General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a method of finding the best straight-line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.
This model is to find a linear relationship between the input variable(s) X and the single output variable y.

- ***Simple linear regression***: When there is only single independent/feature variable X then it is called as simple linear regression.
- ***Multiple linear regression***: When there are multiple independent/feature variables Xi then it is called as Multiple linear regression.
  The independent variable is also known as the predictor variable.
  The dependent variables are also known as the output variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

3. What is Pearson's R?

- ***Pearson's r***, or the bivariate correlation, is a statistic that measures linear correlation between two variables X and Y.
- It has a value between +1 and −1.
- A value of +1 is total positive linear correlation,

- A value 0 is no linear correlation
- A value −1 is total negative linear correlation.

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

Scaling is a technique to standardize the independent features present in the data in a fixed range.
It is performed during the data pre-processing to handle highly varying magnitudes or values or units. So, we use Feature Scaling to bring all values to same magnitudes.

- ***Normalization*** usually means to scale a variable to have a values between 0 and 1, while ***standardization***

transforms data to have a mean of zero and a standard deviation of 1.

- Standardisation over min max is that it doesn't compress the data between a particular range as in Min-Max

scaling.

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

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution..

The **purpose** of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if

the two data sets come from a common distribution, the points will fall on that reference line.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Importance:

a)It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

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?

- ***Year***: A coefficient value of '0.2279' indicated that a unit increase in year variable, increases the bike hire numbers by 0.2279 units. **Dependent variable('count') is positively correlated with year.** , people hire more bikes in the month of January with respect to other months.
- ***Light_rainbow***: A coefficient value of '0.2331' indicated that, a unit increase in light_rainbow variable, decreases the bike hire numbers by 0.2331 units. **Dependent variable('count') is positively correlated with year.,** people tend to hire bikes in light rainbow weather.
- ***Misty***: A coefficient value of ' 0.0517' indicated that, a unit increase in light_rainbow variable, decreases the bike hire numbers by 0.0517 units. **Dependent variable('count') is negatively correlated with year.**
- ***Winter***: A coefficient value of '0.1357' indicated that, a unit increase in winter variable increases the bike hire numbers by

0.1357 units. **Dependent variable('count') is positively correlated with year.,** people tend to hire bikes in winter season.

- ***Holiday***: A coefficient value of '0.0994' indicated that, a unit increase in holiday variable decreases the bike hire numbers by 0.0994 units. **Dependent variable('count') is negatively correlated with year. ,** people do not tend to hire bikes on holidays or weekends.

- ***Summer***: A coefficient value of '0.0803' indicated that, a unit increase in summer variable decreases the bike hire numbers by 0.0803 units. **Dependent variable('count') is negatively correlated with year., .,** people do not tend to hire bikes in summer season.

- ***Sep*** : A coefficient value of '0.0959' indicated that , a unit increase in sep variable increases the bike hire numbers by 0.0959 units. **Dependent variable('count') is positively correlated with year. .,** people tend to hire bikes in the month of September.

- ***July***: A coefficient value of '0.0481' indicated that, a unit increase in july variable decreases the bike hire numbers by 0.0481 units. **Dependent variable('count') is negatively correlated with year.** people do not tend to hire bikes in the month of July.

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

drop_first=True is important to use, it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

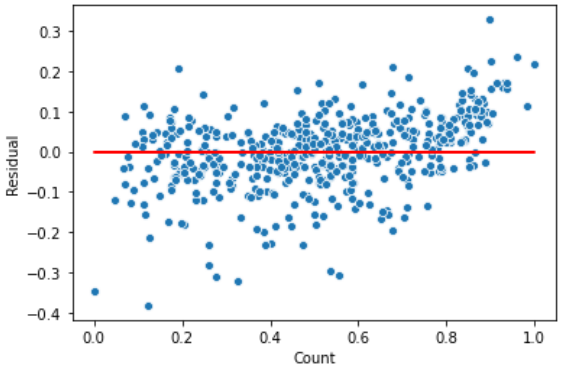
'TEMP' variable has the highest correlation with the target variable ('count').

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

Assumptions of linear regression:

- ***Linear relationship***
  Plotted dependent variable (***"count"***) against each explanatory variables ***("temperature", "windspeed", "humidity"*** and visually inspect the scatter plot for signs of non-linearity.

- ***Homoscedasticity***
  The residual errors should have constant variance.



Plotted scatter plot with target variable ("count") on x-axis and error term.

And as shown in the figure residual error turned out to have constant variance.

- ***Checking for Multi collinearity***
  multicollinearity is when two or more independent variables in a regression are highly related to one another, such that they do not provide unique or independent information to the regression.
  1.plotted heat map between the categorical variable and was observed no multi collinearity between the variables.
  2.Two variables temp and atemp were correlated , so atemp was dropped.
- ***Normality of error***
  Plotted histogram for the error term, to check the distribution of error term and it turned out to be normally distributed.

- ***Independence of residuals***
  To test the independency of residuals, DURBIN WATSON test is used.

The Durbin-Watson value for Final Model is 2.0689.
Hence,there is almost no correlation.

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

Top 3 variables contributing significantly towards explaining the de mand of the shared bikes are:

- ***Temperature (temp)*** - A coefficient value of '0.5978' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5978 units.
- ***Light rainsnow (weathersit_3)*** - A coefficient value of '-0.2331' indicated that, a unit increase in Light rainsnow variable decreases the bike hire numbers by 0.2331 units.
- ***Year*** - A coefficient value of '0.2279' indicated that a unit increase in yearr variable increases the bike hire numbers by 0.2279 units.