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?**

A. From the analysis of categorical variables, we can infer the following

Season : Bike rentals seem to be considerably less in Spring season than all the other seasons. Summer and Fall seem to have higher demand for bikes followed by winter.

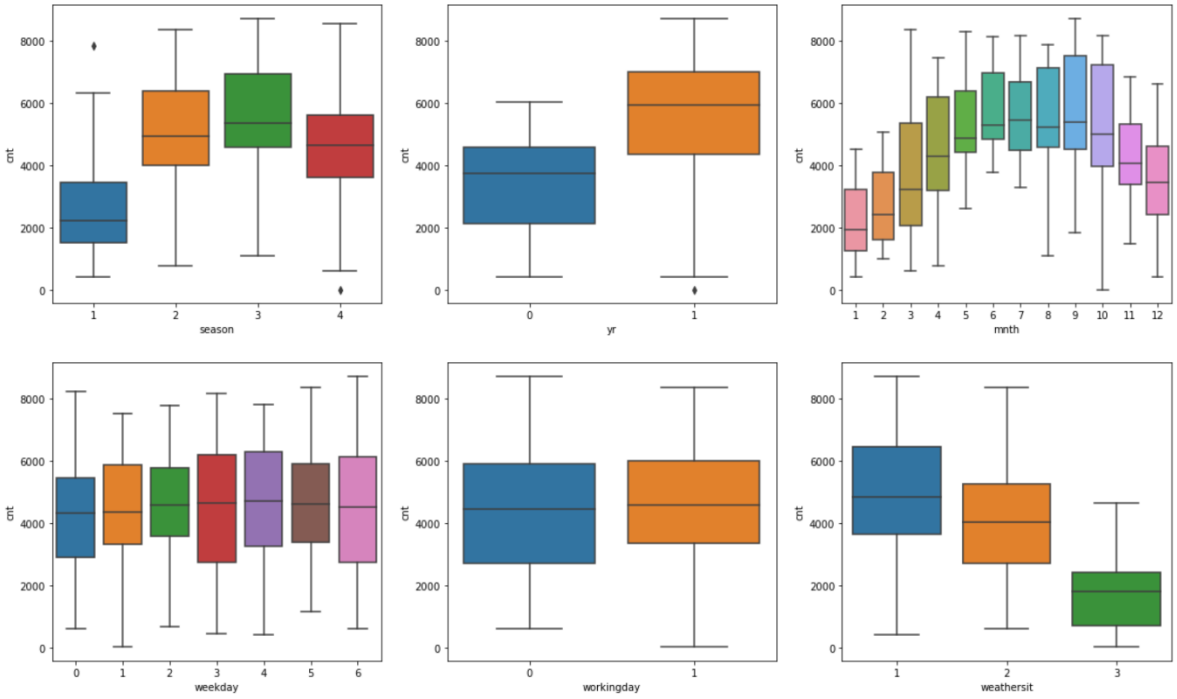
Year : There is a significant increase in bike rentals in the year 2019 compared to 2018. Might be the bike sharing became more popular over the year due to advertisements of the company or may be as people got used to bike sharing year on year.

Month : Bike rentals seem to be higher in the months from May to October then the other months. This seems to be related to the season as the season is summer and fall during may to October.

Weekday: Weekday doesn't seem to have any impact on the bike rentals. Requirement of the bikes doesn't seem to depend on a specific day of a week. All the days of a week seem to have almost the same demand for the bikes.

Working Day : Working day doesn't seem to have any impact on the bike rentals which means the bike rentals are not dependent on if it's a working day, weekend or a holiday.

Weather : Bike Rentals are considerably low when there is thunder storm, light rain and snow and high when the weather is clear with few clouds



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

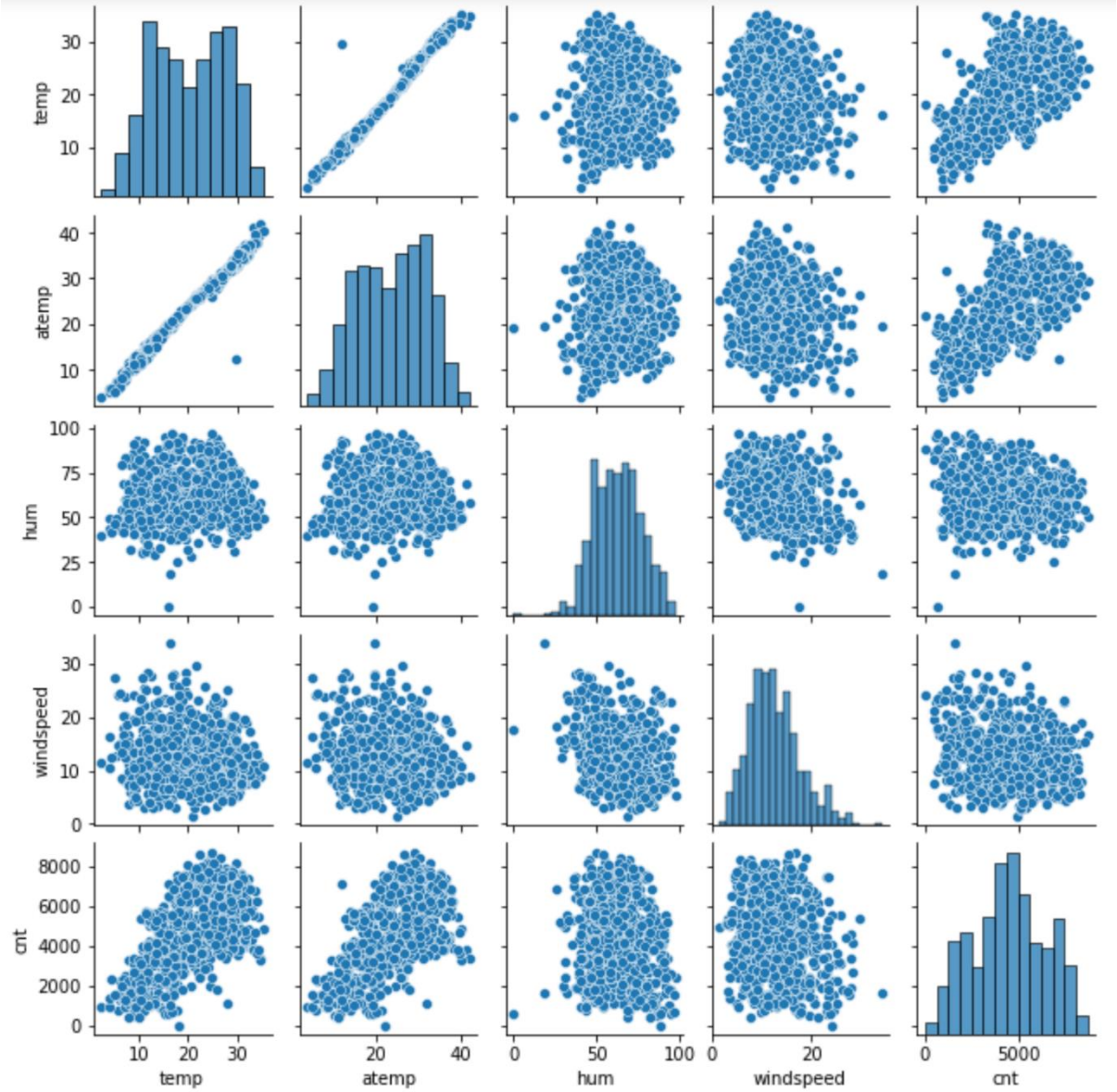
A. It is important to use drop_first =True for the following reasons
   > We can define the categorical variable data by having one column less and also without losing any data.
   > If we have all the columns created by the dummy variables, the columns will be highly correlated and will result in multicollinearity while building the model.

	Features	VIF
29	LightSnow	inf
2	workingday	inf
28	Mist	inf
27	Clear	inf
7	spring	inf
8	summer	inf
9	fall	inf
10	winter	inf
22	Sunday	inf
26	Saturday	inf
1	holiday	inf
3	temp	66.85
4	atemp	55.97
20	Nov	6.78

When the model is built without using drop_first=True, the VIF values turn out to be infinite which means the predictor variables are correlated. For example, the impact of one of the predictor variable could be explained with the other predictor variables

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

A. Looking at the pair-plot among numerical variables, the variables “temp” (temperature) and “atemp” (feeling temperature) have 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?**

A. Following assumptions of Linear Regression are validated after building the model

➔ **There is a linear relationship between the dependant and the independent variables**
After building the model, the R-squared calculation on the test data has a good R-squared value of 77.2% which is very close to the R-squared value on the training data which was 77.9%
From this we can say that there is a linear relation between the dependent and independent variables

R-Squared on training data is as follows

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.779
Model:	OLS	**Adj. R-squared:**	0.774
Method:	Least Squares	**F-statistic:**	146.0
Date:	Tue, 11 Jan 2022	**Prob (F-statistic):**	2.69e-154
Time:	00:11:56	**Log-Likelihood:**	423.93
No. Observations:	510	**AIC:**	-821.9
Df Residuals:	497	**BIC:**	-766.8
Df Model:	12		
Covariance Type:	nonrobust		

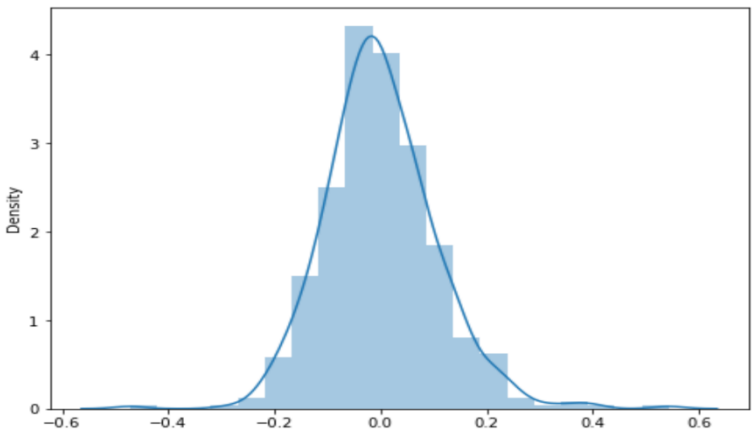
R-squared on test data is as follows

```
1 r2_score(y_true=y_test, y_pred=y_pred)
```

```
0.7720998995152383
```

➔ Residuals should be normally distributed and they have a mean value of zero

Below plot shows that the residuals are normally distributed and they are normally distributed around the value of zero



➔ There is no multicollinearity between the independent variables
The p-values of all the coefficients is almost zero and the VIF values are < 5 which indicates that the independent variables used for building the final model are not correlated and thus there is no multicollinearity

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2346	0.019	12.576	0.000	0.198	0.271
yr	0.2472	0.010	26.018	0.000	0.229	0.266
workingday	0.0573	0.013	4.414	0.000	0.032	0.083
windspeed	-0.1741	0.029	-5.978	0.000	-0.231	-0.117
summer	0.2566	0.014	18.946	0.000	0.230	0.283
fall	0.2755	0.017	15.994	0.000	0.242	0.309
winter	0.1881	0.016	12.118	0.000	0.158	0.219
Aug	0.0413	0.021	1.990	0.047	0.001	0.082
Sep	0.0970	0.020	4.767	0.000	0.057	0.137
Oct	0.0975	0.020	4.792	0.000	0.058	0.137
Saturday	0.0644	0.017	3.853	0.000	0.032	0.097
Mist	-0.0942	0.010	-9.252	0.000	-0.114	-0.074
LightSnow	-0.3148	0.029	-10.891	0.000	-0.372	-0.258

	Features	VIF
1	workingday	3.65
2	windspeed	3.28
4	fall	2.94
5	winter	2.20
0	yr	1.94
6	Aug	1.85
3	summer	1.77
9	Saturday	1.61
8	Oct	1.56
10	Mist	1.55
7	Sep	1.49
11	LightSnow	1.10

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

A. Based on the final model, I feel that the following 3 features contribute significantly towards explaining the demand of the shared bikes

- ➔ Year
- ➔ Season
- ➔ Weather Status

Out of the features which were used to build the final model,

When a model is built with variable "year" alone, the variance in the target variable was explained upto 35%

```
1 X_train_rfe=X_train["yr"]
2 X_train_lm=sm.add_constant(X_train_rfe)
3 lr_model=sm.OLS(y_train,X_train_lm).fit()
4 lr_model.summary()
```

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.350
Model:	OLS	**Adj. R-squared:**	0.349
Method:	Least Squares	**F-statistic:**	273.3
Date:	Wed, 12 Jan 2022	**Prob (F-statistic):**	1.95e-49
Time:	12:18:43	**Log-Likelihood:**	148.67
No. Observations:	510	**AIC:**	-293.3
Df Residuals:	508	**BIC:**	-284.9
Df Model:	1		
Covariance Type:	nonrobust		

When the model was built with "year" and "season" together, the variance in the target variable was explained upto 66.3%

Here season includes all the seasons from the final selected features – summer, fall and winter

```
col=["summer","fall","winter","yr"]
X_train_rfe=X_train[col]
X_train_lm=sm.add_constant(X_train_rfe)
lr_model=sm.OLS(y_train,X_train_lm).fit()
lr_model.summary()
```

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.663
Model:	OLS	**Adj. R-squared:**	0.661
Method:	Least Squares	**F-statistic:**	248.8
Date:	Wed, 12 Jan 2022	**Prob (F-statistic):**	6.73e-118
Time:	12:45:53	**Log-Likelihood:**	316.55
No. Observations:	510	**AIC:**	-623.1
Df Residuals:	505	**BIC:**	-601.9
Df Model:	4		
Covariance Type:	nonrobust		

When the model was build with “year”, “season” and “weather” together, the variance in the target variable was explained upto 73.5%

Here season includes all the seasons from the final selected features – summer, fall and winter

And weather includes “Mist” and “LightSnow” from the selected features

```
col=["summer","fall","winter","yr", "Mist","LightSnow"]
X_train_rfe=X_train[col]
X_train_lm=sm.add_constant(X_train_rfe)
lr_model=sm.OLS(y_train,X_train_lm).fit()
lr_model.summary()
```

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.735
Model:	OLS	**Adj. R-squared:**	0.732
Method:	Least Squares	**F-statistic:**	232.5
Date:	Wed, 12 Jan 2022	**Prob (F-statistic):**	1.55e-141
Time:	12:49:54	**Log-Likelihood:**	377.54
No. Observations:	510	**AIC:**	-741.1
Df Residuals:	503	**BIC:**	-711.4
Df Model:	6		
Covariance Type:	nonrobust		

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

A. Linear Regression is a machine learning algorithm used for predicting an output depending on the previous data for the impacting variables. Here the output variable is expected to be continuous variable. It is a supervised learning algorithm.
For example, we can predict the scores of students in a course depending on their Xth class marks
Another example could be to predict the sales for a particular commodity based on various factors which impact it like the requirement, availability, usability, importance etc.

The main aim of Linear regression is to find the best line which fits most of the given data.

There are 2 types of linear regression models

1. Simple Linear Regression: Here the output variable is predicted using one of the independent variables. Here, we analyse the impact of the predictor variable on the desired target variable.
2. Multiple Linear Regression: Here the output variable is predicted using multiple independent variables from the data. Here, we analyse the impact of multiple predictor variables on the desired target variable

   We can do multiple linear regression with 2 approaches
   1. Build model using all the variables and remove the variables one-by-one which seem to be of not much importance in predicting the target variable
   2. Build model by adding the features one-by-one and using that to predict the target variable

We start Linear Regression with some assumptions

- There exists a linear relation between the dependent and independent variables
- The error terms are normally distributed with their mean at zero
- The error terms are independent of each other
- There is no multicollinearity in the data i.e., the independent variables are not correlated

2. **Explain the Anscombe's quartet in detail**

A. Anscombe's quartet consists of four data sets which seem to be similar when simple statistics like mean, median etc are used to analyse them.

The differences in their distribution can be understood only when they are plotted on a graph.

Anscombe's quartet signifies the importance of visualising the data before actually building a model. Though by analysing the data from different data sets by using statistics like mean, variance etc seem to have the same distribution and we might feel that same machine learning model can be used for predicting them. But in reality the data distribution might be actually very different where one of the data set might have a linear relationship and the other one might have exponential relation and the third might not have any relationship

3. **What is Pearson's R?**

A. Pearson's R is Pearson's correlation coefficient. It is one of the most significant way of calculating the correlation coefficient. Correlation means the strength of the relationship between the variables which may be positive or negative. It shows how much one variable impacts the value of the other variable. It might be that the second variable increases with an increase in the first variable and decreases with a decrease in the first variable or it could be like increase in first variable results in decrease in second variable and vice-versa.

The value of Pearson's R varies between -1 and 1 where 1 indicates strong positive linear relation and -1 indicates strong negative linear relation and 0 means that there is no linear relation.

Formula for Pearson's R is as follows

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

4. **What is Scaling? Why is scaling performed? What is the difference between Normalised scaling and standardized scaling?**

A. Scaling is a process of multiplying or divided a value with a constant value so that we can get the actual values by simply doing the opposite of the applied function.

For Example, When we have a values of a student marks in maths which was calculated for 100 and we want to have the values for 50, we then divide the marks of the students by 2 to obtain their values for 50.

While building a machine learning model, if we have values of different predictor features in different scales, there will be difference in the calculation of the coefficients and it might impact the actual prediction of the target variable. Here the scaling is very useful to have all the values of the predictor variables to be in the same scale say 0 to 1, so that the coefficients will be more realistic.
For example if we have one of the predictor variable having values in range of 0 to 10 and the other one having the values in the range of 0 to 100, the coefficient of first variable might be high say 400 and the second variable might be low say 40 which might give a wrong indication that increase of 1 unit in the first variable impacts the target variable by an increase of 400 units where as increase of second variable by 1 unit impacts the target variable by an increase of 40 units which might not be the actual case.
Also, when values are used without scaling, the model might consider higher values as more significant and lower values as least significant even though that's not the actual case.

Normalised Scaling: It is a technique where the values are compressed between 0 and 1. Here the minimum value in the data is considered as 0 and the maximum value in the data is considered as 1 and the calculation goes as follows
Scaled value = (Actual value – Min value of data)/(Max value of data – Min value of Data)

Standardised Scaling: It is a technique where the values are scaled depending on the mean and standard deviation of the data. Here the values will be scaled to values between -1 and 1 anf they are distributed around the mean.
Scaled value = (Actual value – Mean)/Standard deviation

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

A. In the data the VIF sometimes had infinite value. This happened when there are values which has high correlation between them, once some of the highly correlated independent values are removed, the VIF values were finite.
When the model was built with all the variables, Observed that the VIF was infinity for some of the variables when I skipped using drop_first = True for the dummy variables. In that case there will be a high correlation when all of the dummy variables created for a feature are used.

	Features	VIF
29	LightSnow	inf
2	workingday	inf
28	Mist	inf
27	Clear	inf
7	spring	inf
8	summer	inf
9	fall	inf
10	winter	inf
22	Sunday	inf
26	Saturday	inf
1	holiday	inf
3	temp	66.85
4	atemp	55.97
20	Nov	6.78
17	Aug	6.47
19	Oct	6.47

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

A. Q-Q plot is a graphical way of analysing if the data from two different data sets has the same distribution. It is also called Quantile-Quantile plot.
Q-Q plot is plotted with quantile values from first data set on X- axis and the quantile values from second data set on Y-axis
We consider a straight line passing through 0.
If both the data sets have the same distribution pattern, they lie on the straight line.
If they don't have the same distribution pattern, then they lie away from the straight line

Advantages:

- It can be plotted with data sets having different sample sizes.
- We can detect the presence of outliers, change in location etc