

**Course: DATA522 Solving Big Data Problems**

Class Project

Date: Sep 17, 2021

Name: Wujian Xue

Student ID: 770031176

Content

[1.Introduction 3](#_Toc81517005)

[2.Data 4](#_Toc81517006)

[3.Methodology 8](#_Toc81517007)

[4.Results and further discussion 9](#_Toc81517008)

[5.Reference 9](#_Toc81517009)

[6.Appendix 1: R program 9](#_Toc81517010)

[7.Appendix 2: GitHub, GitHub Desktop, PowerShell 11](#_Toc81517011)

*“The land is the only thing in the world worth working for, worth fighting for, worth dying for, because it's the only thing that lasts."*

*--Gerald O'Hara, Gone with The Wind*

**FHFA HPI Time Series Analysis**

## 1.Introduction

For most people, the largest business in their life is to qualify mortgages from banks and buy houses. The Nobel price winner Robert Shiller states in his book “Irrational Exuberance” that people work for houses and the purpose of work is just to keep paying mortgages for 30 years. On average, the house price in the U.S. is around $300,000. With 20% down payment and 3% interest rate, the loan amout is about $240,000 and monthly payment is around $1,300 plus other fees.

Besides these costs, buying houses maybe the simplest and the most suceessful investment for many people. On average, the annual growth rate for U.S housing price is 3-5%. Although the growth rate itself doesn’t look exciting, considering the large principle, the profits from buying houses are still significant. For example, if a house worth $300,000, the growth rate is 5%, then 10 years later, the target house price will be $488,668. The $188,668 gain from house is much easier and stabler than profit from other kind of investment such as stocks or cryptocurrencies.

Unlike the stock market, which has one price for one stock in a centralized trading platform such as New York Stock Exchange(NYSE), the housing markets are composed of lots of local markets. In the other words, 3-bed room single family houses have different prices in New York, Washington D.C, Altanta or San Fransisco. Therefore, measuring housing market is different from measuring the stock market.

However, just like the stock index, many institutions design, calculate and publish House (or Home) Price Index (HPI) to measure housing prices. The most well-known HPI is S&P/Case-Shiller indexes, which is proposed by Karl Case and Robert Shiller and published by S&P Global Ratings. In addition, the U.S. government also creates HPI. For instance, Federal Housing Financing Agency(FHFA) publishes HPI using data from Fannie Mae and Freddie Mac since 1975.

In this project, I will analyze the HPI from FHFA, demonstrate HPI in differenct areas of US and forcast the future prices using time series model.

## 2.Data

The FHFA House Price Index is a method to measure the movement of single-family house prices. It is a weighted and repeat-sales HPI, which measures average price changes in repeat sales or refinancings on the same properties. This information is obtained by Fannie Mae or Freddie Mac since January 1975.

In this project, I will use FHFA House Price Indexes (HPIs) from Kaggle. It is located at https://www.kaggle.com/tunguz/fhfa-house-price-indexes-hpis. On FHFA website(https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index.aspx), there are more detailed HPI datasets.

Table 1 explains the columns in the dataset. Originally, the dataset has 11 variables and 114,372 observations. Among these variables, yr and period are integers. And index\_nsa (non-seasonal adjusted) and index\_sa (seasonal adjusted) are float. The rests are string. The examples of each variable are also listed in Table 1.

Table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Variable | Type | | Example |
| 1 | hpi\_type | String | Traditional, non-metro, distress-free, developmental | |
| 2 | hpi\_flavor | String | purchase-only, all-transactions | |
| 3 | frequency | String | Monthly, quarterly | |
| 4 | level | String | MSA, State, USA or Census Division, Puerto Rico | |
| 5 | place\_name | String | Washington-Arlington-Alexandria, DC-VA-MD-WV (MSAD)  New York-Jersey City-White Plains, NY-NJ (MSAD) | |
| 6 | place\_id | String | Zip code, state | |
| 7 | yr | Integer | 1975, 1976, …, 2020 | |
| 8 | period | Integer | 1,2,3,4,5….,12 | |
| 9 | index\_nsa | Float | 100, 100.98, 101.22, … | |
| 10 | index\_sa | Float | 100, 101.06, 100.95, …. | |

Basically, the dataset contains non-seasonal adjusted and seasonal adjusted HPI time series data from 1975 to 2020. The HPI datasets can by grouped by several ways such as by state, zip code or Metropolitan Statistical Area (MSA). The original point for both non-seasonal adjusted and seasonal adjusted HPI are 100. From Figure 1 to Figure 4, HPIs in 4 MSAs are shown:

* New York-Jersey City-White Plains, NY-NJ
* Washington-Arlington-Alexandria, DC-VA-MD-WV
* Charlotte-Concord-Gastonia, NC-SC
* Francisco-San Mateo-Redwood City, CA

Figure 1

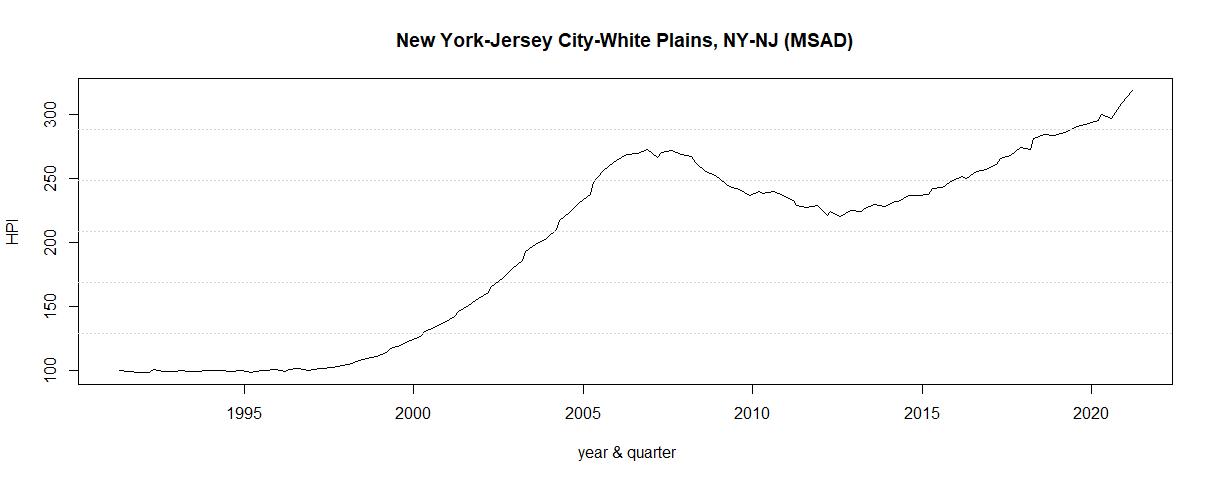


Figure 2

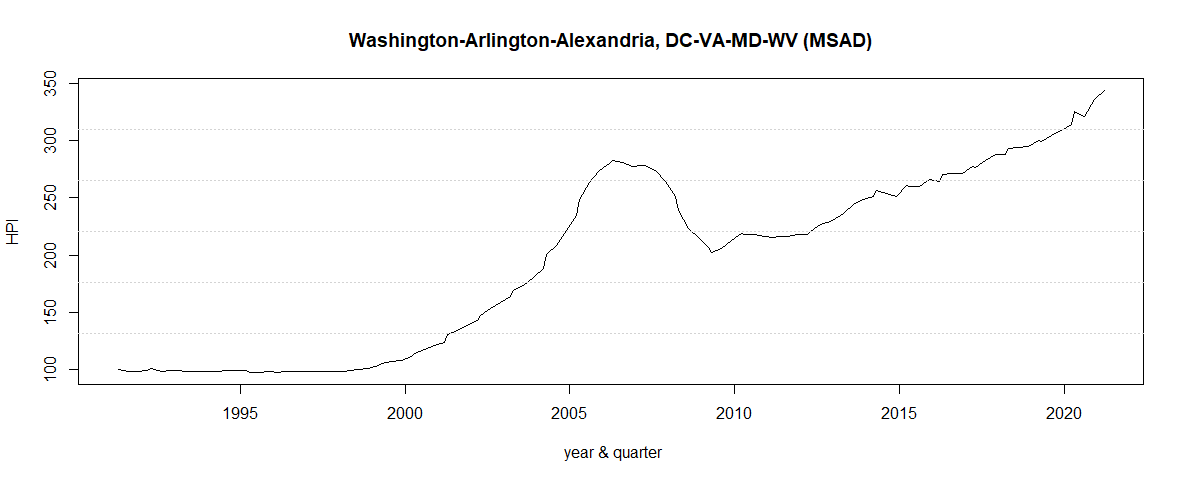


Figure 3

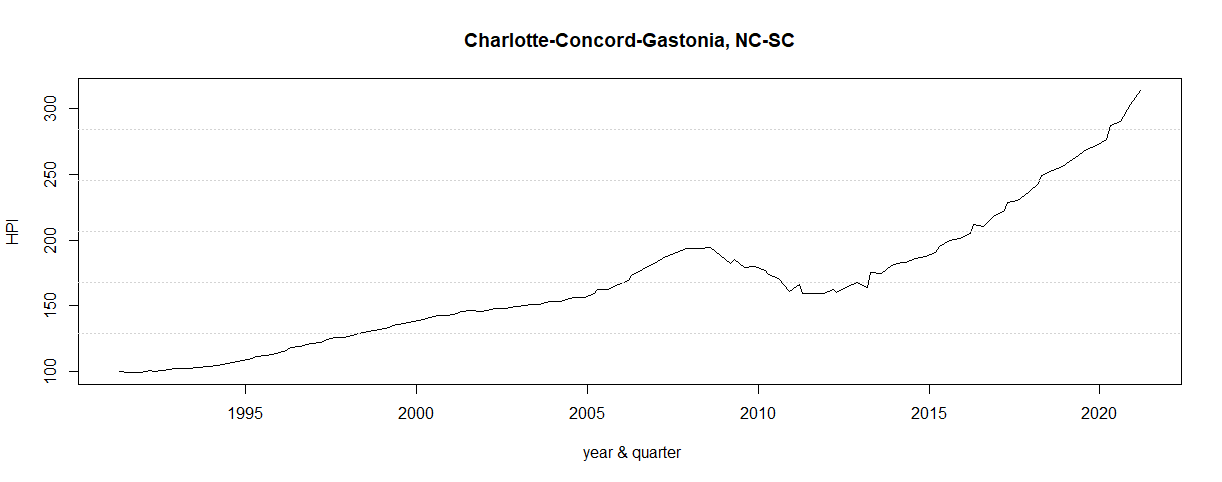
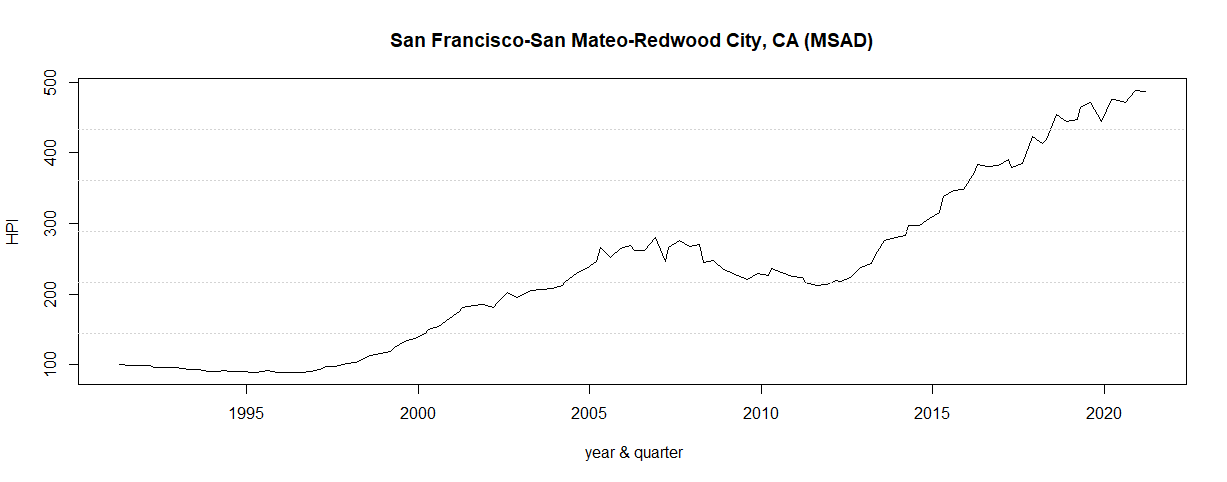
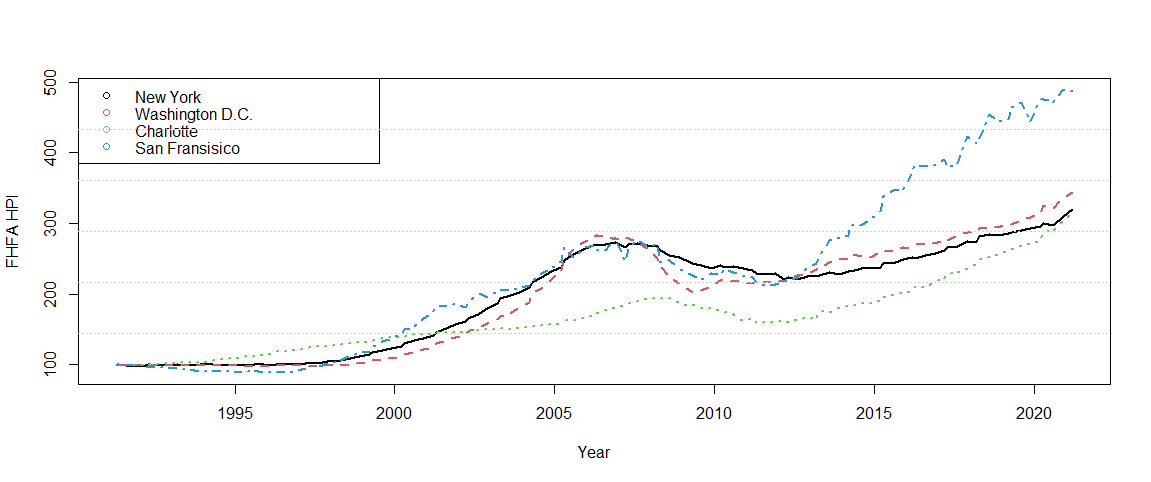


Figure 4



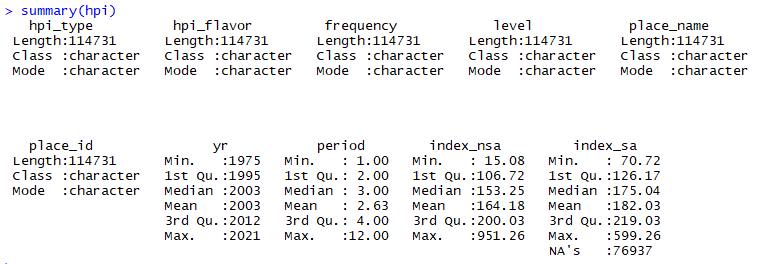
As you can see from these figures, the HPI increased in all the 4 MSAs from 1990 to 2020. And obviously, these 30 years can be divided into 3 periods: 1990-2000, 2001-2007 and 2008-2020. Before 2000, the HPIs are relatively stable in the 4 MSAs, although Charlotte has a little bit increasing trend. In the second period, the HPIs in all MSAs raised dramatically with 2-3 times increment. However, a new cycle started from 2008. The HPI dropped for several years and then create new high.

The HPIs for the 3 MSAs in the east coast have similar pattern. Washington D.C metropolitan area has the highest seasonal adjusted HPI at 344.05, followed by New York at 319.40 and Charlotte at 313.82. However, when it comes to the west coast, San Francisco has an HPI at 486.75. This is significantly higher than those 3 MSAs in the east coast, as shown in Figure 5.

Figure 5

Besides this visualization, I also explore the data by check missing data, computing descriptive statistics, and other methods.



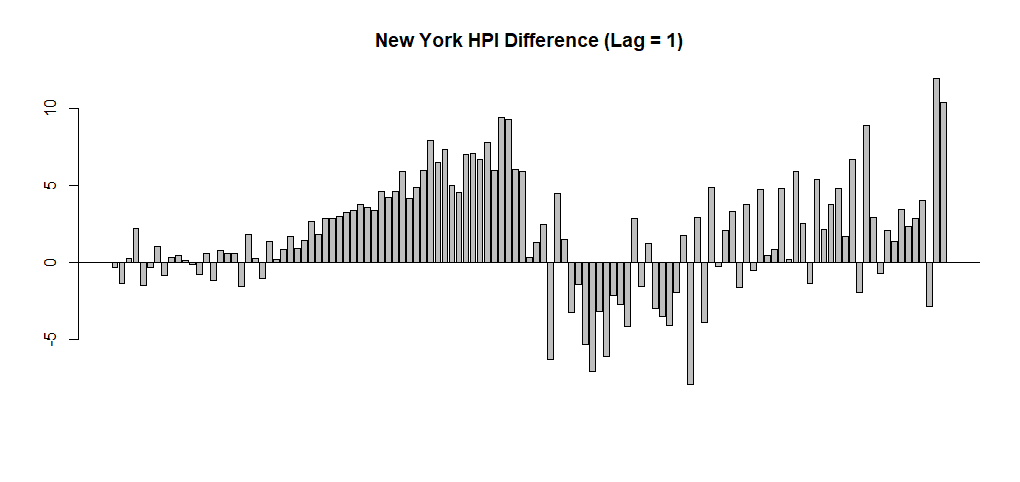


## 3.Methodology

In this section, I use HPI in New York as a time series example and apply ARMA model to identify and describe its structure and forecast future HPI values.

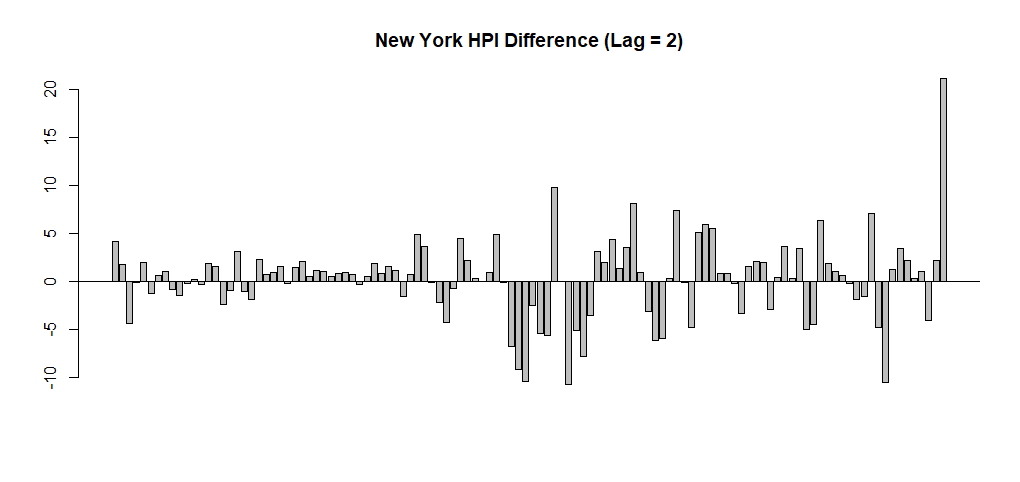
To apply an ARMA model, the dataset needs to be stationary time series. Using diff() function in R, the New York HPI time series is differenced once as shown in Figure 6. But it doesn’t appear to be a stationary process.

Figure 6



By changing lag to 2 and difference to 2 in the diff() function, we can see this issue can be solved. As shown in Figure 7, the differenced time series has a constant mean closed to zero and a fairly constant variance over time. Therefore, a stationary time series has been obtained.

Figure 7



Then, the ACF and PACF for the differenced series are plotted in Figure 8 and Figure 9. As we can see from these figures, AR(1) and MA(1) can be considered.

Figure 8

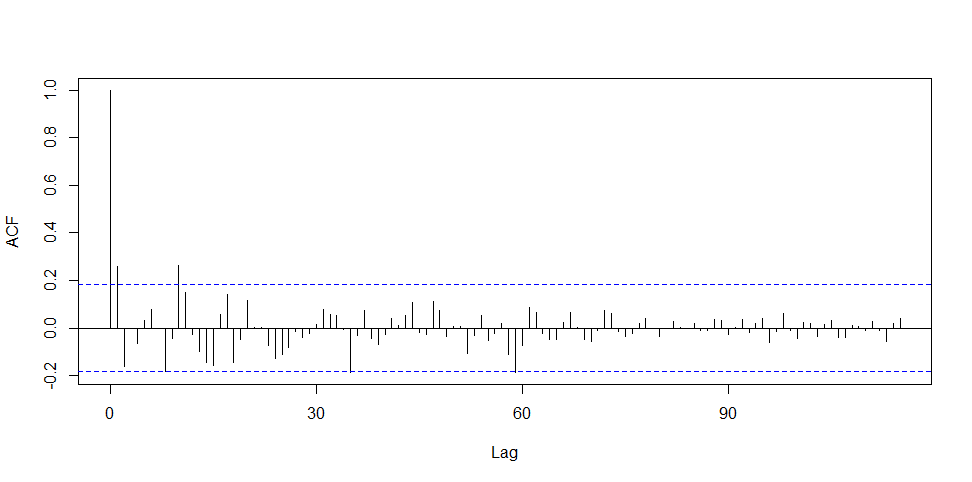
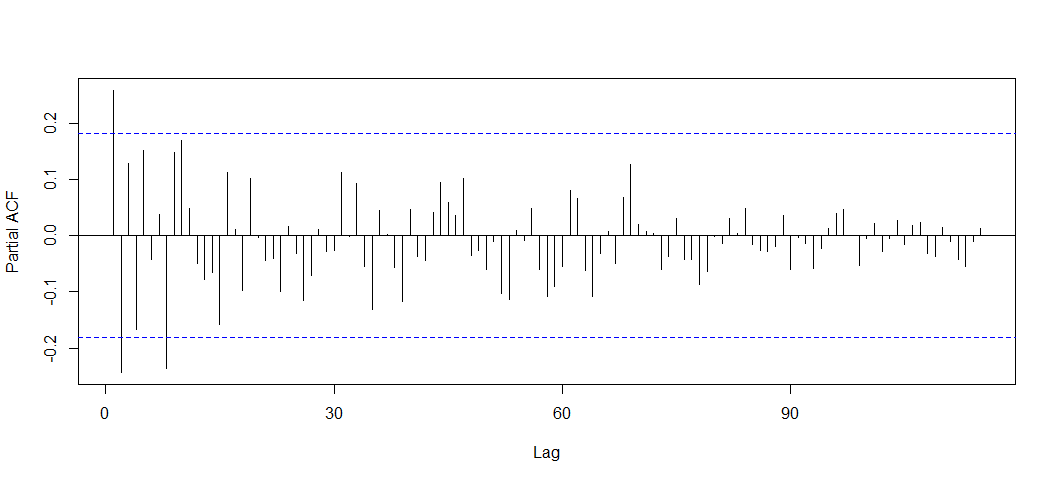


Figure 9



Using arima() function in R to fit this model, parameters can be estimated as follow.

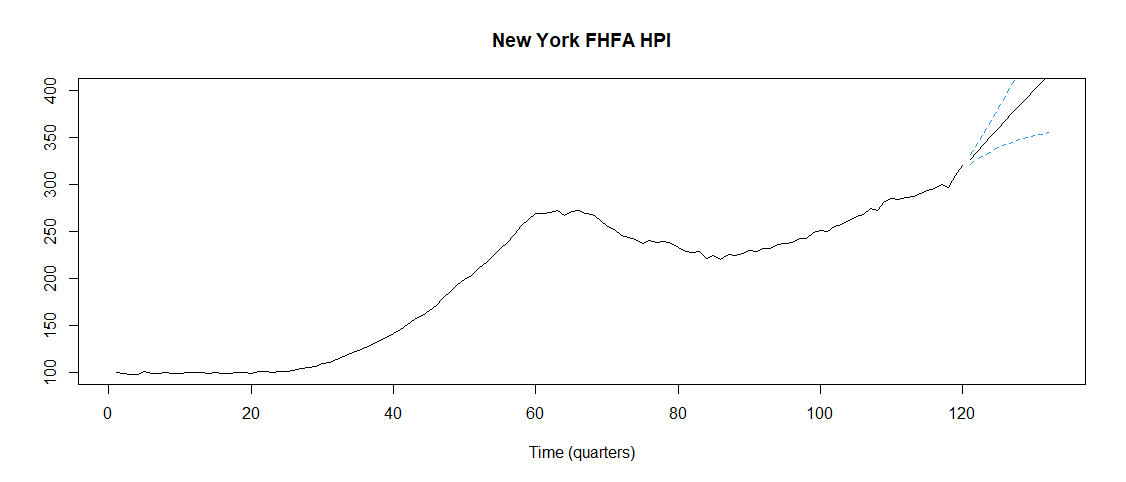


The next step is to forecast the future New York FHFA HPI. This can be obtained using the R predict() function. The predicted values together with lower and upper bounds at 95% confidence intervals are shown in Table 2 and Figure 10. The real New York HPI values in 2021 Q1 and Q2 announced by FHFA are also shown in Table 2.

Table 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Low Bound | Predict Values | FHFA Values | Upper Bound |  |
| 2021Q1 | 321.12 | 326.56 | 329.27 | 332.00 |  |
| 2021Q2 | 326.93 | 335.12 | 341.43 | 343.31 |  |
| 2021Q3 | 330.98 | 343.07 | - | 355.16 |  |
| 2021Q4 | 335.11 | 351.29 | - | 367.46 |  |
| 2022Q1 | 338.64 | 359.39 | - | 380.14 |  |
| 2022Q2 | 341.89 | 367.54 | - | 393.19 |  |
| 2022Q3 | 344.77 | 375.67 | - | 406.57 |  |
| 2022Q4 | 347.34 | 383.81 | - | 420.28 |  |
| 2023Q1 | 349.62 | 391.94 | - | 434.27 |  |
| 2023Q2 | 351.61 | 400.08 | - | 448.56 |  |
| 2023Q3 | 353.32 | 408.22 | - | 463.11 |  |
| 2023Q4 | 354.78 | 416.35 | - | 477.93 |  |

Figure 10



Obviously, the upward trending is similar to the current housing market. The real New York HPI values in 2021 Q1 and Q2 are 329.27 and 341.43 respectively. They are closed to but different from the predicted values, although both of them are in the 95% confidence interval.

One reason results in these differences is that the data on Kaggle is dated since these data are seasonal adjusted. To solve this issue, I can use the non-seasonal adjusted data instead of seasonal adjusted HPI. Another method to solve the issue is to change the data source to FHFA, as it updates the data frequently.

## 5.Reference

* Aderibigbe, T. & Chi, H. (2018). Investigation of Florida Housing Prices using Predictive Time Series Model, Proceedings of the Practice and Experience on Advanced Research Computing, July 2018, Article No. 92, Pages 1–4
* Smola, A. & Vishwanathan, S.V.N. (2008). Introduction to Machine Learning, Cambridge Press.
* Shiller, R., (2015). Irrational exuberance, 3rd edition, Princeton University Press.
* Chacon, S. & Straub, B. (2014). Pro Git, <https://git-scm.com/book/en/v2>

## 6.Appendix 1: R program

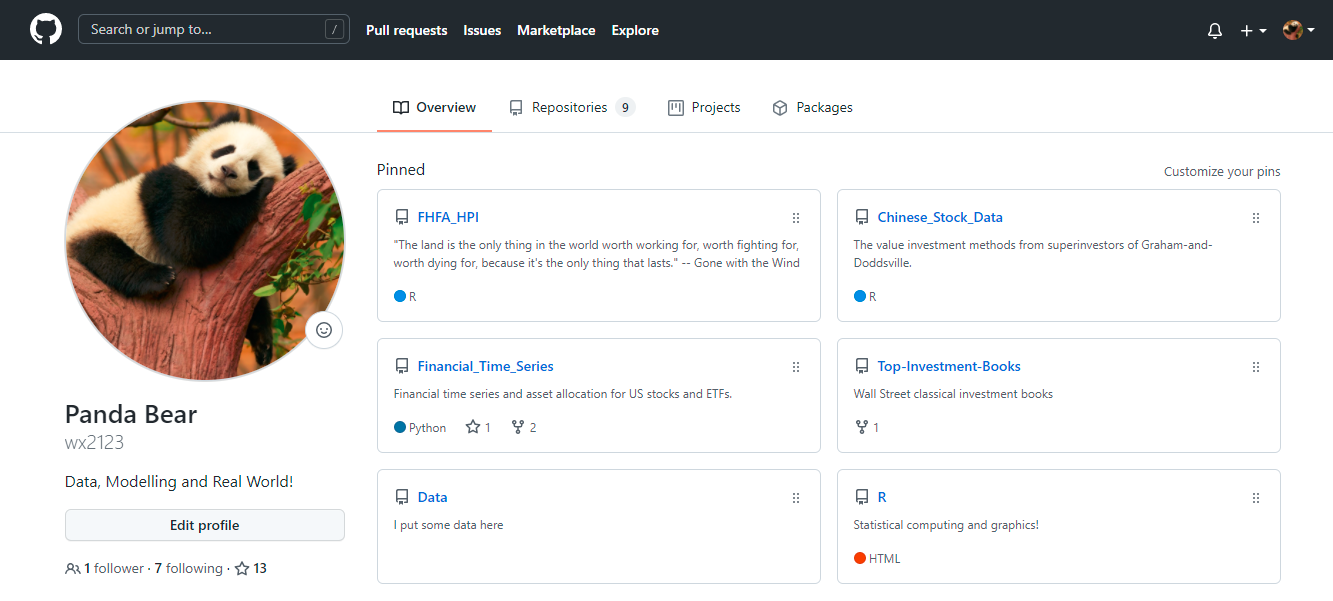
R program for this project is located at

<https://github.com/wx2123/FHFA_HPI/blob/master/D522_project.R>

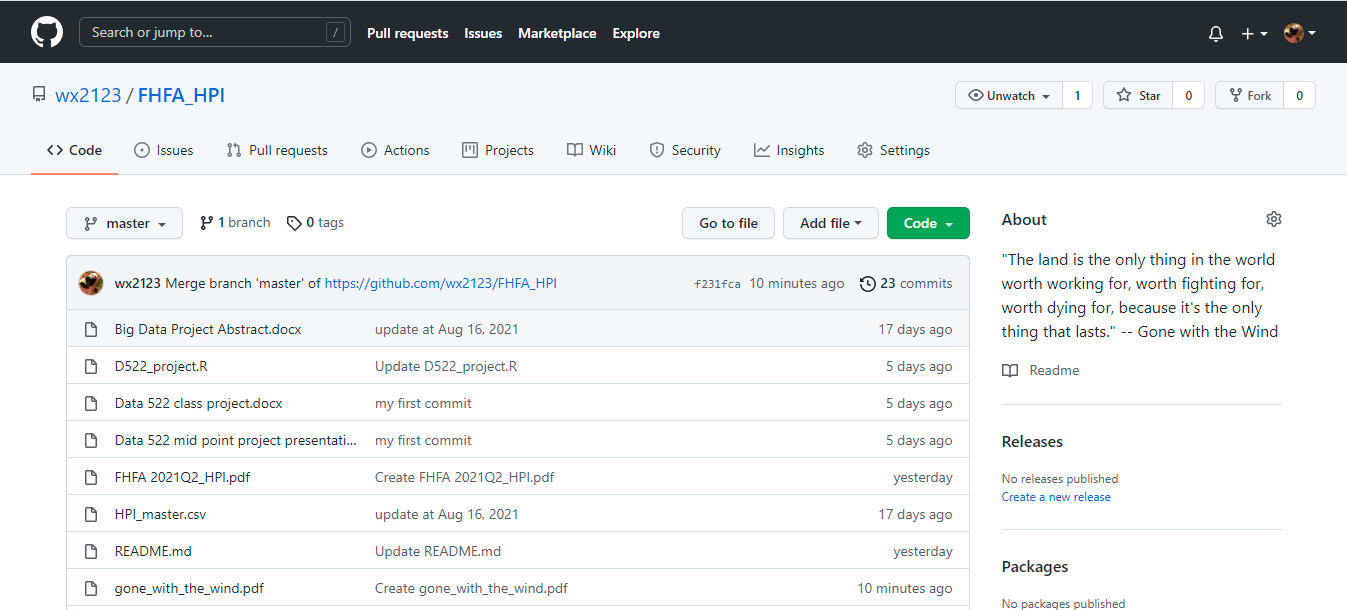
## 7.Appendix 2: GitHub, GitHub Desktop, PowerShell

To complete this project, I used Git tools such as GitHub, GitHub Desktop and PowerShell.

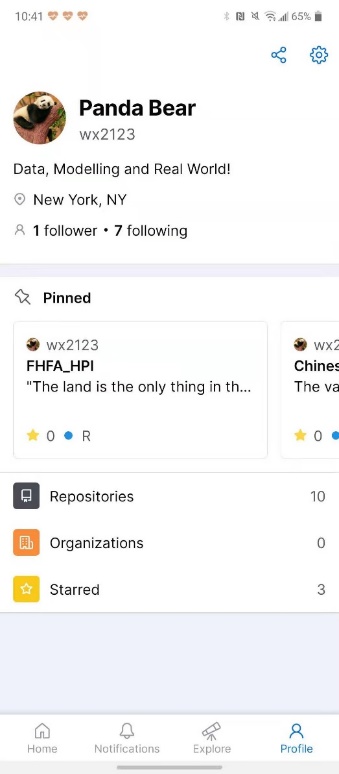
* My GitHub: https://github.com/wx2123



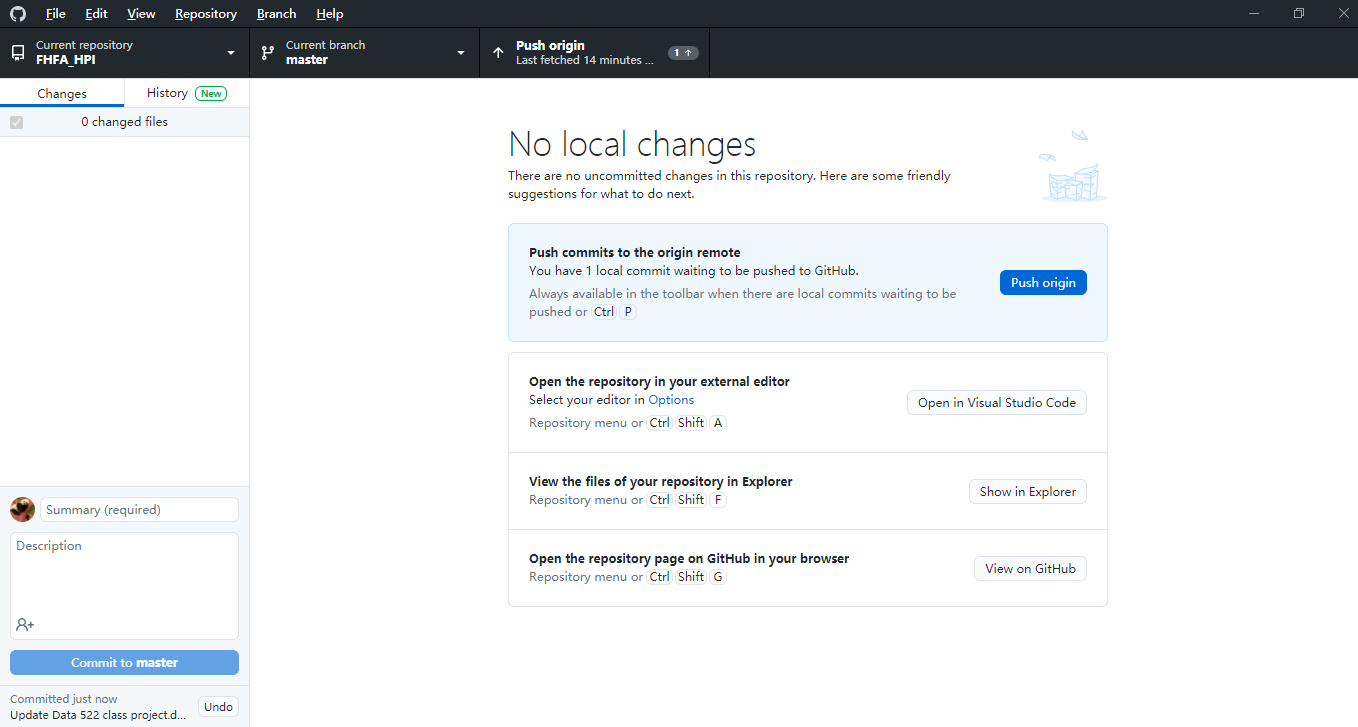
* Page for this project: https://github.com/wx2123/FHFA\_HPI

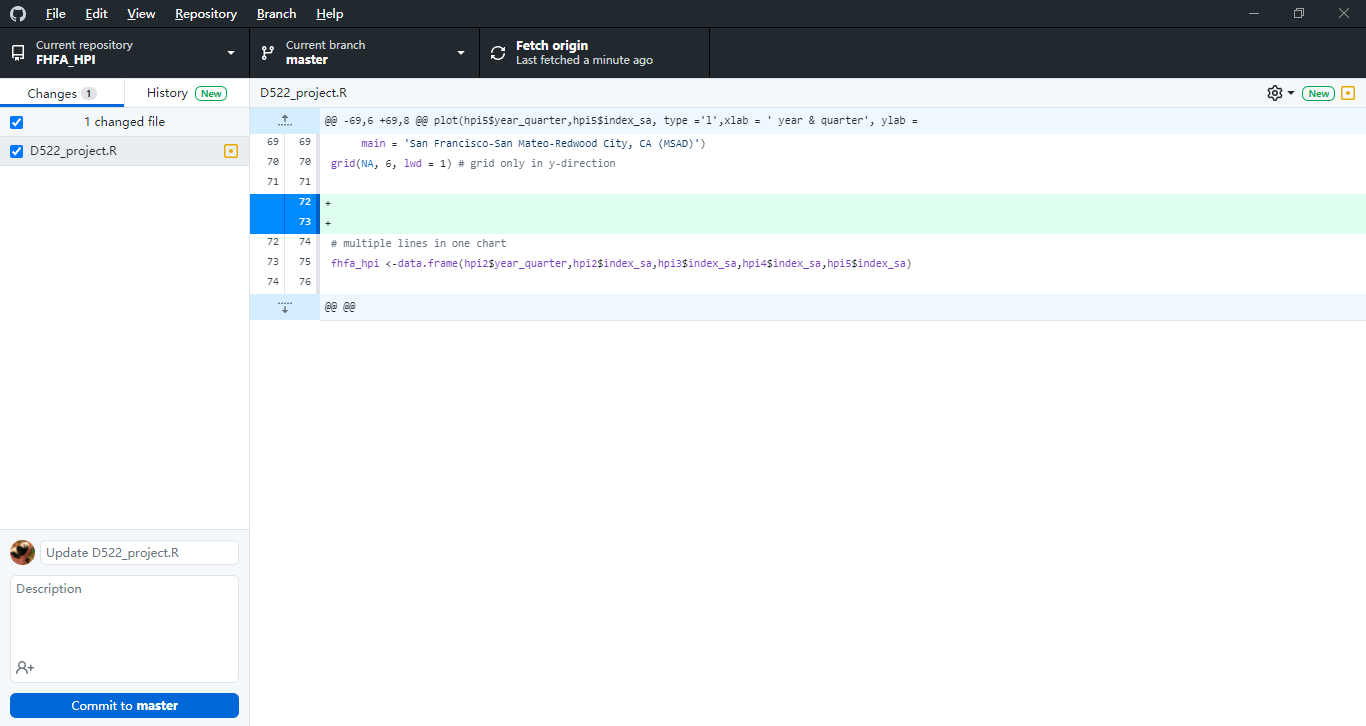


* GitHub APP



* GitHub Desktop





* Git functions in PowerShell

