



Final Project Report ISQS- 6339 Business
Intelligence

CRIME ALONG WITH DRUG ABUSE DATA

Group – 8

Aniket Kumar
Anusha Papineni
Shafia Askari

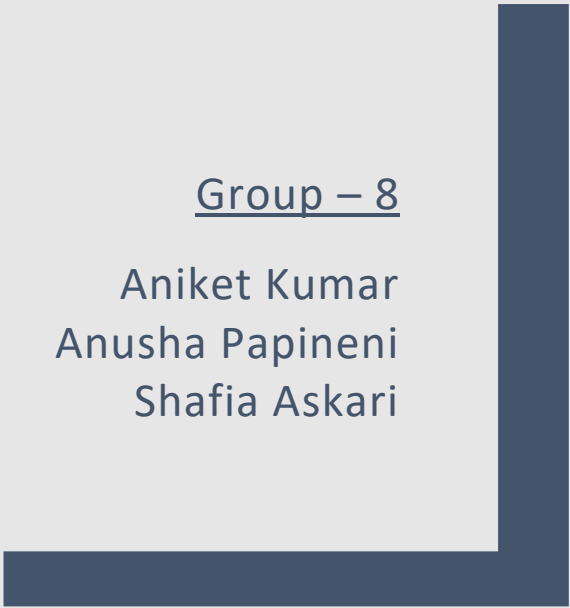


TABLE OF CONTENTS

- 1. Introduction**
- 2. Analysis of Data**
- 3. Data Cleaning (Before Merging)**
 - I. White spaces**
- 4. Data Merging**
- 5. Data Cleaning (After Merging)**
 - I. Duplicate columns**
 - II. Inconsistent data**
 - III. Calculated columns**
- 6. Analysis of Visualization**
- 7. Flow diagram**
- 8. Instructions for code**

Introduction

This project is done to analyze whether drugs, crime had an impact on economic condition in years 2001-2014 in U.S. Our findings highlight how each of these factors impact all states in U.S and to find correlation between drug deaths, crime categories, GDP per capita, employment rates, unemployment rates.

Analysis of Data

This project uses two datasets, "*drug_abuse_data*" & "*ucr_by_state*" to analyze the drug usage and crime statistics across the states in U.S.

The *drug_abuse_data* is an effort of an independent data source to gather data from Center for disease and control prevention (CDC) for drug related cause of death. contains mortality and population counts for all U.S. states 1997-2017. This Data is based on death certificates for U.S. residents by drug abuse.

The *ucr_by_state* is gathered from the Uniform Crime Report (UCR). The UCR has served as the FBI's primary national data collection tools to "acquire, collect, classify, and preserve identification, criminal identification, crime, and other records."

These two datasets provide U.S states statistics for death caused from drug, deaths caused by violent crimes such as robbery, car theft, rape, aggravated assault. In contrast these datasets also compare the GDP, employment rate and unemployment rate for U.S states across years. The factual data for drug abuse, crime numbers and cause valuable to our analysis for comparing the drug, crime and economic statistics for each state in each year.

These datasets variables are numerical and are particularly useful in aggregation functions. This data is can be used in. time series analysis for predicting future drug abuse and crime analysis in U.S states.

Data Cleaning (Before Merging)

At first glance the data quality seemed clean with no missing values. Later few cleaning activities such as stripping off white spaces, data type conversions were found to be necessary.

White spaces:

The columns '*state*' of "*drug_abuse_data*" and '*jurisdiction*' of "*ucr_by_state*" datasets have leading and trailing white spaces which leads to missing rows while merging the datasets and hence the white spaces are stripped off before merging the datasets.

Data Merging:

Both the datasets "*ucr_by_state*" and "*drug_abuse_data*" has factual data for states and years and they uniquely identify each row in the datasets. Therefore, they are merged on the columns

'jurisdiction' and 'year' in "ucr_by_state" and "state" and "year" columns in "drug_abuse_data" datasets by inner join.

There are no issues with the multilevel measurement in the merged dataset as both the datasets are on the same level of detail: state and year.

In our analysis we found that by keeping together employment_rate, unemployment_rate, gdp_per_capita, labor_force have significance in one data set prove useful in analyzing the increment and fall in economic situation related to crime and drugs usage in each state for each year.

After the datasets are merged, the combined datasets contain all the information related to each type of crime, drug abuse, economic situations of each year which can be used by to analyze the economic growth despite the law and order situation in each state and how the economic situation is effected by the drugs usage and crime in each state.

Data Cleaning (After merging):

Duplicate columns:

The column population is present in both the datasets named "population" in "drug_abuse_data" and "state_population" in "ucr_by_state" datasets with the same value.

The merged data contains state and year columns twice since they are present in both the datasets. Hence the columns "state_population", "jurisdiction", "year" are dropped from the merged dataset.

Inconsistent data:

The dataset "ucr_state" contained the character ',' (comma) in few numeric columns (*violent_crime_total*, *murder_manslaughter*, *rape_legacy*, *robbery*, *agg_assault*, *property_crime_total*, *burglary*, *larceny*, *vehicle_theft*) and are treated as object type by pandas. This character is removed from the values and are the columns are converted to numeric datatype.

Also, the columns "state_code" and "year" from "drug_abuse_data" are considered as numeric type instead of object type. These are converted to object type.

Calculated Columns:

Few additional columns "Total_crime" and "Pct_crime" are calculated for more insight into the data.

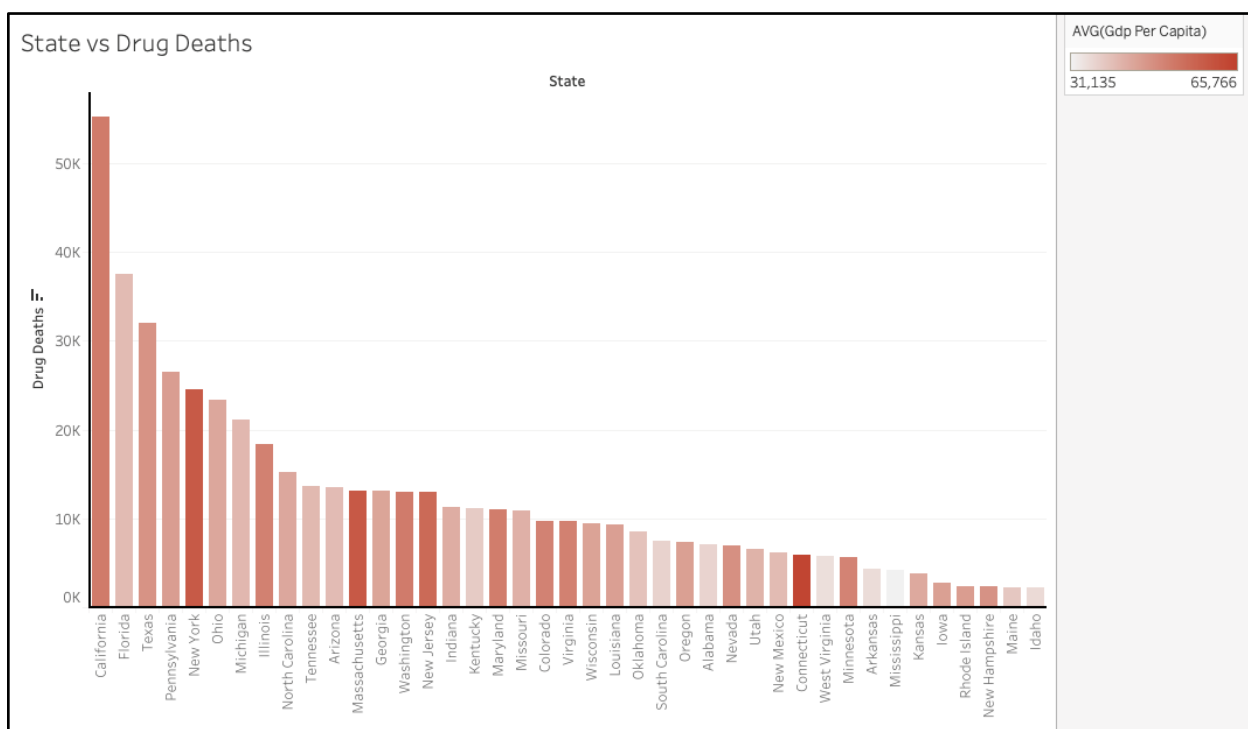
"Total_crime" is the sum of all the crimes present i.e. *violent_crime_total*, *murder_manslaughter*, *rape_legacy*, *rape_revised*, *robbery*, *agg_assault*, *property_crime_total*, *burglary*, *larceny*, *vehicle_theft* for each state and year.

"Pct_crime" is the percentage change in the "Total_crime" every year for each state.

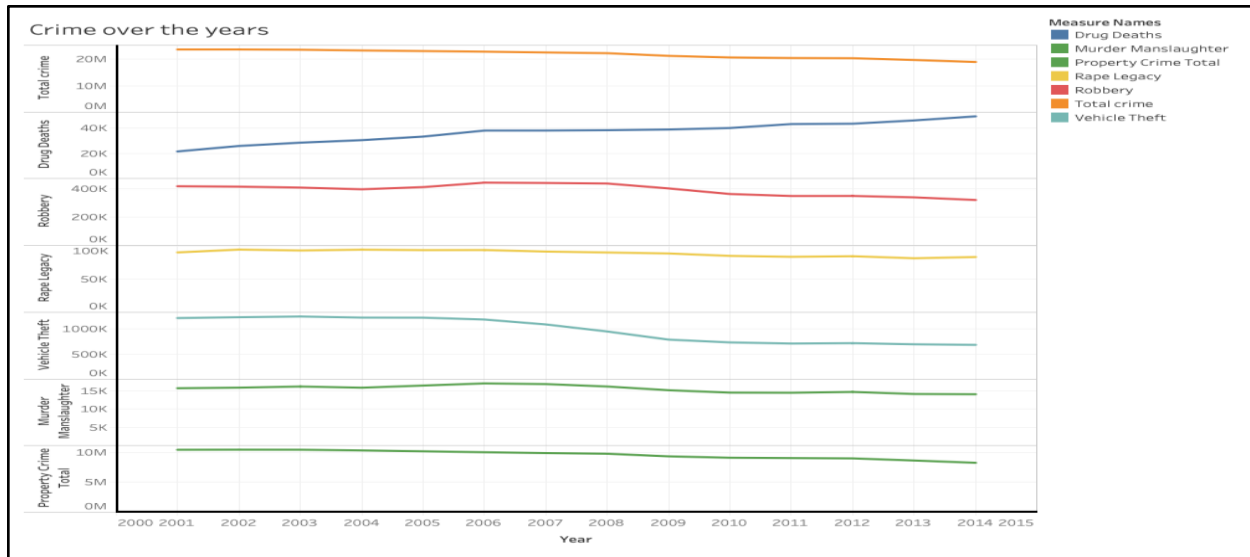
Analysis of Visualization

Dashboards:

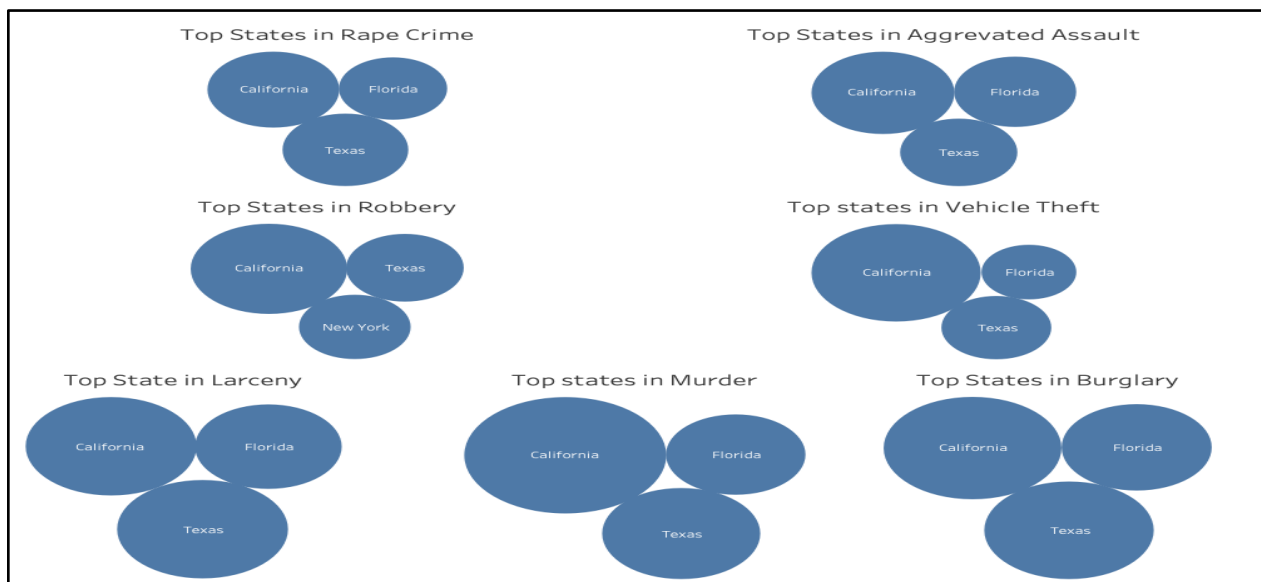
This visualization represents the total drug deaths in each state and the GDP per capita of that state. The length of the histogram bar represents the drug death total, and the color of the bar indicates the GDP per capita of the state (dark: high GDP--light: low GDP). This chart indicates no consistent relationship trend between GDP per capita and drug deaths.



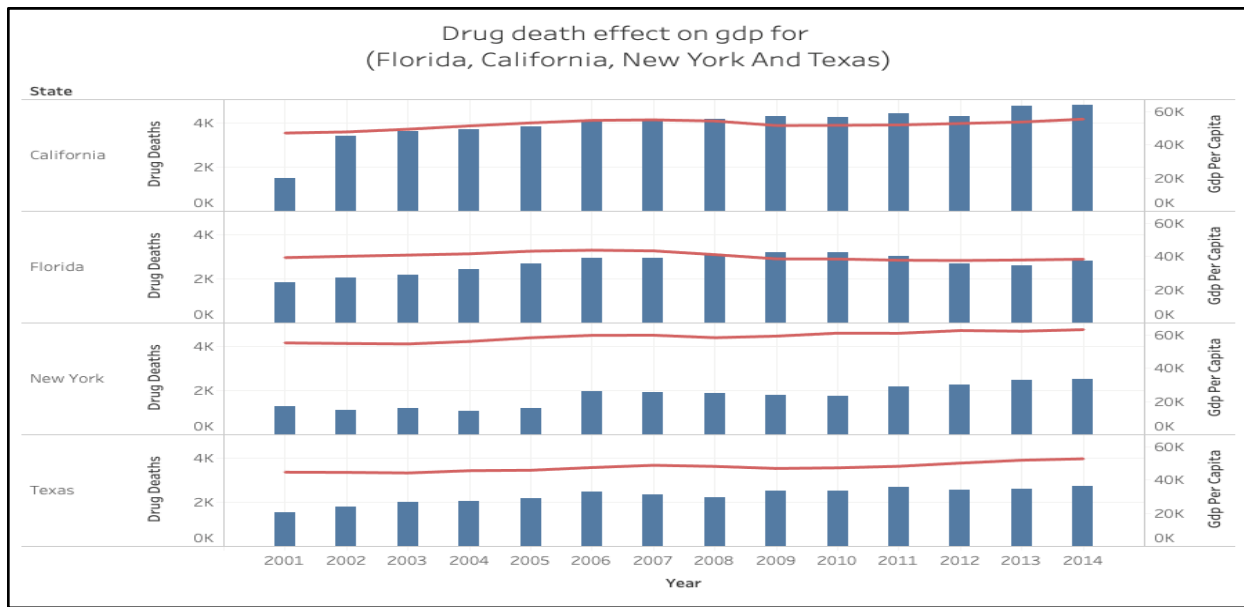
This dashboard visualizes the trend of total crime, Drug death, robbery, rape, vehicle theft, murder, property crime each year from 2000-2014. The trend compares each of category of crime over the years. The overall crime rate is in a decreasing trend. However, if you see the drug deaths is in an increasing trend.



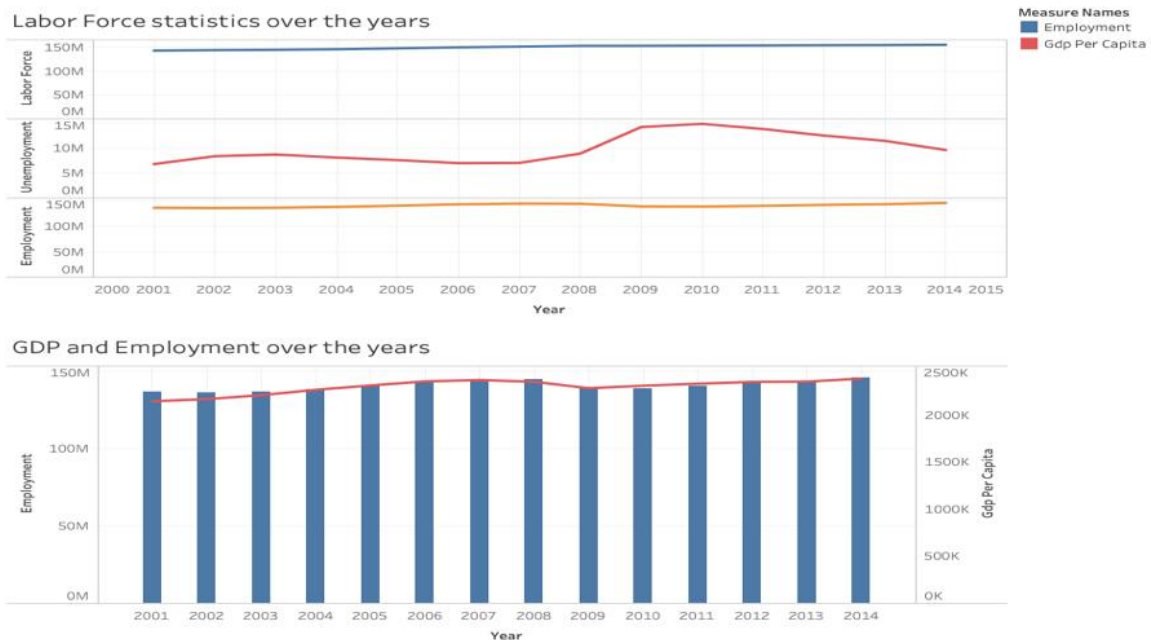
The dashboard represents states which have the highest cases of Rape, Aggravated assault, Robbery, vehicles e theft, larceny, murder and burglary. California, Texas, Florida and New York are the states which are common among all the state categories.



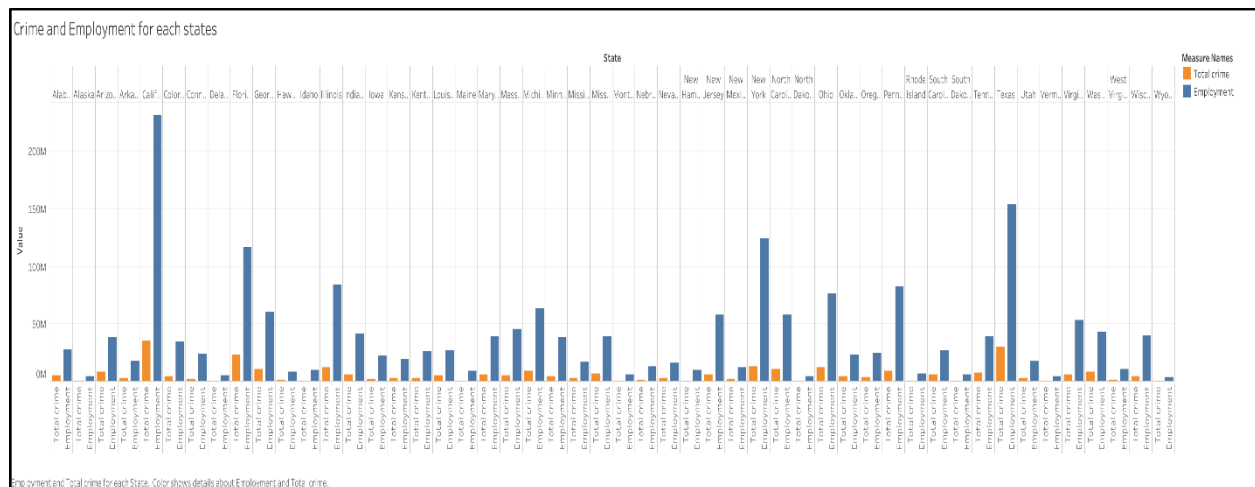
This dashboard represents an analysis of California, Florida, New York and Texas and the trend of drug deaths in these states throughout these years and also comparing its impact on GDP. For states like New York and Texas drug deaths does not have any impact on GDP as it remains more or less constant throughout irrespective of increase or decrease in the drug deaths.



This visualization represents an analysis of GDP and employment. And Labor statistics over the years. The top shows the overview of total labor force, employment, unemployment over the years 2001-2014. The bottom chart shows how employment number represented by histogram changes with GDP which is represented by the line graph. The GDP and employment do not indicate a consistent relationship between GDP and Employment



This visualization represents a comparison of the count of Total crime and Employment of each state in U.S. By looking at the trend in this chart, it is analyzed that cities such as California, New York, Florida and Texas have highest drug deaths and highest employment population too.



Principles of good Visualization:

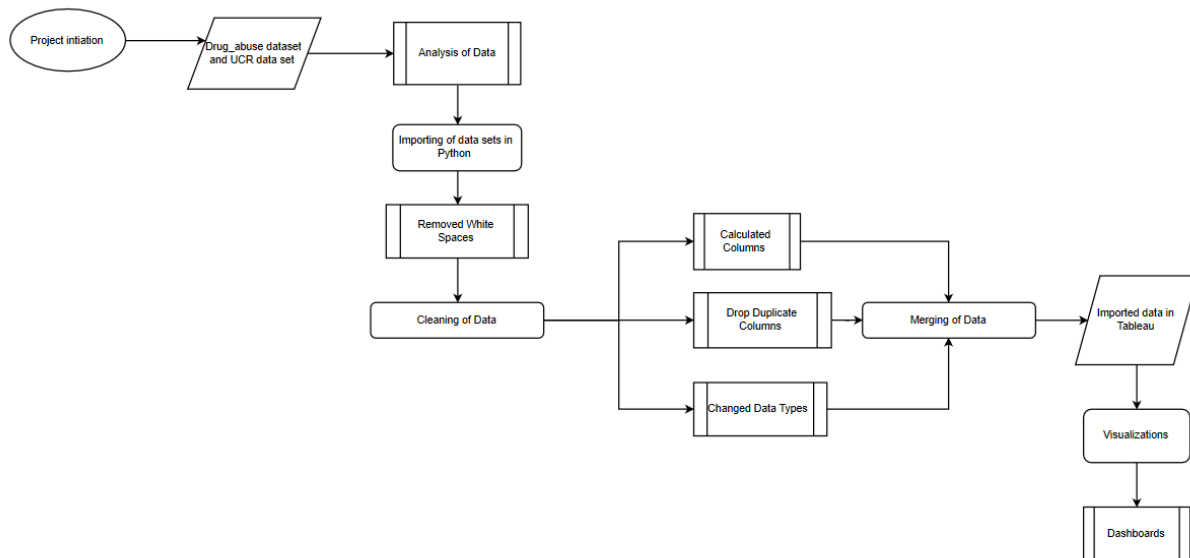
“Visualizing information can give us a very quick solution to problems. We can get clarity or the answer to a simple problem very quickly”.

Similarly, the main goal of visualization was to aggregate of attributes such as total crime and total drug deaths across dimensions: year and states. The key principles that were followed in all visualizations were to create easy to understand charts such as line charts, sorted histograms, and packed bubbles. Distinct bright colors are used in graphs to represent different attributes. Every graph includes labelled axis, legends, and color shade bar to map the color used with attributes and attribute values.

Concept of Natural Processing:

The values used in the graphs are the aggregated values of total crime, and drug deaths, and individual crime count and drug death of each year and state dimensions. This was a simple approach taken to give a holistic view and represent the number of drug death, total crime and individual crime across each state and year. This can be processed by Tableau’s natural language processing tool ask tableau using phrases “what is the total crime over time” OR “sum of drug death by state” OR “top 3 states by sum of total crime”

Flow diagram



Instructions for code

- The input files “*ucr_by_state*” and “*drug_abuse_data*” are downloaded from their respective websites and saved in a particular location of the system.
- Variables to store the path of these input files and output files are created and assigned at the beginning of the code. The value of the variables should be changed as per the location where the input files are located.
- Read the files into variables
- Strip off white spaces from the fields 'jurisdiction' of “*ucr_by_state*” and 'state' of “*drug_abuse_data*” using strip () function
- Merge the two datasets based on 'state' and 'year'
- Drop the duplicate columns using drop()
- Run the for loop to remove the character comma (,) from the numerical columns using str.replace(',', '') which are in object type and then run the for loop convert them to numeric type using astype('float64')
- Create calculated column 'Total_crime' as sum of crimes
- Add calculated column 'Pct_crime' using groupby('state').pct_change()
- export the merged file to csv using to_csv()

Group 8 Members: - Aniket Kumar, Anusha Papineni, Shafia Askari

Types of Attributes: -

Categorical	Discrete	Continuous
state	drug_deaths	unemployment_rate
state_code	population	cpi_all_urban_consumers
year	labor_force	Pct_crime
crime_reporting_change	employment	
crimes_estimated	unemployment	
	gdp_per_capita	
	violent_crime_total	
	murder_manslaughter	
	rape_legacy	
	robbery	
	agg_assault	
	property_crime_total	
	burglary	
	larceny	
	vehicle_theft	
	Total_crime	

Correlation Matrix: -

index	drug_deaths	population	labor_force	employment	unemployment	unemployment	cpi_all_urban_consumers	gdp_per_capita	crime_reporting_change	crimes_estimated	violent_crime_total	murder_manslaughter	rape_legacy	robbery	agg_assault	property_crime_total	burglary	larceny	vehicle_theft	Total_crime	Pct_crime
drug_deaths	1	0.9312055	0.9280287	0.9240654	0.9011565	0.3402135	0.1936834	0.0858229	0.0223276	0.0855419	0.8890923	0.8830328	0.877496	0.8846541	0.868913	0.8777921	0.8746743	0.8744472	0.7784561	0.8828007	-0.0815
population	0.9312055	1	0.9991664	0.998441	0.9286064	0.2437603	0.0315356	0.1332749	0.0239462	0.122519	0.960269	0.9543453	0.9283977	0.9715876	0.9333055	0.9531226	0.9248094	0.9523029	0.8716213	0.9579363	-0.05018
labor_force	0.9280287	0.9991664	1	0.9994897	0.9268458	0.2360126	0.0250561	0.1479501	0.0218985	0.130052	0.9571179	0.950284	0.9272864	0.9709379	0.9288535	0.9504546	0.9196811	0.9502923	0.8702104	0.9551968	-0.05145
employment	0.9240654	0.998441	0.9994897	1	0.9143804	0.2163339	0.015541	0.1500499	0.0202069	0.1292088	0.9590156	0.9517362	0.9294292	0.972416	0.9309306	0.9525495	0.9198215	0.9528661	0.8729264	0.9572739	-0.04907
unemployment	0.9011565	0.9286064	0.9268458	0.9143804	1	0.4486298	0.1349109	0.1115591	0.0400479	0.1296692	0.8590548	0.8579979	0.8287036	0.8767121	0.8309192	0.8506009	0.8452371	0.8448225	0.7694079	0.8551781	-0.0749
unemployment_rate	0.3402135	0.2437603	0.2360126	0.2163339	0.4486298	1	0.4173523	-0.104888	0.0738408	0.0724825	0.1930175	0.2117871	0.2142861	0.1853683	0.1896183	0.1970629	0.2466703	0.1940935	0.1060296	0.1972825	-0.11291
cpi_all_urban_consumers	0.1936834	0.0315356	0.0250561	0.015541	0.1349109	0.4173523	1	0.1407654	0.0635765	-0.039335	-0.046004	-0.034808	-0.042828	-0.048885	-0.050136	-0.062589	-0.025128	-0.056253	-0.141167	-0.060842	-0.17001
gdp_per_capita	0.0858229	0.1332749	0.1479501	0.1500499	0.1115591	-0.104888	0.1407654	1	-0.001303	0.0429698	0.0903546	0.0550782	0.029005	0.1494182	0.062204	0.0438928	-0.007979	0.0509667	0.08729	0.0504068	-0.02164
crime_reporting_change	0.0223276	0.0239462	0.0218985	0.0202069	0.0400479	0.0738408	0.0635765	-0.001303	1	-0.010335	0.0147237	0.0042382	-0.009381	0.0194212	0.0145129	-0.000383	-0.003539	0.0052068	-0.02051	0.0016952	0.018910
crimes_estimated	0.0855419	0.122519	0.130052	0.1292088	0.1296692	0.0724825	-0.039335	0.0429698	-0.010335	1	0.1305472	0.1531185	0.1334055	0.1494613	0.1166937	0.0948419	0.0754359	0.1091975	0.0478053	0.1001113	-0.06459
violent_crime_total	0.8890923	0.960269	0.9571179	0.9590156	0.8590548	0.1930175	-0.046004	0.0903546	0.0147237	0.1305472	1	0.9711909	0.9387204	0.9790277	0.9939001	0.9654613	0.9412939	0.9576192	0.9073561	0.9741316	-0.01848
murder_manslaughter	0.8830328	0.9543453	0.950284	0.9517362	0.8579979	0.2117871	-0.034808	0.0550782	0.0042382	0.1531185	0.9711909	1	0.9202119	0.9691124	0.9535878	0.9530657	0.9313155	0.941263	0.9106605	0.9594382	-0.01827
rape_legacy	0.877496	0.9283977	0.9272864	0.9294292	0.8287036	0.2142861	-0.042828	0.029005	-0.009381	0.1334055	0.9387204	0.9202119	1	0.9080722	0.9275796	0.961533	0.9464991	0.9594545	0.863772	0.9623186	-0.02209
robbery	0.8846541	0.9715876	0.9709379	0.972416	0.8767121	0.1853683	-0.048885	0.1494182	0.0194212	0.1494613	0.9790277	0.9691124	0.9080722	1	0.9518322	0.9455059	0.9138897	0.9366886	0.9061463	0.9539729	-0.02169
agg_assault	0.868913	0.9333055	0.9288535	0.9309306	0.8309192	0.1896183	-0.050136	0.062204	0.0145129	0.1166937	0.9939001	0.9535878	0.9275796	0.9518322	1	0.9541959	0.9338633	0.9463446	0.8916628	0.9635298	-0.0135
property_crime_total	0.8777921	0.9531226	0.9504546	0.9525495	0.8506009	0.1970629	0.062589	0.0438928	-0.000383	0.0948419	0.9654613	0.9530657	0.961533	0.9455059	0.9541959	1	0.9868004	0.9957062	0.9040985	0.9993648	-0.00232
burglary	0.8746743	0.9248094	0.9196811	0.9198215	0.8452371	0.2466703	-0.025128	-0.007979	-0.003539	0.0754359	0.9412939	0.9313155	0.9464991	0.9138897	0.9338633	0.9868004	1	0.9786781	0.8685977	0.9846171	-0.0011
larceny	0.8744472	0.9523029	0.9502923	0.9528661	0.8448225	0.1940935	-0.056253	0.0509667	0.0052068	0.1091975	0.9576192	0.941263	0.9594545	0.9366886	0.9463446	0.9957062	0.9786781	1	0.8677863	0.9945566	-0.00699
vehicle_theft	0.7784561	0.8716213	0.8702104	0.8729264	0.7694079	0.1060296	-0.141167	0.08729	-0.02051	0.0478053	0.9073561	0.9106605	0.863772	0.9061463	0.8916628	0.9040985	0.8685977	0.8677863	1	0.9082835	0.017172
Total_crime	0.8828007	0.9579363	0.9551968	0.9572739	0.8551781	0.1972825	-0.060842	0.0504068	0.0016952	0.1001113	0.9741316	0.9594382	0.9623186	0.9539729	0.9635298	0.9993648	0.9846171	0.9945566	0.9082835	1	-0.00442
Pct_crime	-0.0815	-0.050181	-0.051457	-0.049077	-0.07497	-0.112918	-0.170014	-0.021645	0.0189107	0.064594	-0.018488	-0.018271	-0.022095	-0.021697	-0.01352	-0.002328	-0.00115	-0.006999	0.0171725	-0.004428	1