Project Step 2

Joshua Greenert

2022-08-06

How to import and clean my data

1. Minimum Wage (Minimum Wage Data.csv)
    - Deleted footnotes column from the data set provided.
2. Unemployment Data (Unemployment.csv)
    - Deleted rows with empty data.
    - Reformatted data from columns to rows so that years can be segmented and shown over time.
    - Deleted area name as state is sufficient and fine-grain data will not be useful in this broad search.
    - Adjusted column names to be "Civilian_labor_force", "Employed", "Unemployed", and "Unemployment_rate".
    - Delete US total rows to not affect results.
    - Removed Puerto Rico from the data set since there isn't any information for income available.
3. Home Price Index (HPI_master.csv AND HPI_master2.csv)
    - Deleted all data besides USA for place_id.
    - Kept a copy of the original dataset as it includes state data we will drill down later.
4. US Dollar Historical Data (US Data Index Historical Data.csv)
    - No changes made.

What does the final data set look like?

Home Price Index

```
## 'data.frame':    815 obs. of  10 variables:
##  $ hpi_type  : chr  "traditional" "traditional" "traditional" "traditional" ...
##  $ hpi_flavor: chr  "purchase-only" "purchase-only" "purchase-only" "purchase-only" ...
##  $ frequency : chr  "monthly" "monthly" "monthly" "monthly" ...
##  $ level     : chr  "USA or Census Division" "USA or Census Division" "USA or Census Division" "USA o
##  $ place_name: chr  "United States" "United States" "United States" "United States" ...
##  $ place_id  : chr  "USA" "USA" "USA" "USA" ...
##  $ yr        : int  1991 1991 1991 1991 1991 1991 1991 1991 1991 1991 ...
##  $ period    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ index_nsa : num  100 100 101 101 101 ...
##  $ index_sa  : num  100 100 100 100 100 ...
```

Unemployment Data

```
## 'data.frame':    1122 obs. of  9 variables:
##  $ Year                        : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ FIPS_code                   : int  1000 2000 4000 5000 6000 8000 9000 10000 11000 12000 ...
##  $ State                       : chr  "AL" "AK" "AZ" "AR" ...
##  $ Area_name                   : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ Civilian_labor_force        : chr  "2,147,173" "319,777" "2,510,611" "1,260,507" ...
##  $ Employed                    : chr  "2,047,731" "299,590" "2,410,825" "1,207,833" ...
##  $ Unemployed                  : chr  "99,442" "20,187" "99,786" "52,674" ...
##  $ Unemployment_rate           : num  4.6 6.3 4 4.2 4.9 2.7 2.1 3.6 5.7 3.8 ...
##  $ Median_Household_Income_2020: chr  "53,958" "79,961" "64,652" "51,146" ...
```

US Dollar Historical Data

```
## 'data.frame':    5000 obs. of  6 variables:
##  $ Date    : chr  "Jan 02, 2001" "Jan 03, 2001" "Jan 04, 2001" "Jan 05, 2001" ...
##  $ Price   : num  109 110 109 108 109 ...
##  $ Open    : num  109 109 110 109 108 ...
##  $ High    : num  110 110 110 109 109 ...
##  $ Low     : num  109 108 109 108 108 ...
##  $ Change..: chr  "-0.72%" "1.29%" "-1.23%" "-0.36%" ...
```

Minimum Wage Data

```
## 'data.frame':    2862 obs. of  15 variables:
##  $ Year                                              : int  1968 1968 1968 1968 1968 1968 1968 1968
##  $ State                                             : chr  "Alabama" "Alaska" "Arizona" "Arkansas"
##  $ State.Minimum.Wage                                : num  0 2.1 0.468 0.156 1.65 ...
##  $ State.Minimum.Wage.2020.Dollars                   : num  0 15.61 3.48 1.16 12.26 ...
##  $ Federal.Minimum.Wage                              : num  1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15
##  $ Federal.Minimum.Wage.2020.Dollars                 : num  8.55 8.55 8.55 8.55 8.55 8.55 8.55 8.55
##  $ Effective.Minimum.Wage                            : num  1.15 2.1 1.15 1.15 1.65 1.15 1.4 1.25 1
##  $ Effective.Minimum.Wage.2020.Dollars               : num  8.55 15.61 8.55 8.55 12.26 ...
##  $ CPI.Average                                       : num  34.8 34.8 34.8 34.8 34.8 34.8 34.8 34.8
##  $ Department.Of.Labor.Uncleaned.Data                : chr  "..." "2.1" "18.72 - 26.40/wk(b)" "1.25/
##  $ Department.Of.Labor.Cleaned.Low.Value             : num  0 2.1 0.468 0.156 1.65 ...
##  $ Department.Of.Labor.Cleaned.Low.Value.2020.Dollars : num  0 15.61 3.48 1.16 12.26 ...
##  $ Department.Of.Labor.Cleaned.High.Value            : num  0 2.1 0.66 0.156 1.65 ...
##  $ Department.Of.Labor.Cleaned.High.Value.2020.Dollars: num  0 15.61 4.91 1.16 12.26 ...
##  $ Footnote                                          : chr  "" "" "(b)" "(b)" ...
```

What information is not self-evident?

While the data do follow similar trends (i.e. the US dollar and Home Price Index), they fail to make a one to one connection with the data points in how they establish a connection; furthermore, they fail to provide a clean correlation where the data shares any dependency upon one another. Unemployment may play a larger role in the housing market issues previously but do not directly indicate a certainty.

What are different ways you could look at this data?

These data will need to be compared within the same charts to announce any key findings or missing connections. Would the reduction of the US dollar impact the value of homes overall? Questions about the US dollar need to be defined further as the dollar and home prices do have a direct relationship; without looking into the US dollar further, we may overlook that the value of the dollar going down can mean that the value of the home price would also reduce.

How do you plan to slice and dice the data?

Certain states have different reported numbers for unemployment, minimum wage, and their respective home price index. These states may contain data that has been overlooked as to why they may not have been impacted as harshly as others. The data is dependent on time, so the timelines will need to be trimmed within the code to match up appropriately with one another. These key considerations will be crucial to understanding their connection with one another.

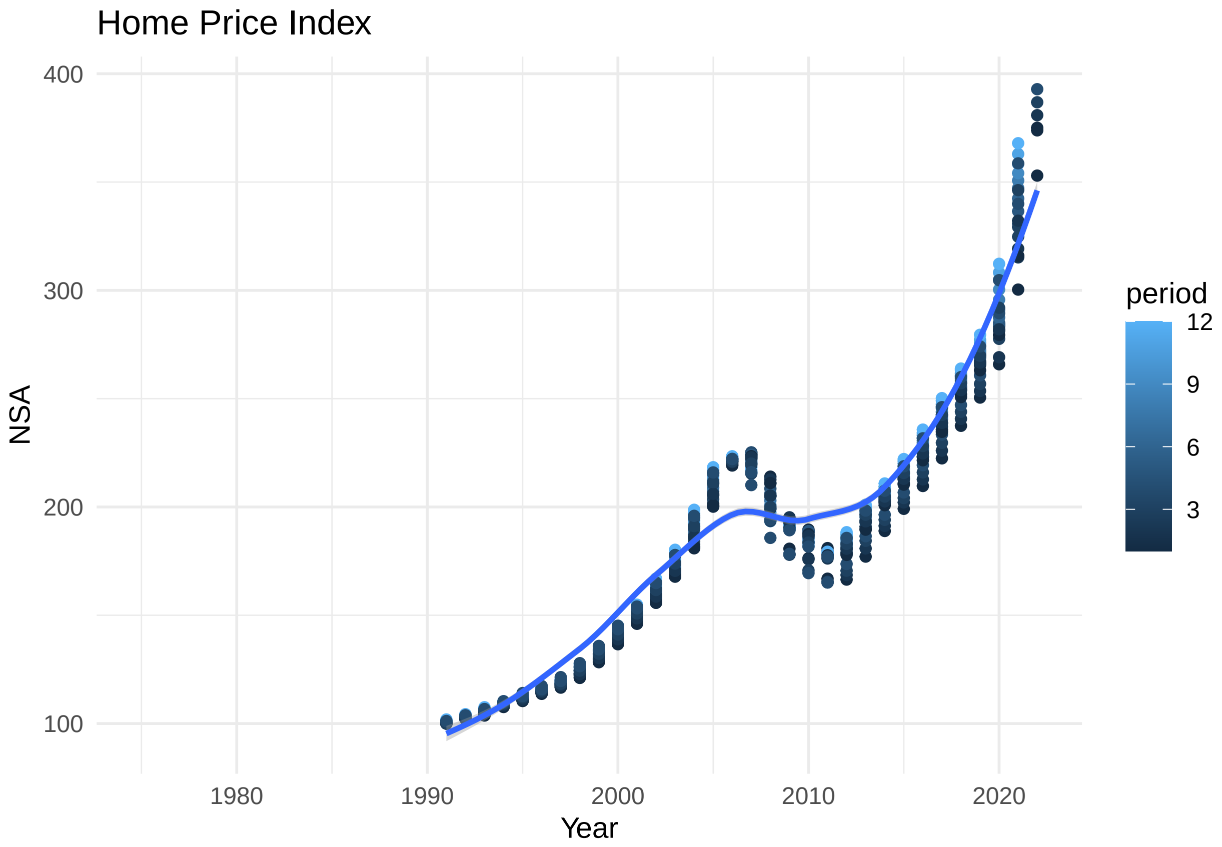
How could you summarize your data to answer key questions?

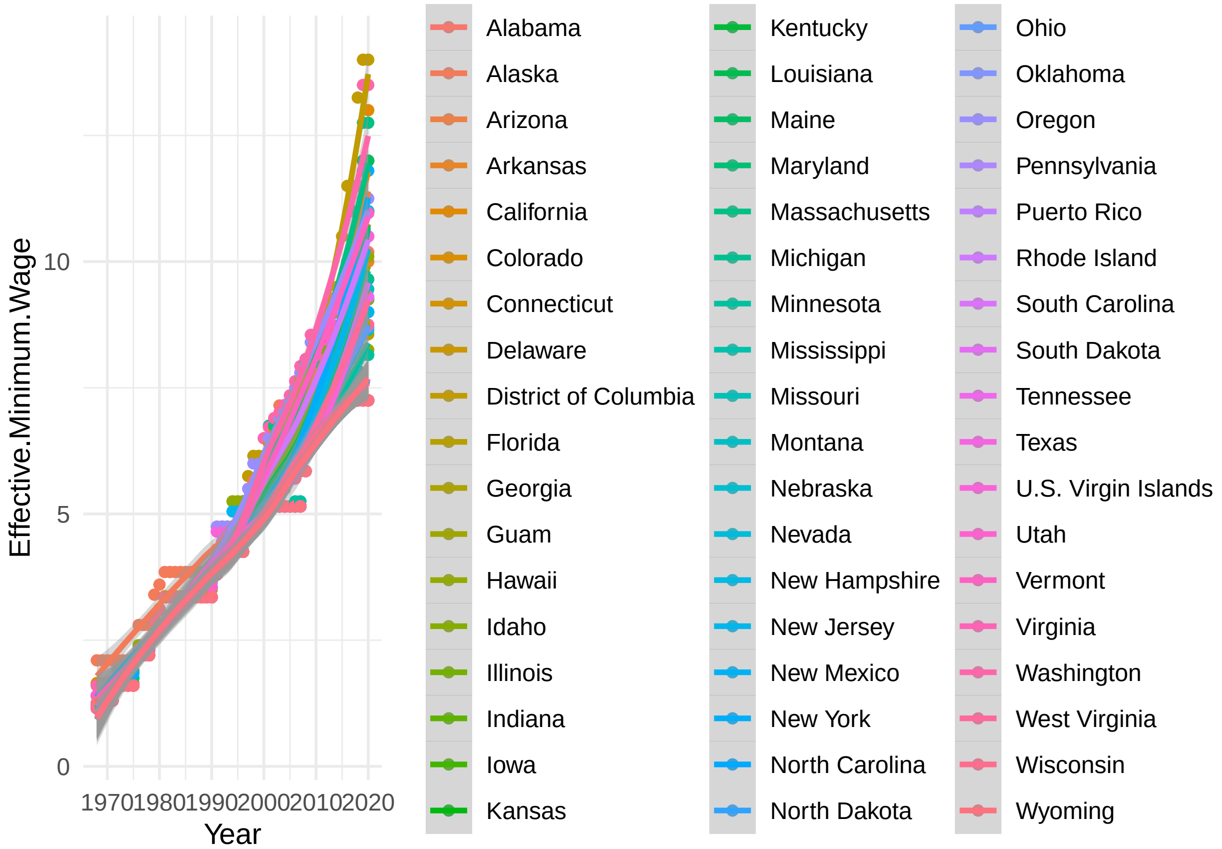
By comparing the data from the US dollar to the home price index, we can relate the impact taken on the US dollar and how that may effect the housing market soon. For example, the home price index is at all-time highs at the moment while the US dollar is drastically down. Will this be significant enough to enforce a downward trend? The data from the minimum wage dataframe, when compared with the unemployment and home price index may shed light on states that weren't impacted nearly as much; this can show us states that were less effected. Essentially, this data will be like a Sudoku puzzle where you solve problems with data you know you don't know.

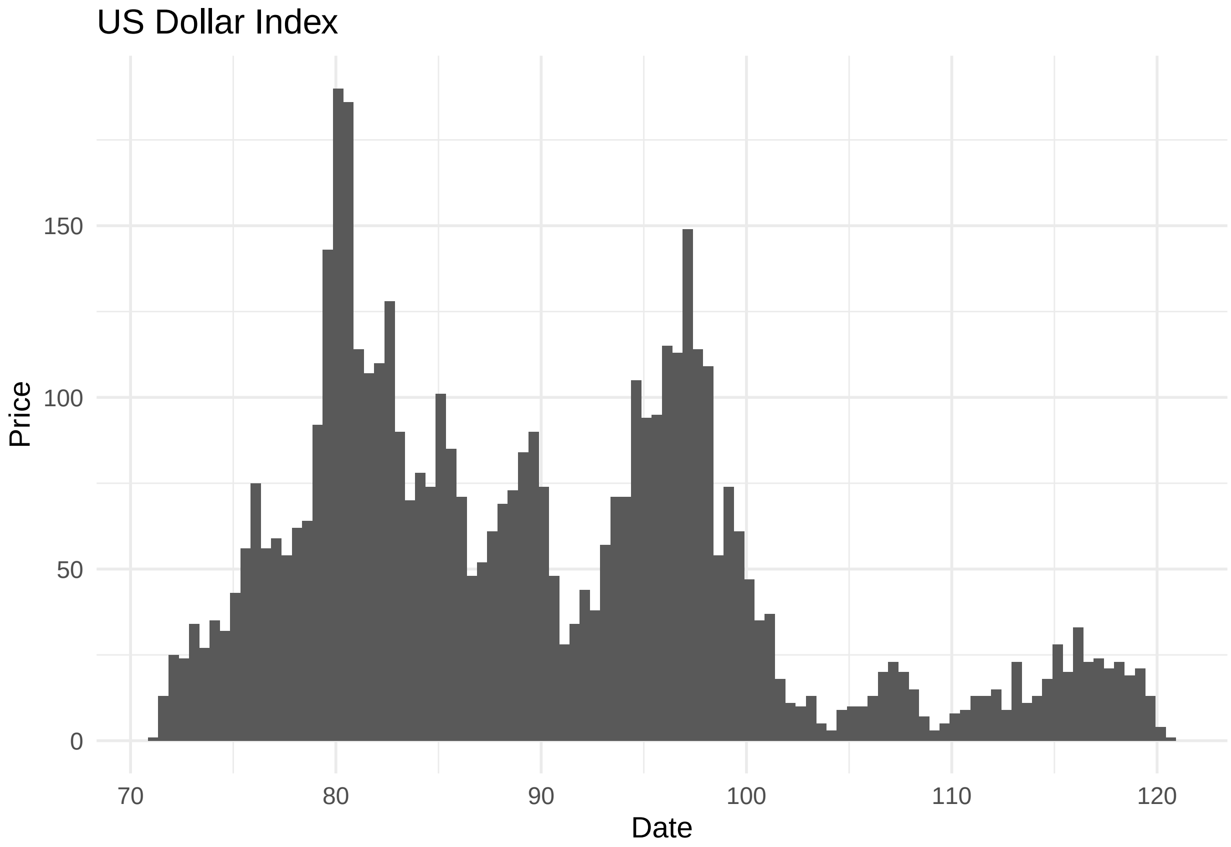
What types of plots and tables will help you illustrate findings?

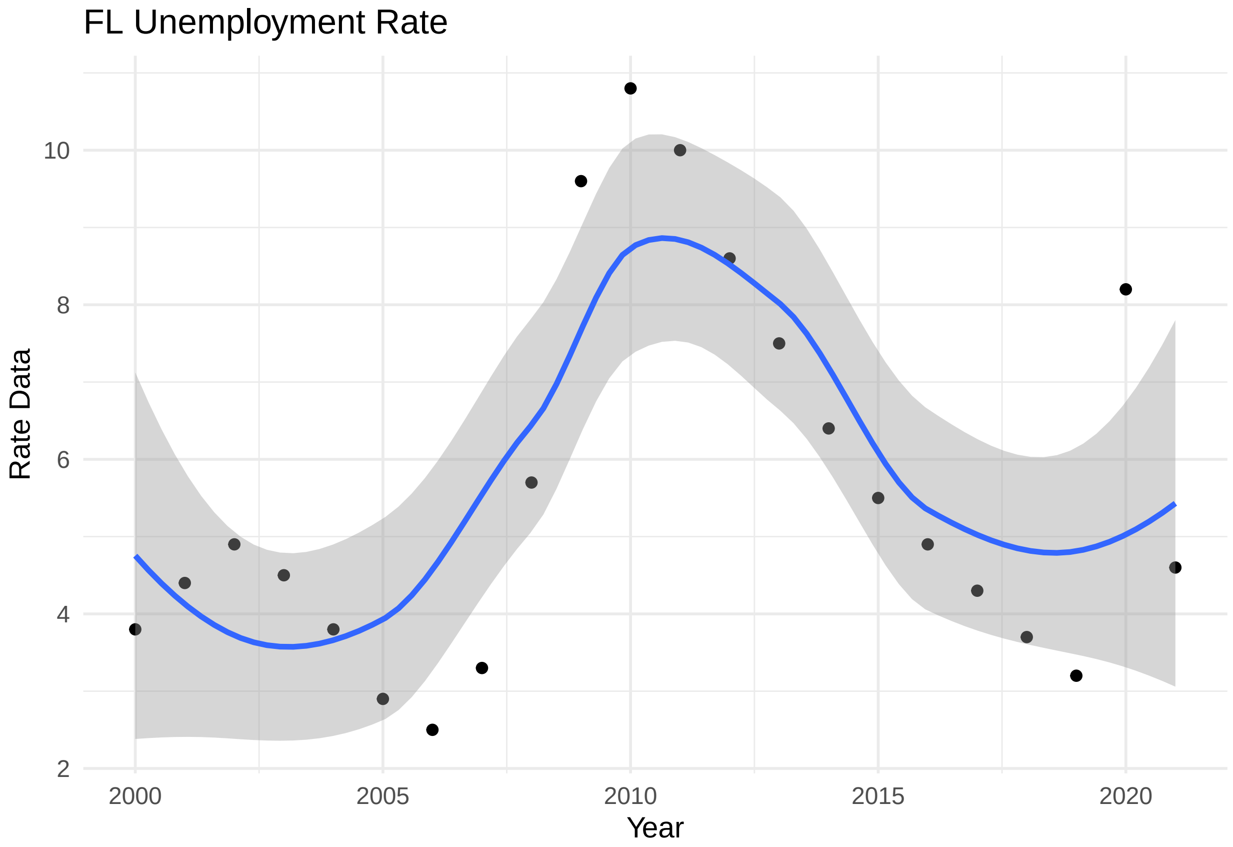
Histograms, scatterplots, and Bar plots are the main plots that will be used to display the data. These are used primarily for their visibility into the data as well as their visible trend lines to make connections with the datasets.

Examples of Plots






US Dollar Index
Price
150
100
50
0
70
80
90
100
110
120
Date



Do you plan on incorporating any machine learning techniques to answer questions?

This problem seems to have enough data surrounding it, so the answers to the data should be fairly easy to find. No machine learning techniques or algorithms will need to be used to surmise an appropriate conclusion.

Questions for future steps

```
+ Will more data need to be aggregated once other connections have been made?
+ Are there other dependent factors for the HPI that are not available to general public?
+ Have new banking regulations reduced the impact of the housing market bubble that happened in 2008?
```