Social Statistics

Introducing Association and Correlation

October 3, 2023

Assignment 2 Recap

General Thoughts

- Suppress messages with `#| message: false`
- Remember to render as a pdf
- Headings require a space between hashtags and text
  - → This will not render a heading: `##Heading`
  - → This will render a heading: `## Heading`
- Answer the question in full sentences (don't just show the code and output)
- Leaving comments for me is helpful

Assignment 2 Recap

Question 1 - Create a histogram showing the distribution of the `hrs1` variable.

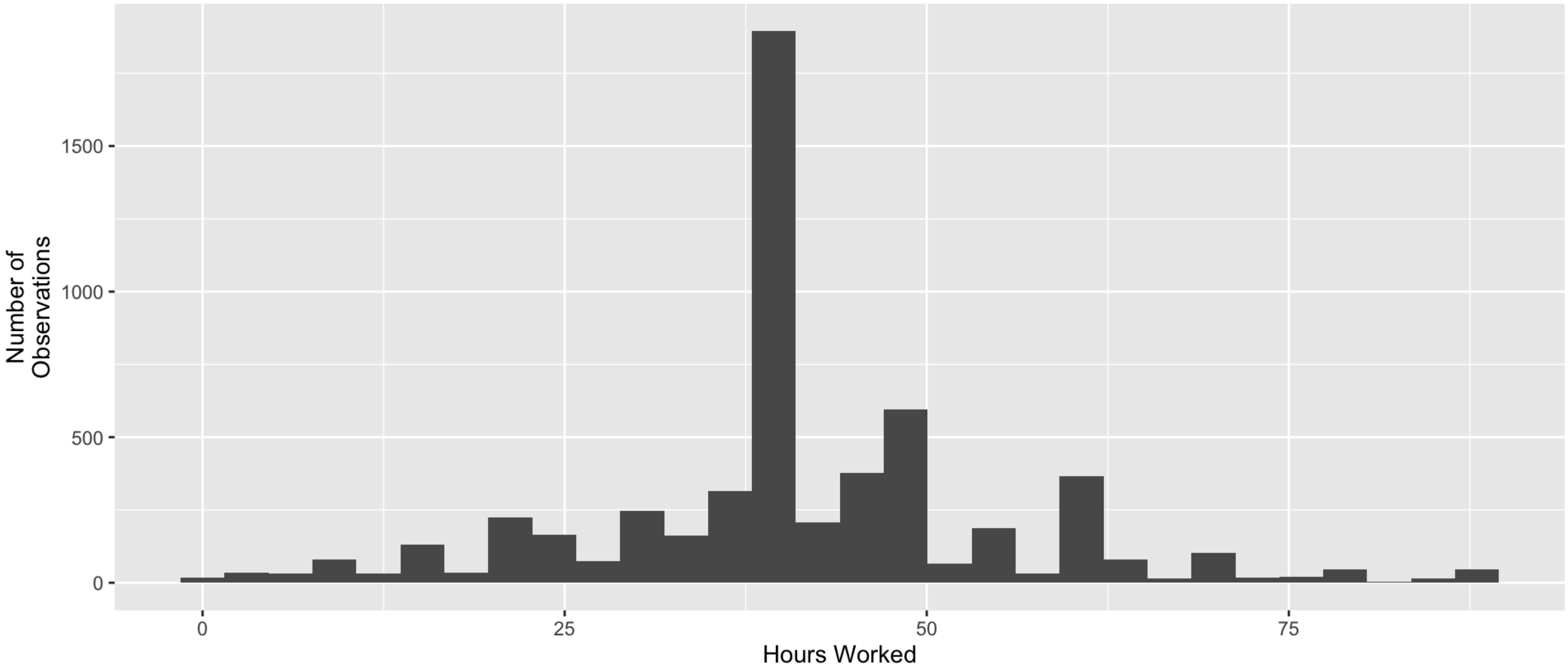
```r
# Create plot as an object
hours_histogram <- ggplot(assignment_02, aes(x = hrs1))

# Add geom layer
hours_histogram + geom_histogram() +
# Add title
  labs(x = "Hours Worked",
       y = "Number of\nObservations", # \n = new line
       title = "Distribution of Hours Worked",
       subtitle = "GSS, 2012 - 2021",
       caption = "ML for SOCI 385 Fall 2023")
```

Assignment 2 Recap

Distribution of Hours Worked

GSS, 2012 - 2021



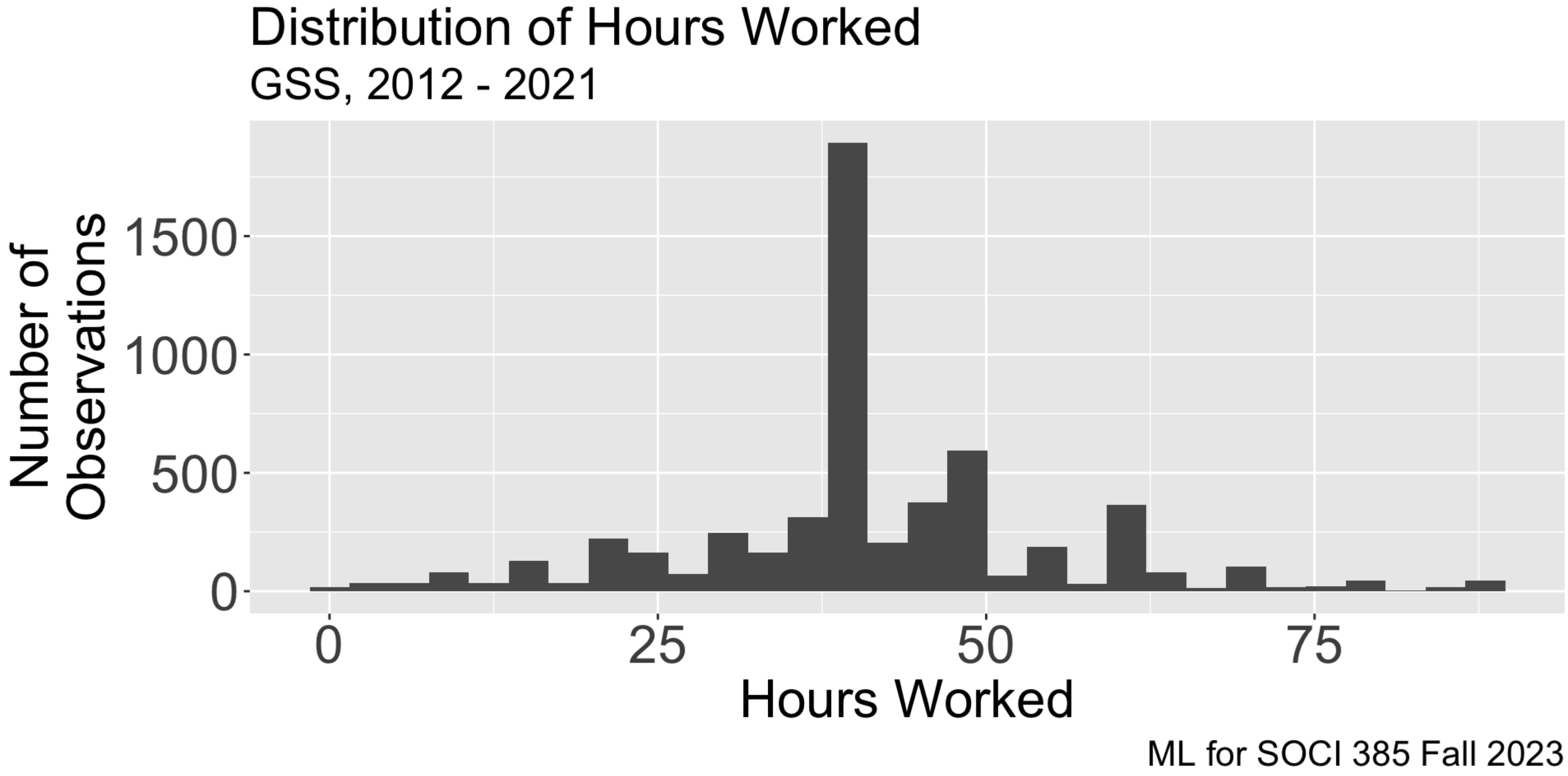
ML for SOCI 385 Fall 2023

Assignment 2 Recap

Question 1 - Change title fonts?

```r
hours_histogram + geom_histogram() +
  labs(x = "Hours Worked",
       y = "Number of\nObservations", # \n = new line
       title = "Distribution of Hours Worked",
       subtitle = "GSS, 2012 - 2021",
       caption = "ML for SOCI 385 Fall 2023") +
# Add theme to change text size
     theme(axis.text.x = element_text(size = 24),
           axis.text.y = element_text(size = 24),
           axis.title = element_text(size = 24),
           plot.title = element_text(size = 24),
           plot.subtitle = element_text(size = 20),
           plot.caption = element_text(size = 16))
```

Assignment 2 Recap



Assignment 2 Recap

Question 2 - Create a figure with overlapping density plots for each social class.

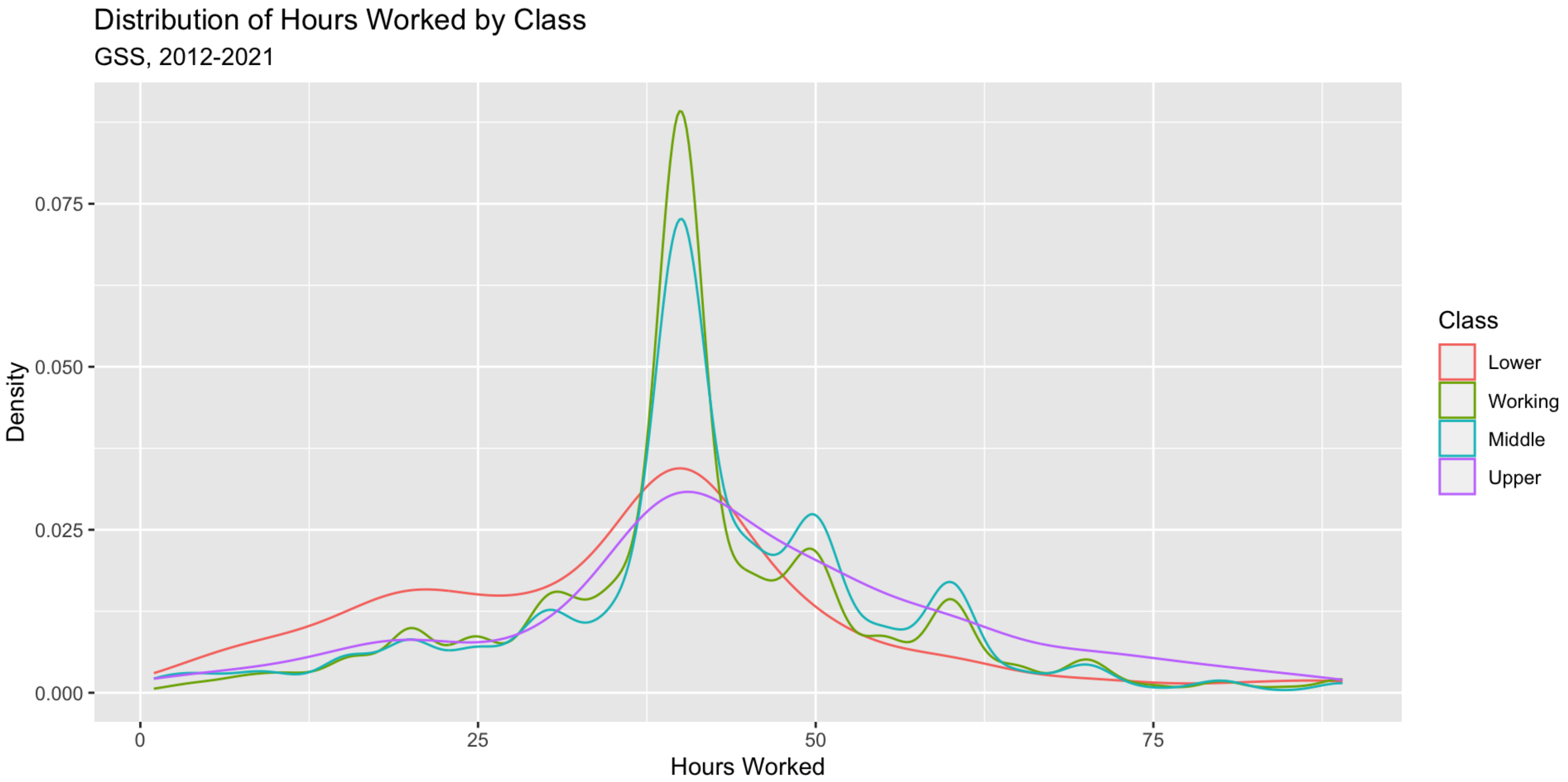
```r
# Create factor variable
assignment_02 <- assignment_02 |>
  mutate(class = factor(class,
    levels = c("Lower", "Working", "Middle", "Upper")))

# Create plot as object
hours_class_plot <- ggplot(assignment_02,
  aes(x = hrs1, color = class))

# Add geom layer and title
hours_class_plot + geom_density() +
      labs(x = "Hours Worked", y = "Density",
            title = "Distribution of Hours Worked by Class",
            subtitle = "GSS, 2012-2021",
            color = "Class")
```

Assignment 2 Recap

Question 2 - Create a figure with overlapping density plots for each social class.



Assignment 2 Recap

Question 3 - Create a new factor variable for hours worked with the following categories: less than 20, 20-39, 40, more than 40. The levels should be in order from least to most hours.

```r
# Create new variable
assignment_02 <- assignment_02 |>
     mutate(hours_cat = ifelse(hrs1<20, "Less Than 20",
                               ifelse(hrs1 %in% 20:39, "20-39",
                                      ifelse(hrs1==40, "40",
                                             "More Than 40"))),
# Order levels
            hours_cat = factor(hours_cat,
               levels = c("Less Than 20", "20-39", "40", "More Than 40")))
```

Assignment 2 Recap

Question 4 - Create a table showing the proportion of respondents in each region who are in each category of hours worked. Which region has the highest proportion of respondents who work more than forty hours a week?

```
round(prop.table(table(assignment_02$region,
                       assignment_02$hours_cat),1),3) # 1 for row proportions!
```

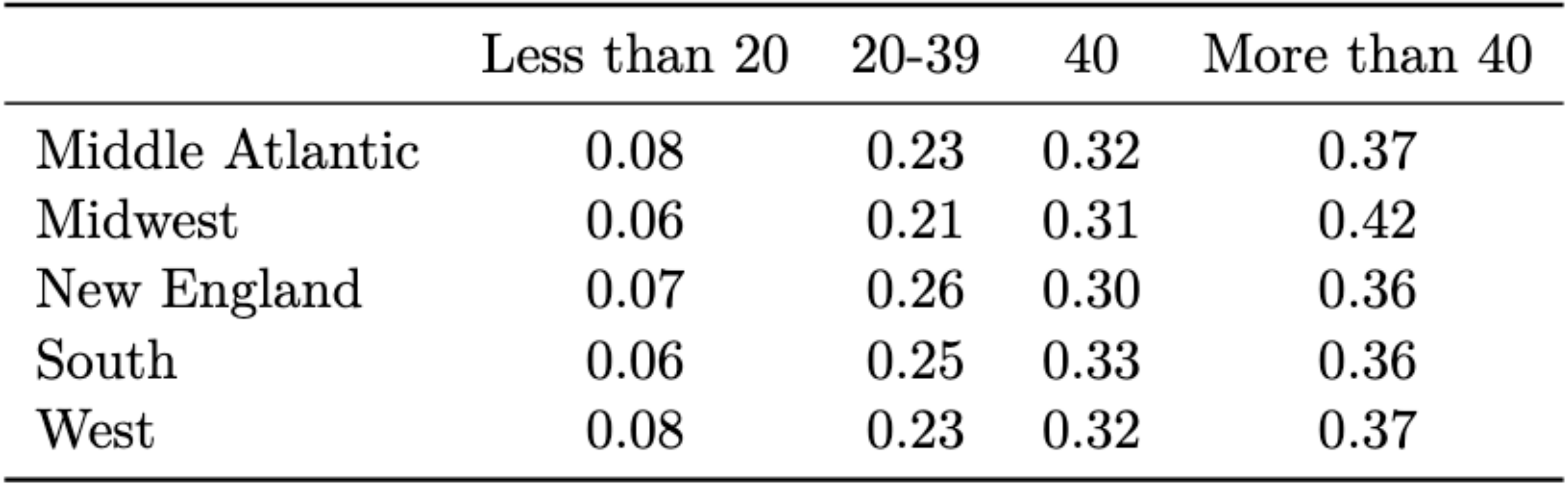
```
                Less Than 20 20-39    40 More Than 40
Middle Atlantic        0.076 0.232 0.320        0.371
Midwest                0.057 0.212 0.313        0.417
New England            0.070 0.264 0.303        0.363
South                  0.056 0.251 0.331        0.363
West                   0.080 0.229 0.323        0.368
```

Using Kable To Improve Tables

```r
# install and load package
install.packages("kableExtra") # hashtag this line after installing
library(kableExtra)

# Use the table as a dataframe and add the kable functions
round(prop.table(table(assignment_02$region,
                       assignment_02$hours_cat),1),3) |>
    kable(booktabs = TRUE, # For basic formating
          align = rep('c', 4)) # For *center* alignment in the *4* columns
```

Using Kable To Improve Tables



	Less than 20	20-39	40	More than 40
Middle Atlantic	0.08	0.23	0.32	0.37
Midwest	0.06	0.21	0.31	0.42
New England	0.07	0.26	0.30	0.36
South	0.06	0.25	0.33	0.36
West	0.08	0.23	0.32	0.37

Using Kable To Improve Tables

	Less Than 20	**20-39**	**40**	**More Than 40**
Middle Atlantic	0.076	0.232	0.320	0.371
Midwest	0.057	0.212	0.313	0.417
New England	0.070	0.264	0.303	0.363
South	0.056	0.251	0.331	0.363
West	0.080	0.229	0.323	0.368

Assignment 2 Recap

Question 5 - Using tidyverse functions, find the standard deviation of hours worked for each race/ethnicity category in the New England region.

```r
assignment_02 |>
  filter(region == "New England") |> # to choose only specific *rows*
  group_by(racehisp) |> # do something for each value of this variable
  summarise(sd_hrs = sd(hrs1)) # what to do for each group
```

```
# A tibble: 4 × 2
  racehisp sd_hrs
  <chr>     <dbl>
1 Black      9.53
2 Hispanic  14.7
3 Other     16.5
4 White     13.6
```

Measures of Association

- Last week: describing two categorical variables
- This week: describing relationship between two numerical variables
- Remember a hypothesis: How variables *tend to move together*
  - → How close or tight are the values? How well do they resemble a straight line?
  - → The spread that they tend to share = covariance
  - → The change they tend to share = correlation

Interpreting Correlations

- Easier to interpret correlations than covariances
- Always bounded by -1, 1
- Association is linear (for now)
- Positive correlation > 0
  - → When X is larger than its mean, likely that Y is larger than its mean
- Negative correlation < 0
  - → When X is larger than its mean, unlikely that Y is larger than its mean

Interpreting Correlations

- Correlation of X, Y = Correlation of Y, X
  - → But still think of axes
- Not affected by changes in scale
  - → Can multiply all the values by a constant and the correlation is still the same
  - → Temperature degrees, currencies, etc.
- But can be affected by outliers

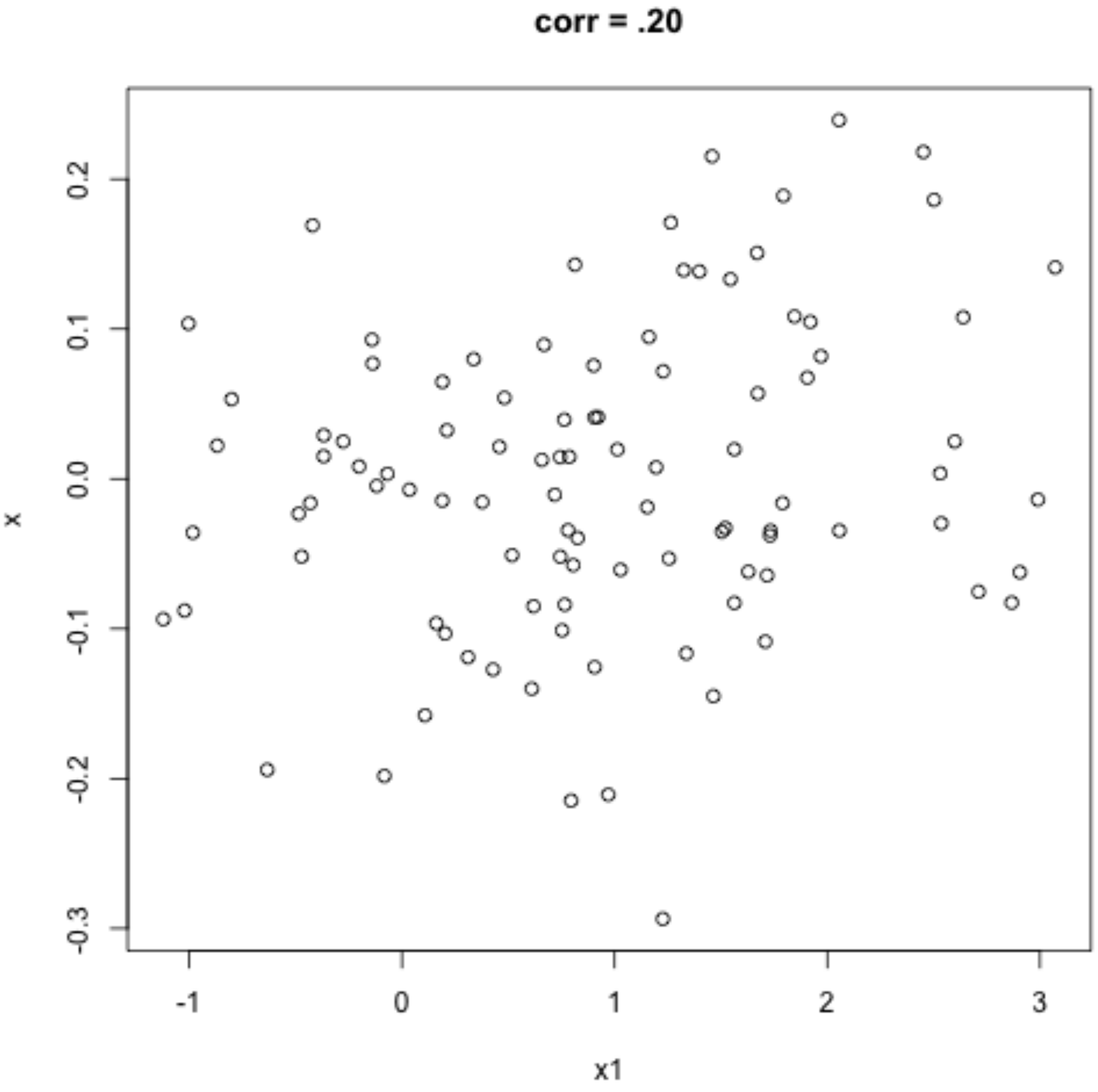
Strength of Associations

- Positive correlation coefficients look like proportions but they are not
- Strong association: knowing a value of one variable helps predict a value of the other variable
- Weak association: too much variability to use the value of one variable to make a good guess about the value of the other variable
- Remember: Not causal!
- Keep in mind: *strong* is not always *better*

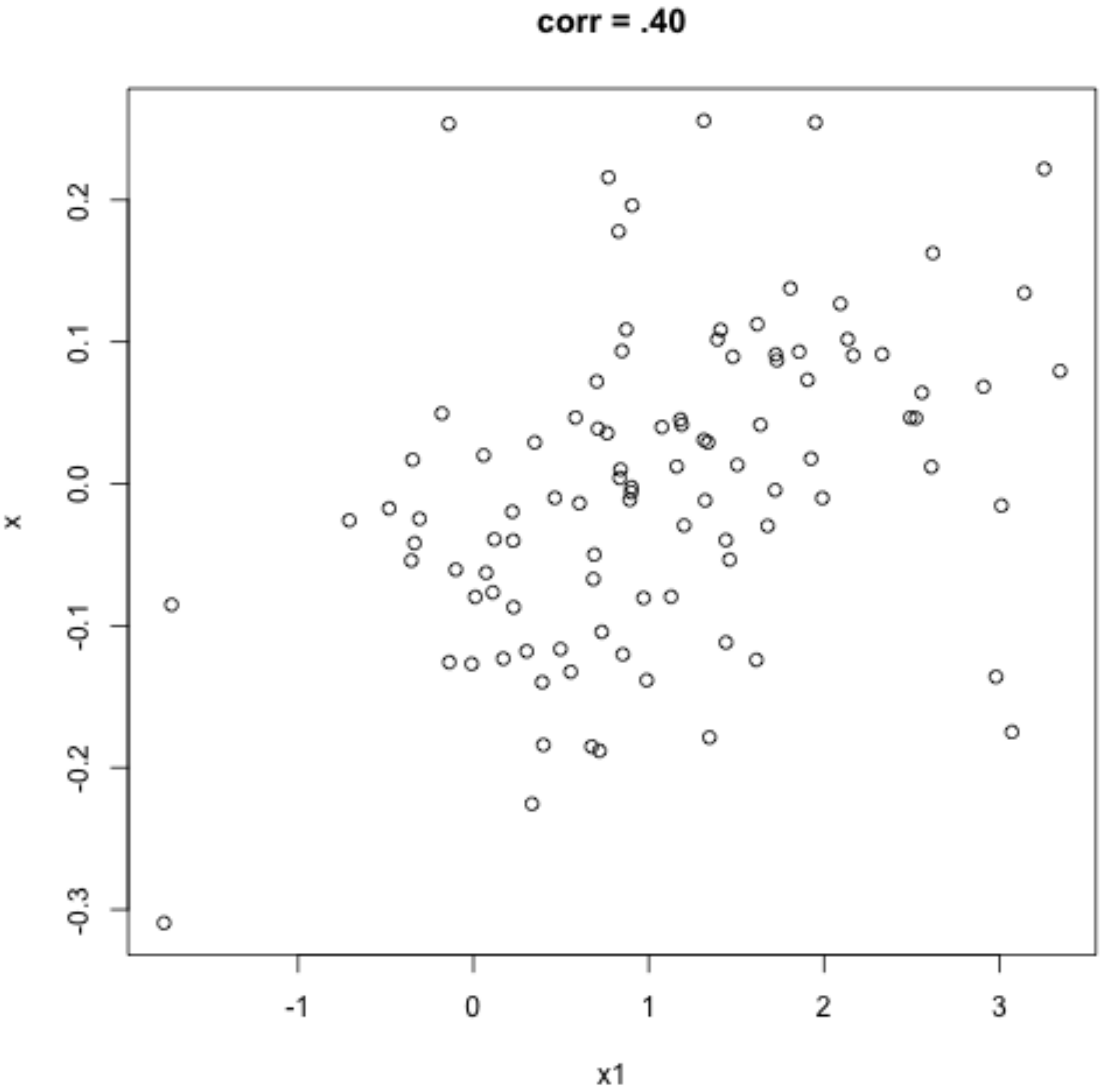
Strength of Associations

- No association = 0 to .19 or 0 to -.19
- Weak association = .20 to .29 or -.20 to -.29
- Moderate association = .30 to .49 or -.30 to -.49
- Strong association = .50 to .69 or -.50 to -.69
- Very strong association = greater than .70 or less than -.70

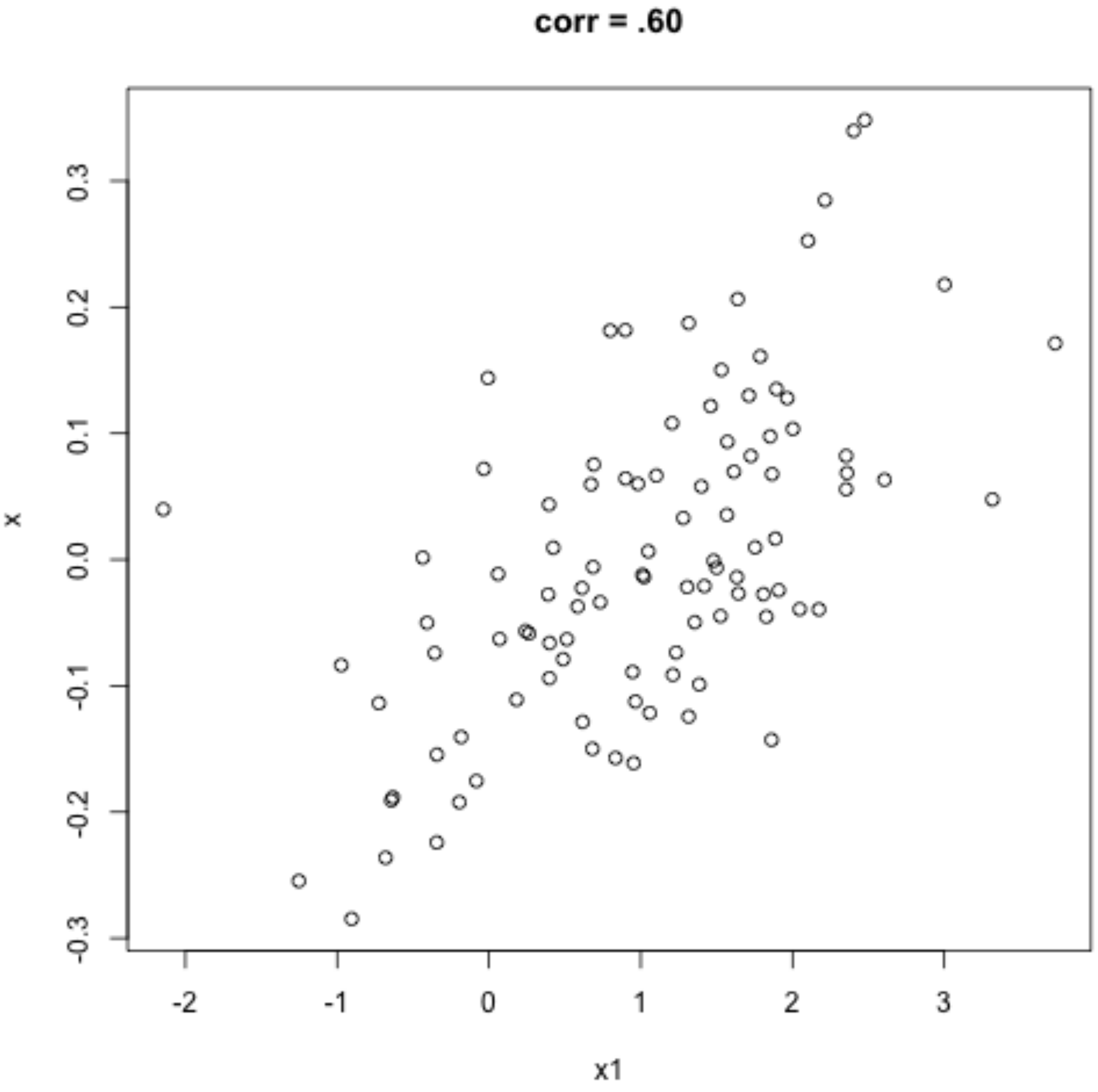
Scatterplots



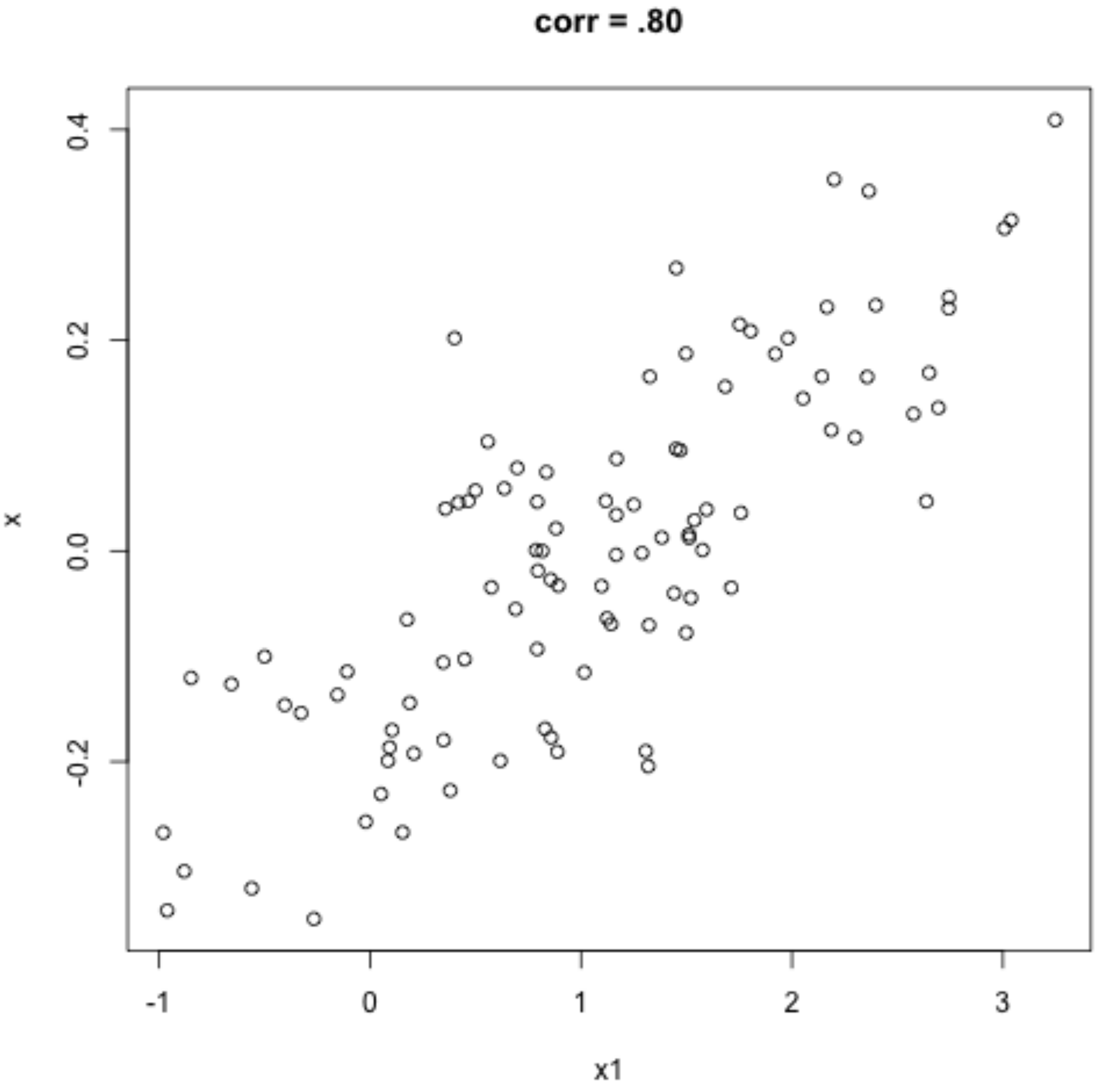
Scatterplots



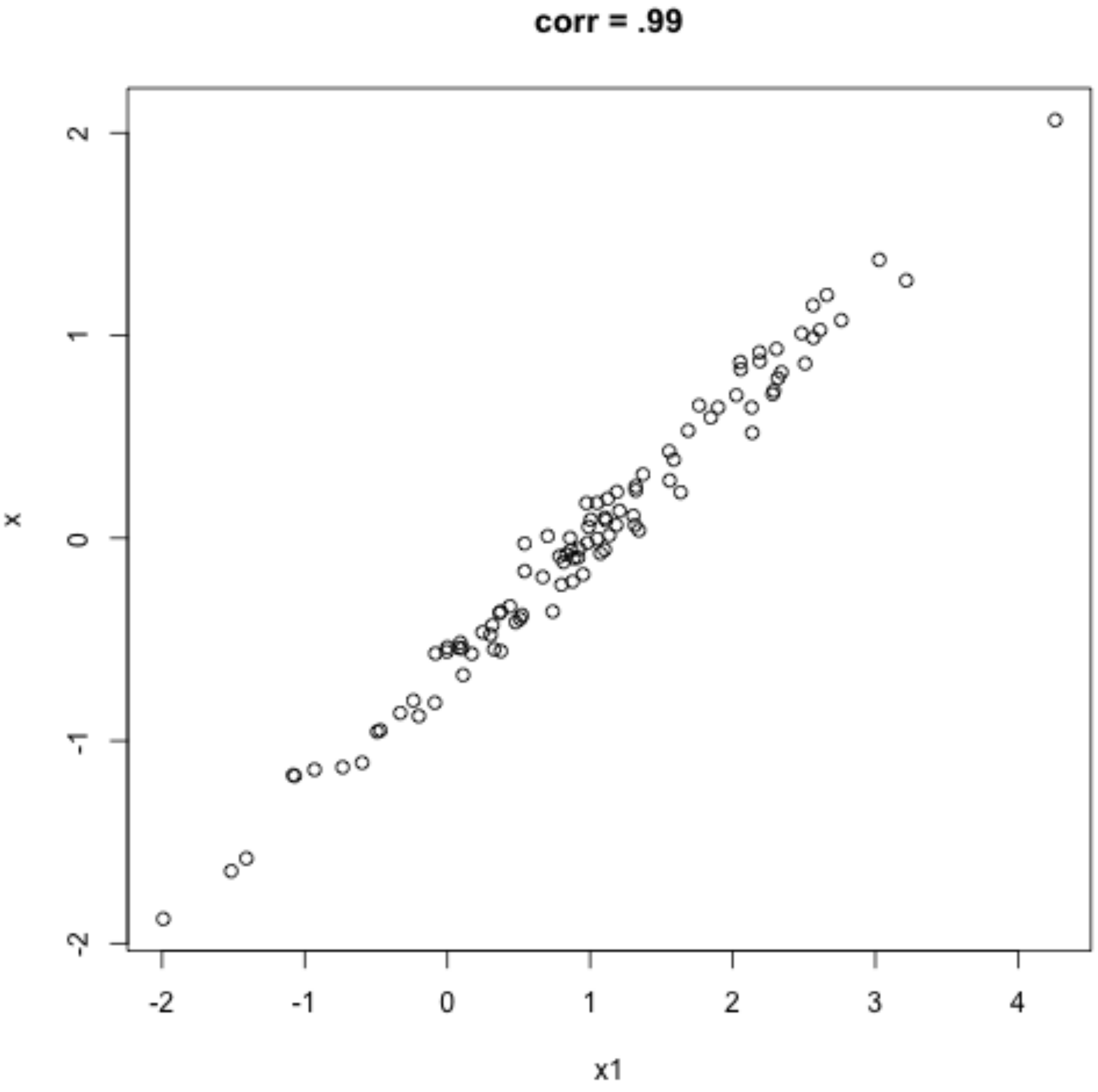
Scatterplots



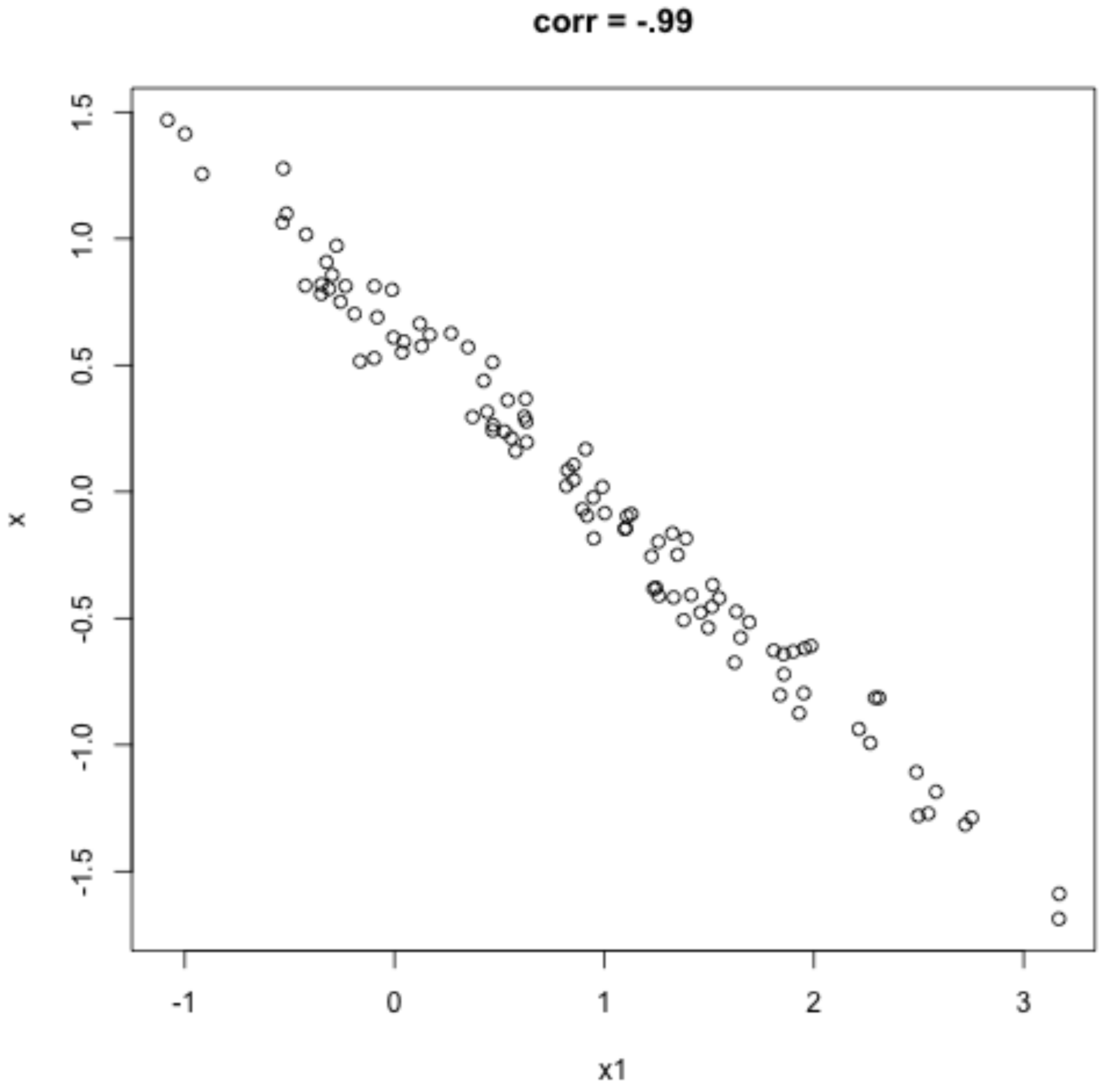
Scatterplots



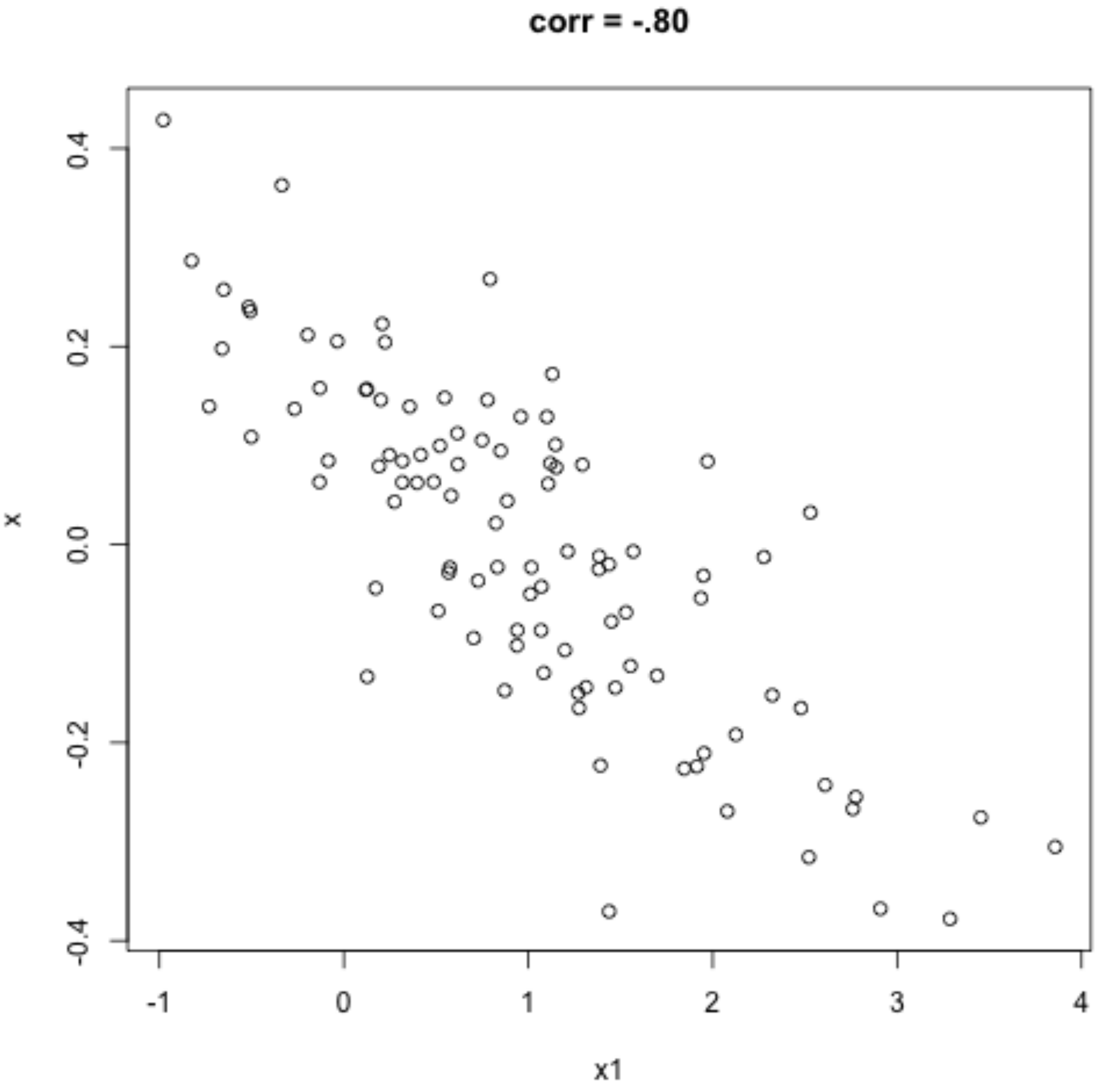
Scatterplots



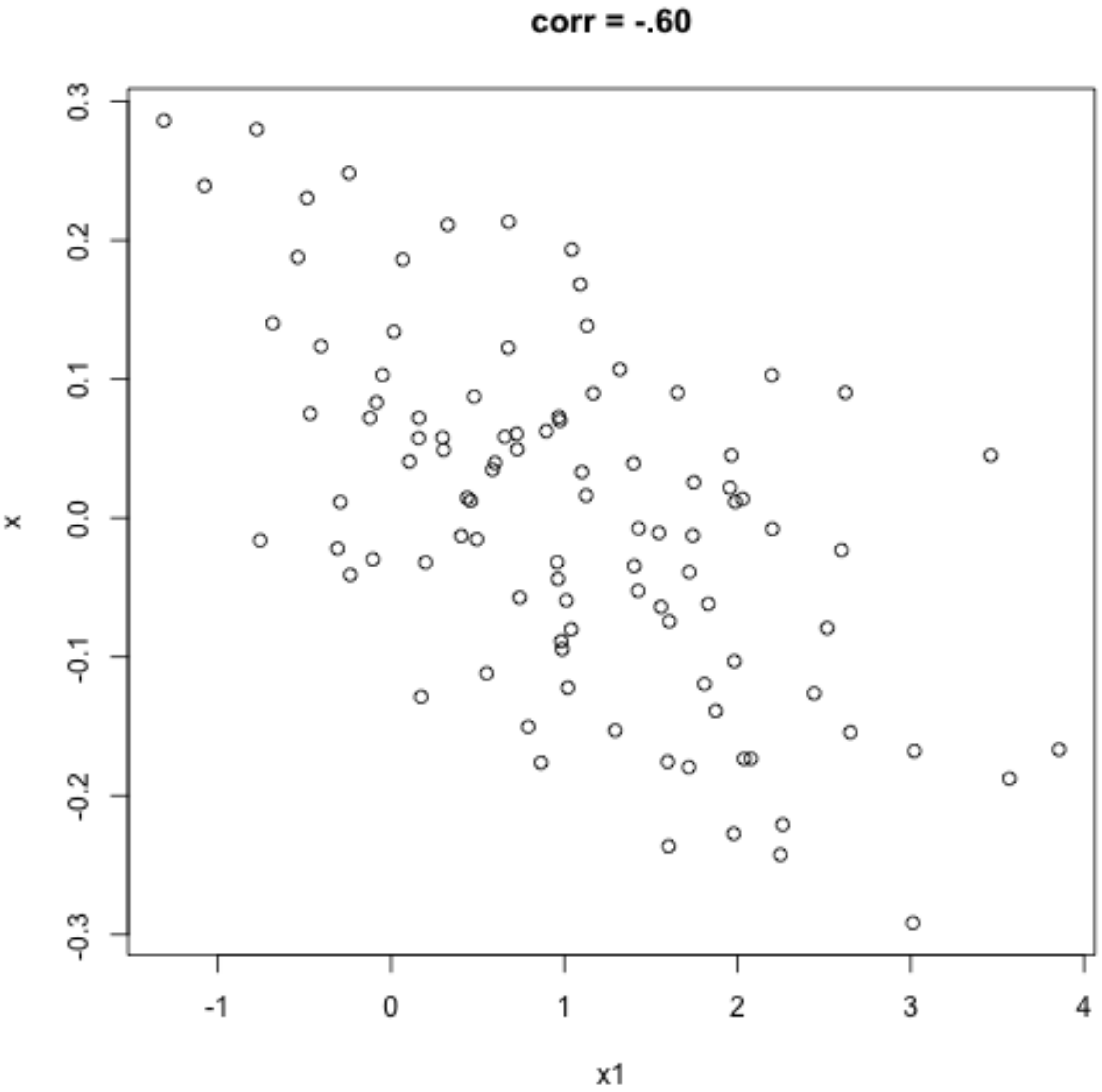
Scatterplots



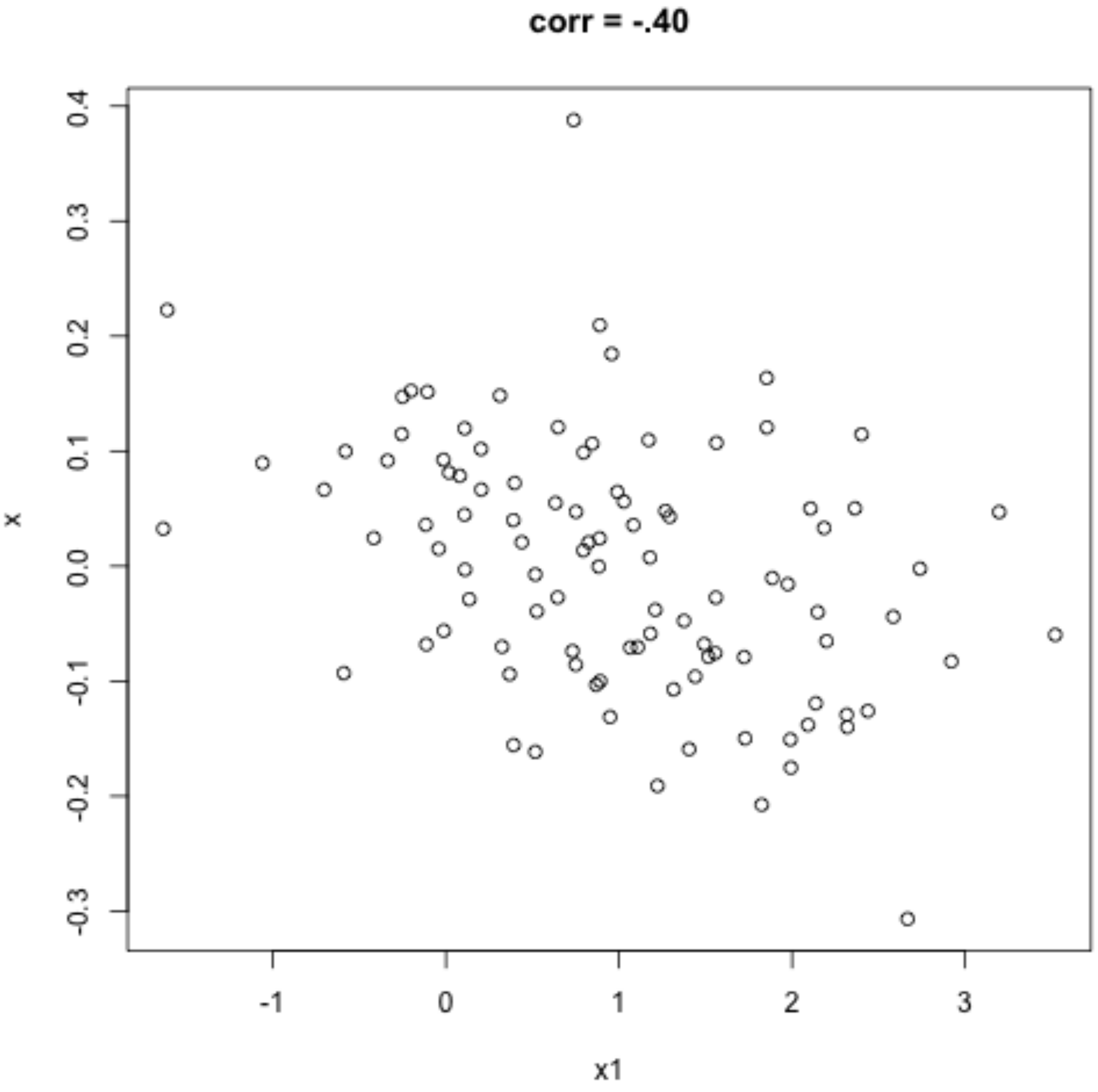
Scatterplots



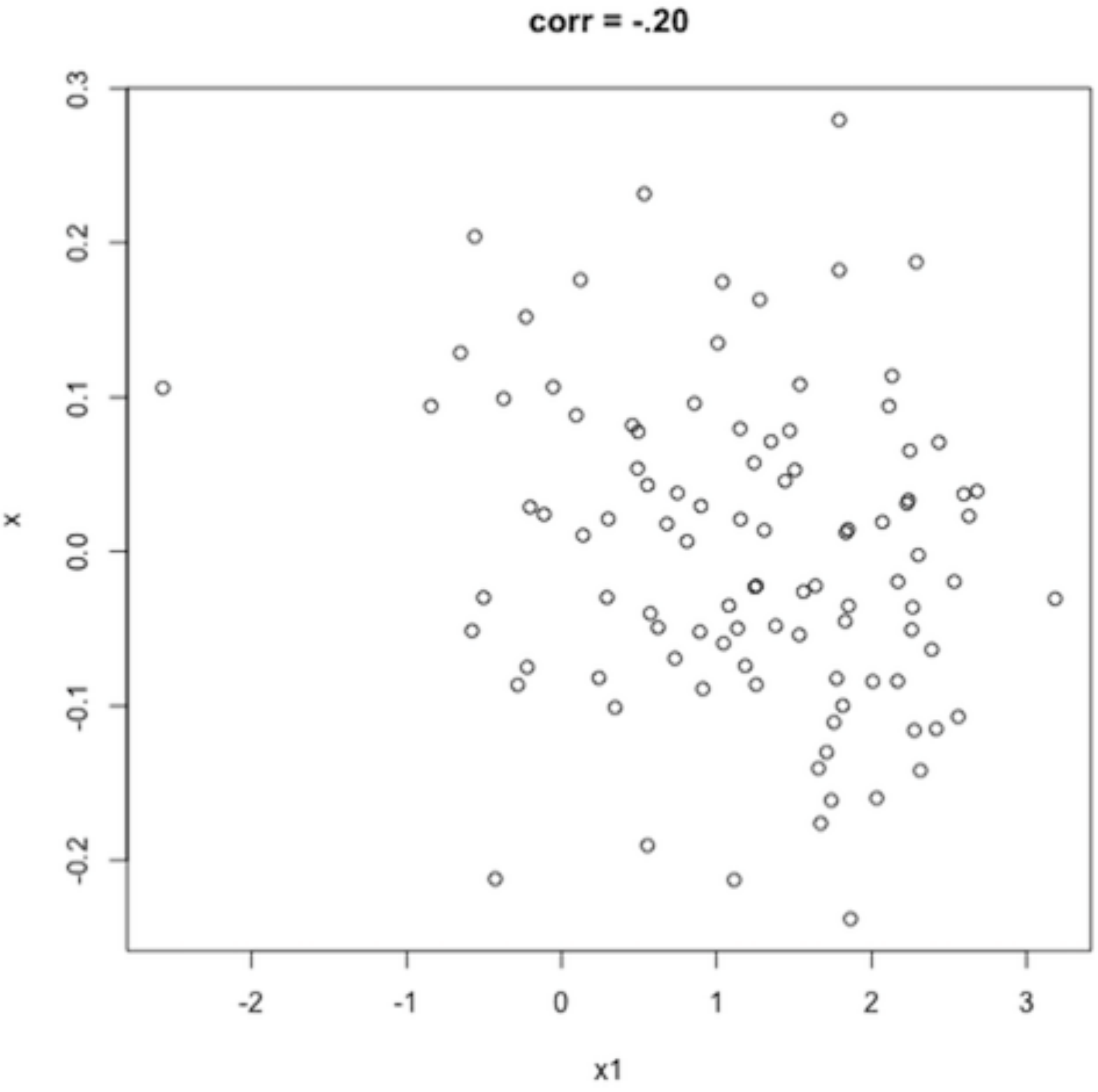
Scatterplots



Scatterplots



Scatterplots



An Example: Chetty et al (2014)

- Comparing income mobility rates across regions
- Regional variation across 741 “Commuting Zones”
  - → Cover all counties (not just metropolitan statistical areas)
  - → We are in the Burlington, VT commuting zone
- Tax data from parents of 6.3 million children born in 1980-2
  - → Identify CS and parent’s income when child was age 16
  - → Compare to child’s income around age 30
- Main measure of upward mobility = the average income rank in the children’s income distribution for children growing up at the 20th percentile in the parent’s income distribution

An Example: Chetty et al (2014)

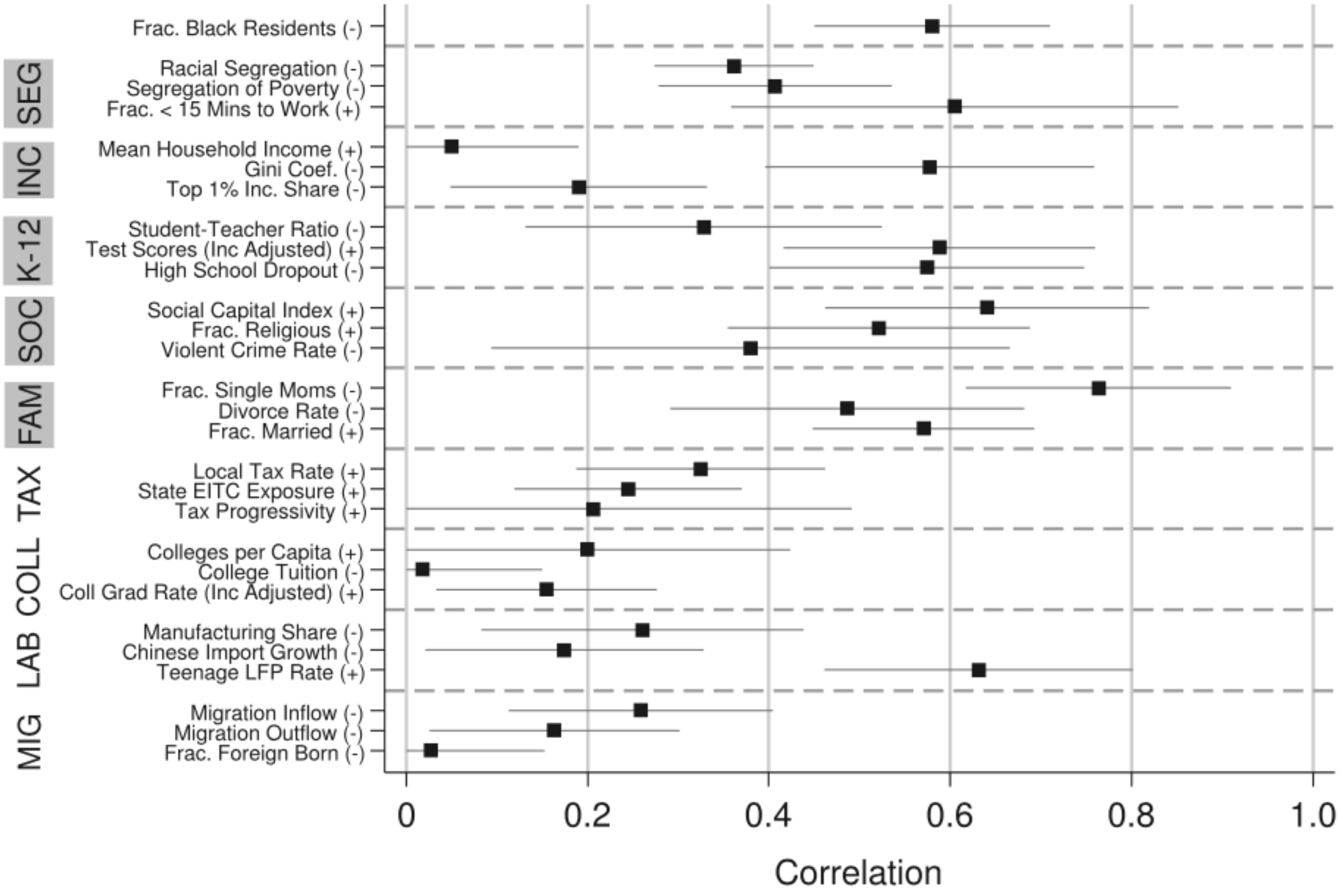


FIGURE VIII

Correlates of Spatial Variation in Upward Mobility