Assignment-based Subjective Questions

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

   Answer:
   The following conclusions were reached after a careful examination of the dataset provided:

- Demand for motorcycles was higher in 2019 and lower in 2018.
- The demand for the bikes was roughly same on working and non-working days.
- Holiday demand for bikes was lower than non-holiday demand.
- The month of September had the most demand for bicycles, followed by October and August, while January saw the lowest demand.
- The demand for motorcycles was lowest during the spring season.
- The demand for the bikes is very consistent throughout the week.
- The demand for motorcycles is greatest when the weather is bright and there are few clouds, and it is lowest when there is light snow and rain.

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

   Answer:
   The drop first=True command aids in reducing the excess column formed during the fake variable generation process. As a result, it lowers the correlations formed between dummy variables. We have a few categorical columns in our dataset where drop first might be utilised.

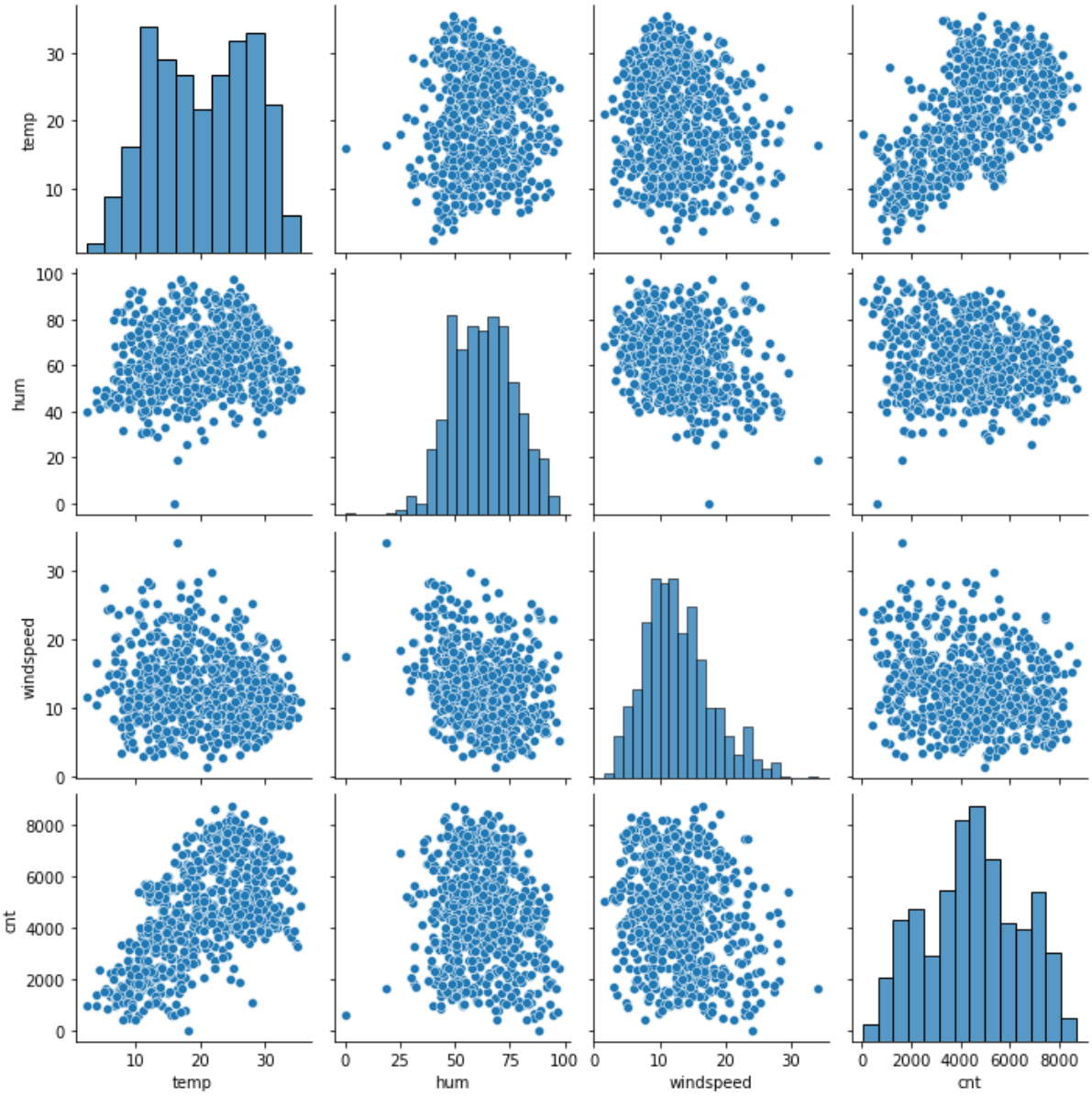
   Example:

   df = pd.get_dummies(data[['season','weekday','mnth','weathersit']],drop_first=True)

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

   Answer:

   The snapshot of the pair-plot plotted during our study is shown below. As the pair-plot below plainly indicates, the variable "temp" (temperature) has the strongest correlation with the target variable "cnt."



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

   Answer:

   After developing the model on the training set, we tested the assumption of Linear Regression using residual analysis between predictions and actual values.

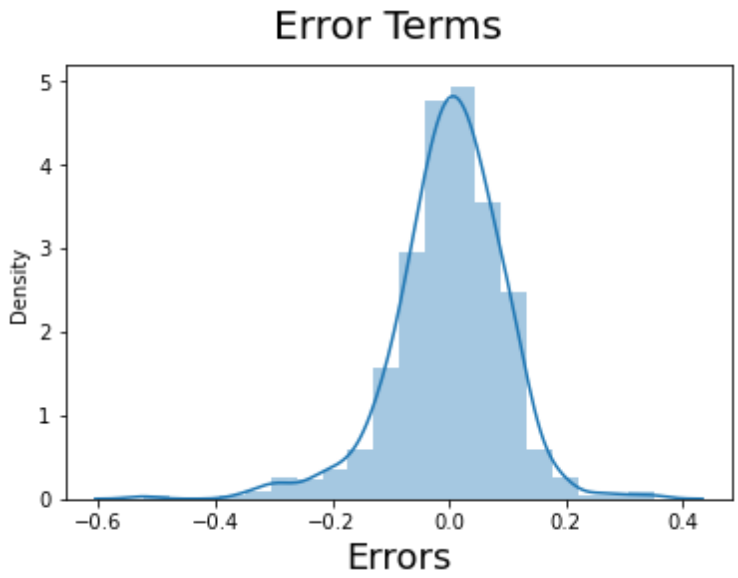


Figure: In this plot, the error terms are observed to be regularly distributed, with their mean centered at 0.

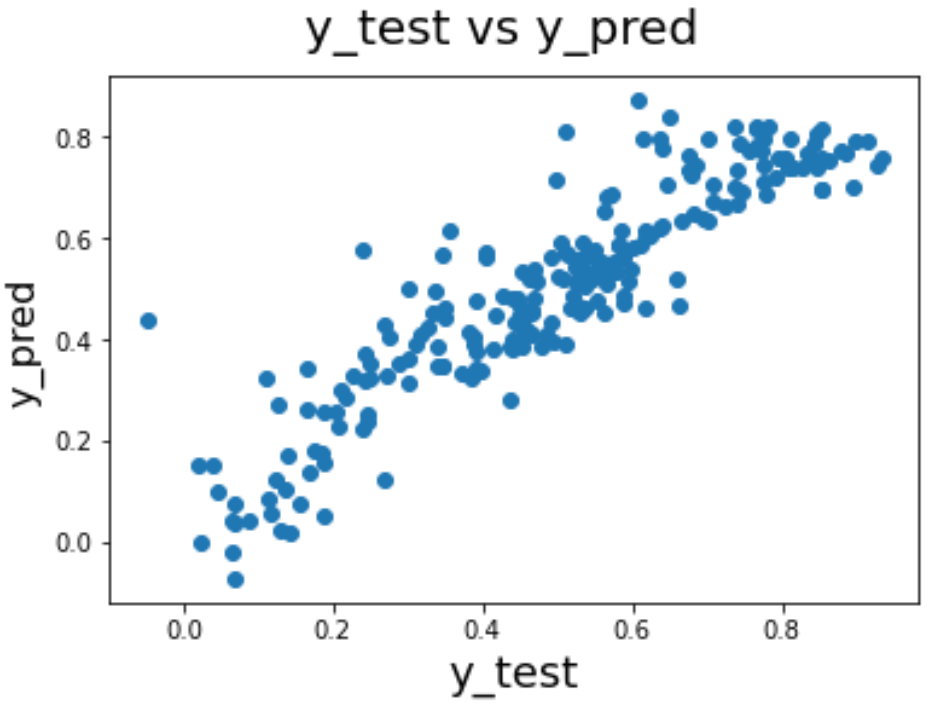


Figure: In addition, using this linearly distributed plot

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

Answer:

The top three factors that contribute considerably to understanding the demand for shared bikes are:

1. Temperature ("temp") - there is a positive relationship.

2. The year ("yr") has a positive connection.

3. Weathersit LightSnow – has an inverse relationship.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

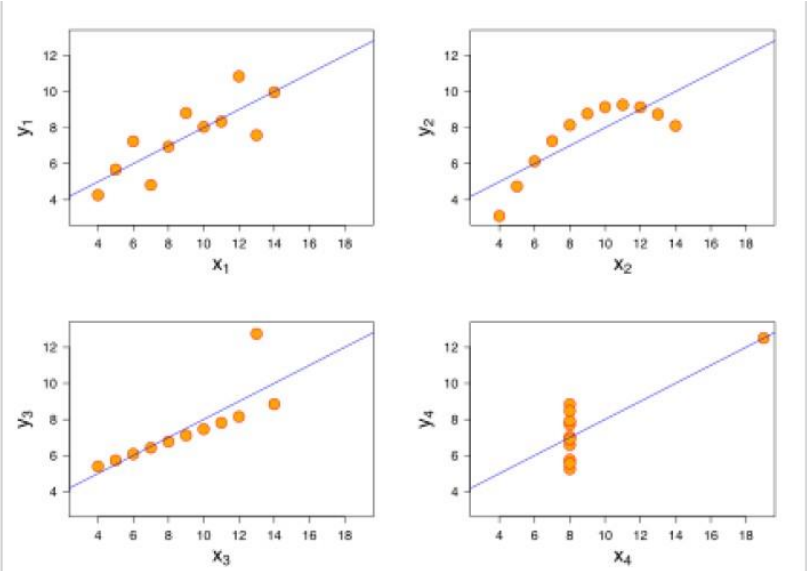
Linear Regression is a fundamental algorithm in Machine Learning that falls under the category of supervised learning. It is a statistical model that determines whether or not there is a linear connection between a dependent variable and a set of independent factors. Increases or decreases in the values of one variable have the same effect on the other variable in a linear relationship. It is mostly used for forecasting. Simple Linear Regression describes the connection between a dependent variable (Y) and an independent variable (X).

Its mathematical equation is given as: $\mathbf{Y = \beta 0 + \beta 1} X$
Where,

- Y is target/dependent variable
- X is independent variable
- $\beta 1$ is the coefficient of X
- $\beta 0$ is the intercept

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is made up of four datasets that have almost similar simple statistical features but seem substantially different when displayed. Each dataset has eleven (x, y) points. They were built in 1973 by statistician Francis Anscombe to highlight the significance of charting data before analysing it, as well as the impact of outliers on statistical features.



• The first scatter plot (top left) looks to show a straightforward linear connection, corresponding to two correlated variables, where y might be described as gaussian with mean linearly dependent on x.

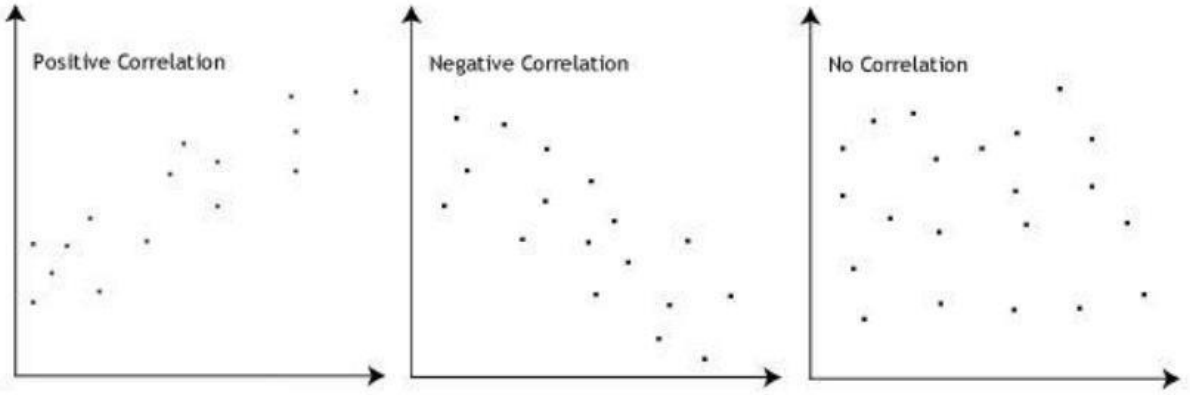
• Although there is no normal distribution in the second graph (top right), it is clear that there is some link between the two variables x and y, and the Pearson correlation coefficient is irrelevant. A broader regression and the accompanying coefficient of determination would be preferable.

• The distribution in the third graph (bottom left) is linear, but it should have a different regression line (a robust regression would have been called for). The calculated regression is offset by one outlier that appears to be far from the line.

• In the fourth graph (bottom right), we observe that a high-leverage point is enough to yield a high correlation coefficient, even while the other points show no link between the variables.

3. What is Pearson's R?

Answer: Pearson's R is a numerical representation of the strength of the linear relationship between two variables. Pearson's correlation coefficient ranges between -1 and +1 in the following cases:



• If R = 1, the data is fully linear and has a positive slope (both variables vary in the same direction, whether positive or negative).

• If R = -1, the data is completely linear with a negative slope (i.e., both variables change in different directions).

• If R = 0, there is no linear relationship.

• If R is greater than zero, the relationship is weak.

• If R is more than 5, it indicates a moderate relationship.

• And if R > 8, it indicates a significant relationship.

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

Answer: Scaling is a technique for normalising the range of independent variables. It aids in the acceleration of computations in any algorithm. It is done in regression to bring all of the independent variables on the same scale. Without scaling, the model considers higher values to be significant and lower values to be non-significant. Scaling has no effect on the p-value, r-squared value, or other parameters.

Normalization scaling or Min-Max scaling brings all the data in the range of 0 and 1.

Min-Max scaling: $X = (\mathbf{x} - \mathbf{min}(\mathbf{x})) / (\mathbf{max}(\mathbf{x}) - \mathbf{min}(\mathbf{x}))$

Standardization scaling replaces the values by z-scores. It brings all the data into standard normal
distribution which as mean(μ) as 0 and standard deviation(σ) as 1.

Standardization scaling: $\mathbf{x} = (\mathbf{x} - \mathbf{mean}(\mathbf{x})) / \mathbf{sd}(\mathbf{x})$

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

Answer:

We know that if there is perfect correlation, VIF equals infinite. This demonstrates an exact correlation between two independent variables. In the event of perfect correlation, R2 = 1, resulting in 1/(1-R2) infinite. To remedy this issue, we must remove one of the variables from the dataset that is producing the perfect multicollinearity.

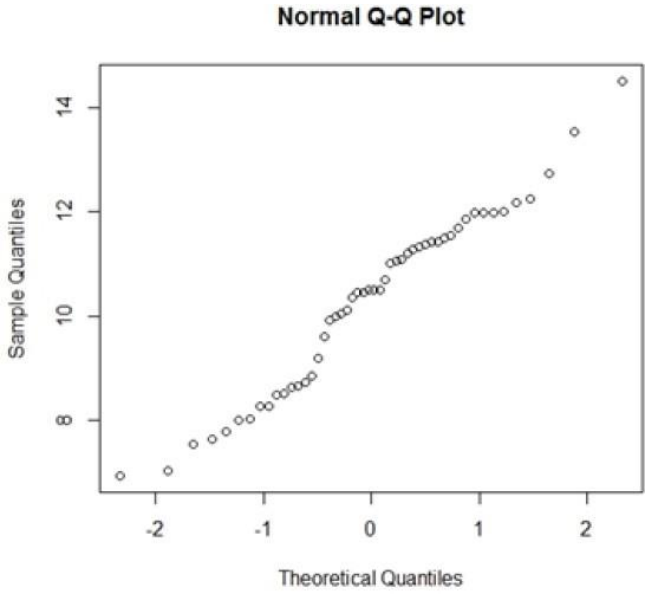
An infinite VIF value suggests that a linear combination of other variables may represent the relevant variable perfectly (which show an infinite VIF as well).

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

Answer:

The Q-Q plot, also known as the quantile-quantile plot, is a graphical tool for detecting if two data sets are from populations with a similar distribution.

A Q-Q plot is a scatterplot formed by charting two quantile sets against each other. If both sets of quantiles originate from the same distribution, the points should form a relatively straight line. An example of a Normal Q-Q plot is shown below, where both sets of quantiles are drawn from Normal distributions.



The Application of the Q-Q Plot in Linear Regression: The Q-Q plot is used to determine if the points are roughly on the line. If they don't, it suggests our residuals aren't Gaussian (Normal), and our errors aren't either.

The Importance of the Q-Q Plot:

• Sample sizes do not have to be equal; and

• Many distributional features may be investigated at the same time. For instance, variations in position, scale, symmetry, and the existence of outliers.

• The q-q plot, rather than analytical approaches, can give more information about the nature of the difference.