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.

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 dataframe 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 dataframe 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 dataframe 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 dataframe into a regular dataframe 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 dataframe 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 dataframe 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 dataframe 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 dataframe 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 dataframe 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 dataframe 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 dataframe 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 dataframe before merged into one. The resulting dataframe 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 dataframes. One was the mean aggregate grouped by Year and State. The other one was the Month data. They are saved to dataframes 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 dataframe and deleted the EVENT_TYPE

column. I made two separate dataframes 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 dataframe.

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 dataframe and merged the previous dataframe with this one on ‘State’ using a left join so that only the states in the original dataframe 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