BIG DATA ANALYTICS REPORT - INDIVIDUAL PROJECT 1

STUDYING ABROAD: FACTORS INVOLVED

December 20, 2016

Pratik Shirish Kamath (N14671569)

Contents

1 CHICAGO CRIME HEAT MAP ANALYSIS

I was fortunate to get a huge dataset for Chicago which allowed me to do lots of analysis in information visualization. I also happened to get the shapefiles for Chicago from the internet. In the dataset for Chicago, there were many such records where there was no values for longitude and latitude and therefore first I had to clean those up and when the dataframe was ready to go I started off with my visualization part.

In order to get acquainted I started off with the dot distribution plot using the ggplot2 library. It is a very simple way of showing things as to how everything is lined up in the map.

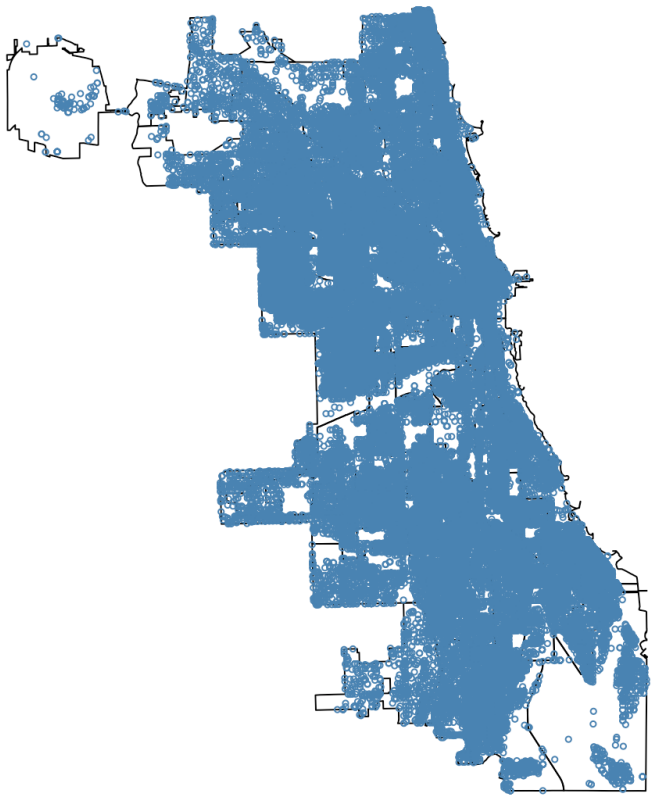


Figure 1

However, I figured it is not good for doing any sort of analysis. Therefore, I went ahead and thought of plotting the dot distribution graph along with longitude, latitude values and

gave it a title and x-axes and y-axes co-ordinates as follows using ggplot2 library and also plotted a dot distribution graph on the map using ggmap library.

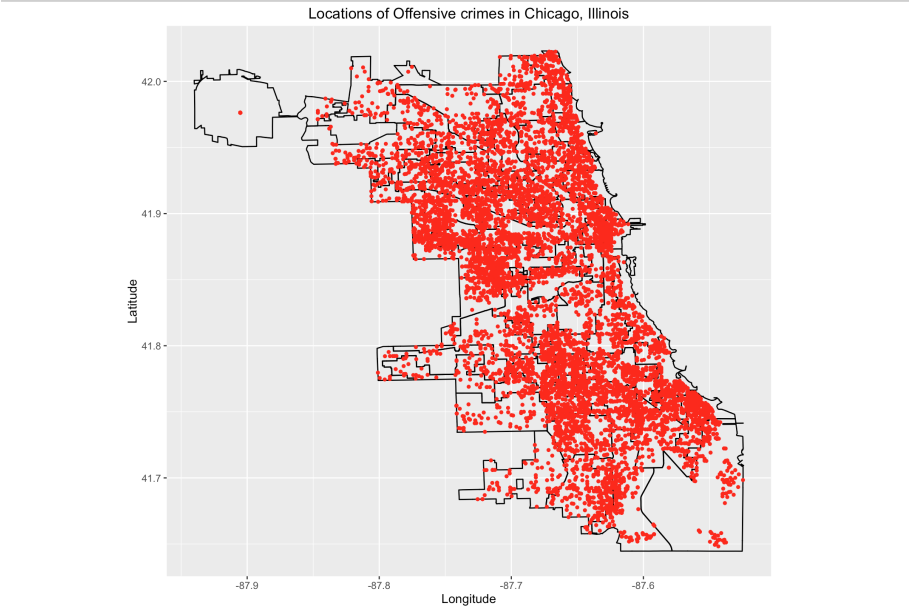


Figure 2

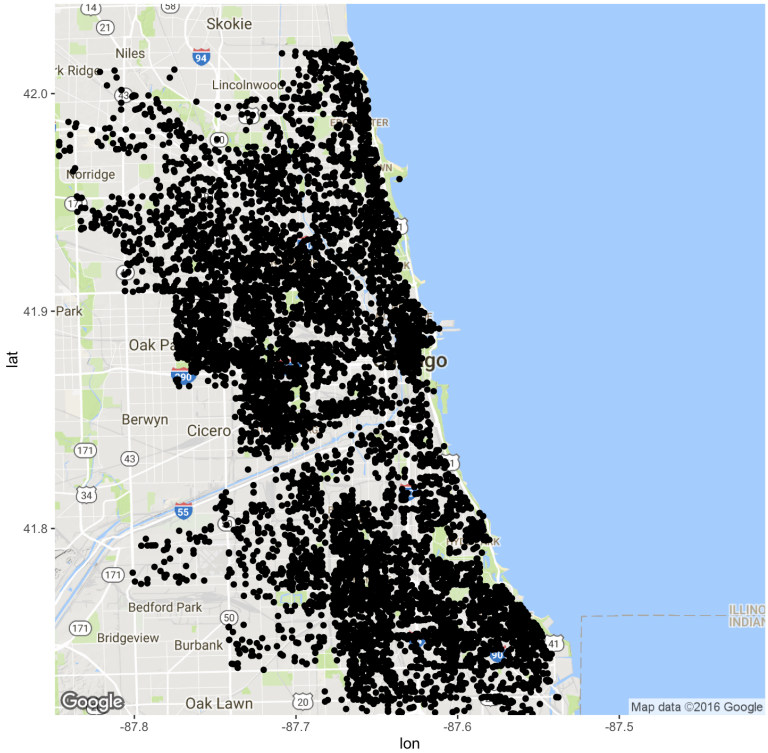


Figure 3

In this graph, I have adjusted the values of alpha in the plot to make it a lot clear.

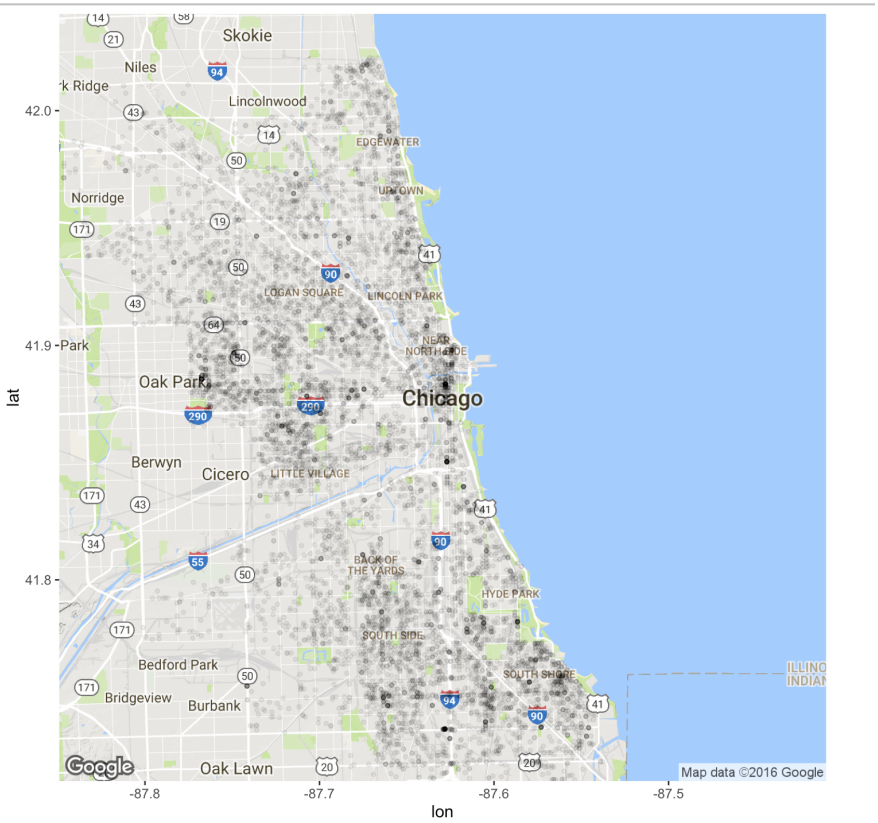


Figure 4

Although now all this is very pleasing to the eye, it still does not give any suitable crime density related conclusions and is therefore not very useful. Therefore I tried to do a very cool thing. I used shapefile of Chicago and plotted it over the actual Chicago map[5] and over that I performed a dot distribution graph as follows:

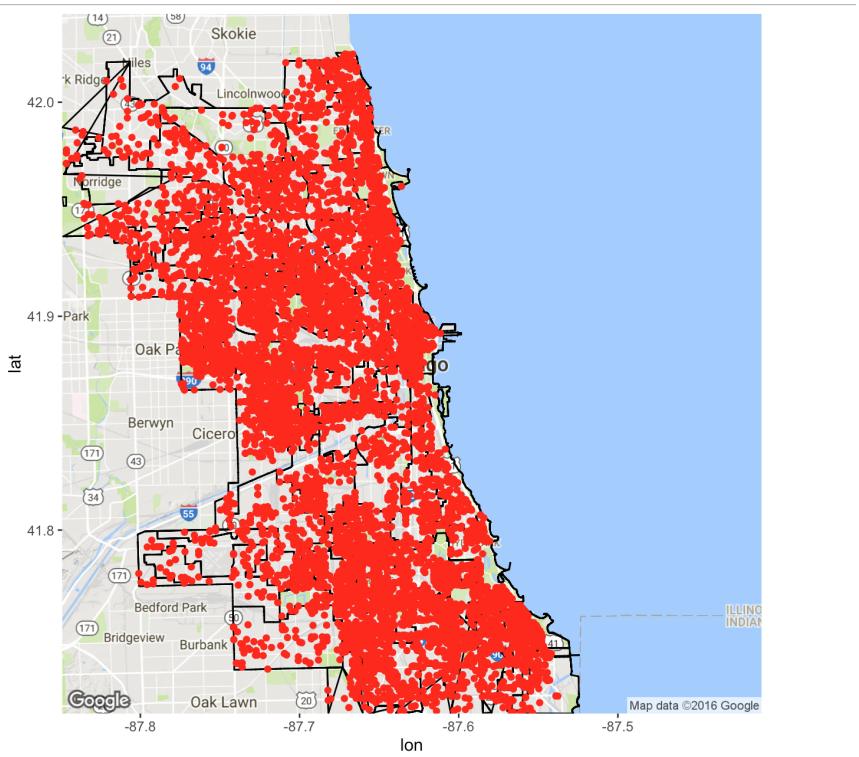


Figure 5

Now this can be used for some analysis. However it still doesnâĂŹt give any crime heat density. Therefore I then decided to move on to the cool stuff i.e the Crime Heat Map. A Crime heat map can effectively show the density of crimes and by the intensity of colors on the map we can find out where the crime heat intensity is highest.

I started off with my heatmap analysis as follows in which I used blue color to show places where there is high crime intensity and red color to show where there is lower crime rates. It was a good start and it gave me an entire picture as to how to start off with the crime heat map analysis.

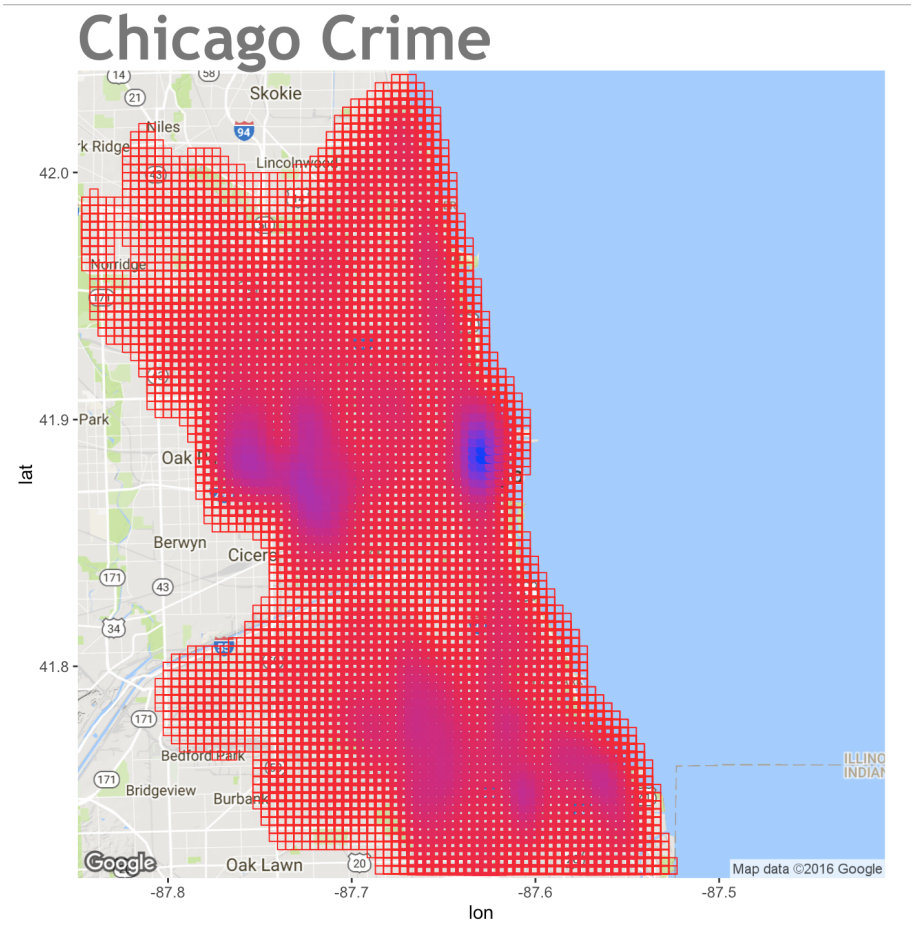


Figure 6

However the problem with this heat-map was it was not at all transparent. It accurately showed heavy crime-infested areas. However it was difficult to see the areas shown on the map. Therefore I thought of a good improvisation by manipulating the alpha values in the plot functions to get a very clear view in the following fashion:

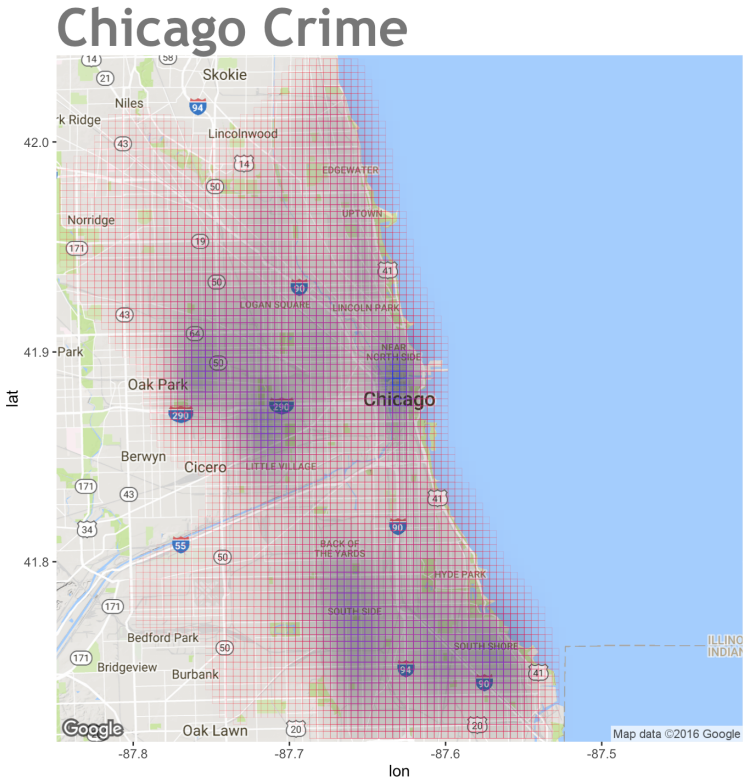


Figure 7

By manipulating the alpha values I got the following graph. However it does not accurately show the levels of crime in the heavily infested regions. That is by looking at the graphs we cannot directly predict as to whether crime rate is higher in Near Northside region or Southside region or South Shore region.

Therefore I thought more I adjusted the alpha values more so that I can clearly see where there is higher crime rate and this graph clearly illustrates that:

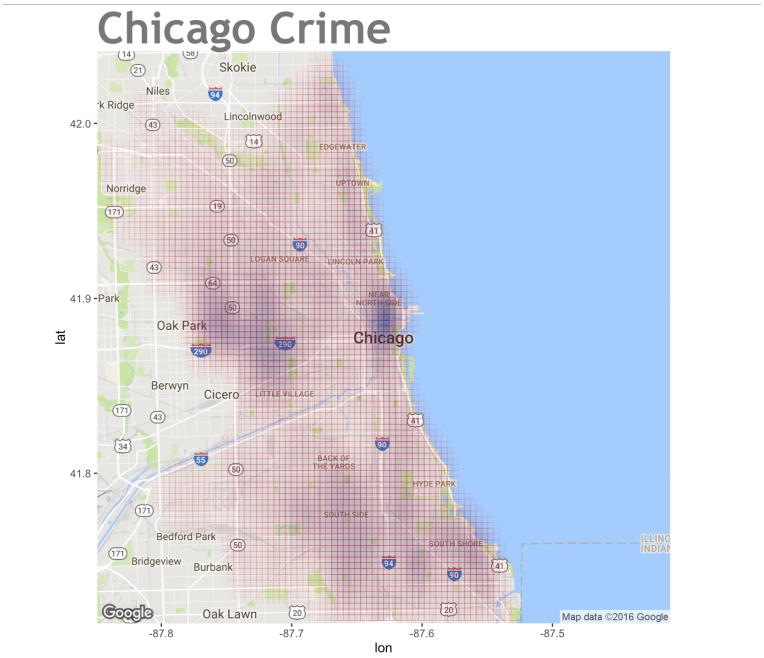


Figure 8

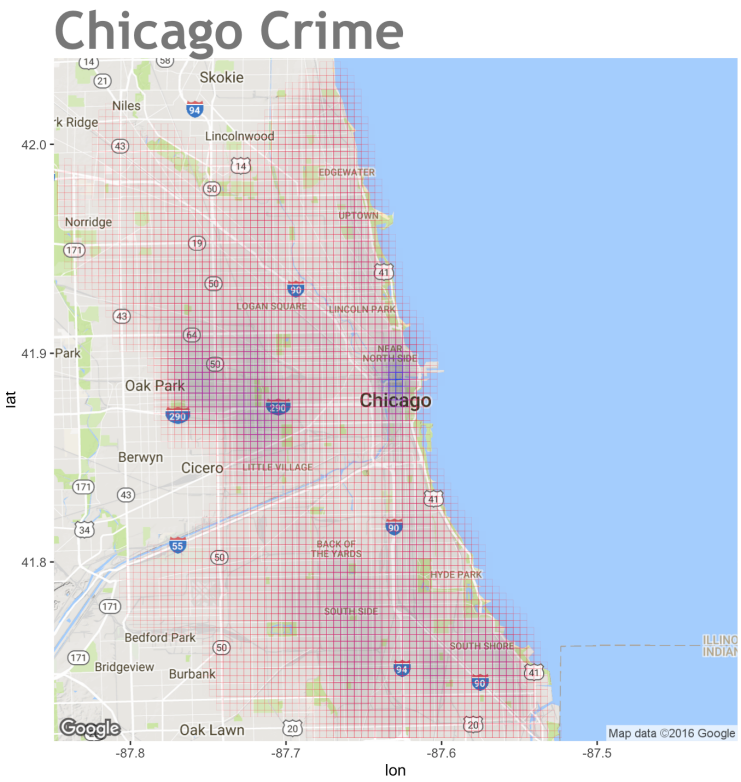


Figure 9

These graphs clearly show that places near Near Northside region have very high crime

heat intensity. Places near Little Village and Oak park have also been susceptible to crime. In the South Side, the crime rate is high but not high enough. Now in this way we can start to make some very good observations from the information visualization.

Once I got the crime heat-map through which I could deduce good observations I thought of a very new idea. In order to get a very accurate analysis I did the following:

After that, I did some work in splitting the whole Chicago into precincts. I figured that by showing the ward number on the heat-map we can make the work of information visualization a lot better. This involved a lot of work for me. First I split the Chicago map into wards. After that I made a new dataset for that. After that, I plotted the heatmap and in conjunction with that using âĂIJgeom_textâĂİ, I even plotted the numbers of wards on the Chicago city. That is one of the new things I thought of because the more human-readable a map becomes more it becomes effective.

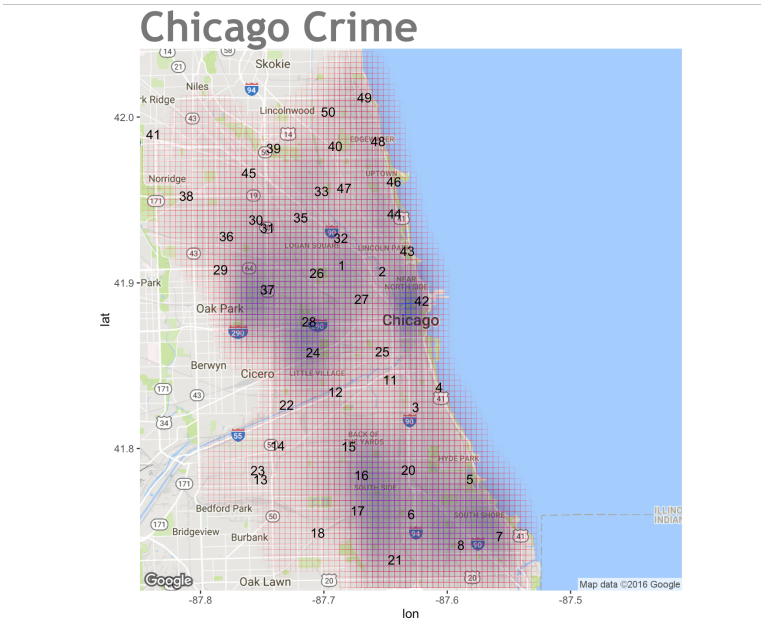


Figure 10

Looking at the graph we can say that wards 37, 28, 24, 42, 6,17,21,7,5 etc. have higher crime density as opposed to the others

So through information visualization I have achieved very easy analysis of data and also I believe these technologies will be useful to law enforcers who want an idea as to which areas are more crime-born and where security should be kept higher.

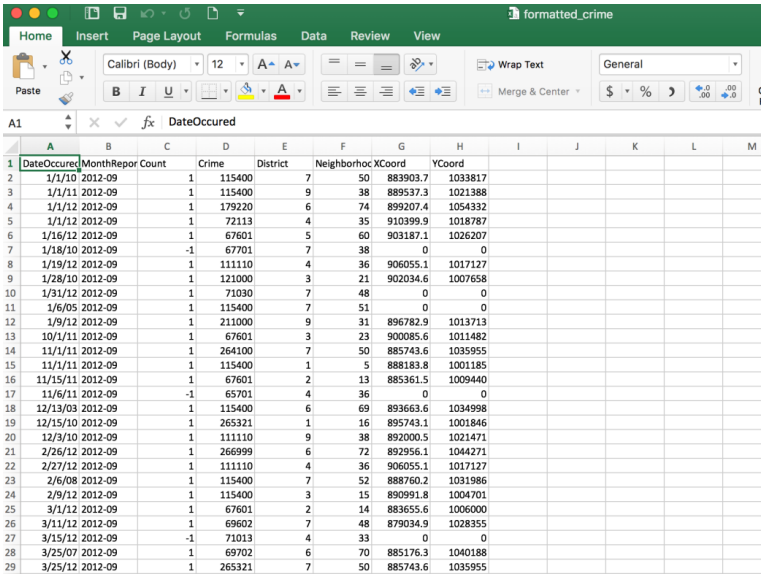
Now taking the analysis to a new level, I felt that maybe weather is a predictor for the crime rates[4]. I used St. Louis crime and weather datasets in my next part of my project to do some cool Machine Learning analysis[3].

2 WEATHER VS CRIME COUNT ANALYSIS FOR ST. LOUIS DATASETS.

After doing information visualization in R using Chicago dataset, I moved on to the St.Louis datasets. I had two main datasets for St. Louis. The crime dataset for St. Louis and the weather datasets for St. Louis. I wanted to do a crime vs. weather analysis using the two datasets. First of all I had to do data cleaning, lots of it. I decided to use Python because I had heard a lot about the âĂIJscikit-learn packageâĂİ as being one of the best for machine learning analysis. In Python, there is a package called pandas. It allows lots of versatility in Python enabling us to use data-frames, allowing us to export to csv files wherever needed and lots of other features making Python work like a âĂIJdata-based languageâĂİ. Therefore first of all I set up Pandas using âĂIJsudo pip install pandasâĂİ and set up the pandas environment. I considered using Pandas over MongoDB because I wanted to export my dataframes to CSV easily and use dataframes in my analysis and Pandas makes it lot easier than PyMongo. After setting up Pandas, I wrote a program named âĂIJdataprep.pyâĂİ in which I created two classes, one of them which processed Crime data differently and fed it to a dataframe and another one which processed wetaher data separately and fed it to a dataframe. I extracted only the columns I felt were good enough for the analysis. This meant a lot of cleaning work, but using operations on panda dataframes I could at last get the two different dataframes as I wanted for the analysis.

After that, it was time to join the two datasets. I thought a lot as to how the two dataframes could be joined. I wanted crime count on a particular date and the maximum temperature and minimum temperature for that date. The dates were in two diffeenet formats. And everything was becoming a problem. Therefore after that I formatted the date columns so that the date column of both the dataframes could match and after that I can go on to merge the two datasets.

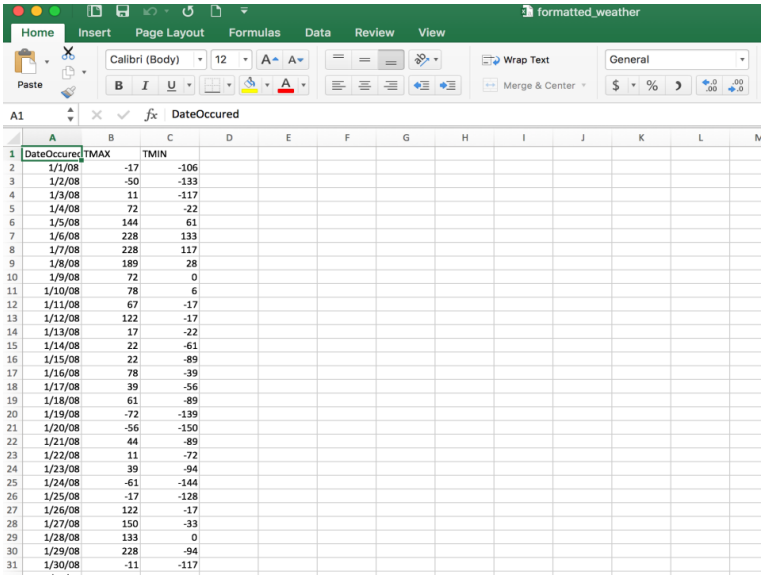
I extracted specific columns and did data cleaning by changing date columns to match by making teir format same as follows: I have exported it to csv for your perusal.



formatted_crime

DateOccured	MonthRepor	Count	Crime	District	Neighborhoc	XCoord	YCoord
1/1/10	2012-09	1	115400	7	50	883903.7	1033817
1/1/11	2012-09	1	115400	9	38	889537.3	1021388
1/1/12	2012-09	1	179220	6	74	899207.4	1054332
1/1/12	2012-09	1	72113	4	35	910399.9	1018787
1/16/12	2012-09	1	67601	5	60	903187.1	1026207
1/18/10	2012-09	-1	67701	7	38	0	0
1/19/12	2012-09	1	111110	4	36	906055.1	1017127
1/28/10	2012-09	1	121000	3	21	902034.6	1007658
1/31/12	2012-09	1	71030	7	48	0	0
1/6/05	2012-09	1	115400	7	51	0	0
1/9/12	2012-09	1	211000	9	31	896782.9	1013713
10/1/11	2012-09	1	67601	3	23	900085.6	1011482
11/1/11	2012-09	1	264100	7	50	885743.6	1035955
11/1/11	2012-09	1	115400	1	5	888183.8	1001185
11/15/11	2012-09	1	67601	2	13	885361.5	1009440
11/6/11	2012-09	-1	65701	4	36	0	0
12/13/03	2012-09	1	115400	6	69	893663.6	1034998
12/15/10	2012-09	1	265321	1	16	895743.1	1001846
12/3/10	2012-09	1	111110	9	38	892000.5	1021471
2/26/12	2012-09	1	266999	6	72	892956.1	1044271
2/27/12	2012-09	1	111110	4	36	906055.1	1017127
2/6/08	2012-09	1	115400	7	52	888760.2	1031986
2/9/12	2012-09	1	115400	3	15	890991.8	1004701
3/1/12	2012-09	1	67601	2	14	883655.6	1006000
3/11/12	2012-09	1	69602	7	48	879034.9	1028355
3/15/12	2012-09	-1	71013	4	33	0	0
3/25/07	2012-09	1	69702	6	70	885176.3	1040188
3/25/12	2012-09	1	265321	7	50	885743.6	1035955

Figure 11

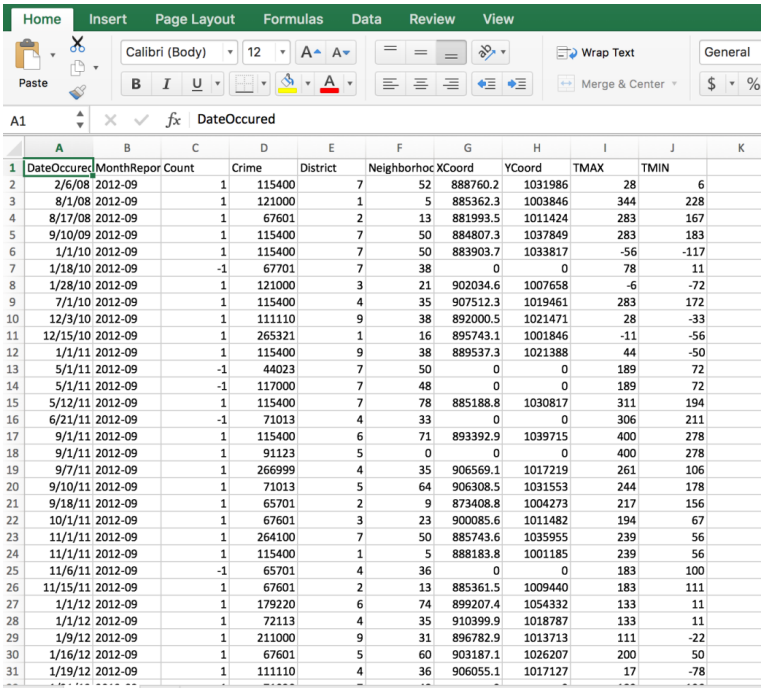


formatted_weather

DateOccured	TMAX	TMIN
1/1/08	-17	-106
1/2/08	-50	-133
1/3/08	11	-117
1/4/08	72	-22
1/5/08	144	61
1/6/08	228	133
1/7/08	228	117
1/8/08	189	28
1/9/08	72	0
1/10/08	78	6
1/11/08	67	-17
1/12/08	122	-17
1/13/08	17	-22
1/14/08	22	-61
1/15/08	22	-89
1/16/08	78	-39
1/17/08	39	-56
1/18/08	61	-89
1/19/08	-72	-139
1/20/08	-56	-150
1/21/08	44	-89
1/22/08	11	-72
1/23/08	39	-94
1/24/08	-61	-144
1/25/08	-17	-128
1/26/08	122	-17
1/27/08	150	-33
1/28/08	133	0
1/29/08	228	-94
1/30/08	-11	-117

Figure 12

After formatting the date column in weather dataframe to be in same format as crime dataframe, I renamed the Date in weather dataframe as DateOccured to match the crime dataframe Date column. After that I merged the two dataframes based on Date. After that I had crime count w.r.t to date and the corresponding maximum and minimum temperature values for that date too. I merged the two dataframes as df_merged in the âĂIJloaddata.pyâĂİ file and I exported dataframe as csv so that I can find out whether the merge is perfect or not am posting the screenshot for the merged csv as follows:



	A	B	C	D	E	F	G	H	I	J	K
1	DateOccured	MonthRepor	Count	Crime	District	Neighborhoc	XCoord	YCoord	TMAX	TMIN	
2	2/6/08	2012-09	1	115400	7	52	888760.2	1031986	28	6	
3	8/1/08	2012-09	1	121000	1	5	885362.3	1003846	344	228	
4	8/17/08	2012-09	1	67601	2	13	881993.5	1011424	283	167	
5	9/10/09	2012-09	1	115400	7	50	884807.3	1037849	283	183	
6	1/1/10	2012-09	1	115400	7	50	883903.7	1033817	-56	-117	
7	1/18/10	2012-09	-1	67701	7	38	0	0	78	11	
8	1/28/10	2012-09	1	121000	3	21	902034.6	1007658	-6	-72	
9	7/1/10	2012-09	1	115400	4	35	907512.3	1019461	283	172	
10	12/3/10	2012-09	1	111110	9	38	892000.5	1021471	28	-33	
11	12/15/10	2012-09	1	265321	1	16	895743.1	1001846	-11	-56	
12	1/1/11	2012-09	1	115400	9	38	889537.3	1021388	44	-50	
13	5/1/11	2012-09	-1	44023	7	50	0	0	189	72	
14	5/1/11	2012-09	-1	117000	7	48	0	0	189	72	
15	5/12/11	2012-09	1	115400	7	78	885188.8	1030817	311	194	
16	6/21/11	2012-09	-1	71013	4	33	0	0	306	211	
17	9/1/11	2012-09	1	115400	6	71	893392.9	1039715	400	278	
18	9/1/11	2012-09	1	91123	5	0	0	0	400	278	
19	9/7/11	2012-09	1	266999	4	35	906569.1	1017219	261	106	
20	9/10/11	2012-09	1	71013	5	64	906308.5	1031553	244	178	
21	9/18/11	2012-09	1	65701	2	9	873408.8	1004273	217	156	
22	10/1/11	2012-09	1	67601	3	23	900085.6	1011482	194	67	
23	11/1/11	2012-09	1	264100	7	50	885743.6	1035955	239	56	
24	11/1/11	2012-09	1	115400	1	5	888183.8	1001185	239	56	
25	11/6/11	2012-09	-1	65701	4	36	0	0	183	100	
26	11/15/11	2012-09	1	67601	2	13	885361.5	1009440	183	111	
27	1/1/12	2012-09	1	179220	6	74	899207.4	1054332	133	11	
28	1/1/12	2012-09	1	72113	4	35	910399.9	1018787	133	11	
29	1/9/12	2012-09	1	211000	9	31	896782.9	1013713	111	-22	
30	1/16/12	2012-09	1	67601	5	60	903187.1	1026207	200	50	
31	1/19/12	2012-09	1	111110	4	36	906055.1	1017127	17	-78	

Figure 13

2) Machine Learning using Neural Network: Now after I got the dataframe in this format, I was ready to go for my Machine Learning analysis[6]. After thorough research work I figured I could use a neural network[2] in order to predict the Crime rate from the maximum temperature and minimum temperature values. I started off with Scikit Learn Neural network model. I decided to use a LBFGS solver(Limited Memory Broyden Fletcher Goldfarb Shanno) in my neural network because it was a problem of parameter estimation I.e Crime count estimation. I used TMAX(Maximum temperature) and TMIN(Minimum Temperature) as the predictors for crime count. I used MLPClassifier using Theano based on Multi Layer Perceptron feed-forward artificial neural network model. I used MLP classifier because I wanted to do logistic regression classification with my data and wanted to predict crime count as a function of maximum and minimum temperatures. I wanted a one-dimensional array for crime count as the output. Therefore, I had to use the fit function in order to train my neural network so that it is âĂIJintelligentâĂİ and has idea about previous input and output combinations. A MLPClassifier has hidden layers, the higher the number of hidden layers higher the time it takes to compute but it gives more accurate results. I used hidden layer size of 15 ax it will be enough as I was getting accurate results. After all these analysis, I finally trained my neural network and allowed it to predict crime count as function of TMAX and TMIN as follows. The 2d array is TMAX and TMIN values and 1d array is Crime Count I have used to train the neural network and below that is the answer for [200., 150.]:

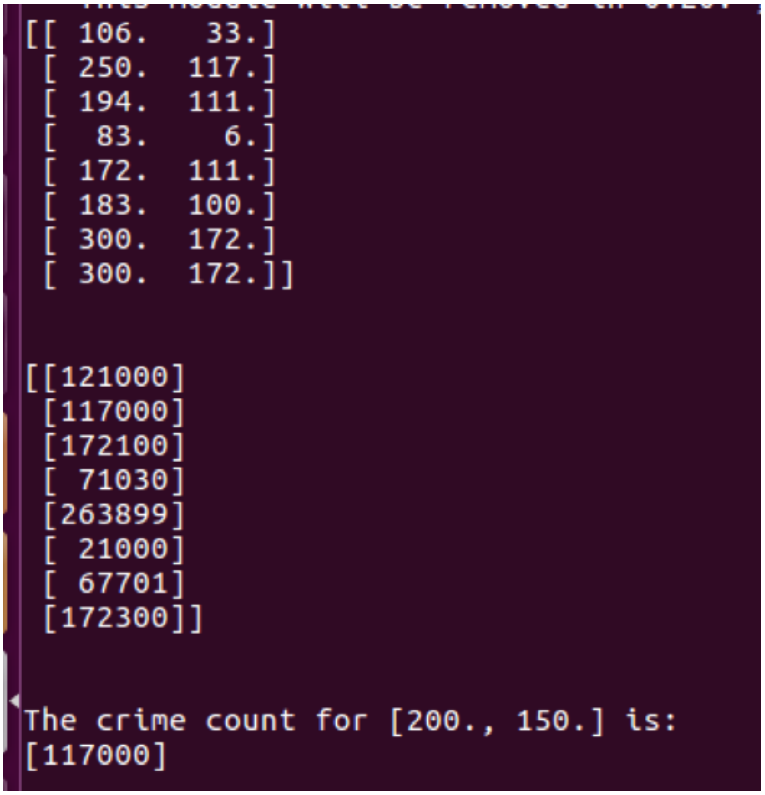


Figure 14

After that we can expand the neural network to cover the entire data and it will accurately predict the Crime count values as a function of Crime. Therefore in this way I created a neural network using MLPClassifier and lgbfs solver using hidden layer size of 15.

3) Machine Learning using Regression Analysis: After building the neural network I just explored the field of Machine learning[1] a little bit more as to finding out how does crime count vary with TMAX and TMIN respectively and performed a visualization of that using Seaborn library in python.TMIN and TMAX at some places seem to have ambiguous and very high values however on the wole they can assumed to be true I figured this can be useful if we want to look into these factors individually. Using principles of regression I found Crime count vs TMAX and Crime count vs TMIN as follows:

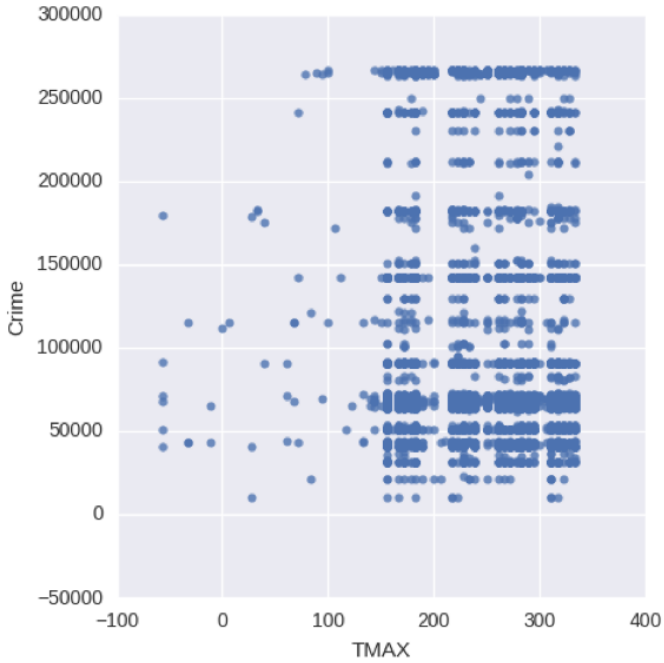


Figure 15

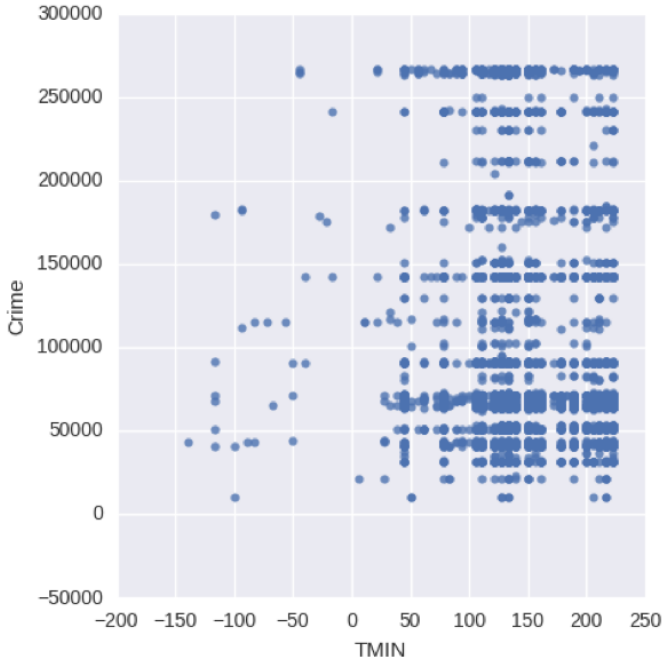


Figure 16

From these graphs we can find that for TMIN range from 100 to 200 and TMAX range of 150 to 300 the crime rates reported are highest and for the lower temperatures the crime

count is lower. This is clear indication that in higher temperatures there is higher occurrence of crime.

In this way, I have analyzed the St. Louis datasets and found out inestimable deductions from it.

3 CONCLUSION

In this way I created a heatmap for Chicago and segregated it on basis of wards for easy data analysis and I did a Machine learning analysis using St.Louis datasets wherein I created a neural network for prediction and also did a regression analysis to find out how temperature affects the crime rate.

REFERENCES

[1] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. An overview of machine learning. In *Machine learning*, pages 3–23. Springer, 1983.

[2] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin, and Michael Chau. Crime data mining: a general framework and some examples. *Computer*, 37(4):50–56, 2004.

[3] John L Cotton. Ambient temperature and violent crime1. *Journal of Applied Social Psychology*, 16(9):786–801, 1986.

[4] Simon Field. The effect of temperature on crime. *British Journal of criminology*, 32(3):340–351, 1992.

[5] Daniel A Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[6] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.