Assignment – 4

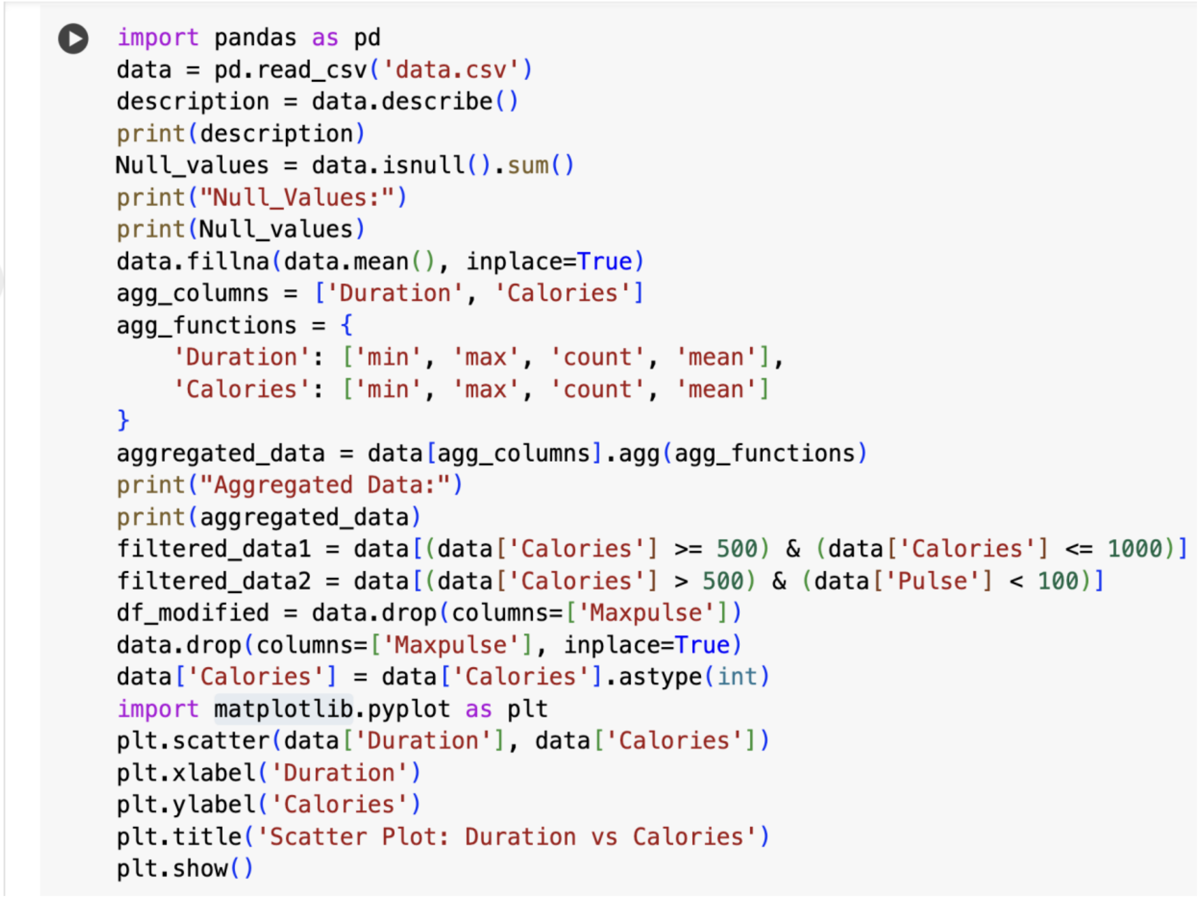
Mohith Degala

700746278

GitHub Link: https://github.com/Mohith700/Assignment_4.git

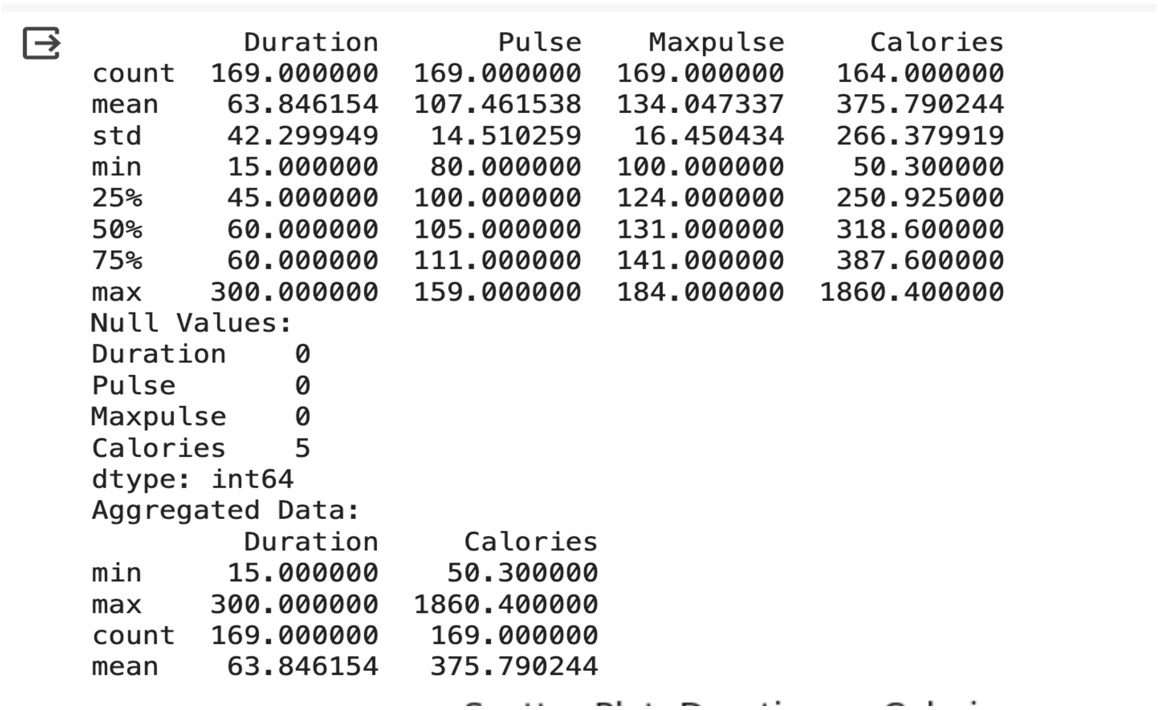
Video Link:
https://drive.google.com/file/d/1meTc25OHMBsmgQDJUgd2aNGpEYMRtXTN/view?usp=drive_link

1) Data Manipulation :

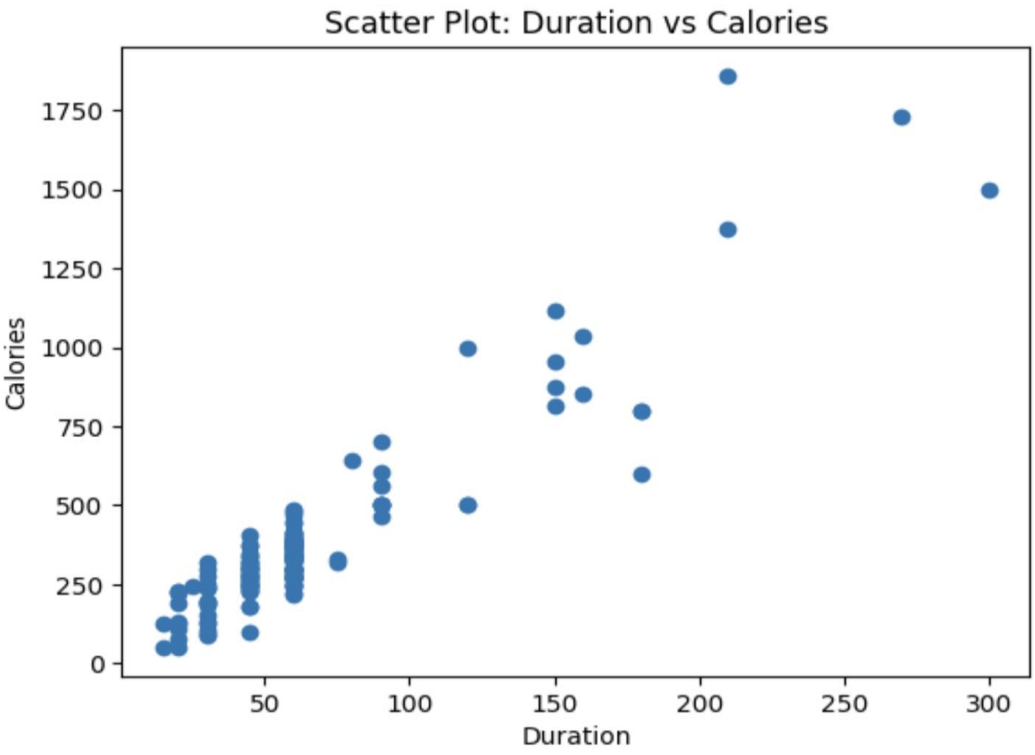


```
import pandas as pd
data = pd.read_csv('data.csv')
description = data.describe()
print(description)
Null_values = data.isnull().sum()
print("Null_Values:")
print(Null_values)
data.fillna(data.mean(), inplace=True)
agg_columns = ['Duration', 'Calories']
agg_functions = {
    'Duration': ['min', 'max', 'count', 'mean'],
    'Calories': ['min', 'max', 'count', 'mean']
}
aggregated_data = data[agg_columns].agg(agg_functions)
print("Aggregated Data:")
print(aggregated_data)
filtered_data1 = data[(data['Calories'] >= 500) & (data['Calories'] <= 1000)]
filtered_data2 = data[(data['Calories'] > 500) & (data['Pulse'] < 100)]
df_modified = data.drop(columns=['Maxpulse'])
data.drop(columns=['Maxpulse'], inplace=True)
data['Calories'] = data['Calories'].astype(int)
import matplotlib.pyplot as plt
plt.scatter(data['Duration'], data['Calories'])
plt.xlabel('Duration')
plt.ylabel('Calories')
plt.title('Scatter Plot: Duration vs Calories')
plt.show()
```

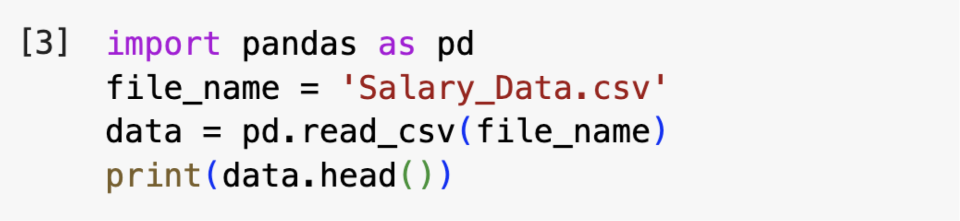
O/P:



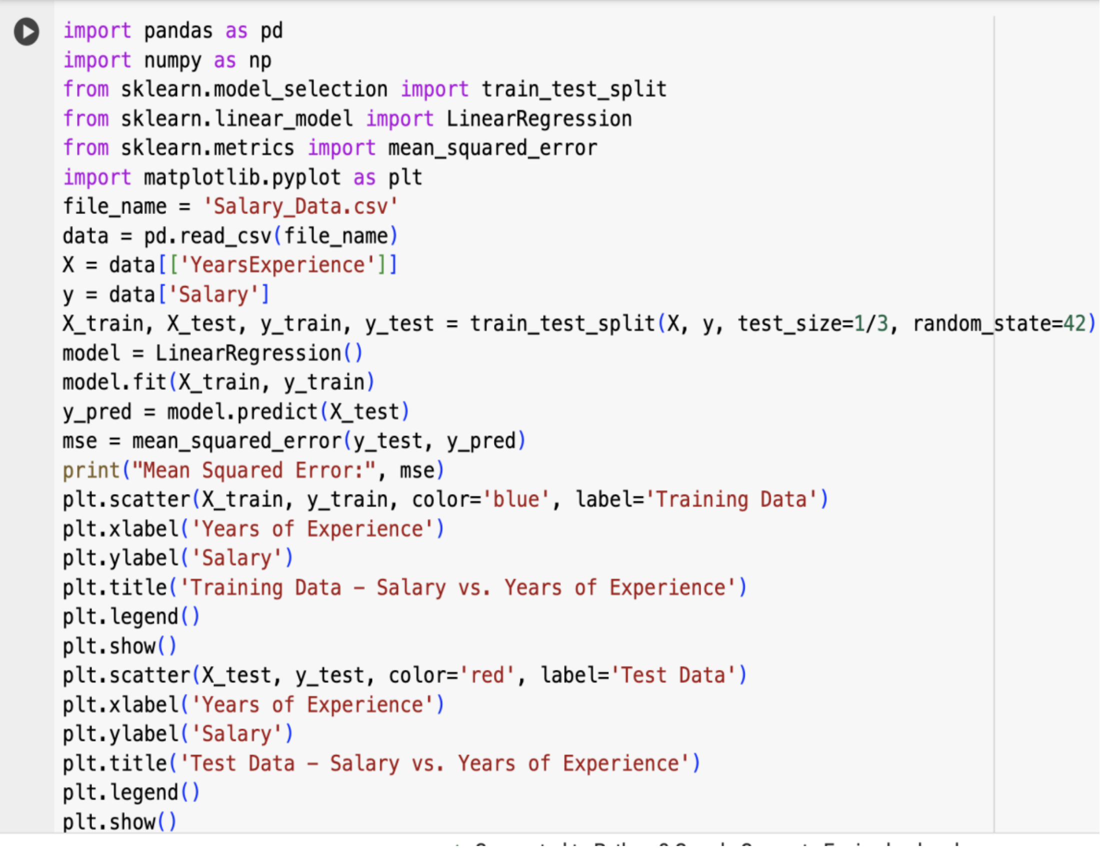
```
          Duration       Pulse    Maxpulse     Calories
count   169.000000  169.000000  169.000000   164.000000
mean     63.846154  107.461538  134.047337   375.790244
std      42.299949   14.510259   16.450434   266.379919
min      15.000000   80.000000  100.000000    50.300000
25%      45.000000  100.000000  124.000000   250.925000
50%      60.000000  105.000000  131.000000   318.600000
75%      60.000000  111.000000  141.000000   387.600000
max     300.000000  159.000000  184.000000  1860.400000
Null Values:
Duration    0
Pulse       0
Maxpulse    0
Calories    5
dtype: int64
Aggregated Data:
          Duration     Calories
min      15.000000    50.300000
max     300.000000  1860.400000
count   169.000000   169.000000
mean     63.846154   375.790244
```



2) Linear Regression :

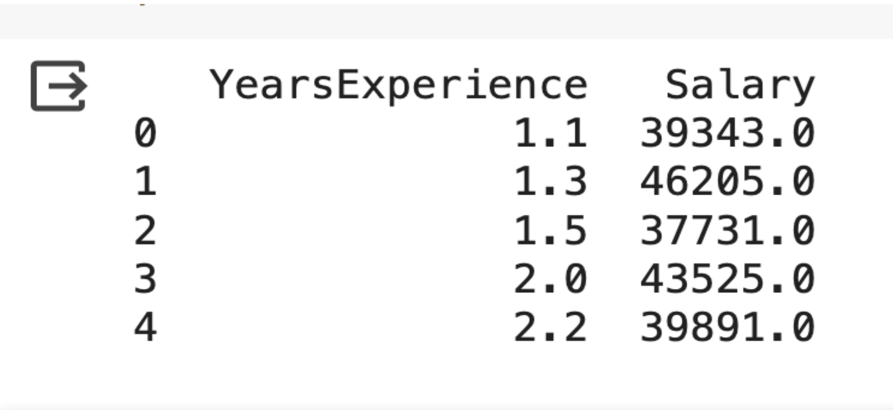


```
[3] import pandas as pd
    file_name = 'Salary_Data.csv'
    data = pd.read_csv(file_name)
    print(data.head())
```

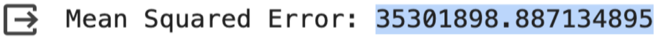


```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
file_name = 'Salary_Data.csv'
data = pd.read_csv(file_name)
X = data[['YearsExperience']]
y = data['Salary']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=1/3, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
plt.scatter(X_train, y_train, color='blue', label='Training Data')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.title('Training Data - Salary vs. Years of Experience')
plt.legend()
plt.show()
plt.scatter(X_test, y_test, color='red', label='Test Data')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.title('Test Data - Salary vs. Years of Experience')
plt.legend()
plt.show()
```

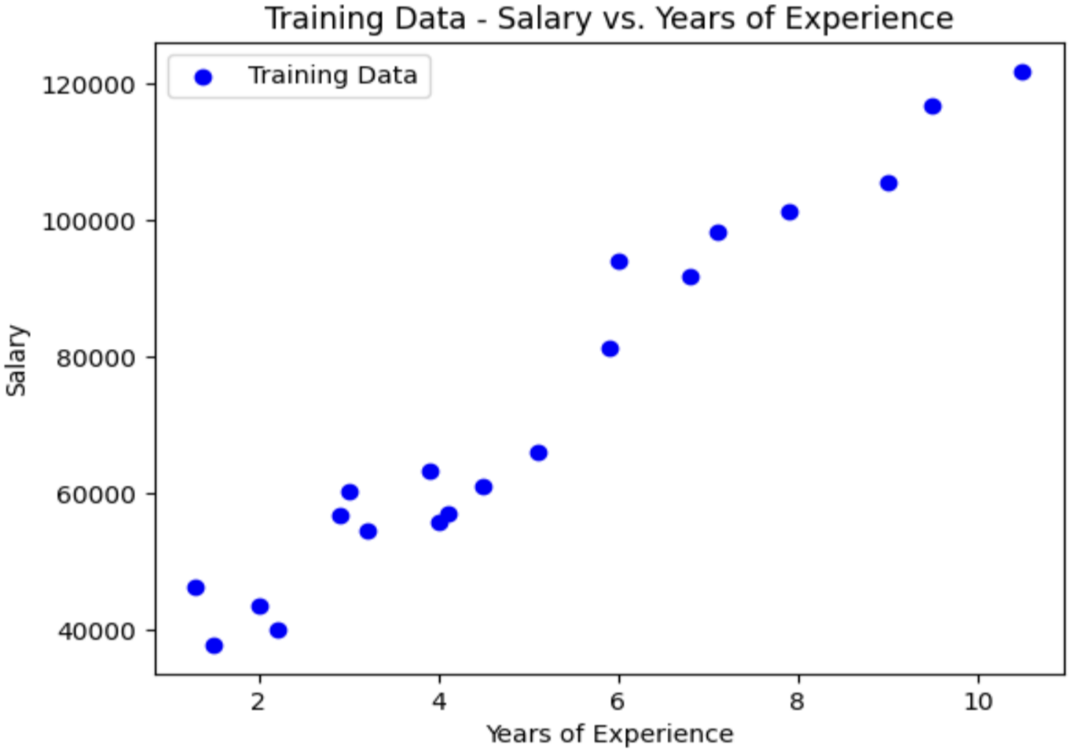
O/P:

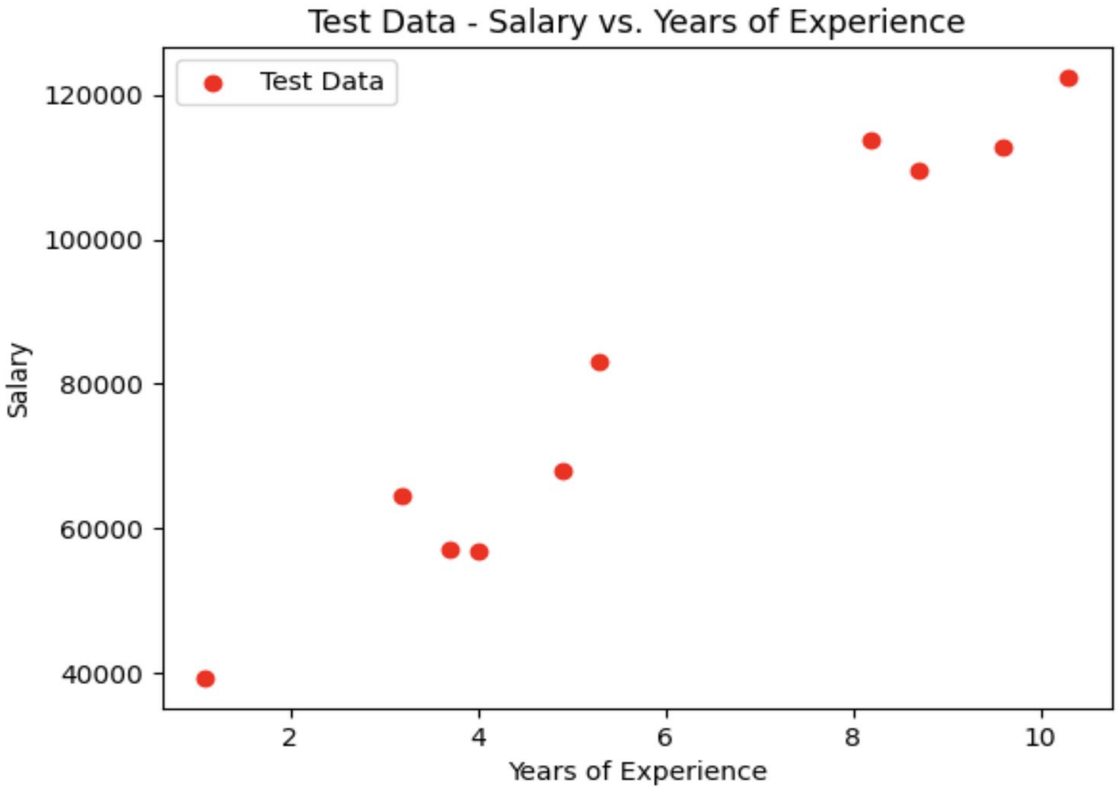


```
   YearsExperience   Salary
0              1.1  39343.0
1              1.3  46205.0
2              1.5  37731.0
3              2.0  43525.0
4              2.2  39891.0
```



```
Mean Squared Error: 35301898.887134895
```




Test Data - Salary vs. Years of Experience
Test Data
Salary
120000
100000
80000
60000
40000
2
4
6
8
10
Years of Experience