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? (3 marks)

   Some of the categorical variables are significant.
   Some months have high demand for Bikes such as September.
   Some Seasons have high demand for Bikes such as Winter.
   Bad Weather(Weather Type 3) has negative effect on demand of Bikes .

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

   drop_first=True reduces the number of independent variables which needed to be dealt with while building the Linear Regression Model. drop_first=True would not change the resulting linear model. First column can still be explained by other Columns. So, it is good to drop the first column while creating dummy variables .

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   tmp and atemp have the highest correlation with the target variable cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   We validate the assumptions of Linear Regression by drawing the histogram of error terms.
   And we check that whether drawn histogram follows a normal distribution(Gaussian distribution) or not.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   The top 3 features contributing significantly towards explaining the demand of the shared bikes are temperature, year, and weather .

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Linear regression algorithm is basically focused on developing a best possible linear equation which can explain the relationship between the dependent variable and independent variables .
   In this algorithm , we study the relationship between dependent variable and independent variables and select the most important variables which influence the target variable positively or negatively .

Best fit line is drawn based on where the sum of the squared differences between y_predicted and y_actual is least .

We perform following steps in Sequence.

1) Data cleaning steps such as removing invalid values or filling invalid values with appropriate values , removing unwanted columns , removing columns having same value in all rows, removing rows which have very little data.
2) Examine the outliers and then either remove the outliers or replace outliers with appropriate values.
3) After performing cleanup , we introduce dummy variables .
4) We study correlation between dependent variables and independent variables.
5) Now, we divide the dataset into training and test data set .
6) We need to scale some values as some column values may be either too small or too big.
7) After scaling , we build the linear model using training dataset . We use the recursive feature elimination.
8) We Check the model stats such as R-squared and Adjusted R-squared.
9) Remove insignificant variables with p values less than 0.05.
10) Check VIF for multicollinearity. Exclude variables with VIF less than 7. The choice of VIF cutoff can vary.
11) Rebuild the model with refined set of features.
12) Check whether error terms between predicted y_train and actual y_train is a normal distribution.
13) Apply Linear model on Test Data Set.
14) Plot the spread between predicted y_test and actual y_test to check how well is the linear model working on test data set.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet consists of 4 datasets with similar statistics parameters such as mean, variance and correlation but differ vastly when drawn on the graph. This quartet highlights the importance of visualizing data. One dataset when drawn on the graph may be linear while the other one having similar statistics may have entirely different non-linear curve.

3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient measures the strength of linear relationship between 2 variables . Its value can change from -1 to +1 . +1 indicates a perfect positive linear Relationship whereas -1 indicates a perfect negative relationship and 0 means there is no linear relationship .

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

When we are dealing with multiple independent variables , then their values may be on different scales, one variable may have values in millions whereas other variable may have values less than 1 . In that case , coefficient values will highly differ . Therefore, to bring the coefficient values of variables to comparable values , we need to perform scaling of dataset.

Normalized scaling is MinMax Scaling Technique. We scale the values of variable using the below formula in case of normalized scaling.
(X – Xmin)/(Xmax-Xmin) where X is actual value of variable.
It is must to use normalization technique only after addressing outliers.

In the case of Standardized Scaling, we use the formula
(X-Xmean)/Standard deviation where X is actual value of variable.

While it's advisable to use Standardization after handling outliers but the repercussions of not addressing outliers in case of Standardization are far lesser as compared to normalized scaling.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   Value of VIF is infinite when there is perfect collinearity between 2 independent variables . Existence of perfect collinearity between 2 independent variables means that one independent variable can accurately determine the value of other variable or you can also say that there is correlation of 1 or -1 between 2 independent variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

   Q-Q plot is a Quantile-Quantile plot . It is drawn between quantiles of 2 data sets . Quantiles of one data set is drawn on x axis and quantiles of other data set is drawn on y axis. Q-Q plot is often drawn between below quantiles while developing a Linear regression model.

   1) Target variable on Test data set and Target variable on training dataset.
   2) Target variable on Test data set and predicted values of Target variable by linear model.
   3) Target variable on Training data set and predicted values of Target variable by linear model.

   Resultant curve is often a curve wrapped around y=x straight line which is at 45 degree to x axis. Such a kind of resultant curve proves that the 2 dataset follow the same distribution . When the graph is drawn between values derived from linear model and actual values, a resultant curve wrapped around y=x straight line proves that the linear model developed is a good model.

   Q-Q plot can also be drawn between residuals or error terms of 2 data sets . In this case , Quantiles of error terms of one data set is drawn on x axis and quantiles of error terms of other data set is drawn on y axis.