**Predicting Crime Rates Using Weather data**

Team 18

By:   
Inamdar, Seema  
Savva, Andrew

December 16, 2016

The University of texas at arlington

Table of Contents

[1: Executive Summary 2](#_Toc469330055)

[2: Research Question and Project Implication 3](#_Toc469330056)

[3: Data Sets Used 3](#_Toc469330057)

[4: Data Pre-Processing 4](#_Toc469330058)

[5: Model and Algorithm Description 6](#_Toc469330059)

[6: Model Results and Evaluation 7](#_Toc469330060)

[Austin Master Data Set 7](#_Toc469330061)

[In-Sample Poisson Data Model 7](#_Toc469330062)

[Binomial Data Model 9](#_Toc469330063)

[Los Angeles Master Data Set 10](#_Toc469330064)

[In-Sample Poisson Data Model 10](#_Toc469330065)

[Binomial Data Model 12](#_Toc469330066)

[7: Conclusions and Lessons Learned 14](#_Toc469330067)

[Data 14](#_Toc469330068)

[Model 15](#_Toc469330069)

[Appendix 16](#_Toc469330070)

[4.1 Chicago Crime Sample Data Set 16](#_Toc469330071)

[4.2 Austin Crime Data Sample 16](#_Toc469330072)

[4.3 Los Angeles Crime Data Sample 17](#_Toc469330073)

[4.4 Weather Data Sample 17](#_Toc469330074)

[5.1 Austin Data Set JSON Exert 18](#_Toc469330075)

[5.2 Crime Master Data Set JSON Exert 18](#_Toc469330076)

# 1: Executive Summary

# 2: Research Question and Project Implication

In this project, we seek to analyze whether the daily crime rates in various metropolitan areas have a statistical relationship with weather conditions. We gathered crime rates data and daily weather data for metropolitan areas of Chicago, Austin and Los Angeles. As we are doing analysis on daily crime rates, we have assumed other factors such as average age, average household income, and gender distribution remains constant throughout this period. These factors are likely to have an impact on crime rates, but the goal of our study is to isolate and study the effect of weather factors.

We modelled daily crime rates, so this is a reasonable assumption to make that all the population factors will remain constant over the year. For this project, we have performed analysis on data between January 2015 and January 2016.

Goals of the project are:

1. Come up with a model relationship between weather conditions and daily crime rates
2. Perform in-sample training and see how good the fit is in out-of-sample testing
3. Perform statistical tests to see if we have got a good model

We aim to come up with a statistical relationship between weather conditions and crime rates. If we find a meaningful pattern between crime rates and weather data, the model predictions could be useful for city police department to better allocate limited resources at disposal to tackle crimes in the city. Furthermore, within a city we could have different zones and each zone could have its own model to predict crime rates against weather to better inform city police department about likelihood of crimes.

# 3: Data Sets Used

To accurately predict crimes rates by weather conditions across geographic regions, a robust data set was needed that reported on daily, accurate crimes and weather conditions. Our search for accurate information crossed many different online sources. Little information was available reporting on daily crime rates within the year of 2015 across geographic boundaries (in more than one city). Naturally, it was determined early on that multiple crime data sets and sources would need to be combined to generate enough information to support our hypothesis. An ideal source for crime information available online is the data.gov website. Using its keyword search we can identify three strong data sets that reported on daily crime rates within the calendar year of 2015. These datasets included the City of Chicago Crime Dataset, the Austin City Crime Dataset, and the Los Angeles City Crime Dataset. These cities were ideally suited for our analysis purposes because all three cities are geographically dispersed across the continental United States.

The City of Chicago Crime Dataset was supplied from data.gov and was published by the data.cityofchicago.org website. The original data was gathered by the City of Chicago’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system which is used by the Chicago police departments for tracking incidents of crime. The information was uploaded to data.gov on February 4th of 2015 and was downloaded as Crime\_Data\_Chicago.csv to a local staging environment by our team on December 1st of 2016. The data source reported crime incidents from 11/27/2015 to 11/25/2016. The raw data source contains a total of 264,335 records and 17 attributes. These attributes included a case number, the date of the occurrence, a primary description of the incident, geographic coordinates about the incident’s location and more (see appendix 4.1). Values for each of these attributes included the following data types: String, Boolean, Float, DateTime, and Integer.

The City of Austin Crime Dataset was supplied from data.gov and was published by the data.austintexas.gov website. The original data was gathered by the Austin Police Department, and the original system(s) from which the dataset was acquired is unknown. The information was uploaded to the data.gov website on July 12, 2016 and was downloaded as Crime\_Data\_Austin.csv to a local staging environment by our team on December 2nd of 2016. The data source reported crime incidents from 1/1/2015 to 12/31/2015. The raw data source contains a total of 38,573 records and 13 attributes. These attributes include a Primary Key, Offense Description, Incident Location Address, geographic coordinates about the incidents location and more (see appendix 4.2). Values for each of these attributes included the following data types: String, Boolean, Float, DateTime, and Integer.

The City of Los Angeles Crime Dataset was supplied from data.gov and was published by the data.lacity.org website. The municipal department from which the data was sourced is unknown, although it can reasonably be assumed by that it was gathered by information systems used by police departments within the City of Los Angeles. The information was uploaded to the data.gov website on December 1st of 2016, and was download as Crime\_Data\_Los\_Angeles.csv to a local staging environment by our team on December 2nd of 2016. The data source reported crime incidents from 1/1/2013 to 12/3/2015. The raw data source contains a total of 935,259 records and 13 attributes, making it the largest data source used in this study. The attributes within the dataset include Date Reported, Driver’s License Number, Date of Occurrence, Time of Occurrence, Location, Crime Description, Crime Status, and more (see appendix 4.3). Value for each of these attributes included the following data types: String, Boolean, Float, DateTime, and Integer.

Finally, the weather data used for this study was collected from Weather Underground ([www.wunderground.com](http://www.wunderground.com)) website. Weather Underground is a leading supplier of weather information shared with the public since 1993, and has been collected and supplied by their expert meteorologists from around the globe. Using their historical weather finder our team was to generate a report with the date range of January 1st of 2015 to January 1st of 2016 for each of the crime dataset locations: Chicago, Austin, and Los Angeles. The generated reports were then exported into three separate csv files Weather\_Data\_Austin.csv, Weather\_Data\_Chicago.csv, and Weather\_Data\_Los\_Angeles.csv which were then stored into a local staging environment. Each data set contains a total of 366 records and 23 attributes. These attributes include Max and Mix temperature, Max and Mix humidity levels, perception, cloud cover, and reported date. Values for each of these attributes included the following data types: String, Boolean, Float, DateTime, and Integer.

# 4: Data Pre-Processing

As explained above, multiple data sets were collected for the use of this study. One of the unique challenges of working with four different datasets was combining the data in an accurate and meaningful way.

To begin this process, the crime data sets were inserted into three separate MongoDB database collections from each location’s corresponding crime csv files. Each attribute for each dataset was tested and converted into its corresponding data type as either Float, DateTime, Boolean, or String. Several date time formats were used in each csv file. The conversion of date time strings was handled by matching each unique string date time formats using datetime’s strptime function. Values stored as either N or Y, which stood for no or yes respectively, were converted into a Boolean value of False or True respectively. Finally, any csv string that could casted as a floating point number was stored as floating point, numeric value. In addition, its corresponding attribute name was modified to upper case and was stripped of any none alphanumeric characters. The new attribute names and their corresponding values were then inserted into a JSON document for each record that was then stored in one of the three MongoDB crime data collections by location. The pymongo library used for interfacing to and from the MongoDB database. Source code for the process described above can be found in section three of the Final\_Project python notebook. A sample of an Austin Crime Data Mongo Collection record can be seen in section 5.1 of the appendix.

Once all the data had been converted to JSON and stored in MongoDB, the next phase was to combine these datasets in a meaningful way that could then be used for analysis. Common fields within all three data sets included a description of the incident that occurred, the type of incident, the date and time in which the incident was reported, and the city in which the incident had occurred. These fields were sourced from multiple different attributes in each dataset and were combined in one complete crime data set known as Crime Master Data. The following table displays the Crime Master Data fields and their corresponding field in each source Crime Data Mongo Collection by location. Source for the process described above can be found in section four of the Final\_Project python notebook. A sample of a Crime Master Data Mongo Collection record can be seen in section 5.2 of the appendix.

|  |  |  |  |
| --- | --- | --- | --- |
| Crime Master Data Collection Attribute Name | Chicago Crime Data Collection Attribute Name | Austin Crime Data Collection Attribute Name | Los Angeles Crime Data Collection Attribute Name |
| Reported\_Date | DATE\_OF\_OCCURENCE | GO\_REPORT\_DATE | DATE\_RPTD |
| Month | DATE\_OF\_OCCURENCE.Month | GO\_REPORT\_DATE.Month | DATE\_RPTD.Month |
| Day | DATE\_OF\_OCCURENCE.Day | GO\_REPORT\_DATE.Day | DATE\_RPTD.Day |
| Year | DATE\_OF\_OCCURENCE.Year | GO\_REPORT\_DATE.Year | DATE\_RPTD.Year |
| Description | SECONDARY\_DESCRIPTION | GO\_HIGHEST\_OFFENSE\_DESC | CRMCD\_DESC |
| Type | PRIMARY\_DESCRIPTION | HIGHEST\_NIBRS\_UCR\_OFFENSE\_DESCRIPTION | CRMCD\_DESC |
| City | “CHICAGO” | “AUSTIN” | “LOS\_ANGELES” |

The weather data for each city location was already conformed to have the same attributes. Hence, merging the datasets was a much easier process. First the dataset for each csv file was read in using python. These datasets were then processed similarly to the original crime data sets. Numeric fields were converted to floating point numbers, and date time fields were converted to date time objects using the strptime function of the datetime module. String values were not stored from this data set, and additional cleaning was needed when processing columns that were meant to store numeric values, but contained alpha characters. These values were replaced with a default floating point value of 0.0 before they were inserted into the Weather Master MongoDB collection. In addition, a City field was added to each JSON document to indicate from which the weather data set was sourced.

Finally, the two master data sets for crime and weather were combined by date time at run time before performing any kind of statistical analysis. These two master data sets were also filtered by location to drill down into specific regions of the data set for more granular statistical regressions.

# 5: Model and Algorithm Description

As described in previous sections, we collected crime rates by city. We are interested in modelling count of crimes per day as a function of weather parameters.

We know that crime rates per day have following properties:

1. They are between none to theoretically infinity.
2. The probability of having a given crime rate per day decreases with increasing crime rate.

Thus, the crime rate data nicely fits preconditions of Poisson hypothesis, i.e. number of crimes is likely to have a Poisson distribution with a different Poisson mean every day. Hence, we choose generalized linear model and setup Poisson regression to regress average of crime rates per day as a linear combination of weather parameters.

Our algorithm is as follows:

1. Fit a Poisson regression on all weather parameters.
2. Look at p-values of coefficients and select weather parameters where p-value for significance is < 0.05
3. Refit Poisson regression in-sample with subset of columns that are significant.
4. Get the in-sample plots to see predicted averages vs observed counts in training period
5. Perform goodness of fit test to see if we have a statistically significant relationship
6. Plot out-of-sample predicted values against observed counts to see if we have good predictions
7. Fit binomial model to categorize crime rate into High, Low and Medium categories
8. Plot out-of-sample/in-sample differences in categories

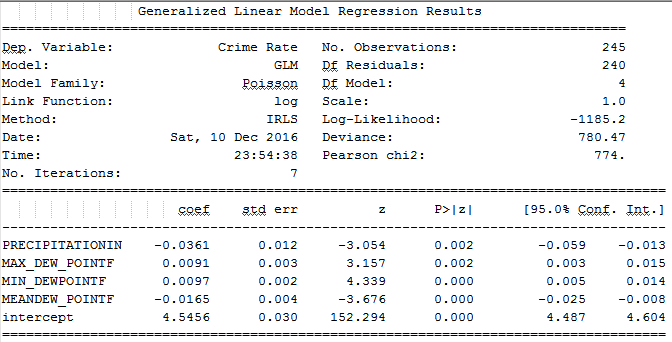
We also tried to fit multinomial logistical regression model with 3 levels and negative binomial model for correcting over-dispersion in Poisson regression fit. However, the model fitting library couldn’t converge with given data. We also tried fitting in crime rates by category but the model had convergence failures when we tried to fit Poisson regression of crime rates for categories such as theft against weather parameters.

In order to fit Binomial model, we assign number 0, 1, and 2 for bottom 33%, middle 33 percentile and >66 percentile categories. Then with response variable having only values in {0, 1, 2}, we fir Multinomial Logistic regression to fit a classification model in in-sample training data (Jan – Aug) and then test out fit in out-of-sample testing. To plot the multinomial logistic regression fit, we plot the difference between predicted and actual value. Value of zero indicates accurate prediction. We look at Poisson regression as limiting case of multinomial Binomial regression with infinite categories.

# 6: Model Results and Evaluation

## Austin Master Data Set

### In-Sample Poisson Data Model



From the model, it is clear that the only significant predictors are P>|z| < 0.05.

Austin crime rates seem largely unrelated to weather conditions. Austin mostly has dry weather and only weather factor affecting crime rates is mean dew and precipitation levels which inversely affect crime rates. Thus, we find that crime rates in Austin are largely independent of weather parameters.

#### Poisson Goodness of Fit Test

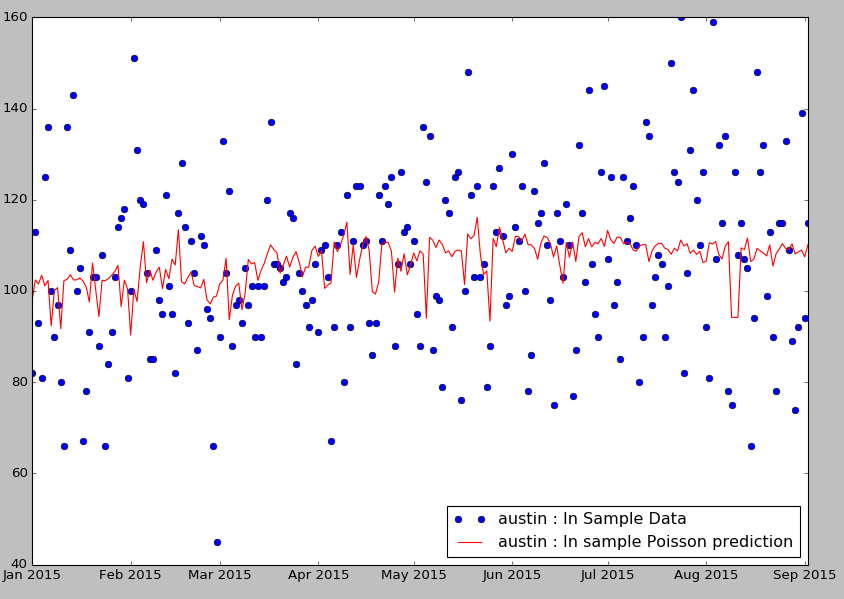
Degrees of freedom of residuals in 239. Pearson Statistic is 774. Thus p-value is:

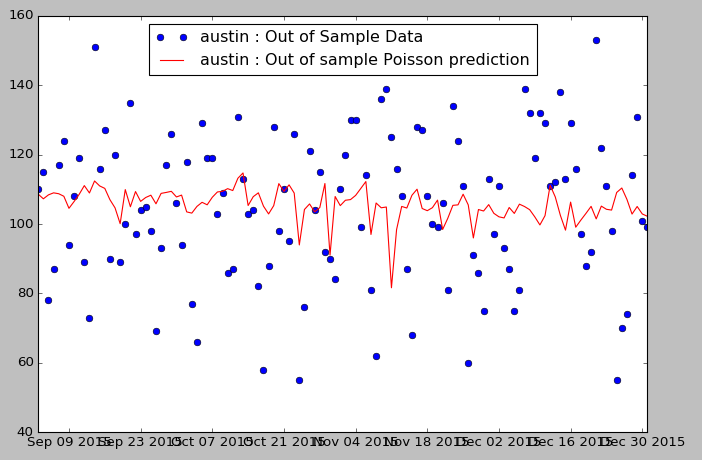
> 1-pchisq(774,239)

[1] 0

Hence, we don’t have sufficient evidence to reject the null hypothesis and over model is not fitted well.

#### In Sample Vs. Out Sample Poisson Regression Plot





### Binomial Data Model

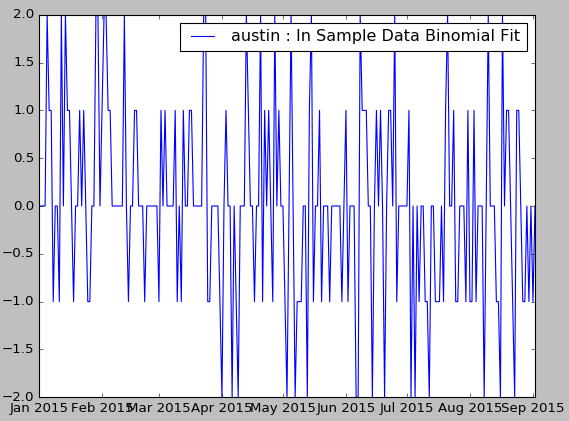
We have provided plots for difference between predicted level and actual percentile level of crime rates per day. Difference of zero indicates accurately predicted value.

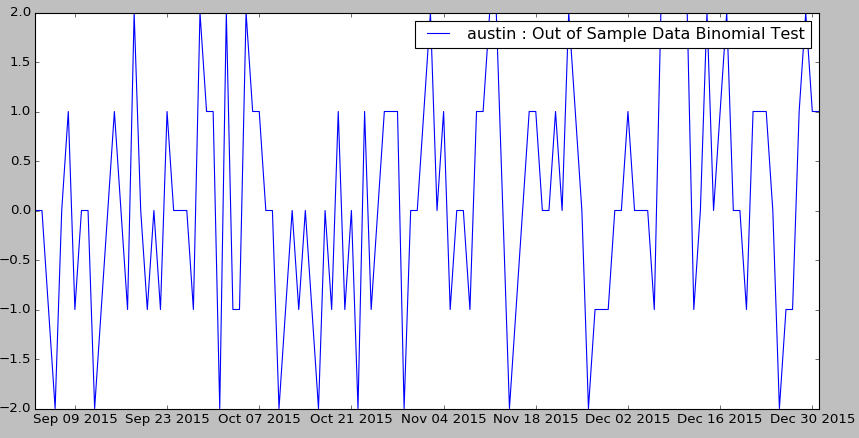
Binomial success In-Sample = 49.39 %

Binomial success Out-of-Sample = 33.33 %

Thus, it’s amply clear that Austin crime rates are unrelated to weather conditions.

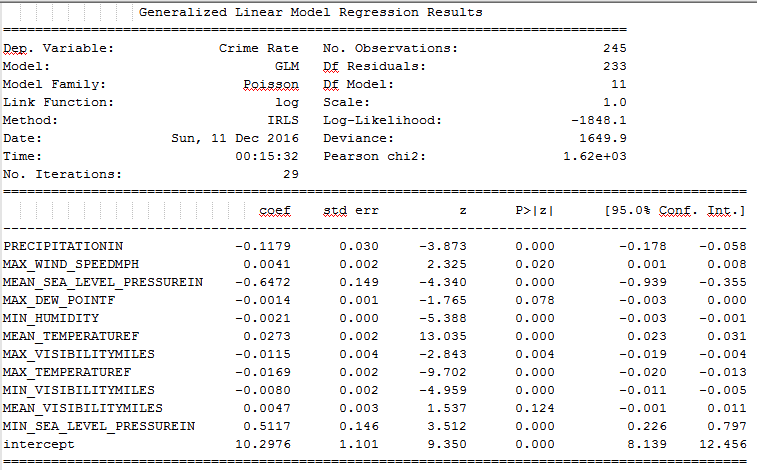
#### In Sample vs. Out Sample Binomial Regression Plot





## Los Angeles Master Data Set

### In-Sample Poisson Data Model



From the model, it’s clear that only significant predictors are P>|z| < 0.05. Looking at the model, we can draw following conclusions:

1. Crime rates are positively correlated with Mean temperature levels.
2. Increasing level of precipitation has negative effect on rate of crimes.

#### Poisson Goodness of Fit Test

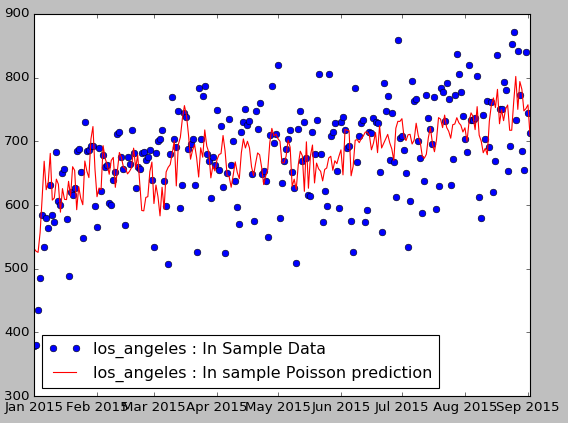
Degrees of freedom of residuals in 233. Pearson Statistic is 1620. Thus p-value is:

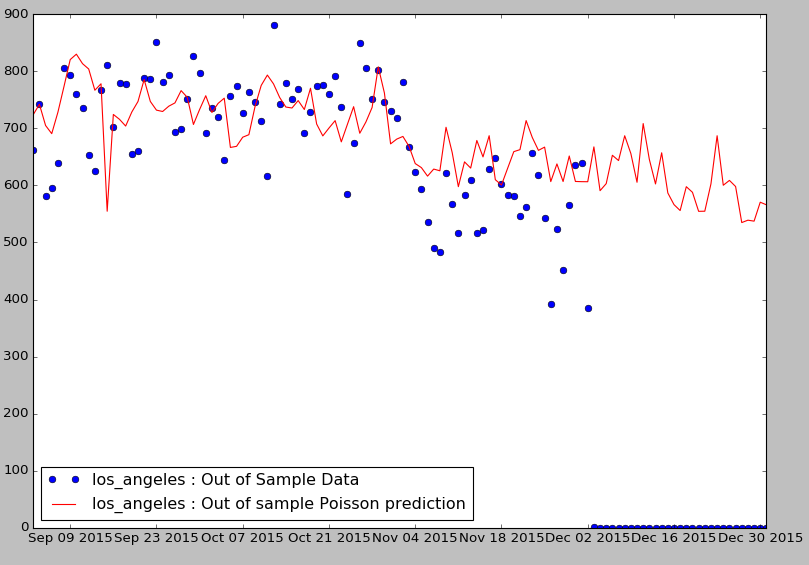
> 1-pchisq(1620,233)

[1] 0

Hence, we don’t have sufficient evidence to reject the null hypothesis and over model is not fitted well.

#### In Sample Vs. Out Sample Poisson Regression Plot





### Binomial Data Model

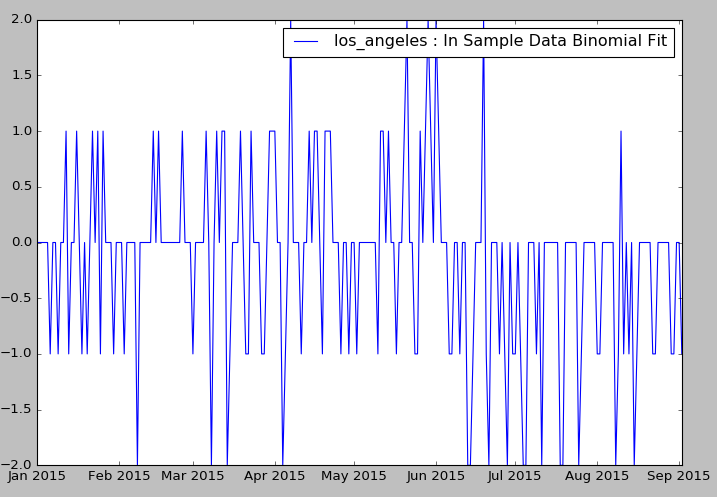
We fit multinomial Logistic Regression:

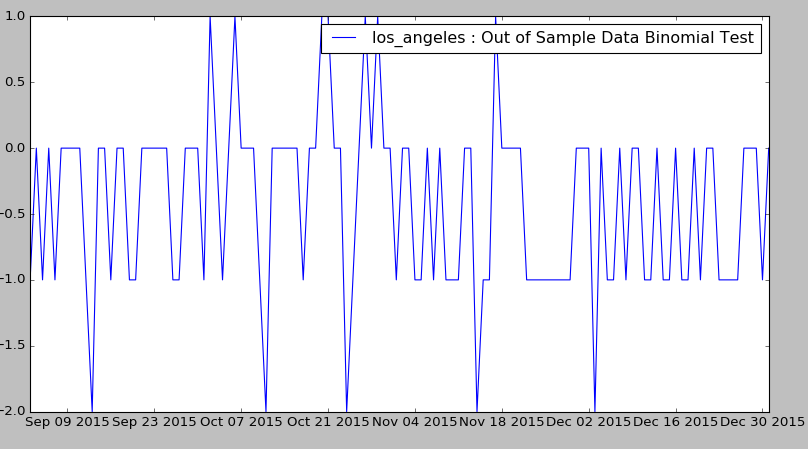
Binomial success In-Sample = 59.18%

Binomial success Out-of-Sample = 51.67%

This looks like a much better fit than Austin.

#### In-Sample vs Out of Sample Binomial Regression Plot

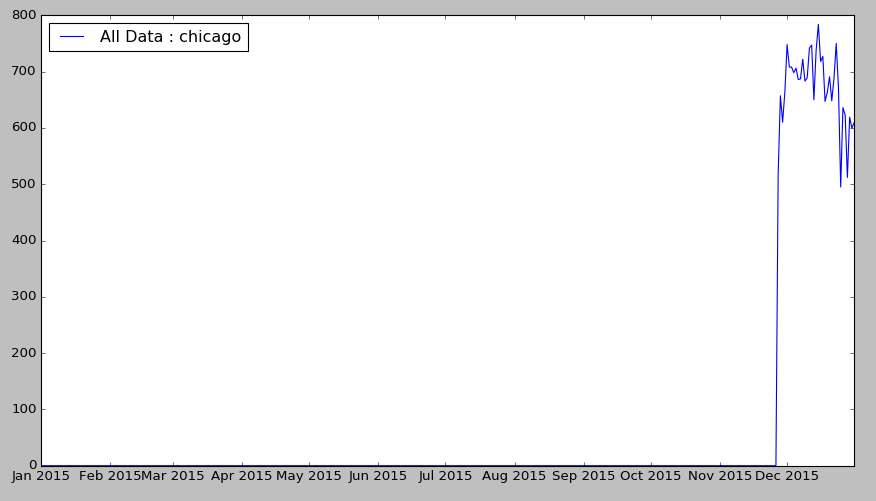


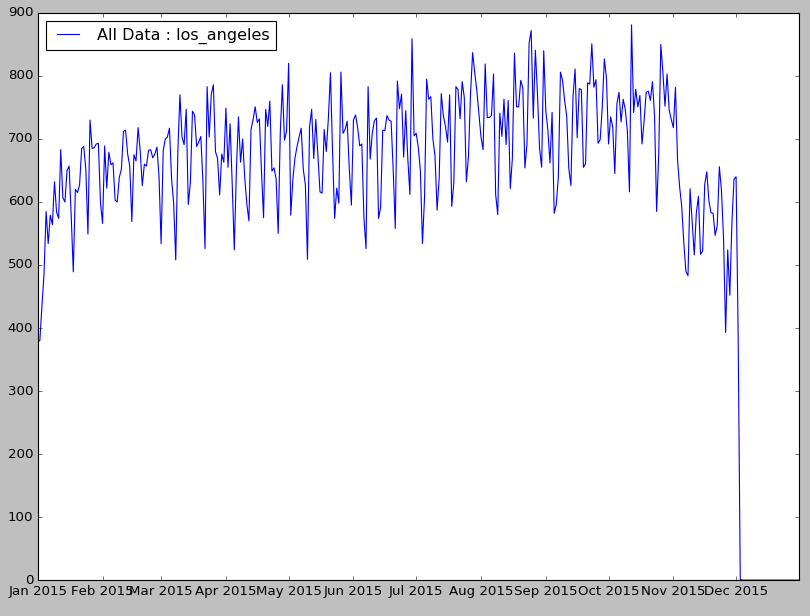


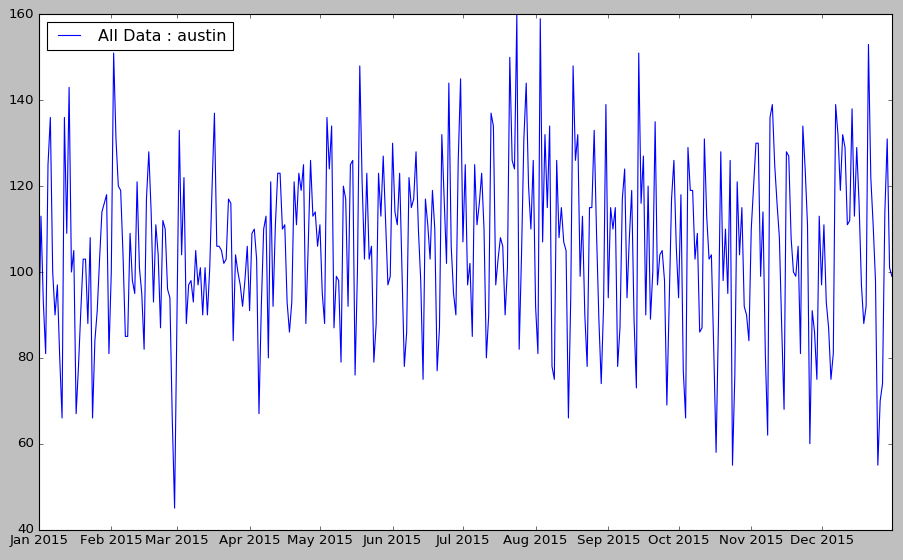
# 7: Conclusions and Lessons Learned

### Data

We plotted complete data for all 3 cities towards the end of the project and we noticed the following plots:







We realized that our Poisson model worked as we had a bug in the code to check for just day of crime as opposed to day and month. Hence, we realized we never had complete dataset for Chicago only when we plotted the complete count of crimes dataset.

### Model

We started with an assumption on distribution of crimes data and came up with factors that could affect the crime rates and nature of distribution. After having reasonable intuition about nature of likely model, we assumed that over the year of model data, population factors such as age, gender, and average income distribution remain constant.

We then came up with Generalized Linear Model Poisson regression of daily crime rates against weather parameters. We then pruned the parameters based on strength of relationship using coefficient p-values and tried to further correct the deficiencies in model. We got significant linear relationship but not a statistically significant good fit for Poisson regression. We also realized that in Austin, crime rates are by and large unrelated with weather conditions.

To correct the deficiencies in model fit, we attempted following approaches:

1. Try fitting model of crime rates by category against weather conditions
2. Try to categorize overall crime rates into high, low, medium category and try to come up with Multinomial Logistic regression (binomial regression) against weather conditions

However, in both these approaches, the model failed to provide a statistically good fit. We perhaps needed longer dataset than the in-sample data size of 244 to fit the model. While fitting the model, we also encountered bugs due to unclean data and had to insert code to guard against data dimension mismatch in handling multi-dimensional data as this is crucial for model code to fit.

# Appendix

## 4.1 Chicago Crime Sample Data Set

|  |  |
| --- | --- |
| Attribute Name | Value |
| CASE # | HZ209325 |
| DATE OF OCCURRENCE | 3/26/2016 2:00 |
| BLOCK | 002XX S HALSTED ST |
| IUCR | 810 |
| PRIMARY DESCRIPTION | THEFT |
| SECONDARY DESCRIPTION | OVER $500 |
| LOCATION DESCRIPTION | STREET |
| ARREST | N |
| DOMESTIC | N |
| BEAT | 1232 |
| WARD | 27 |
| FBI CD | 6 |
| X COORDINATE | 1171090 |
| Y COORDINATE | 1899016 |
| LATITUDE | 41.87837335 |
| LONGITUDE | -87.6472551 |
| LOCATION | (41.878373345, -87.647255095) |

## 4.2 Austin Crime Data Sample

|  |  |
| --- | --- |
| Attribute Name | Value |
| GO Primary Key | 201510782 |
| Council District | 1/4/1900 0:00 |
| GO Highest Offense Desc | AGG ROBBERY/DEADLY WEAPON |
| UCR Offense Description | Robbery |
| GO Report Date | 1-Jan-15 |
| GO Location | 9001 N IH 35 SVRD NB |
| Clearance Status | N |
| Clearance Date | 28-Jan-15 |
| GO District | E |
| GO Location Zip | 78753 |
| GO Census Tract | 18.13 |
| GO X Coordinate | 3130483 |
| GO Y Coordinate | 10102366 |

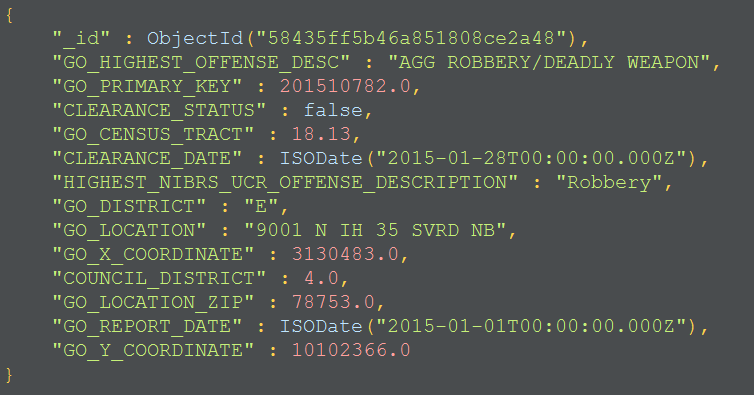
## 4.3 Los Angeles Crime Data Sample

|  |  |
| --- | --- |
| Attribute Name | Value |
| Date.Rptd | 3/20/2013 |
| DR.NO | 132007717 |
| DATE.OCC | 3/20/2013 |
| TIME.OCC | 2015 |
| AREA | 20-Jan-00 |
| AREA.NAME | Olympic |
| RD | 2004 |
| Crm.Cd | 23-Sep-02 |
| CrmCd.Desc | TRAFFIC DR # |
| Status | UNK |
| Status.Desc | Unknown |
| LOCATION | OXFORD |
| Cross.Street | OAKWOOD |
| Location.1 | (34.0776, -118.308) |

## 4.4 Weather Data Sample

|  |  |
| --- | --- |
| Attribute Name | Value |
| CST | 1/1/2015 |
| Max TemperatureF | 40 |
| Mean TemperatureF | 2/6/1900 |
| Min TemperatureF | 34 |
| Max Dew PointF | 8-Feb-00 |
| MeanDew PointF | 35 |
| Min DewpointF | 29 |
| Max Humidity | 9-Apr-00 |
| Mean Humidity | 82 |
| Min Humidity | 64 |
| Max Sea Level PressureIn | 30.38 |
| Mean Sea Level PressureIn | 30.25 |
| Min Sea Level PressureIn | 30.15 |
| Max VisibilityMiles | 10 |
| Mean VisibilityMiles | 3 |
| Min VisibilityMiles | 1 |
| Max Wind SpeedMPH | 9 |
| Mean Wind SpeedMPH | 5 |
| Max Gust SpeedMPH | 13 |
| PrecipitationIn | 0.63 |
| CloudCover | 8 |
| Events | Rain-Thunderstorm |
| WindDirDegrees | 359 |

## 5.1 Austin Data Set JSON Exert



## 5.2 Crime Master Data Set JSON Exert

