Dataset:

The dataset I chose for the project is **"Iris Dataset".** Dataset contains columns like sepal length, petal length, sepal width, petal width and species type. The numerical values in the dataset are in **cm** measurements. The dataset contains 150 rows and 5 columns.

Study Design:

The purpose of this study is to compare the four characteristics of the three distinct species of iris flowers in terms of their lengths and widths of sepal's and petal's (iris-setosa, iris-versicolor, and iris-virginica).

Data Collection:

The data for this study is already available in the iris dataset. The dataset contains 150 observations of iris flowers, with 50 observations for each of the three species.

Themes to be explored:

From this dataset we can explore themes like finding the significant differences between 3 type of iris species by using sepal length, petal length, sepal width and petal width.

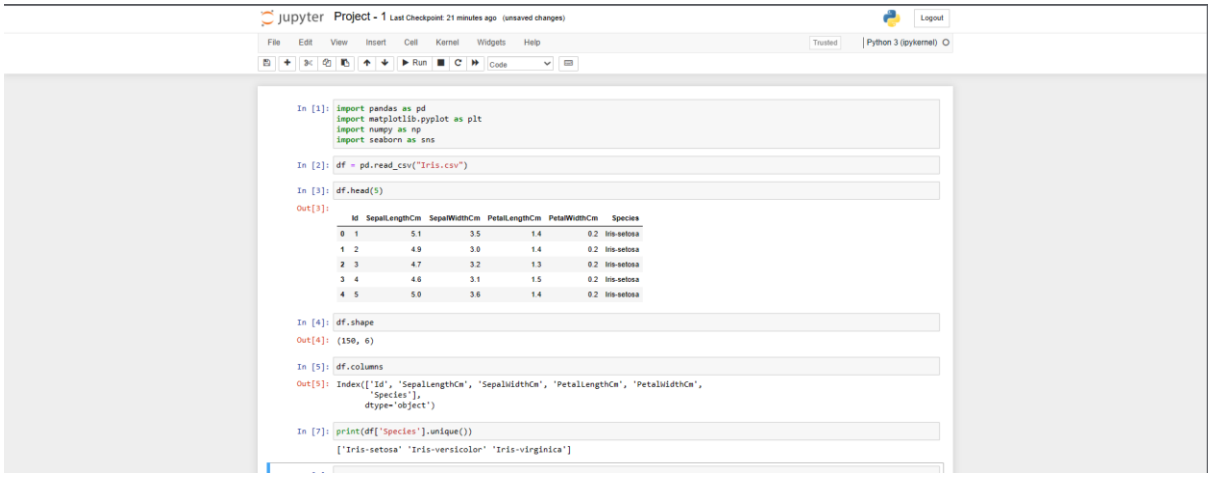
Data analysis:

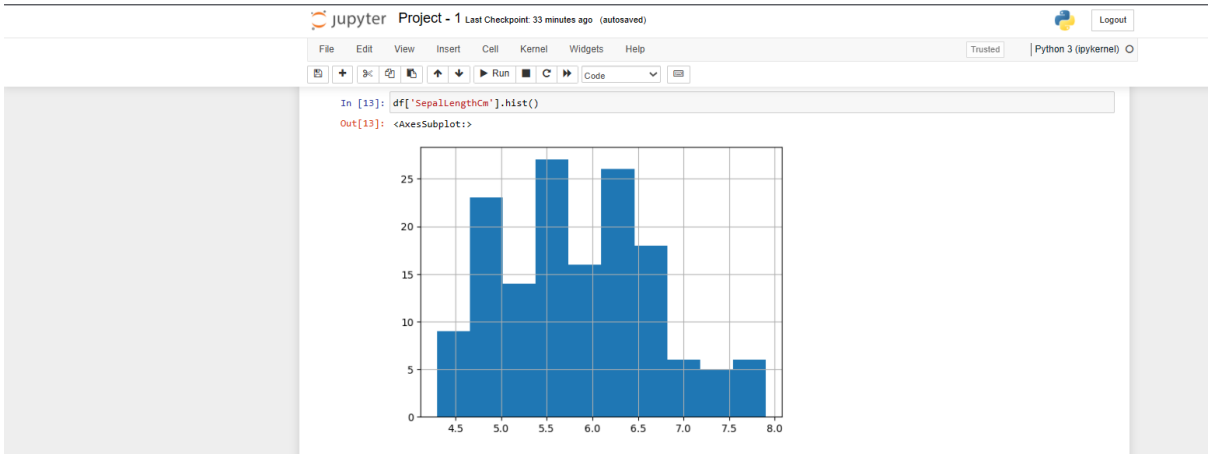
After data collection, descriptive statistics like mean, median, and standard deviation will be used to summarize the data. The variations in characteristics across the three iris species will be investigated using exploratory data analysis (EDA) approaches such frequency distribution, box plots, and scatter plots. The hypotheses will be tested, and the significance of the differences in the traits among the three species will be assessed, using statistical methods including analysis of variance (ANOVA) and post-hoc tests. We will undertake pairwise comparisons across each species to pinpoint the precise variations in each attribute.

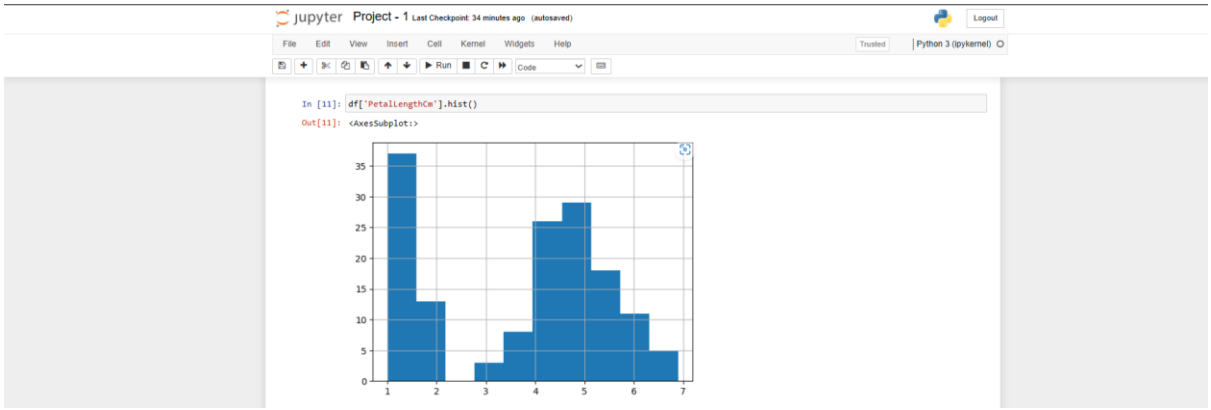
Hypothesis:

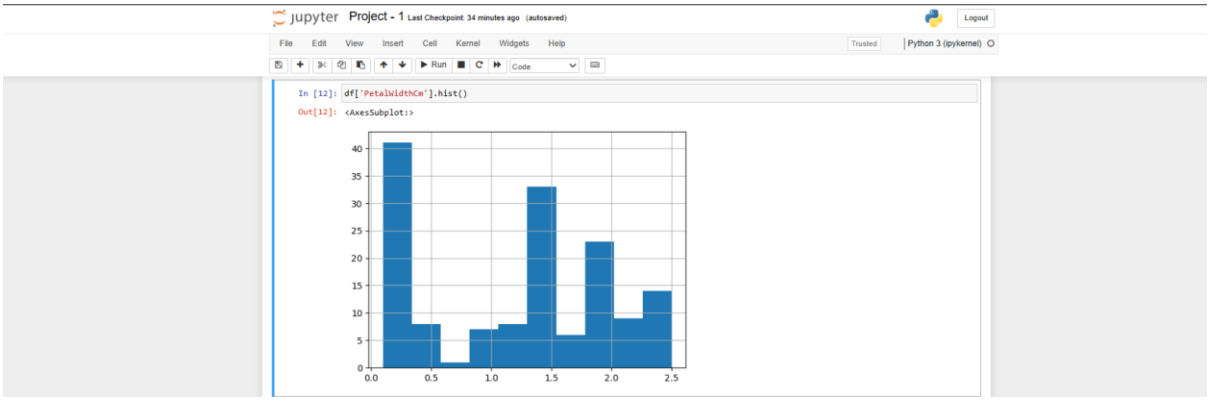
The main hypothesis of this study can be finding the significant differences in lengths and widths of sepal and width among all the three species of iris flowers (iris-setosa, iris-versicolor, and iris-virginica).

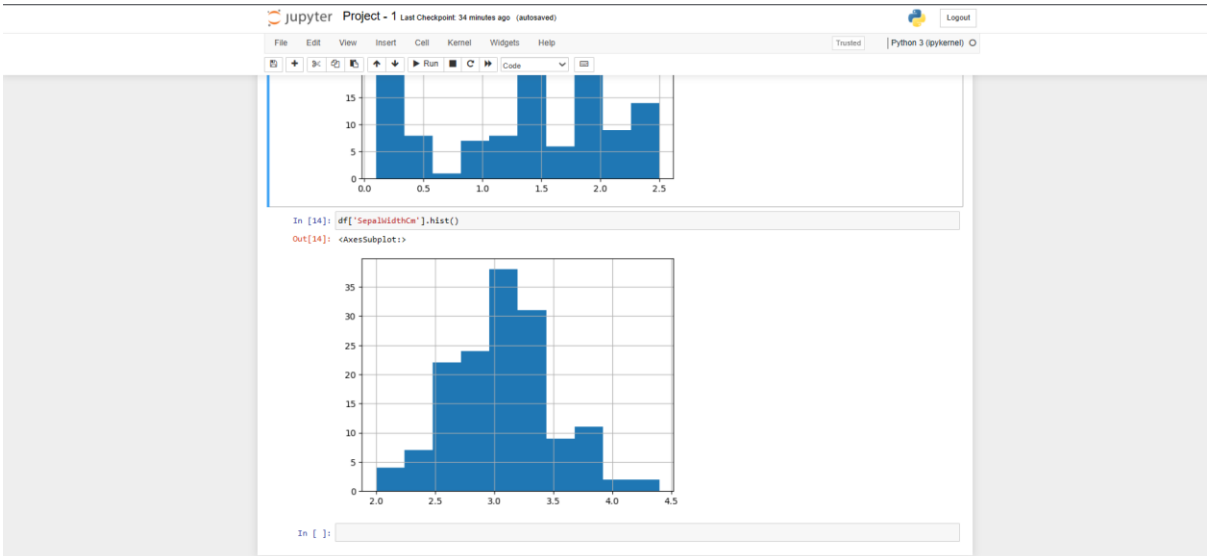
Python Code:


jupyter Project - 1 Last Checkpoint: 21 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [1]:
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
In [2]:
df = pd.read_csv("Iris.csv")
In [3]:
df.head(5)
Out[3]:
Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species
0 1 5.1 3.5 1.4 0.2 Iris-setosa
1 2 4.9 3.0 1.4 0.2 Iris-setosa
2 3 4.7 3.2 1.3 0.2 Iris-setosa
3 4 4.6 3.1 1.5 0.2 Iris-setosa
4 5 5.0 3.6 1.4 0.2 Iris-setosa
In [4]:
df.shape
Out[4]:
(150, 6)
In [5]:
df.columns
Out[5]:
Index(['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm',
'Species'],
dtype='object')
In [7]:
print(df['Species'].unique())
['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']


jupyter Project - 1 Last Checkpoint: 33 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [13]:
df['SepalLengthCm'].hist()
Out[13]:
<AxesSubplot:>
25
20
15
10
5
0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


jupyter Project - 1 Last Checkpoint: 34 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [11]:
df['PetalLengthCm'].hist()
Out[11]:
<AxesSubplot:>
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7


jupyter Project - 1 Last Checkpoint: 34 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [12]: df['PetalWidthCm'].hist()
Out[12]: <AxesSubplot:>


jupyter Project - 1 Last Checkpoint: 34 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [14]: df['SepalWidthCm'].hist()
Out[14]: <AxesSubplot:>
In [ ]: