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?

   Ans: The following inferences can be made about effect of categorical variables on cnt(dependent variable) which is the number of bikes rented on a day, from the data set

   1. Maximum number of bikes are rented during fall and summer.
   2. The number of people using shared bikes have increased from year 2018 to 2019 significantly.
   3. More number of bikes are rented towards the last working days of a week, i.e Friday and Saturday, as compared to other days.
   4. Months 6, 7, 8 and 9 see a maximum usage of rented bikes.
   5. Usage of rented bikes service is less on a holiday as compared to other days.
   6. Similar to above usage of rented bikes is more on a working day than on a non working day.
   7. More bikes are rented under lighter weather conditions like clear skies, partly cloudy or lightly misty, than on more severe conditions like rainy or snowy or during thunder storms.

---

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

   Ans: It is important to use drop_first=True because in order to represent a categorical variable with degree n we need only n-1 columns. Here degree is the number of distinct values a categorical variable can take. To illustrate this let us take an example.

   Say season is the categorical variable in our case with four values namely spring, summer, fall, winter.

   No let us say we create dummy variables for season, which will split it into following,

spring	summer	fall	winter	
1	0	0	0	→ spring
0	1	0	0	→ summer
0	0	1	0	→ fall
0	0	0	1	→ winter

   In order to represent winter we can take spring, summer and fall values to be zero rather than creating a separate column for winter. Which then can be represented as

spring	summer	fall	
1	0	0	→ spring
0	1	0	→ summer
0	0	1	→ fall
0	0	0	→ winter

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

   Ans: From the pair plot we see that registered variable has the highest correlation to target variable, but this is because registered and casual add up to cnt which is the target variable and hence the correlation.

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

   Ans:
   a. First we plot a distplot of the error terms i.e y_train – y_train_pred
      This plot should be a normal distribution with mean centred at zero.
   b. Next we can take any one X variable say 'windspeed' in our case and plot error terms in y axis and 'windspeed' in X axis. If the error terms are randomly distributed then it means there is no relationship between the error terms.
   c. Then we plot y_test vs y_test_pred and see if the variance in error terms is constant from the plot.

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

   Ans: by looking at the absolute values of coefficients we see that the top three features contributing significantly are
   a. cas_to_registered, which the ratio of the number of casual to registered users, as this increases the demand for bike sharing increases.
   b. Next is a derived dummy variable spring which has negative coefficient which is surprising. It might mean that holidays of school kids fall during spring break and hence people move to their hometowns or spend time with their family which leads to less usage of bike service.
   c. Then we have yr that is year, more number of people are using bike sharing service from 2018 to 2019, that is it is gaining popularity over the years.

General Subjective Questions

6. Explain the linear regression algorithm in detail.

Ans: **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

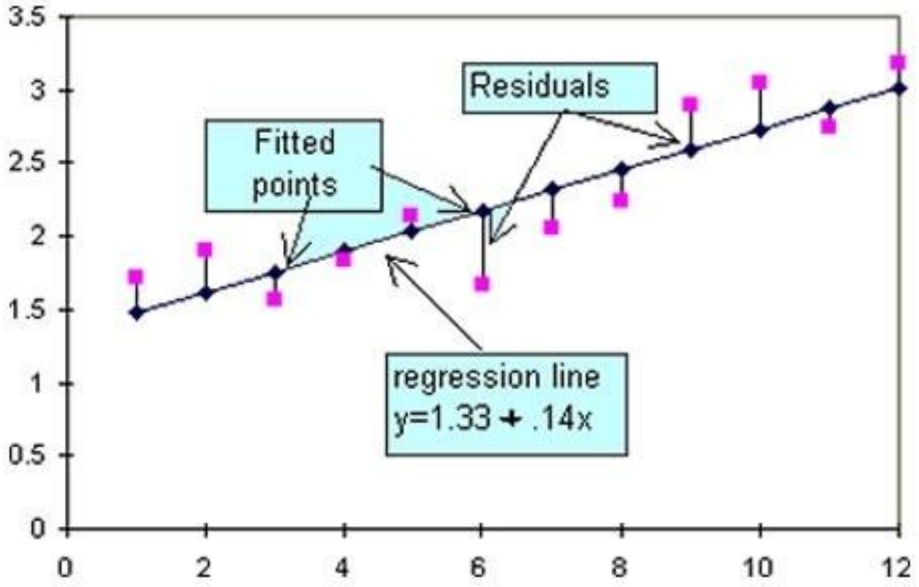
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. The regression line is the best fit line for our model.

Hypothesis function for Simple Linear Regression:

$$y = \theta_1 + \theta_2.x$$

While training the model we are given:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x

Let us suppose we have following plot and the LR algorithm fits a best fit line as shown below

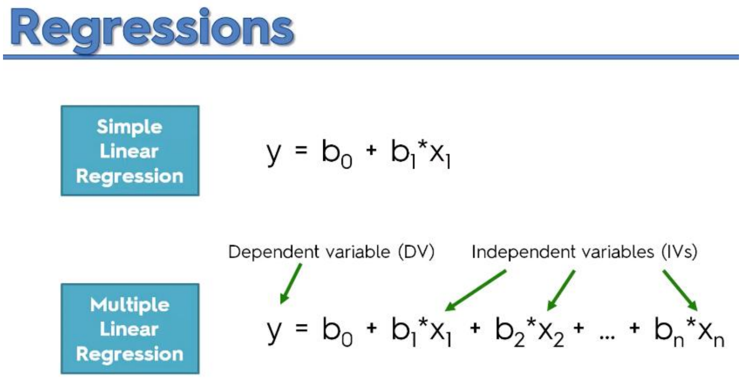


We see that pink dots are actual point and black ones are predicted points which fall on the line. As we can see from the plot there is some difference between actual points and predicted points which is called the error. LR algorithm aims to minimize the error by minimizing the sum the of the squares of the error terms, which then will give us the best fit line. The sum of the squares of the error terms is called residual sum of squares.

$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Where predi is the predicted value and yi is the actual value of dependent variable, which depends on independent variable x. The algorithm begins with a assumed value of slope and intercept and using **gradient descent** it finds out the optimal value of slope and intercept.

This can be extended to multiple independent variables and single dependent variable as well using the similar hypothesis by considering the best fit line to be,



Regressions

Simple Linear Regression

$y = b_0 + b_1 * x_1$

Dependent variable (DV)

Independent variables (IVs)

Multiple Linear Regression

$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$

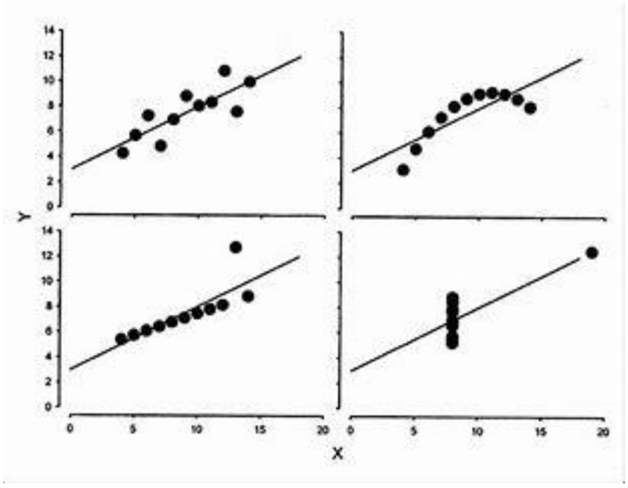
While applying LR there are some assumptions made about the data points which are as follows:

- Linear relationship. One of the most important assumptions is that a linear relationship is said to exist between the dependent and the independent variables.
- No auto-correlation or independence. The residuals (error terms) are independent of each other. In other words, there is no correlation between the consecutive error terms.
- No Multicollinearity. The independent variables shouldn't be correlated. If multicollinearity exists between the independent variables, it is challenging to predict the outcome of the model.
- Homoscedasticity. Homoscedasticity means the residuals have constant variance at every level of x. The absence of this phenomenon is known as heteroscedasticity.
- Normal distribution of error terms. The last assumption that needs to be checked for linear regression is the error terms' normal distribution.

  If any of these assumptions are later found to be invalid then LR is not the best model for the underlying prediction.

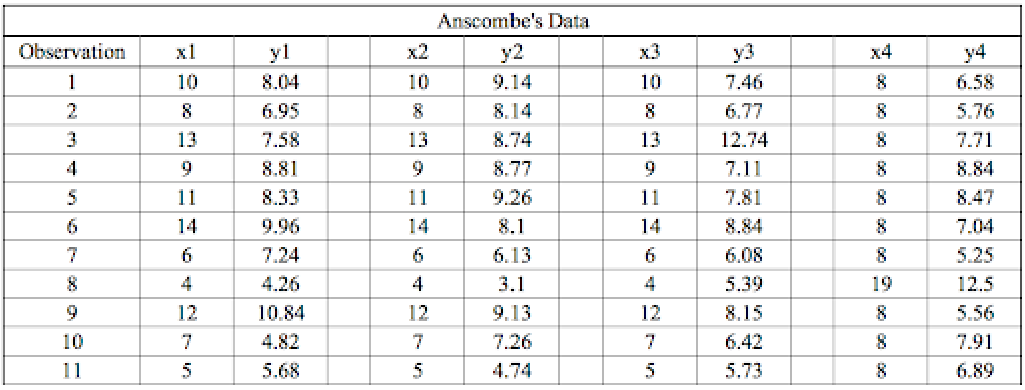
7. Explain the Anscombe's quartet in detail.

   Ans: Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.



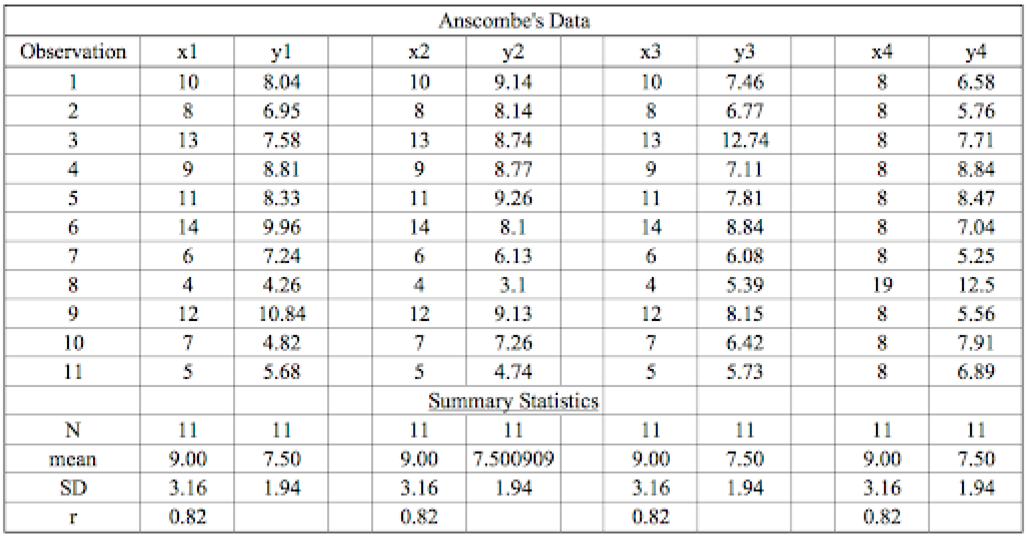
Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

We can define these four plots as follows:



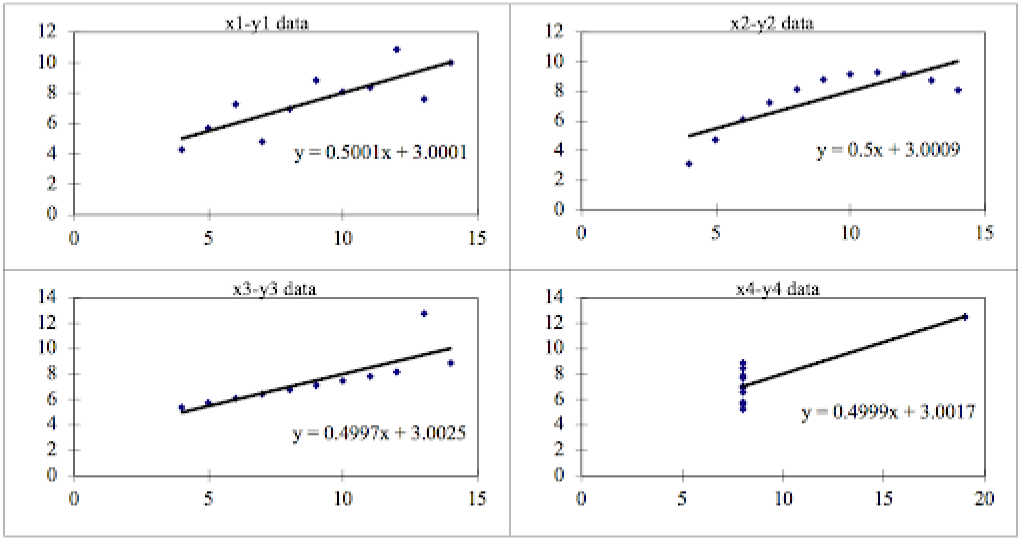
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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

The statistical information for these four data sets are approximately similar. We can compute them as follows:



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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

ANSCOMBE'S QUARTET FOUR DATASETS

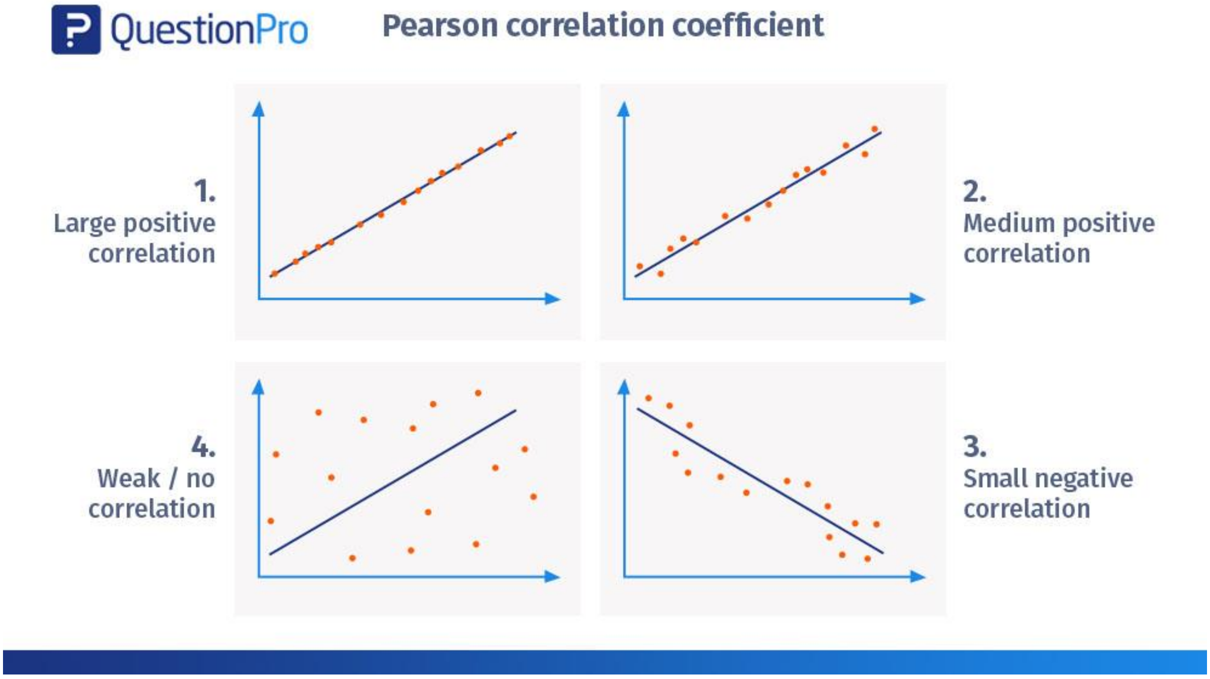
- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

8. What is Pearson's R?

Ans: The Pearson correlation coefficient is a statistic measure, that summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Another way to think of the Pearson correlation coefficient (*r*) is as a measure of how close the observations are to a line of best fit. 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. When *r* is 1 or –1, all the points fall exactly on the line of best fit:



Below is a formula for calculating the Pearson correlation coefficient (*r*):

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

x and y are the data points and n is the number of data points.

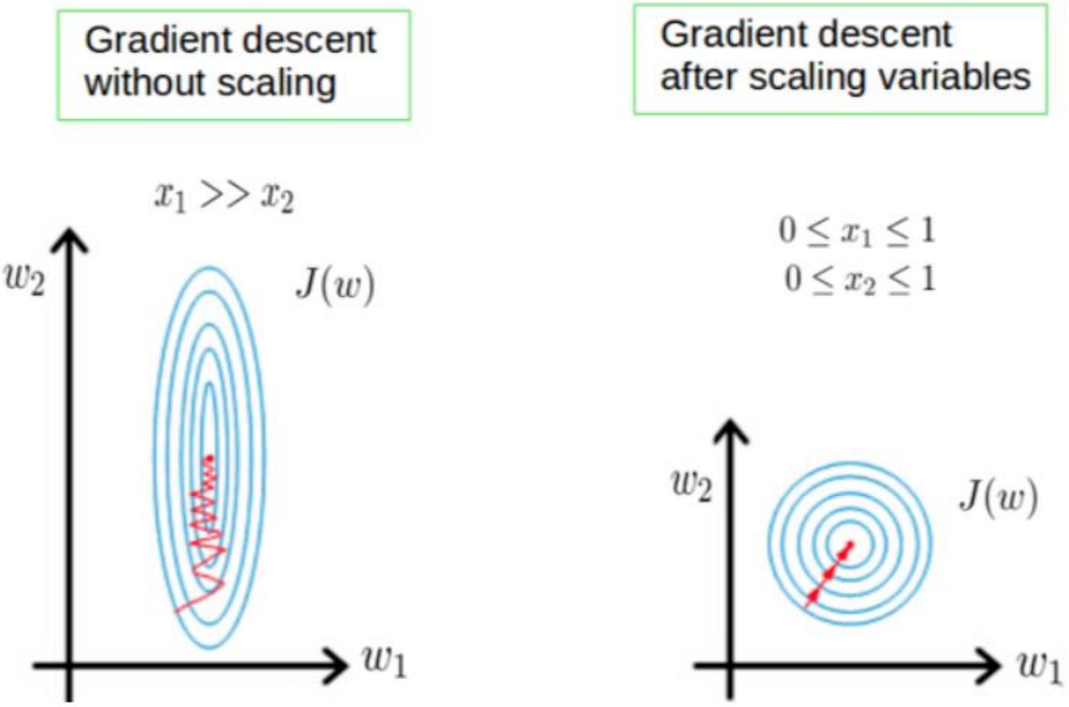
9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:

What is scaling?

When your data is comprised of attributes with varying scales, many machine learning algorithms can benefit from rescaling the attributes to all have the same scale. Often this is referred to as normalization and attributes are often rescaled into the range between 0 and 1.

It is a technique to standardise the independent variables present to a fixed range in order to bring all values to same magnitudes. Generally performed during the data pre-processing step and also helps in speeding up the calculations in an algorithm. **Used in Linear Regression, K-means, KNN, PCA, Gradient Descent etc.**



Why scaling?

Feature Scaling should be performed on independent variables that vary in magnitudes, units, and range to standardise to a fixed range.
If no scaling, then a machine learning algorithm assign higher weight to greater values regardless of the unit of the values. As the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. **For example:** ML considers the value 1000 gram > 2 kilogram or the value 3000 meter greater than 5 km and hence the algorithm will give wrong predictions.

Many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature.
Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it. **In short we scale down to same scale.**

Algorithm which is NOT distance based are not affected by feature scaling. eg. Naive Bayes

Mainly there are two types of feature scaling :-

Min-Max Scaling (Scaling): It differs from normalisation in the sense that here sole motive to change range of data whereas as in Normalization/standardization , the sole motive is to normalise the distribution shape curve and to make it perfect Gaussian curve. the data is scaled to a fixed range — usually 0 to 1. The cost of having this bounded range in contrast to standardization is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization (Z-score normalization): transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1. **μ=0 and σ=1. Mainly used in KNN and K-means.** where μ is the mean (average) and σ is the standard deviation from the mean; standard scores (also called ***z*** scores) of the samples are calculated as follows:

$$x' = \frac{x - x_{mean}}{\sigma}$$

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

Ans: 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.

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

Ans: 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. In Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyse and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

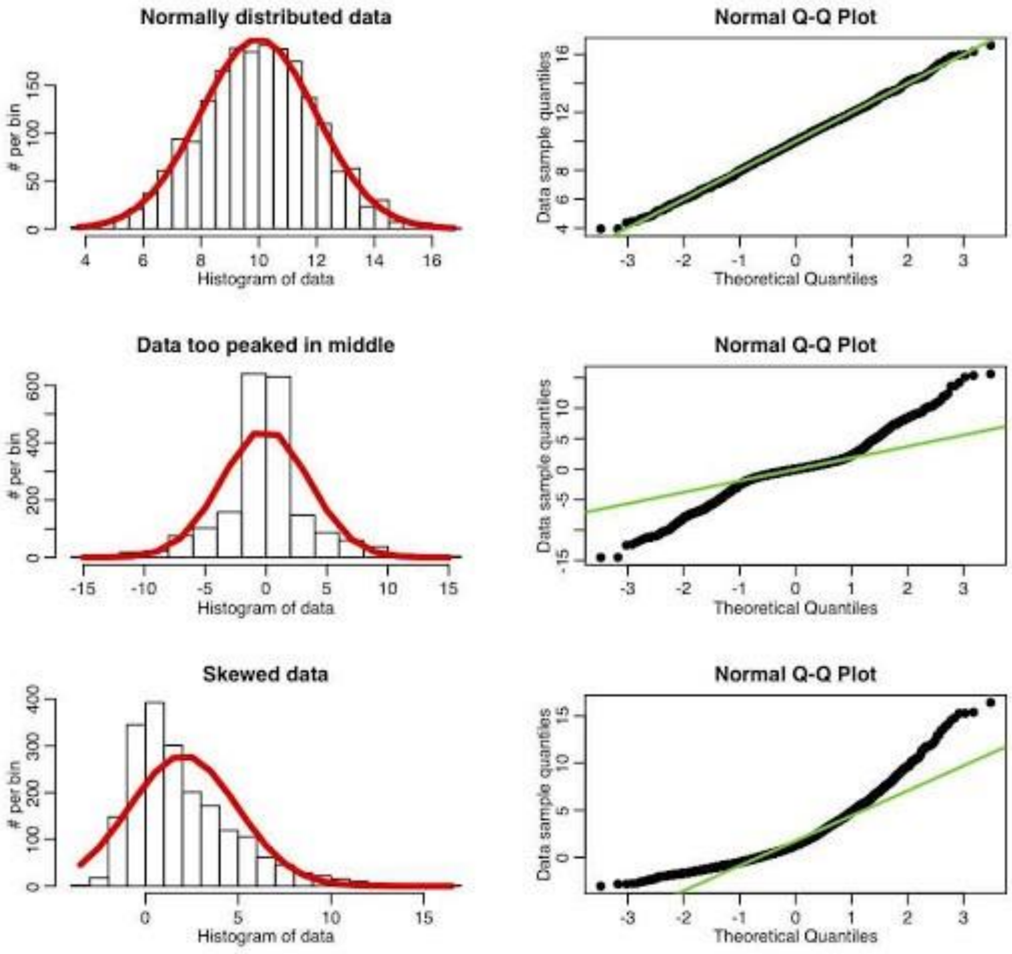
Importance:

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. In general, we are talking about **Normal distributions** only because we have a very beautiful concept of 68–95–99.7 rule which perfectly fits into the normal distribution, So we know how much of the data lies in the range of first standard deviation, second standard deviation and third standard deviation from the mean. knowing if a distribution is Normal opens up new doors for us to experiment with the data easily. Secondly, Normal Distributions occur very frequently in most of the natural events which have a vast scope.

We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1)on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph.

Now we have to focus on 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 indeed are scattered significantly from the positions then we cannot conclude a relationship between the x and y axes which clearly signifies that our ordered values which we wanted to calculate are 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 which is the simple concept of Q-Q plot.



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 —

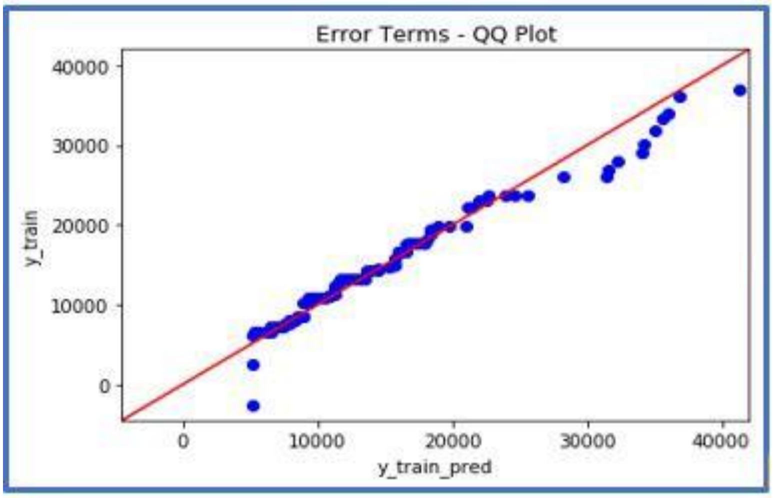
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour

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

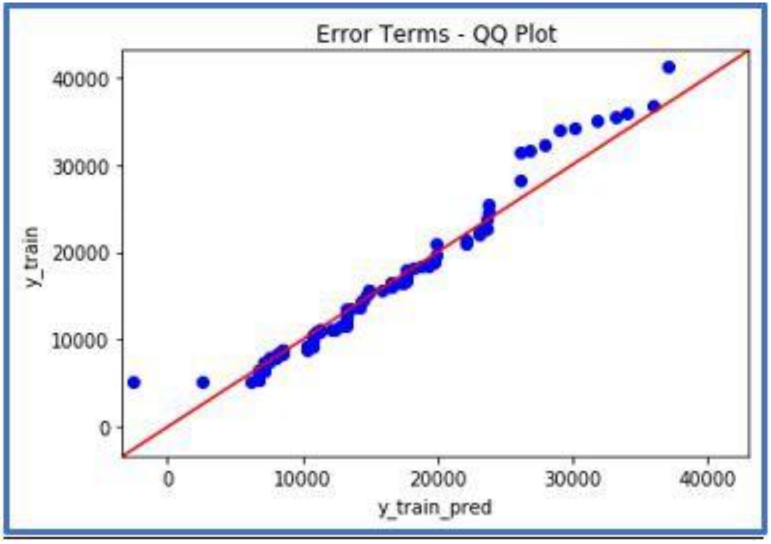
Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis.

References for subjective questions:

www.geeksforgeeks.org

Anscombe’s Quartet: What Is It and Why Do We Need It? | Built In

Pearson Correlation Coefficient (r) | Guide & Examples (scribbr.com)

Feature Scaling :- Normalization, Standardization and Scaling ! | by Nishant Kumar | Analytics Vidhya | Medium

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

Q-Q plot in linear regression-Explained | by PremalMatalia | Medium

THANK YOU