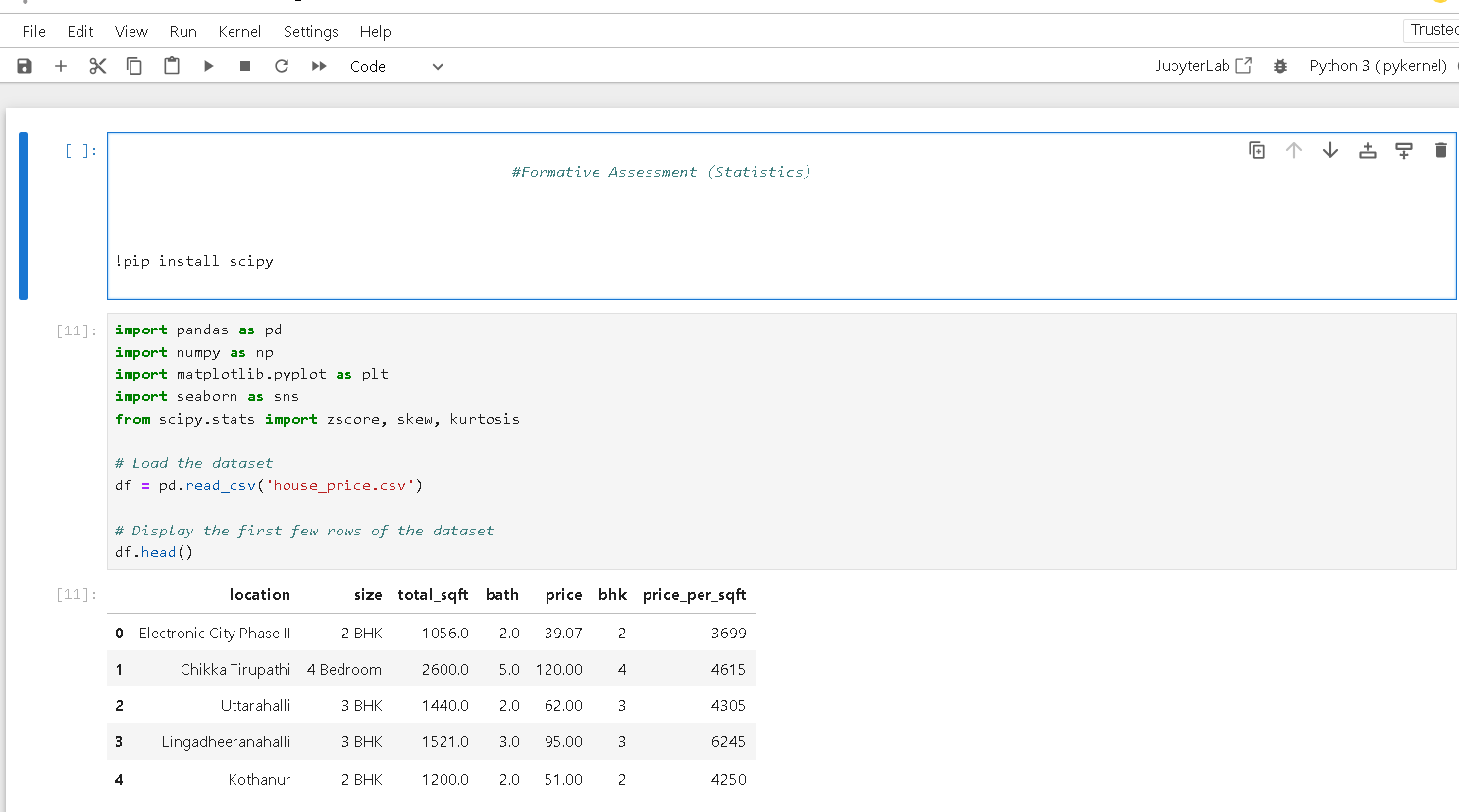
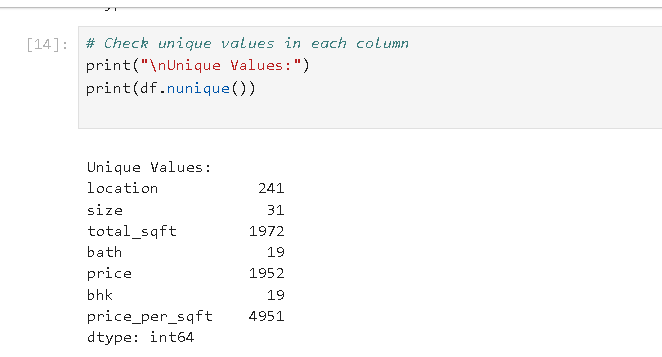
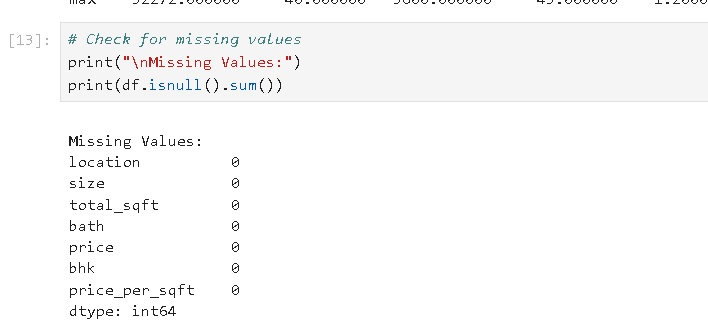
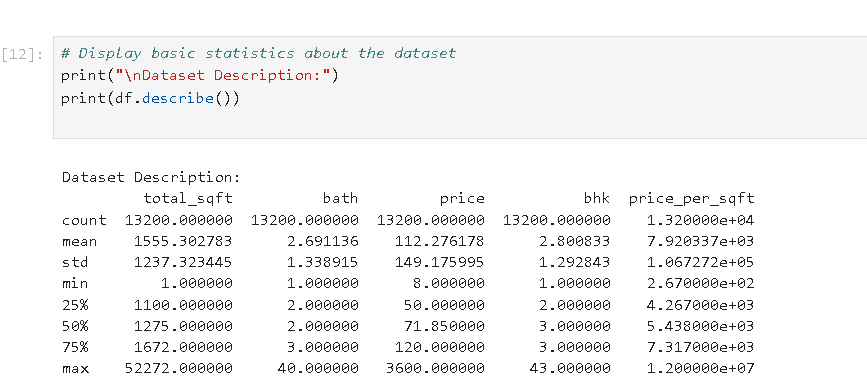
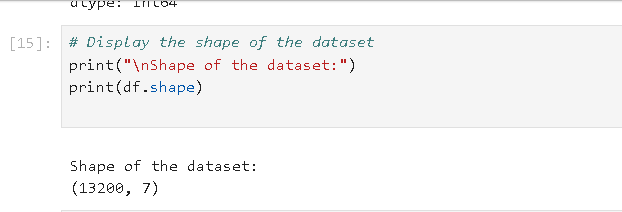
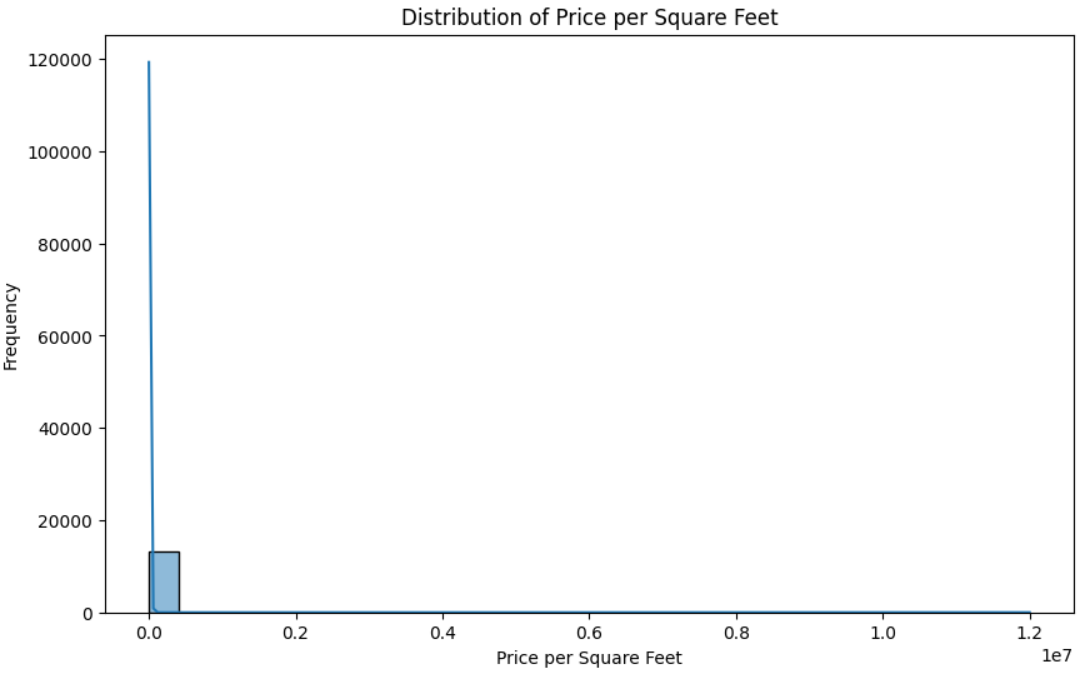
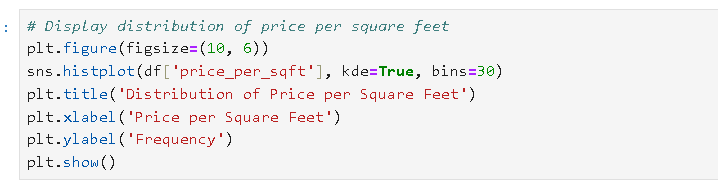
Formative Assessment (Statistics)

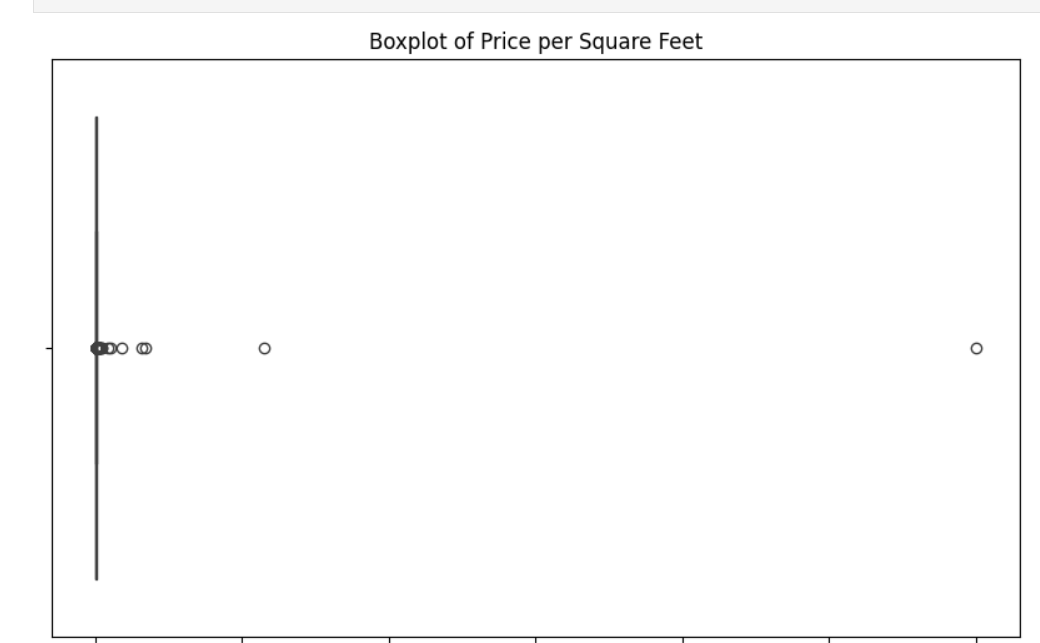
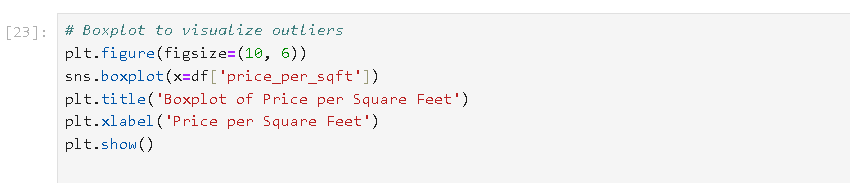
You are given house\_price.csv which contains property prices in the city of Bangalore. You need to examine price per square feet do the following:  
  
**Q1**. Perform basic EDA (**Score:1**)

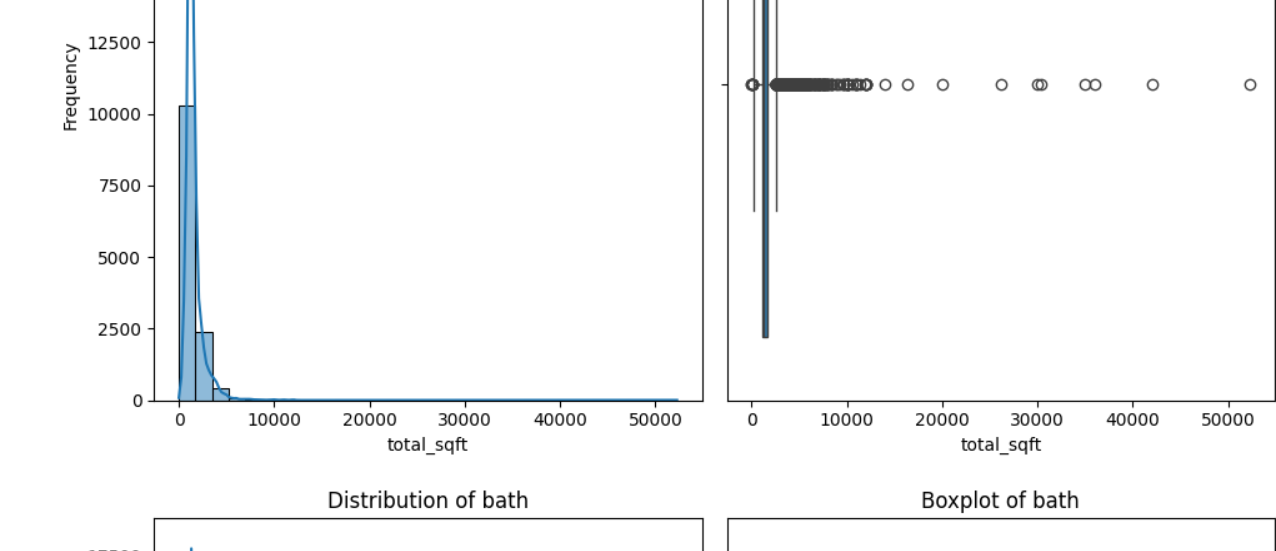
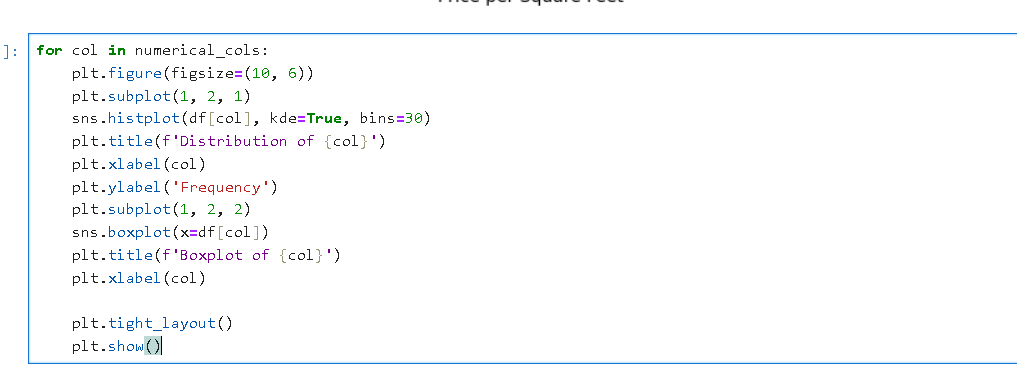








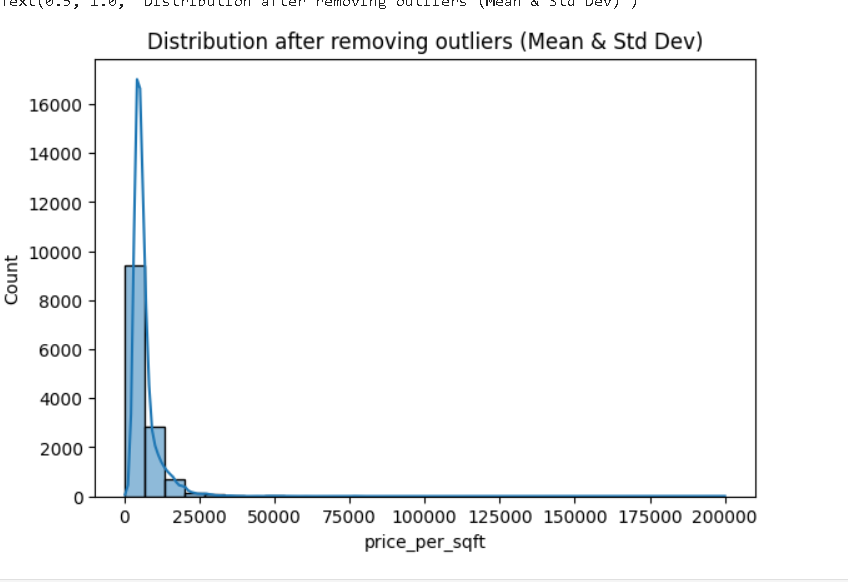
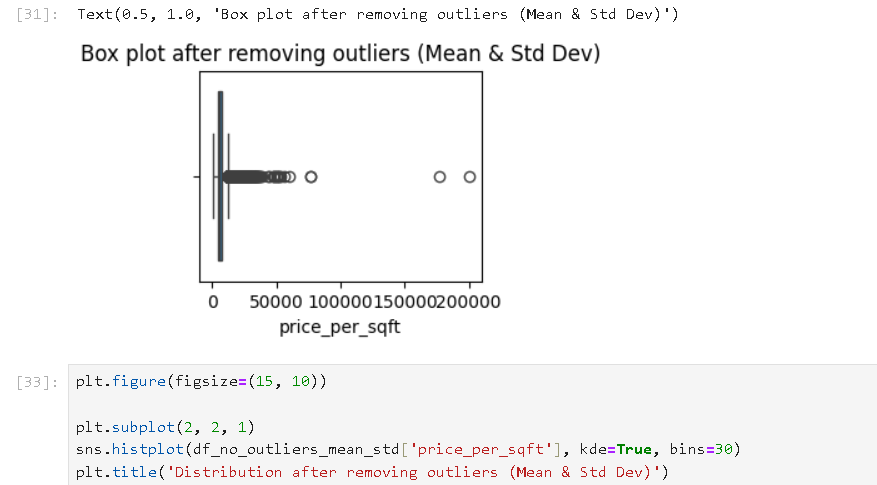
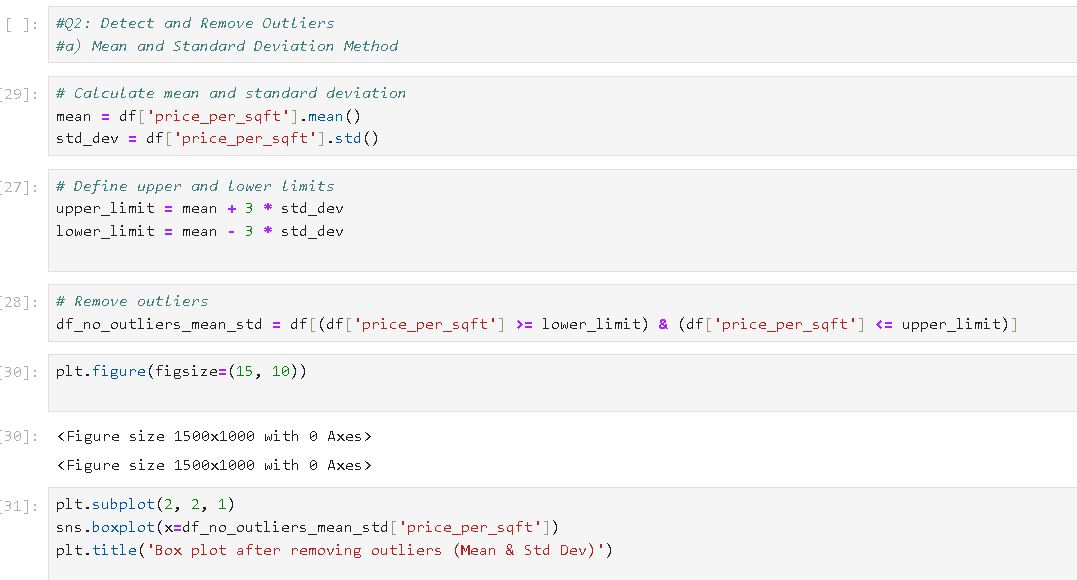


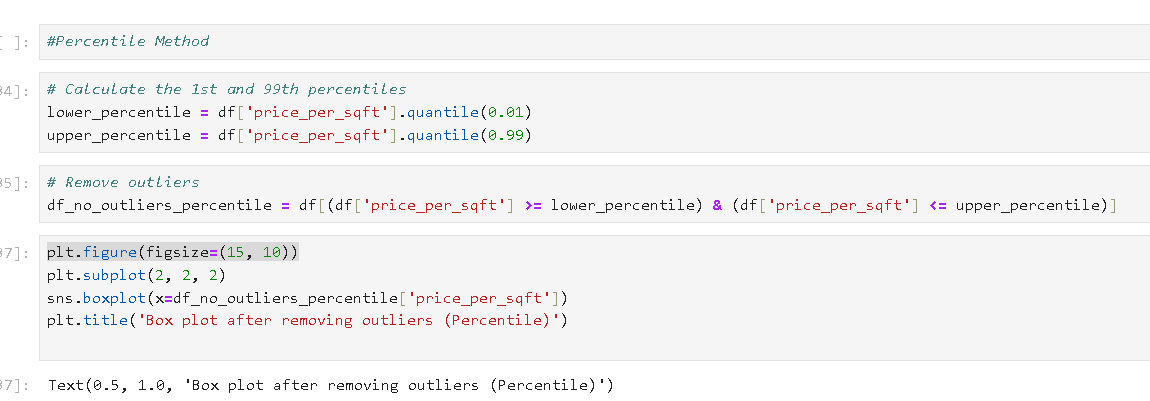


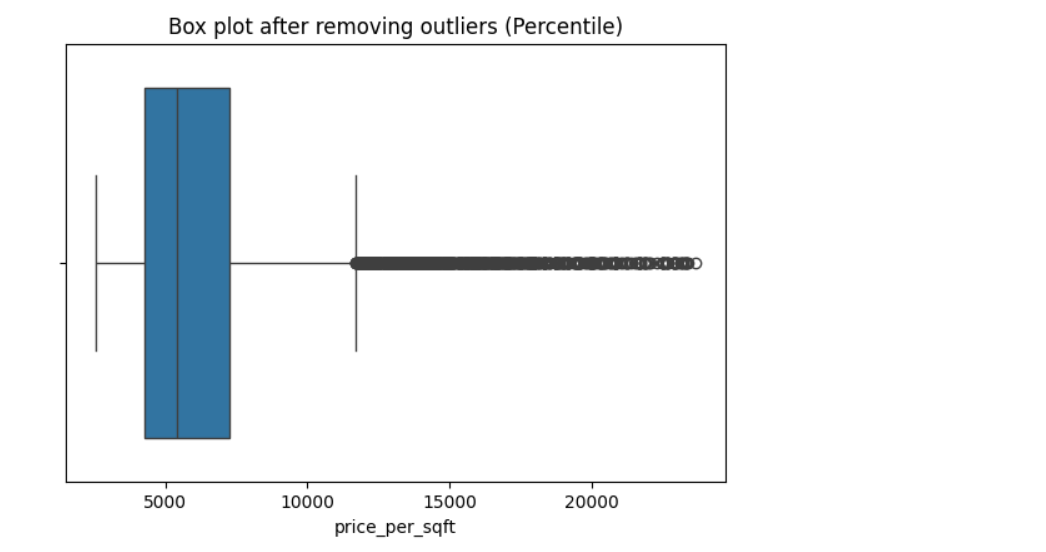
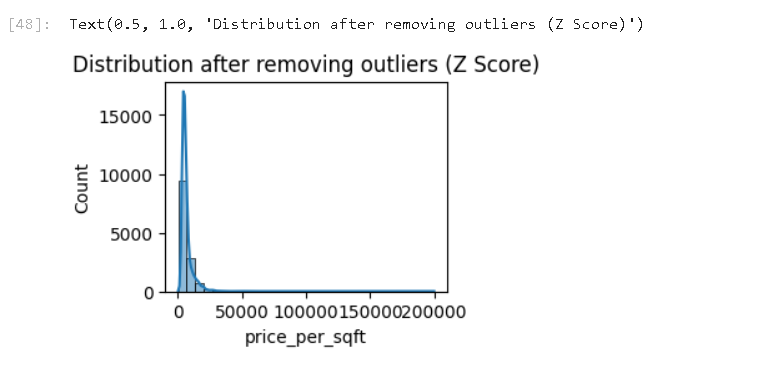
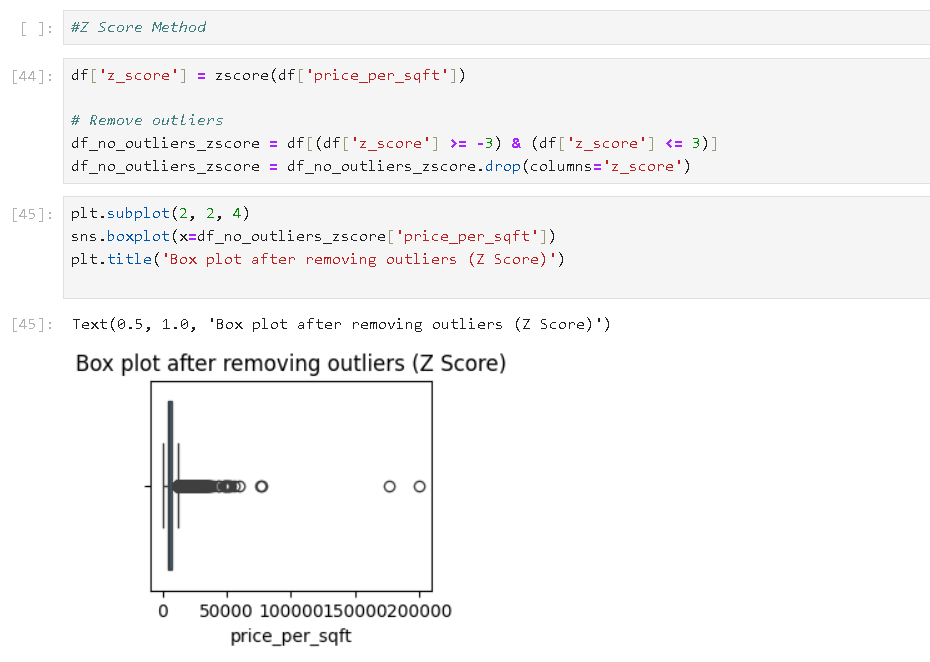
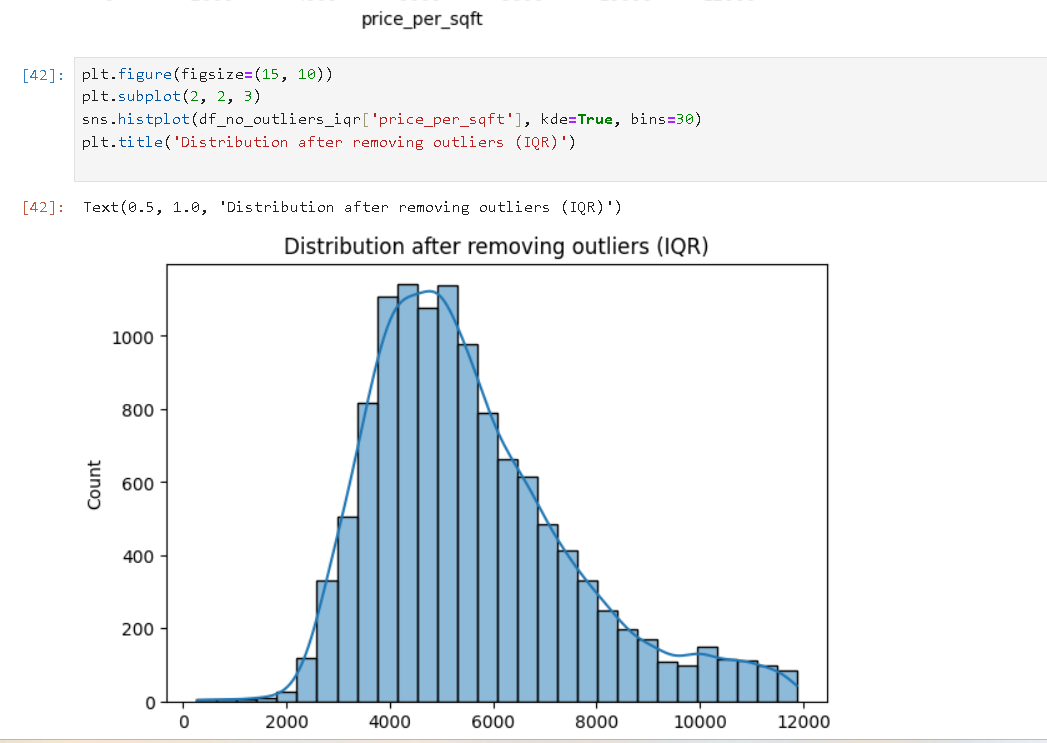
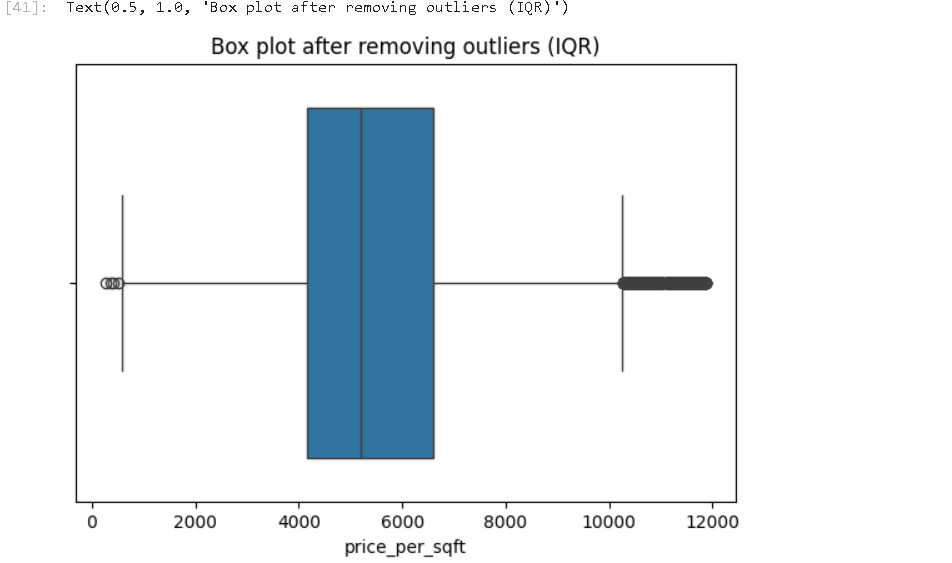
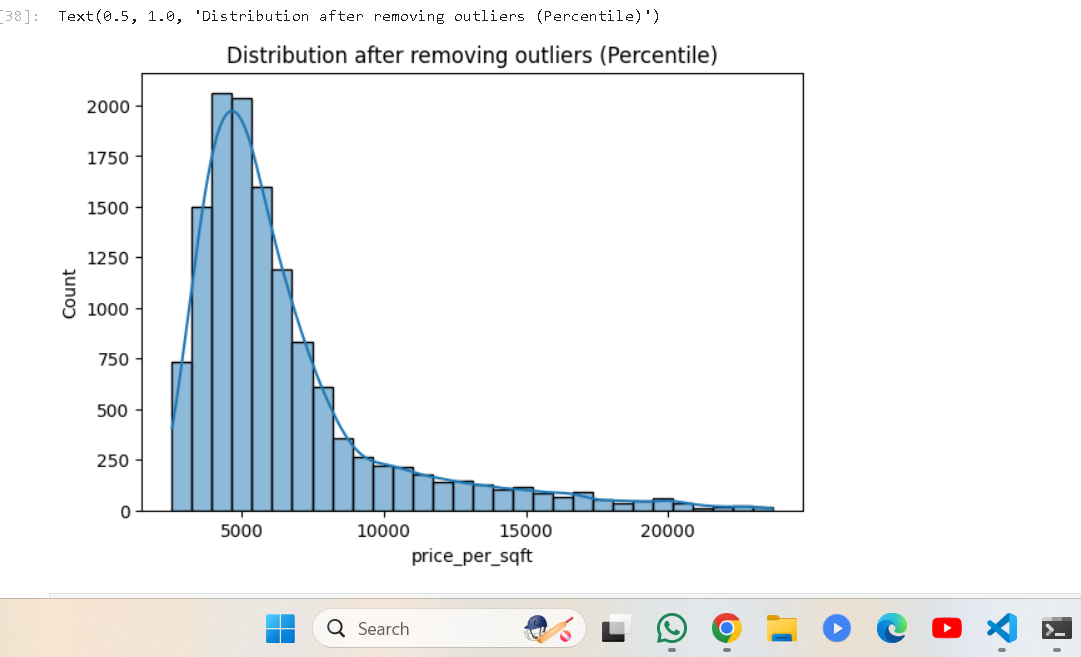
This code covers:

* Viewing the first and last few rows of the dataset.
* Checking the shape and column names of the dataset.
* Identifying numerical and categorical columns.
* Checking for duplicate rows.
* Displaying basic statistics of numerical columns.
* Creating histograms and box plots for the price\_per\_sqft column.
* Creating histograms and box plots for all numerical columns.

**Q2**. Detect the outliers using following methods and remove it using methods like trimming / capping/ imputation using mean or median (**Score: 4**)  
a) Mean and Standard deviation  
b)Percentile method  
c) IQR(Inter quartile range method)  
d) Z Score method



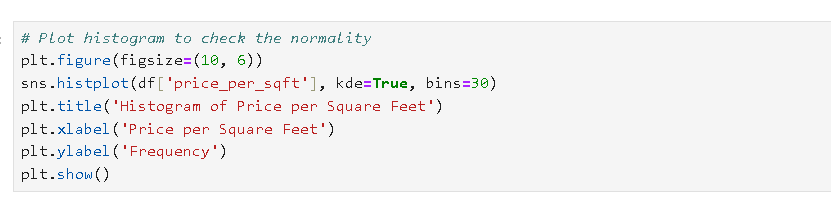
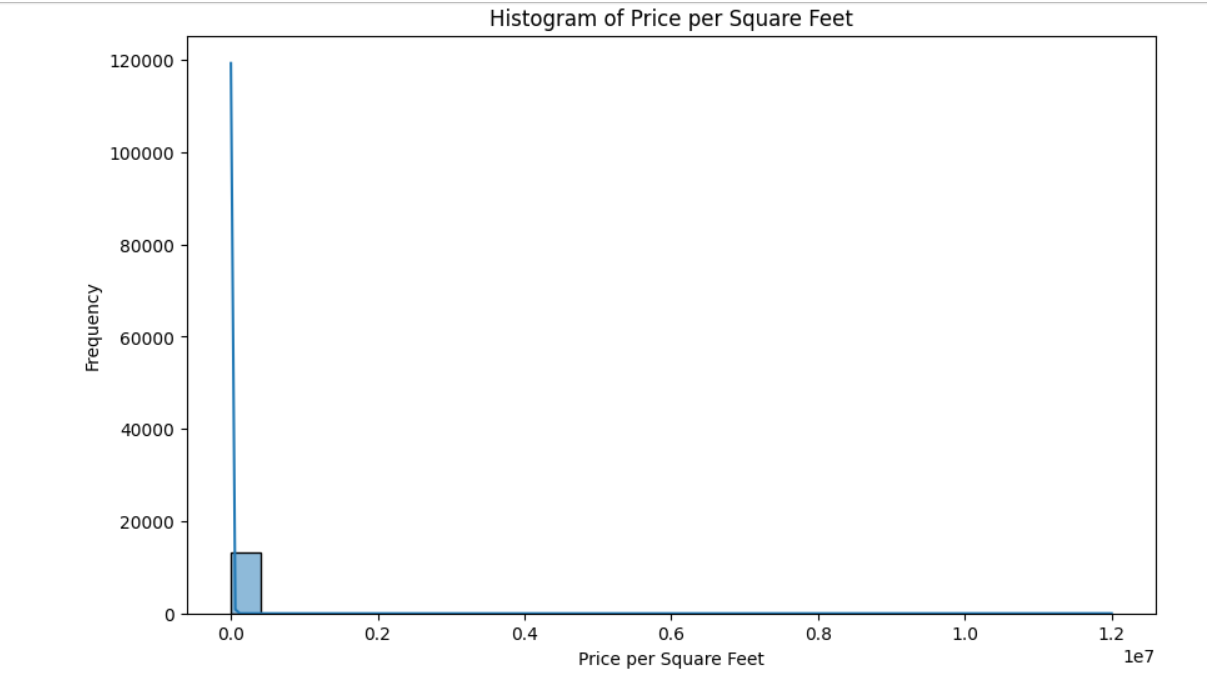
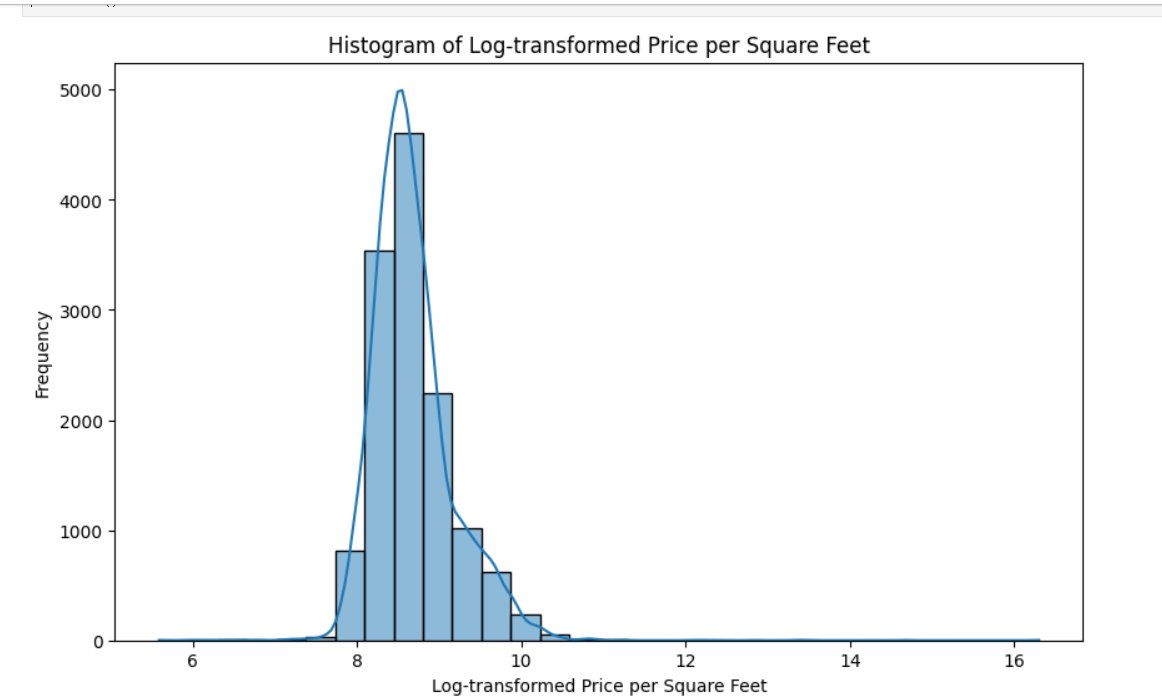
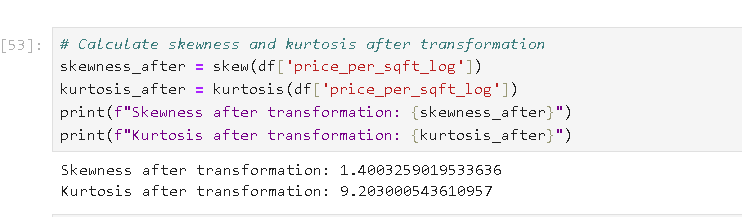


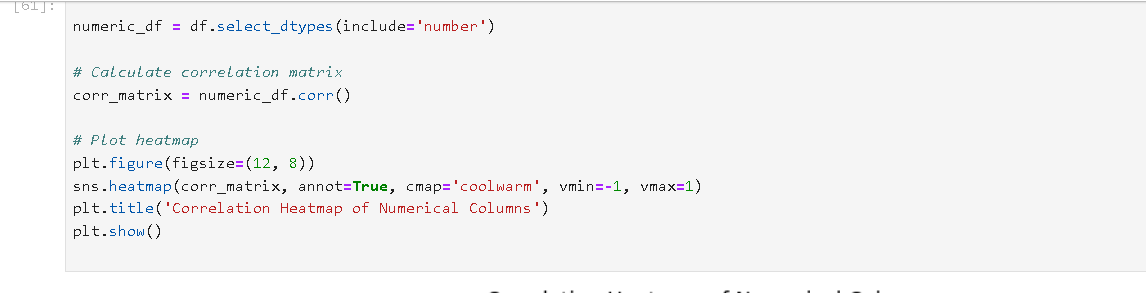
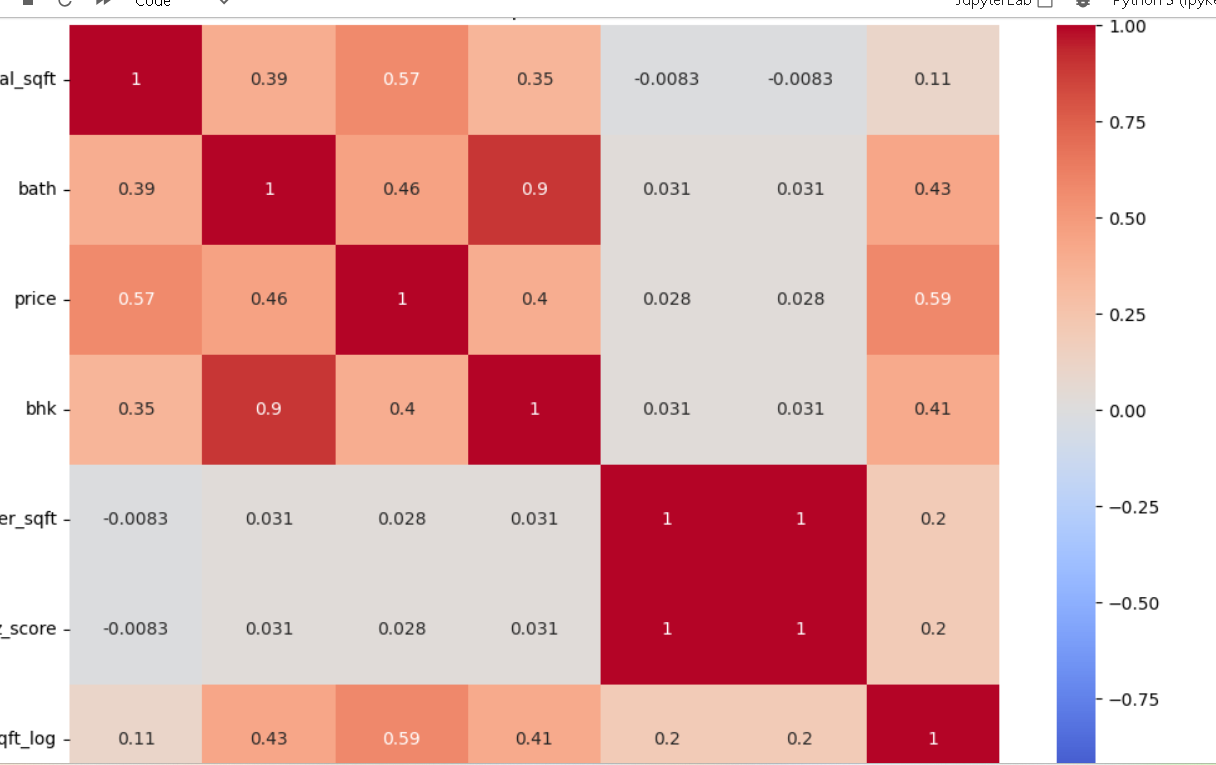
****  
**Q3**. Create a box plot and use this to determine which method seems to work best to remove outliers for this data? (**Score:1**)

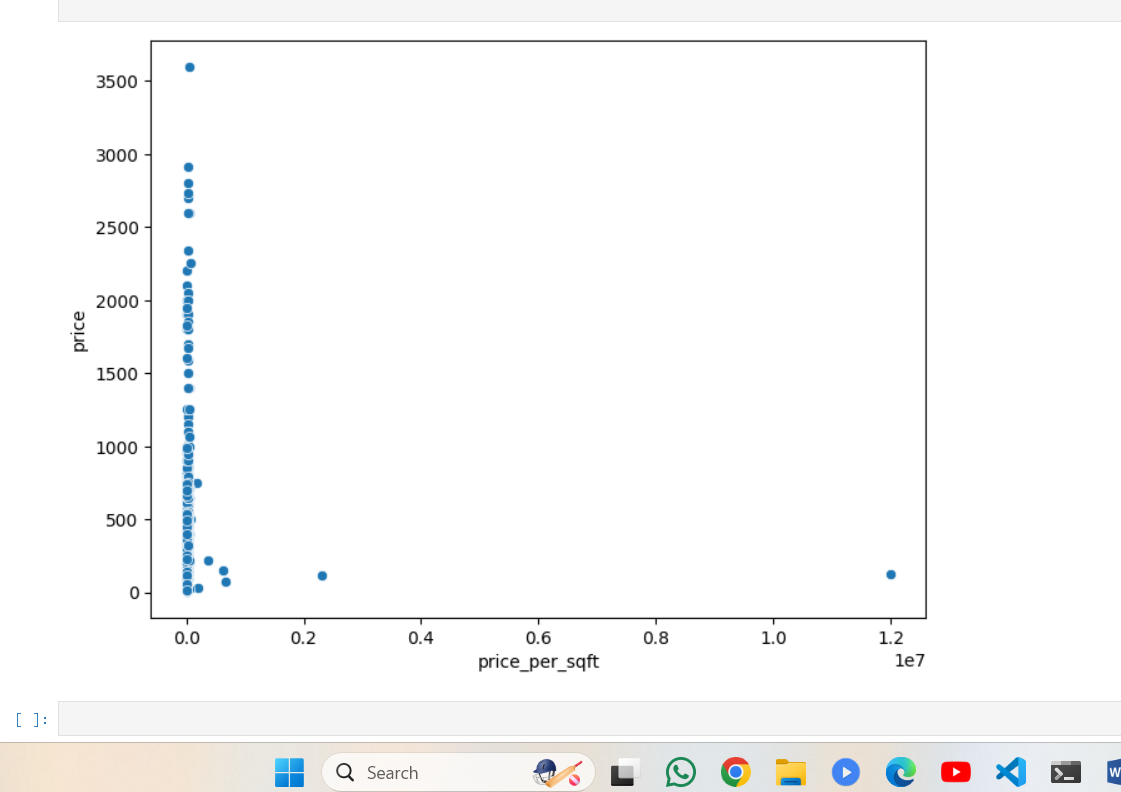
### Analysis:

* **Mean and Standard Deviation Method**: If the box plot still shows significant outliers, this method might not be effective for this dataset.
* **Percentile Method**: By removing the extreme 1% of data on both ends, this method can effectively remove outliers but might also remove valid data points.
* **IQR Method**: This method typically works well if the data follows a normal distribution, as it removes points outside the typical spread of the middle 50% of data.
* **Z Score Method**: This method standardizes data and removes points that are more than 3 standard deviations from the mean. It works well for normally distributed data.

**Q4**. Draw histplot to check the normality of the column(price per sqft column) and perform transformations if needed. Check the skewness and kurtosis before and after the transformation. (**Score:1**)

      
  
**Q5**. Check the correlation between all the numerical columns and plot heatmap. (**Score:1**)

   
  
**Q6**. Draw Scatter plot between the variables to check the correlation between them. (**Score:1**)

   
  
**Timely Submission**(**Score:1**)  
**Total Marks : 10**  
  
**Dataset :**<https://drive.google.com/file/d/1UlWRYU0UglE2ex3iFse0J6eCLEU8g98P/view?usp=sharing>