**2b. Graphical Representation**

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_ Batch ID:** \_\_\_\_\_\_\_\_\_\_\_

**Topic: Data Visualization**

**Guidelines:**

**1. An assignment submission is considered complete only when the correct and executable code(s) is submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered a correct submission.**

**2. Ensure that you submit your assignments correctly. Resubmission is not allowed.**

**3. Post the submission you can evaluate your work by referring to the keys provided. (will be available only post the submission).**

**Hints: Follow CRISP-ML(Q) methodology steps, where were appropriate.**

1. **Data Understanding: work on each feature of the dataset to create a data dictionary as displayed in the image below:**

Table

Description automatically generated

**Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

**Problem Statements:**

**1. Univariate plots for UNIV data (Plot must have Title, X & Y label)**

**A) Plot numerical column with 3 different plots ?**

**B) What are bin parameters? What are the methods to define the number of bins and bin sizes ?**

* + Ans:)
* Number of bins, Bin width, Bin edges, Starting point.
* Number of bins - Square root method , Sturges' formula , Freedman-Diaconis' rule
* Bin width = (max value - min value) / number of bins

**C) Why do density plots exceed the range values of the column ?**

* + Ans:)
* The density plot can extend beyond the actual values because it is an estimate of the probability density function of the underlying distribution that the data comes from.

**D) Plot categorical columns by taking unique values ?**

**2. Bivariate graphs for UNIV data (Plot must be readable [use rotation], have all labels)**

**A) Plot 2 numerical columns with scatter plot [use grid] ?**

**B) 2 Different plots for plotting a numerical column with a categorical column (bar, line) ?**

**C) How are bar plots different from histogram?**

Ans:)

| Bar Plots | Histograms |
| --- | --- |
| Used to compare the values of different categories or groups. | Used to show the distribution of a single variable. |
| Can be vertical or horizontal. | Can only be vertical. |
| Bars are separated and do not touch each other. | Bars are touching each other and form a continuous distribution. |
| Suitable for categorical or discrete data. | Suitable for continuous or numerical data. |
| X-axis represents categories or groups. | X-axis represents the range of the variable being measured. |
| Y-axis represents a numerical value. | Y-axis represents the frequency or count of the data in each bin. |

**3. Plot multivariate graphs (correlation heatmap, pairplot)**

**A) Plot for only numerical data ?**

**B) Plot multivariate graphs for both numerical and categorical columns ?**

**C) What does it mean when a correlation value says 1? When it is negative? When it is zero?**

Ans:)

Correlation value of 1: It means that there is a perfect positive correlation between the two variables. This means that when one variable increases, the other variable also increases proportionally, and vice versa.

Correlation value of -1: It means that there is a perfect negative correlation between the two variables. This means that when one variable increases, the other variable decreases proportionally, and vice versa.

Correlation value of 0: It means that there is no correlation between the two variables. This means that there is no relationship between the variables. However, it's important to note that a correlation of 0 does not necessarily imply independence between the variables, as there may be other types of relationships that are not captured by correlation.

**4. Plot Skewness & Probability distribution for each column of marks data. (Hist, box, density)**

**A) What is normally distributed and What will be the relationship between mean, median & mode ?**

Ans:)

* Math column is normally distributed. Generally, Mean = Median = Mode when data is normally distributed.

**B) Which data variables are positively skewed and What will be the relationship between mean, median & mode ?**

Ans:

* Science column is positively distributed. Generally, Mean > Median > Mode when data is positively distributed.

**C) What are negatively skewed/distributed and What will be the relationship between mean, median & mode**

Ans:)

* Social studies column is positively distributed Generally Mean < Median < Mode when data is negatively distributed.

**D) What are the distinctive differences between skewness and distribution?**

Ans:)

Distinctive differences between skewness and distribution:

* Distribution refers to the shape or spread of the data points in a dataset, while skewness refers to the degree and direction of the asymmetry of the distribution.
* Skewness measures how far the data is spread out on one side of the mean compared to the other side, while distribution refers to the values of the data points themselves.
* A distribution can be symmetrical, positively skewed, or negatively skewed, while skewness can be positive, negative, or zero.
* Skewness can be affected by extreme values, while distribution is not.