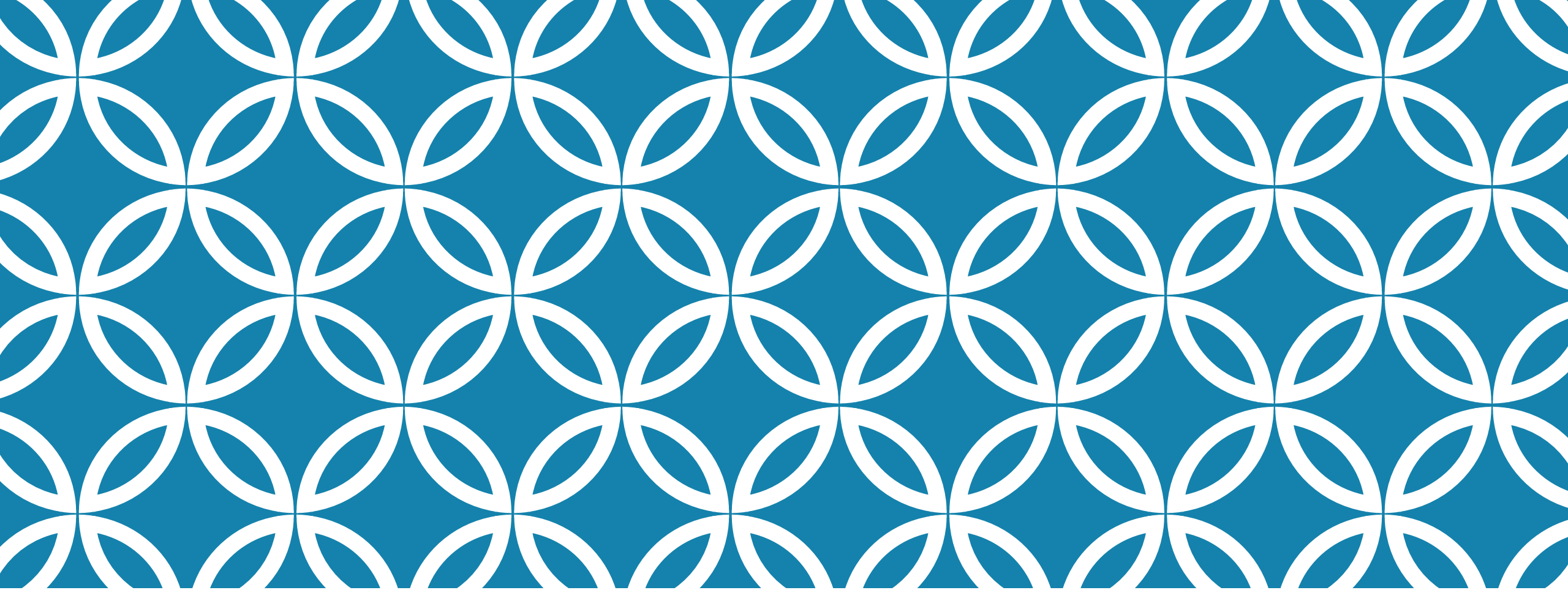


USING R MARKDOWN TO PRODUCE OFFICIAL STATISTICS

Aoife O'Neill
Client Statistics
Aoife.O'Neill@dwp.gsi.gov.uk
Twitter: stats_o

INCLUDING

The importance of official statistics

How they are produced

Automating production using R Markdown

What is R/R Studio/R Markdown

How I'm using R Markdown to automate official statistics

WHY ARE OFFICIAL STATISTICS SO IMPORTANT?

They’re statistics which are presented impartially.

Equal access for all.

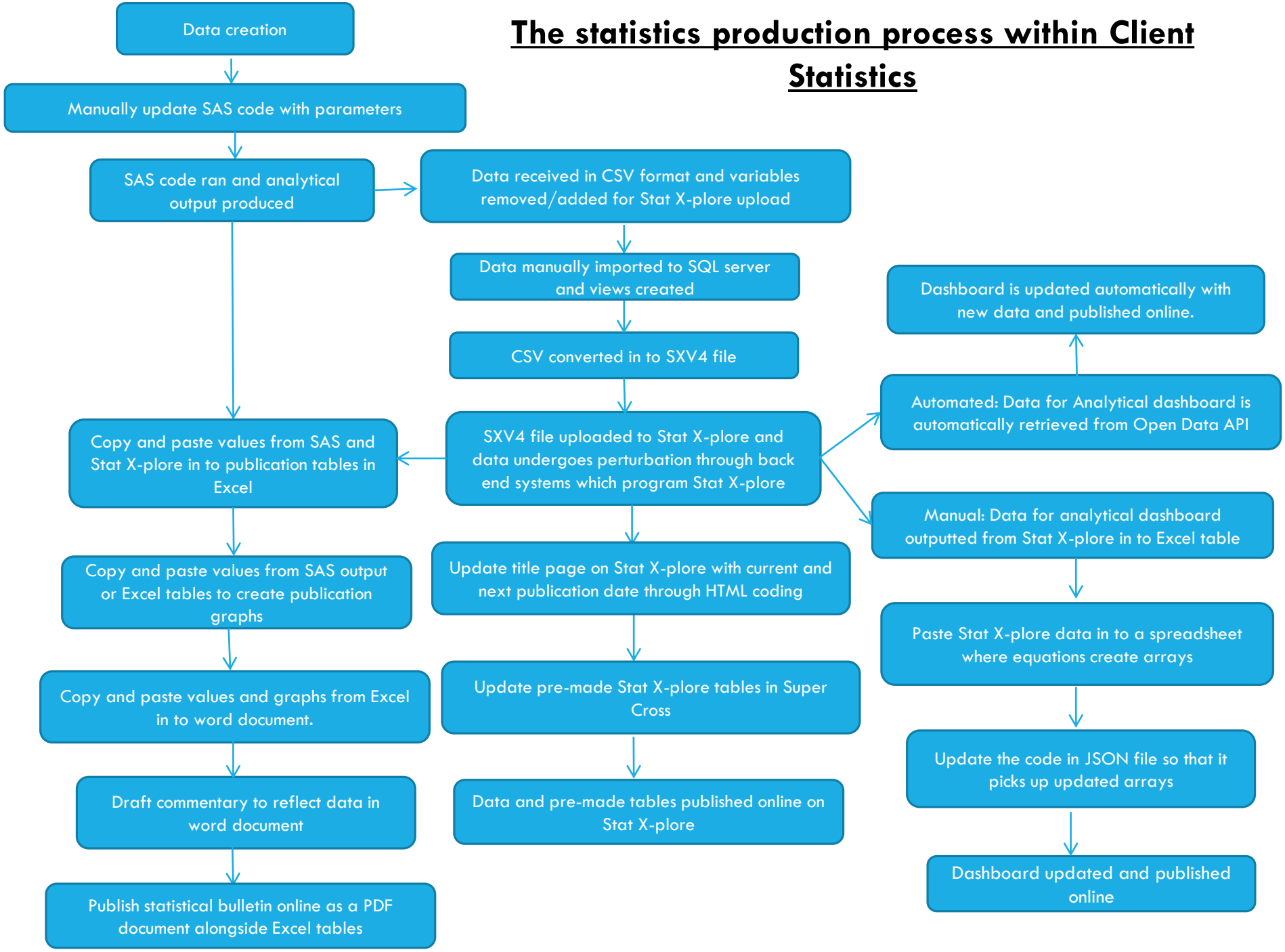
Academics, charities, local authorities and press office use these figures to inform their researches and strategise how they provide services.

Help to hold the government accountable.

“The single source of truth”

An essential public asset.

The statistics production process within Client Statistics



BUT THERE IS A BETTER WAY..

Department for Work & Pensions

National Insurance Numbers Allocated to Adult Overseas Nationals

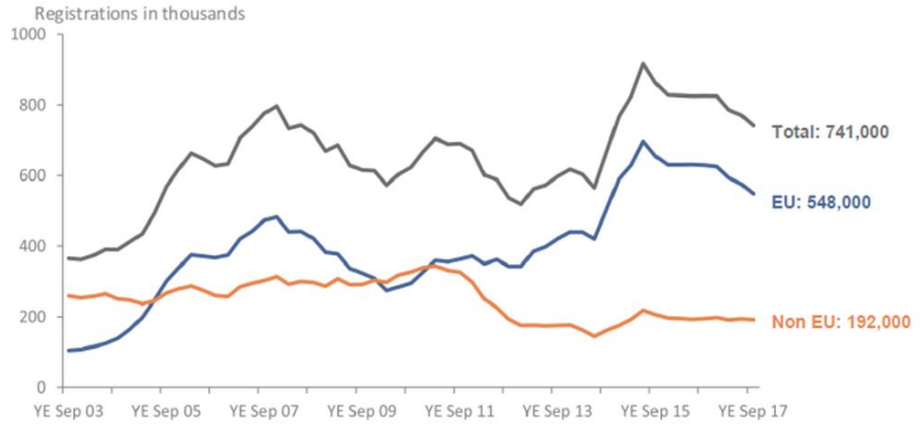
Registrations to September 2017

Quarterly | Published 30th November 2017 | United Kingdom | Official Statitics

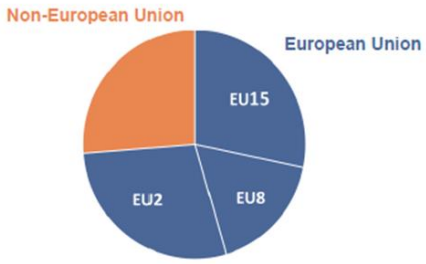
What do these statistics show? A National Insurance Number (NINo) is generally required by any adult overseas national looking to work or claim benefits / tax credits in the UK. As such they show a count – by individual nationality - of the number of NINo's registered to adult non-UK nationals each quarter, irrespective of the length of stay in the UK. The NINo statistics are not a count of the number of adult non-UK nationals currently residing in the UK

Main stories

741 thousand NINo's registered in the year to September 2017: a 10% decrease on the previous year

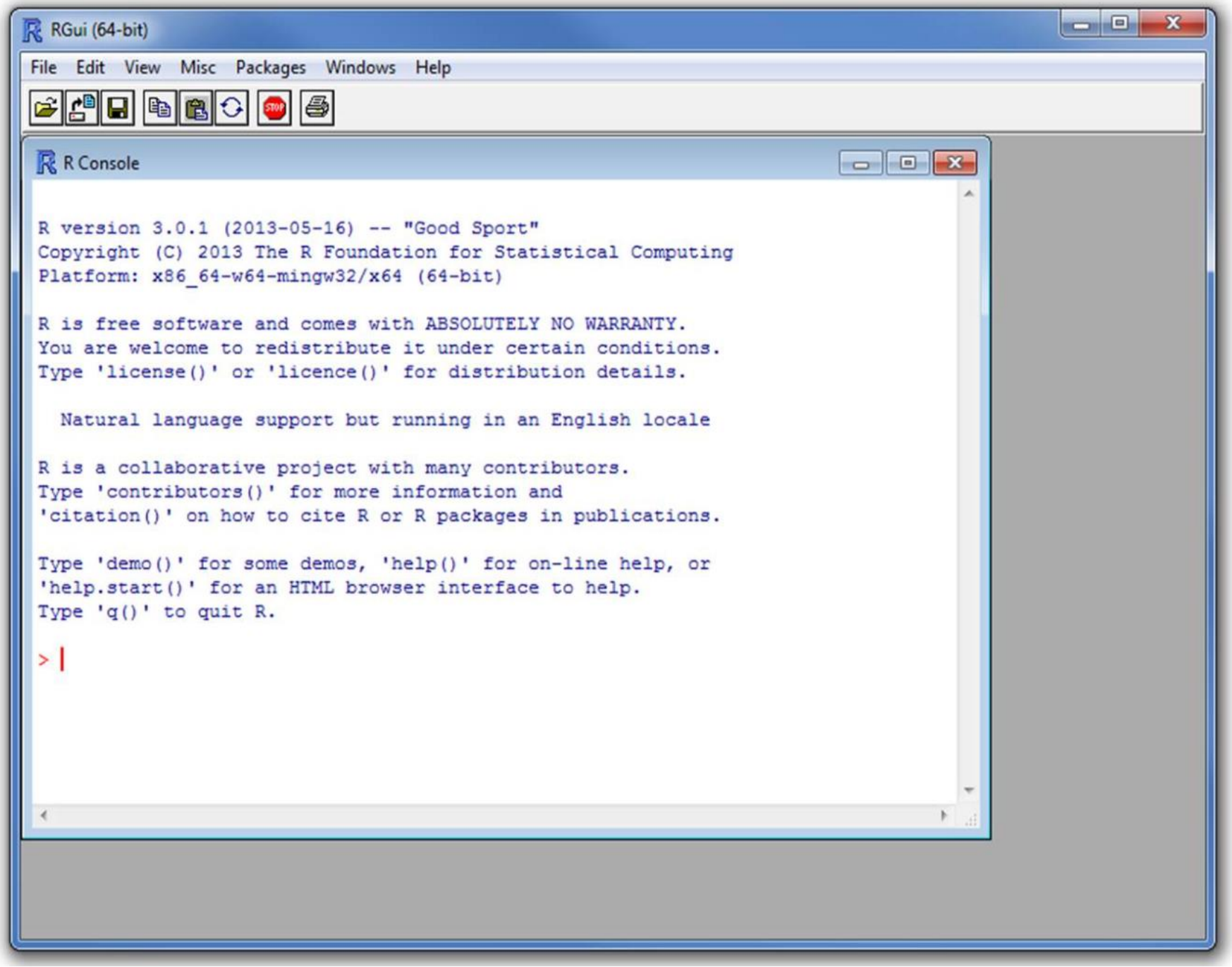


…of these, 74% are from within the EU




R
Studi


rmarkdown


RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (Non

Population_data_wrangling.R*

Source on Save | Run | Source

```
1  library(dplyr)
2  library(readr)
3  library(reshape2)
4  getwd()
5
6  ##before begining to read, check if any previous population figures have been revised!!
7
8
9  #### Local Authority #####
10 ##########################
11
12 ###before reading in, remove the commas in the numbers within excel! Or find the r code to do this
13 #year 2012
14 population_male_2012 <- read.csv("population_2012_m.csv", header = TRUE)
15 poulation_female_2012 <- read.csv("population_2012_f.csv", header = TRUE)
16 #year 2016
17 population_male_2013 <- read.csv("population_2013_m.csv", header = TRUE)
18 poulation_female_2013 <- read.csv("population_2013_f.csv", header = TRUE)
19 #year 2014
20 population_male_2014 <- read.csv("population_2014_m.csv", header = TRUE)
21
```

10:27 (Untitled) R Script

Console Terminal

~/Population data/

```
> poulation_female_2013 <- poulation_female_2013[,-3]
> population_male_2014 <- population_male_2014[,-3]
> poulation_female_2014 <- poulation_female_2014[,-3]
> population_male_2015 <- population_male_2015[,-3]
> poulation_female_2015 <- poulation_female_2015[,-3]
> population_male_2016 <- population_male_2016[,-3]
> population_female_2016 <- population_female_2016[,-3]
> head(population_male)
Error in head(population_male) : object 'population_male' not found
> str(population_female)
Error in str(population_female) : object 'population_female' not found
> Male_melt_2012 <- population_male_2012 %>%
+                   melt(measure.vars = c(3:93))%>%
+                   mutate(Gender= "M")%>%
+                   mutate(Year = "2012")
> Male_melt_2013 <- population_male_2013 %>%
+                   melt(measure.vars = c(3:93))%>%
+                   mutate(Gender= "M")%>%
+                   mutate(Year = "2013")
> Male_melt_2014 <- population_male_2014 %>%
+                   melt(measure.vars = c(3:93))%>%
+                   mutate(Gender= "M")%>%
+                   mutate(Year = "2014")
```

Environment History Connections

Import Dataset | List

Global Environment

Female_melt_2012	40040 obs. of 6 variables
Female_melt_2013	40040 obs. of 6 variables
Female_melt_2014	40040 obs. of 6 variables
Female_melt_2015	40040 obs. of 6 variables
Female_melt_2016	40040 obs. of 6 variables
final_data	355810 obs. of 5 variables
Male_melt_2002	35581 obs. of 5 variables
Male_melt_2003	35581 obs. of 5 variables
Male_melt_2004	35581 obs. of 5 variables
Male_melt_2005	35581 obs. of 5 variables
Male_melt_2006	35581 obs. of 5 variables
Male_melt_2007	35581 obs. of 5 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > Population data

Name	Size	Modified
..		
data.csv	8.5 MB	Apr 9, 2018, 4:06 PM
data_part1.csv	2.6 MB	Apr 9, 2018, 4:14 PM
data_part2.csv	5.9 MB	Apr 9, 2018, 4:14 PM
old_data.csv	3.3 MB	Jun 14, 2018, 2:24 PM
ParlC		
Population data.Rproj	217 B	Apr 12, 2018, 11:35 AM
population_2012_f.csv	185.4 KB	Apr 9, 2018, 3:21 PM
population_2012_m.csv	184.3 KB	Apr 9, 2018, 3:22 PM
population_2013_f.csv	185.6 KB	Apr 9, 2018, 3:22 PM
population_2013_m.csv	184.5 KB	Apr 9, 2018, 3:23 PM
population_2014_f.csv	185.7 KB	Apr 9, 2018, 3:23 PM
population_2014_m.csv	184.7 KB	Apr 9, 2018, 3:24 PM
population_2015_f.csv	185.9 KB	Apr 9, 2018, 3:25 PM
population_2015_m.csv	184.8 KB	Apr 9, 2018, 3:25 PM
population_2016_f.csv	186 KB	Mar 20, 2018, 2:44 PM
population_2016_m.csv	185 KB	Mar 20, 2018, 2:45 PM
Population_data_wrangling.R	6.6 KB	Jun 12, 2018, 2:25 PM
RAW_data.csv	3.4 MB	Jun 14, 2018, 2:48 PM
Scottish data zones		

File Edit Code View Plots Session Build Debug Profile Tools Help

u10096444

Go to file/function | Addins | Project: (Non

NiNo script.R | NiNo document.Rmd | NiNo_document.html

Knit | Insert | Run

```
555
556     + EU12 `r filter (EU2_data, Yr.to.Sep == "YE Sep 17") %>% select (change, Percent)`% to `r filter (EU2_data,
        Yr.to.Sep == "YE Sep 17") %>% select (reg_th)`,000
557
558 * Registrations to EU8 nationals now account now account for 23% of all EU registrations, this is down from 74% in
    2007.
559
560 Note: There was a sharp rise in NINo registrations in 2014 following the removal of restrictions on work for EU2
    citizens. Some of this is reflecting individuals who were residing in the UK prior 2014 and would only have been able
    to register for a NINo after that date (see arrival to registration analysis on EU2 nationals).
561 </div>
562
563 </div>
564
565 <div class ="row">
566 <div class="col-md-7"><!--Column left -->
567 <font size ="4">**Outside of the EU, South Asia has the most registrations** </font>
568 <img src="/scratch/working/clientstats/output/graph_4.png" alt = "A bar chart which shows the number of national
    insurance numbers allocated to nationals from world areas for the year ending September 2017>
569
570 </div>
```

559:1 # **What you need to know ** | R Markdown

Console | Terminal | R Markdown

.../output/NiNo document.Rmd

```
  |.....................................................          |  80%
label: Graphs 5 and 6 (with options)
List of 1
 $ echo: logi FALSE

  |.....................................................          |  87%
  ordinary text without R code

  |.....................................................          |  93%
label: NiNO_table2 (with options)
List of 2
 $ echo   : logi FALSE
 $ results: chr "asis"

  |.....................................................          | 100%
  ordinary text without R code
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
all_nino	16 obs. of 3 variables
all_ninos_time_...	16 obs. of 5 variables
data	9829677 obs. of 10 variables
data2	9829677 obs. of 17 variables
data3	8093344 obs. of 22 variables
geog_lookup	232184 obs. of 6 variables
graph_1	Large gtable (2 elements, 690.2 Kb)
graph_1_data	32 obs. of 4 variables
graph_2	List of 9
graph_2_data	4 obs. of 2 variables
nationality_loo...	235 obs. of 6 variables
Percent	2 obs. of 5 variables
title.grob	List of 11
yr_lookup	75 obs. of 3 variables
Functions	

Files | Plots | Packages | Help | Viewer

New Folder | Upload | Delete | Rename | More

/ > scratch > working > clientstats

Name	Size	Modified
..		
.gitignore	54 B	Mar 13, 2018, 3:40 PM
.Rprofile	117 B	Mar 14, 2018, 8:58 AM
clientstats.Rproj	205 B	Apr 3, 2018, 10:00 AM
core.8816	131.4 GB	Apr 27, 2018, 3:26 PM
core.9598	4.9 GB	May 1, 2018, 10:01 AM
data		
figure		
NiNo_document.html	1.4 MB	Apr 18, 2018, 12:09 PM
output		
script		

R MARKDOWN- BEING USED TO REPLICATE A PUBLICATION

Main Stories

`r filter (all_ninos_time_series, Yr.to.Sep == "YE Sep 17") %>% select(reg_th)` thousand NINo's registered in the year to September 2017: a `r filter (all_ninos_time_series, Yr.to.Sep == "YE Sep 17") %>% select(Percent)` % `r filter (all_ninos_time_series, Yr.to.Sep == "YE Sep 17") %>% select (change)` on the previous year

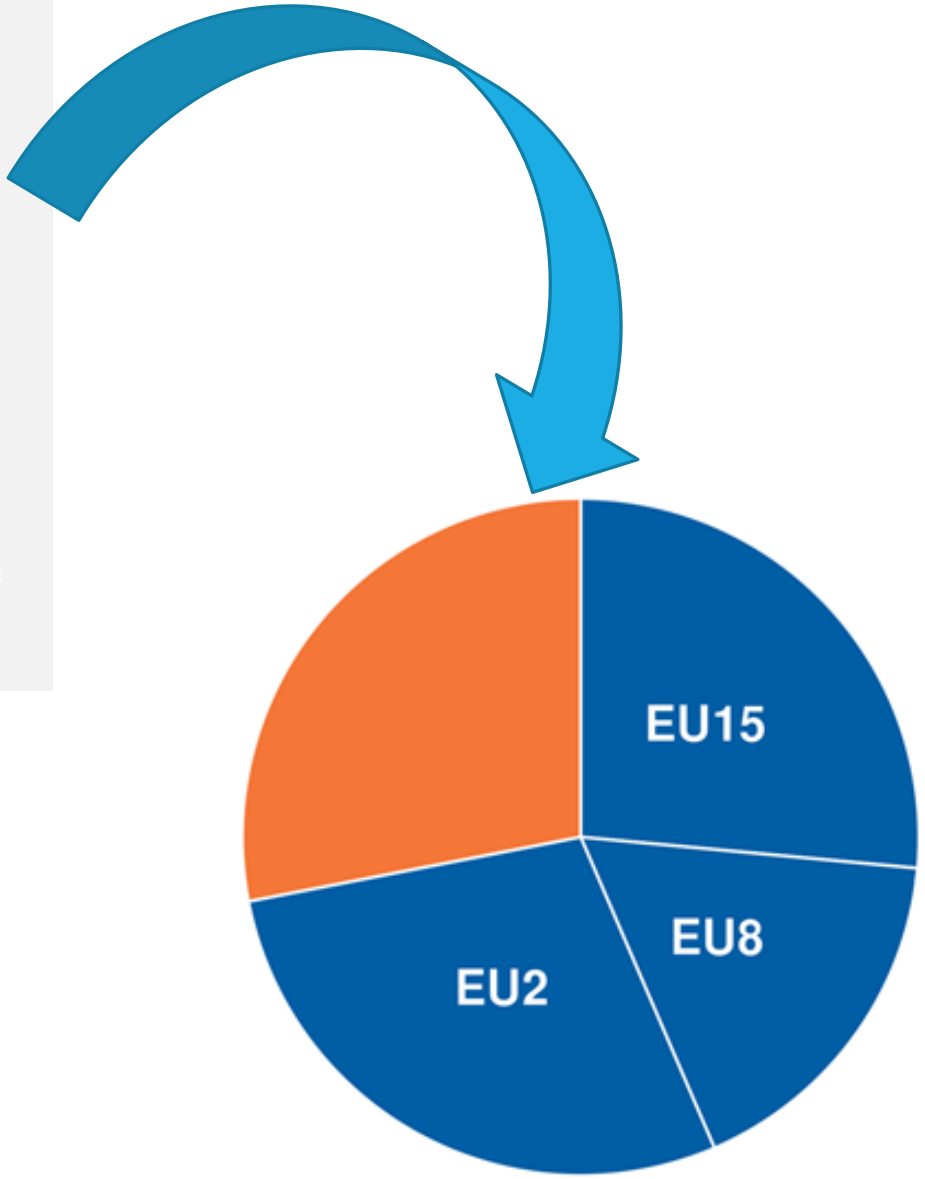
...of these, `r filter (Percent, eu_or_non == "EU") %>% select (Percent)` % are from within the EU

741 thousand NINo's registered in the year to September 2017: a 10 % decrease on the previous year

...of these, 74 % are from within the EU

```r
## GRAPH 2 ##

graph_2 <- graph_2_data %>%
  ggplot+
  (aes
   (x="", y= n, fill=eu_or_non))+
  geom_bar(width = 1, stat = "identity", colour = "white")+
  geom_text(aes(label = c(
    "",
    "EU2",
    "EU8",
    "EU15"
  )), position = position_stack(vjust = 0.5),
  fontface = "bold",
  colour = "white",
  size = 6
  )+
  coord_polar("y", start=0) +
  theme_gov(base_size = 12, base_colour = "white", axes = "xy")+
  scale_fill_manual(
    values = gov_cols[c("Non-European Union" = "orange", "EU15" = "govuk_blue", "EU2" =
"govuk_blue", "EU8" = "govuk_blue" )] %>% unname
  )
```

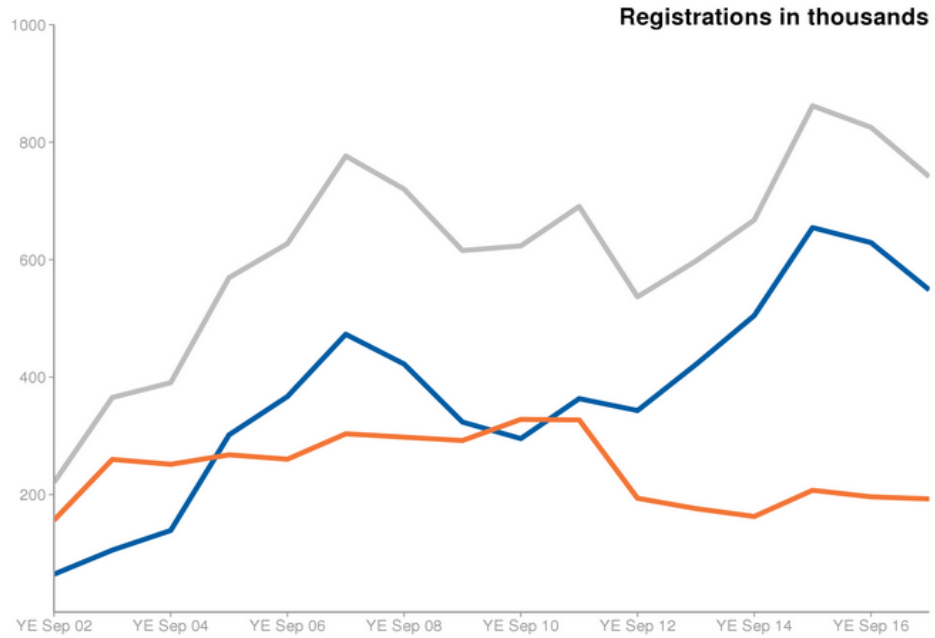


AND HTML CAN BE USED TO FORMAT…

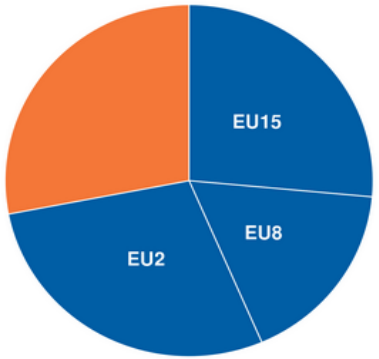
What do these statistics show? A National Insurance Number (NINo) is generally required by any adult overseas national looking to work or claim benefits / tax credits in the UK. As such they show a count - by individual nationality - of the number of NINo's registered to adult non-UK nationals each quarter, irrespective of the length of stay in the UK. The NINo statistics are not a count of the number of adult non-UK nationals currently residing in the UK.

Main stories

741 thousand NINo's registered in the year to September 2017: a 10 % decrease on the previous year



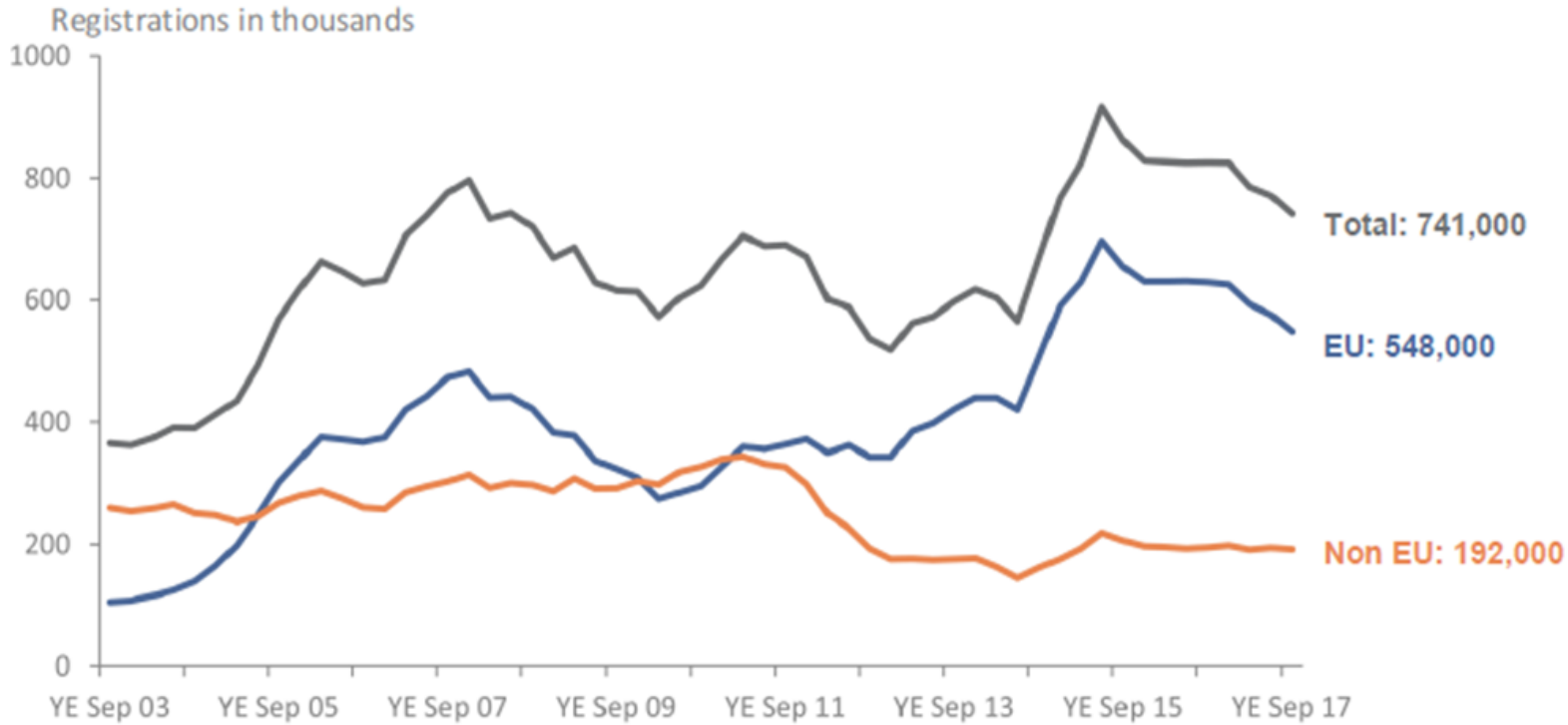
…of these, 74 % are from within the EU



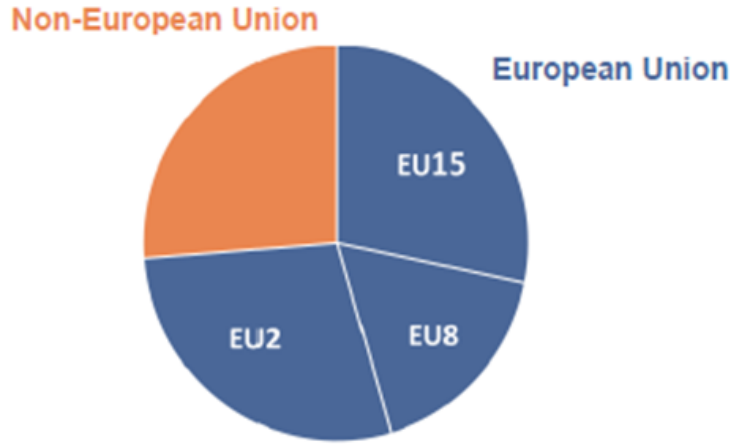
What do these statistics show? A National Insurance Number (NINo) is generally required by any adult overseas national looking to work or claim benefits / tax credits in the UK. As such they show a count – by individual nationality - of the number of NINo's registered to adult non-UK nationals each quarter, irrespective of the length of stay in the UK. The NINo statistics are not a count of the number of adult non-UK nationals currently residing in the UK

Main stories

741 thousand NINo's registered in the year to September 2017: a 10% decrease on the previous year



…of these, 74% are from within the EU



SUMMARY

The importance of official statistics

How they are produced

Automating production using R Markdown

What is R/R Studio/R Markdown

How I'm using R Markdown to automate official statistics

ANY QUESTIONS…?

AOIFE'S FAVOURITE LEARNING RESOURCES:

- R for Data Science e-book: http://r4ds.had.co.nz/index.html
- Data camp: https://www.datacamp.com/home
- R package which makes graphs in the gov.uk style: https://github.com/ukgovdatascience/govstyle
- RAP e-book: https://ukgovdatascience.github.io/rap_companion/
- Blog on RAP: https://dataingovernment.blog.gov.uk/2017/11/27/transforming-the-process-of-producing-official-statistics/
- RAP online learning: https://www.udemy.com/reproducible-analytical-pipelines/
- Advanced R: http://adv-r.had.co.nz/
- Happy Git with R: http://happygitwithr.com/