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?
   - **The provided data has seasonal effect such as Spring season has less demand compared to other seasons.**
   - **The demand has inclining trend compared to 2018 and 2019.**
   - **The demand was more on non-holidays.**
   - **Clear weather condition contributed more renting**
2. Why is it important to use **drop_first=True** during dummy variable creation?
   - **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(n-1).**
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - **Registered and cnt has 0.95% correlation**
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   - **There should be linear relationship between dependent and independent variable(cnt)**
   - **The independent variables should not be correlated.**
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   - **Positive correlation in year2019 and temperature**
   - **Negative correlation in weather in snow**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   - **Linear Regression is ML algorithm based on supervised learning. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.**
   - **Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.**
   - **The function for linear regression is Y = mx+c where m and c are intercept and coefficients**

2. Explain the Anscombe's quartet in detail.

   - **Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.**
   - **They have very different distributions and appear differently when plotted on scatter plots.**

3. What is Pearson's R?

- **Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.**
- **The form of the definition involves a "product moment", that is, the mean of the product of the mean-adjusted random variables. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.**

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

- **Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.**
- **Example if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique.**

S.NO.	Normalisation	Standardisation
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

- **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 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.**
- **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.
   - **The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot.**
   - **If the two data sets come from a common distribution, the points will fall on that reference line. ... This particular type of Q Q plot is called a normal quantile-quantile (QQ) plot.**