Chicago Crime Scenes

**Project Team:** 1. Spoorthi Gunuganti

2. Srishti Kabtiyal

3. Trisha Chandra

*ORIGINAL WORK STATEMENT*

*We the undersigned certify that the actual composition of this proposal was done by us and is original work.*

Signatures:

|  |  |
| --- | --- |
| **Name** | **Signature** |
| Spoorthi Gunuganti | SG |
| Srishti Kabtiyal | SK |
| Trisha Chandra | TC |

Table of Contents

[I. EXECUTIVE SUMMARY 2](#_Toc483082609)

[II. DATA DESCRIPTION 3](#_Toc483082610)

[III. RESEARCH QUESTIONS 4](#_Toc483082611)

[IV. METHODOLOGY 4](#_Toc483082612)

[Binomial Linear Regression 5](#_Toc483082613)

[V. FRONT END 6](#_Toc483082614)

[VI. RESULTS AND FINDINGS 7](#_Toc483082615)

[Prediction and Performance Parameters 7](#_Toc483082616)

[VII. CONCLUSION 7](#_Toc483082617)

[VIII. LIMITATIONS AND IMPROVEMENTS 7](#_Toc483082618)

[IX. APPENDIX 8](#_Toc483082619)

[ Charts on Tableau 8](#_Toc483082620)

[ Charts on Python 9](#_Toc483082621)

[ Data Cleaning on R 10](#_Toc483082622)

[ Data pre-processing using Pyspark 12](#_Toc483082623)

[ Model creation in Pyspark 13](#_Toc483082624)

[ Prediction and performance parameters 14](#_Toc483082625)

[X. REFERENCES 15](#_Toc483082626)

# EXECUTIVE SUMMARY

The dataset has been collected from Kaggle website and is a compilation of crimes as reported for Chicago from the year 2012 to 2017. The data has been collected from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. Chicago’s overall crime rate has been substantially higher than the US average. It is thus imperative that studies be done in this arena to shorten or improve upon the statistics reported or recorded. Our project is a small attempt in dealing with this issue that has been plaguing the society for so long. Through the big data techniques we have learnt through the semester, we aim towards identifying the number of arrests for a predicted set of conditions or circumstances reported (independent variables discussed later). We tried to predict the number of arrests with utmost accuracy possible.

# DATA DESCRIPTION

**Data Source:** Our data source was acquired from Kaggle.com and has been collected from Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. It comprises of crimes reported for the city of Chicago from the year 2012 to 2017 (most recent).

The various fields in our data set are described below:

|  |  |  |
| --- | --- | --- |
| Sr No. | Field | Description |
| 1 | ID | Self-Explanatory |
| 2 | Case Number | Self-Explanatory |
| 3 | Date | Date on which the incident was recorded |
| 4 | Block | Block in which the incident occurred |
| 5 | IUCR | The Illinois Uniform Crime Reporting code. |
| 6 | Primary Type | The primary description of the IUCR |
| 7 | Description | Description of the crime reported |
| 8 | Location Description | Description of the location in which the crime was committed |
| 9 | Arrest | Whether an arrest happened or not |
| 10 | Domestic | Whether the arrest was a domestic arrest or not |
| 11 | Beat | Indicates the smallest police geographical area where the crime was reported |
| 12 | District | The district in which the crime was reported |
| 13 | Ward | Ward in which the crime was reported |
| 14 | Community Area | Community area in which the crime was reported |
| 15 | FBI Code | FBI code assigned to the crime |
| 16 | X Coordinate | X-coordinate of the crime location |
| 17 | Y Coordinate | Y-coordinate of the crime location |
| 18 | Year | Year in which the crime was committed |
| 19 | Updated On | Date on which the crime details were last recorded |
| 20 | Latitude | Latitude details corresponding to the crime location |
| 21 | Longitude | Longitude details corresponding to the crime location |
| 22 | Location | Location of crime |

The dataset consisted of a total of about **1 million** records with 31 variables each with some instances having occurrence of null values.

# RESEARCH QUESTIONS

The fundamental goal of this project was to determine whether an arrest would happen following a crime or not. The dataset we used for our project was from years 2012 to 2017, and had a set of about 22 variables. We cut down this to the most relevant variables which were taken as our independent variables. We recognized five primary independent variables for our project viz. domestic, primary type, location description, hour and district. Basis this, we predicted the outcome of our dependent variable – arrest. The outcome of this prediction would be either yes or no. In binary terms, it was a 0 or 1. 0 for no arrest and 1 for arrest.

We had to perform data cleaning at various levels in order to transform our dataset into a feasible solution which could thus be used for our desired predictions. We then performed multinomial linear regression for our model creation and finally obtain the desired prediction result.

# METHODOLOGY

We started off with creating visualization charts using Python and Tableau. These charts were mostly created using the entire dataset. We did not perform any substantial data cleaning before creating charts. This is so because we directly took the variables across which we wanted to check the relationship. Charts gave us a clarity on various details such as the most prevalent types of crimes, or the most crime prone locations, the trend of crime rates over the years and across months, etc. These have been added in appendix at the end of the report.

Next, we performed a data cleaning operation to eliminate any columns that did not add value or were posing a hindrance to the functioning of the predictive model that we were trying to build. This data cleaning involved the following steps:

* **Elimination of Null values:** This stage involved the elimination of all tuples that had one or more attributes associated with null occurrences or missing data.
* **Date and time transformation in R and parsing on date in Python:** The date and time formats in our dataset were not consistent. We thus had to convert them into appropriate format in order to be able to use them for our project.
* **Parsed dates**: The time variables were converted from an absolute time to buckets specifying the hour of day. This would be a more interpretable outcome.
* **Elimination of Cause variables:** Certain variables had little to no contribution towards the predictive analysis of our model. We thus decided to get rid of these variables and
* **Binning of independent variable:** The independent hour variable was binned or bucketed based upon which business scenario was catered to by the model in question.

We employed all the below mentioned model when performing our analyses.

## Binomial Linear Regression

This is a predictive analysis method used to describe the relationship between data. This explains the relationship between one dependent and one or more independent variables in our dataset. As mentioned above, the five key independent variables we took for predicting our outcome were - domestic, primary type, location description, hour and district. And the dependent variable or the variable we were predicting was arrest. We performed the analysis using Pyspark.

To accomplish the same, we split our dataset into 70:30, for training and test respectively. Prior to training the model, we converted a couple of string variables into factors using the string indexer. We then grouped all the input variables into a single vector column called features. This was done using a vector assembly function. Logistic regression model was then used to train the model and then transform prediction on test data.

We also converted both output dataset and test dataset to pandas to perform merge. The dataframe obtained was then relocated to AWS in the form of a CSV.

# FRONT END

We created out front end through tableau public. We created several charts and dashboards in Tableau, which were then hosted on an HTML webpage. Following is the link to the charts and dashboard that were created.

<http://terpconnect.umd.edu/~sgunugan/>

As can be seen, the front end is a user-friendly page that can be navigated through various horizontal tabs corresponding to a chart or a dashboard created. The charts and dashboards can be manipulated by altering or changing the values of the filters added.

Following is a brief explanation of the charts/dashboards:

1. Predictions based on location: Number of predicted arrests for different primary types ad locations for each hour- morning, afternoon and evening. Details can be seen by hovering over the graph, which will make a tool-tip appear that will contain details of the bar in question.
2. Predictions based on primary type: Chart that shows the number of predicted arrests based on primary type.
3. Dashboard primary type and location: This is a dashboard that shows the number of predicted arrests for three filters- primary type, location and hours. One can see that we have added an “activity” feature on the charts such that the two charts can be manipulated together through the mentioned filters.
4. Predictions – all parameters- This is a chart that shows the number of predicted arrests corresponding to the five input parameters or independent variables that are discussed above in the report. This is a detailed view or analysis of the predictions based on the independent variables.
5. Pred- hourly basis: This is a chart that shows the count of predictions on hourly basis- morning, afternoon and evening. The yellow bar shows the number of predicted arrests whereas the blue bar shows the non-arrest events.

1. Predicted Arrests: This is a dashboard that will show the predicted number of arrests vs non-arrest for a set of input parameters or variables.
2. Cumulative View: Dashboard that shows the number of predicted arrest corresponding to the five independent variables. One can select filters based on will and can check the predicted number of arrests in the output graph on the dashboard.

Note: The number of predicted arrest is a cumulative figure for all years for which the dataset is taken- 2012-2017

# RESULTS AND FINDINGS

## Prediction and Performance Parameters

We compared our prediction results to actual test data and obtained the following performance evaluation metrics.

* Accuracy – 71%
* Sensitivity- 0.84
* Error Rate- 0.28

# CONCLUSION

In conclusion, it is evident from the above analysis that this was indeed a very challenging data set to work with, especially considering the nature of the question that we had posed for analyses. Crime rate has been on the rise for a long time now and it is now more than ever that we need to put all our efforts into curbing the same.

In an era where technology provides a solution to almost all our problems, there is no reason for it to be ignored in providing a solution for such a significant social cause. Big data solutions can be put to good use in predicting the outcomes of an event based on a certain set of input parameters. Appropriate actions can thus be taken based on the results achieved.

# LIMITATIONS AND IMPROVEMENTS

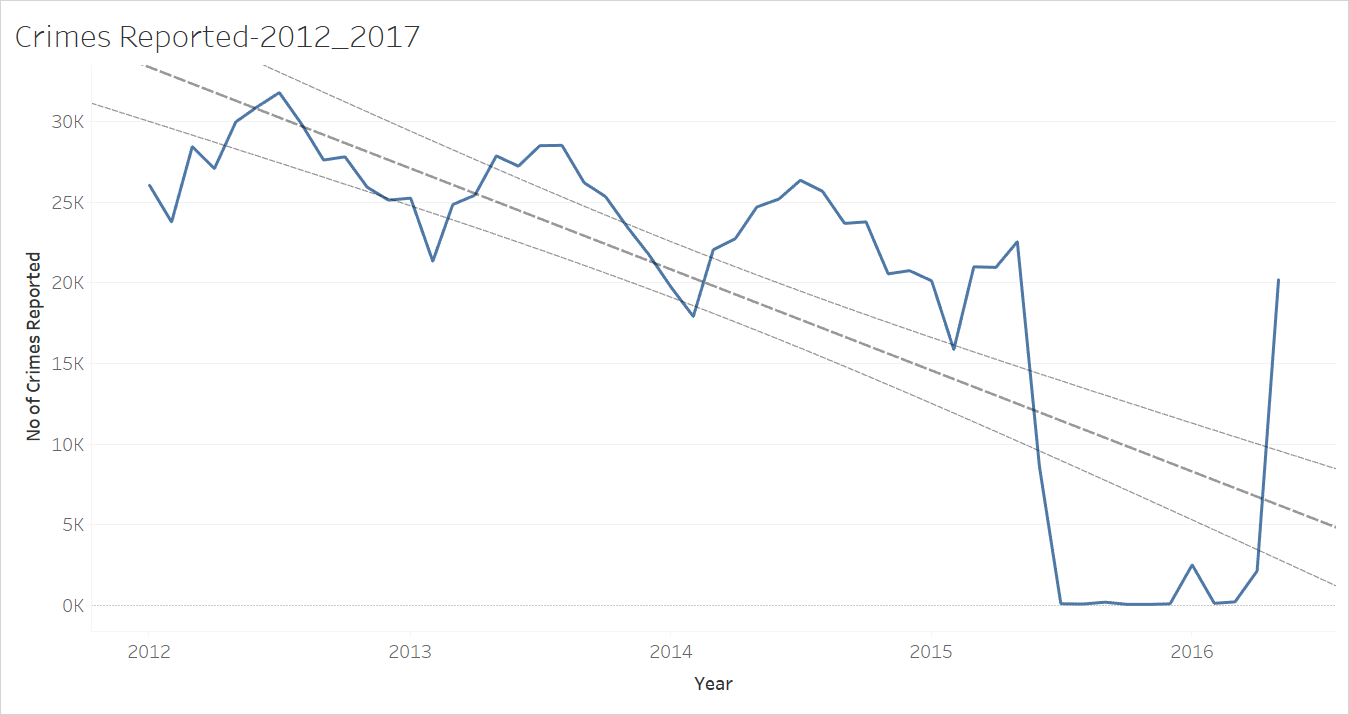
Improvements can be made in terms of obtaining better specificity for our model. There are very few instances with the success class while training the model and therefore the model does not accurately predict their occurrences.

We could probably perform over sampling and provide the training data set with more such instances. Also, diverse algorithms would have helped us gain a comparative analysis of the best solution possible.

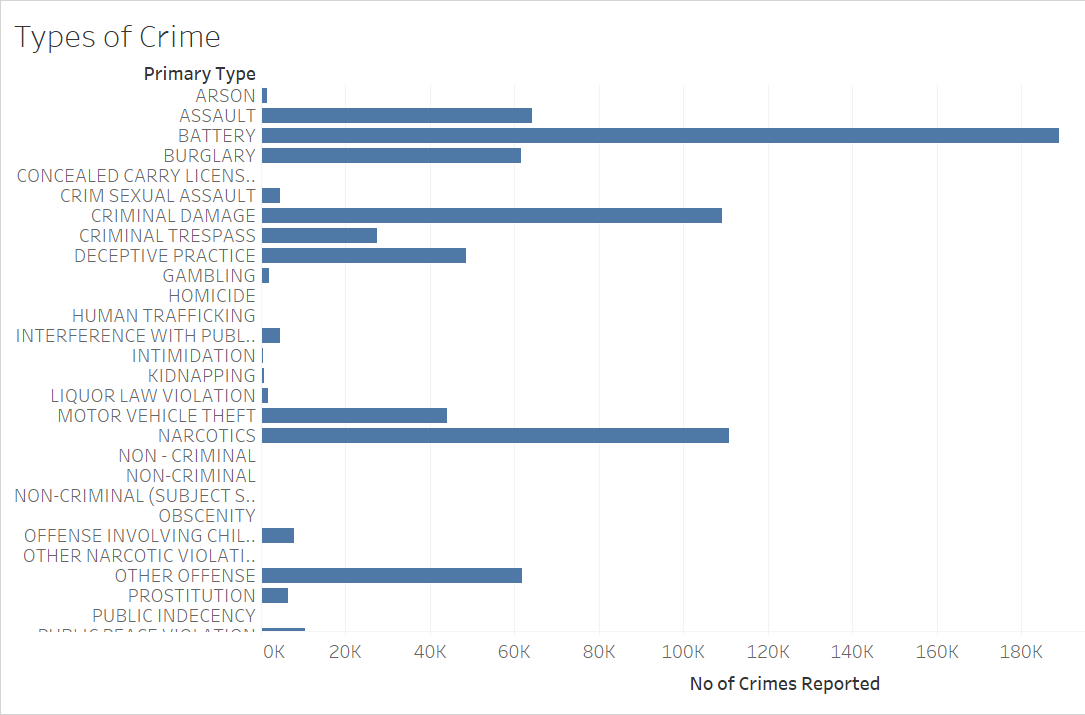
# APPENDIX

## Charts on Tableau

No of crimes reported over the years in Chicago from 2012 to 2017.

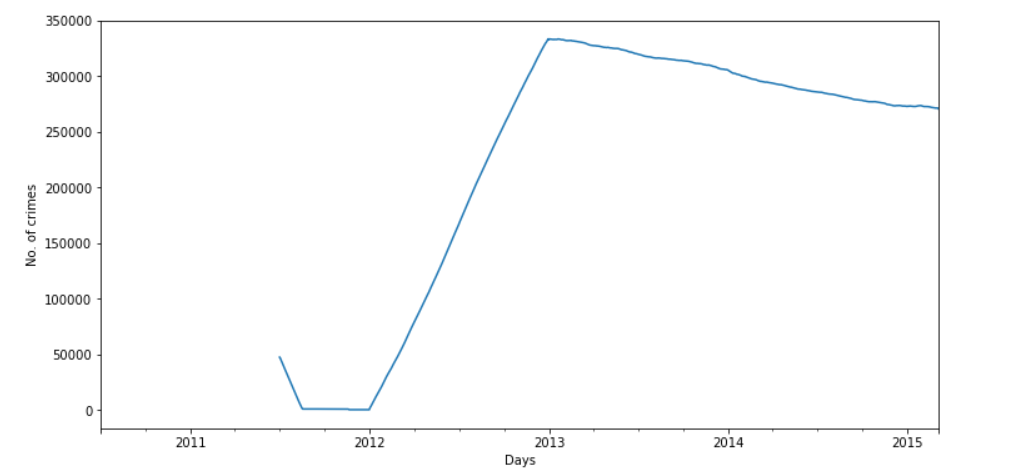


Types of crimes reported based on their frequency.

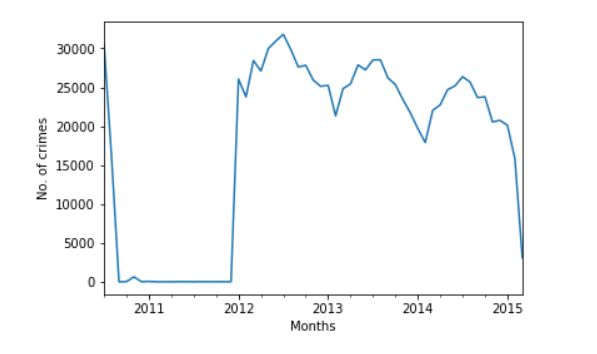


## Charts on Python

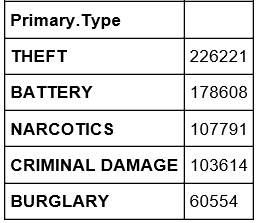
No of crimes reported over years 2012 to 2016.



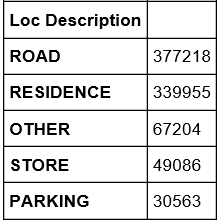
No of crimes reported over months across years in Chicago from 2012 to 2017.



Most prevalent types of crimes with respect to the primary type

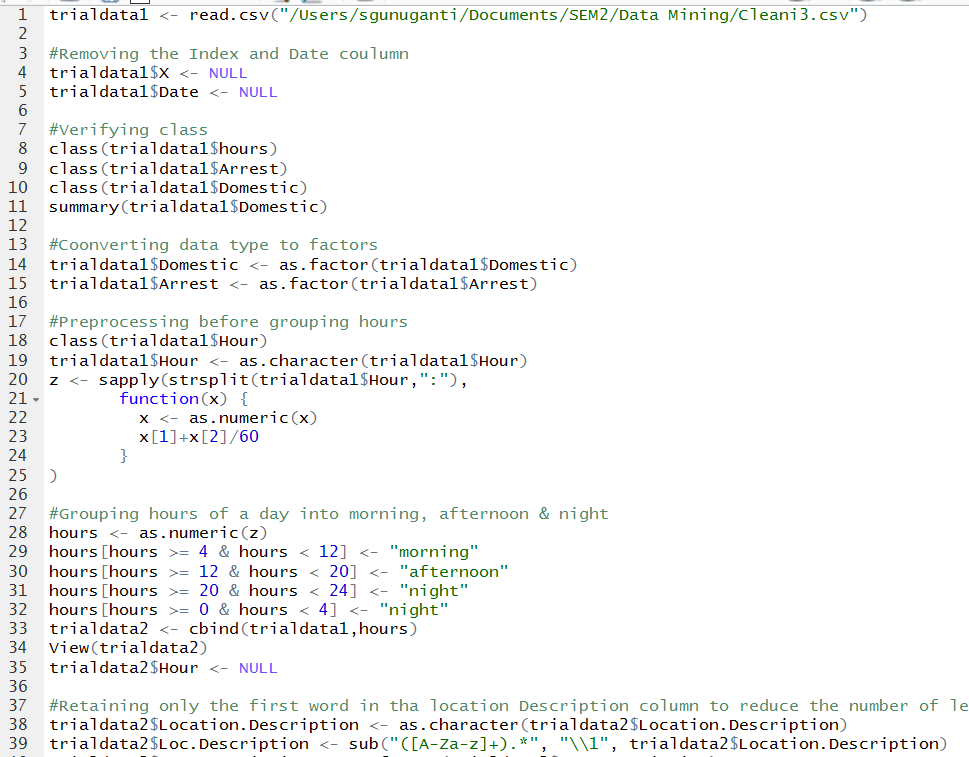


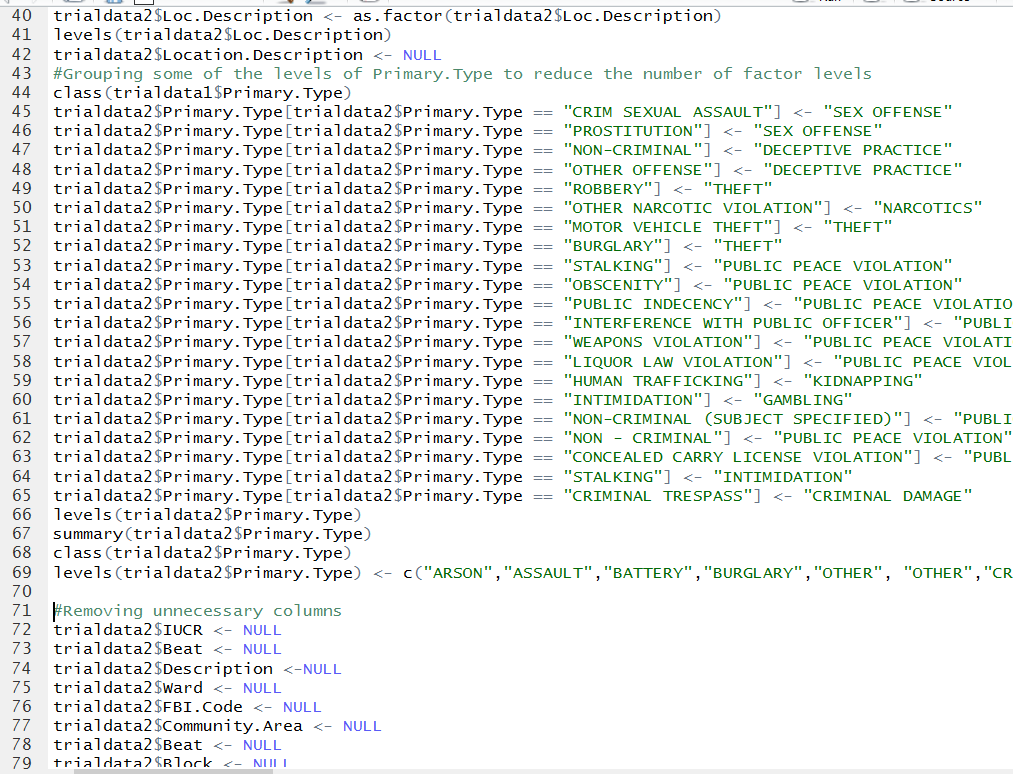
Most prevalent types of crimes with respect to location



## Data Cleaning on R









## Data pre-processing using Pyspark



## Model creation in Pyspark



## Prediction and performance parameters



# REFERENCES

* Our data source was acquired from Kaggle.com - https://www.kaggle.com/currie32/crimes-in-chicago
* <http://www.statisticssolutions.com/mlr/>
* <https://www.analyticsvidhya.com/blog/2016/02/multinomial-ordinal-logistic-regression/>