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)

   Answer:
   The demand for bike is as follows:
   Seasons: Spring < Winter < Summer < Fall (Spring has the least and Fall and Summer have high)
   Year: 2018 < 2019 (Increased from 2018 to 2019)
   Month: July has the highest and January has the least
   Weather Situation: The demand is least during rain, snow and thunderstorm

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

   Answer:
   'drop_first=True' is important to use, because it reduces the extra column created during dummy variable creation. This will reduce 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? (1 mark)

   Answer:
   'temp' and 'atemp' have 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? (3 marks)

   Answer:
   The assumptions of linear regression:
   1. Linear relationship between the feature and target: This can be observed from the final result.
   2. Less multicollinearity between the features: Less correla
   3. Homoscedasticity assumption: Residual values is same across all independent variables, confirmed by drawing the relevant plot.
   4. Normal Distribution of error terms: This was confirmed by drawing the plot of error terms.
   5. Less Autocorrelation in the residuals: This was also confirmed by checking the plot of residuals vs the individual variables.

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

   Answer:
   Top 3 features contributing significantly towards explaining the demand of the shared bikes are:
   1. The year (whether 2018 or 2019)
   2. Holiday (whether the given day is a holiday or not)

3. Spring (whether the given season is spring or not)

General Subjective Questions

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

   Answer:
   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Based on individual variables, it tries to predict a linear relation to the dependent variable, which has to be predicted.

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

   Answer:
   Anscombe’s Quartet is a set of 4 data sets whose simple descriptive statistics, when compared, appears the same, but upon plotting them, we realise the distributions are entirely different. Some peculiarities in the dataset fool the regression model if built.

   It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building. The four data set plots which have nearly same statistical observations including variance, and mean of all x,y points in all four datasets.

3. What is Pearson’s R? (3 marks)

   Answer:
   It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. It lies between −1 and 1. It takes into account linear correlation between the variables in question.

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

   Answer:
   It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

   Usually, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not the units thereby resulting in incorrect modelling. To solve this issue, scaling has to be done to bring all the variables to the same level of magnitude.

   Normalization/Min-Max Scaling:

   It brings all of the data in the range of 0 and 1.

   X = [X – min(X)]/[max(X) – min(X)]

Standardization Scaling:

Standardization Scaling is done by replacing the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

X = [X – mean(X)]/sd(X)

The disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer:
If there is perfect correlation, then VIF = infinity. This means a perfect correlation between two independent variables. In the case of perfect correlation, we get R^2 =1, which means 1/(1-R^2) is infinity. One of the variables from the dataset which is causing this perfect multicollinearity needs to be dropped to prevent this.

An infinite VIF value means that the corresponding variable can be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:
Quantile-Quantile (Q-Q) plot, is a graphical tool to aid us in assessing whether a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to predict if two data sets come from populations with a common distribution.

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.

Uses:
a) It can be used with sample sizes
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.

It is used to check following scenarios:
If two data sets —
1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior