

**AMERICAN INTERNATIONAL**

**UNIVERSITY-BANGLADESH**

**Faculty of Science and Technology**

**Project Report Cover Page**

|  |  |  |
| --- | --- | --- |
| Assignment Title: | Mid Term Project: Titanic Dataset Preparation | |
| Assignment No: | 01 Date of Submission: July 17,2023 | |
| Course Title: | INTRODUCTION TO DATA SCIENCE[B] | |
| Course Code: | 02399 | Section: B |
| Semester: Summer |  | Course Teacher: TOHEDUL ISLAM |

**Declaration and Statement of Authorship:**

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student’s work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

*\* Student(s) must complete all details except the faculty use part.*

\*\* Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.

:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** |  | **Name** | **ID** | **Program** | **Signature** |
| 1 | NAHIUN,AHNAF HASNAIN |  | 20-42359-1 | BSc [CSE] |  |
| 2 |  |  |  | **Choose an item.** |  |
| 3 |  |  |  | **Choose an item.** |  |
| 4 |  |  |  | **Choose an item.** |  |
| 5 |  |  |  | **Choose an item.** |  |
| 6 |  |  |  | **Choose an item.** |  |
| 7 |  |  |  | **Choose an item.** |  |
| 8 |  |  |  | **Choose an item.** |  |
| 9 |  |  |  | **Choose an item.** |  |
| 10 |  |  |  | **Choose an item.** |  |

|  |  |  |
| --- | --- | --- |
| ***Faculty use only*** | | |
| FACULTYCOMMENTS | **Marks Obtained** |  |
|  |
|  |
|  | **Total Marks** |  |
|  |
|  |

**Project Overview :-**

The Titanic Dataset contains 301 instances with 10 attributes. This dataset provides information on various factor such as gender,age,class and survival status.Each instance represents a passenger details and attributes describe different factor those are related with titanic tragedy. All this information making it excellent resource for analyzing patterns and drawing.

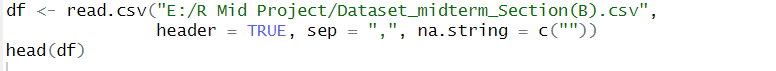
The ten attributes are given below:

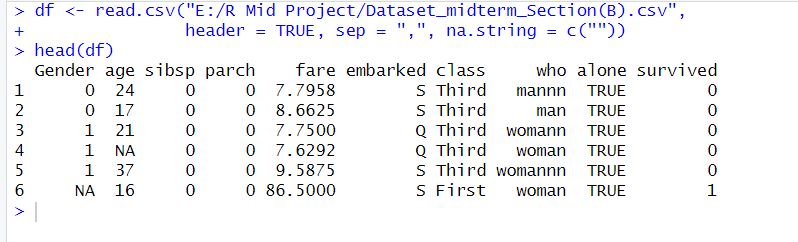
* Gender: the gender of passengers(integer)
* age: the age of the passengers(numeric)
* sibsp: sibling of passengers (integer)
* parch: Parents or children aboard with passenger (integer)
* fare: Each passenger fare (numeric)
* embarked: Port of embarkation (Char)
* calss: ticket class (Char)
* who: categories to passengers (char)
* alone: passenger was alone in ship or no (logi)
* survived: passenger survived or not (integer)

**Data Preparation :**

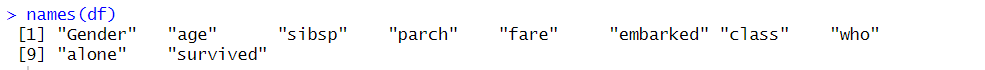
**# Data Exploration :**

Load the Data



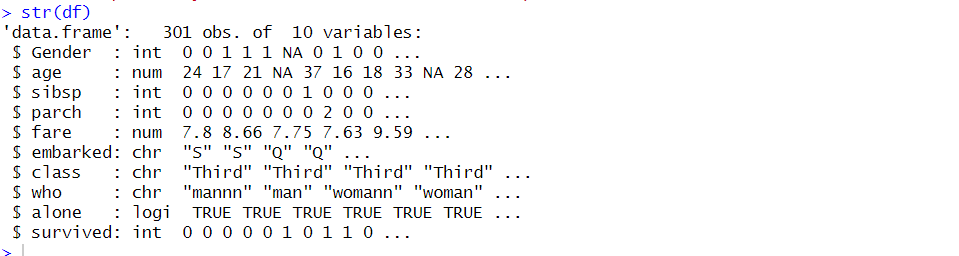


View column Name

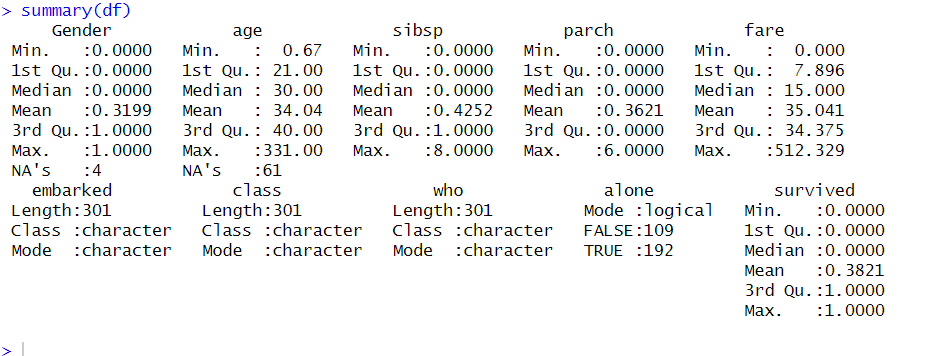


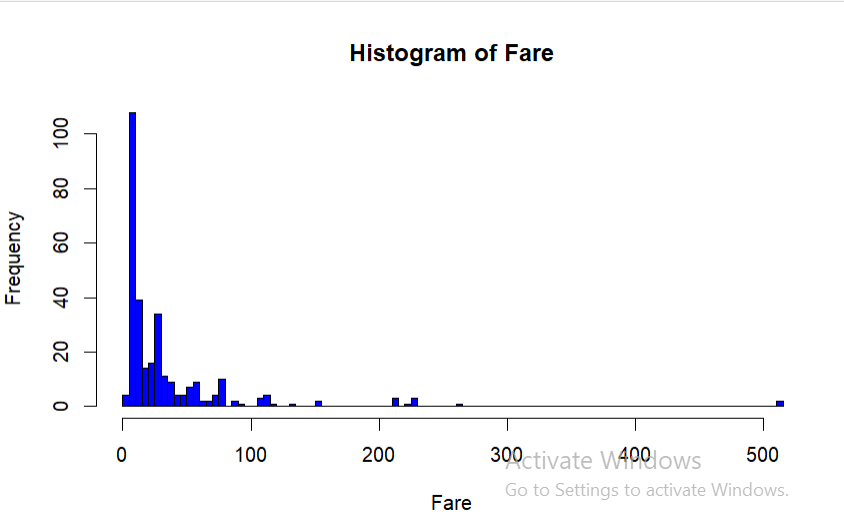
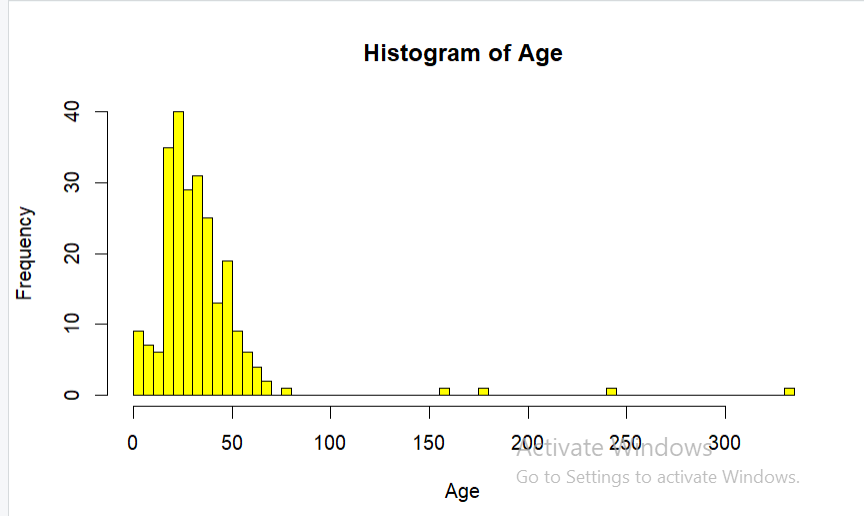
Summary of the data

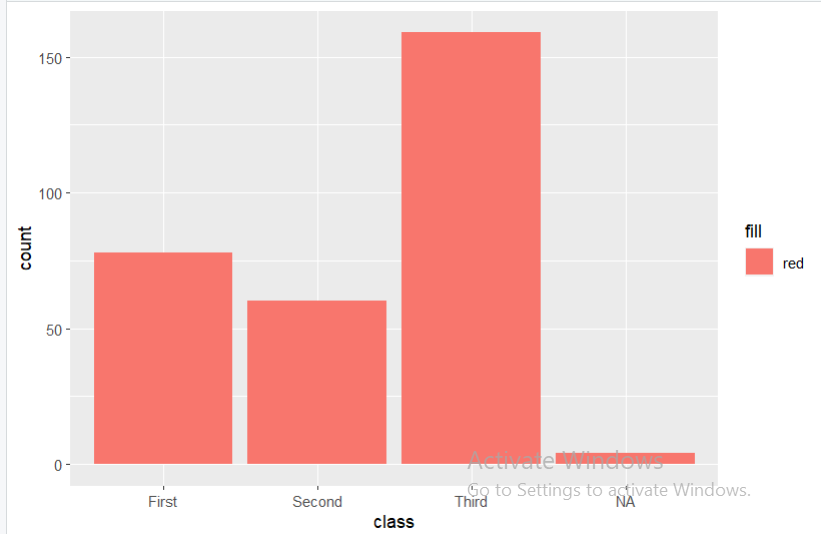




Descriptive Statistics



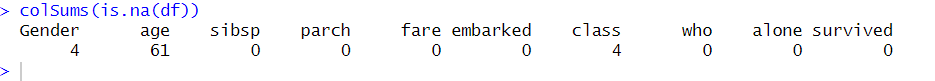




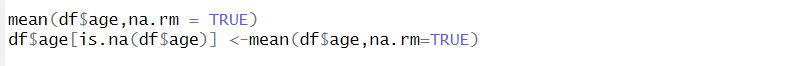
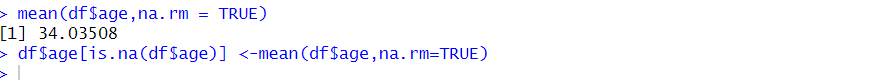
**# Missing and Duplicate Values**

Counting number of NULL values in each column

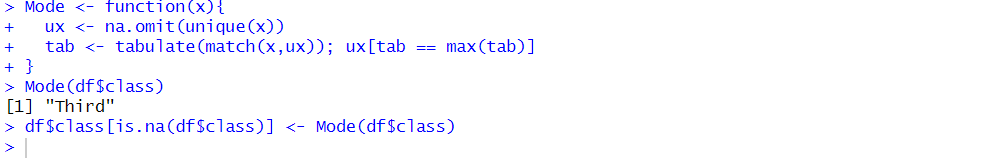
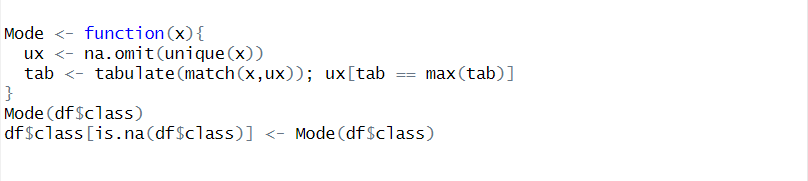




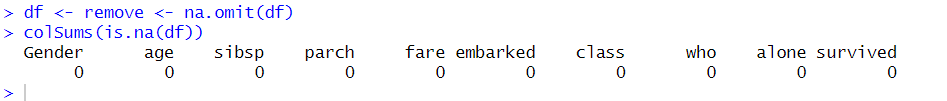
Null Value replace by Mean value



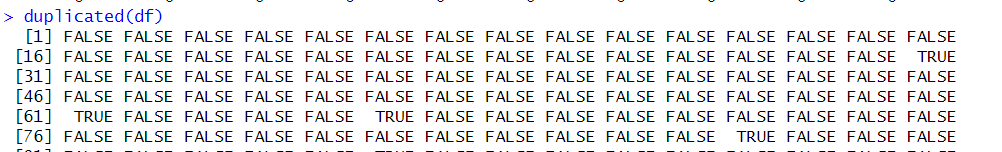
Null Value replace by Mode value



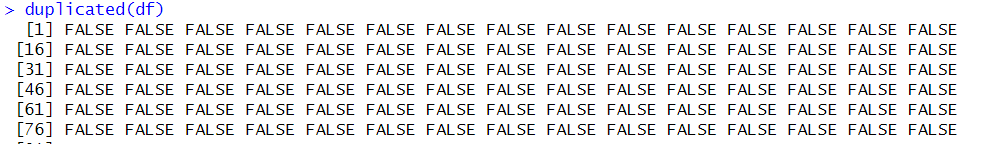
Removing Null and Missing Values



Cheacking Duplicates value

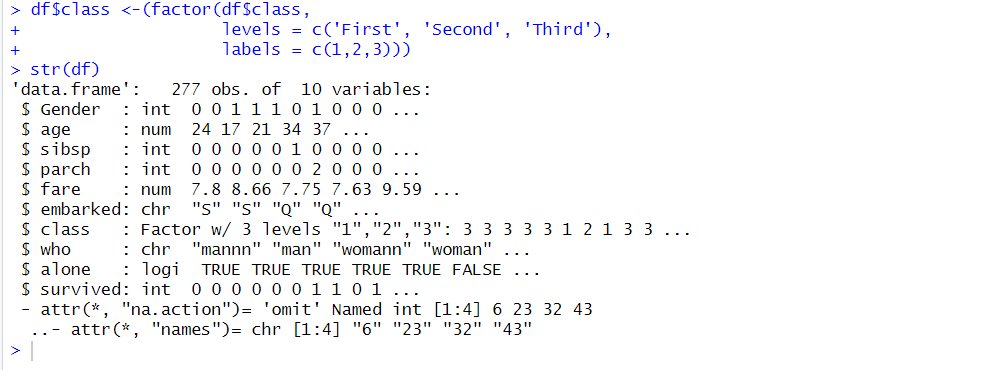
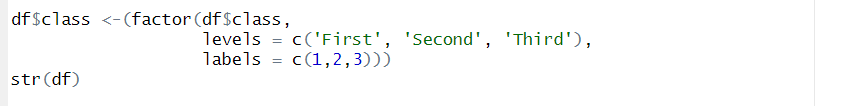


Removing Duplicates values

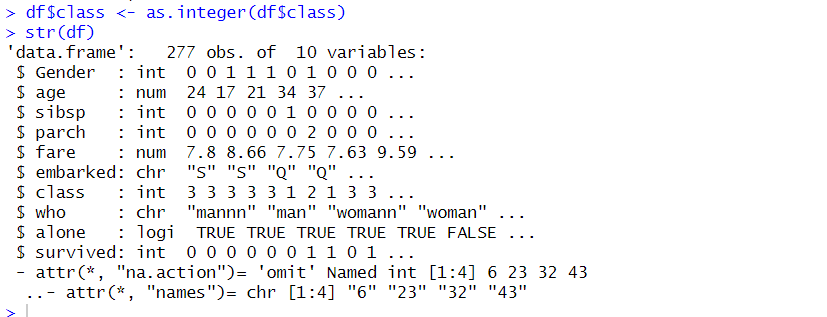


**# Data types and Conversion :**

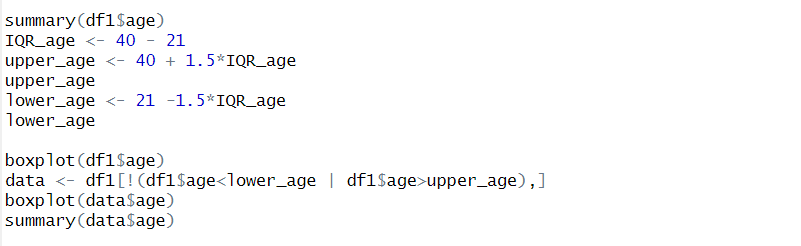
Annoting Class column First =1, Second =2, Third =3

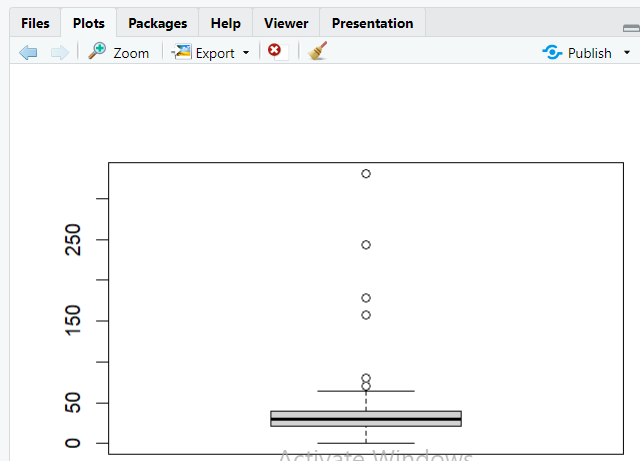


Converting Char to Integer Data type

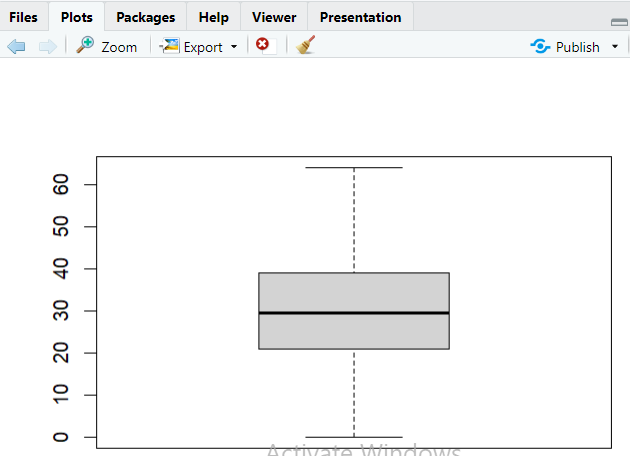


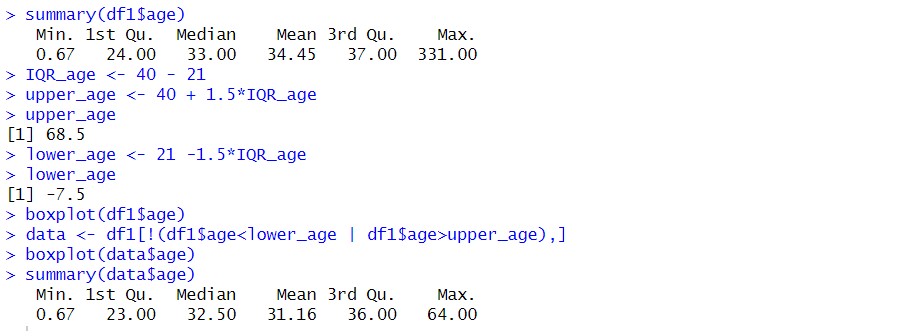
**#Outlier :**

Detecting Outlier for age attribute

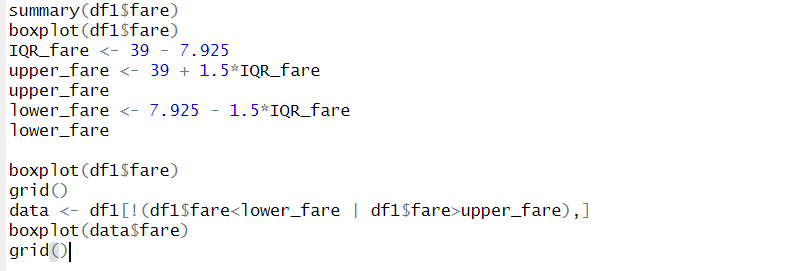


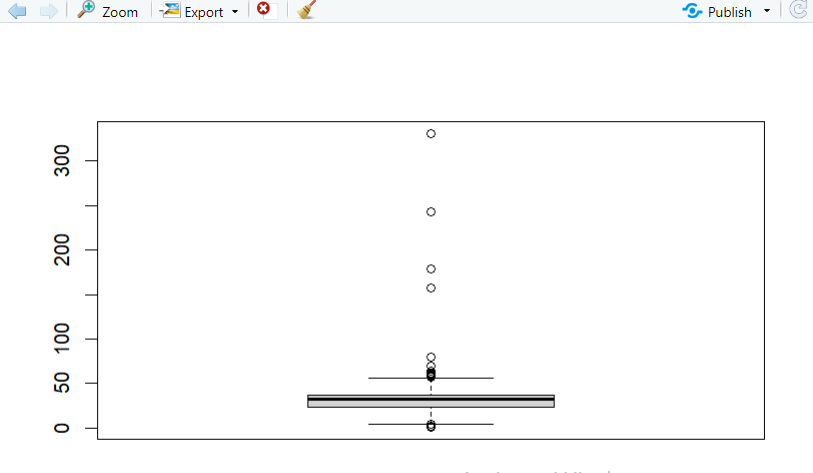
Removing Outliers

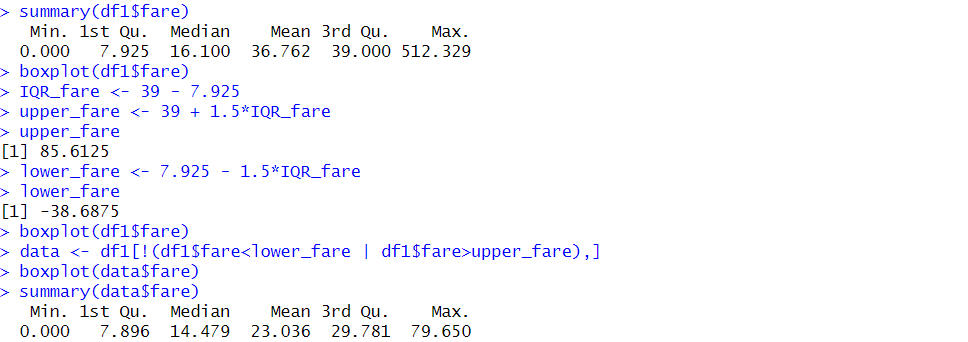




Detecting Outlier for fare attribute



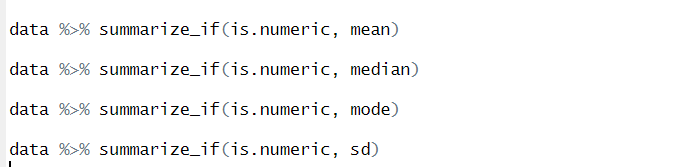


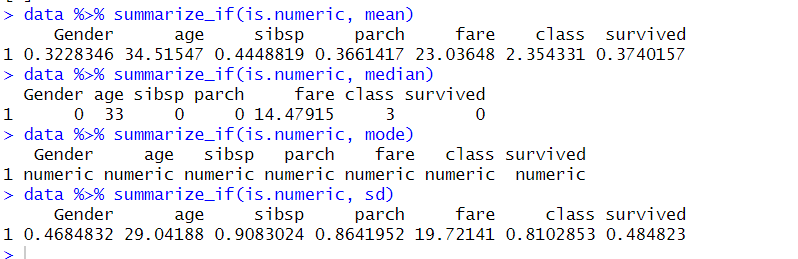


Standard deviation of ‘age’ and ‘fare’ attributes



**# Univariate Exploration(mean,median,mode, Standard Deviation) :**





**# Feature Selection :**

Feature selection

