Data Incubator Cap Stone Project Proposal

Do High Pay States Attract More Skilled Public Employee?

Predicting Financial Crisis?

Zefan Wang

University of Delaware

April 25, 2018

Contents

1 Introduction

For the cap stone project proposal, I prepared two project proposals. One relates microeconomic labor economics, the other is of macroeconomic concern. I found both topics on the Data Incubator's blog. The reason I have done two is because that I am interested in both of these in terms of data and topic. I am hesitant to drop one over the other. This proposal is divided into two sections, one for each topic. Each section contains a description of the project, kind of questions might be answered, what will be achieved in the end, strength and weakness of the data and project, and several basic exploratory graphs. The Jupyter Notebook contains codes, more analysis and graphs. I am sorry that if two project proposals would create any inconvenience. I am interested in both data, it is hard for me to drop one.

2 Microeconomic Project

This section will briefly go through the state salary project, which is related to labor economics in some extent. The project aim to collect and analyze data on public state employee salary data for each state. The data is on individual level. The volume of data is enormous, Georgia solely contains around 360 MB data and 4 million observations. The size of data for all states will much more. The data conveys a tremendous amount of information across states and years. One goal of this project is to see if states with higher public employee salary attract more high skilled public employee over time. Other questions such as posted on TDI's blog can also be answered, such as the correlation among professor's salary, school ranking and citation number. The final product of this proposal would be an interactive U.S. map which can tell detailed information on professions within and across states.

2.1 Questions to be Answered

1. What is highest pay positions within a state?

2. Which public positions have highest growth potential over time?

3. What is average salary for each state? Is salary in some states higher than other state? And for a specific position?

4. Is there difference in salary of same position across adjacent states?

5. Is salary of professor related to school or department ranking, and the cited times?

6. Do high pay states attract more high skilled public employees over time?

2.2 Strength & Weakness

The strength of the project and data is that it contains a large amount of information across individuals, regions and time. The disadvantage is that the lack of features in the data. The only information the data provides is quite basic such the salary, name, and position, with some variation in detail in each state. Other than those, it is hard to find more information on the individual. I am not sure if there is any other public information on the individual which can be merged into the data set. Therefore, it is hard to use machine learning or econometrics to build a model which gives deeper analysis. And the time span might not be long enough for a time-series analysis (generally from 2010 - 2017). Correlation analysis and so on can be performed on the data, but it can not be concluded as a causal relationship. Data for some states is harder to collect than other, for example, California provides data on the website, but I think one might need to do some web scraping to get it into the right form, which I am currently not sure how to do.

2.3 Data & Graphs

This subsection will present some information on data and graph. I did some analysis for Georgia due to the reason that the data is cleaner than other state's.

Some key information on data:

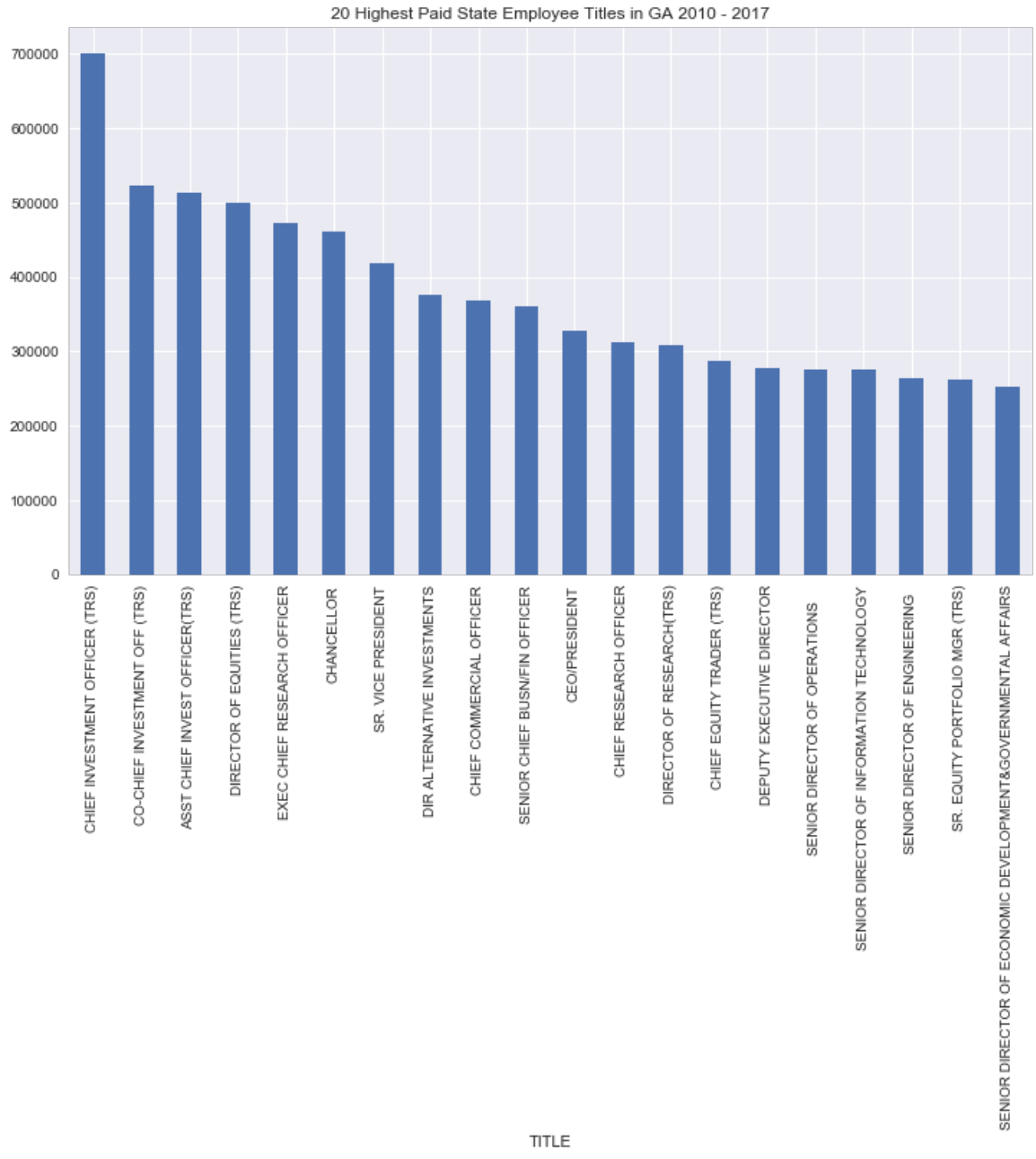
1. some states don't have data on salary of professors or higher education.

2. Washington State: 2012 - 2016

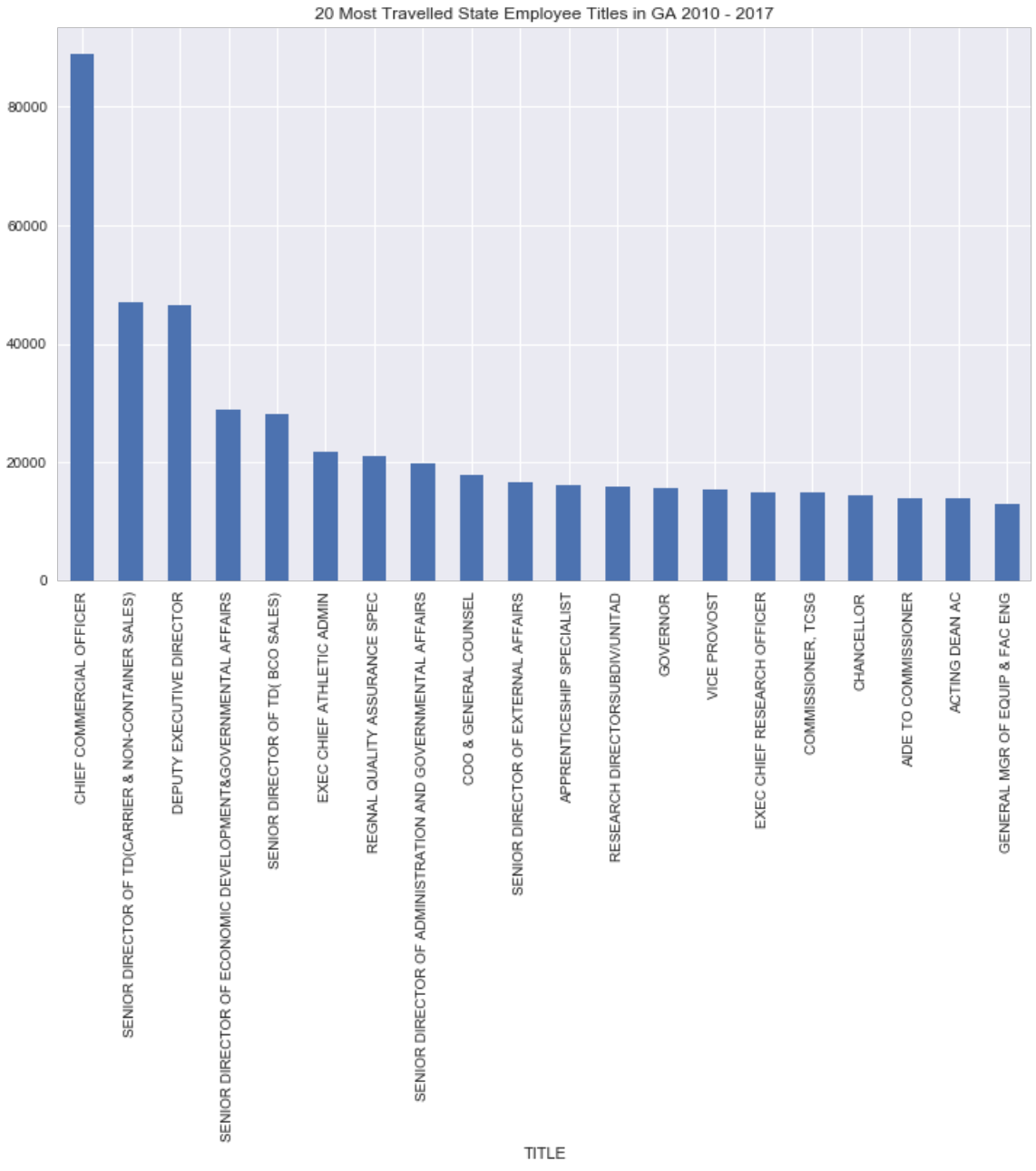
3. Ohio: 2010 - 2017: psychiatrist has dominant pay, from mental health department. Do not contain data on higher education salary.

4. Georgia: 363 MB, 4,012,395 observations from 2010 - 2017, has data on salary of professor, contains travel information as well. This information can used to be answer questions such as: which titles travel the most? Is there a relationship between travel and salary? Pearson Correlation between salary and travel: 0.326.

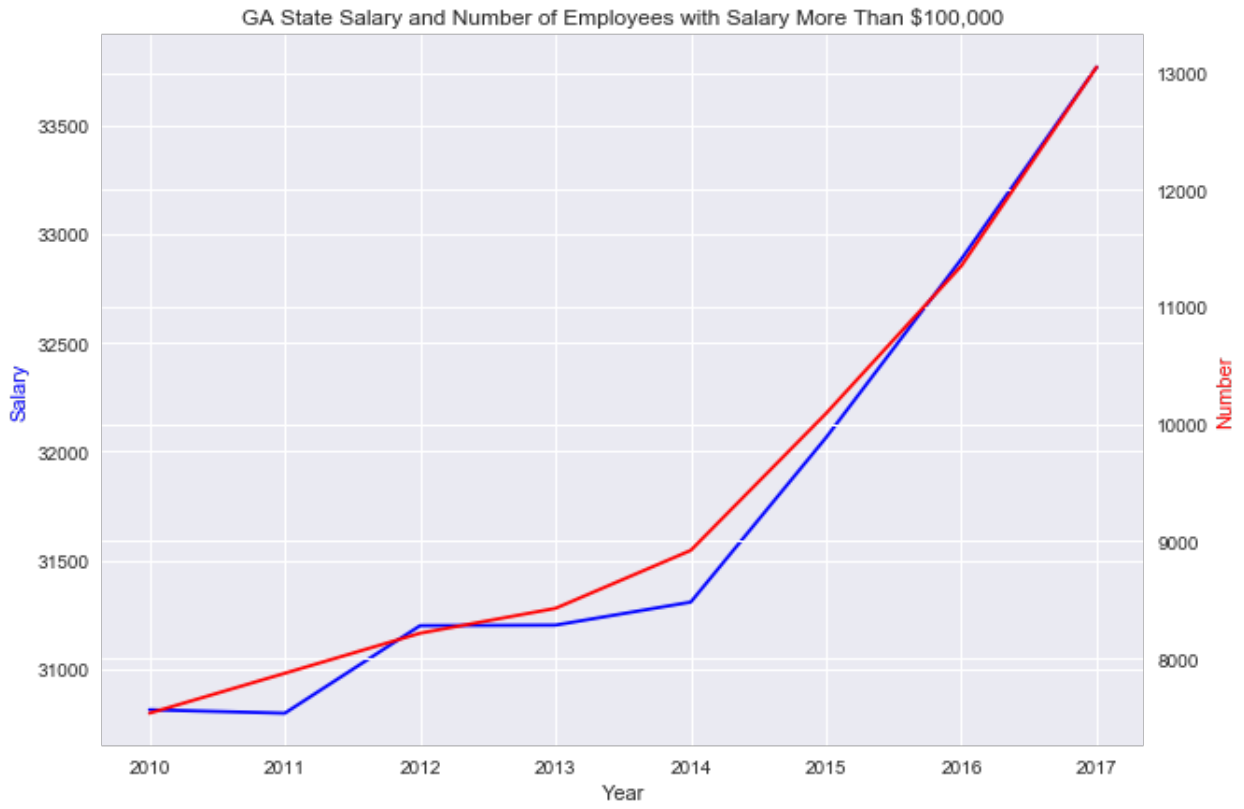
20 Highest Paid Title in Georgia



20 Most Travelled Title in Georgia



GA Average State Salary & Number of Employee with Salary over $100,000



2.4 Future Improvement

It would be much better if there is more information on the individuals and build machine learning or econometric model to generate deeper analysis, and more solid results.

3 Macroeconomic Project

One of the questions people tend to ask is that how is the economy, or is there going to be a recession soon? I have always thought of doing a classification project on financial crisis using machine learning.

I am uncertain if this can be accomplished under the current stage in terms of data availability and transparency. It is a hard question to answer or answer correctly: is there a way to predict and prevent a crisis? There is a general public question that why did lots of professionals or economists miss to foresee the 2008 financial crisis. Several reasons could be: 1. Everyone believes that the economy is in a good shape. 2. Every crisis comes in a different shape, appearance or form, which means that every crisis is unique, so it is hard to predict. 3. A lot other things going on before and during a crisis, so it is hard to identify which factors are critical for causing a crisis, especially when data was not as available as today, or the way to collect and deal with data was not as advanced as today.

The idea was generated based on a book I read last year, "This Time is Different: Eight Centuries of Financial Folly" by Reinhart & Rogoff (2009). They studied episodes of different types of crisis for several hundreds of years across different countries. They performed some data analysis and identify several key indicators of a crisis such as: asset price, government debt, real economy activity, housing index and so on. I was intrigued by the idea that if it is possible to build a model using machine learning to predict a crisis. It could be a supervised classification project or unsupervised grouping project. I found the macroeconomic data on Jord-Schularick-Taylor Macrohistory Database when I read TDI's blog, which reminds me the book.

3.1 Questions to be Answered

1. Is is possible to build a machine learning model to predict a financial crisis or other form of crisis? Or would it be accurate?

2. How are monetary policy and fiscal policy affecting economic activities and inflation?

3. Is there a relationship between monetary policy and house price?

4. Has the well being improved for countries over year?

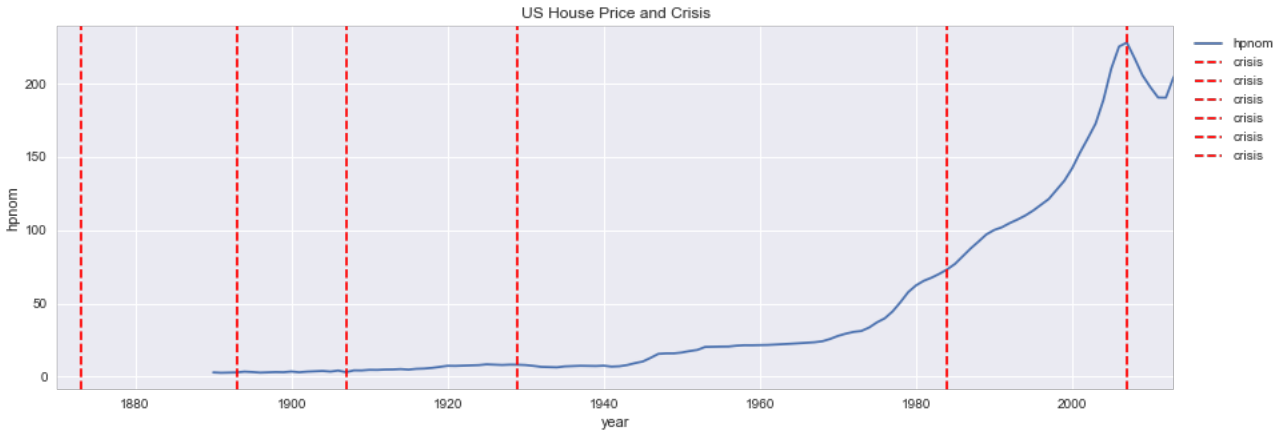
3.2 Strength & Weakness

The strength of the project is that it has a great potential, and it can provide information which could have some social impact. The concern I have are from several reasons: first, I do not consider myself an expert on the subject, so there might be factors/features I am not aware of; second, some government data or domestic data are not so transparent, so it is difficult to learn information from that; third, the sample size might not be big enough to generate a valid conclusion. On the other hand, this could be an on-going process that new data and valid features can always be added into model to improve results in future. This project has a great potential but might be too big, it might not generate a accurate conclusion. The good thing is that we can always tear the whole problem into smaller piece and start from these small pieces. For this data, even a small piece could provide some meaningful information. There is also some weakness on the data. The data only contains wealthy countries, and some data are missing for some features.

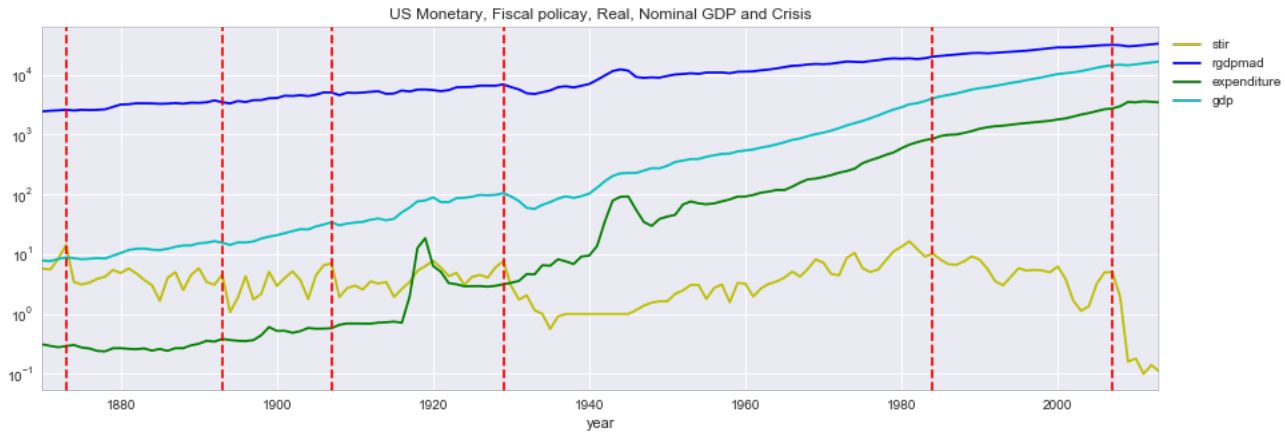
3.3 Data & Graphs

Jord-Schularick-Taylor Macrohistory Database contains data on gdp, house price, stock price and so on for wealthy countries over years. Below are some graphs for the analysis I did. Jupyter Notebook contains more information on other countries.

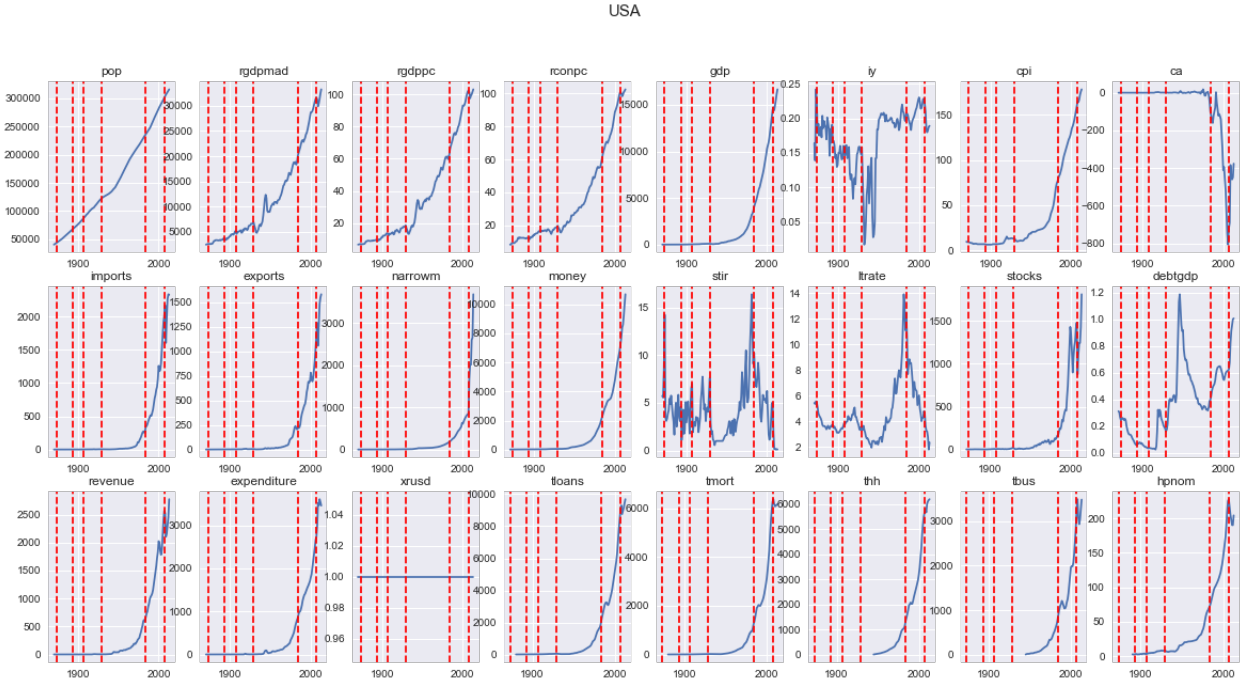
US House Price and Crisis Year



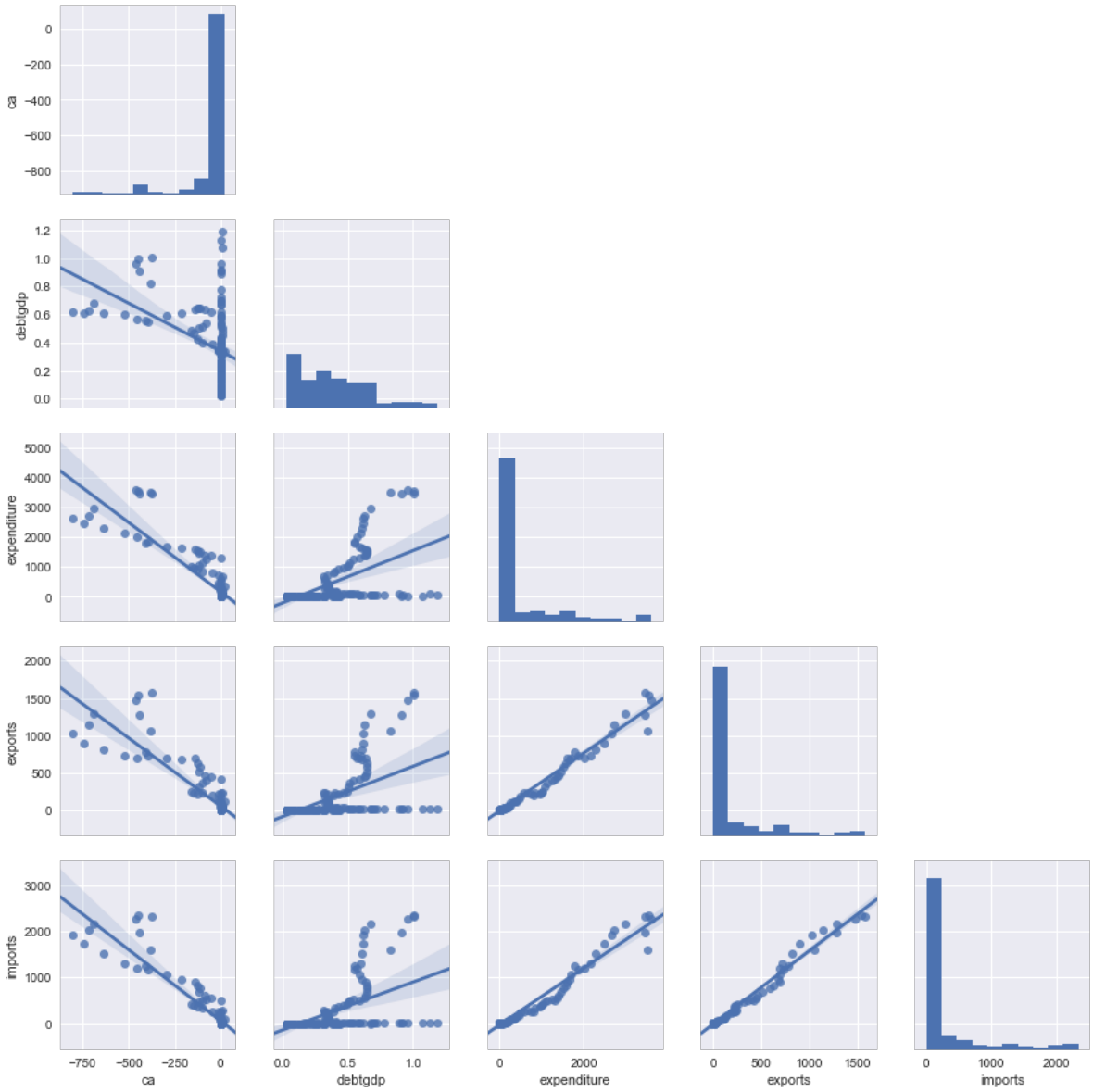
US Monetary Policy, Fiscal Policy, Real and Nominal GDP



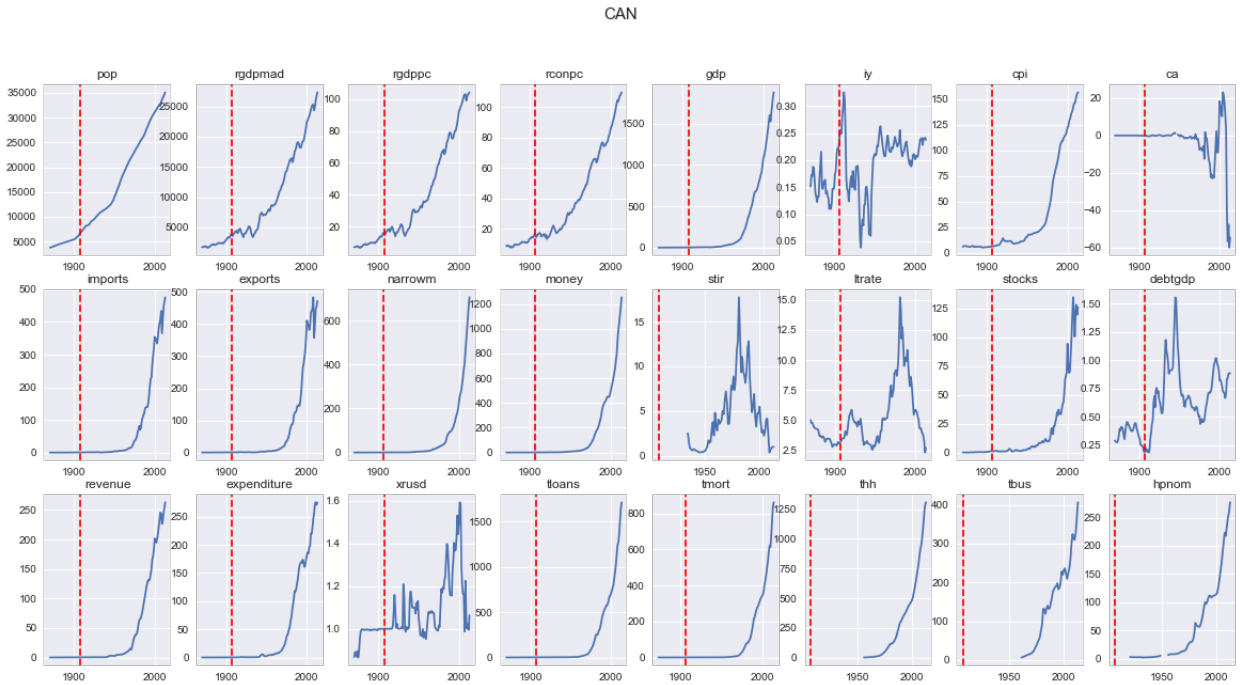
US Time Trend for All the Variables and Crisis Year



US Correlation Among Several Variables



Canada Time Trend for All the Variables and Crisis Year



Canada, which is believed having one of the most stable financial system in the world, has only one financial crisis over years.

3.4 Future Improvement

More data on other countries could be added to the data set. More data and valid features can be incorporated into model to improve the model.

4 Conclusion

Both micro and macro project are interesting and informational. However, they do have weakness respectively. Microeconomic project is short of features. Macroeconomic project might be too big to be accomplished and accurate.

References

Reinhart, C. M., & Rogoff, K. S. (2009). *This time is different: Eight centuries of financial folly*. princeton university press.