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 sale of bikes is less in the month of spring when compared to other seasons.
   - The sale of bikes increased in the year 2019 when compared to year 2018.
   - The months from Jun to Sep is the period when bike demand is high.
   - January has the lowest sales of all the months.
   - Sales are less in holidays in comparison to holidays.
   - The demand of bikes is almost the same during the weekdays.
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 which helps to reduce the complexity . It also 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: 'atemp'(feeling temperature in Celsius) has the highest correlation with the target variable as per the pairplots
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Ans:
   - By plotting the residual terms we can see that it follows a normal distribution
   - When we have time series data (e.g. yearly data), then the regression is likely to suffer from autocorrelation because demand next year will certainly be dependent on demand this year. Hence, error terms in different observations will surely be correlated with 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)
   - Top 3 features are: year, season_spring and weather_sit_bad

General Subjective Questions

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

   Ans: The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.
   The strength of the linear regression model can be assessed using 2 metrics:
   1. R^2 or Coefficient of Determination

2. Residual Standard Error (RSE)

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

   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. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
3. What is Pearson's R? (3 marks)

   Ans: The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit). The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
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**. 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 Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q 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. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.