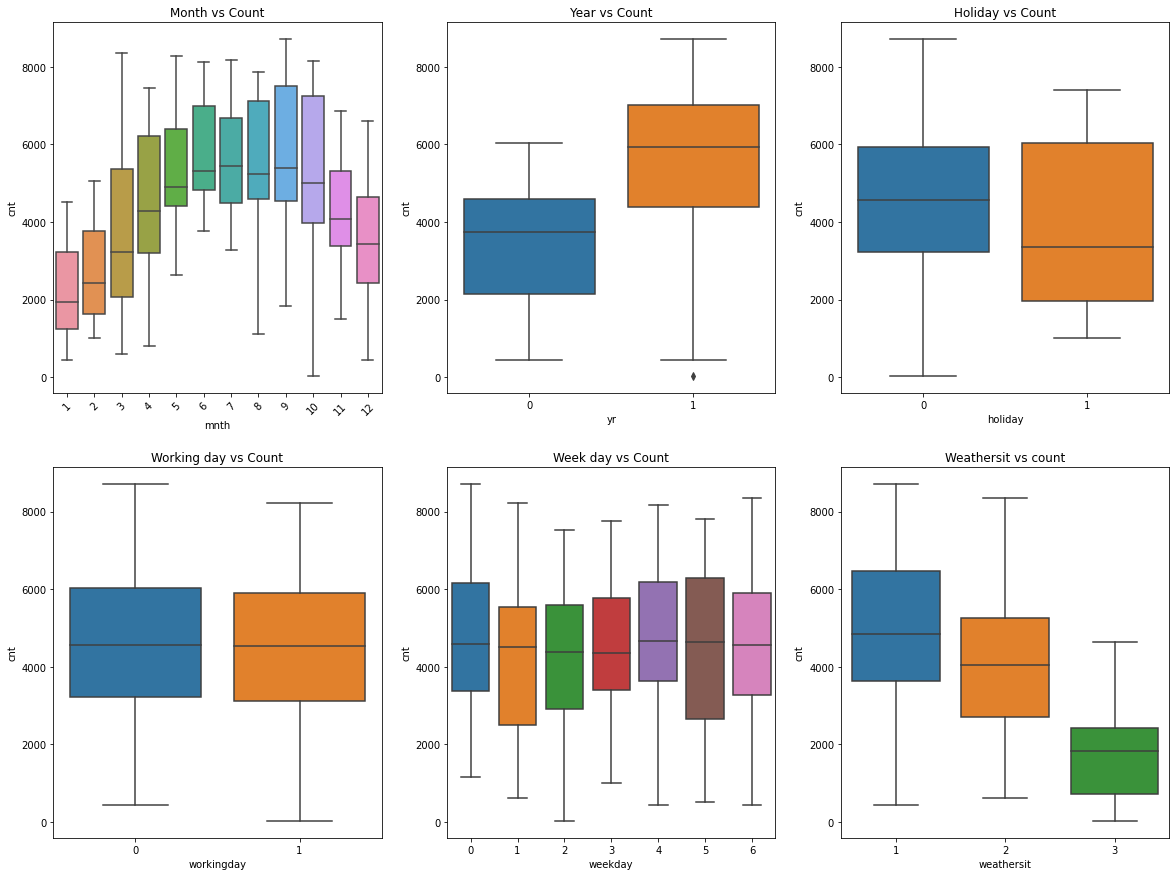
**Assignment-based Subjective Questions**

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

Categorical variables like Season, year, month, working day, weekday, weathersit etc has an impact on the target variable.

When plotting a boxplot against cnt we can visualize these impacts.



1. Why is it important to use drop\_first=True during dummy variable creation?

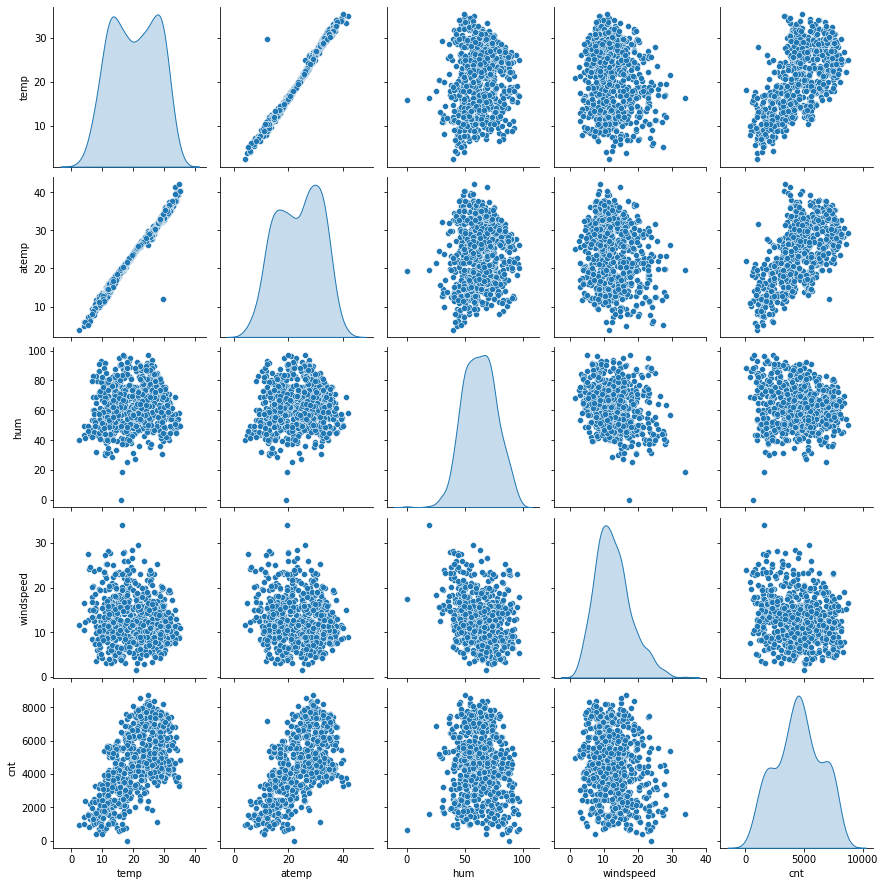
It helps in reducing the extra variable created during dummy variable creation, for n columns, only n-1 dummy variables are needed. Hence reducing the correlation between the dummy variables.

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

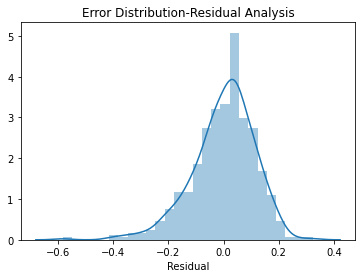
Variables temp and atemp have the highest correlation with cnt variable.

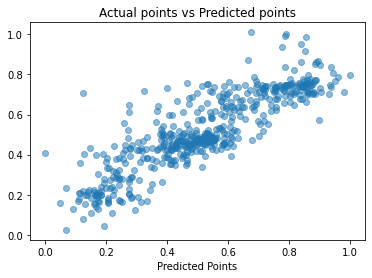
1. How did you validate the assumptions of Linear Regression after building the model on the training set?

* Linearity – There should be linear relationship between dependent and independent variable.



* Normal Error Distribution – The Error term is distributed normally around zero



* Homoscedasticity - Constant Error Variance, i.e, the variance of the error term is same across all values of the independent variable.

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

season\_Fall, mnth\_May,mnth\_June have the highest coefficients for the multiple linear regression model we’ve built. Thus we can conclude that the demand for shared bikes are more during the season of Fall and months of May and June.

The company can investigate contributing factors in these season(s) and months so as to increase the hiring of shared bikes.

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

**Linear Regression** is a machine learning algorithm based on **supervised learning**.

It performs a **regression task**. Regression models a target prediction value based on independent variables.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

The equation for Linear Regression is:

y=β₀+β₁X₁+β₂X₂+….

Here β₀ is the intercept,

Xi represents the different independent variables

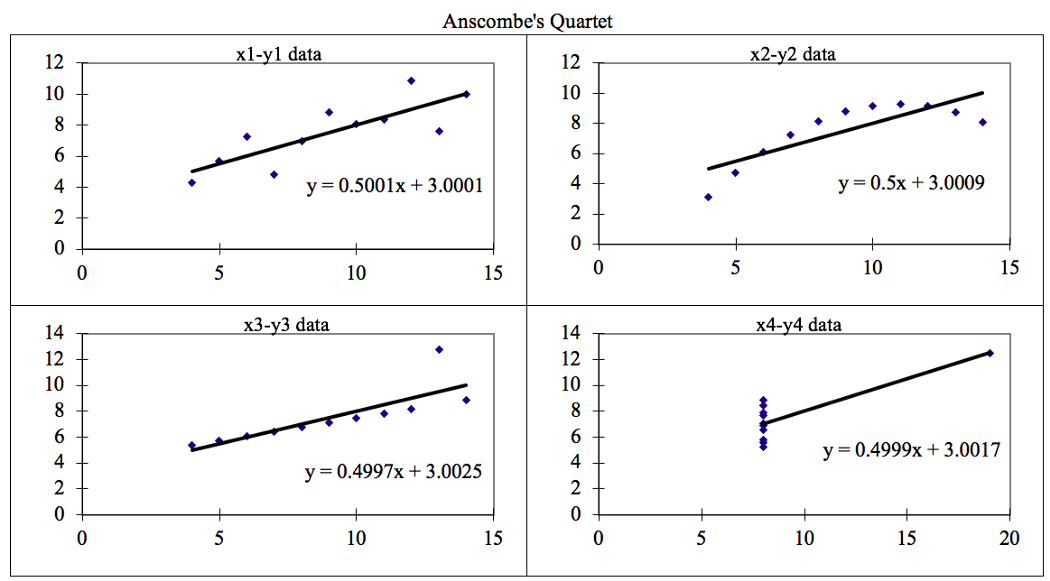
y is the target variable i.e. predicted using the linear regression model.

The linear model changes value of βi such that the error between the predicted y and actual target variable is minimized.

1. Explain the Anscombe’s quartet in detail.

**Anscombe’s quartet** comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

When a linear regression model is built with these 4 datasets, whose descriptive statistics show a very similar trend even though, when plotted, shows a completely different story.



**Anscombe’s quartet is used to show the importance**of **plotting the graphs**before analyzing and model building, and the effect of other **observations on statistical properties**. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean**of all (x, y) points in all four datasets.

1. What is Pearson’s R?

**Pearson’s correlation** (also called Pearson’s R) is a **correlation coefficient** commonly used in [linear regression](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/). **It is** used to measure how strong a relationship is between two [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/).

It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.  It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

Feature Scaling is performed when we are dealing with Gradient Descent Based algorithms like Linear and Logistic Regression as these are very sensitive to the range of the data points.

Standardization: This technique transforms the data to have a mean of zero and a standard deviation to 1.

Normalization: This technique transforms the values in variables between 0 and 1.

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

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

A large value(infinity) of VIF indicates that there is a correlation between the variables.

A general rule of thumb is that if VIF > 10 then there is multicollinearity.

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

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight.

The use of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.