**Data Appendix – Andrew Vo**

**I: Data Appendix**

1, CPI:

* Definition: Monthly Consumer Price Index. Values are U.S. city averages for all items
* Unit: Number
* Original file: month\_CPI.csv – Original Variable name: Index
* Number of valid values: 568
* Number of missing values: 0
* Comments: Quite normally distributed, constant growth through time

2, CPI\_inflation:

* Definition: Monthly Consumer Price Inflation. Values are for U.S. city averages for all items.
* Unit: Number
* Original file: month\_CPI.csv – Original Variable name: Inflation
* Number of valid values: 568
* Number of missing values: 0
* Comments: Quite normally distributed, but fluctuated growth, there is a big drop near year 2007

3, avg\_HPI:

* Definition: Monthly average House Price Index accross the states.
* Unit: Number
* Number of valid values: 395
* Original file: month\_housePrice.csv – Original Variable name: avgHousePrice
* Number of missing values: 170
* Comments: Quite left skewed, constant growth but big drop around year 2007

4, long\_term\_interest:

* Definition: 10 year US Government Bond Yields (long-term interest rate)
* Unit: Percent per year
* Original file: month\_longtermInterest.csv – Original Variable name: Rate
* Number of valid values: 568
* Number of missing values: 0
* Comments: Quite normally distributed, grow to peak near 1980 but then constantly drop through time

5, federal\_funds\_rates:

* Definition: Monthly effective Federal Reserve Interest Rates.
* Unit: Percentage
* Original file: month\_interestRates.csv – Original Variable name: Effective Federal Funds Rate
* Number of valid values: 420
* Number of missing values: 148
* Comments: Quite normally distributed, more missing data but seem to have SAME trend as long\_term\_interest

6, budget\_on\_education:

* Definition: Yearly US budget on education.
* Unit: millions of dollars
* Original file: year\_educationBudget.csv – Original Variable name: BUDGET\_ON\_EDUCATION
* Number of valid values: 568
* Number of missing values: 0
* Comments: Quite rightly skewed, slowly grow till 2000, then grow faster but has a drop since 2007

7, gdp:

* Definition: Yearly US GDP.
* Unit: millions of dollars
* Original file: year\_educationBudget.csv – Original Variable name: GDP
* Number of valid values: 568
* Number of missing values: 0
* Comments: Normally distributed and constant growth over time

8, population:

* Definition: US yearly population.
* Unit: thousands of people
* Original file: year\_empRate.csv – Original Variable name: population
* Number of valid values: 568
* Number of missing values: 0
* Comments: Normally distributed and constant growth over time

9, employed\_percent:

* Definition: Yearly employment rate.
* Unit: percentage
* Original file: year\_empRate.csv – Original Variable name: employed\_percent
* Number of valid values: 568
* Number of missing values: 0
* Comments: quite lefty skewed, decently grow but has 2 drops near 2000 and 2007

10, unemployed\_percent:

* Definition: Yearly unemployment rate.
* Unit: percentage
* Original file: year\_empRate.csv – Original Variable name: unemployed\_percent
* Number of valid values: 568
* Number of missing values: 0
* Comments: quite normally distributed, unstable drop, seem to have a cycle every 10 years (Up then Down)

11 – 14, lowest, second, third, fourth:

* Definition: Yearly Upper income limits for each Fifth of all US households
* Unit: dollars
* Original file: year\_houseIncome.csv – Original Variable name: Lowest, Second, Third, Fourth
* Number of valid values: 568 each
* Number of missing values: 0
* Comments: Quite normally distributed, constant growth through time

15, top 5 percent:

* Definition: Yearly Lower income limits for Top 5% of US households
* Unit: dollars
* Original file: year\_houseIncome.csv – Original Variable name: Top 5 percent
* Number of valid values: 568
* Number of missing values: 0
* Comments: Quite normally distributed, constant growth through time

**II: Data Process codes and Graphs:**

**Basic Summary Statistics:**





The link below would show how I use python pandas library to process and combine the data from multiple files to one dataset for model. I included graphs codes for histogram and line plots at the bottom

<https://gitlab.com/hiep.vo/econ342_RecessionUs/-/blob/master/Data_Process/DataGroup.ipynb>

**Histograms:**

****

**Scatter Plots:**

****