A picture containing shape, arrow

Description automatically generated

**Imputation**

Instructions:

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

Please ensure you update all the details:

**Name: Vishvash C**

**Batch Id: 23012024**

**Topic: Data Pre-Processing**

**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**

**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 vehicle Insurance data. Research the Data fields and perform preliminary analysis.
3. Research and perform all possible steps for obtaining the 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.

import numpy as np

import pandas as pd

# Load modified ethnic dataset

df = pd.read\_csv(r'C:/Users/Lenovo/Downloads/Study material/EDA/InClass\_DataPreprocessing\_datasets/claimants.csv') # for doing modifications

# Check for count of NA's in each column

print(df.isna().sum())

# For Mean, Median, Mode imputation we can use Simple Imputer or df.fillna()

from sklearn.impute import SimpleImputer

import seaborn as sns

df.info()

print(df.shape)

# casewise deletion with respect to the column 'CLMSEX'

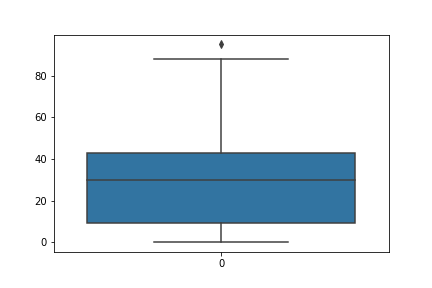
df.dropna(subset=['CLMSEX'], inplace=True)

df.CLMSEX.isna().sum()

df.reset\_index(drop=True, inplace=True) #mandatory step inorder not to lose data in further imputation steps

print(df.shape)

sns.boxplot(df.CLMAGE)



# Median Imputer

median\_imputer = SimpleImputer(missing\_values = np.nan, strategy = 'median')

df["CLMAGE"] = pd.DataFrame(median\_imputer.fit\_transform(df[["CLMAGE"]]))

df["CLMAGE"].isna().sum() #using median impuation to not get affected by outlier

# Mode Imputer since there are nominal data

mode\_imputer = SimpleImputer(missing\_values = np.nan, strategy = 'most\_frequent')

# df = pd.DataFrame(mode\_imputer.fit\_transform(df))

df["CLMINSUR"] = pd.DataFrame(mode\_imputer.fit\_transform(df[["CLMINSUR"]]))

df["SEATBELT"] = pd.DataFrame(mode\_imputer.fit\_transform(df[["SEATBELT"]]))

df.isnull().sum()

mode\_imputed\_df = df.fillna(df.mode().iloc[0])

df.CASENUM.value\_counts()

median\_imputed\_SEATBELT = df['SEATBELT'].fillna(df['SEATBELT'].median())

median\_imputed\_SEATBELT.isna().sum()

# Random Imputer

from feature\_engine.imputation import RandomSampleImputer

random\_imputer = RandomSampleImputer(['CLMAGE'])

df1 = pd.DataFrame(random\_imputer.fit\_transform(df[["CLMAGE"]]))

df1["CLMAGE"].isna().sum() # all records replaced by median

df1.describe()

df1["CLMAGE"].median()

print(df.isna().sum())

**Output:**

CASENUM 0

ATTORNEY 0

CLMSEX 12

CLMINSUR 41

SEATBELT 48

CLMAGE 189

LOSS 0

dtype: int64

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1340 entries, 0 to 1339

Data columns (total 7 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 CASENUM 1340 non-null int64

1 ATTORNEY 1340 non-null int64

2 CLMSEX 1328 non-null float64

3 CLMINSUR 1299 non-null float64

4 SEATBELT 1292 non-null float64

5 CLMAGE 1151 non-null float64

6 LOSS 1340 non-null float64

dtypes: float64(5), int64(2)

memory usage: 73.4 KB

(1340, 7)

(1328, 7)

CASENUM 0

ATTORNEY 0

CLMSEX 0

CLMINSUR 0

SEATBELT 0

CLMAGE 0

LOSS 0

dtype: int64

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| **ID** |  | **Quantitative/ Nominal** | **Irrelevant (ID does not provide useful information)** |
| CASENUM | Case number | Nominal | Irrelevant |
| ATTORNEY | Attorney indicator (1 if insured has attorney; 0 otherwise) | Nominal | Relevant |
| CLMSEX | Gender of the insured (1 = male; 2 = female) | Nominal | Relevant |
| CLMINSUR | Insurance indicator (1 if the insured has insurance; 0 otherwise) | Nominal | Relevant |
| SEATBELT | Seatbelt use indicator (1 if the insured was wearing a seatbelt; 0 otherwise) | Nominal | Relevant |
| CLMAGE | Age of the insured | Quantitative | Relevant |
| LOSS | Total loss incurred by the insurance company | Quantitative | Relevant |