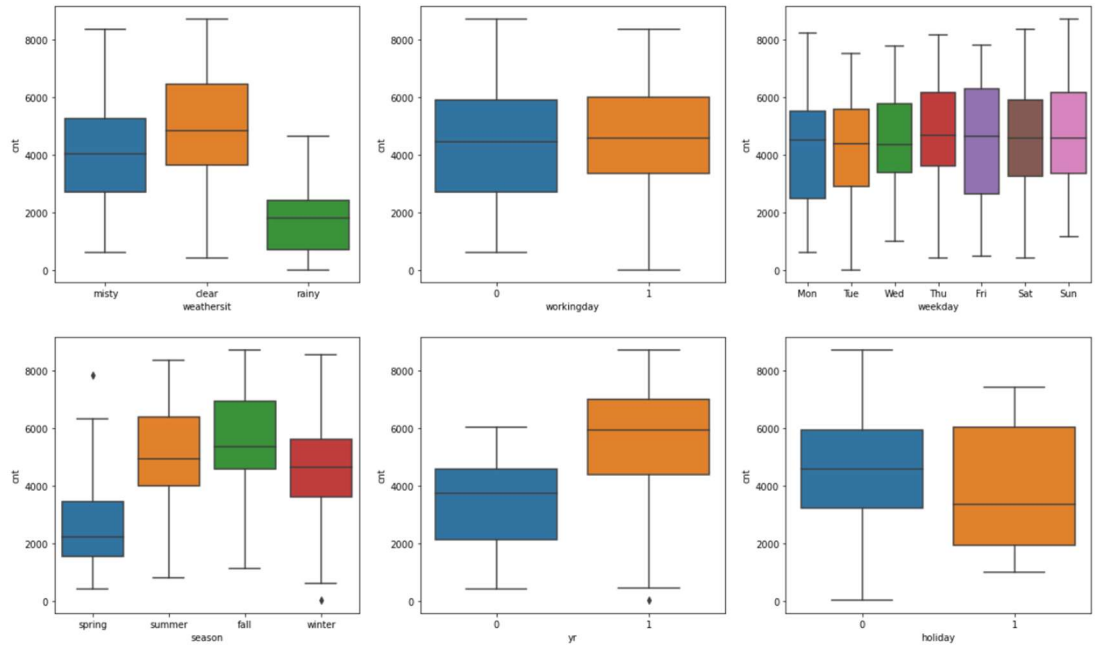
NAME: Sujith Kumar

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



- **Weathersit** – Count of users increases for weathersit 1 (clear) and 2(Mist)
- **Workingday** – does not seem to create significant impact on count of users
- **Season** – count of users are low in spring and high during fall
- **Year** – Count of users have increased during 2019 when compared to previous year
- **Workingday** – More users are in 50-75% range during a working day

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

For a linear regression model, categorical variables should be converted into numerical variables. One hot encoding is used in regression models which will create 'n' new attributes (dummy variables) corresponding to 'n' number of categories in a particular categorical variable. Dummy variables have values of '0' or '1'. It is important to use drop_first to drop the first dummy column for a category to avoid Dummy Variable Trap.

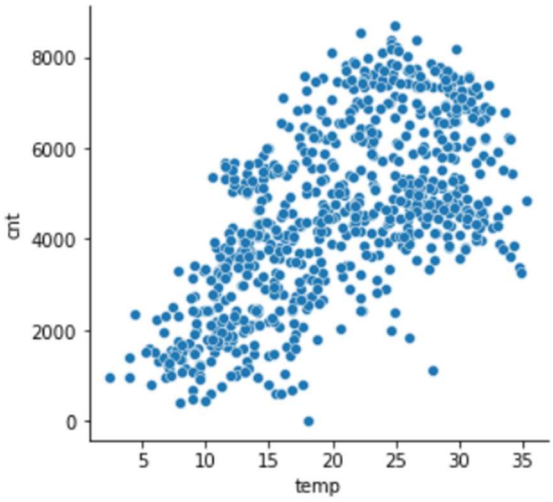
Dummy Variable Trap:

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. In the model this will lead to a higher R square value and the overall model fit F-statistic would not be high.

Ex: in the given model - weathersit is converted to dummy variables and one column is dropped to avoid impact of multi-collinearity

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

Temperature has the highest correlation with cnt



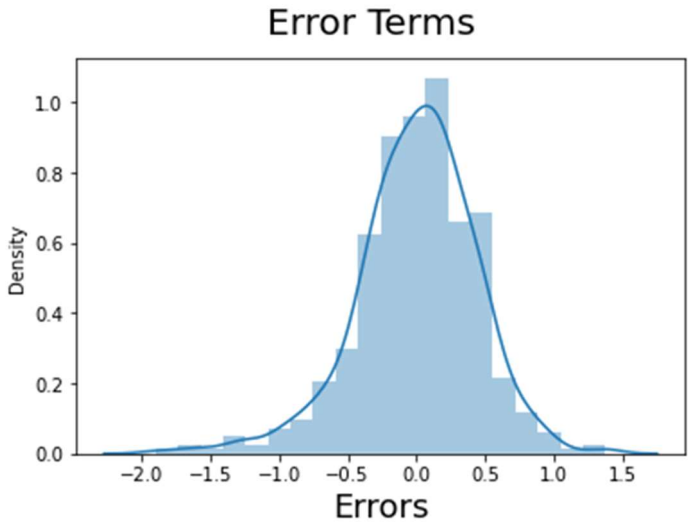
It is also observed that the derived variable days of business also shows high correlation

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

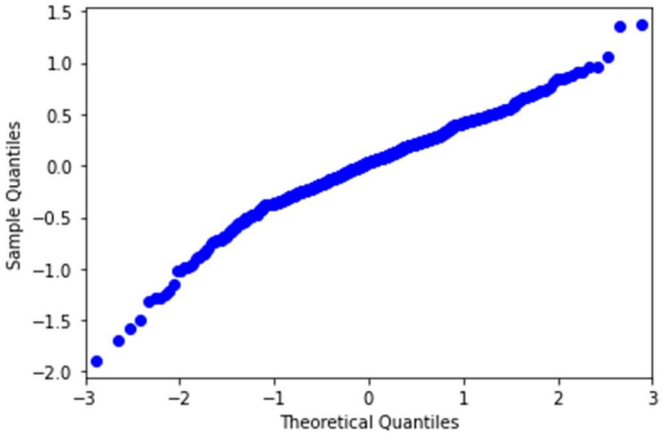
The below assumptions where validated

i. There is linear relationship between X and y
ii. Error terms are normally distributed with mean zero(not X, Y)
iii. Error terms are independent of each other
iv. Error terms have constant variance (homoscedasticity)

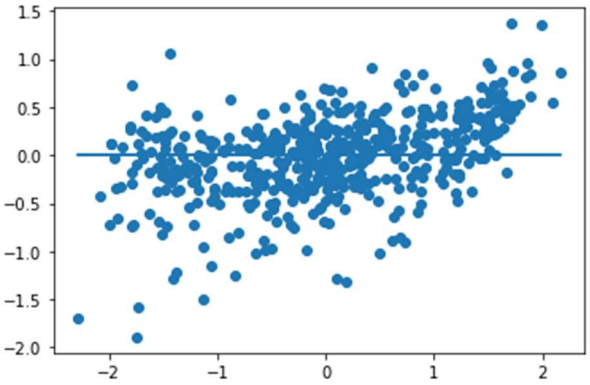
Plotting Error Terms showed normal distribution



QQ Plot also verifies normal distribution of residuals



Error Terms are independent of each other and are homoscedastic



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

When all other variables are kept constant,

- **Weathersit_clear** – for a unit improvement on clear weather count of user increases by 1.35
- **Yr –** for continued business operations, each year count would increase by 1.03 times
- **Weathersit_misty** – for a unit improvement on misty weather count of user increases by 1.01
- **Temp** – for a unit improvement in temperature, count of users increases by 0.59 times

1. Explain the linear regression algorithm in detail.

Linear regression is a method for modelling the relationship between two scalar values: the input variable x and the output variable y. The model assumes that y is a linear function or a weighted sum of the input variable.

$y = f(x)$

Or, stated with the coefficients

$y = b_0 + b_1 x_1$

The model can also be used to model an output variable given multiple input variables called multivariate linear regression

$y = b_0 + (b_1 x_1) + (b_2 x_2) + ..$

The objective of creating a linear regression model is to find the values for the coefficient values (b) that minimize the error in the prediction of the output variable y.

Ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares, minimizing the sum of the squares of the differences between the observed dependent variable and those predicted by the linear function of the independent variable.

Gradient Descent is an optimisation algorithm which optimises the objective function (for linear regression it's cost function) to reach to the optimal solution. Gradient Descent Algorithm iteratively calculates the next point using gradient at the current position, then scales it (by a learning rate) and subtracts obtained value from the current position (makes a step.

This process can be written as:

$$p_{n+1} = p_n - \eta \nabla f(p_n)$$

There's an important parameter η which scales the gradient and thus controls the step size. In machine learning, it is called learning rate and have a strong influence on performance. The smaller learning rate the longer GD converges, or may reach maximum iteration before reaching the optimum point. If learning rate is too big the algorithm may not converge to the optimal point (jump around) or even to diverge completely.

Gradient Descent method's steps are:

1. choose a starting point (initialisation)
2. calculate gradient at this point
3. make a scaled step in the opposite direction to the gradient (objective: minimise)
4. repeat points 2 and 3 until one of the criteria is met:
    a. maximum number of iterations reached
    b. step size is smaller than the tolerance.

2. Explain the Anscombe's quartet in detail.

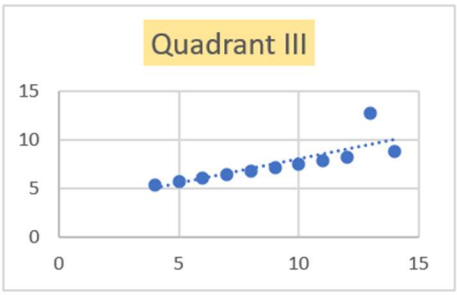
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

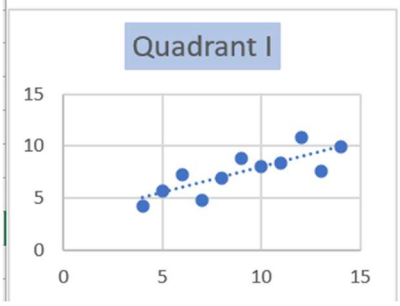
Each dataset consists of eleven (x,y) points.

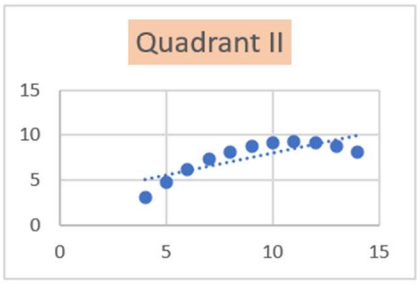
They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

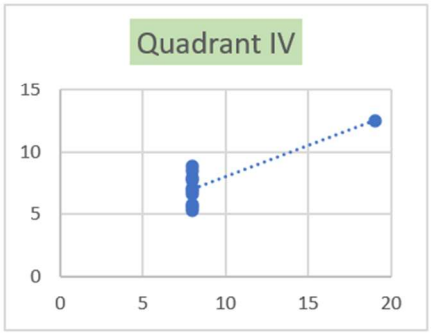
Below is a plotting of the 11 data points in excel

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









Below is the calculation of various statistical values for the above Quadrants in excel

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

	I	II	III	IV
Mean X	9	9	9	9
Mean Y	7.500909	7.500909	7.5	7.500909
Sample Variance X	11	11	11	11
Sample Variance Y	4.127269	4.127629	4.12262	4.123249
Correlation X and Y	0.816421	0.816237	0.816287	0.816521
R Squared	0.666542	0.666242	0.666324	0.666707

3. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables

r value > 0.5 means strong positive correlation

r value < -0.5 means strong negative correlation

r value = 0 means no correlation

Below is a sample illustration of R calculation done in excel

X	Y	x^2	y^2	X*Y
1	2	1	4	2
2	4	4	16	8
3	6	9	36	18
4	8	16	64	32
5	10	25	100	50

$\sum xy$ 110

$\sum x$ 15

$\sum y$ 30

$\sum x^2$ 55

$\sum y^2$ 220

$$r = \frac{n\sum xy - (\sum x)*(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Calculated r

1

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

Scaling is the process of normalizing features in a data set which have higher magnitudes in different units. Not scaling the features would result in incorrect model as the algorithm would consider higher magnitudes without considering units.

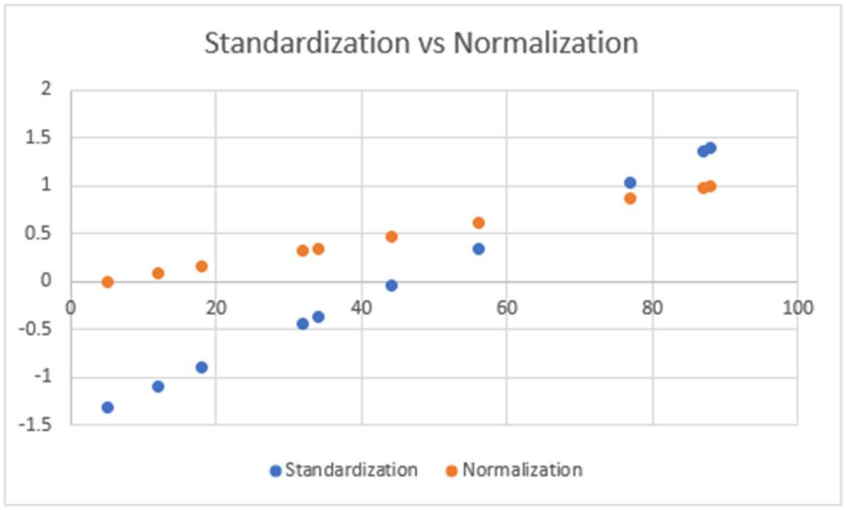
standardization (or Z-score normalization) will rescale features such that they follow a standard normal distribution with μ=0 and σ=1 where μ is the mean and σ is the standard deviation from the mean; standard scores (also called z scores) of the samples are calculated as follows: $z=(x-\mu)/\sigma$

Min-Max scaling ("normalization") scales features to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is smaller standard deviations, which could suppress the effect of outliers. A Min-Max scaling is typically done via the following equation: $X_{norm}=(X-X_{min}) / (X_{max}-X_{min})$

Below is an excel example of scaling

X	Standardization	Normalization
12	-1.085734883	0.084337349
34	-0.368432558	0.34939759
5	-1.313967441	0
77	1.033567441	0.86746988
88	1.392218604	1
56	0.348869767	0.614457831
87	1.359613952	0.987951807
32	-0.43364186	0.325301205
44	-0.042386046	0.469879518
18	-0.890106976	0.156626506

σ	30.67047078
μ	45.3
Min	5
Max	88



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

If there is a perfect correlation between features then VIF value is observed as infinity. In the below equation if correlation is 1 (max) then VIF will be infinite

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

where *'i'* refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

The common heuristic we follow for the VIF values is:

> 10: Definitely high VIF value and the variable should be eliminated.

> 5: Could be ok to have this feature, but it is worth inspecting.

< 5: Good VIF value. No need to eliminate this variable.

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

A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles.

The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q–Q plot follows the 45° line y = x. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line y = x.

Q-Q plots helps check that the data meet the assumption of normality. They compare the distribution of data to a normal distribution by plotting the quartiles of data against the quartiles of a normal distribution. If the data is normally distributed then they should form an approximately straight line

Below is the Q-Q plot output from case study

In [75]:
```
# Plot QQ plot
#import scipy.stats as stats
sm.qqplot(res)
plt.show()
```