Project 1: Descriptive Statistics

1. Forced Expiratory Disease (FEV) is an index of pulmonary functions that measures the volume of air expelled after 1 sec of constant effort. Data set FEV.DAT contains determination of FEV in 1980 on 654 children ages 3 to 19 who were seen in the Childhood Respiratory Disease (CRD) study in East Boston. These data are part of a longitudinal study to follow the change in pulmonary function over the time in children.

i. For each of the variable (other than ID), obtain appropriate descriptive statistics (both numeric (mean, sd, quantiles etc) and graphic (bar, hist, boxplot, lineplot etc)).
ii. Use both numerical summary and graphic summary to assess the relationship of FEV to age, height, and smoking status. (Do this separately for both boys and girls). (Use the hint)
iii. Compare the patterns of growth of FEV by age for boys and girls. Are there any similarities? Any differences? Explain the pattern in details.
Hint: Compute the mean FEV by age group (3-4, 5 – 9, 10-14, 15-19) separately for boys and girls and plot the mean FEV by age.

2.

This “Water Potability” dataset contains water quality measurements and assessments related to potability, which is the suitability of water for human consumption. The dataset's primary objective is to provide insights into water quality parameters and assist in determining whether the water is potable or not. Each row in the dataset represents a water sample with specific attributes, and the "Potability" column indicates whether the water is suitable for consumption.

Columns:

pH: The pH level of the water.
Hardness: Water hardness, a measure of mineral content.
Solids: Total dissolved solids in the water.
Chloramines: Chloramines concentration in the water.
Sulfate: Sulfate concentration in the water.
Conductivity: Electrical conductivity of the water.
Organic_carbon: Organic carbon content in the water.
Trihalomethanes: Trihalomethanes concentration in the water.
Turbidity: Turbidity level, a measure of water clarity.
Potability: Target variable; indicates water potability with values 1 (potable) and 0 (not potable).

Site: samples collected from three different sites (1,2,3)

Objective:
The main objective of this dataset is to assess and predict water potability based on water quality attributes. It can be used for evaluating the safety and suitability of water sources for human

consumption, making informed decisions about water treatment, and ensuring compliance with water quality standards.

i. For each of the variable obtain appropriate descriptive statistics (both numeric (mean, sd, quantiles etc) and graphic (bar, hist, boxplot, lineplot etc)).
ii. Use both numerical summary and graphical summary to assess the relationship of ‘potability’ to the different qualities of the water.

iii. Use both numeric and graphic measures to assess the relationship of ‘potability’ to the different qualities of the water for different ‘sites’.
iii. Compare the patterns of potability based on the results. Are there any patterns? Any differences? Explain the patterns, if any.