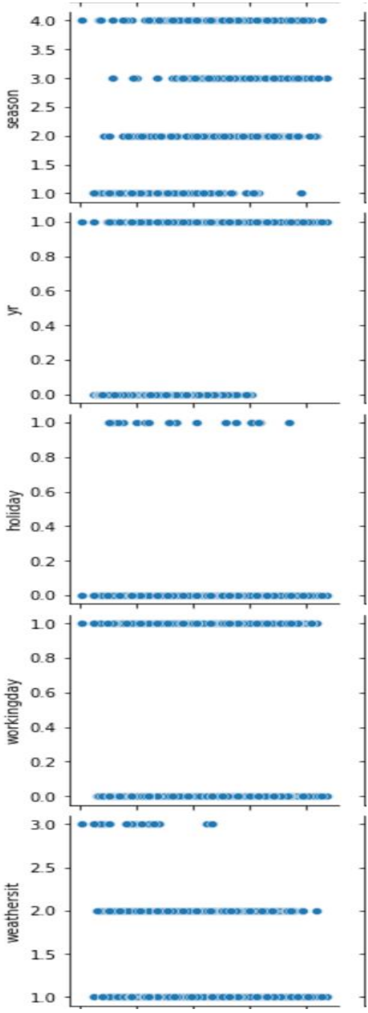
Boom Bikes Linear Regression Assignment

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

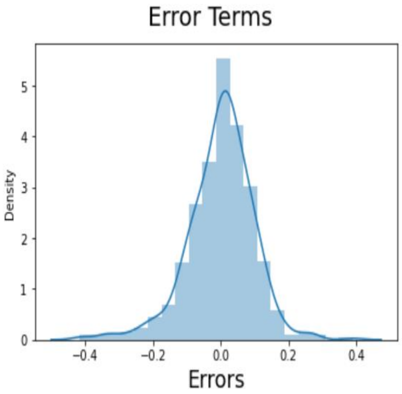


As we can see in the above plot of the categorical variables related to dependent driver(cnt) it shows that specifically for weather situation & season, as the weather changes the count of total rental bikes including both casual and registered bikes changes and same of season that as the season changes it shows its impact on count of total rental bikes including both casual and registered bikes. But if we look at holiday or working day weather it's a working day or non-working day or holiday or non-holiday it doesn't show much impact on the count of total rental bikes including both casual and registered bikes.

2. Why is it important to use **drop_first=True** during dummy variable creation?
   It is important to drop_first=True during variable creation because it helps reduce extra columns created while creating dummy variables. Therefore, it reduces the correlations produced between dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Temperature has the highest correlation with target variable of **64.4%**

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



   With the help of residual analysis, I analyzed that the variance of residual is same for every value of X, it is perfectly aligned at mean=0 and for any fixed value of X, Y is normally distributed and during this analysis I validated the assumptions of linear regression.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Top 3 features contributing significantly towards the demand of the shared bikes are-:
   Temperature with positive coefficient 0.362729
   Weather Situation -Light **Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds** with negative coefficient -0.271313
   Year with positive coefficient 0.236372

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Linear regression can be defined as a statistical model that analyzes the linear relationship between a dependent variable and a given set of independent variables. A linear relationship between variables means that a change (increase or decrease) in the value of one or more independent variables causes a corresponding change (increase or decrease) in the value of the dependent variable.
   Mathematically, the relationship can be expressed using the formula −
   Y=mX+b where Y is the dependent variable you are trying to predict.
   X is the independent variable used to make the prediction.
   m is the slope of the regression line representing the effect of X on Y.

   b constant known as Yintercept. If X = 0, Y is equal b.
   Furthermore, linear relationships can be positive or negative in nature, as explained below. −
   Positive Linear Relationship

A linear relationship is said to be positive if both the independent and dependent variables are increasing.
A linear relationship is said to be positive when the independent variable increases and the dependent variable decreases.
Assumptions

- Below are some assumptions about the data set produced by the linear regression model. Multicollinearity occurs when independent variables or characteristics have dependencies.
- Autocorrelation - Another assumption made by linear regression models is that the data have little or no autocorrelation. Basically, autocorrelation occurs when there is a dependency between residual errors.
- Relationships Between Variables - A linear regression model assumes that the relationship between the response and characteristic variables must be linear.

2. Explain the Anscombe’s quartet in detail.
Anscombe's Quartet is a mode l example that demonstrates the importance of data visualization. This consists of his 4 datasets, each consisting of 11 (x,y) points. The basic analysis of these datasets is that the descriptive statistics (mean, variance, standard deviation, etc.) are all the same, but the graphical representations are different. Each graph shows different behavior independent of statistical analysis.
    - Dataset I - Consists of a set of (x,y) points representing a linear relationship with variance.
    - Dataset II - Waveforms are shown but linear relationships are not shown.
    - dataset III - A close linear relationship between x and y, except for one large outlier.
    - Dataset IV - The values of x appear to remain constant except for one outlier.

3. What is Pearson’s R?
Pearson's correlation method is the most commonly used method for numeric variables. Assign a value between −1 and 1. 0 is no correlation, 1 is completely positive correlation, -1 is completely negative correlation. This is interpreted as: A correlation value of 0.7 between two variables indicates a significant positive relationship between the two variables. A positive correlation means that as the variable A increases, so does B; a negative value of correlation means that as A increases, B decreases.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
This is a data preprocessing step applied to the independent variables to normalize the data within a certain range. It also helps in speeding up the computation of algorithms.
In most cases, collected datasets contain features that vary widely in size, units, and range. Without scaling, the algorithm only considers the size and not the units, which is incorrect modeling. To solve this problem, all variables should be scaled to have the same size.

    It is important to note that scaling only affects coefficients, not other parameters such as t-statistics, F-statistics, p-values, and R-squared.
    Normalization/Min-Max Scaling:
    Makes all the data between 0 and 1. sklearn.preprocessing.MinMaxScaler helps implement normalization in Python.

Standardised Scaling:
Normalization replaces values with Z-scores. Fit all data to a standard normal distribution with zero mean (μ) and 1 standard deviation (σ).

sklearn.preprocessing.scale helps implement standardization in Python.
The disadvantage of normalization compared to Standardised is that some information in the data is lost. In particular, information about outliers is lost.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
If there is perfect correlation, VIF = infinity. This shows perfect correlation between the two independent variables. A perfect correlation would have R2 = 1 and 1/(1-R2) infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
A Quantile-Quantile (Q-Q) chart is a graphical tool that allows you to assess whether a data set reasonably derives from a theoretical distribution such as normal, exponential, or uniform. It also helps determine whether two datasets come from populations with a common distribution.

This is useful in linear regression scenarios when you receive separate training and test data sets, and Q-Q plots can be used to confirm that both data sets come from the same distributed population.
Used to check that if two datasets come from populations with common distribution have similar tail behavior, have common location and scale & have similar distributional shapes.