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: DHEERAJ MISHRA**

**Batch Id: DS\_01072021**

**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**

**SOLN:-**

1. **Business understanding:**

**Max:- Claimants convenience**

**Min:- Churn rate**

**Constraints:- Competition of market**

1. **Data understanding**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| **CASENUM** | **Case number** | **Nominal , discrete** | **It does not provide useful information** |
| **ATTORNEY** | **Having lawyer or not** | **Discrete** | **Relevant** |
| **CLMSEX** | **Claimants gender** | **Discrete** | **Relevant** |
| **CLMINSUR** | **Claimant insured or not** | **Discrete** | **Relevant** |
| **SEATBELT** | **Wearing seat belt or not** | **Discrete** | **Relevant** |
| **CLMAGE** | **Claimants age** | **Discrete** | **Relevant** |
| **LOSS** | **Economic loss** | **Continuous** | **Relevant** |

**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

**A picture containing shape, arrow

Description automatically generatedGrading Guidelines:**

**Note: 1. An Assignment submission is considered complete only when successful executable code(s), and documentation explaining the applied solution and results are provided. Failing to submit either of them will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline date will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline date
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of questions asked in assignments are completed. (or)
  + All assignments were submitted, however, after the given deadline
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of questions asked in assignments are completed. (or)
  + Less than 80% of questions asked in assignments are submitted after the deadline
* **Grade D: (>= 60 and < 70):** Assignments submitted after the Deadline and with 50% or less of questions
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of questions asked in the assignments are submitted after the deadline (OR)
  + Less than 30% of questions asked in the assignments are submitted before deadline
* **Grade F: (< 50):** Copied submission or No submission