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:
The demad for ikes is less in the month of spring when compared to other seasons
The demand of bike increased in the year 1 than the previous year.
Jun to Sep is the peak demand. Lowest demanding month is Jan.
Demand of bike is almost similar all the weekdays.
he bike demand is high when weather is good and Few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any data when rain/snow so can not conclude anything.
Bike demand is less in holiday seasons

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces 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)
Ans: Temp and atemp has the highest correlation

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans:
One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on histogram

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 are temp, year and weather conditions

General Subjective Questions

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

An interpolation technique is used to predict the correlation between variables and how independent variables are influenced by the dependent variable(s), is linear regression. After looking into the data and cleaning it using EDA, we had split dataset into training set and the testing set. After checking the collinearity of variables and using the requisite variables to train the model and checking the R-value of the model and the p-values of dependent variables, after dealing/dropping the necessary columns and reiterating the steps (feature elimination), we come to a final model. According to the conditions of linear regression which states that the error curve must be a normal one, we proceed to testing the model with the test dataset. The conclusion hence drawn on the model would be used to provide valuable insights/predictions on datapoints in the range of the model.

2. Explain the Anscombe's quartet in detail. (3 marks)
A regression model is not always necessarily a perfect one, it can also be mistaken by some data. In some cases, there are multiple datasets which are completely different but after training, the regression model looks the same. A group of four such datasets having identical descriptive statistics but with some peculiarities, is the Anscombe's quartet.

3. What is Pearson's R? (3 marks)
In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
ANS:
Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a pre-set range. Typically used in Neural networks broadly.
Standardized scaling: The example given above is of standardized scaling. Here, the values of variable(s) is/are compressed into a specific range to suit the model.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
ANS: If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Ans: The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.