Lab #2: Plotting and Inferring Empirical Distribution

A. Objectives

- Identify variable types.
- Plot **empirical distributions using histograms, Kernel Density Estimates (KDE), and Empirical Cumulative Distribution Functions (ECDF).**
- **Interpret** these plots to describe the underlying properties of a dataset, such as central tendency, spread, skewness, modality, and tail behavior.
- **Infer** potential underlying theoretical distributions (e.g., Normal, Exponential, Uniform) from the empirical data.
- **Compare** multiple empirical distributions visually to draw preliminary conclusions.
- **Implement** the above using Python's core data science libraries (Pandas, Matplotlib, Seaborn).
- Distinguish between numeric and categorical variables.
- Plot empirical distributions (histogram, ECDF, bar chart).
- Compute summary statistics and interpret shape, spread, skewness.
- Compare empirical distribution with simple theoretical distributions.

B. Theory

Write in short about each type of plots used and their significance**.**

C. Datasets Used & description of dataset

- marks_sem1.csv – Numeric variables like marks & attendance.
- transport_mode.csv – Categorical variable for transportation mode.

D. Observations

1. Summary of each variable types- numeric(discrete / continuous), categorical(nominal/ordinal). Explain why you categorized those variables as such.
2. Numerical Variable Distribution Analysis

2.1 summary stats of data

```
=== DETAILED SUMMARY STATISTICS ===
```

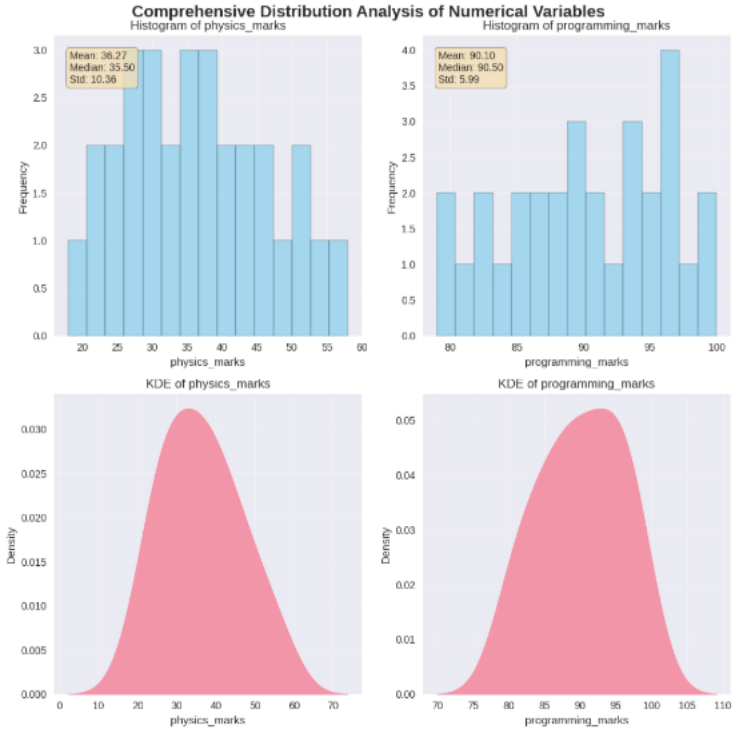
	math_marks	physics_marks	programming_marks	attendance_percent
count	20.000000	20.000000	20.000000	20.000000
mean	67.700000	65.800000	74.750000	83.700000
std	15.235347	16.452803	15.983133	9.744634
min	39.000000	35.000000	44.000000	60.000000
25%	58.000000	57.500000	62.500000	77.750000
50%	68.000000	66.500000	75.500000	84.500000
75%	78.000000	74.500000	86.250000	90.250000
max	92.000000	95.000000	99.000000	98.000000

2.2 Additional stats like variable variance skewness kurtosis IQR

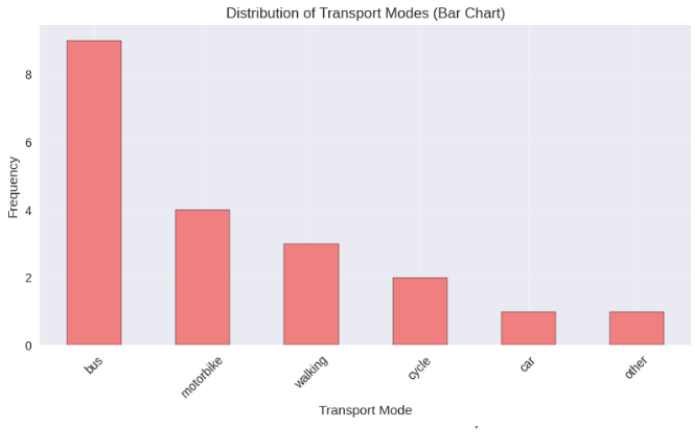
```
=== Additional Statistics ===
```

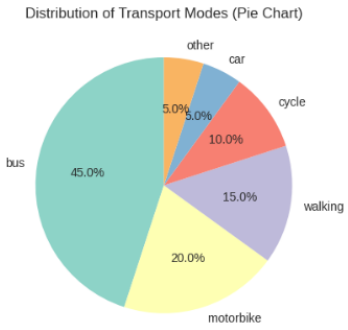
	variable	variance	skewness	kurtosis	IQR
0	math_marks	232.115789	-0.094859	-0.838875	20.00
1	physics_marks	270.694737	-0.100269	-0.612689	17.00
2	programming_marks	255.460526	-0.209887	-0.900030	23.75
3	attendance_percent	94.957895	-0.598484	-0.044569	12.50

2.3 plots for each numerical columns



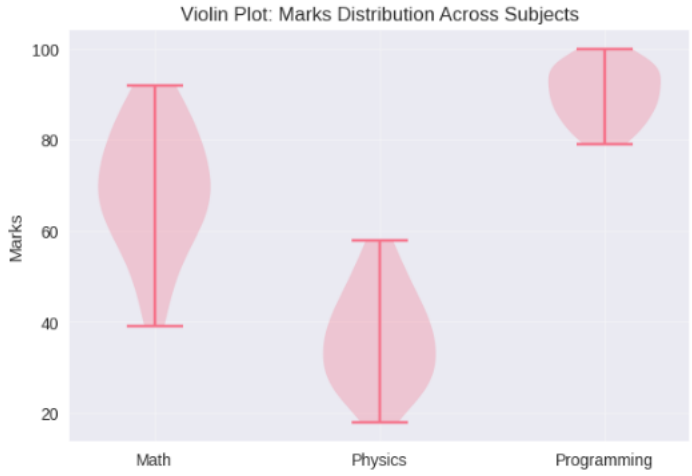
3. Categorical variable analysis

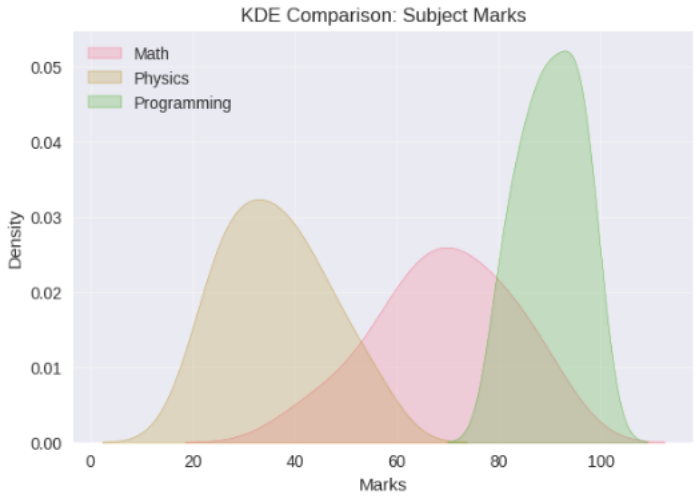




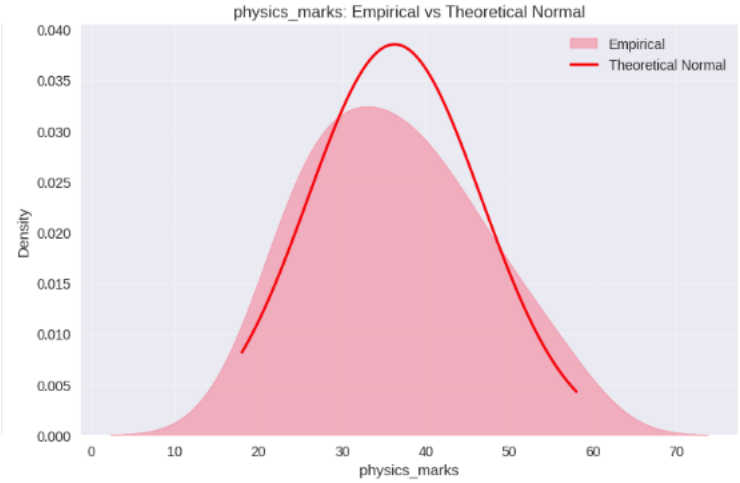

```
=== TRANSPORT MODE DISTRIBUTION ===
transport_mode
bus          9
motorbike    4
walking      3
cycle        2
car          1
other        1
Name: count, dtype: int64

Most common transport mode: bus (9 students)
Least common transport mode: other (1 students)
```

4. Comparative analysis

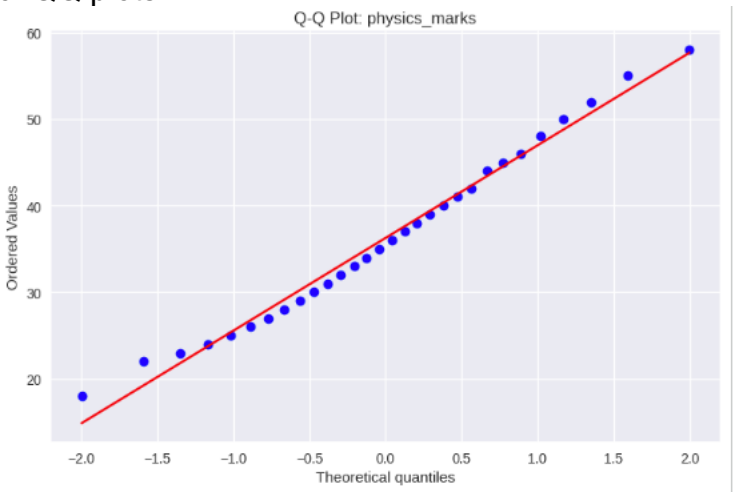




5. Theoretical distribution inference for each of them



6. QQ plots



7. Normality tests (Shapiro-Wilk test) results
8. Comprehensive distribution insights for each of them

8.1. Mathematics marks:
- Mean: 69.20, Median: 69.50
- Standard Deviation: 13.48
- Skewness: -0.29
- Moderately left-skewed distribution with average performance

8.22. Physics marks:
- Mean: 36.27, Median: 35.50
- Standard Deviation: 10.36
- Skewness: 0.30
- Similar to math but slightly lower average with moderate negative skewness

E. Conclusion

Submission Guidelines

Complete the Jupyter notebook with observations. Document each step with comments and explanations. Submit the well documented notebook file (.ipynb) to bidur(@)gces.edu.np with email

subject: BECE2022 - CMP 360 – **Lab#2**.

The email body must have your: Name, Class Roll Number, Lab Number and Lab Title.

Use your gces email to complete the submission.

Hardcopy Submission (Individual **Handwritten**):

- Lab Title
- Objectives
- Theory
- Observation (provide plots and other observations for '**programming',** you can submit all if you wish)
- Conclusion