Introduction

Background

The real estate markets, like those in the USA, present an interesting opportunity for data scientists to analyze and predict where property prices are moving towards. Prediction of property prices is becoming increasingly important and beneficial. Property prices are a good indicator of both the overall market condition and the economic health of a country.

The problem

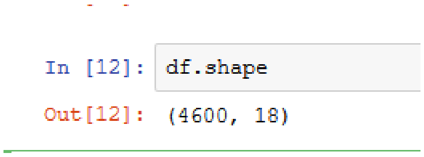
In this project we want to estimate the price of an house that have to been sold based on information regarding houses that have been previously sold. These information include the location of the house, its surface, its condition, the number of rooms and the number of floors and the year of construction.

Interest

Obviously the price of an house is of interest for both potential people that want to buy it and want to have a correct estimation of its price and moreover it's a good indicator of the status of the economy of a country

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.

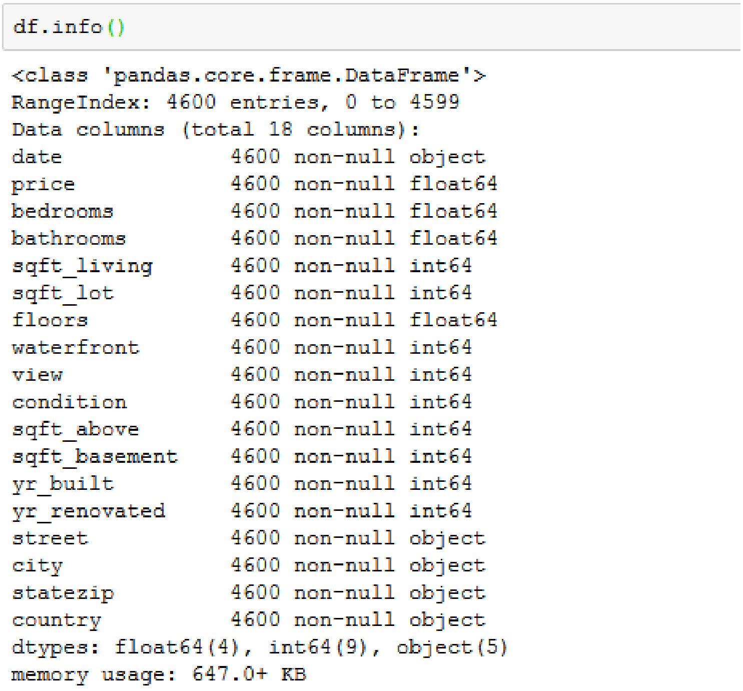
[1] You can find the data set at this website `https://www.kaggle.com/shree1992/housedata`

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

- 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.
- **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).

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



```
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

Methodology

Data preparation

First of all I clean up the set of data. I observed that the following

- The last column of the data-frame, the column *country* contains always the same information, namely *USA*. Therefore I will drop it from the data set since its information does not seem to be relevant.

- The first column of the data frame, the column *date*, has a limited set of values. The minimal values is *2014-05-02* while the maximal value is *2014-07-10*. I think that in a couple of two months the economical situation should not change too much. Therefore I will also drop this column during the data analysis.

- Finally I focused my attention on the price column, that is going to be the target variable. I report in figure 2 the corresponding histogram It's possible to see that, according to the data set, some houses have been sold with a price equal to zero. This clearly does not make sense. Therefore I decided to remove the corresponding rows from the initial data set. Moreover the distribution of the *price* column shows also the presence of some

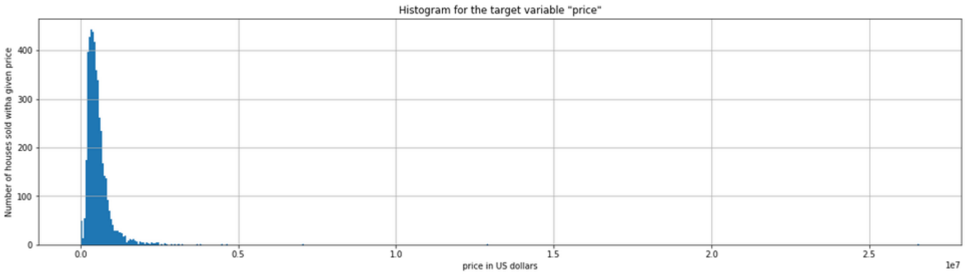


Figure 2: Price distribution

outliers. There is an house sold with a price of 26.590.000 dollars, another with a price of 12.899.000 dollars and finally a last one with a price of 7.062.500 dollars. Also in this case I decided to remove the corresponding rows

Data inspection

After the preliminary analysis described above I decided to get a general overview of the data and understand which of them could be related to target variable *price*. To this regard I created the *heatmap* reported in figure 3 that displays the correlation between the data.

Specifically the linear correlation coefficient of the most correlated columns are (see figure 4)

therefore it seems that the variable price is linear correlated with the column *sqft living* and the column *sqft above*.

In the following a will test this hypothesis using only the above mentioned columns. In order to this I randomly divide the data using the function *train test split* of the package *sklearn* in a train test and in test set, whose dimension is 20 per cent of the total. I then perform a linear regression on train set and I use the test set to make prediction I discuss the corresponding results in the following section.

Results

- Lets' first consider the correlation of the living space of the house and the price. I fit the data using a linear regression model. I reported the corresponding result in figure 5 the corresponding linear correlation coefficient is 0.484.

- Lets' then consider the correlation of the only space above in the house and the price. I fit the data using a linear regression model. I reported the corresponding result in figure 6 the corresponding linear correlation

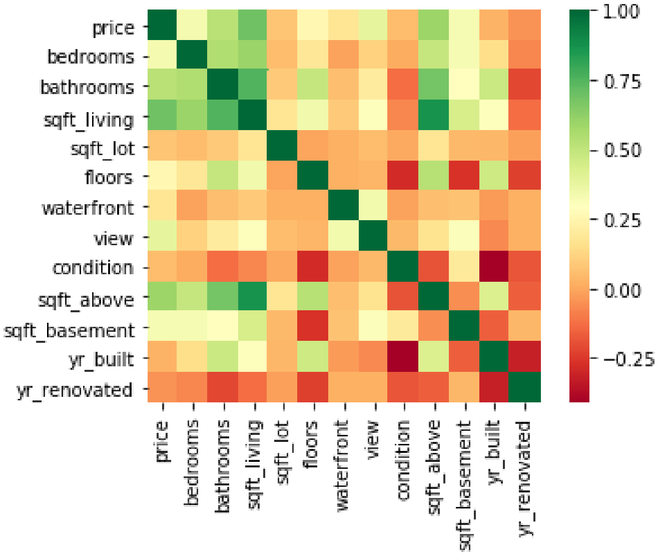
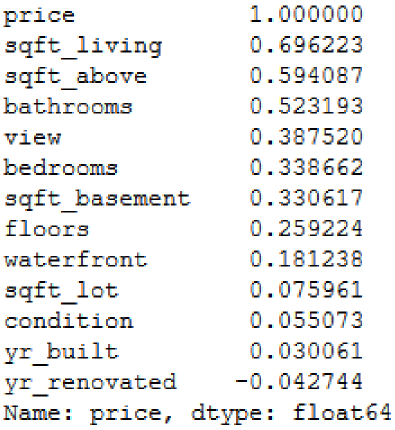


Figure 3: Heat map displaying the correlation between the data. A strong green is related to a positive linear correlation, while a strong a red to a negative linear correlation



```
price             1.000000
sqft_living       0.696223
sqft_above        0.594087
bathrooms         0.523193
view              0.387520
bedrooms          0.338662
sqft_basement     0.330617
floors            0.259224
waterfront        0.181238
sqft_lot          0.075961
condition         0.055073
yr_built          0.030061
yr_renovated     -0.042744
Name: price, dtype: float64
```

Figure 4: List of the correlation coefficients

coefficient is 0.352.

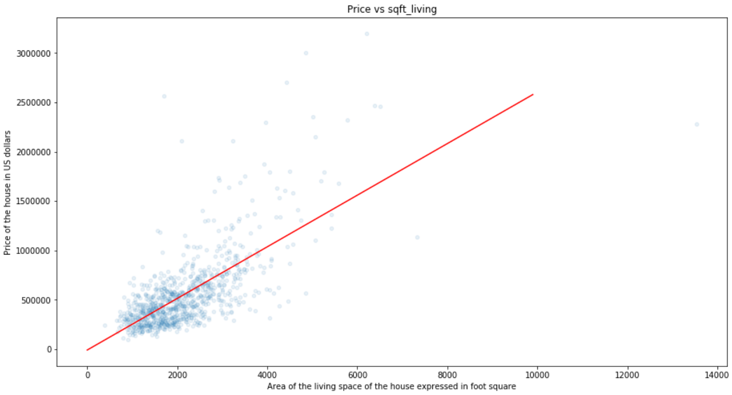


Figure 5:

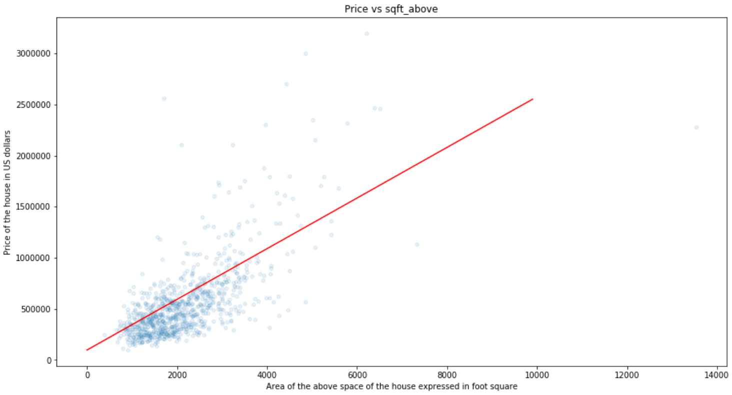


Figure 6:

Discussion

The above analysis seems to suggest that, as expected, there is a strong linear correlation between the price of the house and the corresponding area (both the living space of the house and the above space). it would be interesting to extend such analysis with a bigger data set.

Conclusions

In this study I analyzed the relation between the price of houses in THE Washington state and the other independent variables. In particular I focused my attention on the living surface of the full house and of the surface of the ground floor. it would be interesting to extend such analysis taking into account also other variables and try to fit the data with more complex model than linear regression.