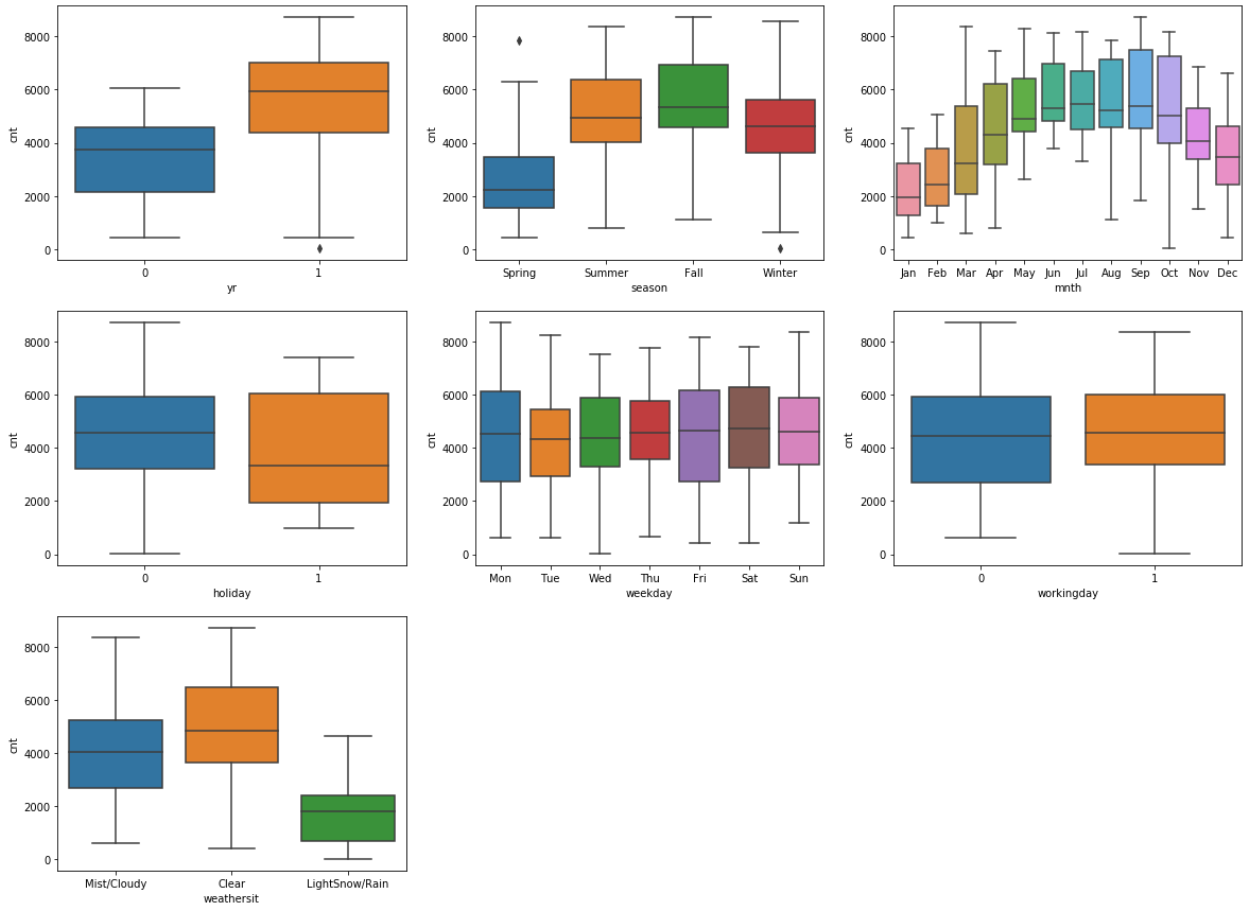
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



We observe the impact of the Different Categorical Variables on the target:

Year with a symbol of ***yr*** shows a huge difference in their rentals based on the year. In 2018 we see lower rentals than 2019.

Season has a drastic impact on sales. There are four categories in season: Spring, Summer, Fall and Winter. We observe that the rentals are lowest in the month of Fall, with much better rentals during the month of Summer and Fall. With a slight decrease in the rentals in the month of winter.

Months, with a symbol of ***mnths*** has shown significant impact with rental being high during mid year months.

The category ***holiday*** shows lower consistency for the days there are holidays, where the IQR has a wider spread than the days there are no holidays.

Across the ***weekday***'s and ***workingday***'s there is no variation seen in the number of rentals.

Weatersit visualize the significant impact of the weather on the bike rentals.

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

The reason why we drop the first created dummy variable is because we want to avoid dummy variable traps.

Dummy variable trap is induced multicollinearity due to the creation of equal numbers of dummy variables as the number of categories.

Let's say we have a Column with 3 categories *low*, *med*, *high*.

cat	low	med	high
low	1	0	0
med	0	1	0
high	0	0	1
med	0	1	0

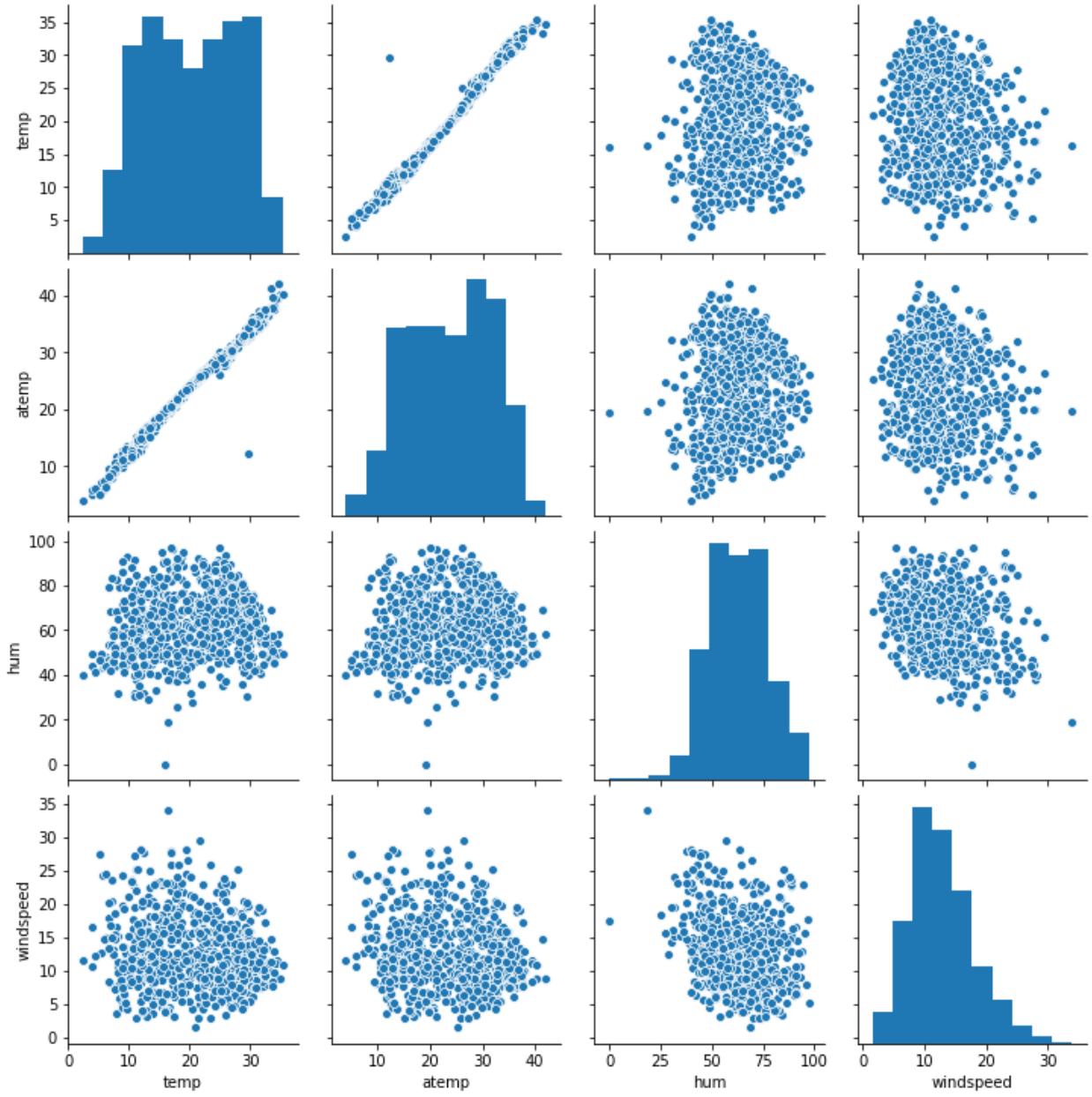
cat	med	high
low	0	0
med	1	0
high	0	1
med	1	0

From the first table we can derive a perfect collinearity by the equation $1 = low + med + high$
This leads to induced multicollinearity.

Moreover, same amount of information is retained in the second table as the first table, where 0 in *med* and 0 in *high* represent the category *low* value. The effect of *low* on the target is baked into the base equation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

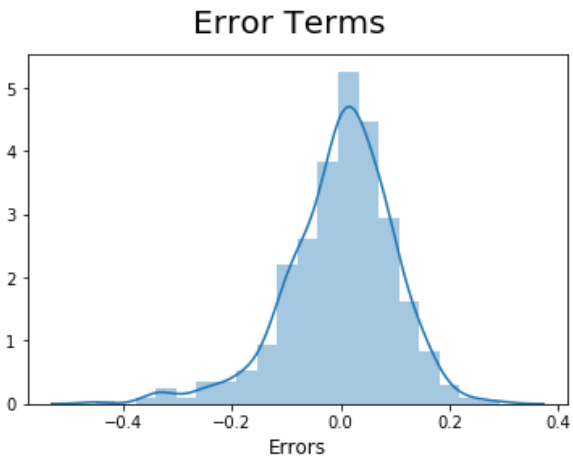
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
temp and ***atemp*** have a high correlation



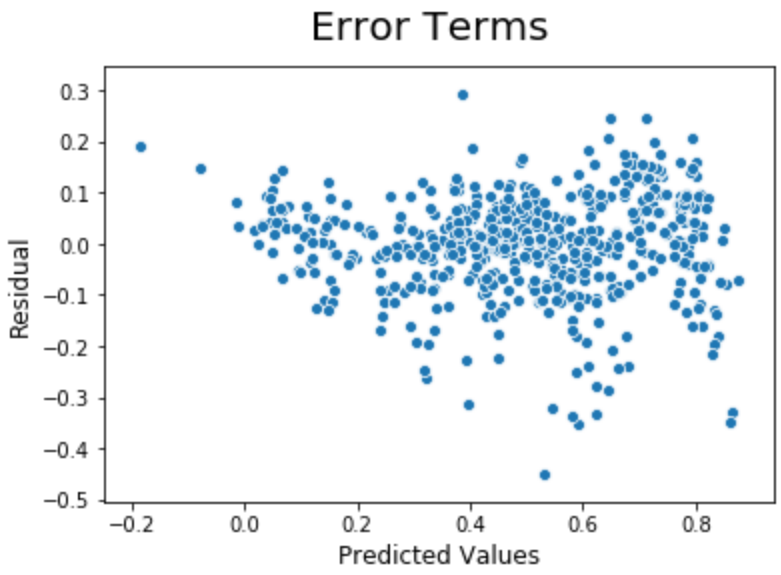
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

After fitting the model to the training we had to check the following three assumptions :

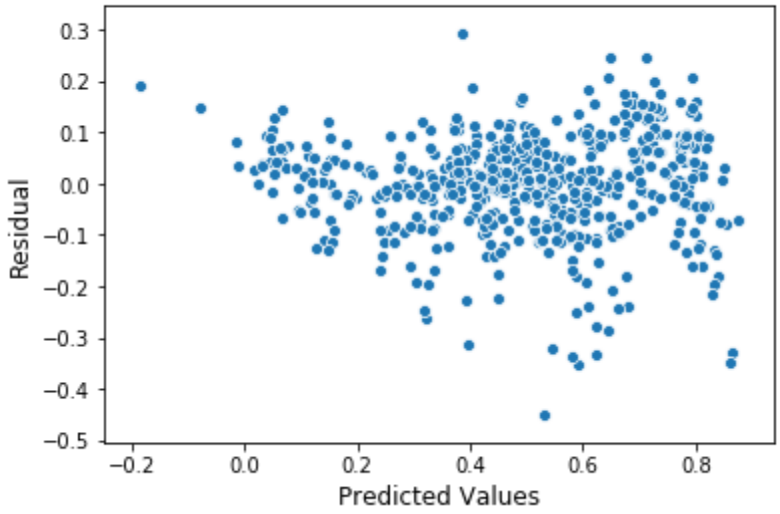
1. Error terms are normally distributed



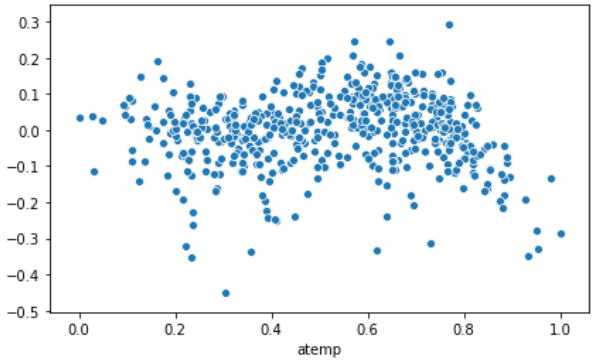
2. Error terms are independent of each other (no pattern) (no sign of auto-correlation)

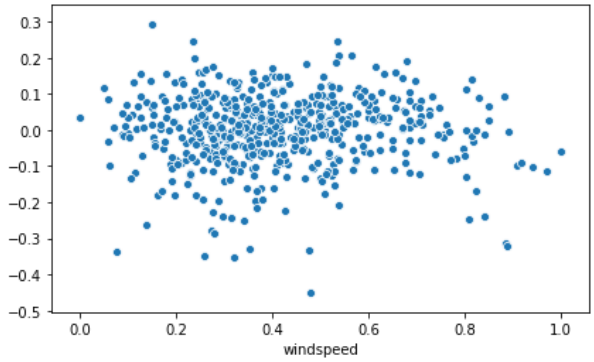


3. Error terms are showing Uniform Variance




```
mean of residuals when yr 3.5562444234017825e-16
mean of residuals when not yr 6.949407379519209e-17

mean of residuals when holiday 1.491862189340054e-16
mean of residuals when not holiday 2.2230595686114834e-16

mean of residuals when season_Spring 1.3655743202889425e-17
mean of residuals when not season_Spring 2.7781884218108184e-16

mean of residuals when season_Winter 2.3592239273284576e-16
mean of residuals when not season_Winter 2.1263362351236106e-16

mean of residuals when mnth_Mar 8.405974329304756e-17
mean of residuals when not mnth_Mar 2.2676088437534325e-16

mean of residuals when mnth_Sep 3.8510861166685117e-16
mean of residuals when not mnth_Sep 2.0296612308903788e-16

mean of residuals when weathersit_LightSnow/Rain 5.387847031269142e-17
mean of residuals when not weathersit_LightSnow/Rain 2.2461068644421207e-16

mean of residuals when weathersit_Mist/Cloudy 1.3818733391610992e-16
mean of residuals when not weathersit_Mist/Cloudy 2.586912423673937e-16
```





5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1) ***Atemp*** - coeff of 0.45
2) ***weather_LightSnow/Rain -*** coeff of -0.2666
3) ***Yr -*** coeff of 0.2456

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2140	0.027	7.824	0.000	0.160	0.268
yr	0.2456	0.009	28.673	0.000	0.229	0.262
holiday	-0.0849	0.025	-3.334	0.001	-0.135	-0.035
atemp	0.4505	0.033	13.665	0.000	0.386	0.515
windspeed	-0.0739	0.024	-3.066	0.002	-0.121	-0.027
season_Spring	-0.1545	0.017	-9.211	0.000	-0.187	-0.122
season_Winter	0.0499	0.013	3.868	0.000	0.025	0.075
mnth_Mar	0.0592	0.018	3.268	0.001	0.024	0.095
mnth_Sep	0.0744	0.015	4.821	0.000	0.044	0.105
weathersit_LightSnow/Rain	-0.2666	0.025	-10.552	0.000	-0.316	-0.217
weathersit_Mist/Cloudy	-0.0799	0.009	-8.779	0.000	-0.098	-0.062

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

It is expressed as the equation below,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Where,

Y is the target/dependent Variable
Xi is the Independent
βi Is the weight of impact the independent variable has on the target

2. Explain the Anscombe's quartet in detail. (3 marks)

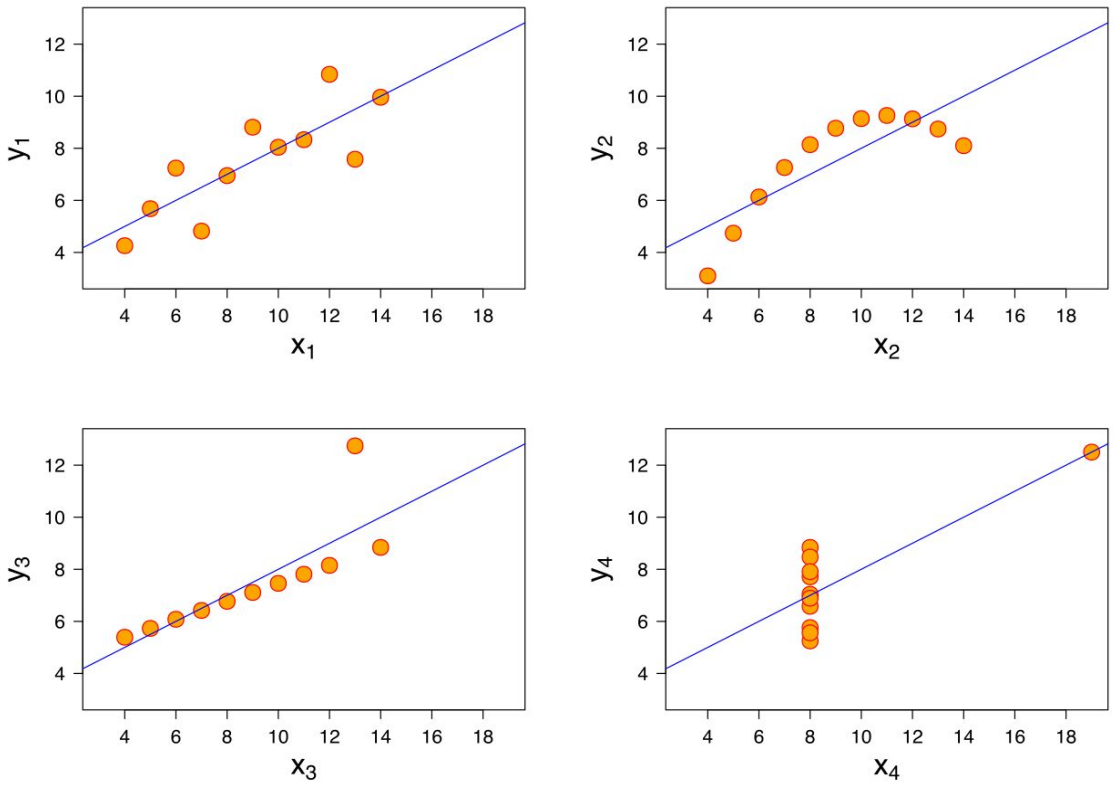
Anscome Quartet is a set of Four (x,y) data sets which have the same variance, mean and

Descriptive statistics. Although they have the same descriptive statistics, they are entirely different data when visualized. The difference between them is explain below!

It was made to prove to the statisticians of the time that data needs to be visualized irrespective of the descriptive statistics!

Anscombe's quartet

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



The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R? (3 marks)

Pearson's R or Pearson correlation coefficient, is a statistic that measures the degree to which two variables affect each other linearly. It is a measure of the linearity between two variable

It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$	Where, $Cov(x,y)$ is the Covariance of x,y σy is the Standard Deviation of y σx is the Standard Deviation of x

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

What is scaling :
Scaling is the method of bringing the features within the same range.
Why do we use scaling :
This helps in increasing the interpretability of the model. Since after scales the features are within the same range, the coefficients are comparable. This helps in gaining useful business insights! Higher coefficients have higher impact on the target variable

Standardization vs Normalization (Scaling) :
Normalization is a type of scaling where the data is converted to a range within 0 to 1. This can be achieved by subtracting the data from it mean and dividing it by the range of the feature

$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$	Where, $min(x)$ is the minimum value of the feature x is the value at the i^{th} data point $max(x)$ is the maximum value of the feature

Standardization is a type of scaling which scales the feature to the z-score. Thus it may range beyond 1 and 0.

$z = \frac{x_i - \mu}{\sigma}$	Where, μ is the mean of the data x_i is the value at the i^{th} data point σ is the Standard Deviation

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

This happens when one feature is entirely explained by the other features linearly.
For ease of understanding, let's assume three variables A,B and C.

B and C can explain the entire variation in A. In such a case the R^2_A value will be 1. And the Given formula for VIF_A will become Infinite

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

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks).

Q-Q plot is a graphical approach to comparing the probability distribution of two samples/features. It uses the quantile values of the feature and compares it with the quantile of the second feature. The second feature can be a standard distribution such as Gausian, Normal, Uniform, Posson etc.

In linear Regression :
We can use a Q-Q normal distribution plot to compare the residual to a standard normal distribution and if the QQ normal plot is a straight line. It means that our residual has a normal distribution and it satisfies one of the assumptions.