RuthEDA

Ruth Walters

2025-02-14

```
# 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)
```

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

library(ggplot2)
library(dplyr)
library(corrplot)
library(tidyr)
library(cowplot)
library(ggpubr)
```

```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:cowplot':
##
##     get_legend
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
theme_set(theme_bw())
theme_update(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  plot.title = element_text(size = 12, face = "italic")
)
```

From datacamp

[https://www.datacamp.com/tutorial/linear-regression-R][How to Do Linear Regression in R]

> "When a regression takes into account two or more predictors to create the linear regression, it's called multiple linear regression. In R, to add another coefficient, add the symbol"+" for every additional variable you want to add to the model.

Linear model: `lm([target] ~ [predictor], data = [data source])`

Data preparation

```
mobility <- read.csv("mobility-all.csv", header = TRUE, stringsAsFactors = TRUE)
```

Drop all non-quantitative rows

```
quals <- c("ID","Name", "State", "Latitude", "Longitude")

mobility <- mobility[,!(names(mobility) %in% quals)]
```

Drop low-quality columns

```
print(colSums(is.na(mobility)))
```

```
##                  Mobility                Population                     Urban
##                        12                         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
##                         0                         0
```

```
bad_cols <- c("Colleges","Tuition", "Graduation", "HS_dropout") # +100 NULL
mobility <- mobility[,!(names(mobility) %in% bad_cols)]
```

Drop remaining NULLS

```
before <- nrow(mobility)
mobility <- drop_na(mobility)
dropped <- before - nrow(mobility)

print("Data reduced by: ")
```

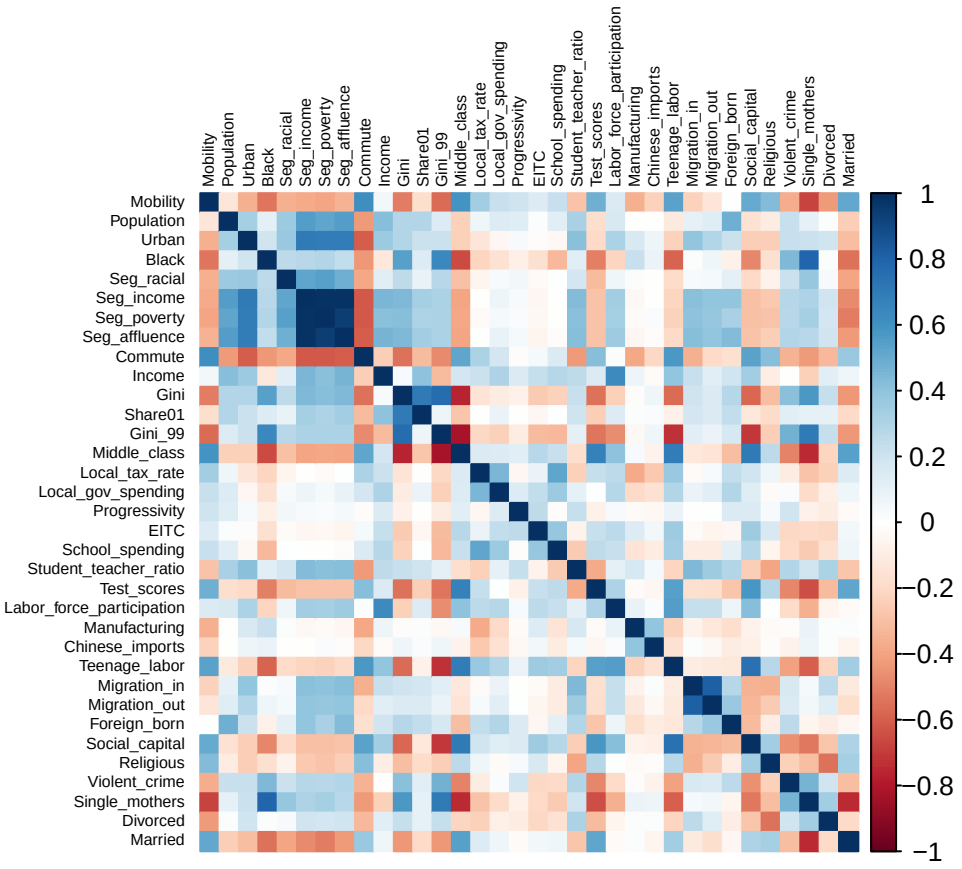
```
## [1] "Data reduced by: "
```

```
print((dropped/before))
```

```
## [1] 0.145749
```

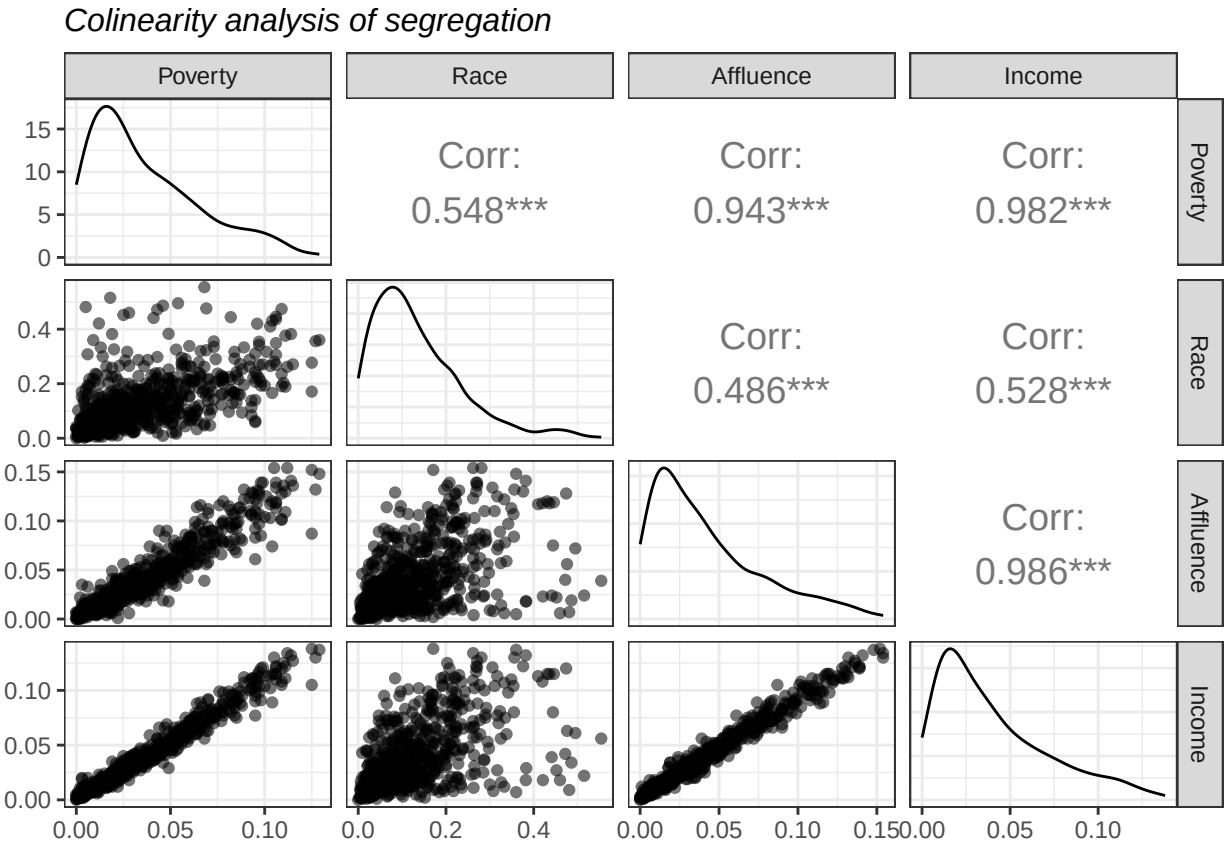
Exploratory data analysis

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



Explore highly correlated variables

```
mobility[c("Seg_poverty", "Seg_racial", "Seg_affluence", "Seg_income")] %>%
  ggpairs(aes(alpha = 0.5),
          upper = list(continuous = wrap("cor", size = 5)),
          columnLabels = c("Poverty", "Race", "Affluence", "Income"),
          title = "Colinearity analysis of segregation",
          progress = FALSE)
```

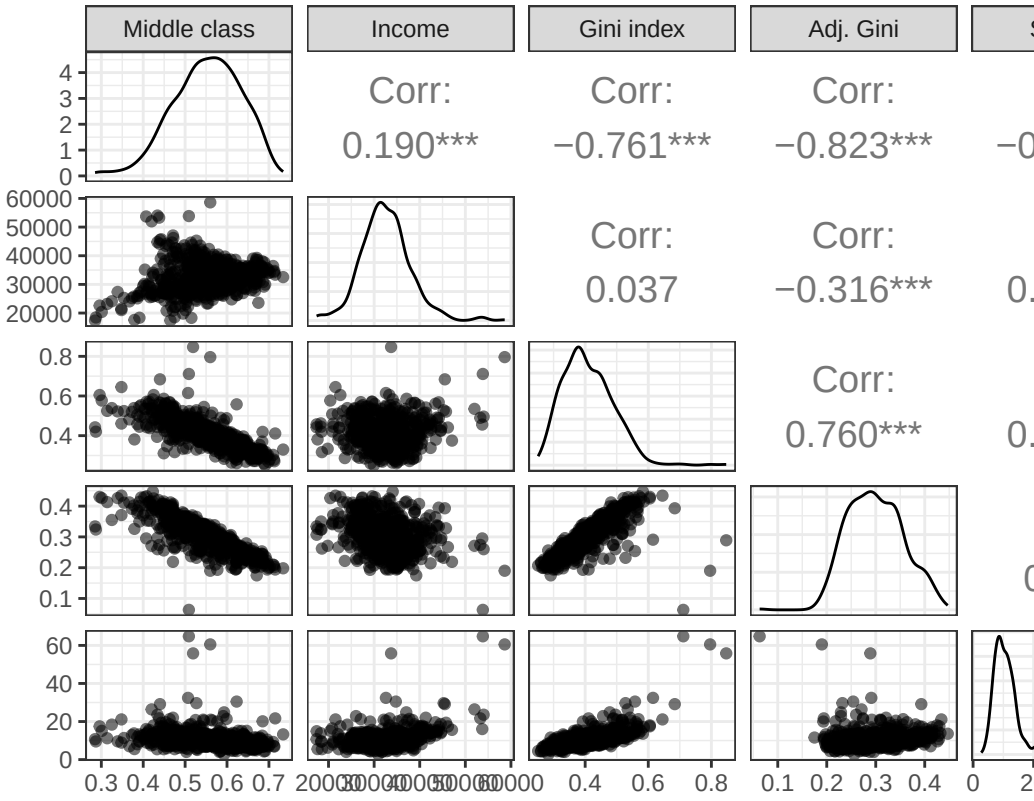


Segregation

While segregation on poverty lines is not particularly well correlated with segregation on racial lines, is is highly associated with segregation by affluence and segregation by income, which are also highly associated with each other. Since `Seg_poverty`, `Seg_affluence` and `Seg_income` are so strongly co-linear, `Seg_affluence` and `Seg_income` will be removed from the model.

```
mobility[c("Middle_class", "Income", "Gini", "Gini_99", "Share01")] %>%
  ggpairs(aes(alpha = 0.5),
          upper = list(continuous = wrap("cor", size = 5)),
          columnLabels = c("Middle class", "Income", "Gini index", "Adj. Gini", "Share01"),
          title = "Colinearity analysis of income and income inequality",
          progress = FALSE)
```

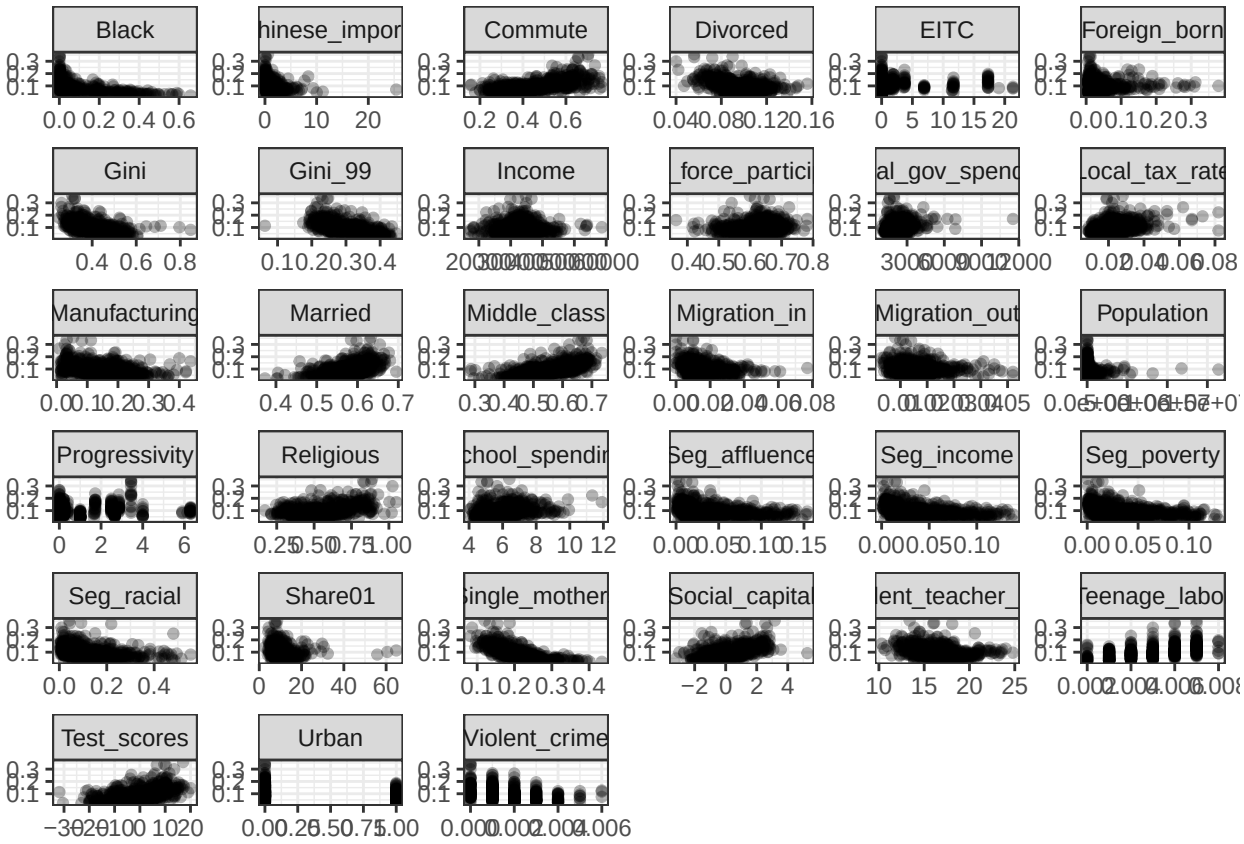
Colinearity analysis of income and income inequality



Income and income inequality

Explore non-linear variables

```
mobility %>%
  gather(-Mobility, key = "var", value = "value") %>%
  ggplot(aes(x = value, y = Mobility)) +
  geom_point(alpha = 0.3) +
  facet_wrap(~ var, scales = "free")
```



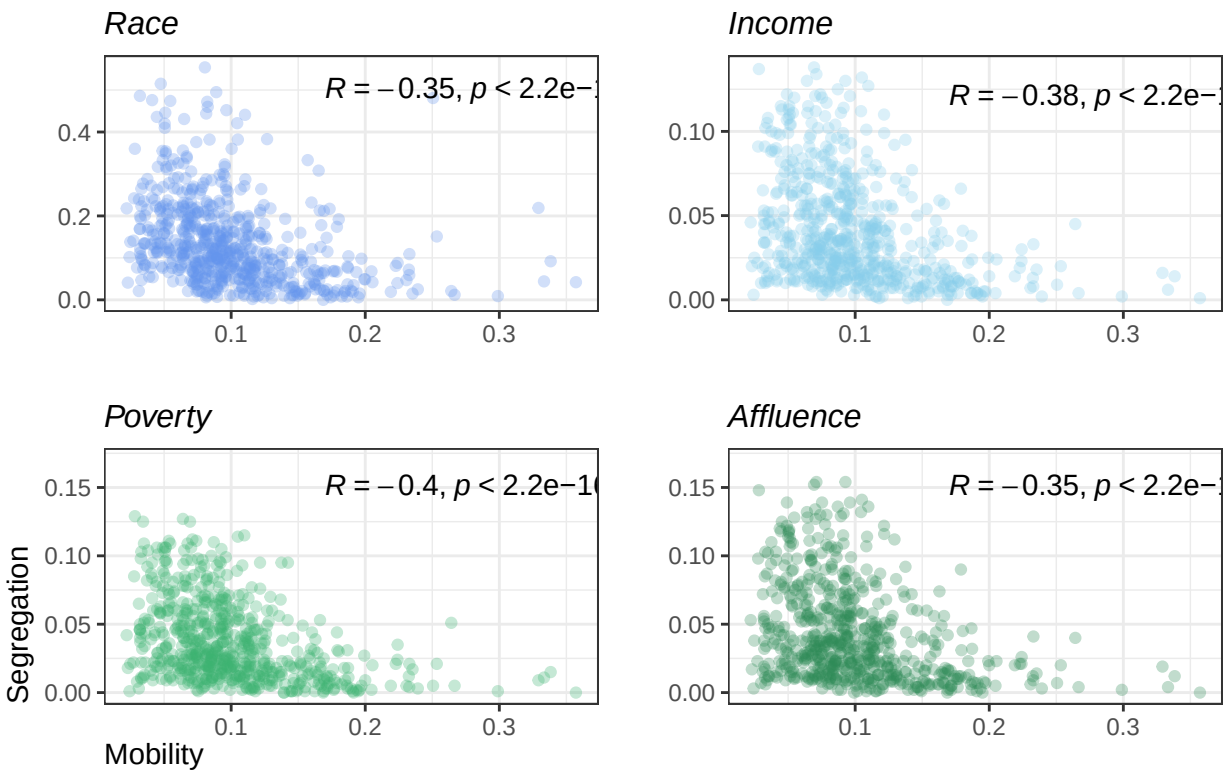
Social determinants of mobility

Goal: explore potential social determinants of mobility

Selected variables: - segregation variables `Seg_racial`, `Seg_income`, `Seg_poverty`, and `Seg_affluence` - educational variables `School_spending`, `Student_teacher_ratio`, and `Test_scores` - family dynamic variables `Single_mothers`, `Divorced`

```
a <- ggplot(data = mobility, aes(x = Mobility, y = Seg_racial)) +
  geom_point(color = "cornflowerblue", alpha = .3) +
  #stat_smooth(method = "lm", formula = y ~ x, geom = "line", color = "darkorange") +
  stat_cor(label.x=.17, label.y=.5) +
  ggtitle("Race")

b <- ggplot(data = mobility, aes(x = Mobility, y = Seg_income)) +
  geom_point(color = "skyblue", alpha = .3) +
  stat_cor(label.x=.17, label.y=.12) +
  ggtitle("Income")

c <- ggplot(data = mobility, aes(x = Mobility, y = Seg_poverty)) +
  geom_point(color = "mediumseagreen", alpha = .3) +
  stat_cor(label.x=.17, label.y=.15) +
  ylim(0,.17) +
  ggtitle("Poverty") +
  xlab("Mobility") +
  ylab("Segregation") +
  theme(axis.title.x = element_text(hjust = 0),
        axis.title.y = element_text(angle=90, hjust = 0, margin = margin(r = 5)))
```

```
d <- ggplot(data = mobility, aes(x = Mobility, y = Seg_affluence)) +
  geom_point(color = "seagreen", alpha = .3) +
  stat_cor(label.x=.17, label.y=.15) +
  ylim(0,.17) +
  ggtitle("Affluence")

plot_row <- plot_grid(a,b,c,d, align = "hv")

title <- ggdraw() +
  draw_label(
    "Segregation as a predictor of mobility",
    fontface = 'bold',
    x = 0,
    hjust = 0) +
  theme(plot.margin = margin(0, 0, 0, 7))
plot_grid(
  title, plot_row,
  ncol = 1,
  rel_heights = c(0.1, 1)
)
```



Remove highly correlated variables:

```
mobility <- mobility[,!(names(mobility) %in% c("Seg_income", "Seg_affluence", "Gini_99", "Share01"))]
```

```
mobility$Urban <- as.factor(mobility$Urban)

a <- ggplot(data = mobility, aes(x = Mobility, y = Gini, col = Progressivity)) +
```

```
  geom_point(alpha = .3) +
  ggtitle("Progressivity") +
  xlab(" ") +
  theme(axis.title.x = element_text(hjust = 0))

b <- ggplot(data = mobility, aes(x = Mobility, y = Single_mothers)) +
  geom_point(color = "skyblue", alpha = .3) +
  stat_cor(label.x=.17, label.y=.3, label.size = 0.05) +
  ggtitle("Proportion of single mothers")
```

```
## Warning in stat_cor(label.x = 0.17, label.y = 0.3, label.size = 0.05): Ignoring
## unknown parameters: `label.size`
```

```
c <- ggplot(data = mobility, aes(x = Mobility, y = Gini, col = Urban)) +
  geom_point( alpha = .3) +
  ggtitle("Urban communities") +
  xlab("Mobility") +
  ylab("Gini index") +
  theme(axis.title.x = element_text(hjust = 0),
        axis.title.y = element_text(angle=90, hjust = 0, margin = margin(r = 5)))

d <- ggplot(data = mobility, aes(x = Mobility, y = Violent_crime)) +
  geom_point(color = "seagreen", alpha = .3) +
  stat_cor(label.x=.17, label.y=.004) +
  ggtitle("Violent crime incidence")

plot_row <- plot_grid(c,a,b,d, align = "none")

title <- ggdraw() +
  draw_label(
    "Community factors associated with mobility",
    fontface = 'bold',
    x = 0,
    hjust = 0) +
  theme(plot.margin = margin(0, 0, 0, 7))
plot_grid(
  title, plot_row,
  ncol = 1,
  rel_heights = c(0.1, 1)
)
```

Community factors associated with mobility