

**DATA ANALYTICAL PROGRAMMING (DAP)**

**CT050-3-M**

**Individual Assignment**

SUBMITTED TO

Dr. Kalai Anand Ratnam

PREPARED BY:

NAME : FOO YEONG JIN

STUDENT ID : TP044538

INTAKE : UCMP1701DSBA

DATE : 15TH MAY 2017

Contents

[Acknowledgement 2](#_Toc482304407)

[Introduction 3](#_Toc482304408)

[Defining Target 4](#_Toc482304409)

[Data Acquisition 5](#_Toc482304410)

[Data Cleansing and Preparation 6](#_Toc482304411)

[Data Transformation 12](#_Toc482304412)

[Analysis 13](#_Toc482304413)

[Conclusion 22](#_Toc482304414)

[References 23](#_Toc482304415)

[Appendix 24](#_Toc482304416)

# Acknowledgement

I would like to express my deep gratitude to Dr. Kalai Anand Ratnam, my lecturer of this module (Data Analytical Programming), for his patient guidance and assistance from the very beginning of this assignment, without which the completion would not be possible. Apart from the guidance, Dr. Kalai also broadens my knowledge in the Data Science by giving a lot of example on how the Data Science works in the industry and prepare me for the bigger challenge.

I would also like to extend my thanks to fellow classmates, who contributed to this assignment directly or indirectly.

Finally, I wish to thank my wife for her support and encouragement throughout my study.

# Introduction

The Uniform Crime Reporting (UCR) program was form in 1929 by the Internal Association of Chief of Police in order to collect reliable data for the crime that reported to the law enforcement in the United States. Since 1930, The Federal Bureau of Investigation (FBI) is tasked to gather the crime related statistic for the Uniform Crime Reporting (UCR) from the law enforcement agencies in United States.

Today, the data that gathered through this program has enabled the Federal Bureau of Investigation (FBI) to produce four annual publications which includes Crime in the United States, National Incident-Based Reporting System, Law Enforcement Officers Killed and Assaulted, and Hate Crime Statistics. These publications were produced from data received from over 18,000 city, university and college, county, state, tribal, and federal law enforcement agencies voluntarily participating in the program. In addition to the publications, the data collected through the Uniform Crime Reporting (UCR) also enabled the law enforcement analyst, students, researchers, media, and the public to analyze the crimes in the United States.

In this assignment, I am responsible to analyze the crime data of the United States together with the members of the data science team in Federal Bureau of Investigation (FBI) by using the data collected from the Uniform Crime Reporting (UCR) mentioned above.

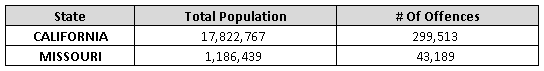
For this assignment, our team would like to achieve the following objectives:

The objective:

1. Analyze and understand the overall crime rate across United States.
2. Top 3 types of crimes in United States.
3. Identify the top 3 states with highest crime rate in year 2014 and their crime rate in year 2015
4. Identify the top 3 states with lowest crime rate year 2014 and their crime rate in year 2015
5. Compare the Police Employee Ratio in the Top 3 States with lowest crime rate (Objective 4) and the top 3 states with highest crime rate (Objective 3) in Year 2014.
6. Compare the crime rate for the states close to Mexico border and the states close to Canada border.

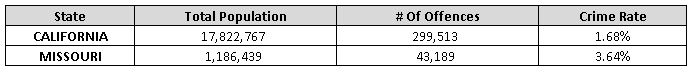
# Defining Target

In this assignment, crime rate will be used as the target instead of the absolute number of offences/crimes. This is because the number of offences is direct correlate to the population size (linear relationship), which means the bigger the population the higher the number of offences. Comparing the number of offences between 2 selected samples can be misleading.



\*Data retrieve from table 4

Based on the example above, if we measure the target by number of offences, it seems like California is more than 5 times more dangerous than Missouri. However, this is not true because if we measure by the crime rate, it shows that the probability of experiencing an offence/crime is much higher in Missouri.



\*Data retrieve from table 4

The basis of crime rate is to measure the probability of experiencing an offence/crime in every 100 citizens and it calculated based on the formula below:

Crime Rate = # of Offences / Total Population

By using this approach, it can eliminate the population effect on the number of offences/crimes and reveal the actual security status of a state or city.

# Data Acquisition

There are 4 tables provided for this assignment, detail as follow:

**Table 1**

This table provides the percent change in offenses known to law enforcement, by population group, for the first 6 months of 2014 and 2015.

**Table 2**

This table provides the percent change in offenses known to law enforcement, by region, for the first 6 months of 2014 and 2015.

**Table 3**

This table provides the percent change in offenses known to law enforcement for the nation for the first 6 months in consecutive years from 2011 through 2015.

**Table 4**

This table provides the number of offenses known to law enforcement for the first 6 months of 2014 and 2015 in cities with populations of 100,000 and over.

The primary data that used for this analysis is the table 4, which contain the detail of offences including the cities, states and population. Additional data also being acquired from the Uniform Crime Reporting (UCR) and United States Census Bureau in order to achieve the objectives. Below are the details of the additional data:

1. Law enforcement employee (Table 78)

This data is retrieved from the Uniform Crime Reporting (UCR), it contains number of officers which is the police employee who has the arresting power and also the civilian employee by states and cities. In addition to the number of officers, this data also contain the population of each city.

1. State code, division and region data (Region)

This data is acquired from United States Census Bureau, it enables the team to map the cities and states to respected division and regions.

# Data Cleansing and Preparation

Data cleansing and preparation is a very important step in any kind of analysis. In fact, this step consume most of the time (~50% - 80%) in the whole project. Through this step, the analyst will have a detail understanding of the input data in term of the *Accuracy, Integrity, Cleanliness, Correctness, Completeness* and *Consistency* before we start the analysis in SAS.

In this project, we split the data cleansing and preparation in to two parts. The first part of data cleansing and preparation is done in Excel. The team prepares and cleans the table 4 and additional data acquired, then convert the tables into a structured table format to import to SAS. The details as follow:

The first step is to remove the unnecessary items and rename the columns in the table4, which includes:

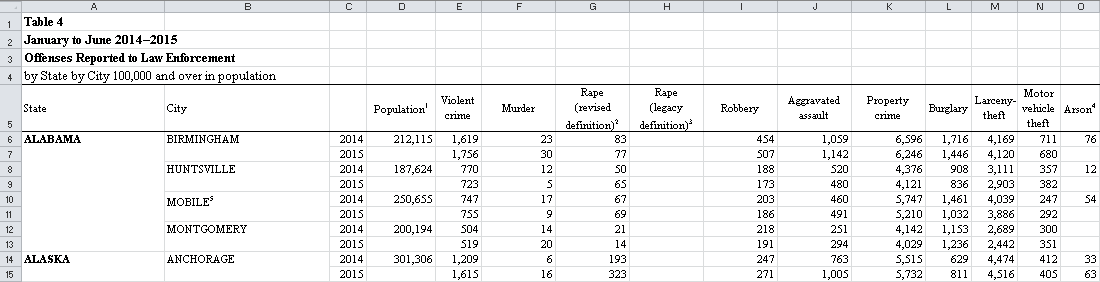


Figure 1: Table 4 header description

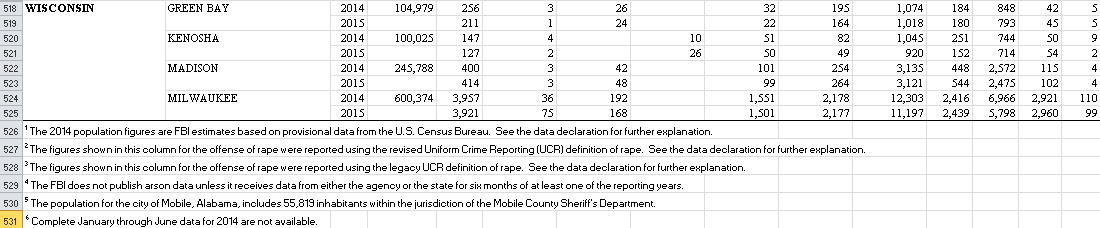


Figure 2: Table 4 footnote

* Remove the description of the table (highlighted in **blue** in figure 1) and footer (highlighted in **orange** in figure 2). The description is not in the structure format and will not be used in the analysis.
* Remove the special characters and remarks in the columns header, i.e. bracket and numbering (highlighted in **green** in figure 1).
* Unmerge the merged cells (highlighted in **red**) as SAS is not able to convert the merge cell properly during the import process. Without removing the merge cell, the output table will be inaccurate.
* Rename the columns name by replacing the space to underscore (‘\_’).

After remove the unnecessary items and rename the columns, the team then follow the steps in excel (Refer to: Appendix 1) to tidy up the table4. After all the steps, now the gaps are filled in with the correct information as below (Figure 3). We follow the same steps for the additional data and turn those data into a structured table.

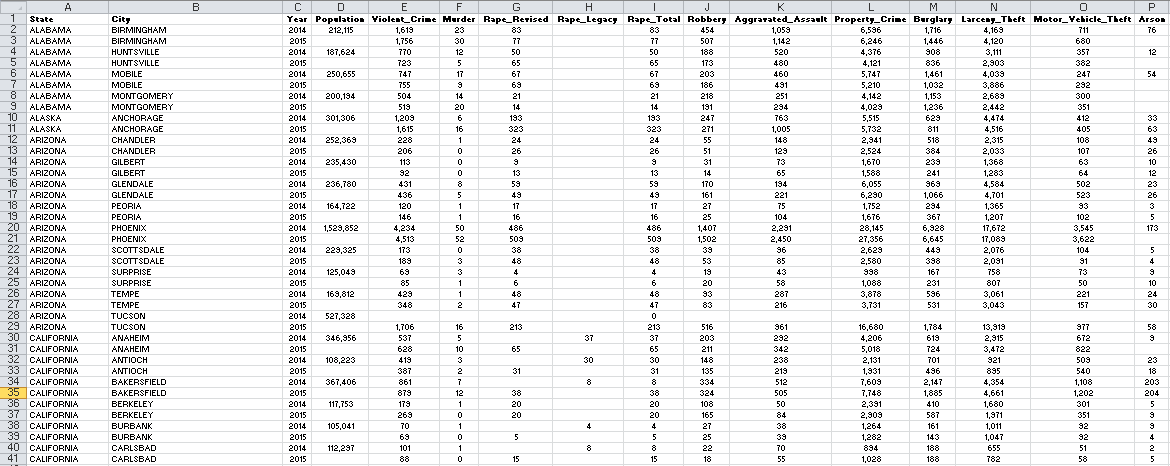


Figure 3: Table 4 output

4 excel tables are created through this process:

1. Table\_4\_Cleaned.xls
2. Region.xls
3. Table\_78\_2014\_Cleaned.xls
4. Table\_78\_2015\_Cleaned.xls

The second part of the data cleansing and preparation is done in the SAS. In this project, we are using the SAS® OnDemand for Academics, a web-based SAS development environment. 4 tables created in the part one were uploaded into the SAS® OnDemand for Academics using the code below. SAS macro was used in this step as 4 tables are using the same code to import.

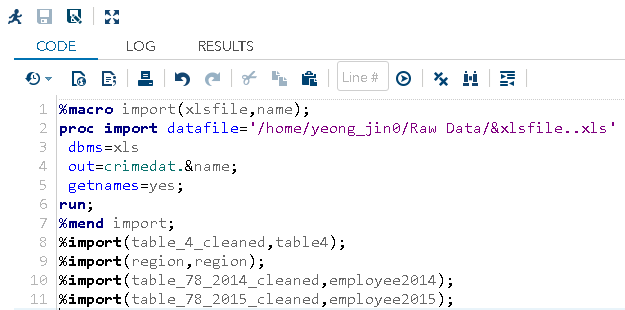


Figure 4: Import data

In this stage of cleansing and preparation, we zoom in into the data structure to ensure the data are imported correctly. The team is reviewing the result of the **Proc Contents** of the 4 tables to ensure the number of records, data type, length and formats are assigned correctly. (Refer to: Appendix 2)

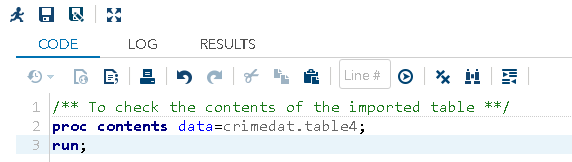


Figure 5: Sample code for proc contents



Figure 6: Proc contents output for Table4

The team then zooms in into the values inside the table4, which is the main table for this analysis. The acceptance level of the missing value in this project is set as 5% for each crime, anything more than 5% will be removed as it might impact the accuracy of the final result. Proc Means was applied to produce the output.

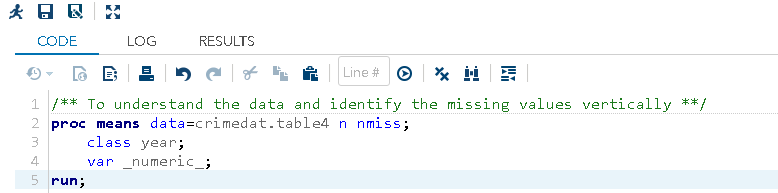


Figure 7: code for proc means

The next step is to access the missing value horizontally to review the missing crimes for each city. Simple transformation was used in this step to analyze the missing value for each city (Refer to appendix 4). In this assignment, anything more than 50% missing crime horizontally will be removed from the final analysis.

The last step is to verify the data consistency. At this stage, we’ve noticed that the population is missing for 2015. So, we will measure the consistency based on 2014 population (Refer to appendix 5).

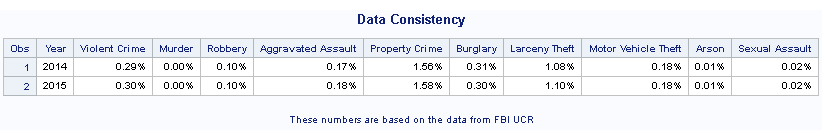


Figure 8: Consistency result

Based on the output above, the data is consistent across both years as there is no significant spike or drop in the percentage for every crime.

From the steps above, the team has summarized some observations which need to be addressed before we start the analysis.

1. Additional columns accidentally created due to the excel formatting during the proc import. These columns will be removed when the team prepare the final data for the analysis

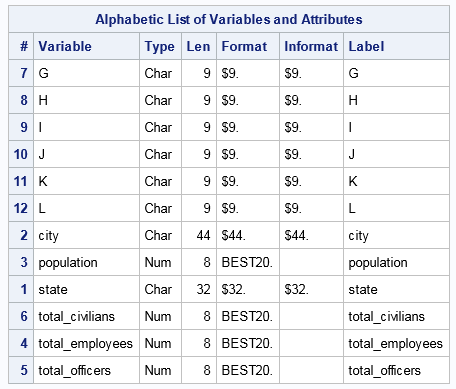


Figure 9: Proc contents output for table employee2015

1. Missing Population for all the cities in year 2015

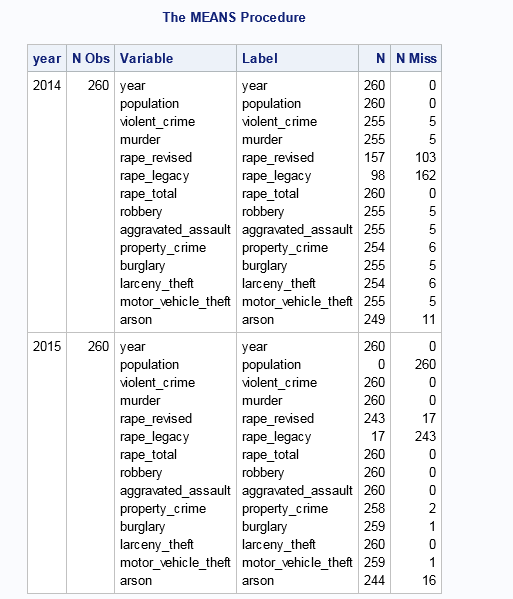


Figure 10: Proc means output for table 4

In table4, the populations are missing for all the cities. To estimate the population for year 2015, the team is using the information from United States Census Bureau below:

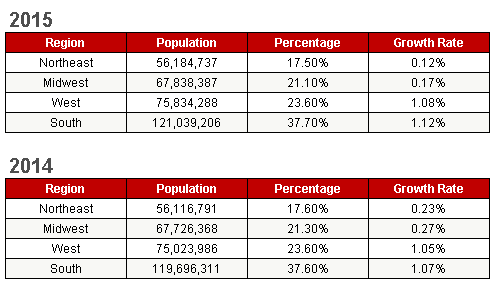


Figure 11: Based on the data publish in United States Census Bureau.

1. Both rape 2 types of rape in the report (legacy and revised)

Based on the output in the proc means above, the team identified high missing values in both rape categories. However, this is mainly due to the change of the rape definition in 2013. The team then decided to combine both rape categories in the transformation stage.

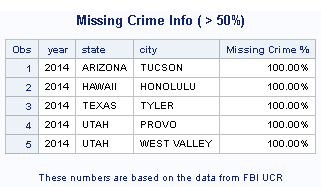
1. In year 2014, several cities are having missing records for all the crimes.  
   

Figure 12: Proc print output for missing crimes cities

The team decided to remove the cities with all missing crimes in 2014 as it will underestimate the state’s crime rate. To ensure the consistency, the same cities will also be removed from 2015.

1. During the process, the team found out that the property Crime in the table4 is the summation of burglary, larceny theft and motor-vehicle theft. However, 3 cities are having a blank value in the property crime. To solve this issue, we will drop the property crime in the table4 and re-create the property crime by sum up all the 3 types of crimes.
2. To create a total crime columns by sum up all the crime type.

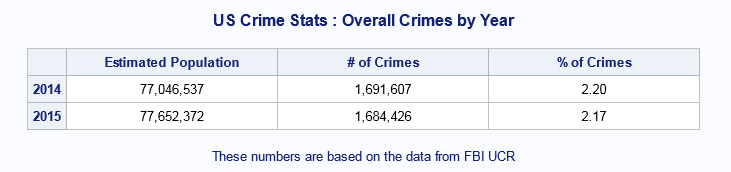
# Data Transformation

After the data review, the team proceeds to compile the final data for the analysis. The purpose is to prepare a dataset that incorporate all the changes on the finding above. In this stage, we also merge in the region data. Please refer to the appendix 6 for the detail.

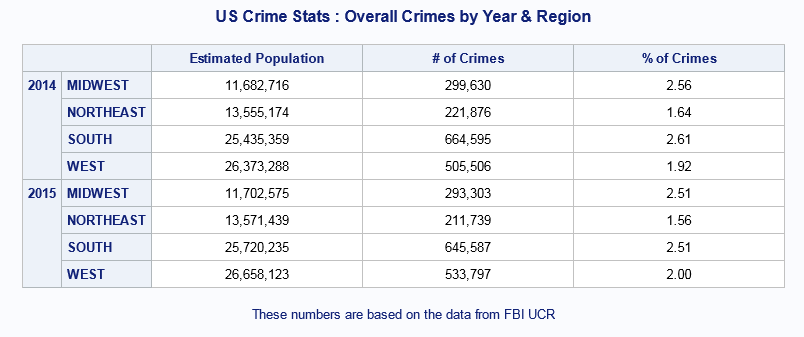
# Analysis

1. **Analyze and Understand the Overall Crime Rate (Excluding Arson) Across the United States.**

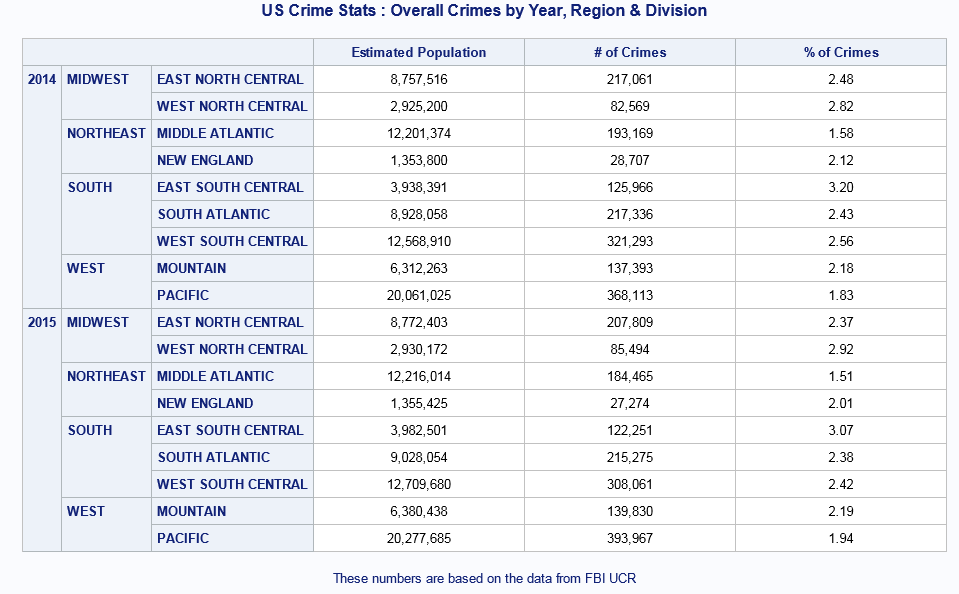
**(Excluding Tucson, Honolulu, Tyler, Provo and West Valley)**



Overall crime rate in United States is around 2.10%-2.20% per population. In year 2014, the overall crime rate is at 2.20% with the estimated population of 77.05 million. Compare to 2014, the overall crime rate for the year 2015 is further improved; it dropped from 2.20% to 2.17%. The improvement is not only come from the increase of the population, the number of crimes is also dropped compare to year 2014.



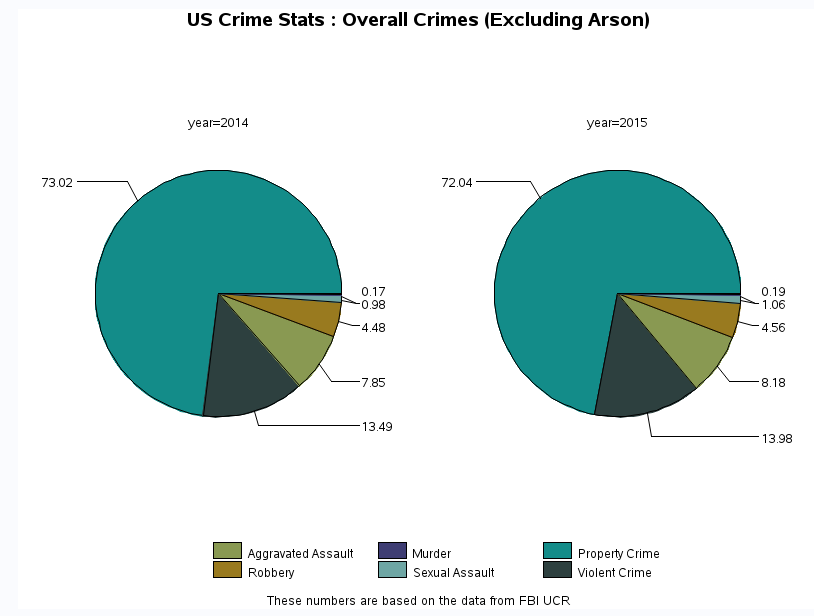
Although the overall crime rate is decreased in year 2015, the team is notice that there is a slight increase in the crime rate in the west side of United States (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, Washington). The crime rate of the West side of United States is increased from 1.92% in 2014 to 2.00% in 2015.



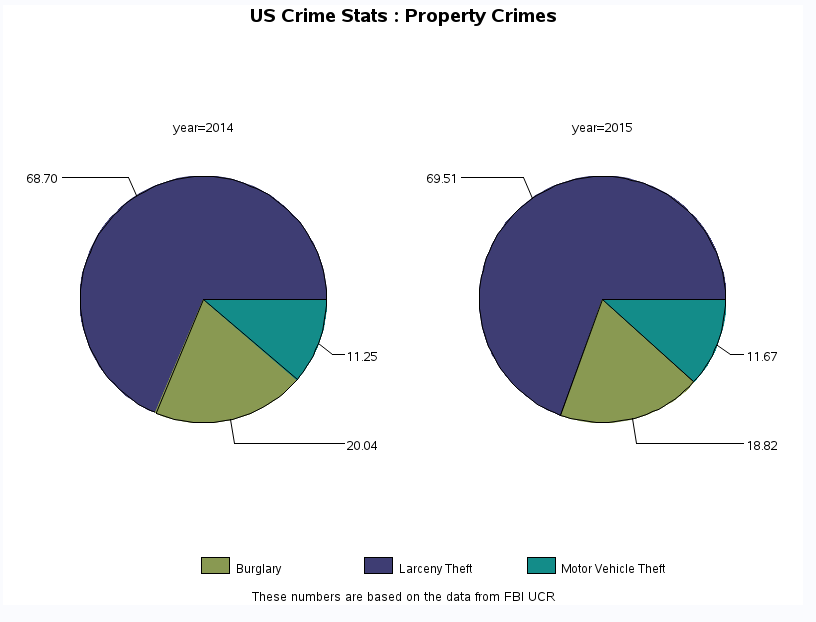
The team then further segment the data by the division and notice that the crime rate of 3 divisions across 2 regions are slightly increase compare the year 2014. The 3 divisions are West North Central in Midwest and all the divisions in the West.

\*Please refer to appendix 7 for the full code.

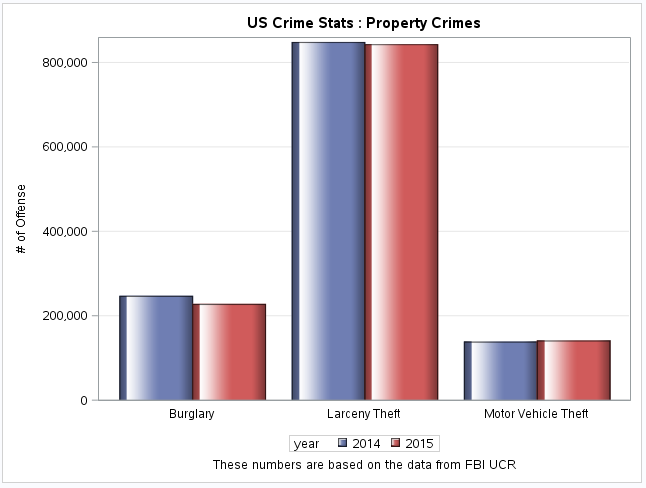
1. **Top 3 Types of Crime (Excluding Arson) in United States**.



Based on the data in table4, the top 3 type of crime in the United States are Property Crime, Violent Crime and Aggravated Assault. The distributions are 73.02%, 13.49% and 7.85% respectively in 2014. In 2015, the top 3 are remaining the same with a slight improvement in property crime (73.02% in 2014 vs. 72.04% in 2015).



As the property crime consists of burglary, larceny theft and motor vehicle theft, the team decided to zoom in to the property crime to see which type of sub-crime is causing improvement. From the chart above, the team is notice that the improvement in the property crime is mainly due to decrease in burglary.

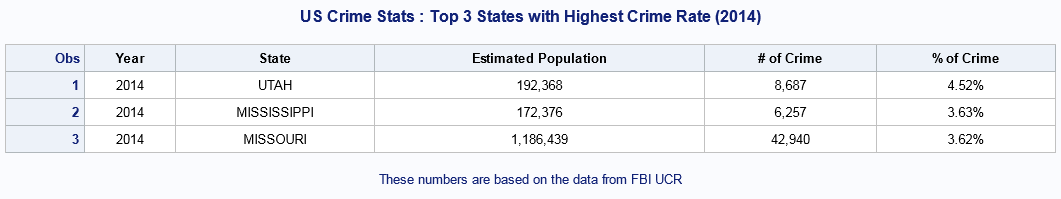


The improvement in burglary is not only on the distribution, the chart above shows that the number of offence is also lower compares to year 2014.

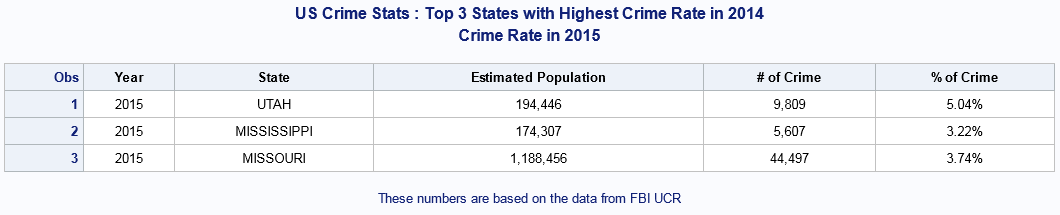
\*Please refer to appendix 8 for the full code.

1. **Identify the Top 3 States with Highest Crime Rate in Year 2014 and Their Crime Rate in Year 2015**

**(Excluding Tucson, Honolulu, Tyler, Provo and West Valley)**



The top 3 states with highest crime rate in year 2014 are Utah, Mississippi and Missouri. The crime rates are 4.52%, 3.63% and 3.62% respectively. The highest crime rate state, Utah, is more than 2X higher than the average crime rate in United States.

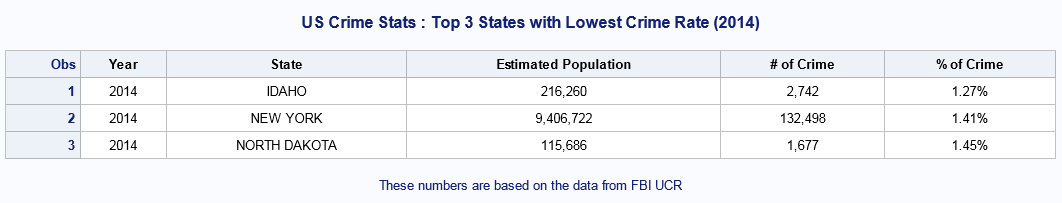


In year 2015, the crime rate for Utah is further increase to 5.04% even the overall crime rate in United States is improving. Mississippi is the only city in the top 3 that having an improvement in the crime rate. The crime rate for Mississippi improved from 3.63% in 2014 to 3.22% in 2015.

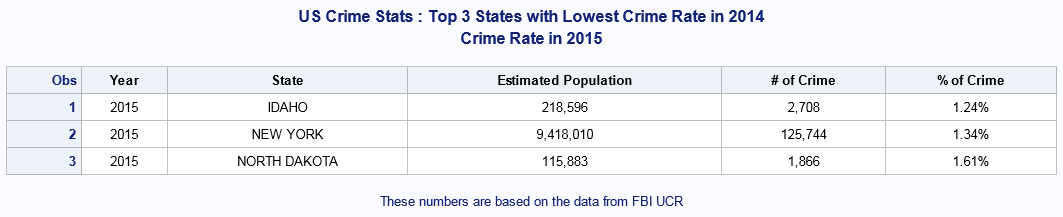
\*Please refer to appendix 9 for the full code.

1. **Identify the Top 3 States with Lowest Crime Rate in Year 2014 and Their Crime Rate in Year 2015**

**(Excluding Tucson, Honolulu, Tyler, Provo and West Valley)**



The top 3 states with lowest crime rate in year 2014 are Idaho, New York and North Dakota. The crime rates are 1.27%, 1.41% and 1.45% respectively. The lowest crime rate state, Idaho, is more than 3.5X lower than the crime rate of the highest crime rate state, Utah.



In year 2015, the crime rate for Idaho is further improved to 1.24%. New York is also further improved from 1.41% to 1.34%. North Dakota is the only city in the top 3 that having a deterioration in the crime rate. The crime rate for North Dakota increased from 1.45 % in 2014 to 1.61 % in 2015.

\*Please refer to appendix 10 for the full code.

1. **Compare the Police Employee Ratio of the Top 3 States with Lowest Crime Rate and the Top 3 States with Highest Crime Rate in Year 2014.**

The law enforcement employee data is retrieve from the Uniform Crime Reporting (UCR). In the raw data, the law enforcement employee is categorized into 2 which are:

1. Officer

Individuals who ordinarily carry a firearm and a badge, have full arrest powers, and are paid from governmental funds set aside specifically to pay sworn law enforcement.

1. Civilian

Include full-time agency personnel such as clerks, radio dispatchers, meter attendants, stenographers, jailers, correctional officers, and mechanics.

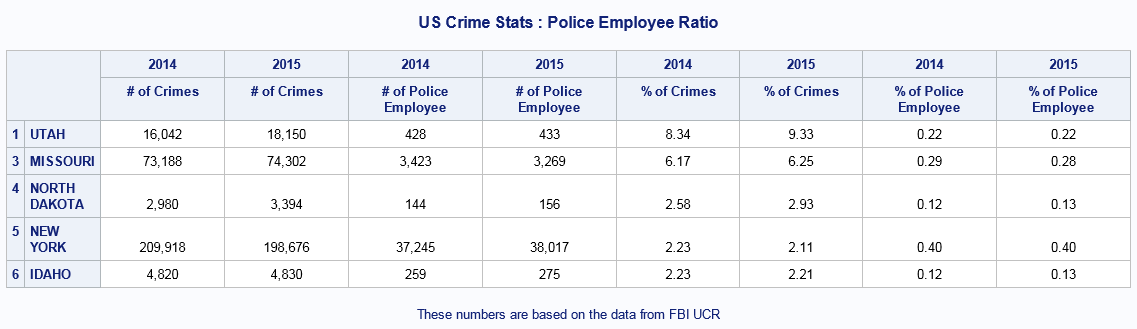
In the following analysis, only officer is considered as the civilian will not have any direct impact to the crimes. In addition, the state will be removed if the police employee data is missing or not available.

6 states below were selected for this analysis:

Top 3 highest crime rate states : Utah, Missouri and Mississippi

Top 3 lowest crime rate states : Idaho, New York and North Dakota

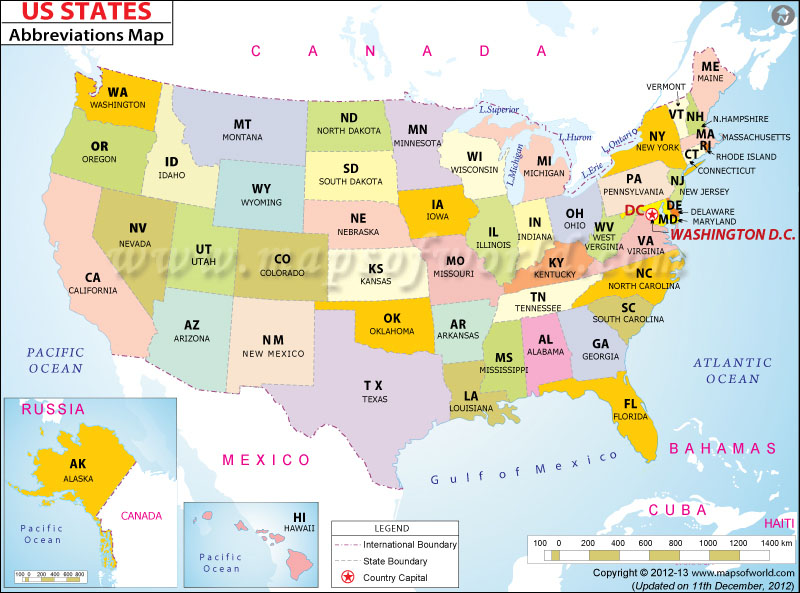
In the final report, Mississippi is excluded as the police employee data for the city Jackson is missing for year 2015 and this is the only city in Mississippi exceeds 100K population.



Based on the proc tabulate above, it shows that the police employee ratio for the highest crime rate states are generally more than the lowest crime rate states except for New York. New York states has highest police employee ratio within the 6 states.

\*Please refer to appendix 11 for the full code.

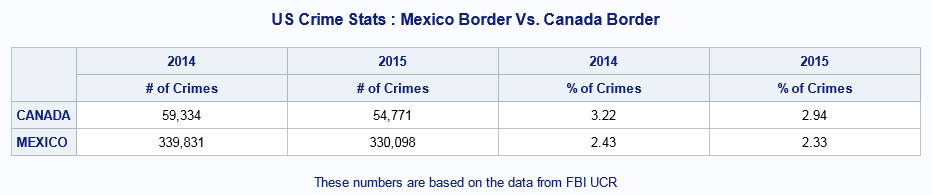
1. **Compare the Crime Rate for The States Close to Mexico Border and The States Close to Canada Border.**

****

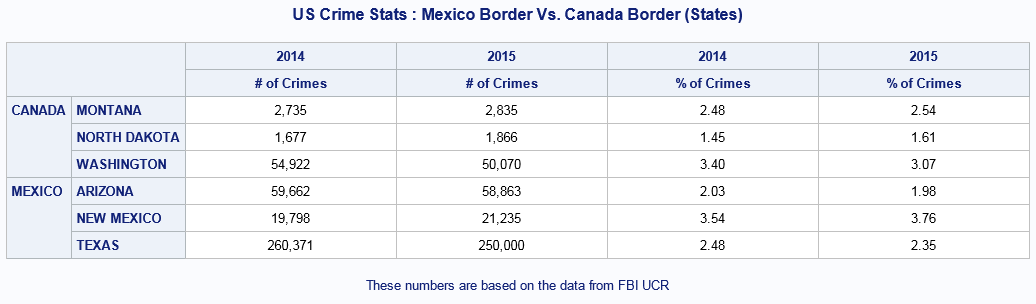
In this objective, 6 states are chosen for the comparison:

States close to Canada border : Washington, Montana and North Dakota

States close to Mexico border : Arizona, New Mexico and Texas



Surprisingly, the crime rate of the states close to Mexico border is actually lower than the states that close to the Canada border. The overall crime rate of the states close to Mexico is 2.33% while the states close to Canada is 2.97%.



However, when the team zooms into the states, we found that New Mexico is having the highest crime rate among the 6 states. The crime rate of New Mexico is 3.54% in year 2014 and increase to 3.76% in year 2015. The second highest crime rate state in this analysis is Washington. The crime rate for Washington is 3.40% in year 2014, and improved to 3.07% in year 2015.

\*Please refer to appendix 12 for the full code.

# Conclusion

Through this analysis, we understand that the overall crime rate of United States is in an improving trend. Overall crime rate has been reduced from 2.20% to 2.17%. This means that the current crime fighting strategy is performing well.

Property crime remains as the highest offence among all the crime types. However, we saw an improvement in burglary. We should identify the drivers that causing the improvement and try to implement for the larceny theft and motor vehicle theft to further reduce the property crime distribution.

As for the police employee, the data is not conclusive. However, New York is having a high police employee ratio (0.40%) and the crime rate is significant lower than the high crime rate states. In order to improve the crime rate for Utah, Missouri and Mississippi, local government can increase the police employee ratio to New York level if the states’ financial is allowed. The increase in police employee will surely help in reducing the crime rate.

Crime rate of the states close to Mexico border is lower than the states close to Canada border. However, the highest crime rate among the 6 states that were chosen for this analysis is New Mexico (Mexico Border). In this case, building a barrier across the states close to the Mexico border might not be necessary. Focus on the New Mexico might be a good alternative to the barrier.

# References

SAS 9.4 Programming Documentation. Available at: <http://documentation.sas.com/?cdcId=pgmmvacdc&cdcVersion=9.4&docsetId=basewn&docsetTarget=helpcenterfeedback.htm&locale=en>

SAS Support Communities . Available at:<https://communities.sas.com/>

FBI: Uniform Crime Reporting. Available at: <https://ucr.fbi.gov/>

United States Census Bureau. Available at: <https://www.census.gov/>

# Appendix

**Appendix 1: Excel Preparation**

**Step 1**

On the **Home** tab, in the **Editing** group, choose the **Find & Select** drop-down list and then click **Go To Special...**: and select the **Blanks** option and click **OK.**

**Step 2**

On the **Formula bar**, type an equal sign (=) followed by the address of the first cell with an entry in the column (=A6) and press **Ctrl + Enter**.

**Step 3**

Reselect the original range and press **Ctrl + C** to copy the selection. On the **Home** tab, in the **Clipboard** group, choose the **Paste** drop-down list and then click **Paste Values** to convert the formulas to values.

**Appendix 2: Proc Contents Codes**

/\*\* To check the contents of the imported table \*\*/

**proc** **contents** data=crimedat.table4;

**run**;

**proc** **contents** data=crimedat.region;

**run**;

**proc** **contents** data=crimedat.employee2014;

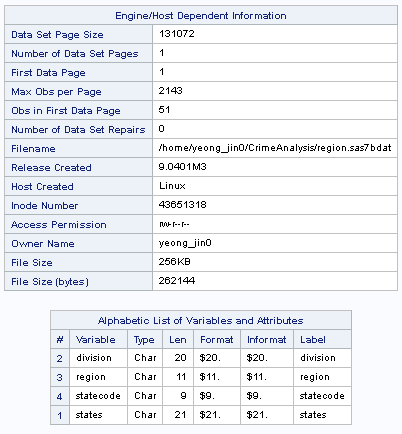
**run**;

**proc** **contents** data=crimedat.employee2015;

**run**;

**Appendix 3: Proc Contents Output**

Output for table Region



Output for table Employee2014



Output for table Employee2015



**Appendix 4: Identify the Missing Values Horizontally**

/\*\* To understand the data and identify the missing values horizontally \*\*/

**data** miss\_crimes;

set crimedat.table4;

/\*\* To combine 2 types of rape definition avoid the inaccurate missing value due to the change of definition \*\*/

if rape\_revised=**.** and rape\_legacy=**.** then sexual\_assault=**.**;

else

sexual\_assault=sum(rape\_revised,rape\_legacy);

drop rape\_revised rape\_legacy;

/\*\* To calcultate number of missing in each crime \*\*/

all\_crimes=sum(violent\_crime=**.**,murder=**.**,robbery=**.**,aggravated\_assault=**.**,property\_crime=**.**,burglary=**.**,larceny\_theft=**.**,

motor\_vehicle\_theft=**.**,arson=**.**,sexual\_assault=**.**);

/\*\* To derive the missing % \*\*/

crimes\_miss\_rate=all\_crimes/**10**;

/\*\* To output the row with missing %>=50% \*\*/

if crimes\_miss\_rate>=**0.5** then output;

**run**;

/\*\* To print the row with missing %>=50% \*\*/

**proc** **print** data=miss\_crimes label;

format crimes\_miss\_rate percent10.2;

label crimes\_miss\_rate='Missing Crime %';

var year state city crimes\_miss\_rate;

**run**;

**Appendix 5: Consistency Check**

/\*\* Summarized the data by year and crime \*\*/

**proc** **summary** data=work.all\_crimes nway;

class year;

var population violent\_crime murder robbery aggravated\_assault property\_crime burglary larceny\_theft motor\_vehicle\_theft arson sexual\_assault;

output out=work.summary\_all\_crimes(drop=\_freq\_ \_type\_) sum=;

**run**;

/\*\* To retain 2014 population and calculate the crime type \*\*/

**data** work.crime\_consistency;

length new\_population **8**;

set work.summary\_all\_crimes;

retain new\_population;

if population ne **.** then new\_population=population;

/\*\* To calculate the % of each crime type \*\*/

p\_violent\_crime = violent\_crime/new\_population;

p\_murder = murder/new\_population;

p\_robbery = robbery/new\_population;

p\_aggravated\_assault = aggravated\_assault/new\_population;

p\_property\_crime = property\_crime/new\_population;

p\_burglary = burglary/new\_population;

p\_larceny\_theft = larceny\_theft/new\_population;

p\_motor\_vehicle\_theft = motor\_vehicle\_theft/new\_population;

p\_arson = arson/new\_population;

p\_sexual\_assault = sexual\_assault/new\_population;

**run**;

/\*\* To print the consistency result \*\*/

title 'Data Consistency';

**proc** **print** data=work.crime\_consistency label;

label year = 'Year'

p\_violent\_crime = 'Violent Crime'

p\_murder = 'Murder'

p\_robbery = 'Robbery'

p\_aggravated\_assault = 'Aggravated assault'

p\_property\_crime = 'Property Crime'

p\_burglary = 'Burglary'

p\_larceny\_theft = 'Larceny Theft'

p\_motor\_vehicle\_theft = 'Motor Vehicle Theft'

p\_arson = 'Arson'

p\_sexual\_assault = 'Sexual Assault';

format p\_violent\_crime percent10.2

p\_murder percent10.2

p\_robbery percent10.2

p\_aggravated\_assault percent10.2

p\_property\_crime percent10.2

p\_burglary percent10.2

p\_larceny\_theft percent10.2

p\_motor\_vehicle\_theft percent10.2

p\_arson percent10.2

p\_sexual\_assault percent10.2;

var year p\_violent\_crime p\_murder p\_robbery p\_aggravated\_assault p\_property\_crime p\_burglary p\_larceny\_theft p\_motor\_vehicle\_theft p\_arson p\_sexual\_assault;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

**Appendix 6: Compile Data for Analysis**

/\*\* Finalized the final data set for analysis \*\*/

**data** work.table4\_tmp\_v1;

set crimedat.table4(drop=property\_crime);

/\*\* To remove the cities with >=50% missing crimes\*\*/

if (state='ARIZONA' and city='TUCSON') or

(state='HAWAII' and city='HONOLULU') or

(state='TEXAS' and city='TYLER') or

(state='UTAH' and city='PROVO') or

(state='UTAH' and city='WEST VALLEY') then delete;

/\*\* to combine 2 types of rape definition \*\*/

if rape\_revised=**.** and rape\_legacy=**.** then sexual\_assault=**.**;

else sexual\_assault=sum(rape\_revised,rape\_legacy);

property\_crime=sum(burglary,larceny\_theft,motor\_vehicle\_theft);

/\* To calculate the total crimes \*\*/

total\_crimes=sum(violent\_crime,murder,robbery,aggravated\_assault,

burglary,larceny\_theft,motor\_vehicle\_theft,

sexual\_assault);

drop rape\_revised rape\_legacy arson;

**run**;

/\*\* Merge the table4 with the divion, region and state code \*\*/

**proc** **sort** data=work.table4\_tmp\_v1;

by state;

**run**;

**proc** **sort** data=crimedat.region;

by states;

**run**;

**data** work.table4\_tmp\_v2;

merge work.table4\_tmp\_v1(in=a)

/\*\* Rename the column states to state for merging \*\*/

crimedat.region(in=b rename=(states=state));

by state;

if a;

**run**;

/\*\* Assign the estimated population to year 2015 \*\*/

**proc** **sort** data=work.table4\_tmp\_v2;

by state city year;

**run**;

**data** crimedat.table4\_final1;

set work.table4\_tmp\_v2;

by state city;

/\*\* Retain the population for year 2014 for the calculation \*\*/

retain new\_population;

/\*\* Assign the 2014 population to the new column new\_population \*\*/

if first.city then new\_population=population;

/\*\* Based on the 2014 population and region to estimate the 2015 population \*\*/

else do;

if region='NORTHEAST' then new\_population=round(new\_population+(new\_population\***0.0012**));

else if region='MIDWEST' then new\_population=round(new\_population+(new\_population\***0.0017**));

else if region='WEST' then new\_population=round(new\_population+(new\_population\***0.0108**));

else if region='SOUTH' then new\_population=round(new\_population+(new\_population\***0.0112**));

end;

label state = 'State Name'

city = 'City'

year = 'Year'

violent\_crime = 'Violent Crime'

murder = 'Murder'

robbery = 'Robbery'

aggravated\_assault = 'Aggravated Assault'

property\_crime = 'Property Crime'

burglary = 'Burglary'

larceny\_theft = 'Larceny Theft'

motor\_vehicle\_theft = 'Motor Vehicle Theft'

sexual\_assault = 'Sexual Assault'

total\_crimes = 'Total Crimes';

/\*\* New\_population will be used for the following analysis and drop the original column \*\*/

drop population;

**run**;

**Appendix 7: Objective 1**

/\*\* Objective 1 \*\*/

ods html body='/home/yeong\_jin0/ODS Output/ Overall United States.htm' style=HTMLBlue;

title 'US Crime Stats : Overall Crimes by Year';

**proc** **tabulate** data=crimedat.table4\_final1 S=[foreground=black cellwidth=**200** just=c];

class year;

var new\_population total\_crimes;

table year='', new\_population='Estimated Population'\*sum=''\*f=comma14. total\_crimes='# of Crimes'\*sum=''\*f=comma14. total\_crimes='% of Crimes'\*pctsum<new\_population>=''\*f=**8.2**;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

title 'US Crime Stats : Overall Crimes by Year & Region';

**proc** **tabulate** data=crimedat.table4\_final1 S=[foreground=black cellwidth=**200** just=c];

class year region;

var new\_population total\_crimes;

table year=''\*region='', new\_population='Estimated Population'\*sum=''\*f=comma14. total\_crimes='# of Crimes'\*sum=''\*f=comma14. total\_crimes='% of Crimes'\*pctsum<new\_population>=''\*f=**8.2**;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

title 'US Crime Stats : Overall Crimes by Year, Region & Division';

**proc** **tabulate** data=crimedat.table4\_final1 S=[foreground=black cellwidth=**200** just=c];

class year region division;

var new\_population total\_crimes;

table year=''\*region=''\*division='', new\_population='Estimated Population'\*sum=''\*f=comma14. total\_crimes='# of Crimes'\*sum=''\*f=comma14. total\_crimes='% of Crimes'\*pctsum<new\_population>=''\*f=**8.2**;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

ods html close;

**Appendix 8: Objective 2**

/\*\* Objective 2 \*\*/

/\*\* To summarized the crime type \*\*/

**proc** **summary** data=crimedat.table4\_final1 nway;

class year;

var violent\_crime murder robbery aggravated\_assault property\_crime burglary larceny\_theft motor\_vehicle\_theft sexual\_assault total\_crimes;

output out=summary\_year(drop=\_freq\_ \_type\_) sum=;

**run**;

/\*\* To transpose the data \*\*/

**proc** **transpose** data=summary\_year out=table4\_2014\_trans;

where year=**2014**;

**run**;

**proc** **transpose** data=summary\_year out=table4\_2015\_trans;

where year=**2015**;

**run**;

/\*\* To rename the transposed data \*\*/

**data** table4\_2014\_trans\_v2;

set table4\_2014\_trans;

where \_label\_ not in ('Year','Total Crimes');

year=**2014**;

rename \_label\_ = crime\_type

col1 = no\_of\_offence;

label \_label\_ = 'Crime Type'

col1 = '# Offence';

drop \_name\_;

**run**;

**data** table4\_2015\_trans\_v2;

set table4\_2015\_trans;

where \_label\_ not in ('Year','Total Crimes');

year=**2015**;

rename \_label\_ = crime\_type

col1 = no\_of\_offence;

label \_label\_ = 'Crime Type'

col1 = '# Offence';

drop \_name\_;

**run**;

/\*\* To calculate the crime type distribution \*\*/

**proc** **freq** data=table4\_2014\_trans\_v2 noprint;

where crime\_type not in ('Burglary','Larceny Theft','Motor Vehicle Theft');

tables year\*crime\_type/out=table4\_2014\_trans\_v3;

weight no\_of\_offence;

**run**;

**proc** **freq** data=table4\_2015\_trans\_v2 noprint;

where crime\_type not in ('Burglary','Larceny Theft','Motor Vehicle Theft');

tables year\*crime\_type/out=table4\_2015\_trans\_v3;

weight no\_of\_offence;

**run**;

/\*\* To combine table 2014 and 2015 \*\*/

**data** table4\_1415\_trans\_v3;

set table4\_2014\_trans\_v3

table4\_2015\_trans\_v3;

**run**;

ods html path='/home/yeong\_jin0/ODS Output/'

body='Overall Crimes.html';

/\* Define the titles \*/

title 'US Crime Stats : Overall Crimes (Excluding Arson)';

/\* Define pattern color for each crime type \*/

pattern1 color=day; /\* Aggravated assault \*/

pattern2 color=mob; /\* Murder \*/

pattern3 color=vibg; /\* Property Crime \*/

pattern4 color=deoy; /\* Robbery \*/

pattern5 color=libg; /\* Violent Crime \*/

pattern6 color=dabg; /\* Sexual assault \*/

legend1 label=none

position=(bottom)

offset=(**4**,)

across=**3**

order=("Aggravated Assault" "Murder" "Property Crime" "Robbery" "Sexual Assault" "Violent Crime" )

value=(color=black)

shape=bar(**4**,**1.5**);

/\* Create the pie chart \*/

**proc** **gchart** data=table4\_1415\_trans\_v3;

pie crime\_type / sumvar=percent

descending

other=**0**

legend=legend1

value=none

across=**2**

value=arrow

coutline=black

noheading

group=Year;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

**quit**;

ods html close;

/\* To calculate the % within property crime \*/

**proc** **freq** data=table4\_2014\_trans\_v2 noprint;

where crime\_type in ('Burglary','Larceny Theft','Motor Vehicle Theft');

tables year\*crime\_type/out=table4\_2014\_trans\_propc;

weight no\_of\_offence;

**run**;

**proc** **freq** data=table4\_2015\_trans\_v2 noprint;

where crime\_type in ('Burglary','Larceny Theft','Motor Vehicle Theft');

tables year\*crime\_type/out=table4\_2015\_trans\_propc;

weight no\_of\_offence;

**run**;

**data** table4\_1415\_trans\_propc;

set table4\_2014\_trans\_propc

table4\_2015\_trans\_propc;

**run**;

ods html path='/home/yeong\_jin0/ODS Output/'

body='Property Crimes.html';

/\* Define the titles \*/

title 'US Crime Stats : Property Crimes';

/\* Define pattern color for each crime type \*/

pattern1 color=day; /\* Burglary \*/

pattern2 color=mob; /\* Larceny Theft \*/

pattern3 color=vibg; /\* Motor Vehicle Theft \*/

legend1 label=none

position=(bottom)

offset=(**4**,)

across=**3**

order=("Burglary" "Larceny Theft" "Motor Vehicle Theft" )

value=(color=black)

shape=bar(**4**,**1.5**);

/\* Create the pie chart \*/

**proc** **gchart** data=table4\_1415\_trans\_propc;

pie crime\_type / sumvar=percent

descending

other=**0**

legend=legend1

value=none

across=**2**

value=arrow

coutline=black

noheading

group=Year;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

**quit**;

/\*\* To create bar chart \*\*/

title 'US Crime Stats : Property Crimes';

**proc** **sgplot** data=work.Table4\_1415\_trans\_propc;

format COUNT comma10.;

label COUNT = '# of Offense';

vbar crime\_type / response=COUNT group=year groupdisplay=cluster

stat=sum dataskin=gloss;

xaxis display=(nolabel noticks);

yaxis grid;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

ods html close;

**Appendix 9: Objective 3**

/\*\* Objective 3 \*\*/

/\*\* To summarize the population and total crimes by year and states\*\*/

**proc** **summary** data=crimedat.table4\_final1 nway;

class year state;

var new\_population total\_crimes;

output out=summary\_crime\_population(drop=\_freq\_ \_type\_) sum=;

**run**;

/\*\* To select only 2014 \*\*/

**data** crimehigh\_population\_2014;

set summary\_crime\_population;

where year=**2014**;

crime\_rate=total\_crimes/new\_population;

**run**;

**proc** **sort** data=crimehigh\_population\_2014;

by descending crime\_rate;

**run**;

ods html path='/home/yeong\_jin0/ODS Output/'

body='Top 3 Highest Crimes.html';

/\*\* To print the result for the Top 3 \*\*/

title 'US Crime Stats : Top 3 States with Highest Crime Rate (2014)';

**proc** **print** data=crimehigh\_population\_2014(obs=**3**) label style(header)={just=c foreground=black}

style(table)={width=**100**%};

format new\_population comma14.

total\_crimes comma14.

crime\_rate percent7.2;

label year = 'Year'

state = 'State'

new\_population = 'Estimated Population'

total\_crimes = '# of Crime'

crime\_rate = '% of Crime';

var year / style(data)={just=c};

var state / style(data)={just=c};

var new\_population / style(data)={just=c};

var total\_crimes / style(data)={just=c};

var crime\_rate / style(data)={just=c};

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

/\*\* To select only the top 3 2014 in the 2015 data \*\*/

/\*\* To summarize the population and total crimes by year and states\*\*/

**data** crimehigh\_population\_2015;

set summary\_crime\_population;

where year=**2015** and state in ('UTAH','MISSISSIPPI','MISSOURI');

if state='UTAH' then rank=**1**;

else if state='MISSISSIPPI' then rank=**2**;

else rank=**3**;

crime\_rate=total\_crimes/new\_population;

**run**;

**proc** **sort** data=crimehigh\_population\_2015;

by rank;

**run**;

/\*\* To print the result for the Top 3 \*\*/

title1 'US Crime Stats : Top 3 States with Highest Crime Rate in 2014';

title2 'Crime Rate in 2015';

**proc** **print** data=crimehigh\_population\_2015(obs=**3**) label style(header)={just=c foreground=black}

style(table)={width=**100**%};

format new\_population comma14.

total\_crimes comma14.

crime\_rate percent7.2;

label year = 'Year'

state = 'State'

new\_population = 'Estimated Population'

total\_crimes = '# of Crime'

crime\_rate = '% of Crime';

var year / style(data)={just=c};

var state / style(data)={just=c};

var new\_population / style(data)={just=c};

var total\_crimes / style(data)={just=c};

var crime\_rate / style(data)={just=c};

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

ods html close;

**Appendix 10: Objective 4**

/\*\* Objective 4 \*\*/

**proc** **summary** data=crimedat.table4\_final1 nway;

class year state;

var new\_population total\_crimes;

output out=summary\_crime\_population(drop=\_freq\_ \_type\_) sum=;

**run**;

/\*\* To select only 2014 \*\*/

**data** crimelow\_population\_2014;

set summary\_crime\_population;

where year=**2014**;

crime\_rate=total\_crimes/new\_population;

**run**;

**proc** **sort** data=crimelow\_population\_2014;

by crime\_rate;

**run**;

ods html path='/home/yeong\_jin0/ODS Output/'

body='Top 3 Lowest Crimes.html';

/\*\* To print the result for the Top 3 \*\*/

title 'US Crime Stats : Top 3 States with Lowest Crime Rate (2014)';

**proc** **print** data=crimelow\_population\_2014(obs=**3**) label style(header)={just=c foreground=black}

style(table)={width=**100**%};

format new\_population comma14.

total\_crimes comma14.

crime\_rate percent7.2;

label year = 'Year'

state = 'State'

new\_population = 'Estimated Population'

total\_crimes = '# of Crime'

crime\_rate = '% of Crime';

var year / style(data)={just=c};

var state / style(data)={just=c};

var new\_population / style(data)={just=c};

var total\_crimes / style(data)={just=c};

var crime\_rate / style(data)={just=c};

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

/\*\* To select only the top 3 2014 in the 2015 data \*\*/

**data** crimelow\_population\_2015;

set summary\_crime\_population;

where year=**2015** and state in ('IDAHO','NEW YORK','NORTH DAKOTA');

if state='IDAHO' then rank=**1**;

else if state='NEW YORK' then rank=**2**;

else rank=**3**;

crime\_rate=total\_crimes/new\_population;

**run**;

**proc** **sort** data=crimelow\_population\_2015;

by rank;

**run**;

/\*\* To print the result for the Top 3 \*\*/

title1 'US Crime Stats : Top 3 States with Lowest Crime Rate in 2014';

title2 'Crime Rate in 2015';

**proc** **print** data=crimelow\_population\_2015(obs=**3**) label style(header)={just=c foreground=black}

style(table)={width=**100**%};

format new\_population comma14.

total\_crimes comma14.

crime\_rate percent7.2;

label year = 'Year'

state = 'State'

new\_population = 'Estimated Population'

total\_crimes = '# of Crime'

crime\_rate = '% of Crime';

var year / style(data)={just=c};

var state / style(data)={just=c};

var new\_population / style(data)={just=c};

var total\_crimes / style(data)={just=c};

var crime\_rate / style(data)={just=c};

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

ods html close;

**Appendix 11: Objective 5**

/\*\* Objective 5 \*\*/

/\*\* Merge the law enforcement data to the dataset \*\*/

**data** employee2014;

set crimedat.employee2014;

/\*\* the city in table4 is upper case \*\*/

city=upcase(city);

/\*\* create a year for the merging \*\*/

year=**2014**;

/\*\* drop the duplicate or additional columns \*\*/

drop population;

**run**;

**data** employee2015;

set crimedat.employee2015;

/\*\* the data in table4 is upper case \*\*/

city=upcase(city);

/\*\* create a year for the merging \*\*/

year=**2015**;

/\*\* drop the duplicate or additional columns \*\*/

drop population g h i j k l;

**run**;

**proc** **sort** data=employee2014 out=employee2014\_tmp(keep=state city);

by state city;

**run**;

**proc** **sort** data=employee2015 out=employee2015\_tmp(keep=state city);

by state city;

**run**;

/\*\* Create a lookup table and only select the cities that have 2 years of records \*\*/

**data** employee\_lookup;

merge employee2014\_tmp(in=a)

employee2015\_tmp(in=b);

by state city;

if a and b;

**run**;

/\*\* Combine 2014 and 2015 law enforcement table \*\*/

**data** total\_employee;

set employee2014

employee2015;

**run**;

**proc** **sort** data=total\_employee;

by state city;

**run**;

/\*\* Select the cities that have 2 years of records based on the lookup table \*\*/

**data** total\_employee\_final;

merge total\_employee(in=a)

employee\_lookup(in=b);

by state city;

if a and b;

**run**;

**proc** **sort** data=total\_employee\_final;

by state city year;

**run**;

**proc** **sort** data=crimedat.table4\_final1 out=table4\_final1\_tmp;

by state city year;

**run**;

/\*\* Merge with the transformed table4 and create the final data for analysis \*\*/

**data** crimedat.table4\_final2;

merge table4\_final1\_tmp(in=a)

total\_employee\_final(in=b);

by state city year;

if a and b;

**run**;

/\*\* Summarized the population total officers total crimes by year and state \*\*/

**proc** **summary** data=crimedat.table4\_final2 nway;

class year state;

var new\_population total\_officers total\_crimes;

output out=summary\_police\_population(drop=\_freq\_ \_type\_) sum=;

**run**;

/\*\* Select the highest top 3 and lowest top 3 states\*\*/

**data** police\_population;

set summary\_police\_population;

if state in ('UTAH','MISSISSIPPI','MISSOURI','IDAHO','NEW YORK','NORTH DAKOTA');

if state='UTAH' then rank=**1**;

else if state='MISSISSIPPI' then rank=**2**;

else if state='MISSOURI' then rank=**3**;

else if state='NORTH DAKOTA' then rank=**4**;

else if state='NEW YORK' then rank=**5**;

else rank=**6**;

**run**;

**proc** **sort** data=police\_population;

by rank year;

**run**;

ods html path='/home/yeong\_jin0/ODS Output/'

body='Poloce.html';

/\*\* Print the result \*\*/

title 'US Crime Stats : Police Employee Ratio';

**proc** **tabulate** data=police\_population S=[foreground=black cellwidth=**200** just=c];

class state rank year;

var new\_population total\_crimes total\_officers;

table rank=''\*state='',year=''\*total\_crimes='# of Crimes'\*sum=''\*f=comma14. year=''\*total\_officers='# of Police Employee'\*sum=''\*f=comma14. year=''\*total\_crimes='% of Crimes'\*pctsum<new\_population>=''\*f=**8.2** year=''\*total\_officers='% of Police Employee'\*pctsum<new\_population>=''\*f=**8.2**;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

ods html close;

**Appendix 12: Objective 6**

/\*\* Objective 6 \*\*/

/\*\* To summarize the population and total crimes by year and states\*\*/

**proc** **summary** data=crimedat.table4\_final1 nway;

where state in ('WASHINGTON', 'MONTANA', 'NORTH DAKOTA', 'ARIZONA', 'NEW MEXICO', 'TEXAS');

class year state;

var new\_population total\_crimes;

output out=summary\_crime\_border(drop=\_freq\_ \_type\_) sum=;

**run**;

/\*\* Assign the border to each states\*\*/

**data** crime\_border\_v2;

set summary\_crime\_border;

length border $6.;

if state in ('WASHINGTON', 'MONTANA', 'NORTH DAKOTA') then border='CANADA';

else border='MEXICO';

**run**;

ods html path='/home/yeong\_jin0/ODS Output/'

body='Border.html';

/\*\* Print the result \*\*/

title 'US Crime Stats : Mexico Border Vs. Canada Border';

**proc** **tabulate** data=crime\_border\_v2 S=[foreground=black cellwidth=**200** just=c];

class year border;

var new\_population total\_crimes;

table border='',year=''\*total\_crimes='# of Crimes'\*sum=''\*f=comma14. year=''\*total\_crimes='% of Crimes'\*pctsum<new\_population>=''\*f=**8.2** ;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

/\*\* Print the result \*\*/

title 'US Crime Stats : Mexico Border Vs. Canada Border (States)';

**proc** **tabulate** data=crime\_border\_v2 S=[foreground=black cellwidth=**200** just=c];

class year border state;

var new\_population total\_crimes;

table border=''\*state='',year=''\*total\_crimes='# of Crimes'\*sum=''\*f=comma14. year=''\*total\_crimes='% of Crimes'\*pctsum<new\_population>=''\*f=**8.2** ;

footnote "These numbers are based on the data from FBI UCR" ;

**run**;

ods html close;