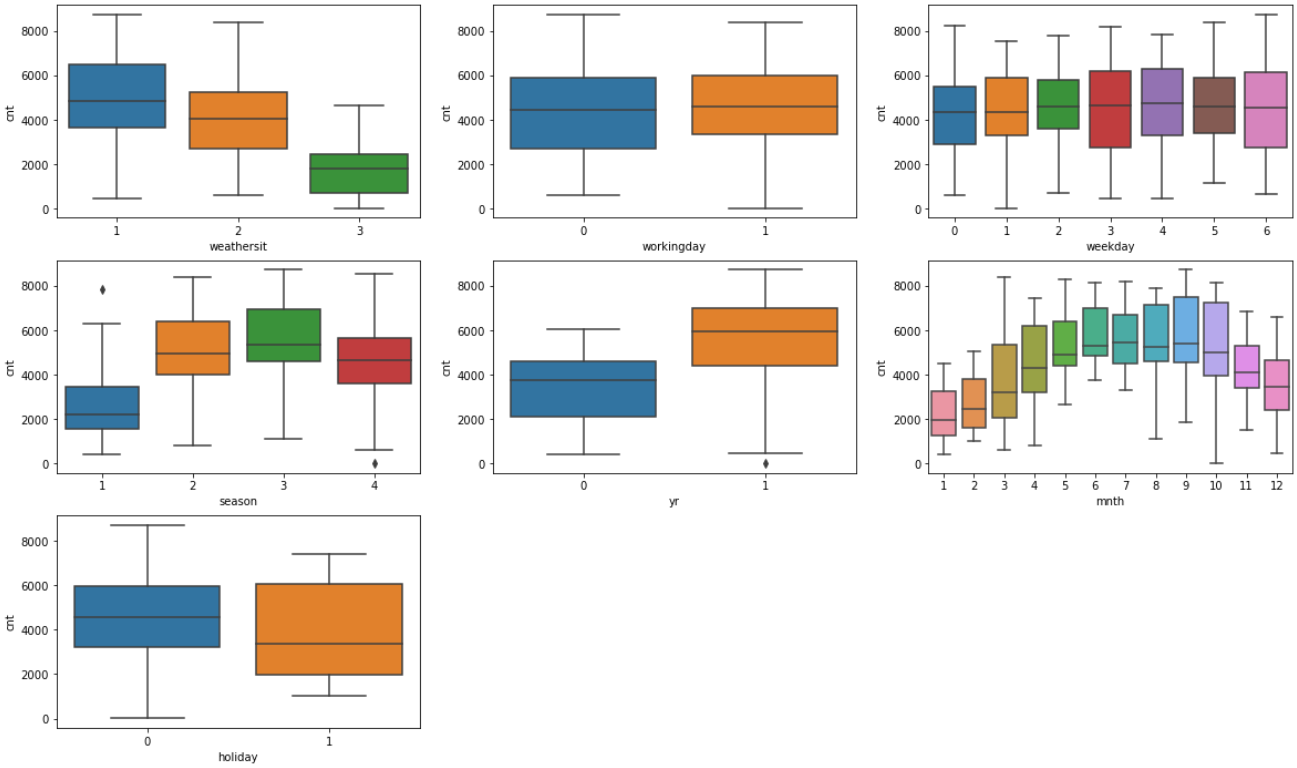
Assignment-based Subjective Questions

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

   The categorical variables present in the dataset are:

   1. **Weathersit**
   2. Workingday
   3. Weekday
   4. **Season**
   5. **Yr**
   6. **Mnth**
   7. **Holiday**



- As can be observed in box plot, for workingday and weekday, there is no variation observed in mean value of cnt. Hence, they were dropped from the model in pre-processing step.

- The other 5 variables were significant in some form when broken into dummy variables in the model.

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

If there are “n” levels in a categorical variable, then pd. Get_dummies creates n dummy variables. However, 1 of those dummy variables can be predicted from the rest of n-1

dummy variables, thus exhibiting **100% R*2 or VIF = inf**. Hence, to avoid **multicollinearity** it is important to drop one of those dummy variables. Drop_first provides an easy parameter to drop the first of those dummy variables.

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

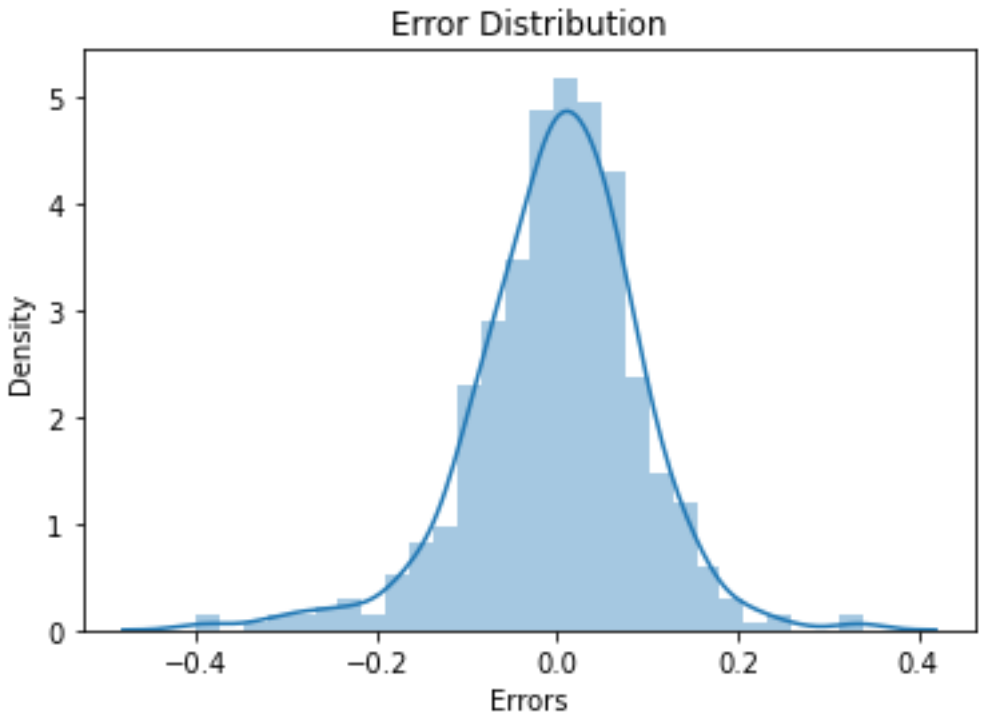
Both atemp and temp have the highest correlation with the target variable. Though, both are correlated with each other, hence one of them is dropped in model. So, in my model, atemp has the highest correlation with 'cnt' or target variable

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

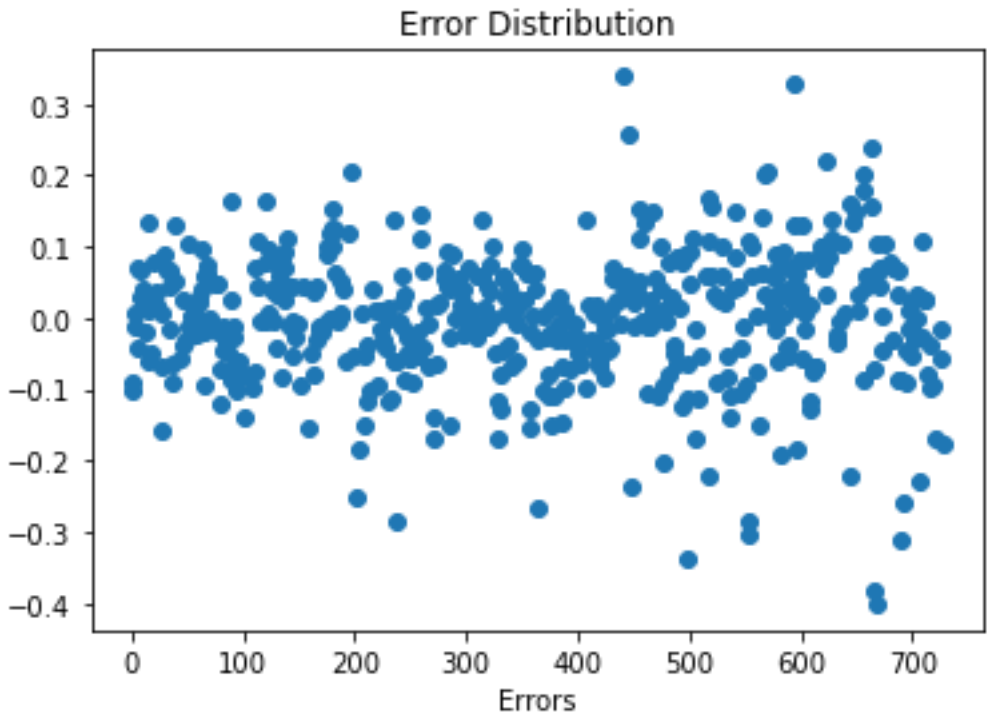
To validate the assumptions of Linear Regression, the following were completed:

- **Normal Distribution Validation** for Residuals of Training Set

  Histogram of errors (y pred – y train) was plotted. A normal distribution cantered around mean was obtained.



- **Homoscedasticity,** where the variance of errors is constant for all the data.



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

The final model's summary is shown below:

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.829
Model:	OLS	Adj. R-squared:	0.825
Method:	Least Squares	F-statistic:	241.2
Date:	Tue, 10 May 2022	Prob (F-statistic):	6.05e-184
Time:	15:31:21	Log-Likelihood:	488.67
No. Observations:	510	AIC:	-955.3
Df Residuals:	499	BIC:	-908.8
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1275	0.017	7.429	0.000	0.094	0.161
yr	0.2334	0.008	27.867	0.000	0.217	0.250
holiday	-0.0934	0.027	-3.513	0.000	-0.146	-0.041
atemp	0.5370	0.024	22.850	0.000	0.491	0.583
windspeed	-0.1319	0.026	-5.132	0.000	-0.182	-0.081
weathersit_2	-0.0822	0.009	-9.184	0.000	-0.100	-0.065
weathersit_3	-0.2760	0.025	-10.953	0.000	-0.326	-0.227
season_2	0.0990	0.011	8.750	0.000	0.077	0.121
season_4	0.1311	0.011	12.114	0.000	0.110	0.152
month_8	0.0679	0.017	4.065	0.000	0.035	0.101
month_9	0.1209	0.017	7.226	0.000	0.088	0.154

As can be seen above, the top three features are:

1. **atemp** with coeff 0.53
2. **weathersit_3** with coeff -0.27

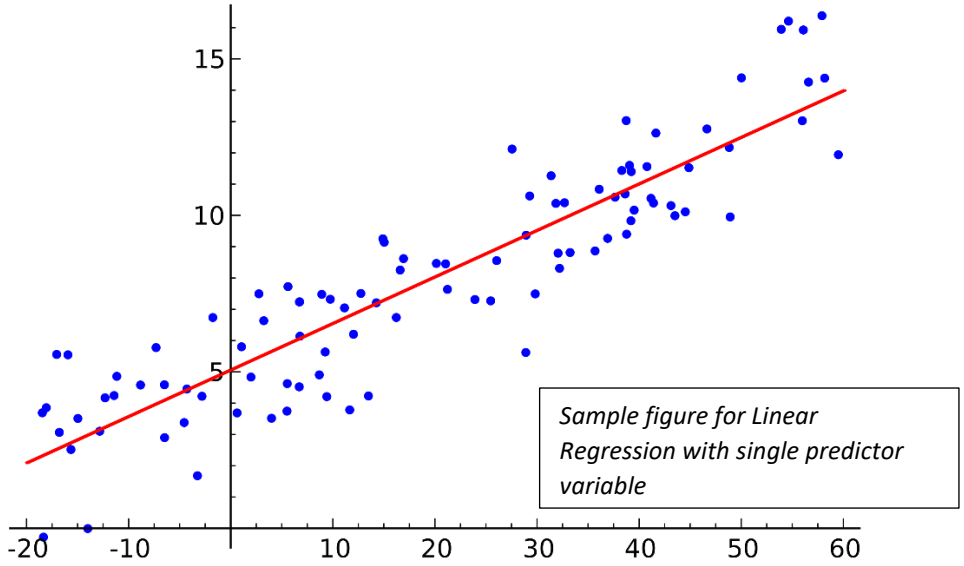
- This means that in light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds, there is less demand for bikes

3. **yr** with coeff 0.23

General Subjective Questions

1. Explain the linear regression algorithm in detail.

In the linear regression algorithm, the objective is to minimize the square of the variance of the predicted values and the actual training set values. This is also known as **Least Square method of prediction.**

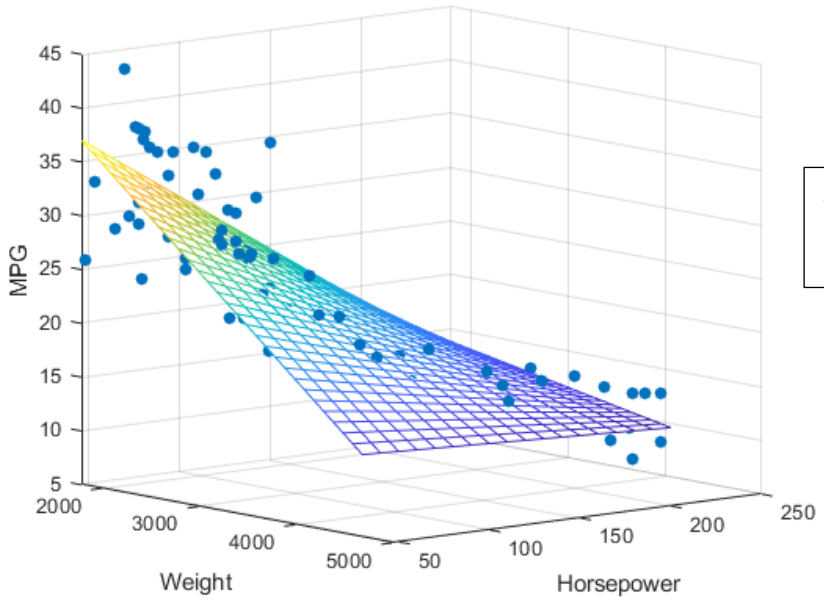


Sample figure for Linear Regression with single predictor variable

Through this algorithm, a linear equation forms the output, which is of the form as below:

Y = β0 + β1*X1 + β2*X2+ β3*X3+ β4*X4+ β5*X5……+ βn*Xn

Where Y: Target Variable; β0: Constant; X1,X2,X3: Independent Variables or Predictor Variables; β1, β2 etc.: Coefficients of corresponding predictor variables.



Sample figure for Linear Regression with multiple predictor variables

Assumptions for Linear Regression Model:

- **Linearity:** Target Variable has linear relationship with predictor variables
- **Independence:** The residuals follow a normal distribution with mean at zero
- **Homoscedasticity:** The variance of residual is similar for any observation

Evaluation Metric:

The linear regression model is evaluated basis on adjusted R^2 , where a value greater than 60% is considered a decent model.

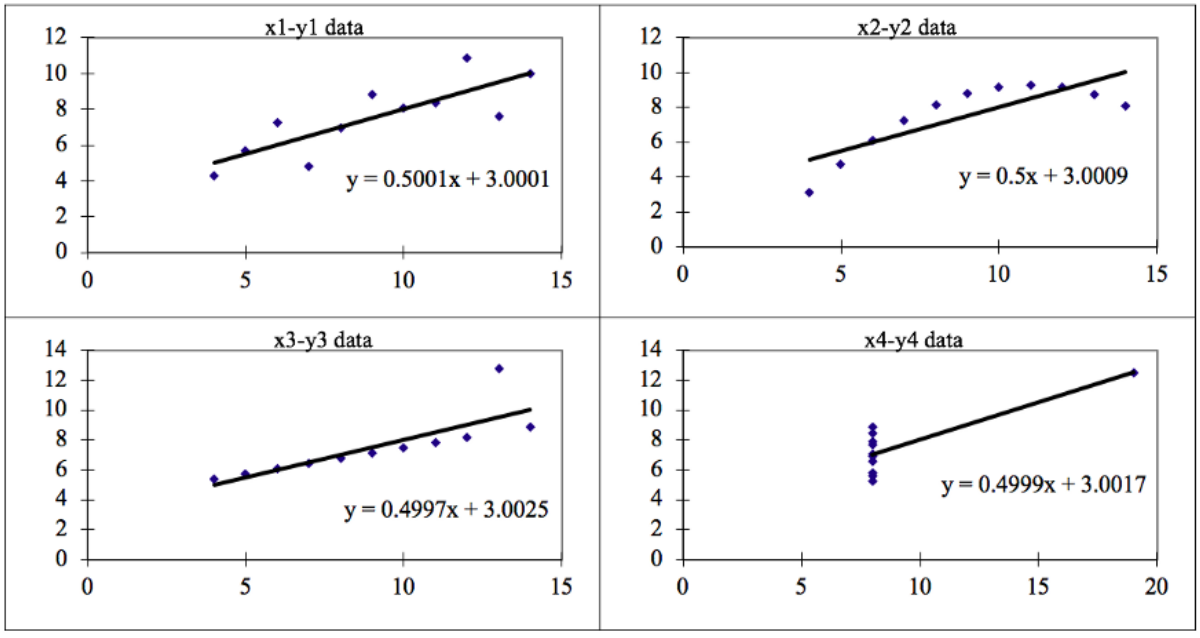
2. Explain the Anscombe's quartet in detail.

Statistician Francis Anscombe in a bid to demonstrate that just descriptive statistics in isolation are no means to analyze a dataset, **created 4 different X Y datasets** which were similar in their descriptive summary stats yet **had completely different distributions.**

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Fig Ref: https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

As you can see on the above table, we have 4 different XY datasets which have the same Summary Stats. Let's now observe their distribution through graphs.



As can be observed from 4 scatter plots, the regression line plotted is the same for the 4 datasets. **This can be misleading** as only the first dataset follows a linear relationship between X and Y (Dataset 3 has one outlier which hampers the trendline).

To conclude, **one should apply linear regression model to only those datasets which follows a linear relationship between target and predictor variables.** Additionally, linear regression models are susceptible to outliers.

3. What is Pearson's R?

Developed by Karl Pearson, Pearson's R represents the **correlation coefficient** between two variables. **It is a standardized metric and takes a value between -1** and 1. Mathematically, it's a ratio of covariance of two variables divided by the product of the standard deviation of those two variables.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Pic ref: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

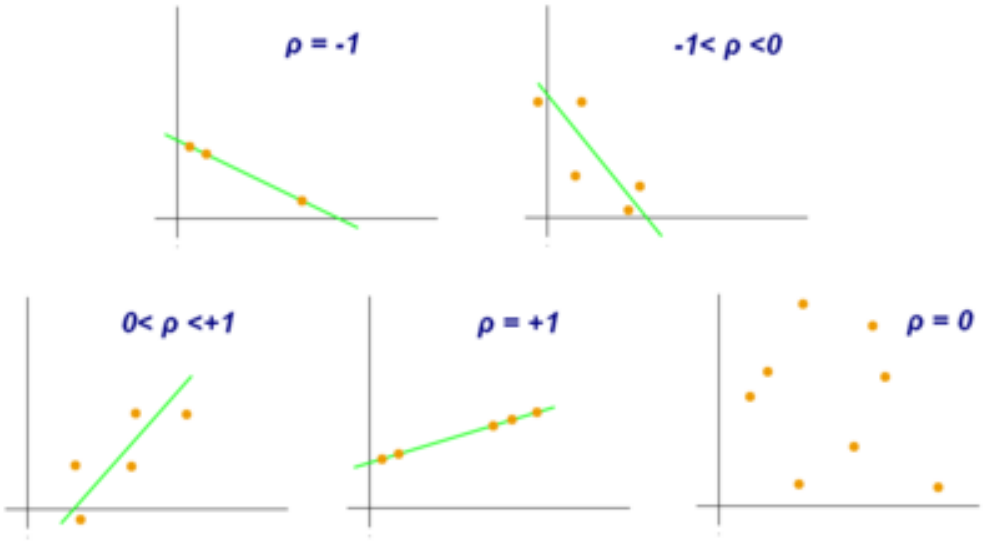


Fig ref: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

As can be seen on the above image, there can be different pearson's coefficient depending upon the scatter plot of the variables.

Machine Learning Modelling Applications

- A positive correlation denotes increase in value of target variable with increase in value of predictor variable, and vice versa for negative correlation
- If correlation coefficient > 0.7, then it's a strong correlation. If between 0.3-0.7, then medium correlation, else it's a low correlation.
- Generally, **strongly correlated variables influence the target variables the most** and will be found as significant predictor variables.

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

Scaling refers to transformation of data from one scale to a different scale. **This is done so as to:**

- **Better compare the distributions of multiple variables**.
- Increase computational speed or **efficiency** of the model.

For ex – If there are two variables Age and Salary, with below descriptive stats:

	Age	Salary
Mean	25	1,00,000
Min	10	20,000
Max	60	15,00,000

If we try to predict a target variable for ex – Retention Period in a firm (years) using Age and Salary, we will run into a problem where the correlation coefficients of both variables will not be comparable since both have different ranges.

To solve this problem we scale the variables using either of the below two approaches:

a) Normalized Scaling or Min Max Scaling

In this approach, every data is transformed using the below mathematical formula:
$X_{new} = (X - X_{min}) / (X_{max} - X_{min})$
X_{new} will lie between 0 and 1b)

b) Standardized Scaling

Compared to normalization, in Standardization each data is transformed basis the below formula.

$X_{new} = (X - X_{mean}) / (\text{Std. Deviation})$

The range of X_{new} is not bounded, however the distribution of the variable is now a normal distribution with mean zero.

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

VIF is defined as Variance Influence Factor, which is a measure of how much is a variable can be predicted from all the other variables in the dataset. The formula for VIF is as follows:

$VIF = 1/1-R^2$

where, R^2 is defined as how much can the predictor variables are able to account for the variance for the target variable.

If $R^2 = 1$ or 100%, then it means the target variable can be explained to the exact 100% through the independent variables. **This is a case of perfect correlation** and leads to VIF of infinite.

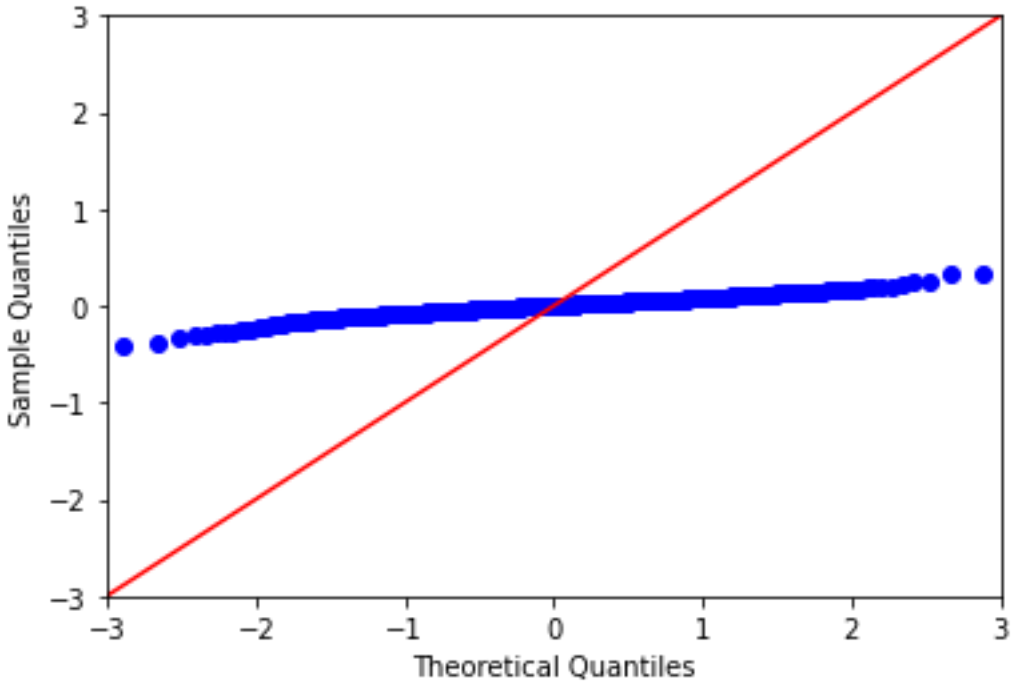
In such scenarios, we need to drop the variables where VIF is infinite.

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

A Q-Q plot or Quantile-Quantile plot compares the distribution of a variable with a desired statistical distribution such as Gaussian or Normal Distribution, Uniform Distribution or any other type of standard distribution.

In Linear Regression, we can evaluate the errors for a normal distribution centred around mean zero using a Q-Q plot. The Y axis will have Quantiles of Error Terms while the X axis will have Theoretical Quantiles of a Normal Distribution. This is similar to plot of a histogram of error terms (with KDE).

In our assignment, below was the graphical result of the error terms from the training data set.



The above Q-Q plot shows that the residuals are stacked up at zero and have very narrow tails