Problem:

Crime is something that is unpredictable and once it happens, it is too late. Predicting the exact time or type of crime committed is impossible, but what if we can determine if crime is more likely to happen under certain conditions? I would like to determine if the number of crimes is related to other factors such as population, weather, gas prices, health, income, or location. My hypothetical client would be any kind of law enforcement agency or government in general. They would care about this particular issue because they can assign more officers on patrol or be warier when certain observations are met. For an example, if it was discovered that after a major disaster such as a hurricane struck the number of theft-related crimes increased in that area. Increasing the number of officers on patrol in that particular area could help lower that number. Or if it was found that the average number of days of poor mental health in a month is correlated with a high number of crimes, allocating funds to mental health programs could indirectly lower crime rate.

To help with exploratory analysis, I will create an interactive visualization using the Dash package from Python. This interactive visualization could be used by a law enforcement agency to explore further or by a person who wishes to do additional research in this area. Next, I will build a statistical model to determine what factors best contributes to the total number of crimes.

Datasets

Crimes

I will be using multiple datasets from a variety of different sources.

The first type of data I will use is crime data. One dataset is located at https://crime-data-explorer.fr.cloud.gov/downloads-and-docs
It is the Summary(SRS) Data with estimates CSV file under the additional datasets. It contains aggregated crime count data at the state and national level from 1995 to 2016.

It includes variables such as year, state, population, violent crimes, homicide, rape, robbery, aggregated assault, property crime, burglary, larceny, and motor vehicle theft.

Weather

The second type of data I will use is weather data. I manually collected 50 datasets, one from each state in the USA at https://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp#
I selected the “State” tab, with the period from 01/1994 to 12/2017, and comma delimited text output. I saved each file as the name of the state.

The variable names can be found at the bottom of the page under Indices Output. A more extensive variable definition is available at: ftp://ftp.ncdc.noaa.gov/pub/data/cirs/climdiv/divisional-readme.txt

I am mainly interested in Year/Month, TMIN, TMAX, and TAVG. The latter three are in Fahrenheit.

Storm Events

I use another dataset located at ftp://ftp.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles
I downloaded the StormEvents_details file from 1994 to 2017. Some of the variable definitions can be found at the top in the Storm Data Export Format word document.

The variables of interest here other than the Year and Date are: EVENT_TYPE, INJURIES_DIRECT, INJURIES_INDIRECT, DEATHS_DIRECT, DEATHS_INDIRECT, DAMAGE_PROPERTY, and DAMAGE_CROPS. EVENT_TYPE is the type of event such as 'Tornado'. The other variables are self-explanatory.

Standard Precipitation Index

I use another dataset located at https://ephtracking.cdc.gov/download
I downloaded the Standard Precipitation Index (Pearson) file. It includes county-level data from 1895 to the present.

The only unique variable of interest is the Standard Precipitation Index (SPI).

Population

The third type of data I will use is population data. The dataset is located at
https://seer.cancer.gov/popdata/download.html#19
I downloaded the adjusted 1969-2016 White, Black, Other data County-level data. The adjusted data takes into account Hurricane Katrina and Rita.

The variable dictionary can be found at: https://seer.cancer.gov/popdata/popdic.html
The unique variables of interest are: population, race, sex, and age.

Income

The fourth type of data I will use is Income data. The dataset is located at
https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html
I downloaded the Table H-8 data representing the median household income by state. The data ranges from 1984 to 2016.

The only unique variable here is the median income of households.

Gas Prices

The fifth type of data I will use is Gas price data. The dataset is located at
https://www.eia.gov/state/seds/seds-data-complete.php?sid=US
Under data files, Prices and expenditures, 1970-2015, all states, I downloaded the CSV file for Prices.

The unique variable of interest is the gas price in dollars per million british thermal units.

Health

The last type of data I will use is the Behavioral Risk Factor Surveillance System data. It is available at
https://www.cdc.gov/brfss/annual_data/annual_1994.htm
For the year 1994. I use the data ranging from 1994 to 2016. I downloaded the XPT transport SAS files under each year. The extensive variable definitions are also viewable from this page.

Other Datasets

There other datasets that we can use here, but I did not find any more. There were several resources that required certain clearances to access. Any data that is aggregated by the year and state level can be used because that is the type of data I gathered. Data aggregated by the month, day, hour, minute, etc can also be used. By aggregating at the year level or downsampling would achieve the same effect.

Data Cleaning

Data cleaning is divided into two steps. Full details are in the jupyter notebooks.

Part 1

Weather

I load in each csv file using pandas, concatenating each consecutive data frame onto the previous. I Created a new column to store which what state the data represents.

I subsetted the data to include only relevant columns: State, YearMonth, TVG, TMIN, and TMAX.

I extracted the year and month fro the YearMonth column as separate columns and stored YearMonth as a datetime variable. I converted each column into their proper datatype and saved the cleaned data into a new csv file called Weather.

Income

I loaded in the data and noticed that the years were represented as individual columns. I melted the data frame and stored the years all into one column. I fixed the datatypes of each column and converted the date column to a datetime object. The cleaned data frame was saved to a new csv file called Income.

Population

I loaded in the data and used the variable dictionary to determine how to extract the data. The data was stored into one column and had no delimiter. I created new columns for each relevant variable: Year, State_Abbrev, Race, Sex, Age, and Population.

I subsetted the data to include only years past 1993. I created a new column called “Age_Group” to pool the given age categories into fewer categories. Ages from 0-19 are “Child/Teen”, 20-39 are “Young Adult”, 40-59 are “Older Adult”, and the last one is 60+. I corrected all datatypes.

First, I grouped by Year, State_Abbrev, Race, Sex, and Age_Group finding the aggregate sum Population. This will be used later and grouped in different ways to create different types of visualizations. I converted the multi indexed data frame into a regular data frame with each variable in its own column. I saved this to a csv file called Population2

Next, I grouped by Year and State_Abbrev finding the aggregate sum Population. This is the data frame that would be merged to the other datasets . I saved this to a csv file called Population

Storm Events

I loaded in the data files using a for loop, concatenating each consecutive data frame to the previous.

The data was subsetted to include only relevant variables: BEGIN_YEARMONTH, STATE, YEAR, EVENT_TYPE, INJURIIES_DIRECT, INJURIES_INDIRECT, DEATHS_DIRECT, DEATHS_INDIRECT, DAMAGE_PROPERTY, and DAMAGE_CROPS.

I extracted the month from the BEGIN_YEARMONTH column and stored it in a column called Month. All null values were dropped from the data frame so that the DAMAGE_PROPERTY and DAMAGE_CROPS columns could be cleaned.

Both of these problem columns contained values such as “10M”, “3.00K”, “K”, “05”, etc. I checked through the value_counts to find all the cases to account for. In general, I removed the letter from each row and depending on what letter was removed, the number remaining was multiplied by a number. For “K”, the multipler was 1000 for an example. If there was only a “h” in the column, it was treated as 0 because there were values that were “1.00K” meaning that the “h” did not represent 100 otherwise it would be written as “1.00h”.

Incorrect datatypes were converted to their proper types and the cleaned data frame was saved to a csv file called Storm_Events

Gas Prices

I loaded in the data and subsetted the data so that it included only 'MGACD' under the MSN column. This represents motor gas usage. Since the years were stored in multiple columns, I melted it so that it was stored in one column. The data was subsetted again to include only years past 1993. The numbers in the Gas column represented $ per million btu (british thermal unit). I converted it into $ per gallon.

I deleted irrelevant columns, fixed column labels to match other datasets, and saved the cleaned data frame to a csv file called Gas_Price.

Crimes

I loaded in the the data and included only non null values of state_abbr. This removed the nationwide aggregated data. I fixed column labels so that they match other datasets of the same information.

I compared the population value for 1995 in this data with the Population data. They don't match. I deleted this population column and will use the population values in the other data frame since the population came from a more reliable source. I removed the rape_revised column because there is already another column containing counts of rape. The cleaned data was saved to a csv file called Crimes.

Standard Precipitation Index

I loaded in the data and subsetted the data to include only years past 1993. I removed the 'fips' column before melting the data. Months were stored in multiple columns instead of one. The month column included months in their three letter abbreviation. I made a function to change them into their respective number representations.

I grouped by state, year and month and aggregated using the mean for SPI values. I rename columns to match other datasets and then save the cleaned data frame to a csv file called SPI.

Health

I loaded in the XPT files using pandas read_sas. The data was immediately subsetted to include only the variables: _STATE, PHYSHLTH, MENTHLTH, and Year. I removed certain values of _STATE so that values such as "virgin islands" was not included as a state. I created three lists to help load in the data and clean it. Years was created to match the data from 1994 through 2016. 'states' was created with the states in alphabetical order to help create a State column to extract the correct states for each observation. The 'numbers' column was created to help determine what state each observation was. The _STATE variable contained numerical values which represented a certain state. The numbers used

skipped numbers frequently. The numbers in the numbers list I created are listed with their state representation in alphabetical order.

The data was further subsetted so that it includes MENTHLTH and PHYSHLTH values 30 and below. This excludes responses such as “refuse to answer” and “don’t know”. I aggregated by mean and grouped by Year and State. Each consecutive XPT file was stripped down to the point of the aggregated data frame before merged into one. The resulting data frame is saved to a csv file called Health.

Part 2

I loaded in Weather, Income, Population, Population2, Storm_Events, Gas_Prices, Crimes, SPI, and Health.

Weather

I created two separate data frames. One was the mean aggregate grouped by Year and State. The other one was the Month data. They are saved to data frames called Weather_Year and Weather_Month.

Storm Events

I renamed the columns to match the other datasets. I converted the EVENT_TYPE column to several dummy variables. I combined this result with the original data frame and deleted the EVENT_TYPE column. I made two separate data frames as I did with weather. One contains the year level data and the other one contains the month aggregated data by state.

SPI

I renamed the state column to match other datasets. I saved the original dataset as the month aggregated data and made the year aggregated data as a separate data frame.

Year Combined

I used a left join to combine on ‘State’ and ‘Year’ for Weather_Year, Storm_Year, and Health.

I created a dictionary of the states with their abbreviation as the values. I converted this into a data frame and merged the previous data frame with this one on ‘State’ using a left join so that only the states in the original data frame are included.

I merged the remaining datasets on ‘State_Abbrev’ and ‘Year’. I saved the resulting data to a file called Year_df.

Month Combined

I repeated similar steps with the Month aggregated data. This only included Weather_Month, Storm_Month, and SPI_Month. The resulting csv file is called Month_df

Visualization Application Description

Crime Type Checkbox

What crimes to include in the scatter plots and overall time series. Depending on what is selected, you can find relationships between all crimes, a subset of crime types, or a single crime type.

Select Variables

There is a linear or log option and a average or sum option. The dropdown menu selects what variable to plot a time series and a scatter plot of that variable versus the total number of crimes committed from selected crimes in the checkbox. The sum option will compute the sum of that selected variable from every selected state. The average option will compute the average of that selected variable depending on what states are selected. The linear/log option changes the scale of the scatter plot so that it is easier visualized when points are clustered.

Choropleth map

Hovering over each state will list the total number of crimes committed depending on what is selected for crime type. It shows how many of each crime contributes to the overall number of crimes for that state. It is heat mapped as well to easily visualize which states have a higher number of crimes committed.

Variable Time Series

Shows the distribution of selected variables over time averaging or summing each state at each year depending on selected states. There are two time series so that you can compare one variable to another with ease.

Scatter Plots

Shows the scatter plot of the variable versus the total number of crimes committed based on selected crime types. You can use this to visualize each variable against violent crime, rape, etc individually. You can also use this to visualize each variable against a subset of crimes or all of them. Two scatter plots are included to to able to visualize two variables at once against total number of crimes committed. You can hover over the points to determine what year and state that point is.

Select States

This is a multi-select dropdown menu. Depending on what is selected affects the appearance and results of all visualizations and the table at the bottom.

Overall Time Series

Plots the total number of crimes from selected crimes over time. There is a slider to select the range of desired years. There is also an 'Input Year to draw line' option to draw a vertical line. This may be helpful in case there are certain years with important events to see before/after effects of the event. For an example, Hurricane Katrina was in 2005. We can draw a vertical line at the year 2005 to indicate this event and easily visualize trends before 2005 and after 2005.

Table

We can filter additional variables and change what data is included in the time series and scatter plots. For an example, filtering median income > 40000 and then visualizing the average number of days in a month of poor mental health versus crimes committed or just the time series of mental health over time.

Initial Findings

All States/All Crimes

Choropleth map

California, Texas, and Florida have the highest number if total crimes committed in that order from 1995 to 2016.

Variable Time Series

Gas prices have increased and decreased from 1994 to 2015, but overall it has increased.

The average number of days of poor mental health in a month has increased from 1994 to 2016. In 1994, a person had about 11 days of poor mental health in any given month on average. In 2016, this number increased to about 13.

The same has happened to the average number of days of poor physical health in a month from 1994 to 2016. In 1994, a person had about 10 days of poor phydsical health in any given month on average. In 2016, this number increased to about 13.

The median income of a household in the USA has increased from 1984 to 2016.

The average temperature of the USA has fluctuated from 1994 to 2016. The minimum and highest temperatures see the same peaks. The SPI seems to have an inverse relationship with all three temperature variables. When SPI decreases, average temperature increases and vice versa. This is not always the case.

The population of each state on average is on a very steady linear increase from 1994 to 2016. In 1994, the population of each state on average was 5.25 million. In 2016, the population of each state on average was 6.45 million.

Overall Time series

The overall number of crimes for the USA is on a steady decrease by sum and by average, despite that the population is on a steady increase.

Scatter Plot

Population is positively correlated with the total number of crimes committed despite that the total number of crimes has been on a steady decrease and population has been on a steady increase over the years. This is likely due to Simpson's paradox and the way the data is structured. Since the data is actually comprised of the 50 states, it is likely just population differences between the states forming this contradictory relationship. Using the table, I filtered population > 30000000 and found that all these belonged to California. Overall, the total number of crimes decrease as the population decreases.

There are no clear trends for other plots.