DATA ANALYTICAL PROGRAMMING

CT050-3-M

ASSIGNMENT

CRIME ANALYSIS IN UNITED STATES

SUBMITTED TO

DR. KALAI ANAND

PREPARED BY

STUDENT NAME : NG WEN GE

STUDENT ID : TP042400

INTAKE CODE : UCMP1606DSBA

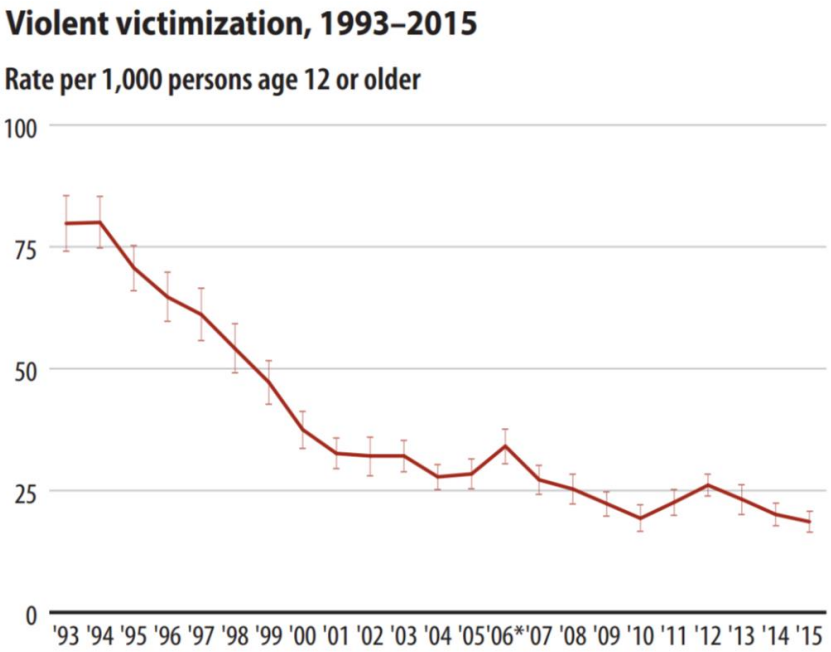
SUBMISSION ON

15/05/2017

Table of Contents

1. INTRODUCTION

Crime is an action of expressing violence and aggression to either individual or the society that violates the law and regulatory. According United States Bureau of Justice Statistics, nearly 5 million of criminal violent victimizations happened to United States residents aged 12 or older in 2015, with the crime rate of 18.6 per 1000 persons. Despite statistics indicated a huge decline in crime rate since the past decades, it is apparent that crime is inevitable in human society. It is the matter of minimizing it from happening, by countering criminals in an efficient manner.



With the continuous advancement of technology and integrated system, people managed to collect and integrate more data and make more reliable data-driven decisions, and that created hype of ‘Big Data’ and ‘Big Data Analytics’. Big data analytics have been commonly used in almost every domain today, as it facilitates the decision-making process with historical evidence. In efforts to counter crimes, big data analytics have also been widely used in studying criminals’ behaviour, recognizing the patterns and help avoiding it from happening. Up to certain extent, it also allows expertise to perform statistical modelling and make prediction on the occurrence of crime in a specific location.

2. PROBLEM STATEMENT

As a data scientist at the Headquarters of Federal Bureau of Investigations (FBI), Washington, D.C. in United States, I am responsible to explore the crime-related data and deep dive into diagnostic analytics to facilitate the crime investigation process. A comprehensive analysis of violent crime and property crime is required to present the criminal trend in United States of America and what can be done to facilitate the investigation process.

A preliminary data set from the Uniform Crime Reports of the FBI is provided to carry out the analysis. The preliminary data set contains the first 6 months (January to June) of 2014 and 2015 in cities with populations of 100,000 and over.

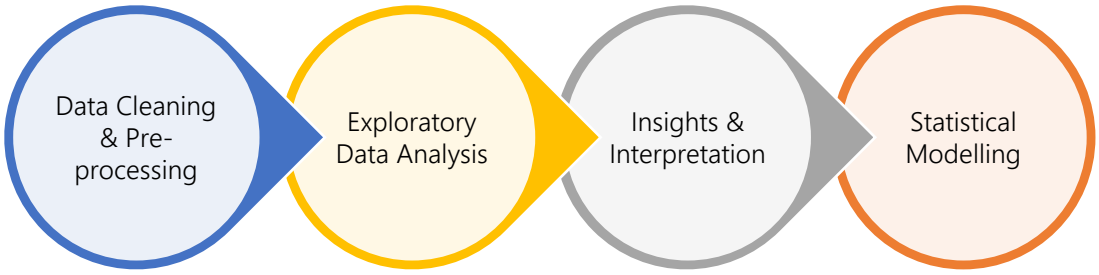
3. AIM & OBJECTIVE

The aim of the assignment is to perform a comprehensive analysis and reporting on the crimes in United States given the preliminary data in efforts to help reduce crime. Statistical Analysis System (SAS) which is a well-known statistical tool will be used to conduct the analysis. Followed by the primary aim of the assignment are the objectives as below:

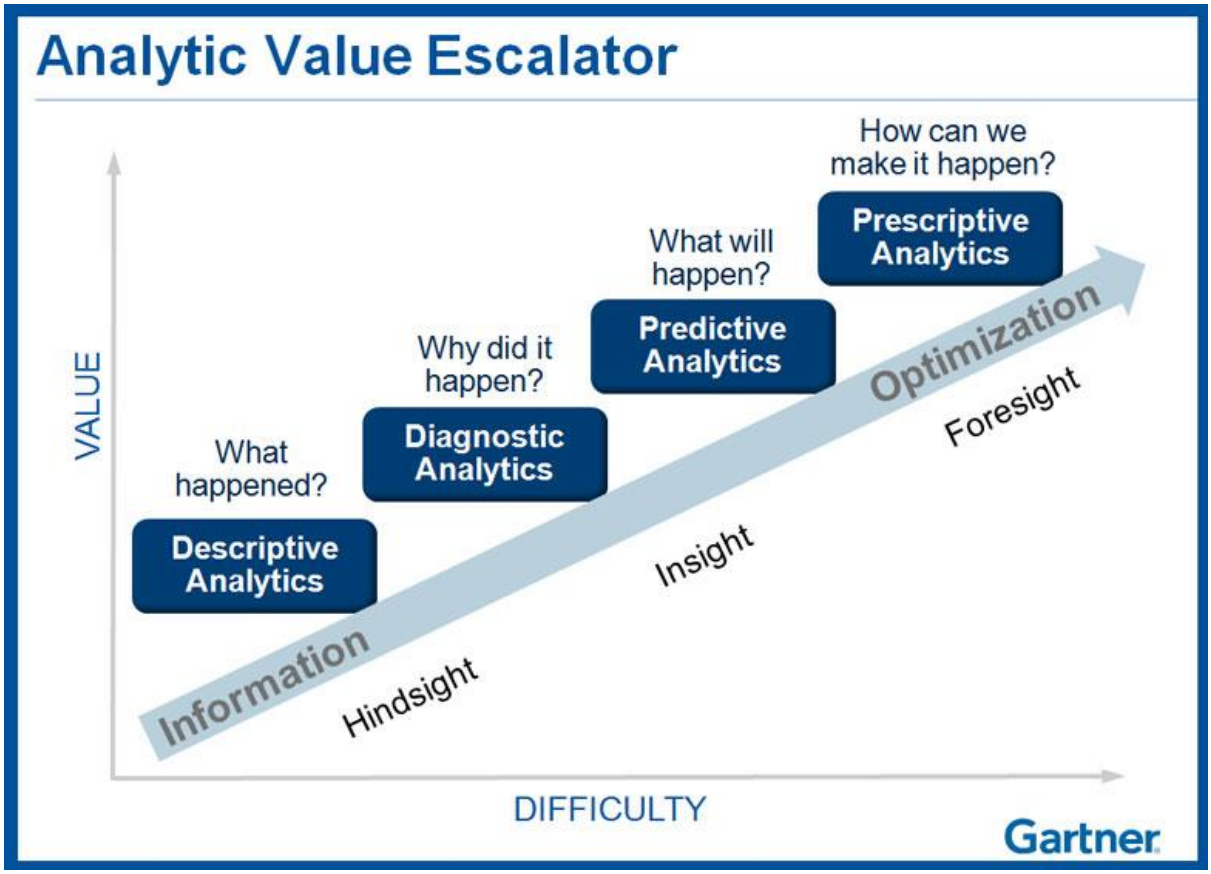
- ✓ Objective 1: To explore the overall crime trend in United States.
- ✓ Objective 2: To identify the top 3 states with highest and lowest crime rate.
- ✓ Objective 3: To analyze the crime details in the selected top states.
- ✓ Objective 4: To identify the top high/low risk city in the top states with highest/lowest crime rate.
- ✓ Objective 5: To analyze the crime details in the selected top cities.
- ✓ Objective 6: To analyze how poverty relates to property crime rate.
- ✓ Objective 7: To analyze how unemployment rate relates to overall crime rate.

4. METHODOLOGY

The basic process of big data analytics starts with data cleaning and data pre-processing right after acquiring the data. Most of the time, the raw data collected are not clean and ready for performing data analysis. Before data analysis and data modelling can be carried out, data scientist always spends significant time and efforts in cleaning the data, by imputing missing values, treating outliers in proper way, transforming data, etc. where applicable.



With structured and well-processed data, data can be explored via graphs, and analysis and interpretations can be done through the visualization This is commonly known as exploratory data analysis (EDA), which allows data scientist and data analyst to derive insights and recommendations. Essentially in this stage, there are 4 important phases of analytics, each describing different level of depth in analytics, they are descriptive, diagnostic, predictive and prescriptive analytics (as described by Gartner Analytics Maturity Model). Descriptive comes at the very first place as it gives the surface-level information about the data with the aids of descriptive statistics. Analysts typically explore the general information of the data and ask questions like what had happened regarding the problems in a particular domain. In this phase, descriptive statistics like total, mean, median, quartile, and range are commonly used to identify the trend described by the data. For example, the average crime rate in United States or the total crime rate trend in United States over the years can be explored. Data can be further discovered from different angle or dimension, addressing questions like which types of crime are averagely high in certain states in United States.



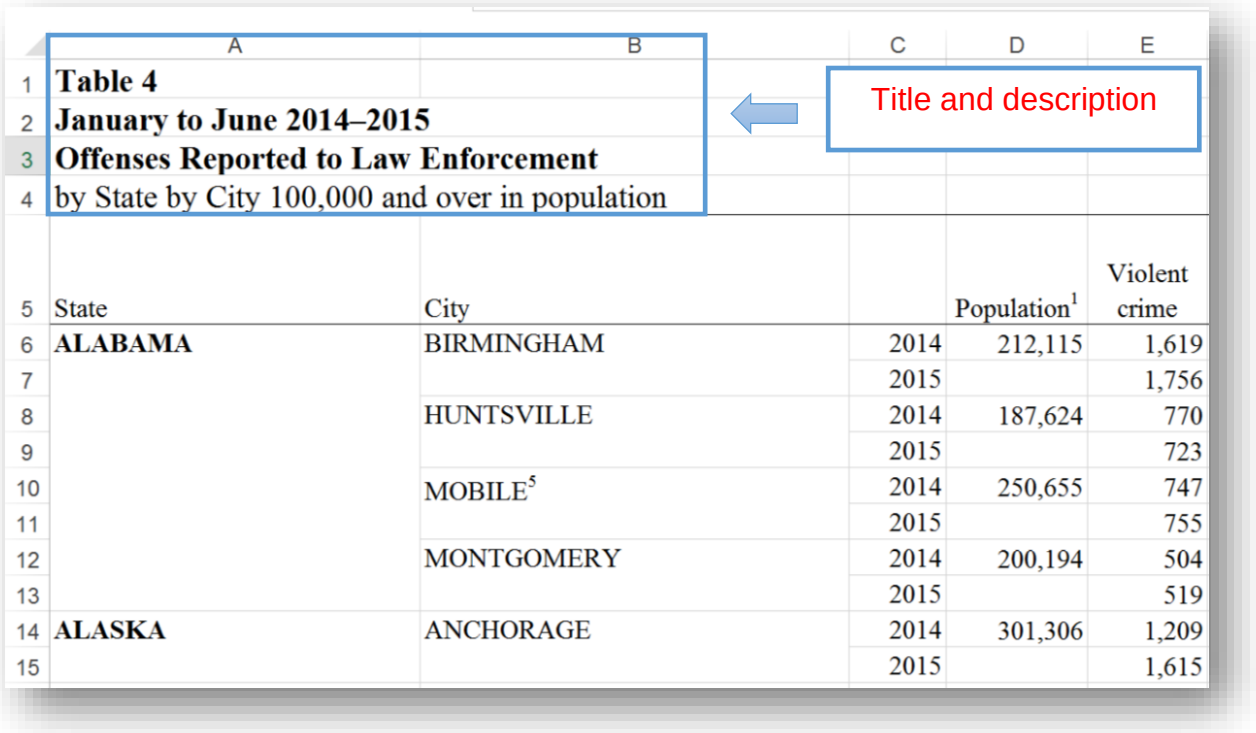
After getting some intuitions about the data, diagnostic analytics can be applied to deep dive into the data to discover the patterns within. The general outcome of this phase enables the capability to unveil the insights and explain why did it happen. For example, the property crime rate in a state in US was particularly high as there exists high poverty in the state too. For instance, correlation test plays important role to help recognize the pattern and relationship between variables.

On top of performing descriptive and diagnostic analytics that help understand the business situation and insights, subsequently, predictive analytics is the phase that can take business insights further and unveil the potential of data analytics to facilitate decision-making process and make better decision. This is also where the statistical and predictive modelling takes part. Predictive analytics are commonly categorized under advanced analytics as it enables foresights on some business scenarios or problems in advance. Essentially, a truly accurate prediction can result in effective optimization to the business, which is where prescriptive analytics take part in. In this assignment, only descriptive and diagnostic analytics will be covered, and further works on predictive and prescriptive analytics can be done with the acquisition of more comprehensive data.

4.1 Data Exploration in Excel

Before go deep into analysis, it is always a necessity to explore the data and verify if it is in a well-structured format before pumping into analytics tool like SAS. A well-structured format is the table format that contains rows as the data with columns as the fields for describing the data. For the given crime-related data in excel format, it is not in the well-structured table format. Let's identify the problems:

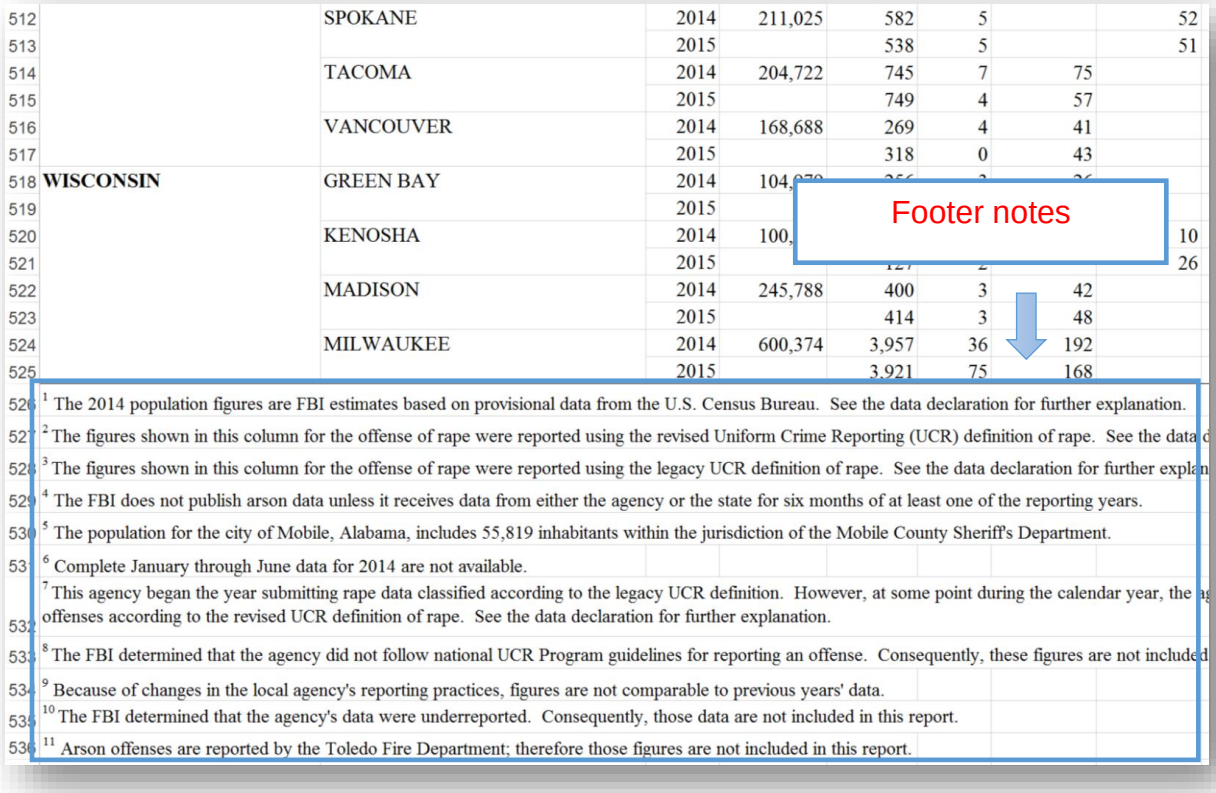
i. *Problem 1: Title and description above the headers*



	A	B	C	D	E
1	**Table 4**				
2	**January to June 2014–2015**				
3	**Offenses Reported to Law Enforcement**				
4	by State by City 100,000 and over in population				
5	State	City		Population[1]	Violent crime
6	**ALABAMA**	BIRMINGHAM	2014	212,115	1,619
7			2015		1,756
8		HUNTSVILLE	2014	187,624	770
9			2015		723
10		MOBILE[5]	2014	250,655	747
11			2015		755
12		MONTGOMERY	2014	200,194	504
13			2015		519
14	**ALASKA**	ANCHORAGE	2014	301,306	1,209
15			2015		1,615

A table format should contain only the header and the data itself. The title and data description should be removed before loaded into SAS.

ii. *Problem 2: Footer notes*



512		SPOKANE	2014	211,025	582	5		52
513			2015		538	5		51
514		TACOMA	2014	204,722	745	7	75	
515			2015		749	4	57	
516		VANCOUVER	2014	168,688	269	4	41	
517			2015		318	0	43	
518	**WISCONSIN**	GREEN BAY	2014	104,[illegible]				
519			2015					
520		KENOSHA	2014	100,				10
521			2015					26
522		MADISON	2014	245,788	400	3	42	
523			2015		414	3	48	
524		MILWAUKEE	2014	600,374	3,957	36	192	
525			2015		3,921	75	168	

[1] The 2014 population figures are FBI estimates based on provisional data from the U.S. Census Bureau. See the data declaration for further explanation.

[2] The figures shown in this column for the offense of rape were reported using the revised Uniform Crime Reporting (UCR) definition of rape. See the data d

[3] The figures shown in this column for the offense of rape were reported using the legacy UCR definition of rape. See the data declaration for further expla

[4] The FBI does not publish arson data unless it receives data from either the agency or the state for six months of at least one of the reporting years.

[5] The population for the city of Mobile, Alabama, includes 55,819 inhabitants within the jurisdiction of the Mobile County Sheriff's Department.

[6] Complete January through June data for 2014 are not available.

[7] This agency began the year submitting rape data classified according to the legacy UCR definition. However, at some point during the calendar year, the a offenses according to the revised UCR definition of rape. See the data declaration for further explanation.

[8] The FBI determined that the agency did not follow national UCR Program guidelines for reporting an offense. Consequently, these figures are not included

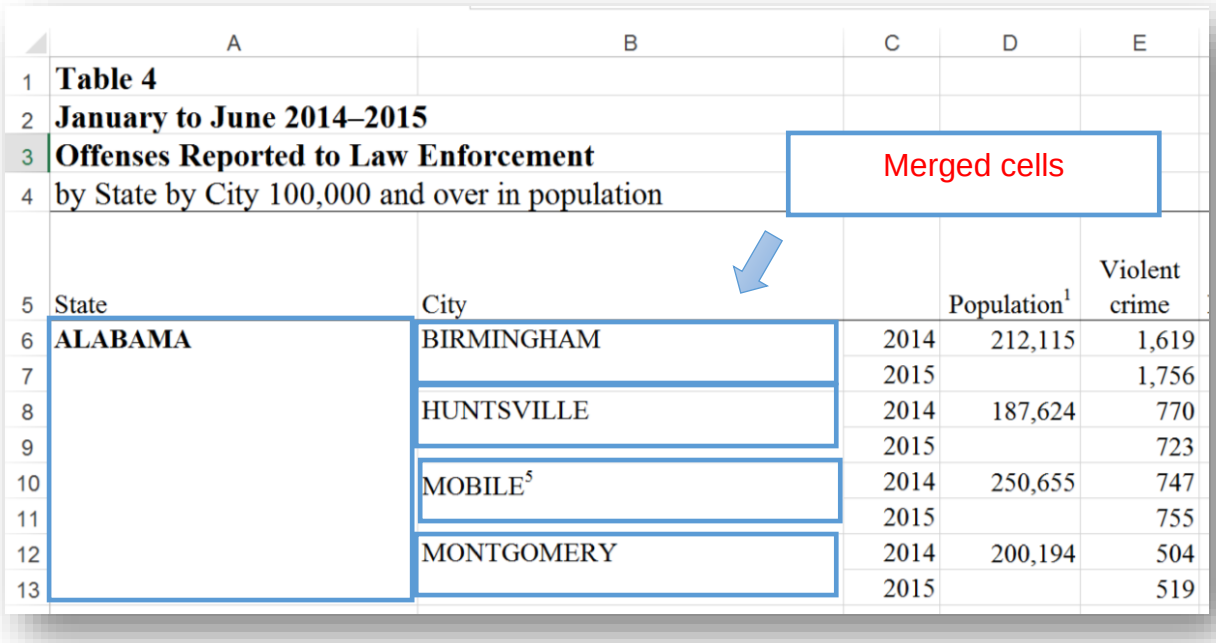
[9] Because of changes in the local agency's reporting practices, figures are not comparable to previous years' data.

[10] The FBI determined that the agency's data were underreported. Consequently, those data are not included in this report.

[11] Arson offenses are reported by the Toledo Fire Department; therefore those figures are not included in this report.

Other than title and data description, footer notes that describe about the data remarks should also be removed before loaded into SAS.

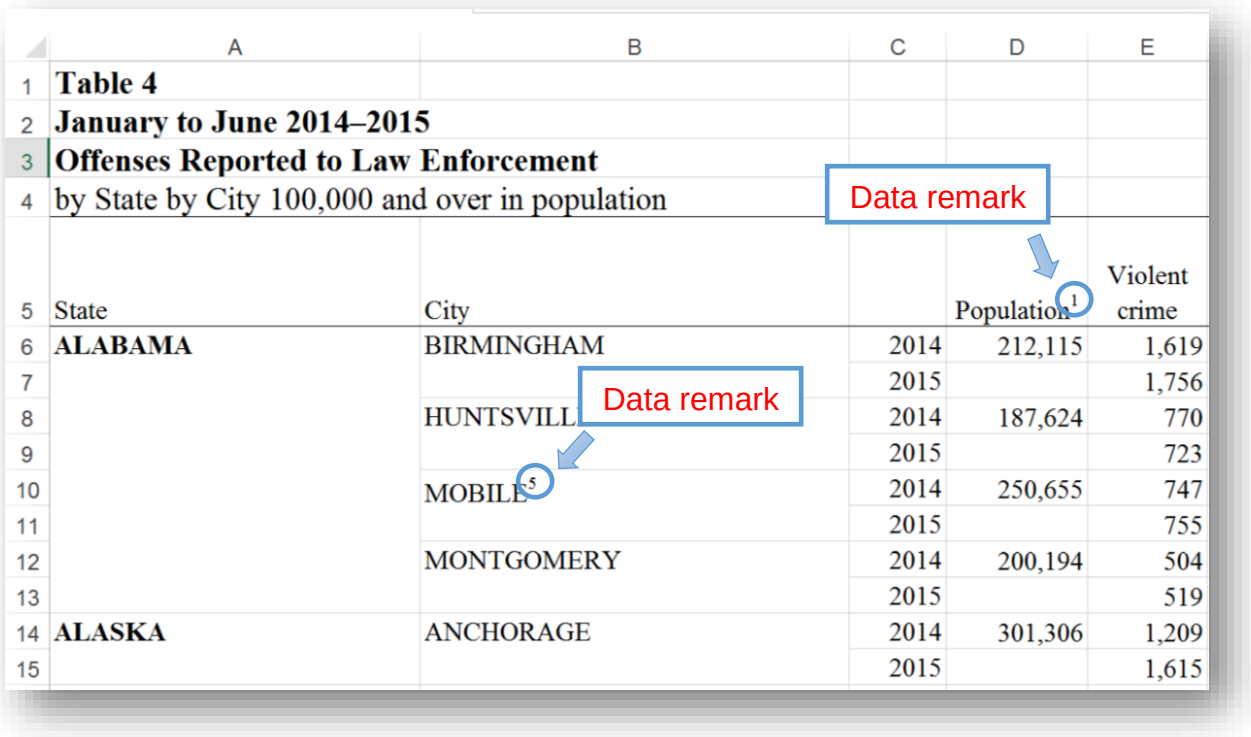
iii. Problem 3: Merged cells



	A	B	C	D	E
1	**Table 4**				
2	**January to June 2014–2015**				
3	**Offenses Reported to Law Enforcement**				
4	by State by City 100,000 and over in population				
5	State	City		Population[1]	Violent crime
6	**ALABAMA**	BIRMINGHAM	2014	212,115	1,619
7			2015		1,756
8		HUNTSVILLE	2014	187,624	770
9			2015		723
10		MOBILE[5]	2014	250,655	747
11			2015		755
12		MONTGOMERY	2014	200,194	504
13			2015		519

It can easily be recognized that the first 2 columns (columns A and B) contain merged cells, which are certainly not appropriate as the rest of the cells (except the first one) will turn empty after unmerging the cells. In this case, further treatment should be done by filling up the values repeatedly.

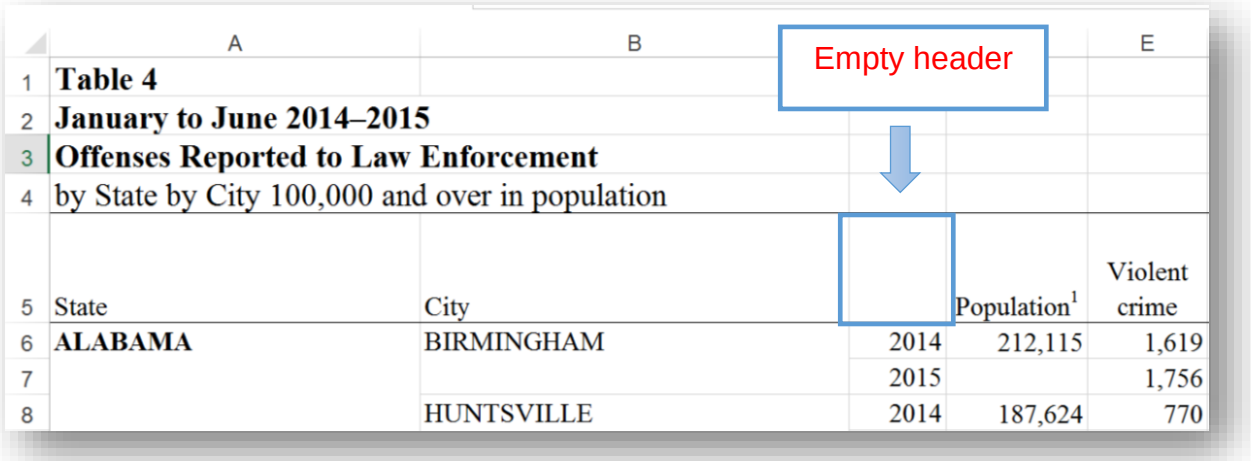
iv. Problem 4: Data remarks



	A	B	C	D	E
1	**Table 4**				
2	**January to June 2014–2015**				
3	**Offenses Reported to Law Enforcement**				
4	by State by City 100,000 and over in population				
5	State	City		Population[1]	Violent crime
6	**ALABAMA**	BIRMINGHAM	2014	212,115	1,619
7			2015		1,756
8		HUNTSVILLE	2014	187,624	770
9			2015		723
10		MOBILE[5]	2014	250,655	747
11			2015		755
12		MONTGOMERY	2014	200,194	504
13			2015		519
14	**ALASKA**	ANCHORAGE	2014	301,306	1,209
15			2015		1,615

The data remarks in superscript format should also be removed as these will be read as normal wording instead of superscript after loading into SAS.

v. Problem 5: Empty header



	A	B			E
1	**Table 4**				
2	**January to June 2014–2015**				
3	**Offenses Reported to Law Enforcement**				
4	by State by City 100,000 and over in population				
5	State	City		Population[1]	Violent crime
6	**ALABAMA**	BIRMINGHAM	2014	212,115	1,619
7			2015		1,756
8		HUNTSVILLE	2014	187,624	770

Missing header should be filled up properly instead of leaving it blanked. In this case, the empty header is recognized as representing year (2014 & 2015), so it can be filled up with 'Year'.

4.2 Data Cleaning & Pre-Processing

Data Cleaning with the Aids of Python

To resolve the few problems identified above, Python has been used as it allows high flexibility in manipulating the data and can help to resolve the problems mentioned without much manual works. Python is a widely used open source programming language in data science and analytics as it provides many libraries that allows user to manipulate data, clean messy data, process data into proper form, visualize the data and perform analysis, establish statistical analysis and modelling, etc. In this assignment, Python is utilized to only achieve the objective of turning the raw data file into a well-structured table format, which involves solving the few problems mentioned above.

To cater with the problems of addition titles, descriptions, and footer notes (Problem 1 & 2), a Python library called 'pandas' can be used. Removing these additional notes can be achieved upon reading the Excel file using 'pandas' in Python. Let's explore the script required to perform this action.

```
import pandas as pd
df = pd.read_excel(r'C:\Prelim Semiannual Report to FIOU - Dec 14 2015\Table 4\Table_4_January_to_June_2015.xls',
                   sheetname='15table4',
                   header=4,
                   skipfooter=11)
```

Script explanation:

- Step 1: import 'pandas' library in Python
- Step 2: use read_excel function exists in the library to import Excel file
    i. First argument: file path
    ii. Argument 'sheetname': the worksheet name in the Excel file
    iii. Argument 'header': index of header **(resolve Problem 1)**

iv. Argument ‘skipfooter’: last n rows to skip with **(resolve Problem 2)**

After running the script, let’s print out the data frame that we have just import:

```
df
```

	State	City	Unnamed: 2	Population1	Violent crime	Murder	Rape (revised definition)2	Rape (legacy definition)3	Robbery	Aggravated assault	Property crime	Burglary	Larcen theft
0	ALABAMA	BIRMINGHAM	2014	212115.0	1619.0	23.0	83.0	NaN	454.0	1059.0	6596.0	1716.0	4169.0
1	NaN	NaN	2015	NaN	1756.0	30.0	77.0	NaN	507.0	1142.0	6246.0	1446.0	4120.0
2	NaN	HUNTSVILLE	2014	187624.0	770.0	12.0	50.0	NaN	188.0	520.0	4376.0	908.0	3111.0
3	NaN	NaN	2015	NaN	723.0	5.0	65.0	NaN	173.0	480.0	4121.0	836.0	2903.0
4	NaN	MOBILE5	2014	250655.0	747.0	17.0	67.0	NaN	203.0	460.0	5747.0	1461.0	4039.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	NaN	NaN	2015	NaN	127.0	2.0	NaN	26.0	50.0	49.0	920.0	152.0	714.0
516	NaN	MADISON	2014	245788.0	400.0	3.0	42.0	NaN	101.0	254.0	3135.0	448.0	2572.0
517	NaN	NaN	2015	NaN	414.0	3.0	48.0	NaN	99.0	264.0	3121.0	544.0	2475.0
518	NaN	MILWAUKEE	2014	600374.0	3957.0	36.0	192.0	NaN	1551.0	2178.0	12303.0	2416.0	6966.0
519	NaN	NaN	2015	NaN	3921.0	75.0	168.0	NaN	1501.0	2177.0	11197.0	2439.0	5798.0

520 rows × 15 columns

It can clearly be seen that ‘pandas’ skipped the title, description and footer and reads only the data itself. Next, it can be noticed that the first 2 columns ‘State’ and ‘City’ contain many missing values (presented as ‘NaN’ in ‘pandas’ data frame). This was caused by the merge cells as discussed above in Problem 3. A function has been written to fill up the values accordingly:

```
def fill_in_na(df, arr):
    for iarr in arr:                        # Loop columns
        check = ~df[iarr].isnull()          # Identify row index that is NOT missing
        for i in range(len(df[iarr])):      # Loop every single row
            if check[i]:
                val = df[iarr][i]           # Store the value in that row if it is NOT a missing value / 'NaN'
            else:
                df[iarr][i] = val           # Fill up with the stored value if it is a missing value / 'NaN'
    return df                               # Return the data frame

df = fill_in_na(df, ['State', 'City'])      # Execute the function on columns 'State' and 'City'
```

Script explanation:

- Step 1: Define function that take data frame and array as inputs
    i. Argument ‘df’: Input of data frame to perform the function
    ii. Argument ‘arr’: The column names selected to perform the function
- Step 2: Loop all input columns to perform following actions
    i. Identify the row index that is not a missing value / ‘NaN’
    ii. Loop every single row to perform following actions
        - Store the value in the row if it is not a missing value / ‘NaN’
        - Fill up with the stored value if it is a missing value / ‘NaN’

- Step 3: Return the data frame
- Step 4: Execute the function

Performing this function will give us the following result, and notice that missing values in columns ‘State’ and ‘City’ have been filled up accordingly. **This resolved the issue resulted by the merged cell (Problem 3).**

```
df
```

	State	City	Unnamed: 2	Population1	Violent crime	Murder	Rape (revised definition)2	Rape (legacy definition)3	Robbery	Aggravated assault	Property crime	Burglary	Larc theft
0	ALABAMA	BIRMINGHAM	2014	212115.0	1619.0	23.0	83.0	NaN	454.0	1059.0	6596.0	1716.0	4169
1	ALABAMA	BIRMINGHAM	2015	NaN	1756.0	30.0	77.0	NaN	507.0	1142.0	6246.0	1446.0	4120
2	ALABAMA	HUNTSVILLE	2014	187624.0	770.0	12.0	50.0	NaN	188.0	520.0	4376.0	908.0	3111
3	ALABAMA	HUNTSVILLE	2015	NaN	723.0	5.0	65.0	NaN	173.0	480.0	4121.0	836.0	2903
4	ALABAMA	MOBILE5	2014	250655.0	747.0	17.0	67.0	NaN	203.0	460.0	5747.0	1461.0	4039
...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	WISCONSIN	KENOSHA	2015	NaN	127.0	2.0	NaN	26.0	50.0	49.0	920.0	152.0	714.
516	WISCONSIN	MADISON	2014	245788.0	400.0	3.0	42.0	NaN	101.0	254.0	3135.0	448.0	2572
517	WISCONSIN	MADISON	2015	NaN	414.0	3.0	48.0	NaN	99.0	264.0	3121.0	544.0	2475
518	WISCONSIN	MILWAUKEE	2014	600374.0	3957.0	36.0	192.0	NaN	1551.0	2178.0	12303.0	2416.0	6966
519	WISCONSIN	MILWAUKEE	2015	NaN	3921.0	75.0	168.0	NaN	1501.0	2177.0	11197.0	2439.0	5798

520 rows × 15 columns

For removing the data remarks in superscript form (when showed in Excel), it requires to make changes in the string itself. Regular Expression is a common programming technique used to identify patterns in the string, which is helpful in resolving this problem. It is identified that the data remarks in superscript form only exist in the headers and column ‘City’.

a) Remove data remarks exist in the headers

```
def clean_colnames(df):
    colnames = []                                                    # Initialize an array
    for col in df.columns.tolist():                                  # Loop every header
        try:
            col = re.compile(r'(\D+)\d+$').findall(col)[0]           # Finding the header with digit(s) as last character
        except AttributeError:
            pass                                                     # Skip if no digit(s) found
        colnames.append(col.replace('\n', ' ').replace('-','').strip())  # Trim the header and remove '-'
    df.columns = colnames                                            # Update the headers
    return df                                                        # Return the data frame

df = clean_colnames(df)                                              # Execute the function
```

Script explanation:

- Step 1: Take data frame as the input
- Step 2: Initialize an empty array for storing modified headers
- Step 3: Loop every header to perform following actions

i. Extract only the character and ignore the digits if there is any
ii. Skip if there is no digit in the header
iii. Append the modified header into the array

- Step 4: Replace the headers of the data frame with the modified headers
- Step 5: Return the data frame with new headers
- Step 6: Execute the function

b) Remove data remarks exist in columns ‘State’ and ‘City’

```
df['State'] = df['State'].map(lambda x: x[:[m for m in re.compile('\d').finditer(x+'1')][0].start()])
df['City'] = df['City'].map(lambda x: x[:[m for m in re.compile('\d').finditer(x+'1')][0].start()])
```

Script explanation:

- ‘map’ function has been used to perform looping on each row of ‘State’ and ‘City’ columns to remove the digits exist in the words. This works the same as the previous action of removing data remarks in the headers.

Concretely by running the scripts, it gives us the following result, which all the numbers exist in the headers and column ‘State’ and ‘City’ have been taken out (as illustrated in the circles below. **This has fixed the issue of data remarks in the data itself (Problem 4).**

```
df
```

	State	City	Unnamed:	Population	Violent crime	Murder	Rape (revised definition)	Rape (legacy definition)	Robbery	Aggravated assault	Property crime	Burglary	Larceny theft
0	ALABAMA	BIRMINGHAM	2014	212115.0	1619.0	23.0	83.0	NaN	454.0	1059.0	6596.0	1716.0	4169.0
1	ALABAMA	BIRMINGHAM	2015	NaN	1756.0	30.0	77.0	NaN	507.0	1142.0	6246.0	1446.0	4120.0
2	ALABAMA	HUNTSVILLE	2014	187624.0	770.0	12.0	50.0	NaN	188.0	520.0	4376.0	908.0	3111.0
3	ALABAMA	HUNTSVILLE	2015	NaN	723.0	5.0	65.0	NaN	173.0	480.0	4121.0	836.0	2903.0
4	ALABAMA	MOBILE	2014	250655.0	747.0	17.0	67.0	NaN	203.0	460.0	5747.0	1461.0	4039.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	WISCONSIN	KENOSHA	2015	NaN	127.0	2.0	NaN	26.0	50.0	49.0	920.0	152.0	714.0
516	WISCONSIN	MADISON	2014	245788.0	400.0	3.0	42.0	NaN	101.0	254.0	3135.0	448.0	2572.0
517	WISCONSIN	MADISON	2015	NaN	414.0	3.0	48.0	NaN	99.0	264.0	3121.0	544.0	2475.0
518	WISCONSIN	MILWAUKEE	2014	600374.0	3957.0	36.0	192.0	NaN	1551.0	2178.0	12303.0	2416.0	6966.0
519	WISCONSIN	MILWAUKEE	2015	NaN	3921.0	75.0	168.0	NaN	1501.0	2177.0	11197.0	2439.0	5798.0

520 rows × 15 columns

For the last problem, which is related to empty headers, this can be achieved by simply rename the headers accordingly **(resolved Problem 5).**

```
df = df.rename(columns = {'Unnamed:':'Year'})
```

```
df
```

	State	City	Year	Population	Violent crime	Murder	Rape (revised definition)	Rape (legacy definition)	Robbery	Aggravated assault	Property crime	Burglary	Larceny theft	Mot vehi thef
0	ALABAMA	BIRMINGHAM	2014	212115.0	1619.0	23.0	83.0	NaN	454.0	1059.0	6596.0	1716.0	4169.0	711.
1	ALABAMA	BIRMINGHAM	2015	NaN	1756.0	30.0	77.0	NaN	507.0	1142.0	6246.0	1446.0	4120.0	680.
2	ALABAMA	HUNTSVILLE	2014	187624.0	770.0	12.0	50.0	NaN	188.0	520.0	4376.0	908.0	3111.0	357.
3	ALABAMA	HUNTSVILLE	2015	NaN	723.0	5.0	65.0	NaN	173.0	480.0	4121.0	836.0	2903.0	382.
4	ALABAMA	MOBILE	2014	250655.0	747.0	17.0	67.0	NaN	203.0	460.0	5747.0	1461.0	4039.0	247.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	WISCONSIN	KENOSHA	2015	NaN	127.0	2.0	NaN	26.0	50.0	49.0	920.0	152.0	714.0	54.0
516	WISCONSIN	MADISON	2014	245788.0	400.0	3.0	42.0	NaN	101.0	254.0	3135.0	448.0	2572.0	115.
517	WISCONSIN	MADISON	2015	NaN	414.0	3.0	48.0	NaN	99.0	264.0	3121.0	544.0	2475.0	102.
518	WISCONSIN	MILWAUKEE	2014	600374.0	3957.0	36.0	192.0	NaN	1551.0	2178.0	12303.0	2416.0	6966.0	292
519	WISCONSIN	MILWAUKEE	2015	NaN	3921.0	75.0	168.0	NaN	1501.0	2177.0	11197.0	2439.0	5798.0	296

520 rows × 15 columns

Lastly, we export the well-structured data frame out as a file in Excel format using the 'ExcelWriter' and 'to_excel' functions in 'pandas' library.

```
writer = pd.ExcelWriter(r'C:\Prelim Semiannual Report to FIOU - Dec 14 2015\Table 4\crime data_processed.xlsx')
df.to_excel(writer, 'Sheet1', index=None)
writer.save()
```

Data Pre-processing with SAS

i) Data Import

- *Crime data*
- *2015 population data*
- *Poverty data*
- *Unemployment data*

As far as it goes, the messy structure of the raw data has been cleaned up to a proper table in csv format, which is ready to be imported into SAS for further processing and analysis. By running the following scripts, SAS reads the pre-uploaded crime-related data (which has been undergone initial processing using Python, named "crime data_processed.xlsx") located at "/home/ng.wge0731/DAP Assignment/". The function PROC IMPORT with argument DBMS specified as "xlsx" is the function to perform

data import on excel file into the system. Next, PROC CONTENTS output the summary description about the imported crime data, describing the number of observations, number of variables, data types, etc. PROC PRINT will print out the entire data for reference purpose.

```
/* Data Import : Crime Data */
proc import datafile="/home/ng.wge0731/DAP Assignment/crime data_processed.xlsx"
        out=work.crimedata
        dbms=xlsx
        replace;
run;

title 'Imported Crime Data';
proc contents data=crimedata;
run;

proc print data=crimedata;
run;
title;
```

Table of Contents

Imported Crime Data

The CONTENTS Procedure

Data Set Name	WORK.CRIMEDATA	**Observations**	520
Member Type	DATA	**Variables**	15
Engine	V9	**Indexes**	0
Created	05/14/2017 14:32:57	**Observation Length**	152
Last Modified	05/14/2017 14:32:57	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
10	Aggravated_assault	Num	8	BEST.		Aggravated_assault
15	Arson	Num	8	BEST.		Arson
12	Burglary	Num	8	BEST.		Burglary
2	City	Char	29	$29.	$29.	City
13	Larceny_theft	Num	8	BEST.		Larceny_theft
14	Motor_vehicle_theft	Num	8	BEST.		Motor_vehicle_theft
6	Murder	Num	8	BEST.		Murder
4	Population	Num	8	BEST.		Population
11	Property_crime	Num	8	BEST.		Property_crime
8	Rape_legacy	Num	8	BEST.		Rape_legacy
7	Rape_revised	Num	8	BEST.		Rape_revised
9	Robbery	Num	8	BEST.		Robbery
1	State	Char	14	$14.	$14.	State
5	Violent_crime	Num	8	BEST.		Violent_crime
3	Year	Num	8	BEST.		Year

Imported Crime Data

Obs	State	City	Year	Population	Violent_crime	Murder	Rape_revised	Rape_legacy	Robbery	Aggravated_assault	Property_crime	Bur
1	ALABAMA	BIRMINGHAM	2014	212115	1619	23	83	.	454	1059	6596	
2	ALABAMA	BIRMINGHAM	2015	.	1756	30	77	.	507	1142	6246	
3	ALABAMA	HUNTSVILLE	2014	187624	770	12	50	.	188	520	4376	
4	ALABAMA	HUNTSVILLE	2015	.	723	5	65	.	173	480	4121	
5	ALABAMA	MOBILE	2014	250655	747	17	67	.	203	460	5747	
6	ALABAMA	MOBILE	2015	.	755	9	69	.	186	491	5210	
7	ALABAMA	MONTGOMERY	2014	200194	504	14	21	.	218	251	4142	
8	ALABAMA	MONTGOMERY	2015	.	519	20	14	.	191	294	4029	
9	ALASKA	ANCHORAGE	2014	301306	1209	6	193	.	247	763	5515	
10	ALASKA	ANCHORAGE	2015	.	1615	16	323	.	271	1005	5732	

From the above PROC PRINT output, it is known that population data in 2015 are missing. According to World Bank publication [2], statistics regarding U.S. census showed that population in U.S. grew 0.784% from 2014 to 2015. With this information on hand, the missing 2015 data for population can be estimated by taking the multiplication of 2014 data and 1.00784 (indicating the growth of 0.784%). The data have been estimated using Excel and uploaded into SAS. The same procedure of PROC IMPORT has been carried out to import the excel file into the system. Similarly, we use PROC CONTENTS and PROC PRINT to understand the data.

```
/* Data Import : 2015 Population Data */
/* According to World Bank publication, U.S. population growth from 2014 to 2015 is 0.784% */
/* 2015 population data was calculated by taking (2014 data)*1.00784 */
proc import datafile="/home/ng.wge0731/DAP Assignment/population_2015.xlsx"
        out=work.population15
        dbms=xlsx
        replace;
run;

title 'Imported Population Data (2015)';
proc contents data=population15;
run;

proc print data=population15;
run;
title;
```

Table of Contents

Imported Population Data (2015)

The CONTENTS Procedure

Data Set Name	WORK.POPULATION15	Observations	260
Member Type	DATA	Variables	4
Engine	V9	Indexes	0
Created	05/14/2017 14:32:57	Observation Length	64
Last Modified	05/14/2017 14:32:57	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
2	City	Char	29	$29.	$29.	City
4	Population15	Num	8	BEST.		Population15
1	State	Char	14	$14.	$14.	State
3	Year	Num	8	BEST.		Year

Imported Population Data (2015)

Obs	State	City	Year	Population15
1	ALABAMA	BIRMINGHAM	2015	213777.9816
2	ALABAMA	HUNTSVILLE	2015	189094.97216
3	ALABAMA	MOBILE	2015	252620.1352
4	ALABAMA	MONTGOMERY	2015	201763.52096
5	ALASKA	ANCHORAGE	2015	303668.23904
6	ARIZONA	CHANDLER	2015	254347.57296
7	ARIZONA	GILBERT	2015	237275.7712
8	ARIZONA	GLENDALE	2015	238636.3552
9	ARIZONA	PEORIA	2015	166013.42048
10	ARIZONA	PHOENIX	2015	1541846.0397

The data set contains 260 number of observations with 4 variables, which will later be merged to the crime data set to replace the missing values.

In coming sections, we will be utilizing poverty and unemployment rate in U.S. to enrich the analysis of crime, which will be covered in Objective 6 & 7. To achieve this, poverty data was acquired via U.S. Census Bureau [4] and unemployment data was obtained from Bureau of Labor Statistics [3]. Similarly, the 3 procedures (i.e. PROC IMPORT, PROC CONTENTS and PROC PRINT) were used to import the data, view the data summary and the data itself respectively.

For poverty data:

```
/* Data Import : Poverty Rate Data */
proc import datafile="/home/ng.wge0731/DAP Assignment/poverty data.xlsx"
        out=work.poverty
        dbms=xlsx
        replace;
run;

title 'Imported Poverty Rate Data';
proc contents data=poverty;
run;

proc print data=poverty;
run;
title;
```

Table of Contents

Imported Poverty Rate Data

The CONTENTS Procedure

Data Set Name	WORK.POVERTY	**Observations**	12
Member Type	DATA	**Variables**	3
Engine	V9	**Indexes**	0
Created	05/14/2017 15:40:55	**Observation Length**	32
Last Modified	05/14/2017 15:40:55	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
3	Poverty_rate	Num	8	BEST.		Poverty_rate
1	State	Char	13	$13.	$13.	State
2	Year	Num	8	BEST.		Year

Imported Poverty Rate Data

Obs	State	Year	Poverty_rate
1	CALIFORNIA	2014	16.4
2	CALIFORNIA	2015	15.3
3	TEXAS	2014	17.2
4	TEXAS	2015	15.9
5	NEW YORK	2014	15.9
6	NEW YORK	2015	15.4
7	NORTH DAKOTA	2014	11.5
8	NORTH DAKOTA	2015	11
9	NEW HAMPSHIRE	2014	9.2
10	NEW HAMPSHIRE	2015	8.2
11	IDAHO	2014	14.8
12	IDAHO	2015	15.1

For unemployment data:

```
/* Data Import : Unemployment Rate Data */
proc import datafile="/home/ng.wge0731/DAP Assignment/unemploy data.xlsx"
        out=work.unemploy
        dbms=xlsx
        replace;
run;

title 'Imported Unemployment Rate Data';
proc contents data=unemploy;
run;

proc print data=unemploy;
run;
title;
```

Table of Contents

Imported Unemployment Rate Data

The CONTENTS Procedure

Data Set Name	WORK.UNEMPLOY	**Observations**	102
Member Type	DATA	**Variables**	3
Engine	V9	**Indexes**	0
Created	05/14/2017 15:40:55	**Observation Length**	40
Last Modified	05/14/2017 15:40:55	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	State	Char	20	$20.	$20.	State
3	Unemployment_rate	Num	8	BEST.		Unemployment_rate
2	Year	Num	8	BEST.		Year

Imported Unemployment Rate Data

Obs	State	Year	Unemployment_rate
1	NORTH DAKOTA	2014	2.7
2	NEBRASKA	2014	3.3
3	SOUTH DAKOTA	2014	3.4
4	UTAH	2014	3.8
5	VERMONT	2014	4
6	MINNESOTA	2014	4.2
7	WYOMING	2014	4.2
8	IOWA	2014	4.3
9	NEW HAMPSHIRE	2014	4.3
10	HAWAII	2014	4.4

ii) Data pre-processing

- *Identify missing values*
- *Missing values imputation*
- *Merge data where applicable (e.g. 2015 population data)*
- *Create or modify variables (also commonly called features) where applicable*

For instance, data comes in unclean forms, where missing values exists. Missing values are inevitable and mostly increase the challenges to analysts as significant existence of missing values might lead to misinterpretation of the findings and unreliable analyses, which can be huge impact to the business. To start off the data pre-processing process, we first identify the missing value existence in each variable with PROC MEANS procedure as below.

```
/* Check missing value count */
title 'Overview of Missing Values';
proc means data=work.crimedata nmiss mean;
run;
title;
```

Table of Contents

Overview of Missing Values

The MEANS Procedure

Variable	Label	N Miss
Year	Year	0
Population	Population	260
Violent_crime	Violent_crime	5
Murder	Murder	5
Rape_revised	Rape_revised	120
Rape_legacy	Rape_legacy	405
Robbery	Robbery	5
Aggravated_assault	Aggravated_assault	5
Property_crime	Property_crime	8
Burglary	Burglary	6
Larceny_theft	Larceny_theft	6
Motor_vehicle_theft	Motor_vehicle_theft	6
Arson	Arson	27

The PROC MEANS with argument ‘NMISS’ will output a summary table containing each variables with number of missing values exist. From the summary table, variable ‘Population’ has 260 of missing values as expected (2015 data are missing). Other variables contain only small amount of missing values, except variables ‘Rape_revised’ and ‘Rape_legacy’, with 120 and 405 counts of missing values respectively. These variables contain large number of missing values because the data for raping cases were recorded with 2 UCR (Uniform Crime Reporting) definitions. However, both of these describe rape cases, which can be added up to form a new variable called ‘Total_rape’, indicating rape cases in general. Since there are missing values exist in both variables ‘Rape_revised’ and ‘Rape_legacy’, the 2 variables cannot be simply undergone summation as missing values (presented in ‘.’) plus any number will output a missing value. To resolve this, a macro function named ‘replace_value’ has been created to replace all the missing values with zeros (in numeric form) so the 2 variables can be added up.

```
/* Macro function : Replace missing value */
%macro replace_value(variable_, find_, replacement_, arrname_);
        array &arrname_ _numeric_;
                do over &arrname_;
                        if &variable_ = &find_ then &variable_ = &replacement_;
                end;
%mend;
```

Generally, the macro function will perform looping throughout the entire data of the specified variable and find the matching input (indicated as ‘find_’) and replace with the replacement input (indicated as ‘replacement_’). In rape cases, we would like to replace all the missing values presented in ‘.’ with 0 in numeric form. The following codes utilize the macro function to create new variable called ‘Total_rape’ and drop the original variables ‘Rape_revised’ and ‘Rape_legacy’:

```
/* Calculate variable 'Total_rape' */
data work.crimedata_v2;
        set work.crimedata;

        *replace missing with 0 for variable 'rape_revised' & 'rape_legacy';
        %replace_value(rape_revised, ., 0, rape_revise_new);
        %replace_value(rape_legacy, ., 0, rape_legacy_new);

        *create variable 'Total_rape';
        Total_rape = rape_revised + rape_legacy;

        drop rape_revised rape_legacy;
run;
```

Next, the population data in year 2015 should be filled up by leveraging the estimated data that we have imported earlier. PROC SQL is the SAS procedure that allows us to perform various SQL queries in SAS. To perform the merging of population data, SQL query was performed to first join both data sets (i.e. ‘crimedata’ and ‘population15’) by state, city and year. After that, we utilized again the macro function ‘replace_value’ to replace all the missing values exist in both columns ‘Population14’ and ‘Population15’ so summation of both columns can be executed. After the summation of population data in both years, we should again perform the macro functions to replace all the 0 figures exist in ‘Population’ back to missing values in ‘.’ because these are cases where population in both years were missing and should be undergone proper imputation (after all, 0 population in particular state is not logical at all).

```
/* Merge data to fill up 2015 population data */
proc sql;
        create table temp_joined as
        select crimedata_v2.State,
                crimedata_v2.City,
                crimedata_v2.Year,
                crimedata_v2.Population,
                crimedata_v2.Violent_crime,
                crimedata_v2.Murder,
                crimedata_v2.Robbery,
                crimedata_v2.Aggravated_assault,
                crimedata_v2.Property_crime,
                crimedata_v2.Burglary,
                crimedata_v2.Larceny_theft,
                crimedata_v2.Motor_vehicle_theft,
                crimedata_v2.Arson,
                crimedata_v2.Total_rape,
               population15.population15
        from crimedata_v2
        left join population15
        on crimedata_v2.state = population15.state and
           crimedata_v2.city = population15.city and
           crimedata_v2.year = population15.year;
quit;
```

```
data work.crimedata_v3;
        set work.temp_joined(rename=(Population=Population14));

        *replace missing with 0 for variable 'Population14' & 'Population15' ;
        %replace_value(population14, ., 0, pop14_new);
        %replace_value(population15, ., 0, pop15_new);

        *Sum up to become complete variable 'Population';
        Population = population14 + population15;

        *replace 0 as missing value as they should be imputed with average in coming step;
        *(p/s: population with 0 does not make any sense!);
        %replace_value(population, 0, ., pop_new);

        drop population14 population15;
run;
```

The next action required in data pre-processing stage is missing value imputations. As we see that there is not much of missing values exist in the rest of the variables, mean imputation method will provide simple and sufficiently good result in overwriting the missing values.

```
/* Impute missing values with average values */
proc standard data=work.crimedata_v3
        out=work.crimedata_v4
        replace
        print;
run;
```

Lastly, identifying the crime rate in certain states or cities with total crime and crime ratio might also be useful for analysis in later stage. The following codes help create 2 new variables, which are ‘Total_crime’ by taking the summation of violent, property and arson crimes and ‘Total_crime_ratio_per_100k’ by taking the ratio of total crime to 100000 population.

```
/* Calculate total crime & crime ratio */
data work.crimedata_v5;
        set work.crimedata_v4;
        Total_crime = violent_crime + property_crime + arson;
        Total_crime_ratio_per_100k = total_crime / population * 100000;
run;
```

At the end of the stage, these executions and data processing steps will result in the new data set as below:

Table of Contents

New Data Set: Pre-processed Data

The CONTENTS Procedure

Data Set Name	WORK.CRIMEDATA_V5	**Observations**	520
Member Type	DATA	**Variables**	16
Engine	V9	**Indexes**	0
Created	05/14/2017 16:13:18	**Observation Length**	160
Last Modified	05/14/2017 16:13:18	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
7	Aggravated_assault	Num	8	BEST.		Aggravated_assault
12	Arson	Num	8	BEST.		Arson
9	Burglary	Num	8	BEST.		Burglary
2	City	Char	29	$29.	$29.	City
10	Larceny_theft	Num	8	BEST.		Larceny_theft
11	Motor_vehicle_theft	Num	8	BEST.		Motor_vehicle_theft
5	Murder	Num	8	BEST.		Murder
14	Population	Num	8			
8	Property_crime	Num	8	BEST.		Property_crime
6	Robbery	Num	8	BEST.		Robbery
1	State	Char	14	$14.	$14.	State
15	Total_crime	Num	8			
16	Total_crime_ratio_per_100k	Num	8			
13	Total_rape	Num	8			
4	Violent_crime	Num	8	BEST.		Violent_crime
3	Year	Num	8	BEST.		Year

From the data summary above, variables ‘Rape_revised’ and ‘Rape_legacy’ have been taken out, variables ‘Total_rape’, ‘Total_crime’ and ‘Total_crime_ratio_per_100k’

have been added into the data set. Besides, we can also re-explore the missing values at this stage and all missing values have been disappeared with proper treatment.

```
/* Re-explore the missing value count */
title 'Overview of Missing Values';
proc means data=work.crimedata_v5 nmiss;
run;
title;
```

Overview of Missing Values

The MEANS Procedure

Variable	Label	N Miss
Year	Year	0
Violent_crime	Violent_crime	0
Murder	Murder	0
Robbery	Robbery	0
Aggravated_assault	Aggravated_assault	0
Property_crime	Property_crime	0
Burglary	Burglary	0
Larceny_theft	Larceny_theft	0
Motor_vehicle_theft	Motor_vehicle_theft	0
Arson	Arson	0
Total_rape		0
Population		0
Total_crime		0
Total_crime_ratio_per_100k		0

4.3 Exploratory Data Analysis & Interpretations

Objective 1: To explore the overall crime trend in United States

During the initial step in data analysis, we should explore the overall crime trend in general, which helps us to get some intuitions about the crime in U.S. from 2014 to 2015. Furthermore, we can further explore the crime trend by different types of crimes, which are violent crime, property crime and arson.

Number of crimes by year

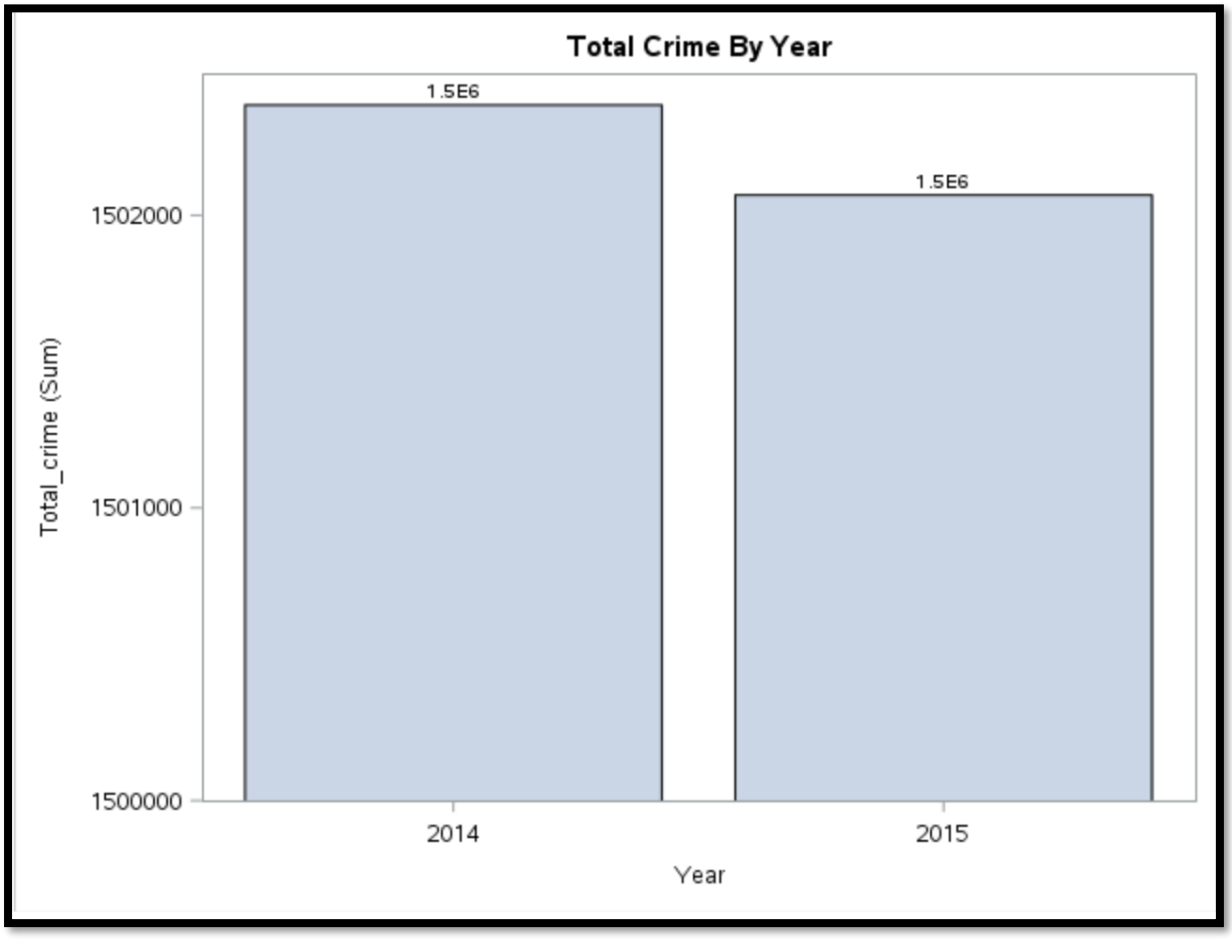
The MEANS Procedure

Year=2014

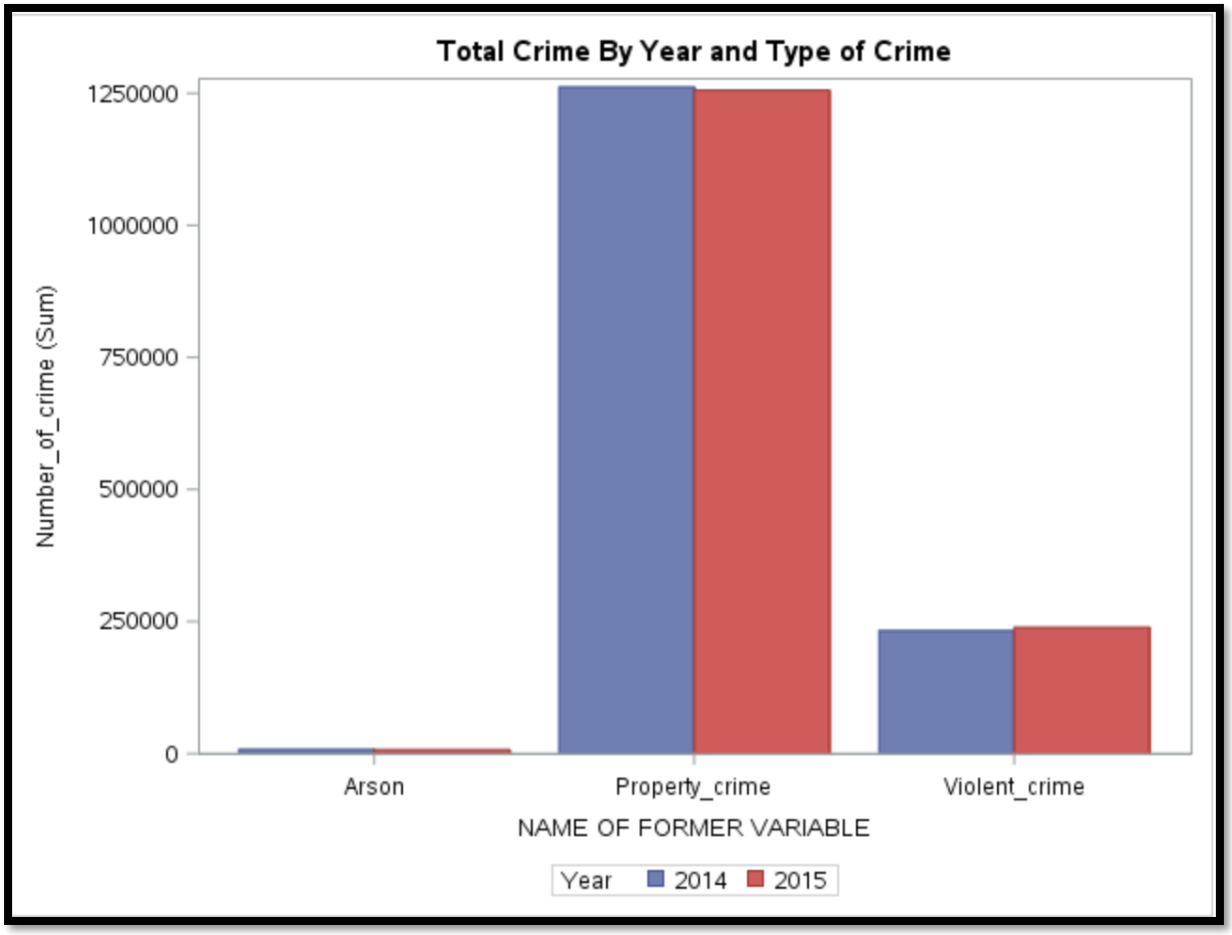
Variable	Label	Sum
Violent_crime	Violent_crime	232722.11
Property_crime	Property_crime	1261714.81
Arson	Arson	7939.66
Total_crime		1502376.58

Year=2015

Variable	Label	Sum
Violent_crime	Violent_crime	239033.00
Property_crime	Property_crime	1255675.27
Arson	Arson	7360.78
Total_crime		1502069.05



From the summary table, it is known that the total crime rate in U.S. slightly decreased from 2014 to 2015, which incurred about 300 criminal cases decreased. Next, we can explore further to see the total crimes by year and the 3 general types of crimes (Property, Violent and Arson).



Among the 3 types of crime, property crimes incurred the highest frequency followed by violent crimes. By comparing the trend of these 3 types of crime, we see that number of arson and property crime are decreasing over the years. However, we should also notice that violent crime increased about over 6000 cases from 2014 to 2015, despite the downward trend of total crime rate. In other words, despite other crimes have been getting less from 2014 to 2015, violent crimes have been getting more instead.

Objective 2: To identify the top 3 states with highest and lowest crime rate

After understanding the overall trend of crime in U.S., some states can be picked to further study and analyze the crime. Here, we will select the top states with highest crime ratio and top state with lowest crime ratio occurred in 2014 and 2015. Crime ratio was calculated with the division of total number of crime by population. It gives more meaningful figures as it takes population into account to measure the crime rate better. Concretely in our case, crime ratio per 100,000 was calculated to measure in 100,000 crowd of people in the location, how many occurrences of criminal activities.

Top 3 States with Highest Crime Rate

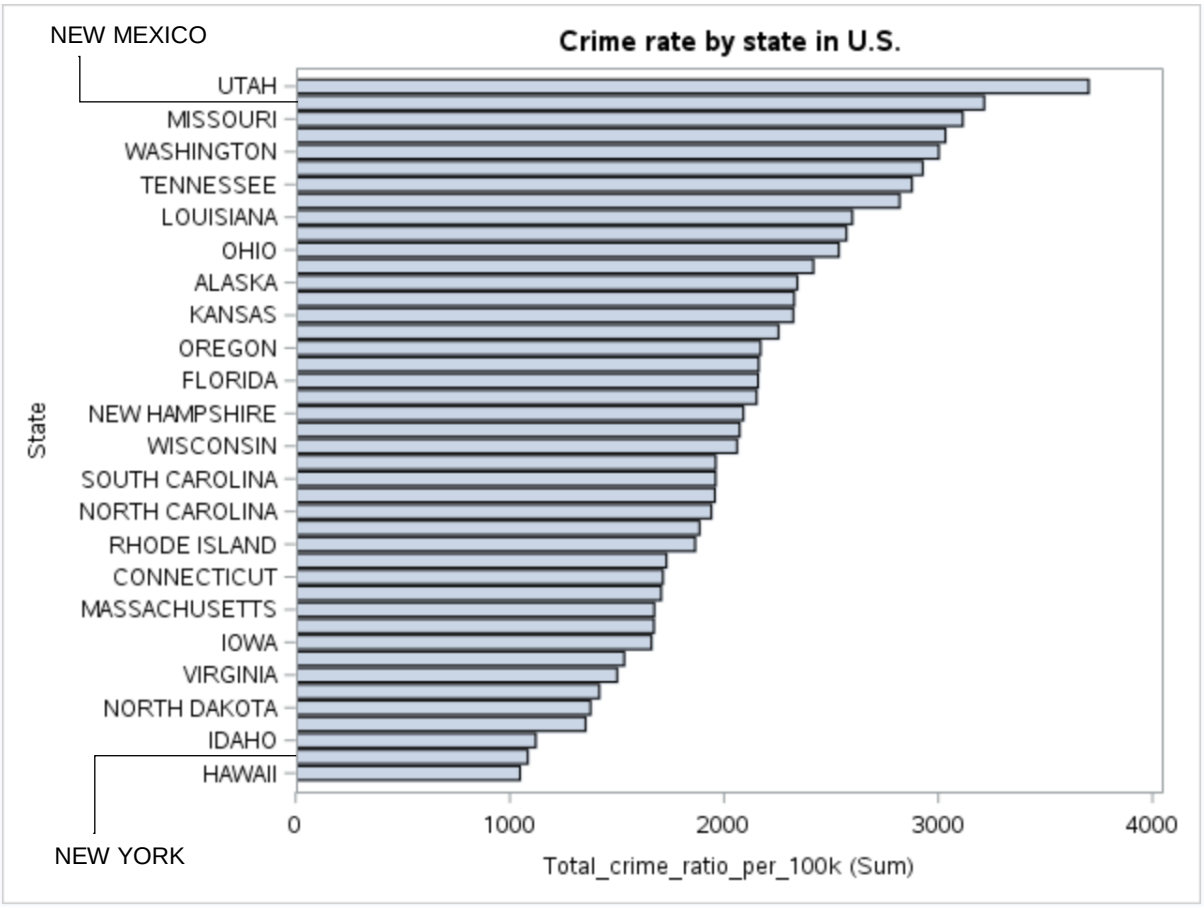
Obs	State	Population	Total_crime	Total_crime_ratio_per_100k
1	UTAH	891713.87	33008.56	3701.70
2	NEW MEXICO	1122129.57	36067.00	3214.16
3	MISSOURI	2382179.68	74184.00	3114.12

From the summary tables, top 3 states with highest crime rate (particularly crime ratio per 100,000) in 2014 and 2015 are Utah > New Mexico > Missouri, where Utah and New Mexico have over 30,000 of total criminal activities throughout the years, indicating high risk of crime in the states. Despite we see Missouri has much higher total number of crimes among the 3 states, we also see that it has much higher number of population in that state. Taking crime ratio as the measurement of crime rate, every 100,000 people in Utah will find approximately 3700 criminal activities in 2014 and 2015. On the other hand, Hawaii, New York and Idaho are the states with lowest crime rate over the years, following the order: Hawaii > New York > Idaho. Among the states in U.S., Hawaii has the lowest crime rate in 2014 and 2015, with about 1047 criminal activities per 100,000 people in the state.

Top 3 States with Lowest Crime Rate

Obs	State	Population	Total_crime	Total_crime_ratio_per_100k
1	HAWAII	1995861.23	20897.78	1047.06
2	NEW YORK	18887192.70	204616.27	1083.36
3	IDAHO	434215.48	4866.00	1120.64

The top states with highest and lowest crime rate can also be visualized with a bar plot of crimes ratio by states:



Objective 3: To analyze the crime details in the selected top states

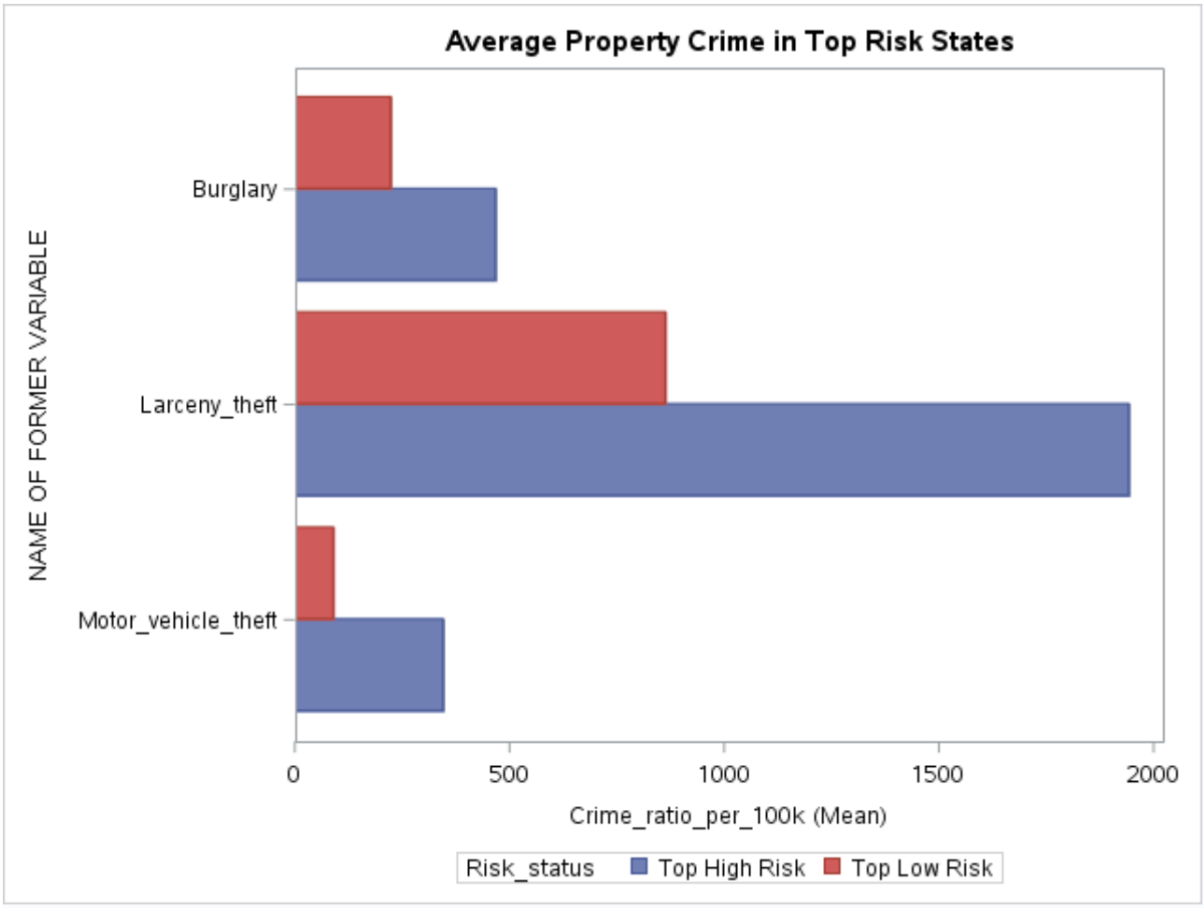
At this stage, top states with highest and lowest crime rate have been identified. Now, we can further analyze what type of crimes happened more frequent in high-risk and low-risk states.

- Top 3 high-risk states: Utah > New Mexico > Missouri
- Top 3 low-risk states: Hawaii > New York > Idaho

Based on FBI's Uniform Crime Reporting (UCR) definition, criminal activities like burglary, larceny theft and motor vehicle theft are categorized as property crime, which are related to activities of obtaining money, property and other benefits without any agreement and permissions of the owner. For violent crime, it covers murder, rape, robbery and aggravated assault, which are the criminal activities related to forcible violence upon the victims. In this section, analyses will be carried out on property, violent and arson crimes to see which crime components tend to have higher crime rate in both low and high risk states.

Average Property Crime in Top Risk States by Year

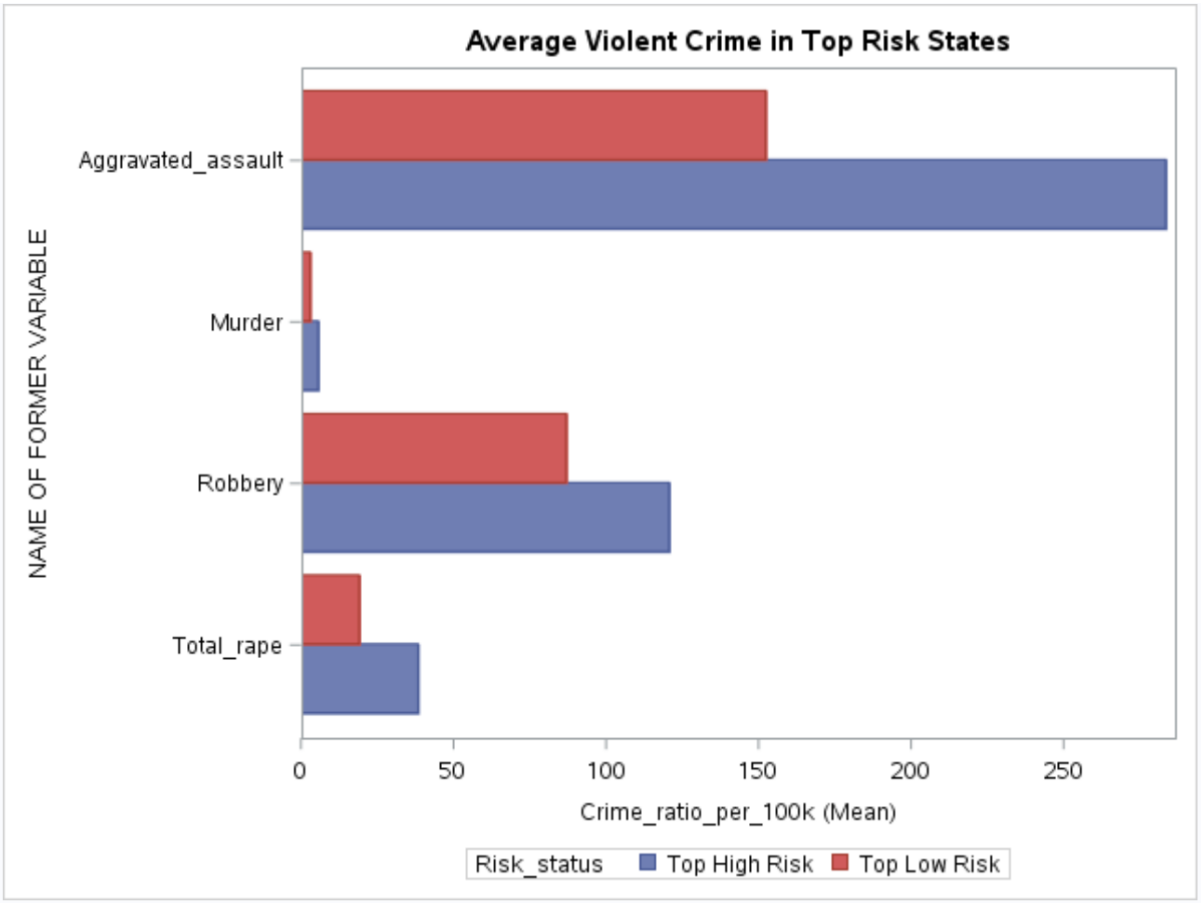
		Year						All		
		2014			2015					
		Burglary	Larceny_theft	Motor_vehicle_theft	Burglary	Larceny_theft	Motor_vehicle_theft	Burglary	Larceny_theft	Motor_vehicle_theft
		Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio
Risk_status	State									
Top High Risk	MISSOURI	454.01	1786.35	323.72	463.30	1693.28	335.77	458.65	1739.82	329.75
	NEW MEXICO	553.79	1833.69	307.94	515.40	1909.44	389.70	534.59	1871.56	348.82
	UTAH	624.07	2787.58	435.60	292.93	1817.99	303.65	458.50	2302.78	369.62
Top Low Risk	HAWAII	94.06	336.73	55.55	206.82	1012.55	161.70	150.44	674.64	108.63
	IDAHO	161.38	761.12	38.38	124.80	788.23	60.56	143.09	774.68	49.47
	NEW YORK	275.04	960.86	87.50	221.68	856.49	98.28	248.36	908.68	92.89



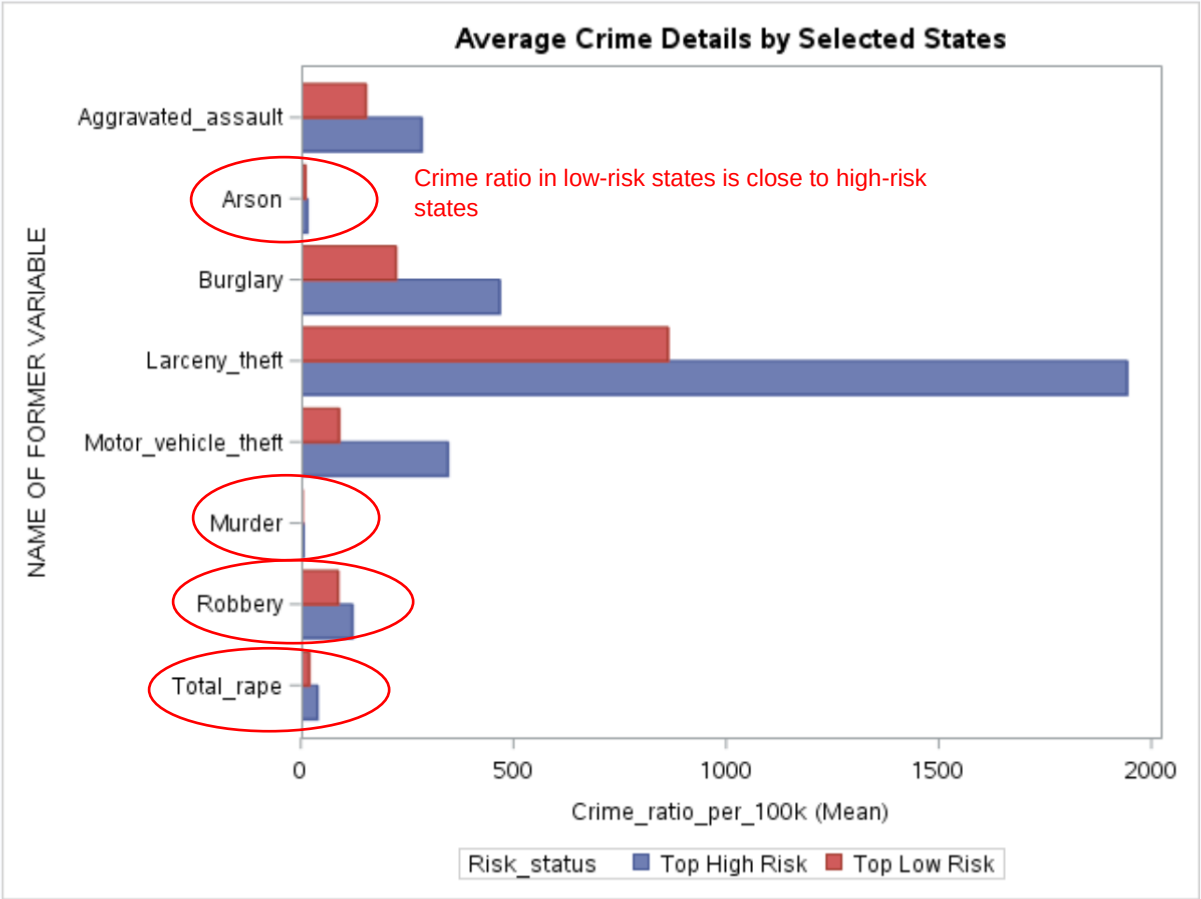
Among all the property crimes, larceny theft is the most common criminal activity regardless if the states are high or low risk states in United States.

Average Violent Crime in Top Risk States by Year

		Year								All			
		2014				2015							
		Aggravated_assault	Murder	Robbery	Total_rape	Aggravated_assault	Murder	Robbery	Total_rape	Aggravated_assault	Murder	Robbery	Total_rape
		Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio	Average Crime Ratio
Risk_status	State												
Top High Risk	MISSOURI	285.33	6.25	105.26	43.05	334.10	8.44	127.15	42.35	309.71	7.34	116.21	42.70
	NEW MEXICO	275.38	2.15	110.76	35.25	285.48	4.08	152.86	35.33	280.43	3.12	131.81	35.29
	UTAH	343.75	6.42	192.24	19.58	137.89	0.76	56.55	44.75	240.82	3.59	124.39	32.17
Top Low Risk	HAWAII	53.22	1.18	30.02	0.00	56.00	1.20	43.72	14.67	54.61	1.19	36.87	7.34
	IDAHO	107.74	0.92	14.80	30.06	101.40	0.00	7.80	25.23	104.57	0.46	11.30	27.65
	NEW YORK	180.18	4.02	111.43	19.72	173.05	3.43	104.40	19.55	176.61	3.72	107.92	19.64



For violent crimes, aggravated assault is the main contributor to the crime rate in both low and high risk states, followed by robbery, rape and murder.



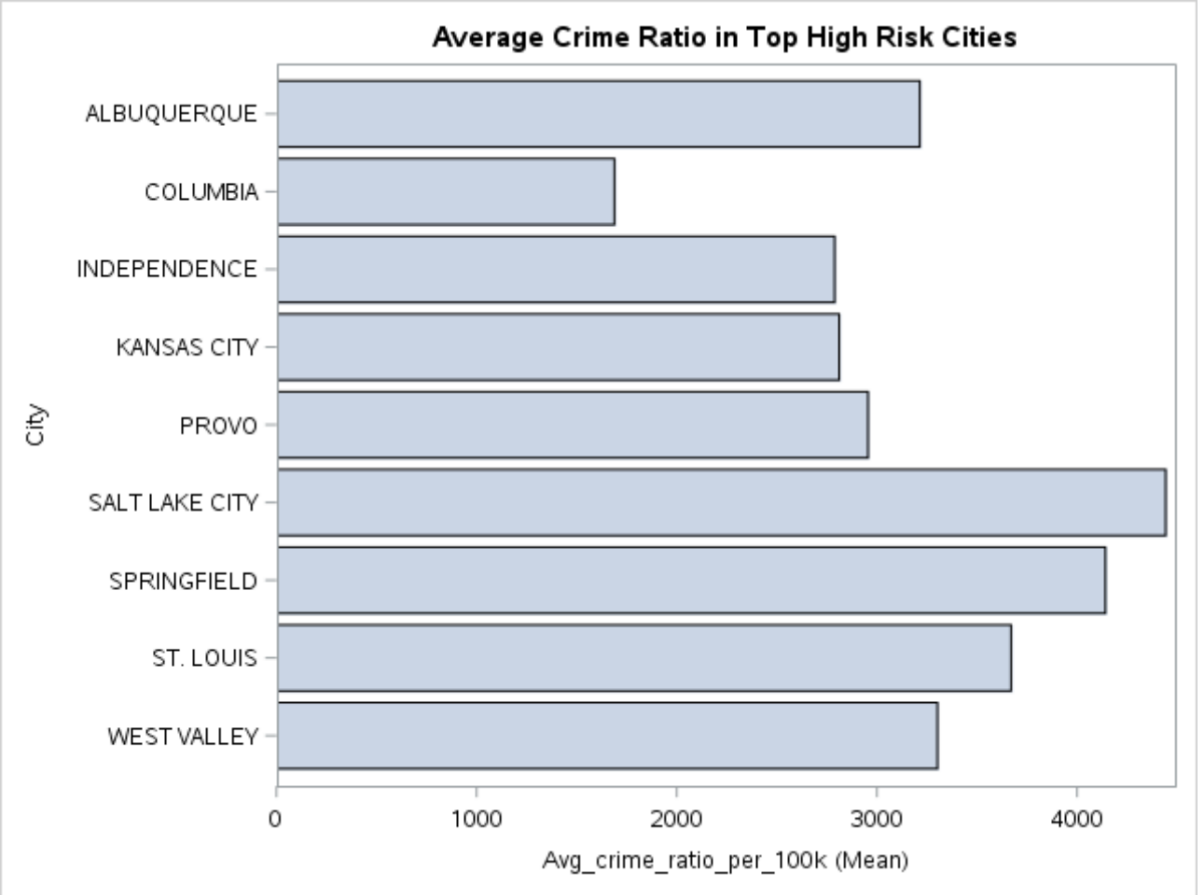
By putting all criminal activities together, we notice that all types of crime in high-risk states are higher than low-risk states (this is somehow expected as high-risk states should have higher crime rates in general regardless of the type of crime). However, we should also notice that for certain types of crime, the crime ratio in low-risk states tend to be close to the crime ratio in high-risk states. We can further analyze the crime ratio comparison between low and high risk states by taking the ratio of crime ratio in low-risk states to crime ratio in high-risk states. This means that despite we see the crime ratio in low-risk states is lower, for these types of crimes (arson, murder, robbery and rape) tend to have higher ratio to those in high-risk states (among the types of crimes).

Objective 4: To identify the top high/low risk city in the top states with highest/lowest crime rate

By utilizing PROC SQL and PROC TABULATE procedures in SAS, the crime data can be aggregated by city level to view the crime rate in each city in both low and high risk states.

Average Crime Ratio in Top High Risk Cities

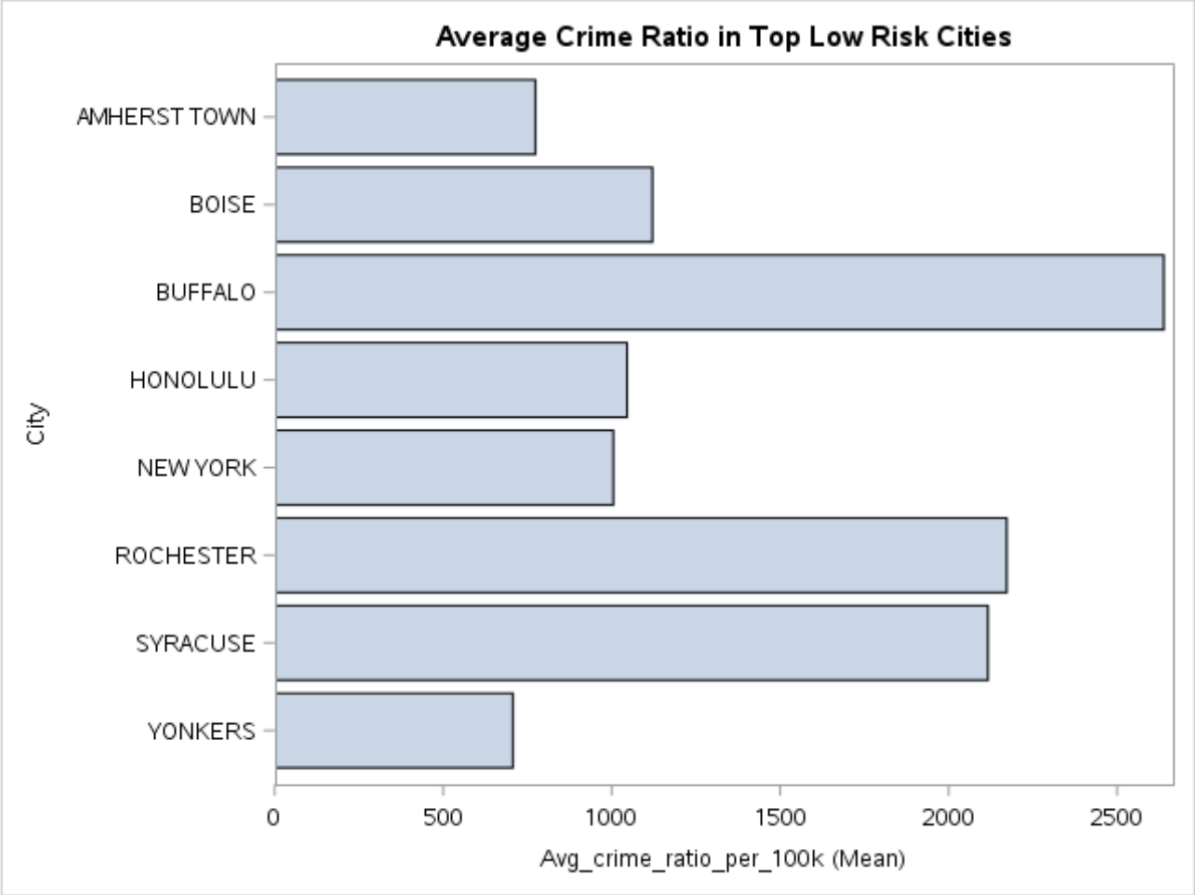
		Total_crime
		Average Crime Ratio per 100,000
State	City	
MISSOURI	COLUMBIA	1688.25
	INDEPENDENCE	2788.31
	KANSAS CITY	2811.03
	SPRINGFIELD	4141.19
	ST. LOUIS	3669.76
NEW MEXICO	ALBUQUERQUE	3213.82
UTAH	PROVO	2955.32
	SALT LAKE CITY	4442.33
	WEST VALLEY	3302.03



Based on the aggregated table and graph above, the top 3 high risk cities in Utah, New Mexico and Missouri are Salt Lake City, Springfield, St. Louis where these are cities with highest crime ratio in overall (during 2014 and 2015).

Average Crime Ratio in Top Low Risk Cities

State	City	Total_crime Average Crime Ratio per 100,000
HAWAII	HONOLULU	1045.24
IDAHO	BOISE	1120.65
NEW YORK	AMHERST TOWN	773.26
	BUFFALO	2638.09
	NEW YORK	1004.73
	ROCHESTER	2171.94
	SYRACUSE	2116.24
	YONKERS	706.51



Among the cities in the top low-risk states (i.e. Hawaii, New York and Idaho), Yonkers, Amherst Town and New York are the safest cities with lowest crime ratio as compared to other cities.

Objective 5: To analyze the crime details in the selected top cities

In this section, we analyze the crime details for the top cities that have been identified with highest and lowest crime rate.

- Top 3 high-risk cities: Salt Lake City > Springfield > St. Louis
- Top 3 low-risk cities: Yonkers > Amherst Town > New York

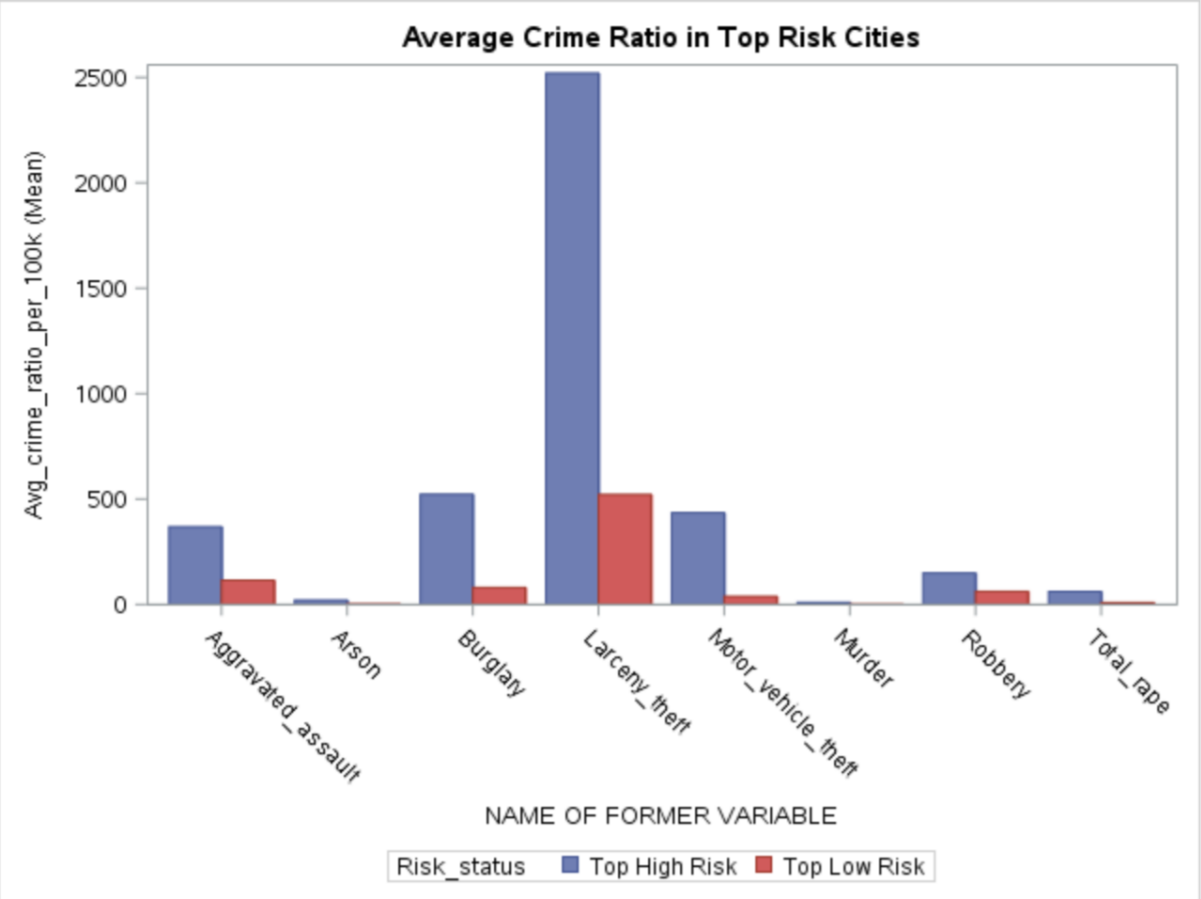
Average Property Crime in Top Risk Cities

			Burglary	Larceny_theft	Motor_vehicle_theft
			Avg Crime Ratio	Avg Crime Ratio	Avg Crime Ratio
Risk_status	State	City			
Top High Risk	MISSOURI	SPRINGFIELD	542.03	2631.13	346.21
		ST. LOUIS	578.81	1792.22	477.49
	UTAH	SALT LAKE CITY	446.35	3136.42	480.05
Top Low Risk	NEW YORK	AMHERST TOWN	63.68	638.35	17.59
		NEW YORK	81.39	603.99	40.30
		YONKERS	95.33	321.74	56.55

Average Violent Crime in Top Risk Cities

			Aggravated_assault	Murder	Robbery	Total_rape
			Avg Crime Ratio	Avg Crime Ratio	Avg Crime Ratio	Avg Crime Ratio
Risk_status	State	City				
Top High Risk	MISSOURI	SPRINGFIELD	402.48	3.61	118.39	79.56
		ST. LOUIS	506.84	23.43	222.79	42.38
	UTAH	SALT LAKE CITY	195.70	1.03	104.58	61.09
Top Low Risk	NEW YORK	AMHERST TOWN	27.67	0.42	17.19	5.03
		NEW YORK	174.35	1.82	89.86	12.68
		YONKERS	141.80	1.24	77.69	8.18

For all the cities involved, larceny theft contributed the most in all property-related crime activities, followed by burglary and motor vehicle theft. Among all violent crimes, aggravated assault is the main contributor to the crime rate in the cities, with average crime ratio per 100,000 as high as 402 cases in Springfield.

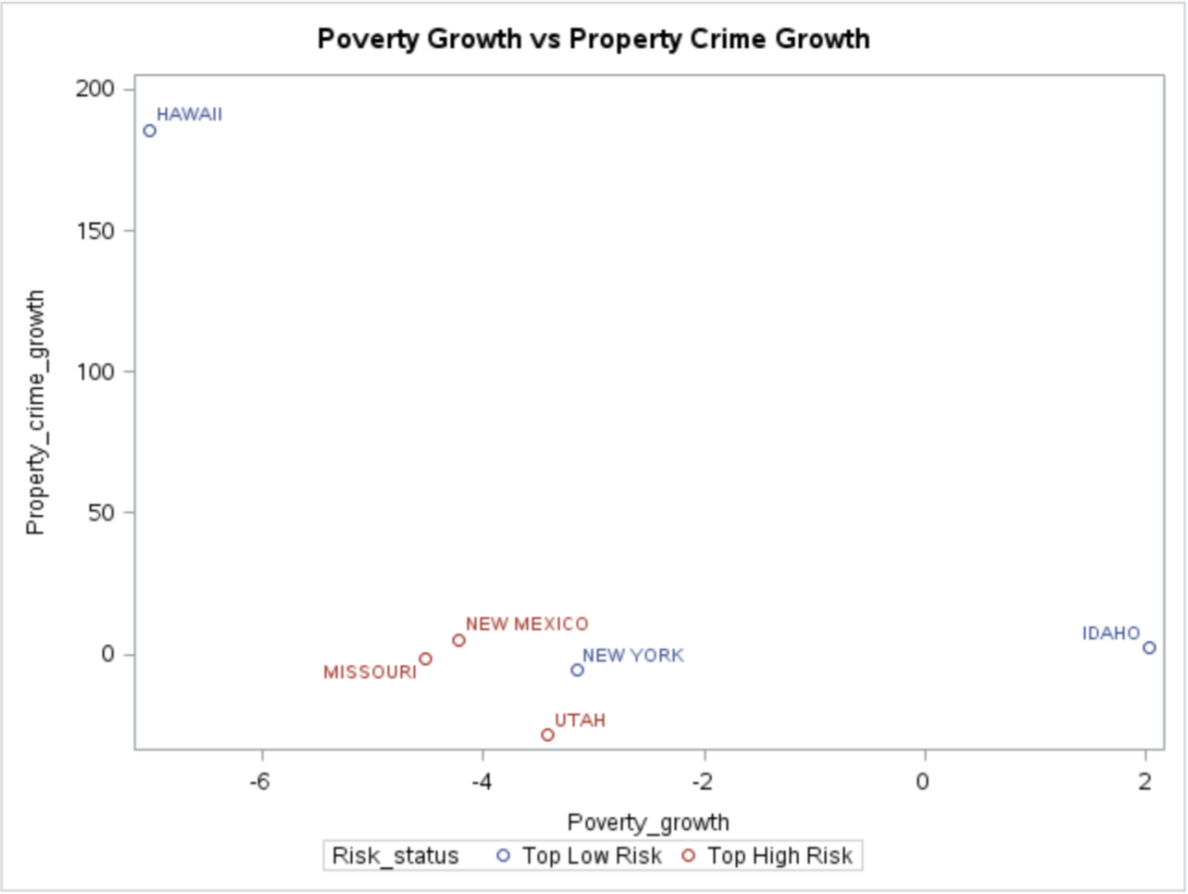


For all the cities, larceny theft seems to be very frequent to happen over the years. However, it is known that high-risk cities tend to have much higher crime ratio on average compared to low-risk cities. Besides, it is also known that arson, murder and rape activities have less frequency in both high and low risk cities compared to other types of crime like larceny theft and burglary.

Objective 6: To analyze how poverty relates to property crime rate

As far as the analyses have been carried out to achieve objectives 1 to 6, descriptive analytics have been covered. In previous section, it is discussed that descriptive analytics provide surface level analysis about the data while diagnostic analytics intended to discover the potential factors that have some impacts on the variables. In this section, we would like to discover and analyze how was the impact of poverty rate on the property crime rate, since a typical perception about high property crime rate is when people are poor. By comparing the top states with highest and lowest crime rate, we manage to figure out if the statement is true in this case.

Obs	State	Risk_status	Property_crime_growth	Poverty_growth
1	HAWAII	Top Low Risk	185.801	-7.01754
2	IDAHO	Top Low Risk	2.117	2.02703
3	MISSOURI	Top High Risk	-1.465	-4.51613
4	NEW MEXICO	Top High Risk	5.238	-4.22535
5	NEW YORK	Top Low Risk	-5.797	-3.14465
6	UTAH	Top High Risk	-28.562	-3.41880



The property crime growth represents the growth of property crime from 2014 to 2015, while the poverty growth represents the growth of poverty rate from 2014 to 2015. The states that fulfill the hypothetical statement “it tends to have higher property crime rate when poverty rate increases” are Idaho, Missouri, and Utah, which it does not apply to New York, New Mexico and Hawaii.

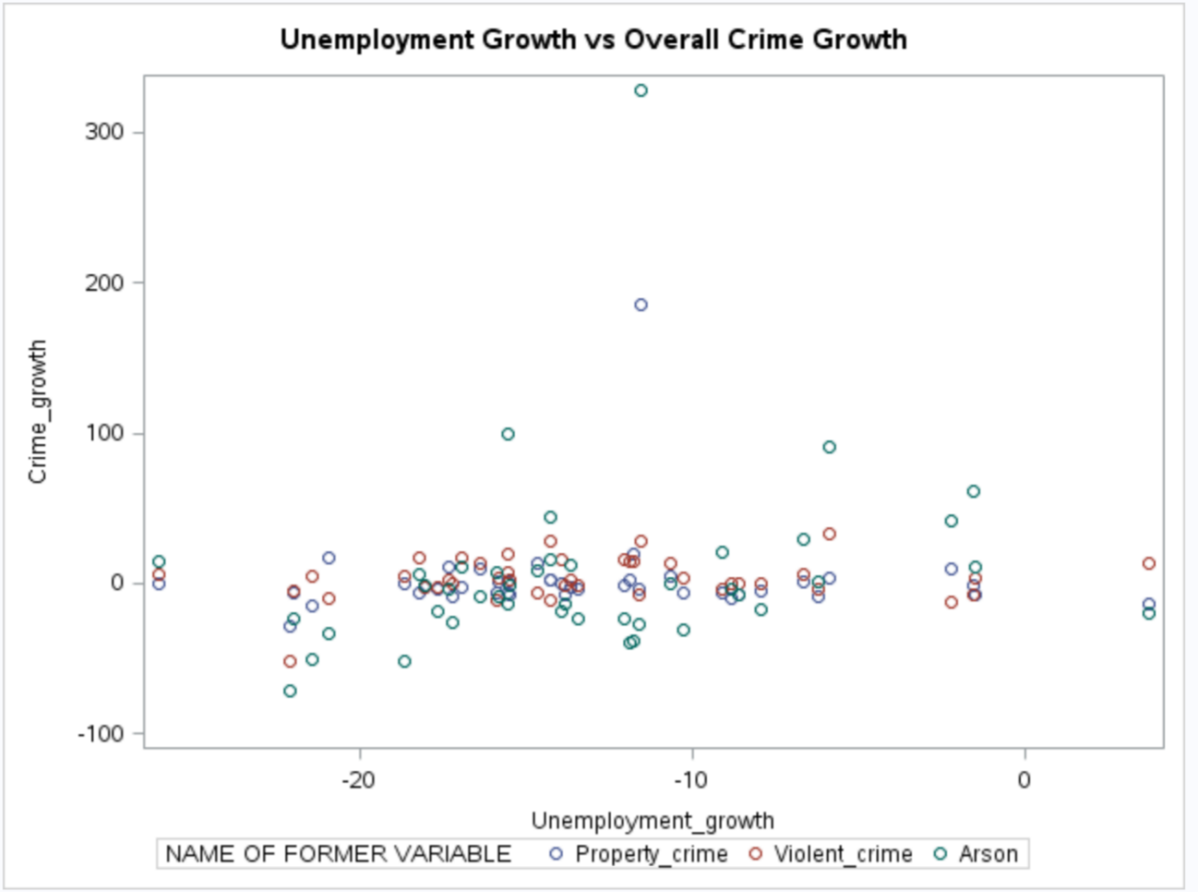
- For Idaho: when poverty rate increases, it tends to have a little increase in property crime rate, which incurred about 2%.
- For Missouri & Utah: when poverty rate decreases, it tends to have decline in property crime rate, which incurred about 1% and 29% respectively.
- For New York, New Mexico and Hawaii: when poverty rate decreases, it tends to have higher property crime rate instead, indicating that poverty might not be the important factor to explain property crime rate.

In addition, the entire scatterplot does not show a clear upward or downward trend. This means the 2 variables do not have correlation or strong relationship between each other. However, conclusion can hardly be made until data are sufficient and

statistical testing is carried out to test if the relationship between poverty and property crime is statistically significant.

Objective 7: To analyze how unemployment rate relates to overall crime rate

Lastly, we would also like to find out if there is any relationship between unemployment rate and overall crime rate. In other words, when there is a rise on unemployment and more people are getting unemployed from 2014 to 2015, does it tend to have an increase in overall crime rate. In fact, no obvious pattern (upward or downward trend) can be observed from the scatterplot, which indicates that unemployment might not have significant impact to the overall crime rate.



5. DISCUSSION

The data analysis has been carried out to analyze the criminal activities in 43 states and 260 cities of U.S. in first half year of 2014 and 2015. The total number of crimes is identified to be in decreasing trend over the years, amounted 1502376 in 2014 and 1502069 in 2015 respectively. Despite the decreasing trend in overall crime, violent crime has increased throughout the years. To measure the crime rate better, crime ratio was used instead of the total number of crime. Top 3 states / citys in U.S. with highest and top 3 states / cities with lowest crime ratio have been identified in the following order:

- Top 3 High Risk States: Utah > New Mexico > Missouri
- Top 3 Low Risk States: Hawaii > New York > Idaho
- Top 3 High-risk cities: Salt Lake City > Springfield > St. Louis
- Top 3 Low-risk cities: Yonkers > Amherst Town > New York

Among various types of crime, larceny theft is the most common criminal activity happen in either high or low risk states and cities. Despite the highest crime rate is larceny theft, it is also important to note that for crimes like arson, murder, robbery and rape tend to have higher ratio of crime rate in low-risk to high-risk states, indicating that these crimes appeared to be more frequent compared to other types in low-risk states.

Finally, external factors like poverty rate and unemployment rate have also been taken into the analysis to unveil the impact of these factors to the crime rate. Both external factors does not seem to have direct and significant impact to the crime rate based on the initial exploratory data analysis. However, further statistical testing can be performed to test the relationship, which is out of the scope for this assignment.

6. CONCLUSION

The primary aim of the assignment is to apply data analytics given the preliminary data of crime-related activities to come out with meaningful insights and direction that facilitates the investigation process. The data pre-processing and analysis have been carried out using SAS. For initial stage of data cleaning, Python has been used to clean up the raw data and output a well-structured table so further processing and analysis can be done in later stage. The Python and SAS scripts that have been used for achieving the objectives of this assignment have been uploaded on Github and accessible via https://github.com/tp042400/DAP-Assignment_CT050-3-M. Last but not least, this assignment also illustrates how data analytics can be applied in countering crimes, which discovers the potential to reduce the crime rates in United States.

7. REFERENCE

[1] SAS Product Documentation [WWW Document], n.d. URL http://support.sas.com/documentation/ [Accessed 5.1.2017].

[2] World Bank Organization, URL http://data.worldbank.org/indicator/SP.POP.GROW?locations=US [Accessed 5.1.2017].

[3] Bureau of Labor Statistics, URL https://www.bls.gov/lau/lastrk14.htm [Accessed 5.9.2017].

[4] Bishaw, A., Glassman, B. (2016) *Poverty: 2014 and 2015.* [Online] Available from: https://www.census.gov/content/dam/Census/library/publications/2016/demo/acsbr15-01.pdf [Accessed 5.9.2017].