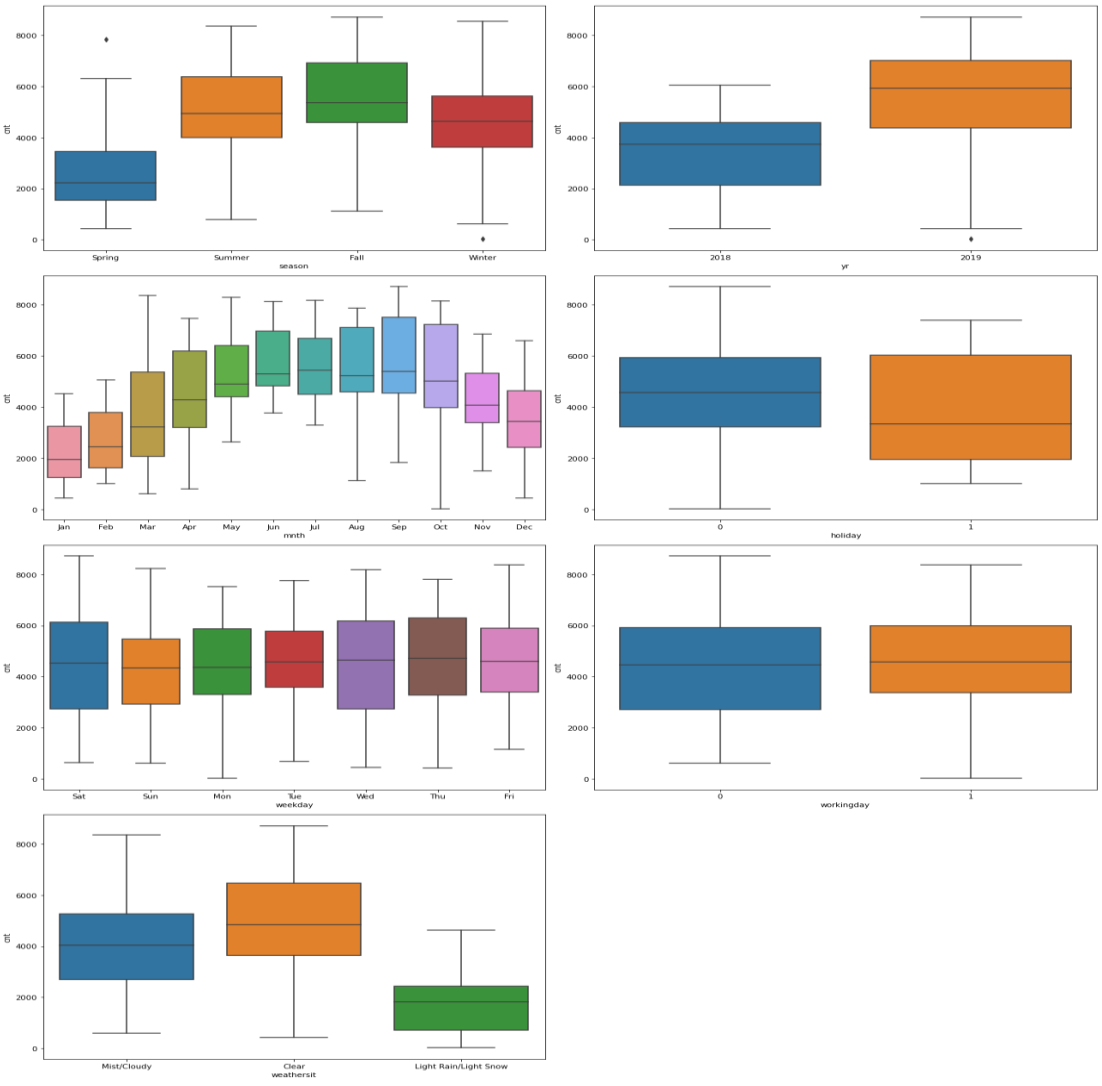
Linear Regression Subjective Questions

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? (3 marks)

Ans: Below are the analysis of the categorical variables:

1. More users in Fall/Autumn season, followed by summer and winter.
2. Spring has least number of users.
3. Demand is increasing yearly.
4. Demand is consistent between Months June to October, basically in Summer and Fall seasons (As per US).
5. Demand is dropped during Jan and Feb months, ie in Winter (As per US).
6. On an average, demand is consistent everyday. But more users are there during non-holidays and saturdays.
7. Demand is more during clear weather situations and least in Rain/Snow.



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

Ans: Dummy variable creation method created dummy variable for each of the values in the categorical variable. Drop_first=True, drops one of the dummy variables for the categorical variable, which will reduce the number of features and also the correlation between features [multicollinearity].

For example: Analyse yr by creating dummy variable.

```
In [88]: bike_pre_process.yr.value_counts()

Out[88]: 2018    365
         2019    365
         Name: yr, dtype: int64
```

```
In [96]: dummy_var = pd.get_dummies(bike_pre_process[['yr']])
         dummy_var.head(5)
```

Out[96]:

	yr_2018	yr_2019
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0

```
In [87]: dummy_var.corr()
```

Out[87]:

	yr_2018	yr_2019
yr_2018	1.0	-1.0
yr_2019	-1.0	1.0

'yr' has two values, 2018 and 2019. When dummy variable is created without drop_first=True, it gives 2 variables. The below are the combination possible:

yr_2018	yr_2019	
1	0	Indicates year = 2018
0	1	Indicates year = 2019

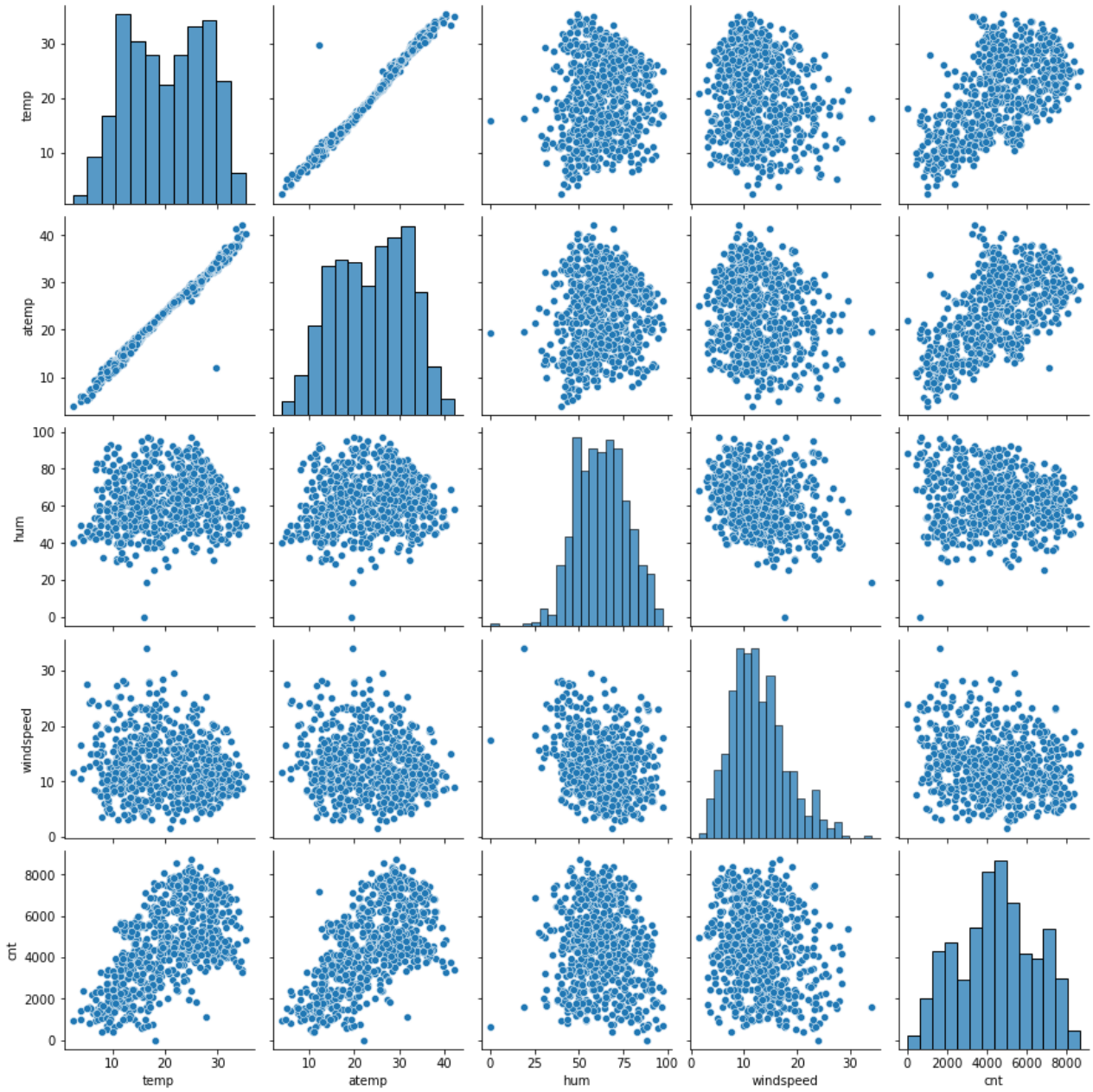
With one dummy variable also, the above information can be derived.

yr_2019	
0	Indicates year = 2018
1	Indicates year = 2019

Hence, the correlation between both the variables are high (=1 in both +ve and -ve side) as shown above in correlation matrix. Dropping one of the dummy variable is suggested. If n values are there n-1 dummy variables suffices.

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

Ans: temp and atemp has highest correlation with target variable

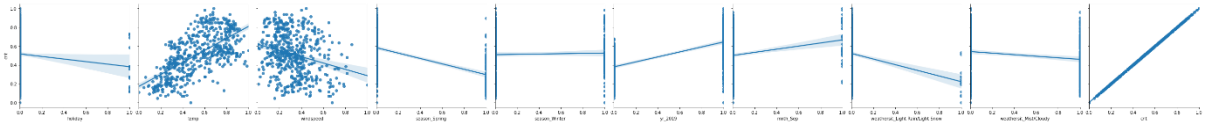


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: Below are the assumptions of Linear Regression:

1. Linear Relationship between dependent and independent variables

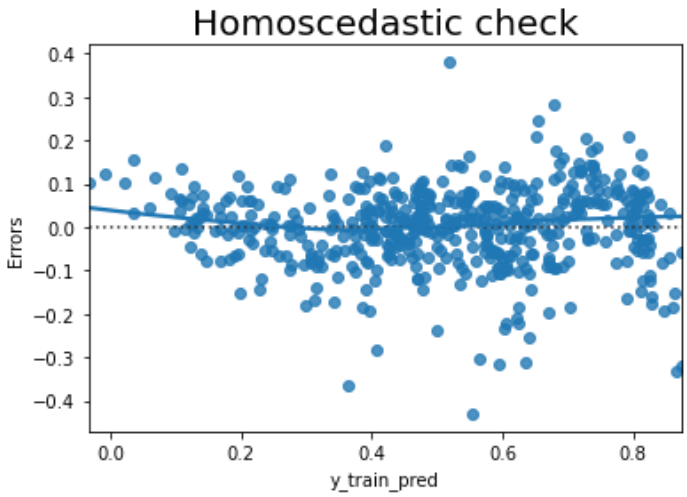
- Pairplot (scatter plot) on the selected features against the target variable, on the train set, with a regression line.



- Since categorical variables are not properly explainable, plotted pairplot for numeric variables for each category to understand if the relationship of the numerical variable changing for category

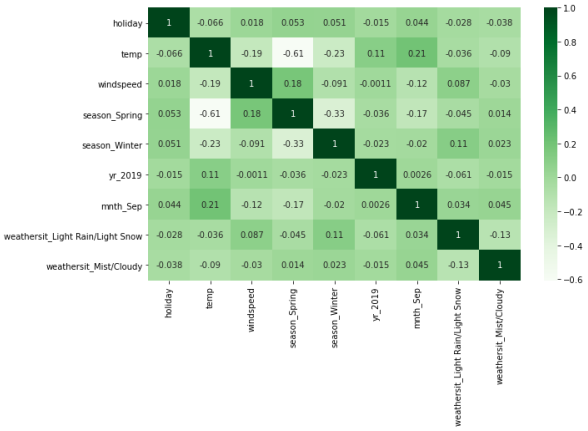
2. Homoscedasticity : The residuals/error terms have a constant variance

- Residplot on predicted y value and residual was done. Analyse the pattern of the residuals against the line at 0. The distance between the points to the line is similar for each predicted outcome.



3. Multicollinearity

- Heatmap on all selected features were done to understand the correlation between features. If the value is <0.4 (both +/-) is considered less correlation.

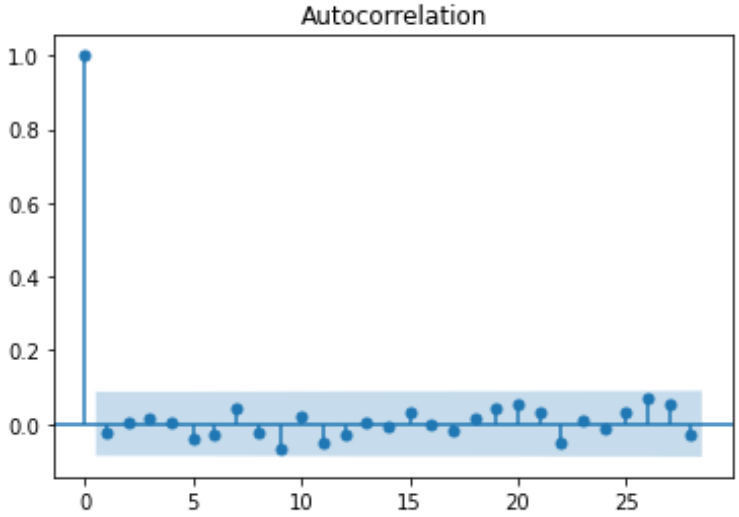


- VIF on the features also helps to identify multicollinearity. VIF <5 are considered less correlation.

Features	VIF
windspeed	3.94
temp	3.75
yr_2019	2.05
season_Spring	1.65
weathersit_Mist/Cloudy	1.50
season_Winter	1.37
mnth_Sep	1.16
weathersit_Light Rain/Light Snow	1.08
holiday	1.04

4. Independence of errors or absence of autocorrelation

- Used statsmodels acf plot to identify the correlation between residuals. The observations are within the blue box represents absence of autocorrelation.

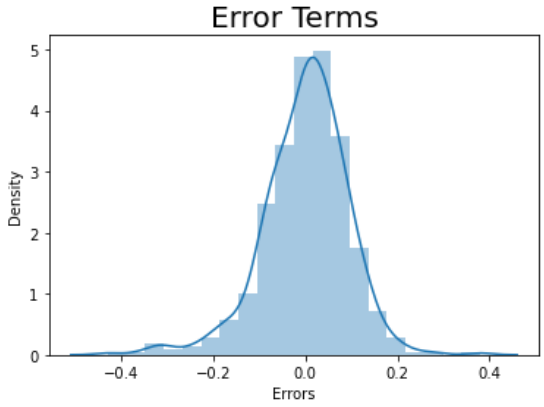


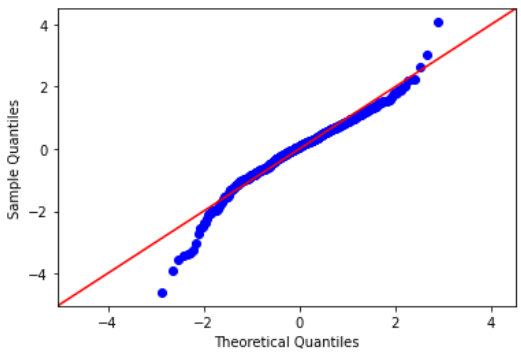
- Durbin-Watson metric in statsmodels result summary, represents autocorrelation. Value 2.0 represents no autocorrelation

```
Durbin-Watson:                   2.041
```

5. The residuals are normally distributed

- Used distplot and qqplot of residuals to identify normal distribution





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

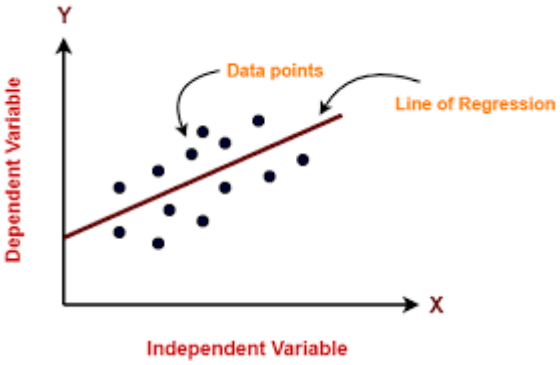
Ans: Below are the 3 top features contributing to the Shared bike demands.

- **Temperature (in Celsius)** - One unit increase in temperature will increase the Bike Share users by 0.40783 unit
- **Weather Situation: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds** - One unit increase in Light Rain/Light Snow weather situation will decrease the Bike Share users by 0.28846 unit
- **year_2019** - One unit increase in temperature will increase the Bike Share users by 0.40783 unit

General Subjective Questions

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

Ans: Linear Regression is a machine learning algorithm used to identify the relationship between the output variable and other variables. This is a regression model and is a supervised learning. Regression models to be used only when the target variable is continuous. This is a predictive model which gives the relationship between the dependent (output/target variable) and independent variables (predictors).



There are two types of Linear Regression:

1. Simple Linear Regression
   Explains the relationship between dependent variable and one independent variable.
2. Multiple Linear Regression
   Explains the relationship between dependent variable and more than one independent variable.

<u>Simple Linear Regression</u>

Simple linear regression line is defined by:

$Y = \beta_o + \beta_1 X$

Where, β_o is the intercept of the line

β_1 is the slope of the line

Y is the dependent variable and X is the independent variable.

Example : Y = bike_sharing_users , X = temperature

Suppose the equation formed is bike_sharing_users = 0.1234+0.2456*temperature

Indicates that one unit increase in temperature will increase the bike_sharing_users by 0.2456 unit.

<u>Multiple Linear Regression</u>

Multiple linear regression is defined by:

$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p$

Where β_o is the intercept

β_1 to β_p are the coefficients of the independent variables.

Y is the dependent variable and X_1 to X_p are the independent variables.

This equation indicates the impact of change in a unit of one of the independent variable, by keeping all the other independent variables constant, by β unit in dependent variable.

- ➔ Below are the assumptions of linear regression:
  - Linear relationship between dependent and independent variables
  - Homoscedastic – Constant variance between errors/residuals
  - No multi collinearity
  - No auto correlation between error terms/residuals
  - Linear distribution between error terms/residuals
- ➔ Linear regression tries to reduce the error terms between the residuals to identify the coefficients of the features. If the standard error is less, the coefficient is significant.

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

Ans: *Anscombe's Quartet* demonstrates the importance of data visualization. This was developed by the statistician *Francis Anscombe* in 1973 to signify the importance of plotting data along with analyzing it with statistical properties.

Anscombe's Quartet comprises of four data-set and each data-set consists of eleven (x,y) points. All the data sets share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph show different behaviour, irrespective of the same statistical question.

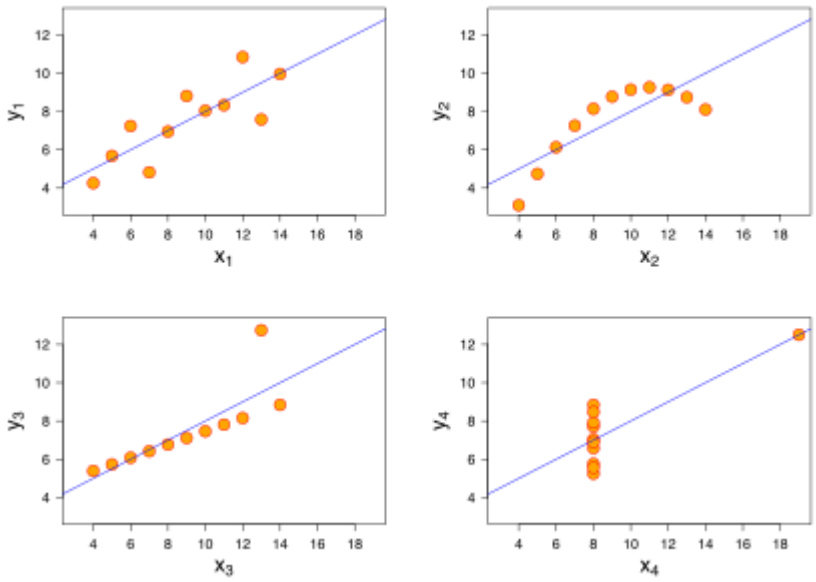
Dataset:

	I		II		III		IV	
	x	y	x	y	x	y	x	y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics of the datasets:

Mean of x values	9.0
Mean of y values	7.5
Correlation between x and y	.816
Linear equation	y1=3.0+0.5*x1
R square value	0.67

Plots of the datasets:



All of the 4 datasets are identical with the summary statistics, but the graphical representation varies.

Data set1: Shows a linear relationship between x and y, with a slight deviation

Data set2: Shows a curve shape, non linear relation.

Data set3: Shows a linear relation with an outlier

Data set4: Shows a constant relation with an outlier

“Summary statistics may mislead the model interpretation and visualisation is important.”

3. What is Pearson’s R? (3 marks)

Ans: *Pearson’s R* is a correlation coefficient, measuring the linear correlation between two variables.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

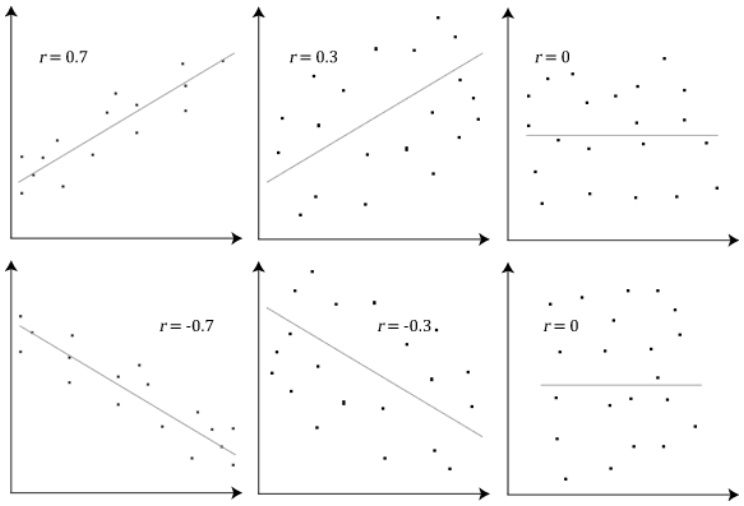
It is the covariance of x and y divided by the product of standard deviation of x and standard deviation of y.

The values range between -1 to 1.

-1 – indicates negative correlation between x and y, ie when one increases the other decreases.

1 – indicates positive correlation between x and y, ie when one increases the other also increases.

0 – indicates no correlation between x and y, ie when one increase, no effect on other



Examples of Pearson's R values and its behaviour

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

Ans: Scaling is a technique to standardise the features to a fixed range. For example: convert the features to have values between 0 to 1. Or if the features are of similar types ,for example – all features are of length, and represent in different scales ie. One in cm, another in meter, one in km etc, convert all of them to same scale, eg: to meter. When the features are of completely different magnitudes, and of different types, then there are few scaling techniques available.

Scaling is needed to analyse the various features in the same magnitude, for better feature interpretation. If not scaled the model will give inappropriate coefficients. This is also helps to adjust the step size in gradient descent method of linear regression model.

Scaling techniques:

Standardised Scaling – mean of the variables to 0 and standard deviation to 1

x = (x-mean(x))/sd(x)

Min-Max scaling /Normalised Scaling– Value of the variables will be between 0 and 1, by using min and max values of the variables

x = (x-min(x))/(max(x)-min(x))

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

Ans: VIF is a measure of collinearity between the independent variables.

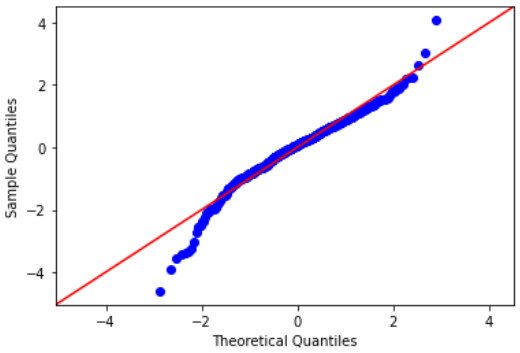
VIF = 1 indicates, that there is no correlation between the variables

VIF = high value, indicates, a high correlation between variables

VIF = infinity, indicates perfect correlation with the variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) as infinity.

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

Ans: **Q-Q or Quantile-Quantile** plot is a probability plot. It compares the probability distribution by plotting quantiles against each other.



If the distributions are linearly related, then the points in the plot will approximately lie on the line. Q-Q plot helps in linear regression to analyse the normal distribution of error terms/residuals, which is one of the assumption of linear regression.