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: As per the result from the linear regression, below are some categories which are **negatively correlated** to dependent variables

   - Holiday
   - weathersit_2
   - weathersit_3

   And below are the categories which are **positively correlated**

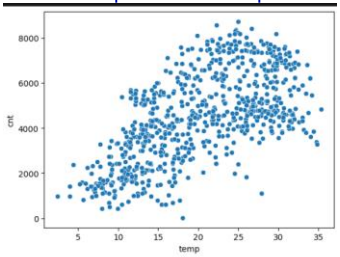
   - yr
   - workingday
   - season_2
   - season_4
   - mnth_8
   - mnth_9
   - mnth_10
   - weekday_6

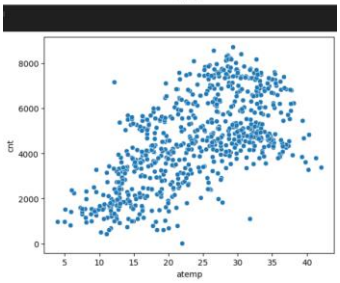
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   Ans : When we create dummy variable, It drops the 1st dummy after scaling a variable, as we don't need all the variables. It ensures that the model remains properly identified and interpretable by setting n-1 category as a reference and estimating the effects of the dropped category relative to it.

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: By comparing the R2 value of the trained model with the test model whether they are close enough to conclude the assumptions made.

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:

   Top 3 features contributing significantly
   Positive correlation
   1. Temp – The temperature.
   2. Yr - The year
   Negative Correlation
   3. weathersit_3 (light snow, Light rain etc)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans : Imagine you want to understand how the car speed varies with car weight. With the data given, you can do the scatter plot and see that as the weight of the car decreases the speed of the car increases. Linear regression is like drawing a straight line through all the points on a graph for this relationship.

   Now with the given data we have, we now must predict the speed of the car for a given weight.

   Hence linear regression is like finding the best straight line to make good guesses based on what data we already have.

2. Explain the Anscombe's quartet in detail. (3 marks)
   Ans: Anscombe's quartet is a set of four datasets created in 1973 by statistician Francis Anscombe to illustrate the importance of graphing data before analysing it. The datasets have similar statistical observations but look very different when graphed.
   The datasets have the same mean, standard deviation, and regression line, but have different distributions and appear qualitatively different. This shows how summary metrics can be misleading and that you need to visualize your data to avoid missing important trends. Anscombe intended the quartet to counter the idea that "numerical calculations are exact, but graphs are rough"

3. What is Pearson's R? (3 marks)
   Ans: Pearson's R also known as Pearson correlation coefficient, is a statistical model to measure the strength and direction of linear relationship between two continues variables.
   The value nearer to -1 denotes negative correlation and the value closer to 1 denotes positive correlation.
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Ans: Scaling is required to adjust the variables in to a range which is comparable. Certain times a variable has large values compared to others and when they get fit in to the equation of linear regression, they look to be contributing more percentage in the prediction. Hence its is necessary to scale all of them in to a comparable state.

   Normalized scaling is also called as Min-Max scaling and the scaled data of a variable always lies between [0,1].

When to Use:

- When you need all features to lie within a specific range.
- Particularly useful when the data does not follow a Gaussian distribution or when the scale of data is important (e.g., image processing).

Standardization scales the features so that they have the properties of a standard normal distribution with a mean of 0 and a standard deviation of 1

When to Use:

- When features have different means and variances.
- Standardization is more robust for algorithms that assume normally distributed data, such as linear regression.

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

   Ans : VIF becomes infinite when there is **perfect multicollinearity** in the dataset. This situation occurs when one predictor variable is an exact linear combination of one or more other predictor variables. In other words, if you can perfectly predict one variable using others, the VIF for that variable will be infinite. It can happen if a variable is exact duplicate of another variable too. Also, if we fail to drop_first=true while creating a dummy variable VIF becomes infinite.

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

   Ans: After fitting the model if we need to check whether the residuals are normally distributed. Normality of residuals are key assumption in linear regression. A Q-Q plot can help diagnose whether this assumption is met. We can get the residuals after fitting the model and can use statmodel for qqplot.