

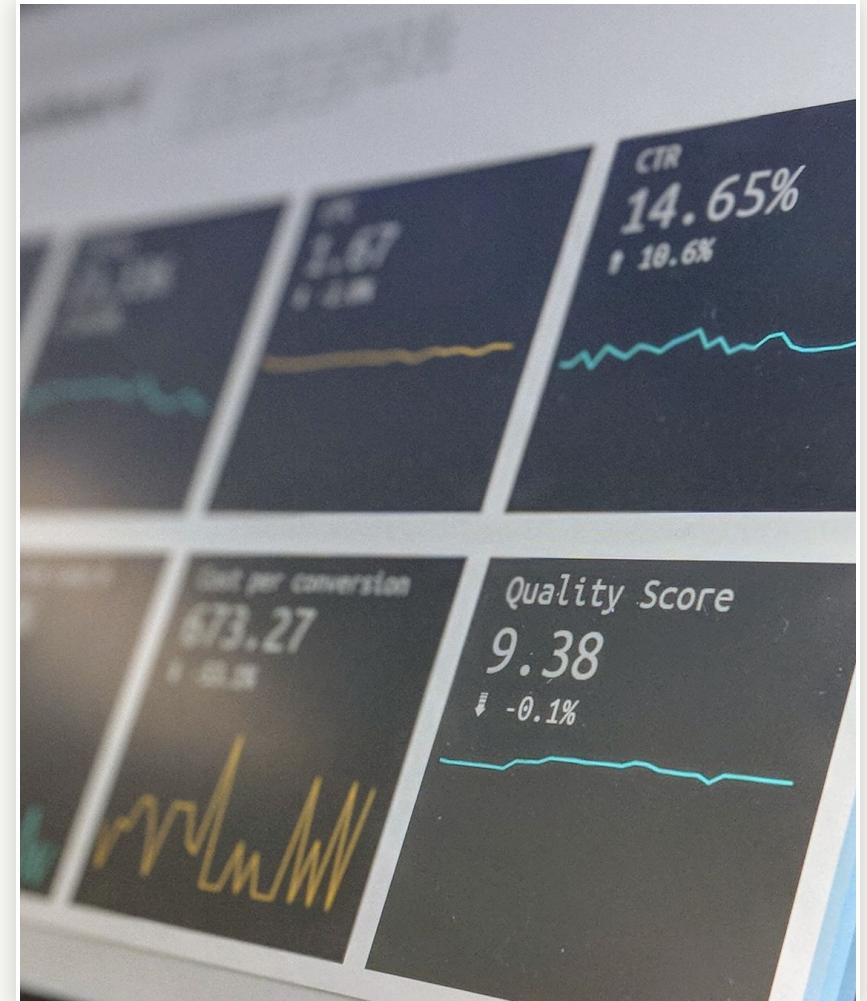


# Assignment 1

Mike Solis

# The data set

- 16 columns
- Types: Int, float and object
- 5000 instances

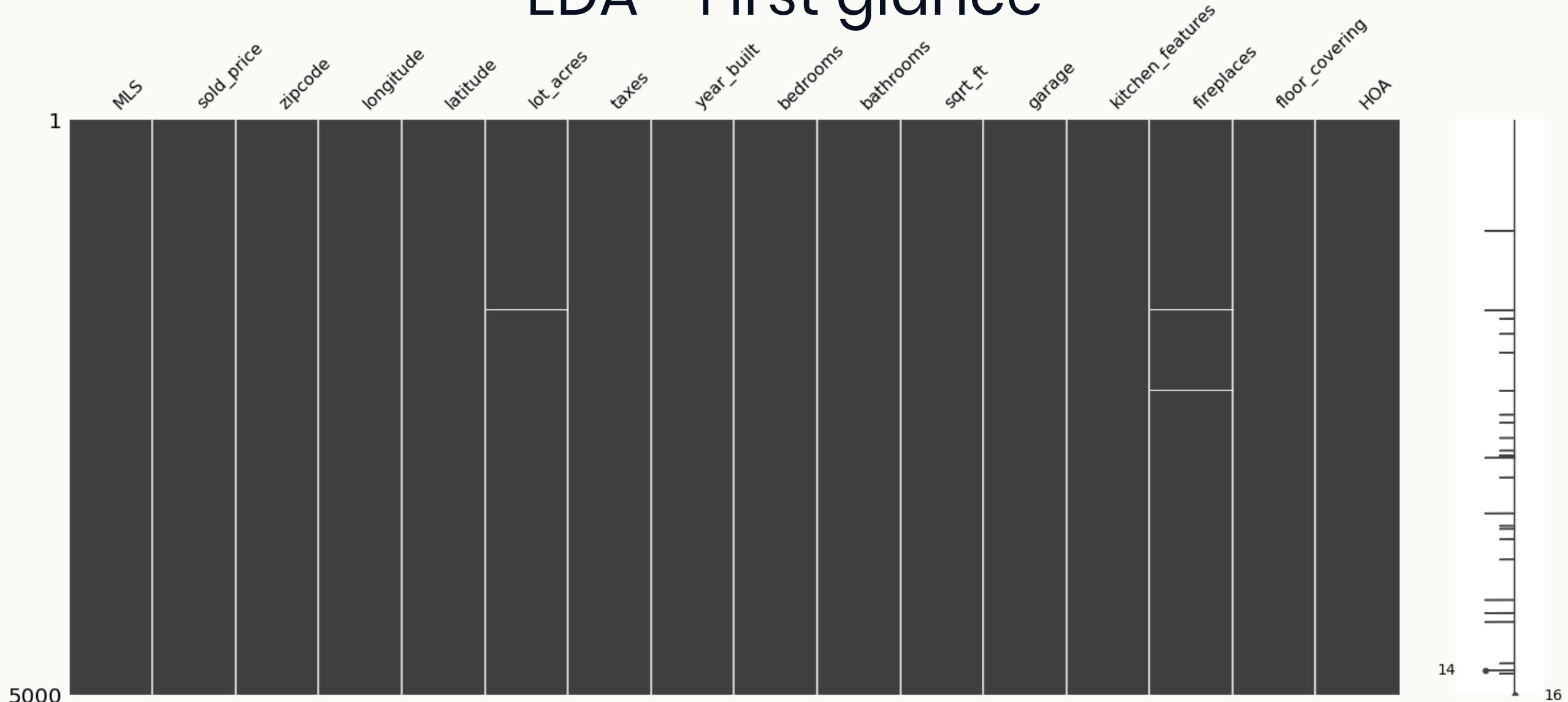


# EDA – First glance

```
MLS           int64
sold_price    float64
zipcode       int64
longitude     float64
latitude      float64
lot_acres     float64
taxes         float64
year_built    int64
bedrooms      int64
bathrooms     object
sqrt_ft       object
garage        object
kitchen_features object
fireplaces    float64
floor_covering object
HOA           object
dtypes: object
```

```
MLS           0
sold_price    0
zipcode       0
longitude     0
latitude      0
lot_acres     10
taxes         0
year_built    0
bedrooms      0
bathrooms     0
sqrt_ft       0
garage        0
kitchen_features 0
fireplaces    25
floor_covering 0
HOA           0
```

# EDA – First glance



# Interesting!

- 1 There are some 'None' values
- 2 These 'None' values are considered as strings in the object type columns
- 3 Some of these object types should be integers
- 4 Some others should be floats
- 5 Some others should be strings





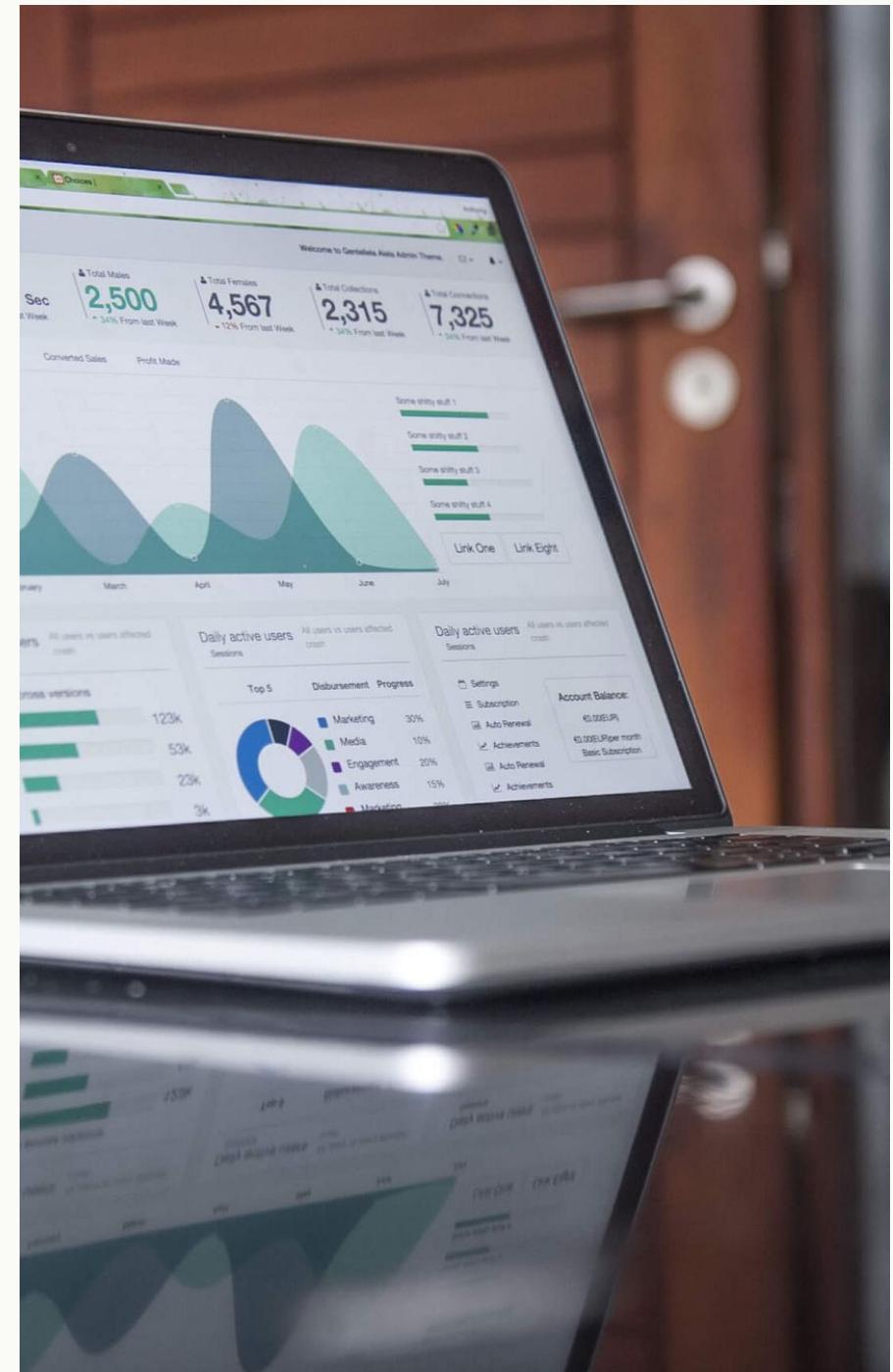
## ✓ Preliminary Data Exploration

## ✗ Data Cleaning

- Changing object columns to string, int or float depending the case.
- Delete rows and/or Predict missing values

## ✗ More Data Exploration

- Correlation
- Distribution



# Object type columns

The columns who has object type are the following:

bathrooms  
sqrt\_ft  
garage  
kitchen\_features  
floor\_covering  
HOA

# Some Things To Consider

In the 'floor\_covering' column, there are two strings to consider as nan: 'None' and 'Other: None'. In this project both will be considered as nan due to both means the same.

In the 'kitchen\_features' column there are 'None' values and actual string values who have the 'none' status in microwave, so only the first one will be consider as nan since in the second case 'none' is considered part of the kitchen information.

Object columns who have string and int values have any extraordinary situation and will be considered as their respective values after replacing 'None' for nan.

# Missing Values Comparison

New one

MLS	0
sold_price	0
zipcode	0
longitude	0
latitude	0
lot_acres	10
taxes	0
year_built	0
bedrooms	0
bathrooms	6
sqrt_ft	56
garage	7
kitchen_features	33
fireplaces	25
floor_covering	2
HOA	562
other_features	0

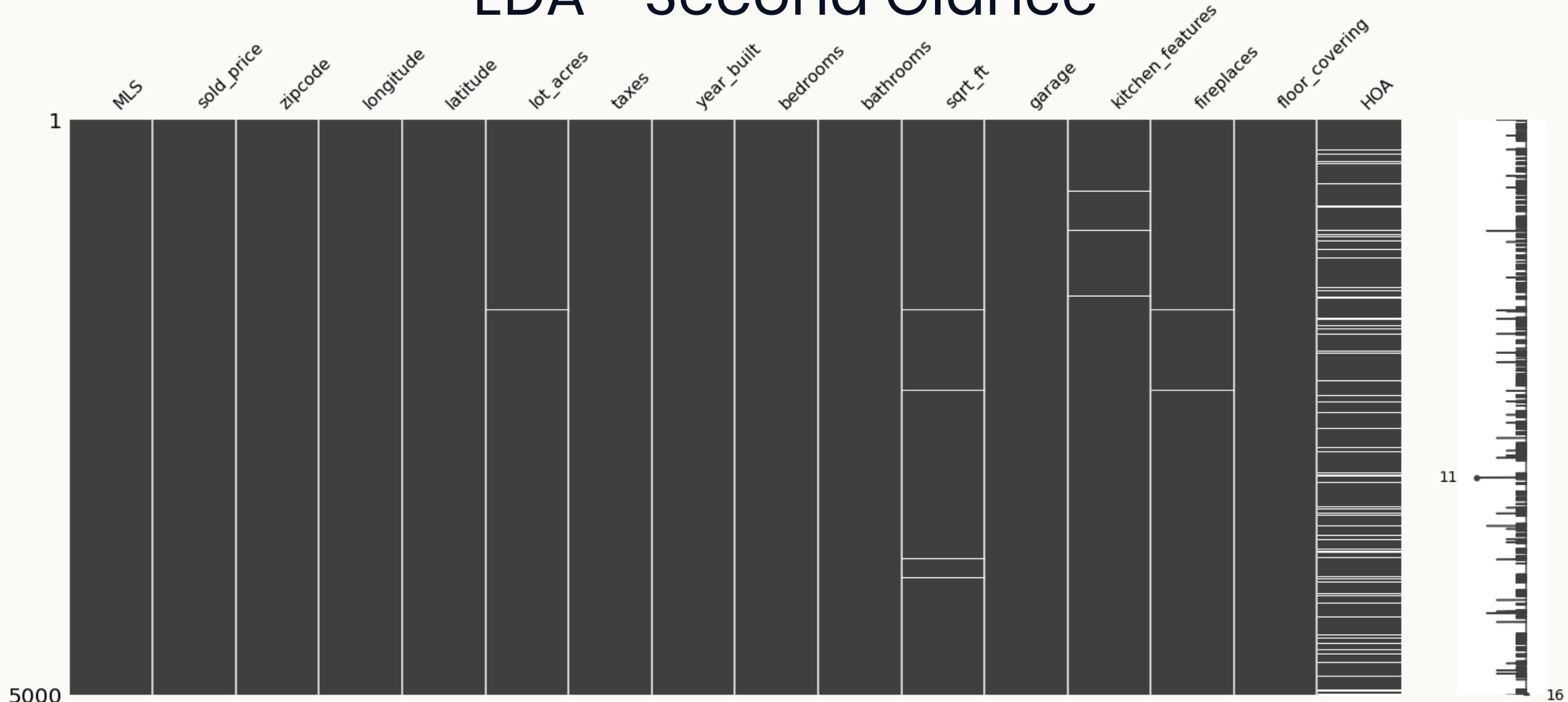
Old one

MLS	0
sold_price	0
zipcode	0
longitude	0
latitude	0
lot_acres	10
taxes	0
year_built	0
bedrooms	0
bathrooms	0
sqrt_ft	0
garage	0
kitchen_features	0
fireplaces	25
floor_covering	0
HOA	0
other_features	0

# Now we went from 35 nan values to 701 nan values.



# EDA - Second Glance



# Changing Object Types

```
MLS          int64
sold_price   float64
zipcode      int64
longitude    float64
latitude     float64
lot_acres    float64
taxes        float64
year_built   int64
bedrooms     int64
bathrooms    float64
sqrt_ft      float64
garage       float64
kitchen_features string
fireplaces   float64
floor_covering string
HOA          float64
```

Now all columns have its  
respective types.

※＼(^o^)／※



# Deletion and prediction of missing values

# Something to consider

There are two important thing to remember. The dataset has 701 nan values. 562 of them are from 'HOA' column, making it the column with the most missing values. The other 139 are spread within 'lot\_acres', 'sqrt\_ft', 'garage', 'kitchen\_features', 'fireplaces' and 'floor\_covering' columns.

# Correlation

HOA	1.000000
sold_price	0.171170
latitude	0.030892
year_built	0.015036
fireplaces	0.006481
bathrooms	0.005243
taxes	0.004560
sqrt_ft	0.002485
lot_acres	-0.008533
MLS	-0.018158
longitude	-0.021703
zipcode	-0.024722
garage	-0.039678
bedrooms	-0.067988

# Deletion of missing values

MLS	0
sold_price	0
zipcode	0
longitude	0
latitude	0
lot_acres	0
taxes	0
year_built	0
bedrooms	0
bathrooms	0
sqrt_ft	0
garage	0
kitchen_features	0
fireplaces	0
floor_covering	0
HOA	0
driveway_attached	0

# Now our data is complete

※＼(^o^)／※

# Some Things To Consider

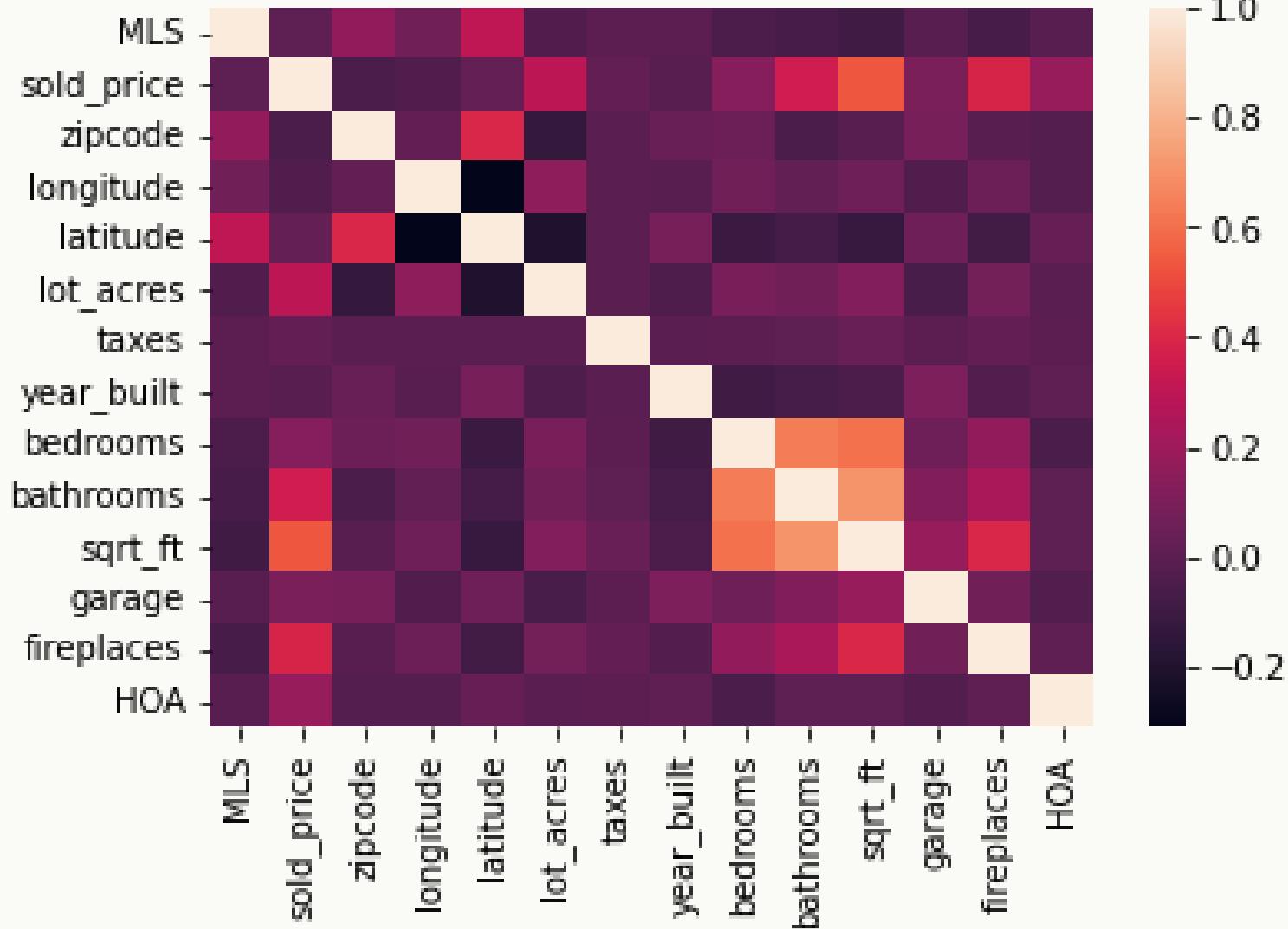
93 instances deleted

562 values were predicted for 'HOA' column.

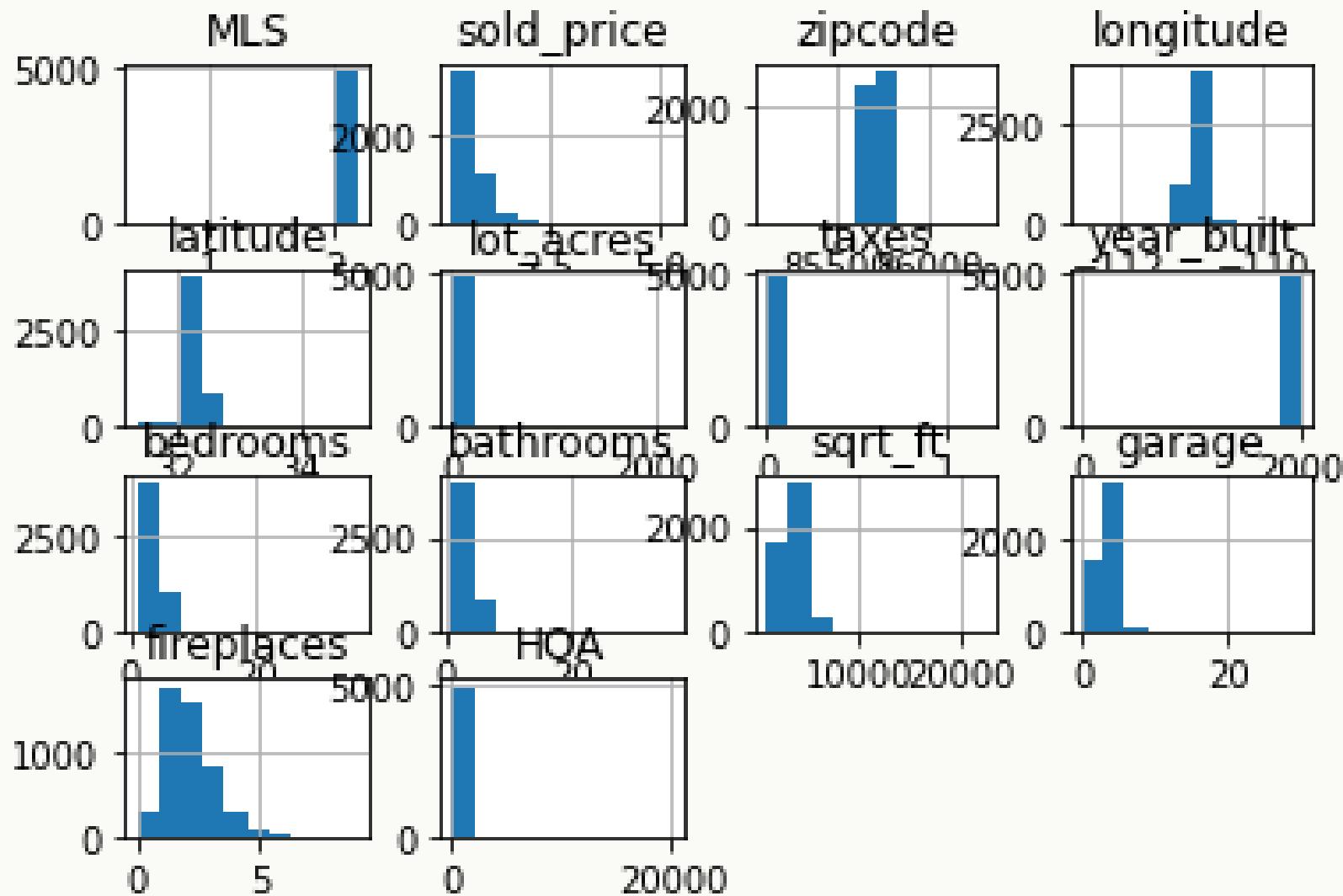


- ✓ Preliminary Data Exploration
- ✓ Data Cleaning
  - Changing object columns to string, int or float depending the case.
  - Delete rows and/or Predict missing values
- ✗ More Data Exploration
  - Correlation
  - Distribution

# Correlation



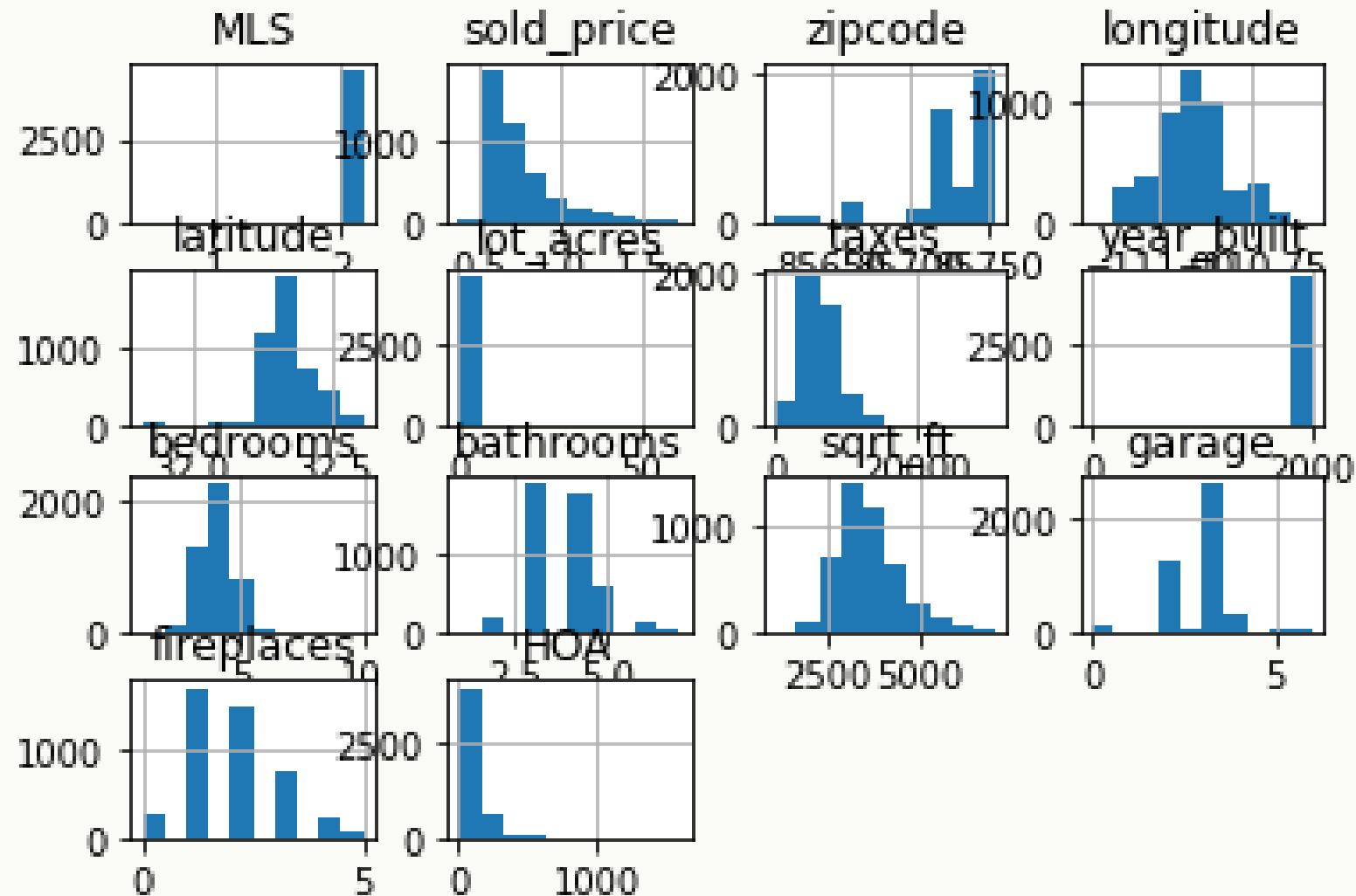
# Distribution



# A wild outlier appears!

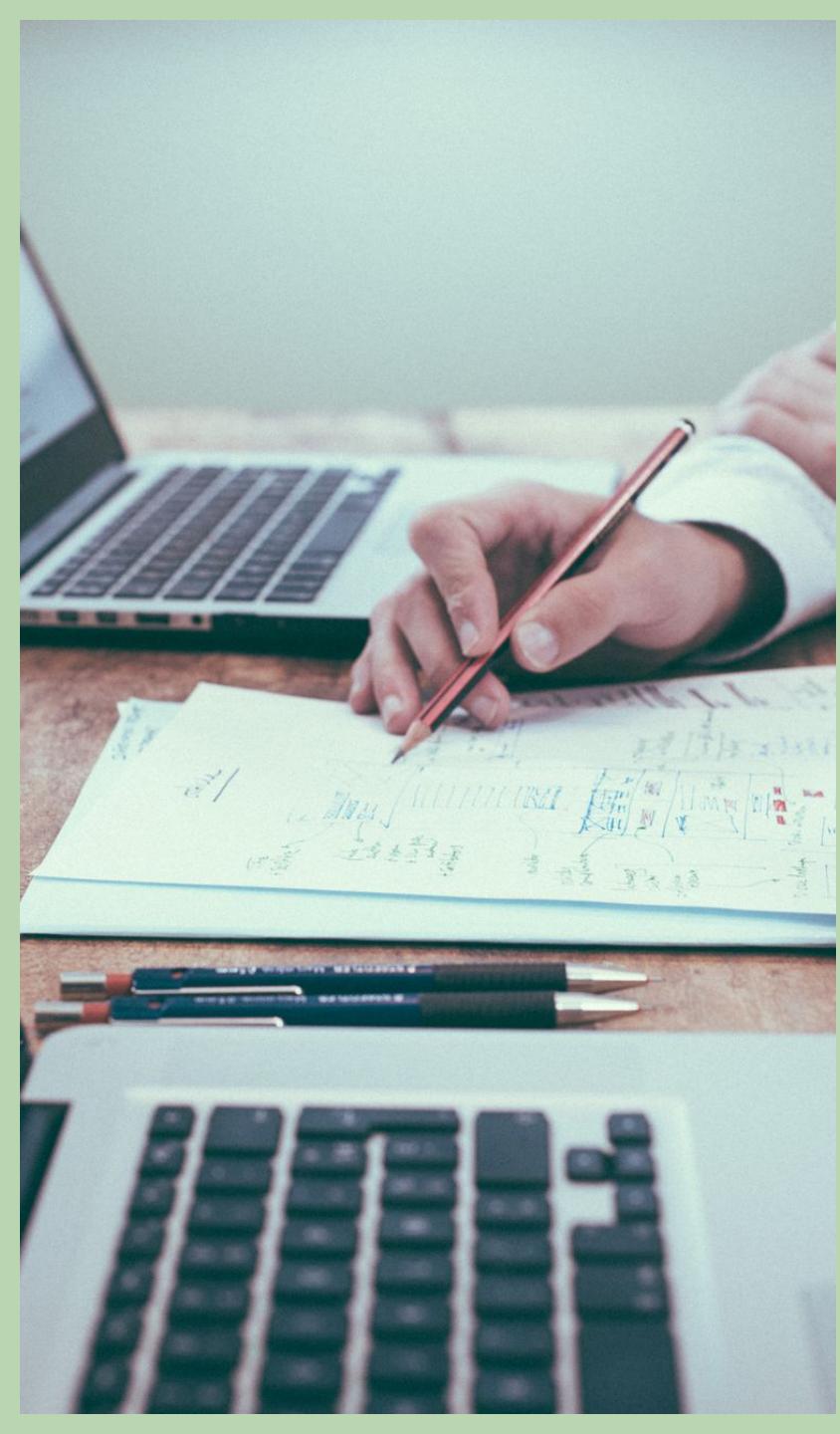
(.•\_•.)

# Distribution after outlier deletion





- ✓ Preliminary Data Exploration
- ✓ Data Cleaning
  - Changing object columns to string, int or float depending the case.
  - Delete rows and/or Predict missing values
- ✓ More Data Exploration
  - Correlation
  - Distribution

A photograph showing a person's hands writing in a notebook with a red pen. The notebook contains various sketches and diagrams. In the background, a laptop is open on a desk, and a white keyboard is visible in the foreground.

# Conclussions

# Conclusions

- The original dataset had 5000 instances.
- The data types were: 4 ints, 10 floats and 2 strings.
- In the data exploration phase, there were some challenges dealing with missing data like: Finding object type columns, dealing with strings, etc.



# Conclusions

- Data deletion and data prediction were needed in this project to clean the dataframe.
- Some outliers have to be deleted.
- After the cleaning process, the dataset has 4570 instances.



# Conclusions

- 430 instances had to be deleted.
- Correlations between variables tend to be poor.
- ( ツ~ツ) ノシ The world is full of dirty data!



# Thank You!

(૩ ૦ - ૩ ૩)