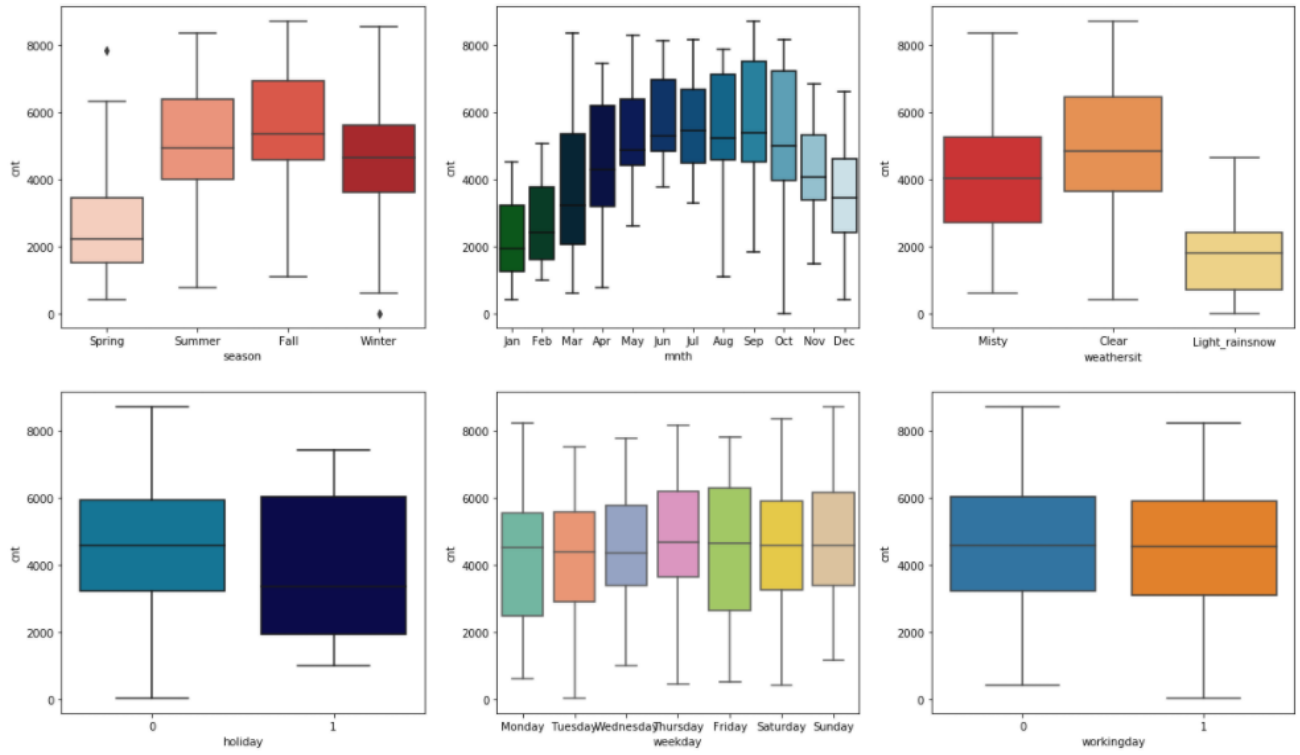
Assignment Questions (Subjective)

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



season: Most of the bike booking were happening in Fall with a median of over 5000 booking.This was followed by Summer & Winter. This indicates, season can be a good predictor for the dependent variable.

mnth: Most of the bike booking were happening in the months May, June, July, August and September with a median of over 4000 booking per month. This indicates, mnth has can be a good predictor for the dependent variable.

***Weathersit**: Most of the bike bookings were done on a clear weather situation with mean over 4000. This was followed by misty weather and the least was light_rainsnow.*

holiday: Major percentage of the bike booking were happening when there is no holiday which means this data is clearly biased. This indicates, holiday cannot be a good predictor for the dependent variable.

Weekday: The bike bookings were similar on all days of the week. So it may or may not be a good predictor of the count variable. It is not as significant as other variables.

workingday: Majority of the bike booking were happening in 'working day' with a median of close to 5000 booking This indicates, working day can be a good predictor for the dependent variable.

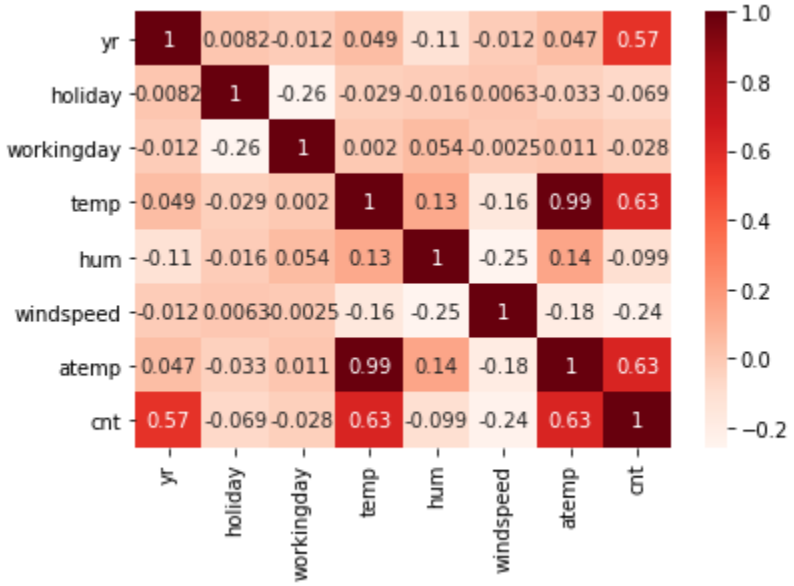
Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer: Dummy variable creation is a type of the approach to encode Categorical data. pandas get_dummies() method takes categorical feature as an argument. Then it creates a Dummy Variable for every label in the feature, such that each dummy variable holds data as 1 or 0. 1 indicates the presence of a particular label and 0 indicates the absence of a particular label.

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: Temperature

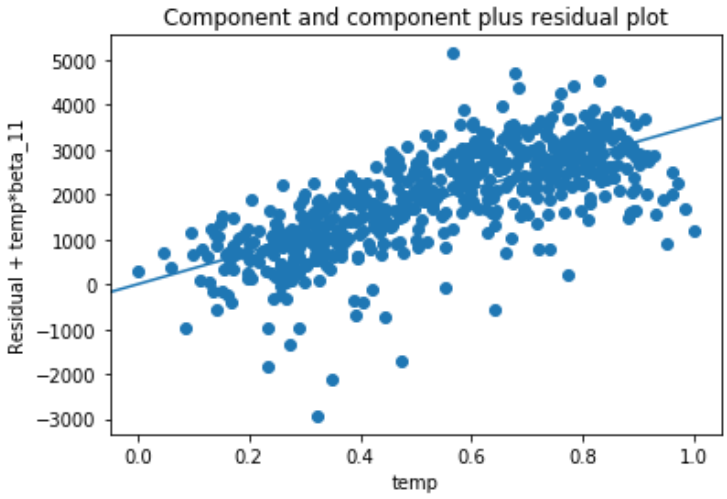


Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

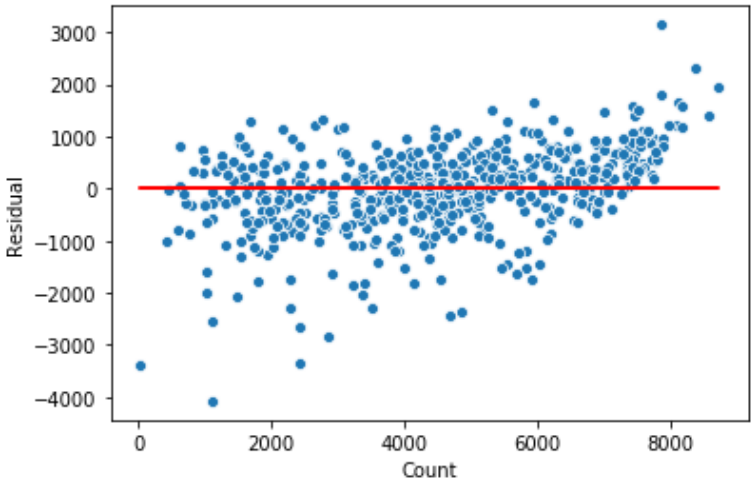
Answer: Validating the assumption of Linear Regression Model :

- Linear Relationship
- Homoscedasticity
- Absence of Multicollinearity
- No Autocorrelation in residuals.
- Normality of Errors

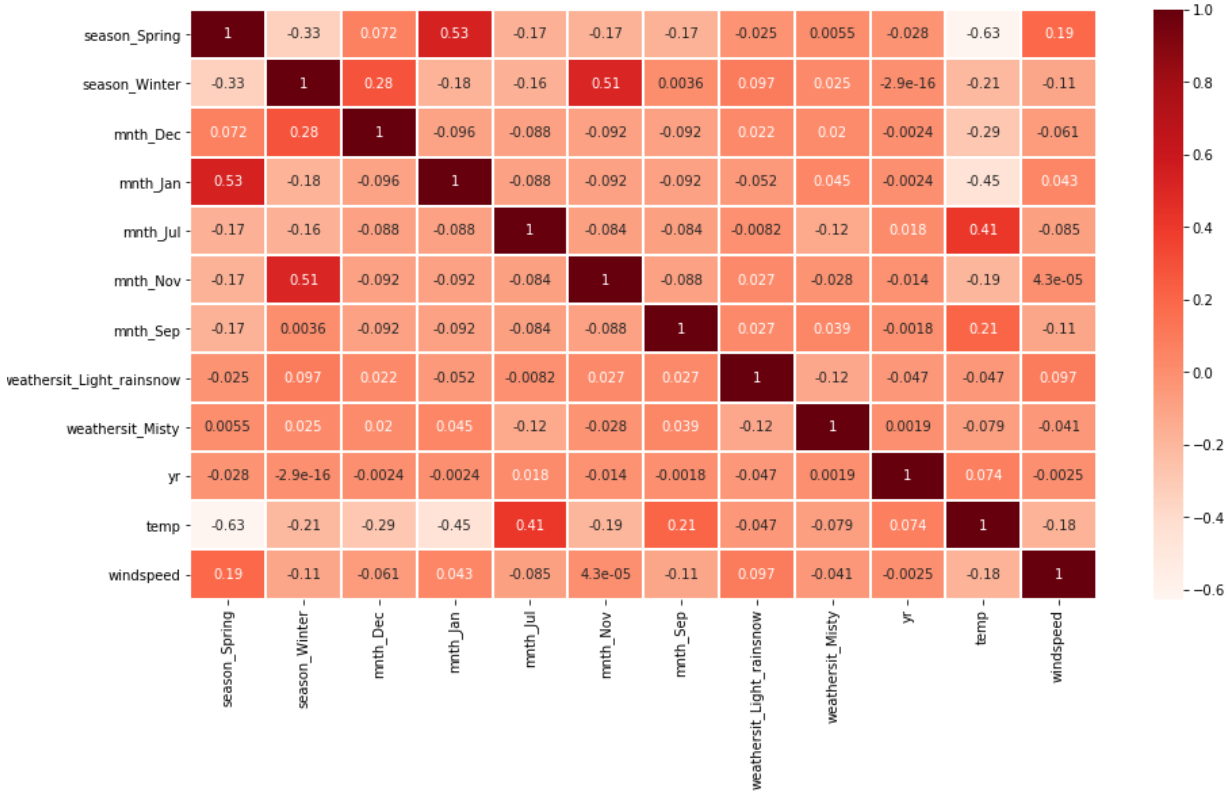
1. **Linear Relationship:** The above plots represents the relationship between the model and the predictor variables. As we can see,there is linear relationship in the model



2. **Homoscedasticity:** There is no visible pattern in residual values, thus homoscedasticity in the model



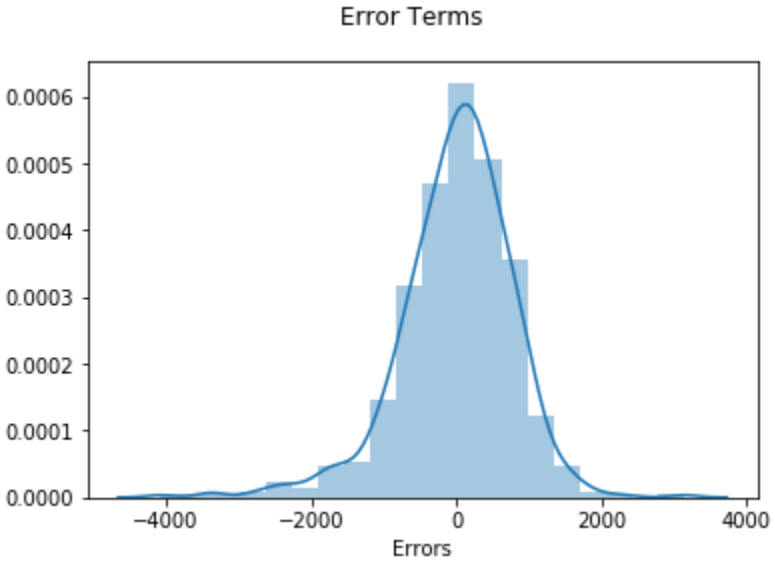
3. **Absence of Multicollinearity** : All the predictor variables have VIF value less than 5. So we can consider that there is insignificant multicolinearity among the predictor variables.



	Features	VIF
10	temp	4.73
11	windspeed	4.14
1	season_Winter	2.36
0	season_Spring	2.34
9	yr	2.07
5	mnth_Nov	1.67
3	mnth_Jan	1.61
8	weathersit_Misty	1.52
2	mnth_Dec	1.42
4	mnth_Jul	1.36
6	mnth_Sep	1.21
7	weathersit_Light_rainsnow	1.07

4. **Independence of residuals :**
   To check Independence of residuals/No Autocorrelation in residuals we can Use Durbin-Watson Test. The test output values will be between 0 and 4.
   The closer it is to 2, the less auto-correlation there is between the various variables.
   DW = 2 would be the ideal case here (no autocorrelation)
   - 0 < DW < 2 -> positive autocorrelation
   - 2 < DW < 4 -> negative autocorrelation
   - Our model's Durbin-Watson is 1.939 which is less then 2. Hence we can conclude that there is almost no autocorrelation.

5. **Normality of error**: Based on the histogram, we can conclude that error terms are following a normal distribution



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: As per the final model, the top 3 predictor variables that influences bike booking are:

1. **Temperature (Temp):**A coefficient value of " 3541.810723" indicated that a temperature has significant impact on bike rentals

2. **Light Rain & Snow (weathersit_Light_rainsnow**): A coefficient value of -2561.987118 indicated that the light snow and rain decrease people from renting out bikes

3. **Year (yr):**A coefficient value of 2021.916809 indicated that a year wise demand of rental bikes are increasing

As per Final model outcome we can conclude that these three variables are one of most important parameter for Bike rental booking. It can be recommended to give importance to these three variables while planning to achieve maximum bike rental booking.

General Questions (Subjective)

Q1 : Explain the linear regression algorithm in detail.

Answer: Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables.

Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

Y=mX+bY=mX+b

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

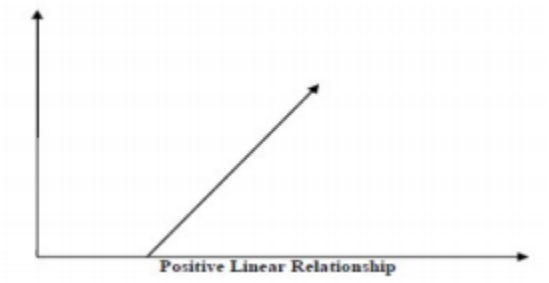
b is a constant, known as the YY-intercept.

If X = 0,Y would be equal to bb.

Furthermore, the linear relationship can be positive or negative in nature as explained below –

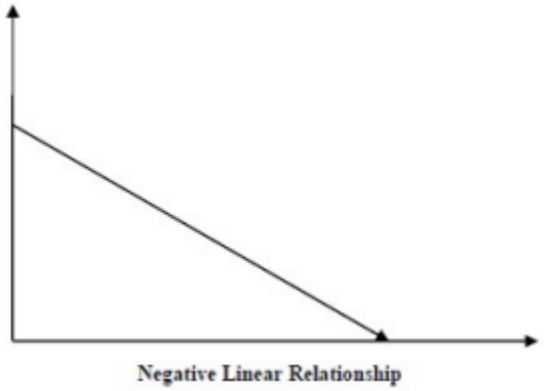
Positive Linear Relationship

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph-



– Negative Linear relationship

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Negative Linear Relationship

Types of Linear Regression

Simple regression

Simple linear regression uses traditional slope-intercept form, where mm and bb are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

$$y=mx+by=mx+b$$

Multivariable regression

A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$f(x,y,z)=w1x+w2y+w3zf(x,y,z)=w1x+w2y+w3z$$

The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation.

Q2 : Explain the Anscombe's quartet in detail.

Answer: 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 can mislead the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which

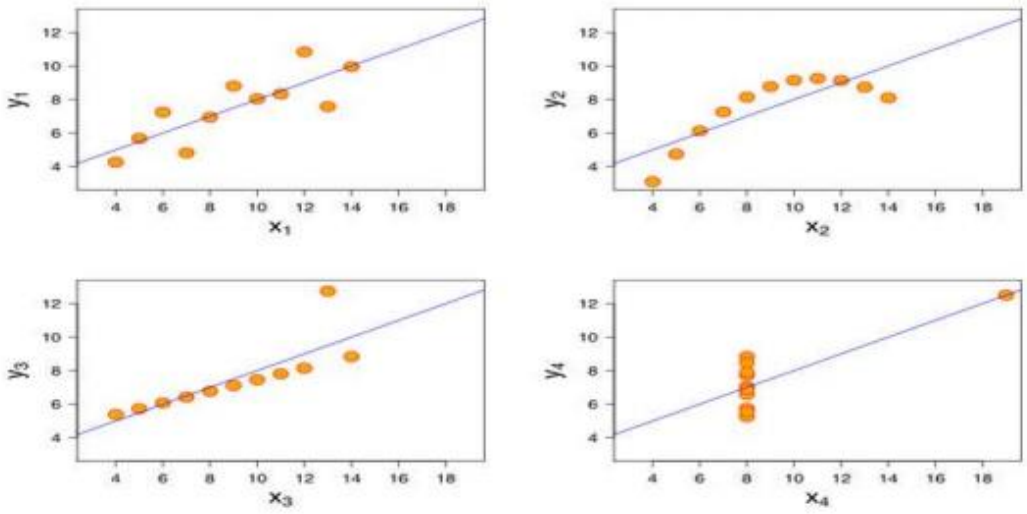
provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

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
				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	

Image by Author

This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear reparability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is mislead by these peculiarities and can be seen as follows:



The four datasets can be described as:

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

2. Dataset 2: this could not fit the model well as the data is non-linear.

3. Dataset 3: shows the outliers which cannot be handled by linear regression model

4. Dataset 4: shows the outliers which cannot be handled by linear regression model

Conclusion: All the Four dataset was having nearly identical variance ,mean correlation and regression lines .But Once we Visualize dataset we can actually understand the different between each dataset be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

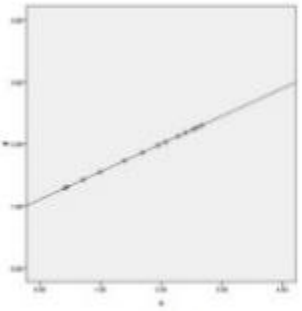
Q 3. What is Pearson's R?

Answer: Pearson's r (Pearson's correlation coefficient) is a statistical measure of the strength of a linear relationship between paired data. In a sample it is denoted by r and is by design constrained as follows

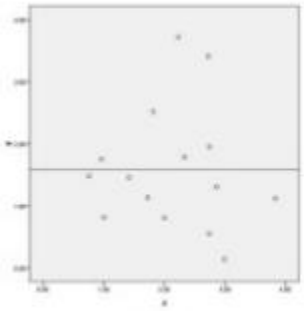
The Pearson's correlation coefficient varies between -1 and +1 where:

- Positive values denote positive linear correlation.
- Negative values denote negative linear correlation.
- A value of 0 denotes no linear correlation.
- The closer the value is to 1 or –1, the stronger the linear correlation.

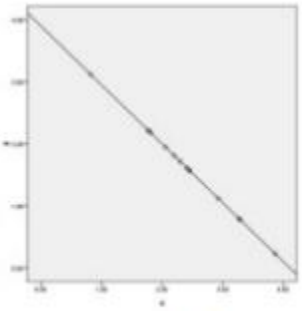
In the figures various samples and their corresponding sample correlation coefficient values are presented. The first three represent the "extreme" correlation values of -1, 0 and 1



$r = -1$
perfect -ve correlation

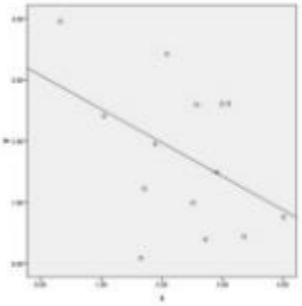


$r = 0$
no correlation

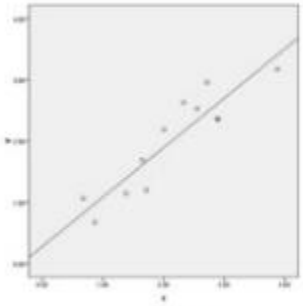


$r = 1$
perfect +ve correlation

When we say we have perfect correlation with the points being in a perfect straight line. Invariably what we observe in a sample are values as follows:



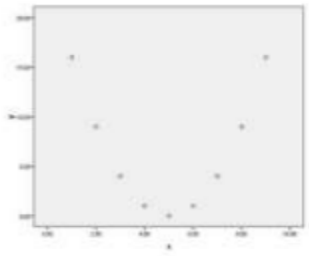
$r = -.45$
moderate -ve correlation



$r = .92$
very strong +ve correlation

- The correlation coefficient does not relate to the gradient beyond sharing its +ve or –ve sign!

- The correlation coefficient is a of does not imply there is no relationship between the variables. For example in the following scatterplot which implies no (linear)



$r = 0$
perfect quadratic relationship

The Pearson correlation coefficient is computed as:

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

As we can see, the correlation coefficient is just the covariance (cov) between 2 features x and y "standardized" by their standard deviations (σ), where the standard deviation is computed as

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Similarly, the covariance is computed as

$$cov(X,Y) = \sum_{(i=1)}^{N} \frac{1}{N}(x_i - \mu_x)(y_i - \mu_y).$$

Q4.What is scaling? Why is difference between normalized scaling and standardized scaling?

Answer: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalized scaling

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1

• If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardized scaling

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well). This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior