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?**

   From the final model, I could infer that the categorical variables yr 2019, holiday, Spring, Summer, Dec, Jan, Nov, Sep, Light Rain/Snow, Misty has been only used in model creation without any continuous variables. It was sufficient to predict the values with adj.R2 of 76.7% with train dataset.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.772
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	169.0
Date:	Sat, 24 Oct 2020	Prob (F-statistic):	3.79e-153
Time:	19:57:18	Log-Likelihood:	415.91
No. Observations:	510	AIC:	-809.8
Df Residuals:	499	BIC:	-763.2
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5309	0.011	49.274	0.000	0.510	0.552
yr 2019	0.2452	0.010	25.469	0.000	0.226	0.264
holiday	-0.0879	0.031	-2.844	0.005	-0.149	-0.027
Spring	-0.2616	0.015	-17.771	0.000	-0.291	-0.233
Summer	-0.0521	0.013	-3.988	0.000	-0.078	-0.026
Dec	-0.1109	0.018	-6.108	0.000	-0.147	-0.075
Jan	-0.1062	0.021	-5.178	0.000	-0.146	-0.066
Nov	-0.1189	0.019	-6.349	0.000	-0.156	-0.082
Sep	0.0662	0.019	3.461	0.001	0.029	0.104
Light Rain/Snow	-0.3349	0.029	-11.639	0.000	-0.391	-0.278
Misty	-0.0861	0.010	-8.412	0.000	-0.106	-0.066

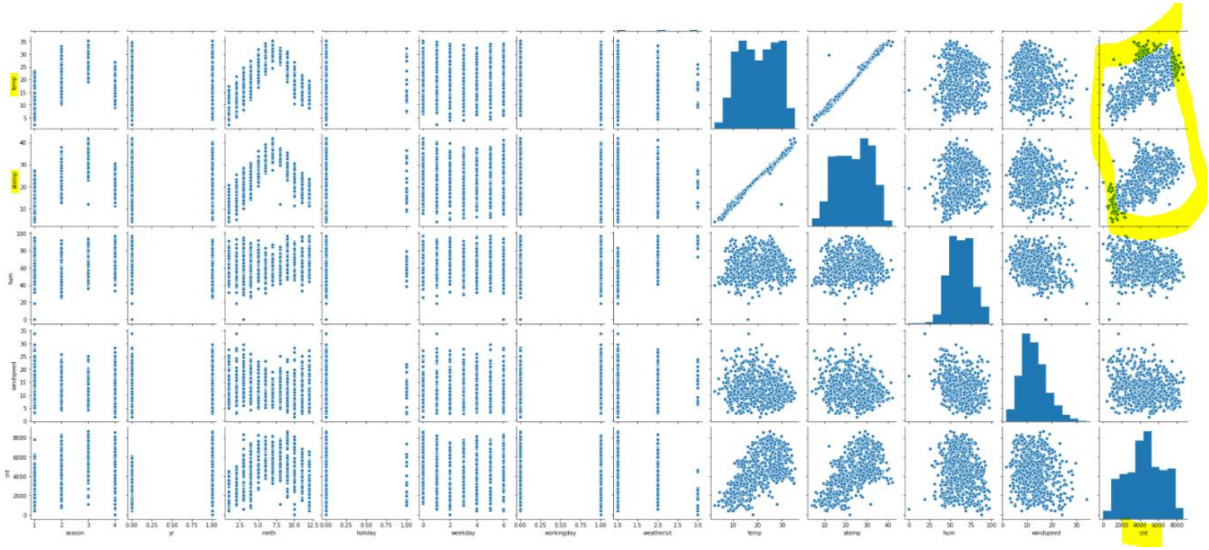
Omnibus:	68.305	Durbin-Watson:	1.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	198.368
Skew:	-0.637	Prob(JB):	8.41e-44
Kurtosis:	5.777	Cond. No.	8.28

- From the weathersit column, the days which have clear sky will have higher cnt compared to other two categories in the given dataset.
- From the season type, Spring has negative impact comparatively to other seasons.
- During a holiday, the model projects lesser demand than working day.
- The cnt has improved in year 2019 and hence we see a positive coefficient.

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

   The drop_first is important in columns with lesser categories because, the data will get negative correlation of each other. If number of categories are more then the drop _first will not be required as negative correlation will not be much affected.

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



   Temp and aTemp are the variables equally and highly correlated to count.

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

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.772
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	169.0
Date:	Sat, 24 Oct 2020	Prob (F-statistic):	3.79e-153
Time:	19:57:18	Log-Likelihood:	415.91
No. Observations:	510	AIC:	-809.8
Df Residuals:	499	BIC:	-763.2
Df Model:	10		
Covariance Type:	nonrobust		

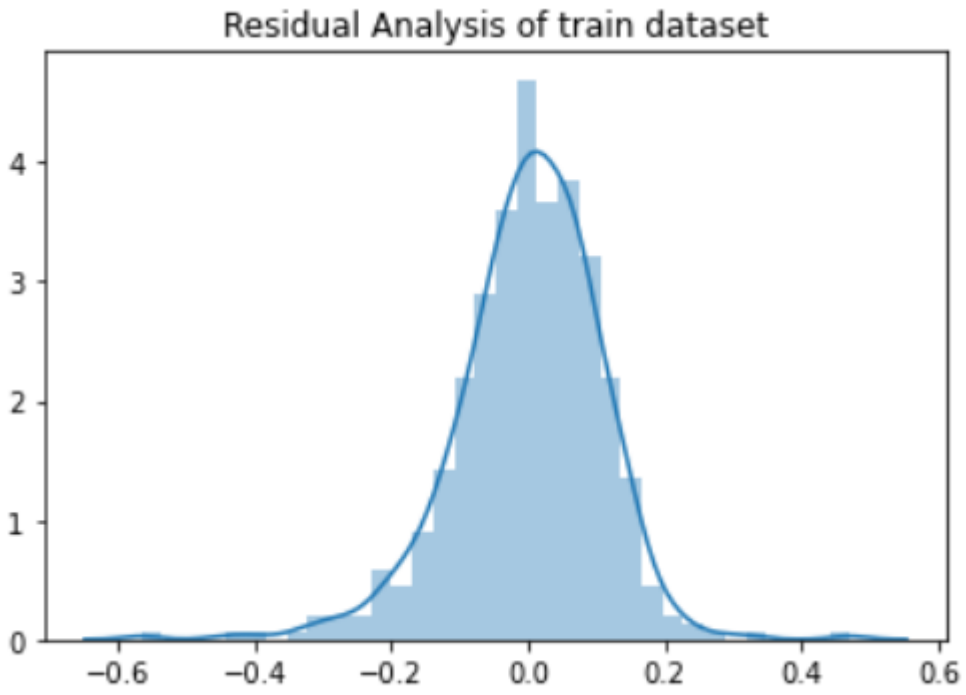
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5309	0.011	49.274	0.000	0.510	0.552
yr 2019	0.2452	0.010	25.469	0.000	0.226	0.264
holiday	-0.0879	0.031	-2.844	0.005	-0.149	-0.027
Spring	-0.2616	0.015	-17.771	0.000	-0.291	-0.233
Summer	-0.0521	0.013	-3.988	0.000	-0.078	-0.026
Dec	-0.1109	0.018	-6.108	0.000	-0.147	-0.075
Jan	-0.1062	0.021	-5.178	0.000	-0.146	-0.066
Nov	-0.1189	0.019	-6.349	0.000	-0.156	-0.082
Sep	0.0662	0.019	3.461	0.001	0.029	0.104
Light Rain/Snow	-0.3349	0.029	-11.639	0.000	-0.391	-0.278
Misty	-0.0861	0.010	-8.412	0.000	-0.106	-0.066

Omnibus:	68.305	Durbin-Watson:	1.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	198.368
Skew:	-0.637	Prob(JB):	8.41e-44
Kurtosis:	5.777	Cond. No.	8.28

In the OLS regression results, We check adj.R squared(76.7%) and Prob(F-statistic) is less than 1(3.79 e-153)

	Features	VIF
2	Spring	1.84
5	Jan	1.61
0	yr 2019	1.55
9	Misty	1.44
3	Summer	1.31
6	Nov	1.12
7	Sep	1.12
4	Dec	1.10
1	holiday	1.06
8	Light Rain/Snow	1.03

VIF of all variables less than 2.



The residual is distributed normally.

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

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.772
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	169.0
Date:	Sat, 24 Oct 2020	Prob (F-statistic):	3.79e-153
Time:	19:57:18	Log-Likelihood:	415.91
No. Observations:	510	AIC:	-809.8
Df Residuals:	499	BIC:	-763.2
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5309	0.011	49.274	0.000	0.510	0.552
yr 2019	0.2452	0.010	25.469	0.000	0.226	0.264
holiday	-0.0879	0.031	-2.844	0.005	-0.149	-0.027
Spring	-0.2616	0.015	-17.771	0.000	-0.291	-0.233
Summer	-0.0521	0.013	-3.988	0.000	-0.078	-0.026
Dec	-0.1109	0.018	-6.108	0.000	-0.147	-0.075
Jan	-0.1062	0.021	-5.178	0.000	-0.146	-0.066
Nov	-0.1189	0.019	-6.349	0.000	-0.156	-0.082
Sep	0.0662	0.019	3.461	0.001	0.029	0.104
Light Rain/Snow	-0.3349	0.029	-11.639	0.000	-0.391	-0.278
Misty	-0.0861	0.010	-8.412	0.000	-0.106	-0.066

Omnibus:	68.305	Durbin-Watson:	1.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	198.368
Skew:	-0.637	Prob(JB):	8.41e-44
Kurtosis:	5.777	Cond. No.	8.28

Yr 2019- 0.2452 , if the data is from 2019 it is positively correlated to cnt
Light Rain/snow - -0.3349, if the weather is with light rain/ snow it reduces the demand of bikes
Spring - -0.2616 – if the season is spring(initial months of an year) it is negatively correlated, thereby it has a less demand.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   It is the relationship between one dependant variable and one or many independent variables formulated to form a linear model

$$Y = c + m_1x_1 + m_2x_2 + m_3x_3 + m_4x_4 + m_5x_5 \ldots\ldots + m_nx_n$$

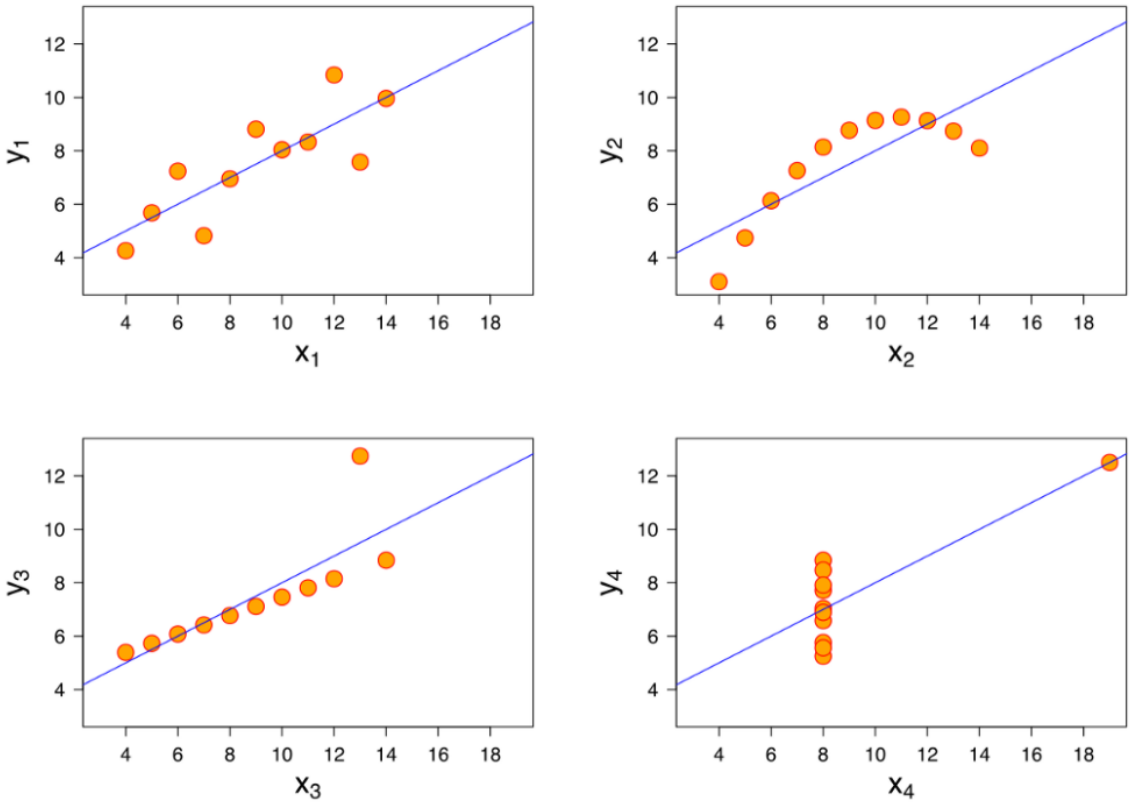
Y = dependant variable
c = intercept
m = slope or coefficient
x = independent variables

For a good linear regression model:
Adj R2 between train and test should be in absolute difference of 5% and greater than 75%.
F statsistic probability should be very less than 1

The distribution of residual should be normally distributed and with mean 0.
Variable Inflation Factor(VIF) is should be at least of 5 for all the variables in the model.

Variable selection method in Multiple linear regression:
RFE- Recursive Feature Exclusion- automated approach
Manual approach - by manually adding/ excluding in a recursive manner.

2. **Explain the Anscombe's quartet in detail.**
   It comprises of four different datasets in which the statistical properties like mean, correlation and standard deviation remain same on all datasets yet the distribution remains different among the each and every dataset. And as the result of this the data visualization remains an important part in the data analysis without which the model selection can't be significant.



The summary statistics show that the means and the variances were identical for x and y across the groups :

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

In the above plots, the linear relationship is well fit for X1,Y1 whereas the same model doesn't fit the other datasets even though the statistical properties remain same among four.

3. **What is Pearson's R?**
   It is a correlation coefficient whose values are between -1 to 1 is related between two variables.
   If the coefficient is 1 then the variables have positive linear relationship and if the coefficient is -1 then the variables have negative linear relationship and in case of 0 there is no relationship between the variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   It is a feature used especially in multiple linear regression where multiple independent variables are present in different scale and with this method the variables are normalized and brought under same scale. Without this feature model will bring strange coefficients and hard to interpret.
   This is done by two methods:
   Normalized scaling: the values normalized and will be in the range of 0 – 1.

$$x = \frac{x - min}{max - \min}$$

   It is also known as minmax scaling.
   Standardized scaling: the variables are scaled based on mean=0 and sd=1.

$$x = \frac{x - mean}{sd}$$

The advantage of Normalized scaling over Standardized is the outliers are handled by default.

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

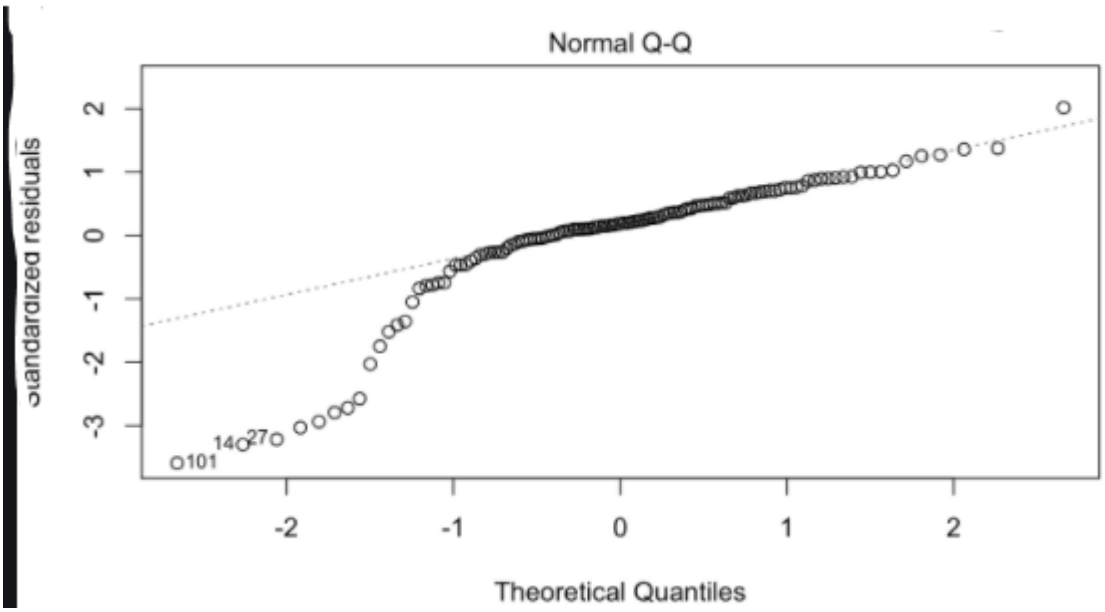
   This happens when two independent variables are highly correlated, which results the r- squared value of 1.

$$VIF = \frac{1}{1 - R^2}$$

   if R-squared =1, VIF = ∞ . In this case one of the variables has to be eliminated and work on the model.

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

   Q-Q plot is comparison plot to check two datasets are of same distribution. The size of the dataset may differ as the comparison is done based on quantiles i.e., the data after 10, 20,30,40,50,60,70,80,90,100 percentile of data. And when these points are plotted over scatter plot, it should pass through the line having slope of 45 degrees over the x axis.



   If the most of the points passes through the 45 degree line then both datasets are in same distribution.

   In linear regression this is used when the train and test datasets are given separately, we could use the Q-Q plot to check that the train and test have the same distribution.