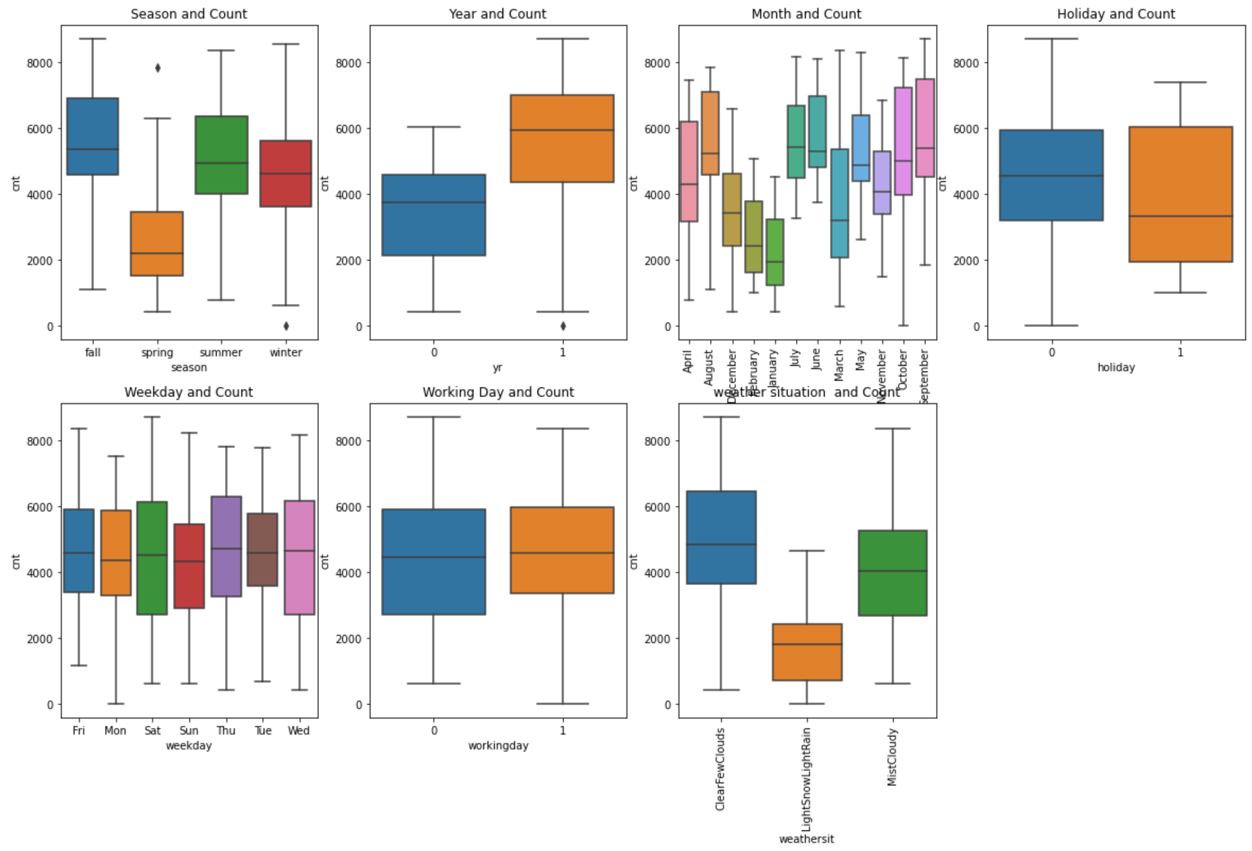
Assignment-based Subjective Questions

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

   Answer:
   Before answering the question let's draw box plots for different categorical variables used in this assignment



Observations from the above boxplots for categorical variables:

- The year box plots indicate that more bikes were taken during 2019 (yr=1).

- The season box plots indicate that more bikes were taken during fall season.
- The working day and holiday box plots indicate that more bikes were taken during normal working days than on weekends or holidays (non-working day).

  This finding was derived using the following observations

  Though the mean seems almost the same

  The 25% line of working is higher than of non-working day.

  The 75% line is the same for both

  This can also be validated using the following command

  ```
  bikes['workingday'].value_counts()
            1     499

            0     231
  ```

Note: Here 1 is a working day and 0 is a non-working day

- The month box plots indicate that more bikes are rented during September month.
- The weathersit box plots indicate that more bikes are rented during Clear, Few clouds weather.

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

   Answer: Suppose we have two values in columns yr 0 and 1. When we create dummy variables it will create two variables say yr_0, yr_1; we can drop the first variable (yr_0) as a second variable (yr_1) can hold both values such as when 1 then yr=1 otherwise yr=0.

   In a nutshell for n number of levels we need n-1 columns therefore we drop the First Column

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

   with the target variable? (1 mark)

   Answer: `atemp` has the highest correlation. This can be validated by using the following command

   ```
   bikes.corr()[cnt].sort_values(ascending=False).reset_index()
   ```

	index	cnt
0	cnt	1.000000
1	atemp	0.630685
2	temp	0.627044
3	yr	0.569728
4	workingday	0.062542
5	hum	-0.098543
6	windspeed	-0.235132

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

Answer: The three assumptions of Linear Regression are explained below:

1. **Linear Relationship between the features and target**

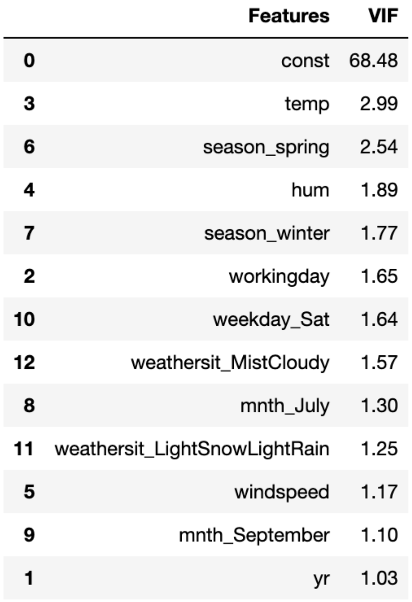
   From R-Sqaured and adj R-Sqaured value of the trained dataset, we could conclude that the above variables can well explain more than 80% of bike demand.

   R- Sqaured train: 0.84

   Adj. R-Squared train: 0.838

2. **Little or no Multicollinearity between the features**

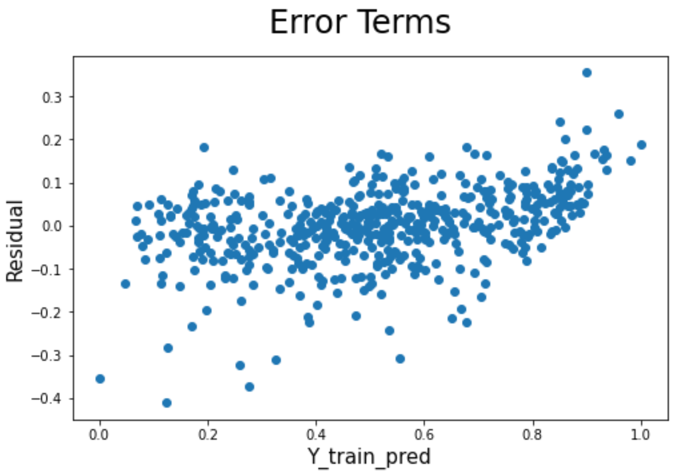
   This can be validated by viewing the VIF for the training set. All the features' VIF values are less than 5.



	Features	VIF
0	const	68.48
3	temp	2.99
6	season_spring	2.54
4	hum	1.89
7	season_winter	1.77
2	workingday	1.65
10	weekday_Sat	1.64
12	weathersit_MistCloudy	1.57
8	mnth_July	1.30
11	weathersit_LightSnowLightRain	1.25
5	windspeed	1.17
9	mnth_September	1.10
1	yr	1.03

3. **Homoscedasticity**

   Homoscedasticity has been validated by drawing a scatter plot between the training set and residuals.



Insights:
- It seems like the corresponding residual plot is reasonably random.
- The error terms satisfy to have reasonably constant variance (homoscedasticity)

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Answer:

Based on final model top three features contributing significantly towards explaining the demand are:

1. Temperature: (0.479)
2. weathersit: LightSnowLightRain+MistCloudy (-0.309)
3. year: (0.231)

General Subjective Questions

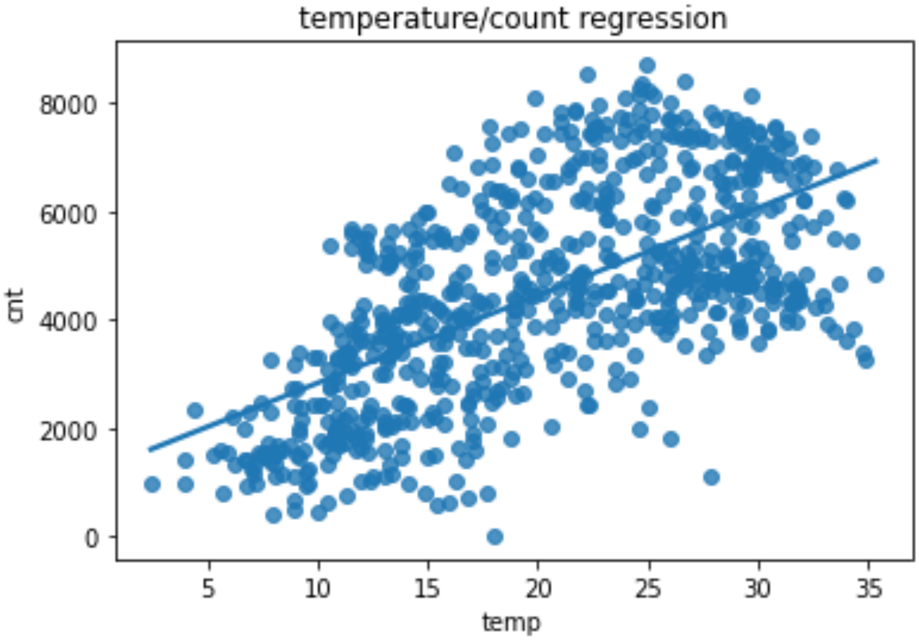
1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression (LR) is a Machine Learning (ML) algorithm used for supervised learning. LR can be used to predict dependent variable by using independent variables. Since the technique finds out the linear relationship hence the name is LR

Following is an example (taken from this assignment) to show linear regression between count (cnt) and temperature (temp)

Figure: temperature/count regression



In the figure above, we have ‘temp’ on x-axis and ‘cnt’ on Y-axis. The line shows the linear regression between the two variables.

Here are the pros and cons of RL:

Pros:

1. Linear Regression is very simple to implement.
2. Linear Regression may lead to overfitting, which can be avoided by using different techniques such as:

- dimensionality reduction techniques
- regularization techniques,
- and cross-validation.

Cons:

1. Outliers can have negative effect on this algorithm so should be removed before running the algorithm

2. Explain the Anscombe's quartet in detail.

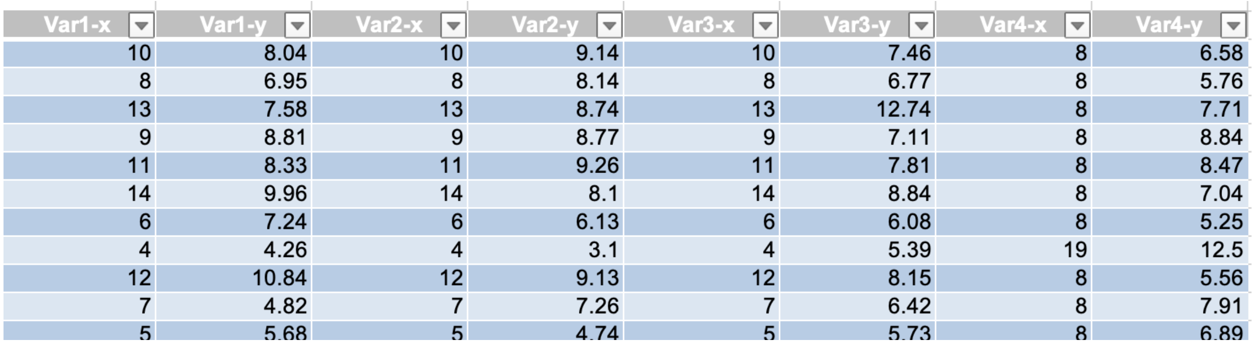
Answer:

Anscombe's quartet

In 1973 a statistician name Francis Anscombe demonstrated how descriptive statistics could be and the importance of graphical distributions. Francis took 4 different datasets with 11 rows.

If we create an excel sheet the dataset will look like this:

Figure: Dataset

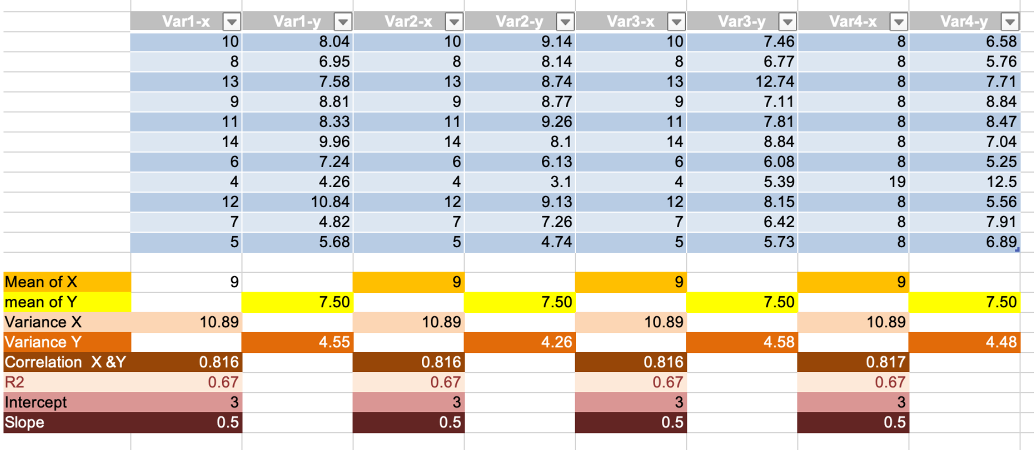


Var1-x	Var1-y	Var2-x	Var2-y	Var3-x	Var3-y	Var4-x	Var4-y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	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

Let's calculate the mean, variance, correlation, R2, intercepts, and slopes

It will look like this:

Figure: Dataset with statistical values



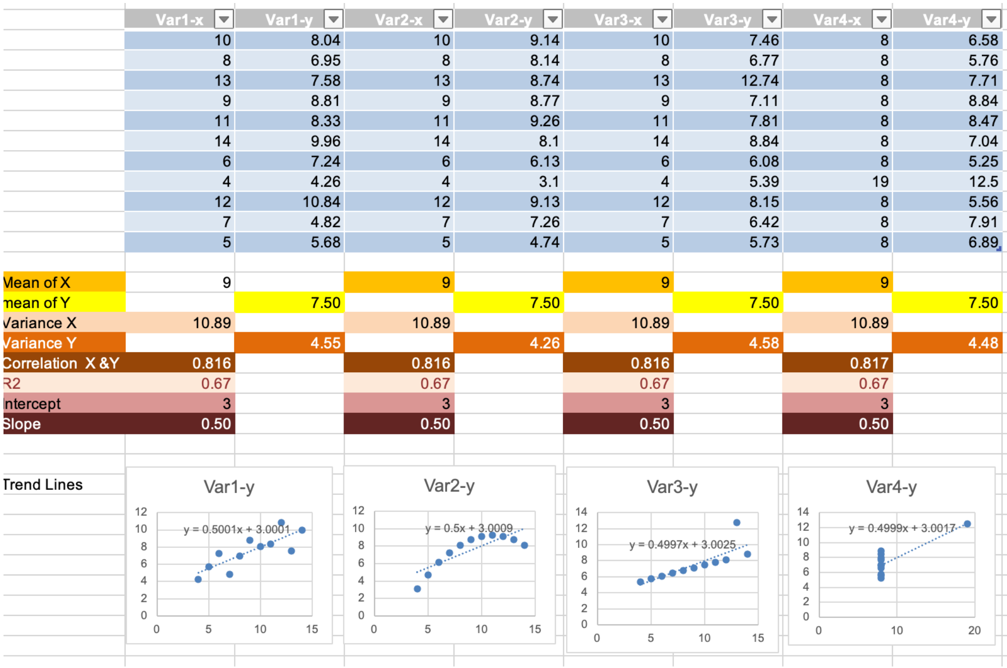
	Var1-x	Var1-y	Var2-x	Var2-y	Var3-x	Var3-y	Var4-x	Var4-y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	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
Mean of X	9		9		9		9	
mean of Y		7.50		7.50		7.50		7.50
Variance X	10.89		10.89		10.89		10.89	
Variance Y		4.55		4.26		4.58		4.48
Correlation X &Y	0.816		0.816		0.816		0.817	
R2	0.67		0.67		0.67		0.67	
Intercept	3		3		3		3	
Slope	0.5		0.5		0.5		0.5	

Note: The values are color codes so that you can easily focus on one thing e.g., Slope

From the figure above we can see:

1. The mean of X and y is the same (two decimal places)
2. Variance of X is same (two decimal places)
3. Variance of Y is very similar (though has some differences)
4. The correlation between X & Y is very similar (only one difference)
5. R2 is same (two decimal places)
6. Intercept is the same (two decimal places)
7. Slope is the same (two decimal places)

Let’s draw scatter graphs for all four (4) datasets



	Var1-x	Var1-y	Var2-x	Var2-y	Var3-x	Var3-y	Var4-x	Var4-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
Mean of X	9		9		9		9	
mean of Y		7.50		7.50		7.50		7.50
Variance X	10.89		10.89		10.89		10.89	
Variance Y		4.55		4.26		4.58		4.48
Correlation X &Y	0.816		0.816		0.816		0.817	
R2	0.67		0.67		0.67		0.67	
Intercept	3		3		3		3	
Slope	0.50		0.50		0.50		0.50	

Trend Lines

The data looks very different in the scatter graph which is because of outliers

Please see the brief explanation of each chart:

Var1-Y	Var2-Y	Var3-Y	Var4-Y
Looks like a linear regression	Non-linear	very linear -Almost all the points are on a linear line	almost a straight line

References:

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

3. What is Pearson’s R?

Answer: Pearson R is a system that is used to:

1. identify a linear relationship numeric variable.
2. The direction
3. Strength

A value greater than 0.5 shows that an increase in one variable will increase the value of the other as well. In this assignment for example there is a positive correlation between variables temp and ‘cnt’

A value less than -0.5 means an increase in one feature will cause a decrease in the other. For example, in this assignment, there is a negative correlation between ‘cnt’ and windspeed

Some of the conditions for using Pearson’s R:

1. All features are quantitative
2. features are normally distributed
3. The data has no outliers
4. The relationship is linear – the relationship can be described well by a line

The value and corresponding strength and direction for Pearson’s Rare explained in the following table:

Table: correlation values, strength, and the direction

Value	Strength	Direction
>0 .5	Strong	Positive
>0.3 and <0.5	Moderate	Positive
> 0 and <0.3	Weak	Positive
0	None	None
<0 and >–0.3	Weak	Negative
<–0.3 and > –0.5	Moderate	Negative
< –0.5	Strong	Negative

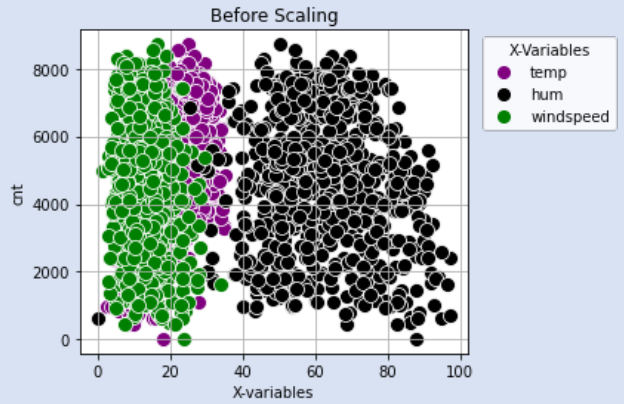
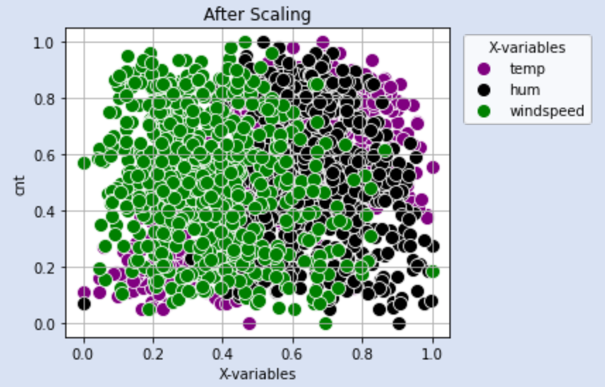
Note: the value of R can range between -1 and +1

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

Answer:

Scaling is a process of unifying different scale variables on a single scale. For example, in this Assignment, we saw variables such as temp, hum, and windspeed. Now all these variables have a very different scale and if we try to create a scatter graph they will almost be shown as silos, and it is very difficult to find relationships.

However, when we applied to scale (min-max in this case) data seems to be more unified as shown below

Before applying scaling (Min,Max)	After
	

What is Normalization (Min-Max Scaling)?

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here’s the formula for normalization:

Xs =(X-Xmin)/(Xmax-Xmin)

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively and Xs=scaled value of the feature.

What is Standardization?

Standardization is a scaling technique which transformed the values in such a way that:

1. The values are centered around the mean which is zero
2. Values have a unit standard deviation.

Here's the formula for standardization:

$$Xs= (X – Xmean) / Xsd$$

Xmean and **Xsd** are the mean and standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

   (3 marks)

To answer this question let's review the formula of VIF:

VIF =1 / (1-R2)

Now the VIF will be infinite if the denominator becomes zero which is only possible when R2 will become 1 (a perfect positive correlation) as illustrated below:

VIF=1/(1-R2)
VIF=1 / (1-1)
VIF=1/0
VIF=infinity

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 an abbreviation for Quantile- Quantile plot. Q-Q Plot helps us to find out if a set of data possibly came from a distribution such as:

1. Normal
2. Exponential
3. Uniform

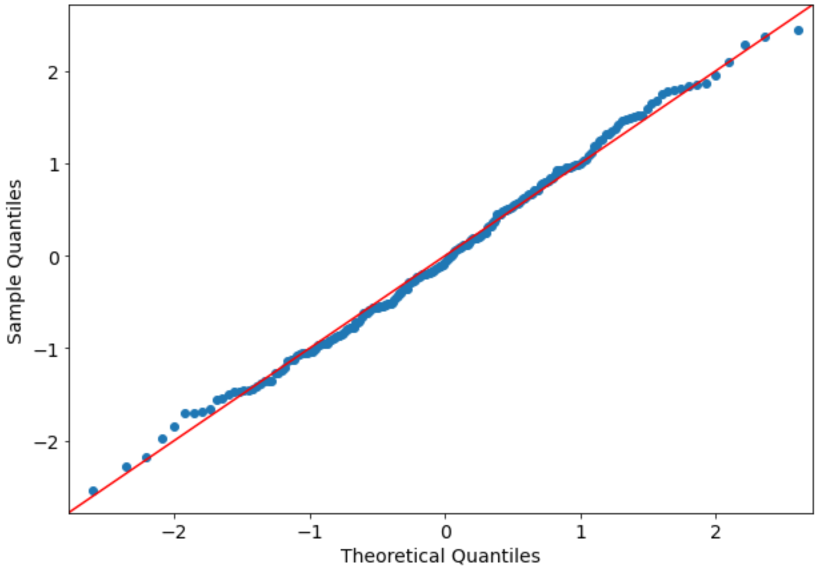
This plot helps us verify in scenarios, such as this assignment, that training and test datasets are from populations with the same distributions.
In this plot, we may compare the theoretical quantiles of a population with the test dataset quantiles.

If the test dataset has been generated from the population, we expect this chart to be close to the 45-degree line, because the sample quantiles will be like the population quantiles.

Here is an example of generating a qqplot in python using ` hum` variable of our test dataset:

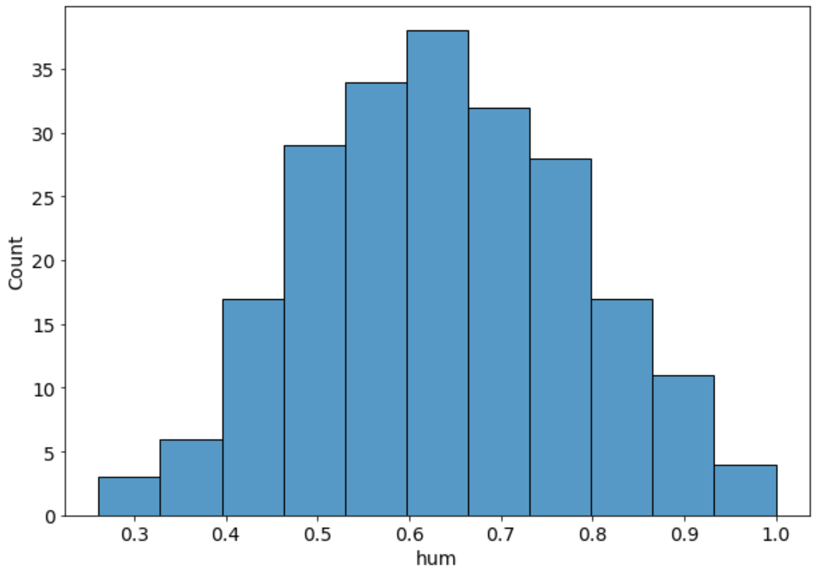
```
from statsmodels.graphics.gofplots import qqplot
from matplotlib import pyplot as plt
plt.rcParams['figure.figsize'] = [10, 7]
plt.rc('font', size=14)
from scipy.stats import norm, uniform
import numpy as np

qqplot(df_test['hum'],norm,fit=True,line="45")
plt.show()
sns.histplot(df_test['hum'])
plt.show()
```



From the Figure above most of the points are very close to 45-degree line (the red line). This shows the normal distribution of the data.

This was further validated by creating a histogram plot for the same data and you can observe the distribution is normal



Interpretation:

Below are the possible interpretations for two data sets.

Similar distribution:	If all point of quantiles lies on or close to the red line
Y-values < X-values:	If y-quantiles are lower than the x-quantiles.
Y-values > X-values:	If x-quantiles are lower than the y-quantiles.
Different distribution:	If all point of quantiles lies away from the red line