**University of** 

**Asia Pacific**

**Course Code:** CSE 404

**Course Title:** Artificial Intelligence Lab

**Project - 02**

**Subject:** Multivariable Linear Regression using open source dataset.

**Submitted By: Submitted To:**

**Name:** Al Khaled Rayhan **Name:** Dr. Nasima Begum

**ID:** 17201045 Assistant professor, CSE

**Sec:** A University of Asia Pacific

4th year 1st semester

**Implement the Multivariable Linear Regression Using Open Source Dataset without SK-Learn.**

**Description :**

Language: python

Dataset : [Medical Cost Personal Datasets](https://www.kaggle.com/mirichoi0218/insurance)

Environment: Windows XP

**Introduction of Linear Regression :**

Linear regression is probably the most simple machine learning algorithm. It is very good for starters because it uses simple formulas. So, it is good for learning machine-learning concepts.

**Concepts and Formula :**

Linear regression uses the simple formula

**Y = C+AX**

Just as a reminder, Y is the output or dependent variable, X is the input or the independent variable, A is the slope, and C is the intercepT.

For the linear regression, we follow these notations for the same formula



If we have multiple independent variables, the formula for linear regression will look like



**Implementation of the Algorithm :**

In this dataset, there are only seven variables but I developed the algorithm for any number of variables.

I will use numpy and pandas library in python. All these rich libraries of python made the machine learning algorithm a lot easier.

Import the packages and the dataset.

Print the first 5 row of the dataset

This method prints information about a DataFrame including the index dtype and columns, non-null values and memory usage

And show the information of my dataset.

I'm using df.isnull().sum() to get a count of NaN value in a pandas dataframe.

Is there a way to only show the value count that isn't zero (i.e. if the column has 0 NaNs then don't show in the value count.

This is the result and I would like to remove the 0 values

This returns a Series with the data type of each column

from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()

This approach is very simple and it involves converting each value in a column to a number

I did the encoding on sex, smoker and region variable Because linear regression does not work with strings. Linear regression recognizes only integer data.

I put all of mean value of the variables in Thetas.

And put all the variables in X’s.

NumPy works with Python objects called multi-dimensional **arrays**. Arrays are basically collections of values, and they have one or more dimensions.

I created 2 functions of distance and prediction Which I left blank

I called df.head () and got the chart.

We will fill in the predictions with theta values.

Then I will count the distance.

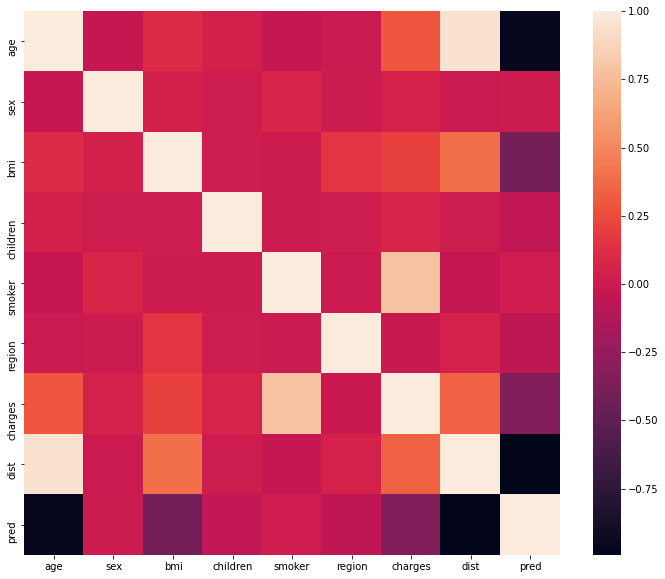
cost = (df['dist'].sum())/(2\*m)

I'll get out of trouble.

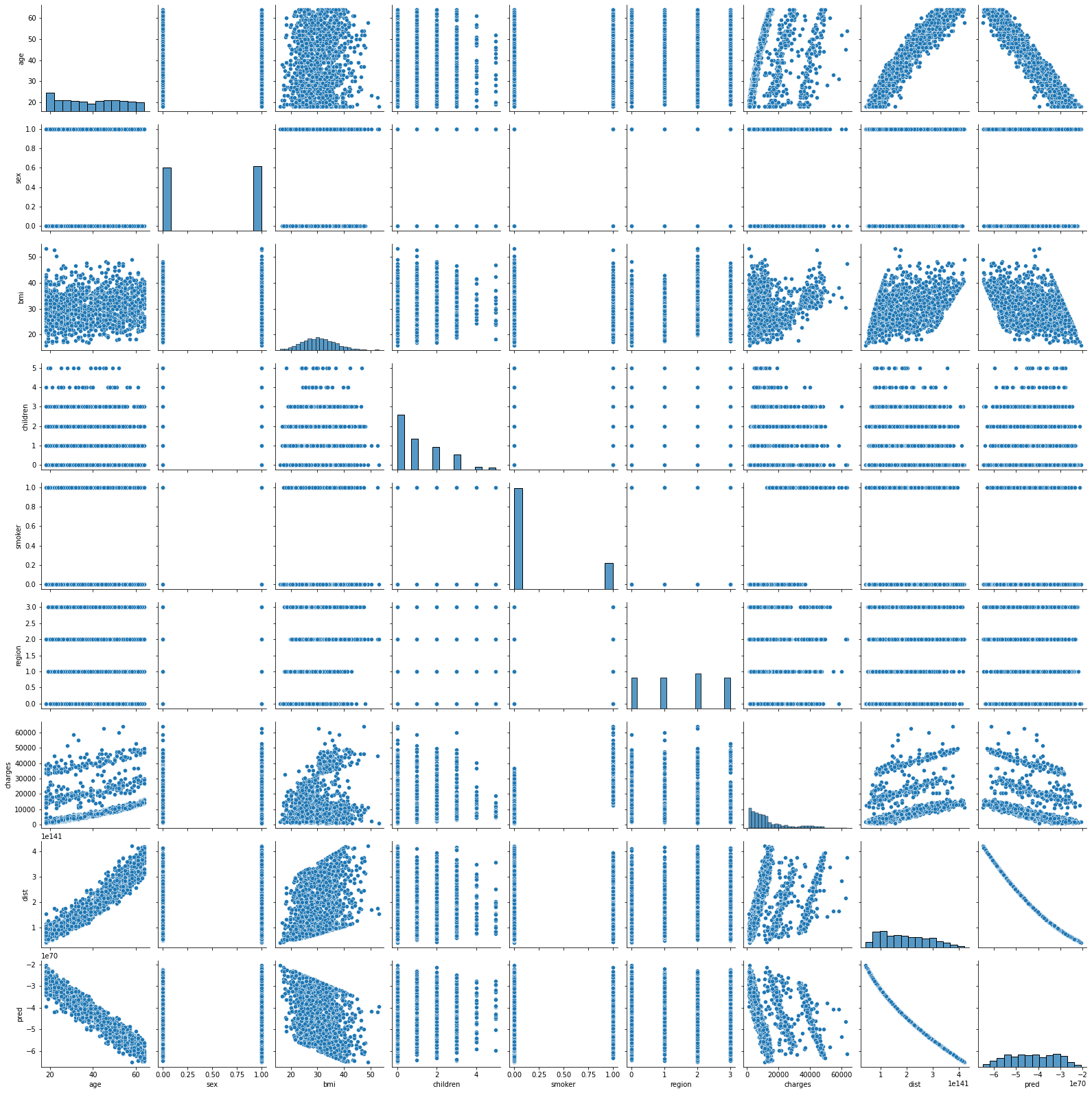
I will put all the theta’s in the formula to find the gradient.

**Graph:**

**Correlation (Pearson):**

****

**Bivariate Analysis:**

****

Git link: <https://github.com/Rayhan1996/Linear_Regression/>