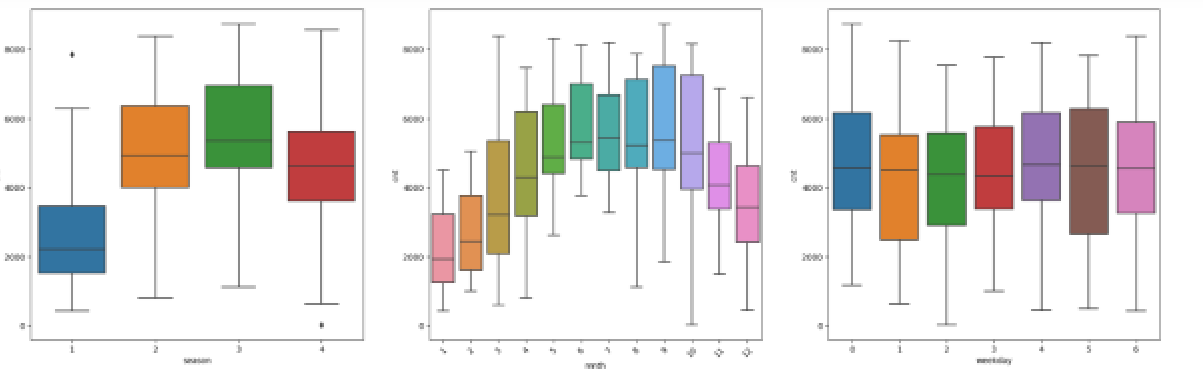
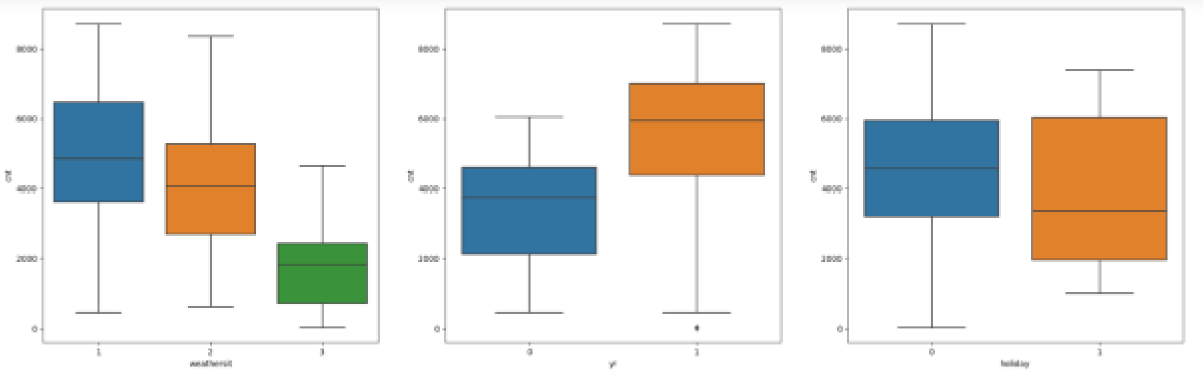
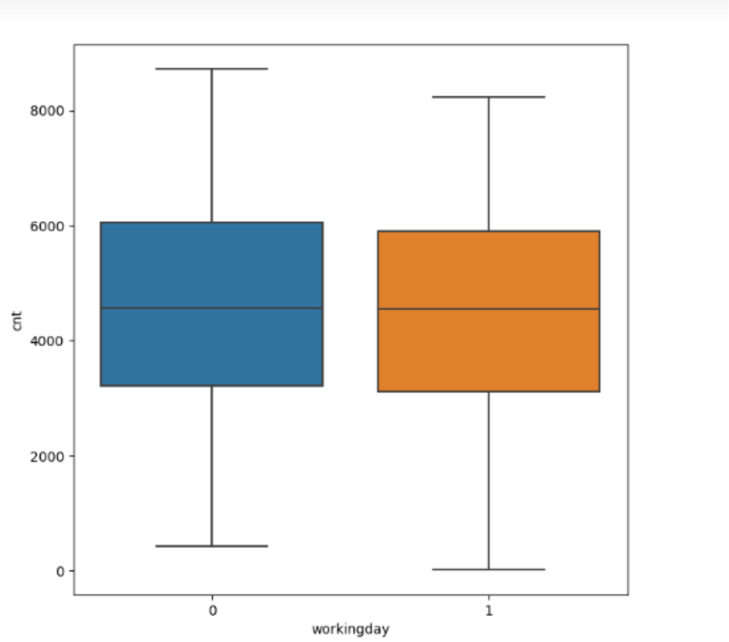
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1. From categorical variables we infer that-







a) People are more likely to take bike on rent in the season of Summer and Fall.
b) Bike rental rates are more in month of September and October.
c) More bikes were rented on Thursday, Friday, Saturday and Sunday.
d) Count of rented bikes increases with clear weather. (Clear, Few clouds, partly cloud.)
e) More bikes were rented in 2019.
f) Rental rates are higher during holidays.

Q2. Why is it important to use drop_first=True during dummy variable creation?

A2. The pandas "**get dummies"** function creates a dummy variable from pandas in python. The syntax used is as follows-

Import pandas as pd

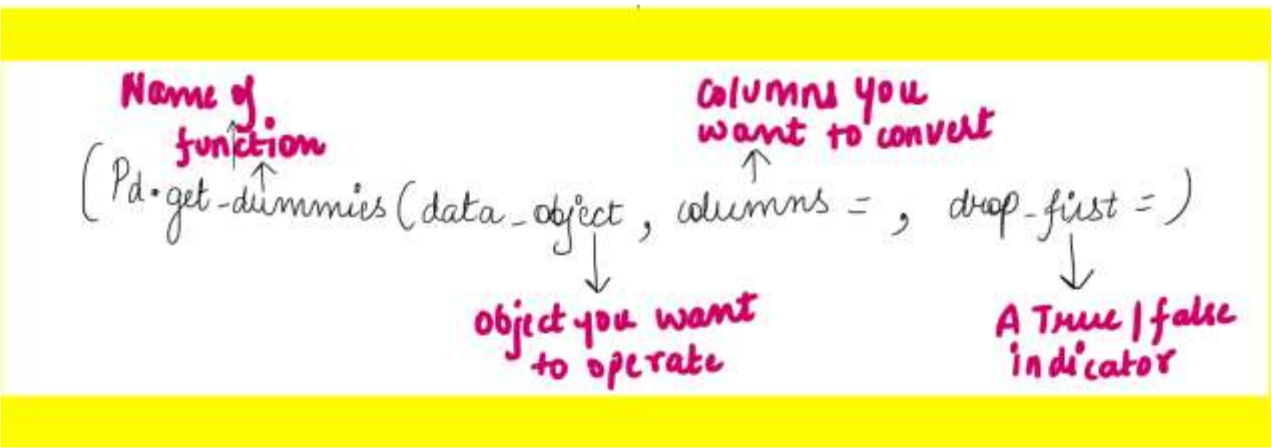
pd.get_dummies (df. column_name, drop_first = True)

df=dataframe

Dummy variable- is a numeric variable that encodes categorical information.

Drop_first = specifies whether you want to drop the first category of the categorical variable. By default, it is set to False which will cause get_dummies to create one dummy variable for every level of the input categorical variable.

If you set drop_first =True, then it will drop the first category. So if you have n categories, it will give you n-1 dummy variables.


Name of function
Columns you want to convert
(Pd.get-dummies(data_object, columns = , drop-first =)
object you want to operate
A True|false indicator

It is important to drop_first dummy variable because

a) By dropping it avoids multicollinearity.
b) Interpretability – when we drop one variable out of k levels, it serves as a reference category and the coefficients of other remaining dummy variables represents how much category differs from the reference category. Thus, making interpretation more meaningful.
c) It reduces model complexity.
d) It prevents redundancy in the model.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

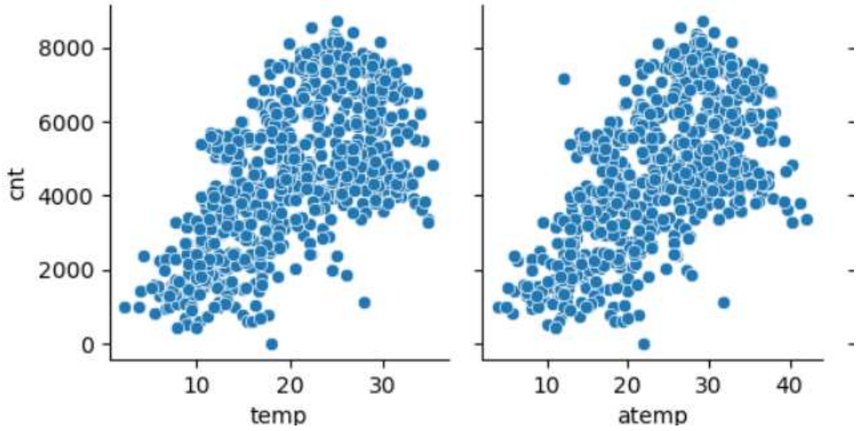
A3. While looking at the pair-plot among the numerical variables temperature (temp) has highest correlation with target variable that is cnt. (Coefficient value 0.5174)

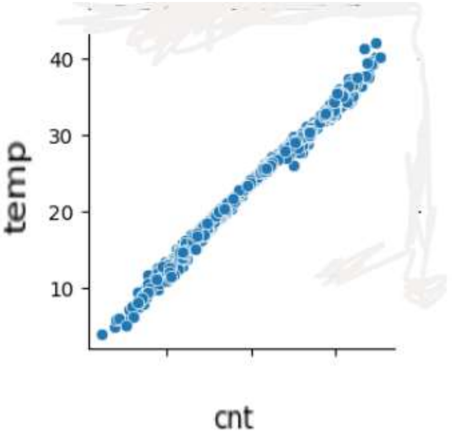
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A4. Assumption -

a) **Linearity** – There should be linear relationship between dependent(Y) and independent variable(X).

This assumption is validated by plotting a scatter plot





We can clearly see there is a linear relationship between dependent and independent variable.

b) **Independence** - The observations in a data set should be independent of each other. Error terms are independent of each other.

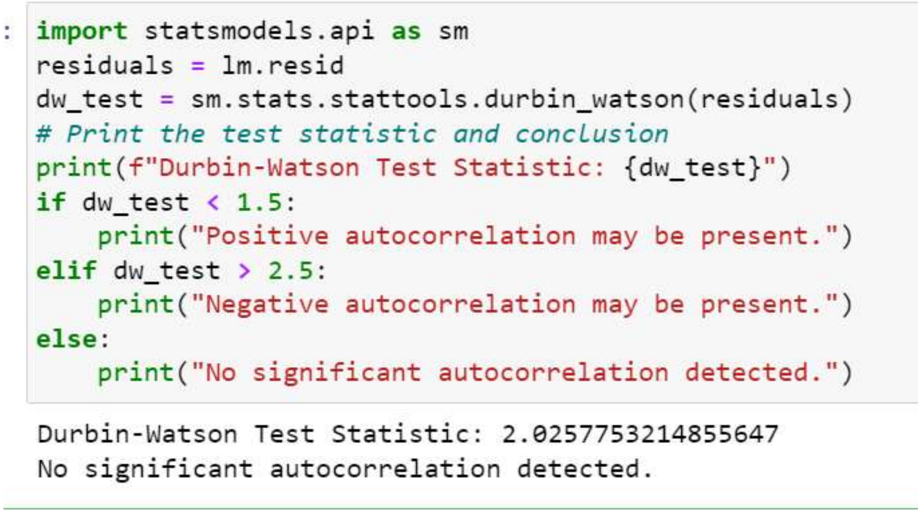
Test for presence of autocorrelation in residual (errors) of a regression model.

For this durbin_watson(dw) test has been performed.

Autocorrelation occurs when there is a pattern or correlation between the error terms.

1. In this test if DW value is close to 2 it indicates errors are independent of each other.
2. If DW value is significantly less than 2 (around 1.5) it signifies positive correlation between errors.

3. If DW value is significantly more than 2 (around 2.5) it signifies negative correlation between errors.



```
: import statsmodels.api as sm
  residuals = lm.resid
  dw_test = sm.stats.stattools.durbin_watson(residuals)
  # Print the test statistic and conclusion
  print(f"Durbin-Watson Test Statistic: {dw_test}")
  if dw_test < 1.5:
      print("Positive autocorrelation may be present.")
  elif dw_test > 2.5:
      print("Negative autocorrelation may be present.")
  else:
      print("No significant autocorrelation detected.")
```

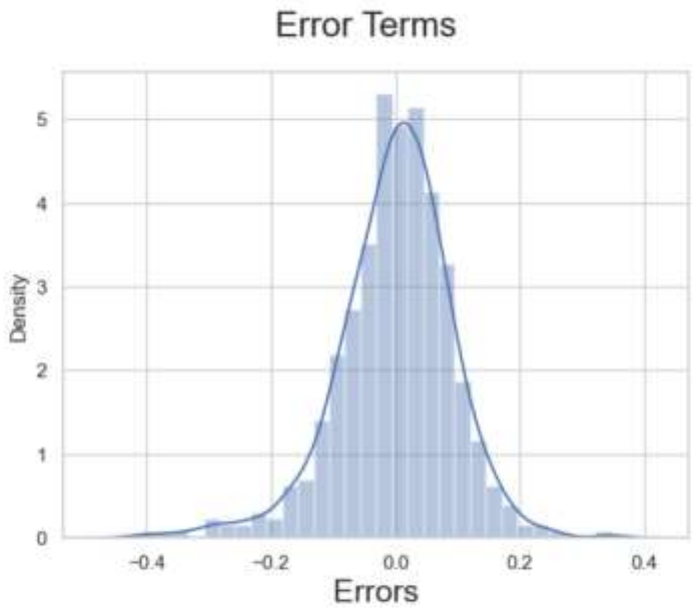
```
Durbin-Watson Test Statistic: 2.0257753214855647
No significant autocorrelation detected.
```

Here we can see the value is coming out to be 2.025 which shows no correlation in error terms.

c) **Normalisation** – Error terms are normally distributed between X & Y.

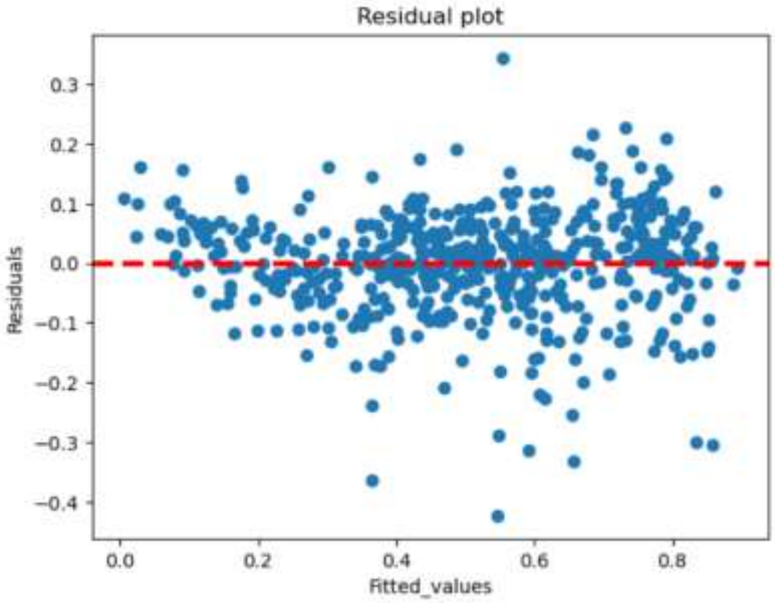
For this we have plotted a graph

(y_train) – (y_train_pred)   where : y_train_pred = y_train_cnt



We can clearly see error terms are distributed normally.

d) **Homoscedasticity** – Error terms should have constant variance.



We can clearly see there is no pattern which shows variance is constant for error terms and since there is no pattern so error terms are also independent.

e. Multicollinearity is checked by VIF. If value is more than 5 it should be dropped.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

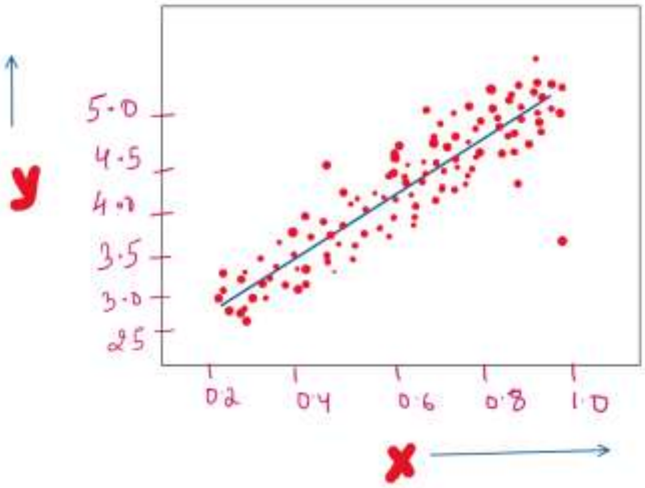
A5. Based on the final model the top 3 features contributing significantly towards explaining the demand of the shared bikes are

a. Temp
b. Yr. (year)
c. w3_lightsnow

General Subjective Questions

Q1. Explain the linear regression (LR) algorithm in detail.

A1. Linear regression uses ML (Machine learning) algorithm in which machine learns from the given data. ML consists of supervised and unsupervised learning and Linear regression falls under supervised category. As the name suggests Linear regression finds a best fit linear relationship between independent and dependent variable. Supervised learning is the one in which target variable is given. In case of linear regression target column is numerical or continuous.



Here Y-axis contains a dependent variable and X-axis contains independent variable. Regression is of two types-

a) **SLR – Simple linear regression**
b) **MLR- Multiple linear regression**

SLR – It explains the relationship between a dependent and only one independent variable. Here a straight line is plotted using a scatter plot.

Formula : $Y = \beta_o + \beta_1X + \mathcal{E}$

Where Y_i = dependent variable , β_o = Intercept ,

β_1 = Slope , X = independent variable , $\mathcal{E}$ = Error

MLR- It explains the relationship between dependent variable and more than one independent variable. It's a hyperplane instead of straight line.

Formula : $Y= \beta_o + \beta_1X_1+ \beta_2X_2 + \beta_3X_3 +$ ………………..$+ \beta_iX_i + \varepsilon$

LR is a very powerful tool to predict the behaviour of variables.

Hypothesis testing in LR –

$Y = \beta_o + \beta_1X$

Here we start by saying β_1 is not significant that is it has no relationship between X and Y.

So Null Hypothesis (H_0) : $\beta_1 = 0$

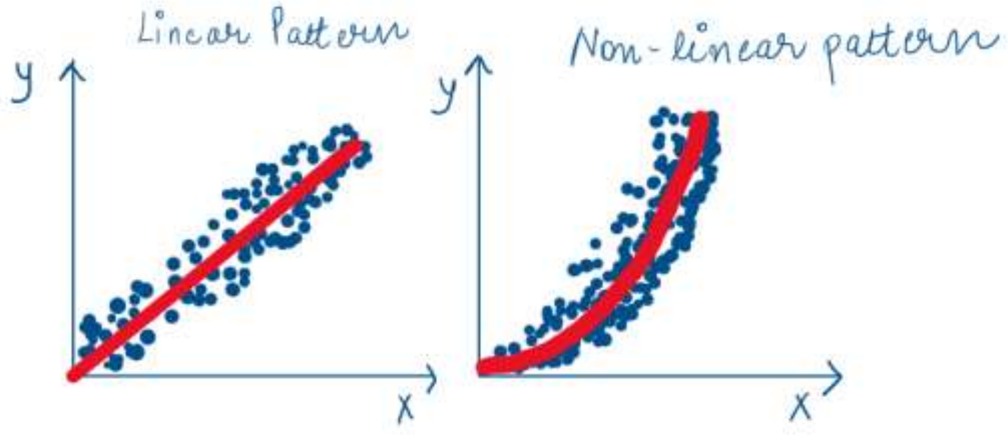
Alternate Hypothesis (H_1) : $\beta_1 != 0$

Hence if we fail to reject null hypothesis, β_1 is insignificant and no use of model.

And if you reject Null hypothesis it would mean that β_1 is not zero and the line fitted is a significant one.

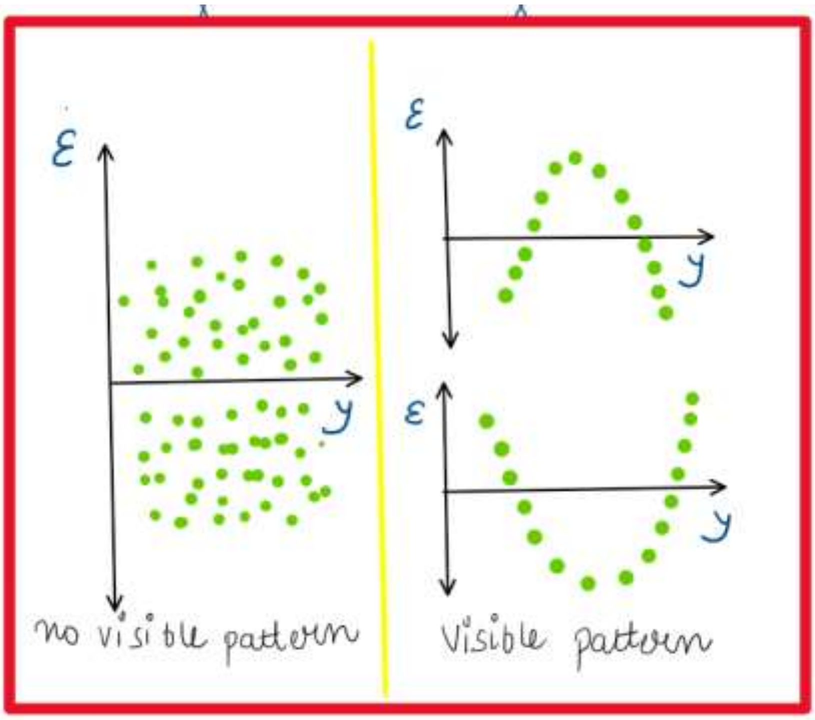
Assumptions of LR-

a) **Linearity** – There should be linear relationship between dependent(Y) and independent variable(X).

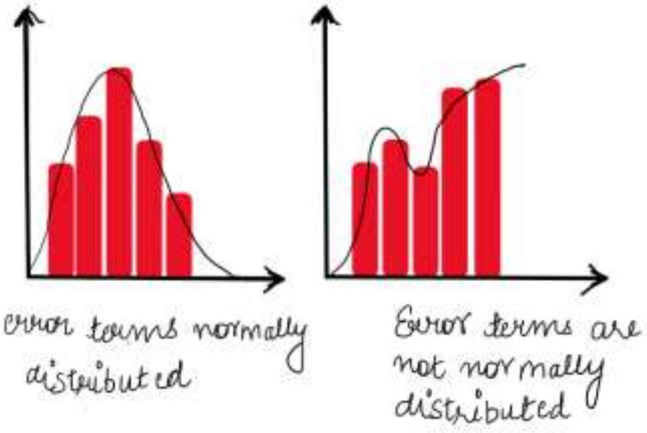


b) Independence - The observations in a data set should be independent of each other.

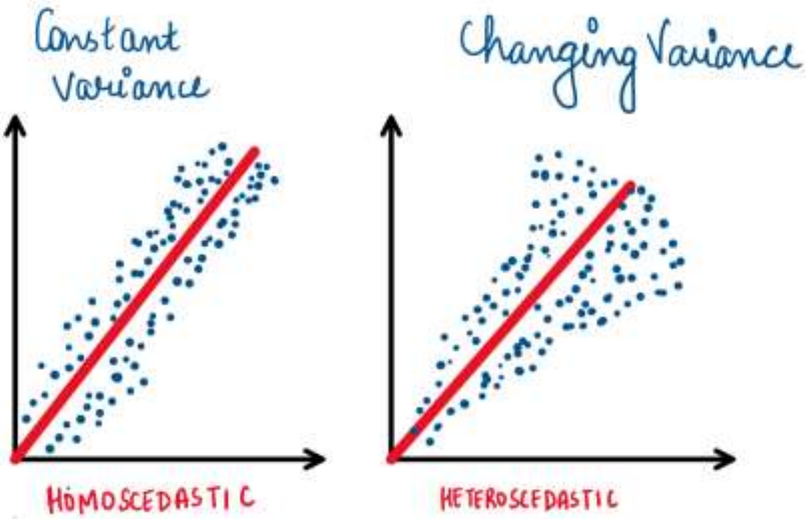
Error terms are independent of each other.



c) Normalisation – Error terms are normally distributed between X & Y.

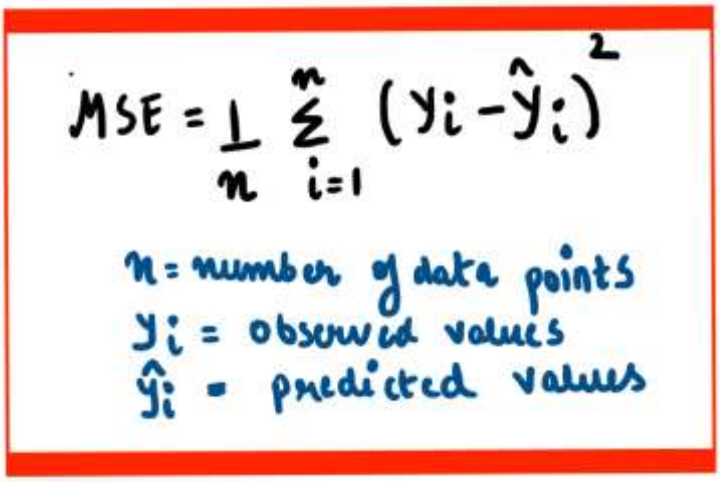

error terms normally distributed
Error terms are not normally distributed

d) Homoscedasticity – Error terms should have constant variance.



e) No multicollinearity – There is no high correlation between the independent variables.

Finding best fit line- To find out the best fit line the values of β_o , β_1 needs to be calculated.
It is done by using cost function.
COST FUNCTION – It optimises the regression coefficients. It is a mathematical function which calculate the error. We use MSE (Mean squared error)



$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

n = number of data points
y_i = observed values
$\hat{y}_i$ = predicted values

The difference between the observed value and predicted value is known as **Residuals**.

If observed points are far from regression line then residual will be high and hence high cost function and if observed points are near regression line then residual will be small and hence less cost function.

Gradient Descent

Gradient descent method is used to minimize MSE by calculating gradient of cost function. A regression model uses Gradient descent to update the coefficients of line by reducing the cost function. It is an iterative method which is done by random selection of a values of variables and then reaching to minimum cost function.

Model Performance

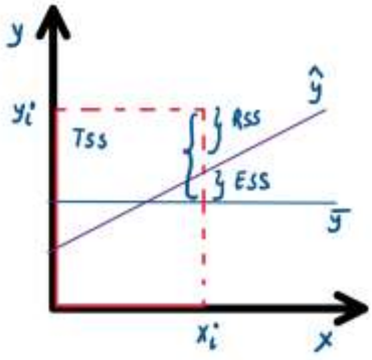
The process of finding best fit model out of various models is known as optimization. It can be done by R-squared method.

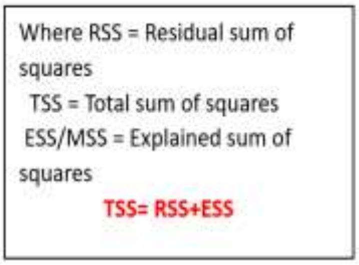
R-squared method –

This method mesures the strength of the relationship between the dependent and independent variables on a scale of 0-100% or 0 to 1.

Formula - R-squared = Explained Variation / Total Variation

R^2 = 1- RSS/TSS



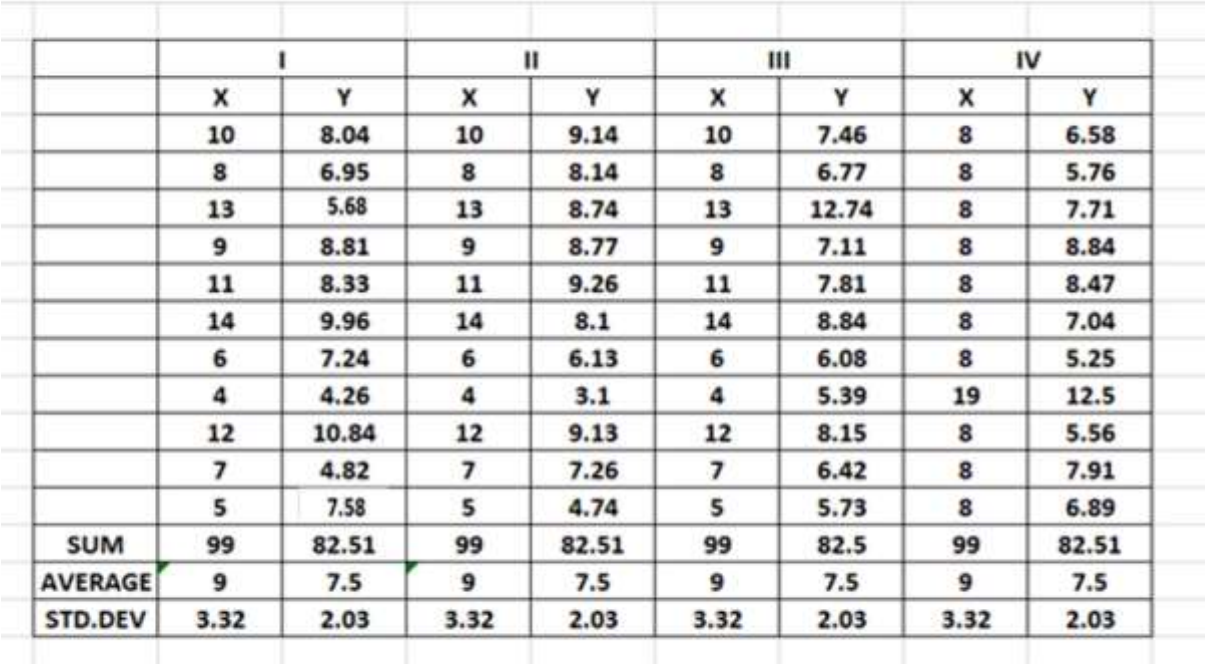

Where RSS = Residual sum of squares
TSS = Total sum of squares
ESS/MSS = Explained sum of squares
TSS= RSS+ESS

Q2. Explain the Anscombe’s quartet in detail.

A2. Anscombe’s quartet consists of four dataset having identical statistical properties in terms of variance, mean, R-squared, linear regression lines, correlations but becomes different when we plot those data points using scatter plot. It is used in data visualisation to spot trends, outliers and other crucial details that might not be visible by just looking summary of statistics alone.

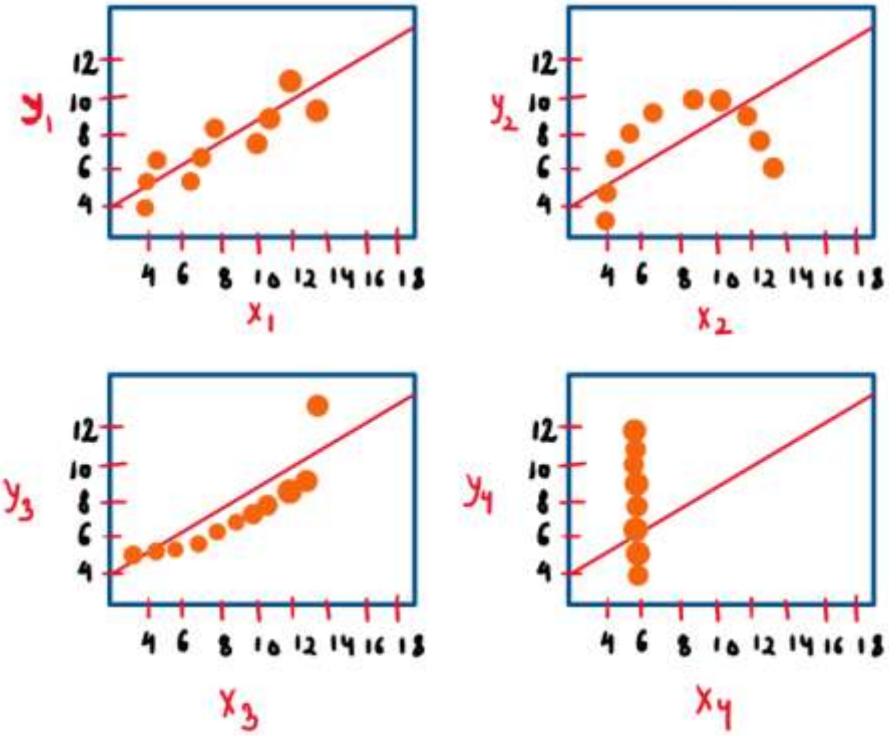
It was constructed in 1973 by statistician Francis Anscombe to describe the importance of plotting graphs before analysing the model building.

Suppose we have a data



	I		II		III		IV	
	X	Y	X	Y	X	Y	X	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	5.68	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	7.58	5	4.74	5	5.73	8	6.89
SUM	99	82.51	99	82.51	99	82.5	99	82.51
AVERAGE	9	7.5	9	7.5	9	7.5	9	7.5
STD.DEV	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Here what we can see that sum , average and standard deviation is almost same. So by just watching the data and there statistical properties we can infer that the graphs(scatter plot) would be same.But when we plotted these data set then what we can observe that



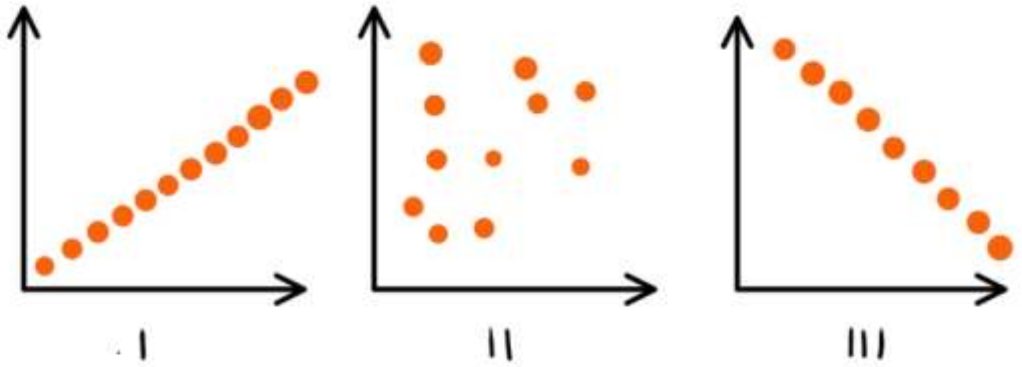
1. Scatter plot 1 - is clean and well fitted linear models.
2. Scatter plot -2 – is not distributed properly.
3. Scatter plot- 3 – is linear but the calculated regression is not lying in straight line because of an outlier.
4. Scatter plot - 4 – Shows that one outlier is enough to produce a high correlation coffiecient.

Thus we can say that those details which we were not able to see by statistical data was visible by plotting them. So both statistical and data visualisation method are important in order to jump to a solution.

Q3. What is Pearson's R?

A3. To check the degree of association correlation coefficient is used. It is denoted by **'r'** or sometime **'R'.** It is also known as Pearson 's correlation coefficient. The correlation coefficient is measured on a scale of -1 to 1.

When one variable increases with other variable it shows positive correlation and if one variable decreases with another variable it shows negative correlation.



Graph I – It shows positive correlation. That is if value of one variable increases then the value of other variable also increases. Example – the more you will study more knowledge you will gain .

Graph II - It shoes no correlations between two variables. Example More is the price of books as the rainfall decreases.

Graph III – It shows negative correlation that is if one variable increases other will also increase. Example- The more mobile usage will lead to weak battery life.

Pearson 's correlation method is known by many names such as

1. Bivariate correlation
2. Pearson's product -moment correlation coefficient(PPMCC)
3. The correlation coefficient

Pearsons correlation coefficient(R) value	Strength	Direction
Less than – 0.5	Strong	Negative
Between -0.3 -0.5	Moderate	Negative
Between 0 and -0.3	Weak	Negative
0	None	None
Between 0 and 0.3	Weak	Positive
Between 0.3 -0.5	Moderate	Positive
Greater than 0.5	Strong	Positive

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4. **SCALING**

It is a pre-processing step applied in model building applied to independent variables to normalize the data within a given range. It helps in optimisation that is speed of analysis increases while scaling.

Why Scaling?

a) Sometimes the given data set contains features which highly varies from each other in terms of magnitude, units and hence it becomes difficult to do analysis as in this case it will only consider the magnitude value not the given units which will lead to wrong analysis.
b) Scaling will lead to improved performance of a model
c) Scaling will reduce the impact of outlier.

Types of Scaling

A. Standardized Scaling
B. Normalized Scaling

Standardized Scaling

Standardization is a scaling method where we centre the values around the mean with a unit standard deviation. This means the mean of attributes is centred around zero and the resultant distribution has a unit of standard deviation.

Formula

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard deviation}}$$

It is used by importing it from sklearn.

Syntax -From sklearn.preprocessing import **StandardScaler**

Normalization Scaling

It is used to adjust the value of features to a common scale. In this the values are rescaled between 0 to 1. It is also known as Min-Max Scaling.

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

We use Min-Max Scaling by using Sklearn library.

Syntax – from Sklearn.preprocessing import MinMaxScaler

Which scale is important and how to decide which scale should be used????.........

Normalization	Standardization
Sensitive to outliers.	Not sensitive to outliers.
Retains the shape of original distribution.	Changes the shape of original distribution.
May not preserve the relationship between the data points.	Preserves the relationship between data points.
Useful when distribution of data is unknown or not a gaussian.	Useful when the distribution of data is Gaussian or unknown

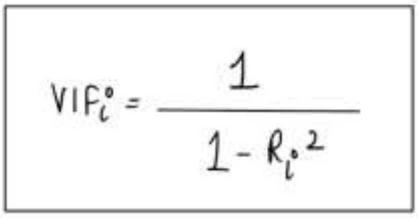
Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5. VIF= Variance Inflation Factor

VIF is a measure of amount of multicollinearity in a analysis of regression. Sometimes there is a high correlation in independent variables in multiple linear regression which adversely affects the regression results. Thus, VIF estimate how much variance of a regression coefficient is inflated due to multicollinearity.

To ensure that model is functioning correctly some tests are run for multicollinearity and VIF is one of them.

FORMULA FOR CALCULATION OF VIF –



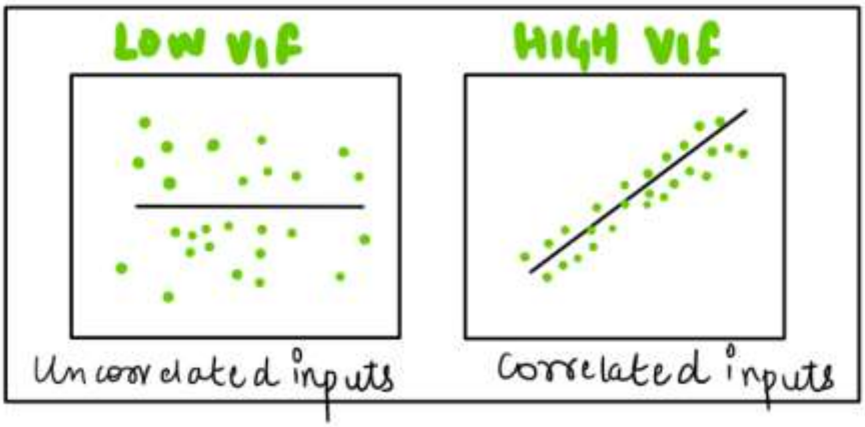
$$VIF_i = \frac{1}{1 - R_i^2}$$

Where R_I^2 = unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones.

When $R_I^2 = 0$, VIF =1 means the ith independent variable are not correlated to the remaining ones, meaning that multicollinearity does not exists.

VIF VALUE	CORRELATION
Equals to 1	**Variables are not correlated.**
Between 1-5	**Variables are moderately correlated.**
More than 5	**Variables are highly correlated.**

Sometimes the value of VIF is infinite which means there is a perfect correlation between the two values. In this case the value of $R_I^2 = 1$ hence VIF = infinity. Since there is a perfect correlation between the two variables one has to be dropped in order to get better model else it will lead to wrong interpretations.



2 graphs are shown above first one has low VIF value that is independent variables are uncorrelated whereas in the second one the Vif value is high which means that variables are highly correlated.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6. Q-Q plot is also known as Quantile-Quantile plot. It is a graphical method for determining whether two samples of data came from the sample population or not. Here quantiles of first data set are plotted against quantile of second data set.

Quantiles – Quantile means the fraction (or percent) of points below the given value.

Here quantile of a sample distribution is plotted against the quantile of theoretical distribution. It helps us to determine if a dataset follows any particular type of probability distribution like uniform, exponential, normal etc.

Q-Q plots helps in determining-

a) If 2 populations are of same distribution.
b) Skewness of distribution.
c) If residuals follow a normal distribution can be verified.

Libraries required to plot Q-Q plot-

i) NumPy
ii) Matplotlib & Seaborn
iii) Statsmodel.api
iv) SciPy. Stats

Importance of Q-Q plot:

1. It is used to validate assumptions about the distribution of residuals.
2. It also helps in detection of outliers.
3. It helps us to identify whether a dataset follows a particular probability distribution.