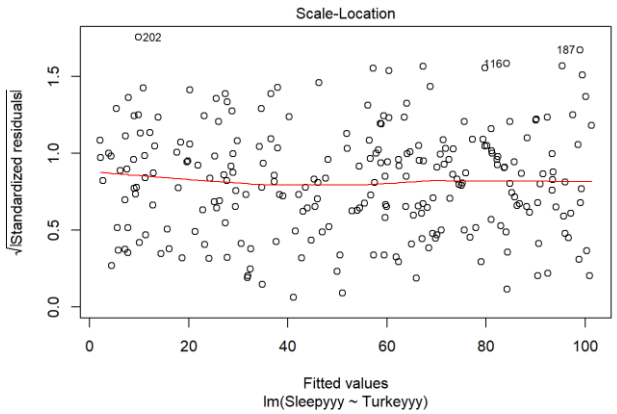
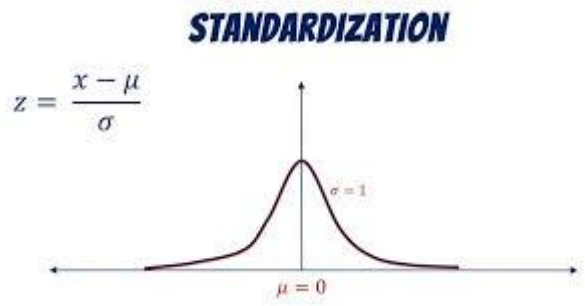
1. Explain the linear regression algorithm in detail.
   - Linear regression is a statistical relationship to build a model using target variable (dependent variable) and the variables affecting the target variables (independent variables)
2. What are the assumptions of linear regression regarding residuals?
   - No correlation between the residuals
   - It doesn't follow any pattern at it, most likely it is randomly distributed

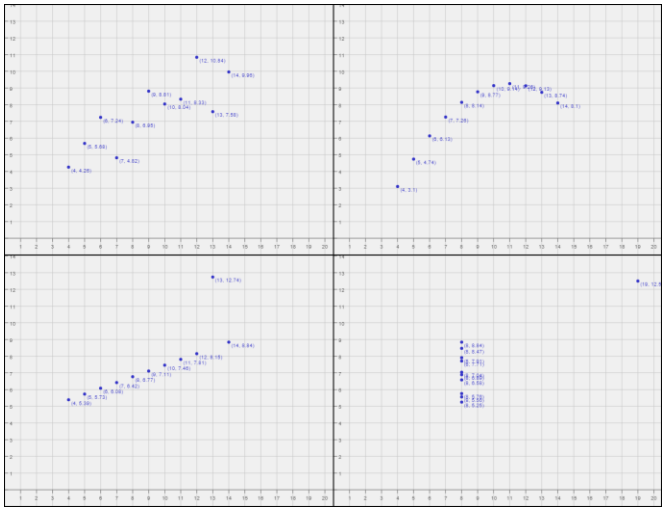


   along the best fit line, showing equal variance for the complete data
   - It deals with the prediction which is within the range (interpolation) but cannot guarantee extrapolation which is out of its reach
   - The model is given by a line which says the model is linear (y= mx+c)
   - The independent variables and the variables effecting the target variables are not correlated
   - The number of observations made are greater than the no.of variables which are present in the data
   - The regression model is well specified by showing the relation between target and independent variables effecting
   - There is no perfect linear relationship between any two independent variables



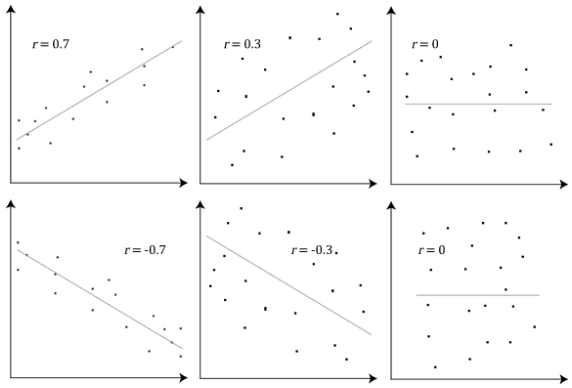
3. What is the coefficient of correlation and the coefficient of determination?
   - The coefficient of correlation is the correlation between the target variable and independent variables which is represented by R
   - The coefficient of determination is percentage of variation in the target variable explained by the independent variables and represented as R-Squared
   - R- Squared value may decrease some time when the independent variables have a correlation

4. Explain the Anscombe's quartet in detail.
   - The Anscombe's quartet is of a pictorial representation how the data shows the information in different aspects



- One will show the variance of the data and the other shows some other kind of pattern which says a different story
- The other showing a plot where the data points lie on the residual line which is having outliers
- The last one shows than a single outlier can make a best fit line

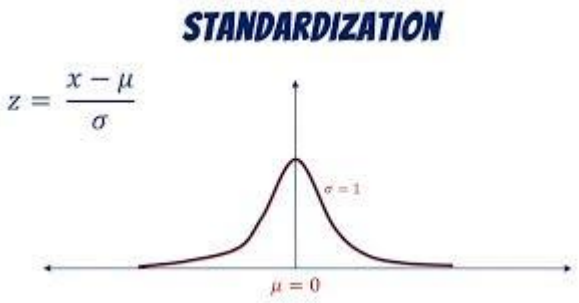
5. What is Pearson's R?
   - Pearson's correlation coefficient is the correlation between the two continuous variables. The correlation lies from – 1 to +1
   - When the value is close to 0 says the correlation is less



6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   - Bringing all values to common unit can be stated as scaling.
   - Normalization rescales every value within the range of 0 to 1

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

   - standardization rescales the value to have a mean of 0 and standard deviation of 1



7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   - When the R-Squared value reaches close to 1 then the VIF becomes

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

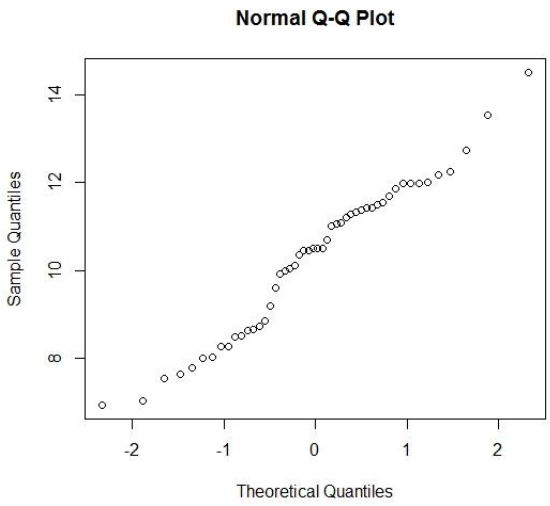
     infinite
   - It also says that the variables are highly correlated

8. What is the Gauss-Markov theorem?
   - Gauss Markov theorem is of linear regression model which states the errors in the model are uncorrelated in the method of OLS (ordinary least squares)
   - The variables are not perfectly correlated

- The sample must be taken randomly from while splitting the test and train data sets
- The variance is constant no matter what variables are taken

9. Explain the gradient descent algorithm in detail.
   - Gradient descent is the another method to optimize in iterative method by minimizing the value and bringing towards 0
   - Firstly we will take an assumed value and do second degree differentiation for an equation and substitute the value again and again until there is no difference in the value of Y

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    - Q-Q plot is a scatter plot to know how two variables are related to each



other if those two variables are very well positively correlated the plot will be as below