**Programming Assignment01**

**Seungjun lee 15146317**

|  |
| --- |
| **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.

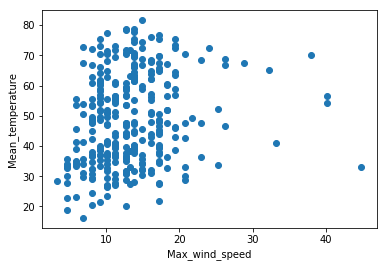
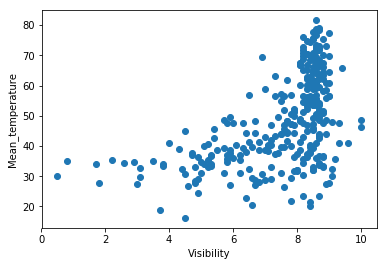
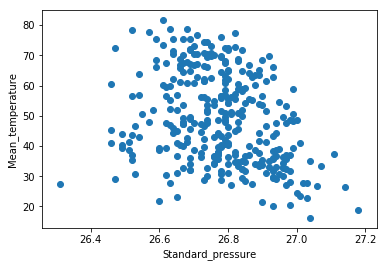
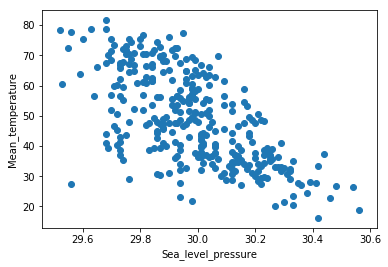
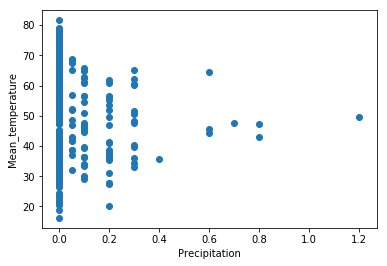
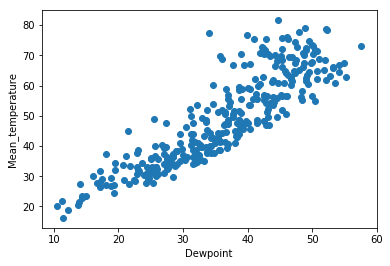
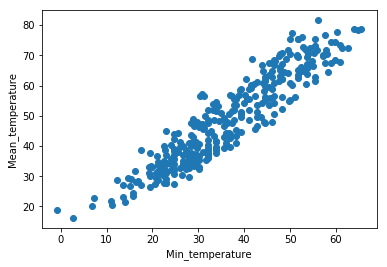
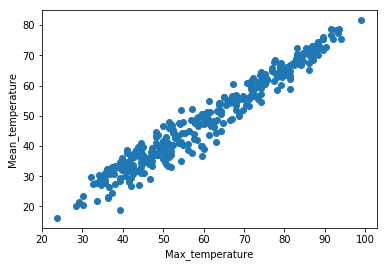
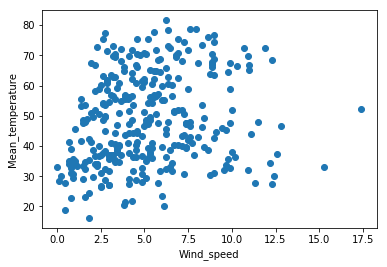
1. Explanatory analysis
2. Build a multiple linear regression model

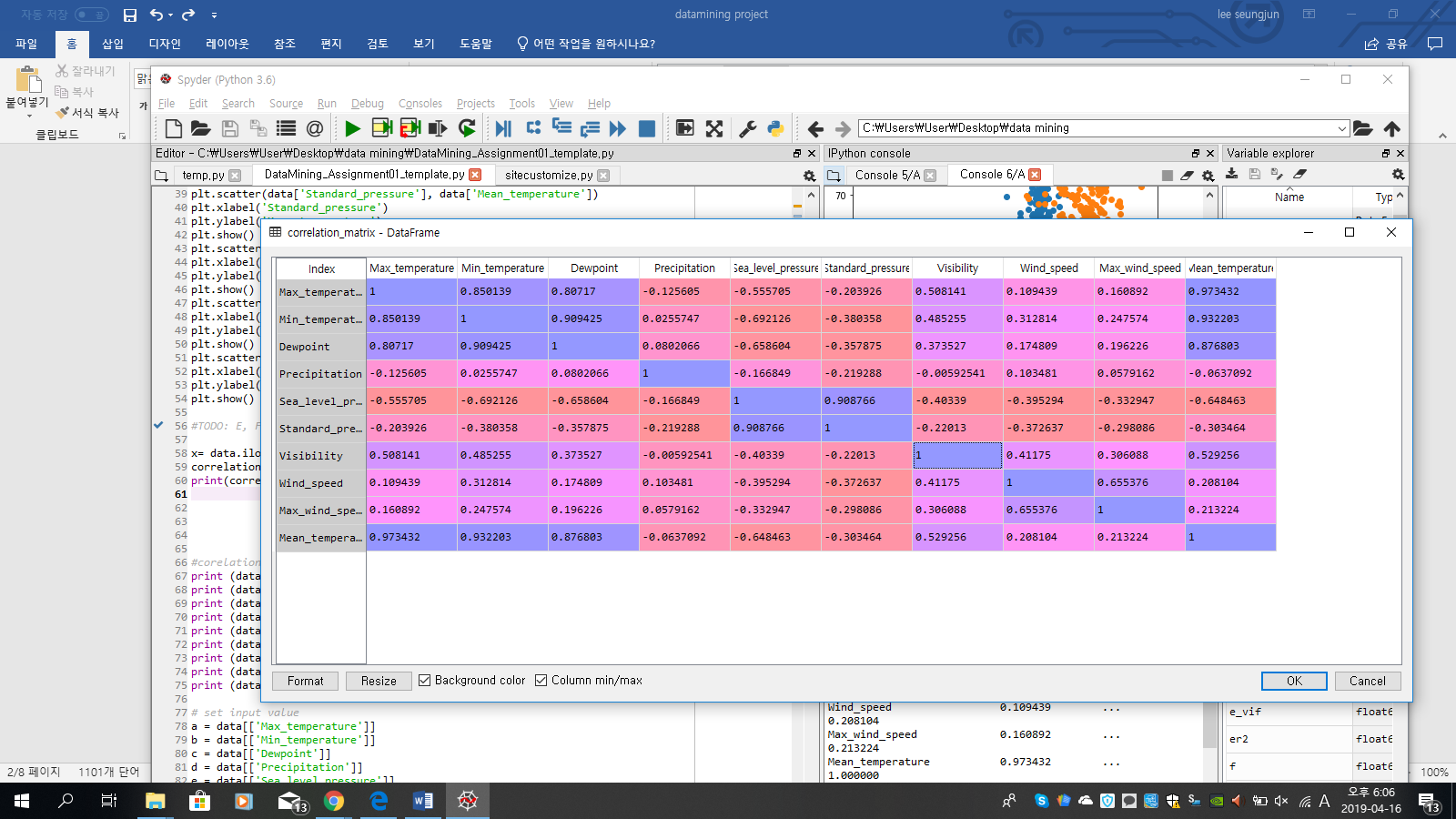
(1) Explanatory analysis (40pts)

* 1. Complete code to get the answers of the following questions (15pts)

In the python script

* 1. 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 and Min\_temperature are useful to predict the target. Because that scatter plots are similar with some linear line. In other words, Error rate with some linear line is smaller than the other value. That mean 2 value has high correlation. We can predict target value by using max and min temperature. Sea\_level\_pressure has negative relation with mean\_temperature.
  1. Which input variable does seem to be most irrelevant for multiple linear regression? Why? (5pts)
* I guess the most irrelevant input variable is Standard\_pressure. Because it plot seem to be scattered no relation with Mean\_temperature. That’s mean is we can’t estimate predict value from this value.
  1. Calculate correlation matrix for input variables. (5pts)  
     
  2. Describe what you can get from the correlation matrix. (5pts)

First, We can know degree of Relation between each variable. Specially, we know that how much associate between input and output. In other words, correlation between input data and Mean\_temperature. Or we can get relation between independent variables.

Max\_temperature has highest association with Mean\_temperature and Min\_temperature has second high association with Mean\_temperature. Also precipitation is almost no correlation with Mean\_temperature.

Second we can know that whether relation is negative or positive. Negative means that if input value will increase, output value will be decreased. Positive means that if input value will increase, output value will be increased. In other words, negative relation is inverse relation between input and output

Dewpoint has high negative correlation with Mean\_temperature.

1. Build a multiple linear regression model (60pts)
   1. Complete code to get the answers of the following questions (20pts)

In the python script

* 1. Calculate VIF of all input variables. (5pts)
     1. a\_vif(Max\_temperature)
     2. Out[104]: 7.501383007784356
     3. b\_vif(Min\_temperature)
     4. Out[105]: 10.487994307663547
     5. c\_vif(Dewpoint)
     6. Out[106]: 7.350460624234334
     7. **d\_vif(Pricipitation)**
     8. **Out[107]: 1.1686546204554096**
     9. e\_vif(Sea\_level\_pressure)
     10. Out[108]: 52.60136561647
     11. f\_vif(Standard\_pressure)
     12. Out[109]: 32.814262488565085
     13. **g\_vif(Visibility)**
     14. **Out[110]: 1.7214581086687133**
     15. **h\_vif(Wind\_speed)**
     16. **Out[111]: 2.4837932235253044**
     17. **i\_vif(Max\_wind\_speed)**
     18. **Out[112]: 1.8310595448884444**
  2. 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)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficient | T-statistics | p-value | Standard error |
| Pricipitation | -6.5702629 | -1.2429685 | 0.10740046 | 5.28594483 |
| Visibility | 4.84429037 | 10.08628481 | 0 | 0.48028491 |
| Wind\_speed | -0.34722997 | -1.03249489 | 0.15131495 | 0.33630188 |
| Max\_wind\_speed | 0.26641892 | 1.59425685 | 0.05593909 | 0.16711167 |

se

Out[674]: array([5.28594483, 0.48028491, 0.33630188, 0.16711167])

t

Out[675]: array([-1.2429685 , 10.08628481, -1.03249489, 1.59425685])

1-tdist.cdf(np.abs(t), n-p-1) =p-value

Out[720]: array([0.10740046, 0. , 0.15131495, 0.05593909])

new\_reg.coef\_

Out[141]: array([-6.5702629 , 4.84429037, -0.34722997, 0.26641892])

new\_reg.intercept\_

Out[142]: 11.010741432417959

tdist.ppf(1-alpha/2,n-p-1) = x-value of significance level(alpha) in the t-distribution

Out[732]: 1.9674995188238251

np.abs(t)

Out[731]: array([ 1.2429685 , 10.08628481, 1.03249489, 1.59425685])

- alpha=0.05, p-value of visibility is 0 so that is smaller than alpha so only visibility is significance variable. Other is insignificance variable.

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

adj\_r2

Out[284]: 0.28051654638686874

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

f-statistics

Out[630]: 32.19088228958253

1-fdist.cdf(f,p,n-p-1) # f-statistics P-value

Out[897]: 1.1102230246251565e-16

- usually insignificance level is 0.05 or 0.1 in the F-test. eventually p-value of F-test is smaller than 0.05. this overall event is not occurred by coincidence so this linear regression model is significant model in the overall regression model. We select this alternative hypothesis, instead of null hypothesis.

* 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)

Vi\_vif

Out[33]: 1.15925073182583

Dew\_vif

Out[34]: 1.171393221487938

Pre\_vif

Out[35]: 1.007992992111563

SSE first Linear Regression

Out[47]: 50990.696864255275

SSE1 second Linear Regression

Out[48]: 12057.684772230774

These value of vif are almost 1. Each value doesn’t have relation with other value. These mean that these value are reliable value in regression model.

SSE of first model is bigger than second model. It means that error of second model between predict value and target value is decreased. I guess This set of value is appropriate value for regression model

* 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)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficients | Standard errors | t-statistics | p-value |
| Visibility | 2.10952566 | 0.47087381 | 9.21284547 | 0 |
| Dewpoint | 1.1815122 | 0.07566587 | 32.1108415 | 0 |
| Precipitation | -14.01917353 | 5.27068014 | -5.46977247 | 0.000000004 |

sec\_reg.intercept\_

Out[808]: -8.850624834617449

tdist.ppf(1-alpha/2,n-p-1)

Out[835]: 1.9674756580272579

no insignificant variables to predict the target. Because p-values of these input variable are almost 0, it means that this event is not coincidence events. Also t-statistics is high. So these values are reliability value in this regression model

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

adj\_r2

Out[887]: 0.2827862102783929

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

JB ------JB statistics value

Out[196]: 84.43370160357776

1-chi2.cdf(JB,2)----------p-value of chi2

Out[199]: 0.0

P-value is 0. P-value is smaller than significance level. So this residual of the new model doesn’t follow normal distribution

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

LM------------------🡪Lagrange Multiplier statistic

Out[203]: 3.609704425237884

1-chi2.cdf(LM,p)----------🡪p-value of LM

Out[204]: 0.30681006706428304

In this case, null hypothesis is this model is satisfy homoskedasticity

p-value of LM is bigger than significance level. This null hypothesis will not be rejected. So it is satisfy homoskedasticity.