Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans1. After final model evaluation following categorical variables were identified having impact on the target variable.

Categorical Variable	Coefficients	Comments
Year	0.233	In 2019 demand increased compared to 2018. This means that business is growing each year.
Season	Winter: 0.129 Summer: 0.089	Winter and Summer are having the positive coefficients , which means they have a positive impact on the demand of rented bikes. They are in more demand in winter than summer.
Weather	Misty: -0.078 Snowy: -0.283	Bad weather has a negative effect on the demand as we can see from the coefficients of Mist and snow. Which means when weather is either Misty + Cloudy or when it's snow or when there is a thunderstorm demand goes down.
Weekday	Sunday: -0.283	Sunday has the negative coefficient which means demand is slow on Sunday compared to other days of the week.

Month	Sep: 0.095	September has the positive coefficient that means Sep month plays an important role in the increase of demand for rented bikes compared to other months.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. It is important to use drop_first = True during dummy variable creation. For example there is a categorical variable for marital_status. The unique values of the variable are as follows:

- Married
- Unmarried
- Divorced

If we get dummy values for marital_status without using drop_first = True then it will look like this:

Married	Unmarried	Divorced
1	0	0
0	1	0
0	0	1

Here the level for the categorical variable is 3 .

Here we are unnecessary storing values for level - Married

Even if we remove Married still values of the category can be decoded as below:

00 – Married

10 – Unmarried

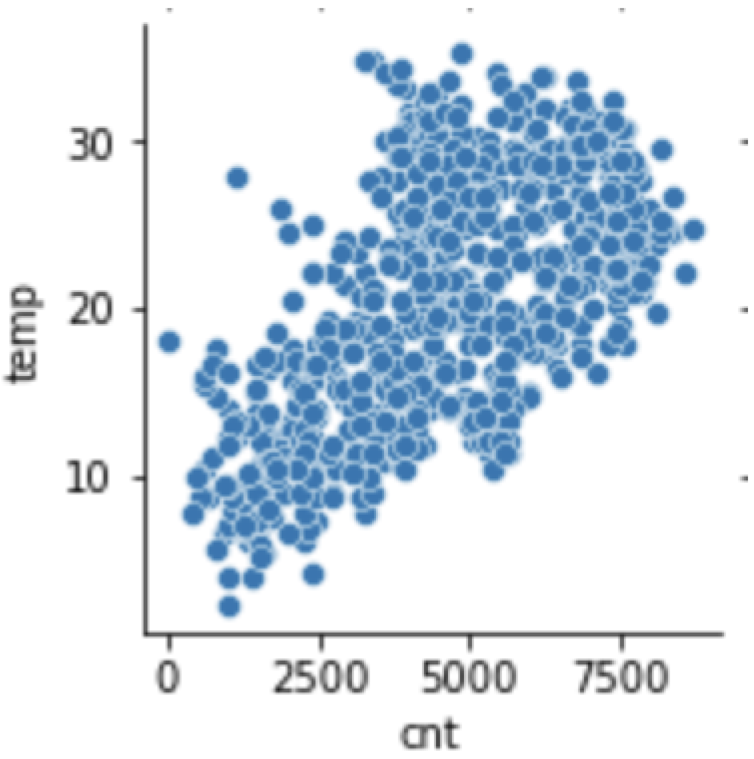
01 – Divorced

So if there are n levels of categorical variables it can be stored in n-1 dummy variables. This way if we drop the first variable we can save the creation of an extra variable.

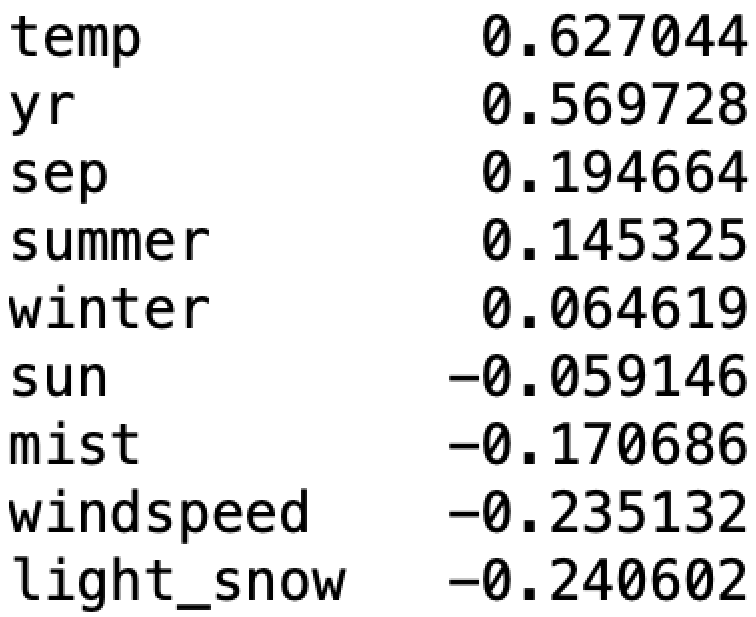
Drop_first = True – indicates python to drop the first variable and keep the other categories.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. Looking at the pair plot among the numerical variables it is observed that the temp variable has the highest correlation (0.63) with the target variable. Below is the scatter plot showing relation between temp and cnt.



Following is the correlation of all the independent variables in final model with target variable:



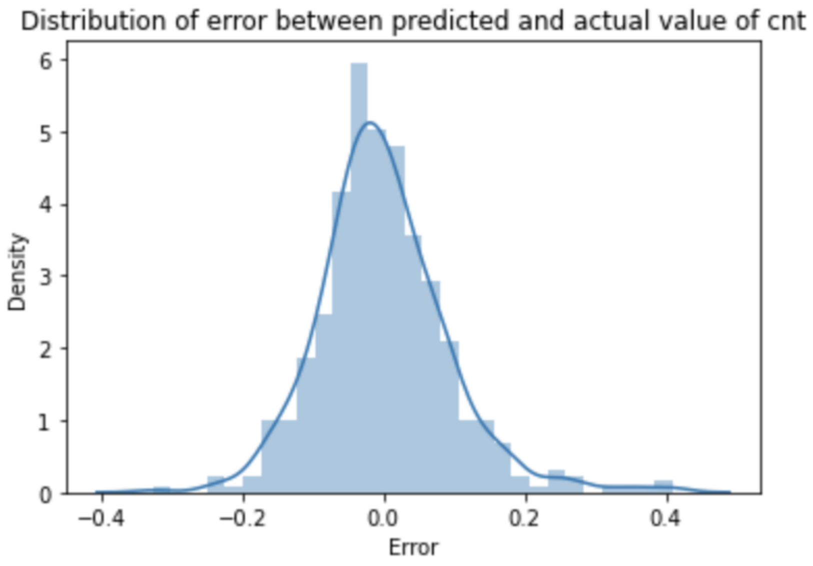
```
temp          0.627044
yr            0.569728
sep           0.194664
summer        0.145325
winter        0.064619
sun          -0.059146
mist         -0.170686
windspeed    -0.235132
light_snow   -0.240602
```

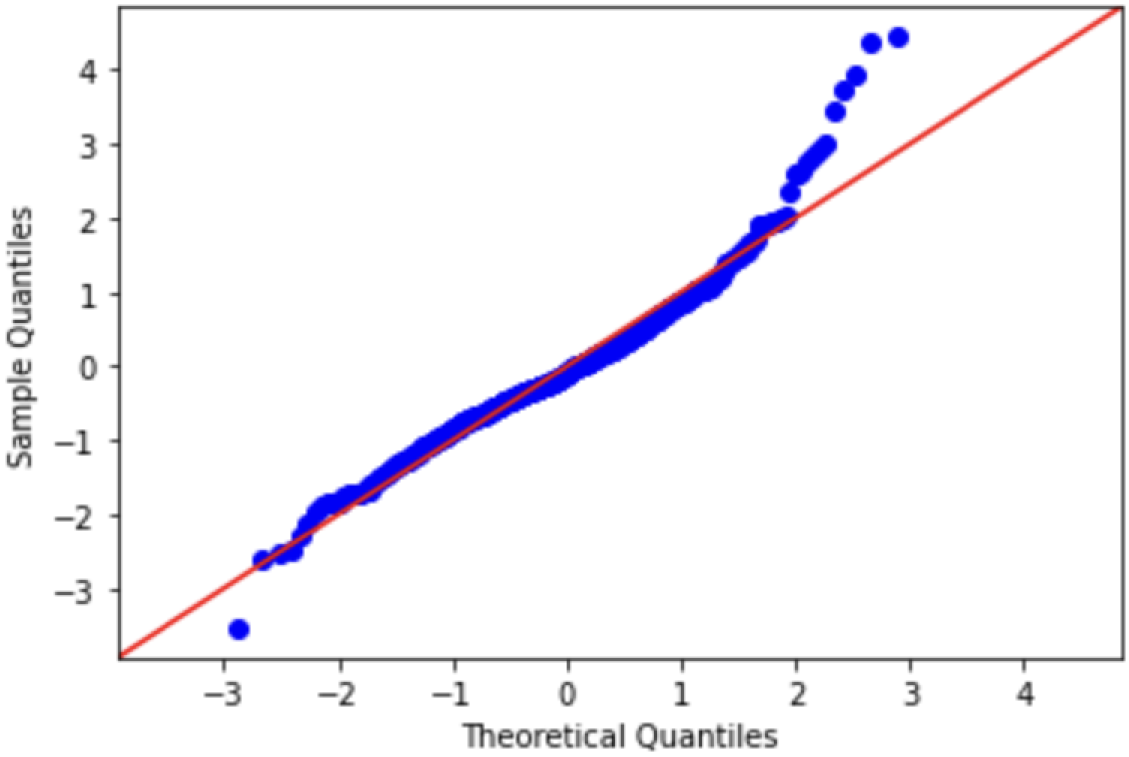
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans.

Linear Regression Assumption 1: Error terms are normally distributed with mean equal to 0:

Plotted a histogram plot on the residual (predicted values - actual values)





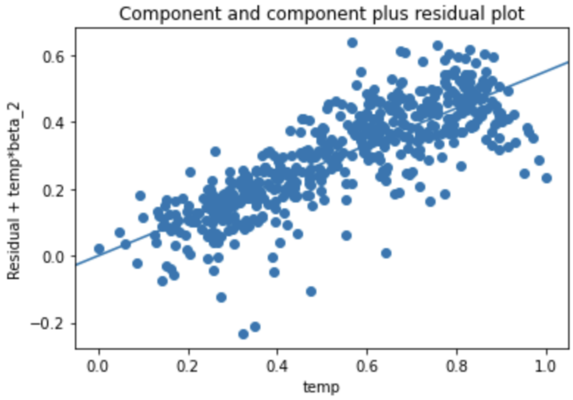
The above QQ plot is drawn on the error between actual values and predicted values.

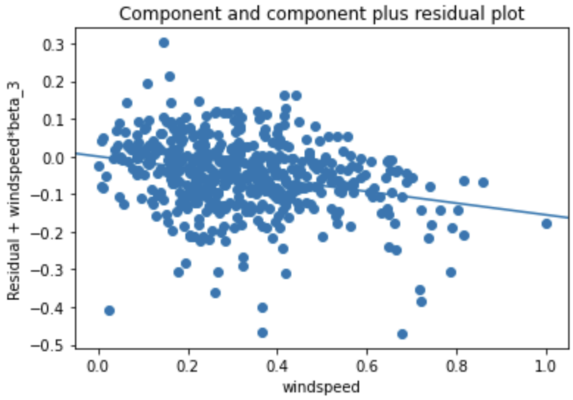
From the above plot it's proved that the residuals are normally distributed and have a mean of 0.

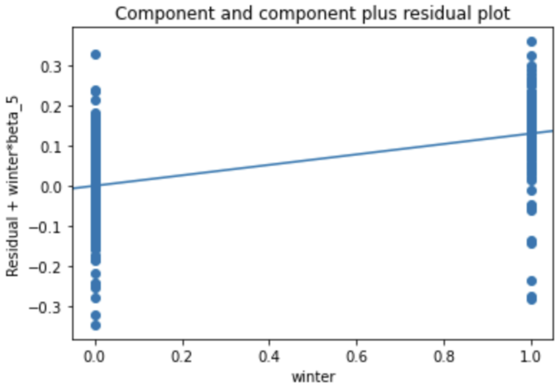
Linear Regression Assumption 2: Independent variables have linear relationship with target variable

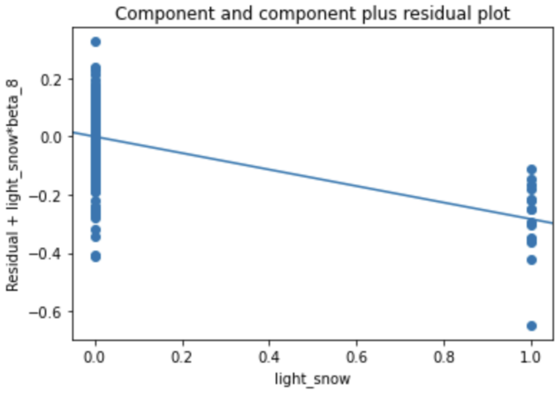
To prove this assumption we plotted the CCR plot using statsmodels on variables temp , windspeed , winter and light snow.

These plots are showing that there is a linear relationship between target variable and features.



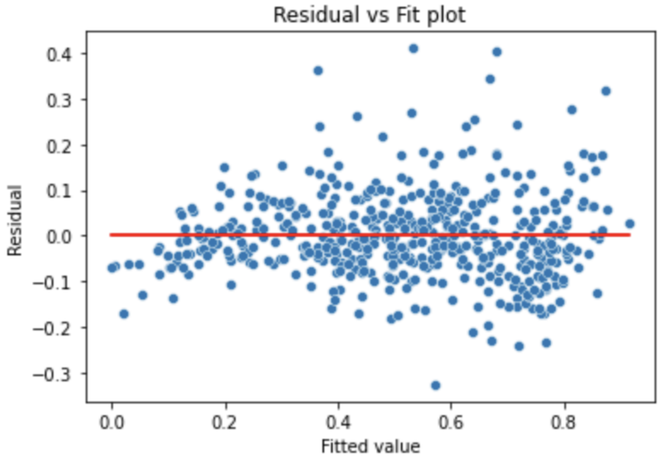






Linear Regression Assumption 3: Homoscedasticity - Error terms have constant variance

To prove this we plotted a scatter plot between predicted values and error terms.



No visible pattern between the errors hence we can say that variance is constant across X and Y.

Linear Regression Assumption 4: Homoscedasticity - Error terms are independent of each other.

In the summary of the final model the value of **Durbin-Watson is 2.052** , which means there is no correlation between residuals hence it proves that the error terms are independent of each other.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. Based on the final model, top 3 features contribute significantly towards explaining the demand of the shared bikes.

Temperature – Among all the features in the final model temp has the highest coefficient value of 0.55. That shows temperature plays a significant role in the demand of shared bikes. If the temperature is good and there are no chances of mist or rain or thunderstorm then the demand increases.

Year – The demand for shared bikes increased in 2019 significantly compared to 2018. Which gives an indication that business is growing compared to when it was started.

Snow conditions – The coefficient for snow conditions is -0.283 , this is having a negative correlation with the target variable. Hence it is quite evident that on the days when it snows or the road conditions are not good the demand for the shared bikes decreases significantly.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear regression is a supervised regression algorithm which is helpful in predicting real or continuous dependent variables. It tries to best fit the line between independent and dependent variables by minimising the sum of squares between predicted values and actual values (also called residuals) of independent variables. If only one dependent variable is available then it's called simple linear regression , when dependent variables are more than one then it's called multiple linear regression.

Mathematical equation of simple linear regression:

Y = B0 + B1X1 + E

Y = Independent variable
B0 = Constant (Interceptor)
B1 = Coefficient constants for predictor variable (this tells the increase/decrease in independent variable on 1 unit increase of dependent variable)
E = Error

Mathematical equation of multiple linear regression:

Y = B0 + B1X1 + B2X2 + B3X3 + ... BnXn + E

Y = Independent variable
B0 = Constant (Interceptor)
B1 - Bn = Coefficient constants for each dependent variable (this tells the increase/decrease in independent variable on 1 unit increase of respective dependent variable considering all the other dependent variables kept constant)
E = Error

The best fit line between dependent and independent variable is drawn by minimising the cost function using gradient descent method. In linear regression the task is to find best values of coefficients which minimises the RSS (Residual sum of squares). It is also called the ordinary least square method.

To evaluate the strength of the linear regression model R-square metric is used. Value of R-square liest between 0 and 1, where 0 being the worst model and 1 being the best model.
This explains what portion of the given data variation can be explained by the model build.

R2 = 1 - (RSS / TSS)

RSS = residual sum of squares (sum of squares of difference between actual and predicted values of independent variable)

TSS = Total sum of squares (sum of squares of differences between actual and average value of independent variable)

Assumptions of linear regression:

1. Independent variables have a linear relationship with dependent variables.
2. Error terms are normally distributed.
3. Error terms have constant variance (Homoscedasticity).
4. Error terms are independent of each other.

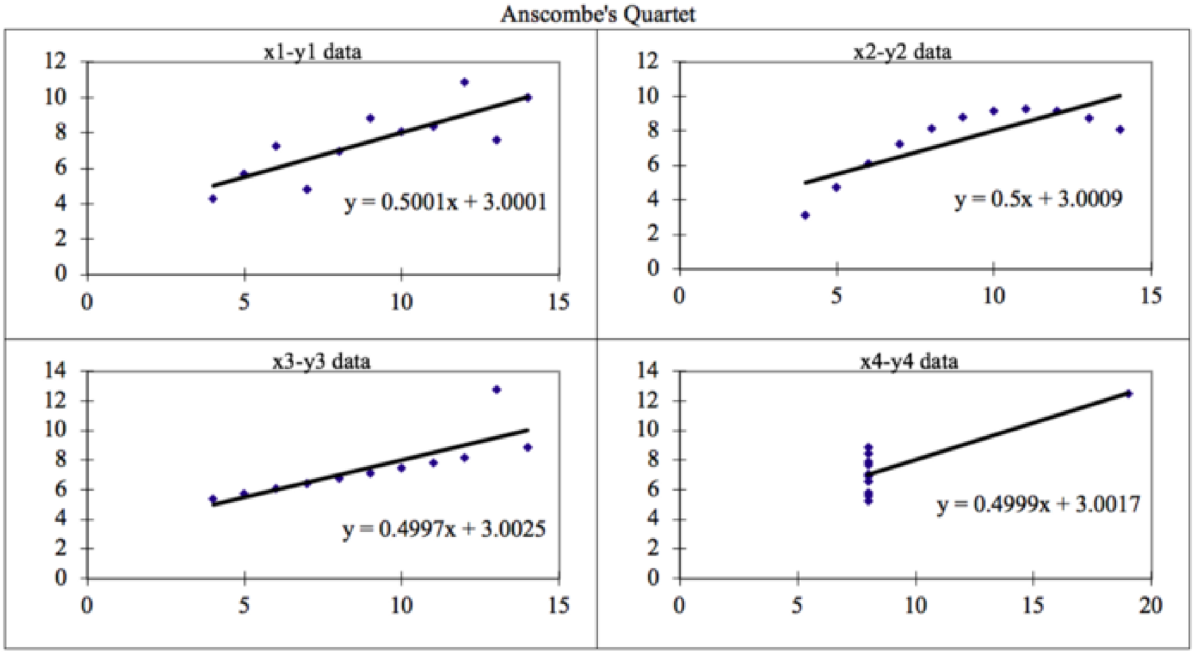
Some examples where linear regression can be used to predict:

1. Salary prediction of an employee based on employee skills and other emp data
2. Temperature prediction
3. Sales prediction

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet can be defined as a group of four data sets which gives nearly identical simple descriptive statistics but there are some peculiarities in the data set which fools the model. All 4 dataset have very different distributions when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting graphs before analysing and building the model.



As can be seen in the above image showing the 4 scatter plots , the distribution of points X and y is different in each plot but the linear regression best fit line is almost identical in all 4 scenarios.

1. Dataset 1: fits the linear regression model
2. Dataset 2: data is not linear , can't fit the linear regression model well
3. Dataset 3: there is an outlier in the dataset which can't be handled by linear regression model
4. Dataset 4: there is no linear relationship between the data and there is an outlier which is giving incorrect predictions by making a perfect line.

This tells the importance of visualising the data before applying any machine learning algorithm to build models. It says that data features must be plotted to see the distribution of the samples that can help identify the various anomalies in the data like outliers , diversity of the data , linear separability of the data etc. Linear Regression can only be considered a fit for the data with linear relationship (which is a basic assumption of building linear regression model) and is incapable of handling any other kind of dataset.

Q3. What is Pearson’s R?

It is also known as Pearson product-moment correlation coefficient (PPMCC). Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

In statistics Pearson correlation coefficient also known as Pearson’s R is a measure of linear correlation between two sets of data.It is the ratio between the covariance of 2 variables and the product of their standard deviations. It is a normalised measure of covariance such that the result always has a value between -1 and 1.

For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

Formula:

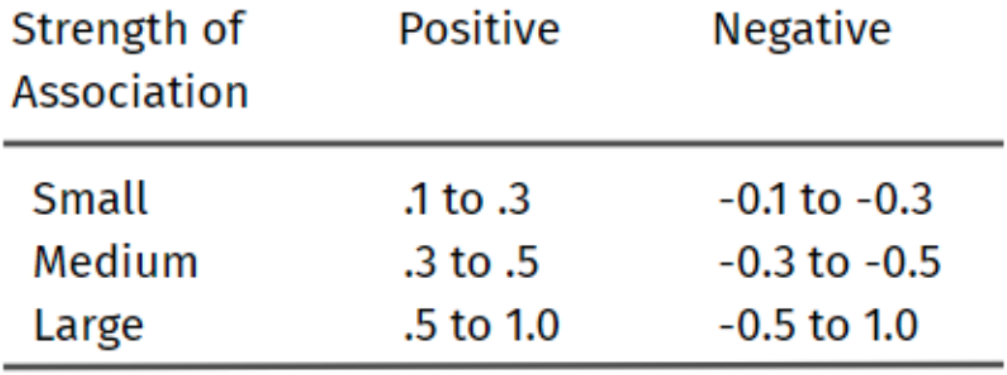
$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where,
N = the number of pairs of scores
∑xy = the sum of the products of paired scores
∑x = the sum of x scores
∑y = the sum of y scores
∑x2 = the sum of squared x scores
∑y2 = the sum of squared y scores

The positive value of Pearson’s correlation coefficient implies that if we change either of these variables, there will be a positive effect on the other. For example, if we increase the age there will be an increase in the income.

Strength of pearson-correlation:

As we have learned from the definition of the Pearson product-moment correlation coefficient, it measures the strength and direction of the linear relationship between two variables.
The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables.
Below, we have shown the guidelines to interpret the Pearson coefficient correlation :



Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

Q4 . What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Ans. Scaling is a technique to standardise the independent features present in the data in a fixed range. It is performed during the data pre-processing stage to handle highly varying magnitudes or values or units.
The machine learning algorithm works on numbers and it doesn’t know what that number represents. For example a weight of 10 grams and price of 10 dollars represents totally different values but for a model as a feature both the values are same. Feature scaling is essential for machine learning algorithms that calculate distances between data. If not scale, the feature with a higher value range starts dominating when calculating distances.

Why do we need scaling?
1) Concept of Gradient Descent — In linear regression, we aim to find the best fit line. To do so, we first have to find global minima with the concept of gradient descent. And, we can reach this global minima faster if we scale the data.

2) Distance Based Algorithms –In algorithms like KNN, K-means and Hierarchical clustering we find the nearest points using Euclidean distance and hence the data should be scaled

for all features to weigh in equally. If not done so, the features with high magnitude will weigh a lot more in the distance calculations than features with low magnitude.

Note: It is important to perform feature scaling post splitting the data into training and testing. If not done so, there will be data leakage from test data to train data.

Standardised scaling:

In this approach, we bring all the features to a similar scale centring the feature at 0 with a standard deviation of 1. In the case of outliers, this scaler technique will be affected. Hence, it is used when the features are normally distributed.

Formula for standardisation:

$$x_{new} = \frac{x - \mu}{\sigma}$$

Normalisation or min-max scaling:

Transform features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g., between zero and one. This Scaler shrinks the data within the range of -1 to 1 if there are negative values.

Formula for normalisation:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

This Scaler responds well if the standard deviation is small and when a distribution is not Gaussian. This Scaler is sensitive to outliers.

Difference between Standardisation and Normalisation:

Normalisation	Standardisation
Min and max values are used for scaling	Mean and standard deviation is used for scaling
It is really affected by outliers	It is much less affected by outliers
It is useful when the distribution of the	It is useful when the distribution of data is

data is unknown	Gaussian or Normal distribution
It is often called as scaling normalisation	It is often called as z-score normalisation
It squishes n-dimensional data into n-dimensional unit hypercube	It translates the data to the mean vector of original data to the origin and squishes or expands.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. Following is the formula to calculate VIF :

$$\textbf{VIF = 1/(1-R^2)}$$

VIF is calculated on the basis of R-square calculated for each variable such that the variable is a dependent variable and all other variables are considered as independent variables.

Consider a scenario where the variable gets perfect R-square of 1 when considered as dependent variable along with other variables considered as independent variable.

In this case the VIF will be :

$$\textbf{VIF = 1/(1-1) = 1/0 = infinite}$$

To solve this problem we should drop this variable.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. Q-Q plots are also known as quantile-quantile plots. They plot the quantiles of sample distribution against quantiles of theoretical distribution. By doing this it can be determined if the dataset follows any kind of particular probability distribution like Normal or Gaussian , Uniform or Exponential distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
- Skewness of distribution

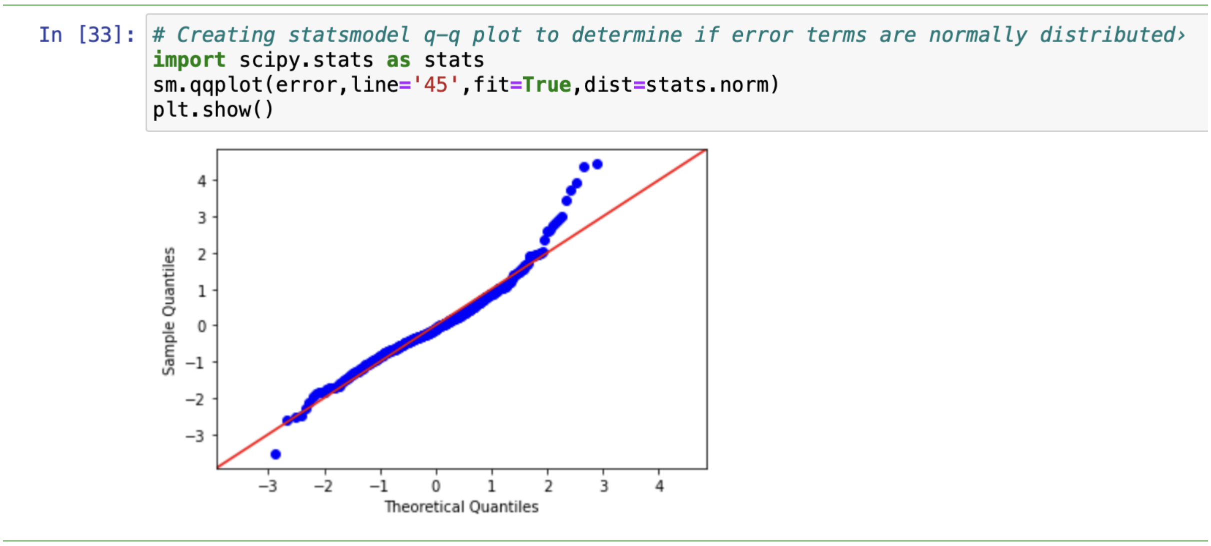
In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line

statsmodels.api provides qqplot and qqplot_2samples to plot Q-Q graphs for single and two different data sets respectively.

For Normal distribution:

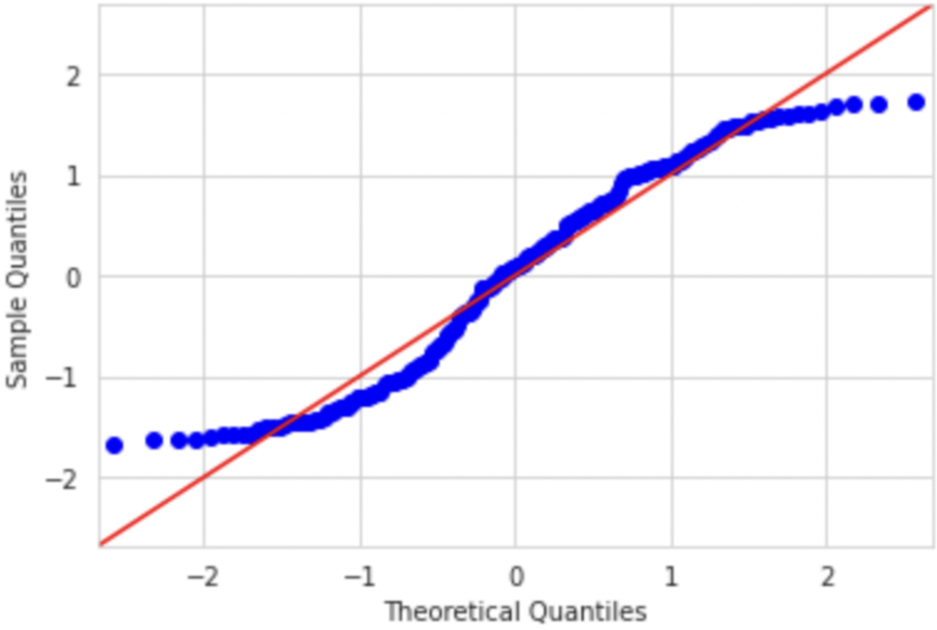
If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution which we used to validate the linear regression assumption that error terms are normally distributed in Bike sharing case study.



In [33]:

```
# Creating statsmodel q-q plot to determine if error terms are normally distributed›
import scipy.stats as stats
sm.qqplot(error,line='45',fit=True,dist=stats.norm)
plt.show()
```

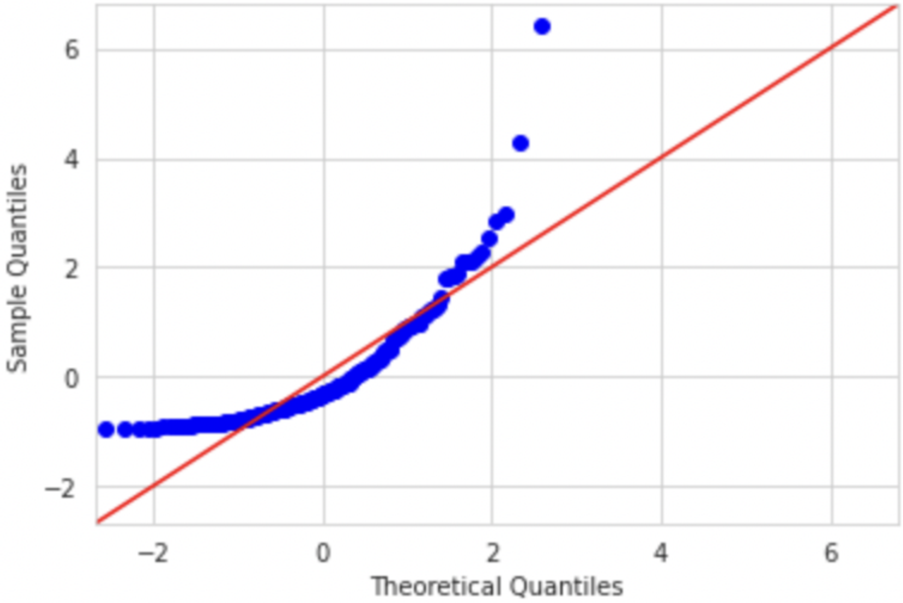
For Uniform Distribution:



Since the dataset has a uniform distribution, both the right and left tails are small and the extreme values in the above plot are falling close to the centre. It follows S-shape for uniform distribution.

For Exponential distribution:

If the data follows exponential distribution graph looks like below:


Sample Quantiles
Theoretical Quantiles