Assignment-based Subjective Questions

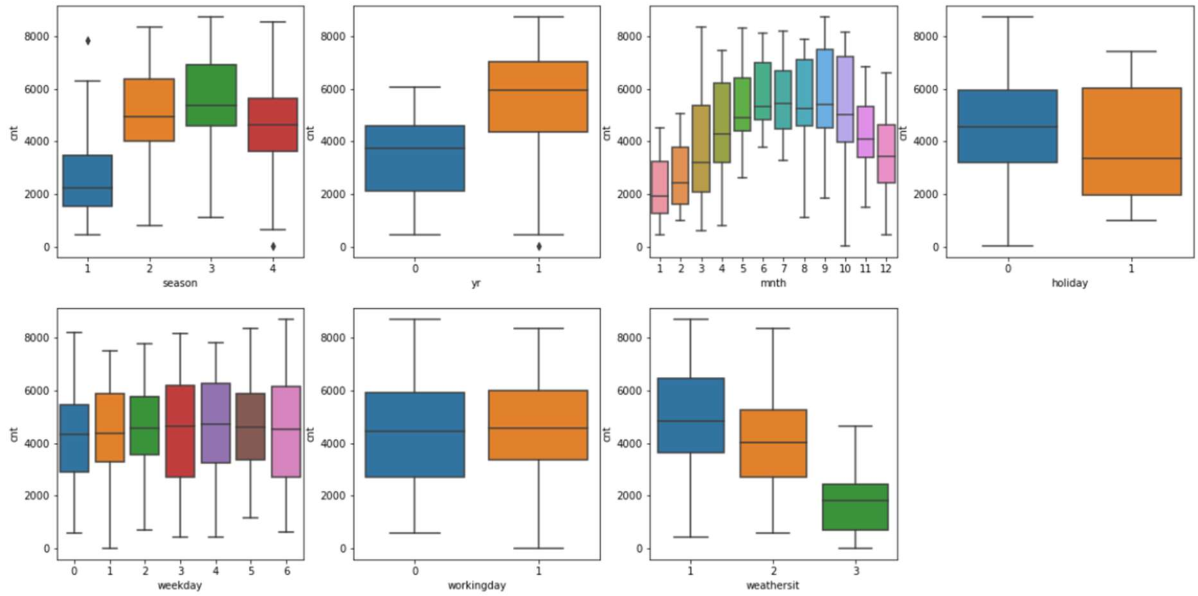
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:

Following are the categorical variables in the dataset.

Season, year, month, weekday, holiday, weathersit, workingday

Following box plots shows the effect of categorical variables vs count



Following are the inferences from the above plots:

1. Season was shown impact on bike rentals. More rentals observed during 3(fall) season and then in 2(summer) season.

2. We can observer high number of rentals in second year(2019) compared to first year(2018).

3. Bike rentals are high from 5th(May) month to 10th(October) month. May be those months are comes under fall and summer seasons.

4. Bike rentals are more on non-holidays.

5. Weather condition is impacting bike rentals. When weathre condition is in normal state(Clear, Few clouds, Partly cloudy, Partly cloudy days) people are renting the bikes more.

6. Weekdays doesn't shown much impact on rentals.

7. Working day or holiday doesn’t seem to have much effect on bike rentals.

8. According to above plots there are no much outliers in the data.

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

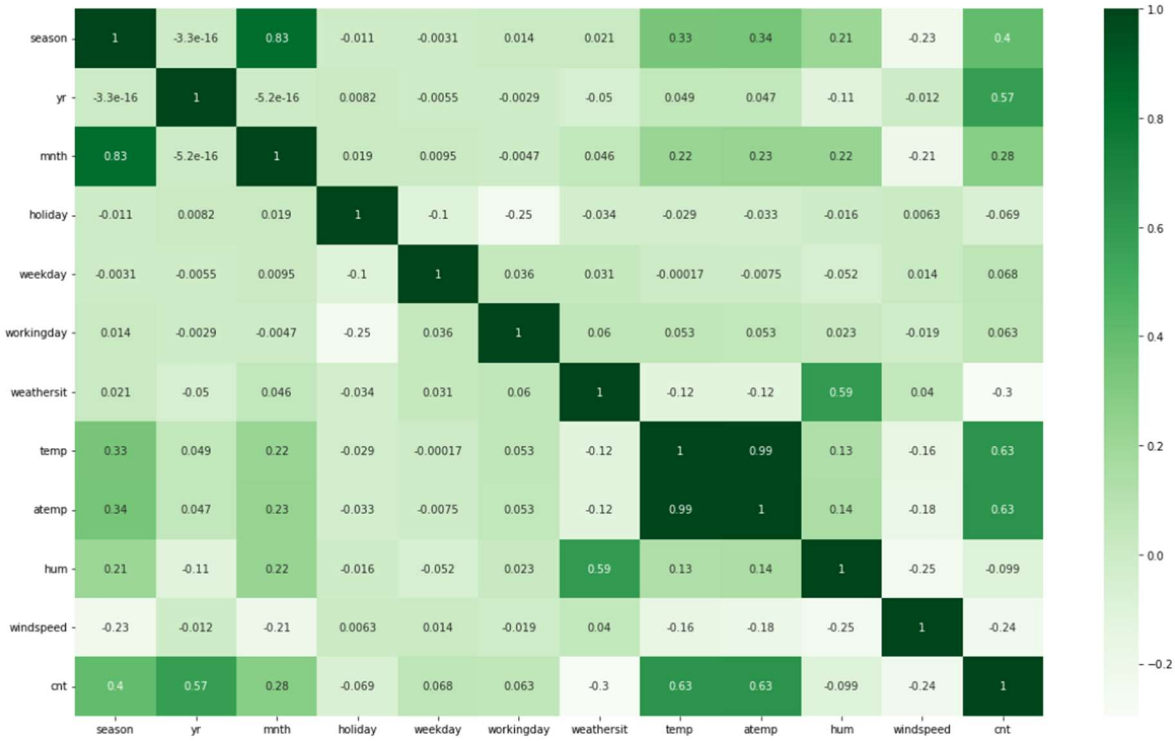
Ans:

It is important to use drop_first=True because, It helps to reduce the extra column which is created while creating dummies. Also, It reduces the correlations created among dummy variables.

By dropping the first variable we won't loose the data because if every other dummy column is 0 then this means first values would have been 1.

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

Ans:



Above plot is the correlation plot that we get from the given data.

'temp' and 'atemp' variables has very strong correlation with 'count'(target variables).

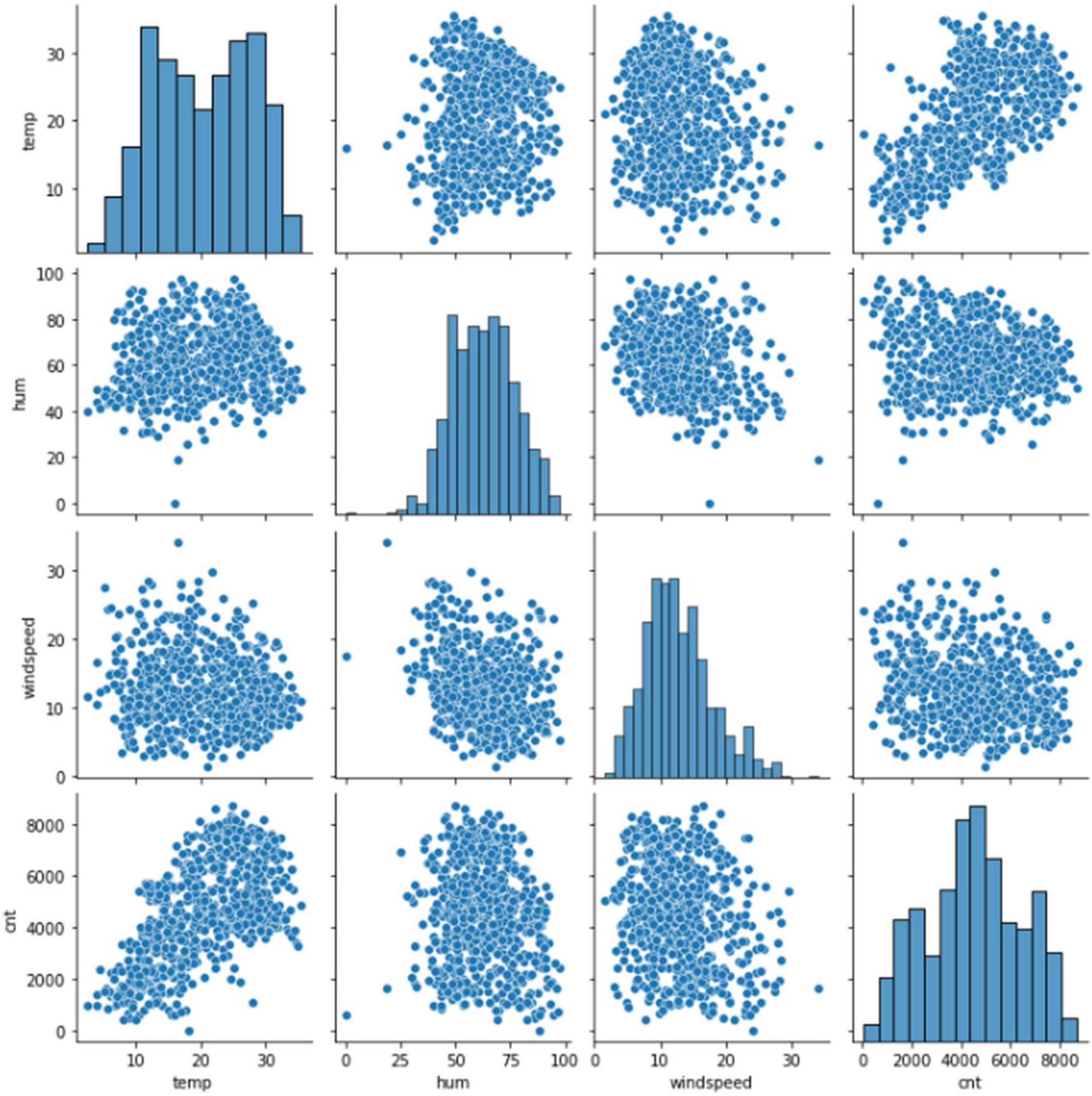
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

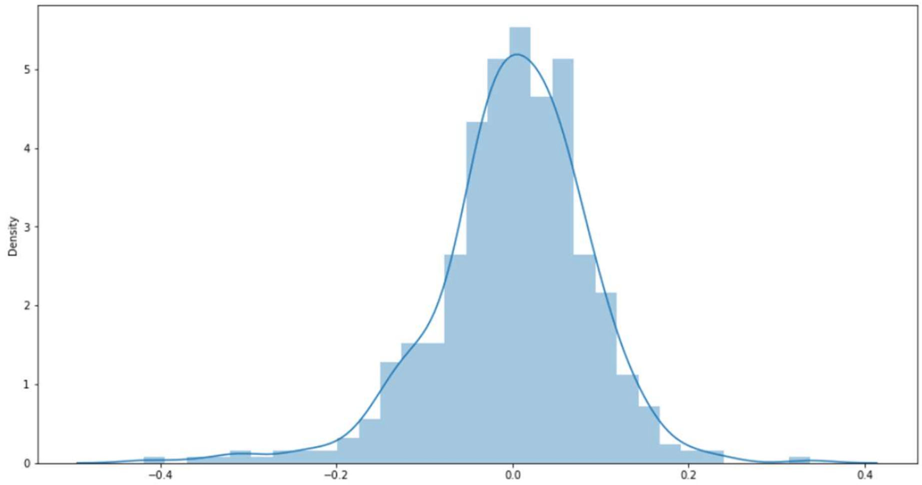
There are five assumptions of the linear regression

1. Linear relationship
2. Residuals must be normally distributed
3. No Perfect Multicollinearity
4. No Heteroskedasticity
5. No auto correlation

Linear relationship: Linear relationship can be determined by scatter plots. From the following plot we can observe linear relationship between 'temp' and 'cnt'(predictor variable) variables. Also, there is a similar variable 'atemp' has linear relationship with 'cnt'.

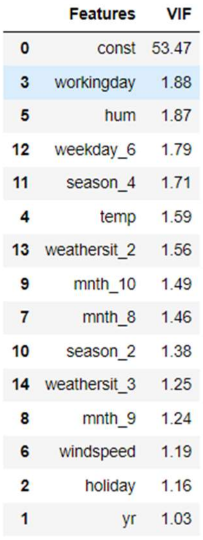


Residuals must be normally distributed: From the below plot we can observe the normal distribution on error terms



No Perfect Multicollinearity: Multicollinearity generally occurs when there are high correlations between two or more predictor variables. This can be checked by using correlation coefficients or by using VIF values of predictor variables.

In following image, we can see the VIFs of our final model. VIF of all final predictor variables is less than 5, so we can say that there is very less or moderate multicollinearity between variables.

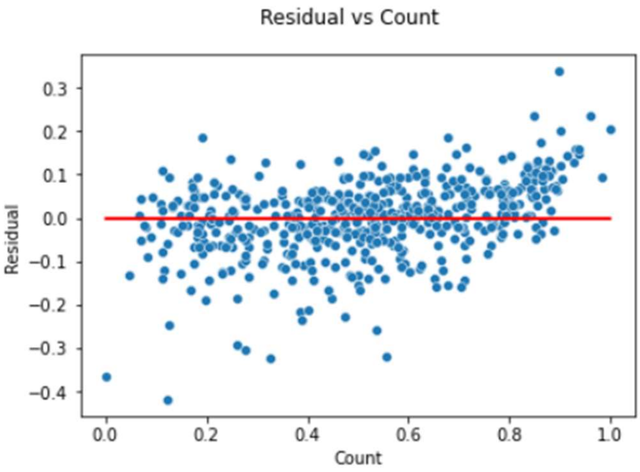


	Features	VIF
0	const	53.47
3	workingday	1.88
5	hum	1.87
12	weekday_6	1.79
11	season_4	1.71
4	temp	1.59
13	weathersit_2	1.56
9	mnth_10	1.49
7	mnth_8	1.46
10	season_2	1.38
14	weathersit_3	1.25
8	mnth_9	1.24
6	windspeed	1.19
2	holiday	1.16
1	yr	1.03

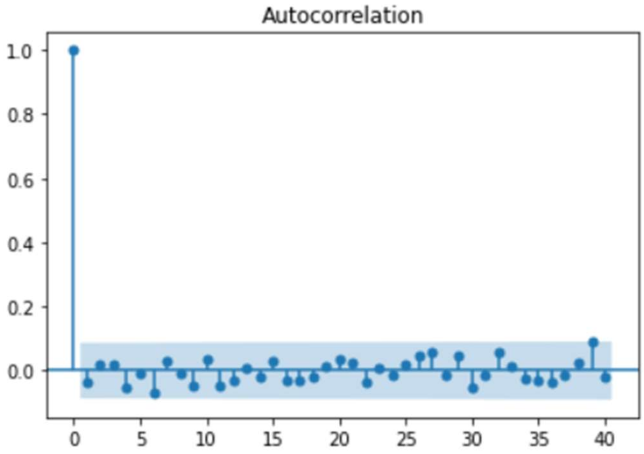
No Heteroskedasticity: Which means that the error is constant along the

values of the dependent variable. We can clearly observe that there is a

constant deviation from the zero line in the following plot. Hence we can conclude our assumption of Homoscandasticity valid true.



No auto correlation: Auto-correlation will check whether there are any patterns between the errors or not i.e error depending on y value or previous error values or not. Since there are no much error components crossing the confidence So, we can say that there is no pattern in the error and hence No auto-correlation of 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)

Ans:

Following are the top three variables that are contributing for the demand of shared bikes.

temp(temparature) with coefficient **0.5309**

yr(Year) with coefficient **0.2292**

weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) with coefficient **-0.2470**

General Subjective Questions

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

Ans:

Linear Regression Algorithm is a machine learning algorithm based on supervised learning and it is part of regression analysis. Linear regression finds the best linear-fit relationship between independent and dependent variables on a given data. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

Assumptions in a Linear Regression:

1. Linear relationship
2. Multivariate normality
3. No or little multicollinearity
4. No auto-correlation
5. Homoscedasticity

There are two types of Linear Regression models:

1. **Simple Linear Regression:** In simple linear regression, we aim to reveal the relationship between a single independent variable, or you can say input, and a corresponding dependent variable or output.

$$y = \beta 0 + \beta 1x + \varepsilon$$

2. **Multiple Linear Regression:** In this type of linear regression, we always attempt to discover the relationship between two or more independent variables or inputs and the corresponding dependent variable or output and the independent variables can be either continuous or categorical.

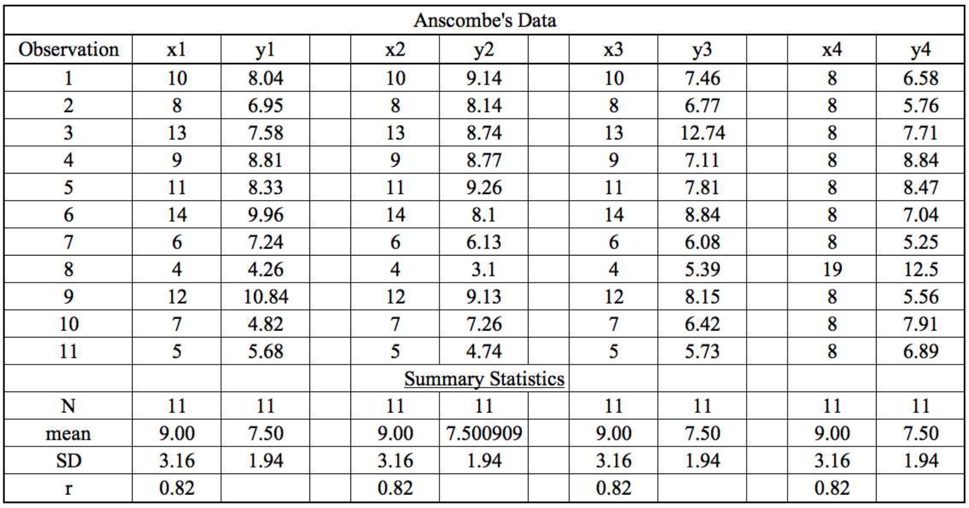
$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p \text{ ,}$$

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

Ans:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Below are the four datasets and its summary

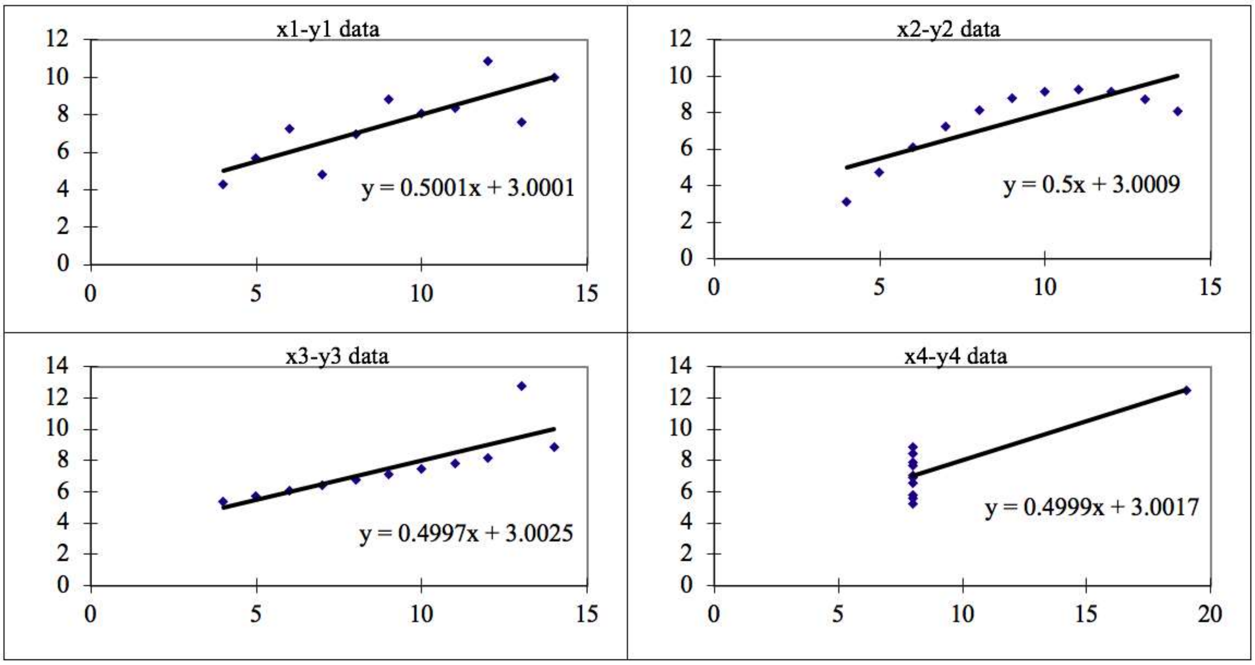


Anscombe's Data

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	

The statistical information for all these four datasets is approximately similar.

But when these datasets are plotted on a scatter plot, they generate different kind of plots which cannot be interpreted by any regression algorithms.



The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: This couldn't able to fit the linear regression model on data because data is non-linear.

Dataset 3: Shows few outliers involved in the dataset which are not able to handle by linear regression model.

Dataset 4: Shows the outliers involved in the dataset which are not able to handle by linear regression model.

Conclusion:

Above four datasets describe the importance of data visualisation and how any regression algorithm can be fooled by the same. So, It is important to visualise features in the dataset before implementing machine learning algorithm.

Source: https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

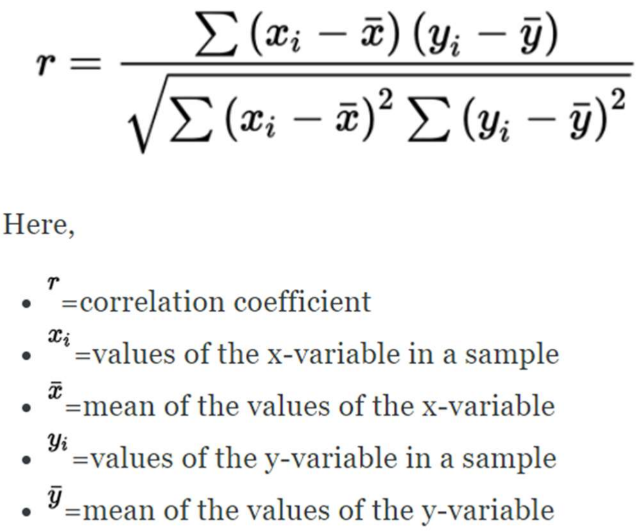
3. What is Pearson's R? (3 marks)

Ans:

In statistics, the Pearson's r, also referred to as, Pearson correlation coefficient (PCC), the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation.

It is a measure of linear correlation between two datasets. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Formula:

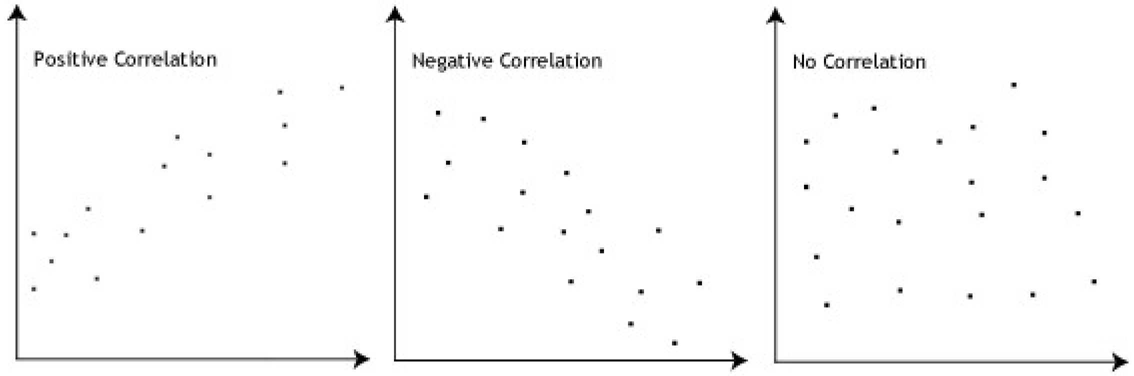


$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

The figure below shows some data sets and their correlation coefficients.



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

Ans:

It is part of date Pre-Processing which is applied to independent variables to normalize particular data with in a particular range. It helps to speed up the calculations in algorithm.

Most of the times, data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units, we may end up with incorrect modelling. So, we have to do scaling to bring all the variables to the same range.

Scaling doesn’t effect the parameters like t-statistic, F-statistic, p-values, R-squared, etc. It only affects the coefficients.

Majorly, we use two types of scaling techniques

1. Normalization/Min-Max Scaling
2. Standardization Scaling

Normalization/Min-Max Scaling: This brings the data into the 0 and 1 range.

Following is the formula for min-max scaling:

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling: Standardization replaces the values by their Z scores. Data is scaled into a range that consists mean(μ) zero and standard deviation(σ) to one.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans:

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

$$\text{VIF} = \frac{1}{1-R^2}$$

Perfect correlation between independent variables is ‘VIF = Infinity’

In such case we get R sq = 1, which lead to 1/(1-R sq) = infinity

To solve this problem 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. (3 marks)

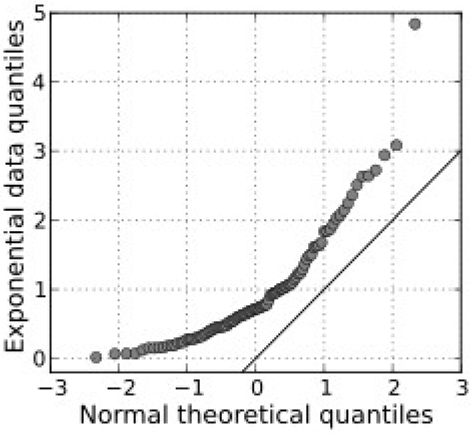
Ans:

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

This helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, and exponential.

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 quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. 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.