

**Individual Assignment**

**DATA ANALYTICAL PROGRAMMING**

**(DAP)**

**(CT050-3-M)**

Name: Aliff Marzuki Bin Mahmud Marzoki

Student ID: TP019595

Intake: UCMF1704DSBA

Lecturer: Dr. Kalai Anand Ratnam

Submission Date: 18 December 2017

Contents

[Introduction 3](#_Toc501666253)

[Methodology 4](#_Toc501666254)

[Dataflow Diagram 4](#_Toc501666255)

[Dataset 4](#_Toc501666256)

[Data Preparation 6](#_Toc501666257)

[Additional information removed 6](#_Toc501666258)

[Merging of cells 6](#_Toc501666259)

[Heading changes 6](#_Toc501666260)

[Merged columns are detached 7](#_Toc501666261)

[Missing population values 7](#_Toc501666262)

[Missing values 8](#_Toc501666263)

[Dropping Columns 8](#_Toc501666264)

[Creation of Library 8](#_Toc501666265)

[Analysis 9](#_Toc501666266)

[Conclusion 14](#_Toc501666267)

[References 15](#_Toc501666268)

[Appendix 16](#_Toc501666269)

# Introduction

The assignment is the analysis of violent crimes, property crimes and arson crimes within the United States for the first 6 months of 2014 and 2015. By applying analytics to crime data, we can benefit from its where we can help detect and reduce crime rate with crime prediction and more efficient ways to enhance the police force. The dataset was obtained from the Uniform Crime Reports of the FBI for cities with population of 100,000 and more. The data is provided in excel format and needs to be imported to SAS to carry out analytics.

SAS enables organisations to transform data into insights which results into action. It allows the users to explore and analyse data and it also allows for modelling and predicting the outcome of the data also. SAS can access different forms of data and allow it to be accessed and changed to the SAS format. Once all that would be done, we can extract the data and make summary reports in a form of PDF and the organisation can look upon it.

There are many reasons to why we need to analyse the crime data that has been given. We need to look at the crime reduction as the police force were not able to deal with certain crime scenes as maybe they were unprepared for it or even are unable to handle the situation as they have limited amount of personnel which are able to take care of the incidents. Hence with the data, the police force would be able to decide on where resources should be spent and how they should use them.

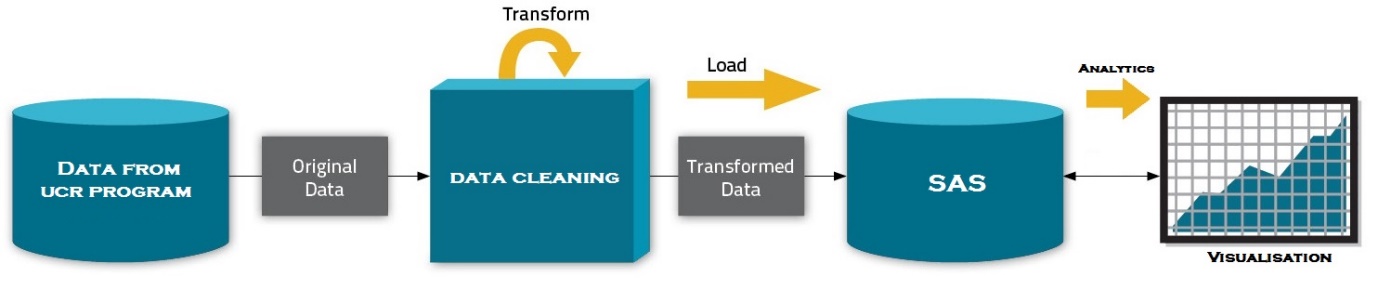
# Methodology

Data mining is an important aspect of the entire operation as it is the process of finding outliers, patterns and correlation within a large dataset to predict outcomes (SAS, 2017). Getting all these information from the dataset is crucial into understanding the situation as a whole. With that being said, we will be using the Cross-Industry Standard Process for Data Mining (CRISP-DM) method as the base to complete the analysis.

CRISP-DM is a methodology that provides a structed approach to planning a data mining project (Smart Vision, 2017). CRISP-DM is built upon these six processes which are business understanding, data understanding, data preparation, modelling, evaluation and deployment. Each of these processes are important to the work that will be given as at each step we can evaluate the data that is represented in each aspect.



## Dataflow Diagram



## Dataset

The dataset given consist of 15 columns which consist of State, City, Year, Population, and the type of crimes where cases have been registered by the FBI. The table consists of 520 rows of data in which there are some that are incomplete and with missing values. We found that there are several footnotes to the data, which are:

* The 2014 population figures are FBI estimates based on provisional data from the U.S. Census Buruea
* Arson offenses are reported by the Toledo Fire Department, hence Arson figures are not entirely included
* The data that is available from January until June 2014 is not complete

There are several columns within the dataset that will be listed below. This dataset also shows cities within the United States with more than 100,000 in population size, also the dataset will only have reports that have been recorded by the FBI and only within the first half of the year of 2014 & 2015.

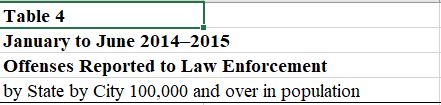
|  |  |
| --- | --- |
| State | The state within the United States |
| City | The city within the state |
| Year | Year it was published |
| Population | The population of the city |
| Violent Crime | Violent Crime that was reported |
| Murder | Murder that was reported |
| Rape (Legacy definition) | Rape that was reported |
| Rape (Revised definition) | Rape that was reported |
| Robbery | Robbery that was reported |
| Aggravated Assault | Aggravated Assault that was reported |
| Burglary | Burglary that was reported |
| Larceny Theft | Larceny Theft that was reported |
| Motor Vehicle Theft | Motor Vehicle Theft that was reported |
| Property Crime | Property Crime that was reported |
| Arson | Arson crime that was reported |

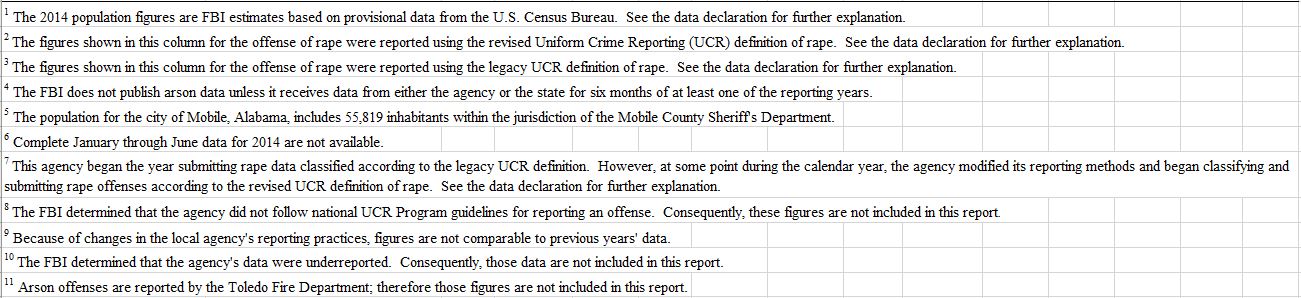
# Data Preparation

We need to prepare the data and pre-process it to be able to able to get a good significant amount of data. Unnecessary data are included with the excel file and this might cause problems to the outcome if not stopped at all. In this stage, the data would be cleaned, missing data substituted and several changes would happen to the dataset that will be done without violating th provided data values.

## Additional information removed

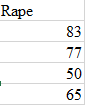
There were additional data that was present within the dataset. This additional data shows what the dataset is about and what it actually is the meaning behind them. Since they are non-essential, it will be removed.





## Merging of cells

Rape (revised definition) and Rape (legacy definition) are merged into one variable as this would make it easier to classify rape cases, as it would just a slight definition change that occurred.



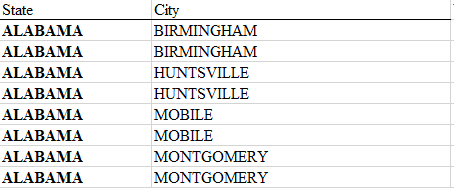
## Heading changes

The headers are relabelled and the columns that consists years 2014 and 2015 have been changed to Year. This is to streamline the process of labelling.



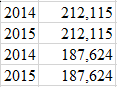
## Merged columns are detached

Merged column that are merged would be detach so that it would be easier to provide all the necessary data within the dataset.



## Missing population values

The missing population values for 2015 is assumed unchanged and that we will use the year 2014 as the reference.



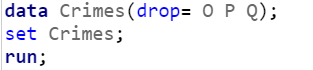
## Missing values

The missing crime values will be replaced with 0. This is to ensure that the missing values are represented at a minimum as the Toledo Fire Department has not reported them yet.



## Dropping Columns

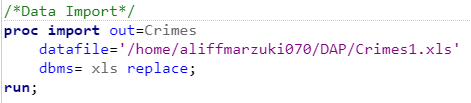
We dropped some columns from the dataset as they are just there being some redundant columns with no value. This also helps us to streamline the dataset.



## Creation of Library

We created a folder within the SAS and uploaded the data onto it. A library is created and the data is imported to the library. It is showed as follows:

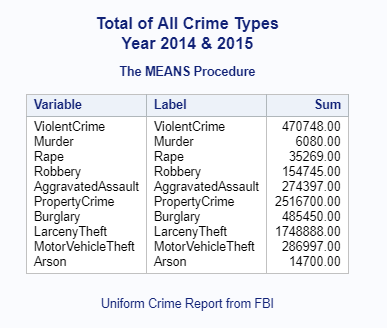




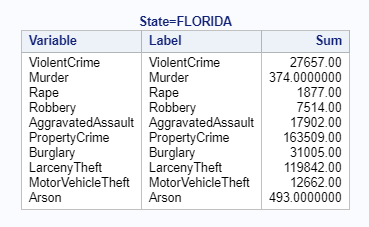
This is to ensure that the dataset is within the SAS library and that no issues of non-existing data table cannot be formed.

# Analysis

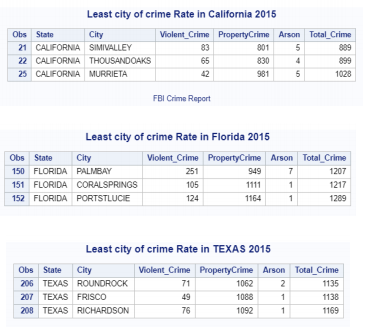
Below are some of the reports that have been figured and shown within the dataset and the overall outcome. Do note that this is not the final version as there are some issues with the dataset.



The above figure is the total crime shows that Property Crime has more cases than the second highest recorded figure which is the Violent Crime and the third worst crime would be Larceny Theft. This is quite significant as we can see that Property Crime is a huge deal and that the police force would need to take specific action against these kinds of cases.

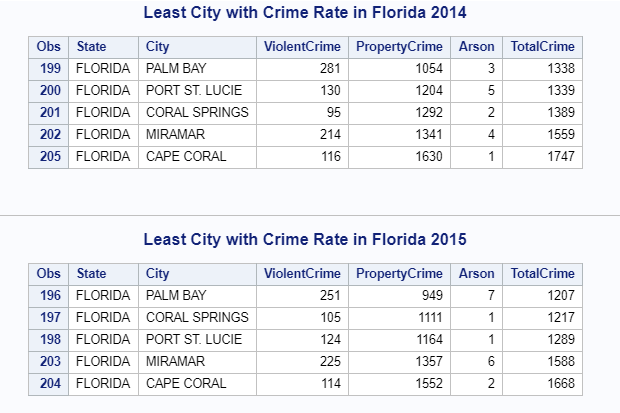


In the above figure we look at the state of Florida and the total sum amount crimes that have happened there. We see that the worst case would be the Property Crime, second would be Larceny Theft and third would be Burglary. We see here that there is a high amount of property crime and larceny theft compared to the rest of the crimes that has occurred over within the state.

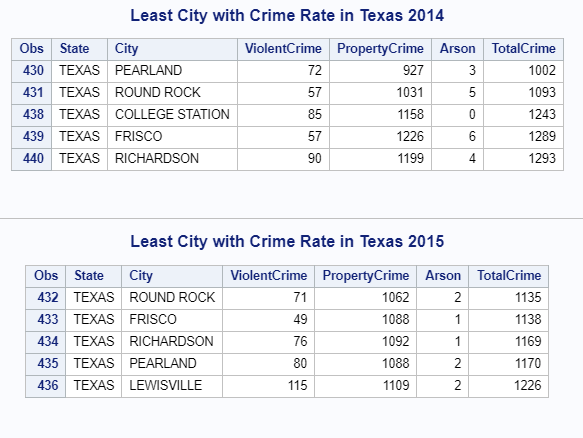


The above figure shows the report of the three less risky cities for crime in the year 2015 listed based on each state. We see here that the 3 least cities with the total crime value amongst all the three states of United States would be Simmi Valley, California with 889 Crimes, Roundrock, Texas with 1135 crimes and Palmbay, Florida with 1207 crimes.

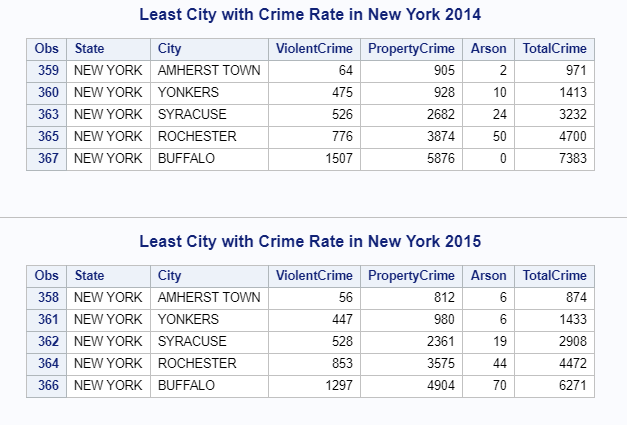
The data visualisation would be included within the updated documentation and it would include a more thorough approach towards the system itself.



The figure above shows the least crime rates in Florida for both 2014 and 2015. In 2014, we see that Palm Bay is the lowest total crime in both years. In 2014, it had a total crime rate of 1338 but in 2015 there was a reduction in crime until 1207. We can also see the other cities with the Florida state have also have a decrease in reported cases, other than Miramar. This is really great as it shows that the police force is having a great impact in Florida.



The figure above shows the least crime rates within Texas for both 2014 and 2015. In 2014, we see that Pearland has the least amount of crimes that have appeared within the city at 1002 total crimes, while in 2015 we see that the least city with rimes is Round Rock with 1135. We see that there was an increase in crimes in Pearland of 132 crimes and also an increase of crimes at Round Rock from 1093 to 1135 crimes. But overall there was an increase of crimes with these five cities and that it is important to note that the police forces there seem to have trouble in containing the crimes that are occurring there, so a look at the system in Texas is important.



The figure above shows the least crime rates in New York cities in the year 2014 & 2015. We see that in 2014, Amherst Town is rated the city with least crime with 971 crimes while it also maintained its status in 2015 with a crime rate of 874. We see here that overall there has been a decline in crimes cases in 2015 compared to 2014. The police force has taken liberty into making sure that there is less crime within the city as New York is one of the smaller states with a huge population.

# Conclusion

Analytics would most definitely help whichever organisation would to use it and it can create and outcome that they are able to view and make decisions of their own to fix whichever problem that has occurred and prevent. SAS has all the tools to make it work which involves around the pre-processing to the post-processing to the visualisation of the outcome also. It helps that the tool that is available must include all the relevant tools to make it work properly.

We found out that that within the above dataset that the police force should prioritize slightly more towards Violent Crime and Property Crime as they are the highest recording crimes within the United States for the year 2014 & 2015. This is important as they are able to redistribute resources evenly amongst the cases and change them accordingly when crimes increase or decrease. It is also easier to create a task force that will specialise in the highest recorded cases, so that they are able to reduce the crimes.

# References

Federal Bureau of Investigation. (2016) FBI Releases 2015 Crime Statistics. [Online] Available from: https://www.fbi.gov/news/pressrel/press-release/fbi-releases-2015-crime-statistics [Accessed 10 December 2017]

SAS. (2017) Data Mining. What is it and why it matters. [Online 14 December 2017] Available from: https://www.sas.com/en\_us/insights/analytics/data-mining.html [Accessed 14 December 2017]

Smart Vision. (2016) What is the CRSIP+DM Methodology. [Online] Available from: https://www.sv-europe.com/crisp-dm-methodology/ [Accessed 14 December 2017]

# Appendix

