1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

>> I have set two variable as categorical variables from the given set of data i.e., "season" and "weathersit" and by making this we see good significan and VIF with respective train and test model.

| | | | | | | |
|---|---|---|---|---|---|---|
| **season_2** | 0.1168 | 0.016 | 7.488 | 0.000 | 0.086 | 0.147 |
| **season_3** | 0.0695 | 0.023 | 3.013 | 0.003 | 0.024 | 0.115 |
| **season_4** | 0.1529 | 0.022 | 6.907 | 0.000 | 0.109 | 0.196 |
| **weathersit_2** | -0.0566 | 0.011 | -5.203 | 0.000 | -0.078 | -0.035 |
| **weathersit_3** | -0.2463 | 0.027 | -9.024 | 0.000 | -0.300 | -0.193 |

2. Why is it important to use drop_first=True during dummy variable creation?
>> Is to drop the redundant dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

>>temp show a goof and high correlation with cnt
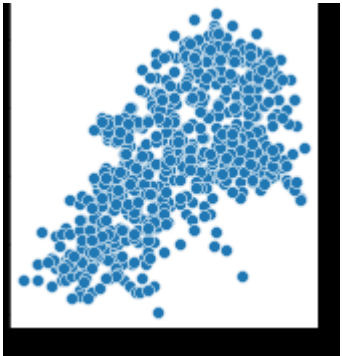
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

>> Validate with LOS regression by seeing the p type whether the variables are significant and also checked VIF whether the VIF of each variable is less than the 0.5

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

>>         temp = 0.6199 → It indicates that a unit of temp variable increase with 0.6199 units

         yr_1  = 0.2306→ It indicates that a unit of yr_1 variable increase with 0.2306 units

         season_4 =0.1039→ It indicates that a unit of season_4 variable increase with 0.1039 units

**General Subjective Questions**

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

>> Linear regression algorithm is based on supervised learning. It is part of regression analysis in machine learning.
It is a technique which is used to find out the best relation with respective target variable and the input variable.
This regression is used to find out the effect of Input variables on Target variable and also to find out the change in Target variable with respect to one or more input variable. Till now we have see simple linear regression and multiple linear regression. In linear model used two different libraries: statsmodels and SKLearn.

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

Quartet which four and four datasets that have identically statistical properties and appeared different when plot the graph.
So this was explained or constructed by Anscombe on the importance of plotting graph before analysis and the effect's of the outliers.

3. What is Pearson's R? (3 marks)

It is also know as PCC person's correlation coefficient, linear correlation between data sets. The covariance of two variables, divided by the product of their standard deviations and its result always lies between -1 and 1

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

>> Scaling means to transfer the available data to fit into a specific scale for ex: (0-10). It is used to normalize the data. 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)

>> VIF will infinite when there is perfect correlation between two independent variables and also these variables expressed exactly by the linear combination

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

>>Q-Q plots are the two quantiles which are against each other.  The use case of Q-Q is to find out whether the values comes in between two quantiles are from same distribution. If it is same then there will be 45degree Q- Q plot is plotted.