Housing Market in the U.S.

PRESENTED BY:

SUNMIN LEE, THIDAR SWE TIN, WEN YIN

Project Description

- **There may be some factors that affect the housing prices in the US.**
- **This project will examine correlations between the various factors and their impact on the housing prices in New York.**

Research Questions to Answer

What is the trend in house prices over 25 years (1991 – 2016)?

Are there any correlation between US house prices (1991-2016) and

- **Household income**
- **GDP**
- **Interest rates**

Track the correlation of New York house prices for 8 years (2009-2016) with

- **Household income**
- **Weather**
- **Crime rate**

Predict house prices in New York for the 8-year period (2009 – 2016)

Compare the model with actual house prices and analyze the impact of each factor

US House Prices

(1991 – 2016)

US House Prices

Data source: CSV

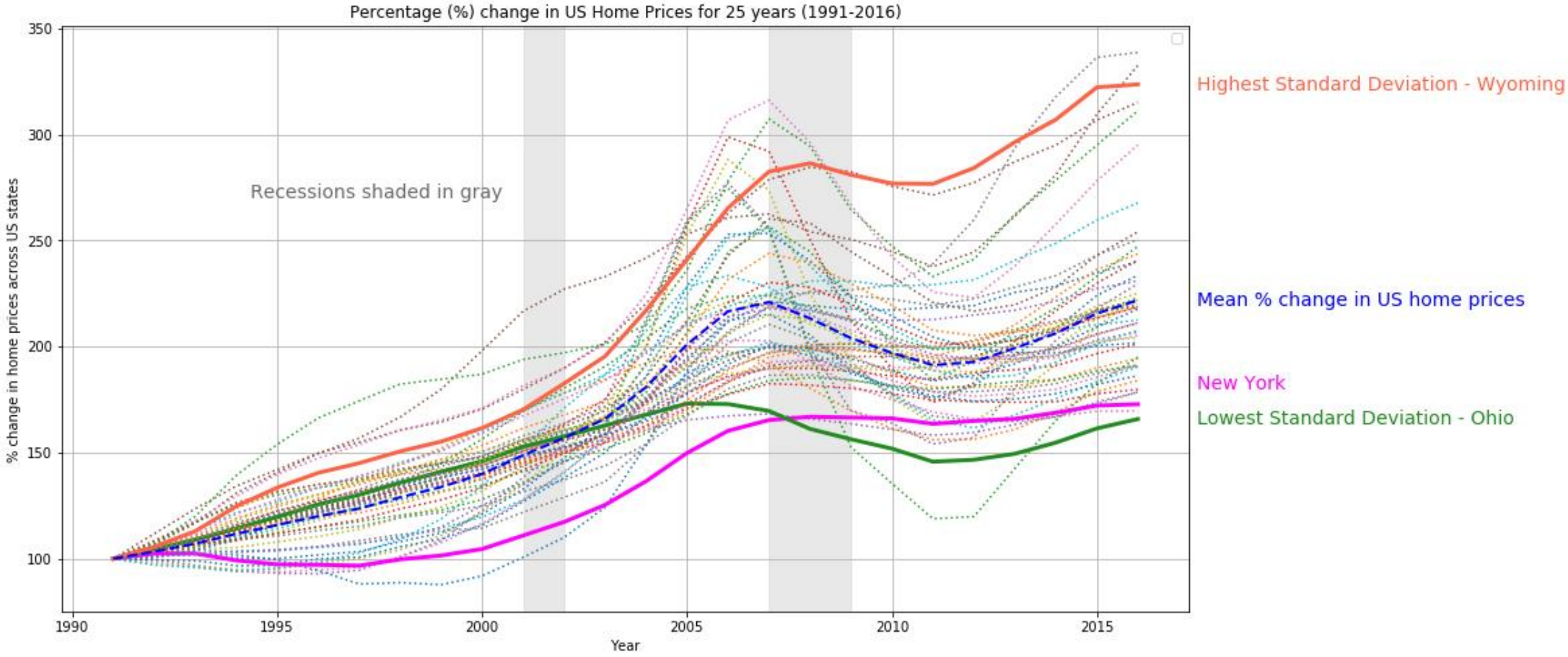
Objective: Analyze the trend in US house prices across all states over 25 years (1991 – 2016)

Data manipulation:

1. Normalize all house prices in all US states to start from 100 in 1991 to analyze the trend
2. Highlight the states with highest and lowest standard deviation, New York and compare them with the mean house prices over the years

Line plot: matplotlib

US House Prices



US House Prices

Observations:

1. House prices have doubled over the 25-year period
2. Wyoming is the state with the highest standard deviation in house prices – most change
3. Ohio is the state with the lowest standard deviation in house prices – least change
4. Great Recession from Dec 2007 - June 2009
   - The subprime mortgage crisis led to the collapse of the United States housing bubble
   - House prices in most states went down.
   - New York is an exception; the house prices remained stable

US Household Income

(1991 – 2016)

US Household Income

Data source: CSV and Census API

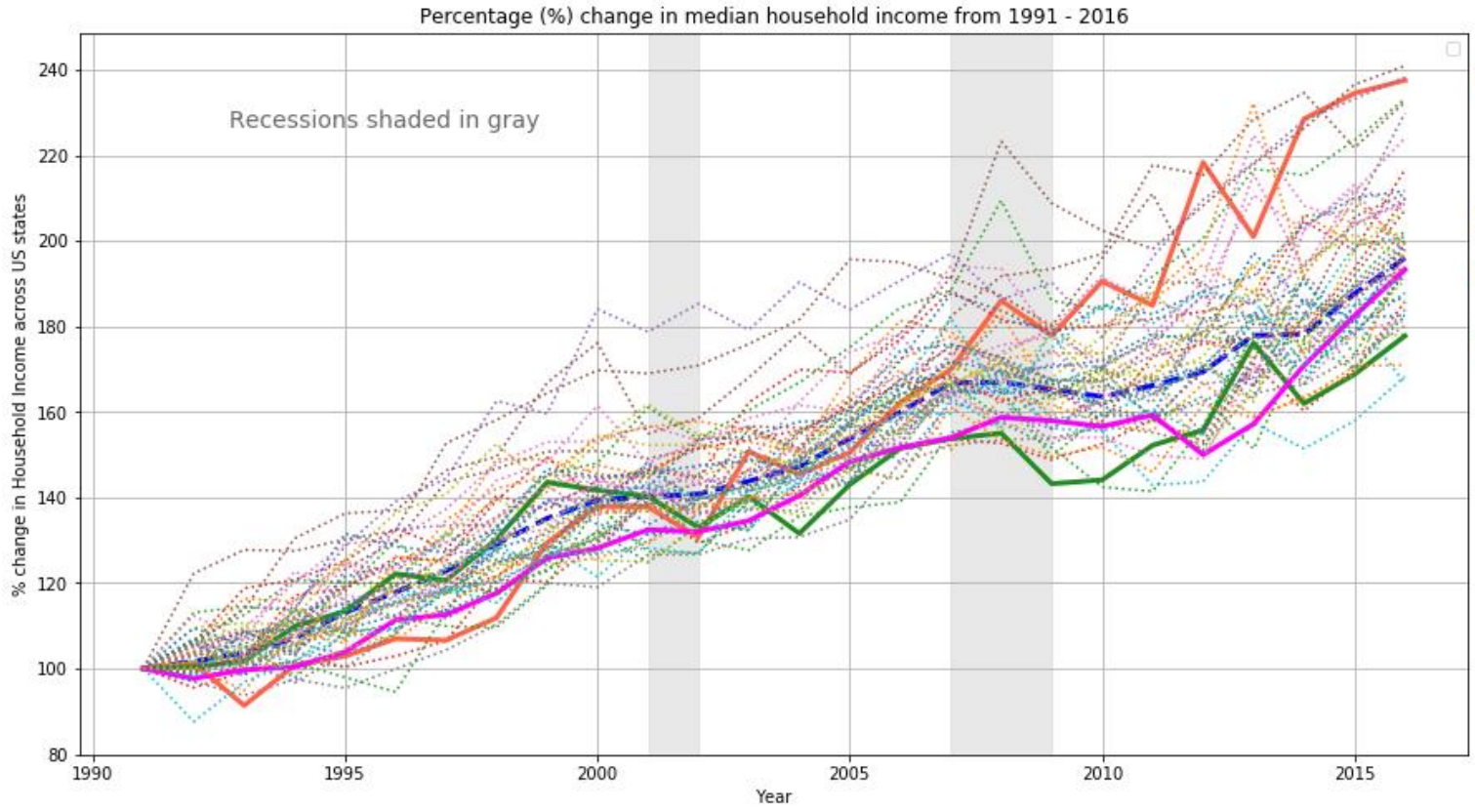
Objective: Analyze the trend in US household income across all states over 25 years (1991 – 2016)

Data manipulation:

1. Normalize all median household income in all US states to start from 100 in 1991 to analyze the trend
2. Highlight the states with highest and lowest standard deviation, New York and compare them with the mean household income over the years

Line plot: matplotlib

US Household Income



Highest Standard Deviation - District of Columbia

Mean % change in US household incomes

New York

Lowest Standard Deviation - Michigan

US Household Income

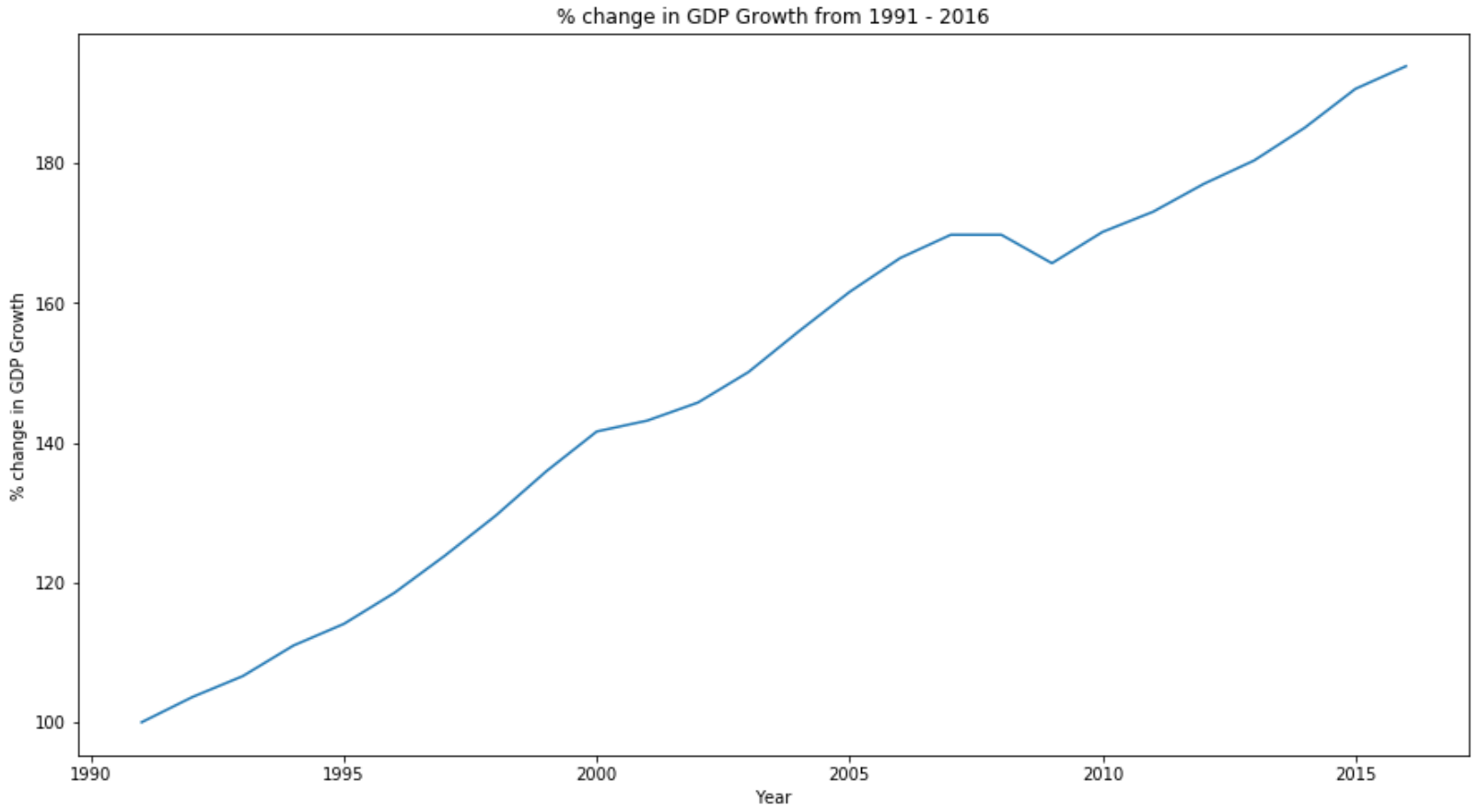
Observations:

1. House prices have doubled over the 25-year period
2. District of Columbia is the state with the highest standard deviation in household income – most change
3. Michigan is the state with the lowest standard deviation in house prices – least change
4. Great Recession from Dec 2007 - June 2009
   - Household income in most states went down.
   - New York has a delayed impact; the household income dropped a couple of years later in 2013

US GDP & Interest Rates

(1991 – 2016)

GDP



Data source: CSV

Objective: Analyze the trend in US GDP over 25 years

Data manipulation:

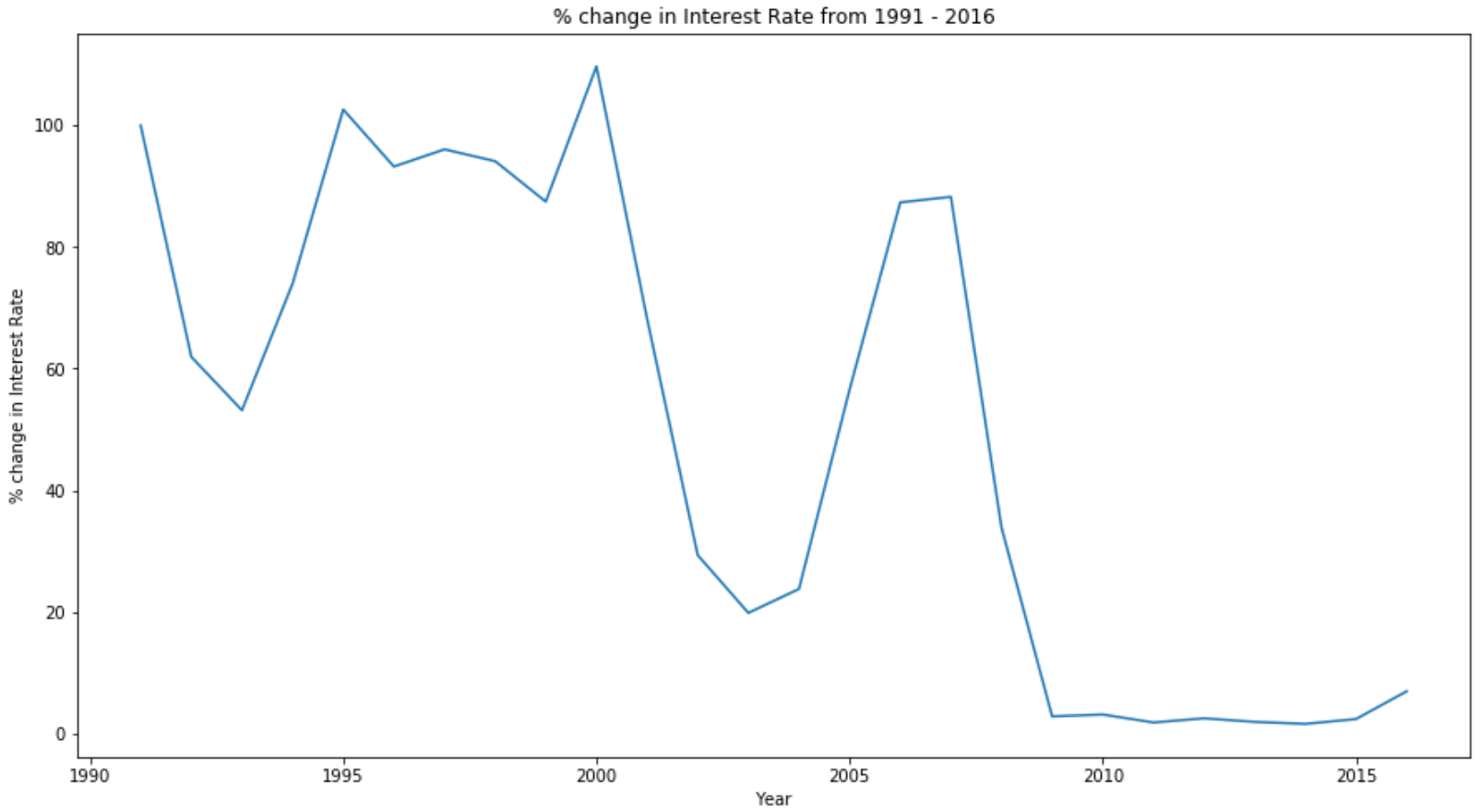
1. Convert GDP growth rate from year after year to start from 100 in 1991 to analyze the trend

Line plot: matplotlib

Observations:

1. GDP has grown linearly except for a slight dip after the great recession

Interest Rates



Data source: CSV

Objective: Analyze the trend in interest rates over 25 years

Data manipulation:

1. Normalize interest rates to start from 100 in 1991 to analyze the trend

Line plot: matplotlib

Observations:

1. Interest rates dipped significantly after both recessions and remained stable until 2015

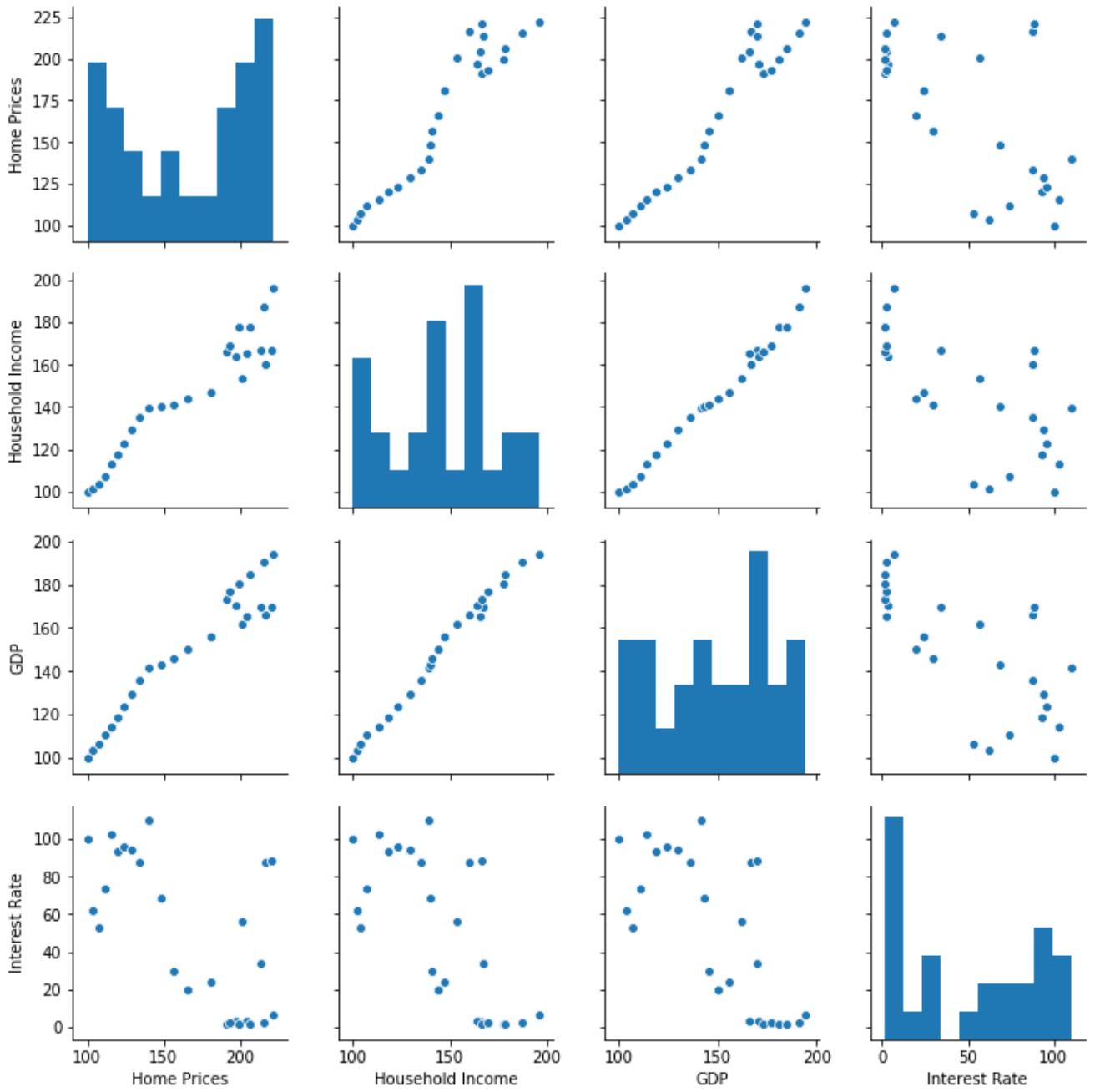
Analysis - US

(1991 – 2016)

Analysis

Pairplot: Seaborn

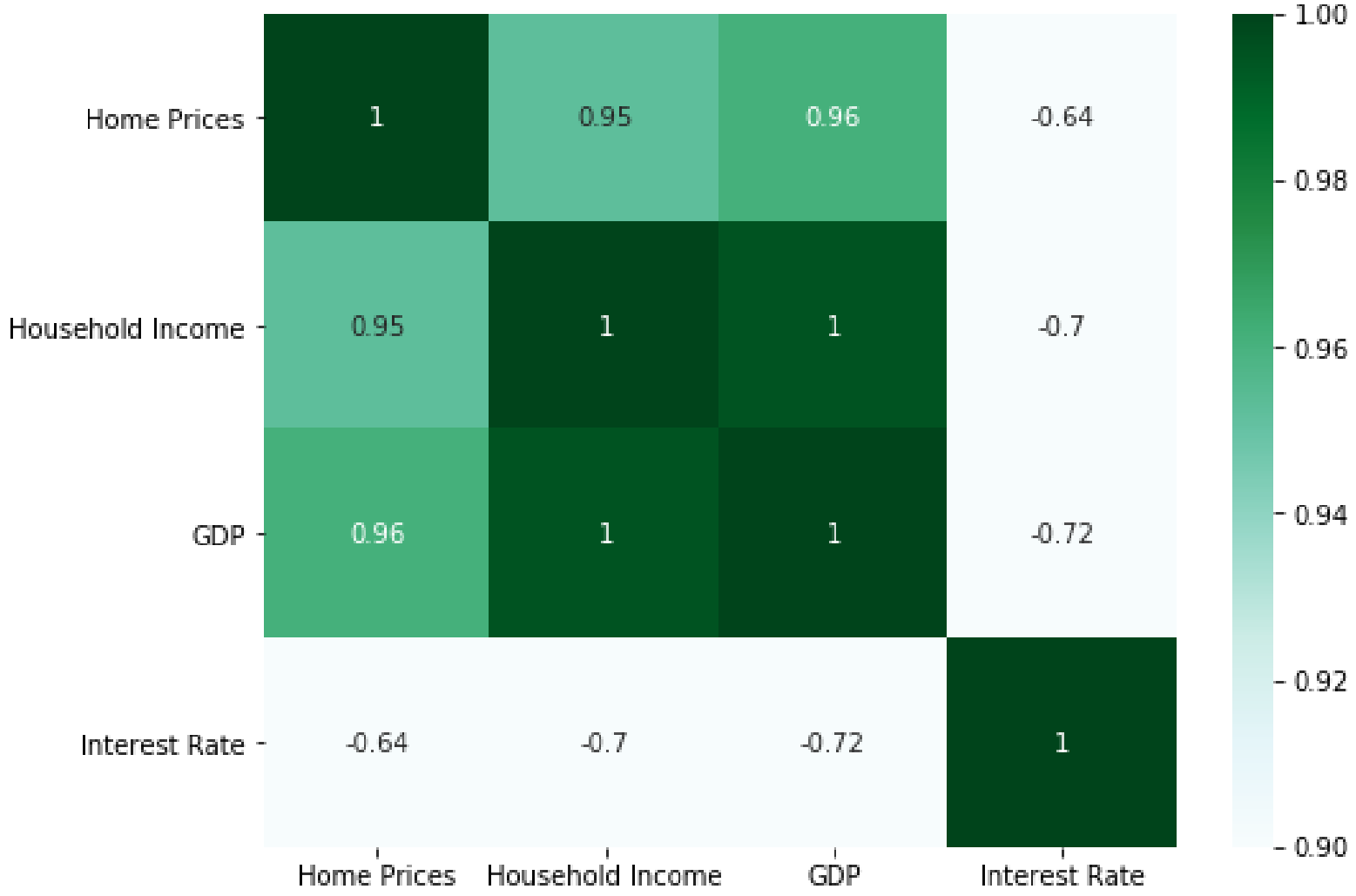
1. House Prices
2. Household Income
3. GDP
4. Interest Rate



Analysis

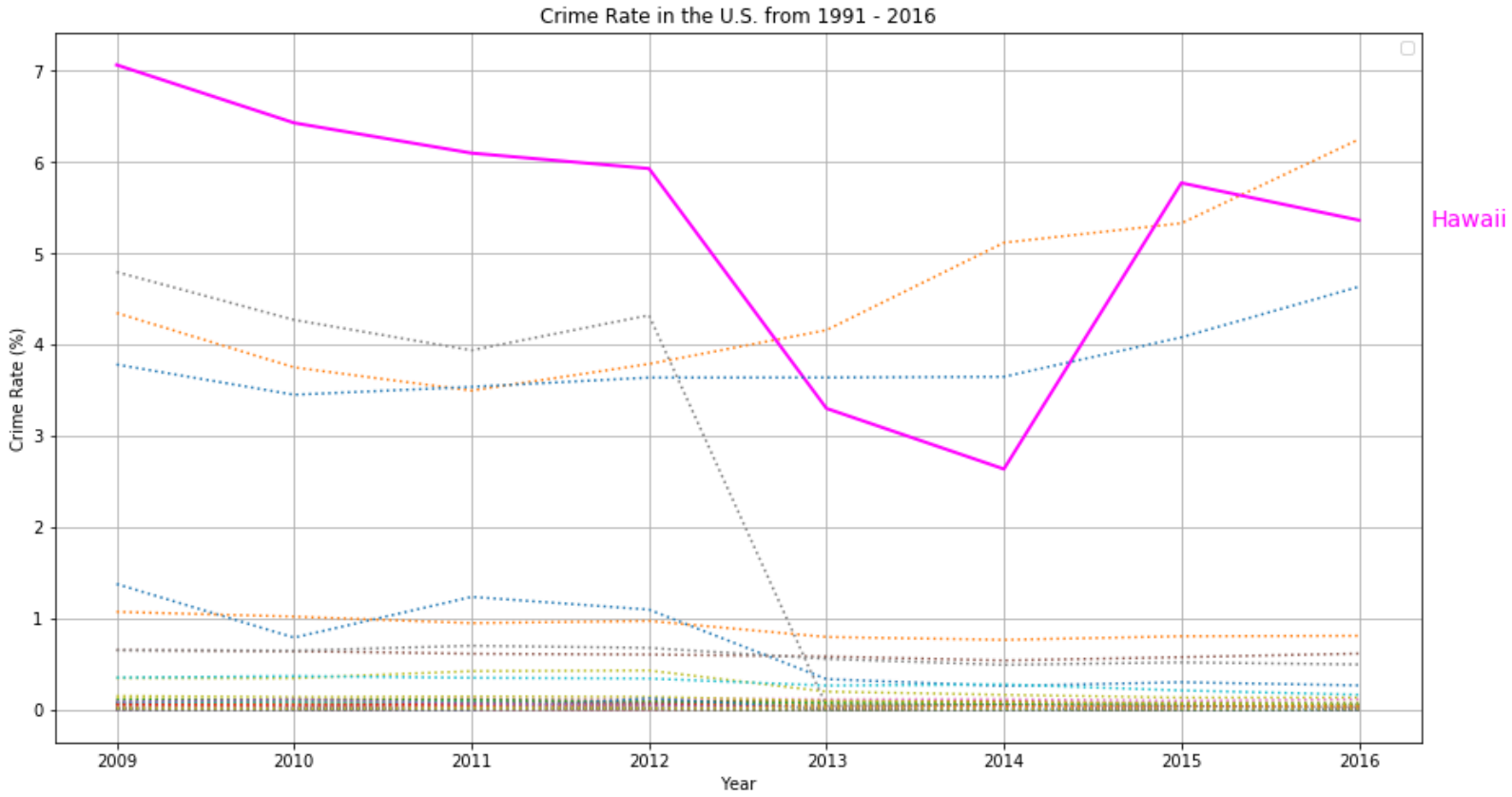
Heatmap: Seaborn

1. House Prices
2. Household Income
3. GDP
4. Interest Rate



US Crime Rate

(2009 - 2016)



Crime Rate

1. Data source: API from FBI
2. Data retrieval: Convert into JSON and export in CSV file
3. Data manipulation: Merge with "state" and "population" dataframe and clean data
4. Data wrangling: Calculate crime rate and include in the dataframe
5. Line plot: matplotlib

Crime Rate 2016 - Heat Map

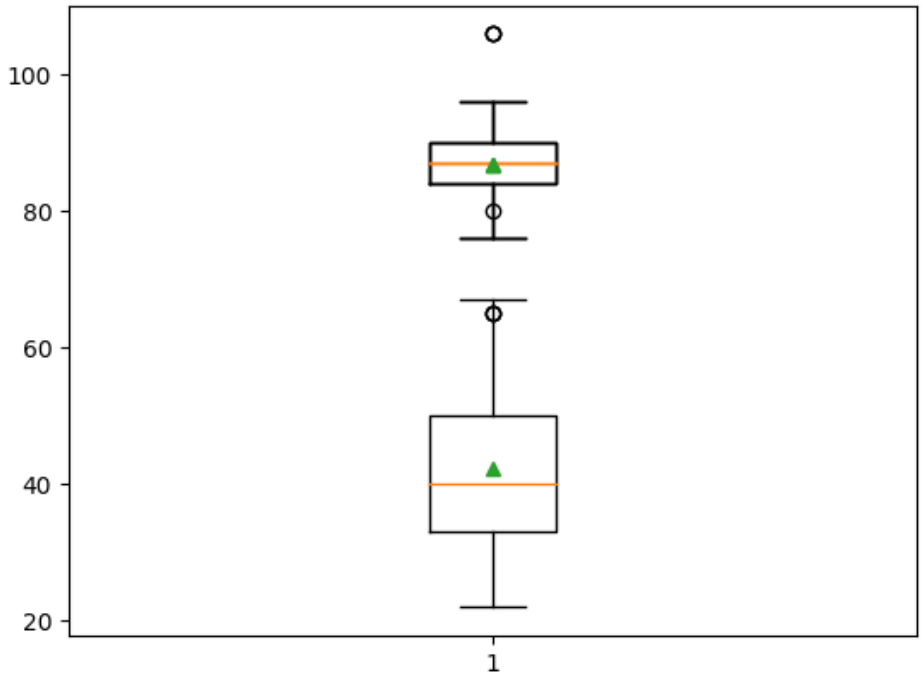
US Weather

(1999 – 2017)

US Weather

- **Data source:** CSV
- **Objective:** Analyze the trend in US TEMPERATURE across all states around 20 years (1999 – 2017)
- **Data manipulation:**
  1. Computerize the temperature range for each state
  2. Calculate the correlation between temperature and housing price in some cold-winter (<30 F) or hot-summer (> 90 F) states
- **Line plot:** matplotlib

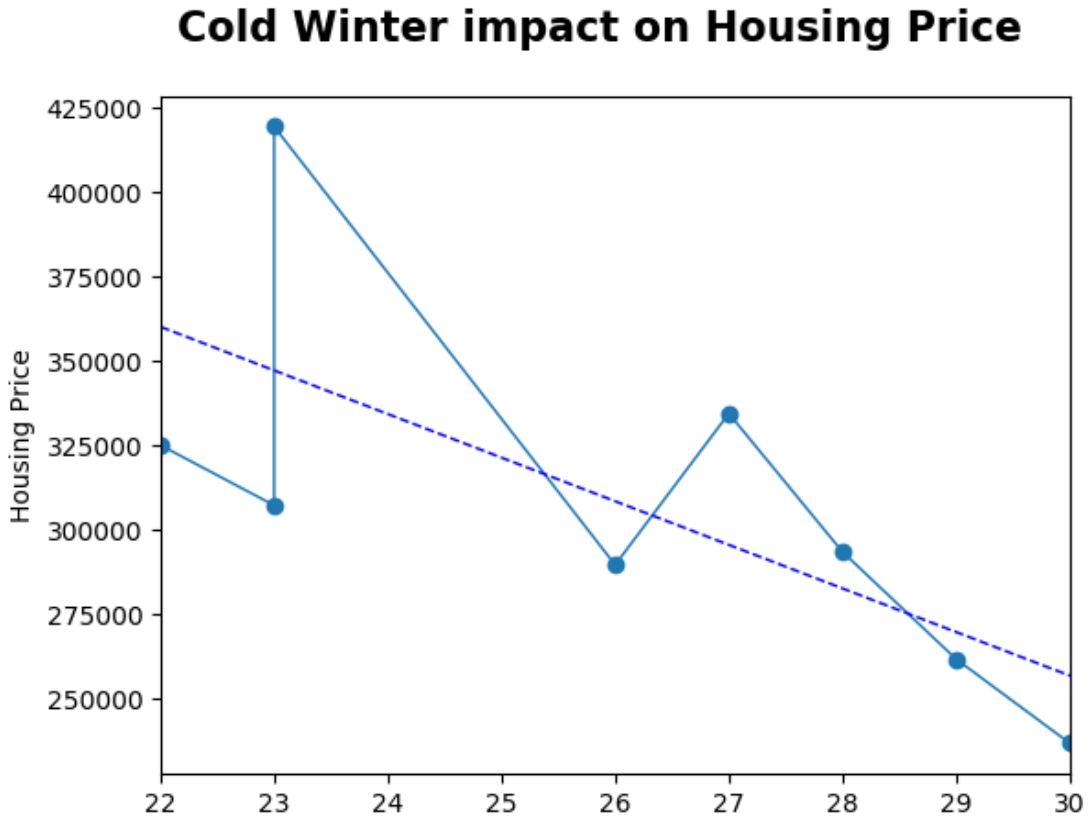
Weather Data



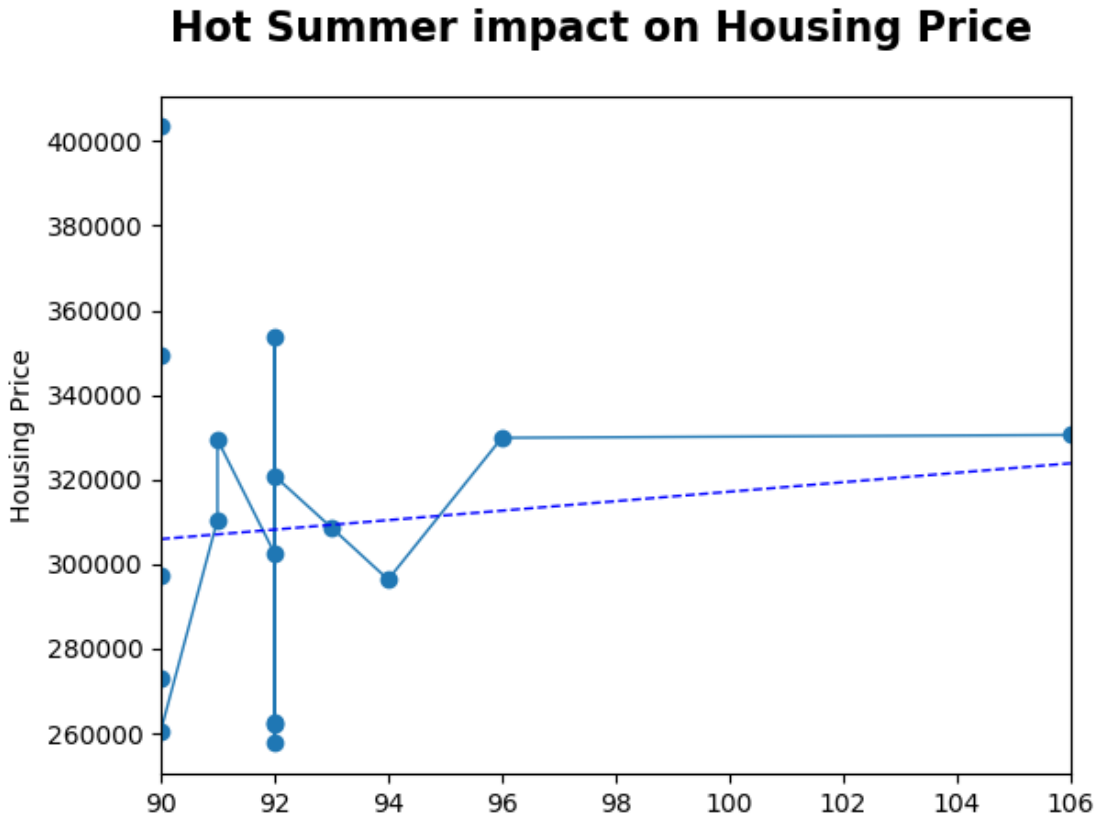
Summer (25%-75%) : 84 F to 90 F
(6 F difference)

Winter (25% - 75%) : 33 F to 50 F
(17 F difference)

Weather Data Correlation = - 0.7



Weather Data Correlation = 0.1



Why President Trump try hard to build a Barrier to protect Texas inside of a Wall to protect the Alaska?

Because Nobody wants to come to Alaska

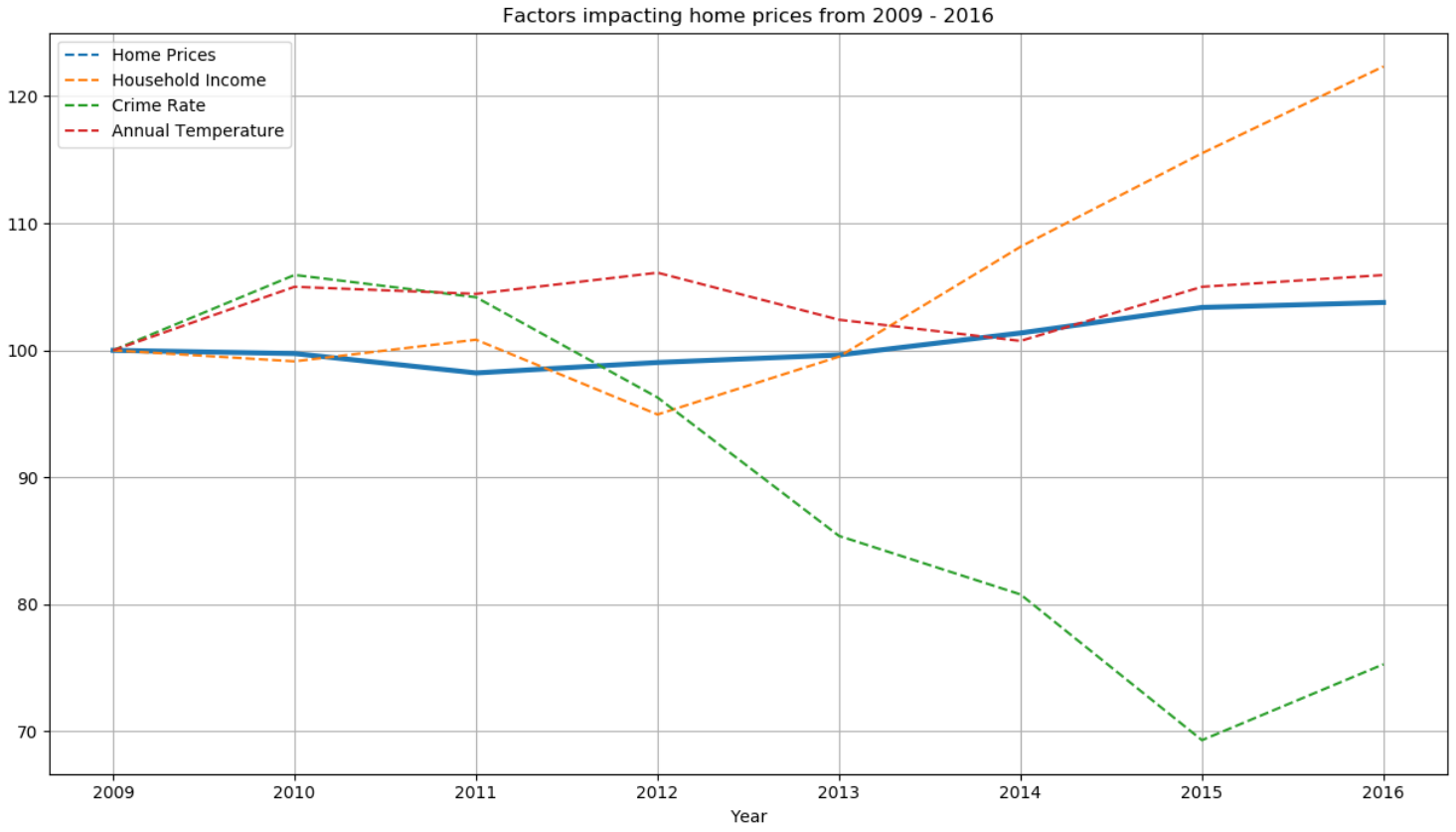
High housing price + Less job = Homeless

Homeless + Cold Winter = Froze to death

Analysis - New York

(2009 – 2016)

Analysis – New York



Data source: CSV

Objective: Track the correlation of New York house prices for 8 years (2009-2016) with household income, crime rate, and annual temperature.

Data manipulation:

1. Normalize all factors to start from 100 in 2009 to analyze the trend

Line plot: matplotlib

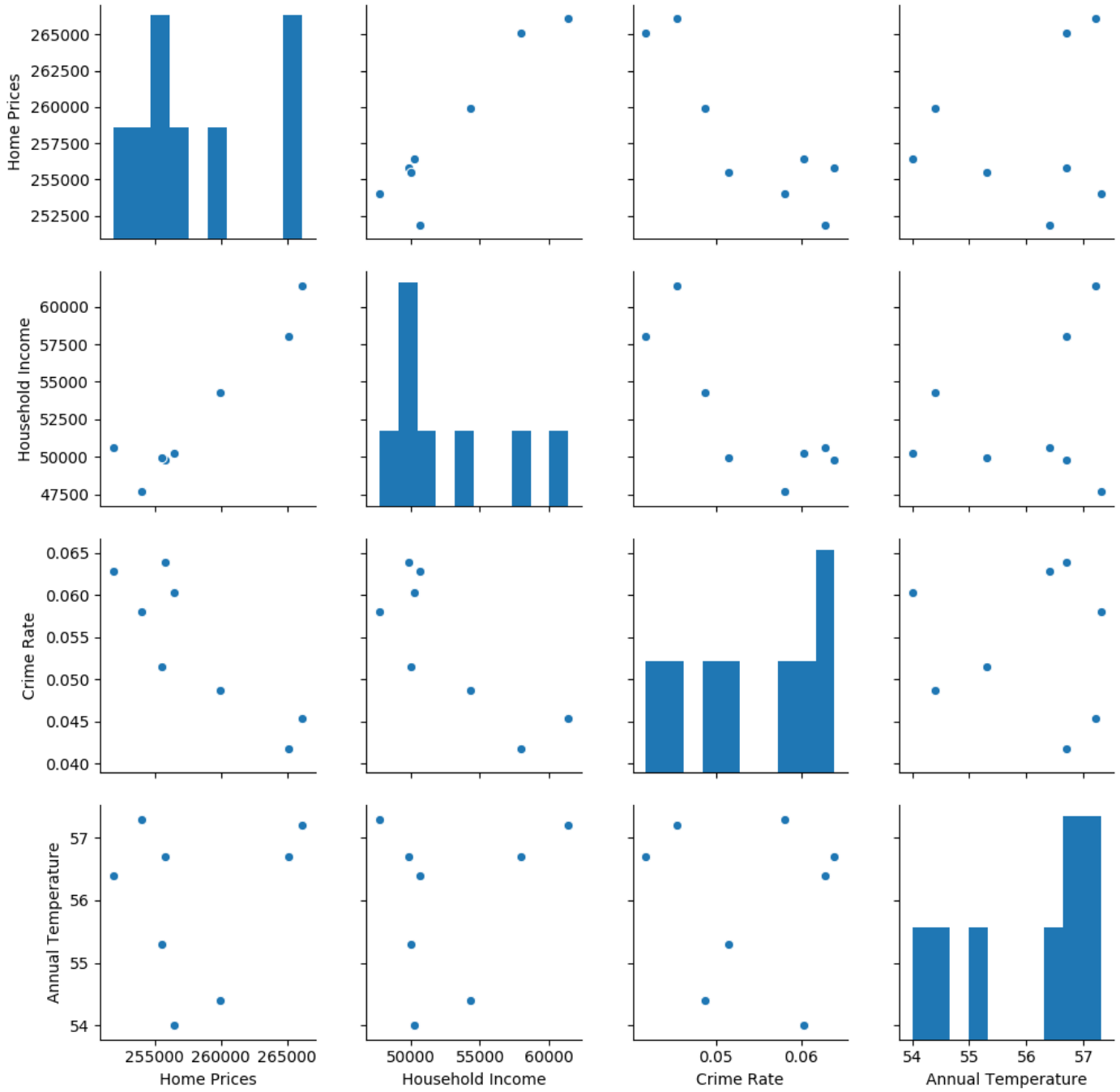
Observations:

1. Crime rate decreased> 30% significantly from 2010 – 2015
2. Household Income increased > 20% from 2012 - 2016

Analysis

Pairplot: Seaborn

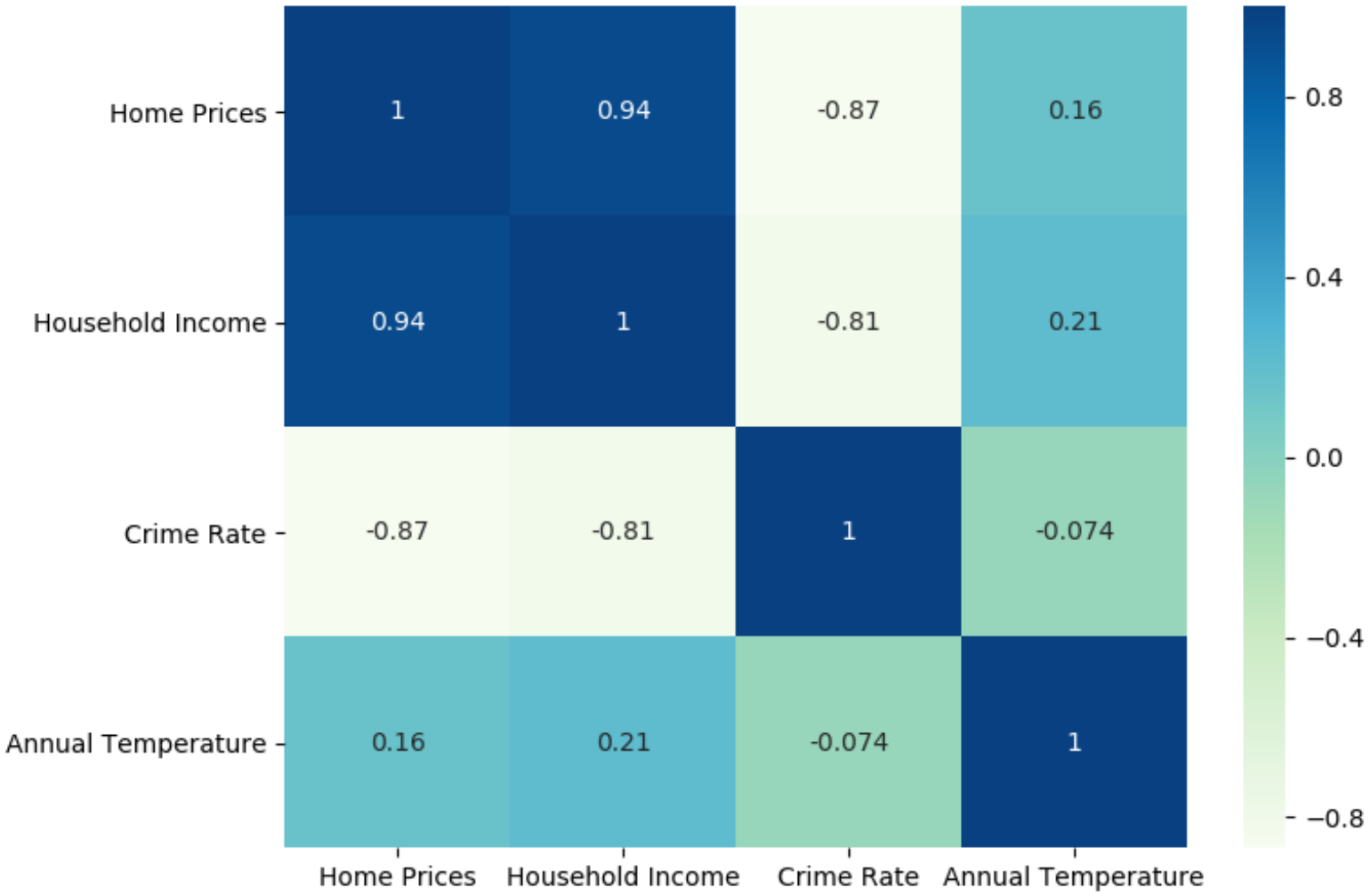
1. House Prices
2. Household Income
3. Crime Rate
4. Annual Temperature



Analysis

Heatmap: Seaborn

1. House Prices
2. Household Income
3. Crime Rate
4. Annual Temperature



Prediction – New York

(2009 – 2016)

Prediction – New York

Objective: Predict house prices in New York for the 8-year period (2009 –2016) and compare the model with actual house prices and analyze the impact of each factor

2 ways used to make our model:

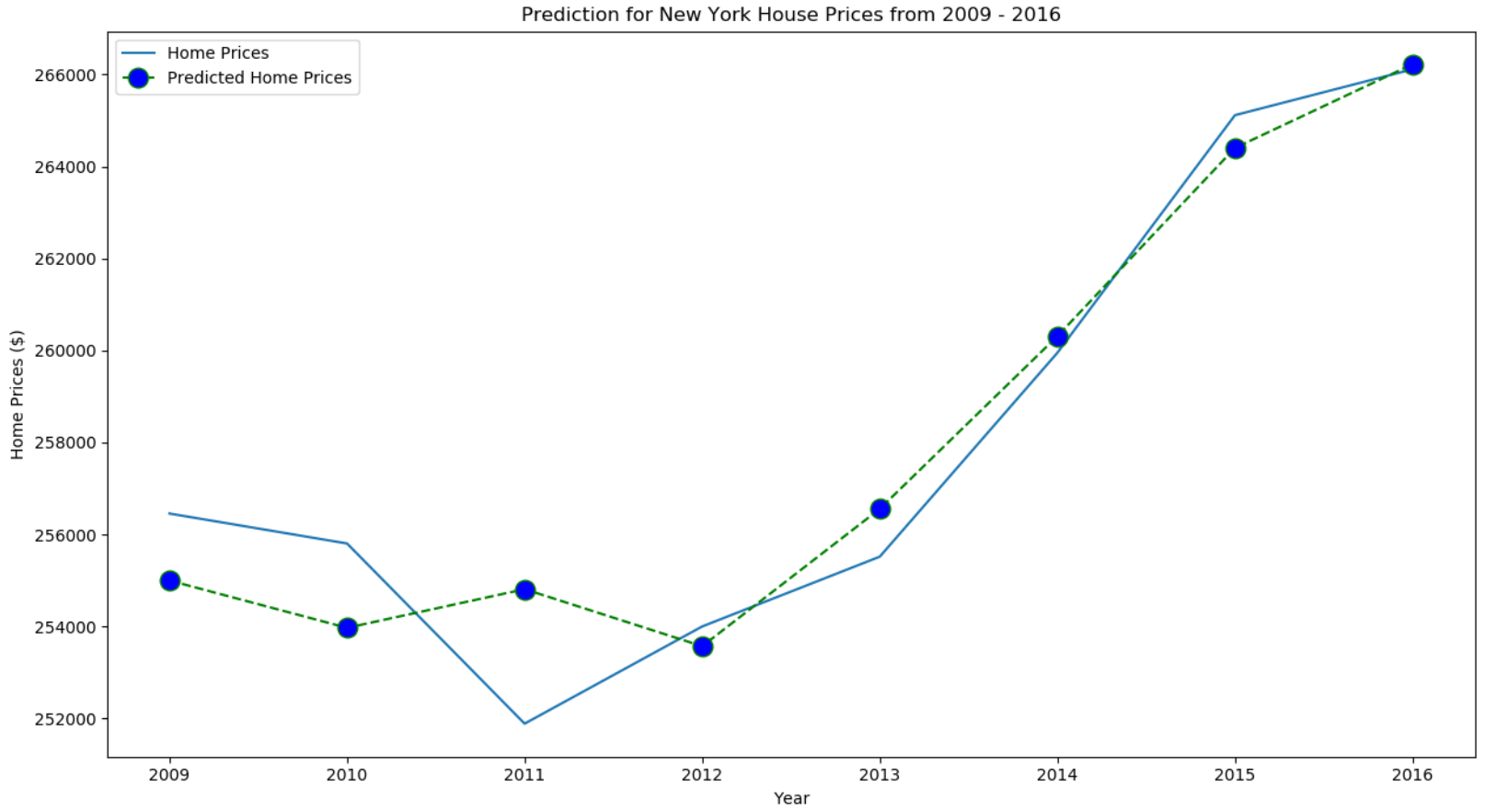
1. Statsmodels.api
   - An ordinary least square model
2. Statsmodels.formula.api
   - A formula (y ~ HI + CR)

The results are the same for both linear regressions.

In our model,

1. X denotes factors, which are
   - Household income (HI)
   - Crime rate (CR)
2. y denotes a target variable, which is house prices in New York

Prediction – New York



R-squared score is used to check the importance of factors:

1. Crime Rate (CR)
2. Household Income (HI)

Model Type	R-squared
Base Model	0.915
Model Without HI	0.753
Model Without CR	0.879

Conclusion

1. Both household income and crime rate are factors that are important to our prediction model.
2. Household income has more impact on house prices in New York than crime rate.

Suggestions to improve the prediction model

1. More factors (such as mortgage rates, GDP) can be analyzed and added
2. No. of observations in the training data can be increased, currently only 8 years
3. Future home prices (2019 upwards) can be predicted if estimated values of the factors are added to the training data

GitLab URL

Codes are available in the project GitLab: main.ipynb

https://columbia.bootcampcontent.com/thidar/project1_wts.git