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?

Ans- In the Bike Sharing Dataset, the important categorical varibales are:

1. Season
2. Month
3. Weather Sitution
4. Year

And our dependent varibale is cnt which means count of total bike rent.
Now we are infering the effect of categorical variable with the dependent variable one by one:-

1. Season:- The maximum demand for the bikes are in the Fall season.
2. Month :- The demand for bike start increasing from the April month till September month. And the maximum demand for the bike is in September month.
3. Weather Situation:- As it is clearly visible from the Boxplot that weathersit 1 which means clear weather attracts more customer than any other weather situations.
4. Year:- Year 2019 generate more demand for customers as comapred to previous year.

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

Ans- While Dummy variable creation , it is the basic rule that if there are n categorical varibales , the there should be (n-1) Dummy variable . Let me clear this with the help of following example:

If there is column known as gender which is Male and Female . And if you create Dummy for these varibale then it is obvious if a person is male then he is not female and if the person is not male then she is female. So we have to just create 1 column as Dummy variable which requires less memory and smoothens our modeling process.

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

Ans- According to my model , temp(temperature) 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?

Ans- The assumptions and their validations of the Linear Regression are as follows:

1. Linear Relationship: There must be Linear Relationship between the dependent variable and their predictors.
   Validation: For this we can make the Pair Plot for our DataFrame.

2. Error Terms are independent of each others
   Validation: For this we have to use Durbin – Watson (DW) statistic.

3. Absence of Multicollinearity
Validation: Here we can find the correlation between the independent variable and them make heatmaps .
After this we can also look for VIF (variance Inflation Factor) among the variables and eliminate those variables having more than 5 VIF.

4. Homoscendasticity:
Validation: We can plot the residual plot vs the Fitted value plot. If the plot shows the funnel shape then there is no Homoscendasticity means it has Heterocendasticity.

5. Error terms are normally distributed:
Validation: For this we have to plot quantile-quantile plot (q-q plot) , if the q-q plot shows straight line then its normally distributed else not.

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

Ans-According to my modal, the top 3 features contributing significantly towards determing the demand for shared bikes are as follows:

1. Temp: Temperature has the coefficient value of '0.4362' , which is highest in my modal. So according to my modal temp is the most prominent factor that affects the bike's demand positively.
2. Yr: Year has the coefficient value of '0.2344', which is second highest in my modal. So it indicates that the bike's demand increases year by year.
3. Season_winter: Searon_winter has the coefficient of '0.0901', which shows that there is good demand for bikes in winter.
4. Mnth_sept: September month has the coefficiennt of '0.0692', which is also the significant factor.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans-Before understanding the meaning of linear regression, we have to understand the meaning of Regression.

Regression: Regression helps us to determine the strength of the Relationship between one dependend Variable and one or more Independent variable.

Linear: Linear here doesn't mean a perfect straight line but it means linearity in the parameters.

Dependent
Variable: Dependent Variable is that Variable on which modal has to be build .Dependent variable is located at the y-axis and also known as the target variable.

Independent
Variable: Independent varibale is known as the Predictor variable. It is located on the x-axis and it helps in our modal prediction.

If we combine all the above 4 terminologies, we can get Linear regression.

Some important things to note:

1. Output variable that has to be predicted must be a numerical/continuos one.
2. Linear Regression is considered as the Supervised learning method because labels are present.

Formula: $Y= a + bX$ (Y=Dependent variable ,X=Independent variable, b=slope , a= constant)

Types:

1. Simple Linear Regression : Here predictor variable is one.

2. Multiple Linear Regression: Here predictor variable are more than one.

Some assumptions and their validation:

1. Linear Relationship: There must be Linear Relationship between the dependent variable and their predictors.
   Validation: For this we can make the Pair Plot for our DataFrame.

2. Error Terms are independent of each others
   Validation: For this we have to use Durbin – Watson (DW) statistic.

3. Absence of Multicollinearity
   Validation: Here we can find the correlation between the independent variable and them make heatmaps .
   After this we can also look for VIF (variance Inflation Factor) among the variables and eliminate those variables having more than 5 VIF.

4. Homoscendasticity:
   Validation: We can plot the residual plot vs the Fitted value plot. If the plot shows the funnel shape then there is no Homoscendasticity means it has Heterocendasticity.

5. Error terms are normally distributed:
   Validation: For this we have to plot quantile-quantile plot (q-q plot) , if the q-q plot shows straight line then its normally distributed else not.

This is the basics for Linear Regression Algorithm.

2. Explain the Anscombe's quartet in detail.

Ans- Anscombe's quaret was discoverd by Francis John "Frank" Anscombe in 1973. He has plotted 11 Datapoints on the graph and find some summary statistics about that dataset like mean , variance , correlation coefficient and best fit line and then he made 3 more datasets having exactly the same summary statistics ,what he plot the all three other datasets on the graph, he was completely shocked. Because all the four datsets have same summary statistics but have different graphs. So form this Theory, we have concluded that the graphs are very important if we want to draw any conclusion about any data. That is the reason why matplotlib is very much important to visualize the data.

3. What is Pearson's R?

Ans- Pearson's R is the correlation coefficient which is named after Karl Pearson.Inspired by Francis Galton he formulated this. The basic definition of Pearson's R is the covariance of 2 variables divided by product of their standard deviation. The value of the Pearson 's R lies between -1.0 and +1.0. Negative Pearson's R mean if the value of one variable increases ,the other variable value decreases proportionately. Positive Pearson's R mean if the value of one variable increases ,the other variable value increases proportionately.

Formula= r= NΣxy-(Σx)(Σy)/root([NΣx^2 - (Σx)^2] [NΣy^2 - (Σy)^2])

Where,

N = the number of pairs

Σxy = the sum of the products

Σx = the sum of x

Σy = the sum of y

Σx2 = the sum of squared x

Σy2 = the sum of squared y

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

Ans- Scaling= Scaling is a pre - processing technique which is used to scale the value of the data within specified range.

Why
Scaling= First, it should be clear that scaling is performed only on Independent Variables.Why are we scaling variables? The scaling is optional in case of simple linear regression while it must be compulsory for multiple linear regression modals. Suppose we have 2 columns like one is bedroom which has values between (1 to 6) and another column is area having values between (1000 to 5000). So after training the modal we get very high coefficients for bedrooms and very low coefficients for area columns because area value is high. So to cope up these kind of situation we can do scaling. So that every variables are treated equally.

Normalised
Scaling= Normalised scaling is also known as Min-Max scaling. Here scaling is done in such a way so that all the values of all the variables lie between 0 and 1. Here 0 is minimum and 1 is maximum.

Formula=(X-Xmin)/(Xmax-Xmin)

Standardized
Scaling= In this caling mean is 0 and sigma=1.

Fromula=(X-mu)/sigma

Difference= Here the major difference between two is that Normalised scaling handles outliers while Standardized one can't handle . So Normalised scaling is better than standardized one.

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

Ans-Here first we have to understand the meaning of VIF. Full form of VIF is Variance Inflation Factor. VIF is the method that shows the relationship of one independent varibale with other independent varibale. If for a particular variable , if VIF is high then it means it has high association with other varibale and vice versa.

Formula=1/(1-r^2)

r^2= r^2 shows the correlation .

Here it is given that VIF is infinite that means r^2 is 1 which means that one variable is perfectly Correlated with other variable.VIF generally shows multicollinearity.

Proof= 1/(1-1) = 1/0 = infinity

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

Ans-Q-Q plot also known as Quantile-Quantile plot.

Use= It is the graphical technique which is used to check whether two datasets that are coming from the same Population have same distribution or not.

Importance=The major imporatnce advantage of the Q-Q plot is to validate the assumptions of the multiple linear regression which is Error Terms should be normally distributed. And we can validate this with the help of Q-Q plot , if the plot shows straight line then it is normally distributed otherwise not.