Name: Arya Nair
Roll Number: 16010421063

```
import pandas as pd


def mean(arr):
    res=0
    for i in arr:
        res+=i
    return res/len(arr)

df=pd.read_csv('data.csv')
followers=df['followers']
views=df['Views']
```

[48]

```
m_followers=mean(followers)
m_views=mean(views)
m_views
```

[49]

```
27059170214.84831
```

```
numerator=0
denominator=0

for i in range(len(followers)):
    numerator+=(followers[i]-m_followers)*(views[i]-m_views)
    denominator+=(followers[i]-m_followers)**2
b1=numerator/denominator
b1
```

[50]

```
828.7676283650984
```

```
c1=m_views-(b1*m_followers)
if c1>0:
    print(f"y={b1}x+{c1}")
else:
    print(f"y={b1}x{c1}")
```

```
c1=m_views-(b1*m_followers)
if c1>0:
    print(f"y={b1}x+{c1}")
else:
    print(f"y={b1}x{c1}")
```

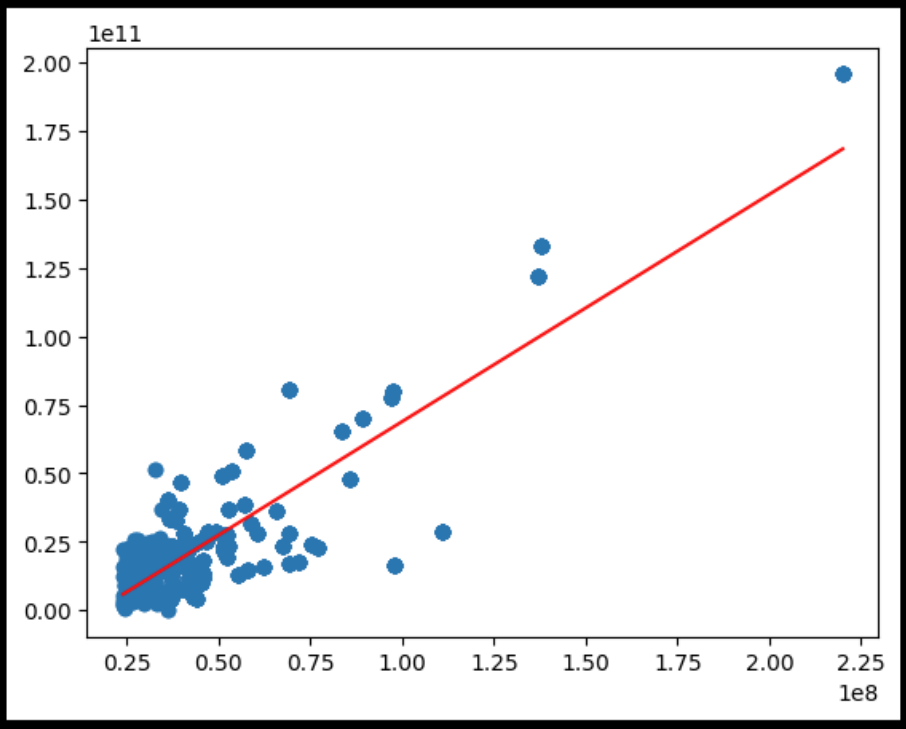
[51]

```
y=828.7676283650984x-13989487336.159077
```

```
import matplotlib.pyplot as plt

plt.scatter(followers,views)
plt.plot(followers,b1*followers+c1,color='red')
plt.show()
```

[52]



1. **How will you choose between linear regression and non-linear regression?**
   Ans: The general guideline is to use linear regression first to determine whether it can fit the particular type of curve in our data. If we can't obtain an adequate fit using linear regression, then we might need to choose nonlinear regression. Linear regression is easier to use, simpler to interpret, and we obtain more statistics that help you assess the model. While linear regression can model curves, it is relatively restricted in the shapes of the curves that it can fit. Sometimes it can't fit the specific curve in our data. Nonlinear regression can fit many more types of curves, but it can require more effort both to find

the best fit and to interpret the role of the independent variables. Additionally, R-squared is not valid for nonlinear regression, and it is impossible to calculate p-values for the parameter estimates.

2. **Explain the nature or characteristics of a dataset where we can apply regression imputation.**

Ans: Regression imputation consists of two subsequent steps:

1. A linear regression model is estimated on the basis of observed values in the target variable Y and some explanatory variables X.
2. The model is used to predict values for the missing cases in Y. Missing values of Y are then replaced on the basis of these predictions.

Relationships of X and Y (i.e. correlations, regression coefficients etc.) are preserved, since imputed values are based on regression models. This is a big advantage over simpler imputation methods such as mean imputation or zero substitution.

Regression imputation is classified into two different versions: deterministic and stochastic regression imputation.

Deterministic regression imputation replaces missing values with the exact prediction of the regression model. Random variation (i.e. an error term) around the regression slope is not considered. Imputed values are therefore often too precise and lead to an overestimation of the correlation between X and Y.

Stochastic regression imputation was developed in order to solve this issue of deterministic regression imputation. Stochastic regression imputation adds a random error term to the predicted value and is therefore able to reproduce the correlation of X and Y more appropriately.

Outcomes: CO2: Comprehend descriptive and proximity measures of data

Conclusion- Successfully understood and implemented Linear and Multiple Regression