

# Lesson 13: Purposeful model selection

Nicky Wakim

2024-03-04

# Learning Objectives

1. Understand the overall steps for purposeful selection as a model building strategy
2. Apply purposeful selection to a dataset using R
3. Connect purposeful selection steps back to tests of coefficients in Class 8
4. Use different approaches to assess the linear scale of continuous variables in logistic regression

# Learning Objectives

1. Understand the overall steps for purposeful selection as a model building strategy
2. Apply purposeful selection to a dataset using R
3. Connect purposeful selection steps back to tests of coefficients in Class 8
4. Use different approaches to assess the linear scale of continuous variables in logistic regression

“Successful modeling of a complex data set is **part science**, **part statistical methods**, and **part experience and common sense**.”

Hosmer, Lemeshow, and Sturdivant Textbook, pg. 101

# Overall Process

0. Exploratory data analysis

1. Check unadjusted associations in simple linear regression

2. Enter all covariates in model that meet some threshold

- **One textbook** suggest  $p < 0.2$  or  $p < 0.25$ : great for modest sized datasets
- PLEASE keep in mind sample size in your study
- Can also use magnitude of association rather than, or along with, p-value

3. Remove those that no longer reach some threshold

- Compare magnitude of associations to unadjusted version (univariable)

4. Check scaling of continuous and coding of categorical covariates

5. Finalize main effect model

6. Check for interactions

7. Assess model fit

- Model assumptions, diagnostics, overall fit

# Process with snappier step names

**Pre-step:** Exploratory data analysis (EDA)

**Step 1:** Simple linear regressions

**Step 2:** Preliminary variable selection

**Step 3:** Assess change in coefficients

**Step 4:** Assess scale for continuous variables

**Step 5:** Finalize main effect model

**Step 6:** Check for interactions

**Step 7:** Assess model fit

# Learning Objectives

1. Understand the overall steps for purposeful selection as a model building strategy
2. Apply purposeful selection to a dataset using R
3. Connect purposeful selection steps back to tests of coefficients in Class 8
4. Use different approaches to assess the linear scale of continuous variables in logistic regression

# Pre-step: Exploratory data analysis

- Things we have been doing over the quarter in class and in our project
- I will not discuss some of the methods mentioned in our lab and data management class
  - I am only going to introduce additional exploratory functions


A few things we can do:

- Check the data
- Study your variables
- Missing data?
- Explore simple relationships and assumptions



# Pre-step: Exploratory data analysis: Check the data

- Get to know the potential values for the data
  - Categories
  - Units
- Then make sure the summary of values makes sense
  - If minimum or maximum look outside appropriate range
  - For example: a negative value for a measurement that is inherently positive (like population or income)



[Donate](#) [Resources](#) [About](#) [Log in](#)

[Home](#) > [Download the data](#) > [Documentation](#)

## Documentation

*Gapminder combines data from multiple sources into unique coherent time-series that can't be found elsewhere.*

Most of our data are not good enough for detailed numeric analysis. They are only good enough to revolutionize people's worldview. But we only fill in gaps whenever we believe we know roughly what the numbers would have been, had they existed. The uncertainties are often large. But we comfort ourselves by knowing the errors in peoples worldview are even larger. Our data is constantly improved by feedback in our [data forum](#) from users finding mistakes.

**We fill in all gaps:** Our data is more consistent over time and space than most other sources, because we dare to fill all the gaps in the sources. We dare this because our purpose is to show people the big picture, and they won't understand it if its full of holes.

**We use current geographic boundaries:** We show the world history as if country borders had always been the same as today. Read more [here](#).

Below are links to documentation describing how we have combined the sources in each case. For the sake of transparency, whenever allowed to share the underlying data, we make our complete calculations available for download, often in Excel files. In most of these files the details are not documented, as we haven't had time to describe every little step in our data process. But our data is constantly being improved by people who help find problems. If you have questions, we will try to answer them in our [data-forum](#).

Each documentation page has a version number and links to the previous versions. Whenever we update the data, or make other significant changes in the documentation, we make a new version.

**Data combined by Gapminder**  
[Average age at 1st marriage \(girls\)](#)  
[Babies per woman \(total fertility rate\)](#)  
[Child Mortality Rate, under age five](#)  
[GDP per capita in constant PPP dollars](#)  
[Gini](#)  
[HIV/AIDS](#)  
[Income Mountains](#)  
[Infant Mortality Rate, under age one](#)  
[Legal slavery](#)  
[Life Expectancy at Birth](#)  
[Maternal mortality](#)  
[Population](#)  
[World Health Chart, data sources](#)

*This list only includes data that we have somehow modified or calculated ourselves. The complete list of data we use is [here](#) »*

https://www.gapminder.org/data/documentation/

# Pre-step: Exploratory data analysis: Check the data

- Look at a summary for the raw data
- Typical use:

```
1 library(skimr)
2 skim(gapm)
```

- Some `skim()` help

# Pre-step: Exploratory data analysis: Check the data

- Look at a summary for the raw data
- Typical use:

```
1 library(skimr)
2 skim(gapm)
```

- Some `skim()` help
- Note that `skim(gapm)` looks different because I had to create factors
- I am breaking down the `skim()` function into the categorical and continuous variables only because I want to show them on the slides

```
1 skim(gapm_sub1) %>% yank("factor")
```











Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
four_regions	0	1.00	FALSE	4	Asi: 57, Afr: 54, Eur: 49, Ame: 35
income_levels1	1	0.99	FALSE	4	Hig: 56, Upp: 55, Low: 52, Low: 31
income_levels2	1	0.99	FALSE	2	Hig: 111, Low: 83

# Pre-step: Exploratory data analysis: Check the data

```
1 skim(gapm_sub1) %>% yank("numeric")
```

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
CO2emissions	4	0.98	4.55	6.10	0.03	0.64	2.41	6.22	41.20	
ElectricityUsePP	58	0.70	4220.92	5964.07	31.10	699.00	2410.00	5600.00	52400.00	
FoodSupplykcPPD	27	0.86	2825.06	443.59	1910.00	2490.00	2775.00	3172.50	3740.00	
IncomePP	2	0.99	16704.45	19098.61	614.00	3370.00	10100.00	22700.00	129000.00	
LifeExpectancyYrs	8	0.96	70.66	8.44	47.50	64.30	72.70	76.90	82.90	
FemaleLiteracyRate	115	0.41	81.65	21.95	13.00	70.97	91.60	98.03	99.80	
WaterSourcePrct	1	0.99	84.84	18.64	18.30	74.90	93.50	99.07	100.00	
Latitude	0	1.00	19.11	23.93	-42.00	4.00	17.33	40.00	65.00	
Longitude	0	1.00	21.98	66.52	-175.00	-5.75	21.00	49.27	179.14	
population_mill	0	1.00	35.95	136.87	0.00	1.73	7.57	24.50	1370.00	

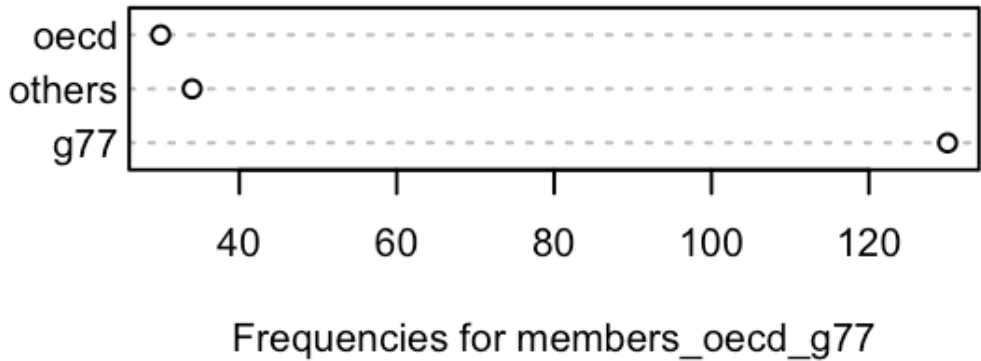
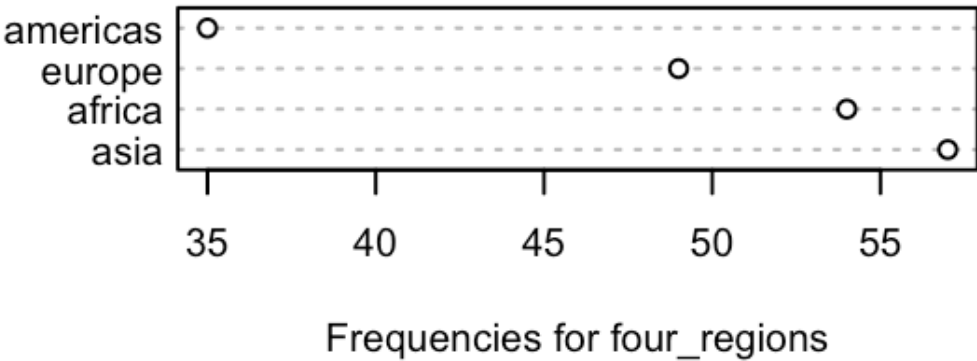
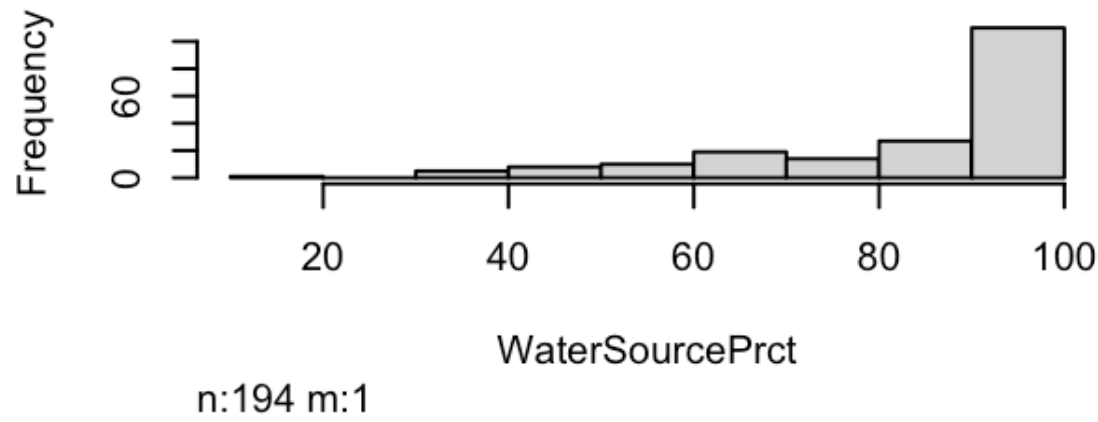
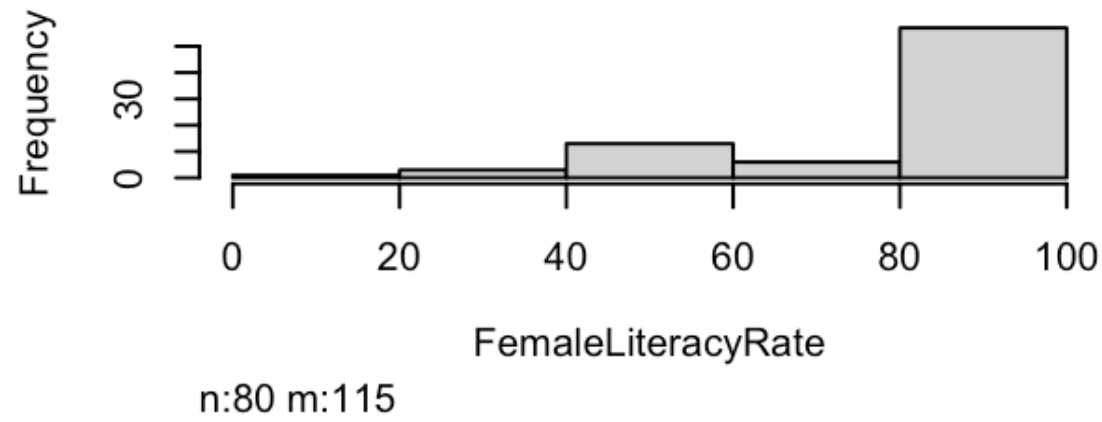
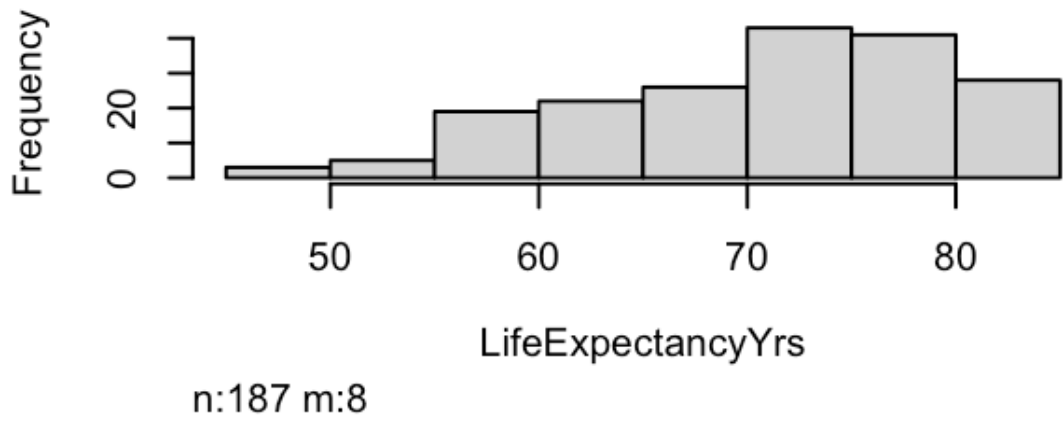
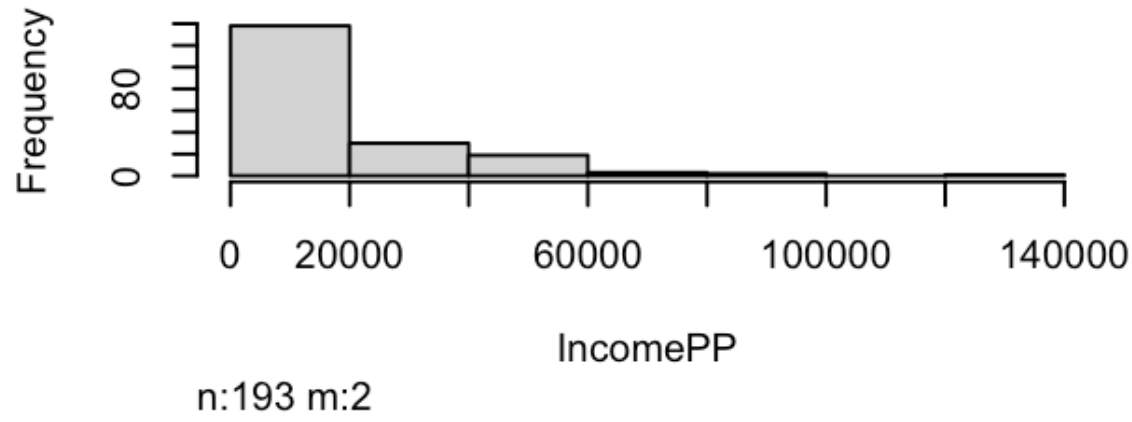
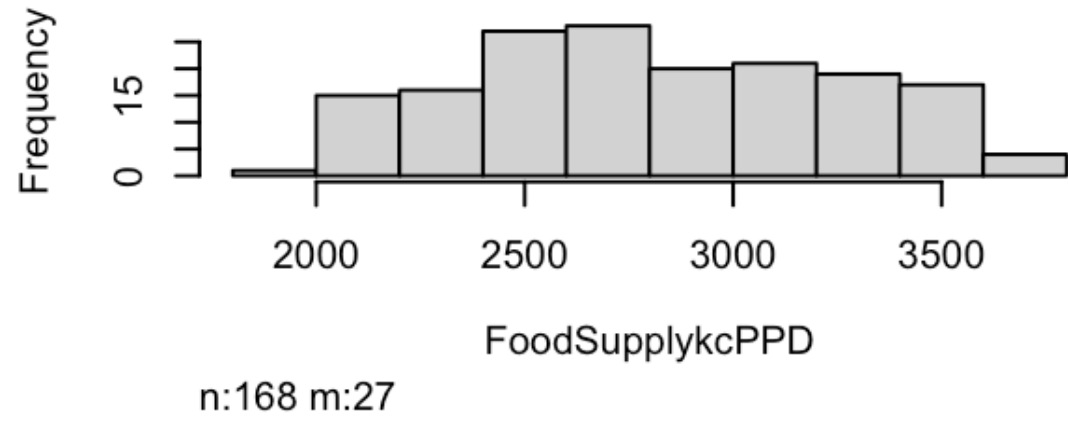
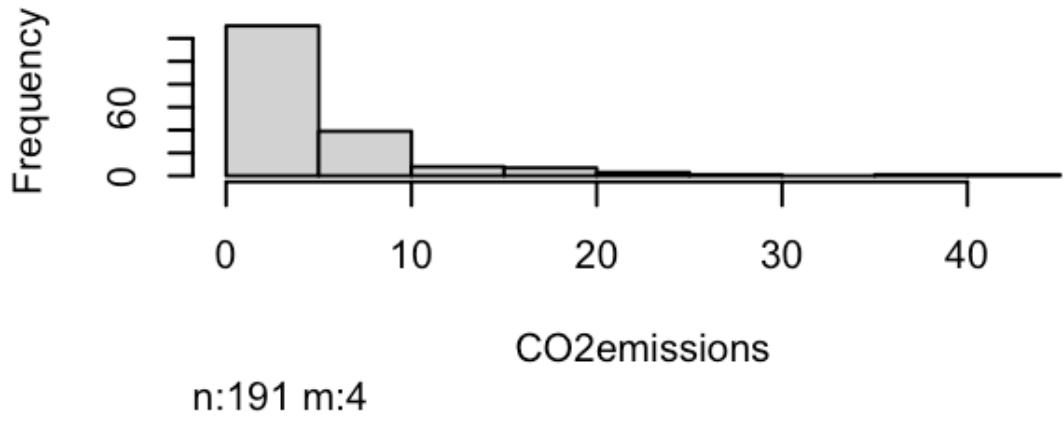
# Poll Everywhere Question 1

# Pre-step: Exploratory data analysis: Study your variables

- Started this a little bit in previous slide (`skim()`), but you may want to look at things like:
  - Sample size
  - Counts of missing data
  - Means and standard deviations
  - IQRs
  - Medians
  - Minimums and maximums
- Can also look at visuals
  - Continuous variables: histograms (in ``skimr()`` a little)
  - Categorical variables: frequency plots

# Pre-step: Exploratory data analysis: Study your variables

```
1 library(Hmisc)
2 hist.data.frame(gapm %>% select(-Longitude, -Latitude, -eight_regions, -six_regions, -geo, -`World bank, 4 income groups`
```



# Poll Everywhere Question 2

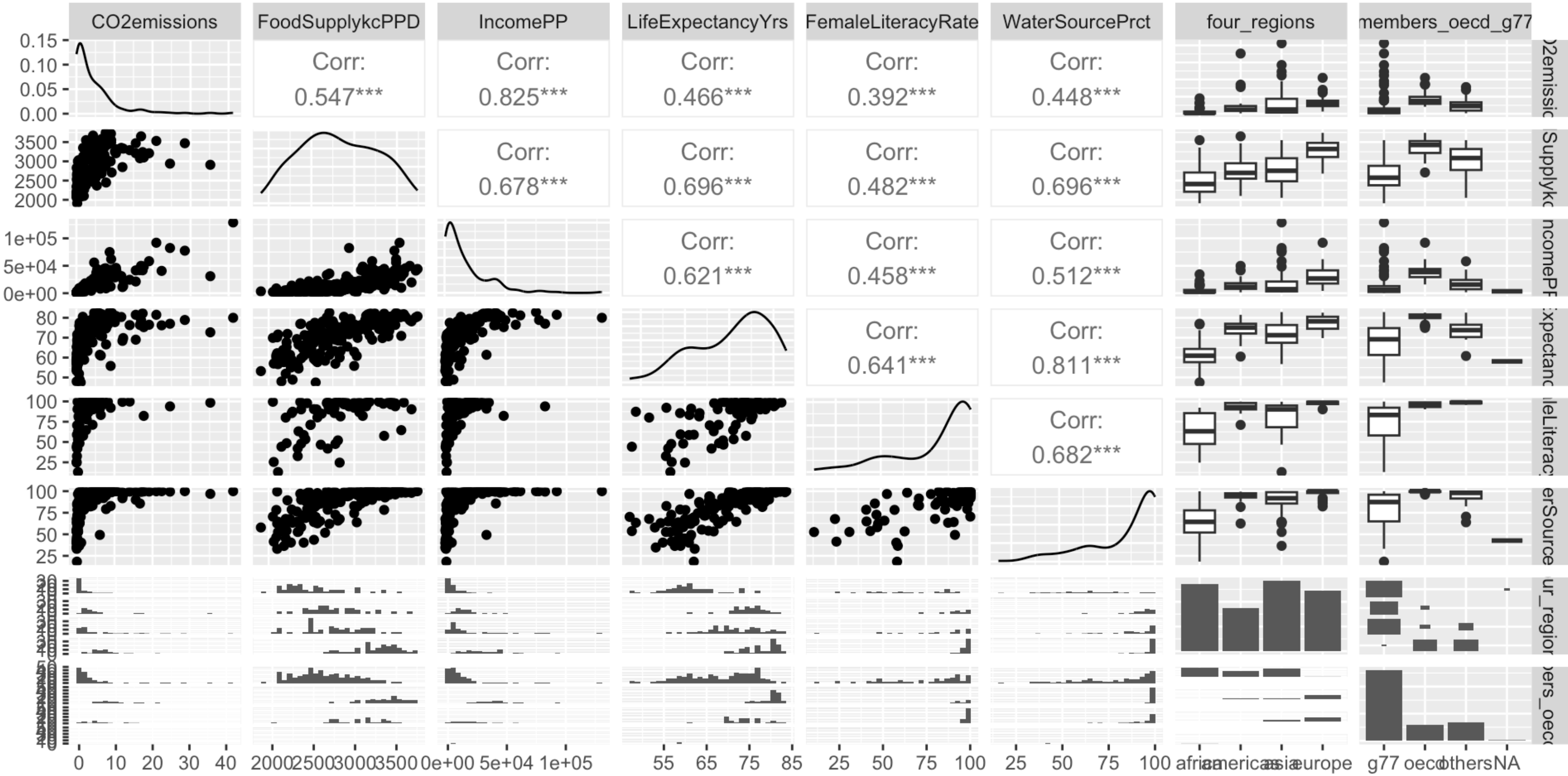


# Pre-step: Exploratory data analysis: Missing data

- Why are there missing data?
  - Which variables and observations should be excluded because of missing data?
  - Will I impute missing data?
- 
- Unfortunately, we don't have time to discuss missing data more thoroughly
  - I will try to cover this topic more thoroughly in BSTA 513
- 
- For the Gapminder dataset, we chose to use complete cases

# Pre-step / Step 1 : Explore simple relationships and assumptions

```
1 gapm2 %>% ggpairs() # gapm2 is a new dataset with some variables selected
```



# Poll Everywhere Question 3

**End of 3/4 class**

# Step 1: Simple linear regressions

## Step 2: Preliminary variable selection

## Step 3: Assess change in coefficients

## Step 4: Assess scale for continuous variables



## Step 5: Finalize main effect model

## Step 6: Check for interactions

# Step 7: Assess model fit

