On Whether Weather Withers Violence

Introduction to Data Science, Final Report

Designed by: Richard Bravo, Alan Jeanpierre, Evan Adams, Lexi Vessels, Eduardo Mata

29th November, 2017

Abstract

Weather has no effect on crime in the city of Philadelphia. Crime rates in Philadelphia are not influenced by temperature, humidity, or pressure. While there is an increase in crime during the summer, there isn't a correlation between the time of day and crime rates in Philadelphia.

Table of Contents

Introduction

Weather has a profound effect on our daily lives. Anecdotal evidence suggests that weather conditions can influence one's mood. One may reflect in their emotions the overcast sky outside, or feel extra chipper on the first day of Spring. Our goal is to see if there is any evidence that particular weather conditions can affect the incidence of violent crimes. Specifically, we want to look at whether miserable weather might see a drop in violent crime. Miserable weather will be further defined as we explore the data, ranging from hot and humid to cold and wet. While data is readily available for weather patterns, crime is more difficult to find. Each policing district has both their own method of data disbursal, encoding, and offenses that are actually crimes. While there are central repositories that catalogue aggregate crime data, there is no such repository for the resolution we are seeking, which is the time of the crime down to the minute. To facilitate the exploration, we decided to select a set of cities with similar factors that influence violent crime, namely poverty, unemployment, and population. To accomplish this, we examined a number of datasets for county level data (as city was unavailable) to select counties with similar factors, and from there we chose to use the county seat as the city to examine.

We chose Philadelphia PA, Albany GA, Memphis TN, Toledo OH, Pine Bluff AR, Detroit MI, Baltimore MD, Flint MI, and St. Louis MO. Of those, there was a lack of crime data for Albany and Pine Bluff. As our data is gathered from a non-governmental agency, the reports are biased towards recent years due to the availability of information. As such, we will only examine each city's crime and weather data for recent years, even if there is historical data available.

Datasets

NOAA ISD

The National Oceanic and Atmospheric Administration publishes their Integrated Surface Database (ISD) on their website. It contains comprehensive weather data for participating stations across the world. Included with the actual data is the confidence level of the measurement to help identify reliable data. The database contains yearly data for each participating station, starting from 1901 to today. Included is a history file, which lists all participating stations, their names and locations, and timeframe in which they supplied data. Weather measurement stations are commonly components of airports or military bases, although there are some dedicated weather stations in more remote areas.

We identified the stations of interest using a combination of their USAF ID and WBAN ID, because one of them might not be supplied and be filled with a null-value. Specifically, after selecting the cities we wished to study, we found the closest station with current data using longitude and latitude using Pythagoras' Theorem.

Each station publishes data that is comprised of three sections: metadata about the station and measurement itself, required data for standard measurements, and optional data. The metadata are things like the name and location of the station, along with the time of the weather observation. The required data are standard measurements like temperature and wind-speed. Optional data are anything under the sun, including abnormal weather and particularly precipitation. The data is in fixed-width format, but because of the optional data section, only the first two sections are actually fixed width. The optional section is fixed-width, but each measurement category is prefixed with a unique identifier, and that subsection is addressable by index, instead of the entire line being indexable, as with properly fixed-width data..

The data we used from the dataset are observation date, observation time, temperature, pressure, sky condition, and mean relative humidity. Each entry of data was supplied roughly every hour, with multiple hour gaps between recordings of the optional data. The temperature was measured in degrees Celsius, scaled by a factor of 10, which we converted to Fahrenheit. Pressure was measured in hectopascals, also scaled by a factor of 10. It's curious that they didn't use decapascals and no scaling. The humidity was accompanied by a period measurement over which the humidity was measured. The sky condition is a measurement in Oktas, which are eighths of the sky which are covered by clouds. This was converted from a scale of 0 to 8 to a decimal percentage value. Because oftentimes the optional data was supplied at a different measurement time than the mandatory data, we backfilled all the values to account for superficial missing data.

SpotCrime

Due to every city publishing its crime differently, it was difficult to find accurate crime data for any arbitrary city. Some cities have an open-data policy, in which they publish a large amount of their data publicly. New York City, for example, has a very comprehensive set of data for many things in the city, crime included. Smaller cities though may only have the current county jail inmates, or even just the last few days worth of new county jail inmates. Other times, the arrest records are behind a paywall. We utilized the website SpotCrime, which has the most comprehensive dataset for crimes that we could find. Their data comes from scraping those temporal arrest notices on jail websites, comprehensive crime databases, news sources, and verified user submissions.

Each city has an associated daily crime ticker that lists, for each crime committed that day, the type of crime, the date (and sometimes time), the address, and a link to a more detailed (if available) description. The crime type is not as comprehensive as regular police records. While a police record may have a crime vary from “assault” to “aggravated assault with a deadly weapon” to “domestic violence”, SpotCrime only uses the keyword “Assault”. Specifically, the crime types are arrest, arson, assault, burglary, robbery, shooting, theft, vandalism, and other. In the FBI’s UCR, the FBI provides a definition of “violent crime”: one of murder, rape, assault, or robbery. We mapped those terms to SpotCrime’s and selected as our violent crime types: arson, assault, robbery, and shooting. It is not clear if murder is contained in assault or shooting, or if rape is included in assault. It is also not clear whether arrest is a catchall for any crime not accurately reported or if it specifically means a police officer arrested someone who had a warrant put out on them or any other minor offense, such as public drunkenness or possession of a controlled substance. Despite the flaws of this dataset, it is comprehensive and uniform and for most cities is superior to any other dataset.

We used BeautifulSoup to iterate through each day on the daily crime blotter, download the table of crimes, and load it into a dataframe. Because the script took an exceedingly long time to run due to timeouts and request overhead, we saved each dataset to disk as well. Because the date was provided as a string read from a website, we also had to convert that to a proper datetime format in Python. One flaw in SpotCrime’s dataset was that many crimes only have the day associated with it, and not the actual time. During the conversion to a date, if the time isn’t specified, the time defaults to 00:00, midnight. For portions of our analysis that dealt with the precise time of the offense, we culled those times that were not specific. We did not focus as much on cities which lost a significant amount of data from that culling operation.

County-Level Employment and Income

County-level unemployment rates are sourced from the Bureau of Labor Statistics Local Area Unemployment Statistics. Income rates are from the U.S. Census Bureau’s Small Area Income and Poverty Estimate. We averaged the data values for the unemployment rates between the years 2007 and 2015, and used the median household income for each county in 2015.

Because the dataset was already prepared by the source, there was no preprocessing to do.

County-Level Population Estimates

The county-level population estimates were sourced from the 1990 and 2000 Censuses of Population, and from the Census Bureau's Population Estimates Program for the subsequent years between censuses. We chose to average the population estimates of recent years, specifically 2010-2016. We also used averaged the change in population of those years as well. Because the dataset was already prepared by the source, there was no other preprocessing to do.

County-Level Poverty Estimates

This aggregate data is sourced from the U.S. Census Bureau's Small Area Income and Poverty Estimate. The data is listed by FIPS code, which is a code that identifies each county in the United States, and has columns estimating the poverty for each year sectioned by age. We used the section for all ages for the year 2015 for our analysis. Because the dataset was already prepared by the source, there was no preprocessing to do aside from renaming the columns.

County-Level Educational Attainments

Educational attainments are divided into four categories: those with no high school diploma, those who only have a high school diploma, those who completed some college, but no diploma, and those with a college diploma. It is not clear if "some college" includes a two-year degree such as an associate's degree, or if it denotes no degrees earned whatsoever. In 1970 and 1980, the category for "no high school diploma" includes those who didn't complete the 12th grade, which implies that adults with a GED are considered a part of that category. For the decades up until 2000, the data is sourced from the Censuses of Population. The data from 2011-2015 is a five-year average sourced from the Census Bureau's American Community Survey. We chose to use the years 2011-2015 to better represent the timeframe of data the other sets are dealing with.

This dataset was chosen to be the progenitor for the final dataset. It had a column called "Area Name" which was the name of the county, and a FIPS code that needed no preprocessing. We renamed the FIPS and area column to FIPS and County respectively, and also renamed all the columns for educational attainments to shorter version. We then set the index to the FIPS code. There was no other preprocessing needed to be done.

Uniform Crime Reporting Data

The FBI annually collects data from local police agencies in their Uniform Crime Reporting Program. It is effectively a voluntary program that each agency submits their records to. It is accompanied by a coverage indicator that represents the amount of individual agencies who completely report their data. The data provided is for the year 2012, and includes separate datasets for arrests and reports of offenses. The offenses are categorised into part I offenses, murder, rape, robbery, aggravated assault, burglary, larceny, auto theft, and arson; and part II offenses, forgery, fraud, embezzlement, vandalism, weapons violations, sex offenses, drug and alcohol abuse violations, gambling, vagrancy, curfew violations, and runaways. When talking about the division of violent versus nonviolent crimes, rape is included in the violent section. However, according to the UCR, rape does not include such cases as statutory rape (classified in the UCR as “sex offenses”) or sexual assault. We chose to accept those further definitions and did not attempt to sort sex offenses cases as violent or non-violent and chose only to include the category rape, as defined by the UCR.

The UCR offers datasets for both physical arrests as well as crime reports. Because we are trying to map temporal data of crime and temperature, we felt that the offset between when the crime was committed and when an officer would actually arrest them would obscure the relationship too much. As such, we chose to use the less accurate report data rather than arrest data. From this dataset, we only extracted violent crimes per county, specifically murder, rape, robbery, and assault.

The crime was grouped among multiple columns by the FIPS code, which denotes the specific county. The FIPS code is split in two sections, the state portion and the county portion. We merged those two numbers into a single FIPS column that represented the full code for that county. We created a column for violent crime that consisted of the sum of the columns for the number of rapes, murders, robberies, and assaults for that county. We set the index of the dataframe to the FIPS code in anticipation of the merging of the datasets.

Data Preprocessing

City Selection

We built a dataframe with which to select cities to further examine. We merged the UCR violent crime reports, educational attainments, poverty estimates, unemployment estimates, and income censuses, associated all with their county. The data was grouped by the 3-quantile of each factor, which categorized each value as Low, Medium, or High. We chose four factors to select out final cities, violent crime rates, percentage of the population below the poverty line, percentage of the population who are unemployed, and the average population over five years. We selected a random sample of cities from a group such that the violent crime rate per 100,000 people was greater than 800, and they had a relatively high level of violent crime, impoverished population, unemployment rate, and population. We chose a high population so that particularly bad days wouldn't skew the dataset and also to increase the likelihood of effective crime reporting (in regards to datasets). We chose poverty and unemployment as we believed that they are the greatest factors towards a tendency to commit a violent crime. Other factors such as income or education are correlated with these factors as well. After selecting the counties we wished to study, we looked up the county seats and made note of their name and geographic location (longitude and latitude).

The NOAA ISD dataset contains a record of every station that contributes to the dataset, including their station IDs, locations, and measurement timeframe. For each city we selected, we searched through that list of stations to find the nearest station that had an adequate date range of measurements. Making note of those station IDs, we could begin to create a dataset for each city, including the weather reports and crime reports.

	p_no_HS_dip	p_HS_dip	p_some_college	p_college_dip	avgpop	p_impoverished	p_unempl	med_income	COVIND	p_dpop	vcrime_rate
count	3142.00	3142.00	3142.00	3142.00	3.14e+03	3141.00	3144.00	3141.00	3178.00	3.14e+03	3134.00
mean	14.57	34.75	30.26	20.41	1.01e+05	16.27	7.01	48600.60	98.18	6.56e-04	235.21
std	6.64	7.07	5.17	9.02	3.22e+05	6.47	2.31	12355.27	8.26	8.00e-03	200.79
min	1.60	7.50	11.40	1.90	8.90e+01	3.40	2.00	22894.00	0.00	-3.42e-02	0.00
25%	9.50	30.30	26.70	14.20	1.10e+04	11.50	5.41	40426.00	100.00	-4.17e-03	94.50
50%	13.10	35.10	30.30	18.20	2.58e+04	15.20	6.92	46800.00	100.00	-7.35e-04	185.84
75%	18.70	39.60	33.80	24.20	6.75e+04	19.70	8.37	54153.00	100.00	4.47e-03	321.75
max	53.70	54.80	47.80	78.80	1.00e+07	47.40	24.97	125900.00	100.00	9.35e-02	1800.32

Table 1. Descriptive statistics for the county-level dataset.

	p_no_HS_dip	p_HS_dip	p_some_college	p_college_dip	avgpop	p_impoverished	p_unempl	med_income	COVIND	p_dpop	vcrime_rate
p_no_HS_dip	1.00	0.21	-0.52	-0.60	-0.05	0.68	0.45	-0.56	-0.03	-0.22	0.19
p_HS_dip	0.21	1.00	-0.31	-0.76	-0.31	0.20	0.23	-0.47	-0.08	-0.46	-0.13
p_some_college	-0.52	-0.31	1.00	0.06	-0.07	-0.35	-0.24	0.18	0.02	0.09	-0.09
p_college_dip	-0.60	-0.76	0.06	1.00	0.32	-0.46	-0.38	0.68	0.08	0.47	0.01
avgpop	-0.05	-0.31	-0.07	0.32	1.00	-0.07	0.01	0.24	0.05	0.23	0.22
p_impoverished	0.68	0.20	-0.35	-0.46	-0.07	1.00	0.58	-0.78	-0.09	-0.31	0.28
p_unempl	0.45	0.23	-0.24	-0.38	0.01	0.58	1.00	-0.48	-0.05	-0.23	0.24
med_income	-0.56	-0.47	0.18	0.68	0.24	-0.78	-0.48	1.00	0.10	0.49	-0.15
COVIND	-0.03	-0.08	0.02	0.08	0.05	-0.09	-0.05	0.10	1.00	0.10	0.05
p_dpop	-0.22	-0.46	0.09	0.47	0.23	-0.31	-0.23	0.49	0.10	1.00	0.05
vcrime_rate	0.19	-0.13	-0.09	0.01	0.22	0.28	0.24	-0.15	0.05	0.05	1.00

Table 2. Correlations between factors for the county-level dataset

```
Out[137]: <matplotlib.axes._subplots.AxesSubplot at 0x7fe8c1bacda0>
```

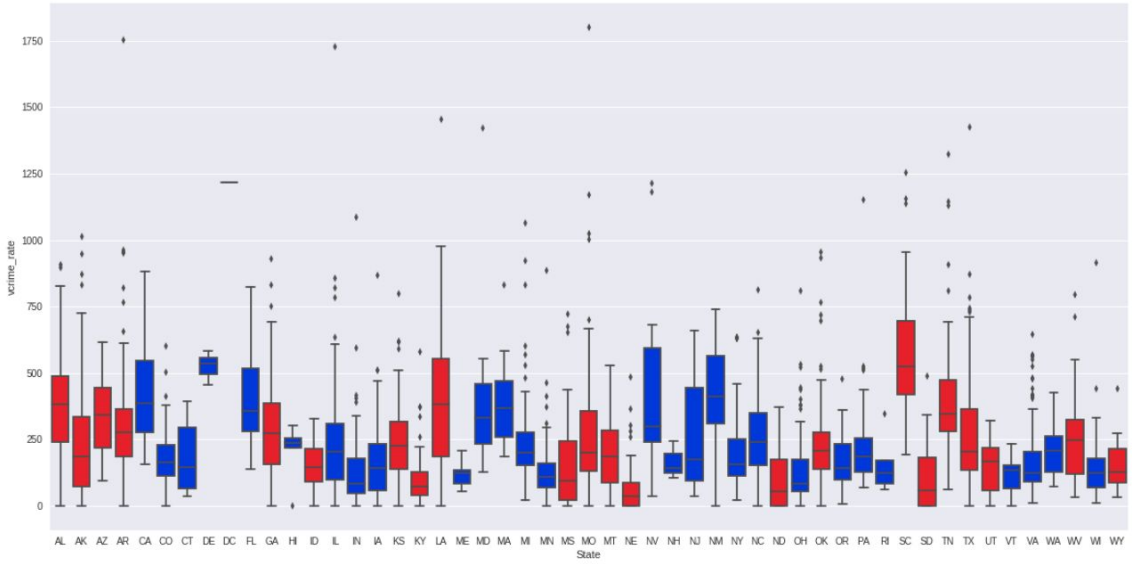

Figure 1. Box plot for each state's violent crime rate, coloured according to their political leaning in the 2008 presidential election.

Figure 2. Scatter plot of each county's percentage of residents unemployment rates vs the violent crime rate. There appears to be a very slight positive correlation.

Figure 3. Scatter plot of each county's percentage of residents poverty rates vs the violent crime rate. There appears to be a very slight positive correlation.

Per-City Analysis

NOAA ISD stations publish their measurements extremely regularly, in contrast to the crime data which occurs randomly. We chose to group the dates by hour in order to have meaningful groups. We also limited our date range to between 2014 and 2017, the range of which captures a saturated amount of submissions to SpotCrime. We split each cities crime data into two tables: total crime and violent crime. We only merged violent crime with the weather set, as opposed to all crime types. Each city's set of dataframes was accessible

through a Python class, which allowed all the dataframes to have concise names. It also allowed elegant initialization and loading of data into the dataframes.

When we loaded the crime dataset, some values were lacking specific times of day, so when we converted the raw string to a datetime type, we only allowed a specific format that included this time of day. This not only made the conversion 100 times faster, but also solved the issue of times defaulting to midnight masking genuine midnight crimes. The downside is that some of our data was culled, with some cities experiencing more data loss than others. Specifically, Memphis and Philadelphia lost almost no data, St. Louis, Baltimore lost some data, and Detroit, Flint, and Toledo lost a great deal of data.

Out[6]:

	Baltimore, MD	Detroit, MI	Flint, MI	Memphis, TN	Philadelphia, PA	St. Louis, MO	Toledo, OH
Arrest	0.994262	0.369246	0.037559	0.995202	0.991354	0.755795	0.525197
Arson	0.515254	0.454158	0.025937	0.994919	0.980000	0.999352	0.682870
Assault	0.637649	0.258590	0.073566	0.997853	0.997054	0.773747	0.472545
Burglary	0.428944	0.264795	0.061728	0.996855	0.996788	0.735583	0.343376
Other	0.999275	0.267180	0.225458	0.998993	0.999803	0.913931	0.989402
Robbery	0.503708	0.570339	0.596610	0.995363	0.995316	0.786775	0.361761
Shooting	0.849033	0.944949	0.383013	0.835112	0.911980	0.749254	0.846970
Theft	0.398938	0.278902	0.004967	0.997238	0.998257	0.832724	0.580260
Vandalism	0.995617	0.195425	NaN	0.996840	0.999750	0.573438	0.765670

Table 3. Table of data record counts as a percentage of processed/total crime. Memphis and Philadelphia retained nearly all of their records, whereas Flint lost nearly all of it.

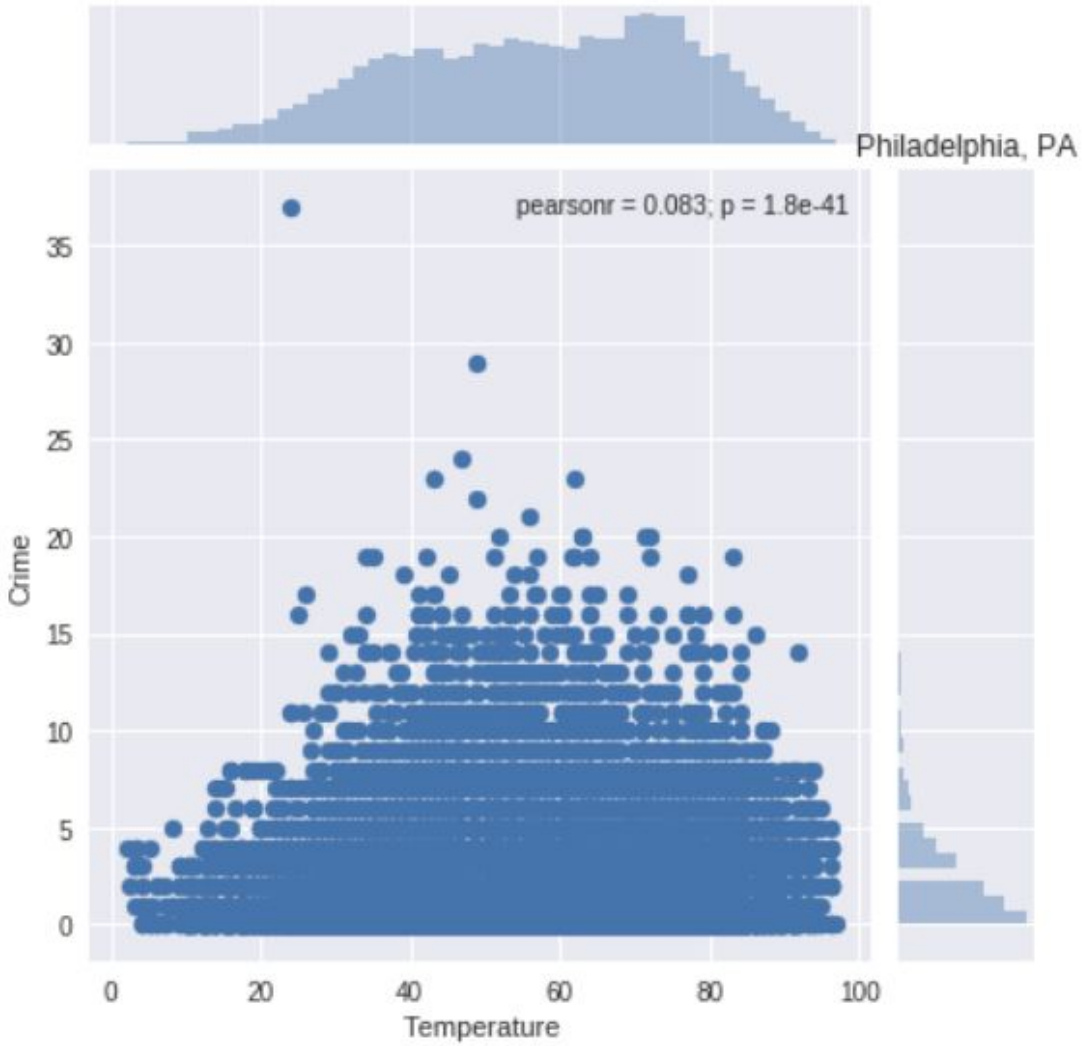

Figure 4. Scatter plot of the number of crimes in a given hour and the temperature in that hour. There is a roughly normal distribution, which suggests that crime is evenly distributed and the normal distribution represents the temperature range of Philadelphia.

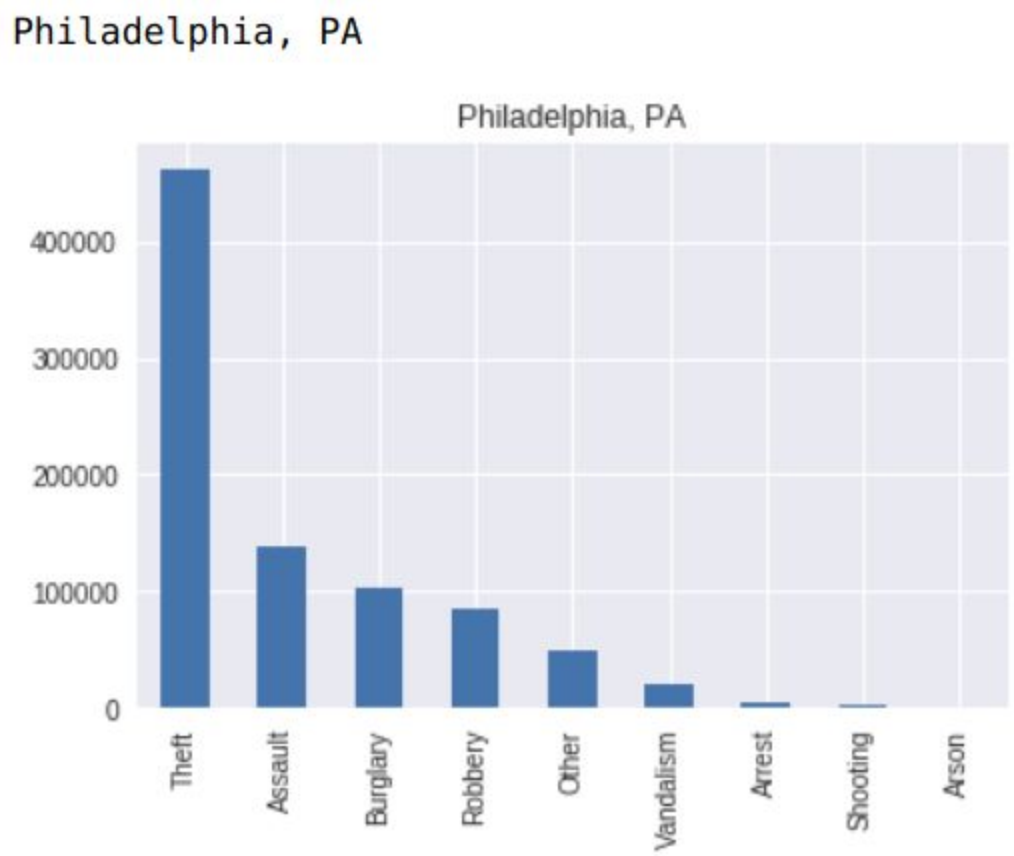

Figure 5. Distribution of the types of crime in Philadelphia.

Programming

We attempted to fully automate the majority of the data acquisition and analysis. While most of the datasets are cached in the data folder, they are available to download through Python from their sources. The exception of which is the discovery of the county seats, and therefore their longitude and latitude location used in determining their nearest weather station.

Of note is the section which crawls a large quantity of webpages from SpotCrime. Although they have an API, it does not appear to be public so we chose to webscrape using BeautifulSoup. Web scraping turned out to be a tedious task because of the sheer time component associated with it. The overhead associated with multiple separate HTTP requests along with the server timing out because of too many successive requests resulted in the script taking multiple days to run. It had to accommodate timeouts and methods to circumvent them while still being gentle on the server using a try/except statement paired with a ten second delay.

While Jupyter notebooks are a valuable tool, they proved to be one of the most challenging programming aspects. Because they are formatted documents and not raw text, working with teammates on a Git repository proved difficult. We settled on a system where the

one who pushed their changes must clear all the output data so that the code is at a bare minimum of difference to others who want to pull the repository. On top of that, while Github renders the notebooks very nicely on their website, it doesn't handle file differences with commits at all. As such, when two people were going to commit at the same time, it was impossible to merge them and one person just had to stash their changes somehow and manually merge them in Jupyter. In the future, Git branches may have proved helpful, or working in separate, personal Jupyter notebooks until one is ready to merge.

Results

We had five hypotheses we wanted to test. We specifically used Philadelphia because that city had the highest quality data. The hypotheses are:

1. Temperature is positively correlated with violent crime rates
2. Humidity and pressure have no impact on violent crime
3. Violent crime rates are higher in the summer compared to the winter
4. Violent crime is higher during the midnight hours between 11:00pm and 2:00am
5. There is a particularly large spike at 2:00am when bars traditionally close for the night.

When we examined the correlation between weather and crime in Philadelphia, we found no significant correlations. There is a slight correlation with the highest values of crime, but overall there is no correlation between weather and crime in these cities, between temperature, pressure, or humidity. Perhaps in other cities there is such a correlation, and the code is modular enough to investigate other cities as needed.

There appears to be a higher incidence of crime during the summer than in other seasons. We defined summer to be months between June and September. This could be attributed to the fact that many criminals are teenages and young adults, and schools aren't in session during the summer.

```
Out[175]: <matplotlib.axes._subplots.AxesSubplot at 0x7fe8c39cfa90>
```

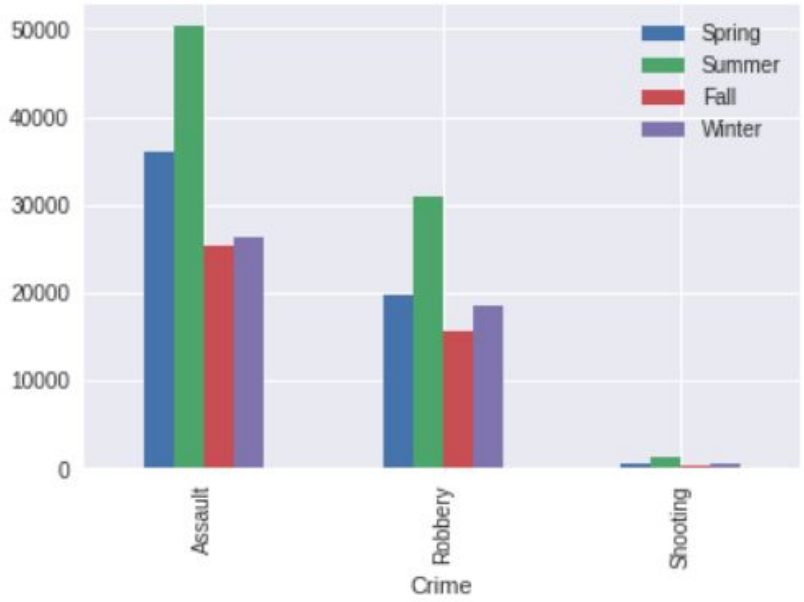

Figure 6. Bar chart of the number of crimes per season per crime.

There is no data to suggest that violent crime is particularly higher at night in Philadelphia. In Baltimore and St. Louis, spikes particularly during the late morning, around 10:00am to 12:00pm, which suggests that criminals act on the fact that many people are away at school or work. Also, there is no spike in crime at 2:00am, in fact crime drops drastically after that time. Perhaps because criminals are rowdy during operating hours and then all head home afterwards, instead of the assumption that they would become even more rowdy.

```
Out[176]: <matplotlib.text.Text at 0x7fe8c38ff550>
```

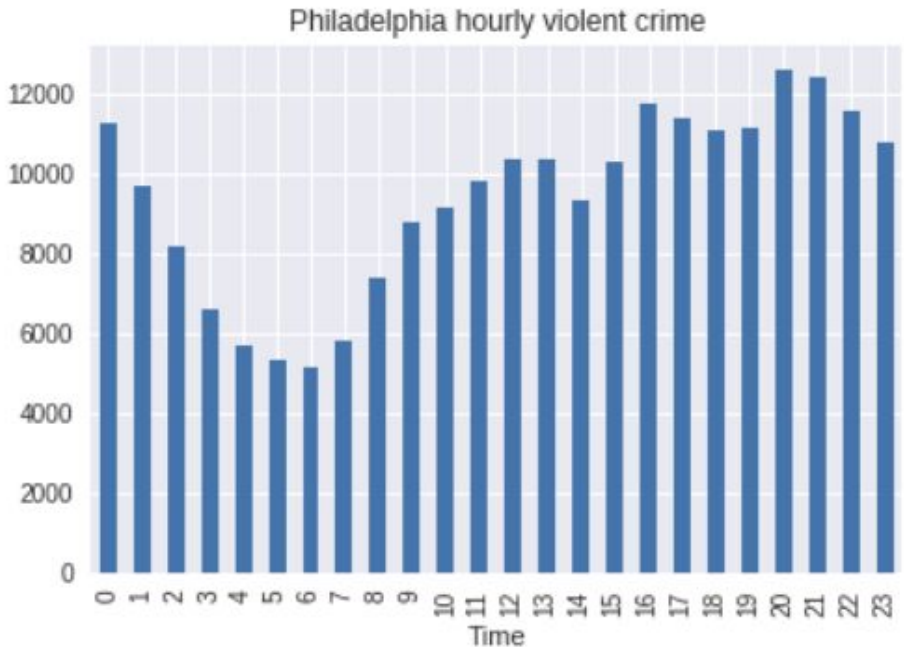

Figure 7. Distribution of violent crime during the hours of the day. It is fairly steady except for the hours around dawn when people sleep.

All in all, we were incorrect in every single hypothesis except that crime is higher in the summer. If we would continue this project, we would like to examine cities that are less extremely dangerous and more average and representative of the United States. The fact that we chose those cities may have hidden any relation between weather and violence simply because of the sheer number of crimes committed at all hours of the day.

References

Mood Oscillations and Coupling Between Mood and Weather in Patients with Rapid Cycling Bipolar Disorder
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2651091/

Temperature and Violent Crime in Dallas, Texas: Relationships and Implications of Climate Change
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3415828/

Unemployment and median household income for the U.S., States, and counties, 2007-16
https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

Population estimates for the U.S., States, and counties, 2010-16
https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

Poverty estimates for the U.S., States, and counties, 2015
https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

Educational attainment for the U.S., States, and counties, 1970-2015
https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

Uniform Crime Reporting Program Data: County-Level Detailed Arrest and Offense Data, 2012
http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/35019

Integrated Surface Database
ftp://ftp.ncdc.noaa.gov/pub/data/noaa