Assignment - Linear Regression

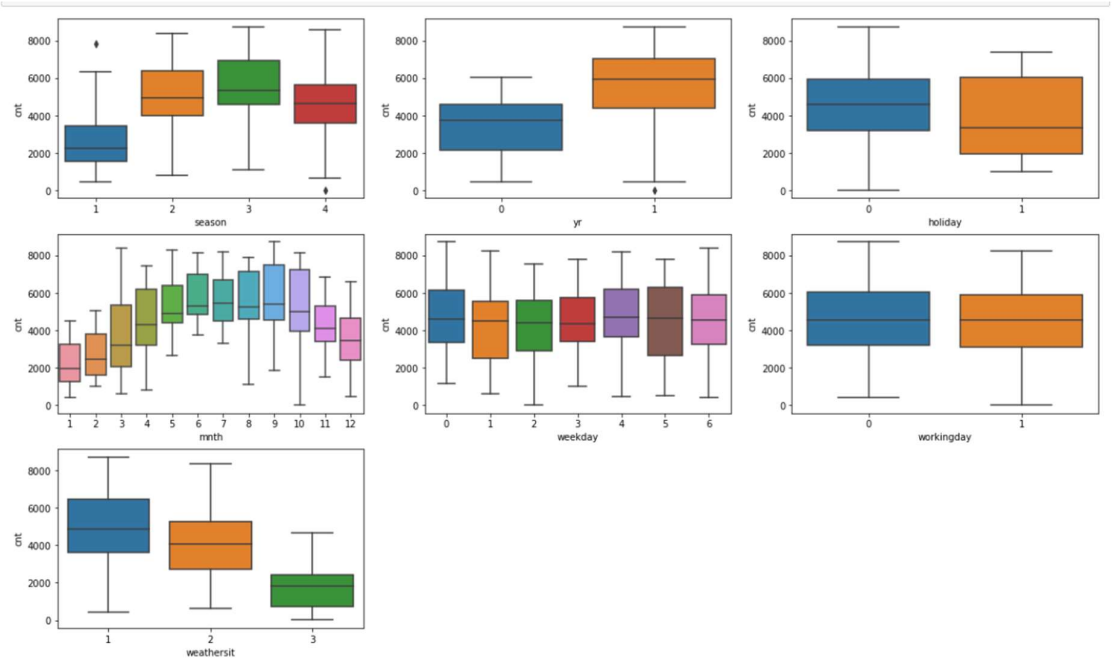
Assignment-based Subjective Questions

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

   Categorical variables do have an effect on dependent variables.
   Inference -

   - Spring season has least bike demands while fall has the most.
   - Yr 2019 has huge bike demands as compared to 2018.
   - During holidays, the bike demand is less which is reasonable as more people tend to stay home.
   - Mnth Jan has least bike demands in the US while July & September seems to have highest demands.
   - Bike demand seems to be similar for the weekdays.
   - Bike demands are the same for working and non-working days.
   - When the weather is clear, the demand for bikes is high while in case of Light Snow & Rain, it is least.

Reference screenshot -



Some of the variables which were created as part of Dummy Variables creation step do have an effect on the dependent variable (cnt). Below are the dummy variables -

1. mnth - Sep, Jul
2. weathersit - Light Snow & Rain, Mist & Cloudy
3. season - spring

These feature variables are having very less p-value and also acceptable VIF (<5) due to which our final model is a good model with R-square value as 83% approximately.

Attaching the reference screenshot of the final model which shows the variables along with their coefficients.

Dep. Variable:	cnt	**R-squared:**	0.828
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	267.5
Date:	Sat, 08 May 2021	**Prob (F-statistic):**	8.40e-185
Time:	23:13:00	**Log-Likelihood:**	487.41
No. Observations:	510	**AIC:**	-954.8
Df Residuals:	500	**BIC:**	-912.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3086	0.019	16.274	0.000	0.271	0.346
yr	0.2353	0.008	28.011	0.000	0.219	0.252
holiday	-0.0924	0.027	-3.477	0.001	-0.145	-0.040
temp	0.3896	0.026	14.839	0.000	0.338	0.441
windspeed	-0.1516	0.025	-5.993	0.000	-0.201	-0.102
spring	-0.1454	0.012	-11.712	0.000	-0.170	-0.121
Jul	-0.0736	0.018	-4.154	0.000	-0.108	-0.039
Sep	0.0565	0.016	3.511	0.000	0.025	0.088
Light Snow & Rain	-0.2783	0.025	-11.067	0.000	-0.328	-0.229
Mist & Cloudy	-0.0820	0.009	-9.166	0.000	-0.100	-0.064

Omnibus:	46.958	**Durbin-Watson:**	2.018
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	103.026
Skew:	-0.514	**Prob(JB):**	4.25e-23
Kurtosis:	4.947	**Cond. No.**	11.2

2. Why is it important to use drop_first=True during dummy variable creation?

For any categorical variable which gets converted to a dummy variable of k distinct values (in the form of 0 & 1), k-1 levels are sufficient to identify the columns with one value being considered as the base value hence we drop this base value column from the dataset using **drop_first = True** command.

Reference screenshot as follows -

```
In [31]: # Get the dummy variables for the feature 'season' and store it in a new variable - 'season_type'
         season_type = pd.get_dummies(df['season'])

In [32]: # Check the new dataset 'season_type'
         season_type.head()
```

Out[32]:

	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

Now, we drop one of the column lets say `fall` column, as the type of fall can be identified with the last there columns where —

- `000` will correspond to `fall`
- `001` will correspond to `winter`
- `010` will correspond to `summer`
- `100` will correspond to `spring`

Here, for the season variable, we have 4 distinct values - fall, winter, summer & spring. Now we can drop the ‘fall’ column as it can be identified using the ‘000’ value of winter, summer & spring.

Reference screenshot is as follows -

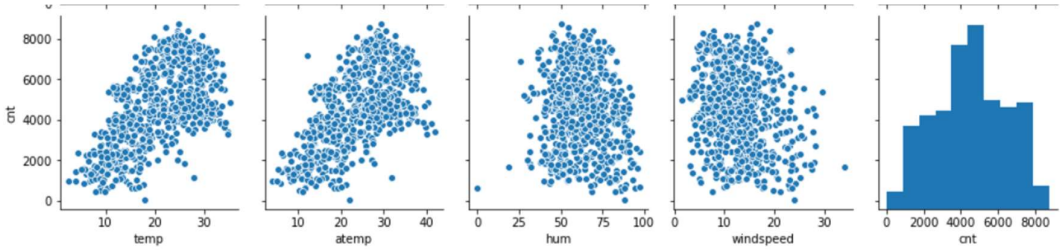
```
In [33]: # Let's drop the first column from season_type df using 'drop_first = True'
         season_type = pd.get_dummies(df['season'], drop_first=True)

In [34]: # Add the results to the original dataframe
         df = pd.concat([df, season_type], axis=1)

In [35]: # Now let's see the head of our dataframe.
         df.head()
```

Out[35]:

	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	cnt	spring	summer	winter
0	spring	0	Jan	0	Mon	1	Mist & Cloudy	14.110847	18.18125	80.5833	10.749882	985	1	0	0
1	spring	0	Jan	0	Tue	1	Mist & Cloudy	14.902598	17.68695	69.6087	16.652113	801	1	0	0
2	spring	0	Jan	0	Wed	1	Clear	8.050924	9.47025	43.7273	16.636703	1349	1	0	0
3	spring	0	Jan	0	Thu	1	Clear	8.200000	10.60610	59.0435	10.739832	1562	1	0	0
4	spring	0	Jan	0	Fri	1	Clear	9.305237	11.46350	43.6957	12.522300	1600	1	0	0

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

temp/atemp variable has the highest correlation among the numerical variables with 'cnt' target variable. Reference screenshot of cnt variable with highlighted plots-



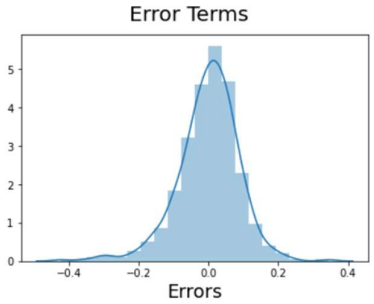
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Using Residual Analysis on the training set - If error terms are normally distributed with mean value = 0 and standard deviation =1 then we can say that our model is a good fit for linear regression problem.
   It is one of the major assumptions of linear regression which is validated by plotting a histogram of the error terms.

Residual Analysis of the train data

So, now to check if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression), let us plot the histogram of the error terms and see what it looks like.

```
In [132]: y_train_pred = lm2.predict(X_train_lm2)

In [133]: # Plot the histogram of the error terms
          fig = plt.figure()
          sns.distplot((y_train - y_train_pred), bins = 20)
          fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
          plt.xlabel('Errors', fontsize = 18)

Out[133]: Text(0.5, 0, 'Errors')
```



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

The top 3 features are -

a. temp (0.3896)
b. weathersit - Light Snow & Rain (-0.2783)
c. yr (0.2353)

Reference screenshot of the final model -

In [130]:
```
lm7.summary()
```

Out[130]:

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.828
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	267.5
Date:	Sat, 08 May 2021	**Prob (F-statistic):**	8.40e-185

Dep. Variable:	cnt	**R-squared:**	0.828
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	267.5
Date:	Sat, 08 May 2021	**Prob (F-statistic):**	8.40e-185
Time:	23:13:00	**Log-Likelihood:**	487.41
No. Observations:	510	**AIC:**	-954.8
Df Residuals:	500	**BIC:**	-912.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3086	0.019	16.274	0.000	0.271	0.346
yr	0.2353	0.008	28.011	0.000	0.219	0.252
holiday	-0.0924	0.027	-3.477	0.001	-0.145	-0.040
temp	0.3896	0.026	14.839	0.000	0.338	0.441
windspeed	-0.1516	0.025	-5.993	0.000	-0.201	-0.102
spring	-0.1454	0.012	-11.712	0.000	-0.170	-0.121
Jul	-0.0736	0.018	-4.154	0.000	-0.108	-0.039
Sep	0.0565	0.016	3.511	0.000	0.025	0.088
Light Snow & Rain	-0.2783	0.025	-11.067	0.000	-0.328	-0.229
Mist & Cloudy	-0.0820	0.009	-9.166	0.000	-0.100	-0.064

Omnibus:	46.958	**Durbin-Watson:**	2.018
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	103.026
Skew:	-0.514	**Prob(JB):**	4.25e-23
Kurtosis:	4.947	**Cond. No.**	11.2

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   - Regression is one of the machine learning algorithm/method which is used for predictive analysis.
   - The output variable which is to be predicted is a continuous variable like price of a stock, student marks etc.
   - It falls under the category of Supervised learning methods which has the defined labels from the previous dataset.
   - The industry applicable to common use cases of linear regression are health, education, real-estate, sports, business, entertainment.
   - Linear Regression explains the relationship between the variables using a straight line.

- It has one/multiple independent variables called feature variables and 1 output/dependent variable called as target variable which is to be predicted.
- There are two type of linear regression:-
    a. Simple Linear Regression
    b. Multiple Linear Regression
- Simple Linear Regression - An algorithm when the target variable is dependent on a single independent/feature variable.
- Multiple Linear Regression - An algorithm when the target variable is dependent on more than one independent/feature variable.
- Standard equation of simple linear regression is -

    $y = mx + c$

    OR

    $y = \beta_1 X + \beta_0$

    m/β_1 - slope of the line
    c/β_0 - Intercept/constant
    y - target variable
    X - feature variables(one/many)

- Our aim is to predict/find a linear equation/line which has optimum coefficients (m,c) so that our target variable(y) is close to their actual values.
- The line which we obtain/plot for a dataset has some errors i.e. all the data points do not coincide with the predicted line in a real world scenario.
- So, the best fit line (of the dataset) is obtained by minimising the error/residual using different techniques one of which is RSS (Residual Sum of Squares) whose formula is -

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_i)^2$$

- The strength of the linear regression model can be assessed using 2 metrics:
    a. R-Square or Coefficient of Determination
    b. Residual Standard Error (RSE)

- R-square is calculated/represented as

    $R2 = 1 - (RSS / TSS)$

    Where RSS - Residual Sum of Squares
    TSS - Total Sum of Squares

    R2 = 1 means perfect correlation
    R2 = 0 means no correlation
    R2 = +ve means positively correlated
    R2 = -ve means negatively correlated
- In multiple linear regression model, target variable is dependent on more than one feature variable.

- Standard Equation of multiple linear regression is -

  $y = c + m1x1 + m2x2 + \ldots + m(n)x(n)$

  OR

  $y = \beta_0 + \beta1X1 + \beta2X2 + \ldots + \beta(n)X(n)$

- Significance of the variable is validated using p-value (p-value of the coefficients should be less than 0.005)
- One of the major assumptions of linear regression model is residual analysis wherein the error terms should be normally distributed with mean value = 0.
- In the case of multiple regression model, the independent variables could be correlated with each other which is known as multicollinearity. To validate it, we compute VIF (Variance Inflation Factor) represented as -

  VIF = 1/(1-R2)
  Where R2 is R-Square
- For model building, we need to check/validate below parameters to get the desired model -

  a. p-value of the coefficients of feature variables < 0.005
  b. VIF < 5 is acceptable (may vary depending on use case)
  c. R-square and adjusted R-square should be close (+5%, -5%).

- RFE (Recursive Feature Elimination) - is an automated technique to select the no of desired features variables (eg 10, 15 etc) and build a model based on that which can further be reduced/increased using manual selection of the variables to build the desired model.

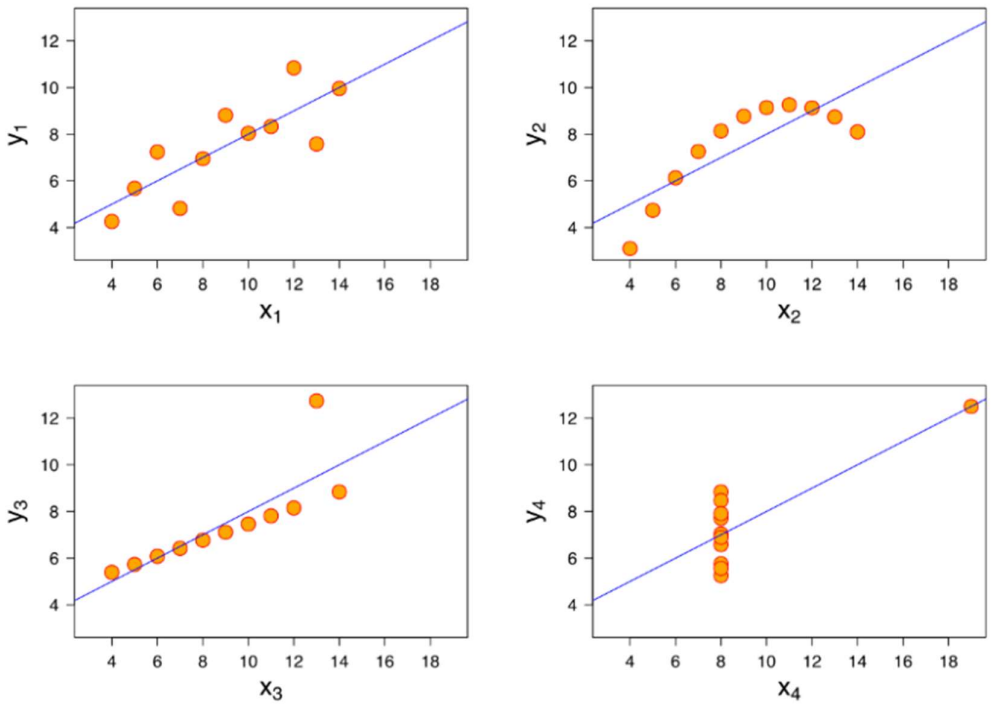
2. Explain the Anscombe's quartet in detail

- Anscombe's quartet can be defined as a group of four datasets which are nearly identical in simple descriptive statistics (Sum/Mean/Std Dev etc) but appear differently when plotted/graphed.
- It was developed by Francis Anscombe in 1973 hence the name Anscombe's quartet.
- It illustrates the importance of Data Visualisation before analysing and model building.
- It comprises four datasets each containing (x,y) pairs which share nearly the same descriptive statistics but things change when they are plotted.
- Below is the sample dataset and their corresponding graphs -

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

a. Sum of the datasets are same
b. Mean and Standard Deviation are also same

Plotting the above datasets -



After plotting these - we can clearly see the difference in the graphs which could have never been identified while only looking at the data and their statistical summary. And Each dataset is communicating a different story -

a. Plots 1 and 3 are linear with difference being an outlier in Plot 3.
b. Plot 2 is Non-linear
c. Plot 4 had no correlation but only 1 outlier has created a high correlation.

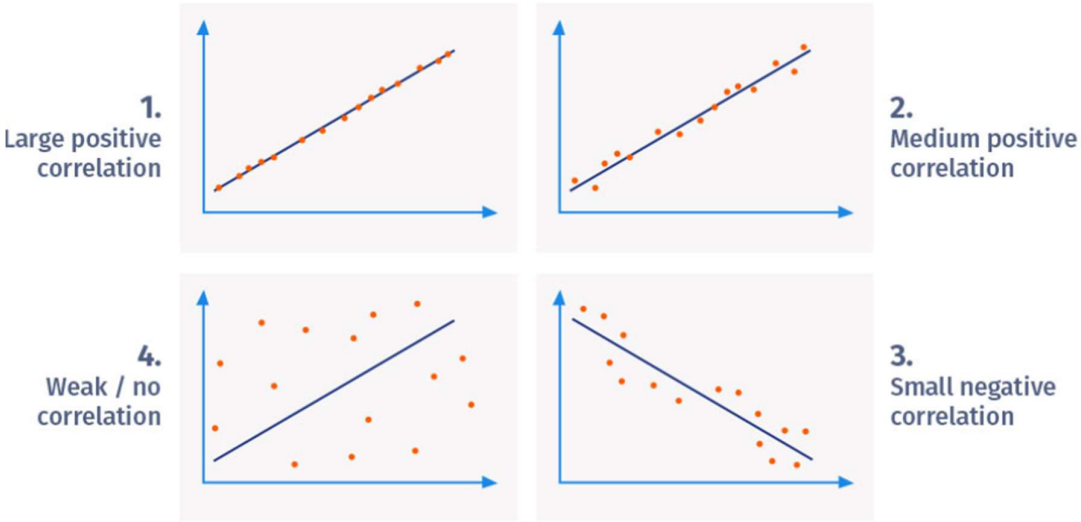
- Hence Anscombe's quartet states the importance/significance of Data Visualisation in Data Science/Analysis.

3. What is Pearson's R?

- Correlation coefficients are used to measure the relationship between two variables.
- The method most popularly used to find the correlation coefficient is Pearson's correlation, also known as Pearson's R method.
- It was founded by Karl Pearson hence referred to as Pearson's R.
- It is commonly used in Linear Regression algorithm/model.
- It is the covariance of the two variables divided by the product of their standard deviations.
- The value of correlation coefficient varies between +1 and -1.

  value = +1,-1 means perfect correlation
  value = 0 means no correlation
  value = +ve means positively correlated
  value = -ve means negatively correlated

  Reference screenshot as follows -



- Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

where -

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

- Scaling is another important factor to be considered in the case of multiple linear regression model.
- When there are more feature/independent variables, it might be the case when many of them might be on different scales due to which our coefficients will not be appropriate (too high/too low etc) and hence it causes difficulty in interpretation.
- So, in order to avoid such scenarios, we scale these feature variables also known as Feature Scaling.
- There are two main reasons to perform scaling -
    a. Ease of interpretation
    b. Speeding Up/Faster Convergence of Gradient Descent method
- There are two type of scaling: -
    a. Standardized Scaling
    b. Normalized Scaling/Min-Max Scaling

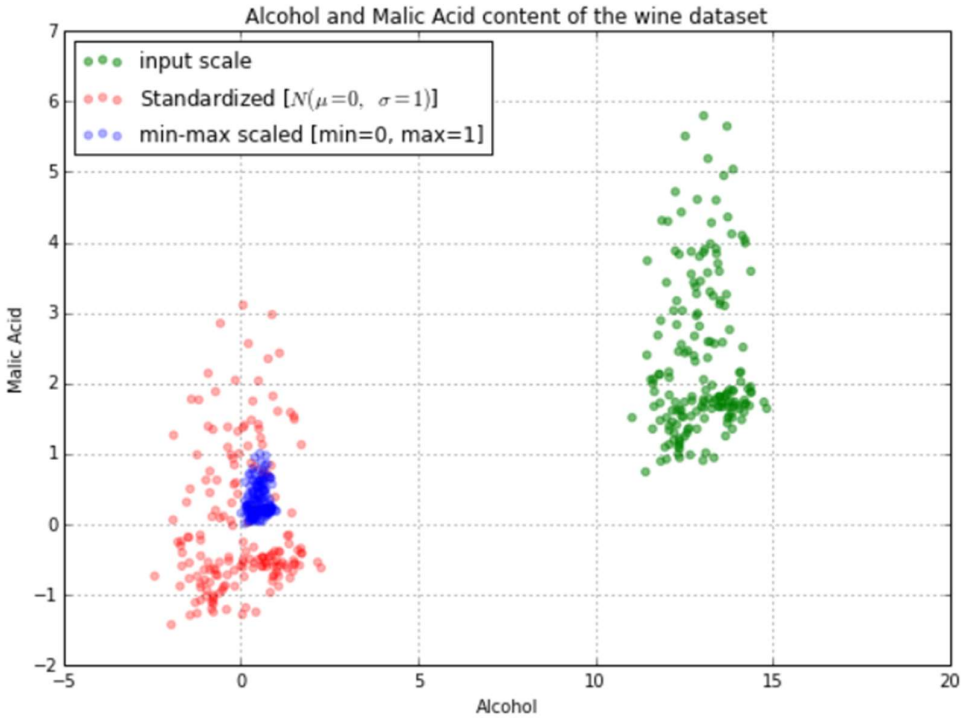
- Standardized Scaling - In this case, scaling is done in such a way that their mean = 0 and standard deviation = 1. Formula is -

    x = (x - mean(x))/std(x)

- Normalized/Min-Max Scaling - In this case, scaling is done in such a way that all the values lie between 0 and 1 with 0(min value) being lowest and 1(max value) being highest. Formula is -

    x = (x - min(x))/(max(x) - min(x))

Reference screenshot as follows -



Green dots are input data/sample. When feature scaling is applied to it -

1. The blue dots are Min-Max scaling where all the values lie between 0 & 1.
2. Red ones are standardised scaling where mean of the data points = 0 (equal/balanced data points lie towards left of 0 and right of 0 hence mean = 0).

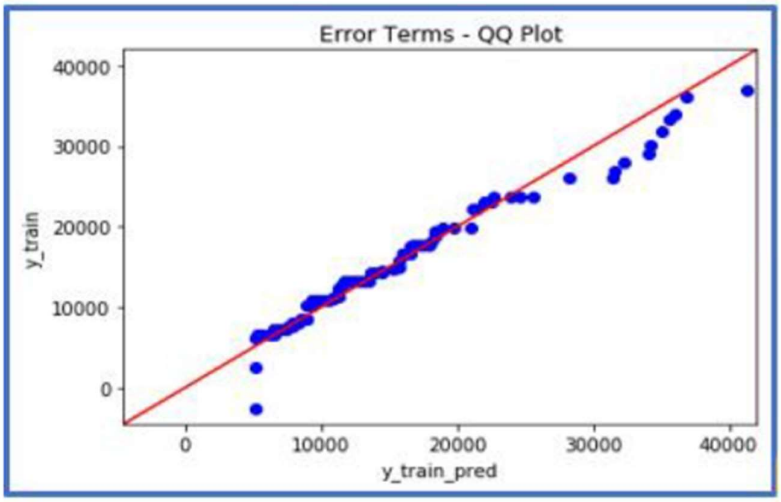
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- It happens when one of the feature variables is highly correlated with another or it has perfect correlation with another variable.
- In other terms, the feature variable is duplicate of the one of the other independent variables used in our model building.
- VIF value between 4-5 means moderate correlation while VIF = 0 means no correlation.
- In case of infinite/high VIF value, we need to drop one of the feature variables and rebuild the model.
- Other methods are picking the business interpretable variable, creating new variables by either adding new features to it or transformation of existing features (PCA).

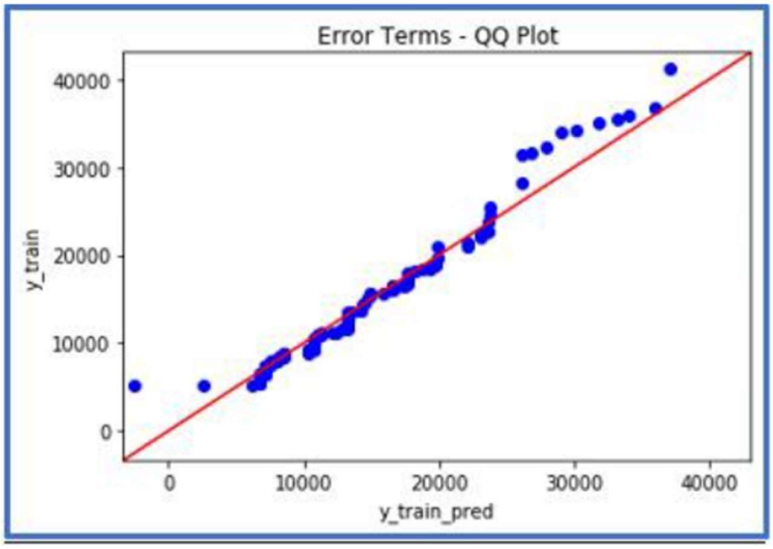
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Q-Q stands for Quantile-Quantile Plot which compares the quantiles of our data against the quantiles of the desired distribution.
- Quantiles divide our dataset into four parts - 0-25%, 25-50%, 50-75% and 75-100%.
- Each quantile of the first dataset is compared with the quantile of the second dataset.
- Q-Q plot helps us to identify if the dataset came from some distribution like Normal, Uniform or exponential distribution.
- It also helps to determine if two datasets come from populations with a common distribution or not.
- This is used in case of linear regression where we have two datasets - train and test dataset and Q-Q plot can be used to confirm if both these datasets are from the same population or not.
- It can be used with Sample Sizes as well.
- Few of the aspects like outlier's presence, changes in symmetry, shifts in scale/location etc can be identified from this plot.
- Below are the possible interpretations for two datasets -

    a. Similar distribution: When all the points of quantiles lie on or close to a straight line at an angle of 45 degree from x -axis.
    b. Y-values < X-values: If y-quantiles are lower than the x-quantiles.



    c. X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d. Different distribution: If all points of quantiles lie away from the straight line at an angle of 45 degree from x -axis.