PCA

Principal Component Analysis

JOIA ZHANG

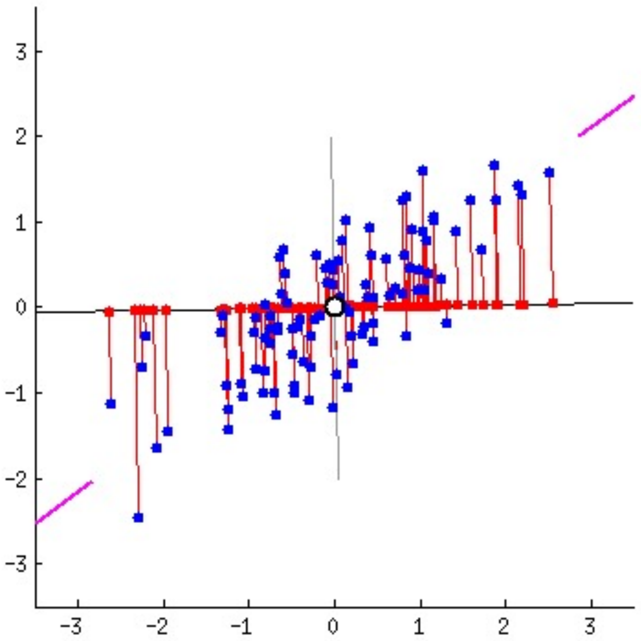
MENTOR: JERRY WEI

WHY PCA?

- Invented 1901 by Karl Pearson
- Still relevant: most widely used dimension reduction technique

THE BIG PICTURE

- Reduce number of variables in your data set while preserving as much information as possible
- Project data onto directions that account for the most variance



Source: stack exchange

1) PREPROCESS DATA

- Centralize data to calculate covariance
- Optional: standardize to prevent large scale variables from dominating others

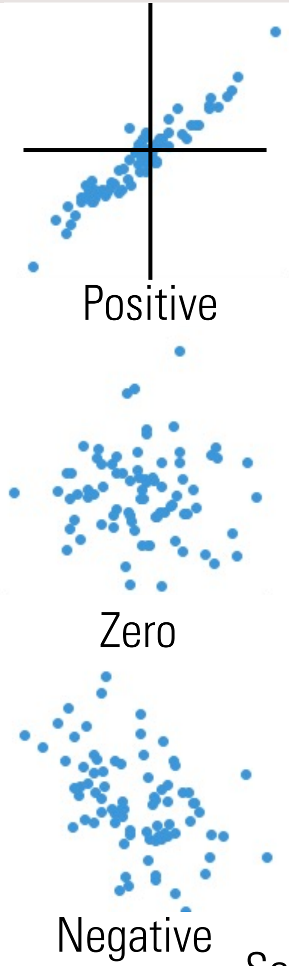
$$\mathrm{X} = \frac{value - mean}{standard\ deviation}$$

2) COMPUTE COVARIANCE MATRIX

- Determines covariance between variables

$$S = XX^T = \begin{bmatrix} \mathrm{Cov}(x, x) & \mathrm{Cov}(x, y) & \mathrm{Cov}(x, z) \\ \mathrm{Cov}(y, x) & \mathrm{Cov}(y, y) & \mathrm{Cov}(y, z) \\ \mathrm{Cov}(z, x) & \mathrm{Cov}(z, y) & \mathrm{Cov}(z, z) \end{bmatrix}$$

$$\mathrm{Cov}(x,y) = \Sigma \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$



Source: chartio

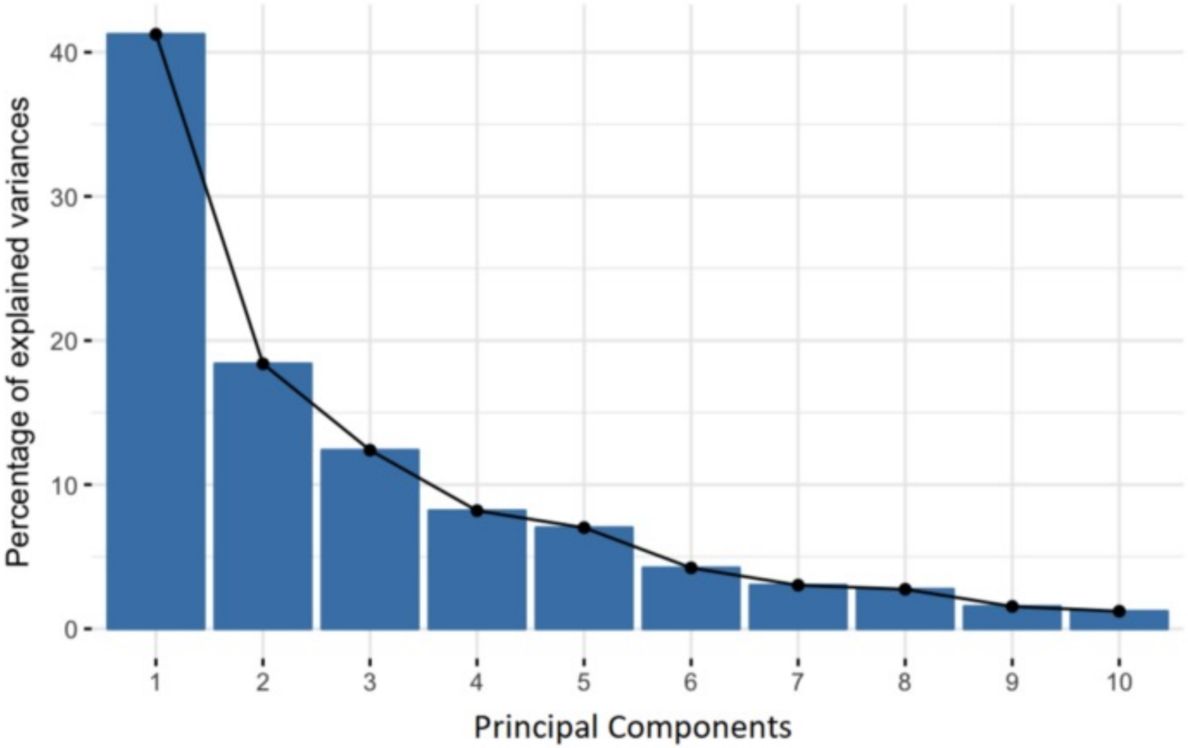
3) EIGENVECTORS & EIGENVALUES

- PCs are eigenvectors of S
- Eigenvalues reflect amount of variance carried in each PC
- Pick eigenvectors with largest eigenvalues

- $\mathrm{S} = \mathrm{P}D\mathrm{P}^{\mathrm{T}}$
- S symmetric covariance matrix
- P projection matrix of eigenvectors
- D diagonal matrix of eigenvalues

4) SELECT PCS TO KEEP

- Choose how many PCs you want to keep based on the variance contained by the PCs



Source: builtin.com

5) RECAST DATA ALONG PCS

- PCs become new axes
- PCs explain a maximal amount of variance
- PCs create a new basis for the data of lower dimension

$$X' = PX$$

- X original data
- P projection matrix of eigenvectors
- X' transformed data

ECOLOGICAL FALLACY

The incorrect assumption that associations identified between group-level variables hold at the individual-level

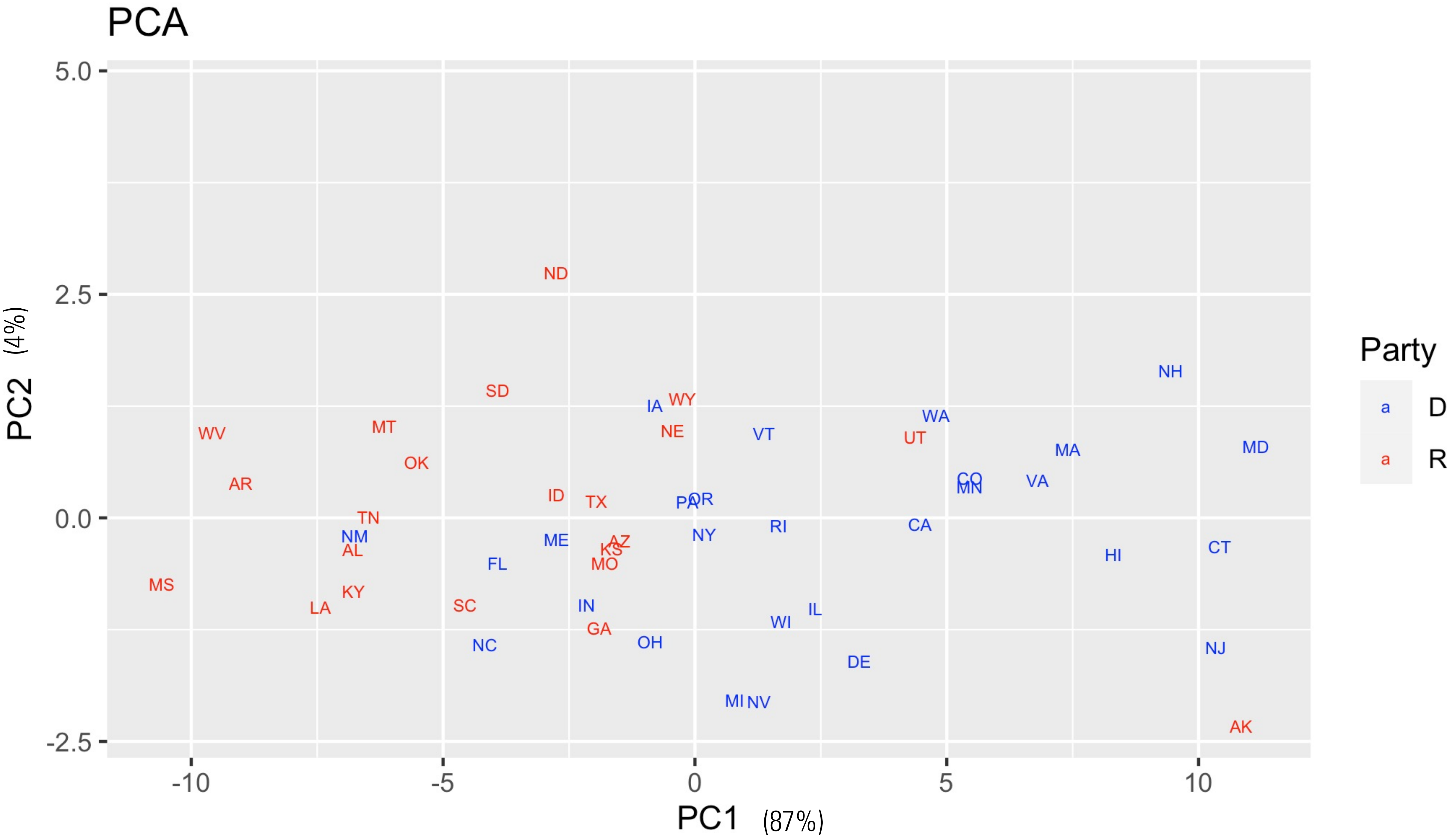
Example: In the U.S., wealthier states tend to favor Democratic candidates, while wealthier individuals tend to favor Republican candidates

DATA

- Data: household income data for each state (1984-2018)
  - Party based on 2008 election

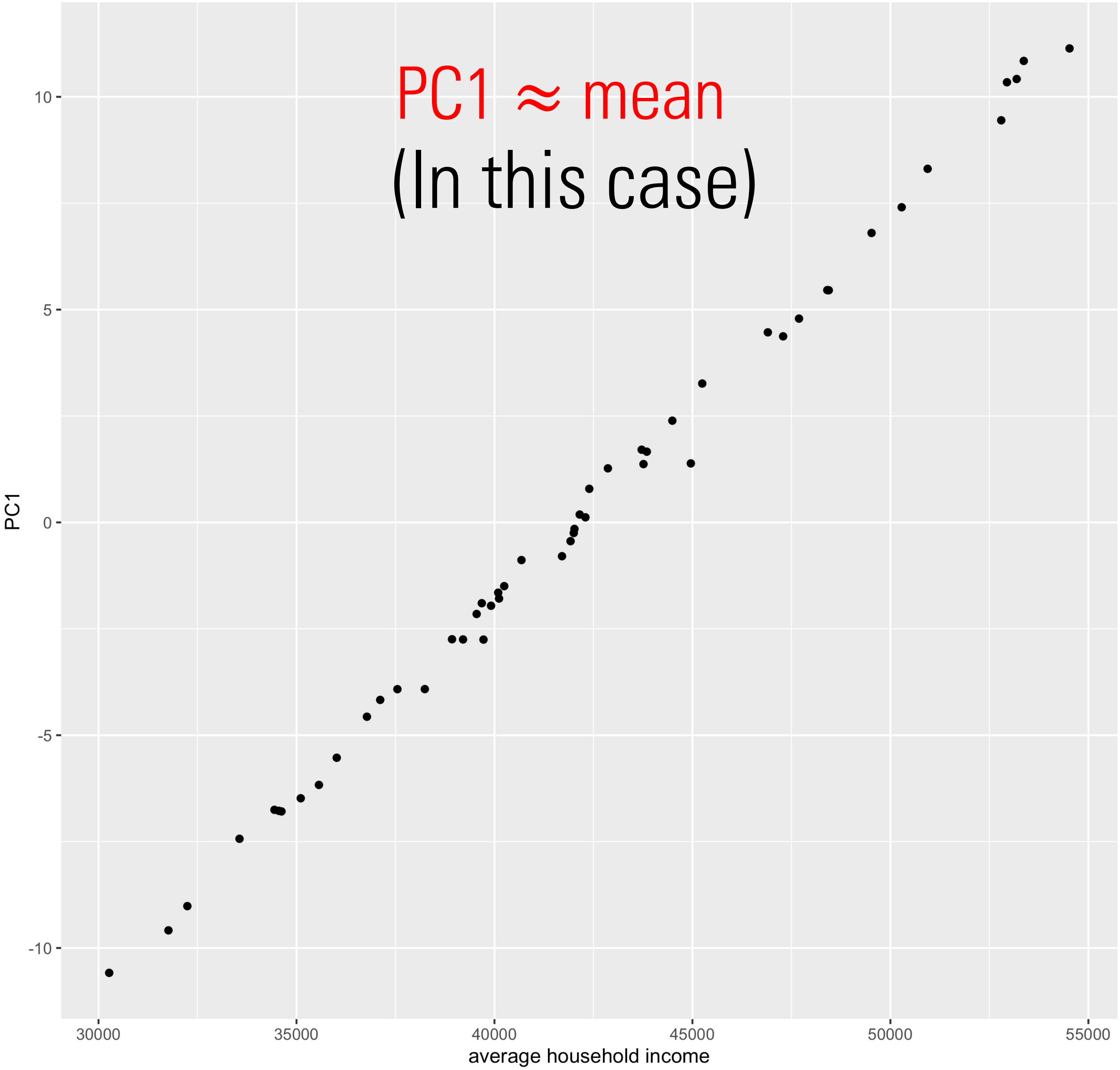
State	Party	HI1984	HI1985	…	HI2017	HI2018
AL	R					
AK	R					
. . .						
WI	D					
WY	R					

Source: US Census Bureau


PCA
PC2 (4%)
PC1 (87%)
5.0
2.5
0.0
-2.5
-10
-5
0
5
10
Party
a D
a R
ND
NH
SD
IA
WY
WA
MT
NE
VT
UT
MD
WV
MA
OK
AR
CO
MN
VA
ID
TX
PA
OR
TN
RI
CA
NM
ME
AZ
NY
CT
AL
KS
HI
MO
FL
MS
KY
LA
SC
IN
IL
WI
GA
OH
NC
NJ
DE
MI
NV
AK

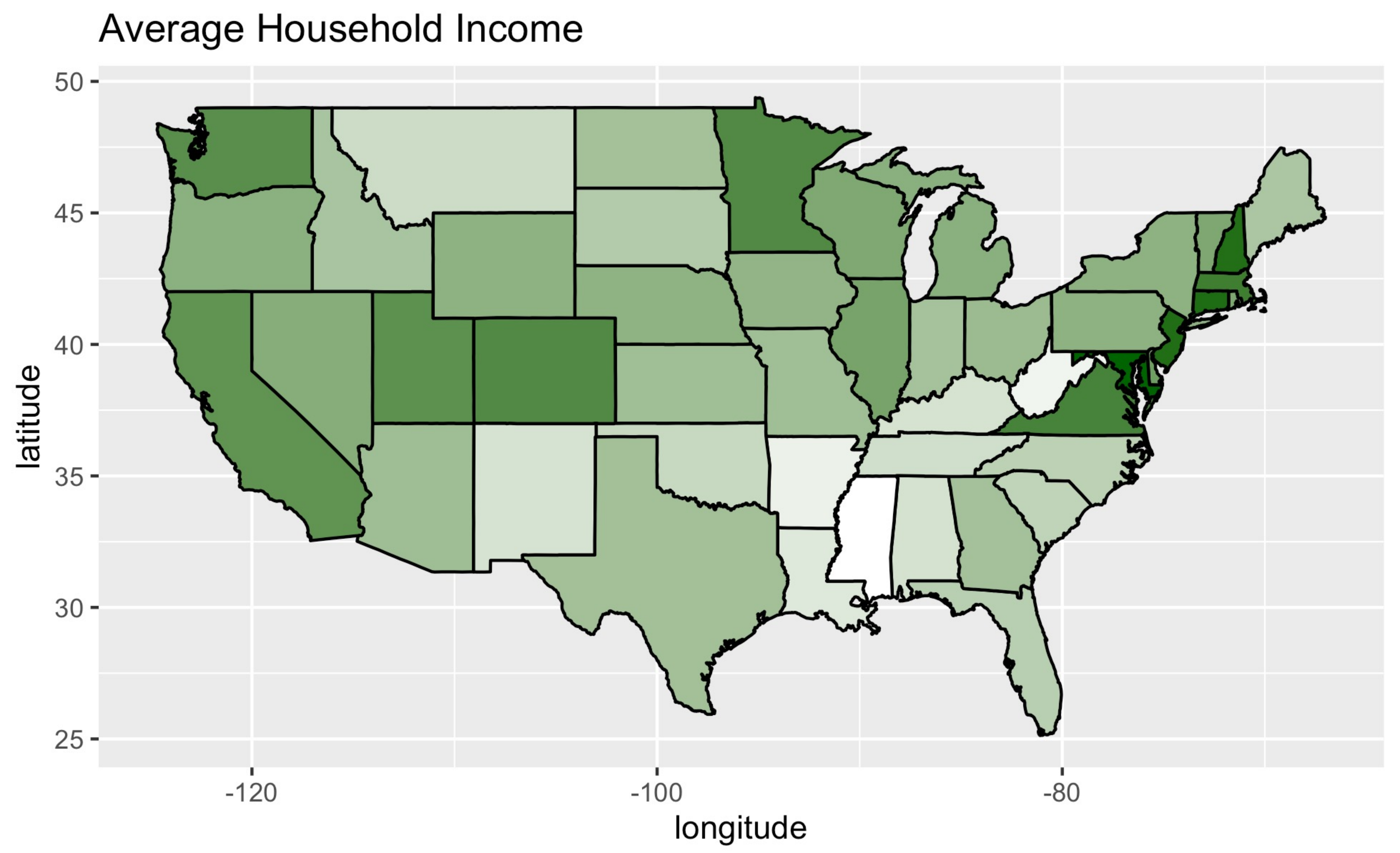
LOADINGS

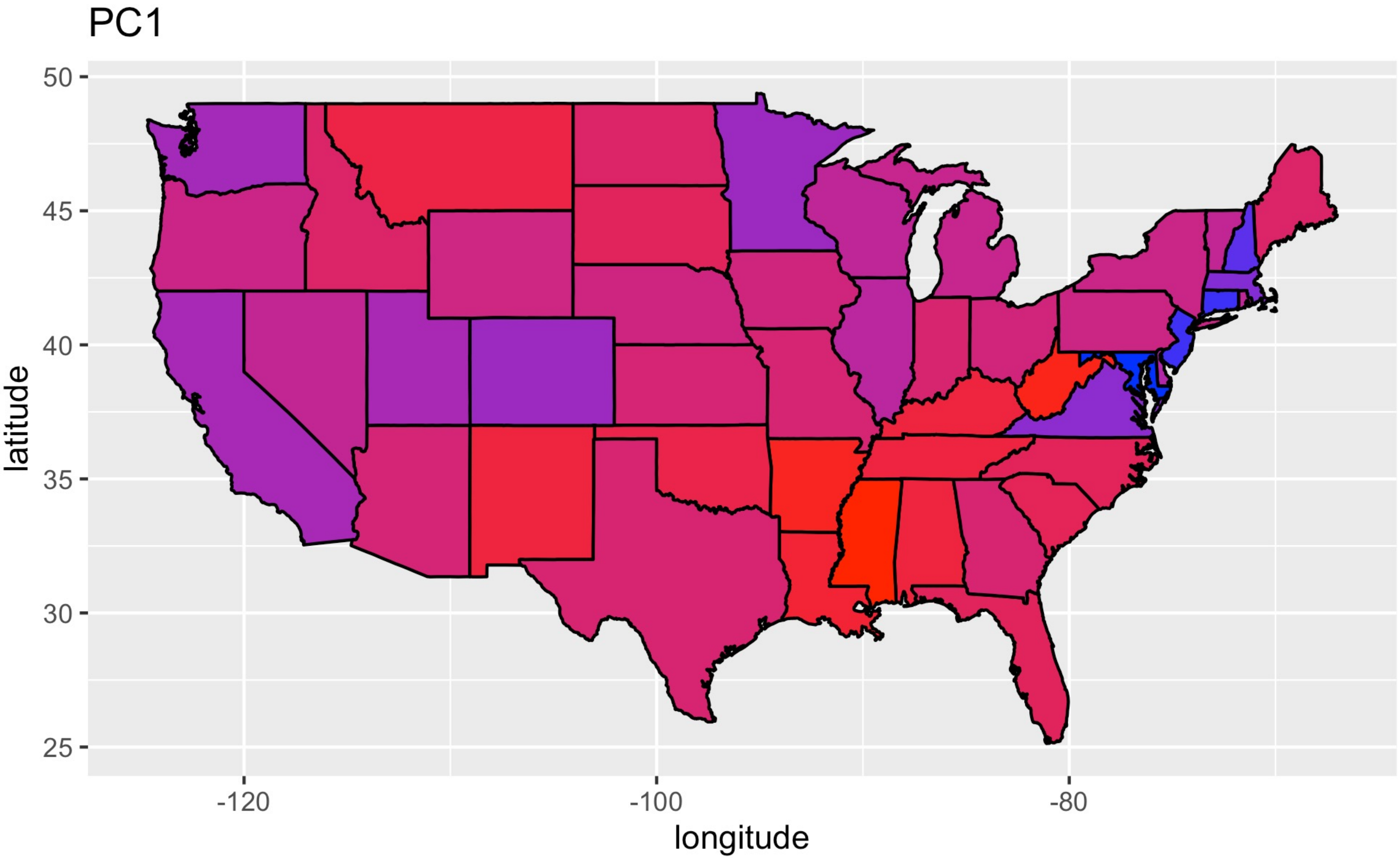
	PC1	PC2
HI2018	0.16	0.31
HI2017	0.16	0.23
HI2016	0.17	0.20
HI2015	0.17	0.22
HI2014	0.16	0.25
HI2013	0.16	0.20
HI2012	0.16	0.29
HI2011	0.16	0.22
HI2010	0.17	0.18
HI2009	0.17	0.12
HI2008	0.17	0.12
HI2007	0.17	0.03
HI2006	0.17	0.00
HI2005	0.17	-0.01
HI2004	0.17	0.00
HI2003	0.17	-0.01
HI2002	0.17	-0.09
HI2001	0.17	-0.12
HI2000	0.17	-0.14
HI1999	0.17	-0.16
HI1998	0.16	-0.19
HI1997	0.16	-0.22
HI1996	0.16	-0.26
HI1995	0.16	-0.20
HI1994	0.17	-0.15
HI1993	0.17	-0.15
HI1992	0.17	-0.09
HI1991	0.17	-0.12
HI1990	0.17	-0.10
HI1989	0.17	-0.08
HI1988	0.17	-0.07
HI1987	0.17	-0.08
HI1986	0.17	-0.10
HI1985	0.16	-0.14
HI1984	0.16	-0.13

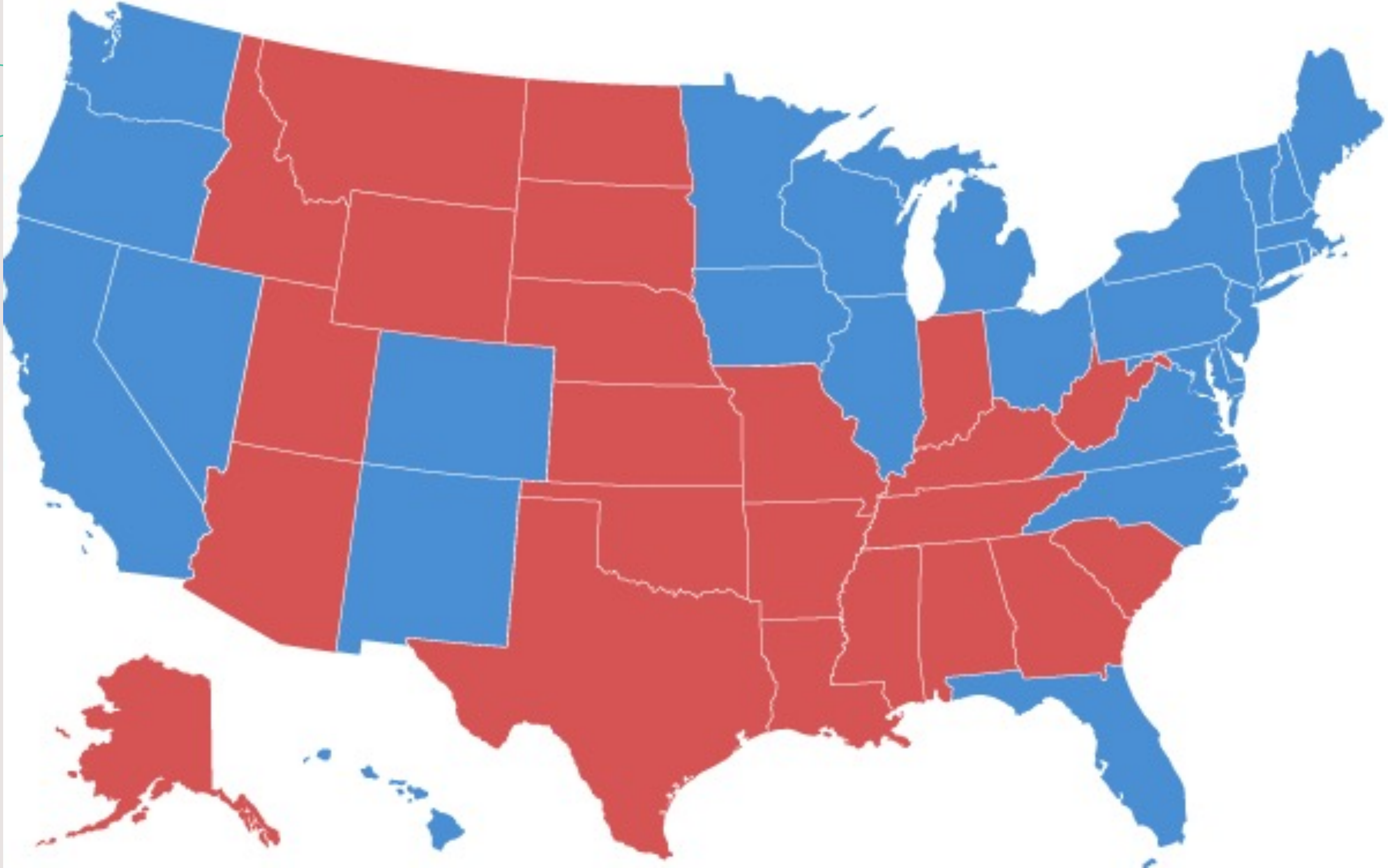


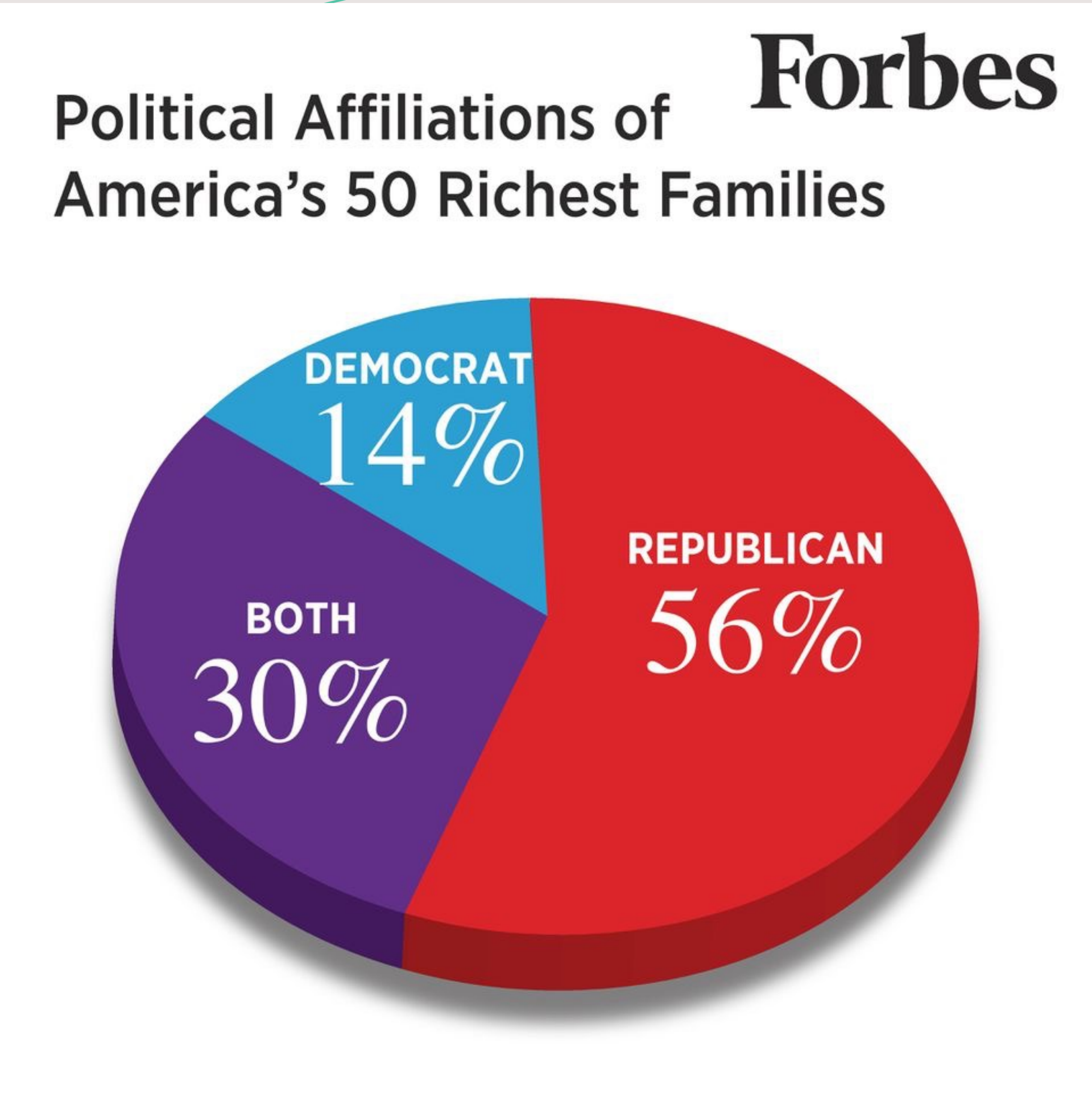
PC1 ≈ mean

(In this case)


Average Household Income
50
45
40
35
30
25
latitude
-120
-100
-80
longitude


PC1
50
45
40
35
30
25
latitude
-120
-100
-80
longitude

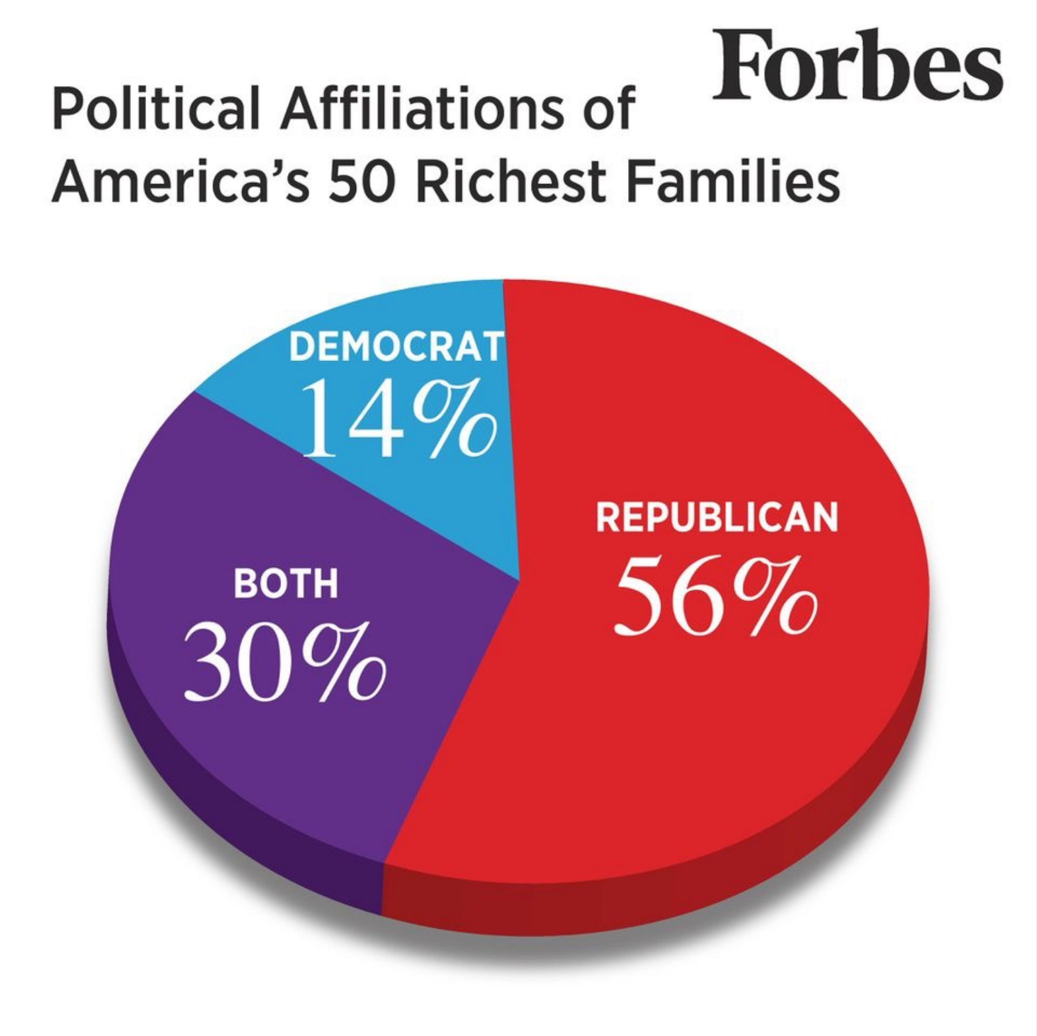


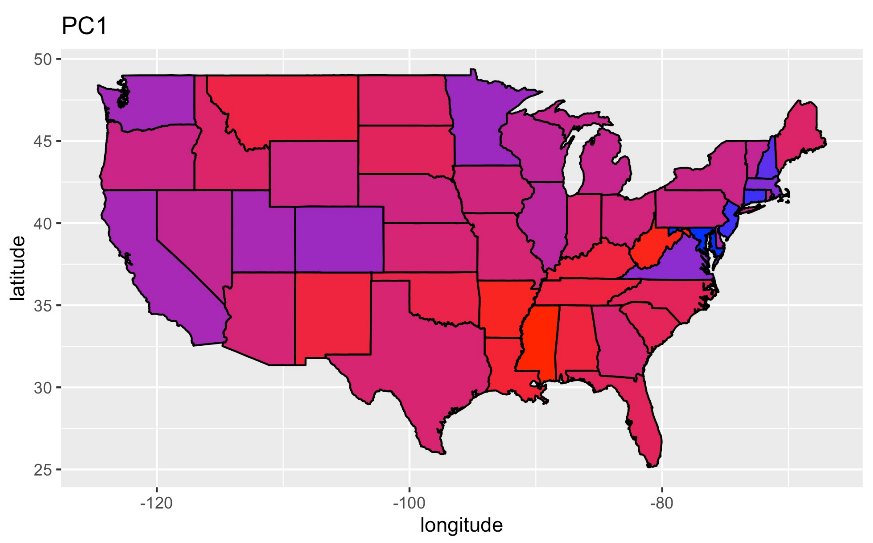


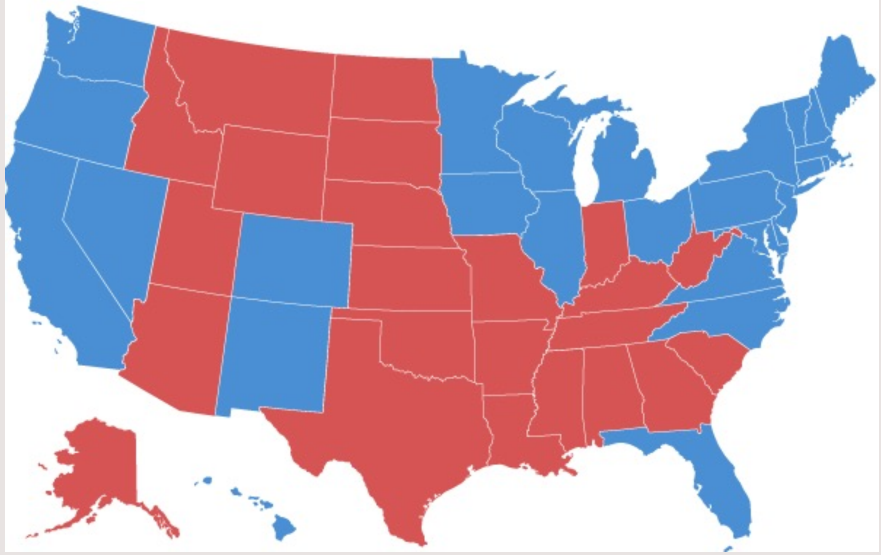
Forbes

Political Affiliations of America's 50 Richest Families

INDIVIDUAL AND GROUP REVERSED







Individual-level

Group-level

CONCLUSION

PCA on household income demonstrates:
The association identified at the group-level DOES NOT hold at the individual-level

THANK YOU!
ANY QUESTIONS?

SOURCES

- https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf
- https://en.wikipedia.org/wiki/Principal_component_analysis#:~:text=PCA%20was%20invented%20in%201901,Harold%20Hotelling%20in%20the%201930s
- https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues
- https://chartio.com/learn/charts/what-is-a-scatter-plot/
- https://builtin.com/data-science/step-step-explanation-principal-component-analysis
- https://www.nytimes.com/interactive/2016/05/04/upshot/electoral-map-trump-clinton.html
- https://www.forbes.com/sites/katiasavchuk/2014/07/09/are-americas-richest-families-republicans-or-democrats/?sh=12c46ef73e83
- Data: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjMs7agzJjvAhVfJzQIHSIgAu8QFjAAegQIBBAD&url=https%3A%2F%2Fwww2.census.gov%2Fprograms-surveys%2Fcps%2Ftables%2Ftime-series%2Fhistorical-income-households%2Fh08.xls&usg=AOvVaw0MqyTNSJP8VVd2wWKeoHeo