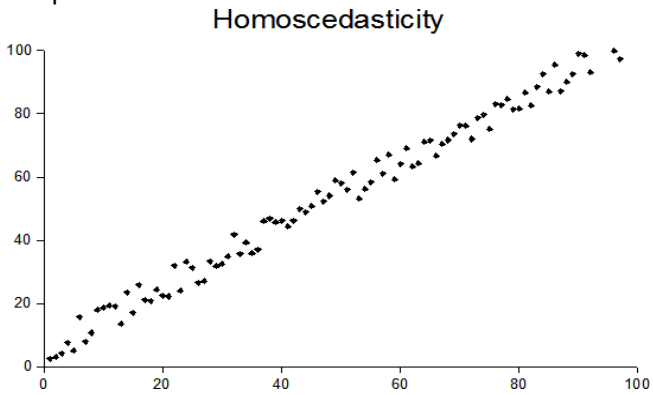
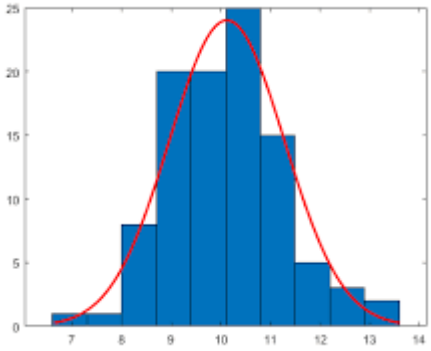
1. What are the assumptions of linear regression regarding residuals?
    - linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other. For example stock market data is Autocorrelated.
    - residuals are equal across the regression line. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables

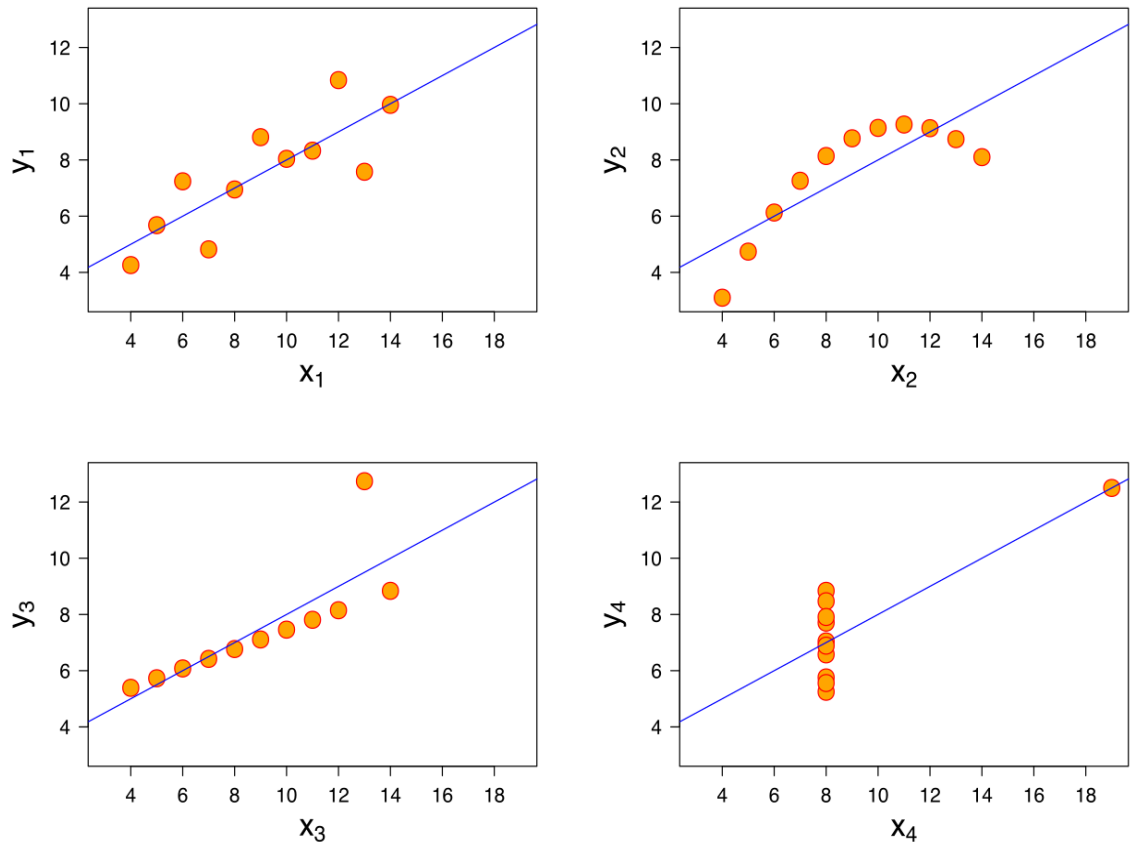


    - The residuals are normally distributed.



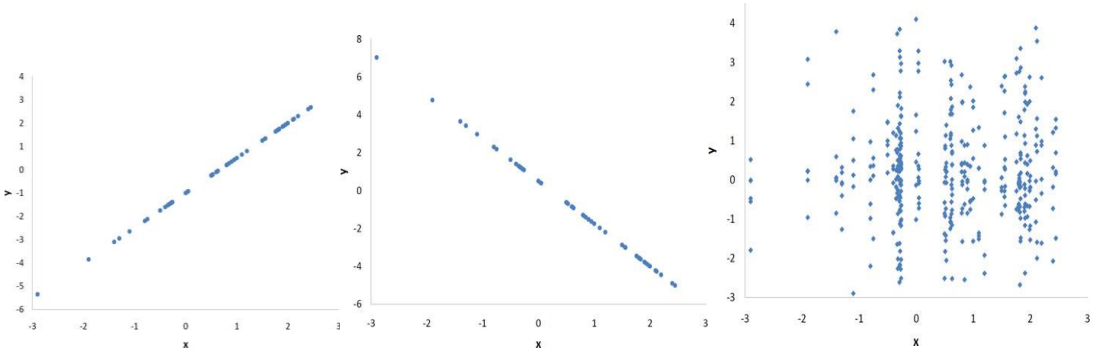
2. What is the coefficient of correlation and the coefficient of determination?
   Coefficient of correlation is “R” value which is given in the summary table in the Regression output. R square is also called coefficient of determination. Multiply R times R to get the R square value. In other words Coefficient of Determination is the square of Coefficeint of Correlation.
   R square or coeff. of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.
   Coefficient of Correlation: is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison.

3. Explain the Anscombe's quartet in detail.
   Anscombe's quartet comprises four data sets that have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally
- the third graph (bottom left), the distribution is linear, but should have a different regression line
- the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

4. What is Pearson's R?
    - The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.
    - Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables



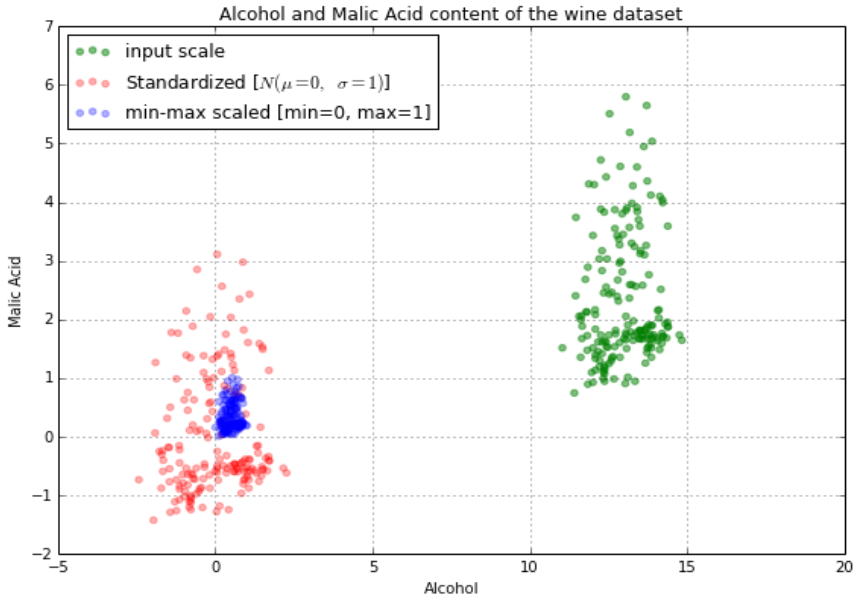
    - First one Shows a perfect positive relationship. R=1
    - Second one shows perfect negative relationship. R=-1
    - Third one shows no relationship. R=0

5. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Scaling is a method used to normalize the range of independent variables or features of data. the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.
   Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.
   Normalization : Scaling some data to a confined range. It changes the shape of Data.
   For example {1,2,3} → {0, 0.5, 1} →by normalization in the reference of [0.0, 1.0] range.
   For example Min-Max scaling

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

   Standardized Scaling : Moving the data to a centered location. It represent the number of deviation s each value from the mean value. It rescales a feature to have mean of zero and a unit variance.

$$z = \frac{x_i - \mu}{\sigma}$$



Green color datapoints represents input data.
Red color represents standardized scaled data.
Blue color represents Normalized Data.

6. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. VIF is represented by

$$VIF_i = \frac{1}{1 - R_i^2}$$

If two X's are perfectly correlated. VIF = 1/(1-1)= 1/0 = infinity that is the estimate it as imprecise as it can be.