Principal Component Analysis on U.S. Gini Index Group 10 - Assignment 2

Alexandra Goh
agoh0008@student.monash.edu

Evan Ginting
egin0003@student.monash.edu

Huyen-Anh Pham
hpha0042@student.monash.edu

Pei Ni Lim
plim0015@student.monash.edu

Report for
ETF5500 High Dimensional Data Analysis

22 September 2023

MONASH
BUSINESS
SCHOOL

**Department of
Econometrics &
Business Statistics**

☎ (03) 9905 2478
✉ BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

Contents

1 Introduction

This report examines the Gini Index, a vital measure of income inequality, across all 52 U.S. states, as documented by Farris (2010). Our analysis investigates two significant historical periods: the era of the Great Depression and World War II (WWII) (1928 - 1945) and the aftermath of the Global Financial Crisis (GFC) (2007 – 2015). The report's objective is to understand how these historical events influenced income inequality trends among the states.
The first section includes data cleaning, summarising, and data visualisation. The second part focuses on Principal Component Analysis (PCA) as the primary dimensionality reduction method, with the addition of cluster analysis and Multidimensional Scaling (MDS).

2 Preliminary Analysis

2.1 Data

Our processed dataset comprises of three variables: 'State name', 'Year', and the 'Gini Index'. According to the Bureau (2021), the Gini Index operates within a value range of 0 to 1, where 0 implies a state of perfect income equality, while 1 indicates perfect income inequality. To facilitate a rigorous comparative analysis of individual states, we deliberately excluded observations pertaining to the mean Gini Index of the "United States" in our dimension reduction procedures.

Within our dataset, a distinctive outlier value emerged in the Gini Index for Oregon in 1934, and we encountered a total of 10 missing values in the year 2010. In addressing these data gaps, we navigated through various missing values handling techniques, such as mean imputation and deletion. However, we opted for a more strategic approach by substituting the missing values with accurate data sourced from Farris (2015) - the research website, thereby preserving the integrity of the dataset to the greatest extent possible.

2.2 Exploratory Data Analysis

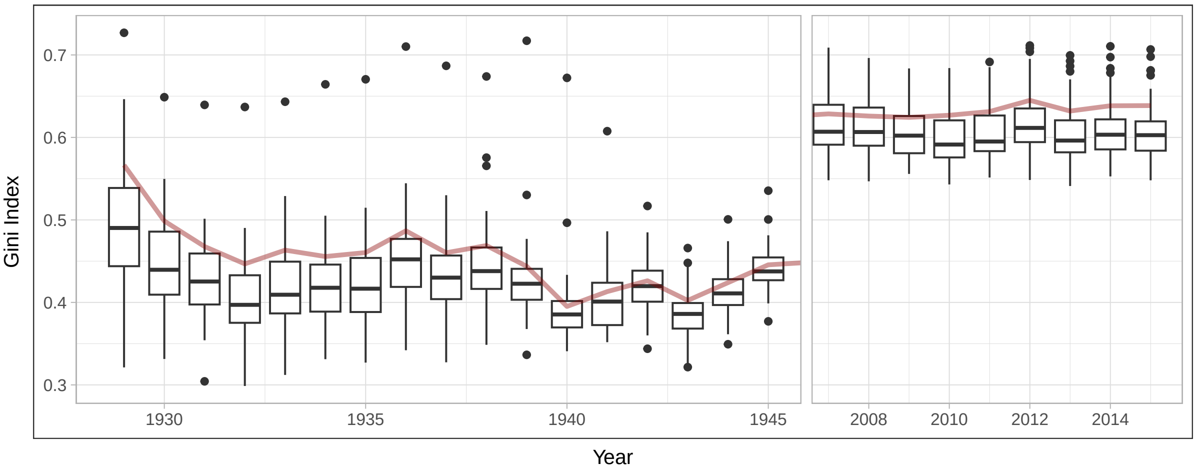


Figure 1: *Distribution of Gini Index over the year*

Figure 1 illustrates the distribution of the Gini Index over the years (boxplots), accompanied by a comparison to the U.S.'s average Gini Index (red line). Our plot reveals an apparent decreasing trend in Gini Index during the initial period from 1929 to 1945, while there is a sustained presence of relatively high Gini Index levels during the subsequent period.

The boxplot also provides insights into the distribution characteristics of each year. The distributions display a degree of symmetry, accompanied by a slight right-skewness as the mean is marginally larger than the median in certain years. We observe that during the Great Depression and WWII, the variance of the Gini Index experiences a significant reduction, while it remains rather stable across the GFC period.

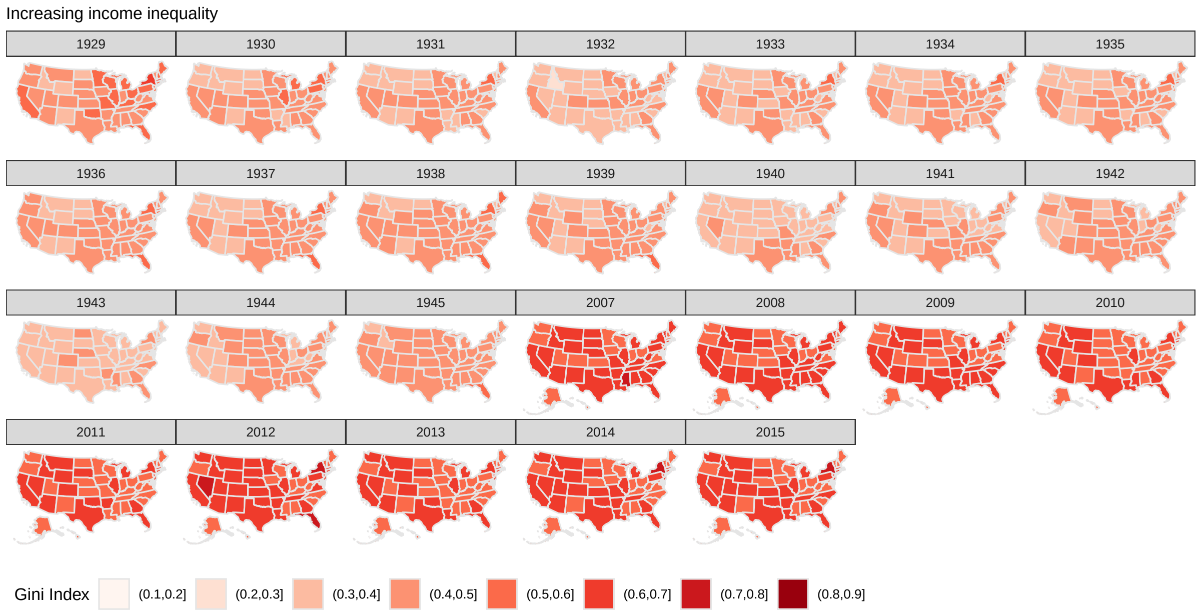


Figure 2: *Gini Index across U.S. states over time*

Figure 2 visually illustrates the link between geographic location and income inequality across U.S. states. Initially, North-Eastern metropolitan states like New York, Massachusetts, Connecticut, Delaware, and Florida consistently displayed higher income inequality. Factors contributing to this include their status as economic hubs, especially in the finance sector, which offers high-paying jobs that contribute to income disparities. Additionally, the high cost of living exacerbates income gaps, particularly for lower-income individuals. Significant migration patterns involving both high and low-skilled workers further contribute to income disparities. In the subsequent period, income inequality variance among states reduced, coinciding with a nationwide worsening of income inequality. Nevertheless, New York and Florida remained notable for higher income inequality.

3 Analysis

3.1 Assumptions

There are some assumptions that we employ when performing the analysis:

- We did not scale the data since the Gini Index already has consistent units of measurement.
- The Euclidean method was used to calculate distances, capturing differences in Gini index values.
- Hierarchical clustering was chosen due to the lack of prior data characteristics, allowing us to determine the appropriate cluster number during analysis.
- The D2.Ward method was employed to minimise within-cluster variance and create compact clusters, with later assessments of its robustness and comparison with alternative methods.
- We reduced dimensions to two in MDS for enhanced interpretability and visualization.

3.2 Cluster Analysis

3.2.1 Period 1

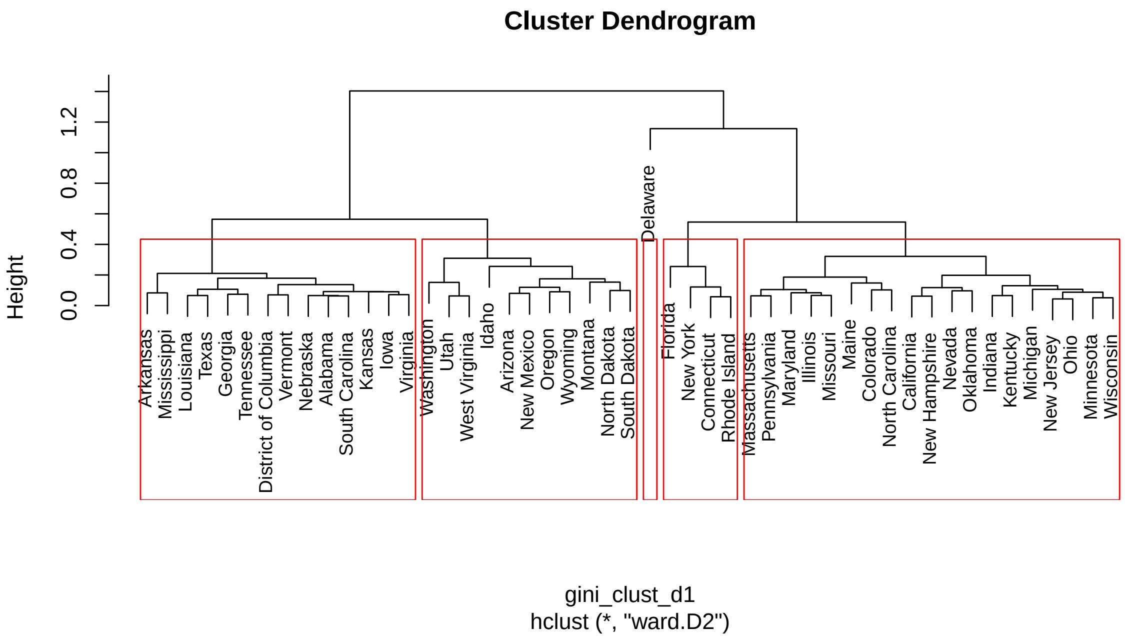


Figure 3: *Cluster Analysis of the 1st Period (1929 - 1945)*

From Figure 3, a stable solution with five clusters is observed within a tolerance range of 0.35 to 0.44. This configuration provides adequate grouping without sacrificing the detail within each cluster.

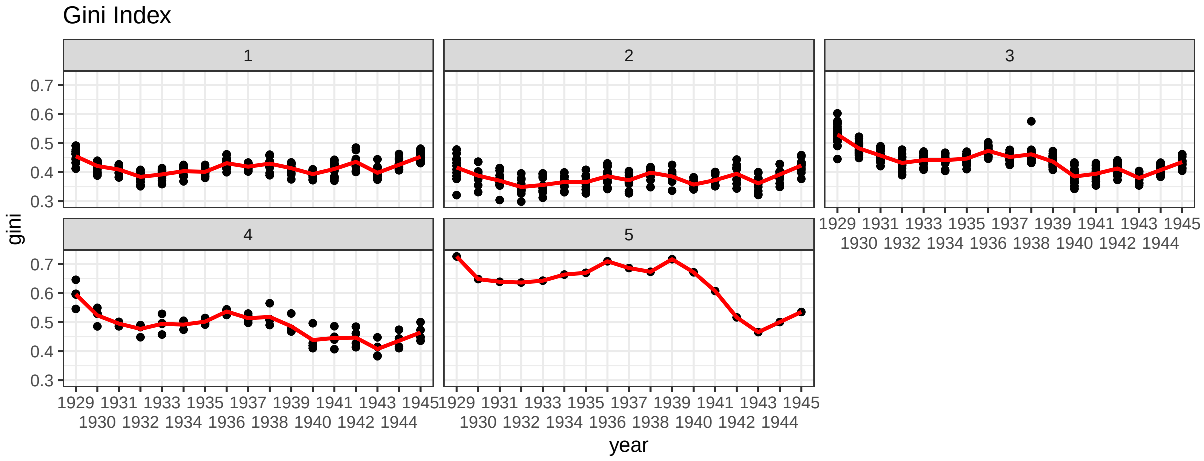


Figure 4: *Cluster Profiles for the 1st Period: Five-cluster Solution*

Figure 4 visually illustrates the Gini Index for each cluster, along with an average line. Notably, during the Great Depression and WWII era, Delaware, part of cluster five, exhibited significantly higher inequality compared to the others, making it stand out.

Lastly, we assess the robustness of Ward's method against others by calculating the Adjusted Rand Index. Table 5 in the appendix shows that Ward's method exhibits fairly strong robustness compared to alternative methods.

3.2.2 Period 2

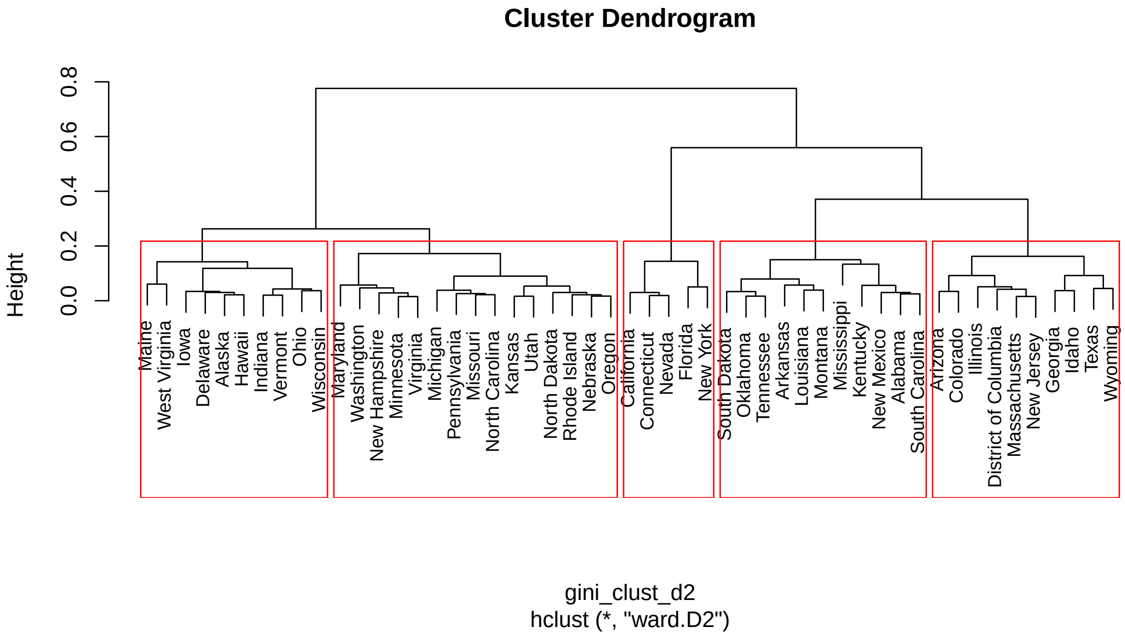


Figure 5: *Cluster Analysis of the 2nd Period (2007 - 2015)*

Figure 5 also portrays a stable five-cluster solution within a tolerance range of 0.18 to 0.3.

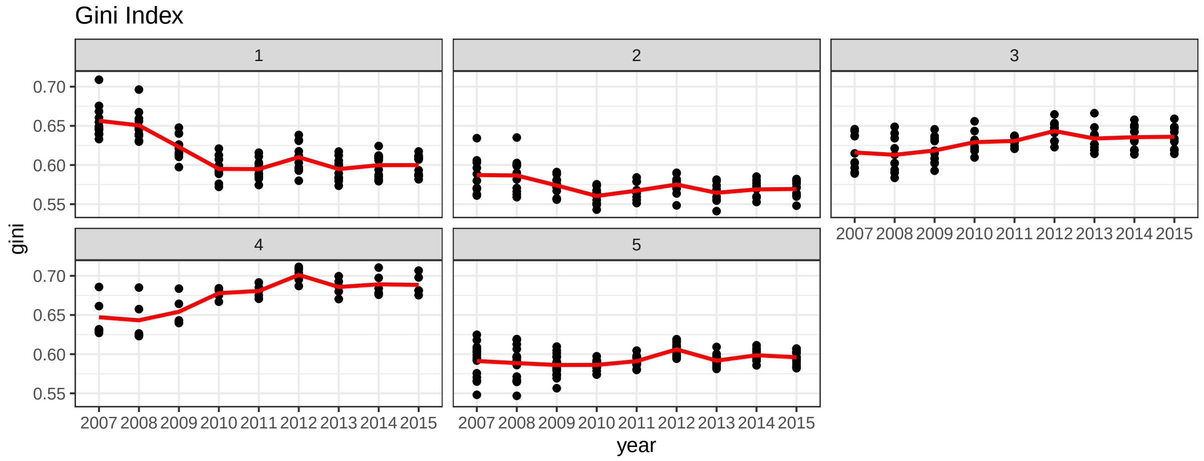


Figure 6: *Cluster Profiles for the 2nd Period: Five-cluster Solution*

Figure 6 highlights cluster four as having the highest inequality. Table 1 provides a list of states within this cluster.

Table 1: *Cluster 4: states with the highest Gini Index*

Cluster	State	2007	2008	2009	2010	2011	2012	2013	2014	2015
4	California	0.630	0.626	0.640	0.667	0.671	0.687	0.670	0.678	0.675
4	Connecticut	0.627	0.624	0.640	0.675	0.675	0.695	0.680	0.684	0.681
4	Florida	0.686	0.685	0.684	0.682	0.685	0.711	0.693	0.697	0.698
4	Nevada	0.632	0.623	0.643	0.682	0.681	0.704	0.686	0.676	0.681

Cluster	State	2007	2008	2009	2010	2011	2012	2013	2014	2015
4	New York	0.661	0.657	0.664	0.684	0.691	0.708	0.699	0.710	0.707

Regarding the robustness analysis, as shown in Table 6 in the Appendix, we can conclude that Ward's method is relatively robust compared to other methods.

3.3 Multidimensional Scaling (MDS)

3.3.1 Period 1

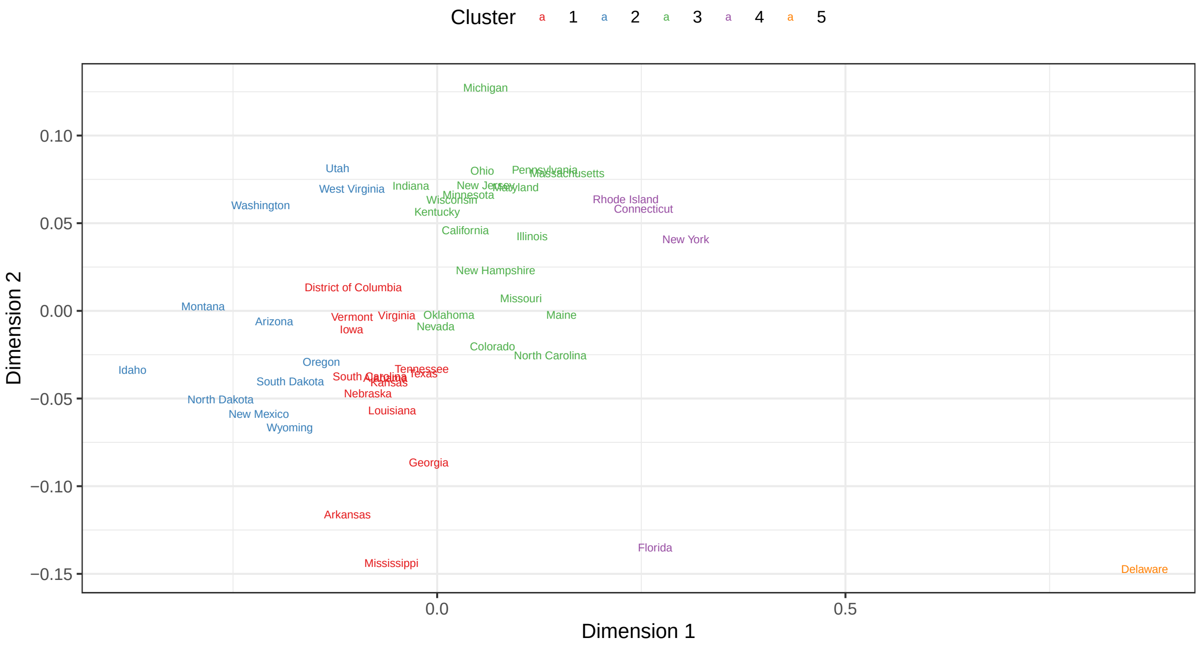


Figure 7: *MDS by U.S. state for the 1st period (1929 - 1945)*

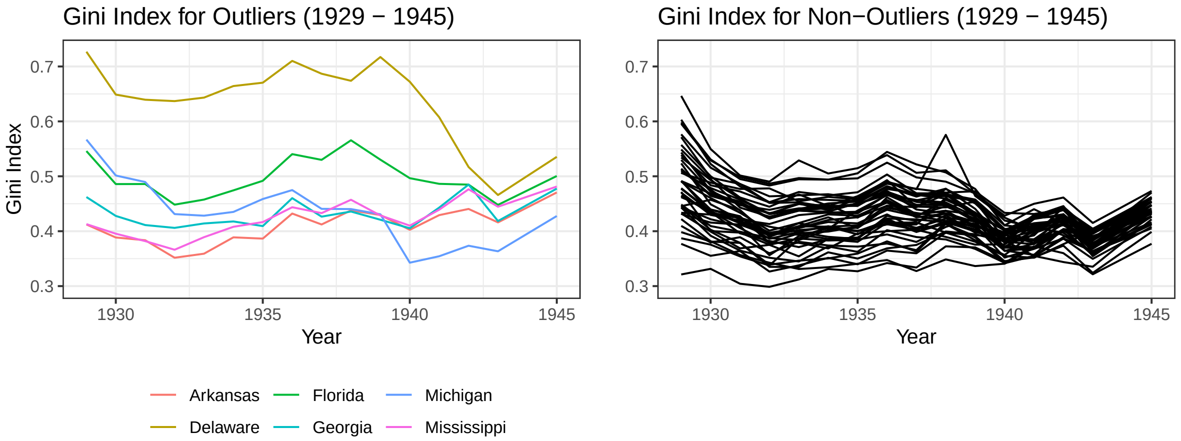


Figure 8: *Gini Index of the outlier states between 1929 and 1945*

Figure 7 reveals that states cluster near the center of the plot on both axes, suggesting lower Gini indices during the Great Depression/WWII in comparison to states located further from the center.

3.3.2 Period 2

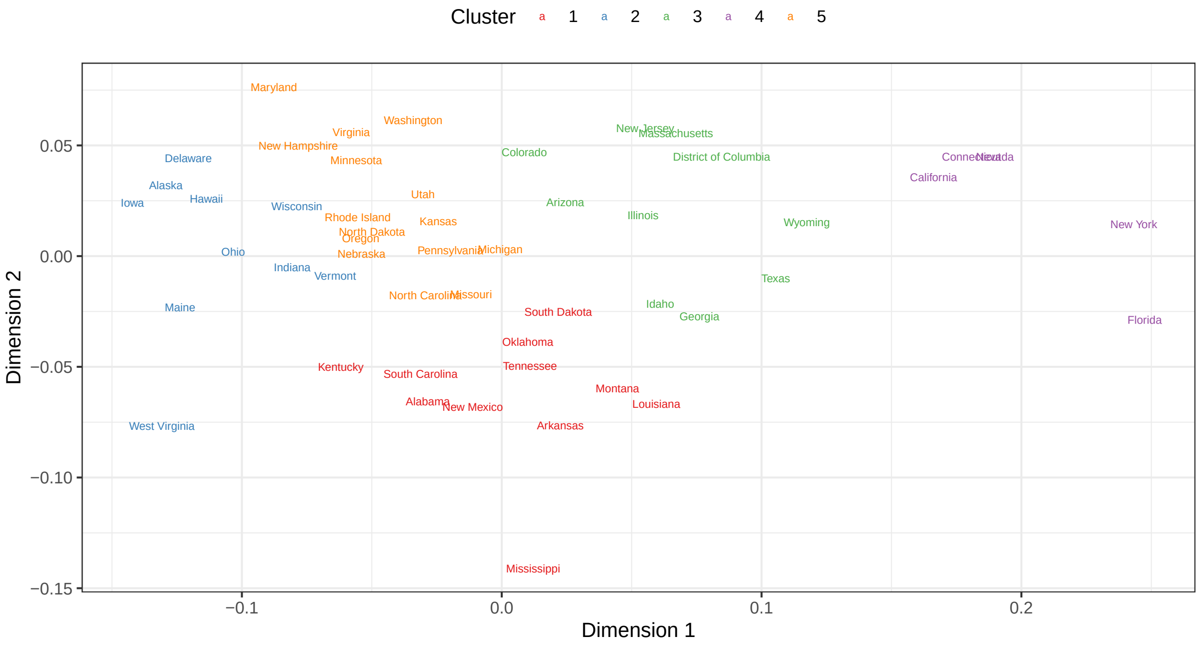


Figure 9: *MDS by U.S. state for the 2nd period (2007 - 2015)*

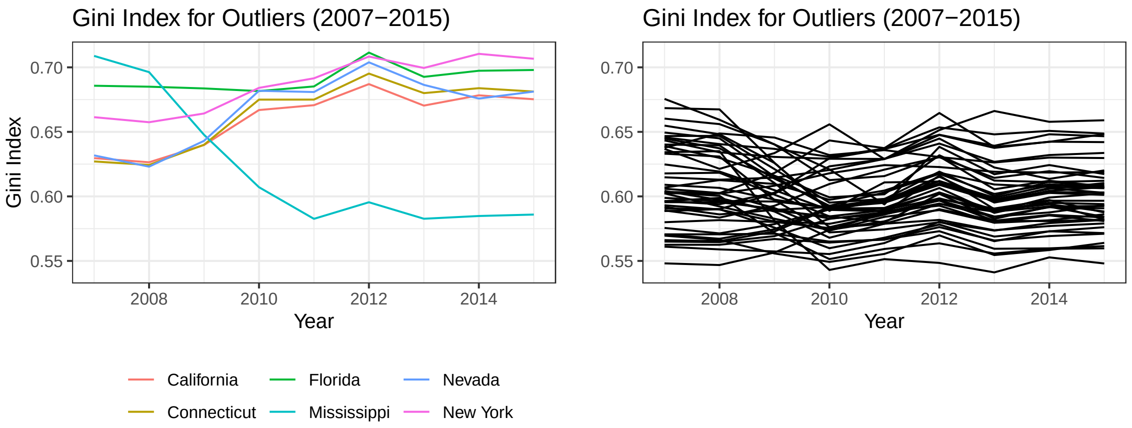


Figure 10: *Gini Index of the outlier states between 2007 and 2015*

Figure 9 reveals increased state dispersion away from the origin, implying higher income inequality during the GFC compared to the Great Depression/WWII era. In particular, New York, Florida, and Mississippi are outliers, while California, Connecticut, and Nevada cluster near New York's position.

3.4 Principal Component Analysis (PCA)

3.4.1 Period 1

3.4.1.1 Kaiser's rule

Table 2: *Importance of components - 1st Period*

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.5896	1.6142	0.7831	0.5699	0.3549	0.3501	0.2993	0.2405	0.2173	0.2005
Proportion of Variance	0.7580	0.1533	0.0361	0.0191	0.0074	0.0072	0.0053	0.0034	0.0028	0.0024
Cumulative Proportion	0.7580	0.9112	0.9473	0.9664	0.9738	0.9810	0.9863	0.9897	0.9925	0.9948

According to Table 2, following Kaiser's rule, we select 2 principal components (PCs) since both the variance and standard deviation are greater than 1, and the cumulative proportion of variance is 91.12%.

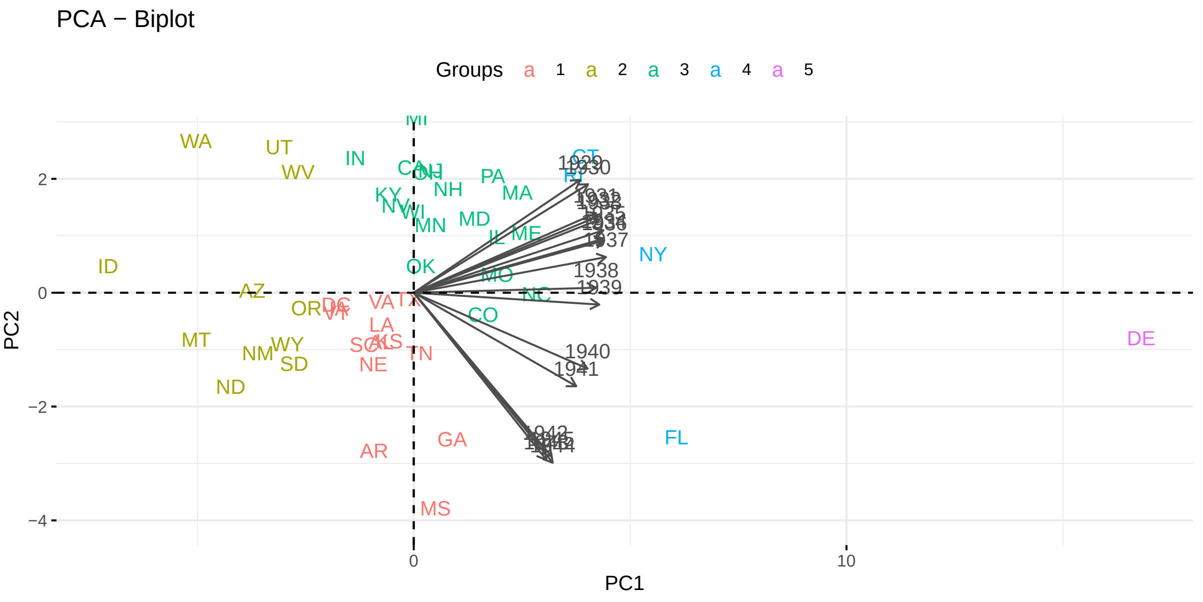


Figure 11: *Distance Biplot - PCA of the 1st period*

3.4.1.2 Distance bi-plot

In Figure 11, the years 1937-1939 closely align with PC1, coinciding with the start of WWII. Due to increased military spending, Pells and Romer (2023) shows a 9% GDP growth from 1933 to 1937 whilst Waiwood (2013) notes a 10% drop in unemployment rates. Therefore, PC1 characterises higher GDP and lower unemployment, representing the US's recovery from the Great Depression.

Delaware's (DE) high PC1 values, indicating a strong GDP and employment rate at the start of WWII, are linked to its key role in gunpowder manufacturing and shipyards (Rowe (1980)). However, this led to increased income inequality as top industry leaders prospered. The influx of migrants for jobs during WWII further widened income disparities due to limited opportunities and lower wages for these migrants. As a result, Delaware had the highest Gini index, with its top earners making two to three times more than their counterparts in other states (Schmitz and Fishback (1983); Figure 8)

Meanwhile, Idaho (ID) has low PC1 values, indicating low average real GDP and employment rates at the onset of WWII. According to Boyd A. Martin (n.d.), this is unsurprising given Idaho's heavy reliance on agriculture and limited industrial capacity in the 1930s. This economic dependence on agriculture, coupled with falling food prices and reduced agricultural income during the Great Depression, likely worsened its income inequality.

Figure 11 associates high PC2 values with GDP decline and unemployment surge during the Great Depression (1929-1933), while lower PC2 values correspond to economic growth and job increase in WWII's later years (1942-1945). As noted by Duignan (2018), US' industrial production fell by nearly 47% between 1929 and 1933, leading to a 30% GDP reduction and unemployment rates peaking at over 20%. In contrast, extensive military production drove up US' real GDP by 72% from 1940-1945, with unemployment rates falling from 9.5% in 1940 to below 2% from 1943-1945 (Fishback (2019)).

Mississippi's (MS) low PC2 values reflect the significant economic benefits reaped during WWII. As per Skates (2017), income per capita increased substantially by 11%, and urban employment opportunities grew as the farm population decreased by 26%, with many joining the military.

States such as Tennessee, Alabama, Kansas, Nebraska, Louisiana, and Arkansas experienced lower PC2 values during the 1929-1933 period. These Southern and Midwestern states faced economic difficulties due to widespread poverty and heavy reliance on farming (Davis (1978)). Roosevelt's New Deal policies played a pivotal role in revitalising agriculture, stimulating economic growth and creating jobs (Nourse (1936)), which aided these regions in recovering from the Great Depression and stabilising income inequality.

3.4.1.3 Correlation bi-plot

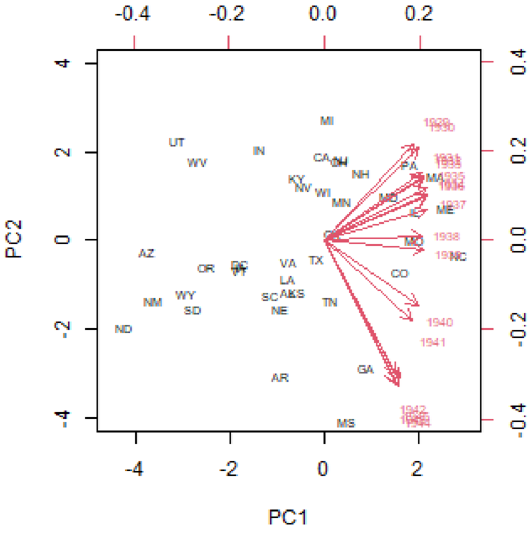


Figure 12: *Correlation Biplot - PCA of the 1st period*

Based on Figure 12, some years show little correlation. For example, the years 1929 and 1941, as well as 1934 and 1942, have angles close to 90 degrees between them. This lack of correlation is because the Great Depression lasted from 1929 to 1939, while WWII occurred from 1939 to 1945. In contrast, 1929 and 1945 have a strong negative correlation, with an angle close to 180 degrees, as the former marks the start of the Great Depression and the latter the end of WWII. These correlations reflect the distinct historical contexts and economic conditions of these years.

3.4.2 Period 2

3.4.2.1 Kaiser's rule

Table 3: *Importance of components - 2nd Period*

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.6752	1.3084	0.2128	0.1829	0.1613	0.1121	0.0835	0.0742	0.0401
Proportion of Variance	0.7952	0.1902	0.0050	0.0037	0.0029	0.0014	0.0008	0.0006	0.0002
Cumulative Proportion	0.7952	0.9854	0.9904	0.9942	0.9970	0.9984	0.9992	0.9998	1.0000

According to Table 3, we choose two PCs based on Kaiser's rule, as both the variance and standard deviation are greater than 1, and the cumulative proportion of explained variance is 98.54%.

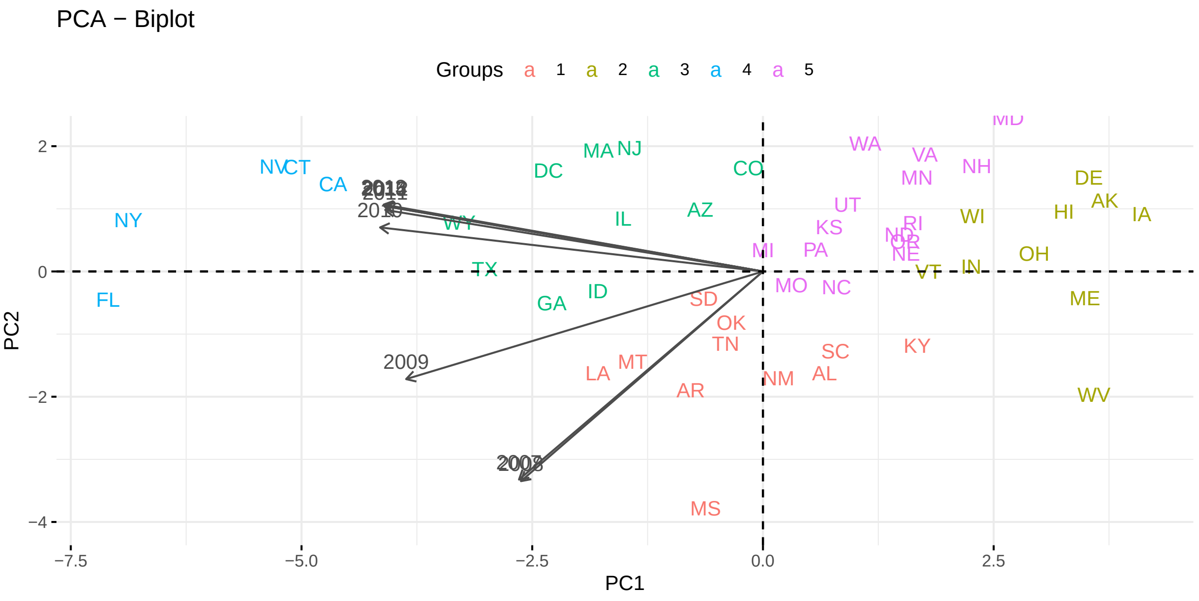


Figure 13: *Distance Biplot - PCA of the 2nd period*

3.4.2.2 Distance bi-plot

Figure 13 displays a distance biplot, highlighting the association between PC2 and the GFC, while PC1 corresponds to the 'recovery period.' According to data from Bank (2023a), the U.S. experienced a GDP decline during the GFC, reflecting the economic downturn characterized by business struggles, reduced consumer spending, and shrinking investments. Subsequent GDP growth indicates a shift away from recession.

The employment rate in the U.S. significantly dropped in the early 2007 period and began a gradual increase (Bank (2023c)). The GFC's global impact exacerbated the situation, with decreased global demand for U.S. exports contributing to rising unemployment. Additionally, a decline in inflation during 2007-2008 (Bank (2023b)) can be attributed to the GFC, as economic contraction led to reduced demand, business challenges, and job losses.

The biplot highlights significant income inequality in states like Florida, New York, and Nevada, where affluent individuals tend to cluster in urban areas such as Miami, New York City, and Las Vegas. This concentration of wealth resulted in high Gini Index values during and after the GFC. In Florida and New York, the richest 1% of households accounted for a substantial 61% of the adjusted gross income, as indicated by CBPP (2012).

California and Texas were also heavily affected by the GFC, primarily due to their reliance on manufacturing and construction, as reported by U.S. Bureau of Labor Statistics (2010). In contrast, Delaware showed a declining Gini Index during both the GFC and the recovery period. This trend might be attributed to its progressive tax rates, potentially leading to wealthier households migrating out, as suggested by Vuoccolo (2018).

3.4.2.3 Correlation bi-plot

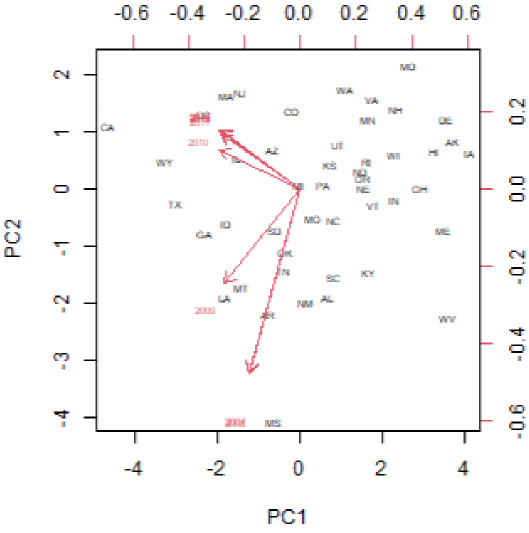


Figure 14: *Correlation Biplot - PCA of the 2nd period*

In Figure 14, strong positive correlations are observed for specific years, notably 2007 and 2008, reflecting the impact of the GFC. 2011-2015 exhibit similar patterns, likely associated with economic recovery. However, 2009 and 2012 appear uncorrelated, indicating the U.S. transition from recession to growth. Notably, there are no significant negative correlations among the years.

4 Limitations

Several limitations that surfaced during our analysis include:

Cluster Analysis

- Optimal cluster number determination can be subjective and challenging due to methodological variations.
- Outliers and data noise can impact cluster quality and sensitivity.

Multidimensional Scaling (MDS)

- MDS reveals data similarities/differences but does not offer causal explanations for state-specific income inequality patterns.
- MDS may not capture all nuances in the original data, potentially leading to information loss. To address this, we calculated Goodness of Fit (GOF) values, which were approximately 0.934 in period 1 and 0.985 in period 2 (see Appendix's Table 5). These indicate varying degrees of alignment between MDS plots and the original data, with period 2 showing a higher degree of accuracy.

Principal Component Analysis

Interpreting PCA results can be challenging without sufficient domain knowledge. To address this, we conducted extensive research on relevant historical events to provide context and support our interpretation of the PCA findings.

5 Conclusion

Our analysis of income inequality during the Great Depression, WWII, and the GFC revealed significant findings:

The Great Depression/WWII

- A five-cluster approach effectively groups states for detailed analysis.
- States closer to the center in our MDS analysis show lower income inequality during these periods.
- PC1 reflects higher GDP and lower unemployment, indicating recovery from the Great Depression.
- PC2 strongly correlates with economic conditions during these events.

The Global Financial Crisis (GFC)

- A five-cluster solution is again useful for meaningful groupings.
- States shifted away from the center in our MDS analysis, indicating increased income inequality overall compared to the previous era.
- PC1 aligns with the GFC recovery period (2009–2015), while PC2 relates to the GFC itself (2007 - 2008), reflecting changes in US' GDP.

In summary, our analysis provided valuable insights into income inequality during critical historical events. To enhance our understanding, we recommend incorporating Factor Models and additional economic indicators. This will help us better interpret the complex economic dynamics shaped by these significant events.

6 Appendix

6.1 Appendix 1: Analysis

Table 4: *Summary by Year*

Year	Min	1st Qu.	Median	3rd Qu.	Max
1929	0.321	0.444	0.490	0.539	0.727
1930	0.331	0.409	0.440	0.486	0.649
1931	0.304	0.397	0.425	0.459	0.639
1932	0.299	0.375	0.397	0.433	0.637
1933	0.312	0.387	0.409	0.449	0.643
1934	0.331	0.389	0.418	0.446	0.664
1935	0.327	0.388	0.417	0.454	0.670
1936	0.342	0.419	0.452	0.477	0.710
1937	0.327	0.404	0.430	0.457	0.687
1938	0.349	0.416	0.438	0.467	0.674
1939	0.336	0.403	0.423	0.441	0.717
1940	0.341	0.370	0.385	0.402	0.672
1941	0.352	0.372	0.401	0.424	0.608
1942	0.344	0.401	0.420	0.438	0.517
1943	0.322	0.368	0.386	0.399	0.466
1944	0.349	0.397	0.411	0.428	0.501
1945	0.377	0.427	0.438	0.455	0.535
2007	0.548	0.591	0.607	0.640	0.709
2008	0.547	0.590	0.607	0.636	0.696
2009	0.556	0.581	0.602	0.626	0.684
2010	0.543	0.576	0.591	0.621	0.684
2011	0.551	0.583	0.595	0.627	0.691
2012	0.548	0.594	0.612	0.635	0.711
2013	0.541	0.582	0.596	0.621	0.699
2014	0.553	0.585	0.603	0.622	0.710
2015	0.548	0.584	0.603	0.619	0.707

6.1.1 Cluster Analysis

Table 5: *Evaluating the robustness of the model - 1st Period*

	Average Linkage	Centroid	Complete Linkage
Ward's Method	0.715	0.203	1

Table 6: *Evaluating the robustness of the model - 2st Period*

	Average Linkage	Centroid	Complete Linkage
Ward's Method	0.63	0.198	0.395

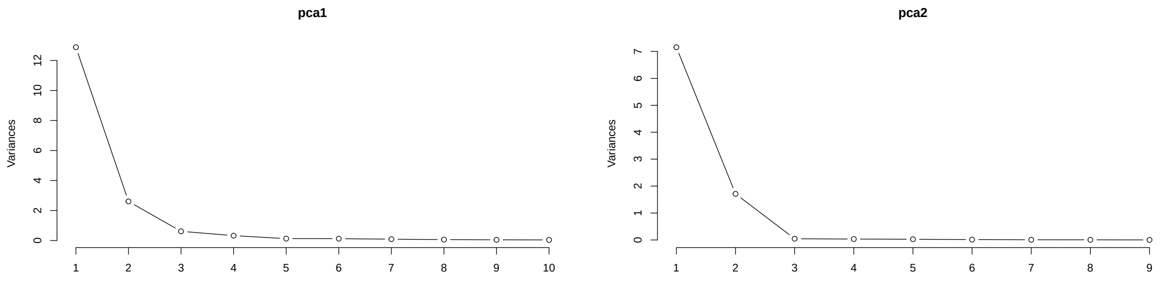
6.1.2 MDS

Table 7: *MDS Goodness of Fit*

Period	GOF1	GOF2
1st	0.9340	0.9340
2nd	0.9855	0.9855

6.1.3 PCA

Scree plot



The Scree plot of both periods show that the “elbow point” is in the third point. These three point is sufficient for dimensionality reduction.

Weights of the first two PCs

1st Period

Table 8: *Weights of the first two PCs - 1st Period*

	PC1	PC2
1929	0.2350	0.2690
1930	0.2459	0.2586
1931	0.2571	0.1897
1932	0.2604	0.1821
1933	0.2617	0.1744
1934	0.2681	0.1289
1935	0.2674	0.1489
1936	0.2688	0.1243
1937	0.2711	0.0848
1938	0.2570	0.0118
1939	0.2617	-0.0283
1940	0.2452	-0.1810
1941	0.2292	-0.2230
1942	0.1856	-0.3754
1943	0.1871	-0.3985
1944	0.1960	-0.4054
1945	0.1944	-0.3902

Based on Table 8, in 1937, PC1's high weight suggests a strong link between that year and PC1. PC2's high weight for 1929 indicates a strong connection between 1929 and PC2. The Great Depression, starting in 1929, had a significant impact on 1937, likely due to enduring economic hardships, including high unemployment and income disparities. From 1939 to 1945, PC2 consistently had negative values, implying reduced income inequality in other years, likely driven by increased job opportunities and higher incomes during WWII (1939-1945).

2nd Period

Table 9: *Weights of the first two PCs - 2nd Period*

	PC1	PC2
2007	-0.2318	-0.5958
2008	-0.2299	-0.6004
2009	-0.3391	-0.3082
2010	-0.3638	0.1260
2011	-0.3589	0.1762
2012	-0.3586	0.1868
2013	-0.3603	0.1915
2014	-0.3594	0.1895
2015	-0.3601	0.1874

Based on Table 9, negative PC1 values imply decreased income inequality, particularly during the GFC, while positive PC1 values highlight the significant impact of the crisis on income disparities. The shift from negative PC2 values (2007-2009) to positive (2010-2015) suggests changing income inequality patterns, possibly influenced by the Great Recession and subsequent economic factors.

6.2 Appendix 2: Code

Data Cleaning & Preliminary Analysis

```
# DATA CLEANING ------------------------------------
# Get Data
gd <- read_csv(here::here("data/Inequality_GD.csv"))
gr <- read_csv(here::here("data/Inequality_GR.csv"))

# Bind two raw datasets
gini <- rbind(
  gd %>%  dplyr::select(-"...1") %>%
    pivot_longer(-State, names_to = "year", values_to = "gini"),
  gr %>% dplyr::select(-"...1")%>%
    pivot_longer(-State, names_to = "year", values_to = "gini")
) %>%
  rename("state_name" = "State") %>%
  mutate(year = as.numeric(year))

# Add state abbreviation
state <- tibble(state_name = c(unique(gini$state_name)))
state_abb <- data.frame(state_name = state.name, state = state.abb)
state <- left_join(state, state_abb, by = c("state_name" = "state_name"))
state <- state %>%
  mutate(state = case_when(
    state_name == "United States" ~ "US",
    state_name == "District of Columbia" ~ "DC",
    TRUE ~ state))
gini <- left_join(gini, state, by = c("state_name" = "state_name"))

# Get updated dataset from research website
updated <- read_csv(here::here("data/Frank_Gini_2018.csv")) %>%
  dplyr::select(State,Year,Gini) %>%
  rename_all(tolower) %>%
  rename(state_name = state, gini_updated = gini)

# Handle Outlier and Missing Values
gini <- gini %>%
  left_join(., updated, by = c("year"="year","state_name"="state_name")) %>%
  mutate(gini = case_when(
    state_name=="Oregon" & year==1934 ~ gini_updated, # Outlier
    is.na(gini) ~ gini_updated,                       # Missing Values
    TRUE ~ gini)) %>%
  dplyr::select(1:4)

# DATA WRANGLING ------------------------------------
## Subset data
gini1 <- gini %>%
  select(state_name, state, year, gini) %>%
  filter(year %in% 1929:1945,
         state_name != "United States")
gini2 <- gini %>%
  select(state_name, state, year, gini) %>%
  filter(year %in% 2007:2015,
         state_name != "United States")

## Transforming the table into wide format
gini_period1 <- gini %>%
  select(state_name, state, year, gini) %>%
  filter(year %in% 1929:1945,
         state_name != "United States") %>%
  pivot_wider(names_from = year,
              values_from = gini)

gini_period2 <- gini %>%
  select(state_name, state, year, gini) %>%
  filter(year %in% 2007:2015,
         state_name != "United States") %>%
  pivot_wider(names_from = year,
              values_from = gini)

# Assigning first column as the rownames (for visual purpose)
rownames(gini_period1) <- gini_period1$state_name
rownames(gini_period2) <- gini_period2$state_name
```

```
# Box Plot: Distribution of Gini Index over the year
gini %>%
  filter(state != "US") %>%
  ggplot(aes(x = year)) +
  geom_boxplot(aes(y = gini, group = year)) +
  geom_line(data = gini %>% filter(state == "US"),
            aes(y = gini, group = 1),
            color = "darkred", size = 1.2, alpha = 0.4) +
  scale_x_break(c(1945,2007)) +
  theme(axis.title = element_text(size = 8),
        legend.position = "top") +
  theme_light() +
  labs(x = "Year", y = "Gini Index")
```

```
# Choropleth Map: Gini Index across U.S. states over time
df <- gini %>%
  select(state, year, gini) %>% filter(state != "US") %>%
  mutate(gini_cut = cut(gini,breaks = c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)))

# Map Plot: Gini Index across states over time
plot_usmap(data = df, values = "gini_cut", color = "grey90") +
  scale_fill_brewer(palette = "Reds", drop = FALSE, name = "Gini Index") +
  theme(legend.position = "bottom", aspect.ratio = 5.5/9) +
  facet_wrap(~year, ncol=7) +
  guides(fill=guide_legend(nrow = 1)) +
  labs(subtitle = "Increasing income inequality")
```

Cluster Analysis

```
# CLUSTER ANALYSIS 1  ------------------------------------
## Calculating the distance matrix using euclidean distance
gini_clust_d1 <- gini_period1 %>%
  dist(method = 'euclidean')
## Hierarchical clustering
gini_clust_v1 <- hclust(gini_clust_d1, method = "ward.D2")
# Plotting the Cluster Dendrogram
plot(gini_clust_v1, cex = 0.8)
rect.hclust(gini_clust_v1, k = 5, border = "red")
# Store Ward.D2 cluster membership in a new variable.
gini_v1_ward <- cutree(gini_clust_v1, k = 5)
# Other cluster methods
gini_v1_al <- hclust(gini_clust_d1, method = 'average') %>% cutree(k = 5)
gini_v1_cm <- hclust(gini_clust_d1, method = 'centroid') %>% cutree(k = 5)
gini_v1_cl <- hclust(gini_clust_d1, method = 'complete') %>% cutree(k = 5)
# Map cluster membership to dataset
gini_period1 <- gini_period1 %>% mutate(cluster = gini_v1_ward)
gini_clustered_v1 <- gini_period1 %>%
  mutate(cluster = gini_v1_ward) %>%
  pivot_longer(-c(state_name,state, cluster),names_to = "year",values_to = "gini")
# Period 1: Plot by Cluster
gini_clustered_v1 %>%
  ggplot(aes(x = year, y = gini)) +
  geom_point() +
  stat_summary(mapping=aes(x=year,y=gini,group = 1),fun="mean", geom="line", color="red",size = 1) +
  ggtitle("Gini Index") +
  facet_wrap(~cluster)
```

```
# CLUSTER ANALYSIS 2  ------------------------------------
## Calculating the distance matrix using euclidean distance
gini_clust_d2 <- gini_period2 %>% dist(method = 'euclidean')
## Hierarchical clustering
gini_clust_v2 <- hclust(gini_clust_d2, method = "ward.D2")
# Plotting the Cluster
plot(gini_clust_v2, cex = 0.8)
rect.hclust(gini_clust_v2, k = 5, border = "red")
# Store Ward.D2 cluster membership in a new variable.
gini_v2_ward <- cutree(gini_clust_v2, k = 5)
# Other three methods
gini_v2_al <- hclust(gini_clust_d2, method = 'average') %>% cutree(k = 5)
gini_v2_cm <- hclust(gini_clust_d2, method = 'centroid') %>% cutree(k = 5)
gini_v2_cl <- hclust(gini_clust_d2, method = 'complete') %>% cutree(k = 5)
# Assigning the cluster
gini_period2 <- gini_period2 %>% mutate(cluster = gini_v2_ward)
gini_clustered_v2 <- gini_period2 %>%
  pivot_longer(-c(state_name, state, cluster), names_to = "year", values_to = "gini")
# Plotting the cluster profile (with average line)
gini_clustered_v2 %>%
  ggplot(aes(x = year, y = gini)) +
  geom_point() +
  stat_summary(mapping=aes(x=year,y=gini,group = 1),fun="mean", geom="line", color="red",size = 1) +
  ggtitle("Gini Index") +
  facet_wrap(~cluster)
```

MDS

```
### MDS 1  ------------------------------------
d1 <- gini_period1 %>% dplyr::select("1929":"1945") %>% dist(method = 'euclidean')
attributes(d1)$Labels <- gini_period1$state_name
cmds1 <- cmdscale(d1, eig = T)
df1 <- cmds1$points %>%
  as.data.frame() %>%
  rownames_to_column(var = 'state_name')
## Plot MDS 1 by State, with Cluster Membership
ggplot(df1, aes(x = V1, y = V2, label = `state_name`, color = as.factor(gini_period1$cluster))) +
  geom_text(size = 2) +
  scale_color_brewer(type = "div", palette = "Set1",name = "") +
  labs(x = "Dimension 1", y = "Dimension 2") +
  guides(color = guide_legend(title = "Cluster")) +
  theme(legend.position = "top")
## Side-by-side Line Plots
common_y_limits <- c(min(min(gini1$gini), min(gini1$gini)),
                     max(max(gini1$gini), max(gini1$gini)))

p1 <- ggplot(
  data = gini1 %>%
    filter(state_name %in% c("Delaware", "Michigan", "Georgia", "Arkansas", "Mississippi", "Florida")),
  aes(x = year, y = gini, col = state_name)) +
  geom_line() +
  labs(title = "Gini Index for Outliers (1929 - 1945)",
       x = "Year", y = "Gini Index") +
  scale_color_discrete(name = NULL) +
  guides(color = guide_legend(title = NULL)) +
  ylim(common_y_limits) +
  theme(legend.position = "bottom",aspect.ratio = 2/4)

p2 <- ggplot(
  data = gini1 %>%
    filter(!state_name %in% c("Delaware", "Michigan", "Georgia", "Arkansas",  "Mississippi", "Florida")),
  aes(x = year, y = gini, group = state_name), color="grey70") +
  geom_line() +
  labs(title = "Gini Index for Non-Outliers (1929 - 1945)",
       x = "Year", y = "") +
  guides(color = guide_legend(title = NULL)) +
  ylim(common_y_limits) +
  theme(legend.position = "none",aspect.ratio = 2/4)

p1 + p2 +
  plot_layout(ncol = 2, guides = "keep") &
  theme(legend.position = "bottom")
```

```
# MDS 2  ------------------------------------
d2 <- gini_period2 %>% dplyr::select("2007":"2015") %>% dist(method = 'euclidean')
attributes(d2)$Labels <- gini_period2$state_name
cmds2 <- cmdscale(d2, eig = T)
df2 <- cmds2$points %>%
  as.data.frame() %>%
  rownames_to_column(var = 'state_name')

## Plot MDS by State, with Cluster Membership
ggplot(df2, aes(x = V1, y = V2, label = `state_name`, color = as.factor(gini_period2$cluster))) +
  geom_text(size = 2) +
  scale_color_brewer(type = "div", palette = "Set1",name = "") +
  labs(x = "Dimension 1", y = "Dimension 2") +
  guides(color = guide_legend(title = "Cluster")) +
  theme(legend.position = "top")

## Side-by-side Line Plots
common_y_limits1 <- c(min(min(gini2$gini), min(gini2$gini)),
                      max(max(gini2$gini), max(gini2$gini)))

p3 <- ggplot(
  data = gini2 %>%
    filter(state_name %in% c("New York", "Florida","Mississippi", "California", "Connecticut", "Nevada")),
  aes(x = year, y = gini, col = state_name)) +
  geom_line() +
  labs(title = "Gini Index for Outliers (2007-2015)",
       x = "Year", y = "Gini Index") +
  scale_color_discrete(name = NULL) +
  guides(color = guide_legend(title = NULL)) +
  ylim(common_y_limits1)  +
  theme(legend.position = "bottom",aspect.ratio = 2/4)

p4 <- ggplot(
  data = gini2 %>%
    filter(!state_name %in% c("New York", "Florida", "Mississippi", "California", "Connecticut", "Nevada")),
  aes(x = year, y = gini, group = state_name), color="grey70") +
  geom_line() +
  labs(title = "Gini Index for Outliers (2007-2015)",
       x = "Year", y = "") +
  guides(color = guide_legend(title = NULL)) +
  ylim(common_y_limits1) +
  theme(legend.position = "none",aspect.ratio = 2/4)

p3 + p4 +
  plot_layout(ncol = 2, guides = "keep") &
  theme(legend.position = "bottom")
```

PCA

```
# PCA 1  ------------------------------------
gini_period1 %>%
  select_if(., is.numeric) %>% select(-cluster) %>%
  scale() %>%
  prcomp() -> pca1
StateSEPC1 <- augment(pca1, gini_period1)
## Kaiser's rule
data.frame(summary(pca1)$importance)[,1:10] %>%
  kable(booktabs = TRUE, longtable = TRUE, digits=4,
        caption = "Importance of components - 1st Period",
        align = "c") %>%
  kable_styling(font_size = 8)
## PCA Period 1: Distance Biplot (with Cluster Membership)
fviz_pca_biplot(pca1, geom = "text",
                habillage = gini_period1$cluster,
                col.var = "grey30",
                title = "Distance Biplot - PCA Period 1") +
  theme(legend.position = "top") +
  labs(x="PC1",y="PC2")
## Correlation Biplot
rownames(pca1$x) <- use_series(gini_period1, state)
biplot(pca1, cex = 0.6, xlim = c(-4.5, 3), ylim = c(-4, 4), scale = 0)
```

```
# PCA 2  ------------------------------------
gini_period2 %>%
  select_if(., is.numeric) %>% select(-cluster) %>%
  scale() %>%
  prcomp() -> pca2
StateSEPC2<-augment(pca2, gini_period2)
## Kaiser's rule
data.frame(summary(pca2)$importance) %>%
  kable(booktabs = TRUE, longtable = TRUE, digits=4,
        caption = "Importance of components - 2nd Period",
        align = "r") %>%
  kable_styling(font_size = 8)

rownames(pca2$x) <- use_series(gini_period2, state)
## Distance Biplot (with Cluster Membership)
fviz_pca_biplot(pca2, geom = c("text"),
                habillage = as.numeric(gini_period2$cluster),
                col.var = "grey30",
                title = "Distance Biplot - PCA Period 2") +
  theme(legend.position = "top") +
  labs(x="PC1",y="PC2")
## Correlation Biplot
rownames(pca2$x)<-use_series(gini_period2,state)
biplot(pca2,cex=0.5,xlim=c(-4.6,4),ylim=c(-4,2.2),scale=0)
```

References

Bank, W (2023a). *GDP growth (annual %) - United States*. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2015&locations=US&start=2007.

Bank, W (2023b). *Inflation, consumer prices (annual %) - United States*. https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?end=2015&locations=US&start=2007.

Bank, W (2023c). *Unemployment, total (% of total labor force) (modeled ILO estimate) - United States*. https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?end=2015&locations=US&start=2007.

Boyd A. Martin, GLM (n.d.). *Economy of Idaho*. https://www.britannica.com/place/Idaho/Government-and-society.

Bureau, UC (2021). *Gini Index*. Website. https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/gini-index.html.

CBPP (2012). *Income Inequality Has Grown in Florida*. https://www.cbpp.org/sites/default/files/atoms/files/Florida.pdf.

Davis, S (1978). The South as "the Nation's No. 1 Economic Problem": the NEC Report of 1938. *The Georgia Historical Quarterly* **62**(2), 119–132.

Duignan, B (2018). *Causes of the Great Depression*. Encyclopedia Britannica. https://www.britannica.com/story/causes-of-the-great-depression.

Farris, FA (2010). The Gini Index and Measures of Inequality. *The American Mathematical Monthly* **117**(10), 851–864. eprint: https://www.tandfonline.com/doi/pdf/10.4169/000298910X523344.

Farris, FA (2015). *Update on Other Measures of Income Inequality (Atkinson Index, Gini Coefficient, Relative Mean Deviation, Theil Index), 1917-2015*. English.

Fishback, P (2019). *World War II in America: Spending, deficits, multipliers, and sacrifice*. https://cepr.org/voxeu/columns/world-war-ii-america-spending-deficits-multipliers-and-sacrifice.

Nourse, EG (1936). Fundamental Significance of the Agricultural Adjustment Concept. *Journal of Farm Economics* **18**(2), 244–255.

Pells, RH and CD Romer (2023). *Great Depression*. Encyclopedia Britannica. https://www.britannica.com/money/topic/Great-Depression.

Rowe, GS (1980). History of Delaware. *Journal of American History* **67**(3), 655–656.

Schmitz, M and PV Fishback (1983). The Distribution of Income in the Great Depression: Preliminary State Estimates. *The Journal of Economic History* **43**(1), 217–230.

Skates, JR (2017). *World War II*. Date of Last Update: April 15, 2018. http://mississippiencyclopedia.org/entries/world-war-ii/ (visited on 09/21/2023).

U.S. Bureau of Labor Statistics (2010). *Job Availability During a Recession: An Examination of the Number of Unemployed Persons Per Job Opening*. U.S. Bureau of Labor Statistics. https://www.bls.gov/opub/btn/archive/job-availability-during-a-recession-an-examination-of-the-number-of-unemployed-persons-per-job-opening.pdf.

Vuoccolo, A (2018). *Viewpoint: Delaware Income Inequality Through the Ages*. https://delawarebusinesstimes.com/news/delaware-income-inequality/.

Waiwood, P (2013). *Recession of 1937–38*. Federal Reserve History. https://www.federalreservehistory.org/essays/recession-of-1937-38.