Christian Campbell - Milestone 2

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
og_crime_data = pd.read_csv("C:/Users/Owner/Documents/Important/Bellevue/5 - Data Preparation/Project/cri
og_crime_data
```

Out[2]:

	jurisdiction	includes_jails	year	prisoner_count	crime_reporting_change	crimes_estimated	state_population	violent_crim
0	FEDERAL	False	2001	149852	NaN	NaN	NaN	
1	ALABAMA	False	2001	24741	False	False	4468912.0	1
2	ALASKA	True	2001	4570	False	False	633630.0	
3	ARIZONA	False	2001	27710	False	False	5306966.0	2
4	ARKANSAS	False	2001	11489	False	False	2694698.0	1
...	...	...	...	...	...	...	...	
811	VIRGINIA	False	2016	29882	False	False	8414380.0	1
812	WASHINGTON	False	2016	17228	False	False	7280934.0	2
813	WEST VIRGINIA	False	2016	5899	False	False	1828637.0	
814	WISCONSIN	False	2016	23163	False	False	5772917.0	1
815	WYOMING	False	2016	2352	False	False	584910.0	

816 rows × 17 columns

1.

In [3]:
```
# In this step I will be deleting the rape_revised column.
# The database contains data from 2001 to 2016 however the rape_revised column has null values till 2012.
# As such, I find that column to be incomplete. Which is why I choose to delete it.
```

In [4]:
```
step1_crime_data = og_crime_data.drop('rape_revised', axis=1)
step1_crime_data
```

Out[4]:

	jurisdiction	includes_jails	year	prisoner_count	crime_reporting_change	crimes_estimated	state_population	violent_crim
0	FEDERAL	False	2001	149852	NaN	NaN	NaN	
1	ALABAMA	False	2001	24741	False	False	4468912.0	1
2	ALASKA	True	2001	4570	False	False	633630.0	
3	ARIZONA	False	2001	27710	False	False	5306966.0	2
4	ARKANSAS	False	2001	11489	False	False	2694698.0	1
...	...	...	...	...	...	...	...	
811	VIRGINIA	False	2016	29882	False	False	8414380.0	1
812	WASHINGTON	False	2016	17228	False	False	7280934.0	2
813	WEST VIRGINIA	False	2016	5899	False	False	1828637.0	
814	WISCONSIN	False	2016	23163	False	False	5772917.0	1
815	WYOMING	False	2016	2352	False	False	584910.0	

816 rows × 16 columns

2.

In [5]:
```
# The rows labelled Federal contain only 4 columns of data.
# The rest of the row contains null values.
# For the purposes of this project, the "Federal" rows are not needed. Which is why I'm deleting them.
```

In [6]:
```
step2_crime_data = step1_crime_data[step1_crime_data['jurisdiction'] != 'FEDERAL']
step2_crime_data
```

Out[6]:

	jurisdiction	includes_jails	year	prisoner_count	crime_reporting_change	crimes_estimated	state_population	violent_crim
1	ALABAMA	False	2001	24741	False	False	4468912.0	1
2	ALASKA	True	2001	4570	False	False	633630.0	
3	ARIZONA	False	2001	27710	False	False	5306966.0	2
4	ARKANSAS	False	2001	11489	False	False	2694698.0	1
5	CALIFORNIA	False	2001	157142	False	False	34600463.0	21
...	...	...	...	...	...	...	...	
811	VIRGINIA	False	2016	29882	False	False	8414380.0	1
812	WASHINGTON	False	2016	17228	False	False	7280934.0	2
813	WEST VIRGINIA	False	2016	5899	False	False	1828637.0	
814	WISCONSIN	False	2016	23163	False	False	5772917.0	1
815	WYOMING	False	2016	2352	False	False	584910.0	

800 rows × 16 columns

3.

In [7]:
```
# I have decided to convert the values in the include_jails column to integers.
# False = 0 and True = 1
```

In [8]:
```
step3_crime_data = step2_crime_data.copy()
step3_crime_data['includes_jails'] = step3_crime_data['includes_jails'].astype(int)
step3_crime_data
```

Out[8]:

	jurisdiction	includes_jails	year	prisoner_count	crime_reporting_change	crimes_estimated	state_population	violent_
1	ALABAMA	0	2001	24741	False	False	4468912.0	
2	ALASKA	1	2001	4570	False	False	633630.0	
3	ARIZONA	0	2001	27710	False	False	5306966.0	
4	ARKANSAS	0	2001	11489	False	False	2694698.0	
5	CALIFORNIA	0	2001	157142	False	False	34600463.0	
...	...	...	...	...	...	...	...	
811	VIRGINIA	0	2016	29882	False	False	8414380.0	
812	WASHINGTON	0	2016	17228	False	False	7280934.0	
813	WEST VIRGINIA	0	2016	5899	False	False	1828637.0	
814	WISCONSIN	0	2016	23163	False	False	5772917.0	

4.

In [9]:
```
# I have decided to convert the values in the crime_reporting_change column to integers.
# False = 0 and True = 1
```

In [10]:
```
step4_crime_data = step3_crime_data.copy()  # Create a copy of the original DataFrame
step4_crime_data['crime_reporting_change'] = step4_crime_data['crime_reporting_change'].map({False: 0, Tr
step4_crime_data
```

Out[10]:

	jurisdiction	includes_jails	year	prisoner_count	crime_reporting_change	crimes_estimated	state_population	violent_
1	ALABAMA	0	2001	24741	0.0	False	4468912.0	
2	ALASKA	1	2001	4570	0.0	False	633630.0	
3	ARIZONA	0	2001	27710	0.0	False	5306966.0	
4	ARKANSAS	0	2001	11489	0.0	False	2694698.0	
5	CALIFORNIA	0	2001	157142	0.0	False	34600463.0	
...	...	...	...	...	...	...	...	
811	VIRGINIA	0	2016	29882	0.0	False	8414380.0	
812	WASHINGTON	0	2016	17228	0.0	False	7280934.0	
813	WEST VIRGINIA	0	2016	5899	0.0	False	1828637.0	
814	WISCONSIN	0	2016	23163	0.0	False	5772917.0	

5.

In [11]:
```
# I have decided to convert the values in the crime_estimated column to integers.
# False = 0 and True = 1
```

In [12]:
```
final_crime_data = step4_crime_data.copy()
final_crime_data['crimes_estimated'] = step3_crime_data['crimes_estimated'].map({False: 0, True: 1})
final_crime_data
```

Out[12]:

	jurisdiction	includes_jails	year	prisoner_count	crime_reporting_change	crimes_estimated	state_population	violent_crim
1	ALABAMA	0	2001	24741	0.0	0.0	4468912.0	1
2	ALASKA	1	2001	4570	0.0	0.0	633630.0	
3	ARIZONA	0	2001	27710	0.0	0.0	5306966.0	2
4	ARKANSAS	0	2001	11489	0.0	0.0	2694698.0	1
5	CALIFORNIA	0	2001	157142	0.0	0.0	34600463.0	21
...	...	...	...	...	...	...	...	
811	VIRGINIA	0	2016	29882	0.0	0.0	8414380.0	1
812	WASHINGTON	0	2016	17228	0.0	0.0	7280934.0	2
813	WEST VIRGINIA	0	2016	5899	0.0	0.0	1828637.0	
814	WISCONSIN	0	2016	23163	0.0	0.0	5772917.0	1
815	WYOMING	0	2016	2352	0.0	0.0	584910.0	

800 rows × 16 columns

Short Paragraph

In [13]:
```
"""
The data for this database was pubished by the Bureau of Justice Statistics and the FBI Uniform Crime Rep
found it very difficult to determine which values/columns/rows or parts of the dataset needed to be trans
It is clear by looking at the original dataset that meticulous effort was put into publishing a well poli
only minor transformations were needed on my part.

As stated above, the data in this dataset was published by well respected US institutions. And for the mo
realities on the ground. As such, any data wrangling that is done to this dataset must be done in a way t
the data that has been provided to us.

"""
```

Out[13]:
```
"\nThe data for this database was pubished by the Bureau of Justice Statistics and the FBI Uniform Crim
e Reporting Program. I\nfound it very difficult to determine which values/columns/rows or parts of the
dataset needed to be transformed or cleaned. \nIt is clear by looking at the original dataset that meti
culous effort was put into publishing a well polished dataset. As such\nonly minor transformations were
needed on my part.\n\nAs stated above, the data in this dataset was published by well respected US inst
itutions. And for the most part reflect \nrealities on the ground. As such, any data wrangling that is
done to this dataset must be done in a way that it doesn't skew\nthe data that has been provided to u
s.\n\n"
```

In [14]:
```
import pandas as pd
import sqlite3 as sql
```

In [16]:
```
conn = sql.connect('flatfile_crime_data')
```

In [17]:
```
final_crime_data.to_sql("flatfile_crime_data", conn, index=False)
```

Out[17]:
```
800
```

In [18]:
```
conn.close()
```

In [ ]:
```

```