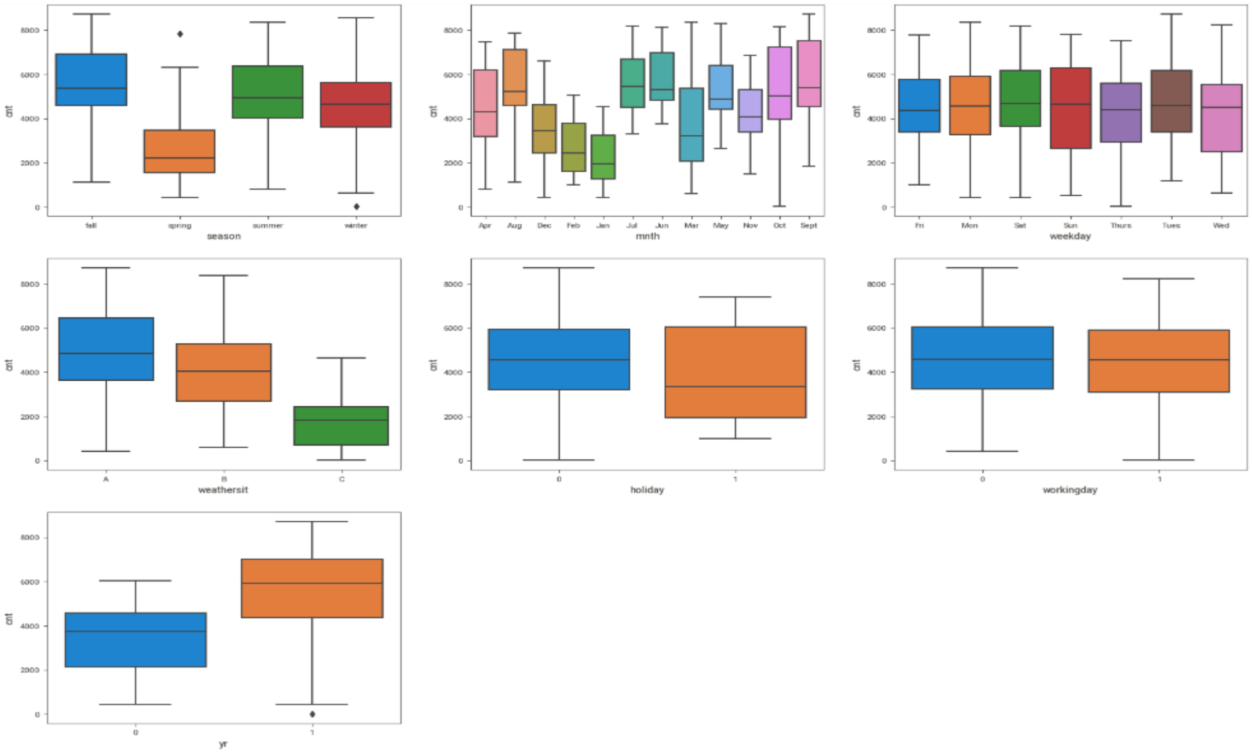
Linear Regression Q & A

A Anurag

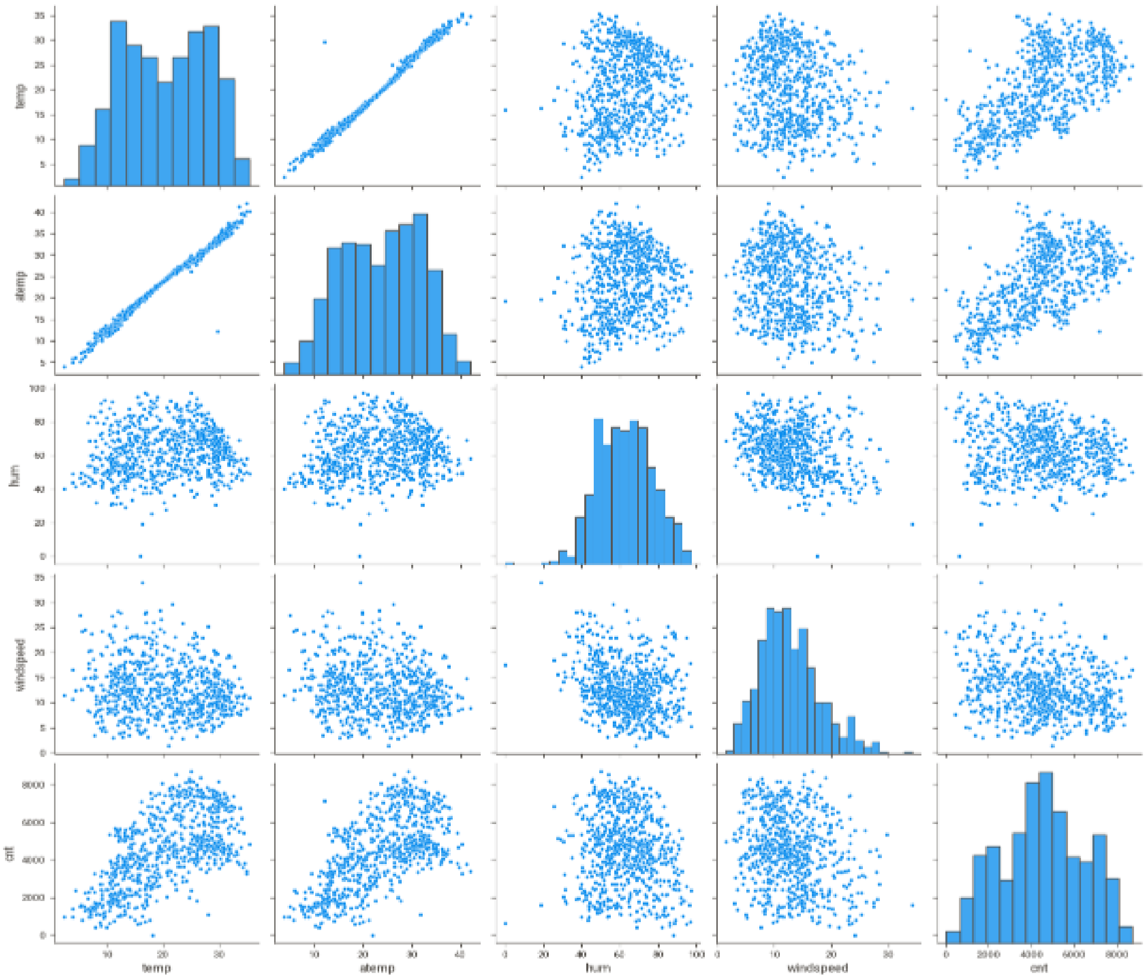
Assignment-based Subjective Question

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
    - Based on my analysis:
        - The fall season appears to have garnered greater interest in bookings, with a significant year-over-year increase in reservations from 2018 to 2019 across all seasons.
        - The majority of bookings occurred in the months of May, June, July, August, September, and October. The trend showed a rise from the beginning of the year until the middle of the year, followed by a decline towards the end of the year. The number of bookings for each month saw an increase from 2018 to 2019.
        - Thursday, Friday, Saturday, and Sunday exhibit a higher volume of bookings compared to the other days of the week.
        - It's evident that clear weather conditions attracted more bookings, which is quite expected. Furthermore, in comparison to 2018, bookings increased for each type of weather situation in 2019.
        - When it's not a holiday, the number of bookings tends to be lower, which is a reasonable observation since people may prefer to stay at home and spend quality time with their families during holidays.
        - Bookings appeared to be nearly equal on both working days and non-working days. However, there was an increase in the booking count from 2018 to 2019.



2. **Why is it important to use drop_first=True during dummy variable creation?**
    - The purpose of creating a dummy variable is to represent a categorical variable with 'n' levels by generating 'n-1' new columns, each indicating whether a specific level exists or not through binary values (0 or 1).
    - By using this approach, if all these binary values are zero, it implies that the observation belongs to the category that was dropped.
    - Therefore, 'drop_first=True' is often used to ensure that the resulting dummy variable set corresponds to 'n-1' levels, thus reducing the correlation among the dummy variables, as it helps avoid multicollinearity.
    - Multicollinearity is a situation where two or more predictor variables in a regression model are highly correlated, which can make it challenging to determine the individual effect of each predictor on the response variable. Dropping one level and using 'n-1' dummy variables helps mitigate this issue.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    - “temp” and “atemp” are the columns which have highest correlation to the target variable that is “cnt”. As atemp and temp are highly correlated we will eventually drop atemp. The correlation of numerical variables looks like this:



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - Validated using below mentioned ways:
        - Normality of Error terms, that is the errors should be normally distributed.
        - Multicollinearity, it should be insignificant among variables.
        - Homoscedasticity, there should be no such visual patterns in the residual values.
        - Independence of the residuals, that is there is no auto-correlation.
        - Linearity, there should be linearity among variables.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    - The top 3 features that significantly influence the demand for shared bikes are season, year, and temperature.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. A basic and popular supervised machine learning approach called linear regression is used to describe the connection between one or more independent variables (predictors or features) and a dependent variable (sometimes called the target or response). The dependent variable and the independent variables are assumed to have a linear relationship. The best-fitting linear equation to describe this relationship is the main objective of linear regression.

It has following procedure:-

- The linearity of the relationship between the variables, the independence of the errors, the constant variance of the errors(homoscedasticity), and the normally distributed errors are the main presumptions of linear regression. Verifying these assumptions is crucial prior to using linear regression on a dataset.

- Preparing the data involves gathering and preprocessing it. Ensure that you have a dataset with values for the independent variable (X) and the dependent variable (Y) in pairs.

- The following is a representation of the simple linear regression model:

$Y = \beta 0 + \beta 1X + \varepsilon$.

Y: The dependent variable

X: Independent variable

The value of Y when X is zero is represented by the intercept (y-intercept), or β0.

Slope coefficient, or β1, shows how much Y changes for every unit change in X.

ε: The residual error term that accounts for the unexplained variation in Y.

- Finding the values of β0 and β1 that minimize the sum of squared residuals (SSE), or the sum of the squared discrepancies between the actual and projected Y, is the aim of linear regression.

- Usually, you use something like the least squares method to select the line that fits the best. The least squares approach modifies the values of β0 and β1 in an effort to reduce SSE.

- In basic linear regression, common measures for assessing the model's performance are R-squared (R^2): A measurement of the percentage of the dependent variable's variance that the independent variable accounts for. RMS, also known as Root Mean Squared Error, is the mean square error. Calculates the mean squared difference between the expected and actual values. Q-Q and residual plots to evaluate the model's suppositions.

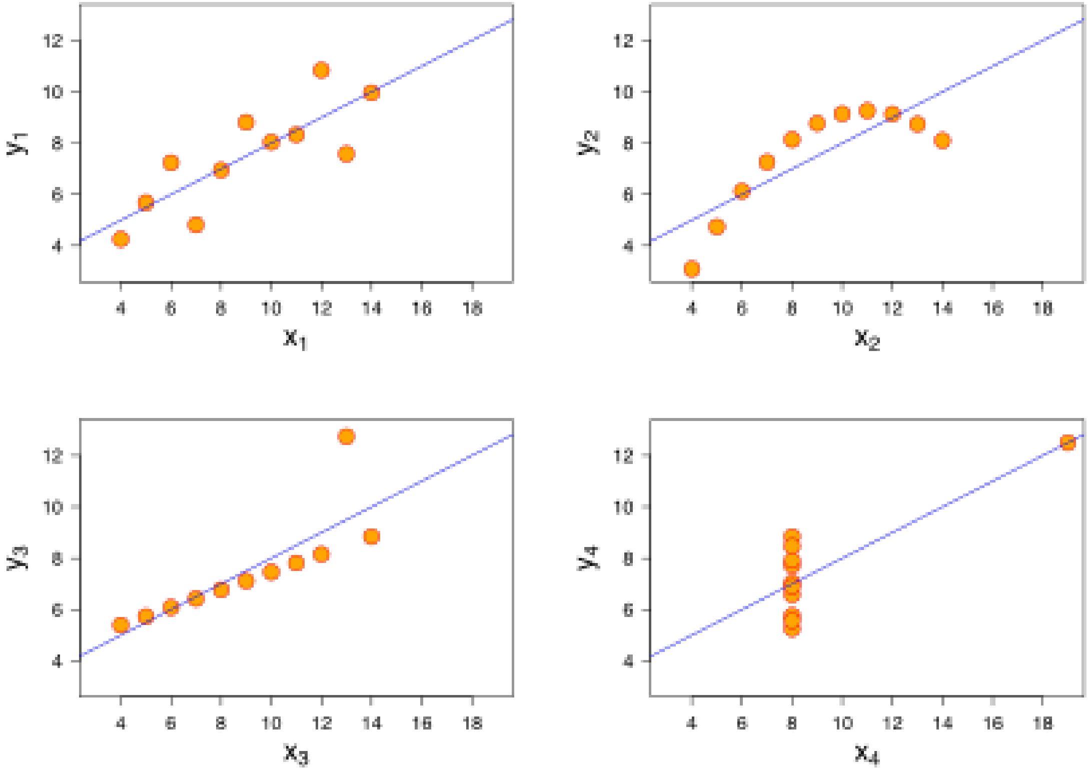
- You can use the model to make predictions once it has been trained. The linear equation can be used to determine the value of the dependent variable given a new value of the independent variable. For tasks like forecasting stock prices, evaluating product demand, and examining the effects of factors on an outcome, linear regression is utilized in a wide range of disciplines, including economics, finance, social sciences, and many branches of science and engineering.

- It's important to note that when there are numerous independent variables, linear regression can be expanded to multiple linear regression, but the fundamental ideas stay the same.

2. Explain the Anscombe’s quartet in detail.

Ans. Anscombe's quartet consists of four datasets with almost similar simple descriptive statistics that, when shown graphically, show clear differences. The British statistician Francis Anscombe composed this quartet in 1973 to highlight the limitations of depending only on summary statistics and the significance of data visualization in comprehending the nature of data. The idea that data should be examined and visualized before any statistical or modeling conclusions is frequently emphasized through the usage of Anscombe's quartet.

Following are the four datasets visualization:



NOTE:- Image reference from wikipedia

Dataset I: This dataset exhibits an obvious linear relationship between X and Y, with some random scatter around the line. For these data, a linear regression model might be suitable.

Dataset II: The link between X and Y seems to be curved and quadratic. Here, a linear model would be completely out of place.

Dataset III: The linear regression line is greatly affected by one outlier, which is present but otherwise comprises all the same X values. The significance of identifying and managing outliers is highlighted by this dataset.

Dataset IV: There is no simple relationship between X and Y in this dataset. There is some variability in the Y values, but the majority of the X values remain constant. This dataset demonstrates how patterns can still exist in data even in circumstances when summary statistics point to a lack of association.

You will discover that the summary statistics (such as mean, variance, and correlation) for X and Y in all four datasets are almost exact when you calculate them for each dataset. This highlights how summary statistics are not very useful for completely comprehending the features of a dataset. Anscombe's quartet shows how deceptive it may be to rely solely on summary data. It emphasizes how crucial data visualization is for exposing the underlying relationships and structure in data. Even though the summary statistics of two datasets are the same or nearly equivalent, a visual inspection can reveal radically distinct patterns. Therefore, judgments drawn from statistical models or decisions made simply on the basis of summary statistics may be inaccurate or inappropriate.

The quartet is a powerful reminder that before performing statistical analysis or creating models, data must be fully explored and visualized. To find out about outliers, hidden relationships, and other details that may not be visible from summary statistics alone, graphic tools and data visualization are crucial.

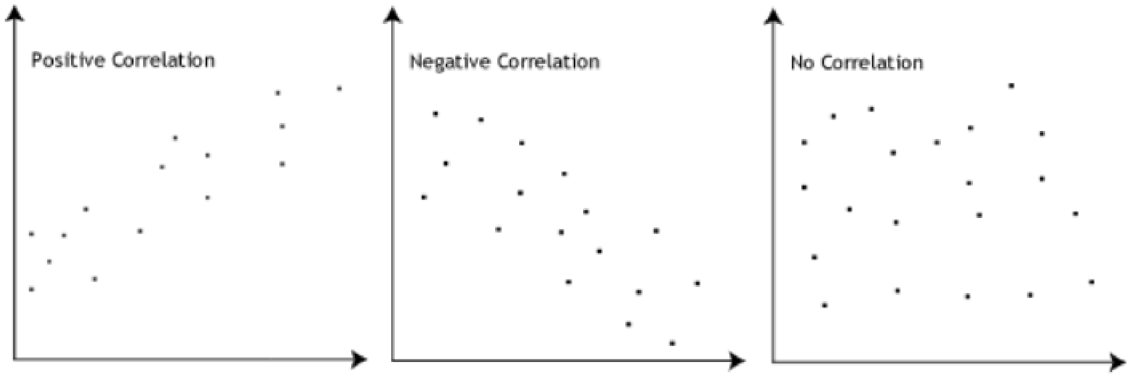
3. What is Pearson’s R?

Ans. The statistic known as Pearson's correlation coefficient, or simply "r" or "Pearson's r," measures the linear relationship between two continuous variables. It gauges the direction and intensity of a linear relationship between two variables. The range of Pearson's r is -1 to 1, where:

An r value of 1 indicates a perfect positive linear relationship. This means that as one variable increases, the other also increases proportionally.

An r value of -1 indicates a perfect negative linear relationship. In this case, as one variable increases, the other decreases proportionally.

An r value of 0 suggests no linear relationship between the variables.



NOTE:- The above image is from website:- https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

Ans. Scaling is a preprocessing technique in data analysis and machine learning that involves transforming the features (independent variables) of a dataset to bring them to a common scale or distribution. It's done to ensure that the variables contribute equally to the analysis, as some machine learning algorithms and statistical methods are sensitive to the scale of the input features.

1. The lowest and maximum values of the features are utilized in normalized scaling, whereas the mean and standard deviation are used in standardize scaling.

2. Standardized scaling is employed to guarantee zero mean and unit standard deviation, while normalized scaling is utilized when features are of different scales.

3. In contrast to standardized scaling, which lacks or is not limited in a certain range, normalized scaling scales values between (0,1) and (-1,1).

4. Outliers have an impact on normalized scaling, but they have no influence on standardized scaling.

5. When we don't know anything about the distribution, we use normalized scaling; when the distribution is normal, we use standardized scaling.

6. Standardized scaling is referred to as Z Score Normalization, while normalized scaling is called scaling normalization.

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

Ans. In a multiple regression study, a Variance Inflation Factor (VIF) quantifies the extent to which multicollinearity increases the variance of the predicted regression coefficients. When two or more independent variables in a regression model have a high degree of correlation, this is known as multicollinearity, and it makes it difficult to separate out the distinct effects of each variable. A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. And anything below 5 is acceptable range.

There is a perfect correlation between two independent variables when the VIF value is infinite. R-squared (R2) = 1 in the situation of perfect correlation leads to 1/(1-R2) infinity. In order to resolve this, we must remove the variable producing this perfect multicollinearity from the dataset. (Perfect Multicollinearity and Mathematical Dependency)

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

Ans. A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics and data analysis to assess if a dataset follows a particular theoretical distribution, such as the normal distribution. It is a type of probability plot that helps you visualize whether the quantiles (percentiles) of your dataset match the expected quantiles of the chosen theoretical distribution. In the context of linear regression, Q-Q plots are valuable for examining the distribution of residuals.

Use of a Q-Q Plot in Linear Regression:

One of the primary uses of a Q-Q plot in linear regression is to check the assumption of normality for the residuals (the differences between the observed and predicted values) of your regression model. Linear regression models often assume that the residuals are normally distributed. A Q-Q plot helps you assess whether this assumption holds.

When you create a Q-Q plot for the residuals, it displays the residuals' quantiles on the x-axis and the expected quantiles of a normal distribution on the y-axis. If the data points fall along a straight line, it suggests that the residuals are normally distributed. Deviations from a straight line indicate departures from normality.

Q-Q plots can reveal the presence of skewness or heavy tails in the data. If the points in the plot deviate from a straight line, it may suggest that the data is not normally distributed and might have skewness or outliers.

Importance of Q-Q Plots in Linear Regression:

Linear regression assumes that the residuals are normally distributed. A Q-Q plot provides a clear and visual way to check this assumption. If the residuals are not normally distributed, it might impact the validity of the regression analysis and the reliability of statistical inferences.

By examining the Q-Q plot, you can assess whether the linear regression model is appropriate for the data. If the residuals follow a normal distribution, it indicates that the model is well-suited for the data, and the parameter estimates are likely more reliable.

Q-Q plots are part of a set of diagnostic tools used in linear regression to identify issues like heteroscedasticity, nonlinearity, and outliers. Deviations from the expected straight line in the Q-Q plot can prompt further investigation into the nature of the residuals.