Studying the Relationships Between Economic Factors and Crime in Maine's Metropolitan Areas

Ryan C. Reed

University of Southern Maine

COS 422: Computing for Data Science

Professor Bruce MacLeod

December 14, 2022

Problem Domain

The problem domain for this project covers the topics of economy, crime, population statistics, and analysis of statistical relationships from different domains. The primary challenge was identifying key features from reliable recorded economic and crime data. Another challenge that related to the domain included utilizing exploratory data analysis, primarily visualization, to determine whether relationships between data features from the two different statistical fields existed. The scope of the project was kept to Maine's metropolitan statistical areas, to provide a key baseline for granularity between the collected datasets.

My reasoning for choosing my project over this domain is easily explainable by relevance, and importance. Data science as a field is built upon statistics, discovering relationships, and reporting unbiased results, and is trademarked by analysis for the key purpose of providing actionable insights. The relationship, or lack thereof, of economy and crime, two important societal issues, provides an ample opportunity to practice the python, data manipulation, and analysis skills that we've spent the semester honing. The issues of economy and crime hold a significant importance in American life & politics, and in a polarized era of American politics, it's more important than ever that data science holds an astute role in analyzing these issues (Schaeffer & Green, 2022).

Problem Statement

The key goal of this project is to study potential relationships between select economic features & crime rates as they pertain to Maine's three primary metropolitan areas: the Lewiston Auburn, the Portland-South Portland-Biddeford, and the Bangor statistical areas. The economic features included within this project consist of two separate categories: quarterly & annual recordings. The economic features cover quarterly housing price index, and annual total GDP,

unemployment rate, per capita personal income, and resident population. The selected crime features included the total property crimes, and the total violent crimes, both annually recorded, with the specific goal of computing annual crime rates for the expressed goal of this project.

This is a problem that I defined and personally sought out sources for data, for the added purpose of gaining experience in setting my own problem domain, finding reputable sources for datasets, and adapting the data science foundational techniques that we've been focusing on throughout the semester to perform a unique analysis with minimal guidance. Ideally, I wanted to cover as many sections of our course as I could over this self-created domain, from the ground up, without the aid of other notebook sources as commonly featured on sites such as Kaggle.

Datasets and Inputs

This project ultimately spanned 14 crime datasets, 1 quarterly economic dataset, and 4 annual economic datasets. Dataset importation was handled through manually mounting my Google Drive, as API access to the crime sets, and the economic datasets, was not possible. Below, I go through the various datasets. Crime data was made available as legacy format (2006-2010), and modern excel files (2011-2019). While the economic datasets were compiled into TSV formats.

The primary dataset for this analysis project was for the combined annual economic data, spanning yearly records for resident population, per capita personal income, total gross domestic product, and unemployment rates, with yearly entries for each of the three metropolitan areas as seen in Figure 1. Altogether, there was a total of 53 recorded years, with 13 columns, although in reality the years differed based upon the specific feature.

DATE	BANG723PCPI	BANPOP	LAUMT237075000000003A	LAUMT237465000000003A	LAUMT237675000000003A	LEWI623PCPI	LWAPOP	NGMP12620	NGMP30340	NGMP38860	PORT723PCPI	PTLPOP
1969-01-01	2974			3417	3704							
1970-01-01	3282			3713	3985							
1971-01-01	3488			3773	4202							
1972-01-01	3822			3985	4562							
1973-01-01	4185			4451	4908							
1974-01-01	4614			4798	5270							
1975-01-01	5067			5298	5752							
1976-01-01	5666			5962	6397							
1977-01-01	6061			6338	6906							
1978-01-01	6624			6914	7628							
1979-01-01	7383			7535	8335							
1980-01-01	8237			8475	9385							

Figure 1: Annual Economic Dataset (United States Federal Reserve, 2022).

The quarterly dataset for the housing price indexes is shown in Figure 2, and it had a total of 162 recorded quarters, and 4 total columns. For clarity, housing price indexes within this dataset are defined as a broad measure of the changes in single-family property prices. Just as with the annual economic data, the start-points for the recorded years vary based on statistical area.

DATE	ATNHPIUS12620Q	ATNHPIUS30340Q	ATNHPIUS38860Q
1982-04-01		51.22	
1982-07-01		49.72	
1982-10-01		51.55	
1983-01-01		51.10	
1983-04-01		53.59	
1983-07-01		55.47	
1983-10-01		58.66	

Figure 2: Quarterly Housing Price Index Dataset (United States Federal Reserve, 2022).

The crime data consists of 14 crime datasets with annual totals of crimes by state, by city. The general format is shown within Figure 3. There was no unified collection to manipulate, and each file representing a year had to be manually downloaded and combined into one dataset indexed by year. Technically, each dataset has unique characteristics by way of changing crime definitions for the recorded year as well as having data that includes and excludes different towns over time, as towns merge together or form across Maine. The former appears to represent a major challenge, but as each year applies the federal definition for those crimes, the universal precedent is simply the current federal definition. The shapes of these datasets ranged from 110-117 records but kept a constant 12 columns. The key features of these datasets included the annual totals for violent and property crime.

Table 8												
Offenses Known to Law Enforcement												
by State by City, 2019												
State	City	Population	**Violent crime**	Murder and nonnegligent manslaughter	Other	Robbery	Aggravated assault	**Property crime**	Burglary	Larceny-theft	Motor vehicle theft	Arson[2]
ALABAMA[3]	Hoover	85,670	114	4	15	27	68	1,922	128	1,694	100	2
ALASKA	Anchorage	287,731	3,581	32	540	621	2,388	12,261	1,692	9,038	1,531	93
	Bethel	6,544	130	1	47	3	79	132	20	84	28	12
	Bristol Bay Borough	852	2	0	0	0	2	20	5	8	7	0

Figure 3: Annual crime data by state, by city, for the year of 2019

All datasets contained 'cleaned' data unless explicitly handled in the next section. The primary attribute to note was that for the range of 2006-2019, there were no missing values.

Data Cleaning and Wrangling

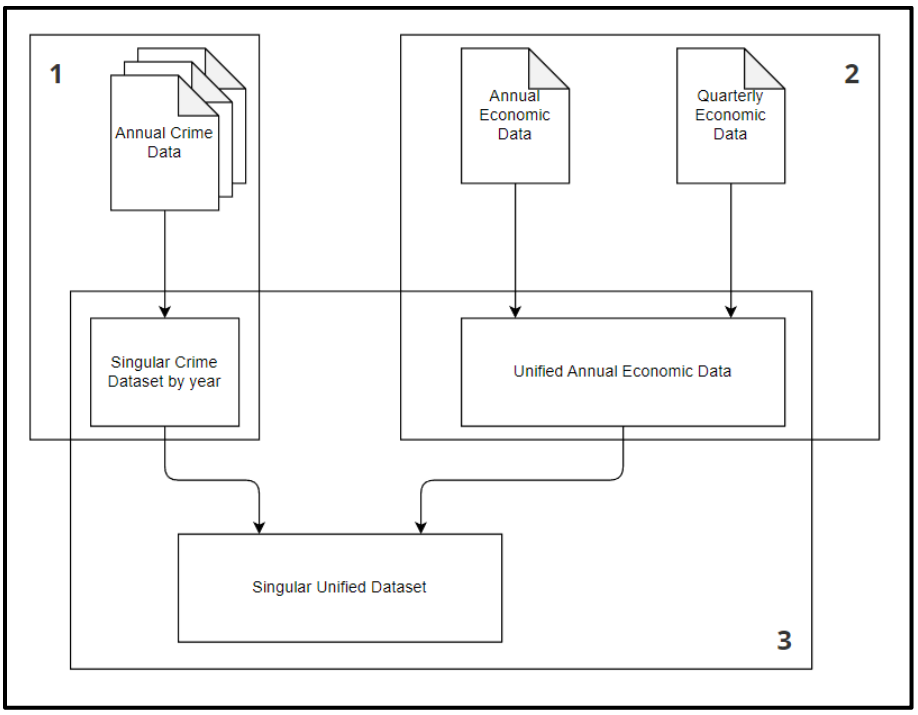


Figure 4: A breakdown of data cleaning & wrangling

Developed or Adapted Techniques

I find it easiest to define the process of data cleaning & wrangling in stages as diagrammed in Figure 4. In stage 1, we manipulate crime data through the following steps. We first drop all state data except for the state of Maine and then remove the state column, reducing the crime datasets to only Maine cities. By manually compiling lists of towns, and former towns, of each statistical area based on recency-precedence in the form of tsv files, the cities are summed, or dropped

completely, to form accurate statistical area records. For each dataset, the year of the data is applied as a column, and all 14 sets are combined into a single crime dataset with 6 columns, representing annual total property & violent crimes for each of the three areas indexed by year.

Stage 2 was much shorter due to the nature of the economic data, with the first step being the averaging of the quarterly housing price index into yearly values, and secondly, appending that data to the annual data, creating a unified economic set. In stage 3, the economic & crime sets are concatenated into a single dataset, and all feature columns are renamed appropriately, as seen in Figure 5.

	Bangor VC	Lewiston VC	Portland VC	Bangor PC	Lewiston PC	Portland PC	Bangor PI	Bangor pop	Bangor UR	Lewiston UR	...	Lewiston PI	Lewiston pop	Bangor GDP	Lewiston GDP	Portland GDP	Portland PI	Portland pop
DATE																		
2006	90.0	128.0	623.0	3208.0	2304.0	11852.0	30103	148.197	4.7	4.7	...	30836	106.971	5304.131	3524.866	22942.558	40039	512.073
2007	77.0	142.0	598.0	3349.0	2275.0	11092.0	31196	148.603	4.7	4.6	...	32369	106.695	5462.341	3749.462	23600.482	41405	513.791
2008	96.0	149.0	615.0	3458.0	2117.0	10768.0	32557	149.268	5.2	5.4	...	33633	107.061	5410.995	3877.941	23859.239	42823	516.026
2009	103.0	135.0	602.0	3764.0	2200.0	10580.0	33367	149.419	7.8	8.7	...	33672	106.539	5571.966	3786.188	24268.556	42177	516.826

Figure 5: The unified dataset containing all desired features from the collected data.

Cleaning was largely unnecessary, with no missing values, solely numerical features, and no need to convert data types except for utilizing datetime for the years and filtering them. Desired features engineered for analysis were computed using the following formulae:

$$Total\ Crime\ Rate_{x,y} = Property\ Crimes_{xy} + Violent\ Crimes_{xy}$$

$$Yearly\ Crime\ Rate\ per\ 1000\ People = \frac{Total\ Crime\ Rate_{x,y}}{Population_{x,y} * 1000}$$

$$Per\ Capita\ Property\ Crimes = \frac{Total\ Crime\ Rate_{x,y}}{Population_{x,y} * 1000}$$

$$Per\ Capita\ Violent\ Crimes = \frac{Total\ Crime\ Rate_{x,y}}{Population_{x,y} * 1000}$$

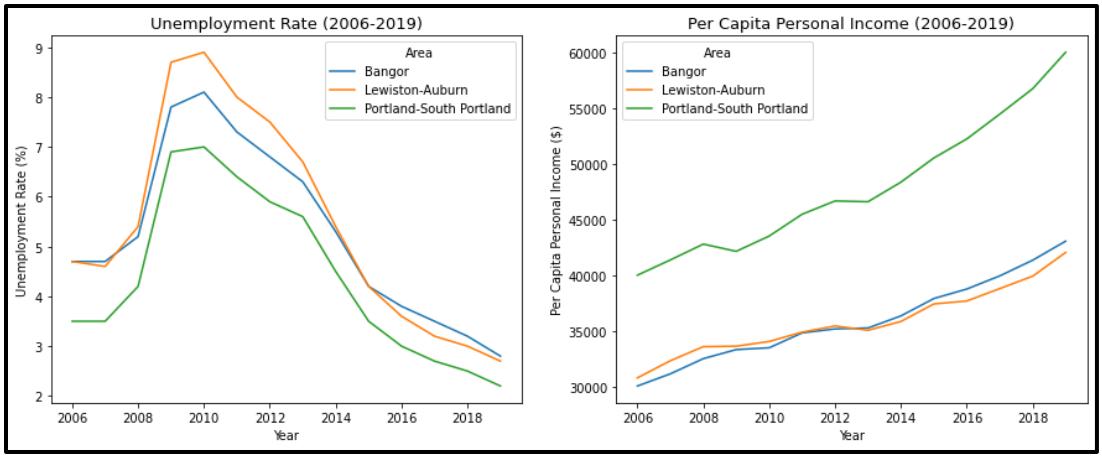
$$Per\ Capita\ GDP = \frac{Gross\ Domestic\ Product_{x,y}}{Population_{x,y} * 1000}$$

Where *x* and *y* indicate statistical area and recorded year.

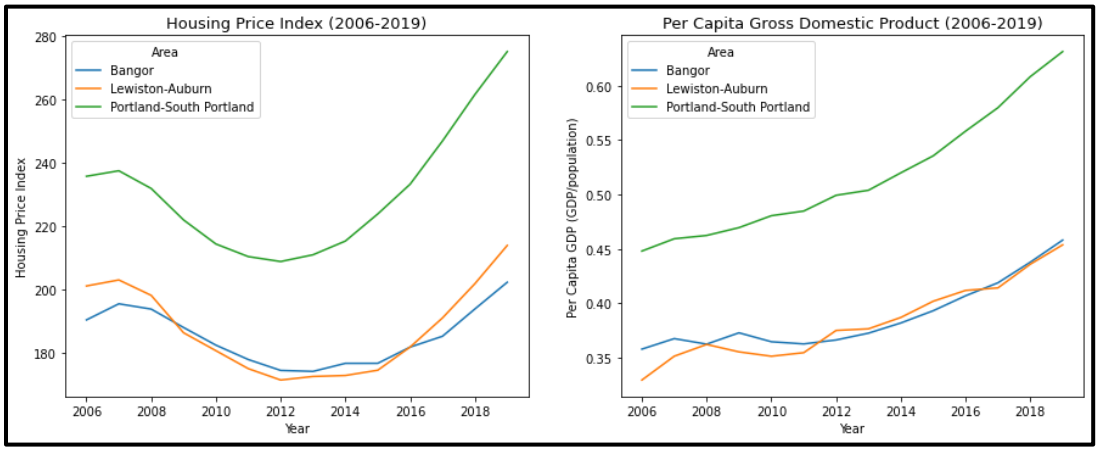
Copied Techniques

All of the techniques used in the data cleaning & wrangling were adapted from the foundations of our course textbook, and no other source was used for that portion of this project, but explicitly referenced in this project are the sources that I used for implementing & understanding the machine learning models of ridge regression, multiple linear regression, and random forest regression. That being said, I want to point explicitly to the Titanic Workflow notebook for the adapted code that I used for correlative heatmaps, as well as the general structure of sections in my Jupyter notebook.

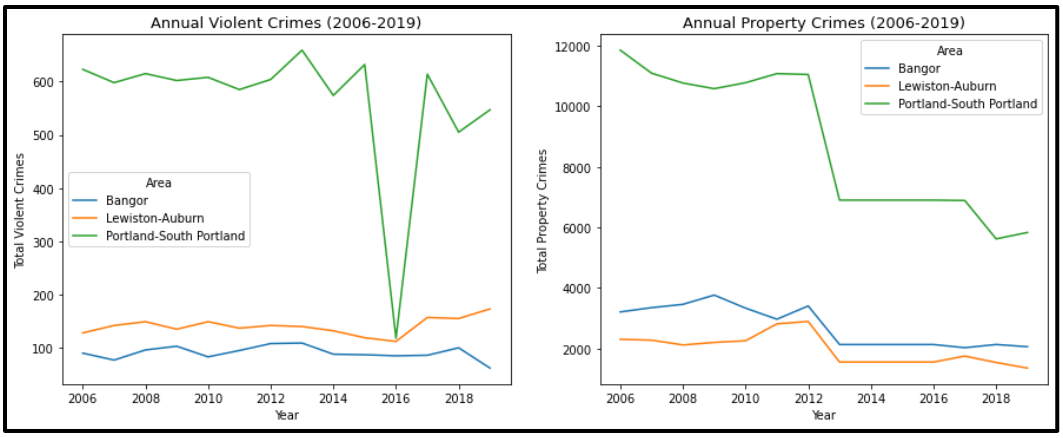
Results



Figures 6 & 7: Visualization of the annual changes of unemployment rates and annual per capita personal incomes respectively.



Figures 8 & 9: Visualization of the annual changes of housing price indexes and per capita gross domestic product respectively.



Figures 10 & 11: Visualization of the annual changes of total violent crimes and total property crimes respectively.

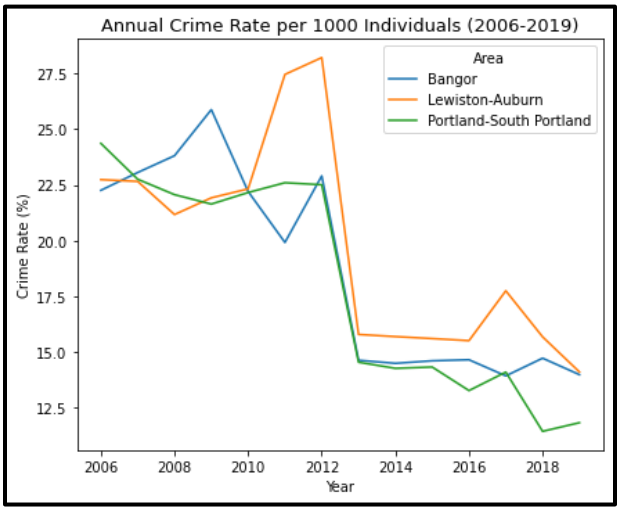
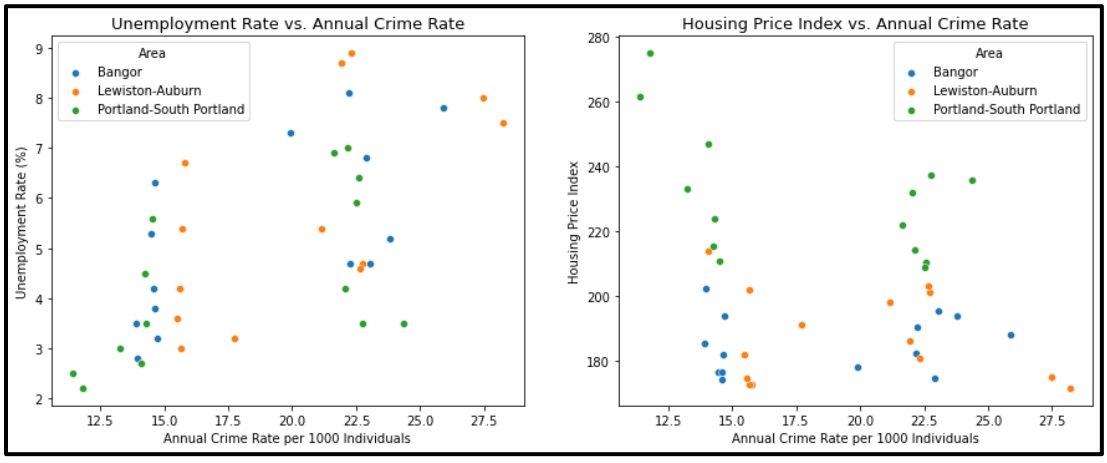
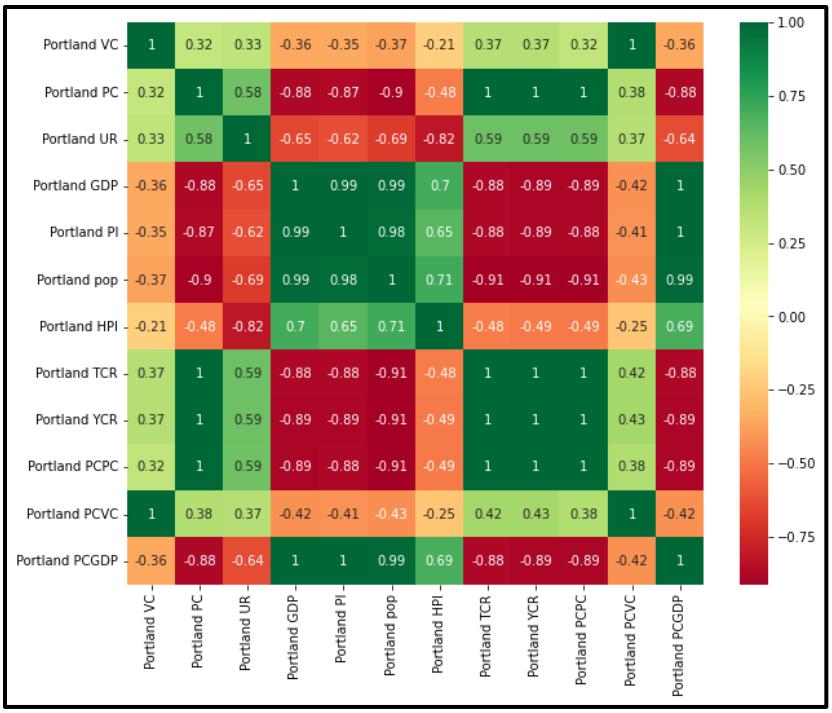


Figure 12: Visualization of the annual changes of the annual crime rate per 1000 persons.



Figures 13 & 14: Visualization of the relationships between unemployment rate, housing price index and annual crime.



Figures 15: Visualization of the correlation between the features for the Portland-South Portland-Biddeford area.

Figures 6-9 reveal annual visualizations of the economic data, and indicate positive economic growth, in contrast to the general decrease in crime shown in Figures 10-12. Building off of this observation, Figures 13 & 14, visualize the relationship between some economic factors and the annual crime rate, with each displaying a positive correlation, in the case of unemployment rate, or a negative correlation, as seen with housing price indexes. Figure 15 paints a clearer image of the correlation between features using the Portland-South Portland-Biddeford area as an example, and we see notable relationships between the observed economic factors, and crime rate. Namely, this analysis attributes a moderate positive correlation between annual crime rates and unemployment rates, moderate negative correlation between annual crime rates and housing price indexes, and strong negative correlations for population, gross domestic product, and per capita personal income with annual crime rates.

This evidence of relationships satisfies the key goal of this project, to analyze data to find evidence of relationships between economic factors & crime. As a secondary goal, we delve into

the evaluation of multiple ML predictive models using only economic data for the Portland-South Portland-Biddeford area, with the annual crime rate as the target.

Model	Mean Absolute Error (MAE)
Random Forest Regression	~1.4691
Ridge Regression	~1.6696
Multiple Linear Regression	~4.8186

Table 1: Evaluation results for various machine learning models.

As seen in Table 1, random forest regression performed the best with a mean absolute error of ~1.4691%. To give this result a bit of clarity, the annual crime rate for the statistical area around Portland typically ranged around 18-24%, and so this is a remarkable success rate. Ridge regression performed only slightly worse, and ultimately multiple linear regression performed the worst with a resulting MAE of around ~4.8186%.

Next Steps

One of the biggest challenges I faced with this project stemmed from stepping outside of the competition sites, specifically finding reliable and accurate data from reputable sources. The good, and bad, side of this major challenge is that federal entities keep records of this data, making them a very convenient and reliable source, at the expense of being limited to certain datasets. Another issue was finding datasets that had appropriate granularity with the data already collected. Altogether, I found it difficult to reliably seek out new sources for economic data, other than what the Federal Reserve made available, unfortunately, the United States Bureau of Economic Analysis carried the same datasets due to their relationship as economic national entities. My search for additional data typically ended unfortunately, only finding state or national level records, such as inflation or consumer price index. If I had more time, I would have searched more vigorously for metropolitan level data, specifically containing records for

the areas under study. Ultimately, I would have liked to include homeless rates and inflation metrics but was unable to find anything beneath state level data.

Due to the unfortunate circumstances regarding law enforcement agencies following the national directive to change from the *Universal Crime Reporting System* to the *National Incident Based Reporting System*, and the previously aforementioned challenges, I couldn't compile a larger time range than 2006 through 2019. Just like the issues I had with securing additional economic data, the only datasets from reputable sources I was able to uncover, were state level data, and would not have been nearly similar to the granularity that I intended for this project.

Initial Project Proposal

Disclaimer: I'm unsure of which portions of the project proposal that you want included, so I'll include the most important portions: description, criterion for success, and the plan of work.

Introduction & Description: The exact relationship between economic factors and crime rates have long been debated. Theoretically, as economic growth and opportunity increase, crime should generally decrease. This theoretical pattern should be easy to detect, based upon relevant economic & crime data collection. As a personal experience, during my employment as a correctional officer, I was taught repeatedly about the economic factors of crime, but in reality, the links of a poorer economy and higher crime rates, as well as vice versa, are quite understudied and largely based upon assumptions.

In order to facilitate a data science project studying the relationships between economy & crime, a focus should be on minimizing the impact of other factors of crime. Sociological factors that impact crime should therefore be considered, these include topics such as cultural, age, educational, and wealth demographics. While no U.S. state has perfect conditions for this particular kind of data project, I've identified Maine's demographics as being ideally somewhat balanced or heterogenous in terms of demographics. More ideally, the general idea of this project is to identify which economic factors might be correlated to crime rates.

Criterion for Success:

1. Identification of the correlation between economic factors and crime rates.

2. A prediction model with a significant rate of target prediction.

Plan of Work:
11/31 - Data Wrangling – **Completed on Schedule**
12/02 – Exploratory Data Analysis - **Completed on Schedule**
12/04 – Machine Learning Application - **Completed on Schedule**
12/07 – Class Presentation – **Completed on Schedule**

References

1. Reed, R. (2022). *Final Project Proposal – Assignment 4* [Unpublished Manuscript]. University of Southern Maine.

2. United States Federal Reserve. (n.d.). *Economic Data*. https://fred.stlouisfed.org/

3. Federal Bureau of Investigation. (n.d.). *UCR Publications*. https://www.fbi.gov/how-we-can-help-you/need-an-fbi-service-or-more-information/ucr/publications

4. MacLeod, B. (2018, September 13). *Titanic Survival Prediction: Workflow*. https://github.com/usm-cos422-522/courseMaterials/blob/main/Labs/titanic-workflow.ipynb

5. McKinney, W. (2022, October 19). *Python for Data Analysis: Data Wrangling with pandas, NumPy & Jupyter (3rd ed)*. 978-1-098-10403-0.

6. Brownlee, J. (2020, October 11). *How to Develop Ridge Regression Models in Python*. https://machinelearningmastery.com/ridge-regression-with-python/

7. IBM Cloud Education. (2020, December 7). *Random Forest*. https://www.ibm.com/cloud/learn/random-forest

8. Piepenbreier, N. (2022, July 1). *Linear Regression in Scikit-Learn (sklearn): An Introduction*. https://datagy.io/python-sklearn-linear-regression/

9. Schaeffer, K. V. Green, T. (2022, November 3). *Key facts about U.S. voter priorities ahead of the 2022 midterm elections.* https://www.pewresearch.org/fact-tank/2022/11/03/key-facts-about-u-s-voter-priorities-ahead-of-the-2022-midterm-elections/

10. United States Bureau of Economic Analysis. (n.d.). *Regional Data*. https://www.bea.gov/itable/regional-gdp-and-personal-income