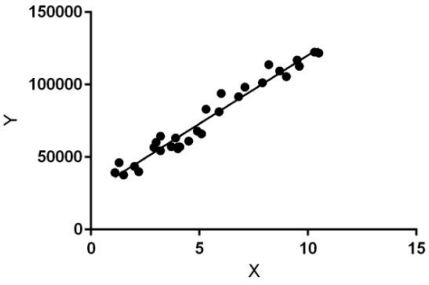
Name - Anusha Kogta
Email - anukogta008@gmail.com

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 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 being 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
In the figure above, X (input) is the work experience and Y (output) is the salary of a person.
The regression line is the best fit line for our model.While training the model we are given :
x: input training data (univariate – one input variable(parameter))

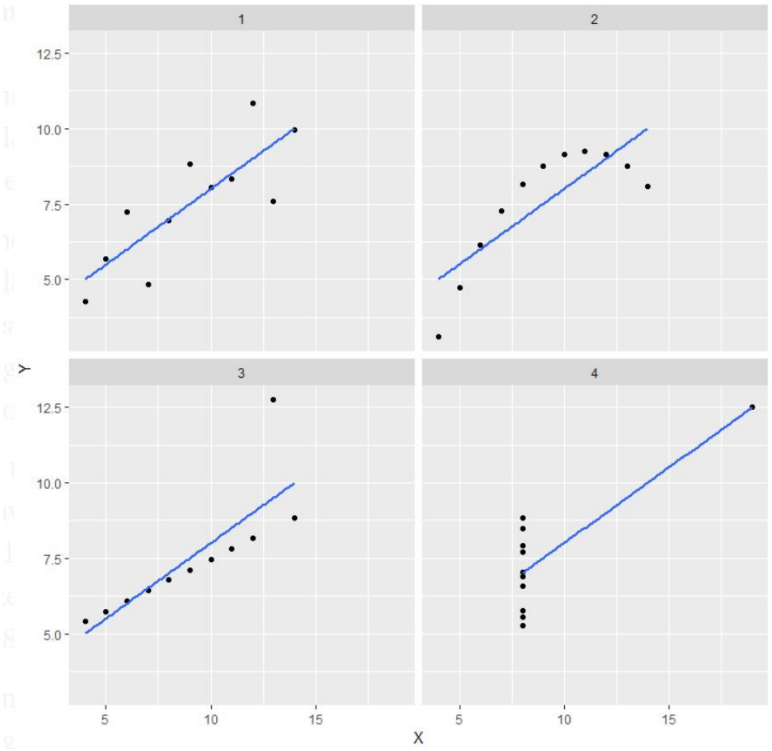
$$y = \theta_1 + \theta_2.x$$

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 θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

2. Explain the Anscombe’s quartet in detail

Ans - 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.



Explanation of this output:
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.

3. What is Pearson's R?

Ans -The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.
The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y.
Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship.
The Pearson coefficient shows correlation, not causation.
English mathematician and statistician Karl Pearson is credited for developing many statistical techniques, including the Pearson coefficient, the chi-squared test, p-value, and linear regression.

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

Ans - 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.
Example: If an algorithm is not using the feature scaling method then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.
Difference between Normalisation and Standardisation
1- Normalisation - a- Is used to transform features to be on a similar scale.
b- Minimum and maximum value of features are used for scaling
c- It is used when features are of different scales.
d- Scales values between [0, 1] or [-1, 1].
e- It is really affected by outliers.
f- Scikit-Learn provides a transformer called MinMaxScaler for Normalization
g- This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.
h- It is useful when we don't know about the distribution

i - It is a often called as Scaling Normalization
j- X_new = (X - X_min)/(X_max - X_min)

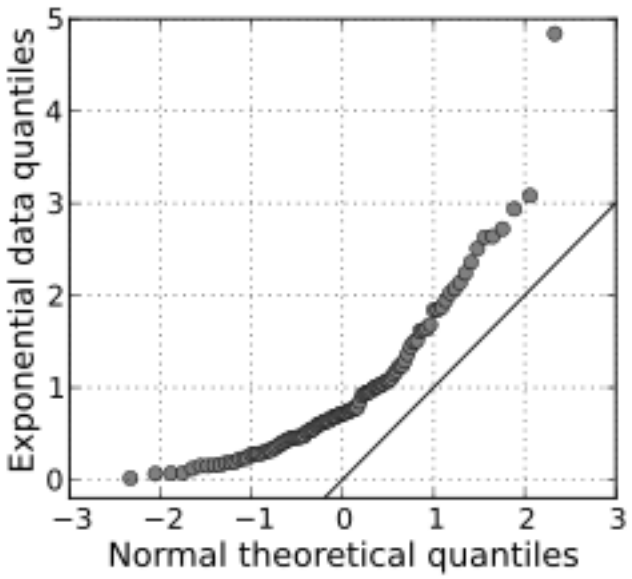
2-Standardisation- a- the transformation of features by subtracting from mean and dividing by standard deviation
b- Mean and standard deviation is used for scaling.
c- It is used when we want to ensure zero mean and unit standard deviation.
d- It is not bounded to a certain range.
e- It is much less affected by outliers.
f- Scikit-Learn provides a transformer called StandardScaler for standardization.
g- It translates the data to the mean vector of original data to the origin and squishes or expands.
h- It is useful when the feature distribution is Normal or Gaussian.
i- It is a often called as Z-Score Normalization.
j- X_new = (X - mean)/Std also called as Z-score

5. 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.
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 - 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.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a normal quantile-quantile (QQ) plot. The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

- How to Make a Q Q Plot

Sample question: Do the following values come from a normal distribution? 7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

Step 1: Order the items from smallest to largest.

:- 3.77,4.25,4.50,5.19,5.89,5.79,6.31,6.79,7.19

Step 2: Draw a normal distribution curve. Divide the curve into n+1 segments. We have 9 values, so divide the curve into 10 equally-sized areas. For this example, each segment is 10% of the area (because 100% / 10 = 10%).

Step 3: Find the z-value (cut-off point) for each segment in Step 3. These segments are areas, so refer to a z-table (or use software) to get a z-value for each segment.

The z-values are:10% = -1.28,20% = -0.84,30% = -0.52,40% = -0.25
50% = 0,60% = 0.25,70% = 0.52,80% = 0.84,90% = 1.28,100% = 3.0

Step 4: Plot your data set values (Step 1) against your normal distribution cut-off points