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:**

**Batch Id:**

**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**

**Answer:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| CASENUM | Number of the case | Quantitative, Discrete, Ratio | Relevant |
| ATTORNEY | Gender of the claimant | Quanlitative, Discrete, Binary | Relevant |
| CLMSEX | Insurance is done by claimants or not | Quanlitative, Discrete, Binary | Relevant |
| CLMINSUR | Charles River dummy variable | Qualitative, Discrete,  Binary | Relevant |
| SEATBELT | Seatbelt used by claimant or not | Quanlitative,  Discrete, Binary | Relevant |
| CLMAGE | Age of the claimant | Quantitative, Discrete, Count | Relevant |
| LOSS | Total loss happened | Quantitative,  Continuous, Ratio | Relevant |

**Dataset:**

The dataset is that of an Insurance claimants data, wherein the objective is to figure out whether a person insured does hire an attorney or not based on the features available in the data.

Objective: Whether the person would hire an attorney

Constraint: The data has missing values

**Working with the Data**

* In order to start working with the dataset, we need to import the data first and the necessary libraries
* We use either Spyder, Jupyter, Colab, as per your system requirement, ease of use and availability.

**Pre processing and EDA**:

* We check the data description using “info” or “describe”
* Additionally checking the shape shows that the data consists of 1340 rows and 7 columns/ features
* We check the data for missing values using the function “isna()” and summing it up with “sum()”
* It describes that there are missing values present in the data to the frequencies of
  1. CLMSEX – 12 missing values
  2. CLMINSUR – 41 missing values
  3. SEATBELT – 48 missing values
  4. CLMAGE – 189 missing values
* Now we know to replace the missing values we use imputation, We have various methods to impute the missing values and we use them for the problem. We first import the function as

“ fromsklearn.impute import SimpleImputer” and do the following:

* + 1. For CLMAGE, since it is a continuous data and not discrete, we can impute the data using mean values.
    2. For the discrete categorical data we use the median and mode (most-frequent)
    3. Post applying the same we check again for the missing values in the data and find that the same are now not present and have been imputed successfully.
* Now we see that the data has the first column of CASENUM, which is basically nominal in nature. Hence we can easily drop the column/ feature using “drop” function.
* We can either check the four moments of business decision by implementing individually or use define a EDA function and do so as well.
* The new data now we have has all values intact and no meaningless data into the dataset. But, the data has two features that have scales varying to that of others, viz, “CLMAGE” and “LOSS”.
* In order to make the two feature scale free we might use Normalization or standardization techniques.
* Herein we use the Normalization technique which needs to be defined. We define the norm function which is also called the min max scaler and apply on the two features.
* Post applying the normalization we can see that the data in the two features have now been set between a range of 0 to 1 which coincide with the rest of the data and the data is hence normalized.
* Now we also need to check whether the data has any outliers present. In order to do so we use boxplots and see that the data has outliers in present. But, we need to understand that the data provided might require the same values as they are not actually outliers but values that are relevant to the dataset and provide a meaningful insight to the data.
* Hence the data is now pre-processed and EDA performed. It is ready for further processing!

**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