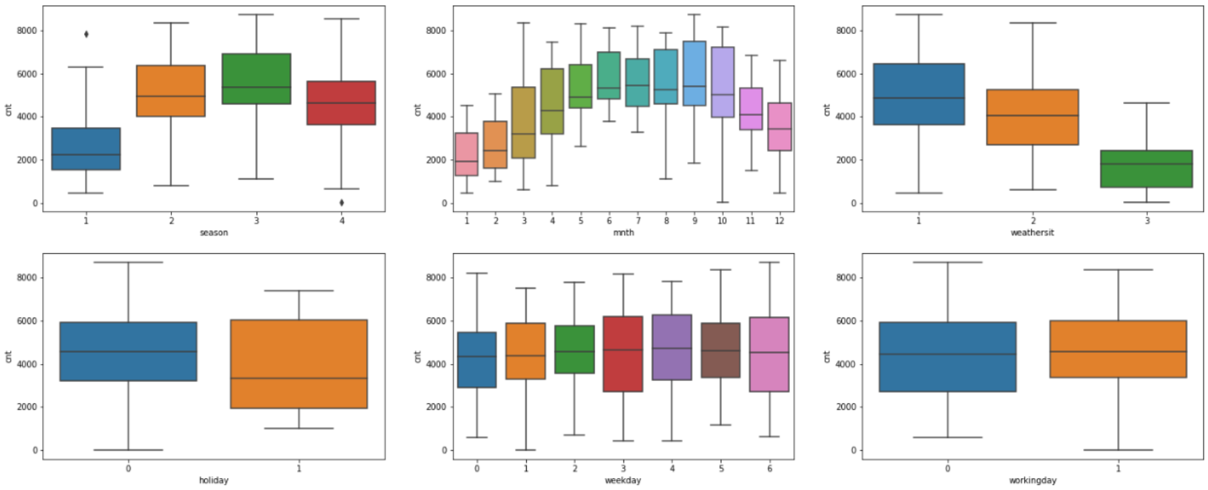
Assignment-based Subjective Questions

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



season: Maximum bike booking were happening in season3 with a median of over 5000 booking). This was followed by season2 & season4
mnth: Most of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month.
weathersit: Most of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking
holiday: Most of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
weekday: weekday variable shows very close trend having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
workingday: Most of the bike booking were happening in 'workingday' with a median of close to 5000 booking. This indicates, workingday can be a good predictor for the dependent variable

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   It removes the extra column which produced during dummy variable creation

Ex: If a column with 3 categories and if we want to create dummy variables it creates 3 columns but if you give drop_first= True then it only creates 2 columns

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

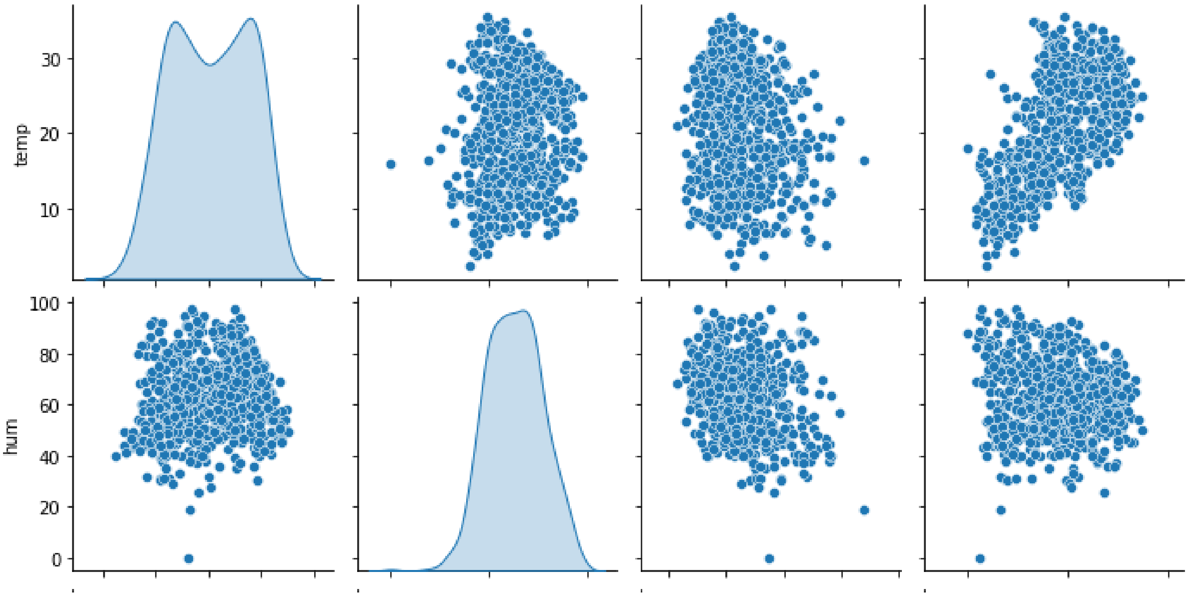
   'temp' and 'atemp' columns high correlation between the target variable 'cnt'

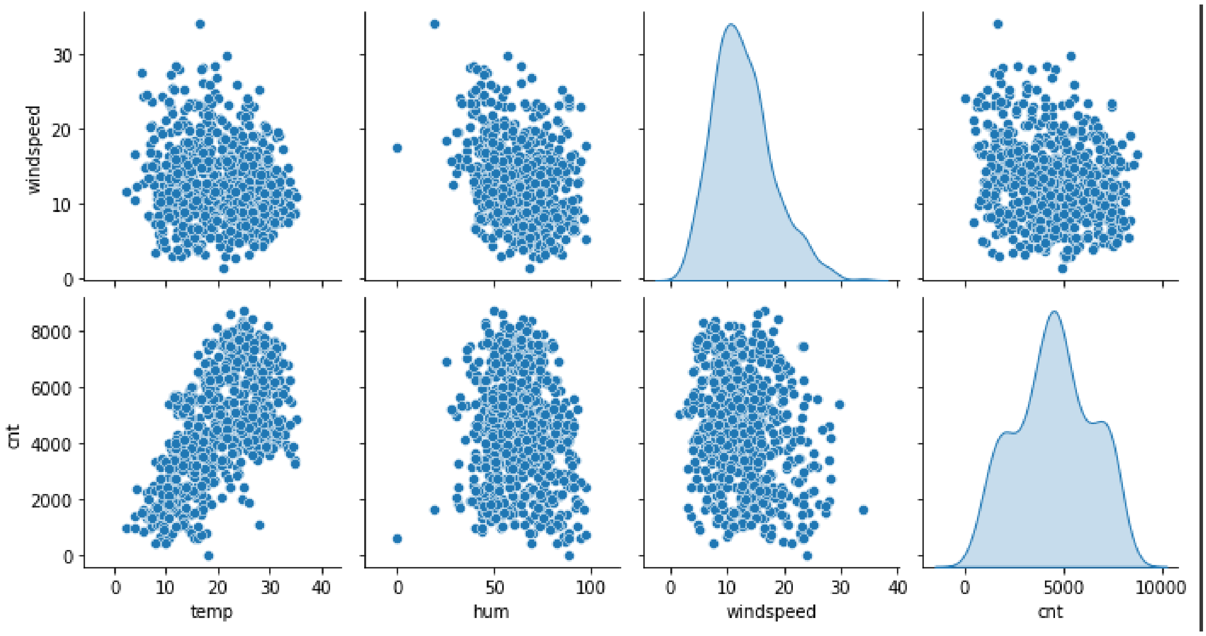
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   There are 5 basic assumptions of Linear Regression Algorithm:

   1. Linear Relationship between the features and target:

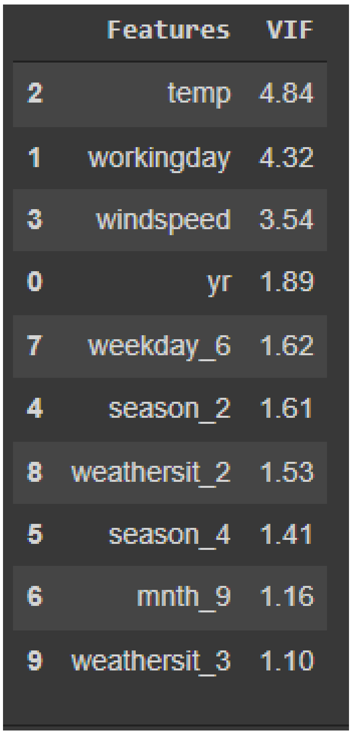
   According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.





2. Little or no Multicollinearity between the features:

Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables.

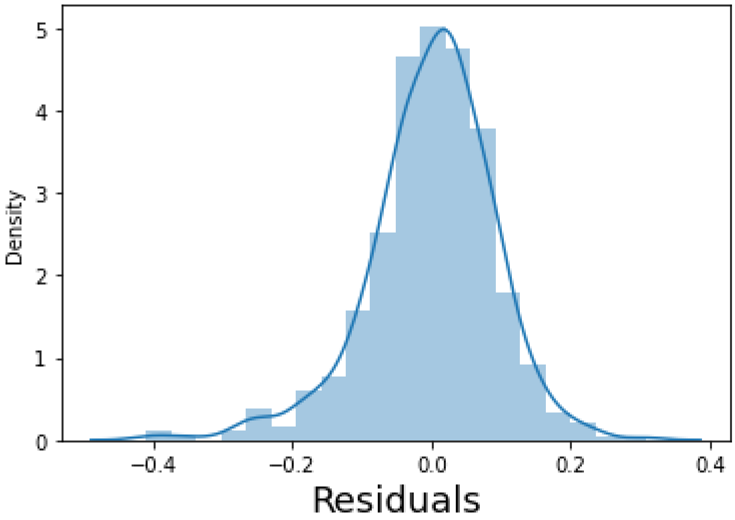


	Features	VIF
2	temp	4.84
1	workingday	4.32
3	windspeed	3.54
0	yr	1.89
7	weekday_6	1.62
4	season_2	1.61
8	weathersit_2	1.53
5	season_4	1.41
6	mnth_9	1.16
9	weathersit_3	1.10

From this we can find that there is no correlation between the independent variables

3. Homoscedasticity Assumption:

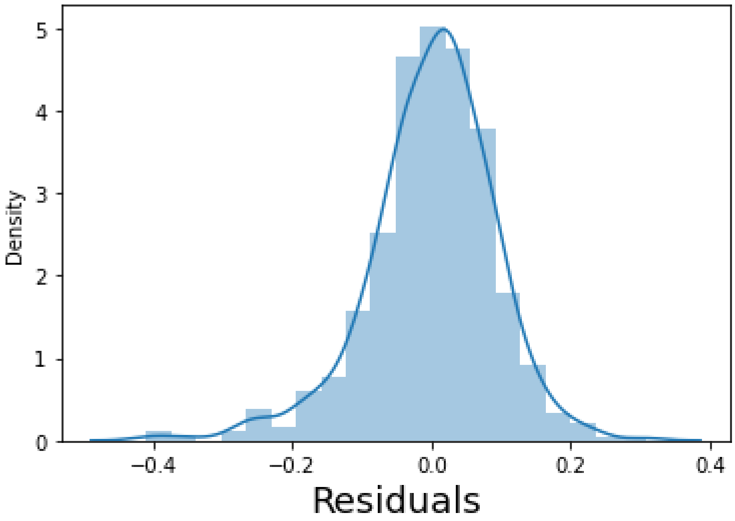
Homoscedasticity describes a situation in which the error term is the same across all values of the independent variables.A scatter plot of residual values vs predicted values is a goodway to check for homoscedasticity.



Here error term is normally distributed and the mean is zero. Hence assumption is satisfied

4. **Normal distribution of error terms:**

The fourth assumption is that the error(residuals) follow a normal distribution.



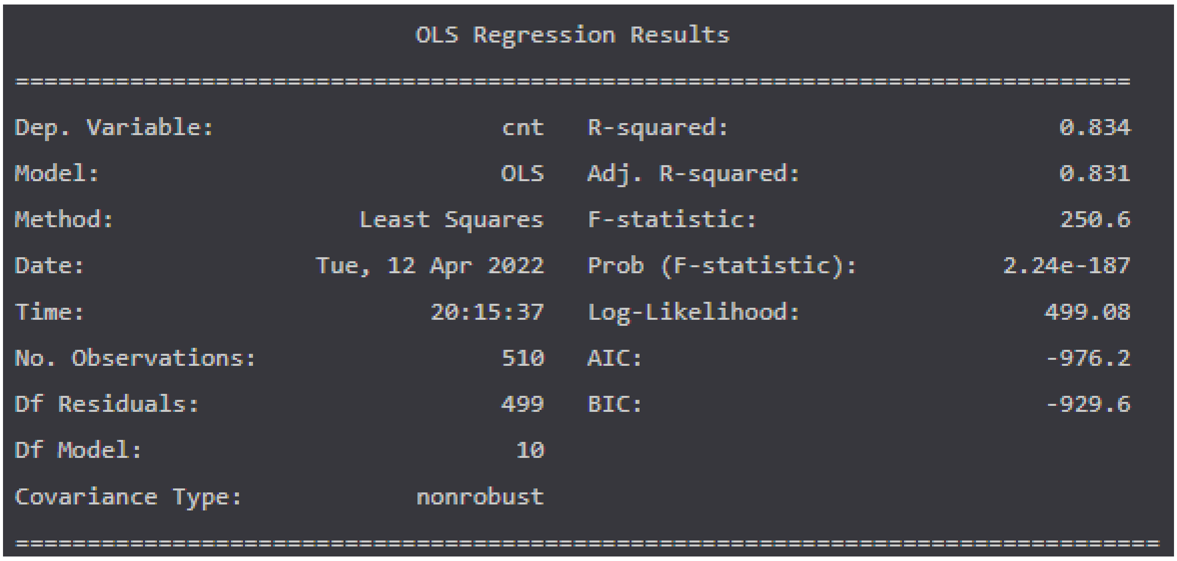
This is the Residual plot obtained from the final model. Here it shows the normal distribution and it satisfies the assumption

5. **Little or No autocorrelation in the residuals:**

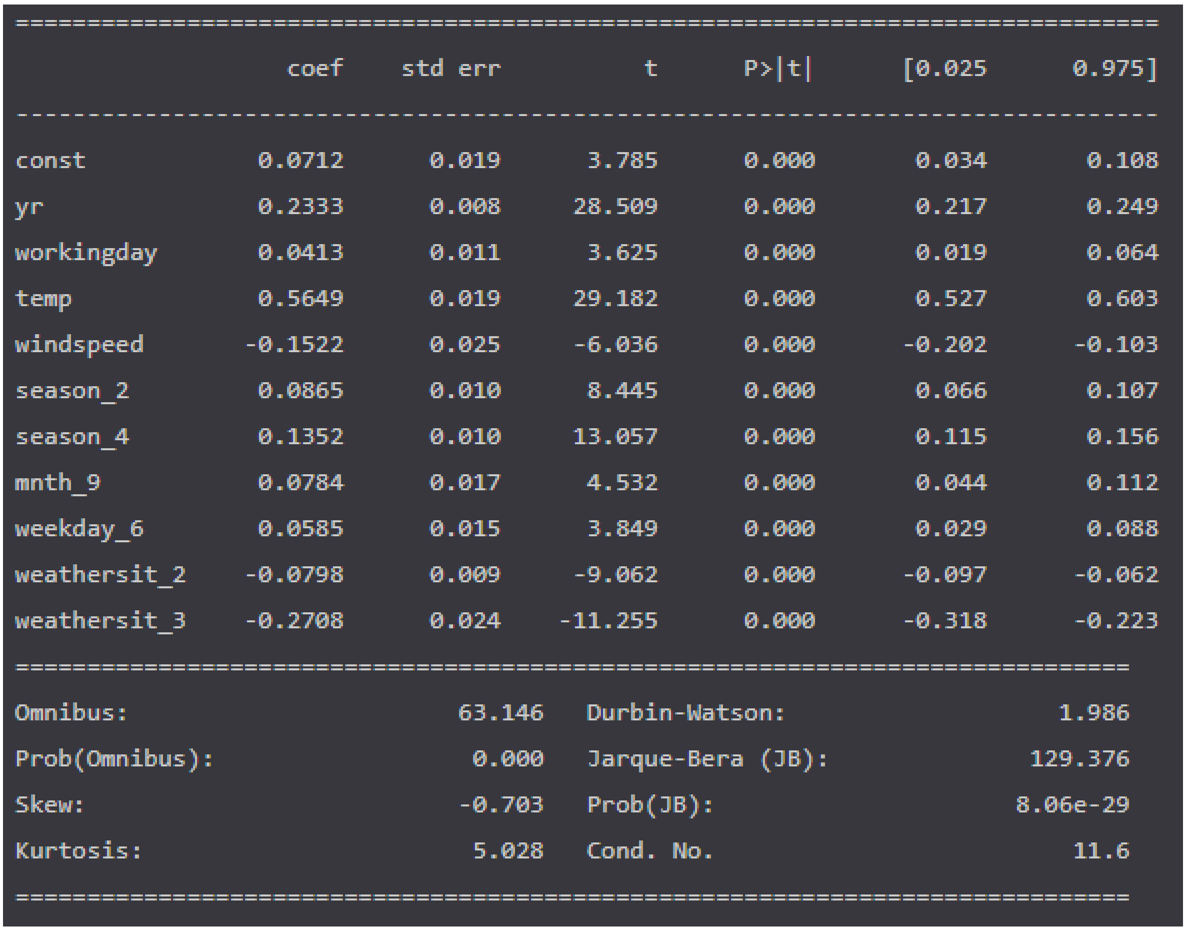
   The Durbin Watson statistic is a test for autocorrelation in a regression model's output.

The DW statistic ranges from zero to four, with a value of 2.0 indicating zero autocorrelation.

Values below 2.0 mean there is positive autocorrelation and above 2.0 indicates negative autocorrelation.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.831
Method:                 Least Squares   F-statistic:                     250.6
Date:                Tue, 12 Apr 2022   Prob (F-statistic):          2.24e-187
Time:                        20:15:37   Log-Likelihood:                 499.08
No. Observations:                 510   AIC:                            -976.2
Df Residuals:                     499   BIC:                            -929.6
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
```



```
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.0712      0.019      3.785      0.000       0.034       0.108
yr               0.2333      0.008     28.509      0.000       0.217       0.249
workingday       0.0413      0.011      3.625      0.000       0.019       0.064
temp             0.5649      0.019     29.182      0.000       0.527       0.603
windspeed       -0.1522      0.025     -6.036      0.000      -0.202      -0.103
season_2         0.0865      0.010      8.445      0.000       0.066       0.107
season_4         0.1352      0.010     13.057      0.000       0.115       0.156
mnth_9           0.0784      0.017      4.532      0.000       0.044       0.112
weekday_6        0.0585      0.015      3.849      0.000       0.029       0.088
weathersit_2    -0.0798      0.009     -9.062      0.000      -0.097      -0.062
weathersit_3    -0.2708      0.024    -11.255      0.000      -0.318      -0.223
================================================================================
Omnibus:                       63.146   Durbin-Watson:                   1.986
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              129.376
Skew:                          -0.703   Prob(JB):                     8.06e-29
Kurtosis:                       5.028   Cond. No.                         11.6
==============================================================================
```

From our final model Durbin-Watson is 1.986 which is very close to 2 since there is zero correlation.

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

   Based on the final model top 3 features are

Features	**VIF**
temp	4.84
workingday	4.32
windspeed	3.54

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   - ML algorithm which is based on supervised learning
   - It attempts to explain the relationship between dependent variable(y) and independent variable(x) using a straight line Y=mx+c
   - Creates best fit line to the provided data to find the best linear relationship between the independent and dependent variables.
   - The best-fit line is obtained by minimising a quantity called Residual Sum of Squares (RSS), this is the best time to be introduced to what is known as the cost function.
   - The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$
     - RSS: Residual Sum of Squares
     - TSS: Total Sum of Squares
   - It is of 2 types
     - Simple Linear regression : When the number of independent variables is 1
       Y= mx+c
     - Multiple linear regression : When the number of independent variables is more than 1
       Y = c+m1.x1+m2.x2+m3.x3

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

Anscombe’s Quartet was developed by statistician Francis Anscombe. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics.
Anscombe’s Quartet warns the dangers of outliers in data sets.

3. What is Pearson’s R? (3 marks)
   Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

   Pearson’s correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a “product moment”, that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Scaling is the process to normalize the data within a particular range. Many times, we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

   The two scaling methods
   - Normalization
     Normalization typically scales the values into a range of 0 to 1
     x = x-min(x)/(max(x) - min(x))
   - Standardization
     Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).
     x = x-mean(x)/Sd(x)

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

Formula: VIF = 1/ (1-R^2)
If R^2=1(Residual sum of squares) then VIF is infinite. It says there is perfect correlation in variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.
   Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.
   Importance of Q-Q plot: Below are the points:
   I. The sample sizes do not need to be equal.
   II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
   III. The q-q plot can provide more insight into the nature of the difference than analytical methods.