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: Categorical variables in the datasheet (season, month, weekday, weather set) had effect on dependent variable as a dependent variable is what you measure in the experiment and what is affected during the experiment. The dependent variable responds to the independent variable. It is called dependent because it "depends" on the independent variable. In a scientific experiment, you cannot have a dependent variable without an independent variable.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Answer: Because we have already created the Dummies of the particular variable with 0 and 1, so it's good to delete the drop the un-necessary categorical variables as it's not needed in the model now.
   Example: for column "Month" we have Dummified to creates month wise columns
   df['mnth']= df['mnth'].map({1:"Jan", 2:"Feb", 3:"March", 4:"April",5:"May",6:"June",7:"July",8:"Aug",9:"Sep",10:"Oct",11:"Nov",12:"Dec"})
   And now the month column is not needed for the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer: atemp seems to have 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?
   Answer: scatterplots is helpful for validating the linearity assumption as it is easy to visualize a linear relationship on a plot. So we plot graph between
   y_test and y_pred
   In addition and similarly, a partial residual plot that represents the relationship between a predictor and the dependent variable while taking into account all the other variables may help visualize the true nature of the relationship between variables y_test and y_pred

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer: Features atemp , year and holiday are contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.?

Answer: linear regression is a method of finding the best straight line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Residual Sum of Squares Method.
In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.
One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.
Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as
$y = a + bx$

b = Slope of the line.
a = y-intercept of the line.
x = Independent variable from dataset
y = Dependent variable from dataset

2. Explain the Anscombe’s quartet in detail.

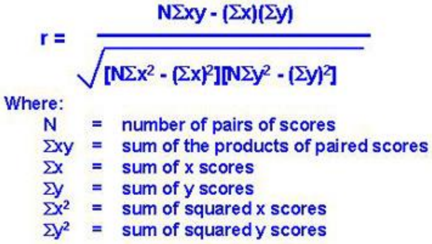
Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before 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) is not distributed normally; 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 distribution 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.

3. What is Pearson’s R?

Answer: 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 (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter “r” for a sample.



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

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

Answer: Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

In scaling (also called min-max scaling), you transform the data such that the features are within a specific range e.g. [0, 1].

difference between normalized scaling and standardized scaling

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a z-score, and data points can be standardized with the following formula:

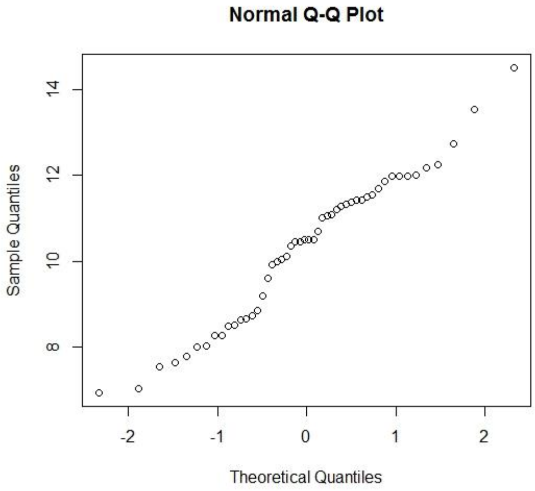
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: If VIF is infinity that means there is perfect correlation A large value of VIF indicates that there is a correlation between the variables.
it shows a perfect correlation between two independent variable.
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 variable 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

Answer: The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It’s just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles