**Experiment Number : 3 - Predicting missing data values using regression modeling**

**Batch: FDS-2 Roll Number: 16010422234 Name: Chandana Ramesh Galgali**

**Aim of the Experiment:** Predict missing data values using regression modeling.

**Program/ Steps:**

Activity 0:

Problem statement:

Let's predict temperature through interpolation techniques . The Data given points are as follows.

Recorded temperatures:

Time: 2.0 hours, Temperature: 9.0°C

Time: 3.0 hours, Temperature: 6.0°C

Time: 4.0 hours, Temperature: 12.0°C

Predict the temperature: 3.5 hours?

Activity 1:

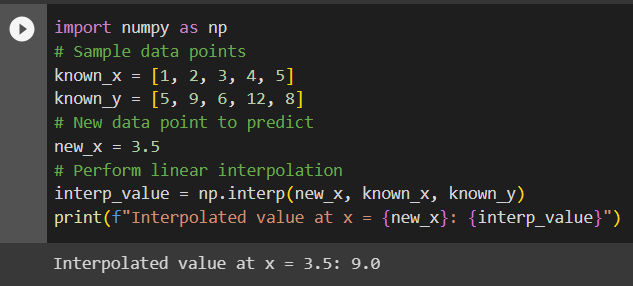
Identify attributes suitable for applying Linear regression. Construct a linear regression model for your dataset and predict the missing values in your data set. Evaluate the accuracy of prediction.(usage of built in package for prediction is not expected)

Activity 2:

Identify attributes suitable for applying Multiple Linear regression. Construct a linear regression model for your dataset and predict the missing values in your data set. Evaluate the accuracy of prediction.(usage of built in package for prediction is not expected)

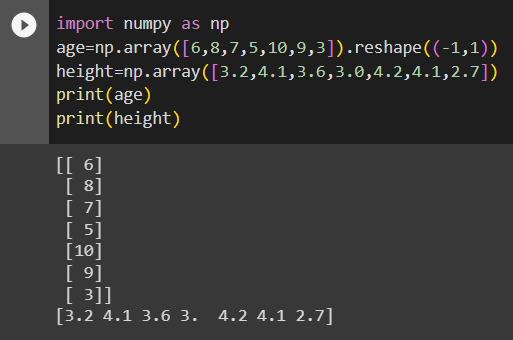
**Code with Output/Result:**

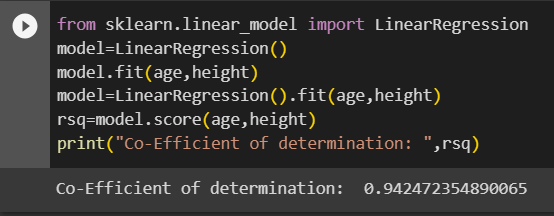
Activity 0:

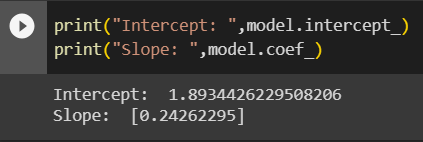
****

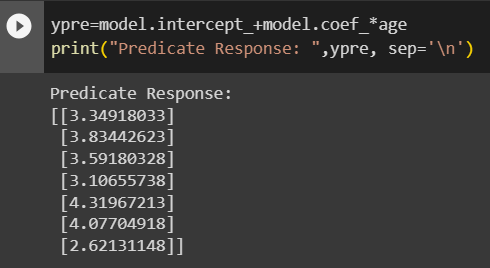
Activity 1:

Single Linear Regression:

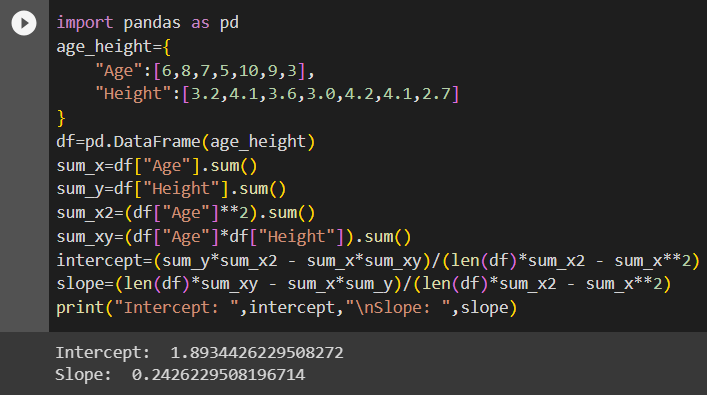


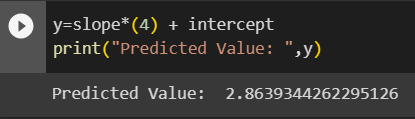






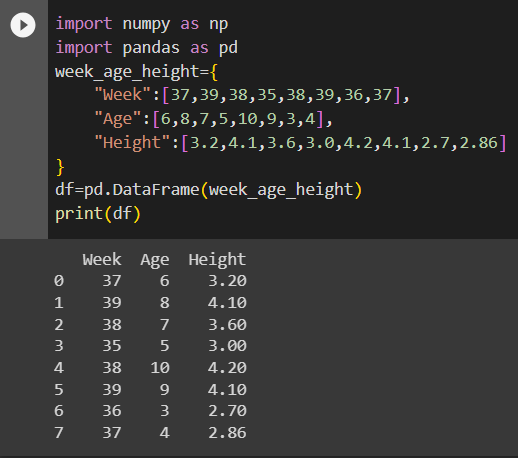
Single Linear Regression without readymade functions:

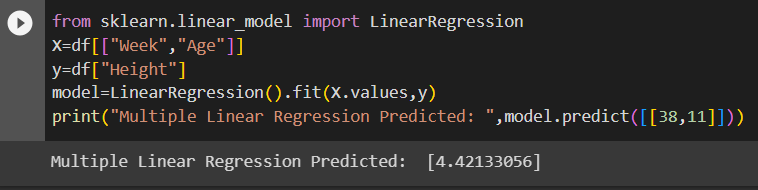


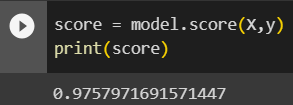


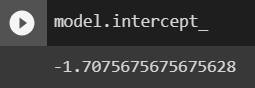
Activity 2:

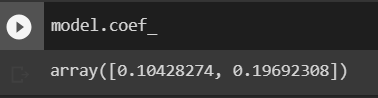
Multiple Linear Regression:



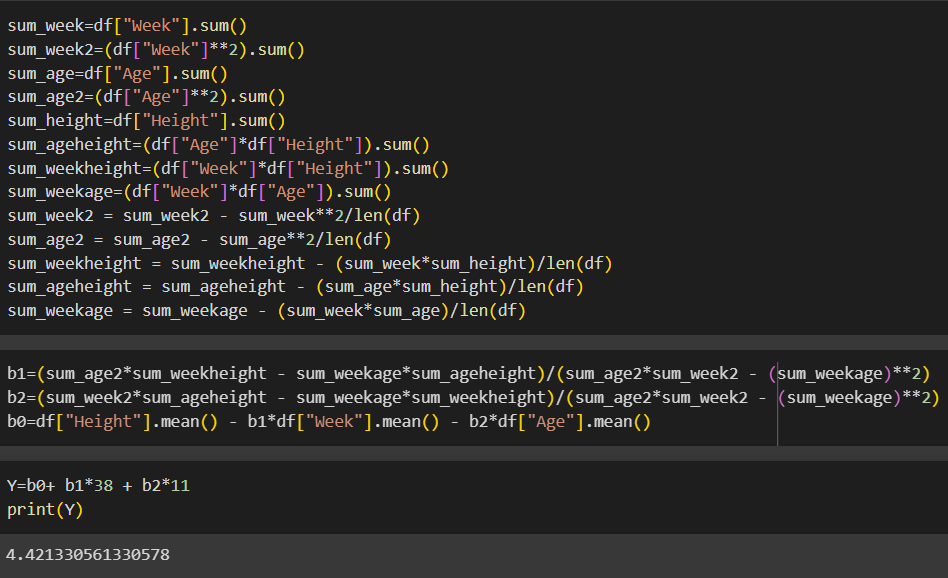








Multiple Linear Regression without readymade functions:



**import numpy as np**

**import pandas as pd**

**data=pd.read\_csv(r'C:\Users\daxay\Downloads\Flight\_delay.csv')**

**data\_array=data.to\_numpy()**

**print("Dataframe:\n",data\_array)**

**arrdelay=data\_array[:,12]**

**depdelay=data\_array[:,13]**

**print("\nArrival Delay:\n",arrdelay)**

**print("\nDepature Delay:\n",depdelay)**

**from sklearn.linear\_model import LinearRegression**

**model=LinearRegression()**

**model.fit(arrdelay.reshape(-1, 1),depdelay.reshape(-1, 1))**

**rsq=model.score(arrdelay.reshape(-1, 1),depdelay.reshape(-1, 1))**

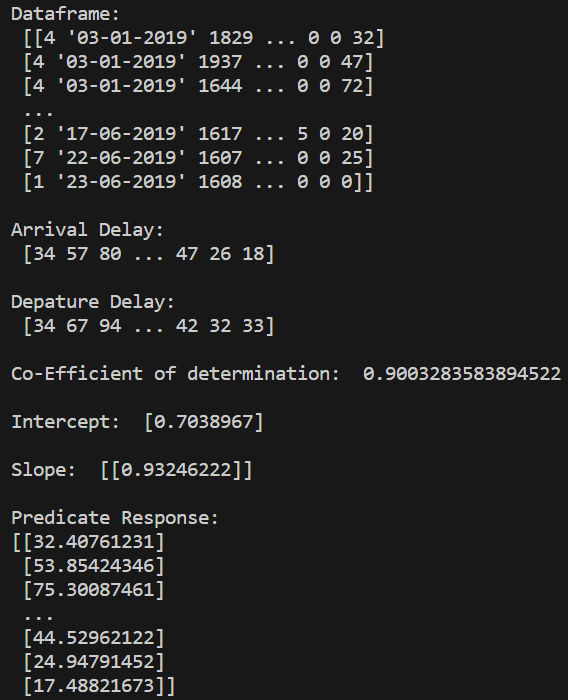
**print("\nCo-Efficient of determination: ",rsq)**

**print("\nIntercept: ",model.intercept\_)**

**print("\nSlope: ",model.coef\_)**

**ypre=model.predict(arrdelay.reshape(-1,1))**

**print("\nPredicate Response: ",ypre, sep='\n')**

****

**import numpy as np**

**import pandas as pd**

**data=pd.read\_csv(r'C:\Users\daxay\Downloads\Flight\_delay.csv')**

**data\_array=data.to\_numpy()**

**print("Data:\n",data\_array)**

**x={**

**"airtime":data\_array[:,11],**

**"arrdelay":data\_array[:,12],**

**"depdelay":data\_array[:,13]**

**}**

**df=pd.DataFrame(x)**

**print("\nDataframe:\n",df)**

**from sklearn.linear\_model import LinearRegression**

**X=df[["arrdelay","depdelay"]]**

**y=df["airtime"]**

**model=LinearRegression().fit(X.values,y)**

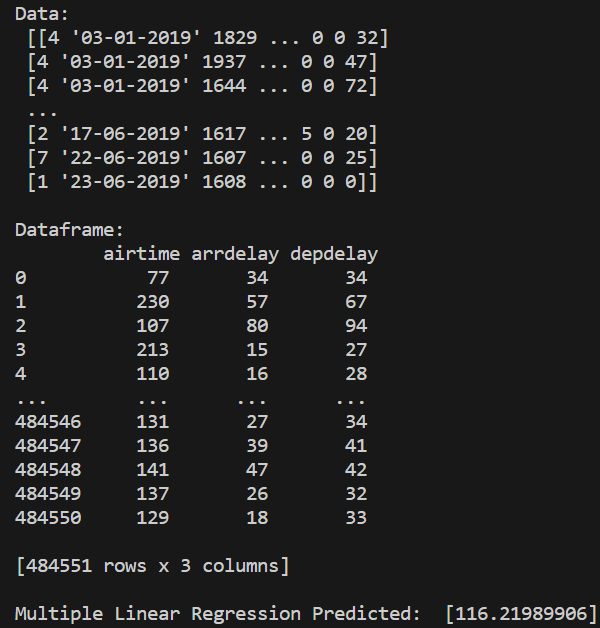
**print("\nMultiple Linear Regression Predicted: ",model.predict([[38,11]]))**

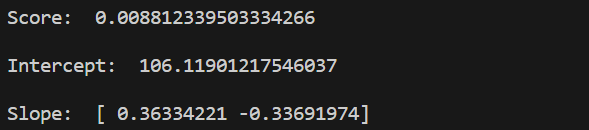
**score = model.score(X,y)**

**print("\nScore: ",score)**

**print("\nIntercept: ",model.intercept\_)**

**print("\nSlope: ",model.coef\_)**

****

****

**Post Lab Question-Answers:**

**1. How will you choose between linear regression and non-linear regression?**

**Ans:** When choosing between linear regression and non-linear regression, you need to consider the nature of your data and the relationship between the variables you are analyzing. Here are some factors to consider:

1. Linearity of the relationship: If the relationship between the independent and dependent variables appears to be linear, with a straight-line pattern, then linear regression is appropriate. On the other hand, if the relationship seems to follow a curved or non-linear pattern, non-linear regression may be more suitable.

2. Complexity of the model: Linear regression models are simpler and easier to interpret since they assume a linear relationship between variables. Non-linear regression models can capture more complex relationships but may be more challenging to interpret and require more computational resources.

3. Domain knowledge: Consider your understanding of the underlying process or theory related to the data. If you have prior knowledge suggesting a specific non-linear relationship, it may be more appropriate to use non-linear regression.

4. Data availability: Non-linear regression models often require more data points to estimate the parameters accurately. If you have a limited dataset, linear regression may be a more viable option.

5. Model performance: Evaluate the performance of both linear and non-linear regression models using appropriate metrics (e.g., R-squared, mean squared error). Compare their predictive accuracy and choose the model that performs better on your data.

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

**Ans:** Regression imputation is a technique used to fill in missing values in a dataset by predicting them based on the relationship between the variables. It assumes that the missing values are related to other variables in the dataset and can be estimated using regression analysis.

The characteristics of a dataset where regression imputation can be applied are as follows:

1. Missing at random (MAR): Regression imputation assumes that the missing values are not systematically related to the missing values themselves. In other words, the missingness is related to other observed variables in the dataset. If the missingness is random or can be explained by other variables, regression imputation can be effective.

2. Linear relationship: Regression imputation assumes a linear relationship between the variable with missing values and the other variables used for prediction. If the relationship is non-linear, other techniques like non-linear regression imputation or machine learning algorithms may be more appropriate.

3. Sufficient predictor variables: Regression imputation requires having other variables in the dataset that are strongly correlated with the variable with missing values. These predictor variables should be able to explain a significant portion of the variation in the variable with missing values.

4. Independence of errors: Regression imputation assumes that the errors in the regression model are independent and normally distributed. This assumption ensures that the imputed values are unbiased and have reasonable variability.

5. Adequate sample size: Regression imputation performs better with larger sample sizes since it relies on estimating the regression coefficients accurately. With a small sample size, the imputed values may be less reliable.

**Outcomes:**

**Comprehend descriptive and proximity measures of data**

**Conclusion (based on the Results and outcomes achieved):**

The experiment demonstrated the potential of regression modeling for predicting missing data values, providing researchers with a valuable tool for data imputation and analysis. Further research and experimentation can explore the limitations and applicability of regression imputation in different contexts and datasets.

**References:**

Books/ Journals/ Websites

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition