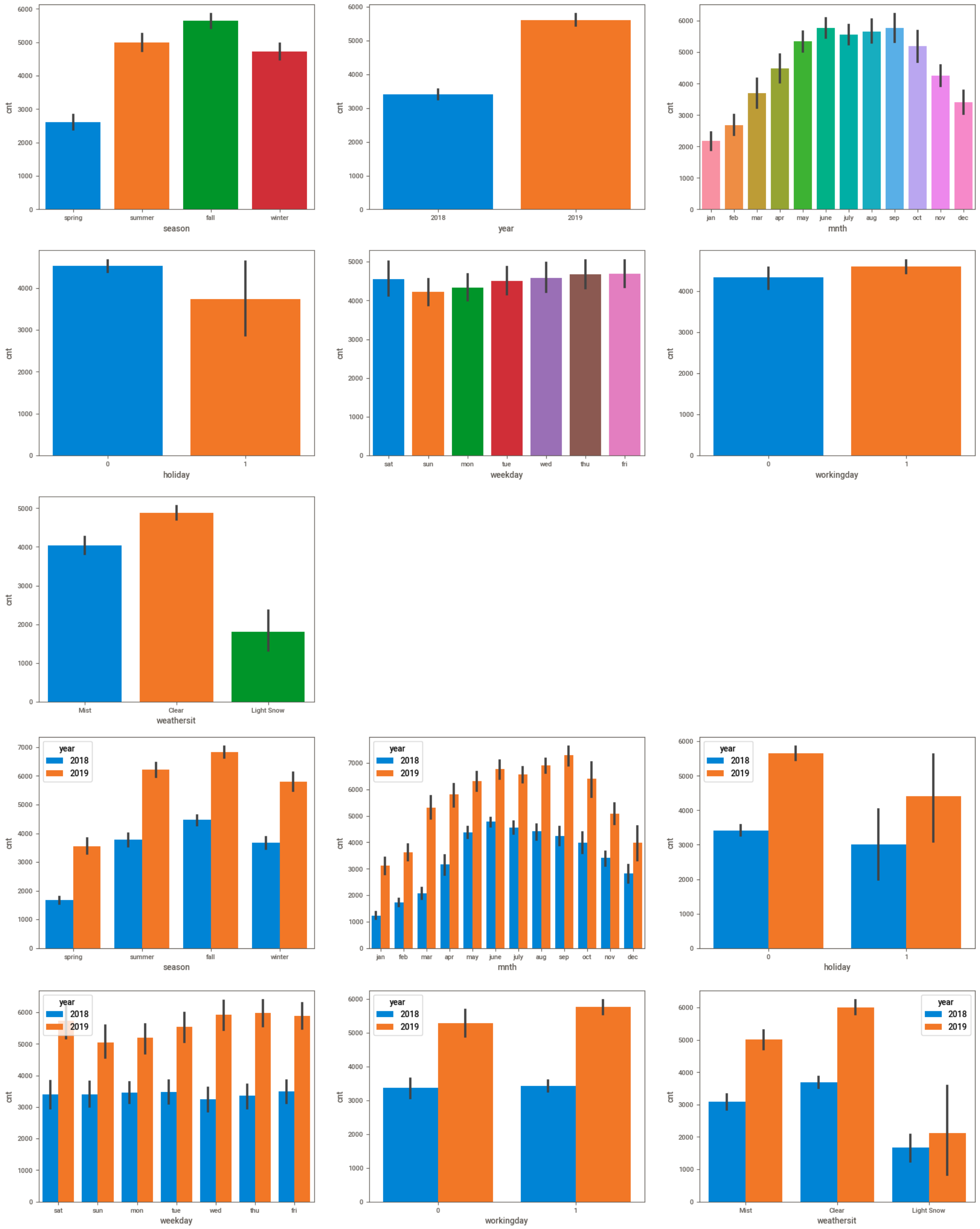
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**



We can infer following points from the analysis of categorical variables:-

- In Fall seasons, there are more bookings.
- Bookings has increased in good numbers from 2018 to 2019. In fact, For each categorical variable, booking has increased drastically from 2018 to 2019
- There is some trend in month wise bookings. In first and last few months, booking are less. Generally, bookings are higher during months of May to Oct.
- On holiday, there seems to be less booking.
- As weekend approaches, there is slight rise in bookings.
- There is marginal rise in booking on working day.
- On Clear weather , there are more bookings.

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

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

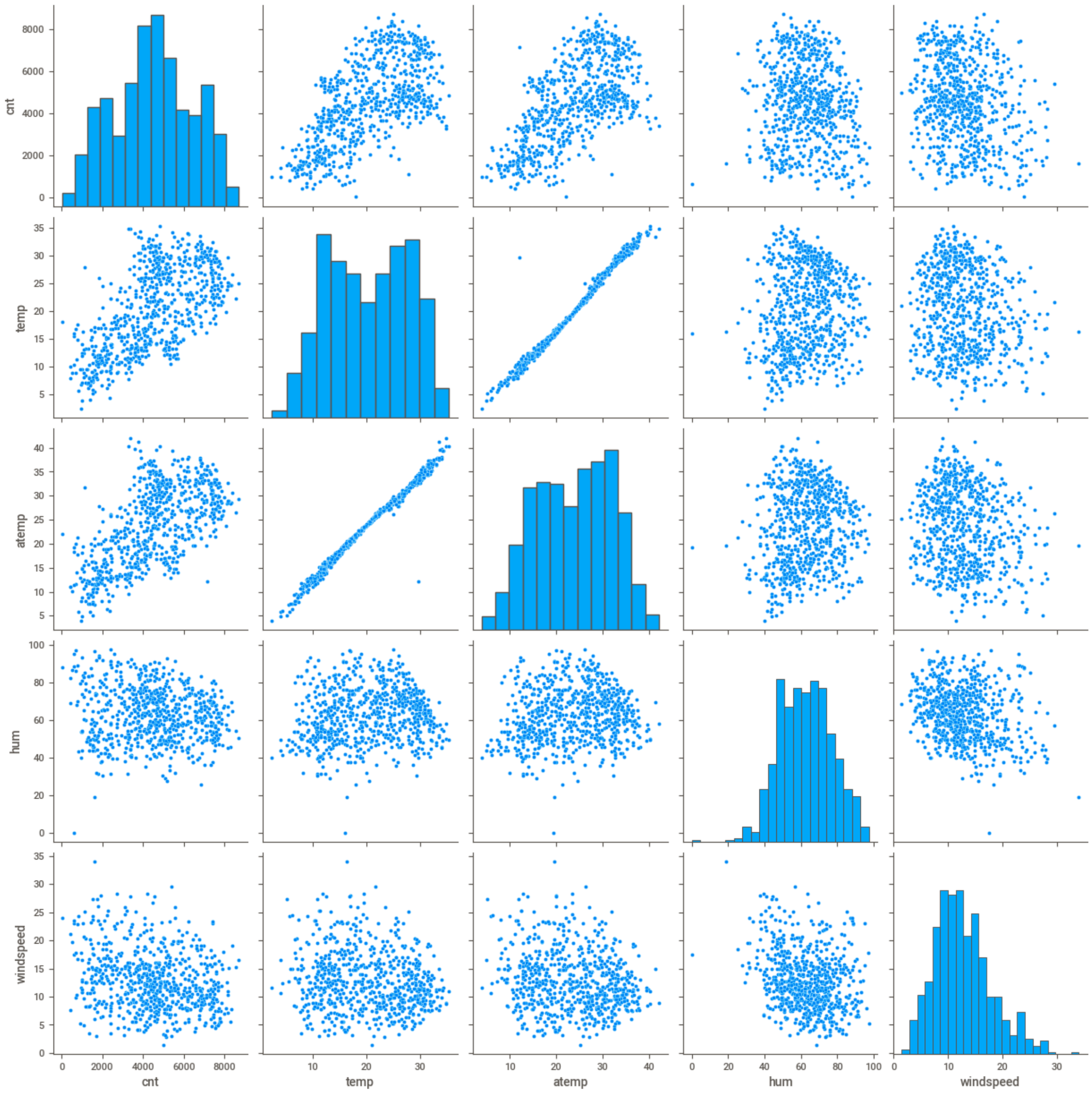
Syntax -
drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's take an example of furnishingstatus having 3 types of values as Furnished, unfurnished and Semi Furnished. We need only 2 dummy variables. Let's see below

Unfurnished	Semi Furnished	
0	0	This is Furnished, because both values are 0.
0	1	This is Semi-Furnished
1	0	This is Unfurnished

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Looking at pair plot, temp as high correlation with the target variable.

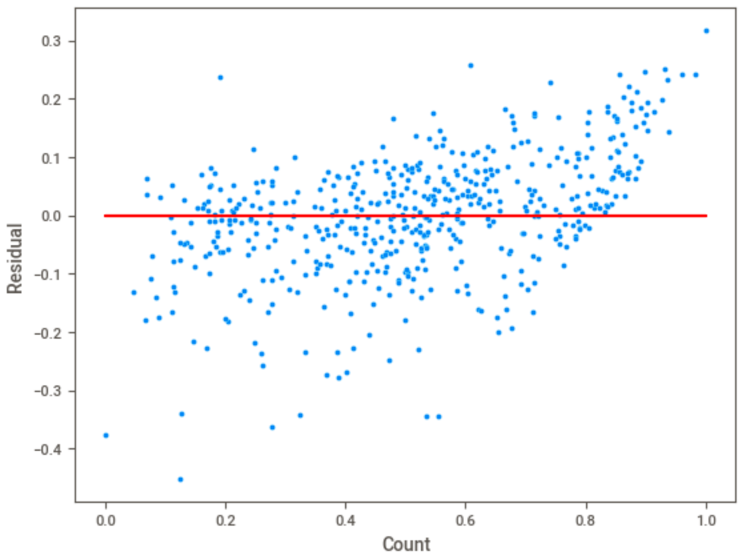


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

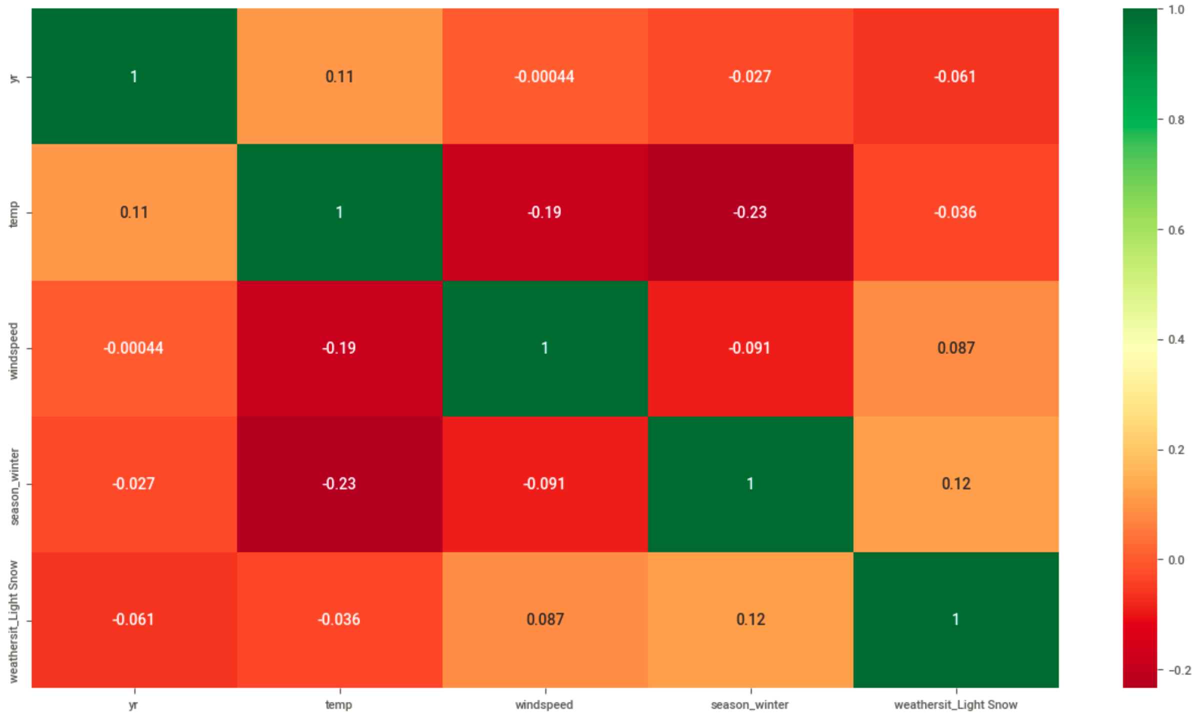
I have validated the assumptions of Linear Regression based on 5 assumptions:-

1. **Linear Relationship** :- This assumes that there exists a linear relationship between the dependent variable and target variable. This we checked by drawing a pair plot. Refer to above pair plot

2. **Homoscedasticity** :- This means residual have constant variance no matter the level of dependent variable. We verified this using scatter plot on residual.

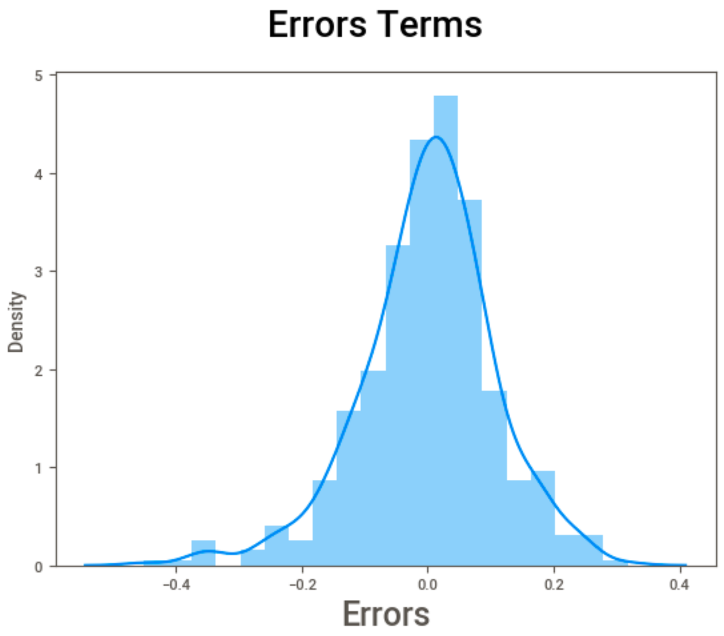


3. **Absence of Multicollinearity** :- Multicollinearity refers to having a high correlation between two or more independent variables. This we have verified checking the VIF of independent variables of final model. As a rule of thumb, VIF less that 5 generally indicates absence of multicollinearity.



	yr	temp	windspeed	season_winter	weathersit_Light Snow
yr	1	0.11	-0.00044	-0.027	-0.061
temp	0.11	1	-0.19	-0.23	-0.036
windspeed	-0.00044	-0.19	1	-0.091	0.087
season_winter	-0.027	-0.23	-0.091	1	0.12
weathersit_Light Snow	-0.061	-0.036	0.087	0.12	1

4. **Normality of Errors** :- If the residuals are not normally distributed, model may become biased. We have verified by drawing a distribution of residuals against levels of dependent variable. It is normally distributed.



5. **Independence of residuals** :- There should be no auto-correlation. auto-correlation occurs when there is dependency between residual errors.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes :-

- temp
- windspeed
- yr

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is one of the very basic forms of machine learning in the field of data science where we train a model to predict the behaviour of your data based on some variables. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form, i.e., numerical value.

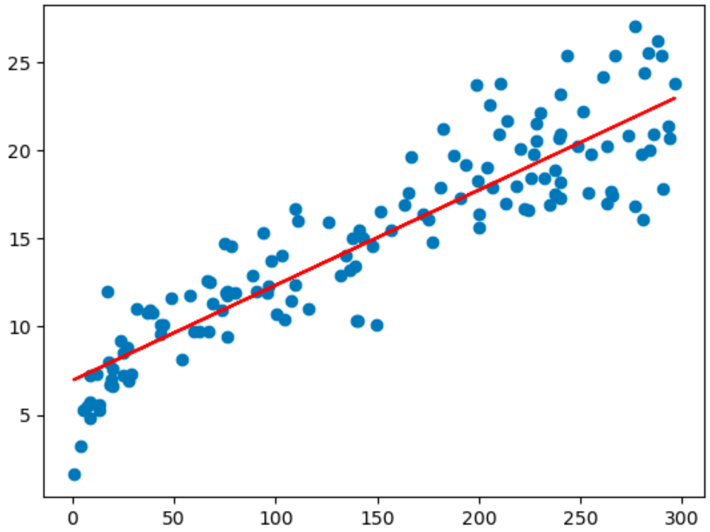
Linear regression can be expressed mathematically as:
$Y = mX + c$

where,

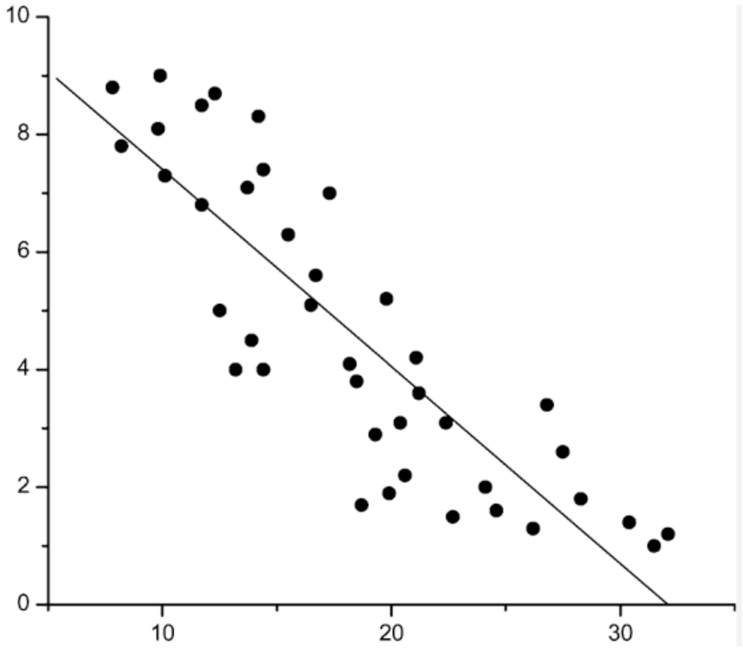
- Y = target variable.
- X = independent variable.
- m = slope of the regression line which represents the effect X has on Y
- c = intercept of line. If X = 0, Y would be equal to c.

Relationship can be of 2 ways:-

I. **Positive Linear Relationship** : In Positive linear Relationship, if one unit increase in independent variable increases the target variable. i.e slope is positive.



II. **Negative Linear Relationship** :- In Negative linear Relationship, one unit increase in independent variable decreases the target variable i.e slope is negative.



Further , Linear Regression is of following 2 types :-

a. **Simple Linear Regression** – A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.

   Mathematically it represents as y=mX+c

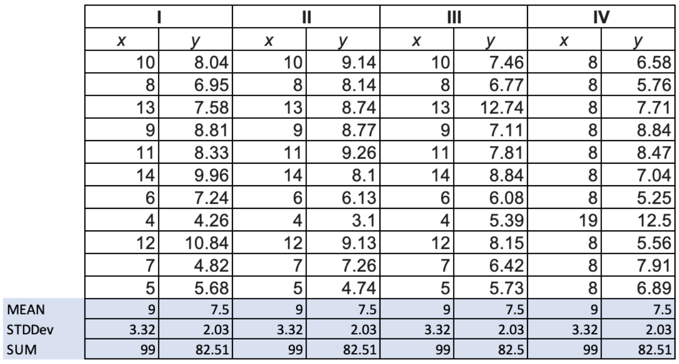
b. **Multiple Linear Regression** – A multiple Linear Regression represents the relationship between two or more independent variables and a target variable. It is needed when one variable might not be sufficient to create a good model and make accurate predictions.

   Mathematically it represents as y=c+m1X1+m2X2 + ……

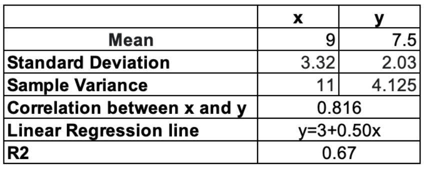
2. **Explain the Anscombe's quartet in detail.** **(3 marks)**

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x , y) points. It was demonstrated by statistician Francis Anscombe to show the effect of outliers and other influential observations on statistical properties.

Let's see below 4 datasets:-



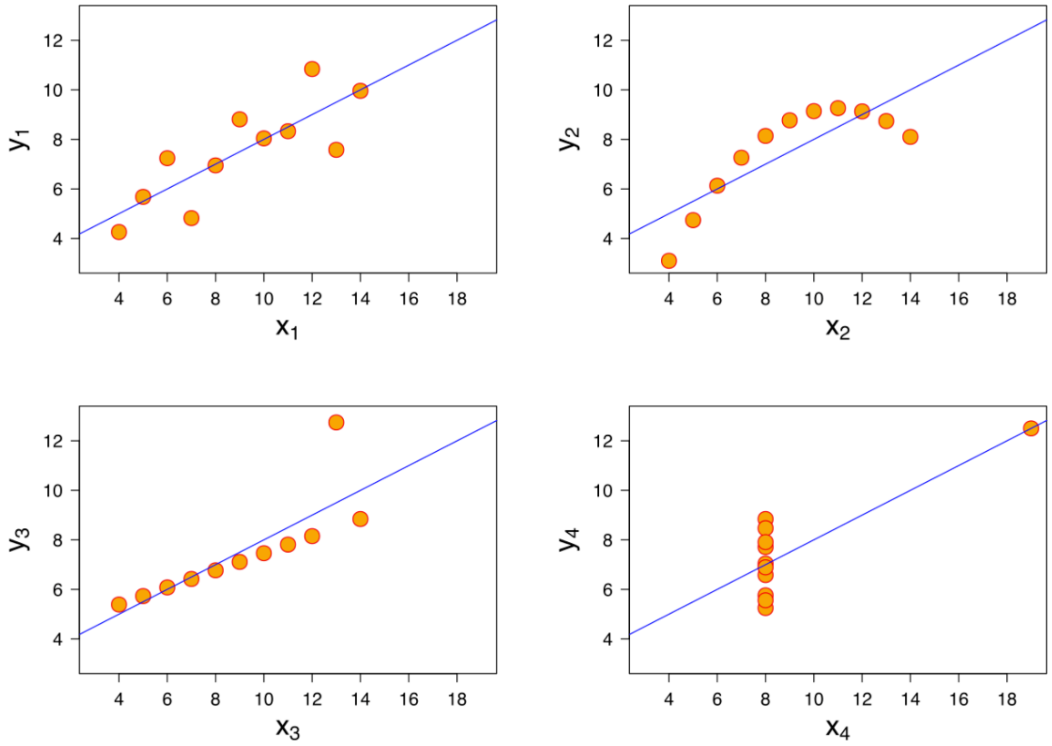
	I		II		III		IV	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	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
MEAN	9	7.5	9	7.5	9	7.5	9	7.5
STDDev	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
SUM	99	82.51	99	82.51	99	82.5	99	82.51



	x	y
Mean	9	7.5
Standard Deviation	3.32	2.03
Sample Variance	11	4.125
Correlation between x and y	0.816	
Linear Regression line	y=3+0.50x	
R2	0.67	

All four datasets are identical when examined using simple summary statistics.

When we plot these four datasets on x/y coordinate, we observe that they show the same regression lines but have different distribution of data.



- 1st dataset appears to be simple linear relationship.
- 2nd graph , relationship between 2 variables are not linear.
- 3rd graph , relationship is linear but should have different regression line. Calculated regression is thrown off by an outlier.
- 4th graph, appears to show that one outlier is enough to produce a high correlation coefficient.

This quartet shows the importance of illustration of data graphically before starting to analyse data.

3. What is Pearson's R? (3 marks)

Pearson's R is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

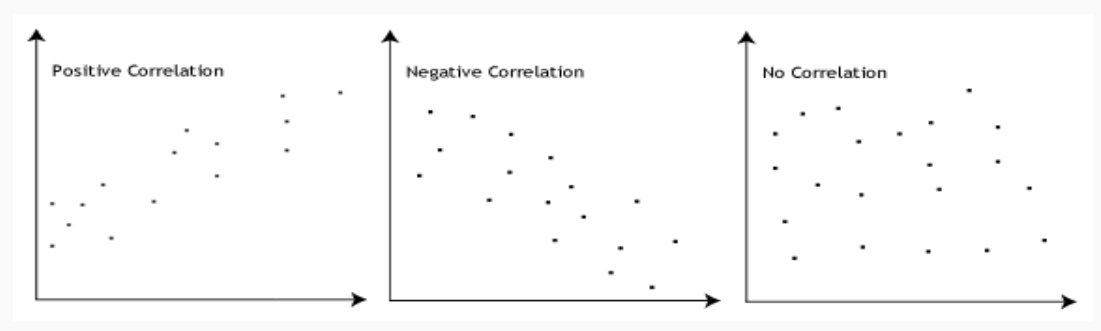
The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

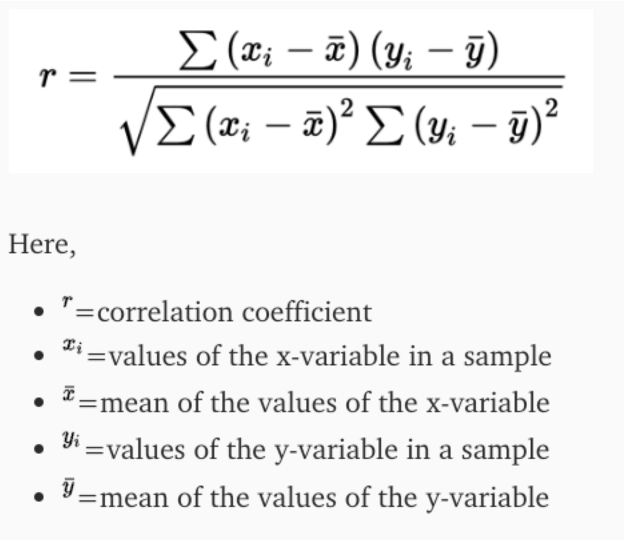
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

- r = 0 means there is no linear association

- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r=correlation coefficient
- x_i=values of the x-variable in a sample
- $\bar{x}$=mean of the values of the x-variable
- y_i=values of the y-variable in a sample
- $\bar{y}$=mean of the values of the y-variable

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

   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. 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.

Sr No.	Normalized Scaling	Standardized scaling

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	Math formula : X_new = (X - X_min)/(X_max - X_min)	Math formula :- X_new = (X - mean)/Std

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

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. (3 marks)**

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

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.

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.