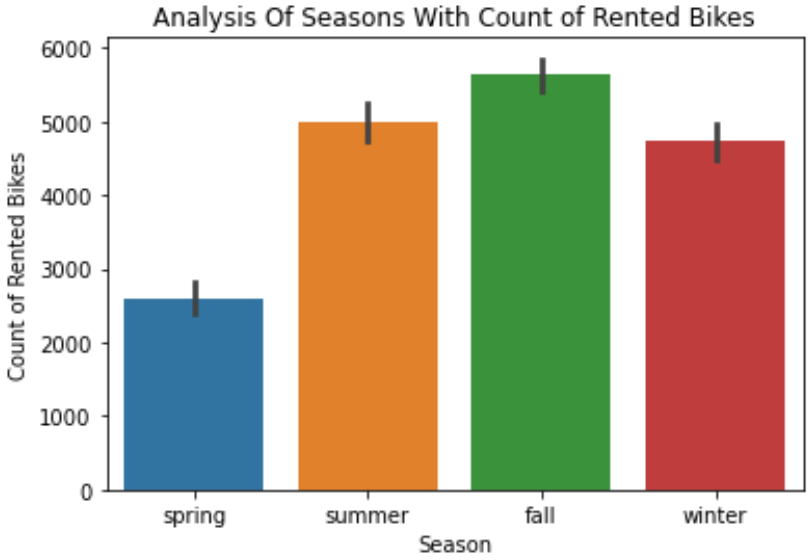
BoomBikes (Bike Sharing Assignment) Subjective Questions

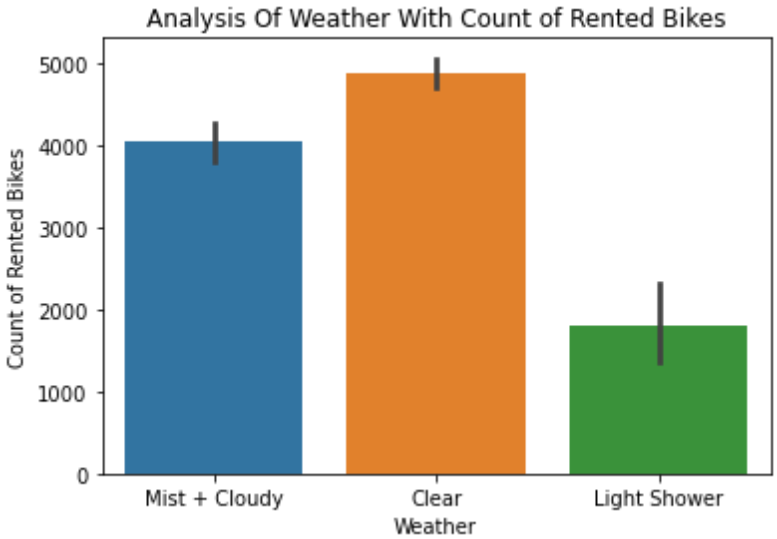
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Ans.

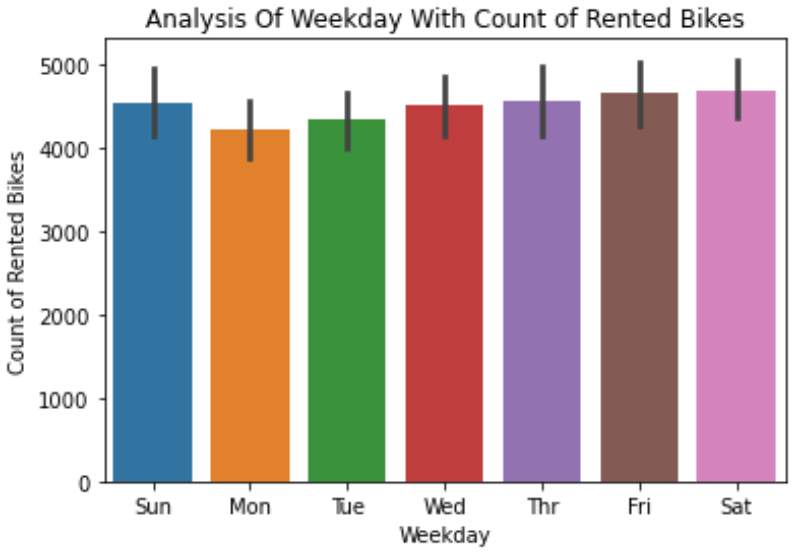
   - The categorical variables for the provided dataset **are Season, Month, Weekday, Year, Weather, Working day and Holiday.**
   - Let's look at the following graphs to infer about their effect on the dependent variable Count



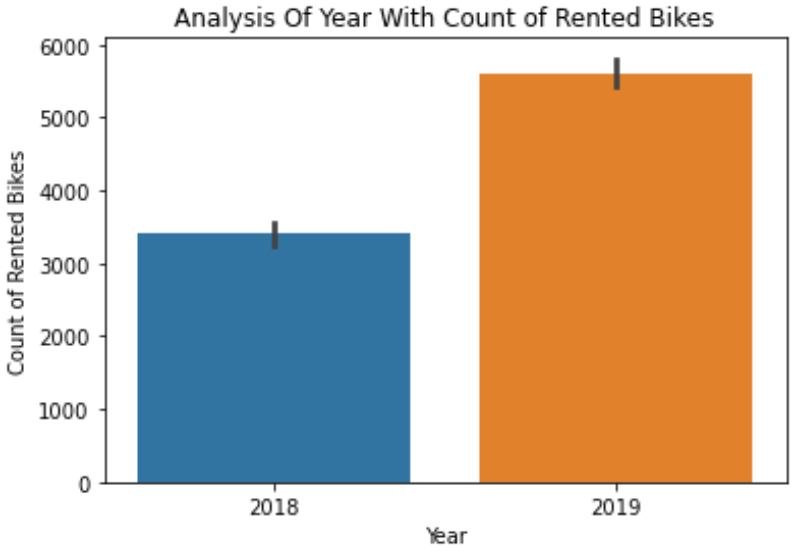
   - From the above graph we can say that, **People prefer to rent bikes in Fall Season mostly**



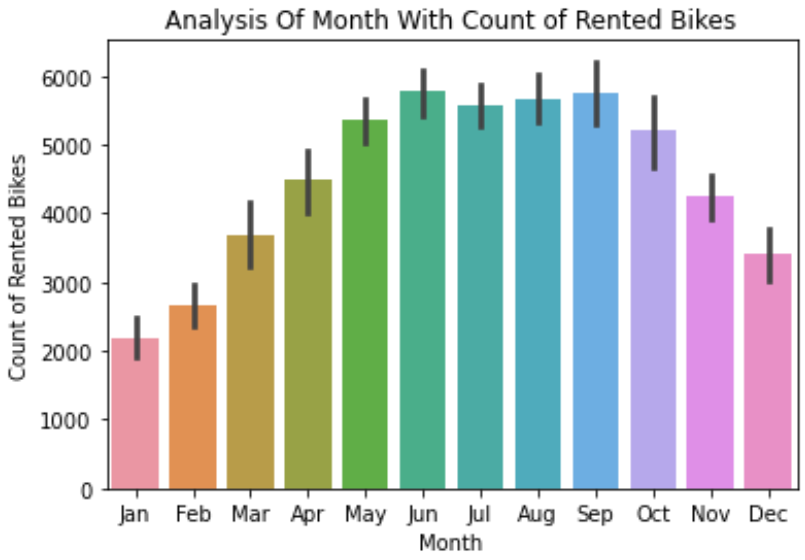
   - From the above graph we can say that, **People prefer to rent bikes in Clear Weather**



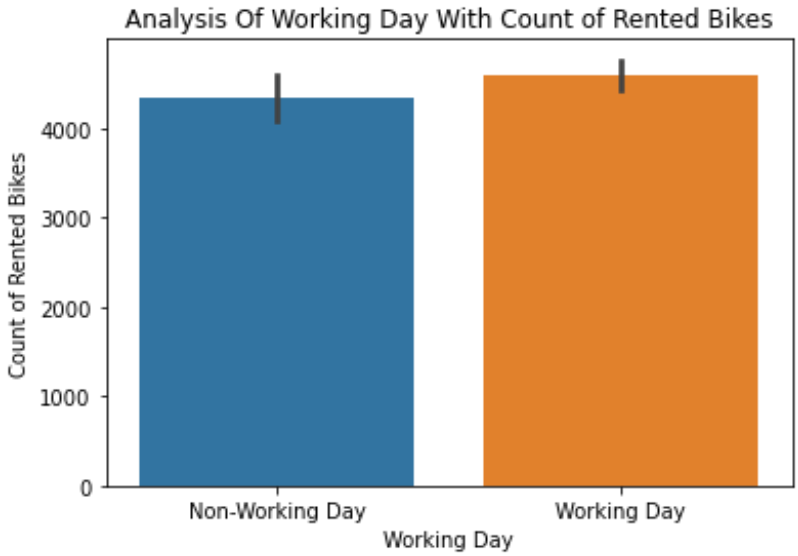
- From the above graph, we can say that, **People prefer to use rented bikes on almost everyday**



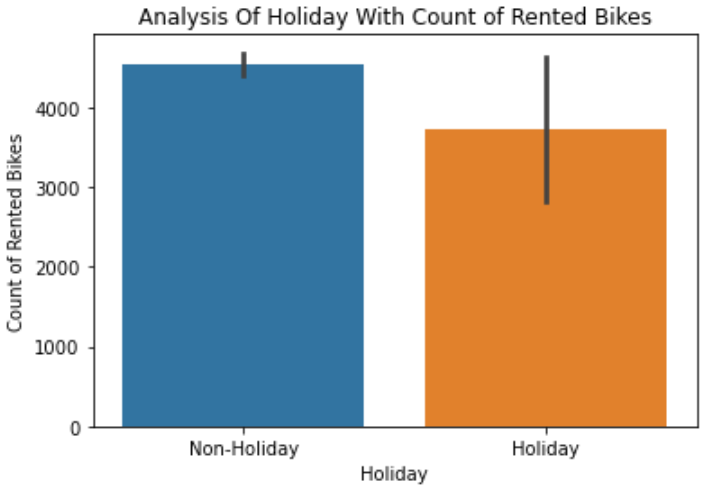
- From the graph above we can say that, **People have started renting bikes more recently and the demand in increasing for the same**



- From the above graph we can say that, **People prefer to rent bikes mostly in Jun to Sep period as that is season of fall**



- From the above graph we can say that, **People prefer to use rented bikes on Working Day then the Non-Working Days**



- From the above graph we can say that, People prefer to use the rented bikes more on working days i.e. non-holidays

Inference Summary:

- **It can be said that from all above inferences that the demand for rented bikes is increasing rapidly year on year at the rate of 65% approx .**
- **People would love to rent and drive bike everyday while going for work, more likely in the season of Fall starting from month June to September with Clear weather conditions**

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

Ans.

- Creation of dummy variables without specifying the drop_first=True parameter, creates the total of N columns where N is the no of unique values in a column.
    - For e.g. season column in the current dataset has 4 unique values (1-4), so if we create dummy variables without specifying the drop_first=True parameter in the function pd.get_dummies(bikes['season']), it will create 4 columns each representing the unique value of the column season.
- Creation of dummy variables by specifying this parameter (drop_first = True) will create N-1 columns and will have same significance as earlier.
    - For e.g. column season has 4 unique values (spring, summer, fall, winter),
    - so if we mention drop_first=True in pd.get_dummies(bikes['season'], drop_first=True),
        - it will create three columns summer, fall & winter.
        - For any record if the values of all this 3 columns will be 0 then it will mean that the value for that record will be spring
- Creation of columns to signify the same value creates more complexity also unnecessary memory utilization in a Dataframe. To reduce the complexity and memory consumption it is important to use drop_first=True

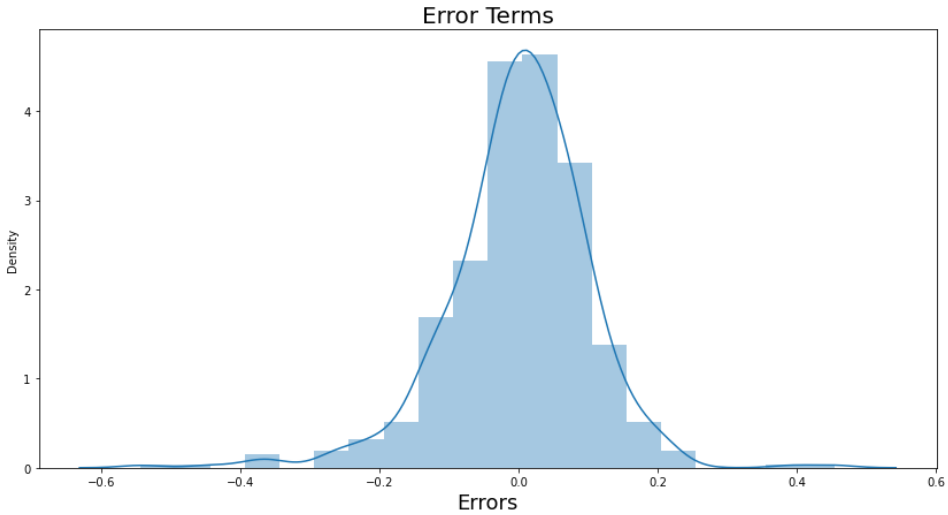
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Ans. Column **aTemp** has the highest correlation with the target variable **cnt**.

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

   Ans. Following steps were performed to validate the assumptions of Linear Regression

    - **Step 1: Performing Residual Analysis**
        - **Residual Analysis to validate that the errors (y_train values – y_predicted values) follow a normal distribution as shown in the diagram below**



- **Step 2: Performing test data validation**
  - **Evaluating the test data using the final model. Checking whether the predicted values and actual test values match and fall on the regression line.**

- **Step 3: R Square Value and Adjusted R Squared Value Matches the model metrics**
  - R Square Value 0.7878976207612222 is matching the R Square value of model (0.799)
  - Adj R Square Value 0.7702217499999999 is matching the Adj R Square vale of model(0.770)

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

Ans.

Top 3 Features contributing to the demand of shared bikes are **Year 2019, August & June month**

- This can be inferred from the Final Model's OLS Summary which provides the co-efficients.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.794
Model:	OLS	Adj. R-squared:	0.788
Method:	Least Squares	F-statistic:	136.1
Date:	Wed, 07 Apr 2021	Prob (F-statistic):	1.75e-159
Time:	14:38:45	Log-Likelihood:	441.10
No. Observations:	510	AIC:	-852.2
Df Residuals:	495	BIC:	-788.7
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2599	0.015	16.949	0.000	0.230	0.290
winter	0.1620	0.017	9.812	0.000	0.130	0.194
2019	0.2453	0.009	26.594	0.000	0.227	0.263
Apr	0.0766	0.021	3.640	0.000	0.035	0.118
Aug	0.2131	0.020	10.727	0.000	0.174	0.252
Dec	-0.1182	0.019	-6.064	0.000	-0.157	-0.080
Feb	-0.1351	0.022	-6.112	0.000	-0.178	-0.092
Jan	-0.1844	0.020	-9.064	0.000	-0.224	-0.144
Jul	0.1832	0.021	8.619	0.000	0.141	0.225
Jun	0.2064	0.021	9.624	0.000	0.164	0.249
May	0.1794	0.021	8.698	0.000	0.139	0.220
Nov	-0.1025	0.021	-4.908	0.000	-0.144	-0.061
Sep	0.2122	0.020	10.715	0.000	0.173	0.251
Clear	0.0818	0.010	8.267	0.000	0.062	0.101
Light Shower	-0.2216	0.028	-7.838	0.000	-0.277	-0.166

Omnibus:	84.425	Durbin-Watson:	1.922
Prob(Omnibus):	0.000	Jarque-Bera (JB):	332.665

General Subjective Questions

1. Explain the linear regression algorithm in detail
   Ans.

- To understand Linear Regression, we would understand Regression.
  - Regression simply means to perform some task recursively to get better results or closer results.
  - In Terms of Machine Learning, Regression is a method to model or predict the target variable or result based on dependent variables or predictors
  - Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.
- Linear Regression is a type of regression that involves one or more independent variables and there is linear relationship between independent and dependent variables.
- Linear Regression always has the defined target variable and many dependent variables and hence it forms the part of Supervised Learning Algorithm
- Linear Regression can be represented by the graph below:

- The red line referred in the graph above is the best fit line. The line can be modelled based on the linear equation shown below:
  - $\mathbf{y = \beta_0 + \beta_1 x}$
- Linear Regression tries to identify the best values of β_0 & β_1 using techniques like Residual Sum of Squares (RSS), Total Sum of Squares TSS, Gradient Descent Regression Algorithm
- This has two important concepts
  - **Cost Function:**
    - This function helps us identify the best possible values for β_0 & β_1 , so that we can find the best fit line

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

- This function can be minimized. The difference between the predicted and actual values of y is known as error difference. We square the error difference and sum all data points and divide that by total no of data points. This provides the average squared errors over all the data points. Therefore this cost function is known as Mean Squared Error(MSE) function.
- Using this function we can find the minimum values.

- **Gradient Descent Algorithm:**
  - In this method, we start with some random values of β_0 & β_1 and descent iteratively to find the β_0 & β_1 near the minima.
  - Gradient Descent Algorithm performs partial derivatives of the coefficients to reduce the cost and the stepping rate
  - For a linear regression gradient descent uses the following cost function

$$J(m,c) = \sum_{i=1}^{N}\left(y_i - (mx_i + c)\right)^2$$

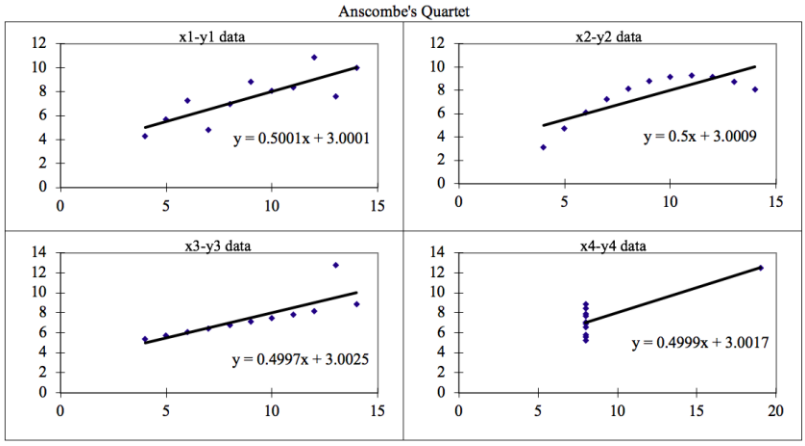
  - It reduces the cost function using partial derivative over m and c both

$$\frac{\partial J}{\partial m} = 2\sum_{i=1}^{N}\left(y_i - (mx_i + c)\right)(-x_i)$$

$$\frac{\partial J}{\partial c} = 2\sum_{i=1}^{N}\left(y_i - (mx_i + c)\right)(-1)$$

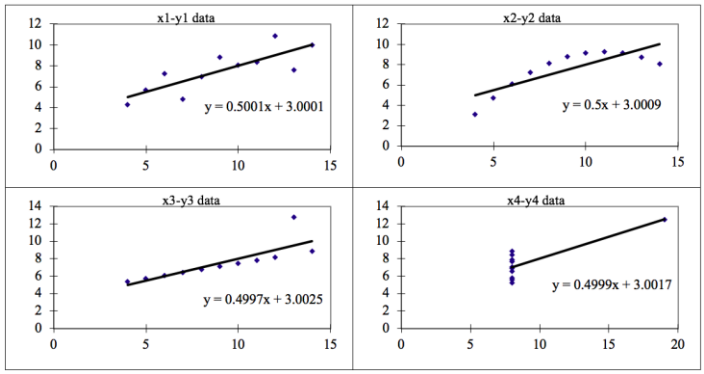
2. Explain the Anscombe's quartet in detail.

Ans.

- Anscombe quartet is a group of four data sets which are almost identical in descriptive statistics, but have different distributions and it appears differently when plotted on scatter plots.
    - For e.g. Refer the diagram below:
    - 

- It was constructed by statistician **Francis Anscombe** in 1973
- It was created to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.
- Lets see below the four data sets which have almost same statistical observations that involves variance and mean, of all x, y points in all the sets
    -

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	

- When we plot these models on scatter plot, all datasets generates different kind of plot that cannot be inferred to the output from any of the regression algorithm.
- This four data sets when plotted on scatter plot is displayed as below:
    - 

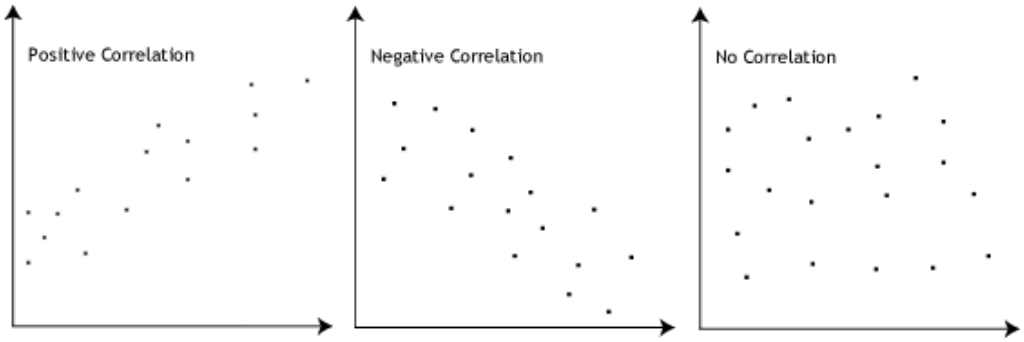
    - Dataset 1: this fits the linear regression algorithm

    - Dataset 2: this does not fit the linear regression algorithm as the data is non-linear
    - Dataset 3: this shows the outliers that cannot be handled by linear regression
    - Dataset 4: this shows the outliers that cannot be handled by linear regression
- This type of data graphs shows us the importance of data visualizations before modelling any regression algorithm so that such anomalies are not part of model and helps to create best fit model

3. What is Pearson's R?

Ans.

- Pearson's R is known as Pearson Correlation coefficient and denotes the measure of strength of linear association between two variables.
- This coefficient indicates how far away all the data points are from the best fit line.
- The values of this coefficient range from +1 to -1.
- A value of 0 indicates that there is no association between the two variables. Positive value denotes positive association. Negative value denotes negative association
- This is shown in the diagram below:



- The Strength of Association can be determined as follows:

Strength of Association	Coefficient, *r*	
	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

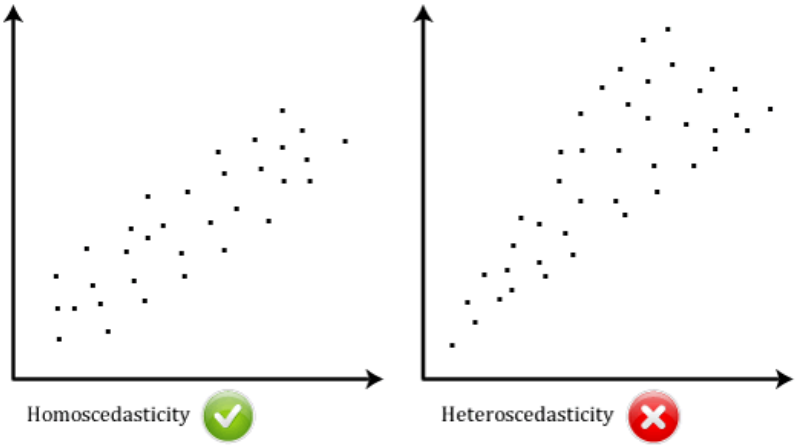
- Before selecting the Pearson's Coefficient to determine the statistical test between the two variables, the following assumptions are to be made :
    - Assumption 1: Two variables should be measured on a **continuous** scale (i.e., they are measured at the **interval** or **ratio** level). Examples of continuous variables include revision time (measured in hours), intelligence (measured using IQ score),etc

- Assumption 2: Two continuous variables must be paired, which means each case has two values: one for each variable. These values are also referred to as data points

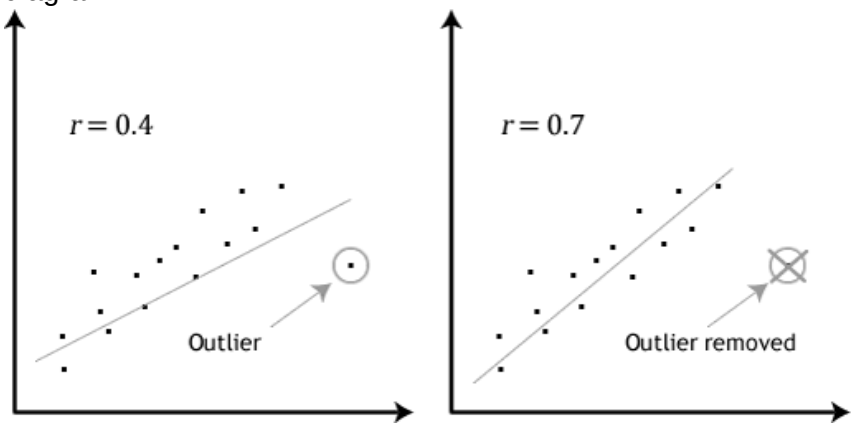
- Assumption 3: There should be independence of cases, which means two observations for one case (e.g., the scores for revision time and exam performance for "student #1") should be **independent** of the two observations for any other case (e.g., the scores for revision time and exam performance for "student #2", or "student #3", or "student #50", for example).

- Assumption 4: There should be linear relationship between two continuous variables. We can check this simply by plotting them on scatterplot

- Assumption 5 : There should be homoscedasticity, i.e. variance should be along the line of best fit. If variances are not similar then there is heteroscedasticity.



- Assumption 6: There should be no univariate or multivariate outliers. Outliers can have a very large effect on the line of best fit and the Pearson correlation coefficient, which can lead to very different conclusions regarding your data. This is demonstrated in below diagram



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

   Ans.

   - Scaling is a technique to standardize the independent features present in data in the fixed range
   - It is done in machine learning, while pre-processing the data to handle varying multitudes of values or units
   - If scaling is not done then machine learning algorithm tends to weigh greater values higher and lower values lower, regardless of the unit of values. For e.g. 100 meters will be considered higher than 1 km, although 1 km is the higher value
   - We have to use scaling in few algorithms like Neural network gradient descent where convergence is much faster with feature scaling then without using scaling
   - There are two types of Scaling:
     - **Normalized Scaling :**
       - It is technique in which the values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max Scaling

       $$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

       - X_{max} and X_{min} are the maximum and minimum values of the feature respectively.

     - **Standardized Scaling :**
       - It is the technique where the values are centered around the mean with a unit standard deviation.
       - The formula for Standardized Scaling is given by:

       $$X' = \frac{X - \mu}{\sigma}$$

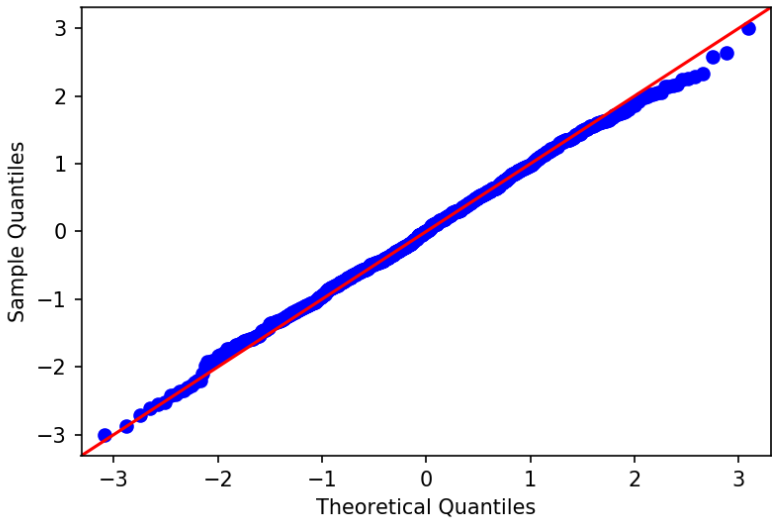
       μ Is the mean of feature values and σ is the standard deviation of the feature values
   - Normalization distribution can be used when distribution of data does not follow Gaussian distribution. This can be used in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks
   - Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans.
   - Variance Inflation Factor(VIF) provides a number summary description of collinearity for each model term
   - VIF is given by **VIF = 1 / 1 – R^2**
   - R^2 is the coefficient of determination of determination of an regression model
   - If the R^2 value is 1 then the VIF value can be infinity. i.e. when the coefficient of determination of an regression model is perfect fit it has VIF value of Infinity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans.
   - Q-Q plots are plots of two quantities against each other
   - The purpose of Q-Q plots is to find out if two sets of data come from same distribution.
   - It is used to visualize whether the data is normally distributed
   - If the data or errors are normally distributed then the value of data will fall say 95% of the time between -1.96 and +1.96 (standard deviations of the mean)
   - Let's make up some data that we already know is normally distributed:

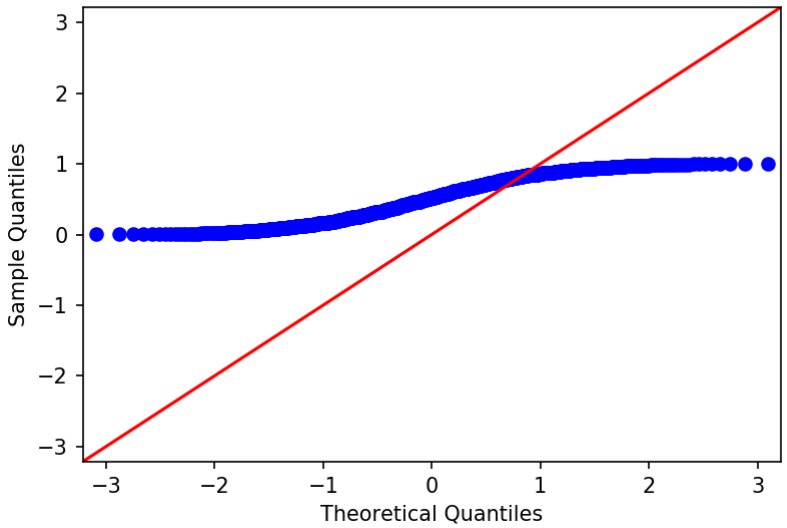
```
import numpy as np
# Generate some normally distributed random numbers
random_normals = [np.random.normal() for i in range(1000)]

import statsmodels.api as sm
from matplotlib import pyplot as plt# Create QQ plot
sm.qqplot(np.array(random_normals), line='45')
plt.show()
```

This code above creates the following graph:



Let's take a look at the QQ plot for something that's not normal:

```
import random# Generate some uniformly distributed random variables
random_uniform = [random.random() for i in range(1000)]# Create QQ plot
sm.qqplot(np.array(random_uniform), line='45')
plt.show()
```



- How do QQ plots work?
  - **QQ plot compares the quantiles of our data against the quantiles of the desired distribution** (defaults to the normal distribution, but it can be other distributions too as long as we supply the proper quantiles).
  - **Quantiles are breakpoints that divide our numerically ordered data into equally proportioned buckets.** For example, you've probably heard of percentiles before — percentiles are quantiles that divide our data into 100 buckets (that are ordered by value), with each bucket containing 1% of observations.
- What do QQ plots infer?
  - If the dots fall on 45 degree line, then the data is normally distributed
  - The slope tells us whether the data is too small or too big
  - A steepy slope means the data is more the data is more spread out