Data Cleaning in R

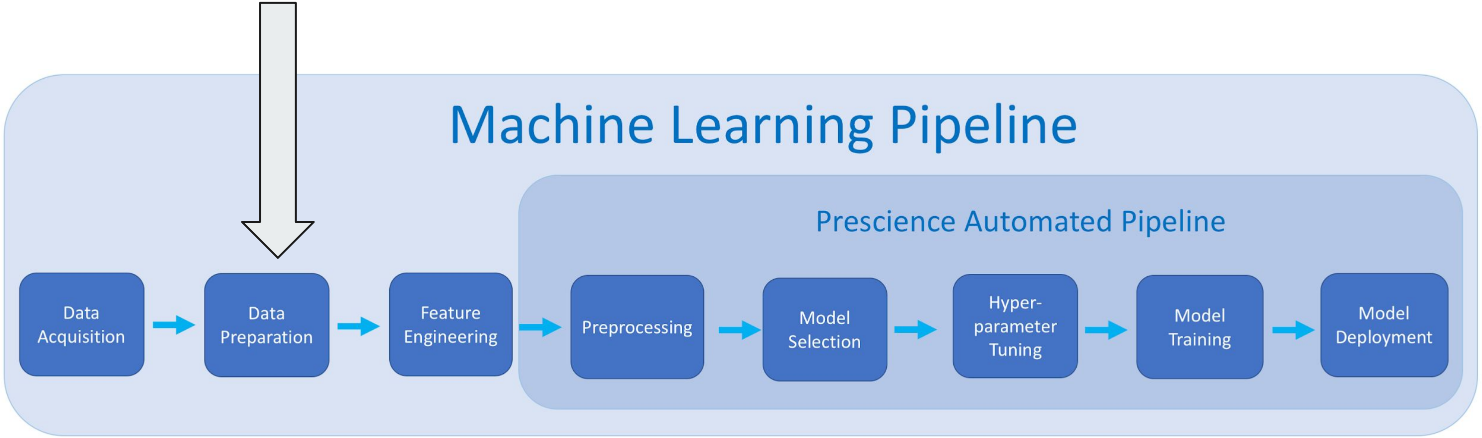
Dataset

- 5000 **houses for sale** data in the Arizona real estate market
- **16 variables** describing every house:

  Multiple Liste Service (MLS), price that the house have been sold, locations (longitude, latitude, zipcode), acres of the lots, taxes, year that the houses have been built, number of bedrooms and bathrooms, size of the houses (in square feet), number of garage slots, kitchen and floor description, number of fireplaces and

Motivation

- Dirty data **disable** the ability to do exploratory data analysis.
- **Outliers** of a dataset can enhance the **bias** of an analysis.

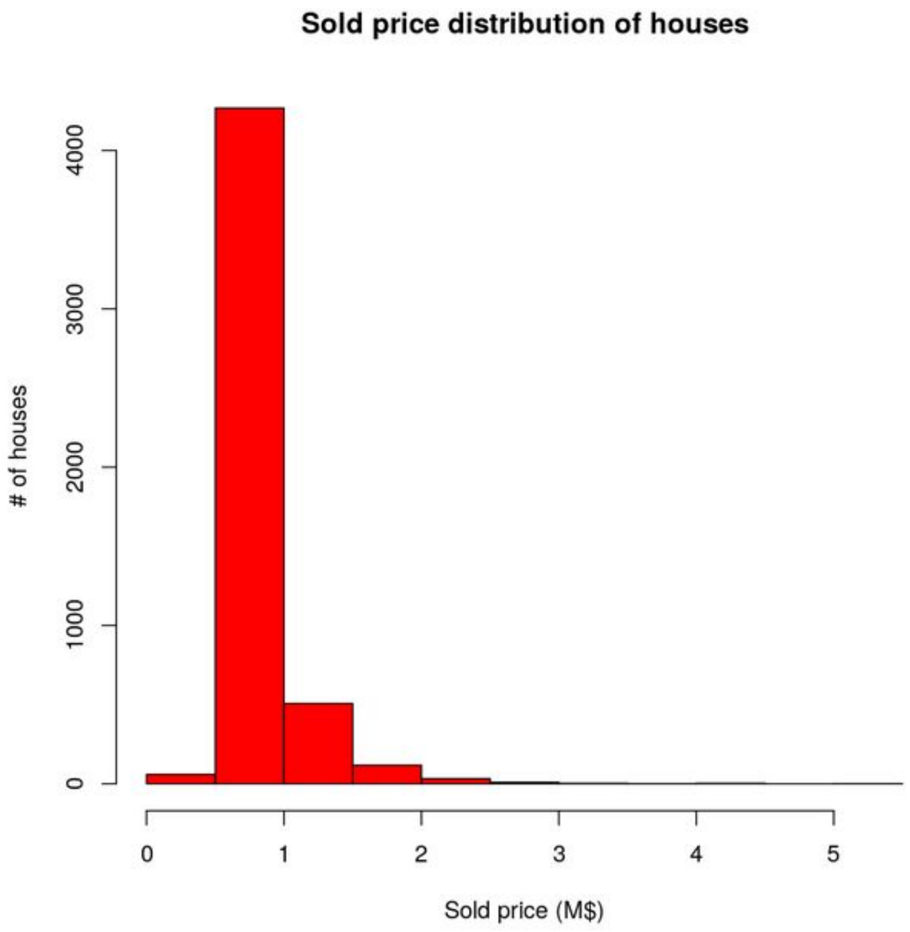


Tools

- Programming language : R
- Packages: **tidyverse**; readr (tibble data type), revgeo

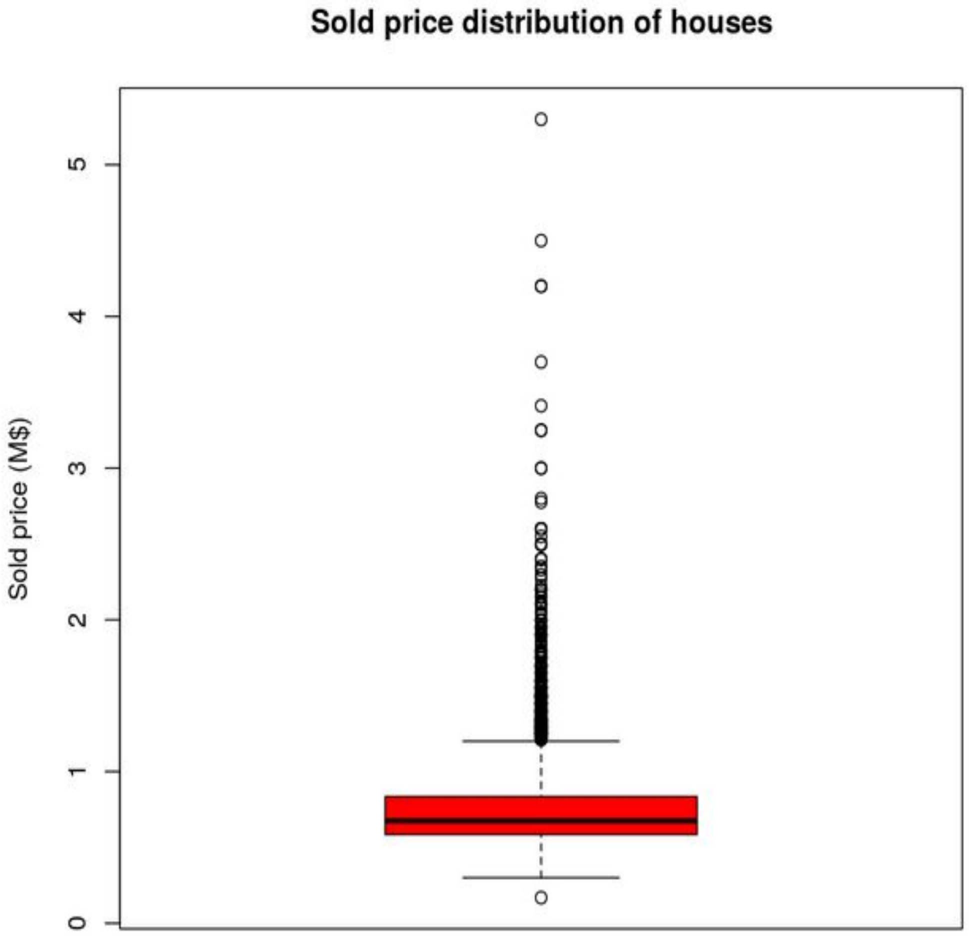
Outliers

- Using the **sold price** to remove outliers.
- Distribution doesn’t look too bad from Histogram.



Outliers

- 394 data points above the maximum
- 1 point below the minimum

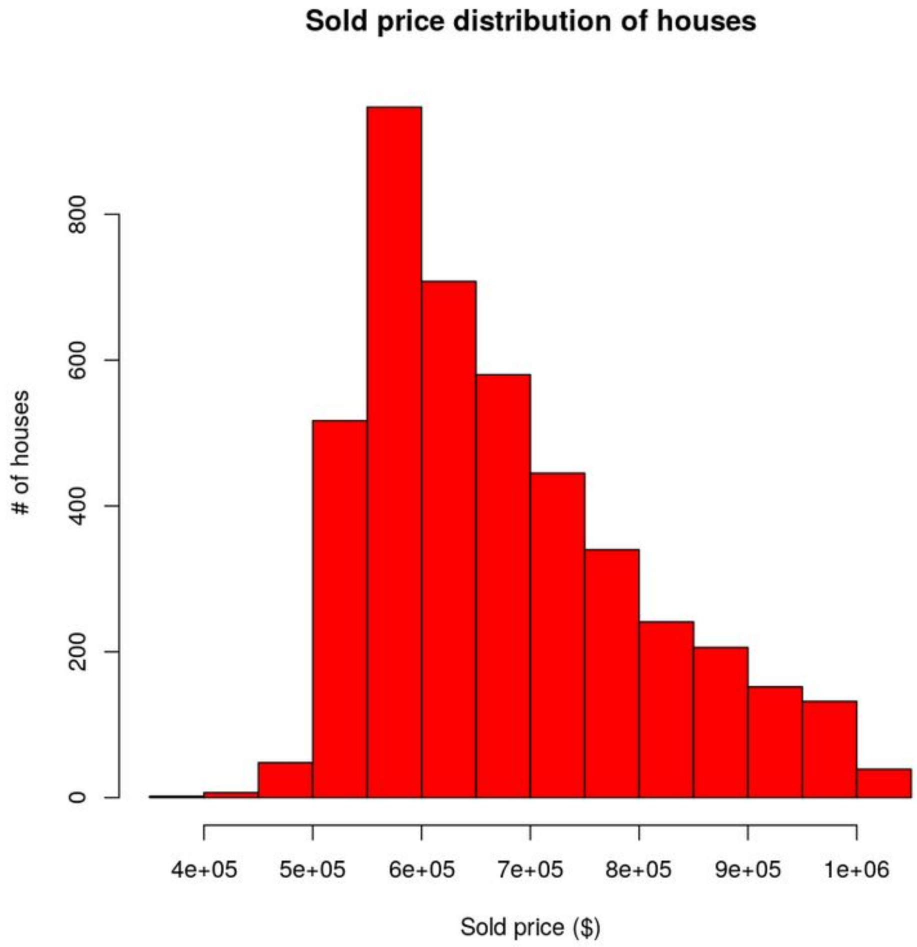


Outliers

- After removing the outliers, all data stays in the in the same scale.

Outliers

- Distribution looks better (closer to a bell curve).



Variable type

- We need to make sure that the variables are of the correct type (i.e.: numeric, character).
- **HOA** misclassified as character due to the commas separating thousands.
- **Garage** and **square feet** and misclassified as character due of values of “None” (and not NA)

garage
0
0
None

HOA
1,000
250
1,200
1,200
None
1,200

sqrt_ft
10500
7300
None

Dirty data - bathrooms

- 5 missing values for 3 and 2 bedrooms houses.
- Mean of 3 bedrooms houses = 3.28
- Mean of 2 bedrooms houses = 2.84
- Impute with the **mean** of the houses with the same number of bedrooms.

Dirty data - lot of acres

- 10 missing values for lot of acres
- 24 lots of acres were equal to 0
- Mean of 2.8395 and median of 0.9400
- 2308 values (>50% between 0 and 1)
- Impute with **median.**

Dirty data - square feet

- 55 missing values for square feet.
- **Correlation** between square feet and sold price (p-value ~= 0.41)
- Impute with the **closest sold price** (KNN, k=1).

Dirty data - taxes

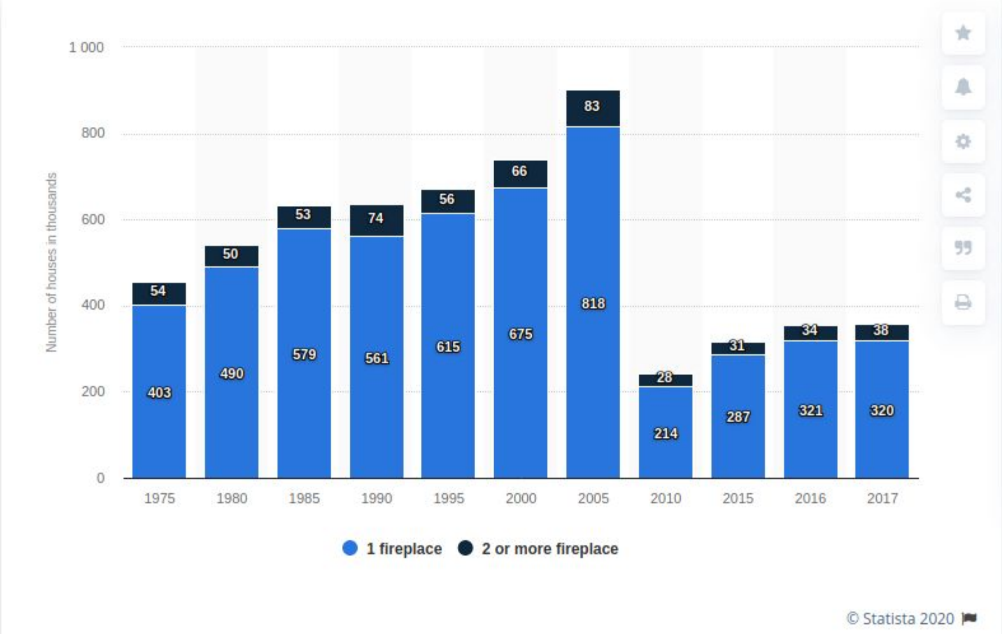
- 17 data points of taxes were equal to 0.
- API calls to a geocoder provider confirmed location of the houses in Arizona.
- Impute with the **mean** as we can expect the taxes to correlate with the location.

Dirty data - HOA

- 497 missing values of HOA (different of values of 0)
- Imputate with the **mean**
- ~10% of data: can generate bias during analysis

Dirty data - fireplaces

- 25 values missing
- No correlation with the year built (p-value ~= -0.03)
- Mean of ~1.75 (looks high to impute with 2 fireplaces by house).
- Adding data: Statista shows a majority of 1 fireplace/house in the US.
- Imputate with 1



Conclusion

- Histogram and boxplot to remove outliers (395 removed)
- Make sure that the variables are of the correct type (3 variables corrected)
- Think about possible correlations and impute missing values with the mean, median, KNN,etc.