**PHASE 4**

**DEVELOPMENT PART 2**

**Steps:**

* We continue to build the analysis by creating visualizations and building a predictive model.
* Using Visualization libraries to create histograms, scatter plots, and correlation matrices.
* Here we Build a predictive model such as logistic regression and random forest.
* Using these models to determine water potability based on water quality parameters.

**# IMPORTING PACKAGES**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

**#Visualization using Seaborn**

**#Histogram**

Value = analysis['Hardness'].mean()

plt.figure(figsize=(12,6))

plt.subplot(1,2,1)

plt.hist(analysis['Hardness'],bins=30,color='blue',alpha=0.7)

plt.axvline(Value, color='red', linestyle='dashed', linewidth=2)

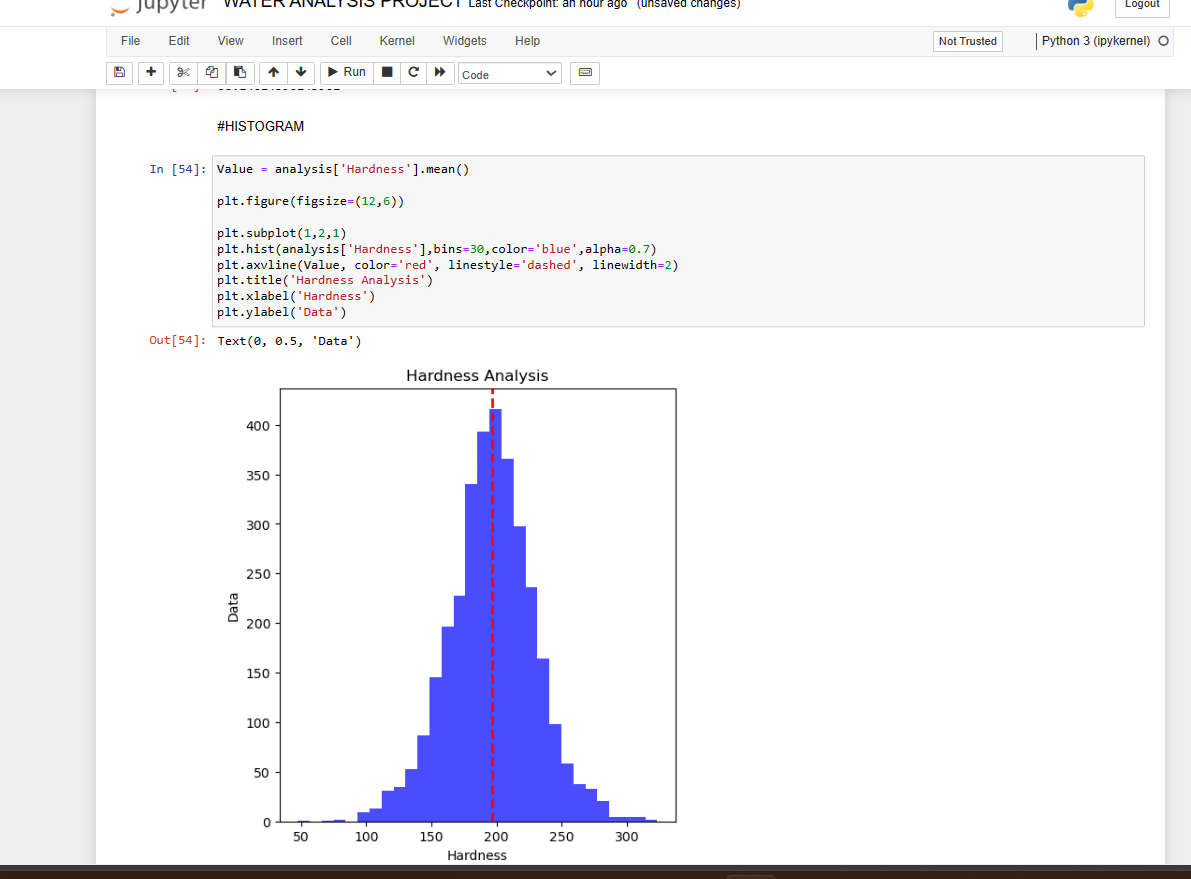
plt.title('Hardness Analysis')

plt.xlabel('Hardness')

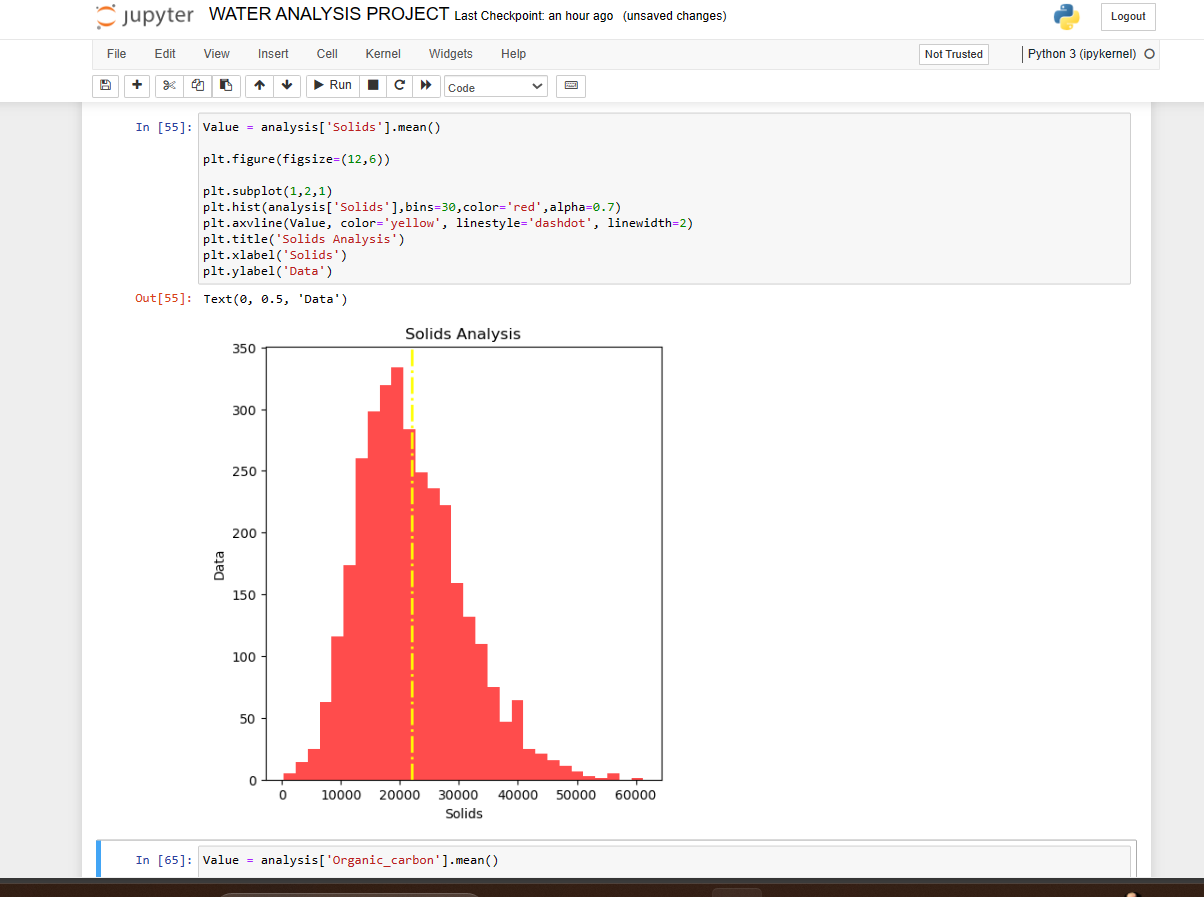
plt.ylabel('Data')

**#Output**

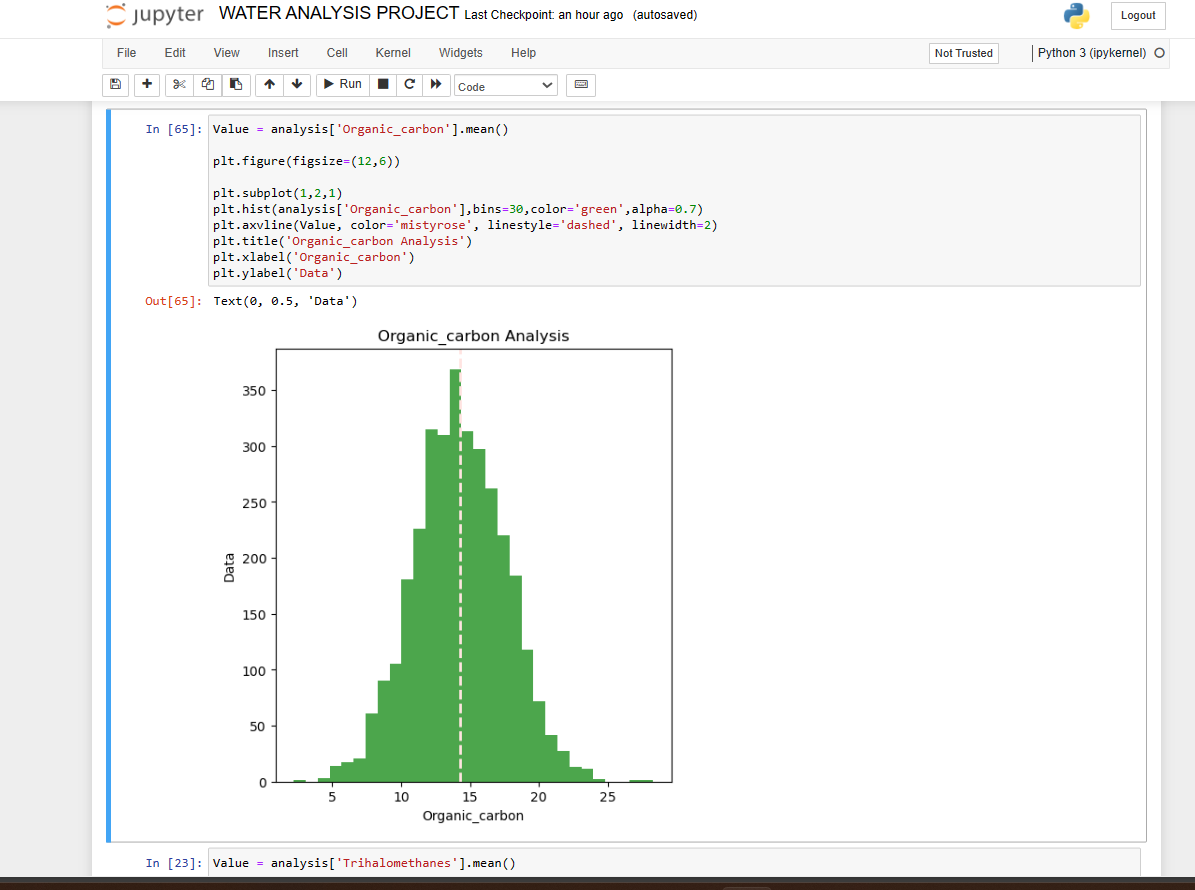
**#Hardness**

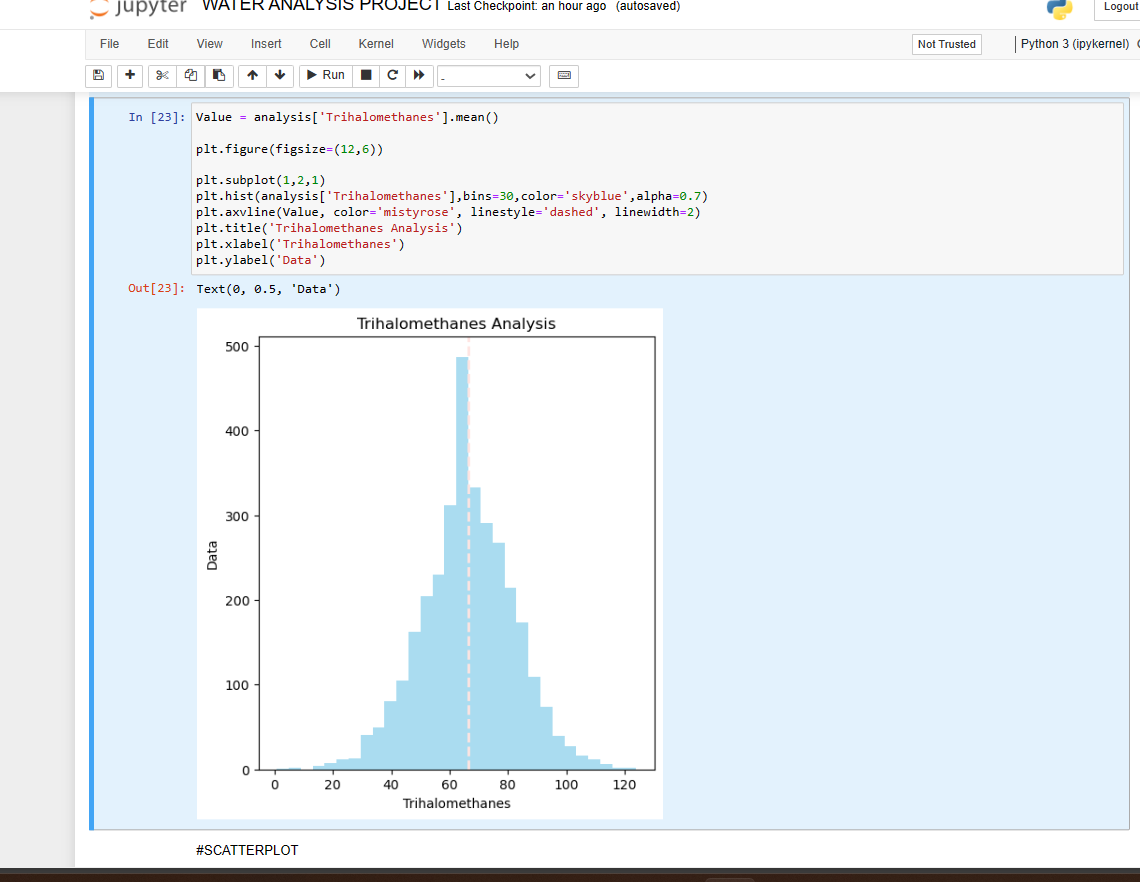
****

**#Solids**

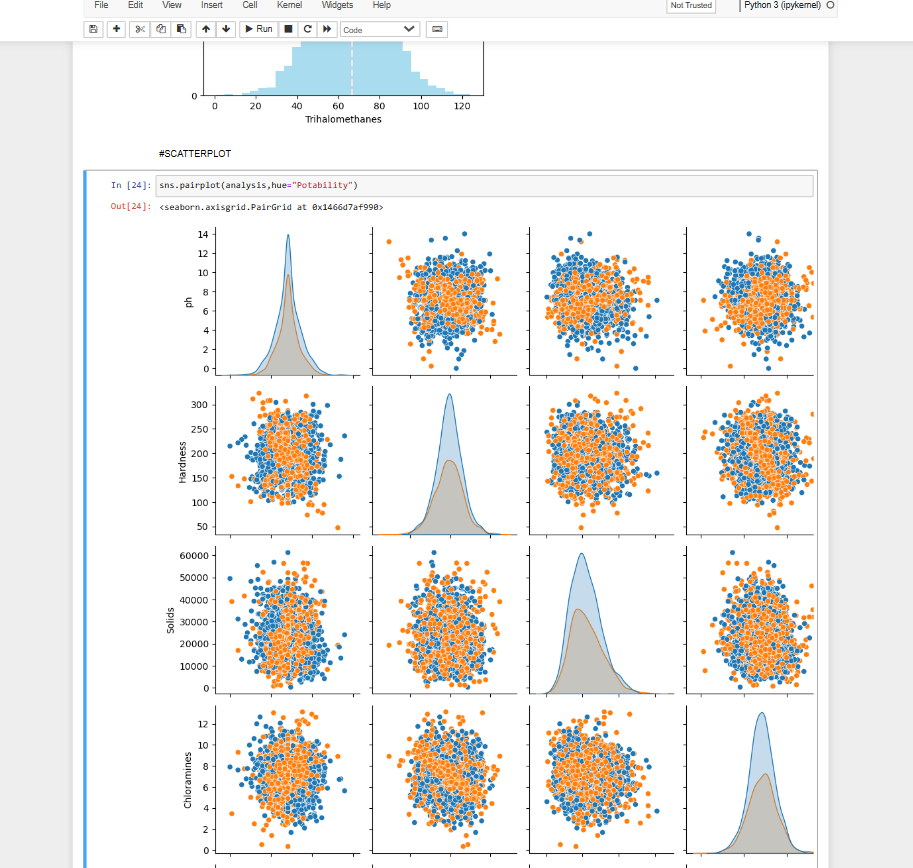
****

**#Organic carbon**

****

**#Trihalomethanes**

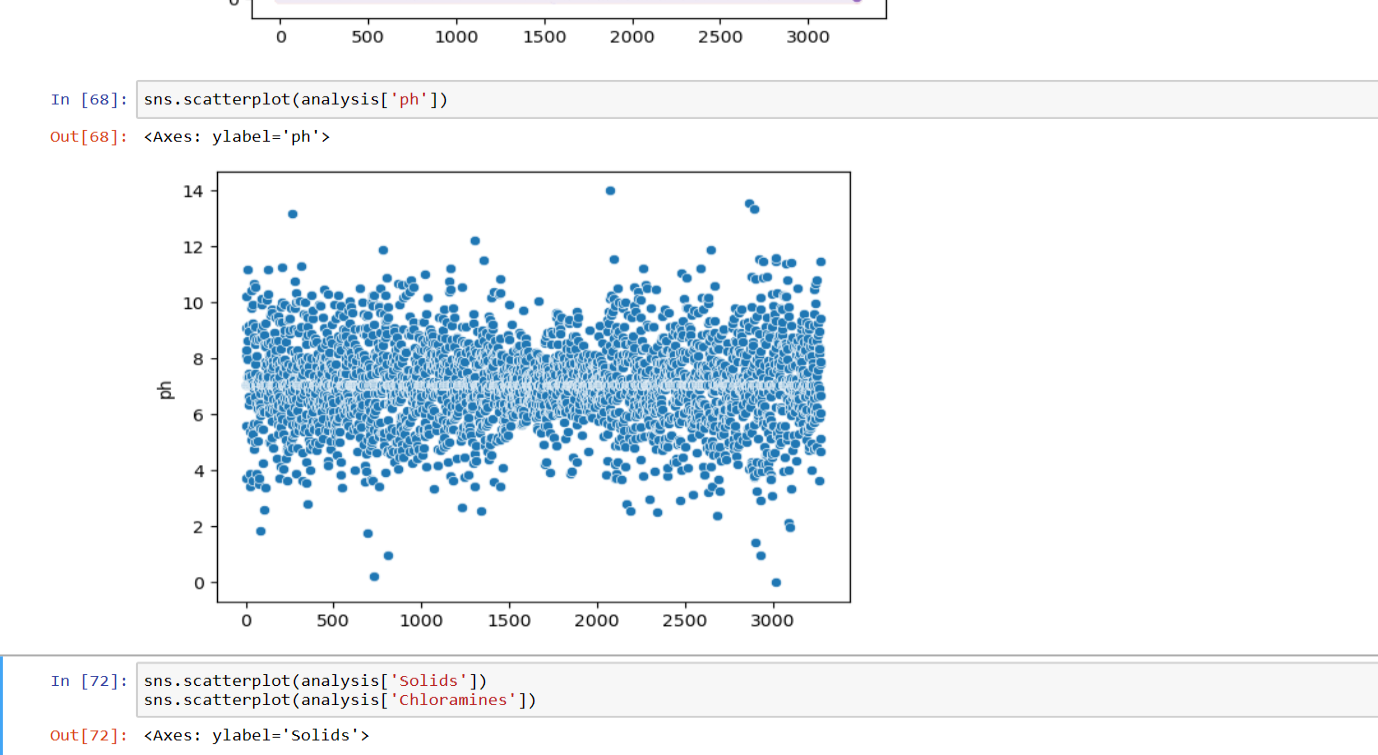
**#Scatterplots**

****

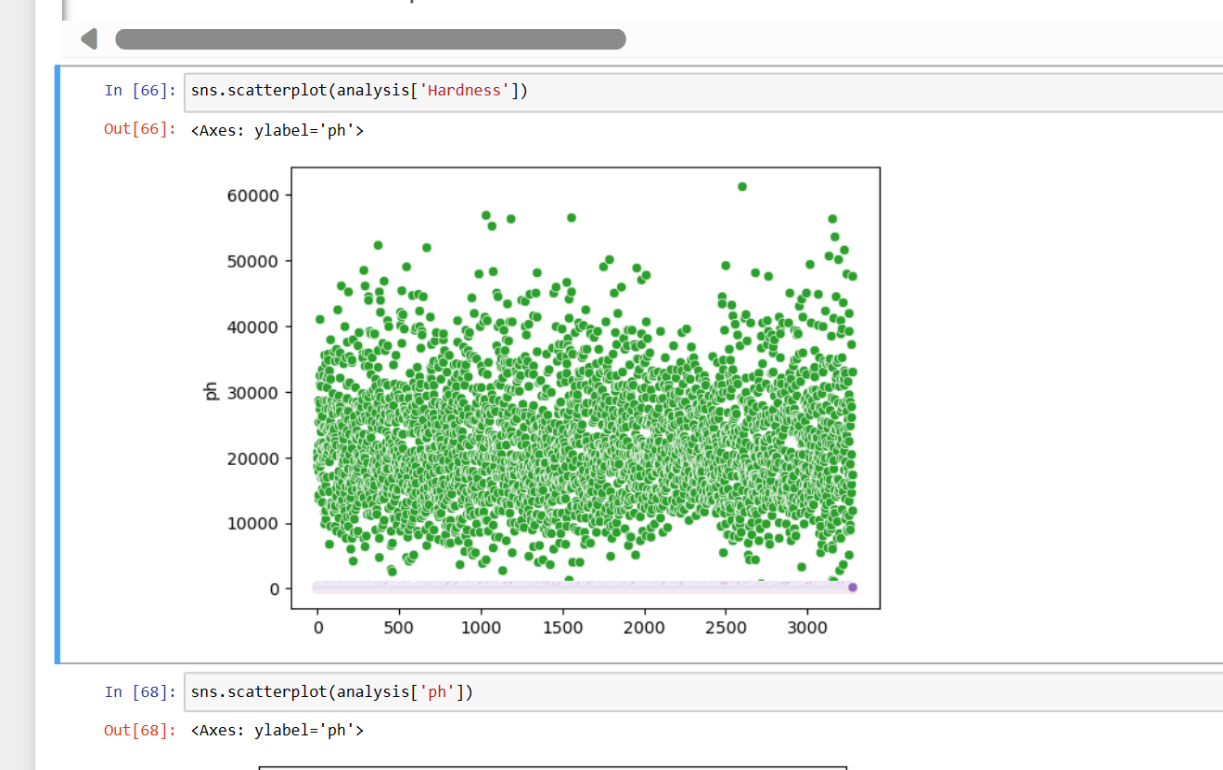
sns.pairplot(analysis,hue="Potability")

**#PH**

sns.scatterplot(analysis['Hardness'])

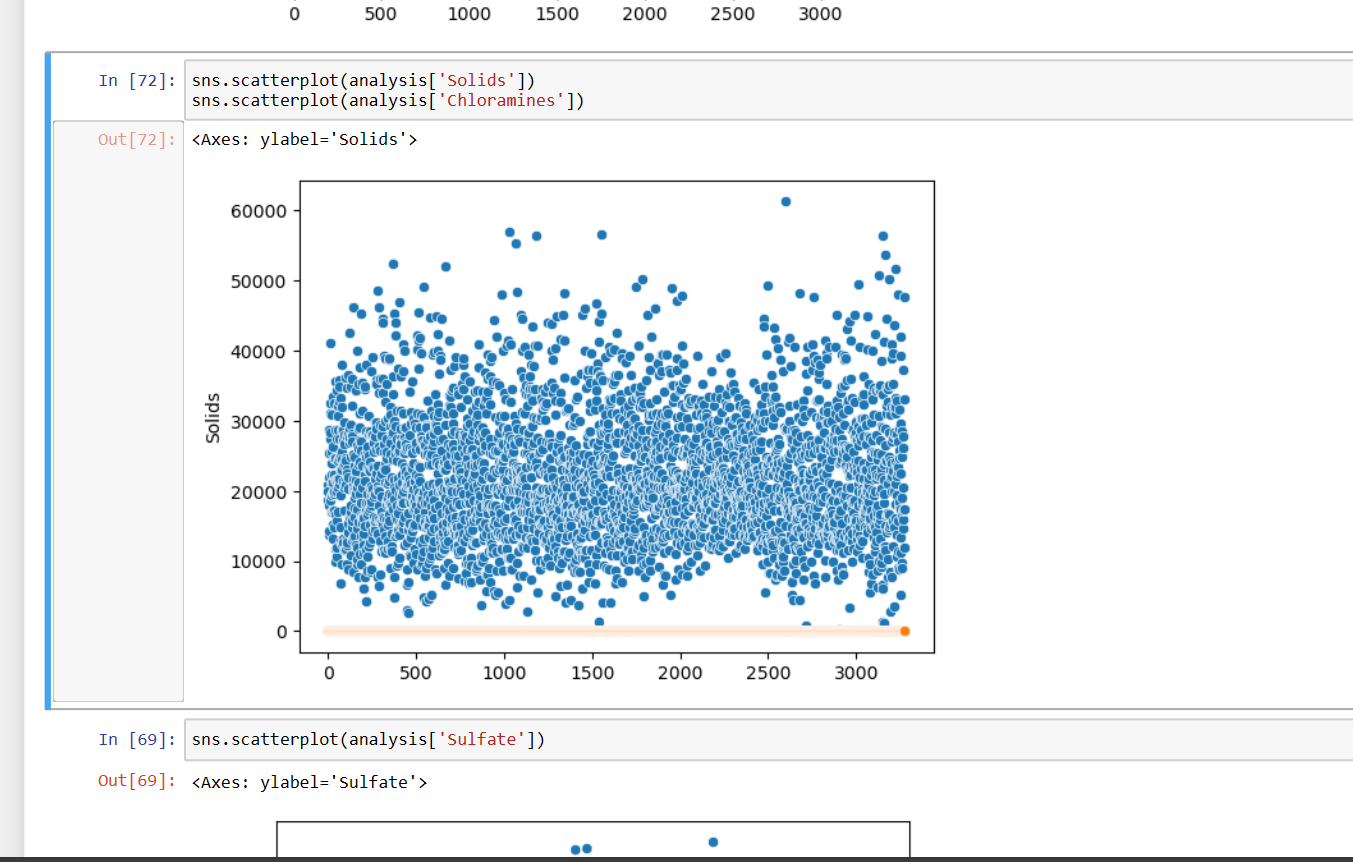
****

**#Hardness**

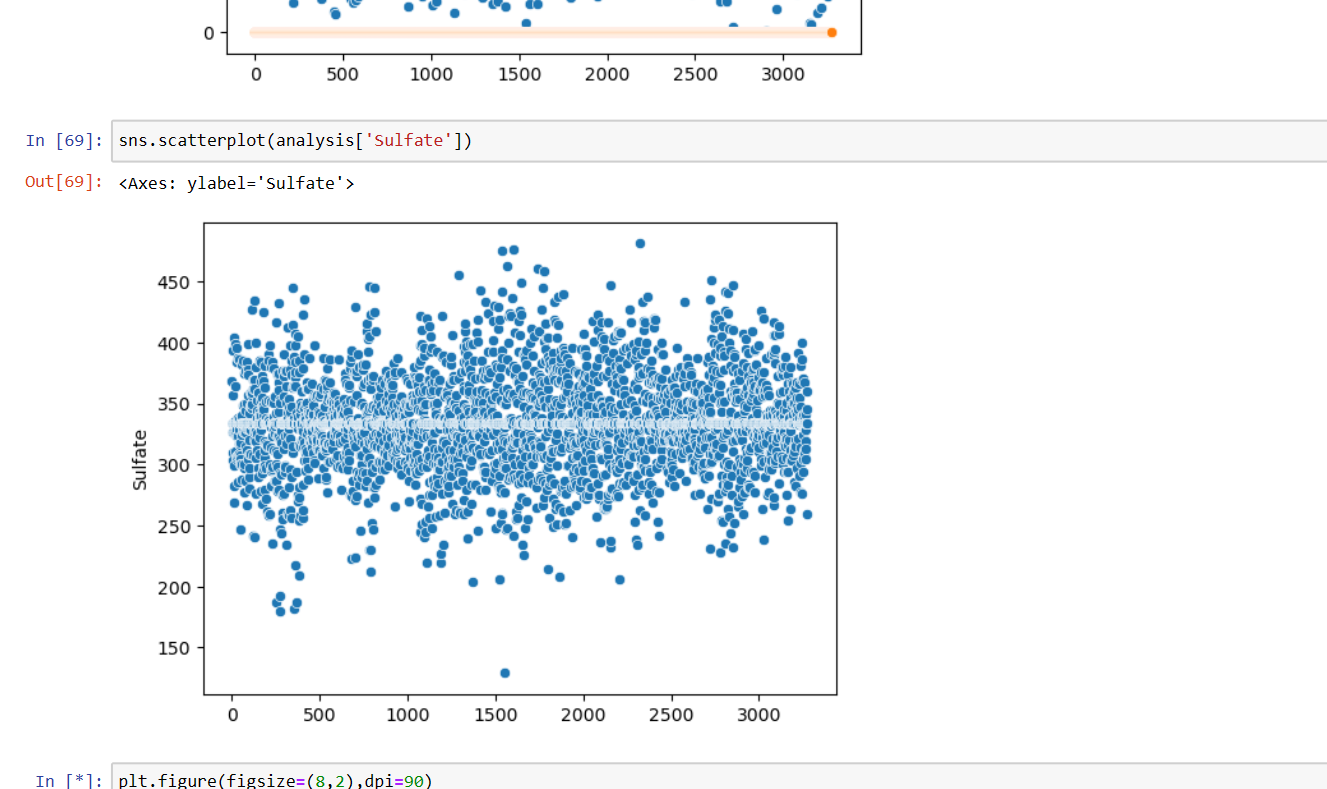
****

**#Solids**

**#Chloramines**

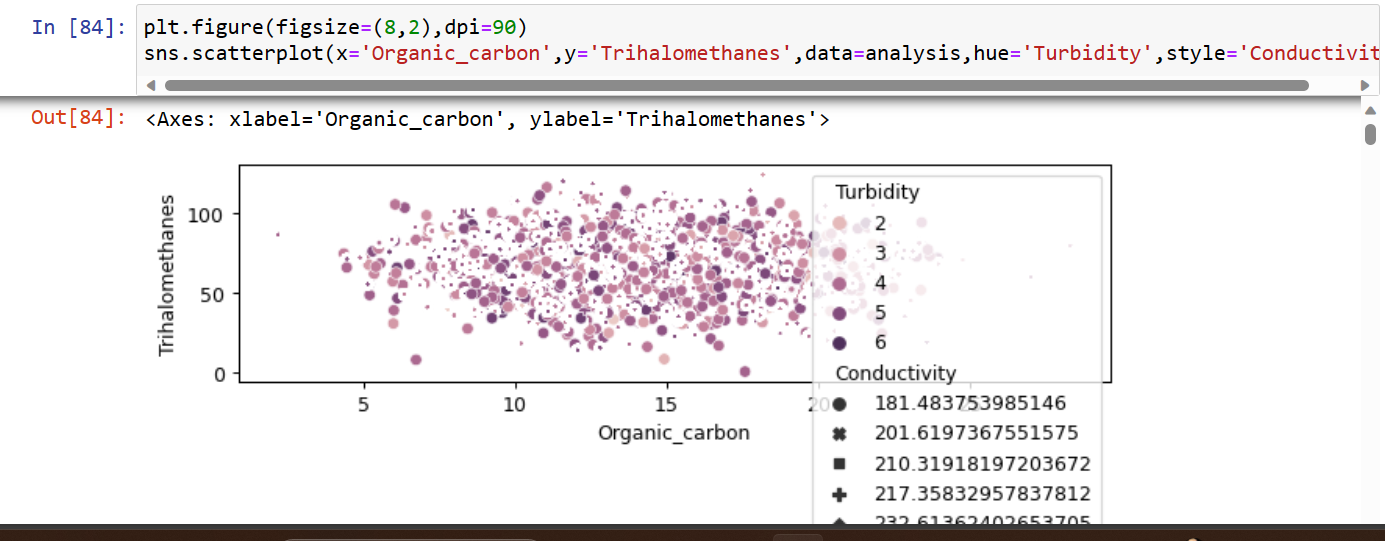
****

**#Sulfates**

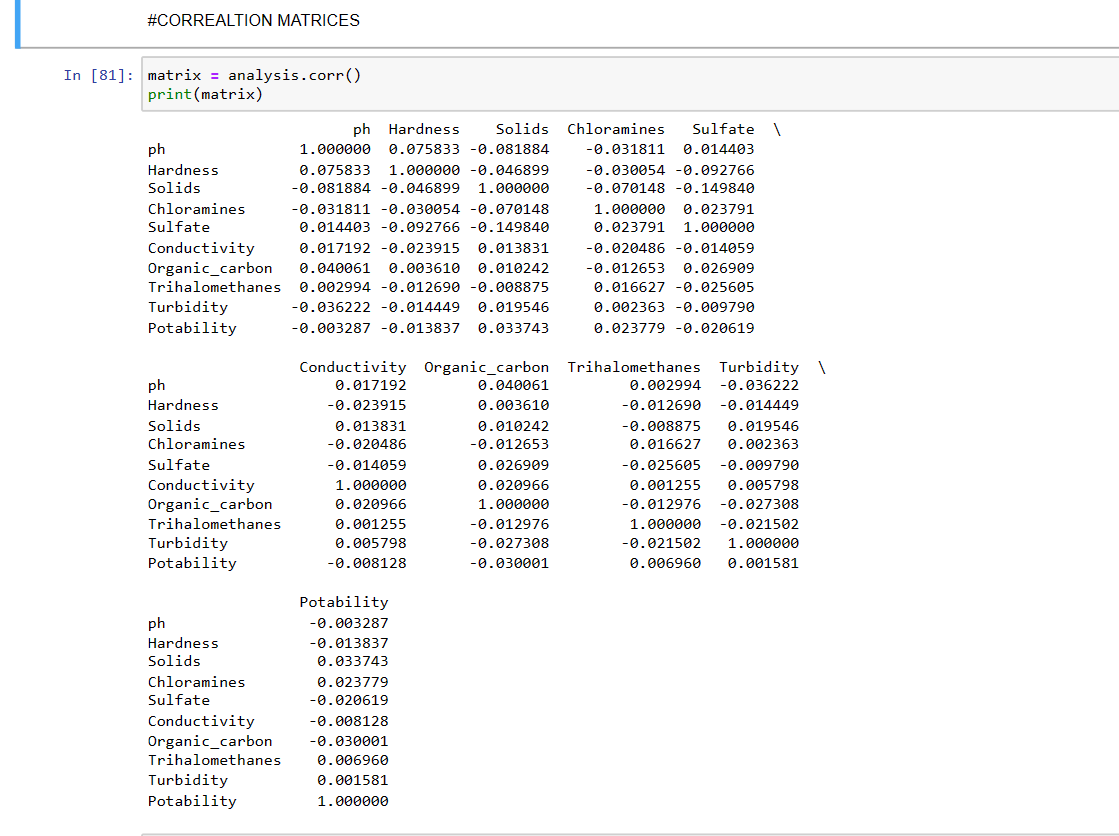
****

plt.figure(figsize=(8,2),dpi=90)

sns.scatterplot(x='Organic\_carbon',y='Trihalomethanes',data=analysis,hue='Turbidity',style='Conductivity')

****

**#Correlation Matrices**

****

**#Logistic Regression**

from sklearn.linear\_model import LogisticRegression

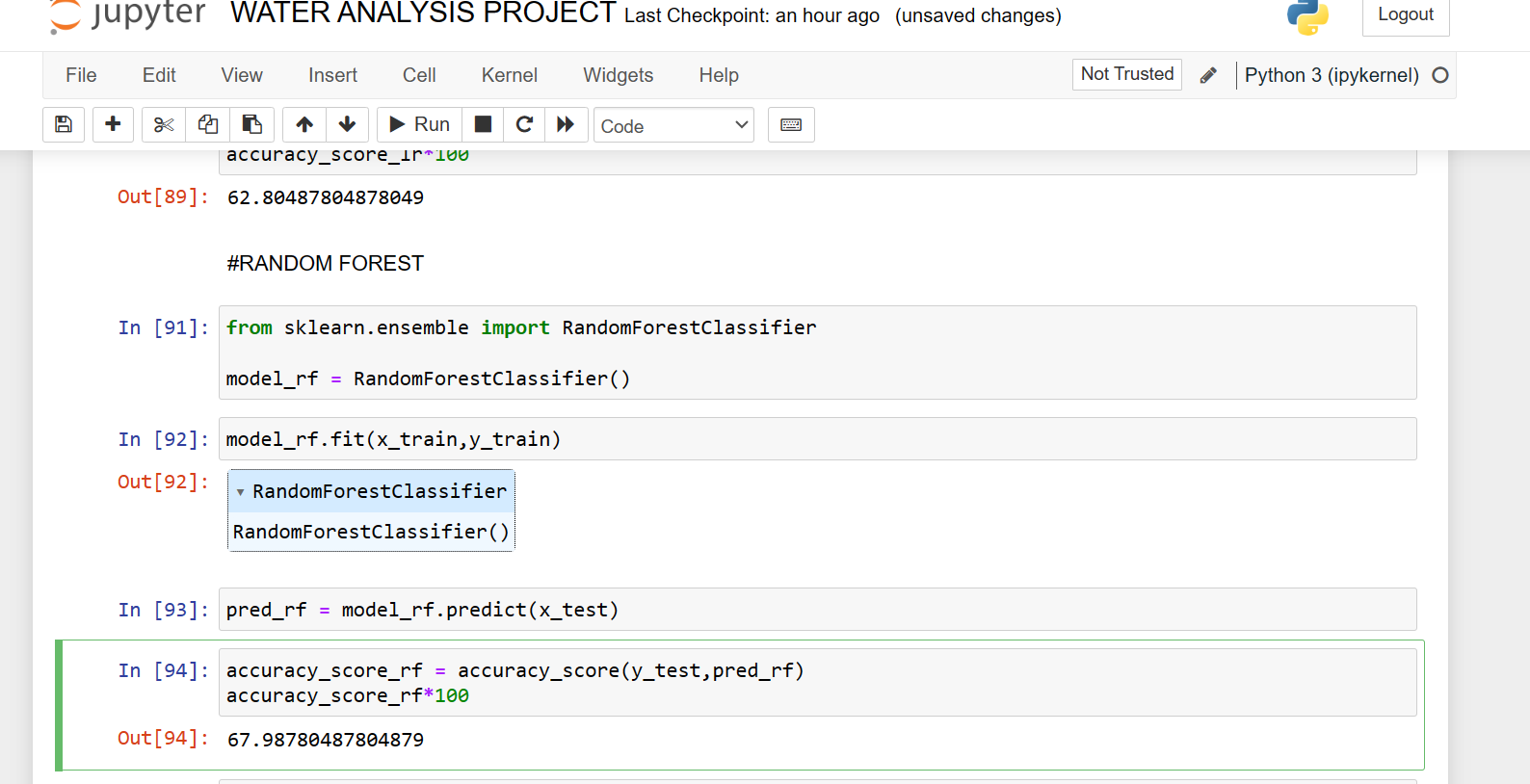
model\_lr = LogisticRegression()

model\_lr.fit(x\_train,y\_train)

pred\_lr = model\_lr.predict(x\_test)

accuracy\_score\_lr = accuracy\_score(y\_test,pred\_lr)

accuracy\_score\_lr\*100

****

**#Output**

**67.98780487804879**

**#Random Forest**

from sklearn.ensemble import RandomForestClassifier

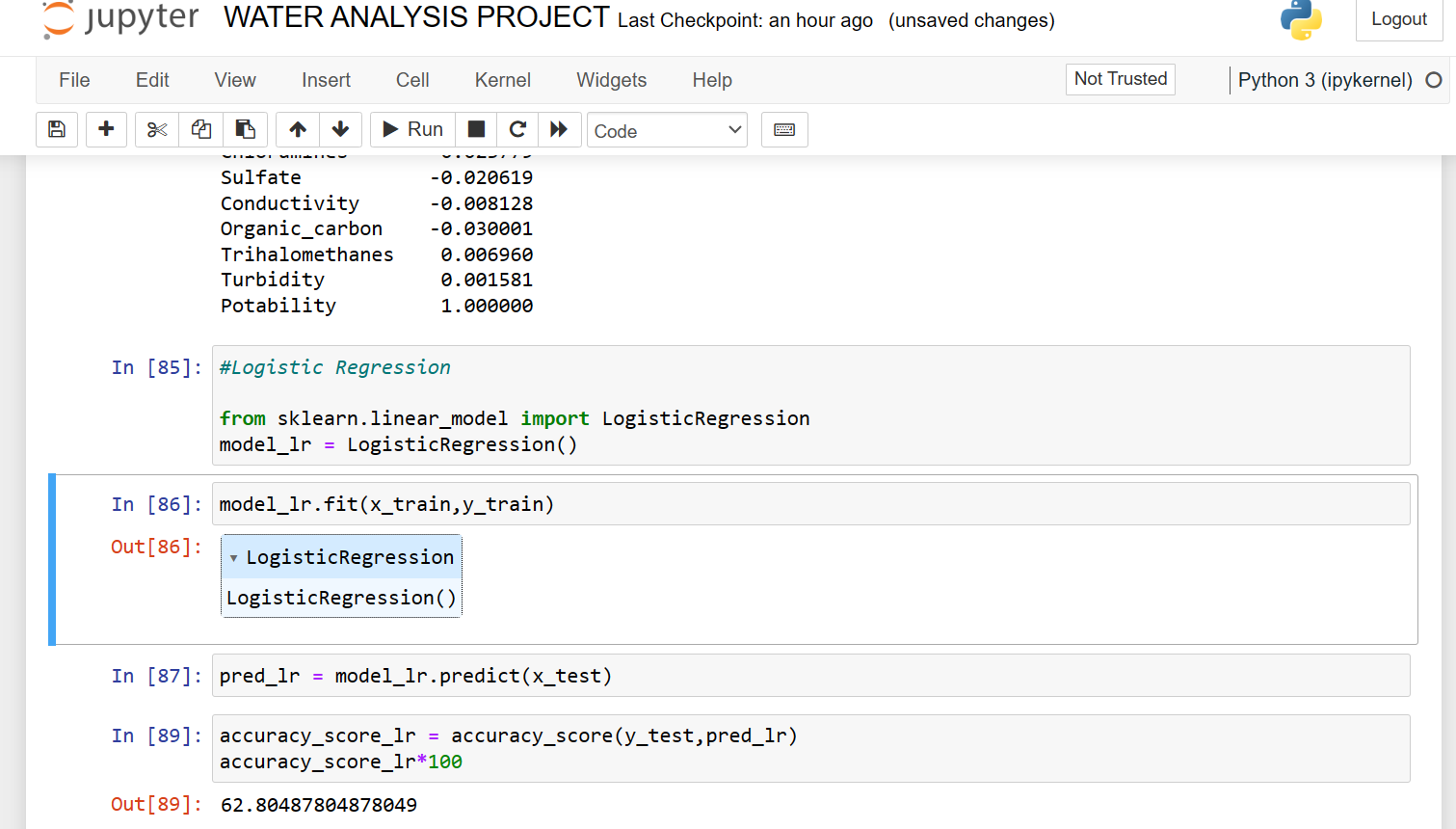
model\_rf = RandomForestClassifier()

model\_rf.fit(x\_train,y\_train)

pred\_rf = model\_rf.predict(x\_test)

accuracy\_score\_rf = accuracy\_score(y\_test,pred\_rf)

accuracy\_score\_rf\*100

****

**#Output**

**62.80487804878049**

In this WATER QUALITY ANALYSIS, we represent various Histogram, Scatterplots and Correlation Matrices, Then we use machine learning models Logistic Regression and Random Forest.