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?**

   Answer:

   - Spring season is good for renting more bikes.
   - Year 2019 has more demand for bike sharing and hence after the pandemic, the demand will increase.
   - Months May-Sept are better months for renting bikes.
   - Holidays are having more count of bike sharing data.
   - Weekends have comparatively higher bike renting data.
   - Clear weather is more appropriate for bike sharing.

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

   Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

   For example, In our case, we have created dummy variables for the weather situation variable, so when new has created three new variables as per the values in the weather_sit, drop_first=True dropped the main variable.

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

   Answer: temp, atemp variables have highest correlation with the target variable named as cnt.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Answer: After building the model, we can consider following assumptions:

   - Check Normality of the residuals
   - Check homoscedasticity

- Check multicorrelation
- Check linearity between residuals(errors) and actual values.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

   Answer: The top three features contributing significantly towards explaining the demand of the shared bikes are as follows:

   - Season- spring
   - Month- september
   - Weather- Clear, Few clouds, Partly cloudy, Partly cloudy

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Answer:

   - Linear regression is a simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables.
   - It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.
   - If there is a single input variable (x), such linear regression is called simple linear regression.
   - And if there is more than one input variable, such linear regression is called multiple linear regression.
   - The linear regression model gives a sloped straight line describing the relationship within the variables.
   - The goal of the linear regression algorithm is to get the best values for intercept and coefficients to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

2. **Explain the Anscombe's quartet in detail.**

   Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

   Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

3. **What is Pearson's R?**

   Answer: The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Answer: 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.

   Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

   Normalized scaling brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardized scaling 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?**

   Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

   Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

   Pictorial representation of the QQ plot :