RuthEDA

Ruth Walters

2025-02-12

```
# Import dependencies
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.1
```

```
## corrplot 0.95 loaded
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.1
```

```
theme_set(theme_bw())

# Import data
mobility <- read.csv("mobility-all.csv", header = TRUE)
```

Exploratory data analysis

```
# Check which columns have NA/null values
print(colSums(is.na(mobility)))
```

```
##                      ID                    Name                Mobility
##                       0                       0                      12
##                   State              Population                   Urban
##                       0                       0                       0
##                   Black             Seg_racial              Seg_income
##                       0                       0                       0
##             Seg_poverty          Seg_affluence                 Commute
##                       0                       0                       0
##                  Income                    Gini                 Share01
##                       0                       0                      32
##                Gini_99            Middle_class          Local_tax_rate
```

```
##                         32                         32                          1
##        Local_gov_spending                Progressivity                       EITC
##                          2                          0                          0
##           School_spending      Student_teacher_ratio                Test_scores
##                         10                         30                         36
##                HS_dropout                   Colleges                    Tuition
##                        148                        157                        161
##                Graduation Labor_force_participation              Manufacturing
##                        160                          0                          0
##           Chinese_imports               Teenage_labor               Migration_in
##                         19                         32                         17
##             Migration_out               Foreign_born            Social_capital
##                         17                          0                         19
##                 Religious              Violent_crime            Single_mothers
##                         0                         27                          0
##                  Divorced                    Married                  Longitude
##                         0                          0                          0
##                  Latitude
##                         0
```

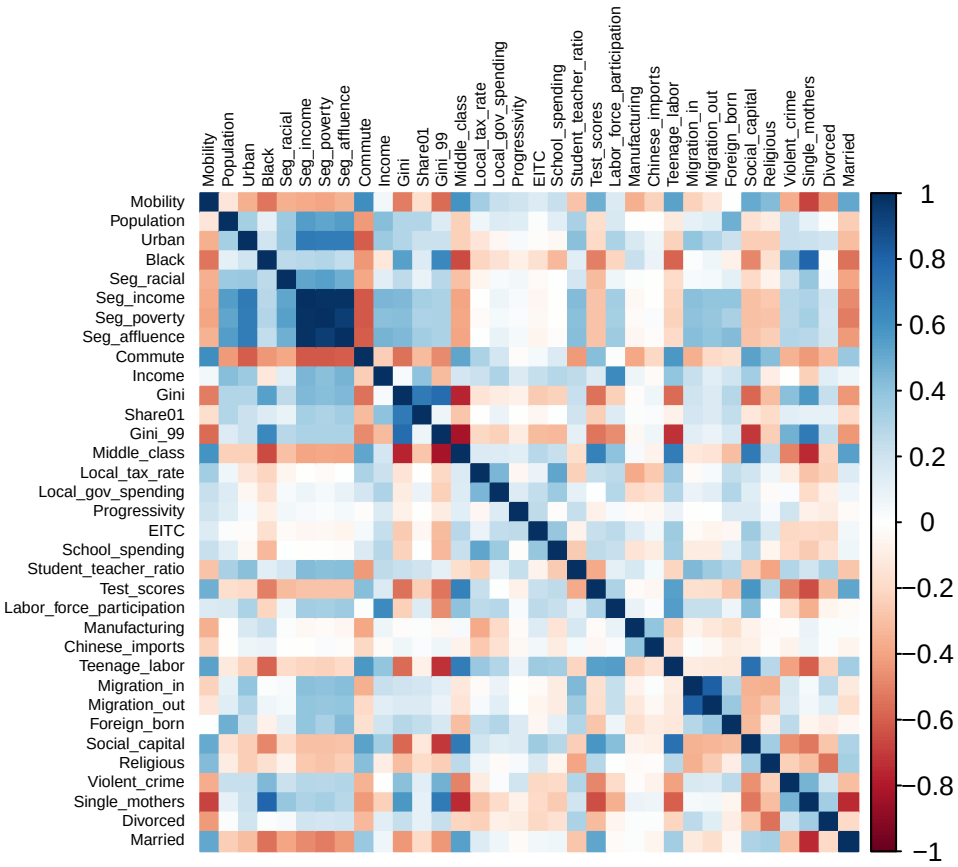
```
# Drop columns with >100 NA values
mobility <- mobility[,!(names(mobility) %in% c("Colleges","Tuition", "Graduation", "HS_dropout"))]

# Drop rows with NAs
mobility <- drop_na(mobility)
```

```
# Recheck NA
for (i in colSums(is.na(mobility))) {
  if (as.numeric(i) != 0) {print(i)}
}
```

```
mobility_numeric <- mobility[,!(names(mobility) %in% c("ID","Name", "State", "Latitude", "Longitude"))]

corrplot(cor(mobility_numeric),
         tl.col = "black",
         tl.cex = .5,
         method = 'color')
```



Mobility appears to be highly positively correlated with the cluster of variables that measure segregation

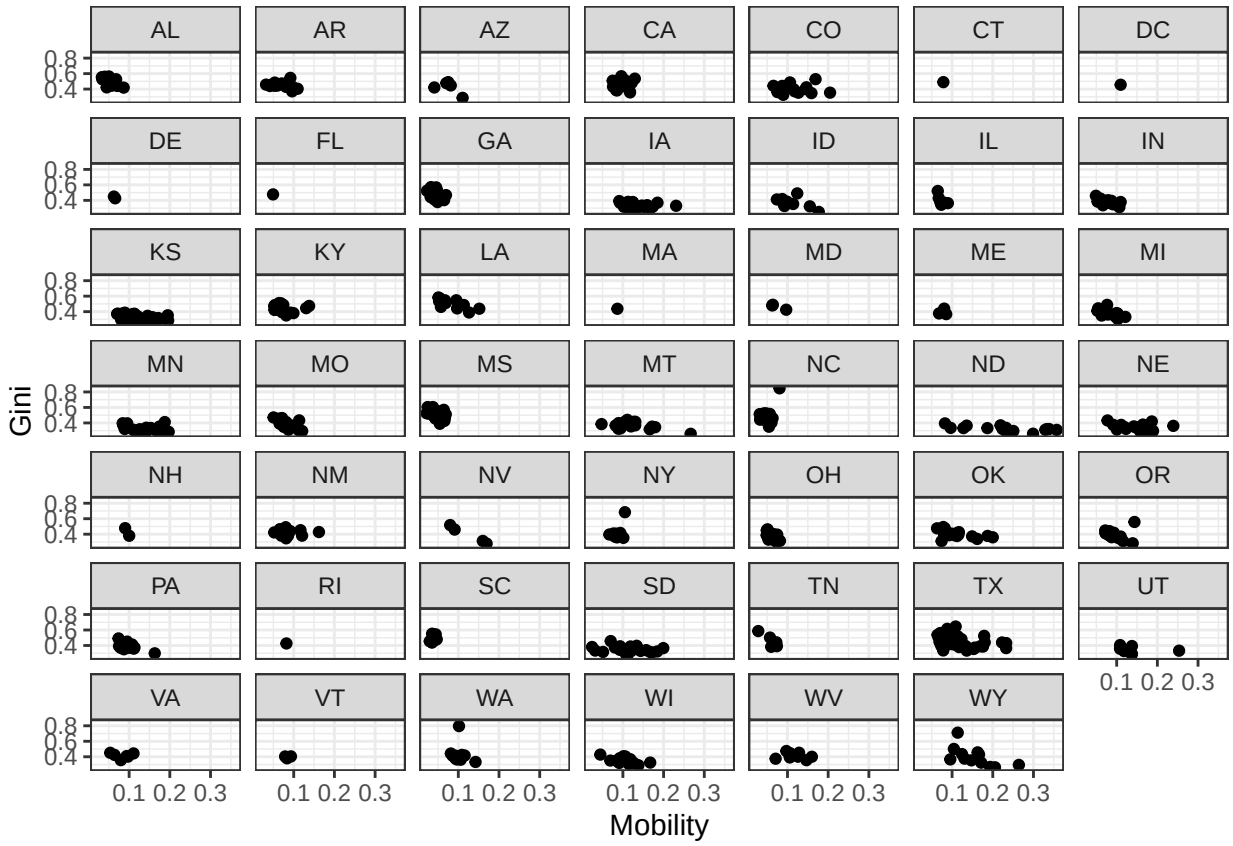
We can further identify three clusters of highly correlated variables:

- measures of segregation (`seg_racial`, `seg_income`, and `seg_affluence`)
- measures of the Gini index (`Gini`, `Share01`, `Gini_99` and `middle_class`)
- measures of migration (`migration_in` and `migration_out`)

```
# Drop highly correlated variables
mobility <- mobility[,!(names(mobility) %in% c("Seg_income","Seg_affluence", "Share01", "Gini_99"))]

mobility_numeric <- mobility[,!(names(mobility) %in% c("ID","Name", "State", "Latitude", "Longitude"))]
```

```
ggplot(mobility, aes(Mobility, Gini)) +
  geom_point() +
  facet_wrap(~State)
```


AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
NM
NV
NY
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Gini
0.8
0.6
0.4
0.1 0.2 0.3
Mobility