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

We infer from the above information that

- More bikes go on rent in fall
- Bad Weather conditions affect the count (Light Snow, Misty)
- September month has more count
- Holiday decreases the count
- Jan and dec has least count

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.

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

The numerical variable 'registered' has the highest correlation with the target variable 'cnt' , if we consider all the features.

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

- Weathersit_Light_now
- Yr
- Temp

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

- Temp
- Season
- Hum

- Windspeed
- Weathersit

1.Explain the linear regression algorithm in detail

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.
Linear regression is used to predict a quantitative response Y from the predictor variable X.

Some key points about Linear Regression:

- It assumes a linear relationship between the independent and dependent variables.
- It is sensitive to outliers and multicollinearity.
- It is a parametric model because it makes assumptions about the underlying data distribution.
- Regularization techniques like Lasso and Ridge Regression can be used to handle overfitting and improve model performance.

In summary, Linear Regression is a powerful and interpretable algorithm for solving regression problems by modeling the linear relationship between the target variable and one or more independent variables. It serves as the basis for more advanced regression models and is widely used in various data analysis and predictive modeling tasks.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a famous statistical dataset that consists of four sets of data, each containing 11 data points. What makes Anscombe's quartet remarkable is that despite the vastly different patterns and distributions in the four datasets, they share almost identical statistical properties. This dataset was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and the potential pitfalls of relying solely on summary statistics.

Despite having distinct visual patterns when plotted, here are some astonishing statistical properties that are nearly identical across all four datasets:

Mean of X: Each dataset has a mean of approximately 9 for the X values.

Mean of Y: Each dataset has a mean of approximately 7.5 for the Y values.

Variance of X: Each dataset has a variance of exactly 10 for the X values.

Variance of Y: Each dataset has a variance of approximately 3.75 for the Y values.

Correlation between X and Y: Each dataset has a correlation coefficient of approximately 0.816 for the X and Y values.

Linear Regression: Each dataset has a linear regression line of Y on X with an equation of $y \approx 0.5x + 3$.

Anscombe's quartet serves as a compelling reminder that when analyzing data, visualization is crucial to understand the underlying patterns and relationships, as summary statistics alone may not tell the full story. By visualizing the data, we can detect outliers, identify potential nonlinear relationships, and gain deeper insights into the nature of the data, which may lead to more appropriate model selection and interpretation.

3.What is Pearson’s R?

Pearson's correlation coefficient, commonly denoted as "Pearson's R," is a statistical measure that quantifies the linear relationship between two continuous variables. It is used to determine how closely two variables are related to each other and to what extent they change together.

The Pearson's correlation coefficient is a value between -1 and 1, where:

If the value is close to +1, it indicates a strong positive linear relationship, meaning that as one variable increases, the other variable also tends to increase proportionally. If the value is close to -1, it indicates a strong negative linear relationship, meaning that as one variable increases, the other variable tends to decrease proportionally. If the value is close to 0, it indicates a weak or no linear relationship, meaning that there is little to no systematic association between the variables.

The formula for calculating Pearson's correlation coefficient between two variables X and Y with n data points is:

$r = (\Sigma[(X_i - \bar{y}) * (Y_i - \bar{y})]) / [(\Sigma(X_i - \bar{y})^2 * \Sigma(Y_i - \bar{y})^2)]{}^\wedge(1/2)$ Where:

X_i and Y_i are individual data points of X and Y, respectively. $\bar{y}$ (read as "y-bar") is the mean of Y values. The summation symbol Σ represents the sum over all data points. In words, Pearson's correlation coefficient measures the covariance of X and Y (how the variables change together) normalized by the product of their standard deviations (how much each variable varies from its mean).

Some key points about Pearson's correlation coefficient:

It is sensitive to the scale of measurement of the variables. It assumes a linear relationship between the variables. If the relationship is nonlinear, the correlation may not accurately reflect the association. It does not imply causation. Even if two variables have a high correlation, it does not necessarily mean that one causes the other.

Pearson's correlation coefficient is widely used in statistics, data analysis, and machine learning to assess the strength and direction of relationships between variables and to identify potential dependencies in the data.

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

Scaling, in the context of data preprocessing for machine learning, refers to the process of transforming the features (independent variables) of a dataset to bring them within a specific range. The goal of scaling is to standardize the features so that they all have a similar scale, which can be important for certain machine learning algorithms and optimization processes.

In both cases, you're transforming the values of numeric variables so that the transformed data points have specific helpful properties. The difference is that: in scaling, you're changing the range of your data, while. in normalization, you're changing the shape of the distribution of your data.

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.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50 of the data fall below that point and 50 lie above it. 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