Name: Le Thuc Anh
Email: thucanhle149@gmail.com
Course: Executive PG Programme in Machine Learning & AI - March 2023
Module: Linear Regression Assignment
Github link: https://github.com/1darknight/Bikeshares-LinReg

Part I : 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? (3 marks)

After doing the analysis, which contains EDA, Feature Selection, some categorical variables are found to have some impacts on demand of bike sharing:

Categorical columns that have positive impacts on demand are:
- **weekday_Sun** (from column weekday)
- **Month 3, 6, 9, 10**

It seems that people travel by sharing bike more on working day and Sunday.

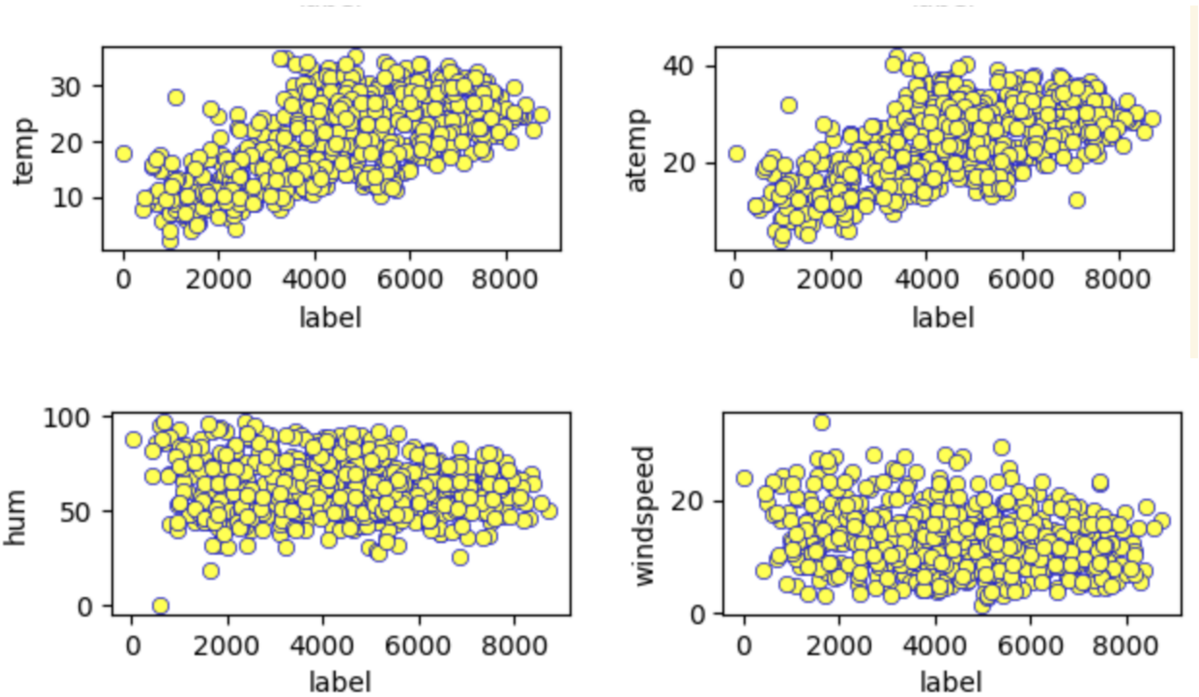
Columns that have negative impacts on demand are:
- **season_Spring** (from column season)
- **season_Winter** (from column season)
- **weathersit_Lightsnow** (from column weathersit)
- **weathersit_Mist** (from column weathersit)
- **Month 4** (from column mnth)

Combined with the columns has positive impacts, it seems that people use sharing bike less in Spring and Winter and when the weather is not encouraging (strong wind, snow, mist which reduce visibility)

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

When creating dummies, the function get_dummies will default created a boolean for each value in the column stated. But **if a column has n unique values, n-1 columns of boolean already cover all possible outcomes** as if all n-1 columns are False it means that left the n value is True. So that, **drop_first = True is to avoid created more variables than we need to describe the data.**

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



After plotting all 4 numerical variables against column 'label' (target variable), **column temp and atemp has the highest correlation with target variable**. There are a clear trend that all points lies from the lower right corner to the upper right corner which indicates **a positive correlation**.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

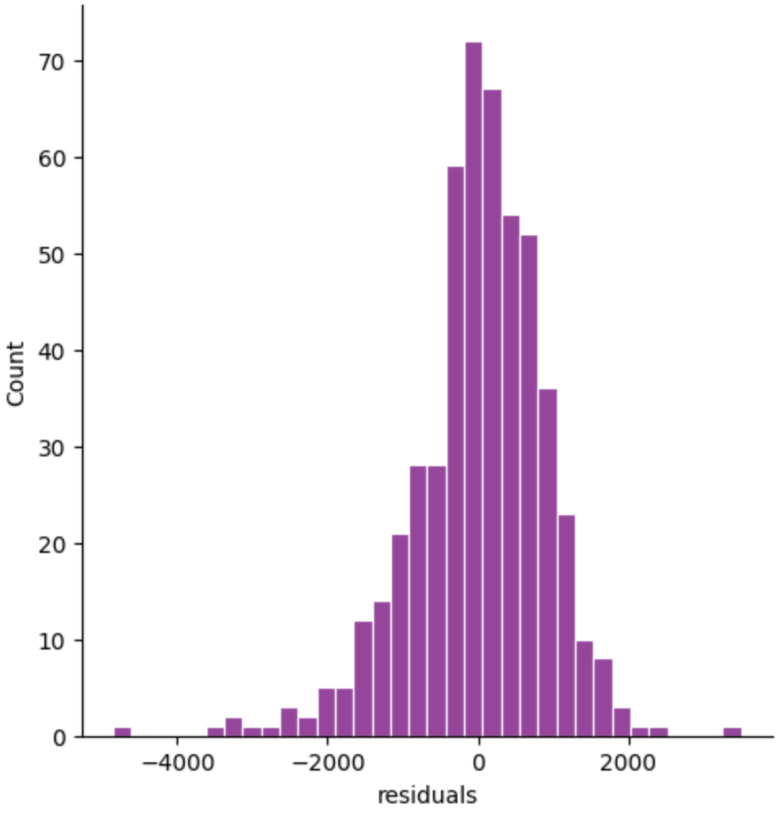
There are 4 assumptions of Linear Regression:

1. **There is a linear relationship between X and Y**:

After plotted the chart of numerical variables versus the target variables, there are clear that they have linear relationships with each other which means that we could fit a straight line through all the datapoints.

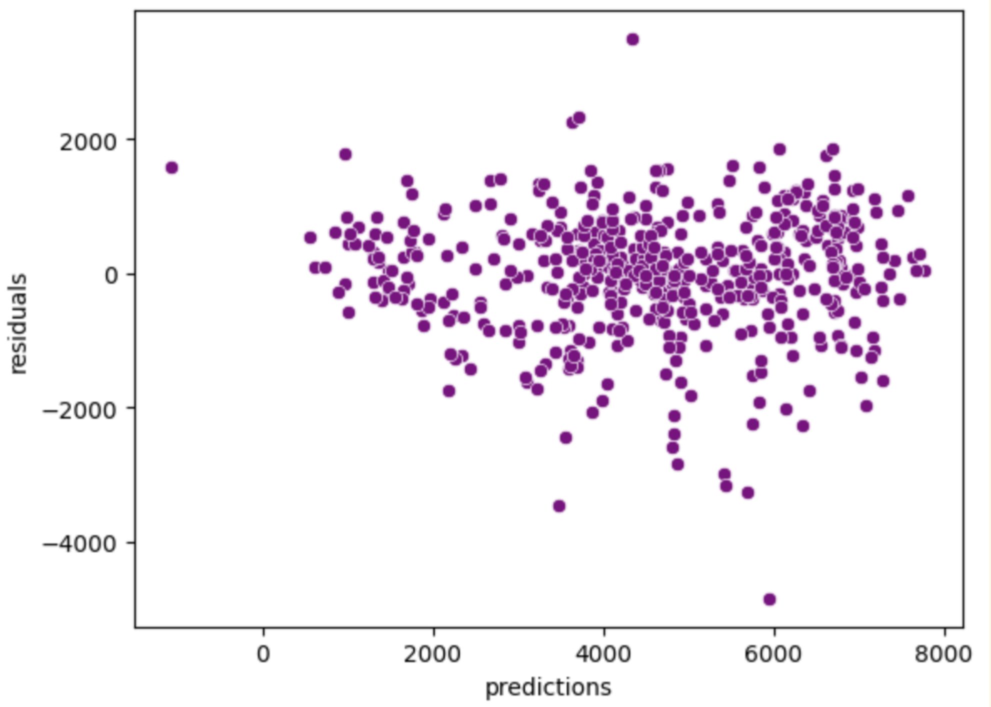
2. **Error terms are normally distributed with mean zero (not X, Y)**:

I plotted the histogram of residuals, the mean is approximately at 0 and the distribution is similar to a normal distribution.



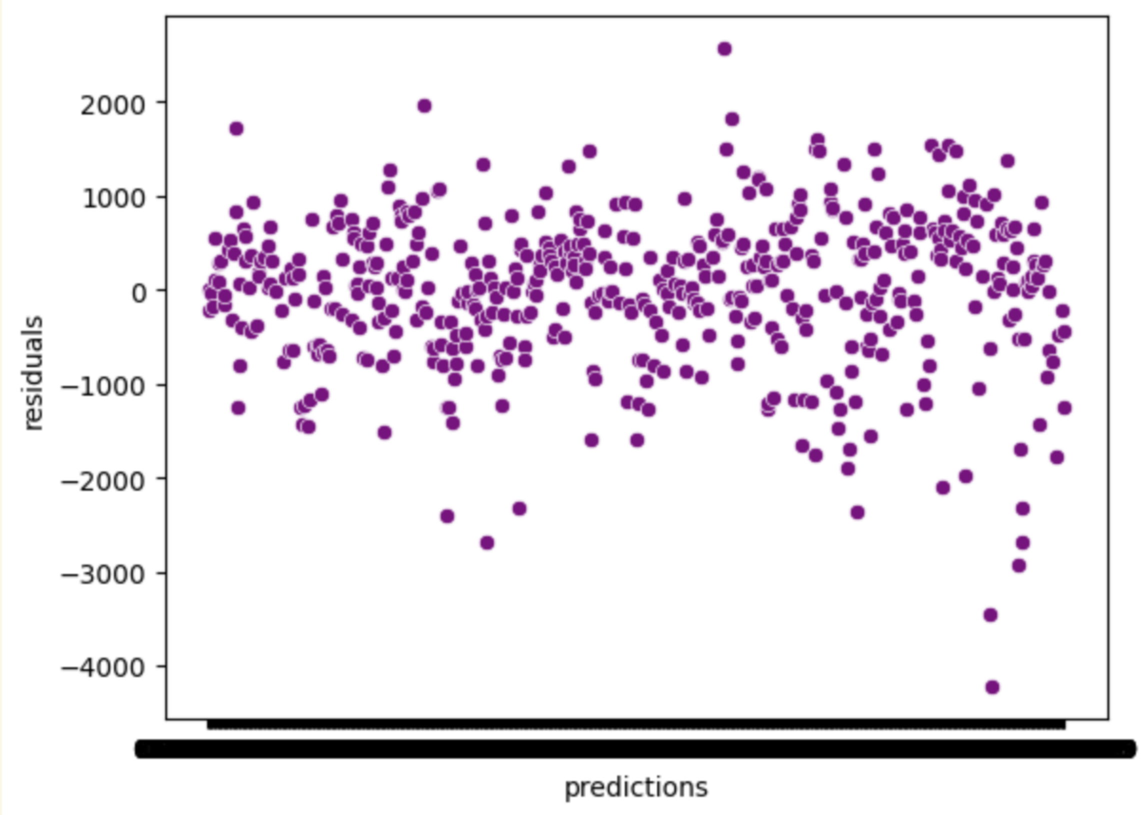
3. **Error terms are independent of each other**:

I plotted the residuals and predictions of y in train dataset, there are no visible patterns or trends in the below chart, so that the error terms is independent.



4. **Error terms have constant variance (homoscedasticity)**:

Plotting the residuals against the predictions, the variance of residuals through time is quite constant in the chart below.

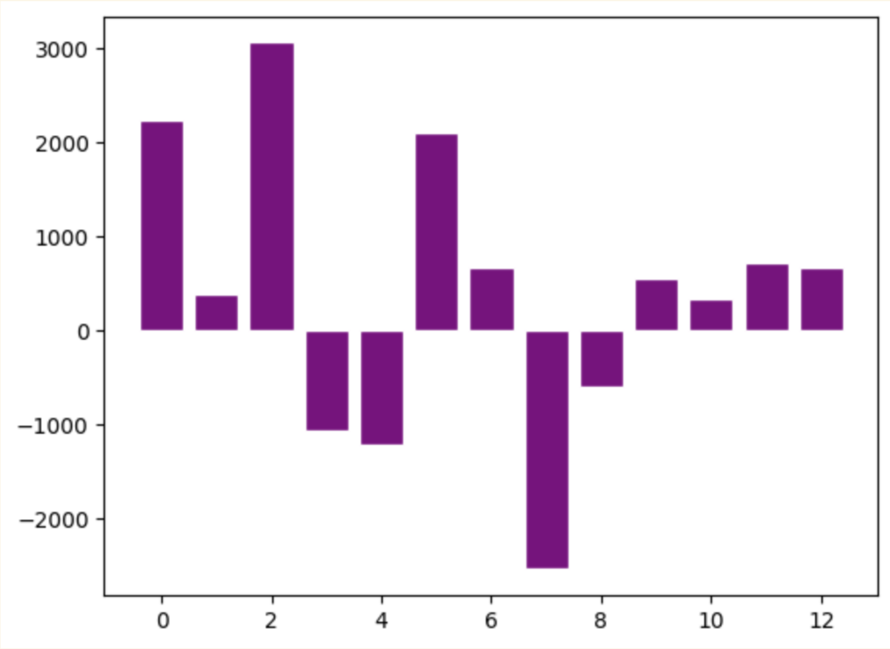


5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

So that the **top 3 features that impacts 'label' column (demand of shared bikes)** are:

- **temp** : coefficients of **3060.417** (column 2)
- **weathersit_Lightsnow**: coefficients of **-2544.739** (column 7)
- **yr_2019**: coefficients of **2101.623** (column 5)

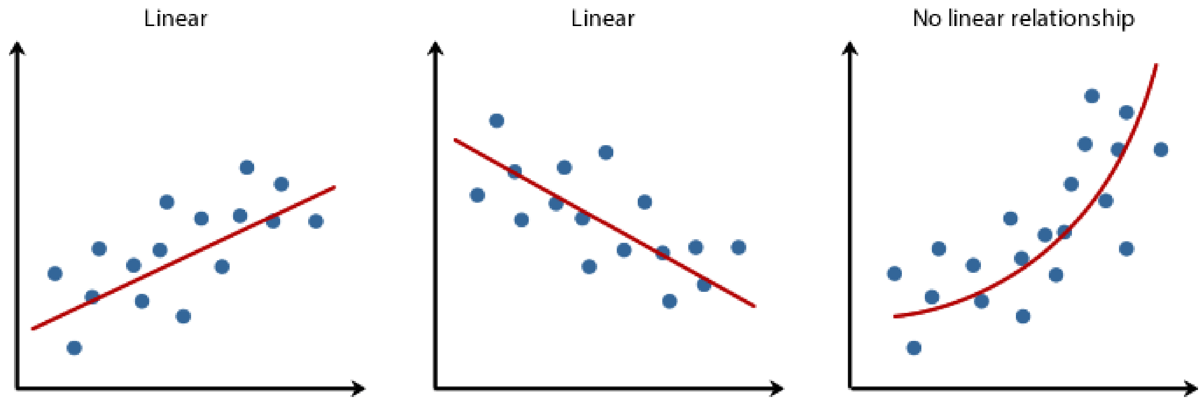
**Noted that column 0 is constant variable*



Part II : General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

A linear regression is trying to find linear relationship that indicate a correlation between independent variables and dependent variables. By fitting a straight line that representting the linear relationship between two variables that could either be positive (first chart) or negative (second chart).



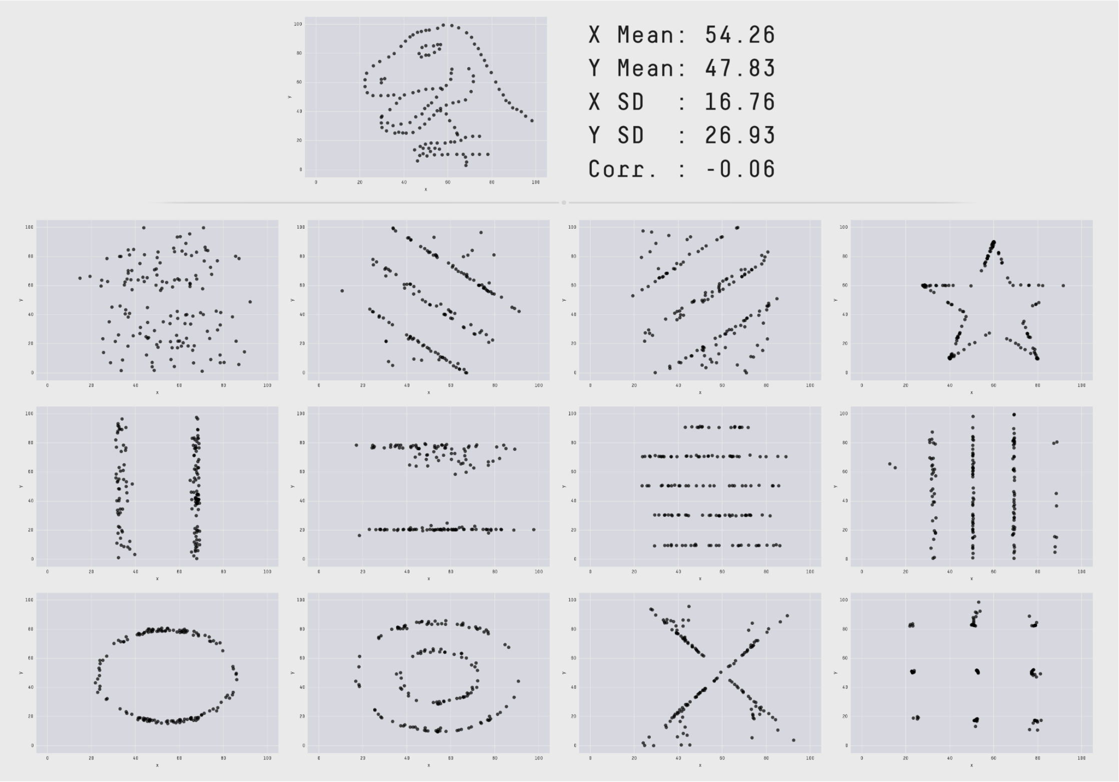
Copyright 2014. Laerd Statistics.

To find the best fitted line, linear regression used an algorithm called **Least Squares that find the smallest possible distance from each data point to the line**. And the formula (included constant, coefficients) for the best fitted line will be summarized along with others metrics (R squared, Adjusted R Squared…) iin the models summary function in Python.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a situation that **different dataset have the same descriptive statistics** such as Mean, Standard deviation, even Correlation Coefficients but have a really **different distributions** and when plotting, they have **different graphs.**

Here is an interesting example of Anscombe's quartet:



3. What is Pearson's R? (3 marks)

Pearson's R is a way to **measure correlation of a linear relationship between two variables**.

This metric ranges from -1 to 1 with below meaning in each interval:

- **From -1 to 0:** two variables has a **negative correlation** (two variables moving in opposite directions)
- **Equal 0**: Two variables is **not correlated or having any relationship**
- **From 0 to 1**: Two variables has a **positive correlation** (two variables moving in same direction)

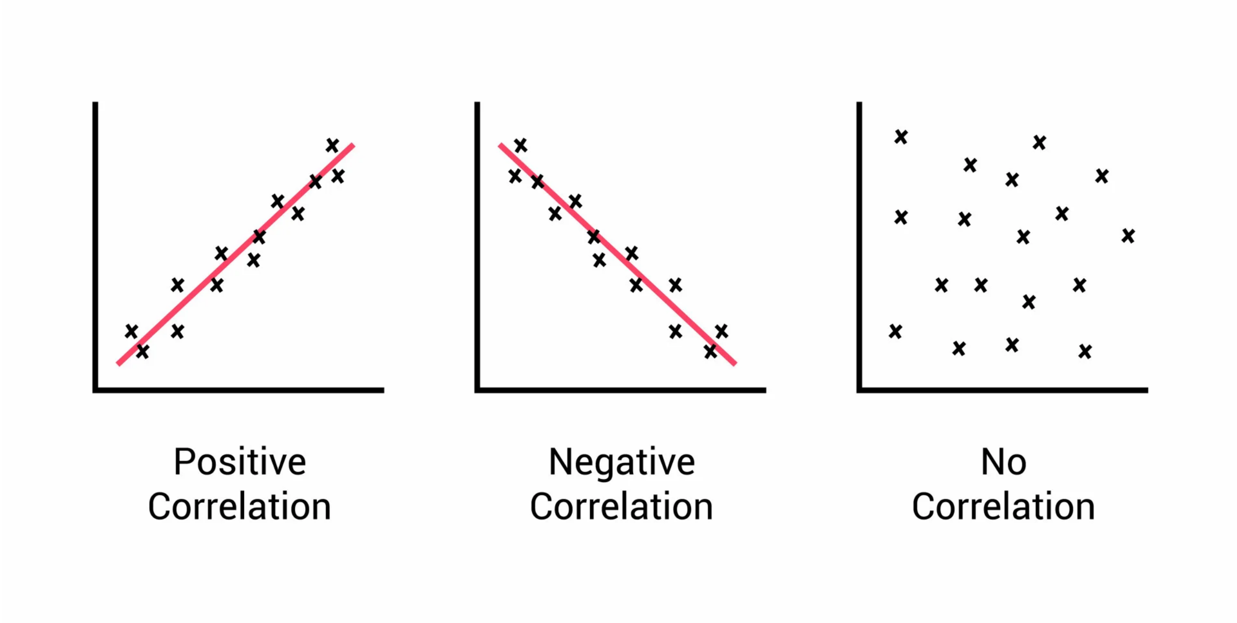
The formula to calculate Pearson' R (sometimes write in lowercase "r") is:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y}^2 - (\sum \mathbf{y})^2]}}$$

With x,y is two variables that we want to calculate the correlation.

It is included sum of x, sum of y, sum all cross products of x and y, sum of x squared and y squared.

Here is an example of three type of correlation above:



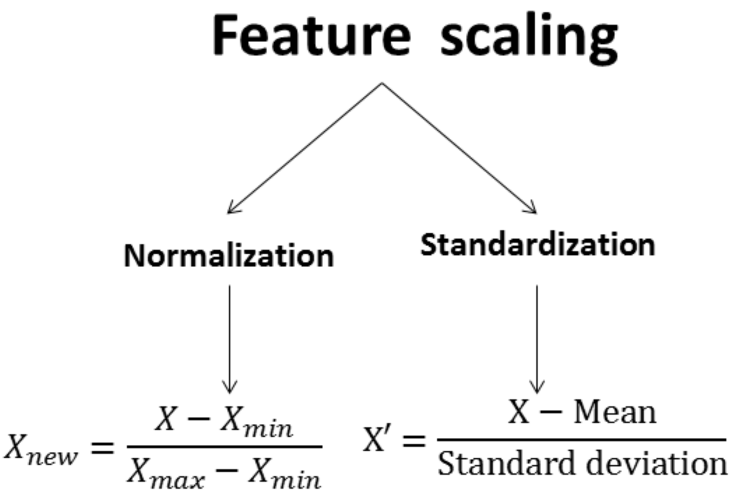
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a method in **data preprocessing for machine learning models** that **normalize the distance between values in independent variables**.
To **achieve a less bias and having equal variance** in values of all variables, scaling is performed.
Each type of scaling have different method to normalize the values:

- **Normalized scaling**: using **min and max values of each variable** as limits and the distance between values is scaled down or scaled up according to the two limits and still keep the original distance variance, with **the scaled range is [-1,1]**
- **Standardized scaling**: **shifting the whole distribution** of the variables to a Z distribution with variable's **mean = 0** and the variable's **standard deviation = 1**

The key differences of two methods is the output, **normalized scaling** is having a range form -1 to 1 which is **useful when the variable's distribution is not sure or unclear**, while **standardized scaling** is used when the variables is **similar to a normal distribution** that mean and standard deviation can describe the whole distribution.

The difference between two methods could be easier understand through this chart:

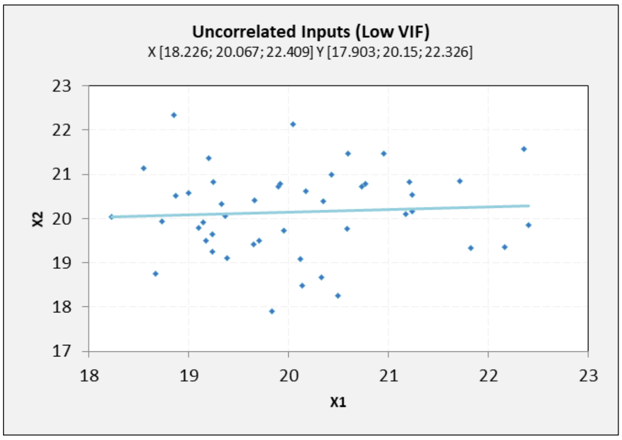


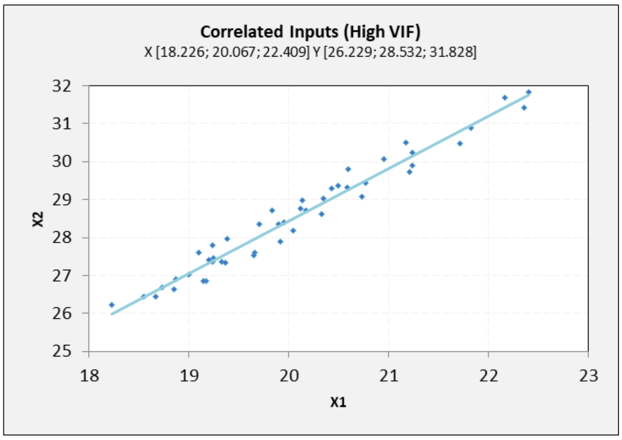
$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad \mathrm{X}' = \frac{\mathrm{X} - \mathrm{Mean}}{\mathrm{Standard\ deviation}}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The range of VIF is between 1 and infinity. When **VIF = 1**, two variables is **not having a collinearity** or in a simpler word, they are not correlated. In contrast, when **VIF is infinity**, two variables are perfectly correlated.

The chart below is plotting two situations of VIF between X1 variable and X2 variable:

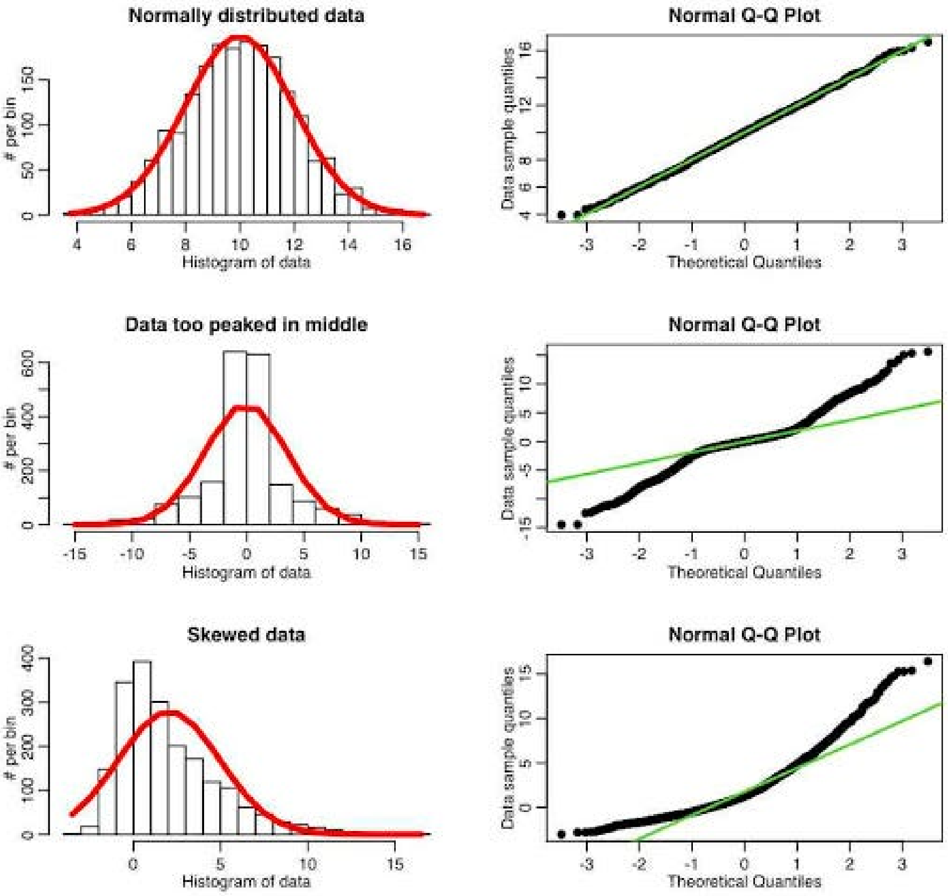




6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plot, also known as **Quantile plot,** is a plotting method to **see if a variable have the same quantile range as a normal distribution**.

The below chart is showing if a variable is similar to a normal distribution, the **data point will be align with the green line (representing a normal distribution).** So that we could conclude that each part of the distribution is similar to a normal distribution or not based on alignment with the according part of the normal line. The **min and max of the variable will be putting at lower left corner and upper right corner**.



--The End--