Assignment 1

Mike Solis

The data set

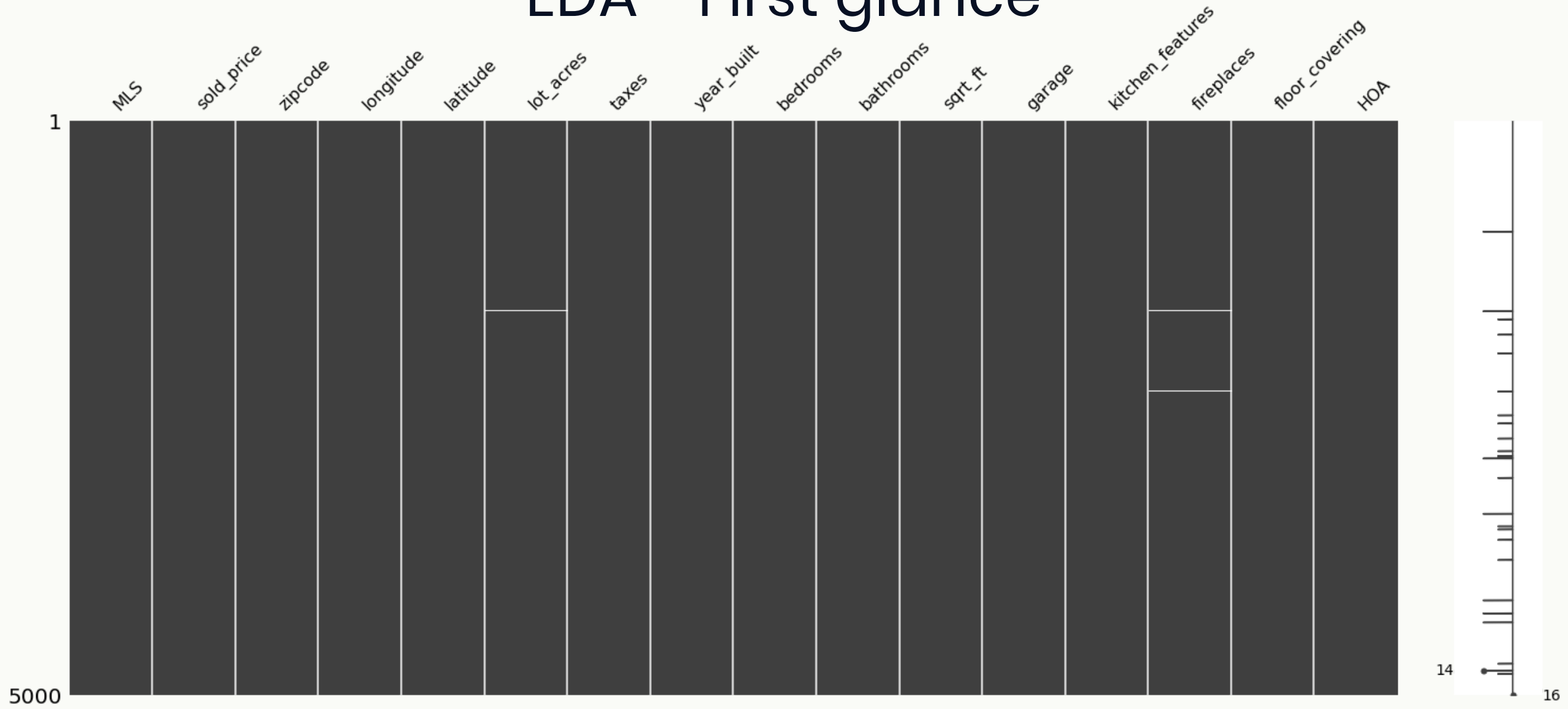
- 16 columns
- Types: Int, float and object
- 5000 instances

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
dtype: 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

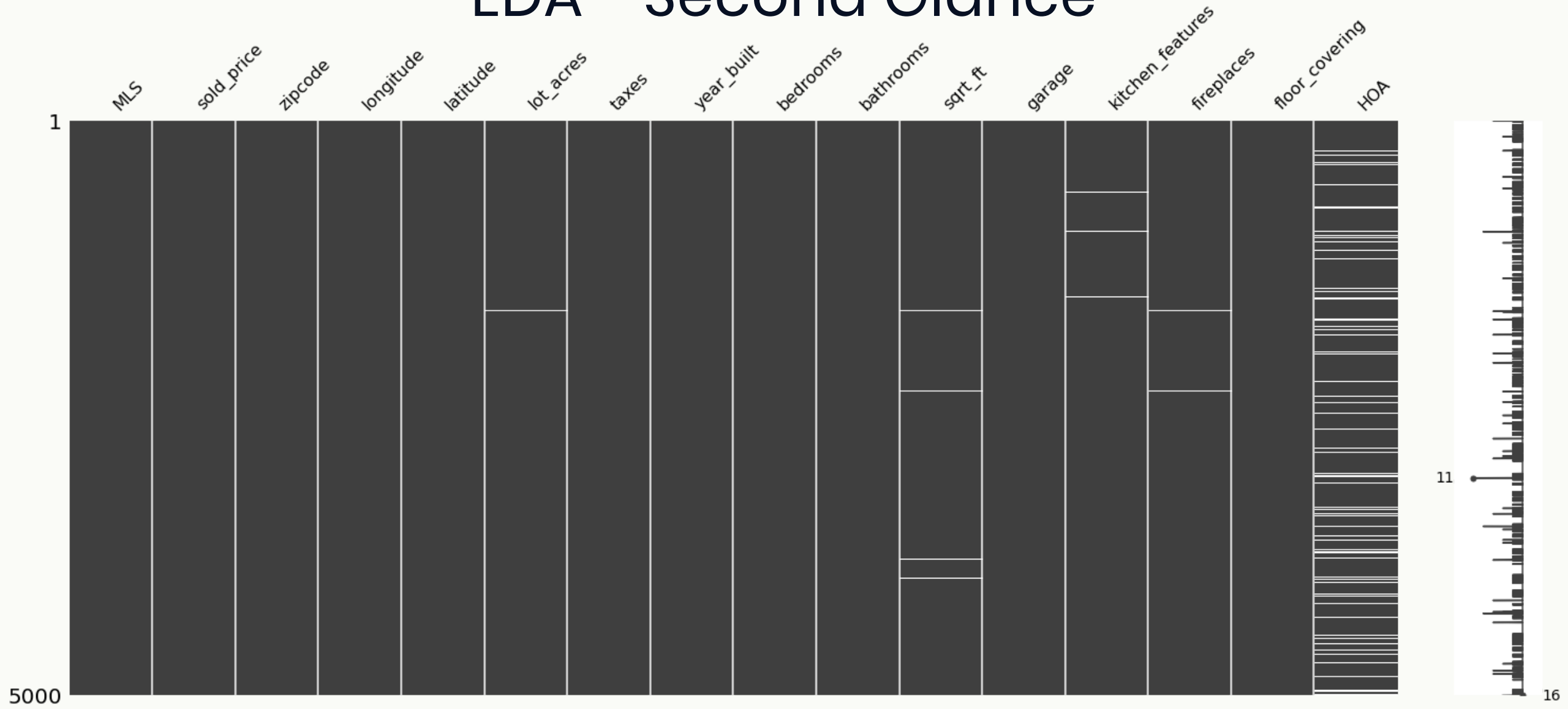
MLS	0
sold_price	0
zipcode	0
longitude	0
latitude	0
lot_acres	10
taxes	0
year_built	0
bedrooms	0
bathrooms	6
sqrt_ft	56
garage	7
kitchen_features	33
fireplaces	25
floor_covering	2
HOA	562

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

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
```

Linear regression

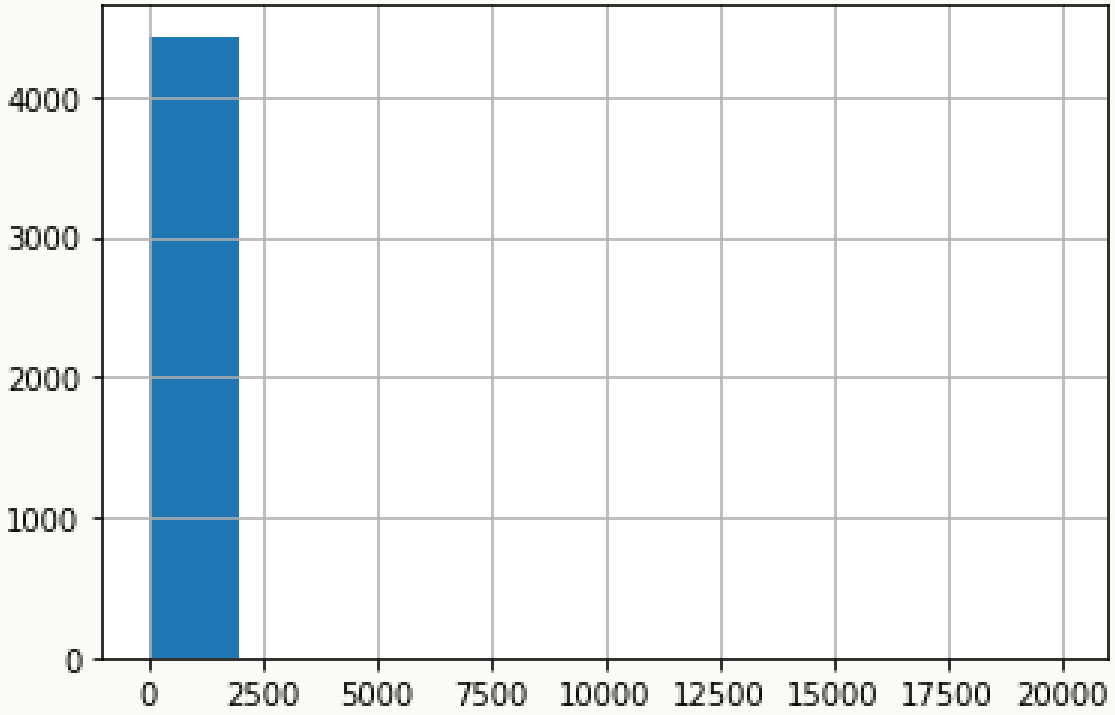
OLS Regression Results

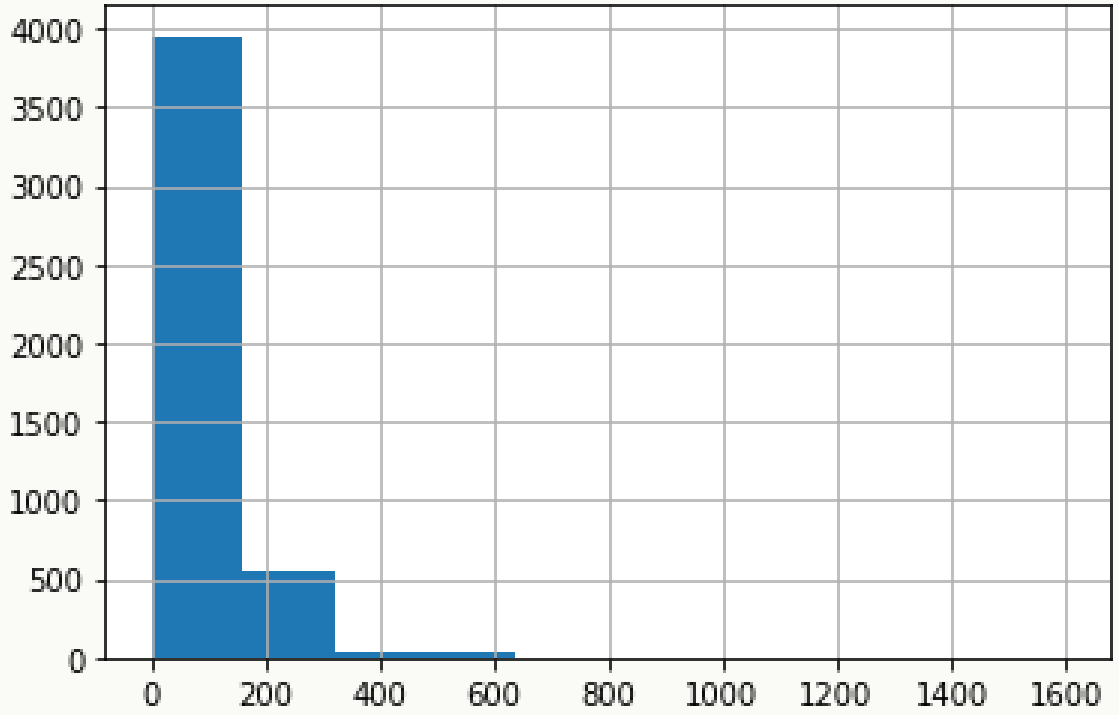
Dep. Variable:	HOA	R-squared:	0.029
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	133.9
Date:	Tue, 16 Aug 2022	Prob (F-statistic):	1.57e-30
Time:	03:47:04	Log-Likelihood:	-34213.
No. Observations:	4438	AIC:	6.843e+04
Df Residuals:	4436	BIC:	6.844e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-127.4913	21.505	-5.928	0.000	-169.653	-85.330
sold_price	0.0003	2.56e-05	11.571	0.000	0.000	0.000

Omnibus:	11871.064	Durbin-Watson:	2.022
Prob(Omnibus):	0.000	Jarque-Bera (JB):	233310353.826
Skew:	32.102	Prob(JB):	0.00
Kurtosis:	1124.420	Cond. No.	2.23e+06

Data replacement





Median = 56

Now our data is complete

※\(^o^)/※

Some Things To Consider

93 instances deleted

562 values were replaced for 'HOA' column by its median.

Preliminary Data Exploration

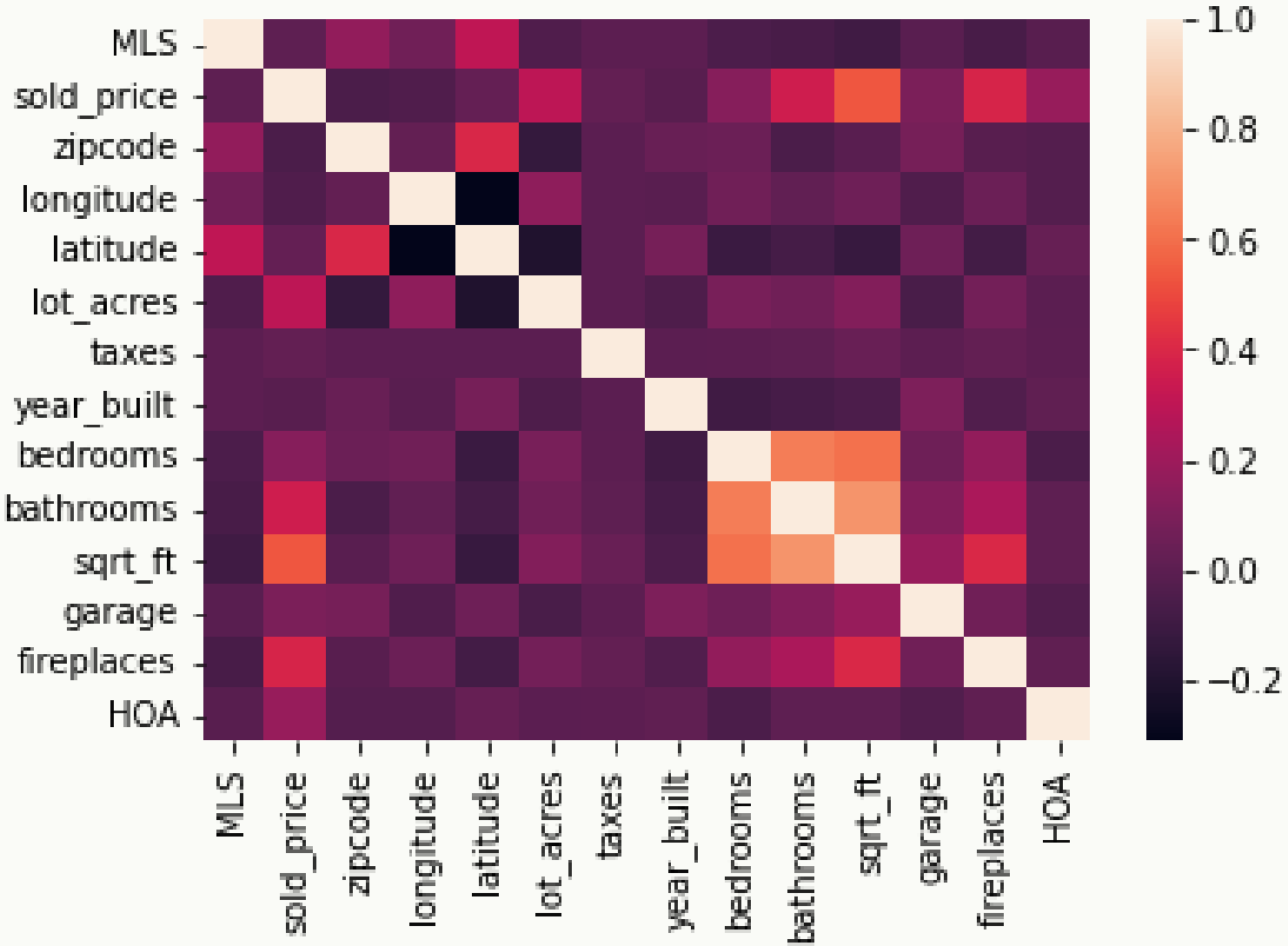
Data Cleaning

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

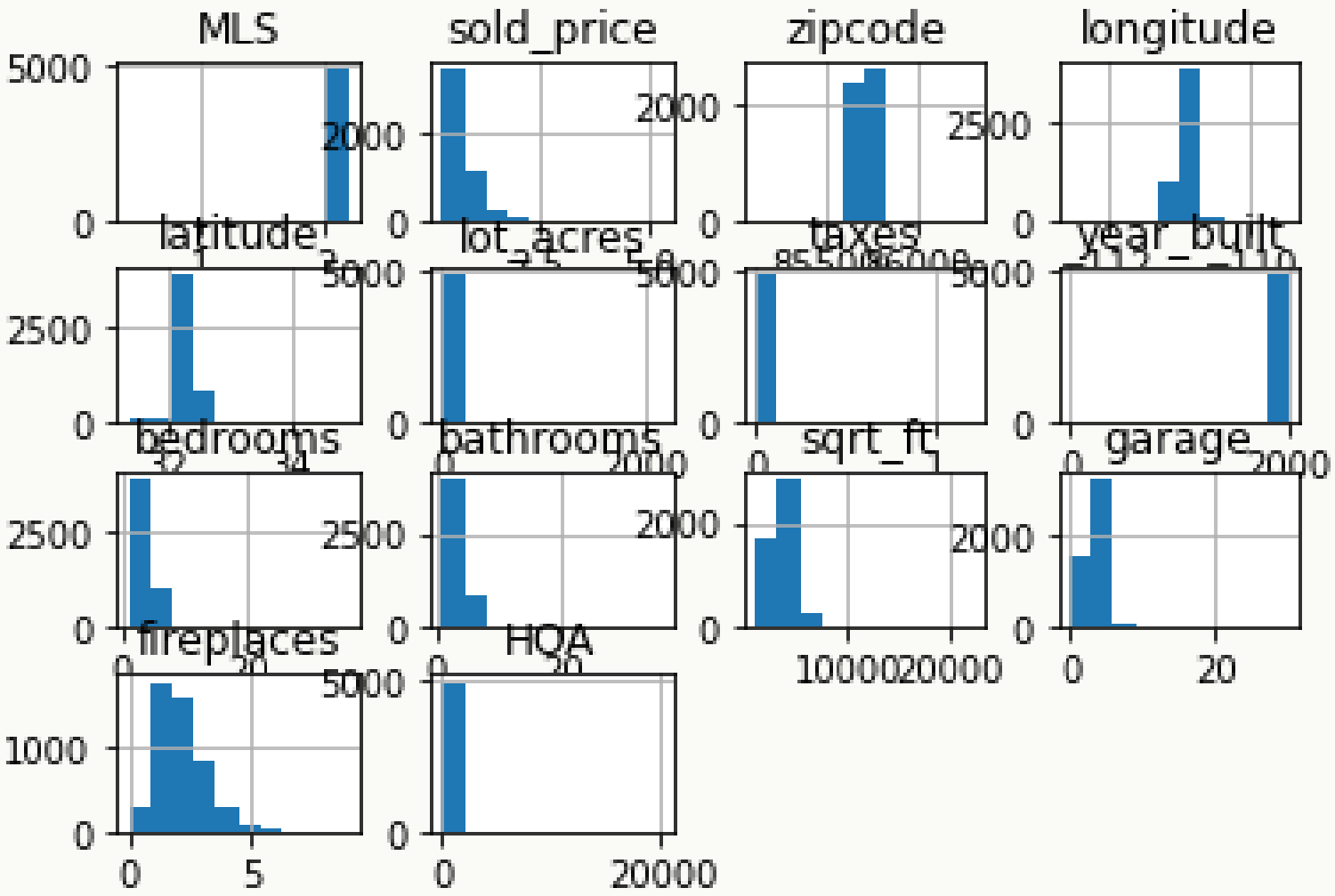
More Data Exploration

- Correlation
- Distribution

Correlation



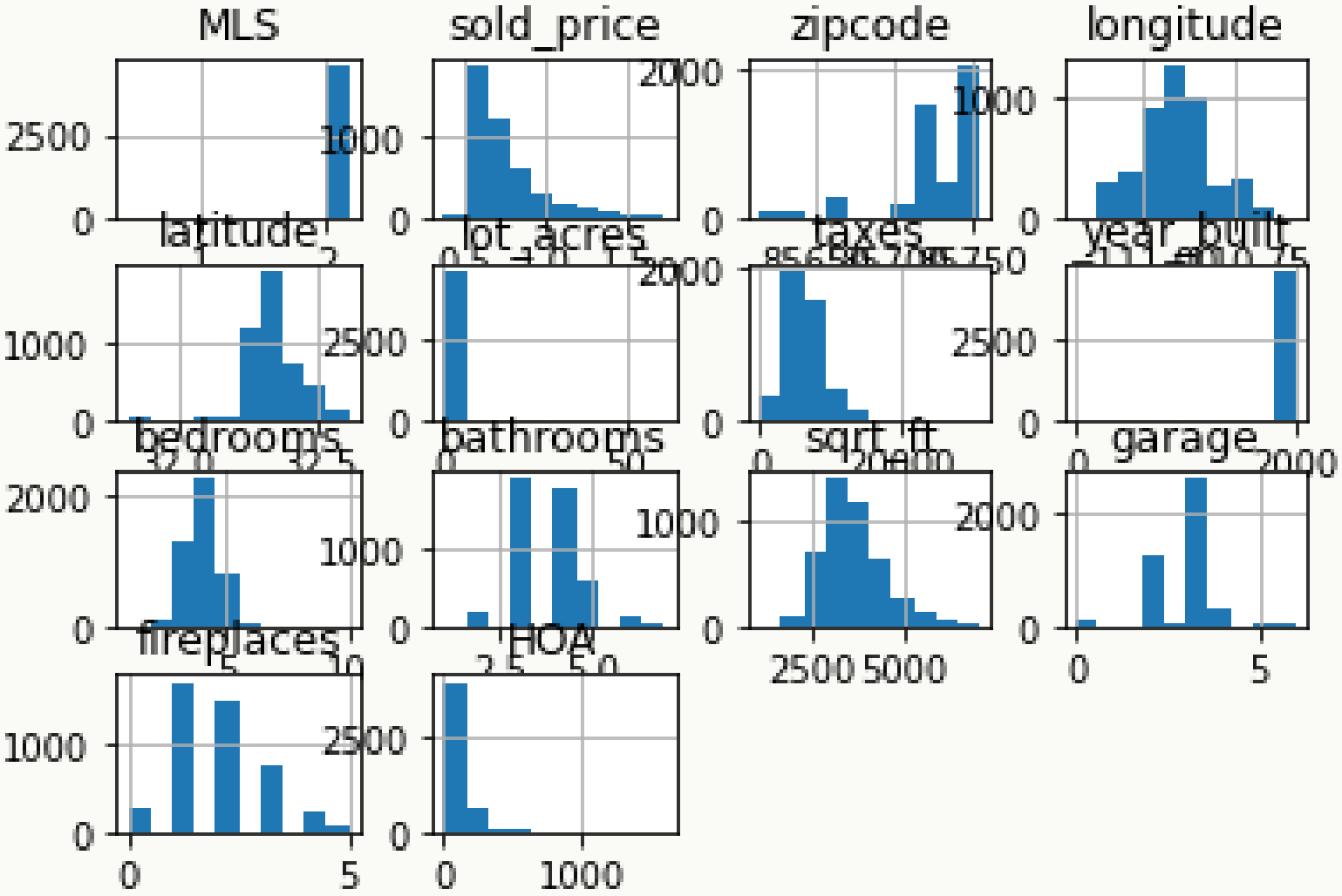
Distribution



A wild outlier appears!

(｡◕‿‿◕｡)

Distribution after outlier deletion



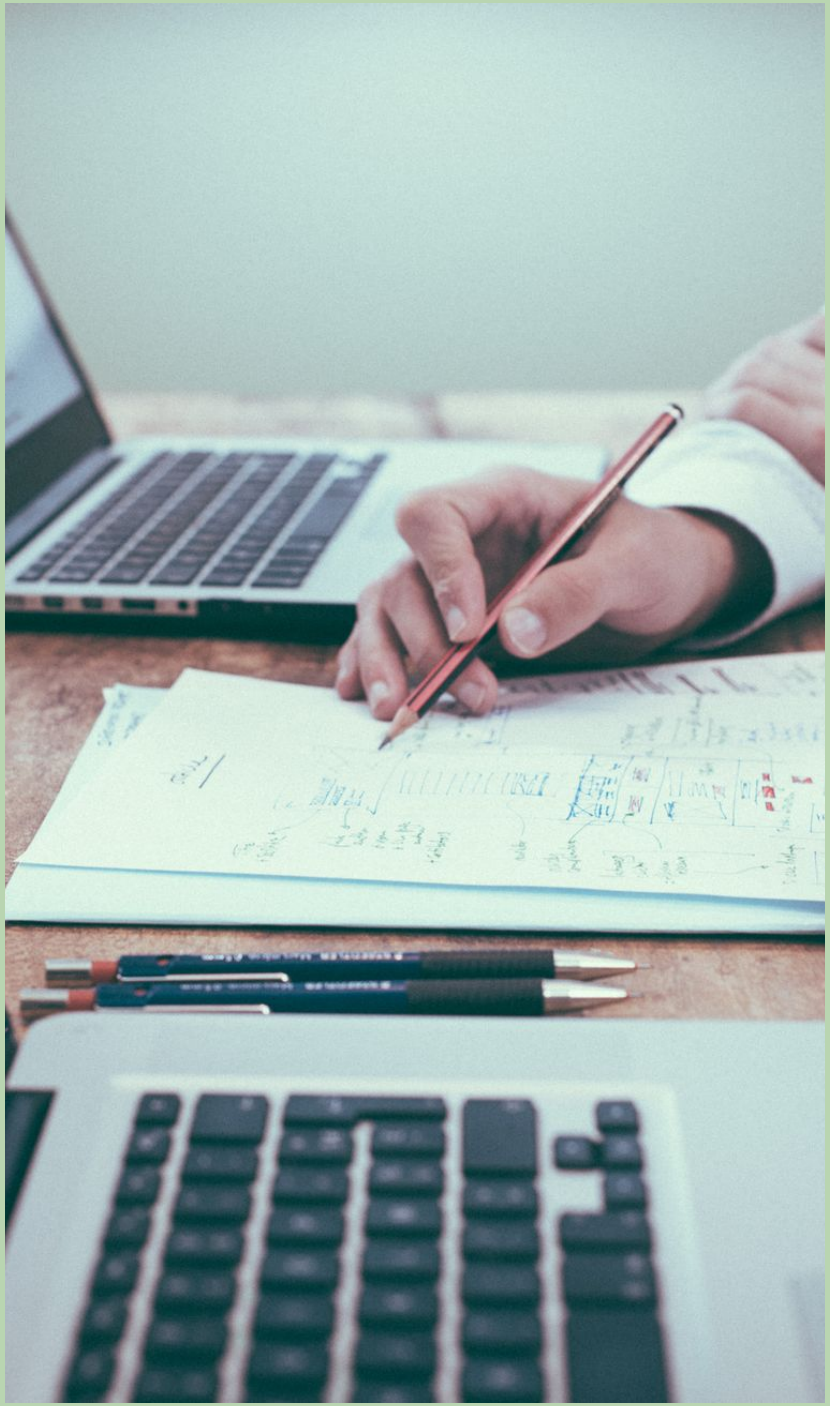
Preliminary Data Exploration

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



Conclussions

Conclusions

- The original dataset had 5000 instances.
- The data types where: 4 ints, 10 floats an 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.
- ‘HOA’ feature has the most missing values. More than 10% were missing.
- Linear regression model for predicting the 'HOA' feature values has a poor performance so, it was necesary to replace nan values with the median due to outliers.

Conclusions

- Some outliers have to be deleted.
- After the cleaning process, the dataset has 4570 instances.
- 430 instances had to be deleted.

Conclusions

- Correlations between variables tend to be poor.
- (╥﹏╥) ノシ The world is full of dirty data!

Thank You!

(ʃƪ˘⌣˘ʃƪ)