Assignment-based subjective questions

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

Answer – Categorical variables from the datasets are – Season, year, month, holiday, weekday, workingday, and weather.

Inferences from these variables –

- Fall season has highest demand while spring season has least demand of bikes

- Demand increased from year 2018 to 2019

- Demand continuously increasing from Jan to June, highest in September. After that demand starts decreasing

- in holiday, demand decreases

- more demand during working day

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

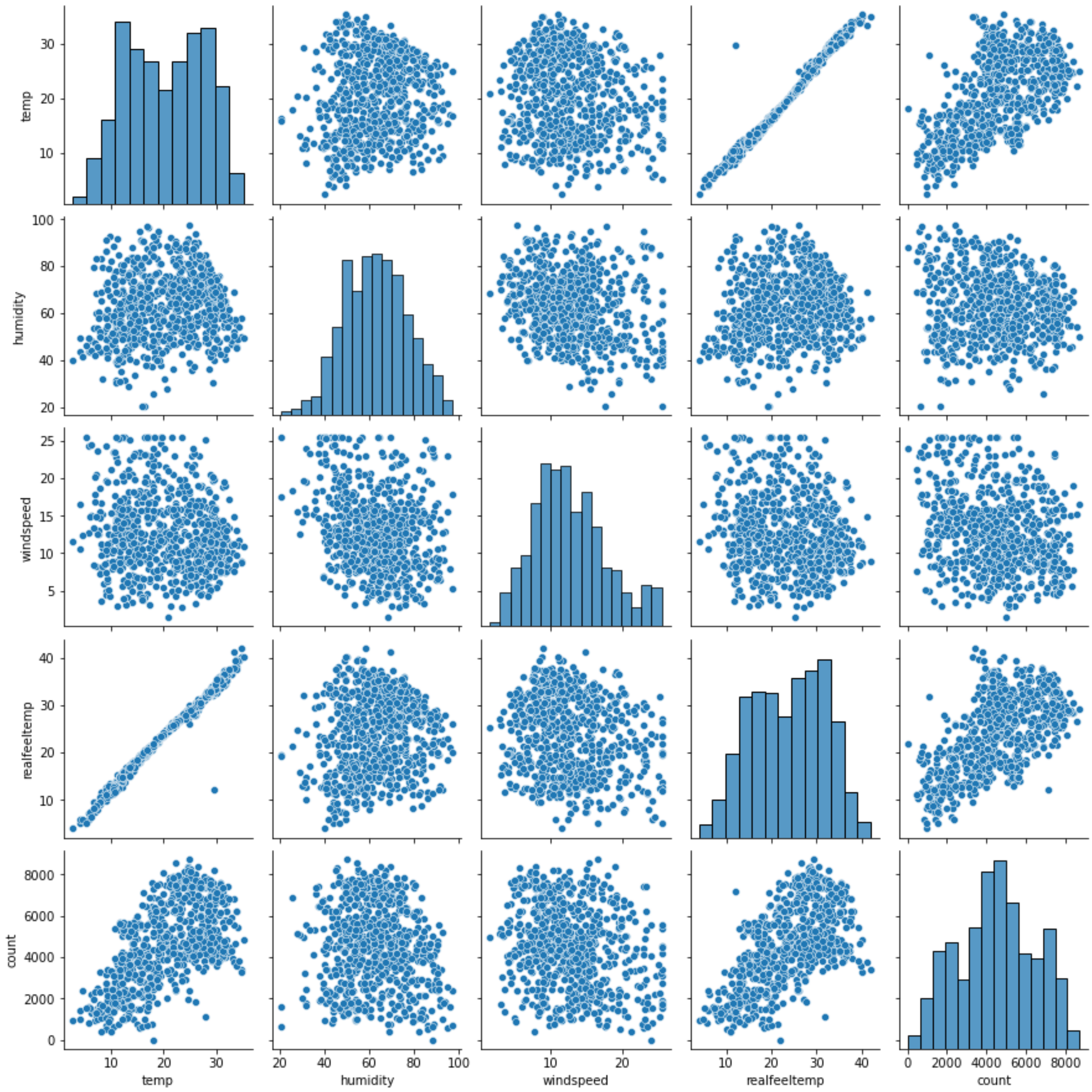
Answer – `drop_first=True` helps to reduce the extra created column during the dummy variable creation. So it will help in reducing correlation among dummy variable.

For e.g. if we have x level of categorical variable then we need to use (x-1) columns to represent the dummy variable.

In the given assignment, there are two possible values for holiday – yes and no, but there is no need of using 2 dummy variables. We just need to use one of them and other will be opposite of that.

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

Answer –

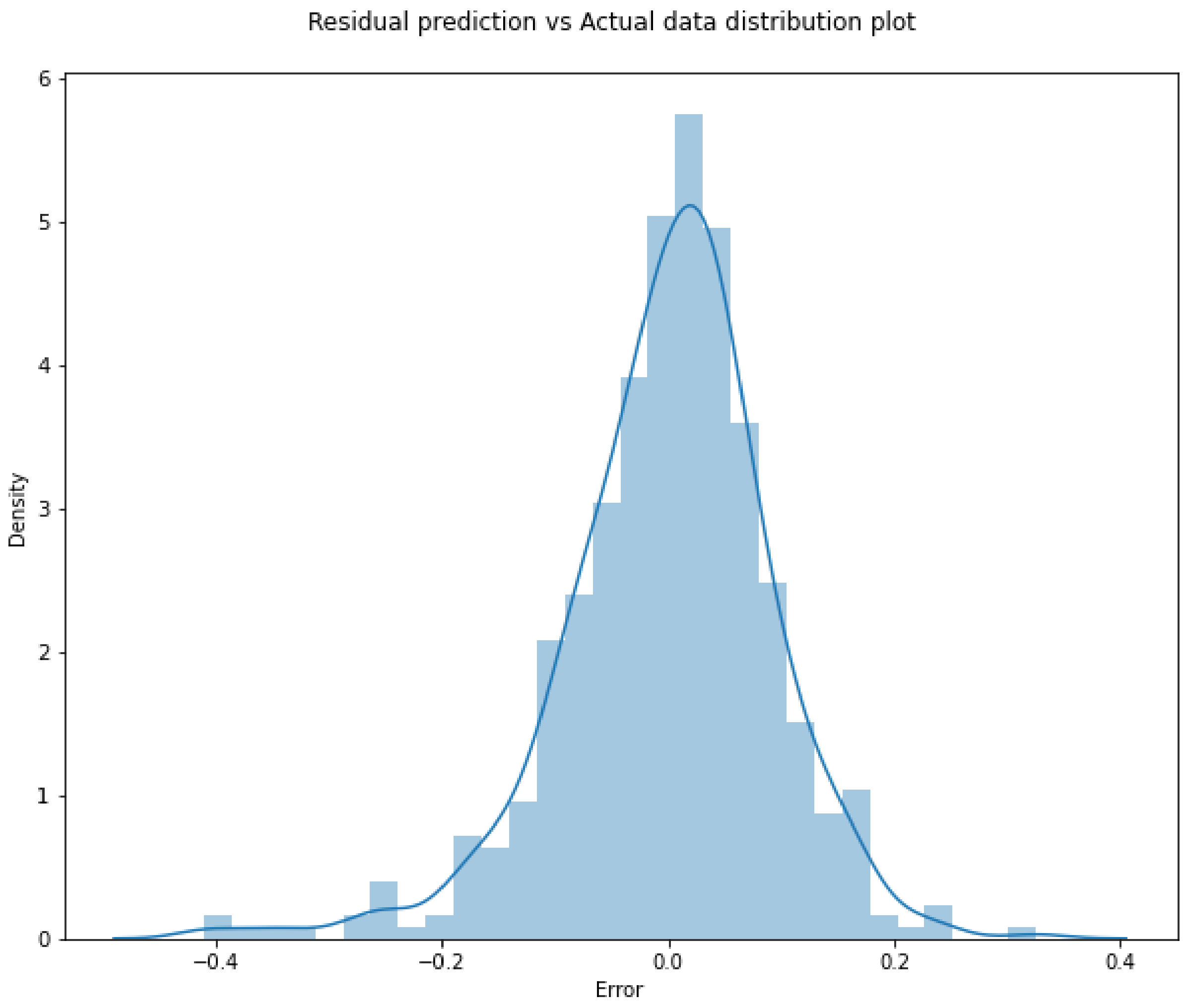


From the pair-plot it's clear that temp and atemp (realfeeltemp) are highly correlated. They are almost linear with the target variable count.

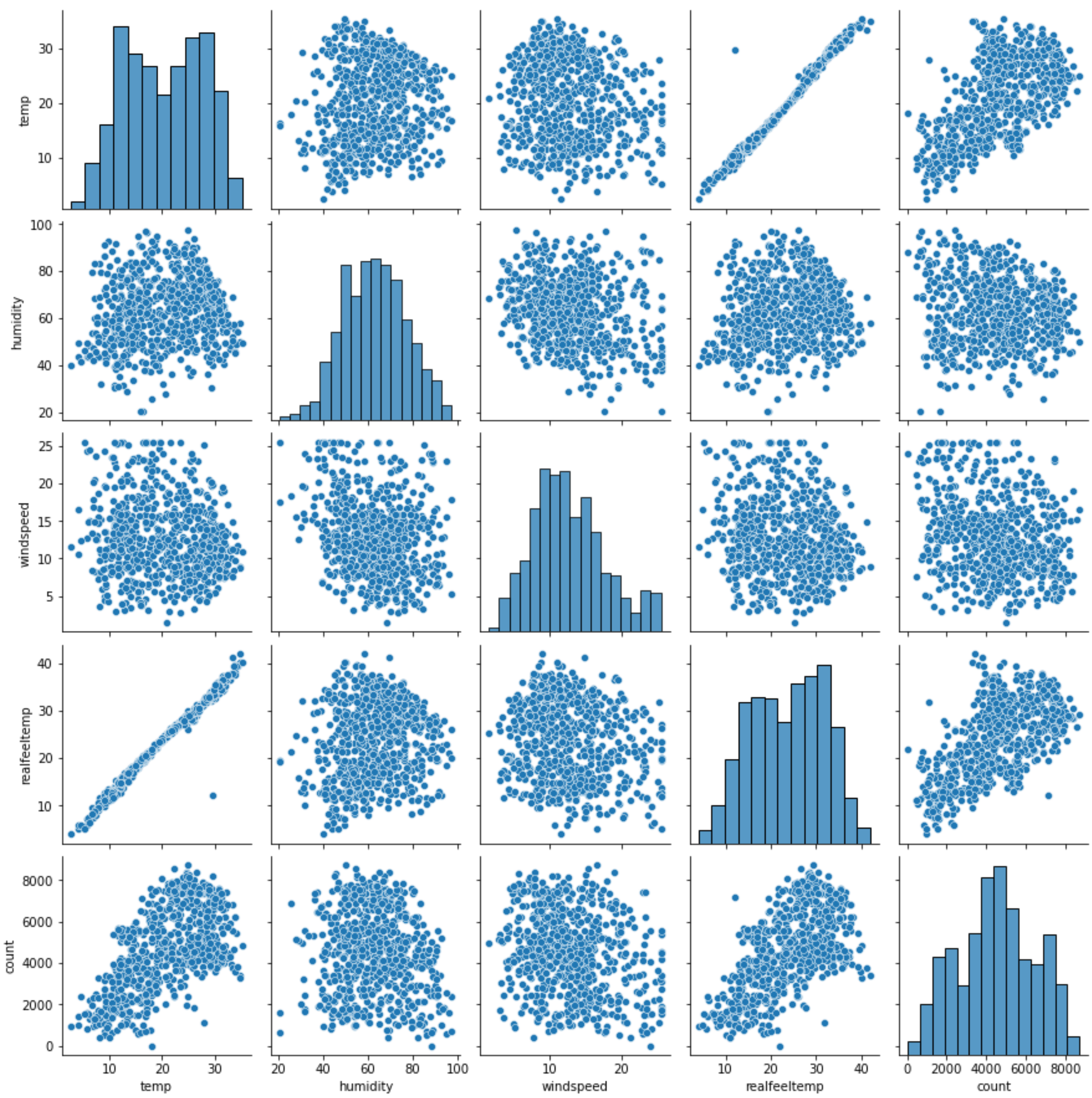
Q.4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer – There are 5 assumptions of linear regression

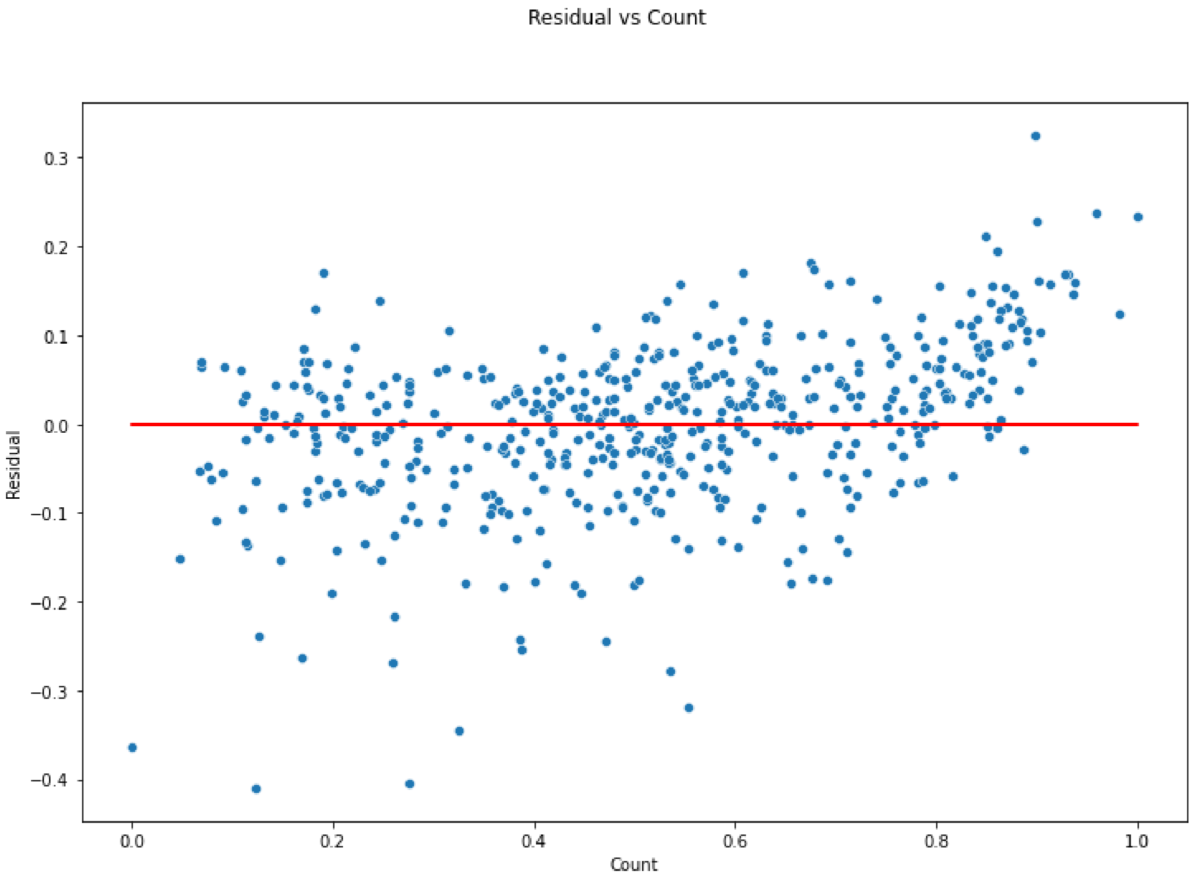
1. Normality – error term normally distributed



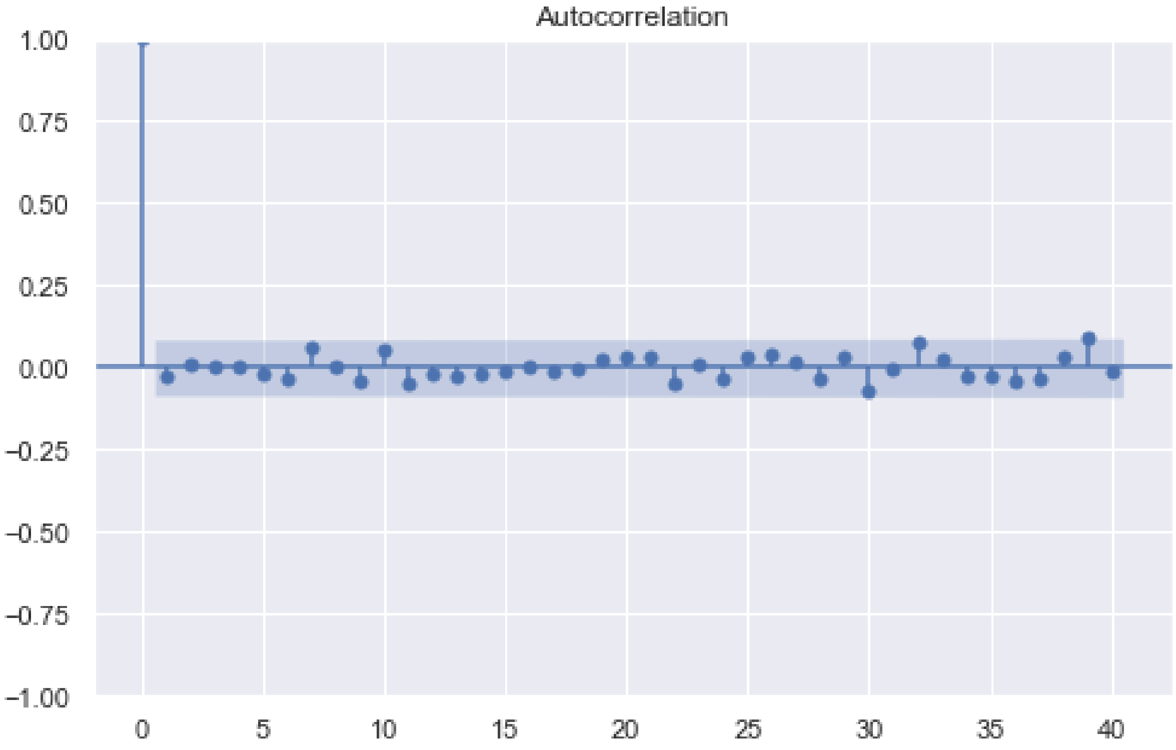
2. Linearity – there is linear relationship between temp, realfeeltemp (atemp) and windspeed with target variable – count



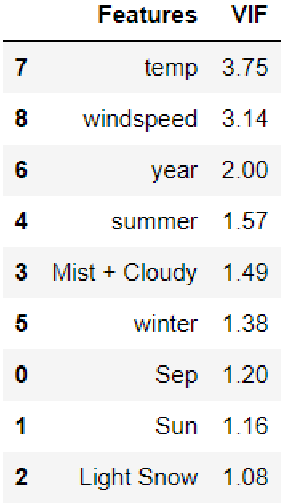
3. Homoscedasticity – error is constant along the independent variable



4. Auto-correlation – no much error components crossing the confidence interval, which means no pattern in error which means no auto-correlation



5. Multicollinearity – since VIF (variance inflation factor) value of all the final predictor variables is less than 5, which means that there is very low or zero multicollinearity between predictor variables.



	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

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

Answer – According to the dataset provided, top 3 contributing features are –

- Temp (0.5525)
- Weather situation – light snow (-0.2330)
- Year (0.2330)

General subjective questions

Q.1. Explain the linear regression algorithm in detail

Answer – Linear regression – it is a statistical regression method. It is used for predictive analysis of continuous variables. It shows relationship between the continuous variables by fitting best fit straight line through plot. It is a supervised machine learning algorithm.

It shows relationship between – X-axis features (independent variables) and Y-axis predictor variable (dependent variable). There are two types of linear regression model –

1. Simple linear regression – single independent variable
2. Multiple linear regression - >1 independent variable

It can be represented by linear equation –

$y = \beta_0 + \beta_1 x$

where, y = dependent variable

β_0 = Intercept

β_1= Slope

x = independent variable

Assumptions in Linear Regression

Linearity – linear relationship between x & y

Normality – Error terms are normally distributed

Homoscedasticity – error terms are independent of each other

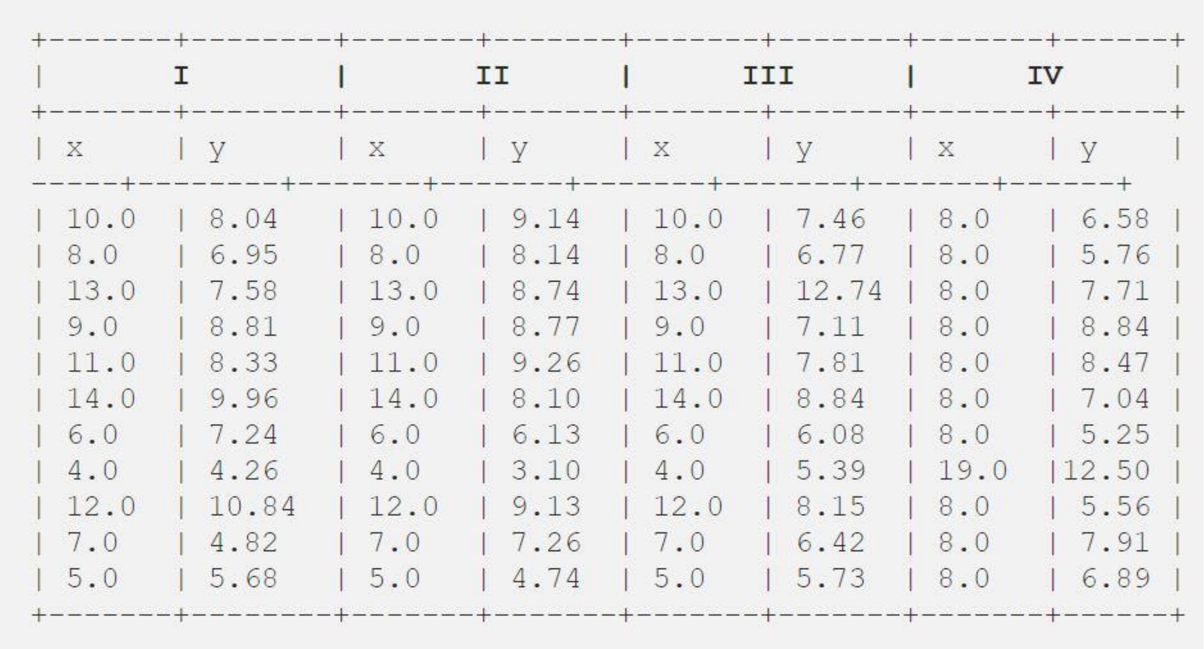
No auto-correlation – error terms have constant variance

Multicollinearity – there should not be multicollinearity in the data. It's tested by VIF

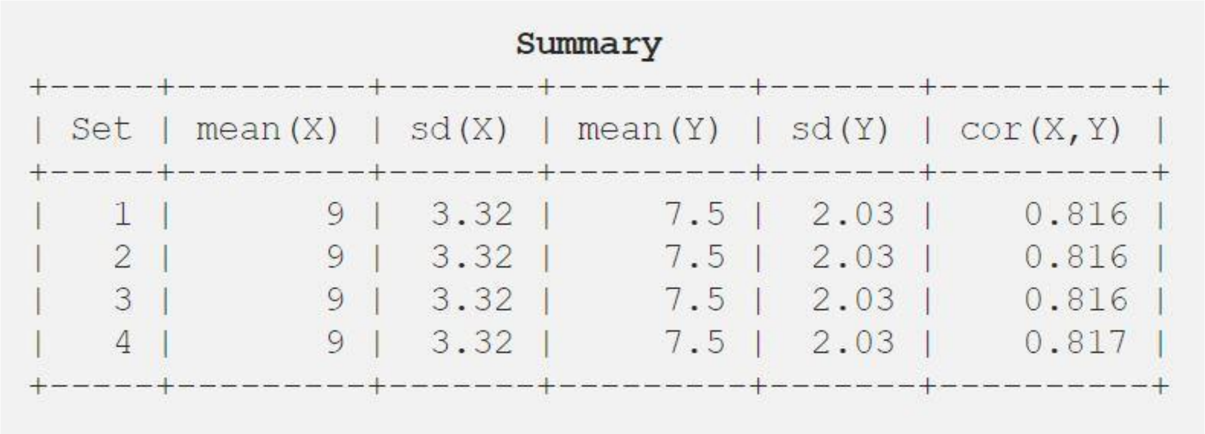
Q.2. Explain the Anscombe's quartet in detail.

Answer – Anscombe's quartet – It comprises of a group of four datasets which might have nearly identical simple statistical properties, but they still appear very different when put on graph. It emphasise both the importance of plotting before analysis and the effect of other influential observation on the statistical properties. These were constructed by statistician Anscombe.

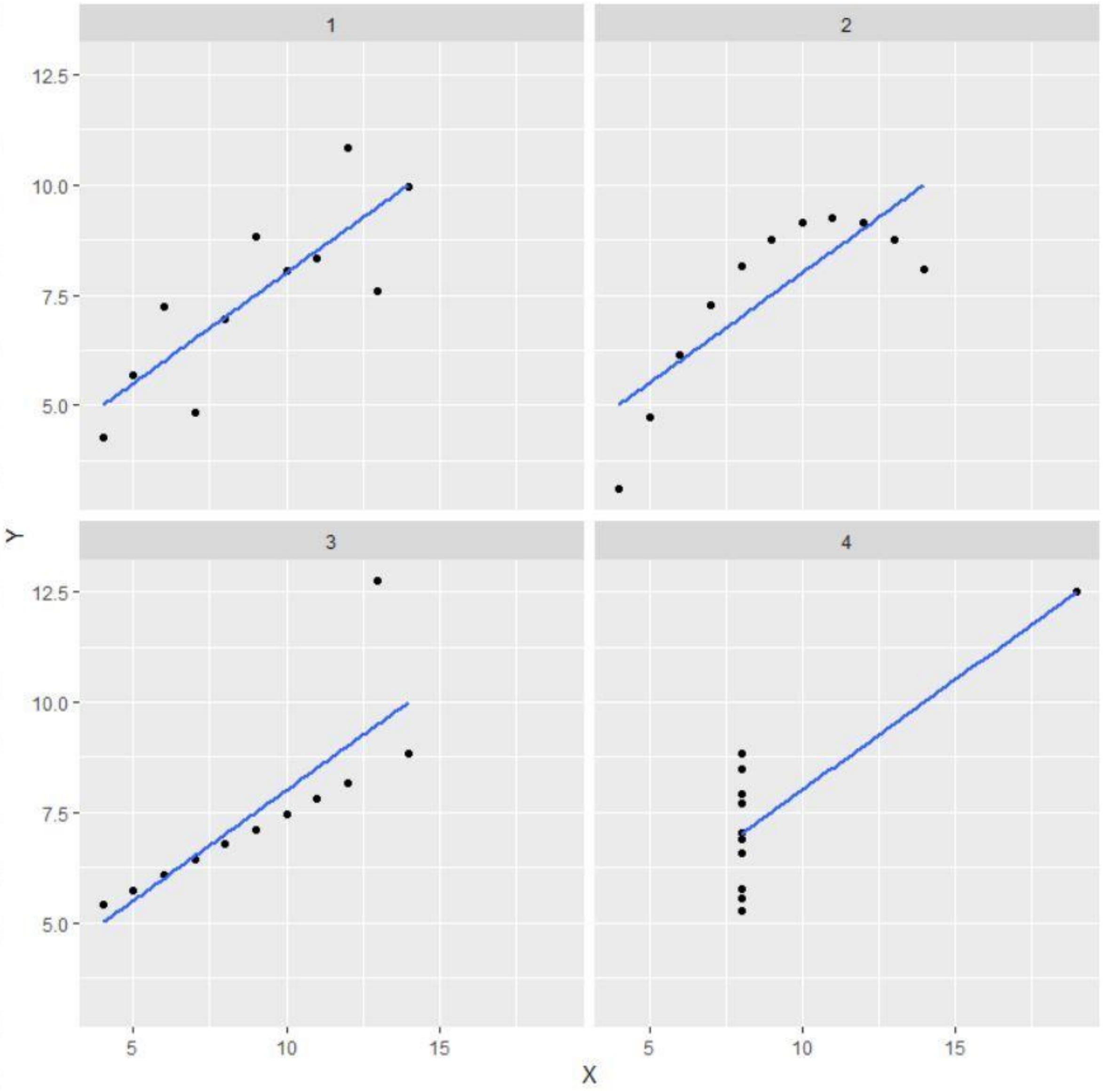
e.g. 4 sets of 11 data-points as shown below (Source: https://www.geeksforgeeks.org/anscombes-quartet/)



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



```
                          Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```



Explanation of the output –

- First one (top left) – linear relationship between x & y

- 2nd one (top right) – non-linear relationship between x & y
- 3rd one (bottom left) – perfect linear relationship except one outlier
- 4th one (bottom right) – one high-leverage point enough to produce a high correlation coefficient

Q.3. What is Pearson's R?

Answer – Pearson's R -In statistics, pearson's correlation coefficient is also called as Pearson's R, PPMCC (pearson product moment correlation coefficient) measures linear correlation between two variables. It lies between -1.0 to +1.0

Requirements for pearsons correlation coefficient

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

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx^2 = the sum of squared x scores

Σy^2 = the sum of squared y scores

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

Answer – Scaling – one of the most important data pre-processing steps in machine learning. It is a technique to normalize the range of independent variables. In dataset we have different features highly vary in magnitude and range, while algorithms just consider magnitude. So, if the scaling is not performed, the constructed model would be incorrect.

Therefore, scaling is performed to bring all the features to the same level of magnitude. It doesn't affect statistical properties like p-value, R-square, F-statistics etc. It affects coefficient.

Generally scaling methods used are –

- Standardized scaling – scaling technique where values cantered around 0 with std. 1. In python we use StandardScaler()
- Normalized scaling – also known as min max scaling. In this, data is scaled in such a way that all values lies between 0 and 1. In python we use MinMaxScaler()

Difference between Normalization and Standardization

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	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.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

Source - https://www.geeksforgeeks.org/normalization-vs-standardization/

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

Answer – Whenever there will be a perfect correlation then VIF will be infinite. Large value of VIF indicate there will be correlation between variables.

VIF = 1 / (1 – R^2) = 1 / Tolerance

From the formula, if the value of R squared is 1 then VIF becomes infinite.

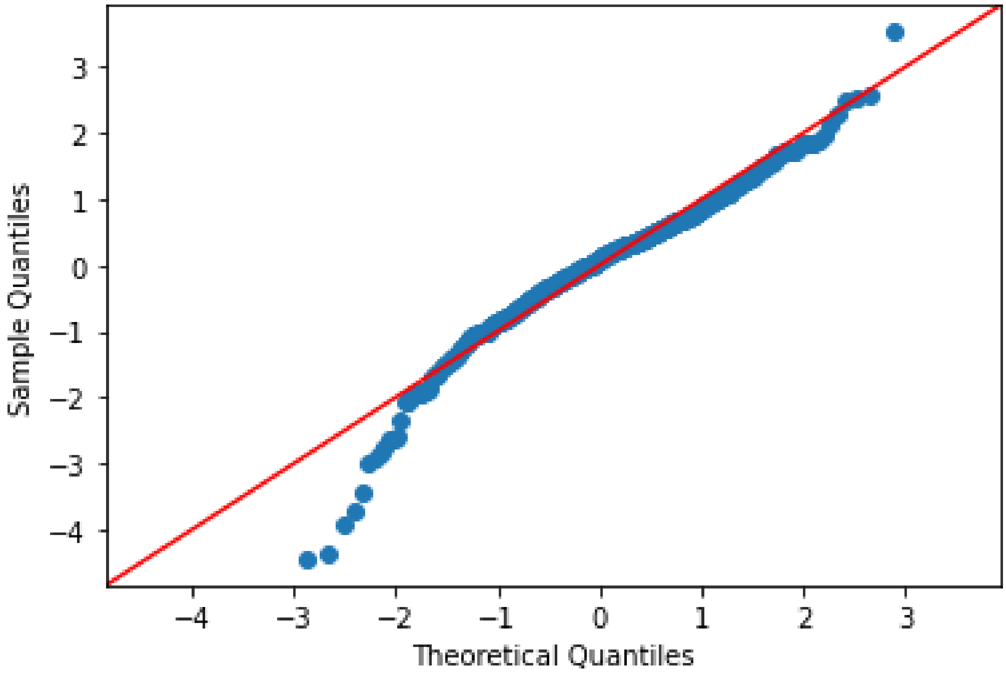
If VIF is infinite, then there is perfect correlation. In this case, we should drop the variable to avoid multicollinearity.

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

Answer – QQ Plot – also known as Quantile – quantile plot. It's plot of two quantile each other. It is graphical technique for determining if the two datasets come from population with common distribution. 45 degrees reference line is also plot if two set come from population from same distribution. It is used to find type of distribution for random variable whether it's uniform, gaussian, exponential etc.

In python, if statsmodel.api is imported as sm, then the command for qq plot would be -

sm.qqplot(res, fit=True, line='45')



Importance of QQ Plot in Linear Regression

- We can confirm that both train and test data set are from the population with same distribution or not
- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry and presence of outliers can be detected from this plot.