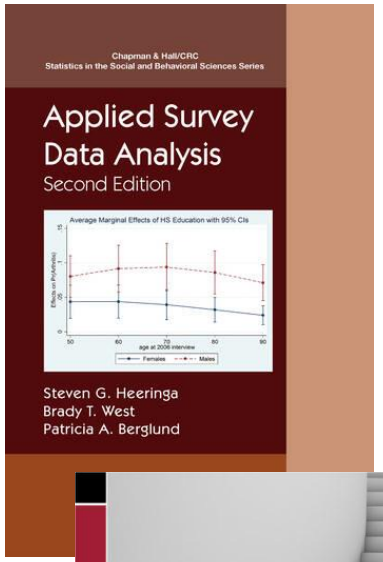
Visualizing Survey Data Analysis Results:
Marrying the Best from Stata and R

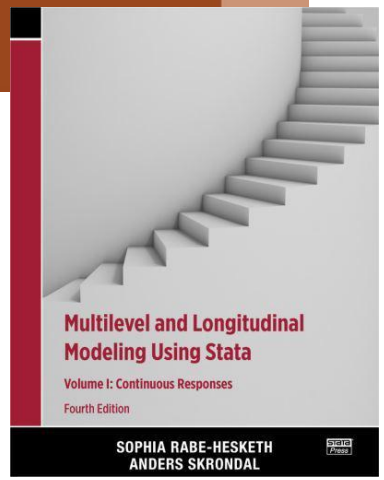
2022 Stata Conference
4 – 5 August 2022 Washington DC

Nel Jason (Jason) L. Haw, MS
PhD Student, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
nhaw1@jh.edu @jasonhaw_ linkedin.com/in/neljasonhaw

Easy-to-use survey data analysis tools with excellent resources

stata.com/manuals/svy.pdf



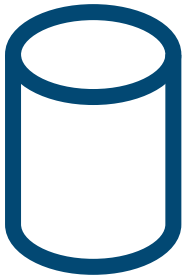


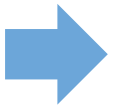
Highly flexible execution of the Grammar of Graphics

Image from Kirsten Packer's ggplot2 presentation last February 2019:
http://www.seec.uct.ac.za/ggplot2-grammar-graphics

Easy-to-use survey data analysis tools with excellent resources

Highly flexible execution of the Grammar of Graphics





Processed survey data

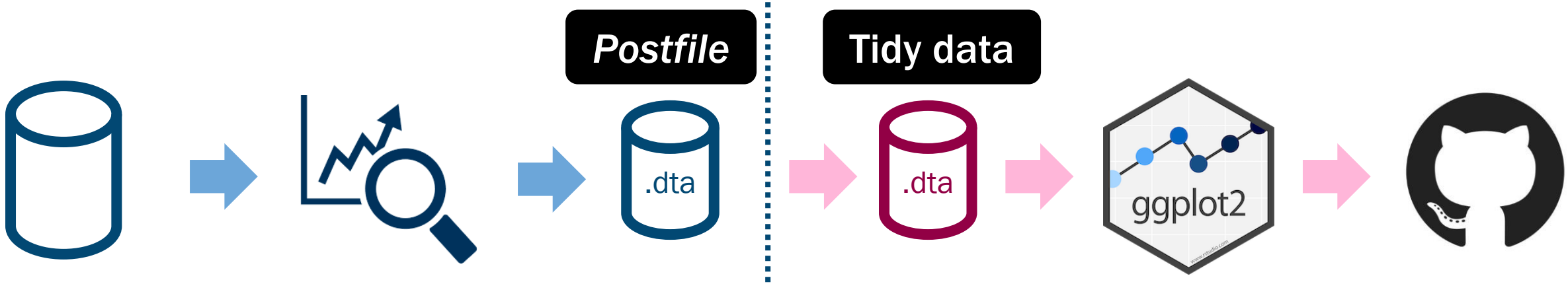
Survey data analysis

Ensure reproducibility during handoff

Create plots using *ggplot2*

Easy-to-use survey data analysis tools with excellent resources

Highly flexible execution of the Grammar of Graphics



Processed survey data

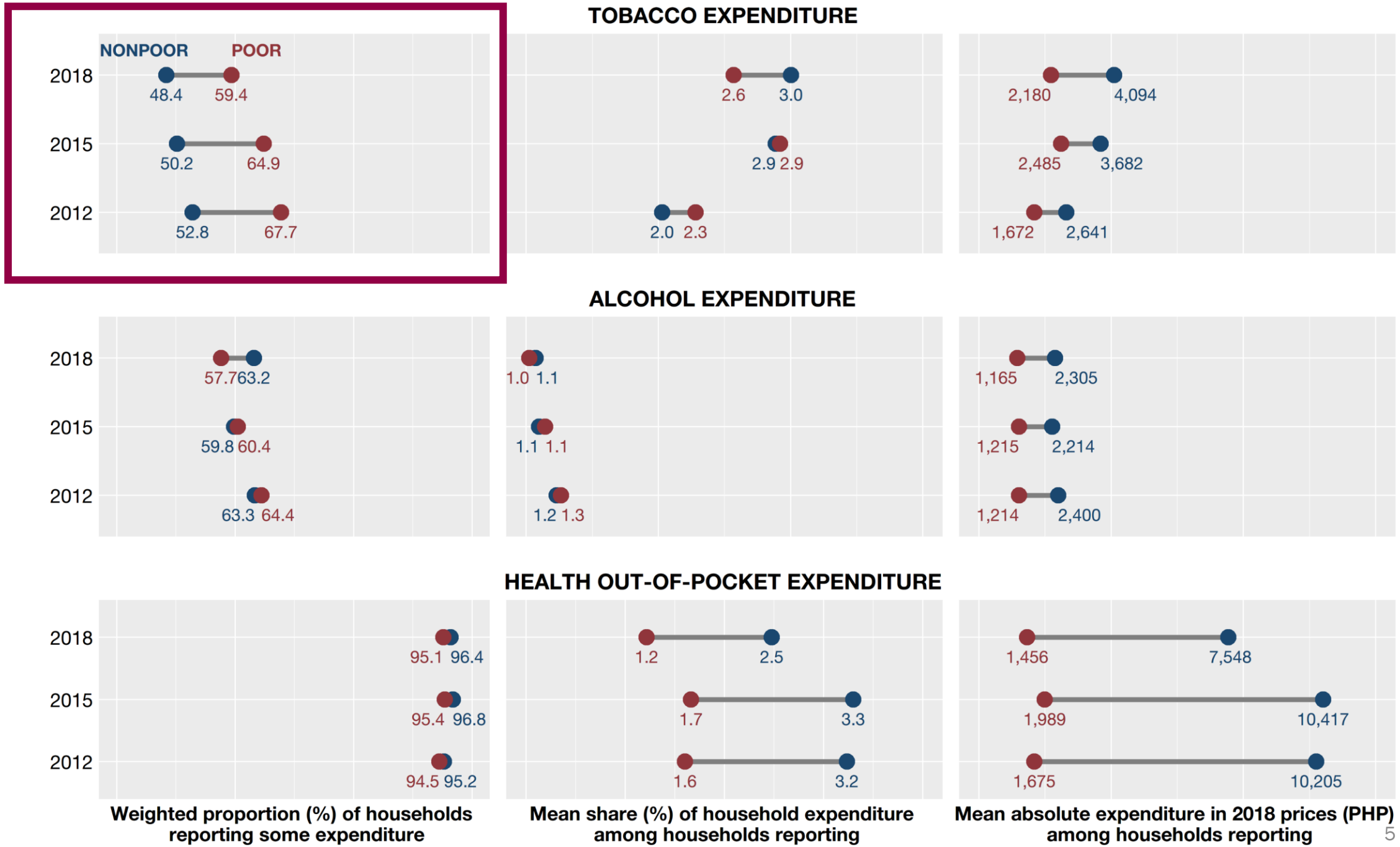
Survey data analysis

Analysis results stored using *postfile*

Read .dta using *haven*

Create plots using *ggplot2*

Upload repository to Github


TOBACCO EXPENDITURE
NONPOOR
POOR
2018
2015
2012
48.4
59.4
50.2
64.9
52.8
67.7
2.6
3.0
2.9
2.9
2.0
2.3
2,180
4,094
2,485
3,682
1,672
2,641
ALCOHOL EXPENDITURE
57.7
63.2
59.8
60.4
63.3
64.4
1.0
1.1
1.1
1.1
1.2
1.3
1,165
2,305
1,215
2,214
1,214
2,400
HEALTH OUT-OF-POCKET EXPENDITURE
95.1
96.4
95.4
96.8
94.5
95.2
1.2
2.5
1.7
3.3
1.6
3.2
1,456
7,548
1,989
10,417
1,675
10,205
Weighted proportion (%) of households reporting some expenditure
Mean share (%) of household expenditure among households reporting
Mean absolute expenditure in 2018 prices (PHP) among households reporting

Easy-to-use survey data analysis tools with excellent resources

Processed survey data

EXAMPLE
Philippines expenditure data from 229,432 households

Survey data analysis

EXAMPLE
Proportion reporting tobacco expenditure between poor and non-poor households

Analysis results stored using *postfile*

Highly flexible execution of the Grammar of Graphics

Read .dta using *haven*

Create plots using *ggplot2*

EXAMPLE
Dumbbell plots

Upload repository to Github

Data description	Stata variable name
• Family Income and Expenditure Survey 2012, 2015, 2018	`survey`
• Two-stage sampling design:	
◦ Primary sampling units (PSU): villages (*barangays*)	`psu`
◦ Secondary sampling units: households within villages	
• Stratified by subnational boundaries (region/province by urban/rural status)	`stratum`
• Post-stratification survey weights	`weight`
• Outcome: reported tobacco expenditure during the year (yes/no)	`prev_tobacco`
• Exposure: Poverty status based on the provincial poverty line (poor/non-poor)	`poverty`

1. Declare survey design using *svyset*

```
svyset psu [pweight = weight], strata(stratum) singleunit(centered)
```

clustering variable **psu** (primary sampling unit)

weight variable **weight**

strata variable **stratum**

single unit stratum are summarized using the overall mean

Stata output

```
Sampling weights: weight
             VCE: linearized
     Single unit: centered
        Strata 1: stratum
 Sampling unit 1: psu
             FPC 1: <zero>
```

2. Calculate proportions of tobacco expenditure over survey round and poverty status

```
svyset psu [pweight = weight], strata(stratum) singleunit(centered)
```

```
svy: mean prev_tobacco, over(survey poverty)
```

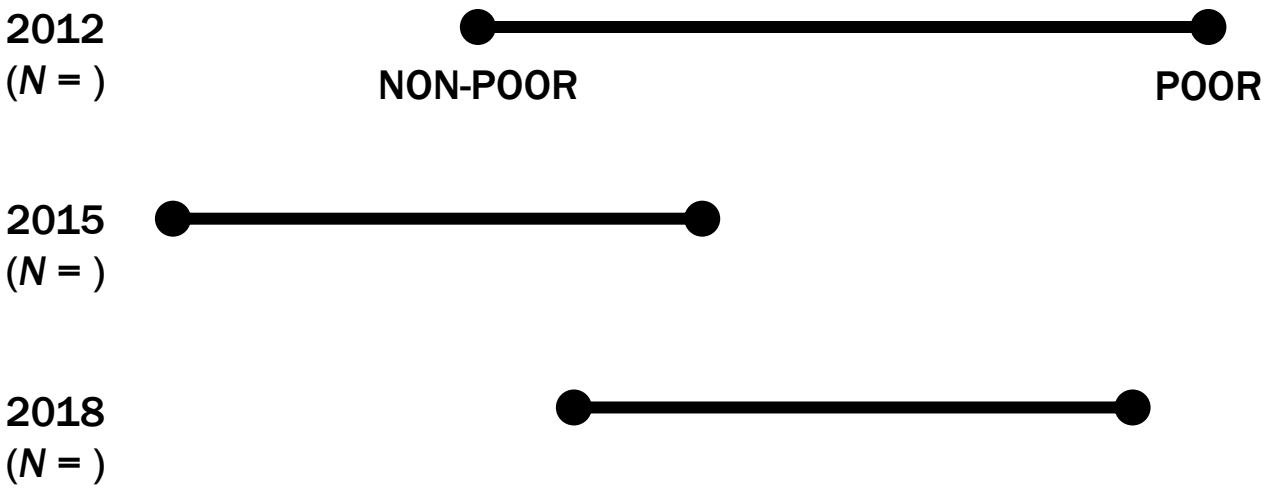
Stata output

```
------------------------------------------------------------------------------
                             |             Linearized
                             |      Mean   std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
c.prev_tobacco@survey#poverty|
                      2012#No|   .5281905   .0039855      .5203784    .5360025
                     2012#Yes|   .6774919   .0064277      .6648929    .6900909
                      2015#No|   .5018166   .0037622      .4944423     .509191
                     2015#Yes|   .6485802   .0064525      .6359325    .6612278
                      2018#No|   .4838312   .0025353      .4788616    .4888007
                     2018#Yes|   .5937887   .0050016      .5839849    .6035925
------------------------------------------------------------------------------
```

1. Identify the data structure needed for the plot

Dumbbell plot

visualizes the difference between two groups



Survey	Non-poor	Poor	N
2012			
2015			
2018			

2. Declare the *postfile* dataset

```
postfile dumbbell int survey float (nonpoor poor N) using "dumbbell.dta", replace
```

Name assigned to Stata in internal memory

integer format to new variable **survey** (contains the survey year)

float format to new variables **nonpoor** (value among non-poor), **poor** (value among poor), and **N** (sample size)

filename of resulting file stored on disk

overwrites file every run

3. Declare survey design using *svyset*

```
postfile dumbbell int survey float (nonpoor poor N) using "dumbbell.dta", replace
svyset psu [pweight = weight], strata(stratum) singleunit(centered)
```

clustering variable **psu** (primary sampling unit)

weight variable **weight**

strata variable **stratum**

single unit stratum are summarized using the overall mean

Stata output

```
Sampling weights: weight
             VCE: linearized
     Single unit: centered
        Strata 1: stratum
 Sampling unit 1: psu
             FPC 1: <zero>
```

4. Calculate proportions of tobacco expenditure over survey round and poverty status

```
postfile dumbbell int survey float (nonpoor poor N) using "dumbbell.dta", replace
svyset psu [pweight = weight], strata(stratum) singleunit(centered)
svy: mean prev_tobacco, over(survey poverty)
```

Stata output

```
------------------------------------------------------------------------------
                             |             Linearized
                             |       Mean   std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
c.prev_tobacco@survey#poverty|
                      2012#No|   .5281905   .0039855      .5203784    .5360025
                     2012#Yes|   .6774919   .0064277      .6648929    .6900909
                      2015#No|   .5018166   .0037622      .4944423     .509191
                     2015#Yes|   .6485802   .0064525      .6359325    .6612278
                      2018#No|   .4838312   .0025353      .4788616    .4888007
                     2018#Yes|   .5937887   .0050016      .5839849    .6035925
------------------------------------------------------------------------------
```

`ereturn list` Stata output (selected)

```
matrices:
                  e(b) :  1 x 6
                  e(V) :  6 x 6
              e(_N_subp) :  1 x 6
             e(V_srssub) :  6 x 6
                e(V_srs) :  6 x 6
                   e(_N) :  1 x 6
```

Contains the means

Stata output

```
------------------------------------------------------------------------------
                             |             Linearized
                             |      Mean   std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
c.prev_tobacco@survey#poverty|
                      2012#No|   .5281905   .0039855      .5203784    .5360025
                     2012#Yes|   .6774919   .0064277      .6648929    .6900909
                      2015#No|   .5018166   .0037622      .4944423     .509191
                     2015#Yes|   .6485802   .0064525      .6359325    .6612278
                      2018#No|   .4838312   .0025353      .4788616    .4888007
                     2018#Yes|   .5937887   .0050016      .5839849    .6035925
------------------------------------------------------------------------------
```

matrix list e(b) Stata output

```
 c.prev_tob~o@  c.prev_tob~o@  c.prev_tob~o@  c.prev_tob~o@  c.prev_tob~o@  c.prev_tob~o@
         1.survey#      1.survey#      2.survey#      2.survey#      3.survey#      3.survey#
       0.poverty      1.poverty      0.poverty      1.poverty      0.poverty      1.poverty
y1     .52819046      .67749186      .50181664      .64858015      .48383119       .5937887

         e(b)[1,1]      e(b)[1,2]      e(b)[1,3]      e(b)[1,4]      e(b)[1,5]      e(b)[1,6]
```

Stata output

```
------------------------------------------------------------------------------
                             |             Linearized
                             |      Mean   std. err.      [95% conf. interval]
-----------------------------+------------------------------------------------
c.prev_tobacco@survey#poverty|
                   2012#No   |   .5281905   .0039855       .5203784    .5360025
                  2012#Yes   |   .6774919   .0064277       .6648929    .6900909
                   2015#No   |   .5018166   .0037622       .4944423     .509191
                  2015#Yes   |   .6485802   .0064525       .6359325    .6612278
                   2018#No   |   .4838312   .0025353       .4788616    .4888007
                  2018#Yes   |   .5937887   .0050016       .5839849    .6035925
------------------------------------------------------------------------------
```

`ereturn list` Stata output (selected)

```
matrices:
                  e(b) :  1 x 6
                  e(V) :  6 x 6
             e(_N_subp) :  1 x 6
            e(V_srssub) :  6 x 6
               e(V_srs) :  6 x 6
                  e(_N) :  1 x 6
```

Contains the sample size (not in main output)

`matrix list e(_N)` Stata output

	c.prev_tob~o@ 1.survey# 0.poverty	c.prev_tob~o@ 1.survey# 1.poverty	c.prev_tob~o@ 2.survey# 0.poverty	c.prev_tob~o@ 2.survey# 1.poverty	c.prev_tob~o@ 3.survey# 0.poverty	c.prev_tob~o@ 3.survey# 1.poverty
r1	31173	8998	32712	8832	126074	21643

`e(_N)[1,1] + e(_N)[1,2]` for 2012 sample size

`e(_N)[1,3] + e(_N)[1,4]` for 2015 sample size

`e(_N)[1,5] + e(_N)[1,6]` for 2018 sample size

5. Call the results from the relevant matrices and post on the file

```
postfile dumbbell int survey float (nonpoor poor N) using "dumbbell.dta", replace
svyset psu [pweight = weight], strata(stratum) singleunit(centered)
svy: mean prev_tobacco, over(survey poverty)
post dumbbell (2012) (e(b)[1,1]) (e(b)[1,2]) (e(_N)[1,1] + e(_N)[1,2])
post dumbbell (2015) (e(b)[1,3]) (e(b)[1,4]) (e(_N)[1,3] + e(_N)[1,4])
post dumbbell (2018) (e(b)[1,5]) (e(b)[1,6]) (e(_N)[1,5] + e(_N)[1,6])
```

6. *Postclose* when done

```
postfile dumbbell int survey float (nonpoor poor N) using "dumbbell.dta", replace
svyset psu [pweight = weight], strata(stratum) singleunit(centered)
svy: mean prev_tobacco, over(survey poverty)
post dumbbell (2012) (e(b)[1,1]) (e(b)[1,2]) (e(_N)[1,1] + e(_N)[1,2])
post dumbbell (2015) (e(b)[1,3]) (e(b)[1,4]) (e(_N)[1,3] + e(_N)[1,4])
post dumbbell (2018) (e(b)[1,5]) (e(b)[1,6]) (e(_N)[1,5] + e(_N)[1,6])
postclose dumbbell
```

Analysis results stored using *postfile*

Postfile commands post results in a Stata dataset

7. Check output

```
use dumbbell.dta, clear
list
```

Stata output

```
     +-------------------------------------+
     | survey    nonpoor       poor      N |
     |-------------------------------------|
  1. |   2012   .5281904   .6774918   40171 |
  2. |   2015   .5018166   .6485801   41544 |
  3. |   2018   .4838312   .5937887  147717 |
     +-------------------------------------+
```

```
svy: mean prev_tobacco, over(survey poverty)
```

Stata output (selected)

```
------------------------------------------------
                                |
                                |       Mean
--------------------------------+---------------
c.prev_tobacco@survey#poverty   |
                     2012#No    |    .5281905
                    2012#Yes    |    .6774919
                     2015#No    |    .5018166
                    2015#Yes    |    .6485802
                     2018#No    |    .4838312
                    2018#Yes    |    .5937887
------------------------------------------------
```


```
# install.packages("haven")     # install if doing this for the first time
library(haven)                  # For opening Stata dta files
data <- read_dta("dumbbell.dta")
data
```

R console output

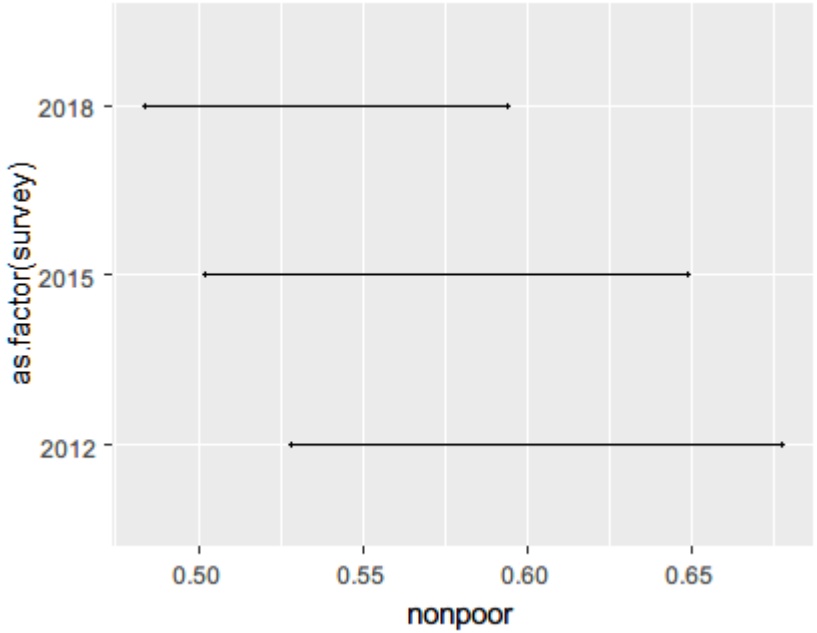
```
# A tibble: 3 × 4
  survey nonpoor  poor      N
   <dbl>   <dbl> <dbl>  <dbl>
1   2012   0.528 0.677  40171
2   2015   0.502 0.649  41544
3   2018   0.484 0.594 147717
```

use dumbbell.dta, clear /// list Stata output

```
     +-------------------------------------+
     | survey    nonpoor       poor        N |
     |-------------------------------------|
  1. |   2012   .5281904   .6774918    40171 |
  2. |   2015   .5018166   .6485801    41544 |
  3. |   2018   .4838312   .5937887   147717 |
     +-------------------------------------+
```

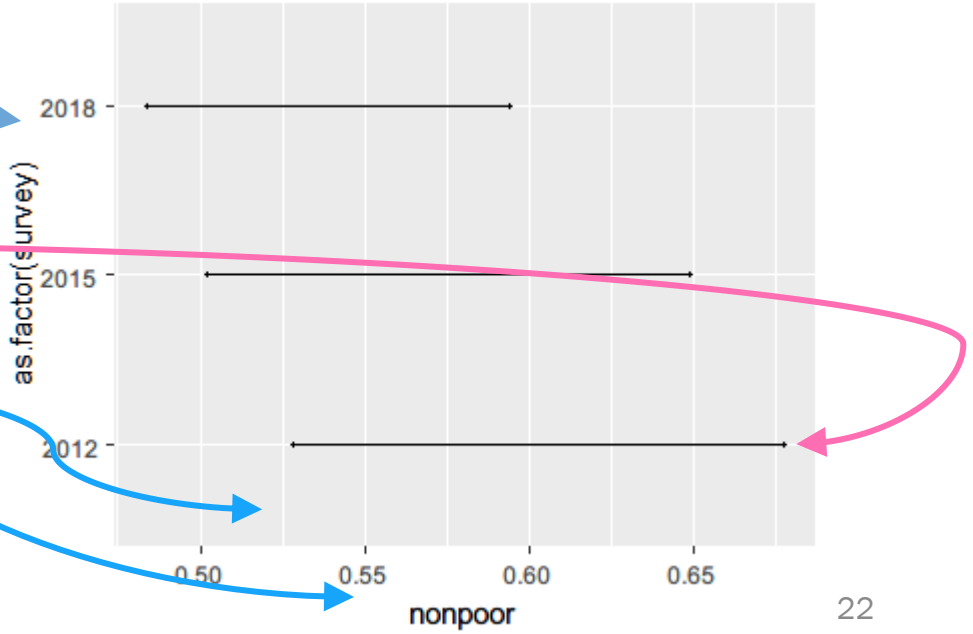
1. Start with the basic plot

```
# install.packages(c("tidyverse", "ggalt"))
library(tidyverse)      # Includes ggplot2
library(ggalt)          # Dumbbell plot extension of ggplot2
dumbbell <- ggplot(data = data,
                   aes(y = as.factor(survey),
                       x = nonpoor,
                       xend = poor)) +
            geom_dumbbell()
dumbbell
```



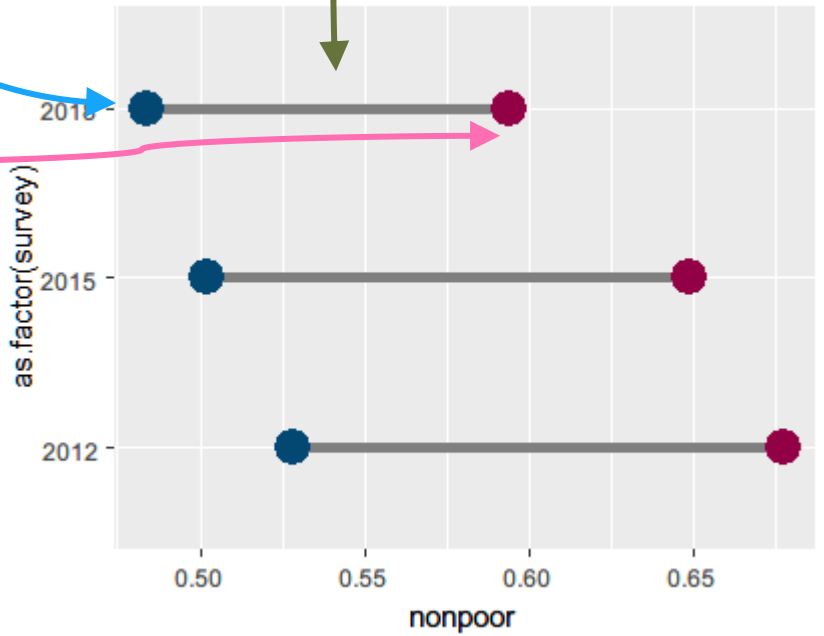
1. Start with the basic plot

```
# install.packages(c("tidyverse", "ggalt"))
library(tidyverse)      # Includes ggplot2
library(ggalt)          # Dumbbell plot extension of ggplot2
dumbbell <- ggplot(data = data,
                   aes(y = as.factor(survey),
                       x = nonpoor,
                       xend = poor)) +
  geom_dumbbell()

dumbbell
```



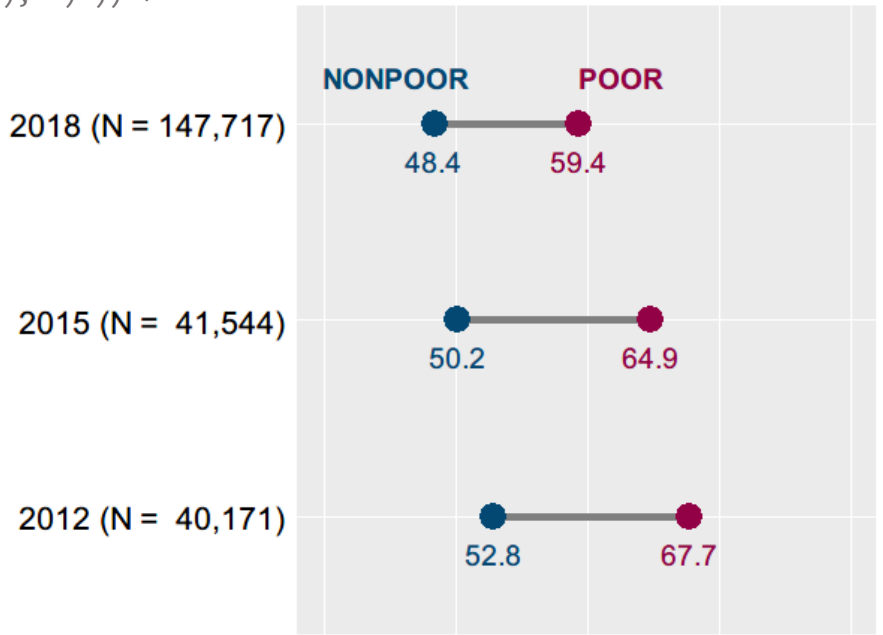
2. Emphasize the plot by adjusting aesthetics

```
dumbbell <- dumbbell +
  geom_dumbbell(size = 2, color = "gray50",
                size_x = 6,
                colour_x = "#024873",
                size_xend = 6,
                colour_xend = "#920045")

dumbbell
```

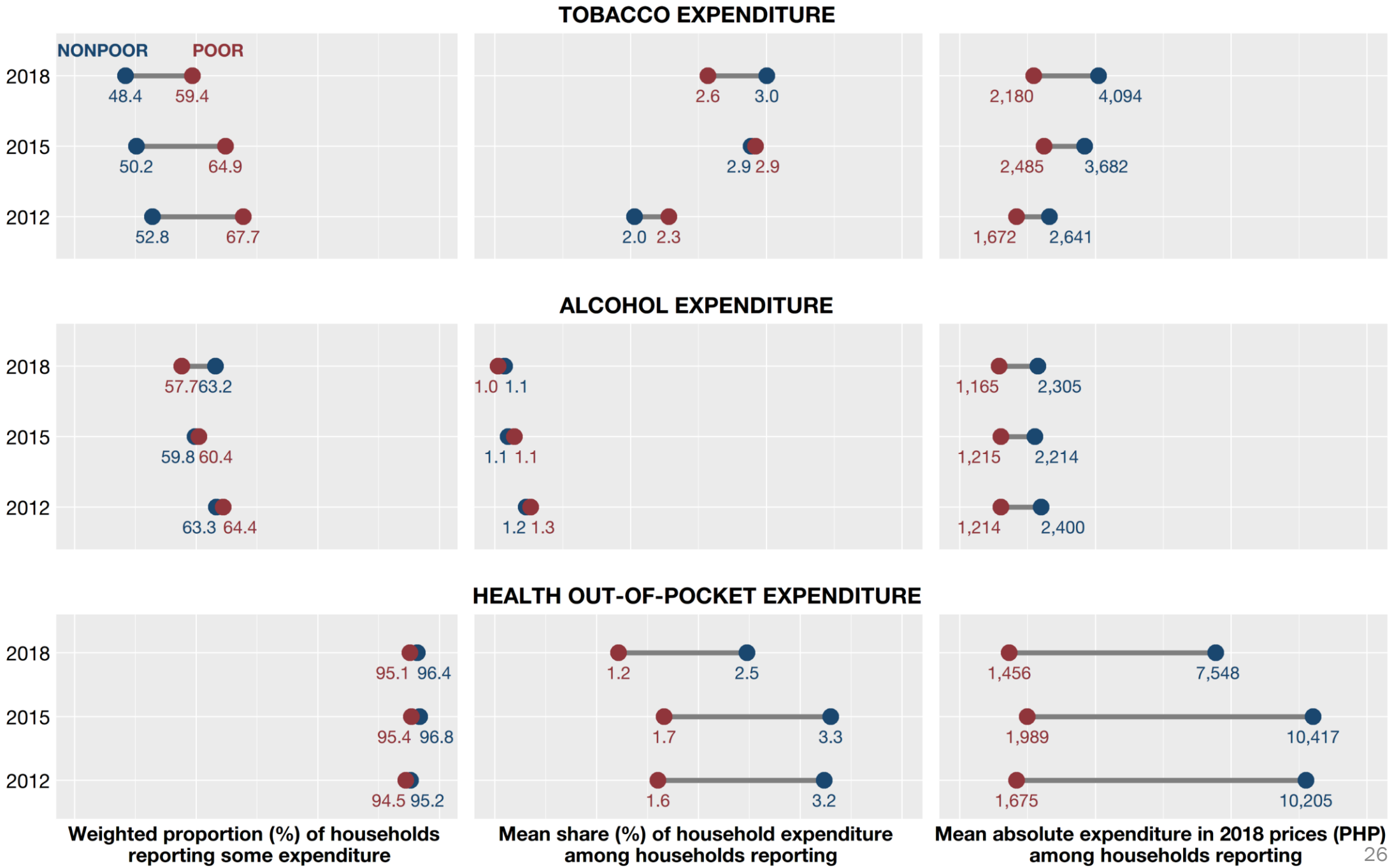


3. Finish the plot by applying a theme and additional modifications

```
dumbbell <- dumbbell +
  # Fix the axes
  scale_y_discrete(name = "“, labels = paste0(data$survey, " (N = ", format(data$N, big.mark = ","), ")")) +
  scale_x_continuous(name = "", limits = c(0.4, 0.8), breaks = seq(0.4, 0.8, 0.2)) +
  # Add percentage labels
  geom_text(aes(x = nonpoor, y = as.factor(survey), label = format(nonpoor*100, digits = 3)),
            color = "#024873", size = 5, vjust = 2) +
  geom_text(aes(x = poor, y = as.factor(survey), label = format(poor*100, digits = 3)),
            color = "#920045", size = 5, vjust = 2) +
   # Add the dot legends
  geom_text(data = filter(data, as.factor(survey) == 2018),
            aes(x = nonpoor, y = 3, label = "NONPOOR"),
            color = "#024873", size = 5, vjust = -1.5, hjust = 0.75, fontface = "bold") +
  geom_text(data = filter(data, as.factor(survey) == 2018),
            aes(x = poor, y = 3, label = "POOR"),
            color = "#920045", size = 5, vjust = -1.5, hjust = 0, fontface = "bold") +
  # Add the theme
  theme(plot.title = element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_text(size = 16, color = "black"),
        axis.title.x = element_text(face = "bold", size = 16, color = "black"),
        axis.title.y = element_blank(),
        axis.ticks = element_blank(),
        panel.background = element_blank())
dumbbell
```

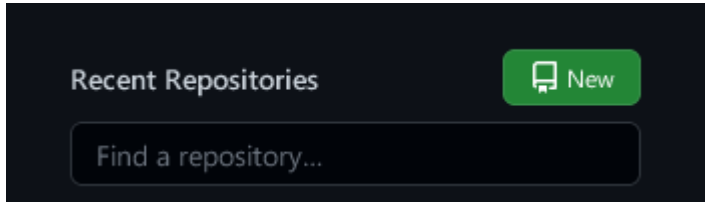


```
*************************************************************************
******** Store summary data in a new Stata .dta file and this will be fed into the R code to generate the dumbbell plot
postfile dumbbell str20 (outcome poverty) int survey float (nonpoor poor N) ///
        using "02_Figures\dumbbell.dta", replace
// Run summary statistics commands for prevalence
foreach var of varlist prev_tobacco prev_alcohol prev_health {
        foreach poor of varlist poor_new2015 poor_old2015 {
                svy: mean `var', over(survey `poor')
                post dumbbell ("`var'") ("`poor'") (2012) (e(b)[1,1]) (e(b)[1,2]) (e(_N)[1,1] + e(_N)[1,2])
                post dumbbell ("`var'") ("`poor'") (2015) (e(b)[1,3]) (e(b)[1,4]) (e(_N)[1,3] + e(_N)[1,4])
                post dumbbell ("`var'") ("`poor'") (2018) (e(b)[1,5]) (e(b)[1,6]) (e(_N)[1,5] + e(_N)[1,6])
        }
}
// Run summary statistics commands for share and absolute value, subsetting on prevalence
local outcomes_group "share_tobacco_totex share_alcohol_totex share_health_totex tobacco_2018 alcohol_2018 health_2018"
local subsets_group "prev_tobacco prev_alcohol prev_health prev_alcohol prev_alcohol prev_health"
local n: word count `outcomes_group'
forvalues i = 1/`n' {
        local outcomes: word `i' of `outcomes_group'
        local subsets: word `i' of `subsets_group'
        foreach poor of varlist poor_new2015 poor_old2015 {
                svy, subpop(`subsets'): mean `outcomes', over(survey `poor')
                post dumbbell ("`outcomes'") ("`poor'") (2012) (e(b)[1,1]) (e(b)[1,2]) (e(_N)[1,1] + e(_N)[1,2])
                post dumbbell ("`outcomes'") ("`poor'") (2015) (e(b)[1,3]) (e(b)[1,4]) (e(_N)[1,3] + e(_N)[1,4])
                post dumbbell ("`outcomes'") ("`poor'") (2018) (e(b)[1,5]) (e(b)[1,6]) (e(_N)[1,5] + e(_N)[1,6])
        }
}
postclose dumbbell
```

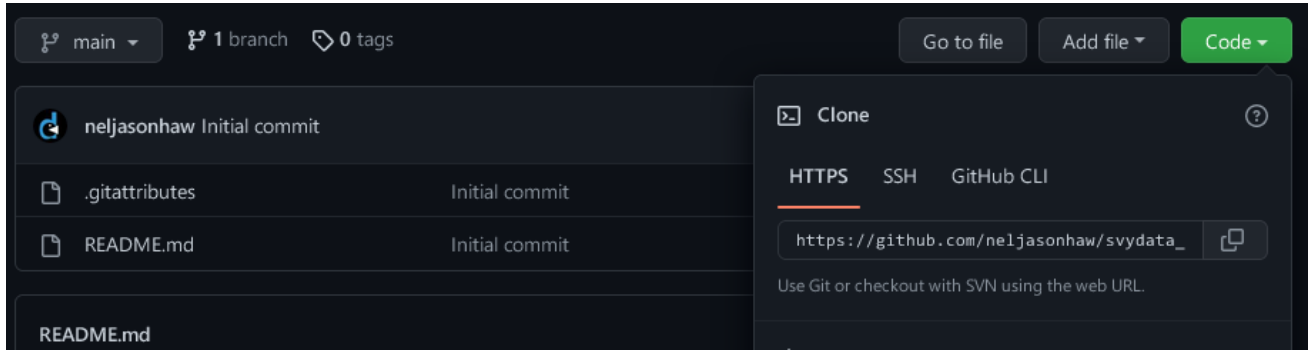

TOBACCO EXPENDITURE
NONPOOR
POOR
2018
2015
2012
48.4
59.4
50.2
64.9
52.8
67.7
2.6
3.0
2.9
2.9
2.0
2.3
2,180
4,094
2,485
3,682
1,672
2,641
ALCOHOL EXPENDITURE
57.7
63.2
59.8
60.4
63.3
64.4
1.0
1.1
1.1
1.1
1.2
1.3
1,165
2,305
1,215
2,214
1,214
2,400
HEALTH OUT-OF-POCKET EXPENDITURE
95.1
96.4
95.4
96.8
94.5
95.2
1.2
2.5
1.7
3.3
1.6
3.2
1,456
7,548
1,989
10,417
1,675
10,205
Weighted proportion (%) of households reporting some expenditure
Mean share (%) of household expenditure among households reporting
Mean absolute expenditure in 2018 prices (PHP) among households reporting

1 Set up pre-requisites

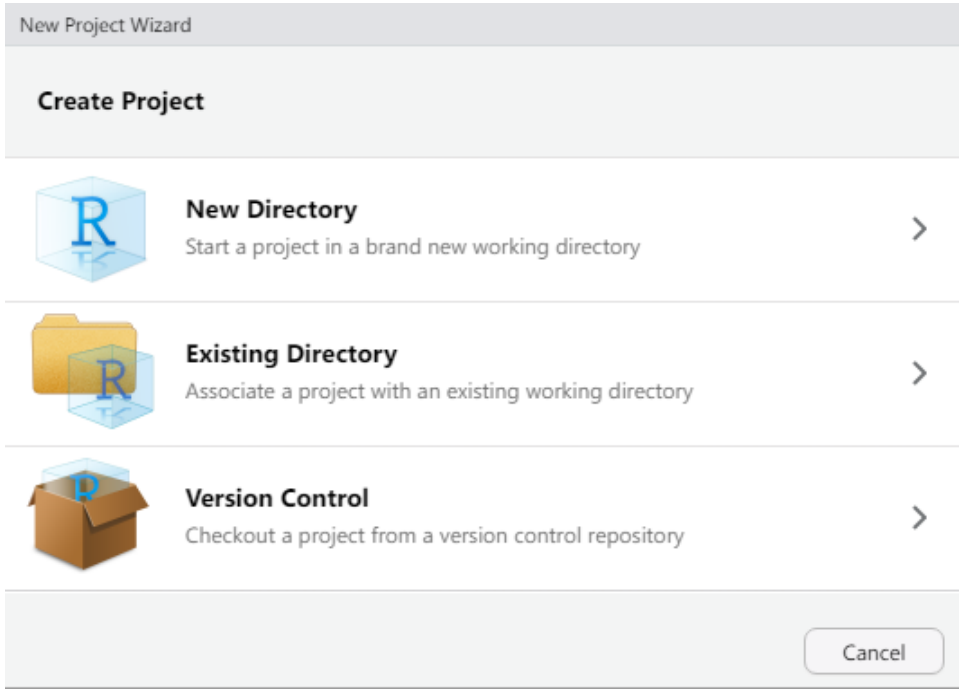
a. Install Git (git-scm.com/downloads)
b. Check for the installation of Git in the shell
   `where git (Windows) / which git (Linux/Mac)`
c. Register / sign in on Github (github.com)
d. Create a new repository



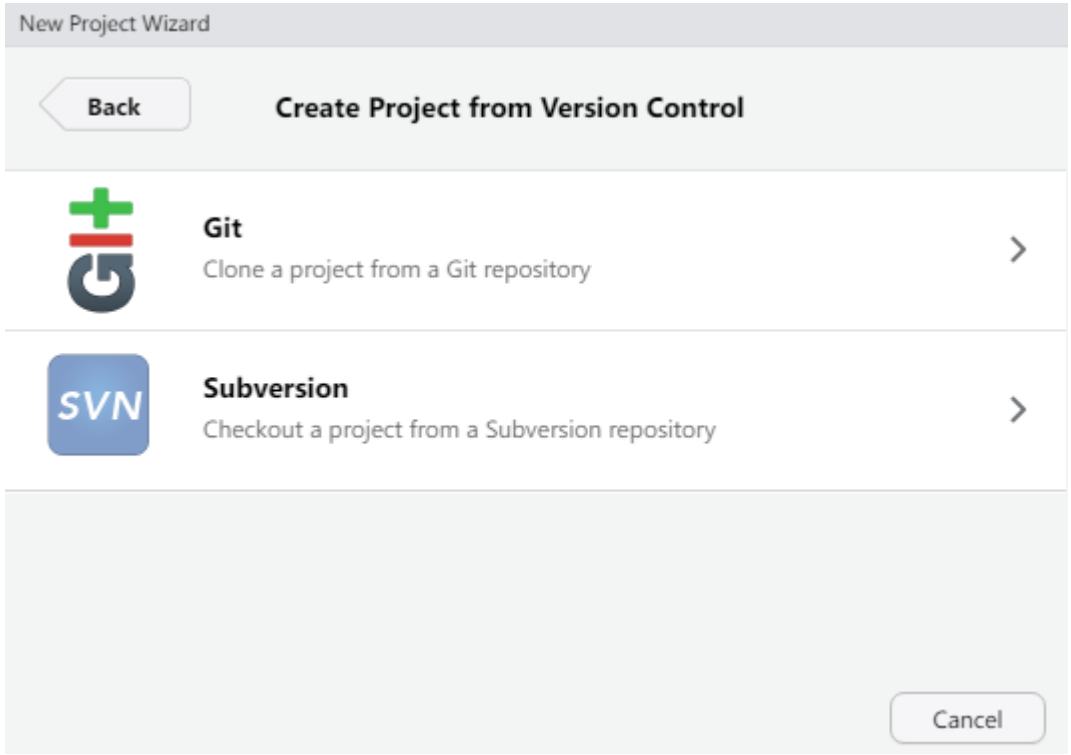
e. Copy the URL via the “Code” button



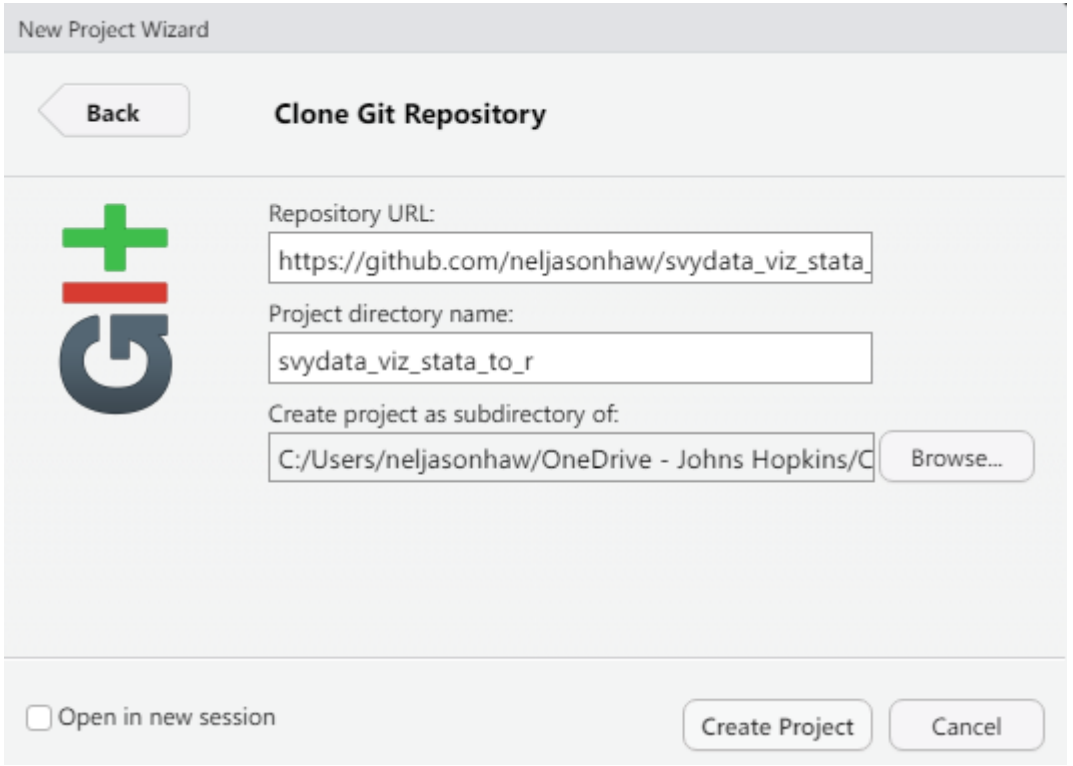
2 On RStudio, File > New Project… then select Version Control



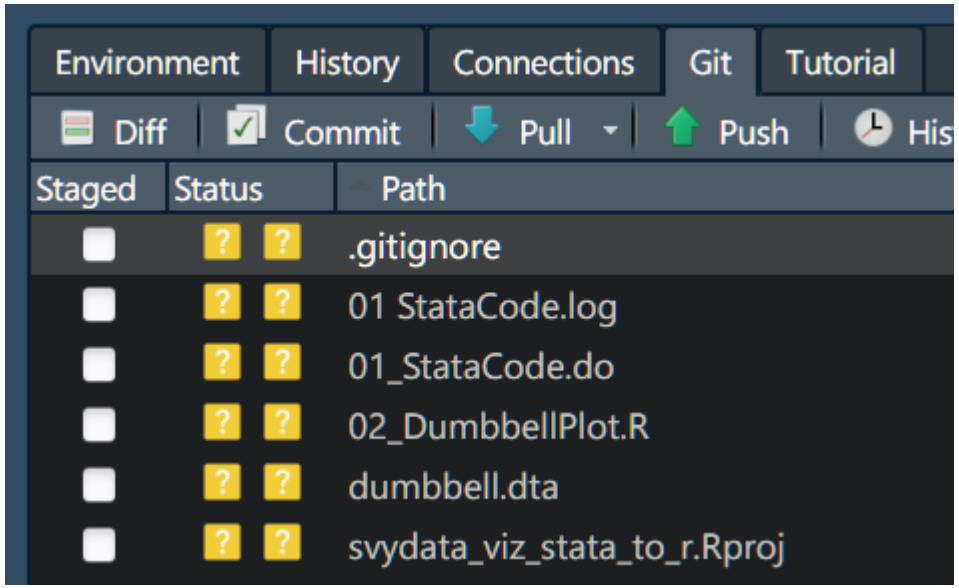
3 Select Git –
Clone a project from a Git repository



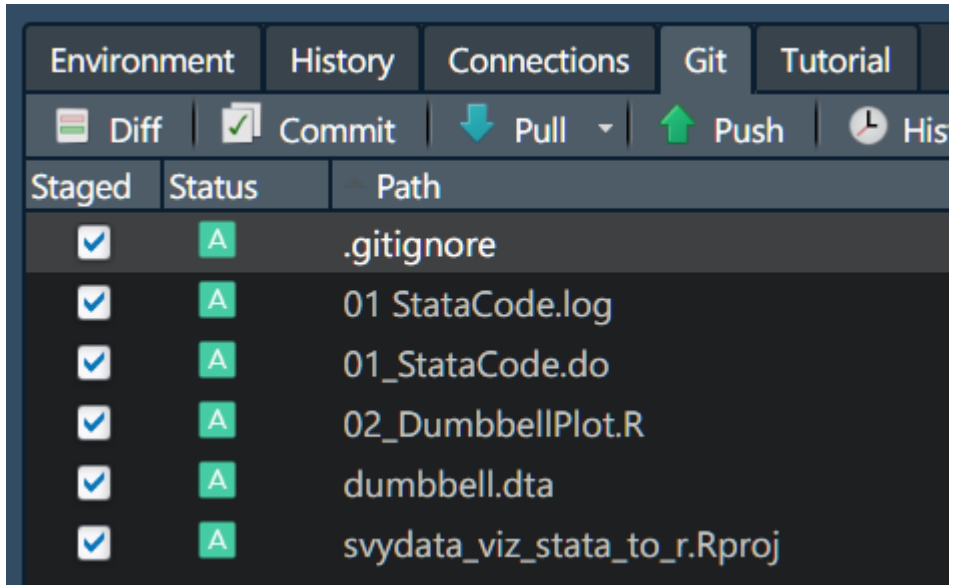
4 Copy the repository URL and identify the local subfolder where the files are housed and click Create Project



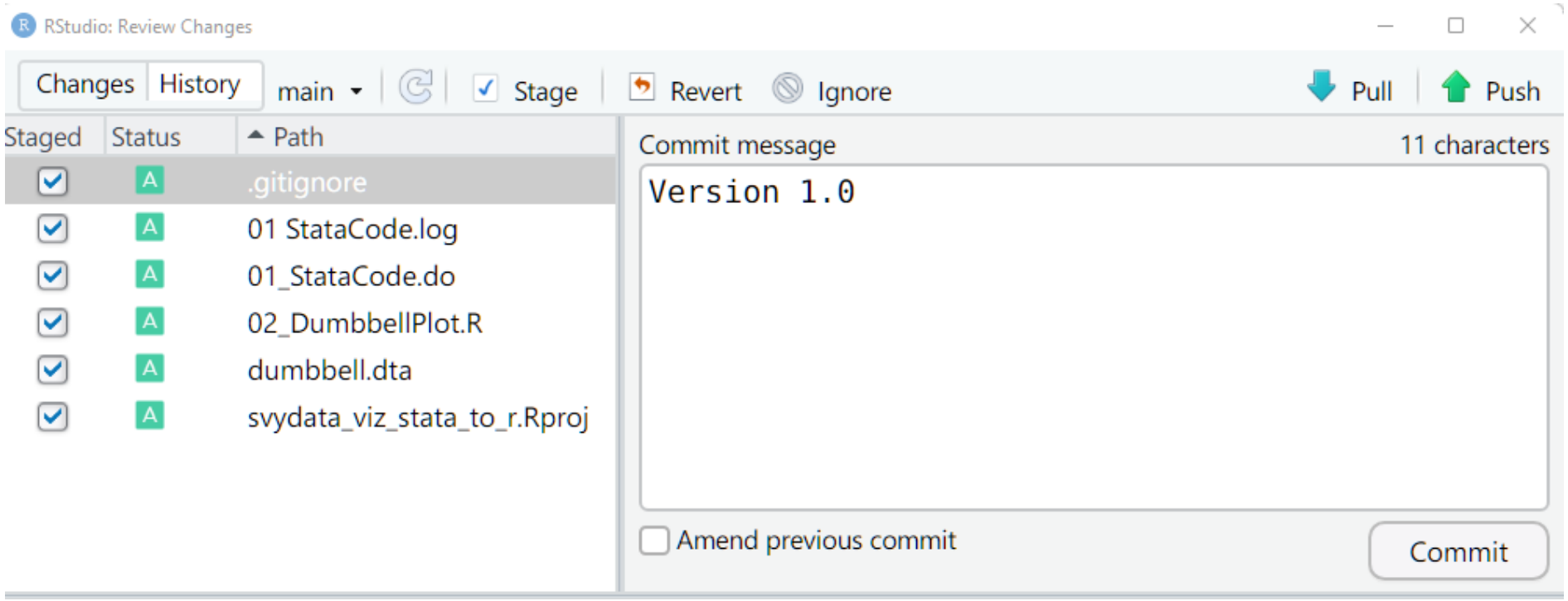
5 Copy any relevant local files into the local subdirectory of the Git repository and they will all appear on the Git tab on RStudio



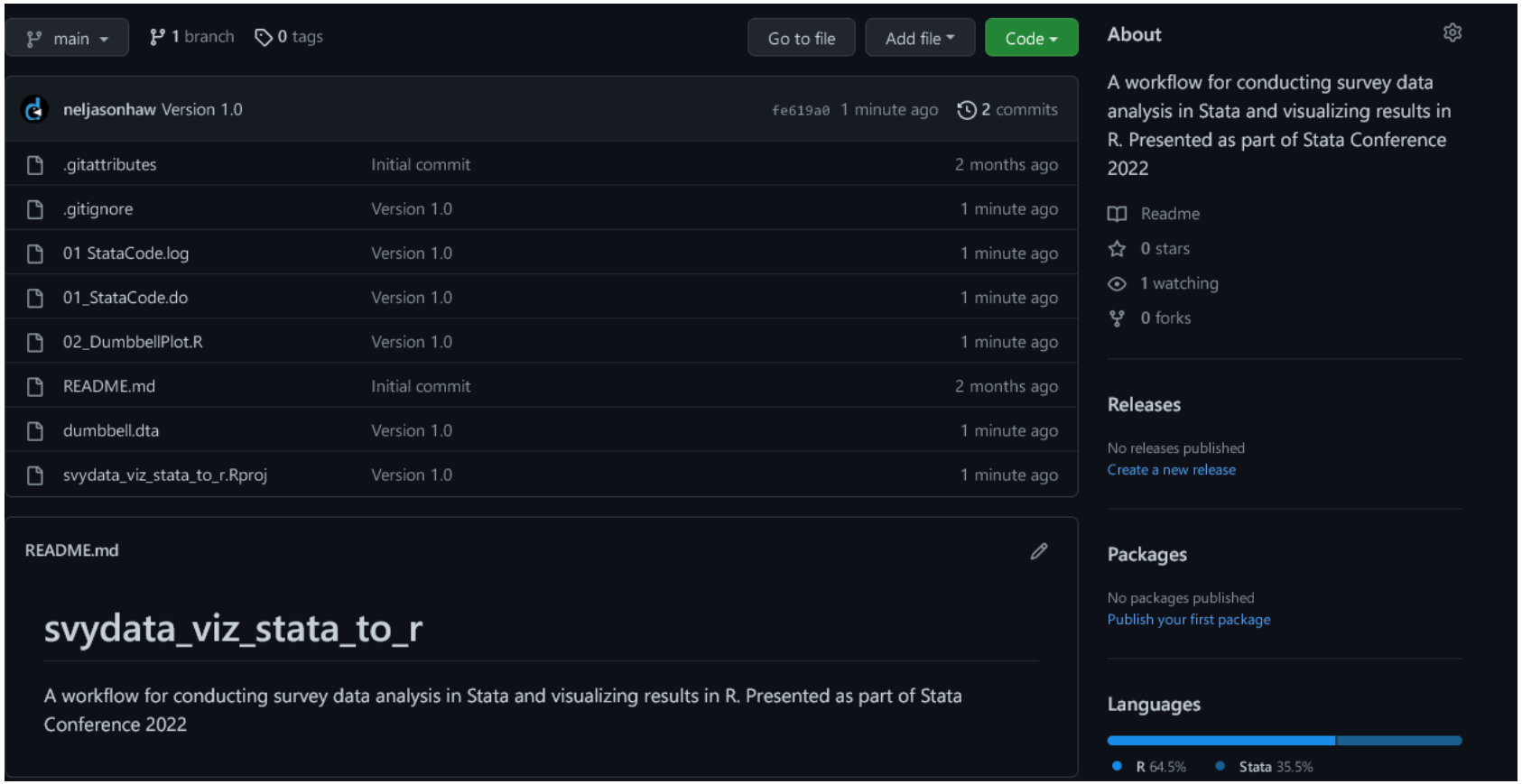
6 Stage any file to be uploaded on Github then click Commit



7 Add a commit message the click Commit then Push



8 The files will appear on your Github repository



Easy-to-use survey data analysis tools with excellent resources

Processed survey data

Survey data analysis

Analysis results stored using *postfile*

Highly flexible execution of the Grammar of Graphics

Read .dta using *haven*

Create plots using *ggplot2*

Upload repository to Github

Visualizing Survey Data Analysis Results:
Marrying the Best from Stata and R

The Github repository for this demonstration is found at github.com/neljasonhaw/svydata_viz_stata_to_r

The Github repository for the entire research project used in this demonstration is found at github.com/neljasonhaw/fies_health_inequalities

Nel Jason (Jason) L. Haw, MS
PhD Student, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
nhaw1@jh.edu **@jasonhaw_** **linkedin.com/in/neljasonhaw**