**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 : When we look at the categorical variables in the US bike sharing rental dataset, we can draw some interesting conclusions about how they affect the number of bike rentals. Check out these possible insights:

**Season:** It seems like the season has a big impact on bike rentals. For example, more people might rent bikes during the spring and summer months when the weather is nicer, compared to the fall and winter.

**Weather Situation:** The weather conditions, like whether it's clear, cloudy, rainy, or snowy, can also affect bike rentals. If the weather is bad, people might be less likely to rent bikes, which leads to changes in rental numbers depending on the weather situation.

**Holiday:** Holidays can have different effects on bike rentals. On one hand, rentals might go up during holidays when people have more free time. On the other hand, they could go down during major holidays when people are traveling or doing indoor activities.

**Weekday:** Whether it's a weekday or a weekend can also make a difference in bike rentals. Weekdays might see more people using bikes for commuting, while weekends might have more people using them for fun. This leads to variations in rental numbers depending on the day of the week.

**Workingday:** Similarly, whether it's a working day or a non-working day can impact bike rentals. Working days might have more rentals from commuters, while non-working days might see more people using bikes for leisure.

**Month:** Different months of the year can show different patterns in bike rentals. For example, rentals might go up during the warmer months and go down during the colder ones.

**Hour:** The time of day also plays a role in bike rentals. Peak hours, like morning and evening commute times, might see more rentals, while late-night hours might have fewer.

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

Ans : I think it depends on the model. If you don't drop the first column then your dummy variables will be correlated (redundant as Dimitre shows in the post below). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importances may be distorted.

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

Ans : Temp(0.63), atemp(0.63) and year(0.57) are highly correlated with CNT

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

Ans:

R square, adjusted r2 square , F-statistic:, and prob F statistic value using these parameter will validate.

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

Ans: As per our final Model, the top 3 predictor variables that influences the bike booking are: --- - Monthly_Sep - A coefficient value of ‘0.0950’ indicated that a unit increase in Monthly_Sep variable increases the bike hire numbers by 0.0950 units. - weathersit_Light_Snow_Rain (Light snow Rain) - A coefficient value of ‘-0.3004’ indicated that, w.r.t Weathersit, a unit increase in weathersit_Light_Snow_Rain variable decreases the bike hire numbers by 0.3004 units. - Year (yr) - A coefficient value of ‘0.2445’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2445 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

Ans:

**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression is a method of modelling a target value based on independent predictors. Regression mainly used in Forecasting and finding out the cause and effect relationship between variables.

Linear regression differed based on the number of independent variables and type of relationship between the independent and dependant variables.

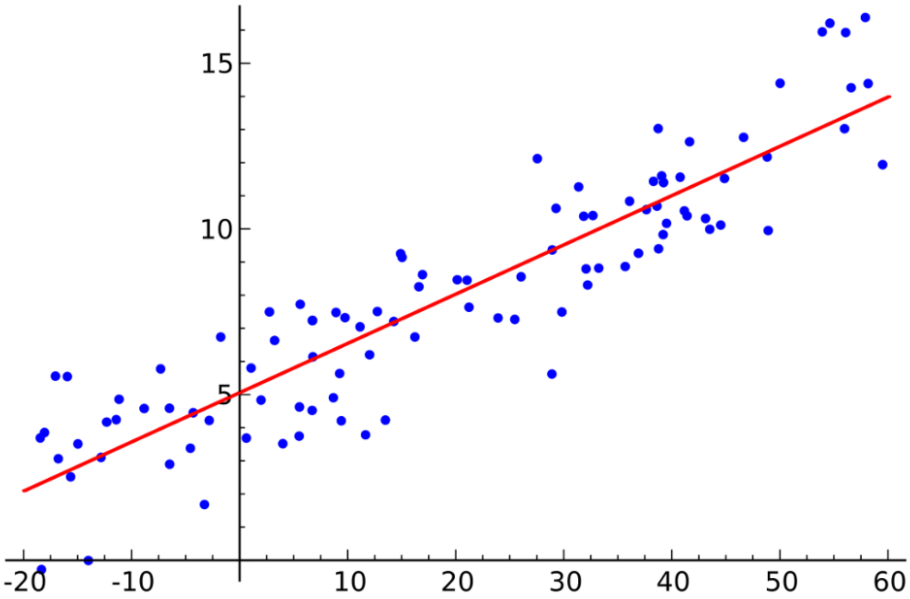
Linear Regression Equation:

$$y = \beta 0 + \beta 1 * x$$

Where

β0 is the intercept of the fitted line

β1 is the coefficient for the independent variable x.



Linear Regression

Linear Regression mainly two types:

**Simple linear regression:**

Simple linear regression uses traditional slope-intercept form, where **β0** and **β1** are the variables our algorithm will try to “learn” to produce the most accurate predictions. **‘x’** represents our input data and **‘y’** represents our prediction.

$$y = \beta 0 + \beta 1 * x + \varepsilon i$$

x : independent variable or predictor variable

y: dependent variable or response variable

β1 is the slope of the line:

This is one of the most important quantities in any linear regression analysis. A value very close to 0 indicates little to no relationship; large positive or negative values indicate large positive or negative relationships, respectively.

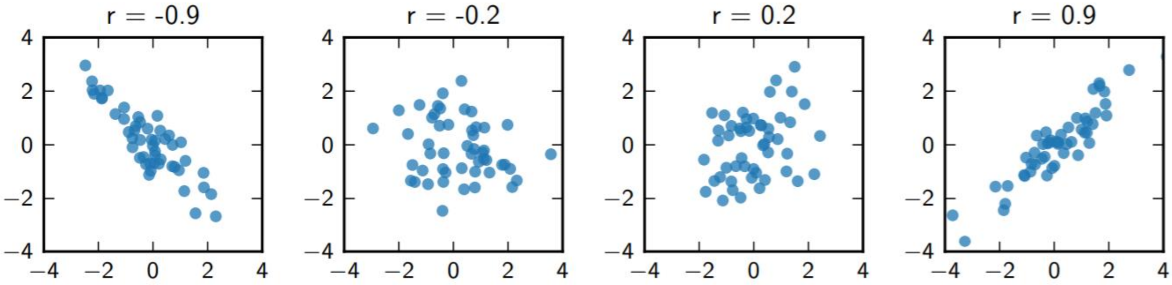
β0 is the intercept of the line

εi is the error term (Gaussian noise)

**correlation coefficient:**

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where ¯x, ¯y, Sx and Sy are the sample means and standard deviations for x values and y values, respectively, and r is the correlation coefficient.



Each plot shows data with a particular correlation coefficient r. Values farther than 0 (outside) indicate a stronger relationship than values closer to 0 (inside). Negative values (left) indicate an inverse relationship, while positive values (right) indicate a direct relationship.

The square of the correlation coefficient r 2 will always be positive and is called the **coefficient of determination.**

**Interpolation vs. extrapolation**

In practice, when we do prediction for some value of x we haven't seen before, we need to be very careful. Predicting y for a value of x that is within the interval of points that we saw in the original data (the data that we fit our model with) is called **interpolation.**

Predicting y for a value of x that's outside the range of values we actually saw for x in the original data is called **extrapolation.**

**Multiple Linear Regression:**

Instead of just a single scalar value x, we have a vector (x1, . . ., xp) for every data point i. So, we have n data points (just like before), each with p different predictor variables or features. We'll then try to predict y for each data point as a linear function of the different x variables.

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

**Multiple dependent variables:** for example, suppose we're trying to predict medical outcome as a function of several variables such as age, genetic susceptibility, and clinical diagnosis. Then we might say that for each patient, x1 = age, x2 = genetics, x3 = diagnosis, and y = outcome.

**Nonlinearities:** Suppose we want to predict a quadratic function y = ax$^2$+bx+c: then for each data point we might say x1 = 1, x2 = x, and x3 = x 2. This can easily be extended to any nonlinear function we want.

**Cost Function:**

The cost function helps us to figure out the best possible values for β0 and β1 which would provide the best fit line for the data points. Since we want the best values for β0 and β1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

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

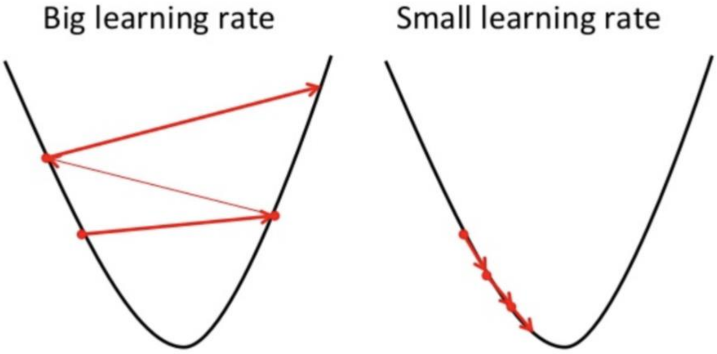
$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Minimization and Cost Function

The difference between the predicted values and ground truth measures the error difference. square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Cost function is nothing but Mean Squared Error(MSE) function.

**Gradient Descent:**

Gradient descent is a method of updating β0 and β1 to reduce the cost function (MSE). The idea is that we start with some values for β0 and β1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.



Gradient Descent

**Optimisation Methods:**

There are two types of Optimisation methods,

1. Closed form solution
2. Iterative form solution

Gradient descent is a Closed form Solution, cost function for gradient decent given as

$$\frac{\partial}{\partial \theta} J(\theta)$$

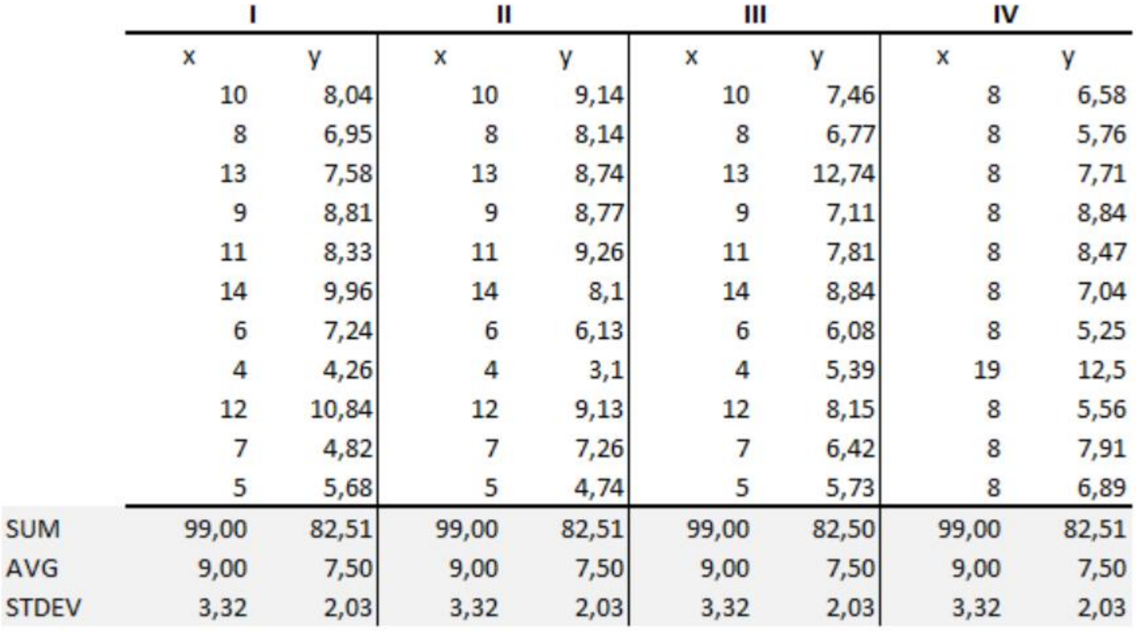
For Gradient descent is an iterative method, $\theta^1$ can be calculated as

$$\theta^1 = \theta^0 - \eta \frac{\partial}{\partial \theta} J(\theta)$$

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

Ans:

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize **COMPLETELY,** when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



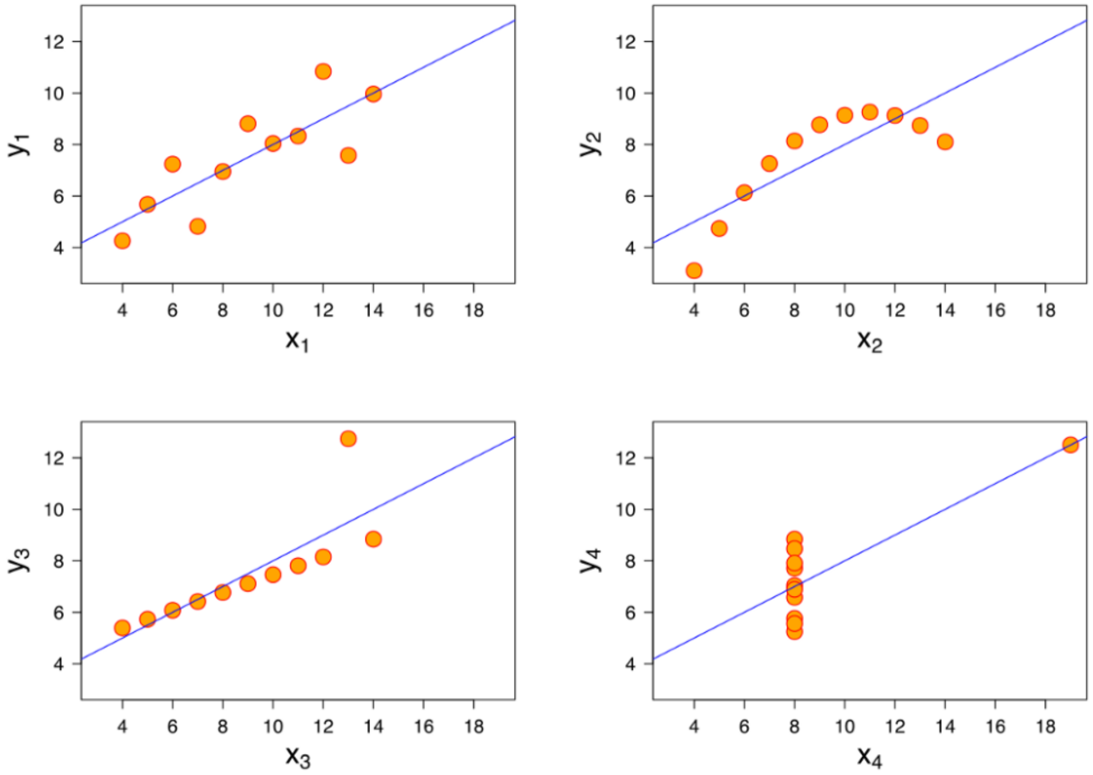
|  | I |  | II |  | III |  | IV |  |
|---|---|---|---|---|---|---|---|---|
|  | x | y | x | y | x | y | x | y |
|  | 10 | 8,04 | 10 | 9,14 | 10 | 7,46 | 8 | 6,58 |
|  | 8 | 6,95 | 8 | 8,14 | 8 | 6,77 | 8 | 5,76 |
|  | 13 | 7,58 | 13 | 8,74 | 13 | 12,74 | 8 | 7,71 |
|  | 9 | 8,81 | 9 | 8,77 | 9 | 7,11 | 8 | 8,84 |
|  | 11 | 8,33 | 11 | 9,26 | 11 | 7,81 | 8 | 8,47 |
|  | 14 | 9,96 | 14 | 8,1 | 14 | 8,84 | 8 | 7,04 |
|  | 6 | 7,24 | 6 | 6,13 | 6 | 6,08 | 8 | 5,25 |
|  | 4 | 4,26 | 4 | 3,1 | 4 | 5,39 | 19 | 12,5 |
|  | 12 | 10,84 | 12 | 9,13 | 12 | 8,15 | 8 | 5,56 |
|  | 7 | 4,82 | 7 | 7,26 | 7 | 6,42 | 8 | 7,91 |
|  | 5 | 5,68 | 5 | 4,74 | 5 | 5,73 | 8 | 6,89 |
| SUM | 99,00 | 82,51 | 99,00 | 82,51 | 99,00 | 82,50 | 99,00 | 82,51 |
| AVG | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 |
| STDEV | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 |

Quartet's Summary Stats

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

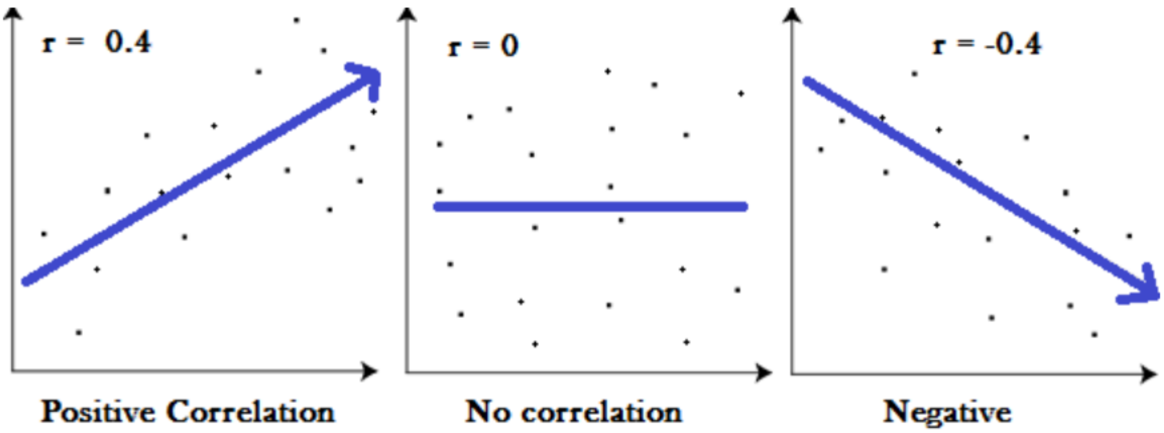
**3. What is Pearson's R?**
Ans:

**Correlation coefficients** are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation commonly used in linear regression.

The Pearson correlation coefficient also referred to as Pearson's R, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1.

A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Example:
Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson's correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

**Sample correlation coefficient:**

$\mathbf{r_{xy} = s_{xy}/s_x s_y}$

$s_x$ and $s_y$ are the sample standard deviations, and $s_{xy}$ is the sample covariance.

**Population correlation coefficient:**

$\boldsymbol{\rho_{xy} = \sigma_{xy}/\sigma_x\sigma_y}$

The population correlation coefficient uses $\sigma_x$ and $\sigma_y$ as the population standard deviations, and $\sigma_{xy}$ as the population covariance.

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

Ans:

**Scaling:**

**Feature scaling** (also known as **data normalization**) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

In some machine learning algorithms, objective functions will not work properly without normalization.

For example, 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.

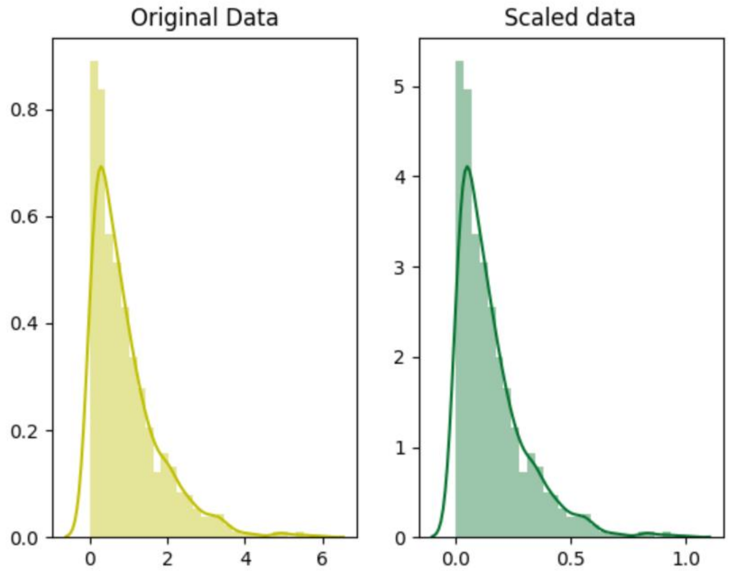
In scaling *(also called* **min-max scaling***)*, you transform the data such that the features are within a specific range e.g. [0, 1].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x' is the normalized value.

This means that you're transforming your data so that it fits within a specific scale, like 0–100 or 0–1. You want to scale data when you're using methods based on measures of how far apart data points, like support vector machines, or SVM or k-nearest neighbours, or KNN. With these algorithms, a change of "1" in any numeric feature is given the same importance.

example, you might be looking at the prices of some products in both Yen and US Dollars. One US Dollar is worth about 100 Yen, but if you don't scale your prices methods like SVM or KNN will consider a difference in price of 1 Yen as important as a difference of 1 US Dollar. This clearly doesn't fit with our intuitions of the world. With currency, you can convert between currencies.

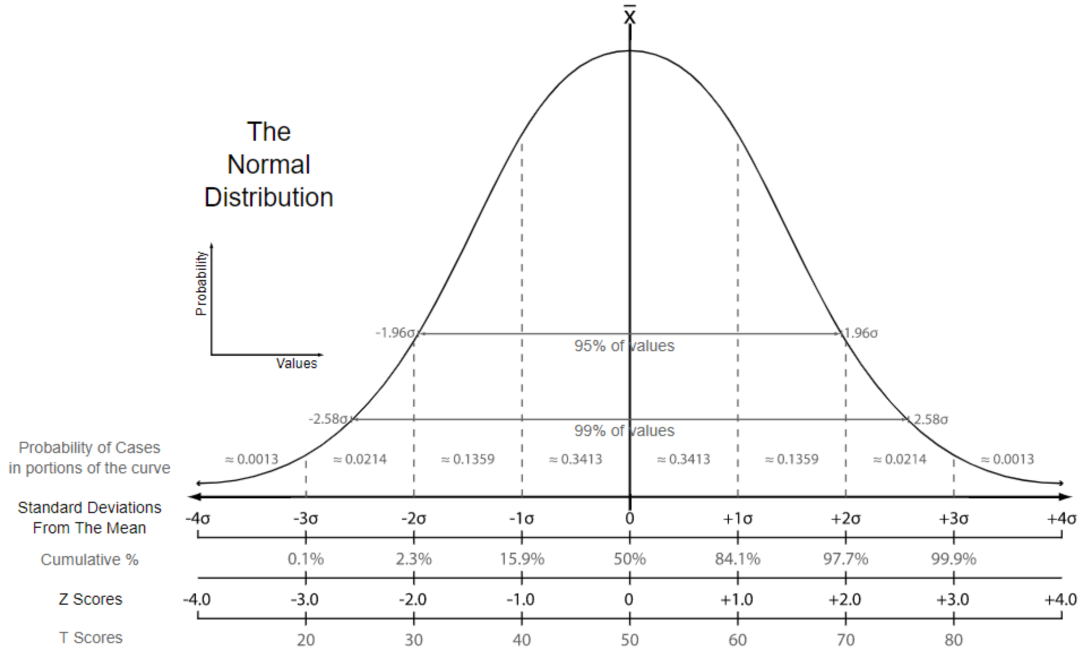


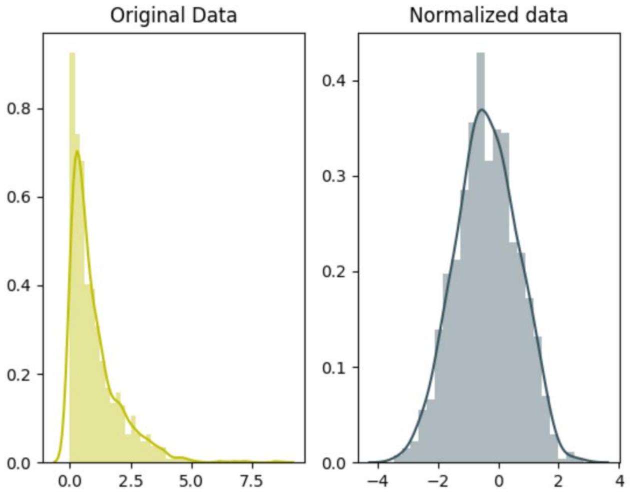
**Normalization and Standardization:**

The point of normalization is to change your observations so that they can be described as a normal distribution.

Scaling just changes the range of your data. Normalization is a more radical transformation. The point of normalization is to change your observations so that they can be described as a normal distribution.

Normal distribution (Gaussian distribution), also known as the **bell curve**, is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean.





**Difference between normalized scaling and standardized scaling:**

In both cases, you’re transforming the values of numeric variables so that the transformed data points have specific helpful properties. The difference is that, in scaling, you’re changing the *range* of your data while in normalization you’re changing the *shape of the distribution* of your data.

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

**Variance Inflation Factor:**

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

We would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$$X_1=C+ \alpha_2 X_2+\alpha_3 X_3+\cdots$$

$$\mathrm{VIF}_1=1/(1-R_1^2 )$$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$$X_2=C+ \alpha_1 X_1+\alpha_3 X_3+\cdots$$

$$\mathrm{VIF}_2=1/(1-R_2^2 )$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).

The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:

- One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.

- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
- Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

In conclusion, when you are building a multiple regression model, always check your VIF values for your independent variables and determine if you need to take any corrective action before building the model.

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.**

Ans:

**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.

QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don’t, your residuals aren’t Gaussian and thus your errors aren’t either. This implies that for small sample sizes, you can’t assume your estimator **β.**

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.

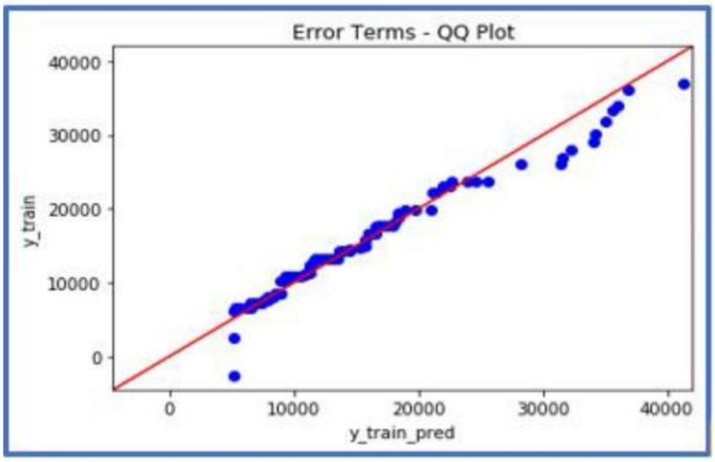
Interpretation:

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

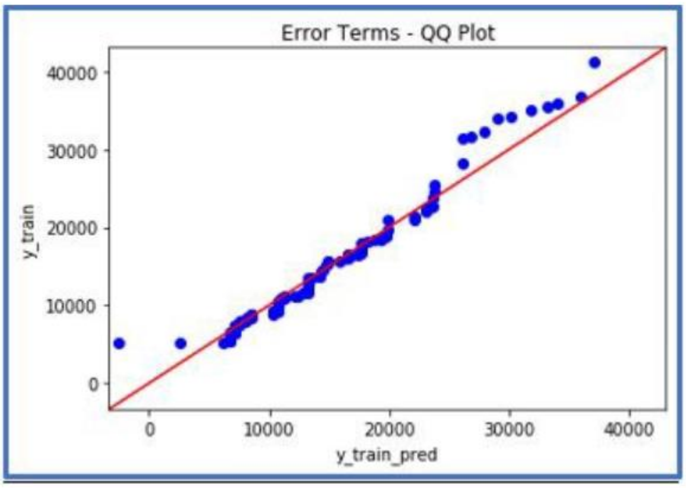
Below are the possible interpretations for two data sets.

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

Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.