ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

1).FROM YOUR ANALYSIS OF THE CATEGORICAL VARIABLES FROM THE DATASET, WHAT COULD YOUINFER ABOUT THEIR EFFECT ON THE DEPENDENT VARIABLE?

Answer:In the dataset provided we have multiple categorical variables like season', 'yr', 'holiday', 'weekday', 'workingday', 'weathersit' where as each show Seasons, Year, Day of the Week,Working Day or not, Weather situation respectively.

Based on model prepared after pruning based on P-value and VIF we found following observation regarding effect of categorical variables on dependent variable:

season: Almost 34% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. Thisindicates, season can be a good predictor for the dependent variable.

mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictorfor the dependent variable.

weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this datais clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days ofthe week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

- Coefficients of variables of model created:

	coef
const	0.1580
yr	0.2341
temp	0.4828
windspeed	-0.1620
season_summer	0.0956
season_winter	0.1119
mnth_Aug	0.0568
mnth_Jan	-0.0461
mnth_Oct	0.0394
mnth_Sep	0.1162
weathersit_Cloudy	-0.0809

As you can see the correlation with the help of coefficient where sign of coefficient indicates positive or negative correlation.

2). WHY IS IT IMPORTANT TO USE DROP FIRST=TRUE DURING DUMMY VARIABLE CREATION?

Answer:

During dummy variable creation, the reference column which is a categorical variable is broken into binary columns such that their combination represents all values in reference column.

For Example: We have column type which has 3 values shown below and this is converted to dummy variables:

Type
First
Second

First	Second	Third
1	0	0
0	1	0

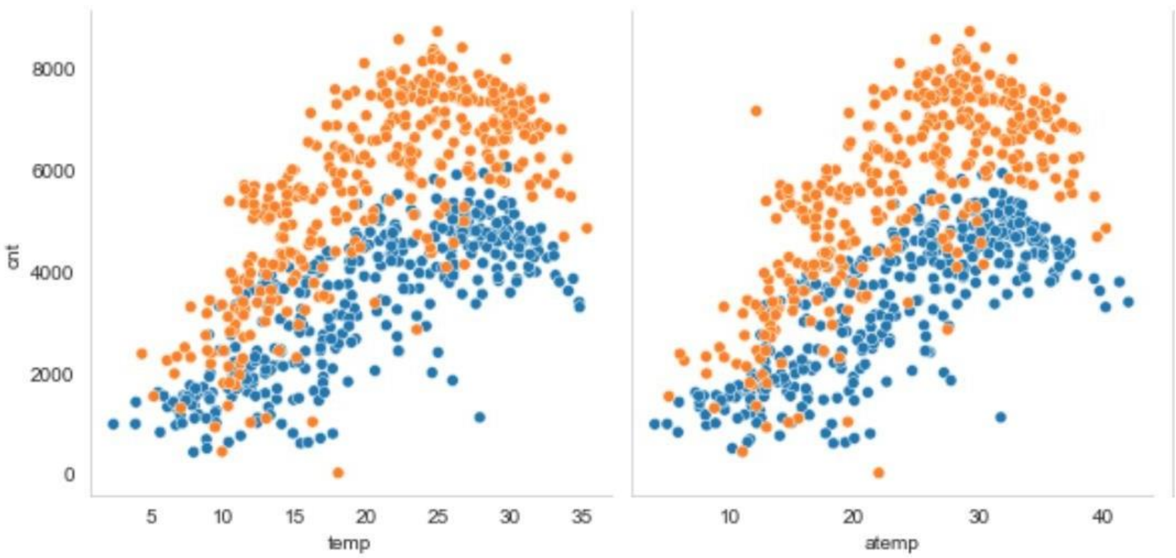
In such case drop_first is used to drop the reference variable as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Here once column can be dropped without loss to R square, thus it is recommended to use drop_first.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent thedummy variables.

3).LOOKING AT THE PAIR-PLOT AMONG THE NUMERICAL VARIABLES, WHICH ONE HAS THE HIGHESTCORRELATION WITH THE TARGET VARIABLE?

Answer:

Among all the numerical variables ‘temp’ and ‘atemp’ has highest correlation with the target variable and they are also highly correlated with each other thus both can be assumed too similar.



Also We will decide which parameters to keep based on VIF and p-value w.r.t other variables

4).HOW DID YOU VALIDATE THE ASSUMPTIONS OF LINEAR REGRESSION AFTER BUILDING THE MODEL ON THE TRAINING SET?

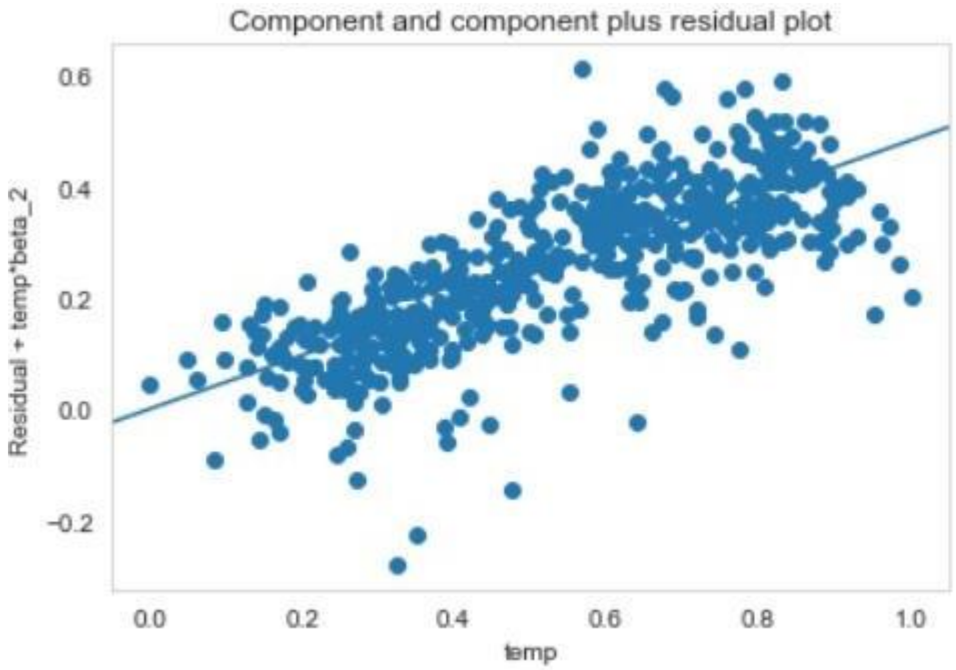
Answer:

The assumptions of Linear Regression are as follows:

- Linear Relationship
- Homoscedasticity
- Absence of Multicollinearity
- Independence of residuals (absence of auto-correlation)

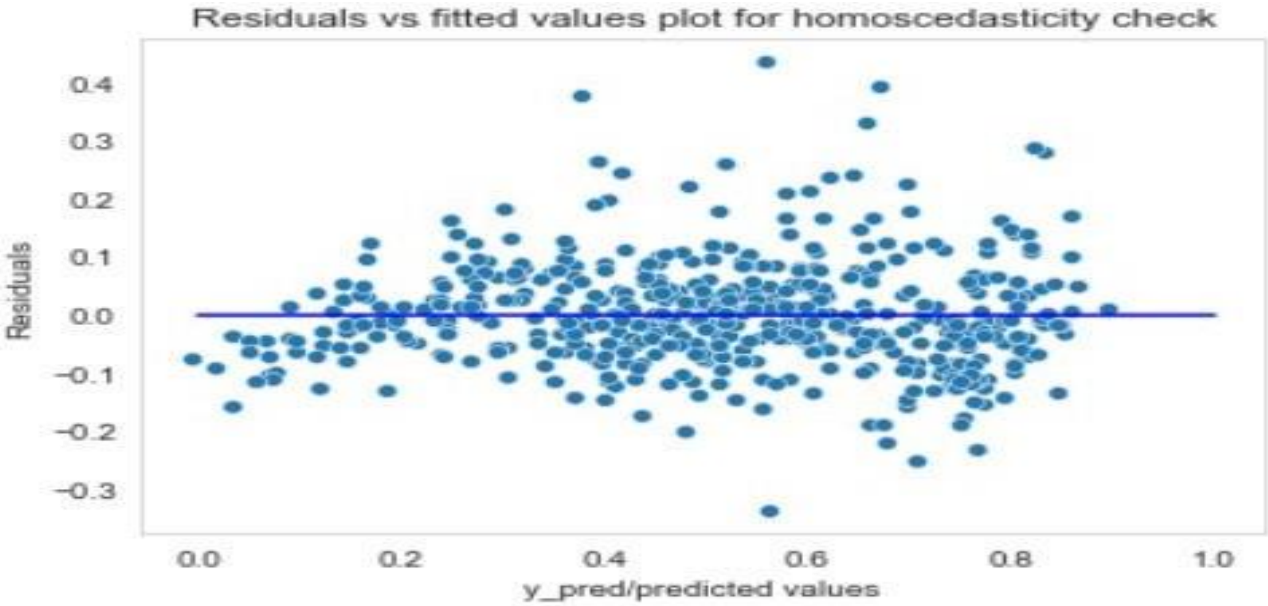
They were tested via following method with evidence in similar order as above:

- ✓ Linear Relationship was checked via partial residual plot (CCPR) in statsmodels library. The CCPR plot provides a way to judge the effect of one regressor on the

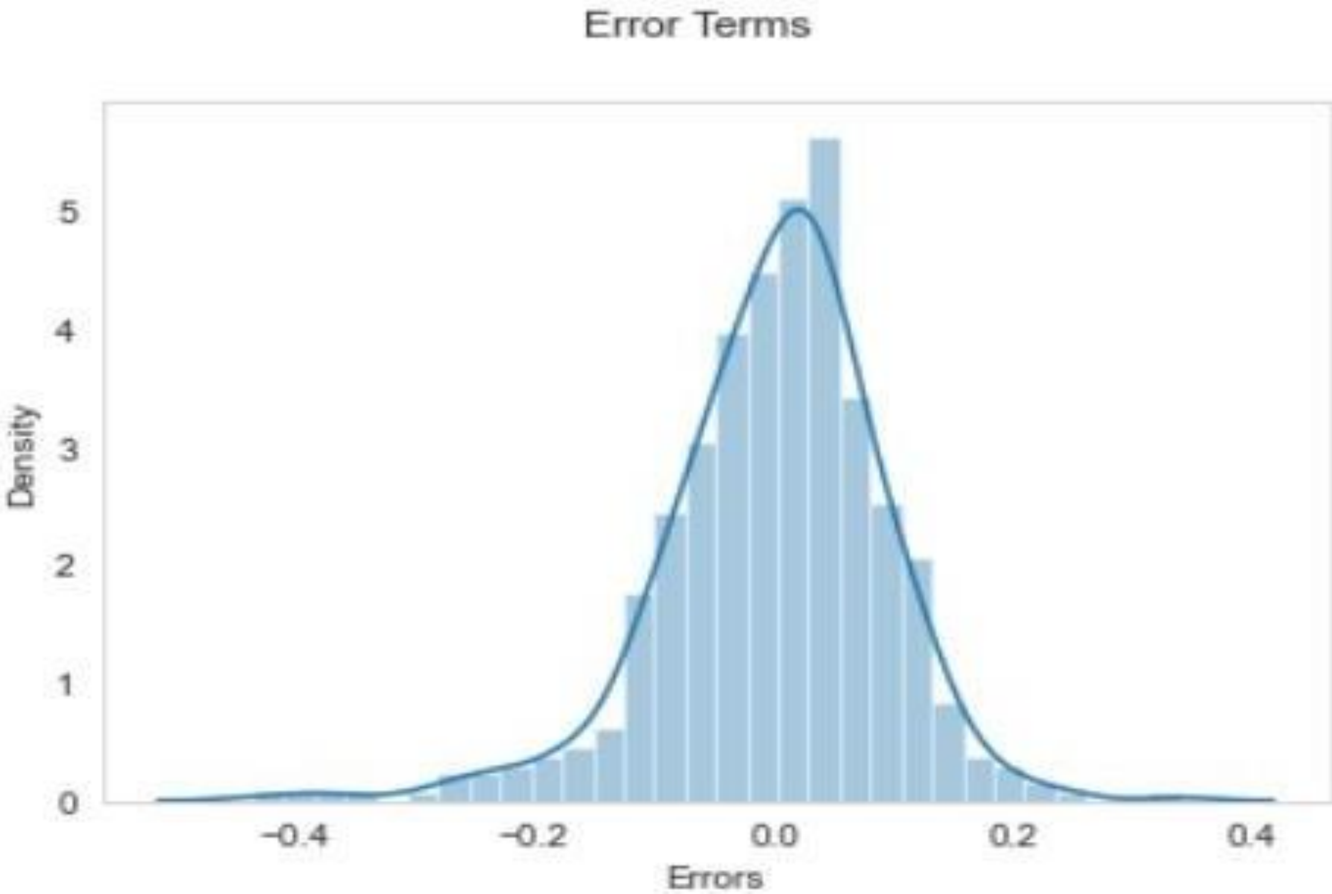


  response variable by taking into account the effects of the other independent variables.
- ✓ Here we plot target variable and ‘temp’ showing linear relation taking all other variables into account.

- ✓ Homoscedasticity was tested by plotting residual vs predicted values and it shows no pattern in scatterplot thus verifying Homoscedasticity.
- ✓ Multicollinearity was checked via heatmap and VIF where no column had high correlation or VIF after pruning.



- ✓ Independence of residual was verified by Durbin-Watson statistic where value of final model is 1.9896 which is close to 2 which indicates non-autocorrelation.
- ✓ The distribution of residual was checked using histogram which is normally



distributed.

5).BASED ON THE FINAL MODEL, WHICH ARE THE TOP 3 FEATURES CONTRIBUTING SIGNIFICANTLY TOWARDS EXPLAINING THE DEMAND OF THE SHARED BIKES?

Answer:

As per our final Model, the top 3 predictor variables that influences the bike booking are:

The mentioned top 3 features contributing significantly towards explaining the demand of the shared bikesare as follows:

- 'temp': 0.4828
- 'weathersit_Rainy: -0.2867
- 'yr': 0.2341

Remaining features:

	coef
const	0.1580
yr	0.2341
temp	0.4828
windspeed	-0.1620
season_summer	0.0956
season_winter	0.1119
mnth_Aug	0.0568
mnth_Jan	-0.0461
mnth_Oct	0.0394
mnth_Sep	0.1162
weathersit_Cloudy	-0.0809
weathersit_Rainy	-0.2867

GENERAL SUBJECTIVE QUESTIONS

1).EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL.

Answer:

Linear Regression is a type of supervised machine learning algorithm, it is used for predictive analysis. Linear regression makes predictions for continuous variables such as sales, salary, age, product price, etc.

It shows a linear relationship between dependent variable (y) and independent variables (x1,x2…xn) in the form of following equation where (β0,β1…βn) are coefficients/weights to depict the relationship:

Linear Regression: Multiple Variables

$$\boxed{\hat{y}} = \underbrace{\beta_0 + \beta_1}\boxed{x_1} + \dots + \underbrace{\beta_p}\boxed{x_p} + \boxed{\epsilon}$$

Linear regression can be further divided into two types of the algorithm:

- Simple Linear Regression: one independent variable
- Multiple Linear regression multiple independent variables

The Algorithms finds the best fit line by using a cost function which is least for best fit line which is given by R square:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

- It is a measure of goodness of fit.
- It is a relative value.
- Its value varies from 0 to 1, where is 1 is best.

There are some assumptions related to linear regression given as follows:

- Linear relationship between the features and target
- Small or no multicollinearity between the features
- Homoscedasticity Assumption: Error terms should not follow any pattern.
- Normal distribution of error term
- No autocorrelations

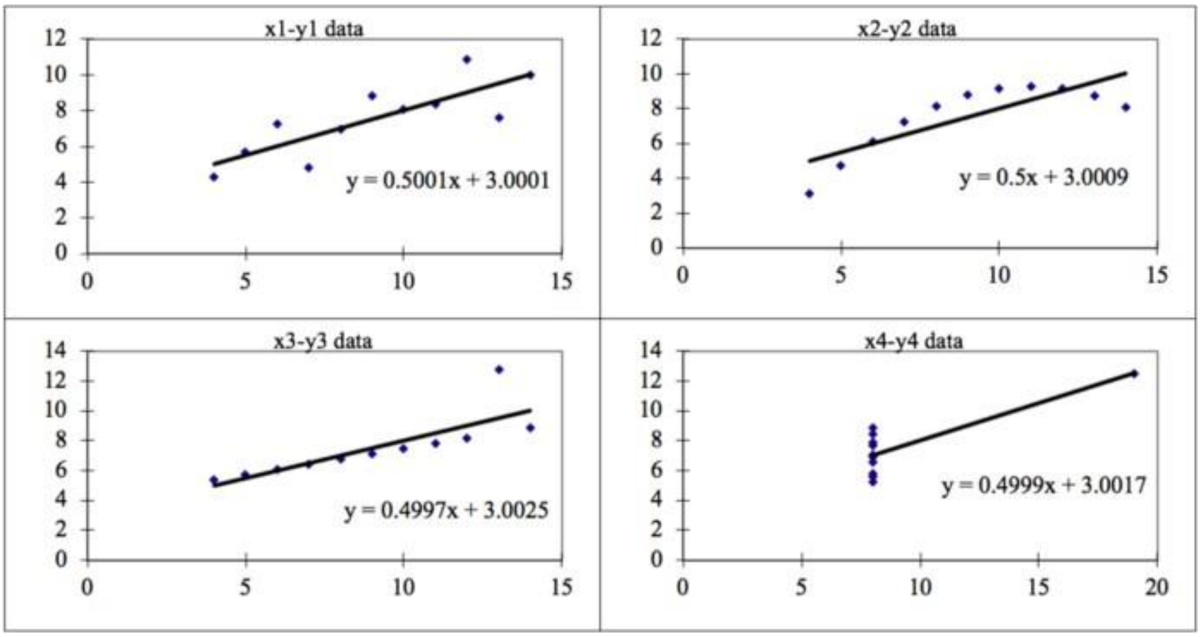
2).EXPLAIN THE ANSCOMBE'S QUARTET IN DETAIL.

Answer:

Anscombe's quartet contains four datasets with nearly identical simple descriptive statistics, such as mean and median, but the distributions are very different and look very different when plotted. It was created in 1973 by statistician Francis Anscombe to show the importance of plotting graphs prior to analysis and model building, and the impact of other observations on statistical characteristics. These four dataset plots have about the same statistical observations that provide the same statistics, including the variances and means of all x and y points in all four datasets.

The datasets are given below followed by them graphs:

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	



As shown here the statistics are similar but graphs are completely different which promotes importance of plotting graph before analysis.

The four datasets can be described as:

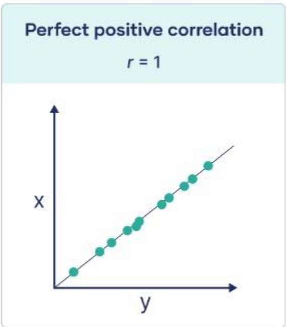
- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers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

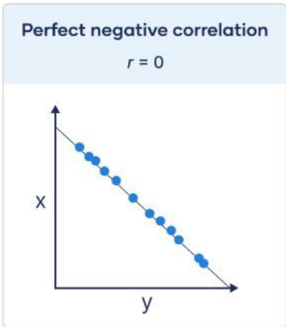
3).WHAT IS PEARSON'S R?

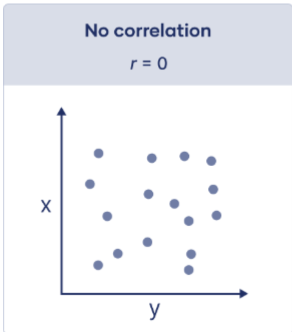
Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be

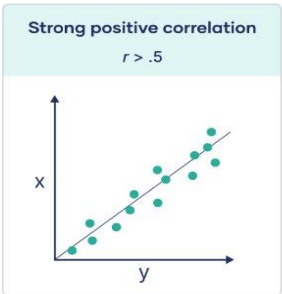
Positive, The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction ofthe relationship between two variables.

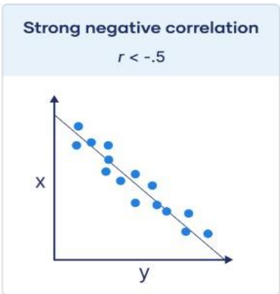
Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

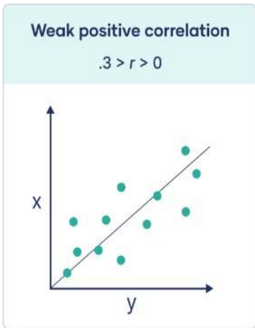


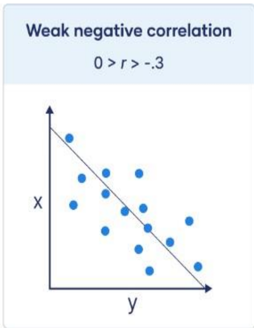












Assumptions with the Pearson's R are as follows:

- Both variables are quantitative
- The variables are normally distributed
- The data have no outliers
- The relationship is linear

4).WHAT IS SCALING? WHY IS SCALING PERFORMED? WHAT IS THE DIFFERENCE BETWEENNORMALIZED SCALING AND STANDARDIZED SCALING?

Answer:

Scaling is a step of data Pre-Processing which is applied to independent variables to normalizethe data within a particular range, it also called Feature Scaling is a method to standardize the independent features present in the data in a fixed range. It is part of data pre-processing to handle highly varyingfeatures which can cause bias in the model towards large values regardless of units which lead to wrong predictions.

Scaling is used to reduce the columns range to similar range across all columns to prevent bias generation due to very high values.

Two most common Scaling method are as follows:

- Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

5).YOU MIGHT HAVE OBSERVED THAT SOMETIMES THE VALUE OF VIF IS INFINITE. WHY DOES THISHAPPEN?

Answer:

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model, its the indicator of correlation between the independent variables, where it is calculated from R2 of model which trains on remaining independent variable to predict the remainingone.

In case of perfect correlation, R2=1 and we get VIF=1/(1-R2) which is Infinity, this means that variable can be explained by linear of other variables. This is perfect multicollinearity, which is resolved by removing one of the features causing this and iterating till VIF is lowered enough.

6).WHAT IS A Q-Q PLOT? EXPLAIN THE USE AND IMPORTANCE OF A Q-Q PLOT IN LINEARREGRESSION.

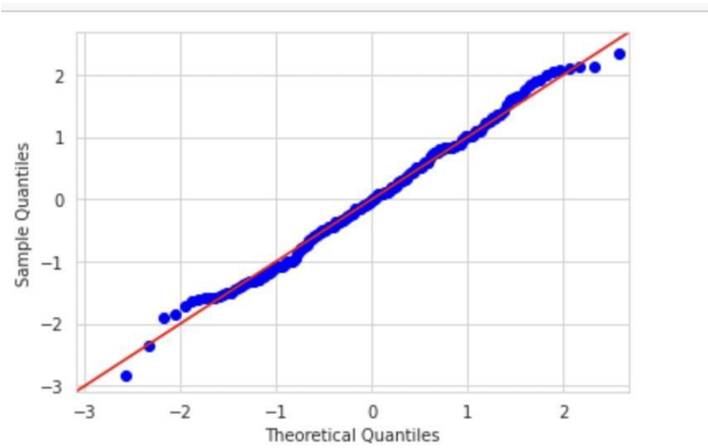
Answer:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. It plots quantile of sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The slope tells us whether the steps in our data are too big or too small. for example, if we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes (a.k.a. quantiles) compare between our data and the normal distribution.

One example cause ofthis would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).



We have plotted sample vs theoretical Quantiles against each other and based on points we can predict what distribution is followed by sample.

In this case as it follows a near straight line, we can say it follows normal distribution.

Steps:

- Sort from smallest to largest
- Draw a normal distribution
- Find the z-value (cut-off point) for each segment
- Plot your data set values (Step 1) against your normal distribution cut-off points