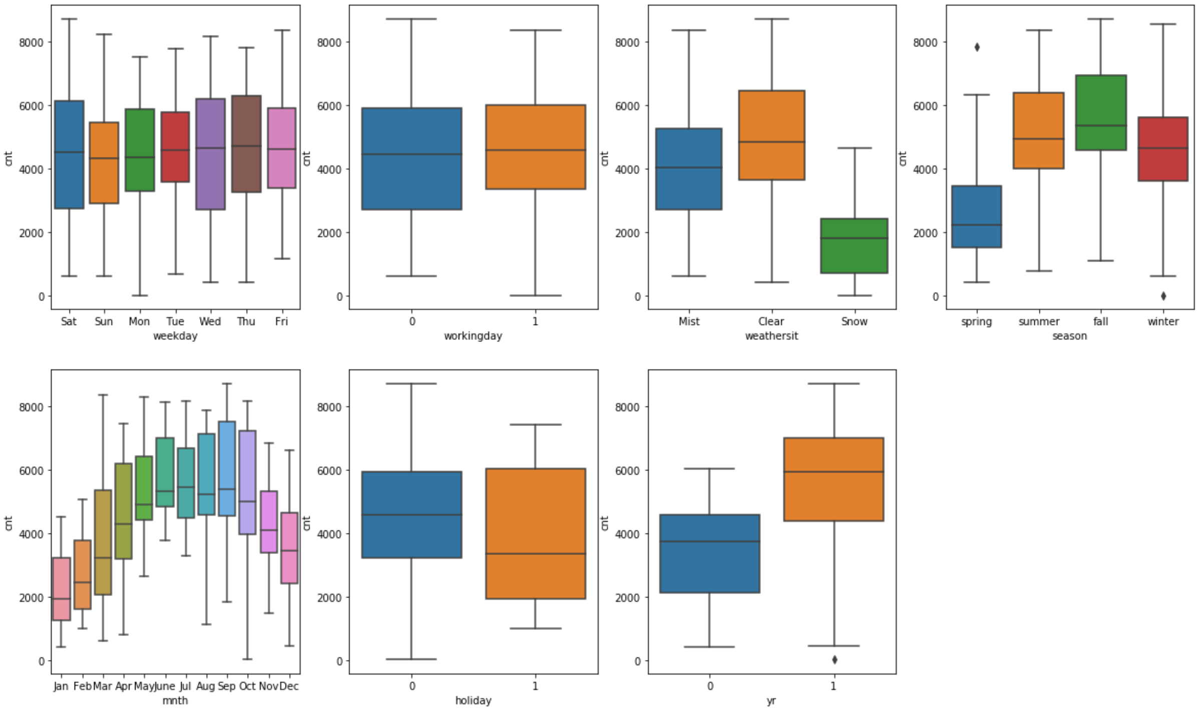
Assignment-based Subjective Questions

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

   We analysed the effect of categorical variable on dependent through box plot between them.



Weekday – Wednesday , Thursday, Saturday,Sunday have most bike bookings.

Workingday – Bike bookings are almost same on working and non working day

Weathersit – Bike booking are highest on clear weather, least in snow weather.

Season – Bike booking are highest on season fall and summer, less in winter and least in spring.

Month – Most Bookings are from may, June, July, august, September,October, November(steady increase in booking is there from may to November).Decreases towards the end of year.

Holiday – Bookings are more on holidays

Year – 2019 year has more bookings than 2018

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

It is important to use drop_first=True as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. For categorical variables with n values, we need n-1 values to represent all states. Ex if we have a categorical variable with A, B , C as 3 values

A - 1 0 0

B - 0 1 0

C – 0 0 1

This can be represented using drop_first=True , only with 2 values

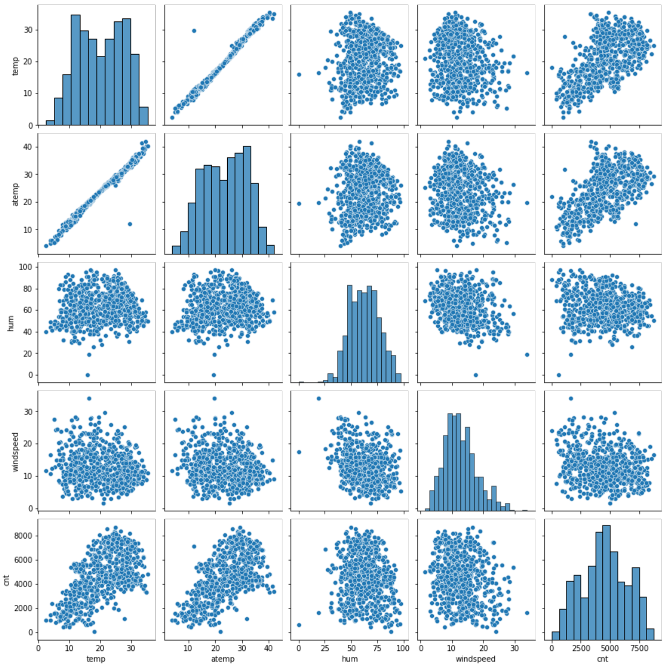
A - 1 0

B - 0 1

C - 0 0

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

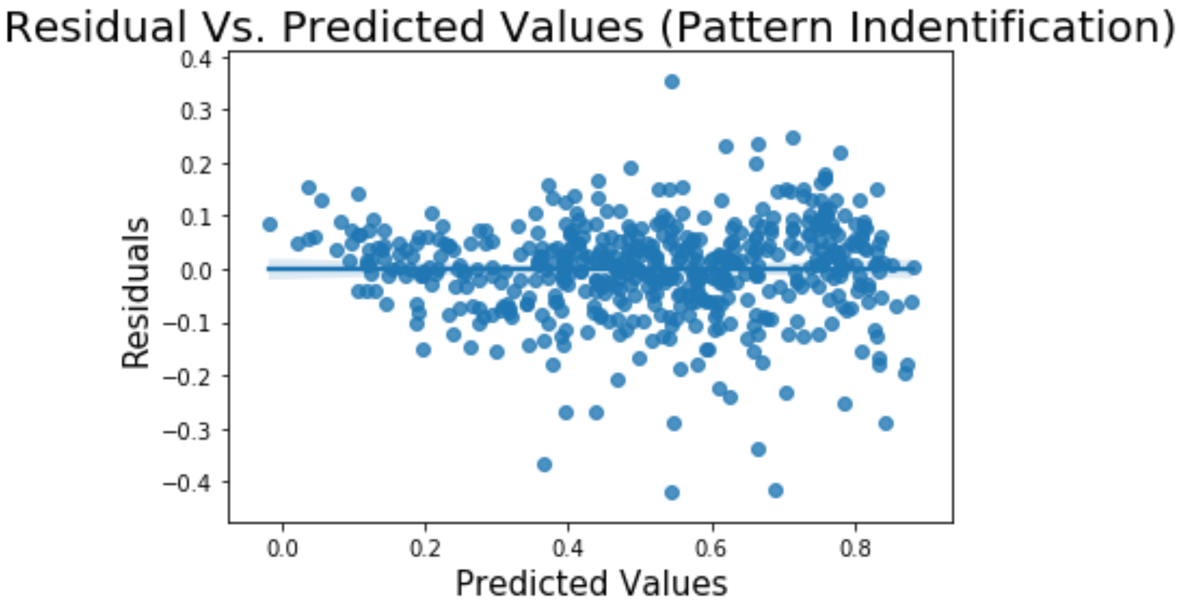
Below is the pair plot of our dataset. temp,atemp shows highest **correlation with cnt.**



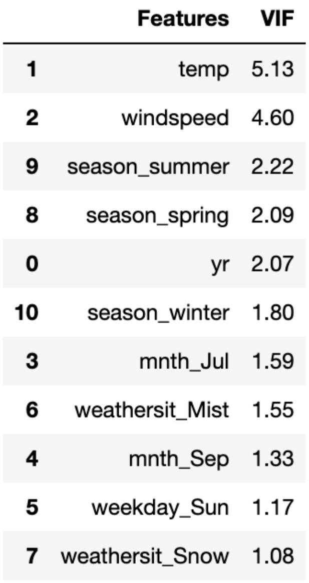
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

We have validated the assumption of linear regression as below

- **Linear Relationship**- We have found linear relationship among the variables in dataset through pairplot visualisation .
- **No Auto correlation or Independence of Error Terms** – We have plotted a graph of residual vs predicted values and found no pattern , hence error terms are independent.

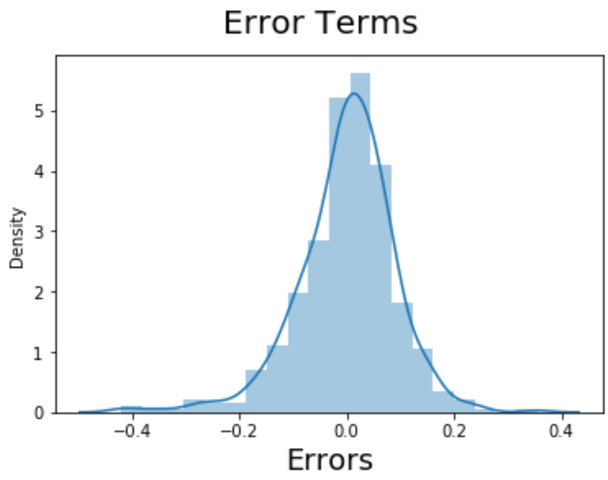


- **No Mulitcollinearity –** There is no collinearity between the independent variables in the final model.
This is verified, since our Model VIF for features is less than 10 .



	Features	VIF
1	temp	5.13
2	windspeed	4.60
9	season_summer	2.22
8	season_spring	2.09
0	yr	2.07
10	season_winter	1.80
3	mnth_Jul	1.59
6	weathersit_Mist	1.55
4	mnth_Sep	1.33
5	weekday_Sun	1.17
7	weathersit_Snow	1.08

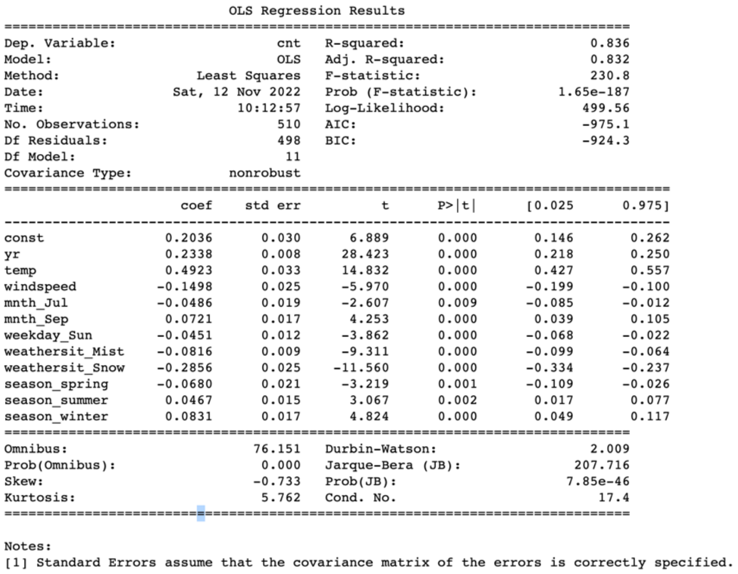
- **Homoscedasticity -** Homoscedasticity means the residuals have constant variance at every level of x. In above graph there is no pattern for residuals vs predicted values and hence confirmed.
- **Normal distribution of error terms –** Error terms are normally distributed when we plot error terms histogram.



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

Below is summary of our model and top 3 influencing factors are

- Temperature -temp(0.479),
- Weather Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds -weathersit_snow(-0.2856),
- Year - yr(0.2338).



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.836
Model:                            OLS   Adj. R-squared:                  0.832
Method:                 Least Squares   F-statistic:                     230.8
Date:                Sat, 12 Nov 2022   Prob (F-statistic):          1.65e-187
Time:                        10:12:57   Log-Likelihood:                 499.56
No. Observations:                 510   AIC:                            -975.1
Df Residuals:                     498   BIC:                            -924.3
Df Model:                          11
Covariance Type:            nonrobust
=====================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               0.2036      0.030      6.889      0.000       0.146       0.262
yr                  0.2338      0.008     28.423      0.000       0.218       0.250
temp                0.4923      0.033     14.832      0.000       0.427       0.557
windspeed          -0.1498      0.025     -5.970      0.000      -0.199      -0.100
mnth_Jul           -0.0486      0.019     -2.607      0.009      -0.085      -0.012
mnth_Sep            0.0721      0.017      4.253      0.000       0.039       0.105
weekday_Sun        -0.0451      0.012     -3.862      0.000      -0.068      -0.022
weathersit_Mist    -0.0816      0.009     -9.311      0.000      -0.099      -0.064
weathersit_Snow    -0.2856      0.025    -11.560      0.000      -0.334      -0.237
season_spring      -0.0680      0.021     -3.219      0.001      -0.109      -0.026
season_summer       0.0467      0.015      3.067      0.002       0.017       0.077
season_winter       0.0831      0.017      4.824      0.000       0.049       0.117
==============================================================================
Omnibus:                       76.151   Durbin-Watson:                   2.009
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              207.716
Skew:                          -0.733   Prob(JB):                     7.85e-46
Kurtosis:                       5.762   Cond. No.                         17.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

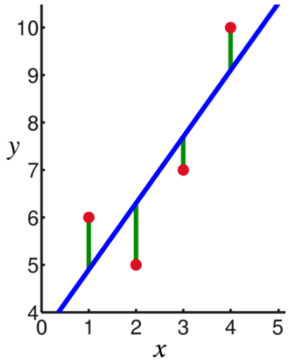
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm is a very common supervised machine learning algorithm technique. We build a predictive model through this algorithm which predicts the behaviour of data based on certain variables. It tells us the relationship between the dependent (target variable) and independent variables (predictors).

The variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression algorithm is represented by the equation.



$Y = \beta_0 + \beta_1 X$

X - independent variable

Y – dependent variable

β_0 - Intercept(where y intercepts x=0)

β_1 - slope of the line

Let us assume that we have drawn the regression line using the following set of x and y values:

(x1,y1),(x2,y2),(x3,y3)......(xn,yn)

The following formulas give the y-intercept and the slope of the equation.

$$\beta_1 = \frac{n\Sigma xy-(\Sigma x)(\Sigma y)}{n\Sigma x2-(\Sigma x)2}$$

β_1= $(\underline{\sum y-m\sum x)}/n$

There are 2 types of linear regression

1. Simple linear regression: Here target variable is dependent on only 1 independent variable.

It is represented by $Y = \beta_0 + \beta_1 X$

2. Multilinear regression: Here target variable depends on multiple independent variable.

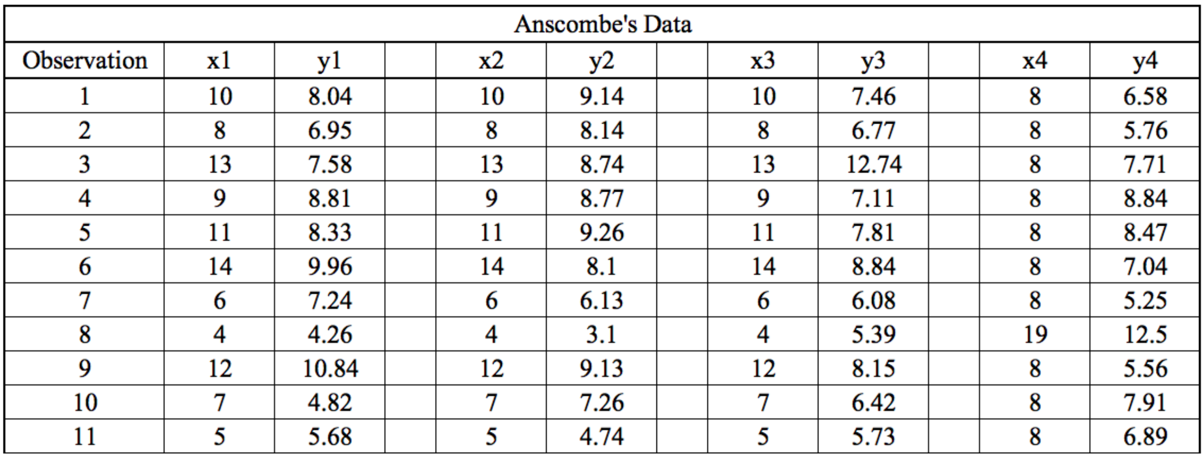
It is represented by Y=β0+β1X1+β2X2+...+βpXp+ϵ

2. Explain the Anscombe’s quartet in detail.

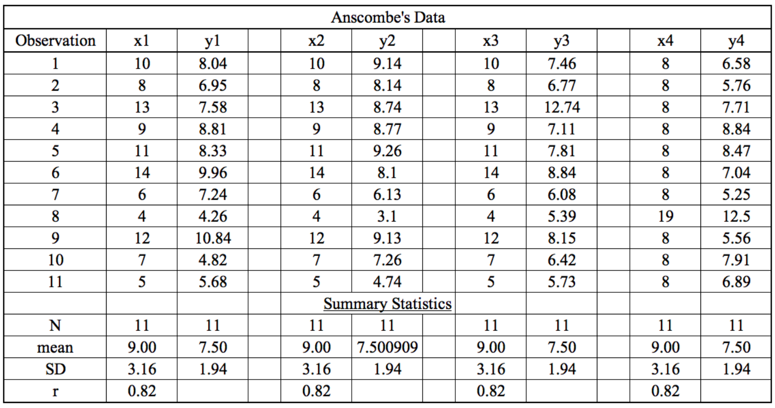
Anscombe’s quartet is a group of four datasets which are nearly identical simple statistical properties yet appear different when plotted on a graph. Each dataset consists of eleven (x,y) points.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

Dataset used is as follows



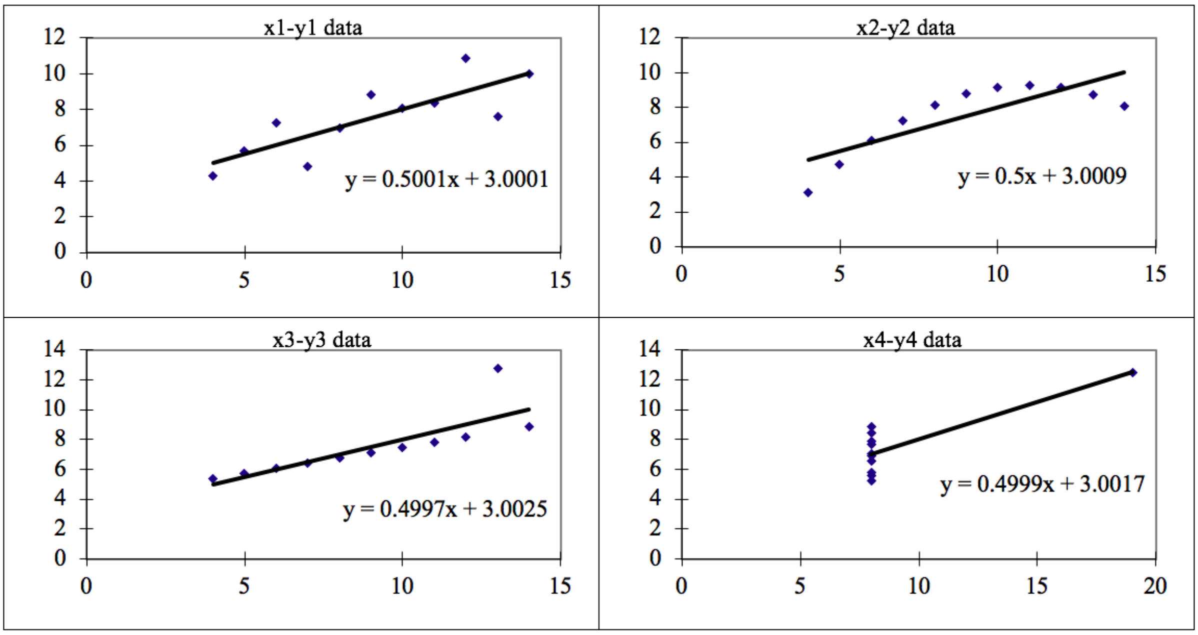
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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



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	

Above is the statistics of the data used.

When visualised on graph they look different.



- The scatter plot of first dataset (top left) we see that there seems to be a linear relationship between x and y.
- The scatter plot of second dataset (top right) we see that there is a non-linear relationship between x and y.
- The scatter plot of third dataset (bottom left) we see that 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.
- The scatter plot of fourth dataset (bottom right) we see that when one high-leverage point is enough to produce a high correlation coefficient.

The quartet is used to demonstrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship. Just statistic properties for describing realistic datasets are not sufficient to analyse the dataset.

3. What is Pearson's R?

Pearson R is the bivariate correlation, measure of linear correlation between 2 sets of data. It is the covariance of two variables, divided by the product of their standard deviations.

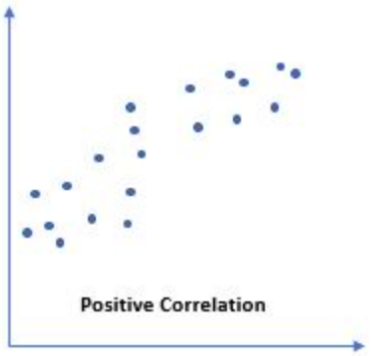
$$\mathbf{r} = \frac{\mathbf{n(\sum xy) - (\sum x)(\sum y)}}{\sqrt{\mathbf{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}}$$

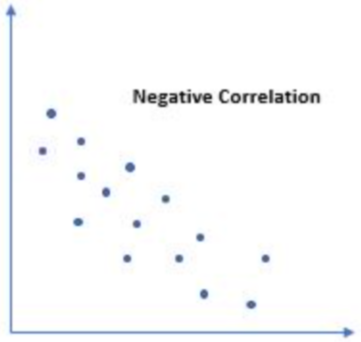
The Pearson's R always lies between -1 and 1

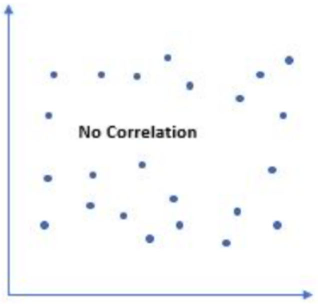
- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

- 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

We can visualise the same in plot as below







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

Scaling is a data pre-processing step applied to independent variables. This is done in order to normalize all the values of data within a certain range.

Normalised scaling

- Normalised scaling method is also known as min/maxscaling is done to get the data values in the range 0 to 1 whereas
- sklearn.preprocessing.MinMaxScaler in python is used for normalized scaling
- NormalizedScaling $x = \dfrac{x\text{-}\min(x)}{\max(x) - \min(x)}$

Standardized scaling

- Standardized scaling replaces values of data by their Z scores.
  All the values are have a standard normal distribution which has mean((µ) zero and standard deviation is 1

- sklearn.preprocessing.scale in python in used for standardisation.
- Standardized Scaling x = $\frac{x-mean(x)}{sd(x)}$

Normalization scaling has 1 disadvantage, it loses some information in the data, mostly the outliers.

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

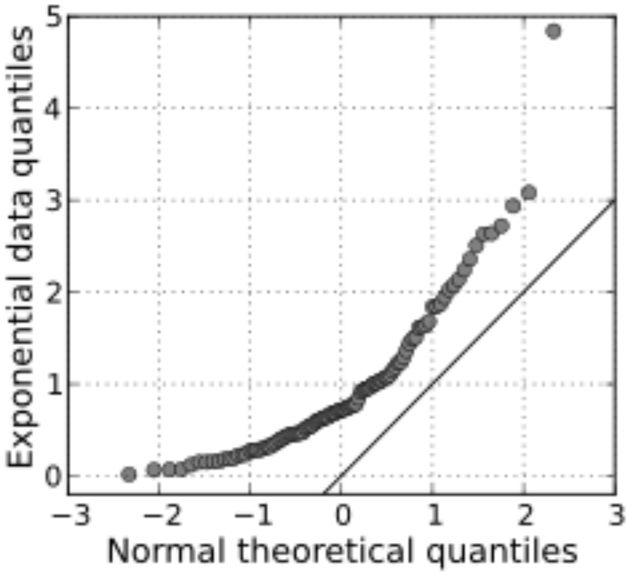
If 2 independent variables in a dataset are correlated 100 percent then we get the VIF as infinity.
The value of R2 will be 1 if 2 variables are 100 percent correlated.
VIF = (1/(1-R2)) hence 1/0 leads to infinity.
This high correlation between 2 independent variables is known as Multicollinearity .
To solve this problem we will need to drop one of the variables when 2 variables have high Multi collinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

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.

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.

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.



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.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.