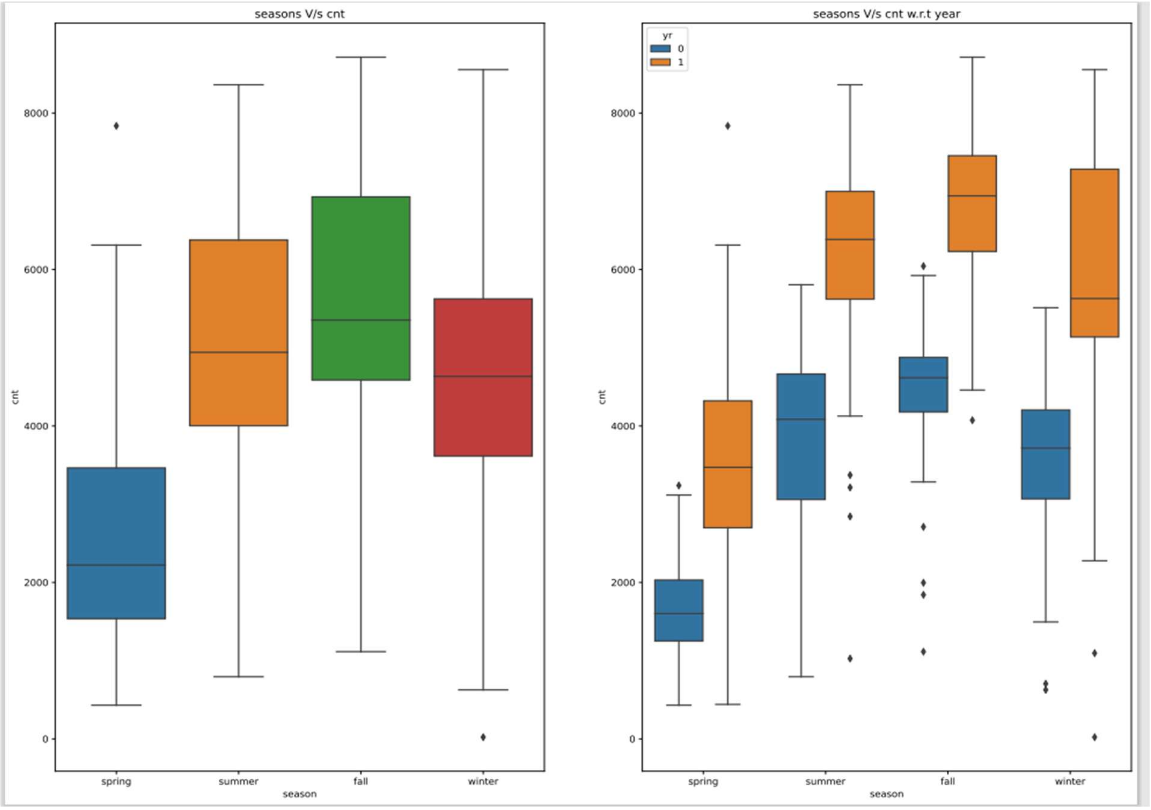
Assignment-based Subjective Questions

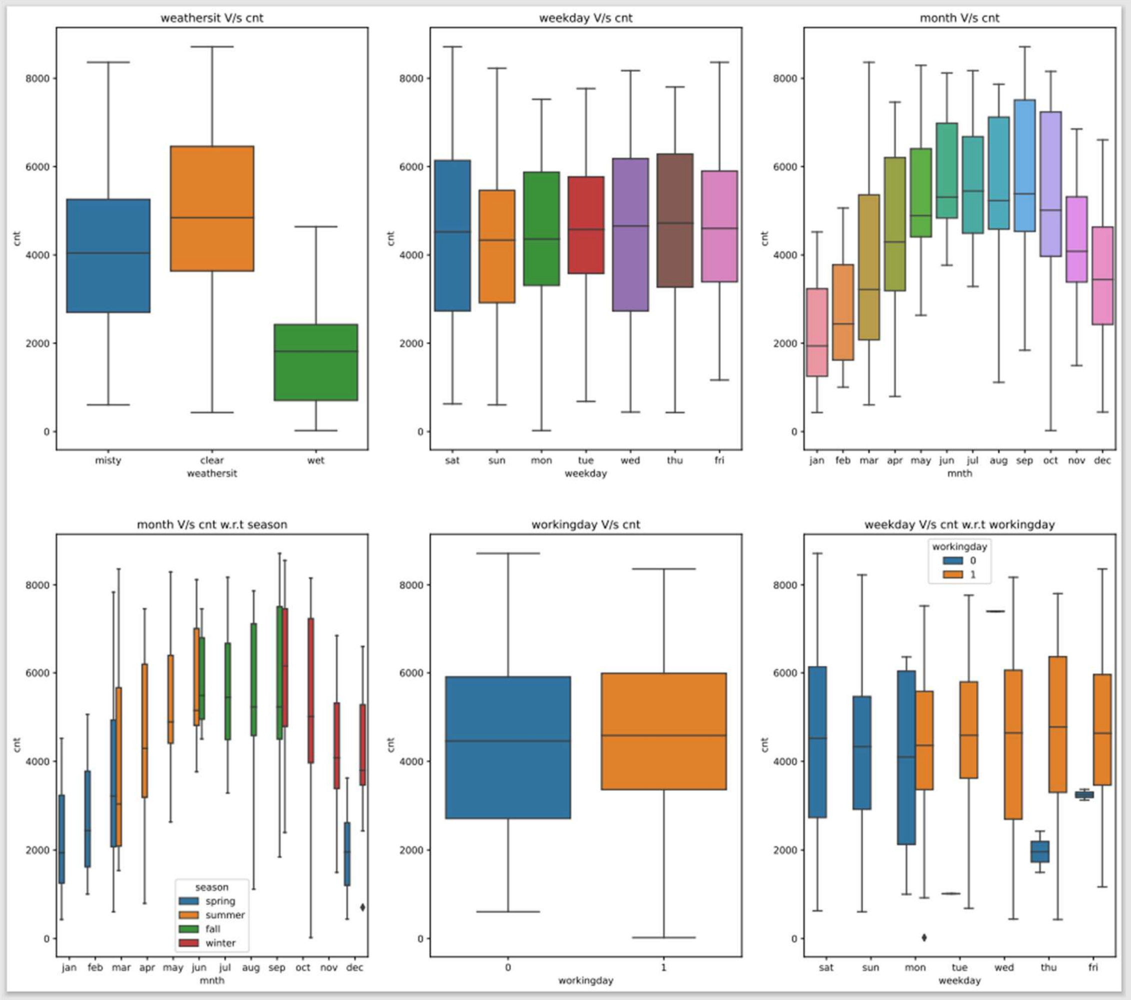
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans : There are basically 4 Categorical Variables – Season, Weathersit, Mnth and Weekday . When we plot a Boxplot with these Categorical Variables V/s cnt below are the observations,

- Seasons V/s Cnt – Boxplot
  - Summary of seasons V/s cnt (Left Figure)
    - Bike Demand is high in Summer and Fall and it decreases in Winter
    - Demand for bike in spring is less and this doesn't sum-up as it should be high due to favourable weather conditions so we plotted "Seasons V/s Cnt w.r.t Year"
  - Summary of Seasons V/s Cnt w.r.t Year (Right Figure)
    - Shows that in 2018-Spring as it was started initially the cnt (due casual + registered) was less and it gradually increased from 2018-summer onwards and the mean has gone Higher in 2019-spring to Winter Year-on-Year
  - These boxplots shows how the cnt got increased right from 2018 till end of 2019 and after that Covid Happened now, once covid is over and during path-to-recovery, company already has registered customers and initial demand pattern may follow similar pattern that as shown below as first Q1 may see less demand BUT demand will definitely increase as Covid situation stabilizes .
  - Bike Sharing Demand has increase Year-on-Year from 2018 to 2019 .



- Boxplot of Weathersit, mnth, weekday , workingday V/s cnt
  - Summary –
    - Mean Demand (cnt) is high with clear > misty > wet > snow (0)
    - Mean Demand is almost similar for all Weekdays .i.e. on average bike demand on all days are similar .
    - Bike demand is high from May to Oct
    - Mean Bike demand (cnt) is similar for working (1) and holiday (0)
    - Mean Bike demand (cnt) is similar for all weekdays irrespective of holidays and working day (workingday variable)



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

Ans : Categorical variables cannot be used directly in the model creation instead it has to be converted into meaningful Numeric values which is done through dummy variables encoding for each Categorical variables .

During the dummy variable creation, each categorical variable with say N levels, we create 'N-1' new indicator variables for each of these levels thus we have to drop one level using drop_first=True.

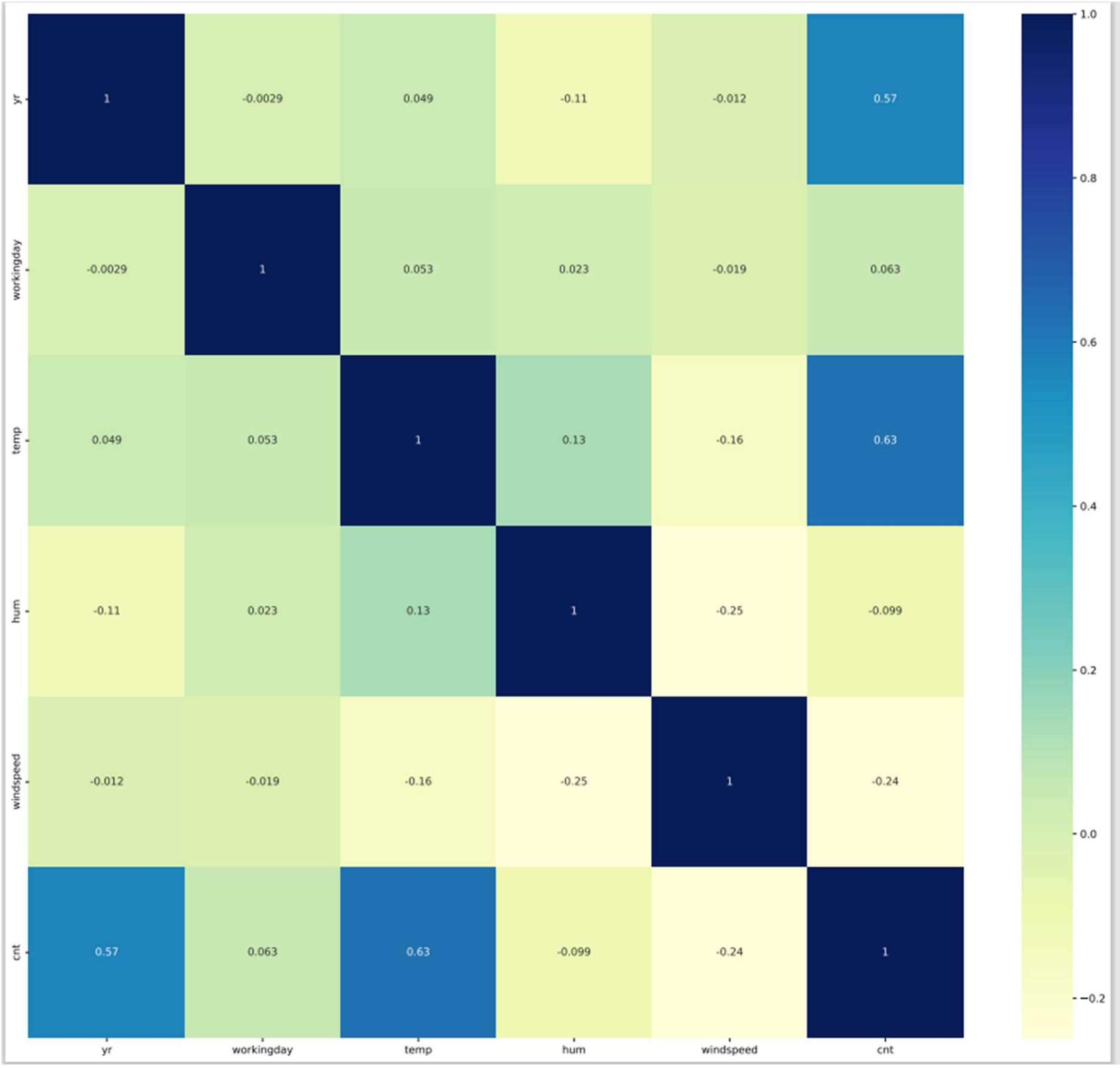
If drop_first = True is not set, then there will "Multi-colinearity" b/w these variables and that can impact model train and prediction.

To overcome Multi-colinearity b/w generated dummy variables we use drop_first=True to drop one column without actually losing an information.

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

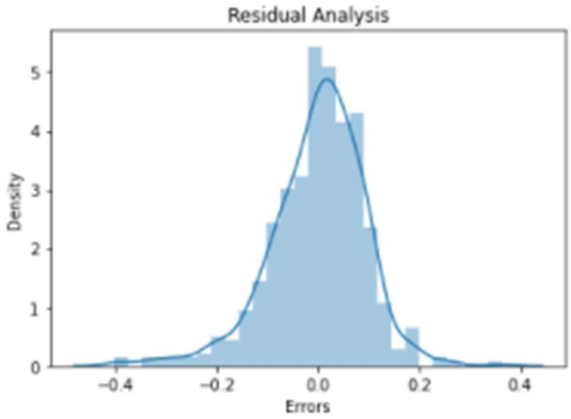
Ans: First Highest correlation with Target variable cnt is from "temp" variable – 0.63

Second higher correlation is from yr variable



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

Ans : Residual Analysis was performed and checked the Distribution of Error Terms . Below Graph shows that Error Terms are Normally Distributed around mean value 0 .



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

Ans : Model shows features which are positively and Negatively impacting the demand as below,

The equation of the best fit line which also describes the impact of each variable is,

cnt = 0.2981 (const) + 0.236 * yr + 0.397 * temp - 0.153 * windspeed - 0.145 * spring - 0.079 * misty - 0.275 * wet - 0.073 * jul + 0.052 * sep + 0.023 * sat

Basically this means,

- Bike demand is impacted positively with below variables,
    - yr - As it is a new concept, YoY there seems to be an increase and can fairly assume higher demands in current year.
    - temp - increases with slight higher temps in US so may be launched in cities which has acceptable temps throughout the year .
    - Bike Demand is slightly higher in Sep Month and during Saturday in a week .
- Bike Demand is impacted negatively with below variables
    - Bike Demand is less during high wind speed .
    - Bike Demand is less in Spring : This may be due to the Dataset as we saw in Boxplot that during Spring time Demand curve slowly increased from first half to Second half . Please refer to Box plot - seasons V/s cnt w.r.t year
    - Bike Demand is less in Misty and Wet weather conditions basically about to rain or rainy sessions

General Subjective Questions

6. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a machine learning algorithm based on supervised learning . Regression is the most commonly used Predictive analysis model .

Linear Regression model is used to find out relationship between the dependent (target Variable) and independent (predictors) variables. There are 2 types of Linear Regression models based on the Number of Predictor variables

- Simple Linear Regression - When one predictor variable is used
- Multiple Linear Regression - When multiple predictors are used

In a Simple Linear Regression, the regression line is given by

y = Beta 0 + Beta1 * x

Here -

Beta0 : Intercept means Value of Y when X=0

Beta1 : Slope

Similarly in Multiple Linear Regression, the Regression line is given by

y = Beta0 + Beta1 * x1 + Beta2 * x2 .... + Betai * xi where i - multiple datapoints

Basically, We find the Best Fit line by minimizing the RSS (Residual Sum of Squares) which is equal to sum of squares of residual for each data point in the plot . Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable

Drawbacks of RSS is it's value will change based on the unit so need to define alternative measure which should be more "Relative" and Not absolute so TSS (Total Sume of Squares) is defined .

Using RSS and TSS , R-Squared or Coefficient of Determination is defined as R-Squared = 1 - (RSS/TSS) . R-Squared statistics provides measures of how well actual data points are replicated by model based on the total variations of outcomes as explained by the model i.e. expected data points. Basically, Higher R-Squared value mean Model fits the data .

Once the Simple or Multiple linear regression Model is built, we need to do the residual analysis and see if it meets the Assumptions of Linear Regression to confirm the model fit

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

In Multiple linear regression Model, need to check for

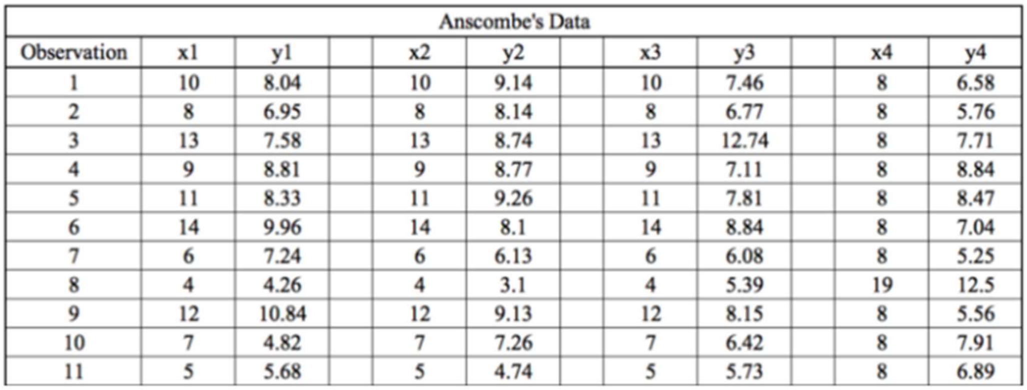
- Overfit
    - Model is complex and provides a near perfect linearity for Actual and predicted values .
    - Basically this show High R2 values for Training set while low R2 for test set
- Multicolinearity
    - This scenario happens when there is co-relation b/w variables used in model
    - This needs to be avoided for a decent model
    - Variance Inflation Factors values can be used to eliminate Multicolinearity withing variables
        - VIF < 5 - acceptable and no need to drop variable
        - VIF > 5 - Ok, but need to rechecked
        - VIF > 10 - Drop the variable

7. Explain the Anscombe's quartet in detail.

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 differences in the dataset that can impact the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

This can be used to demonstrate both, the importance of graphing data when analysis to identify any anomalies and effect of outliers and other parameters on statistical properties .

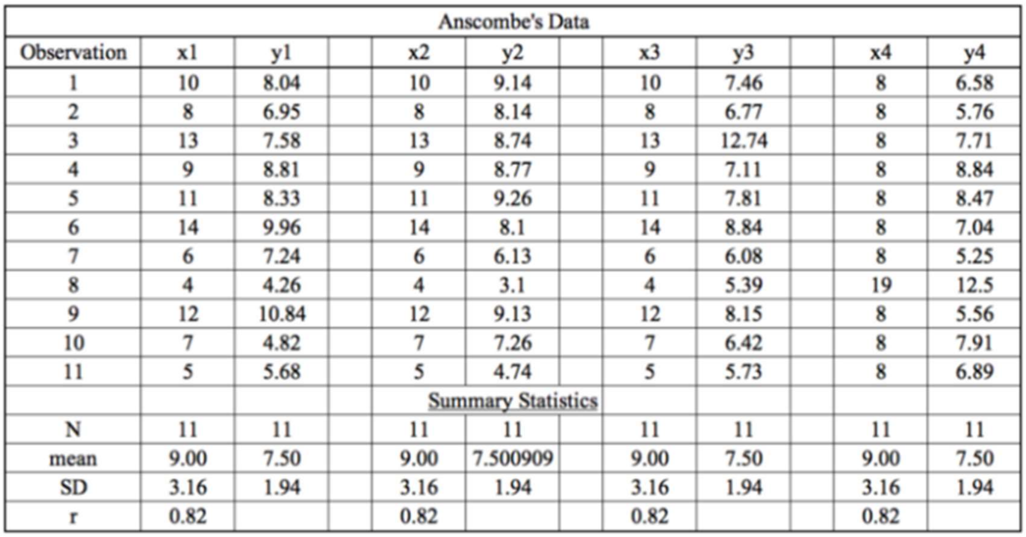
There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



Anscombe's Data

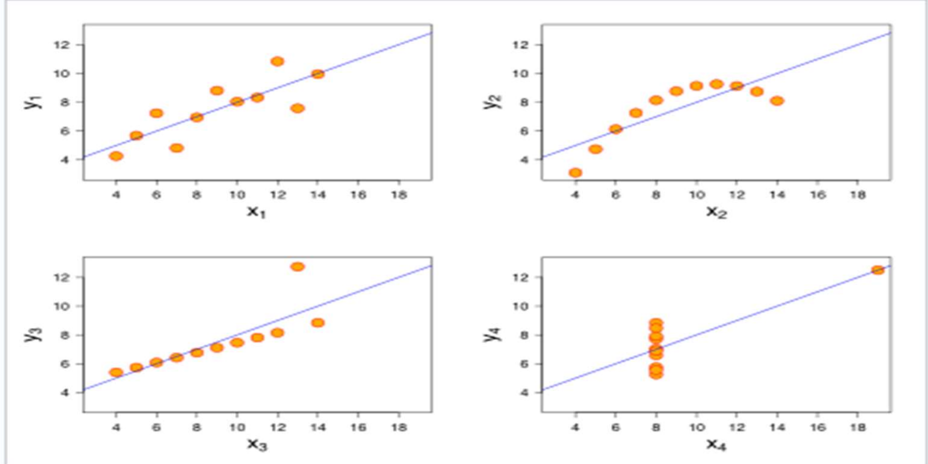
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	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

Statistical Information for all above datasets is approx. similar



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	

But when these models are plotted they generate totally different kind of scatter plots as below



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

The four datasets can be described as:

- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression mode

Conclusion:

- Basically it’s important to visualise the graphs to check the pattern before implementing any model .

8. What is Pearson’s R?

Ans: Pearson’s R or Bivariate correlation, is a statistics that measures the linear correlation between 2 variables and like other correlations it’s numerical value lies b/w -1.0 and +1.0

Peasrson’s correlation coefficient is the Covariance of 2 variables divided by the product of their standard deviations.

Pearson’s correlation coefficient, when applied to population (rho) can be terms as Population correlation coefficient or Population Pearson correlation coefficient .

For Random variables X,Y

Population(X,Y) = Covariance(X,Y) / Stddev-X * Stddev-Y

A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may transform X to a + bX and transform Y to c + dY, where a, b, c, and d are constants with b, d > 0, without changing the correlation coefficient.

The correlation coefficient ranges from −1 to 1. An absolute value of exactly 1 implies that 2 variables are perfectly linear correlated. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1. A value of 0 implies that there is no linear dependency between the variables.

9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Feature Scaling is a method during Data Pre-processing to normalize the range of Independent Variables or features of Data.

Most of times, Data collected contains features with different units and ranges. Without Feature Scaling if model is built then it only considers the Values of the features and not the Units and this leads to Outlier type of behaviour when model is plotted and an incorrect Model which gives very high or low coefficients .

Scaling affects the only coefficients and none of the other parameters like t-stats, F-stats, p-values and R-squares are impacted

There are 2 ways of Scaling

- Standardized Scaling : The variables are scaled in a way that their mean is 0 and Standard Deviation is 1

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

- MinMaxScaling (Normalized Scaling) : Here the variables are scaled in a way that their values lies between 0 and 1 using the Max and Min values in the dataset .

  $$x = \frac{x - min(x)}{max(x) - min(x)}$$

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

Ans : Based on the VIF Formula as below, VIF can be infinite if R2 is 1 .

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

If R2 is 1, this means there is a perfect correlation between 2 independent Variables. To solve this, we have to drop one of the variables from the dataset causing this perfect Multi-colinearity.

But in real time this is very rare to happen.

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

Ans : Q-Q plots (Quantile-Quantile plots) is used to find out if two sets of data come from the same distribution or not. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, then points will fall on or very close to that 45-degree reference line.

In Linear Regression this is helpful 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 or not.

Some Advantages

- it can be used with sample sizes
- Distributional aspects like scale shifts, changes in symmetry, presence of outliers can be detected

Used to check below scenarios,

- Population data with a common distribution
- Have Common scale
- Have Similar distributional shapes
- Have similar tail behaviour