

**Outlier Treatments**

**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:**

Most of the datasets have extreme values or exceptions in their observations. These values affect the predictions (Accuracy) of the model in one way or the other, removing these values is not a very good option. For these types of scenarios, we have various techniques to treat such values.

Refer: <https://360digitmg.com/mindmap-data-science>

1. Prepare the dataset by performing the preprocessing techniques, to treat the outliers.

A picture containing shape, arrow

Description automatically generated**

**Soln:-**

1. **Business understanding:**

**Max:- Living conditions**

**Min:- Attributes specification in boston data set**

**Constraints:- Factors effecting living conditions**

1. **Data understanding**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| **Crim** | Per capita crime rate by town | **Continuous** | **Relevant** |
| **zn** | Proportion of residential land zoned | **Continuous** | **Relevant** |
| **indus** | Proportion of non-retail business | **Continuous** | **Relevant** |
| **chas** | Charles River dummy variable | **Discrete** | **Relevant** |
| **nox** | Nitric oxides concentration | **Continuous** | **Relevant** |
| **rm** | Average number of rooms per dwelling | **Continuous** | **Relevant** |
| **age** | Proportion of owner-occupied | **Continuous** | **Relevant** |
| **dis** | Weighted distances to five Boston employment centres | **Continuous** | **Relevant** |
| **Rad** | Index of accessibility to radial highways | **Continuous** | **Relevant** |
| **Tax** | Full-value property-tax rate | **Continuous** | **Relevant** |
| **ptratio** | Pupil-teacher ratio by town | **Continuous** | **Relevant** |
| **black** | Proportion of blacks by town | **Continuous** | **Relevant** |
| **istat** | Median value of owner occupied homes | **Continuous** | **Relevant** |

**Hints:**

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

1. Work on each feature to create a data dictionary as displayed in the image displayed below:
2. Hint: Boston dataset is publicly available. Refer to Boston.csv file.
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
7. Detailed explanation of your approach is mandatory

**Grading 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