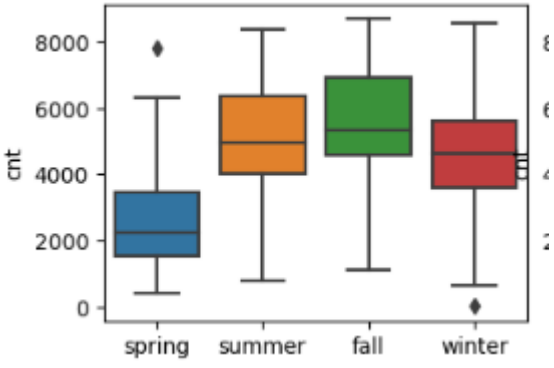
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

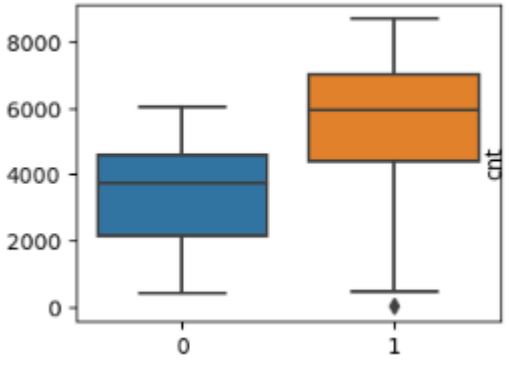
Solution:

Observations when analysis categorical variable.

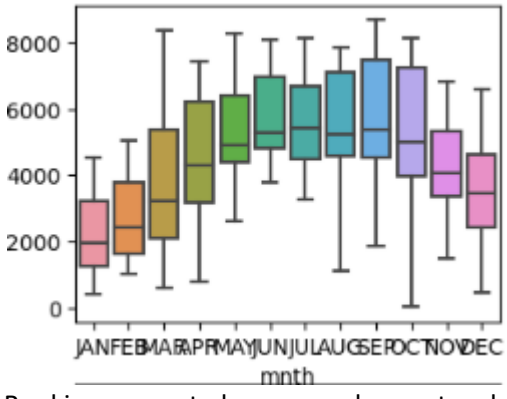
- Fall season has more bookings



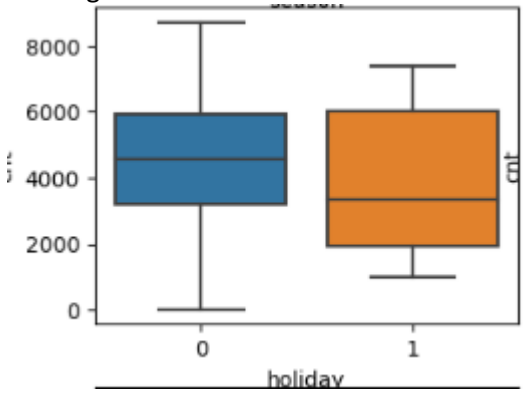
- Year 2019 has seen more booking ---seems business is in good state compared to previous year



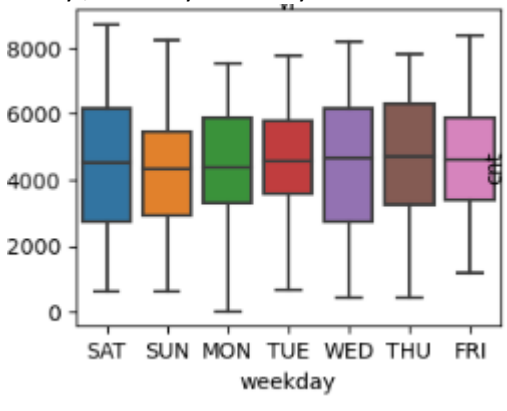
- May ,june ,july ,aug ,sep and oct has more number of bookings the pattern shows bookings increased steadily from starting of year till mid of year and decreased at last couple of months .



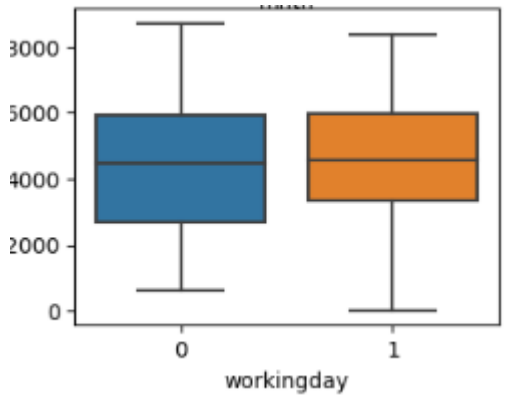
- Booking seems to be more when not an holiday .



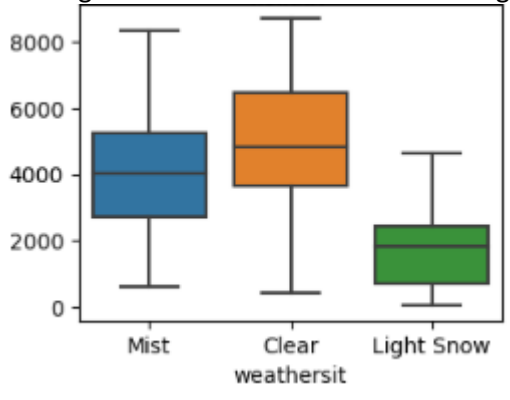
- Friday ,Saturday& Sunday have more number of bookings .



- Booking seems no impact weather its an working day or non –working day



- Booking for clear weather seems to be high as when compared to other



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

Solution:

This kind of approach is used when we have categorical variables with k-values in order to delete extra column when we used to create dummy columns for the categorical variables .
For example if we have one categorical column say water quality which has their values

- Sweet
- Salt
- sour

sweet	salt	sour
1	0	0
0	1	0
0	0	1

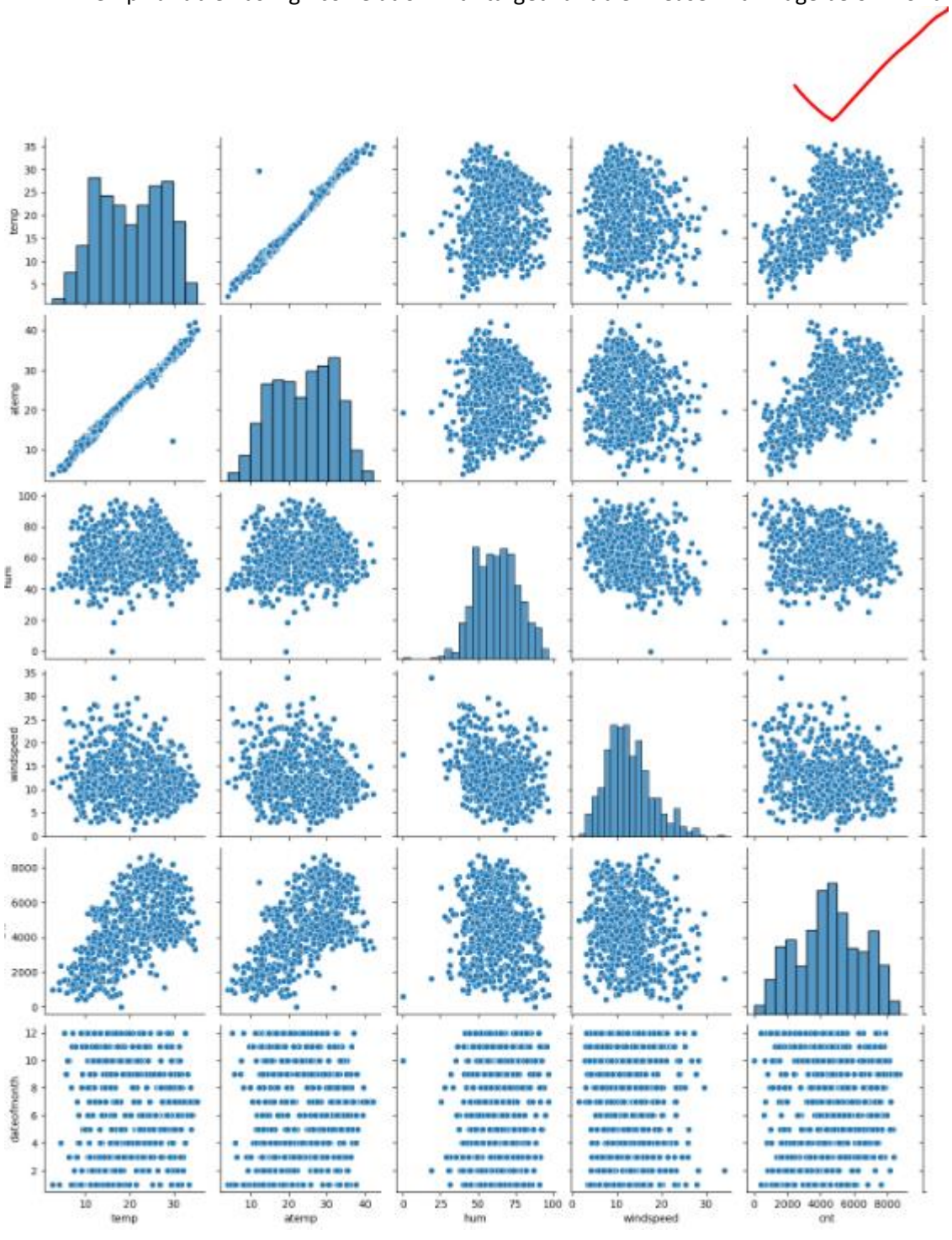
Here if we observe first row if salt and sour have values 0 and 0 then obviously it states water quality is sweet then why we need sweet column.

Dropping one variable helps in reducing correlation between other variables

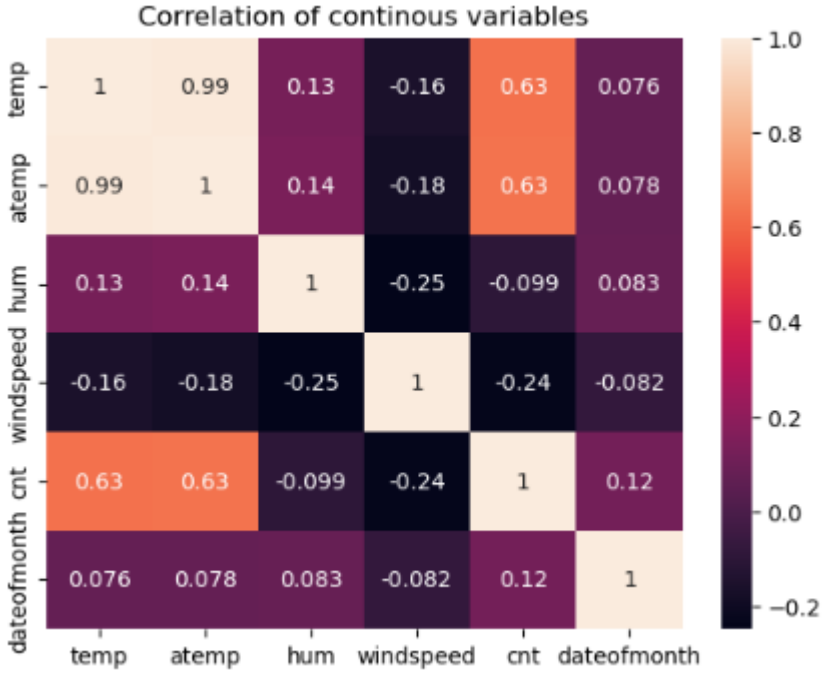
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Solution :

Temp variable has high correlation with target variable .Please find image below for the same.



Additional information

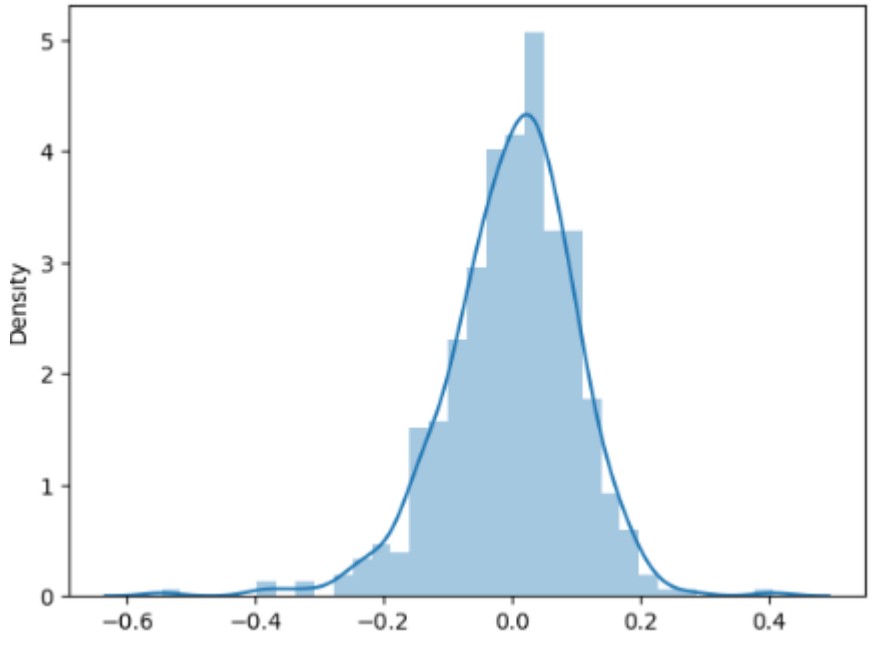


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Solution :

Based on five assumptions :

1)First we check any multi relation between the variables i.e (**multi collinearity** among variables)
2)Is there **any linear relationship** among variables linearity should be observed .
3) **Residuals** should be **normally distributed**.
4) The Durban Watson statistic will always assume a value between 0 and 4. A value of DW = 2 indicates that there is no autocorrelation. When the value is below 2, it indicates a positive autocorrelation
5)check for patterns in residuals –**Homoscedasticity**



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

Solution:

The top variables contributing towards demand of bikes are

1) yr

2) SAT

3) Working day

4) Sep

	coef
const	0.5452
yr	0.2459
workingday	0.0562
windspeed	-0.1920
DEC	-0.1119
JAN	-0.1221
JUL	-0.0149
NOV	-0.1037
SEP	0.0503
spring	-0.2487
summer	-0.0485
winter	-0.0186
SAT	0.0660
Light Snow	-0.3171
Mist	-0.0887

General Subjective Questions

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

Solution :

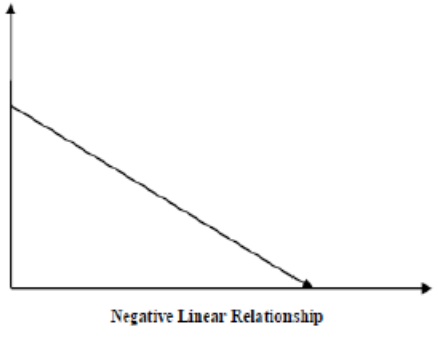
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

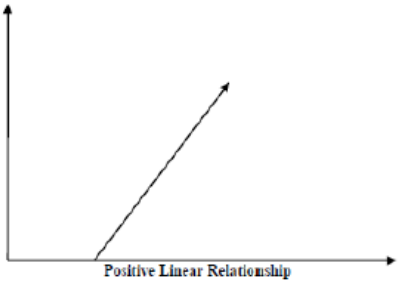
This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Mathematically the relationship can be represented with the help of following equation − Y = c+mX
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
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.

Linear relationship can be classified into positive or negative in nature

- Positive Linear Relationship-A linear relationship will be called positive if both independent and dependent variable increases.
- Negative Linear Relationship- A linear relationship will be called positive if independent increases and dependent variable decreases.


Negative Linear Relationship


Positive Linear Relationship

Assumptions for the Linear regression model :

1) **Linear relationship** between the variables.

2)There should be **no multi co linearity** between the variables

3)Residual or so called error terms should be **distributed normally**

4)Patterns should not be visible in residual values –**Homoscedasticity**

5)There should be no dependency between residual errors-**Auto correlation**

2. Explain the Anscombe's quartet in detail.

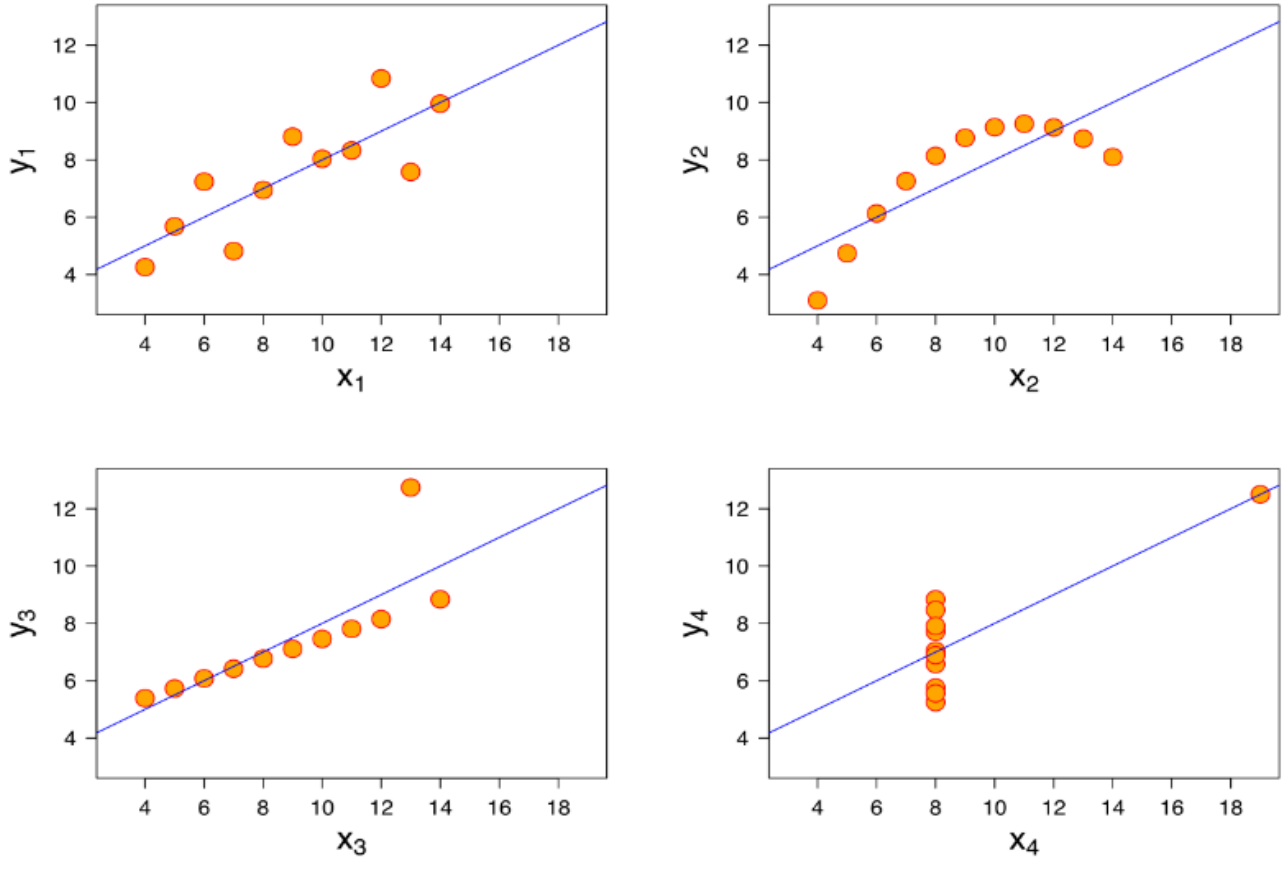
Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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. 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

	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

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:



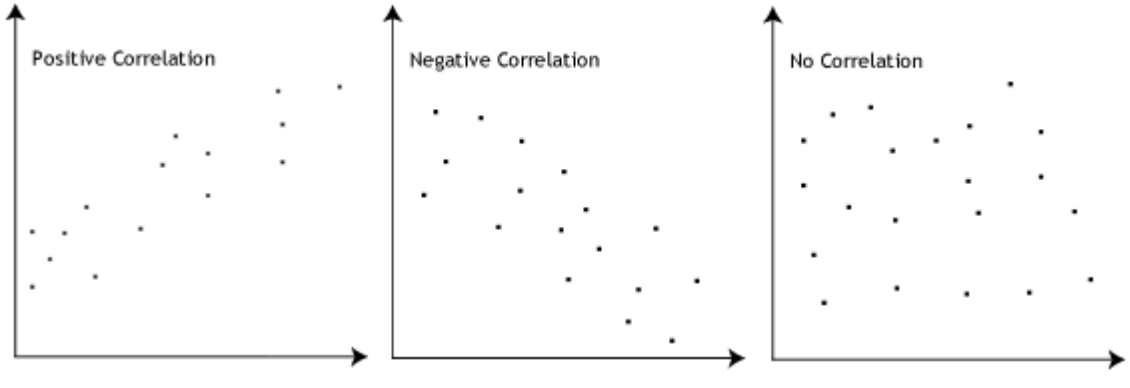
• Dataset I appears to have clean and well-fitting linear models.
• Dataset II is not distributed normally.
• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3) What is Pearson's R?

Solution :

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight:
			The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers:
			The price of a car is not related to the width of its windshield wipers.
Between 0 and -1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure:
			The higher the elevation, the lower the air pressure.



5. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Solution :

Feature 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.

For Example :

Model or algorithm is not using feature scaling method then it can consider the value 300 ml to be greater than 5 litres 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 same magnitudes and thus, tackle this issue.

They are two types of ----min max scaling which usually called as normalized and Standardized scaling

Normalized scaling Features:

1) Minimum and maximum value of features are used for scaling
2) It is used when features are of different scales.
3) Scales values between [0, 1] or [-1, 1].
4) It is really affected by outliers.
5) Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

Standardized Scaling :
1) Mean and standard deviation is used for scaling.
2) It is used when we want to ensure zero mean and unit standard deviation.
3) It is not bounded to a certain range.
4) It is much less affected by outliers
5) Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Solution :

VIF is the one of the factor to measure the correlation between the variables if VIF is infinite then it means it is perfectly correlated

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Solution :

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

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.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.