Assignment1

Dimensio Reducto

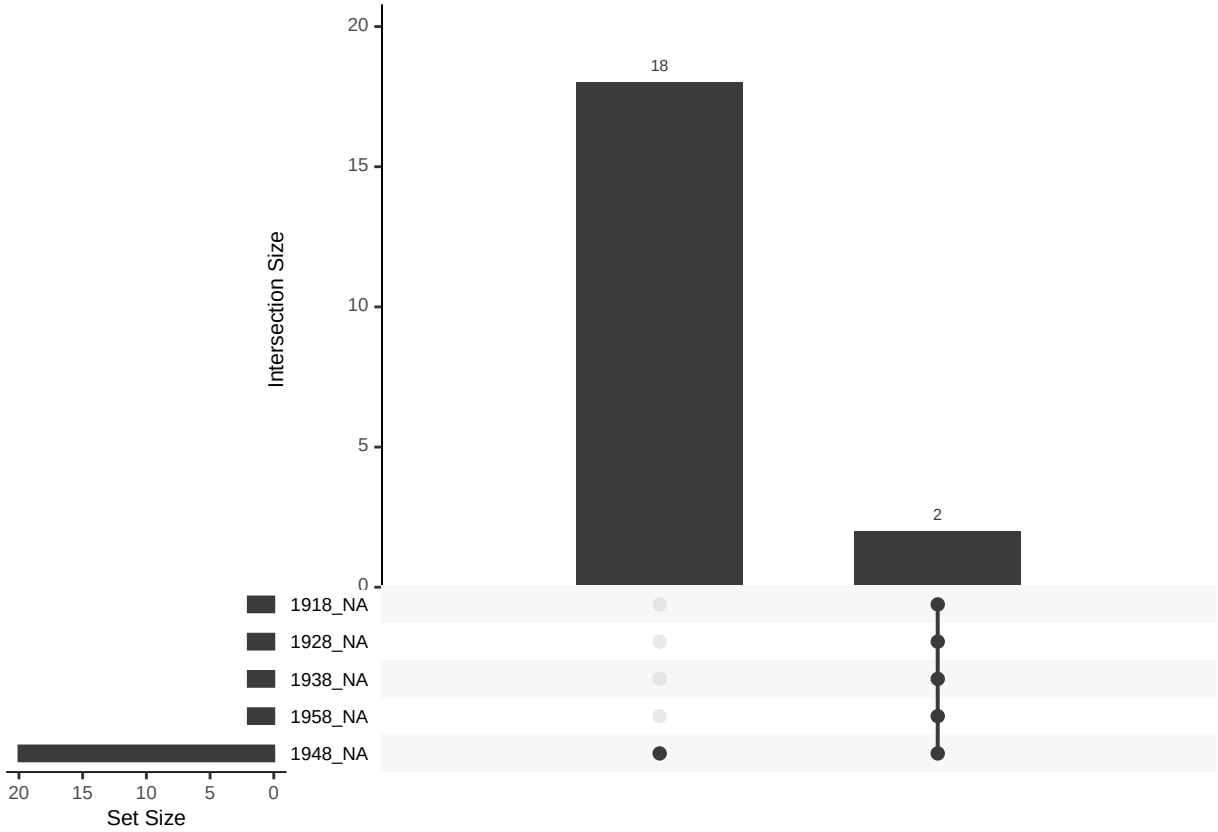
2022-09-21

Data Description

Inequality data has eleven variables, representing the average income inequality in the U.S. for 51 different states and an aggregate of these states. This panel data set variables are the Gini index from 1918 to 2018, recorded every ten years for a total of eleven Gini index variables. In this report, the use of principle components analysis, cluster analysis, and multidimensional scaling will help investigate how inequality has evolved over time for different states in the U.S.

Preliminary Analysis

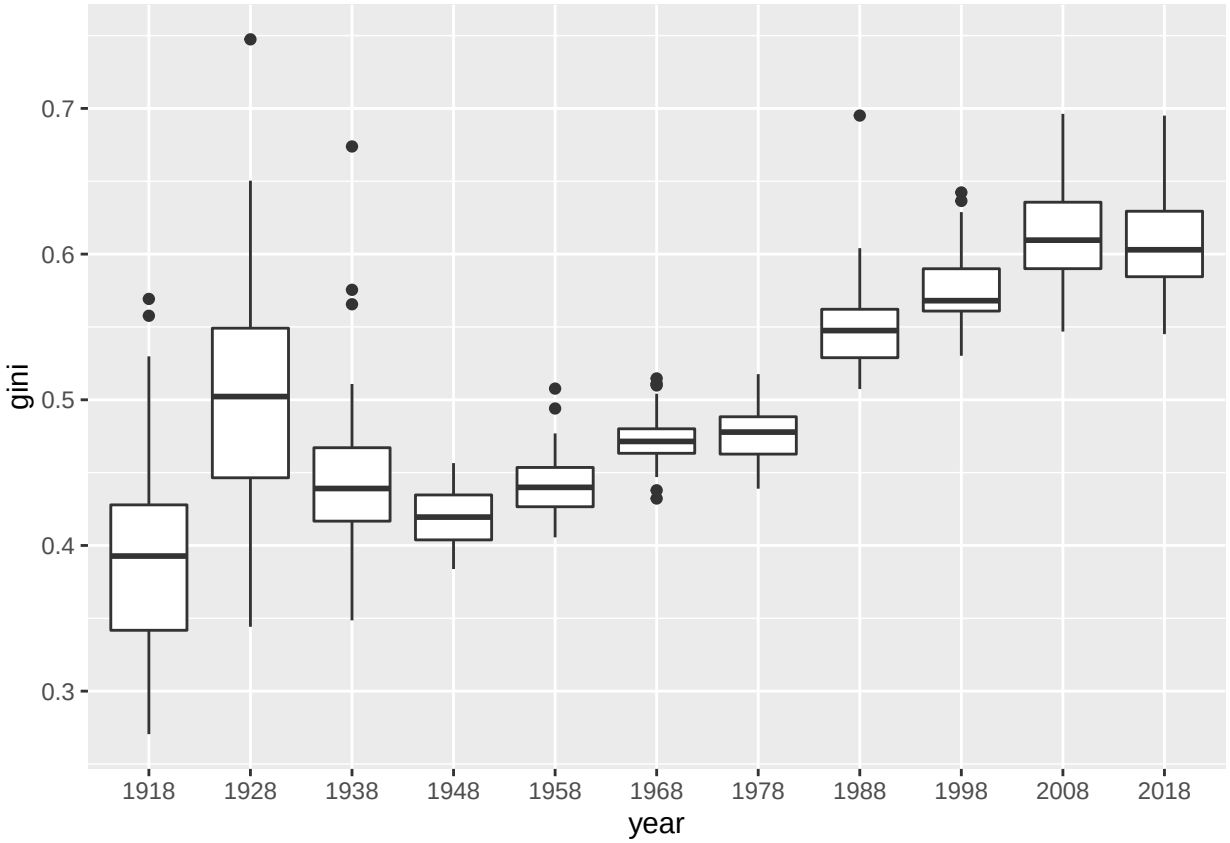
Overview of the raw data



There are a few anomalies such as 2 NA values for all years from 1918 to 1958 with the exception of year 1948 having 20 NA values.

```
##     State                1918             1928             1938
##  Length:52          Min.   :0.2705   Min.   :0.3442   Min.   :0.3486
##  Class :character   1st Qu.:0.3418   1st Qu.:0.4465   1st Qu.:0.4167
##  Mode  :character   Median :0.3927   Median :0.5022   Median :0.4391
##                     Mean   :0.3922   Mean   :0.5009   Mean   :0.4478
##                     3rd Qu.:0.4279   3rd Qu.:0.5492   3rd Qu.:0.4671
##                     Max.   :0.5692   Max.   :0.7474   Max.   :0.6739
##                     NA's   :2        NA's   :2        NA's   :2
##       1948             1958             1968                1978
##  Min.   :0.3838   Min.   :0.4056   Min.   :   0.4322   Min.   :0.4390
##  1st Qu.:0.4038   1st Qu.:0.4266   1st Qu.:   0.4636   1st Qu.:0.4627
##  Median :0.4195   Median :0.4399   Median :   0.4716   Median :0.4778
##  Mean   :0.4199   Mean   :0.4416   Mean   :  19.6947   Mean   :0.4766
##  3rd Qu.:0.4347   3rd Qu.:0.4535   3rd Qu.:   0.4805   3rd Qu.:0.4883
##  Max.   :0.4565   Max.   :0.5077   Max.   :1000.0000   Max.   :0.5176
##  NA's   :20       NA's   :2
##       1988             1998             2008             2018
##  Min.   :0.5074   Min.   :0.5302   Min.   :0.5468   Min.   :0.5450
##  1st Qu.:0.5289   1st Qu.:0.5609   1st Qu.:0.5900   1st Qu.:0.5845
##  Median :0.5475   Median :0.5680   Median :0.6096   Median :0.6030
##  Mean   :0.5504   Mean   :0.5768   Mean   :0.6120   Mean   :0.6075
##  3rd Qu.:0.5621   3rd Qu.:0.5900   3rd Qu.:0.6356   3rd Qu.:0.6294
##  Max.   :0.6951   Max.   :0.6423   Max.   :0.6963   Max.   :0.6951
##
```

Additionally the highest Gini index for the year 1968 is 1000, which is definite a mistake considering the range of Gini index value is between 0 and 1.



We can see from the box plot that the variance of gini index, signifies by the interquartile range, between states was higher for the first 2 periods, 1918 and 1928, and then gradually reduced, reaching lowest variance in 1968 but then started to slightly increase again over the next periods. We can see that quite a few periods there are outliers such as in the years of 1918, 1928, 1938, 1958, 1968, 1988 and 1998. However, only the year 1938 has an extreme value, which is when the outliers is 3*IQR greater than the 1st interquartile. We can see all the outliers below.

State	1918	1928	1938	1948	1958	1968	1978	1988	1998	2008	2018
Delaware	0.5692365	0.7474164	0.6738963	NA	0.4940233	0.5146480	0.4963422	0.5203046	0.5536698	0.5589235	0.5734153
Florida	0.3829387	0.5654918	0.5655543	NA	0.4665656	0.5098608	0.4908346	0.5788947	0.6212426	0.6849992	0.6931326
Maine	0.4131196	0.5487878	0.5754811	NA	0.4614237	0.4469319	0.4686843	0.5119279	0.5624066	0.6004139	0.5570027

The extreme value for the year 1938 is Delaware, with Gini index at 0.674. From the box plot we can also see that the largest Gini index across the whole data set comes from the year 1928, and from the table above, we can see that the point is also Delaware, with Gini at 0.747, this means that during the period 1928 to 1938, the state of Delaware had very wide income gap.

In terms of the distribution of gini index across the years, we can see that the data is quite normally distributed. We can see that the only period which is visibly positively skewed is the year 1998, where the mean is very close to the 1st quantile.

Reassignment of missing values

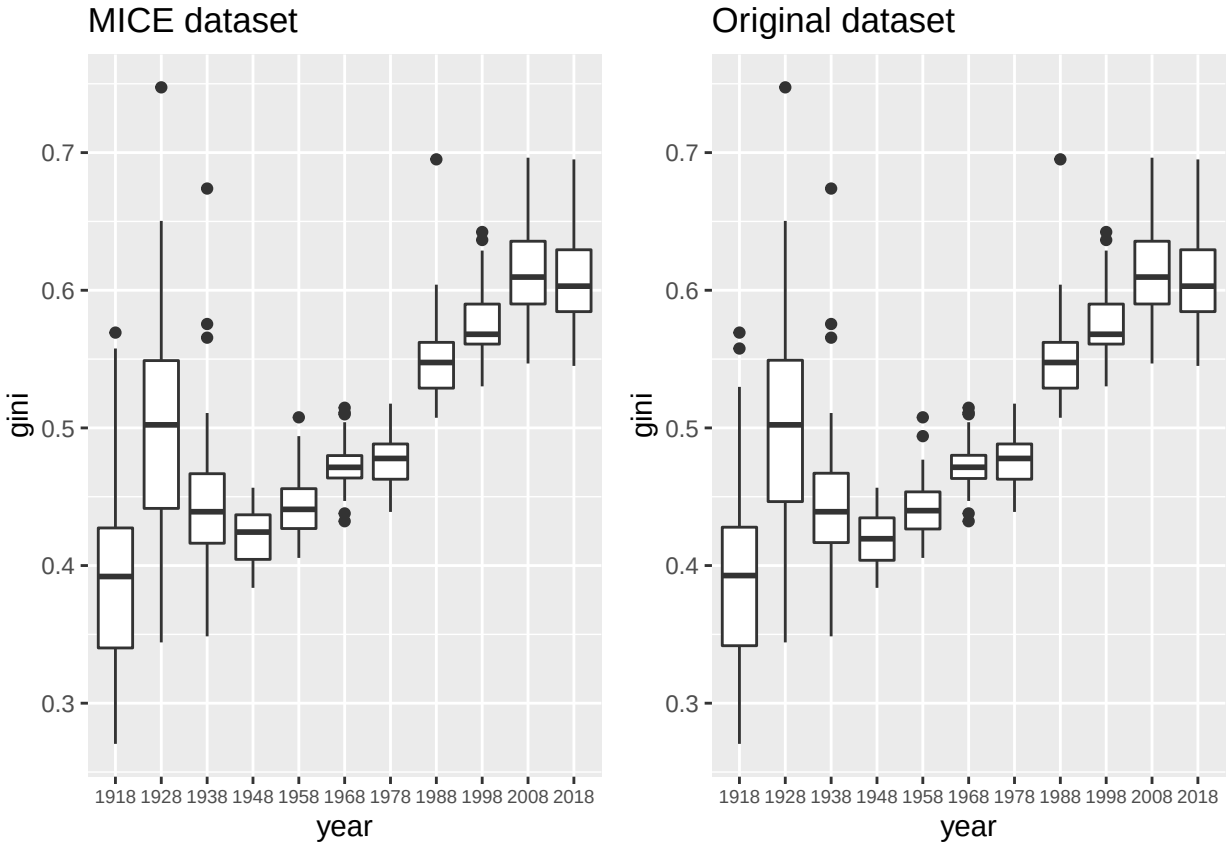
The reason that the aggregate item United States is not removed is that we want to keep this as an observation for reassigning NA. At the same time, although the reason for missing data before 1958 like Alaska and Hawaii

is that two states were only admitted into the union in 1959, in order to ensure the integrity of the data, we chose to predict and fill the NA by regression fitting rather than dropping them directly. Additionally, the highest gini index 1000 for the year 1968 is changed to NA firstly, then will perform 3 different ways to handle all missing values.

First, we have used the MICE (Multiple Imputation by Chained Equations) algorithm. This is a robust, informative method of dealing with missing data in the dataset. The procedure imputes missing data in a dataset through an iterative series of predictive models. In each iteration, each specified variable in the dataset is imputed using the other variables in the dataset. These iterations should be run until it appears that convergence has been met.

raw_1918	mice_1918
0.392153	0.3906969

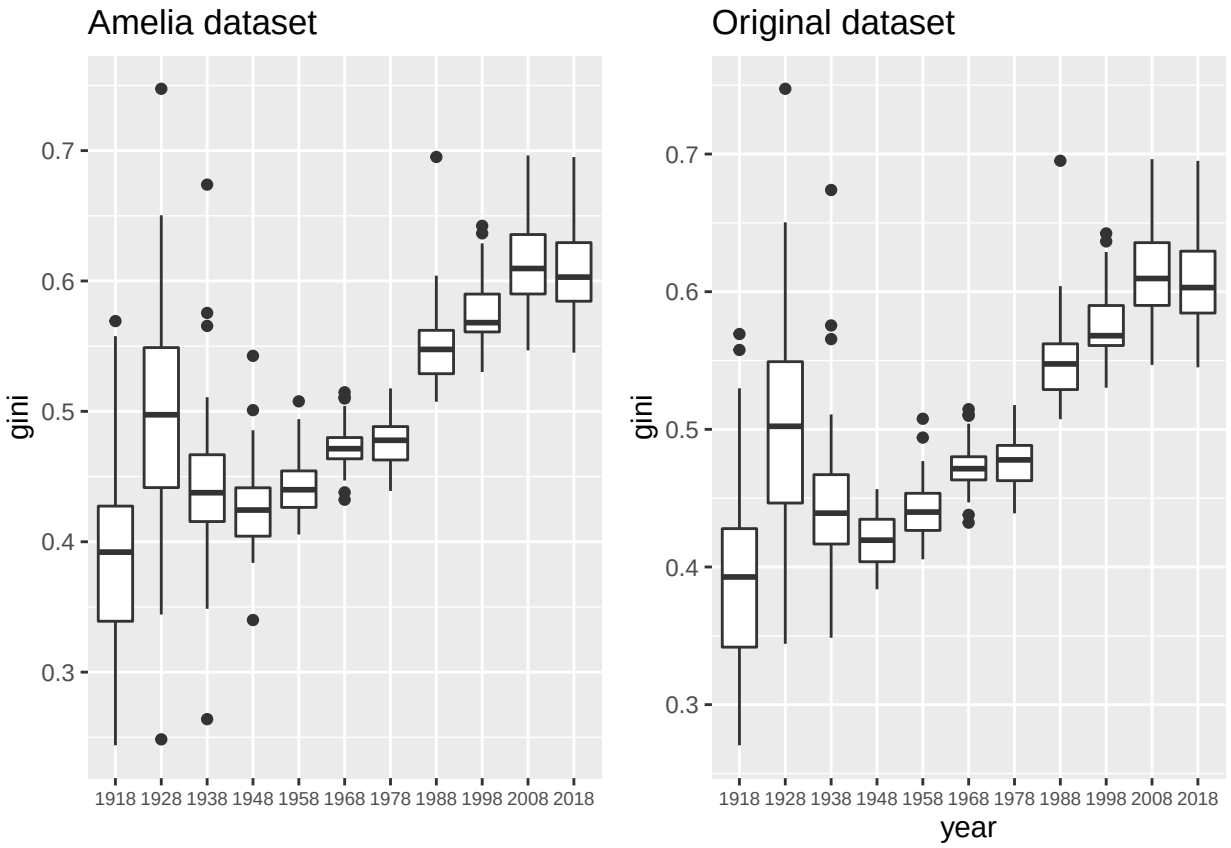
After filling in the missing values with mice algorithm, we can see that the mean of the dataset actually hasn't changed much, this is because the procedure we have used is called predictive mean matching (PMM).



However, the issue with outliers has clearly improved. For the year 1918 and 1958, there were 2 outliers in the original dataset, and now are reduced to 1 after filling in the missing values.

The second method we have tried is to use the package Amelia which impute the missing values probabilistically. This method conducts multiple imputation. It assumes the data is multivariate normal distribution, and it imputes m values for each missing values creating m completed datasets, then analyze each of m completed datasets seperately and finally combine the m results by taking the average and adjust the standard error. This procedure is similar to use bootstrap to simulate by independent variable and the algorithm is called Expectation-maximization with bootstrapping. This method produces unbiased estimates. However, the

limitation is that the values are imputed with uncertainty. In this case, same imputed values are negative which does not comply with our scenario where gini index is between 0 to 1.



After filling in the missing values with amelia package, we can see that the outliers in the year 1918 are reduced to 1, but from 1928 to 1948 there were more new outliers.

The third method is to use the mean or median or nearby values to fill in the missing values. The advantage of this method is easy computing, but of course it has a clear limitation which could lead to biased estimates, and it does not account for uncertainty of the imputed values.

In summary, all 3 methods can to some extend solve the problem of missing values. While the mean method could create biased estimates, the Amelia method also has uncertainty to the imputed values. Hence, we use the mice algorithm to handle the missing value and will use the filled in datasets for the following analysis.

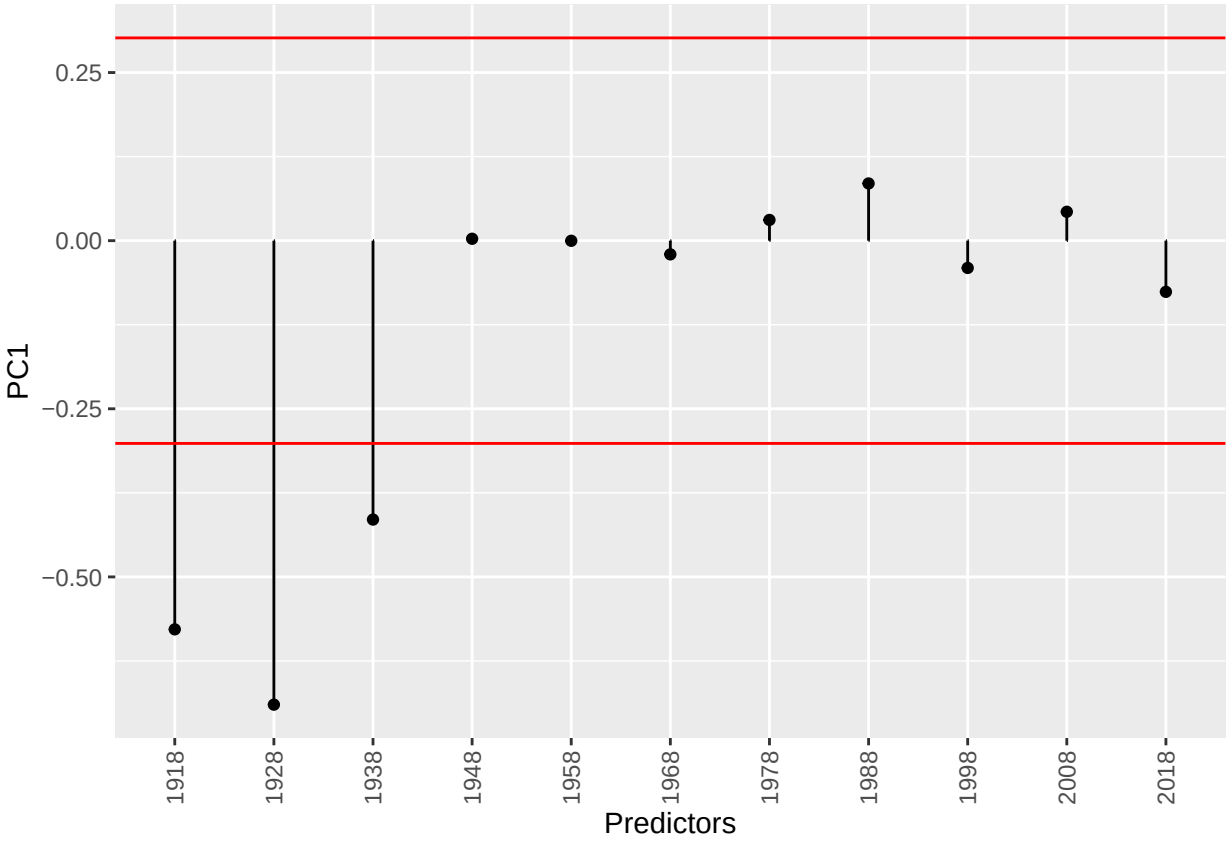
Principle Component Analysis

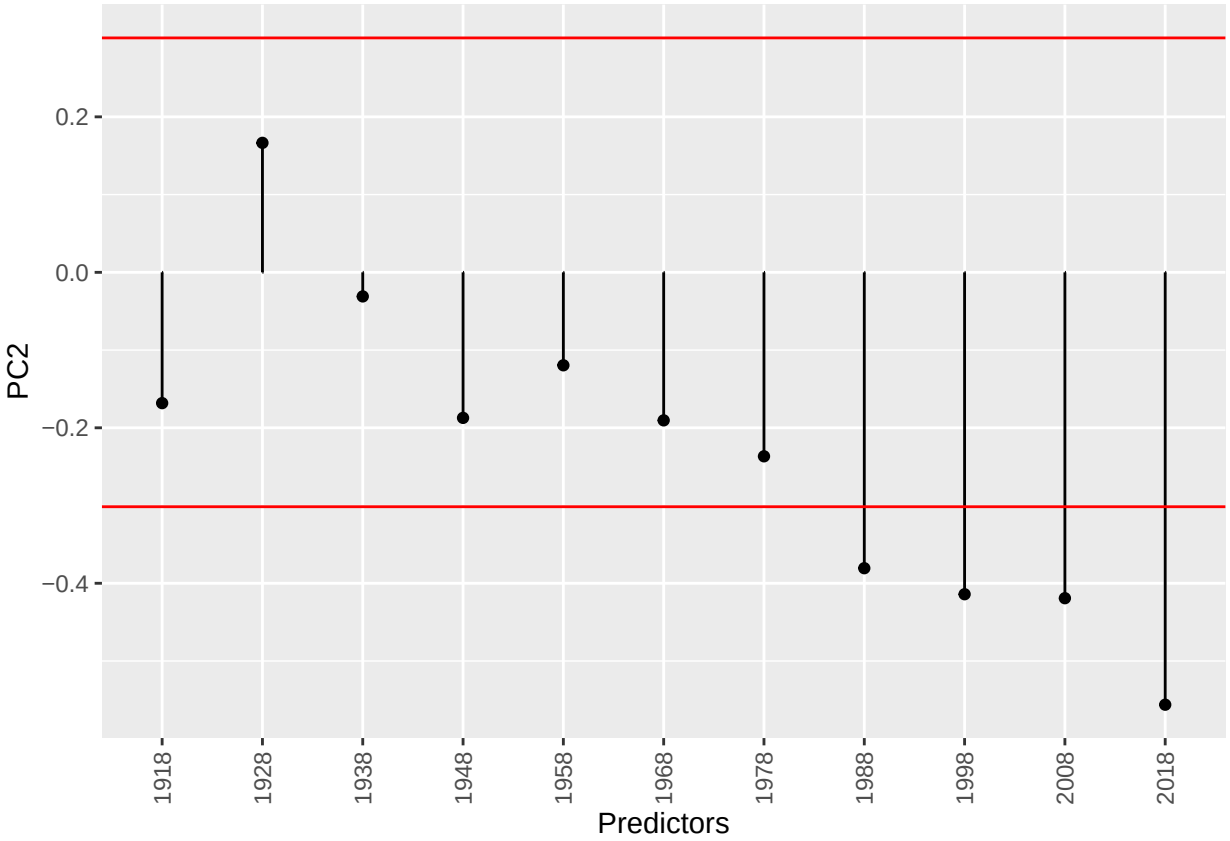
Principal components analysis finds a small number of linear combinations of the original variables that explain a large proportion of overall variation in the data. Since the variables in the dataset under investigation are measured in the same units, we don't need to standardise the data.

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     0.1032 0.05251 0.04293 0.03171 0.03021 0.02313 0.01661
## Proportion of Variance 0.5821 0.15069 0.10072 0.05493 0.04986 0.02924 0.01507
## Cumulative Proportion  0.5821 0.73277 0.83348 0.88841 0.93828 0.96752 0.98259
##                            PC8     PC9     PC10     PC11
```

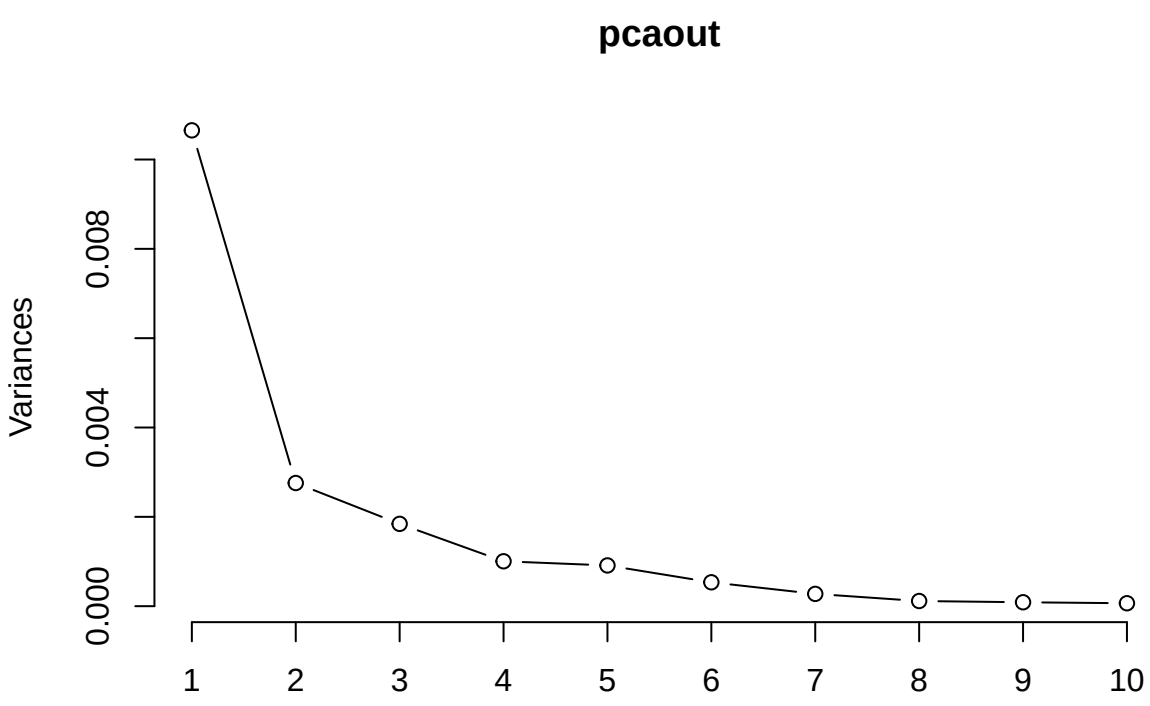
```
## Standard deviation     0.01077 0.00938 0.008134 0.006975
## Proportion of Variance 0.00633 0.00481 0.003620 0.002660
## Cumulative Proportion  0.98892 0.99373 0.997340 1.000000
```

From the summary report, we can see that there are 11 principal components are obtained. Each of these explains a percentage of the total variation in the dataset. That is to say: PC1 explains 58% of the total variance, while PC2 explains 15% of the total variance, as just PC1 and PC2 can explain 73% of the total variance.

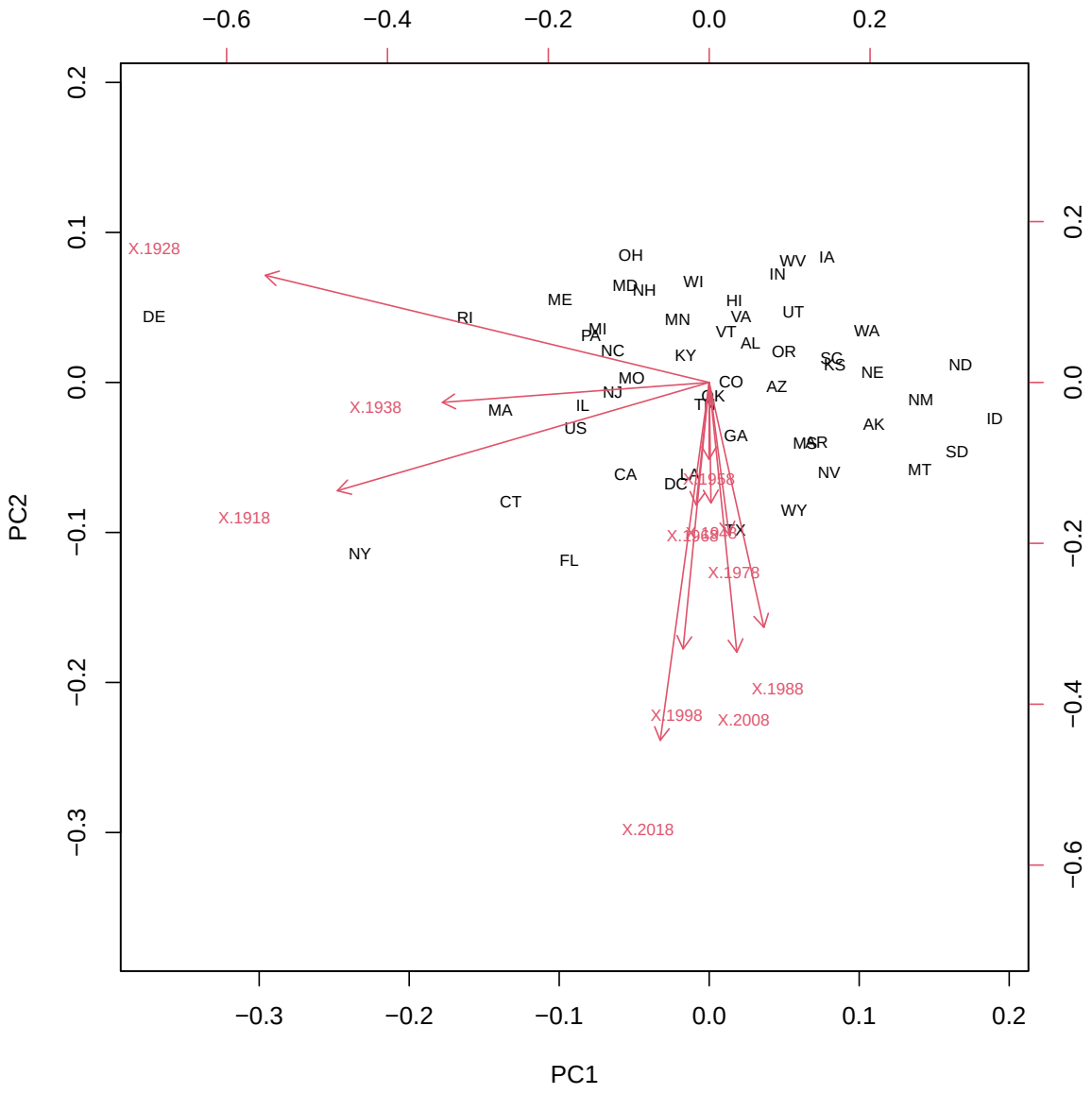




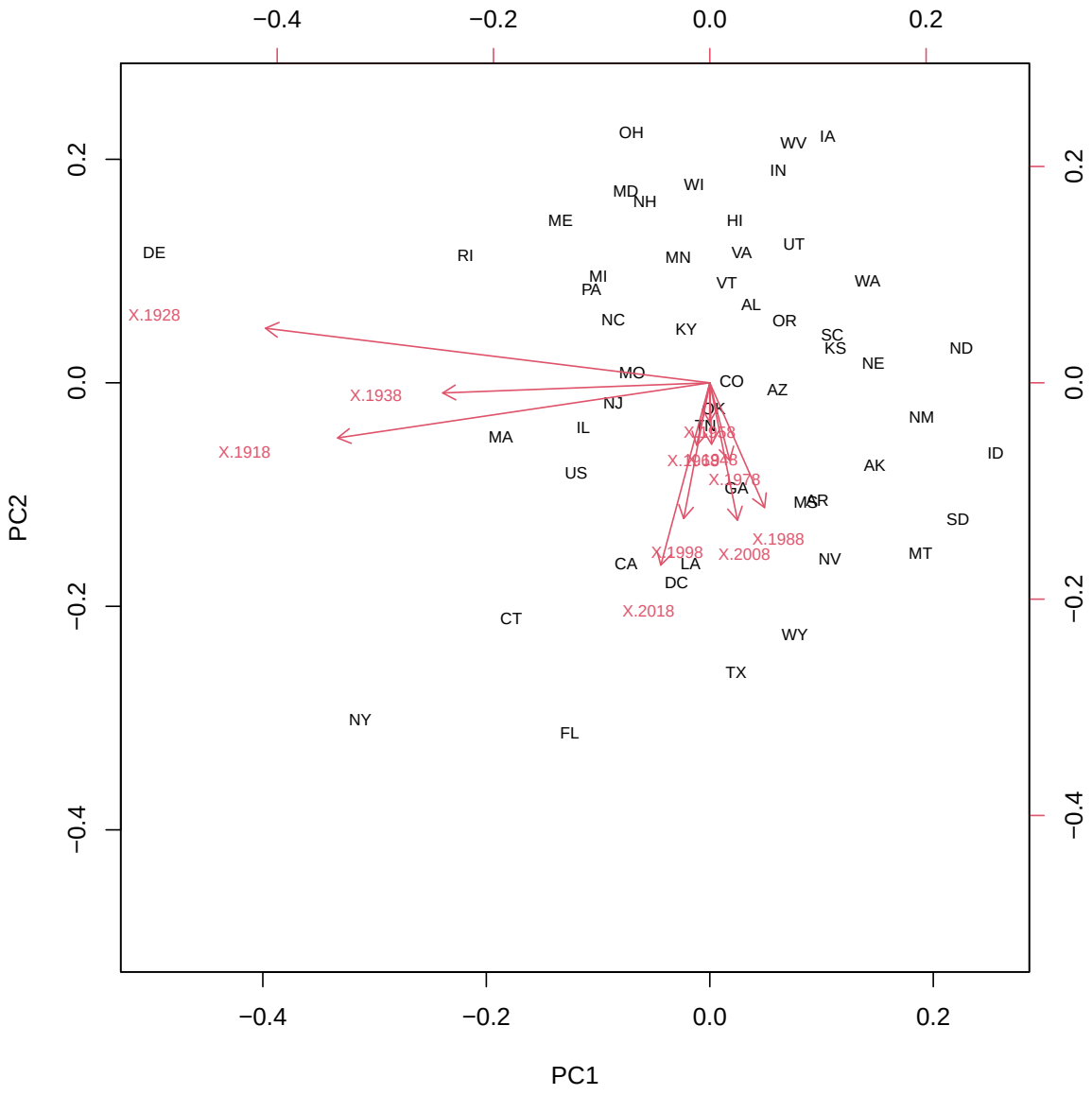
By visualizing the loading plots, we could identify the variable importance to each principal components. While the red lines represents the confidence interval, variables that exceed the red lines are significantly different from zero, which means having more contribution to the principal components. We can see that year 1918,1928 and 1938 are more influential to PC1, while year 1988,1998,2008 and 2018 are more influential to PC2.



A scree plot, on the other hand, is a diagnostic tool to check whether PCA works well on the data. PC1 captures the most variation, PC2 — the second most, and so on, and each of them contributes some information of the data. Here we can see that PC1 and PC2 have captured most of the variation.

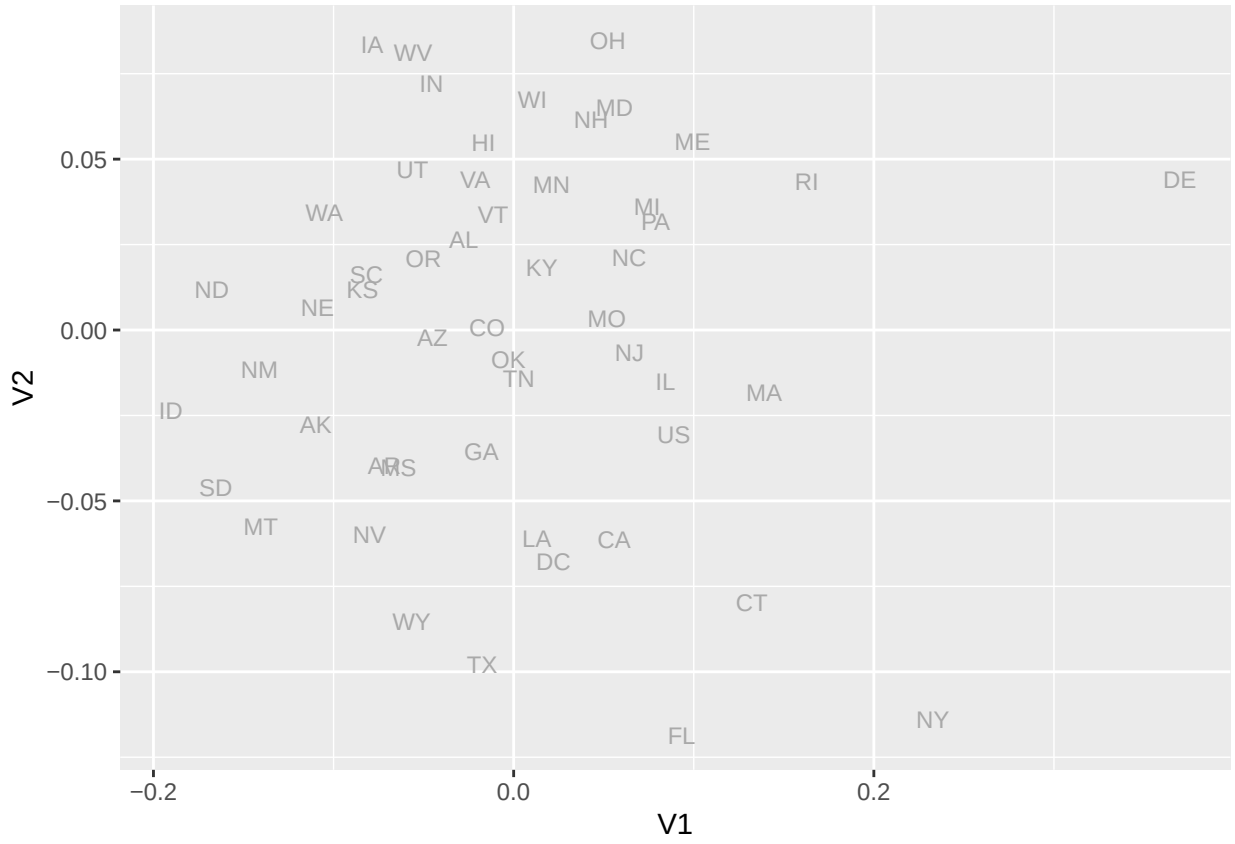


By selecting two principal components we are able to visualize the data using a biplot. The first biplot shows the correlation among variables. We can see year 1928, 1918 and 1938 are positively correlated. One possible explanation could be due to the World War 1 and 2 during that period of time. The rest of the variables are positively correlated.



The second biplot implies the distance. While 1928 contributes the most to the PC1, 2018 contributes the most to the PC2. Also, the closer distance between states, the more similar they are. While most observations are clustered in the top right side of the plot, Delaware, New York and Florida are obviously different from the others.

Multidimensional Scaling

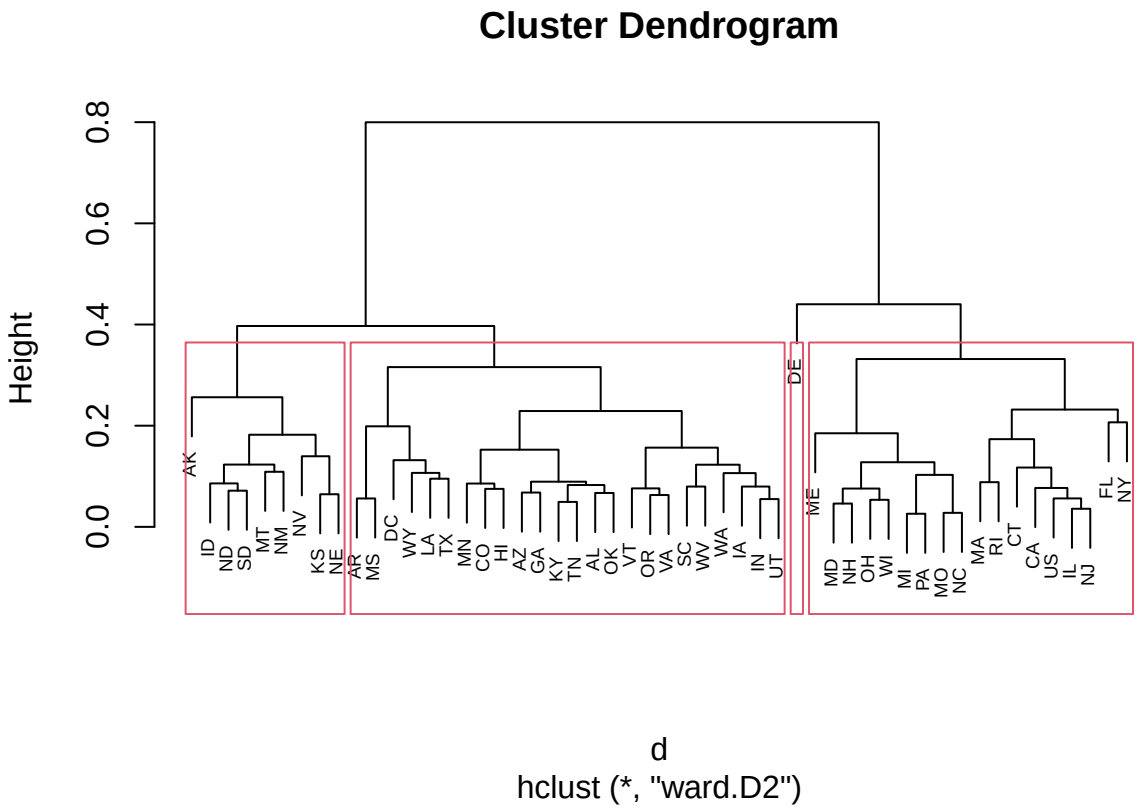


Using the classical Multidimensional Scaling, it finds a low dimensional representation of the dataset by minimizing strain. From the plot above, we can see the potential three outliers of the dataset; Delaware, New York, and Florida. This result is the same as what we have observed in the PCA. Also, the states that are closer together, the closer they are, as this plot represents the similarity between states.

```
## [1] 0.7327662 0.7327662
```

The *GOF* tells us how good the fit is. The numbers are the same for both measures. It is because we use Euclidean distances, which will always give positive eigenvalues, and since the first measure use $|\lambda_i|$ and the second measure uses $max(0, \lambda_i)$. Therefore when using the Euclidean distance, it will give the same results. With the goodness of fit measures being quite high, we can believe that the solution is an accurate representation.

Cluster Analysis

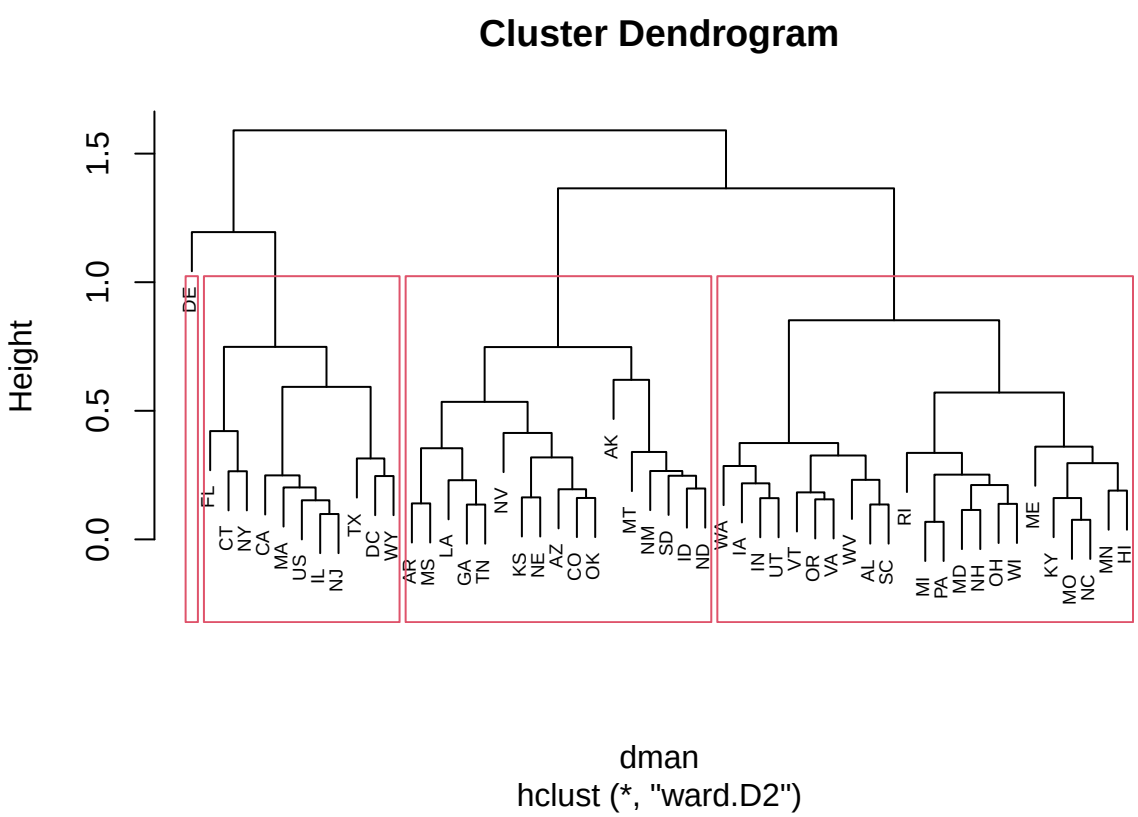


From the dendogram above we can see that using the Ward clustering method, we can cut the tree into 4 clusters for stable solutions. Because 3 clusters is not stable and can change over a short range of tolerance, while 5 and above will not include the cluster with Delaware state. 1 and 2 clusters not a good choice as there are room to form better defined clusters.

What's more interesting to see is that, aside from Delaware, we can identify another less smaller outlier which is Alaska. If we go back to the box plot and identify the second most outlying point in Gini index, which is in the year 1988, we can see that the state is indeed Alaska as below

State	1918	1928	1938	1948	1958	1968	1978	1988	1998	2008	2018
Alaska	NA	NA	NA	NA	NA	0.4763226	0.4943701	0.6951078	0.5890373	0.5625051	0.5589378

If we use manhattan distance instead to cluster the states, the stable solution would also be 4, similar to that as in the case of euclidean distance. More interestingly is we can identify New York and Florida which are previously within the same cluster, if we decided the number of cluster to be higher, and is a potential outlier as compared to the majority of other clusters. These only happens when Euclidean distance is used but using Manhattan distance in clustering, they are not within the same cluster anymore nor do they look like potential outliers



To see if the members of the state are different or not as compared to when euclidean distance is used, we can assume the cluster is 4 and use Rand index:

```
## [1] 0.2084971
```

It's interesting to see how when different distance is used, the number of clusters are still the same but the members of each cluster are all difference, as the adjusted Rand Index is only 0.2

Limitation

- **Data:** the obvious limitation of this data is the null values, although using many different ways to compare and finally choosing the Predictive Mean Matching from the MICE package to reassign, it is also small biased from the actual situation.
- **PCA:** one limitation is the low interpretability of PCA, although the first two PCs have explained 73% of the total variance, it is difficult to generalize what the new index specifically means. On the other hand, PCA is sensitive to the scale of the features and not robust against outliers, so we tried a variety of NA reassignment methods to minimize the bias from the data.