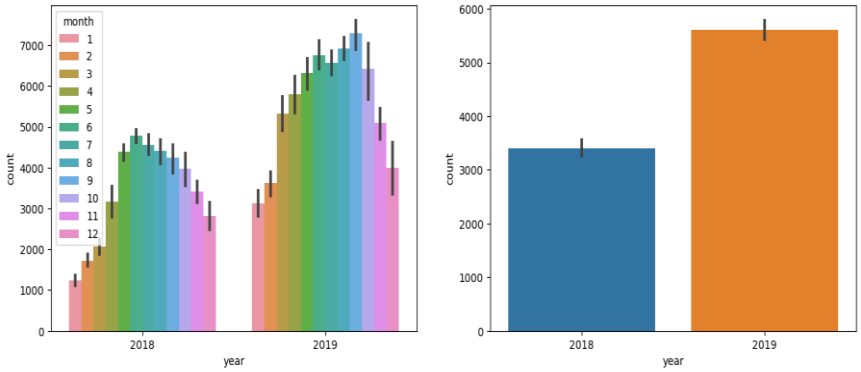
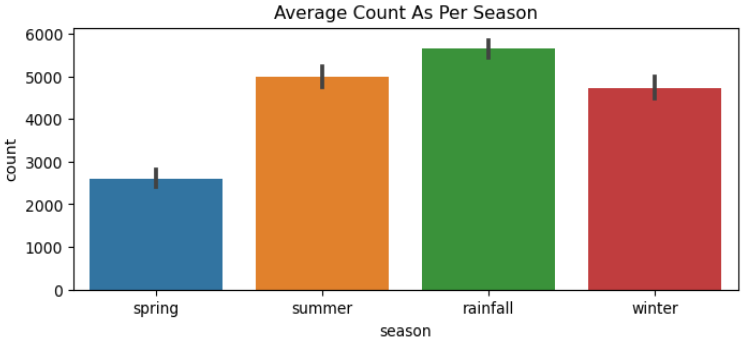
ASSIGNMENT BASED SUBJECTIVE QUESTION

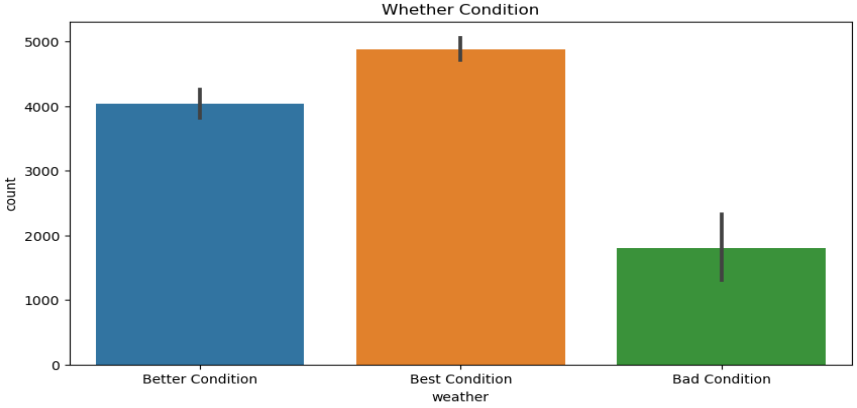
1. **From Your Analysis Of The Categorical Variables From The Dataset, What Could You Infer About Their Effect On The Dependent Variable?** (3 Marks)

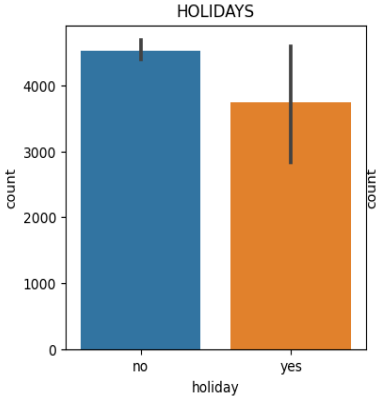
Answer:

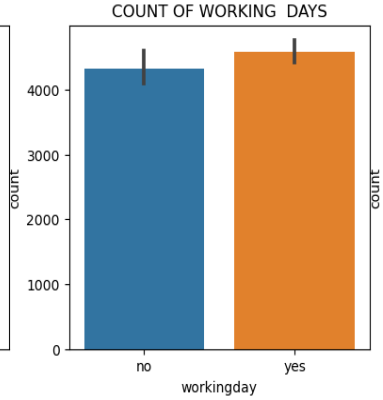
- Season: The Impact Of Season On Bookings Is Significant, With A Notably Higher Number Of Bookings Occurring During The Summer Season, Particularly Due To Increased Rainfall. Summer Exhibits A Greater Volume Of Bookings Com pared To Winter And Spring.
- Weather: The Weather Serves As A Categorical Variable, And Its Influence Is Substantial. Bookings Are Higher In The Best Conditions (Clear, Few Clouds, Partly Cloudy) Compared To Both Better And Worse Conditions.
- Holiday: When There Is A Holiday, The Demand Is Lower Compared To Days Without A Holiday.
- Year: From Our Observation In Terms Of The Year, The Demand Showed An Increase In 2019 Compared To 2018.
- Working day: When It's A Working Day, The Demand Tends To Be Slightly Higher Compared To Non-Working Days.

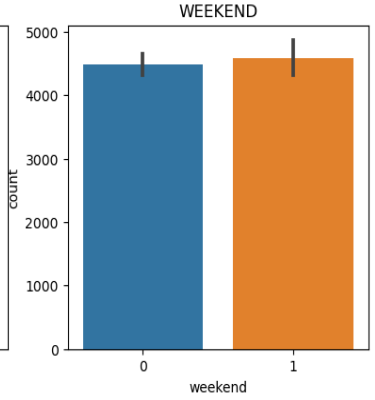












2.Why Is It Important To Use Drop_First=True During Dummy Variable Creation? (2 Mark)

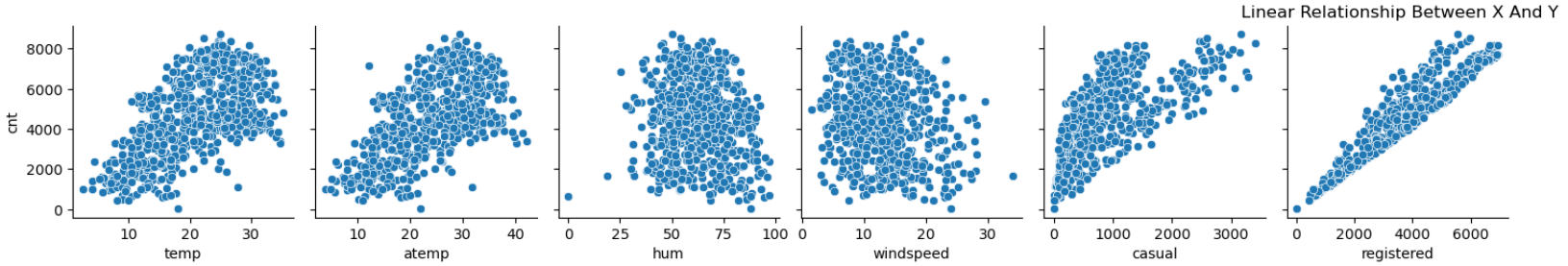
Answer : Using **drop_first=True** in dummy variable creation is crucial in statistical modeling, especially with categorical data. It helps prevent issues like multicollinearity, where variables are too closely linked, causing problems in regression. By skipping the first category when making dummy variables, it solves this problem and improves how we estimate coefficients, understand variable importance with p-values, and overall makes models more reliable. This simple setting not only cleans up redundant data but also makes models easier to understand and better at predicting outcomes.

Multicollinearity Prevention : drop_first=True aids in preventing multicollinearity, a problem where one categorical predictor can be accurately predicted from others.

Reduction of Redundancy: This parameter reduces redundancy among categorical variables, making models more robust and predictive.

3.Looking At The Pair-Plot Among The Numerical Variables, Which One Has The Highest Correlation With The Target Variable? (1 Mark)

Answer :



From The Above Plot We Observed:

a) Highly Positive Correlation Between The Register And Count (0.95)
b) Positive Correlation Between The Casual And Count (0.67)
c) There Is High Positive Correlation Between The Temperature And Count (0.63)
d) Highly Positive Correlation Between The Atemp And Count (0.63)
e) Year And Count Are Positive Correlation (0.57)

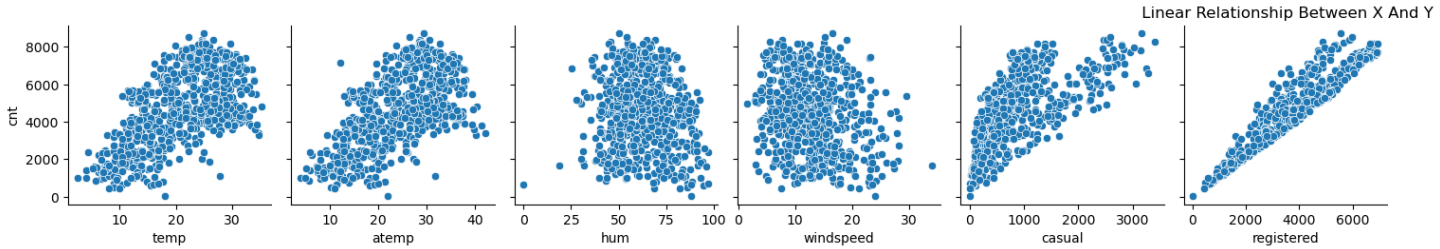
4.Looking At The Pair-Plot Among The Numerical Variables, Which One Has The Highest Correlation With The Target Variable? (3 Marks)

Answer :

Assumptions Of Linear Regression :

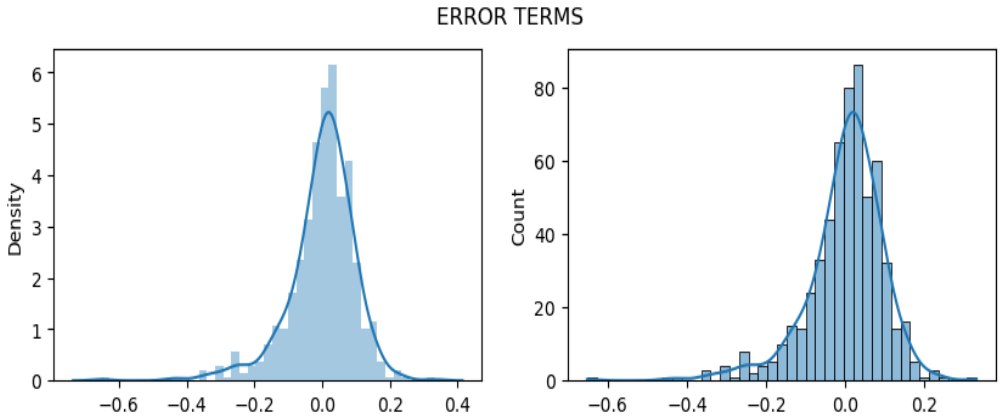
- Linear Relation Is A Relation Between Independent Variable And Dependent Variable
- No Multicollinearity
- Normality Of Residuals
- Homoscedasticity
- No Autocorrelation Of Errors

- **Linear Relationship Between Input And Output:** There Should Be A Linear Relation Between X And Y. From The Data Set We Have Observed That There Is Linear Relation Ship From The Graph Temp Atemp Casual Registered Are Linear And Where As Hum And Windspeed Negative Linear.



- **Normality Of Residuals :**

  - The residuals (errors) are `normally distributed`
  - From The Figure We Come To Know That `Error Terms Are Normally Distributed`



From The Graph We Come To Know That Our Error Terms Are Normally Distributed From That We Come To Know What Our Model Is Good.

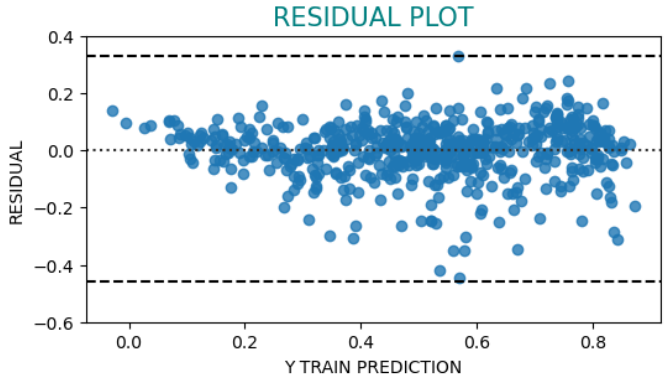
- **NO MULTICOLLINEARITY :**

Multicollinearity Refers to High Correlations Among Predictor Variables in A Regression Model, Leading To Issues In Estimating Coefficients' Impact. VIF Values Below 5 Typically Indicate Moderate Multicollinearity. Severe Multicollinearity, With VIF Values Exceeding 10, Can Distort Coefficient Estimates, Affecting Model Interpretation and Reliability. Regularization Techniques Or Feature Selection Can Mitigate These Issues.

	Features	VIF
0	windspeed	4.57
1	season_spring	2.91
2	season_winter	2.76
3	season_summer	2.29
4	Year	2.08

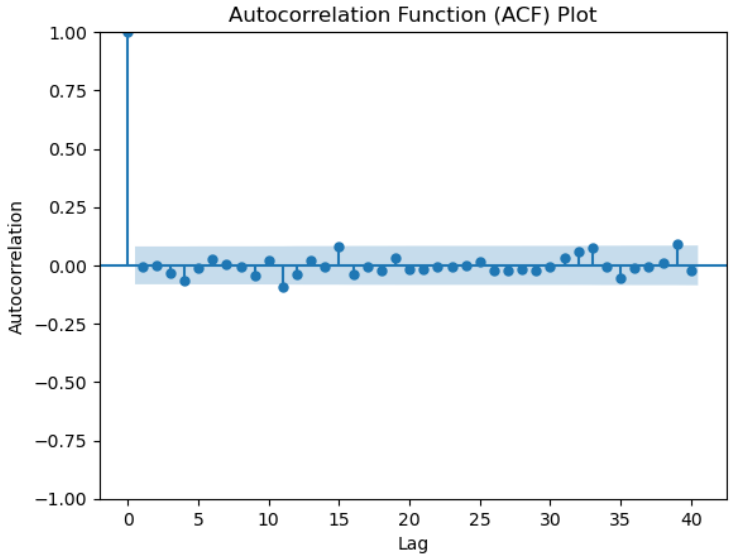
- **Homoscedasticity :**

In Regression Analysis, Homoscedasticity Means the Variance Of The Dependent Variable Is The Same For All The Data. So, In Homoscedasticity, The Residual Term Is Constant Across Observations. Simply, As the Value Of The Dependent Variable Changes, The Error Term Does Not Vary Much. When We Plot .When We Plot A Residual On The Scatter Plot ,The Spread Of The Dots Should Be Equal



- **No AutoCorrelation Of Error:**

There is no autocorrelation of errors. Linear regression model assumes that error terms are independent. This means that the error term of one observation is not influenced by the error term of another observation. In case it is not so, it is termed as autocorrelation.



These Are The Validate The Assumptions Of Linear Regression After Building The Model On The Training Set .

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

(2 Marks)

Answer:

Based On The Final Model, Which Are The Top 3 Features Contributing Significantly Towards Explaining The Demand Of The Shared Bikes Temp Which Has High Demand , Year 2019 Says That There Will Be A High Demand Of Bike And Season Winter Also Explain The High Demand Of The Bike

GENERAL SUBJECTIVE QUESTIONS

1.Explain the linear regression algorithm in detail. (4 Marks)

Answer:

Regression Is Nothing But The Out Put Variable To be Predicted Is A Continuous Variable.

Linear Regression: It Is Machine Learning Technique Which Predict The Variable Based On The Continuous Outcomes. Linear Regression Comes Under Supervised Machine Learning.

Supervised And Unsupervised Learning

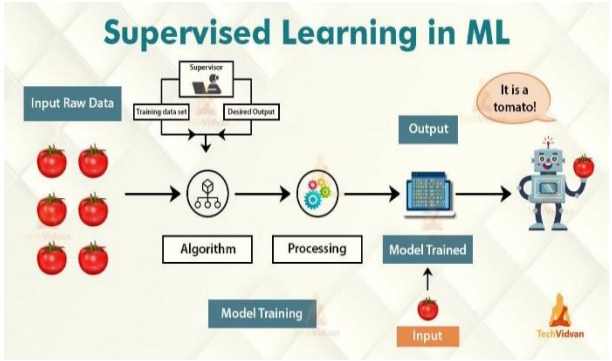
Supervised Machine Learning: It Is Complete Labeled Data Based On The Previous Data The Machine Learning Models Are Formed

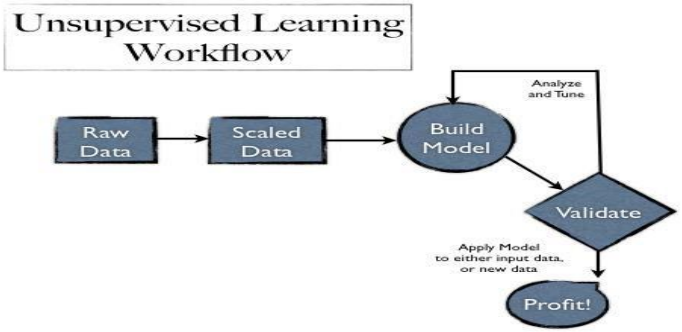
- **Linear Regression** : It Use Label Data And It Should Be Continuous Value
- **Classification** : Classification Refers To A Predictive Modaling Problem Where A Class Label Is Predicted For A Given Example Of Input Data.
    - Example : Detecting The Mail Which Is Spam Or Ham

Unsupervised Machine Learning : The Data Is Unlabeled Data There Is No Previous Data To Find Any Thing

- **Clustering** : Clustering Comes Under Unsupervise Machine Learning Model.

Flow Chart Of Supervised And Unsupervised Machine Learning :





Linear Regression Is Classified In To Two Types

- **Simple Linear Regression**
- **Multiple Linear Regression**

Simple Linear Regression : It Is Mathematicial Straight Line Equation Y=Mx+C Or $Y=B_0+ B_1x$ (Where X Is A Feature Variables And M Is Slope And C Is Intercept) Which Explains The Relation Ship Between The One Independent Variable And One Dependent Variable Using A Straight Line And The Straight Line Is Plot On The Scatter Plot. In Simple Linear Regression There Will Be Linear And Non Linear

Multiple Linnear Regresion : It Is Mathematicial Straight Line Equation $Y=Mx_1mx_2+Mx_3+.....Mx_n +C$ Or $Y=B_0+ Bx_1bx_2+Bx_3+.....Bx_n$ (Where X Is A Feature Variables And M Is Slope And C Is Intercept) Which Explains The Relation Ship Between The Multiple Independent Variable And One Dependent Variable Using A Straight Line And The Straight Line Is Plot On The Scatter Plot. In Multiple Linear Regression There Will Be Linear And Non Linear

2.Explain the Anscombe's quartet in detail. (3 marks)

Answer :

It can be defined as a group of four data sets that are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Because there are two points for the same x value and equidistance y values on the opposite side of the line nullifying the effect so there is no impact on the line.

It tells us the importance of visualizing the data before applying various algorithms.

The four datasets were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

• Applications: The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3.What is Pearson's R? (3 marks)

Answer

It is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviation. Mathematically, we can write Pearson's R can be represented as:

• Correlation = covariance (gen x, gen y) /(std(x) * std(y)) Covariance numbers can be any value between positive and negative infinity. When all the data fall on a straight line then covariance and product of the std terms are the same and division gives us -1 or 1. When data do not fall on a line then covariance and product of the std terms give a value close to zero.

The Pearson's Correlation Coefficient Varies Between -1 And +1 Where:

- R = 1 Means The Data Is Perfectly Linear With A Positive Slope (That Is, Both Variables Tend To Change In The Same Direction)
- R = -1 Means The Data Is Perfectly Linear With A Negative Slope (That Is, Both Variables Tend To Change In Different Directions)
- R = 0 Means There Is No Linear Association
- R > 0 < 5 Means There Is A Weak Association
- R > 5 < 8 Means There Is A Moderate Association
- R > 8 Means There Is A Strong Association

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

VIF is the measure of how much the variance of an estimated regression coefficient increases due to collinearity.

Formula : $\boldsymbol{vif} = \frac{1}{(1-R2^2)}$

where the R2 score is the proportion of the variance in the dependent variable that is predictable from the independent variable. There are three cases of R2 score 1. If the R2 score is 0 In this case, VIF will be

1 as the variables are independent.

2. If the R2 score lies between 0 and 1. In this case, VIF will be greater than 1. The high value of VIF indicates that there is high multicollinearity.

3. If the R2 score is 1. If all the features are dependent then VIF is 1. In the last case, it will be infinite and it indicates a large value of VIF indicates that there is a high correlation between the variables

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 Marks)

Answer

Feature Scaling : Feature Scaling Is A Technique To Standardize The Independent Features Present In The Data In A Fixed Range. It Is Preformed During The Data Preprocessing To Handle Highly Varying Magnitude Or Values Or Unit. If Feature Scaling Is Not Done, Then A Machine Learning Algorithm Tends To Weigh Greater Values, Higher And Consider Smaller Values As The Lower Values, Regardless Of The Unit Of The Values. It Bring The Value In A Fixed Range

Example : If We Have Data Like Iq And Salary Those Who Have High Iq Has High Salary If We Calculate There Distance Between Salray And Iq We Will Get A Far Point Of Salary To Make Them In A Fixed Range Or To Make Them Close We Use Feature Scaling

Feature Scaling Has Two Techniques Named

- Normalization Also Called Has **Min Max Scaling**
- Standardization Normalization

Min-Max Normalization: This Technique Re-Scales A Feature Or Observation Value With Distribution Value Between 0 And 1.

Formula For Min Max Scaling :

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization: It Is A Very Effective Technique Which Re-Scales A Feature Value So That It Has Distribution With 0 Mean Value And Variance Equals To 1.

Formula For Standardization :

$$x_{scaled} = \frac{x - mean}{sd}$$

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression? (3 Marks)

Answer :

A Q-Q plot, short for **Quantile-Quantile plot**, is a graphical tool used to assess if a dataset follows a particular theoretical distribution, often compared to a normal distribution. In linear regression, Q-Q plots are instrumental in **checking the assumption of normality**, which is crucial for reliable predictions and inference.

Use of Q-Q plot in Linear Regression:

Normality Assumption: Linear regression assumes that the residuals (the differences between observed and predicted values) are normally distributed. The Q-Q plot helps validate this assumption by plotting the quantiles of the residuals against the quantiles of a theoretical normal distribution.

Visual Comparison: It provides a visual comparison between the expected distribution (typically a normal distribution) and the actual distribution of the residuals. If the points on the plot lie approximately along a diagonal line, it suggests the residuals are normally distributed.

Identifying Departures: Deviations from a straight line in a Q-Q plot indicate departures from normality. Skewed or heavy-tailed distributions might show as points deviating from the diagonal line.

Model Confidence: Ensuring normality in residuals is crucial for accurate estimation of confidence intervals and hypothesis testing in linear regression. Departures from normality can affect the reliability of statistical inferences.

Importance in Linear Regression:

Assumption Validation: Q-Q plots help validate one of the key assumptions in linear regression, ensuring the reliability of the model's estimates and predictions.

Diagnostic Tool: It serves as a diagnostic tool to identify potential issues with the model, such as outliers or influential data points affecting the normality assumption.

Improving Model Accuracy: By confirming normality, it enhances the trustworthiness of the model's results, aiding in making more accurate predictions and drawing valid conclusions from the data.

In essence, Q-Q plots play a pivotal role in verifying the assumption of normality in linear regression, contributing to the model's reliability and the accuracy of statistical inferences derived from it.