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:- Below are the impact of the categorical values on the dependent variables:

1. In Year 2019 ,the total number of rented bikes are more in number compared to year 2018
2. The end of summer and the start of autumn (fall) season the number of rented bikes is high.
3. On days where the skies are clear or partly cloudy the rentals of bikes are more.
4. On working days(Monday to Friday), the number of bike rentals is high compared to weekends and holidays.

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

Ans:- If the number of categories for a variable is n , then drop_first =True will create n-1 dummies. Hence we will be reducing the number of independent variables by 1 from the dummies created this making the model creation process less complex.

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

Ans:- Feeling Temperature/Actual Temperature (atemp,temp) have 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:- Below are the steps taken to verify the assumptions of Linear Regression:

1. Distribution plot of the error terms is plotted. It is verified that the mean is centered around 0 and the plot follows normal distribution.
2. The plot of residuals and the y actual is plotted. It is verified that the error terms are constant for various values of dependent variable. Homoscedasticity is ensured
3. By plotting the pairplot of various numerical independent variable and also the correlation between the various independent variables. Variables which are highly correlated are not considered for model building process.

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

Ans:- Below are the 3 top features:

1. Feeling temperature increases the demand of shared bikes.
2. Greater the humidity , lesser is the demand of bike rentals.
3. The year also plays a significant role in determining the demand of rented bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:- Linear regression is method of finding the best fit line passing through a given set of independent variables such that sum of squares of the errors is minimum. Error is defined as the difference of the predicted dependent variable and the actual value of the dependent variable.

The general equation of the best fit line is given by the equation $ypred=\beta_0+ \beta_1x_0+ \beta_2x_1+ \beta_3x_2+...+ \beta_nx_{n-1}$ where

β_0 is the constant/offset term
$x_0, x_1,.. x_{n-1}$ are the independent terms
$\beta_0, \beta_1,.. \beta_n$ are the coefficient of the independent terms. They specify how a unit increase of the independent variable will cause the dependent variable to change given that other terms are constant.

The algorithm of linear regression is based on reducing the cost function which is given by minimise $J=(1/n)*SUM(ypred-yi)^2$ where n is the number of data points. Ypred is the predicted value obtained by the assumed best fit line. The gradient decent technique is used to find the local minima of the cost function.

2. Explain the Anscombe's quartet in detail.

Ans:- Anscombe's dataset consists of 4 datasets with each dataset having 11 sets of x and y values. The unique feature of these datasets is that all the 4 datasets have similar statistical characteristics. Properties which are same for all the datasets are mean(x), mean(y), variance(x),variance(y), correlation of x and y, linear regression line equation and R square value. When all the datasets are plotted separately using scatter plot , it explains how the distribution of x,y points can impact the overall regression line even though there may not be any direct relation between the x and y points in the given dataset.

Basically it explains like how the outlier in the data can impact the regression line or how multiple sets of x and y points which are distributed differently can yield the same regression line.

3. What is Pearson's R?

Ans:- Pearson's R or Pearson correlation coefficient is a numerical value which explain how strongly 2 variables are related to each other (is simple words how change in one variable impact the other). The value is between -1 and +1. A negative value implies that the when one of the variable increases, the other will reduce and vice versa. Similarly, a positive value implies that when one variable increases, the other also increases. A value near to 0 means, no correlation exists between the 2 variables. It is obtained by the square root of R^2 value of the linear regression model explaining the 2 variables.

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

Ans:- Scaling is a data pre processing technique which is used to convert the data such that it fits within a particular range. It is performed when we have independent variables whose magnitude ranges differs from each other.

Scaling is performed because of the following reasons:

1. Scaled data speeds up the algorithm and thus computation time for finding the converged values.
2. The coefficients of the independent variable which are scaled are easy to compare with one another and thus analysing the impact of the independent variable becomes easier and more meaningful.

The differences between normalization and standardization scaling are as follows:

SL No	Normalization Scaling	Standardization Scaling
1	The output of this scaling is a value falling between 0 and 1	The output of this scaling is the Z score associated with the variable such that the mean value is 0 and standard deviation is unity.
2	The formula is x=(x-(min x))/(max x -min x)	The formula is x= (x-mean x)/(Standard Deviation of x)
3	Can be used on data which don't have outliers. Presence of outliers will cause loss of information with Normalization scaling	Can be used on data which may or may not have outliers.
4	The output of this scaling will be clustered as the values are limited between 0 and 1 thus useful for finding the insights from the data.	The data is not clustered but will be scattered. Getting insights from the data tend to be difficult.

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

Ans:- The VIF for a independent variable is calculated using the formula $VIF_1=1/(1-(R_1)^2)$ where R_1 is the value obtained by fitting a model between variable X_1 and all the other independent variable. When the value of 1 is obtained for the R^2 score of the model (the fitted lines explains all the variance of X_1 based on the other independent variables) then we get a value of infinite.

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

Ans:- A Q-Q plot is a plot of the quantiles of first dataset to the second dataset. Below are the importance and uses of Q-Q plot in Linear regression:

1. It can be used to compare if the 2 samples (train and test set) are coming from the same population having same distribution.
2. It can be used to identify the scaling of the 2 datasets.
3. Identify the distribution a given dataset follows. It can be Normal, Uniform or exponential distribution.
4. Identify if the error terms are normally distributed by plotting the quantiles of the residuals with the theoretical normal distribution quantiles thus helps validating the linear regression assumption.