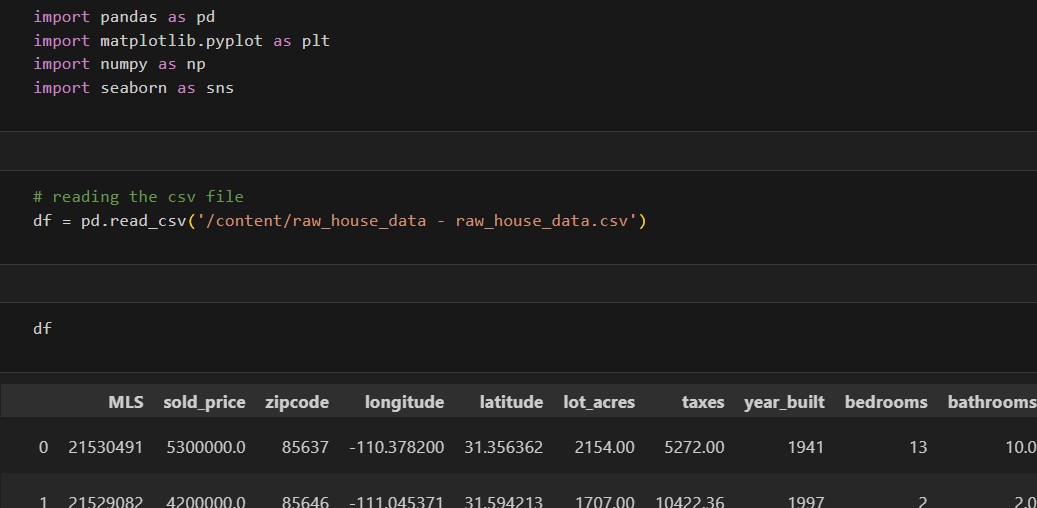
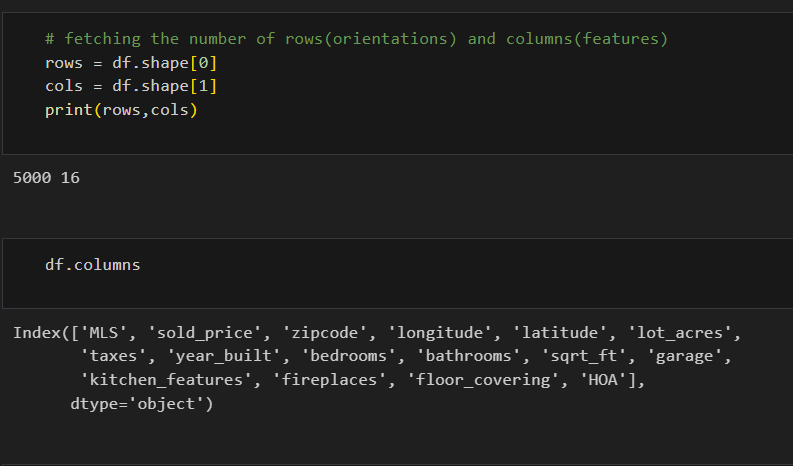
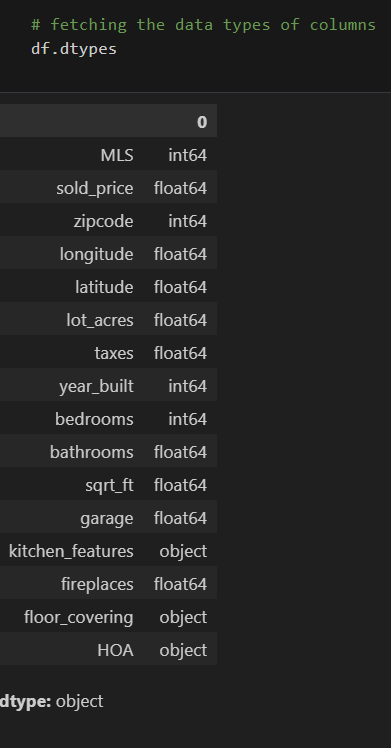
**EDA for raw dataset**

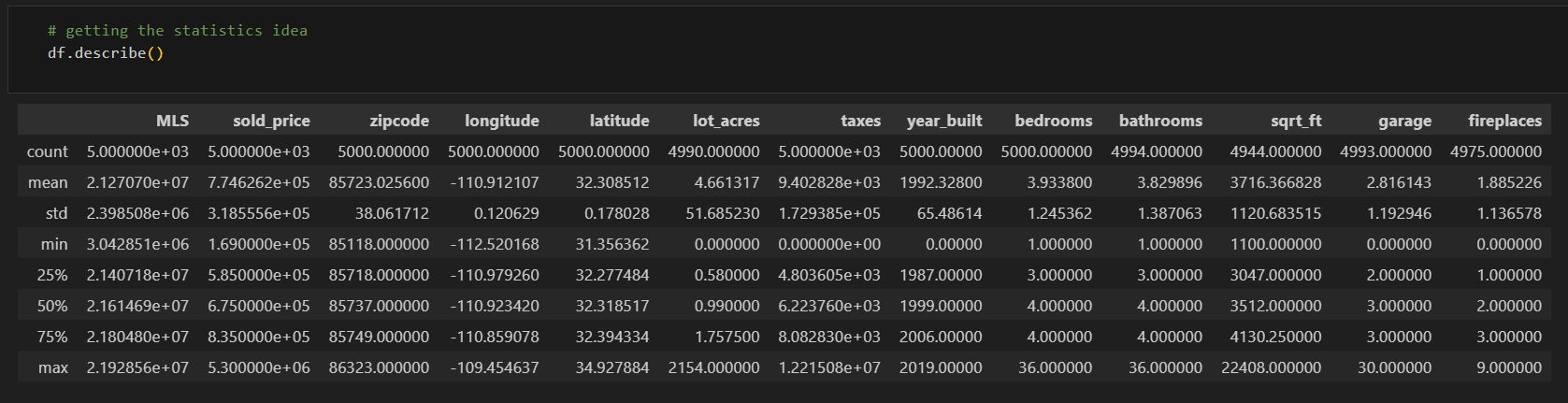
After reading this raw dataset noticed it has 5000 observations and 16 features. Please find the snapshots below:





Done the data types analysis and describing the statistics for more clarity:

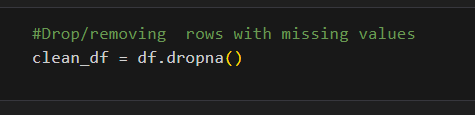




Checking for the missing values:



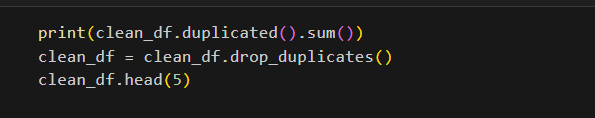
Dropping the rows with missing values:



After dropping missing values, the observations and features are:

4370 rows × 16 columns

Checked if any duplicate items present, then dropped the duplicates ones:



For this dataset no duplicate items found.

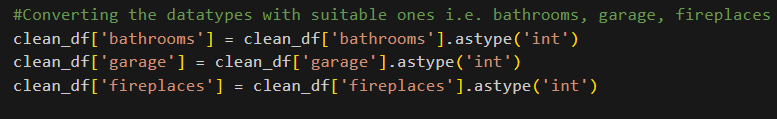
As per my analysis of features, I observed some of them were not in appropriate data types.

Updated the data types for mentioned features:

Bathrooms (float type to int type)

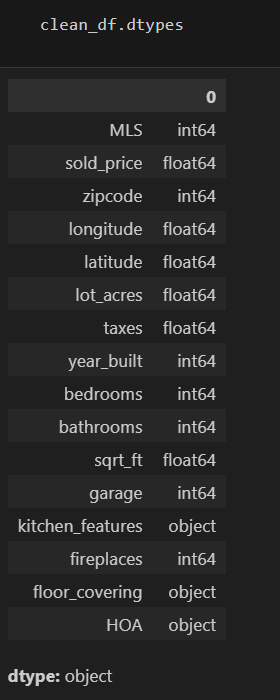
Garage (float type to int type)

Fireplaces (float type to int type)

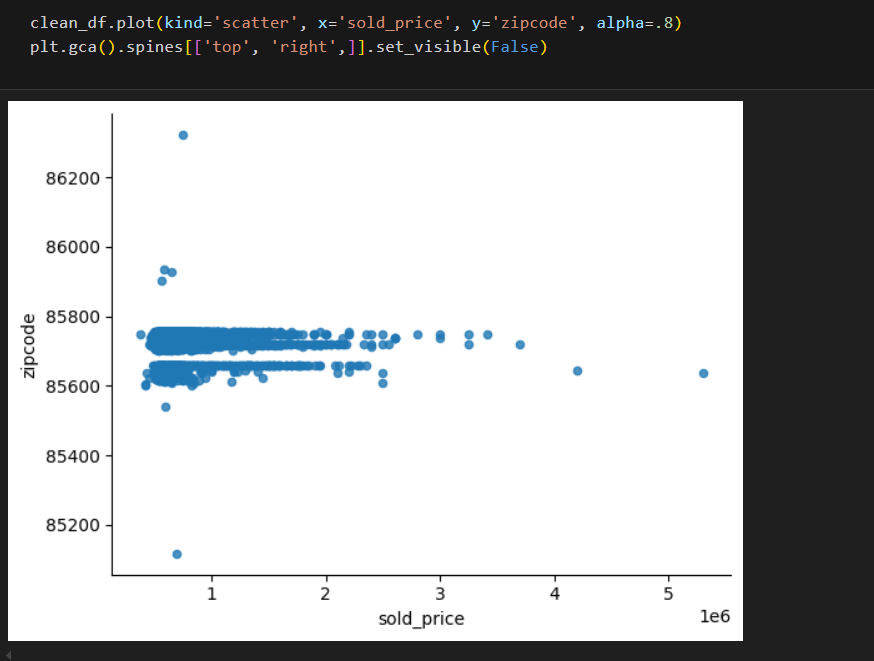


Made these changes for further calculation purposes when needed to increase the efficiency.

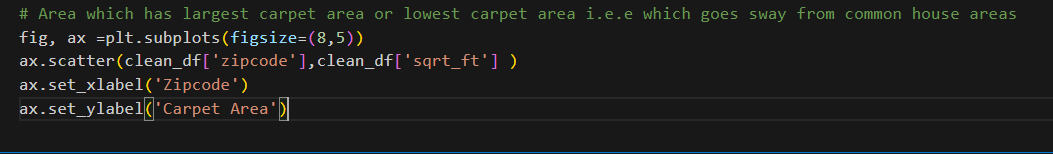
Now after changing the types here is the current data types for data frame:

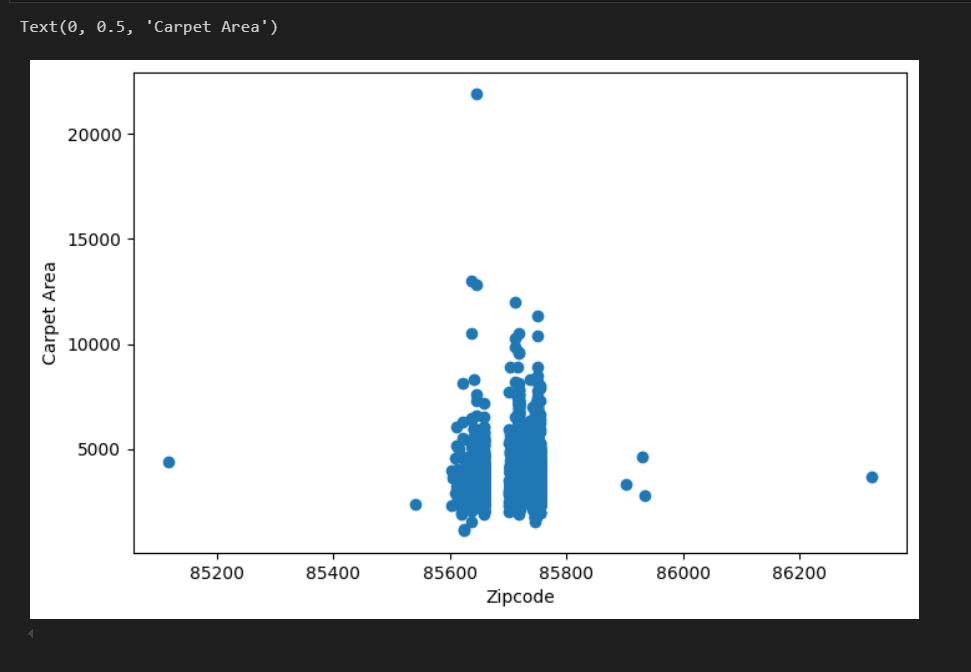


For estimation purpose created scattered graph between sold\_price and zipcode. Which areas has maximum properties sold. And which among them are the outlier one.

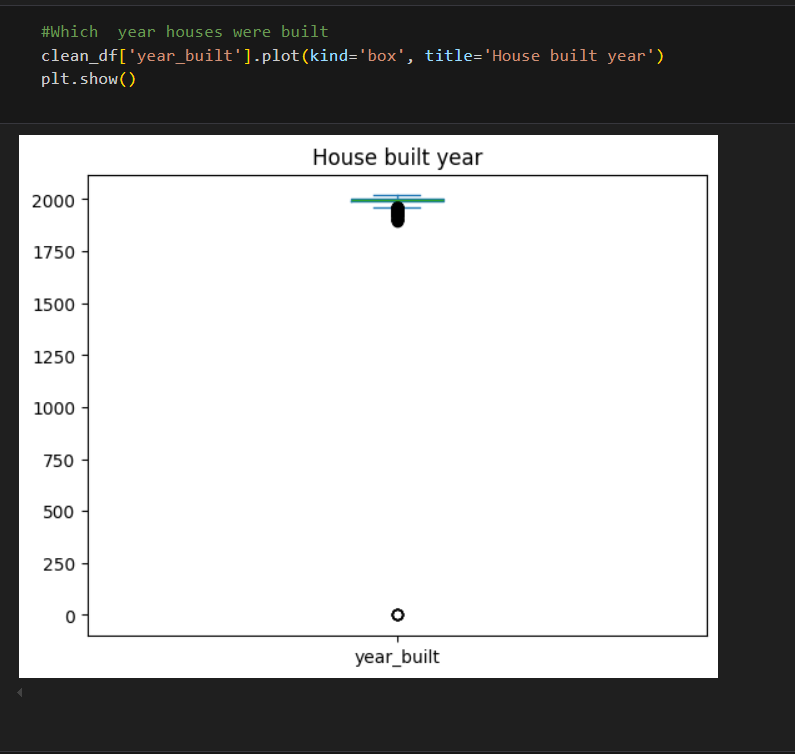


Area which has largest carpet area or lowest carpet area. To find out which areas has largest carpet area for selling purposes or investing purposes.

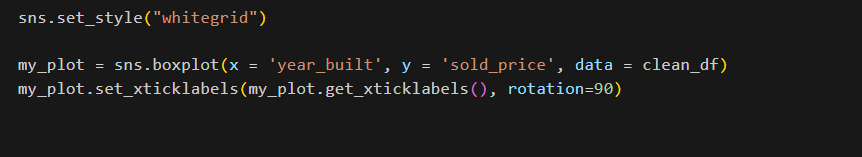


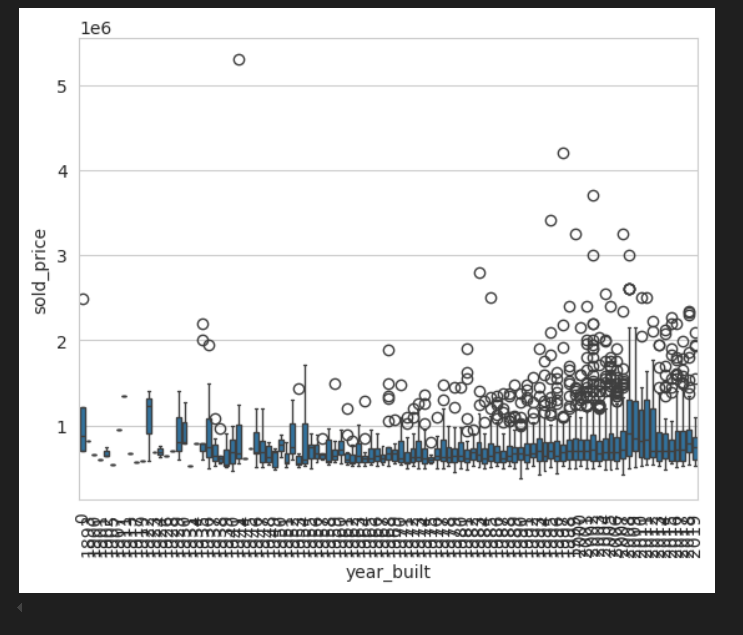


Another box plot chart for estimation of which year the houses were built majorly:



Another box plot chart between sold\_price and year\_built , to find out which year had the most people purchasing the houses.





At the last converted the clean data frame to csv format and saved it in local.

