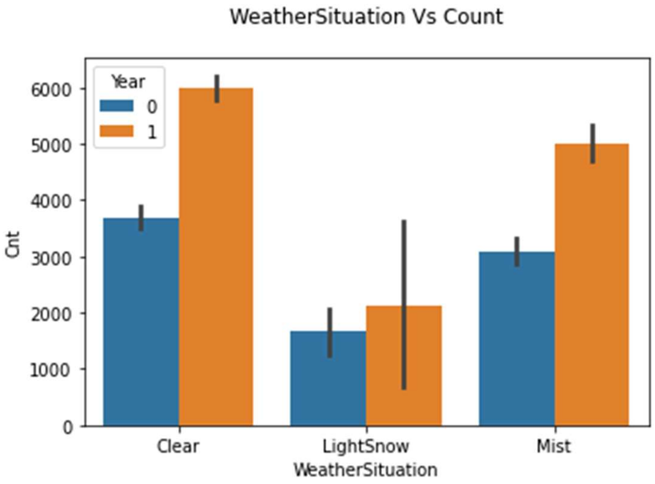
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

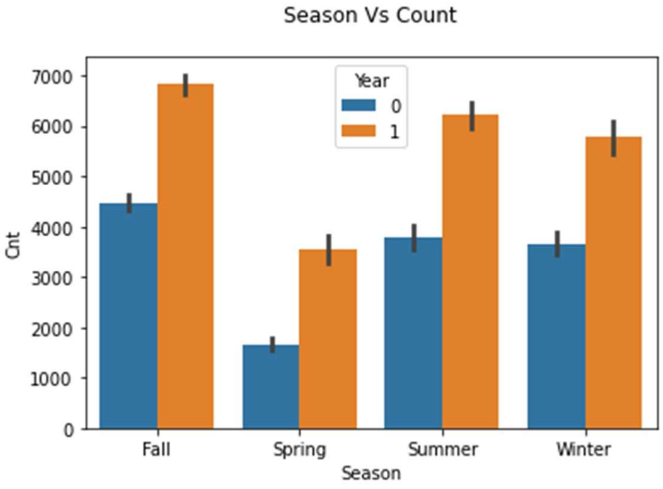
Answer:

Based on the analysis of categorical variables, below points are inferred.

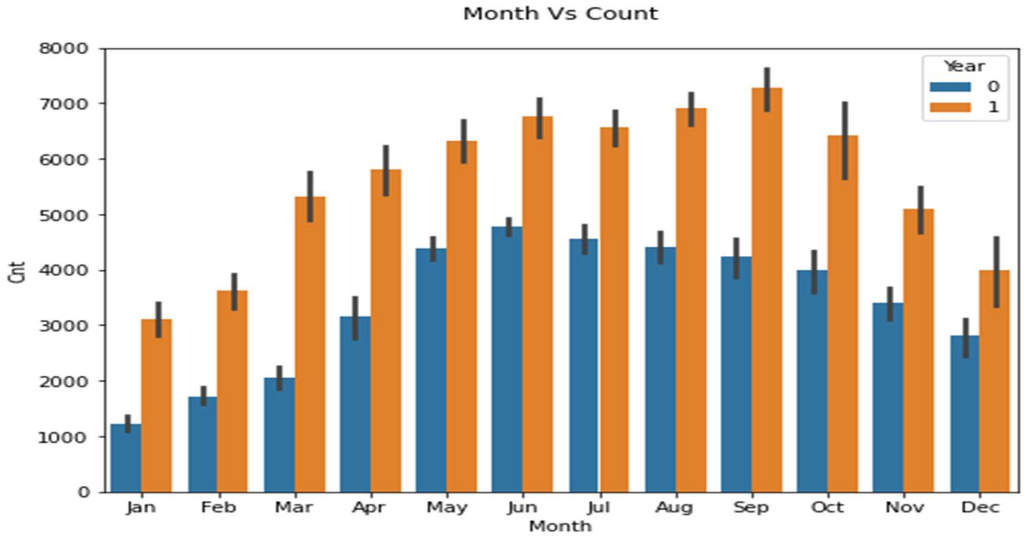
a. When the weather is clear, most customers go for rental. But when it is light snow, it attracts low customers



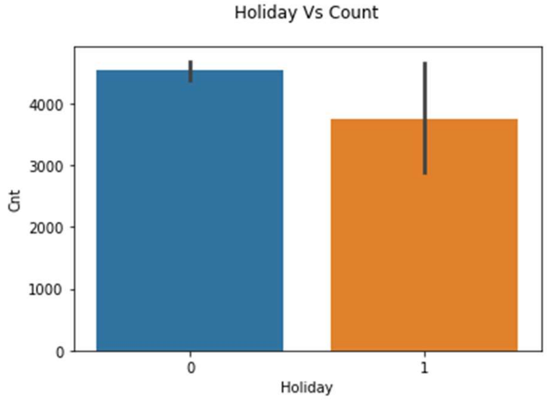
b. During Fall season, lot of customers rented the bike, but during Spring, very low customers used bike rentals



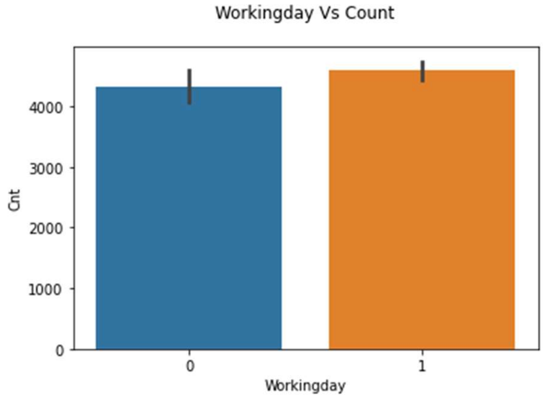
c. The rentals are increasing from Jan to Sept with a minor dip in Jul. And again, started decreasing.



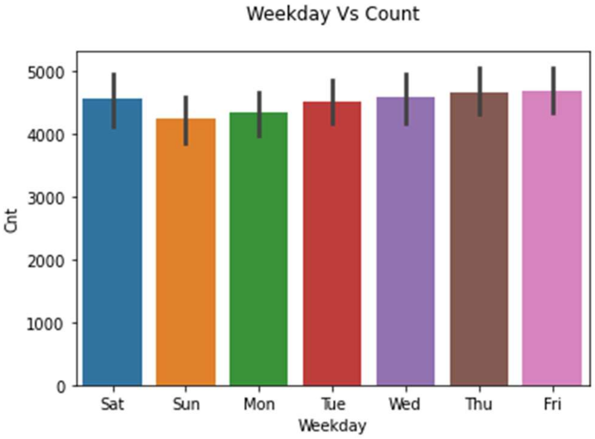
d. During non-holidays, the rentals are high compared to holidays.



e. Also, during working day, the rentals as high as double the non-working day count. This infers that people prefer to spend time with family during non-working day. And people may prefer to hire the bike for going to office.



f. We can see a considerable dip in average number of bookings on Sun & Mon



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

Answer:

Multi collinearity is the case where independent variables depend on each other. When we encode the categorical variables, it creates a new column for each category.
However, even if one category is deleted, we still can identify the category of the data clearly.

For Ex, if we have yes/no column, we don't need to create 2 columns with one for each. We can have only one column which represent either 1 or 0.
Similarly, for n categories, we can create n-1 columns. So, we use drop_first to drop first column.

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

Answer:

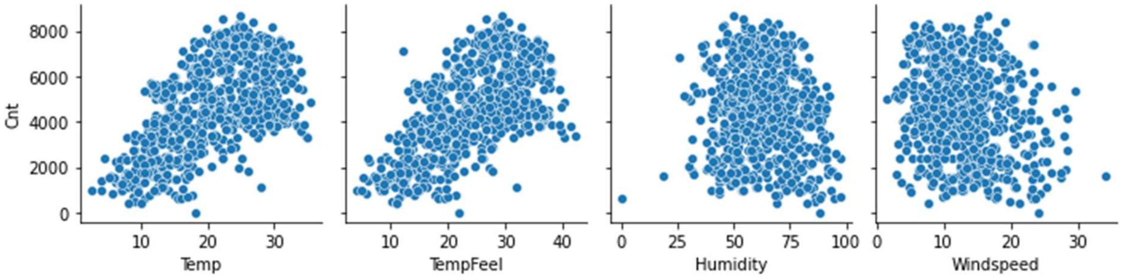
Temp and atemp has high correlation with cnt as 0.63

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

Answer:

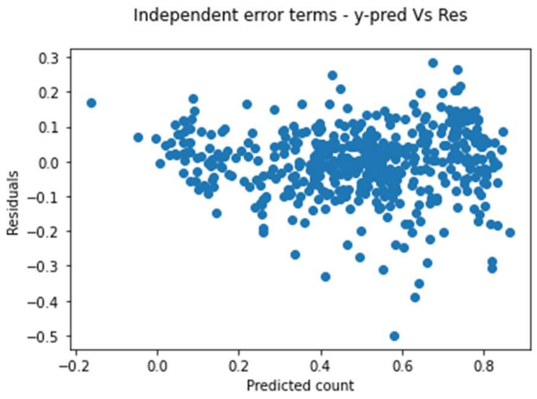
Below are the assumptions that are validated -

- A linear relationship between the dependent and independent variables



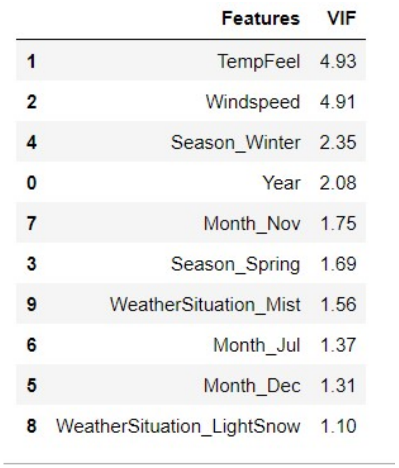
Observations:
- Temp & TempFeel are in linear relation with target variable
- Though Humidity and Windspeed are not in linear relation, based on domain information, as they are important features, they are considered as part of the model.
- However, Humidity was dropped due high collinearity

- Independent error terms



Observations:
- As the error terms does not follow any pattern, we can say that the error terms are independent of each other

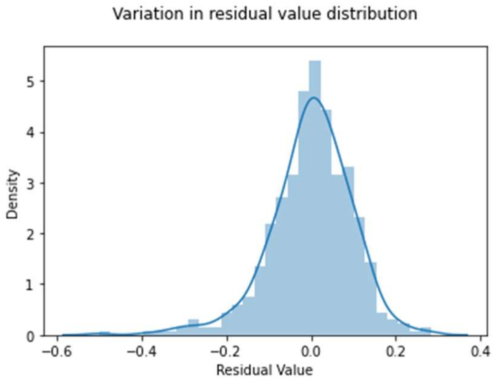
- The independent variables are not highly correlated with each other



	Features	VIF
1	TempFeel	4.93
2	Windspeed	4.91
4	Season_Winter	2.35
0	Year	2.08
7	Month_Nov	1.75
3	Season_Spring	1.69
9	WeatherSituation_Mist	1.56
6	Month_Jul	1.37
5	Month_Dec	1.31
8	WeatherSituation_LightSnow	1.10

Observations:
- The VIF Values indicates the multicollinearity is minimal (General consideration is VIF should be less than 5)

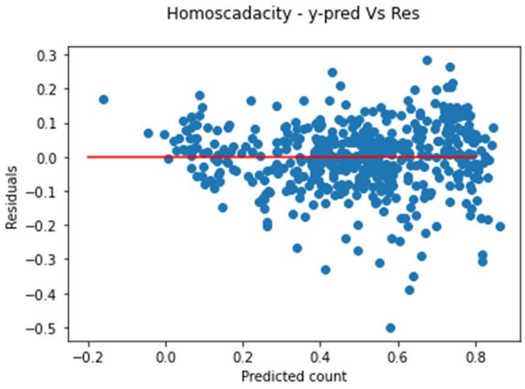
- The Mean of the residuals is centred to zero



Observations:

- The mean value is almost equal to zero
- The residual plot is normal distributed with a mean centred to zero

- Assumption of Homoscedastic



Observations:

- There is no definite pattern (like linear or quadratic or funnel shaped) in the above scatter plot
- This proves that the residuals have equal or almost equal variance across the regression line
- Since p value is more than 0.05 in Goldfeld Quandt Test, we can say that error terms are homoscedastic

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

Below are the 3 variables:

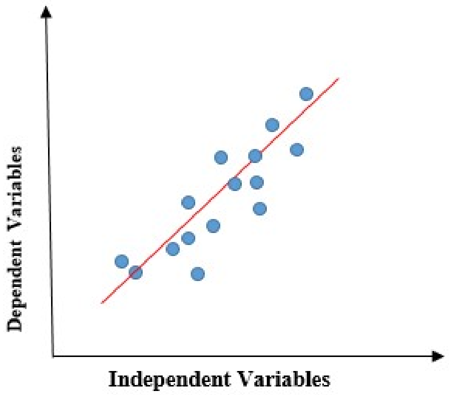
- TempFeel (atemp)
- Year
- WeatherSituation_LightSnow

General Subjective Questions

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

Answer: Regression is a supervised machine learning technique that supports finding the correlation among variables. A regression problem is when the output variable is a real or continuous value.

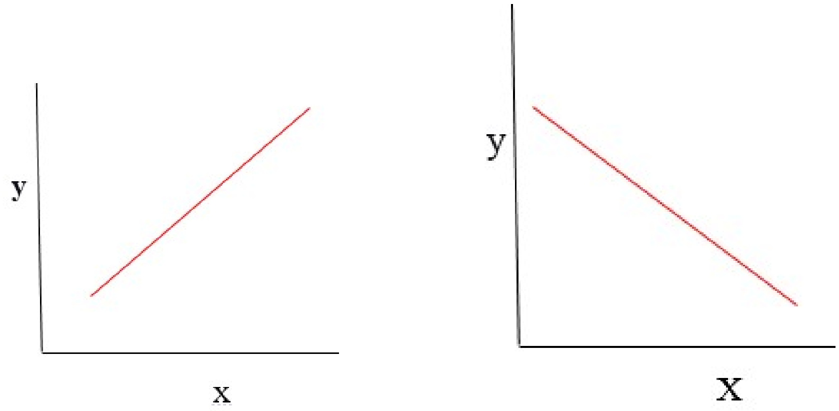
The mathematically, it can be represented as: $y = mX + c$



Here the m signifies how strong the relationship between x & y. m is also called slope or gradient.
m can also be represented as: change in y/change in x
The value of c is called intercept

Generally, regression data is divided into 2 sets of variables – Independent & Dependent
Independent – Based on these variables, the output variable has to be determined
Dependent – This is the variable that needs to be predicted

There are 2 types of relations that we can plot:
Positive – Dependent variable increases if independent variable increases
Negative – Dependent variable increases if independent variable decreases



In linear regression (Simple or multiple), we need to find best fit line which represents the linearity between dependent and independent variables.
And the line should be in such a way that the Mean Squared Error should be minimal

While finding the best fit line, we need to make sure that there is no multicollinearity between independent variables.

Once the best fit line has been built, we have to assess the model by calculating the r-squared which explains the variability.
r-squared takes values from 0 to1. 0 means, no variability is explained whereas 1 means, complete variability is explained.
F-Statistic determines if the model is significant or not. If the F-statistic value is near zero, the model is significant.

However, one drawback of r-squared is, even if we add any insignificant variable to the model, the value will increase or remain constant. To eliminate this drawback, adjusted r squared is introduced by penalizing the model for using high number of predictors.

Assumptions of linear regression:
Apart from linearity, below are the assumptions that we need to consider

- Linear relationship between x and y
- Error terms are normally distributed

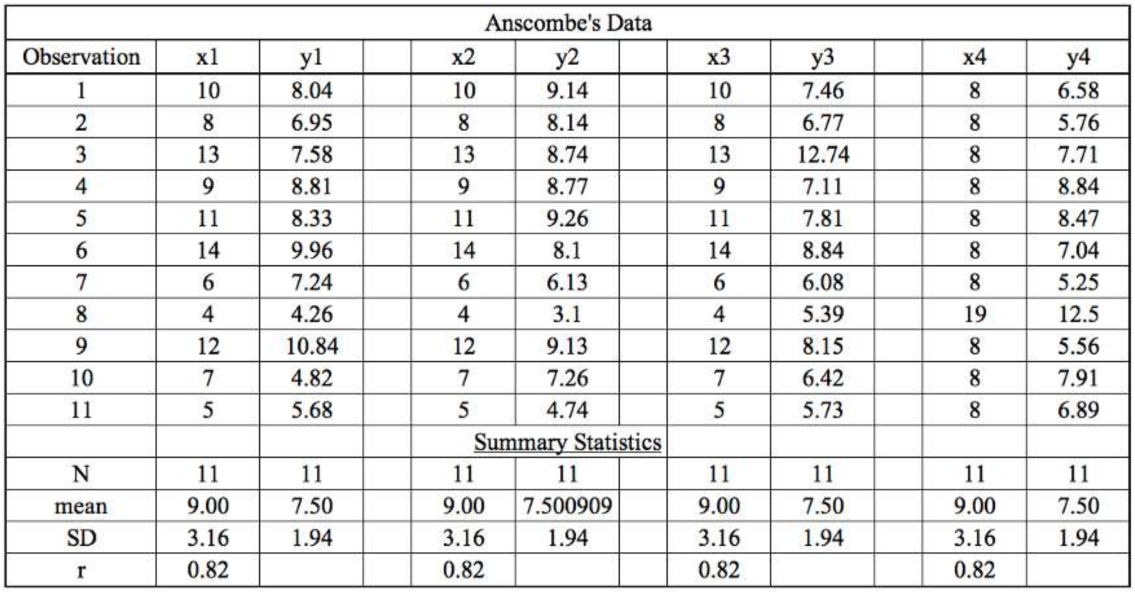
- Error terms are independent of each other
- Error terms have constant variance

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

Answer:

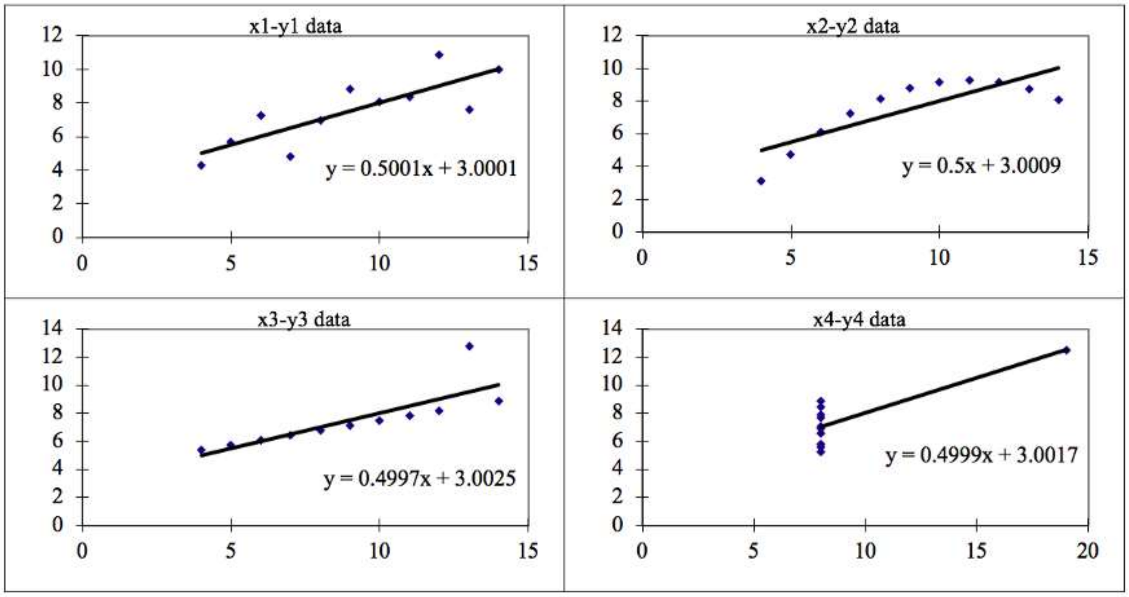
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Suppose if we have datasets like below, the statistical information for all these four datasets are almost similar.



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

But when they are plotted on graph, they generate different type of plots



The four datasets can be described as:

Dataset 1: This fits the linear regression model in a good way.
Dataset 2: This dataset does not follow any linear pattern.
Dataset 3: This dataset shows the outliers which cannot be handled by linear regression model
Dataset 4: This dataset almost constant value on x axis and shows the outliers involved

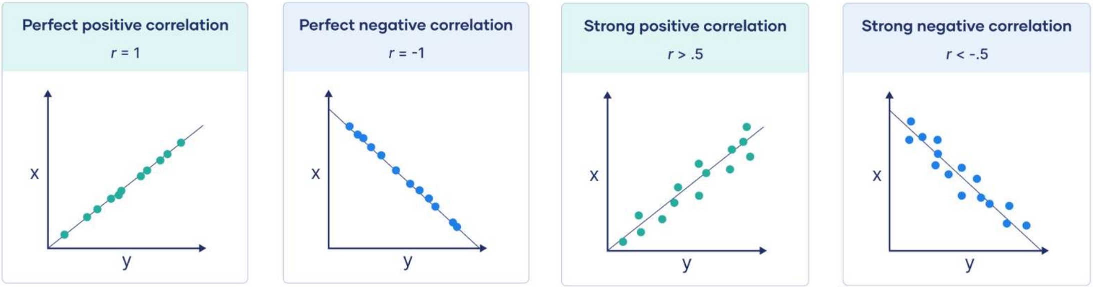
Conclusion: As the final statistical conclusions can deceive us, it's always advisable that all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

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

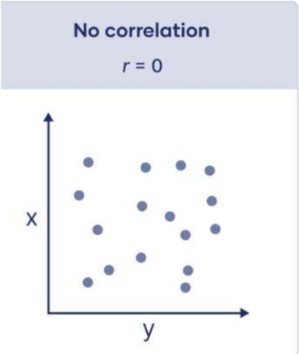
Answer:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It has values between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.



But when r=0, we can say that there is no correlation between the variables.



To use Pearson correlation, below conditions should be satisfied.

- Both variables are quantitative
- Variables are normally distributed
- Data has no outliers
- Relationship between the variables should be at least reasonably linear

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer:
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed to handle highly varying values.
If the features are not scaled, large values are weighed more and small values are weighed less.
For ex., Suppose, salary and age are the features. As the value of salary may go beyond 6 or 7 digits and age can have max value ~120, salary will be considered high weight and age will be considered low weight.

There are two most important scaling techniques:

Min-Max Normalization:
This technique re-scales a feature value between 0 and 1.
X_New = (Xi - min (X)) / (max (X) - min (X))

Xi can have min(X) and max(X) as ends. If X = min(X), the X_New will be 0 and if it is max(X), then the value will be 1. So, all the new values will fall between 0 and 1

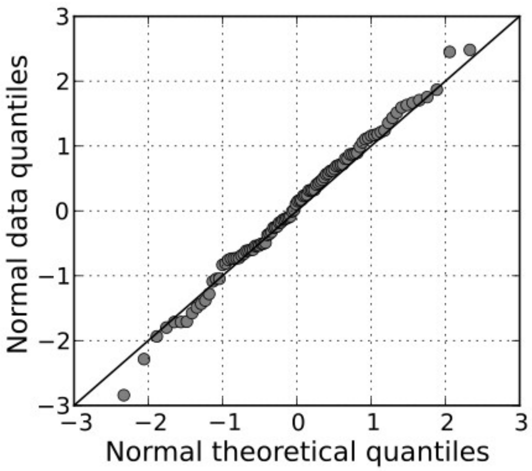
Standardization:
It is a process of re-distributing the values in such a way, the new mean will be 0 and new standard distribution will be 1.
X_New = (Xi - X_mean) / Standard Deviation
It says, how many standard deviations aways from mean. So, most of the values will range from -3 to +3 as 99.7% of data falls between 3 standard deviations.

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

If there is perfect correlation, then VIF = infinity. This indicates that a perfect correlation between two independent variables. In the case of perfect correlation, we get r-squared =1, which lead to 1/(1-r-squared) as infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the variable can be expressed exactly by a linear combination of other variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Q-Q(quantile-quantile) plots are used to compare two probability distributions by plotting their quantiles against each other. If the two distributions are exactly equal then the points on the Q-Q plot will perfectly lie on a straight-line y = x.

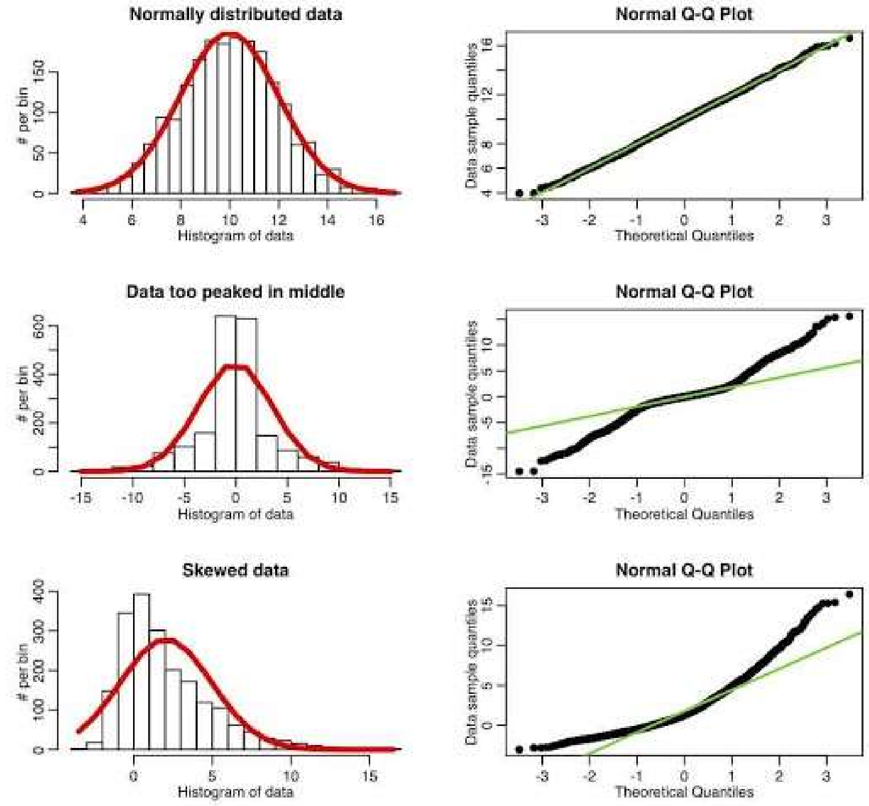


We plot the standard normal distribution quantiles on the x-axis and the normal data quantiles on the y-axis. This gives a smooth straight line like structure from each point plotted on the graph.

Now we need to see the ends of the straight line. If the points at the ends of the curve formed from the points are not falling on a straight line but scattered significantly from the positions then the data is not Normally distributed.

If all the points plotted on the graph perfectly lies on a straight line, then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate.

Below are some of the q-q plots with different distributions.


Normally distributed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Data too peaked in middle
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Skewed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles