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?**
   In the given dataset there are lot of categorical columns like Year, Holiday Windspeed etc.
   As from regression model it is clear that Year has higher coefficient and then followed by others like Holiday & Windspeed

2. **Why is it important to use drop_first=True during dummy variable creation?**
   When we created Dummy variable the dataset contain variable feature So to reduce confusion & repeatability of data, So it make sense to drop the original variable
   So to drop the variable Use command **drop_first=True**

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   The variable 'temp' has higher correlation followed by 'Holiday'

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

   Validating the assumption of Linear Regression Model based on below assumptions:

   1. Normality of error terms : Error terms should be normally distributed
   2. Multicollinearity check : There should be insignificant multicollinearity among variables.
   3. Linear relationship validation :- Linearity should be visible among variables
   4. Homoscedasticity :- There should be no visible pattern in residual values.
   5. Independence of residuals :- No auto-correlation

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   A. Year
   B. Month
   C. Season

General Subjective Questions

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

   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. 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 vazriables getting used

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression

Equation of Linear Regression

Y = M*X + C

2. **Explain the Anscombe's quartet in detail.**
   Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

   - The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
   - The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
   - In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
   - Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.[2][3][4][5][6]

The datasets are as follows. The *x* values are the same for the first three datasets

"Visualization gives you answers to questions you didn't know you had." — Ben Schneiderman

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

In statistics, the **Pearson correlation coefficient** (**PCC**, pronounced /ˈpɪərsən/) — also known as **Pearson's *r***, the **Pearson product-moment correlation coefficient** (**PPMCC**), the **bivariate correlation**,[1] or colloquially simply as **the correlation coefficient**[2] — is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

For a population

Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter *ρ* (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. Given a pair of random variables , the formula for *ρ*[10] is:

$$p(x,y) = cov(X,Y)/ 6x,6y$$

where:

- cov(X,Y) is the covariance
- 6x is the standard deviation of X
- 6y is the standard deviation of Y

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

Scaling:

- Scaling is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor.
- Scaling law, a law that explains how many natural phenomena exhibit scale invariance.

Scaling performed because:

It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range.

the difference between normalized scaling and standardized scaling

The values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

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.**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.

   This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.