

Data Analytical Programming

Soo wai hong  
TP045363

A report submitted in fulfilment of the

requirements for the course

Data Analytical Programming

CT050-3-M-DAP

Asia Pacific University of Technology & Innovation (APU)

School of Computing and Technology

10th April 2017

# 

# **DECLARATION OF ORIGINALITY AND EXCLUSIVENESS**

I declare that this assignment report entitled

Data Analytical Programming  
 is the result of my own research work except as cited in the references.

|  |  |
| --- | --- |
| Signature: |  |
| Name: | Soo Wai Hong |
|  | TP045363 |
| Date: | 30TH APRIL 2017 |

# **Acknowledgment**

First of all, I am grateful The Almighty God for establishing me to complete this Data Analytical Programming (DAP) module and assignment.

I wish to express my sincere thanks to *Dr Kalai Anand Ratnam* for his expert, sincere and valuable guidance and encouragement extended to me.

Lastly, I also place on my sense of gratitude to everyone who is directly or indirectly that has given their helping hand in this venture.

**Table of Contents**

[DECLARATION OF ORIGINALITY AND EXCLUSIVENESS 2](#_Toc481522247)

[Acknowledgment 3](#_Toc481522248)

[Introduction 5](#_Toc481522249)

[Problem Statement 5](#_Toc481522250)

[Objectives 5](#_Toc481522251)

[Discussion on Data Sets 6](#_Toc481522252)

[Methodology 7](#_Toc481522253)

[Data Reduction and Data Cleaning 7](#_Toc481522254)

[Analysis 20](#_Toc481522255)

[Total crime changed in percentage across all states for year 2014 and 2015 21](#_Toc481522256)

[Total number of crime across all states for year 2014 and 2015 26](#_Toc481522257)

[Crime ratio based on population across all states for year 2014 and 2015 28](#_Toc481522258)

[Individual crime statistic for top 5 states with highest crime ratio 34](#_Toc481522259)

[Discussion on overall analysis 35](#_Toc481522260)

[Possible Solution 36](#_Toc481522261)

[Conclusion 37](#_Toc481522262)

[References 38](#_Toc481522263)

[Appendix 39](#_Toc481522264)

[Appendix 1 – Descriptive Statistic in Excel 39](#_Toc481522265)

[Appendix 2 – Import Population Data 41](#_Toc481522266)

[Appendix 3 – Import Table 4 41](#_Toc481522267)

[Appendix 4 – Join Table 4 with Population 42](#_Toc481522268)

[Appendix 5 – Upload File for Analysis and Create Permenant Dataset 43](#_Toc481522269)

[Appendix 6 – Upload Statecode file for Analysis and Create Permenant Dataset 44](#_Toc481522270)

[Appendix 7 – Merge MASTER and STATECODE using PROC SQL 44](#_Toc481522271)

[Appendix 8 – Total Crime Per State on Year 2014 and 2015 45](#_Toc481522272)

[Appendix 9 – Total Crime Changed Per State for Year 2014 and 2015 table view 46](#_Toc481522273)

[Appendix 10 – Total Crime Changed Per State for Year 2014 and 2015 map view 47](#_Toc481522274)

[Appendix 11 – Crime ratio based on population across all states for year 2014 and 2015 48](#_Toc481522275)

[Appendix 12 – Individual Crime in Top 5 States with Highest Crime Ratio 50](#_Toc481522276)

# **Introduction**

The preliminary dataset provided in this assignment contained information related to offenses reported to law enforcement by state and city with the population of 100,000 and over for the period of January till June for the year of 2014 and 2015. The source of this dataset came from FBI through the Uniform Crime Reporting (UCR) Program.

These offenses are classified into two categories which are violent crime and property crime. Violent crime includes murder, rape (both revised and legacy definition), robbery, and aggravated assault. However, property crime includes burglary, larceny-theft, motor vehicle theft, and arson.

# **Problem Statement**

As a new hired Data Scientist at the Headquarters of Federan Bureau of Investigation (FBI), Washington. D.C., United States, I have been requested to analyze this preliminary dataset in detail using SAS program to identify the relationships and patterns among these given crime data. As a conclusion, it is required to produce output results and interpreting them by writing report and or presentation. The ODS Statistical Graphics can be used as a medium to further support the interpretation.

# **Objectives**

* Total crime changed (both violent and property crime, exclude Arson) in percentage across all states for the 1st 6 months of year 2014 and 2015
* Total number of crime (both violent and property crime, exclude Arson) across all states for the 1st 6 months of year 2014 and 2015
* Crime ratio based on population across all states for the 1st 6 months of year 2014 and 2015
* Individual crime statistic for top 5 states with highest crime ratio

# **Discussion on Data Sets**

There are 15 variables in this dataset. The following table contained the variables and its corresponding data type:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Data Type | Variable | Data Type |
| State | Character | Robbery | Numeric (with missing value) |
| City | Character | Aggravated Assault |
| Year | Numeric (with missing value) | Property Crime |
| Population | Burglary |
| Violent Crime | Larceny Theft |
| Murder | Motor Vehicle Theft |
| Rape (Revised Definition) | Arson |
| Rape (Legacy Definition) |  | |

Apart from that, this dataset contained 43 different states. It built up 260 entries when these states are combined with cities. We can see that California is the state that contained the highest number of city which is 66 in total from this dataset. Texas is the 2nd highest which is having 30 cities followed by Florida which is 21 cities. Arizona is the 4th highest which has 9 cities and followed by Colorado and Illinois which are both having 8 cities.

Variable violent crime is the subtotal that contained the total number of cases for murder, rape robbery, and aggravated assault whereby variable property crime is the subtotal contained the total number of cases for burglary, larceny-theft and motor vehicle theft. Variable arson is excluded in property crime subtotal due to fluctuations in reporting. Among 520 observations from this dataset, there is 27 observations that are having missing value in arson variable.

Variables rape is kept in two different columns which classified in legacy definition and revised definition. In year 2013, the revised definition is used by FBI on collecting rape data. This revised definition for rape has the term “forcible” that is used in legacy definition. It is now removed from the offense name in the revised definition. The legacy definition is still used in this dataset because of when the revised definition for rape is used, not all agencies or state have had the ability to change their records management systems to incorporate with the new definition.

By observing each entry in this dataset, we can see that some of the entries are not having any figure for any of the offenses. This is because the complete January through June data are not available at the time when this data is consolidated. For example, crimes data is missing in Arizona state, Tucson city for year 2014. Apart from that, we can observe that there is no population data for year 2015 on all observations.

# **Methodology**

## **Data Reduction and Data Cleaning**

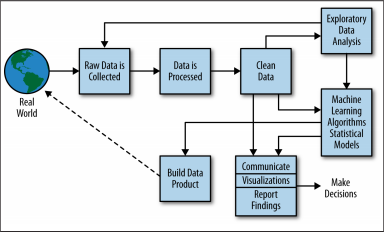


Figure 1 - Data Science Process Flowchart (O'Neil, 2014)

From Figure 1, data reduction and cleaning is the first crucial step in data analysis. In the real word, massive amount of data is created from various sources every day. They can be in the form of structured, unstructured or semi structured. Before these data is useful for performing analysis, the data must be in a clean format.

The cleaning process includes identifying outlier, missing value, scaling, check normalization of data and others. It is important to make sure data is transforming to normal distribution because most of the statistical model only work well with normalized data when we are seeking for relations.

For the crime data given in this task, several steps must be taken to clean up the data before it can be used in SAS for further analysis. These steps are listed as followed:

**Step 1** – Convert the original state code text file that is retrieved from census website (Census State Code, 2013) into csv format using excel.

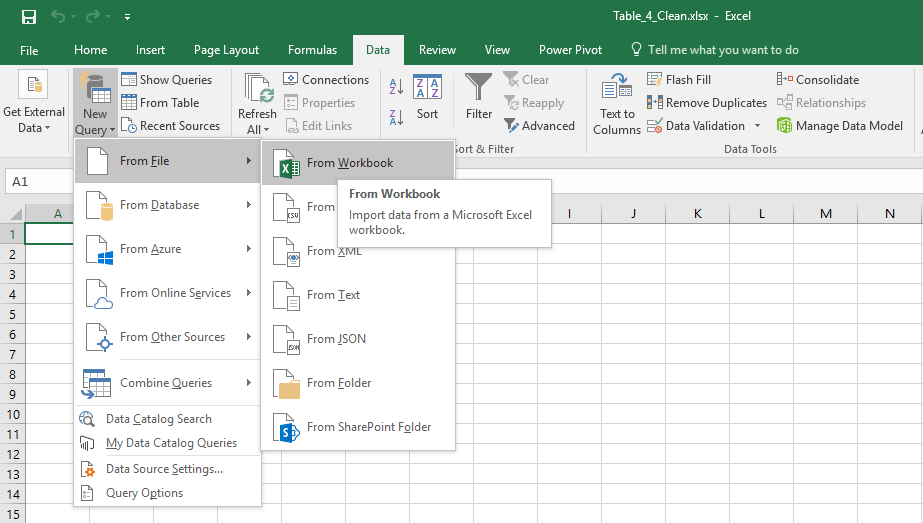
|  |  |  |
| --- | --- | --- |
|  |  |  |

**Step 2** – Rename the given Table 4 long file name in this task from

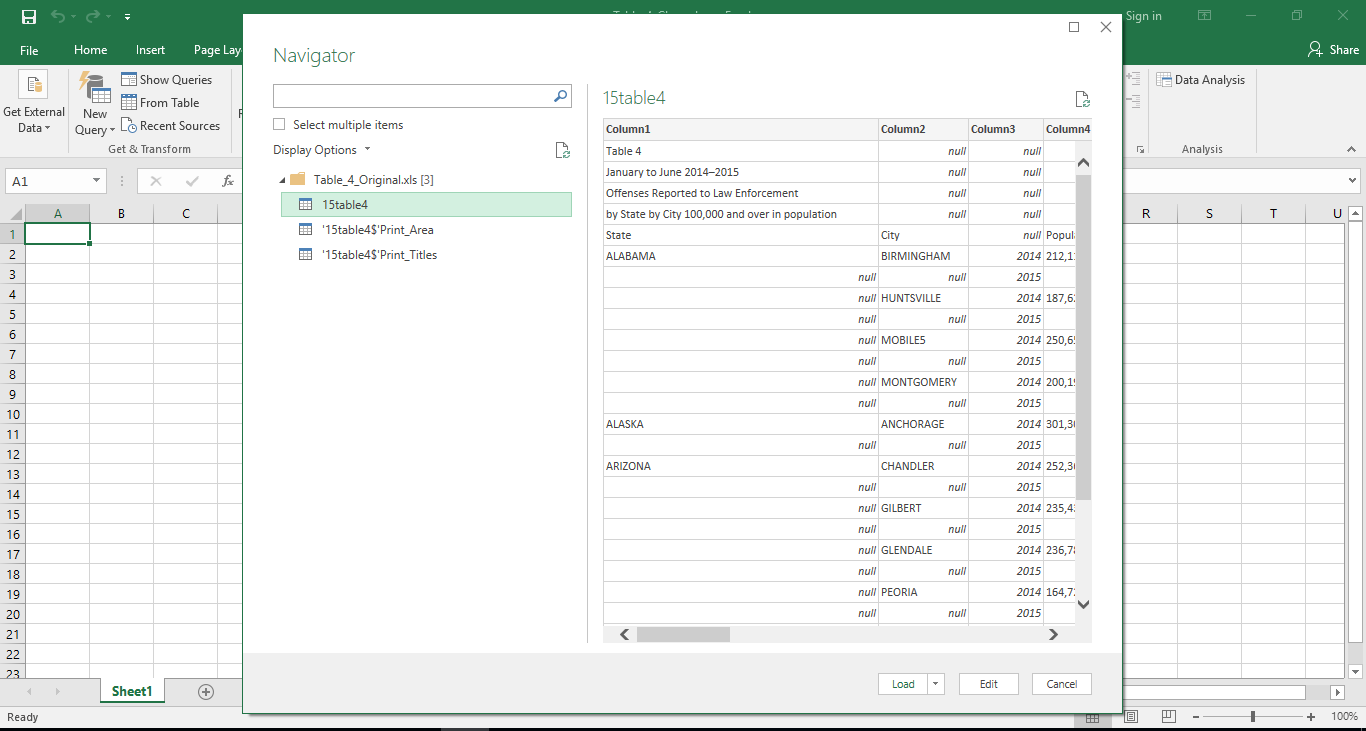
*Table\_4\_January\_to\_June\_2015\_Offenses\_Reported\_to\_Law\_Enforcement\_by\_State\_by\_City\_100,000\_and\_Over\_in\_Population.xls*

to *Table\_4\_Original.xls*

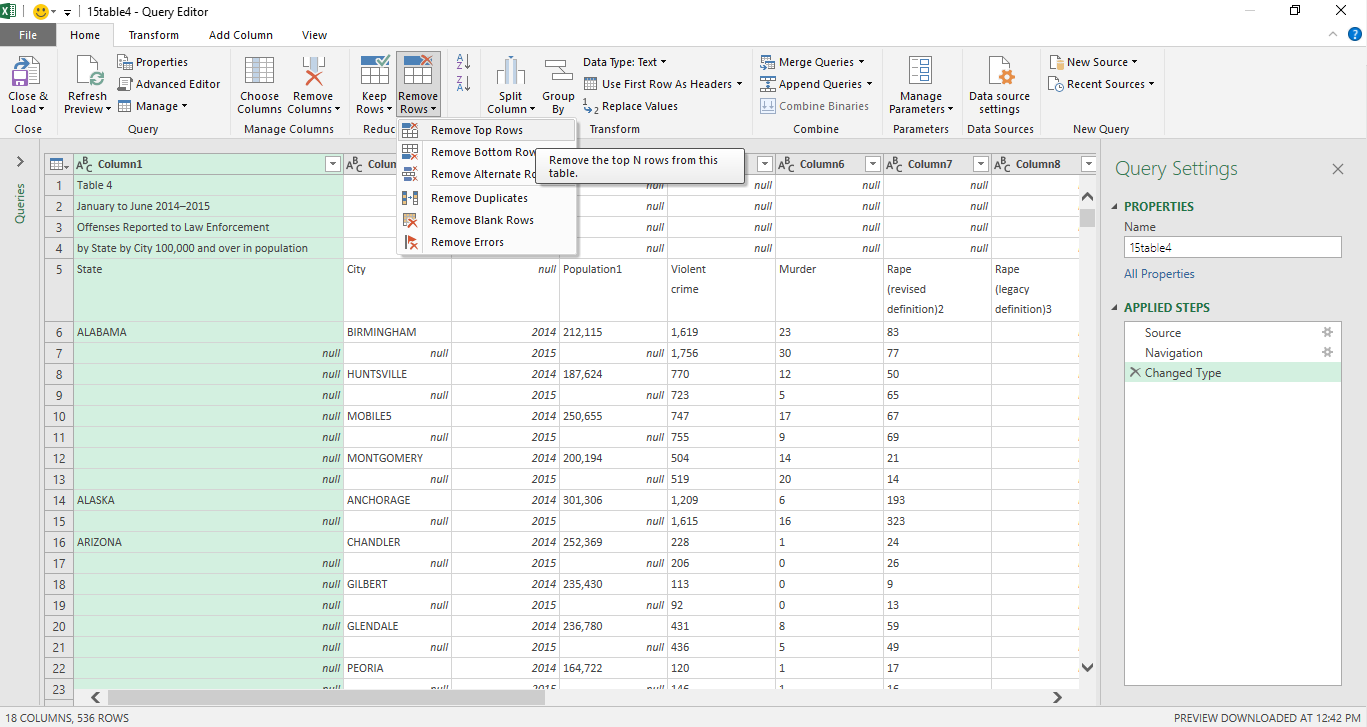
**Step 3** – The given *Table\_4\_Original.xls* file is in report format. It is not recognizable in SAS for its formatting. In this task, Excel will be used to convert the report format data into raw data. In Excel 2016, Data Query can be used to transform an excel Workbook that is in report format to raw data format.



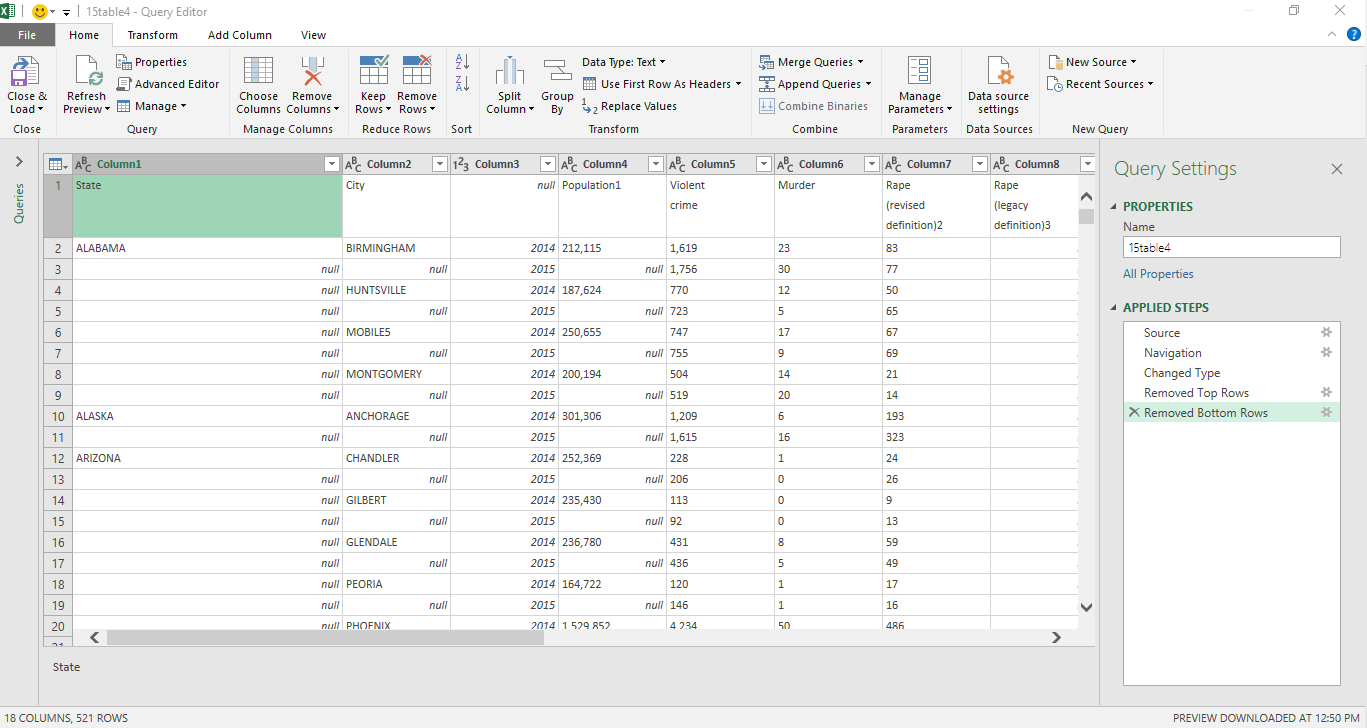
After clicking From Workbook showed on above figure, we can see that the Navigator pop up as showed in the following figure. By clicking the 15table4 in the left panel, we can see that the right panel will preview the data. By clicking the edit button, the Query Editor screen will pop up that can facilitate the data cleaning process.



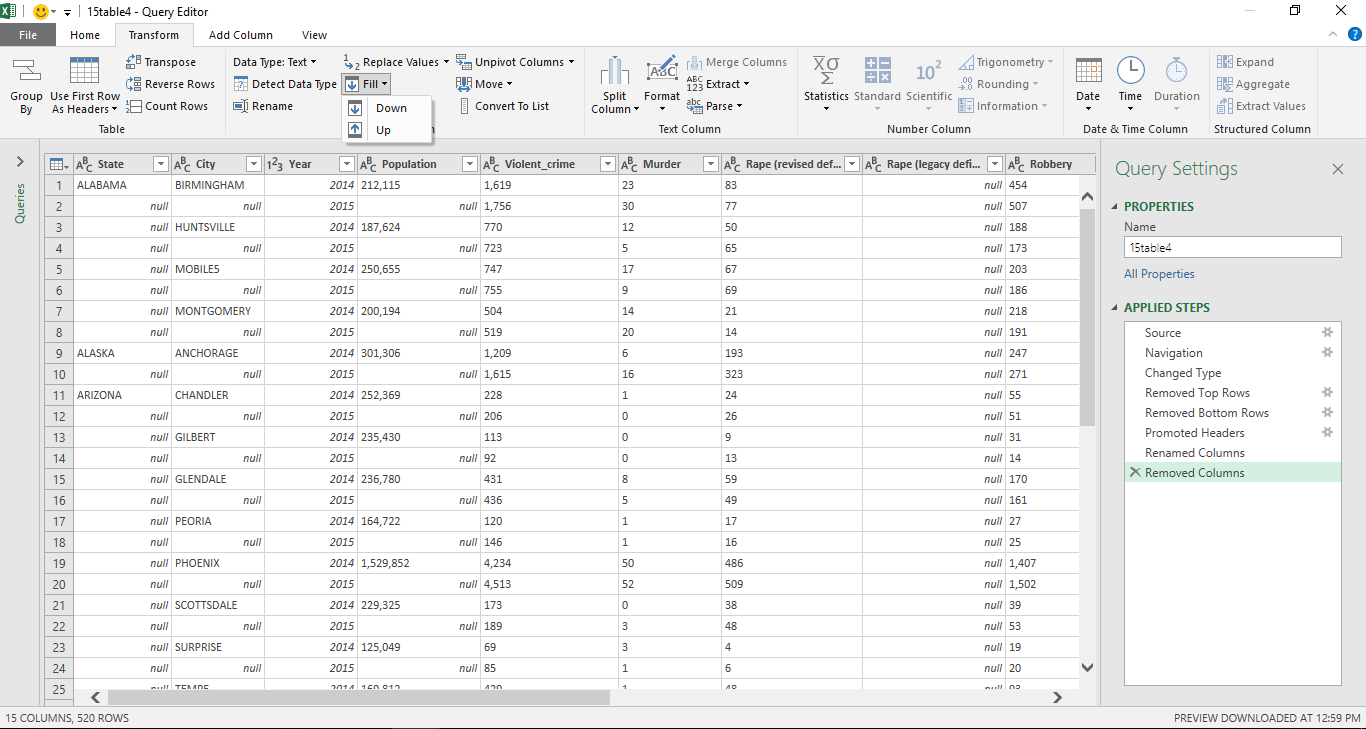
In the Query Editor, Remove Rows feature can be used to remove the report header and footer in this given dataset.



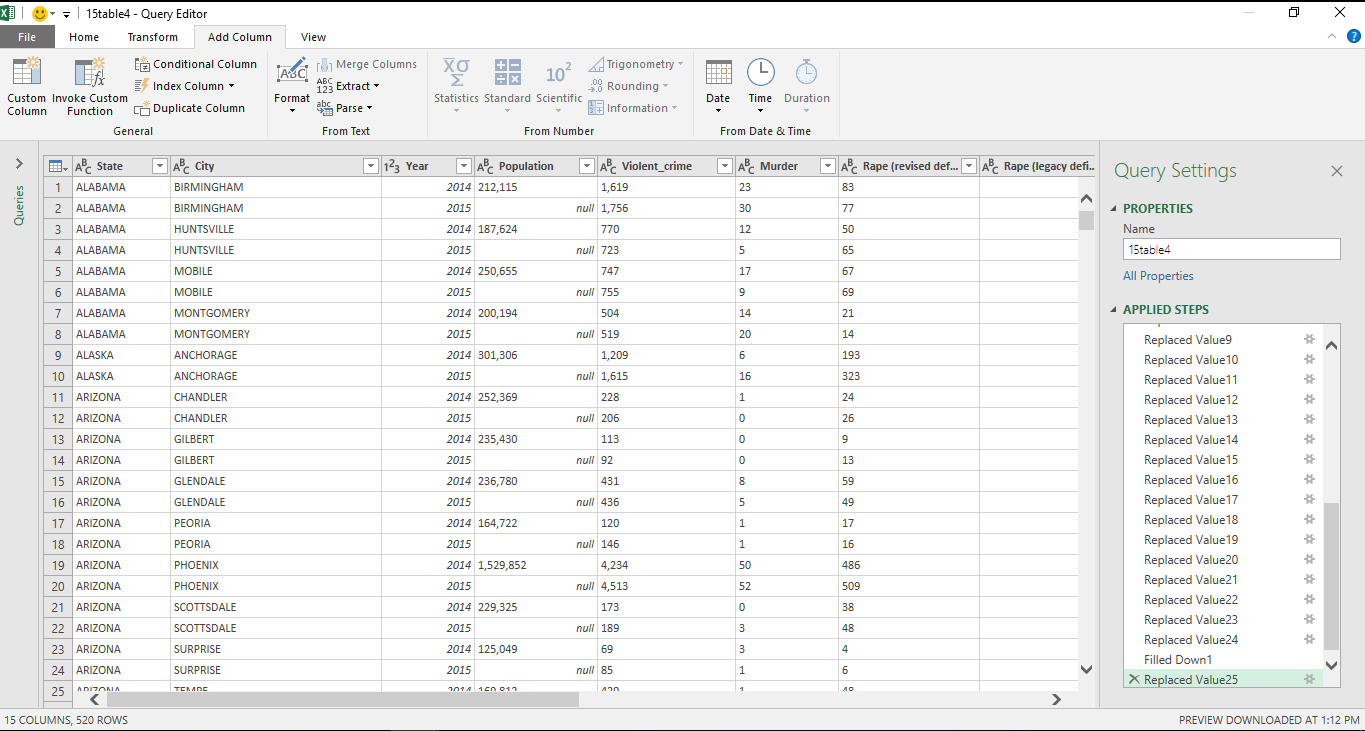
Using the feature of Use First Row As Headers will update the column header based on value in 1st row.

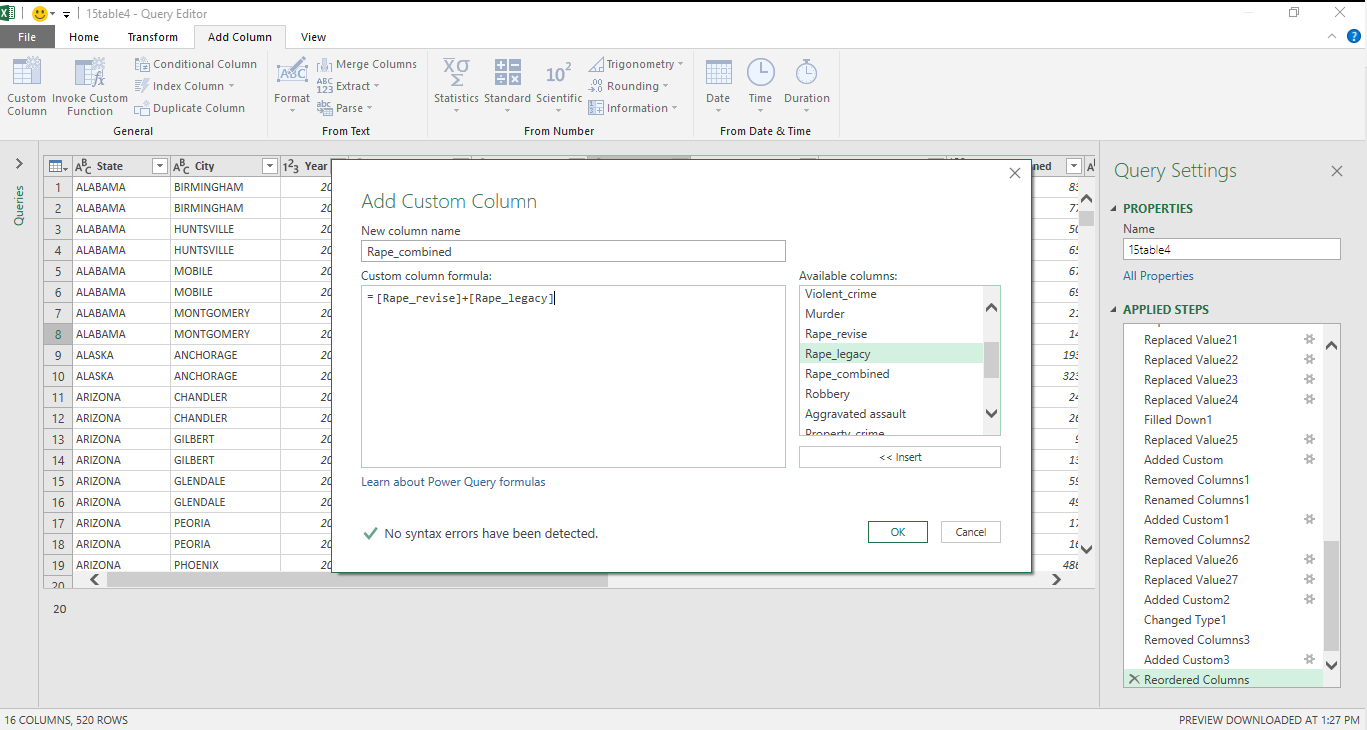


In the Transform tab, there is a feature named Fill that allowed us to fill up the null value for State and City based on the 1st value in the group.

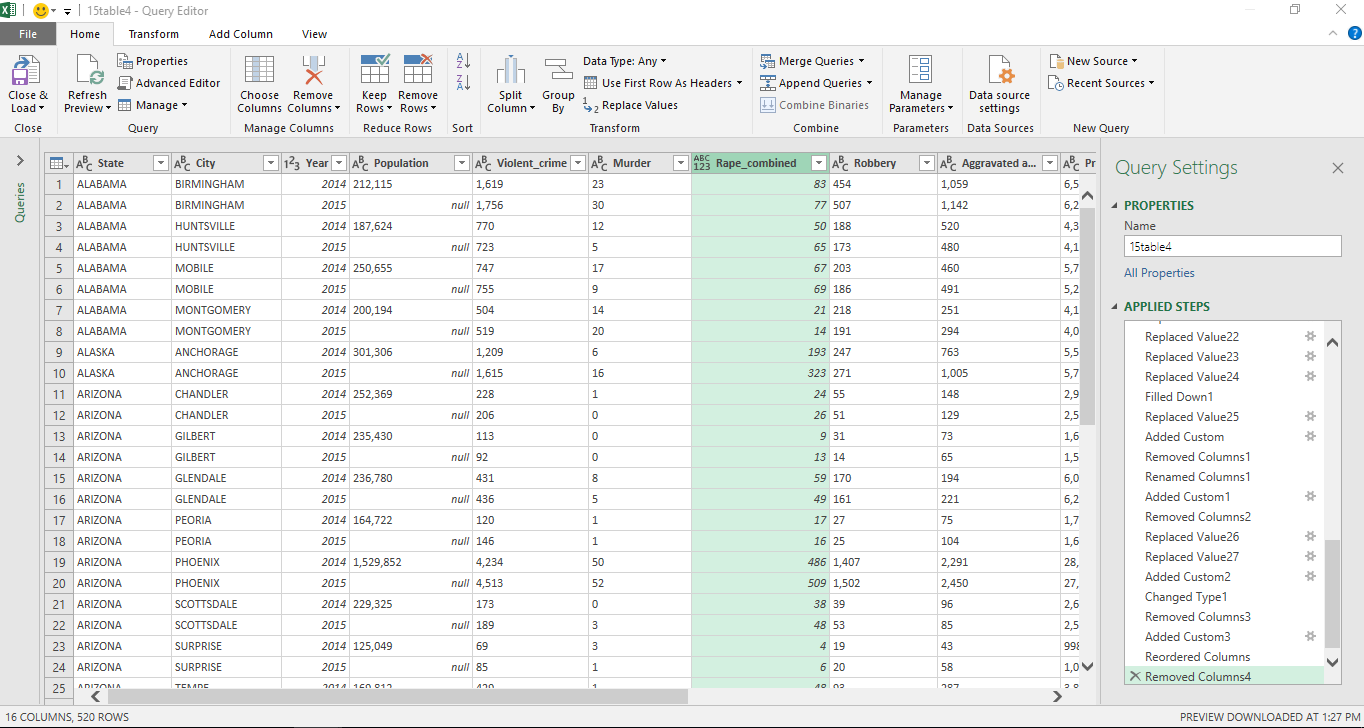


After filling up the null value based on the first value in the group, the edited dataset looks as followed. There is a feature called Custom Column that can be used to add a new column to this dataset. A new column named rape combined will be created to get the sum of rape with revise and legacy definition.

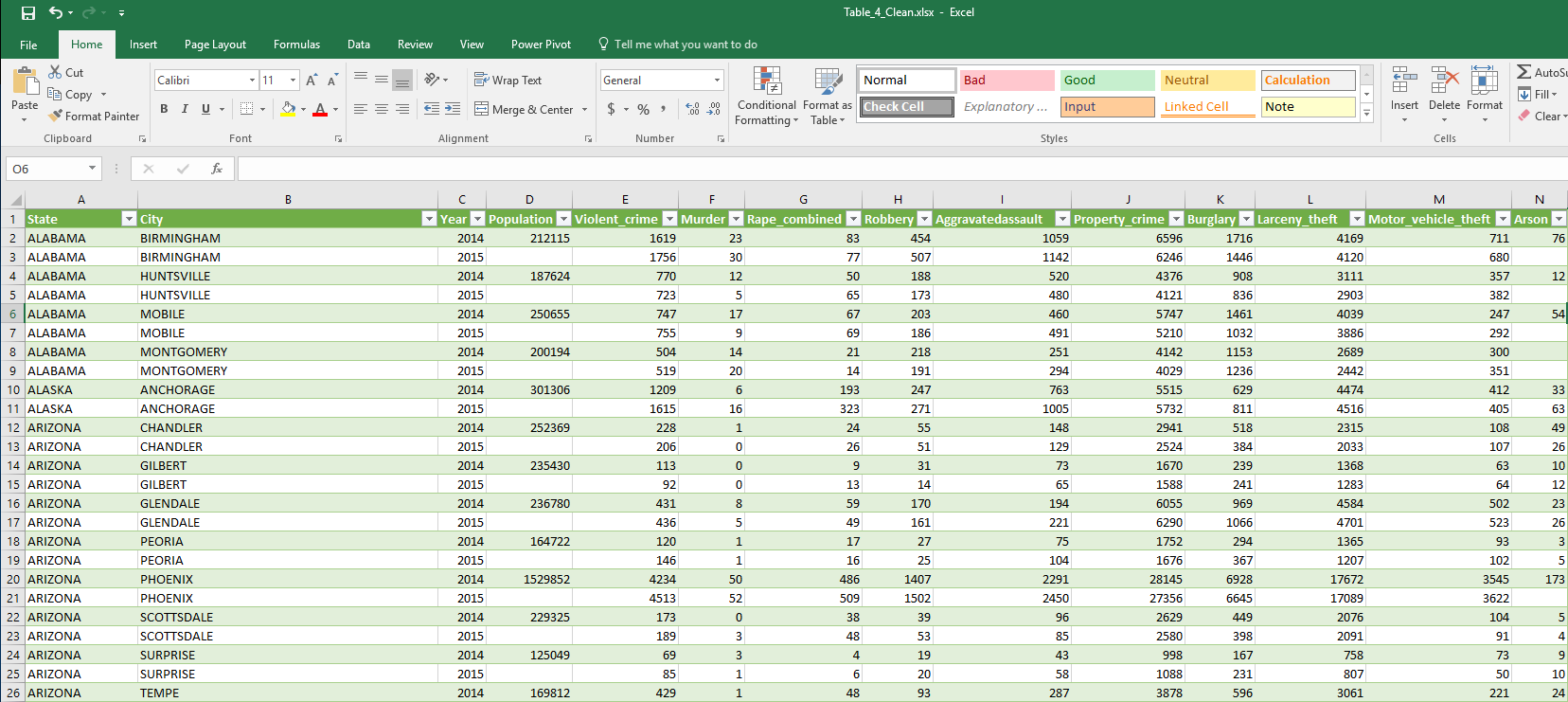




One last step before we load this data to excel, Remove Column feature can be used to delete the columns for rape revise and legacy definition. By clicking of Close & Load, data can now be loaded into Excel.



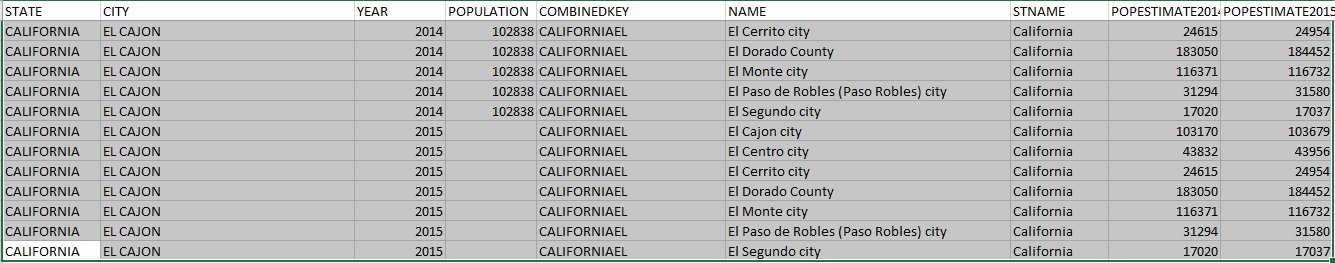
The loaded data showed as followed in Excel. This is a clean file by not having merge cell, header, footer, and subscript. By performing Descriptive Statistic in Excel, it gives us an summary statistic for these data. These statistic details include min, max, mean, mod, range, count and many others (refer to Appendix 1).



**Step 4** – By browsing through this dataset, we can observe that the population for year 2015 is missing for all cities. To obtain the accurate population for year 2015 on all cities, we can use SAS to join this crime dataset with the population data available for year 2015 in Census (Census Population, 2016). Refer to Appendix 3, 4, 5 for loading of Population and Table 4 dataset and joining of these two dataset to obtain the population value for year 2015.

The joining process for these two datasets will be tedious due to these files are came from different sources and the common key is not exactly matching with each other. Several steps are executed in order to get the maximum number of matched between these two files:

1. Create a common key in table 4 by using 1st word of state concatenating 1st word of city
2. Create common key in population file by using 1st word of state concatenating 1st word of city. Transform this key into upper case for matching as the table 4 source file state and city are both in upper case
3. Remove duplicate records in population file by having state, city, estimate population for year 2014 and year 2015 as key
4. Keeping all records in the Table 4 by performing left join with the remove duplicate records from item 3 using the combined key created in item 1 & 2. This step resulted in the Table 4 having 1506 records. This is caused by using the method in item 1 & 2 to form the combined key. We can see that when the city in the population file with multiple rows with the first word is the same, then these matched records are incorrect.



1. To overcome the problem in item 4, we will now match the 2nd word for city in both files. Before continuing to form the key using 2nd word in city, we can observe that the following records will resulted in mismatched after using the combined key with 2 words from city. To match the 2nd word in city field, we should first remove the text “city” in the city field from population file.

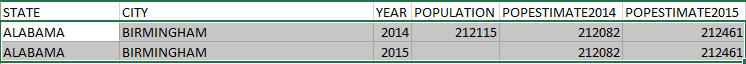


1. After match with the 2nd word of the city using method in item 5, we now left with 568 records. By observing these records in Excel, we can see that the duplicate records come from state and city contained population with different value for year 2015 in the population dataset. We will select the population record that is with the 2014 population value closest to Table 4 file.

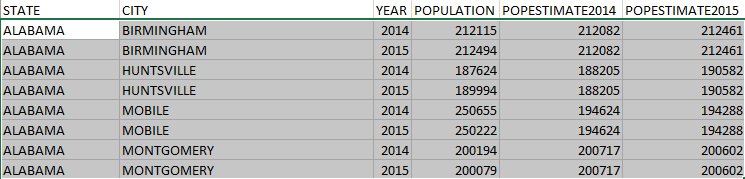
**Step 5** – Recalculate the estimated population for year 2015 in merged Table 4 dataset from STEP 4 (Table\_4\_Population.csv). By having the estimated population for year 2014 and 2015 from Census population file, we will set the Table 4 year 2015 population using the same ratio. The estimated population for 2015 is calculated as

(year 2014 population in Table 4) \* (estimated population in 2015 from Census) / (estimated population in 2014 from Census).

Based on the following example, the population value for year 2015 in Table 4 will be 212115 \* 212461 / 212082 = 212494



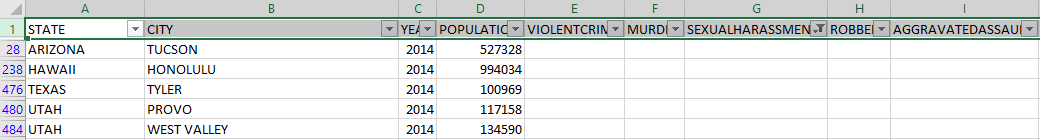
The following figure showed the result of population for year 2015 is filled.

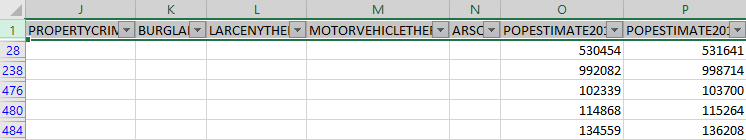


**Step 6** – Handle missing value. From Appendix 1 - Descriptive Statistic created earlier, we can see that each crime variable is having missing value. The summary of missing value for each variable is showed as followed:

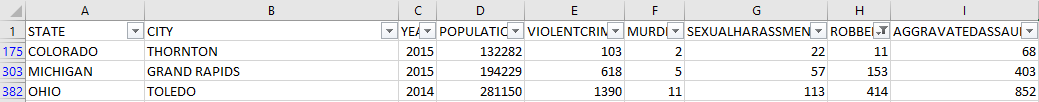
|  |  |  |
| --- | --- | --- |
| Variable | Count | Missing value |
| Violent\_crime | 515 | 5 |
| Murder | 515 | 5 |
| Rape\_combined | 515 | 5 |
| Robbery | 515 | 5 |
| Aggravatedassault | 515 | 5 |
| Property\_crime | 512 | 8 |
| Burglary | 514 | 6 |
| Larceny\_theft | 514 | 6 |
| Motor\_vehicle\_theft | 514 | 6 |
| Arson | 493 | 27 |

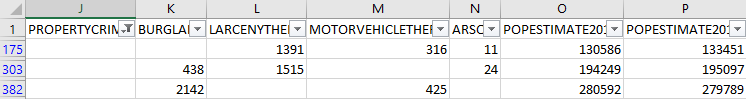
5 observations contained missing value for all crime variables on year 2014 are showed in the following table. We will not handle any missing value for these observations due to there is no reasoning for updating any value into these crime variables. We will then exclude these observations from analysis.





Apart from the 5 records showed from the above table, we can see that there are 3 observations having missing value in property crime and 1 observation with missing value on burglary, larceny theft and motor vehicle theft variables. These observations are identified and showed in the following tables:

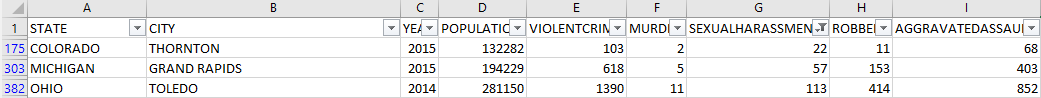


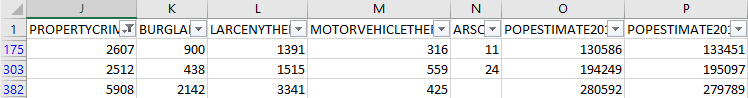


To overcome these missing values, we will perform mean substitution into burglary, larceny theft and motor vehicle theft variables based on the mean for these variables on that particular year. Property crime is a summation on these 3 fields. We can get the property crime value populated once these 3 fields are replaced. The mean value for these variables are calculated as followed:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Year | Sum | Count | Average |
| Burglary | 2015 | 233015 | 259 | 899.6718 |
| Larceny theft | 2014 | 848648 | 254 | 3341.133 |
| Motor vehicle theft | 2015 | 144795 | 259 | 559.054 |

The replaced value for these variables are showed in the following table:



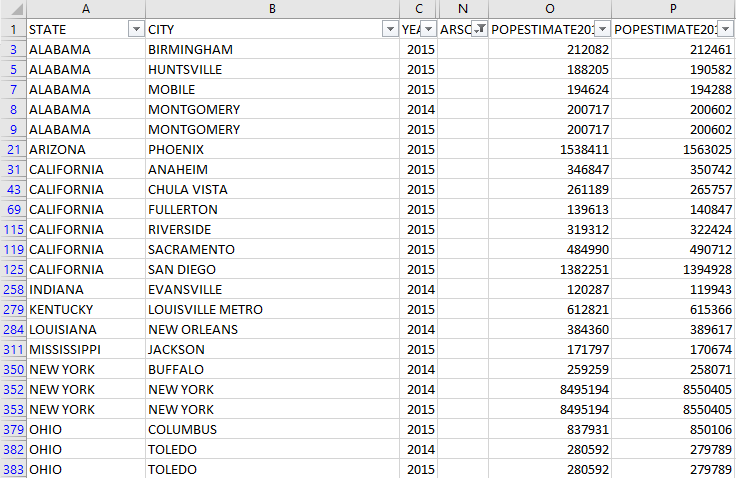


Lastly, we will be looking into variable arson where there are 27 observations with missing value. 27 observations are equivalent to (27 / 520 \* 100) which is 5.19%. Some of the researches showed that by having missing value between 5% to 10% is accepted to perform imputation without having significant impact to the statistic result (Yiran Dong, 2013).

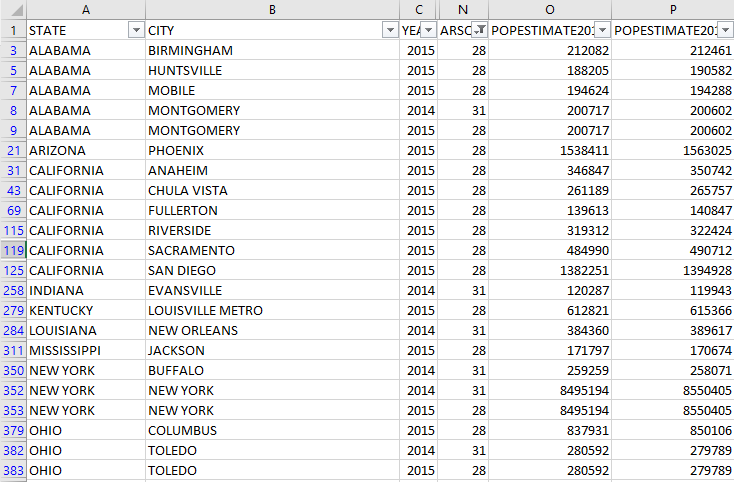
In this case, we will conduct the mean substitution on arson variable since the missing value percentage is within the accepted range. The mean value for arson variable are calculated as followed:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Year | Sum | Count | Average |
| Arson | 2014 | 7616 | 249 | 30.586 |
| Arson | 2015 | 6890 | 244 | 28.237 |

The following figure showed the 22 observations with missing value in arson variable. Another 5 observations with all crime values missing will not have the arson value populated using mean substitution.



The following figure showed the 22 observations with arson variable populated.

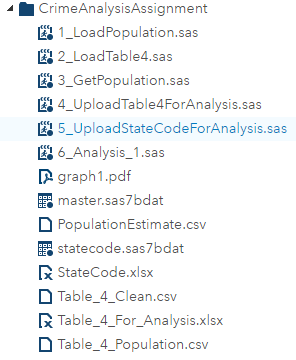


The data reduction and data cleaning process are now done and the data is ready for performing analysis. A new file with named *Table\_4\_For\_Analysis.xlsx* will be created and upload into SAS for further analysis.

## **Analysis**

Firstly, upload the *Table\_4\_For\_Analysis.xlsx* into SAS and create a permenant dataset MASTER.sas7bdat (refer to Appendix 5). All the subsequent analysis will be based on this permenant dataset.

Secondly, upload the *StateCode.xlsx* into SAS and create permenant dataset STATECODE.sas7dat (refer to Appendix 6). This dataset will be used to merge with the MASTER.sas7dat on getting the statecode for each state to build SAS map.



The code for merging MASTER and STATECODE dataset using PROC SQL is showed in Appendix 7. This merge dataset will be used in all the subsequent analysis.

### **Total crime changed in percentage across all states for year 2014 and 2015**

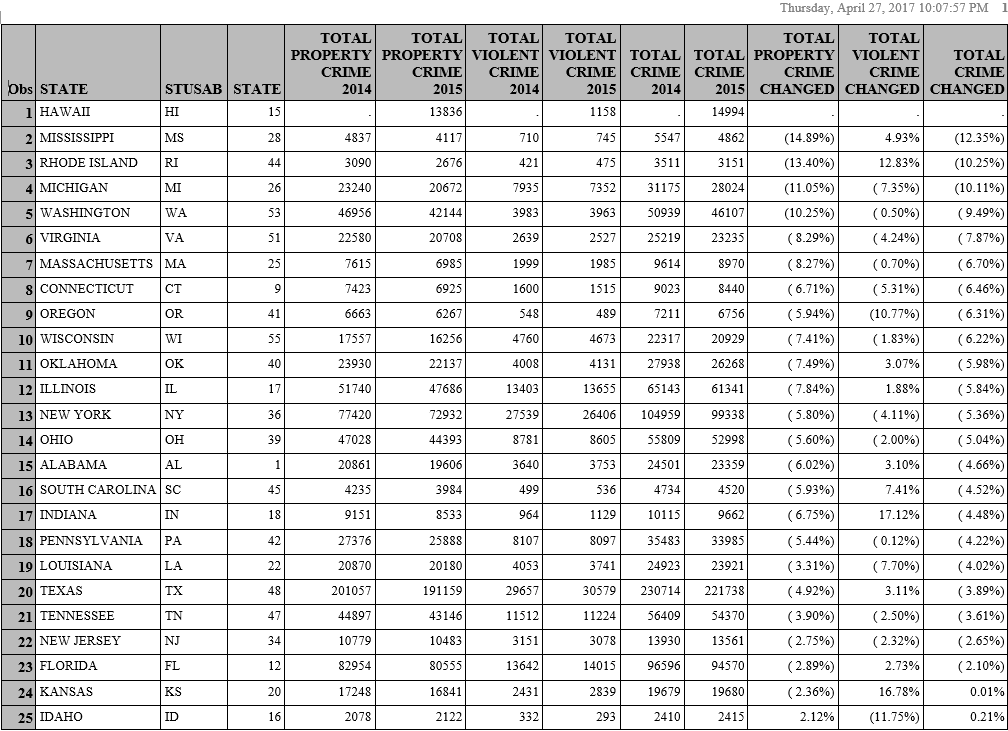
Figure 2 showed the total crime (both property and violent crime, exclude Arson) changed in percentage per state on year 2014 and 2015. This table is sorted in descending order based on column Total Crime Changed.

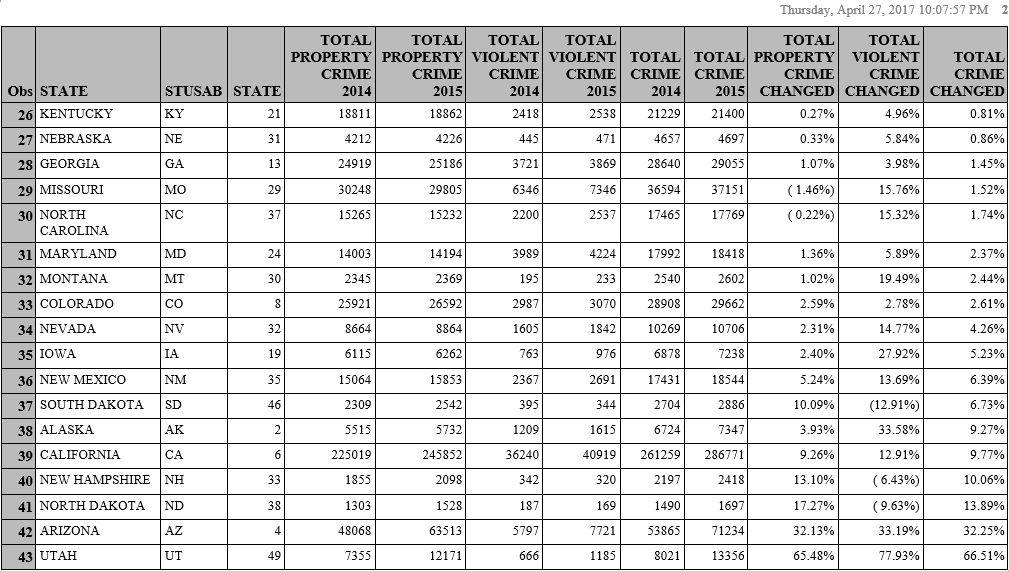
In Data Reduction and Data Cleaning section, we realized that there is missing crime for some cities on year 2014 in state **Hawaii, Utah, Arizona and Texas**. By observing Figure 2, the 1st row showed that Hawaii is having missing value in total crime changed due to missing crime record for year 2014. We can also see that for state Utah and Arizona, the total crime changed in percentage is significantly high in the list due to the missing crime data for year 2014. This showed the symptom of outlier and these states should be excluded for discussion.

|  |  |  |  |
| --- | --- | --- | --- |
| Number of State | Property Crime | Violent Crime | Total Crime |
| Excluded | 3 | 3 | 3 |
| Increase | 15 | 23 | 18 |
| Decrease | 25 | 17 | 22 |

The above table showed the summary of the crime changed for all states. Among the 43 states, we can see that property crime and total crime are on a decreasing trend. Violent crime is on increasing trend in year 2015 compare to year 2014. These statistics complied to the finding known to law enforcement agencies.

The code for generating this report can be found in Appendix 9.



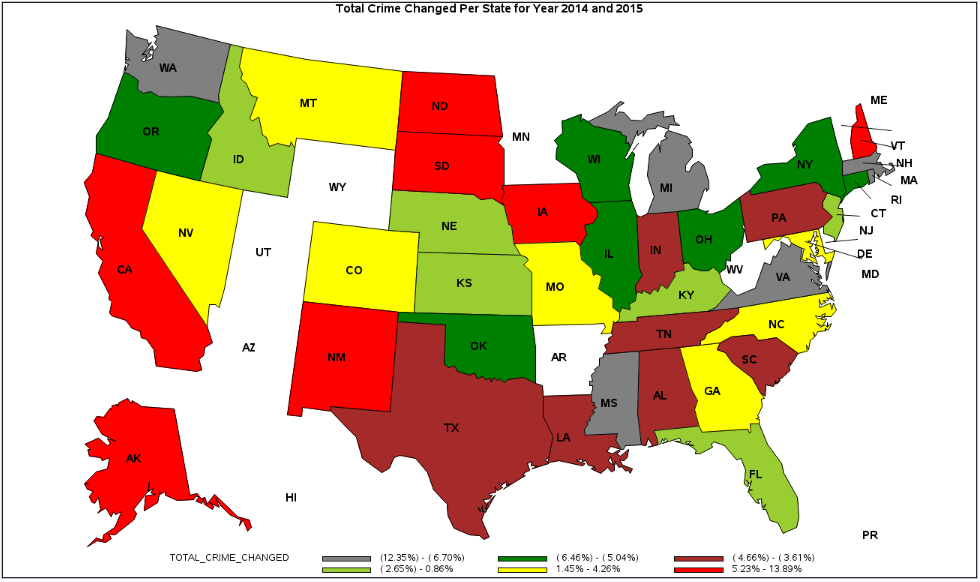


*Figure 2 – Total Crime Changed in Percentage Per State for Year 2014 and 2015*

Figure 3 showed the total crime changed in percentage per state for year 2014 and 2015 in map view. This map view can give a better insight on the location for which the crime increase or decrease. 3 states namely Utah (UT), Arizona (AZ), and Hawaii (HI) that are identified to be outlier from figure 2 are excluded from the map.

This map classified total crime changed percentage per state for year 2014 and 2015 into 6 groups. State labelled with red color showed the highest increase and state labelled dark grey showed the highest decrease. We can see that both highest increase and decrease of crime changed are located at the border of US.

The code for generating this map can be found in Appendix 10. PROC GMAP is used to create this map and the label of each state is plotted using ANNOTATE option.

****

*Figure 3 – Total Crime Changed in Percentage Per State for Year 2014 and 2015 in Map View*

### **Total number of crime across all states for year 2014 and 2015**

Figure 4 showed the total number of crime (both property and violent crime, exclude Arson) per state on year 2014 and 2015.

In this figure, we can clearly see that the top 4 states with highest number of crime are **California, Texas, New York and Florida**. Texas, New York and Florida are the 3 states that have a slightly decrease in number of crime cases for year 2015 compare with year 2014. California is the only state among these four that has a significant increase of crime case for year 2015.

The top 4 states with the lowest number of crime are **Idaho, New Hampshire, North Dakota, and Montana**. By referring to figure 2, these 4 states showed an increase on crime case in year 2015. North Dakota is a state with the highest increase in total crime case which is 13.89% in year 2015.

The SAS code for generating this figure is in Appendix 8.



Figure 4 - Total Number of Crime Per State for Year 2014 and 2015

### **Crime ratio based on population across all states for year 2014 and 2015**

From the previous, figure 4, we can see **California, Texas, New York and Florida** are 4 states with the highest number of crime case. While the crime ratio is calculated based on population in each state showed in figure 5, we can find out that New Mexico is the state with highest crime ratio based on population and not California although California is the state with highest number of crime case. On the other hand, New York is the state with lowest crime ratio calculated based on population in figure 5 and New York is in the list of states with 3rd highest number of crime case.

By observing figure 5, it clearly showed that the chances of crime happened is not felt into a state that are having the highest number of crime case. In figure 5, New Mexico showed that there is around 3% chance that an individual will conduct a crime which is the highest in US and New York is the lowest where the chance is only around 1.2%.

The top 5 states with highest crime ratio are **New Mexico, Missouri, Maryland, Mississippi, and Washington**. New Mexico, Missouri and Maryland have an increase in crime ratio whereby Mississippi and Washington have a decrease in crime ratio.

Apart from that, we can observe the crime ratio is on a decreasing trend in year 2015 as per the red line showed in the graph for majority of the states. This is a good sign that the crime ratio is going down in the country.

Figure 6 & 7 showed the crime ratio in map view for year 2014 and 2015. The crime ratio is categorized into 6 groups. States with highest crime ratio are labelled in red and it remained the same for year 2014 and 2015. Figure 8 showed the crime ratio in table view for year 2014 and 2015.

The SAS code for generating these figures is in Appendix 11.



Figure 5 – Crime Ratio Based on Population Per State for Year 2014 and 2015

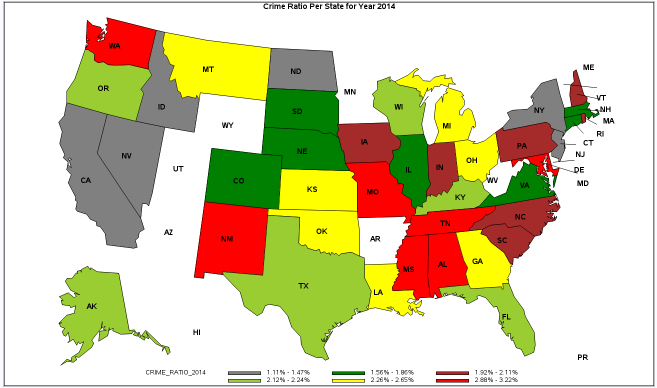
****

Figure 6 – Crime Ratio Per State for Year 2014

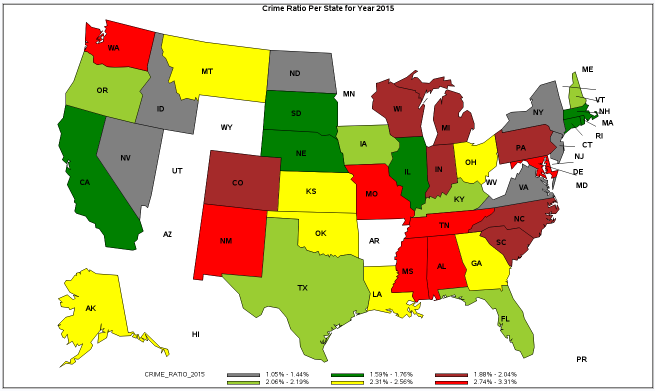
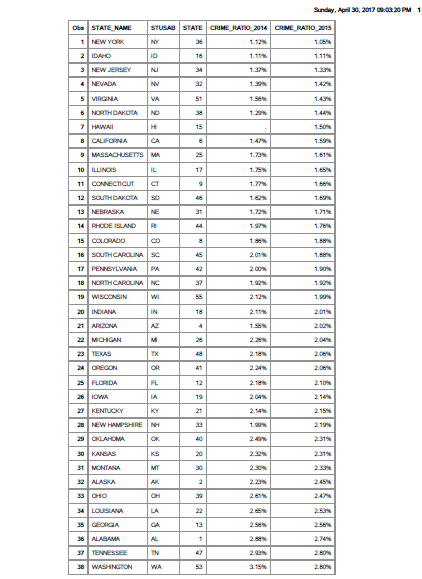
****

Figure 7 – Crime Ratio Per State for Year 2015

****

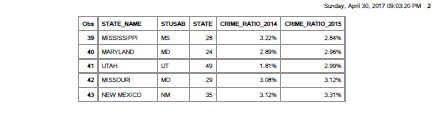
****

Figure 8 – Crime Ratio in Table View for Year 2014 and 2015

### **Individual crime statistic for top 5 states with highest crime ratio**

Figure 9 showed the individual crime statistic for the top 5 states with highest crime ratio. One of the item worth to take note in this chart is the increase of murder case in Maryland which is close to 5%. This is the highest increase for all crimes across all 5 states in year 2014 and 2015.

The SAS code for generating these figures is in Appendix 12. This figure is built with using PROC TABULATE.

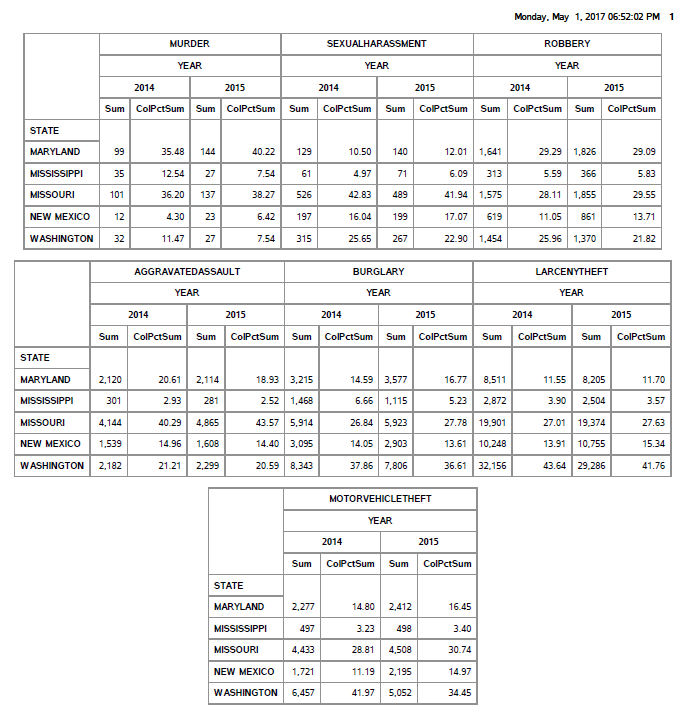


Figure 9 - Individual Crime for Top 5 States with Highest Crime Ratio for Year 2014 and 2015

# Discussion on overall analysis

The overall finding related to the top 5 states with highest crime ratio will be further discuss in this section. Based on the analysis from the previous section, we can summarize the finding for the top 5 states with highest crime ratio in the following tables:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| States | State Code | Total Crime Changed | Total Violent Crime Changed | Total Property Crime Changed | Crime Ratio 2014 | Crime Ratio 2015 |
| New Mexico | NM | 6.39% | 13.69% | 5.24% | 3.12% | 3.31% |
| Missouri | MO | 1.52% | 15.76% | (1.46%) | 3.08% | 3.12% |
| Maryland | MD | 2.37% | 5.89% | 1.36% | 2.89% | 2.96% |
| Mississippi | MS | (12.35%) | 4.93% | (14.89%) | 3.22% | 2.84% |
| Washington | WA | (9.49%) | (0.5%) | (10.25%) | 3.15% | 2.8% |

By referring to this table, we can see that New Mexico, Missouri, and Maryland have an increase in crime ratio whereby Mississippi and Washington have a decrease in crime ratio for year 2015 compare to year 2014. All states except Washington have an increase in violent crime which matched with the trend across the country.

By summarizing the data in figure 9, we can conclude that robbery and aggravated assault are the major cause for the increase of violent crime. Murder case has showed a significant increase in Maryland and Missouri for year 2015 among the top 5 states with the highest crime ratio.

Although California, Texas, New York and Florida are states with highest number of crime, the crime case happened in these states are relatively insignificant compare with these 5 states (New Mexico, Missouri, Maryland, Mississippi, and Washington) with the highest crime ratio based on population. More attention should be given to these 5 states as we will conclude that these states are riskier to live in the US.

# **Possible Solution**

Violent crime has showed an increase throughout this analysis. The possible actions to be taken are:

* Gun control: The increase of murder cases in some states are worrying and gun control will be a practical way to reduce case on gun death.
* Focus on hotspot: Robbery and aggravated assault are the major types of violent crime. By identifying hotspot, police patrol in these hotspots can reduce the chances of crime happened in this area
* Education: Educational initiatives can help to prevent the increase of crime. Parents and teachers play a huge role throughout the development life cycle of a children. They should immediately stop children when their attitude and behavior is not correct and educate them on not joining criminal gang.
* Target inequality: Inequality is a caused for increasing of crime and income inequality could be one of it. When poorer people perceive inequality, they view crime as more acceptable as they feel less of commitment to social norms. Therefore, by reducing the economic inequality will be one of the method to reduce crime.

# **Conclusion**

The major objective of this assignment is to analyze the given dataset from FBI Uniform Crime Reports using SAS to get meaning insight on identifying approach to reduce the crime case. The information collected from the dataset are presented into graph, map or table format for better visualization. The given report can be later used by other respective parties to identify a better solution on reducing the crime case in different state.

The analysis from this report support the preliminary finding from the law enforcement agencies which showed the increase of violent crime in the country. There is lacking information from this report as the original dataset is not attached with the cause of each individual crime happened in the state. By identifying the types of cause, it will be easy to come out with an approach to overcome the issue.

All the copies of this assignment report, program and data can be found in https://github.com/swhong87/DAP.

# **References**

*Census Population*. (2016). Retrieved from Census: https://www2.census.gov/programs-surveys/popest/datasets/2010-2015/cities/totals

*Census State Code*. (2013). Retrieved from Census: https://www2.census.gov/geo/docs/reference

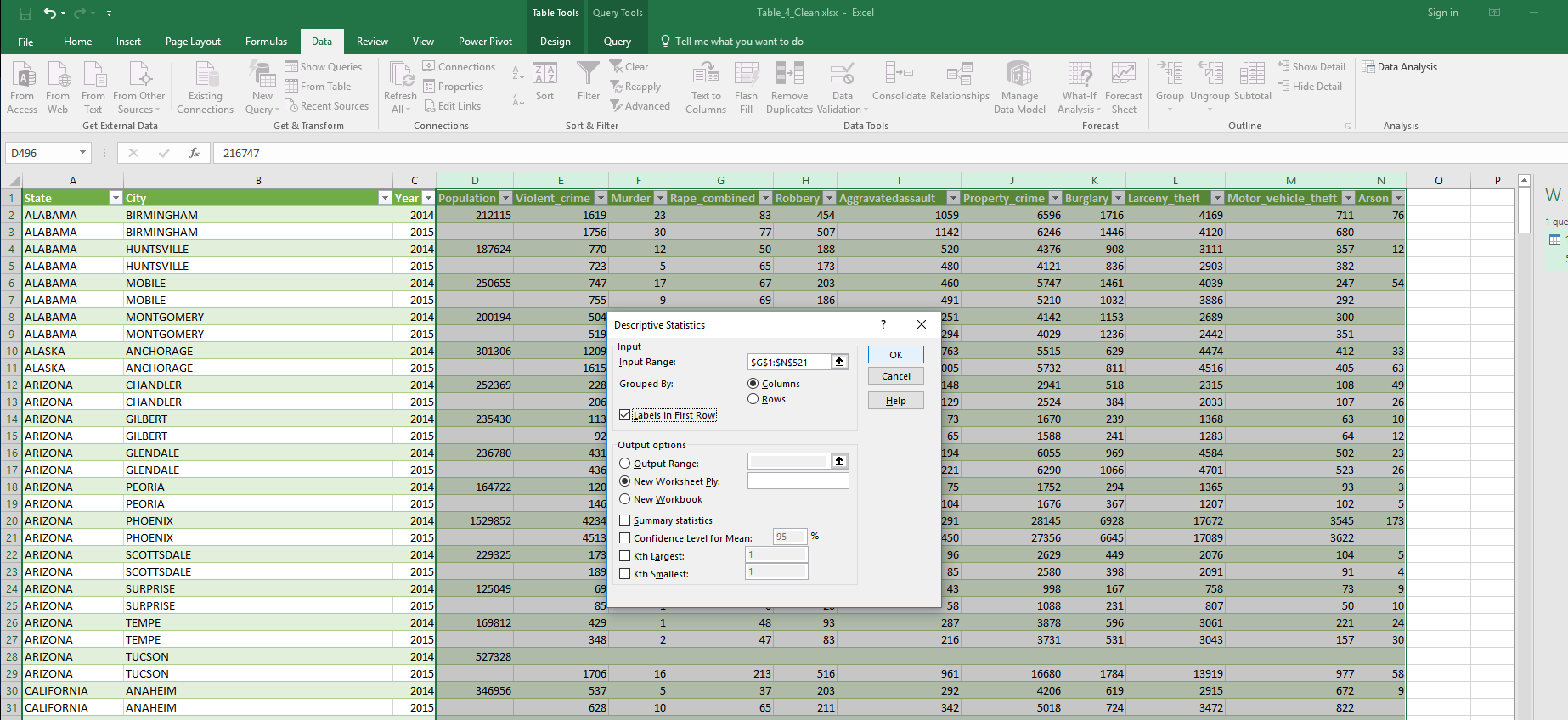
O'Neil, C. S. (2014). *Doing Data Science.* O'Reilly.

Yiran Dong, C.-Y. J. (2013). *Principled missing data methods for researchers*. Retrieved from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/

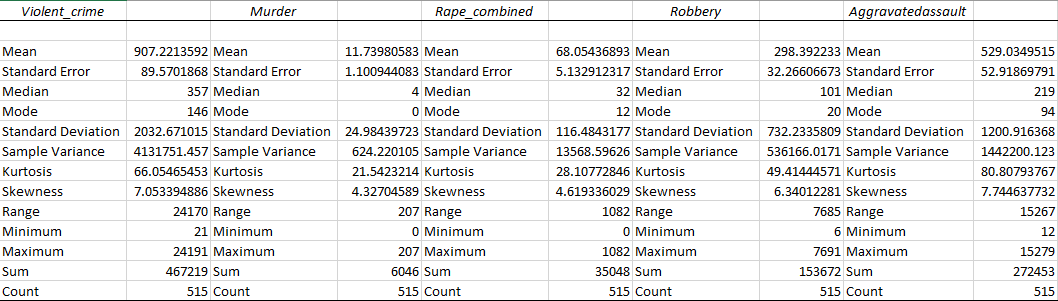
# **Appendix**

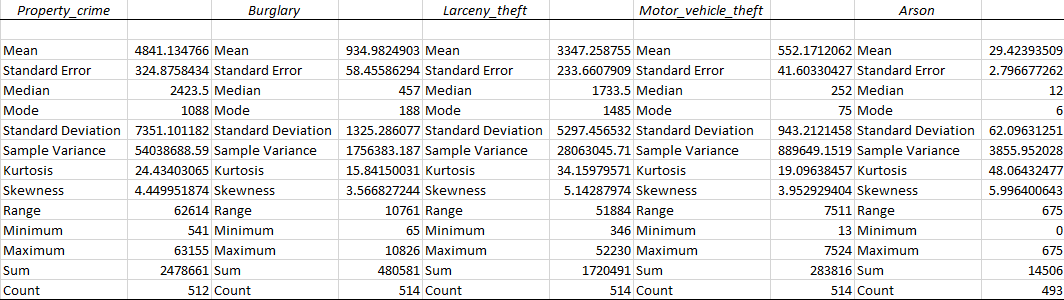
## Appendix 1 – Descriptive Statistic in Excel

Descriptive statistic is a feature provided in Excel that can be used to extract summary statistic about the dataset. It allowed us to find the mean, mod, count and others related information for this dataset. By clicking on Data tab, there is an option for Data Analysis.



By considering into the summary statistic of this dataset showed on the following figure, we can see that all the crime variables are having null value as none of them is giving the total count of 520 which is our records size.





## Appendix 2 – Import Population Data

/\*READ ESTIMATE POPULATION PER CITY INTO SAS\*/

DATA CITYPOP;

INFILE '/home/tp0453630/CrimeAnalysisAssignment/PopulationEstimate.csv' delimiter=',' FIRSTOBS=2;

LENGTH STNAME $40. NAME $40. FUNCSTAT $1. CENSUS2010POP $40.;

INPUT SUMLEV STATE COUNTY PLACE COUSUB

CONCIT PRIMGEO\_FLAG FUNCSTAT NAME STNAME

CENSUS2010POP ESTIMATESBASE2010 POPESTIMATE2010 POPESTIMATE2011

POPESTIMATE2012 POPESTIMATE2013 POPESTIMATE2014 POPESTIMATE2015;

KEEP STNAME NAME POPESTIMATE2014 POPESTIMATE2015;

\*EXCLUDE ROWS WITH PT. AS SUFFIX.THESE ROWS ARE BREAKDOWN RECORDS;

IF NOT FIND(NAME,'(PT.)','I') > 0;

RUN;

## Appendix 3 – Import Table 4

/\*READ TABLE 4 INTO SAS\*/

DATA MASTER;

INFILE '/home/tp0453630/CrimeAnalysisAssignment/Table\_4\_Clean.csv'

DELIMITER=',' DSD FIRSTOBS=2 MISSOVER;

LENGTH STATE $40. CITY $40.;

INPUT STATE CITY YEAR POPULATION VIOLENTCRIME MURDER

SEXUALHARASSMENT ROBBERY AGGRAVATEDASSAULT

PROPERTYCRIME BURGLARY LARCENYTHEFT MOTORVEHICLETHEFT ARSON;

RUN;

## Appendix 4 – Join Table 4 with Population

DATA CITYPOP\_KEY(KEEP=NAME STNAME COMBINEDKEY POPESTIMATE2014 POPESTIMATE2015);

SET CITYPOP;

COMBINEDKEY = UPCASE(SCAN(STNAME,1) !! SCAN(NAME,1));

RUN;

PROC SORT DATA=CITYPOP\_KEY NODUPKEY OUT=CITYPOP\_KEY\_DEDUP;

BY COMBINEDKEY NAME STNAME POPESTIMATE2014 POPESTIMATE2015;

RUN;

DATA MASTER\_KEY;

SET MASTER;

COMBINEDKEY = SCAN(STATE,1) !! SCAN(CITY,1);

RUN;

PROC SQL;

CREATE TABLE WORK.MASTER\_POP

AS

SELECT A.\*, B.NAME, STNAME, POPESTIMATE2014, POPESTIMATE2015

FROM MASTER\_KEY\_SORT A LEFT JOIN CITYPOP\_KEY\_DEDUP B

ON A.COMBINEDKEY = B.COMBINEDKEY

;

DATA WORK.MASTER\_POP\_REMOVE\_WORD (DROP=REMOVECITY1 REMOVECITY2);

SET WORK.MASTER\_POP;

REMOVECITY1=SUBSTR(NAME,1,INDEX(UPCASE(NAME), 'CITY')-1);

COMBINEDNAME1=UPCASE(SCAN(REMOVECITY1,1) !! SCAN(REMOVECITY1,2));

REMOVECITY2=SUBSTR(CITY,1,INDEX(UPCASE(CITY), 'CITY')-1);

COMBINEDNAME2=UPCASE(SCAN(REMOVECITY2,1) !! SCAN(REMOVECITY2,2));

RUN;

PROC SQL;

CREATE TABLE WORK.MATCH\_MASTER\_POP\_NAME

AS

SELECT \*

FROM WORK.MASTER\_POP\_REMOVE\_WORD

WHERE COMBINEDNAME1 = COMBINEDNAME2

;

PROC EXPORT DATA=WORK.master\_pop dbms=CSV OUTFILE="/home/tp0453630/CrimeAnalysisAssignment/Table\_4\_Population.csv" REPLACE;

RUN;

## Appendix 5 – Upload File for Analysis and Create Permenant Dataset

/\* Generated Code (IMPORT) \*/

/\* Source File: Table\_4\_For\_Analysis.xlsx \*/

/\* Source Path: /home/tp0453630/CrimeAnalysisAssignment \*/

/\* Code generated on: 19/04/2017 22:26 \*/

%web\_drop\_table(WORK.IMPORT);

FILENAME REFFILE '/home/tp0453630/CrimeAnalysisAssignment/Table\_4\_For\_Analysis.xlsx';

PROC IMPORT DATAFILE=REFFILE

DBMS=XLSX

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

%web\_open\_table(WORK.IMPORT);

LIBNAME CRIME '/home/tp0453630/CrimeAnalysisAssignment';

DATA CRIME.MASTER;

SET WORK.IMPORT;

RUN;

## Appendix 6 – Upload Statecode file for Analysis and Create Permenant Dataset

/\* Generated Code (IMPORT) \*/

/\* Source File: StateCode.xlsx \*/

/\* Source Path: /home/tp0453630/CrimeAnalysisAssignment \*/

/\* Code generated on: 22/04/2017 19:14 \*/

%web\_drop\_table(WORK.IMPORT);

FILENAME REFFILE '/home/tp0453630/CrimeAnalysisAssignment/StateCode.xlsx';

PROC IMPORT DATAFILE=REFFILE

DBMS=XLSX

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

%web\_open\_table(WORK.IMPORT);

LIBNAME CRIME '/home/tp0453630/CrimeAnalysisAssignment';

DATA CRIME.STATECODE;

SET WORK.IMPORT;

RUN;

## Appendix 7 – Merge MASTER and STATECODE using PROC SQL

LIBNAME CRIME '/home/tp0453630/CrimeAnalysisAssignment';

PROC SQL;

CREATE TABLE WORK.MERGE\_STATECODE

AS

SELECT B.STUSAB, B.STATE AS STATEID, A.\*

FROM CRIME.MASTER A, CRIME.STATECODE B

WHERE A.STATE = UPCASE(B.STATE\_NAME)

;

RUN;

## Appendix 8 – Total Crime Per State on Year 2014 and 2015

PROC SQL;

CREATE TABLE WORK.TEMP

AS

SELECT STATE, YEAR, STUSAB, STATEID,

SUM(PROPERTYCRIME) AS TOTAL\_PROPERTYCRIME,

SUM(VIOLENTCRIME) AS TOTAL\_VIOLENTCRIME,

SUM(PROPERTYCRIME+VIOLENTCRIME) AS TOTAL\_CRIME

FROM WORK.MERGE\_STATECODE

GROUP BY STATE, YEAR, STUSAB, STATEID

ORDER BY STATE, YEAR, 5 ASC

;

options orientation=landscape;

ODS PDF options orientation=landscape;

ODS PDF file='/home/tp0453630/CrimeAnalysisAssignment/TotalCrimePerState.pdf';

ods graphics on / width=12in;

goptions reset=all cback=white border htitle=12pt htext=10pt HPOS=100 VPOS=100;

TITLE 'Total Crime Per State for Year 2014 and 2015';

PROC SGPLOT DATA=WORK.TEMP;

VBAR STATE /RESPONSE=TOTAL\_CRIME GROUP=YEAR GROUPDISPLAY=CLUSTER GROUPORDER=DATA BARWIDTH=0.5

;

RUN;

ODS PDF close;

## Appendix 9 – Total Crime Changed Per State for Year 2014 and 2015 table view

options orientation=landscape;

ODS rtf file='/home/tp0453630/CrimeAnalysisAssignment/TotalCrimeChangedPerState.rtf';

ods graphics on / width=12in;

PROC SQL;

CREATE TABLE WORK.TEMP

AS

SELECT A.STATE AS STATE\_NAME, A.STUSAB, A.STATEID AS STATE,

A.TOTAL\_PROPERTYCRIME AS TOTAL\_PROPERTYCRIME\_2014,

B.TOTAL\_PROPERTYCRIME AS TOTAL\_PROPERTYCRIME\_2015,

A.TOTAL\_VIOLENTCRIME AS TOTAL\_VIOLENTCRIME\_2014,

B.TOTAL\_VIOLENTCRIME AS TOTAL\_VIOLENTCRIME\_2015,

A.TOTAL\_CRIME AS TOTAL\_CRIME\_2014,

B.TOTAL\_CRIME AS TOTAL\_CRIME\_2015,

((B.TOTAL\_PROPERTYCRIME - A.TOTAL\_PROPERTYCRIME) / A.TOTAL\_PROPERTYCRIME)

AS TOTAL\_PROPERTYCRIME\_CHANGED FORMAT PERCENT8.2,

((B.TOTAL\_VIOLENTCRIME - A.TOTAL\_VIOLENTCRIME) / A.TOTAL\_VIOLENTCRIME)

AS TOTAL\_VIOLENTCRIME\_CHANGED FORMAT PERCENT8.2,

((B.TOTAL\_CRIME - A.TOTAL\_CRIME) / A.TOTAL\_CRIME) AS TOTAL\_CRIME\_CHANGED FORMAT PERCENT8.2

FROM

(SELECT \* FROM WORK.TEMP WHERE YEAR = 2014)A

LEFT JOIN (SELECT \* FROM WORK.TEMP WHERE YEAR = 2015)B

ON A.STATE = B.STATE

ORDER BY TOTAL\_CRIME\_CHANGED

;

RUN;

PROC PRINT DATA=WORK.TEMP LABEL SPLIT='\*';

LABEL TOTAL\_PROPERTYCRIME\_2014='TOTAL\*PROPERTY\*CRIME\*2014'

TOTAL\_PROPERTYCRIME\_2015='TOTAL\*PROPERTY\*CRIME\*2015'

TOTAL\_VIOLENTCRIME\_2014='TOTAL\*VIOLENT\*CRIME\*2014'

TOTAL\_VIOLENTCRIME\_2015='TOTAL\*VIOLENT\*CRIME\*2015'

TOTAL\_CRIME\_2014='TOTAL\*CRIME\*2014'

TOTAL\_CRIME\_2015='TOTAL\*CRIME\*2015'

TOTAL\_PROPERTYCRIME\_CHANGED='TOTAL\*PROPERTY\*CRIME\*CHANGED'

TOTAL\_VIOLENTCRIME\_CHANGED='TOTAL\*VIOLENT\*CRIME\*CHANGED'

TOTAL\_CRIME\_CHANGED='TOTAL\*CRIME\*CHANGED'

;

RUN;

ODS rtf close;

## Appendix 10 – Total Crime Changed Per State for Year 2014 and 2015 map view

data center;

length function $ 8;

retain flag 0 xsys ysys '2' hsys '3' when 'a';

set maps.uscenter

(where=(fipstate(state) ne 'DC')

drop=long lat);

style = "'Albany AMT/bold'";

function='label';

text=fipstate(state);

size=2;

position='4';

if ocean='Y' then

do;

position='9';

output;

function='move';

flag=1;

end;

else if flag=1 then

do;

function='draw';

size=.25;

flag=0;

end;

output;

run;

goptions reset=all cback=white border htitle=12pt htext=10pt HPOS=100 VPOS=100 xpixels=1300 ypixels=768;

pattern1 v=s c=gray;

pattern2 v=s c=green;

pattern3 v=s c=brown;

pattern4 v=s c=yellowGreen;

pattern5 v=s c=yellow;

pattern6 v=s c=red;

TITLE 'Total Crime Changed Per State for Year 2014 and 2015';

PROC GMAP DATA=WORK.TEMP MAP=MAPS.US;

ID STATE;

CHORO TOTAL\_CRIME\_CHANGED /ANNOTATE=CENTER;

WHERE STUSAB NOT IN ('HI','UT','AZ');

RUN;

QUIT;

## Appendix 11 – Crime ratio based on population across all states for year 2014 and 2015

PROC SQL;

CREATE TABLE WORK.TEMP1

AS

SELECT STATE, YEAR, STUSAB, STATEID,

SUM(POPULATION) AS TOTAL\_POPULATION,

SUM(PROPERTYCRIME) AS TOTAL\_PROPERTYCRIME,

SUM(VIOLENTCRIME) AS TOTAL\_VIOLENTCRIME,

SUM(PROPERTYCRIME+VIOLENTCRIME) AS TOTAL\_CRIME

FROM WORK.MERGE\_STATECODE

GROUP BY STATE, YEAR, STUSAB, STATEID

ORDER BY STATE, YEAR

;

PROC SQL;

CREATE TABLE WORK.TEMP1

AS

SELECT A.STATE AS STATE\_NAME, A.STUSAB, A.STATEID AS STATE,

(A.TOTAL\_CRIME / A.TOTAL\_POPULATION) AS CRIME\_RATIO\_2014 FORMAT PERCENT8.2,

(B.TOTAL\_CRIME / B.TOTAL\_POPULATION) AS CRIME\_RATIO\_2015 FORMAT PERCENT8.2

FROM

(SELECT \* FROM WORK.TEMP1 WHERE YEAR = 2014)A

LEFT JOIN (SELECT \* FROM WORK.TEMP1 WHERE YEAR = 2015)B

ON A.STATE = B.STATE

ORDER BY 5

;

RUN;

options orientation=landscape;

ODS PDF file='/home/tp0453630/CrimeAnalysisAssignment/TotalCrimeRatioPerState.pdf';

ods graphics on / width=12in;

goptions reset=all cback=white border htitle=12pt htext=10pt HPOS=100 VPOS=100;

TITLE 'Crime Ratio in Population Per State for Year 2014 and 2015';

PROC SGPLOT DATA=WORK.TEMP1;

SERIES X=STATE\_NAME Y=CRIME\_RATIO\_2014;

SERIES X=STATE\_NAME Y=CRIME\_RATIO\_2015;

YAXIS LABEL="Crime Ratio";

WHERE STUSAB NOT IN ('HI','UT','AZ');

;

RUN;

ODS PDF close;

goptions reset=all cback=white border htitle=12pt htext=10pt HPOS=100 VPOS=100 xpixels=1300 ypixels=768 ;

pattern1 v=s c=gray;

pattern2 v=s c=green;

pattern3 v=s c=brown;

pattern4 v=s c=yellowGreen;

pattern5 v=s c=yellow;

pattern6 v=s c=red;

TITLE 'Crime Ratio Per State for Year 2014';

PROC GMAP DATA=WORK.TEMP1 MAP=MAPS.US;

ID STATE;

CHORO CRIME\_RATIO\_2014 /ANNOTATE=CENTER;

WHERE STUSAB NOT IN ('HI','UT','AZ');

RUN;

QUIT;

goptions reset=all cback=white border htitle=12pt htext=10pt HPOS=100 VPOS=100 xpixels=1300 ypixels=768 ;

pattern1 v=s c=gray;

pattern2 v=s c=green;

pattern3 v=s c=brown;

pattern4 v=s c=yellowGreen;

pattern5 v=s c=yellow;

pattern6 v=s c=red;

TITLE 'Crime Ratio Per State for Year 2015';

PROC GMAP DATA=WORK.TEMP1 MAP=MAPS.US;

ID STATE;

CHORO CRIME\_RATIO\_2015 /ANNOTATE=CENTER;

WHERE STUSAB NOT IN ('HI','UT','AZ');

RUN;

QUIT;

options orientation=portrait;

PROC PRINT DATA=TEMP1;

RUN;

## Appendix 12 – Individual Crime in Top 5 States with Highest Crime Ratio

PROC SQL;

CREATE TABLE WORK.INDIVIDUAL\_CRIME

AS

SELECT STATE, YEAR, STUSAB, STATEID,

SUM(MURDER) AS TOTAL\_MURDER,

SUM(SEXUALHARASSMENT) AS TOTAL\_SEXUALHARASSMENT,

SUM(ROBBERY) AS TOTAL\_ROBBERY,

SUM(AGGRAVATEDASSAULT) AS TOTAL\_AGGRAVATEDASSAULT,

SUM(BURGLARY) AS TOTAL\_BURGLARY,

SUM(LARCENYTHEFT) AS TOTAL\_LARCENYTHEFT,

SUM(MOTORVEHICLETHEFT) AS TOTAL\_MOTORVEHICLETHEFT

FROM WORK.MERGE\_STATECODE

WHERE STUSAB IN ('NM','MO','MD','MS','WA')

GROUP BY STATE, YEAR, STUSAB, STATEID

;

RUN;

PROC TABULATE DATA=WORK.INDIVIDUAL\_CRIME;

VAR TOTAL\_MURDER TOTAL\_SEXUALHARASSMENT TOTAL\_ROBBERY TOTAL\_AGGRAVATEDASSAULT TOTAL\_BURGLARY TOTAL\_LARCENYTHEFT TOTAL\_MOTORVEHICLETHEFT;

CLASS STATE YEAR;

TABLE STATE,

(TOTAL\_MURDER='MURDER')\*YEAR\*(SUM='Sum'\*F=COMMA7. colpctsum)

(TOTAL\_SEXUALHARASSMENT='SEXUALHARASSMENT')\*YEAR\*(SUM='Sum'\*F=COMMA7. colpctsum)

(TOTAL\_ROBBERY='ROBBERY')\*YEAR\*(SUM='Sum'\*F=COMMA7. colpctsum)

(TOTAL\_AGGRAVATEDASSAULT='AGGRAVATEDASSAULT')\*YEAR\*(SUM='Sum'\*F=COMMA7. colpctsum)

(TOTAL\_BURGLARY='BURGLARY')\*YEAR\*(SUM='Sum'\*F=COMMA7. colpctsum)

(TOTAL\_LARCENYTHEFT='LARCENYTHEFT')\*YEAR\*(SUM='Sum'\*F=COMMA7. colpctsum)

(TOTAL\_MOTORVEHICLETHEFT='MOTORVEHICLETHEFT')\*YEAR\*(SUM='Sum'\*F=COMMA7. colpctsum)

;

RUN;