**Student Name & ID: AIML Assignment 2: [80 marks]**

**Regression: Linear, Polynomial, Multiple, Scale, Train/Test**

x = [5,7,8,7,2,17,2,9,4,11,12,9,6]  
y = [90,80,87,88,90,86,89,87,94,78,77,85,86]

Figure 1: the x-axis represents a car’s age, and the y-axis represents its speed.

Using the two vectors in Figure 1 above, write code in Python to:

1. Create a Scatter Plot with the values of x and y [4 marks]

#1 scatter plot

x=[5,7,8,7,2,17,2,9,4,11,12,9,6]

y=[90,80,87,88,90,86,89,87,94,78,77,85,86]

plt.scatter(x,y)

plt.xlabel('Age of car')

plt.ylabel('Speed of car')

plt.show()

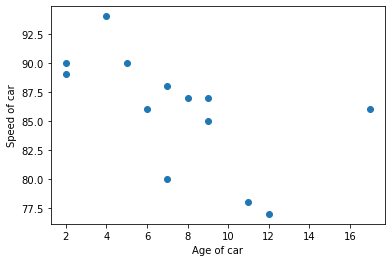


Figure 1 scatter plot for question 1

1. Fit the data to a Linear Regression line, and plot it on the scatter plot [4 marks]

#2 linear regression

slope,intercept,r,p,std\_err = stats.linregress(x,y)

Reg\_LineY = np.array(x)\*slope + intercept

plt.scatter(x,y)

plt.plot(x,Reg\_LineY)

plt.xlabel('Age of car')

plt.ylabel('Speed of car')

plt.grid()

plt.show()

1. Estimate the coefficient of correlation, r, and explain what it means [4 marks]

#3 estimate coefficient of correlation r

print(r)



The coefficient of correlation r ranges from 1 to -1. It is used as a measurement of the relationship between 2 variables usually x and y. Any value of r that is greater than 1 or -1 represents an error. Values of r that are close to 1 and -1 represent better relationship between variables. If they are exactly 1 and -1, they represent perfectly positive and negative relationships between the variables. Values that are closer to 0 represent a poor relationship between 2 variables.

In this case, an r value of -0.5753 represents an adequate or average negative relationship between the car year (x) and the speed of the car(y).

1. Predict the speed of a 13 year old car using the linear reg. model [4 marks]

#4 predict spd of 13 yr old car using lin reg model

Car\_Age=13

Speed\_Predict= Car\_Age \*slope + intercept

print(Speed\_Predict)



1. Fit the data to a third degree polynomial and plot it on the scatter plot[4 marks]
2. Predict the speed of a 13 year old car using the 3rd deg. polynomial [4 marks]
3. Estimate the r-squared value and explain what it means [4 marks]
4. Explain any differences in values obtained in (4) and (6) [4 marks]

Use the attached excel file data1.csv to perform the following:

1. Read in the data1 file into your workspace and store it as a data frame[4 marks]
2. Extract the weight and volume fields and store it in a variable, X [2 marks]
3. Extract the CO2 field and store it as variable, y [2 marks]
4. Fit a Linear Regression model to variables X and y [4 marks]
5. Predict value of CO2 of the Volvo XC70 with weight of 1746 kg & volume of 2000cm3 , and explain why this value is different from 117, as given in the data file. [6 marks]
6. Find the reg. coefficient between X and y and explain its meaning [4 marks]
7. Find the R2 (R-squared) score and explain its meaning [4 marks]
8. Scale the variables in X using the standardization method [4 marks]
9. Repeat question (13) using this scaled X variable [2 marks]
10. Explain any differences in values in ques: (13) & (16) [2 marks]
11. Explain what is the train/test method and why is it important [4 marks]
12. Apply the train/test method as follows: i) use the first 30 values of variables X and y to train a Linear Regression model, ii) repeat question (13), (14), (15) and iii) explain any differences between the values obtained in here and in question (13) for the value of the CO2 (calculate % differences between predicted and actual value for your discussion) for the Volvo XC70, the reg. coefficient and R-squared values. [10 marks]

Please type out your solutions in this word document just after the respective question (so that I can see the marks allocated when I am marking), then convert it to a pdf (file 🡺 save as 🡺 type pdf, make sure you have Adobe Acrobat Reader installed) submit your solutions as a pdf file on canvas