

## 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?**

- A) The season box plot indicates that more bikes are rent during fall season.
- B) In months from May-sept total renting is higher.
- C) when weather is clear total bookings is higher.
- D) On non holiday total bookings is higher.
- E) in 2019 total bookings is higher than 2018.
- F) The working day and holiday plot indicates that more bikes are rent during normal working days than on weekends or holidays.

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

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. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is furnished and semi\_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

The 'temp' variable 'registered' has the highest correlation with the target variable 'cnt' .

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

From R-Squared and adjusted R-Squared value of both train and test dataset we could conclude that the above variables can well explain more than 78% of bike demand. Coefficients of the variables explains the factors effecting the bike demand.

**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 contributing significantly towards the demands of share bikes are: weathersit\_Light\_Snow(negative correlation) and yr\_2019/yr\_1(Positive correlation) and temp(Positive correlation).

## General Subjective Questions

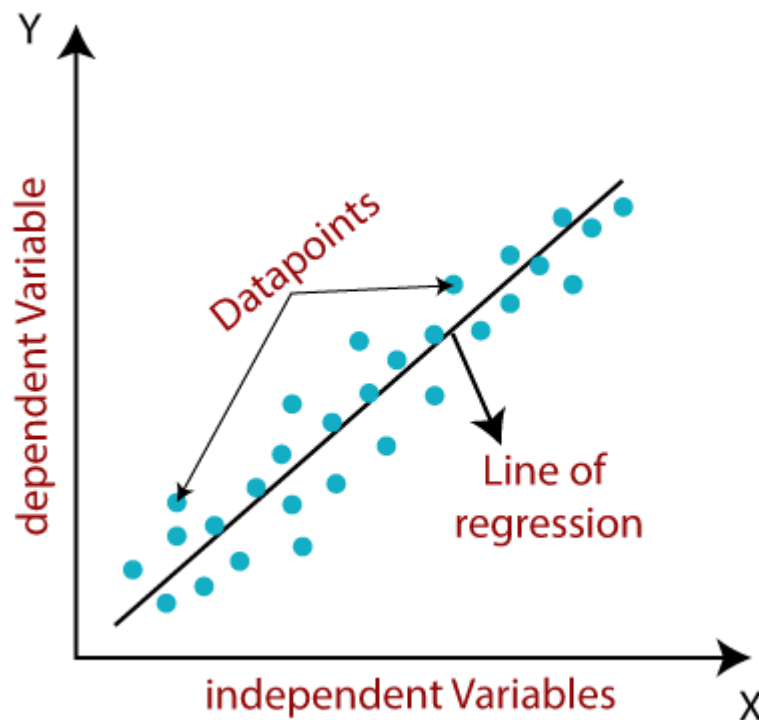
**1. Explain the linear regression algorithm in detail.**

Linear regression is one of the basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables.

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. It is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and yaxis should be linearly correlated.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Linear regression equation can be written as:  $y=mx+c$  ,

$m$  = Slope of the line,  $c$  =  $y$ -intercept of the line,  $x$  = Independent variable from dataset,  $y$  = Dependent variable from dataset.

Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

## 2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different 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 analysing it and the effect of outliers on statistical properties , diversity of the data, linear separability of the data, etc.

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:

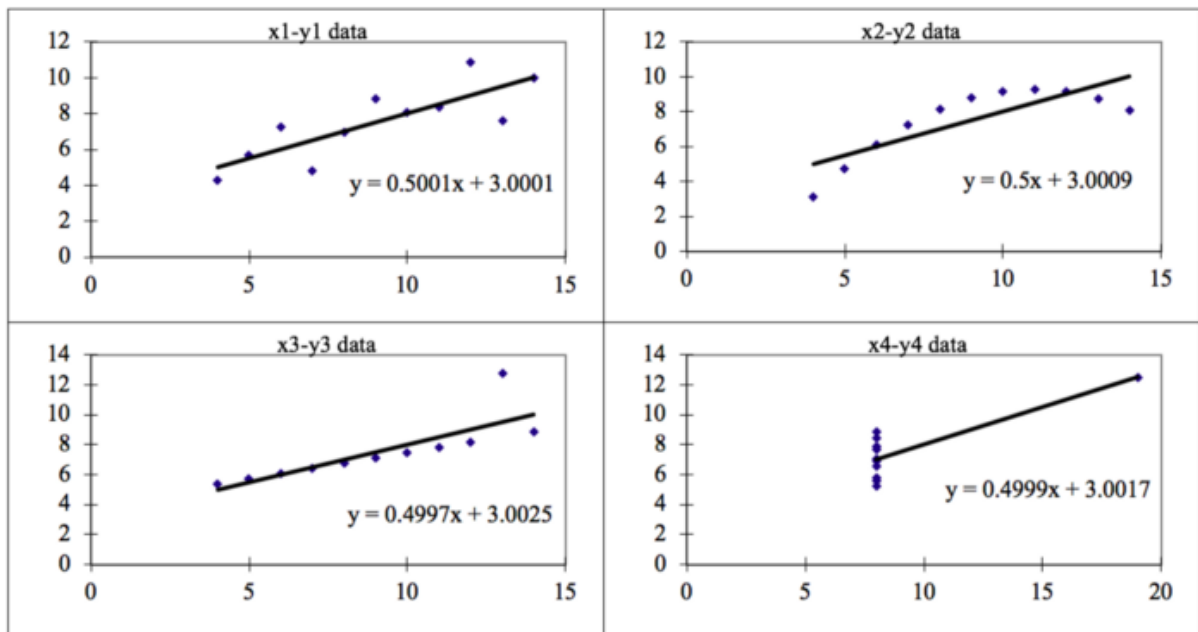


Image by Author

The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

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.

### 3. What is Pearson's R?

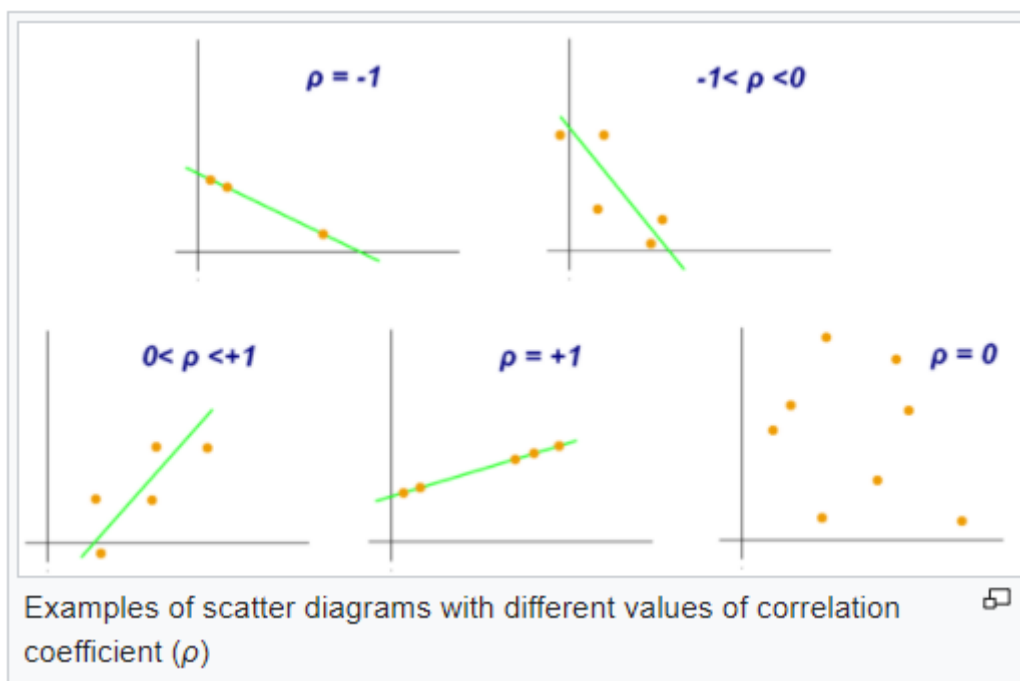
In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's  $r$ , the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. Pearson's correlation (also called Pearson's  $R$ ) is a correlation coefficient commonly used in linear regression.

It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between  $-1$  and  $1$ .

The Pearson's correlation coefficient varies between  $-1$  and  $+1$

where:

- $r = 1$  means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)
- $r = -1$  means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)
- $r = 0$  means there is no linear association
- $r > 0 < .5$  means there is a weak association
- $r > .5 < .8$  means there is a moderate association
- $r > .8$  means there is a strong association



Using the formula proposed by Karl Pearson, we can calculate a **linear relationship** between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient  $r$ . There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

The formula given is:

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

Where,

**N** = the number of pairs of scores

**$\sum xy$**  = the sum of the products of paired scores

**$\sum x$**  = the sum of x scores

**$\sum y$**  = the sum of y scores

**$\sum x^2$**  = the sum of squared x scores

**$\sum y^2$**  = the sum of squared y scores

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

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. 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.

**Normalization or Min-Max Scaling** is used to transform features to be on a similar scale. The new point is calculated as:

$$X_{\text{new}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

This scales the range to [0, 1] or sometimes [-1, 1]

Normalization is useful when there are no outliers as it cannot cope up with them.

**Standardization or Z-Score Normalization** is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X_{\text{new}} = (X - \text{mean}) / \text{Std}$$

Standardization does not get affected by outliers because there is no predefined range of transformed features.

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

If there is perfect correlation, then  $VIF = \infty$ . This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get  $R^2 = 1$ , which lead to  $1/(1 - R^2)$  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.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

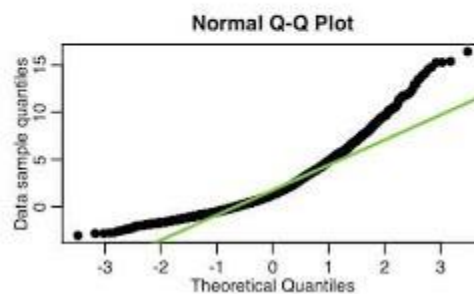
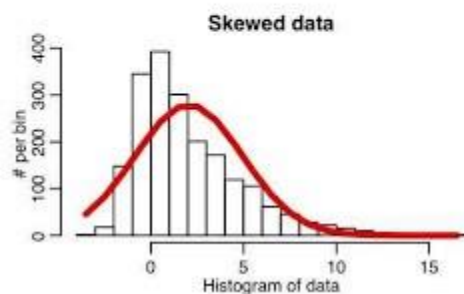
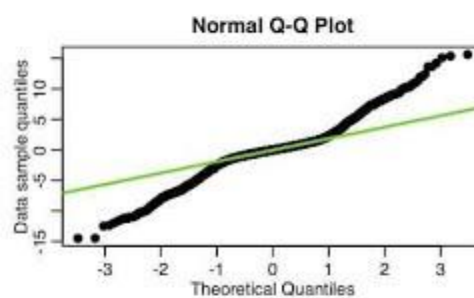
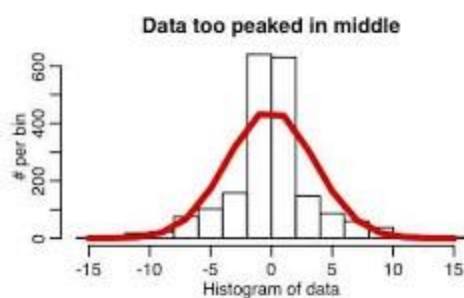
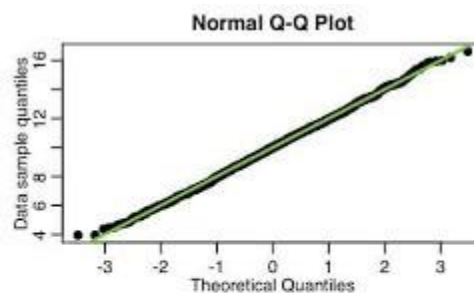
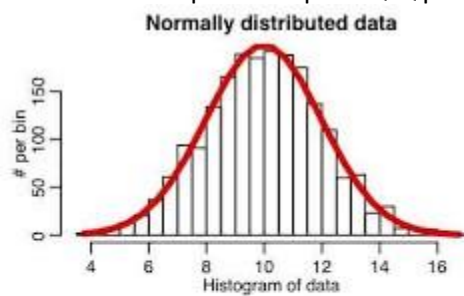
The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. A Q Q plot showing the 45 degree reference line: If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line  $y = x$ . If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line  $y = x$ .

Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

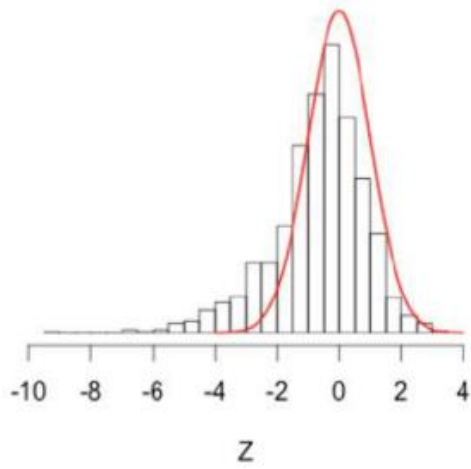
The advantages of the q-q plot are:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, 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.

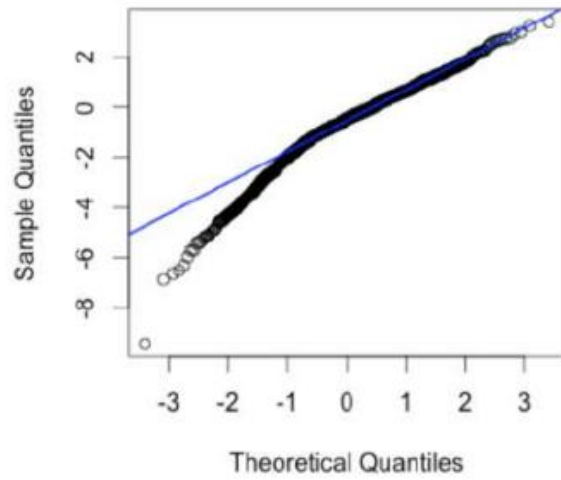
If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.



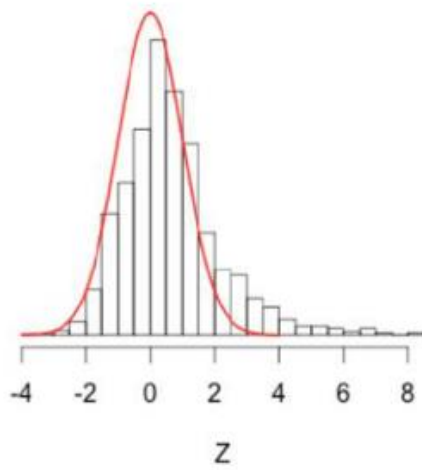
**Skewed Left**



**Normal Q-Q Plot**



**Skewed Right**



**Normal Q-Q Plot**

