Report

Objective:

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

Introduction:

Me DIYA MAHYAVANSHI responsible to commutate the analysis on housing related data, In this project, in this data we have we have following one flow(core analytics flow). This analysis beneficial for multipole 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 the house. The motivation of the analysing the data is came from standard “ Manufacturing Analytical rules for analysis”, This standards are utilised in standardising the null values. As per my knowledge and under the guidance of PRAFUL VINAYAK BHOYAR SIR, I am going to conclude this analysis as the standard followers.

Methodology :

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| Data Overview |  |  | Data Preprocessing |  |  | EDA |  |  | Reporting |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  | Data cleaning |  |  | Univariate analysis |  |  |  |
|  |  |  | Missing value treatment |  |  | Bivariate Analysis |  |  |  |
|  |  |  | Outlier treatment |  |  | Multivariate Analysis |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| Figure 1 Methodology | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

Data Overview :

Q1. Collect the detail information about the data

**Answer** : The data consists of information related to different entities, their corresponding tables, fields, data types, and foreign key references.

Q2. Collect the detail information about the column/Attributes

**Answer** : we have information about the columns of price, crime\_rate, resid\_area, air\_qual, room\_num, age, dist1, dist2, dist3, dist4, teachers, poor\_prop, airport, n\_hos\_beds n\_hot\_rooms, waterbody, rainfall, bus\_ter ans parks

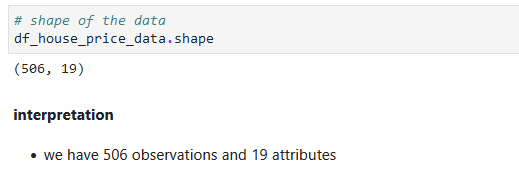
Q3. show the basic statistics of the data

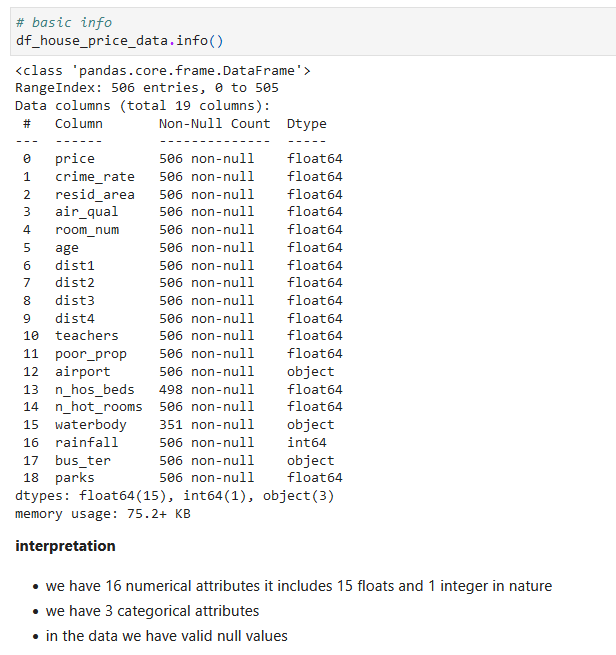
Q3. Answer - Screenshot of shape and the description and information

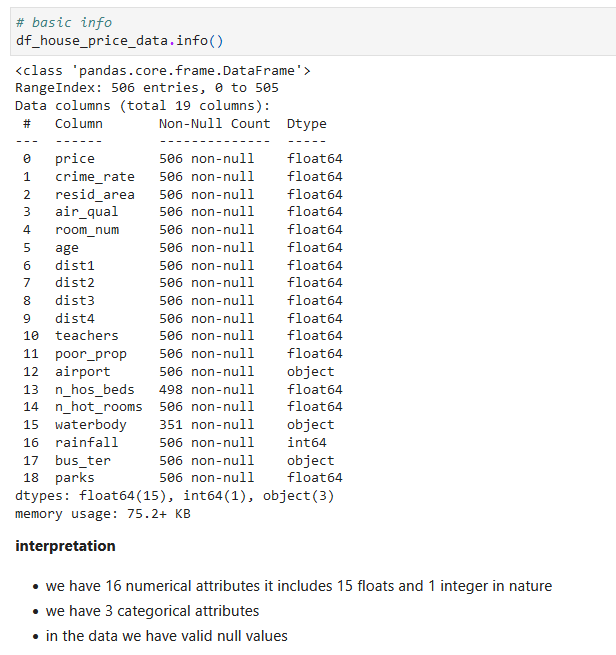
Screenshot of info and the description

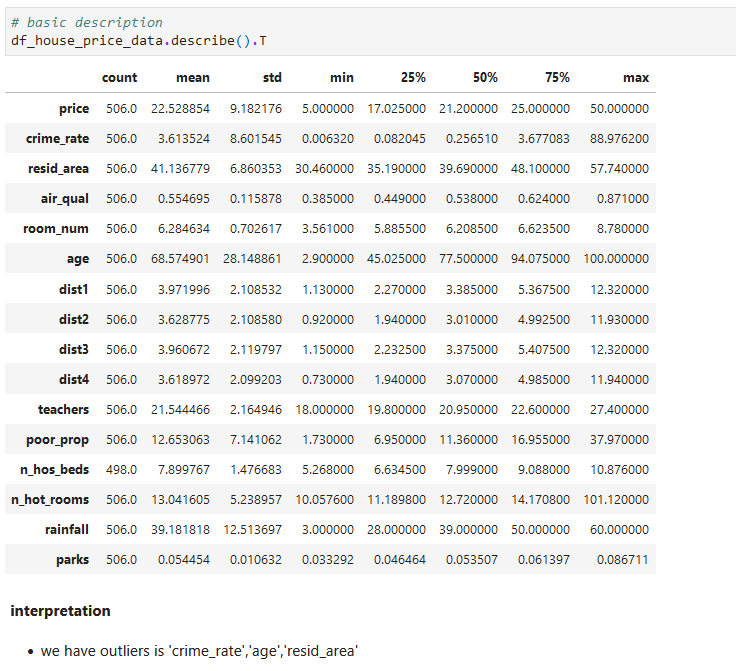
Screenshot of the description and the description

**Answer** :









Data Preprocessing :

In the preprocessing we have few stages that we need to follow , Those stages we can see inside figure 1

1. Data cleaning
2. Missing value treatment
3. Outliler treatment

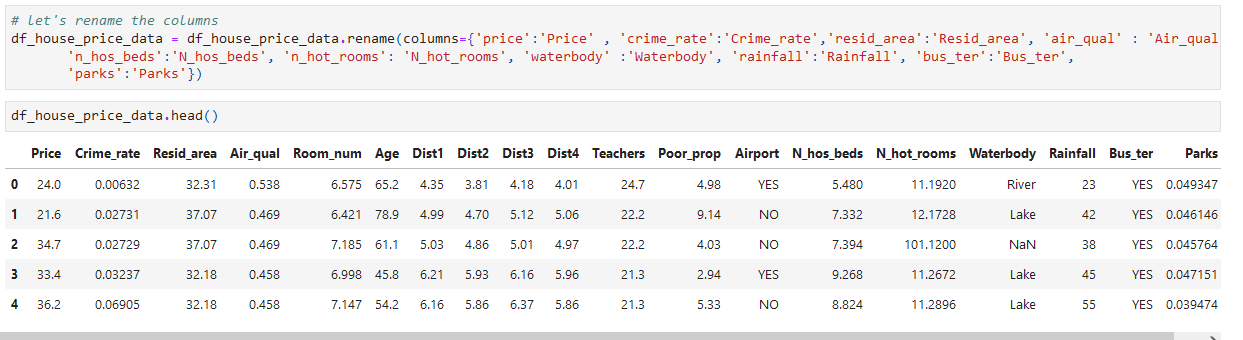
# Data cleaning :

Q1 . What is data cleaning ? (in small paragraph)

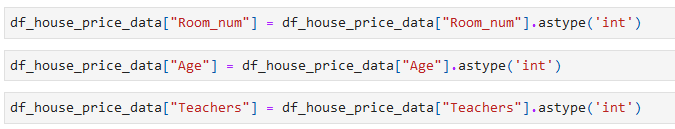
**Answer** : Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

Q2 . What we did in our data (screen shot of the code ,information about effect)

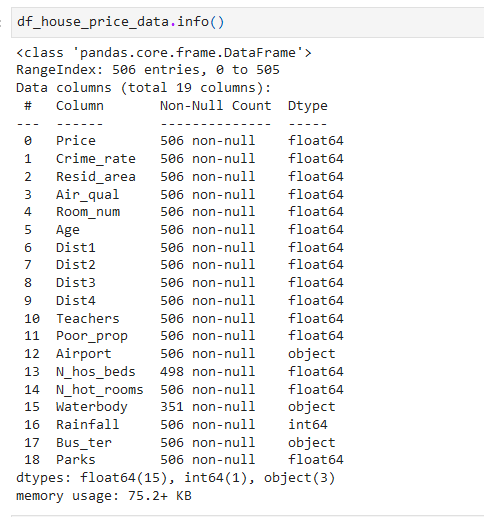
**Answer** : Renaming the columns



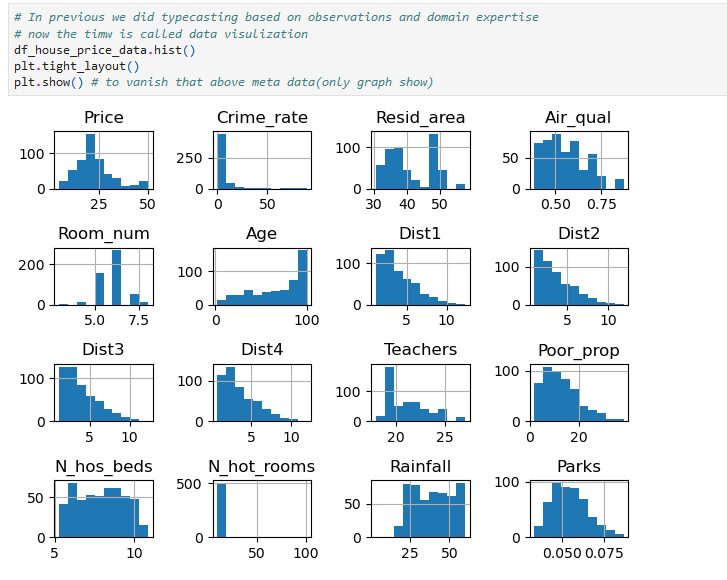
Typecasting the columns Room\_num, Age, Teachers, N\_hot\_rooms.







* Visualised the Data using Hist Map.



# Missing value treatment :

Q1 . What are type of missing values (small paragraph)

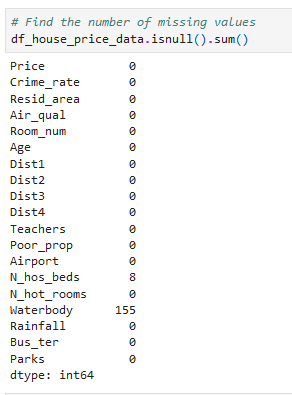
**Answer** : There are Three Types Of Missing Values, 1. Missing data are randomly distributed across the variable and unrelated to other variables., 2. Missing data are not randomly distributed but they are accounted for by other observed variables., 3. Missing data systematically differ from the observed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

Q2. Rules of the missing value as per the manufacturing industry (small paragraph)

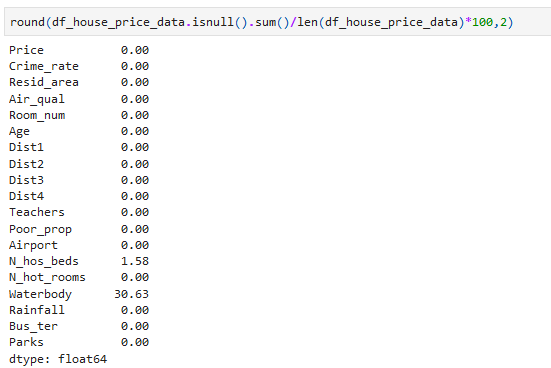
**Answer** : There are mainly Three types Of Missing Values Rules as per the manufacturing Industry. First one is When we have 10% Data Can be missing Than The Data Removed The Observation. Second one is When We Observed The Data Can be Percentage between 30 to 70 Percentage then We need to input the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

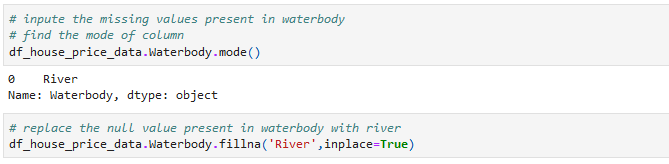
Q3. What we did on our data (screenshot of code and info. About the effect)

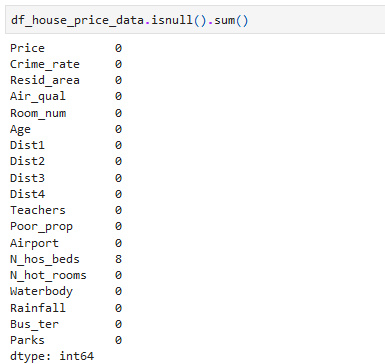
* **Answer** : Find the missing values in our Data



* Check the how many Percentage Data are missing data.

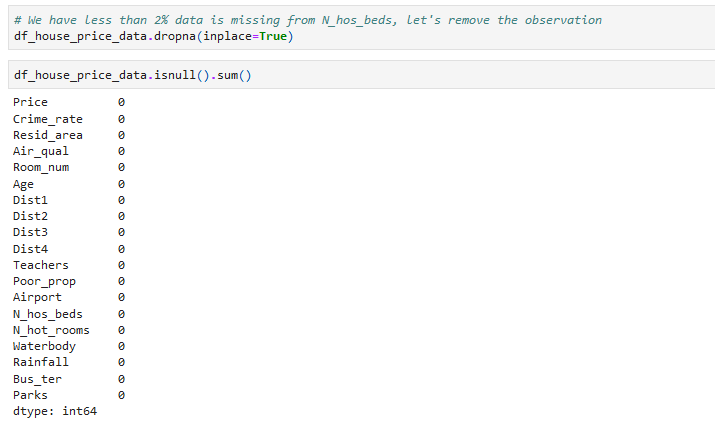






* We Have Less than 2% Data is Missing From N\_Hos\_Beds, Lets the Remove the observation.

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



# Outliler treatment

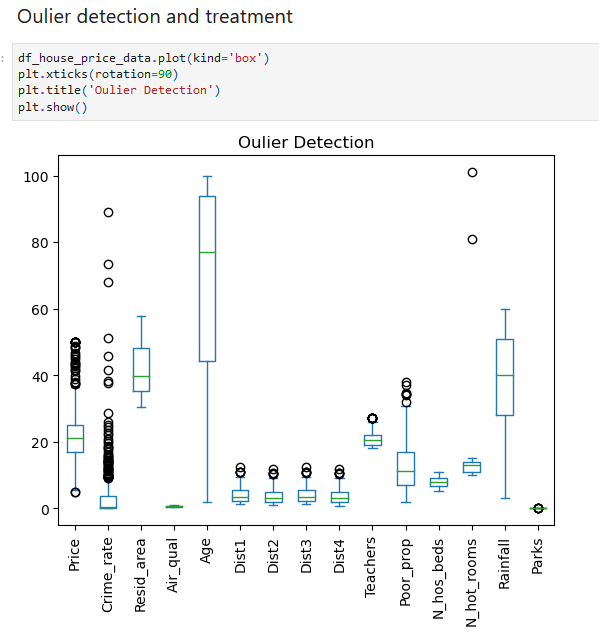
Q1. What is mean by the outliler (small paragraph)

**Answer** : An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

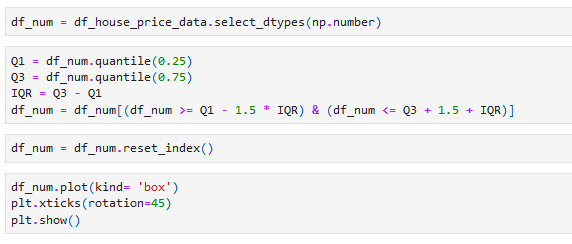
Q2. Which method we used to treat outliers

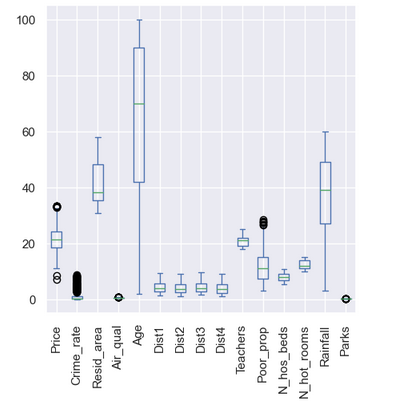
**Answer** : To treat outliers in dataset, I used the Interquartile Range (IQR) method. The IQR method is a robust statistical technique that identifies outliers based on the spread of the middle 50% of the data. Here's how I applied this method:

Q3. What we did on our data (screenshot of code and info. About the effect)

**Answer** : 

* After detection using IQR method for outliers treatment





EDA (Exploratory Data Analysis):

It is a method that we utilize to explore unseen data, Also it is an process of converting any raw data(unseen data) into actionable insights/ information.

Q1. Explain the types of the eda ( small paragraph of each)

**Answer** : There are three types of EDA 1) Univariate analysis 2) Bivariate analysis 3) Multivariate analysis

1. Univariate analysis : Univariate refers to the analysis of a single variable at a time. It involves examining the characteristics and distribution of that variable without considering any other variables. This approach is often used to understand the behavior and patterns within individual variables in a dataset.
2. Bivariate analysis : Bivariate analysis involves the simultaneous analysis of two variables to determine if there is any relationship between them. It helps to understand how one variable behaves in relation to another.
3. Multivariate analysis : Multivariate analysis extends beyond bivariate analysis to include the simultaneous analysis of three or more variables. It aims to understand how multiple variables interact with and relate to each other. This type of analysis is used to explore more complex relationships and patterns within a dataset.

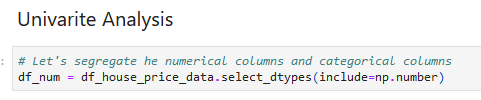
# Univariate Analysis

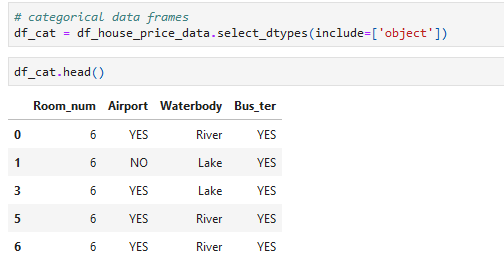
Q1. what is mean by univariate analysis(Small paragraph)

**Answer** : Univariate analysis is the simplest form of analysing data. Uni means "one", so the data has only one variable (univariate). Univariate data requires to analyse each variable separately.

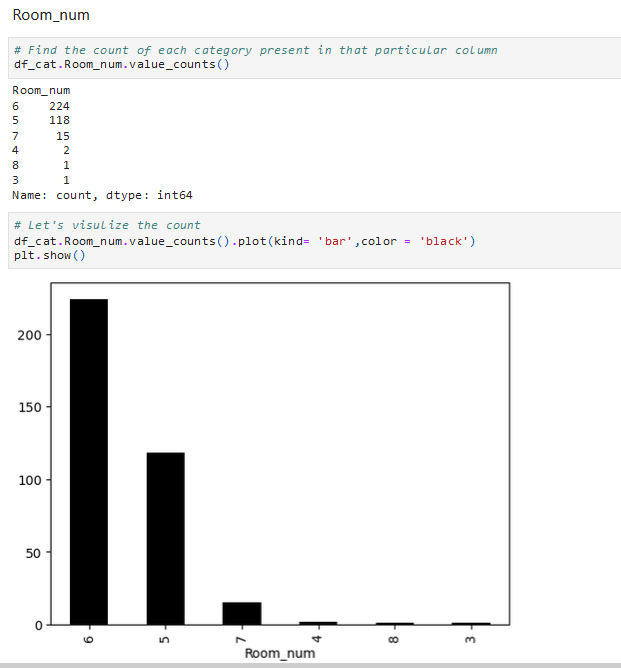
Q2. What we did with our data (screenshot of code, screenshot of the graph, inference column wise<in heading2>)

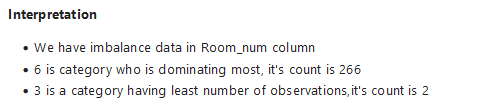
**Answer** :



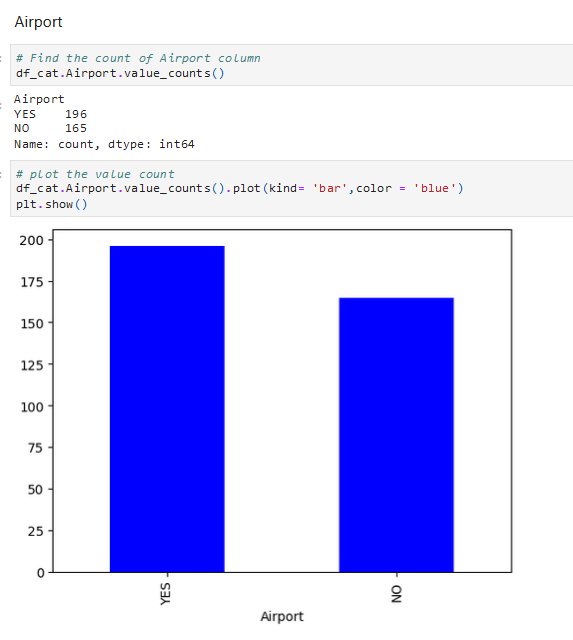


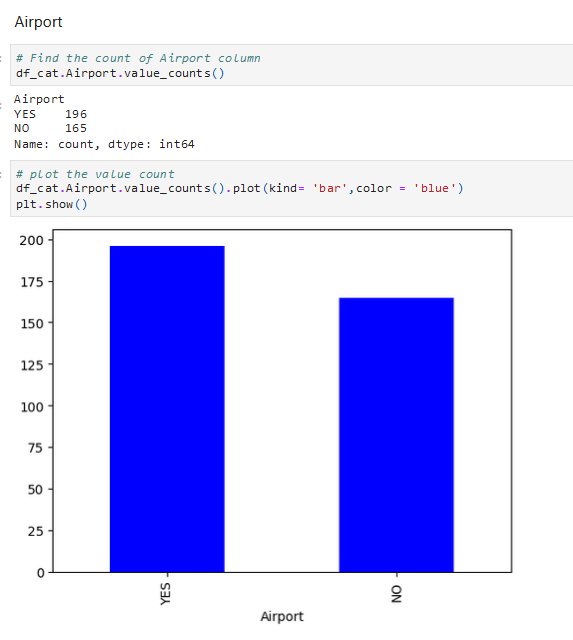
Room\_num

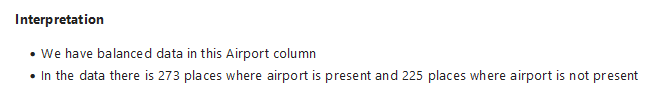




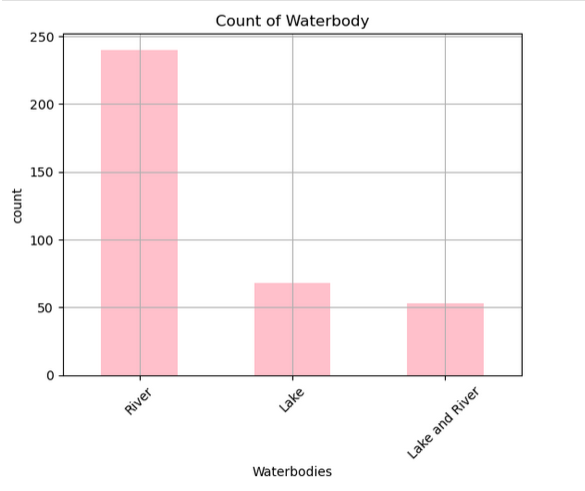
Airport

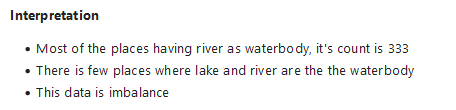




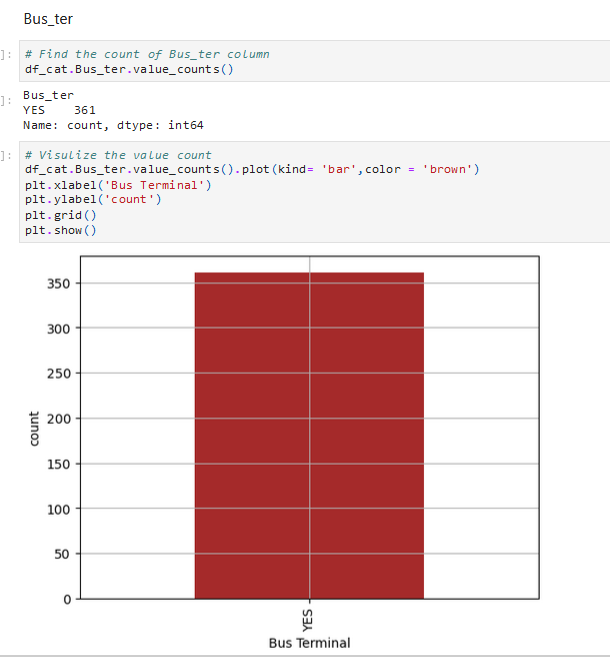


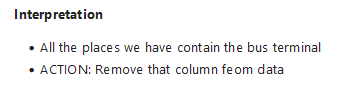






Bus\_ter

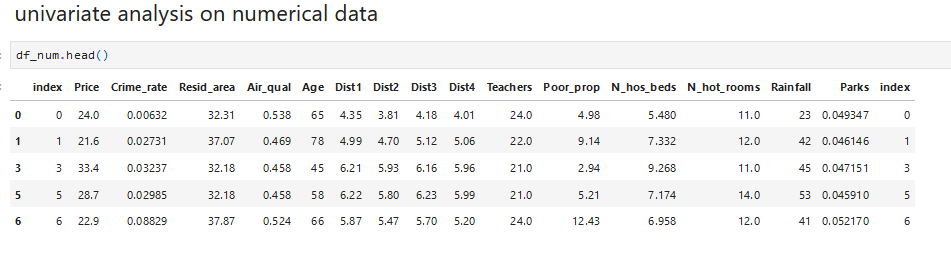




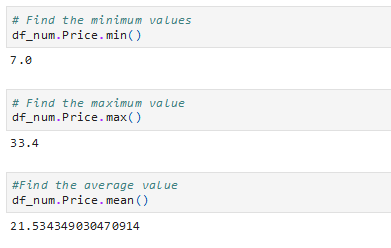
Removing this column



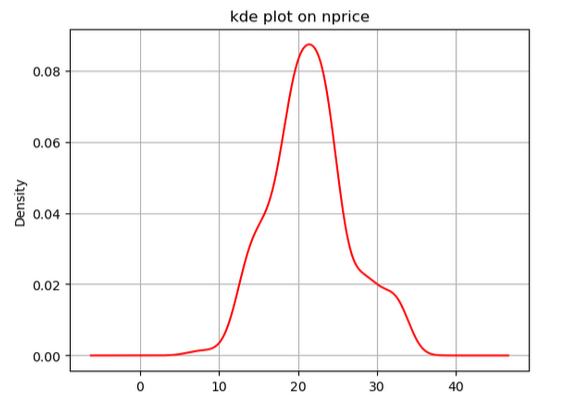
* Univariate analysis on Numerical data

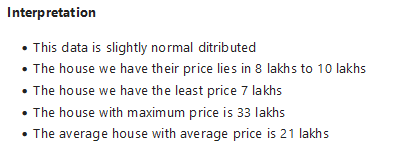


Price



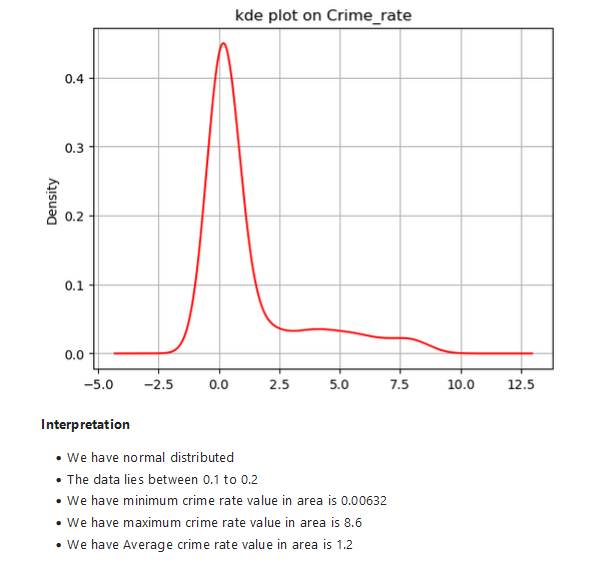






Crime\_rate

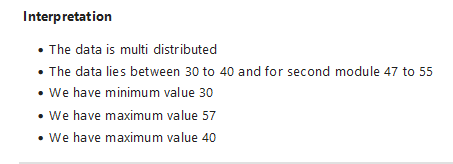




Resid\_area

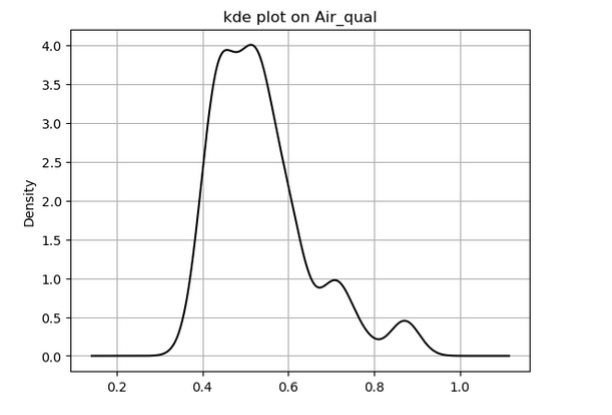


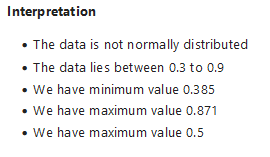




Air\_qual



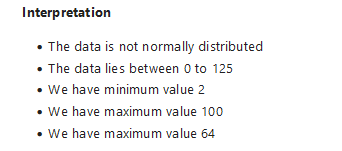




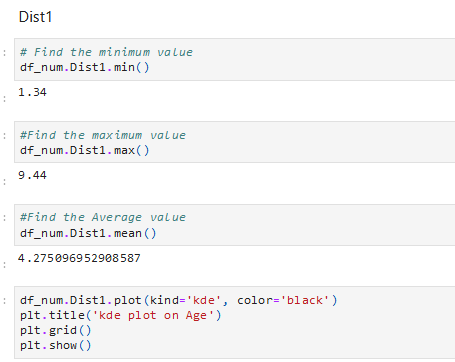
Age

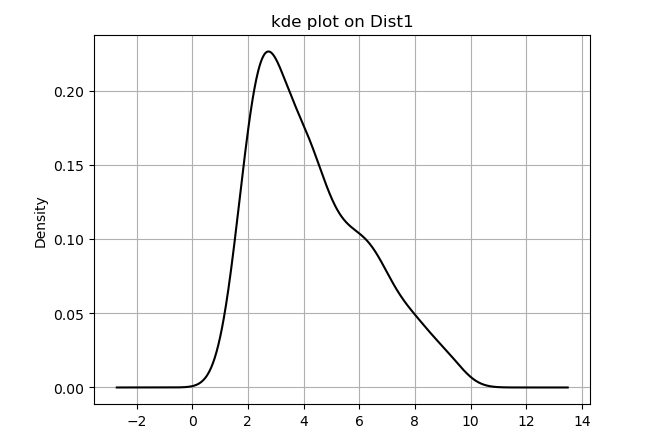


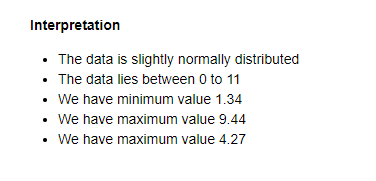




Dist1

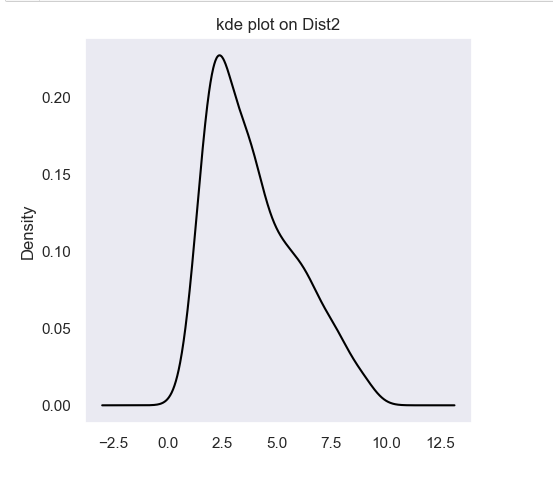


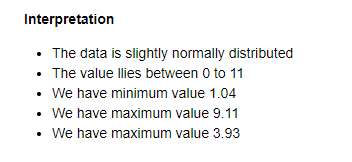




Dist2

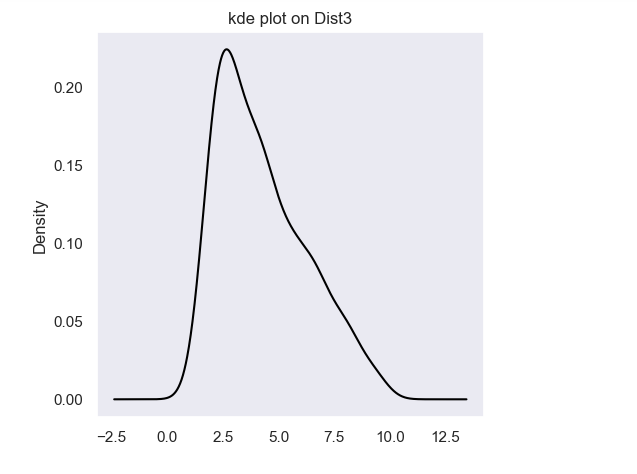


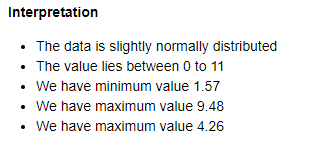




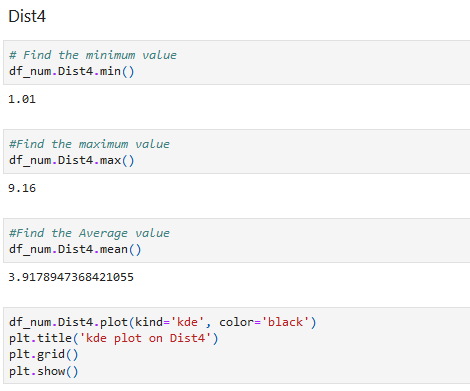
Dist3

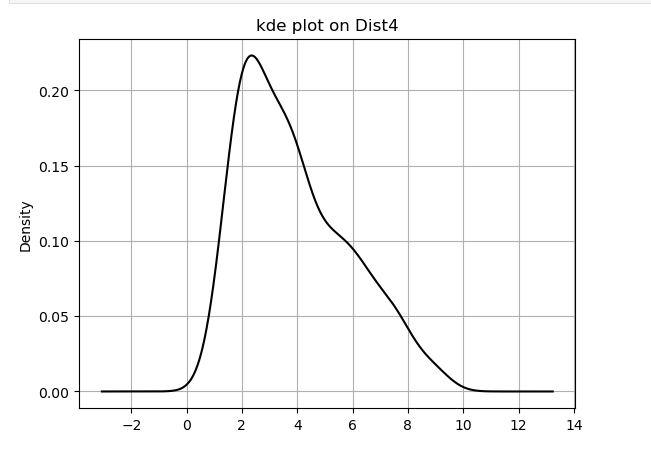


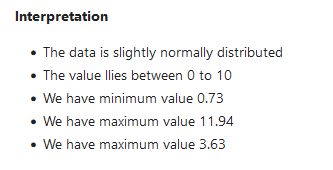




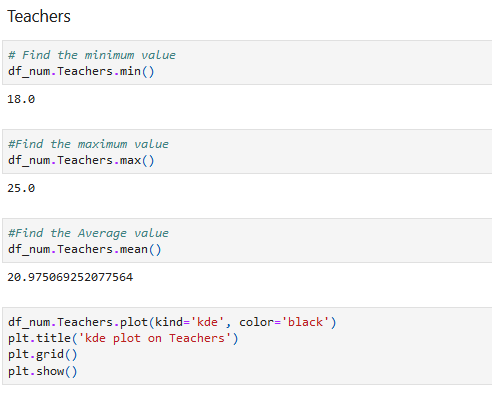
Dist4

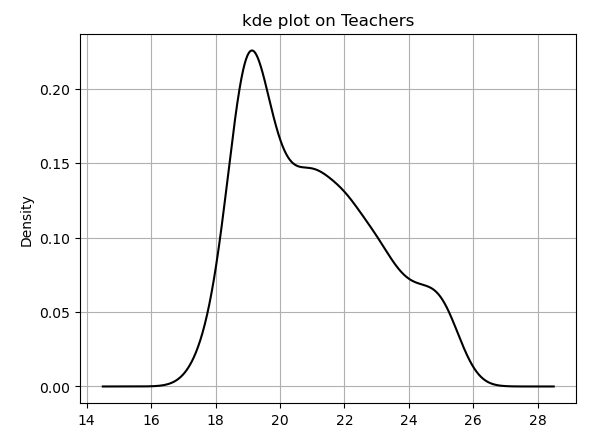






Teachers

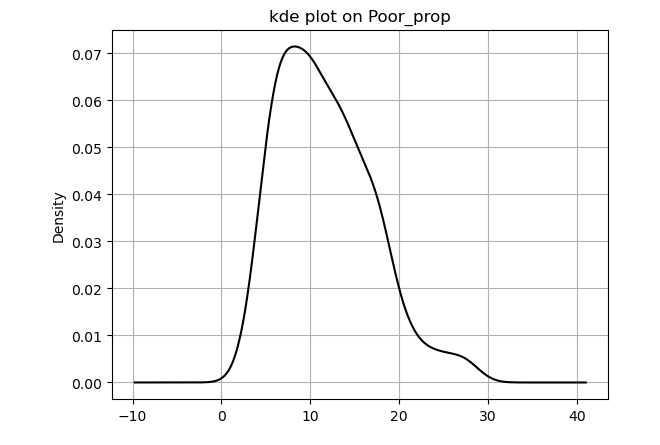


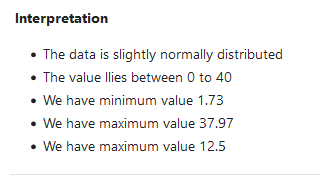




Poor\_prop

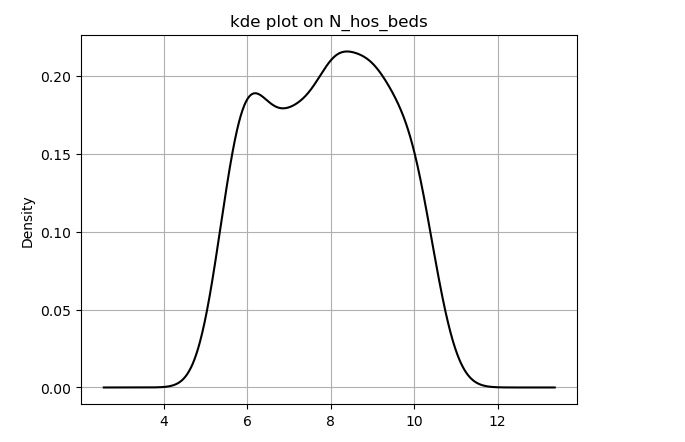


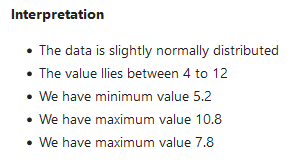




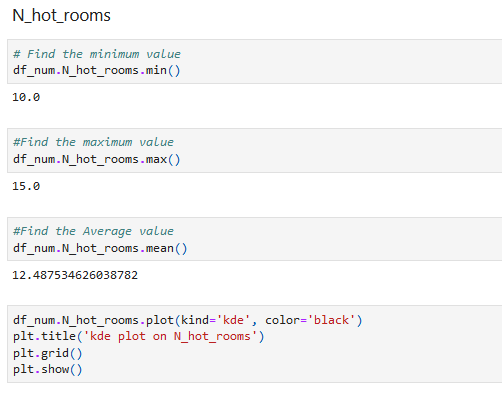
N\_hos\_beds

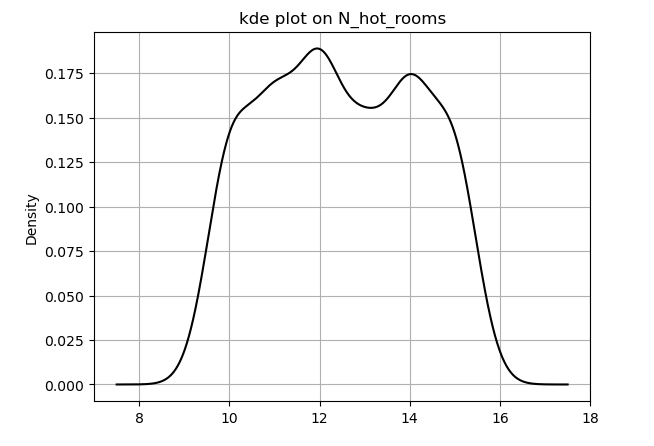


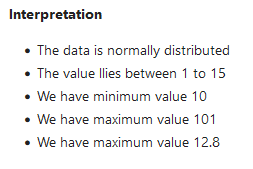




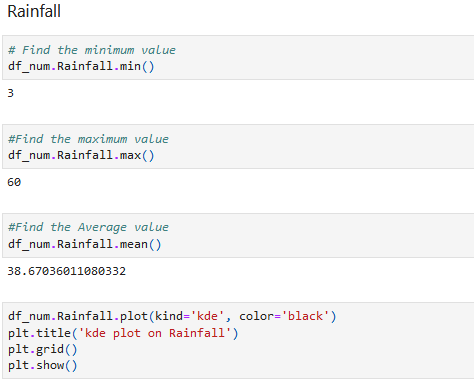
N\_hot\_rooms

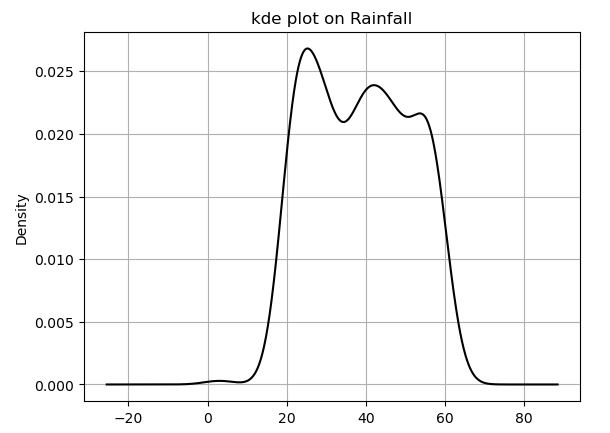


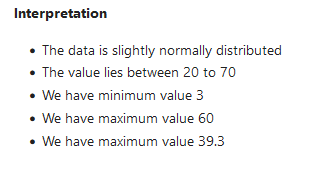




Rainfall

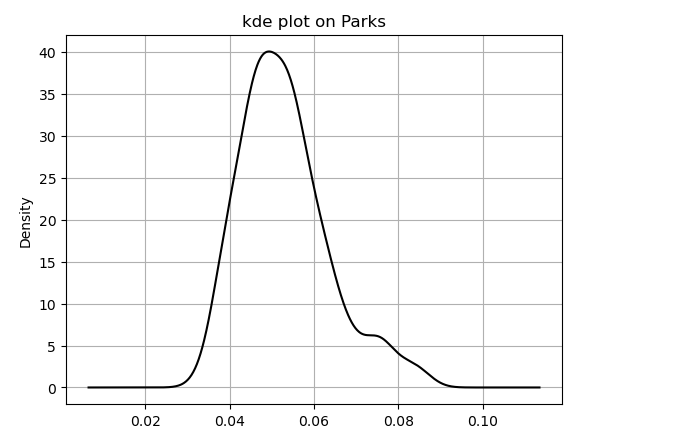


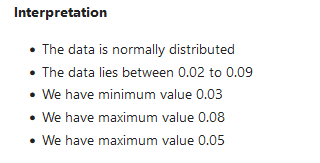




Parks







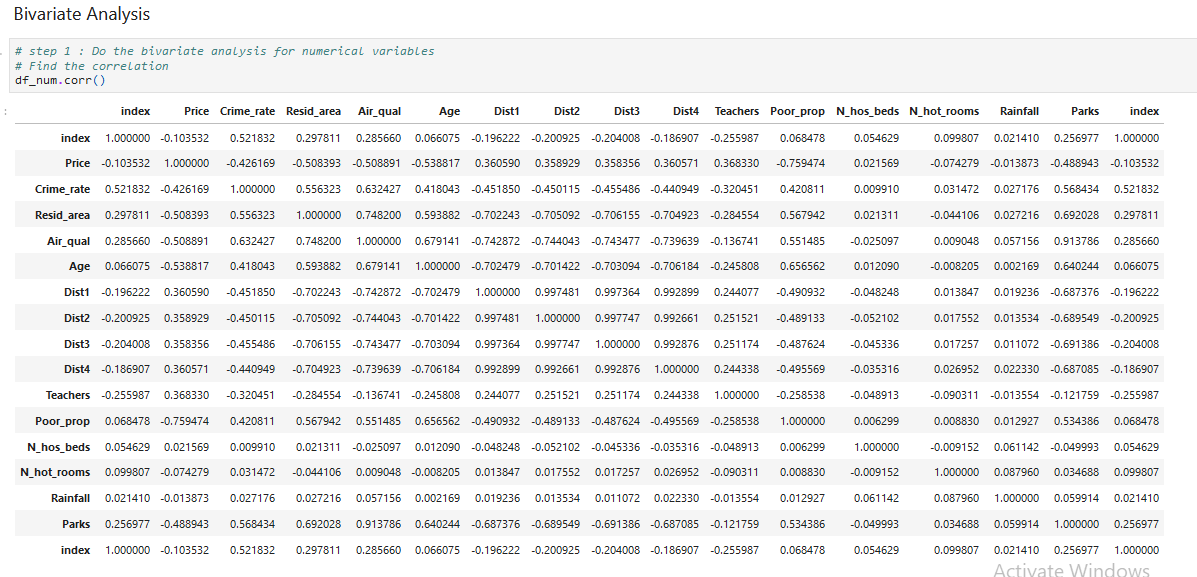
# Bivariate Analysis

Q1. what is mean by Bivariate analysis(Small paragraph)

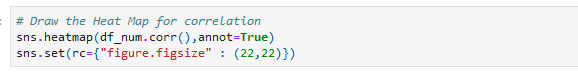
**Answer** : Bivariate analysis is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y.

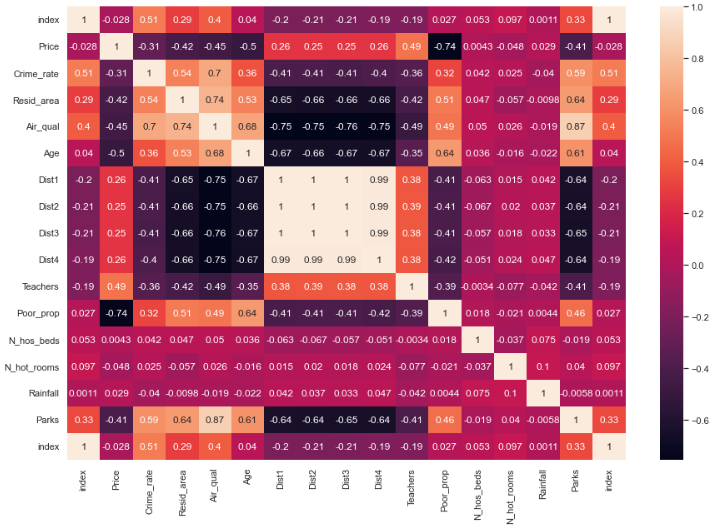
Q2. Method that we use to find the pairs(screenshot of code,screenshot of the graph, inference )

**Answer** :



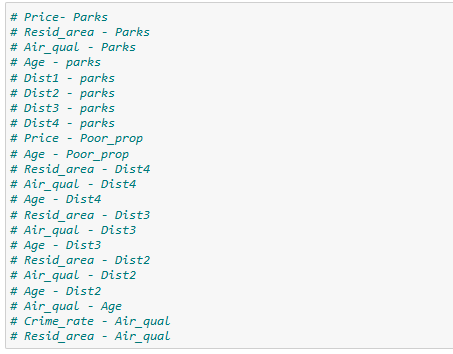
* Heat map for correlation





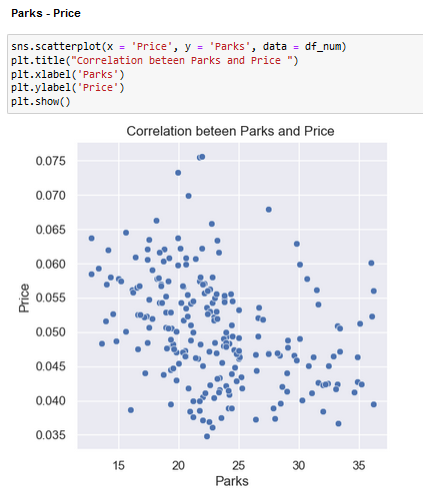
Q3. What we did with our data (screenshot of code,screenshot of the graph, inference column wise<in heading2>)

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

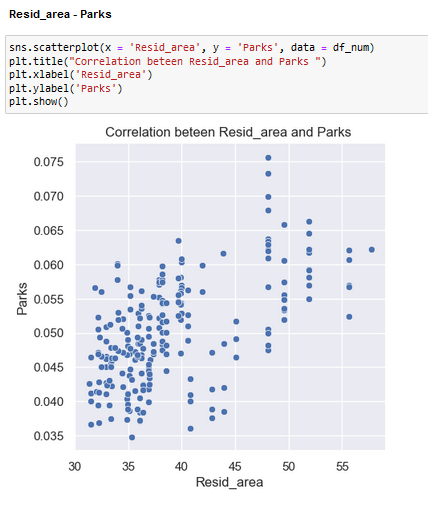


Now, we will find the correlation between our choosen pairs of columns

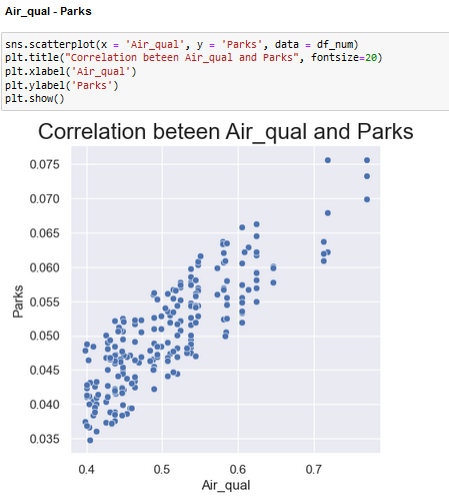
1. # Price – Parks



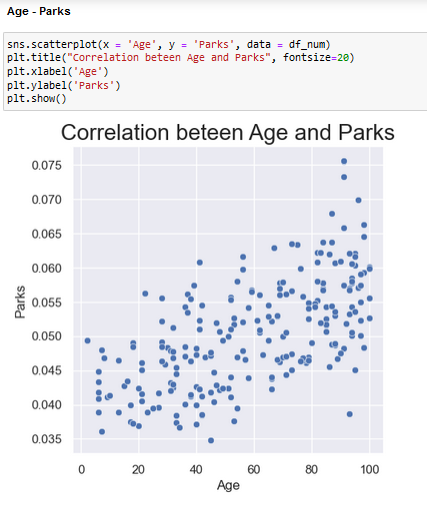
1. # Resid\_area – Parks



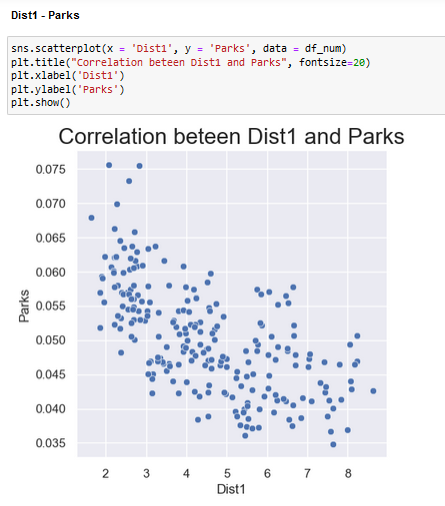
1. # Air\_qual – Parks



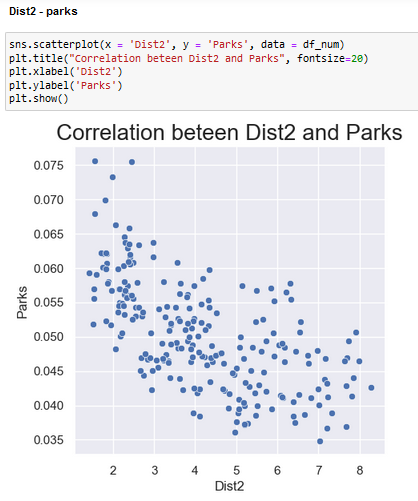
1. # Age – parks



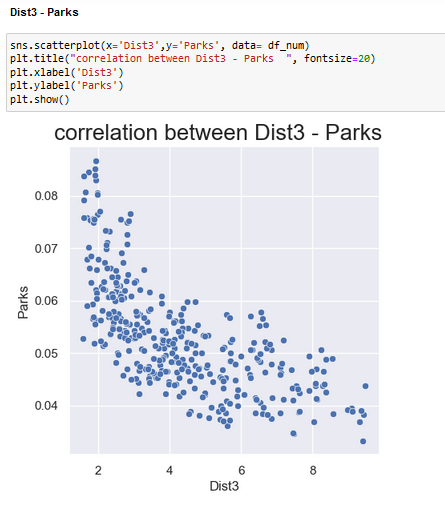
1. # Dist1 – parks

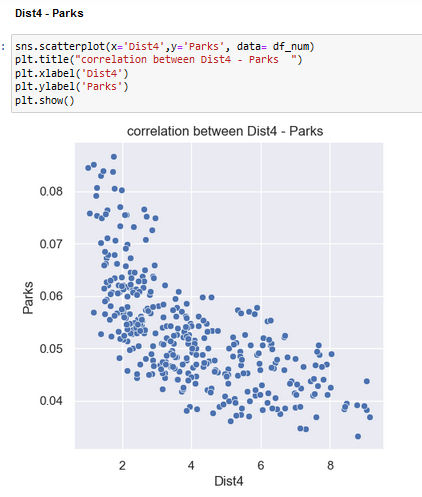


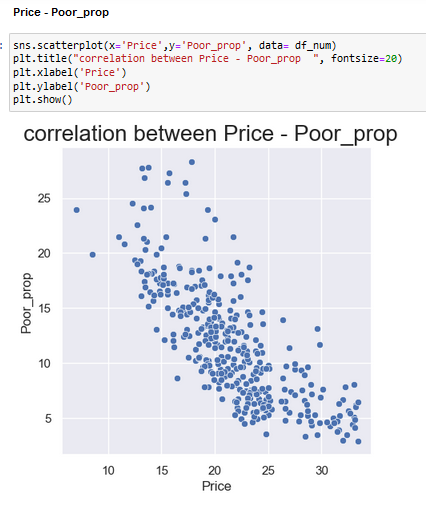
1. # Dist2 – parks



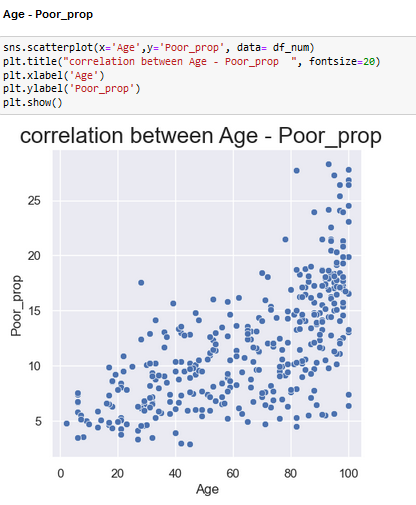
1. # Dist3 – parks



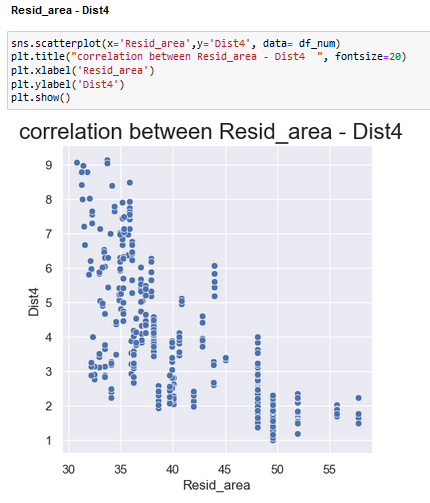
1. # Dist4 – Parks  
   
2. # Price - Poor\_prop



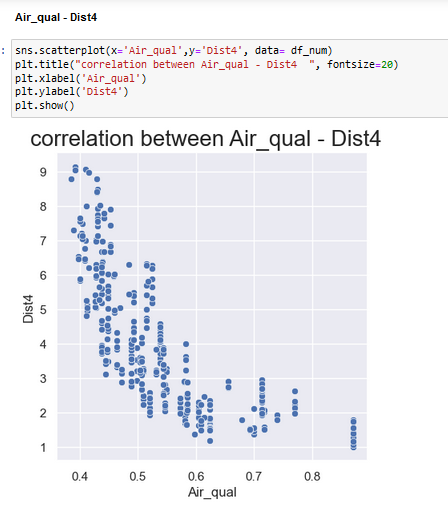
1. # Age - Poor\_prop



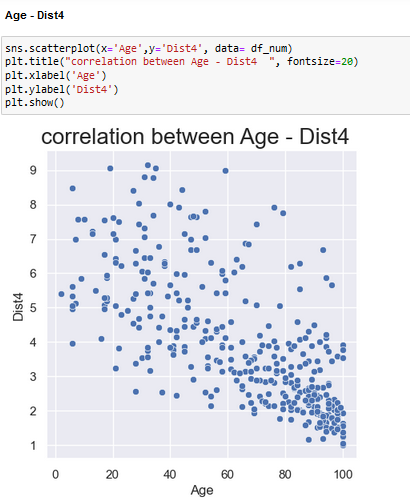
1. # Resid\_area - Dist4



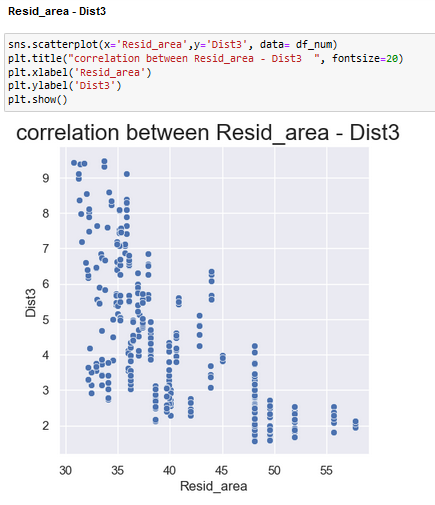
1. # Air\_qual - Dist4



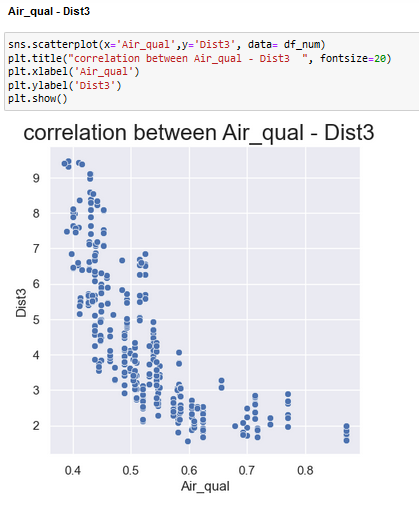
1. # Age - Dist4



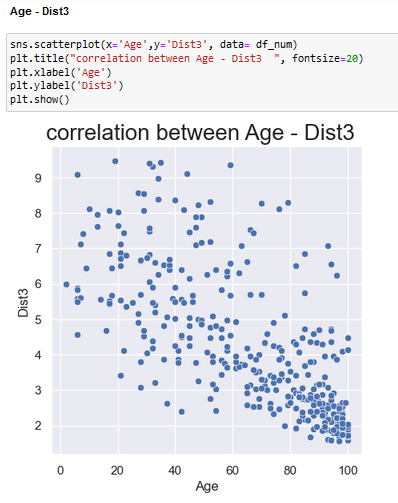
1. # Resid\_area - Dist3



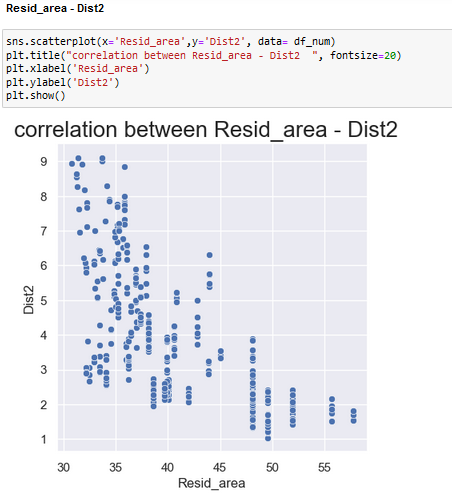
1. # Air\_qual - Dist3



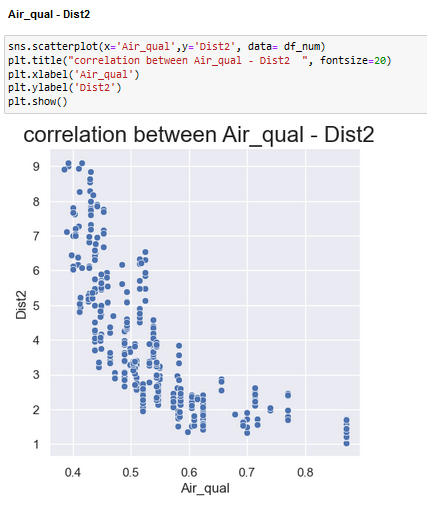
1. # Age - Dist3



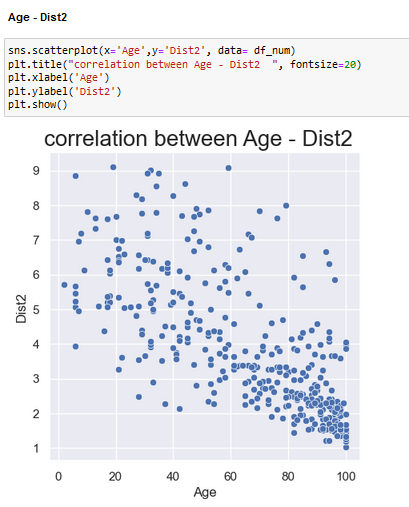
1. # Resid\_area - Dist2



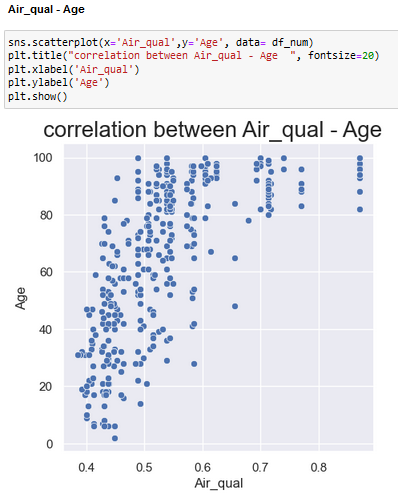
1. # Air\_qual - Dist2



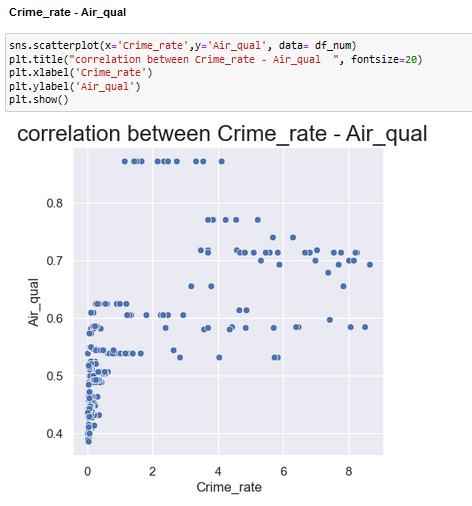
1. # Age - Dist2



1. # Air\_qual – Age



1. # Crime\_rate - Air\_qual



1. # Resid\_area - Air\_qual



Q4. Decide the target variable /attribute/column

**Answer** : After analysing all the pairs of the columns we will decide the target variable/attribute/column.

So, the target variables will be

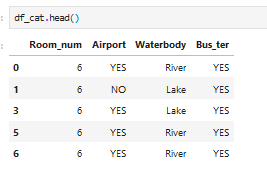
Parks

Price

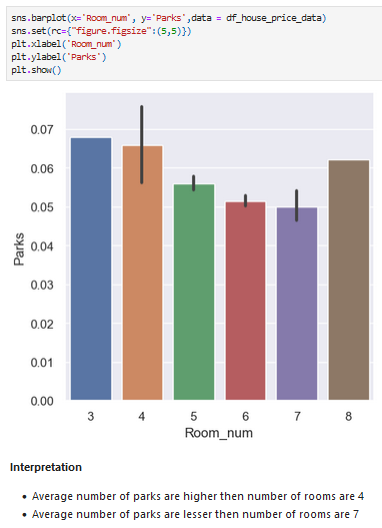
Air\_qual

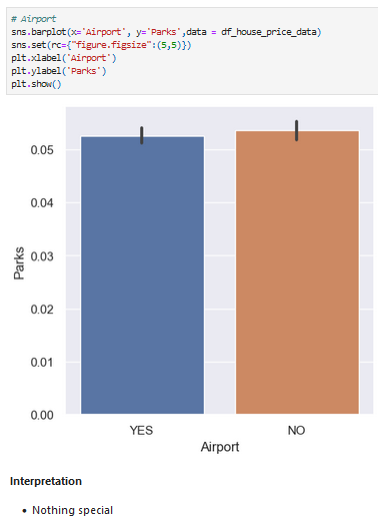
Poor\_prop

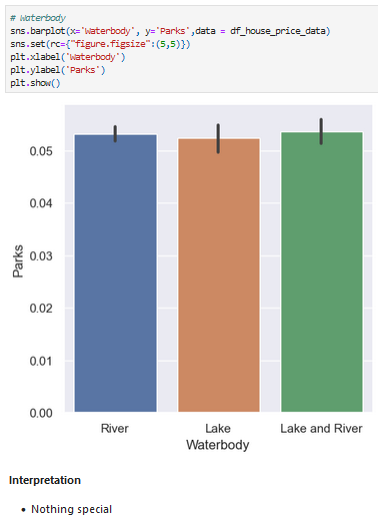
Now, check all this target variables with catetgorical values



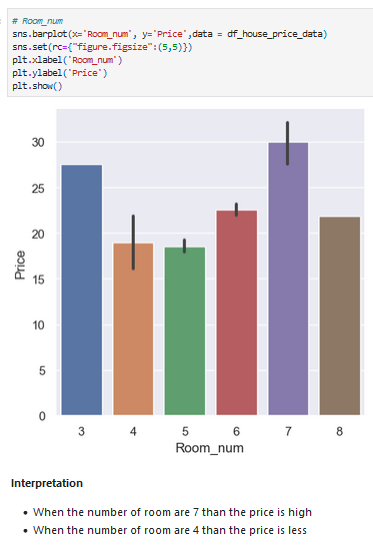
1. Parks

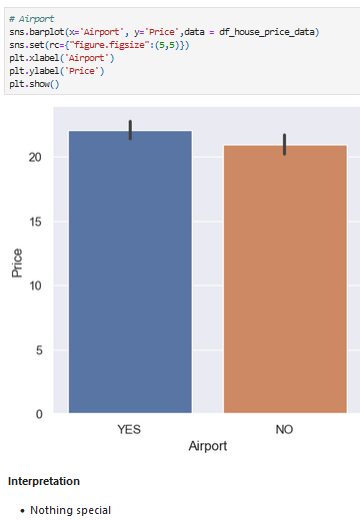


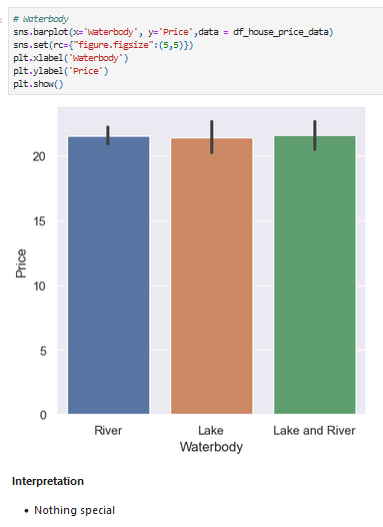




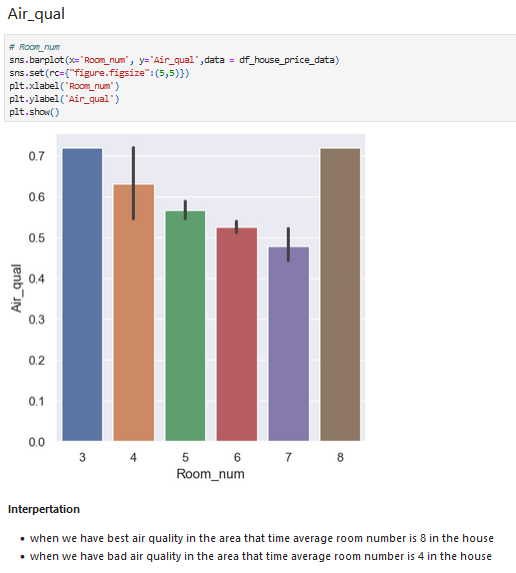
1. Price

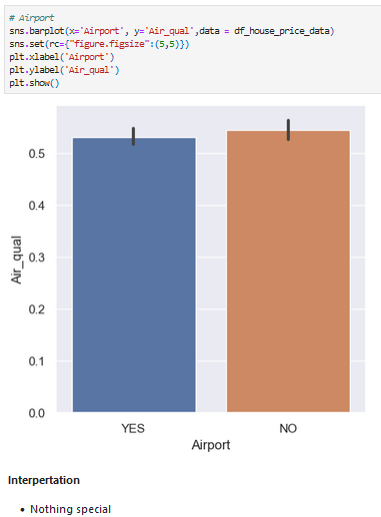


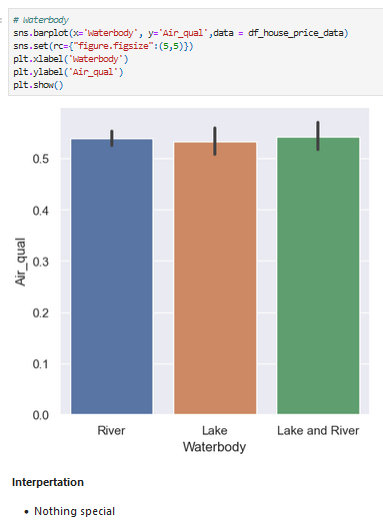


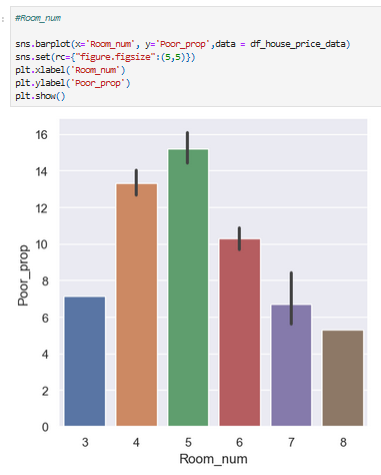


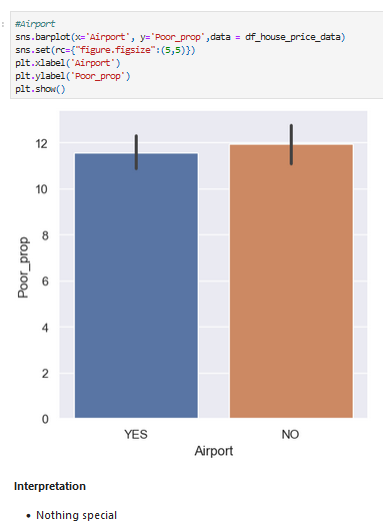
1. Air\_qual

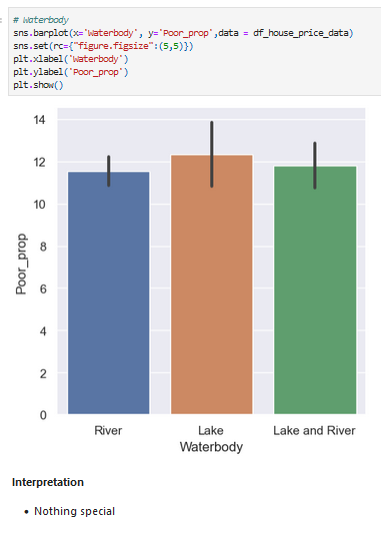






1. Poor\_prop  
   





# Multivariate Analysis

Q1. what is mean by multivariate analysis(Small paragraph)

**Answer** :

Multivariate analysis extends beyond bivariate analysis to include the simultaneous analysis of three or more variables. It aims to understand how multiple variables interact with and relate to each other. This type of analysis is used to explore more complex relationships and patterns within a dataset.

Q2. Methods we used to describe the multivariate analysis(small paragraph for each method)

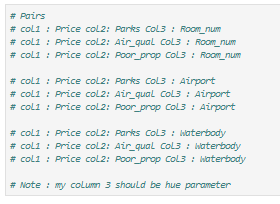
**Answer** : Both Cross Tabulation and Pivot Tables can be instrumental in multivariate analysis by providing a clear view of how multiple variables interact. Here’s how they fit into multivariate analysis:

* **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.

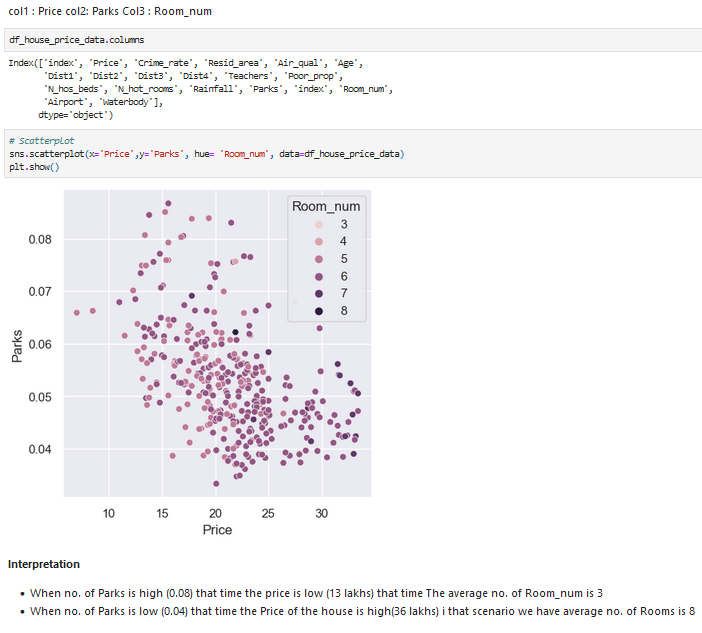
Q3. What we did with our data (screenshot of code,screenshot of the graph, inference column wise<in heading2>) (table ss)

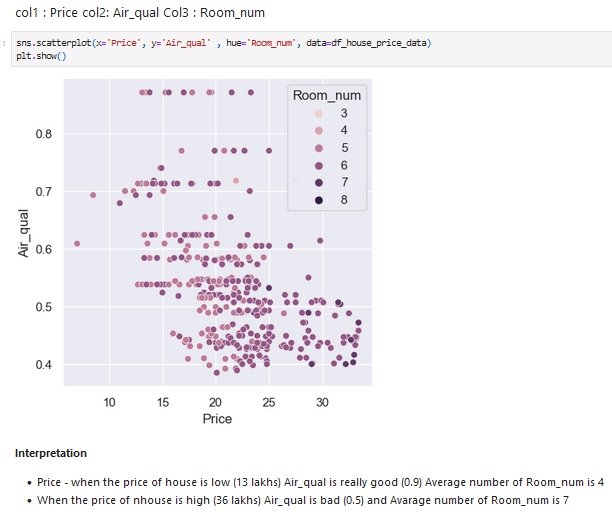
**Answer** : From pairs of target variables and categorical values make multiple columns pairs





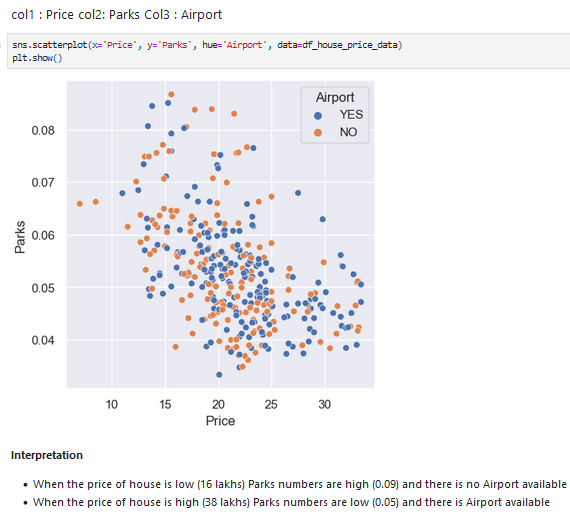
Room\_num



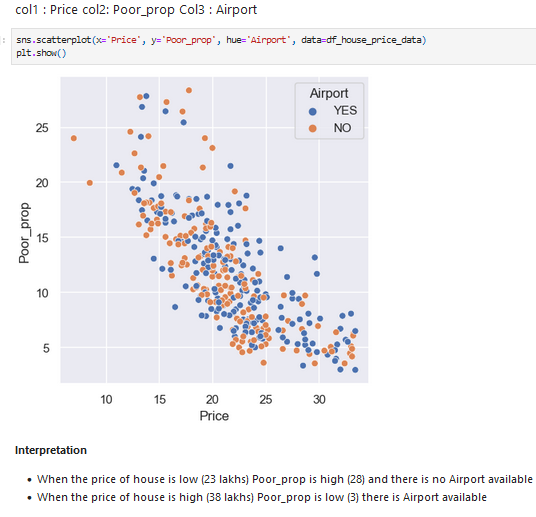




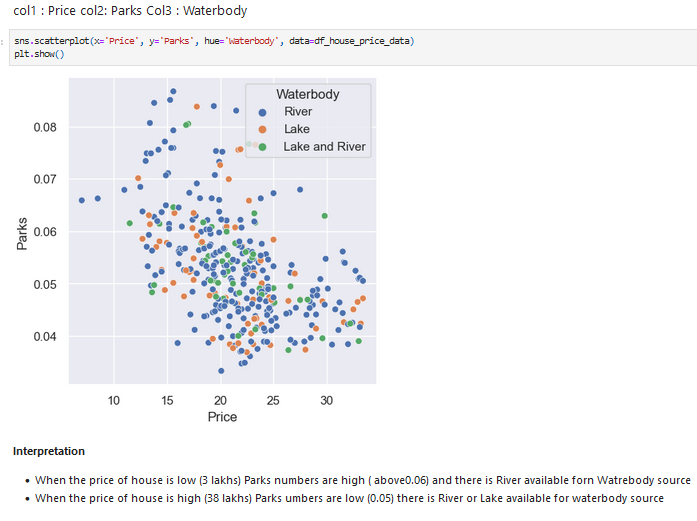
Airport

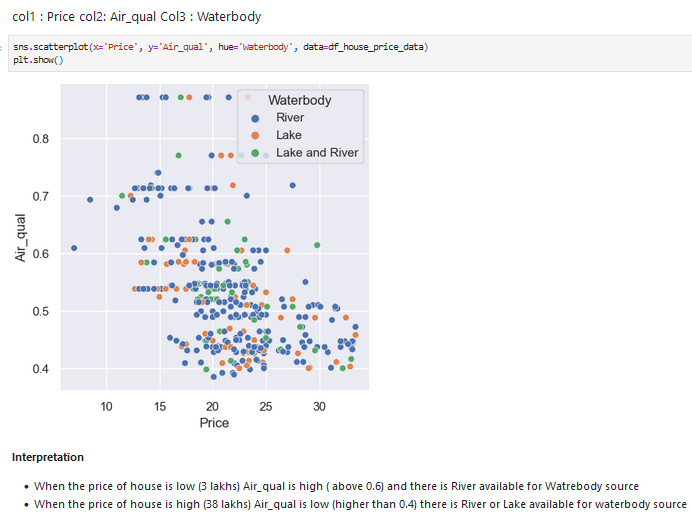


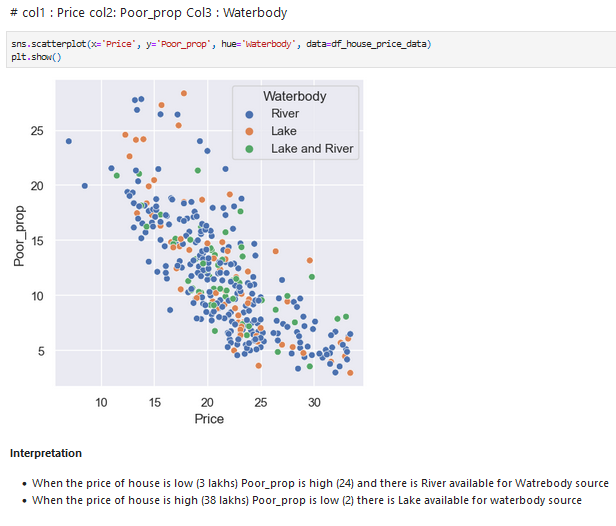




Waterbody



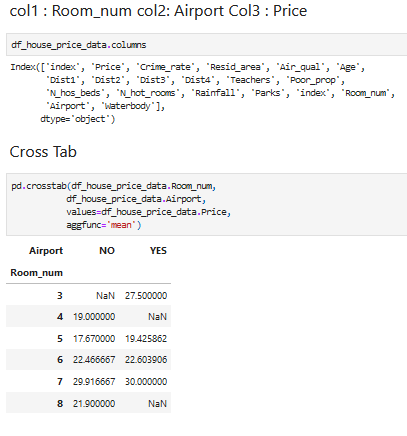


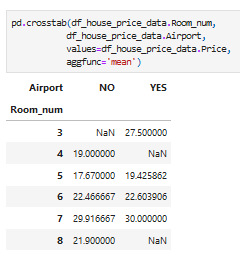


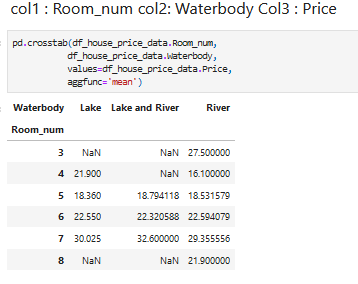
### Now Arrangement of pivot and cross tab

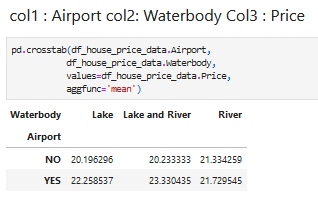


### Cross Tab









Summary

The report analyzes housing-related data to aid builders in understanding and predicting housing prices, develop policies, and identify influencing factors. The data encompasses attributes like price, crime rate, residential area, air quality, number of rooms, age, distance to amenities, and more. Methodologically, it includes data cleaning, missing value treatment following industry rules, and outlier handling using the Interquartile Range method. Exploratory Data Analysis (EDA) covers univariate, bivariate, and multivariate analyses to uncover insights. This structured approach guides informed decision-making for housing pricing and policy development.