Visualizing FBI Crime Data

Capstone Project

Matt DeVay - DSI-ATX-11

Problem Statement

Very few resources exist for examining and modeling crime data, particularly in the context of cannabis policy. Aggregated crime data is only readily available through the FBI Uniform Crime Reporting system.

Problem Background:

- Criminal Justice Reform is a major priority for most US states.
- Drug charges, and in particular, cannabis, account for the majority of arrests in the US
- Drug arrests disproportionately affect people of color
- Drug prohibition does not correlate to a decrease in crime

Despite these points, there are no widespread tools and visualizations to inform potential policy changes!

Even more simply. . .

47%

Of people arrested on drug charges are Black or Latino….

Despite the fact that they only make up

32%

of the population

Data Sources:

Data were downloaded directly from the FBI UCR/NIBRS website, available at https://www.fbi.gov/services/cjis/ucr. Data were available for the years 1999 - 2018.

Data Characteristics - the good:

- Data covers most metropolitan areas
- Data is typically very inclusive
- Raw data is very granular
- The UCR/NIBRS programs are well established

Data Characteristics - the bad:

- Data is reported voluntarily by individual jurisdictions
  - Not all jurisdictions report
  - Not all jurisdictions report all data
- Data only cover approximately 80% of the US population
- Data are published for public consumption via Excel Spreadsheets
- Raw data is in an obscure file format that is extremely complex

Data Summary

- Probably still useful for trend analysis and inference
- Impossible to quantify selection/observation bias of contributors
- Probably should be taken with a grain of salt for more academic purposes
- Incomplete, possibly in ways that we can never know

Problem Recap:

UCR data is complex, has novel and often inconsistent formatting, and is only available in formats that are unfriendly for general data analysis and modelling.

These problems inhibit performing comprehensive EDA and/or modeling.

Tools Used:

- Python
  - Dash/Plotly
  - Pandas
  - Numpy
  - Seaborn/Matplotlib
- Microsoft Excel
- QGIS

Process

- Import and clean data via Pandas
- Amend data with policy status
- Create Dashboard with Dash/Plotly
- Perform some EDA and recommend next steps

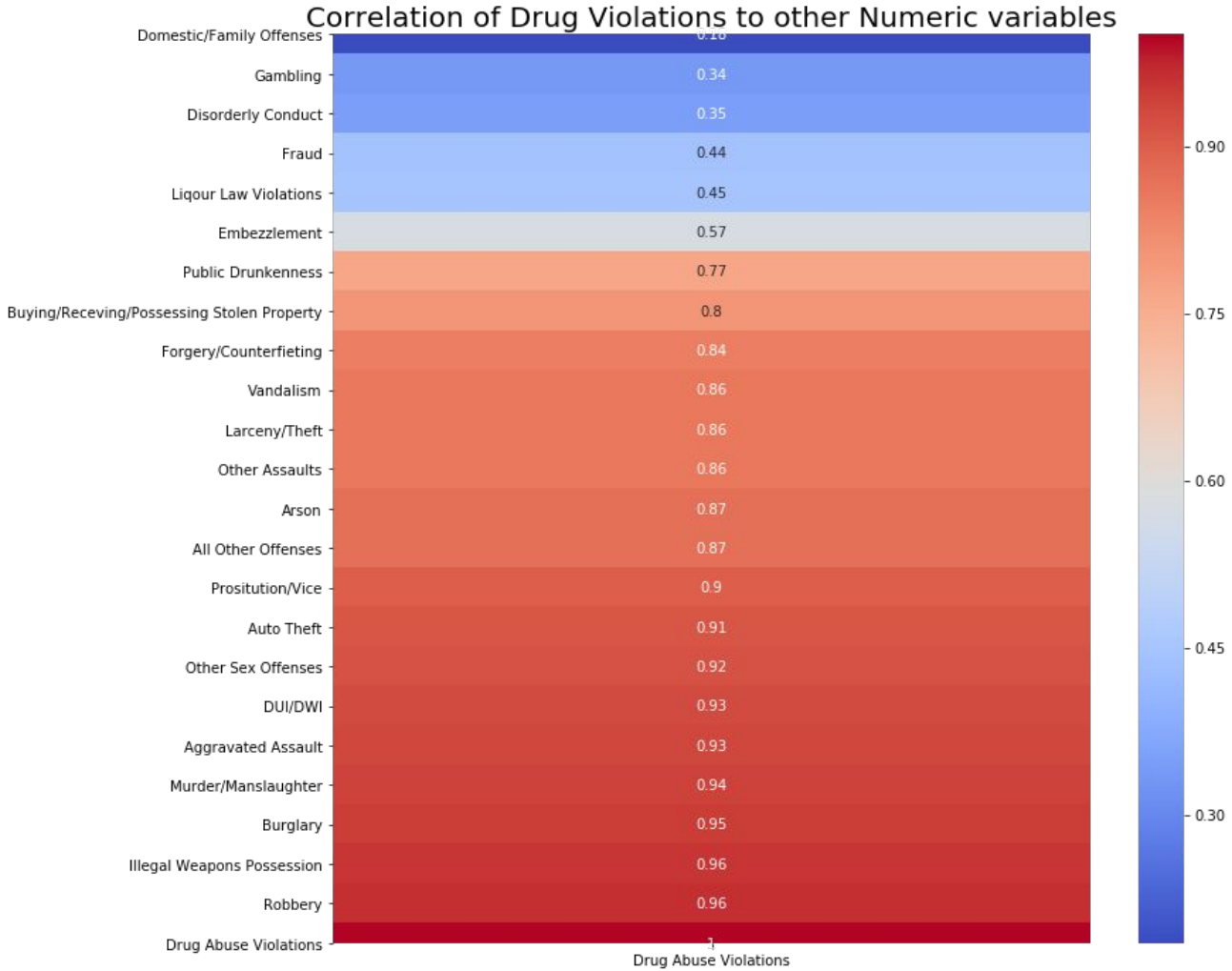
Import and Cleaning:

- Excel. ‘Nuff said
- Flatfiles not natively understood by Pandas, require significant custom delimiting
- Lots of weird landmines (leading spaces, inconsistent formatting, etc)
- Lots of regex and string manipulation

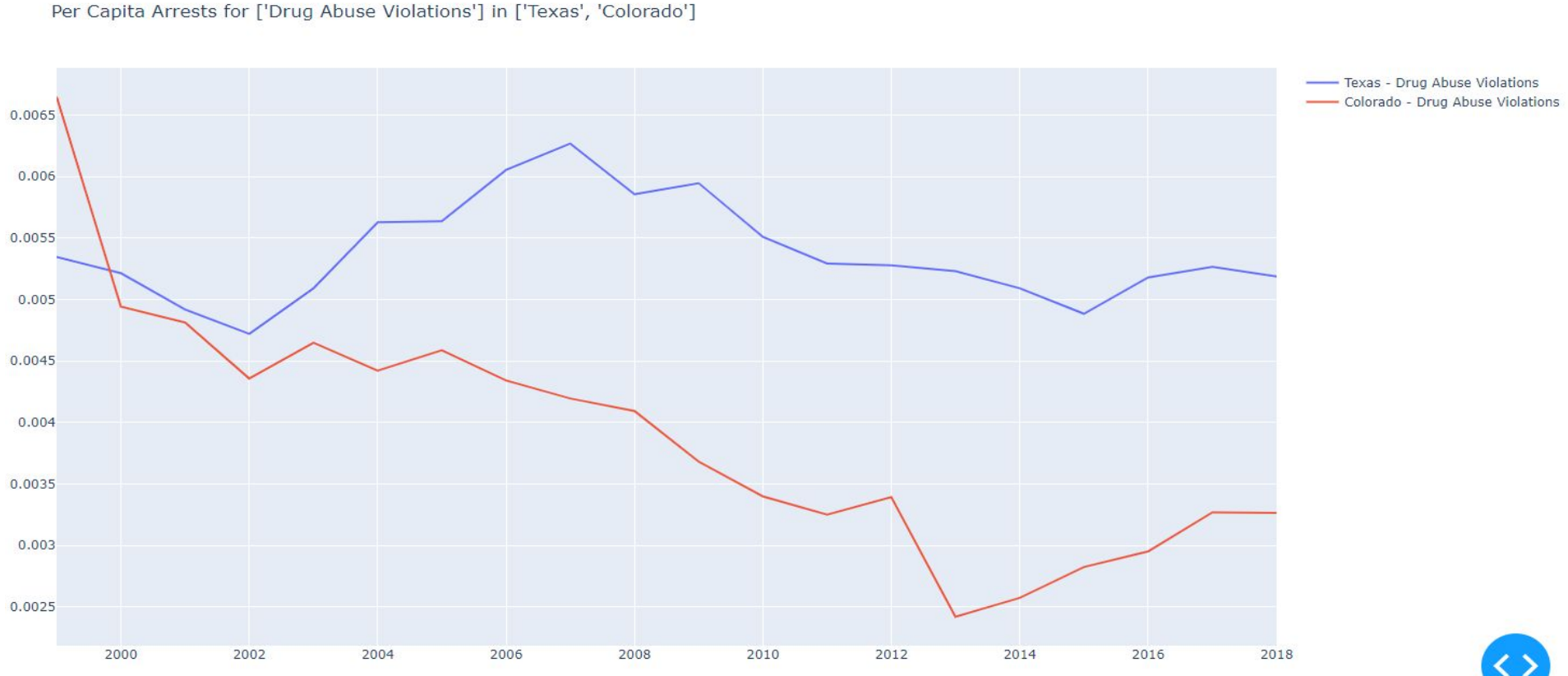
Data Amending:

- Policies were one-hot-encoded manually - Consolidated data for cannabis policies not readily available
- Feature names and strings had to be regex’d and re-formatted
- Two letter state codes were added to facilitate Plotly interactivity

Initial EDA:

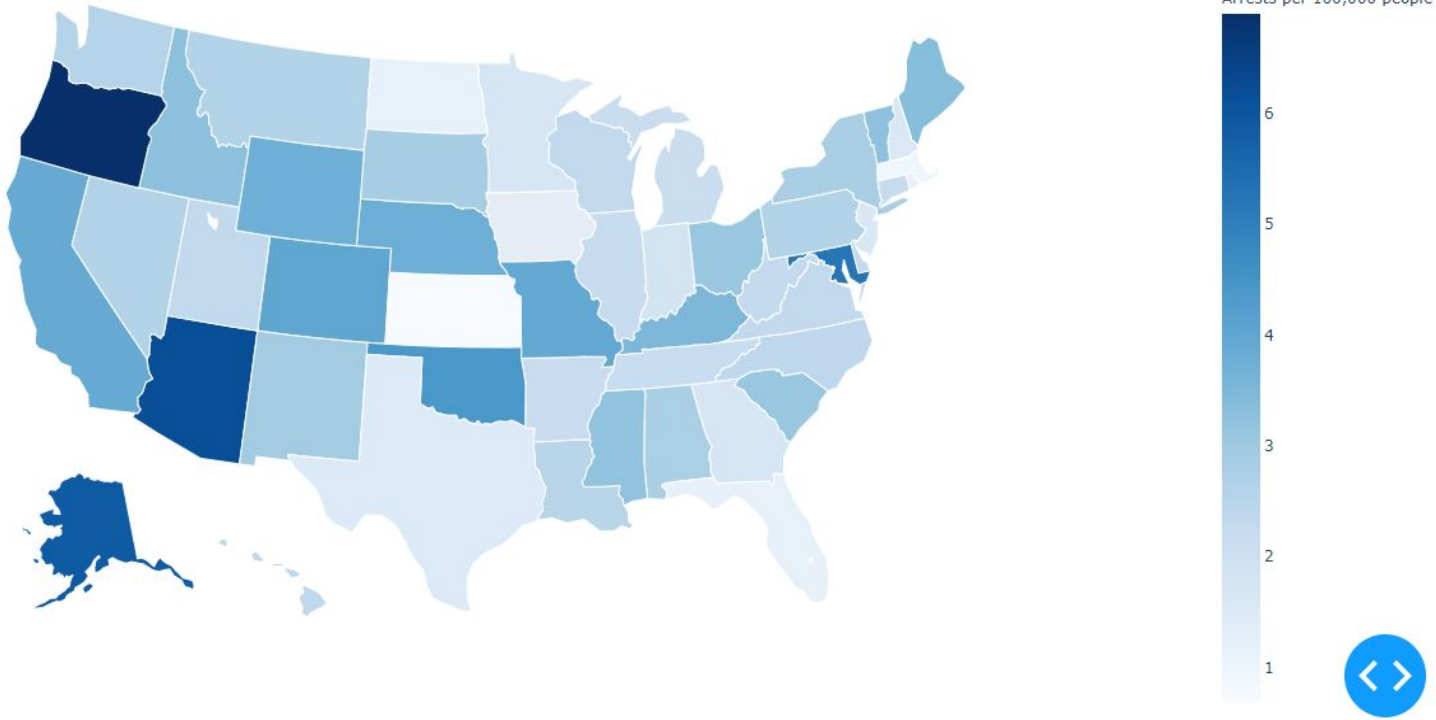

Correlation of Drug Violations to other Numeric variables
Domestic/Family Offenses
Gambling
0.34
Disorderly Conduct
0.35
Fraud
0.44
Liqour Law Violations
0.45
Embezzlement
0.57
Public Drunkenness
0.77
Buying/Receving/Possessing Stolen Property
0.8
Forgery/Counterfieting
0.84
Vandalism
0.86
Larceny/Theft
0.86
Other Assaults
0.86
Arson
0.87
All Other Offenses
0.87
Prositution/Vice
0.9
Auto Theft
0.91
Other Sex Offenses
0.92
DUI/DWI
0.93
Aggravated Assault
0.93
Murder/Manslaughter
0.94
Burglary
0.95
Illegal Weapons Possession
0.96
Robbery
0.96
Drug Abuse Violations
1
Drug Abuse Violations
0.90
0.75
0.60
0.45
0.30

Initial EDA contin:



DASH!!

2018, Arrest Rate for Arson



My App (still in progress)

Next Steps:

- Extract and aggregate the granular data from the raw flatfiles
- Improve functionality and range of dashboard
- Perform more robust EDA
- Perform VAR or RNN modelling on the raw data to forecast

Questions?