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)   
 Some of the categorical variables like season and weather situation had a great affect on cnt.

1. We can conclude that the Count is the highest when the weather is Clear i.e. Clear, Few clouds, Partly cloudy, Partly cloudy and negligible when the day is rainy i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.
2. We can conclude that the most bike rentals happens in Autumn and the least in Spring.

Some of the variables like holiday and working day did not have much difference in the cnt mean.

2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)

We can represent n categories with n-1 dummies. So, its important to remove first dummy variable as it is redundant.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)   
 temp / atemp has 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)

I checked for the p-value and VIF to check if the correlation between the variables used in the model and the target variable was high or not.

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 were :- weathersit\_rainy , year\_2019 and temp.

General Subjective Questions

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

Linear Regression is a type if Supervised learning method . It tries to find a linear equation that can best describe the impact of independent variables on the dependant variables. It achieves this best fit line iteratively using least square method.

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

Anscombe’s quartet tells nus the importance of visualisation before building models.  
Variables must be plotted and the distribution of the samples can help identify anomalies present in the data.

3. What is Pearson’s R? (3 marks)   
Pearson’s R is a way to measure the linear correlation. The correlation can be positive or negative based on the value which swings between -1 to 1. 0 being no correlation , and 1 ,-1 being the highest. It summarises the characteristics of a dataset.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm .  
The data collected can have a highly varying range. If scaling is not done , the algorithm will take time and might model incorrectly.  
Normalization/Min-Max Scaling:  
It brings all of the data in the range of 0 and 1.   
Standardization Scaling:   
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)   
This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression  
Q-Q plots are also known as Quantile-Quantile plots. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.