Guidelines for Mini-Project

Dataset Constraints

- Min: not less than 10 columns(attributes)
- Min: not less than 500 rows or records
- Min: 3 numerical columns and 2 categorical columns
- About 3-5% of data as missing values and NAN.

Marks distribution

1. Data Cleaning:

- All the NAN's for categorical columns to be replaced with its previous row values(**1 mark**)
- All the NAN's for numeric columns to be replaced with average of the column(**1 mark**)
- Interpolation of immediate data before and after it(**1 mark**)
  Please put up screenshot of dataset before and after data cleaning as an example.(Only the necessary part of the dataset)

2. Normalization and Standardization:

- Normalize all the numeric columns, to make mean 0 and variance 1(**1 mark**)
- Why is normalization important? How does it affect dataset? Different graphs used to check whether the data is normal. (**2 marks**)

3. Graph visualization: Visualize the dataset to infer some meaning insights about it.

- Use at least 3 different graph visualization techniques(**2 mark**)
- Come up with two meaningful insights from each of the graphs(**1 mark**)

4. Hypothesis Testing: (**2.5 marks**)

Freedom to make your own hypothesis based on the columns.

5. Correlation:

- Come up with the columns which are most and least correlated and give insights. (Make conclusion based on correlation co-efficient and state reasons/inferences) (**1.5 mark**)

6. Presentation:

- PPT and presentation during class hours. Screen shot of a particular part of dataset and related graphs can be put in PPT along with insights. (**2 marks**)