BIKE SHARING ASSIGNMENT

- Sandhya Purushothaman

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?

Answer: The dataset contains 6 categorical variables, which were analyzed using box plots to understand their impact on the dependent variable 'cnt'.

. Working days contrbtute 5 days a week and non-working days 2 days. This shows that even if its a non-working day, the bikes taken on loan is quite a good number, for just 2 days. The number of bikes taken on rent in 2 days, is half the number of bikes taken on rent in the remaining 5 days

. **Working Day :** Working days constitute 5 days a week and non-working days 2 days. This shows that even if its a non-working day, the bikes taken on loan is quite a good number, for just 2 days. The number of bikes taken on rent in 2 days, is half the number of bikes taken on rent in the remaining 5 days

. On Holidays, people generally don't prefer to travel shorter distances using bikes.

. **Weather :** Highest bike-demand is during a clear weather or a weather which is partly cloudy.

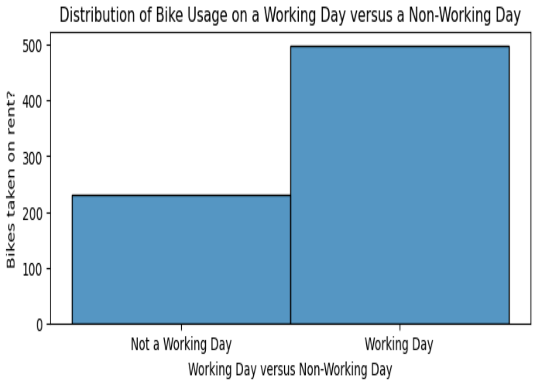
In slight mist, slight demand reduced as compared to when it was clear weather.

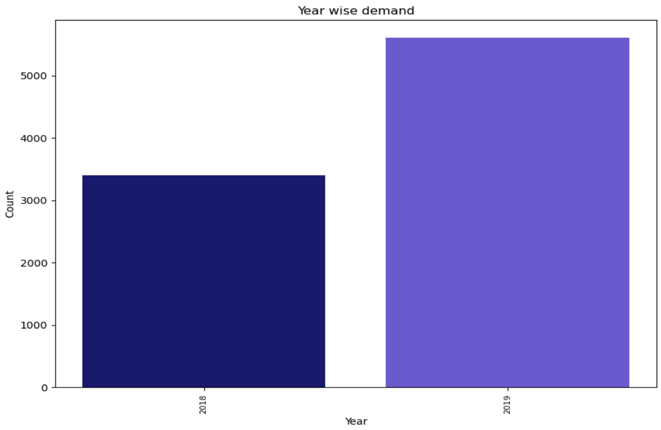
In a weather with partial rains, or partial snowfall, the demand reduced to 1/3rd of when its highest in demand.

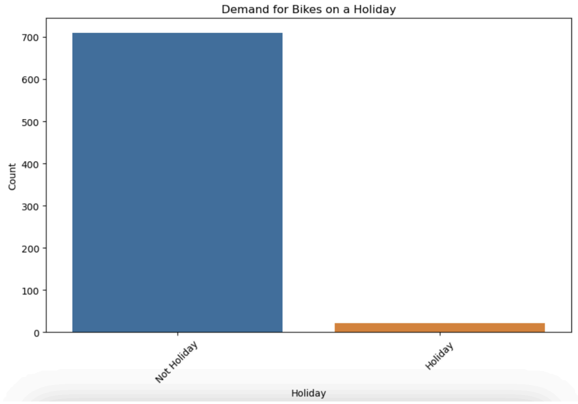
During heavy rain, its quite obvious that nobody rents out the bike, as the data shows.

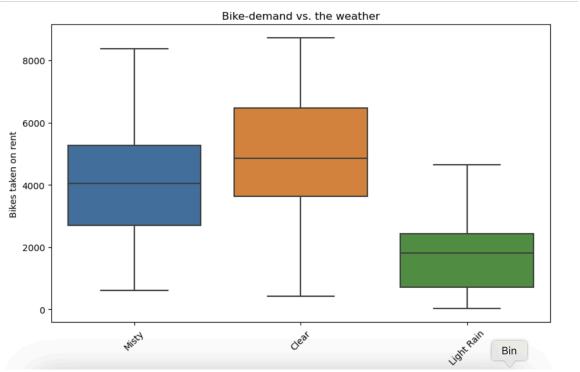
. Year Post Pandemic - Bike sharing has increased in the year post pandemic.

. **Season** No specific useful insights from seasonal inference. In all seasons, there is a good demand, except spring where people are likely to go for longer travels, and use the bikes lesser in number.









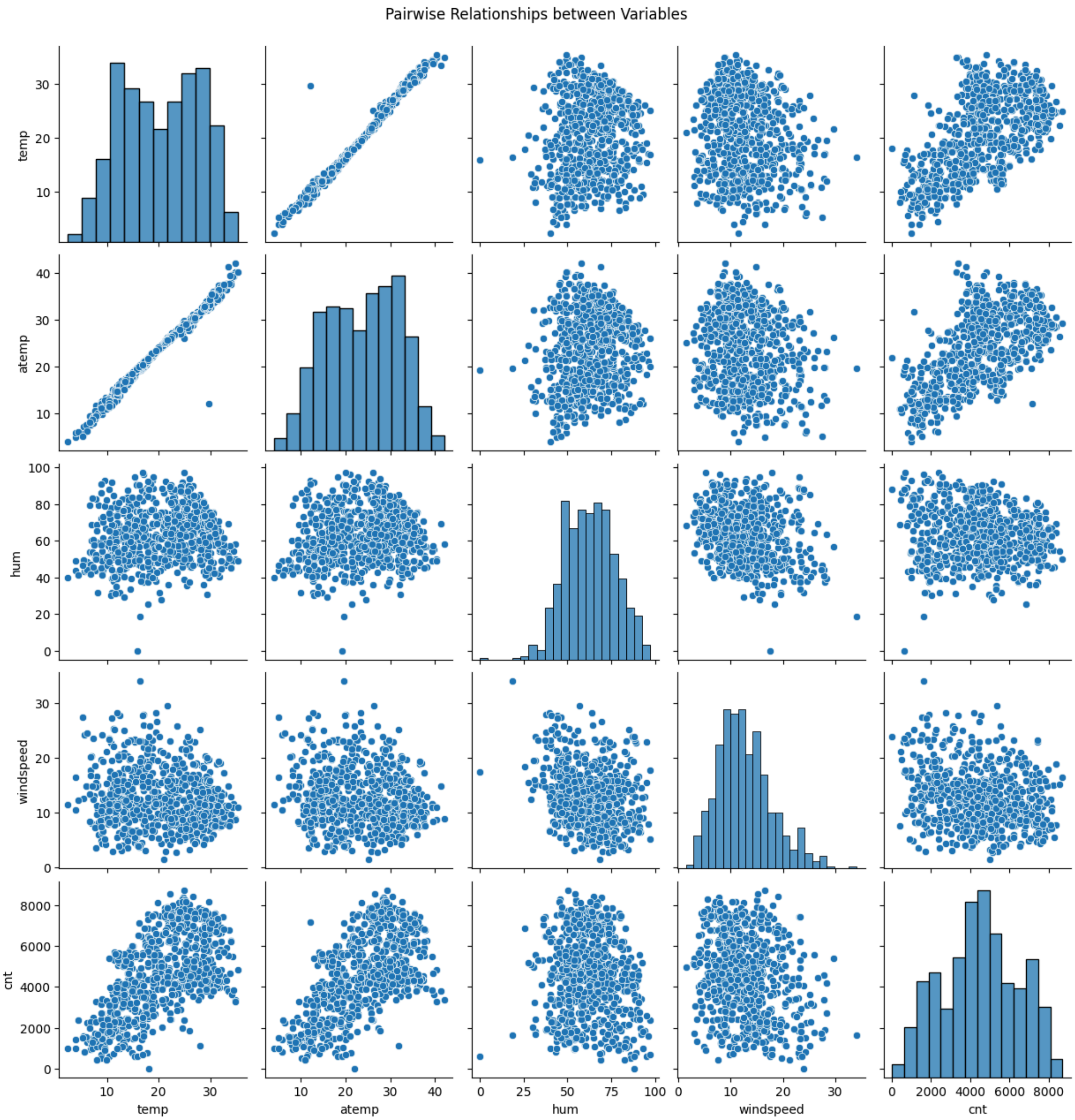
2. Why is it important to use drop_first=True during dummy variable creation?

Answer: It is important to use `drop_first=True` during dummy variable creation **to avoid multicollinearity** issues in the regression model. When creating dummy variables, if you include all levels of a categorical variable, it introduces multicollinearity because one level of the categorical variable can be perfectly predicted from the other levels.

By setting `drop_first=True`, one level of the categorical variable is dropped. This ensures that the model coefficients are properly estimated. Additionally, it **reduces the number of features** in the dataset to **n-1**, making the model efficient.

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

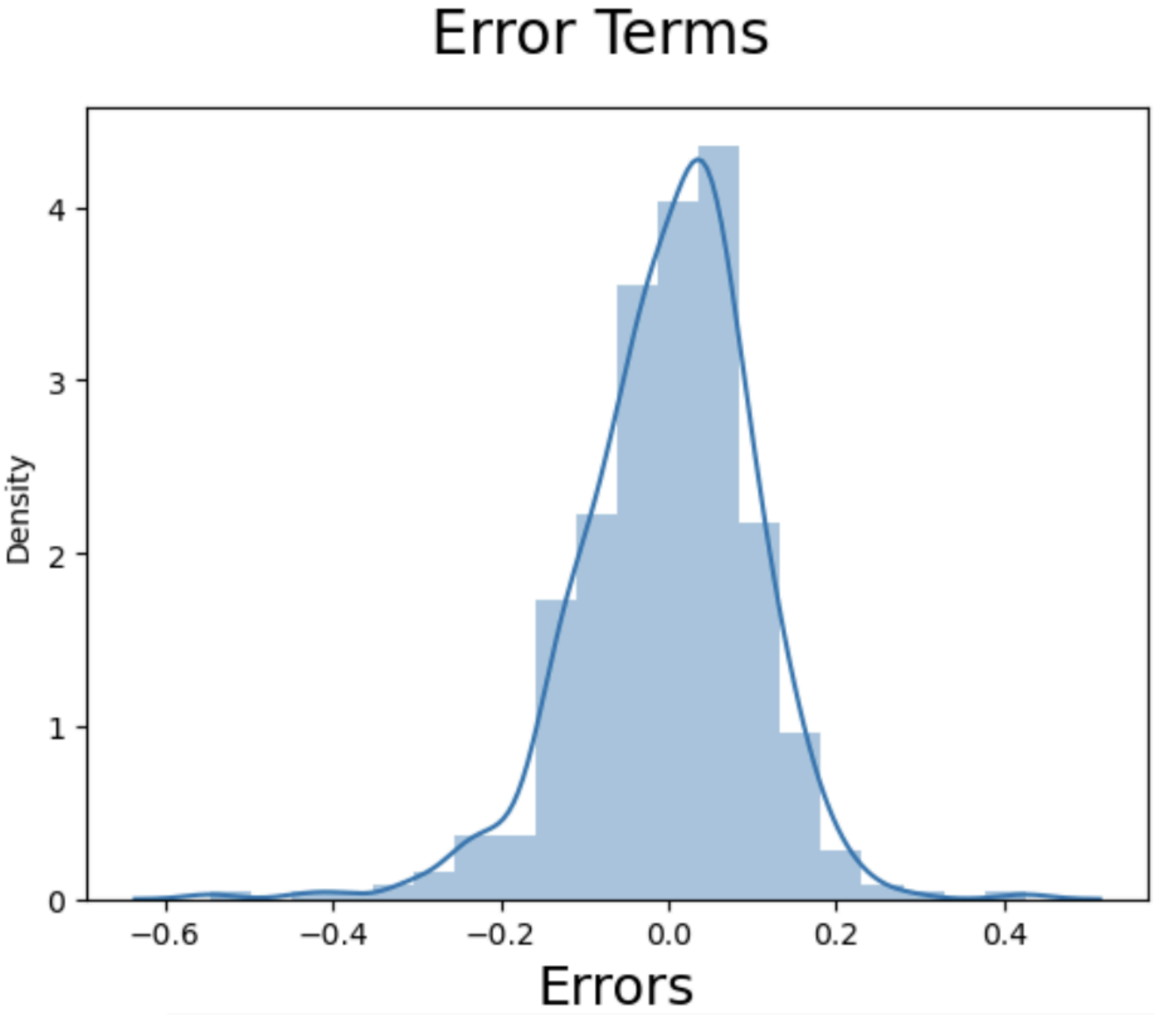
Answer: The variables **'temp' and 'atemp'** have the strongest correlation with the target variable 'cnt' compared to the other variables.

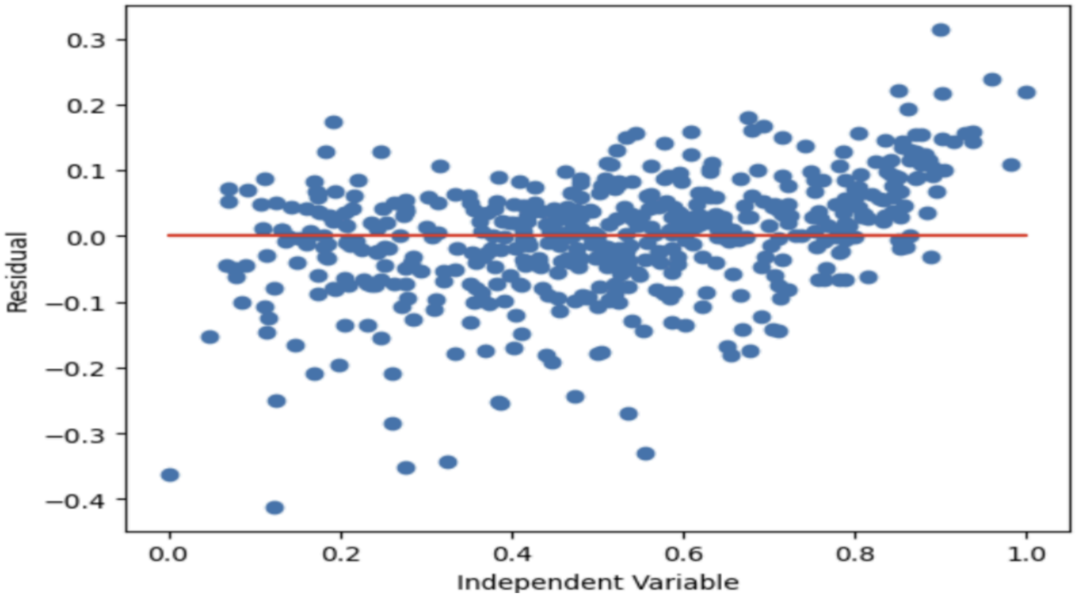

Pairwise Relationships between Variables

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

- **Assumption of Normality of Error Terms**: Error terms are expected to follow a normal distribution with mean at 0, which they are following.
- **Multicollinearity Assessment**: There should be no significant multicollinearity observed among the predictor variables.

- **Validation of Linear Relationship**: The relationship between the predictor variables and the target variable should exhibit linearity.
- **Verification of Homoscedasticity**: Residual values should not display any visible pattern, indicating homoscedasticity.
- **Independence of Residuals**: Residuals should be independent of each other, implying no auto-correlation.

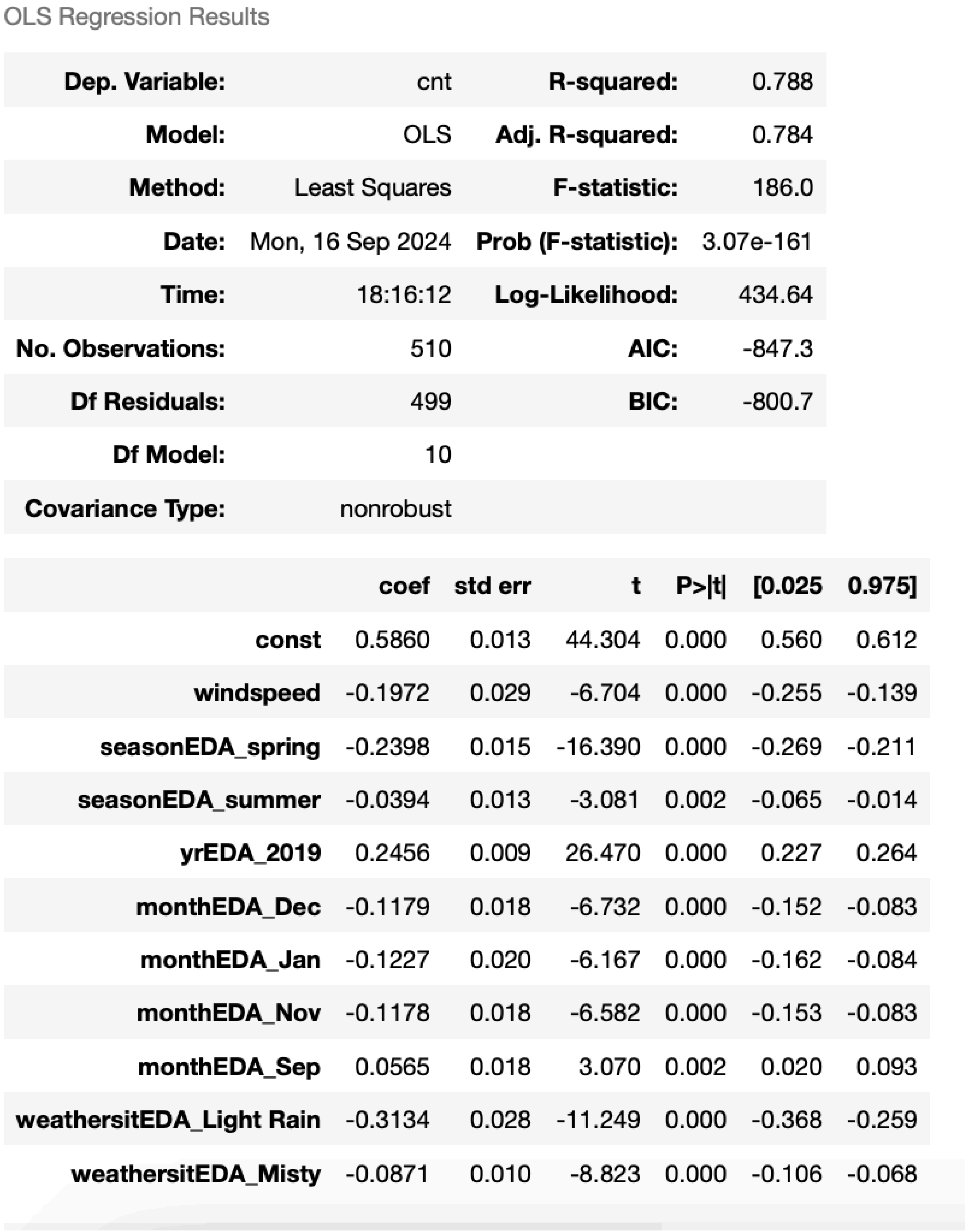




5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: The top 3 features contributing significantly towards explaining the demand of the shared bikes are

- **the year : yr_EDA_2019**
- **the month if its September : monthEDA_Sep**
- **and if there is NO weathersitEDA_Light Rain**



OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.788
Model:	OLS	**Adj. R-squared:**	0.784
Method:	Least Squares	**F-statistic:**	186.0
Date:	Mon, 16 Sep 2024	**Prob (F-statistic):**	3.07e-161
Time:	18:16:12	**Log-Likelihood:**	434.64
No. Observations:	510	**AIC:**	-847.3
Df Residuals:	499	**BIC:**	-800.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.5860	0.013	44.304	0.000	0.560	0.612
windspeed	-0.1972	0.029	-6.704	0.000	-0.255	-0.139
seasonEDA_spring	-0.2398	0.015	-16.390	0.000	-0.269	-0.211
seasonEDA_summer	-0.0394	0.013	-3.081	0.002	-0.065	-0.014
yrEDA_2019	0.2456	0.009	26.470	0.000	0.227	0.264
monthEDA_Dec	-0.1179	0.018	-6.732	0.000	-0.152	-0.083
monthEDA_Jan	-0.1227	0.020	-6.167	0.000	-0.162	-0.084
monthEDA_Nov	-0.1178	0.018	-6.582	0.000	-0.153	-0.083
monthEDA_Sep	0.0565	0.018	3.070	0.002	0.020	0.093
weathersitEDA_Light Rain	-0.3134	0.028	-11.249	0.000	-0.368	-0.259
weathersitEDA_Misty	-0.0871	0.010	-8.823	0.000	-0.106	-0.068

It is evident from the regression equation obtained for the model:

cnt = 0.5860

+ 0.2456 * yr_EDA_2019

+ 0.0565 * monthEDA_Sep

- 0.1972 * windspeed

- 0.2398 * seasonEDA_spring

- 0.0394 * seasonEDA_summer

- 0.1179 * monthEDA_Dec

- 0.1227 * monthEDA_Jan

- 0.1178 * monthEDA_Nov

- 0.3134 * weathersitEDA_Light Rain

- 0.0871 * weathersitEDA_Misty