2a. Graphical Representation Assignment

Name: ULLI VENKATA SAI KUMAR Batch ID: 04072024HYD10AM

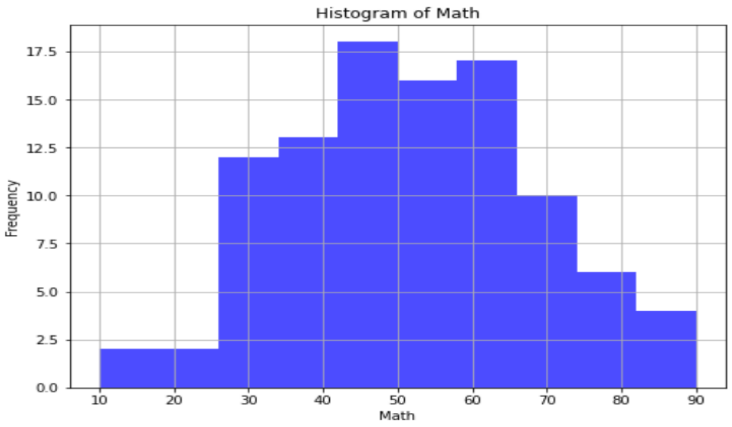
Topic: Data Visualization

Problem Statements:

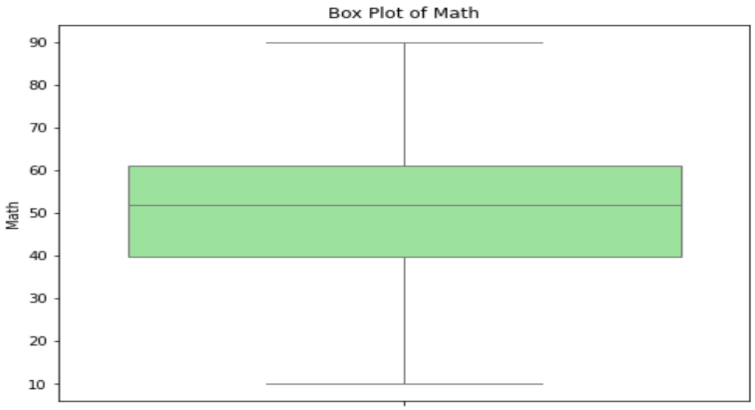
1. Univariate plots for UNIV data (Plot must have Title, X & Y label)

A) Plot numerical column with 3 different plots ?

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
file_path = (r'Marks Data.csv')
df = pd.read_csv(file_path)
numerical_column = 'Math'

# Histogram
plt.figure(figsize=(8, 6))
plt.hist(df[numerical_column], bins=10,
         color='blue', alpha=0.7)
plt.title(f'Histogram of {numerical_column}')
plt.xlabel(numerical_column)
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

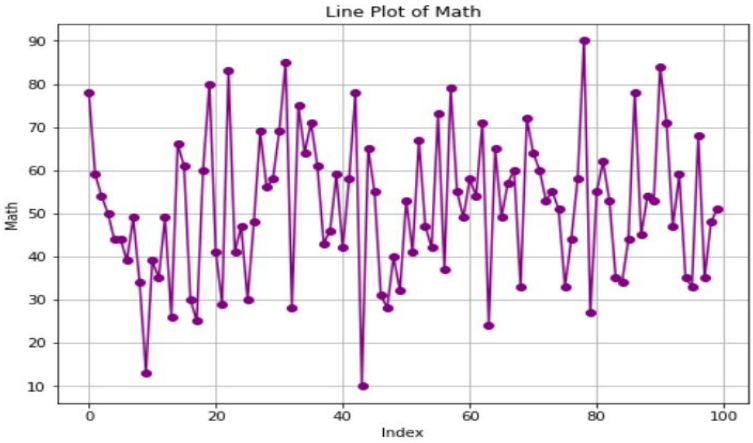



```
# Box Plot
plt.figure(figsize=(8, 6))
sns.boxplot(y=df[numerical_column],
            color='lightgreen')
plt.title(f'Box Plot of {numerical_column}')
```



```
plt.ylabel(numerical_column)
plt.show()
```

```
# Line Plot
plt.figure(figsize=(8, 6))
plt.plot(df[numerical_column], marker='o',
linestyle='-', color='purple')
plt.title(f'Line Plot of {numerical_column}')
plt.xlabel('Index')
plt.ylabel(numerical_column)
plt.grid(True)
plt.show()
```



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

Bins are intervals that represent the range of data in a histogram. When you plot a histogram, the data is divided into these intervals or "bins," and the frequency of data points within each bin is counted and displayed as bars. The choice of bin size and the number of bins can significantly affect the appearance and interpretation of the histogram.

Bin parameters include:

Number of Bins (bins): This is the number of intervals (or bins) into which the data range is divided.

Bin Width: This is the width of each bin, calculated as the range of the data divided by the number of bins.

Bin Edges: These are the boundaries of each bin, determined by the bin width and the range of data.

Methods to Define the Number of Bins and Bin Sizes

Sturges' Rule:

Suggests that the number of bins should be

k=⌈log2(n)+1⌉, where n n is the number of data points.

It works well for smaller datasets and assumes that the data is normally distributed.

Scott's Rule:

The bin width h is calculated as

$h=3.49\sigma n-1/3$, where σ is the standard deviation of the data.

It aims to minimize the integrated mean square error and is useful for continuous data with fewer outliers.

Freedman-Diaconis Rule:

The bin width h is determined by

$h=2\times IQR\times n^{-1/3}$, where IQR is the interquartile range of the data.

It is robust to outliers and adjusts the bin width based on the spread of the data.

Square-root Choice:

The number of bins is chosen as

$k = n$ where n is the number of data points.

It's a simple method, often used for quick visualization, particularly when the distribution is unknown.

Doane's Formula:

An extension of Sturges' Rule, it adjusts the number of bins based on data skewness:

Useful for skewed data.

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

Density plots can exceed the range values of the column due to the nature of how they are calculated, specifically through a process called Kernel Density Estimation (KDE).

Kernel Density Estimation (KDE)

KDE is a technique used to estimate the probability density function (PDF) of a random variable. Unlike a histogram, which counts the number of occurrences within predefined bins, KDE creates a smooth curve to represent the distribution.

Kernels: A kernel is a smooth, symmetric function (often Gaussian, but others like Epanechnikov or Tophat can be used) that is applied at each data point to contribute to the overall density estimate.

Why Density Plots Exceed the Range

Gaussian Kernel:

The Gaussian (normal) kernel, which is commonly used in KDE, extends infinitely in both directions. This means that even though most of the kernel's mass is near the data point, it still has some influence on points far away from it.

As a result, the density estimate at the boundaries of the data might be non-zero, even outside the actual data range.

Smoothing:

KDE involves smoothing the data points across a range, which can cause the curve to stretch beyond the minimum and maximum data points.

This smoothing process allows the density curve to extend into regions where no actual data points exist, especially at the tails.

Bandwidth Selection:

The bandwidth parameter in KDE controls the width of the kernel. A larger bandwidth means more smoothing, which can further extend the density plot beyond the data range.

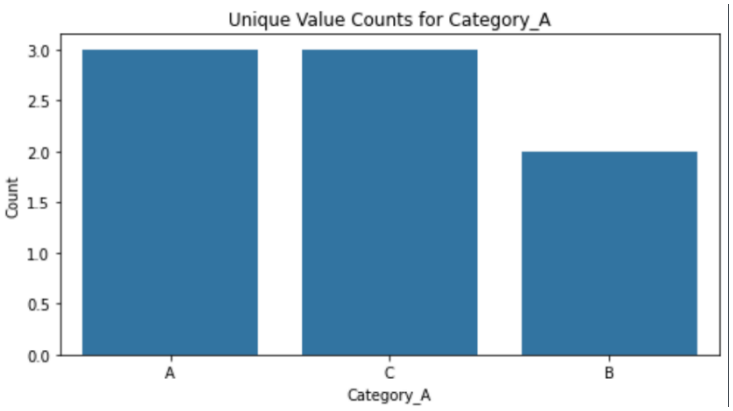
If the bandwidth is too large, the smoothing effect may cause the density plot to cover areas far beyond the actual data points.

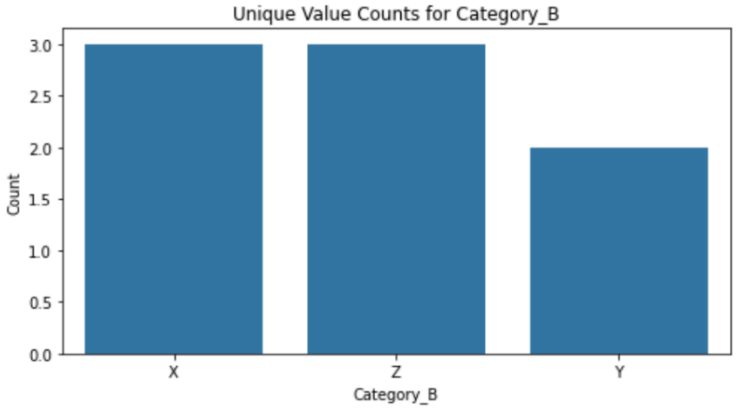
Visual Impact

The extension of the density plot beyond the data range is a visual artifact of the smoothing process and the choice of the kernel. It doesn't imply that the actual data extends into those regions but rather shows the estimated probability distribution based on the available data.

D) Plot categorical columns by taking unique values ?

```
    import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data = {
   'Category_A': ['A', 'B', 'A', 'C', 'B', 'A', 'C', 'C'],
   'Category_B': ['X', 'Y', 'X', 'Z', 'Y', 'X', 'Z', 'Z'],
}
df = pd.DataFrame(data)
categorical_columns = df.select_dtypes
(include=['object', 'category']).columns
for column in categorical_columns:
   unique_value_counts =
df[column].value_counts()
   plt.figure(figsize=(8, 4))
   sns.barplot(x=unique_value_counts.index,
    y=unique_value_counts.values)
   plt.title(f'Unique Value Counts for {column}')
   plt.xlabel(column)
   plt.ylabel('Count')
   plt.show()
```

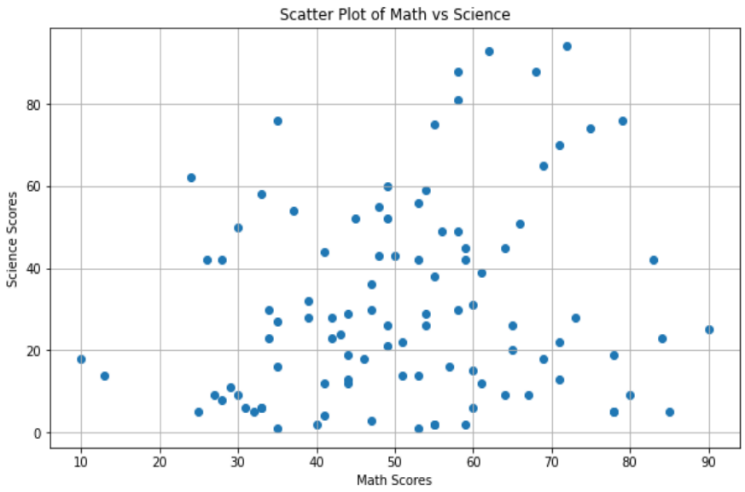




2. Bivariate graphs for UNIV data (Plot must be readable [use rotation], have all labels)

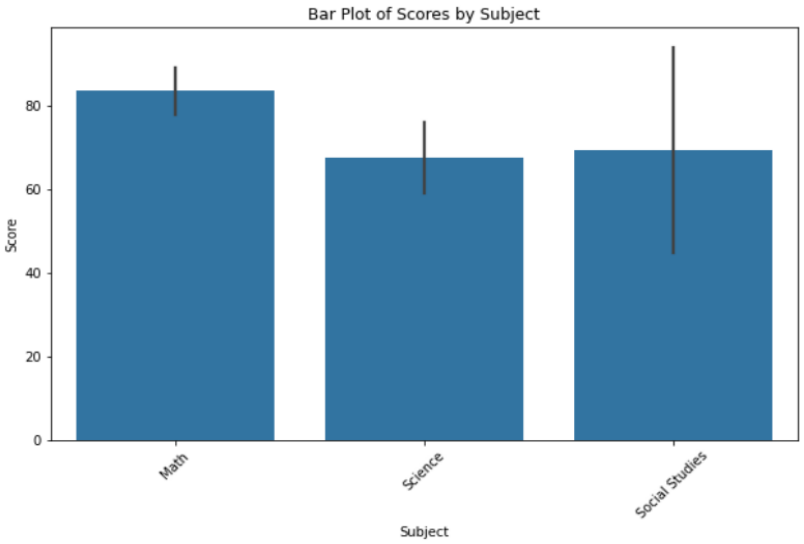
A) Plot 2 numerical columns with scatter plot [use grid] ?

```
import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv(r"Marks Data.csv")
plt.figure(figsize=(10, 6))
plt.scatter(df['Math'], df['Science'])
plt.title('Scatter Plot of Math vs Science')
plt.xlabel('Math Scores')
plt.ylabel('Science Scores')
plt.grid(True)
plt.show()
```

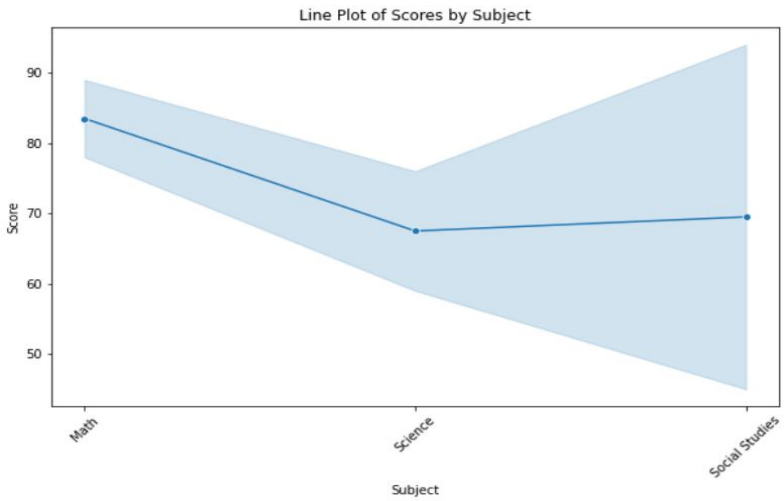


B) 2 Different plots for plotting a numerical column with a categorical column (bar, line) ?

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
# Example data
data = {
    'Subject': ['Math', 'Science',
                'Social Studies', 'Math',
                'Science', 'Social Studies'],
    'Score': [78, 59, 45, 89, 76, 94]
}
df = pd.DataFrame(data)

# Creating the bar plot
plt.figure(figsize=(10, 6))
sns.barplot(x='Subject', y='Score', data=df)
plt.title('Bar Plot of Scores by Subject')
plt.xlabel('Subject')
plt.ylabel('Score')
plt.xticks(rotation=45)
plt.show()
```



```
# Creating the line plot
plt.figure(figsize=(10, 6))
sns.lineplot(x='Subject', y='Score', data=df,
marker='o')
plt.title('Line Plot of Scores by Subject')
plt.xlabel('Subject')
plt.ylabel('Score')
plt.xticks(rotation=45)
plt.show()
```



C) How are bar plots different from histogram?

Bar plots and histograms are both used to visualize data, but they serve different purposes and are structured differently. Here's a breakdown of their key differences:

1. Purpose

- **Bar Plots:**

  Compare Different Categories: Bar plots are used to compare the values of a numerical variable across different categories. Each bar represents a category, and the height of the bar shows the value associated with that category.

- **Histograms:**

  Show Distribution of a Single Variable: Histograms are used to visualize the distribution of a single numerical variable. They show how the data is spread across different intervals (or bins) and how frequently data points fall into those intervals.

2. X-Axis

- **Bar Plots:**

  Categorical Data: The x-axis represents distinct categories or groups (e.g., different subjects, products, etc.).

- **Histograms:**

  Continuous Data: The x-axis represents the range of values of the numerical variable, divided into intervals (bins). The bars touch each other, indicating the continuous nature of the data.

3. Spacing Between Bars

- **Bar Plots:**

  Spaces Between Bars: There are usually spaces between the bars to indicate that the categories are distinct and separate from each other.

- **Histograms:**

No Spaces Between Bars: The bars are adjacent to each other without spaces, indicating the continuous nature of the data. The lack of gaps shows that the intervals are connected and that the data is continuous.

4. Data Type

- **Bar Plots:**

  Categorical vs. Numerical: Bar plots typically have a categorical variable on the x-axis and a numerical variable on the y-axis.

- **Histograms:**

  Numerical Data Only: Histograms are used solely for numerical data, showing the frequency of data points within specific intervals.

5. Use Cases

- **Bar Plots:**

  Comparing Categories: Bar plots are best used when you need to compare different categories or groups, such as comparing sales across different products or test scores across different subjects.

- **Histograms:**

  Analyzing Distribution: Histograms are used when you want to understand the distribution of a dataset, such as determining the frequency of test scores within certain ranges or analyzing the distribution of ages in a population.

Visual Comparison

- **Bar Plot:**

  Example: Comparing the average test scores of students in different subjects like Math, Science, and English.

- **Histogram:**

  Example: Showing the distribution of test scores across a range, like how many students scored between 50-60, 60-70, etc.

3. Plot multivariate graphs (correlation heatmap, pairplot)

A) Plot for only numerical data ?

1. Correlation Heatmap

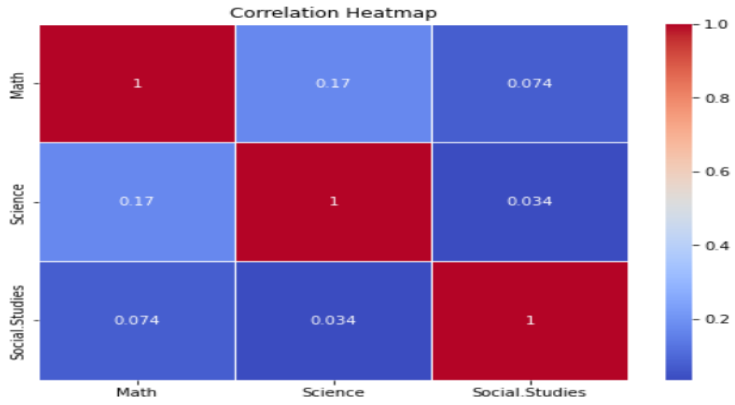
A correlation heatmap visualizes the correlation matrix of numerical variables. Correlation values range from -1 to 1, where:

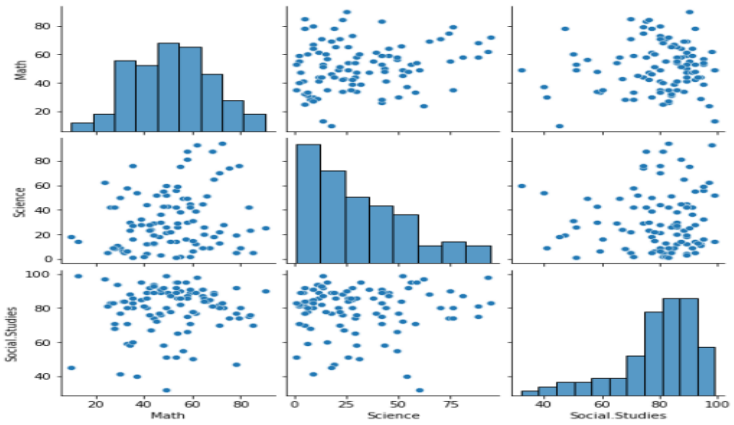
- **1** indicates a perfect positive correlation.
- **-1** indicates a perfect negative correlation.
- **0** indicates no correlation.

2. Pairplot

A pairplot plots pairwise relationships in a dataset. It includes scatter plots for each pair of numerical variables and histograms for each individual variable.

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt


df = pd.read_csv(r"Marks Data.csv")


# Correlation Heatmap

plt.figure(figsize=(8, 6))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Correlation Heatmap')

plt.show()


# Pairplot

sns.pairplot(df)

plt.show()
```

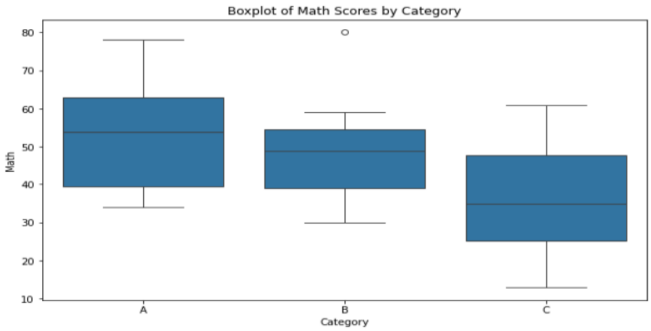




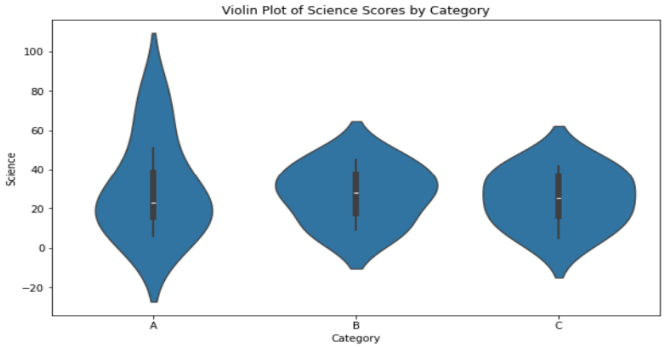
B) Plot multivariate graphs for both numerical and categorical columns ?

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Example Data

data = {

    'Math': [78, 59, 54, 50, 44, 44, 39, 49, 34, 13, 39, 35, 49, 26, 66, 61, 30, 25, 60, 80],

    'Science': [19, 45, 26, 43, 12, 29, 28, 21, 23, 14, 32, 76, 26, 42, 51, 39, 9, 5, 6, 9],

    'Social.Studies': [47, 89, 86, 89, 94, 74, 85, 80, 88, 99, 95, 80, 51, 83, 84, 77, 41, 81, 78, 80],

    'Category': ['A', 'B', 'A', 'B', 'A', 'C', 'B', 'C', 'A', 'C', 'B', 'A', 'B', 'C', 'A', 'C', 'B', 'C', 'A', 'B']

}

# Create DataFrame
```

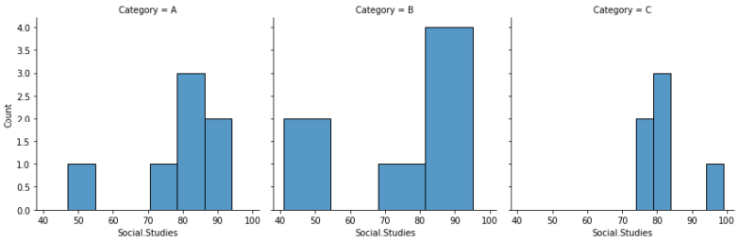
```
df = pd.DataFrame(data)
# 1. Boxplot (Numerical vs Categorical)
plt.figure(figsize=(10, 6))
sns.boxplot(x='Category', y='Math', data=df)
plt.title('Boxplot of Math Scores by Category')
plt.show()
```



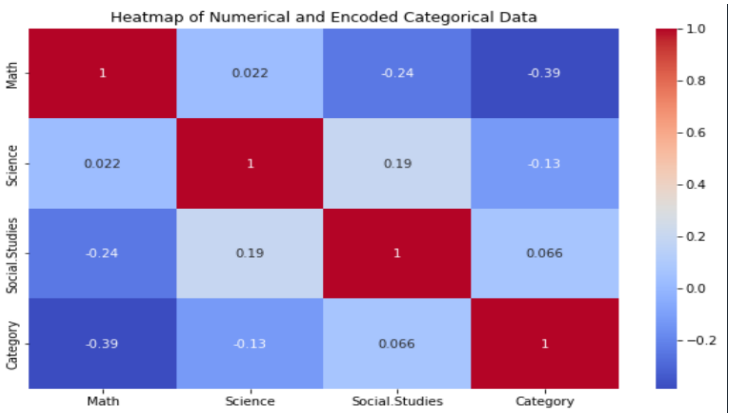

```
# 2. Violin Plot (Numerical vs Categorical)
plt.figure(figsize=(10, 6))
sns.violinplot(x='Category', y='Science', data=df)
plt.title('Violin Plot of Science Scores by Category')
plt.show()
```




```
# 3. FacetGrid (Numerical and Categorical)
g = sns.FacetGrid(df, col='Category', height=4)
g.map(sns.histplot, 'Social.Studies')
plt.show()
```




```
# 4. Heatmap (Categorical vs Numerical Relationship)
# For illustration, we convert 'Category' into
 numerical values to plot heatmap.
df_encoded = df.copy()
df_encoded['Category'] =
df_encoded['Category'].
astype('category').cat.codes
plt.figure(figsize=(10, 6))
sns.heatmap(df_encoded.corr(), annot=True, cmap='coolwarm')
plt.title('Heatmap of Numerical and Encoded Categorical Data')
plt.show()
```



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

Correlation measures the strength and direction of the relationship between two variables. The correlation coefficient, typically denoted as rrr, ranges from -1 to 1 and can be interpreted as follows:

1. Correlation Value = 1

- **Perfect Positive Correlation**: A correlation value of **1** means that there is a perfect positive linear relationship between the two variables.
- **Interpretation**: As one variable increases, the other variable increases proportionally. For example, if variable X increases by a certain amount, variable Y increases by a consistent, proportional amount.
- **Graphical Representation**: On a scatter plot, all data points would lie perfectly on a straight line with a positive slope.

2. Correlation Value = -1

- **Perfect Negative Correlation**: A correlation value of **-1** indicates a perfect negative linear relationship between the two variables.
- **Interpretation**: As one variable increases, the other variable decreases proportionally. For example, if variable X increases, variable Y decreases by a consistent, proportional amount.
- **Graphical Representation**: On a scatter plot, all data points would lie perfectly on a straight line with a negative slope.

3. Correlation Value = 0

- **No Linear Correlation**: A correlation value of **0** indicates that there is no linear relationship between the two variables.
- **Interpretation**: Changes in one variable do not predict any consistent change in the other variable. The variables are uncorrelated in a linear sense.
- **Graphical Representation**: On a scatter plot, the points would be scattered with no discernible pattern or linear trend.

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 ?

Steps to Visualize and Analyze the Data

1. **Histograms:** Visualize the frequency distribution of the scores in each subject.
2. **Box Plots:** Identify the spread, central tendency, and outliers in the data.
3. **Density Plots:** Estimate the probability distribution of the data.

Skewness and Normal Distribution

Skewness:

- Skewness is a measure of asymmetry in the data distribution.
    - **Positive Skew:** The tail on the right side is longer or fatter than the left side. The distribution is skewed to the right.

- **Negative Skew:** The tail on the left side is longer or fatter than the right side. The distribution is skewed to the left.
- **Zero Skewness:** Symmetrical distribution, typical of a normal distribution.

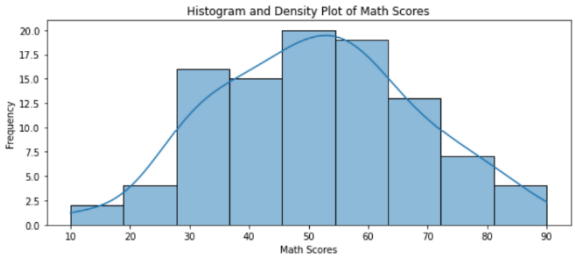
Normal Distribution:

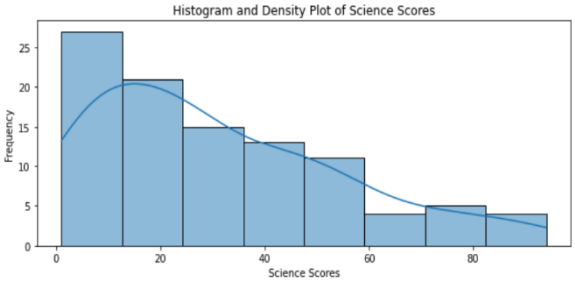
- In a perfectly normal distribution, the **mean**, **median**, and **mode** of the data are all equal and lie at the center of the distribution.

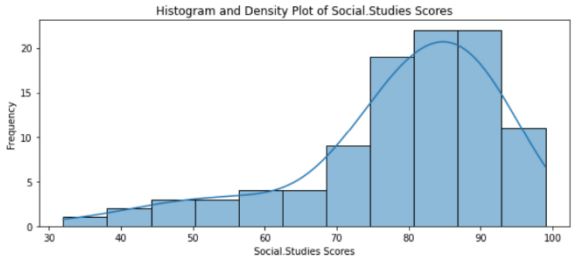
Relationship Between Mean, Median, and Mode

- **In a Normal Distribution:** Mean = Median = Mode.
- **In a Positively Skewed Distribution:** Mean > Median > Mode.
- **In a Negatively Skewed Distribution:** Mean < Median < Mode.

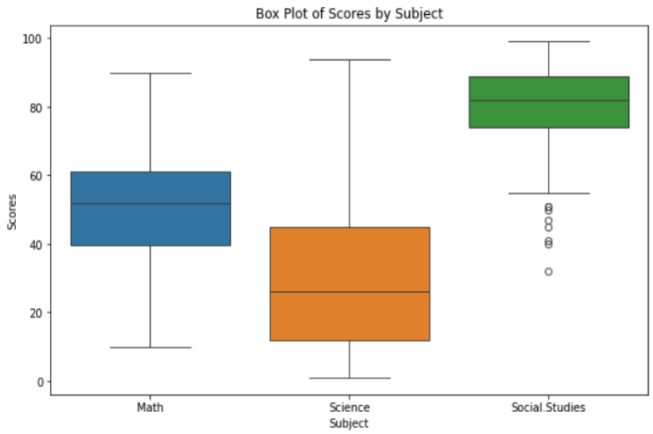
```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt



df = pd.read_csv(r"Marks Data.csv")


# Plotting Histograms

for column in df.columns:

    plt.figure(figsize=(10, 4))

    sns.histplot(df[column], kde=True)

    plt.title(f'Histogram and Density Plot of

    {column} Scores')

    plt.xlabel(f'{column} Scores')

    plt.ylabel('Frequency')
```







```
plt.show()


# Plotting Box Plots
plt.figure(figsize=(10, 6))
sns.boxplot(data=df)
plt.title('Box Plot of Scores by Subject')
plt.xlabel('Subject')
plt.ylabel('Scores')
plt.show()
```



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

- **Positive Skewness** occurs when the distribution of data has a longer right tail. This means that the majority of the data values are concentrated on the left, with fewer values trailing off to the right.

Visual Indicators:

- **Histograms:** In a positively skewed histogram, the right side (tail) is longer or fatter than the left side.
- **Box Plots:** The median will be closer to the lower quartile, with a longer whisker or more outliers on the right side.

```
import pandas as pd
df = pd.read_csv(r"Marks Data.csv")
skewness = df.skew()
print(skewness)
```

OUTPUT:-

```
Math            0.048448
Science         0.841328
Social.Studies  -1.262470
```

Relationship Between Mean, Median, and Mode in a Positively Skewed Distribution

- **Mean > Median > Mode**
    - **Mean:** Is affected by the extreme values in the tail, so it is pulled in the direction of the skew (to the right).
    - **Median:** Lies between the mean and mode and is less affected by the extreme values.
    - **Mode:** The most frequently occurring value, is typically at the peak of the distribution.

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

Negatively Skewed Distribution

Negatively skewed distribution occurs when the distribution of data has a longer tail on the left side. This means that the majority of the data values are concentrated on the right, with fewer values trailing off to the left.

Visual Indicators:

- **Histograms:** In a negatively skewed histogram, the left side (tail) is longer or fatter than the right side.
- **Box Plots:** The median will be closer to the upper quartile, with a longer whisker or more outliers on the left side.

Calculation of Skewness:

- A **negative skewness value** indicates a negatively skewed distribution.

Relationship Between Mean, Median, and Mode in a Negatively Skewed Distribution

- **Mean < Median < Mode**
    - **Mean:** The mean is pulled towards the tail on the left side, making it less than the median.
    - **Median:** The median lies between the mean and mode.
    - **Mode:** The mode is the peak of the distribution and is located to the right of the median and mean.

```
import pandas as pd

df = pd.read_csv(r"Marks Data.csv")

skewness = df.skew()

print(skewness)
```

OUTPUT:-

```
Math            0.048448

Science         0.841328

Social.Studies  -1.262470
```

Interpretation:

- If the skewness for a subject (e.g., Science) is negative, that subject's scores are negatively skewed.
- For negatively skewed subjects, the expected relationship is:

    Mean < Median < Mode

Conclusion:

- **Negatively Skewed Variables:** Any subject with a negative skewness value.

- **Relationship:** For these subjects, the mean will be less than the median, which in turn will be less than the mode.

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

Skewness and distribution are related concepts in statistics, but they describe different aspects of data. Here are the distinctive differences between them:

1. Definition

- **Skewness:**
    - Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. It quantifies the degree to which a distribution leans to the left or right.
    - **Positive Skewness**: The distribution has a longer tail on the right (more values on the lower side).
    - **Negative Skewness**: The distribution has a longer tail on the left (more values on the higher side).
- **Distribution:**
    - A distribution is a general term that describes how the values of a variable are spread or distributed across a range. It provides information about the range of data, central tendency (like mean, median, mode), and the shape of the data (whether it's normal, uniform, etc.).
    - Common types of distributions include **Normal Distribution**, **Uniform Distribution**, **Exponential Distribution**, etc.

2. Focus

- **Skewness:**
    - Focuses on the asymmetry of the distribution.
    - Skewness is a numerical value that can be calculated to describe the extent and direction of skewness.
    - Does not describe the overall shape of the distribution, just the asymmetry.
- **Distribution:**
    - Describes the entire shape and nature of the data, including its central tendency, spread (variance), and form (e.g., bell-shaped, bimodal).
    - Can be visually represented through histograms, box plots, and probability density functions.
    - Includes skewness as one of its characteristics but also encompasses other aspects like kurtosis (how peaked or flat the distribution is), spread (variance), and central tendency.

3. Measurement

- **Skewness:**

  Measured as a single statistic, with values indicating:

  - **0**: Symmetrical (no skewness)
  - **Positive**: Right (positive) skewness
  - **Negative**: Left (negative) skewness

- **Distribution:**

  Measured and analyzed through multiple characteristics, including:

  - **Mean, Median, Mode**: Central tendency
  - **Variance, Standard Deviation**: Spread
  - **Kurtosis**: Peakedness or flatness
  - **Shape**: General form (e.g., normal, skewed, uniform)

4. Impact on Data Interpretation

- **Skewness:**

  - Affects how we interpret the mean and median. In a skewed distribution, the mean may not be the best measure of central tendency because it is affected by extreme values.
  - Helps to understand whether data transformations (e.g., logarithmic) might be needed to normalize data.

- **Distribution:**

  - Gives a complete picture of the data, helping to understand the behavior and properties of the dataset.
  - Guides statistical analysis techniques, including hypothesis testing, confidence intervals, and regression models.

Summary:

- **Skewness** is a measure of the asymmetry of a distribution, providing insight into how much and in what direction the data is skewed.
- **Distribution** is a broader concept that describes the entire structure of data, including its spread, shape, and central tendency. Skewness is just one aspect of distribution.