



```
import pandas as pd
import matplotlib.pyplot as plt

url = 'https://drive.google.com/uc?id={}'.format('16b7H0tAZ5w1-R0p_sZHmsdeplbGGSNFI')
df = pd.read_csv(url)

# Show the basic statistical description about the data.
stat_description = df.describe()

print(stat_description)

# Check if the data has null values and replace with the mean.
df.fillna(df.mean(numeric_only=True), inplace=True)

# Aggregate data for 'Duration' and 'Calories'.
agg_data = df[['Duration', 'Calories']].agg(['min', 'max', 'count', 'mean'])

print(agg_data)

# Filter for calories between 500 and 1000.
filtered_calories_500_1000 = df[(df['Calories'] >= 500) & (df['Calories'] <= 1000)]

print(filtered_calories_500_1000)

# Filter for calories > 500 and pulse < 100.
filtered_calories_pulse = df[(df['Calories'] > 500) & (df['Pulse'] < 100)]

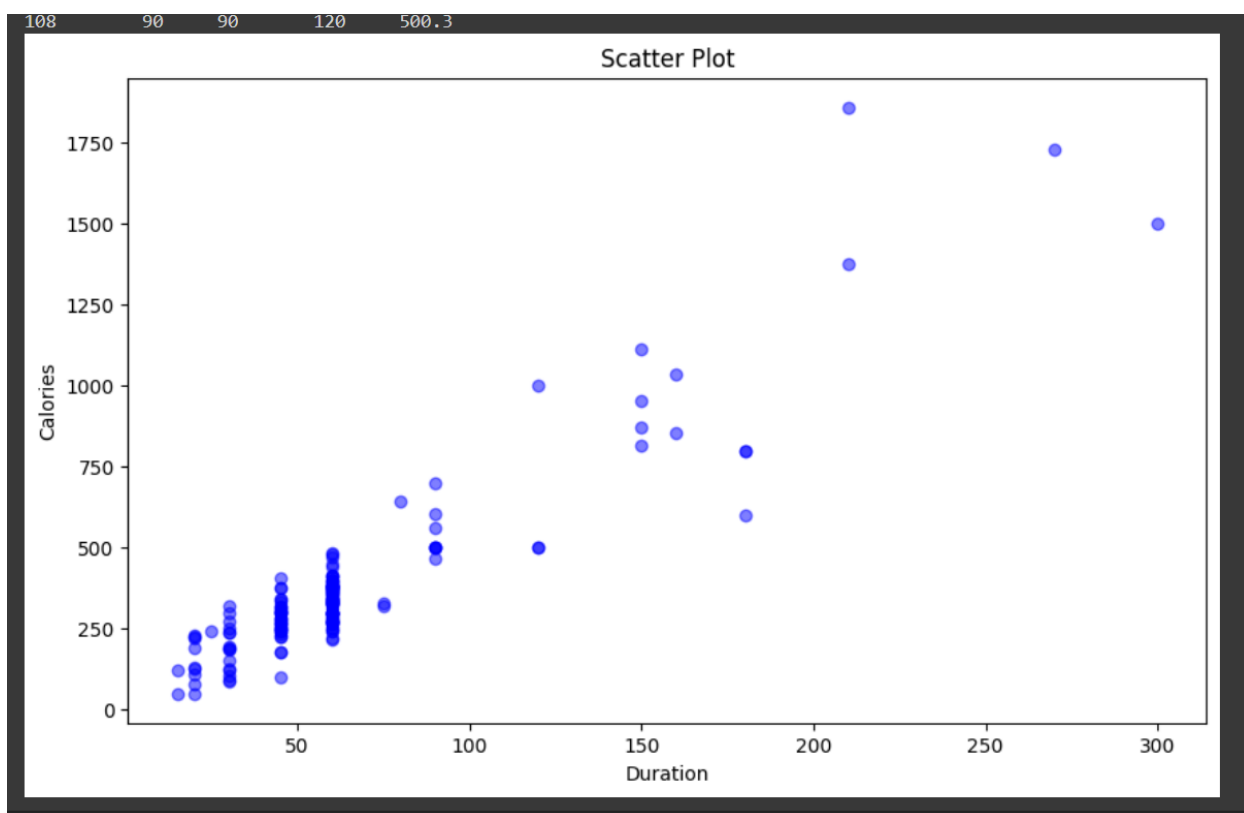
print(filtered_calories_pulse)

# Create a new dataframe without "Maxpulse".
df_modified = df.drop(columns=['Maxpulse'])

# Convert Calories to int.
df['Calories'] = df['Calories'].astype(int)

# Scatter plot for Duration and Calories.
plt.figure(figsize=(10,6))
plt.scatter(df['Duration'], df['Calories'], color='blue', alpha=0.5)
plt.title('Scatter Plot')
plt.xlabel('Duration')
plt.ylabel('Calories')
plt.show()
```

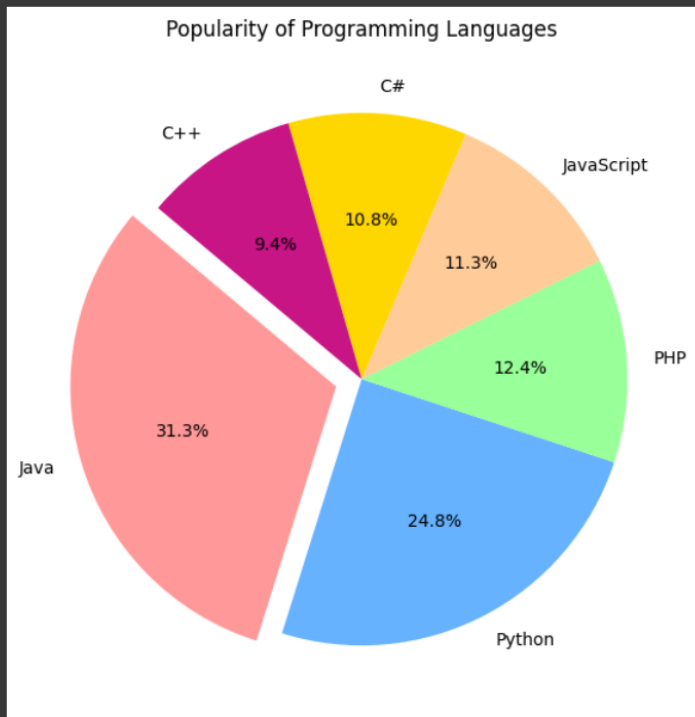
	Duration	Pulse	Maxpulse	Calorie
count	169.000000	169.000000	169.000000	164.000000
mean	63.846154	107.461538	134.047337	375.79024
std	42.299949	14.510259	16.450434	266.37991
min	15.000000	80.000000	100.000000	50.300000
25%	45.000000	100.000000	124.000000	250.92500
50%	60.000000	105.000000	131.000000	318.60000
75%	60.000000	111.000000	141.000000	387.60000
max	300.000000	159.000000	184.000000	1860.40000
	Duration	Calories		
min	15.000000	50.300000		
max	300.000000	1860.400000		
count	169.000000	169.000000		
mean	63.846154	375.790244		
	Duration	Pulse	Maxpulse	Calories
51	80	123	146	643.1
62	160	109	135	853.0
65	180	90	130	800.4
66	150	105	135	873.4
67	150	107	130	816.0
72	90	100	127	700.0
73	150	97	127	953.2
75	90	98	125	563.2
78	120	100	130	500.4
83	120	100	130	500.0
90	180	101	127	600.1
99	90	93	124	604.1
101	90	90	110	500.0
102	90	90	100	500.0
103	90	90	100	500.4
106	180	90	120	800.3
108	90	90	120	500.3
	Duration	Pulse	Maxpulse	Calories
65	180	90	130	800.4
70	150	97	129	1115.0
73	150	97	127	953.2
75	90	98	125	563.2
99	90	93	124	604.1
103	90	90	100	500.4
106	180	90	120	800.3
108	90	90	120	500.3

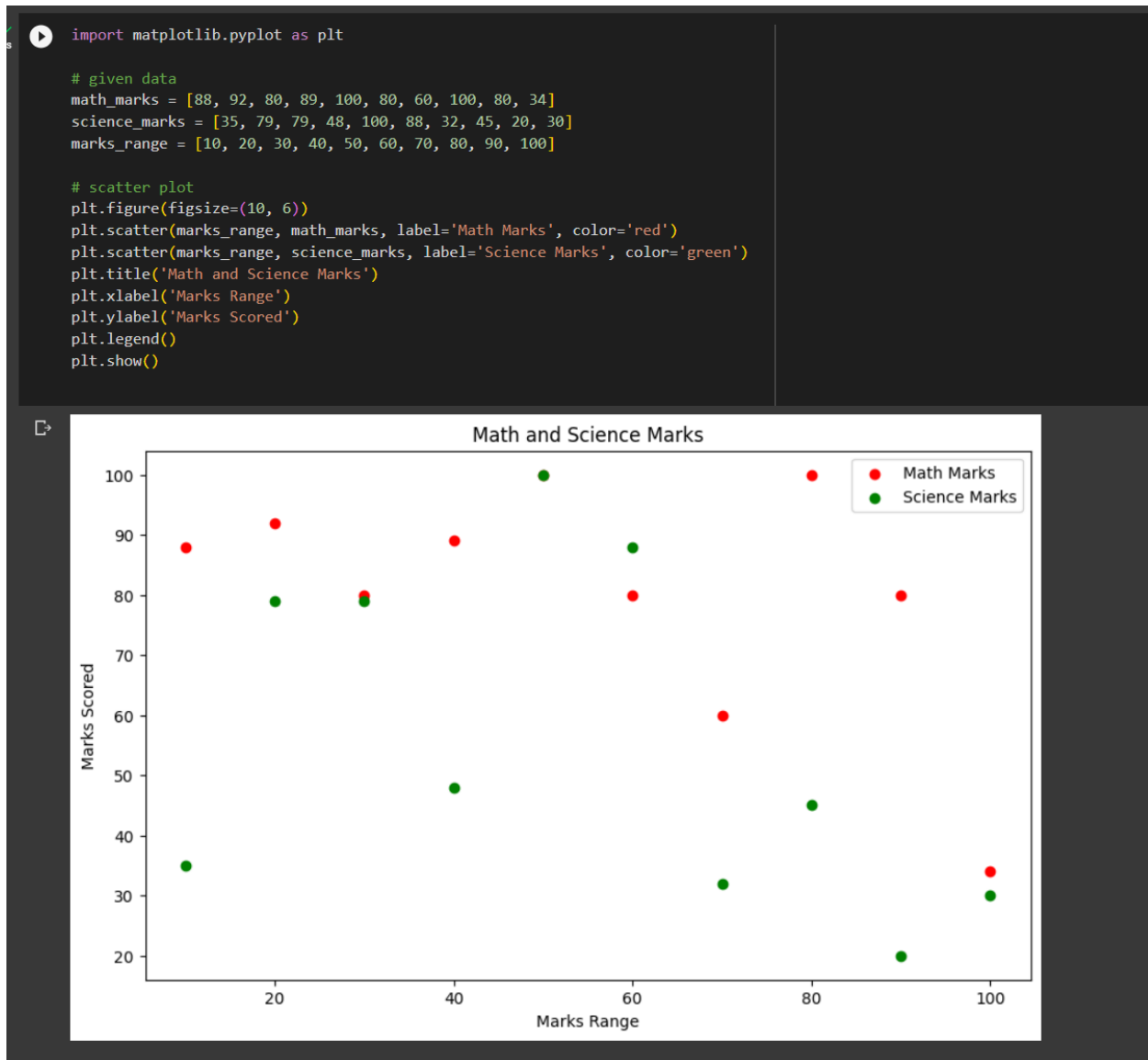


```
import matplotlib.pyplot as plt

# given data
languages = ['Java', 'Python', 'PHP', 'JavaScript', 'C#', 'C++']
popularity = [22.2, 17.6, 8.8, 8, 7.7, 6.7]
colors = ['#FF9999', '#66B2FF', '#99FF99', '#FFCC99', '#FFD700', '#C71585']
explode = (0.1, 0, 0, 0, 0, 0)

# Plotting the pie chart
plt.figure(figsize=(10, 7))
plt.pie(popularity, explode=explode, labels=languages, colors=colors, autopct='%1.1f%%', shadow=False, startangle=140)
plt.title("Popularity of Programming Languages")
plt.show()
```





Github Repo: <https://github.com/dheerukarra/BigDataAnalytics/tree/main/ICP%204>

Video link: <https://youtu.be/5hHXqC7ziHY>