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?**

Answer -:

We have following categorical variables -:

- **Yr** – Represents a financial year
- **Workingday** – represents whether the current day is a working day or not.
- **Weathersit** – Represents the type of weather
- **Season** – Represents the ongoing season
- **Mnth** – Represents the current month
- **Holiday**- Represents the current day is holiday or not
- **Weekday**-represents which day of the week is the current day.
- which we are using for our analysis , and building our linear regression model

1.Category **year** is positively inducing the demand or count of clients i.e. the demand or the count of total client number is increasing by ~0.240287 with increment of a unit year as predicted from our linear regression model.

2.We can see in **working** days the demand is high and in non working days the demand is lower than the working days . In fact if we increase the working day by one unit then the demand increases by ~0.033529 as predicted from our linear regression model.

3. We have created dummy variables for category **weathersit**

- as **VeryGood_W** for very good weather conditions .
- **Good_W** for good weather condition.
- **Bad_W** for bad weather condition.
- And **VeryBad_W** for very bad weather condition.
- And we have seen from our analysis that across all the **season** and **year** ,and **working days the Very good** and **good weather** is giving us a boost in demand. As per our linear regression model we have seen **a unit change of very good weather condition** gives ~**0.320031** increment in demand and **unit change of good weather condition** gives ~ **0.234094** increment in demand .

4. We have created dummy variables for category **season** as

- **fall, spring, summer, winter**.
- And we have seen from our analysis that we have a good amount of client demand on **fall and summer** but we are getting **huge dips in spring** and a **relatively lesser loss of demand on winter** . As our linear regression model gives the coefficients of the **slopes of the winter and fall as negative** we can confirm the fact . As per our linear regression model the coefficient of **spring** is **-0.310422** and coefficient of **winter** is **-0.050412 .** Which says unit increase in **spring** will cost **-0.310422** demand and unit increase in **winter** will cost **-0.050412** to the demand **.**

5**.** The category **month** which ranges from **January to December** , the **weekday** , the **holiday** columns are explained by the **Year** , and **working day** respectively . As the **year** in a **macro level can describe the increment of the demand more better than months** and we can really see **the weekend and holidays are falling in the category of non working days** which is described by the **workingdays** column .Hence to avoid multi collinearity we are dropping those variables from our analysis and modelling.

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

Answer -:

- The removal of first column of the dummy variables of a categorical column is necessary due to the redundancy in explaining the data of this extra variable.
- If we have a binary or any sort of categorical variable , where the n-1 dummy variables can explain the state of the extra 1 dummy variable , then we can reduce or drop one dummy variable , as it is unnecessary and will increase the chance of multicollinearity while building model.

- For example Let's say we have a categorical variable is_raining , where it can either hold yes_raining and no_not_raining.
- When we will convert the is_raining variable to dummy variables we will get two variables as yes_ raining and no_not_raining.
- Now , if we see the state of yes_raining can be either true or false or 1 or 0 , and if it is 0 , then explains it is not raining .
- We can clearly see the state 0 of yes_raining is explaining no_not_raining , hence the no_not_raining is just an extra dummy variable which is explained by the prior one . Removal of such dummy variables are important as they can be explained by the other dummy variables and increase collinearity .
- And form our variance inflation factor formula (VIF=1/1-R^2) we can see a high correlation of an independent variable with other independent variables(multicollinearity) can lead to a lower coefficient of determination (explain ability) of our model.

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

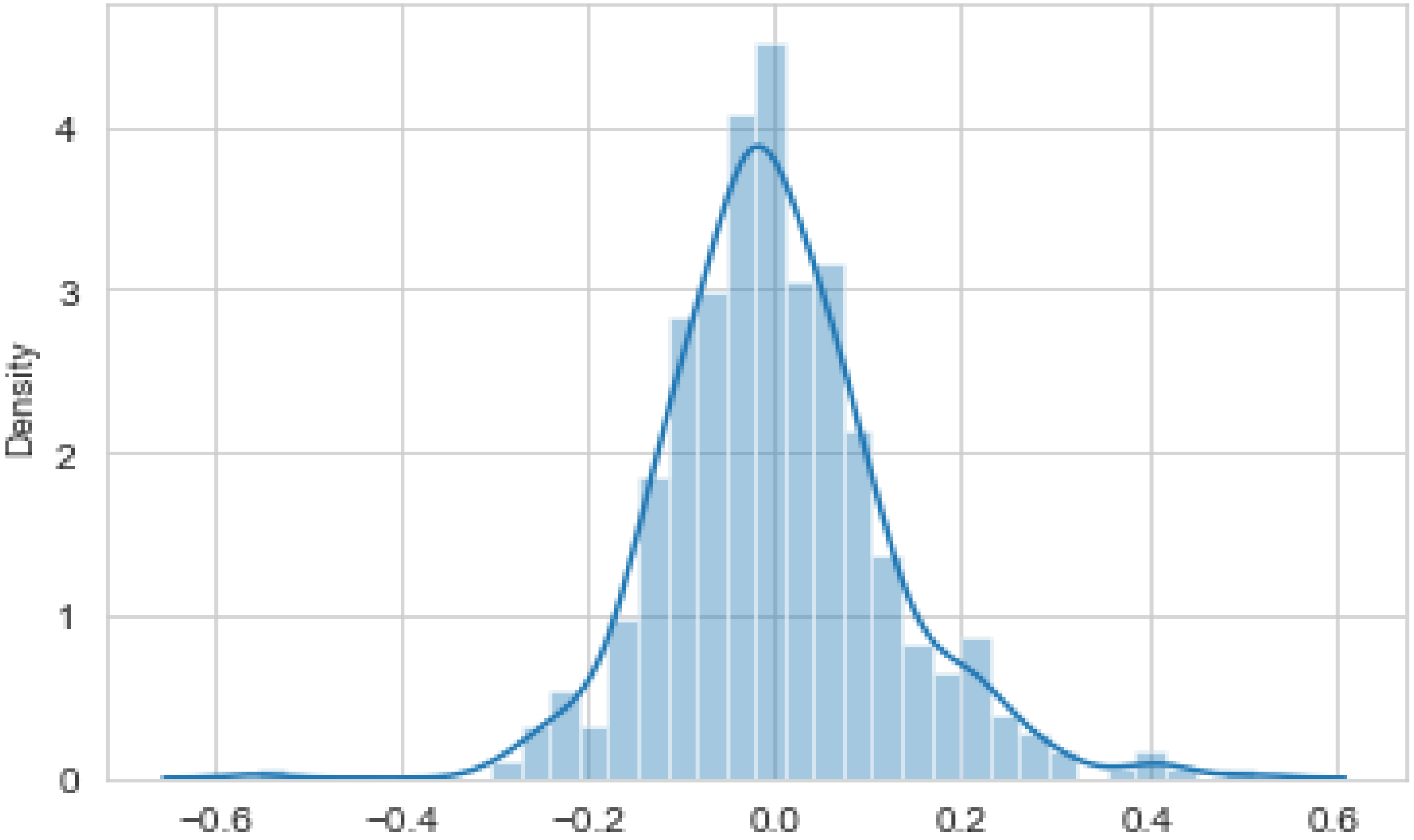
Answer-:

Among the numerical variables the **registered** have a strong ,positive correlation of a **0.95** with **cnt**.

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

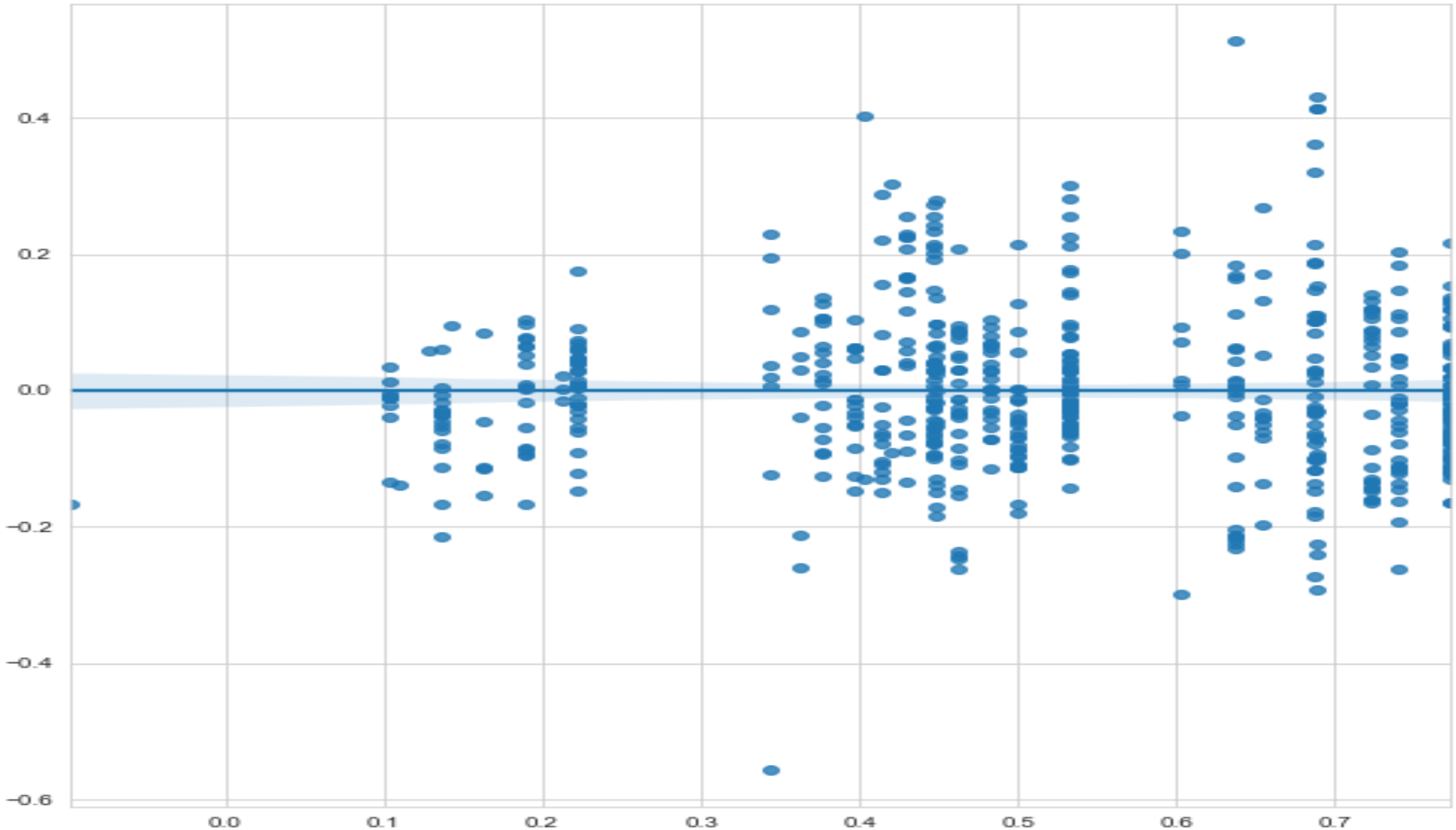
Answer-:

- First we predicted our target variable using the built model on training dataset.
- we subtracted the actual train target value from predicted target value and found the residual terms.
- Then we validated the assumption that **residuals should be distributed normally , as there mean being 0** .
- And as per the following graph we found the residuals are in fact normally distributed.


Density
0
1
2
3
4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

- Then we validated the assumption that **the residuals be randomly distributed and the distribution should follow homoscedasticity .**

- And as per the following figure the residuals are randomly distributed and having **homoscedasticity or equal variance**. Hence we can say our model is pretty accurate**.**



yr

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

Answer -:

- Based on my final model
- **Very good weather** with a coefficient of 0.320031 which is positively and highly impacting the demand of the shared bikes **.**
- **Spring** with a coefficient of -0.310422 which is negatively and highly impacting the demand of the shared bikes.
- **year** with a coefficient of 0.240287 which is positively and highly impacting the demand of shared bikes .

General Subjective Questions

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

- Linear regression model finds a best fit line which fits through the data
- Which means it finds a best linear relationship between dependent and independent variables.
- The best fit line or relationship is evaluated by a cost function called as residual sum of squares or RSS . Which can be defined as
- E = square root of (square of(predicted value of Y1 – actual value of Y1) + square of(predicted value of Y2 – actual value of Y2) +…….+ square of(predicted value of Yn – actual value of Yn))
- This method of finding the residual sum is also called as ordinary least square approximation , which helps to minimize the cost function by which the line may deviate from the best possible fit .

The constructed best fit line should also under go certain validations of assumptions to be considered as a good linear model.

Such assumptions are -:

- Linearity assumption of the model
  - The predictors and the target variables should have a linear relationship.
- Normality assumption of the residuals.
  - The residuals are normally distributed . Hence the distribution of residuals are random .
- Zero mean assumption of residuals.
  - The residuals are normally distributed around the mean being zero.

- Homoscedasticity or equivariance assumption of residuals.

➢ The residuals are randomly distributed and have an equal variance , which means there is no pattern to the distribution of the residuals.

- Independent residuals assumption.

➢ Again this means there is no explainable relation between the residuals .

A linear regression model can be defined by a linear equation -:

$Y = \beta0+\beta1*X1+\beta2*X2+\beta3*X3+\beta4*X4+\ldots\ldots+\beta n*Xn$

Here ,

- Y is the dependent variable .
- X1 , X2 , X3 , X4 , …….,Xn are independent variables or predictors.
- And the coefficients or slope of each variable can be defined as β1 ,β2 ,β3,β4,βn.
- There are some assumptions the independent variables must follow too

➢ It should have a linear relation with the dependent variable and must not have a multi collinearity.

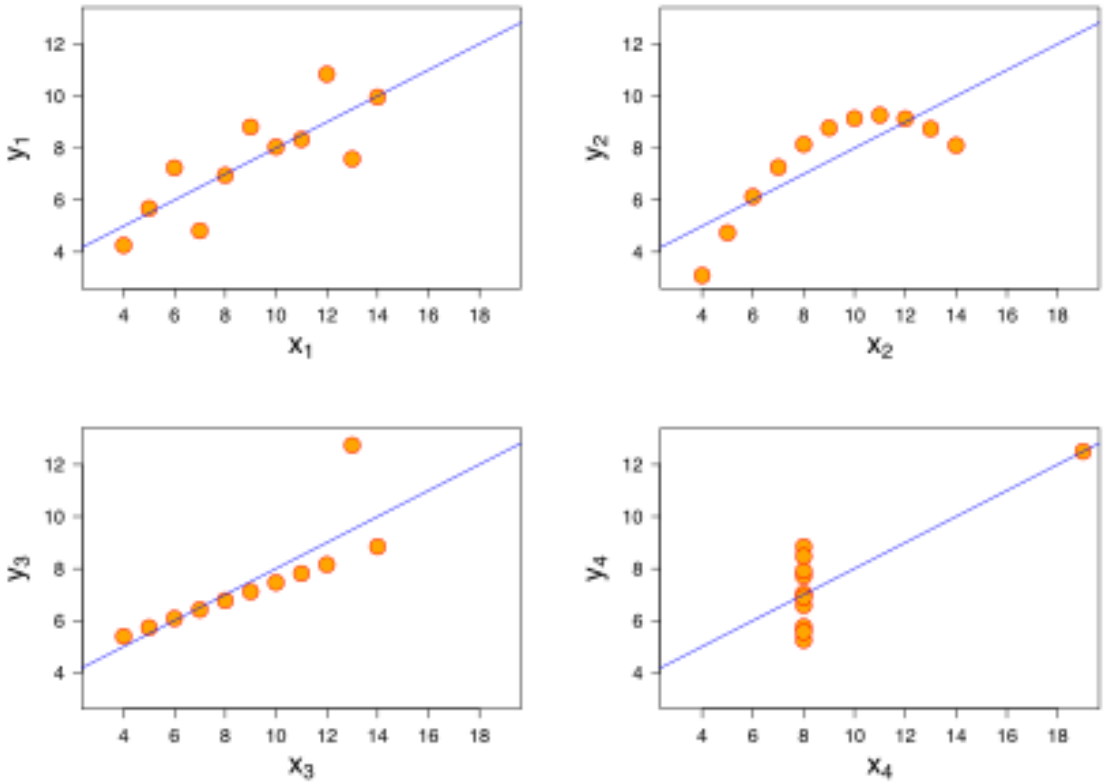
➢ The significance of the variables should be higher which implies the probability value ($p<0.5$).

The interpretation of the equation can be ,the change of y or the target variable , by the coefficients or the slope of the predictor variable , upon unit change in the predictor variable , while keeping the rest of the predictors constant except the predictor currently being predicting the rate of change.

2. Explain the Anscombe's quartet in detail.

Answer-:

- Anscombe's quartet explains about the different distribution of data between 4 data sets , while having equal summary statistics .
- That being said , it means the distribution of data among various data sets can vary even if the summary statistics of them like – mean , standard deviation , r2 score are exactly similar.
- In 1973 a famous statistician named Anscombe came up with 4 uniquely distributed dataset which are having equal summary statistics .
- These images are copied from Wikipedia -:



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s^2 x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s^2 y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : {\displaystyle R^{2}}	0.67	to 2 decimal places

- We can see the summary statistics of the four datasets are equal to the given number while the distribution of them looks quite different.
- The Anscombe' quartet also explains the importance of visualization of the distribution as we can not predict any trend or significance until we do not actually plot them or the statistical summery may fool our analysis about their distribution .

3. What is Pearson's R?

- Pearson's R is other wise known as Pearson's correlation coefficient between two numerical variables plotted in two different axis.
- The higher the correlation is , and closer to 1 or -1 , the higher the chance of ,linear relation ship between two variables to be strong.
- It can range from -1 to 1 , where -1 shows a lowest correlation as a line with a negative slope passes through all the data points.
- And 1 shows a Highest correlation as a line with positive slope passes through all the data points.
- The strength of the correlation can also be measured by the P-value , or probability value .Where the lesser the p value the better the correlation in explaining the linear relationship or we can say the correlation is significant.
- The lesser p- value also means the correlation will not be highly affected by the noise or randomness of the distribution . Hence a higher p-value with a similar correlation will result in a lower significance of the linear relationship , which must be avoided while analyzing the relationship between two variables.

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

Answer-:

- Scaling or feature scaling is the the technique to converge distribution of features ,of different scale , into a common scale .
- The machine learning algorithms which calculates the distance in any sort , will supress the effect of lower scaled feature , over the higher scaled one. Hence the prediction by the model might get wrong .
- Also feature scaling is necessary , while using a gradient descent algorithm to optimize the model, as it makes the algorithm run faster.
- The standardized scaling converges each data point towards a mean of 0 and standard deviation of 1 , and the data points lies beyond the deviation will be considered as outliers . Hence the standardized scaling is performed when we need to evaluate the effect of outliers on data too.
- The formula of the standardized scaling is ($x-\mu /\sigma$)
- While the normalized scaling does not separate the outliers while scaling and the scale may be wrong in case we have outliers in our data. As the denominator of the scale changes .
- The formula of normalized scaling is (X-Xmin/Xmax-Xmin) .
- As we can see the normalized scaling converges all data points in a range of 0 to 1.And are influenced by the outliers.

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

Answer-:

- When a feature explains another feature perfectly the value of coefficient of determinant r-square becomes 1
- Which means the feature is describing another feature/s perfectly.
- So when there is a perfect correlation between two features the denominator of the VIF equation (1/1-(r-square)) becomes zero.
- Which leads to VIF=infinite .

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

Answer-:

- Q-Q plot or quantile-quantile plot is a graphing technique , by which we can see whether or not a set of data came from a normal distribution , or any other distribution , or not.
- We plot theoretical z values along one axis , and we take the z values of quantiles of data in another axis , and plot a graph .
- If we get a straight line then we can say it is a normal distribution. Or other wise we can use this to prove other kinds of distributions as well such as exponential or uniform distribution to name a few.
- For plotting a Q-Q plot we basically sort our data in ascending or descending order and then take the z values related to specific quantile on one axis.
- Then we take the theoretical z values from z-table in another axis and plot points on the intersection of two axis.
- The intersection points then can explain about the distribution , for example if we want to prove the hypothesis that the data is normally distributed we can do so by plotting the intersections and check for a linearity .