

ASSIGNMENT

CT050-3-M-DAP

DATA ANALYTICAL PROGRAMMING

UCMP1606DSBA

HAND OUT DATE: 17 MARCH 2017

HAND IN DATE: 15 MAY 2017

**LECTURENAME: DR. KALAI ANAND RATNAM**

**STUDENT NAME: YIP HOW YUE**

**STUDENT ID: TP042904**

COURSE CODE: UCMP1606DSBA

Contents

[Introduction 3](#_Toc482628258)

[Background 3](#_Toc482628259)

[Objective 3](#_Toc482628260)

[Preliminary Semiannual Uniform Crime Report, January—June, 2015 4](#_Toc482628261)

[Missing data 5](#_Toc482628262)

[External Data Set 6](#_Toc482628263)

[USPS State Code 6](#_Toc482628264)

[State Population 2015 7](#_Toc482628265)

[Ethnicity 7](#_Toc482628266)

[Economy 8](#_Toc482628267)

[Poverty 8](#_Toc482628268)

[Unemployment 9](#_Toc482628269)

[Data Reduction, Cleansing and Transformation 10](#_Toc482628270)

[Table 4 11](#_Toc482628271)

[External Data Set 14](#_Toc482628272)

[Exporting into Comma Delimited File 16](#_Toc482628273)

[Preparing for Data Analysis using SAS 17](#_Toc482628274)

[WORK.MasterByCity 19](#_Toc482628275)

[WORK.MasterByState 21](#_Toc482628276)

[Objective 1: To visualize the population distribution across U.S based on cities reported in the UCR 23](#_Toc482628277)

[Objective 2: To visualize the total crime distribution across U.S 24](#_Toc482628278)

[Objective 3: To visualize the crime rate across United States 26](#_Toc482628279)

[Objective 4: Rank the States by Total Crime, Crime Rate, and Crime Rate Variance. 28](#_Toc482628280)

[Objective 5: Rank the States by crime rate 33](#_Toc482628281)

[Objective 6: Overall analysis on the crime reported on the California, Texas and New York 35](#_Toc482628282)

[Objective 7: Comparing crime rate variation vs population with a scatterplot 40](#_Toc482628283)

[Objective 8: Identify the correlation between the crime types 41](#_Toc482628284)

[Objective 9: Identify the correlation between crime type and economy value 43](#_Toc482628285)

[Objective 10: Identify the correlation between the total crime and external factors 46](#_Toc482628286)

[Ethnicity 46](#_Toc482628287)

[Unemployment 49](#_Toc482628288)

[Poverty 50](#_Toc482628289)

[Data Flow Diagram 52](#_Toc482628290)

[Summary and Recommendation 53](#_Toc482628291)

[Conclusion 54](#_Toc482628292)

[References 56](#_Toc482628293)

[Appendix 57](#_Toc482628294)

# Introduction

Data is touted as the future currency, especially when it comes in huge volume. Data may be worthless by itself, but the potential values and insights one can extract out of it is invaluable. The world of data analysis has evolved from descriptive analysis to predictive analysis. People no longer interested only in what had happened, but what is going to happen. To go further, business person, policy makers, politician or whoever knows the power of predictive analysis would go beyond knowing the future by wanting to what to do about it and that’s prescriptive analysis. Data analysis certainly has a role to play in fighting crime, such as how to effectively prevent crime from happening, accurately predicting the occurrence of crime, assists in hot spot policing etc.

## Background

Being a recent hire as a data scientist at the headquarters of Federal Bureau of Investigation (FBI), Washington, D.C, I am tasked to carry out a comprehensive analysis in the crime investigation department. This analysis will concentrate only in violent crime and property crime. A set of preliminary data from the Uniform Crime Reports (UCR) of the FBI for the first 6 months (January to June) of 2014 and 2015 in the cities with population greater than 100,000 has been provided. To supplement the analysis, external data that is useful and trustworthy can be obtained from credible sources. SAS Output Delivery Systems (ODS) are the expected results from the analysis for demonstration and visualization. The outcome of the analysis is to provide descriptive mean of the current state of crime and to derive any insights that are not immediately available by merely looking at the tabular data.

## Objective

The ultimate objective of the entire analysis is to help the American lives a peaceful and secure life by fighting crime from its underlying social and economic cause, preventing crime with targeted welfare spending and effective allocation of police force. To achieve the goal, we need to understand the current state of crime by extracting data for the following detail level objectives:

1. To visualize the population distribution across U.S based on cities reported in the UCR
2. To visualize the total crime distribution across U.S
3. To visualize the crime rate across U.S
4. Rank the States by Total Crime, Crime Rate, and Crime Rate Increase
5. Rank the States by crime rate
6. Overall analysis on the crime reported for the California, Texas and New York
7. Comparing crime rate variation vs population with a scatterplot
8. Identify the correlation between the crime types
9. Identify the correlation between crime type and economy value
10. Identify the correlation between the total crime and external factors:
    1. Ethnicity
    2. Unemployment
    3. Poverty

## Preliminary Semiannual Uniform Crime Report, January—June, 2015

This is data collected by FBI through the UCR and is the primary data set for the analysis. It provides a statistical view of the offenses reported law enforcement for the first six months of year 2014 and 2015. A total of 43 states and 262 cities are on the record. It contains specific type of offenses that are reported in cities having population of 100,000 and above, by state. Under the violent crime category, there are murder, rape (revised definition), rape (legacy definition), robbery and aggravated assault, whereby property crime includes burglary, larceny theft, motor vehicle theft and arson. All data collection complies with U.S national UCR Program's guidelines. Those identified otherwise were excluded, for example Thornton and Toledo numbers were determined to be noncompliance while Grand Rapids numbers were underreported.

California is the most populated state by having more than 17 million population, while Texas and New York came in second and third at 10 million and 9.4 million respectively in the year 2014.

### Missing data

The data set has high level of completeness but missing values are observed. It is important to note that year 2014 data is missing from Tuscon, Honolulu, Tyler, Provo and West Valley. The following table provide the summary of missing values:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Count** | **Remark** |
| Population | 477 | The population for year 2015 is unavailable |
| Violent crime | 5 | Even though there are missing data for violent crime and property crime, but they are the summary for the respective crime type, hence can be manually calculated when the missing number of individual crime type is fulfilled |
| Property crime | 8 |
| Murder | 5 |  |
| Rape (revised definition) | 5 | Rape is reported in revised definition and legacy definition separately and data is available only in one definition per record. That leaves the missing count to only 5. |
| Rape (legacy definition) |
| Robbery | 5 |  |
| Aggravated assault | 5 |  |
| Burglary | 6 |  |
| Larceny- theft | 6 |  |
| Motor vehicle theft | 5 |  |
| Arson | 27 | The FBI does not publish arson data unless it receives data from either the agency or the state for six months for at least one of the reporting years. |

Table . Missing Data in Table 4

There are not only data missing for metrics, some states are found to be missing from the data set or not reported. These states are:

1. Wyoming
2. Minnesota
3. Arkansas
4. West Virginia
5. Maine
6. Vermont

## External Data Set

In order to conduct more detail analysis, especially for the objective which is to identify correlations of the crime data and external factors, I look to the United States Census Bureau’s website (<https://www.census.gov>) for credible data source. The bureau being part of the U.S Department of Commerce, provides high quality and trustworthy data. The external data sets obtained from the Bureau are the USPS state code, state population 2015, ethnicity, economy, poverty and unemployment.

### USPS State Code

This the Official United States Postal Service (USPS) Code, it is used to label all the available state in Table 4 data with a corresponding state code. With the USPS state code, it enables the use of SAS built-in MAPS library to create map graph to showcase the crime data. Fields acquired are:

|  |  |
| --- | --- |
| No | Field |
| 1 | Code |
| 2 | StateName |

Table . External Source – StateCode

### State Population 2015

To fill up the none available 2015 population data for Table 4, we have obtained the 2015 states’ population estimate from the same Preliminary Semiannual UCR, but for time period January-June, 2016 (<https://ucr.fbi.gov/crime-in-the-u.s/2016/preliminary-semiannual-uniform-crime-report-januaryjune-2016>), where the population for most of the cities are available. The fields collected are:

|  |  |
| --- | --- |
| No | Field |
| 1 | State |
| 2 | City |
| 3 | Year |
| 4 | Population |

Table . External Source – Population2015

### Ethnicity

It’s a known fact that all ethnic groups play a role in defining a society. The languages, social background, cultural tradition, religion, family values and the lifestyle practiced all influence the society in some way. The data collected for the analysis focused on the origin of the resident by state and the proportion to the overall state population. The metrics collected are the estimated population fall into each group:

|  |  |
| --- | --- |
| No | Metric |
| 1 | Estimate; Total: |
| 2 | Estimate; Not Hispanic or Latino: |
| 3 | Estimate; Not Hispanic or Latino: - White alone |
| 4 | Estimate; Not Hispanic or Latino: - Black or African American alone |
| 5 | Estimate; Not Hispanic or Latino: - American Indian and Alaska Native alone |
| 6 | Estimate; Not Hispanic or Latino: - Asian alone |
| 7 | Estimate; Not Hispanic or Latino: - Native Hawaiian and Other Pacific Islander alone |
| 8 | Estimate; Not Hispanic or Latino: - Some other race alone |
| 9 | Estimate; Not Hispanic or Latino: - Two or more races: |
| 10 | Estimate; Not Hispanic or Latino: - Two or more races: - Two races including Some other race |
| 11 | Estimate; Not Hispanic or Latino: - Two or more races: - Two races excluding Some other race, and three or more races |
| 12 | Estimate; Hispanic or Latino: |
| 13 | Estimate; Hispanic or Latino: - White alone |
| 14 | Estimate; Hispanic or Latino: - Black or African American alone |
| 15 | Estimate; Hispanic or Latino: - American Indian and Alaska Native alone |
| 16 | Estimate; Hispanic or Latino: - Asian alone |
| 17 | Estimate; Hispanic or Latino: - Native Hawaiian and Other Pacific Islander alone |
| 18 | Estimate; Hispanic or Latino: - Some other race alone |
| 19 | Estimate; Hispanic or Latino: - Two or more races: |
| 20 | Estimate; Hispanic or Latino: - Two or more races: - Two races including Some other race |
| 21 | Estimate; Hispanic or Latino: - Two or more races: - Two races excluding Some other race, and three or more races |

Table . External Source – Ethnicity

### Economy

One way to identify the main economic activities is by looking at the revenue generated by North American Industry Classification System (NAIC). Some states are perceived to be of a specific dominant economic activity. For example, California is where the Silicon Valley is located may be perceived to generate the most revenue from the Information, scientific and technical services; Massachusetts, home to some of the most prestige universities is expected to generate the most from the education services. However, the data tells a different story and these can be used to supplement our crime data analysis. The metrics collected and used are:

|  |  |
| --- | --- |
| **No** | **Metric** |
| 1 | Meaning of 2012 NAICS code |
| 2 | Value of sales, shipments, receipts, revenue, or business done ($1,000) |

Table . External Source – Economy by NAIC

### Poverty

The reason this is chosen to be factor to investigate is it can lead to property crime. There are states with high number of population falls under the poverty line determined by the American Community Survey 5-year estimates. The metrics collected and used are the population falls under in the group:

|  |  |
| --- | --- |
| **No** | **Metric** |
| 1 | Estimate; Income in the past 12 months below poverty level |
| 2 | Estimate: Income in the past 12months at or above poverty level |

Table . External Source – Poverty

### Unemployment

Unemployment is the first instinctive cause for crime rate. Drilling down to the unemployment rate by state by age group provides a more targeted comparison against the total crime. The metrics collected and used are:

|  |  |
| --- | --- |
| **No** | **Metric** |
| 1 | Unemployment rate; Estimate; Population 16 years and over | |
| 2 | Unemployment rate; Estimate; AGE - 16 to 19 years | |
| 3 | Unemployment rate; Estimate; AGE - 20 to 24 years | |
| 4 | Unemployment rate; Estimate; AGE - 25 to 29 years | |
| 5 | Unemployment rate; Estimate; AGE - 30 to 34 years | |
| 6 | Unemployment rate; Estimate; AGE - 35 to 44 years | |
| 7 | Unemployment rate; Estimate; AGE - 45 to 54 years | |
| 8 | Unemployment rate; Estimate; AGE - 55 to 59 years | |
| 9 | Unemployment rate; Estimate; AGE - 60 to 64 years | |
| 10 | Unemployment rate; Estimate; AGE - 65 to 74 years | |
| 11 | Unemployment rate; Estimate; AGE - 75 years and over | |
| 12 | Unemployment rate; Estimate; RACE AND HISPANIC OR LATINO ORIGIN - White alone | |
| 13 | Unemployment rate; Estimate; RACE AND HISPANIC OR LATINO ORIGIN - Black or African American alone | |
| 14 | Unemployment rate; Estimate; RACE AND HISPANIC OR LATINO ORIGIN - American Indian and Alaska Native alone | |
| 15 | Unemployment rate; Estimate; RACE AND HISPANIC OR LATINO ORIGIN - Asian alone | |
| 16 | Unemployment rate; Estimate; RACE AND HISPANIC OR LATINO ORIGIN - Native Hawaiian and Other Pacific Islander alone | |
| 17 | Unemployment rate; Estimate; RACE AND HISPANIC OR LATINO ORIGIN - Some other race alone | |
| 18 | Unemployment rate; Estimate; RACE AND HISPANIC OR LATINO ORIGIN - Two or more races | |
| 19 | Unemployment rate; Estimate; Hispanic or Latino origin (of any race) | |
| 20 | Unemployment rate; Estimate; White alone, not Hispanic or Latino | |

Table . External Source – Unemployment

# Data Reduction, Cleansing and Transformation

Data cleansing is fundamental to elevate the quality of the data, which has a direct impact to the quality of the analysis outcome. There will rarely be a source data sets that is ready for immediate mining. As Robert Musson, Microsoft Principle Data Scientist, has pointed out a great deal of his time is spent on cleaning data and getting data ready to be trustworthy before it is usable for any data analysis (Musson, 2016). From data warehouse’s Extract Transform Load (ETL) perspective, this is the process to fix dirty data while maintaining an accurate picture of the data as it was captured and striking a balance between these conflicting goals (Kimball & Ross, 2013). The cleansing steps are inclusive of removing irrelevant attributes, filling up missing data, eliminating noise and transforming the source into a clean and consistent data set for statistical analysis. As for our crime data analysis, all data in the Table 4 data set will be maintained, including the outlier as it is a true reflection of the data despite the huge variance of population between the most and least population state.

For this crime data analysis, It is decided to perform the initial data cleansing using R Studio. Prior to importing the data, 4 user-defined reusable function have been created to reduce the coding complexity:

* fCreateDFPopulation – This function is designed around the Table 4 structure to extract population data from Table 4. Return a data frame.
* fConvertDFVariabletoNumeric – This is to convert data frame variable (field) into numeric type. It is worth noting that multiple variables can be passed at one execution. Return a data frame.
* fUpdate2014ValueByPerc – This is to apply a percentage against a variable and returns the calculated result in data frame.
* fUpdate2015ValueByPerc – Like fUpdate2014ValueByPerc but this a reversed version.

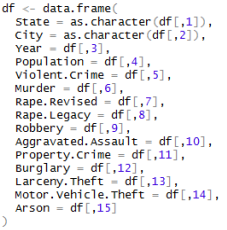
## Table 4

After removing header and footer manually using Microsoft Excel. The data cleansing steps for Table 4.csv are:

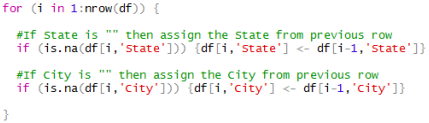
1. Import as data frame. read.csv() command is used and instructed R to treat empty string “”, space “ “ and “NA” as NULL value:



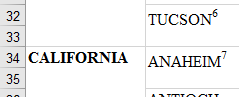
1. Reconstruct the data frame. This is to overwrite the default data type and field name:



1. Fill the missing values for State and City. A loop is used to iterate through every row and update the empty State and City field with the value from the preceding row:



1. Remove footnote number from column State and City. The footnote reference number is carried over from the original data set in xlsx format. For example,



The following code eliminate the footnote number:



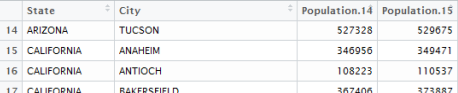
1. Convert variables into numeric datatype. Due to the nature of R programming, values are sometimes defaulted into Factor type in R, hence some variables are required to be explicitly converted into numeric using the User-defined function fConvertDFVariabletoNumeric:



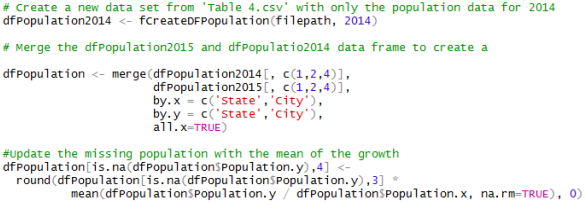
1. Convert NA value into 0. This is applied against Population, Rape.Revised and Rape.Legacy field:



1. Fill up the none available population for 2015. A temporarily data frame with population for 2014 and 2015 pivoted into columns like the following is required:



* 1. Create the temporarily data frame dfPopulation2014 with user-defined function fCreateDFPopulation and fill up the missing Population 2015 by applying a mean growth from population 2014:



* 1. Update the imported data frame df with Population 2015 from the temporarily data frame dfPopulation:



1. Fill the missing values for violent crime (Murder, Robbery, Aggravated Assault) by adding 1.7% against the 2014 total using a user-defined function fUpdate2014ValueByPerc:

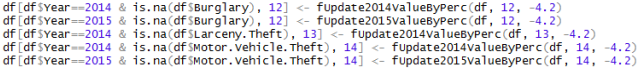


Here’s an example of Robbery for TUCSON before and after. 507+1.7% = 516:

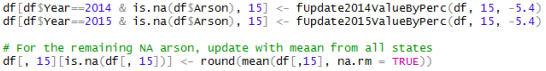




1. Fill the missing values for property crime (Burglary, Larceny Theft, Motor Vehicle Theft) by adding -4.2% against the total of 2014 using a user-defined function fUpdate2014ValueByPerc or fUpdate2014ValueByPerc (The reversed method):



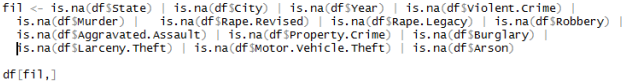
1. Fill the missing values for Arson by adding -5.4% against the total of 2014 using the user-defined function. For the remaining missing value, which are due to no values have been reported for both 2014 and 2015, apply a mean update:



1. Fill up the Violent.Crime and Property.Crime value by recalculating the sum of Murder, Rape.Revised, Rape.Legacy, Robbery, Aggravated.Assault for Violent.Crime and Burglary, Larceny.Theft, Motor.Vehicle.Theft for Property.Crime respectively:

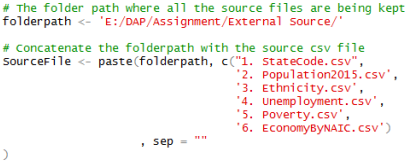


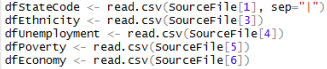
1. At last, verify if there are any missing value (NA) for any of the fields:



## External Data Set

The data acquisition and cleansing steps for external data set are as following:

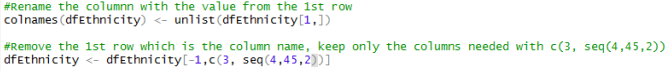
1. Define the source files location: 
2. Import as data frame. There are two different approaches:
   1. read.csv() command is used for the import by reading the file location defined in step 1. Delimiter is specified as pipe ‘|’ for StateCode.csv, while other uses the default comma ‘,’:



* 1. User-defined function fCreateDFPopulation. For Population2015.csv, similar to Table 4.csv, header and footer were manually removed prior to the import:



1. Replace the default header. This is because data sets downloaded from U.S Census Bureau has two line of header, where only the second line is descriptive. To replace the default header with 2nd header coming in as data record, unlist() function is used to create a vector from the 2nd data record before replacing the variable name with the vector:



1. Convert variables into numeric datatype. Due to the nature of R programming, values are sometimes defaulted into Factor type in R, hence some variables are required to explicitly convert into numeric using a User-defined function fConvertDFVariabletoNumeric:



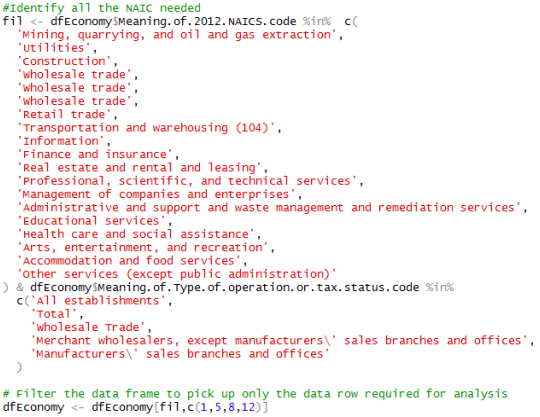
1. Convert total value into percentage. There are some metrics stored as total, in order to provide a fair comparison against other states they need to be converted into percentage:



1. Eliminate irrelevant fields. Unnecessary variables are being removed:

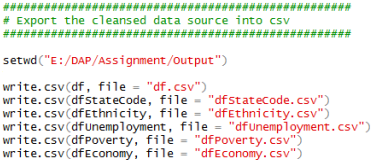


1. Eliminate irrelevant records. Special care is given to the Economy.csv data set, where we remove not only the unnecessary fields, but also unnecessary record to produce a smaller data set:



## Exporting into Comma Delimited File

After all the data cleansing (Table 4) and transformation (external data sets) in R studio, they are ready to be exported into comma delimited files using the write.csv() command:



# Preparing for Data Analysis using SAS

The actual data analysis will be conducted using SAS® OnDemand for Academics release 3.6 (Enterprise). Output Delivery System (ODS) statistical graphs and tables will be produced as part of the output to further support the interpretation.

The csv files produced from R Studio will first be uploaded and import into new SAS library “DAP” using 00.Import Data.sas (https://github.com/howyue/hello-world/blob/master/00.Import%20Data.sas):

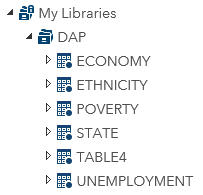


Figure . User defined library in SAS

During the process, labels are created for all variables which the length of the source data set column name is exceeding the permitted length (32 characters) in SAS. For example, DAP.UNEMPLOYMENT.var3 is been labelled “Unemployment rate; Estimate; Population 16 years and over”:

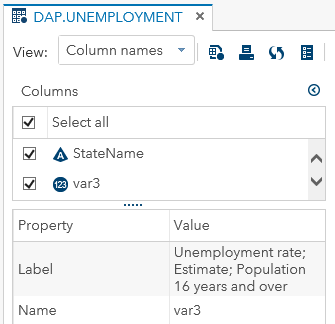


Figure 2. Variable Label

Rape.Revised and Rape.Legacy are dropped from DAP.TABLE4 but two new variables are added:

1. Rape – A sum of Rape.Revised and Rape.Legacy
2. TotalCrime – A sum of Violent.Crime and Property.Crime, which represents the total of all crime reported in Table 4.

Besides, to make the analysis smoother and easier for visual graphs creation, we have derived the following SAS tables into the temporarily SAS library “WORK” using 01.Derive Data.sas(https://github.com/howyue/hello-world/blob/master/01.Derive%20Data.sas) without tampering the data:

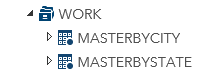


Figure . User defined temporarily library

## WORK.MasterByCity

This is the transposed (unpivot) version of the DAP.TABLE4, which vertically convert variables into observations:

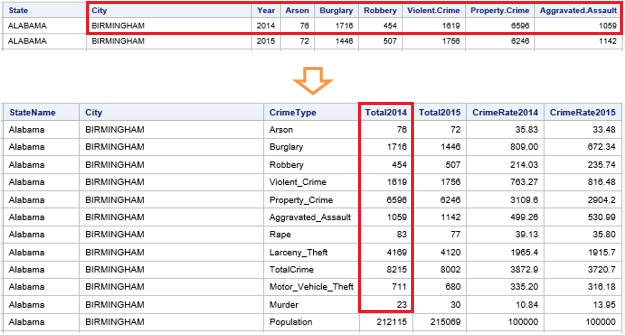


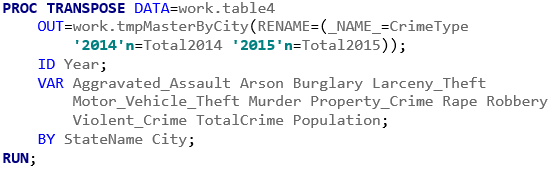
Figure 4. Illustration of Tranpose

DAP.TABLE4 need to go through a few steps before it is transformed into the desired data set above:

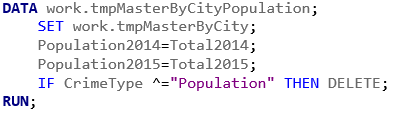
1. PROC SORT – A mandatory step to sort the DAP.TABLE by State, City and Year. The subsequent PROC TRANSPOSE step relies on a sorted data set.



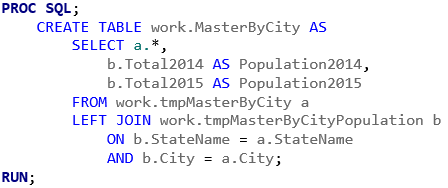
1. PROC TRANSPOSE – The actual transpose happens in this step.



1. DATA – A temporarily data set work.tmpMasterByCityPopulation is to be created. It will be used in the subsequent PROC SQL.

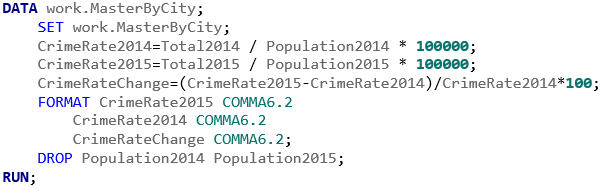


1. PROC SQL – To perform a SQL’s LEFT JOIN to append Population2014 and Population2015 to WORK.MasterByCity as new variables



1. DATA – The final step introduces the following variables after the CrimeRate calculation:
   1. CrimeRate2014 and CrimeRate2015, which are calculated by dividing the number of reported crimes by the total population; the result is multiplied by 100,000.
   2. CrimeRateChange is the difference of CrimeRate between 2014 and 2015 in percentage.

The new variables are FORMAT to COMMA6.2



## WORK.MasterByState

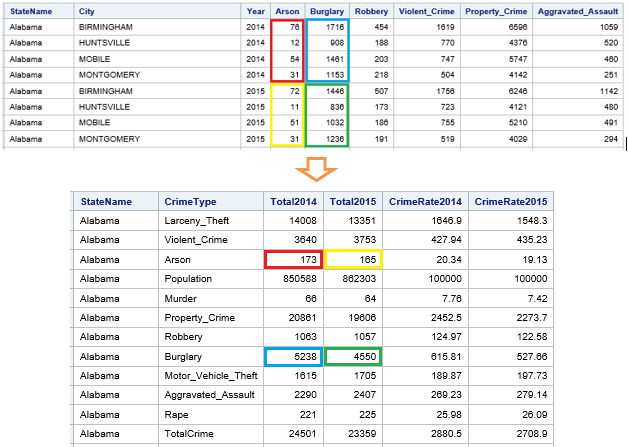
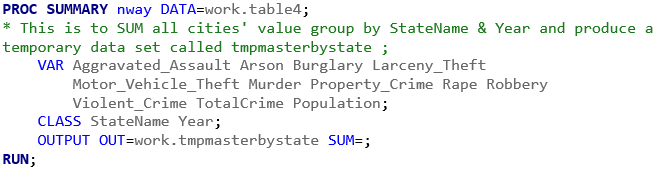
This is a similar to WORK.MasterByCity, but City is removed and grouped by State. All total is summed to state level. For example, the total of Arson for Alabama in year 2014 is 173 (highlighted in RED below):

Figure . Summarizing Total by State

Before the same transformation process that creates WORK.MasterByCity are applied to derive WORK.MasterByState, there is the PROC SUMMARY step applied against DAP.TABLE4

1. PROC SUMMARY – This step sums up all variables defined in the VAR statement and group by the variables defined in CLASS statement. 
2. The same steps 1 to 5 apply to WORK.MasterByCity are to be repeated for WORK.MasterByState.

# Objective 1: To visualize the population distribution across U.S based on cities reported in the UCR

We know for a matter of fact that the crime is closely related to the population density. The more population it is the higher the crime. First objective is to visualize the population distribution. It is very important to note that the population shown in the map below is based only on the cities report in the UCR and NOT the actual population for the state. The reason why we’re not taking the population of the state is because we’re having only subset of the crime data based on the cities with population 100,000 instead of the entire state data.

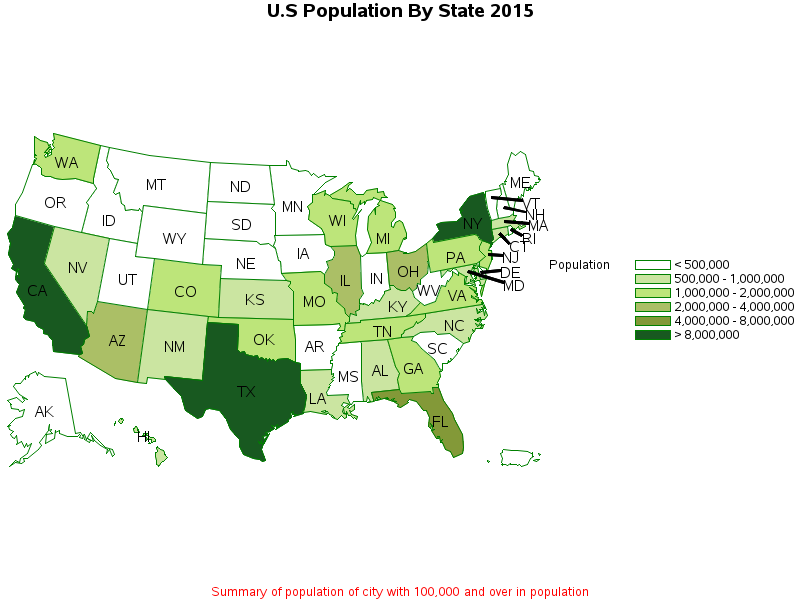


Figure . U.S Population By State 2015

We can see that California (CA), Texas (TX) and New York (NY) are the states with the most population. They are also the states recorded highest number of crime as evident by Objective 2 below.

# Objective 2: To visualize the total crime distribution across U.S

The data provided in tabular format is difficult to picture the status of crime in U.S. Therefore, this objective is to look at how the crime is distributed across the country from a holistic view. PROC GMAP is used to plot the total crime reported for the first 6 months (January-June) between 2014 and 2015 into a U.S map.

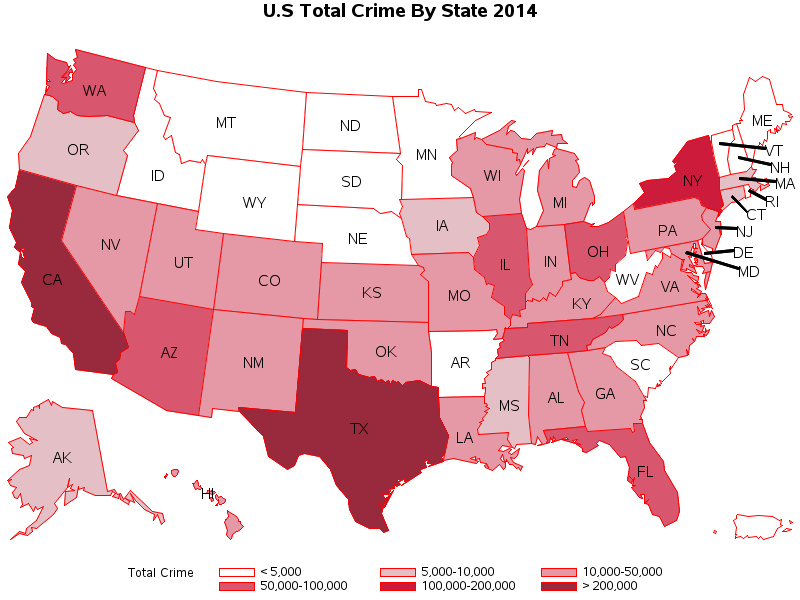


Figure . U.S Total Crime By State 2014

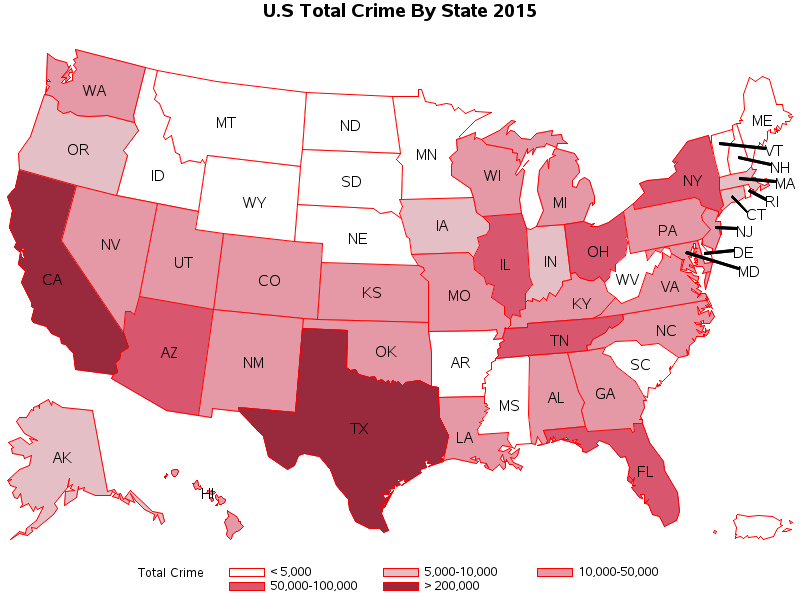


Figure . U.S Total Crime By State 2015

From the Fig 7 & Fig.8 above, beside the 6 states (WY, MN, AR, WV, ME, VT) we have identified to be missing from data set, there are 8 states reported lesser than 5000 crime in total for year 2014:

1. Montana (MT)
2. Idaho (ID)
3. North Dakota (ND)
4. South Dakota (SD)
5. Nebraska (NE)
6. Mississippi (MS)
7. New Hampshire NH)
8. Rhode Island (RI)

The state with total crime < 5000 increased by 1 in year 2015 with South Carolina.

Overall crime is observed to be on reducing trend that the states like New York (NY), Mississippi (MS) and Washington (WA) upgraded from the category of 100k – 200k to 20k - 100k, 5k – 10k to < 5k and 50k – 100k to 10k – 50k respectively. Coincidently there’s no state falls into the category of 100k – 200k in year 2015.

# Objective 3: To visualize the crime rate across United States

If we look at only the total crime reported and neglect the population factor, it will portray a biased view of the states’ security. A more rational assessment would be to use the Crime Rate. A Crime Rate is the number of crimes reported per 100,000 total population. This will provide a reasonable comparison between states.

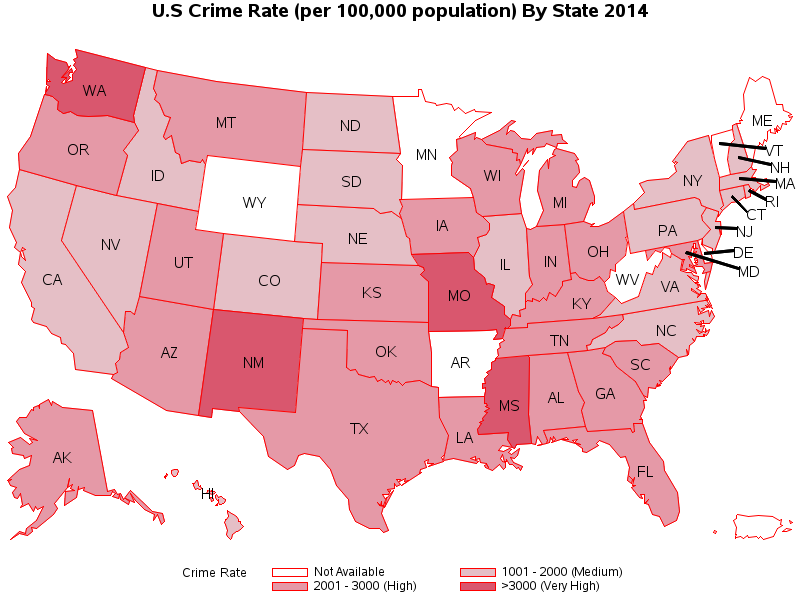


Figure . U.S Crime Rate By State 2014

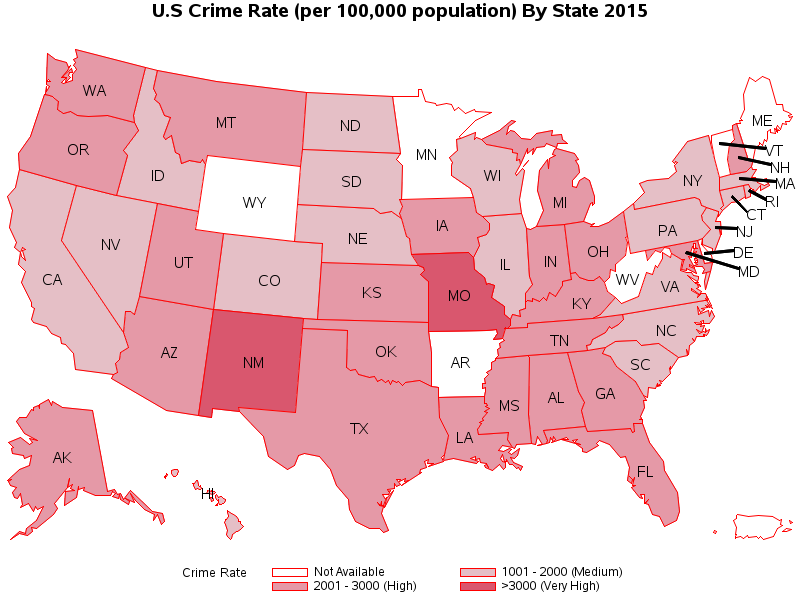


Figure . U.S Crime Rate By State 2015

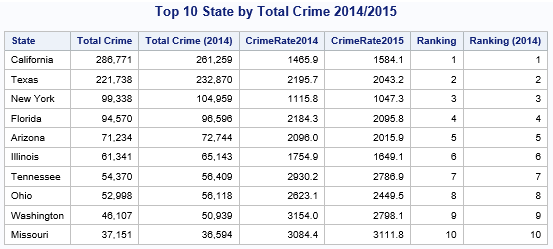
We can see that there’s minimal visible change of crime rate between 2014 and 2015. A closer look find out that Washington (WA) and Mississippi (MS) have improved from the very high to high, whereas New Hampshire worsening.

When the Crime Rate is looked at along with the Total Crime, it is evident that despite California (CA), Texas (TX) and New York (NY) being top crime state, none of them falls into the category of > 3000 (Very High) category.

# Objective 4: Rank the States by Total Crime, Crime Rate, and Crime Rate Variance.

After looking at the overall status of the crime across all states, it’s imperative that we are to identify the extreme states:

1. Top 10 State by Total Crime 2014/2015:



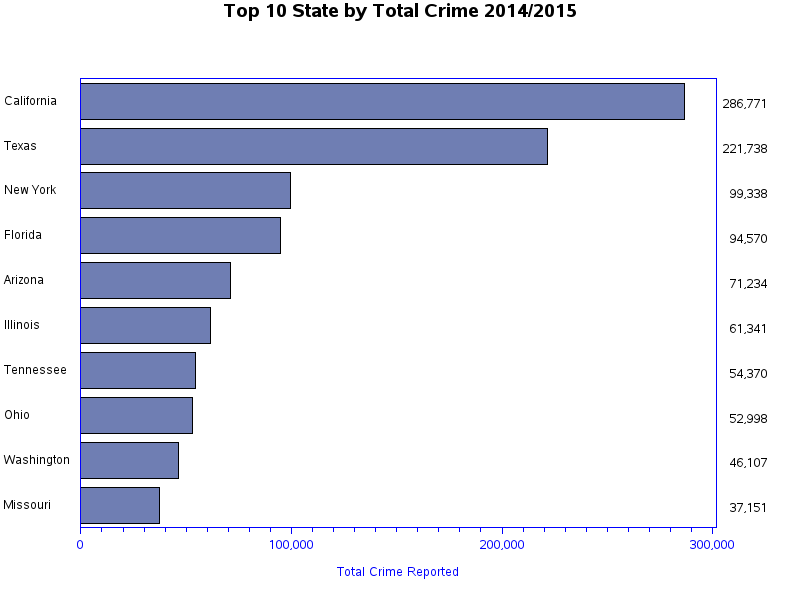
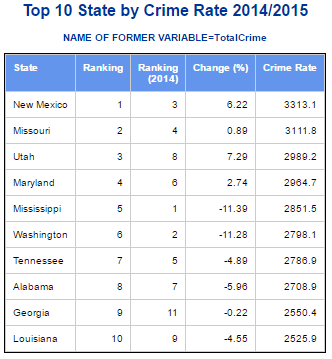


Figure . Top 10 State By Total Crime 2015/2015

It is interesting to find out that the top 10 states have not changed since 2014, as these are the state with the most population and hence the most crime reported.

1. Top 10 State by Crime Rate 2014/2015:



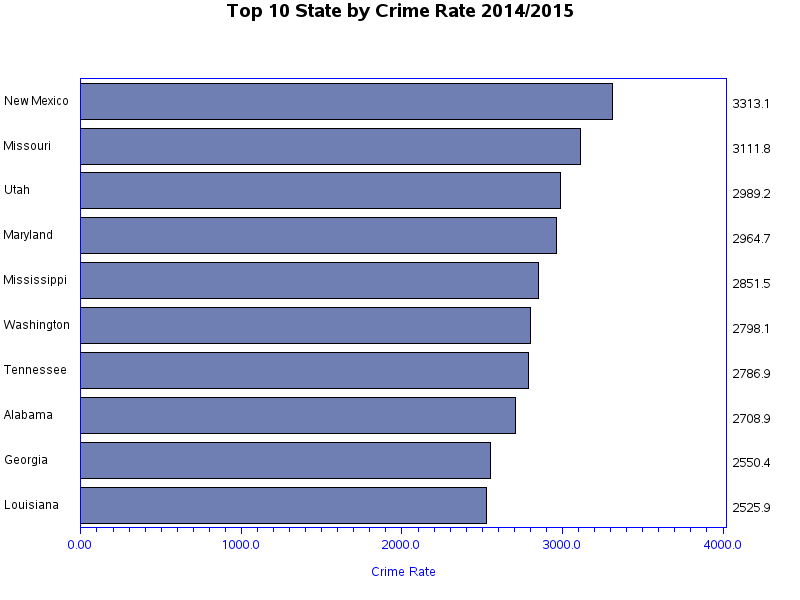
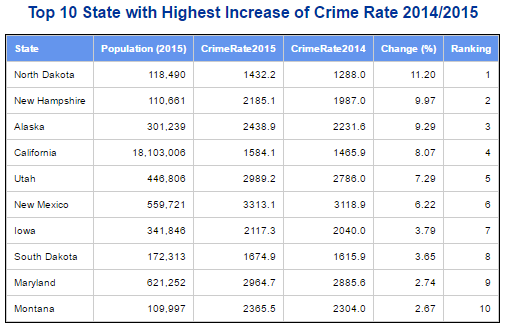


Figure . Top 10 State By Crime Rate 2015/2015

As pointed out in Objective 2, the population factor cannot be neglected. The following chart reinforced the argument that the states with the most reported crime are unnecessary the states with the highest crime rate. Only Washington and Missouri, which are among the top 10 state by total crime also the among the top 10 state by crime rate. It is encouraging that both states reported double figure reduction.

1. Top 10 State with Highest Increase of Crime Rate 2014/2015:



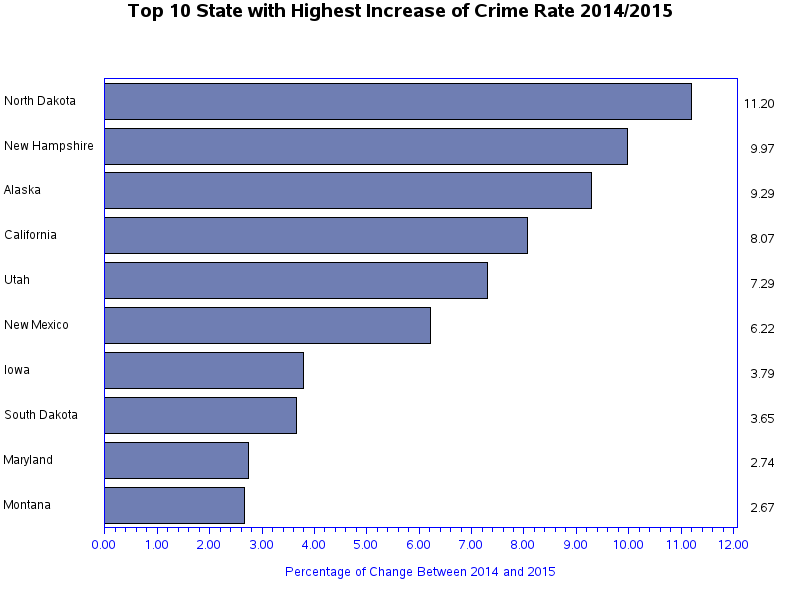


Figure . Top 10 State with Highest Increase of Crime Rate 2014/2015

It is alarming that North Dakota has recorded double figure (11.2%) increase and New Hampshire has a close to 10% of increase since 2014. California is the only state with a population higher than 1,000,000 and recorded a 8.07% increase.

# Objective 5: Rank the States by crime rate

For this objective, we delved deeper into finding out which are the most dangerous state by crime type using PROC RANK. There can be various reasons why a state rank high for one type of crime but doing fine for the rest, it is crucial to not generalize the number so the law enforcement can focus on reducing only the targeted crime types by allocating specialized officers.

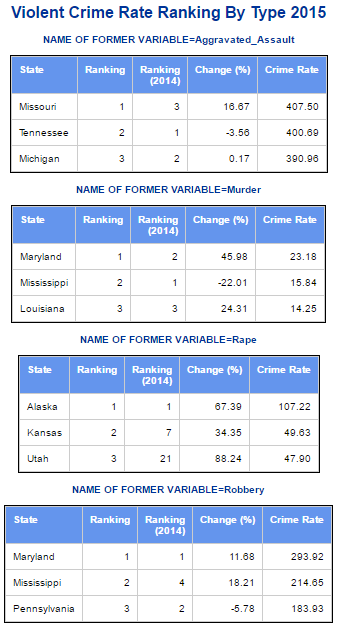
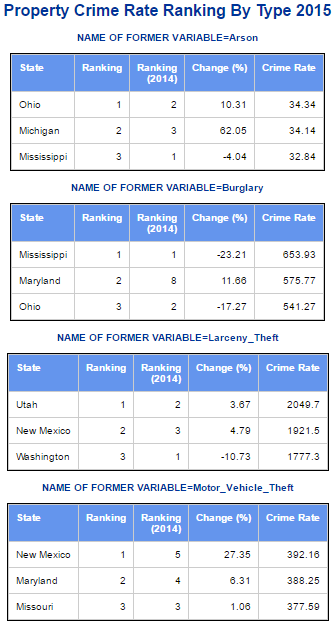
 

Figure . Crime Rate Ranking by Type 2015

Maryland and Mississippi are both among the top three for 4 different crime types. Maryland though is a small state top the ranking in Murder and Robbery, and ranked 2nd in Burglary and Motor Vehicle Theft. Mississippi is improving from 2014 as indicated in the crime rate change but still top Burglary (Reduced by 23%) and ranked 2nd for both Murder (Crime Rate reduced by 22%) and Robbery (Increased by 18.21%) while also ranked 3rd for Arson (Reduce by 4%).

# Objective 6: Overall analysis on the crime reported on the California, Texas and New York

To provide an easier comparison for the top 3 states with the most crime reported, a bar chart is produced with 2 side by side vertical bars group by state representing year 2015 and 2014. We can see from the bar chart below where the increase for California is greater than New York and Texas combined:

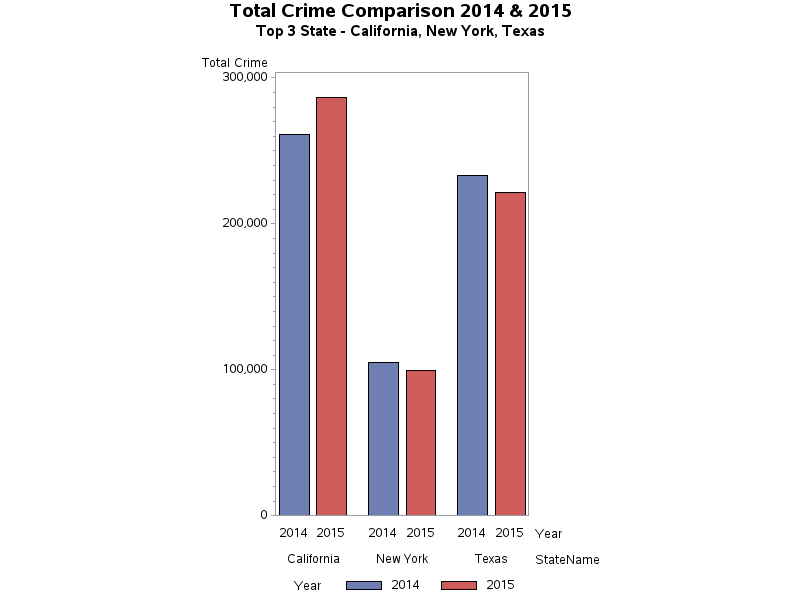


Figure . Total Crime Comparison 2014 & 2015

There are total 102 cities data for the 3 states with the most crime case, California, New York, and Texas. Comparing state level average, min and max is not representative as the population variance is huge, New York have 8.5 million while Tyler have only 100k population. To address this imbalance, a new variable population category is created and the mean, min and max are produced by this new population category.

The following SAS ODS detailed the average, minimum number of case and maximum number of case for overall crime report. The increase of total crime reported in California has overshadow the decreased in the cities for New York and Texas. Out of all city sizes, only city with population between 1 million and 2.5 million and city with more than 5 million have recorded reduction in average.

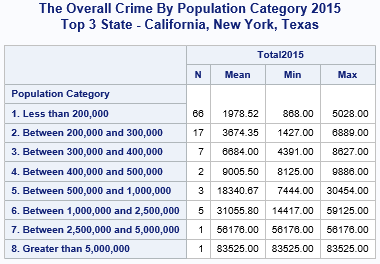
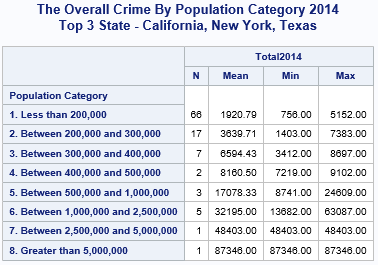
 

Figure . Overall Crime BY Population Category 2015 & 2015

It is important to know the proportion of different crime types. A pie chart is the best graph to meet such requirement. Each state is produced with its own pie chart:

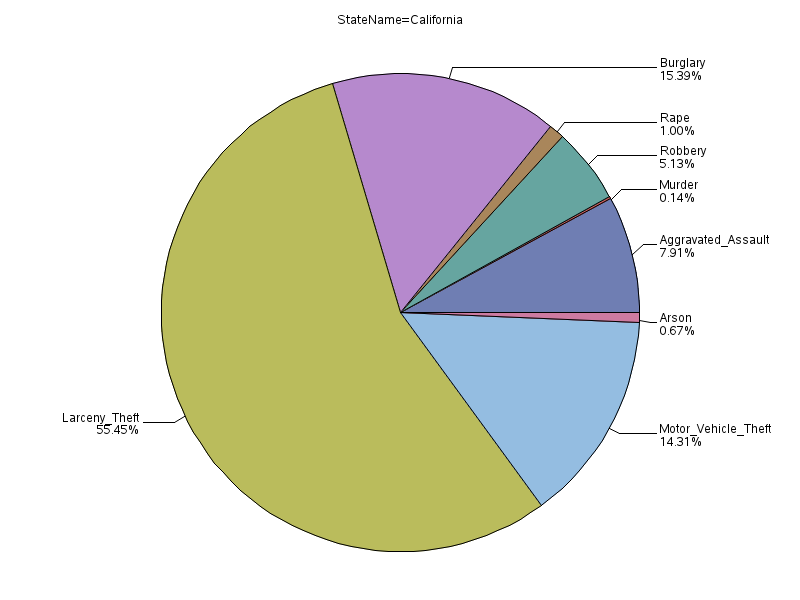


Figure . Proportion of Crime - California

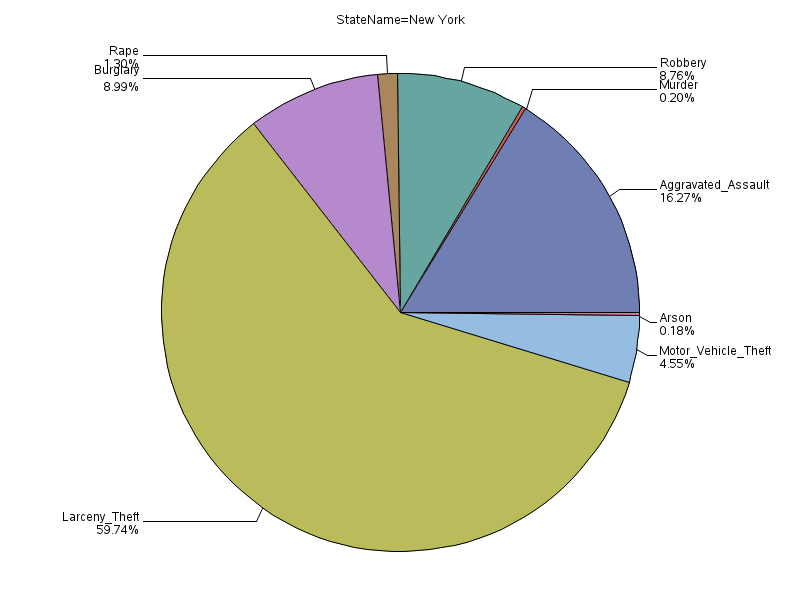


Figure . Proportion of Crime – New York

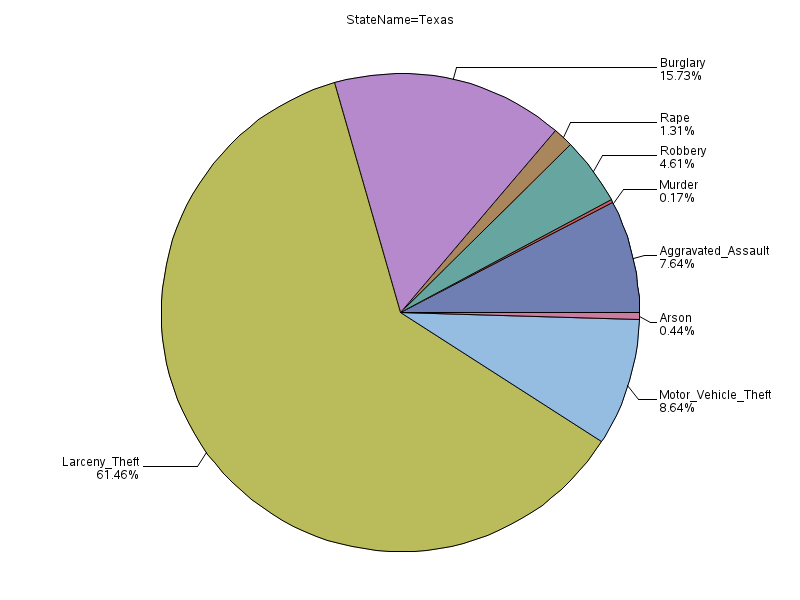


Figure . Proportion of Crime - Texas

Larceny theft is the most common crime, follow by burglary for California and Texas. New York has the highest percentage for aggravated assault 16.27% as its 2nd most committed crime, which is double of both California and Texas. Evidently, New York is the most violent city that it has recorded the highest percentage of violent crime at 26.54%, more than a quarter of the reported crime involves threaten force.

# Objective 7: Comparing crime rate variation vs population with a scatterplot

We want to find out if there’s a relationship between population and variation of crime rate between 2014/2015. Based on the assumption that the total workforce of police remains static between 2014 and 2015. The following bubble chart created using PROC GPLOT have the change percentage as y-axis and population as x-axis, while the size of the bubble represents the total crime reported in 2015.

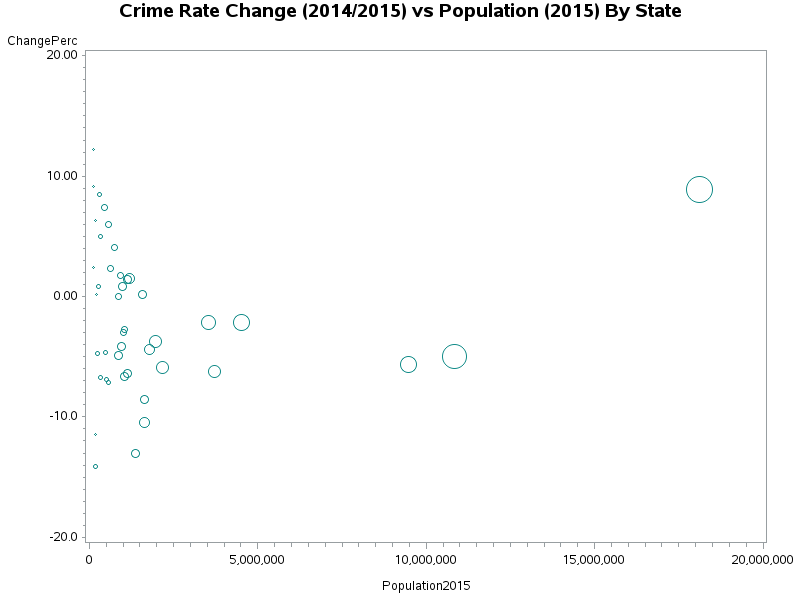


Figure . Crime Rate Change vs Population

There is indication that population size has positive correlation to the crime rate increase. It is however starts only from state with population of 1,000,000 million and becoming more obvious as the population increase. This is evident by the visual, as we can see the smaller states are congregated to the left and the variation range between -14% to 12%.

# Objective 8: Identify the correlation between the crime types

To further understand the different crime type and the reporting figures, we need to look at some of the simple statistics. This is done by PROC CORR procedure. In this objective, we not only want to investigate the simple statistics but also to identify the correlation among all the crime type. This may be able to aid law enforcement in fighting crime by reducing a crime type that has a strong positive correlation to another crime.

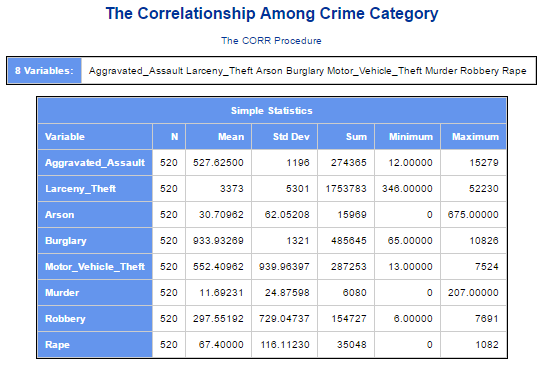


Figure . Simple Statistics

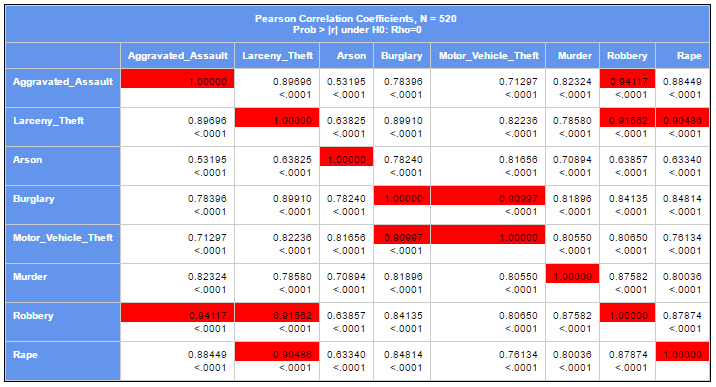


Figure . Correlation Among Crime Type

From the table above, highlighted in RED the strongest positive correlations all above 0.9 with *p*-Value lower than 0.001:

* Robbery & Aggravated Assault – 0.94117
* Robbery & Larceny Theft – 0.91562
* Motor Vehicle Theft & Burglary – 0.90997

# Objective 9: Identify the correlation between crime type and economy value

This objective is to discover the noteworthy correlation between all the crime types and the handpicked (See data cleansing part) economy value generated by the state government classified by NAIC. SAS is instructed to perform PROC CORR to produce Pearson product-moment correlations into a ODS, which highlight the correlation that is greater than 0.8:

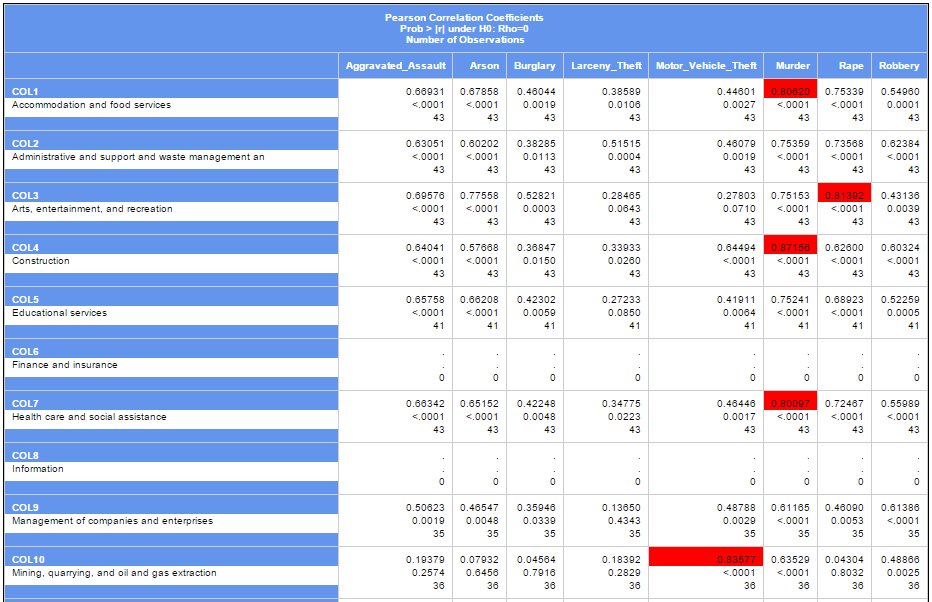


Figure . Correlation between Crime and Economy

The following correlations are identified to be significant and has p-Value < 0.0001:

* Motor Vehicle Theft & Mining, quarrying, and oil and gas extraction - 0.83577
* Murder & Construction - 0.87156
* Rape & Arts, entertainment, and recreation – 0.81392

Murder is discovered to be correlated to many industry, such as Construction, Health care and social assistance, Retail, Transportation and warehousing and etc. The wholesale trades were disregarded intentionally as it is a general industry. They may happen by chance but further investigation need to be done.

These correlations are visualized into ODS with scatterplots:

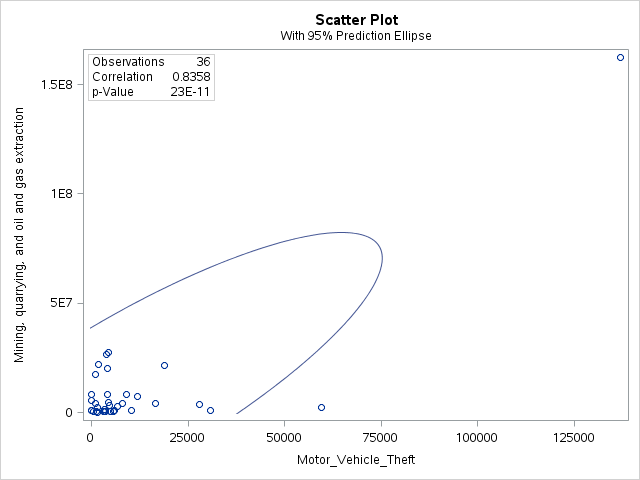


Figure . Scatter plot for Correlation Between Mining and Motor Vehicle Theft

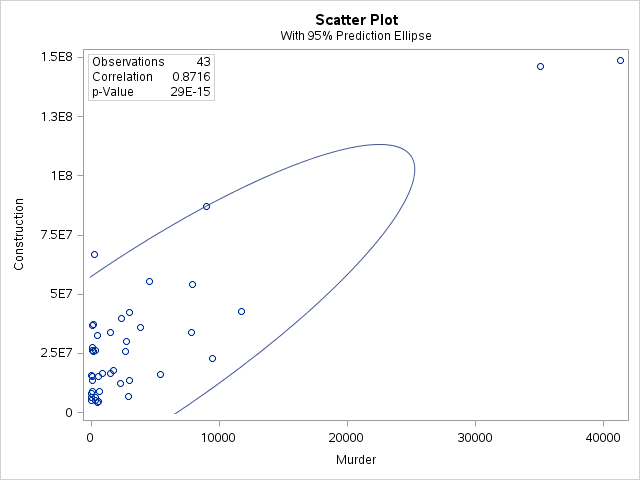


Figure . Scatter Plot for Correlation Between Construction and Murder

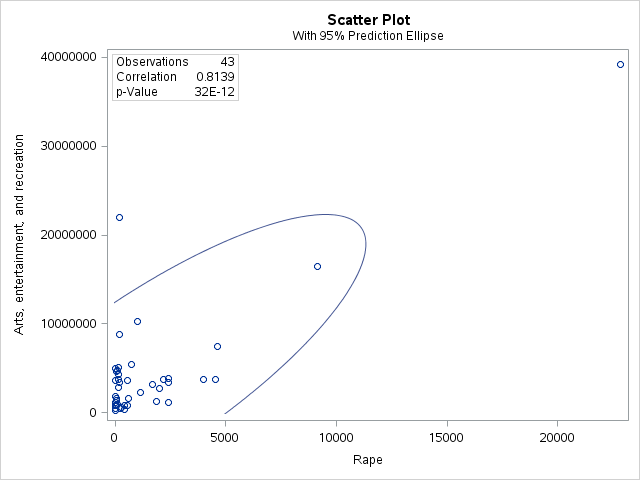


Figure . Scatter Plot for Correlation Between Art, Entertainment and Recreation and Rape

# Objective 10: Identify the correlation between the total crime and external factors

The last objective is to identify the correlation between the total crime and the external factors which we have obtained from U.S Census Bureau. This is achieved in two steps:

1. Using PROC CORR and PROC TEMPLATE to output the result into a formatted ODS.
2. Perform another PROC CORR on selected correlation from the output ODS, where correlation with a p-Value lesser than 0.05 is highlighted in red.

## Ethnicity



Figure . Correlation Between Total Crime and Ethnicity

From the result, it is discovered that the Non-Hispanic (Estimate; Not Hispanic or Latino: - White alone) population has a moderate negative correlation to the total crime while Hispanic population (Estimate; Hispanic or Latino: - White alone, Estimate; Hispanic or Latino: - Some other race alone,   
Estimate; Hispanic or Latino: - Two or more races: - Two races including Some other race) has moderate to strong correlation to the total crime.

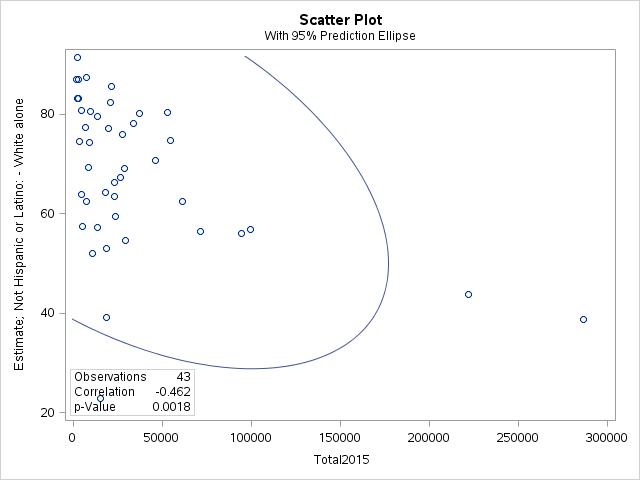


Figure . Scatter Plot for Correlation Between Total Crime and Ethnicity 1

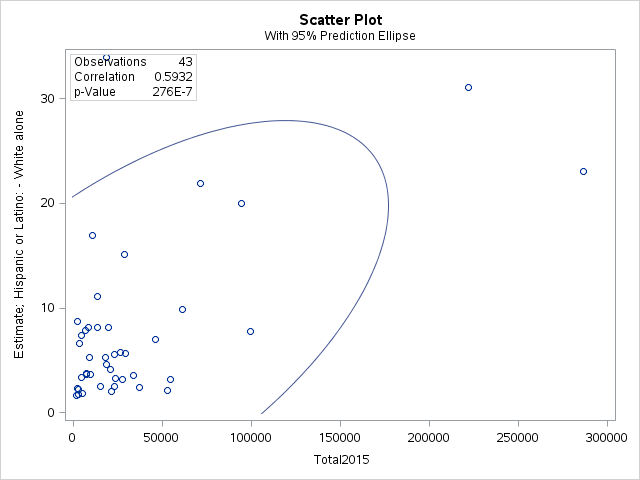


Figure . Scatter Plot for Correlation Between Total Crime and Ethnicity 2

## Unemployment

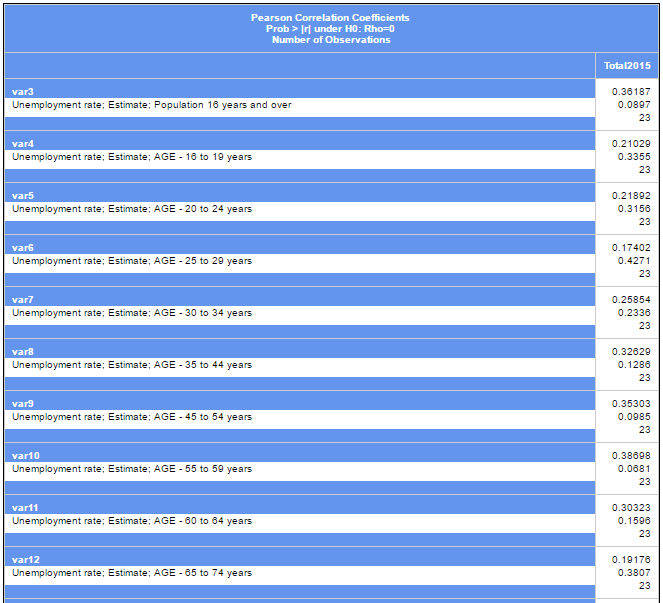


Figure . Correlation Between Total Crime and Unemployment Rate

It is surprising to find out that there’s no strong correlation between crime data and our unemployment rate data set. Further study can be done on the unemployment and crime. The noticeable correlation is the unemployment rate for Hispanic (Unemployment rate; Estimate; RACE AND HISPANIC OR LATINO ORIGIN - White alone):

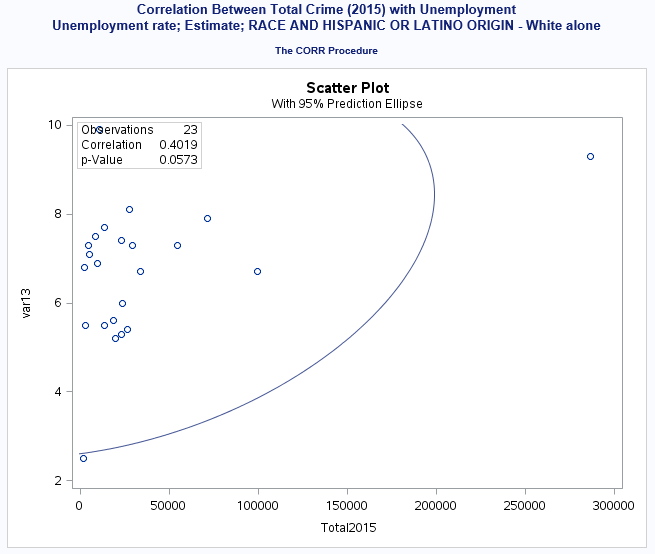


Figure . Scatter Plot for Correlation Between Total Crime and Unemployment

## Poverty

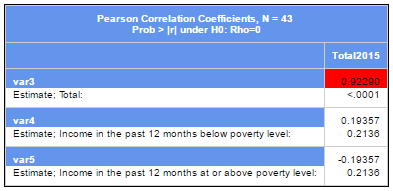


Figure . Correlation Between Total Crime and Poverty

It is intuitive that poverty is the cause of crime and this is evident by our poverty data set that it has a very strong correlation (0.9229) and low p-Value ( <0.001):

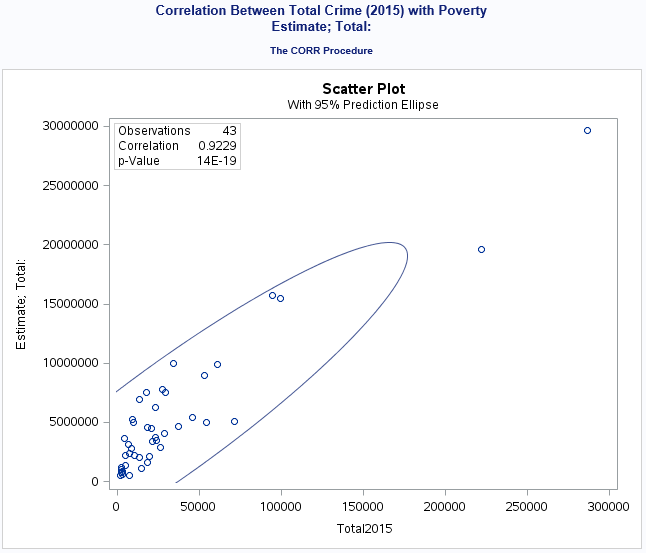


Figure . Scatter Plot for Correlation Between Total Crime and Poverty

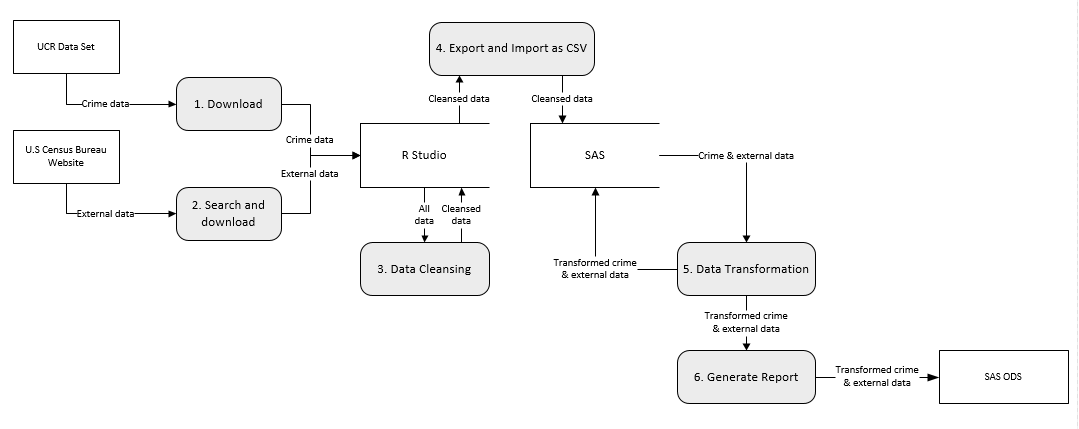
Data Flow Diagram

Figure . Data Flow Diagram for the Crime Analysis

# Summary and Recommendation

From objective 1, we understand that American generally lives near the coastal area spanning from the east coast down to the Gulf of Mexico and to the west coast. As illustrated in the map produced for objective 3, high crime rate is reported in the southern region close to the border of Mexico.

The following table show the top state by different measures elaborated in objective 4:

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Highest Number of Crime** | **Highest Crime Rate** | **Highest Crime Rate Increase** |
| 1 | California | New Mexico | North Dakota |
| 2 | Texas | Missouri | New Hampshire |
| 3 | New York | Utah | Alaska |

Table . Summary of Analysis

The states reported highest number of crime are not the most dangerous city, it is merely due to its population density. Focus should be given to states with the highest crime rate, in New Mexico, there are 3313 crime report in every 100,000 population in year 2015. North Dakota is the most alarming state that the public safety is deteriorating fast with a double figure increase of crime rate.

As pointed out in objective 5, a high murder rate is reported in Maryland with a 45% increase from 2014. Murder is traditionally a hate or organized crime. Based on our finding in objective 8, construction industry has a positive correlation with murder. Perhaps law enforcement need to investigate what contributes to the rise from the construction perspective and if there are expansions or establishments of crime syndicate. Rape case in Alaska, Kansas and Utah is worrying as all top 3 states reported whopping 67%, 34% and 88% increase respectively.

With the scatterplot generated for objective 6, number of population plays a vital role in the crime rate variation between 2014 and 2015. However, the increase of crime is only notable with cities having population 1 million onwards. A close look at the detail crime type in objective 7 indicate that robbery and aggravated assault are closely related, whereas burglary happens with motor vehicle theft.

The correlation analysis conducted in objective 8 and 9, between external factors (economy, ethnicity, poverty, and unemployment) and crime discovers that the higher the proportion of Hispanic population the higher the crime. Pair that with unemployment rate among the Hispanic population leads to higher crime rate. For unemployment rate alone, based on our dataset, there’s no decisive correlation to crime. The only noticeable correlation is when Hispanic population has high unemployment rate. Poverty is always being perceived to be one of the reason for crime, our analysis confirmed that by identifying a strong correlation.

With the above finding, the policy maker can take a closer look at what can be done to reduce the unemployment rate particularly for Hispanic population. Poverty is the perpetual social problem and not only negatively impact the public safety but also the social wellbeing. An improved and more targeted social safety net policy is essential for the states like Mississippi, New Mexico, Louisiana, etc.

All in all, the law enforcement agency need to adopt a data-driven and innovative approach in fighting crime. Mobilizing police according to their expertise and experiences and hot-spot patrolling are the basics. Predicting the location of future crime like the one implemented by Richmond, Virginia police department and Santa Cruz, California is the next generation of solution (Siegel, 2016).

# Conclusion

The source UCR data is not ready for immediate analysis, there are not only missing data but also there are states not accounted for. The structure of the UCR data is also difficult for visualization, deriving data sets need to be created with transposed data.

SAS offers a wide range of analysis features such as PROC FREQ, PROC SUMMARY, PROC CORR, PROC RANK, for vibrant and customizable visual, there are PROC GMAP, PROC FORMAT, PROC GCHART, PROC TABULATE, PROC GPLOT. All these features and SAS ODS helped in the crime analysis.

Results of the analysis discovered some important facts. Besides finding out the top states by various measures, it also uncovered information such as the alarming change in the crime rate for Maryland, Alaska and Michigan’s on Murder, Rape, and Arson respectively due to the drastic increase; Crimes in metropolitan with population greater than 5 million are reducing as compare to smaller city; Larceny theft is accounted for more than 50% of the overall and by far the most committed offense.

The correlation analysis of crime against ethnicity and unemployment rate both led to the finding that Hispanic population influences the crime. Due to the scope of the analysis, the strong correlations discovered are not conclusive. It only signifies more research and investigation is required for future study.

# References

Kimball, R. & Ross, M., 2013. *The Data Warehouse Toolkit.* 3rd ed. s.l.:John Wiley & Sons, Inc.

Musson, R., 2016. *What Does a Data Scientist Do?* [Interview] (20 Apr 2016).

Siegel, E., 2016. *Predictive Analytics.* s.l.:John Wiley & Sons, Inc.

# Appendix

Please refer to GitHub - https://github.com/howyue/DAP