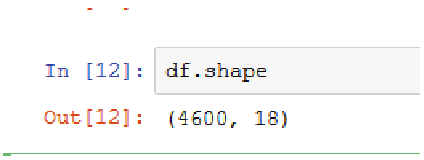
The data

I use the freely available data set present on Kaggle[1]. This data set contains the data regarding 4600 houses that have been sold in the USA. Specifically we have 18 **columns** and 4600 **rows**.[2]



```
In [12]: df.shape
Out[12]: (4600, 18)
```

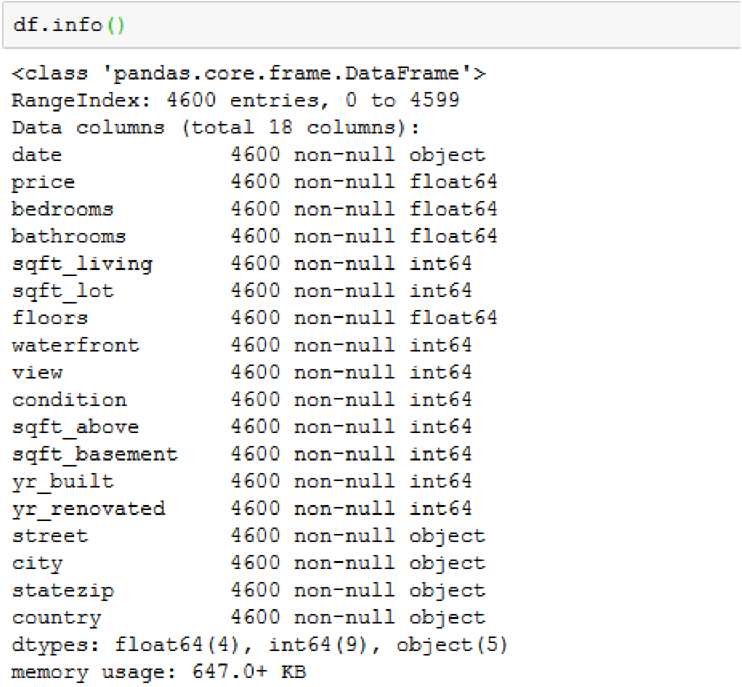
The 18 columns are the following:

- The **date** in which the house has been put up for sale.
- The **price** of the house, that is going to be our target variable, expressed in US dollars.
- The number of **bedrooms** present in the house.
- The number of **bathrooms** present in the house.
- The **sqft-living**, that is to say the living surface of the house expressed in foot square.
- The **sqft-lot**, i.e. the surface of the full house including the garden (if any). This quantity is expressed in foot square.
- The number of **floors** presnt in the house.
- **Waterfront**. This is column contains a categorical variable that takes value 1 if the house is waterfront and 0 otherwise.
- **View**. This is column contains a categorical variable that can take discrete values $\{0, 1, 2, 3, 4\}$. Where 0 is the minimal value for the view (denoting a bad view) and 4 the highest (denoting a very nice view).
- **Condition**. This is column contains a categorical variable that can take discrete values $\{0, 1, 2, 3, 4, 5\}$. Where the value 0 denotes a bad condition of the house while the value 5 denotes the best possible condition of the house.
- **sqft-above**. This is the area of the above floor(s) of the house expressed in foot square.
- **sqft-basament**. This is the area of the basement floor of the house expressed in foot square.

[1] You can find the data set at this website `https://www.kaggle.com/shree1992/housedata`

[2] In the following I will denote by *df* the corresponding pandas DataFrame

- **Year of construction** of the house
- **Year of rennovation** of the house (if any).
- The **street** in which the house is located.
- The **city** in which the house is located.
- The **zip-code** of the house.
- The **country** (even if all the houses taken into account are located in the USA).



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4600 entries, 0 to 4599
Data columns (total 18 columns):
date             4600 non-null object
price            4600 non-null float64
bedrooms         4600 non-null float64
bathrooms        4600 non-null float64
sqft_living      4600 non-null int64
sqft_lot         4600 non-null int64
floors           4600 non-null float64
waterfront       4600 non-null int64
view             4600 non-null int64
condition        4600 non-null int64
sqft_above       4600 non-null int64
sqft_basement    4600 non-null int64
yr_built         4600 non-null int64
yr_renovated     4600 non-null int64
street           4600 non-null object
city             4600 non-null object
statezip         4600 non-null object
country          4600 non-null object
dtypes: float64(4), int64(9), object(5)
memory usage: 647.0+ KB
```

Figure 1: Summary of the different columns of the data set

0.1 The strategy

In order to solve the problem I will use Machine Learning techniques. In particular I will analyze, using a Regression model, how the price of the house (that is my target variable), can be correlated to some of the independent variables present in the above data set. A naive expectation, that have to be checked is that the price will be correlated to the surface of the house, the number of floors, the condition, the number of rooms and the year of construction