**Programming Assignment01**

|  |
| --- |
| **Submission guide**  1. Write answer following questions in this file  2. Write your code using provided python script file   * You have to complete several functions under description * Please check **TODO** |

1. Apply a multiple linear regression on the given dataset

The following code loads a dataset.

|  |
| --- |
| data = pd.read\_csv('https://drive.google.com/uc?export=download&id=1YPnojmYq\_2B\_lrAa78r\_lRy-dX\_ijpCM', sep='\t') |

The given dataset aims to predict Mean\_temperature using several explanatory variables.

[INPUT]

Max\_temperature

Min\_temperature

Dewpoint

Precipitation

Sea\_level\_pressure

Standard\_pressure

Visibility

Wind\_speed

Max\_wind\_speed

[OUTPUT]

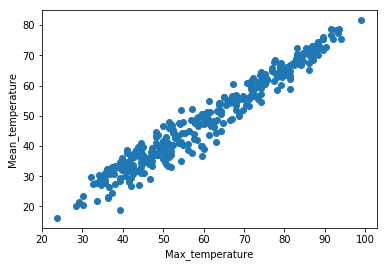
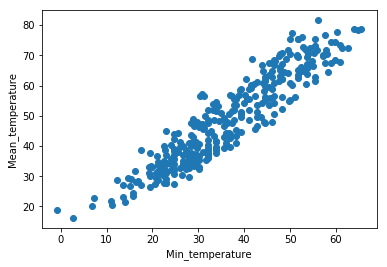
Mean\_temperature

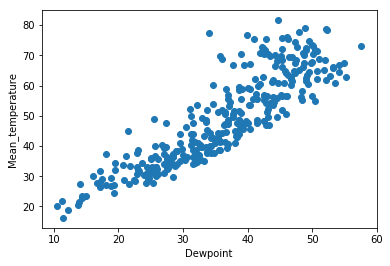
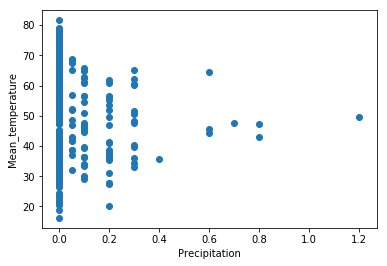
This assignment consists of two parts.

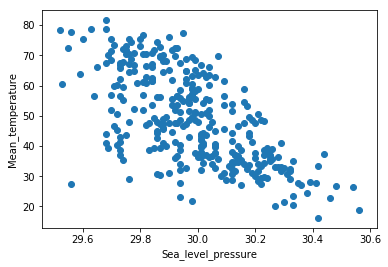
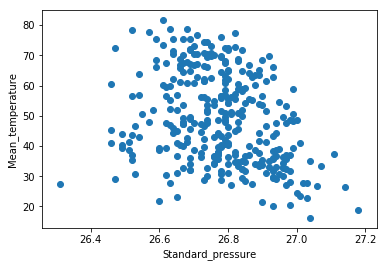
1. Explanatory analysis
2. Build a multiple linear regression model

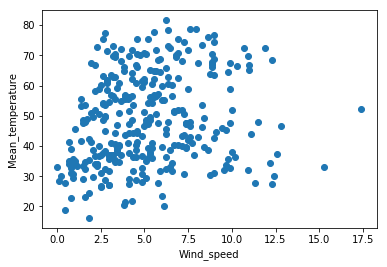
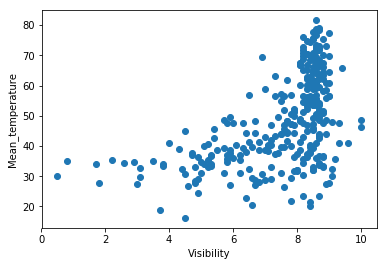
(1) Explanatory analysis (40pts)

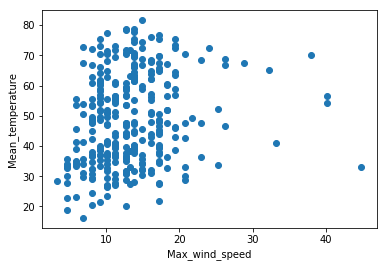
* 1. Complete code to get the answers of the following questions (15pts)
  2. Draw pairwise scatter plot – one scatter plot illustrates the relationship between an input variable and output target (Paste figures here) (5pts)

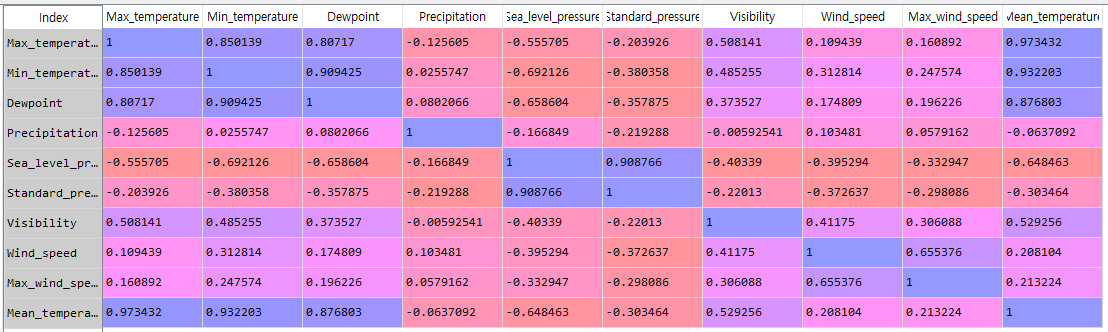




* 1. Which input variable does seem to be useful to predict the target? Why? (5pts)
     + - Max\_temperature, min\_temperature, dewpoint
       - According to the scatter plot, to predict the target, the input-output distribution of the data should show the aspect of the linear regression model. The input showing the linear relationship is max\_temperature, min\_temperature, dewpoint. sea\_level\_pressure and standard\_pressure have a weak linear relationship and the rest can not find a linear relationship.
  2. Which input variable does seem to be most irrelevant for multiple linear regression? Why? (5pts)
     + - Precipitation, wind\_speed, max\_wind\_speed
       - According to the scatter plot, these variables, such as precipitation, wind\_speed, and max\_wind speed, are independent of linear regression analysis. This is because there is no relation between the distribution of y values ​​according to the variation of x variables.
  3. Calculate correlation matrix for input variables. (5pts)

cor\_mat = data.corr()

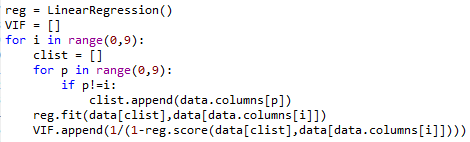
cor\_mat

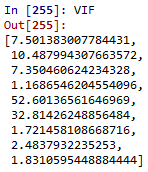


* 1. Describe what you can get from the correlation matrix. (5pts)

Through the correlation matrix, we can get the correlation between variables. The correlation coefficient exists from -1 to 1, and when the two variables are independent, the correlation coefficient is 0, but not all relations in which the correlation coefficient is 0 are independent. When the correlation coefficient is close to 1, the two variables have a positive relationship, and when they are close to -1, they have a negative relationship. As shown in the scatter plot of problem B, Max temperature, min\_temperature, and dewponit have a strong positive correlation with mean\_temperature of y value, and precipitation, wind\_speed and max\_wind\_speed have a correlation coefficient close to zero. And also we can identify the relation between x variables.

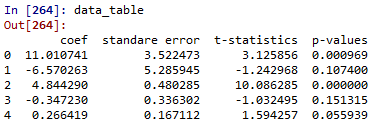
1. Build a multiple linear regression model (60pts)
   1. Complete code to get the answers of the following questions (20pts)
   2. Calculate VIF of all input variables. (5pts)





* 1. Using the variables with VIF values less than 5, build a linear regression model to predict Mean\_temperature and write the table that describes the estimated coefficients, standard errors, t-statistics, p-values of input variables and intercept term. Based on the t-test results, are there any variables insignificant to predict the target? List all the insignificant variables when significance level is 0.05. (5pts)



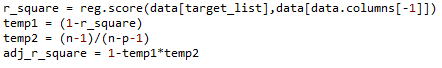


The 0th line of coef is intercept. There are other value of intercept (standard error, t-statistics, p-value) but they are useless variables

To find insignificant variables, we must remove the variable corresponding to the alternative hypothesis. Visibility should be removed because it has a smaller p-value than the significance level.

Since insignificant variables must be greater than the significance level, they are Precipitation, Wind\_speed, and Max\_wind\_speed.

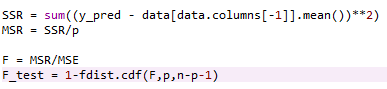
* 1. Calculate adjusted R2 of this linear regression model. (3pts)





Formula is 1-(n-1)/(n-p-1)\*(1-rsquare)

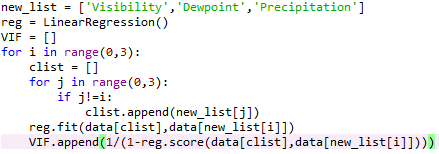
* 1. Calculate F-statistic of F-test for a linear regression model and analyze the result of F-test. (5pts)





F\_test can be regarded as significance because it has a F\_test value less than 0.05(significance level). Therefore, alternative hypothesis is adopted and four variables have a significant effect on dependent variables.

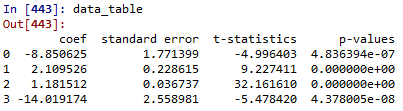
* 1. Instead of some irrelevant variables, retrain a linear regression model using “Visibility”, “Dewpoint”, “Precipitation.” Describe your opinion on whether there is a problem in training a model with this set of variables. (4pts)





When the VIF of each variable is checked, the three variables have a VIF value close to zero for each variable. In other words, it is independent for other variables and is suitable for regression analysis.

* 1. Write the table that describes the estimated coefficients, standard errors, t-statistics, p-values of input variables and intercept term in the new model. Based on the t-test results, are there any variables insignificant to predict the target? List all the insignificant variables when significance level is 0.05. (5pts)



The 0th line of coef is intercept. There are other value of intercept(standard error, t-statistics, p-value) but they are useless variables

To find insignificant variables, we must remove the variable corresponding to the alternative hypothesis. All three variables have significance levels less than 0.05. Therefore, three variables are considered to be significance variables that adopt alternative hypothesis and have a significant effect on dependent variables. There is no insignificant value.

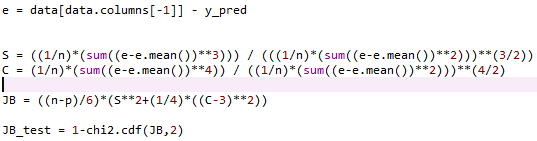
* 1. Calculate adjusted R2 of new linear regression model. (3pts)





Formula is 1-(n-1)/(n-p-1)\*(1-rsquare)

* 1. Do residuals of the new model follow the normal distribution? (significance level is 0.05) (5pts)







The significance level is 0.05. The value of JB\_test is significant at 0.0. Therefore, we should reject null hypothesis and adopt alternative hypothesis. It can also be concluded that this does not follow a normal distribution.

* 1. Do residuals of the new model satisfy homoscedasticty? (significance level is 0.05) (5pts) 



The significance level is 0.05. The value of LM\_test is significant at 0.306~. Therefore, we should adopt null hypothesis and reject alternative hypothesis. It can also be concluded that this follow the homoscedasticity.