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

   Answer – I have analysed the categorical columns using box plot, below are my observation

   - The bike demand increases during summer and in clear weather condition and subsequently decrease in winter
   - The demand is high during May, June, July and September
   - There seems to be no change in the demand pattern between working day or not working day.
   - The demand surged in the year of 2019 as compared to the year 2018.

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

   Answer- drop_first is used to drop a duplicate after creation of dummy variable for a categorical value type column. The drop_first reduces the correlations among the dummy variables.

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

   Answer: temp or actual_temp has the highest correlation with the target variable

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

   Answer- I have validated the assumptions through below point

   a. Normality of the errors – Errors terms should be normally distributed
   b. Multicollinearity check – There should insignificant multicollinearity among the variable
   c. Linear relationship – Linear relationship should be visible among the target and predictor variable
   d. Independence of residual – No auto correlation to be seen.

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

   Answer- Top 3 contributing features

   a. Temperature
   b. Clear weather
   c. September month

1) Explain the linear regression algorithm in detail.

Answer – Linear regression can be interpreted as a statistical model that analyses the relationship a variable(target variable) and independent variable (predictor variables). Linear relationship signifies that the change in the value of an independent variable will result in change of value in dependent variable.

The mathematical representation of the equation → y=mx+c

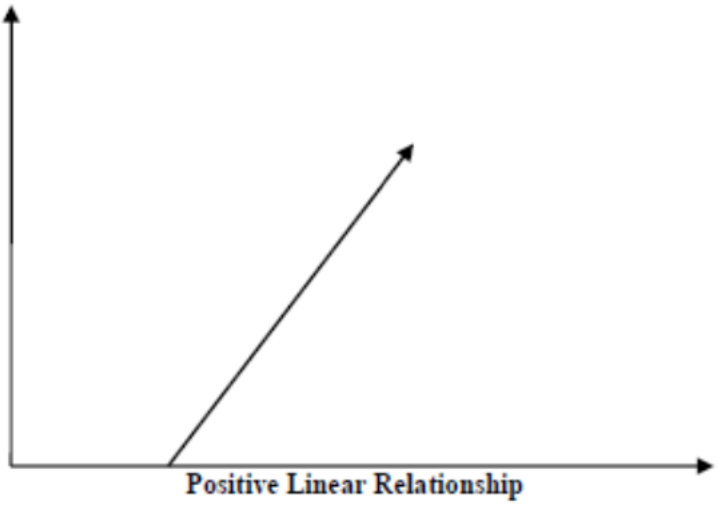
Y -> dependent variable , target variable

X-> independent variable, also known as predicator variable

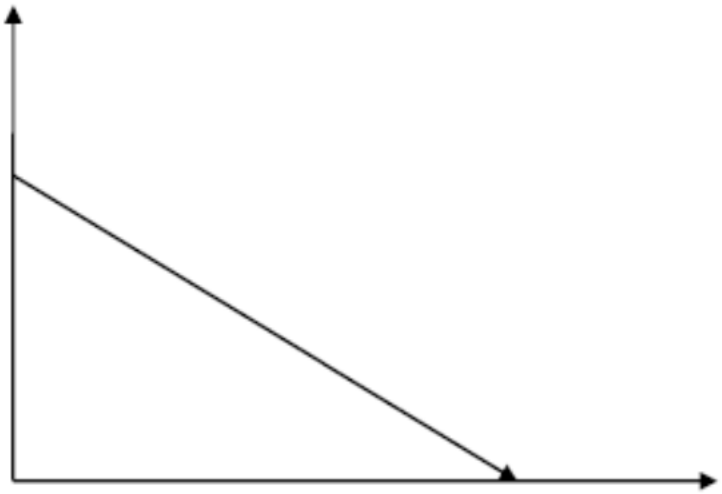
m-> coefficient represent the quantum of change in the value of Y on 1 unit change in the value of c

c-> intercept

1) when value of m>0 the value of y increase with the increase in the value of x



2) when value of m<0 the value of y decreases with the increase in the value of x



Negative Linear Relationship

1. Explain the Anscombe's quartet in detail.

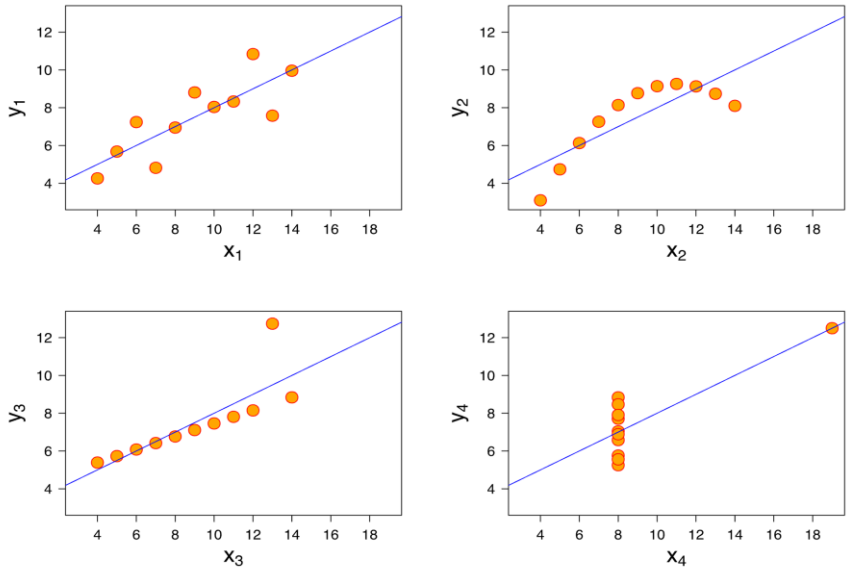
   Answer – Anscombe quartet was developed by statistician Francis Anscombe in the year 1973. It emphasizes that while analysing the data graphical analysis is as important as numerical analysis. The study comprises of 4 dataset of 11 datapoints (x,y) in each set. These datasets look similar when analysed through summary statistics but the perception changes completed when these datasets are plotted a graph.

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50

12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

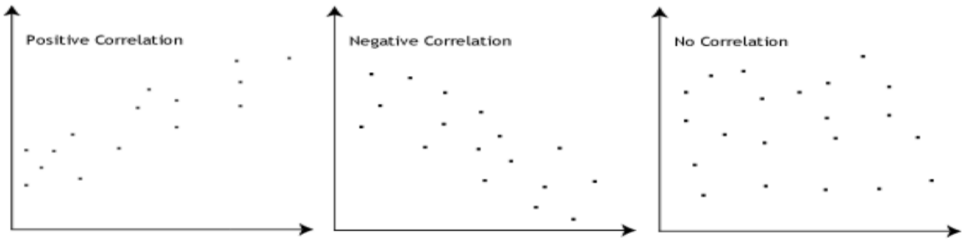


The graphical representation of all datasets are totally different.

2. What is Pearson's R?

Answer – Pearson's R is numerical value represents the strength of linear relationship between two variables. The values ranges between -1 to 1.

The value 0 confirms there is no relationship between the two given variables. A value less than zero denotes a negative association .i.e. if the value of one variable increases other variable value will decrease. Pearson's R value greater than confirms a positive association between two variable .i.e. if one increases the other variable value will also increases.



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

   Answer- Feature scaling is performed to fix the range of independent variable values. The feature scaling is done as part of pre processing of data. The feature is important as it normalizes the coefficients of the linear equation and makes it readable.

   There are two kinds of scaling generally used to scale the features

   Normalized Scaling – Minimum and maximum value of the features are used for scaling, scaled values lies between 0to1 or -1to1. The normalized scaling gets impacted by the outliers.

   Standardized Scaling – This kind of scaling is used when the ask is to ensure zero mean and unit standard deviation. Standardized scaling is not bounded by a range and has lesser impact of outliers.

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

   Answer- Infinite value of VIF indicates high degree of multicollinearity among variables. This can be also a case of overtrained model .i.e. perfect R-squared (R2) value of 1. And VIF is 1/1-R2 → infinite. To overcome the problem of infinite VIF value drop variables from the datasets causing the perfect multicollinearity .

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

   Answer- 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.