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?

    a. Effect of `weathersit`

    - The highest number of bike rentals happen when there are clear skies (weathersit 1).
    - The lowerst number of bike rentals happen when there is light snow (weathersit 3)

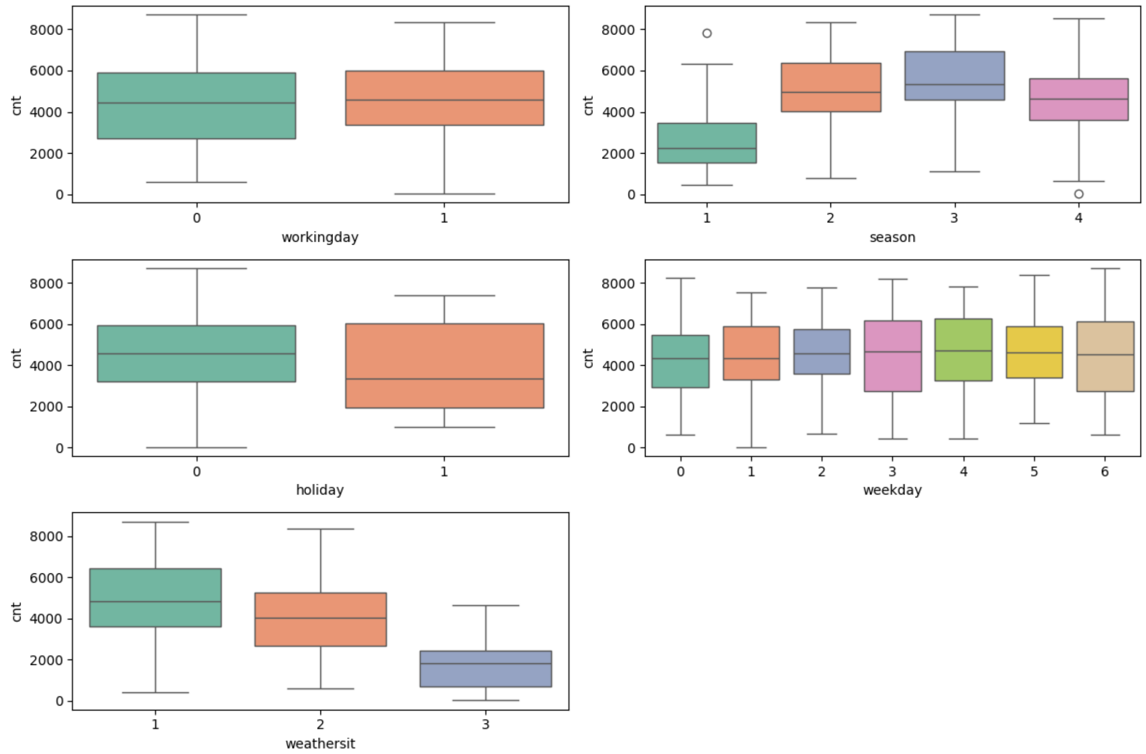
    b. Effect of `season` on rentals

    - The highest number of rentals tend to happen during the fall (season 3)
    - The lowest number of rentals tend to happen during spring (season 1)

    c. Effect of `weekday`

        a. The median number of rentals stays pretty much the same throughout all the days of the week.

        b. Monday (weekday 6) and Friday (weekday 3) show the most spread among all days of the week.



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

   `k` levels of values will generally only require `k-1` levels of encoding.

   For Example: If a categorical variable has 3 possible values, it can be encoded with 2 dummy variables like so

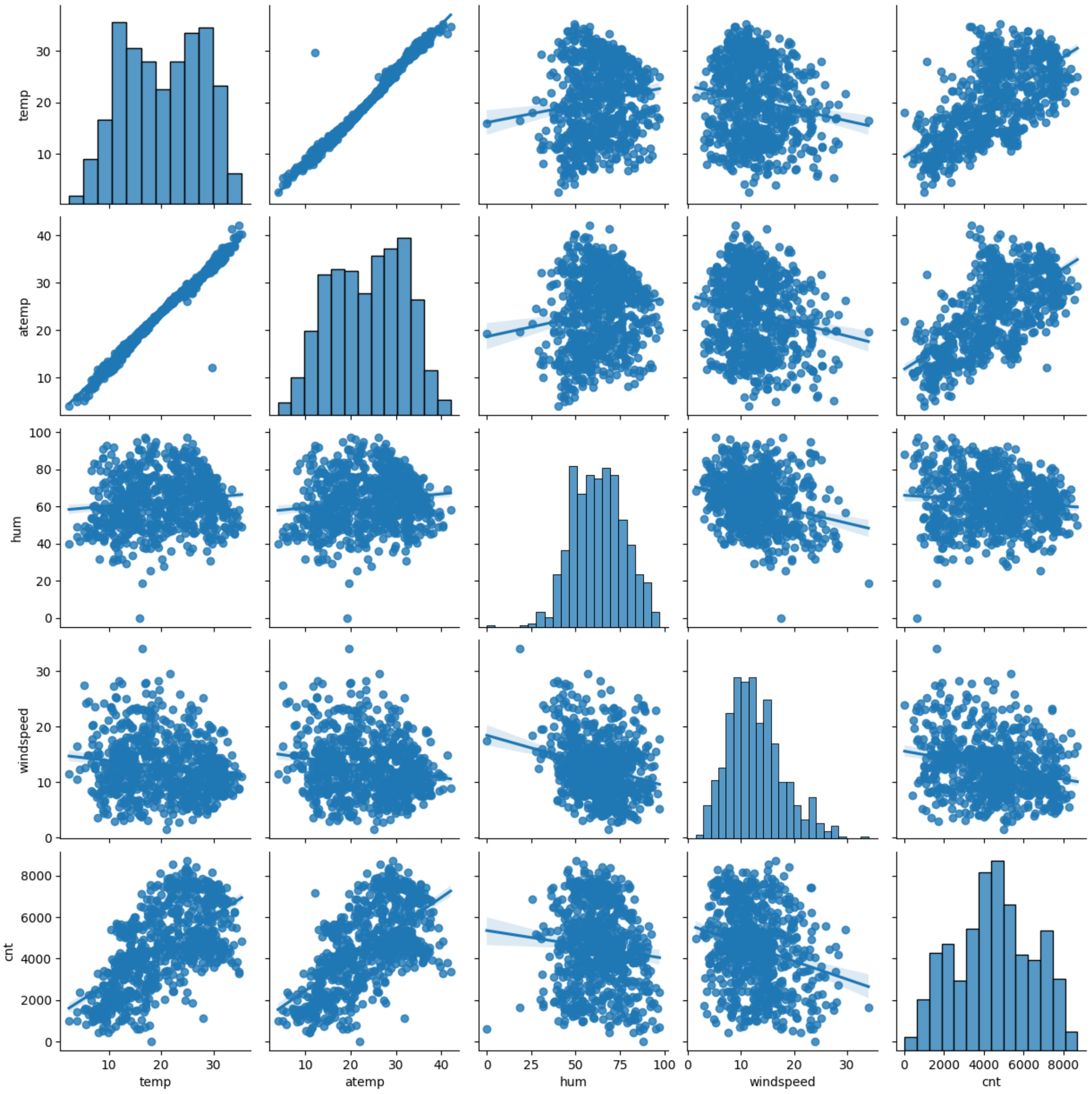
   Category value 1 - `0 1`

   Category value 2 - `1 0`

   Category value 3 - `0 0` (implicit)

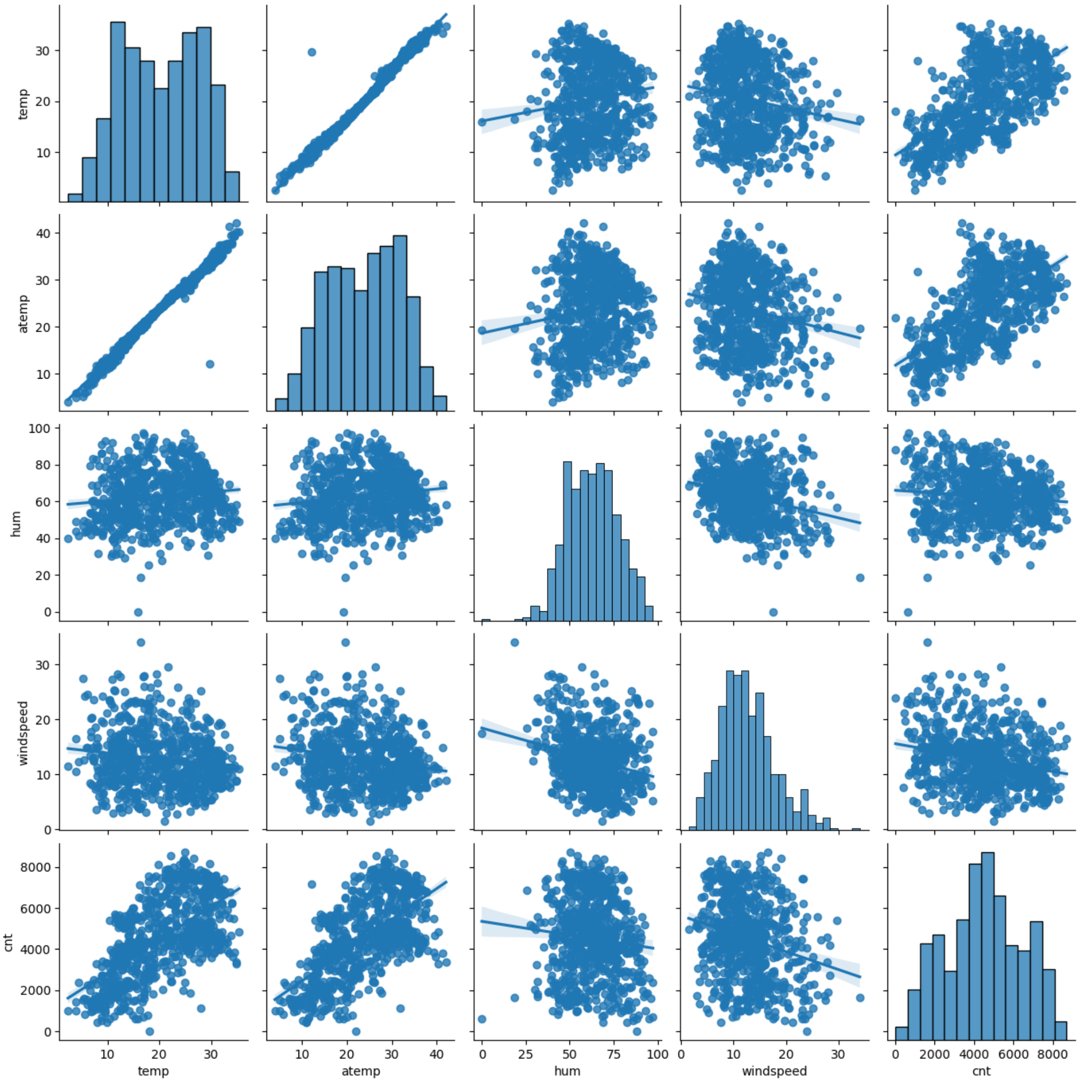
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   From the pairplot, `temp` has the highest correlation with the target variable `cnt`

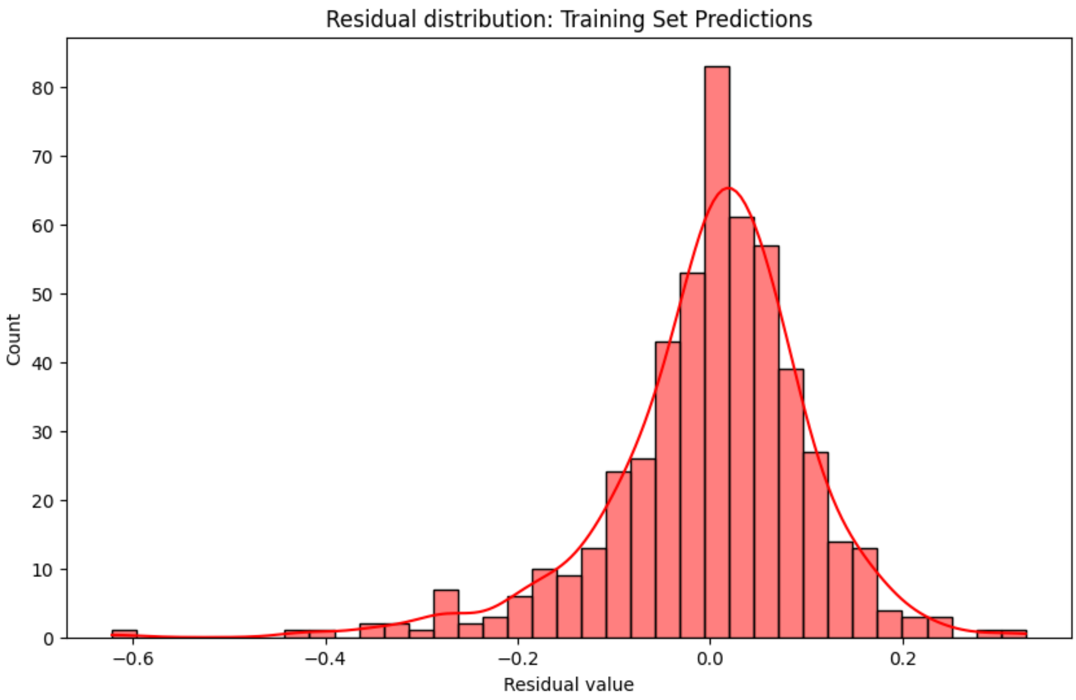


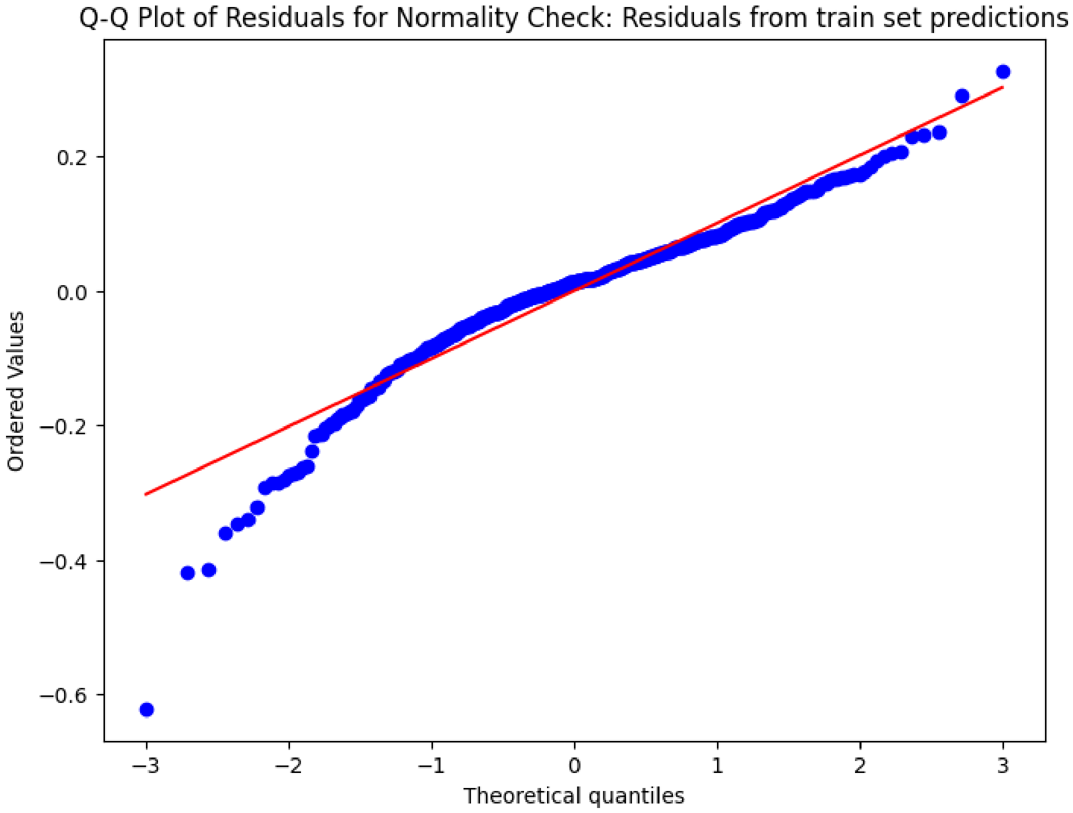
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

    a. **Linear relationship**: There exists a linear relationship between the dependent and the independent variables. This can be verified using the scatter plots from the pairplot which clearly show a linear relationship between `cnt` and the independent variables.

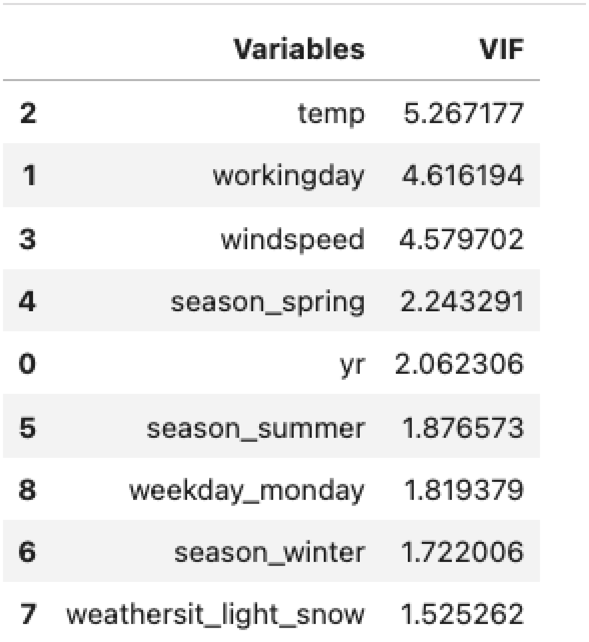


b. **Error terms (residuals) normally distributed**: This has been verified using both the histogram as well as a Q-Q plot.



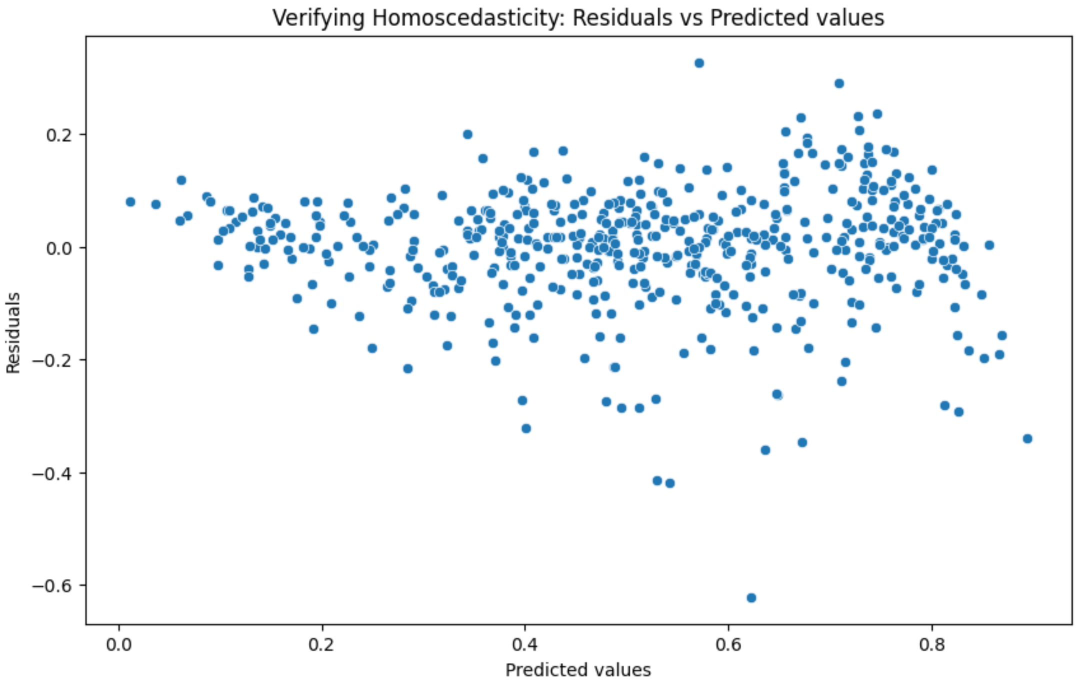


c. **No Multicollinearity**: Predictors which were multicollinear (VIF > 10) have been dropped.



	Variables	VIF
2	temp	5.267177
1	workingday	4.616194
3	windspeed	4.579702
4	season_spring	2.243291
0	yr	2.062306
5	season_summer	1.876573
8	weekday_monday	1.819379
6	season_winter	1.722006
7	weathersit_light_snow	1.525262

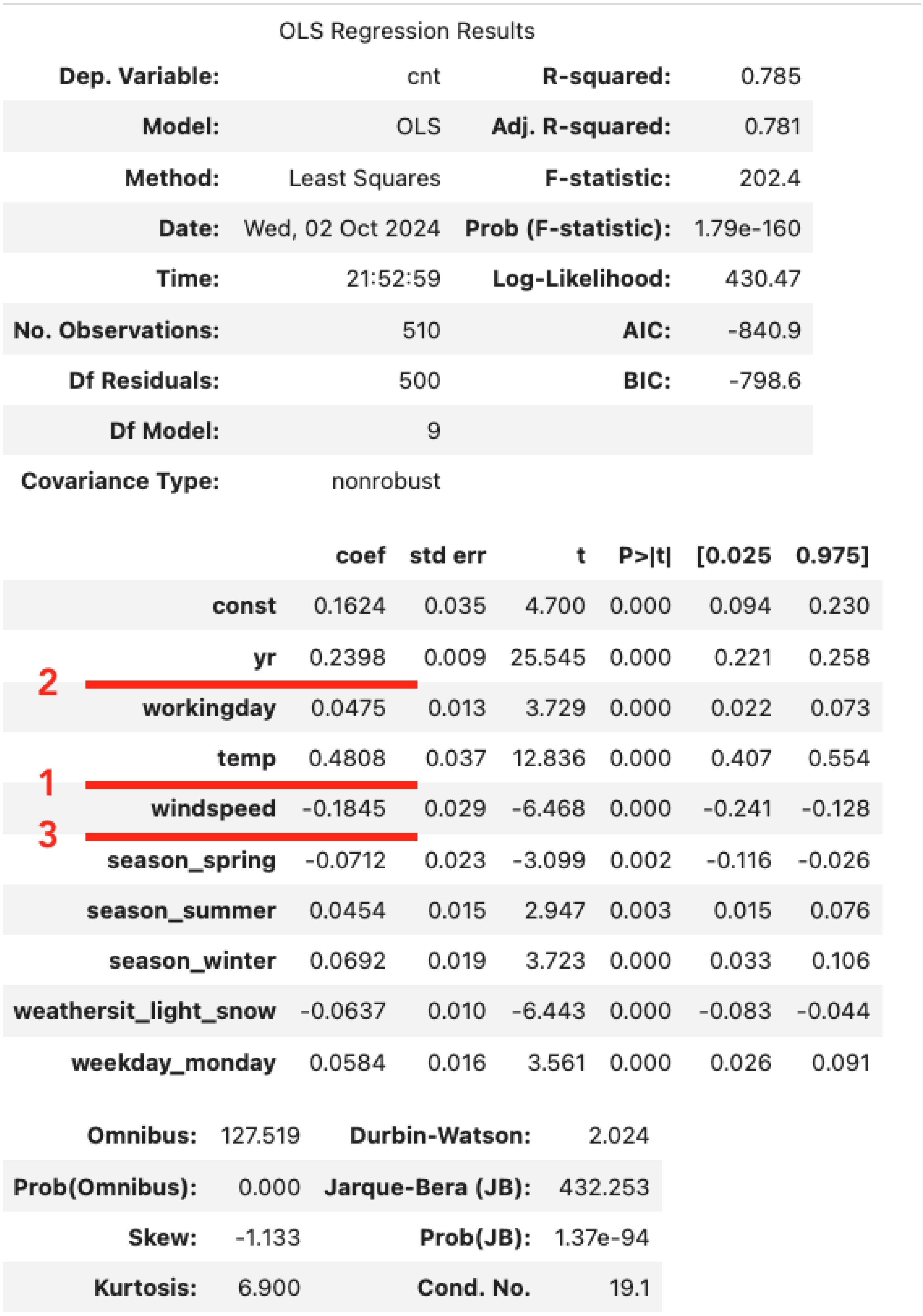
d. **Homoscedasticity**: There should be no discernible pattern when residuals are plotted against the predicted values (i,e) The residuals should have constant variance.



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

The following are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

1. `temp` - strong positive
2. `yr` - string positive
3. `windspeed` - strong negative



OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.785
Model:	OLS	**Adj. R-squared:**	0.781
Method:	Least Squares	**F-statistic:**	202.4
Date:	Wed, 02 Oct 2024	**Prob (F-statistic):**	1.79e-160
Time:	21:52:59	**Log-Likelihood:**	430.47
No. Observations:	510	**AIC:**	-840.9
Df Residuals:	500	**BIC:**	-798.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1624	0.035	4.700	0.000	0.094	0.230
yr	0.2398	0.009	25.545	0.000	0.221	0.258
workingday	0.0475	0.013	3.729	0.000	0.022	0.073
temp	0.4808	0.037	12.836	0.000	0.407	0.554
windspeed	-0.1845	0.029	-6.468	0.000	-0.241	-0.128
season_spring	-0.0712	0.023	-3.099	0.002	-0.116	-0.026
season_summer	0.0454	0.015	2.947	0.003	0.015	0.076
season_winter	0.0692	0.019	3.723	0.000	0.033	0.106
weathersit_light_snow	-0.0637	0.010	-6.443	0.000	-0.083	-0.044
weekday_monday	0.0584	0.016	3.561	0.000	0.026	0.091

Omnibus:	127.519	**Durbin-Watson:**	2.024
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	432.253
Skew:	-1.133	**Prob(JB):**	1.37e-94
Kurtosis:	6.900	**Cond. No.**	19.1

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is the process of fitting a straight line through the data in order to predict the value of a target variable using independent variables.

Depending on whether it's simple linear regression or multiple linear regression, the number of coefficients can either be 2 or more than 2. There are few different ways to calculate the coefficients to arrive at the best fit line. Some of them are

1. Ordinary Least Squares (OLS)
2. Gradient Descent

Assumptions of Linear Regression:

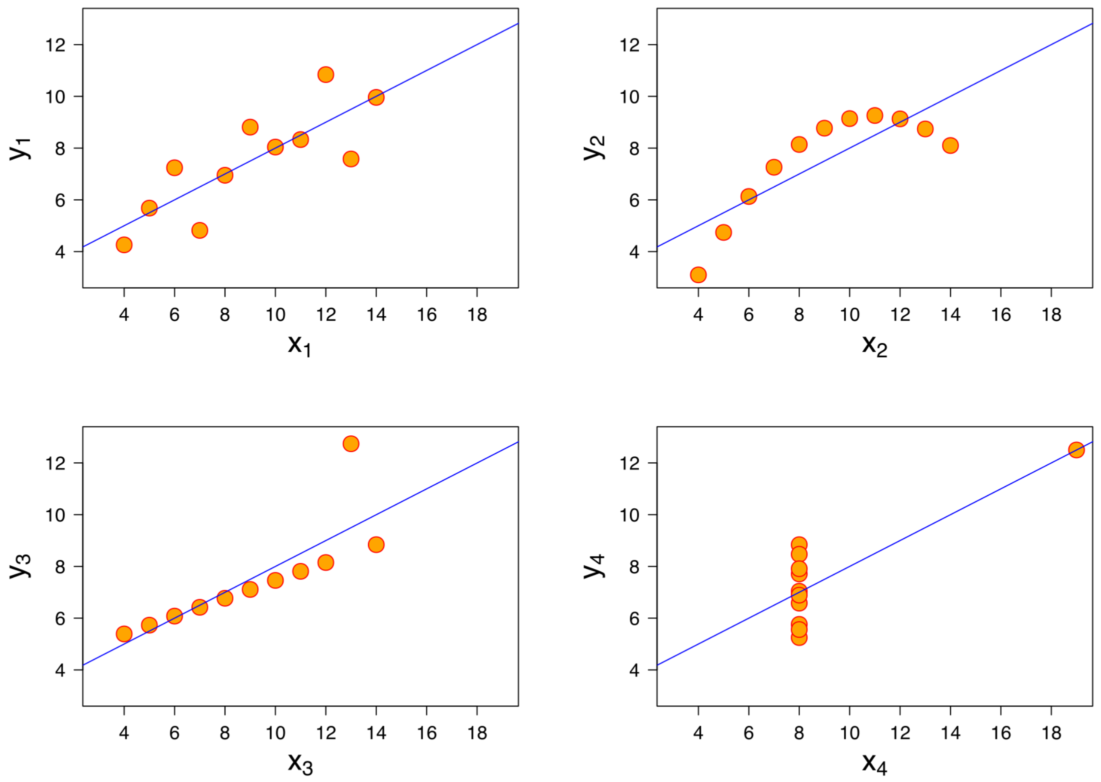
1. **Linearity**: There should be a linear relationship between the independent and target variables.
2. **Homoscedasticity:** The variance of residuals should be constant.
3. **Residuals should be normally distributed**: The residual values (difference between actual and predicted values) should be normally distributed.
4. No Multicollinearity: The indepndent variables should NOT be multicollinear (multicollinearity is when a variable is linearly associated with some other variables).
5. Error terms should be independent of each other.

---

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of 4 datasets that have very similar descriptive statistics but exhibit very different distributions when graphed. The quartet is used to demonstrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistical properties for describing datasets

The mean of all the 4 below datasets is exactly `9` and variance is `11`



3. What is Pearson’s R?

Pearson’s correlation coefficient is a value that measures the strength and direction of linear relationship between 2 set of values. The coefficient is always between -1 (strong negative correlation) and 1 (strong positive correlation). A Pearson correlation coefficient value of `0` indicates that there is no linear relationship between the 2 variables.

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

Scaling is the process of transforming predictor values to a specific range.

Scaling is performed in linear regression because some of the algorithms used to find the coeffiocients are sensitive to feature scales and features having higher scales can distort the model.

In normalized scaling / min-max scaling, the values are rescaled to a fixed range [0,1] whereas in standardized scaling, the values are rescaled in such a way that the mean is 0 and standard deviation is 1.

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

Higher VIF values indicate significant multicollinearity. When the VIF is infinite for a variable, it means that there exists a perfect collinearity where the variable can be linearly derived using other variables.

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

   A Q-Q plot or a quantile - quantile plot can be useful in visualizing values to see if they follow any distribution (such as a normal distribution).

   In Linear regression, it’s important for the residual terms to be normally distributed and hence a Q-Q plot is useful in visually verifying this assumption.