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?

   From analysis it was observed that temperature is highly correlated with bike sharing demand.
   People are more likely to opt bike sharing when the weather is sunny or misty while avoid it on rainy or snowy days.
   On holidays, the count is less.
   And the count is increasing with every year.

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

   When we create dummy variables representing n categorical variables, we can assume that there can be at max n number of those variables available to define that category. So, if we don't consider one of those variables, we can still understand it's property using rest of n-1 variables which helps us keep the calculation simpler. This is why it is important to use drop_first = True during dummy variable creation which gives us n-1 variables and we can use the same to get the features of dropped variable.

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

   Temp 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?

   After building the model on the training set, I performed residual analysis to check linearity, homoscedasticity and normality. If the difference between predicted and actual data is not much, the distplot() shows maximum steep around 0 and we can say the model is correct.

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

   Season, Windspeed and weathersit are top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression algorithm is a supervised algorithm which assumes there is a linear relation between a continuous (or dependent) variable and one or more independent variables.

   It follows y = mx + c to show the relationship between the variables where m and c are coefficients.
   y = dependent variable
   x = independent variable
   If m = 0, then x and y are independent. If m > 0, then there will be positive relationship between x and y. If m < 0 then the relationship between x and y will be negative.

   With the help of Linear regression model we can find solution of many continuous models such as employee attrition report where attrition count is dependent variable and the factors influencing it are independent variables such as, employee benefits, years worked in company, hikes received etc. scikit learn and statsmodel are used to create Linear regression model.

2. Explain the Anscombe's quartet in detail.

   Anscombe's quartet is a group of four datasets which has almost similar summaries i.e. mean, co-effs etc. but when plotted, it shows how each dataset is different from others.

   Anscombe's quartet is a proof that says summary statistics can be misleading unless we visualise the data.

3. What is Pearson's R?

   Pearson's R is a measure of linear correlation between two variables which can range from 1 to -1 where 1 represents a perfect positive correlation and -1 represents a perfect negative correlation. 0 indicates no correlation between the variables.
   The relationship between the variables should be linear and Pearson's R is sensitive to outliers.

Eg. To build any linear regression model, such as finding relation between height and weight of a group of students, Pearson's correlation coefficient can be used.

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

   Scaling is the process in regression which helps to bring all the variables within same range such as the mean of predictors becomes 0.

   Scaling is performed to make the linear regression model more interpretable, and improved. When all the datasets are in same range it is easier for the model to learn the data and it's robust to outliers.

   Normalized scaling is used to transform variables on a similar scale. Usually the scale range is [0,1] or [-1, 1]. The formula to calculate Normalization or Min-max scaling is:
   X_var = (X – X_min)/(X_max – X_min)
   Where, X_var is the new output representing the feature, X and X_min and X_max are the minimum and maximum value of that particular variable present in dataset.

   Whereas Standardized scaling is when the data is transformed such as the mean turns 0 and standard deviation is 1 of all the data of a particular variable in the dataset.
   X_var = (X – X_mean) / X_std

   Normalization is sensitive to outliers and it preserves the shape of data distribution, mean and standard deviation of a feature, unlike Standarddization.

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

   When two or more independent variables are perfectly correlated with each other, then the value of VIF becomes infinite as one variable completely depends on the other.

   For example, in a model to find relation between height and weight of a group of students, we assume there are two independent variables mentioned as 'height in inch' and 'height in cm'. Both these variables will be correlated to each other which will result into infinite VIF.

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

   A Q-Q plot, also known as Quantile-Quantile plot is a graphical representation of two probability distribution which determines it follows normal or exponential or uniform distribution. The quantiles of a distribution that divides the distribution in equal parts.
   For example, 25^{th} quantile divides a value such that 25% of it's distribution is less than or equal to 25^{th} quantile.

   When training and test data is received separately, we can use Q-Q plot to determine if both the sets are from populations of same distribution.
   Shifts in scale, changes in symmetry, presence of outliers can be detected from this plot.