Assignment-based Subjective Questions

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

Ans: Based on the analysis done on the categorical columns, below are variables which are making significant impact on the data set:

- Bikes are in highest demand in fall season and least in demand in spring season.
- The market growth has increased by almost 80 – 85% from year 2018 to 2019.
- Least demand is noticed in the months of January and February, due to weather conditions.
- September month looks the most demanding month for bike booking.
- Rain and snow falls are impacting the bikes booking badly.
- Compared to other days, Saturday's the demand is high.
- Working and non-working days are not impacting much, it is almost the same in both the cases.

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

Ans: Dummy variables are used to convert the categorical variables into a numeric variable in the form of binary data. Drop_first will create one lesser number of levels from the unique levels present in the categorical variable, which can be easily understood by the combination of other generated dummy variables for that categorical column. In simple words, if the dummy variable value is zero in all the dummy variable creations, then by default it will fall to the first level.

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

Ans: 'temp' variable which stands for temperature in Celsius shows the highest correlation with 'cnt' (count of total rental bikes).

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

Ans: Validated these assumptions on the training dataset:

- Multicollinearity check
- Independence of variables
- Normal error term distribution
- Linear relationship validation

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

Ans: These are the top 3 features contributing to significantly towards explaining the demand of the shared bikes:

- temp
- year
- winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a supervised machine learning algorithm which finds the relationship between a dependent and on one or more independent variables. If the dependent variable is one, then it is called univariate linear regression, else called multivariate linear regression. Equation of a straight line is y = mx+ c, where y – how far up, x – how far along, m – gradient and c – value of y when x= 0.

It helps in understanding and predicting the behaviors of variables. These are the assumptions to be followed to get the accurate results from this algorithm:

- Independence of dependent variables.
- Linearity between the dependent and independent variables.
- Normal distribution of the residuals.
- No multicollinearity: There is no high correlation between the independent variables.
- Homoscedasticity: Across all levels of the independent variables, the variance of the errors is constant.

Types of Linear regression:

- Simple linear regression - It involves only one independent variable and one dependent variable.
- Multiple Linear Regression – It involves more than one independent variable and one dependent variable.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, these data sets look very different from one another in the plot. This suggests the data features must be plotted to see the distribution of the samples that can help us identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.) before building the model.

Moreover, linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set. Below are the Anscombe's quartet four data sets:

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

Ans: The Pearson's R displays the Linear relationship between the two datasets. It normalizes measurement of the covariance, such that the result always has a value between −1 and 1.

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

Ans: Scaling is the process of converting the units of different independent variables into the same level or unit. It is important to do because different scales will lead to models with very weird coefficients that might be difficult to interpret, and it fastens the convergence for gradient descent methods. Normalization rescales the values into a range of 0 and 1, while standardization shifts the distribution to have 0 as mean and 1 as a standard deviation.

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

Ans: When there is perfect correlation between the independent variables that gives the R2 values a 1, which leads to VIF value to infinite. This is due to the multicollinearity issue.

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

Ans: The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.

Advantages of Q-Q plot:

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, we don't need to care about the dimensions of values.