2019

Econometrics in R Program

Ian Connors
ian_connors@my.uri.edu

Ian Connors

Professor Malloy

29 April 2019

Econometrics in R Programming

Introduction

In analyzing the economic trends of Rhode Island, I became interested in how the State compares to that of the rest of New England. Additionally, I sought to observe the public opinion of Rhode Island's current economic state as well as the possible implementation of any policy adjustments. To address this issue, I aimed to observe big data using datasets provided by the Federal Reserve Bank of St. Louis and the Bureau of Labor Statistics. In obtaining a greater understanding of how to analyze and correlate these datasets between each other, I began the process of learning R Programming. R is a coding language that allows for statistical computing and design with terminology compatible in other programs (such as S and C++). My main individual objective for this project included gaining a better understanding of data analytics using R-programming. I was not very familiar with R-programming, having only a general concept of how to use it. However, learning this program and implementing it into my project has provided me with the knowledge to develop further economic models in the future. From my research, I have found that R-programming may be my most useful tool in analyzing big datasets, as it allows for facilitated manipulation of data. Using R-programming allows the user to subset data to simplify the observed statistics. I am very interested in data analytics and this project has provided me with a greater understanding of how much work goes into programming.

Before beginning my research, I needed to formulate both a null hypothesis and an alternative hypothesis; My null hypothesis being that if the Rhode Island economy maintains its

levels of economic growth, it will follow the same growth rates as its neighboring New England States; My alternative hypothesis being that if the Rhode Island economy maintains its levels of economic growth, it will fall behind its neighboring New England states on a scale of GDP. Obtaining the datasets for New England's economic indicators allowed for me to test this hypothesis in R. The availability of resources required to perform my project was only slightly limited. I was able to obtain public datasets online through national databases, census and survey records. However, I could not gain access to the URI library dataspark resource. After obtaining my datasets, I installed data extraction software to store my data internally. I then needed to prepare my R-programming using a codebook to sort my datasets. Setting up the structure of my data facilitated the clarification process by sub setting the statistics into various categories. After entering the data into the R, I had to screen for errors in the data to increase the accuracy of my findings. Once the data was loaded into my R-program, I was able to manipulate it into subsets that allow for descriptive statistics. These descriptive statistics categorically involving frequencies of data and numerically describing the mean, standard deviations, min & max, and skewness of the datasets. Using this program, I was able to create, charts, graphs, scatterplots, and other forms of data tables.

To build my statistical model, I used data mining to predict public opinion using classification, regression, and deviation detection techniques. I then clustered my data to describe what exactly it is I am analyzing. After clustering my data, I was able to examine the correlation amongst my numerous variables using ANOVA, logistic regression and factor analysis. I then may compare these groups against one another to find statistical significance in their correlation. After finding the p-values of all economic indicators, I was able to graph these findings using DPLYR and GGPLOT2 within R Program. Next, I was able to data mine hashtags on Twitter to

uncover a positive outlook of social media trends following Rhode Island economics. Lastly, to uncover potential policy enactment of issues that I uncovered in my research I interviewed Judicial Chairman Robert Craven of North Kingstown to observe a State Representative's stance on these issues.

Learning R

To delve into the specifics of learning R, I spent the majority of my time watching instructional videos and taking online courses instructing me the basics of the programming language. To understand how to properly create a model within the program, it was imperative that I understood how the language functioned. This is a very tedious task as a mistake in a single line of code will result in an error report within the entire project. Thus, I first had to learn how to type the language in the correct format so that the functions would perform the tasks I wanted them to perform. This begins with an understanding of data types. There are five data types within R: vectors, matrices, arrays, lists, and data frames. These contain variables (data which is stored to a specific location within the file) that each function differently depending on the data type. Three variables can be created in R using an arrow command, including Integers, Numeric, and Boolean. Beginning with vector data types, this can be described as a sequence of data elements of the same basic type. The five atomic vectors are Logical (True or False functions), Integer (listing digits and their value), Numeric (sequence of numbers including decimals), Complex (sequence of equations) and Character (sequence of words denoted by quotation marks). The next data type of matrices can be described as objects in which the elements are arranged in a two dimensional rectangular layout. You may input a vector into a matrix which becomes the data elements of the matrix. The matrix is then sorted into nrow and

ncolumn to determine the size of the table to be created. A byrow is created as a logical clue to implement header rows or other various techniques to arrange your vectors. The next datatype, arrays, can be described as the R data objects that can store data in more than two dimensions. This computes the vectors of multiple matrices for more advanced calculations. The fourth datatype is a list. Lists are simple as they are objects in R which contain elements of different types, like numbers, strings, vectors, and other lists inside it. Lists let you store multiple vector types and will not default to integer type like others. The last data type, and most important, is a data frame. A data frame is a table or two-dimensional array-like structure in which each column contains values of one-variable and each row contains one set of values from each column. Data frames are created when datasets are imported or uploaded into R environment. These are stored as files which can be manipulated through the use of breaking down rows and columns into specific variables. These variables can then be assigned to any of the data types to create whatever graphic or equation analysis in which you are looking to achieve.

After understanding the different data types, it is imperative to understand what different commands do. An arrow command assigns a value to a variable. This is useful to create shortcuts by assigning broken down data frames into smaller sets. Equal functions are used to equate one of the data types into an operation, such as creating an array of a specific matrix. This is ordered using parentheses to denote which row and column are to be used. The numbers remain the given integer when a comma is used, but form a list between two numbers when a colon is used. Furthermore, mathematical operations may be incorporated into these functions to repeatedly perform the same equation within a dataset.

The next step to understanding R is learning operator commands. Operators are the constructs which can manipulate the value of operands into different types. Arithmetic operators

perform addition, subtraction, multiplication, etc. Relational operators perform equal to, less then, greater than or equal to and etc. commands. Logical operators perform AND, OR, or NOT operations to equate functions into alternative pathways. This then leads to a conditional statement that starts with an IF function. When the function begins, it will perform its task only if a conditional statement instructs it to do otherwise. Conditional coding is done using parentheses to execute a certain statement. This relates to switch coding which performs exactly the same, but instead of preventing the function it switches it to another. When an IF condition is not met, an else IF condition occurs which can be assigned to a True or False value. This can be repeated for numerous statements depending on how many conditions you wish to have in your operation. When an operation does not end but instead repeats itself, it is terms a loop statement. A loop statement allows us to execute a statement or group of statements multiple times. It repeats a statement or group of statement while a given condition is True. It tests the condition after executing the loop body. This can be repeated for a fixed number of times depending on the operation you are performing. Strings may be added into these functions to increase the quantity of objects being operated on at once. These functions are then all applied into different testing methods, such as ANOVA, regression, linear, and T-tests.

When configuring your data, it is imperative that you learn GGPLOT2 to create more dynamic graphics and charts. GGPLOT2 can be uploaded into R using the download.packages command. Using the GGPLOT(df) function will upload your data frame into GGPLOT2 to create your plot. Next you must upload the application from your library and attach your data. You may then implement whatever aesthetic available within the application into your graphic to enhance the quality of its portrayal. It allows for facilitated manipulation of the X and Y axis bounds and logarithms as well as aspects such as colors, sizes, shapes, highlights and layers. This

application allows for the possibility to facet wrap graphics to separate scatterplots into different subcategories. After learning how to use this application to create graphics, it is imperative to understand the different foundations in which it is used for.

Various foundations in which R programming may be used for include linear regression, statistical tests, outlier analysis, logical regressions and ANOVA tests. In my findings, I used linear regression to find the correlation of economic indicators relative between the six states of New England. I typically used a scatterplot graphic to portray the relationship between each indicator in both a faceted and unfaceted format. I was then able to run regression diagnostics to test if the indicators had statistical significance. To find out if variables share statistical significance, one must determine the p-value compared to the significance of your hypothesis. In the case of my project, all but two statistical significance p-values were <2e-16; this is the smallest number that current technology can compute, meaning there was substantial significance between the variables. R also allows the user to equate operations such as mean, median and mode within the datasets. Implementing logarithmic operations into my datasets allowed my graphics to portray less skewed results.

Implementing R into my Project

The first step I took in beginning my new project file in R program was to import or install datasets into the environment. To find the datasets, I used free public data provided by the Federal Bank of St. Louis and the Bureau of Labor Statistics. R studio allows the user to import an application programming interface key (API key). An API key can be obtained by emailing the data sites requesting access to their data for school research. After emailing data sites for API keys, I was able to access a BEAkey and BLSkey. The purpose of the API key is to act as a

password which passes code through the application giving permission to the user to access the data. It is important to set the API key to a variable so that it is easily accessible in the future. When importing the dataset into RStudio, the use must input all the functions of the code to operate the desired function. For example, if calling for unemployment data between the years 2008-2017, the user must bracket the specific coding specifying exactly what they want. Typically, the API, column, row, frequency, dataset, and inputted equation are all included. For the datasets that I could not download because they were outside the bounds of the API key to locate, I had to create an excel file and convert it into a .CSV file. This was done by creating the header row and then repeating each state and national indicator subsequently ten times along with the years 2008 through 2017. When the data is lined up, the it is important to upload the .XLS file into a .CVS file to match the RStudio input. After uploading my data as datasets into the environment and creating variables for them, I was able to begin separating each indicator by state. I would indicate which dataset was which using an arrow command to associate the name of the indicator to the string of vectors it represented (ex Unemploy ~ Percapgdp <- data1[1:60]). I filtered out the indicators by each state using the filter operation and quotation of each state. Since each state was in the header I created in the .XLS file, I was able the simply type the name to pull up the results. After all states' coding was assembled and cleaned up, I was able to run statistical regression on the indicators. Using ANOVA testing for multiple variables, I found the p-values of the data to find very high statistical significance in every instance.

When creating the graphics for my results, I compared each of State's GDP per capita to the different indicators relative to the rest of New England. I used GGPLOT2 and DPLYR to manipulate the graphics using different aesthetic techniques. I set the x-axis as the GDP per capita data and the y-axis as the relative indicator. Additionally, I used logarithmic functions to

unskew my GDP data to give a better representation on the graphic. I used color to represent the state vector on the unfaceted graphs and the year on the faceted graphs. Size equated to population on each graph. Furthermore, I used geom_point and geom_smooth operations to straighten and lighten the scatterplot lines. Lastly, I had to create the titles, fonts, sizes, etc for the headers of the axis.

My findings found my hypothesis to be true: the p-values were extremely significant in my data. The greatest statistical significance I found in Rhode Island's discrepancies relative to the rest of New England was its High School graduation rates relative to per capita GDP. Rhode Island has the greatest inequality in terms of the wealthier districts attaining much higher graduation rates than the lower income neighborhoods. To understand this further, I interviewed my internship supervisor, Judicial Chairman Robert Craven. I asked Representative Craven several questions regarding my data findings. I have provided an example of some questions and answers during the interview (including the question relevant to High School graduation rates).

Q: After looking at this data, do you believe the general public would be satisfied with Rhode Island's results?

A: Rhode Island is better than it was, but is not good enough. While Rhode Island's economic indicators are outperforming the northern New England States, Massachusetts and Connecticut have competitive advantages. The location of these states are relative to major commercial hubs (New York and Boston) and Massachusetts invests more in infrastructure.

Q: Why is there disparity amongst Rhode Island's High School graduation rates compared to its neighbors? How have Massachusetts and Connecticut been more successful in producing high bachelor's degree attainment levels?

A: Rhode Island has many areas of high household income as well as many areas of the opposite. In State policy, the public school is paid for by the income tax of its respective township, meaning low-income neighborhoods do not receive nearly as much aid. If funding was dispersed on the State level, then there would be an increase in high school retention rates of low-income school districts. To maintain high bachelor's degree attainment levels, MA and CT each have very high-income ivy league districts funded by donors. In terms of state loans, Connecticut has one of the best in-state-tuition programs in the country that Rhode Island should aim to model. Universities must be located within a vibrant economy to attract students.

Q: Do you think that raising the minimum wage in Rhode Island to that of Massachusetts would lead to facilitated college availability and therefore higher graduation rates?

A: The minimum wage should be regarded as an entry-level wage that is typically not suitable to sustain a four-year college degree. If the minimum wage is raised, then business backlash may produce unemployment as employers cannot afford to compensate their employees. Thus, this must be regarded as a balancing test to determine where economic equilibrium lies.

Overall, Chairman Craven found that Rhode Island is projected to increase its social spending to promote projects such as infrastructure. He believes that this may attract investment into the state and may even alleviate high school dropout rates in lower income districts. When schools are properly funded, the future of the State will be in more capable hands.

Literature Review

The data surrounding R-programming techniques has shown vast expansion within the previous decade. Using big data to analyze economic trends can be observed as a facilitated method to investigate a State’s economic and business aptitude. In comparing Rhode Island’s

economic sectors, a researcher may correlate any economic faults to their corresponding shortcomings. In analyzing and comparing this data, I have chosen to use R-programming software. R software can be observed as a great tool for this task as it is very flexible for econometric algorithms; it is unrestricted by the functions of the default packages as it is an ever-expanding language of statistics (Farnsworth, 2014). New algorithms are constantly added to the R-programming site to exponentially increase the data processed into the software. The software and data are easy to locate and extract even by somebody with little technical knowledge of coding systems. Learning R-programming includes focusing on topics such as basic data manipulation, sorting data, complex math operations, differing data types, working with dates, merging data frames, and editing data directly. This is also a great program to work with large data files; able to perform linear regressions, time series regressions, plotting, graphing, and statistic outputs, programming new algorithms becomes possible when the language is learned.

To help teach myself the language of R-programing, I researched online courses for beginner learning. Cognitive Class.ai is a website that has teaching modules with lessons ranging the wide topics of data sampling programs (Cognitive Class, 2019). I believe that learning R-programming is not only a useful skill in creating this current project, but in my future occupational career as well. Analyzing large datasets is a lucrative skill in modern labor markets for business predictions, stock projections, and financial advising. This need for data econometrics can be emphasized by the recent 2008 recession; businesses would like to know when markets trends are shifting (Russom, 2011). The main indicators that analysts are focus towards include the management, the visualization, and the modeling of data. The management of data includes loading, coding, and cleaning errors. The visualization of the data describes the statistics in terms of histograms, scatterplots, etc. Lastly, modeling the data uses the previous two

steps to create models for projecting econometric data (Manzan, 2019). Thus, visualizing the data is extremely important because it allows a researcher unskilled in econometrics to easily interpret your statistical findings.

The initial obtainment of the economic data is downloaded through various economic databanks. These online sources are extremely useful and usually free to use. For instance, The United States Bureau of Labor Statistics contains many important datasets relevant to both State and Federal level economic drivers of the United States (Bureau of Labor Statistics, 2019). Additionally, there are sites with directories aimed as a proxy to specific datasets already encoded in r-format. Examples of this include the r-directory website (r-Directory, 2015) and Kaggle (Econometrics, 2018). This data is typically already packaged to the necessary .csv file that is observable in an excel format. When downloaded, it must be unzipped and imported into r-Studio to work with and analyze. In obtaining this data, it is important to first research what other economists have done using r-programming in their own econometric projects.

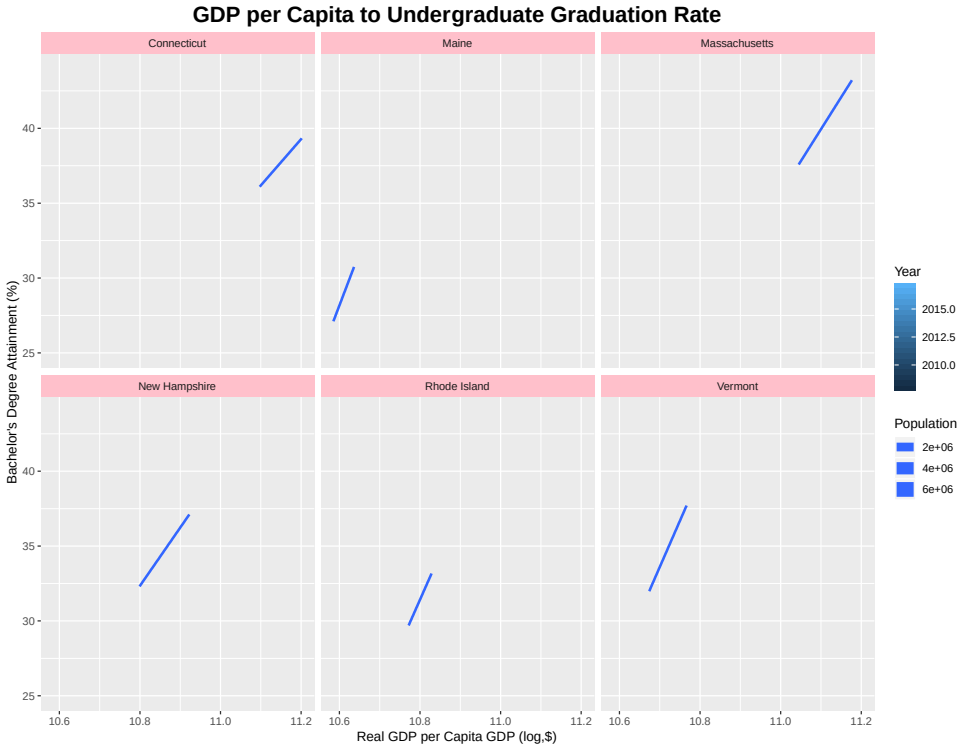
The use of r-programming for econometrics is emerging as a great tool within the world of academia. Even in the mid-twentieth century, configuring inputs to match your datasets was no easy tasks; as researchers Bernard Dzielinski and Ralph Gomory have attempted to trace optimal programming of lot sizes, inventory, and labor allocations (Dzielinski, 1965). This research contributed to the knowledge surrounding the necessity of programming lot sizes for big data that is typically incalculable without computer programming. Creating alternative set-up sequences for computation may aid in understanding the variations of programming issues and solutions. Finding useful solutions to programming questions is shown to forecast economic projects. Professor Ruey Tsay analyzes linear models for financial time series data. Tsay calculates real examples introducing statistical concepts, illustrating step-by-step analysis, and

discusses financial applications (Tsay, 2014). R-programming is able to detect trends that can predict seasonal patterns, which Tsay refers to as "seasonality" trends in time series data. Tsay's ultimate goal is to "study the dynamic dependence of the time series so that proper inference of the series can be made" (Tsay, 2014). This ability to infer where economic shifts occur is imperative in the financial decisions of large businesses. The development of socioeconomic policy is also a major sector in which r-Programming can assist. Dominique van der Mensbrugghe uses R to extract per capita historical GDP for country analysis to summarize growth episodes across regions in a box plot (Van der Mensbrugghe, 2016). Observing national GDP data can aid a researcher in developing economic policy decisions within a government. In my project, I aim to use similar research methods to analyze Rhode Island's economic statistics to project its future aptitude to its surrounding states.

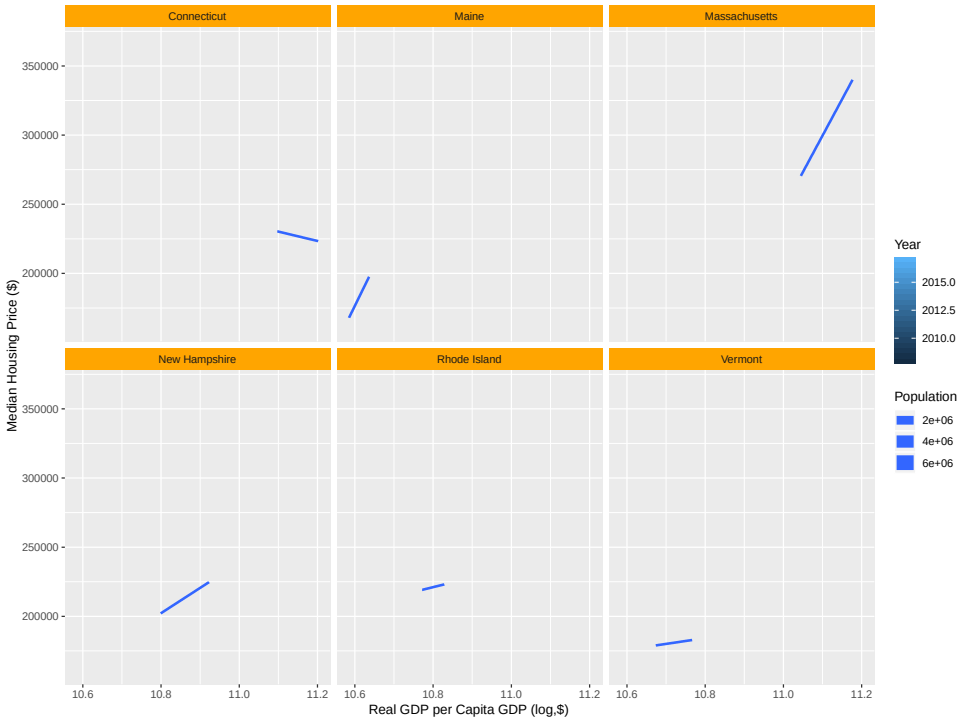
Analyzing Rhode Island's economy must account for numerous factors of growth statistics. The state has a labor force of 556,234 and an unemployment level of 22,181 people (4% unemployment rate) (Bureau of Labor Statistics, 2018). Comparing this to all other New England states recorded in the BLS, Rhode Island in fact has the highest rate of unemployment. New Hampshire, the New England State with the lowest rate of unemployment, has a labor force of 763,611 and an unemployment level of 18,377 people (2.4% unemployment rate) (Bureau of Labor Statistics, 2018). In understanding the unemployment rates and other factors of Rhode Island's economy, one must research why Rhode Island may be falling behind. One outcome may be contributed to Rhode Island's primary dependence on its agricultural sector. As the smallest state in the country, Rhode Island farmland covers 10% of the state spanning 69,000 acres (Farm Flavor, 2018). The loss of employment in this sector may hold responsibility for the lacking GDP growth of the state. However, Mark Maggi of the BLS attributes this growth

lagging to the loss of jobs in education and health services. He states “job losses in the colleges and universities industry, within the educational service sector, and the hospital subsector, within the healthcare and social assistance sector, were responsible for depressed overall growth in this super sector and the state overall (Maggie, 2017). Studying the industries in which Rhode Island is either ahead or behind will be the major indicator of where to focus future policy reform for the state.

Graphics

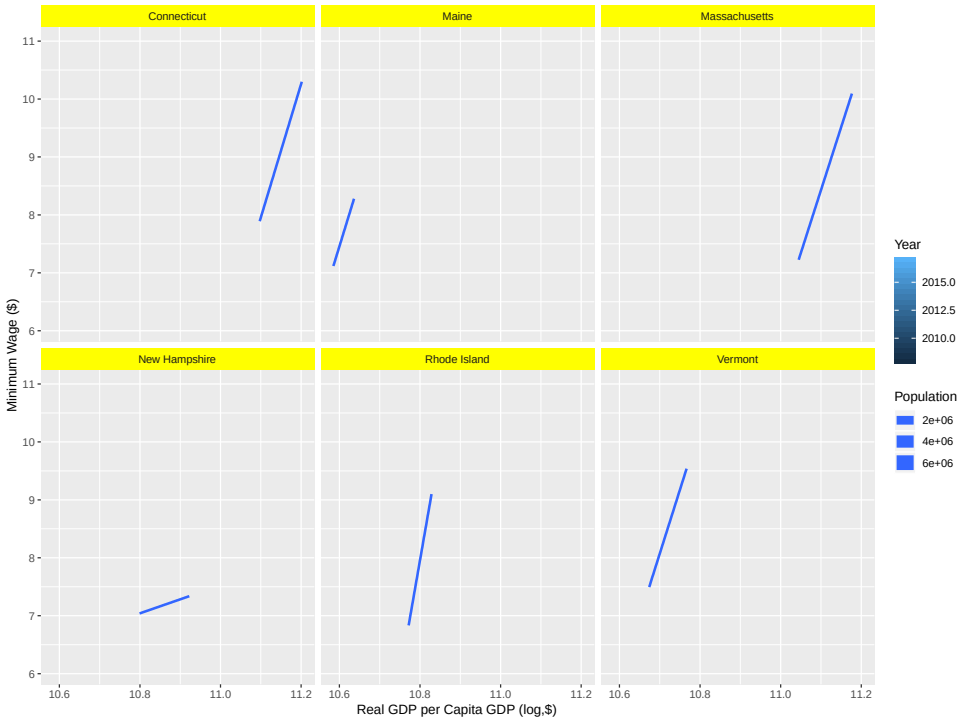


GDP per Capita to Housing Prices

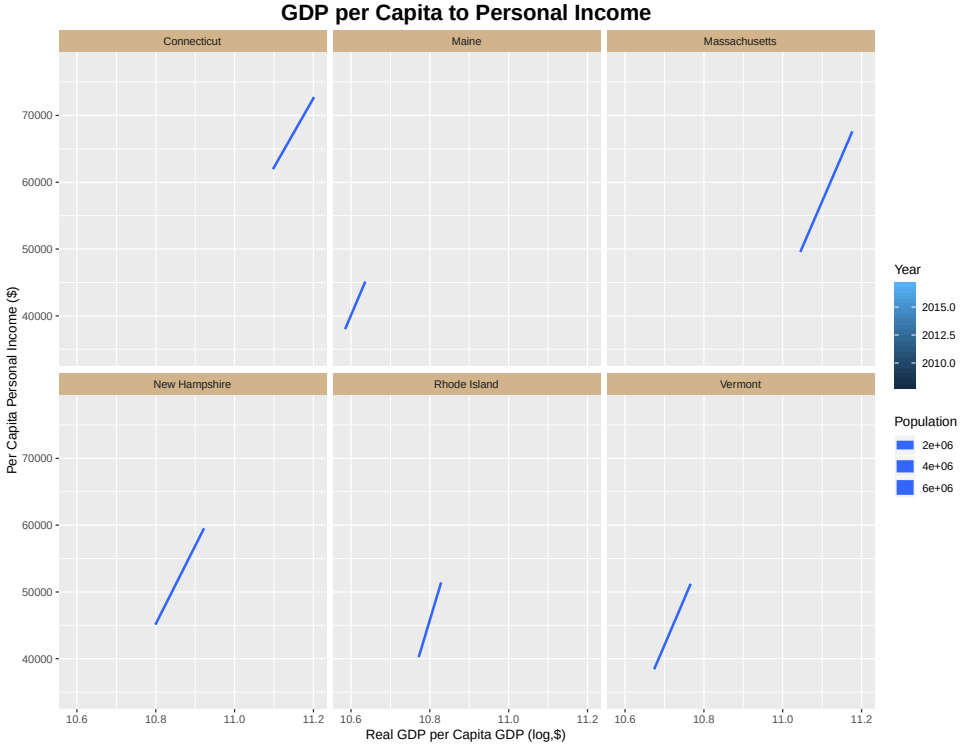


GDP per Capita to High School Graduation Rate

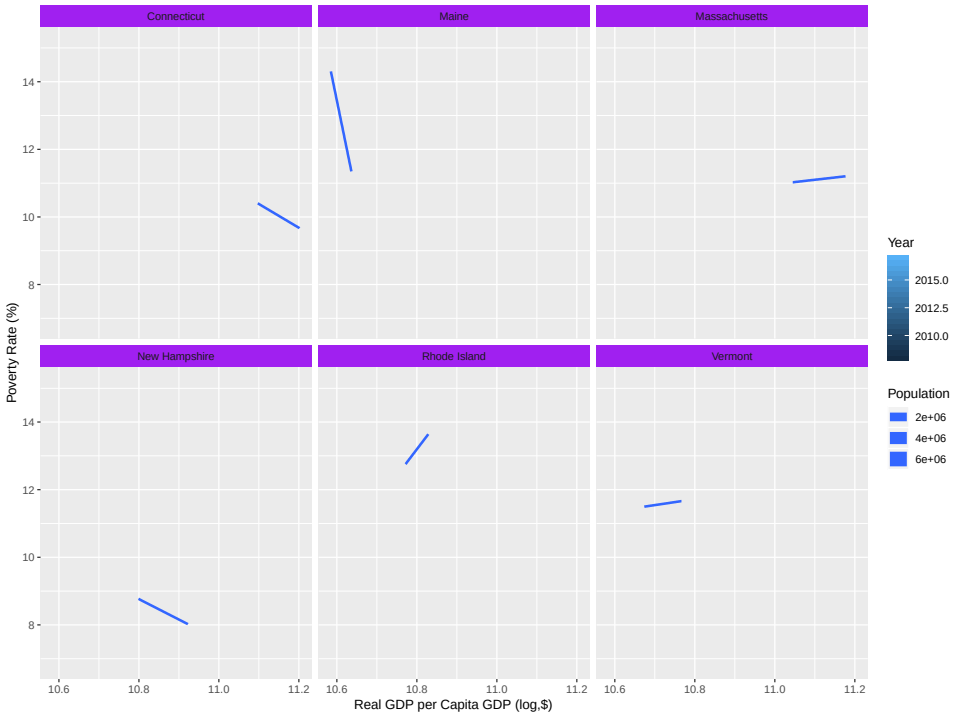
GDP per Capita to Minimum Wage



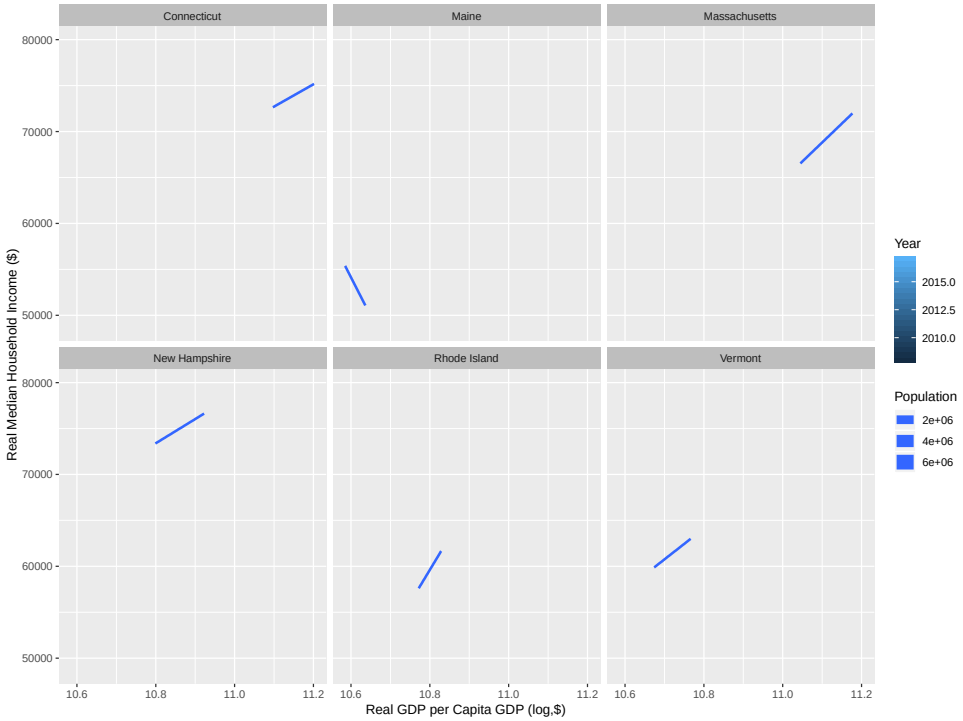
GDP per Capita to Personal Income

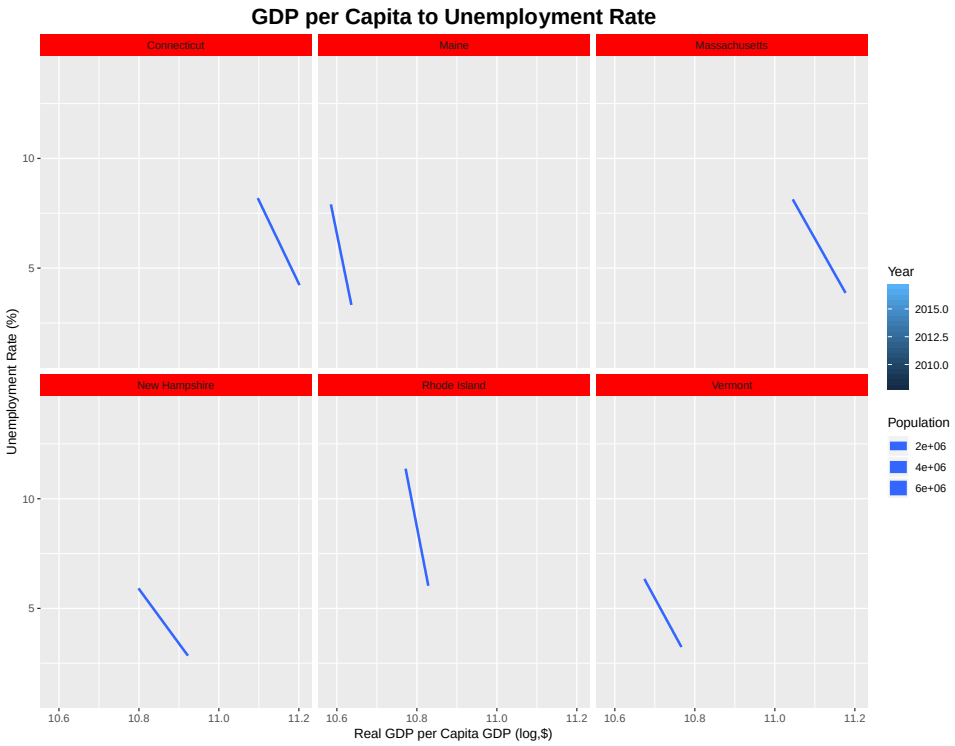


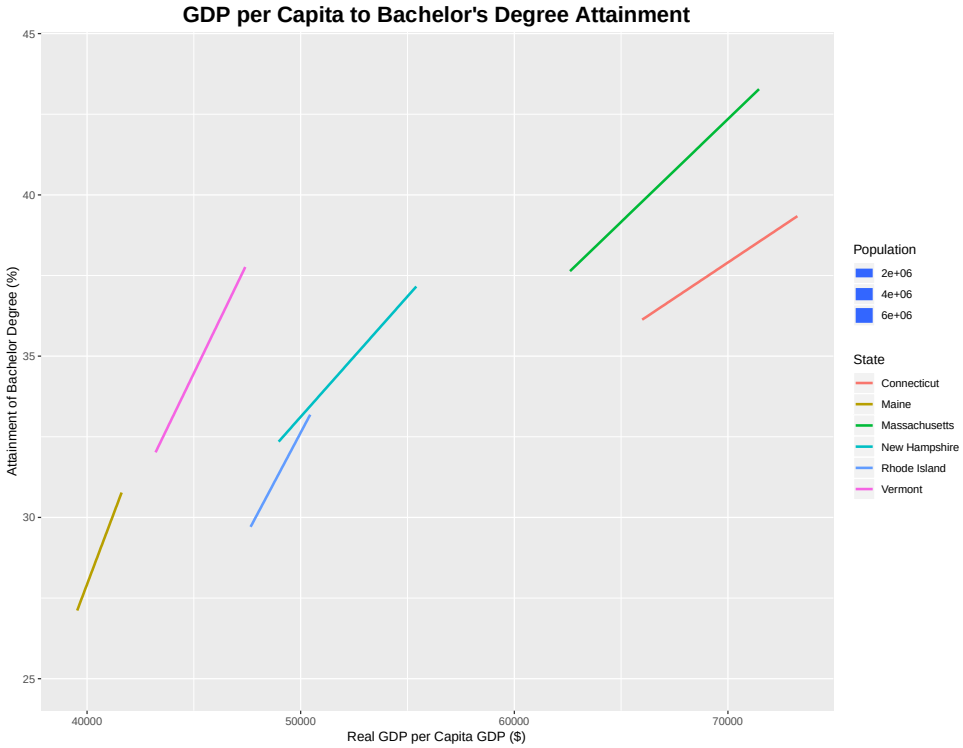
GDP per Capita to Poverty Rate



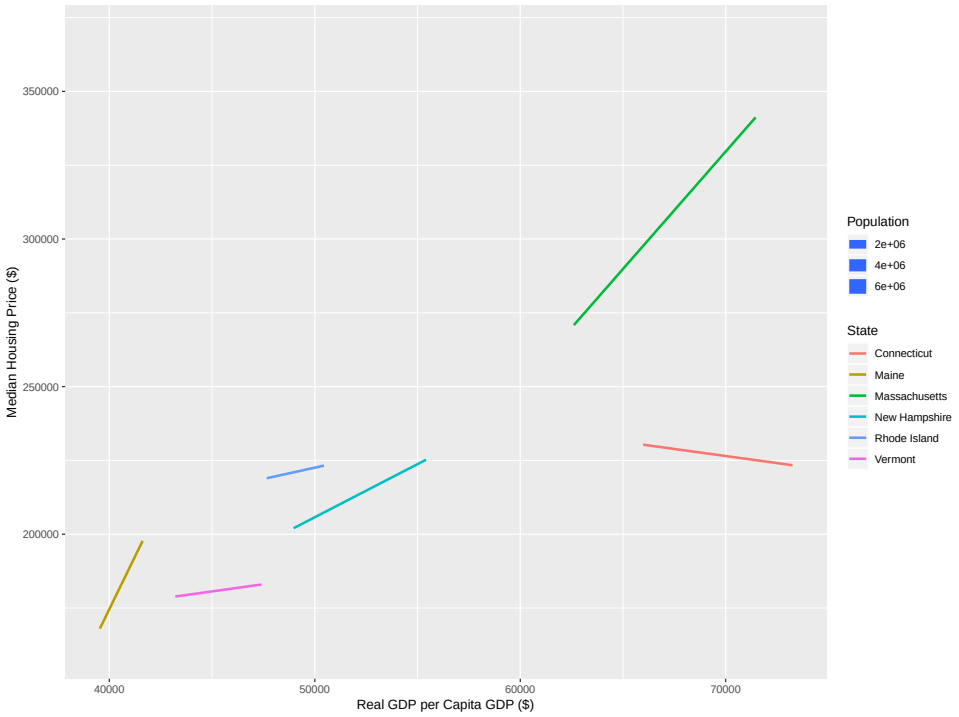
GDP per Capita to Household Income



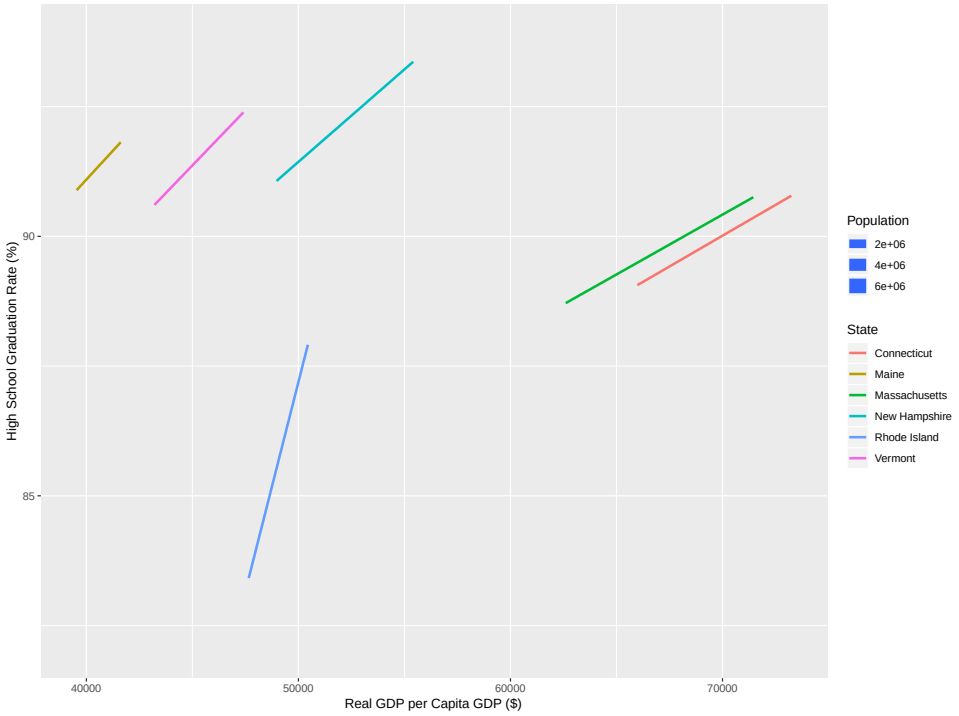

GDP per Capita to Unemployment Rate
Connecticut
Maine
Massachusetts
New Hampshire
Rhode Island
Vermont
Unemployment Rate (%)
10
5
10.6
10.8
11.0
11.2
Real GDP per Capita GDP (log,$)
Year
2015.0
2012.5
2010.0
Population
2e+06
4e+06
6e+06


GDP per Capita to Bachelor's Degree Attainment
Attainment of Bachelor Degree (%)
45
40
35
30
25
40000
50000
60000
70000
Real GDP per Capita GDP ($)
Population
2e+06
4e+06
6e+06
State
Connecticut
Maine
Massachusetts
New Hampshire
Rhode Island
Vermont

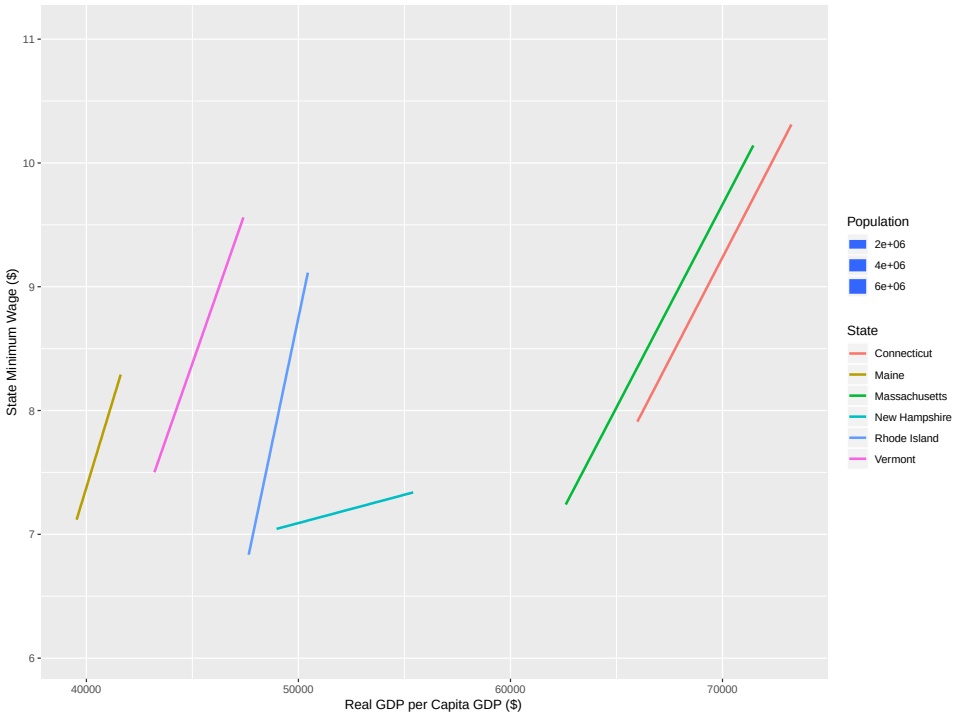
GDP per Capita to Housing Prices



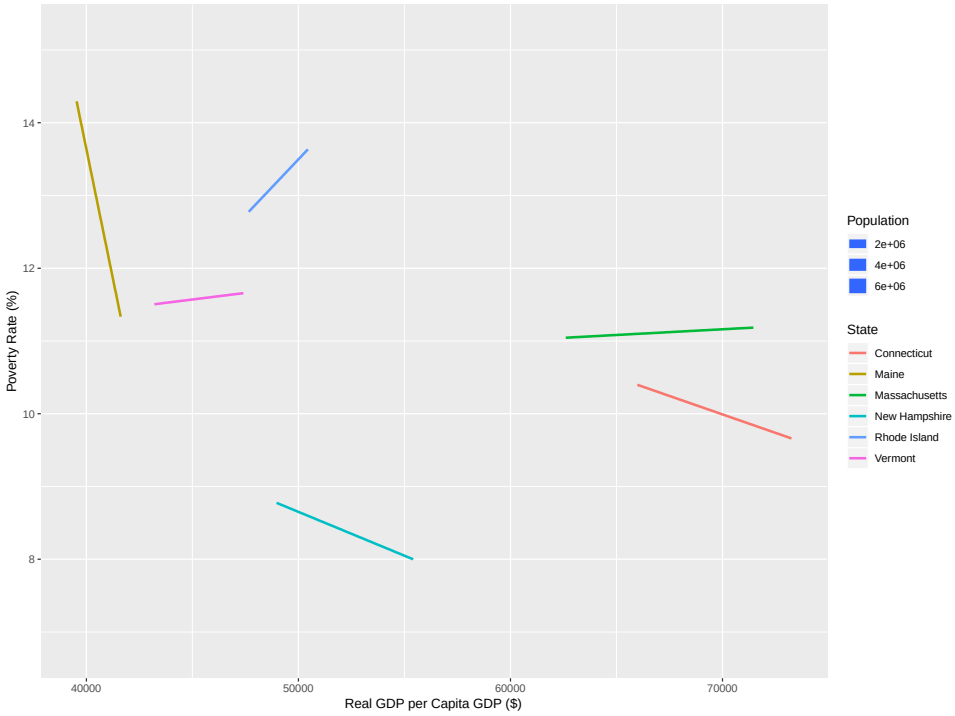
GDP per Capita to High School Graduation



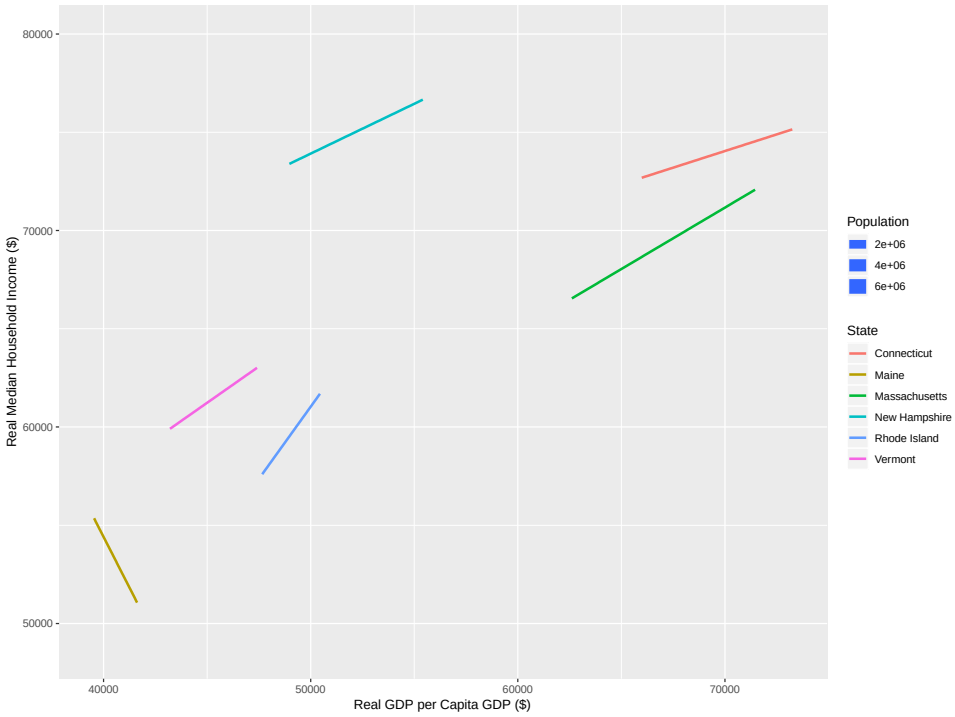
GDP per Capita to Minimum Wage



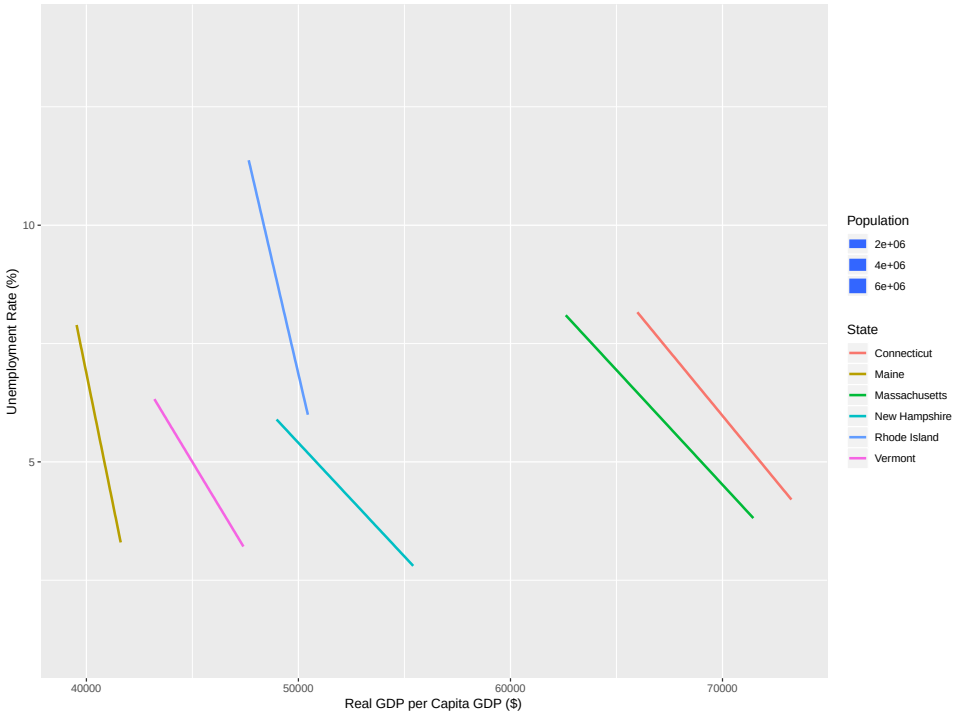
GDP per Capita to Poverty Rate



GDP per Capita to Household Income



GDP per Capita to Unemployment Rate



References

Bureau of Labor Statistics. “Bureau of Labor Statistics Data.” *U.S. Bureau of Labor Statistics*, United States Department of Labor, 2019, data.bls.gov/timeseries/lns14000000.

Cognitive Class. *RP0101ECN*, 2019, courses.cognitiveclass.ai/courses/course-v1:BigDataUniversity RP0101EN 2016/courseware/8dbed036880941c0848fc10aba85daa9/.

Dzielinski, Bernard P., and Ralph E. Gomory. "Optimal programming of lot sizes, inventory and labor allocations." *Management Science* 11.9 (1965): 874-890.

Econometrics, Kaggle, 2018, www.kaggle.com/datasets.

Farm Flavor. *Rhode Island Agriculture*, Journal Communications, Inc., 2018, www.farmflavor.com/rhode-island-agriculture/.

Farnsworth, Grant V. “Econometrics in R.” *MIT.edu*, MIT OCW, 24 Mar. 2014, ocw.mit.edu/courses/economics/14-381-statistical-method-in-economics-fall-2013/study-materials/MIT14_381F13_EcnomtrisInR.pdf.

Federal Reserve Bank of St. Louis. (2019). https://fred.stlouisfed.org

Maggi, Mark. “Why did Rhode Island have slower employment growth than the nation during the recovery period, 2009–16?” *Beyond the Numbers: Regional Economies*, vol. 6, no. 11 (U.S. Bureau of Labor Statistics, September 2017),https://www.bls.gov/opub/btn/volume-6/why-did-rhode-island-have-slower-employment-growth-than-the-nation-during-the-recovery-period-2009-16.htm

Manzan, Sebastiano. *EVOLUTIONARY THEORY AND THE PSYCHOLOGY OF EATING*, 2019, faculty.baruch.cuny.edu/smanzan/bus4093.html.

r-Directory. *Free Datasets*, R-Program, 2015, r-dir.com/reference/datasets.html.

Russom, Philip. "Big data analytics." *TDWI best practices report, fourth quarter* 19.4 (2011): 1-34.

Tsay, Ruey S. *An introduction to analysis of financial data with R*. John Wiley & Sons, 2014.

Van der Mensbrugghe, Dominique. "Using R to Extract Data from the World Bank's World Development Indicators." *Journal of Global Economic Analysis* [Online], 1.1 (2016): 251-283. Web.