

# AN INTRODUCTION TO R FOR POLICY ANALYSIS



# COURSE SCHEDULE:

**Webinar each Monday @ 10am (AEST) ~ 20 July to 17 August**

- ~~• Week 1 ~ An introduction to programming, the R language and Rstudio~~
- ~~• Week 2 ~ Doing stuff with data - Importing, Exploring and Summarizing Data~~
- Week 3 ~ Making pixels pretty (1) - Data Cleaning, Merging and Basic Visualization
- Week 4 ~ Making pixels pretty (2) - Data Cleaning and Visualization
- **Week 5 ~ Bringing together the R programming pipeline**



# WORKSHOP 5 OUTLINE:

- R in the Wild: David Keyes ~ [rfortherestofus.com](http://rfortherestofus.com)
- Reviewing the basics
- Policy Scenario: Poverty and Education in Myanmar
- Workshop evaluation

# DAVID KEYES – R FOR THE REST OF US

- Anthropologist by trade (Phd from University of California)
- Not from a traditionally *quantitative* field
- Picked up R while working as a consultant (self-taught)
- Developed Rfortherestofus.com as a training platform for people who want to learn R, but might not have a statistical/quantitative background
- Has a free 'Getting Started with R' course available on his website @ rfortherestofus.com. For the more advanced paid options, participants in this course can use the coupon code GILES2020 for 20% off.
- He's also worth following on twitter @dgkeyes

See [David's video here](#).



# REFRESHER : THE BASICS

# COMMON ERRORS

**Unmatched parenthesis:** ie not including a close bracket to tell R the arguments passed to a function have been supplied and to apply the function – will tell you with ‘+’

**Misplaced commas:** eg not separating arguments by commas when applying a function

**Applying functions to an incompatible element or data object type** – eg trying to multiply text by 2

**"cannot open"** – attempts to read something that doesn't exist or isn't accessible (eg because the name or directory is wrong or R can't access it due to it being locked/open in another program).

**"could not find function"** – not loading the library needed for a function or misspelling the name of the function (eg Plot instead of plot).

**"Error in eval"** – cause by references to something that doesn't exist (eg misspelling an object name)

**"no applicable method"** – trying to apply a function to an object/element type that doesn't support it

**"subscript out of bounds"** – telling R to access data that doesn't exist/ is out of bounds (eg asking for row 34 in a 20 row vector)

**When data/elements are the wrong type they can be converted using:**

`as.data.frame()`, `as_tibble()`, `as_vector()`, `as_character()`, `as_numeric()`, etc



# DPLYR VERBS

Provides a grammar for basic data manipulation:

- **select** – pick the columns you want
- **filter** – choose the relevant observations
- **mutate** – create new variables
- **group\_by** – define groups you're interested in for summarise
- **summarise** – create summary statistics
- **arrange** – sort your data

enter '?summarise' to see help and a list of summary statistics

## TIDY DATA: **ACCESS VARIABLES VIA COLUMN**

Vehicle	Efficiency	Power	Year
Mazda RX4	21	110	2001
Mazda RX4 Wag	21	110	2015
Merc 230	22.8	95	2015
Merc 280	19.2	123	1983
Datsun 710	22.8	93	2017

- ✓ Each variable is a column.
- ✓ Each observation a row.
- ✓ Each experiment/survey a separate table

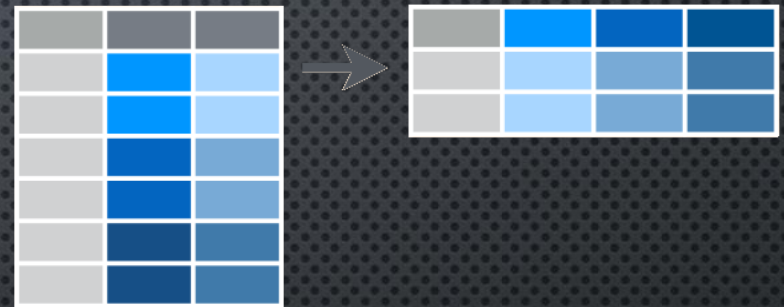


# 'TIDYR' CAN HELP YOU GET DATA IN THE RIGHT SHAPE

`pivot_longer()`



`pivot_wider()`



`separate()`



`unite()`



*(This can also be useful for making data useful for pivot tables in excel)*

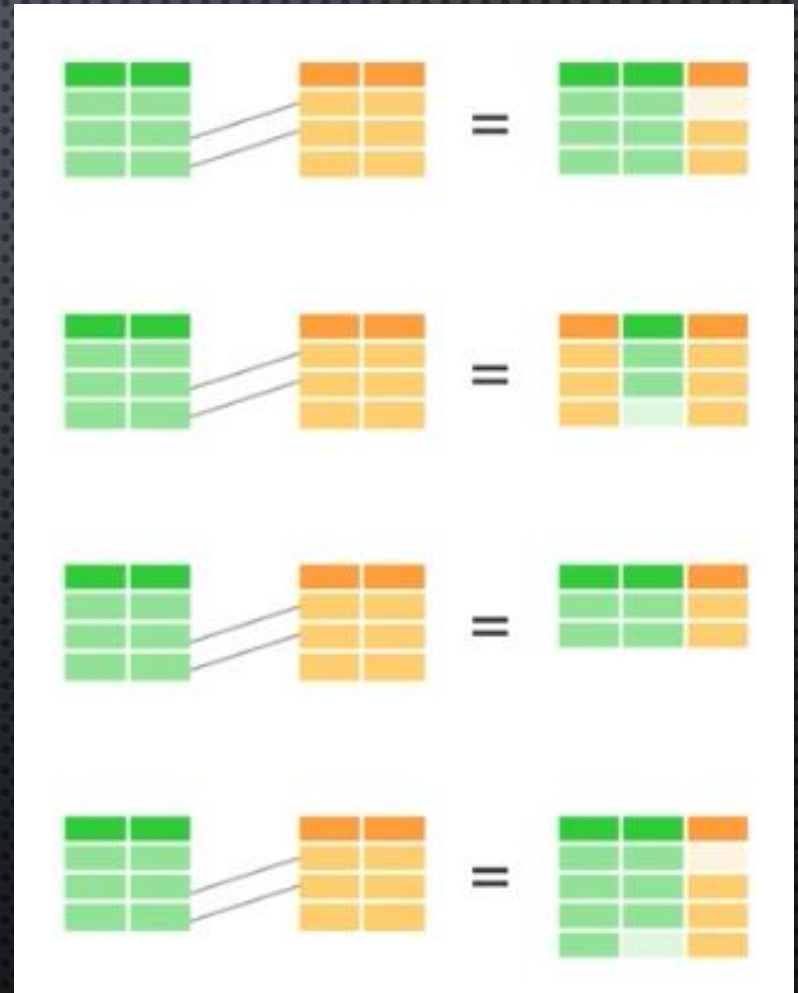
## DPLYR JOINS:

`left_join(x,y)`

`right_join(x,y)`

`inner_join(x,y)`

`full_join(x,y)`





# GGPLOT2 – LAYERS ON A CANVAS

## Define the canvas:

```
ggplot(data=titanic_data, aes(x=age))
```

## Specify a geometry:

```
ggplot(data=titanic_data, aes(x=age)) +  
  geom_histogram()
```

## Add layers:

```
ggplot(data=titanic_data, aes(x=age)) +  
  geom_histogram() +  
  labs(y="A Histogram")
```

# GGPLOT2 – KEY GRAMMAR

## **Geoms**

Defining the geometry of the plot

## **Aesthetics**

Defining the aesthetic mappings

## **Labels**

Defines labelling

## **Themes**

Changes how plots look

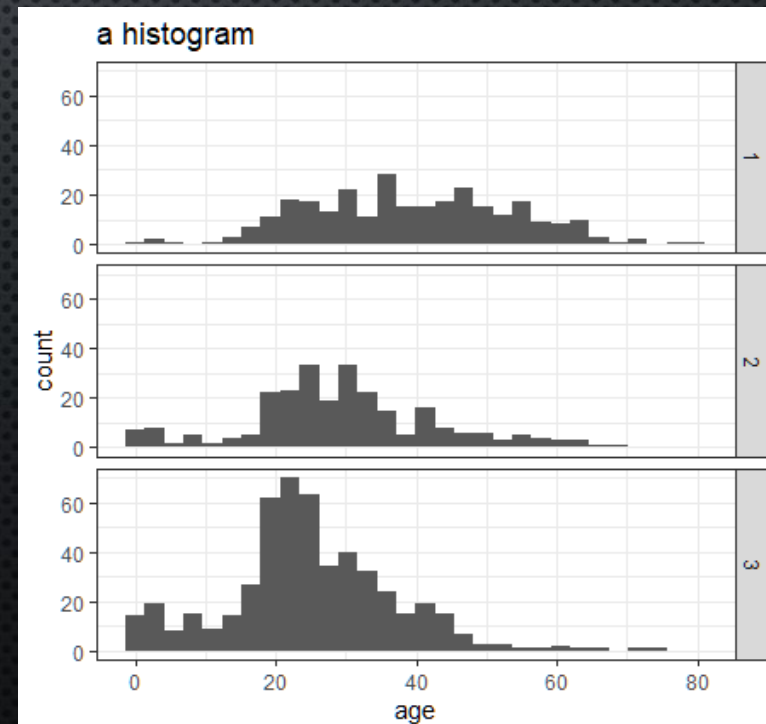
## **Facets**

Dividing plots based on groups



# GGPLOT2 – LAYERS ON A CANVAS

```
ggplot(data=titanic_data, aes(x=age)) +  
  geom_histogram() +  
  labs(title="a histogram") +  
  theme_bw() +  
  facets_grid(pclass ~ .)
```



# POLICY SCENARIO: MYANMAR EDUCATION POLICY



# MYANMAR – EDUCATION POLICY SCENARIO

Education and poverty were central issues in the last election. With voters being particularly concerned about the limited progress that has been made to increase access to education for some of the most disadvantaged communities. Reflecting these concerns, the Minister has asked for your advice on where investments might best be made to address the issue, asking:

1. Which State or Region has seen the strongest growth in the number of primary teachers and primary schools since 2009? How does this relate to wealth?
2. Using the 2015 Wealth Ranking Index and the maternal mortality ratio, which townships appear most worthwhile to target?
3. Is there a relationship between educational attainment, wealth, maternal mortality and urbanization?
4. The Minister has proposed implementing a policy to boost primary enrolment in target rural communities by paying a stipend of 26 USD for all females to enroll in education (between the ages of 5 and 29).
  - How much is this likely to cost per year?
  - Which ten townships are likely to benefit the most from this (in \$ terms)?
  - What if this was restricted only to rural and 'target' communities?



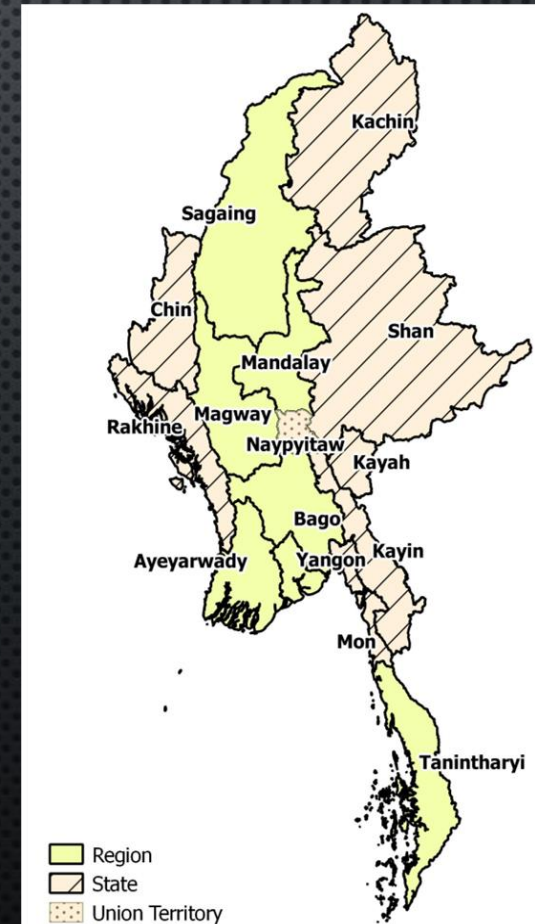
# HOW COULD WE ANSWER EACH QUESTION?

1. Which State or Region has seen the strongest growth in the number of primary teachers and primary schools since 2009? How does this relate to wealth?
2. Using the 2015 Wealth Ranking Index and the maternal mortality ratio, which townships appear most worthwhile to target?
3. Is there a relationship between educational attainment, wealth, maternal mortality and urbanization?
4. The Minister has proposed implementing a policy to boost primary enrolment in targeted rural communities by paying a stipend of 26 USD for all females to enroll in education (between the ages of 5 and 29).
  - How much is this likely to cost per year?
  - Which ten townships are likely to benefit the most from this (in \$ terms)?
  - What if this was restricted only to rural and 'target' communities?



# MYANMAR – EDUCATION POLICY SCENARIO

- Population of 50.3 million (2014)
- 7 States, 7 Regions + 1 Union Territory
- Each state divided into districts then townships
- 330 townships



# MYANMAR – EDUCATION POLICY SCENARIO

## Key Facts:

- Data collated by the Myanmar Information Management Unit (MIMU)
- File includes information on location, year and the source of the data
- Data available at the national (union), state and region and township level (in different sheets)
  - National data is stored in sheet 1 of the excel file, State and Region in sheet 2 and Township data in 3.
- Locations indicated by name and “Pcode” which is a unique ‘place code’ useful for geographic mapping (and merging)
- Data organized in a ‘wide’ format, with years as columns and variables/indicators in rows.



# MYANMAR – EDUCATION POLICY SCENARIO

What do we *at least* need to know to answer the request?

**Question 1 ~ State and Region level information on:**

Number of primary school teachers, number of primary schools and wealth.

**Question 2 ~ Township level information on:**

Wealth and maternal mortality

**Question 3 ~ State and Region level information on:**

Educational attainment, wealth, maternal mortality and urbanization

**Question 4 ~ Union and Township level information on:**

Number of women *between the ages of 5 and 29*, wealth and educational attainment



# MYANMAR – EDUCATION POLICY SCENARIO

- What do we need to do to import the data?
- Have variables been imported in the right format? For instance, are years recognized as numbers to allow correct ordering?
- Are there any duplicates? What about missing values?
- Is the data tidy? If not, how do we need to reshape it so we can easily access the data we need?
- What do we need to do to make it tidy and make the variables we need easily accessible?
- How can we logically check our data measures what we think it does?
- How do we know the variable name is unique enough to ensure we include the right data from the right source?
- How are we going to match up the data needed for questions requiring information from different sources?



# MYANMAR – EDUCATION POLICY SCENARIO

## Key Steps:

- Loading the packages we need (tidyverse + readxl)
- Importing the data from the right excel sheet
- Data cleaning
  - Reshaping data, removing duplicates, dropping missing values, ensuring data is usefully stored, creating useful variable names to allow data we need to be accessed by columns.
  - Picking the right data we need to answer the request
- Data analysis and visualization
  - Summarizing township data to allow us to use it in State/Region level analysis
  - Visualizing the data in a way that responds to the request

# WHAT FUNCTIONS WILL BE USEFUL?

1. *base?*
2. *readxl: ?*
3. *tidyr: ?*
4. *dplyr?*
5. *ggplot2?*
6. *stringr?*

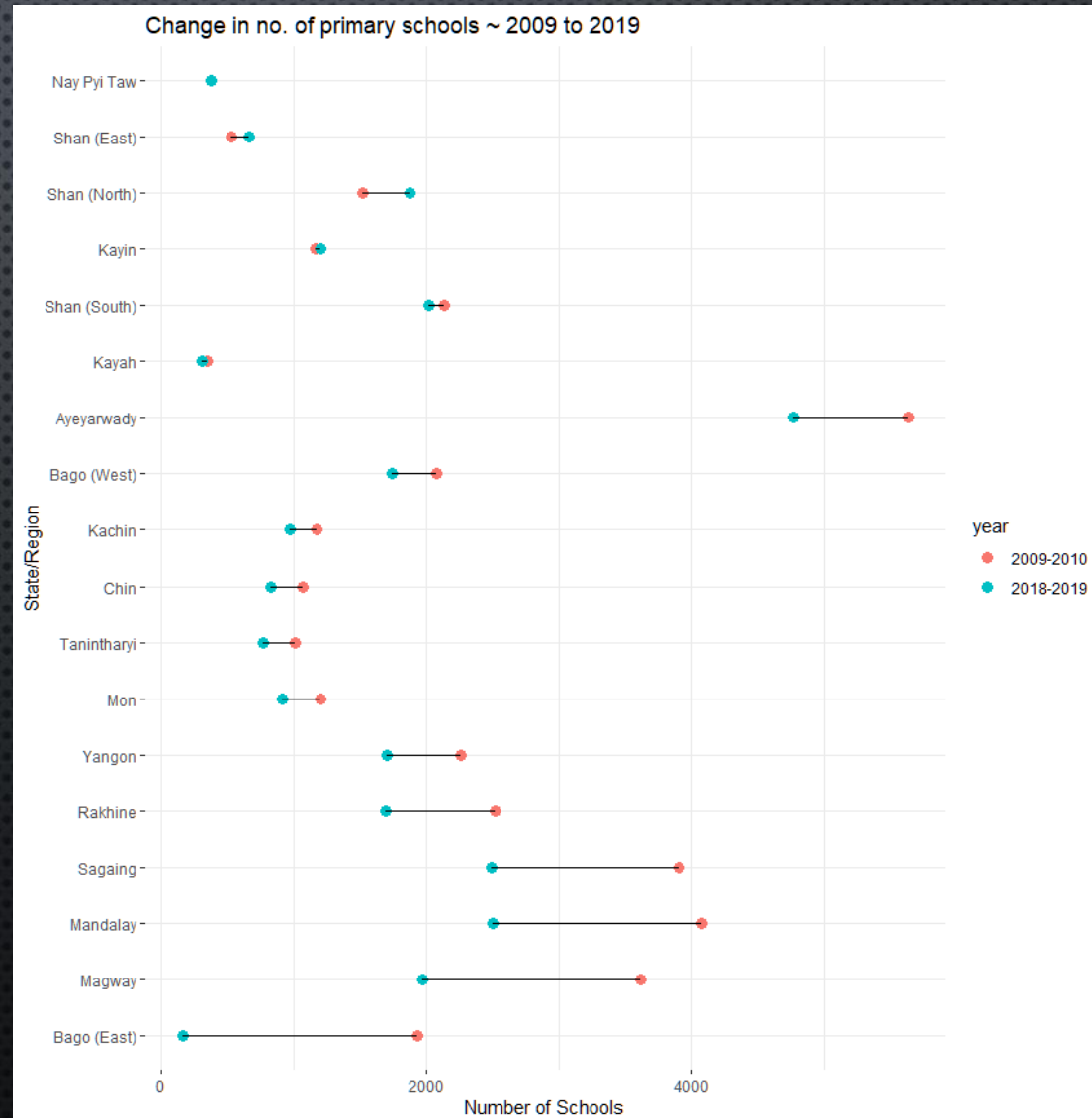


# MYANMAR – EDUCATION POLICY SCENARIO

Education and poverty were central issues in the last election. With voters being particularly concerned about the limited progress that has been made to increase access to education for some of the most disadvantaged communities. Reflecting these concerns, the Minister has asked for your advice on where investments might best be made to address the issue, asking:

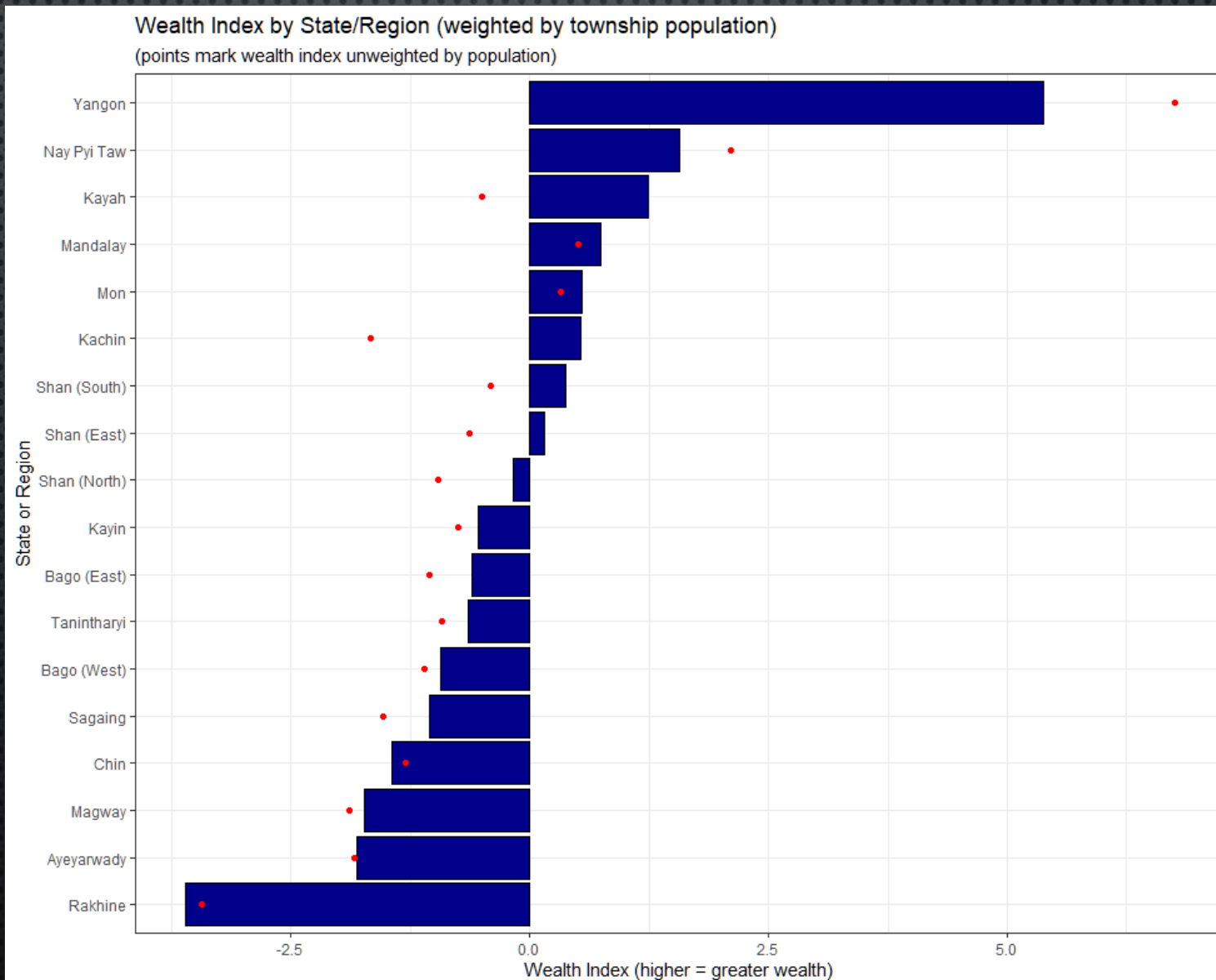
1. Which State or Region has seen the strongest growth in the number of primary teachers and primary schools since 2009? How does this relate to wealth?
2. Using the 2015 Wealth Ranking Index and the maternal mortality ratio, which townships appear most worthwhile to target?
3. Is there a relationship between educational attainment, wealth, maternal mortality and urbanization?
4. The Minister has proposed implementing a policy to boost primary enrolment in target rural communities by paying a stipend of 26 USD for all females to enroll in education (between the ages of 5 and 29).
  - How much is this likely to cost per year?
  - Which ten townships are likely to benefit the most from this (in \$ terms)?
  - What if this was restricted only to rural and 'target' communities?

# QUESTION 1





# QUESTION 1

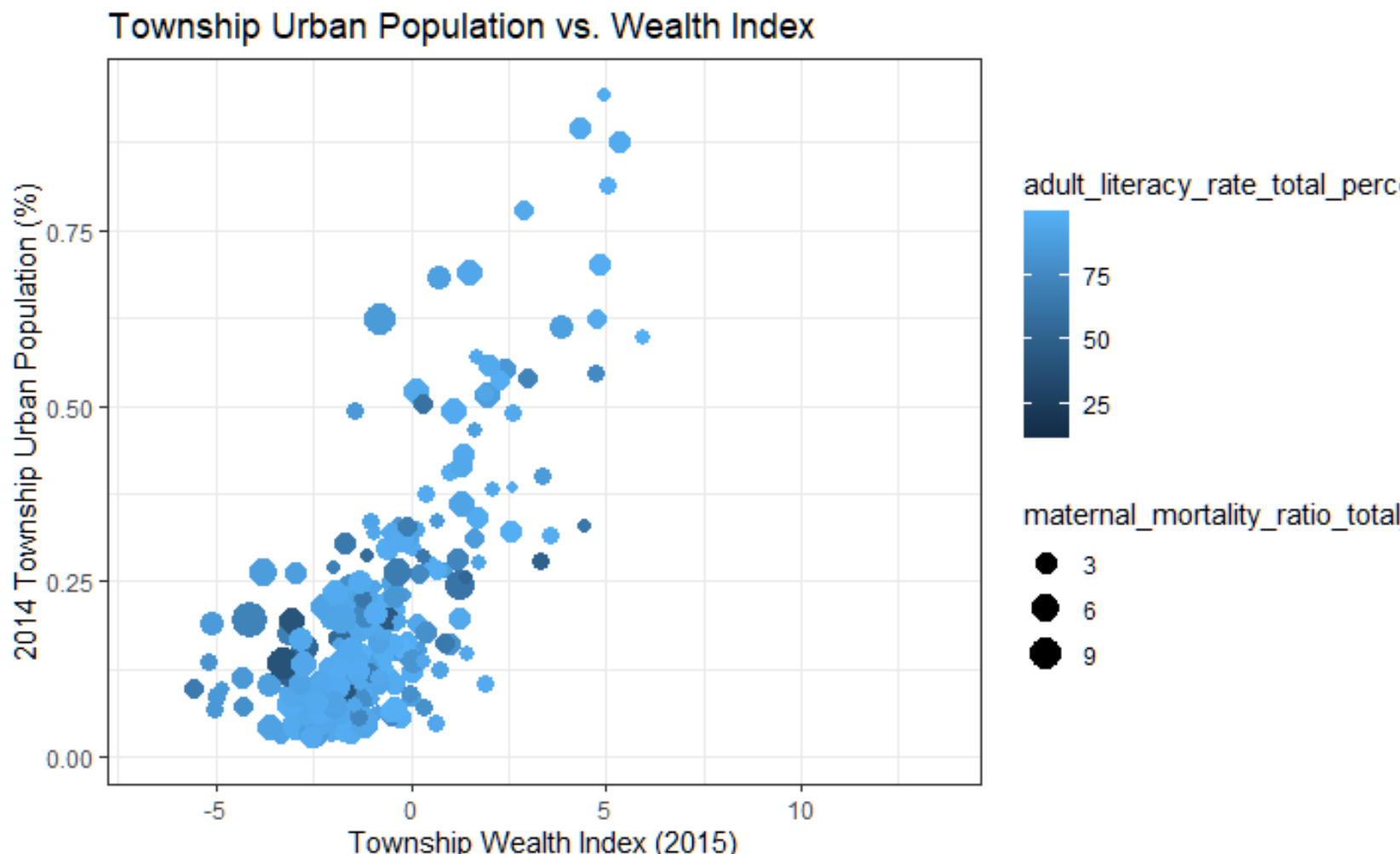


## QUESTION 2

- Develop 'thresholds' to identify target townships and list townships that meet the criteria (eg township 1, 2, 3 etc)
- Can also use mutate() to create a column that tells us if conditions are met

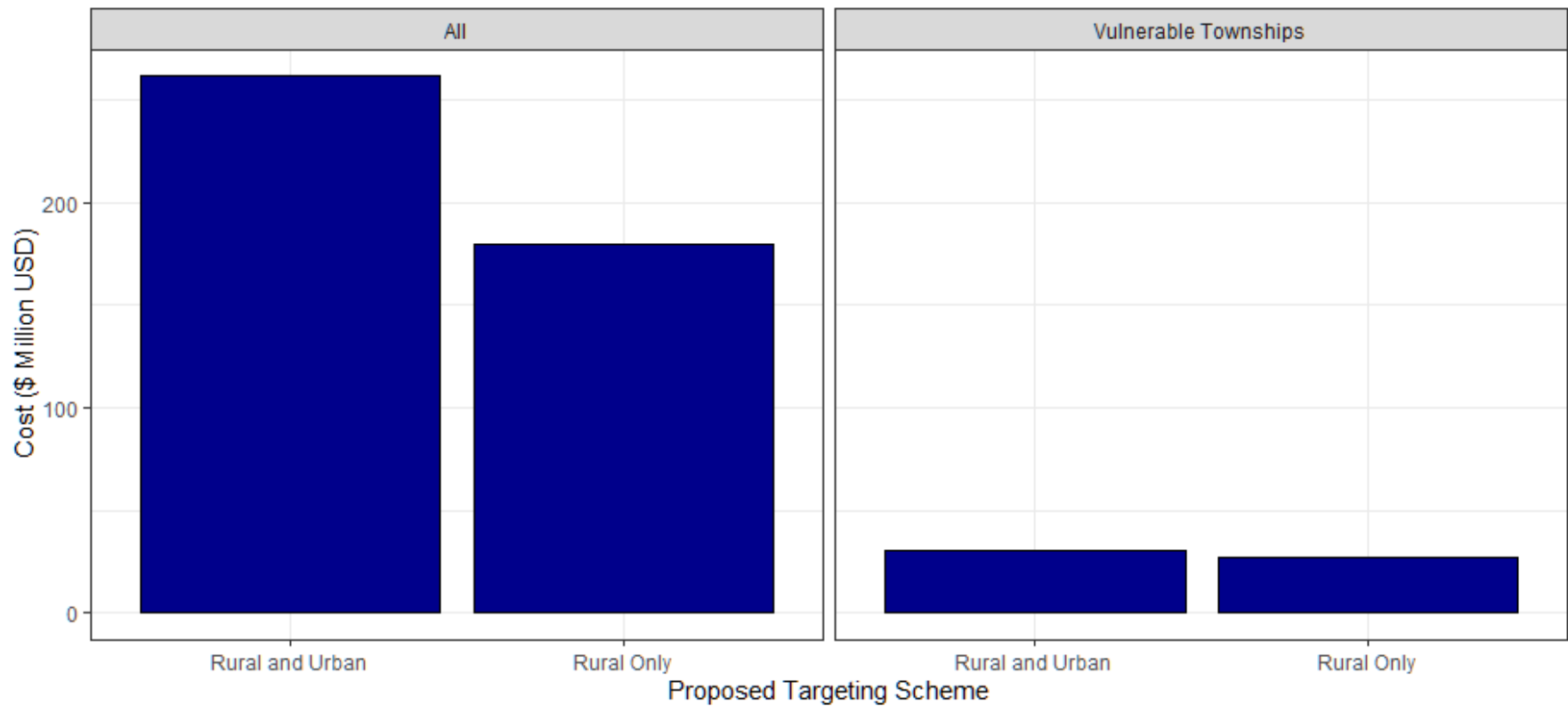


# QUESTION 3



# QUESTION 4

Estimated Cost of Education Policy Proposals





# AN INTRODUCTION TO R FOR POLICY ANALYSIS

Congratulations for surviving!



# CONTINUE LEARNING

Learning R takes time + effort:

- Focus on learning what interests you!
- Commit to using R for a project where you can use a tool you're already comfortable with to check your results (eg excel)
- Get comfortable googling to find solutions
- Enroll in an online course and/or specialization
- Continue with the swirl courses
- Get involved in the community via twitter, reddit, r-bloggers etc



# CONTINUE LEARNING

- More Swirl! - [swirlstats.com](http://swirlstats.com) ~ See the Course Repository:
  - Regression Models
  - Getting and Cleaning Data
  - Statistical Inference
- Online Courses:
  - John Hopkins University - <https://www.coursera.org/specializations/jhu-data-science#courses>
  - Harvard - <https://online-learning.harvard.edu/subject/r>
  - Microsoft - <https://www.edx.org/course/introduction-to-r-for-data-science-2>
- Subscribe to R-bloggers [www.r-bloggers.com](http://www.r-bloggers.com)
- R For the Rest of Us - [rfortherestofus.com](http://rfortherestofus.com)
  - Free introductory course + 20% paid courses for participants
- Youtube:
  - Global Health with Greg Martin ~ Great R 101 videos ([link](#))

# WEBINAR SERIES EVALUATION

<https://forms.gle/ZGkHaPzdS3NZ1m9c8>



# AN INTRODUCTION TO R FOR POLICY ANALYSIS

