Linear Regresson subjective questions

- **Submitted by – Angad Singh Sachdeva**

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?**

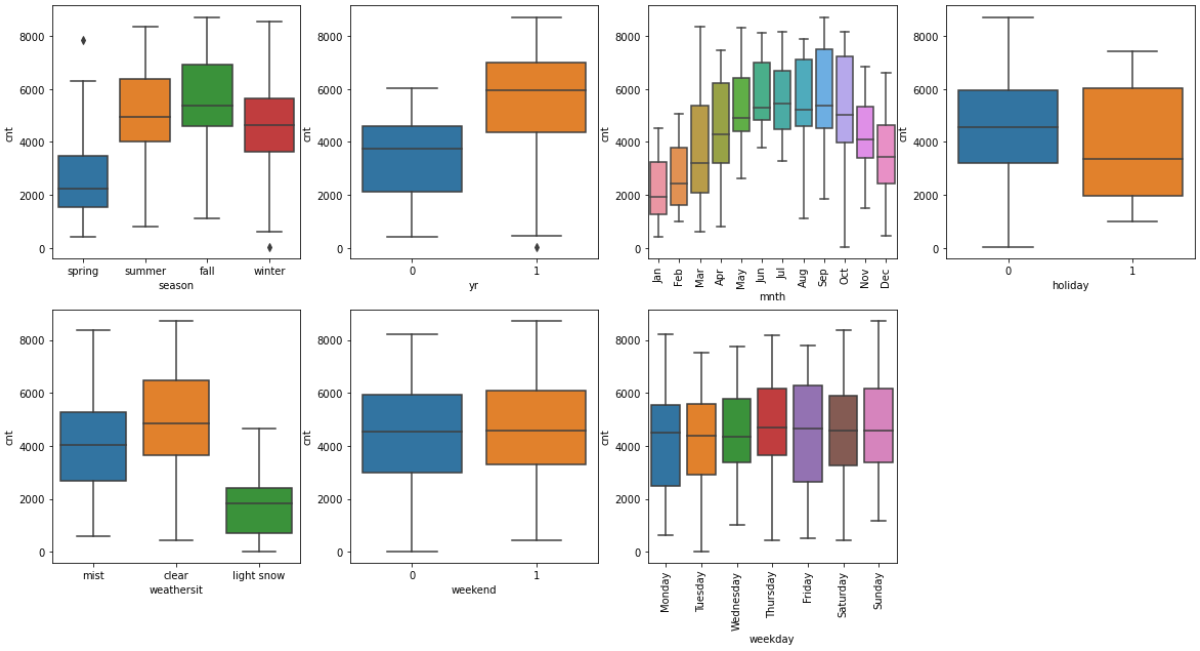
Answer:

Below, is my analysis of the effect of categorical variables on the dependent variable.

The categorical variables are:

i) season
ii) yr
iii) mnth
iv) holiday
v) weathersit
vi) weekday
vii) weekend

The last categorical variable, 'weekend' is derived from weekday.



i) Season seems to have an effect on demand. Spring has the lowest demand while Fall has the highest

ii) Clearly visible that demand grew with the year. 2019 has greater demand than 2018
iii) Also, demand grew mid-year. For the first two months, the demand was low. But, it shot up in March, stayed consistent until the last two months, when it came down again. This is clearly due to seasons.
iv) Holiday has a visible negative effect on demand.
v) light snow weather sees a decline in demand, maybe because people do not go out in the cold weather. Clear weather sees the most demand
vi) Weekend doesn't seem to have a considerable impact on demand.
vii) Demand is almost same on all weekdays

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

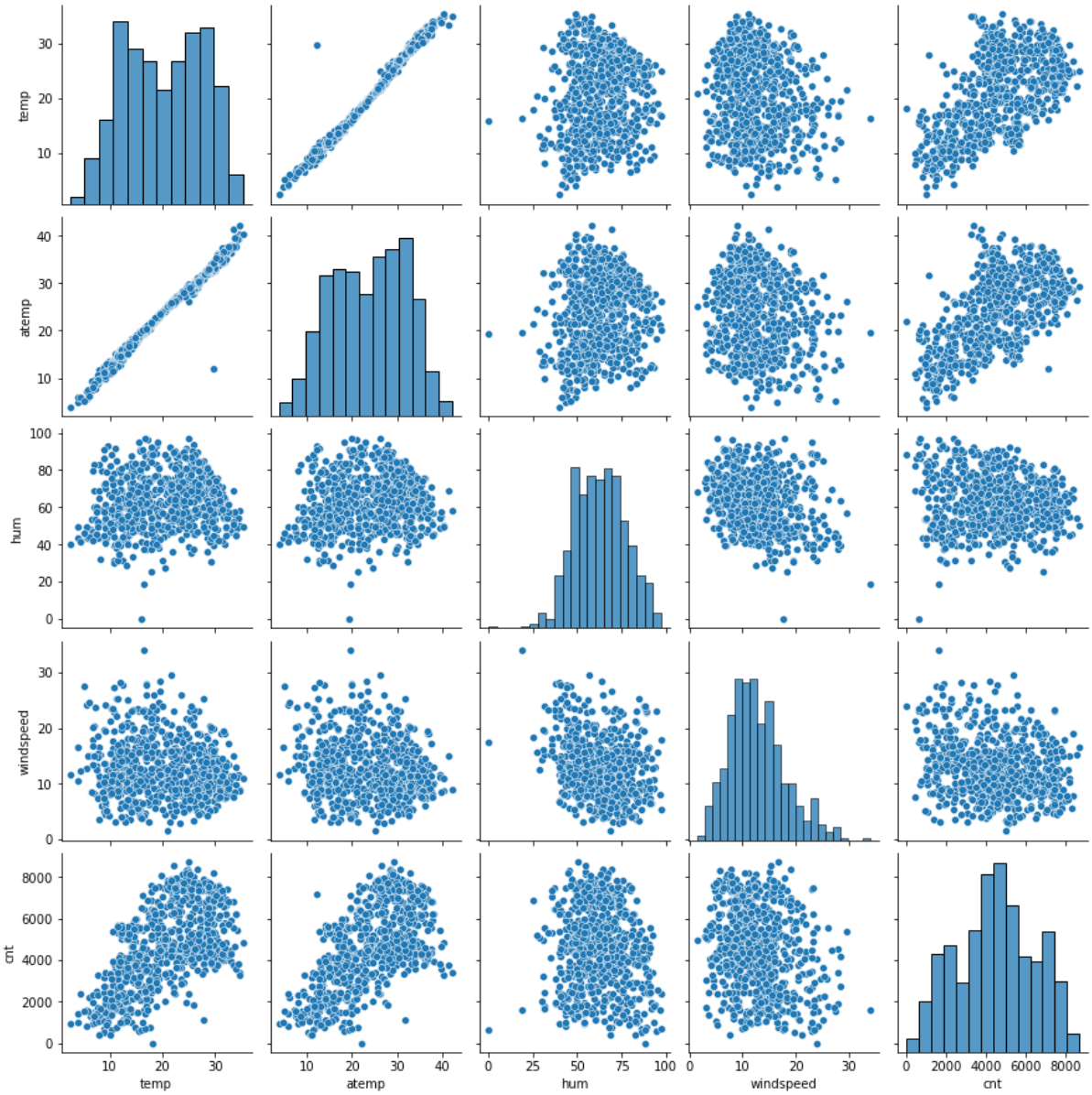
Answer:

Suppose, there is a categorical variable with 3 categories, Furnished, Semi-Furnished, and Unfurnished.
In order to be interpretable by the model, dummy variables need to be created.
So, we will have 3 dummy variables, which will be named after the 3 categories.
The variables will have the value 0(false) or 1(true). If two variables are 0, the third one has to be 1. So, the point is that one variable can be deduced by the other 2 variables. Hence, they become multicollinear.
To avoid multicollinearity, we use drop_first=True, to drop 1 variable out of the 3. Because, it is obvious that if both variables are 0, it is the third category we are talking about.
Multicollinearity is removed because it makes the coefficients of other predictor variables swing which makes it difficult to realize which predictor has more effect on the dependent variable.

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

Answer:

Here, it is clearly visible that temp and atemp have the highest correlation with cnt (target variable)



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

Answer:

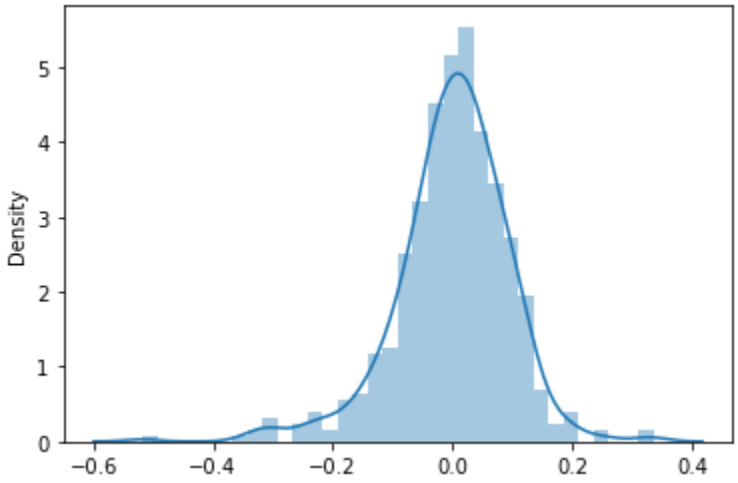
The assumptions of Linear Regression are:

i) Error terms are normally distributed with mean 0.

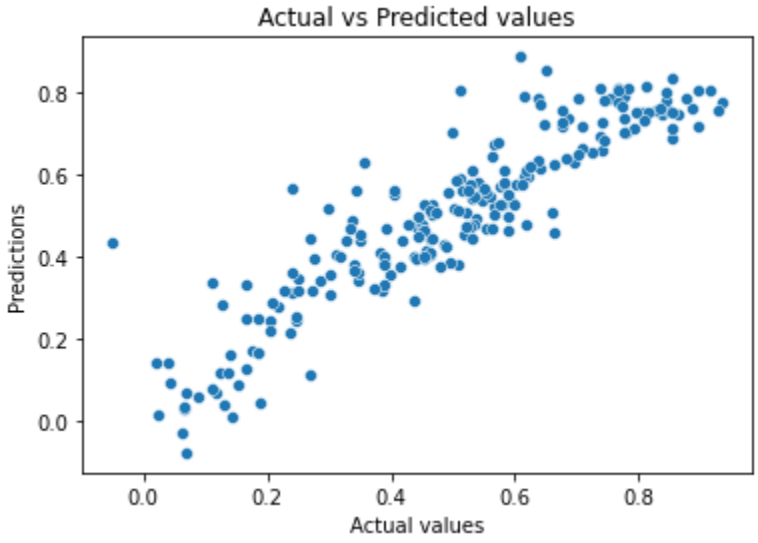
ii) There is a linear relationship between independent and dependent variables.

iii) Homoscedasticity: Residuals are distributed with constant variance.

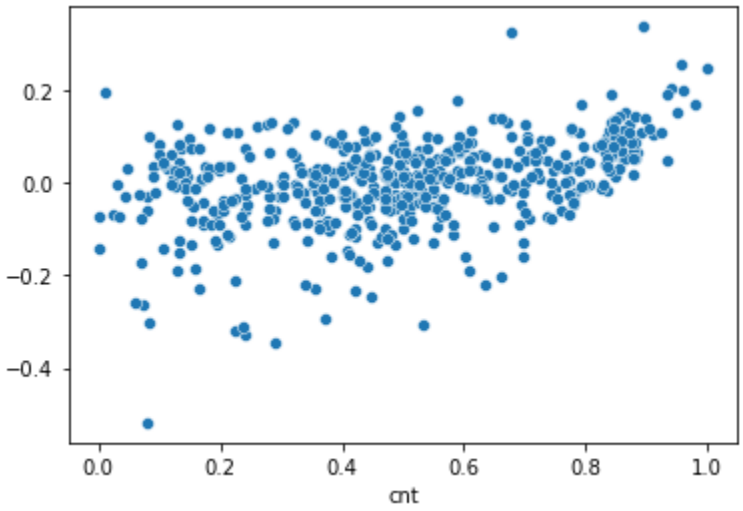
iv) There is no multicollinearity in the data.



This distplot of residuals show that the residuals are normally distributed with mean 0.



This scatterplot of actual vs predicted values shows the linear relationship.



This scatterplot of true values vs residuals shows that the residuals do not follow any pattern, are distributed with constant variance. Hence, they are homoscedastic.

	features	vif
3	windspeed	4.76
2	temp	4.70
5	spring	4.21
6	winter	2.60
12	Jan	2.29
13	Feb	2.14

	features	vif
0	yr	2.05
9	Nov	1.83
7	Dec	1.64
11	mist	1.55
4	weekend	1.39
8	Jul	1.32
10	light snow	1.10
1	holiday	1.06

The VIF (variance_inflation_factor) shows to what extent a variable is multicollinear. A VIF less than 5 is acceptable.

This proves that multicollinearity is not present in the data.

Another metric to prove this is the Durbin-Watson test. (d)

$1.5 < d < 2.5$ says that autocorrelation is likely not a cause for concern.

In this case, $d = 2.066$.

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

Answer:

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.839
Model:	OLS	**Adj. R-squared:**	0.835
Method:	Least Squares	**F-statistic:**	184.7
Date:	Mon, 01 Nov 2021	**Prob (F-statistic):**	3.78e-186
Time:	11:18:39	**Log-Likelihood:**	479.01
No. Observations:	510	**AIC:**	-928.0
Df Residuals:	495	**BIC:**	-864.5
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std errt**		**P>|t|**	**[0.025**	**0.975]**
const	0.2905	0.026	11.322	0.000	0.240	0.341
yr	0.2558	0.009	29.756	0.000	0.239	0.273
holiday	-0.0768	0.024	-3.206	0.001	-0.124	-0.030
temp	0.3752	0.033	11.516	0.000	0.311	0.439
windspeed	-0.1440	0.027	-5.257	0.000	-0.198	-0.090
weekend	0.0226	0.010	2.355	0.019	0.004	0.041
spring	-0.1033	0.019	-5.396	0.000	-0.141	-0.066
winter	0.0858	0.014	5.964	0.000	0.058	0.114
Dec	-0.0896	0.019	-4.825	0.000	-0.126	-0.053
Jul	-0.0652	0.019	-3.441	0.001	-0.102	-0.028
Nov	-0.0972	0.020	-4.940	0.000	-0.136	-0.059
light snow	-0.2633	0.025	-10.744	0.000	-0.311	-0.215
mist	-0.0858	0.009	-9.320	0.000	-0.104	-0.068
Jan	-0.0881	0.023	-3.857	0.000	-0.133	-0.043
Feb	-0.0594	0.023	-2.606	0.009	-0.104	-0.015

Omnibus:	85.220	**Durbin-Watson:**	2.066
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	248.585
Skew:	-0.795	**Prob(JB):**	1.05e-54
Kurtosis:	6.028	**Cond. No.**	15.0

This is the final model.

According to the coefficients of all the variables, the top 3 significant ones are:

i) Temp (coef = 0.3752) Positively affects the demand

ii) Light snow(coef = -0.2633) Negatively affects the demand.

iii) Yr (coef = 0.2558) Positively affects the demand

General subjective questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression may be defined as the statistical model that analyses 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 −

$$Y = mX + c$$

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

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

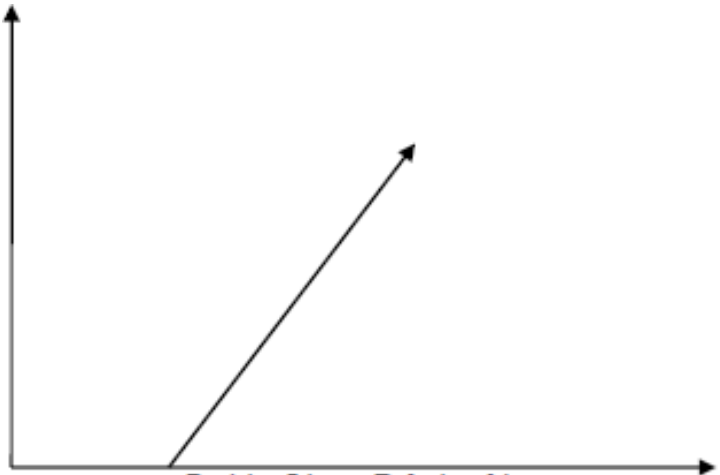
m is the slope of the regression line which represents the

effect X has on Y c is a constant, known as the Y-intercept.

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

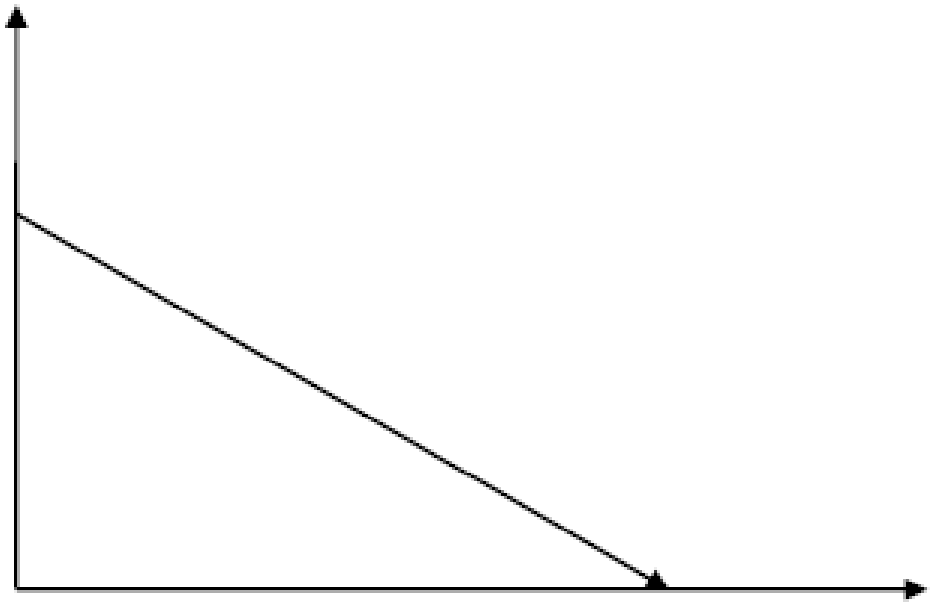
Furthermore, the linear relationship can be positive or negative in nature as explained below−

- 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 −



Positive Linear Relationship

- 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

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model −

ü Multi-collinearity –

- o Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

ü Auto-correlation –

- o Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

ü Relationship between variables –

- o Linear regression model assumes that the relationship between response and feature variables must be linear.

ü Normality of error terms –

- o Error terms should be normally distributed

ü Homoscedasticity –

- o There should be no visible pattern in residual value

2. Explain the Anscombe's quartet in detail.

Answer:

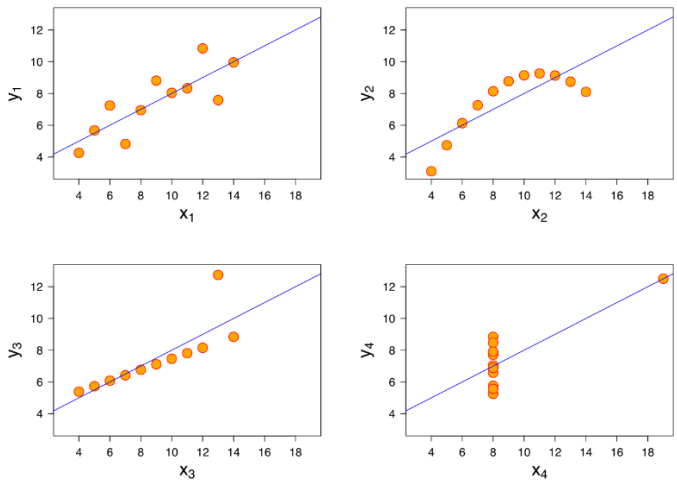
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four
datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

are graphed. Each graph tells a different story irrespective of their similar summary statistics.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

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

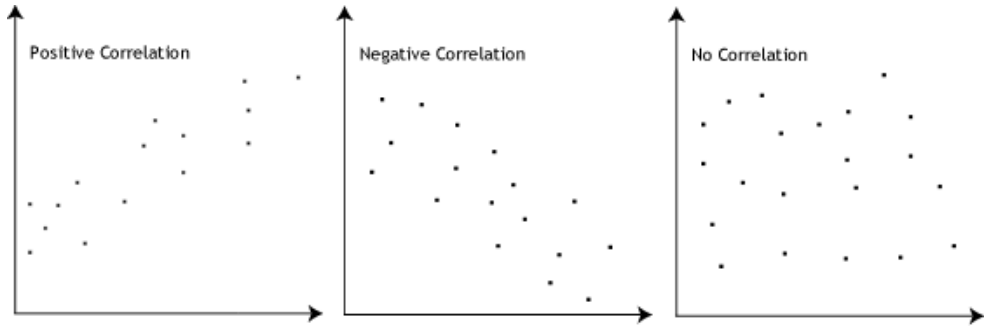
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed 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.

Example**:** If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. 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.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.