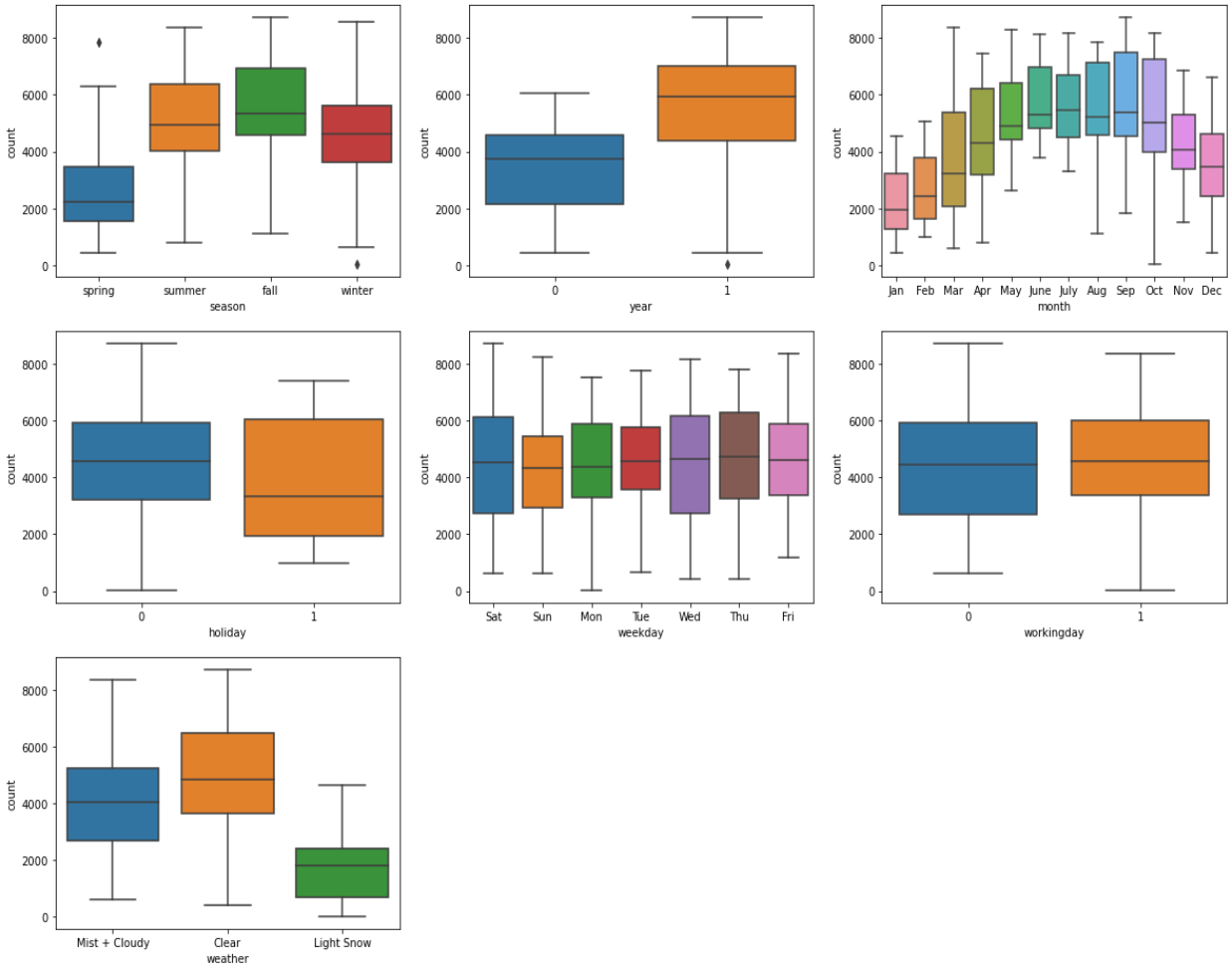
Assignment-based Subjective Questions

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

   Answer : - Categorical variables from the dataset are

   1. **Season 2. Year 3. Month 4. Holiday 5. Weekday 6. Workingday 7. Weather**

   These variables are visualize using bar plot and Box plot both. target variable is count.



Inferences:-

- ✓ fall season (season 3) has Highest demand for rental bikes while, spring season has lowest.
- ✓ its observed that demand for next year (from 2018 to 2019) has increased.
- ✓ Bike Sharing Demand is continuously increasing from January to June, while September month has highest demand among all. after September demand start decreasing.
- ✓ whenever there is holiday started demand has been decreased.
- ✓ weekday has almost same demand. its not much concluding demand during weekday.
- ✓ its looks like similar demand during working day and non-working day, but still working day have more demand.
- ✓ During september month Bike sharing demand is more. while during year begning and end its less demand. it could be due to Extreme weather conditions.

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

Answer: - drop_first=True, its helps to reduce extra created column during dummy variable creation. Which will helps in reducing correlations (redundancy) among dummy variable.

- If we have n level of categorical variable then we need to use (n-1) columns to represent the dummy variable.
- In our data suppose we have data holiday with yes or no in this case we have two features like is holiday yes and is holiday no, but using these two is redundant. We just need to use one of them other will just opposite of the other.
- Another example of house data set as on our content we learnt

  Before not using drop_first = True

```
status = pd.get_dummies(housing_data['furnishingstatus'])
status.head()
```

	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0

Here We don't need unfurnished status column as furnished & semi-furnished both "0" means unfurnished same can be applied to any variable

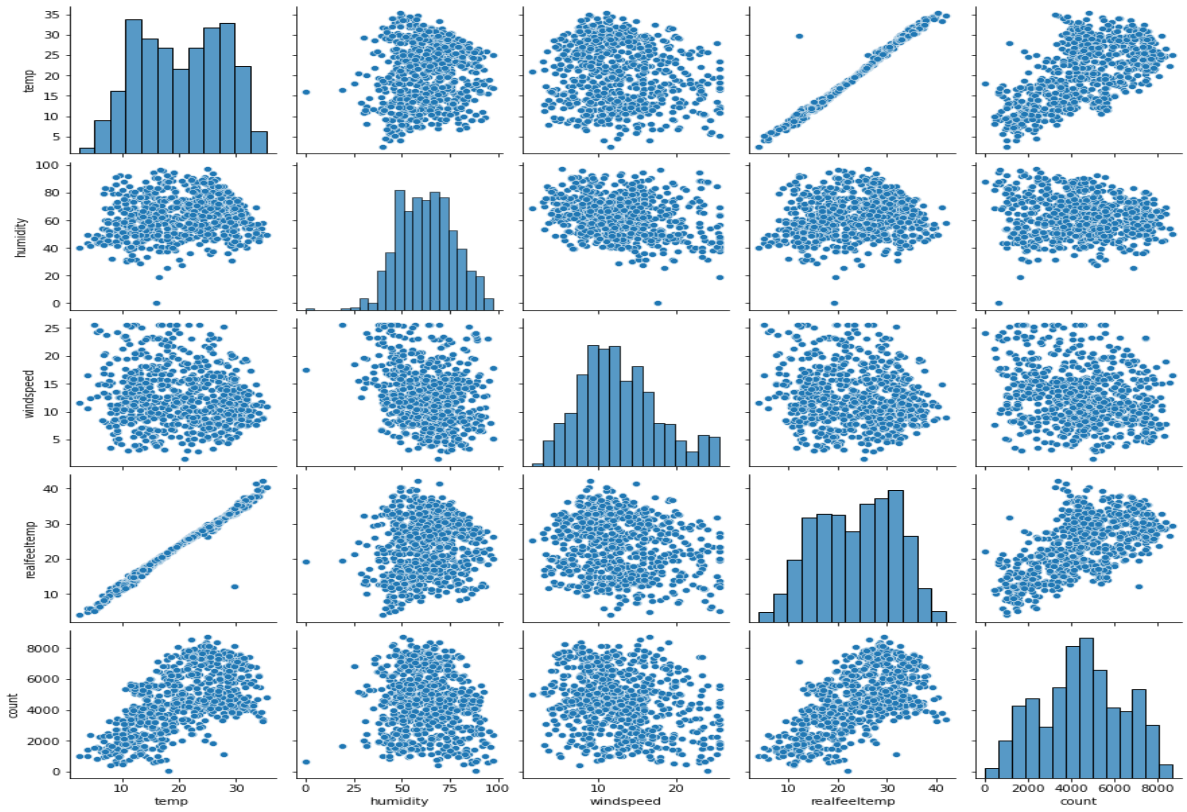
After using drop_first = True

```
_dummy_vars = pd.get_dummies(housing_data['furnishingstatus'], drop_first=True)
_dummy_vars.head()
```

	semi-furnished	unfurnished
0	0	0
1	0	0
2	1	0
3	0	0
4	0	0

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

Answer : -



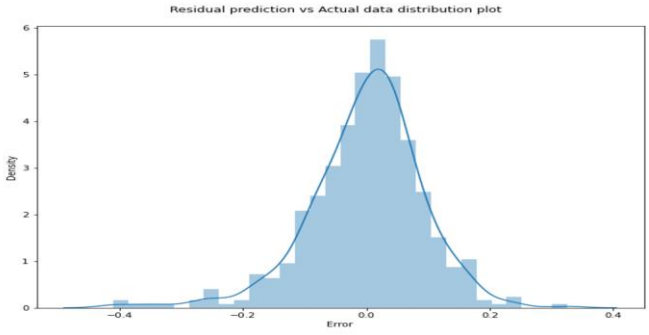
✓ **As From pair plot we can clearly say that temp and realfeeltemp (atemp) is highly coreelated, Almost linear with the target variable count**

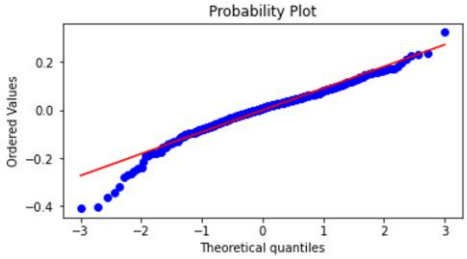
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer : - There are 5 Assumptions of the Linear Regression

1. **Normality : -** The residuals are normally distributed (Error term normally distributed). distplot (Distribution plot) clearly show that error terms are centered at 0 means its normally distributed.

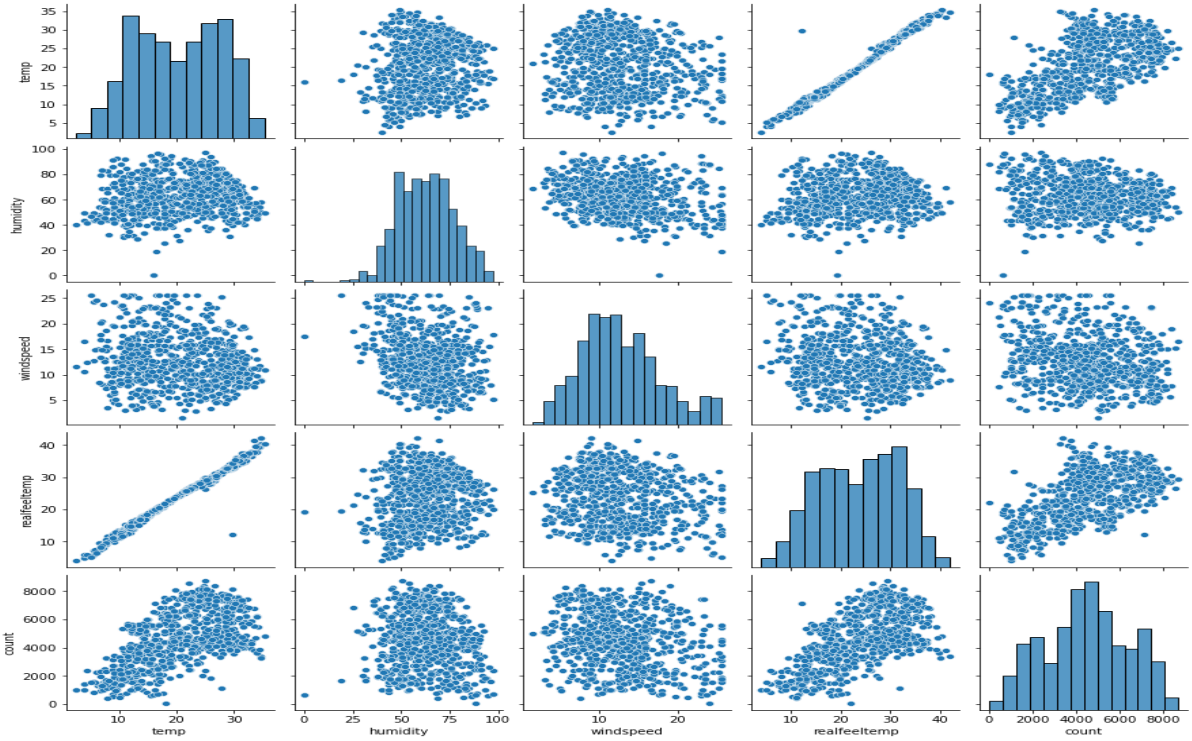
Plotted histogram and qq plot





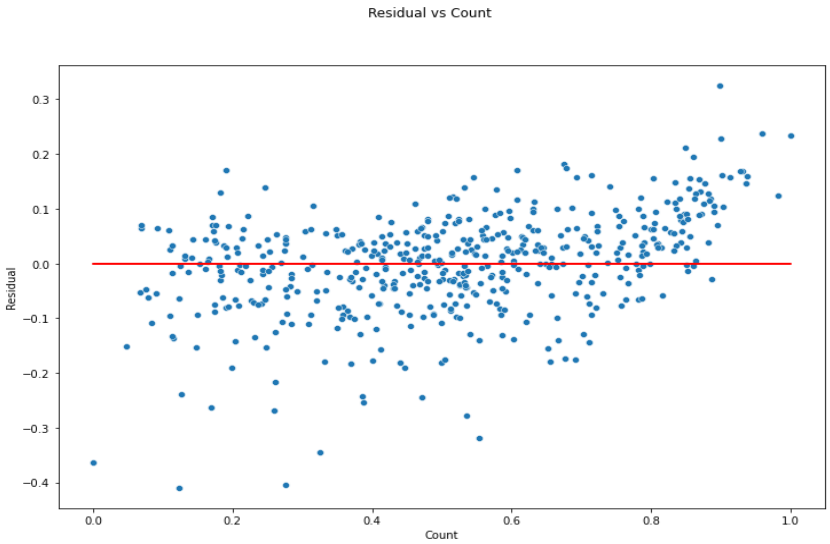
2. **Linearity : -** There is linear relationship between independent variable and dependent variable.

   Here assumption (Linear relationship) is validated through Scatter plot between dependent variables and independent variables.



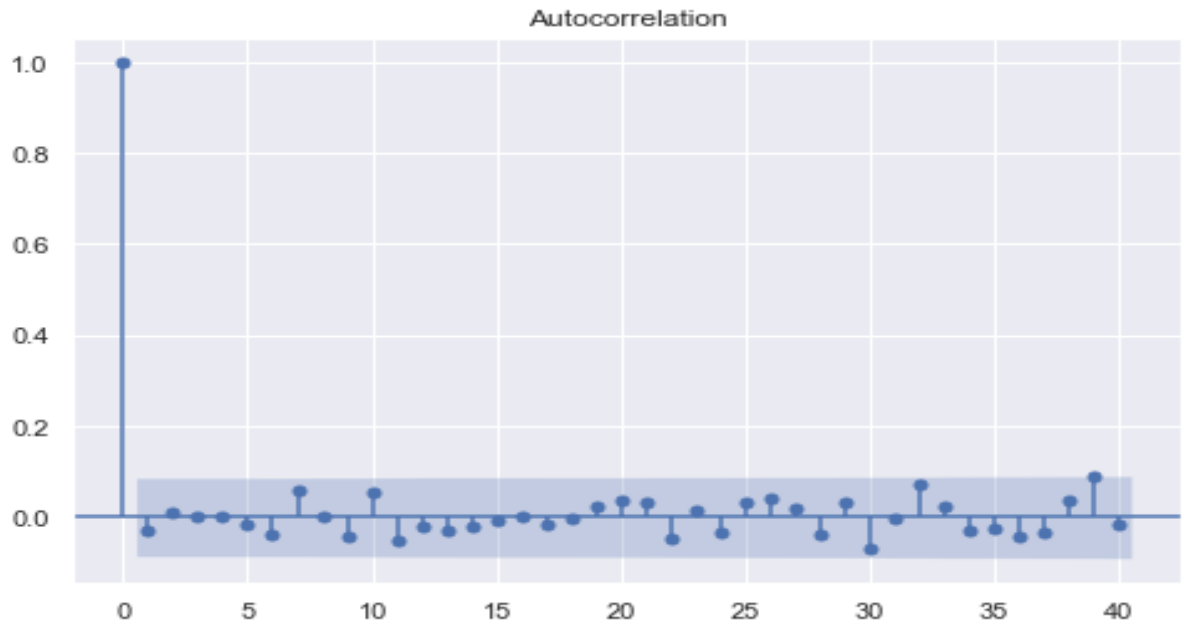
✓ As from pair blot we can say that there are some linear relationship between dependent variables and independent variables. In this case here clearly shows linear relationship between 'temp' , 'realfeeltemp' and windspeed with target variable count.

3. **Homoscedasticity : -** variance of the residual is constant across all of the independent variable x. means error is constant along independent variable.



Here from scatter plot we can observe that the there is constant deviation from the zero line hence our assumption of Homoscedasticity valid and true.

4. **Auto-Correlation Assumption**: - There is no auto correlation between error. Autocorrelation check if there is any pattern between error or not.



- ✓ there are no much error components crossing the confidence interval(greyed out area) and hence we can say that there is no pattern in the error.
- ✓ we can say that No auto-correlation presence. means error term of one observation is not influenced by the error term of another observation.
- ✓ Hence assumption of auto correlation has been preserved.

5. Multicollinearity : There is no correlation (low correlation) between independent variables. It can be verify using correlation coefficient or by variance inflation factor (VIF).

✓

	Features	VIF
7	temp	3.75
8	windspeed	3.14
6	year	2.00
4	summer	1.57
3	Mist + Cloudy	1.49
5	winter	1.38
0	Sep	1.20
1	Sun	1.16
2	Light Snow	1.08

- ✓ Here, Since VIF value of all the final predictor variables is less than 5, we can conclude that there are very less / low or no (zero correlation) zero multicollinearity between predictor variables.

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

Answer : - Top three features contributing significantly towards explaining the demand are:

1. **temp** (0.5525)
2. **Weather Situation** : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.2827)
3. **Year** (0.2330)

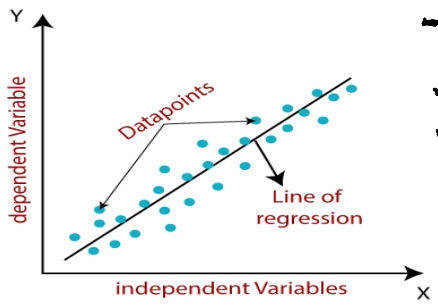
General Subjective Questions

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

Answer :- Linear regression is a simple statistical regression method used for predictive analysis of continuous variables and shows relationship between continuous variables by fitting best fit straight line through plot. Linear regression is a supervised machine learning algorithm. Its show relationship between independent variables (X-axis, features) and dependent variables (Y-axis, predictor variable).

If the input variable x will be single , then its called **Simple Linear Regression.**

If the input variable will be more than 1, then its called **Multiple Linear Regression.**



- plot present linear relationship between dependent variable on y-axis and independent variable on x-axis.
- When value of x increases value of y also increases.
- The line passing through most of the scatter point is referred as best fit straight line.

The line modelled on linear regression equation is

$$Y = a_0 + a_1 x + \varepsilon$$

There are two types of Linear Regression model :

1. **Simple Linear Regression**
2. **Multiple Linear Regression**

1. **Simple Linear Regression :** If single independent variable is use to predict the value of a dependent variable, such Linear regression algorithm is called Simple linear Regression.

It 's represent as

$$y = \beta_0 + \beta_1 X + \varepsilon$$

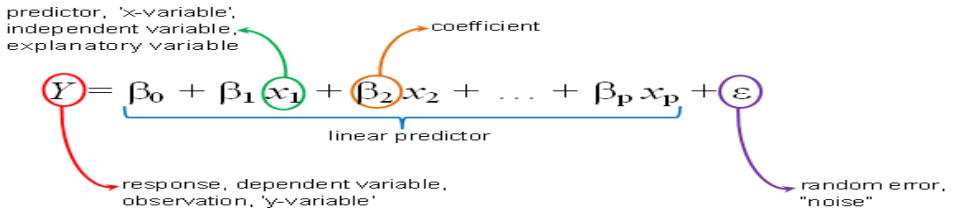
Y = dependent Variable or Target Variable. B_0 = Intercept of line.

X = independent variable or predictor variable. B_1 = linear regression coefficient.

ε = random error.

2. **Multiple Linear Regression :** If there will be more than one independent variable is used to predict the dependent variable, such Linear regression algorithm is called Multiple linear Regression.

   It's represent as



   Source:

   https://medium.datadriveninvestor.com/types-of-linear-regression-89f3bef3a0c7

   Assumptions: -

   - **Linearity**: Linear relationship between x and y
   - **Normality**: - Error terms are normally distributed.
   - **Homoscedasticity** - Error terms are independent of each other.
   - **No auto-correlation** - Error terms have constant Variance.
   - **Multicollinearity** – there shouldn't be multicollinear in data. It happens when indepe ndent variable highly correlated. It can be tested by VIF.

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

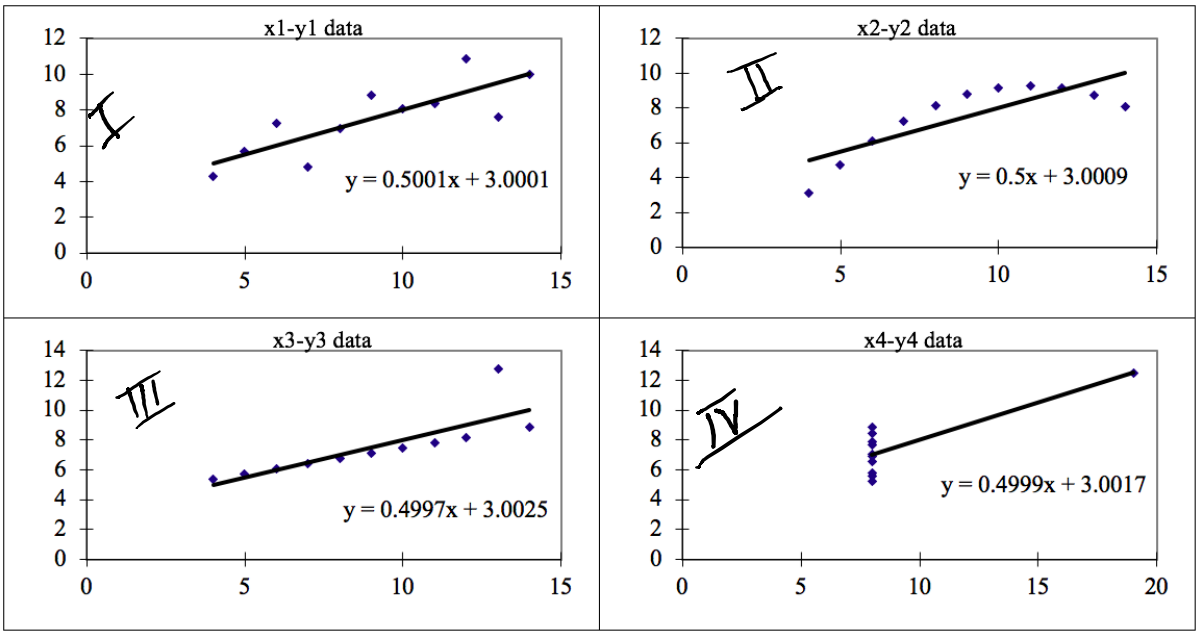
Answer : Anscombe's quartet comprises of a group of four datasets which mig ht have nearly identical in simple statistical properties. It emphasize both the importance of plotting before analysing and the effect of other influential observ ation on statistical properties.

There is 4 datasets and its statistics

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	

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

In the data set its fools linear regression model if model built.It have very different distributio n and it appear very different when plotted on scatter plot.



- ✓ 1st data set fits linear regression model as it seems to be linear relationship between X and y
- ✓ 2nd data set not show linear relationship between X and Y , means its not fit linear regression model.
- ✓ 3rd data set shows some outliers present in dataset which can't be handle by linear regression model.
- ✓ 4th data set has high leverage point means it's produce high correlation coff.

  - Its conclude that regression algorithm can be fooled so, its important to data visualisation before build machine learning model.

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

Answer : In Statistics pearson's correlation coefficient is also called as Pearson's r, PPMCC (pearson product moment correlation coefficient) is a statistics which measures the linear correlation between two variable. Like others correlations it has lies between -1.0 to +1.0 .

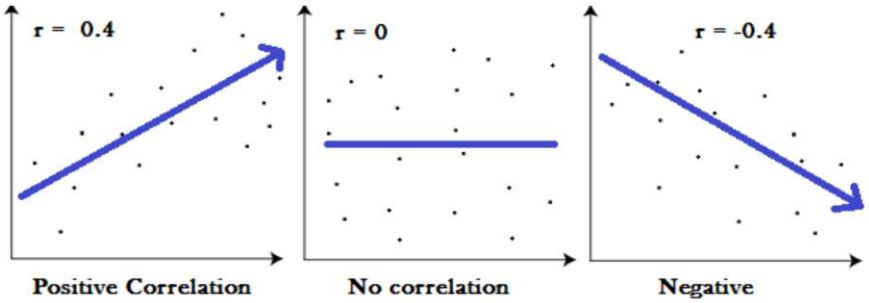
Requirements for pearson's correlation coefficient :

- ✓ Association should be linear.
- ✓ It should have no outliers in data.
- ✓ Variables should be normally distributed (approximately)
- ✓ Scale of measurement should be ratio or interval

$$r = \frac{\Sigma(x - m_x)(y - m_y)}{\sqrt{\Sigma(x - m_x)^2 \Sigma(y - m_y)^2}}$$

where,

- ***r:*** *pearson correlation coefficient*
- ***x and y:*** *two vectors of length n*
- m_x ***and*** m_y***:*** *corresponds to the means of x and y, respectively.*



Source: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

- ✓ r = 0 means no linear association
- ✓ < r < 5 means weak association
- ✓ 5 < r < 8 means moderate association
- ✓ r > 8 means strong association
- ✓ r = 1 means data is perfectly linear with positive slope
- ✓ r = -1 means data is perfectly linear with negative slope

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

Answer : scaling (feature scaling) is technique to normalize range of independent variables. It's generally performed at during of data pre-processing. Scaling of data help model to learn as well as understand the problem within giving range of features.

In dataset we have different features highly vary in magnitude and range, while algorithm considered just magnitude, may be constructed model would be incorrect when scaling is not performed. So, scaling is performed to bring all features to the same level of magnitude.

It's not affect any statisticians properties like p-value, R-square, F statistics, t statistics, Its only effect coefficient.

Generally Methods using for Scaling features are :-

- **Standardized scaling** : scaling technique where values centred around 0 with std. 1. In python we use StandardScaler() method for standardization.

$$x_{stand} = \frac{x - \text{mean}(x)}{\text{standard deviation}\ (x)}$$

- **Normalized scaling** : it also known as min max scaling, in this technique data is scaled in such way that all values lie between 0 and 1 . in python it's use as MinMaxScaler() . for normalization scaling.

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

Difference between normalized and standardized scaling :-

Normalized scaling	Standardized scaling
Used when features are of different scale	Used when we want to ensure 0 mean and unit std
Scales value between [0,1] or [-1,1]	It's not bounded to any range.
It's affected by outliers	Very less affected by outliers
It's also called as scaling normalization	It's called as z-score normalization
it useful when we have no idea about distribution	Useful when features distribution is normal/gausian
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
In python we use MinMaxScaler()	In python we use Standardscaler()

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

Answer : whenever there will be a perfect correlation then **VIF** will be infinite . large value of VIF indicate there will be correlation between variable.

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

From formulae if value of R-squared will 1 then VIF becomes infinity.

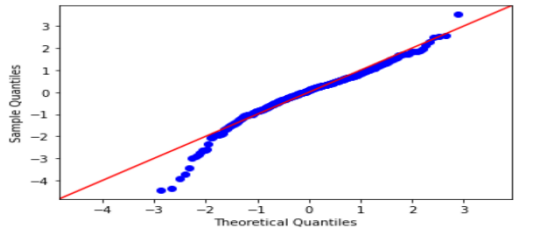
R-squared =1 the VIF = 1/(1-1) means VIF = 1/0 which implie Infinite.

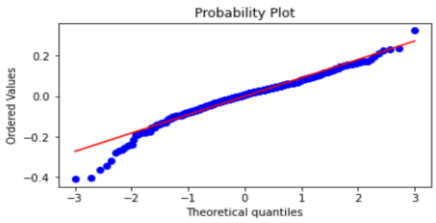
when VIF infinite there is perfect correlation. In this case we have to drop the variable to avoid multicollinearity. Reciprocal of VIF is called as Tolerance. Either of Both used to detect multicollinearity.

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

Answer : Q-Q plot (or Quantile-Quantile) are plot of two quantile each other. Its an graphical technique for determining if two data set come from population with common distribution.
45 degrees reference line is also plot if two set come from population from same distribution.
QQ plot or probability plot used to find type of distribution for random variable whether it's uniform, gaussian, Exponential. Simply we can say that it plots sample quantile against theoretical quantiles distribution. In QQ plot it has been created against by plotting two set of quantiles against each other.
In python used sm.qqplot(res, fit=True, line='45')

```
# plotting qq plot
sm.qqplot(res, fit=True, line='45')
```





Importance of QQ Plot in Linear Regression :

in Linear Regression when we have train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with same distribution or not.

Advantages:

- ✓ It can be used with sample size also
- ✓ *Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot*

Q-Q plot use on two datasets to check

- ✓ If both datasets came from population with common distribution
- ✓ If both datasets have common location and common scale
- ✓ If both datasets have similar type of distribution shape
- ✓ If both datasets have tail behaviour