Lab 3: Data Charts – Univariate Analysis

Prelab Questions

1. **Define univariate analysis and give an example.**

   Univariate analysis involves examining a single variable to understand its distribution, central tendency, and spread. It helps in identifying patterns, trends, and outliers in the data.

   - Example: Analyzing the average height of students in a class, where the variable is "height," and you calculate metrics like mean, median, and standard deviation.

2. **Why are histograms preferred for visualizing continuous data?**

   - Histograms are preferred for visualizing continuous data because they display the distribution of data by grouping it into intervals (bins), allowing for a clear representation of the data's frequency and shape.
   - This helps identify patterns such as skewness, peaks, and spread, which are essential for understanding the underlying distribution of continuous variables.

3. **List the key differences between bar charts and histograms.**

BAR CHART	HISTOGRAM
Used for categorical data	Used for continuous data, showing the distribution of data across intervals (bins).
Bars are separate and do not touch each other.	Bars are adjacent and touch each other, representing continuous ranges.
Compares different categories or groups.	Shows the frequency distribution of continuous data.
Focuses on comparing the count or frequency of different categories.	Focuses on understanding the shape, spread, and central tendency of continuous data

4. **How can outliers affect univariate visualizations?**

   Outliers can significantly affect univariate visualizations by distorting the representation of data. They can:

   - **Skew the Distribution**: Outliers can cause the data to appear skewed, making the distribution look less symmetric.
   - **Inflate Measures of Spread**: Outliers can increase the range, variance, and standard deviation, which may lead to misinterpretation of the data's spread.

- **Mislead Central Tendency**: Outliers can influence the mean, making it higher or lower than the true central tendency. This can give a false impression of the data's center.
- **Mask Patterns**: In some cases, outliers can obscure patterns, making it harder to detect trends or clusters in the data.

5. **What are the benefits of adding a kernel density estimate (KDE) to histograms?**

   Benefits of adding a KDE to histograms:

- **Smoothing**: Makes the distribution smoother and easier to interpret.
- **Better Shape Visualization**: Highlights patterns like peaks or skewness.
- **Density Insight**: Provides a clearer view of data density across the range.
- **Small Data Handling**: Reduces sensitivity to bin width, especially for small datasets.
- **Comparative Analysis**: Allows easy comparison of multiple distributions.

In-Lab Details

Objective:

- Visualize and interpret single-variable distributions using histograms and KDE plots.

Resources:

- Python (Jupyter Notebook).
- Libraries: Matplotlib, Seaborn.
- Dataset: Student erformance.csv with columns for scores and average score.

Program Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

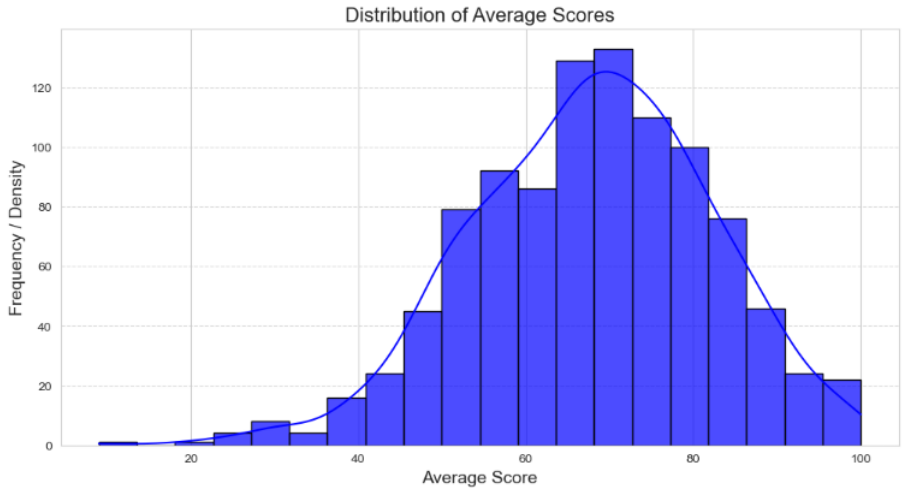
```
# Load the dataset
data = pd.read_csv(r'C:\Users\Rithika\Downloads\archive (9)\Student performance.csv')

# Preview the dataset
data.head()
```

	gender	race_ethnicity	parental_level_of_education	lunch	test_preparation_course	math_score	reading_score	writing_score	total_score	average_score
0	0	group B	bachelor's degree	1	0	72	72	74	218	72.666667
1	0	group C	some college	1	1	69	90	88	247	82.333333
2	0	group B	master's degree	1	0	90	95	93	278	92.666667
3	1	group A	associate's degree	0	0	47	57	44	148	49.333333
4	1	group C	some college	1	0	76	78	75	229	76.333333

```
# Set visual style
sns.set_style("whitegrid")

# Plot the histogram and KDE
plt.figure(figsize=(12, 6))
sns.histplot(data['average_score'], kde=True, bins=20, color='blue', edgecolor='black', alpha=0.7)

# Add titles and labels
plt.title("Distribution of Average Scores", fontsize=16)
plt.xlabel("Average Score", fontsize=14)
plt.ylabel("Frequency / Density", fontsize=14)

# Add grid lines for better readability
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Show the plot
plt.show()
```

Expected Output:

- Histogram showing score distribution.
- KDE overlay for smoother visualization.



Postlab Questions

1. **How does increasing the number of bins in a histogram affect the visualization?**

   Increasing the number of bins in a histogram can:

- **Reveal More Detail**: It can show finer details of the data distribution, highlighting smaller variations.
- **Increase Noise**: Too many bins can make the histogram appear noisy, with fluctuations that may not represent the true distribution.
- **Reduce Smoothness**: A higher number of bins can make the histogram less smooth, potentially obscuring the overall trend or pattern.
- **Overfit Data**: It may lead to overfitting, capturing random variations instead of the underlying distribution.

2. **What does the area under the KDE curve represent?**

   - The area under the Kernel Density Estimate (KDE) curve represents the total probability of the data.
   - Since the curve is a probability density function, the area under the entire curve equals 1, which indicates that the total probability of all possible outcomes is 100%.
   - The area under a specific section of the curve represents the probability of data points falling within that range.

3. **Compare the benefits of using KDE over histograms for data visualization.**

KDE	HISTOGRAM
Provides a smooth, continuous curve, making it easier to identify patterns and the overall shape of the data.	Can be jagged and sensitive to bin width, leading to less smooth visualizations.
Represents the data's probability density, giving a clearer understanding of the distribution's underlying structure.	Represents frequency counts in discrete bins, which may not fully capture the data's distribution.
Does not depend on bin width and is less affected by small dataset sizes.	The choice of bin size can significantly impact the appearance and interpretation of the data.
Easier to interpret complex distributions	Can be harder to interpret for complex distributions
Allows for smoother comparisons between multiple datasets by overlaying curves.	Can be cluttered when comparing multiple datasets, especially if bins don't align well.

4. **Why is it important to consider the scale of the x-axis when plotting univariate data?**

- **Accurate Representation**: Ensures data is displayed correctly and not distorted.
- **Clarity**: Helps highlight relevant patterns, especially with large or varied ranges.
- **Comparison**: Ensures meaningful and fair comparisons between datasets.
- **Outlier Management**: Helps control the impact of outliers on the visualization.
- **Visual Balance**: Improves readability by adjusting data spread for better interpretation.

5. **Suggest scenarios where bar charts are more suitable than histograms.**

Bar charts are more suitable than histograms in the following scenarios:

- **Categorical Data**: When comparing distinct categories (e.g., sales by region or product type).
- **Non-Numerical Data**: For visualizing data like survey responses (e.g., Yes/No answers).
- **Discrete Groups**: When data consists of a limited number of discrete values (e.g., number of students in each grade).
- **Comparing Groups**: When comparing different groups side by side (e.g., revenue in different years or countries).
- **Data with Gaps**: When the data has gaps between categories, such as comparing different brands in a market.

DONE BY

Deepika Ganesan – 22BCS019

Durga S – 22BCS026

Nandhini Priya K K -22BCS073

Nithiyaa T – 22BCS082

Priya Dharshini P – 22BCS094

Rithika S – 22BCS103

Sandhiya G – 22BCS109

Santhiya P – 22BCS111