**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 seaborn as sns

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

data = pd.read\_csv(r'Q1\_a.csv');

numerical\_column = ' --------------'

print(data.head())

plt.figure(figsize=(18, 6))

plt.subplot(1, 3, 1)

sns.histplot(data[numerical\_column], bins=30, kde=False)

plt.title(f'Histogram of {numerical\_column}')

plt.xlabel(numerical\_column)

plt.ylabel('Frequency')

plt.subplot(1, 3, 2)

sns.boxplot(x=data[numerical\_column])

plt.title(f'Box Plot of {numerical\_column}')

plt.xlabel(numerical\_column)

plt.subplot(1, 3, 3)

sns.kdeplot(data[numerical\_column], fill=True)

plt.title(f'Density Plot of {numerical\_column}')

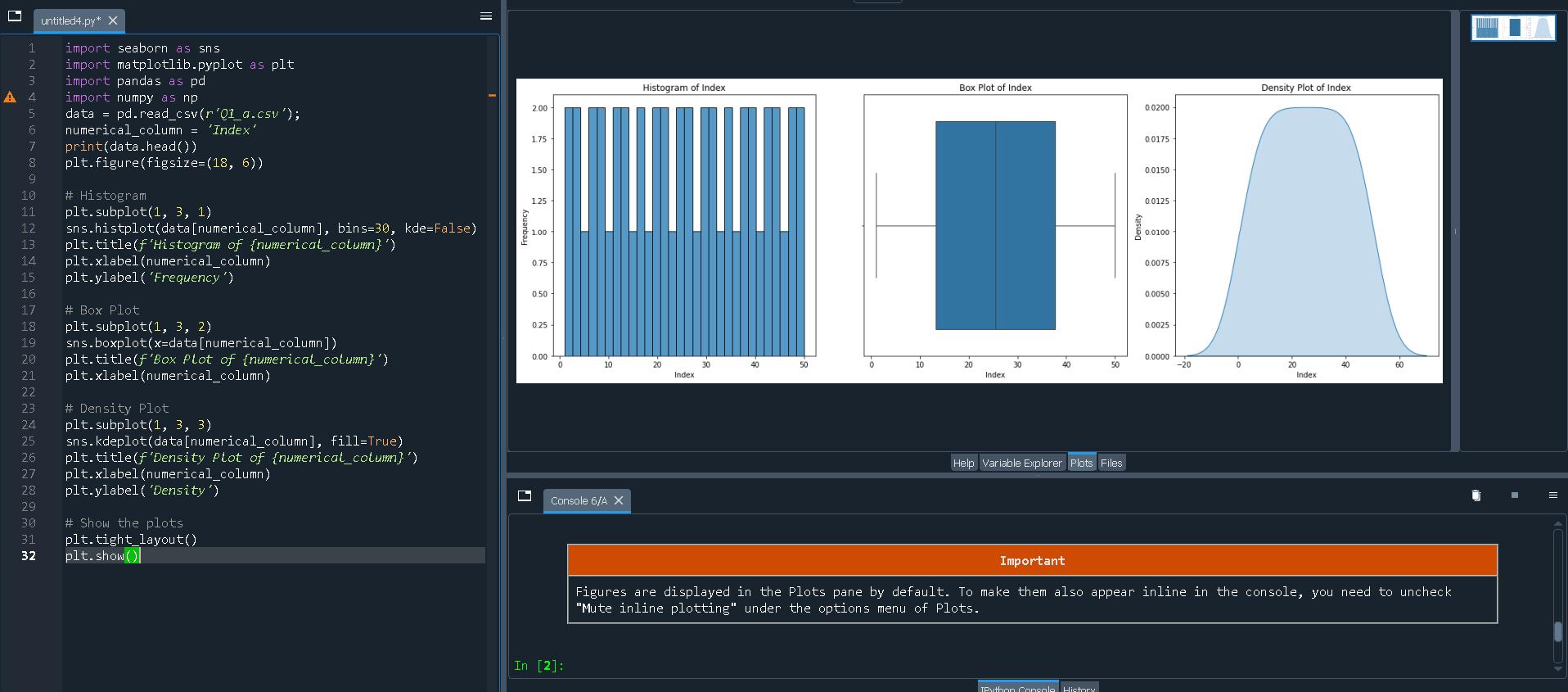
plt.xlabel(numerical\_column)

plt.ylabel('Density')

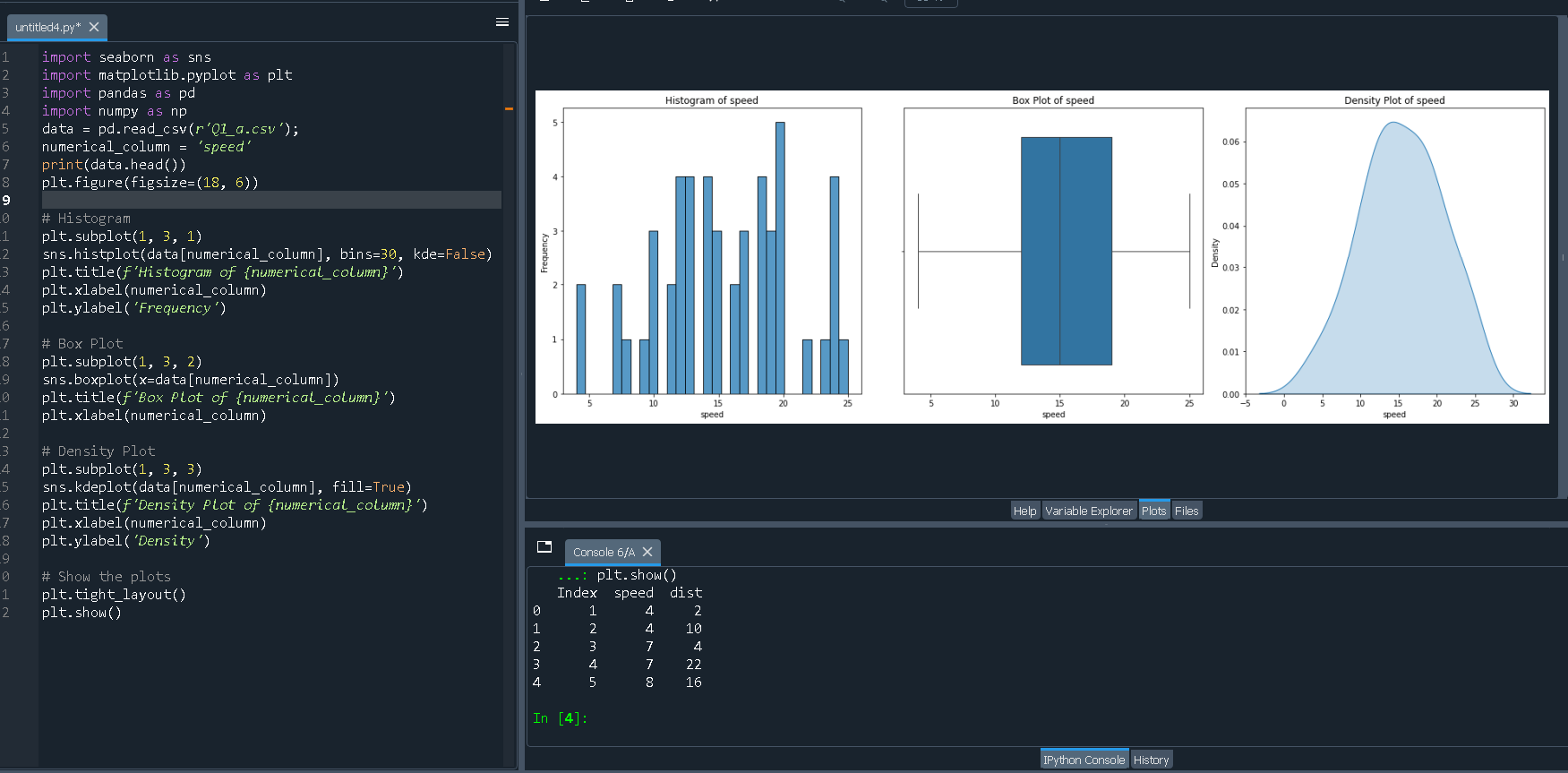
plt.tight\_layout()

plt.show()

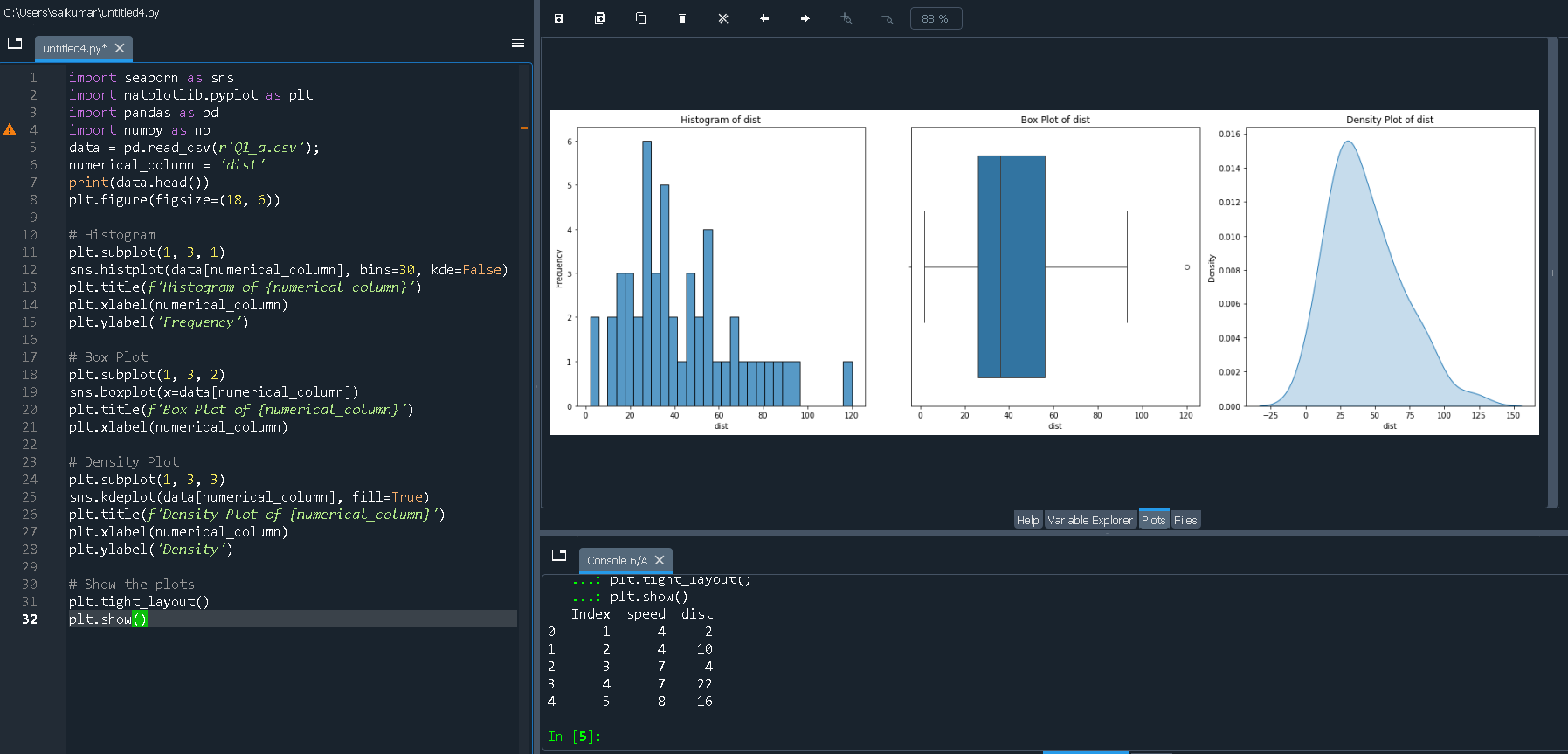
1. numerical\_column = ‘Index’



2. numerical\_column = ‘speed’



3. numerical\_column = ‘dist’



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=3.49𝜎𝑛−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×IQR×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 ?

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?

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?

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 ?

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

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

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