BIKE SHARING ASSIGNMENT

SUBMISSION

Name: Kaliraj Balakrishnan

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?
2. Why is it important to use drop_first=True during dummy variable creation?
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

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

From data dictionary, I have found 4 Categorical variables that are ***'season' , 'mnth', 'weekday', 'weathersit'***

I have mapped those variables with below dictionary.

season_dict = {1:'spring', 2:'summer', 3:'fall',4:'winter'}

month_dict = {1:'Jan', 2:'Feb', 3:'Mar', 4:'Apr', 5:'May', 6:'Jun', 7:'Jul', 8:'Aug', 9:'Sep', 10:'Oct', 11:'Nov', 12:'Dec'}

week_of_day_dict = {0:'Sunday', 1:'Monday', 2:'Tuesday', 3:'Wednesday', 4:'Thursday', 5:'Friday', 6:'Saturday'}

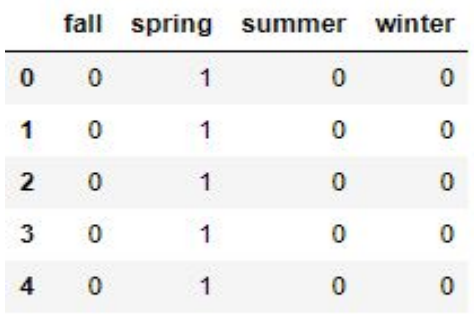
weathersit_dict = {1:'Clear', 2:'Mist + Cloudy', 3:'Light Snow', 4: 'Heavy Rain'}

Independent Categorical variables are nominal, and dependent variables are ordinal. I want to check, there is any effect of independent Categorical variables on each dependent variables.

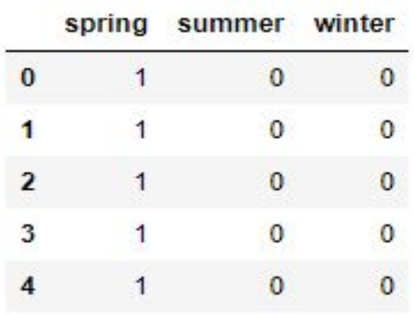
2. Why is it important to use drop_first=True during dummy variable creation?

- **drop_first=True** is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

- Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished. example for season of fall. spring, winter, summer



	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

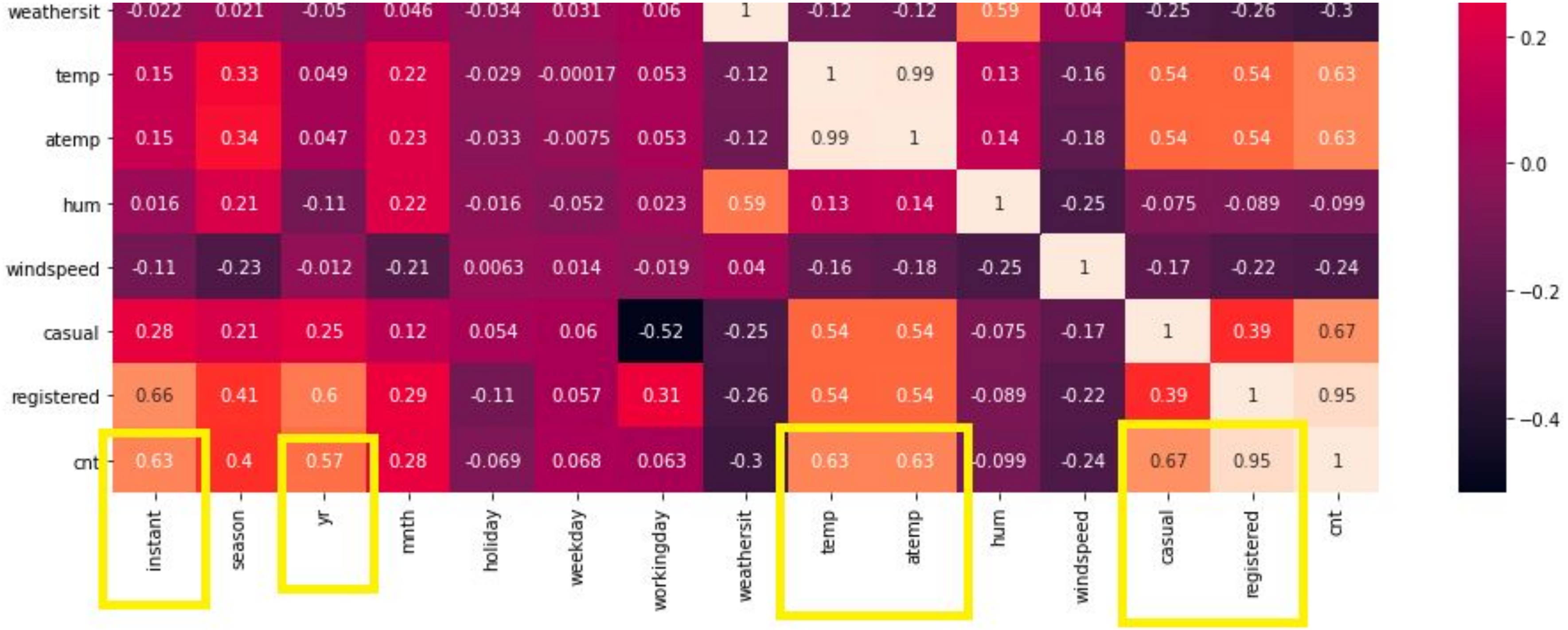


	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

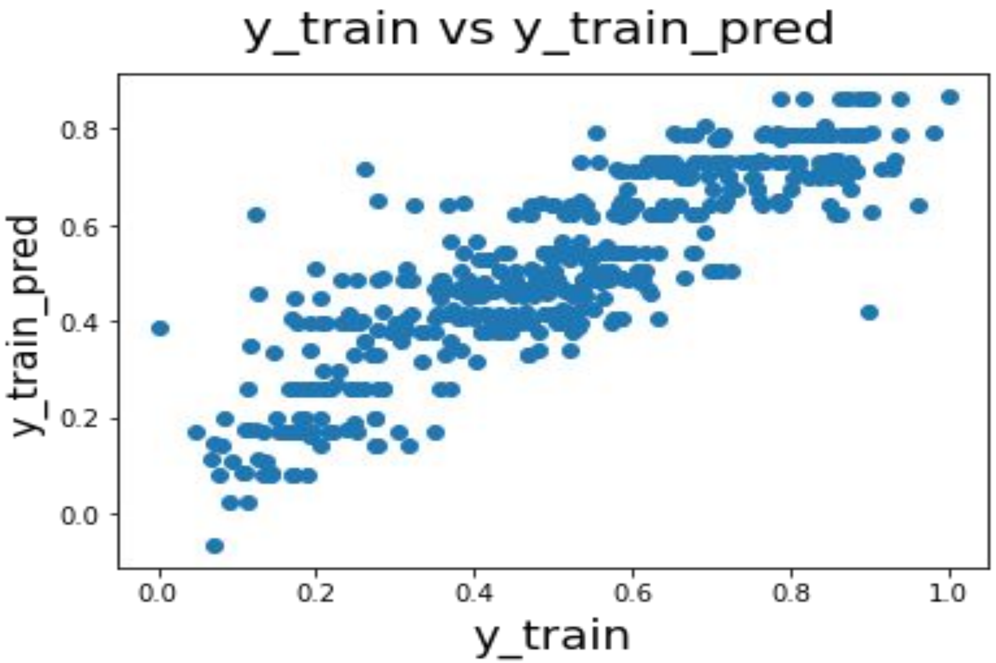
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

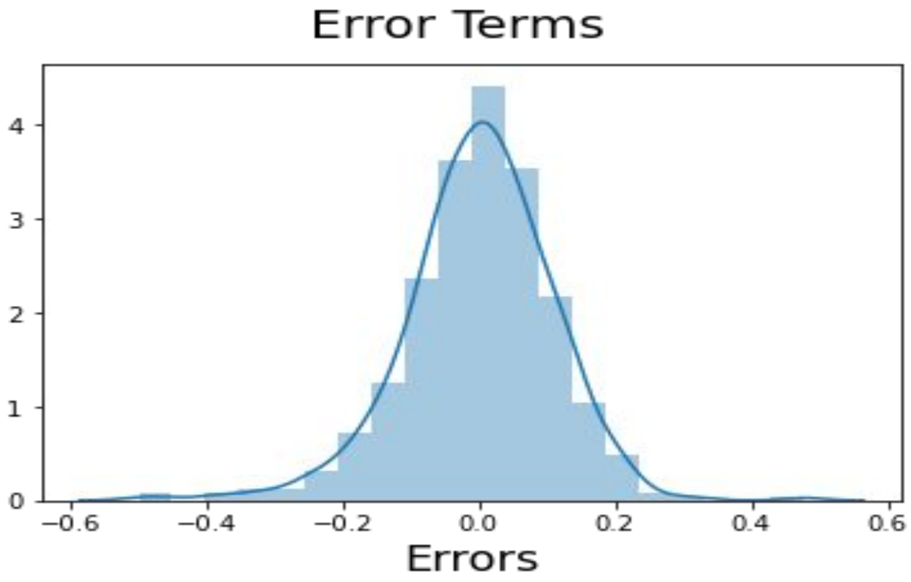
1) Target Variable is ***‘cnt’***
2) Positive correlation variables are ***instant, yr, temp, atemp, causal, registered.*** But, ***instant, causal, registered*** are unused. So, we have ***yr, temp, atemp*** variables. So, in that list ***atemp*** is Highest correlation variable with target variable.



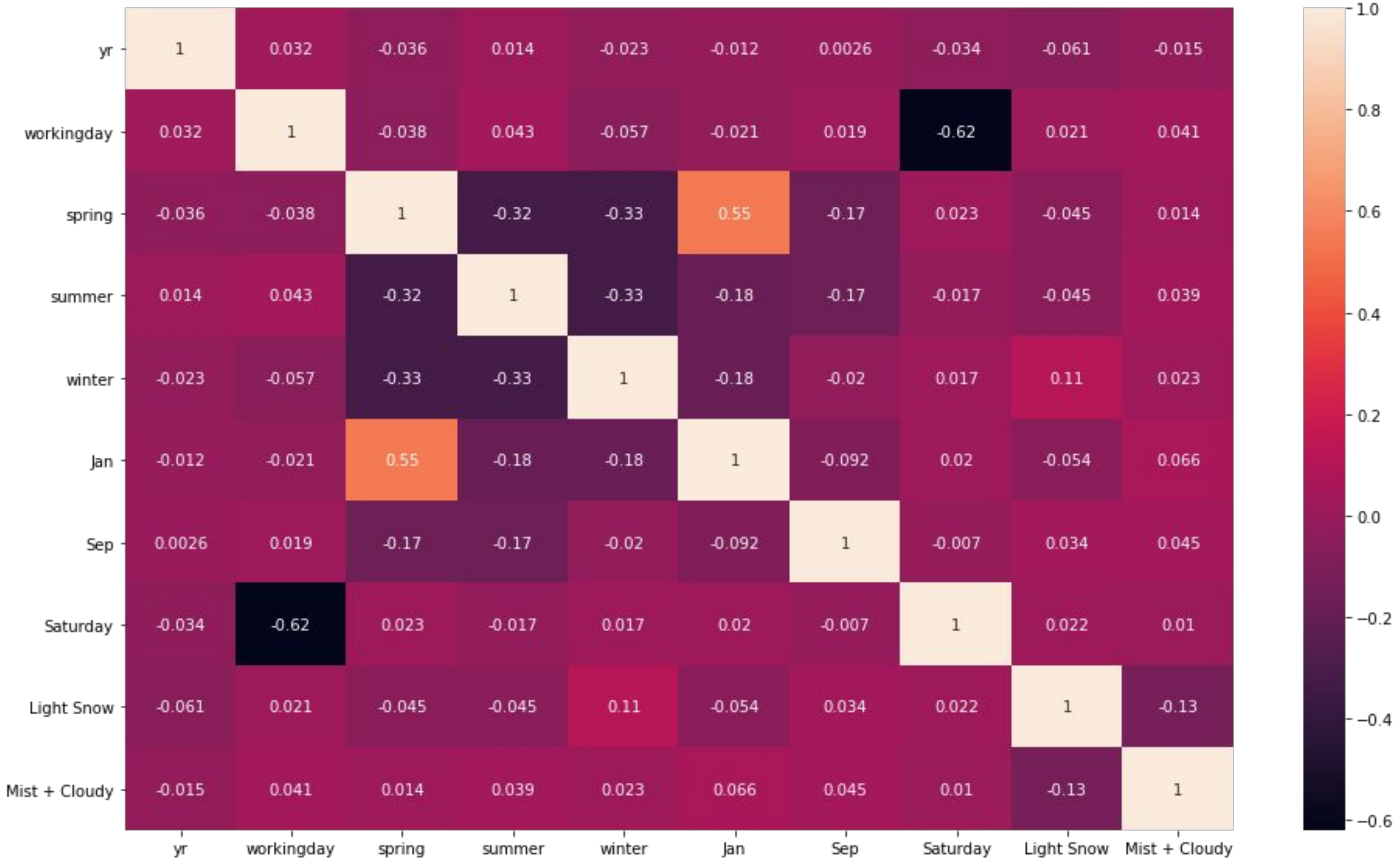
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

1) **Residual Analysis of the train data**
2) **OLS Assumption 2:** The error term has a population mean of zero. The error term accounts for the variation in the dependent variable that the independent variables do not explain. Random chance should determine the values of the error term. For your model to be unbiased, the average value of the error term must equal zero.





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



	yr	workingday	spring	summer	winter	Jan	Sep	Saturday	Light Snow	Mist + Cloudy
yr	1	0.032	-0.036	0.014	-0.023	-0.012	0.0026	-0.034	-0.061	-0.015
workingday	0.032	1	-0.038	0.043	-0.057	-0.021	0.019	-0.62	0.021	0.041
spring	-0.036	-0.038	1	-0.32	-0.33	0.55	-0.17	0.023	-0.045	0.014
summer	0.014	0.043	-0.32	1	-0.33	-0.18	-0.17	-0.017	-0.045	0.039
winter	-0.023	-0.057	-0.33	-0.33	1	-0.18	-0.02	0.017	0.11	0.023
Jan	-0.012	-0.021	0.55	-0.18	-0.18	1	-0.092	0.02	-0.054	0.066
Sep	0.0026	0.019	-0.17	-0.17	-0.02	-0.092	1	-0.007	0.034	0.045
Saturday	-0.034	-0.62	0.023	-0.017	0.017	0.02	-0.007	1	0.022	0.01
Light Snow	-0.061	0.021	-0.045	-0.045	0.11	-0.054	0.034	0.022	1	-0.13
Mist + Cloudy	-0.015	0.041	0.014	0.039	0.023	0.066	0.045	0.01	-0.13	1

As per final model results, **January** Month with **Spring** is High significant for Shared bikes.
In addition to that, Top 3 features as below:

1) **January Month**
2) **Spring Season**
3) **Winter season & Light Snow(weathersit)**

General Subjective Questions

1. Explain the linear regression algorithm in detail.
2. Explain the Anscombe’s quartet in detail.
3. What is Pearson’s R?
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

1. Explain the linear regression algorithm in detail.

Linear regression is a way of find a line which is less distant from each point and the obtained line equation is considered to give the representative value of each dependent variable.

Algorithm:

```
1. mean_x = mean(X)
2. mean_y = mean(Y)
3. n = len(X)

   ##For All values of X & Y calculate

4. Slope = sum (X[i] - mean_x) * (Y[i] - mean_y) )/ sum((X[i] - mean_x) ** 2)
5. Constant = mean_y - (m * mean_x)
```

2. Explain the Anscombe's quartet in detail.

- **Anscombe's** quartet is a representation explaining 4 different set of data having same Mean, Median and Standard Deviation in closest range. The 4 datasets may look alike in terms of final metric measures. But they are spatially different when visualized.

- Anscombe quartet emphasises the importance of visualization needed before making a decision about the data

3. What is Pearson's R?

Pearson R correlation is a **bi-variate** analysis measure. This gives the linear relation between any 2 variables. the values could range from -1 to 1. both the extreme values indicate +ve or -ve correspondence but the with a strong relation.

the value close to 0 indicates no correlation or least significance.

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

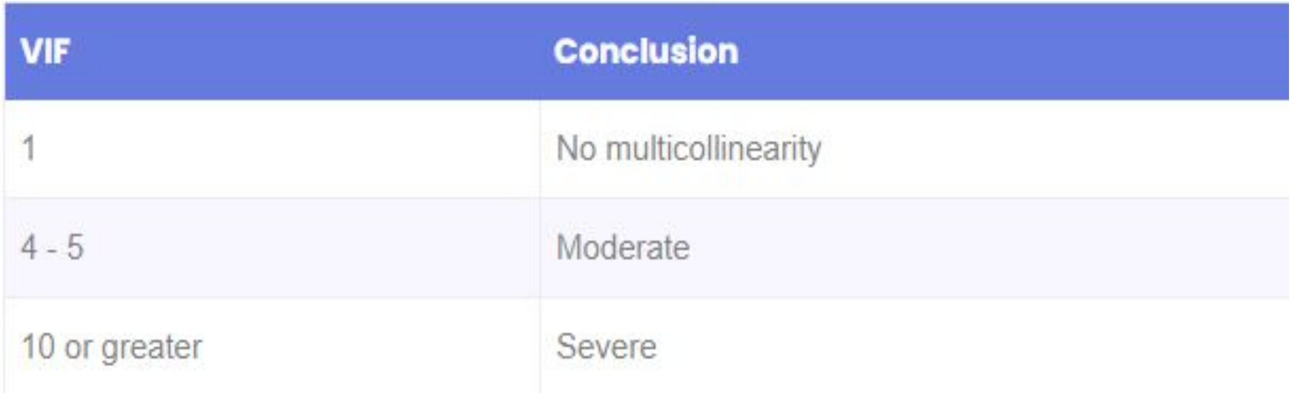
Scaling is a process in data engineering pipeline in which we try to bring all the data under same measurable unit. The most commonly used and sensible form of measurable unit is chosen based on the dataset under study

1. **Normalized scaling** rings down each datapoint with in same range (example 0 to 1, Activation function in neural networks is a best example)
2. **Standardized scaling** refers to creating a cluster of points by within certain radius from mean of all values. the grouping of nearest neighbors is an example.

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.



VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

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

QQ plot is a verification method to check the skewness in the distribution. The data points are plotted on a 2D plane. if the data points align straight with a line that is 45 degrees with X-axis, then the corresponding plot is normally distributed.

The change in orientation or deviation from the line helps measure the skewness of the data points. This graph is a simple comparison to understand between the expected normal distribution vs the existing distribution.