Report

Objective:

In This project analysis the data , that belongs to a builder community they have historical data related to existing houses.

Introduction:

Me JEMIN PRAJAPATI responsible to commuted the analysis on Housing related data in this data we have following one flow(core analytics flow). This analysis beneficial for multiple builder era to conclude the Prices of the Houses of , also it is importins in terms of building the upcoming policies and deciding the factor responsible for understanding the Price of the House. The motivational of analysing the data is came from standard “Manufacturing analytical rules for analysis” this standards I utilized in Standardising the Null values. Asper my Knowledge and under guidance of MR. PRAFUL VINAYAK BHOYER SIR, I am One conclude this analysis as the standard follower.

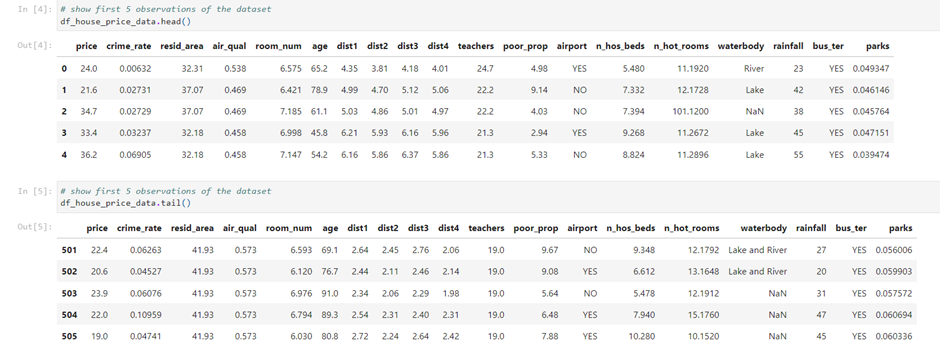
Methodology:-



Data overview:-

Q1. Collect the Detail Information about The Data

ANS:- We have Observe This Data are Historical Existing House Data .



Q2. Collect the Detail Information about The Columns/ Attributes

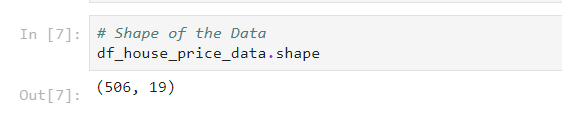
ANS:- This Data we Can This data Different type 19 Columns in their Data.



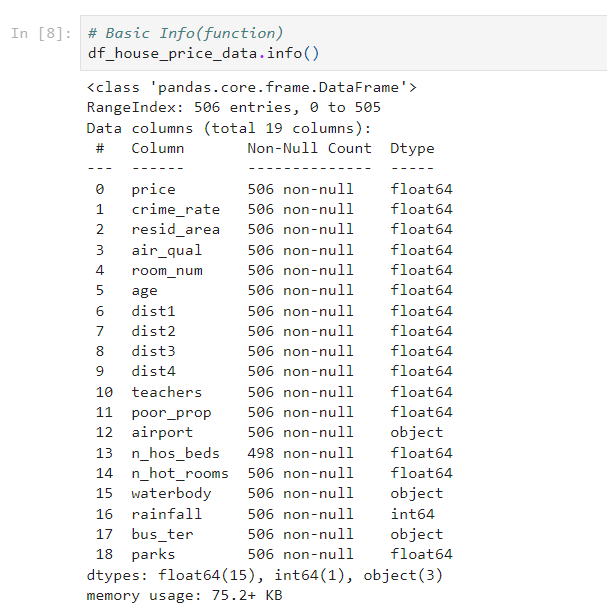
Q3. Show the Basic Statistic Of the Data

ANS:-

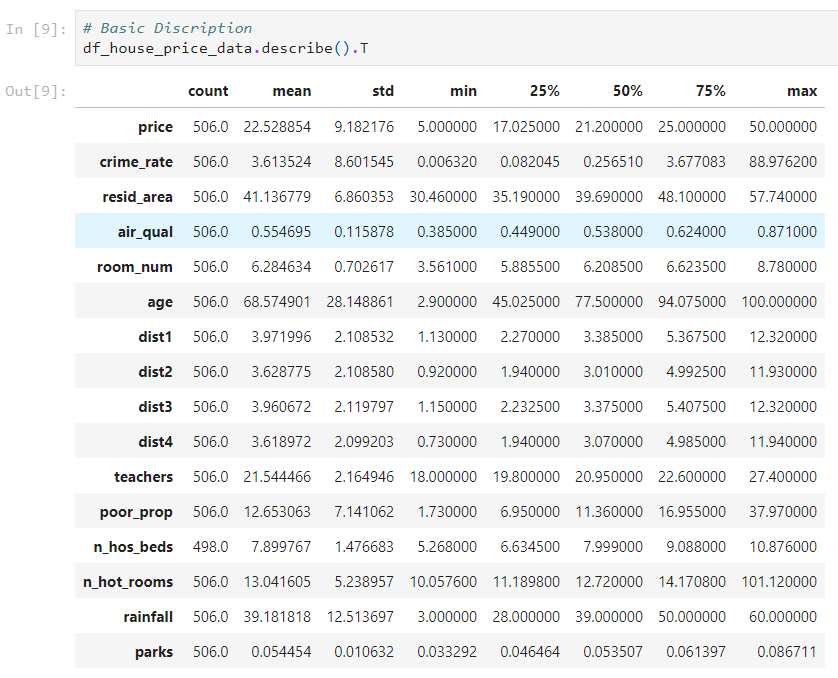
* Screen Short Of Shape and the Discription



* We Can observed This Data 506 Observation and 19 Attributes.
* Screenshot of info and discretion



* We can Observed 16 Attributes are Numerical Data In this Data.
* We can Observed 3 Attributes are Categorical Data In this Data.
* We Can Observed In this Data Valid Null Values.
* Screenshot of Description and The Description



* We can Observed in this Data We can Find The Outlier in the Data are Crime\_Rate, Age, resid\_area, teachers, poor\_prop.

# Data Preprocessing:-

In the Preprocessing we have few stage that we need to follow , those stages we can insides the figure 1

1. Data Cleaning
2. Missing Values Treatment
3. Outlier Treatment

## Data Cleaning.

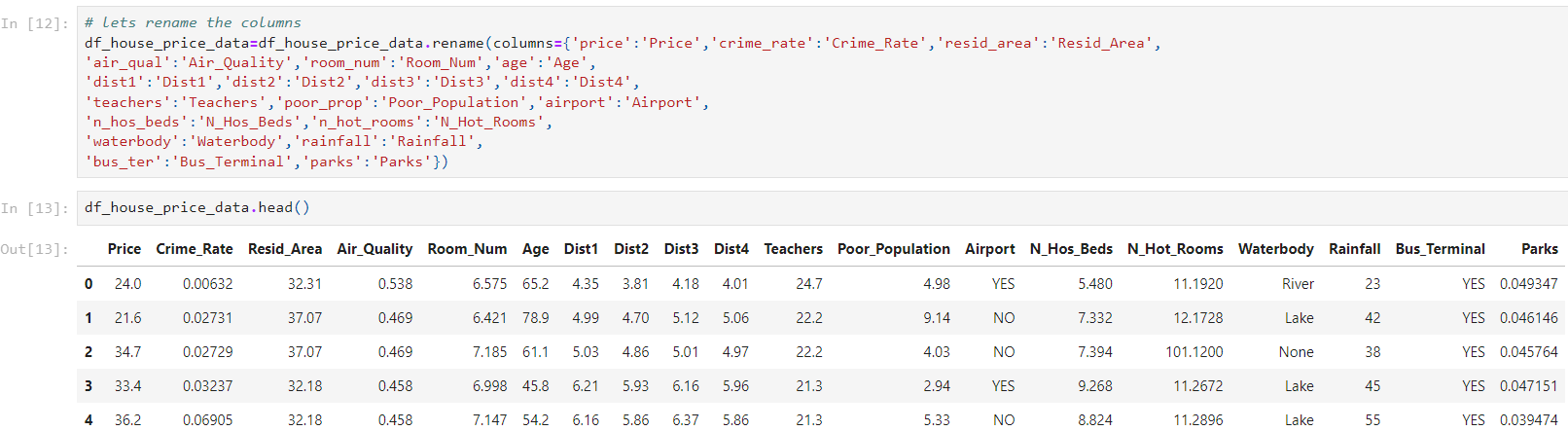
Q1. What is Data Cleaning(Small Paragraph)

ANS:- Data Cleaning Means The To Handle Missing Data ,Correcting Errors, Standardizing Data ,Removing Duplicates, Handling Outliers. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled.

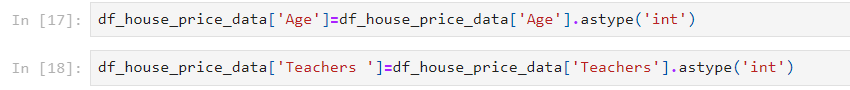
Q2. What we did on out data (Screenshot of the code, information about effect)

ANS:-

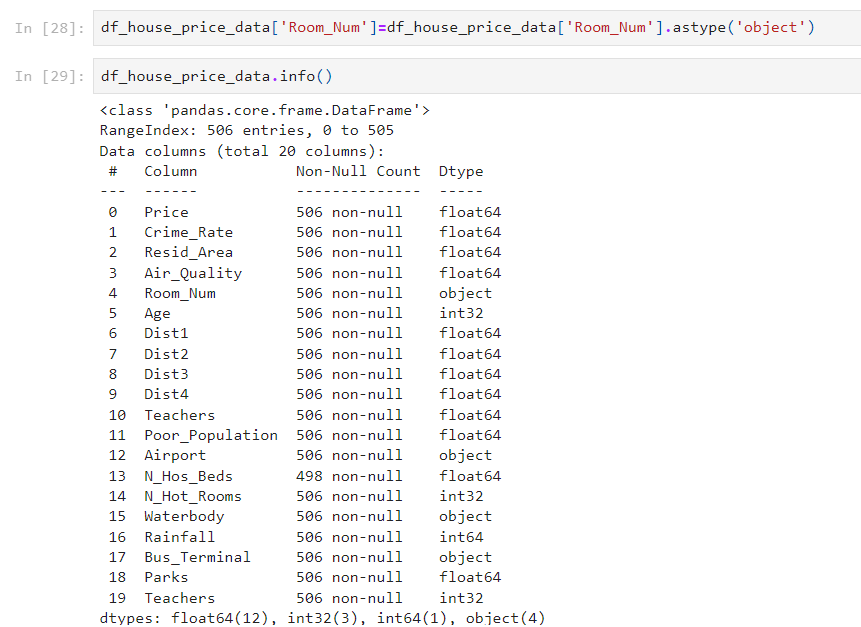
* Rename The All Columns:-



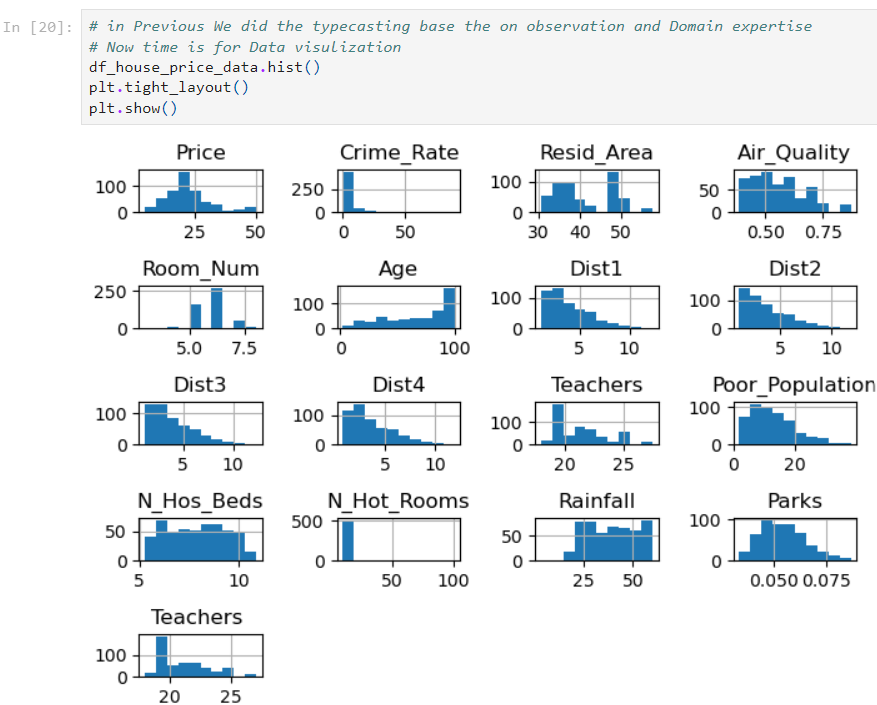
* Typecasting in this data in the Colunms name is Room\_Num, Age, Techers, N\_Hot\_Rooms.







* After typecasting the Visualised the Data usinh hist Map.



## Missing Values Treatment

Q1. What is the Type of missing values (Small Paragraph)

ANS:- There are Three Types Of Missing Values, 1. Missing data are randomly distributed across the variable and unrelated to other variables., 2. Missing data are not randomly distributed but they are accounted for by other observed variables., 3. Missing data systematically differ from the observed values.

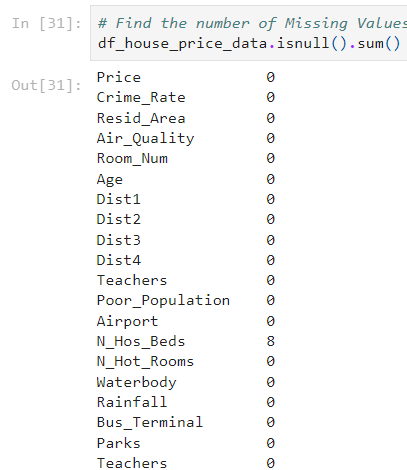
Q2. Rules of missing values as per manufacturing Industry(small Paragraph)

ANS:- There are mainly Three types Of Missing Values Rules as per the manufacturing Industry. First one is When we have 10% Data Can be missing Than The Data Removed The Observation. Second one is When We Observed The Data Can be Percentage between 30 to 70 Percentage then We need to input the Missing Values. Third one is When we have more than 80% values missing in our Data then remove the Attributes.

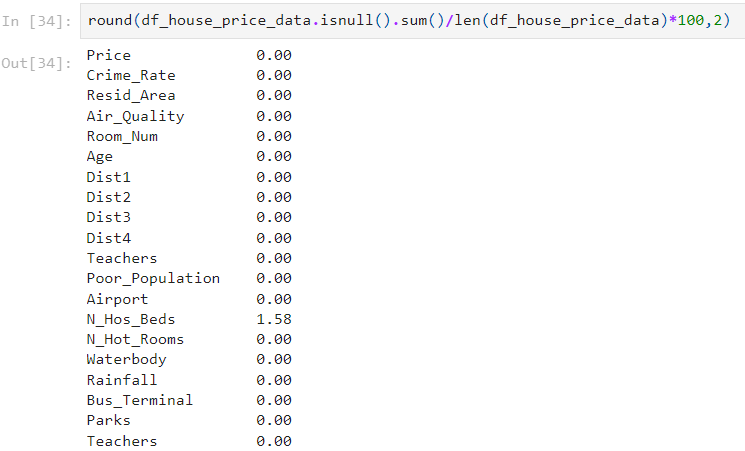
Q3.What we did on our data (Screenshot of code And Information about the effect)

ANS:-

* Find the missing values in our Data

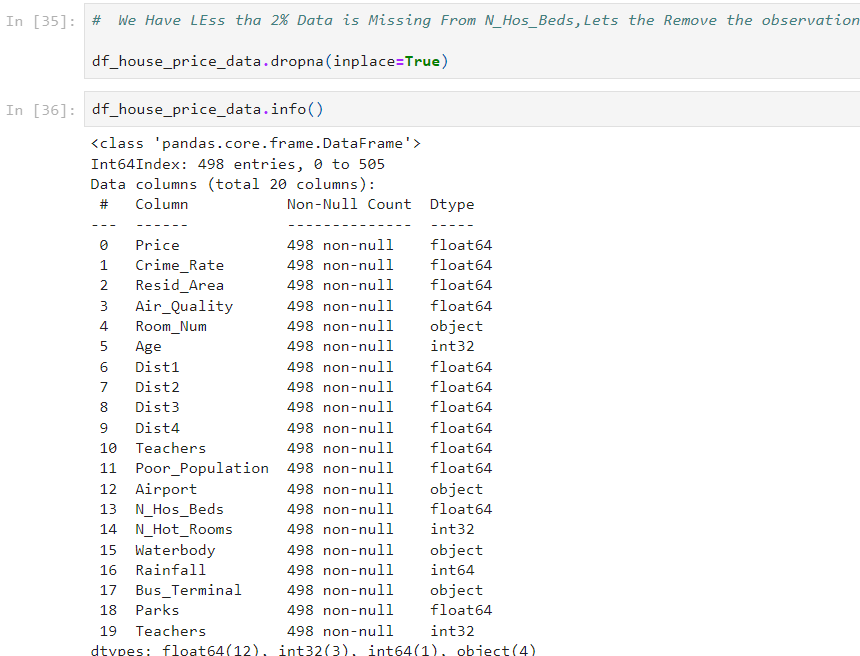


* Check the how many Percentage Data are missing data.



* We Have Less than 2% Data is Missing From N\_Hos\_Beds, Lets the Remove the observation.

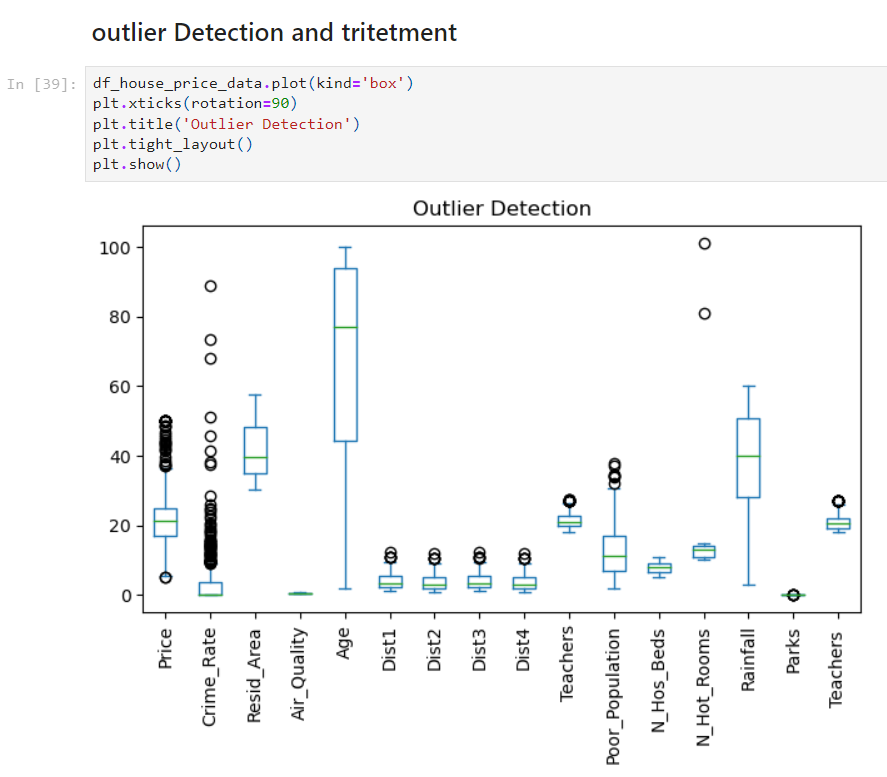
Note(Using Rule number one Rules of missing values as per manufacturing Industry )



# Outlier Treatment

Q1. What is mean by outlier (small Paragraph).

ANS:- An outlier is a datapoint that deviates significantly from a statistical group's average. there might also be outliers that don't belong to specific demographic samples. To sum up, an outlier is a person who, in some way, deviates noticeably from the usual.



* + We can Observed in this data we can show the so many outlier in this data. We can seen the above box plot of the numerical observation of the data we can seen the round circle(O) upper bound and lower bound in this data. It is called Outlier.

Q2. Which method we used to treat the outlier (Small Paragraph).

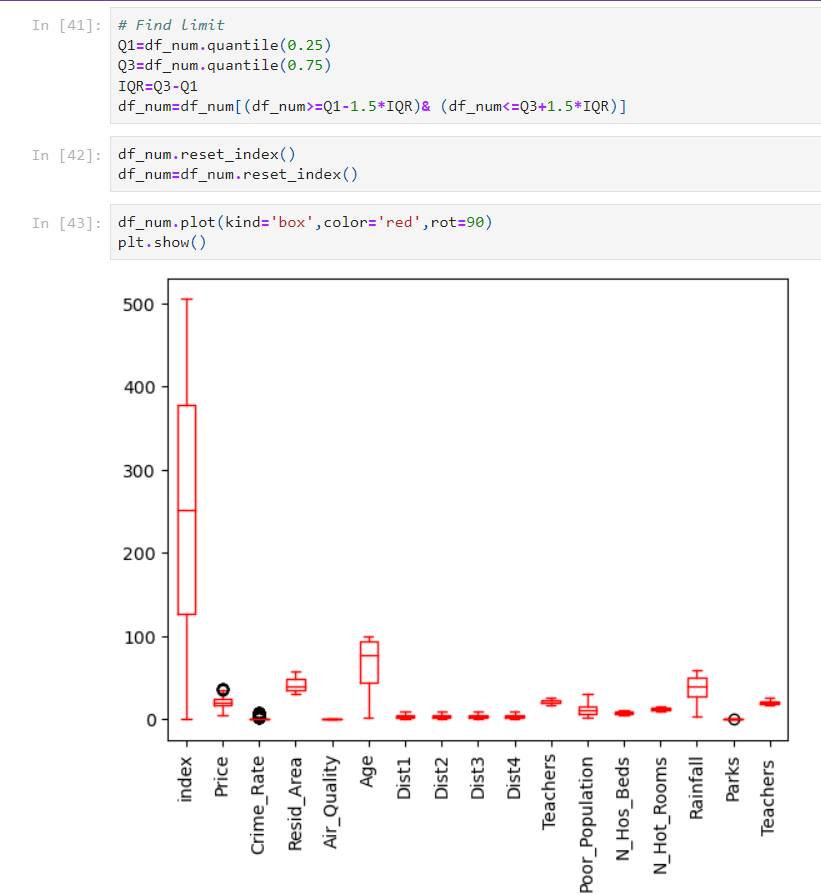
ANS:- When handling outliers in data analysis, there are a few approaches that have advantages and disadvantages. Eliminating them is one method, but this can lead to the loss of potentially important data. An additional choice would be to "trim" the outliers by substituting numbers that are more in line with the main data cluster. This lessens their influence without eliminating them entirely. As an alternative, you can alter the data itself to reduce the impact of the outliers. For example, you can use a log transformation on skewed data. Ultimately, certain statistical methods, such as robust regression, are intended to be less susceptible to outliers in the first place.

Q3.What we did on our data (Screenshot of code And Information about the effect)

* Create the data numeric Dataframe



* Find the Limits ,Reset the index and Plot the Graph.



We can first find the Q1 and Q3. After the find the IQR values. df\_num=df\_num[(df\_num>=Q1-1.5\*IQR)& (df\_num<=Q3+1.5\*IQR)] find the df\_num. after the plot the graph.

# EDA(Exploratory Data Analysis)

It is method that utilize to explore Unseen Data, Also It is an Process of converting Any Raw data (Unseen Data ) into Actionable insights/ Information

Q1. Explain the type of EDA (Exploratory Data Analysis) (Small Paragraph each).

ANS:-

Definition of EDA:- An essential step in data science is exploratory data analysis (EDA), which enables data scientists to learn and comprehend a dataset's primary features. There Three Way to Exploratory Data Analysis, First one is Univariate Analysis, Second one is Bivariate, Third and last one is Multivariate Analysis.

1. Univariate Analysis:- Understanding data analysis starts with a solid understanding of univariate analysis. It focuses on a single variable at a time. To analyze a bag of marbles, for example, a univariate analysis would weigh, characterize, and describe each marble independently. This study uses statistical techniques such finding the mode, or most common hue, the standard deviation, or the degree of size distribution uniformity. Using histograms or box plots to visualize the data, univariate analysis provides a clear picture of each variable's characteristics and distribution before going further into the relationships between them.
2. Bivariate Analysis:- Bivariate analysis digs deeper into your data by analyzing the correlations between two variables. Assume for the moment that you know exam results and study duration. Bivariate analysis would investigate possible correlations between scores and hours rather than just describing each set individually. Plotting the data points on a scatter plot, with each point representing a student's score and hours worked, is one method to accomplish this. By examining the point pattern, you can ascertain whether longer study sessions are consistently associated with improved test scores. Bivariate analysis provides the groundwork for discovering potential connections so that one may understand how changes in one variable may impact changes in another.
3. Multivariate Analysis:- Multivariate analysis extends the scope of data analysis beyond basic pairs or individual variables. Assume you are researching what influences how well students succeed. In addition to examining study hours and test results, multivariate analysis may take into account other factors including sleeping patterns, extracurricular activities, and even class difficulty. It investigates the intricate interactions between all of these factors at once by using methods like regression models. This makes it possible for researchers to spot trends and comprehend how each component affects the final result (performance). Multivariate analysis offers a comprehensive perspective, illuminating the complex network of connections present in your data.

# Univariate Analysis

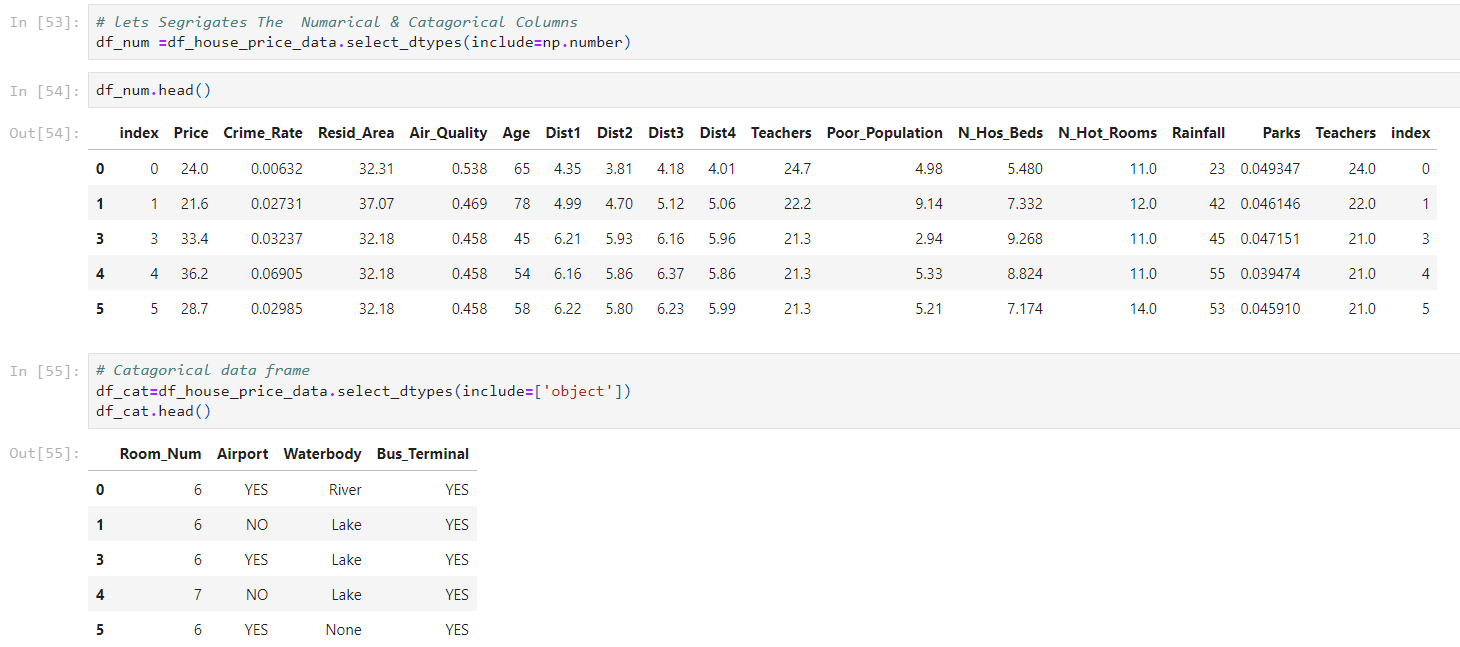
Q1. What is mean by univariate analysis.

ANS:- Understanding data analysis starts with a solid understanding of univariate analysis. It focuses on a single variable at a time. To analyze a bag of marbles, for example, a univariate analysis would weigh, characterize, and describe each marble independently. This study uses statistical techniques such finding the mode, or most common hue, the standard deviation, or the degree of size distribution uniformity. Using histograms or box plots to visualize the data, univariate analysis provides a clear picture of each variable's characteristics and distribution before going further into the relationships between them.

Q2. What we did With Our Data (Screeshot of code ,Screeshot of Graph , Inference) Column wise<Heading 2>

ANS:-

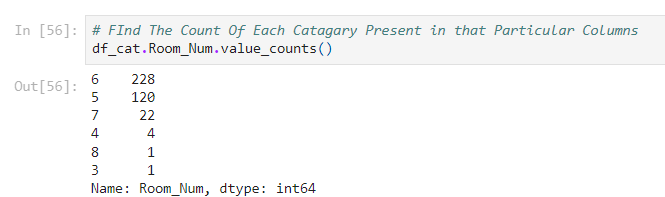
* lets Segregates The Numerical & Categorical Columns



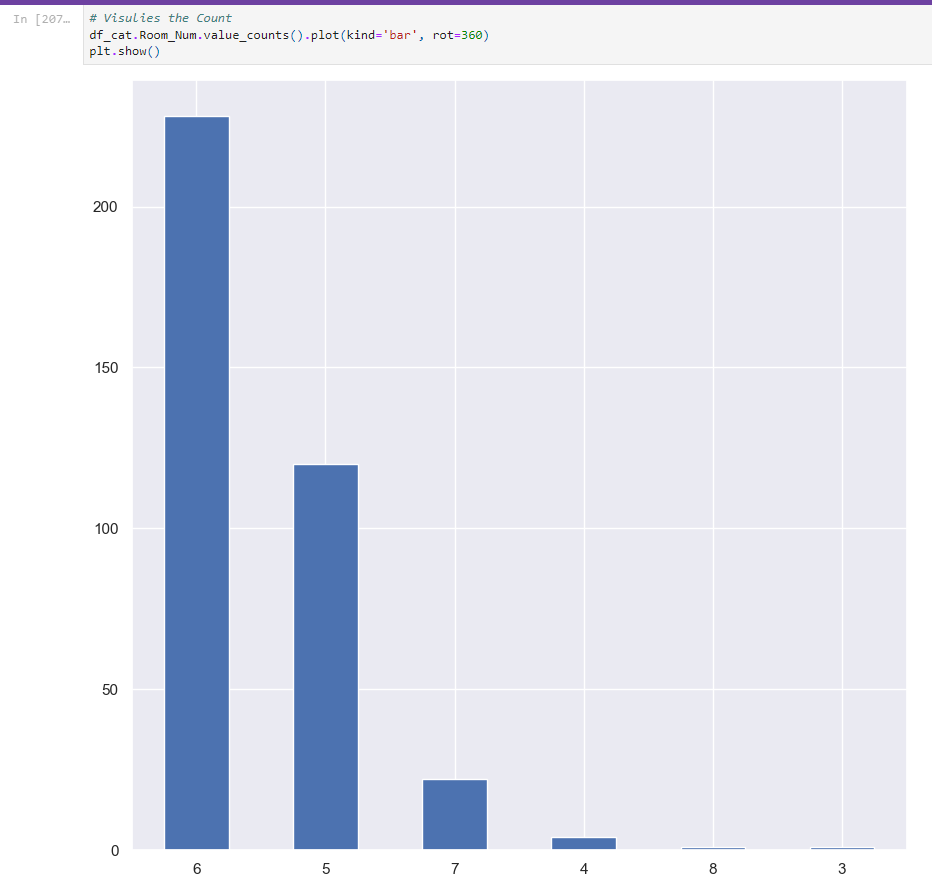
Univarite Analysis on The Categorical data :-

## Room\_Num:-

* FInd The Count Of Each Catagary Present in that Particular Columns.



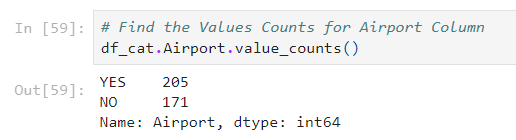
* Visualised the Count:-



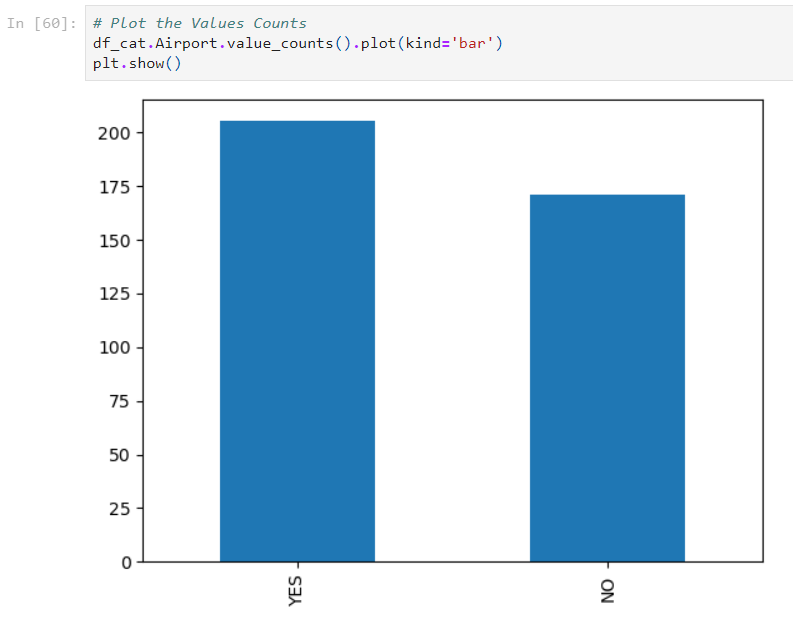
* We have imbalance data in Room\_Num Column
* 6 is Catagary Who is Dominating is More, its counts is 266
* 3 is A Catagary Havind list numbers of Observation ,its counts is 2
* 8 and 4 are catagary Having same Number of Observation

## Airport:-

* Find the Values Counts for Airport Column:-



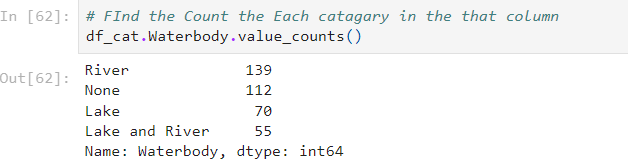
* # Plot the Values Counts:-



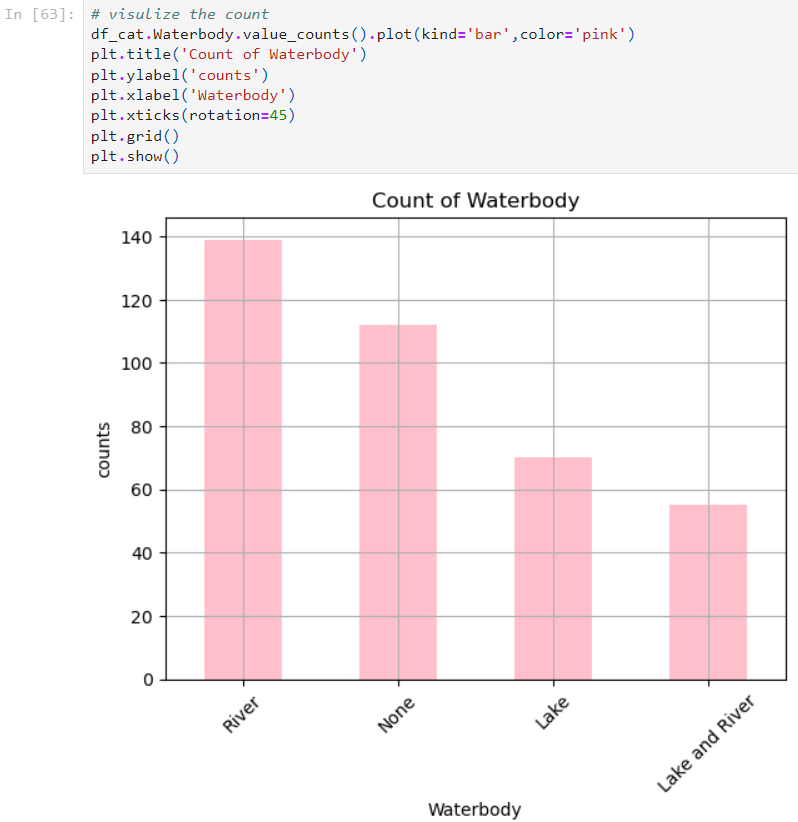
* We Have balecnce Data in this Airport Column
* In The Data 273 Palces Where Airport is Present
* In The Data 225 Palces Where Airport isnot Present

## Waterbody:-

* Find the Count the Each catagary in the that column.



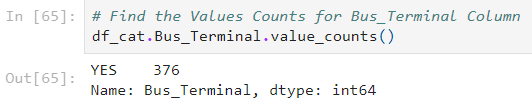
* visualize the count:-



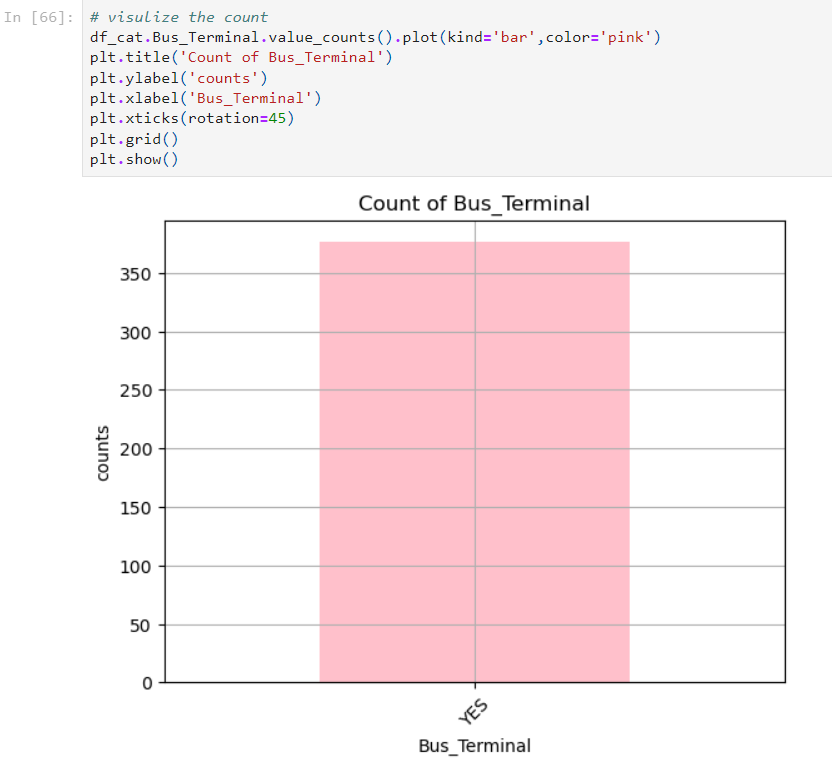
* Most of the Places having River as Water body ,its counts is 178
* where is few Places lacks and river are the waterbody
* The Data is Imbalance
* 155 Place are not Available waterbody

## Bus\_Terminal:-

* Find the Count the Each catagary in the that column.

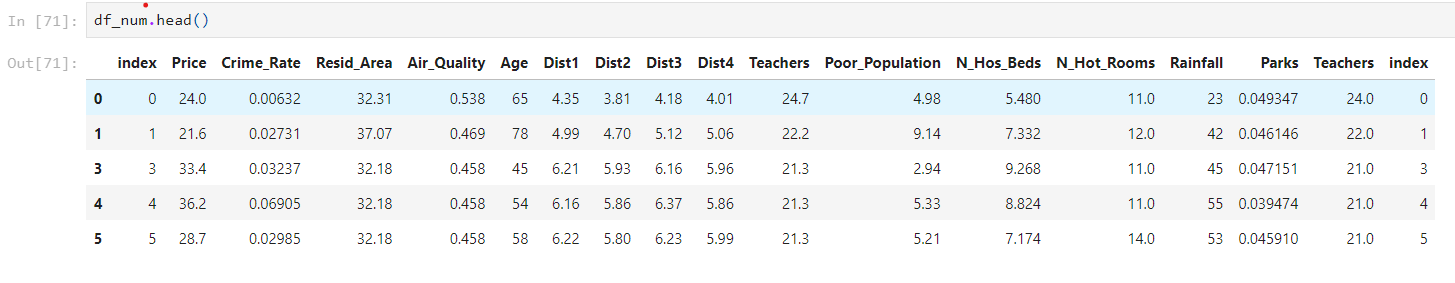


- visualize the count:-



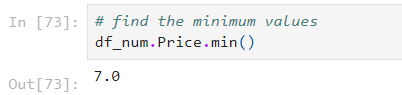
* We Have observed 498 Bus\_Terminal Avalbale
* This Data is Balance Data
* All The Places we have , contains the bus terminal

Univariate Analysis on The Numerical data :-

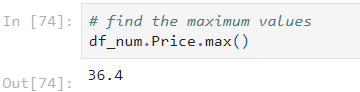


## Price:-

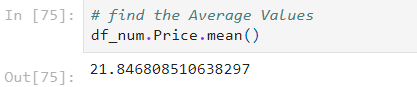
* Find the minimum values:-



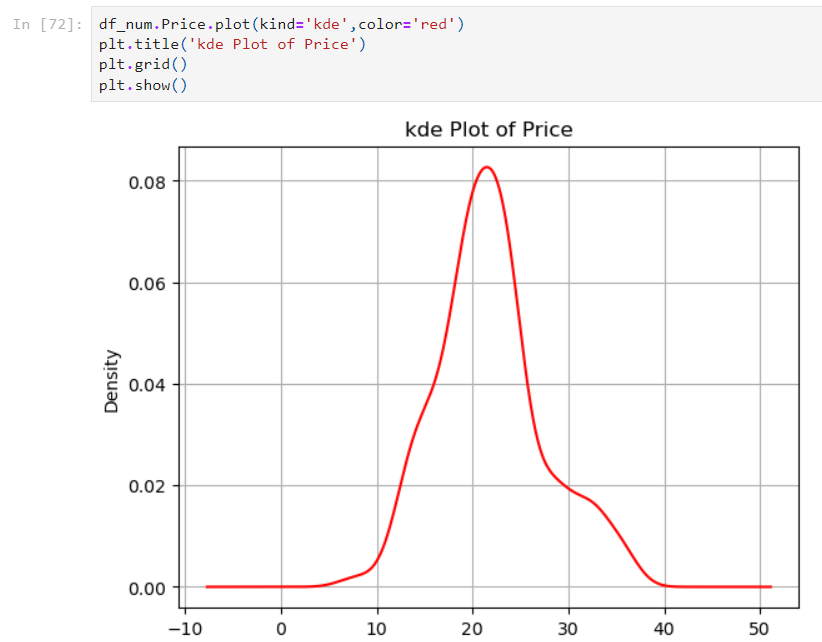
* Find the maximum values:-



* Find the average values:-



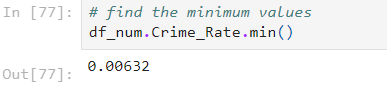
* Visualize the observation Price.



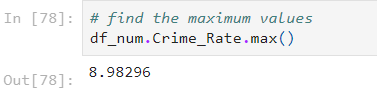
* + This Data slidely Normaly Distributated
  + the Houses We Have Thier Prices lies in between 8 to 40 Lacs
  + the house we have list price is 5 lacs
  + the house we have maximum price is 50 lacs
  + the house we have avarage price is 22.5 lacs

## Crime\_Rate:-

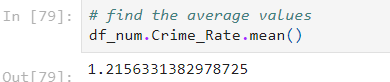
* Find the minimum values:-



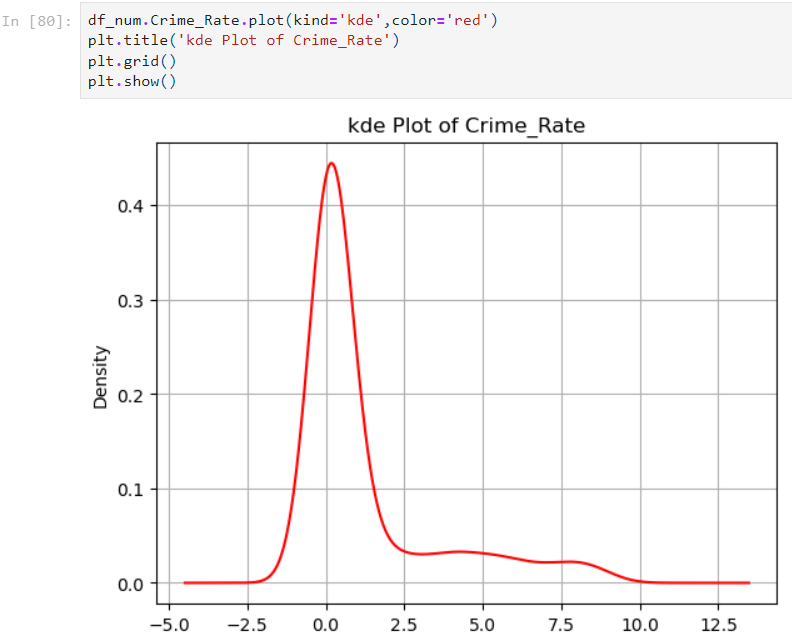
* Find the maximum values:-



* Find the average values:-



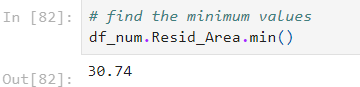
* Visualize the observation Crime\_Rate:-



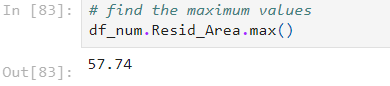
* This Data slidely Normaly Distributated
* the house we have list Crime Rate is 0.00632
* the house we have maximum Crime Rate is 88.9762
* the house we have avarage Crime Rate is 3.6
* the Houses We Have Thier Crime Rate lies in between 0.00632 to 25

## Resid\_Area:-

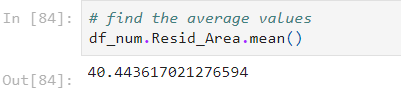
* Find the minimum values:-



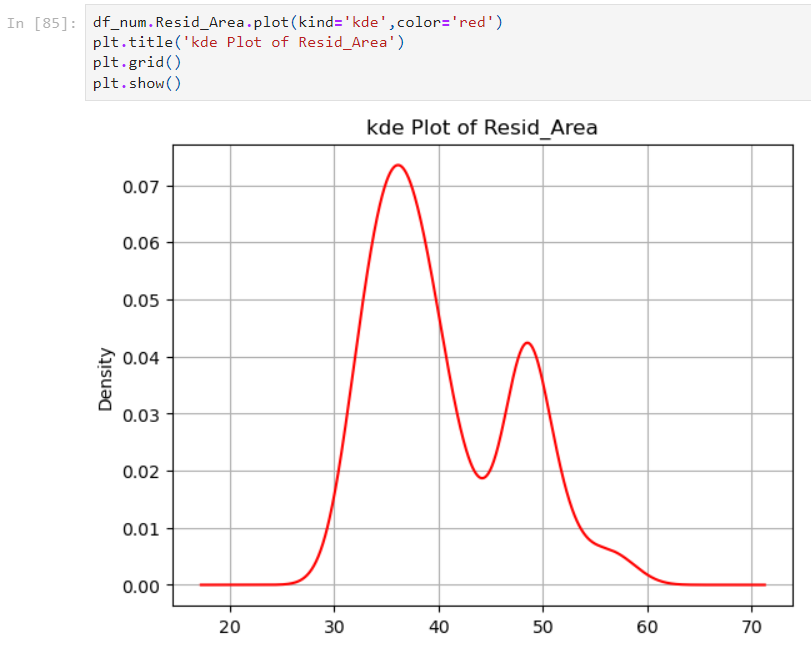
* Find the maximum values:-



* Find the average values:-



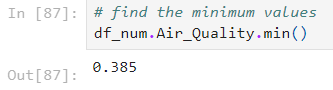
* Visualize the observation Resid\_Area:-



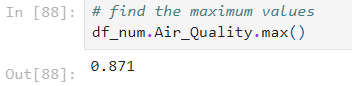
* This Data multi model Distributated
* the house we have list Resid\_Area is 30.46
* the house we have maximum Resid\_Area is 88.9762
* the house we have avarage Resid\_Area is 41.1
* 1 st model the Houses We Have Thier Resid\_Area lies in between 25 to 45
* 2 nd model the Houses We Have Thier Resid\_Area lies in between 45 to 60

## Air\_Quality:-

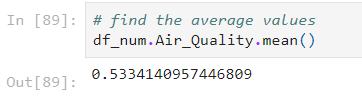
* Find the minimum values:-



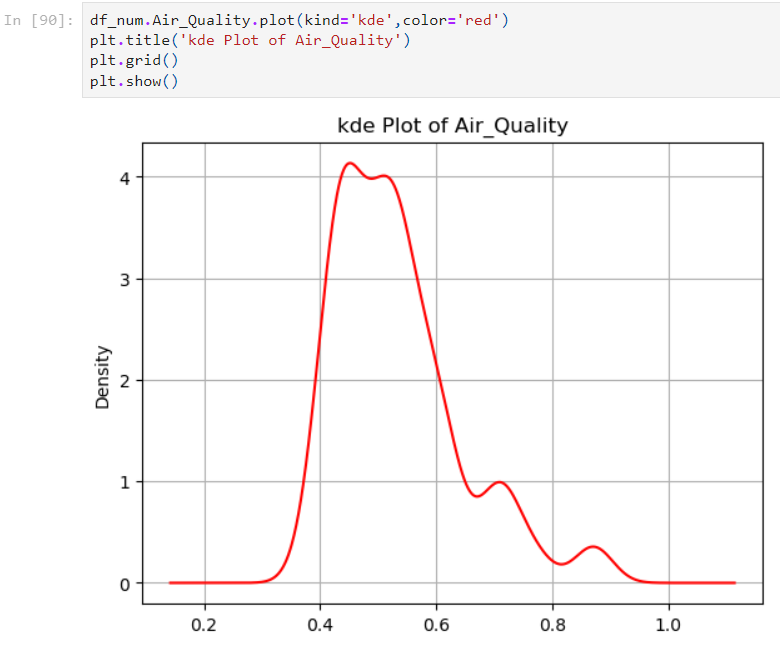
* Find the maximum values:-



* Find the average values:-



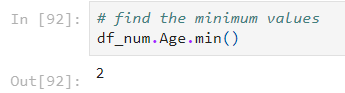
* Visualize the observation Air\_Quality:-



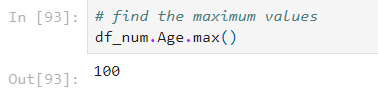
* This Data slitly normaly Distributated
* the house we have list Air\_Quality is 0.385
* the house we have maximum Air\_Quality is 0.871
* the house we have avarage Air\_Quality is 0.555

## Age:-

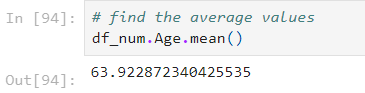
* Find the minimum values:-



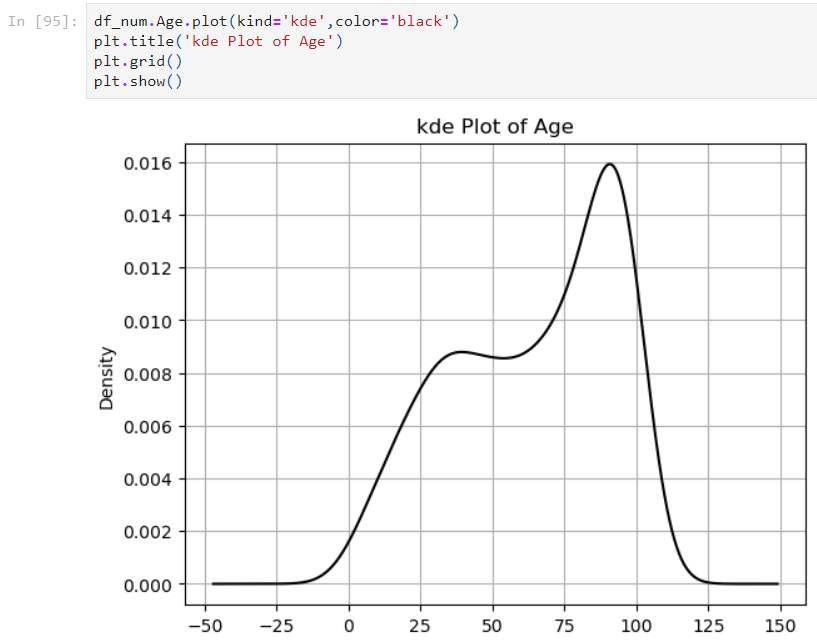
* Find the maximum values:-



* Find the average values:-



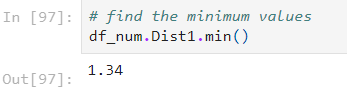
* Visualize the observation Age:-



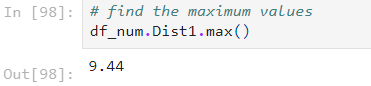
* This Data left skeud Distributated
* the house we have list age is 2
* the house we have maximum age is 100
* the house we have avarage age is 68.04

## Dist1:-

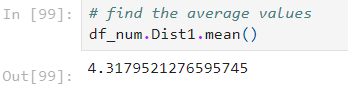
* Find the minimum values:-



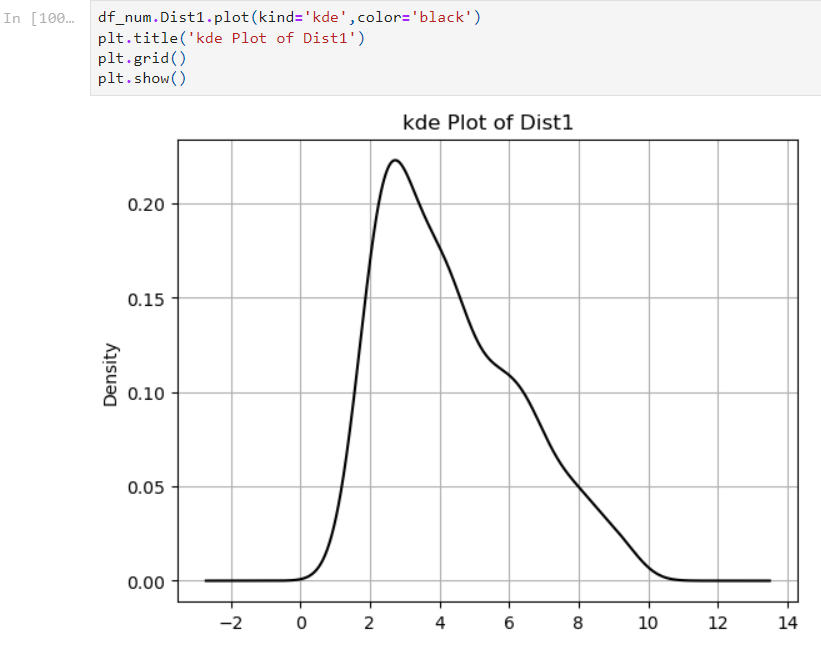
* Find the maximum values:-



* Find the average values:-



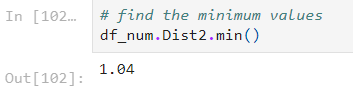
* Visualize the observation Dist1:-



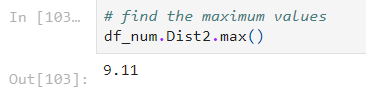
* We have minimum Distence of the industriyal area is 1.13 km
* we have maximum Distence od the industriyal area is 12.32 km
* we have average Distence of the industriyal area is 3.98 km

## Dist2:-

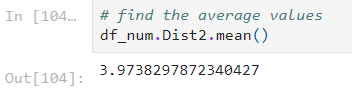
* Find the minimum values:-



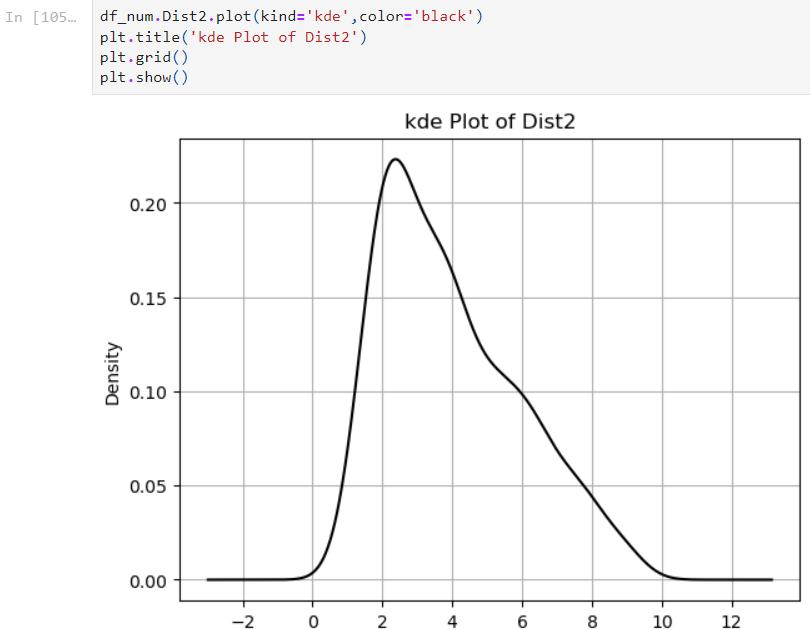
* Find the maximum values:-



* Find the average values:-



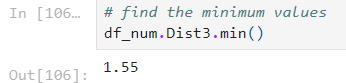
* Visualize the observation Dist2:-



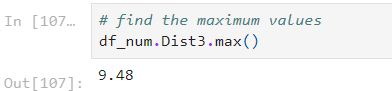
* We have minimum Distence of the industriyal area is 0.92 km
* we have maximum Distence od the industriyal area is 11.32 km
* we have average Distence of the industriyal area is 3.68 km

## Dist3:-

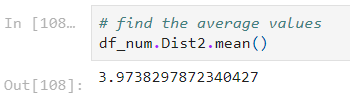
* Find the minimum values:-



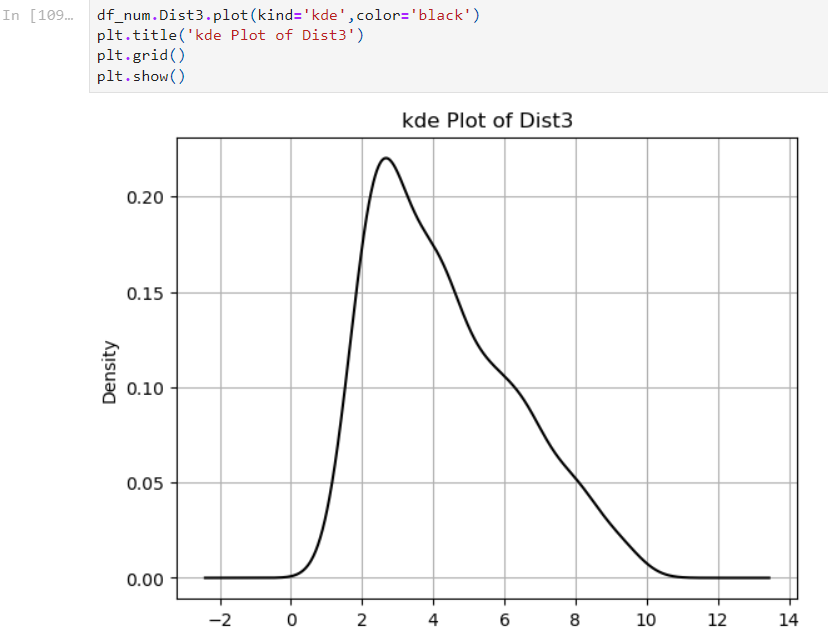
* Find the maximum values:-



* Find the average values:-



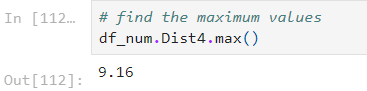
* Visualize the observation Dist3:-



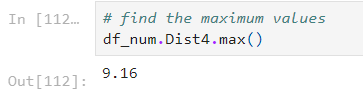
* We have minimum Distence of the industriyal area is 1.15 km
* we have maximum Distence od the industriyal area is 12.32 km
* we have average Distence of the industriyal area is 3.68 km

## Dist4:-

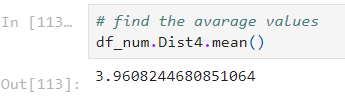
* Find the minimum values:-



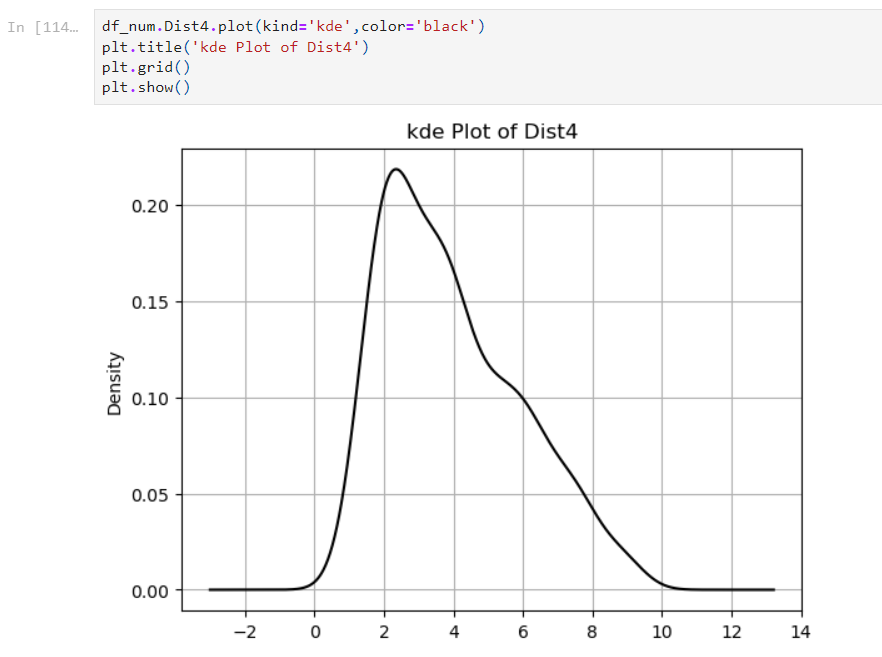
* Find the maximum values:-



* Find the average values:-



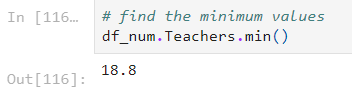
* Visualize the observation Dist4:-



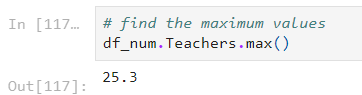
* We have minimum Distence of the industriyal area is 0.73 km
* we have maximum Distence od the industriyal area is 11.94 km
* we have average Distence of the industriyal area is 3.68 km

## Teachers:-

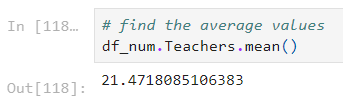
* Find the minimum values:-



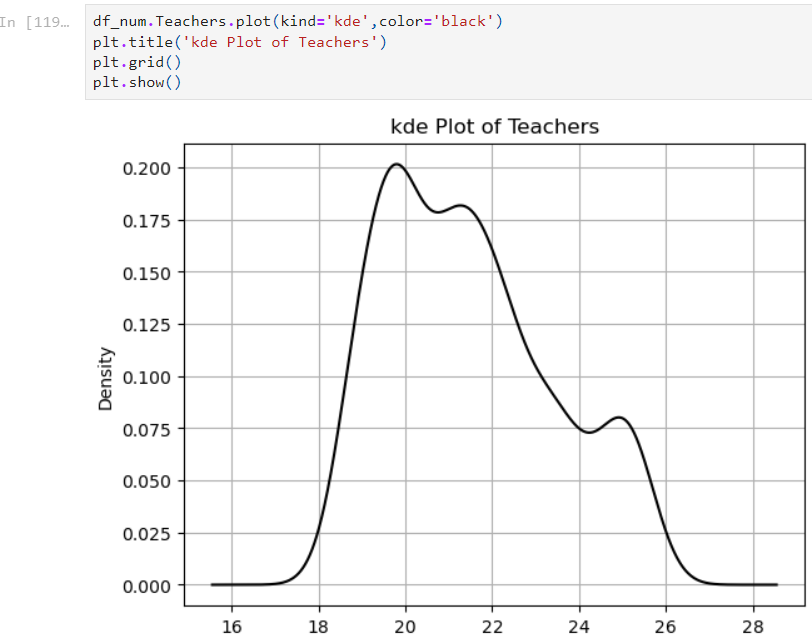
* Find the maxiimum values:-



* Find the minimum values:-



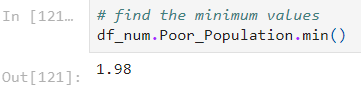
* Visualize the observation Techers:-



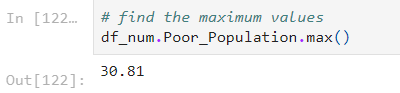
* we have minimum ratio of teachers around the house is 18
* we have maximum ratio of teachers around the house is 27.4
* we have average ratio of teachers around the house is 21.54

## Poor\_Population:-

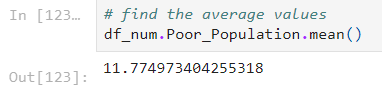
* Find the minimum values:-



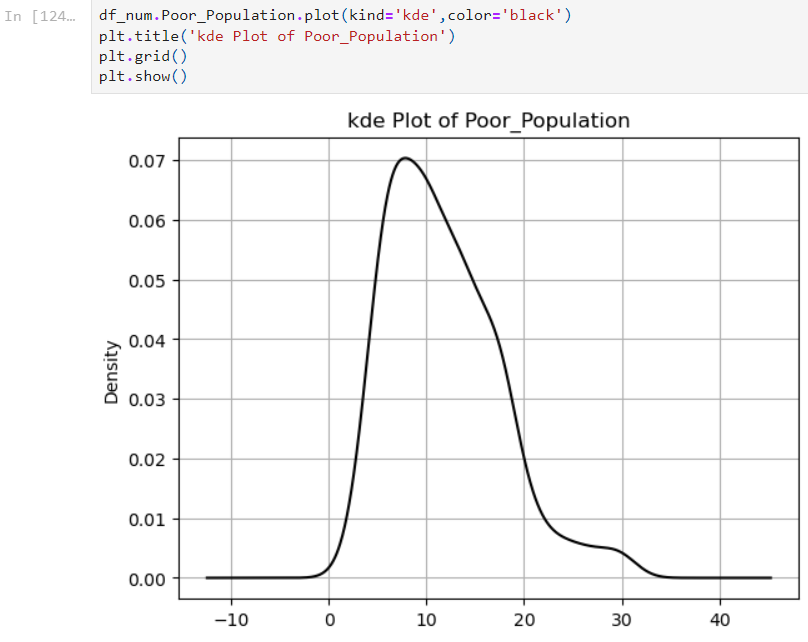
* Find the maximum values:-



* Find the average values:-



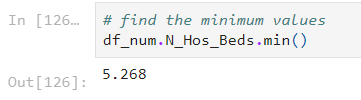
* Visualize the observation Poor\_Population:-



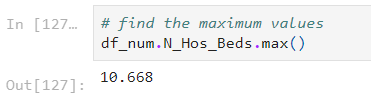
* we have ratio of minium poor population is 1.73
* we have ratio of maximum poor population is 37.97
* we have ratio of average poor population is 12.59
* we have ratio of poor population lies between 0.5 to 40

## N\_Hos\_Beds:-

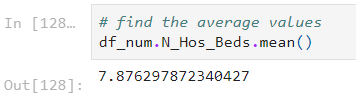
* Find the minimum values:-



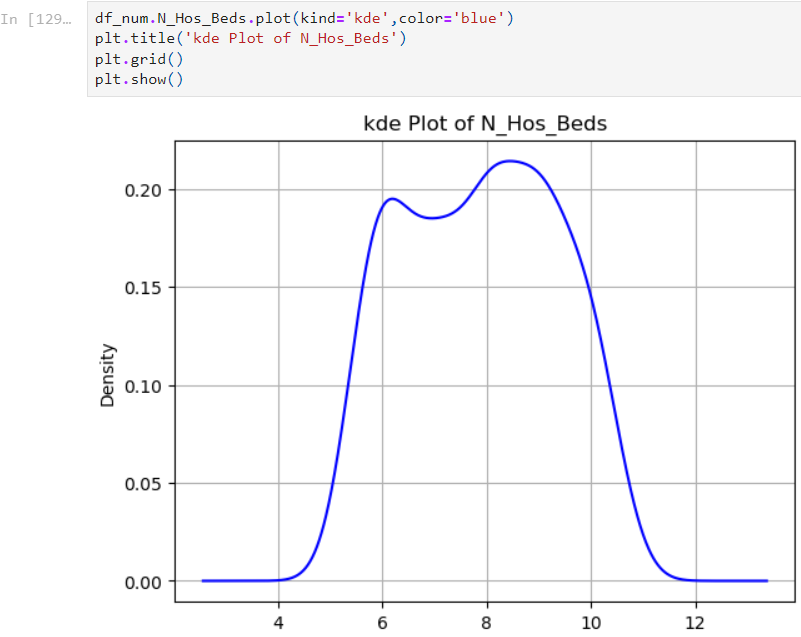
* Find the maximum values:-



* Find the average values:-



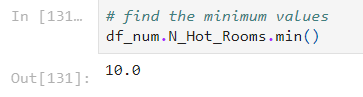
* Visualize the observation N\_Hos\_Beds:-



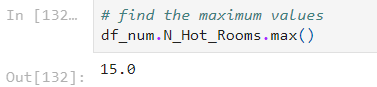
* his data is multi model distributed
* There houses we have list N\_Hos\_Beds is 5.268
* The house we have with maximum N\_Hos\_Beds is 10.876
* The house with average N\_Hos\_Beds is 7.899767068273099
* 1st The house with the N\_Hos\_Beds lies between 4.1 and 7.5
* 2nd The house with the N\_Hos\_Beds lies between 7.5 and 11.6

## N\_Hot\_Rooms:-

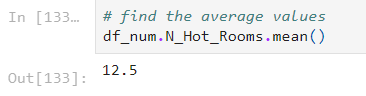
* Find the minimum values:-



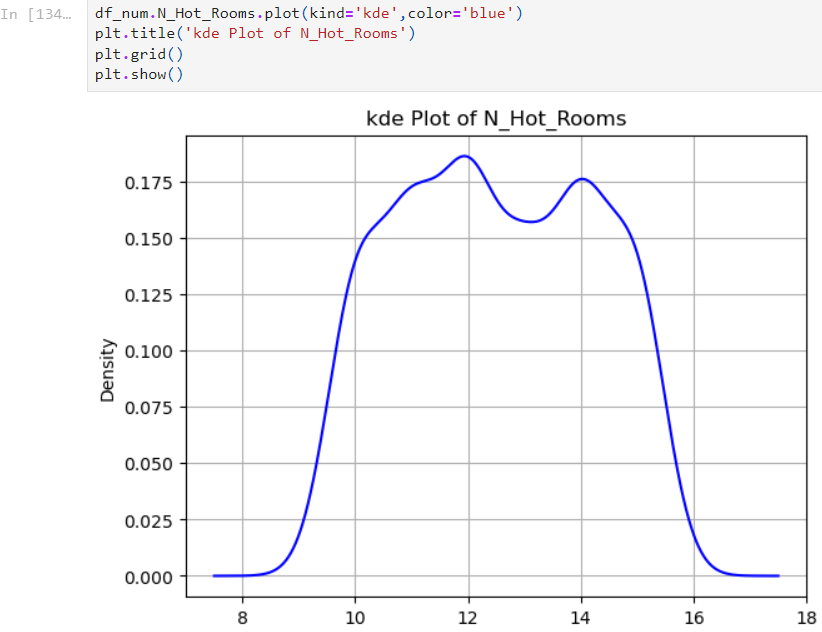
* Find the maximum values:-



* Find the average values:-



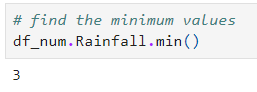
* Visualize the observation N\_Hot\_Rooms:-



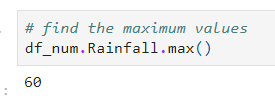
* This data is single model distributed
* There houses we have list N\_Hot\_Rooms is 10
* The house we have with maximum N\_Hot\_Rooms is 101
* The house with average N\_Hot\_Rooms is 12.867469879518072
* The house with the N\_Hot\_Rooms lies between 1 and 22

## Rainfall:-

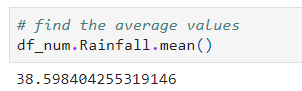
* Find the minimum values:-



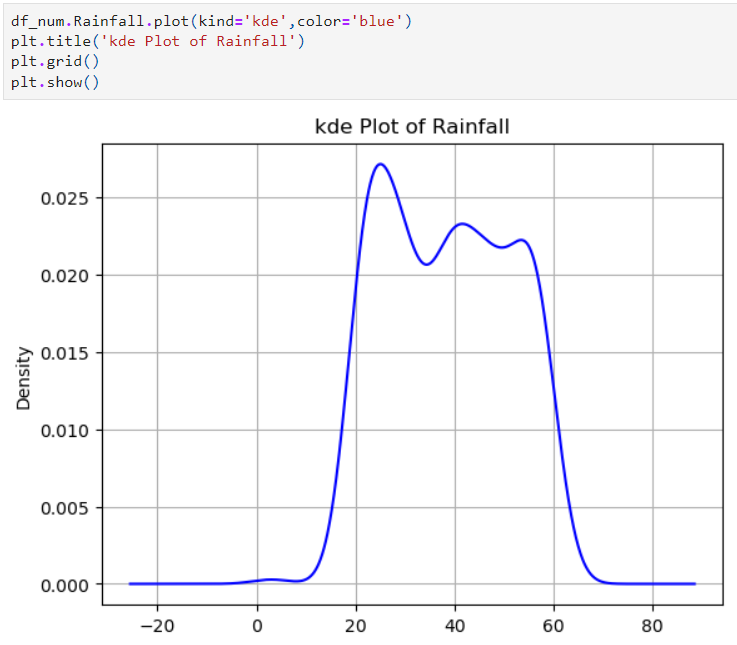
* Find the maximum values:-



* Find the average values:-



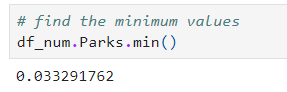
* Visualize the observation Rainfall:-



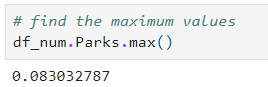
* This data is multi model distributed
* There houses we have list Rainfall is 3
* The house we have with maximum Rainfall is 60
* The house with average Rainfall is 39.30522088353413
* 1st The house with the Rainfall lies between 10 and 30
* 2nd The house with the Rainfall lies between 30 and 50
* 3nd The house with the Rainfall lies between 50 and 70

## Parks:-

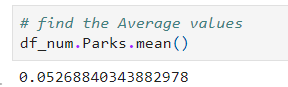
* Find the minimum values:-



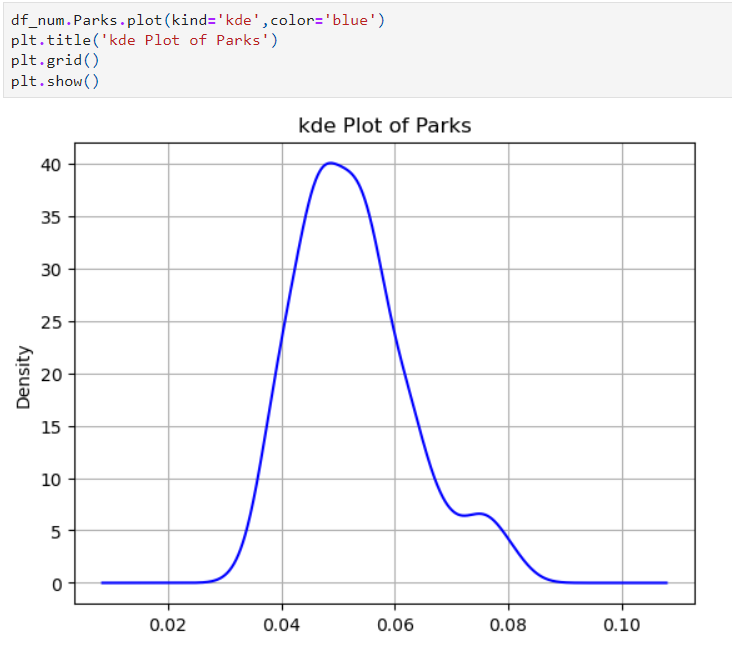
* Find the maximum values:-



* Find the average values:-



* Visualize the observation Parks:-



* This data is single model distributed
* There houses we have list Parks is 0.033291762
* The house we have with maximum Parks is 0.086711473
* The house with average Parks is 0.05436170961044177
* The house with the Parks lies between 0.03 and 0.08

# Bivariate Analysis

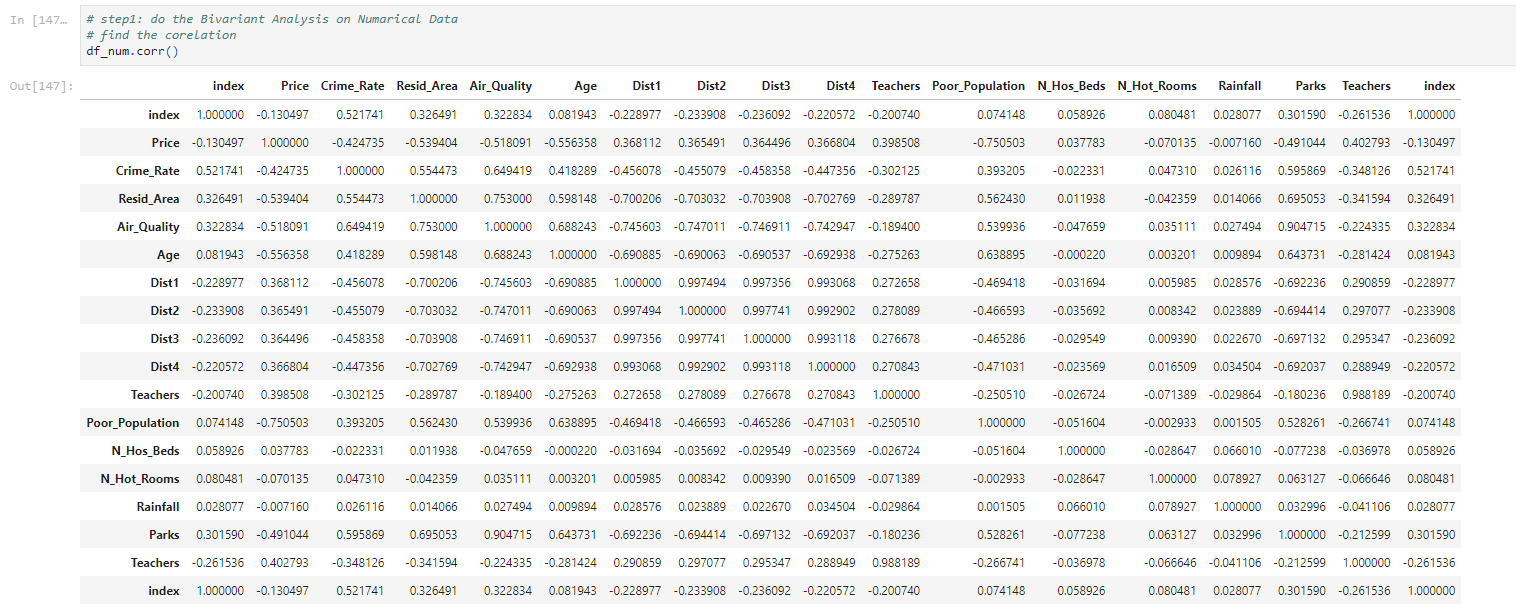
Q1. What is mean by Bivariate analysis(Small Paragraph)

ANS:- Bivariate analysis digs deeper into your data by analyzing the correlations between two variables. Assume for the moment that you know exam results and study duration. Bivariate analysis would investigate possible correlations between scores and hours rather than just describing each set individually. Plotting the data points on a scatter plot, with each point representing a student's score and hours worked, is one method to accomplish this. By examining the point pattern, you can ascertain whether longer study sessions are consistently associated with improved test scores. Bivariate analysis provides the groundwork for discovering potential connections so that one may understand how changes in one variable may impact changes in another.

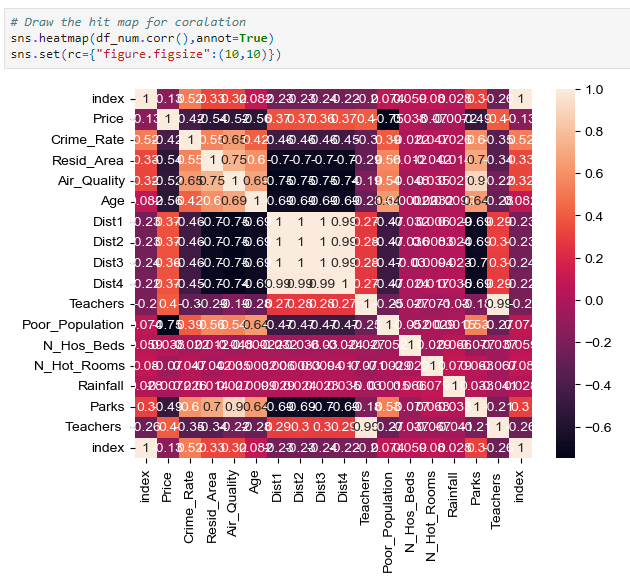
Q2. Method that we use to find the Pairs (Screenshot of code, Screenshot of Graph, Inference)

ANS:-

* step1: do the Bivariate Analysis on Numerical Data.
* find the correlation.



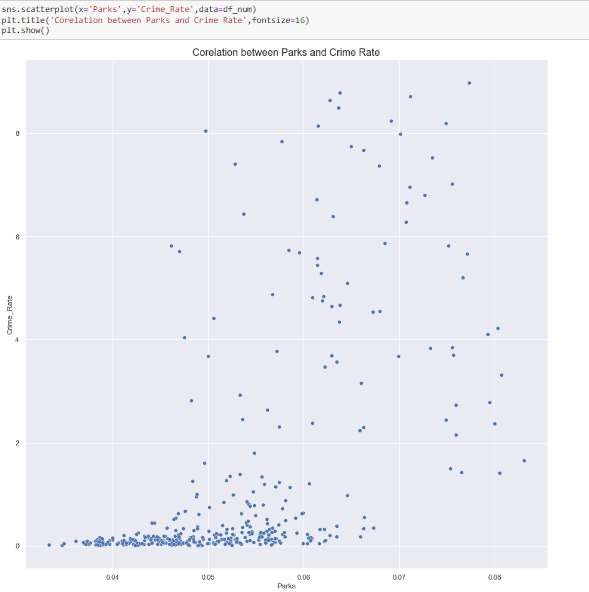
* Draw the hit map for correlation.

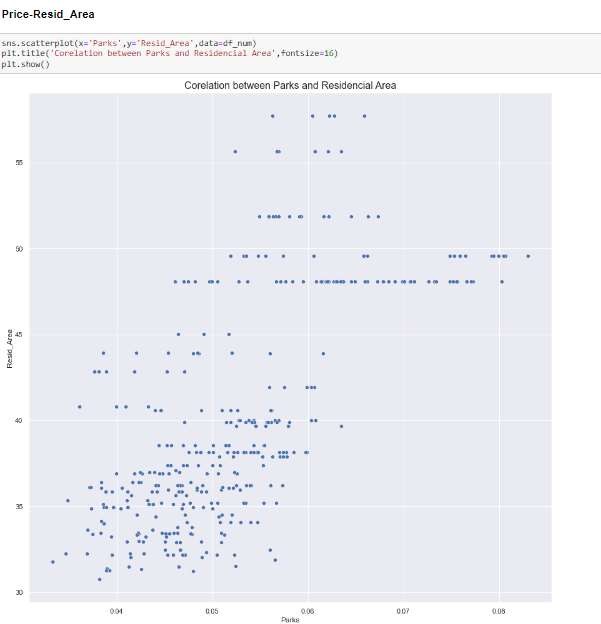


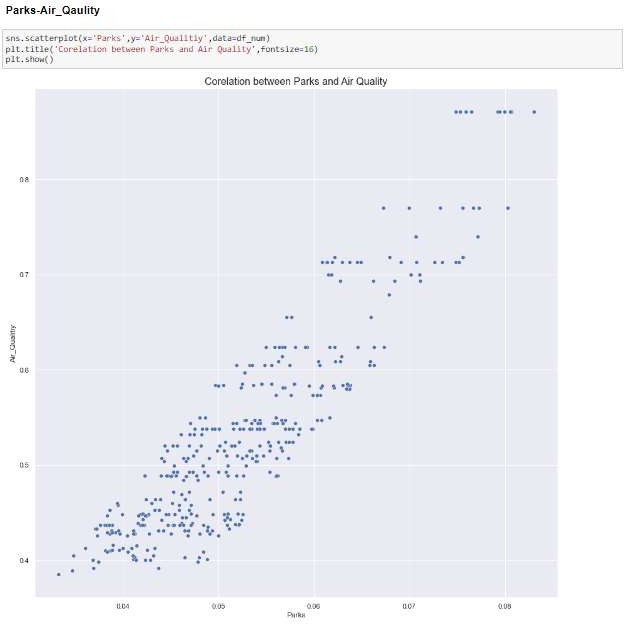
Q3. What we did With Our Data (Screenshot of code, Screenshot of Graph, Inference) Column wise<Heading 2>

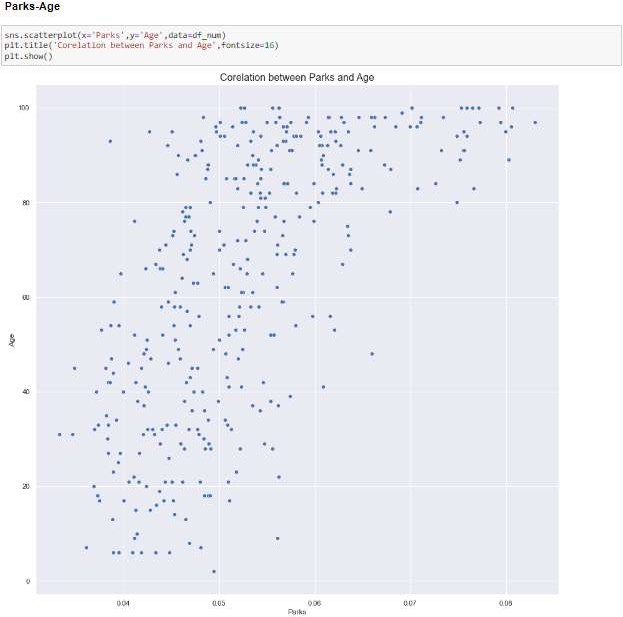
ANS:-

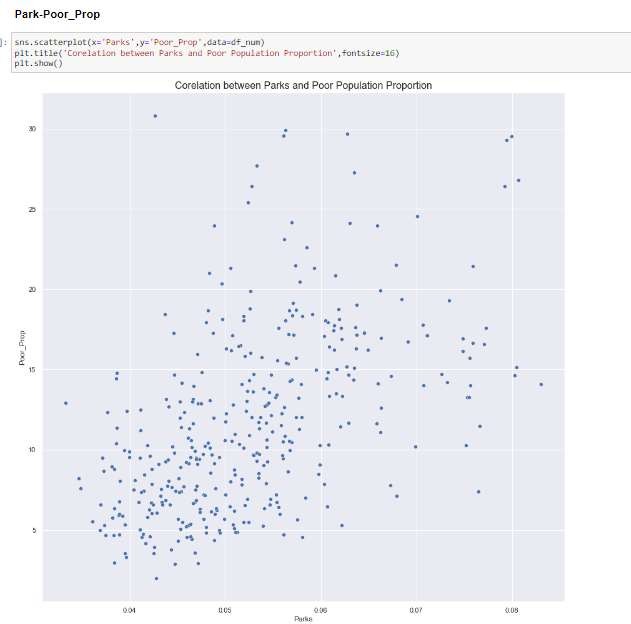
* Correlation between Parks and Crime\_Rate:-

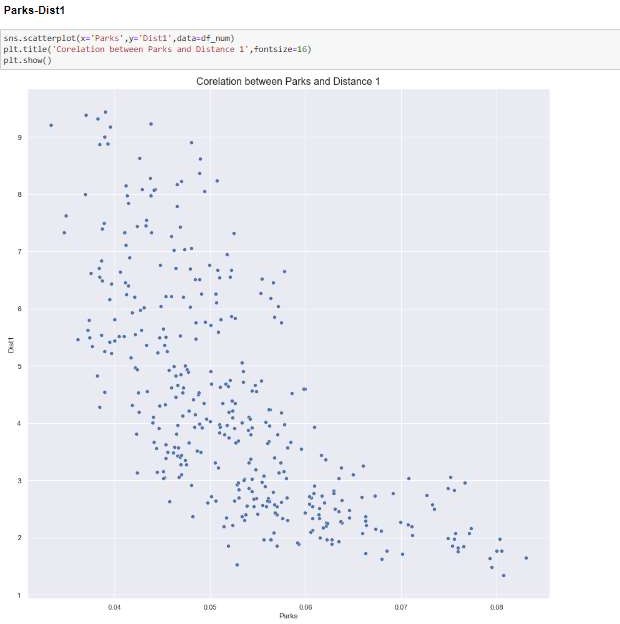


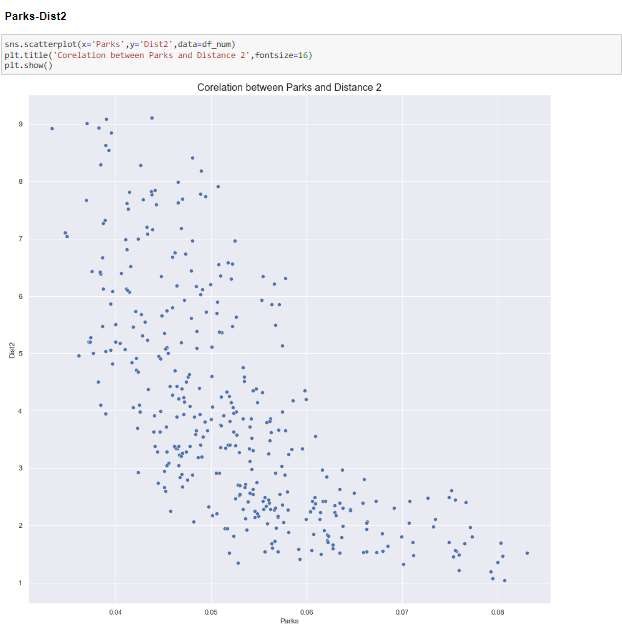


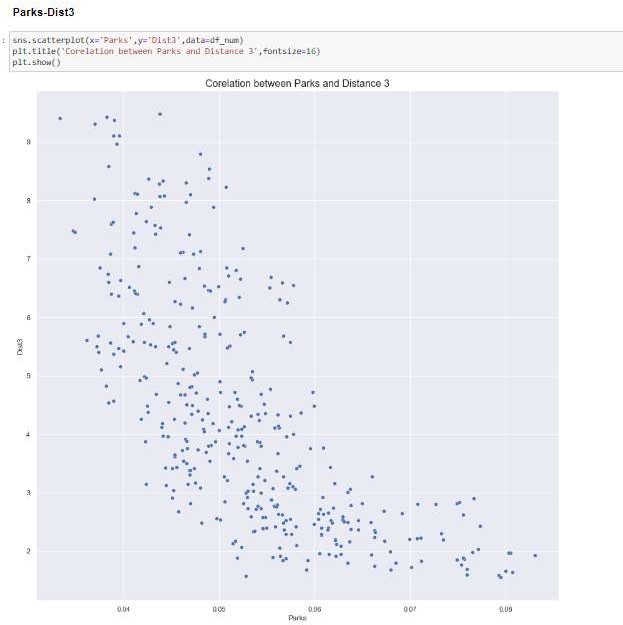


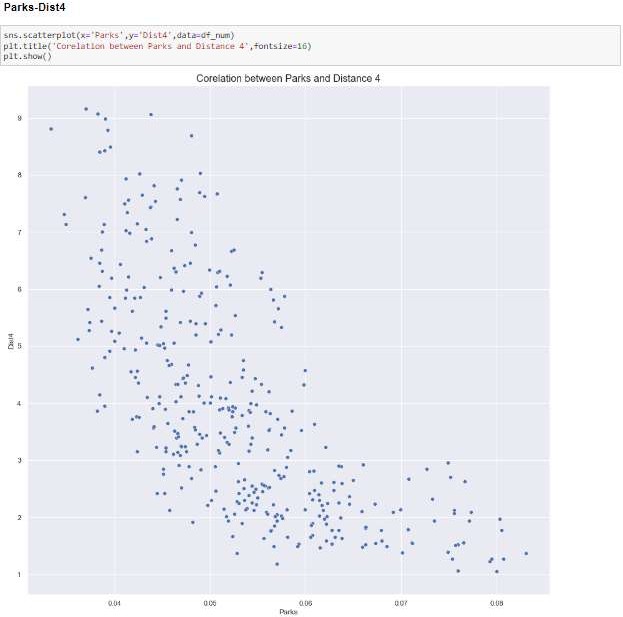


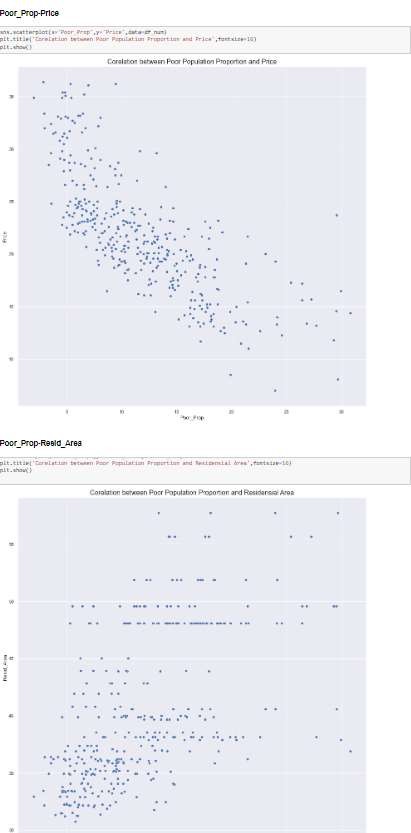


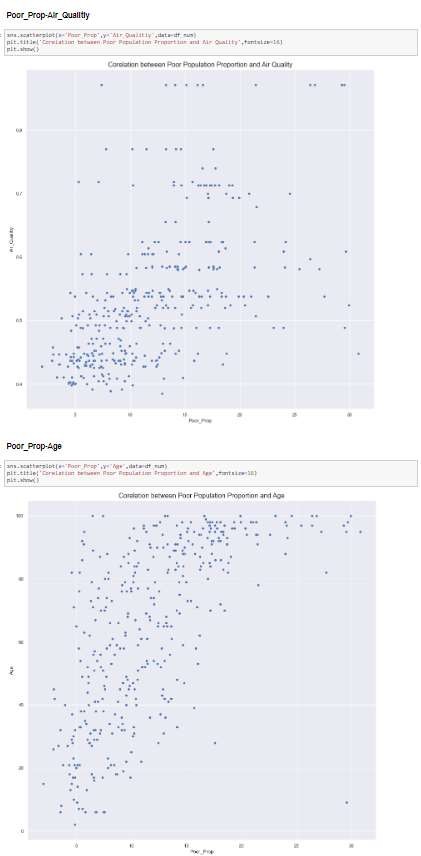


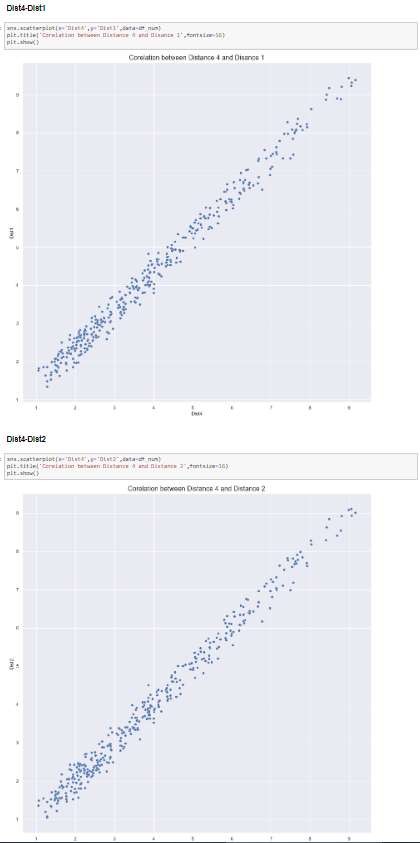


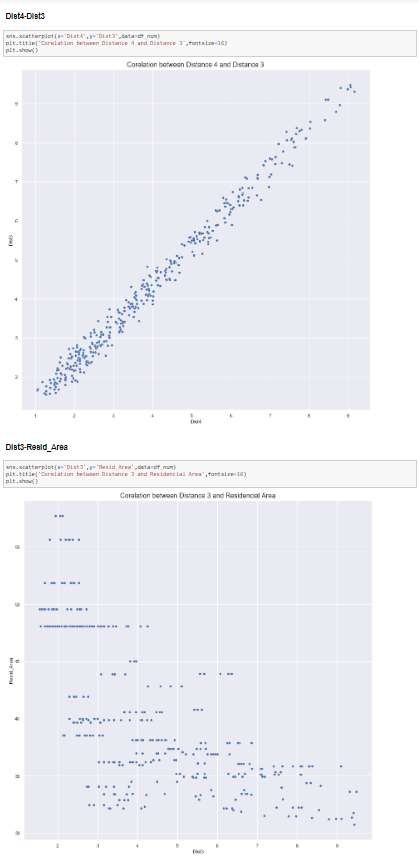


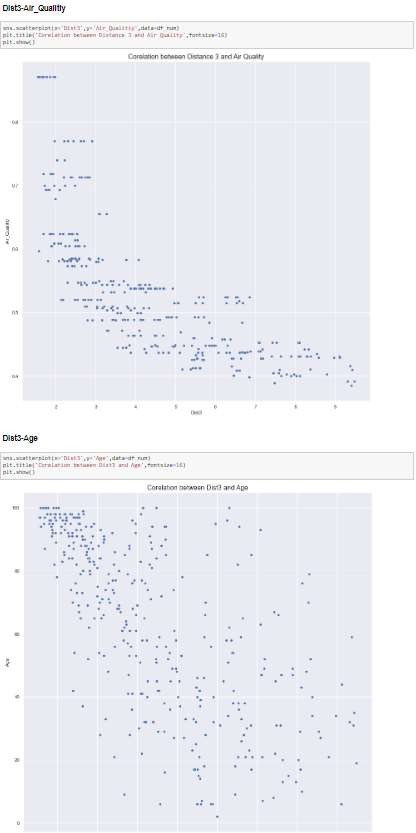


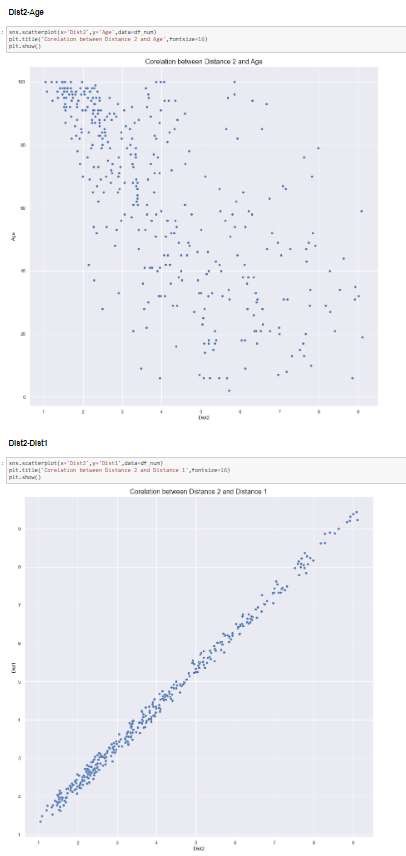












# Multivariate Analysis

Q1. What is mean by Multivariate analysis (Small Paragraph)

ANS:- Multivariate analysis extends the scope of data analysis beyond basic pairs or individual variables. Assume you are researching what influences how well students succeed. In addition to examining study hours and test results, multivariate analysis may take into account other factors including sleeping patterns, extracurricular activities, and even class difficulty. It investigates the intricate interactions between all of these factors at once by using methods like regression models. This makes it possible for researchers to spot trends and comprehend how each component affects the final result (performance). Multivariate analysis offers a comprehensive perspective, illuminating the complex network of connections present in your data.

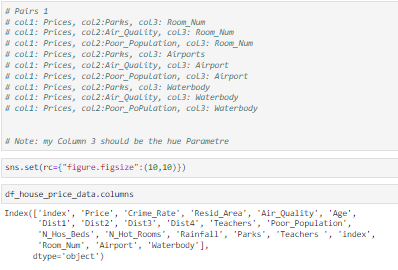
Q2. Method that we use Discribe the Multivariate analysis (Small Paragraph for each method)

ANS:-

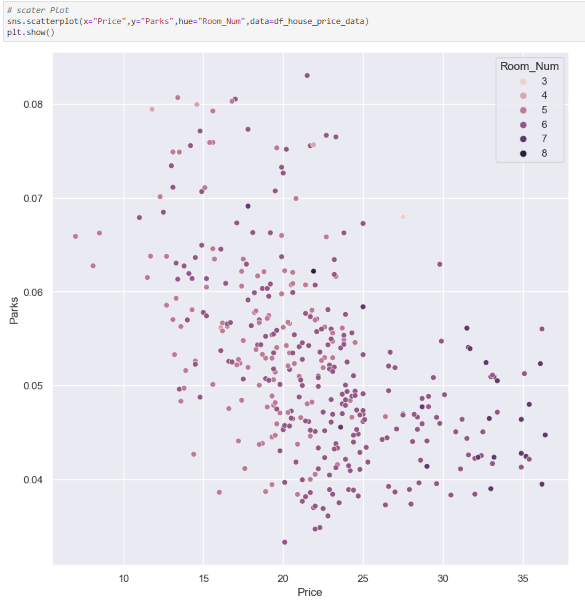
* Add hue
* Cross tab
* Pivot table

Q3. What we did With Our Data (Screeshot of code ,Screeshot of Graph , Inference) Column wise<Heading 2>

ANS:-



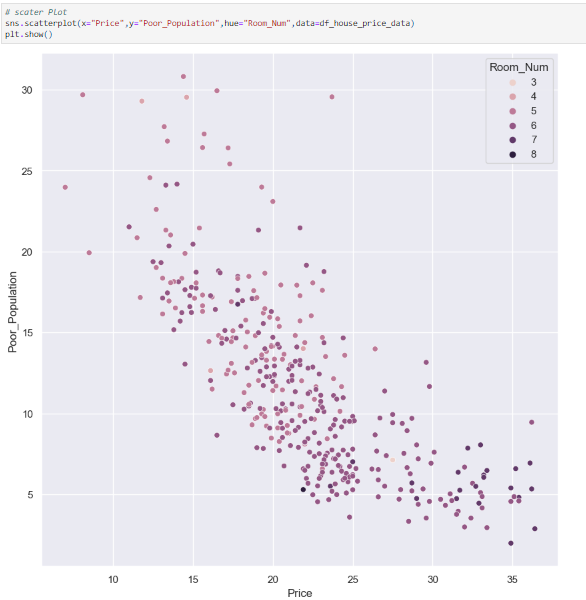
* col1: Price, col2:Parks, col3: Room\_Num:-



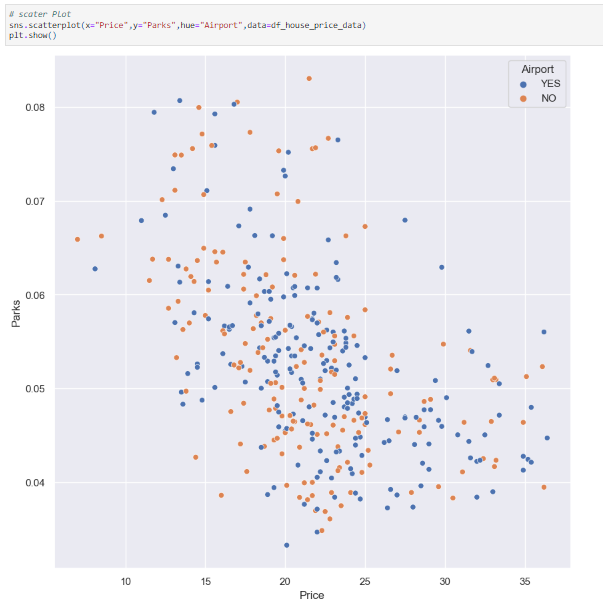
* When number of Parks high (0.08) that time the Price is low (13 lacs) the average numbers of Room\_Num is that senario is 3
* Parks when the no of parks is low (0.04) that time the price of the house is high (36 lacs) in that senario we have average room number is 8.
* col1: Prices, col2:Air\_Quality, col3: Room\_Num:-



* when the price is low(13 lacs) that time the airquality is really good then average number of room in that senario is 4
* when the price is high (36 lacs)that time the airquality is really bed 0.45 then average number of room in that senario is 8
* col1: Prices, col2:Parks, col3: Room\_Num:-



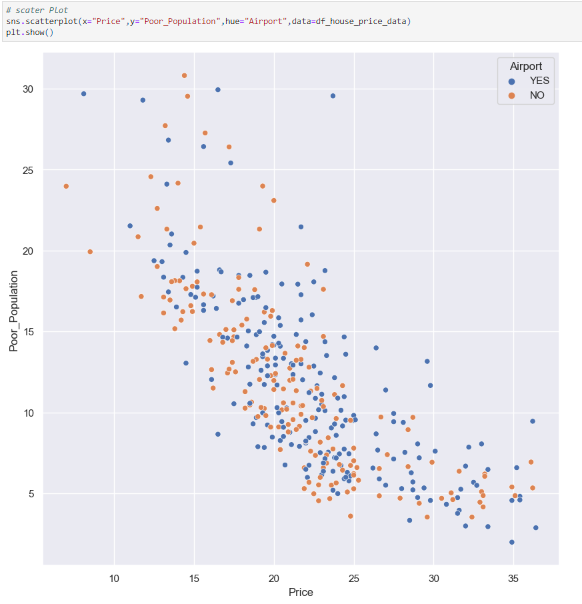
* when the price is low(8 lacs) that time the Poor\_Population is 30 then average number of room in that senario is 4
* when the price is high (36 lacs)that time the Poor\_Population is 8 then average number of room in that senario is 8
* col1: Prices, col2:Parks, col3: Airport:-



* When we have highest Price 37 lacs Then the no of Parks is 0.045 that time ther is Airports is Present.
* When we have lowest Price 17 lacs Then the no of Parks is 0.08 that time ther is Airports is Present
* col1: Prices, col2:Air\_Quality, col3: Airport



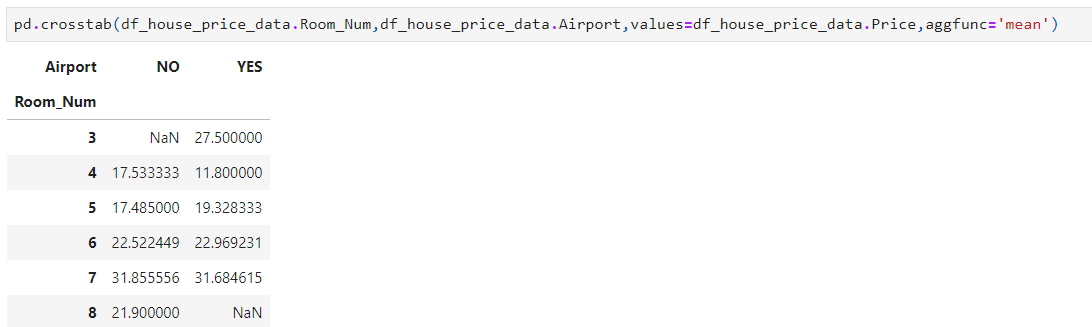
* When we have highest Price 37 lacs Then the Air\_Quality is 0.49 that time ther is Airports is Present.
* When we have lowest Price 17 lacs Then the Air\_Quality is 0.9 that time ther is Airports is Present
* col1: Prices, col2:Poor\_Population, col3: Airport



* when we have highest price (32 lacs)  then the Poor\_Population is (3) and that time we have river has waterbody.
* when we have lowest price (12 lacs)  then the Poor\_Population is (21) and that time we have river has waterbody.

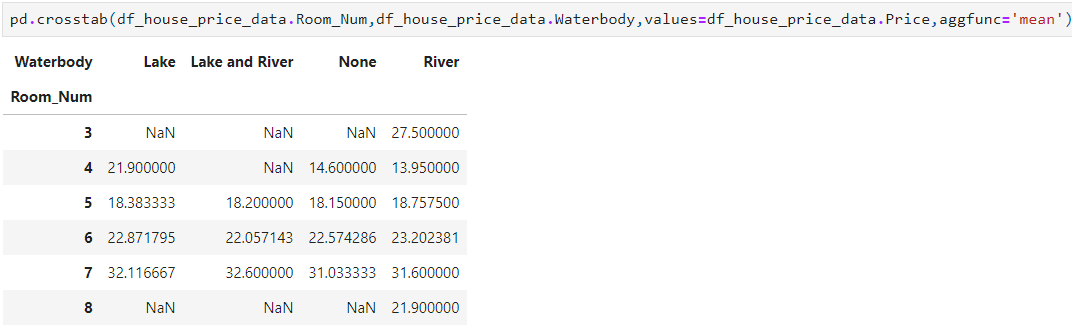
### Cross tab:-

## col1:Room\_Num, col2: Airport, col3: Price:-



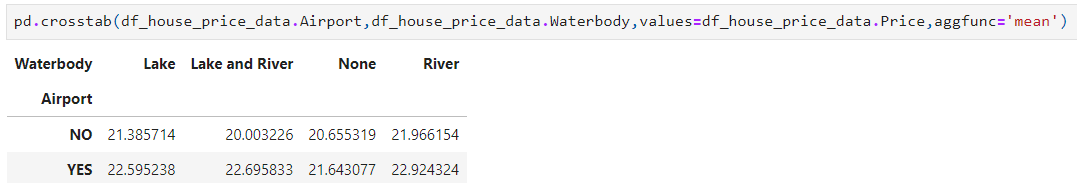
* We can observerd at average Room\_Num are 3 Airport are present and Average price is 27.5 lacs
* We can observerd at average Room\_Num are 7 Airport are present and Average price is 31.68 lacs
* We can observerd at average Room\_Num are 7 Airport arenot present and Average price is 31.85 lacs
* we can observed at average Room\_Num are 8 Airport are not present and avrage price is 21.9 lacs

## col1:Room\_Num, col2: Waterbody, col3: Price:-



* We can observed Average room numbers 3 and waterbody as river present at House price is 27.5 lacs
* we can observed Average room numbers 5 and waterbody as river, Lake and River, river present at house average house price is 18.37 lacs
* we can observed Average room numbers 7 and waterbody as river, Lake and River, river present at house average house price is 31.25 lacs
* We can observed Average room numbers 8 and waterbody as river present at House price is 21.9 lacs

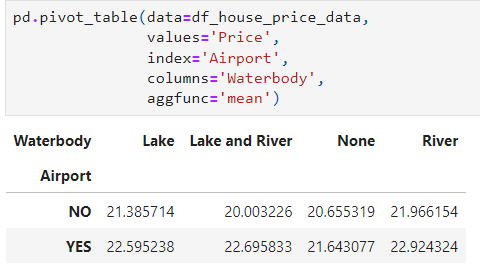
## col1:Airport, col2: Waterbody, col3: Price:-



* We can observed At airport is not present but lake as waterbody is present at average house price is 21.38 lacs
* We can observed At airport is not present but lake and river as waterbody is present at average house price is 20 lacs
* We can observed At airport is not present but River as waterbody is present at average house price is 21.96 lacs
* we can observed at airport is present and waterbody as lake and river present at average house price is 22.69 lacs
* We can observed At airport is present but Lake as waterbody is present at average house price is 22.59 lacs

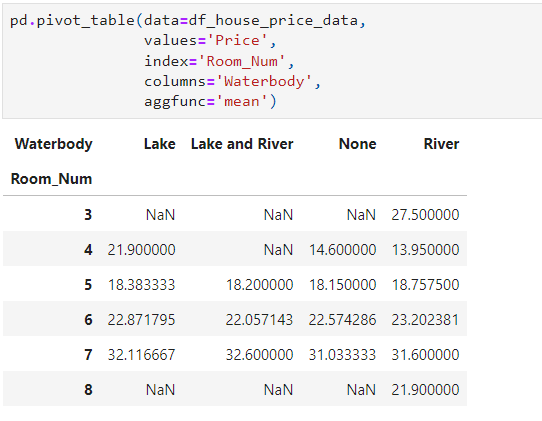
### Pivot\_table:-

## col1:Airport, col2: Waterbody, col3: Price:-



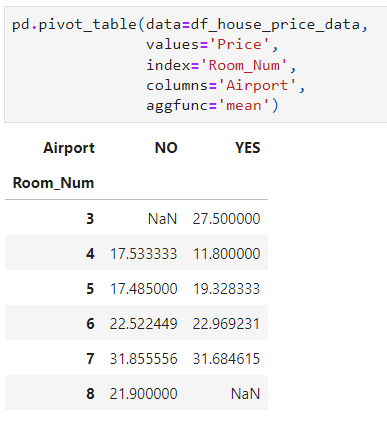
* We can observed At airport is not present but lake as waterbody is present at average house price is 21.38 lacs
* We can observed At airport is not present but lake and river as waterbody is present at average house price is 20 lacs
* We can observed At airport is not present but River as waterbody is present at average house price is 21.96 lacs
* we can observed at airport is present and waterbody as lake and river present at average house price is 22.69 lacs
* We can observed At airport is present but Lake as waterbody is present at average house price is 22.59 lacs

### col1:Room\_Num, col2: Waterbody, col3: Price:-



* We can observed At airport is not present but lake as waterbody is present at average house price is 21.38 lacs
* We can observed At airport is not present but lake and river as waterbody is present at average house price is 20 lacs
* We can observed At airport is not present but River as waterbody is present at average house price is 21.96 lacs
* we can observed at airport is present and waterbody as lake and river present at average house price is 22.69 lacs
* We can observed At airport is present but Lake as waterbody is present at average house price is 22.59 lacs

## col1:Room\_Num, col2: Airporat, col3: Price:-



* We can observerd at average Room\_Num are 3 Airport are present and Average price is 27.5 lacs
* We can observerd at average Room\_Num are 7 Airport are present and Average price is 31.68 lacs
* We can observerd at average Room\_Num are 7 Airport arenot present and Average price is 31.85 lacs
* we can observed at average Room\_Num are 8 Airport are not present and avrage price is 21.9 lacs

Summary:-

The dataset provides information on house prices for builders, encompassing 506 observations with 19 characteristics. It includes 16 numerical attributes and 3 categorical attributes. The data includes valid missing values, outliers in crime rate and age, and an imbalance in room numbers. House prices range from 5 lakhs to 50 lakhs, with a median price of 22 lakhs. Crime rates vary between 0.006 and 20, with an average of 3.59. Residential areas range from 30 to 58 square meters, with an average of 41 square meters. Air quality varies between 0.38 and 0.8, with an average of 0.55. House age varies significantly, with an average of 68 years. Distances to industrial hubs range from 1 to 12 kilometers. The number of teachers varies, with an average of 21. Variations in poor population, hospital beds, hotel rooms, rainfall, and park area are also observed. Areas with higher park areas have fewer rooms on average, while those with lower park areas have more rooms.

Table of Content

|  |  |  |
| --- | --- | --- |
| SR.NO | CONTENT | PAGE.NO |
| 1. | Objective | 1 |
| 2. | Introduction | 1 |
| 3. | Methodology | 1 |
| 4. | Data Overview | 1 |
| 5. | Data Preprocessing | 4 |
| 5.1 | Data Cleaning | 4 |
| 5.2 | Missing Value Treatment | 6 |
| 5.3 | Outlier Treatment | 8 |
| 6. | EDA | 10 |
| 6.1 | Univariate Analysis | 11 |
| 6.2 | Bivariate Analysis | 32 |
| 6.3 | Multivariate Analysis | 49 |
| 6.3.1 | Add Hue Element | 49 |
| 6.3.2 | Cross tab | 56 |
| 6.3.3 | Pivot element | 57 |
| 7. | Summary | 59 |