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 categorical variables has effect on the dependent variables

   - Yr: the dependent variable increase with year
   - Season: Bike demand decrease in winter and spring
   - weekday： Bike demand decreases on Sundays, Mondays
   - Weather: Bike demand decreases when weather is Snowy or misty
   - Months : Bike demand increases in month 6,8,9 and 10

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

   Ans: when we use this option, the first column is dropped, and number of dummy variables is n-1, where n is number of possible values in a original categorical variable

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 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)

   Ans: We did residual analysis and confirmed the error term is normally distributed and mean is around zero.

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: The top 3 features are:

   - Light snow: -0.32
   - Yr : 0.25
   - Month 9: 0.12

General Subjective Questions

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

   Ans: Linear regression is the way of predicting a dependent variable values based on 1 or a group of independent variables.

   e.g. Y = nX + m (this is the equation fo straight line, where Y is dependent variable, X is dependent variable. This is simple linear regression)

   Y = n1 * X1 +n2 * X2 + n3 * X3… This is called multiple linear regression, where multiple independent variables(X1, X2, etc) have impact on dependent variable (Y)

Now, to build a model that can predict the outcome, we use a method called “Finding the Least Squares”. One of the ways to find effectiveness of the model is measuring the r-square, which is calculated as 1 – (RSS/TSS)

RSS = Residual Sum of Sqares

TSS = Total sum of squares

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

   Ans: Anscombe's quartet is a group of 4 datasets, that have almost the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

3. What is Pearson’s R? (3 marks)

   Ans: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

   If it is between 0 to 1, then it is called positive correlation- means, if the independent variable increases, dependent variable value also increases.

   If it is 0 -> then there is no relation

   If id is between 0 to -1, then called -ve coefficient, meaning, any increase in dependent variable would cause a decrease in dependent variable and vice versa

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

   Ans: Scaling means transforming your data so that it fits within a specific scale.

   This is to ensure that no single feature dominates the distance calculations in an algorithm, and can help to improve the performance of the algorithm.

   Normalized scaling scales the data between 0 and 1, whereas standardized scaling makes mean as zero and standard deviation 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)
Ans: VIF is calculated as follows
VIF = 1/(1-R squared).
If R squared value is 1, then VIF value is infinite. And R square would be 1, if there is a perfect correlation.

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

(3 marks)

And: A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.