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)

The categorical variables season, weathersit and months. The output variable cnt is highly dependent on these.

In [ ]:
```
2. Why is it important to use drop_first=True during dummy variable creation? (2 ma

drop_first=True  ensures that a redundant  column is not created, thereby reducing
```

In [ ]:
```
3. Looking at the pair-plot among the numerical variables, which one has the highes
with the target variable? (1 mark)
- atemp
```

In [ ]:
```
4. How did you validate the assumptions of Linear Regression after building the mod
training set? (3 marks)
  If residual errors have a mean value of zero on a plot of residual vs fitted, tha
```

In [ ]:
```
5. Based on the final model, which are the top 3 features contributing significantl
explaining the demand of the shared bikes? (2 marks)
- Year 2019 saw an increase in demand
 - Also clear and misty weathers contributed towards driving demand
```

In [ ]:
```
General Subjective Questions
1. Explain the linear regression algorithm in detail. (4 marks)
Linear regression is an algorithm that provides a linear relationship between an in
variable to predict the outcome of future events.
```

In [ ]:
```
2. Explain the Anscombe's quartet in detail. (3 marks)
Anscombe's quartet is a set of four datasets that have nearly identical simple desc
(mean, variance, correlation, and linear regression lines), yet they are quite dif
This highlights the importance of visualizing data in addition to relying solely on
```

In [ ]:
```
3. What is Pearson's R? (3 marks)
 It is a measure of the strength and direction of a linear relationship between two
    It ranges from -1 to 1, where
r=1 indicates a perfect positive linear relationship (i.e., as one variable increas
r=-1 indicates a perfect negative linear relationship (i.e., as one variable increa

r=0 indicates no linear relationship between the variables.
```

In [ ]:
```
4. What is scaling? Why is scaling performed? What is the difference between normal
and standardized scaling? (3 marks)
When you have a lot of independent variables in a model, a lot of them might be on
1. Ease of interpretation
2. Faster convergence for gradient descent methods
Normalized scaling - This method scales the features so that they fall within a spe
Standardized scaling - transforms the data into a distribution with a mean of 0 and
```

In [ ]:
```
5. You might have observed that sometimes the value of VIF is infinite. Why does th
(3 marks)
- I have seen it happen when there are redundant dummy values. ( where drop_first w
Thi is because there is perfect correlation between two independent variables.
```

In [ ]:

```
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regr
(3 marks)
The quantile-quantile plot is a graphical method for determining whether
two samples of data came from the same population or not.
A Q-Q plot, which stands for Quantile-Quantile plot, is a graphical tool used in s

Here's how a Q-Q plot works:

Theoretical Quantiles: We start by selecting a probability distribution (e.g., nor

Sample Quantiles: We then sort the data from smallest to largest and calculate the

Plotting: We plot the theoretical quantiles on the x-axis and the sample quantiles

If the data perfectly follows the theoretical distribution, the points on the Q-Q

The Q-Q plot is a powerful visual tool for assessing the distribution of data. It'
```