Chicago Crime Pattern Investigation

By Shicheng Huang

The main question of my investigation is if number of crime reports in Chicago are seasonal(repeating the same pattern again and again).

Source:

Chicago: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data
only the crime reports from 2009 to 2015.
Philadelphia :https://www.opendataphilly.org/dataset/crime-incidents
Full dataset
Rockford :
https://data.illinois.gov/dataset/City-of-Rockford-Crime-Offenses-2011-Present/x8xk-7uvk
Full dataset

Sanity Check:

Examine crime date columns from each dataset(and crime type column from the chicago dataset).
The lengths of the dates are the same, mostly likely they are correctly formatted(further proof by not running any errors while I reformatting the dates.
I do find 4 missing values from the Philadelphia dataset but I filter those rows very quickly.

```
: #checking pdd dataframes' date column, confirmed with no length violations
print(chicago["Date"].apply(len).unique())
#nothing out of orindary so far
print(chicago["Primary Type"].unique())
print(rockford["Occurred On Date"].apply(lambda x: x[0:4] + "/" + x[5:7]).unique())
print(pdd["DISPATCH_DATE_TIME"].apply(lambda x: len(x)).unique())
```

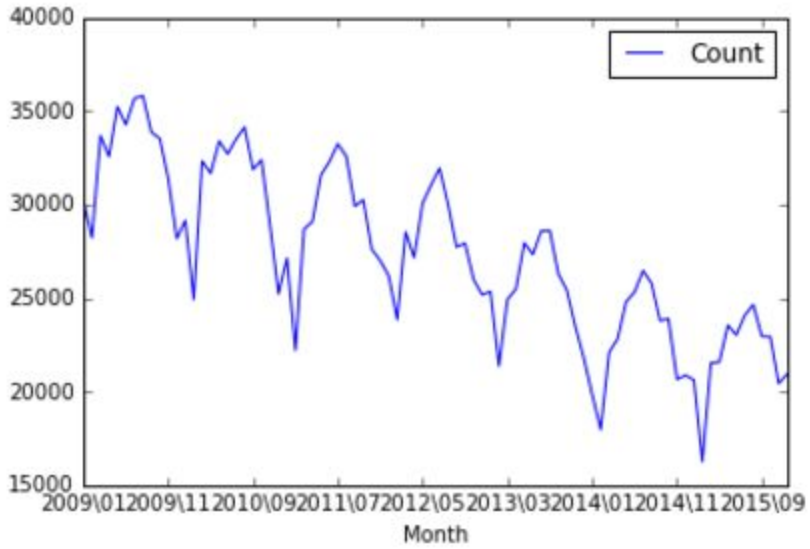
```
[22]
['OTHER OFFENSE' 'DECEPTIVE PRACTICE' 'CRIM SEXUAL ASSAULT'
 'WEAPONS VIOLATION' 'PROSTITUTION' 'BATTERY' 'CRIMINAL DAMAGE'
 'MOTOR VEHICLE THEFT' 'THEFT' 'CRIMINAL TRESPASS' 'ROBBERY' 'NARCOTICS'
 'ASSAULT' 'ARSON' 'PUBLIC PEACE VIOLATION' 'BURGLARY' 'SEX OFFENSE'
 'HOMICIDE' 'OFFENSE INVOLVING CHILDREN' 'INTERFERENCE WITH PUBLIC OFFICER'
 'LIQUOR LAW VIOLATION' 'STALKING' 'KIDNAPPING' 'INTIMIDATION' 'OBSCENITY'
 'NON-CRIMINAL' 'GAMBLING' 'HUMAN TRAFFICKING'
 'CONCEALED CARRY LICENSE VIOLATION' 'PUBLIC INDECENCY' 'NON - CRIMINAL'
 'OTHER NARCOTIC VIOLATION' 'NON-CRIMINAL (SUBJECT SPECIFIED)']
['2016/01' '2016/02' '2016/03' '2016/04' '2013/02' '2012/12' '2012/01'
 '2013/10' '2011/10' '2014/10' '2014/02' '2013/08' '2014/04' '2011/11'
 '2015/05' '2014/03' '2012/04' '2012/05' '2013/03' '2012/07' '2012/06'
 '2011/02' '2015/03' '2015/11' '2013/09' '2014/05' '2013/04' '2014/08'
 '2012/11' '2013/06' '2014/06' '2014/01' '2011/03' '2015/01' '2012/09'
 '2014/07' '2014/12' '2015/10' '2015/09' '2013/07' '2014/11' '2012/10'
 '2011/06' '2012/03' '2013/11' '2015/04' '2011/07' '2012/08' '2015/12'
 '2011/05' '2013/05' '2013/12' '2011/08' '2015/07' '2011/12' '2015/06'
 '2011/01' '2015/02' '2014/09' '2015/08' '2011/04' '2011/09' '2013/01'
 '2012/02' '2016/05' '2016/06' '2016/07' '2016/08']
[22]
```

Terminologies(could refer to this section if I have any confusing terms):

1. Crime : Crime in this report merely represents crime reports or the number of crime reports in the dataset because I do think crimes and crime reports are different. If you believe otherwise you could simply ignore this note.
2. Normalized: for each point of the dataset, subtract by the mean of the dataset, then divide the difference by the standard deviation of the dataset.
3. Crime Rank Distribution: the distribution between the ranks of the crime types(based on frequency), and its corresponding frequency. Given a type of crime is ranked x in terms of its proportion to the total number of crime reports, can we predict the exact proportion?
4. Bolded Sentences: important conclusions/opinions I drawn from my tests and observations, could be wrong and incomplete.

Analysis:

General:

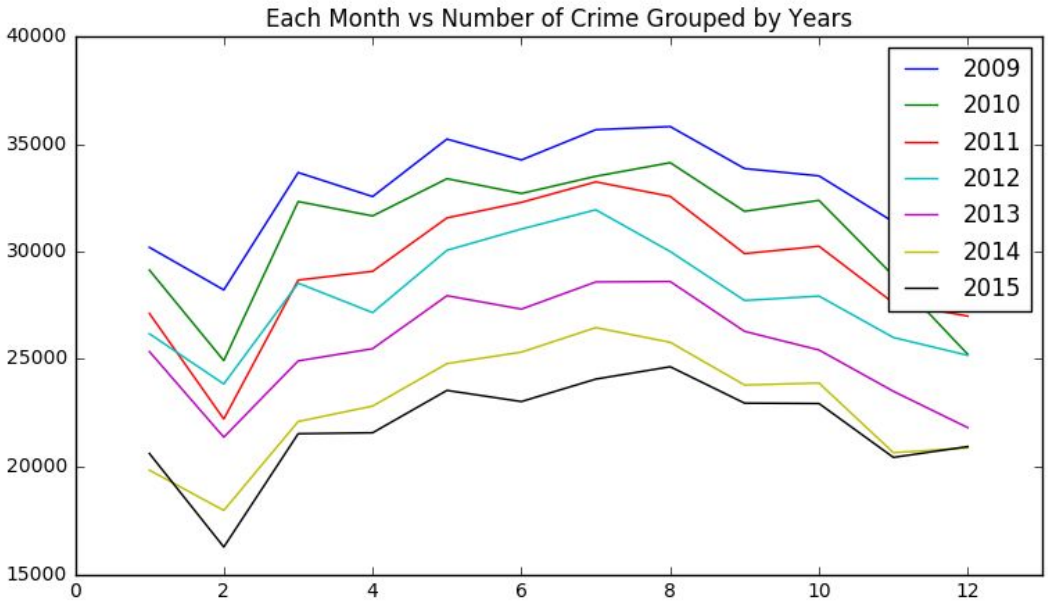


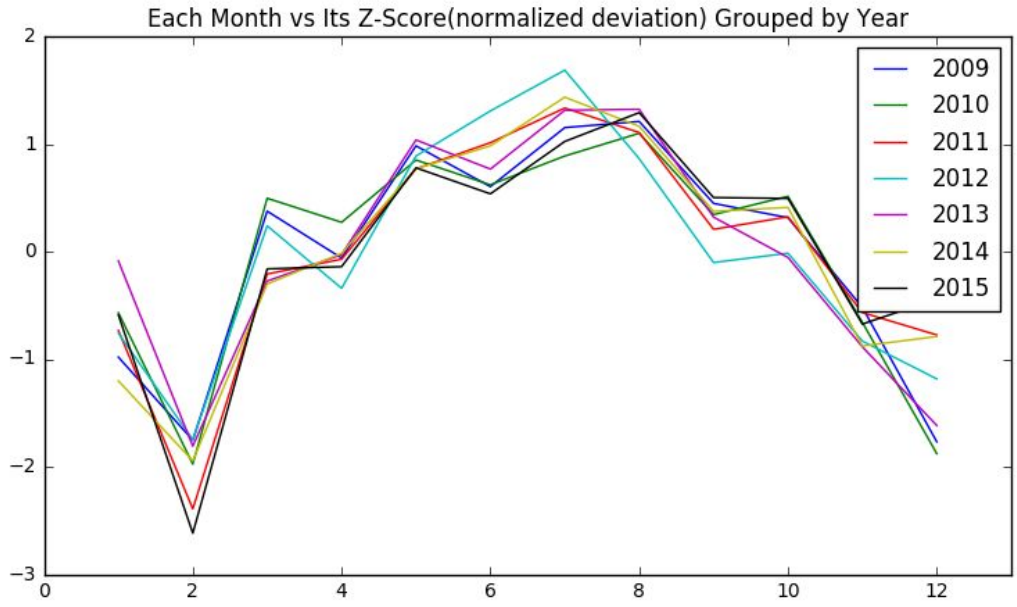
From above graph("Crime vs Time"), we can clearly see the repeating patterns of crimes:

1.The overall crime reports each year is constantly decreasing.

2.Crimes seem to follow a normal distribution(rises, peaks, then falls)

Crimes in Months grouped by Years





Not much pattern could be found by purely graphing the crime numbers in each month(grouped by years) but crimes' seasonality emerges when I scale them based on proportions(to their respective year) and normalize them. All the lines seem to converge to one line.

To have a more scientific analysis whether the distributions of the normalized monthly crime reports in different years, I conducted a chi-square test of homogeneity.

Null Hypothesis:
The distribution of the square of normalized crime data are the same in many years.(in other words, years has no effect on the distribution)

```
#squaring all the z-scores to meet the chi-square test requirement
stats.chi2_contingency(np.apply_along_axis(lambda x : abs(x), 1, normal_count))[:-1]
```

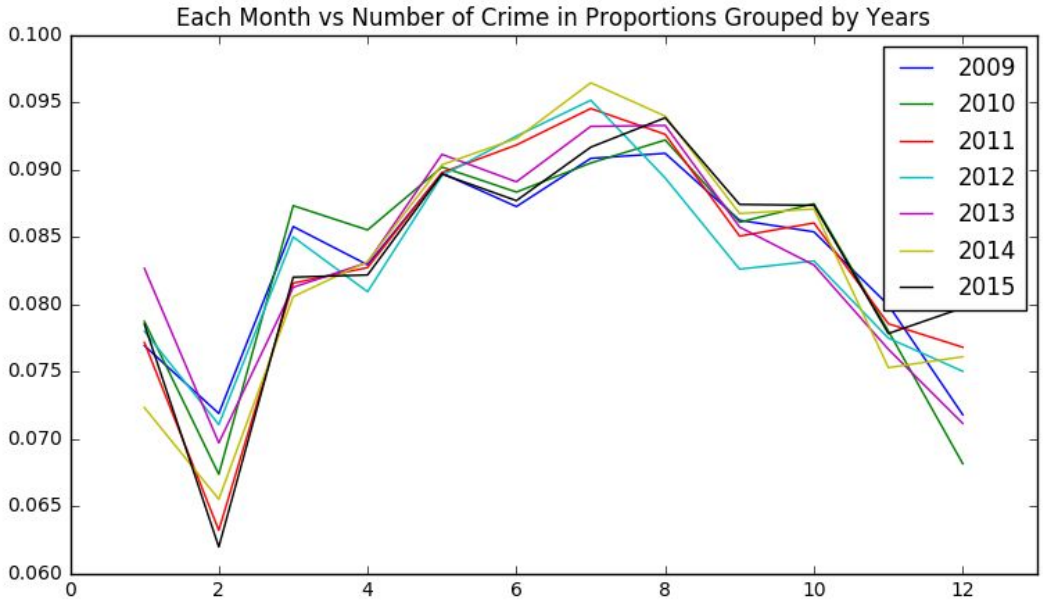
```
(6.2623366066442143, 1.0, 66L)
```

The p-value is actually so high that we cannot reject that the null hypothesis that all the Z^2 distribution of the months on each year, is from one true Z^2 distribution. To validate my results, I checked my codes few times and performed the same test using Rstudio, got the same p-value and test-statistic.

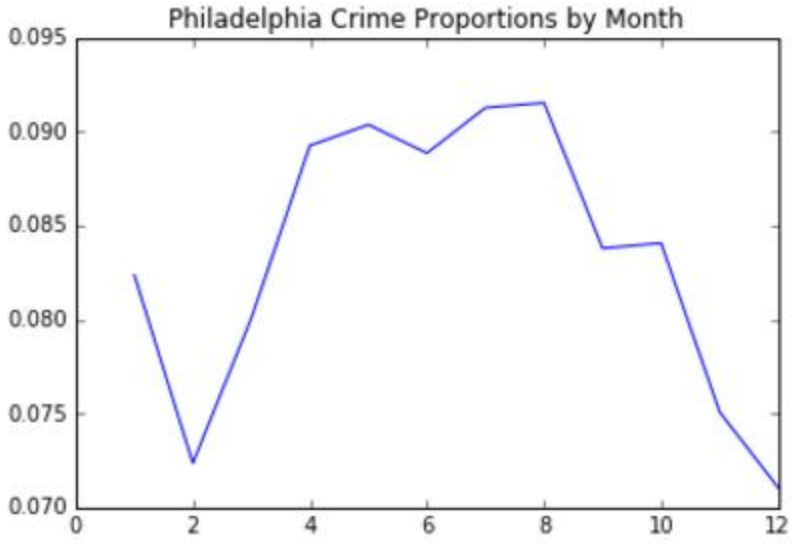
Now I can conclude that the normalized crime report distribution in different months are the same over the years. Assuming the above, I find such distribution seems to follow a normal distribution centering around July or August. By this standard, it is reasonable to think that crimes are highly related to weather/temperature since it follows a similar distribution.

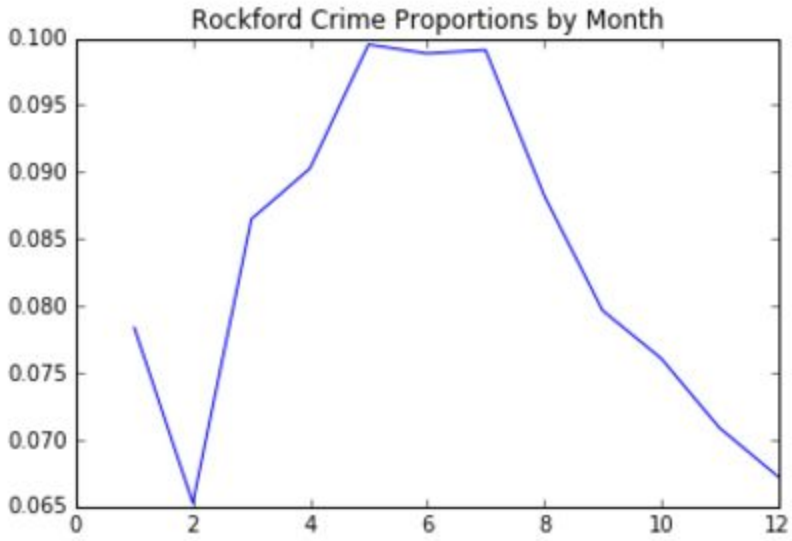
However, the distribution has a drastic fall in crime reports(both in proportions and normalized data during February). To explain this abnormality, my best is because February has 2 or 3 days shorter than other months, if we assume constant temperature, each day is expected to have same number of crimes, then the shorter day theory could explain around 8-10% below where it was supposed to be in the normal curve. Based on our normality assumption, and assuming the expected number of crimes are the same with fixed temperature, the February is expected to be around 0.08% if February has 30 days. Cutting the extra 2 days of crimes to February, then we expect a 0.08% * 28/30 = 0.0747%, which is still at the top of all the empirical data.

Therefore we conclude that the 2 less days reason is solid but not enough to fully explain the low crimes in February.



To see if this February-drastic-fall is only unique to Chicago, I also look at the distributions from Philadelphia and Rockford



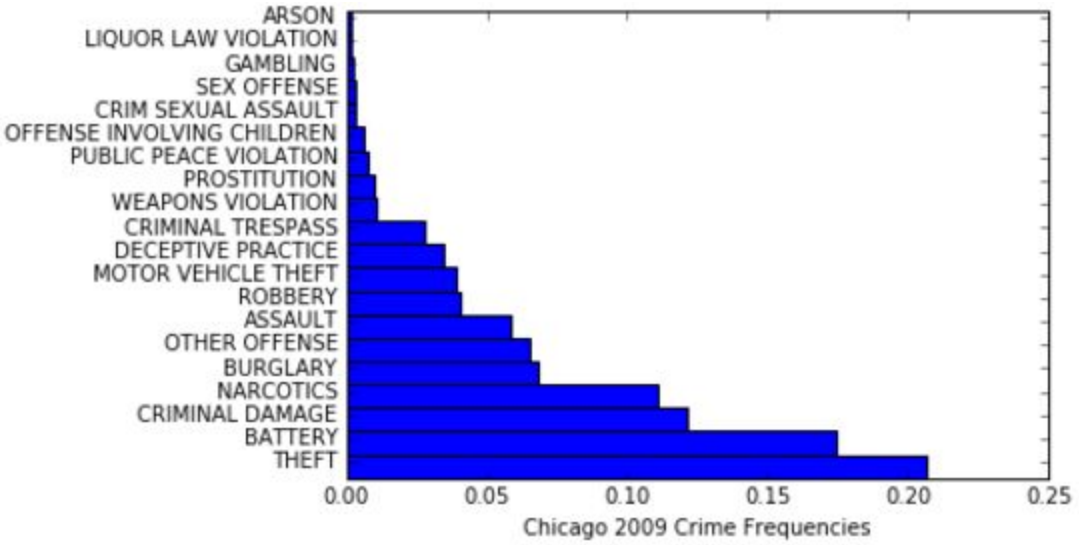


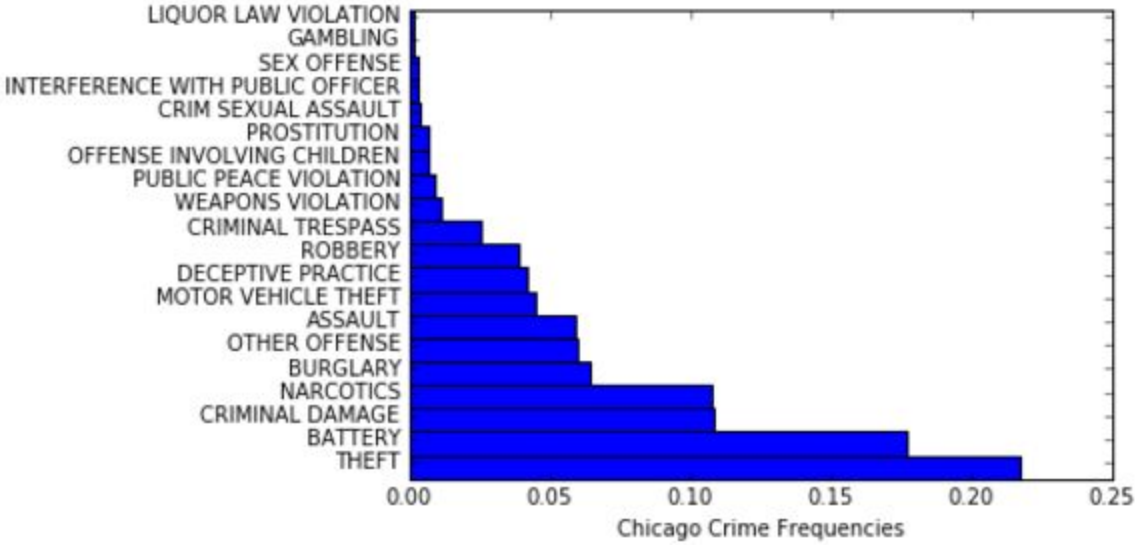
From above two graphs, both Philadelphia and Rockford has the February-drastic-fall phenomenon and the seemingly normally distributed from February to December. **Although I fail to explain this phenomenon fully I do find that it is not unique and most likely universal to many cities.**

Crimes Grouped by Types: Crime Rank Distribution

One of the question I have while analyzing the seasonality of crimes is:
Is the seasonality of crimes only affected by a few influential types of crimes, or most of the crimes are indeed seasonal.

First I graphed the probability distribution of different crimes, both the 2009 distribution and the overall distribution.





From above graph, we can see:

1.Two probability distributions are so similar(almost the same)!

2.The probability distributions seem to follow an exponential distribution, at least extremely right skewed when ranks are in the x-axis.

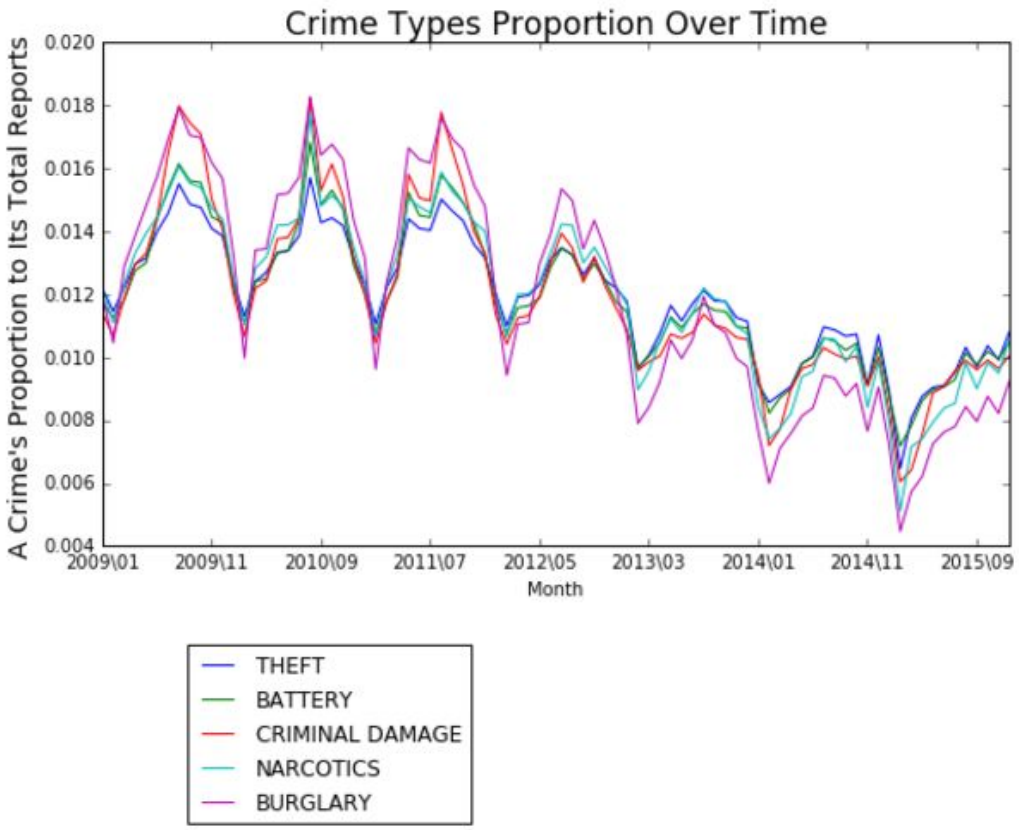
To test the probability distributions are the *same(coming from a true distribution), I make another chi-square test of homogeneity, resulting an almost 1 p-value(not proving the previous hypothesis but denies it was due to something else.)

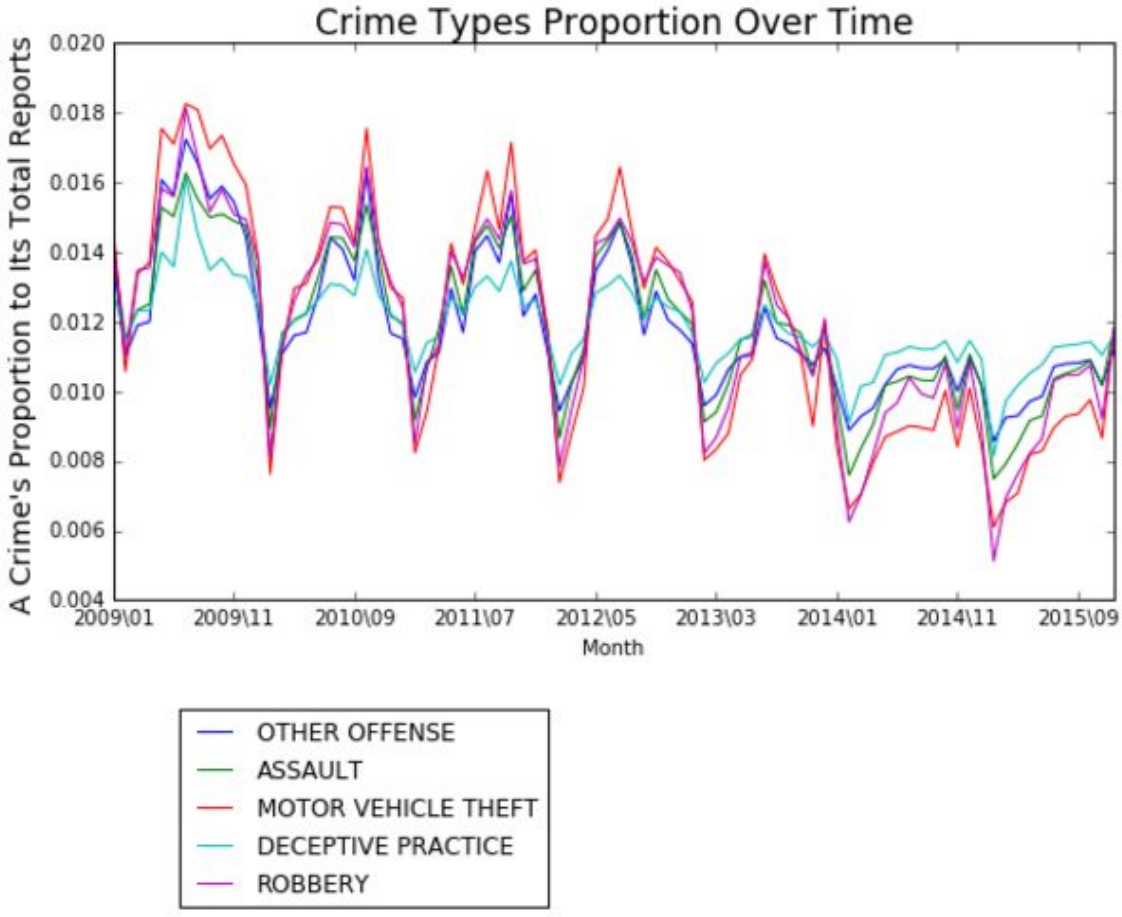
```
stats.chi2_contingency(np.array(crime_dis))[:-1]
```

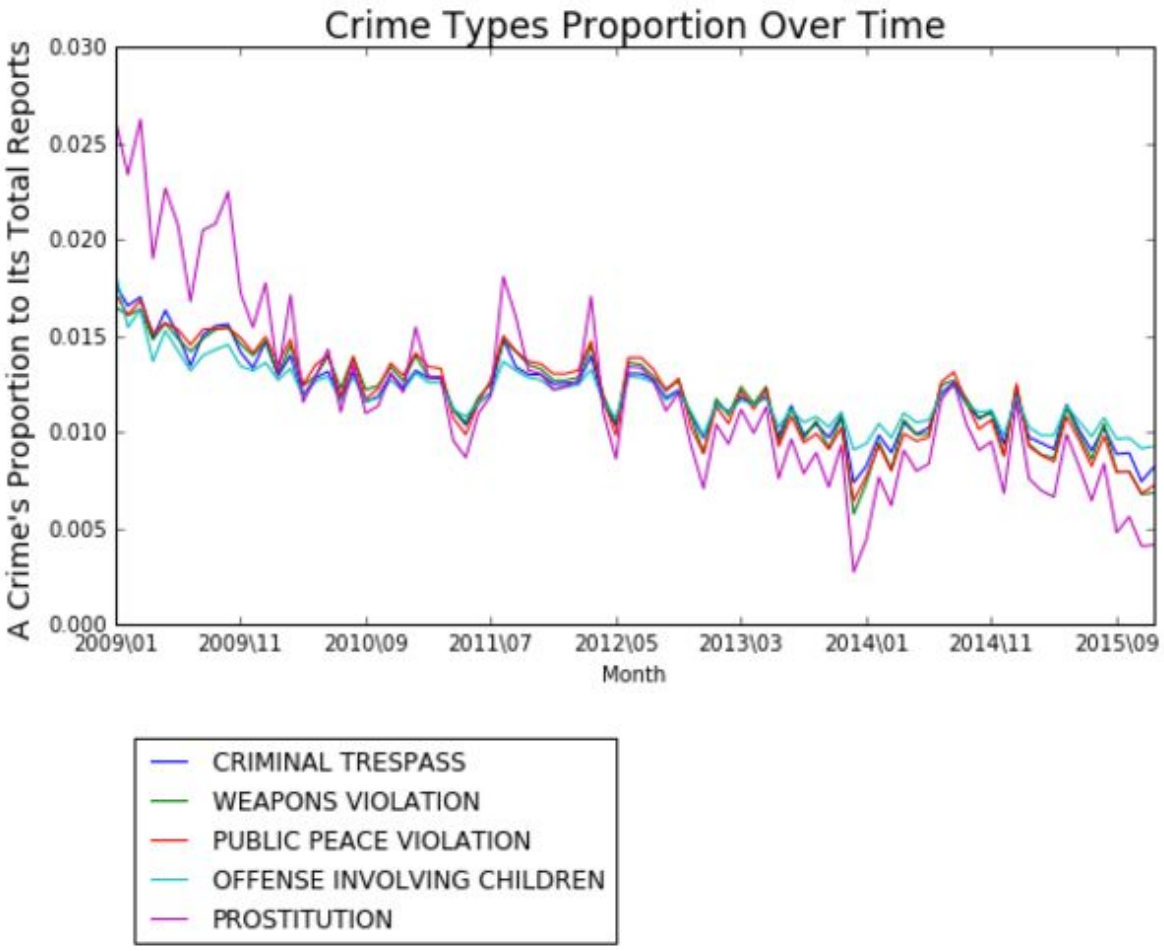
```
(0.020957865200910992, 1.0, 120L)
```

If the crime type frequency distribution is right skewed, top crimes types' crime counts could influence the shape of the overall crime vs time graph a lot.

Below are the the Crime vs Time grouped by Crime Types(from top 1 - 5, top 6 - 10, and top 11 - 15)




Crime Types Proportion Over Time
A Crime's Proportion to Its Total Reports
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
2009\01
2009\11
2010\09
2011\07
2012\05
2013\03
2014\01
2014\11
2015\09
Month
OTHER OFFENSE
ASSAULT
MOTOR VEHICLE THEFT
DECEPTIVE PRACTICE
ROBBERY



From previous three graphs, we can see the top 10 crimes types' which are composed of more than 80% of the total crimes are highly seasonal, but the top 10 - 15 seems a bit chaotic. I conclude that not all the crimes are seasonal, though they don't contribute much to the overall crime vs time pattern(in chicago at least).

Reflections:

Chi-square test of homogeneity: the p-values are too high to be true, although I have checked my codes a few times. It may because the way I manipulate the data makes the p value inflated or the chi-square test doesn't seem to be a good test. But the graphs do convinces me my

hypotheses are very likely to be true even without the chi-square tests. I have thought of the the legitimacy of the chi-square test a lot but still inconclusive. This also makes me realized how little I know about statistics and how helpless I could be when I need guidence.