Project Big Data
30-06-2019

Trends & insights into Washington DC Residentials'

Kelvin Blom -
Jimmy Gijsel
Dragos Pop
Tijmen Stultiens

VRIJE UNIVERSITEIT AMSTERDAM | De Boelelaan 1105

VU UNIVERSITY AMSTERDAM

Introduction

Summary of the report

This report exposes each step taken in order to acquire insights from the Washington DC residential data set and later leverage this knowledge with the final goal of creating profit. In the beginning, the given data is presented together with a definite purpose. Next, the data is examined and cleaned to make its handing faster for analysis. At this point, the report addresses the historical evolution of the prices followed by finding the factors that affect the cost of a house the most. Lastly, three suggestions to increase revenue and decrease expenditure supported by data are provided, without leaving aside the interpretations and the conclusion.

Given data and objective

Residential property descriptions and prices from Washington DC data set has over 150.000 distinct estates, with 48 attributes each, ranging from the gross area of the house to its location within the city, from the selling price to its style.

Based on this information, the objective of the assignment is to search for correlations, discover trends and gather insights that might be helpful to the involved parties. To define the goal in a clearer way, the involved individuals were determined as an American real estate agency or a common person interested in buying or selling a property in Washington DC. Afterward, their personal purpose was assumed to be maximizing their profit. Hence, the points that this report is trying to cover have the purpose of serving the individuals involved in real estate gain a financial advantage by leveraging data.

Data Cleansing

Taking into account the large size of the given data, an exploration was necessary to be performed prior analysis, in order to ensure the lack of significant errors. Accordingly, it was found that the data set contains pairs of columns that have the same information, for instance, GBA and LIVING_GBA, LONGITUDE and Y, or LATITUDE and X. Due to their duality, one element of each pair was eliminated.

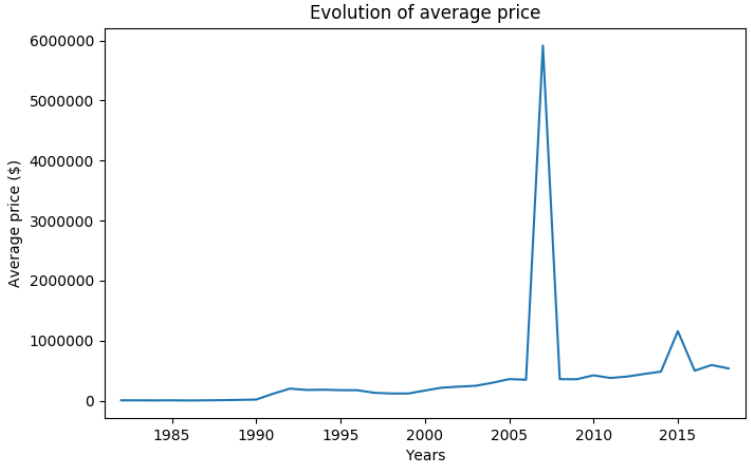
Another discovery was that there are attributes that have the same value for all the entries, such as CITY and STATE which were filled only by Washington and DC respective. Hence, they were erased from the data to simplify its handling in the analysis.

Lastly, NaN values were seen randomly distributed in the attributes of most properties. It was later decided that the most appropriate way to deal with this problem is by not taking into consideration the houses that do not have a value in the columns on which the analysis is based.

Houses price trends

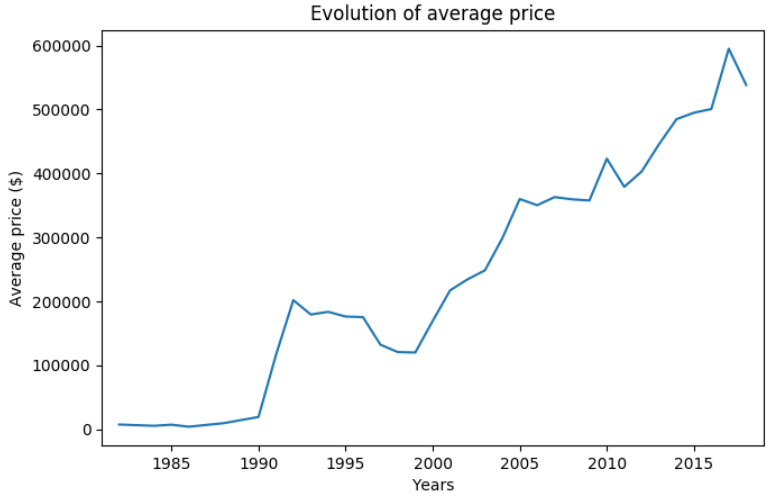
Evolution of the average price

The first step done towards the goal of the analysis was understanding the way houses prices are evolving over time, their trend and periods in which they deviate from the direction. The most relevant and easiest method of aggregation, in this case, is the average, and since the data contains property sales on a time span of over 30 years, the average selling price per year was plotted.



As one can see in the picture above, there is a curious large value between 2005 and 2010, which indicates that the average selling price was around $6.000.000. A possible explanation might be that several rich people bought mansions in Washington DC right before the start of the financial crisis. However, this point is not supported by any evidence so we investigated it further.

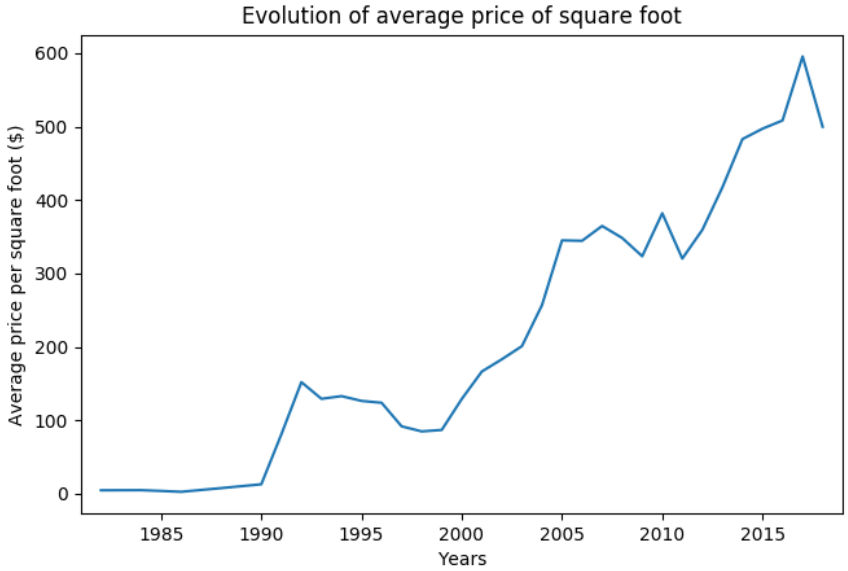
Extracting the most expensive houses from the Washington DC properties data gave the following three values as the top: $137.427.545, $53.969.391 and $53.696.391. In spite of that, an article found on the internet ("Most expensive mansion in Washington DC in 2018 $63 million" – The Billionaire Shop) states that the two most costly villas in the city are a $63 million mansion and Jeff Bezos's $35 million residence. Considering the latter, we decided to eliminate the first three most valuable houses and generate the plot once again.



In this form, the line plot seems to better represent the actual trend of average prices. The image shows an upward tendency in the mean selling price, as expected, taking into account inflation. It has a steady period from 1982 to 1990 where a strange average price of 0 is indicated, followed by a sharp rose until 1992. Then, a large fall had taken place before the beginning of 2000, showing a roughly 50% decrease in the average price. Afterward, there are no other significant changes in the ascending trend, other than the downfall around 2010, which is most likely due to the economic recession.

Evolution of average price per square foot

Based on the previously discovered outcome, one might justify that the rose in average price is caused by the transition towards larger houses over smaller ones. In order to better show the historically values evolution with respect to this point, a more general and popular measure unit was chosen, namely the average price per square foot. Since the given information contains values representing the surface of a property, the same process as before was followed, the only difference being calculating the price per square foot of each house before taking the yearly mean. Consequently, the following plot was drawn.



The illustration looks very similar to 'Evolution of average price', excepting a more accentuated decrease around 2010. Other than that, the same upward trend is remarked, with two significant downfalls, one from 1992 to the beginning of the 21st century and another one during the economic crisis started in 2007. It is also noteworthy that the average price per square foot decrease with $100 (16%) in the last year, which perhaps is due to the fact that the data set was made available before the end of the year, and, thus, does not contains all the sales that were completed during the twelve-month period.

In conclusion, one can easily recognize the ascending tendency of property prices in Washington DC, increasing with over 500% from 2000. However, adjusting the prices to the inflation rate is the next step recommended in case someone wants to take advantage of the uphill trend and invest in real estate.

Influential factors

It was revealed that the prices of properties in the federal capital are going up, so the next paragraphs address the factors that have the largest effect in this transition. To determine these influences, the Kendall rank test was regarded as the safest option to test for correlation because it does not have to check the multiple assumptions that Pearson's test demands and has a distribution with better statistical traits compared to the Spearman's rank test ("Kendall's Tau and Spearman's Rank Correlation Coefficient" – StatisticsSolutions). Therefore, it was used to evaluate the correlation between price and all the other attributes of the property. The most influential factors according to the Kendall test are presented in the table below.

Factor	Correlation with price
1. Gross area of the house	0.34
2. House condition	0.32
3. Sale date	0.30
4. Grade	0.28
5. Style	0.23

The first factor is not a surprise because it is obvious that the price of a regular property is in the first place determined by its size. Moreover, it was already displayed that as late the sale date is, as much the price is increasing and it will be later discussed the possibility that the time of the year also plays a role in the average price. The house condition is a relevant factor for the price, as well, for the reason that someone will be more likely to pay more for a house in a better condition, but this point will be evaluated in more detail in the last part of the report.

On the other side, we expected the location within the city (WARD) to be a decisive factor. However, the Kendall test evaluated its correlation with the price equal to 0.22, surprisingly close to the unpredictable cooling type of the property (AC), 0.18.

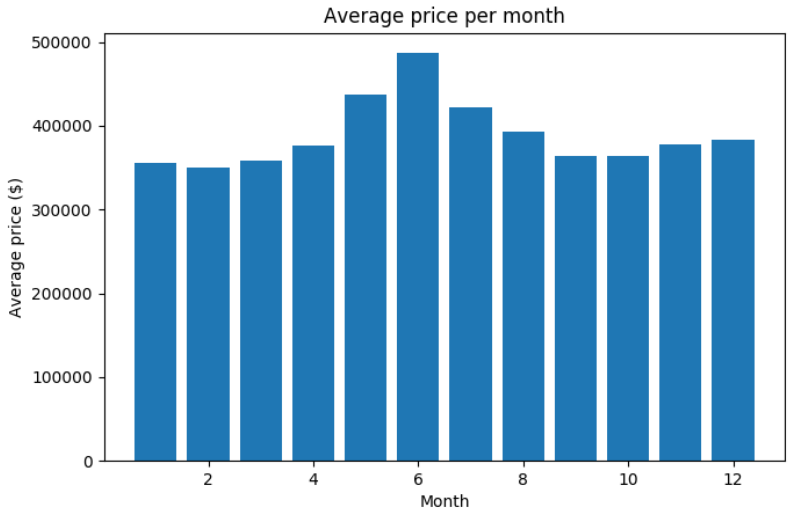
Having the factors with the highest effect on the price, they were later analyzed with the purpose of finding the situation which leads to profit. In other words, the most influential factors were examined to gather insights because they have proportionally the largest probability of gaining great revenue.

Insight gathering to make profit

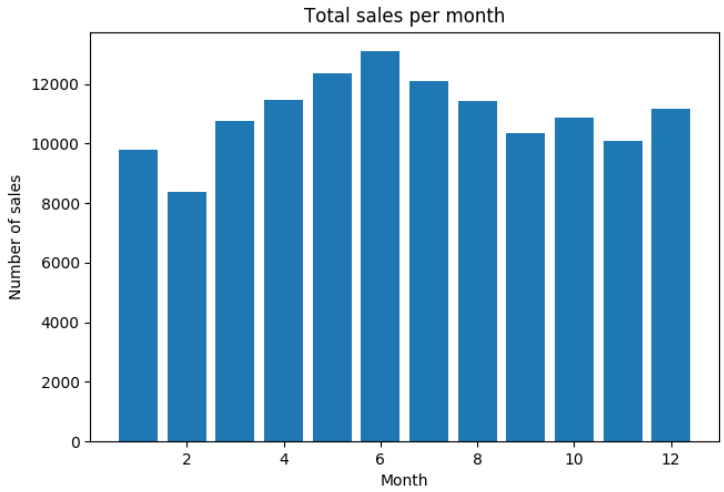
Best time to buy/sell

The sale date has a relatively high correlation with the price, 0.3. It would then be profitable to buy a property in a period where the averages are low or sell a house when they are high. For this reason, it was necessary to aggregate all the sales in a month and compare the outcome to the averages of other months to see if there are certain periods during the year when the prices means are higher or lower.

In chronological order, the sales were first filtered by month and the average price of sales in each month was calculated. To make the contrast visible, the values were represented in a bar chart:

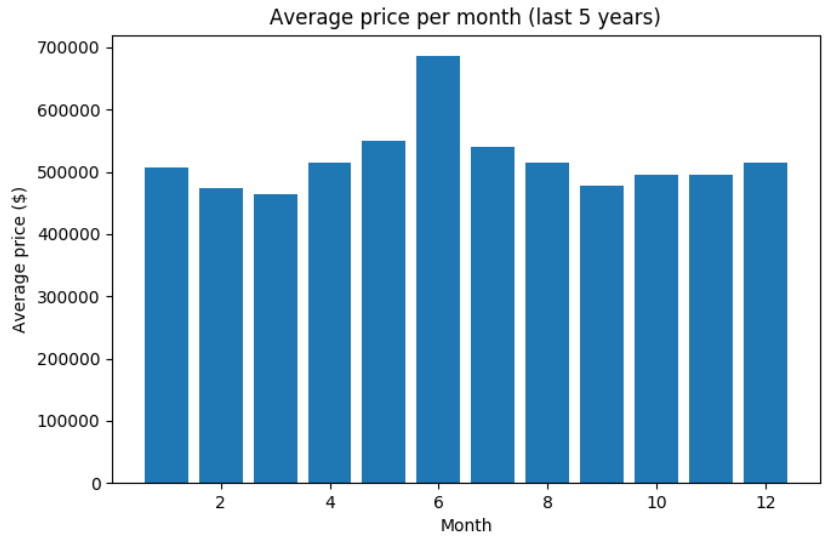


According to the image, the average price in June is the biggest during the year, while February has the smallest value. It is also noteworthy that the prices are increasing during the spring and one possible explanation for this might be that more people are selling their home and, as a result, the competition becomes more fierce and the prices are rising. To check the veracity of the previously mentioned justification, the total number of sales for each month was computed and then inserted in a bar chart.



As anticipated, more homes are sold during the end of spring and beginning of summer than in any other time of the year. This implies that the main reason average price increase has to do with the greater number of houses that sell during that period.

The averages represented in the "Average price per month" graph are calculated using the data over a 30-years duration. However, it would be more financially rewarding to check the trend of a more recent period. For this reason, the same algorithm was applied once more using only the sales done starting from 2014.

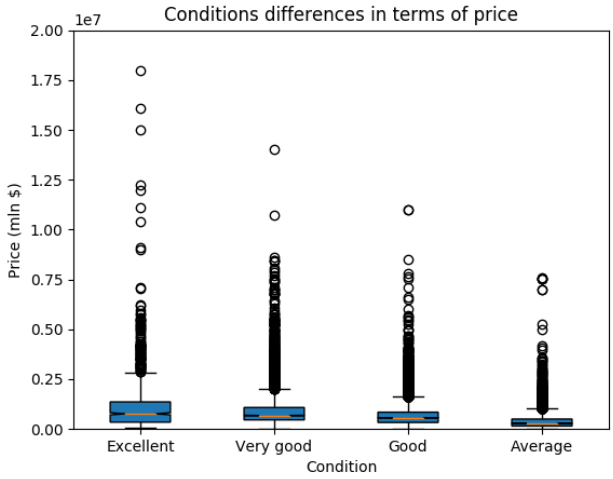


The updated chart looks almost identically with its 'all-time' correspondent, but the lowest average moved one month later, in March. In addition, the difference between the mean price in June and all the other months is even larger in the last 5 years than it is in the other bar chart.

As a conclusion, a person interested to buy a house has a better chance to get a cheaper place in March. On the other hand, the seller would perhaps earn more money if he/she succeeds to trade his/her home in June. However, the respective person should take into consideration that the number of properties for sale will also be higher at the beginning of the summer, therefore, the chance to sell goes down.

Difference between conditions

There are four types of conditions: excellent, very good, good and average. It would be useful from a financial perspective to check if the better conditions are costing more or if there is a way to pay the same price for a superior condition. In order to evaluate that, we tested every two pairs for significant differences in their medians in terms of price. The reason median was chosen over the mean is that the latter is more sensitive to outliers (robust) and all the conditions have several, as it can be seen in the figure below.

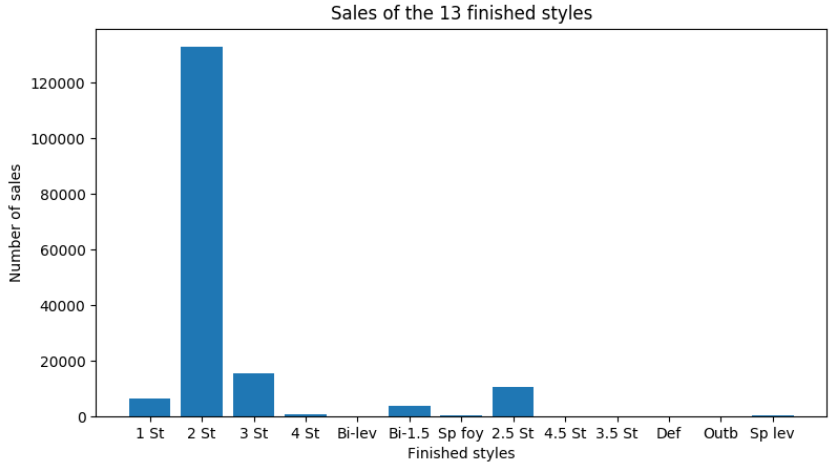


The test performed to evaluate if there are significant differences between the medians of the pairs was Kruskal-Wallis because the other alternative, Mann Withney U, takes into consideration the spread of the sample too, which is not desired in this case. The only pair that resulted in a p-value larger than 0.05 was Excellent and Very good, meaning that there is no significant difference between the medians of the two, while all the other pairs led to substantial distinctions. The same conclusion can be extracted from the zoomed in boxplots.

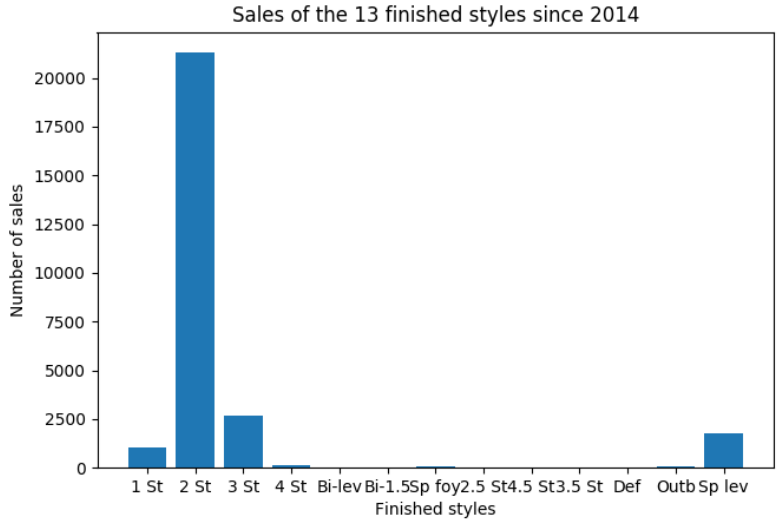
The interpretation of this result is that there are no reasons to buy a house in a very good condition since, for the same price, an excellent property can be purchased. In addition, the strategy according to which a person upgrades his/her residence from very good to excellent to make a profit is discouraged because there is not a significant difference between the market value of both conditions and it is assumed that the renovating costs are not 0. However, since no data concerning the renovating costs are available, it is up to the individual to make a decision.

The style that sells the most

Another profitable way to approach the real estate industry is by matching the offer with the demand. In other words, a real estate agency or an American interested in flipping properties are more likely to sell a house if it satisfies the customers' desires. Regarding this, the buildings in the data set have 18 different styles, with only 13 of them being finished. Thus, it was examined the number of sales for each finished style to see if there is one that sells better than the others.



From the illustration, it is clear that the 2 Story style is the most demanded among customers so it would be advised for someone who wants to flip a house fast to get this respective style. However, it is probably the case that the trends nowadays are not the same with the ones from 20 years ago, hence, the sales of the 13 finished styles were calculated using only the transactions from 2014 onwards.



Again, the 2 Story houses are at the top of the sales, implying the same conclusion as the last bar chart. Comparing the two graphs, one can also notice the recent decrease in the transactions of 2.5 Story buildings and a sharp increase in Split level properties, indicating that many Americans are moving towards smaller homes.

Conclusions

Based on the above-presented figures, it can be concluded that a person interested in buying a house in Washington DC can benefit financially if he/she purchases a residence in the autumn or winter when the average prices were proved to be lower than they are in the summer and spring. Moreover, the respective individual is advised to buy a property in an excellent condition instead of a very good one because the two have a very similar market value.

When it comes to a person that is looking to sell his/her residence in the federal capital, the ideal time to make a transaction is opposite, namely spring and summer, with the highest prices in June. However, it should be noted that in the same period more houses are put on sale so exchanging the property for money becomes more complicated. Besides that, it was demonstrated that upgrading a house from very good to excellent will bring more expenditures than revenues since the renovating services are not free and two conditions trade for a close price.

Lastly, the report offered the numbers that indicate that the most sold type of property is the 2 Story style. Hence, the probability of selling a building with this style is higher than the others. However, the competition is more fierce, so further analysis is suggested in this situation.

References

- “Most expensive mansion in Washington DC in 2018 $63 million” – The Billionaire Shop: https://www.thebillionaireshop.com/most-expensive-mansion-in-washington-dc-in-2018-63-million/
- “Kendall’s Tau and Spearman’s Rank Correlation Coefficient” – StatisticsSolutions: https://www.statisticssolutions.com/kendalls-tau-and-spearmans-rank-correlation-coefficient/