Bike Sharing Case Study

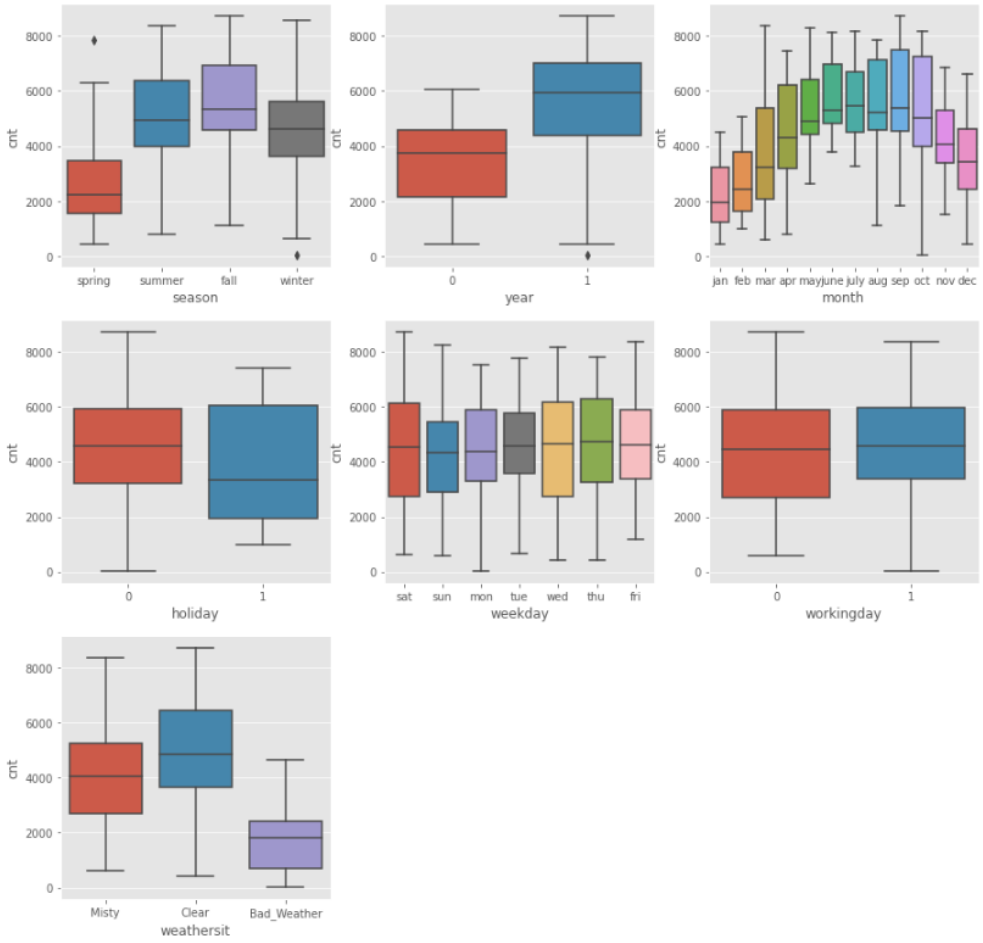
Linear Regression + Subjective Questions

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?

Ans:

- Fall and Summer are the best seasons with Spring being the lowest in demand.
- Demand picked up drastically in 2019. Possible uptick as we see things improve in later years.
- Demand picks up as we move into the middle of the year with the peak in September. End of the year sees the demand fall.
- Lower demand on holidays.
- Bad weather impacts demand adversely.
- Daily demand improves slightly as we move into the week with a slight dip on Sunday.
- Demand picks up during working days. As things normalize may be people are using bikes to take bike rides for health reasons. Also, bikes could be used for office commute for people who are allowed into the office.



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

Ans:

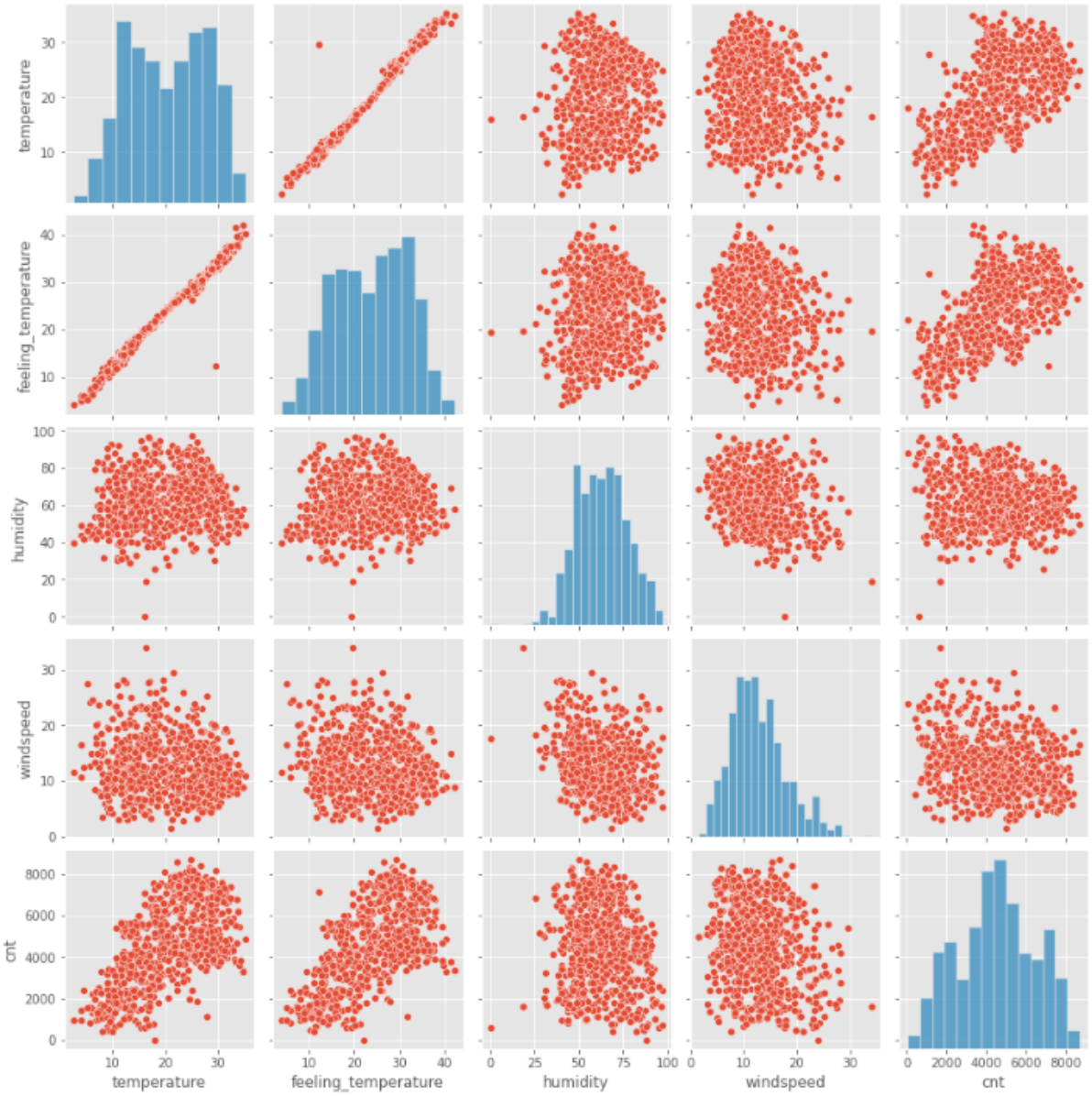
To encode categorical data, one hot encoding is done, where a dummy variable is to be created for each discrete categorical variable for a feature. This can be done by using pandas.get_dummies () which will return dummy-coded data. Here we use parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.

The drop_first is used in order to reduce redundancy in variable creation as the model should be able to understand the pattern based upon n-1 variables. It will also avoid multi-collinearity as these dummy variables are themselves correlated. The drop_first approach will also save a little bit of computational power and model complexity as you would have fewer variables to work with.

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

Ans:

Looking at the below pair-plot we can see that Temperature and Feeling Temperature has highest correlation with the target variable.

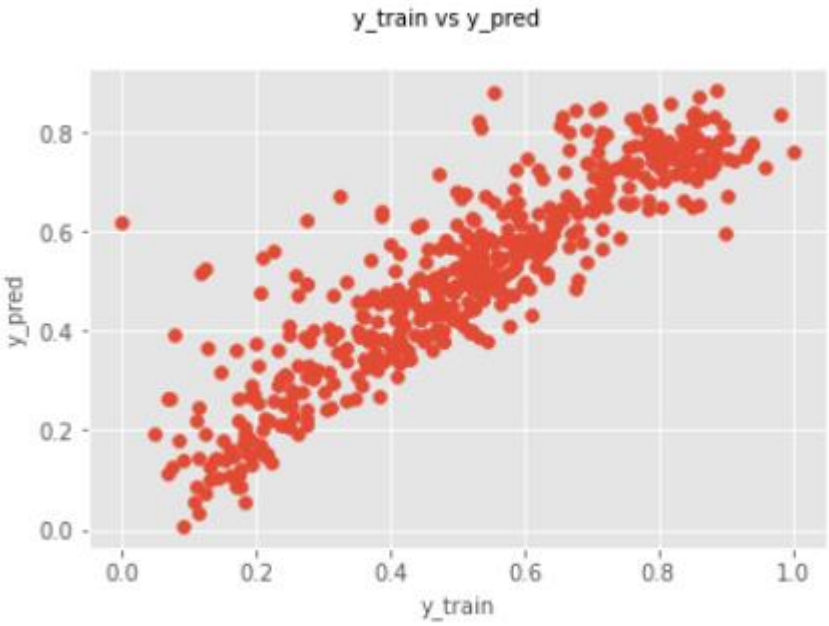


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

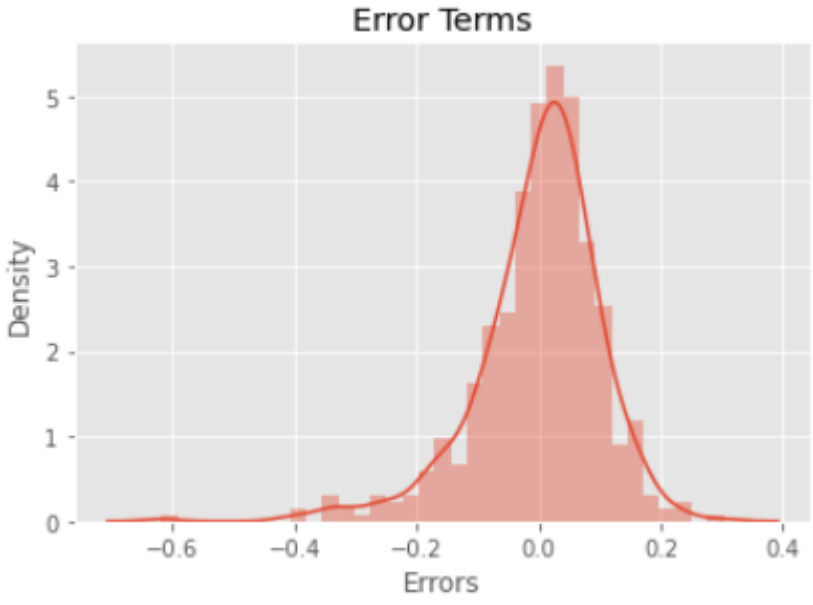
Ans:

The two major assumptions to be validated are:

- Fitted line is linear: This can be done by plotting actual and predicted values for the train data.



- Residual analysis: The error terms are normal and distributed with mean 0 this can be done by calculating the residuals and plotting a histogram if the same.



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

Ans:

Based upon the final model and magnitude of the co-efficient we can say that temperature, year, and winter season are the 3 top features contributing significantly towards the demand if shared bikes.

Variable	Co-efficient
temperature	0.5514
year	0.2388
season_winter	0.1161

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a supervised machine learning algorithm used for predicting a continuous numeric output (dependent variable) based on one or more continuous or categorical input features (independent variables). It models the linear relationship between the features and the target variable by fitting a straight line to the observed data points.

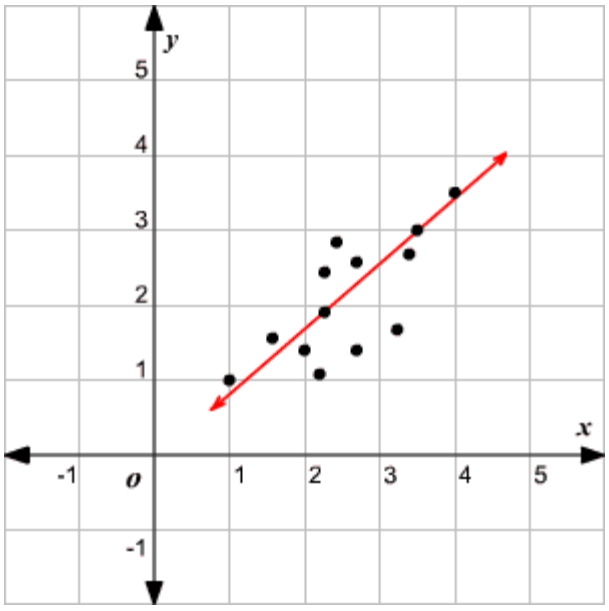
The idea of the linear regression algorithm is to compute a linear relationship between input and output through a simple straight-line equation denoted by:

$$Y = mx + c$$

Where:

- Y is the dependent variable (output)
- x is the independent (input)
- (m) (slope) determines how steep the line is. Which can be interpretated as the magnitude and polarity (+ve or -ve) a unit change in the independent variable will have on the dependent variable.
- (c) (intercept) determines the value of the function at (x=0). Which can be interpretated as value the output will have if the dependent variable is 0.

The line will be considered as **best fit** if the points of the line which are the predictions of the dependent variable arrived through the equation are as close to the actual points as possible.



To achieve the best-fit regression line, the model aims to predict the dependent value such that the error difference between the predicted value and the actual value is minimum.

This calculation of error is called as the **cost function** or the loss function. The Linear Regression algorithm applies the **Mean Squared Error** or **Residual Sum of Squares** as the cost function which is nothing but the average value of the squared errors between predicted and actual values. The MSE is given as:

$$\frac{1}{n}\Sigma\underbrace{\left(y-\hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

So, the idea here is to optimize the values of m and c such that the MSE is minimum. This is normally achieved through a method called **gradient descent**. Gradient descent algorithm iteratively updates the values of m and c to minimize MSE so that we get accurate best-fit line.

The linear regression algorithm involving multiple independent variables used for predicting the dependent variable is called multiple linear regression.

The equation is represented by:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

Here β0, β1… βn are the coefficients of intercept and X1…Xn independent variables respectively.

Assumptions of Linear Regression:

Linearity – The algorithm assumes that there is a linear relationship between the dependent and independent variables.

Independence – The observations of the dataset are independent of each other. This means the dependent variable values does not depend on previous values.

Homoscedasticity – This means the variance of the error terms across all variables is constant and that the number of variables does not impact the variance as that will lead to an inaccurate model.

Normality – This assumes that the residuals are normally distributed with mean 0 as different distribution will lead to inaccurate model.

No multicollinearity – There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it difficult to determine the individual effect of each variable on the dependent variable. If there is multicollinearity, then multiple linear regression will not be an accurate model.

Additivity – The model assumes that the effect of changes in a predictor variable on the response variable is consistent regardless of the values of the other variables. This assumption implies that there is no interaction between variables in their effects on the dependent variable.

Evaluation Metrics:

There are different measures used to evaluate model accuracy we have seen the one widely used which is:

Mean Squared Error: is an evaluation metric that calculates the average of the squared differences between the actual and predicted values for all the data points. The difference is squared to ensure that negative and positive differences don't cancel each other out.

$$\frac{1}{n}\sum\underbrace{\left(y-\hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

Mean Absolute Error: It is similar to the above metric the only change is the absolute term is used instead of the squared term.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - y_j|$$

Root Mean Squared Error (RMSE): The square root of the residuals' variance is the Root Mean Squared Error. It describes how well the observed data points match the expected values, or the model's absolute fit to the data.

RSME is not as good of a metric as R-squared. Root Mean Squared Error can fluctuate when the units of the variables vary since its value is dependent on the variables' units (it is not a normalized measure).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

Co-efficient of Determination R^2: This is a metric which explains how much of the total variation in the dependent variable is explained by the independent variable. It is always in the range of 0 to 1. In general, better the model better is the R^2.

Formula

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

Adjusted R^2: This is the same metric as above however it considers the number pf predictors added to the model and in a way penalizes the addition of redundant/irrelevant variables.

$$R^2_{adjusted} = 1 - (1 - R^2)\left(\frac{n-1}{n-m-1}\right)$$

In which $R^2_{adjusted}$ = the adjusted multiple correlation coefficient
R^2 = the original multiple correlation coefficient
n = the number of cases
m = the number of variables

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet is a set of four datasets, each consisting of eleven (x, y) points. Despite having nearly identical descriptive statistics, these datasets exhibit different distributions and appear distinct when graphed.

These datasets were constructed by Francis Anscombe in 1973

The datasets are as follows:

Data Set I:

x-values: 10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0

y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Appears to have a simple linear relationship.

Data Set II:

x-values: 10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0

y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

Non-linear relationship.

Data Set III:

x-values: 10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0

y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 9.13, 6.42, 5.73

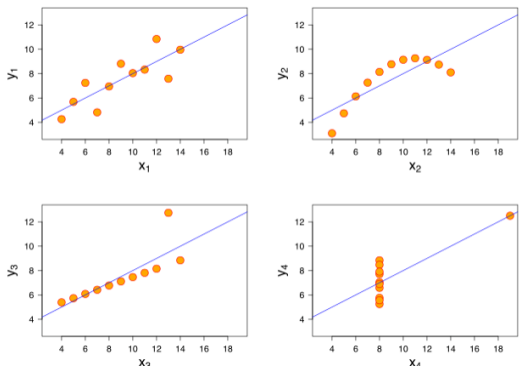
Linear relationship, but an outlier affects the regression line.

Data Set IV:

x-values: 10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0

y-values: 6.58, 5.76, 8.0, 8.84, 8.47, 7.04, 5.25, 5.39, 8.15, 6.89

Outlier point produces high correlation coefficient despite other data points showing no clear relationship.



For all the 4 datasets the following statistics are the same:

Statistic	Value
Mean of X	9
Sample variance of X	11
Mean of Y	7.50
Sample variance of Y	4.125
Correlation between x and y	0.816
Linear Regression Equation	y = 3.00 + 0.500x
R^2 Value	0.67

These is datasets are important as they show the importance of analyzing the data visually and not just draw inference based on descriptive statistics.

The numerical statistics for all are exact but graphs are very different from each other.

3. What is Pearson's R?

Ans:

The Pearson's R is called as the **Pearson correlation coefficient** that provides the strength and direction of the linear relationship between 2 quantitative variables. It ranges between -1 to 1.

High positive value (normally considered above 0.5) indicates strong positive relationship between the 2. Which means that when one variable increases the other tends to increase.

High negative value (normally considered below -0.5) indicates strong negative relationship between the 2. Which means that when one variable increases the other tends to decrease.

0 denotes no relationship.

Formula for Pearson's R is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

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

Ans:

What is scaling?

Scaling is a process of transforming variables of different units and different ranges to bring to a same scale or within a same range.

Why is scaling performed?

Scaling is performed so that we can compare the variables which have different units. For e.g., Temperature and count of bike rentals.

Different ranges of variables are brought within the same range e.g., between 0 and 1 making them easy to interpret.

Many of the machine learning models perform well on scaled data. Scaled data ensures that the model does not give undue importance to one variable because it has a large scale than the other variables.

Some statistical tests like k-means clustering etc. are also sensitive to scale of variables. Scaling ensures parity.

In linear regression coefficients are comparable when they are on the same scale.

Outlier handling can also be done using scaling as all the observations are brought within a same range and on the same scale.

What is the difference between normalized scaling and standardized scaling?

Normalized scaling:

Normalized scaling also called as Min-Max scaling brings the data within the same range of 0 and 1 or sometime -1 and 1. It basically squeezes the data into a smaller range.

Formula for normalized scaling:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

It can be used when you have variables which are on different scales like number of cars vs annual income.

Standardized scaling:

Standardization on the other hand transform by subtracting the mean and dividing by standard deviation to ensure that the data has 0 mean and unit standard deviation.

Since it has 0 mean and unit standard deviation it is insensitive to outliers.

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

Normalization: [0, 1] or [-1, 1].

Standardization: Not bounded to a specific range.

Outliers:

Normalization: Highly affected by outliers.

Standardization: Less affected by outliers.

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

Ans:

Variance Inflation Factor is a measure that measured multicollinearity between independent variables of a regression model. It provides an estimate on how much the regression coefficient of that variables is inflated due to its correlation with other independent variables.

High VIF (normally VIF values above 5 is considered to be high. But in case anything above 10 is also considered) indicates strong multicollinearity.

VIF is calculated as:

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

In this case the R_i^2 is the r-squared of a linear regression model in which the i^{th} or the variable in question is the dependent variable and the rest of the independent variables are the predictors.

If the variation in this variable is very well explained by the rest of variables, then the regression model would be very strong which means R_i^2 might become 1.

If that is the case the denominator in the VIF equation would become 0. This means that one predictor is a perfect linear combination of other predictors. This can be called as **perfect multicollinearity.**

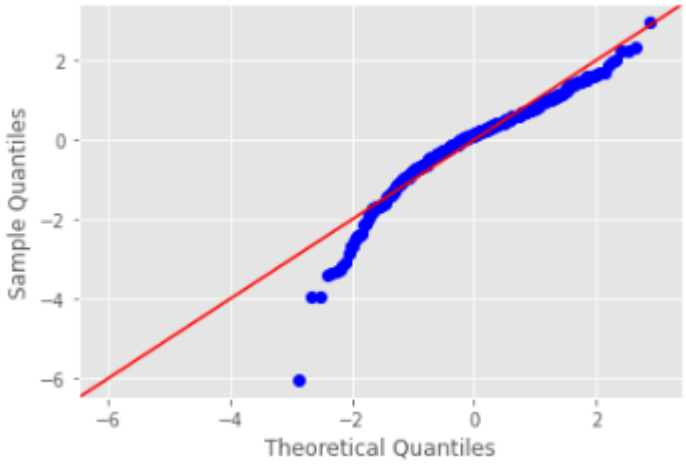
In case of perfect multicollinearity, we need to identify the corelated predictor and remove it from the model.

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

Ans:

A Q-Q plot is called as the **quantile-quantile plot**. It is a graphical representation to validate if a set of data points follow a certain theoretical distribution like normal distribution in case of a linear regression model.

A quantile is group produced by dividing a frequency distribution into equal number of groups. So, Q-Q plot divides the actual data and theoretical data into quantiles and then compares if the data aligns to the theoretical distribution.



In linear regression this plot can be used to:

To validate the normality of the residuals as it is an important assumption of linear regression.

By plotting the residuals against the quantiles of a Normal distribution, we can verify this assumption. If the points in the Q-Q plot deviate significantly from a straight line (usually at a 45-degree angle), it suggests that the residuals do not follow a Normal distribution. This indicates potential issues with the model assumptions.

Q-Q plots work well with small samples.

Interpreting Q-Q Plots:

Similar Distribution:

If all quantile points lie on or close to a straight line at a 45-degree angle from the x-axis, the data sets have similar distributions.

Different Distribution:

If quantile points deviate significantly from the 45-degree line, the data sets likely come from different distributions.