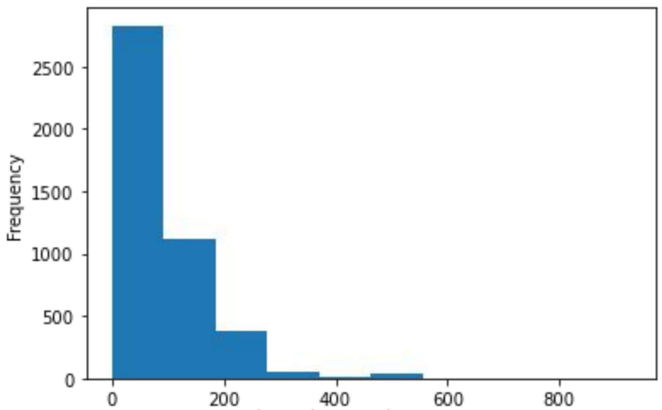
Data Cleaning and EDA

David Lizama

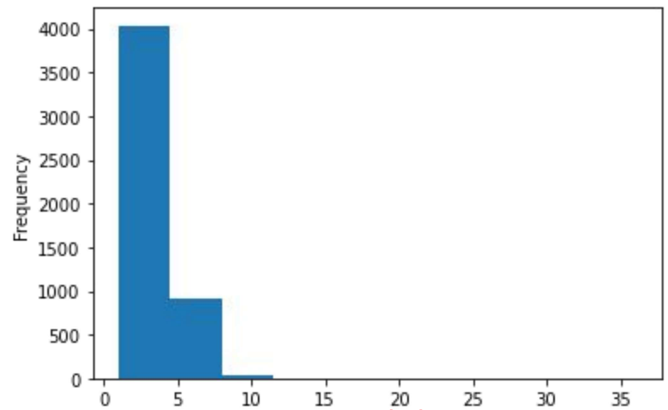
Introduction

- Database of 5000 observations and 16 variables related to houses.
- The first assumption is that our main objective is to predict house prices.
- Three main task: fill in missing values, separate features from strings, and handling bad records.
- No data normalization or correction of outliers in this assignment.
- The most correlated variables with house prices are lot_acres, bathrooms, sqrt_ft, and fireplaces.
- It is recommended to drop MLS and zipcode from the database.

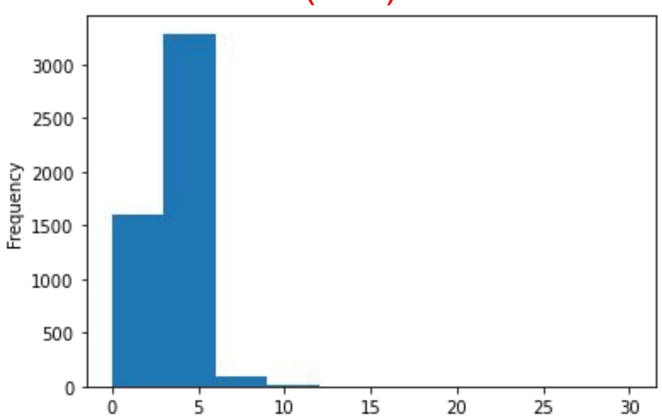
Cleaning data: Fill in missing values



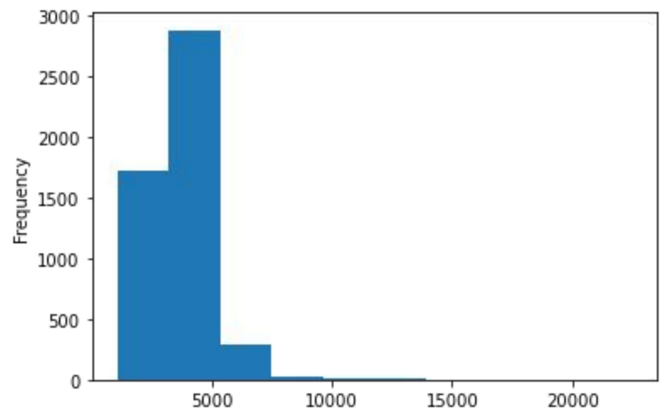
HOA(563): Mode



Bathrooms(6): Mode



SQRT_FT (56): Median



Garage (7): Mean, mode

There are 563 observations with ‘None’ and 35 missing values. It is not possible to drop them (>5%), so we need to impute them.

Cleaning data: Bad records

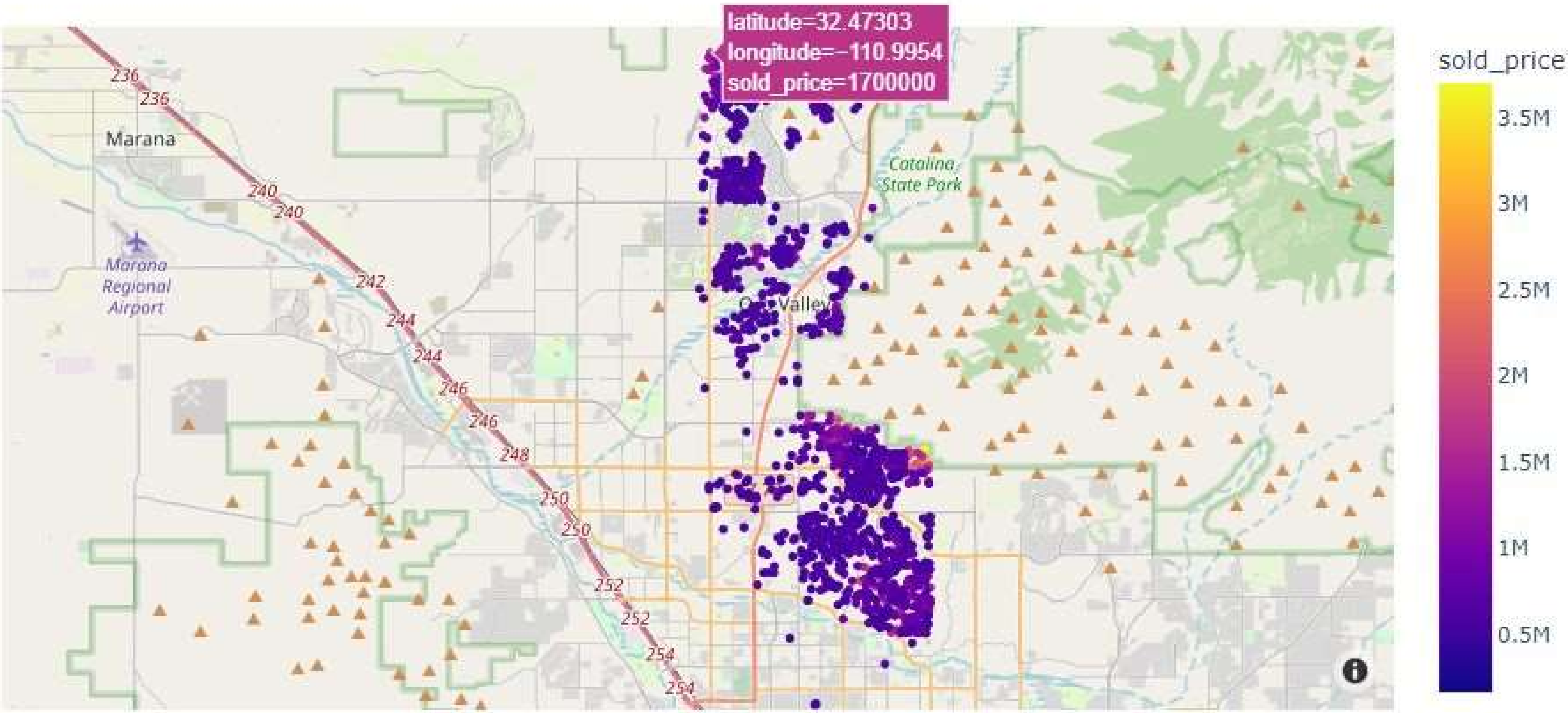
- Numbers separated by commas.
    - df['HOA'] = df.HOA.str.replace(',','.')
    - df['HOA'] = df.HOA.str.replace('None','0')
    - df['HOA'] = df.HOA.astype(float)

```
File ~\anaconda3\lib\site-packages\pandas\core\dtypes\c
   1177     raise ValueError(msg)
   1179 if copy or is_object_dtype(arr.dtype) or is_obj
   1180     # Explicit copy, or required since NumPy ca
-> 1181     return arr.astype(dtype, copy=True)
   1183 return arr.astype(dtype, copy=copy)

ValueError: could not convert string to float: '1,717'
```

- Latitude and longitude with numbers separated by dots.
    - Remove all the dots from columns.
    - Divided by 10^6 all observations.
    - Correct some observations which were out of range.
    - All houses should be located in Houston, Texas. Some of them were located in the middle of the ocean.

Cleaning data: Bad records



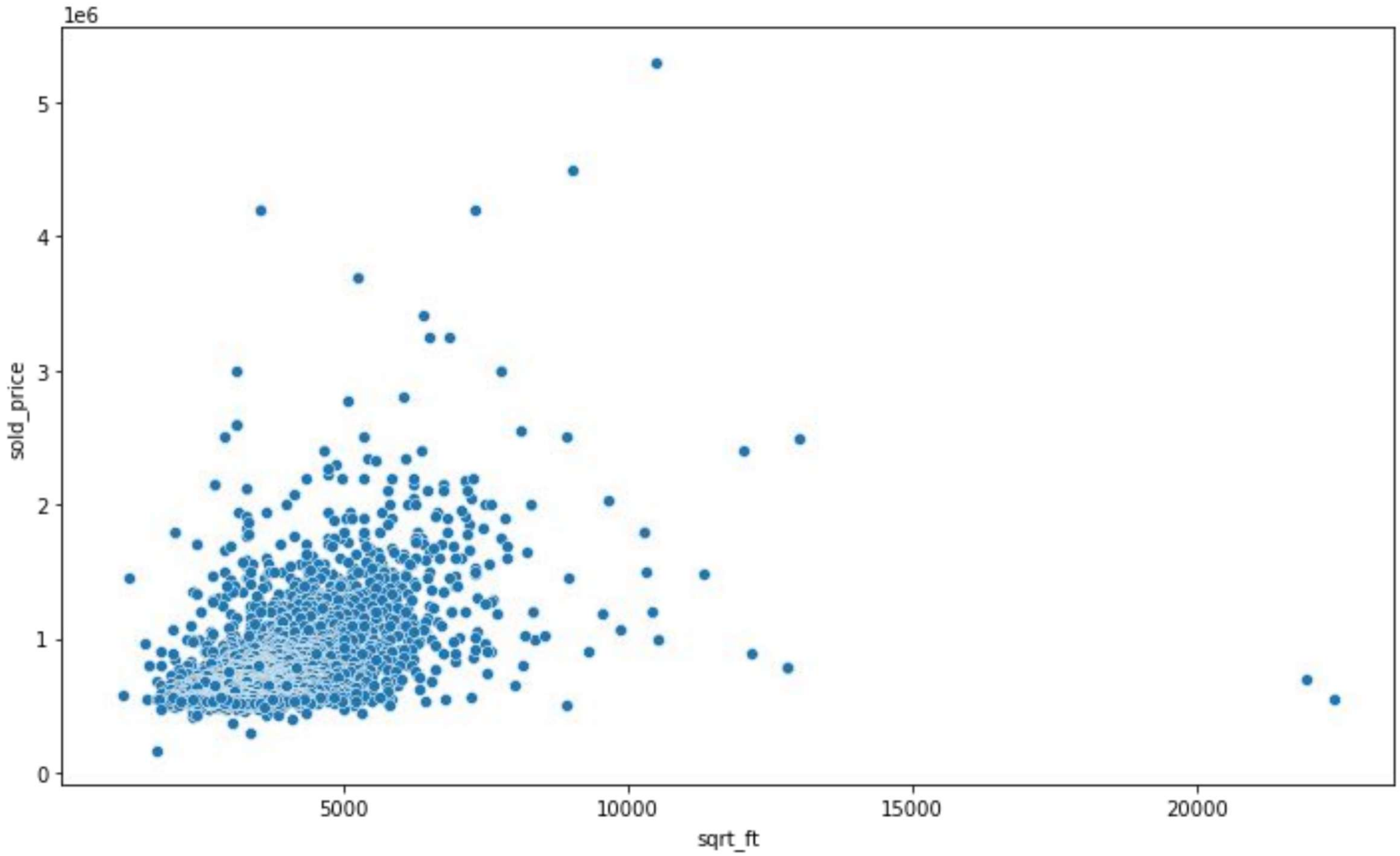
Cleaning data: String variables

- ❑ There are two columns with strings: floor covering and kitchen features.
- ❑ In this case, I detected some keywords in this columns and I created new columns with zeros or ones to count if houses have these keywords.

Floor covering	**Kitchen features**
• FC_Stone 1499 • FC_Ceramic 2527 • FC_Laminate 86 • FC_Wood 1248 • FC_Carpet 3509 • FC_Concrete 756 • FC_MexicanTile 660	• KF_Dishwasher 4857 • KF_GarbageDisposal 4520 • KF_Refrigerator 4234 • KF_DoubleSink 1164 • KF_Microwave 3625 • KF_Oven 3977 • KF_Compactor 432 • KF_Freezer 395 • KF_ElectricRange 401 • KF_Island 1252 • KF_GasRange 1307 • KF_Countertops 1482 • KF_Desk 327

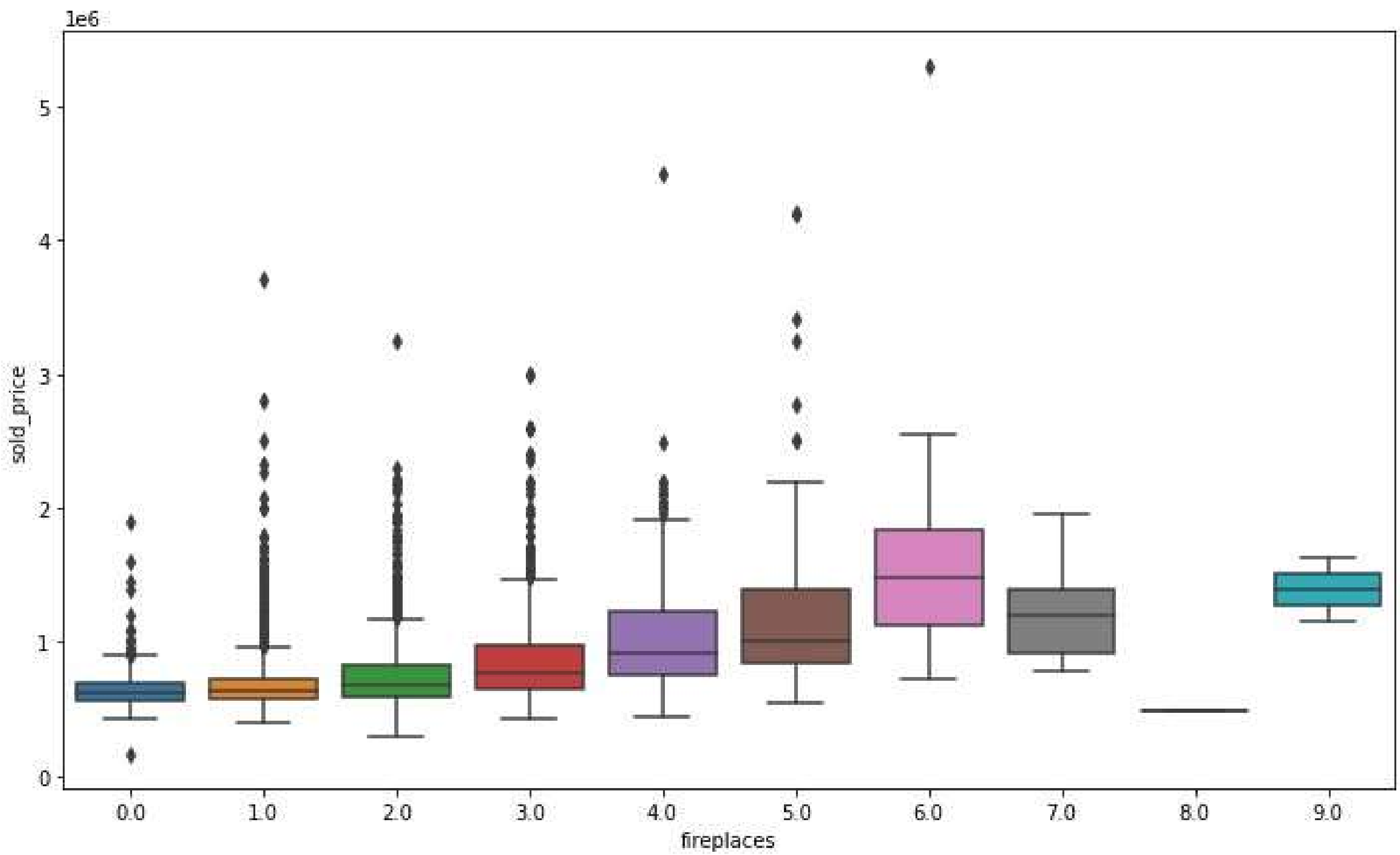
EDA: Price houses and square footage

- High correlation between the house prices and sqrt_ft (0.52).
- When sqrt_ft increases, the price of the house increases.



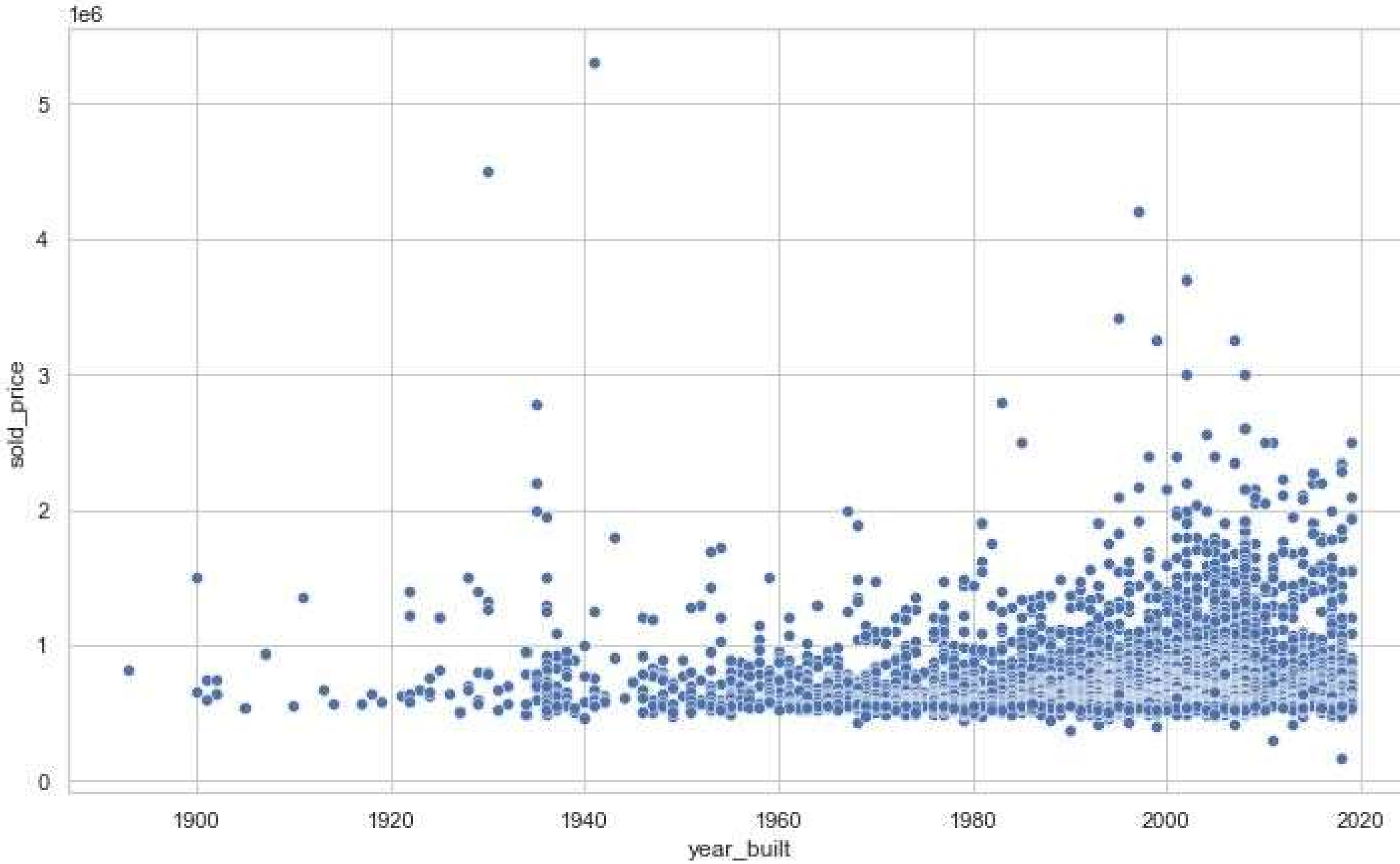
EDA: Price houses and fireplaces

- There is a positive association between the amount of fireplaces and the price of the house.



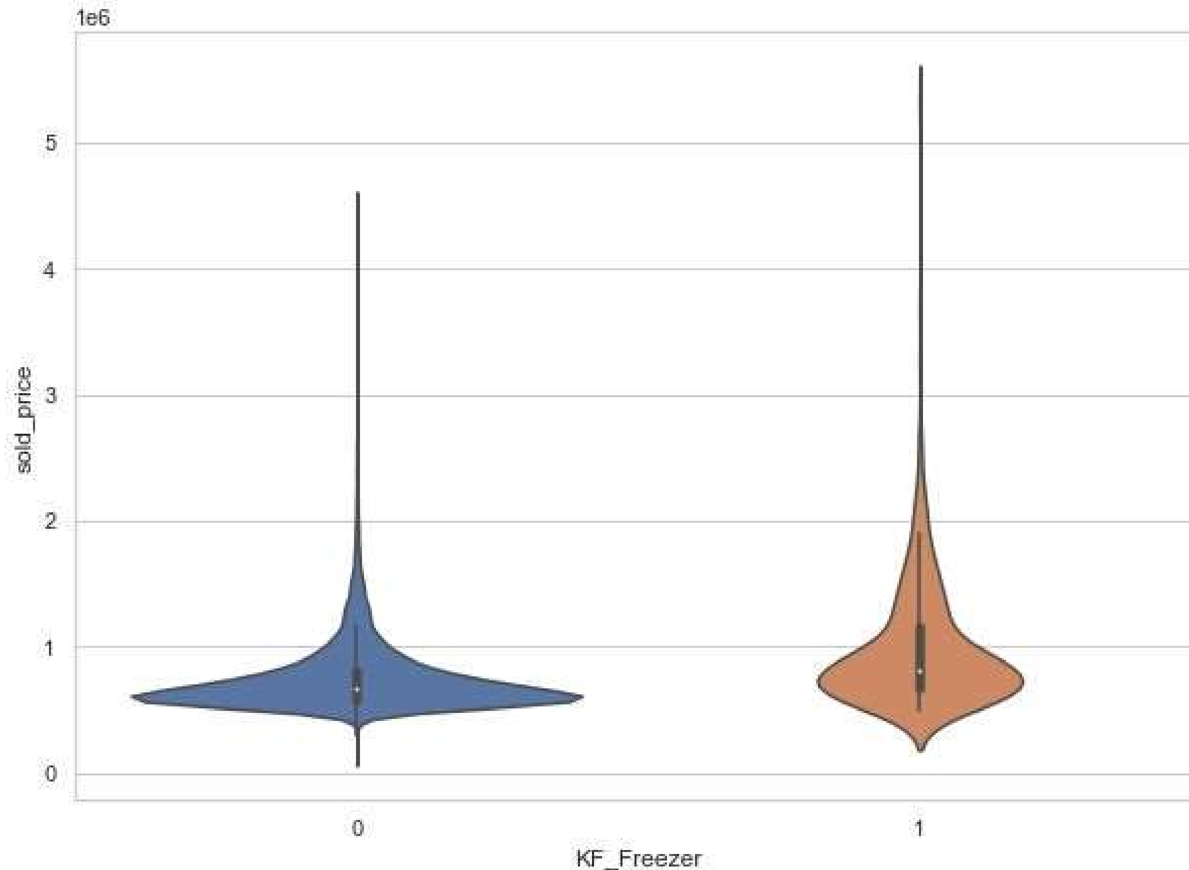
EDA: Price houses and year of construction

- There is a positive association between the year of construction of the house and its price.



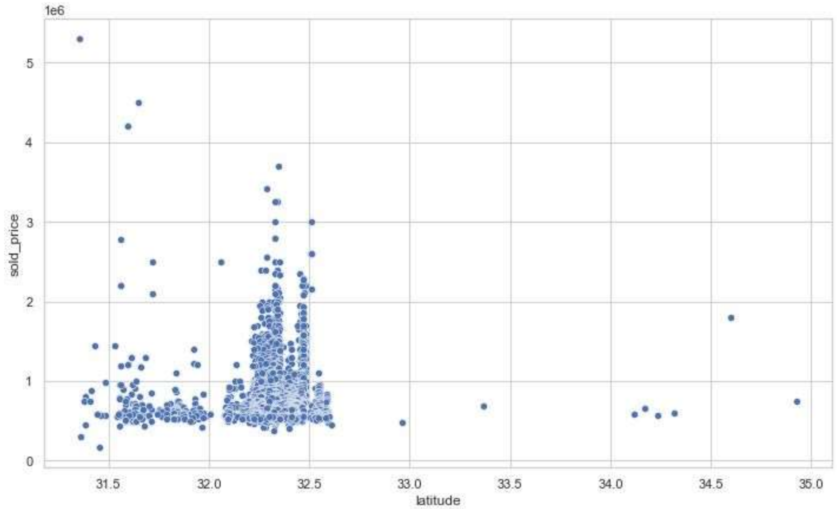
EDA: Price houses and kitchen features

- Most of the features in the kitchen were not significantly at explaining the price of the houses.
- Although the medium price of the houses are similar, houses with a freezer in the kitchen have a higher price that houses without freezer.

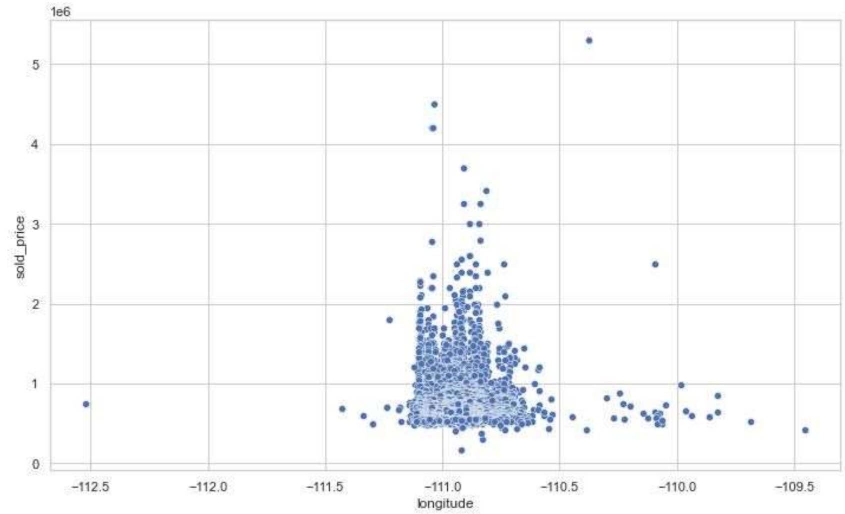


EDA: Price houses, latitude and longitude

Latitude



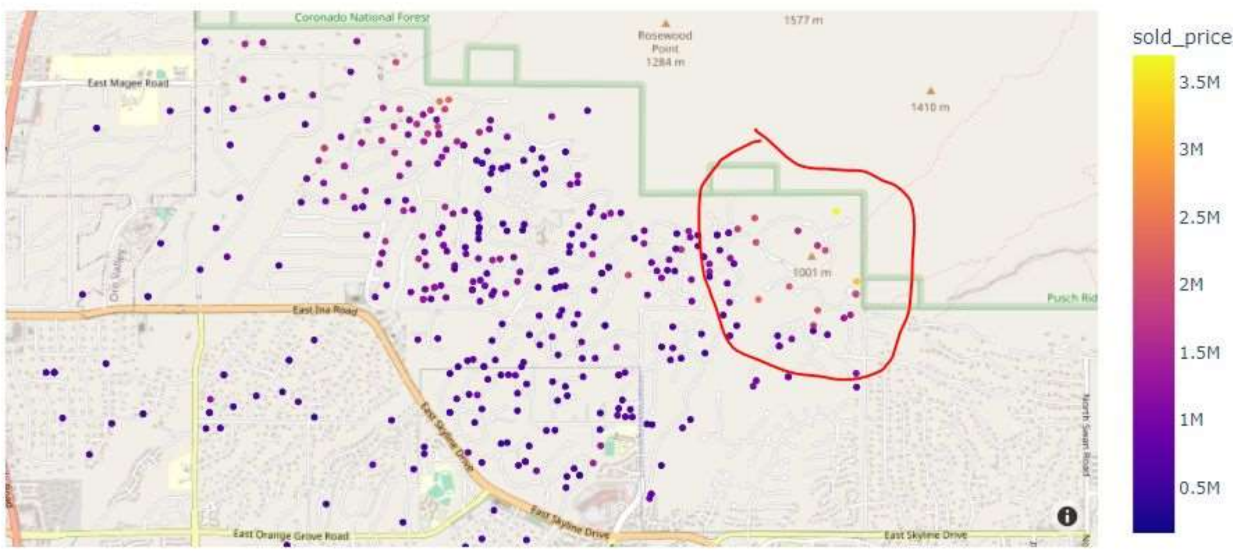
Longitude



- Most of the features in the kitchen were not significantly at explaining the price of the houses.
- Although the medium price of the houses are similar, houses with a freezer in the kitchen have a higher price that houses without freezer.

EDA: Price houses, latitude and longitude

- The most expensive houses are located in a particular area of Houston which is difficult to find with a simple linear regression. A model base tree could obtain a better performance in this case.



Conclusions

- The database was cleaning consist on filling missing information, handling bad records, and separating key features of the kitchen and floor covering.
- The most important information to predict house prices are square footage, year of constructions, bathrooms, lot_acres, and fireplaces.
- There is not a positive association between longitude and latitude, and price house, but we can split specific areas in which we can find high value houses.
- Low association of house prices and the rest of variables. For example, features of kitchen and floor covering do not seem to be relevant.