**Final Year B. Tech. (CSE) – I: 2022-23**

**5CS462: PE5 - Data Mining Lab**

**Assignment No. 1**

**PRN: 2019BTECS00077 Date:20 Aug 2022**

**Full name: Biradar Avinash Vishnu**

**---------------------------------------------------------------------------------------------------------**

**Title: Design the data analysis tool (GUI) to perform data analysis and processing Tasks.**

**Objective: Design a GUI such that data analysis and processing can be done according to each selected attribute from the dataset.**

**Dataset Use: Iris, Brest Cancer**

**Introduction & Theory:**

**Measures of Central Tendency**

**What are the measures of central tendency?**  
  
A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

* Mean
* Median
* Mode

**What is the mean?**  
  
The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

**What is the median?**  
  
The median is the *middle* *value* in distribution when the values are arranged in ascending or descending order.

**What is the mode?**  
  
The mode is the *most commonly occurring* *value* in a distribution.

**What is Range?**

The range tells you the spread of your data from the lowest to the highest value in the distribution. It’s the easiest measure of variability to calculate.

**What is Interquartile Range?**

The Interquartile Range gives you the spread of the middle of your distribution.

**What is variance?**

The variance is the average of squared deviations from the mean. A deviation from the mean is how far a score lies from the mean.

**What is Standard Deviation?**

The standard deviation is the average amount of variability in your dataset.

**Code/implementation:**

# importing packages

import streamlit as st

import pandas as pd

import helper

import matplotlib.pyplot as plt

import seaborn as sns

st.title("Data Mining Asssignment No.1 ")

st.text("")

st.text("")

st.sidebar.title("Select the dataset....")

uploaded\_file = st.sidebar.file\_uploader("Choose a file")

if uploaded\_file :

    df  = pd.read\_csv(uploaded\_file)

    #  st.text(df[''])

    #filterning according to attribute and class

    colums=df.columns

    attribute= st.sidebar.selectbox("Select attribute",colums)

    # data\_class=st.sidebar.selectbox("Select Class",df['diagnosis'].unique())

    st.header("Dataset")

    # df=df.loc[df['diagnosis'] == data\_class]

    st.write(pd.DataFrame(df))

    st.text("")

    # Measures of central tendency

    st.header("Measures of central tendency")

    st.text("")

    col1, col2, col3= st.columns(3)

    data=df[attribute].to\_list()

    with col1:

         # Mean

        st.subheader('Mean')

        st.write(helper.mean(data))

    with col2:

        # Median

        st.subheader('Median')

        st.write(helper.median(data))

    with col3:

        # Mode

        st.subheader('Mode')

        st.write(helper.mode(data))

    st.text("")

    col1, col2, col3= st.columns(3)

    with col1:

        #Mid Range

        st.subheader("Mid Range (max+min)/2")

        st.write(round((max(data)+min(data))/2,3))

    with col2:

        # Variance

        st.subheader('Variance')

        st.write(helper.variance(data))

    with col3:

        # Standard Deviation

        st.subheader('standard deviation')

        st.write(helper.stddeviation(data))

    st.text("")

    #Dispersion of data

    st.header("Dispersion of data")

    length=len(df)

    data=df[attribute].to\_list()

    col1, col2, col3= st.columns(3)

    with col1:

        #Range

        st.subheader("Range (max-min)")

        st.write(round(max(data)-min(data),3))

    with col2:

        #Quartile Q1

        st.subheader("Quartile (Q1)")

        Q1=helper.median(data[0:length//2])

        st.write(round(Q1,3))

    with col3:

        #Quartile Q2

        st.subheader("Quartile (Q2)")

        Q2=helper.median(data)

        st.write(round(Q2,3))

    st.text("")

    col1, col2= st.columns(2)

    with col1:

        #Quartile Q3

        st.subheader("Quartile (Q3)")

        Q3=helper.median(data[length//2:])

        st.write(round(Q3,3))

    with col2:

        #IQR

        st.subheader("Interquartile range (Q3-Q1)")

        Q1=helper.median(data[0:length//2])

        Q3=helper.median(data[length//2:])

        st.write(round(Q3-Q1,3))

    st.text("")

    # Five Number Summary

    st.subheader("Five Number Summary")

    col1, col2, col3 ,col4,col5 = st.columns(5)

    data.sort()

    with col1:

        st.text('Min')

        st.write(min(data))

    with col2:

        st.text('Q1')

        st.write(helper.median(data[0:length//2]))

    with col3:

        st.text('Median')

        st.write(helper.median(data))

    with col4:

        st.text('Q2')

        st.write(helper.median(data[length//2:]))

    with col5:

        st.text('max')

        st.write(max(data))

    st.text("")

    # Verify Results:

    st.subheader('Verify Results:')

    st.write(df.describe())

    #GRaphical Representation:

    st.header('Graphical Representation:')

   #Histogram optin x Count

    plt.rcParams['figure.figsize'] = [8, 4]

    st.write("Histogram")

    fig, ax = plt.subplots()

    plt.locator\_params(nbins = 15)

    plt.xlabel(attribute)

    plt.ylabel("count")

    ax.hist(data)

    st.pyplot(fig)

    plt.clf()

    st.text("")

    st.text("")

    st.text("")

    # Scatter plot

    xlabel = st.selectbox("xLabel",df.columns)

    ylabel = st.selectbox("yLabel",df.columns)

    plt.locator\_params(nbins = 10)

    plt.scatter(df[xlabel],df[ylabel], c ="green", s=5)

    plt.xlabel(xlabel)

    plt.ylabel(ylabel)

    # plt.rcParams['figure.figsize'] = [8, 4]

    st.write("Scatter Plot")

    st.pyplot(plt)

    plt.clf()

    indices = []

    ind=0

    for i in df.columns:

        if df.dtypes[i]!=object:

            indices.append(ind)

        ind+=1

    st.text("")

    st.text("")

    st.text("")

    #box plot

    # df.columns = ["sepal length","sepal width","petal length","petal width", "species" ]

    st.write("Box Plot")

    data= df.iloc[:,indices].values

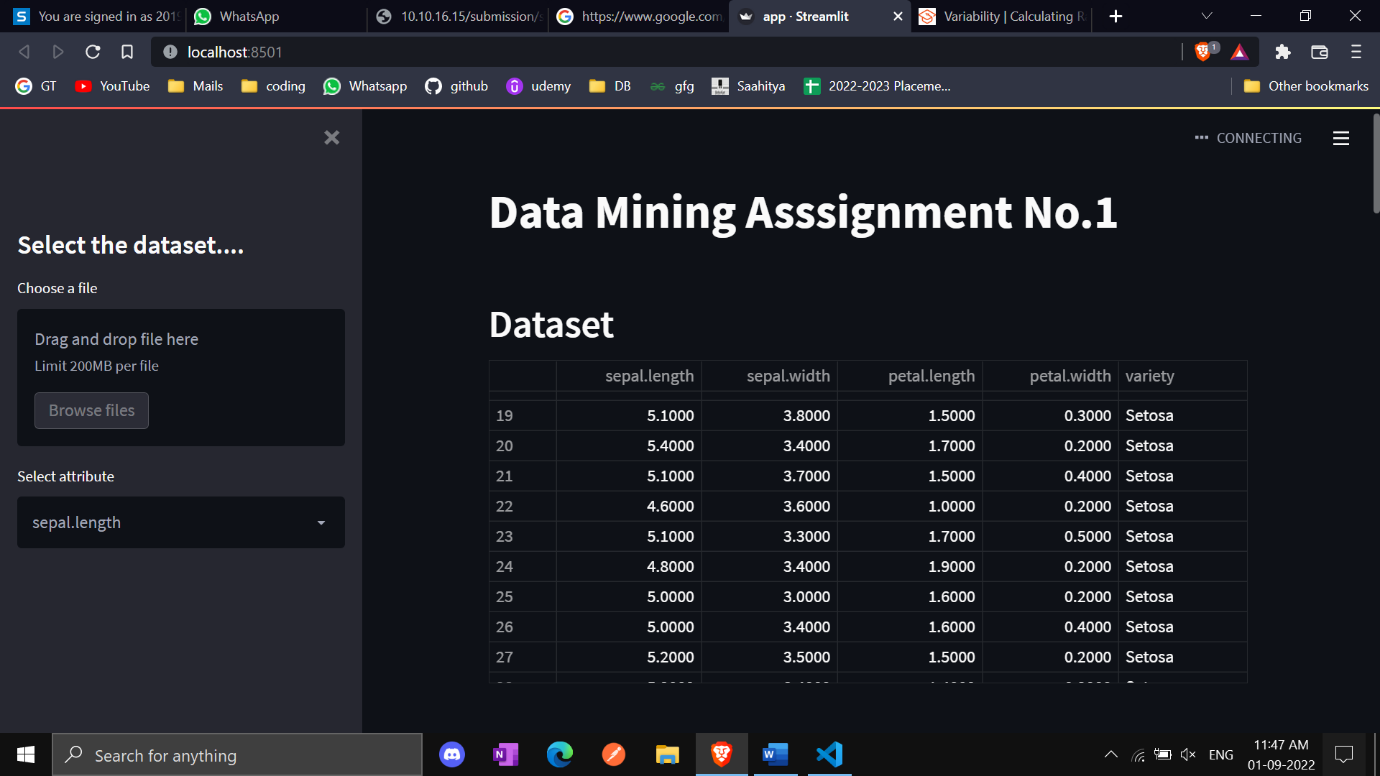
    fig = plt.figure(figsize =(10, 7))

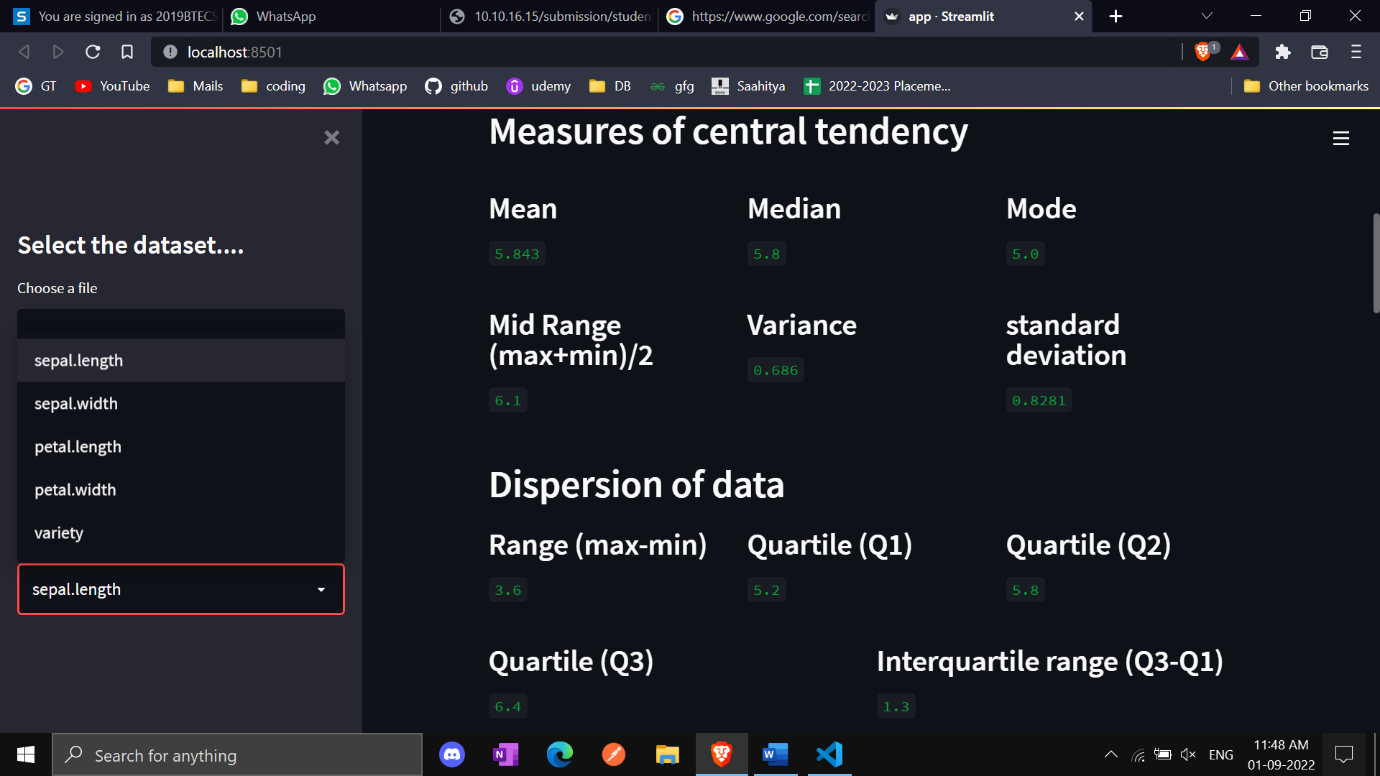
    plt.boxplot(data)

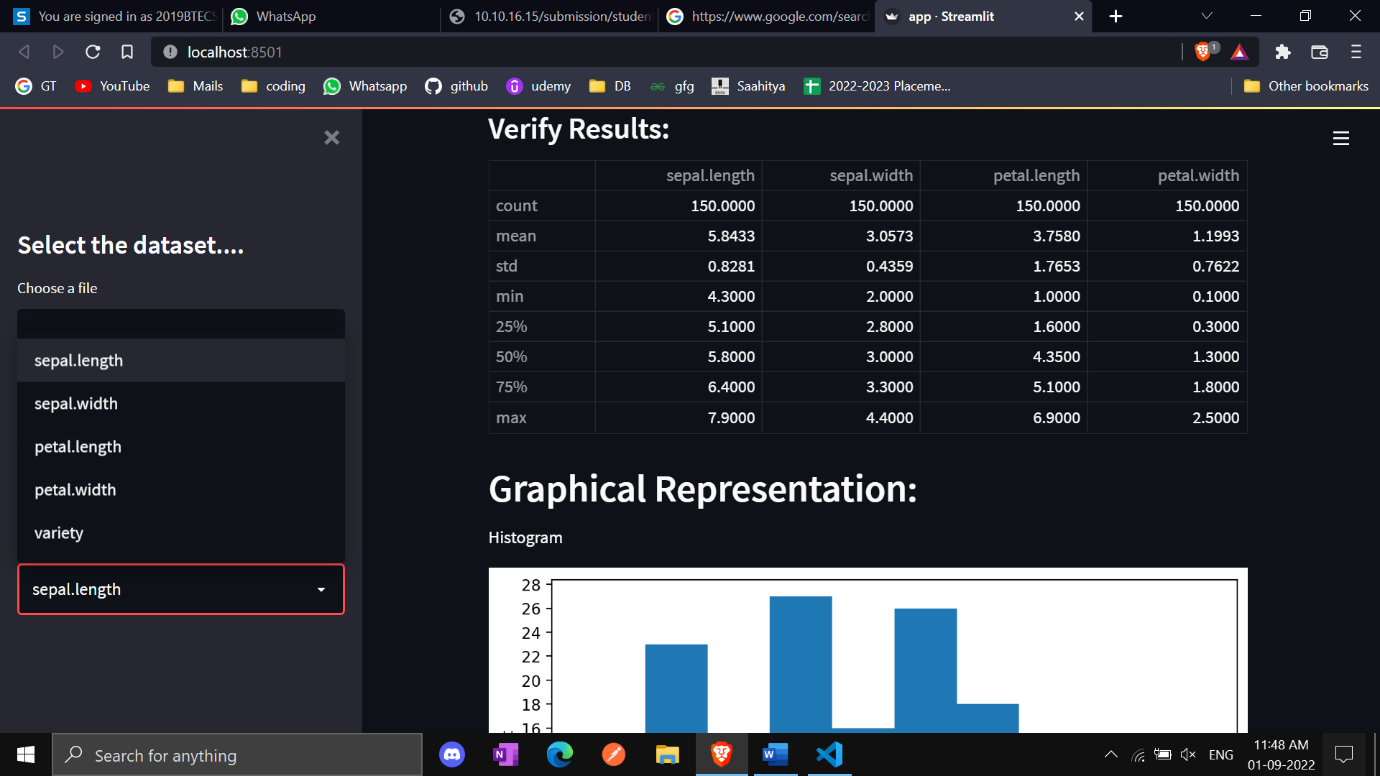
    # plt.show()

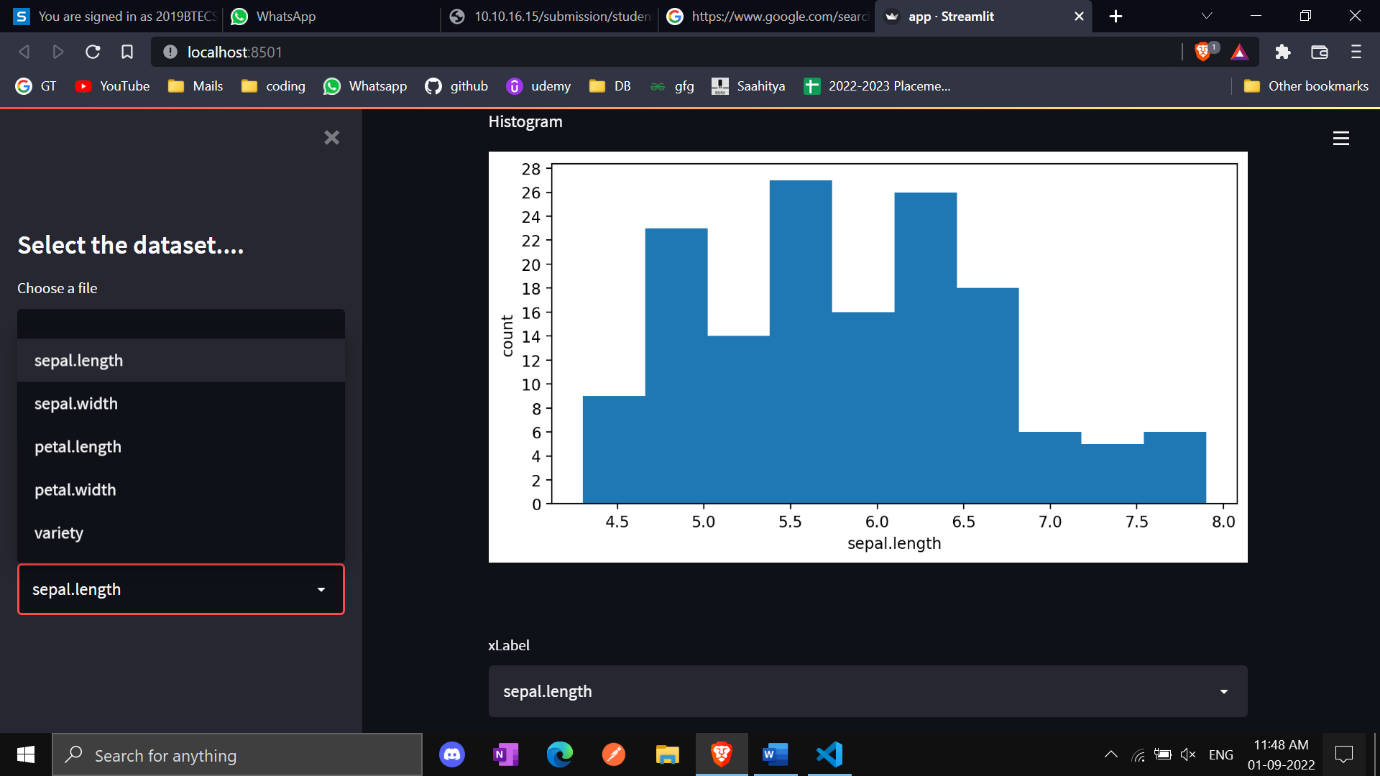
    st.pyplot(plt)

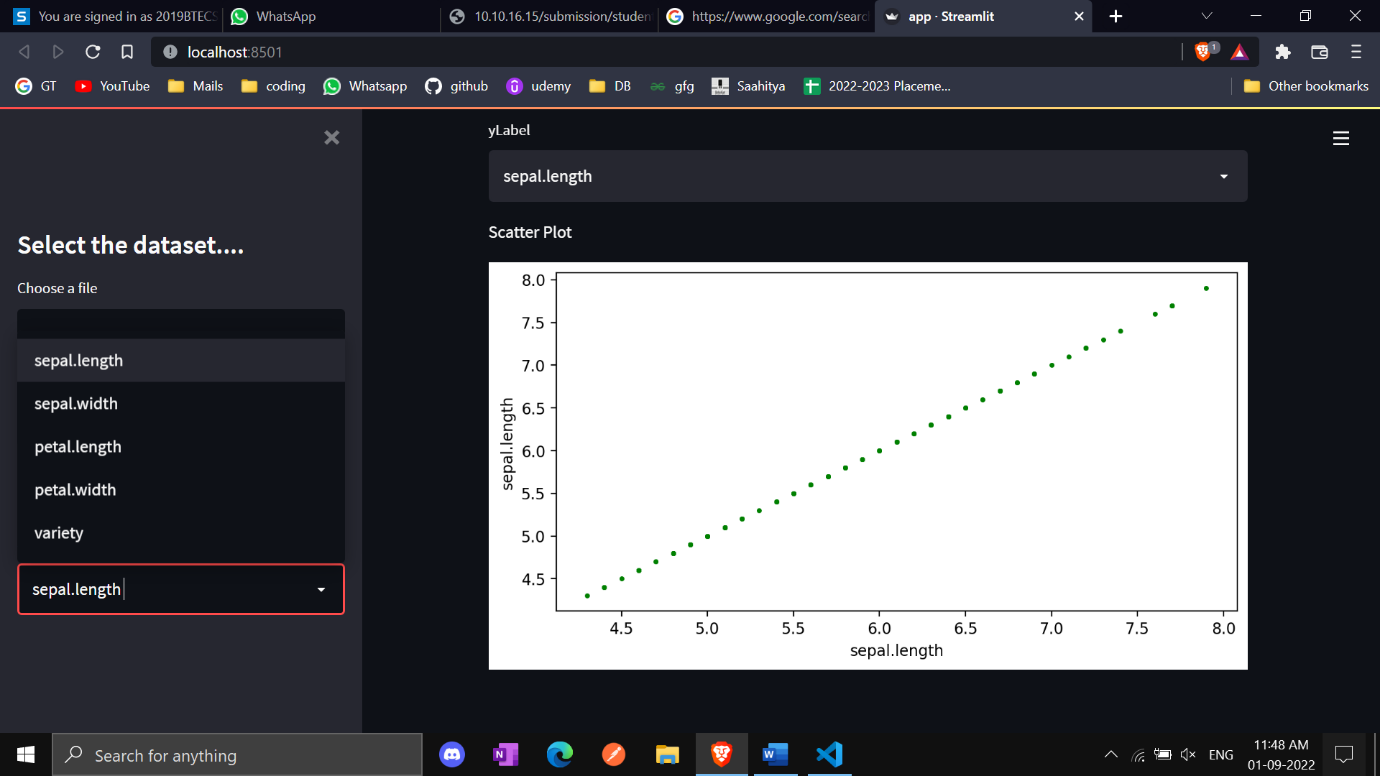
**Output:**

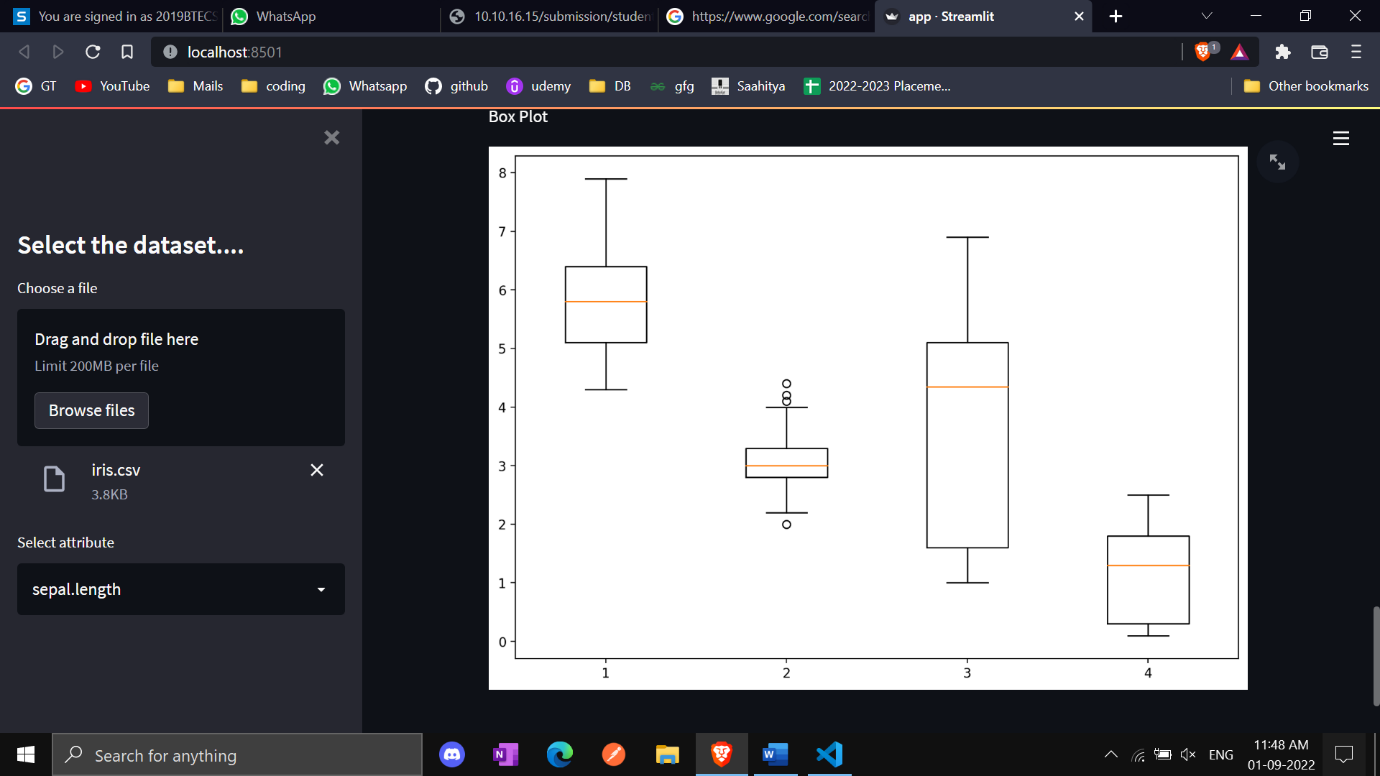
****

****

****

****

****

****