A picture containing shape, arrow

Description automatically generated

**Imputation**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Navya.p**

**Batch Id: WDEO171220**

**Topic: Preliminaries for Data Analysis**

**Problem Statement:**

Majority of the datasets have missing values, that might be because the data collected were not at regular intervals or the breakdown of instruments and so on. It is nearly impossible to build the proper model or in other words, get accurate results. The common techniques are either removing those records completely or substitute those missing values with the logical ones, there are various techniques to treat these types of problems.

1. Prepare the dataset using various techniques to solve the problem, explore all the techniques available and use them to see which gives the best result.

**Hint:**  Go through this link: <https://360digitmg.com/mindmap-data-science>

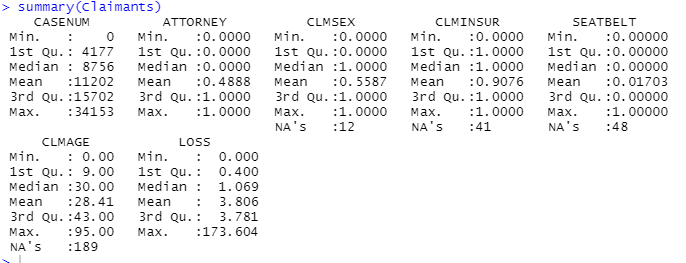
**A picture containing shape, arrow

Description automatically generated**

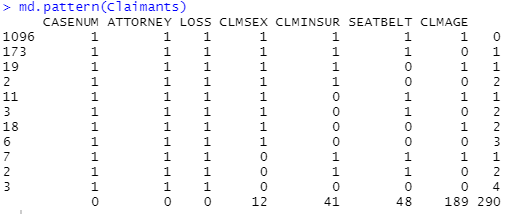
The Claimants data set has 7 columns, out of which only the 4 columns highlighted below have missing data.

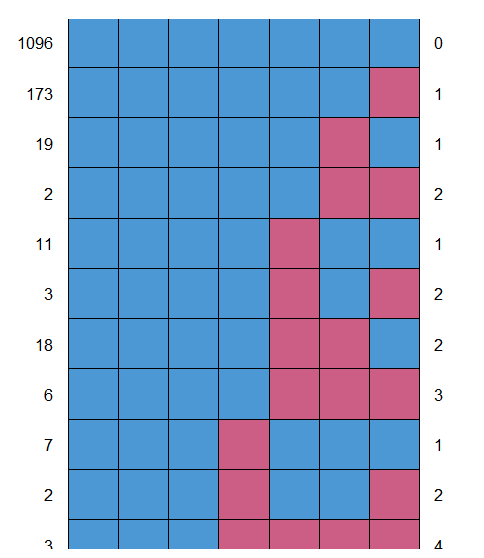
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CASENUM | ATTORNEY | CLMSEX | CLMINSUR | SEATBELT | CLMAGE | LOSS |

The Summary of the data is



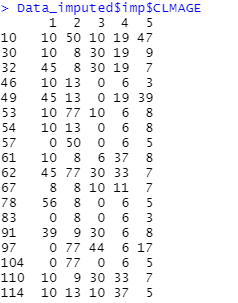
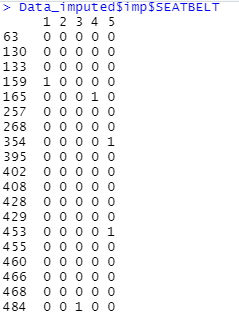
The missing data pattern is as below.

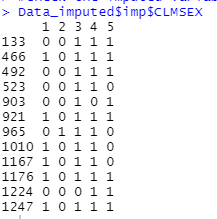
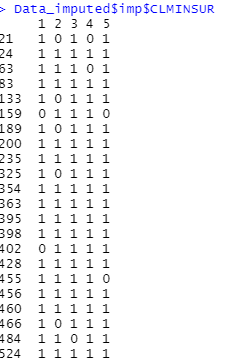




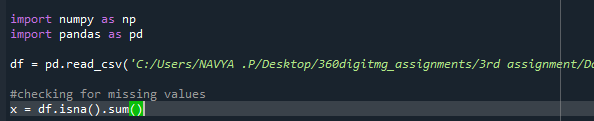
The missing data of 4 columns are imputed using MICE technique with 50 imputations and 50 max. Iterations.

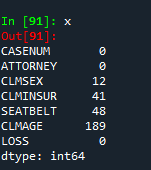
The data after imputation is as follows.

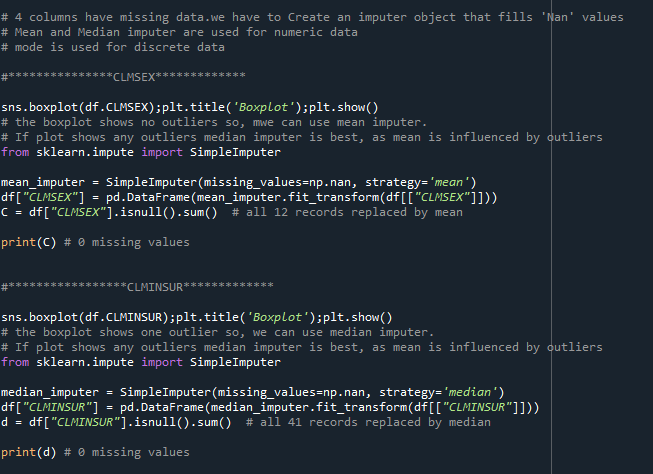
 

In python :

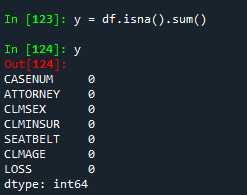




We have 4 columns as shown above with missing values. Out of 4, only CLMSEX have no outliers, so we can use mean imputer. For the rest of 3 columns which have outliers median imputer is calculated.



After replacing missing values:



**Hints:**

For each assignment, the solution should be submitted in the below format

1. Work on every feature of the dataset and create a data dictionary as an example displayed in the image below:



1. Hint: Refer to the file Claimants.csv.
2. The data is a vehicle Insurance data. Research on the Data fields and perform preliminary analysis
3. Research and perform all possible steps for obtaining solution
4. All the codes (executable programs) should execute without errors
5. Code modularization should be followed
6. Each line of code should have comments explaining the logic and why you are using that function