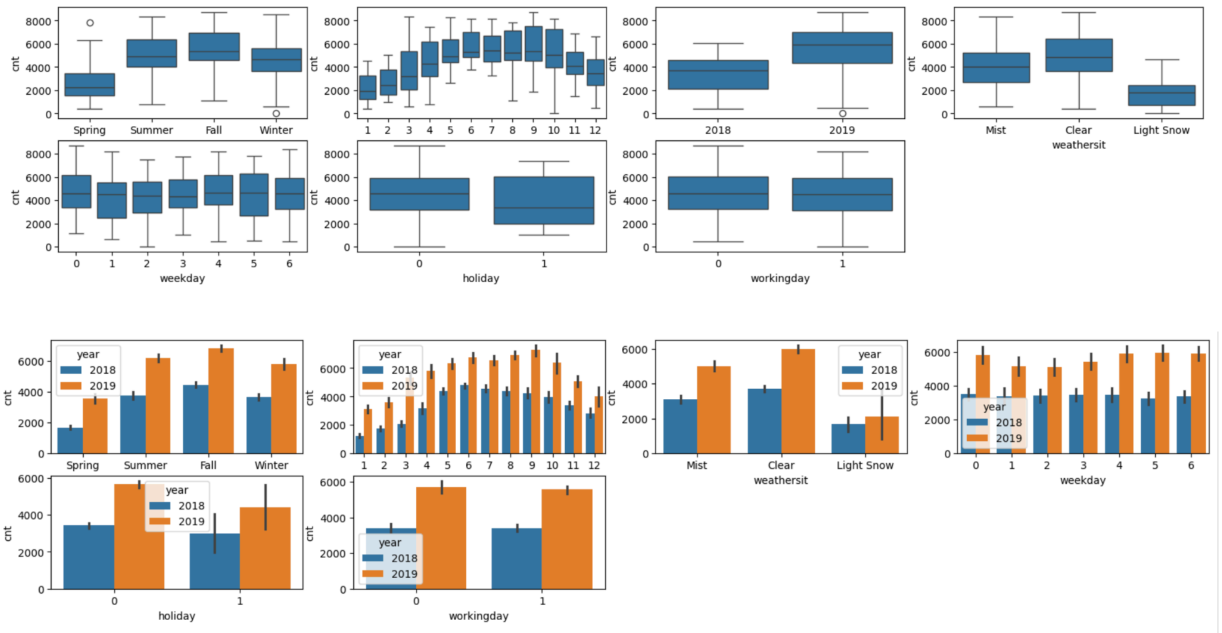
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



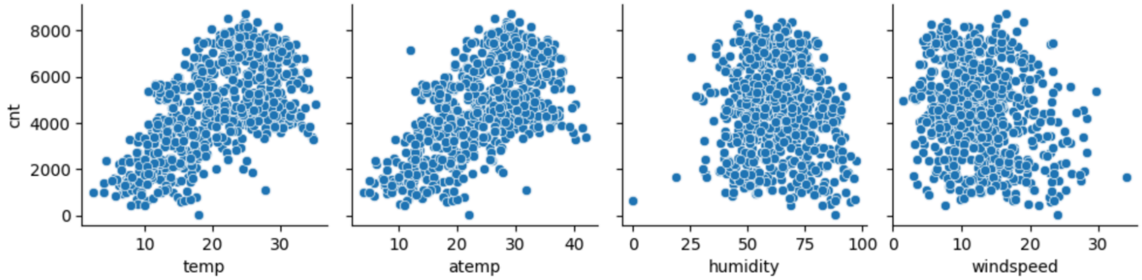
- Weather (Clear) plays vital role in increase of bike rentals
- 2019 is good year compared to 2018 across all categorical variables
- There is not much/negligible difference due to different days
- Months June to Sept has also positive impact on bike rentals

2. Why is it important to use drop_first=True during dummy variable creation?

If we don’t drop first, it will create a multicollinearity issue as all those are highly correlated.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

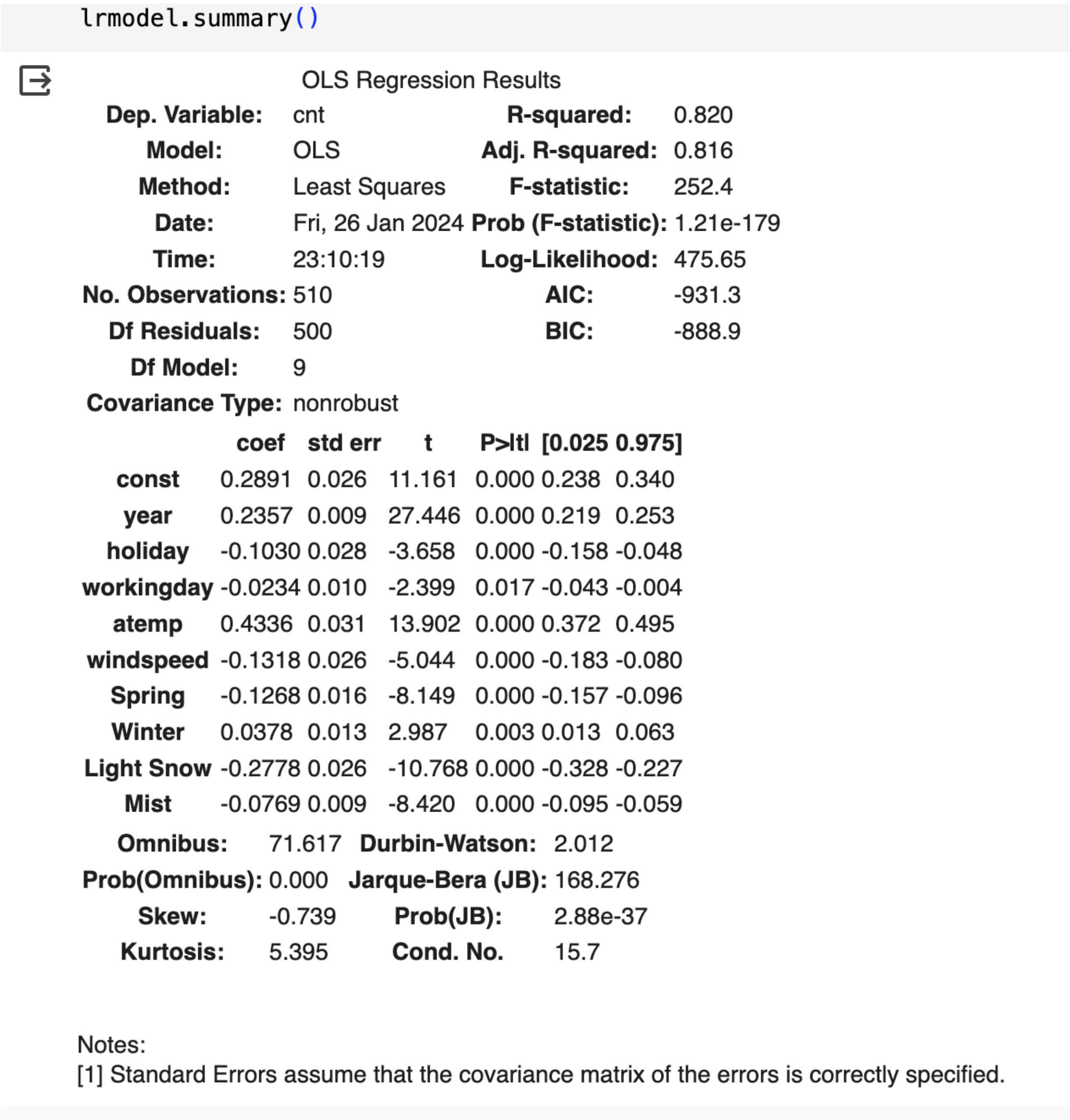
```
<seaborn.axisgrid.PairGrid at 0x7e7e2c067d30>
```



- Temp and atemp variables do have linear relationship with cnt (target variable), whereas other 2 (humidity and wind speed) do not.
- Temp and atemp are multicollinear to each other

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- P-value of all independent variables < 0.05 and R2 score is 82.0

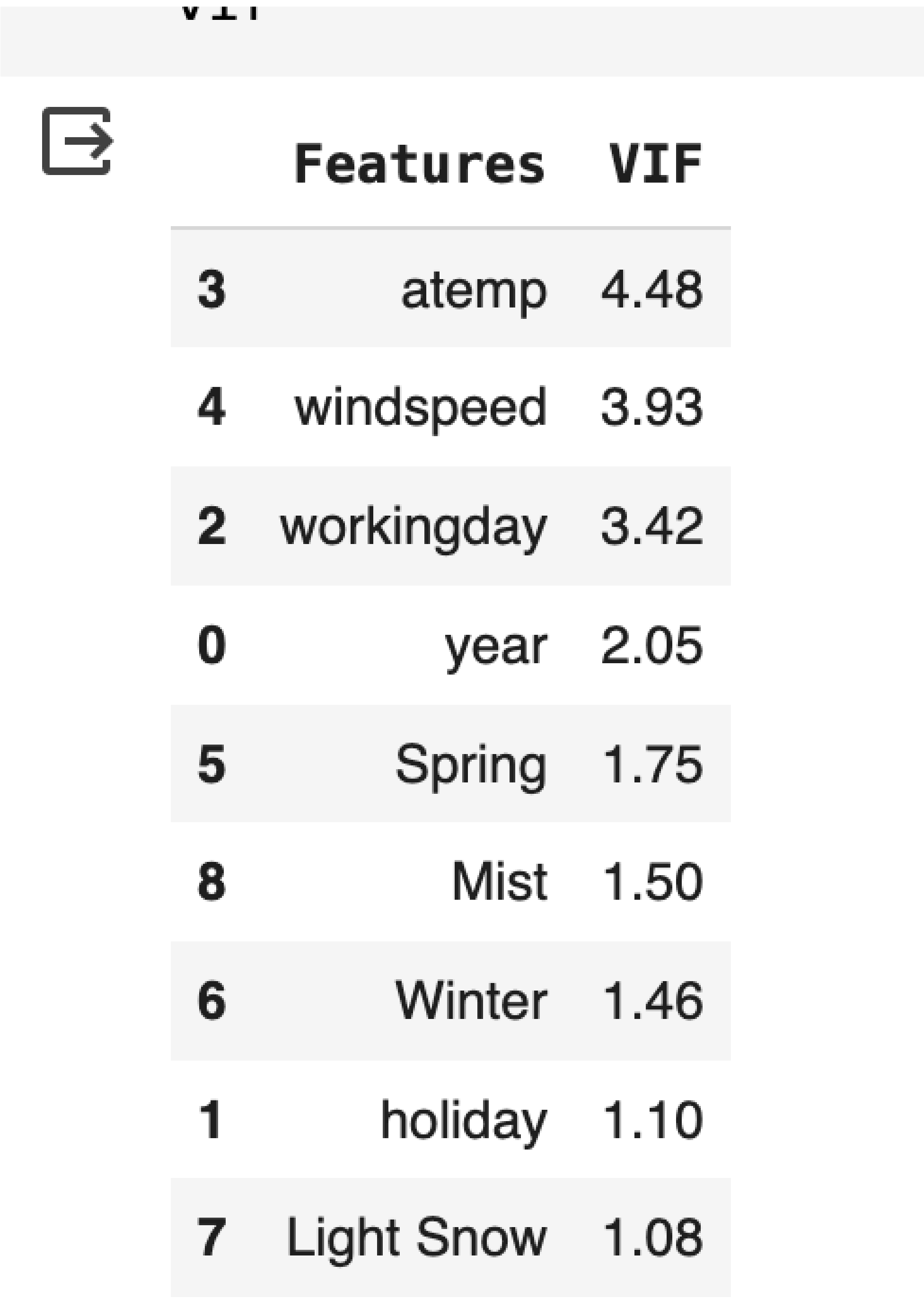


```
lrmodel.summary()
```

```
                        OLS Regression Results
  Dep. Variable:     cnt                   R-squared:    0.820
         Model:      OLS              Adj. R-squared:    0.816
        Method:      Least Squares       F-statistic:    252.4
          Date:      Fri, 26 Jan 2024 Prob (F-statistic): 1.21e-179
          Time:      23:10:19          Log-Likelihood:   475.65
No. Observations:    510                         AIC:    -931.3
   Df Residuals:     500                         BIC:    -888.9
      Df Model:      9
Covariance Type:     nonrobust
               coef   std err    t      P>|t|  [0.025  0.975]
    const     0.2891  0.026   11.161   0.000  0.238   0.340
     year     0.2357  0.009   27.446   0.000  0.219   0.253
   holiday   -0.1030  0.028   -3.658   0.000  -0.158  -0.048
 workingday  -0.0234  0.010   -2.399   0.017  -0.043  -0.004
    atemp     0.4336  0.031   13.902   0.000  0.372   0.495
  windspeed  -0.1318  0.026   -5.044   0.000  -0.183  -0.080
   Spring    -0.1268  0.016   -8.149   0.000  -0.157  -0.096
   Winter     0.0378  0.013   2.987    0.003  0.013   0.063
 Light Snow  -0.2778  0.026   -10.768  0.000  -0.328  -0.227
    Mist     -0.0769  0.009   -8.420   0.000  -0.095  -0.059
     Omnibus:      71.617   Durbin-Watson:     2.012
Prob(Omnibus):     0.000   Jarque-Bera (JB):   168.276
        Skew:      -0.739        Prob(JB):     2.88e-37
    Kurtosis:      5.395        Cond. No.      15.7


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

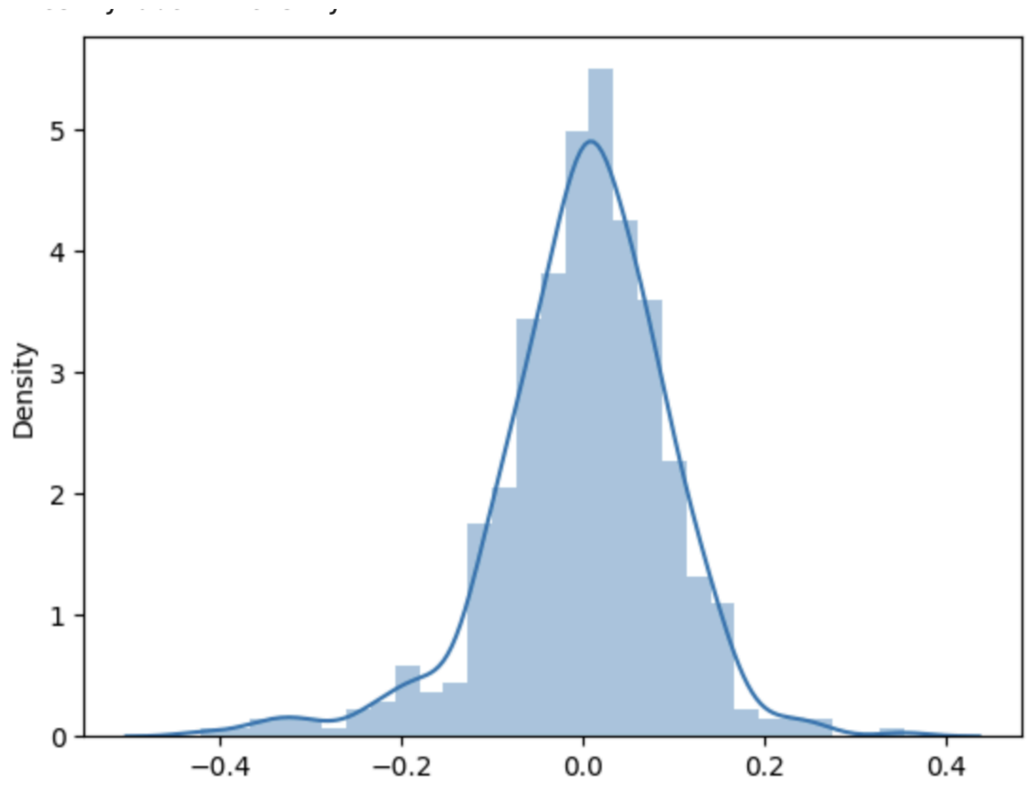
- VIF < 5 for all those independent variables and it implies no problem of multicollinearity



```
vif
```

	Features	VIF
3	atemp	4.48
4	windspeed	3.93
2	workingday	3.42
0	year	2.05
5	Spring	1.75
8	Mist	1.50
6	Winter	1.46
1	holiday	1.10
7	Light Snow	1.08

- Performed residual analysis and residuals (errors) are normally distributed



- Applied the model on test data and got good R2 score **80.6**

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Temp is positively correlated to bike rentals
- Year 2019 has significant contribution compared to 2018
- Light snow and Mist weather are playing negative role in bike rentals

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a supervised machine learning method and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable.

This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.The following is an example of a resulting linear regression equation:

y=b0+b11x1+b2x2

The explanatory variable is also known as the predictor variable, and the dependent variables are also known as the output variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R? (3 marks)

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- When r is between 0 and 1 - When one variable changes, the other variable changes in the same direction.
- When r is 0 - There is no relationship between the variables.
- When r is 0 and -1 - When one variable changes, the other variable changes in the opposite direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling refers to putting the feature values into the same range.A technique to scale data is to squeeze it into a predefined interval.we have 2 types of scaling in ML world, normalization and standardization.

In Normalization, we map the minimum feature value to 0 and the maximum to 1. Hence, the feature values are mapped into the [0, 1] range.

In Standardization, we don't enforce the data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

When R2 score is 1 it is possible, VIF formula = 1/(1-R2) => infinite
This scenario can occur, if strong correlation exists b/n two independent variables (multicollinearity problem).
In our example, temp and atemp.
If we build the model with these two variables we might end up having VIF as infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other

This will help us to find the relationship b/n predictor variables and target variable and also help to choose the algorithm if relationship is linear (LR is the one to choose)