Linear Regression - Subjective Questions

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?

Answer:

The analysis of categorical variable has been done using boxplot and bar plots. Following are the inferences made about the effect of the categorical variable on the dependent variable based on year 2018 and 2019 ("cnt").

- The highest bookings were done during fall season and the booking in each season has increased drastically from 2018 to 2019.
- Most of the bookings have been done during the month of May, June, July, Aug, Sep and Oct. Bookings started slightly increasing from the start of the year till mid of the year and then it started decreasing at the end of year. Number of booking for each month seems to have increased from 2018 to 2019.
- Thu, Fri, Sat and Sun have more number of bookings as compared to the start of the week. There is drastic increase in booking in 2019 for each day when compared to 2018.
- Clear weather increased more booking. In comparison to previous year(2018), booking increased for each weather situation in 2019.
- The bookings are less when it is not an holiday as compared to holiday. Also the bookings are more in 2019.
- The number of bookings is almost same for both working and non-working days but there is increase in booking in 2019 when compared to 2018.
- The bookings have increased in the year 2019 when compared to 2018.

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

Answer:

A dummy variable is a variable that takes values of 0 and 1, where the values are either observed or not observed. Once a dummy variable is created from the categorical variable it can be used in regression like any other quantitative variable.

The drop_first parameter specifies whether or not you want to drop the first category of the categorical variable you are encoding. drop_first=True drops the first column during dummy variable creation. It is useful as it reduces the number of columns. When 2 to n columns are zero that means the first column is 1. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Syntax:

pd.get_dummies(df,drop_first = True)

Example:

Suppose, you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

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

Answer:

The “temp (temperature) and atemp” variable have the highest correlation with the target/dependent variable “cnt (count)”.

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

Answer:

After designing the final Model I have checked whether the data in the model is violating any of the linear regression assumptions. Below are the assumptions:

- **Normally Distributed Error Terms**

  The residuals are normally distributed that is, the errors are distributed across zero which proves that the model has properly handled the assumption of normal distribution of errors.

- **Error Terms Being Independent**

  When we plotted the scatter plot there is almost no relation between residual and predicted values. This means the model does not have any specific pattern which concludes that they are not dependent on each other.

- **Homoscedasticity**

  The residuals are equally distributed across predicted values and this means we see equal variance and there is neither high concentration of data points in one region nor low concentration of data points in other region. This proves the Homoscedasticity of Error Terms.

- **Multicollinearity**
  Multicollinearity happens when independent variables in the regression model are highly correlated to each other. But in the designed final model there is no multicollinearity between the predicted variables as we could see that the VIF values are below 5.

- **Linear Relationship validation**

  Linear regression model assumes that the relationship between target and feature variables must be linear. It has been checked using CCPR plot.

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

Answer:

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

temp (Temperature) - A coefficient value of '0.4777' indicated that a unit increase in temp variable increases the bike hire numbers by 0.4777 units.

winter - A coefficient value of '0.09447' indicated that a unit increase in winter variable increases the bike hire numbers by 0.09447 units.

sep - A coefficient value of '0.0909' indicated that a unit increase in yr variable increases the bike hir e numbers by 0.0909 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis) called linear regression. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

The linear relationship between the dependent and independent variable can be represented by the following equation.

$$Y = mX + c$$

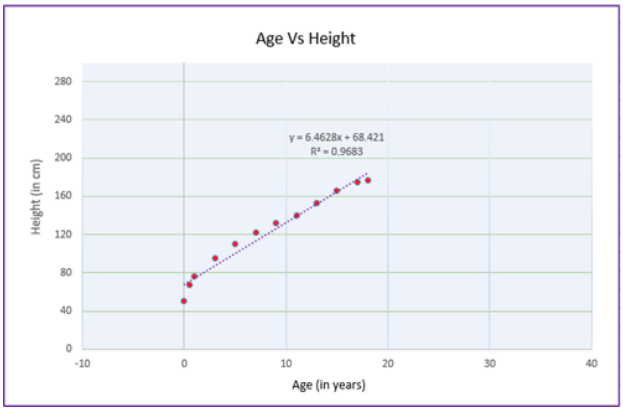
Y is the dependent variable which we have to predict.

X is the independent variable which we are using for prediction.

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Example:



Types of Linear Regression:

There are 2 types of linear regression:

1. Simple Linear Regression
2. Multiple Linear Regressions

Simple Linear Regression:
It is a type of linear regression model where there is only independent or explanatory variable. For e.g., the above scatter plot follows a simple linear regression with age being an independent variable is responsible for any change in height (dependent variable).

Multiple Linear Regressions:
It is similar to simple linear regression but here we have more than one independent or explanatory variable.

Equation for multiple regressions is given by

$$\mathbf{Y = \beta 0 + \beta 1.X1 + \beta 2.\ X2 + \beta 3.\ X3 + \beta 4.\ X4 + \beta 5.\ X5 + \beta 5.\ X6 + \epsilon}$$

Where,

Y = target/response variable

X1, X2, X3… = predictor variables

β0 = Y-intercept (always a constant)

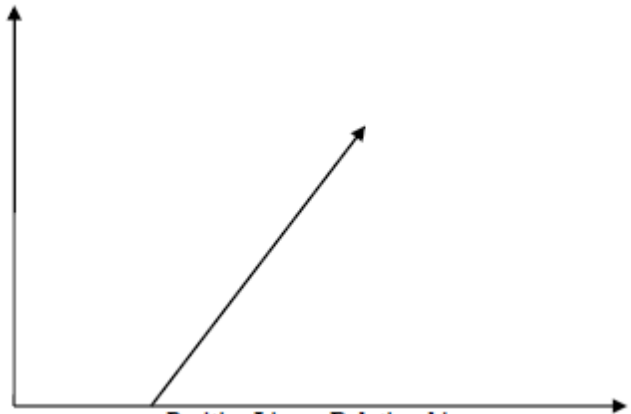
β1, β2, β3, β4, β5 = regression coefficients

ϵ = Error terms (Residuals)

Furthermore, the linear relationship can be positive or negative in nature as explained below−

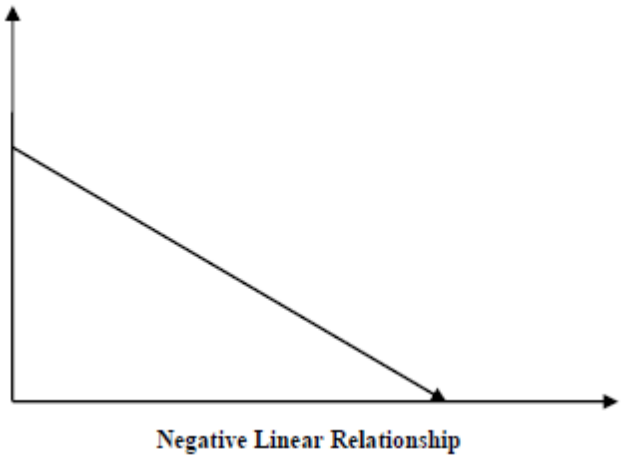
1 . Positive Linear Relationship:

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



2. Negative linear relationship:

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Steps to be followed in Linear Regression Algorithm:

1. Reading and understanding the data

2. Visualizing the data (Exploratory Data Analysis)

3. Data Preparation

4. Splitting the data into training and test sets

5. Building a linear model

6. Residual analysis of the train data

7. Making predictions using the final model and evaluation.

The following are some assumptions about dataset that is made by Linear Regression model :

- **Multi-collinearity:**
  Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

- **Auto-correlation:**
  Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

- **Normality of error terms:**
  Error terms should be normally distributed across zero.

- **Relationship between variables:**
  Linear regression model assumes that the relationship between response and feature variables must be linear.

- **Homoscedasticity:**
  There should be no visible pattern in residual values. The residuals are equally distributed across predicted values and this means we see equal variance and there is neither high concentration of data points in one region nor low concentration of data points in other region. This proves the Homoscedasticity of Error Terms.

Advantage and disadvantage of linear regression algorithm

The advantage and disadvantage of linear regression algorithm:

1. Linear regression provides a powerful statistical method to find the relationship between variables. It hardly needs further tuning. However, it's only limited to linear relationships.
2. Linear regression produces the best predictive accuracy for linear relationship whereas it's little sensitive to outliers and only looks at the mean of the dependent variable.

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 analyzing it and the effect of outliers on statistical properties. Those 4 sets of 11 data-points are given below.

Anscombe's quartet

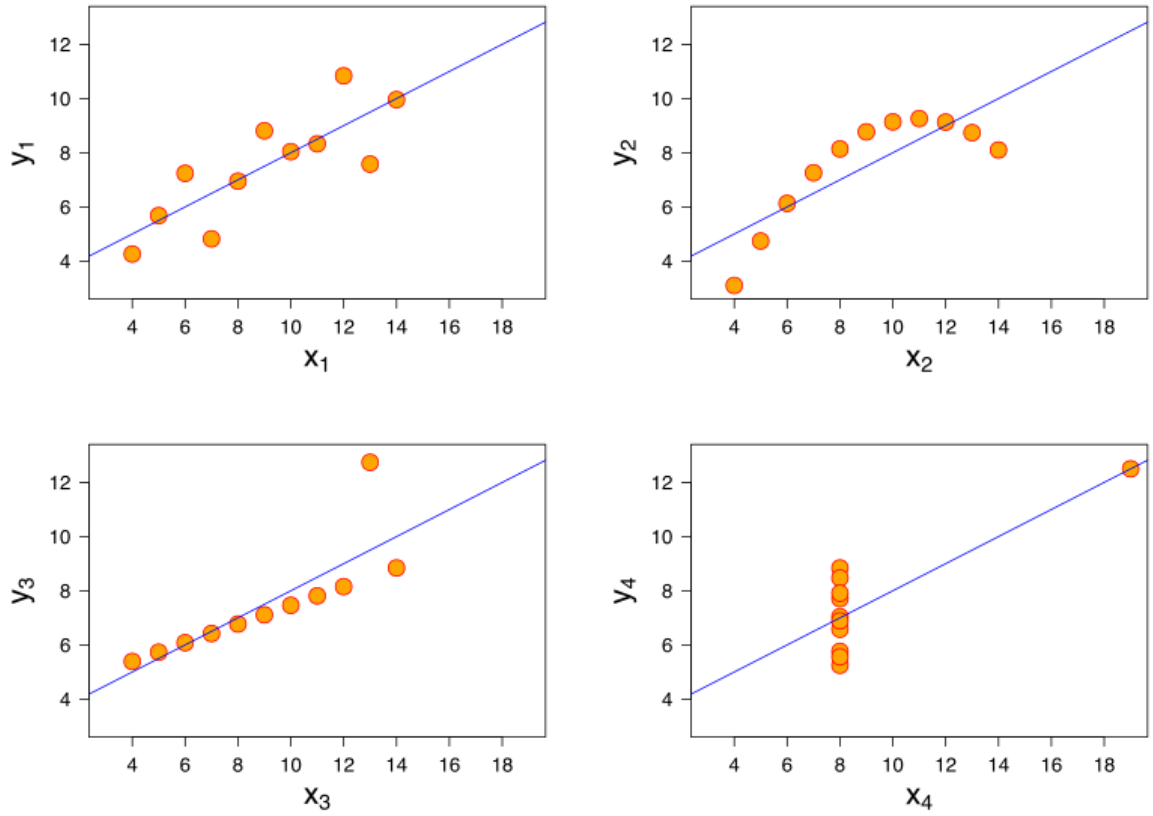
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets.

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	

• 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 0.82 for each dataset

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.



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.

CONCLUSION

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 us 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.

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.

3. What is Pearson's R?

Correlation measures the strength of association between two variables as well as the direction. There are mainly **three types** of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables.

Correlation is a statistic that measures the relationship between two variables in the finance and investment industries. It shows the strength of the relationship between the two variables as well as the direction and is represented numerically by the correlation coefficient. The numerical values of the correlation coefficient lies between **-1.0 and +1.0**.

When the value of the correlation coefficient is exactly 1.0, it is said to be a perfect positive correlation. This situation means that when there is a change in one variable, either negative or positive, the second variable changes in lockstep, in the same direction.

A perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no linear relationship at all. We can determine the strength of the relationship between two variables by finding the absolute value of the correlation coefficient.

Pearson's Correlation Coefficient ®

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

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

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

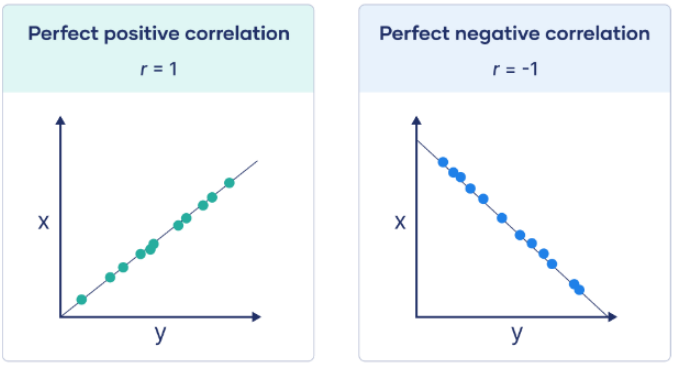
Σy2 = the sum of squared y scores

Visualizing the Pearson correlation coefficient

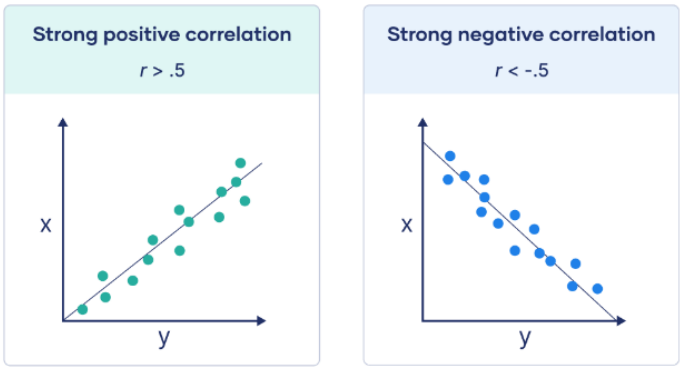
Another way to think of the Pearson correlation coefficient (*r*) is as a measure of how close the observations are to a line of best fit.

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.

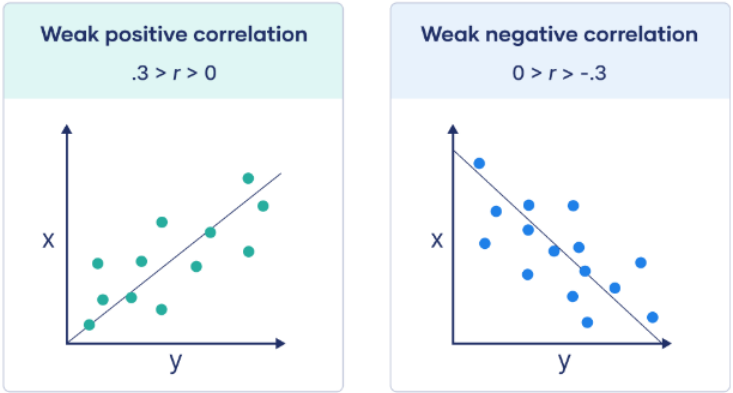
When *r* is 1 or –1, all the points fall exactly on the line of best fit:



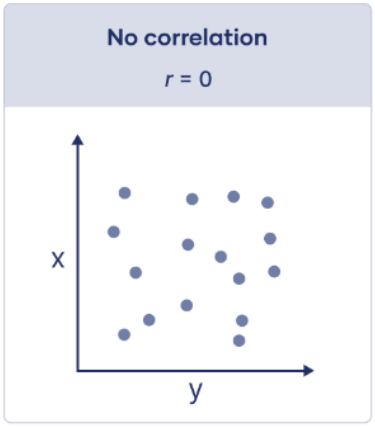
When r is greater than .5 or less than –.5, the points are close to the line of best fit:



When r is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:



When r is 0, a line of best fit is not helpful in describing the relationship between the variables:



When to use the Pearson correlation coefficient

The Pearson correlation coefficient (*r*) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when **all** of the following are true:

- **Both variables are quantitative:** You will need to use a different method if either of the variables is qualitative.
- **The variables are normally distributed:** You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- **The data have no outliers:** Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- **The relationship is linear:** "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

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/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**)

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

For e.g. Male height, female height, and lifespan are the three **features** in the example below.

Male height in Feet	Female height in cm	Life span
7.1	155	36
8.5	160	48
6.9	170	38
8.2	140	41

For understanding feature scaling first imagine a scale in your mind. The scale has a range of values in it. Feature scaling is a method to scale numeric features in the same scale or range (like -1 to 1,0 to 1). This is the last step of data preprocessing and is done before machine model training. We apply feature scaling on independent variables. We fit feature scaling with train data and transform (apply) on train and test data.

In the above example, you have seen that different features have data in different units. As we were having a female height in **cm** and male height in **feet.** As you know in actuality 8.2 feet > 140 cm. But machine learning algorithm will interpret it as 140>8.2 because it is checking according to value only not according to feature.ML algorithms can't understand features, and units, understand numbers only. That's why we need to scale down all the values on one level. In the below figure we have data

of students which is scaled down. Otherwise, the machine learning algorithm would have taken 60>3.0 (considering the value)

	Student	Ratings	Marks
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

	Student	Ratings	Marks
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

Differences between normalized scaling and standardized scaling

S. No	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

Answer:

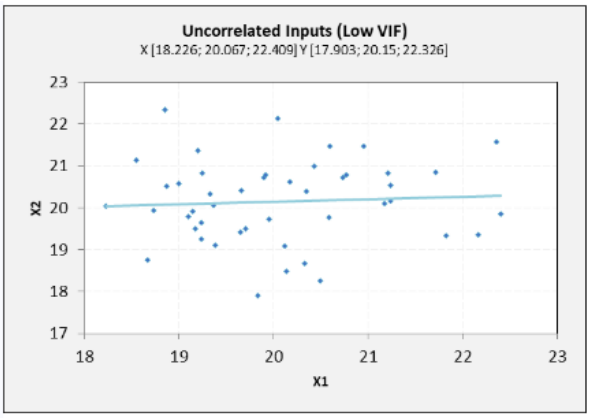
VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

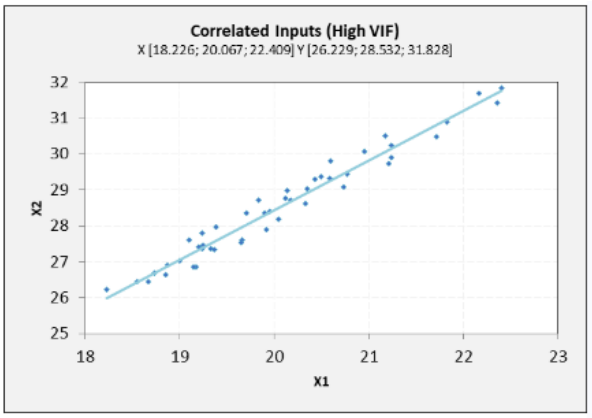
The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.





VIF Threshold

- **VIF > 10** indicates a serious collinearity problem
- **VIF > 5** is cause for concern and
- **VIF ≥ 2.5** indicates considerable collinearity

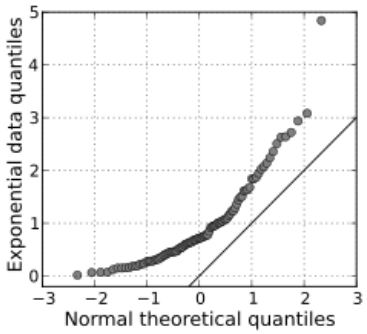
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the

distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

A Q Q plot showing the 45 degree reference line:



Uses of Q-Q plot:

- It can be used with sample sizes also
- 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 if two data sets come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.