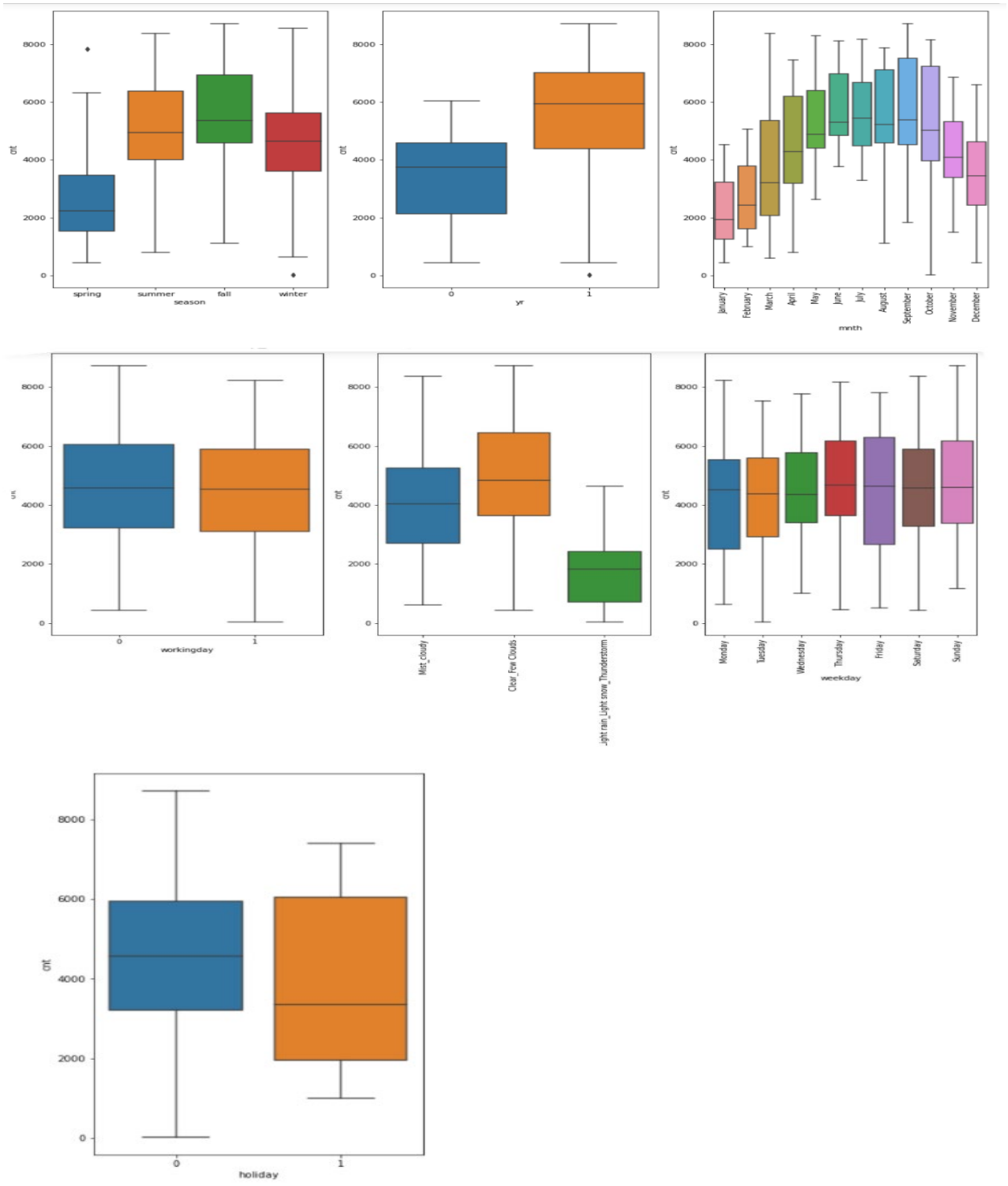
Assignment-based Subjective Questions

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

Answer: As represented the categorical variables affect on the Dependent Variable.



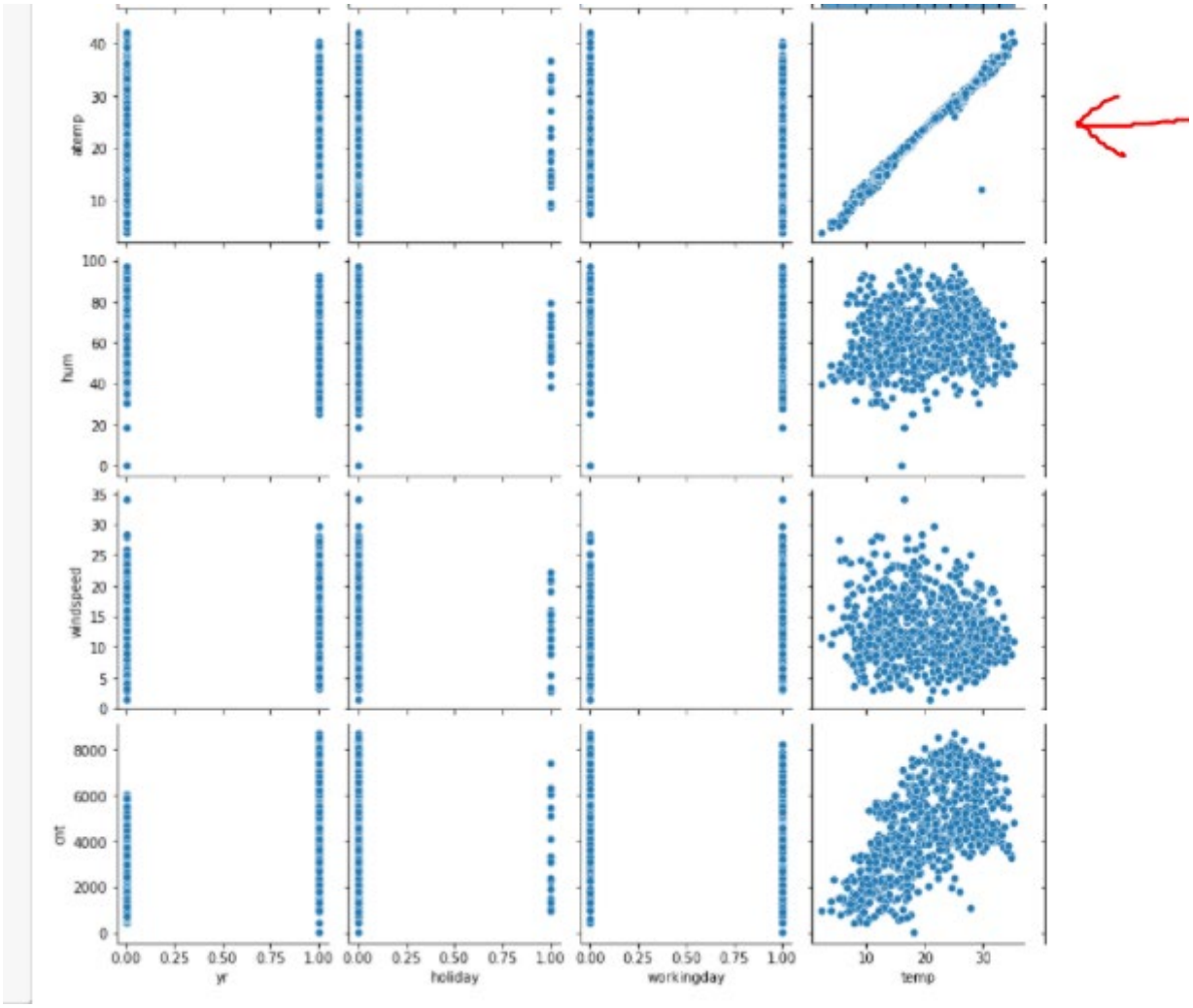
For example for during the Summer, Fall and Winter season the count of total rental bikes is very high compared to Spring Season.

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

Answer: It is important to do drop_first because when we are creating the dummy data for the Categorical variables we only need N-1 number of frequencies from that Categorical variable not all the N frequencies therefore it is best to drop it at the time of creation of the dummy variables.

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

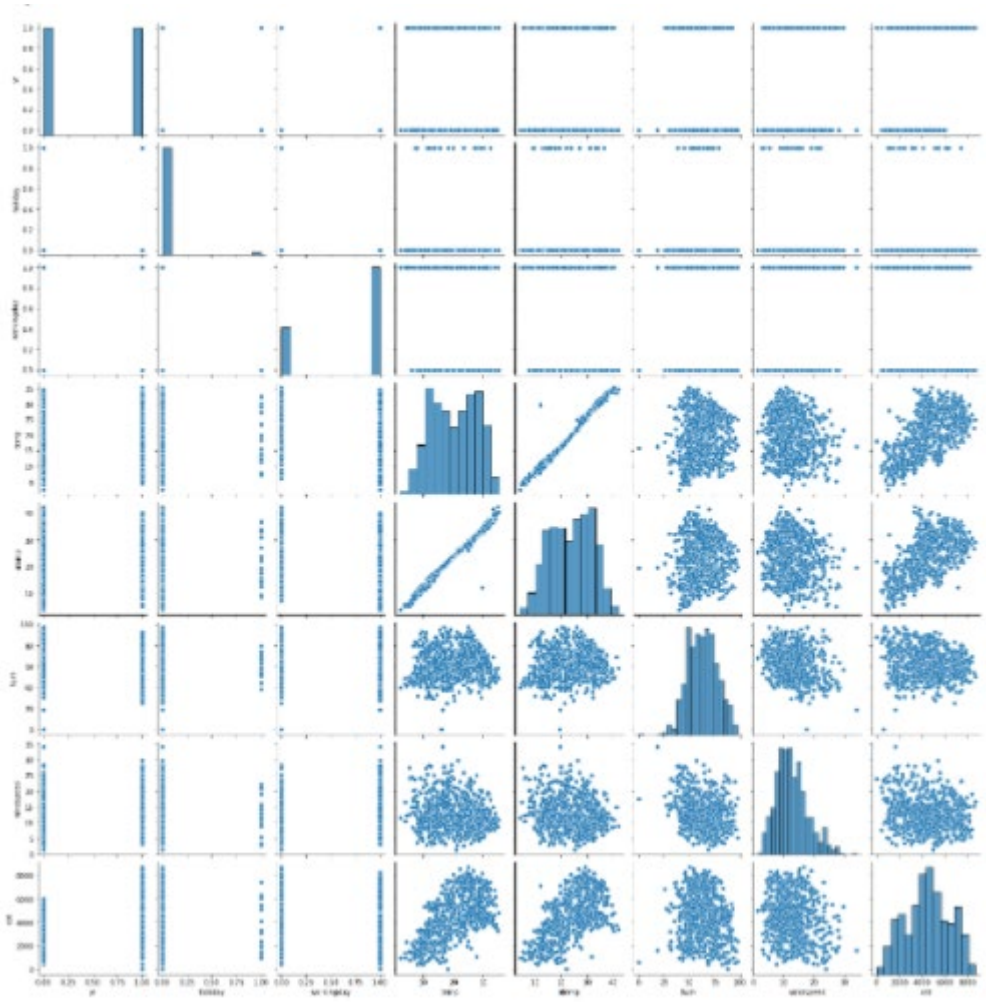
Answer:



The Temp and Atemp have the highest correlation of 0.99 among all the Numerical Variables.

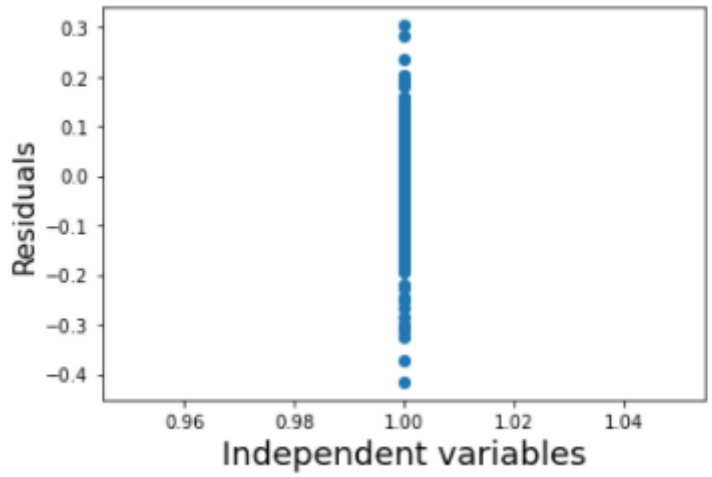
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: a) Linear Relationship between the features and target.

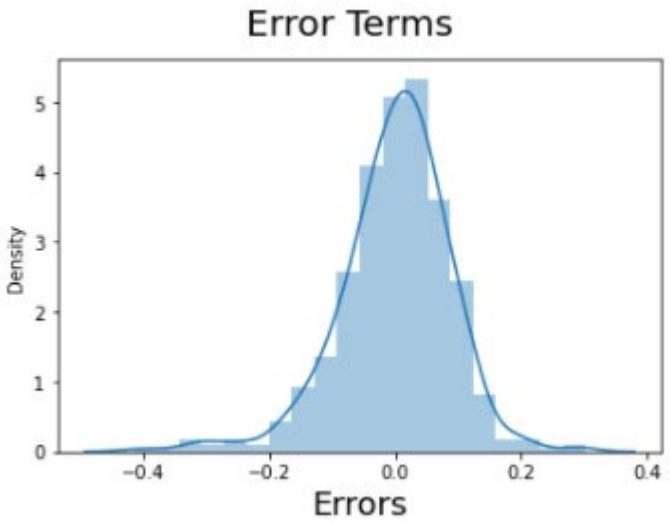


b) Little or no Multi Co-linearity between the features.

c) Homoscedasticity Assumption



d) Normal distribution of error terms.



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

Answer: Year, Atemp and Winter (Seasons) Column are affecting the Count most.

General Subjective Questions

1. 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 −

$Y = mX + b$

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

X is the dependent 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 Y-intercept. If X = 0,Y would be equal to b.

A linear relationship will be called positive if both independent and dependent variable increases. A linear relationship will be called positive if independent increases and dependent variable decreases.

Linear regression is of the following two types –

- Simple Linear Regression:
  It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.

  $Y = mX + b$

- Multiple Linear Regression:
  It is the extension of simple linear regression that predicts a response using two or more features.

  $h(x_i)=b_0+b_1x_{i1}+b_2x_{i2}+...+b_px_{ip}+e_i$

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

  Multi-collinearity – 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.

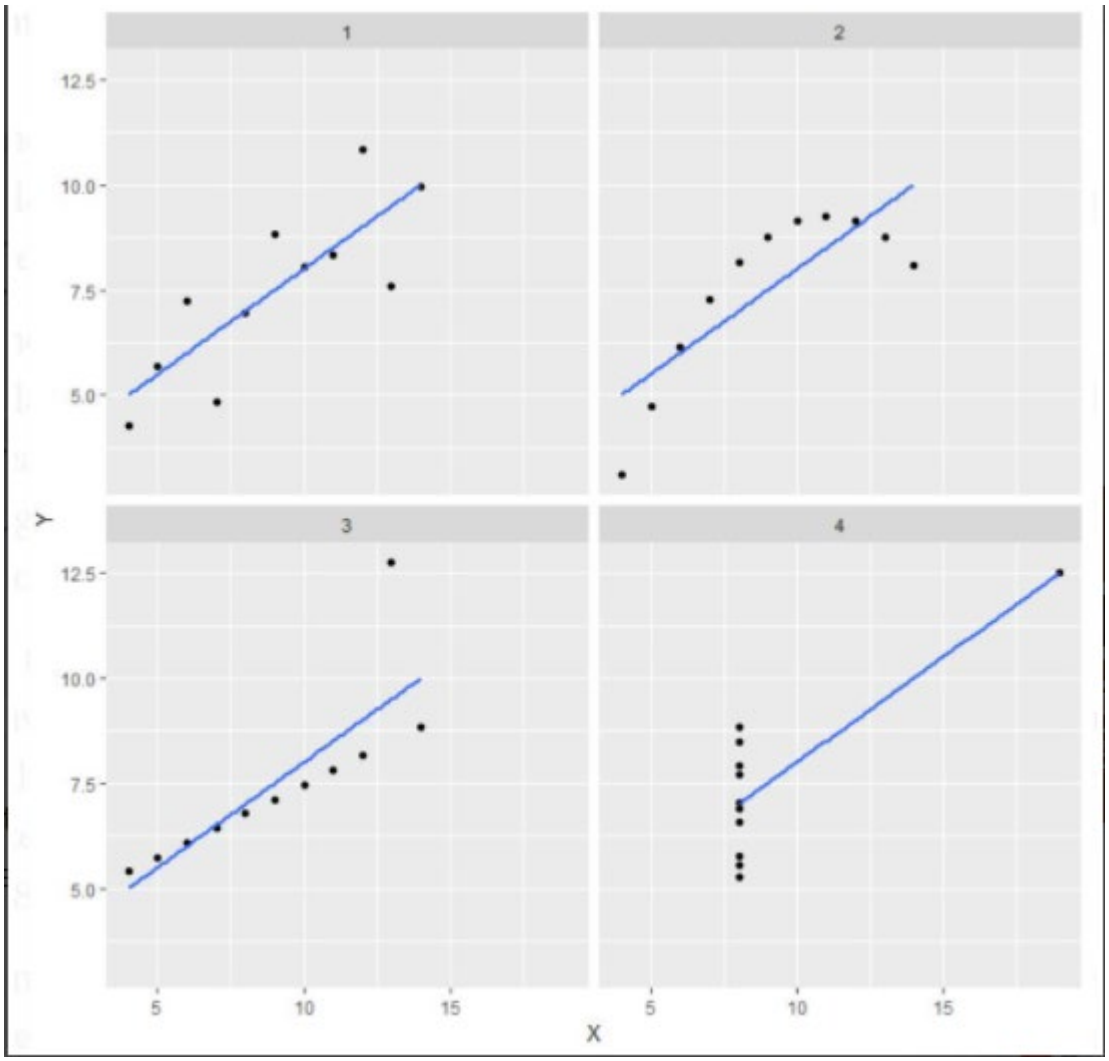
  Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

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

2. Explain the Anscombe's quartet in detail.

Answer: **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

	A	B	C	D	E	F	G	H	I
1	x1	x2	x3	x4	y1	y2	y3	y4	
2	10	10	10	8	8.04	9.14	7.46	6.58	
3	8	8	8	8	6.95	8.14	6.77	5.76	
4	13	13	13	8	7.58	8.74	12.74	7.71	
5	9	9	9	8	8.81	8.77	7.11	8.84	
6	11	11	11	8	8.33	9.26	7.81	8.47	
7	14	14	14	8	9.96	8.1	8.84	7.04	
8	6	6	6	8	7.24	6.13	6.08	5.25	
9	4	4	4	19	4.26	3.1	5.39	12.5	
10	12	12	12	8	10.84	9.13	8.15	5.56	
11	7	7	7	8	4.82	7.26	6.42	7.91	
12	5	5	5	8	5.68	4.74	5.73	6.89	
13									



- In the first one(top left) the scatter plot there seems to be a linear relationship between x and y.
- In the second one(top right) there is a non-linear relationship between x and y.
- In the third one(bottom left) there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

Answer: In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures

the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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

Answer: In scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1].

$x' = (x - x_{min})/(x_{max} - x_{min})$

where x' is the normalized value.
Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important. For example, in the dataset containing prices of products; without scaling, SVM might treat 1 USD equivalent to 1 INR though 1 USD = 65 INR.

a) Standardization *(also called **z-score normalization**)* transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

$x' = (x - x_{mean})/\sigma$

where x is the original feature vector, xmean is the mean of that feature vector, and σ is its standard deviation.
where μ is the mean. By subtracting the mean from the distribution, we're essentially shifting it towards left or right by amount equal to mean i.e. if we have a distribution of mean 100, and we subtract mean 100 from every value, then we shift the distribution left by 100 without changing its shape. Thus, the new mean will be 0. When we divide by standard deviation σ, we're changing the shape of distribution. The new standard deviation of this standardized distribution is 1 which you can get putting the new mean, μ=0 in the z-score equation.

b) Normalization normalizing data.

$x' = (x - x_{mean})/(x_{max} - x_{min})$

For normalization, the maximum value you can get after applying the formula is 1, and the minimum value is 0. So all the values will be between 0 and 1.

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

Answer:

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$X_1=C+ \alpha_2 X_2+\alpha_3 X_3+\cdots$

$〖VIF〗_1=1/(1-R_1^2)$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$X_2=C+ \alpha_1 X_1+\alpha_3 X_3+\cdots$

$〖VIF〗_2=1/(1-R_2^2)$

If all the independent variables are orthogonal to each other, then VIF = 1.0. **If there is perfect correlation, then VIF = infinity.** A large value of VIF indicates that there is a correlation between the variables.

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

Answer:

The quantile-quantile plot or Q-Q plot is a graphical tool to validate if two datasets are coming from populations with common distribution.
Very often we assume given data to be normally distributed for ease of inferring useful information. One way to assess our assumption's correctness is to use Q-Q plot. Not just Normal distribution, we can test for other distributions (for eg. uniform distribution etc.) as well.

Quantiles are the breakpoints that divide the ordered numerical data into equal sized bins.

Percentiles are a type of quantiles that divide the data into 100 equal bins, quartiles divide the data into 4 equal parts and so on.

Q-Q plot compares the quantiles of 2 datasets. We can make Q-Q for any 2 datasets as long as the quantiles can be calculated for both of them.

Importance of Q-Q plot in Linear Regression:

1. Two datasets/sample can be of different size.

2. Q-Q plot can detect outliers, shifts in scale, location, symmetry etc. simultaneously.

3. One of the important assumptions of Linear Regression is that the residual of the model is normally distributed. This can be assessed using Q-Q plot.

Example of Q-Q plot:
Here, first 5000 normally distributed random points are generated.

Then the random points are fed into the Q-Q plot. The blue dots representing the random points are aligning with 45 degree reference straight line in red. This re-confirms the test_data is actually normally distributed. (Left figure)

As test_data is shifted in location, the blue dots have shifted on the left side of the 45 degree reference line. Thus Q-Q plot can show different statistical aspects. (Right figure)

```
import numpy as np
import statsmodels.api as sm
from matplotlib import pyplot as plt

test_data = [np.random.normal() for i in range(5000)]
sm.qqplot(np.array(test_data), line='45')
plt.title("Q-Q plot")
plt.show()
```

```
import numpy as np
import statsmodels.api as sm
from matplotlib import pyplot as plt

test_data = [np.random.normal() for i in range(5000)]
sm.qqplot(6 + np.array(test_data), line='45')
plt.title("Effect of shift in Location on Q-Q plot")
plt.show()
```