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)

Ans: Following points are noted based on my analysis

- We can see that upto 78% of variance can be defined accurately using our model.
- We see 'weekday' or 'workingday' has no significant impact on data as we predicted previously.
- Dependent Variable is highly correlated to 'yr' i.e. our industry is booming and there is high demand each year.
- Temperature plays another major factor in there is high increase in sales when temperature is favorable.
- There is high demand in bike towards the spring season, this is a season you can capitalize highest profits.

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

Ans: We do this in order to reduce number of dummy variables created/used in our model. As we have n levels of categorical data we can manage with (n-1) dummy variables. Our model will perceive that one removed variable when data at rest of the variables is 0.

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

Ans: We see that the variable temp has highest Correlation with our target variable.

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

Ans: I validated this by calculating the Residuals between predicted and actual target variable data. Post that I made as Distplot to validate below 3 assumptions.

- Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.
- Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
- Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.
- Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ2 .

To check Linearity Assumption we plotted pair plots to see if some variance is explained.
For 3rd Assumption I checked VIF of independent variables and made sure they are all independent of each other.

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

Ans: According to my model the top 3 features contributing toward the demand of shared bikes are 'yr', 'temp', 'spring'. Together these 3 describe upto 73% of total variance.

General Subjective Questions

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

Ans: In simple terms, linear regression is a method of finding the best straight line fitting to

the given data, i.e. finding the best linear relationship between the independent and dependent variables.
In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

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

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

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

Ans: In statistics, the Pearson correlation coefficient (PCC, pronounced /ˈpɪərsən/) — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data.

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

Ans: Scaling 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.
Most of the times, 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 units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

- Normalization brings all of the data in the range of 0 and 1.
- Standardization replaces 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).

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

Ans: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may 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)

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine 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.