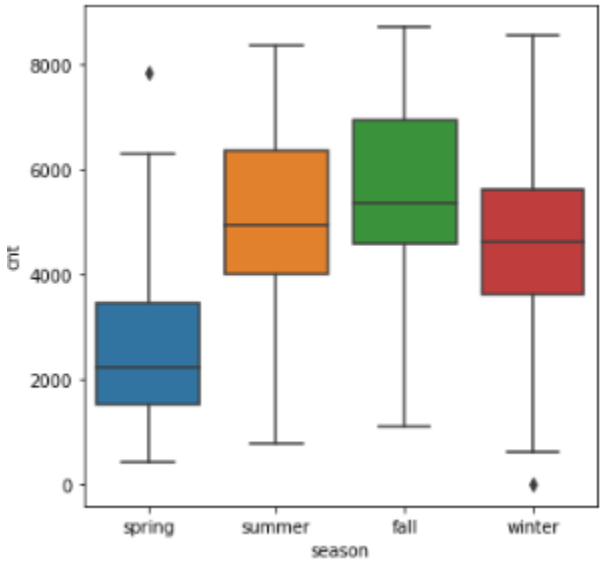
Assignment-based Subjective Questions

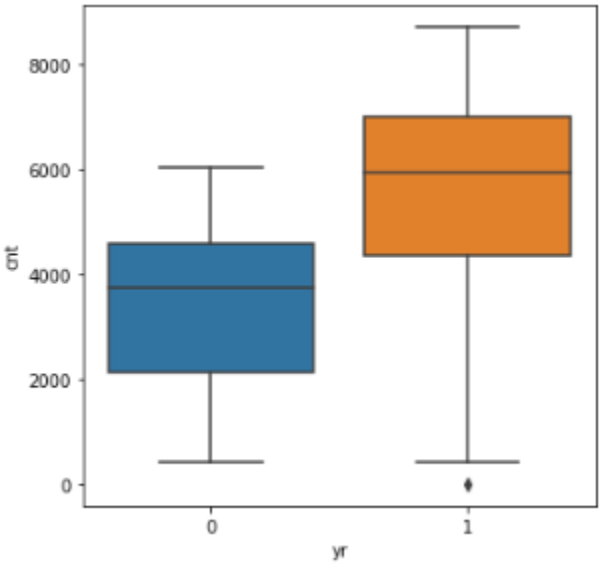
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Observations from the analysis of categorical variables 👍

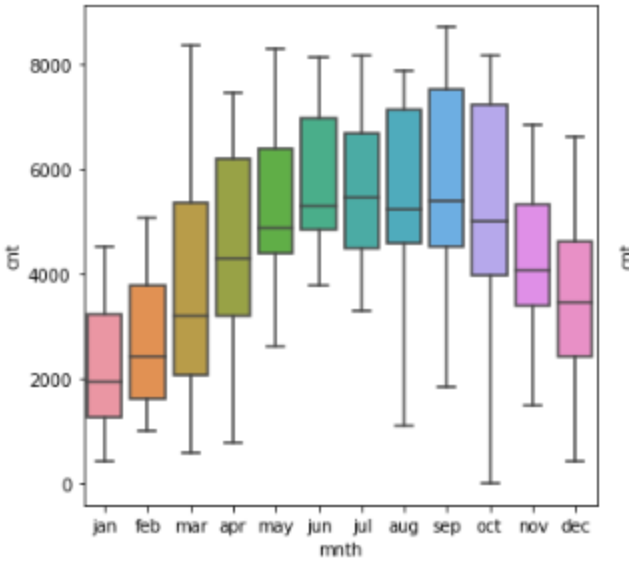
1. Demand for bikes is higher in fall season



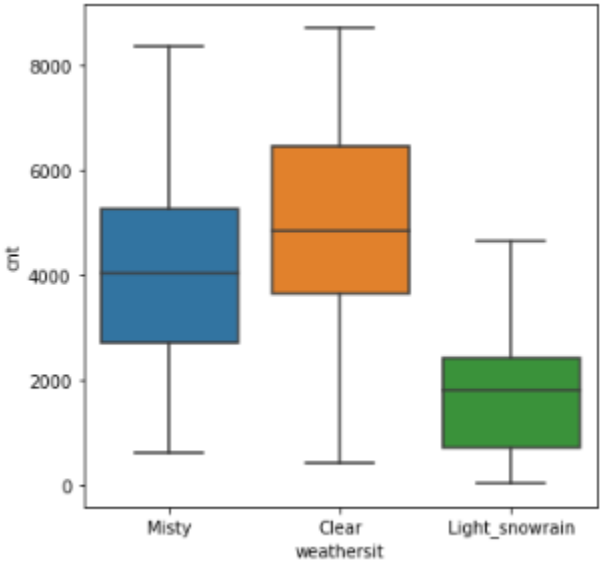
2. Demand for bikes increased significantly from 2018 to 2019.



3. Bike Demand increases initially when the year starts, tops the chart in June, almost remains high till September and then gradually decreases afterwards.



4. Bike Demand is high when weather is clear and very low in light rain



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

Generating dummy variables involves converting categorical variables into binary indicators (0 or 1) to represent different categories. However, including all the generated dummy variables can lead to multicollinearity. This occurs because one dummy variable's presence can be perfectly predicted from the others. By dropping the first generated dummy variable, we eliminate redundancy and mitigate multicollinearity issues.

Also by using drop_first=True, we maintain a clear interpretation of regression coefficients. The dropped dummy variable serves as the reference category or baseline for

comparison. The coefficients of the remaining dummy variables indicate the change in the response variable relative to the baseline category, enhancing interpretability.

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

Variable 'temp' has the highest correlation with the target variable .

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

I validated the assumptions of Linear Regression by the following methods:

Residual Analysis:Plot the residuals against the predicted values. If the residuals exhibit a random scatter around zero, it suggests that the linearity assumption is met.

Assessing Multicollinearity: Multicollinearity occurs when predictor variables are highly correlated. High multicollinearity can make coefficient estimates unstable and unreliable. Calculate the Variance Inflation Factor (VIF) for each predictor variable and check for values above a certain threshold (e.g., VIF > 5 or 10) to identify potential multicollinearity.

Error terms Distribution: Error Terms should be normally distributed.

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

These are the coefficients of the model :

const	- 0.0753
yr	- 0.2331
workingday	- 0.0563
temp	- 0.5499
windspeed	- 0.1552
summer	- 0.0874
winter	- 0.1318
sep	- 0.0972
sat	- 0.0677
Light_snowrain	-0.2880
Misty	- 0.0813

Among these the top 3 features contributing significantly towards explaining the demand of the shared bikes are : yr(0.2331) , temp(0.5499) , Light_snowrain(-0.2880)

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear regression is a popular statistical and machine learning algorithm used to establish a relationship between a dependent variable and one or more independent variables. Its purpose is to determine the best-fit straight line that represents this relationship. By assuming a linear connection between the independent and dependent variables, linear regression proves valuable in solving regression problems.

The linear regression algorithm involves the following steps in detail:

Data Preparation: To begin, you need a dataset comprising independent and dependent variable pairs. The independent variables, often called features, are the factors that influence the dependent variable, which is the target variable we aim to predict. Typically, the dataset is divided into a training set for model development and a test set for evaluation.

Model Representation: In linear regression, the objective is to find a linear equation that accurately represents the relationship between the independent variables (X) and the dependent variable (Y). This equation takes the form Y = b0 + b1X1 + b2X2 + ... + bn*Xn. Here, b0 represents the y-intercept (constant term), b1 to bn denote the coefficients (slopes), and X1 to Xn indicate the independent variables.

Cost Function: The next step involves defining a cost function that measures the model's prediction accuracy. The mean squared error (MSE) is commonly used in linear regression, computing the average squared difference between the predicted values and the actual values from the training set.

Optimization: The aim is to minimize the cost function and determine optimal coefficient values (b0 to bn) that yield the best-fit line. Achieving this involves leveraging an optimization algorithm like gradient descent. By iteratively adjusting the coefficients in the direction that minimizes the cost function, gradient descent facilitates convergence towards optimal coefficient values.

Model Training: With the optimization algorithm in place, the model is trained by feeding the training data into the algorithm. The algorithm calculates predicted values based on the current coefficient values and updates the coefficients using the optimization

algorithm. This process continues until the algorithm converges to optimal coefficients or reaches a predefined stopping criterion.

Model Evaluation: Once trained, the model is evaluated using the test data. Predicted values are generated based on the model, and its performance is measured using evaluation metrics like mean squared error, root mean squared error, or R-squared (coefficient of determination). These metrics assess the model's ability to generalize well to unseen data.

Prediction: Following training and evaluation, the model can be utilized for making predictions on new, unseen data. By inputting values for the independent variables, the model calculates the predicted value for the dependent variable based on the learned coefficients.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that share similar statistical properties but exhibit distinct patterns when visualized. Developed by statistician Francis Anscombe in 1973, the quartet underscores the importance of graphical exploration in data analysis and emphasizes the limitations of relying solely on summary statistics.

Each dataset within Anscombe's quartet possesses identical summary statistics for their x and y variables, including measures like mean, variance, correlation, and linear regression coefficients. However, when plotted, these datasets reveal strikingly different relationships and patterns, challenging the assumption that summary statistics alone provide a comprehensive understanding of the data.

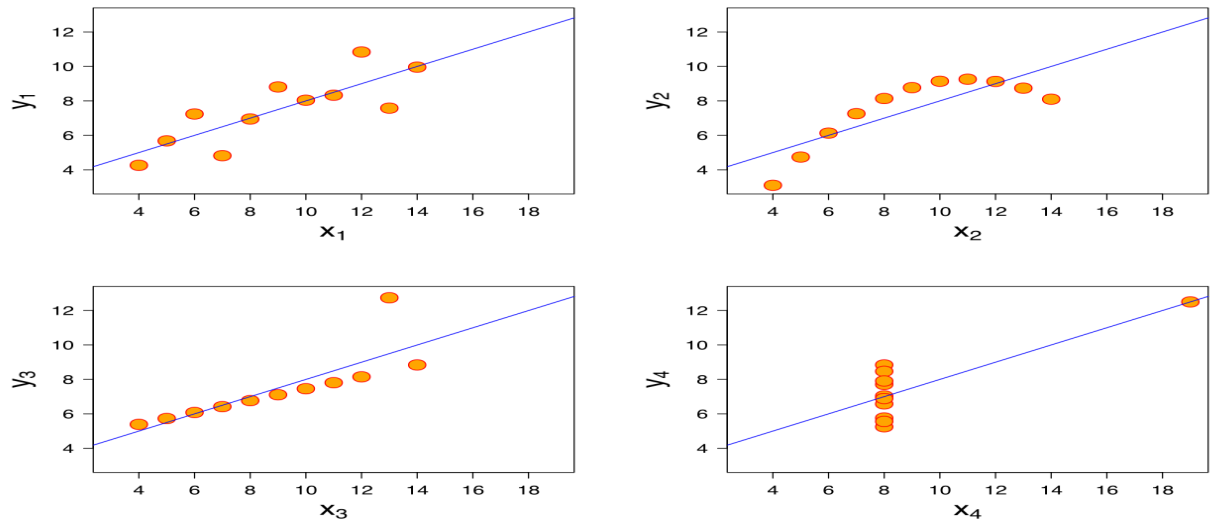


Image Source - Wikipedia

Let's delve into the characteristics of each dataset in Anscombe's quartet:

Dataset I showcases a linear relationship between x and y, with a clear upward trend. It follows the equation y = 3 + 0.5x and exhibits a strong correlation.
The data points tightly align with the linear trend, suggesting that a simple linear regression model would be suitable for analysis.
Dataset II:

Dataset II also demonstrates a linear relationship but with a negative slope. It contains an outlier, an individual point that significantly deviates from the overall pattern.
Despite the presence of the outlier, the summary statistics (mean, variance, correlation) closely resemble those of Dataset I.
Dataset III:

In contrast to the previous datasets, Dataset III showcases a non-linear relationship with a noticeable quadratic curve. The inclusion of an outlier substantially influences the regression line.
While the summary statistics may resemble those of the other datasets, a linear regression model would not effectively capture the underlying pattern.
Dataset IV:

Dataset IV features three distinct groups of points, each forming a linear relationship. However, when considered as a whole, it appears to exhibit a linear trend with a positive slope.
Summary statistics once again resemble those of the other datasets, but the unique grouping pattern within the data is not accurately captured by a simple linear regression model.
Anscombe's quartet underscores the significance of data visualization. It serves as a reminder that relying solely on summary statistics can obscure the underlying intricacies in the data. By graphically plotting the datasets, analysts can gain deeper insights, identify outliers, detect non-linear patterns, and select appropriate models for analysis.

The quartet highlights the importance of exploratory data analysis and visualizations as essential tools for comprehending data. It encourages researchers to question assumptions and promotes the use of graphical techniques to enhance understanding and generate well-informed conclusions about the data.

3. What is Pearson's R?

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson and is widely used in data analysis to assess the degree of association between variables.

Pearson's R ranges between -1 and +1. A value of +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally. Conversely, a value of -1 indicates a perfect negative linear relationship, where as one variable increases, the other variable decreases proportionally. A value of 0 indicates no linear relationship, suggesting that the variables are independent of each other.

To calculate Pearson's R, the following steps are typically followed:

Standardization: Standardize both variables by subtracting their respective means and dividing by their standard deviations. This transformation ensures that the variables have a mean of 0 and a standard deviation of 1.

Covariance Calculation: Multiply the standardized values of the two variables together and calculate their average. This yields the covariance, which measures the joint variability between the variables.

Standard Deviation Calculation: Calculate the standard deviation of each variable based on their standardized values.

Correlation Coefficient: Divide the covariance by the product of the standard deviations of the two variables. This gives Pearson's correlation coefficient (R), which represents the strength and direction of the linear relationship.

Pearson's R has several important properties. It is symmetric, meaning that the correlation between Variable A and Variable B is the same as the correlation between Variable B and Variable A. Additionally, it is sensitive to linear relationships but may not capture non-linear relationships. Therefore, it is essential to complement its analysis with other techniques when exploring complex associations. Interpreting Pearson's R involves considering its magnitude. Values close to +1 or -1 indicate a stronger linear relationship, while values closer to 0 suggest a weaker association.

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

Scaling refers to the process of transforming numerical features or variables to a specific range or distribution during data preprocessing. Its purpose is to ensure that all variables are on a comparable scale, enabling fair comparisons and avoiding biases in subsequent analyses.

Scaling is performed for several reasons:

Equalizing Variable Influence: When variables have different scales, those with larger scales can disproportionately influence machine learning algorithms. Scaling helps balance the influence of variables, preventing any single variable from dominating the learning process.

Enhancing Convergence: Many machine learning algorithms rely on optimization techniques that converge faster when features are on a similar scale. Scaling facilitates quicker convergence and improves the efficiency of the learning process.

Avoiding Numerical Instabilities: Certain algorithms, such as those based on distance calculations, are sensitive to the scale of variables. Scaling ensures that distances or similarities are appropriately represented, preventing numerical instabilities in these algorithms.

Facilitating Feature Interpretation: Scaling enables meaningful comparisons between features by ensuring that differences in feature values are not distorted by varying scales. This allows for easier interpretation and understanding of the relative importance of features.

There are two common methods of scaling: normalized scaling and standardized scaling:

Normalized Scaling (Min-Max Scaling):

Normalized scaling transforms features to a specific range, typically between 0 and 1. It preserves the relative relationships between data points.
The formula for normalized scaling is:
X_scaled = (X - X_min) / (X_max - X_min)
Here, X represents the original value, X_min is the minimum value in the feature, and X_max is the maximum value in the feature.

Standardized Scaling (Z-score Scaling):

Standardized scaling transforms features to have a mean of 0 and a standard deviation of 1. It brings the data closer to a standard normal distribution.

The formula for standardized scaling is:
X_scaled = (X - X_mean) / X_std
Here, X represents the original value, X_mean is the mean of the feature, and X_std is the standard deviation of the feature.

The primary difference between normalized scaling and standardized scaling lies in the resulting distribution of the scaled data. Normalized scaling retains the original distribution and range of the data, while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1, aiming for a standard normal distribution.

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

The formula for calculating the VIF of a predictor variable is:

VIF = 1 / (1 - R^2)

The VIF (Variance Inflation Factor) measures how much the variance of a regression coefficient is inflated due to multicollinearity, which is when predictor variables are highly correlated. When there is perfect multicollinearity, it means that one or more predictor variables can be exactly predicted by a combination of the other variables.

In the case of perfect multicollinearity, the VIF formula leads to an infinite value because it involves dividing by zero. This happens when the coefficient of determination (R^2), which represents the strength of the linear relationship between predictor variables, becomes 1.

Having an infinite VIF means that it is impossible to accurately estimate the regression coefficients because the relationship between the correlated variables cannot be untangled. The issue arises because the coefficient estimates rely on the inverse of the covariance matrix of the predictors, which becomes non-invertible when perfect multicollinearity is present.

To handle perfect multicollinearity, it is necessary to identify and resolve the linear relationships between predictor variables. This can involve actions like removing one of the correlated variables, transforming variables, or reconsidering the model specification.

It's important to note that while an infinite VIF value indicates perfect multicollinearity, even high VIF values (though not infinite) for predictor variables can indicate significant multicollinearity. In such cases, it is advisable to address the multicollinearity to ensure reliable regression analysis and proper interpretation of the model coefficients.

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

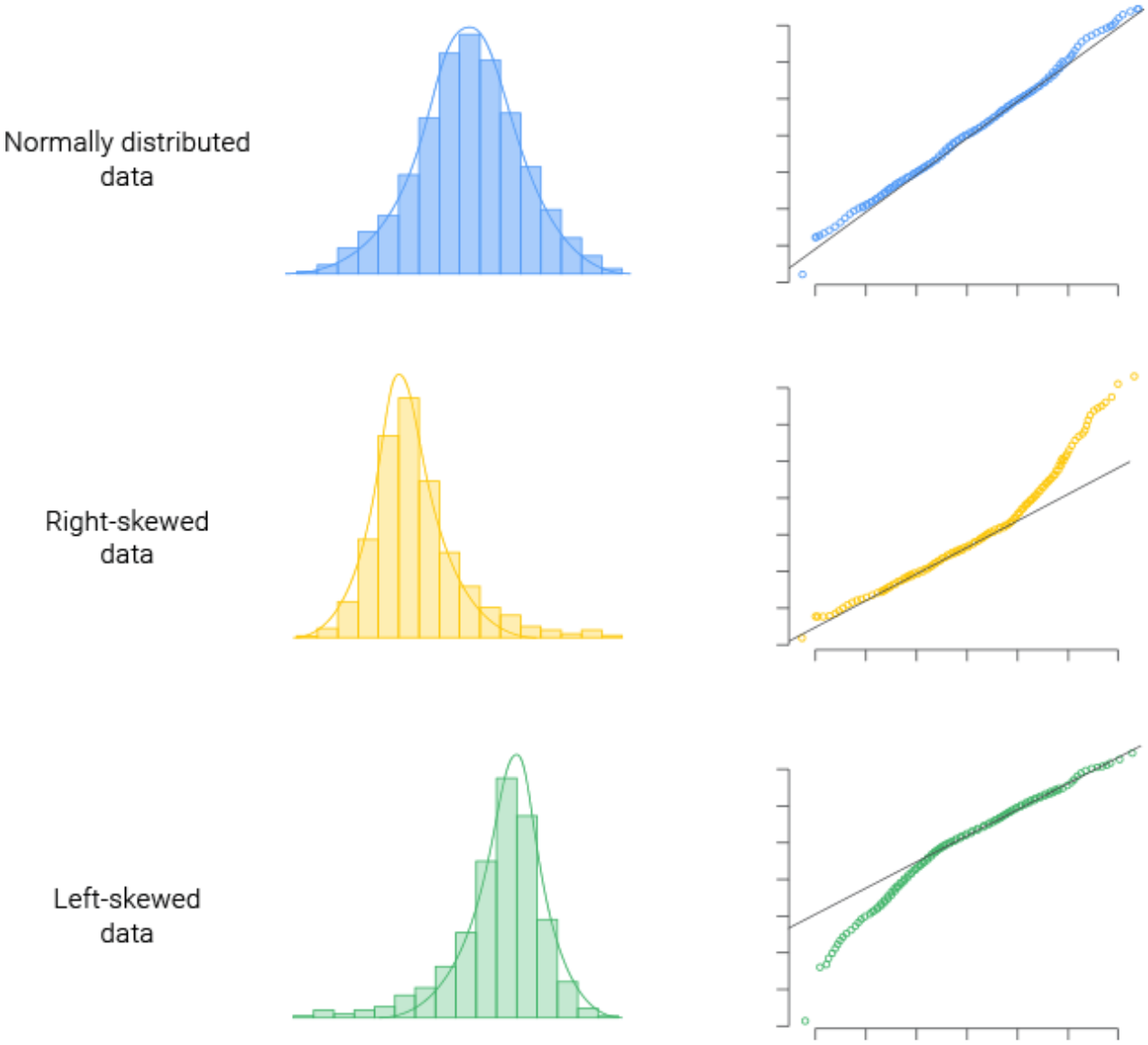


Image Source- https://www.learnbyexample.org/r-quantile-quantile-qq-plot-base-graph/

A Q-Q plot (Quantile-Quantile plot) is a graphical tool that compares the quantiles of a dataset with the quantiles expected from a theoretical distribution. It is commonly employed in linear regression to evaluate whether the residuals (the differences between observed and predicted values) follow a normal distribution, which is an important assumption in regression analysis.

The use and significance of a Q-Q plot in linear regression are as follows:

Checking Normality Assumption: Linear regression assumes that the residuals are normally distributed. A Q-Q plot visually examines this assumption by plotting the observed residuals against the quantiles expected from a normal distribution. If the points on the plot form a roughly straight line, it suggests that the residuals are normally distributed. Deviations from the straight line indicate departures from normality.

Assessing Skewness and Tails: In addition to normality, a Q-Q plot can reveal information about the skewness (lack of symmetry) and tails of the distribution. If the plot deviates from a straight line in the middle, it indicates skewness in the data. Similarly, deviations at the ends suggest heavy or light tails in the distribution.

Detecting Outliers: Outliers can significantly influence the regression model's results. A Q-Q plot helps identify outliers by spotting data points that deviate substantially from the expected line. These points may indicate potential issues like influential observations or violations of the normality assumption.

Validating Model Assumptions: By examining the Q-Q plot, we can validate whether the residuals meet the assumption of normality. If the assumption is violated, it may affect the reliability of statistical inference, confidence intervals, and hypothesis testing associated with the linear regression model. The Q-Q plot serves as a diagnostic tool to validate assumptions and guides us in determining whether transformations or alternative models are necessary.

In summary, a Q-Q plot is an effective tool in linear regression for evaluating the normality assumption of residuals. It enables the comparison of observed data with a theoretical distribution, such as the standard normal distribution. By analyzing the plot, we can detect departures from normality, identify skewness and tails, detect outliers, and validate the assumptions underlying the linear regression model.