|  |  |  |
| --- | --- | --- |
| SR NO. | CONTENT | PAGE NO. |
| 1 | objective |  |
| 2 | Introduction |  |
| 3 | Methodology |  |
| 4 | Data Overview |  |
| 4.1 | Data Cleaning |  |
| 4.2 | Missing Value Treatment |  |
| 4.3 | Outlier Treatment |  |
| 5 | EDA |  |
| 5.1 | Univariate Analysis |  |
| 5.2 | Bivariate Analysis |  |
| 5.3 | Multivariate Analysis |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Objective:

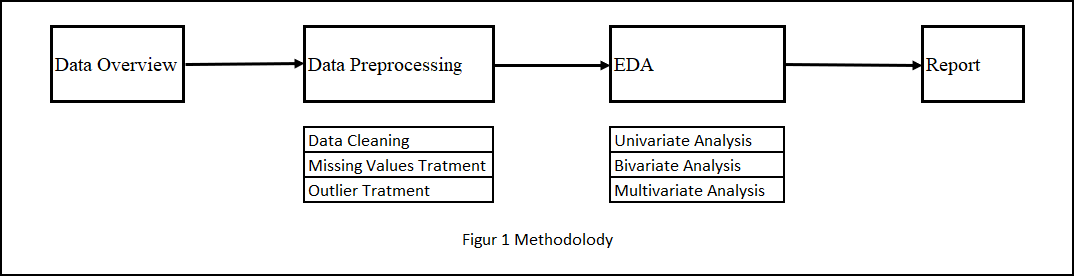
Report

In this project we need to analyses the data, the data belongs to a builder community they have historical data related to existing houses.

## Introduction:

Me **RIVYA MUKESHBHAI TANTI** responsible to commuted the analysis on housing related data in this data we are following one flow (core analytics flow). This analysis is beneficial for multiple builder era to conclude the prices of the houses, also its is important in terms of building the upcoming policies and deciding the factors responsible for understanding the price of houses. The motivation of the analysing the data is came from standard “manufacturing analytics rules for analysis”, this standard utilized in standardizing the null values. As per my knowledge and under the guidance of **VINAYAK BHOYAR SIR**, i am goanna conclude this analysis as the standard followers.

## Methodology:

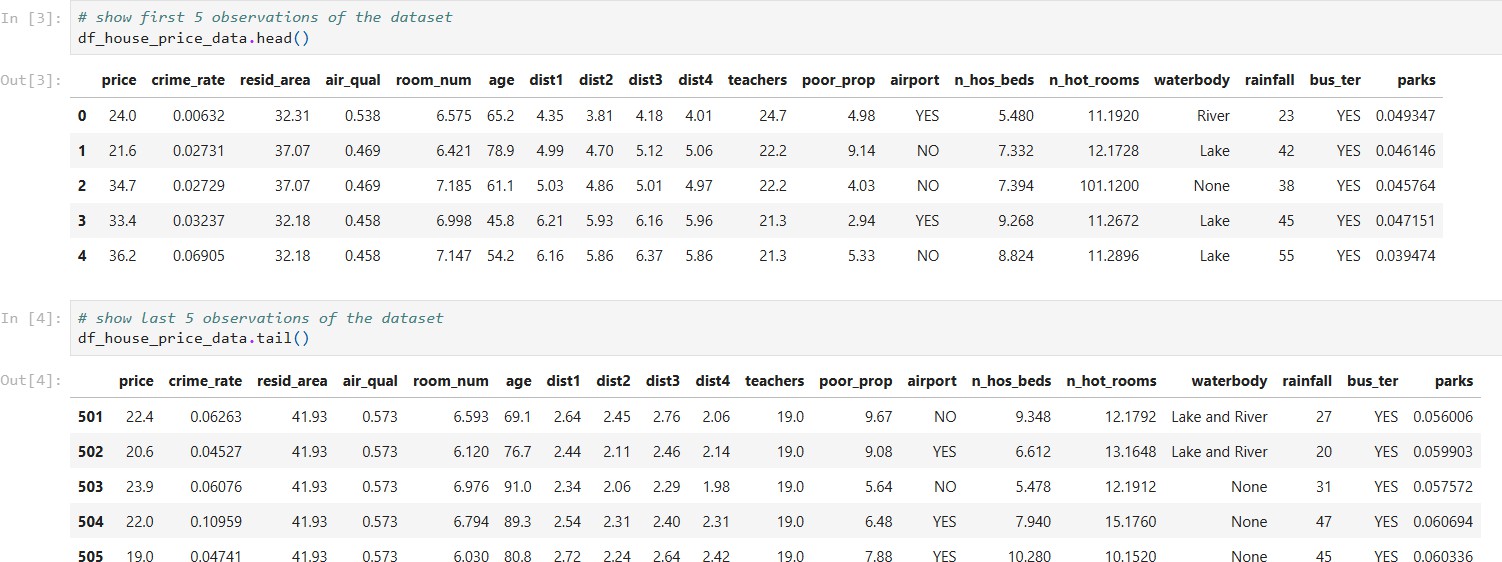


Data overview:

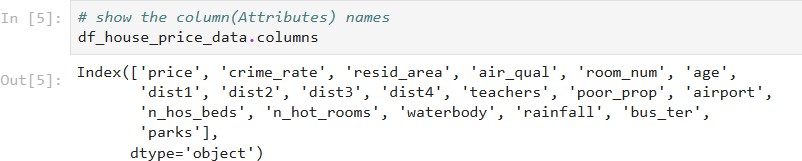
#### Q1. collect the detail information about the data?

**Answer:** The process of gathering and analysing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is Known as Data Collection.

This dataset has 506 rows and 19 columns in which we have different factors(attributes) affecting the price of house like airport, waterbody,room no etc



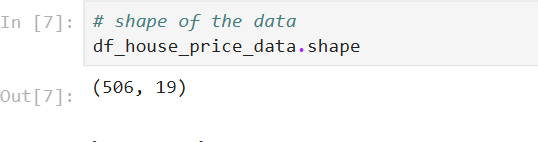
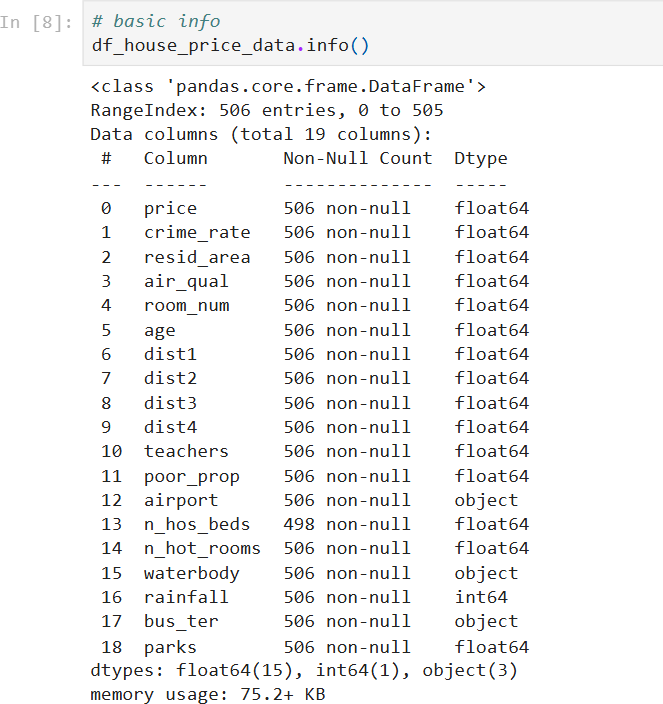
#### Q2. collect the detail information about the column/attributes? Answer:



* price : price of the house.
* crime\_rate : crime rate in that area
* resid\_area : residential of that area
* air\_qual : air quality of that area
* room\_num : residential of that area
* age : How old that house is
* dist1 : distance of area from house for employment 1
* dist2 : distance of area from house for employment 2
* dist3 : distance of area from house for employment 3
* dist4 : distance of area from house for employment 4
* teachers : Numbers of schools near residence
* poor\_prop : Poor population rate near the residental area
* airport : any airport near the residence
* n\_hos\_beds : hospital beds numbers
* n\_hot\_rooms : Number of hotel rooms
* waterbody : source of water near residence
* rainfall : reainfall rate in that area
* bus\_ter : bus terminal near the residence
* parks : Any park near the residence

#### Q3. show the basic statics of the data?

**Answer:** (1) Shape of the data

1. basic info
2. basic description
3. Shape of the data
   * We have 506 observation and 19 attributes
4. Basic info
   * We have 16 numerical attributes it includes 15 floats and 1 integer in nature.
   * We have 3 categorical attributes
   * In the data we have valid null values
5. Basic description



* + We have outliers is crime\_rate, resid\_area, age

## Data preprocessing

In the preprocessing we have few stages that we need to follow , those stages we can see in side figur1

1. data cleaning
2. Missing values treatment
3. Outlier treatment

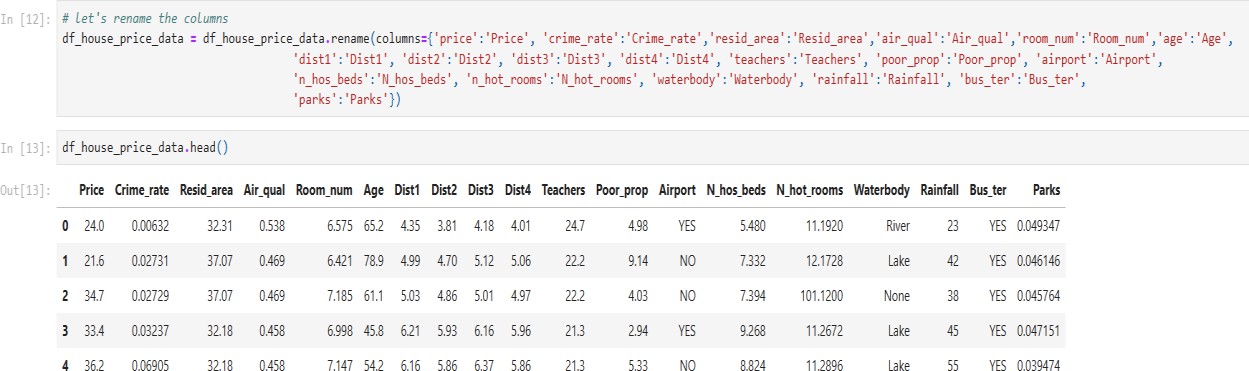
# Data cleaning

**Q1. what is data cleaning?**

**Answer:** 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.

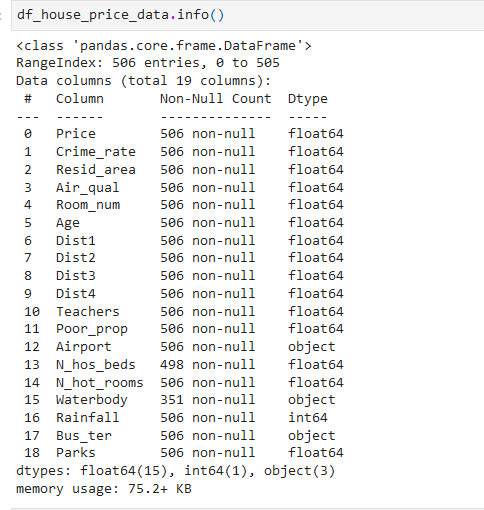
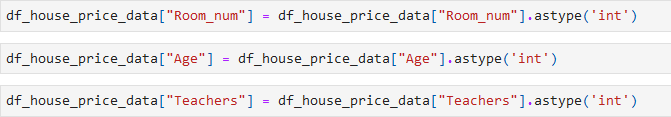
**Q2. what we did on out data?**

#### Answer:

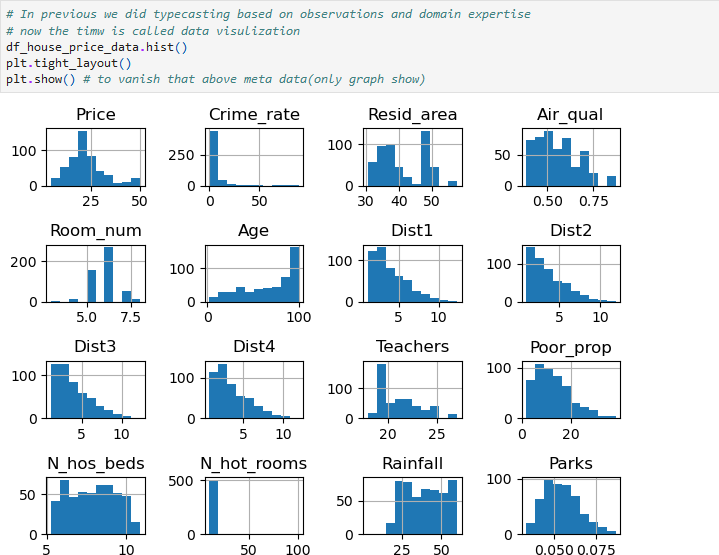


* + We Renamed the attributes of our dataset for better Clarity, Readability, Consistency and Error Reduction.
  + As we can see that ‘price’ to ‘Price’, ‘air\_qual’ to ‘Air\_Quality’ and so on.

Typecasting the columns Room\_num, Age, Teachers, N\_hot\_rooms.



- Visualised the Data using Hist Map.



# Missing value treatment

#### Q1. what are the types of missing values?

**Answer:** There are 2 types of the missing value as described below:

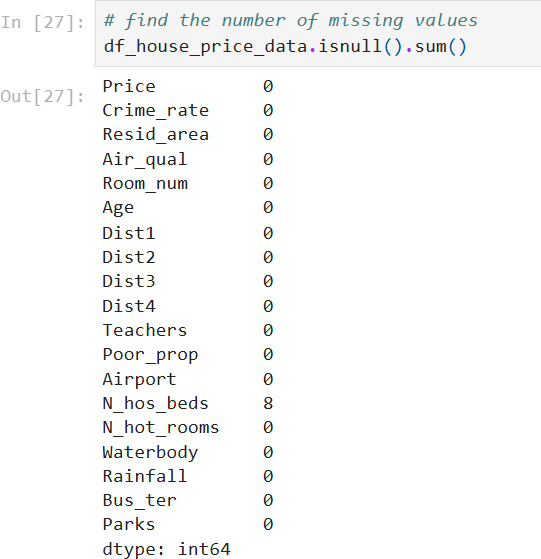
1. Valid Missing Value: When there is empty space instead of data it is called Valid Missing Value
2. Invalid Missing Value: When there is any other value which is irrelevant to the actual data value in case of when we don’t have actual value it is called Invalid Missing Value

#### Q2. Rules of missing values as per manufacturing industry.

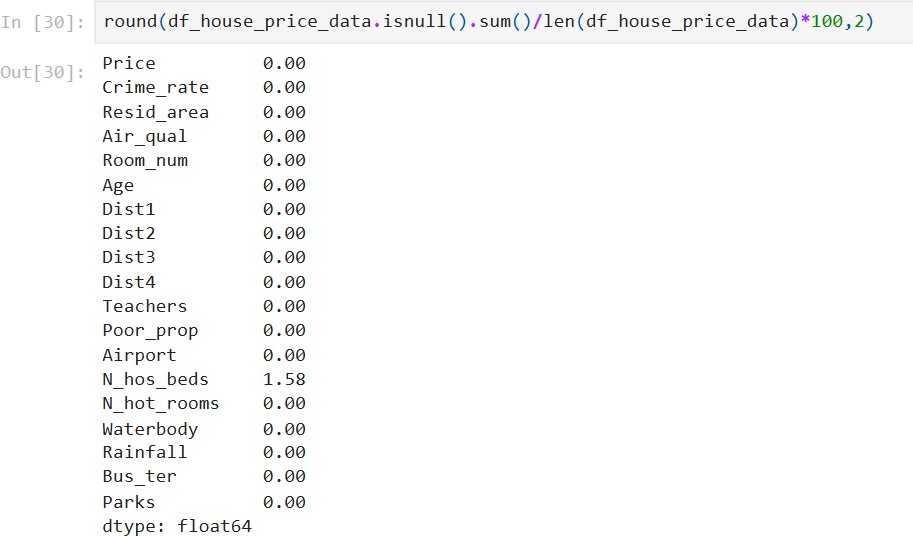
**Answer:** There are 3 Rules of missing values as per manufacture industry namely,

1. When we have less than 10% data is missing, remove the observation
2. When we have missing % values 30 to 70% missing values than we need to inpute the missing values
3. When we have more than 80% data is missing, remove the attributes

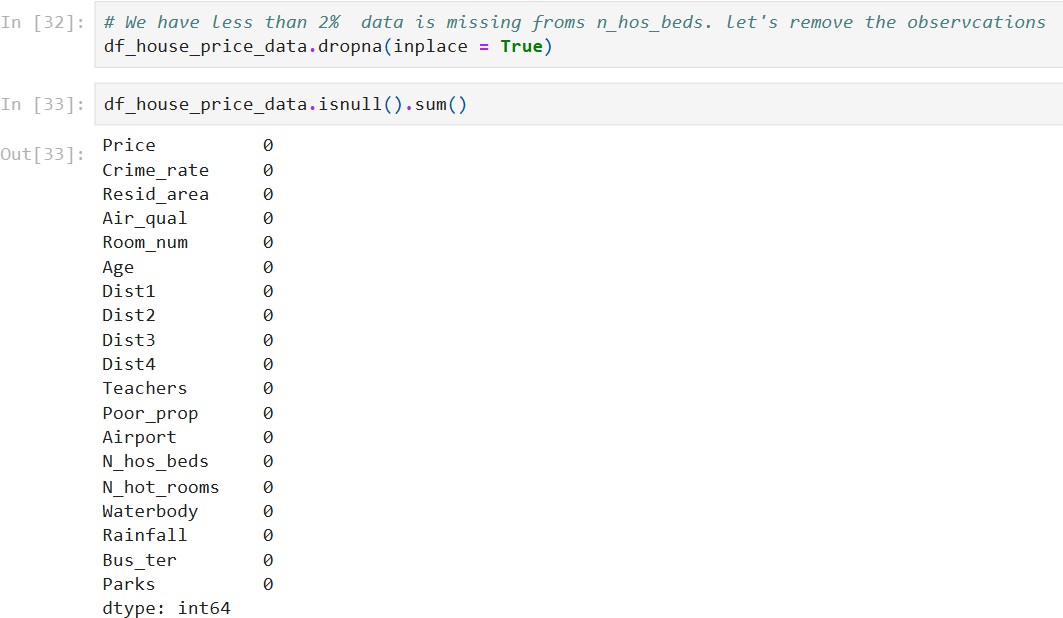
#### Q3. what we did on our data? Answer:



* + We have missing values in number of hospital beds.



* + We have less than 2% data is missing forms Numbers of hospital beds. Then remove the observations



# Outlier treatment

#### Q1. what is mean by outlier?

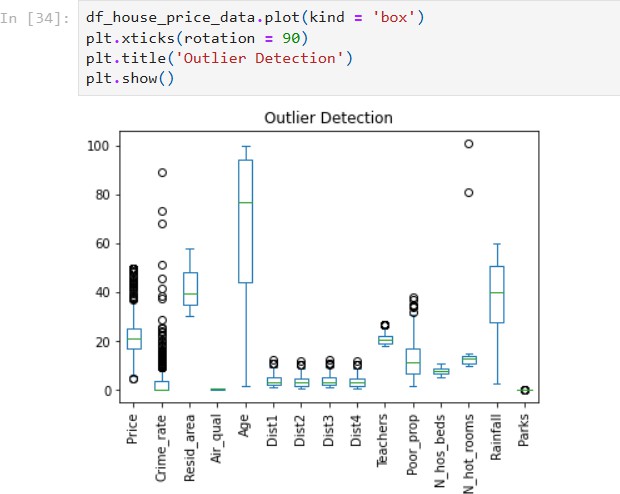
**Answer:** An Outlier is the datapoint that goes far outside the average value of a group of statistics. Outliers may be exceptions that stand outside individual samples of populations as well. In conclusion Outlier is an individual that is markedly different from the norm in some respect.

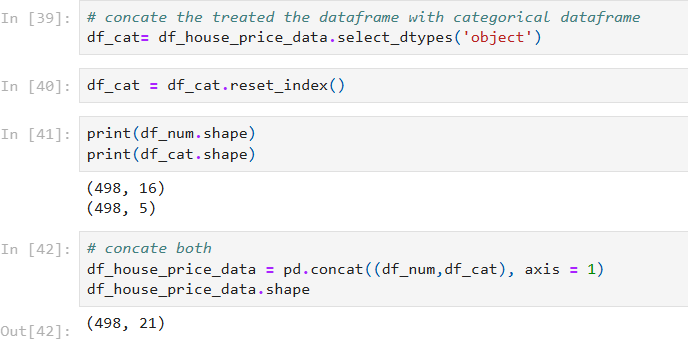
#### Q2. which method we you to treat method?

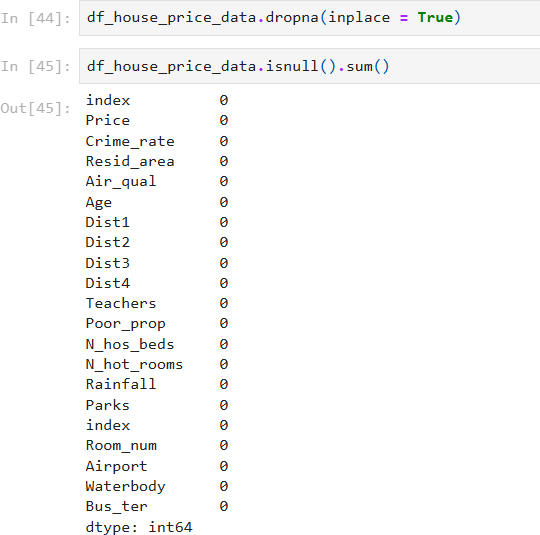
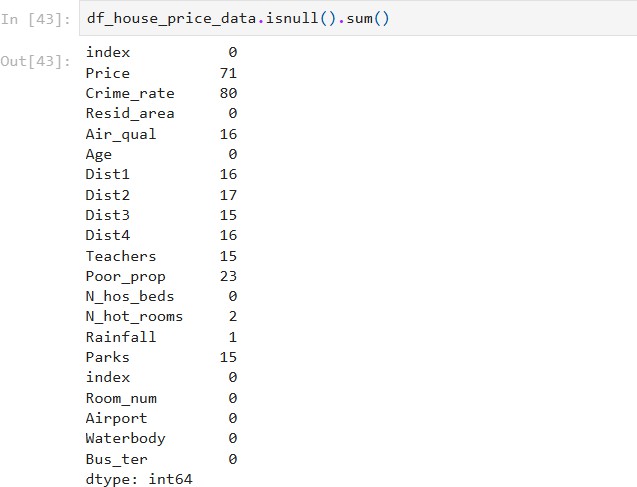
**Answer:** We used the Interquartile Range (IQR) Method to treat the outlier. IQR is the range of the values that resides in the middle of the scores.

When a distribution is skewed, and the median is used instead of the mean to show a central tendency, the appropriate measure of variability is the inter quantile range.

**Q3. what we did on our data? Answer:**







## Exploratory Data Analysis

It is a method that we utilise to explore unseen data, it is an process of converting any row data (unseen data) into actionable inside / information.

#### Q1. explain the types of the EDA.

**Answer:** EDA into three types: (i) Univariate Analysis, (ii) Bivariate Analysis

(iii) Multivariate Analysis

1. Univariate Analysis: Univariate analysis focuses on a single variable to understand its internal structure. It is primarily concerned with describing the data and finding patterns existing in a single feature.
2. Bivariate Analysis: Bivariate evaluation involves exploring the connection between variables. It enables find associations, correlations, and dependencies between pairs of variables. Bivariate analysis is a crucial form of exploratory data analysis that examines the relationship between two variables.
3. Multivariate Analysis: Multivariate analysis examines the relationships between two or more variables in the dataset. It aims to understand how variables interact with one another, which is crucial for most statistical modelling techniques.

# Univariate Analysis

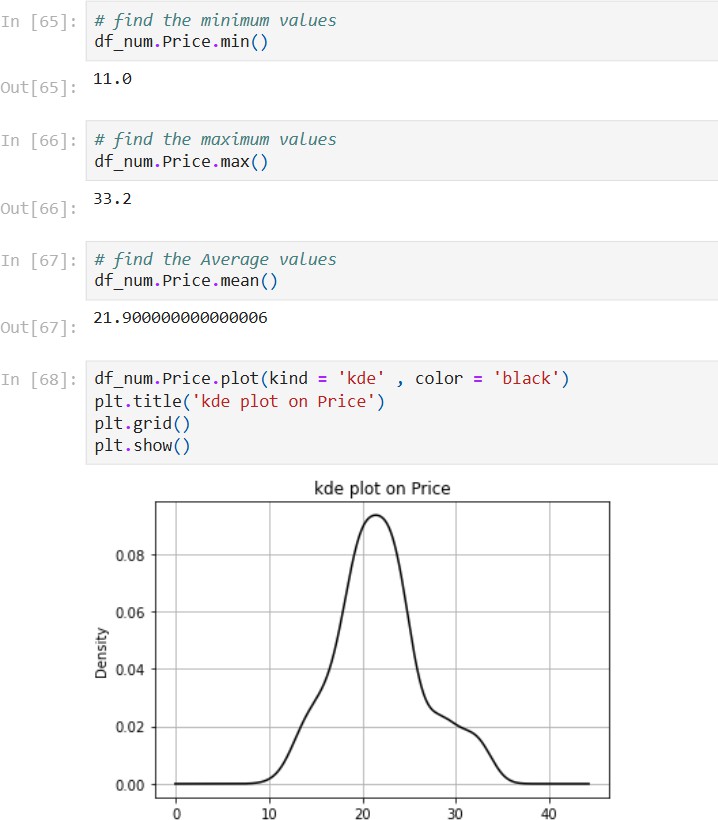
#### Q1. what is mean by univariate analysis?

**Answer:** Univariate analysis focuses on a single variable to understand its internal structure. It is primarily concerned with describing the data and finding patterns existing in a single feature. This sort of evaluation makes a speciality of analysing character variables inside the records set. It involves summarizing and visualizing a unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records.

**Q2. what we did with our data?**

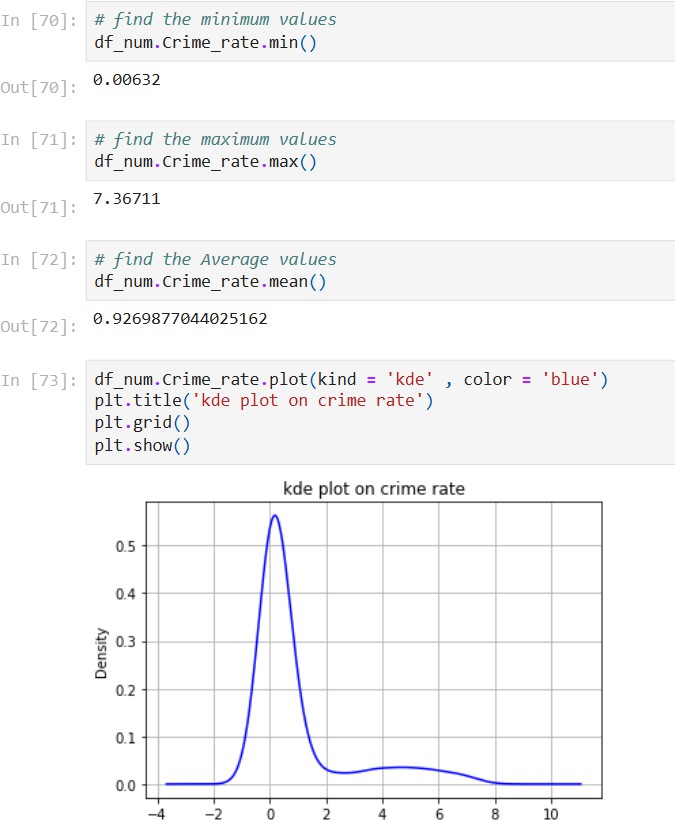
# **Answer:** Univariate analysis on numerical data

Price:



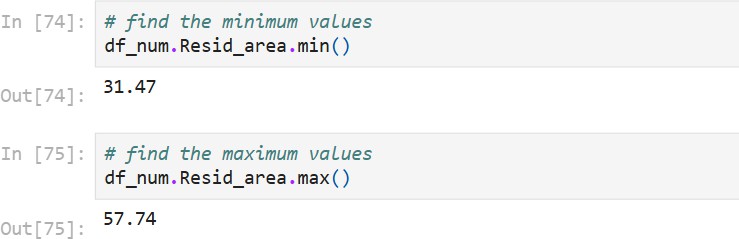
* This data is slightly normal distributed
* The houses we have their prices lies in 8 lac to 40 lacs
* The House we have the least prices 11 lakh
* The House we have the maximum prices 33 lakh
* The Average House with Average prices 21 lakh

# Crime\_rate:



* This data is slightly normal distributed
* This lies in 0.1 to 0.2%
* We have the least crime rate 0.00632 %
* We have the maximum crime rate 88.9762 %
* We have the Average crime rate 3.5889 %

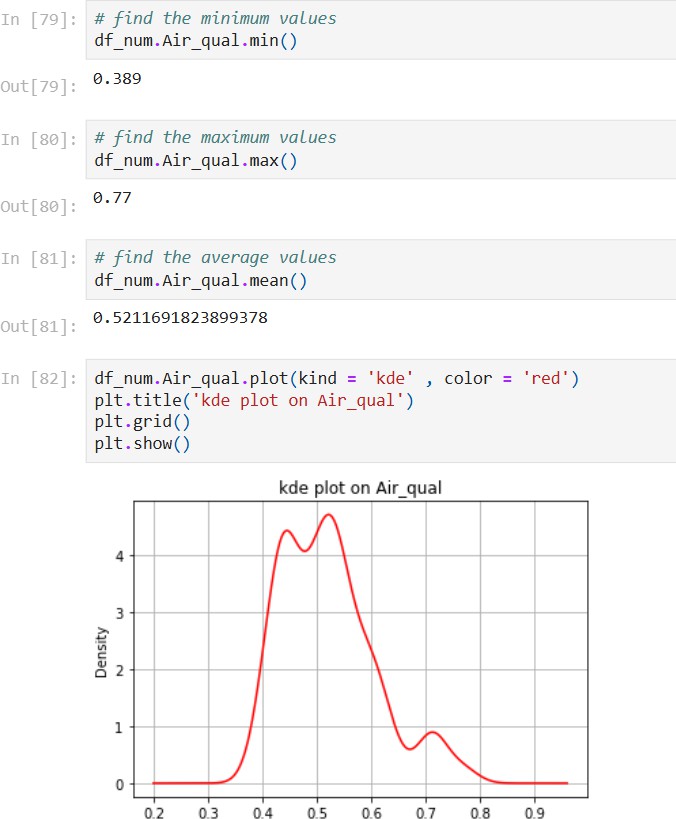
# Resid\_area:





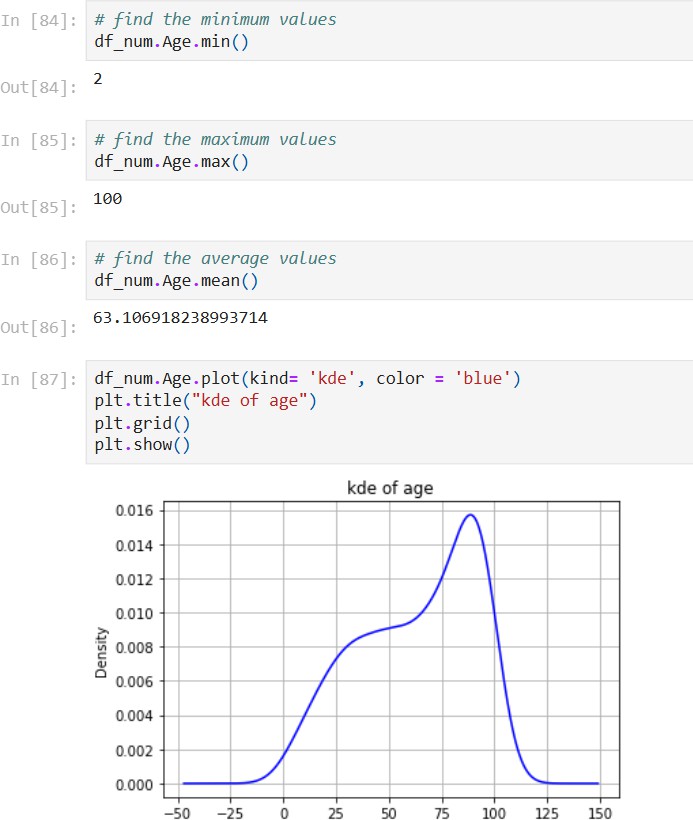
* This data is multi distributed
* First model lies between 30 to 40 and second model lies between 47 to 55
* We have minimum values 31.47
* We have maximum values 57.74
* We have Average values 40.194

# Air\_qual:



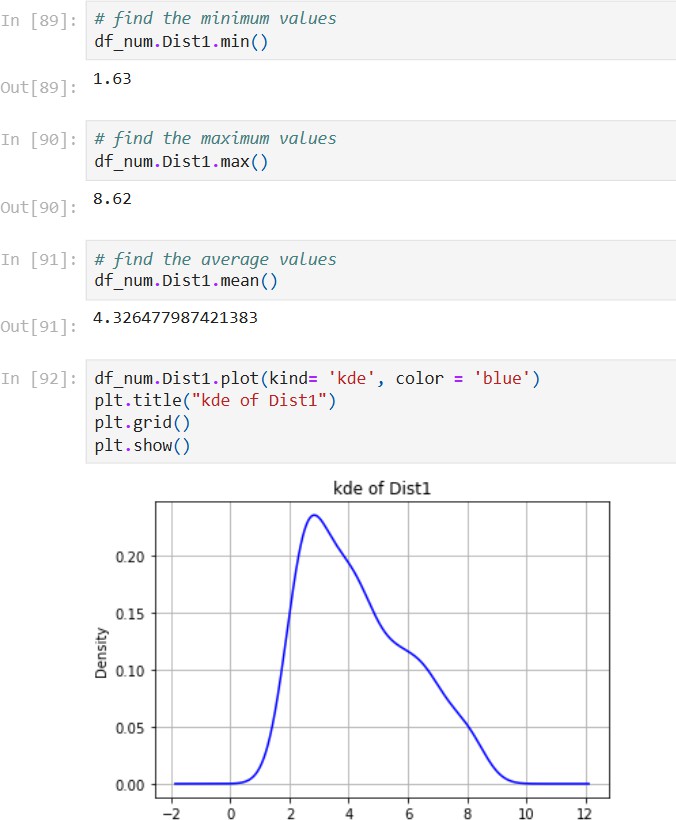
* This data is not normally distributed
* This lies in 0.38 to 0.77%
* We have minimum values 0.389
* We have maximum values 0.77
* We have Average values 0.5211

# Age:



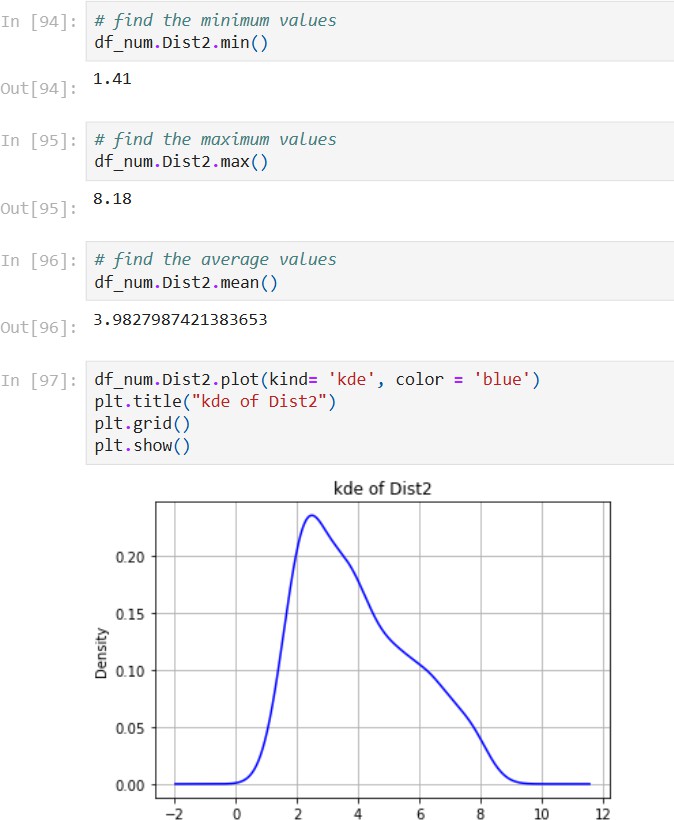
* This data is left distributed
* This lies in 0 to 122
* We have minimum values 2
* We have maximum values 100
* We have Average values 63.1069

# Dist1:



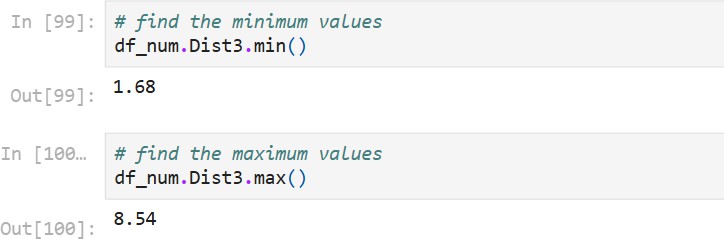
* This data is normally distributed
* This lies in 1 to9
* We have minimum values 1.63
* We have maximum values 8.62
* We have Average values 4.326

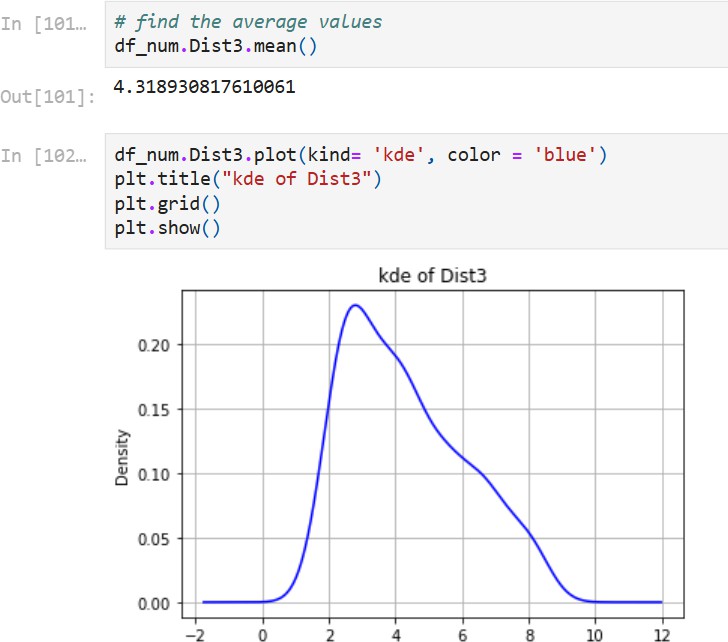
# Dist2:



* This data is not normally distributed
* This lies in 0.6 to 9
* We have minimum values 1.41
* We have maximum values 8.18
* We have Average values 3.9827

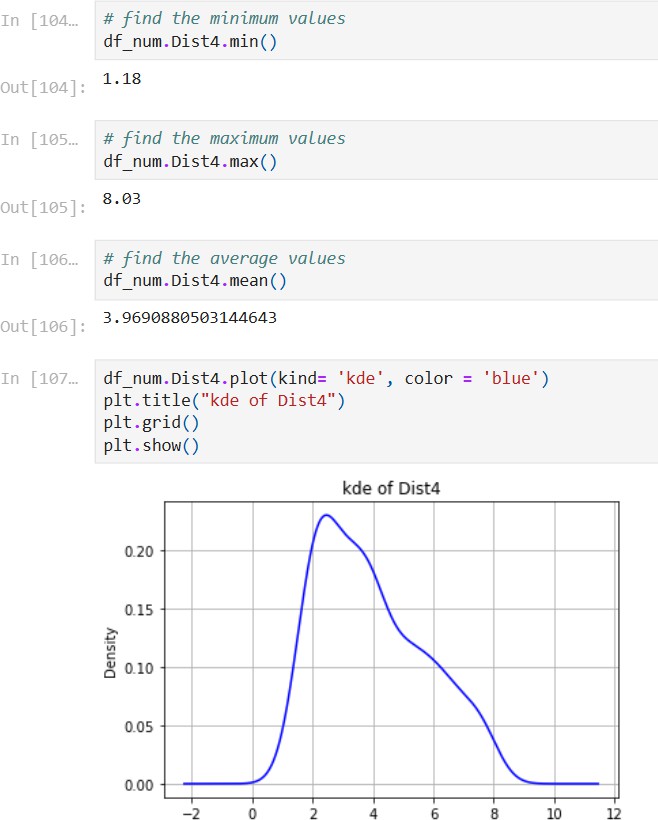
# Dist3:





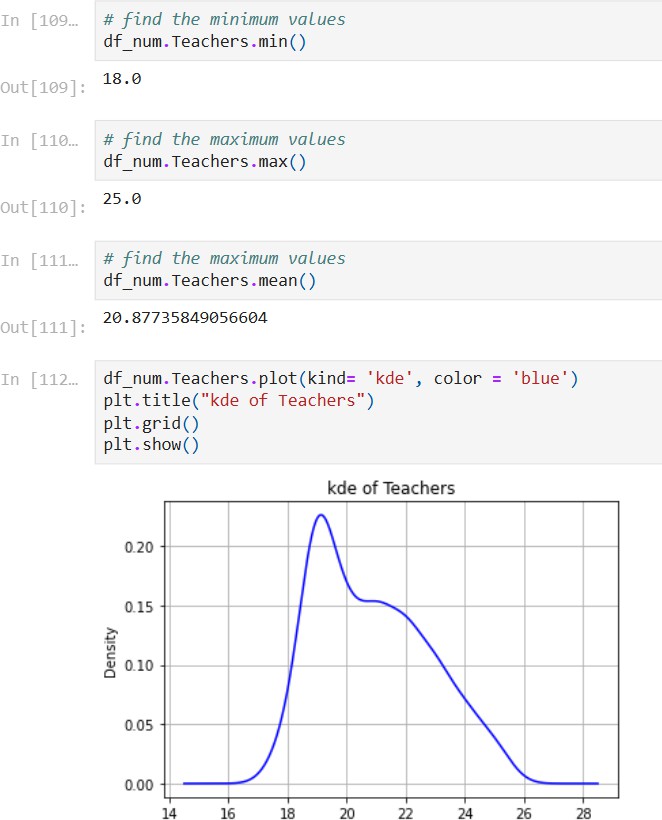
* This data is not normally distributed
* This lies in 0.6 to 9
* We have minimum values 1.68
* We have maximum values 8.54
* We have Average values 4.3189

# Dist4:



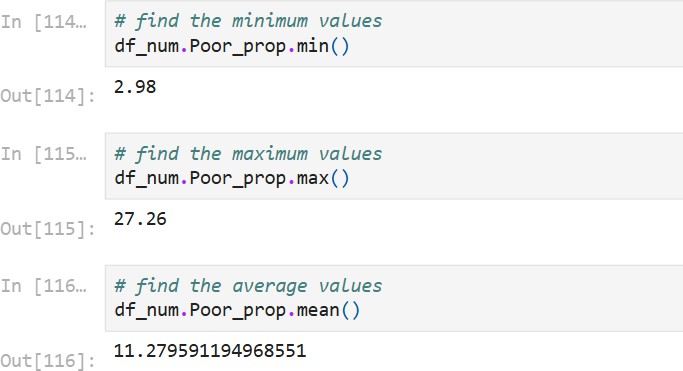
* This data is not normally distributed
* This lies in 0.6 to 9
* We have minimum values 1.18
* We have maximum values 8.03
* We have Average values 3.9690

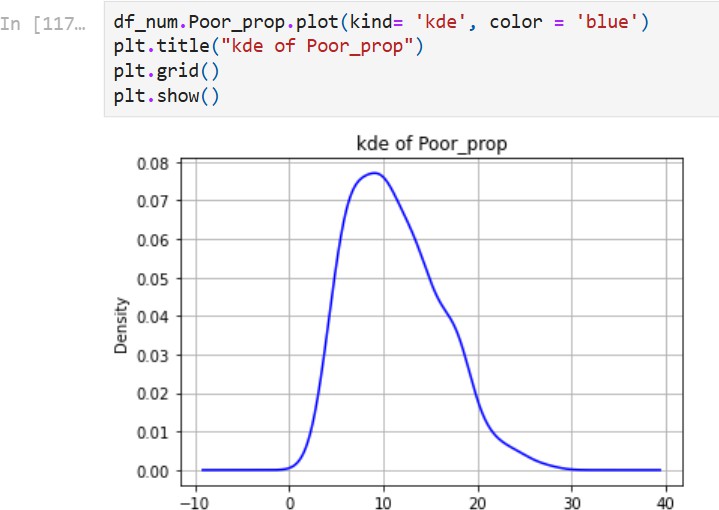
# Teachers:



* This data is not normally distributed
* This lies in 17 to 25
* We have minimum values 18
* We have maximum values 27
* We have Average values 20.8773

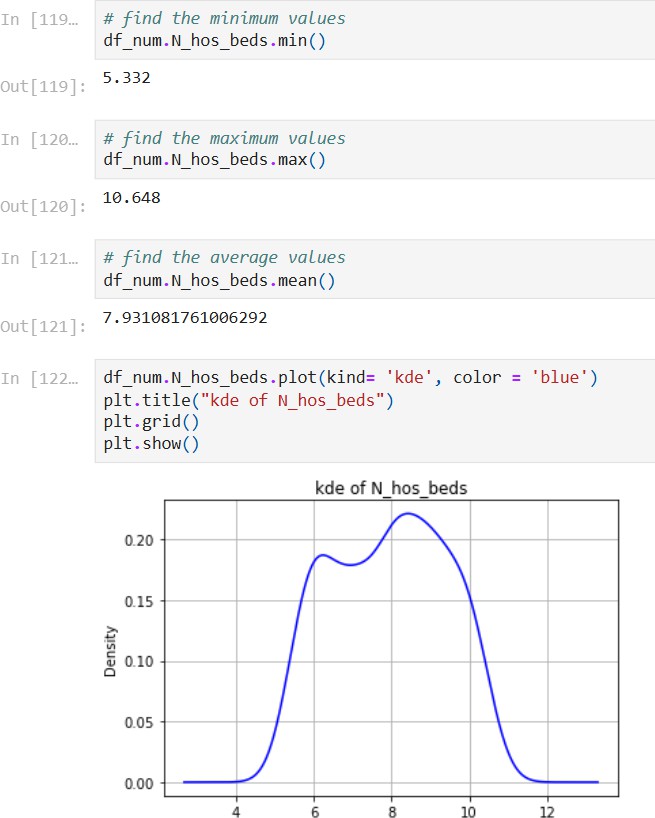
# Poor\_prop:





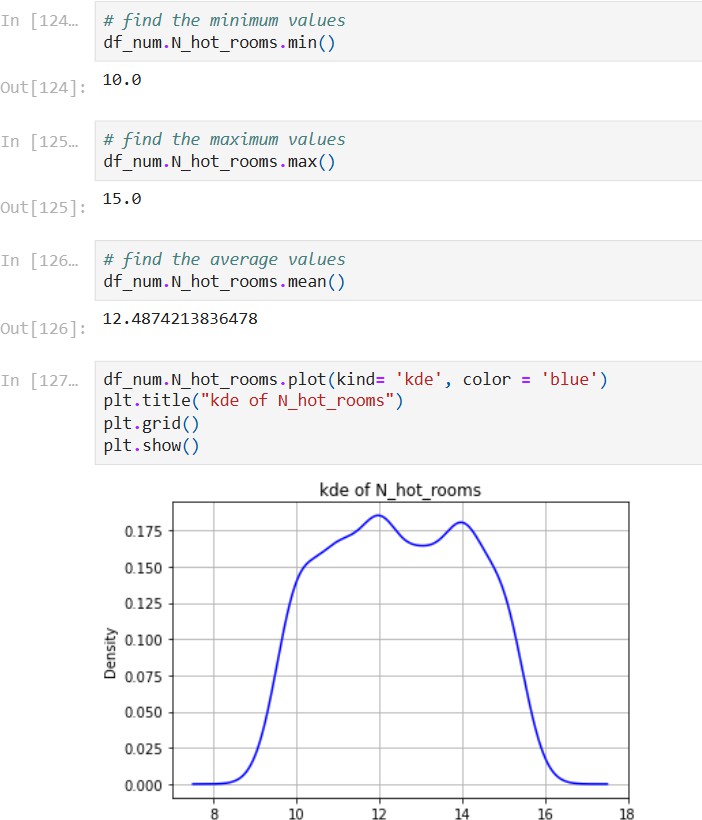
* This data is not normally distributed
* This lies in 1 to 22
* We have minimum values 2.98
* We have maximum values 27.26
* We have Average values 11.2795

# N\_hos\_beds:



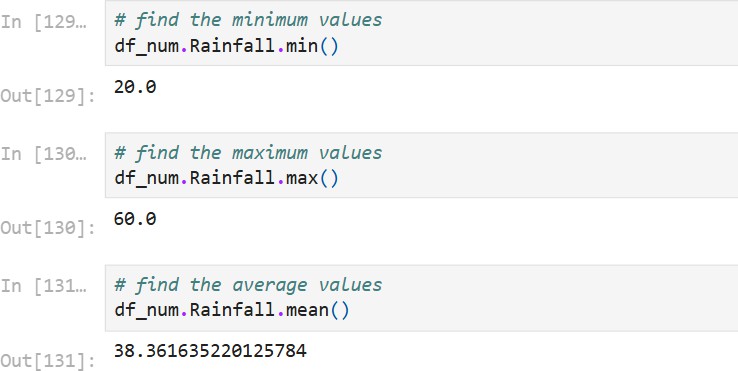
* This data is not normally distributed
* This lies in 5 to 11
* We have minimum values 5.332
* We have maximum values 10.648
* We have Average values 7.9310

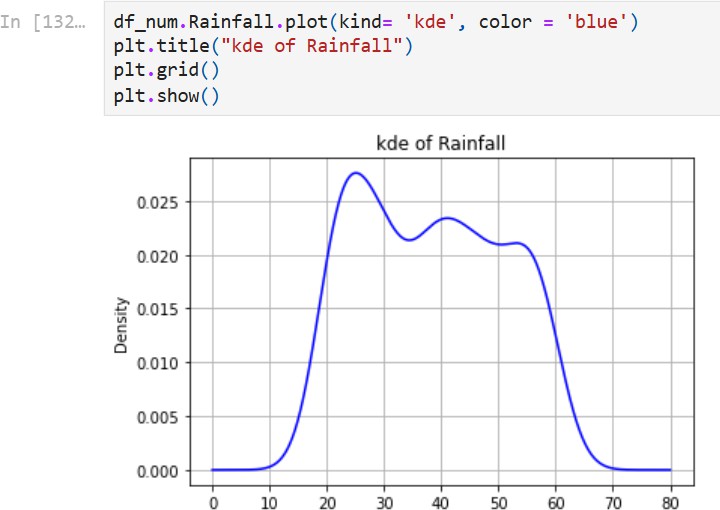
# N\_hot\_rooms:



* This data is not normally distributed
* This lies in 9 to 16
* We have minimum values 10
* We have maximum values 15
* We have Average values 12.4874

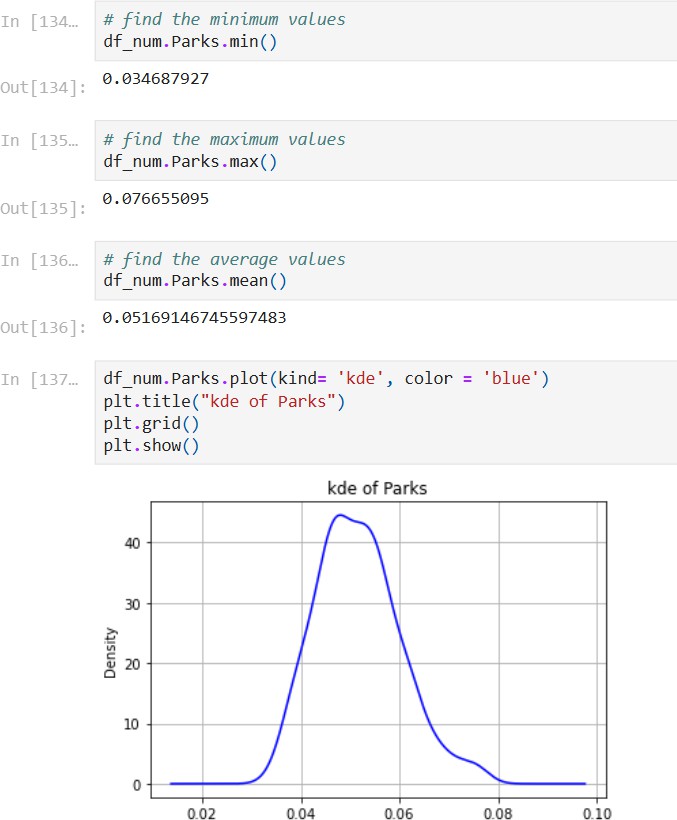
# Rainfall:





* This data is not normally distributed
* This lies in 11 to 65
* We have minimum values 20
* We have maximum values 60
* We have Average values 38.361

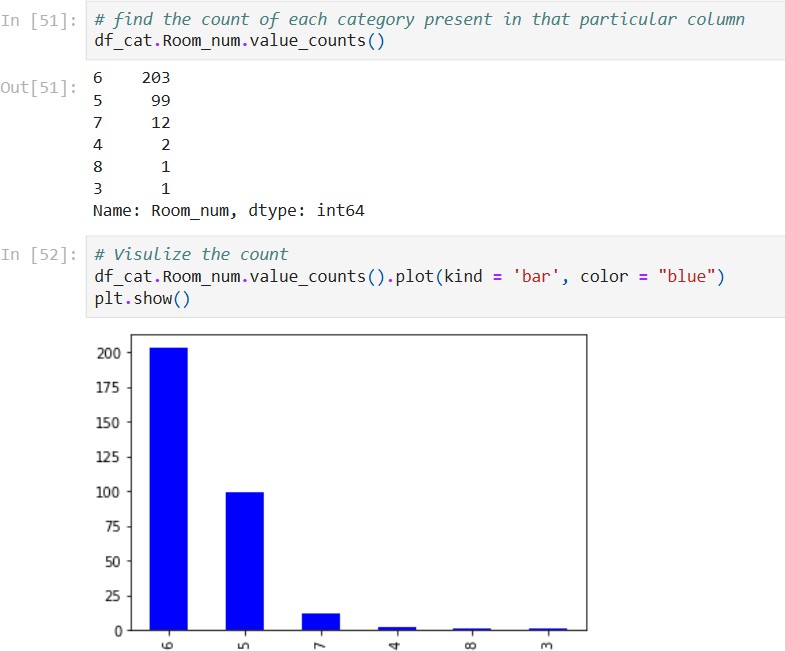
# Parks:



* This data is not normally distributed
* This lies in 0.03 to 0.07
* We have minimum values 0.0346
* We have maximum values 0.0766
* We have Average values 0.0516

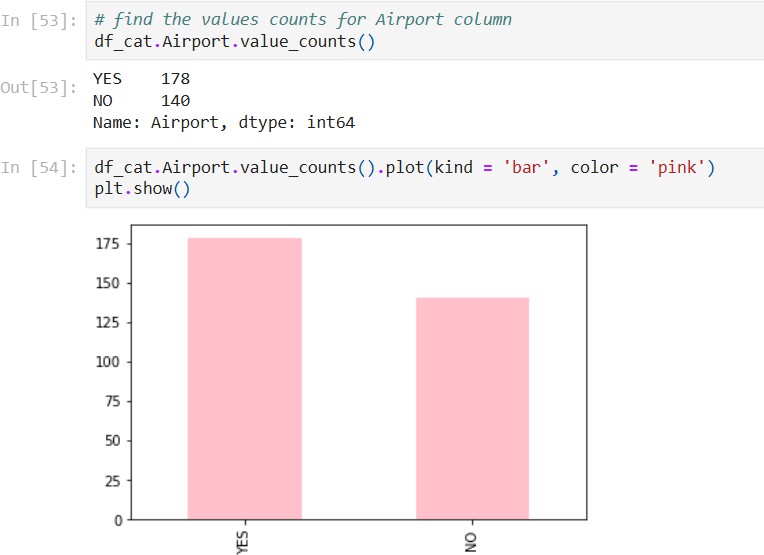
# Univariate analysis on categorical data

Room Number:



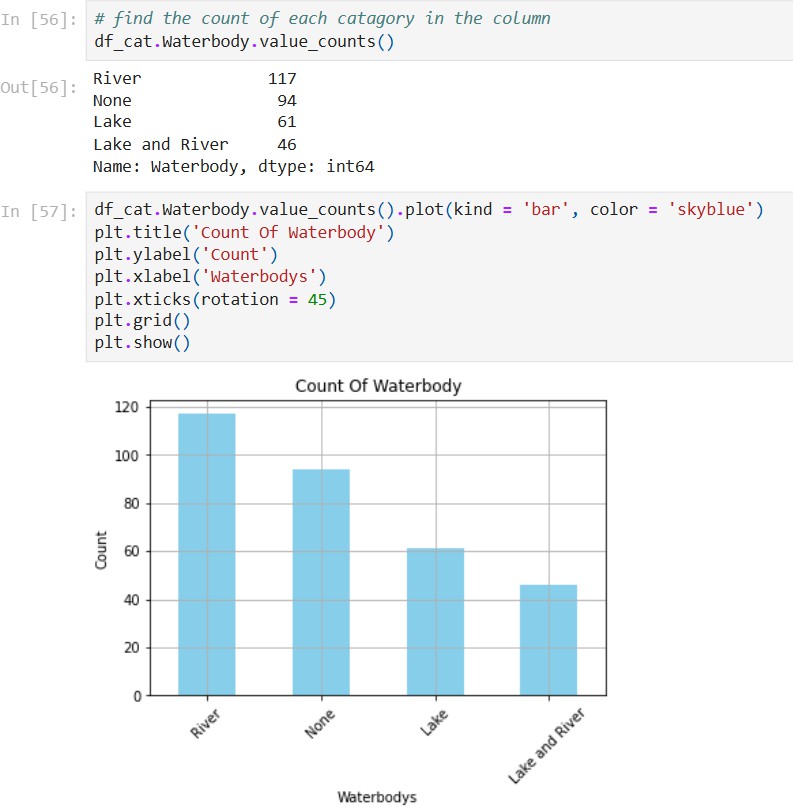
* We have imbalance data in the Room\_num column
* 6 is a category who is dominating most, its count is 266
* 3 is the category having least number of observations, its count is 2

# Airport:



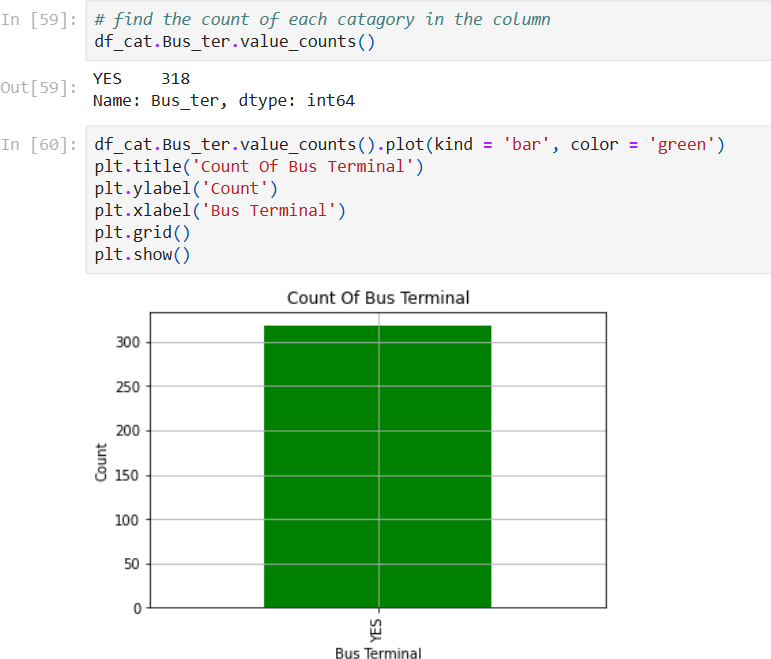
* We have balance data in the Airport columns
* In the data there is 273 places where Airport is present and 225 places where Airport is not

# Waterbody:



* Most of the places having River as waterbody, it's count is 178
* There is few places where lake and River are the waterbody
* This data is imbalance

# Bus\_ter:



* All the places we have contain the bus terminal
* ACTION: remove the column from the data

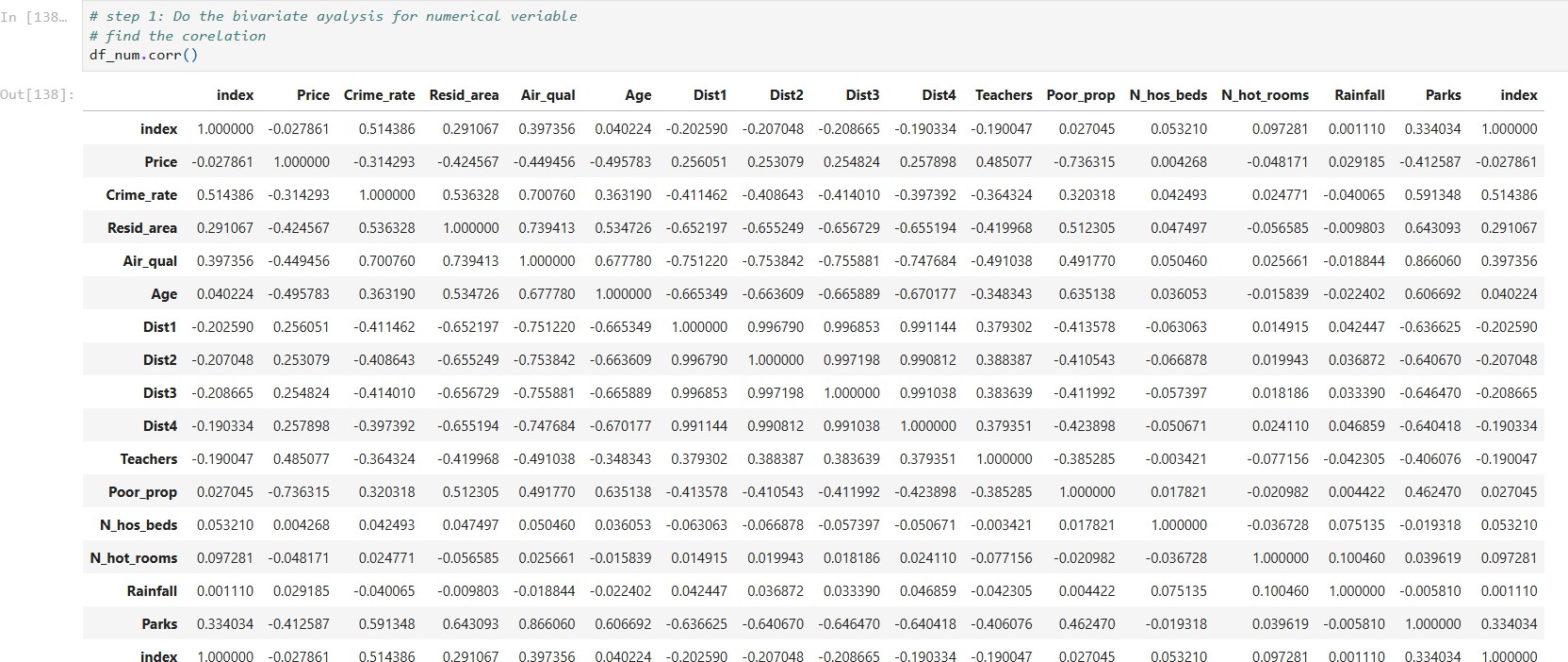
# Bivariate Analysis

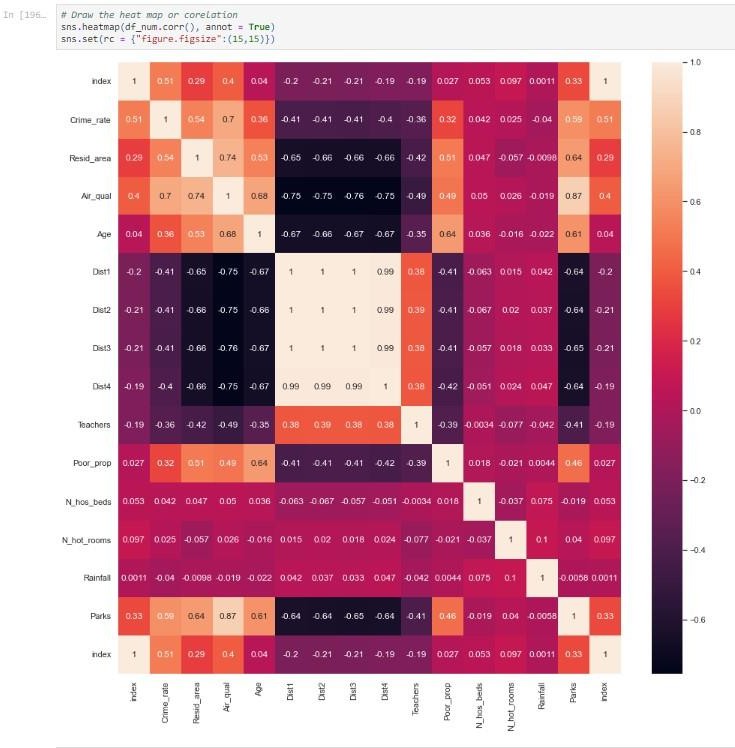
#### Q1. what is mean by bivariate analysis?

**Answer:** Bivariate evaluation involves exploring the connection between variables. It enables find associations, correlations, and dependencies between pairs of variables. Bivariate analysis is a crucial form of exploratory data analysis that examines the relationship between two variables.

It is a helpful technique for determining how two variables are connected and finding trends and patterns in the data. In [statistical analysis](https://www.questionpro.com/blog/statistical-analysis/), distinguishing between [categorical data and numerical](https://www.questionpro.com/blog/categorical-data-vs-numerical-data/) [data](https://www.questionpro.com/blog/categorical-data-vs-numerical-data/) is essential, as categorical data involves distinct categories or labels, while numerical data consists of measurable quantities.

**Q2. method that we you the find the pairs. (ss of the code, ss of the graph, inference) Answer:** We can use Heat map to find the pairs.



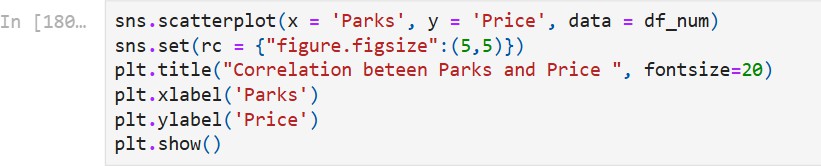


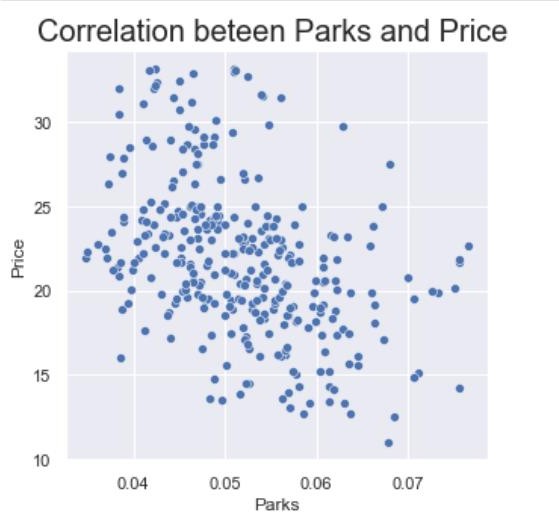
* Here we can consider nearest values for 1 and (-1). Find the pairs using lower triangular matrix.
* *price-parks*
* *Parks - Air\_qual*
* *Dist4 - Dist2*
* *Dist4 - Dist1*
* *Dist4 - Dist3*

#### Q3. what we did with our data? (ss of the code, ss of the graph, inference) column wise (in heading2)

**Answer:** We made pairs of two columns by analysing the heat map

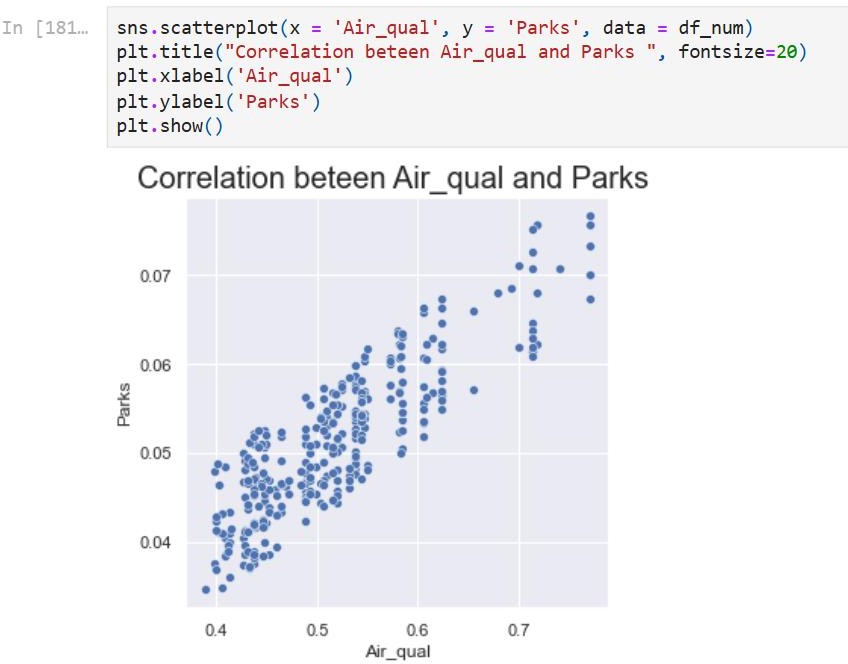
# Parks – Price





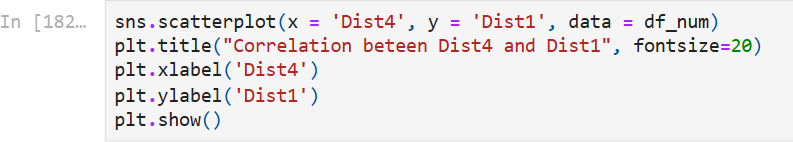
* There is slightly negative correlation
* We can say that if the Price of house is high then the Parks is lesser

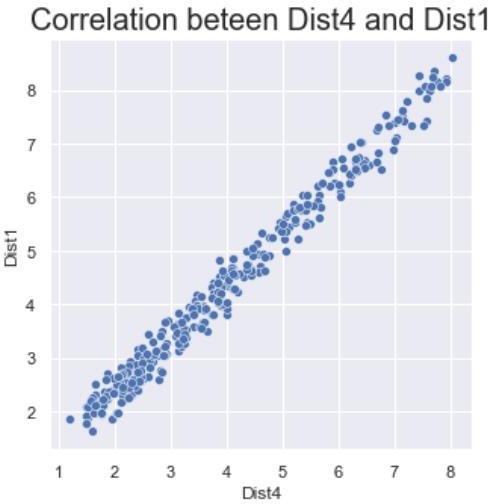
# Parks - Air\_qual



* There is slightly positive correlation
* We can say that if the Air\_qual of house is high then the Parks is higher.

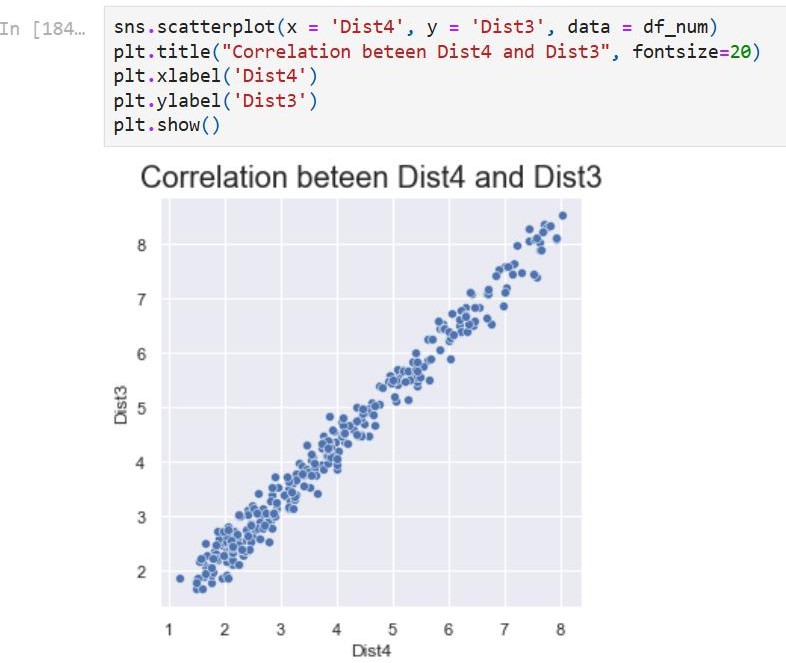
# Dist4 - Dist1





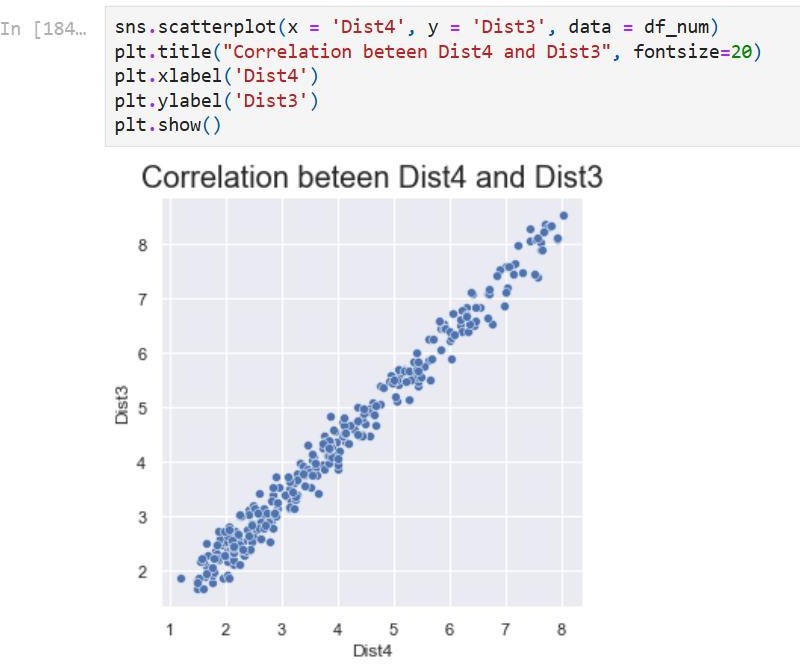
* There is linear correlation
* We can say that if the distance of employment hub 4 and distance of employment hub1 is equal.

# Dist4 - Dist2



* There is linear correlation.
* We can say that if the distance of employment hub4 and distance of employment hub2 is equal.

# Dist4 - Dist3



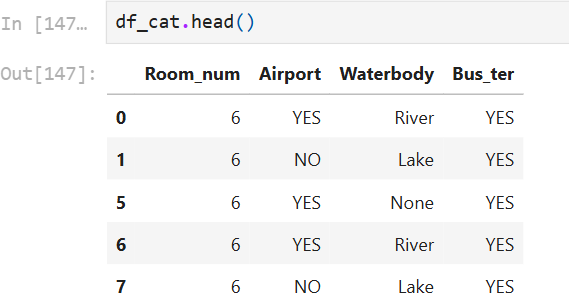
* There is linear correlation.
* We can say that if the distance of employment hub4 and distance of employment hub3 is equal.

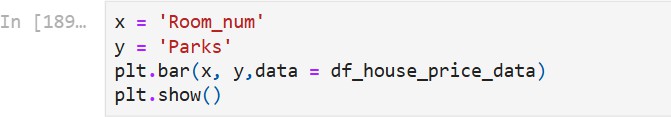
#### Q4. decide the target variable/attributes/column

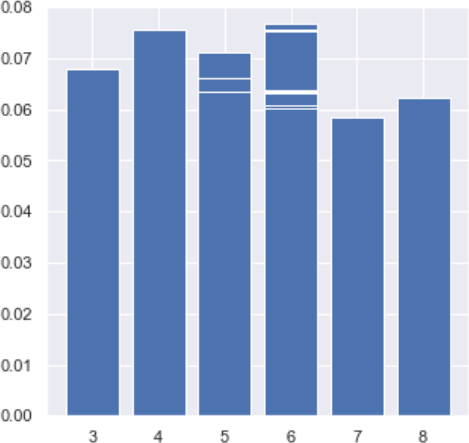
**Answer:** To decide the target variable in bivariate analysis, identify your research objective, classify the variables as independent (predictor) or dependent (outcome), and consider any natural causal relationships.

After analysing all the pairs of the columns we will decide the target variable/attribute/column. So, the target variables will be

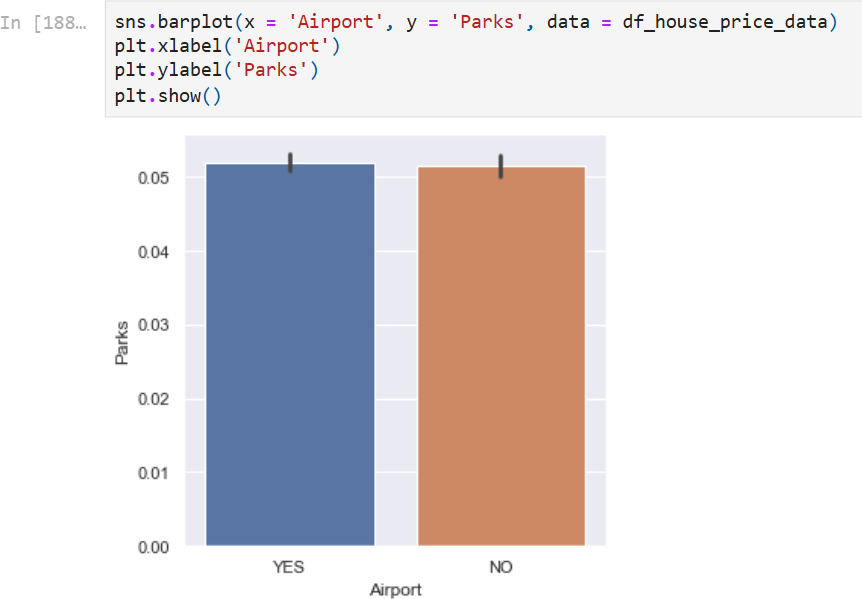
*Parks Price Air\_qual*

Now, check all this target variables with categorical values

1. Parks



* + Average number of parks is high then number of room are 4
  + Average number of parks is leser then number of room are 7

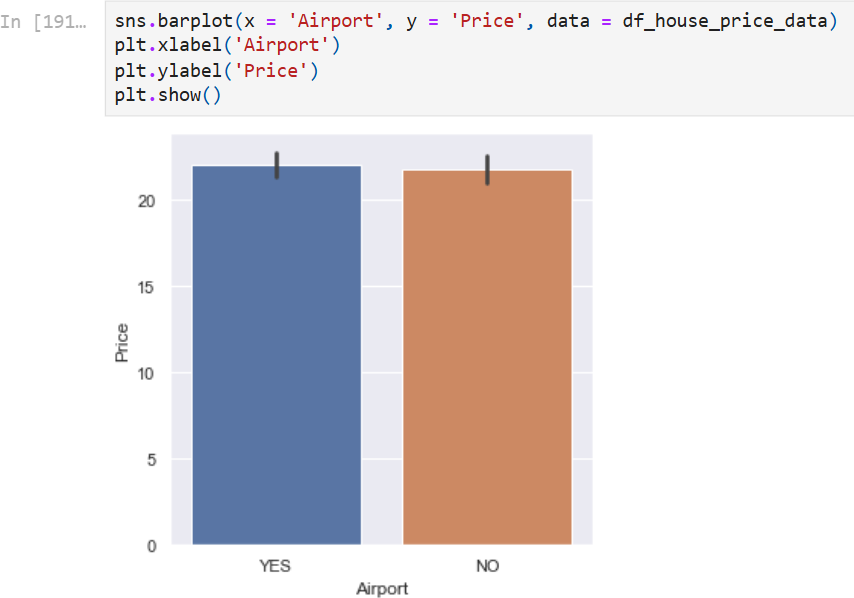


* + Nothing special



* + Nothing special

1. Price
   * When the number of room are 7 than the price is high
   * When the number of room are 4 than the price is less

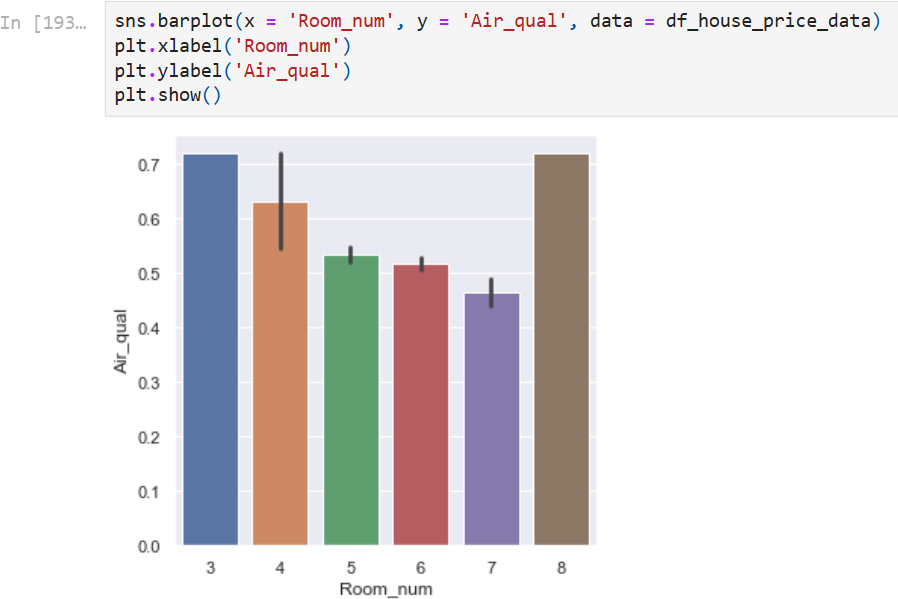


* + Nothing special

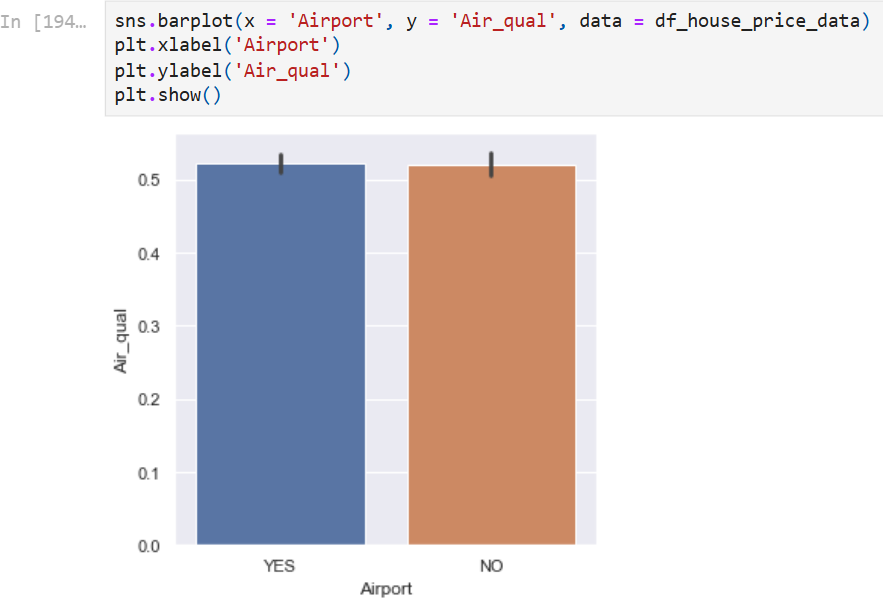


* + Nothing special

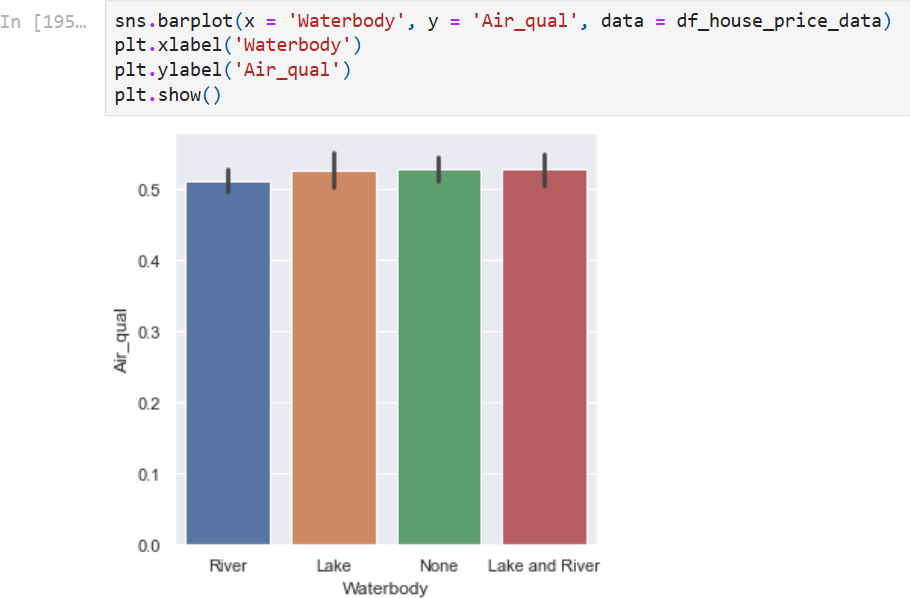
1. Air\_qual



* + when we have best air quality in the area that time average room number is 4 in the house.
  + when we have bed air quality in the area that time average room number is84 in the house.



* + Nothing special



* + Nothing special

# Multivariate Analysis

#### Q1. what is mean by multivariate analysis?

**Answer:** Multivariate analysis examines the relationships between two or more variables in the dataset. It aims to understand how variables interact with one another, which is crucial for most statistical modelling techniques.

#### Q2. method that we use to describe the multivariate analysis.

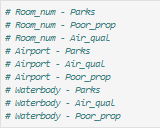
**Answer:** Both Cross Tabulation and Pivot Tables can be instrumental in multivariate

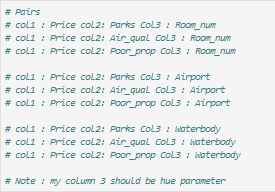
analysis by providing a clear view of how multiple variables interact. Here’s how they fit into multivariate analysis:

* + **Cross Tabulation**: Useful for examining the interaction between two or more categorical variables. For instance, understanding the relationship between education level, occupation, and income category. Cross tabulation can help identify patterns and correlations that might not be immediately obvious.
  + **Pivot Table**: Enables the analysis of data from multiple dimensions simultaneously. It can handle more complex datasets and provide detailed summaries and visualizations. For example, in a sales dataset, a pivot table can help analyze the impact of various factors such as product type, sales region, and sales representative on overall performance.

#### Q3. what we did with our data?(ss of the code, ss of the graph, inference) column wise (in heading2)

**Answer:** From pairs of target variables and categorical values make multiple columns pairs





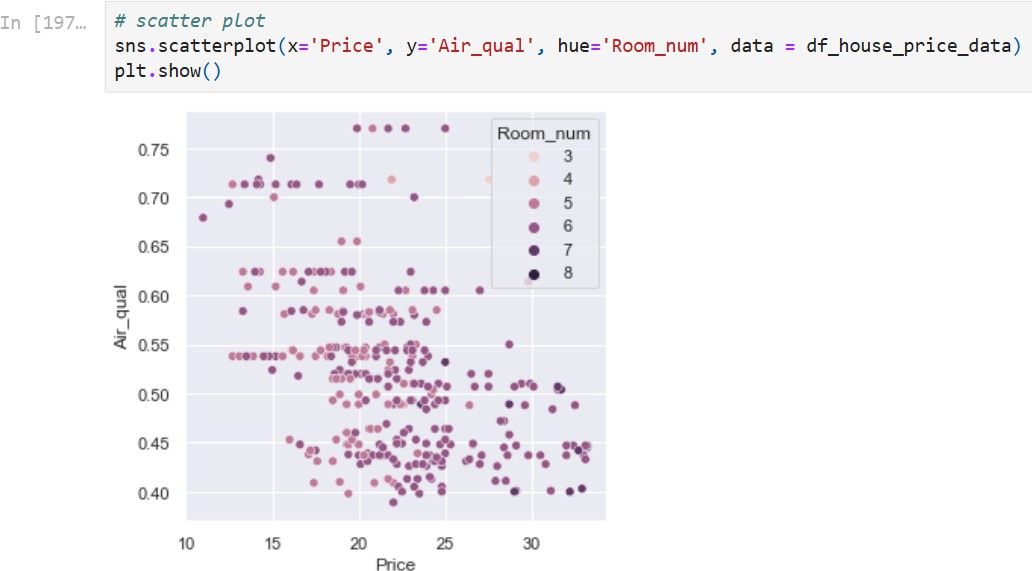
1. Room\_num

col1: Price, col2: Parks col3: Room\_num



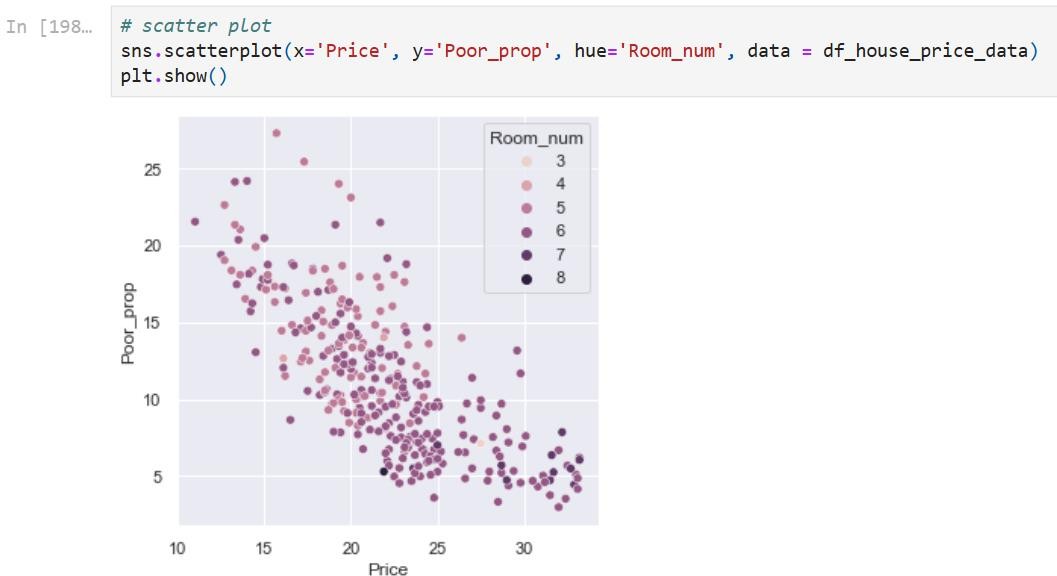
* + When no of Parks is high (0.08) that time the Price is low (13 lacs) the average No of rooms in that scenario is 3.
  + When no of Parks is low (0.04) that time the Price of the house is high (36 lacs) the average No of rooms in that scenario is 8.

col1: Price, col2: Air\_qual col3: Room\_num



* + When the price of the house is low (13 lac) that time air quality is really good(0.9) the average No of rooms in that scenario is 4.
  + Price When the price of the house is high (36 lac) that time air quality is really bad (0.45) the average No of rooms in that scenario is 8.

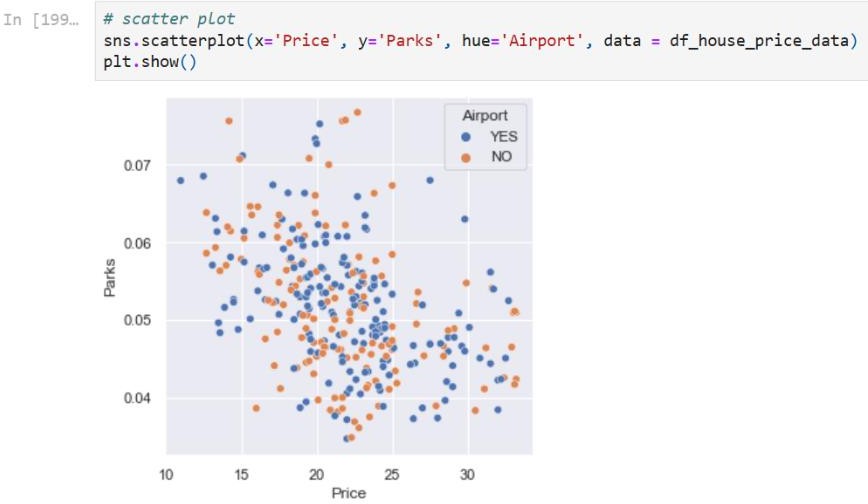
col1: Price, col2: Poor\_prop col3: Room\_num



* + When the price of the house is high (36 lac) that time population of poor people is 4 the average No of rooms in that scenario is 8.
  + Price When the price of the house is low (7 lac) that time population of poor people is 30 the average No of rooms in that scenario is 4.

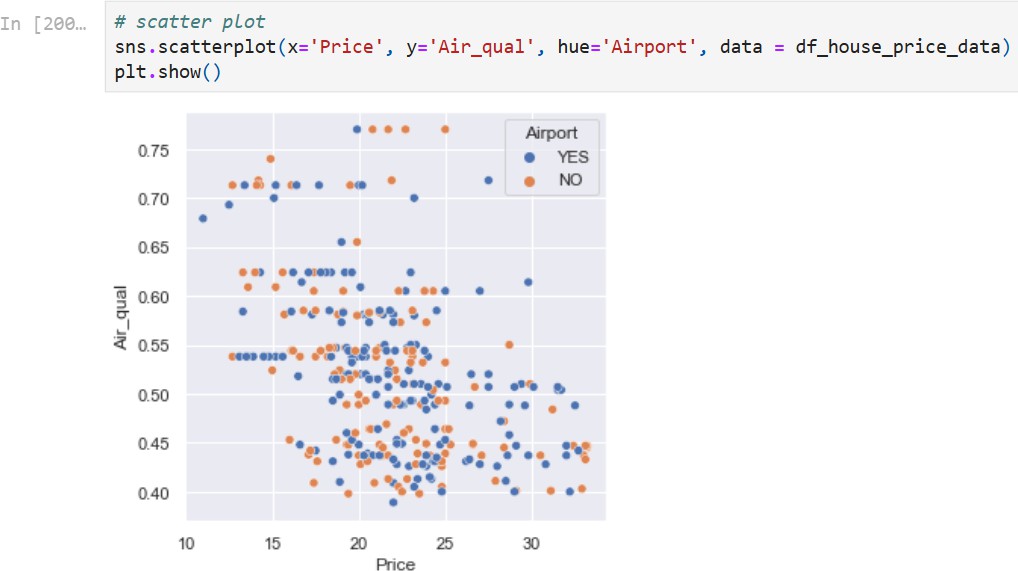
1. Airport

col1: Price, col2: Parks col3: Airport



* + When the price of the house is high (37 lac) that no of parks is (0.045) that time there is airport is not present
  + When the price of the house is low (13 lac) that no of parks is (0.068) that time there is airport is present

col1: Price, col2: Air\_qual col3: Airport



* + When the price of the house is low (12 lac) that time of the air quality is really good (0.69) that time there is airport is present
  + When the price of the house is high (35 lac) that time of the air quality is really bed (0.41) that time there is airport is not present

col1: Price, col2: Poor\_prop col3: Airport



* + When the price of the house is high (37 lac) that time propulation of poor people is 4 that time there is airport is not present
  + When the price of the house is low (12 lac) that time propulation of poor people is 24 that time there is airport is present

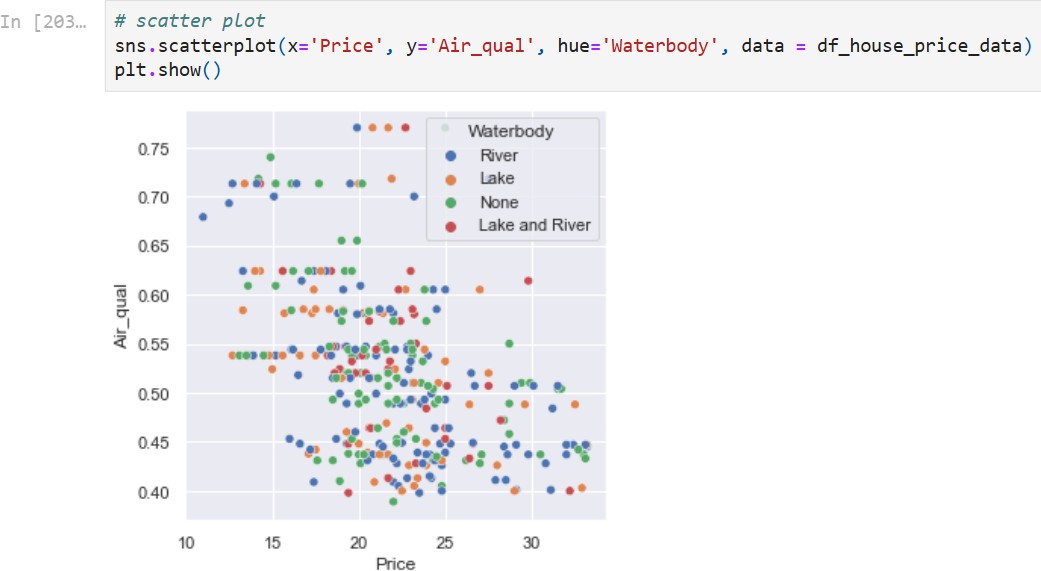
1. Waterbody

col1: Price, col2: Parks col3: Waterbody



* + When the price of the house is high (37 lac) that time no of parks is (0.045) that time they have river as Waterbody
  + When the price of the house is low (12 lac) that time no of parks is (0.0698) that time they have river as Waterbody

col1: Price, col2: Air\_qual col3: Waterbody



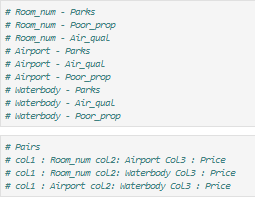
* + When the price of the house is high (35 lac) that time of the air quality is really bed (0.40) that time they have lake and river as Waterbody
  + When the price of the house is low (12 lac) that time of the air quality is really good (0.69) that time they have river as Waterbody

col1: Price, col2: Poor\_prop col3: Waterbody



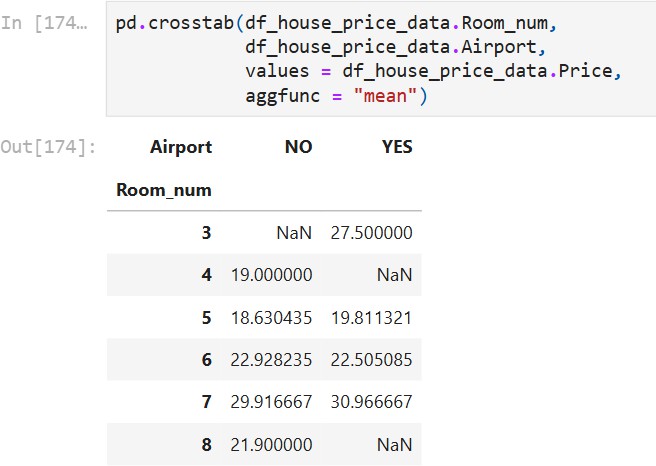
* + When the price of the house is high (35 lac) that time propulation of poor people is 4 that time they have lake as Waterbody
  + When the price of the house is low (12 lac) that time propulation of poor people is 22 that time they have river as Waterbody

### Now Arrangement of pivot and cross tab

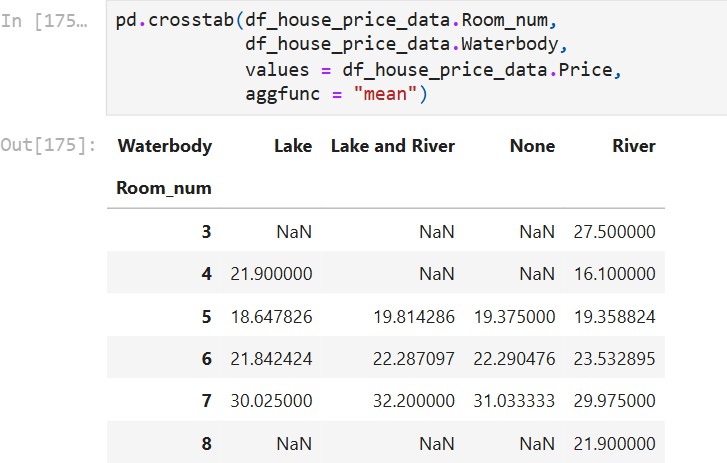


Cross Tab

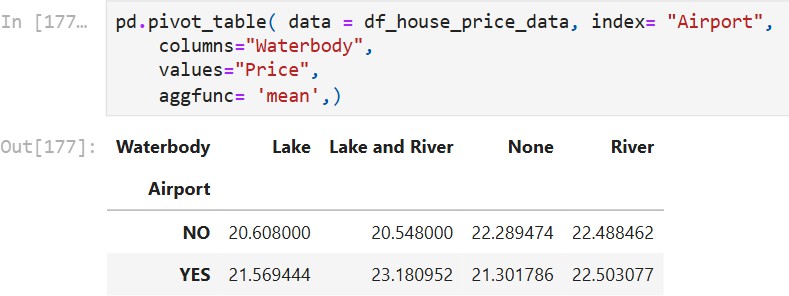
col1: Room\_num, col2: Airport, col3: Price



col1: Room\_num, col2: Waterbody, col3: Price



col1: Airport, col2: Waterbody, col3: Price



## Summary:

This report on the analysis of housing-related data for a builder community. The objective is to analyse historical data to help builders understand and predict housing prices, develop policies, and identify factors influencing house prices. The methodology includes data cleaning, handling missing values, outlier treatment, and exploratory data analysis (EDA).

### Key Points:

#### Data Overview:

* + The data includes various attributes related to houses, such as price, crime rate, residential area, air quality, number of rooms, age, distance to amenities, number of teachers, poverty proportion, airport proximity, number of hospital beds, number of hot rooms, water body proximity, rainfall, bus terminals, and parks.

#### Data Preprocessing:

* + **Data Cleaning**: Involves renaming columns and typecasting certain columns like room number, age, teachers, and number of hot rooms.
  + **Missing Value Treatment**: There are three types of missing values. The report follows manufacturing industry rules for handling missing data: removing observations with less than 10% missing data, imputing missing values if 30-70% of data is missing, and removing attributes with more than 80% missing data.
  + **Outlier Treatment**: The Interquartile Range (IQR) method is used to detect and treat outliers.

#### Exploratory Data Analysis (EDA):

* + **Univariate Analysis**: Examines single variables individually.
  + **Bivariate Analysis**: Studies the relationship between two variables.
  + **Multivariate Analysis**: Explores relationships among three or more variables.

#### EDA Methods:

* + **Univariate Analysis**: Analysis of individual columns like room number, price, crime rate, residential area, air quality, age, distances, teachers, poverty proportion, hospital beds, hot rooms, rainfall, and parks.
  + **Bivariate Analysis**: Analysis includes creating heat maps for correlation and examining pairs of columns like price-parks, residential area-parks, air quality-parks, age-parks, distances-parks, price-poverty proportion, and age-poverty proportion.

The report provides a comprehensive analysis process starting from data collection to preprocessing, and then exploring the data to extract meaningful insights. This structured approach helps builders and policymakers make informed decisions regarding housing prices and related factors.