Choosing a place to live in Washington DC

Abstract

Moving to a new city is always painful. Choosing an area in a new city where to live is mostly done by the impression we take from reading disperse information that we find in Internet or by our own lived experiences. However, big data can analytically help on this issue.

The project here presented deals with this problem. It is assumed that the user would be mostly interested in knowing about the safety, education, economic viability and the entertainment offered by the different areas of the city.

The four building blocks (safety, education, economic viability and the entertainment indexes) of this project will be implemented by scraping data from different sources (such as http://data.octo.dc.gov or yelp repository) and taking into account the following issues:

1) Safety: This index analyzes the criminal activity in the different wards of the city.

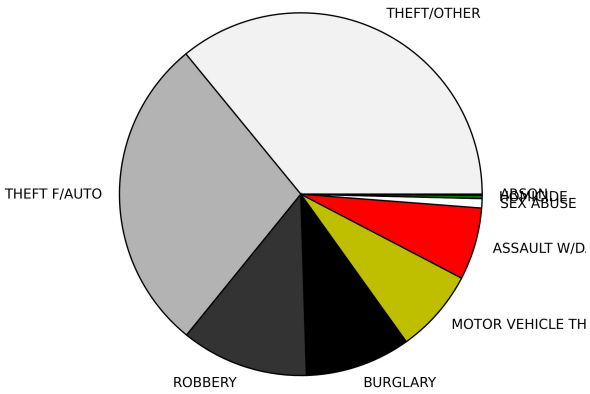
2) Education: The number of schools, kindergartens, and other educational centers in each ward is analyzed to provide an estimation of how easy the access to a good education for the user family members is.

3) Economic viability: The number of available housing and the prices for renting or selling will be addressed by this index.

4) Entertainment: The service offer such as well-rated restaurants, theaters, life-music bars, or the access to green areas, gyms, outdoor activities among others will also give the user one more useful index to choose his/her right place to live.

Data Story I

How safe is Washington DC to their inhabitants? Let's do a statistical analysis in order to shed light into that question. For this investigation, I have selected a data file that comprises the crime reports registered in Washington DC in 2013. This year there were 35896 reported crimes distributed as shown Figure1.



	Percentage_%
THEFT/OTHER	35.923223
THEFT F/AUTO	28.228772
ROBBERY	11.346668
BURGLARY	9.385447
MOTOR VEHICLE THEFT	7.457655
ASSAULT W/DANGEROUS WEAPON	6.438043
SEX ABUSE	0.832962
HOMICIDE	0.289726
ARSON	0.097504

Figure 1 and Table1. Type of offenses distribution reported in DC in 2013.

The Figure1 clearly shows that thefts are the most frequented reported crime: they involve 92.4% of the total.

These insights make us question which is the probability of a citizen to be victim of a crime during this year. Let me define two random variables: X= to be 100% safe in DC and Y= to be attacked. Then the P(X) is the subtraction of P(Y) to 1.

$$P(X)=1-P(Y)$$

P(Y) = (number of reported crime incidents / total of DC inhabitants in 2013[1])

This yields P(X)=94.5%. Now let's see what are the probabilities for any type of incident to occur to an inhabitant. Table2 shows a summary of these probabilities. We can see that the probability of being a homicide victim was 0.02%. These values are wholly reassuring, but let's keep looking for interesting data. What do we know about the reported crime distribution along the 8 different neighborhoods of the city? Figure 2 shows all

Offenses	Probability_%
THEFT/OTHER	1.99
THEFT F/AUTO	1.57
ROBBERY	0.63
BURGLARY	0.52
MOTOR VEHICLE THEFT	0.41
ASSAULT W/DANGEROUS WEAPON	0.36
SEX ABUSE	0.05
HOMICIDE	0.02
ARSON	0.01

Table 2. Probabilities summary per type of crime in DC 2013.

[1] https://es.wikipedia.org/wiki/Washington_D._C.
[2] http://www.neighborhoodinfodc.org/wards/wards.html

the registered incidents classified by ward. This plot indicates that ward2 had the highest number of offenses registered in 2013 while ward3 had the lowest.

And other interesting question to answer would be what was the probability of being attacked in any ward. Table3 summarizes some relevant data that is going to be discussed hereunder. The first column quantifies the percentages shown in Figure2 showing the number of incidents per each ward. In the second column we find the population.[2] Finally the last column contains the probabilities of being involved in a criminal act. In order to make it friendlier, the probabilities are illustrated in bars in Figure 3.

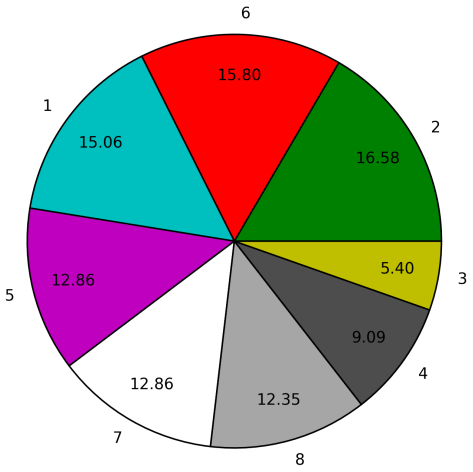


Figure2. Pie plot showing the percentage distribution per ward for all the registered offenses in DC during 2013.

	Number of incidents	Population	Prob_of_any_incident
Wards			
Ward1	5407	74462	7.26
Ward2	5953	76883	7.74
Ward3	1938	78887	2.46
Ward4	3263	75773	4.31
Ward5	4616	74308	6.21
Ward6	5670	76000	7.46
Ward7	4615	71748	6.43
Ward8	4434	73662	6.02

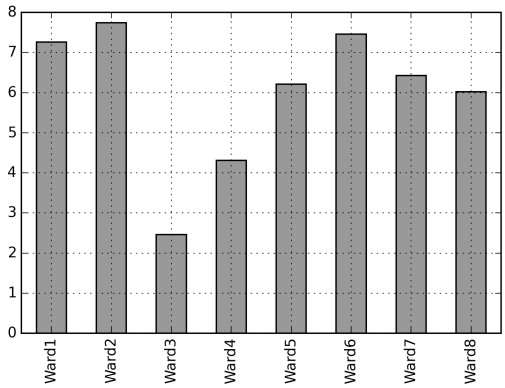


Table 3 and Figure3. (left)Statistical summary for the criminal activity per ward (year 2013) and (right) bar plot with the distribution per ward of the probability (in %) of being involved in a crime situation.

A deeper analysis opens new questions about the frequency of some particular type of offenses. Which was the ward with lower number or homicides or less auto theft registered? Let's see this in the next two plots summarized in Figure4. Ward3 does not appear in the homicide bar plot as there were not reported homicides in this area. On the other side, the auto theft bar plot on the right of Figure4 shows that the wards 3,7 and 8 were the areas with a lower number of auto thefts while ward 1 registered the highest recorded value.

[2] http://www.neighborhoodinfodc.org/wards/wards.html

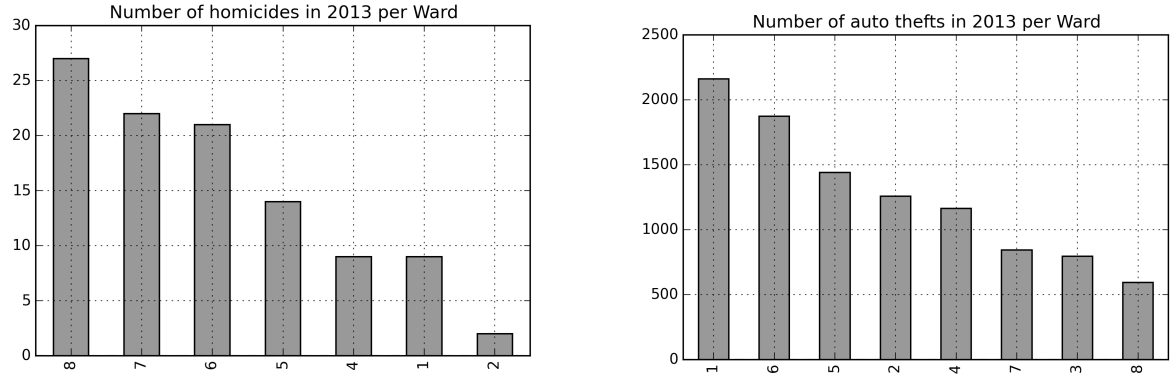


Figure4. Number of homicides (left) and auto thefts (right) filtered by ward.

Now, we are going to analyze the reported activity in function of time. Let's see whether there is a pattern behind of this. The line plot shown in Figure5 (top-left) shows the tendency of the reported number of incidents per month. It is easy to see that the warmer months had more activity than those that were colder.

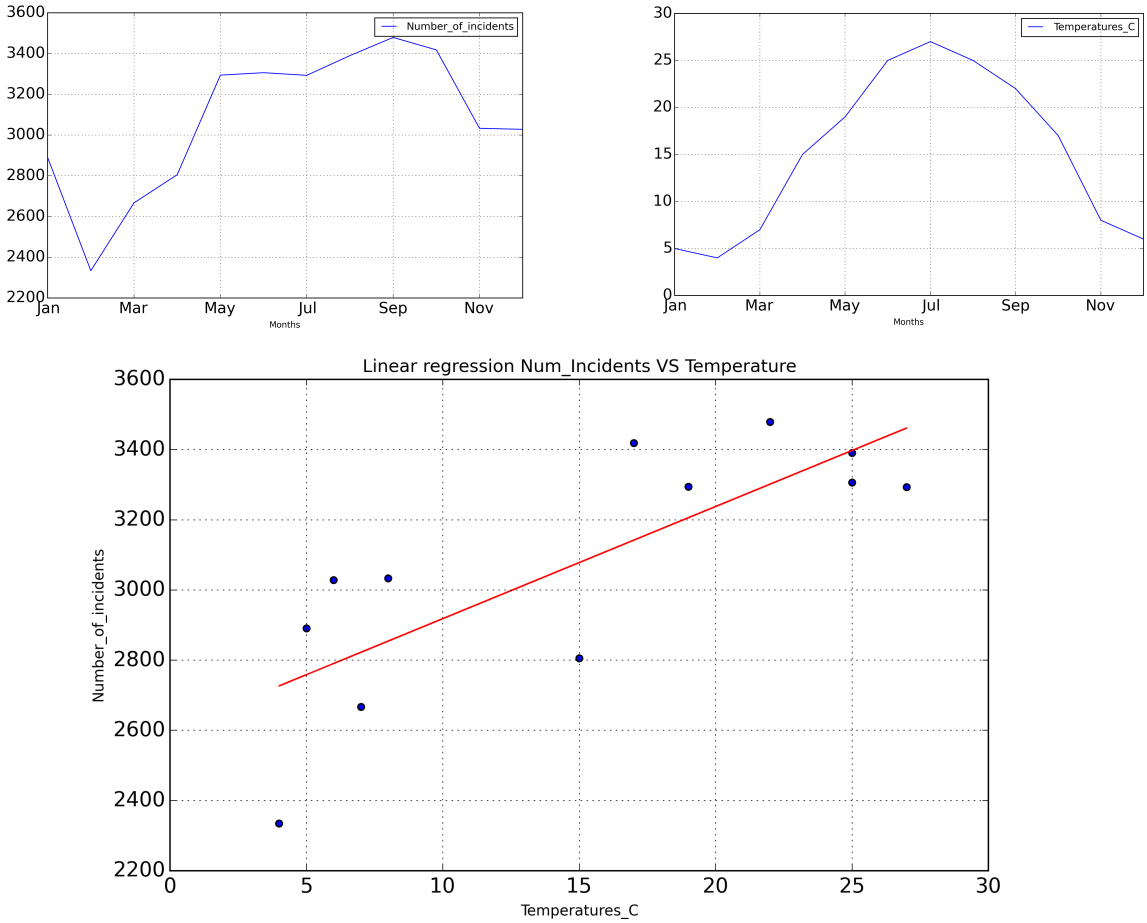


Figure 5. (Top-left) Line plot showing the reported offenses classified by month. (Top-right) Average temperatures per month in 2013. (Bottom) Linear regression between the number of registered crimes and the average temperatures collected in 2013.

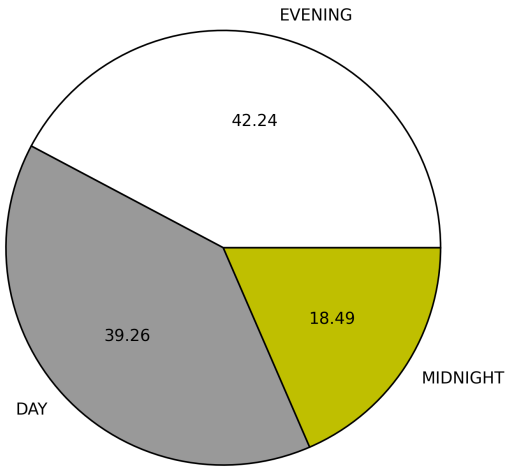
A look over the top-right plot in Figure5 suggests that a correlation between the weather conditions and the criminal activity could exist. The linear regression between the number of incidents reported and the temperatures registered is shown in the bottom plot of Figure5. This graph displays a positive correlation with a R-Sq value of 0.62. This trend is easily observable by classifying the monthly registered crimes into two groups: months with a number of incidents registered above the mean (4487) and those that are below the mean. Table4 summarizes these two groups: the highest temperature recorded with a crime number below the mean is 15ºC while the lowest with a crime number above the mean is 17ºC.

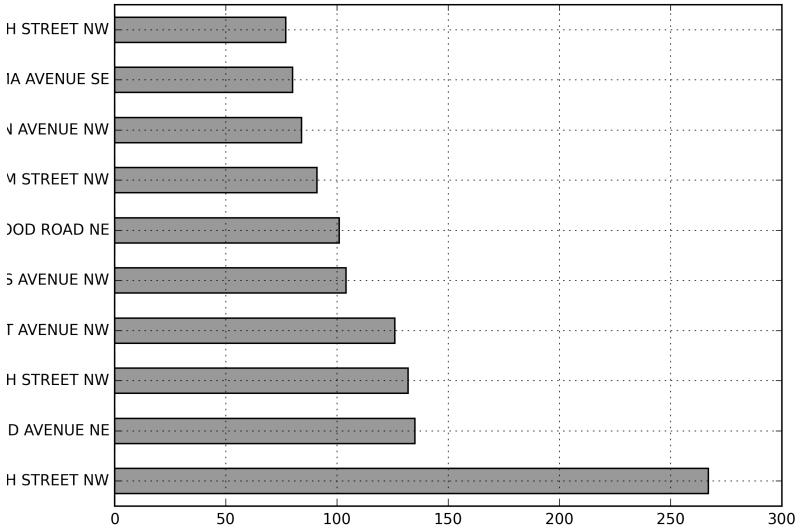
	Number_of_incidents	Temperatures_C
Months		
Jan	2890	5
Feb	2334	4
Mar	2667	7
Apr	2805	15
Nov	3033	8
Dic	3028	6

	Number_of_incidents	Temperatures_C
Months		
May	3294	19
Jun	3306	25
Jul	3293	27
Aug	3390	25
Sep	3479	22
Oct	3418	17

Table 4. Number of incidents and temperatures corresponding to the months with a number of reported crimes below (left) and above (right) the mean.

Two more questions to answer concern the shift of the day with more records and which was the most frequented street. Figure6-left shows that the evening and the day had more incidents than the midnight while the bar plot at the right shows the top-10 most frequented streets for criminal activity during 2013.





Questions answered:

1) What were the reported types of offenses registered in DC at 2013 and their percentage contribution to the total criminal activity recorded?
2) What was the probability of being a citizen of DC and not having reported a crime during 2013?
3) Which percentage of the total criminal activity took place in each ward?
4) How many incidents were registered per ward?
5) What was the probability that a citizen reported an incident in his/her ward?
6) Which is the ward with fewer cases of homicide? And theft autos?
7) Which are the most active months?
8) Is there any trend behind the monthly classified crimes and the temperature?
10) In which shift of the day did the police collect more incident records?
11) Which are the streets with a higher registered criminal activity?