Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you inferabout their effect on the dependent variable?**

Answer:
After carefully analysing dataset, the following findings were made:

- The demand for motorbikes increased in 2019 while declining in 2018.
- The spring season saw the lowest demand for motorbikes.
- All throughout the week, there is a steady demand for the bikes.
- On both working and non-working days, there was nearly the same demand for the bikes.
- Bikes demand during holidays was lower than demand in non-holidays.
- In 2019, January experienced the lowest demand for bicycles, with the highest demand occurring in September, followed by October and August.
- When the weather is bright and there are few clouds, demand for motorbikes is at its highest; when there is light snow and rain, its showing at lowest.

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

Answer:

The surplus columns created during the fake variable formation process can be reduced with the use of the drop_first=True command. It thereby reduces the correlations that are created between dummy variables. Our dataset has a few categorical columns that might benefit from drop first.

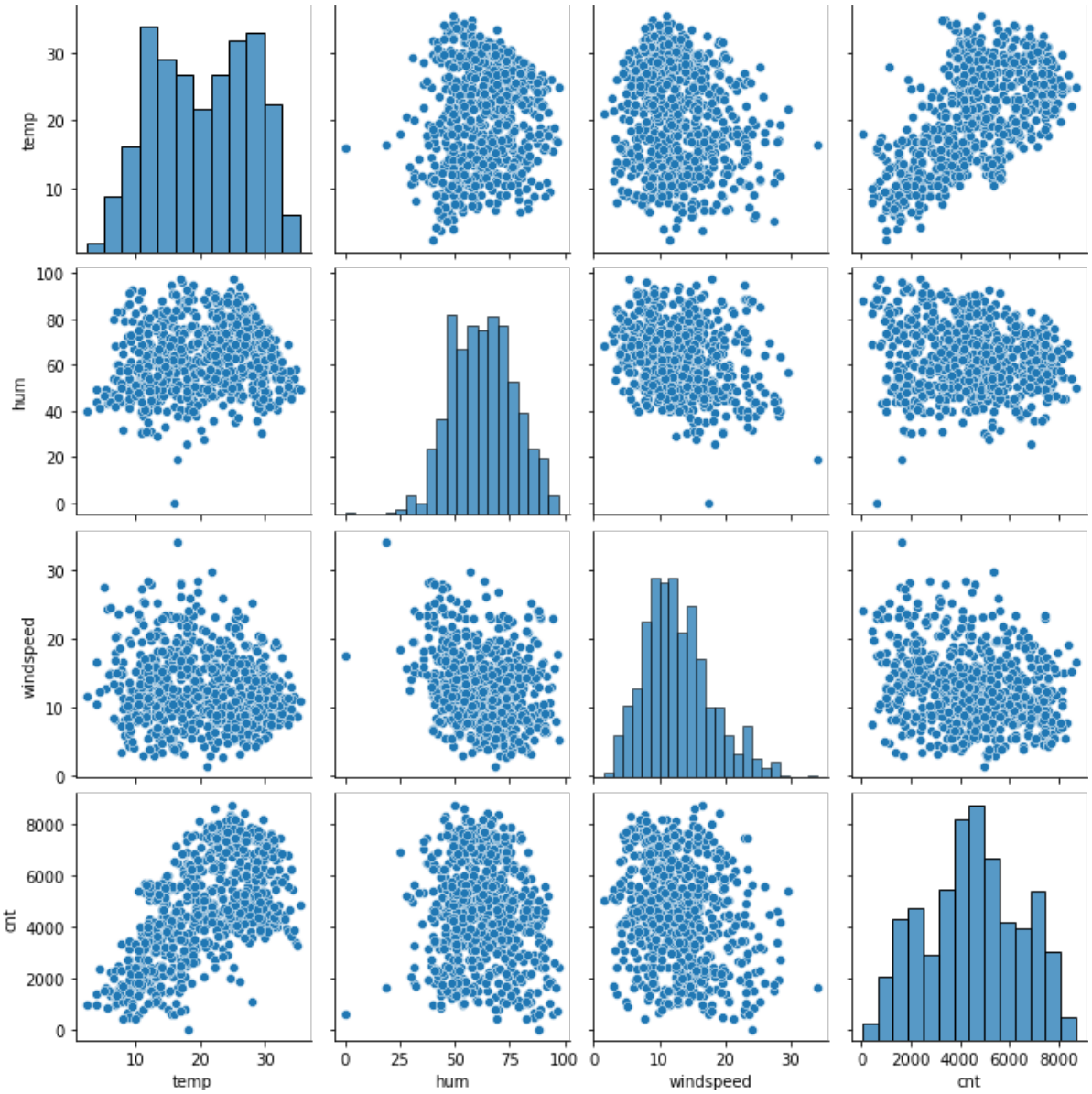
Example:

```
data=pd.get_dummies(data[['season','weekday','month','weather_condition']],drop_first=Tru
e)
```

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

Answer:

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



As we can see in graph, “temp” is having max correlation with “cnt” target variable.

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

Answer:

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

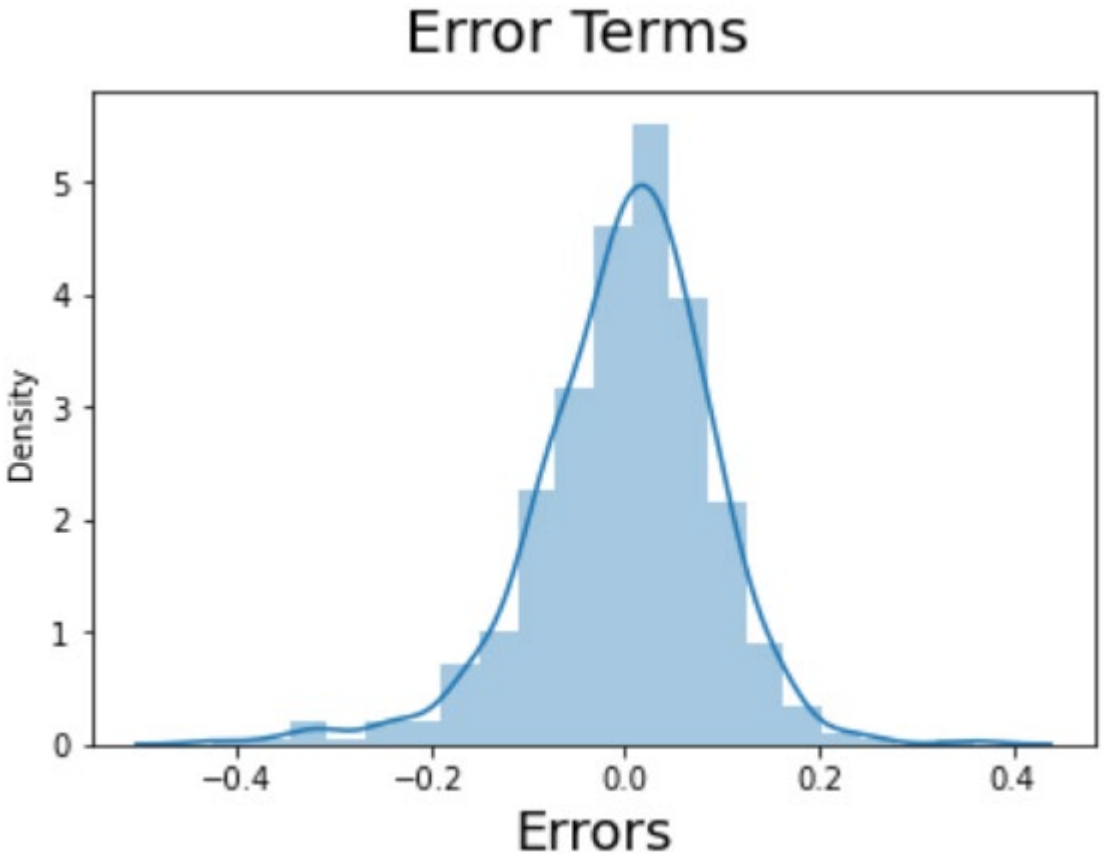


Figure: In this plot, the error terms are observed to be regularly distributed, with their mean centered at 0.

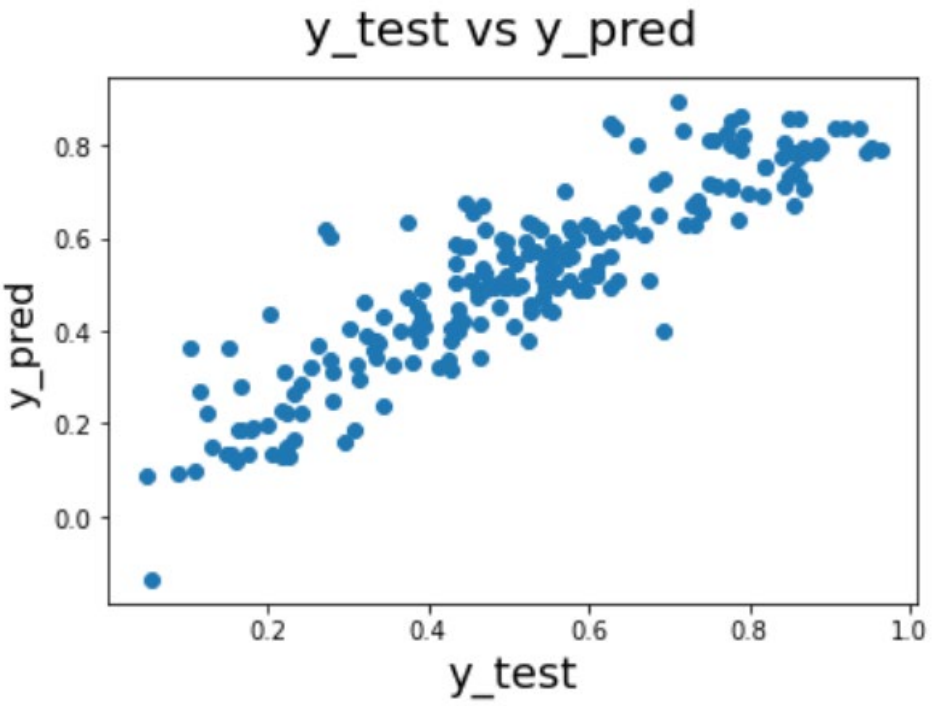


Figure: In addition, using this linearly distributed plot

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

Answer:

The top three elements that significantly aid in comprehending the demand for shared bikes are as follows:

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

2. The Year ("yr") has also a positive relationship.

3. Weathersit Light Snow – has an inverse relationship.

_____xxxx_____

General Subjective Questions

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

Answer:

A fundamental machine learning algorithm that belongs to the supervised learning subcategory is linear regression. It is a statistical model that assesses whether or not a dependent variable and a group of independent variables have a linear relationship. In a linear connection, changes in one variable's value, whether positive or negative, have the same impact on the other variable. Mostly, forecasting is done using it. The relationship between a dependent variable (Y) and an independent variable(X) is described by a simple linear regression.

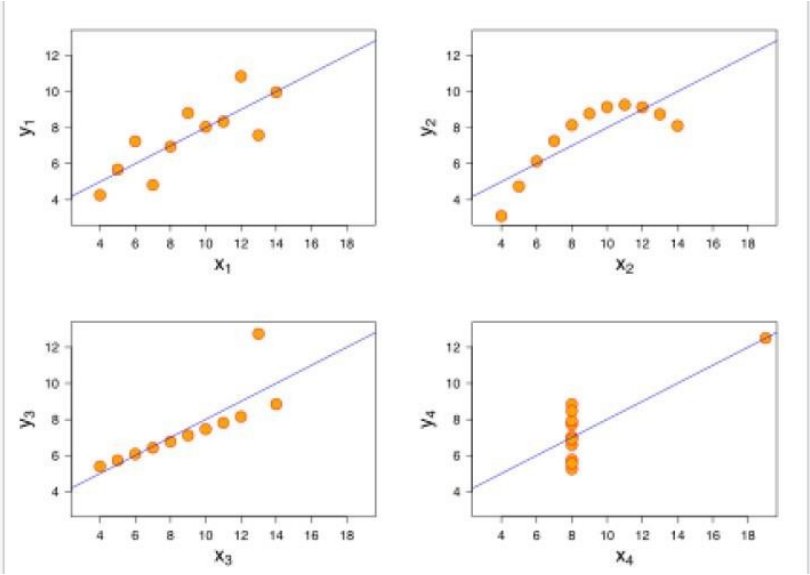
Its mathematical equation is given as: **Y =** β**0 +** β**1**X
Where,

1. Y is target/dependent variable
2. X is independent variable
3. β1 is the coefficient of X
4. β0 is the intercept

2. **Explain the Anscombe's quartet in details:**

Answer:

Anscombe's quartet consists of four datasets that, although they have almost identical fundamental statistical properties, appear to be very different when shown. There are eleven (x, y) points in each dataset. Francis Anscombe, a statistician, designed in 1973 to demonstrate the need of charting data before analysing it and the influence of outliers on statistical traits.



- The first scatter plot (top left) appears to represent a simple linear relationship between two correlated variables, with y being a gaussian distribution with a mean that is linearly dependent on x.

- The Pearson correlation coefficient has no bearing on the relationship between the two variables x and y, even if there is no normal distribution in the second graph (top right). It would be ideal to have a larger regression and associated coefficient of determination.

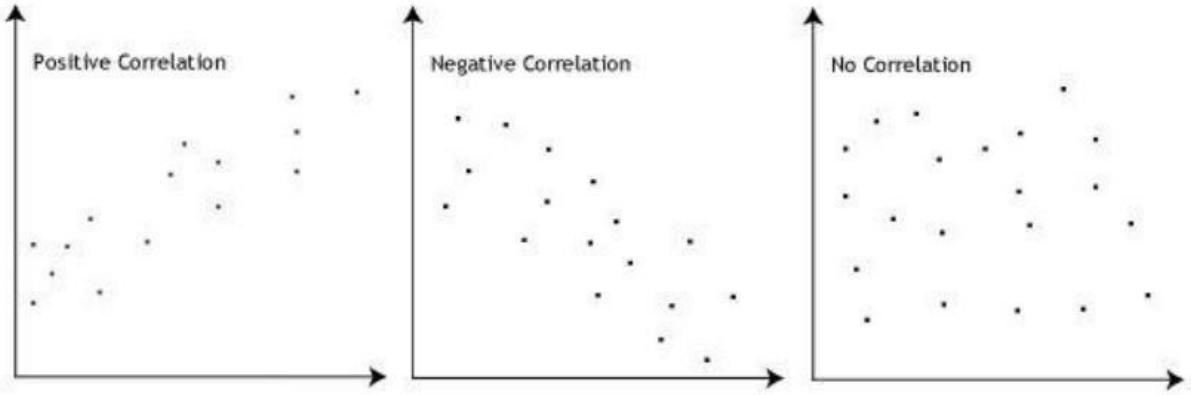
- Although the third graph's distribution is linear, a separate regression line should be used to represent it (a robust regression would have been called for). One apparent outlier that looks to be distant from the line cancels out the estimated regression.

- In the fourth graph (bottom right), even when the other points doesn’t indicate a relationship between the variables, a high-leverage point is sufficient to provide a high correlation coefficient.

3. What is Pearson's R?

Answer:

The strength of the linear relationship between two variables is quantified by Pearson's R. The following situations fall between the range of -1 and +1 for Pearson's correlation coefficient. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



• 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 normalizedscaling and standardized scaling?

Answer:

Scaling is a method for bringing the range of independent variables into regular proportions. It helps any algorithms computation run more quickly. Regression involves bringing all independent variables to the same scale. The model treats higher values as significant and lower values as non-significant in the absence of scaling. The p-value, r-squared value, and other parameters are unaffected by scaling.

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

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

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

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

Answer:

We know that if there is perfect correlation, VIF equals infinite. This demonstrates an exactcorrelation 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.

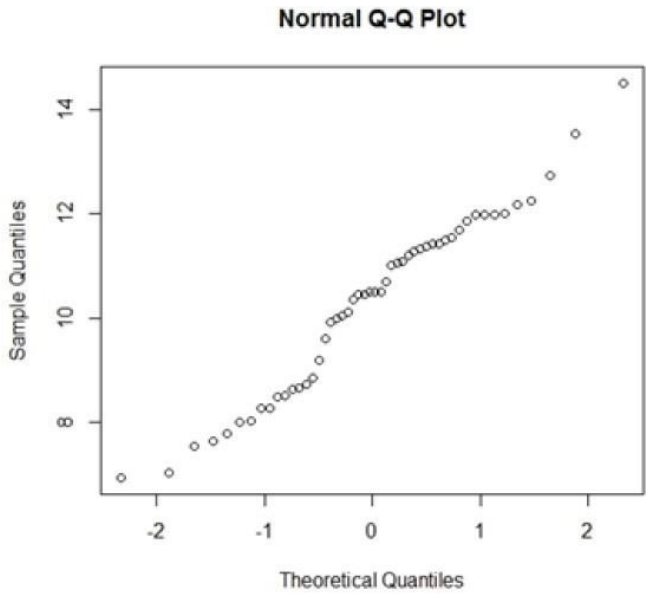
An infinite VIF value suggests that a linear combination of other variables may represent therelevant variable perfectly (which show an infinite VIF as well).

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 ifthe 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.
- 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.s