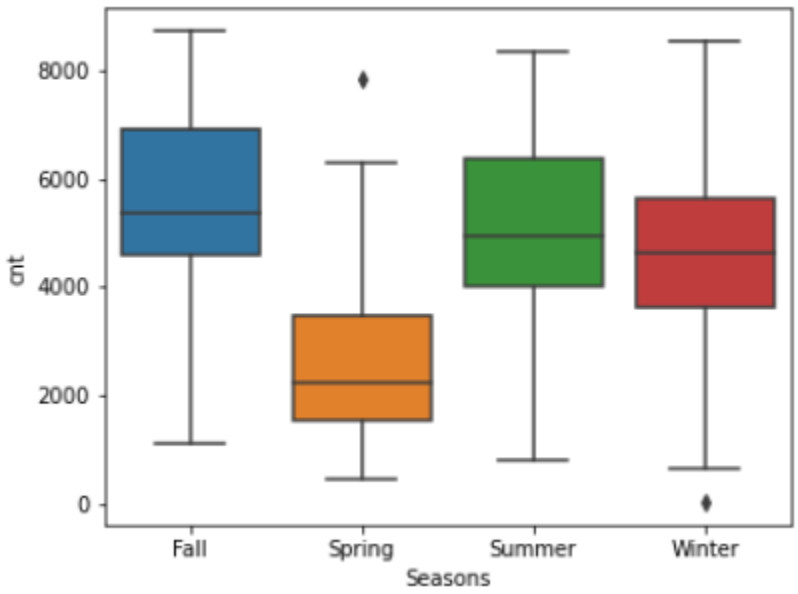
Assignment-based Subjective Questions

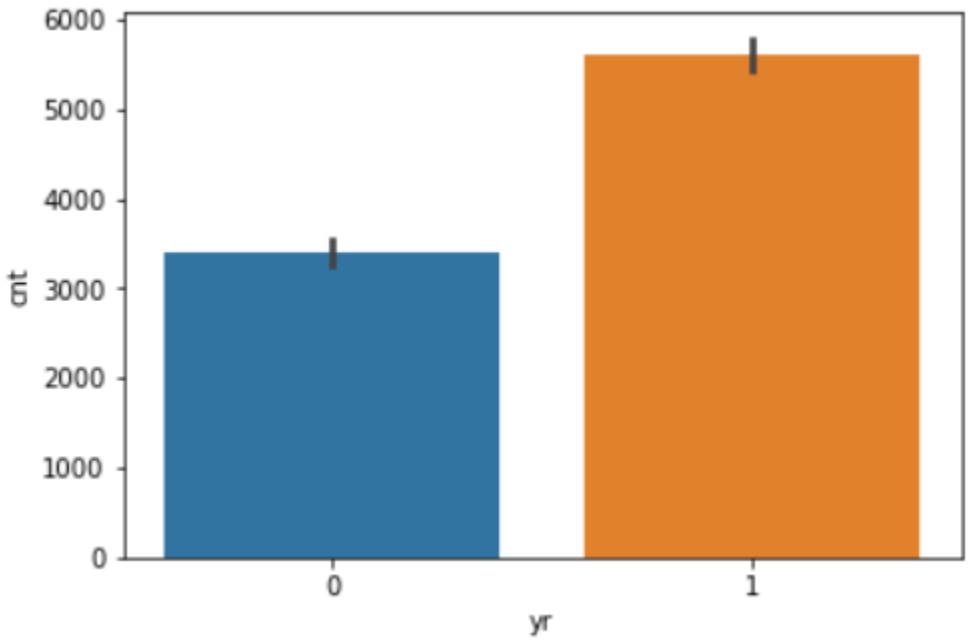
Question: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: The categorical variables in our day.csv dataset is season, yr, mnth, holiday, weekday, workingday, weathersit.

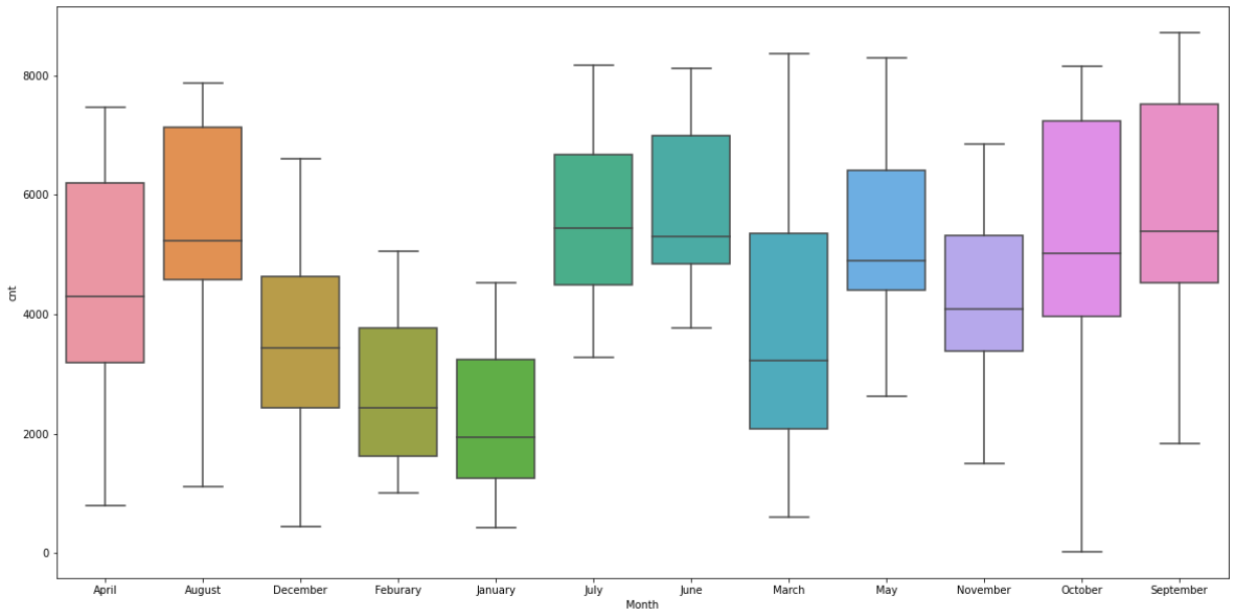
- **Season:** The boxplot shows that Spring has the least bike rentals, when compared to Fall & summer where the bike rentals were higher.



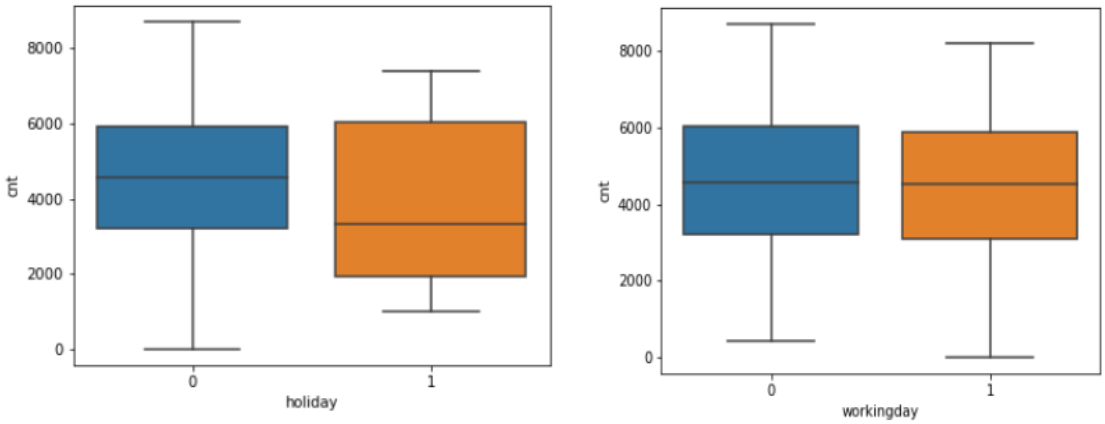
- **Year:** The boxplot shows that 2019 has better rentals than 2018



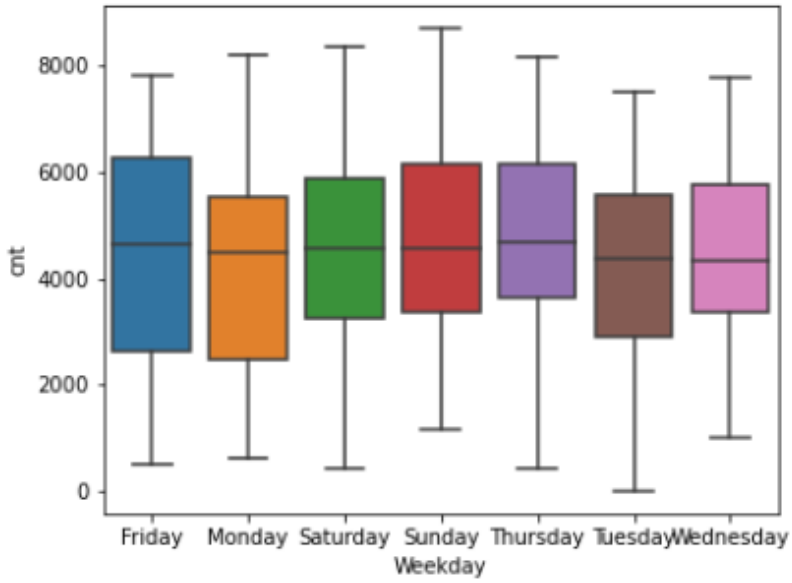
- **Month:** From the boxplot we could infer that September has the highest bike rentals whereas January has the least bike rentals.



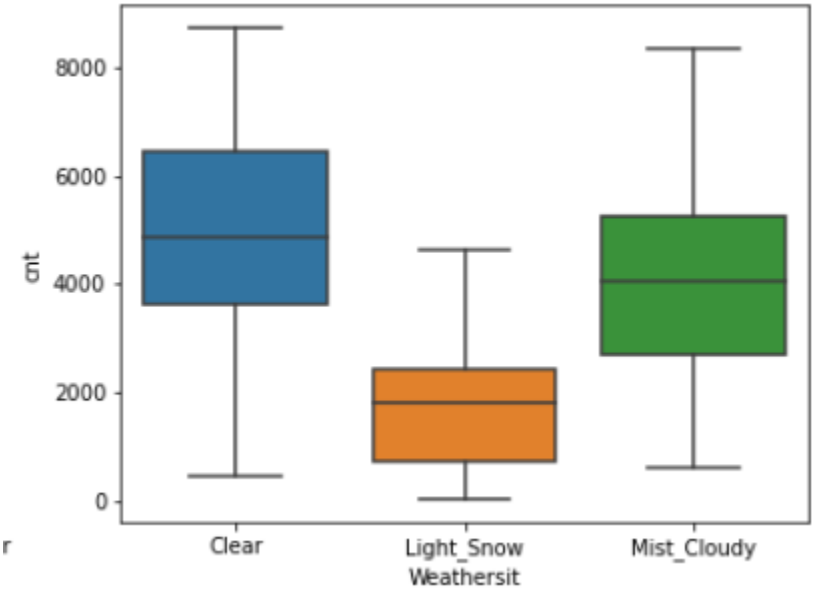
- **Holiday & Workingday:** From the boxplot we could conclude that during working days the rentals were higher & during holidays the bike rentals were relatively lower.



- **Weekday:** From the boxplot we could conclude that bike rentals were higher on Fridays when compared to Tuesdays where its lower.



- **Weathersit:** From the boxplot we can infer that the public prefers riding bikes when the sky is clear and would hardly prefer cycling when there is light snow.

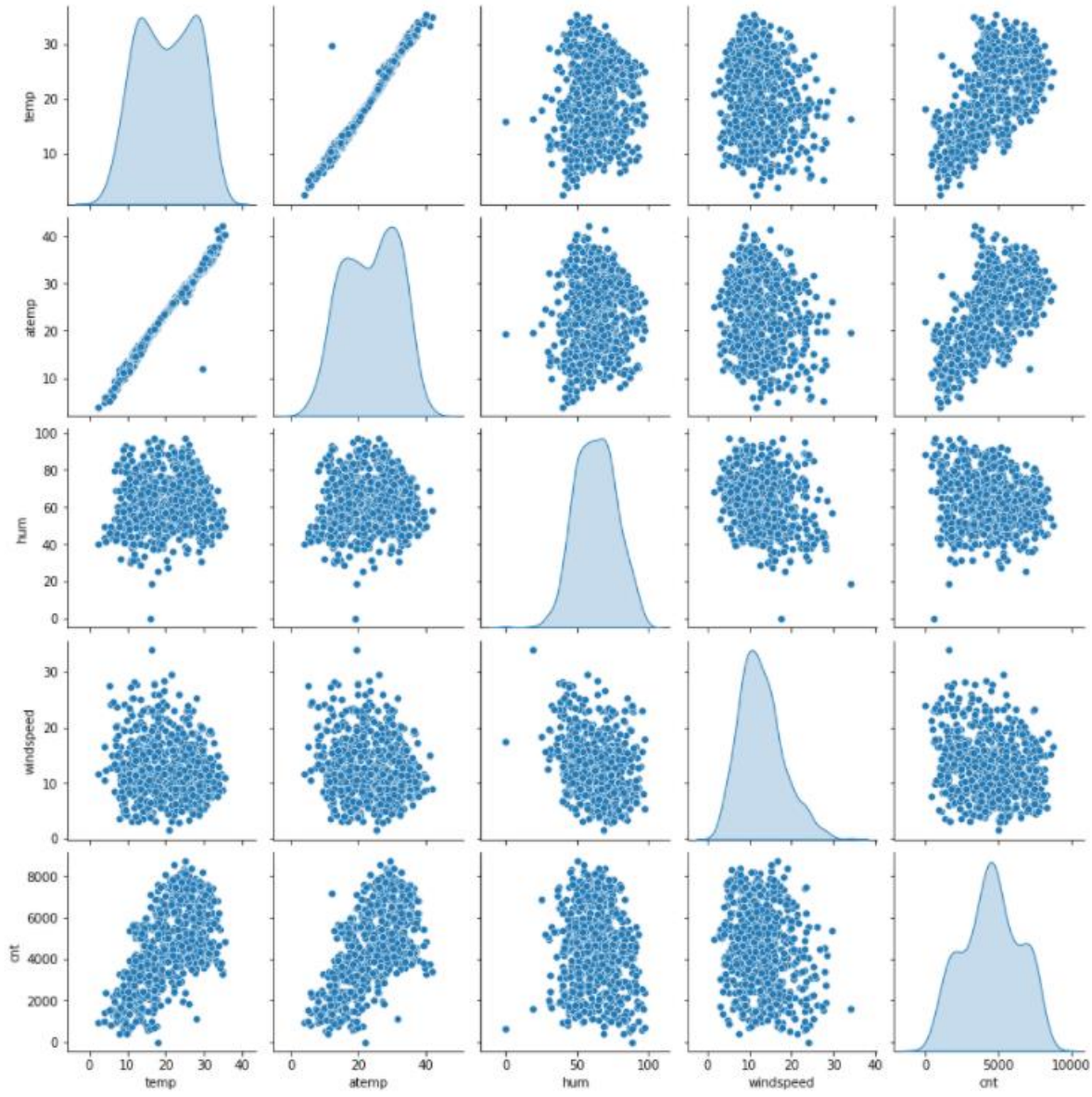


Question: Why is it important to use **drop_first=True** during dummy variable creation?

Answer: In a categorical column there might be ‘n’ levels of variables then we need to use ‘n-1’ columns to represent the dummy variables. Another reason being having all the dummy variables lead to multicollinearity between dummy variables.

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

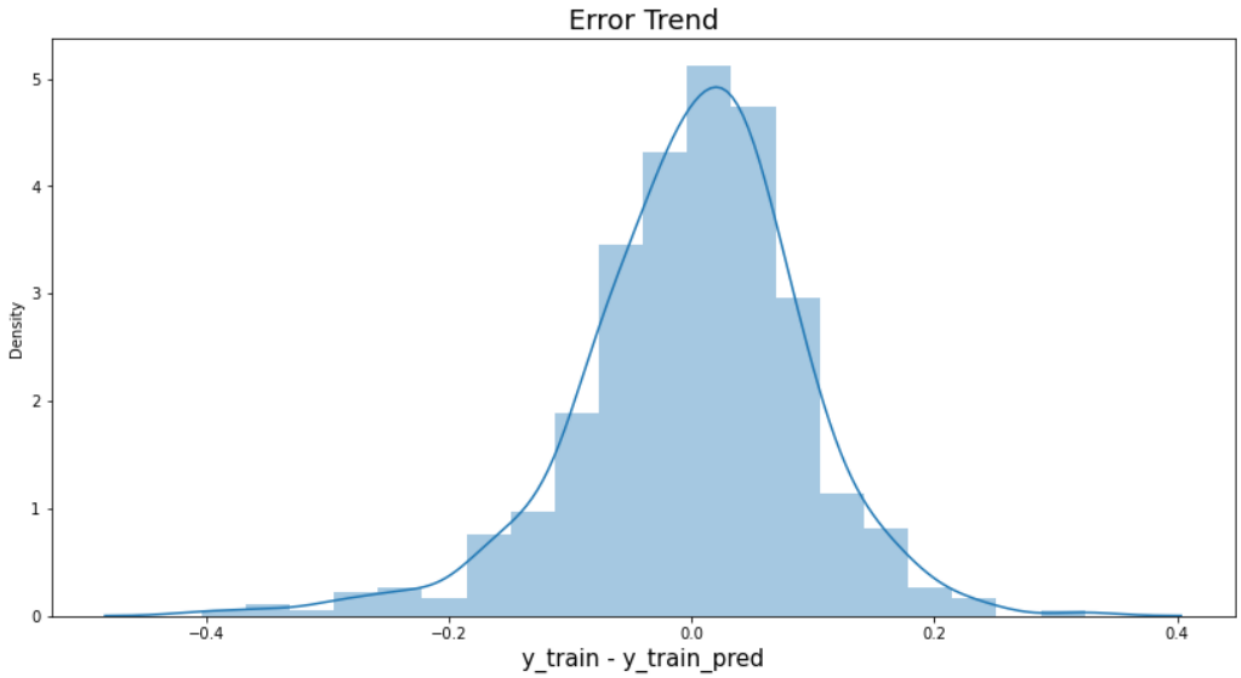
Answer: From the pair-plot we could see that “temp” and “atemp” has the highest correlation with target variable (“cnt”).



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

Answer: Linear Regression is an analysis, measures whether one or more predictor variable predicts the dependent variable. Linear Regression is validated from the results that we obtained from our Residual analysis on the error trends by plotting histogram. If the error trends are centered towards '0' i.e. (normally distributed), the error values must be independent of each other. Scores on a dependent variable can be represented as the sum of independent variable + error $Y_i = m_i X_i + \varepsilon_i$.

From the below histogram we can clearly see that the output of the Error trends are normally distributed.



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

Answer: The top three features contributing significantly towards the demand of the shared bikes are:

Temp: 1.50

Month (July): 1.42

Season (Summer): 1.33

General Subjective Questions:

Question: Explain the linear regression algorithm in detail.

Answer: Linear Regression algorithm is a machine learning algorithm that is based on supervised learning. Regression models a target prediction based on independent variables. Linear Regression is mostly used for finding out relationships between variables and forecasting. There are two types of Regression:

a) Simple Linear Regression: As the name says we will have only one independent variable (easy to find the relationship) and the target variable
b) Multiple Linear Regression: Here we will have two or more independent variables (hence multiple) against the target variable.

We classify the independent variables as "predictor variables" and the dependent variables as "target" variables.

Using linear regression we can find out the equation for the best fit line using Y= mx+c by minimizing the residual sum of squares (RSS)

Basic assumptions of Linear Regression:

- There must be a linear and additive representation between the predictive and the target variable
- There must be no correlation between the residual (error) terms. This is also called as "Auto correlation".
- The independent variables should not be correlated.
- The error terms must be normally distributed and must have constant variance.

Hypothesis: We perform Hypothesis tests to find out whether the fitted line is significant. We use H_0 for Null hypothesis and H_a for Alternate hypothesis.

We could formulate this as $H_0 = 0$ and $H_a \neq 0$ as Null and Alternate hypothesis. We can do this by checking the P-value (Probability value) method. If P is greater than the significance level then we fail to reject null hypothesis, if P is lesser than the significance level then we reject the null hypothesis. The optimum significance level for a P-value is below 5%

Steps for building Linear Regression:

- Load the data
- Perform EDA (Exploratory Data Analysis)
- Build a train and test model using the 70:30 ratio
- Generate the model
- Evaluate the model accuracy by running it on the testing split.

Question: Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet contains four dataset that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of 11 (x,y) points. This model was constructed by a statistician Francis Anscombe to highlight both the importance of graphical data before analyzing and the effect of the outliers on statistical properties.

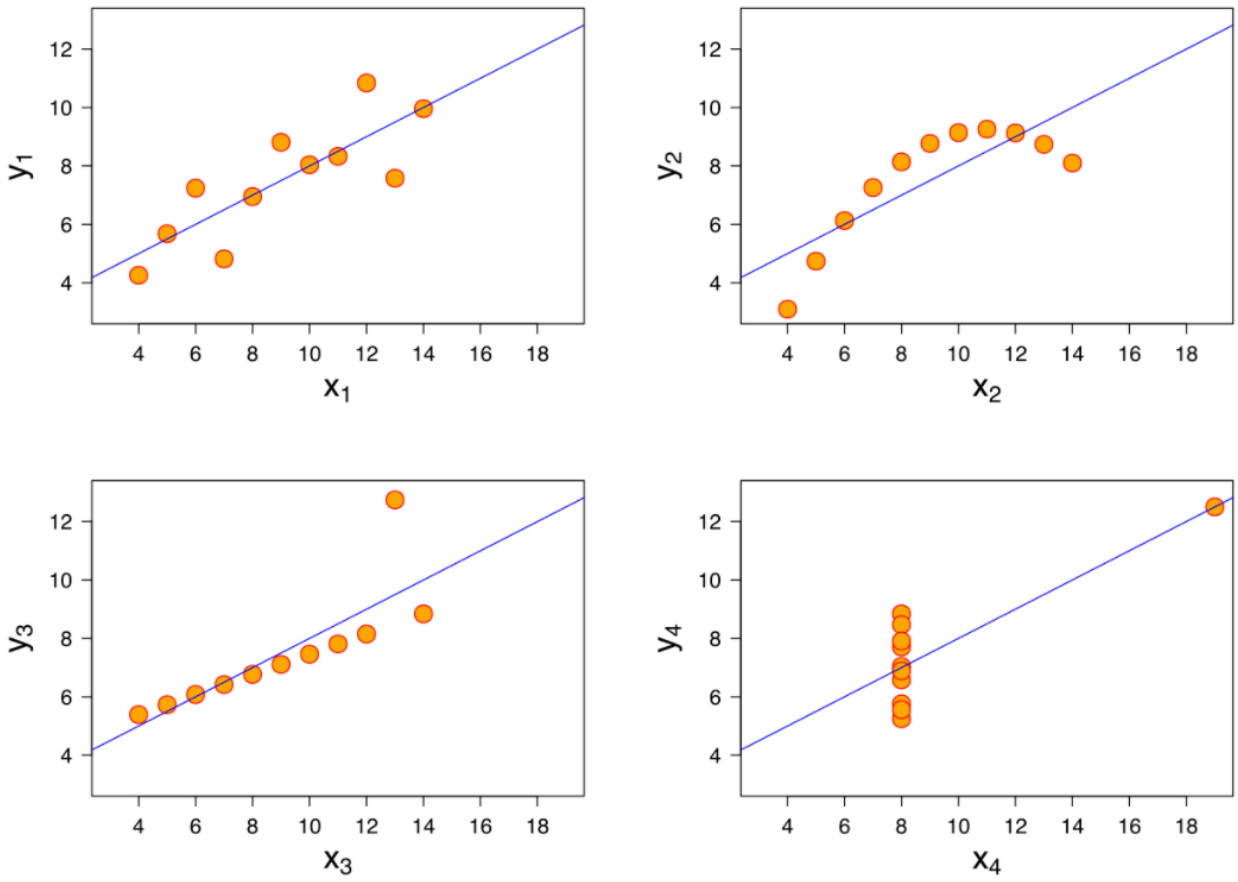
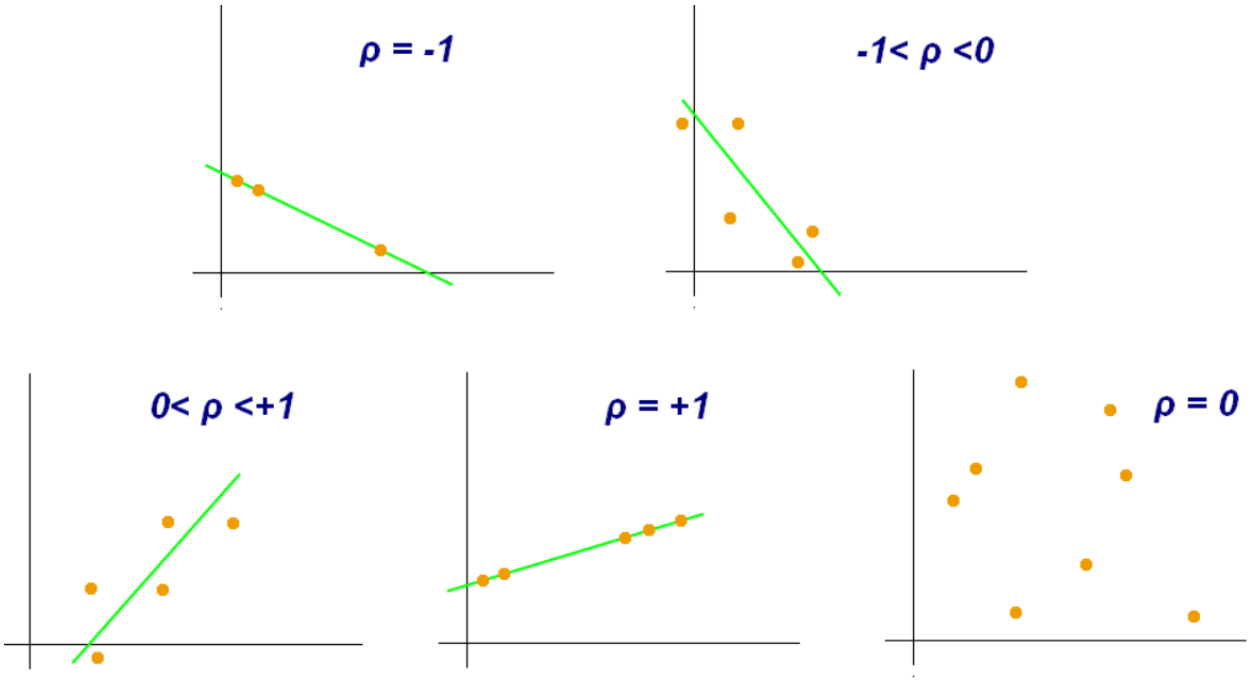


Image source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe's_quartet_3.svg

- The first scatter plot (top left) seems to be a simple linear model that corresponds to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second scatter plot (top right) is not normally distributed but that is a relation between the two variables is obvious. Here Pearson coefficient is not relevant
- The third scatter plot (bottom left) is linear but should have a different regression line.
- Final scatter plot (bottom right) illustrates that one high-leverage point is enough to produce high correlation coefficient, even when do not have any relation between the variables.

Question: What is Pearson's R?

Answer: Pearson's R or Pearson correlation coefficient is a measure of linear correlation between two sets of data. We can represent it as the covariance of two variables divided by the product of their standard deviations. Hence the result always have a value between -1 and 1. This was developed by Karl Pearson, from a related idea introduced by Francis Galton in 1880.



https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#/media/File:Correlation_coefficient.png

For a population:

When Pearson's coefficient is applied to a population commonly represented by ρ the population correlation coefficient. The formula:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#:~:text=In%20statistics%2C%20the%20Pearson%20correlation,between%20two%20sets%20of%20data.

Question: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling is a technique to standardize the independent features that are available in a fixed range. If scaling is not done, the machine learning algorithms tends to weigh higher and smaller values regardless of the unit of the values.

Scaling is important to bring all the variables to the same magnitude level. If scaling is not done then the algorithm takes magnitude into account and would ignore the units. Hence we would see incorrect results.

Techniques to perform scaling:

- **Min-Max normalization**: re-scales the observation value with distribution value between 0 & 1
- **Standardization**: effective technique that rescales the featured value that has distribution with 0 mean values and variance is always 1.

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

Answer: Variance Inflation factor (VIF) detects multicollinearity in regression analysis. Having VIF can adversely affect the regression results. VIF estimates that how much variance of a regression coefficient is inflated due to multicollinearity in the model. VIF formula:

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

Infinite VIF means, the independent variables are perfectly correlated. If the VIF is large we need to take corrective measures ahead of performing the Multiple regression. We can do this by dropping the infinite columns and re-run the model.

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

Answer: Q-Q or Quantile-Quantile plot, is a graphical tool to assess if a dataset was derived from a normal, exponential or a uniform distribution. It also determines if two datasets come from populations with a common distribution. A Q-Q plot is a plot of quartiles of the first data set against the quartiles of the second data.

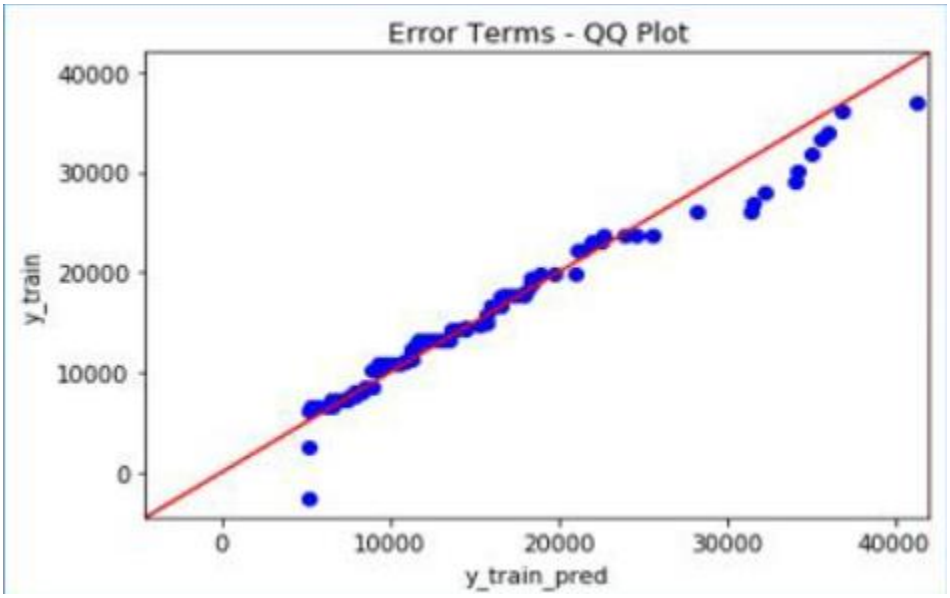
Advantages of Q-Q plot:

- We can use Q-Q plots with sample data
- We can also detect the presence of outliers from this plot.

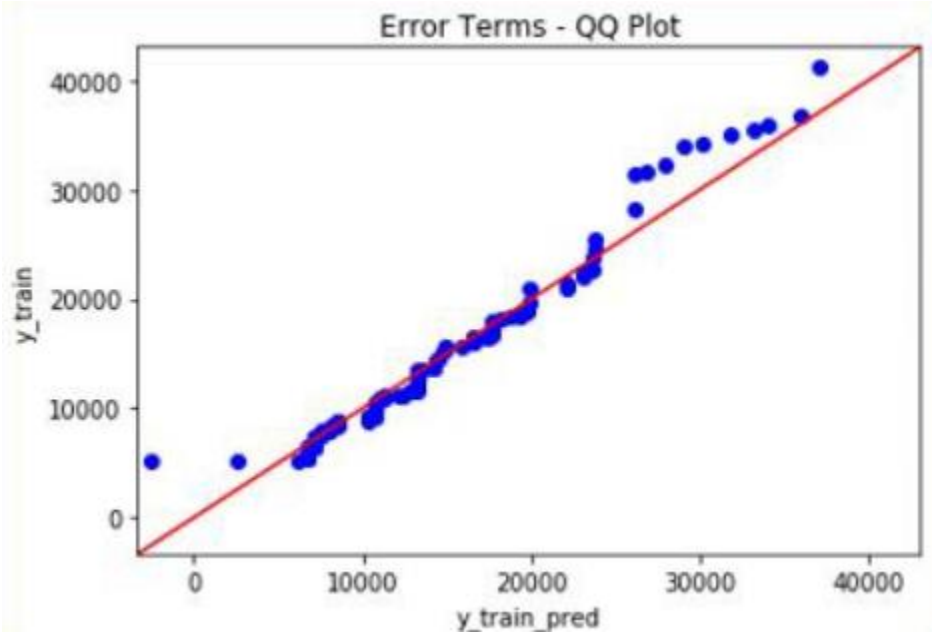
It is also used to check if two datasets come from the populations with a common distribution, have a common scale, have similar shape and have a similar behavior

Possible Interpretations:

- **Similar distribution:** if all point of quartiles forms close to a straight line at an angle of 45 degrees from x axis.
- **Y-values < X-values:** If the Y quartile is lesser than the X quartile



- **X-values < Y-values:** If X quartile is lower than Y quartile



- **Different distributions:** If most of the points lie away from the straight line at an angle of 45 degrees from x-axis