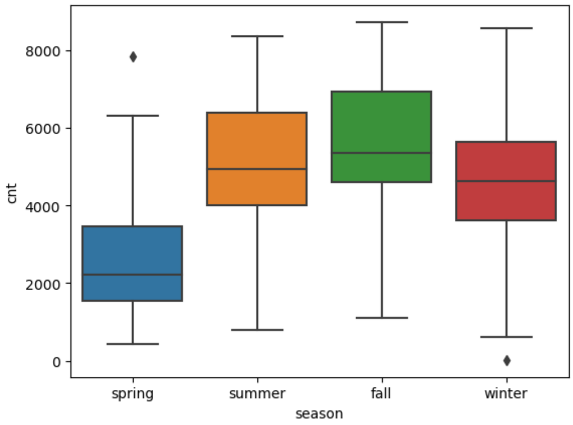
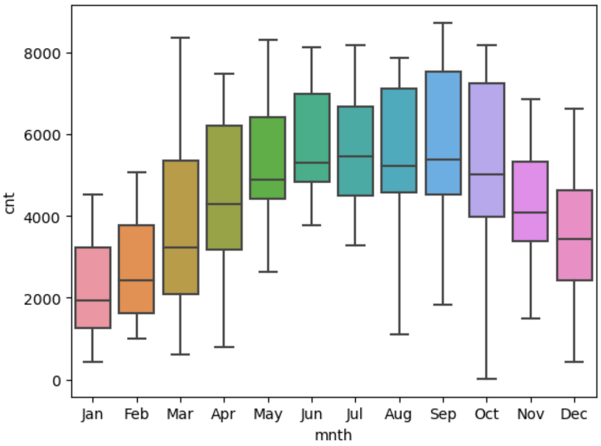
Assignment-based Subjective Questions

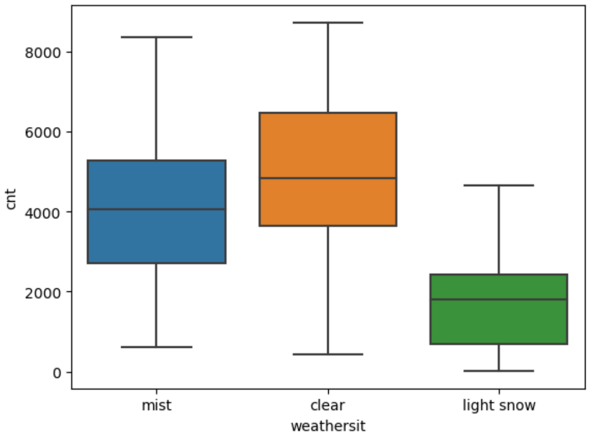
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



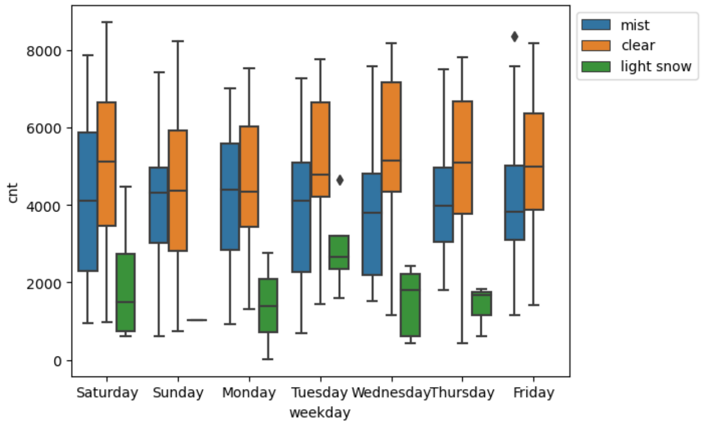
From the boxplot it can be observed that the count of bike rentals is higher for summer and fall season in comparison to winter and spring seasons.

From the boxplot it can be observed that the highest median count of bike rentals is in the month of July and the lowest median count is in January.

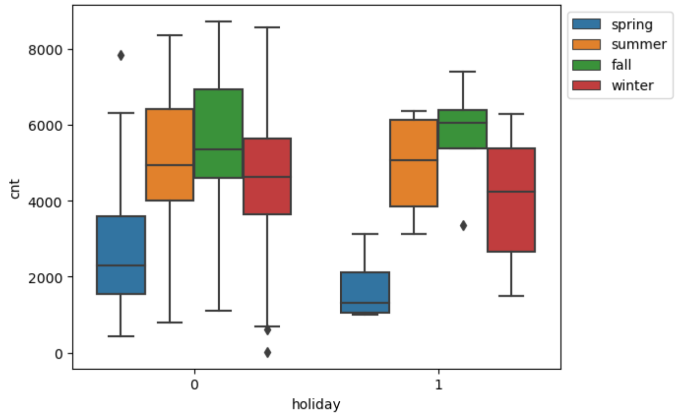




From the boxplot it can be observed that when there is a Light Snow, Light Rain and Thunderstorm and Scattered clouds or Light Rain and Scattered clouds is the median count of bike rentals is the lowest.

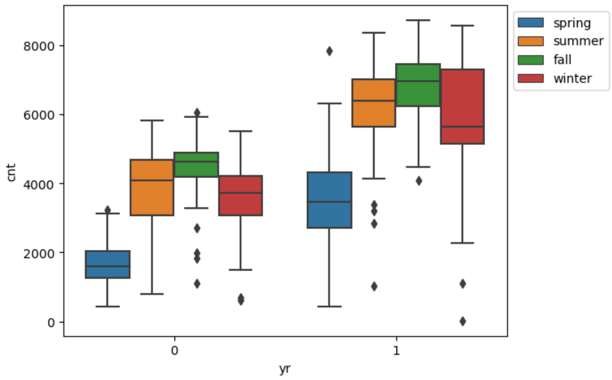


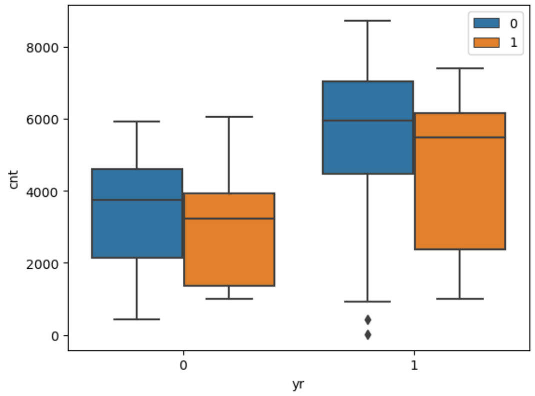
From the boxplot it can be observed that bike rental median count is higher on light snow days if it is not a weekend.



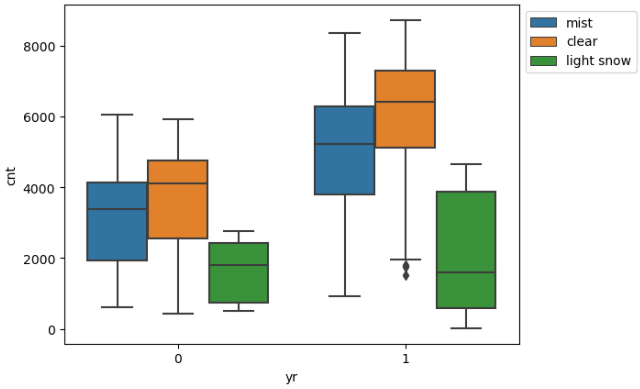
From the boxplot it can be observed that in winter and spring season holidays have a lesser median count of bike rental while in summer and fall seasons the median count of bike rentals is higher on holidays.

From the boxplot it can be observed that irrespective of the season the bike rental median count has increased from 2018 to 2019.

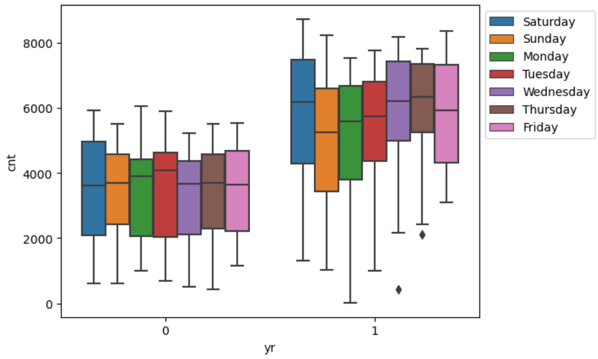




From the boxplot it can be observed that irrespective of whether it is a holiday or not the bike rental median count has increased from 2018 to 2019.

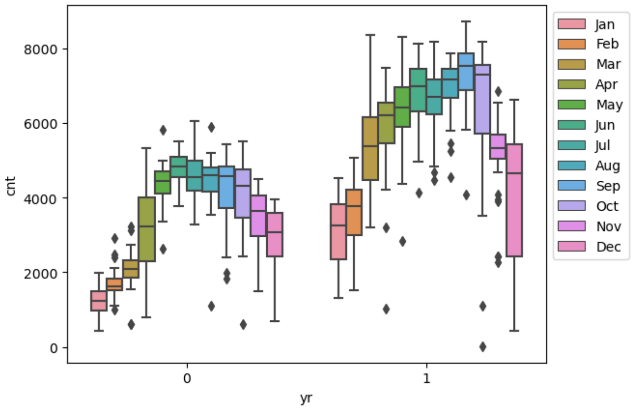


From the boxplot it can be observed that apart from light snow weather condition the bike rental median count has increased from 2018 to 2019 for all other weather situations.



From the boxplot it can be observed that irrespective of the weekday the bike rental median count has increased from 2018 to 2019.

From the boxplot it can be observed that irrespective of the month the bike rental median count has increased from 2018 to 2019.



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

Dummy variables are variables which can take either 0 or 1 as their values. Categorical variables are converted to dummy variables to understand their importance in regression analysis. drop_first = True is important because it ensures that there is no multicollinearity also this ensures that we are using one less feature to evaluate the model by treating it as abase condition so that the regression analysis is easier and less time consuming.

The multicollinearity issue comes because:

$$y = \begin{bmatrix} 65 \\ 50 \\ 70 \\ 120 \\ 80 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 5 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 1 \\ 1 & 5 & 1 \\ 1 & 15 & 0 \end{bmatrix}$$

Consider the above case here the last column in X is a categorical feature having two levels either Yes or No, so if we were to create 2 dummy variables the resulting X matrix will be as below:

$$X = \begin{bmatrix} 1 & 5 & 0 & 1 \\ 1 & 2 & 0 & 1 \\ 1 & 2 & 1 & 0 \\ 1 & 5 & 1 & 0 \\ 1 & 15 & 0 & 1 \end{bmatrix}$$

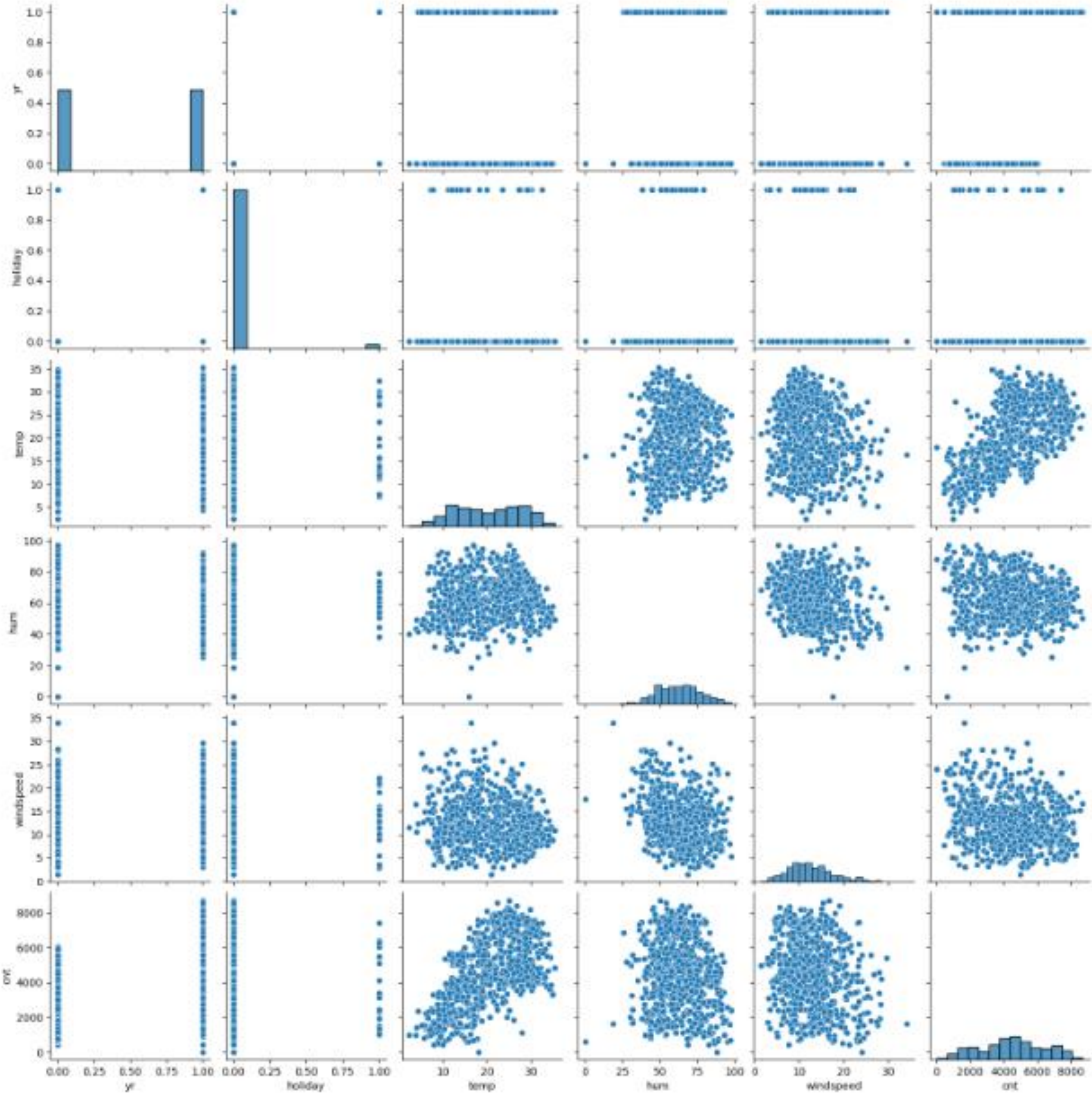
Thus column 1 can be written as a linear combination of column 3 and column 4 as below:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

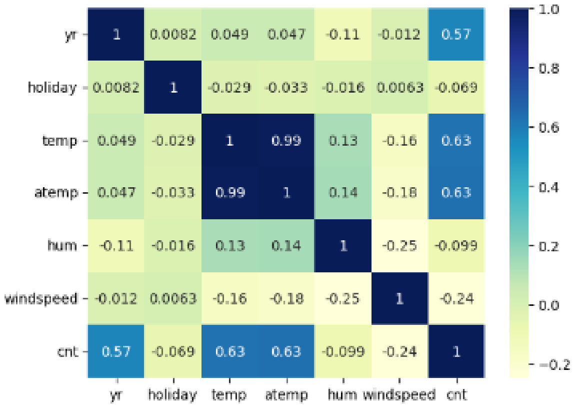
Thus this will create issues as the matrix will become singular and we won’t be able to calculate inverse thus making it difficult to use ordinary least squares method to find the best fit line.
This is the reason to opt for drop_first = True while creating dummy variables.

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

temp has the highest correlation with the target variable cnt as it can be seen from both the correlation heatmap and the pairplot.



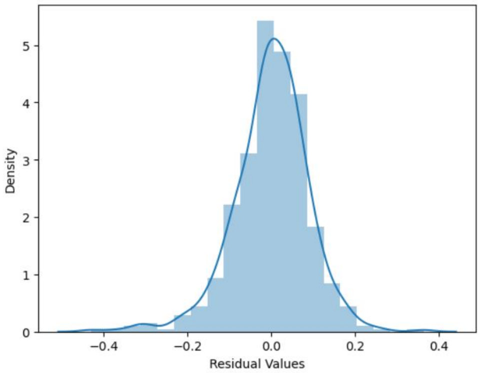
Pair Plot



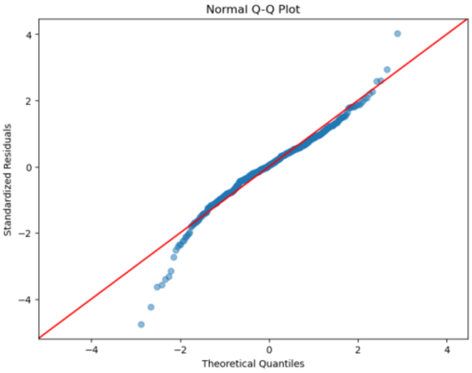
Correlation Heatmap

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

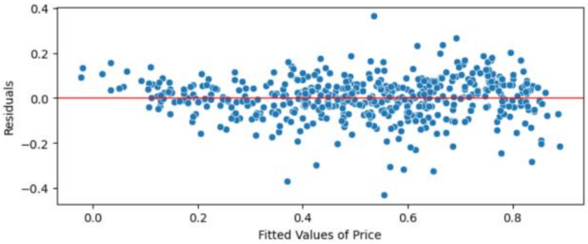
1. The residuals or error terms must follow a normal distribution with mean value as 0.



2. QQ plot to be plotted for checking whether the residuals are adhering to theoretical normal distribution.



3. Homoscedasticity should be observed.



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

From the final model the top 3 features contributing significantly were:

1. temp having coefficient: + 0.4502
2. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds weather situation having coefficient: − 0.2916
3. Yr having coefficient: + 0.2340

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm is a type of supervised machine learning algorithm. When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

The aim of the algorithm is to find the best linear equation for the observed data such that the sum of the squares of the residuals is the minimum.

There are few assumptions which are important for the usage of linear regression:

1. There should be a linear relationship between the predictor variables and the dependent variable.
2. The residuals should be normally distributed with mean 0.
3. The residuals are independent of each other.
4. The residuals should have a constant variance.

The equation for simple linear regression is as below:

$y=\beta_0+\beta_1X$

where β_0 is the intercept, β_1 is the slope or coefficient, X is the independent variable and y is the dependent variable.

The equation for multiple linear regression is as below:

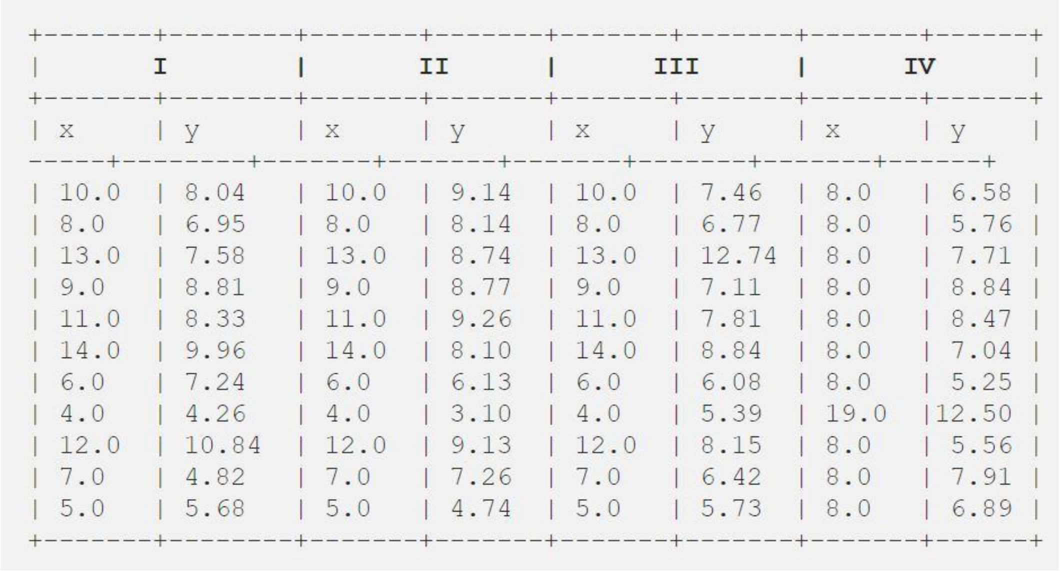
$y=\beta_0+\beta_1X_1+\beta_2X_2+………\beta_nX_n$

where β_0 is the intercept, β_1, β_2,… β_n *are* the slopes or coefficients, X_1, X_2,… X_n are the independent variablea and y is the dependent variable.

2. Explain the Anscombe's quartet in detail.

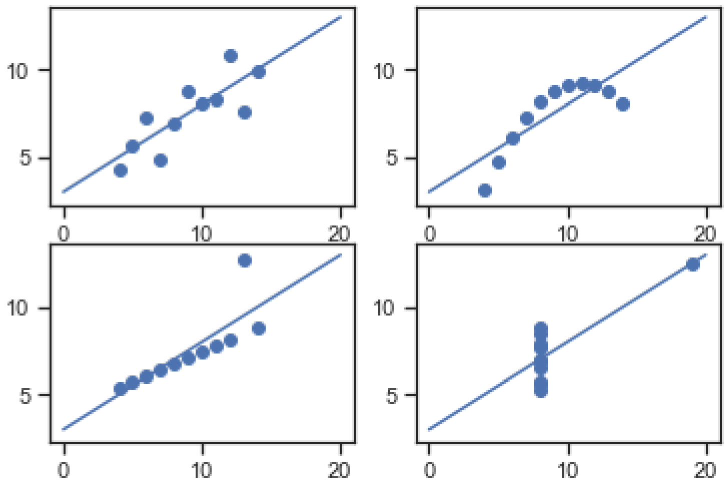
Anscombe's Quartet comprises four datasets with nearly identical summary statistics. The purpose of Anscombe's Quartet is to show the importance of Exploratory Data Analysis and data visualization for understanding data. The datasets, have identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but they have different representations when plotted on a graph.

The Datasets:



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Graphical Representation:



3. What is Pearson’s R?

Pearson’s R (or Pearson correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by r. Pearson’s R, indicates how far away all the data points are to the line of best fit (i.e., how well the data points fit this new model/line of best fit).

Pearson’s 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.

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

Scaling is a technique used to transform the values of features to a similar scale. It is done to ensure that all the features are on the same scale and to avoid giving biased importance to features having larger values. It is used when the dataset used has features which have different ranges, units of measurements or order of magnitude. Scaling helps in smooth performance of the algorithm by improving the model convergence. It also helps in meaningfully comparing the features as all of them are on the same scale. It also removes the effect of higher magnitudes thereby preventing the model to be biased towards features having higher magnitudes.

Normalized Scaling:

Normalization or Min-Max scaling is a scaling technique in which values are rescaled such that they range between 0 and 1.

The formula for normalization:

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

Standardized Scaling:

Standardization is a scaling method 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.

The formula for standardization:

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

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

The formula for VIF (Variance Inflation Factor) is as below:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

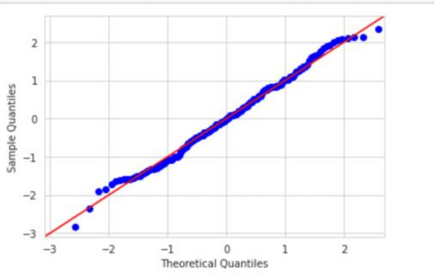
Where R_i^2 stands for the R2 score of the linear combination of ith variable in terms of other predictor variables except the target variable.

Now when VIF is infinite it means that R_i^2 is 1, which means that the ith predictor variable can be completely explained by a linear combination of all the other predictor variables except the target variable and thus can be removed from the set of predictor variables as it will create multicollinearity and the values of the coefficients will swing.

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

A Q-Q plot means Quantile-Quantile plots. It is a plot of the quantiles of a sample distribution against quantiles of a theoretical distribution.

It is used to determine the type of distribution of the sample. This is important in Linear Regression as one of the assumptions of Linear Regression is that the error terms follow a normal distribution. If the theoretical quantile plot and the sample quantile plot coincide it means that the assumption holds and thus the obtained model is linear. If they do not coincide then there is some deviation and the model needs to be reevaluated.



An example of a Q-Q plot of sample distribution vs theoretical distribution. The above example shows that the distribution of the sample is normal.