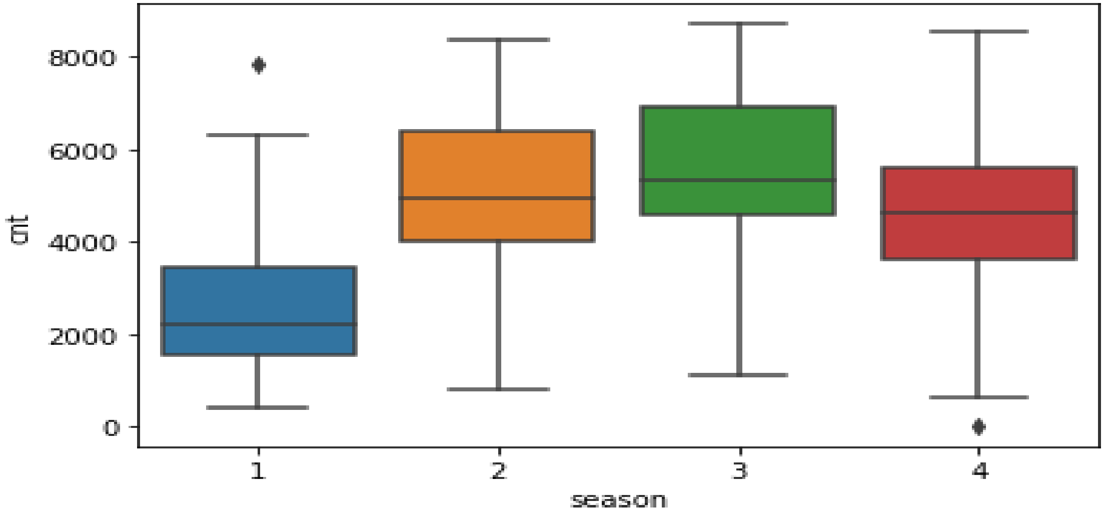
Assignment-based Subjective Questions

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

Ans. Season :-

Highest sales of shared bike is there in season 3 which is fall season.
Second Highest sales of shared bike is there in season 2 which is summer season.
Lowest sales of shared bike is there is season 1 which is spring.

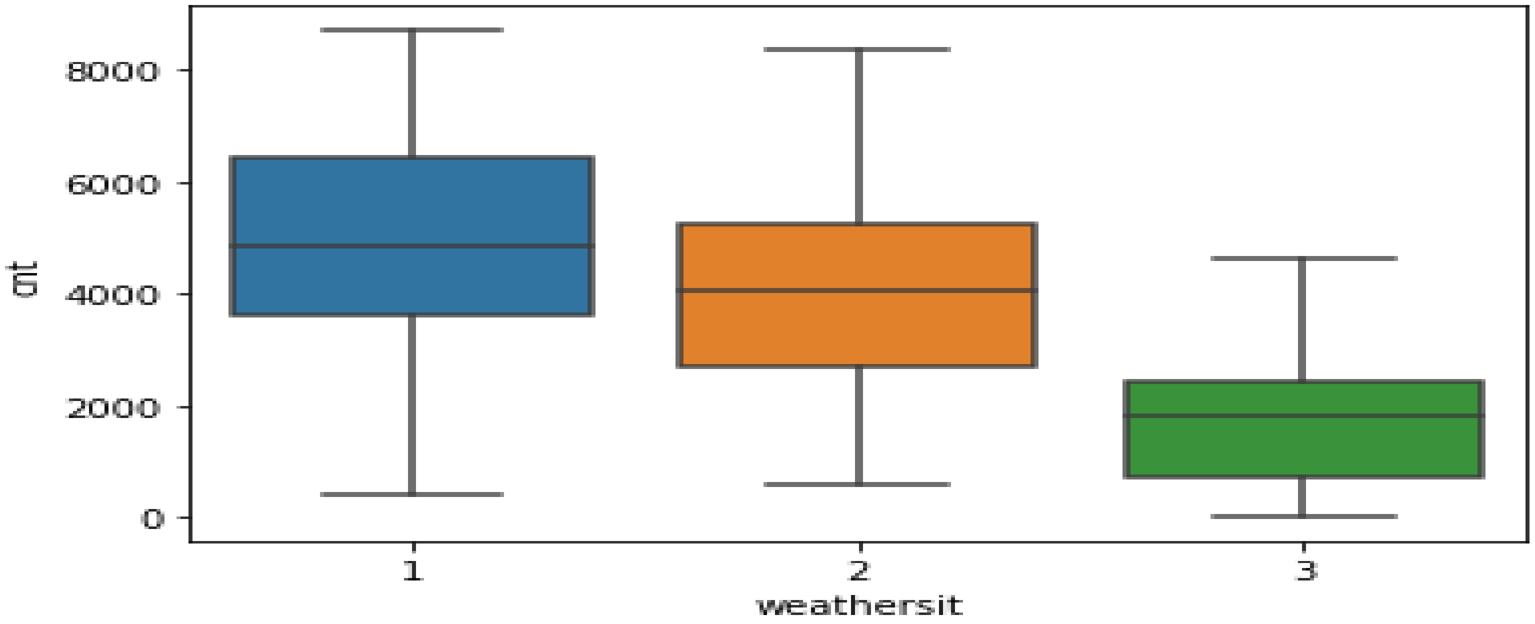


Weather sit :-

Lowest sales of shared bike is there in weather sit 3 which is Light Snow,
Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.

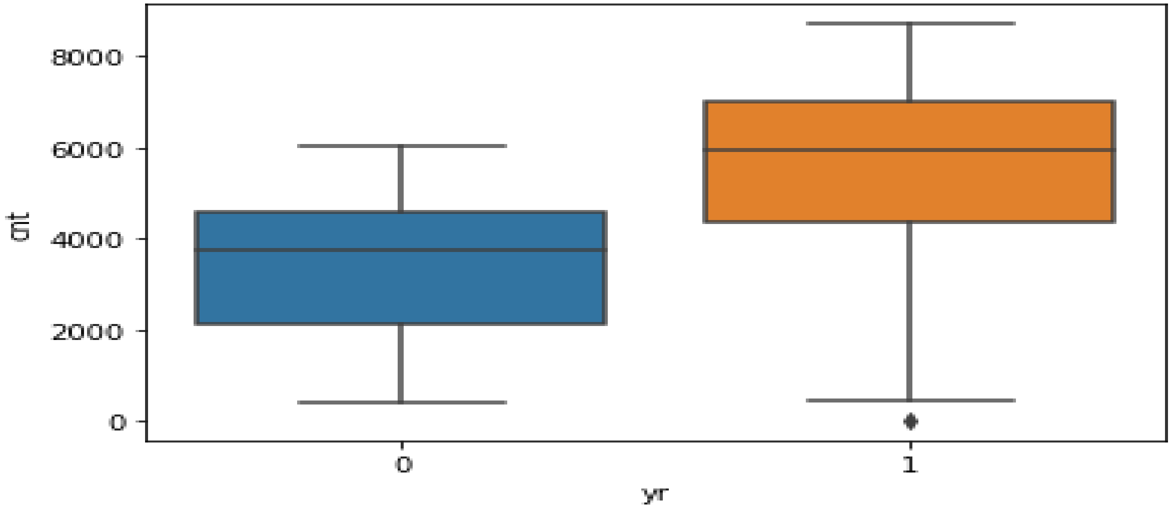
Highest sales of shared bike is there in weather sit 1 which is Clear, Few clouds, Partly cloudy,
Partly cloudy.

Medium sales of shared bike is there is weather sit 2 which is Mist + Cloudy, Mist + Broken clouds,
Mist + Few clouds, Mist.



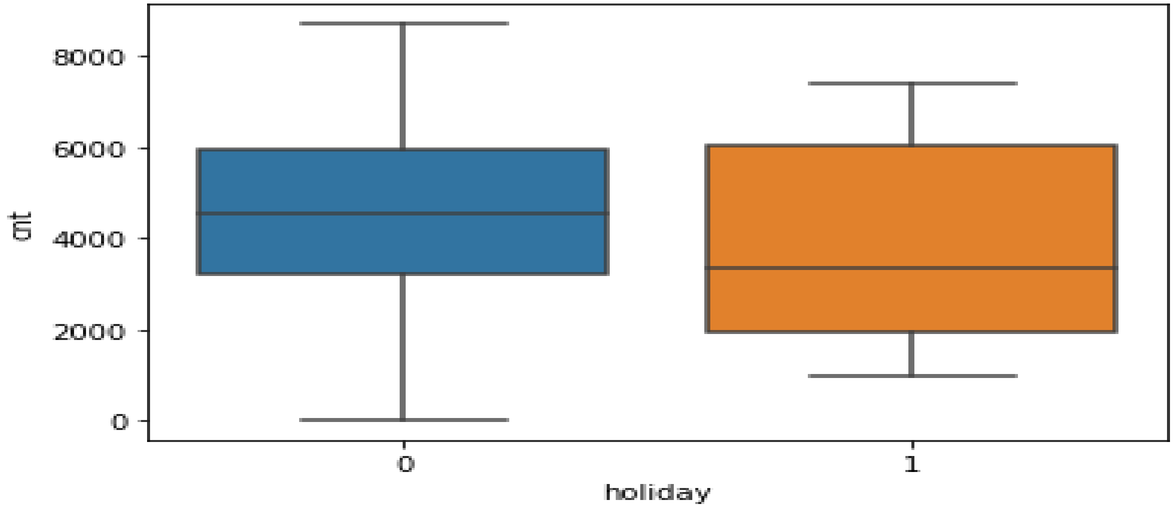
Year :-

Sales has increased from 2018 to 2019.



Holiday :-

On holiday there are less sales as comparative to working day.



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

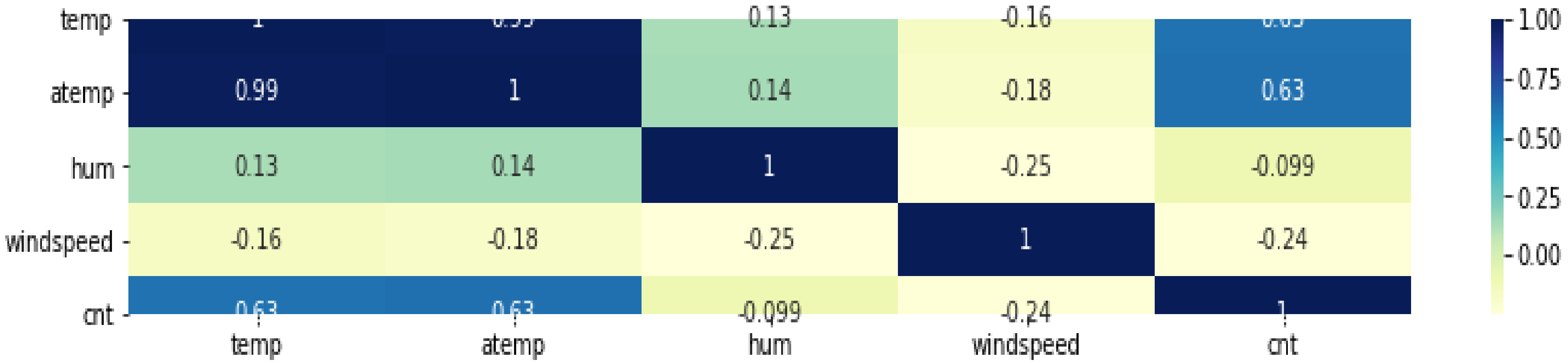
Ans. This is because of correlation. If we keep the reference column then it will be correlated to other columns. For example if we are doing dummy encoding of a variable having two categories male and female. Then male will be highly correlated with female variable. This is the main reason for using drop_first=True.

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

Ans. Temp and aTemp have the highest correlation with the target variable.

Temp :- temperature in Celsius. Correlation Value –> 0.627044

aTemp :- feeling temperature in Celsius . Correlation Value –> 0.630685



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

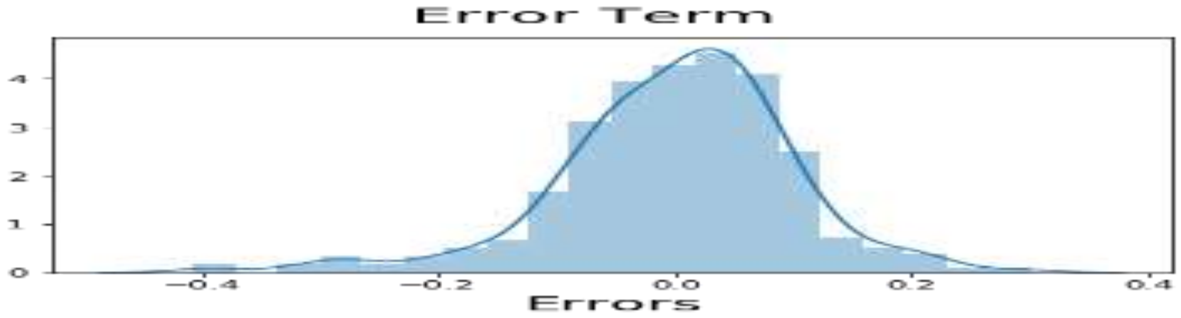
Ans. 1. Assumption of linear relationship between x and y variables.

➔ This can be validated through Durbin-Watson test.

➔ In our case this test value is 1.951 which is between 1.5 to 2.5 which is normal.

2. Error term are normally distributed.

➔Difference between predicted and actual value should be normally distributed.

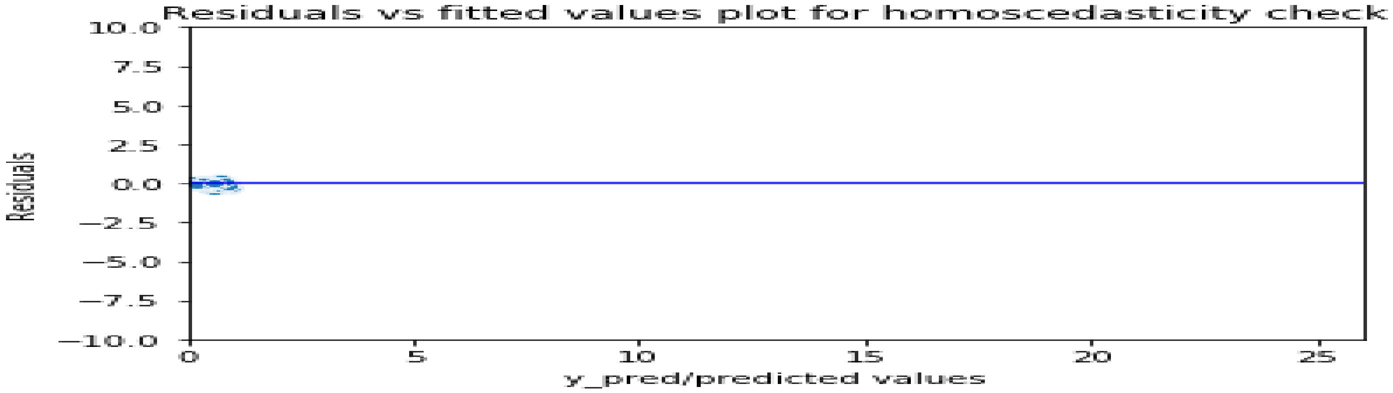


3. Mean of the residual should be almost 0.

➔ In our case mean of the residual is -7.230327447871332e-16.

4. Error term have constant variance.

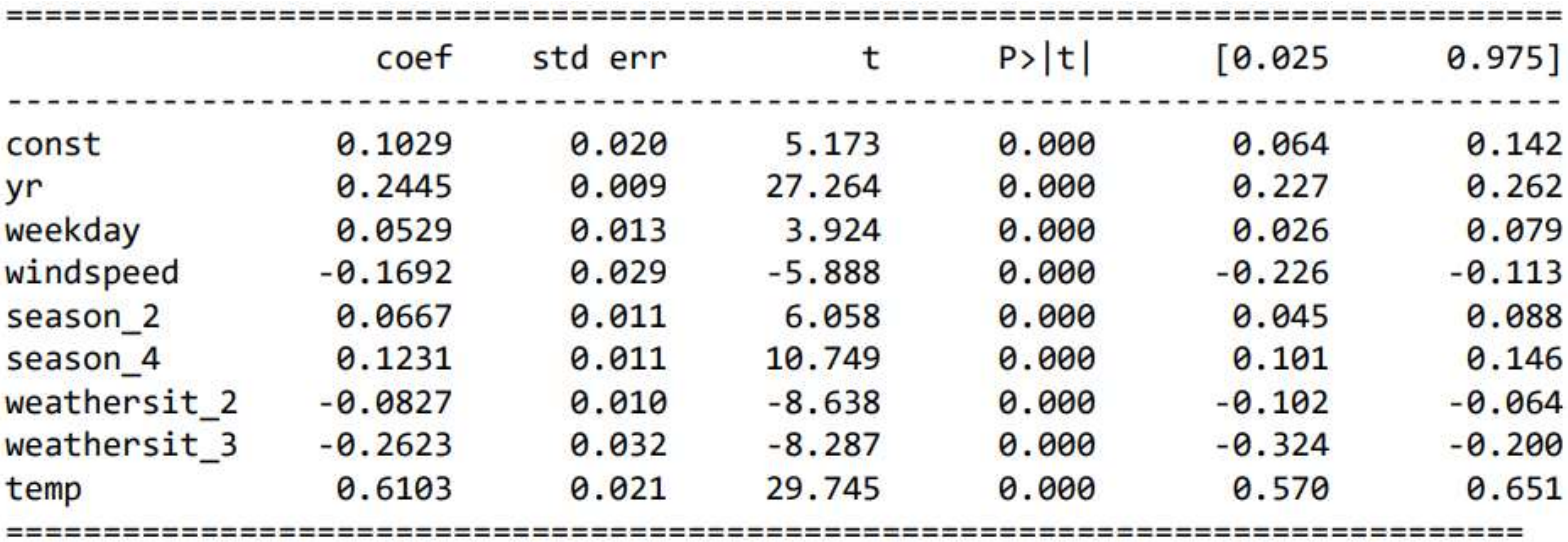
➔ There should not be any pattern in y predicted and residual values.



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

Ans. Top there highest contributing variables are :-

1. Temp :- temperature in Celsius
2. Weather sit :-
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3. Yr :- year (0: 2018, 1:2019)



```
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           0.1029      0.020      5.173      0.000       0.064       0.142
yr              0.2445      0.009     27.264      0.000       0.227       0.262
weekday         0.0529      0.013      3.924      0.000       0.026       0.079
windspeed      -0.1692      0.029     -5.888      0.000      -0.226      -0.113
season_2        0.0667      0.011      6.058      0.000       0.045       0.088
season_4        0.1231      0.011     10.749      0.000       0.101       0.146
weathersit_2   -0.0827      0.010     -8.638      0.000      -0.102      -0.064
weathersit_3   -0.2623      0.032     -8.287      0.000      -0.324      -0.200
temp            0.6103      0.021     29.745      0.000       0.570       0.651
==============================================================================
```

General Subjective Questions

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

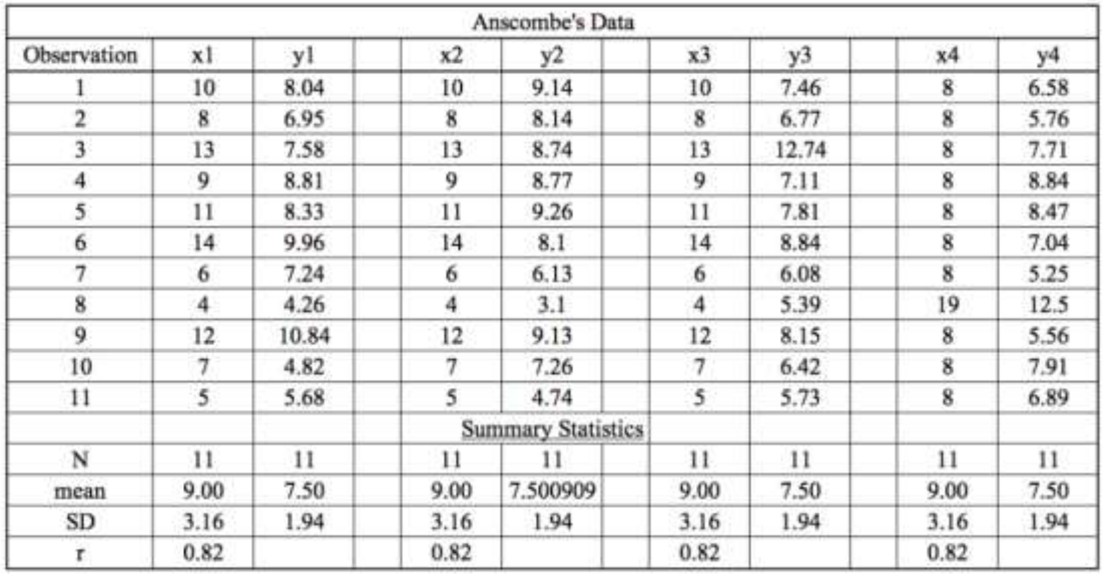
- In linear regression we are trying to predict a continuous variable with combination of continuous and categorical variables.
- The variable which we are trying to predict is called as Dependent variable and all the variable used in predicting the dependent variable are called as Independent variable.
- If we have one independent variable then Linear regression is called as simple linear regression, if we have more than one independent variable we call Linear regression as multiple linear regression.
- Linear regression tries to fit a straight line in such a way that it minimizes the error which is nothing but difference between actual and predicted value of y.
- It is a supervised machine learning algorithm.
- Equation for linear regression goes like
  - Y = B0 + B1X1 + B2X2 + ...
  - Where Y is the dependent variable, B0 is the intercept , B1 to B2 are the weights for individual independent variables, X1 to X2 are independent variables.
- R2 is the parameter which check for model performance. It is highly dependent on number of variables that why we see Adjusted R2 which penalizes the addition of variables in the model.

2. Explain the Anscombe’s quartet in detail.

Ans. Anscombe’s quartet consist of four data sets which have very similar simple descriptive statistics but yet different when graphed. It was constructed in 1973 by statistician Francis Anscombe to illustrate importance of plotting graphs before analyzing and building models.

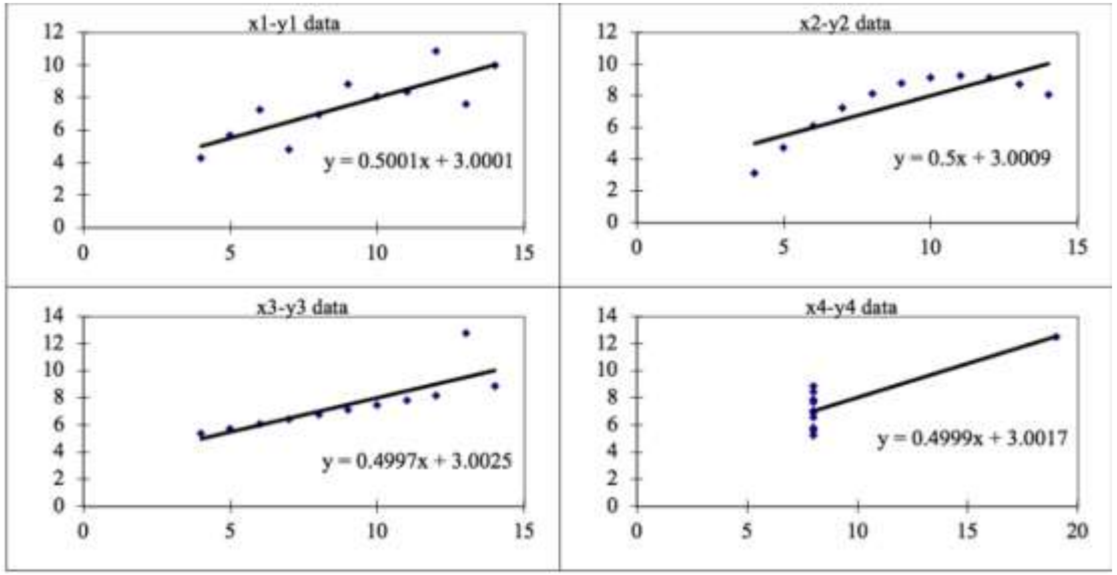
If we look at the mean and standard deviation, there are very close for all the four data set. But when we plot these data they show a complete different results.

1. First data set have a linear relationship between y and x.
2. Second data set have a non linear relationship between x and y.
3. Third data set shows the outlier involved in the dataset which cannot be handled by linear regression.
4. Forth data set have only two value for x variable showing it should be treated as categorical variable rather than continuous.



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	



3. What is Pearson’s R?
Ans. Pearson’s R also known as Pearson correlation coefficient. It measure the linear correlation between two variable. It’s formula is given by

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where cov is the covariance, sigma X is standard deviation of X and sigma Y is the standard deviation of the Y.
Its value falls between -1 to 1. Indication 1 for high positive correlation between variables and -1 for high negative correlation between variables and 0 indicating no correlation between variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans. What?
Scaling is done to normalize the data between certain range. As the values falls under certain range after scaling it helps in decreasing the time of execution of any code.
Why?
So, If we want to compare two variables we need to bring both the variables on same scale as we cant compare age with weight as both of them have different units. If we scale them we can compare the values.
Difference?
Normalize Scaling :- Also known as Min-Max Scaling. Scale the data to a value between 0 and 1. Formula for min-max scaling is x = (x-min(x))/(max(x)-min(x))
Standardized Scaling :- This replaces the value by their z-score. For this kind of scaling mean becomes 0 and standard deviation becomes 1. This can be imagine as how far you point is from mean in terms of standard deviation. Formula is x=(x-mean(x))/standard_deviation(x)

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

Ans. VIF is infinite only when there us perfect correlation between variables. Now we will explain this with example.

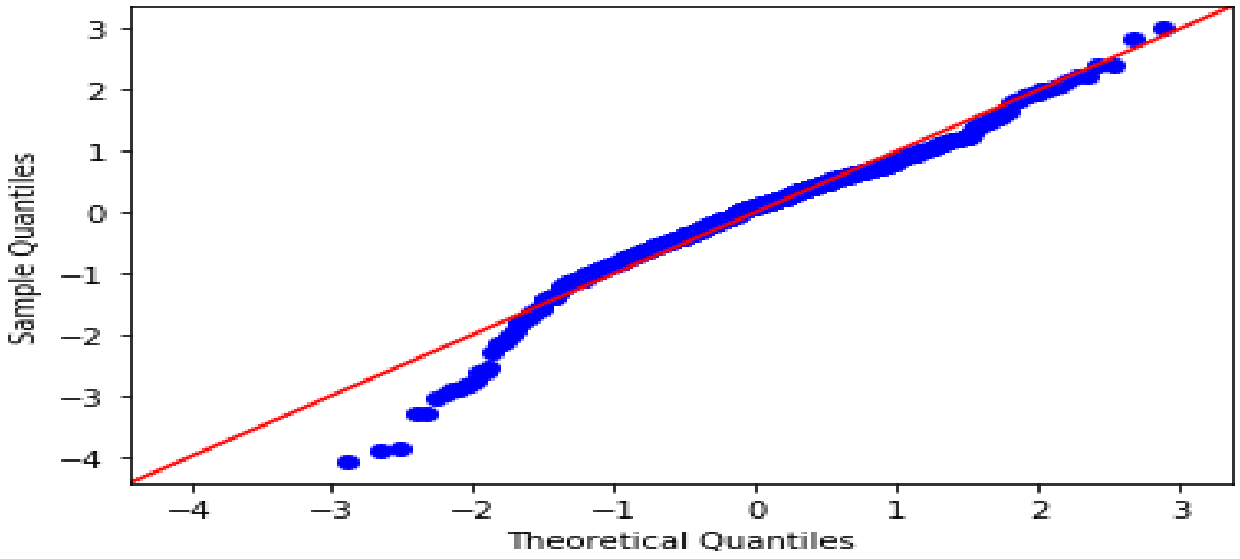
VIF formula is VIF=1/(1-R2)

Where two variables are perfectly correlated then R2 value will be one and in return denomination of VIF become zero and making VIF as infinite.

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

Ans. These are the plot of two Quantile against each other. If two distribution which are being compared are similar then the points on Q-Q plot will approx. lie on the line y=x. This can also be used two compare two distribution having different size.

In Linear regression it can be use to find whether errors are normally distributed or not. Like we have seen in our case.


Sample Quantiles
Theoretical Quantiles