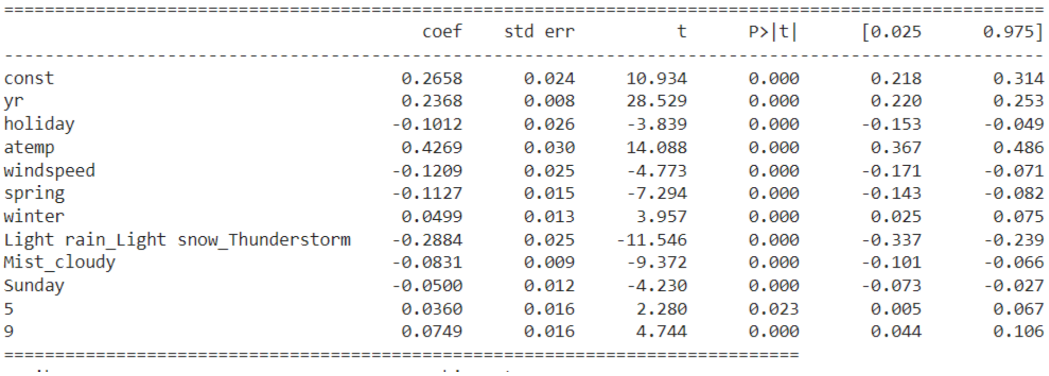
Assignment-based Subjective Questions

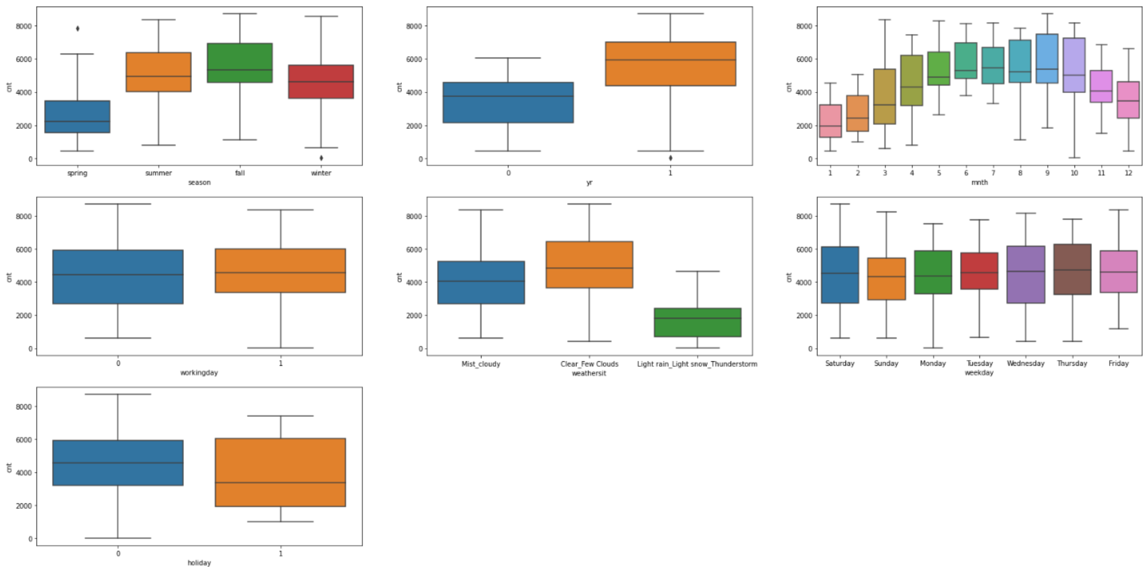
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



```
==============================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                                0.2658      0.024     10.934      0.000       0.218       0.314
yr                                   0.2368      0.008     28.529      0.000       0.220       0.253
holiday                             -0.1012      0.026     -3.839      0.000      -0.153      -0.049
atemp                                0.4269      0.030     14.088      0.000       0.367       0.486
windspeed                           -0.1209      0.025     -4.773      0.000      -0.171      -0.071
spring                              -0.1127      0.015     -7.294      0.000      -0.143      -0.082
winter                               0.0499      0.013      3.957      0.000       0.025       0.075
Light rain_Light snow_Thunderstorm  -0.2884      0.025    -11.546      0.000      -0.337      -0.239
Mist_cloudy                         -0.0831      0.009     -9.372      0.000      -0.101      -0.066
Sunday                              -0.0500      0.012     -4.230      0.000      -0.073      -0.027
5                                    0.0360      0.016      2.280      0.023       0.005       0.067
9                                    0.0749      0.016      4.744      0.000       0.044       0.106
==============================================================================
```

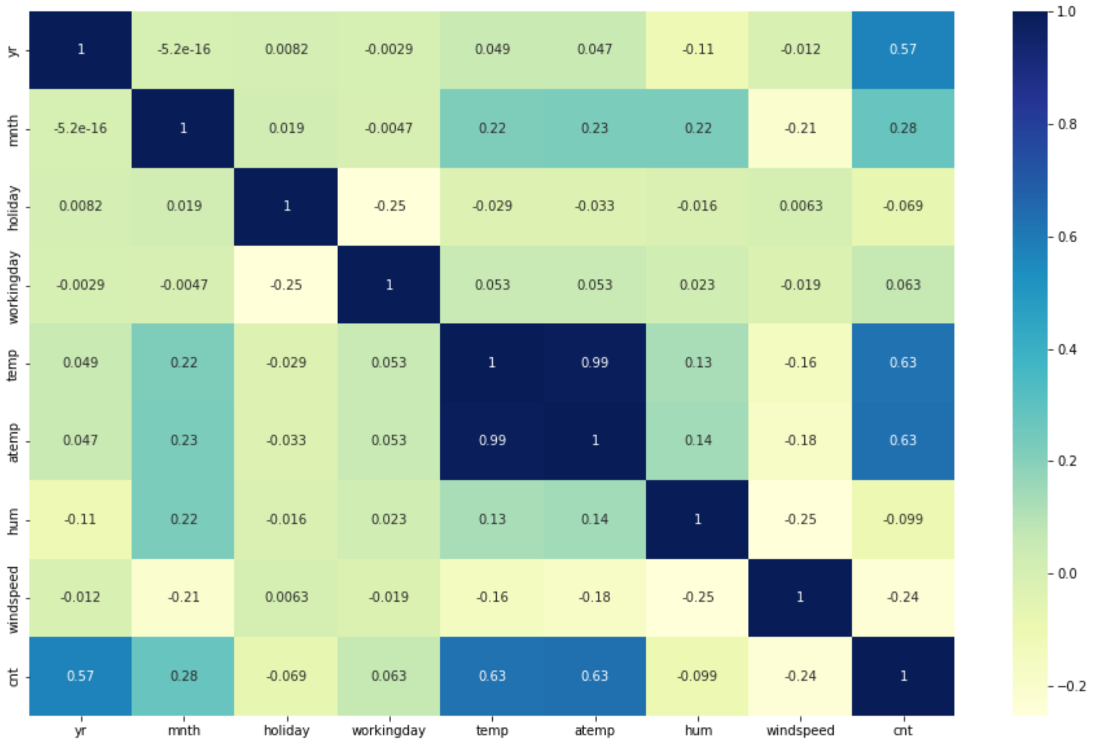
A.

Above Categorical variables from the best fit model that is obtained, are statistically significant as they have p value less than 0.05.



From the above analysis, 'Season' and 'mnth' columns categories have high variation in median. From the best fit model, we can see that some of the categories are removed as they are not significant.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   A. Due to dummy variable trap, which may lead to multicollinearity. Multicollinearity will violate the assumptions of Linear Regression.
   Also it reduces the number of unnecessary features from training the model.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   (1 mark)

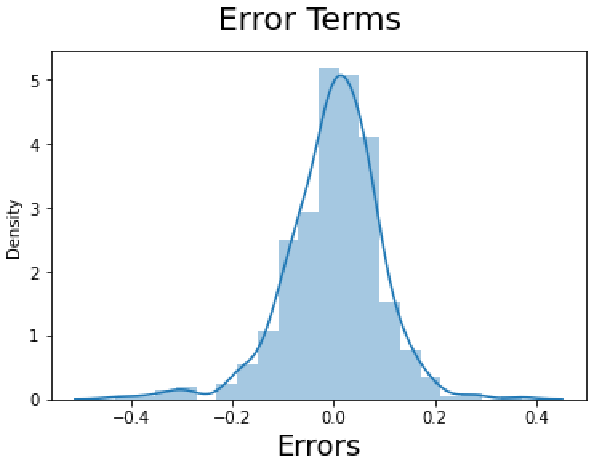


A.

'atemp' and 'temp' features are highly correlated with Target 'cnt' Variable with '0.63' and '0.63' respectively.

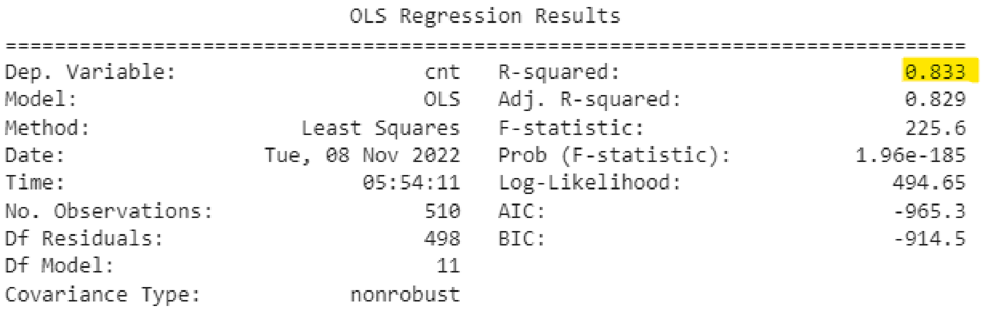
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A. Residual Analysis: (Errors are normal distributed)



Predictors and Target Variables have Linear Relationship:

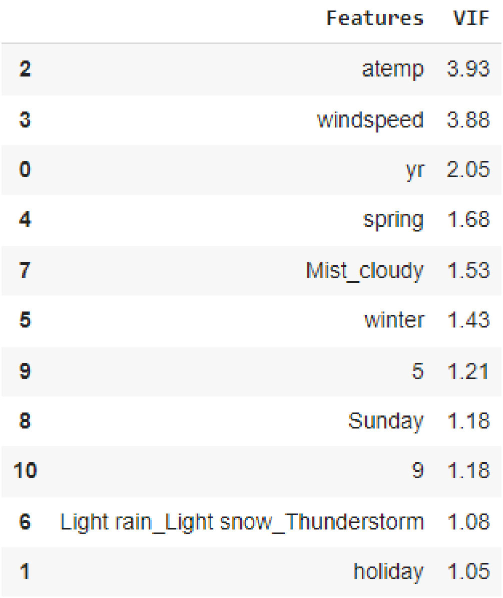
R^2 value = 83.33%



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.833
Model:                            OLS   Adj. R-squared:                  0.829
Method:                 Least Squares   F-statistic:                     225.6
Date:                Tue, 08 Nov 2022   Prob (F-statistic):          1.96e-185
Time:                        05:54:11   Log-Likelihood:                 494.65
No. Observations:                 510   AIC:                            -965.3
Df Residuals:                     498   BIC:                            -914.5
Df Model:                          11
Covariance Type:            nonrobust
```

Multicollinearity:

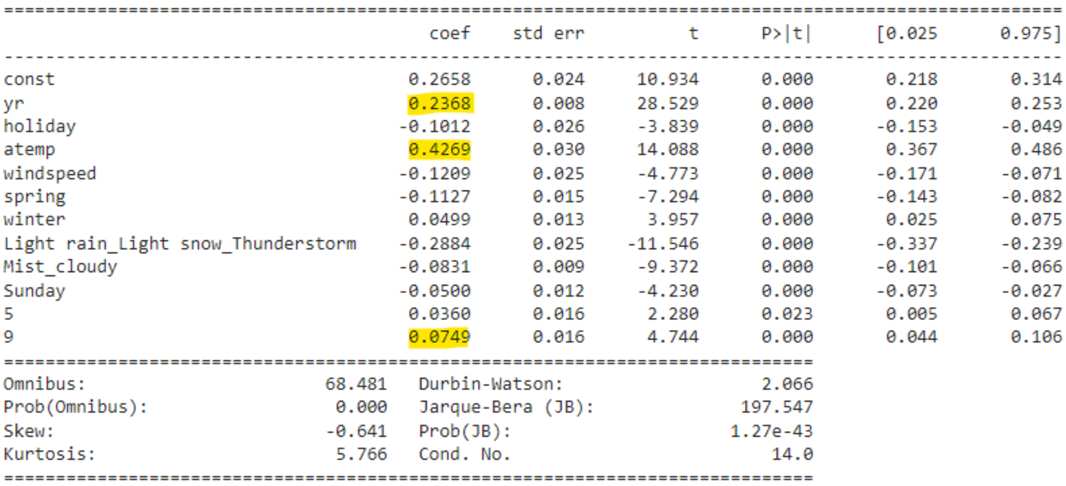
VIF: VIF for all variables are less than 5.



	Features	VIF
2	atemp	3.93
3	windspeed	3.88
0	yr	2.05
4	spring	1.68
7	Mist_cloudy	1.53
5	winter	1.43
9	5	1.21
8	Sunday	1.18
10	9	1.18
6	Light rain_Light snow_Thunderstorm	1.08
1	holiday	1.05

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   A. 'atemp', 'yr' and '9' (month 9) are highly contributing with coefficients '0.4269', '0.2368' and '0.0749' respectively .



```
==========================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                   0.2658      0.024     10.934      0.000       0.218       0.314
yr                                      0.2368      0.008     28.529      0.000       0.220       0.253
holiday                                -0.1012      0.026     -3.839      0.000      -0.153      -0.049
atemp                                   0.4269      0.030     14.088      0.000       0.367       0.486
windspeed                              -0.1209      0.025     -4.773      0.000      -0.171      -0.071
spring                                 -0.1127      0.015     -7.294      0.000      -0.143      -0.082
winter                                  0.0499      0.013      3.957      0.000       0.025       0.075
Light rain_Light snow_Thunderstorm     -0.2884      0.025    -11.546      0.000      -0.337      -0.239
Mist_cloudy                            -0.0831      0.009     -9.372      0.000      -0.101      -0.066
Sunday                                 -0.0500      0.012     -4.230      0.000      -0.073      -0.027
5                                       0.0360      0.016      2.280      0.023       0.005       0.067
9                                       0.0749      0.016      4.744      0.000       0.044       0.106
==============================================================================
Omnibus:                       68.481   Durbin-Watson:                   2.066
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              197.547
Skew:                          -0.641   Prob(JB):                     1.27e-43
Kurtosis:                       5.766   Cond. No.                         14.0
==============================================================================
```

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   A. In Modelling, Linear Regression is used by assuming that the input variables have a linear relationship with the target output variable. The parameters 'slope'(m) and 'intercepts'(c) are found by using least square method.

   Mathematical format of Linear Regression: y=mX+c (Single Linear Regression)

   Y=m1X1+m2X2+m3X3+c (Multi Linear Regression)

   Assuming of Linear Regression:

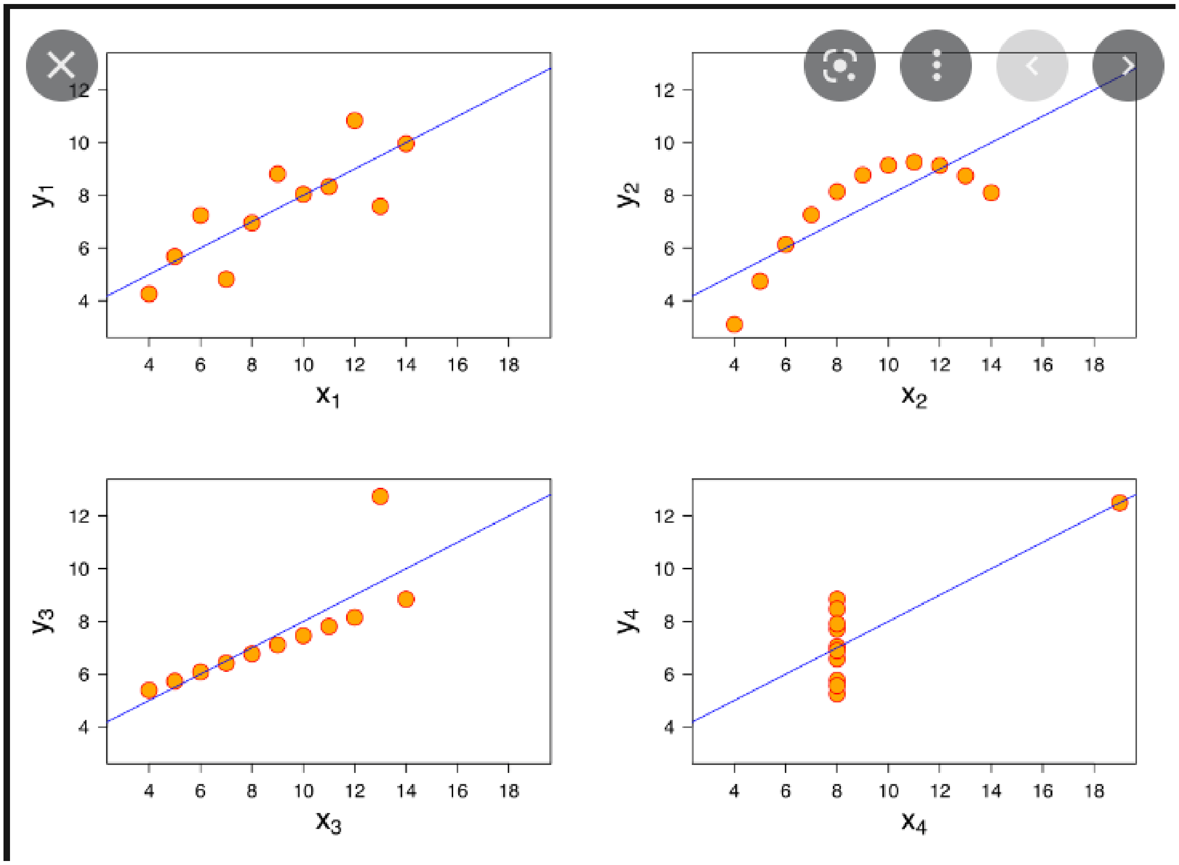
   a) Errors must be Normally distributes.

b) Predictors and Target Variable should have Linear Relationship

c) Error terms should be Independent to each other.

2. Explain the Anscombe's quartet in detail. (3 marks)

A. Anscombe's quartet is a set of 4 datasets which have statistically same properties but when plotted, they tell different story. So it is always suggested to plot the data before driving the analysis where we can see whether the data is linear, nonlinear and if it has any outliers.



Picture from wikipedia

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

A. Pearson R is a statistic that measures the correlation between two variables. It has a numerical range from -1 to +1.

Negatively Correlated: -1 to 0

Positively Correlated: 0 to 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

A. A pre processing step for data that is applied to a predictor variable , to make the data within a specific range.

It helps model to speed up the performance and process of model training. Also it helps to standardize all variables with different units to a single range.

Normalization: It transforms data to a range of 0 to 1

Standardization: It transforms data to a range of -1 to 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A. When there is perfect correlation, R^2 value becomes '1' , the denominator of VIF becomes '0' which leads to 'infinity'

$$VIF = \frac{1}{1 - R_i^{\,2}}$$

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

(3 marks)

A. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.