Project Big Data
Assignment 4
Report on DC Residentials

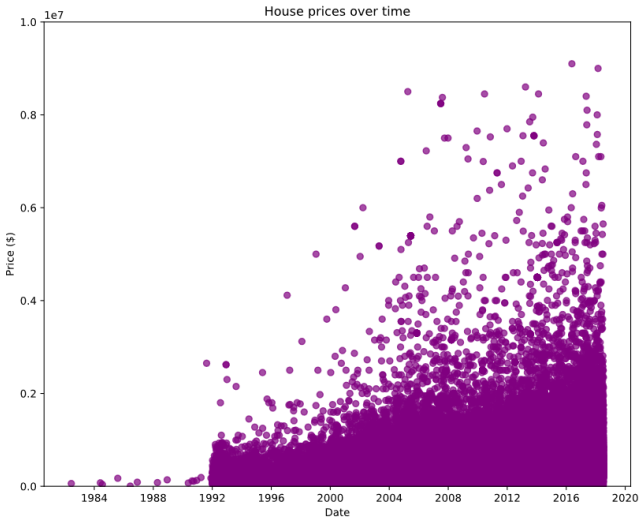
Yasin Aydin - 2657725 - yasinaydin@sabanciuniv.edu
Faruk Simsekli - 2657316 - faruksimsekli.7@gmail.com
Bob Mes - 2650287- b.mes@student.vu.nl
Atal Atmar - 2651146 - atal_atmar@hotmail.com
6/30/2019

Q1-How did the sale price of homes change over time?

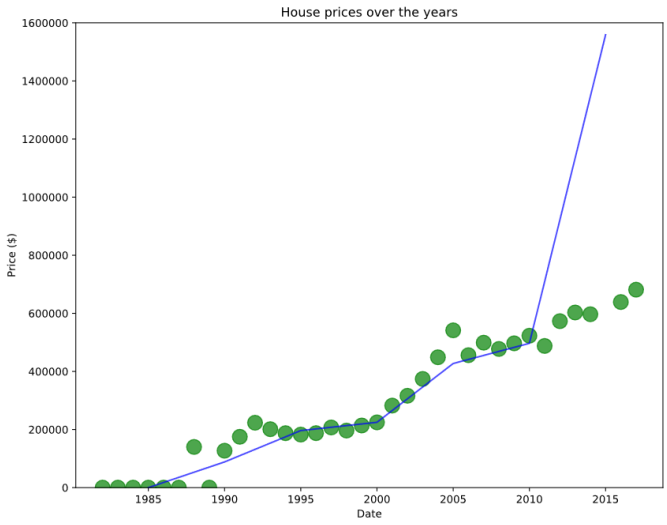
(price_change_each_year.py)

To investigate this question, we started with getting price and sale date columns from the dataset (nan values were ignored). First, we decided to create the graph of each 98.215 house that has been sold from 1982 to 2018 but the graph did not give us any reasonable clue about the mean of the sale prices of houses as seen in the *graph 1*. We can see the increase in maximum prices but the density below $100.000 can change the means.

Graph 1.



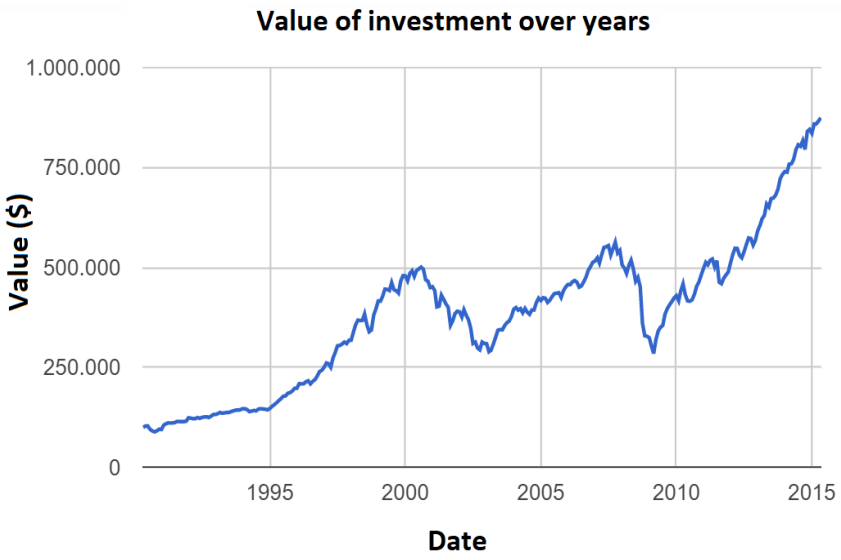
Graph 2.



So, we decided to get the mean sale prices of houses every 5 years. Following that, we created *graph 2* that has both years to year averages(scatter-green) and 5 years averages(plot-blue). There are some outliers that increase last 5 years average to 1.6 million but even without the outliers the increase over the years can be clearly seen. After seeing the increase over the years, we came up with a new sub-question that is useful for the investors that are living in Washington DC: “Was it more profitable to invest on DC residences or S&P 500?”. In order to answer this question, we decided not to use data before the year 1990 since the lack of data might influence the results. Also, we decided to take the average of 2015 without the outliers, which leaves us with 25 years (1990-2015). These constraints lead us to some values that we can work on. We found the historical data of the S&P 500 so that we can calculate the difference.

Our calculations showed that the $100.000 investment in 1990 would be worth $781.022 for S&P 500 investment and $722.784 for the residential investment in Washington DC. G*raph 2* shows that the residential price increases are more stable than the S&P 500 in *graph 3.* So, if the 7.8% difference is not worth risking it would be logical to invest in the residential. On the other side, the residential price means over years does not show that every building value increases that much so there will be some expenses to renew the house in a couple of years in order to keep it in “average house standards”. In conclusion, the residential investment might earn less money than the S&P 500 in the following 25 years but with less risk.

Graph 3.



Q2-How do different room types and land area affect the prices of the residences?

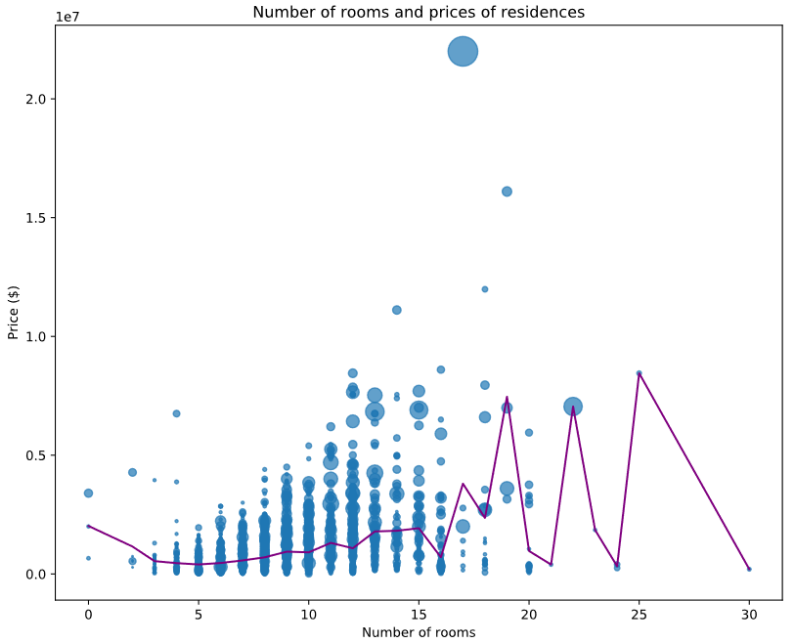
(price_and_room_types.py)

We started with ordinary least-squared (OLS) regression to see if we can create a model that fit and predict. We used the number of bathrooms, bedrooms, kitchens, and the total number of rooms that residence has as predictors. The results showed that the adjusted R-squared is **0.696**. The p-value's of the predictors were below 1% which indicates strong evidence against the null hypothesis so we can reject it. With the model we acquired the coefficients of the factors tell us that the number of rooms (in total), number of bathrooms and land areas are directly proportional to the price of the residence. On the other hand, the number of kitchens and the number of bedrooms are inversely proportional to the price of the house. This can give us a clue about how much buyers care about the number of rooms and room types.

In *graph 4,* the size of the bubble shows the scale of the land area. The model is not a perfectly fitting model but *graph 4* also shows the gradual increase (purple line) in the price with the land area and the total number of rooms. It is possible to predict the price of the house of those attributes with some error rate, but we can have more information with this graph. *Graph 4* shows that as the number of rooms and land area increases the maximum price of the house increases linearly. This does give us some clue of the maximum possible price a residence owner can get by selling his/her residence.

In conclusion, we found a linear model that fits and predicts residences by their number of rooms and land area. We also found out that there is a linearly increasing maximum price limit, which can help owners to predict their maximum profit.

Graph 4.

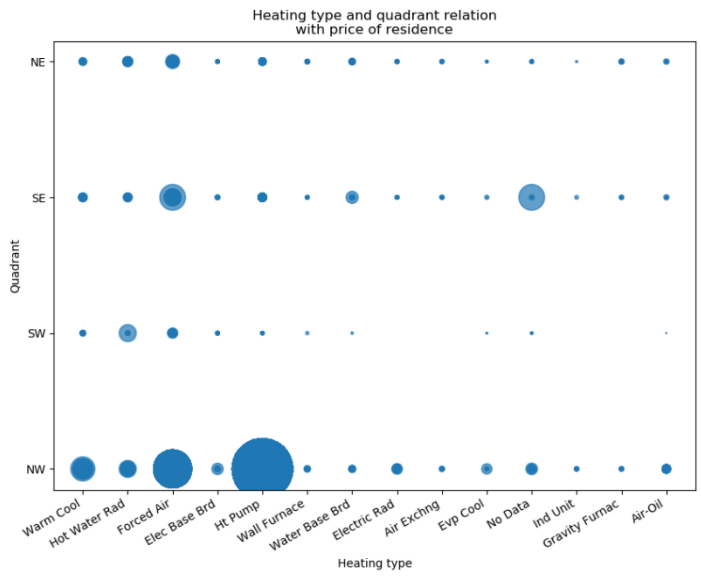


Q3-Do different sides of Washington DC prefers different heating methods and do this affect the price of the residence?

(heating_and_quadrant.py)

We created g*raph 5*, which has larger bubbles as their price increases, to see if we can find any pattern. *Graph 5* shows that the southwest quadrant doesn't have a single residence for some heating types which tells that they are not preferable in that quadrant as other quadrants have residences with those heating types. The larger bubbles show that the first 5 labels in the x-axis are preferable for high price residences in the northwest quadrant. *Graph 6* shows each heating type and the number of residences each has for different quadrants. In this graph, we can see

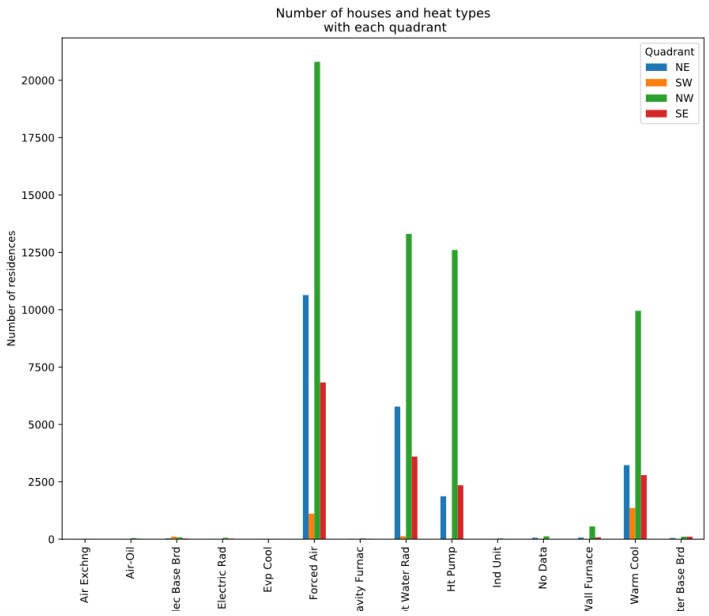
Graph 5.



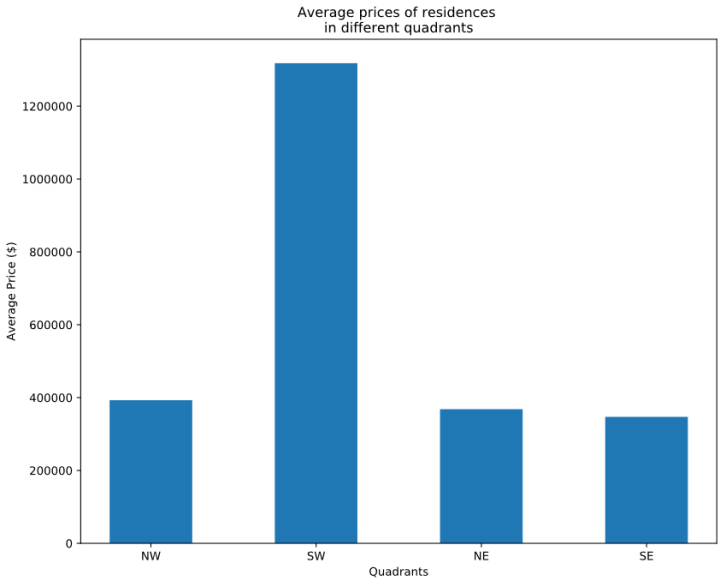
that the "Forced Air" heating type is the most desired heating type for Washington DC. The graph also shows that the number of residences in the northwest is much higher than other quadrants. As expected, the northwest doesn't have the highest average price as seen in *graph 7.* In order to see if there is any correlation between heating types and the price, we can use the highest average price quadrant (Southwest). *Graph 6* shows that the number of residences in the southwest quadrant is highest with the "Forced Air" and "Warm Cool" heating

type which are also preferred by other quadrants. So, this implies that it is not possible to detect house price by looking at “Forced Air” and “Warm Cool” heating types. There is one other deduction we can make from these graphs. As we can see in the *graph 6* the residences in southwest do not prefer “Hot Water Rad” and “Ht Pump” which differentiate average cost residences from more expensive residences.
In conclusion, we can say that the price cannot be determined by looking at residences’ heating types but there are specific 2 heating types that expensive residences prefer so that it is easier to say if a residence is not using those heating types they are more likely to be average priced residences.

Graph 6.



Graph 7.



Q4-How does the neighborhood of the home affect its sale price?

(price_per_neighborhood.py)

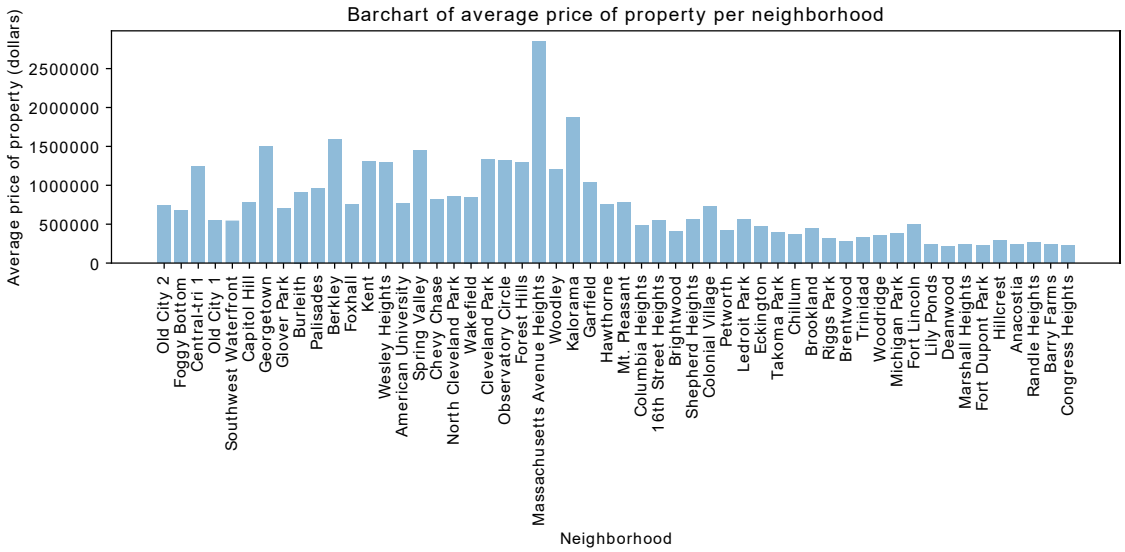
The dataset contains the price of properties in 33 neighborhoods. But before we can compare the neighborhoods with each other, we must investigate if there is enough data per neighborhood.

The neighborhood with the least amount of data is Massachusetts avenue heights. There is information about 112 properties in this neighborhood. This amount is large enough. So, no neighborhoods will be deleted. A one-way ANOVA test is executed to investigate if there are differences between the neighborhoods. The following hypotheses are composed:

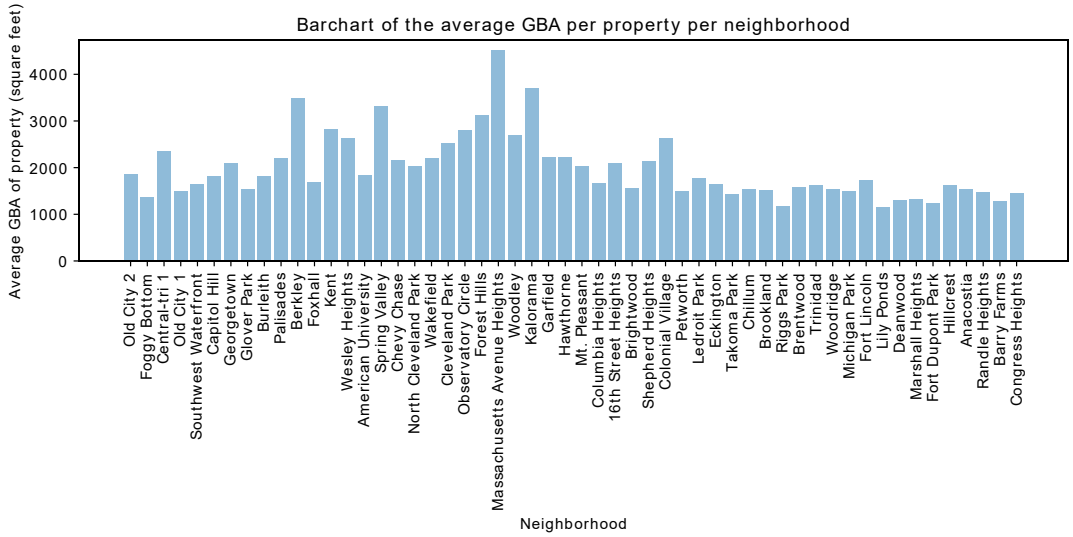
H0 = the price of properties in the neighborhoods isn't significantly different
Ha = the price of properties in the neighborhoods is significantly different

As the significance level, 5% will be taken. The test returned a P-value less than 0.001. Hence, it rejects the null hypothesis. There is a significant difference in price between the neighborhoods. The graph 8 gives the average price per property per neighborhood. This figure illustrates that there are big price differences between neighborhoods. Can we conclude that this price difference is solely because of the neighborhood or are there also other factors that play a role in this difference in price? As the sizes of properties in one neighborhood can be much larger than properties in other neighborhoods. The *graph 9* gives the average gross building area (GBA) per property per neighborhood. It seems to move the same as the bar chart of the average price. So, there is a chance that this is a confounding factor. To further investigate this, three ordinary linear regression models will be made and compared with each other. In *table 1* the R-squared values of these models are given. One model predicting price on only the neighborhood, one model predicting the price on GBA, and one model predicting GBA with a neighborhood.

Graph 8.



Graph 9.



Price and GBA seem to have a moderate linear relationship with each other. However, price and neighborhood and GBA and neighborhood seem to have a relationship equally as strong with each other. This indicates that the relationship between neighborhood and price goes via the relationship of neighborhood and GBA. Hence, we cannot conclude that the neighborhood has a real influence on the price of a property.

Table 1. OLS R-squared results

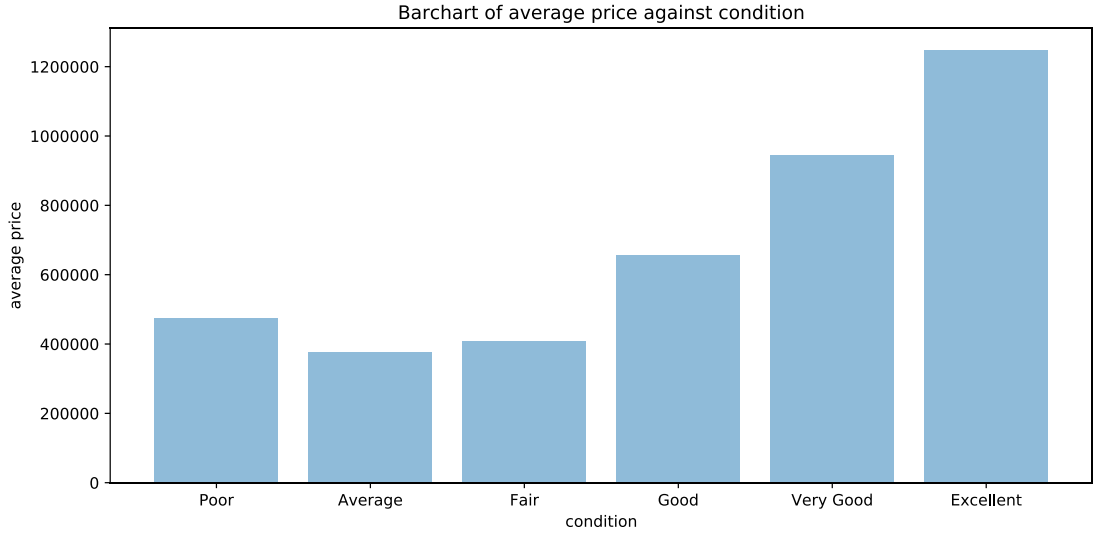
Ordinary linear regression:	R-squared
1. Price on neighborhood	0.337
2. Price on GBA	0.656
3. GBA on the neighborhood.	0.313

Q5- Does the condition of the property influence on the sale price?

(influence-condition-price.py)

This can be an important question for people that are going to sell their house. Is it logical to get their house in a better condition so that they can get a significantly higher price for it? These are the different house conditions encountered in the dataset: poor, average, fair, good, very good, excellent and default. The “Default” will be deleted because the sample size of this condition is too small (4 data points). Graph 10 illustrates the average price per properties with the corresponding condition.

Graph 10.



Graph 10 shows how better the condition, how higher the price of a property is. However, it is striking that houses with the poor condition have a higher average price than houses with an average or fair condition. To investigate these differences, we have done multiple t-tests.

Table 2. (T-tests of conditions)

	Poor	Average	Fair	Good	Very good	Excellent
Poor	1	$5.94*10^{-3}$	0.421	$5.62*10^{-4}$	$3.92*10^{-10}$	$1.45*10^{-10}$
Average		1	$2.88*10^{-2}$	$<1.0*10^{-10}$	$<1.0*10^{-10}$	$<1.0*10^{-10}$
Fair			1	$<1.0*10^{-10}$	$<1.0*10^{-10}$	$<1.0*10^{-10}$
Good				1	$<1.0*10^{-10}$	$<1.0*10^{-10}$
Very good					1	$<1.0*10^{-10}$
Excellent						1

Because of multiple hypothesis testing, the significance will need to be corrected. This will be done via the Bonferroni method. Hence, the new significance level is $0.05/15=3.33*10^{-3}$. With *table 2* it can be concluded that when your house is not in good condition (poor, average, fair) it can be lucrative to upgrade it to a good condition (good, very good, excellent). Also, it can be lucrative to upgrade the condition of your house while the condition is already good.

Q6-How does the age of the home affect its selling price?

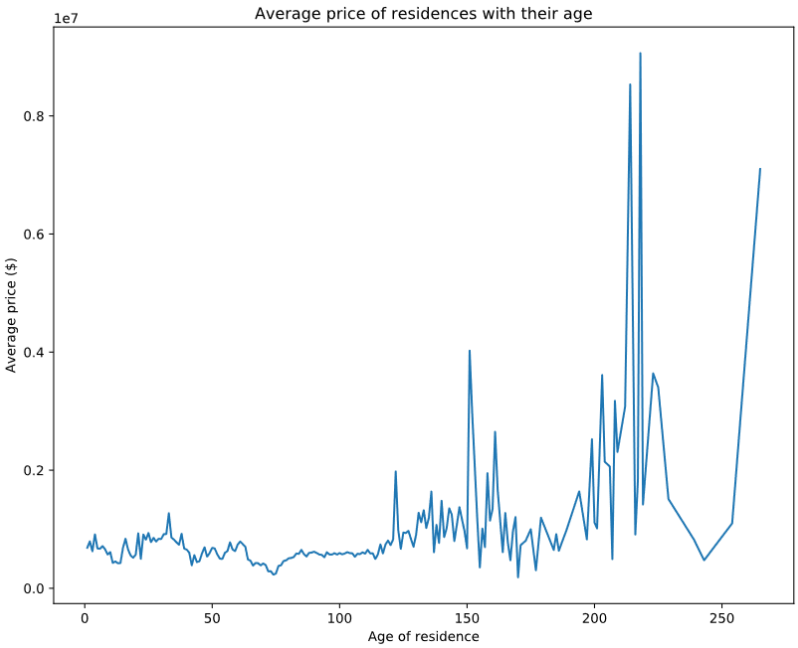
(age_price.py)

To determine what influence the age of a home has on its selling price, we calculated the age of each house. The dataset of DC residential only contains a column which indicates the earliest time (year) the main proportion of the building was built ('AYB'). So, to get a column which indicates the age of a building, we add a column to the dataset that subtracts the column AYB from 2019. This way, we know how many years the houses exist.

Then we performed a linear regression analysis on the age of the houses as an explanatory variable against the price of the houses as a dependent variable. First, we created a linear model and looked at the statistics. This model has an R-squared value of 0,009. This indicates that the model doesn't fit which means that the age of the houses does not affect the prices of the houses.

For the price based on the age, it could be helpful to get the average price per age (in years), because the density of the age against the price is very high. This way, you get a better view of how the price moves against the age.

Graph 11



In graph 11, we see that the price of houses increases with age. This can be because of different reasons. Maybe older houses are renewed and renovated more recently and so the price of the houses are higher because of the new quality. Looking at the graph, (without the errors) we could say that there is some sort of increasing relation between the age and the price, and it could be helpful in a model with other explanatory variables to predict the age.

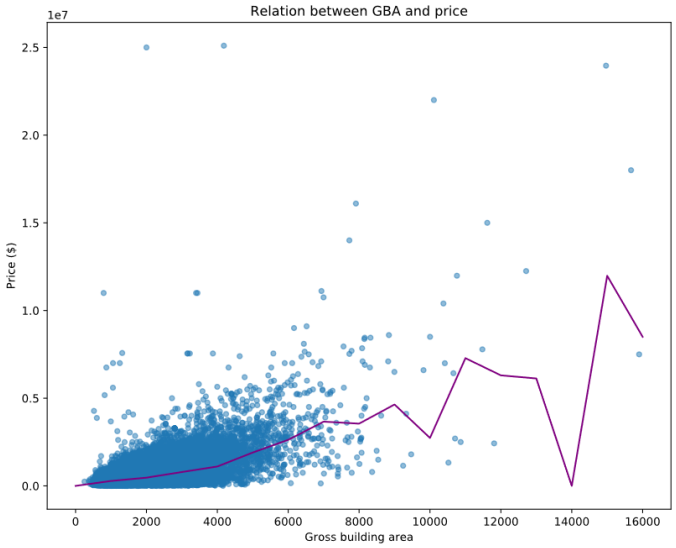
Q7-How does the gross building area of a house affect its selling price?

(GBA_Price.py)

What could also be important for predicting the price of a house, is its gross building area. Assuming that, bigger houses will be more desired than a smaller house. We created a linear model for the gross building area predicting the price of the houses. This model generated an R-squared value of 0,657 and a p-value of less than 1%. This means that the gross building area could be good to predict the price of a house, but it's not great. It can at least give an indication of how the gross building area affects the price.

When we look at *graph 12* of the gross building area against the price, we can see that there is some sort of linear relationship between the gross building area and the price. The line shows that the gross building area could indeed be used to predict the price, and maybe in combination with other factors, it will make for a good model to predict the price.

Graph 12.

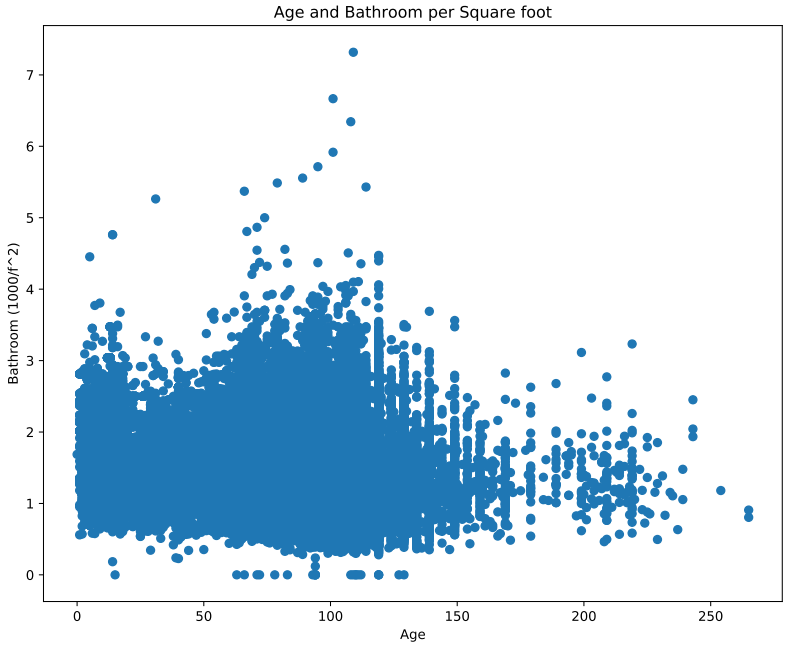


Q8. Do older homes have fewer bathrooms per square foot?

(bathroomPerArea.py)

Before we start the analysis, we think that there is no direct relationship between bathroom per square foot and age of a house, personally I do not even think that there is no correlation at all, they are independent variables.

Graph 13.

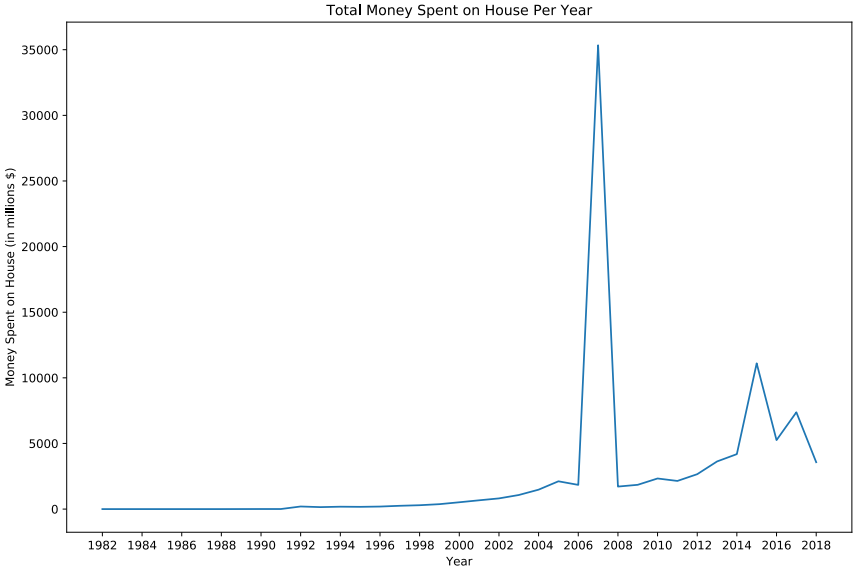


We plotted the variable in *graph 13*. Note that we considered bathrooms as 1 and half bathrooms as 0.5 in our dataset. And we calculated the age using 'AYB' in our dataset, which is the earliest time the main portion of the building was built. We could say that the number of sales of old houses are fewer than the new ones comparing the density of the dots. But, we could not say that there is a clear relationship between the bathroom per unit area and age by just looking at the graph. Yet, we could say that if we were to draw a best line for this graph it would be a constant function. Bathroom per unit area happens to be generally in (0.5-4). Note that we multiplied the actual number with 1000 to get a digit. Then, we created a linear model for these variables mentioned above. The results are as follows: R-squared value is 0.767 and p-value is smaller than 0.0001. In the light of the results, even though it is small, the coefficient suggests there is positive direct proportion. But, looking at the values mentioned above does not support that. Although we may have outliers, we easily see that these variables do not have a linear relationship. There are some new house with more bathrooms per unit area, there are also old houses with more bathrooms per unit area. So, it does not necessarily mean older houses have fewer bathrooms than the new ones. For this model, age is not a factor to predict bathroom per unit area or vice versa.

Q9. What is the relationship between the Great Recession in 2008 and housing marketing in Washington D.C?

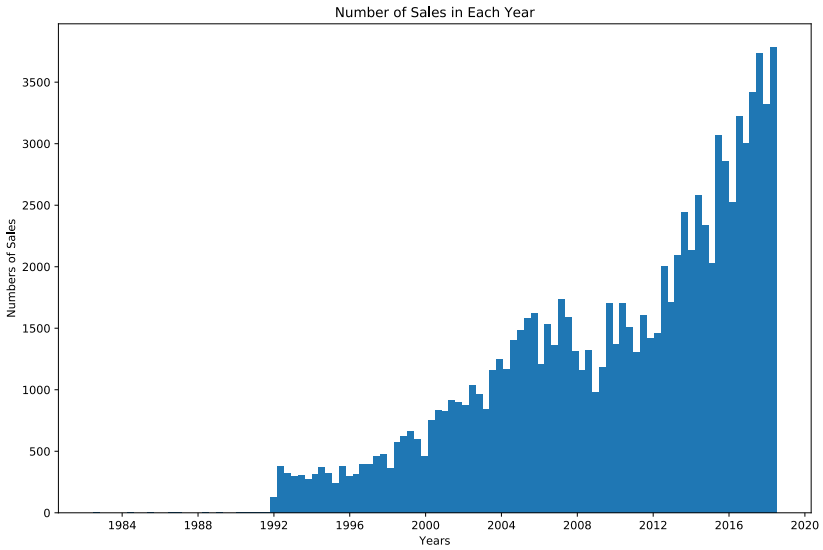
(recession.py)

The fundamentals of the housing market in the US. are rock solid, they said in 2005. After the trend of having the home as assets, a bubble of loans was getting bigger. The banks and government were missing something that would cause one of the most severe economic and financial meltdowns in US history. We will look at the relationship, by means effects and results, between house marketing and the Great Recession in Washington.



Graph 14

When we examine the *graph 2*, we see a decline in the house prices after 2005, which made more people spend much more money to buy houses in the following years. This can be clearly seen in the *graph 14* that while the total money spent on the house was approximately 1.843 billion dollars in 2006, that money became 35.338 billion dollars in 2007. The increase was nearly 1817%, which should have been good for the economy, but it was not. The problem was that most of the money spent was the loan. According to Wall Street, they gave millions of loan without verifying that that money is coming back. That year was the last straw that breaks the camel's back and the bubble blew out. This affected not only the housing market but also the whole economy. According to *graph 15*, the number of house sales has decreased by nearly 40% in the next periods. This decrease became 50% at most in some periods. Also, the money spent on buying houses (*graph 14*), shows that the decrease was almost 95% of the trade volume of the housing market in Washington. The relationship between the housing market and the Great Recession is different in each state. And this was for Washington D.C.



Graph 15

Now, we know that the deeper causes for the Great Recession are not only the decrease in house prices. And the answer is beyond the scope of this paper. But, we will also analyze the relationship between total money spent on buying houses and year as a sub-question.

Just seeing the graph, we can understand the relationship between year and money spent on houses in Washington D.C. Let´s look at the numbers. After performing linear model to the dataset which contains the variables, first we saw that the coefficient of the year, which is an independent variable, is 1.3504. Since it is positive, we gather that those variables have somehow direct proportion. Then the R-squared and p-values are 0.162 and 0.017 respectively. Actually if you assume that there was no recession from the graph you can see that the linear model would have fit better. Nonetheless, these values strongly indicate that the year could be used to predict the money spent on house sales in Washington D.C.

Contribution Table:

Yasin Aydin	Question 1-2-3
Bob Mes	Question 4-5
Atal Atmar	Question 6-7
Faruk Simsekli	Question 8-9

References

Reinvestment calculator of S&P500 (question 1):
https://dqydj.com/sp-500-periodic-reinvestment-calculator-dividends/
Great Recession (question 9):
https://www.curbed.com/2019/1/10/18139601/recession-impact-housing-market-interest-rates
https://www.thebalance.com/the-great-recession-of-2008-explanation-with-dates-4056832