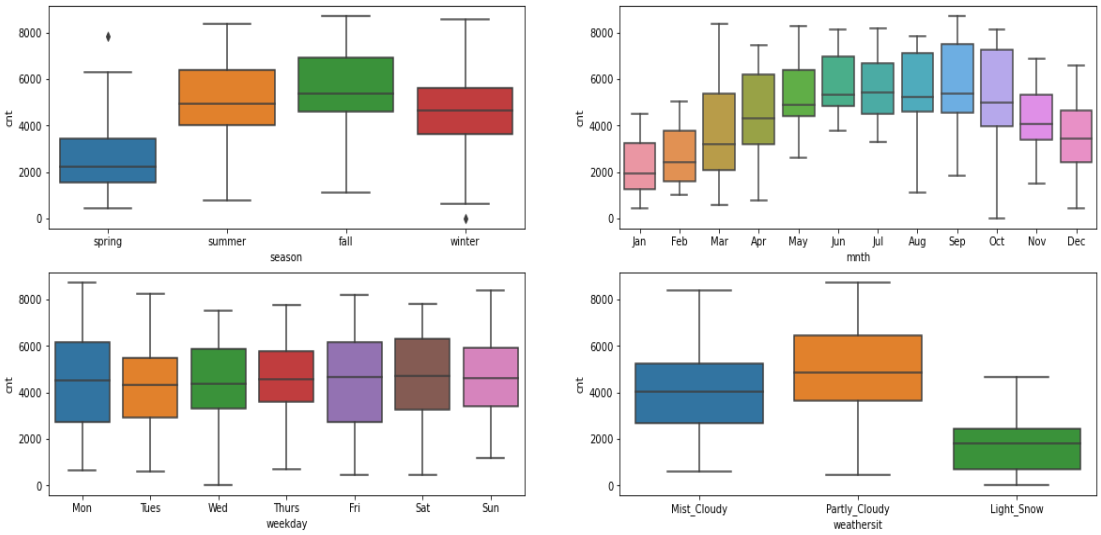
Assignment-based Subjective Questions

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



- There are 4 Categorical columns that are plotted. They are seasons, mnth, weathersit, and weekday.
- From the above plot, it is evident that there is no outliers except spring and winter of seasons.
- My Inference are as follows:
  - Season:
    - More bikes are rent during season 3 which is fall season whereas less bikes are rented during season 1 which is spring.
    - Minimum value in all the four seasons lies in the same range: 0-2000.
    - Maximum value in 'spring' is in between 6000 and 8000 where as in other three seasons, 'summer', 'fall' and 'winter', it is above 8000.
  - Month:
    - Maximum bikes are rent during september month while the least rides happened by October.
  - Weekday:
    - More bikes are rented on Saturday, Monday and Friday.
  - Weathersit:
    - More bikes are rent during Clear, Few clouds, Partly cloudy weather.
    - Average value w.r.t to the target variable is in range of 4000 - 6000.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   - drop_first=True is more common in statistics and often referred to as "dummy encoding", while using drop_first=False is reffered as "one hot-encoding".
   - drop_first=True, helps in reducing the extra column created during dummy variable creation. It allows you whether to keep or remove the reference (whether to keep k or k-1 dummies out of k categorical levels). If provided drop_first = False, then the reference is not dropped and k dummies are created out of k categorical levels, hence there occurs multicolinearity.
   - So drop_first = True, then it will drop the reference column after encoding. Hence it reduces the correlations created among dummy variable.

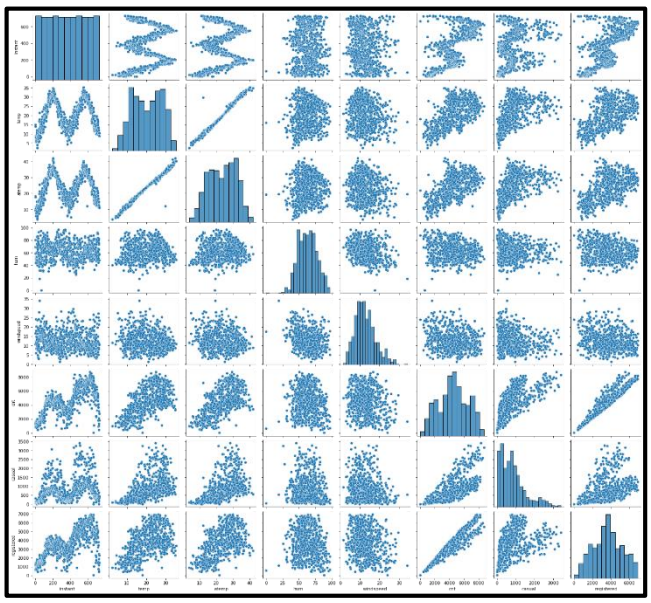
Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

   - As shown in figure, the 'furnishing status' column is processed with dummy variable creation. Since the column unfurnished, according to this figure is not seen in the indicator variable, it is evident that drop_first = True is given and hence it ensures dummy encoding. Also, from the figure, it is evident when both the columns are having the value 0, then it is understood that the value indicates 'unfurnished status'. This reduces the correlations created among the indicator variables.

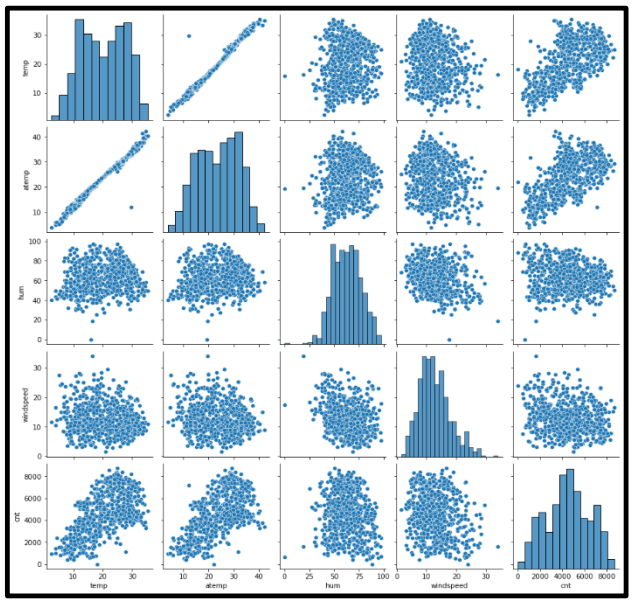
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   The numerical variables 'registered', 'atemp' 'temp' have the highest correlation with the target variable 'cnt', when considering all features. After data preparation, when we drop atemp due to multicollinearity the numerical variable **'temp' has the highest correlation with the target variable 'cnt'.**

Pairplot – With all integers



Pairplot – interger values of the model



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

Linear Regreesion has 4 Assumptions:

1. Errors will be having normal distribution that it's mean = 0 and variance is constant
2. No heteroscedascity for error terms
3. No multicolinearity seen within error terms
4. Linear relationship must exist among the error terms.

I validate all these 4 assumptions by plotting graphs with

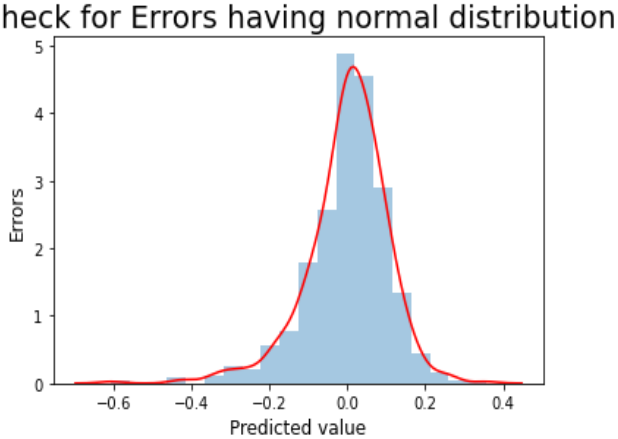
- **residuals:** which are the error terms calculated by formula

  residuals = y_train - y_train_pred

- **y-train:** which are the actual data trained for the model
- **y-train-pred:** which are the predicted data through analysis of the model.

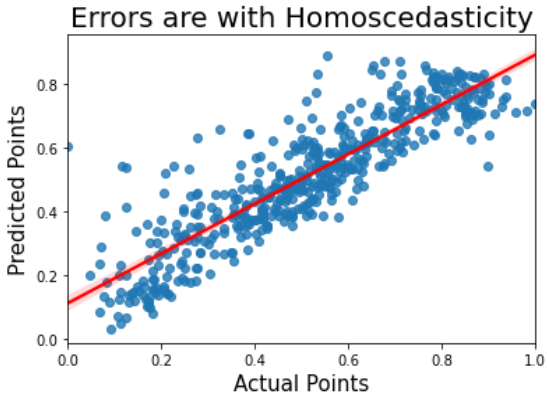
Assumption: 1.

This I validated by plotting distplot with residuals ($Y_{ACTUAL} - Y_{PRED}$).



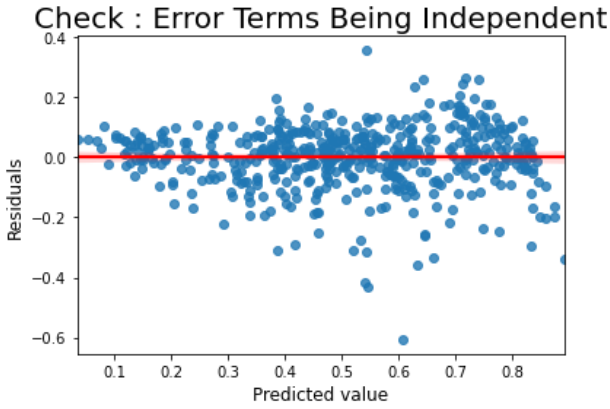
Assumption: 2

This I validated by plotting regplot with y_train vs y_train_pred.



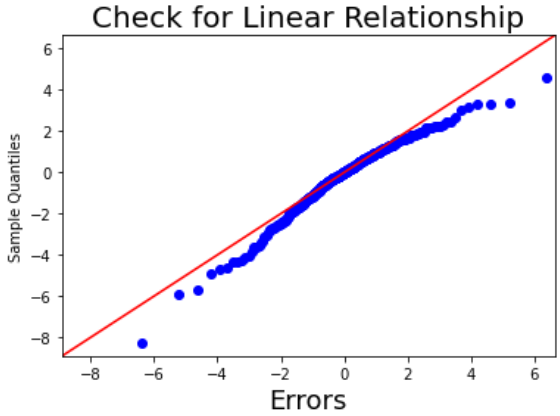
Assumption: 3

This I validated by plotting regplot with y_train_pred vs residuals.



Assumption: 4

This I validated by plotting QQPlot with residuals.



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

windspeed, temperature and year are the top three features *with VIF values 3.59, 3.52 and 2.02 respectively*. These are the variables that contribute significantly towards the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Introduction

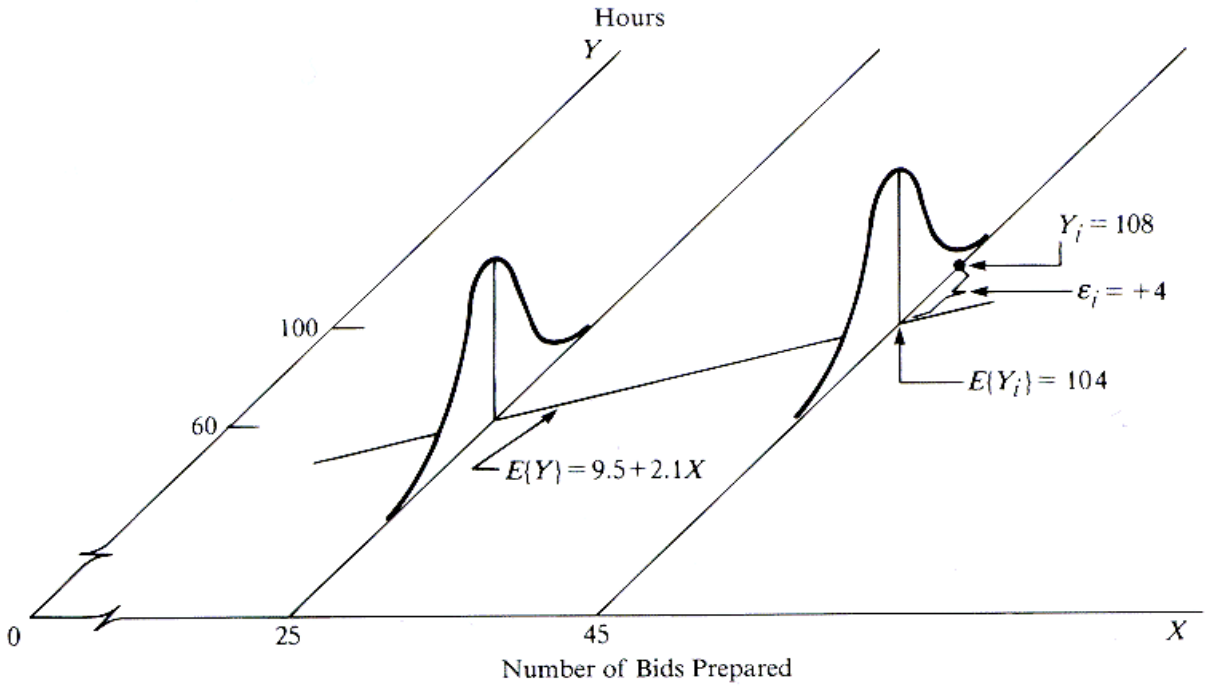
Regression analysis is commonly used for modeling the relationship between a single dependent variable Y and one or more predictors. When we have one predictor, we call this **"Simple Linear Regression"** which is noted as **SLR.** So the equation becomes:

$$E[Y] = \beta_0 + \beta_1 X$$

That is, the ***expected value*** of Y is a straight-line function of X. The betas are selected by choosing the line that minimizing the squared distance between each Y value and the line of best fit. The betas are chose such that they minimize this expression:

$$\sum e_i^2 = (y_i - (\beta_0 + \beta_1 X))^2$$

Illustration of Simple Linear Regression Model



Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable. However, in practice we often have more than one predictor. For example, in the Advertising data, we have examined the relationship between sales and TV advertising. We also have data for the amount of money spent advertising on the radio and in newspapers, and we may want to know whether either of these two media is associated with sales. How

can we extend our analysis of the advertising data in order to accommodate these two additional predictors?

When we have more than one predictors, we call it **multiple linear regression (MLR)**. Thus the equation of **MLR with n predictors** is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \ldots + \beta_n X_n$$

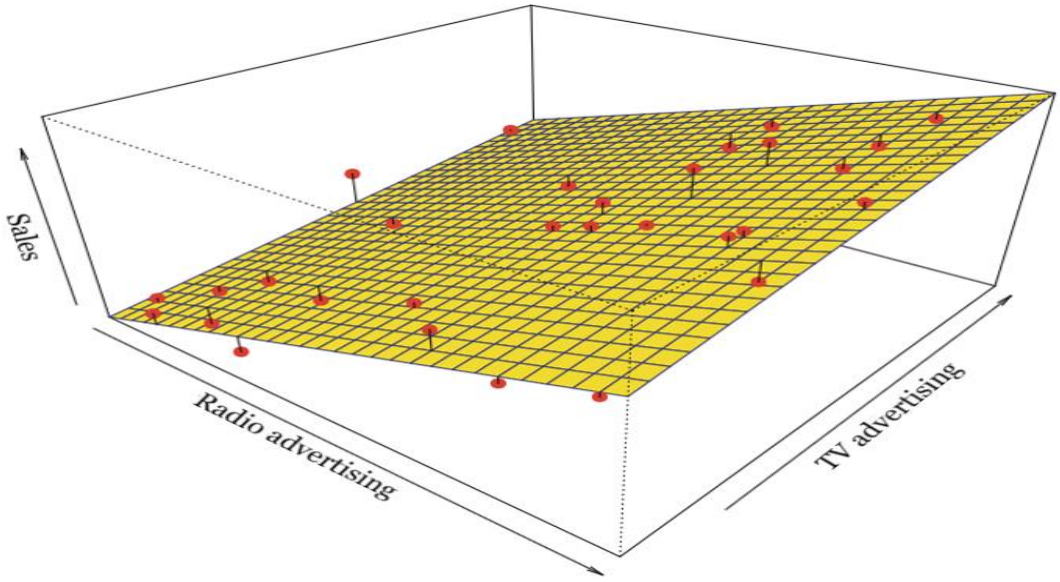
The **fitted** values (i.e., the **predicted** values) are defined as those values of Y that are generated if we plug our X values into our fitted model.

The **residuals** are the fitted values minus the actual observed values of Y.

Here is an example of a linear regression with two predictors and one outcome:

Instead of the "line of best fit" there is a **"plane of best fit"**.

Below figure, shows the pictorial representation of *sales as target variable* and **two predictor variables** as ***TV advertising and Radio advertising***. Hence how the plane is created for MLR.



There are four assumptions associated with a linear regression model:

1. **Linearity:** The relationship between X and the mean of Y is linear.
2. **Homoscedasticity:** The variance of residual is the same for any value of X.
3. **Independence:** Observations are independent of each other.
4. **Normality:** For any fixed value of X, Y is normally distributed.

2. Explain the Anscombe's quartet in detail.

Perhaps the most elegant demonstration of the dangers of summary statistics is Anscombe's Quartet. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed.

Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".

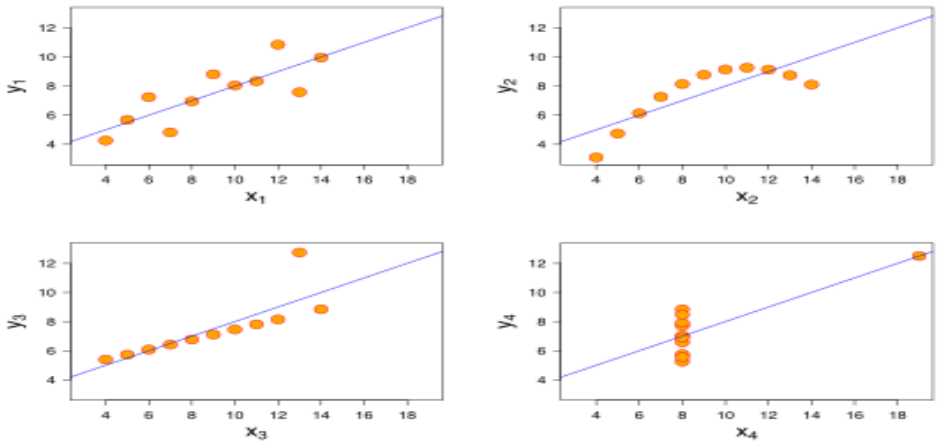
Eg: Consider the 4 datasets with eleven (x,y) pairs values as:

Count pairs in dataset	Dataset I		Dataset II		Dataset III		Dataset IV	
	X	Y	X	Y	X	Y	X	Y
I	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
II	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
III	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
IV	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
V	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
VI	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
VII	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
VIII	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
IX	12.0	10.84	4.65	12.0	9.13	12.0	8.15	8.05
X	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
XI	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

When computed summary statistics, observed results are to be identical as:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:



Explanation for the above figure:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets. Anscombe's quartet suggests that the variables in the dataset must be plotted to observe the sample distribution which help in identifying the anomalies in the data.

3. What is Pearson's R?

Correlation coefficients are used to measure how strong a relationship is between two variables. The most common measure of correlation in stats is the Pearson Correlation. Even though, there are several types of correlation coefficient, the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of

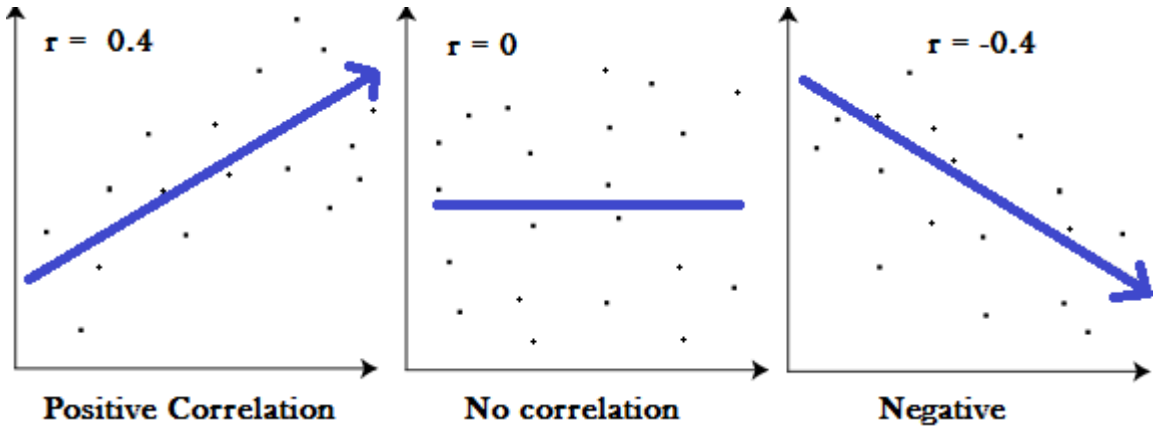
data. Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson's.

Pearson's correlation coefficient (r) or the coefficient of determination are the statistical indices to evaluate the performance of developed models. The coefficient of Determination is the square of Coefficient of Correlation. i.e. r = R2

Pearson's r is usually used to express the correlation between two quantities. i.e., It is a measure of the strength of a linear association between two variables, which is denoted by r.

- Pearson's R is the correlation coefficient that lies between -1 and +1.
- R = -1 and +1 means the data is perfectly linear with negative and positive slopes respectively.
- R = 0 means there is no linear correlation in the data 0<R<5 means there is a weak association.
- 5<R<8 means the association in data is moderate and R>8 means a strong association.



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

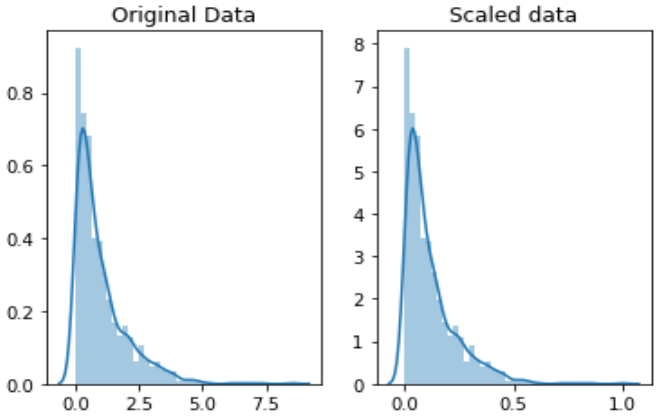
Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbours (KNN) where distance between the data points is important. Scaling is a monotonic transformation. Scaling method should be considered as an important hyper-parameter of analysis model(s).

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to

weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Some examples of algorithms where feature scaling matters are:

- **K-NEAREST NEIGHBORS (KNN)** with a Euclidean distance measure is sensitive to magnitudes and hence should be scaled for all features to weigh in equally.
- **K-MEANS** uses the Euclidean distance measure here feature scaling matters.
- *Scaling* is critical while performing **PRINCIPAL COMPONENT ANALYSIS (PCA).** PCA tries to get the features with maximum variance, and the variance is high for high magnitude features and skews the PCA towards high magnitude features.
- We can speed up **GRADIENT DESCENT** by scaling because θ descends quickly on small ranges and slowly on large ranges, and oscillates inefficiently down to the optimum when the variables are very uneven.



If you the shape of the data doesn't change, but that instead of ranging from 0 to 8 is how changed between 0 to 1 after scaling.

When performing scaling:

- Mean centering does not affect the covariance matrix
- Scaling of variables does affect the covariance matrix
- Standardizing affects the covariance

There are two types of scaling:

1. **Min-Max Normalization.**
2. **Standardization**

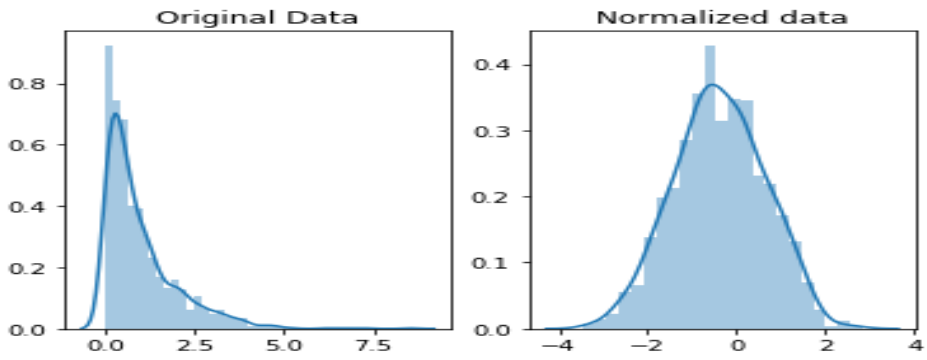
Min-Max Normalization:

- This technique re-scales a feature or observation value with distribution value between 0 and 1.

  MinMaxScaler is used for noramalization.

  MinMaxScaler
  - Scales to range [0, 1], when all the data features are positive.
  - Scales to range [-1, 1] if there are negative values in the dataset.

- This scaling compresses all the inliers in the narrow range [0, 0.005].



Standardization:

- It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.
- StandardScaler is used for standardization.
    - StandardScaler follows Standard Normal Distribution (SND).
    - Therefore, it makes mean = 0 and scales the data to unit variance.
- In the presence of outliers, StandardScaler does not guarantee balanced feature scales, due to the influence of the outliers while computing the empirical mean and standard deviation.
- This leads to the shrinkage in the range of the feature values.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) 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 corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

QQPlot:

A Q-Q plot, short for "quantile-quantile" plot, is a graphical technique for determining if two data sets come from populations with a common distribution. In most cases, this type of plot is used to determine whether or not a set of data follows a normal distribution.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of

a theoretical distribution. A quantile-quantile (Q-Q) plot, shows the distribution of the data against the expected normal distribution.

Purpose/Use: Check If Two Data Sets Can Be Fit With the Same Distribution

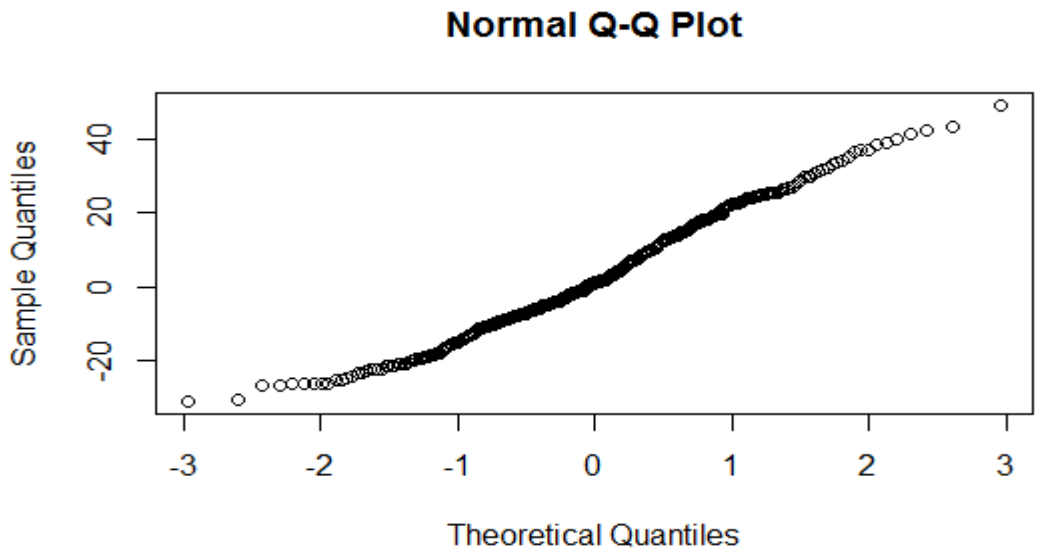
How it is used to check the fit of distribution?

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The q-q plot is formed by:
Vertical axis: Estimated quantiles from data set 1
Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.



If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

The advantages of the q-q plot:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested.
- Example 1: Shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
- Example 2: If the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

Importance of QQPlot: Check for Common Distribution

While checking the common distribution using QQPlot, answers for the following questions shows the result of the distribution.

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale.

If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.