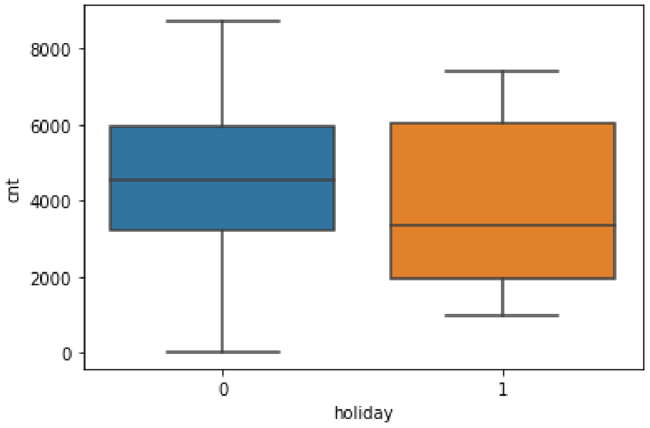
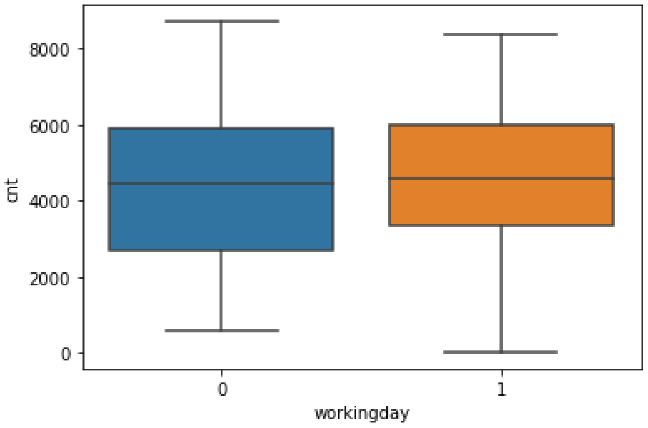
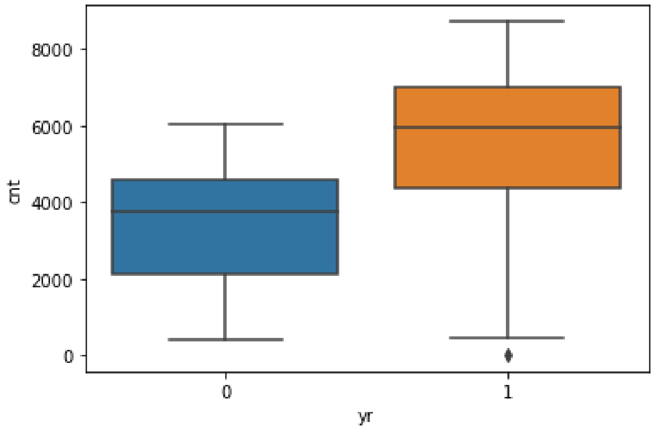
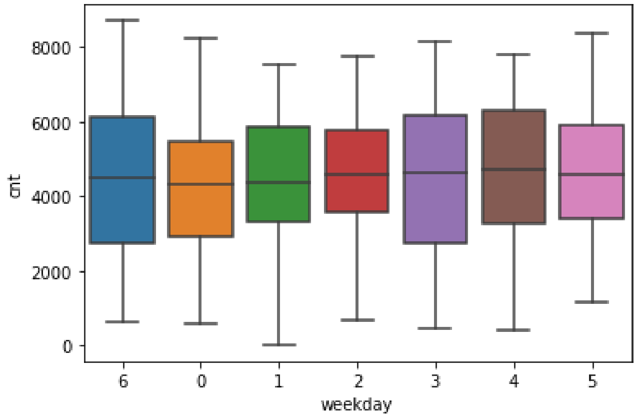
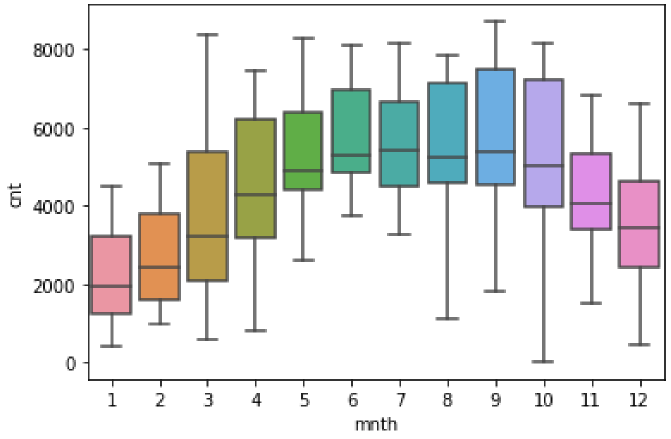
ASSIGNMENT BASED QUESTIONS

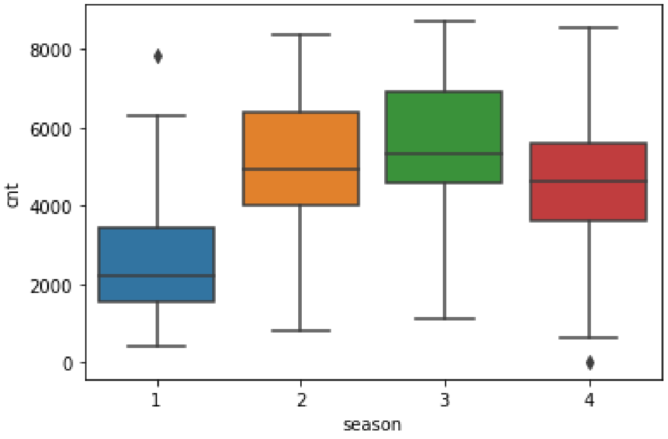


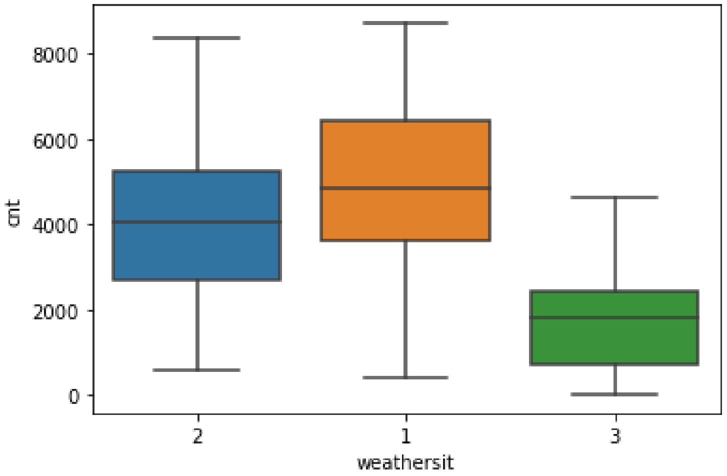










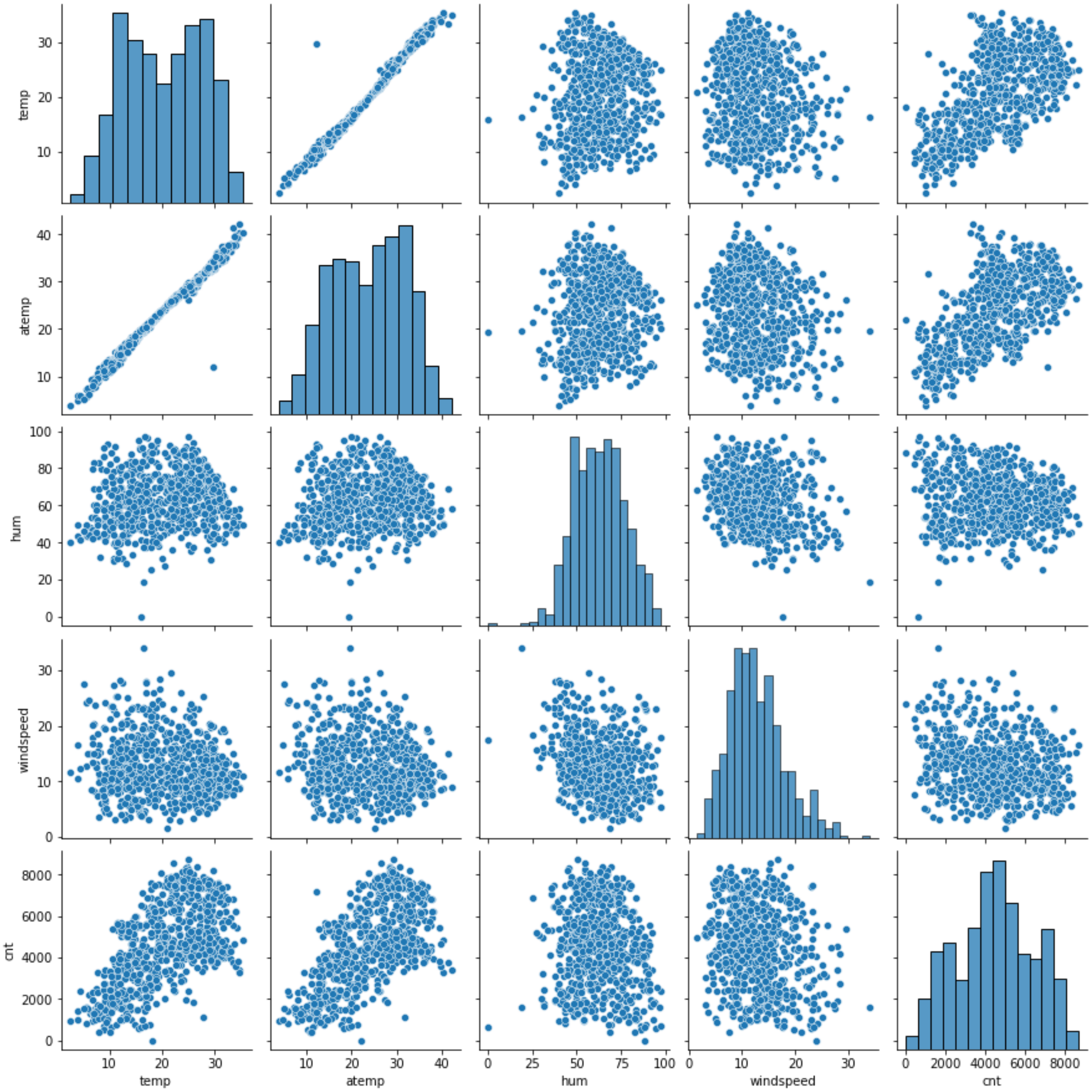


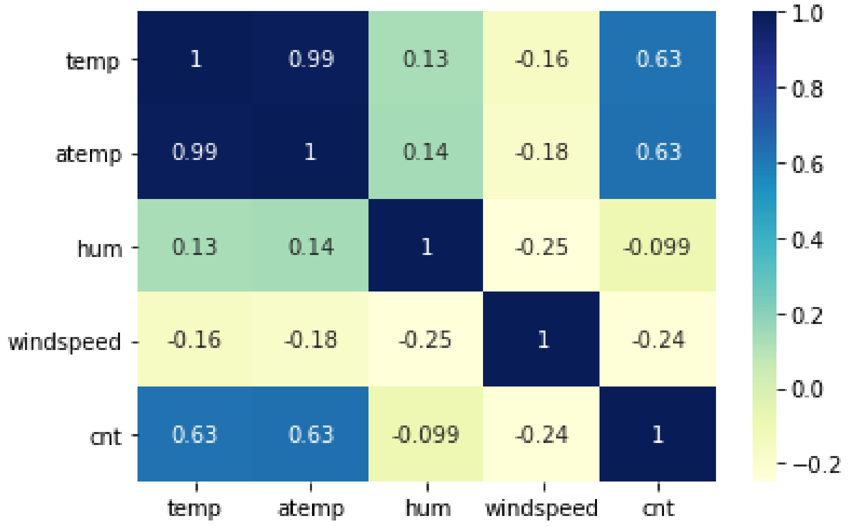
1. The following inferences can be drawn from analysis of influence of categorical variable on dependent variable.

- Season has very solid impact on count as the season changes count of bikes significantly changes for spring(1) the count is least and most for fall(3)
- Year has a significant impact i.e in 2019(1) more count as compared to 2018(0)
- Month has a similar picture as season, which is obvious as seasons are nothing but buckets of various months.
- There is a higher count for non-holiday on average, but the spread is more in holidays.
- Weekday does not print a clear cut picture as almost all have the same median values.
- Workingday also does not print a pretty picture as we get very little inference of count from this category.
- Weathersit impacts the count as it's obvious that when it rains/snows (3) less people are likely to use bikes compared to higher numbers on clearer days.

2. Dummy variables are basically arranging the long format categorical variable to wide format binary encoding. Since every n numbers of data in categorical variable can be explained by n-1 other variable. It is eminent to use drop_first = True to remove the extra column. This helps in faster converging of same data information in fewer columns.

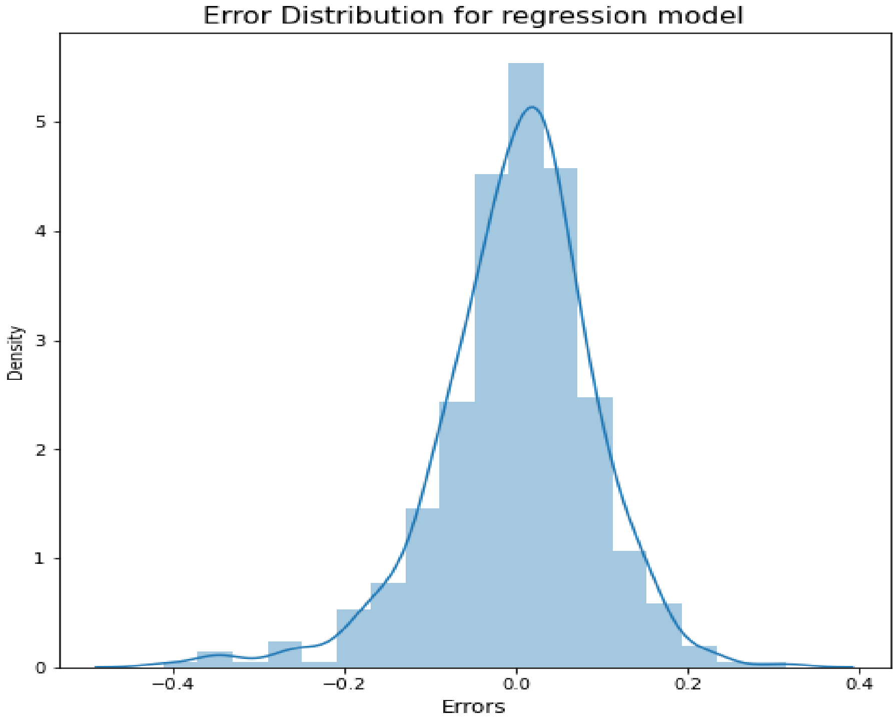
3. From the pair plots of numerical variables it is clear that temp and atemp are the most correlated with target variable as they show a linear positive relationship. Also the correlation coefficient is higher (0.63)

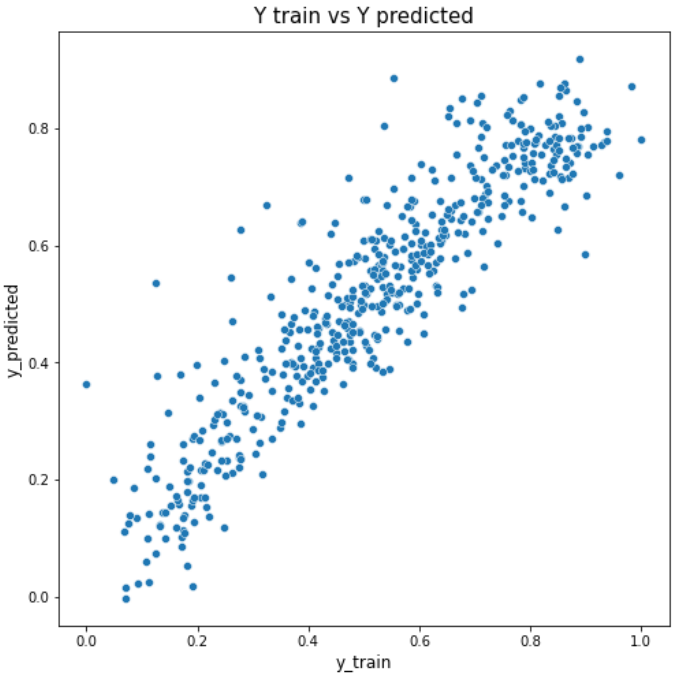


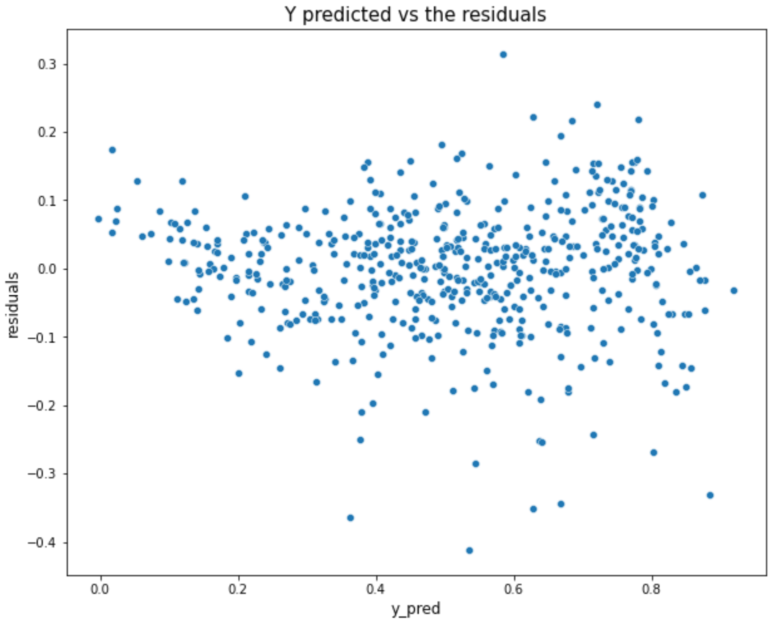


	temp	atemp	hum	windspeed	cnt
temp	1	0.99	0.13	-0.16	0.63
atemp	0.99	1	0.14	-0.18	0.63
hum	0.13	0.14	1	-0.25	-0.099
windspeed	-0.16	-0.18	-0.25	1	-0.24
cnt	0.63	0.63	-0.099	-0.24	1

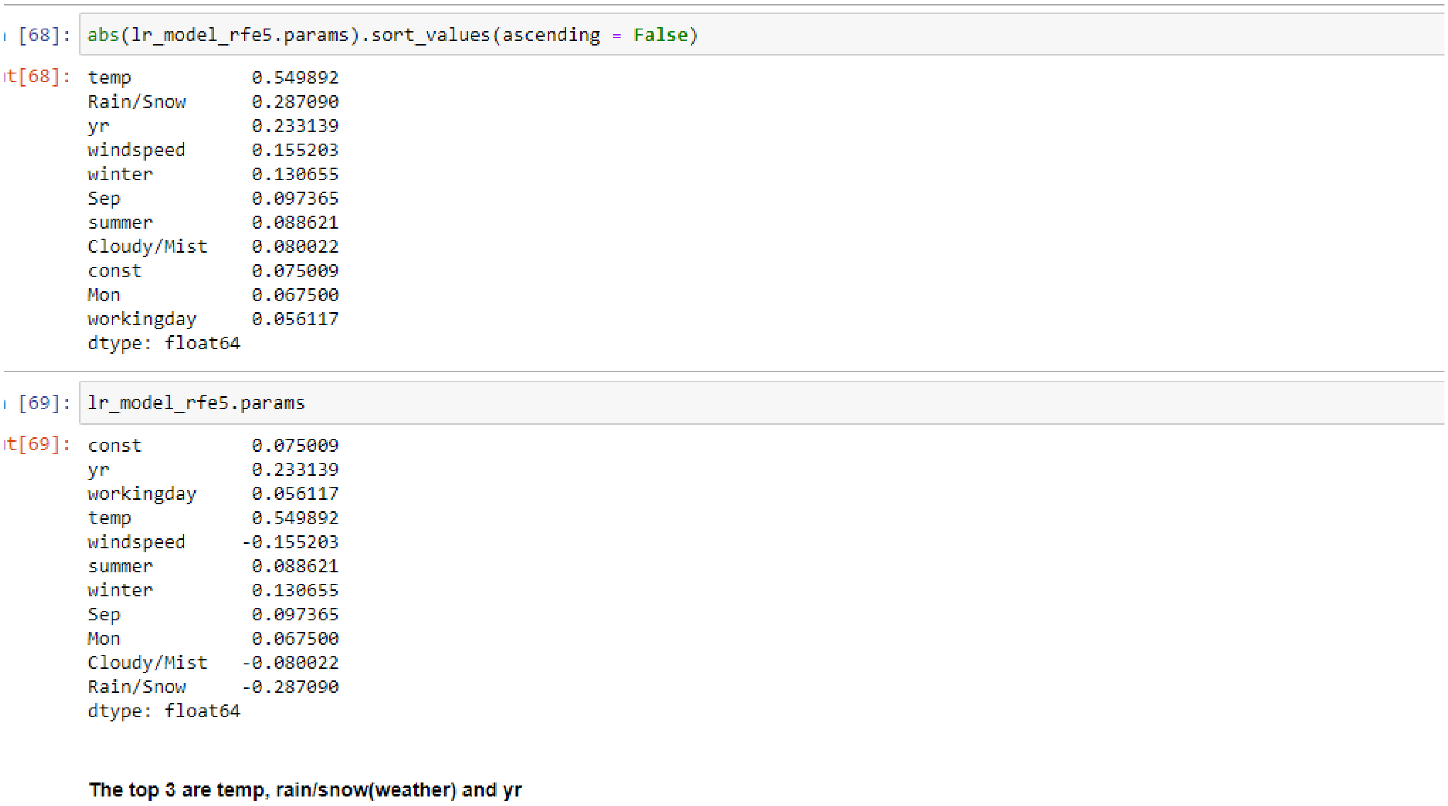
4. Assumptions of linear Regression being validated
   - Errors are normally distributed across a mean 0
   - Predicted model is linear as the y predicted and y train is spread across the diagonal.
   - There is no such clearly recognized pattern to errors. So it can be concluded that the assumptions are valid. This establishes homoscedasticity.







5. The top 3 features contributing significantly to explaining the demand of shared bikes are temp, rain/snow (weather) and yr
Where temp and year are positively related where as bad weather (rain/snow) is negatively related.



```
[68]: abs(lr_model_rfe5.params).sort_values(ascending = False)
```

```
t[68]: temp           0.549892
       Rain/Snow      0.287090
       yr             0.233139
       windspeed      0.155203
       winter         0.130655
       Sep            0.097365
       summer         0.088621
       Cloudy/Mist    0.080022
       const          0.075009
       Mon            0.067500
       workingday     0.056117
       dtype: float64
```

```
[69]: lr_model_rfe5.params
```

```
t[69]: const          0.075009
       yr             0.233139
       workingday     0.056117
       temp           0.549892
       windspeed     -0.155203
       summer         0.088621
       winter         0.130655
       Sep            0.097365
       Mon            0.067500
       Cloudy/Mist   -0.080022
       Rain/Snow     -0.287090
       dtype: float64
```

The top 3 are temp, rain/snow(weather) and yr

GENERAL SUBJECTIVE QUESTIONS

1.

Linear Regression Algorithm – Fits a system of variables to a linear equation in the form:

y = c + m1*x1 + m2*x2 +..., where y is the dependent variable and x1, x2 , etc are independent variables(features) that predict y(target variable).

Unknowns to solve: intercept = c, beta coefficients = m1, m2 ...

The method commonly used-

- **Minimize the cost function(J):**

  Ordinary Least Squares- This is one of the methods used to get the solution of the above equation. It finds the average sum of difference of squares of predicted y from given y. This is the cost function that we have to minimize to get our beta-coefficients of independent variables and the intercept.

- **Methods used for reducing cost function :**
  Differentiation- This differentiates the cost function and equates it to zero. Thus solving the resultant equation results in finding the beta-coefficients. But this method is not feasible for equations of higher degree.

  Gradient Descent- This uses an iterative method to find the minima.
  Step 1. Assumes initial values for the beta-coefficients.
  Step 2. Uses a learning rate alpha that gives the speed of the descent.
  Step 3. Find the difference of alpha*partial differential (J, w.r.t to beta) from beta initial.
  Step 4. The value after finding the difference is the new beta initial.
  Step 5. Step 2, 3 and 4 are repeated until there is no further change in beta value, or the gradient term (differentiation) becomes zero. This is the minima and the value of beta obtained is the required beta coefficient.

This is repeated for every beta corresponding to x and also the constant. Finally the value obtained is called the fit to the data or the result of the logistic regression.

2.

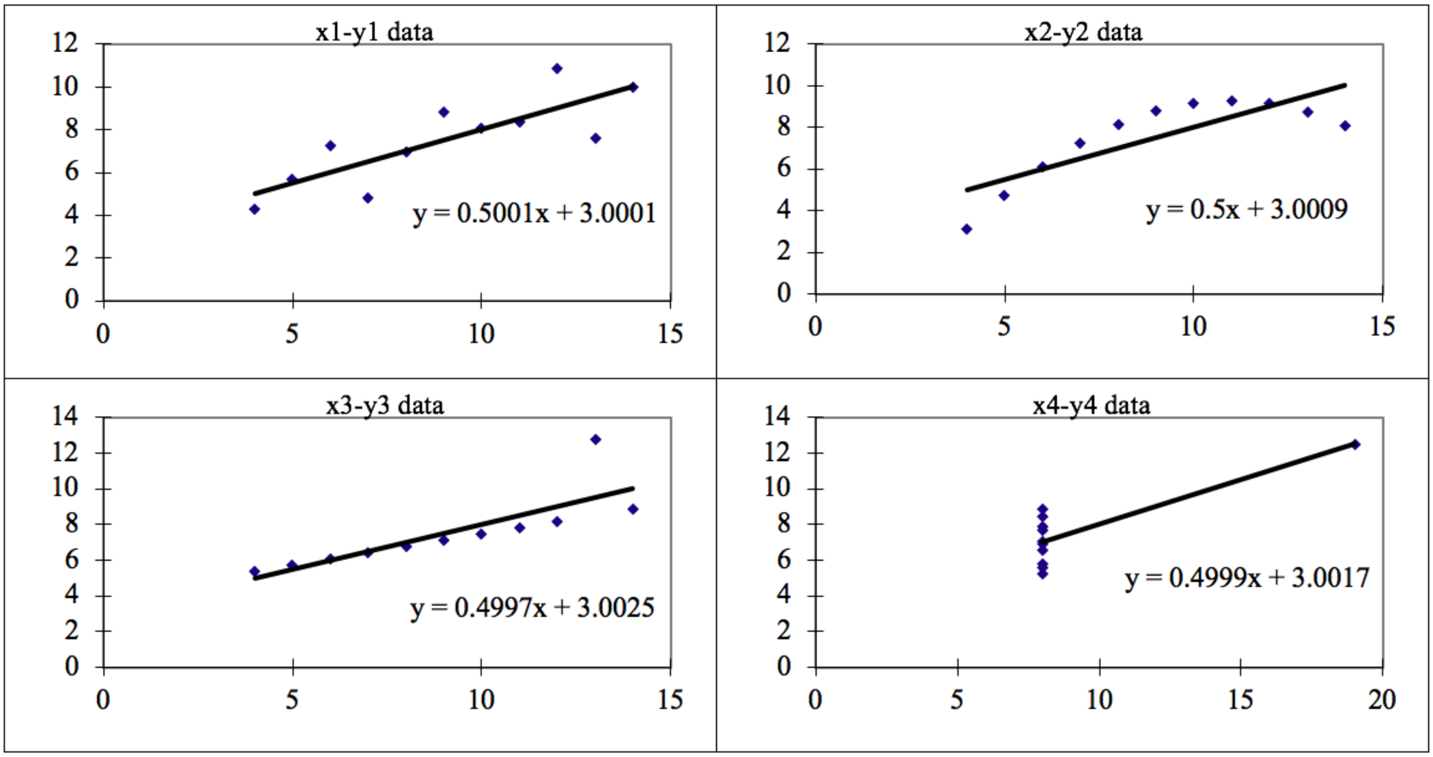
Anscombe’s Quartet is a data representation of four plots which show almost similar regression equation but representation of the data in plots and fitting the line shows very different results.

Dataset 1. Shows the regression model perfectly fits the data

Dataset 2. Shows the data does not follow linear relationship.

Dataset 3. Shows data has outliers.

Dataset 4. The outliers cannot be explained by linear regression.



This is a visual representation of Anscombe’s Quartet. The regression model is almost same for all the dataset but visually the spread shows a different picture.

So it can be concluded that visual representation of the data should be done prior to regression analysis, to deal with outliers or to check whether the data can be fitted to linear regression.

3.

Pearson’s r is a numerical representation of strengths of linear relation of two variables. It lies between -1 and 1. If 1 or -1 is the correlation it means the variables are perfectly positively and negatively correlated respectively. A correlation of 0 states there is no linear association of the variables.

Mathematically:

Corr = covariance(x,y)/product of std. deviations

4.

Scaling is an operation to squeeze the data points or variables to certain boundary limits.

Scaling squeezes all the data of different independent variables to certain boundaries, this closes in the size of coefficients. Suppose in our dataset we have variables which are in range -0.05 to 0.05 and another variable in the range of millions, this significantly impacts the values of coefficients as we get a large coefficient for the former variable and very small for later. This does not necessarily mean former has more significance than other. Thus feature scaling solves this issue. Also it converges the algorithm faster as all the variables are pinned within the boundaries.

- Types:
  1. Standardized scaling = (x-mu)/ (sigma), where mu = mean, sigma = standard deviation. This scales the data centered on the mean with unit standard deviation.
  2. Normalized scaling = (x – min)/ (max-min), this scales the data between zero and 1.

5.

VIF = 1/ (1-R2)

VIF- Variable inflation factor, as the name suggests is the inflation of one variable as a result of presence of other variables. The cause for increase in VIF is the effect of multcollinearity. An infinite VIF means R2 =1.The said variable is in perfect correlation with other variable.

Thus if a particular variable has high correlation with other variables in the data, the presence of the former variable becomes insignificant as the variance of the variable inflates by a factor of VIF and the error by sq.rt(VIF), This increases the confidence interval and reduces the beta coefficient.

6.

Quantile- quantile plots are the plots for various quantile data of two distributions.

A diagonal is plotted and the points of the various quantiles are plotted along with the line. If it falls on the diagonal line, we can say the distributions of two data are pretty close.

In linear regression we use q-q plots to find the y-pred vs y-test data to check whether the data have similar distribution or not, in other words if our model can predict the test data or not. If the plot hovers near the diagonal line the two data have common distribution.