2008 financial crisis in US and its impact on median income and housing prices.

Bryan Berlijn (284091), Marine Latour (2853681),Naomi van Leuvensteijn (2813993),Rocco vitale(28602

2025-06-24

Set-up your environment

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
library(dplyr)
library(ggplot2)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(tigris)
```

```
## To enable caching of data, set 'options(tigris_use_cache = TRUE)'
## in your R script or .Rprofile.
```

```
library(sf)
```

```
## Linking to GEOS 3.13.1, GDAL 3.10.2, PROJ 9.5.1; sf_use_s2() is TRUE
```

```
library(readxl)
```

Title Page

Bryan Berlijn, Marine Latour, Naomi van Leuvensteijn, Rocco vitale, Shruti Ramasamy, Toine van Bakel, Tyfarah Wedervoort

Tutorial 2, Group 3

Name of Tutor: Chantal Schouwenaar

Part 1: Problem Motivation & Literature

1.1: Why was the 2008 Financial Crisis in the US a Social Problem?

The 2008 Financial Crisis, commonly referred to as the Great Recession, originated in the U.S. and lasted from late 2007 to mid-2009. It was a major social problem due to its severe effects on median household income and housing prices, which significantly undermined economic security for millions of Americans. The crisis was mainly caused by excessive household leverage, particularly within the subprime mortgage sector, which triggered a steep decline in housing prices and a wave of mortgage defaults (Mian & Sufi, 2010a). The housing market collapse diminished household wealth and reduced consumer spending, intensifying the recession's severity and duration.

Mian and Sufi (2010a) demonstrate that households with higher leverage before the crisis faced substantially greater financial distress, leading to sharp reductions in durable consumption and higher unemployment rates at the county level. Furthermore, Morrissey et al. (2020) show that during the Great Recession, households with lower education levels, children, or without homeownership experienced significantly higher income and employment instability, an effect closely linked to increased financial hardship (pp. 3–4). Together, these economic and social disruptions illustrate why the Great Recession was not only a financial crisis but also a profound social crisis, affecting housing security, income stability, and overall well-being.

1.2: What sources have identified this topic as a social problem?

Several peer-reviewed studies have recognized the 2008 Financial Crisis as a critical social problem. Mian and Sufi's research (2010a, 2010b) emphasizes the role of household leverage and its consequences on the broader economy and local communities, highlighting the direct connection between financial market failures and deteriorating social conditions such as unemployment and poverty. Morrissey et al. (2020) extend this understanding by detailing how economic instability disproportionately affected vulnerable groups, including less-educated households and renters, thereby exacerbating social inequality during the recession.

1.3: Has a part of this topic not been researched yet?

While extensive research has analyzed household leverage, foreclosures, unemployment, and macroeconomic impacts of the 2008 financial crisis such as by Mian & Sufi, 2010a, 2010b, Morrissey et al., 2020), less academic focus has been placed on the geographic variation in how the relationship between housing price shifts and median income evolved across states during the crisis. What remained unexplored is quantitatively assessing how differently these two indicators interacted on a state-by-state basis, thus creating regional disparities.

1.4: How does the report bring new perspective on the topic?

This report creates a merged, state-level dataset to analyze income and housing prices together. Also, the following analysis will highlight where there was a misalignment between housing prices and income levels. Moreover, this report will visualize the regional variation among states in housing and income trends using

plots and correlations. Furthermore, this report examines these patterns through the wealth effect. The wealth effect demonstrates how changes in prices of assets, especially property values, influences consumer spending. During the 2008 crisis, as house prices were declining, the confidence of consumers in the value of their property also fell. Thus, they perceived a reduction in household wealth. This negative wealth effect, in turn, reduced consumer spending, which consequently contributed to higher unemployment at the country level.

Part 2 - Data Sourcing

2.1 Load in the data

Preferably from a URL, but if not, make sure to download the data and store it in a shared location that you can load the data in from. Do not store the data in a folder you include in the Github repository!

Dataset 1, HPI : hpi_data_set = https://www.fhfa.gov/hpi/download/annual/hpi_at_state.xlsx

Dataset 2: Median income = https://nces.ed.gov/programs/digest/d11/tables/xls/tabn025.xls

```
hpi_at_state = read_xlsx("hpi_at_state.xlsx")

tabn025 = read_xls("tabn025.xls")
```

```
## New names:
## * `` -> `...5`
## * `` -> `...7`
## * `` -> `...9`
## * `` -> `...11`
## * `` -> `...13`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...20`
## * `` -> `...21`
## * `` -> `...22`
## * `` -> `...23`
## * `` -> `...24`
## * `` -> `...25`
## * `` -> `...26`
## * `` -> `...27`
## * `` -> `...28`
## * `` -> `...29`
## * `` -> `...30`
## * `` -> `...31`
## * `` -> `...32`
## * `` -> `...33`
## * `` -> `...34`
## * `` -> `...35`
## * `` -> `...36`
## * `` -> `...37`
## * `` -> `...38`
## * `` -> `...39`
## * `` -> `...40`
```

2.2 Provide a short summary of the dataset(s)

```
head(hpi_at_state)
```

```
## # A tibble: 6 x 8
##   State  Abbreviation FIPS   Year `Annual Change (%)`   HPI `HPI with 1990 base`
##   <chr>  <chr>        <chr> <dbl>               <dbl> <dbl>                <dbl>
## 1 Alaba~ AL           01     1975               NA     100                  48.8
## 2 Alaba~ AL           01     1976                8.21  108.                 52.8
## 3 Alaba~ AL           01     1977               10.4   119.                 58.2
## 4 Alaba~ AL           01     1978                7.9   129.                 62.8
## 5 Alaba~ AL           01     1979               10.9   143.                 69.7
## 6 Alaba~ AL           01     1980                5.15  150.                 73.3
## # i 1 more variable: `HPI with 2000 base` <dbl>
```

```
head(tabn025)
```

```
## # A tibble: 6 x 40
##   State    `1990\\1\\` `2000\\2\\` `2005`  ...5 `2006`  ...7 `2007`  ...9 `2008`
##   <chr>          <dbl>       <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl>
## 1 <NA>             NA          NA     NA    NA     NA    NA     NA    NA     NA
## 2 United ~      51028.      54951.  51600    70  52400    50  53400    50  52700
## 3 Alabama~      40062.      44667.  41200   360  41900   340  42600   270  43200
## 4 Alaska ~      70301.      67483.  62800  1230  64200   950  67700  1020  69300
## 5 Arizona~      46757.      53072.  49500   440  51100   290  52500   320  51600
## 6 Arkansa~      35903.      42111.  39100   410  39600   320  40100   470  39300
## # i 30 more variables: ...11 <dbl>, `2009` <dbl>, ...13 <dbl>, `2010` <dbl>,
## #   ...15 <dbl>, ...16 <lgl>, ...17 <lgl>, ...18 <lgl>, ...19 <lgl>,
## #   ...20 <lgl>, ...21 <lgl>, ...22 <lgl>, ...23 <lgl>, ...24 <lgl>,
## #   ...25 <lgl>, ...26 <lgl>, ...27 <lgl>, ...28 <lgl>, ...29 <lgl>,
## #   ...30 <lgl>, ...31 <lgl>, ...32 <lgl>, ...33 <lgl>, ...34 <lgl>,
## #   ...35 <lgl>, ...36 <lgl>, ...37 <lgl>, ...38 <lgl>, ...39 <lgl>,
## #   ...40 <lgl>
```

This first dataset, published by the **Federal Housing Finance Agency (FHFA)**, provides yearly data on house price trends across all American States using the House Price Index (HPI). The HPI measures changes in single-family home prices over time, constructed using repeat-sales data and financed through mortgages backed by Fannie Mae and Freddie Mac. The repeat-sales methodology enhances accuracy by tracking price changes for the same properties over time.

The second dataset, published by the **National Center for Education Statistics** (NCES), provides annual median household income data for the United States and the individual states. It offers insights into the economic conditions of each state over a specific period.

2.3 Describe the type of variables included

Housing Price Index Dataset

Credibility: FHFA is a U.S. federal agency responsible for regulating the housing finance system. Its data is trusted by researchers, policymakers, and financial institutions due to its transparency, government oversight, and long historical coverage.

Metadata Overview

- **Frequency:** Annually
- **Time Range:** This covers from 1975 to 2024, but our focus point is 2008
- **Geographic Scope:** States
- **Key Variables:**

- **Purchase-Only Index** (focuses on purchase transactions only)

-**All-Transactions Index** (includes both purchases and refinances mortgages)

-**Percent Changes** (year-over-year growth)

-**Seasonally Adjusted & Not Seasonally Adjusted values**

- **Format:** Excel report (XLSX)

Why This Data is Suitable for the Topic

These specific FHFA reports show the progression of U.S. housing conditions during and before the period of financial crisis. They offer reliable, government-backed data with consistent methodology dating back to 1975. The breakdown by state level allows for a better understanding of how different regions were impacted by the price change. Its focus on repeat-sales ensures accurate tracking of real home price changes over time, making it well-suited for studying the dynamics of the housing market crisis.

Data Limitations

1. **Market Coverage Gaps:** Excludes cash sales and jumbo loans, potentially underrepresenting luxury and investor markets, as well as homes not financed through Fannie Mae or Freddie Mac.

2. **Limited detail within reports:** While the reports provide detailed price index, they do not contain demographic, income, or specific housing type data, limiting deeper analysis.

Income Dataset

Credibility: The NCES is the primary federal unit responsible for collecting and analysing data related to U.S. Department of Education, which makes this data highly reliable, objective, and a source of authority for economic statistics.

Metadata Overview

- **Frequency:** Annually
- **Time range:** Covers 1990, 2005-2010, but our focus point is 2008
- **Geographic scope:** National (United States) and all U.S. States
- **Key Variables:**

-Median household income

-Margin of error for estimates

- **Format:** Excel (XLSX)

- **Income adjustment:** All income values are expressed in 2010 dollars

Why This Data is Suitable for the Topic?
This dataset provides a clear and direct view of the median household income trends at state level in America over time, including periods of economic change. It can be used to observe income differences between states and the economic well-being of households. The breakdown by state level allows for a better understanding of how different regions were impacted by the economic crisis.

Discuss data limitations

1. **Limited income detail:** The table provides median household income, but does not break it down by demographic characteristics, for example, race, age or educational attainment.

2. **Limited methodology used**: The data provided lacks indication about the survey methods used and inflation adjustments. For instance, it might be possible that certain states used different survey methods than others, making it difficult to accurately portray differences in cost of living among them.

Explain how these 2 datasets are different or complementary

The datasets are complementary as they both revolve around an important asset, namely housing, while offering state-level data across multiple years. The FHFA dataset provides insight into housing prices, reflecting the market value of a household's asset, while the NCES dataset provides information on median income which is a critical factor which influences a household's ability to afford housing, whether through mortgage payments, rent, or homeownership costs. The connection between the two datasets lies in the fact that income determines a household's ability to purchase or maintain housing. Conversely, fluctuations in housing prices can influence median income level through the wealth effect.

Part 3 - Quantifying

3.1 Data cleaning

```
#Changing both data sets to data frame
tabn025 = as.data.frame(tabn025)
hpi_at_state = as.data.frame(hpi_at_state)
```

```
#Filtering the HPI dataset for 20006-2010
hpi_filtered = hpi_at_state %>%
  filter(Year %in% c(2006,2007,2008,2009,2010)) %>%
  select(State, Year, HPI)
```

```
#Filtering the Income data set for 2006-2010
income_to_use = select(tabn025,State,"2006","2007","2008","2009","2010")
```

```
#Formatting the income data set into a long format
income_long = income_to_use %>%
  pivot_longer(
    cols = -State,
    names_to = "Year",
    values_to = "Income"
  )
```

```r
#Omitting the NA rows since we dont need
income_long_no_NA = na.omit(income_long)
```

```r
#Making the State titles the same
income_long_no_NA$State = income_long_no_NA$State %>%
  str_replace_all("[.]", "") %>%
  str_squish() %>%
  str_trim()
```

```r
#Making the Year the same numeric class
hpi_filtered$Year = as.numeric(hpi_filtered$Year)
income_long_no_NA$Year = as.numeric(income_long_no_NA$Year)
```

```r
#merging the 2 data sets
merge_2 = inner_join(hpi_filtered , income_long_no_NA, by = c("State","Year"))
```

3.2 Generate necessary variables

New variable 1: Housing price to income ratio. Changes in this ratio over time would indicate whether housing prices are rising faster or slower than income.

Housing Price to Income Ratio= House Price Index Value / Median Household Income Value

Why the creation of this variable is informative and useful for your analyses?

We used this ratio to measure the affordability of housing in each state by comparing housing prices to household income. A rising ratio means less affordable housing, while a lower ratio suggests more affordability. We observed that before the 2008 financial crisis, the price_to_income_ratio behaved differently for each state.
The creation of this variable combines the HPI, the income into one indicator as well as data on individual states. Therefore, the variable assesses a household's financial situation, differences in economic and social conditions across different geographic areas or regions, and the recovery post-crisis.

What you intend to analyse/visualize using this variable?

With our new variable, we can visualize a Price- Income divergence. This enables us to look at whether house prices increase faster than the median household incomes. In an economy functioning normally, house prices and income are steadily increasing together. Thus, the ratio will be lower in a normal economy. However, house prices can increase faster than incomes which was common in many states before the crisis of 2008. This was because of expectations about future prices for houses. Moreover, the new variable helps to evaluate which states were better off and which ones were worse off in terms of housing affordability in the midst of the crisis.

```r
#create variable 1: house prices to income ratio
merge_2 <- merge_2 %>%
  mutate(price_to_income_ratio = (HPI / Income)*1000)
```

New variable 2: Link between the housing prices and income value. This link explains the wealth effect and how falling property value led to lower spending and thus lower incomes.

Link between income value and house price: link between housing price change (HPI_t-HPI_{t-1}) and income value change ($Income_t$-$Income_{t-1}$) , where t= year.

Why the creation of this variable is informative and useful for your analyses?

As explained earlier, our report evaluates whether the wealth effect had an impact in the 2008 U.S. financial crisis. This direct relationship between the income value change and the housing prices change helps to analyse whether a drop in housing prices led to a fall in the income value.

What you intend to analyse/visualize using this variable?

We created a new column called link_price_to_income, which labels the relationship between changes in income and housing prices using characters such as "Both Fell", "Income fell, HPI increased", "Income stable, HPI fell", "Income stable, HPI increased", "Income increased, HPI fell", "Income increased, HPI increased". This allows us to directly visualize the link between two variables more easily, for example, we can identify which states experienced "Both fell" during the financial crisis, this helps to establish a direct correlation between the two variables, indicating that the wealth effect worsened the crisis. In this context, the new variable shows how the connection between the two directly contributed to economic and social hardships.

```
# create new variable of income_change and HPI_change

merge_2 <- merge_2 %>%
  arrange(State, Year) %>%  # ensure data is ordered
  group_by(State) %>%       # compute changes within each state
  mutate(
    income_change = Income - lag(Income),
    HPI_change = HPI - lag(HPI)
  ) %>%
  ungroup()

#create variable 2: link between income loss & housing price drop

merge_2 <- merge_2 %>%
  mutate(
    link_price_to_income = case_when(
      income_change < 0 & `HPI_change` < 0  ~ "Both fell",
      income_change < 0 & `HPI_change` > 0 ~ "Income fell, HPI increased",
      income_change == 0 & `HPI_change` < 0 ~ "Income stable, HPI fell",
      income_change == 0 & `HPI_change` > 0 ~ "Income stable, HPI increased",
      income_change > 0 & `HPI_change` < 0 ~ "Income increased, HPI fell",
      income_change > 0 & `HPI_change` > 0 ~ "Income increased, HPI increased",
      TRUE ~ "Other"
    )
  )
#describing missing data for 2006 as NAs
merge_2$link_price_to_income[merge_2$Year=="2006"] = "NA"

merge_2$link_price_to_income[merge_2$Year=="2006"] = "NA"
```

3.3 Visualize temporal variation

```
#Rename merge_2 as data
data = merge_2
```

```
#Rename columns as necessary
colnames(data) <- make.names(colnames(data))
```

```r
#Make the year become an integer
data$Year <- as.integer(data$Year)
```

```r
#Select only necessary columns
plot_data <- data %>%
  select(State, Year, price_to_income_ratio)
```

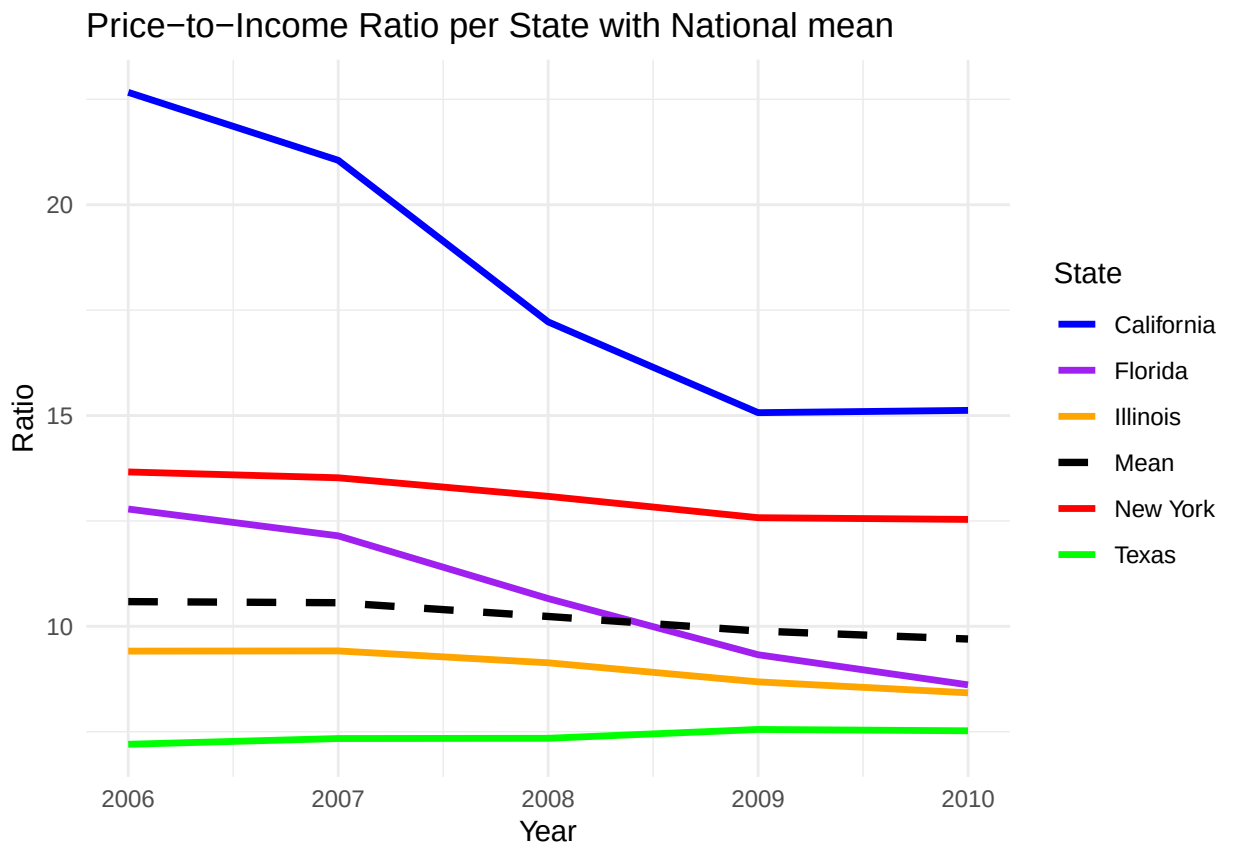
```r
#Calculate average for all states per year
average_data <- plot_data %>%
  group_by(Year) %>%
  summarise(avg_ratio = mean(price_to_income_ratio, na.rm = TRUE))
```

```r
#Filter data for top states
top_states <- c("California", "Texas", "New York", "Florida","Illinois")

filtered_data <- plot_data %>%
  filter(State %in% top_states)
```

```r
#Create the line graph
ggplot() +
  geom_line(data = filtered_data, aes(x = Year, y = price_to_income_ratio, color = State), size = 1.2) +
  geom_line(data = average_data, aes(x = Year, y = avg_ratio, color = "Mean"), size = 1.3, linetype = "d
  scale_color_manual(values = c("California" = "blue", "Texas" = "green", "New York" = "red",
                                "Florida" = "purple", "Illinois" = "orange", "Mean" = "black")) +
  labs(title = "Price-to-Income Ratio per State with National mean",
       x = "Year", y = "Ratio", color = "State") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



How does the problem you've identified vary over time?

The line graph indicates a higher price-to-income ratio in 2006, which declined over time. As the crisis progressed, the line graph illustrates that affordability improved, reflected by the decreasing or steady price-to-income ratio lines.

Include a short description on how the visualization aligns with[your topic]:

As demonstrated in this line graph, the price-to-income ratio of several prominent American states is presented alongside the national mean. This analysis offers a clear indication that the crisis exerted a more significant impact in certain states compared to others. The states which had a higher ratio and experienced lower house affordability are California, New York and Florida. On the other hand, states with a lower ratio, experiencing higher house affordability are Illinois and Texas.

3.4 Visualize spatial variation

```
# mapping state names to STUPS
state_abbr <- tibble::tibble(
  State = state.name,
  STUSPS = state.abb
)
```

```
# Adding manually DC(District of Columbia)
state_abbr <- bind_rows(state_abbr, tibble(State = "District of Columbia", STUSPS = "DC"))
```

```r
# Adding abbreviations to the dataframe
merge_2 <- merge_2 %>%
  left_join(state_abbr, by = "State")
```

```r
# Retrieving US shapefiles and filtering
states_sf <- states(cb = TRUE) %>%
  filter(!STUSPS %in% c("PR", "VI", "GU", "MP", "AS")) %>%
  st_transform(crs = 5070)
```

```
## Retrieving data for the year 2024
```

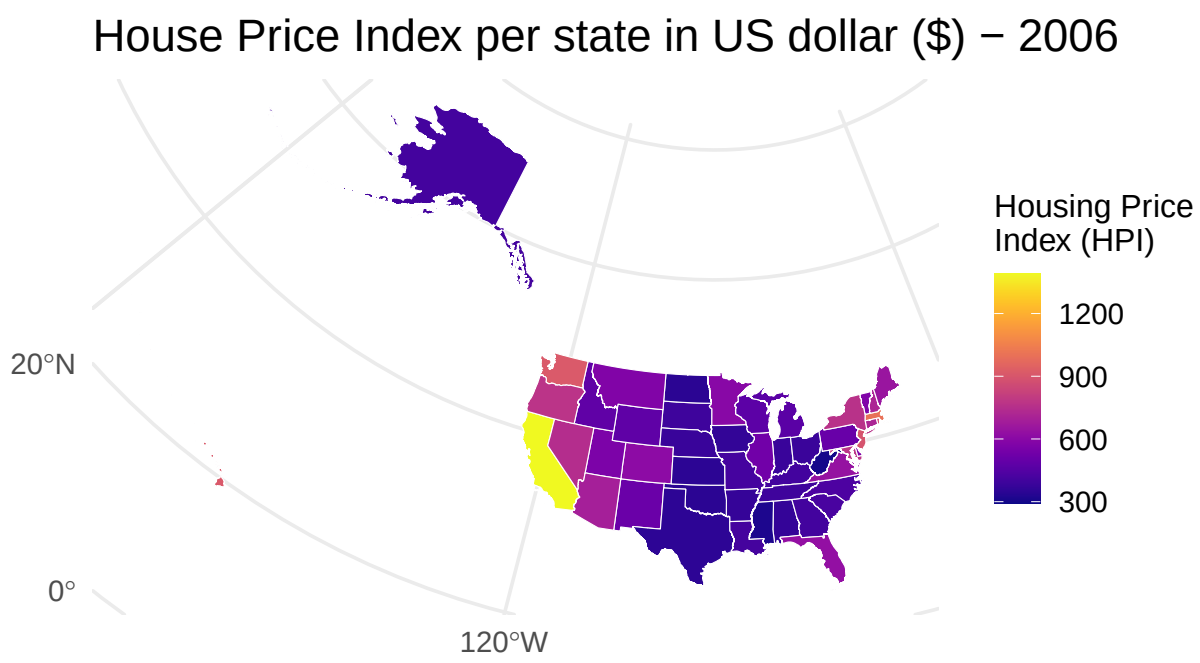
```
##   |                                                                              |
```

```r
# Merging shapefiles and data for all the years
map_data <- left_join(states_sf, merge_2, by = "STUSPS")
```

```r
#setting the limits for the scale of the heat map
hpi_min <- min(map_data$HPI, na.rm = TRUE)
hpi_max <- max(map_data$HPI, na.rm = TRUE)
```

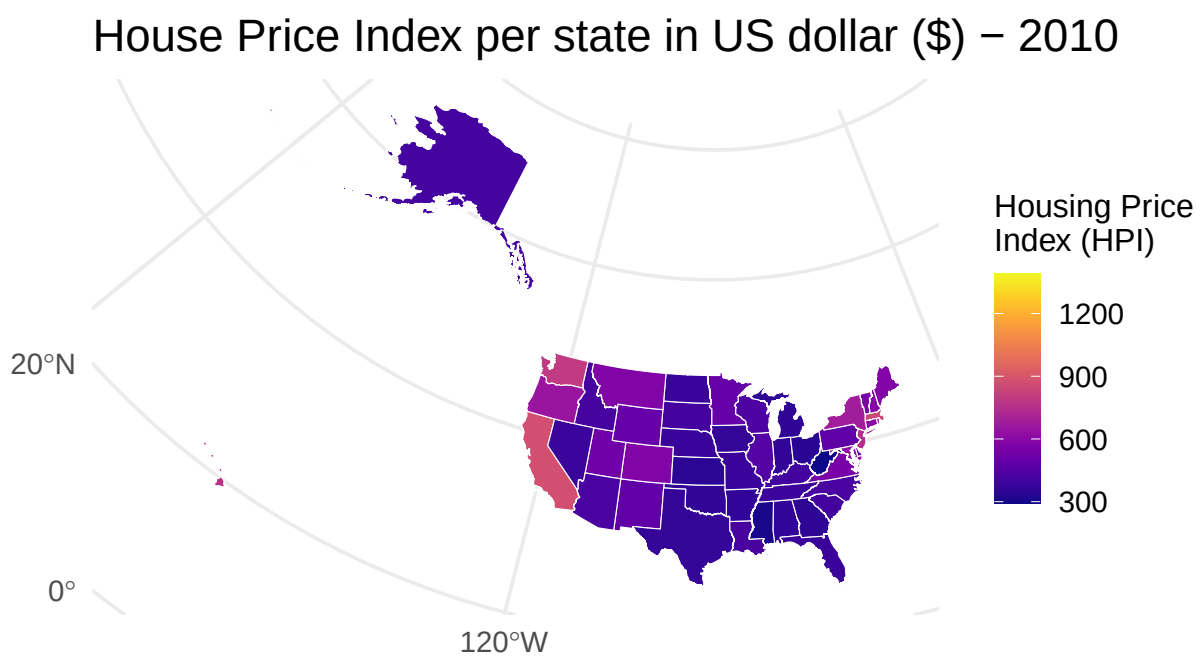
```r
# Filter for 2006
map_data_2006 <- map_data %>% filter(Year == 2006)
```

```r
# Plot 2006
ggplot(map_data_2006) +
  geom_sf(aes(fill = HPI), color = "white", size = 0.2) +
  scale_fill_viridis(option = "plasma", na.value = "grey90", limits = c(hpi_min, hpi_max)) +
  theme_minimal(base_size = 14) +
  labs(
    title = "House Price Index per state in US dollar ($) - 2006",
    fill = "Housing Price \nIndex (HPI)"
  ) +
  theme(
    legend.position = "right",
    legend.title = element_text(size = 12)
  )
```




```r
map_data_2010 <- map_data %>% filter(Year == 2010)

# Plot 2010
ggplot(map_data_2010) +
  geom_sf(aes(fill = HPI), color = "white", size = 0.2) +
  scale_fill_viridis(option = "plasma", na.value = "grey90", limits = c(hpi_min, hpi_max)) +
  theme_minimal(base_size = 14) +
  labs(
    title = "House Price Index per state in US dollar ($) - 2010",
    fill = "Housing Price \nIndex (HPI)"
  ) +
  theme(
    legend.position = "right",
    legend.title = element_text(size = 12)
  )
```



How does the problem you've identified geographically/spatially vary?.

The map helps illustrate how the problem of declining housing affordability varied across American states between 2006 and 2010. By comparing the HPI before and after the 2008 financial crisis. States shaded in yellow or red were the least affordable. In contrast, states shown in shades of purple experienced more affordable housing during this period.

Include a short description on how the visualization aligns with

In 2006, several states of the western region, especially California, Washington, and Massachusetts, had very high HPI values, which indicates rapidly rising home prices. This growth surpassed income, contributing to a decline in housing affordability.

By 2010 the aftermath of the housing market crash is visible in many states due to a drop in HPI values, indicating a decline in home prices. This means that housing became more affordable, but due to economic distress.

Think about how zoomed in or out the map is such that it fits the spread of location of your data

Since our analysis focuses on state variation, using a state level map is the most appropriate. It provides a clear visualization of differences between states, allowing us to observe how the housing affordability problem and the impact of the 2008 financial crisis varied geographically. This scale perfectly fits the scope of our research by capturing the regional trends, making it easier to compare and visualize state-by-state changes in housing price.

3.5 Visualize sub-population variation 1

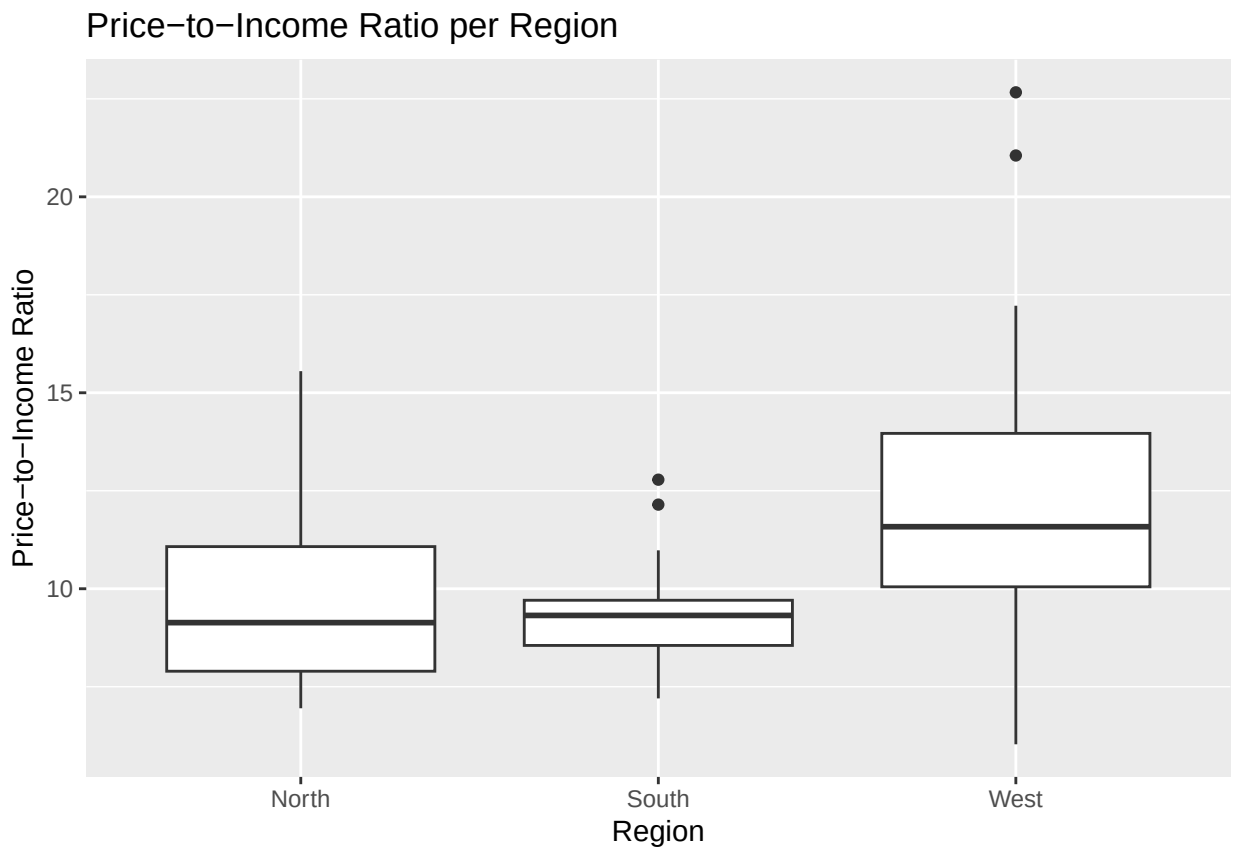
```r
#divide by regions
merge_2 <- merge_2 %>%
  mutate(
    region = case_when(
      State %in% c(
        # Combine Northeast + Midwest as North
        "Maine", "New Hampshire", "Vermont", "Massachusetts", "Rhode Island", "Connecticut",
        "New York", "New Jersey", "Pennsylvania",
        "Ohio", "Indiana", "Illinois", "Michigan", "Wisconsin", "Minnesota", "Iowa",
        "Missouri", "North Dakota", "South Dakota", "Nebraska", "Kansas"
      ) ~ "North",

      State %in% c(
        # South
        "Delaware", "Maryland", "District of Columbia", "Virginia", "West Virginia",
        "North Carolina", "South Carolina", "Georgia", "Florida", "Kentucky", "Tennessee",
        "Mississippi", "Alabama", "Texas", "Arkansas", "Louisiana", "Oklahoma"
      ) ~ "South",

      State %in% c(
        # West
        "Montana", "Idaho", "Wyoming", "Colorado", "New Mexico", "Arizona", "Utah", "Nevada",
        "California", "Oregon", "Washington", "Alaska", "Hawaii"
      ) ~ "West",

      TRUE ~ "Other"
    )
  )
```

```r
#Created the boxplot
ggplot(merge_2, aes(x = region, y = price_to_income_ratio)) +
  geom_boxplot() +
  labs(
    title = "Price-to-Income Ratio per Region",
    x = "Region",
    y = "Price-to-Income Ratio"
  )
```



Include a short description on how the visualization aligns with[your topic].

The boxplot visualizes the distribution of the price-to-income ratio for U.S. states grouped by regions, namely North, South and West. As visualized, the mean of the ratio for the western region being above 10, is higher than the other regions which are below 10. Also, the graph reveals the variability and presence of outliers. In the western region, the two outliers' ratios go beyond 20, which shows that this region greatly suffered a lack of housing affordability.

Visualize sub-population variation 2

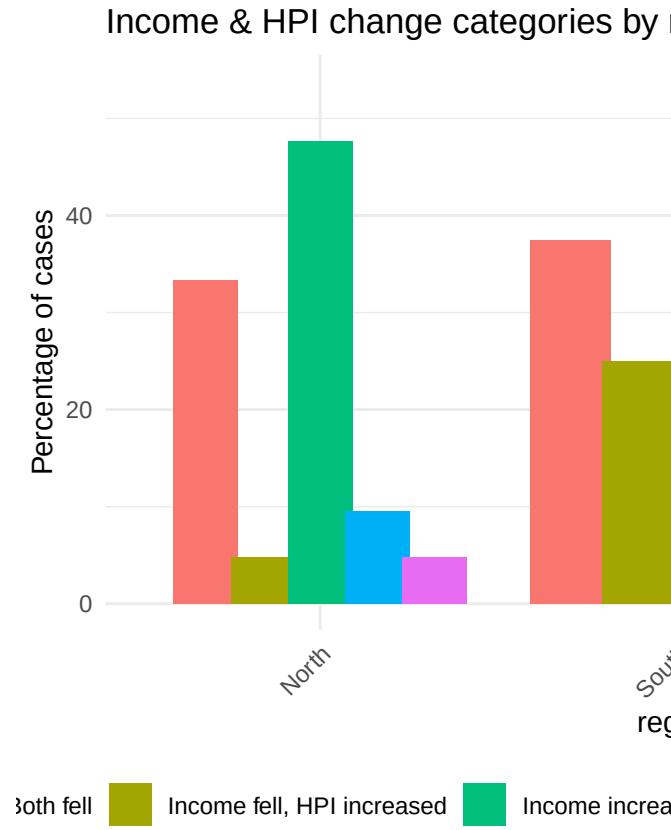
```
# Make sure Year is numeric
merge_2 <- merge_2 %>%
  mutate(Year = as.numeric(Year))
```

```
# Filter for 2008
merge_2_filtered <- merge_2 %>%
  filter(Year == 2008)
```

```
# Summarize counts + compute percentages within each region
summary_data <- merge_2_filtered %>%
```

```
  group_by(region, link_price_to_income) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(region) %>%
  mutate(percent = count / sum(count) * 100) %>%
  ungroup()
```

```
# Plot
ggplot(summary_data, aes(x = region, y = percent, fill = link_price_to_income)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8)) +
  labs(
    title = "Income & HPI change categories by region (2008)",
    x = "region",
    y = "Percentage of cases",
    fill = "Category"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom"
  )
```



Bar chart of Income and HPI change categories by region.

Include a short description on how the visualization aligns with[your topic].

The bar chart illustrates the role of the wealth effect during the financial crisis. It is constructed by using the second variable, which examines the link between changes in income and house price for the year 2008. This

graph groups the states into 3 regions: North, South and West. As shown, during the peak of the crisis, the Western region experienced a stronger negative wealth effect, because most states experienced a fall in both variables. In contrast, the Northern region had a lower percentage of states experience "Both Fell". Instead, a larger share of states experienced the income change variable to either increased or remained stable.

3.6 Event analysis

```
#Rename merge_2 as data
data = merge_2
```

```
#Rename columns as necessary
colnames(data) <- make.names(colnames(data))
```

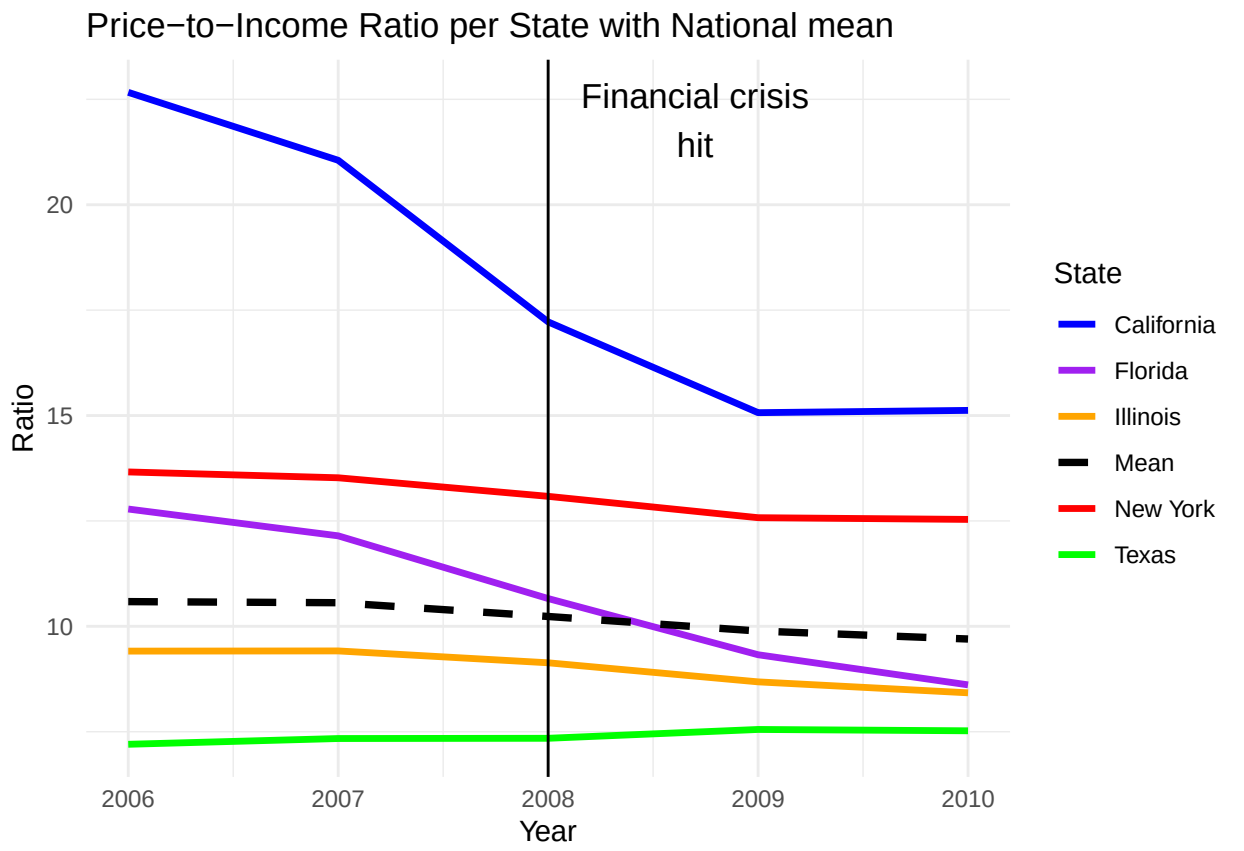
```
#Make the year become an integer
data$Year <- as.integer(data$Year)
```

```
#Select only necessary columns
plot_data <- data %>%
  select(State, Year, price_to_income_ratio)
```

```
#Calculate average for all states per year
average_data <- plot_data %>%
  group_by(Year) %>%
  summarise(avg_ratio = mean(price_to_income_ratio, na.rm = TRUE))
```

```
#Filter data for top states
filtered_data <- plot_data %>%
  filter(State %in% top_states)
```

```
#Create the line graph with the vertical line representing the event analysis
ggplot() +
  geom_line(data = filtered_data, aes(x = Year, y = price_to_income_ratio, color = State), size = 1.2) +
  geom_line(data = average_data, aes(x = Year, y = avg_ratio, color = "Mean"), size = 1.3, linetype = "d
  scale_color_manual(values = c("California" = "blue", "Texas" = "green", "New York" = "red",
                                "Florida" = "purple", "Illinois" = "orange", "Mean" = "black")) +
  labs(title = "Price-to-Income Ratio per State with National mean",
       x = "Year", y = "Ratio", color = "State") +
  theme_minimal() +
  geom_vline(xintercept = 2008, linetype = "solid", color = "black") +
  annotate("text" , x= 2008.7 , y = 22, size = 4.5, label = "Financial crisis\nhit")
```



Here you provide a description of why the plot above is relevant to your specific social problem.

Although we studied the income and price changes and the relationship between them during the years of 2006-2010, the focus point throughout this report is 2008. The vertical line which is added to the temporal visualization at 2008 marks the peak of the financial crisis. It acts as a reference point to compare housing affordability trends before and after the economic collapse.

Part 4 - Discussion

4.1 Discuss your findings

Our analysis of the 2008 financial crisis shows that pre-crisis the house prices were increasing but then fell during and post the recession. This aligns with the negative wealth effect we studied as a key factor worsening the crisis. Our study also concluded that the regional differences were an additional factor that explained differences in the intensity of the consequences of the crisis. For instance, California was concluded to be the leading state of the Western region in being the most affected. This is illustrated by the line graph and the second variable where its price-to-income ratio dropped significantly during the crisis. This is also shown by the heatmap where its 2006 HPI was high pre-crisis but then dropped post-crisis. In conclusion, the crisis' effects were amplified by a combination of housing dynamics and regional disparities.

Part 5 - Reproducibility

5.1 Github repository link

Public Repository Link : https://github.com/Marine444/Group-3-programming

5.2 Reference list

1. MIAN, A., & SUFI, A. (2010). Household Leverage and the Recession of 2007-09. *IMF Economic Review, 58(1)* , p.74–117. https://www.jstor.org/stable/25762071

2. Mian, A., & Sufi, A. (2010b). The Great Recession: Lessons from Microeconomic Data. *The American Economic Review, 100(2)*, p.51–56. https://doi.org/10.1257/aer.100.2.51

3. Morrissey, T. W., Cha, Y., Wolf, S., & Khan, M. (2020). Household economic instability: Constructs, measurement, and implications. *Children and Youth Services Review,* 118. https://doi.org/10.1016/j.childyouth.2020.105502

4. National Center for Education Statistics. (2011). *Median household income, by state: Selected years, 1990 through 2010.* https://nces.ed.gov/programs/digest/d11/tables/dt11_025.asp

5. National Center for Education Statistics. (n.d.). *National Center for Education Statistics (NCES).* https://nces.ed.gov/

6. U.S. Federal Housing. (2025). *"HPI for states (All – Transactions Index): Experimental Indexes Showing Cumulative (Nominal) Annual Appreciation".* https://www.fhfa.gov/hpi/download/annual/hpi_at_state.xlsx


```
setwd("C:/Users/tyfar/OneDrive/Desktop/school/programming/Group-3-programming")
```

```
install.packages(c("renv", "tinytex"))
```

```
## The following package(s) will be installed:
## - renv    [1.1.4]
## - tinytex [0.57]
## These packages will be installed into "C:/Users/tyfar/OneDrive/Desktop/school/programming/Group-3-pr
##
## # Installing packages --------------------------------------------------------
## - Installing renv ...                          OK [linked from cache]
## - Installing tinytex ...                       OK [linked from cache]
## Successfully installed 2 packages in 0.17 seconds.
```

```
# 1. restore the exact package versions used by the course
renv::restore()        # looks at renv.lock
```

```
## - The library is already synchronized with the lockfile.
```

```r
# 2. check whether a LaTeX compiler is already present
has_tex <- tinytex::is_tinytex() || nzchar(Sys.which("pdflatex"))

if (!has_tex) {
  os <- Sys.info()[["sysname"]]
  message("No LaTeX found → installing TinyTeX for ", os)
  tinytex::install_tinytex()     # ~250 MB, 100 % automatic
  message("Restart RStudio after installation finishes.")
}
library(tinytex)
library(renv)
```

```
##
## Attaching package: 'renv'

## The following object is masked from 'package:purrr':
##
##     modify

## The following objects are masked from 'package:stats':
##
##     embed, update

## The following objects are masked from 'package:utils':
##
##     history, upgrade

## The following objects are masked from 'package:base':
##
##     autoload, load, remove, use
```