DAEHYUN KO

OGTIP PROJECT 3
LINEAR REGRESSION ANALYSIS

TABLE OF CONTENTS

1. PROBLEM STATEMENT
2. DATA DICTIONARY
3. LINEAR REGRESSION
4. RESIDUAL ANALYSIS
5. SUMMARY

PROBLEM STATEMENT

BOOMBIKES DEMAND PREDICTION

Boombikes is a US bike sharing provider who wishes to understand which factors affects the demand of the bikes. And if they can determine future demand predictions.

The main problem statements

- Which variables are significant in predicting the demand for shared bikes.
- How well those variables describe the bike demands

DATA DICTIONARY

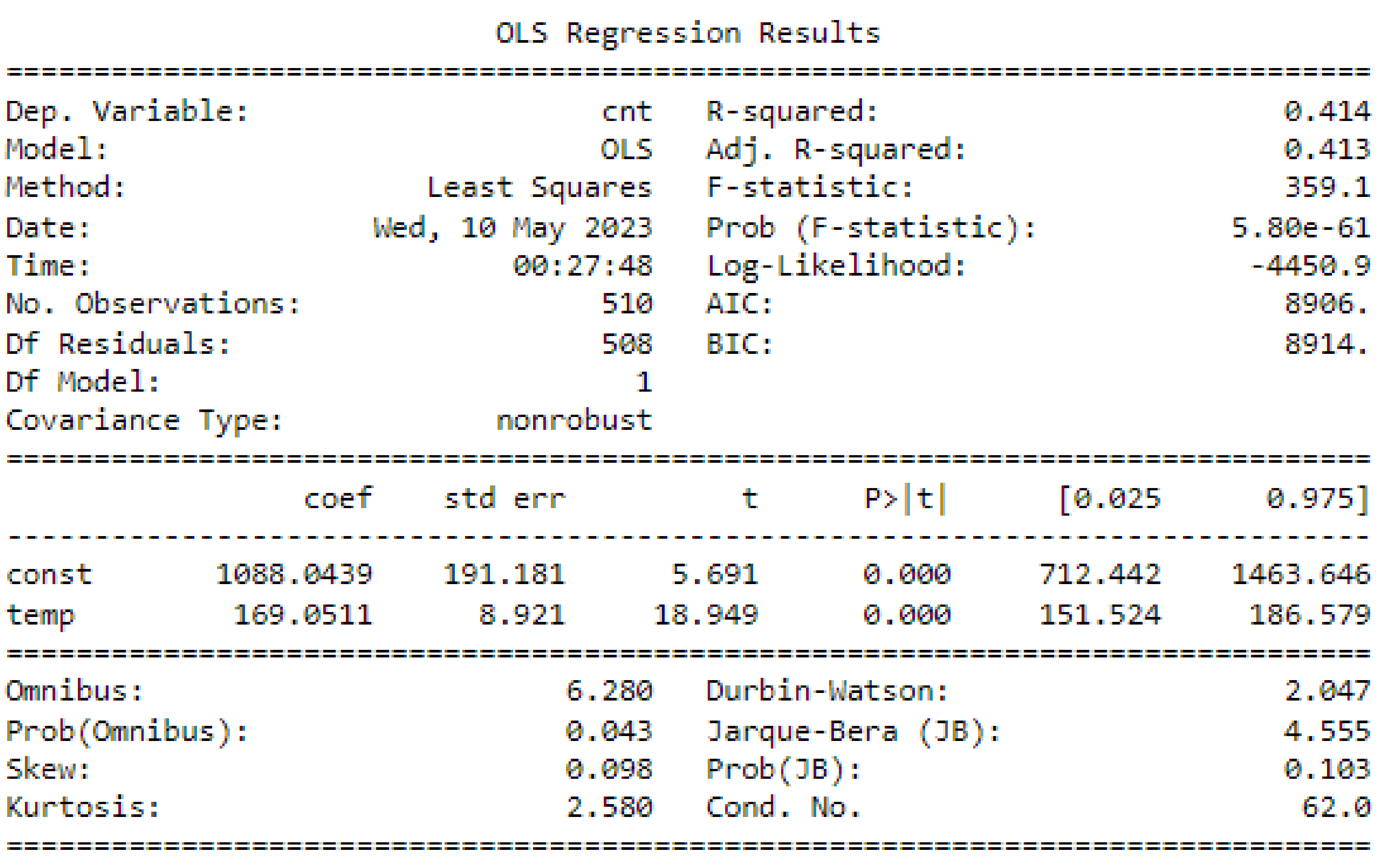
- **INSTANT:** RECORD INDEX
- **DTEDAY :** DATE
- **SEASON :** SEASON (1:SPRING, 2:SUMMER, 3:FALL, 4:WINTER)
- **YR :** YEAR (0: 2018, 1:2019)
- **MNTH :** MONTH (1 TO 12)
- **HOLIDAY :** WEATHER DAY IS A HOLIDAY OR NOT
- **WEEKDAY :** DAY OF THE WEEK
- **WORKINGDAY :** IF DAY IS NEITHER WEEKEND NOR HOLIDAY IS 1, OTHERWISE IS 0.

+ WEATHERSIT :

- 1: CLEAR, FEW CLOUDS, PARTLY CLOUDY, PARTLY CLOUDY
- 2: MIST + CLOUDY, MIST + BROKEN CLOUDS, MIST + FEW CLOUDS, MIST
- 3: LIGHT SNOW, LIGHT RAIN + THUNDERSTORM + SCATTERED CLOUDS, LIGHT RAIN + SCATTERED CLOUDS
- 4: HEAVY RAIN + ICE PALLETS + THUNDERSTORM + MIST, SNOW + FOG
- **TEMP :** TEMPERATURE IN CELSIUS
- **ATEMP:** FEELING TEMPERATURE IN CELSIUS
- **HUM:** HUMIDITY
- **WINDSPEED:** WIND SPEED
- **CASUAL:** COUNT OF CASUAL USERS
- **REGISTERED:** COUNT OF REGISTERED USERS
- **CNT:** COUNT OF TOTAL RENTAL BIKES INCLUDING BOTH CASUAL AND REGISTERED

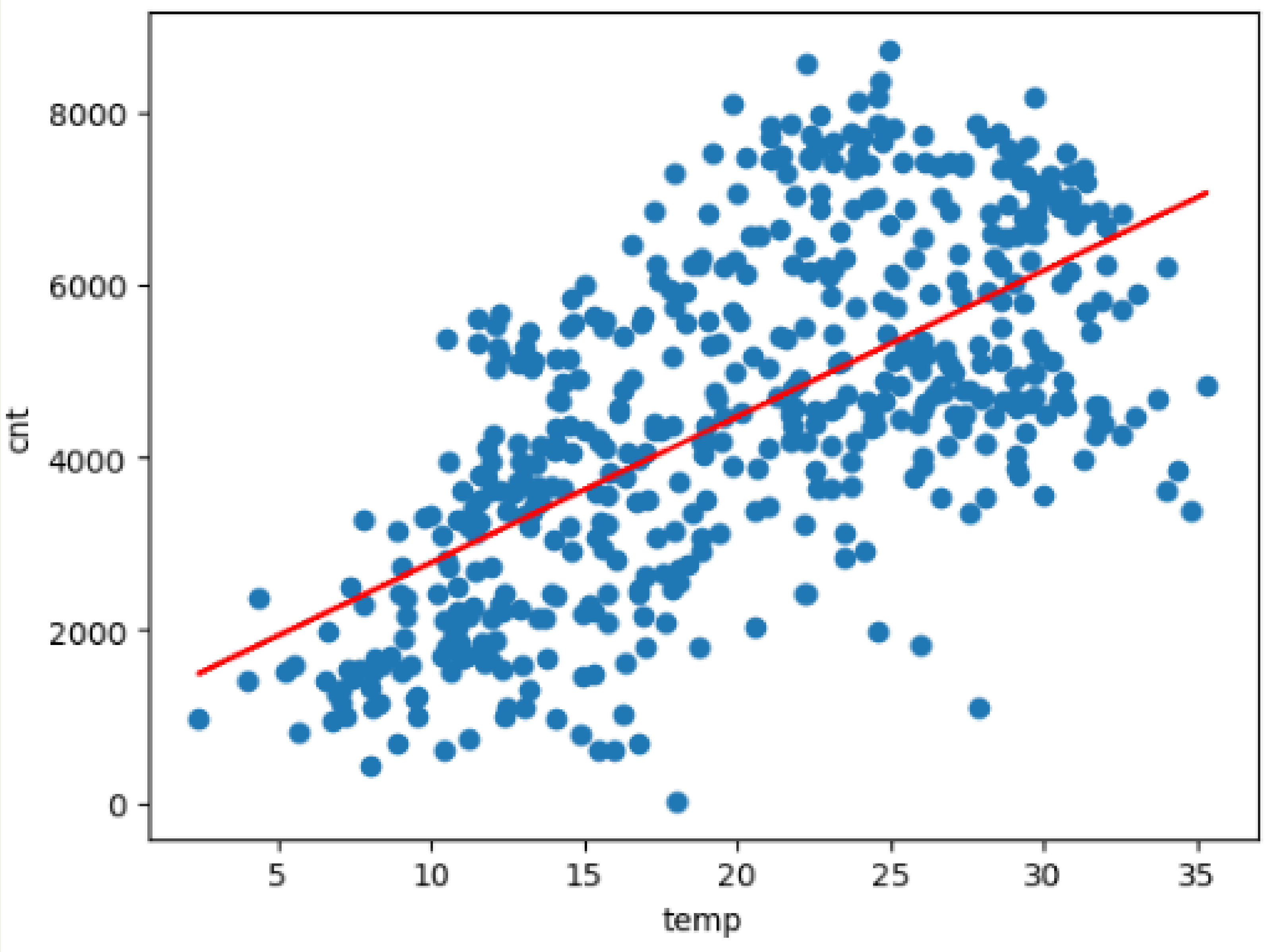
LINEAR REGRESSION ANALYSIS

'TEMP'



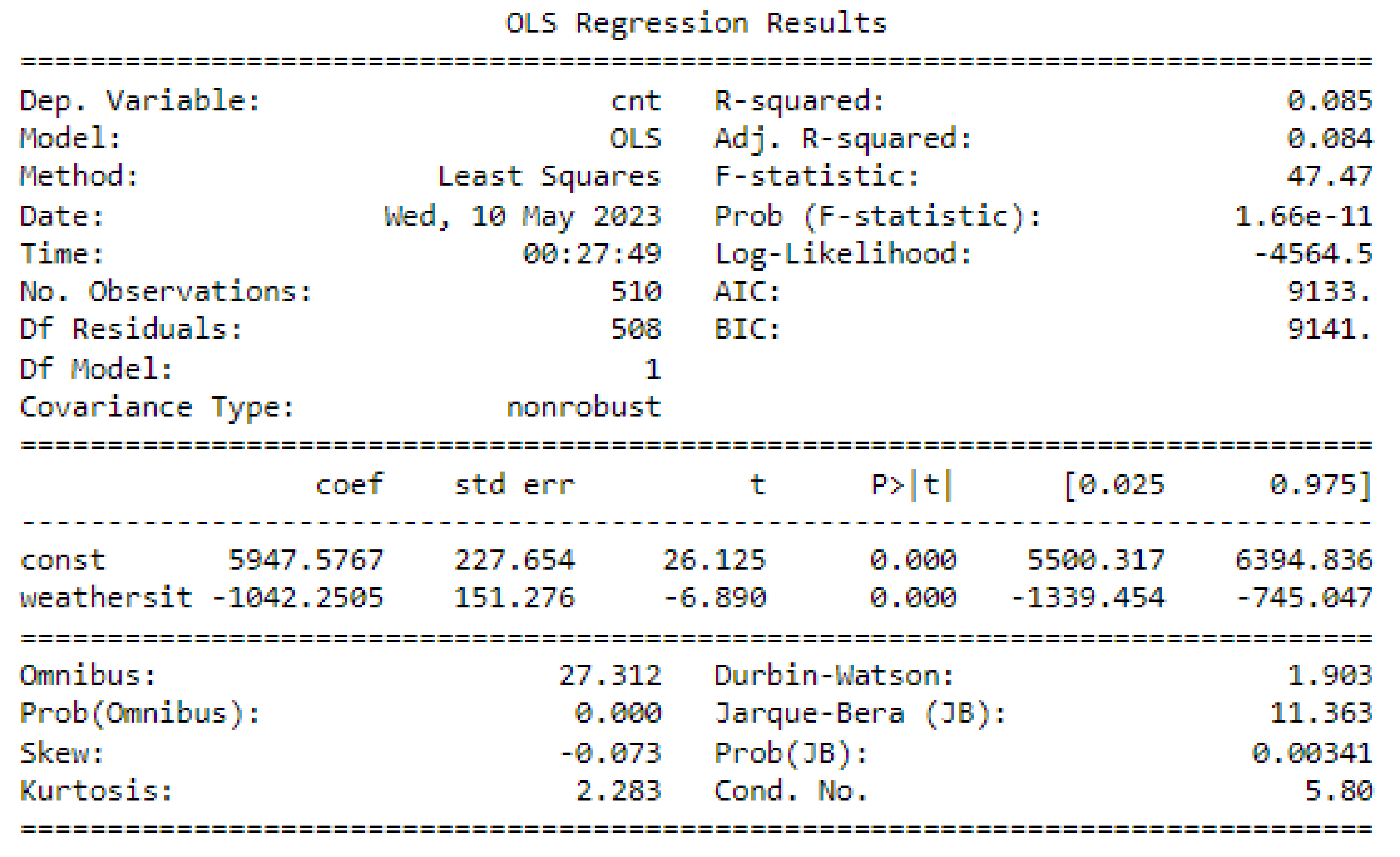
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.414
Model:                            OLS   Adj. R-squared:                  0.413
Method:                 Least Squares   F-statistic:                     359.1
Date:                Wed, 10 May 2023   Prob (F-statistic):           5.80e-61
Time:                        00:27:48   Log-Likelihood:                -4450.9
No. Observations:                 510   AIC:                             8906.
Df Residuals:                     508   BIC:                             8914.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1088.0439    191.181      5.691      0.000     712.442    1463.646
temp         169.0511      8.921     18.949      0.000     151.524     186.579
==============================================================================
Omnibus:                        6.280   Durbin-Watson:                   2.047
Prob(Omnibus):                  0.043   Jarque-Bera (JB):                4.555
Skew:                           0.098   Prob(JB):                        0.103
Kurtosis:                       2.580   Cond. No.                         62.0
==============================================================================
```

MODEL



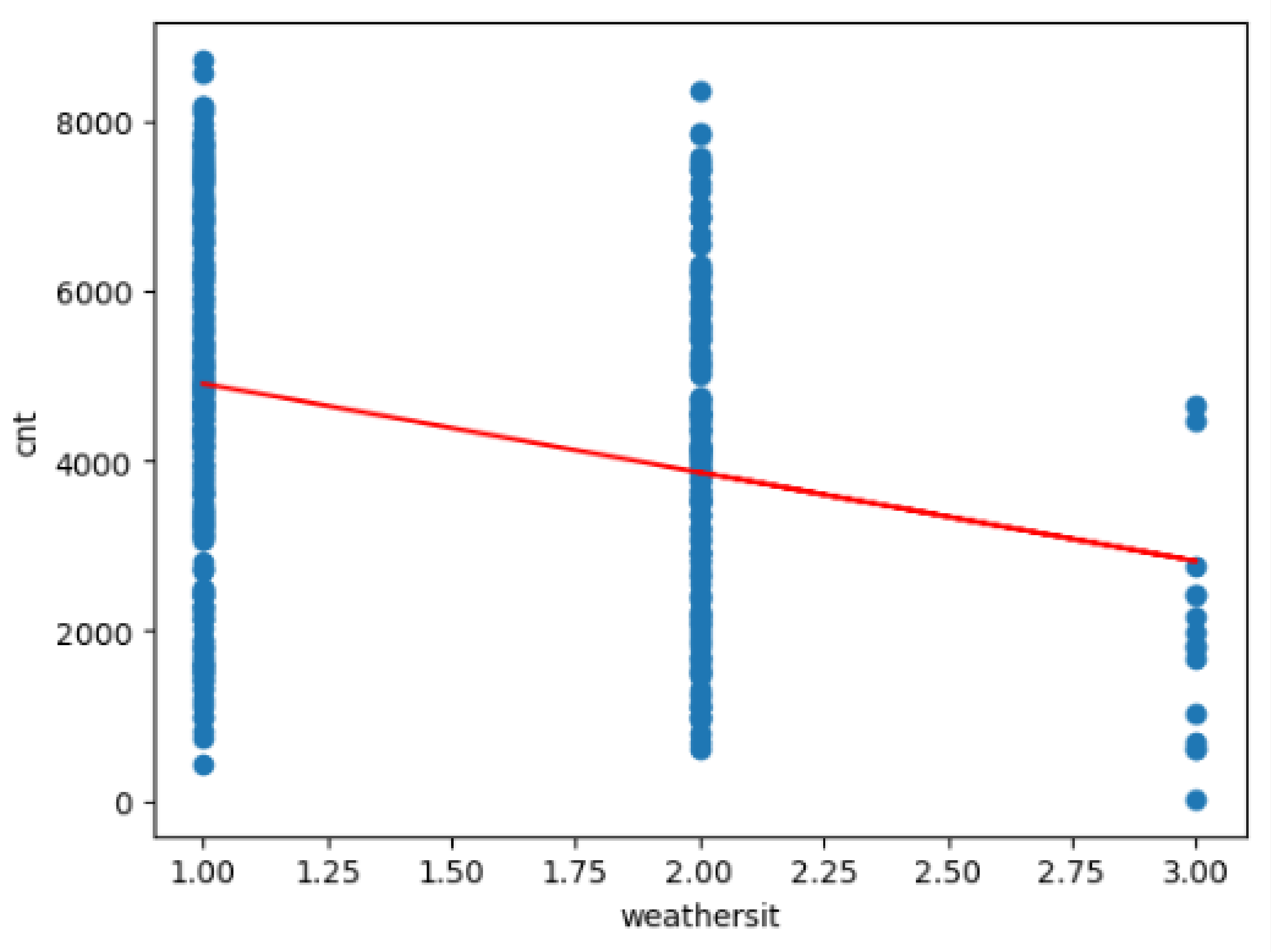
- demand increases as outside temperature increases

'WEATHERSIT'



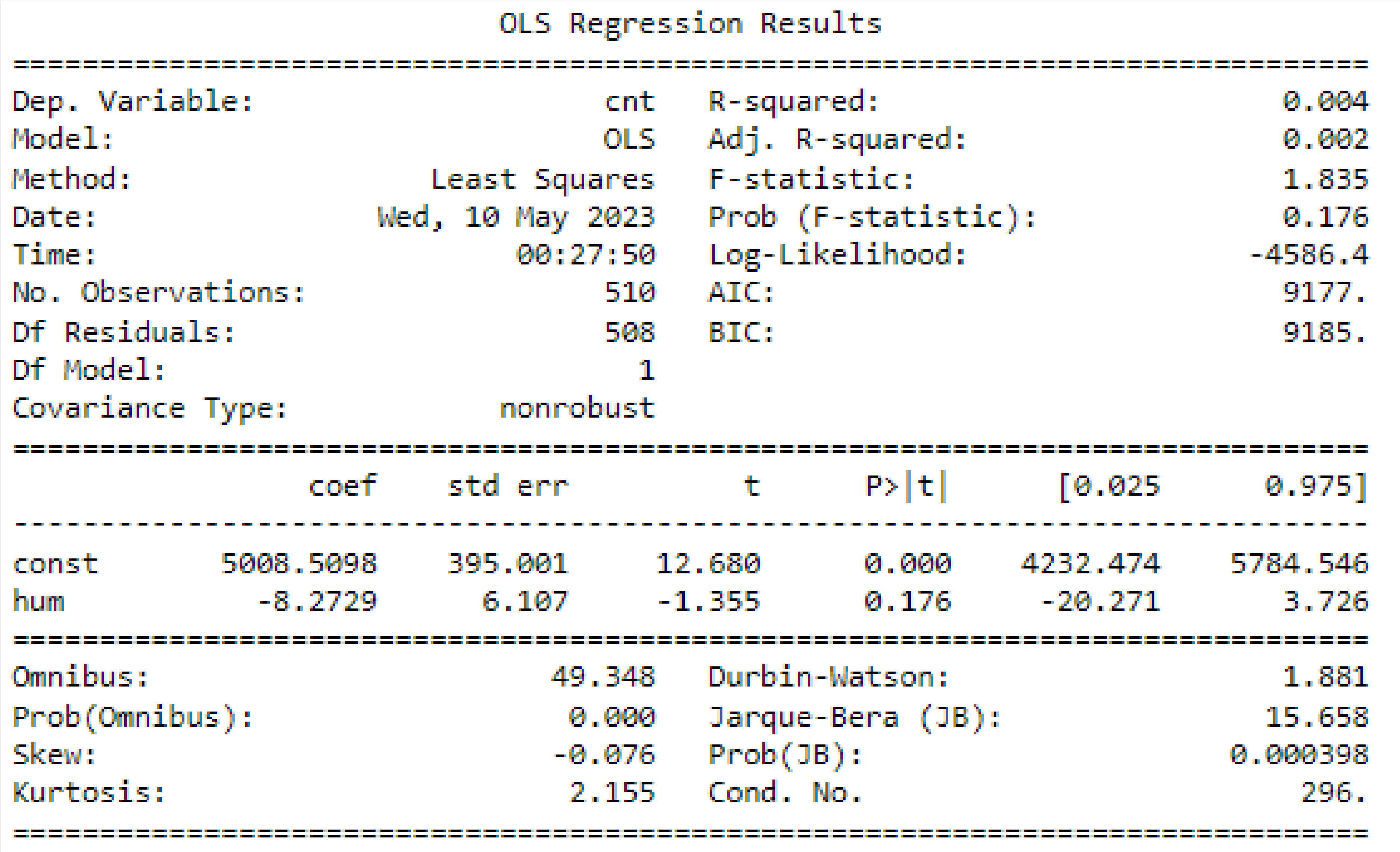
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.085
Model:                            OLS   Adj. R-squared:                  0.084
Method:                 Least Squares   F-statistic:                     47.47
Date:                Wed, 10 May 2023   Prob (F-statistic):           1.66e-11
Time:                        00:27:49   Log-Likelihood:                -4564.5
No. Observations:                 510   AIC:                             9133.
Df Residuals:                     508   BIC:                             9141.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       5947.5767    227.654     26.125      0.000    5500.317    6394.836
weathersit -1042.2505    151.276     -6.890      0.000   -1339.454    -745.047
==============================================================================
Omnibus:                       27.312   Durbin-Watson:                   1.903
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               11.363
Skew:                          -0.073   Prob(JB):                      0.00341
Kurtosis:                       2.283   Cond. No.                         5.80
==============================================================================
```

MODEL



- **demand is highest on clear days, medium on cloudy days and worst on rainy, snowy days**

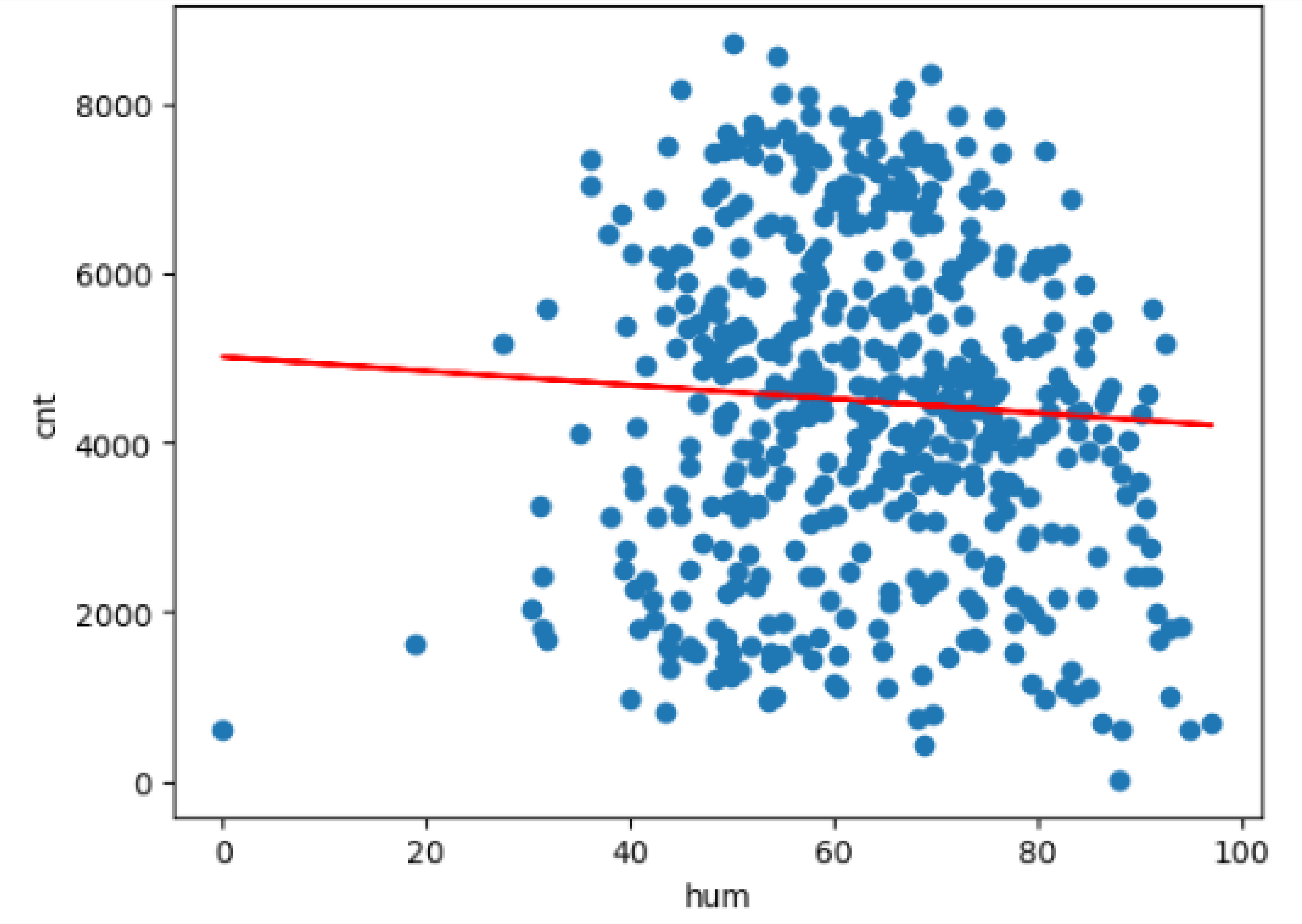
'HUM'(HUMIDITY)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     1.835
Date:                Wed, 10 May 2023   Prob (F-statistic):              0.176
Time:                        00:27:50   Log-Likelihood:                -4586.4
No. Observations:                 510   AIC:                             9177.
Df Residuals:                     508   BIC:                             9185.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       5008.5098    395.001     12.680      0.000    4232.474    5784.546
hum           -8.2729      6.107     -1.355      0.176     -20.271       3.726
==============================================================================
Omnibus:                       49.348   Durbin-Watson:                   1.881
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               15.658
Skew:                          -0.076   Prob(JB):                     0.000398
Kurtosis:                       2.155   Cond. No.                         296.
==============================================================================
```

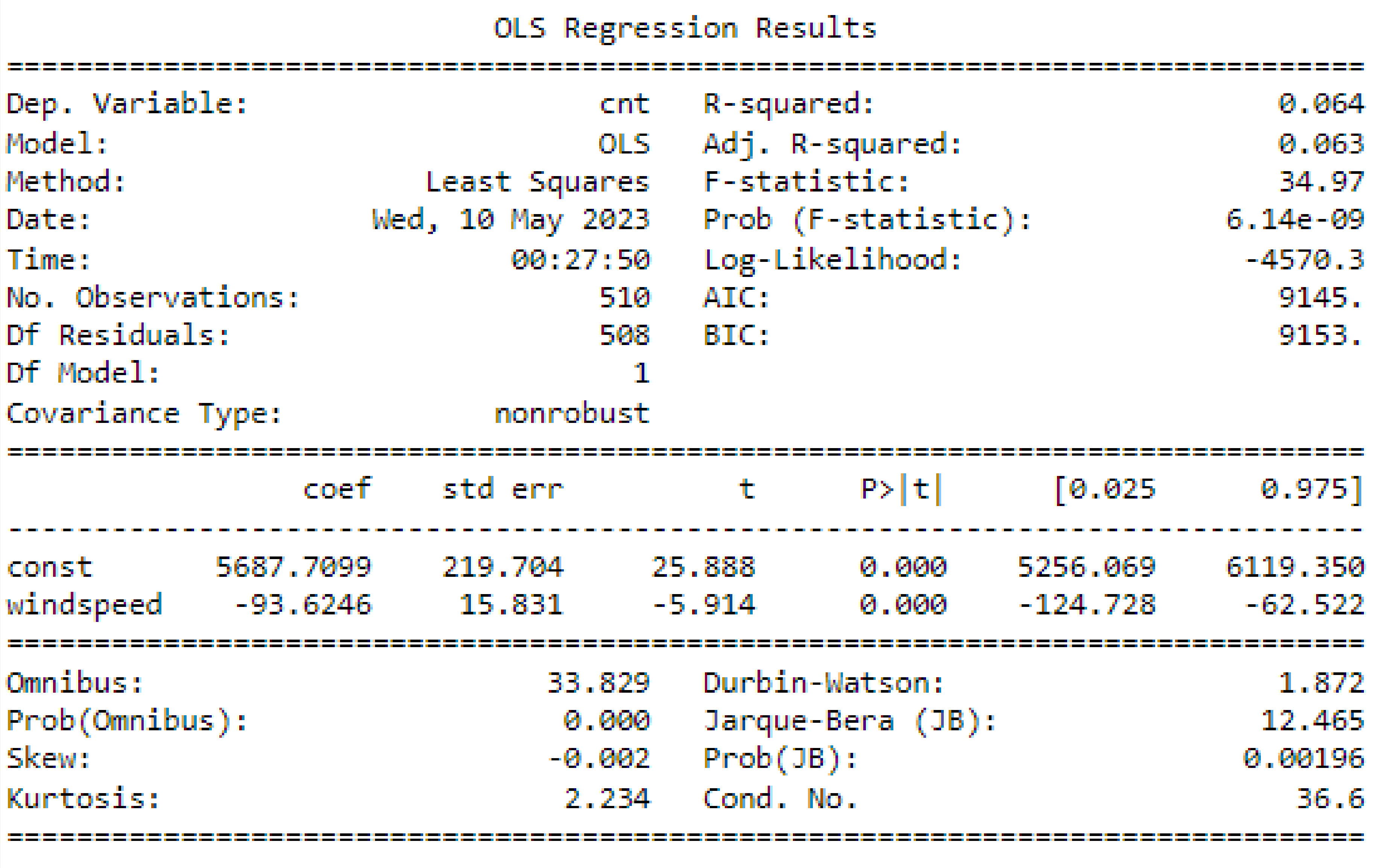
Things to note:
p-value is >0.05 thus not statistically significant

MODEL



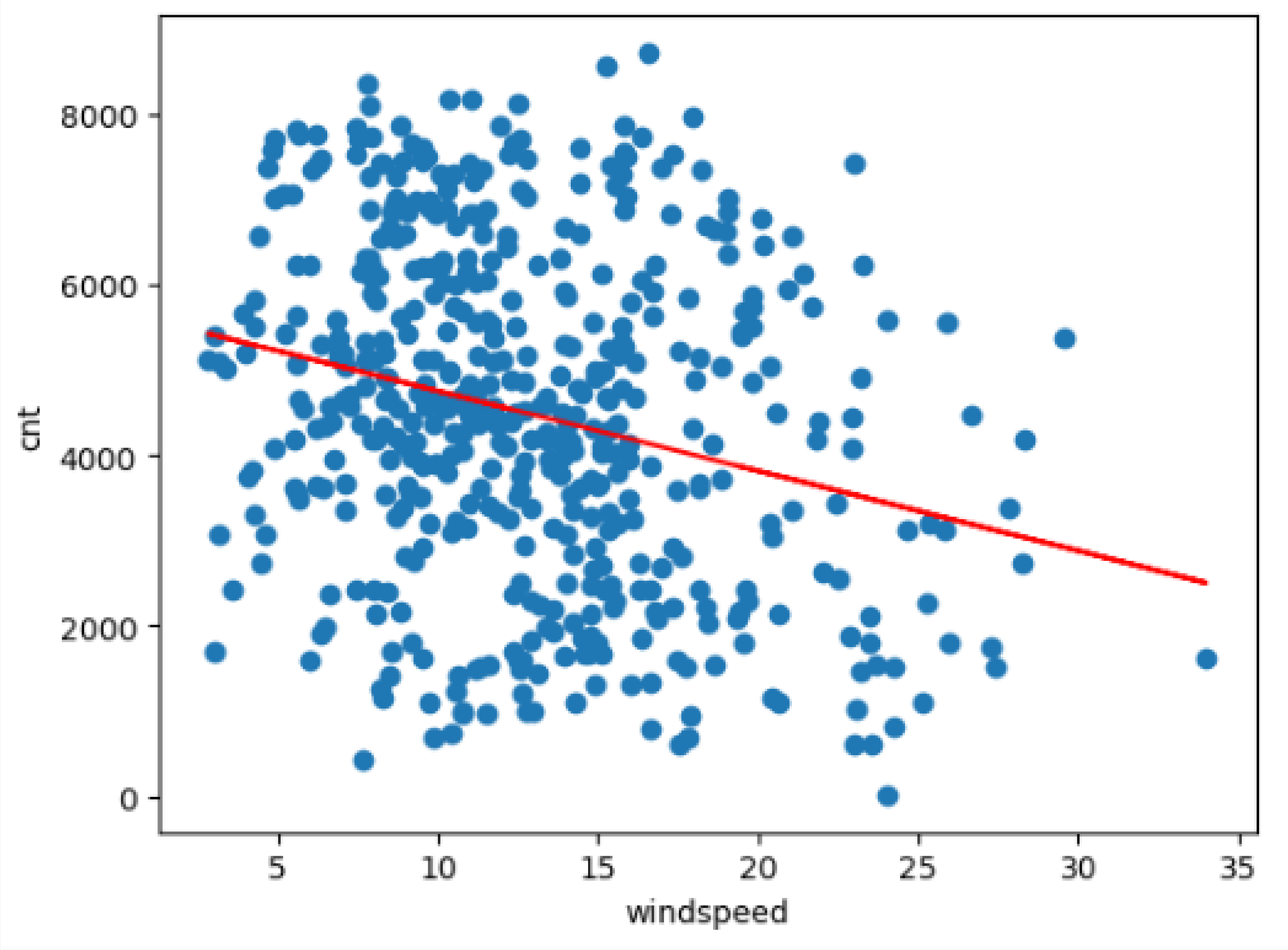
- demand for bikes are slightly less likely on humid days

'WINDSPEED'



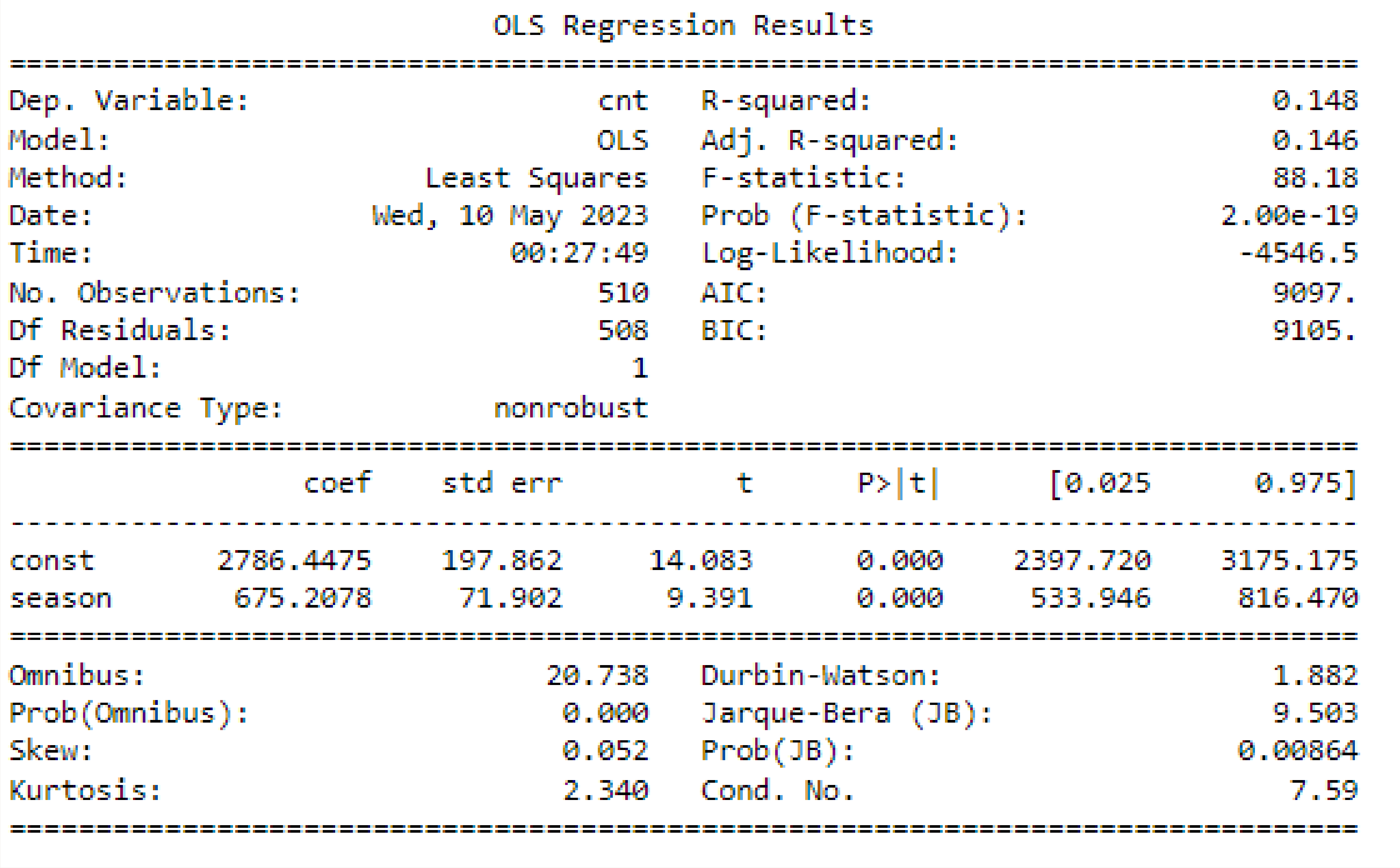
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.064
Model:                            OLS   Adj. R-squared:                  0.063
Method:                 Least Squares   F-statistic:                     34.97
Date:                Wed, 10 May 2023   Prob (F-statistic):           6.14e-09
Time:                        00:27:50   Log-Likelihood:                -4570.3
No. Observations:                 510   AIC:                             9145.
Df Residuals:                     508   BIC:                             9153.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       5687.7099    219.704     25.888      0.000    5256.069    6119.350
windspeed    -93.6246     15.831     -5.914      0.000    -124.728     -62.522
==============================================================================
Omnibus:                       33.829   Durbin-Watson:                   1.872
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               12.465
Skew:                          -0.002   Prob(JB):                      0.00196
Kurtosis:                       2.234   Cond. No.                         36.6
==============================================================================
```

MODEL



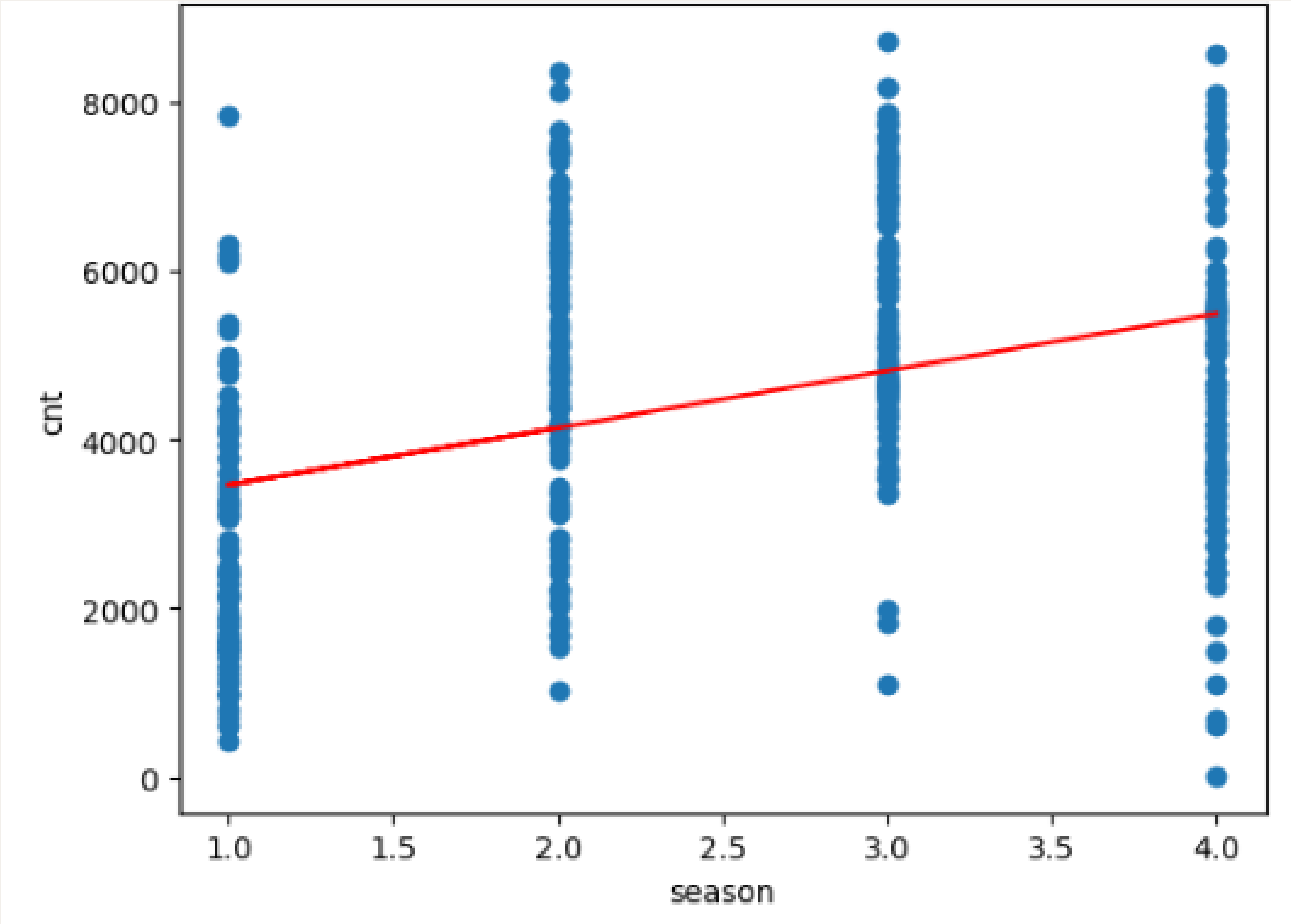
- demand lowers as windspeed increases

'SEASON'



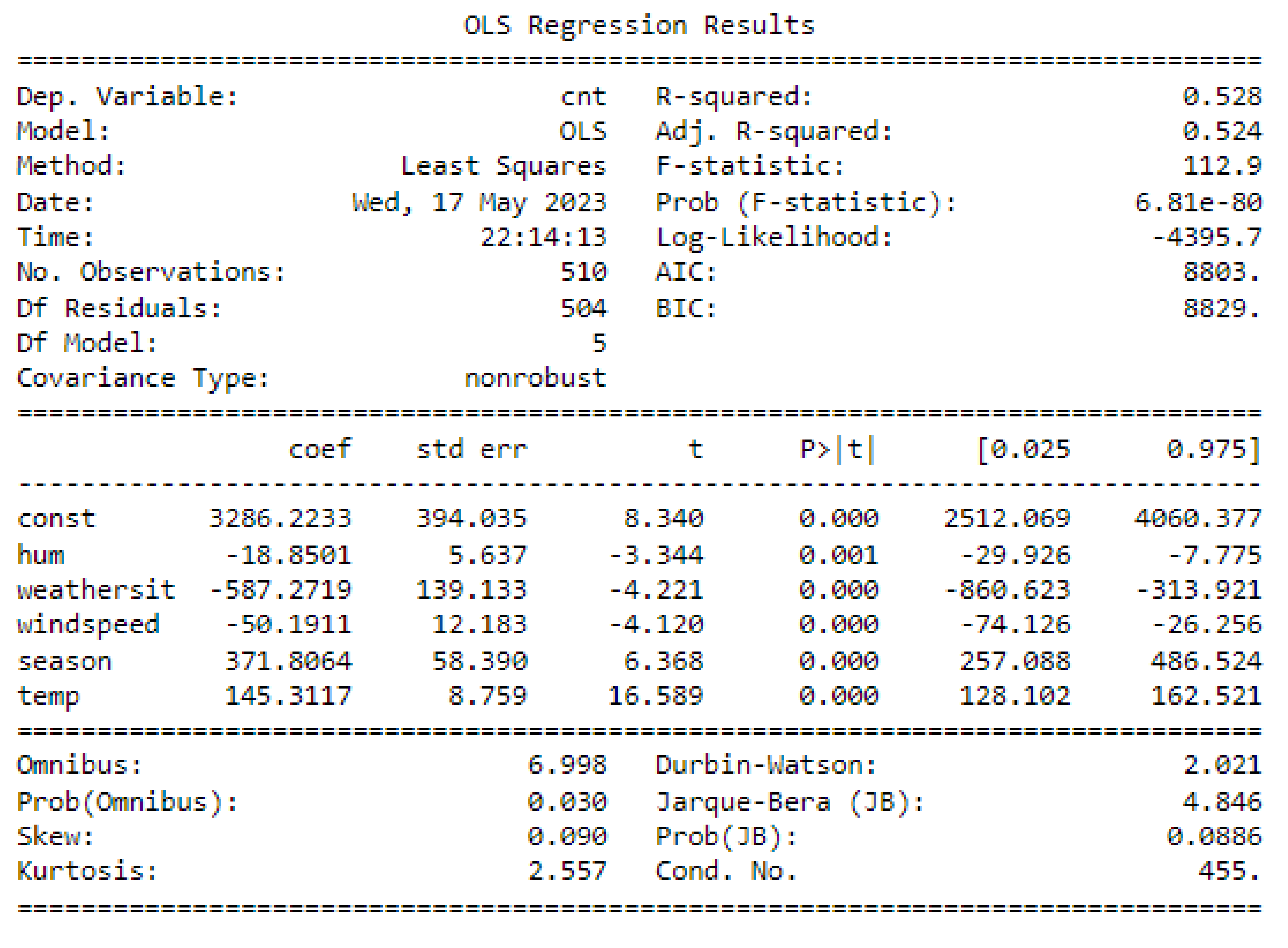
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.148
Model:                            OLS   Adj. R-squared:                  0.146
Method:                 Least Squares   F-statistic:                     88.18
Date:                Wed, 10 May 2023   Prob (F-statistic):           2.00e-19
Time:                        00:27:49   Log-Likelihood:                -4546.5
No. Observations:                 510   AIC:                             9097.
Df Residuals:                     508   BIC:                             9105.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2786.4475    197.862     14.083      0.000    2397.720    3175.175
season       675.2078     71.902      9.391      0.000     533.946     816.470
==============================================================================
Omnibus:                       20.738   Durbin-Watson:                   1.882
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                9.503
Skew:                           0.052   Prob(JB):                      0.00864
Kurtosis:                       2.340   Cond. No.                         7.59
==============================================================================
```

MODEL



- **demand is non-fluctuant through the seasons**

MULTIPLE LINEAR REGRESSION ANALYSIS



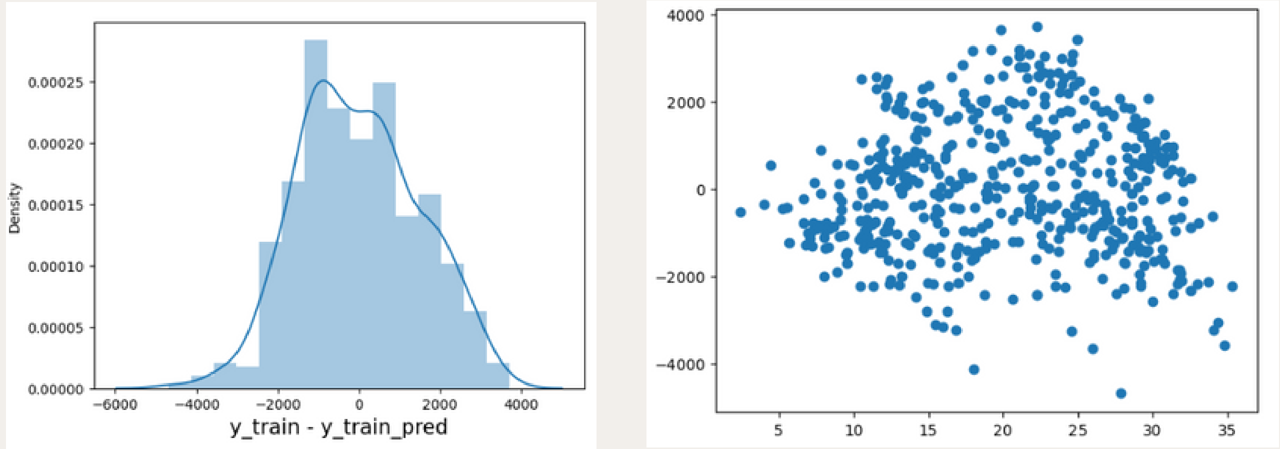
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.528
Model:                            OLS   Adj. R-squared:                  0.524
Method:                 Least Squares   F-statistic:                     112.9
Date:                Wed, 17 May 2023   Prob (F-statistic):           6.81e-80
Time:                        22:14:13   Log-Likelihood:                -4395.7
No. Observations:                 510   AIC:                             8803.
Df Residuals:                     504   BIC:                             8829.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3286.2233    394.035      8.340      0.000    2512.069    4060.377
hum          -18.8501      5.637     -3.344      0.001     -29.926      -7.775
weathersit  -587.2719    139.133     -4.221      0.000    -860.623    -313.921
windspeed    -50.1911     12.183     -4.120      0.000     -74.126     -26.256
season       371.8064     58.390      6.368      0.000     257.088     486.524
temp         145.3117      8.759     16.589      0.000     128.102     162.521
==============================================================================
Omnibus:                        6.998   Durbin-Watson:                   2.021
Prob(Omnibus):                  0.030   Jarque-Bera (JB):                4.846
Skew:                           0.090   Prob(JB):                       0.0886
Kurtosis:                       2.557   Cond. No.                         455.
==============================================================================
```

R-Squared Value is 0.528. 52.8% of the variance in demand is explained by the independent variables.

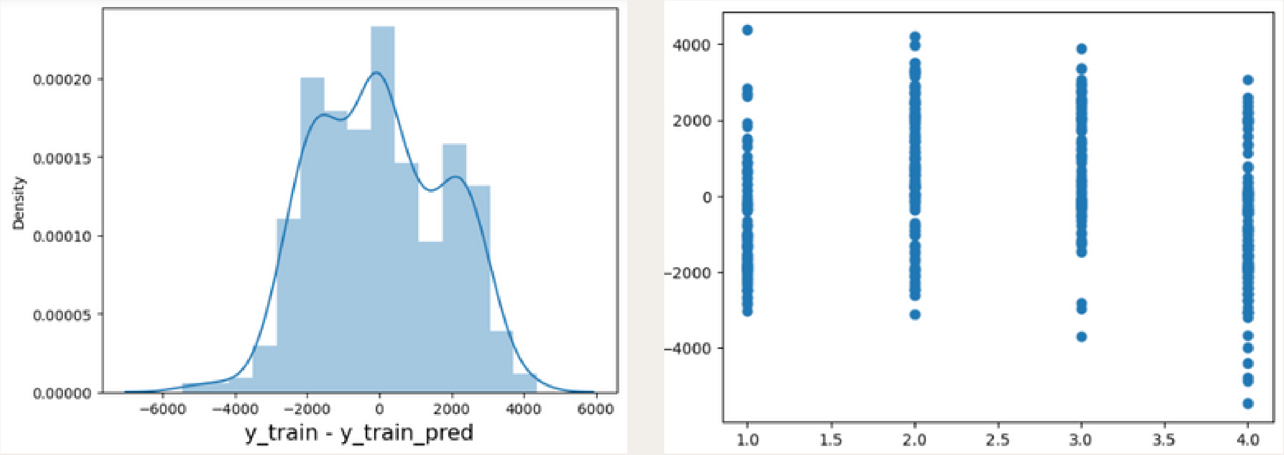
All p values indicate all variables together are statistically significant.

RESIDUAL ANALYSIS

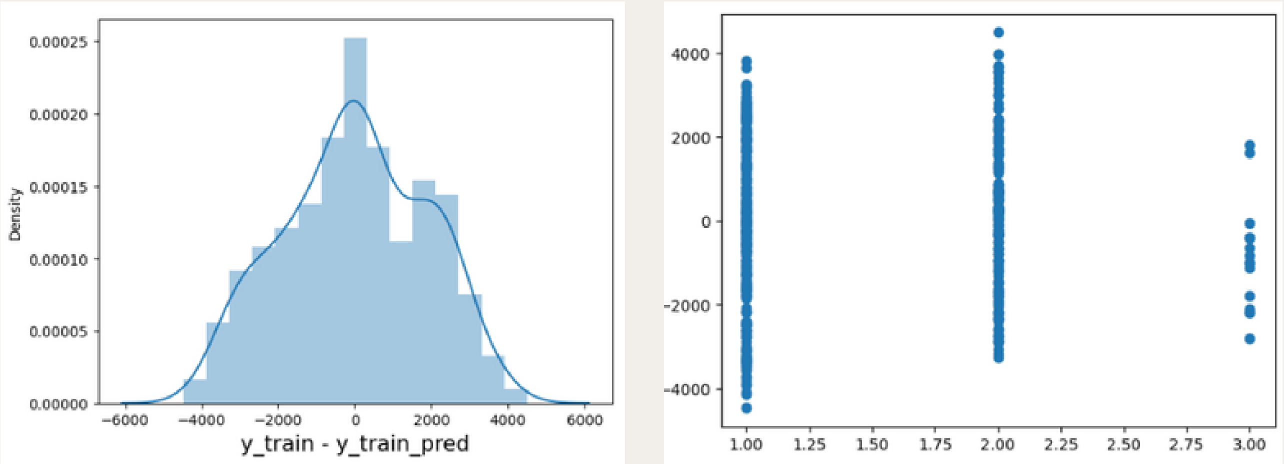
'temp'



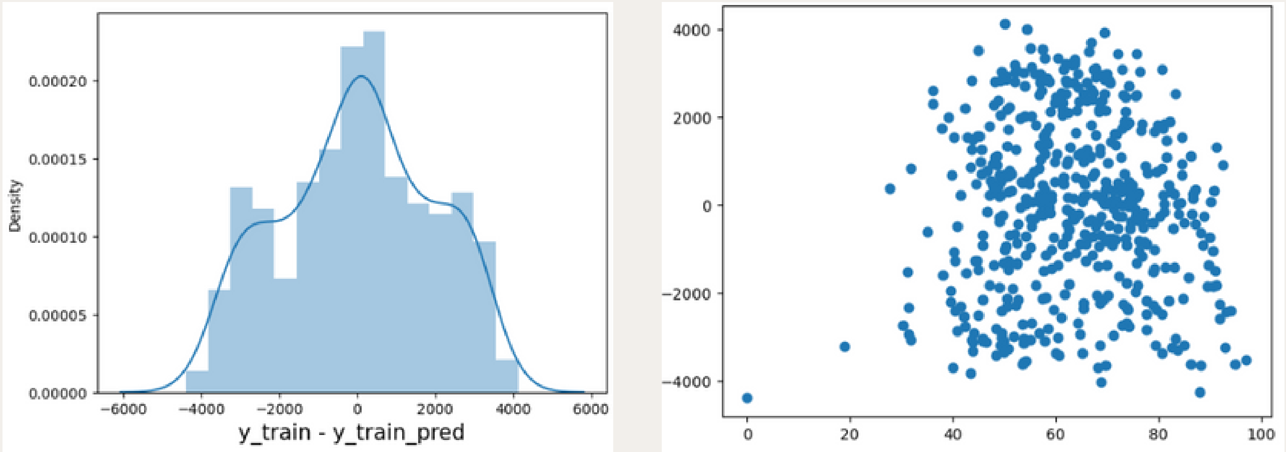
'season'



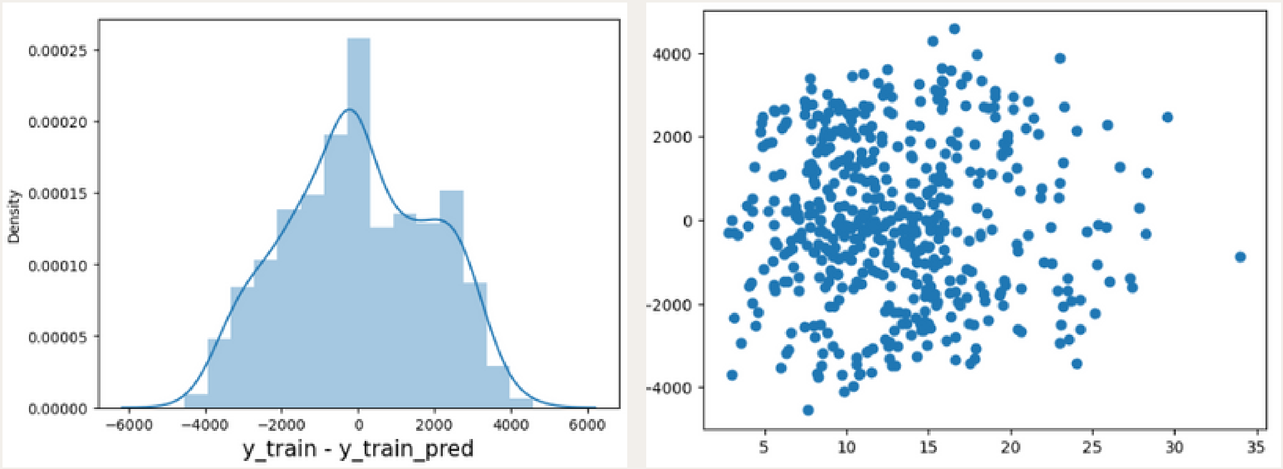
'weathersit'



'hum'



'windspeed'



INFERENCE

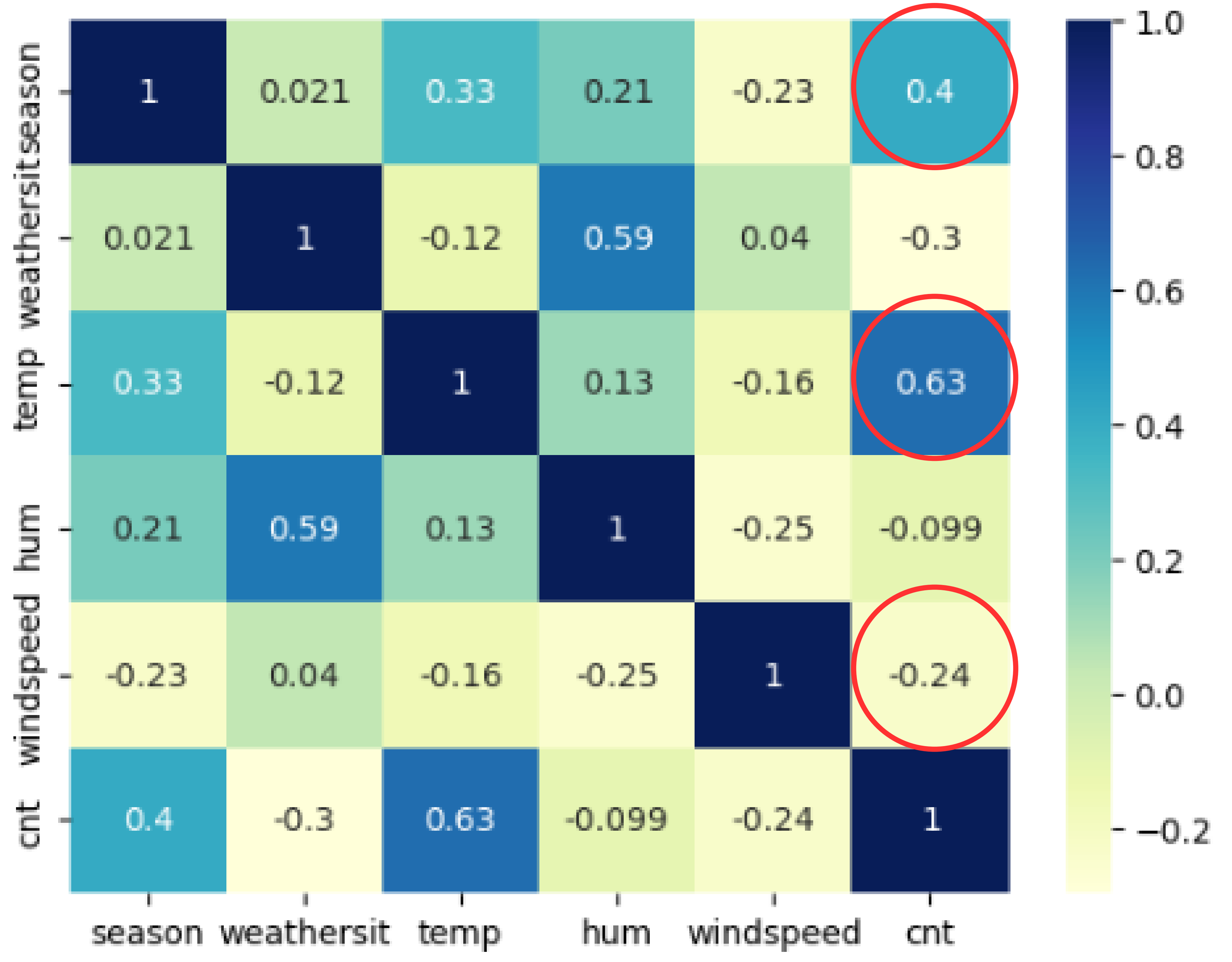
- all variables are normally distributed except 'season' and 'temp'
- all variables has significant amount of residuals

CORRELATION HEATMAP

INFERENCE

The top 3 variables with the highest correlation to the target variable are

1. **temp (temperature)**
2. **season**
3. **windspeed**



	season	weathersit	temp	hum	windspeed	cnt
season	1	0.021	0.33	0.21	-0.23	0.4
weathersit	0.021	1	-0.12	0.59	0.04	-0.3
temp	0.33	-0.12	1	0.13	-0.16	0.63
hum	0.21	0.59	0.13	1	-0.25	-0.099
windspeed	-0.23	0.04	-0.16	-0.25	1	-0.24
cnt	0.4	-0.3	0.63	-0.099	-0.24	1

OVERVIEW

Based on the findings, we can conclude that the outside temperature is the variable with the highest predictor for demand of bikes. If outside temperatures are hot, people are more likely to rent out bikes. Furthermore, the demand is significantly affected by the season as well. The company tend to have higher demand during hot seasons compared to cold seasons. Lastly, outside factors such as the weather and wind speed also affects the demand of bikes. Bikes are more likely to be rented out on clear less windy days. Boombikes have to take account all these variables in order to predict the their future demands.