THE UNIVERSITY OF MELBOURNE

QuantLunch #2: Longitudinal Data using Stata

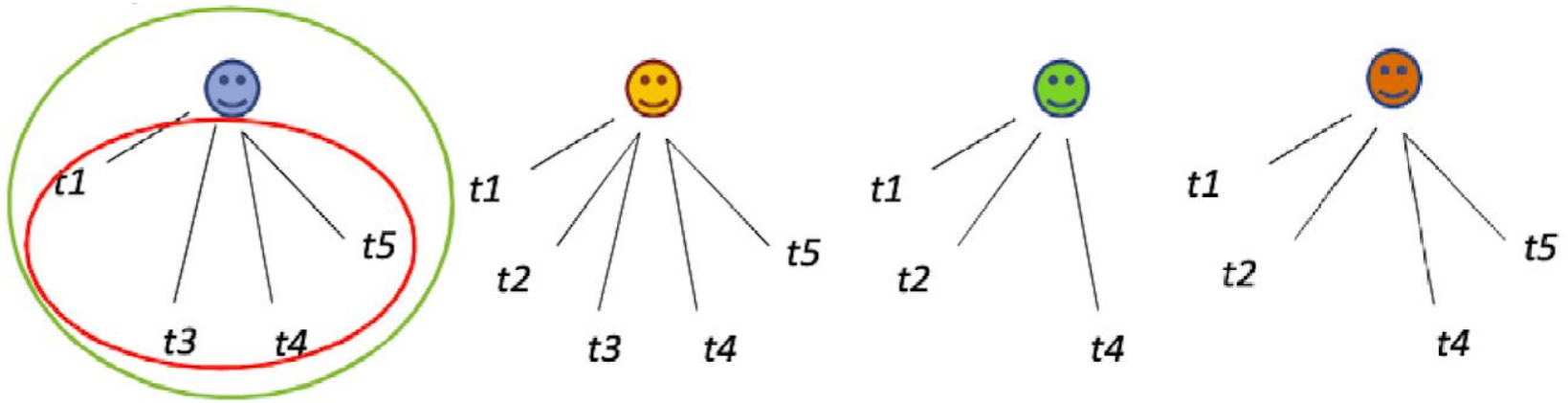
Irma Mooi-Reci

Outline

1. Longitudinal Data & Structure
2. Stata
3. Visualisation

Longitudinal Data & Structure

- Data collected over time within the same units such as individuals, countries, households, etc.



- **Purpose**: Captures changes within the same units over time.

Longitudinal Data & Structure

Distinction Between Wide and Long Format

- **Wide Format**:
  - **Structure**: Single row per unit (e.g., one row per individual).
  - **Example**: Each time point is represented by a separate column (e.g., BMI_2019, BMI_2020, etc.).

Wide format

	pid	esempst2001	esempst2002	esempst2003	esempst2004	
1	100003	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3]
2	100005	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1]
3	100010	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed
4	100014	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
5	100015	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
6	100016	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
7	100024	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed
8	100028	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
9	100029	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
10	100038	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
11	100043	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
12	100048	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
13	100052	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
14	100053	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
15	100057	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
16	100058	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
17	100059	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
18	100060	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
19	100071	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
20	100075	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
21	100078	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
22	100079	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
23	100083	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
24	100085	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
25	100087	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
26	100088	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed
27	100097	[2] Employee of own business	[2] Employee of own business	[2] Employee of own business	[2] Employee of own business	[2] Employee of own business
28	100099	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
29	100107	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
30	100113	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
31	100114	[1] Employee	[1] Employee	[1] Employee	[1] Employee	[1] Employee
32	100128	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed
33	100129	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed	[3] Employer/Self-employed

1 row per unit

Longitudinal Data & Structure

Distinction Between Wide and Long Format

- **Long Format**:
  - **Structure**: Multiple rows per unit, with each row representing a different time point.
  - **Example**: One column for time, another for the variable of interest (e.g., BMI), with each row representing a specific time point for the individual.

Long format: Person-Period Format

	pid	hgage	hgsex	esempst	female	lnwage2	year
1	100001	50	[1] Male	[1] Employee	0	2.47997	2002
2	100001	51	[1] Male	[1] Employee	0	2.681021	2003
3	100001	52	[1] Male	[1] Employee	0	2.630089	2004
4	100002	48	[2] Female	[1] Employee	1	2.336987	2001
5	100002	49	[2] Female	[1] Employee	1	2.033965	2002
6	100002	50	[2] Female	[1] Employee	1	2.518502	2003
7	100002	51	[2] Female	[1] Employee	1	2.694627	2004
8	100003	48	[1] Male	[3] Employer/Self-employed	0	4.60517	2001
9	100003	49	[1] Male	[3] Employer/Self-employed	0	4.60517	2002
10	100003	50	[1] Male	[3] Employer/Self-employed	0	4.60517	2003
11	100003	51	[1] Male	[3] Employer/Self-employed	0	1.965112	2004
12	100003	52	[1] Male	[3] Employer/Self-employed	0	4.60517	2005
13	100003	54	[1] Male	[1] Employee	0	2.568788	2007
14	100003	56	[1] Male	[1] Employee	0	2.99906	2009
15	100003	57	[1] Male	[1] Employee	0	3.284663	2010
16	100004	38	[2] Female	[1] Employee	1	2.219203	2001
17	100004	39	[2] Female	[1] Employee	1	2.222459	2002
18	100004	40	[2] Female	[1] Employee	1	4.169421	2003
19	100004	41	[2] Female	[1] Employee	1	2.458905	2004
20	100005	16	[2] Female	[1] Employee	1	1.526056	2001
21	100005	17	[2] Female	[1] Employee	1	1.976624	2002
22	100005	18	[2] Female	[1] Employee	1	1.477049	2003
23	100005	19	[2] Female	[1] Employee	1	2.439735	2004
24	100005	20	[2] Female	[1] Employee	1	2.264537	2005
25	100005	30	[2] Female	[1] Employee	1	2.721295	2015
26	100005	31	[2] Female	[1] Employee	1	3.234524	2016
27	100005	32	[2] Female	[1] Employee	1	3.451375	2017
28	100005	33	[2] Female	[1] Employee	1	3.267666	2018
29	100005	34	[2] Female	[1] Employee	1	3.24443	2019
30	100005	35	[2] Female	[1] Employee	1	3.55612	2020

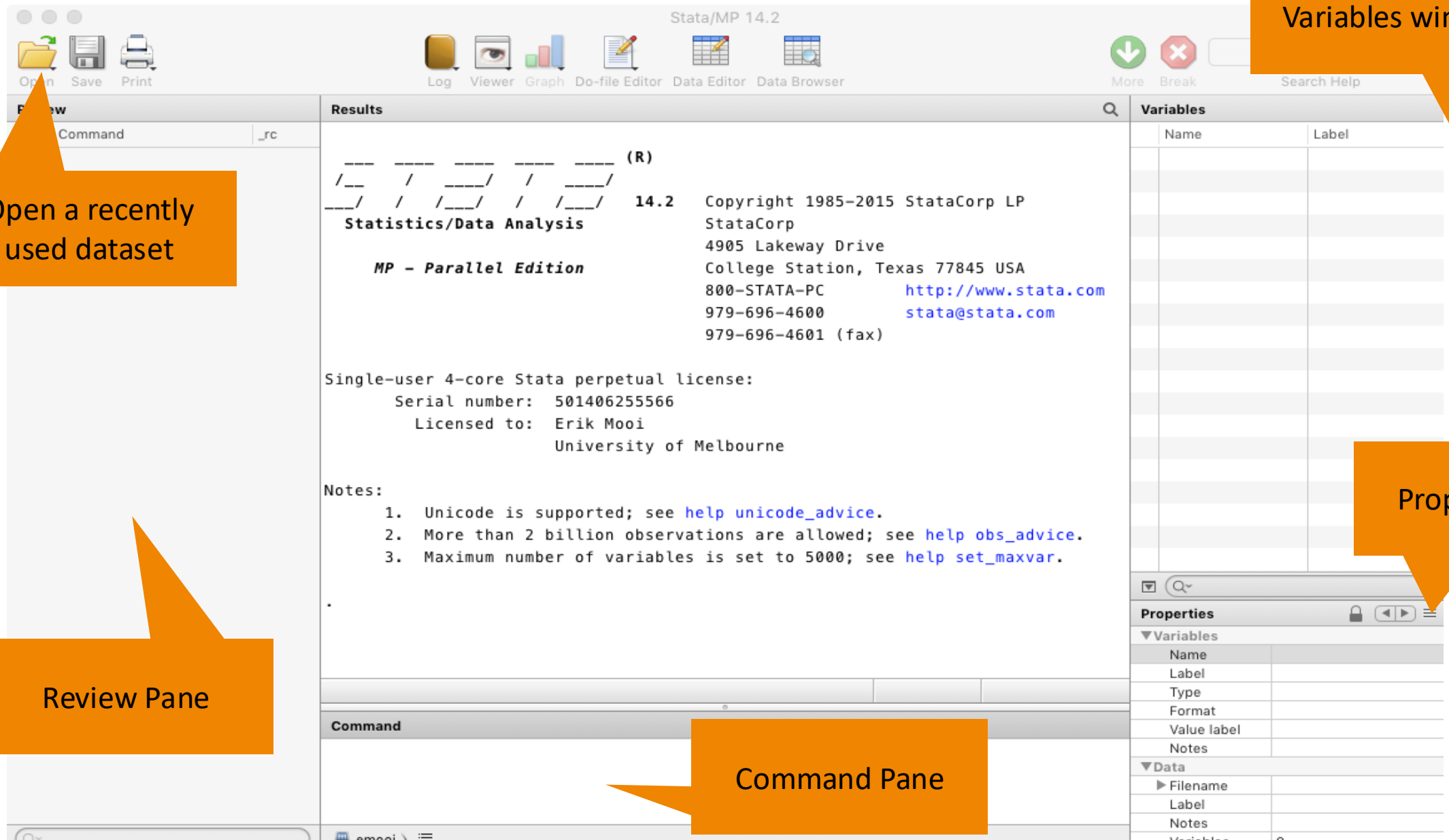
Multiple rows per unit

Outline

1. Longitudinal Data & Structure
2. Stata
3. Visualisation

Introduction to Stata: The Interface

Market Research Using Stata (Mooi et al. 2018)



Open a recently used dataset

Variables window

Properties Pane

Review Pane

Command Pane

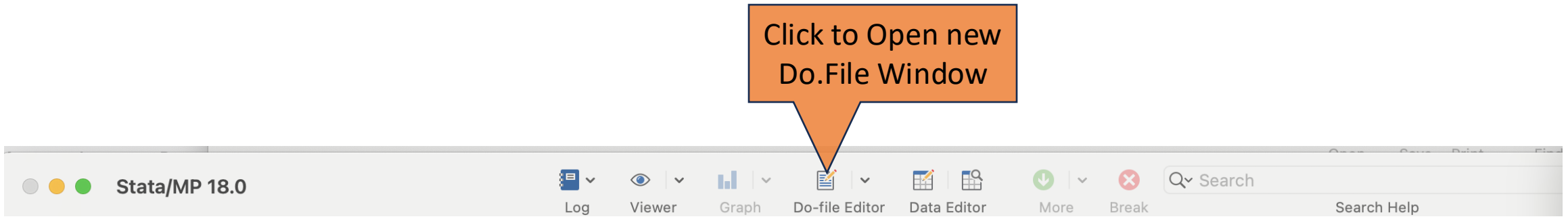
Do-files

- Do-files are text files that contain a series of Stata commands, written in the same way you would enter them interactively in the Command pane.

Why Use Do-Files?

- **Efficiency**: They save time by automating repetitive tasks.
- **Consistency**: They ensure that the same commands are run in the same order every time, reducing the risk of errors.
- **Documentation**: They serve as a clear and detailed record of your analysis process.

`Go to Window → Do-File Editor`



Structuring do-files (II)

```
*==============================================================================
*                              Description of dofile
*                               created by IMR
/*==============================================================================
This dofile generates Figures and Tables for the paper "Working from Home and
the Consequences for Labour Turnover and Career Progression". The structure of
this dofile is as follows:

  1. Setting paths
  2. Opening data
  3. Define the two samples based on the 2 survey components (PQ and SCQ)
  4. Replication descriptive statistics (Table 1)
  5. Replication results (Tables 2-4, Figures 2,3)
  6. Additional robustness checks

*==============================================================================
*                1. Setting paths and installing ado-files
*==============================================================================

    global ddta  "/Users/imooi/Documents/DATA/ECR/ddta/HILDA/HILDA RESTRICTED"
    global dtemp "/Users/imooi/Documents/DATA/ECR/ddta/HILDA/TEMP"
    global doutput "/Users/imooi/Documents/DATA/ECR/ddta/HILDA/OUTPUT"

    clear all
    clear matrix
    set more off
    set maxvar 120000

*==============================================================================
*                            2. Opening the data
*==============================================================================

    use  "$dtemp/HILDA_wfh.dta", clear

*==============================================================================
*                3. Define the two samples (PQ and SCQ)
*               Sample 1: PQ; with valid observations for control variables
*               Sample 2: SCQ; with valid observations for control variables
*==============================================================================
```

Combining different waves

```
*****************************************
*** WFH AND PROMOTIONS, LABOUR TURNOVER ****
*** created by: IMR                     ****
*****************************************

**** VARIABLES ****

    clear all
    clear matrix
    set more off
    set maxvar 120000

    global ddta  "/Users/imooi/Documents/DATA/ECR/ddta/HILDA/HILDA DATA NEW"
    global dtemp "/Users/imooi/Documents/DATA/ECR/ddta/HILDA/TEMP"
    global doutput "/Users/imooi/Documents/DATA/ECR/ddta/HILDA/OUTPUT"

    local varstokeep hhstate hhwtrp hhwtsc hhwte hhidate hhrhid hhrpid hhrih hhiage hgsex mrcurr esbrd esdtl losat jbmhrha jbmhrhw
jbmhrh jbmh jbn gh9i jbmwpsz jbmemsz jbmmwp ///
                    lsvol lsod lsocd lshw lserr lsemp lscom lschd lscar pjljrea jbmplej jbmpgj ehtuj pjsemp pjmsemp jbmo06 ///
                    lshrcom lshrcar lshrvol lshrchd lshrod lshrhw lshrerr ///
                    jbhruc hhtup hhtuh hhfty hhpers hhtype hhpxid hhfxid hhmxid hhstate lsrush lsrelsp pawkfle pawklte lebth lejob
///
                    lsrlrel lshhdiv patird jomcd jomwi jomus jomini jomfd jomflex jbempt ///
    local i = 0
    foreach w in a b c d e f g h i j k l m n o p q r s t u v {
        use "$ddta/Combined_`w'220u.dta", clear
        renpfix `w'      // Strip off wave prefix
        local i = `i'+1  // Increase (wave) counter by 1
        gen wave = `i'   // Create wave indicator (1, 2, ...)
        // select variables needed
        if ("`varstokeep'"!="") {
            local tokeep                                // empty to keep list
            foreach var of local varstokeep {           // loop over all selected variables
                capture confirm variable `var'          // check whether variable exists in current wave
                if (!_rc) local tokeep `tokeep' `var'   // mark for inclusion if variable exists
                }
            keep xwaveid wave `tokeep' // keep selected variables
            }
        // Save temporary data file
        tempfile tempdata_`w'
        save "`tempdata_`w'`"
    }

    clear
    foreach w in a b c d e f g h i j k l m n o p q r s t u v {
        append using "`tempdata_`w'`"
        }
    sort  xwaveid wave

    destring xwaveid , gen(pid)
    sum pid
```

Outline

1. Longitudinal Data & Structure
2. Reshape function in Stata
3. Visualisation

Long format: Person-Period Format

	pid	hgage	hgsex	esempst	female	lnwage2	year
1	100001	50	[1] Male	[1] Employee	0	2.47997	2002
2	100001	51	[1] Male	[1] Employee	0	2.681021	2003
3	100001	52	[1] Male	[1] Employee	0	2.630089	2004
4	100002	48	[2] Female	[1] Employee	1	2.336987	2001
5	100002	49	[2] Female	[1] Employee	1	2.033965	2002
6	100002	50	[2] Female	[1] Employee	1	2.518502	2003
7	100002	51	[2] Female	[1] Employee	1	2.694627	2004
8	100003	48	[1] Male	[3] Employer/Self-employed	0	4.60517	2001
9	100003	49	[1] Male	[3] Employer/Self-employed	0	4.60517	2002
10	100003	50	[1] Male	[3] Employer/Self-employed	0	4.60517	2003
11	100003	51	[1] Male	[3] Employer/Self-employed	0	1.965112	2004
12	100003	52	[1] Male	[3] Employer/Self-employed	0	4.60517	2005
13	100003	54	[1] Male	[1] Employee	0	2.568788	2007
14	100003	56	[1] Male	[1] Employee	0	2.99906	2009
15	100003	57	[1] Male	[1] Employee	0	3.284663	2010
16	100004	38	[2] Female	[1] Employee	1	2.219203	2001
17	100004	39	[2] Female	[1] Employee	1	2.222459	2002
18	100004	40	[2] Female	[1] Employee	1	4.169421	2003
19	100004	41	[2] Female	[1] Employee	1	2.458905	2004
20	100005	16	[2] Female	[1] Employee	1	1.526056	2001
21	100005	17	[2] Female	[1] Employee	1	1.976624	2002
22	100005	18	[2] Female	[1] Employee	1	1.477049	2003
23	100005	19	[2] Female	[1] Employee	1	2.439735	2004
24	100005	20	[2] Female	[1] Employee	1	2.264537	2005
25	100005	30	[2] Female	[1] Employee	1	2.721295	2015
26	100005	31	[2] Female	[1] Employee	1	3.234524	2016
27	100005	32	[2] Female	[1] Employee	1	3.451375	2017
28	100005	33	[2] Female	[1] Employee	1	3.267666	2018
29	100005	34	[2] Female	[1] Employee	1	3.24443	2019
30	100005	35	[2] Female	[1] Employee	1	3.55612	2020

Multiple rows per unit

convert the dataset into a wide format

convert the dataset into a wide format.

identifies the unique units

Indicates data is organised by year

```
.         reshape wide hgage esempst lnwage2, i(pid) j(year)      // put time-varying vars
> first followed by time constant variables
(j = 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2
> 018 2019 2020 2021 2022)

Data                               Long   ->   Wide
-----------------------------------------------------------------------------
Number of observations          213,499   ->   26,581
Number of variables                   5   ->   67
j variable (22 values)             year   ->   (dropped)
xij variables:
                                  hgage   ->   hgage2001 hgage2002 ... hgage2022
                                esempst   ->   esempst2001 esempst2002 ... esempst2022
                                lnwage2   ->   lnwage22001 lnwage22002 ... lnwage22022
-----------------------------------------------------------------------------
```

Fewer observations, but more variables

Variables with adjusted names

Reshape data: From wide to long

```
.          reshape long lnwage2 esempst hgage, i(pid) j(year)
(j = 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
>  2016 2017 2018 2019 2020 2021 2022)

Data                                  Wide   ->   Long
-----------------------------------------------------------------------------
Number of observations              4,504    ->   99,088
Number of variables                    67    ->   5
j variable (22 values)                       ->   year
xij variables:
lnwage22001 lnwage22002 ... lnwage22022      ->   lnwage2
esempst2001 esempst2002 ... esempst2022      ->   esempst
      hgage2001 hgage2002 ... hgage2022      ->   hgage
-----------------------------------------------------------------------------

.
```

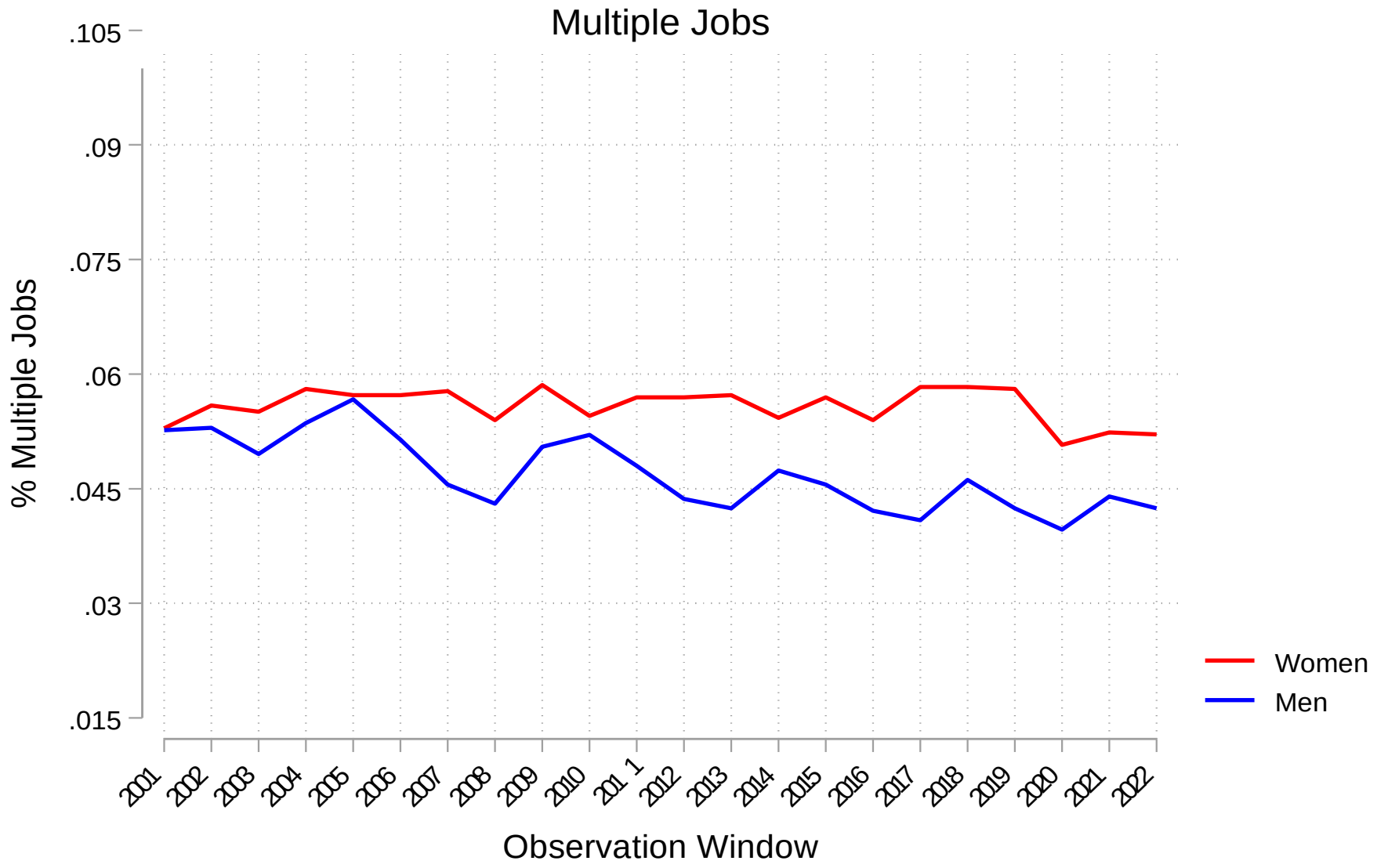
More observations, but fewer variables

A new “year” variable now appears again

Outline

1. Longitudinal Data & Structure
2. Stata & Reshape function
3. Visualisation

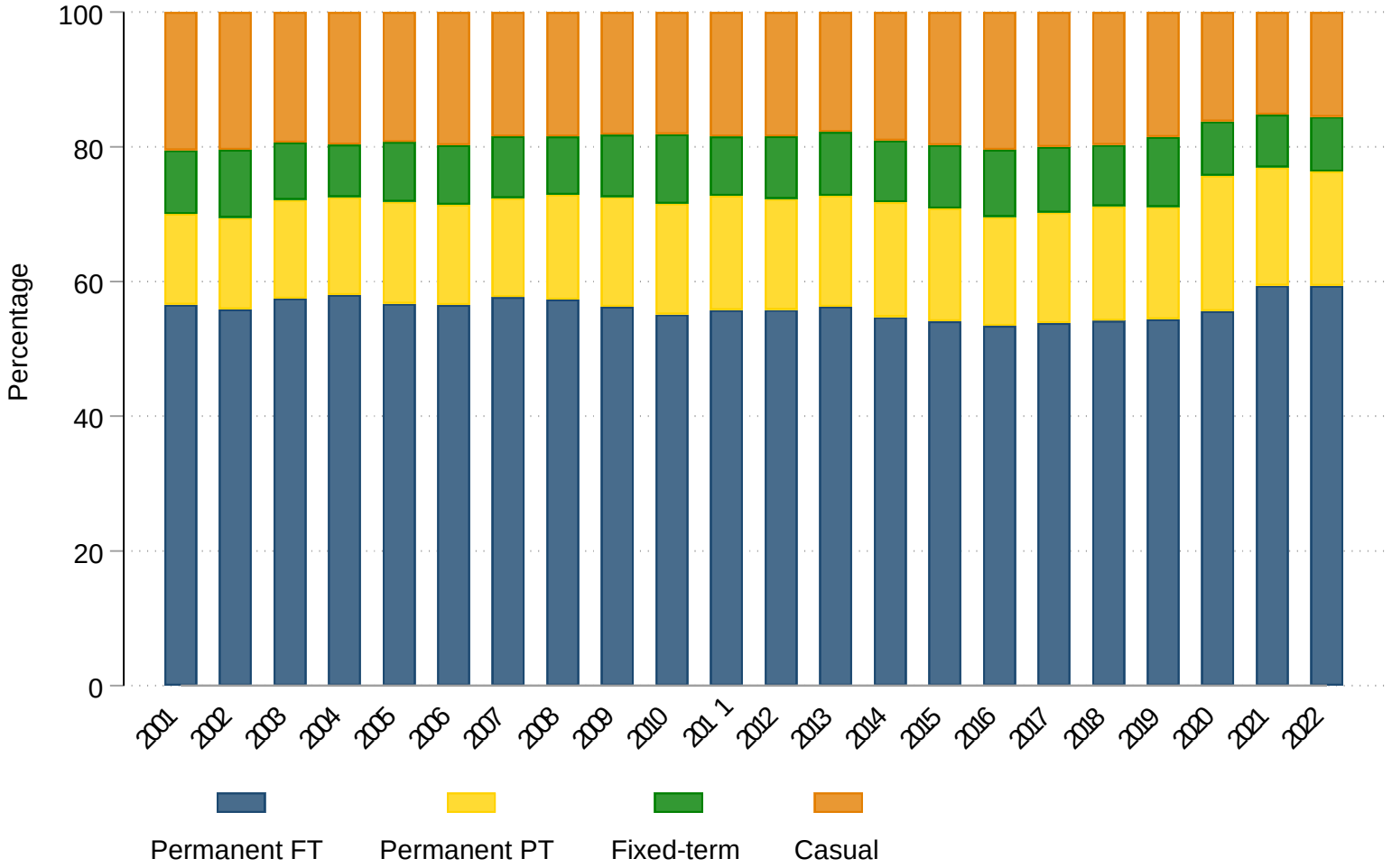
Illustrating Trends (I)



Illustrating Trends - code

```
*==================================================================================*/
*                  2. Figures; cross sectional data                                *
*==================================================================================*

//Graph 1
    use "$dtemp/quantlab.dta", clear

    gen mjobs=jbn==1

    collapse mjobs, by(year female)

    graph twoway (line mjobs year if female==1, lcolor(red) lw(medthick)) (line mjobs year if female==0,
lcolor(blue) lw(medthick)), xtitle("Observation Window") ytitle("% Multiple Jobs") title("Multiple Jobs")
xlabel(2001(1)2022) ylabel(.015(.015).1) legend(order(1 "Women" 2 "Men")) graphregion(fcolor(white))
```

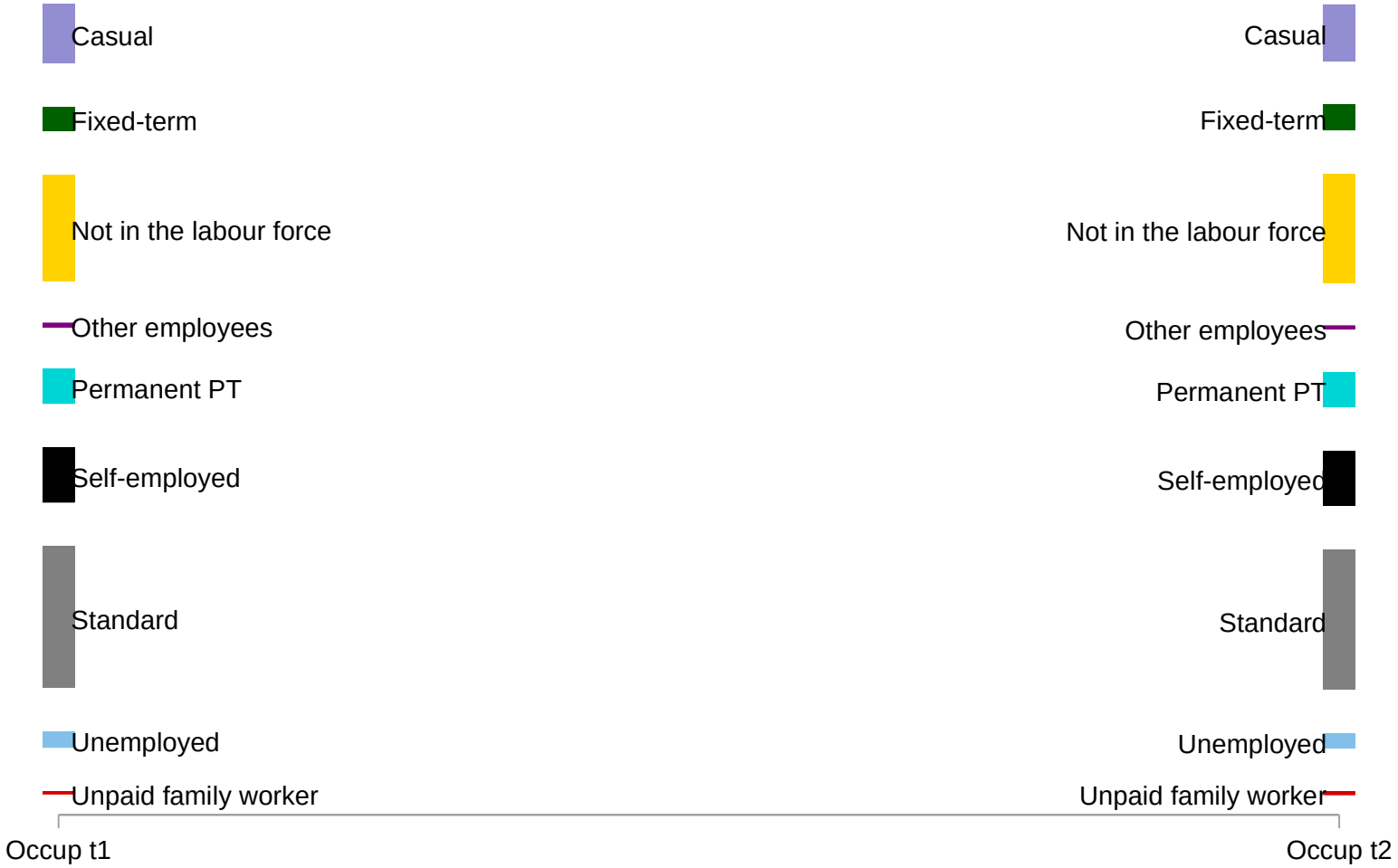
Illustrating Trends (II)



Illustrating Trends - Code

```
*===============================================================================*/
// Graph 2
    ** re-open saved data and proceed with next figures and analyses
    use "$dtemp/quantlab.dta", clear

    // keep only those in dependent employment
    keep if esempst==1

    // keep non-gig workers
    keep if employment_arrangement<5

    // keep only the attached workforce
    keep if hgage > 17 & hgage < 65

    // cross-sectional figure

    set scheme cleanplots
    colorpalette s2, n(7) nograph

    catplot employment_arrangement year, percent(year) asyvars stack ///
    var1opts(label(labsize(small))) var2opts(label(labsize(small) ang(45))) recast(bar) ///
    ytitle("Percentage", size(small)) ///
    graphr(ic(white) fc(white) lc(white)) plotr(ic(white) fc(white) lc(white)) ///
    bar(1, color(navy) fintensity(inten80)) ///
    bar(2, color(gold) fintensity(inten80)) ///
    bar(3, color(green) fintensity(80)) ///
    bar(4, color(dkorange) fintensity(80)) ///
    bar(5, color(cranberry) fintensity(80)) ///
    bar(6, color(lavender) fintensity(80)) ///
    bar(7, color(teal) fintensity(inten80)) ///
    legend(rows(1) stack size(small) ///
    order(1 "Permanent FT" 2 "Permanent PT" 3 "Fixed-term" 4 "Casual") ///
    symplacement(center) pos(7))
```

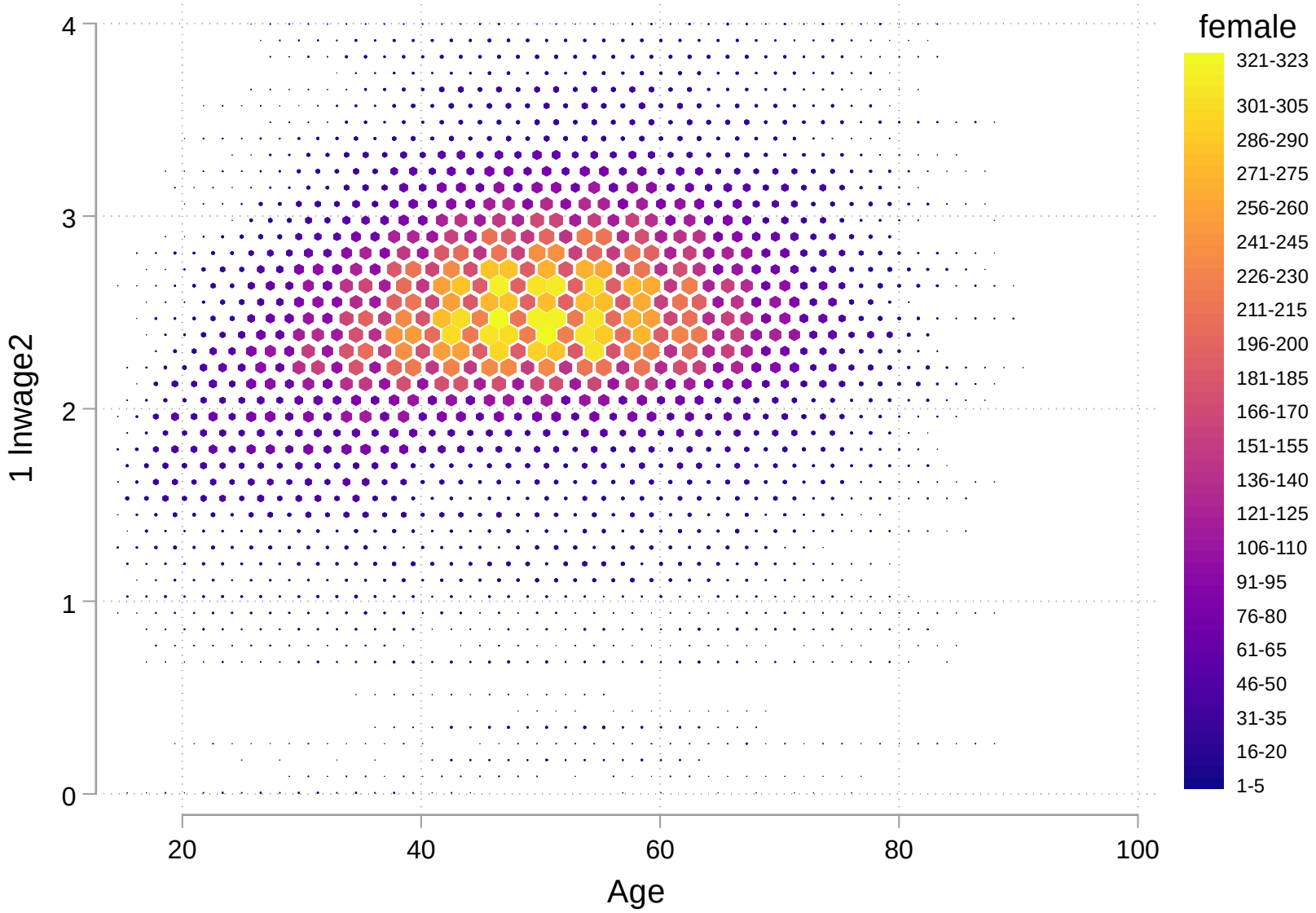
Sankey Plots - Transitions



Sankey Plots - Code

```
*==================================================================================*/
*                  3. Change structure of data                                     *
*==================================================================================*

// Reshape to wide
    use "$dtemp/quantlab.dta", clear

    keep time pid employment_arrangement                // relevant variables

    reshape wide employment_arrangement, i(pid) j(time)
    order pid* employment_arrangement*, first           // order variables

// check transitions between employment arrangements (e.g., at t1 & t22)
    preserve
    drop if employment_arrangement1==.
    drop if employment_arrangement2==.
    sankey_plot employment_arrangement1 employment_arrangement2, wide fillcolor(%50) gap(0.1)
noline xlabel(1 "LFS t1" 2 "LFS t2", nogrid)
    restore

    // add more variables if necessary
    merge 1:m pid using "$dtemp/quantlab.dta"
    keep if _merge==3
```

Heat Plots


female
321-323
301-305
286-290
271-275
256-260
241-245
226-230
211-215
196-200
181-185
166-170
151-155
136-140
121-125
106-110
91-95
76-80
61-65
46-50
31-35
16-20
1-5
1 lnwage2
0
1
2
3
4
20
40
60
80
100
Age

Heat Plots - Code

```
*==========================================================================*/
*                4. Heat Plots                                             *
*==========================================================================*
    // Heatplots
    ** reference year = average wages

    use "$dtemp/quantlab.dta", clear
    keep time pid lnwage2
    sum lnwage2, de

    reshape wide lnwage2, i(pid) j(time)
    order pid* lnwage2*, first          // order variables

    // add more variables if necessary
    merge 1:m pid using "$dtemp/quantlab.dta"
    keep if _merge==3

    // different representation by relative frequency (keep wide format)

    hexplot female lnwage21 hgage, statistic(count) color(plasma) cut(1(5)@max) keylabels(,
range(1)) size
```

That’s all for today!