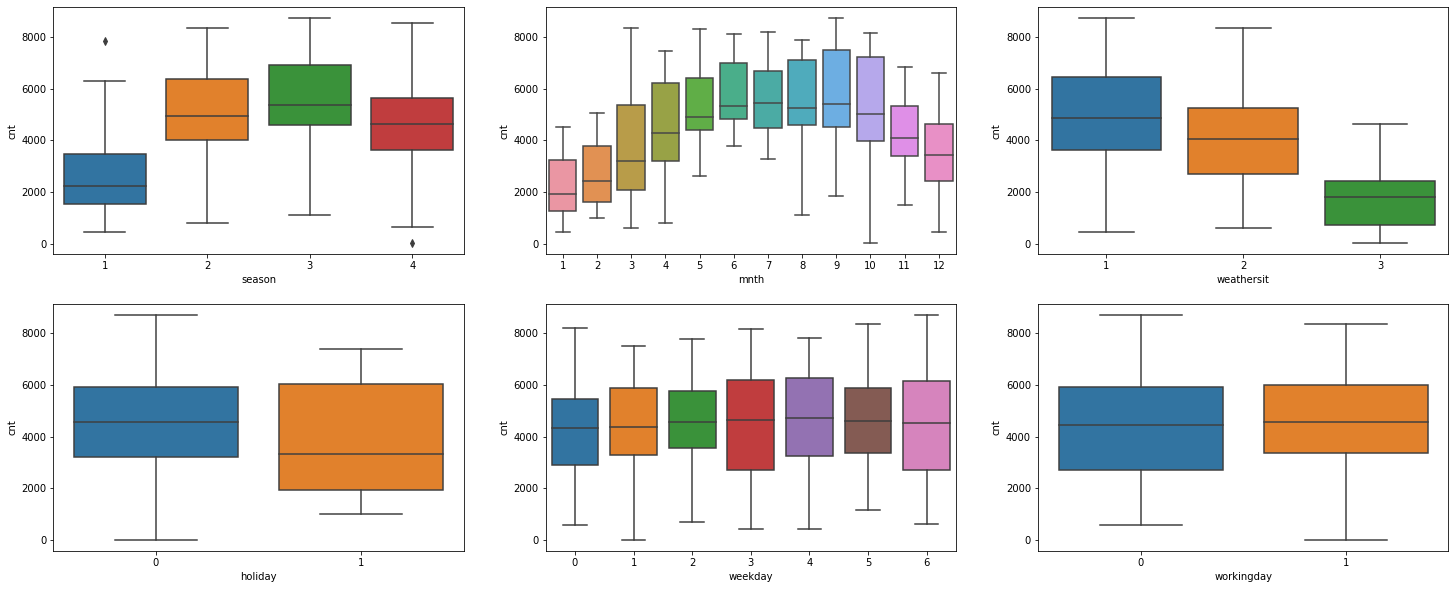
# Assignment-based Subjective Questions

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



There are 6 categorical variables in the dataset. With Box plot we study their effect on the dependent variable (‘**cnt’**).

The inference one can derive is:

- **season**: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

- **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

- **weathersit**: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings and can be a good predictor for the dependent variable.

- **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

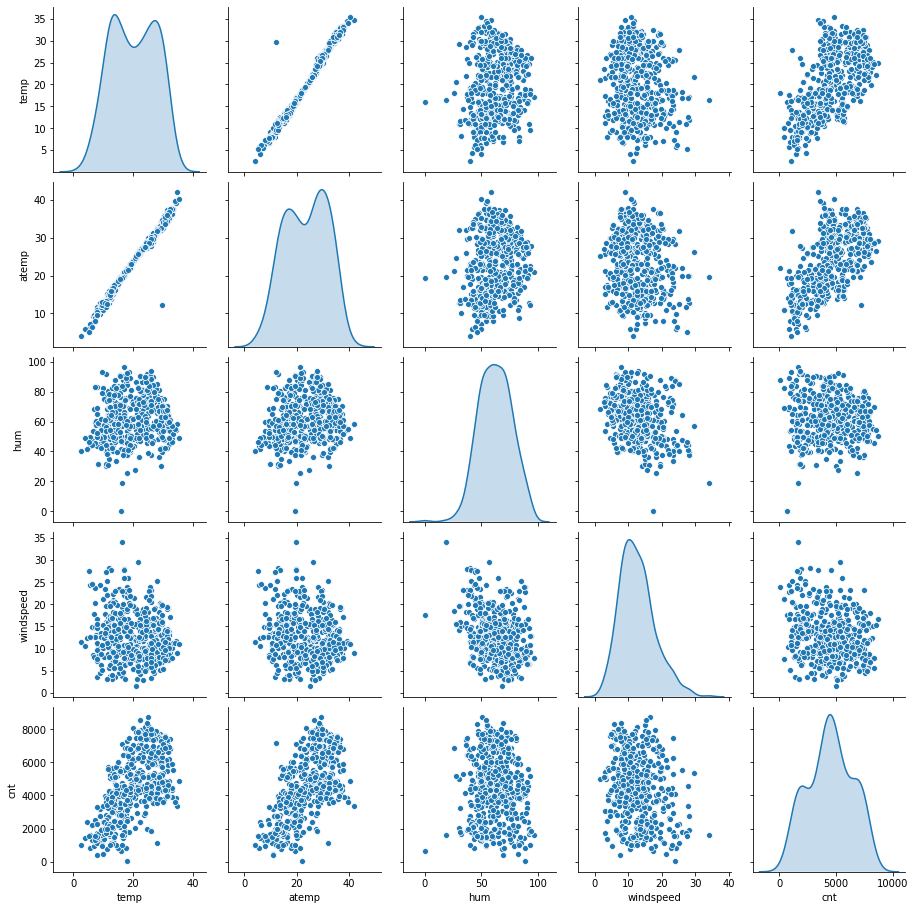
- **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

- **workingday**: Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable. ---

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

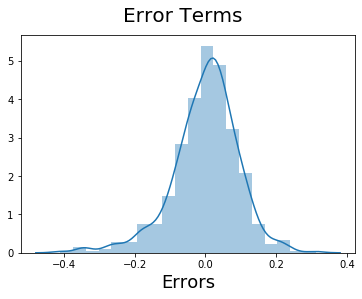
**drop\_first**=**True** is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among **dummy variables**

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



The target variable ‘**cnt**’ seems to be correlated to the numerical variable ‘**temperature**’ - the most.

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



We validated the assumptions of Linear Regression by doing **a Residual Analysis of the training data**. A histogram of the error terms is plotted and one can see the Error terms are normally distributed with mean zero which means the assumptions of Linear Regression are valid.

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

As per our final Model, the top 3 predictor variables that influences the bike booking are: ---

cnt=0.1312 + ( yr × 0.2325 ) - ( holiday × 0.0971 ) + ( temp × 0.5174 ) − ( windspeed × 0.1497 ) + ( season2 × 0.1000 ) + ( season4 × 0.1383 ) + ( mnth8 × 0.0542 ) + ( mnth9 × 0.1162 ) − ( weathersit2 × 0.0825 ) − ( weathersit3 × 0.2828 )

- **Temperature (temp)** - A coefficient value of ‘0.5174’ indicating that a unit increase in temp variable increases the bike hire numbers by 0.5174 units.

- **Weather Situation 3 (weathersit\_3)** - A coefficient value of ‘-0.2828’ indicates that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2828 units.

- **Year** A coefficient value of ‘0.2325’ indicates that a unit increase in yr variable increases the bike hire numbers by 0.2325 units.

--- **SO IT IS RECOMMENDED TO GIVE THESE VARIABLES UTMOST IMPORTANCE WHILE PLANNING, TO ACHIEVE MAXIMUM BOOKING**.

General Subjective Questions

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

Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things:

1. does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are

(1) determining the strength of predictors,

(2) forecasting an effect, and

(3) trend forecasting.

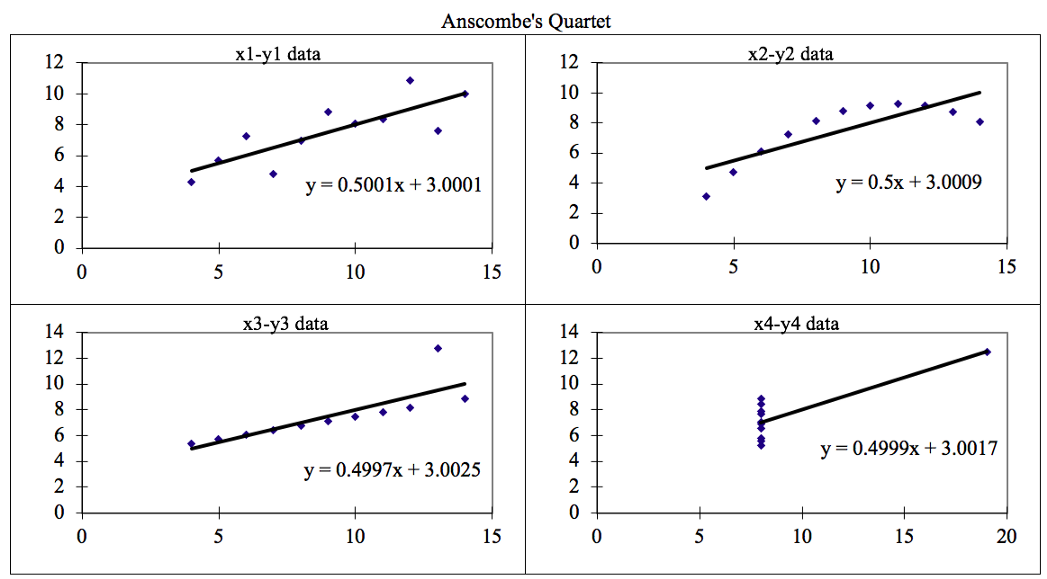
First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable.  Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes.

That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.  A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values.  The regression analysis can be used to get point estimates.  A typical example being, “what will the price of gold be in 6 months?

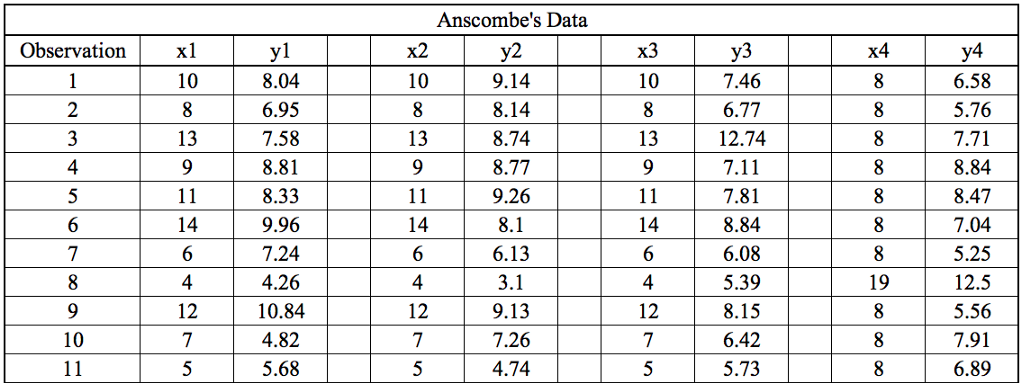
1. **Explain the Anscombe’s quartet in detail. (3 marks)**



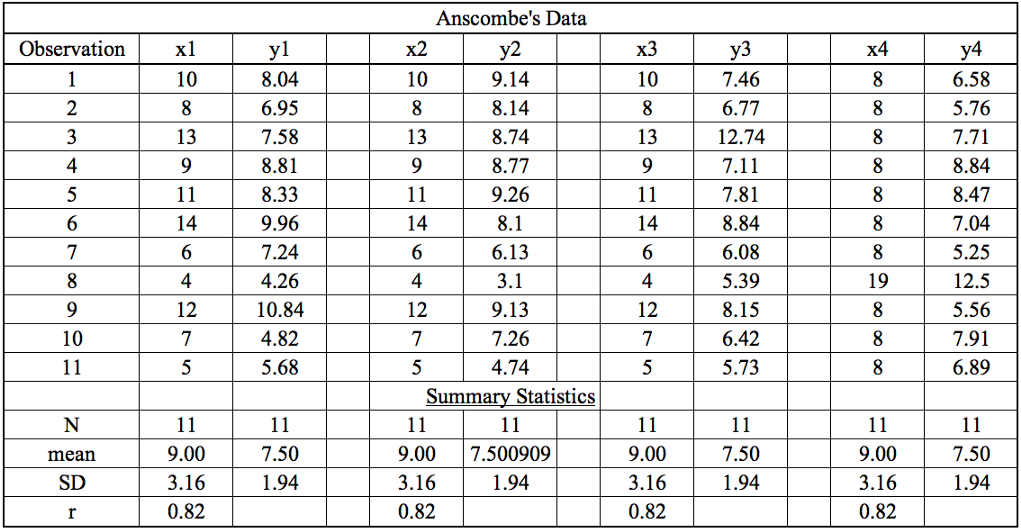
**Anscombe’s Quartet**can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model**if built. They have very different distributions and **appear differently**when plotted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

There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean**of all x,y points in all four datasets.

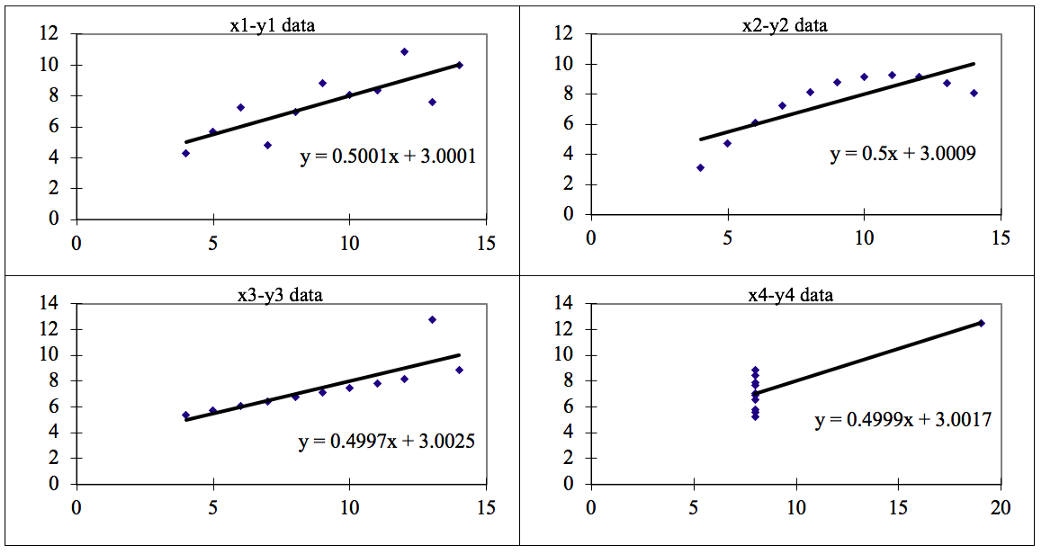
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships**and is incapable of handling any other kind of datasets. These four plots can be defined as follows:



The statistical information for all these four datasets are approximately similar and can be computed as follows:



When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

**Dataset 1:**this **fits**the linear regression model pretty well.

**Dataset 2:** this **could not fit**linear regression model on the data quite well as the data is non-linear.

**Dataset 3:**shows the **outliers**involved in the dataset which **cannot be handled**by linear regression model

**Dataset 4:**shows the **outliers**involved in the dataset which **cannot be handled**by linear regression model

Conclusion:

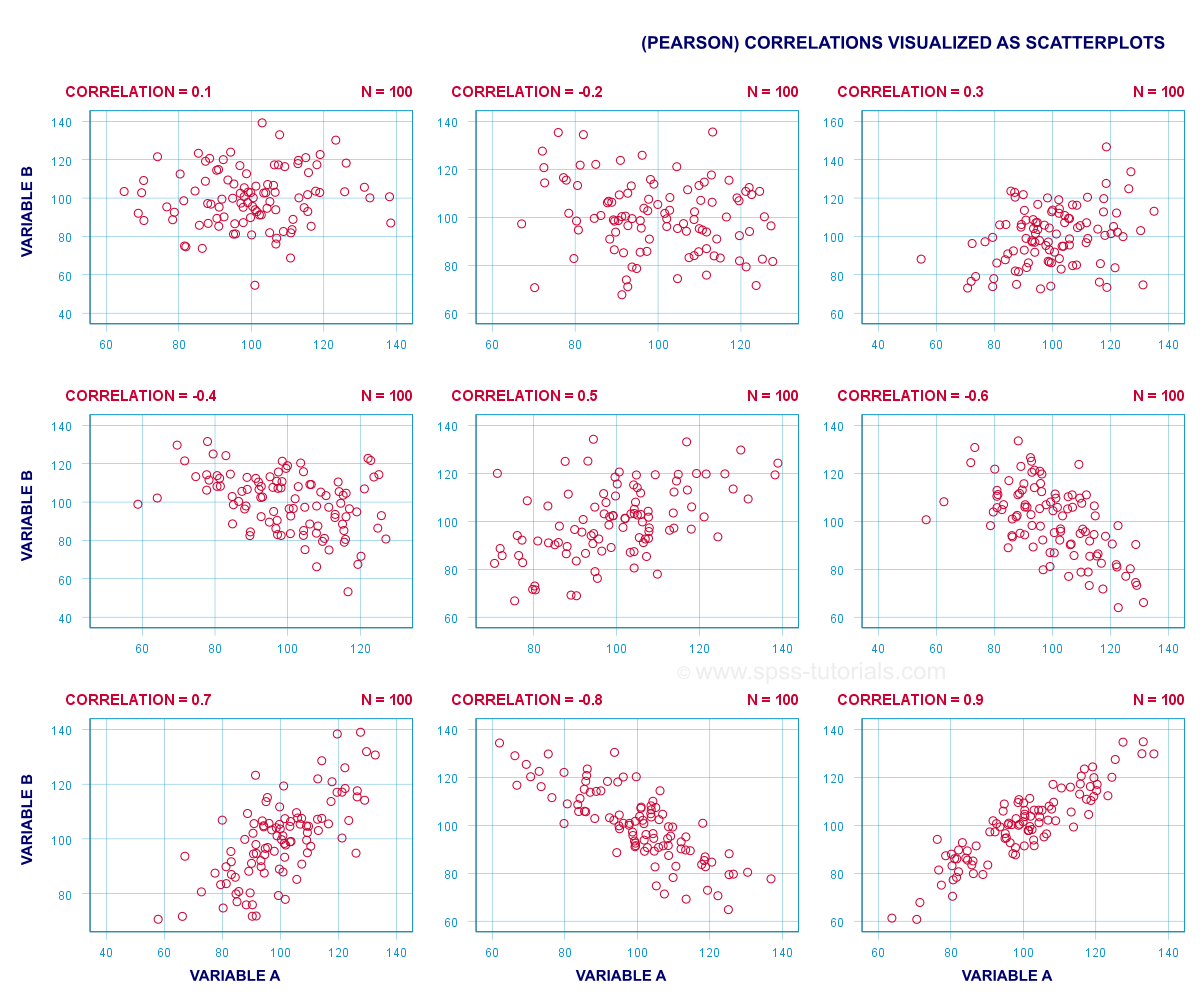
We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

1. **What is Pearson’s R? (3 marks)**

**Correlation(Pearson’s R)** is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. **Pearson's correlation** coefficient (**r**) is a measure of the strength of the association between the two variables.

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

A correlation coefficient indicates the extent to which dots in a scatterplot lie on a straight line. This implies that we can usually estimate correlations pretty accurately from nothing more than scatterplots. The figure below nicely illustrates this point.



## **Correlation Coefficient - Basics**

Some basic points regarding correlation coefficients are nicely illustrated by the previous figure. The least you should know is that

* **Correlations are never lower than -1.** A correlation of -1 indicates that the data points in a scatter plot lie exactly on a straight descending line; the two variables are perfectly negatively linearly related.
* A **correlation of 0** means that two variables don't have any linear relation whatsoever. However, some non linear relation may exist between the two variables.
* **Correlation coefficients are never higher than 1.** A correlation coefficient of 1 means that two variables are perfectly positively linearly related; the dots in a scatter plot lie exactly on a straight ascending line.

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

What?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.*…*

The terms **normalization** and **standardization** are sometimes used interchangeably for scaling but they usually refer to different things. **Normalization** usually means to scale a variable to have values **between** 0 and 1, while **standardization** transforms data to have a mean of zero and a standard deviation of 1.

Normalisation typically allows us to transform the data with varying scales so that no specific dimension will dominate the statistics, and it does not require making a very strong assumption about the distribution of the data. However, [Normalisation](https://www.codecademy.com/articles/normalization) does not treat outliners very well. On the contrary, standardisation allows users to better handle the outliers and facilitate convergence for some computational algorithms like gradient descent. Therefore, we usually prefer standardisation over Normalisation.

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

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

).  If there is perfect correlation, then **VIF** = **infinity**. A large value of **VIF** indicates that there is a correlation between the variables.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

* 1. come from populations with a common distribution
  2. have common location and scale
  3. have similar distributional shapes
  4. have similar tail behavior

Interpretation:A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.