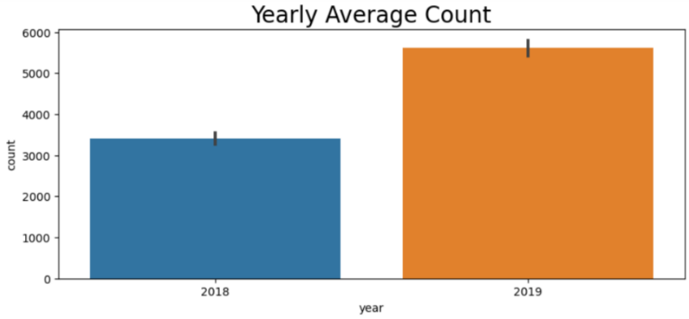
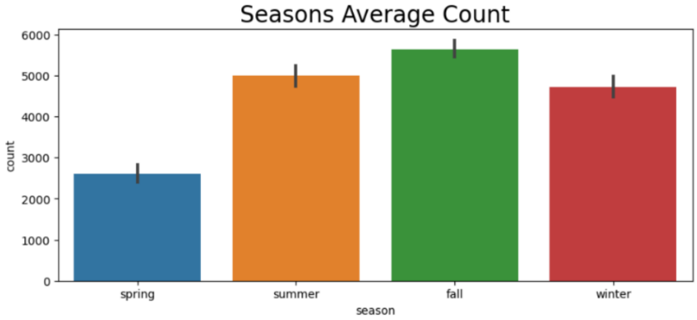
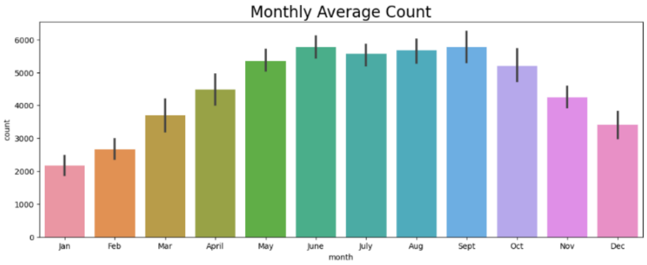
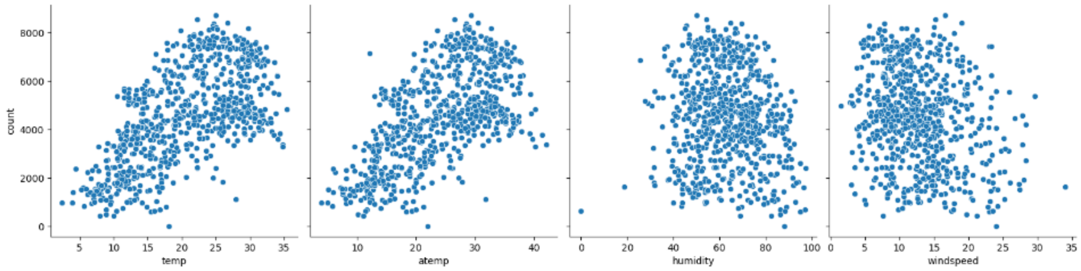
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   In the given Bike data set a couple of categorical variables namely season, mnth, yr, weekday, and weathersit
   The below scatter plots/bar plots show the correlation between the variables.



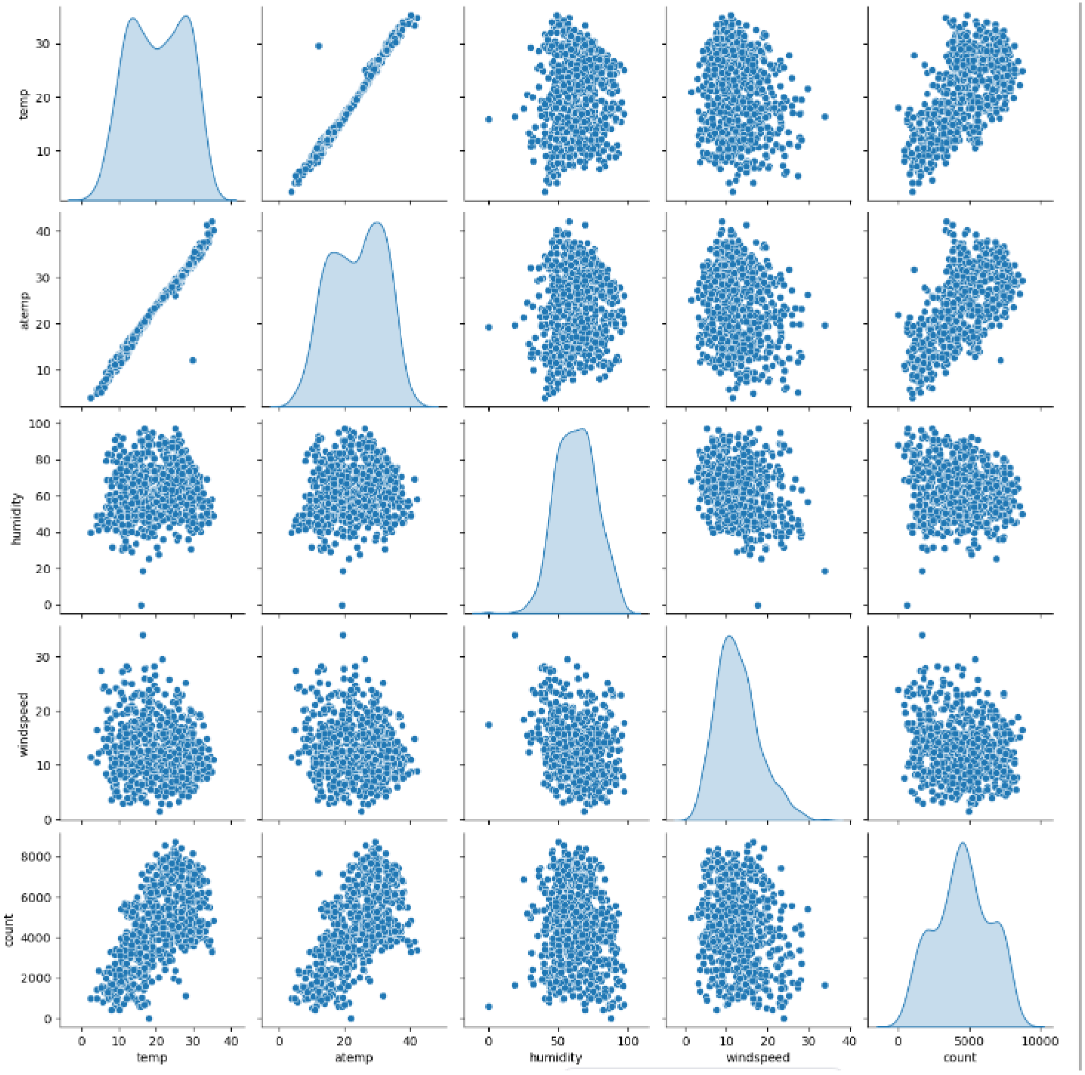






2. **Why is it important to use drop_first=True during dummy variable creation?**
   The intention behind the creation of the dummy variables for the categorial variable of ‘n’ levels you create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. Hence drop_first = true is used so that the resultant can match up n-1 levels. It helps in reducing extra columns created during dummy variable creation.
   Hence it reduces the correlation among the dummy variables.
   E.g. If there are 4 levels, the drop_first will drop the first column.

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

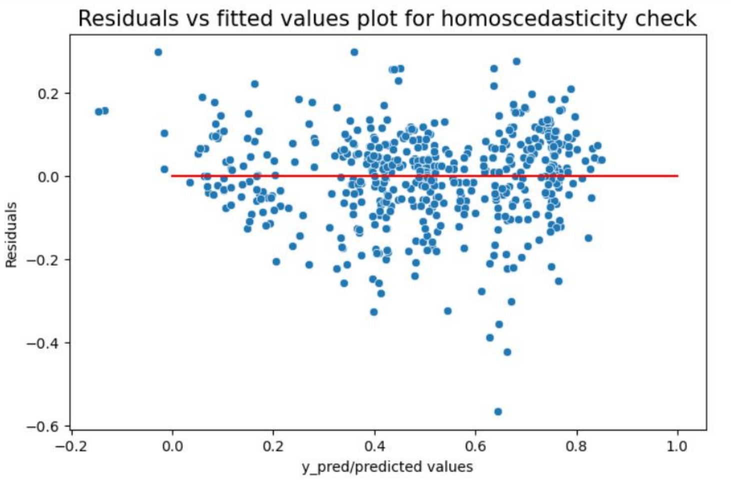


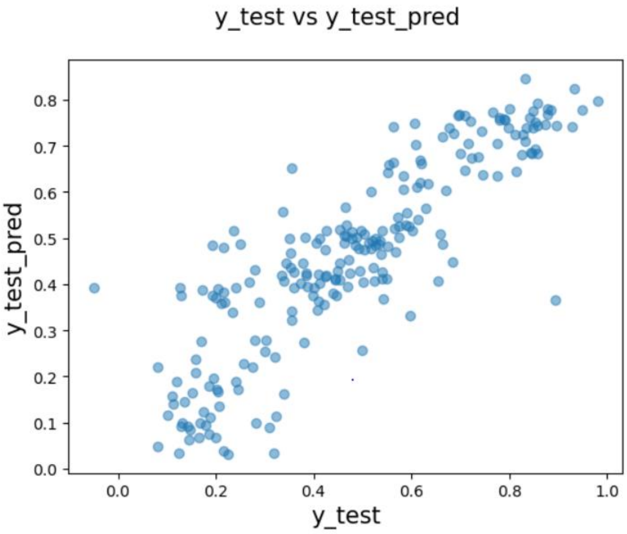
‘temp’ and ‘atemp’ have the highest correlation when compared to the rest of the variables with the target variable ‘cnt’

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

Linear regression models are validated based on linearity, no autocorrelation, normality of error, homoscedasticity, and multicollinearity.

The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.





5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?**
Year, wind speed, and weather are the top 3 features contributing significantly.
Year 2019 bike rentals increased by almost double the amount in 2018 post-pandemic effect.
Windspeed also influences the bike rentals. If wind speed is more bike rentals are less.
Weather also influences bike rentals, Good and clear weather increases bike rentals. If the Weather is moderate/Misty also has a good amount of bike rentals.

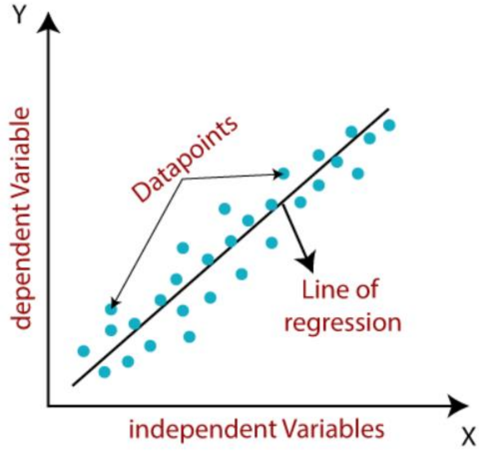
General Subjective Questions

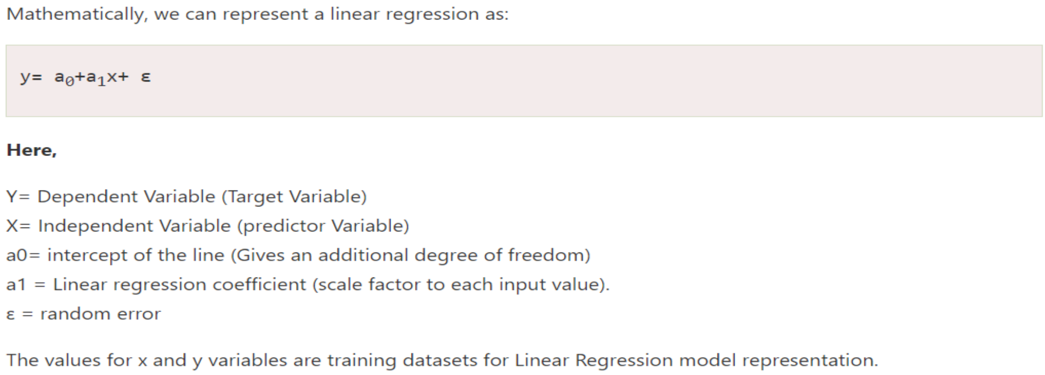
1. **Explain the linear regression algorithm in detail.**
Linear regression is a statistical method that is used for predictive analysis. Linear regression makes predictions for the continuous/real or numeric variables.
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. The algorithm shows how the value of the dependent variable changes according to the value of the independent variable.

The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.As shown below





Mathematically, we can represent a linear regression as:

```
y= a0+a1x+ ε
```

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

A regression line can be a positive linear relationship or a negative linear relationship. The goal of the linear regression algorithm is to get the best value for a0 and a1 to find the best-fit line and the best-fit line should have the least error.
In Linear Regression, RFE or Mean squared error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best-fit line for the data points.

Assumptions of Linear Regression:

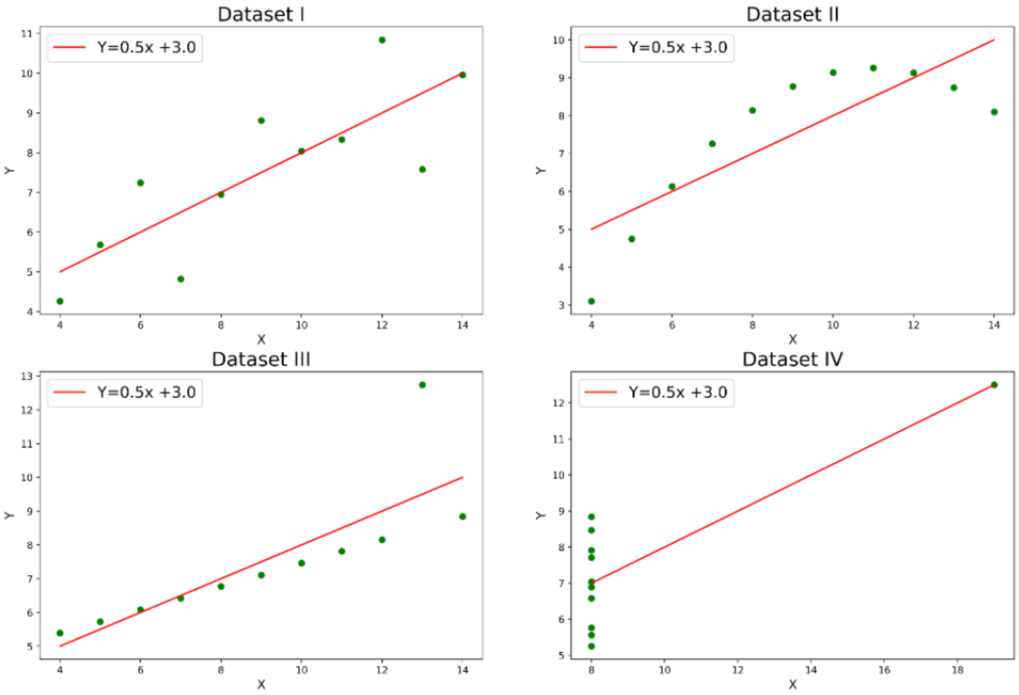
- Linear regression assumes the linear relationship between the dependent and independent variables.
- Small or no multicollinearity between the features.
- Homoscedasticity assumption - Homoscedasticity is a situation when the error term is the same for all the values of independent variables.
- Normal distribution of error terms.
- No autocorrelations.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's Quartet, comprising four datasets with nearly identical summary statistics, underscores the limitations of relying solely on numerical metrics. However, some peculiarities fool the regression model once you plot each data set. As you can see, the

data sets have very different distributions, so they look completely different from one another when you visualize the data on scatter plots.
Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

We can define these four plots as follows:



Anscombe’s quartet Plot

Explanation of this plot:

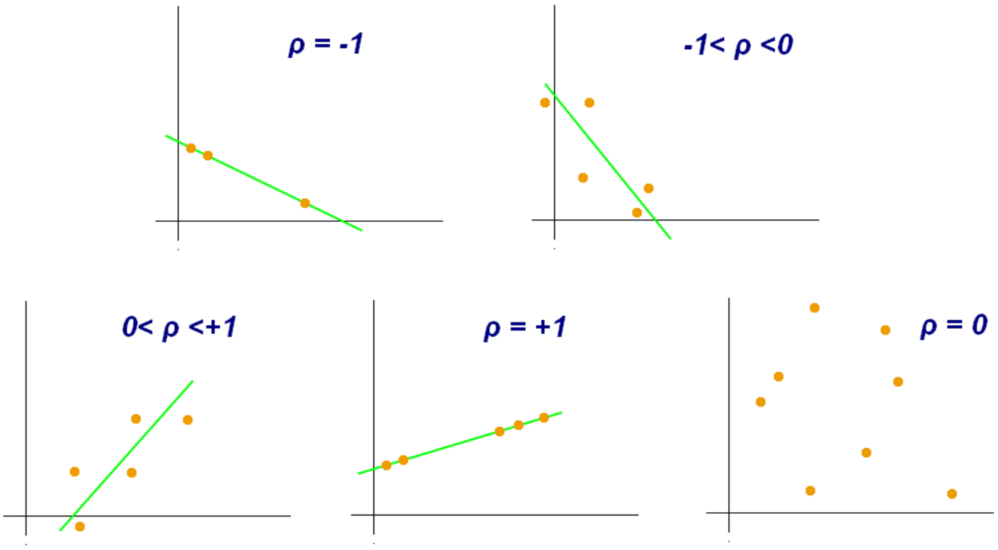
- Dataset I - If you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- Dataset II - If you look at this figure you can conclude that there is a non-linear relationship between x and y.
- Dataset III - You can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated to be far away from that line.
- Dataset IV - Shows an example of when one high-leverage point is enough to produce a high correlation coefficient.

Conclusion:

While the descriptive statistics of Anscombe’s Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

3. What is Pearson’s R?

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment of the origin) of the product of the mean-adjusted random variables; hence the modifier product moment in the name.
The Pearson correlation coefficient (PCC) is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviation; thus, it is essentially a normalized measurement of the covariance, such that that result always has a value between -1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationships or correlations.
As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

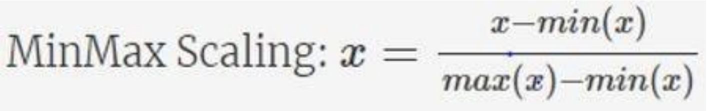


Examples of scatter diagrams with different values of correlation coefficient (ρ)

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
Scaling is a step of data pre-processing that is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Most of the time, the collected data set contains features highly varying in magnitudes, units, and ranges. If scaling is not done, then the algorithm only takes magnitude into account and not units hence incorrect modeling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.
It's important to note that scaling just affects the coefficients and none of the other parameters like **t-statistic**, **F-statistic, p-values**, **R-squared,** etc.

Normalization/Min-Max Scaling:

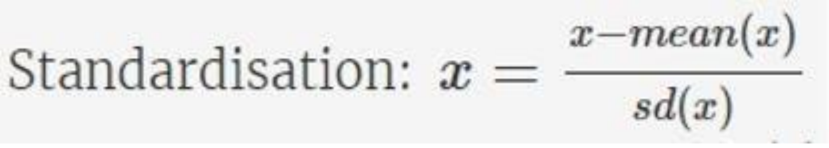
- It brings all the data in the range of 0 and 1.
- *sklearn.preprocessing.MinMAxScaler* helps to implement normalization in python



$$\text{MinMax Scaling: } x = \frac{x-min(x)}{max(x)-min(x)}$$

Standardization Scaling:

- Standardization replaces the value by their Z scores. It brings all the data into a standard normal distribution which had a mean (μ) zero and a standard deviation (σ) of one.



$$\text{Standardisation: } x = \frac{x-mean(x)}{sd(x)}$$

- *sklearn.preprocessing.scale* helps to implement standardization in Python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

   This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   In our given bike data set, there was almost ‘temp’ and ‘atemp’ were closely correlated.

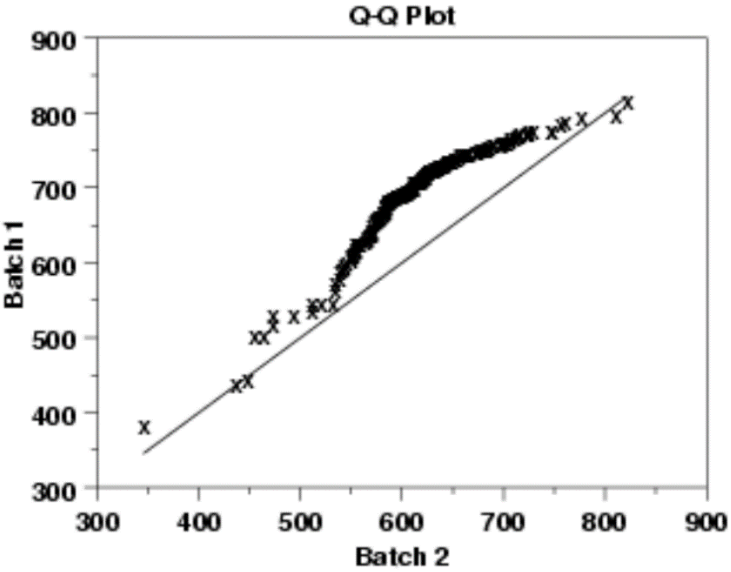
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

   The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% falls above that value.

   The advantages of the q-q plot are:

   - The sample sizes do not need to be equal.
   - Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.



Sample Q-Q plot.