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

Ans: From the analysis of the categorical variable it is identified that the dependent variable i.e bike rental count

- Is more for the year 2019
- When the weather situation is clear
- Is more during the fall of season
- Is more for the month of september

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

Ans: It is recommended to use drop_first=True so that the extra colum that is being created will be dropped an we will be having n-1 columns (where n is the no of level in a column).Also it helps in reducing correlation between the dummy columns created.It also reduces redundancy.

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

Ans: We can see that cnt has the highest correlation with the registered variable

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

Ans: We have validated the below assumptions for the model on training data set:

- Linear relationship: by plotting the model we built (lm5) with the independent variables and we can see that linearity is being preserved
- Checking for Homoscedasticity: we confirmed this by plotting residuals against the target variable where we saw that there was no pattern being followed by the residuals
- Checking for multicolinearity :we confirmed this with the help of VIF(variance_inflation_factor) where we maintained the thumb rule of keeping the VIF<5.
- Normalized mean error: we made sure the residual is normally distributed i.e mean=0.

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

Ans: based on the model built we can see that below feature with high significance and positive coefficients:

- Temp-> p value less than 0.05 and with highest positive coefficient
- Yr-> p value less than 0.05
- Winter Season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression(comes under supervised learning method) is a linear model that assumes a linear relationship between a dependent/target/y variable and independent/features/x variable such that y/target can be calculated/predicted from linear combination/s of input variables. For single input variable it is referred as 'simple linear regression' and for multiple input variable it is referred as 'multiple linear regression'.

In a Simple linear regression the model is given by the below formula for the line:

Y=b0+b1x where b0 is the constant , b1 represents the slope and x is the input variable

For multiple linear regression where there is more than one input variable the line is replaced by plane and it's formula is given as below:

Y=b0+b1X1+b2X2+....+bnXn

There are basically 4 assumptions which justify the use of linear regression model:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

The strength of the linear regression model can be determined using 2 metrics:

- R^2 or Coefficient of Determination
- Residual Standard Error (RSE)

The R^2 explains what portion of the given data variance is explained by the developed model. Higher the value of R^2 better the model fits our data.

$R^2 = 1 - (RSS / TSS)$

Because the Linear Regression is a vast algorithm and it will be difficult to cover all of it. We can improve the model in various ways could be by detecting co linearity, by transforming predictors to fit nonlinear relationships, dropping the feature which are high in co-linear or dropping the feature which is less significant etc.. The advantage and disadvantage of linear regression algorithm:

- Linear regression gives us a powerful statistical method to find/predict the relationship between variables. However, it's only limited to linear relationships between the variables.
- Linear regression are little sensitive to outliers and only it looks at the mean of the dependent variable,however it produces the best predictive accuracy for linear relationship.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe’s quartet comprises of four datasets which does have nearly identical simple statistical properties, yet appear different when graphed/plotted. Each particular dataset consists of eleven (x,y) points. It was constructed in 1973 by the statistician Francis Anscombe to demonstrate the importance of plotting the data before analyzing it and the what is the effect of outliers on statistical properties.

We have these four data set plots which consists of nearly same statistical observations, provides same statistical information that involves variance, and mean of all x,y points present in all four datasets.

It tells us about the significance of visualising the data before applying various algorithms that are out there to build models, out of them it suggests that which data features must be plotted in order to see the distribution of the samples that can help us identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc

Also, it is known to us that the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson’s R?

Ans:Pearson’s correlation coefficient is the test statistics which helps in determining the statistical relationship, or how it is associated, between two continuous variables. This is the best method to measure the association between variables which are of interest because it is based on the method of covariance. It tells us about the magnitude of the association, or correlation, as well as the direction of the relationship.

Coefficient ranges from +1 to -1 where +1 reflects positive relationship(target variable increases with the increase in independent variable),-1 reflects negative relationship of between independent and target variable and 0 reflects no relationship at all.

We should always consider the absolute value for correlationship Higher the absolute value stronger the relationship between the variables.

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

Ans: When there are lot of independent variables in the dataset then the coefficients derived from the model can be very weird or strange as the values of these independent variables can be on very different scales or range from each other which might not help with interpretation. Hence, scaling is the process of bringing all the values to a common scale which will help us easily interpreting the model.

Standardized scaling is a process where the mean of the variable would be 0 and deviation would be 1 whereas the normalized scaling would make the variable range lie between 0 and 1 i.e the max would be 1 and min would be 0.

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

Ans: An infinite VIF value indicates that the particular feature can be strongly represented by an other feature or the corresponding feature is expressed exactly by a linear combination of other variables (which show an infinite VIF as well). When there is perfect correlation,
then VIF = infinity. In case of perfect colinearity the R2 value is 1 which lead to VIF= (1/1-r2) to infinity. To resolve this problem we need to drop one of the columns from the dataset which is causing this multicolinearity.

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

Ans: Q or quantile-plot is a scatter plot which helps us to validate the assumption of normal distribution in a given data set. With the help of this plot we can infer if the data comes from a normal distribution. If the data is normally distributed, the plot would show us fairly straight line whereas no normality in the errors can be seen with deviation in the straight line.

'Quantile', can be defined as points in our data below which a certain proportion of our data falls. Quantile can also be referred to as percentiles. For example: when it is being said that the value of 50th percentile is 130, then it means that half of the data lies below 130.