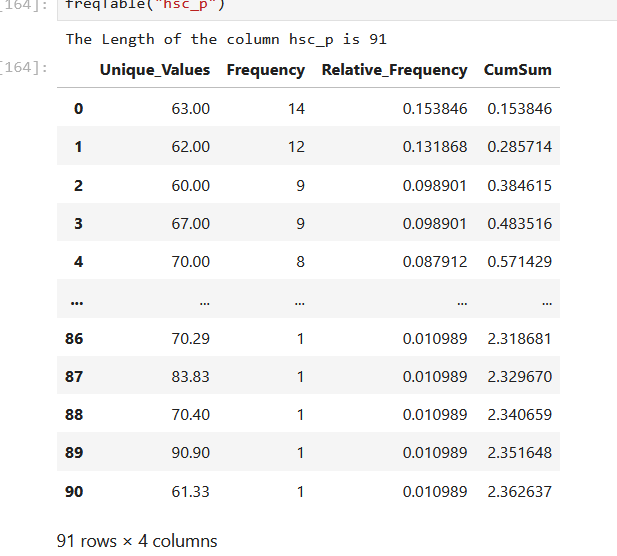
# 

* Replacing outlier with less and greater outlier range calculated from IQR and 1.5 rule

# Measure of Location of Data

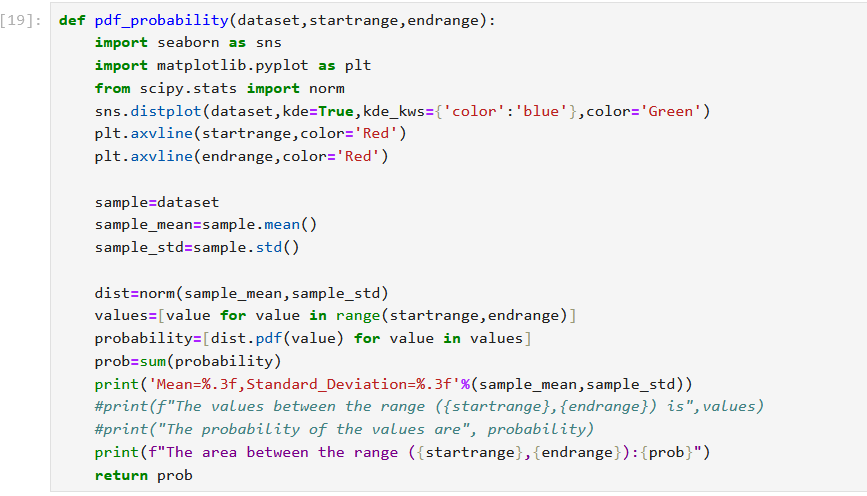


* Python to get frequency, Relative frequency, cumulative frequency
* .value\_counts gives the repeated values (frequency) and no of repetations.
* .cumsum gives cumulative density.

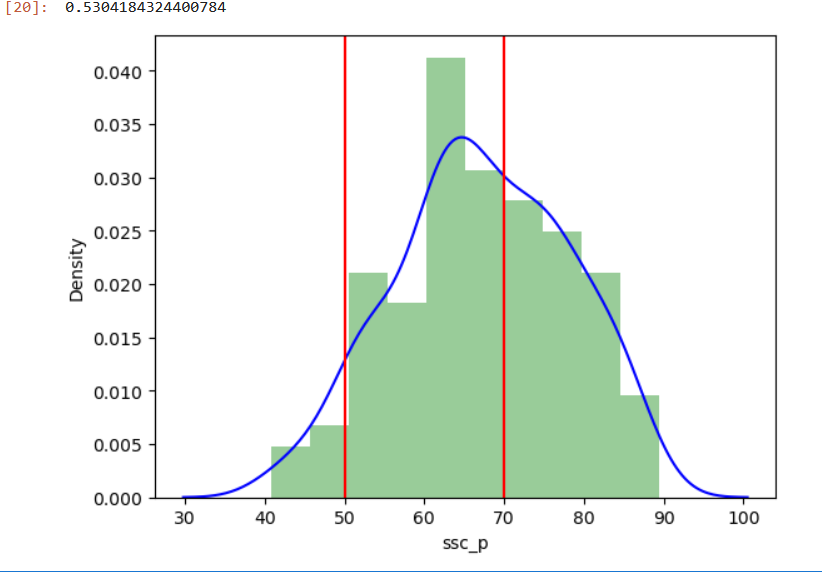


* Frequency table for column “hsc\_p”
* Here 63 repeated 14 times and it’s the highest repeated value.
* Relative Frequency: how often a value has occurred 14/91=0.153845, where 14 occurrence of 63 and 91 total no of values
* Cum Sum : sum of values under the range

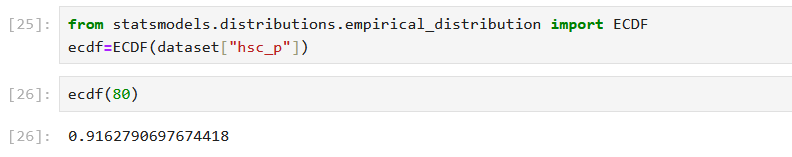
# Normal Distribution



* Seaborn to graph histogram with curve using KDE (Kernel Density Estimation) with given blue color to the curve and green color to the histogram(plotting of frequency).
* Matplotlib.pyplot to bring start and end range line with given red color.
* Since Normal Distribution is characterised by mean and standard deviation, so calculated using .mean() and .std() function
* Then probability of values between the range and cumsum is calculated and brought inside the list using one line for loop.

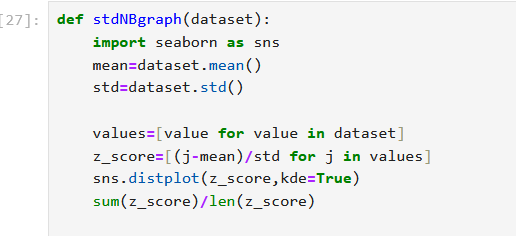


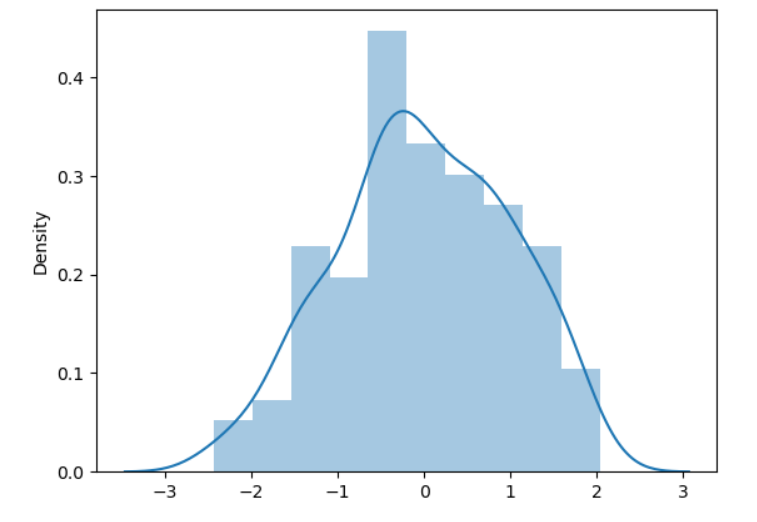
# Empirical Rule



* Percentage of occurrence data at a range

# Standard Normal Distribution





* stdNB is measured using Z-Score and helpful to compare two different type of data (eg. Height in cm and feet).
* Z=(X-mean)/Standard Deviation.
* And plots the x-axis using standard deviation and the y-axis has the density.