**Fall 2023 DATA 220 Mathematical Methods for Data Analytics**

**Homework – 1**

**Name :- Prayag Nikul Purani**

**SJSU Id :- 017416737**

**Problem 5 - Data Analysis using Titanic Dataset (Coding):-**

The important step for the rest of the homework is the importing the data so it can be done using pandas.

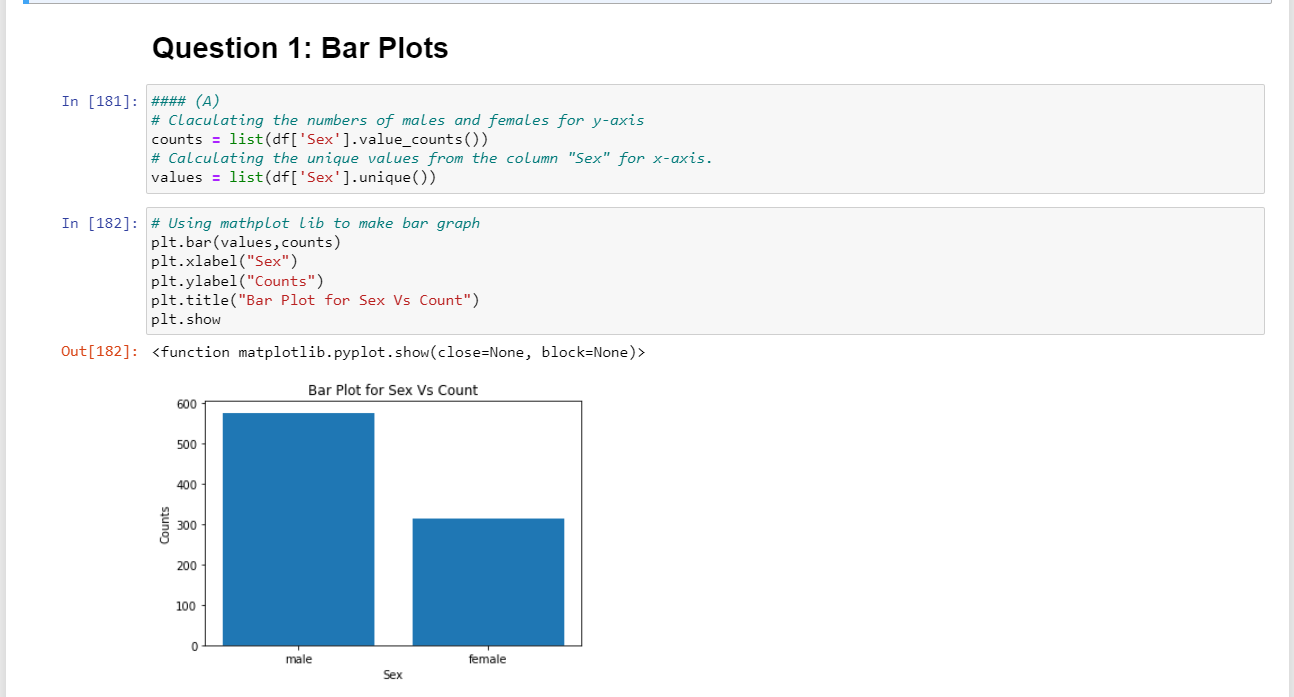
****

**Fig 1 :- Importing csv using pandas**

**Question 1:**

So, bar plot is the best representation for the visualization of the different entities which provides the knowledge of the data in quick access and the users or the viewer can easily get what is the relation between values and can get the idea of the graph or the distribution of data is going to be

1. First bar chat shows the distribution of gender, so on x-axis it will have the unique values in the column which is male and female and for y-axis it will have the counts. This bar chart will tell us how many numbers of male and female where there in the database.



**Fig 2 :- Bar chart for distribution of gender**

1. Second bar chart show the number of passengers boarded from each stop with the ship starting point as Southampton and so on. So, the graph will have the stops on the x-axis and the count of people on y-axis.

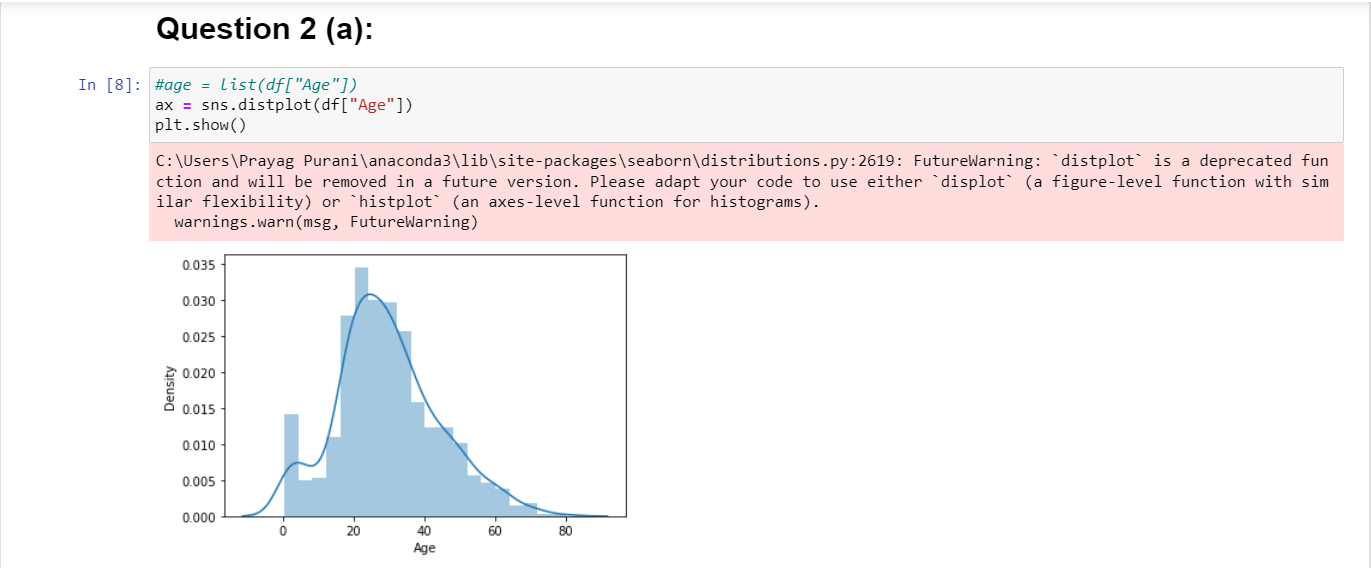


**Fig 3 :- Bar chart for the distribution of Embarked from different stops**

**Question 2:-**

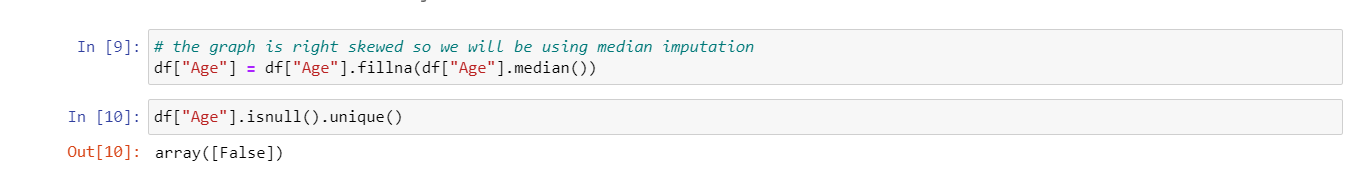
The main aim of the question is to remove the outlines form the data so they will not affect the calculation and out dataset is reduced. Outlier detection is a process of identifying the data points in a dataset that lie far away from the majority of the data. These data points can either be significantly higher or lower than the other observations and can impact the overall results of the analysis. Second thing to do in this data set is to remove null values to make the data-frame more cleaned and can be used for better visualization with efficient reading and values inputs for better predication.

1. The "Age" variable has missing values. Based on distribution (symmetric / non -symmetric), perform an appropriate missing value imputation technique to fill in the missing values. So, to know what’s the pattern for the age column we need to make a histogram and then we if the graph is symmetric then we have to use mean imputation and if the graph is skewed right or left then we need to use median imputation method to fill the null values in the age column.



**Fig 4 :- Histogram for the visualization of graph pattern**

So, from the histogram we can see that the data is right skewed so we have to apply median imputation technique to fill the missing values.

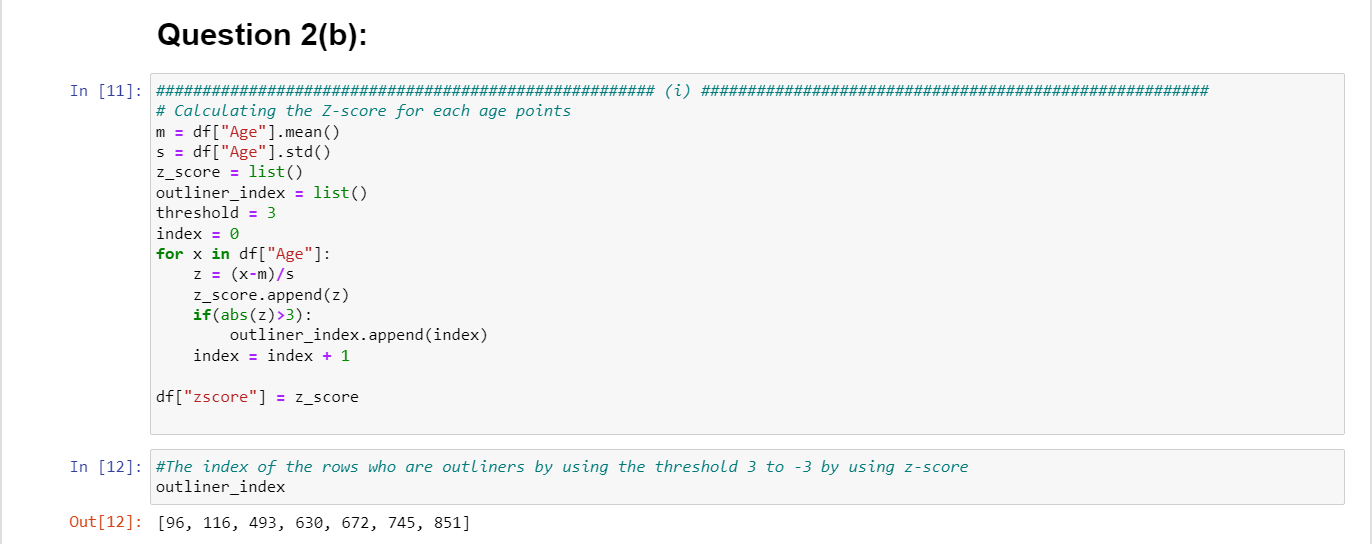


**Fig 5 :- Meian imputation**

So, after filling the age’s nan with the median we can check that if it still has any nan left so the second line of code does the same thing and the result says the unique value in Age after Boolean function of isnull() is only false this means no null values are left in the data.

1. Now the detection of outliners can be done by three different methods and we can’t use any inbuilt libraries of python.
2. z-score:-

We are going to apply this formula z = (x-μ)/σ to find z-score in of the age column. The threshold specified in the question is between 3 to -3. So, after calculation we will print the index of outliners so to understand the number of outliners present in the data set according to the Z score method.



**Fig 6 :- Z-score**

1. Modified z-score :-

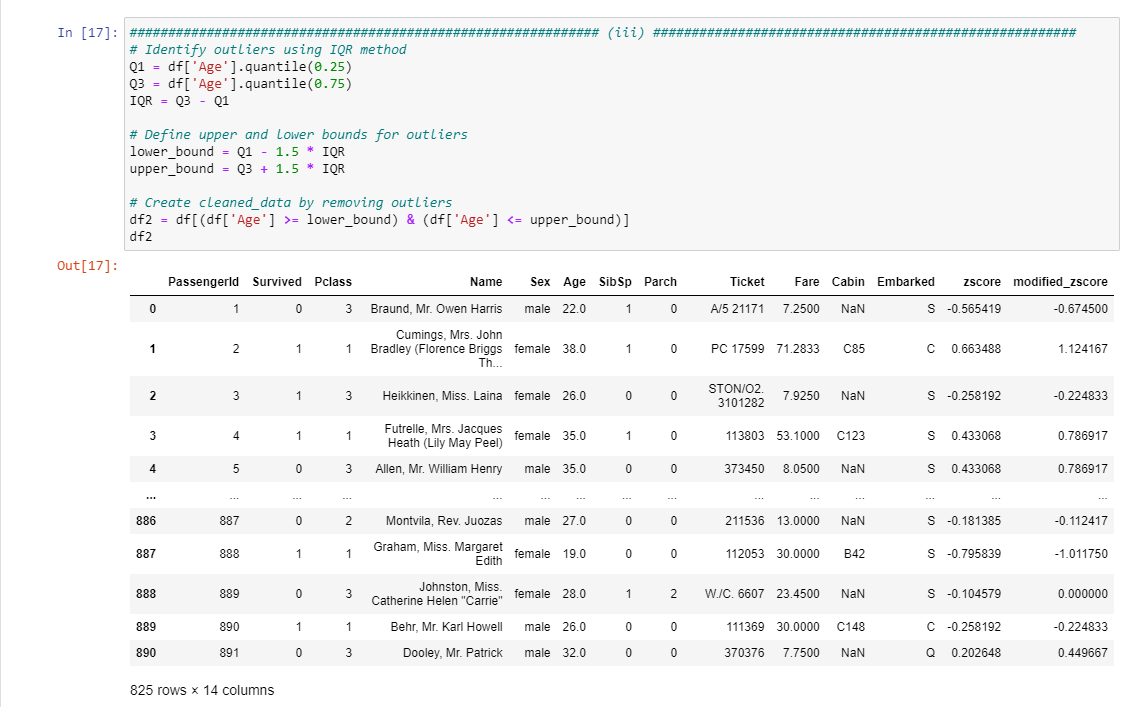
So, the formula which we will be using is 0.6745(xi – x̃) / MAD



**Fig 6 :- Modified Z-score**

1. Inter-Quartile range (IQR):-

The interquartile range (IQR) is a measure of statistical dispersion, or the spread of data. It is the range of values within which the middle 50% of scores reside. The IQR is defined as the difference between the 75th and 25th percentiles of the data.



**Fig 7 :- Interquartile range**

So this is the new dataframe which is cleaned according to IQR and we can see the number of row is reduced from 892 to 825 and this dataset will be used in next of the question to solve them.

**Question 3:-**

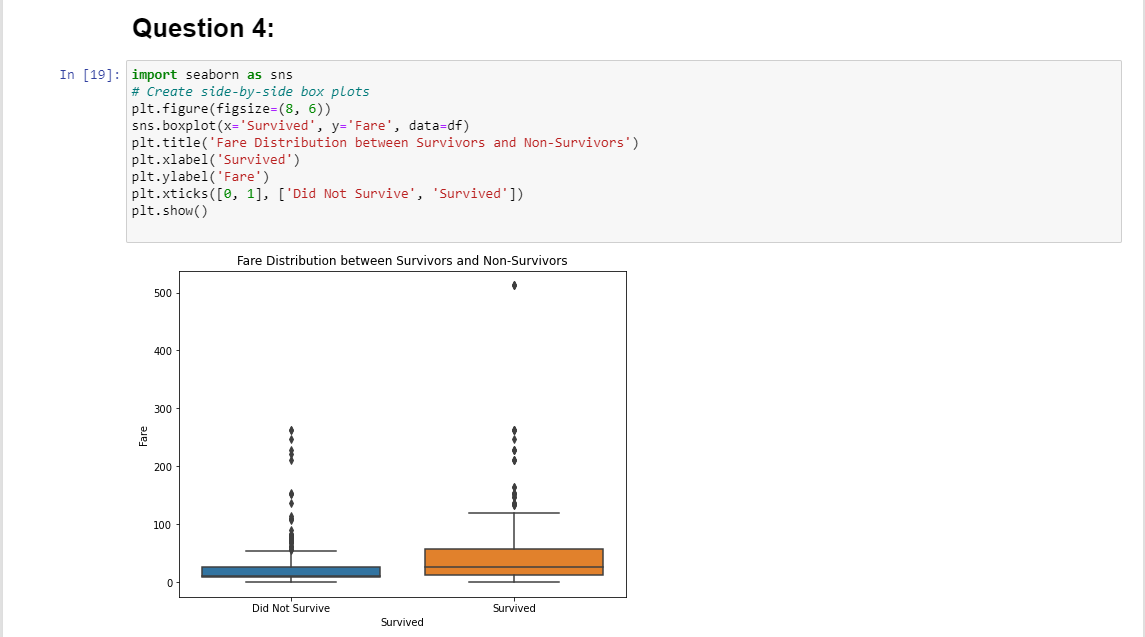
Based on the IQR method – remove all the observations that contain outlier for Age variable and create a new dataset: cleaned\_data.



**Fig 8 :- Cleaned - data**

**Question 4:-**

A box plot, also known as a box-and-whisker plot, is a data visualization technique used to represent the distribution of a set of continuous or numerical data. The box plot displays the median, quartiles, and outliers of the data, allowing for a quick and easy assessment of the spread and skewness of the data. The box is drawn from the lower quartile to the upper quartile, and the median is represented as a line inside the box. Whiskers are drawn from either end of the box to the minimum and maximum data points, excluding outliers. Outliers are plotted as individual points outside the whiskers. Box plots are commonly used in exploratory data analysis and hypothesis testing.



**Fig 9 :- Box plot**

Five number summery:

Minimum :- So the minimum value in both the graphs are nearly the same.

Q1 :- The values of q1 are also same in both the graphs

Median/Q2 :- The value of Q2 in didn’t survive is nearly overlapping with its Q1 but

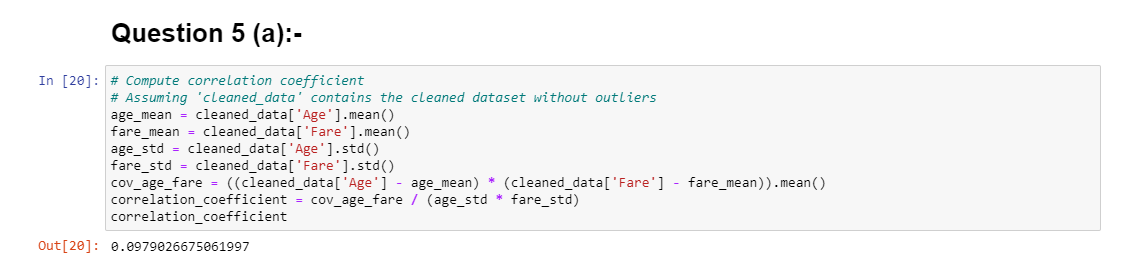
the Q2 of the survived is distant and is defiantly ore then Q2 of the other graph.

Q3 :- Q3 of didn’t survive is less then survived.

Maximum :- The max value of survived is more than the value of max in didn’t survived.

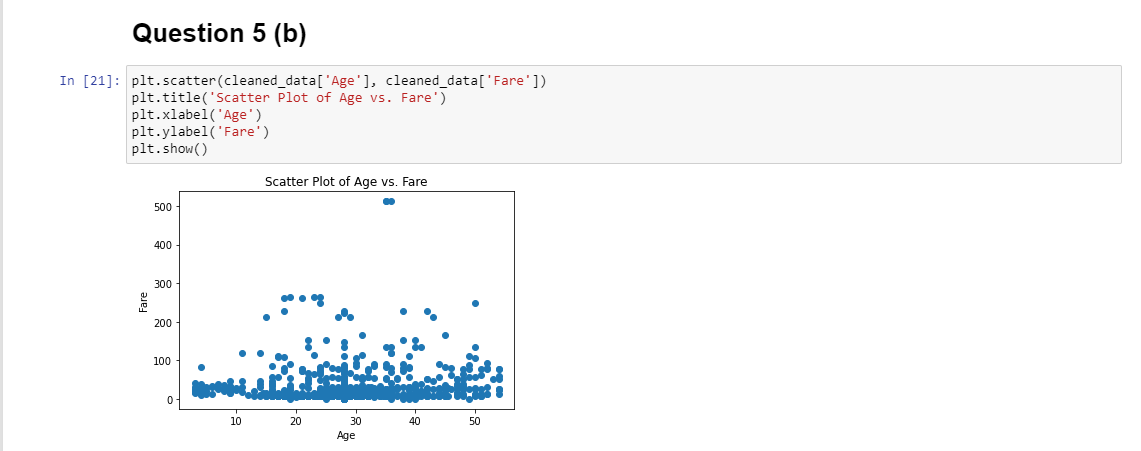
**Question 5 :-**

1. Compute the correlation coefficient between the "Age" and "Fare" variables in the titanic dataset.



**Fig 10 :- Correlation Coefficient**

1. Use a scatter plot to visually inspect the relationship between the two variables: Age and Fare. Interpret the strength and direction of the relationship based on the computed correlation coefficient.



**Fig 11 :- Scatter plot**

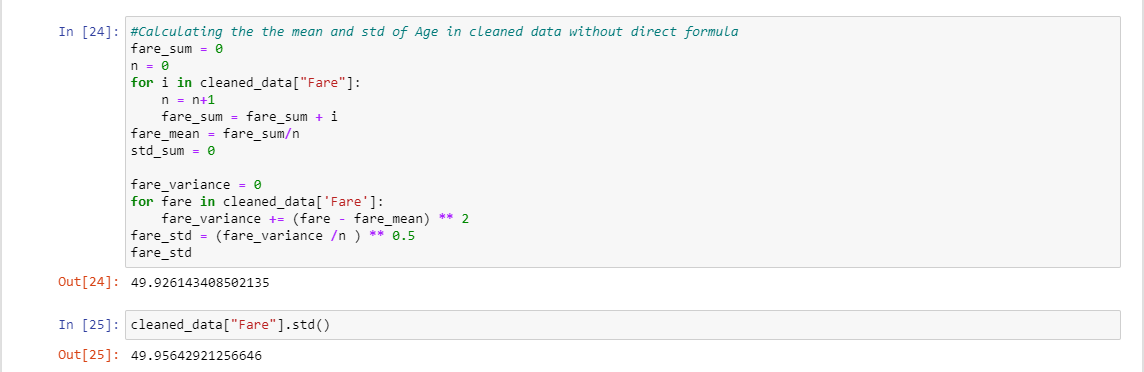
The scatter plot visually shows how "Age" and "Fare" are related. You can interpret the direction (positive or negative) and the strength of the relationship based on the correlation coefficient. Positive values indicate a positive correlation, negative values indicate a negative correlation, and the magnitude indicates the strength.

So, the correlation is nearly 0.097 so it is positive but the correlation is not so strong as it near to the 0 and the graph also tells the same.

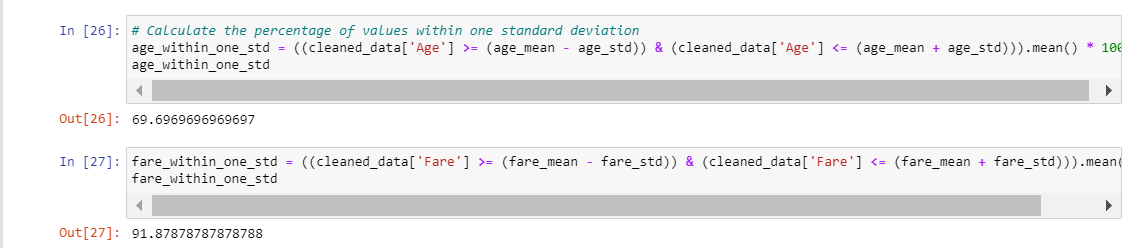
**Question 6 :-**



**Fig 12 :- Age mean and standard deviation**



**Fig 13 :- Fare mean and standard deviation**



**Fig 14 :- Percentage of data between standard deviation**

**References :-**

<https://saturncloud.io/blog/how-to-use-python-pandas-to-get-unique-values-ignoring-nan/>

<https://www.geeksforgeeks.org/bar-plot-in-matplotlib/>

<https://www.w3schools.com/python/matplotlib_histograms.asp>

<https://www.w3schools.com/python/matplotlib_line.asp>

<https://pythonbasics.org/seaborn-distplot/>

<https://seaborn.pydata.org/generated/seaborn.distplot.html>

<https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/>

<https://www.geeksforgeeks.org/find-median-of-list-in-python/>

<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html>