Social Care Survey Open Data

Descriptive Stats

David Henderson

2 March 2017

# Introduction

In part 1 I imported and wrangled the the Social Care Survey open data made available by the Scottish Government. I also assessed the missingness of each of the variables across the 3 years of data. This part of the analysis is to describe each variable in more detail.

# Conclusions

This document is in narrative format and documents the step-by-step analysis of the open datasets. As you we saw in the Import and Wrangle analysis,it was evident that most of the missing data was found in the cohort of individuals that do not receive home care (i.e. hours of carer attending the individual's home) but *only* recieve meals or other services (telecare, laundry etc). The analysis in this document, therefore, concentrates on individuals that *did* receive homecare. It is a long document with lots of graphs. For this reason I have summarised the main findings here.

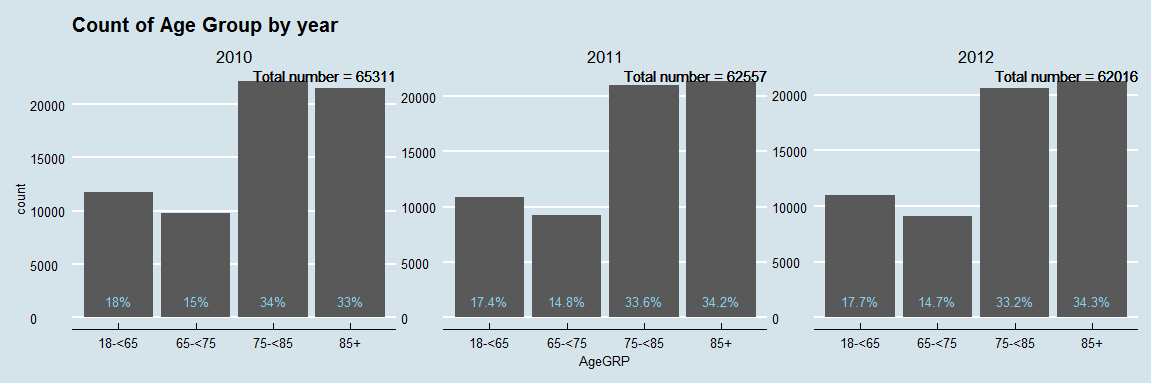
* The Home care cohort has very little missing data, however some variables are of questionable quality.
  + Client Group levels (Dementia/Physical Disability/Infirmity due to Age etc.) are completed very differently across LAs.
  + Meals data are difficult to distinguish between "No Meals" and missing data (NA).
  + The open data has pre-banded, uneven levels of hours of home care making it difficult to identify good or better bandings. The data is completed very well so "real" data when we get it should allow sensitivity analysis of bandings.
  + Living arrangements data has the most missingness in this cohort - levels reported vary significantly across LAs making quality questionable.
  + The Housing Type variable is of poor quality and unreliable.
  + The optional (i.e. optional for LAs to return to the SG) variables; Laundry, Shopping, and housing support show big variations is levels of delivery across LAs. Whether this variation is due to service provision or quality of returned data is difficult to tell - again a little unreliable.
* The Staffing variable (Single or Multi-staffed) appears to be completed well and could possibly act as a proxy for increased need. There is a wide variation across LAs for levels of multi-staffed indiviudals.
* Telecare data is only collected in 2011 and 2012 and appears of good quality. A large proportion of those receiving home care also have a community alarm - almost half of all individuals. There is wide variation across LAs in provision of telecare so will be interesting to see any differences in outcomes.
* Significantly more females than males receive Home care services. I think greater than the proportion of females in the general population although I would need to check this out. Does this influence outcomes? I think worth investigating further.

# Descriptive Statistics

## Age

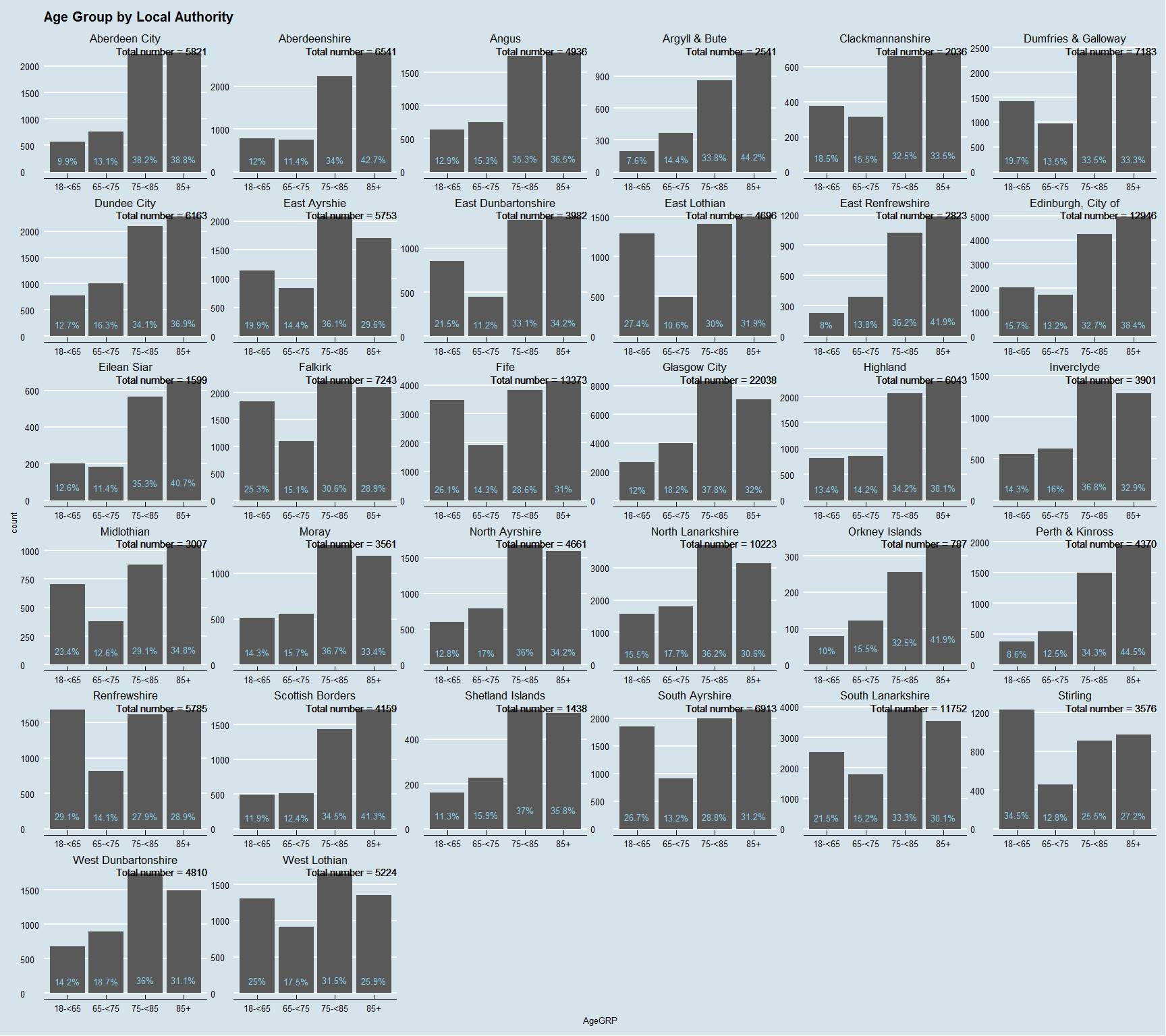
I'll start with age.

### National distribution

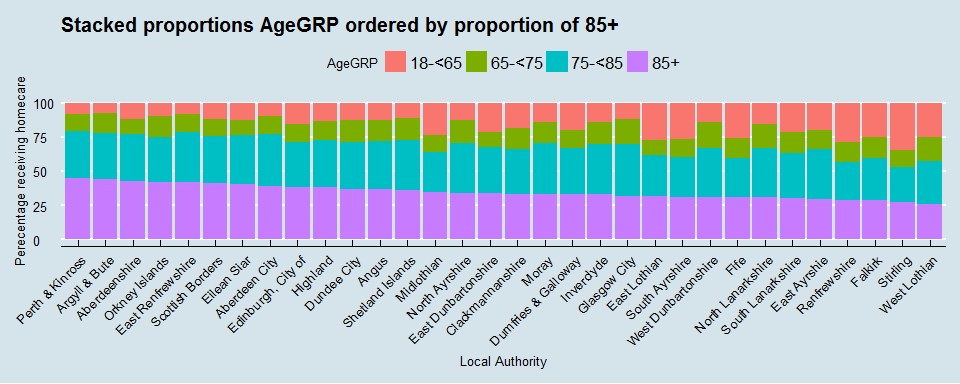
First of all I'll visualise the distribution home care received by age across the country. The open data release pre-banded this variable into 5 categories. 

Here we can see fairly similar numbers and distributions across years. 65311 people received home care in 2010, 62557 in 2011, and 62016 in 2012. so a slight decrease over time.

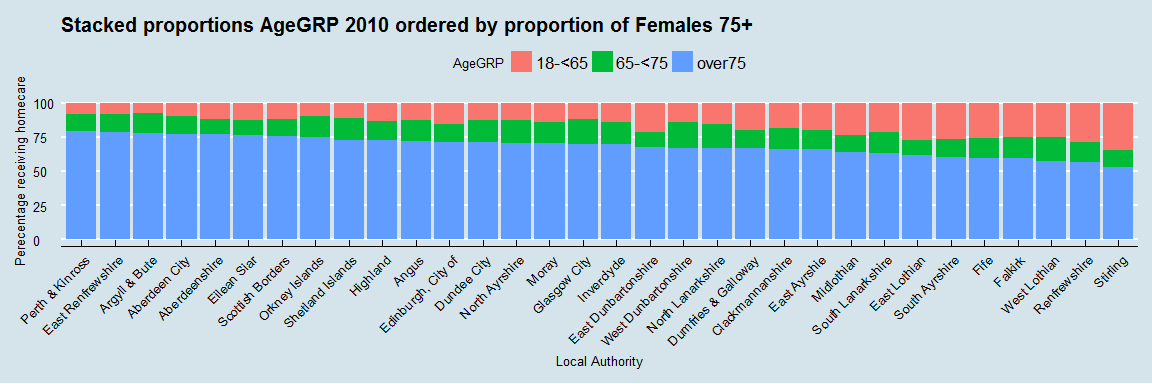
### Distribution by Local Authority



In the above plot we can see large variations in the proportions of Age groups receiving home care. Whilst this plot is handy to see exact figures, it is diffcult to make comparisons - a stacked bar plot should be better:-



So a lot easier to make comparisons here, I wonder if it would be even easier if I collapsed the 75-85 and 85+ age groups together?

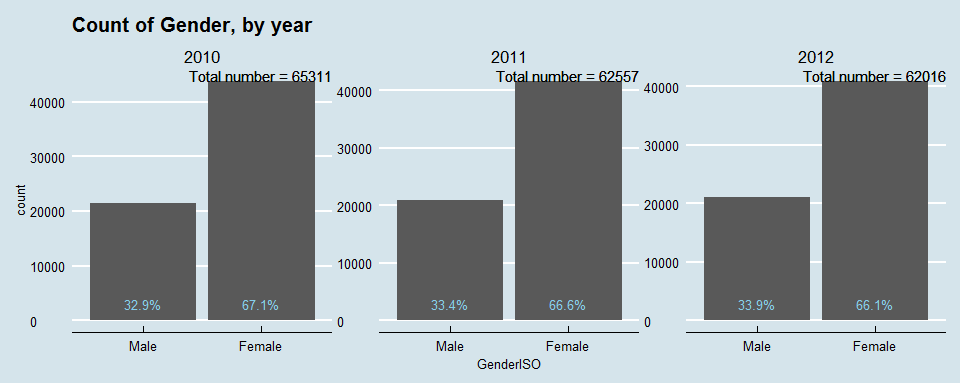


So certainly a lot easier (I think) when we compare over 75s with the other 2 populations. I've ranked this plot by proportion of Females over the age 75.

## Gender

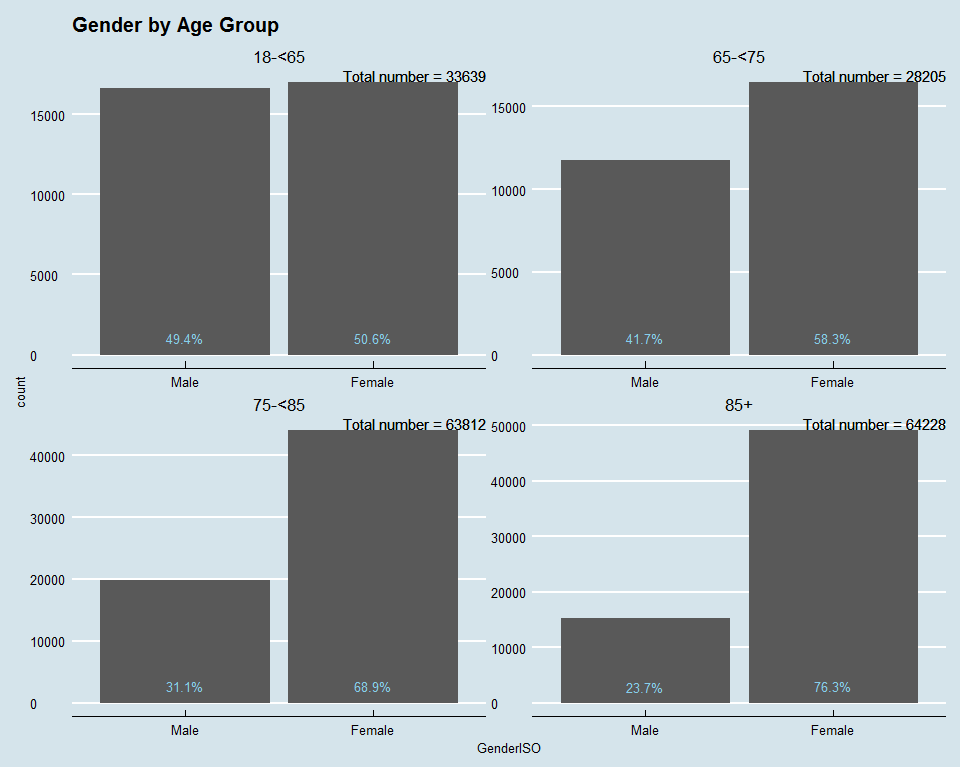
Next up to to investigate is Gender.

### National Distribution



So fairly similar distributions across years and significantly more females recieving home care. Why? Because they live longer?

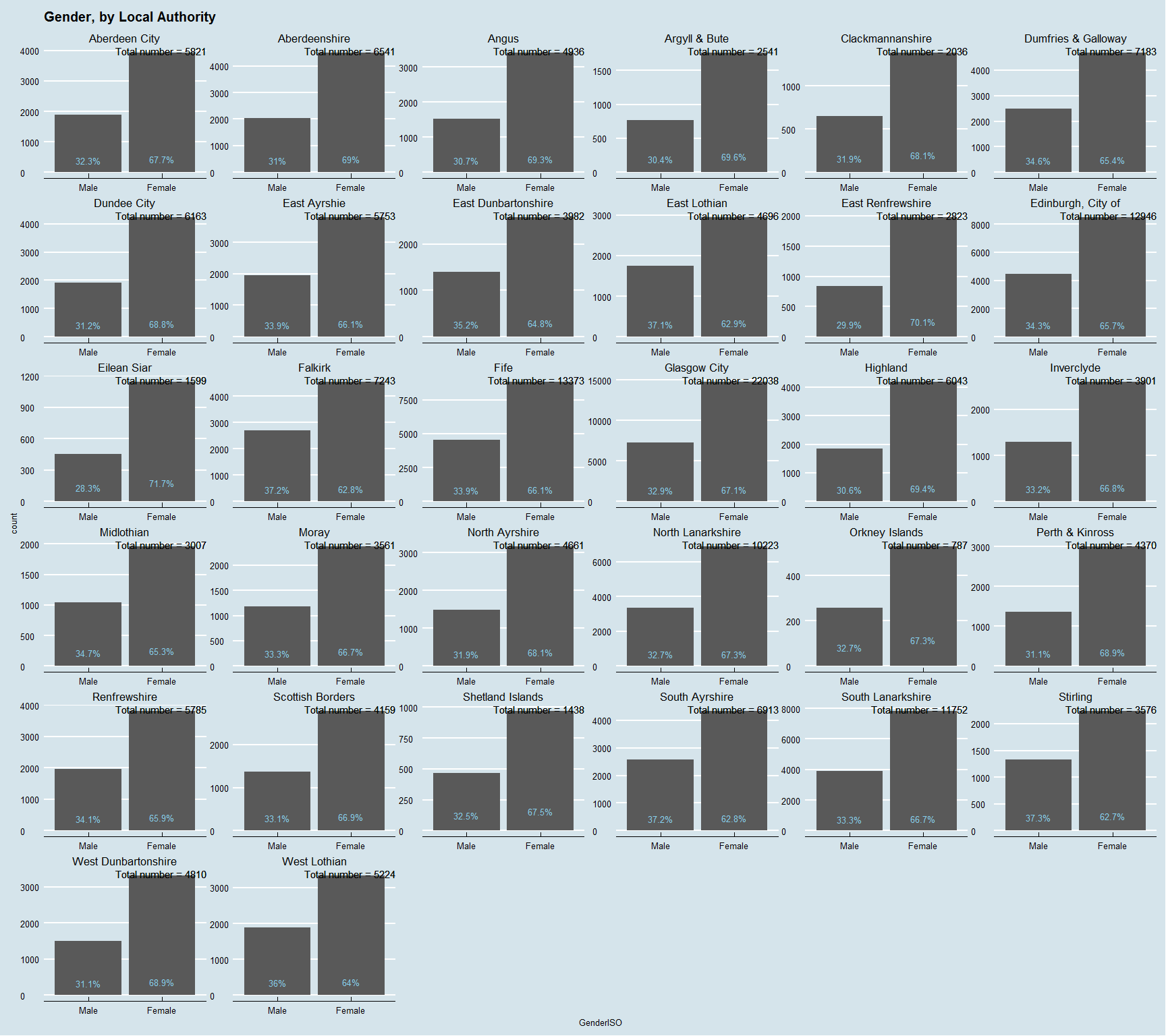
### Distribution by Age



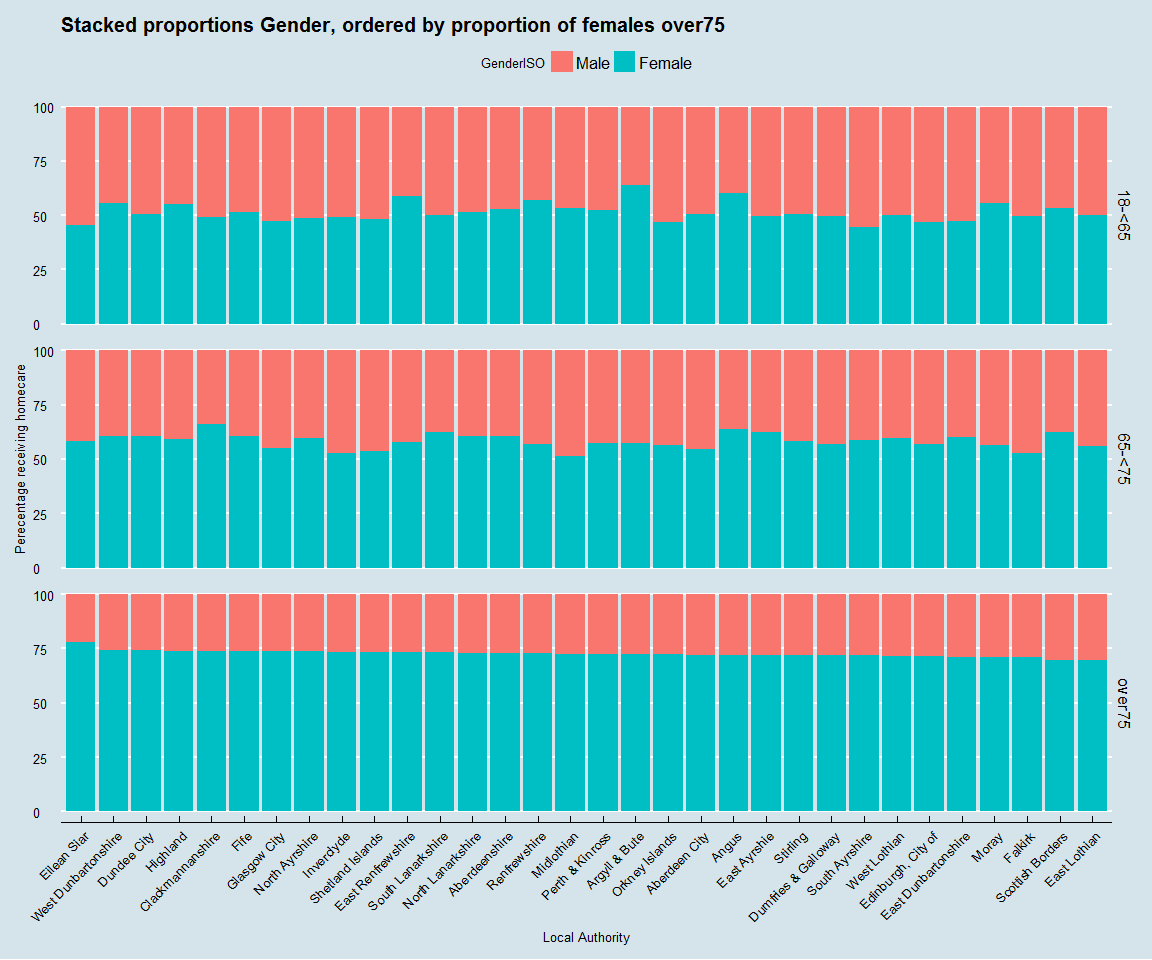
So, pretty easy to see the the gradual increase in proportions of females through the age groups. There is also a disparity in absolute numbers - There are big increases in the size of age groups with females accounting for the majority of this. Comparing absoulte numbers we see there are less men in the 65-75 group compared to 18-65 and also less men in 85+ compared to 75-85. For females these absolute numbers increases.So I guess the question is - do females make up e.g. 76% of the over 85 population? I don't think so (need to check). If not - why do they receive a higher proportion of social care?

### Distribution by LA

I don't expect there to be much variation across LAs but I should check.



Actually, there is a little more than I thought - between 62% and 70% female. I'll stack this again and add age as a facet.



So a little variation across LAs when comparing Gender and Age Group but no major outliers I don't think. I've arranged this by proportion of females over the Age of 75 receiving home care. No obvious pattern between rural/urban or rich/poor (maybe some more poorer council areas over to the left, but richer ones like Aberdeen of Shetland in the middle).

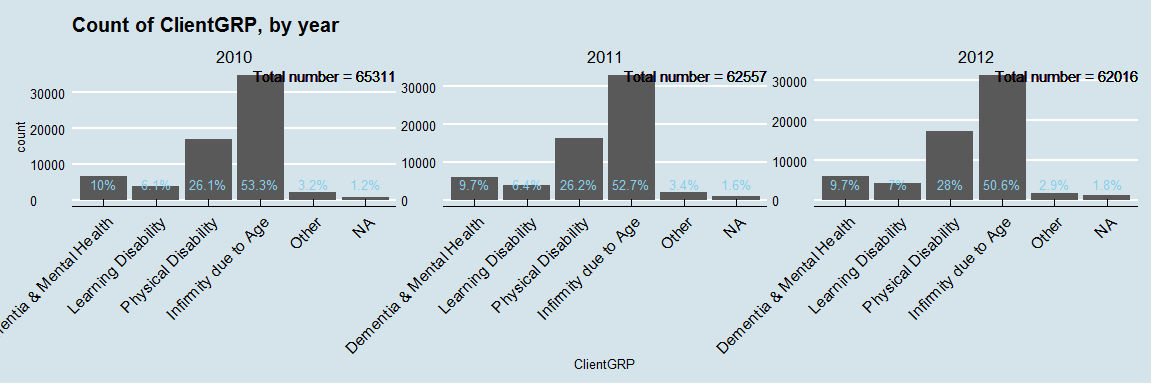
## Client Group

Next up in the dataframe is Client Group. Using the metadata file published alongside the open data we can see this variable has been pre-grouped before release - presumably due to small numbers.

* Dementia and Mental Health are combined.
* "Other" is a combination of the Addiction, Palliative Care, Carers, and Other vulnerable groups designations. "Carers" itself is defined as home care provided to aid a family carer."Other vulnerable groups" includes; HIV/AIDs, Aquired Brain Injury, Homeless, and Women escaping domestic violence.
* The remainin groups are; Learning Disablility, Physical Disability, Problems arising from infirmity due to age.

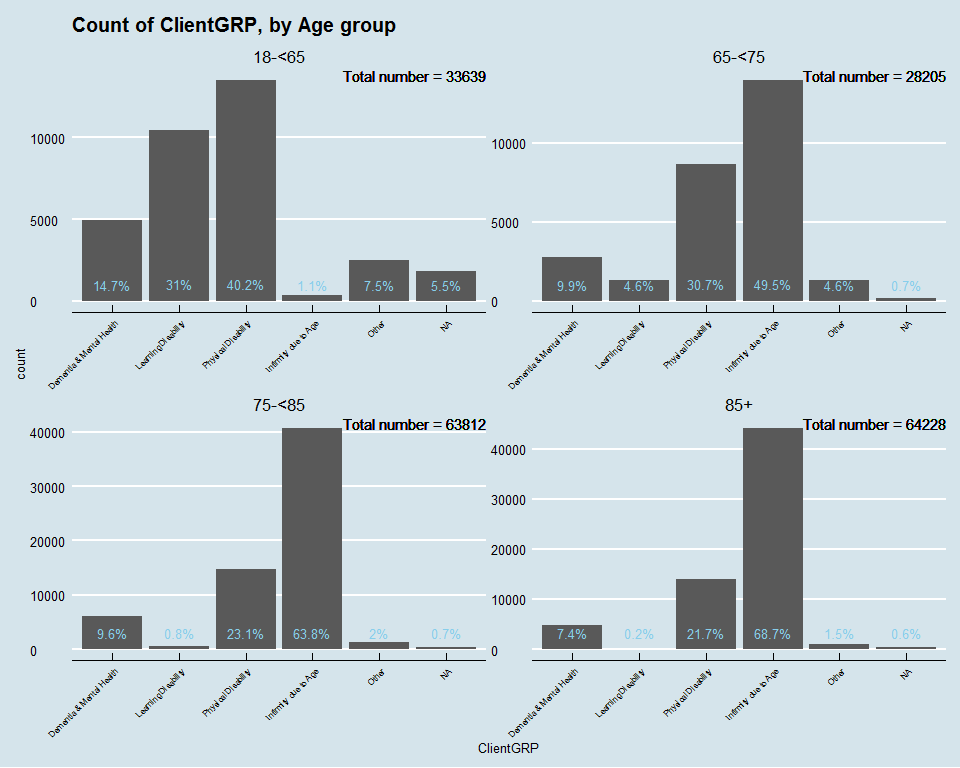
Published Social Care Survey reports identify the potential for poor classification with this variable. Dementia is known to be underecorded. Also Physical Disability and Infirmity are often misclassified and can often be interchangeable - can we see this in the data?

### National distribution



Here we can see similar distributions across the years. The highest proportion of clients are classified as "Infirmity due to Age", followed by"Physical Disability". Next up will be to see how these vary according to Age.

### Distribution by Age



Personally I would not have grouped Mental health and Dementia together. I think it would be fair to say these are likely to be 2 very different groups. As we can see the highest proportion for this group is in the 18-65 age bracket - where mental health is far more likely to be the reason. We know dementia is poorly recorded which explains the lower proportions of this bracket across the other age bands.

As expected, Learning disability has a much higher propotion in the youngest age band, with virtually no clients recorded over the age of 75 reflecting life expectancy for those with LD.

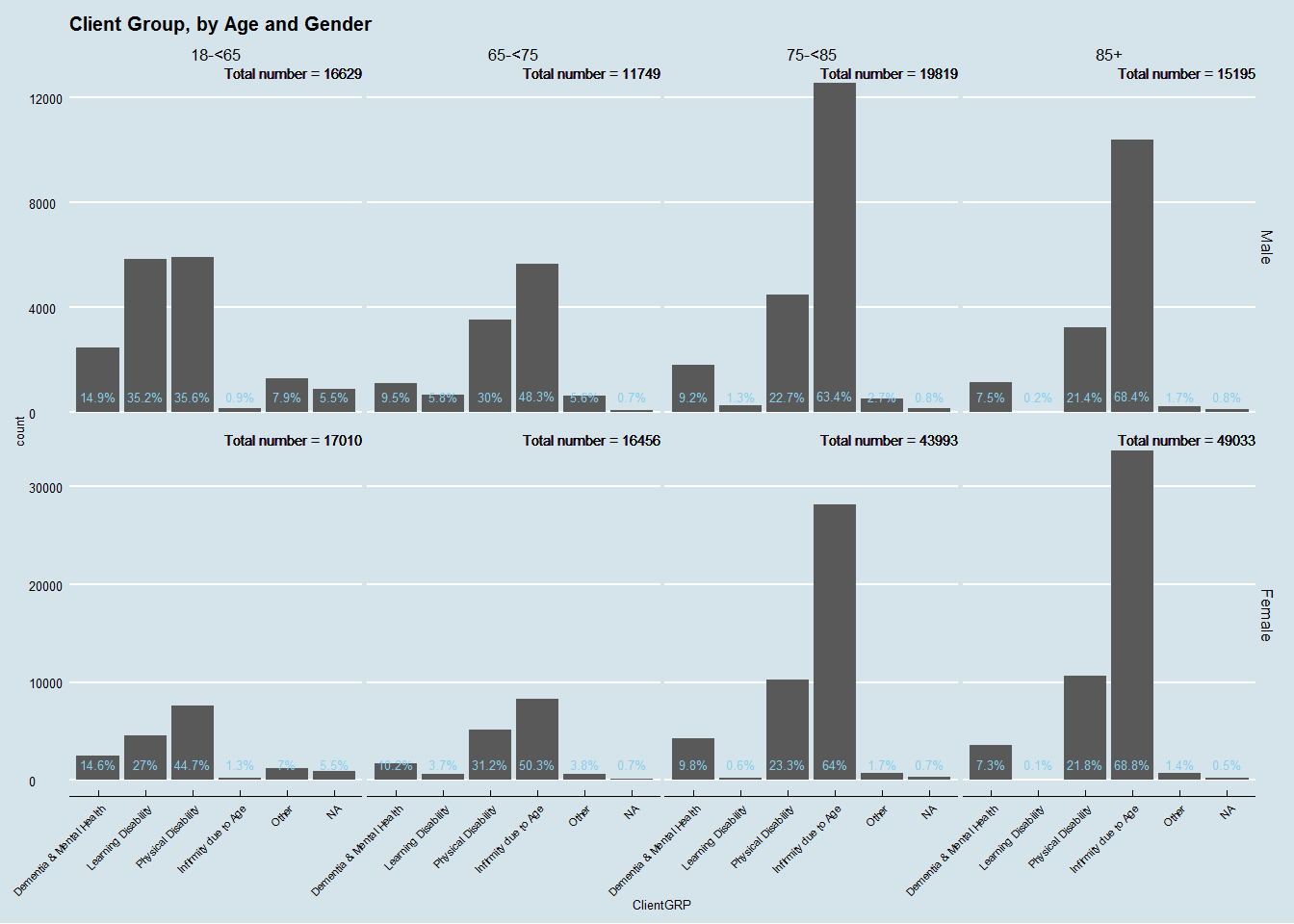
Physical Disability has a higher ratio in the 18-65 group also. Does this reflect thise with e.g. MS or Aquired Brain injury? Difficult to say.

Unsurprisingly, Infirmity due to age has increasing proportions through the Age bands. Other (Addictions, Palliative Care, Carers etc), has decresing propotions as age increases.

Almost all the missing data is in the 18-65 Age group. I'm not sure why this would be - perhaps confined to 1 LA??

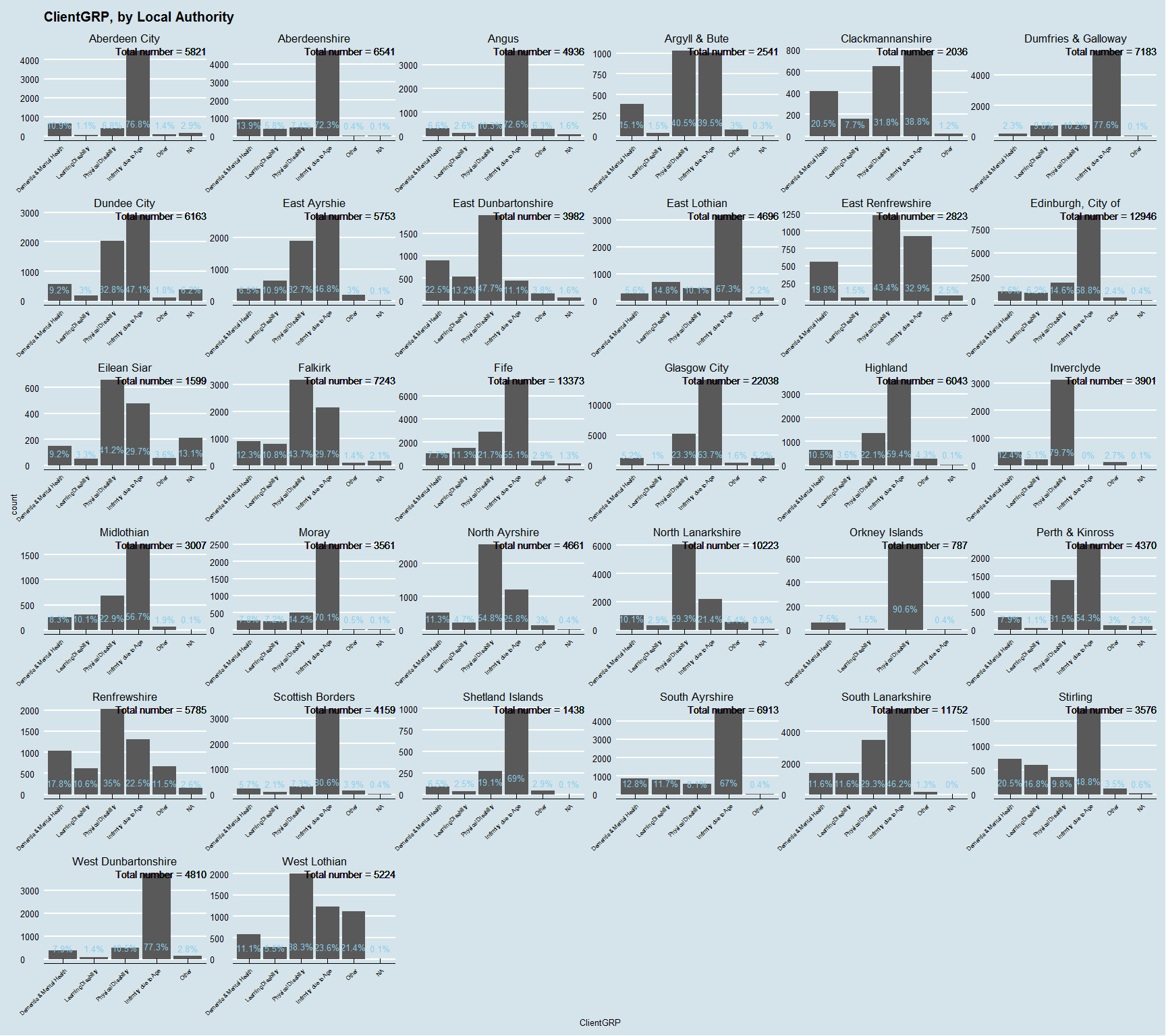
### Distirbution by Gender

Before I look at LA better check differences by Gender.



Higher proportions of Males for LD and higher proption females due to Infirmity. The latter reflects the age pattern. I'm not aware of higher prevelance LD in Males but that may very well be the case.

### Distribution by Local Authority.



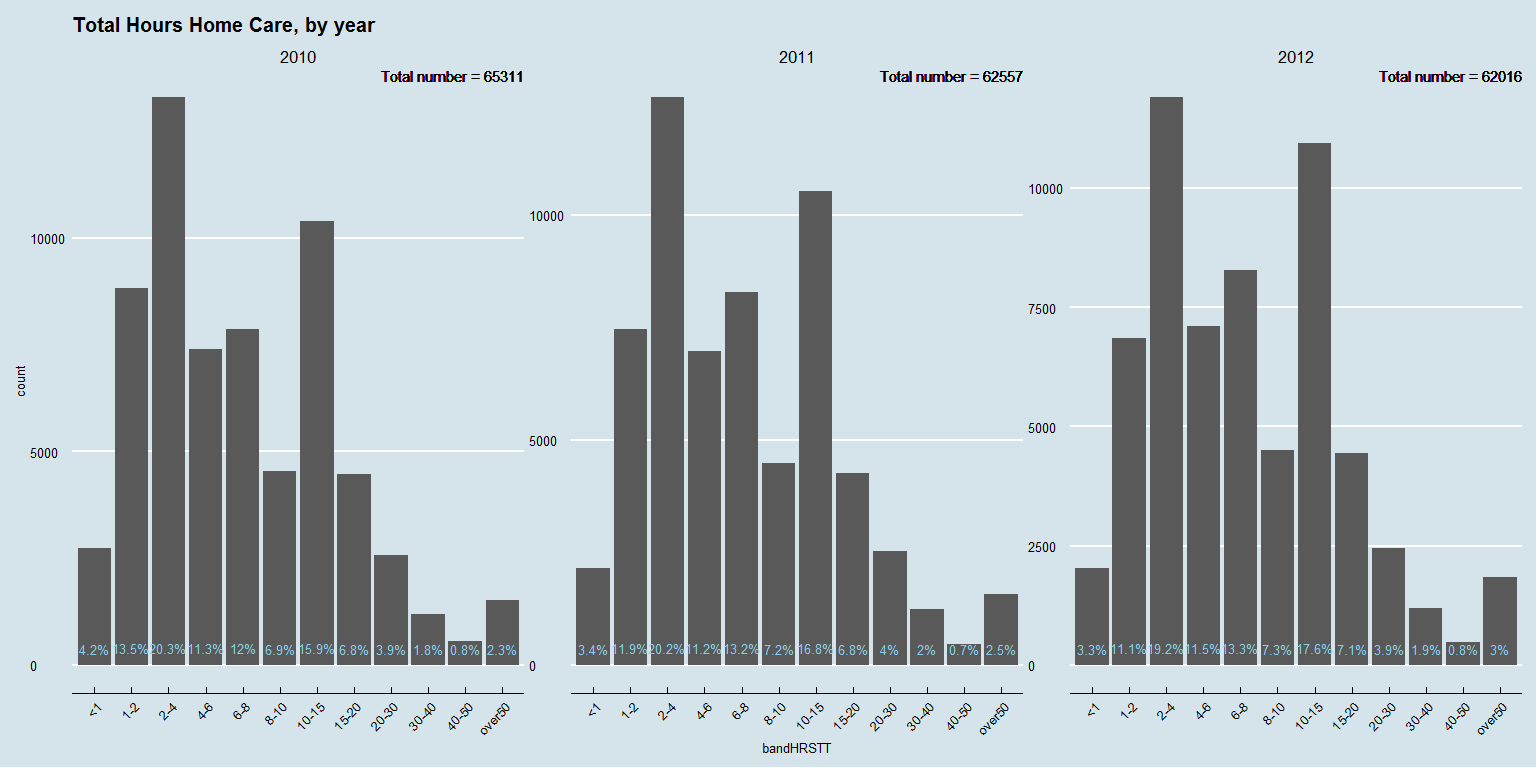
A bit of a messy plot here due to axis labels being big but anyhow we can see the huge variation across Local Authorities. This is almost certainly down to classification practcies at the local level. Some authorities have virtually no clients recorded as Infirm whereas others have 75% recorded as Infirm. Given this sort of variation I think it would be difficult to rely on this variable for any meaningful analysis.

## Home Care Hours

Variable of most interest! Again this variable was banded pre-publication of the data. If I'm honest they aren't the most helpful bandings - something more ordinal would have been better e.g. some bands describe a value between 2hrs of home care, others between 5hrs of home care ("2-4", "5-10"). The home care hours data is provided with 5 variables; number of Local Authority provided hours, Private provided hours, Voluntary Organisation provided hours, Personal Care Hours, and Total hours. There are some people who will receive a variety of these and I may turn to this as a sub-group later. For the moment I will concentrate on total figures.

### National Distribution

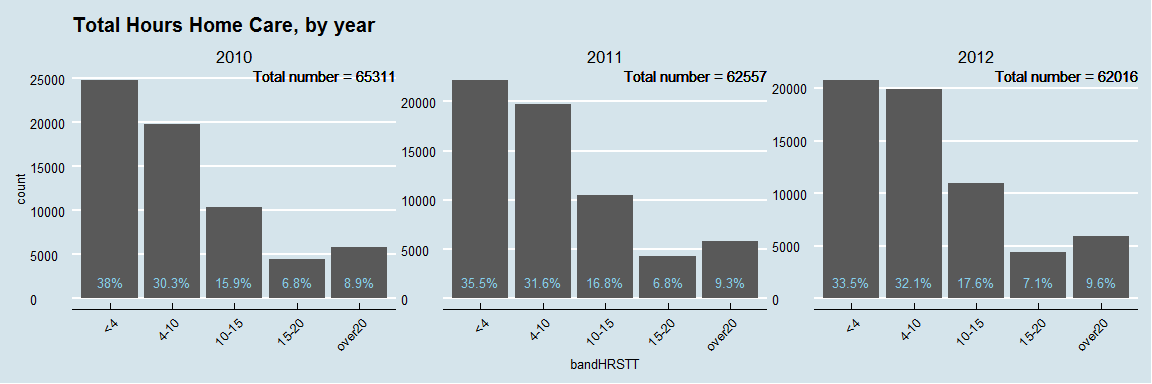
1st of all I will plot the bandHRSTT variable which is the cumulative total of all hours of care regardless of provider.



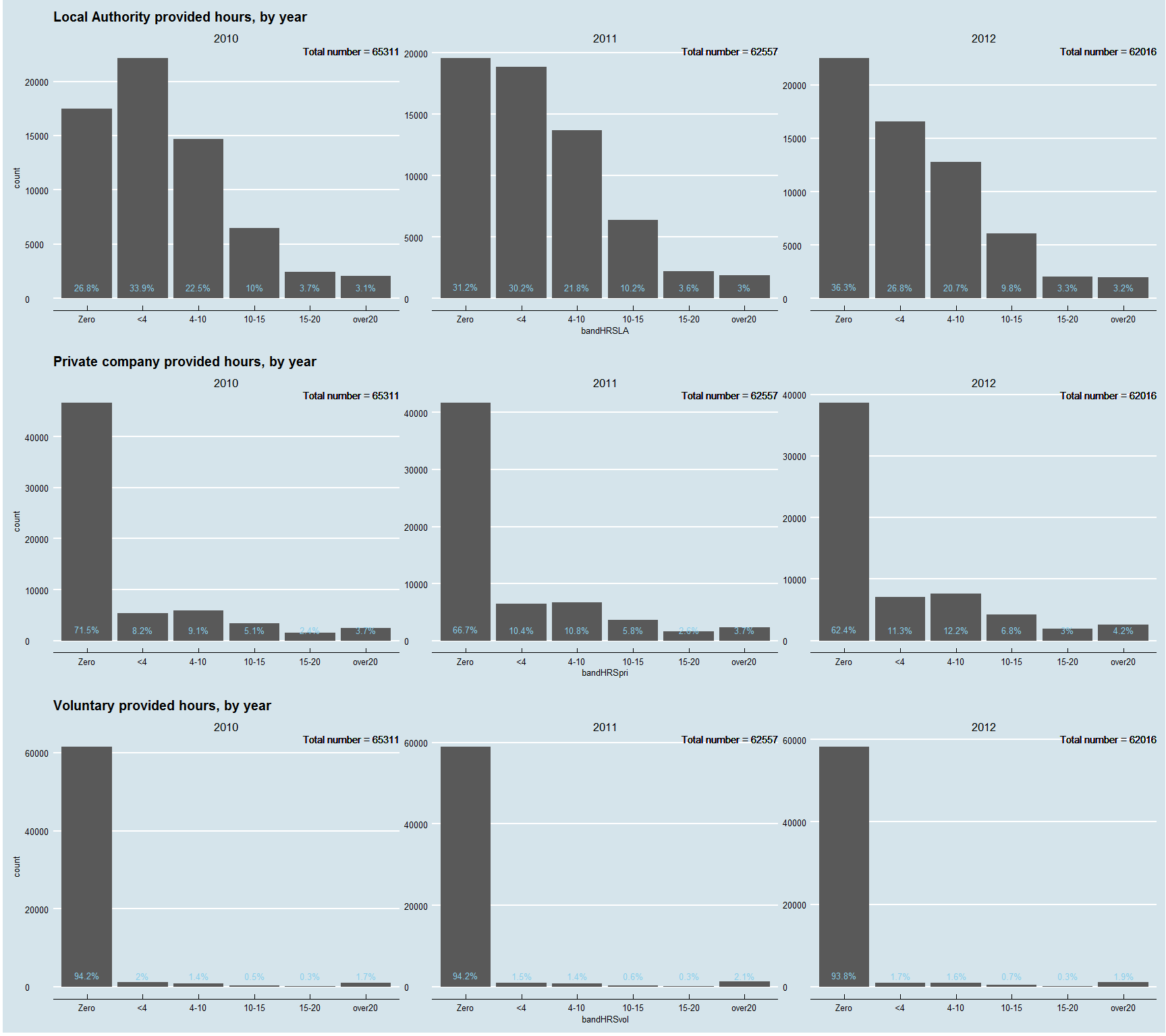
Difficult to compare. I am going to collapse some of these bands together for easier analysis. Before I looked at the data I had thought I'd like to bin the hours into 0-5, 5-10, 10-15, 15-20, and 20 plus. Because of the way this data is grouped I can't do that. For now I'll go for 0-4 and 4-10 then carry on. I will do this for all the bandHRS variables in the alternative, simplified, dataframe.

### National Distribution (2)

OK, I'll retry with the collapsed levels.



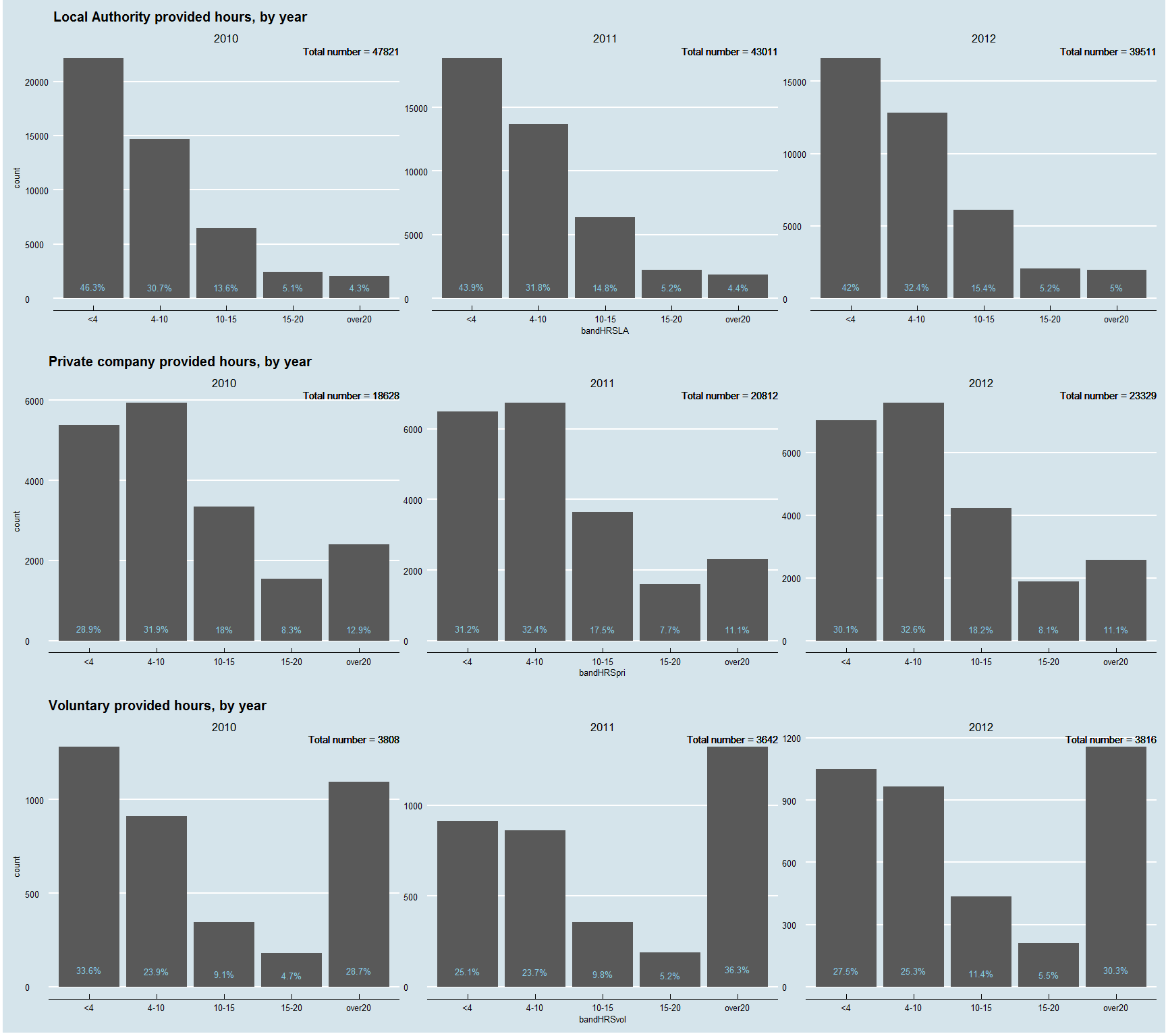
So, a little easier to visualise with collapsed bandings. Similar distributions over years. The Social Care Survey report suggests that private and voluntary organisations tend to pick up larger care packages whereas Local Authorities tend to concentrate on smaller packages of care. Can we visualise this and see if it has changed over time?



So there are a couple of errors here that I will get rid of but before I do there is some useful information that we will lose in the next plot.

In order to get better axis sizes (particularly in the voluntary plot) I will remove the "Zero" hours observations. What is interesting, before I do that, is in the LA plot at the top we see the percentage of "Zero" hours increasing over time. Essentially this is the proportion of Home care that the LA is purchasing from other organisations rather than providing the care in house. LAs are gradually farming more care out. The main beneficiary seems to be private organisations who have steadily decreasing proportions of "Zero" hours.I'm pretty sure LAs are still providing the majority of care - we'll see better on the next plot.

I'll replot.



Local Authorities still handle the majority of home care in-house, by quite a margin. However, the absolute numbers confirm that LAs are providing less in-house care and purchasing that care from private companies instead.

Comparing the proportions of hours of care delivered between LAs and private companies we can see that LAs do indeed have a higher proportion of very low (<4hr) care packages, and that private companies have higher proportions of larger (>10hr) packages.

Voluntary organisations provide a significantly higher porportion of very large (>20hr) packages with much smaller numbers (bearing in mind 590 observations with missing data for voluntary hours have been removed).

### Distribution by Age and Gender

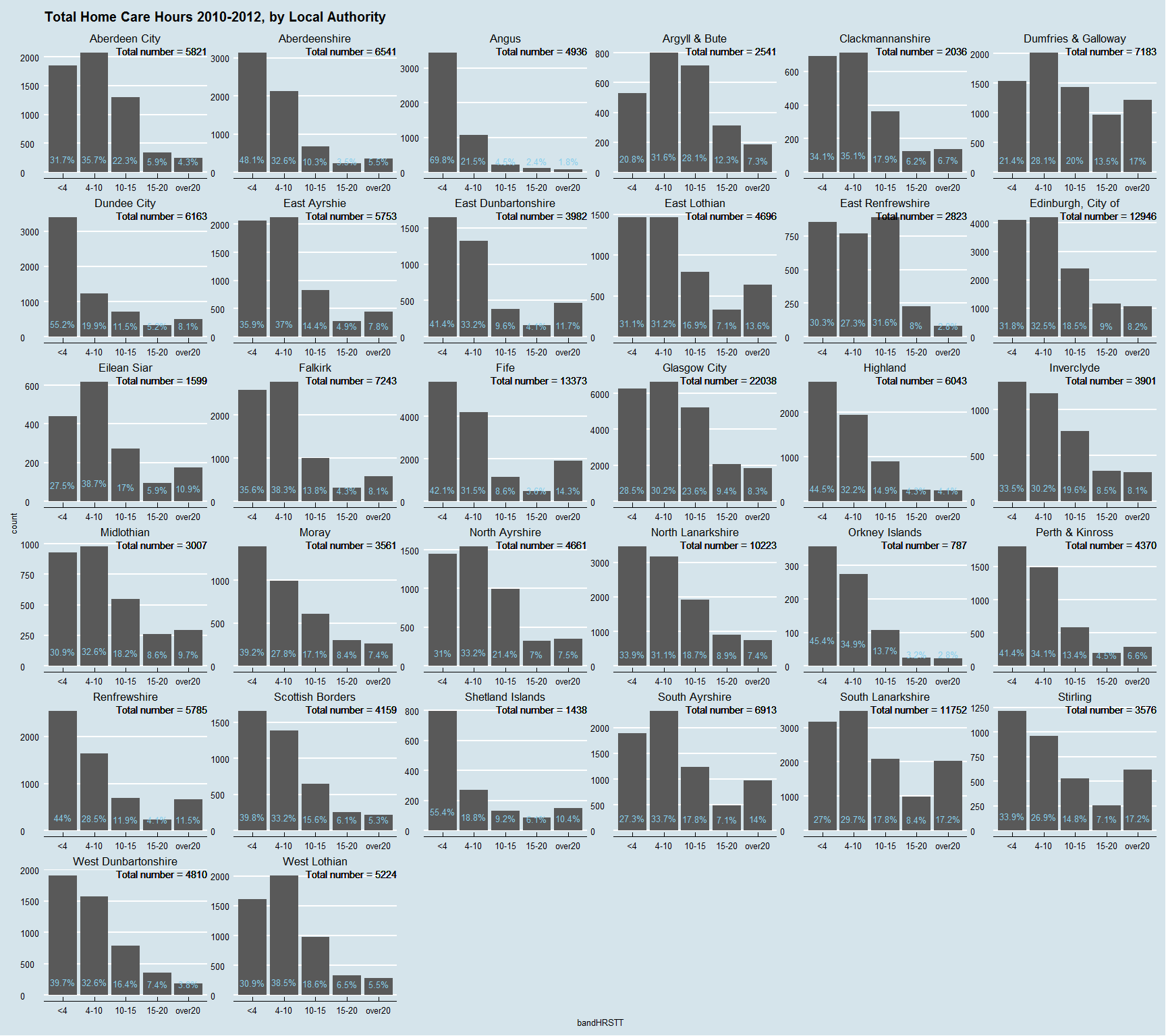
Focusing on the Total hours of home care again, I'll look at the distribution by Age and Gender.

gender-1.png)

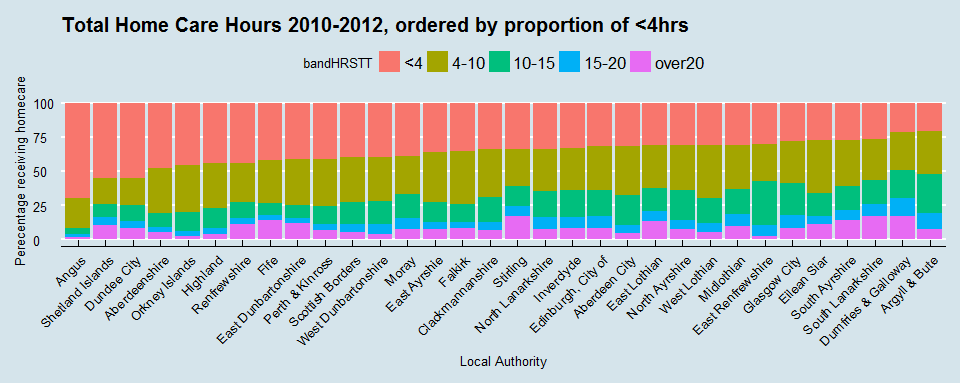
So fairly similar distributions across genders and age (The y-axis makes this a little awkward to see).

### Distribution by Local Authority

Again using the total hours provided across all 3 years worth of data I'll look at the distribution by LA



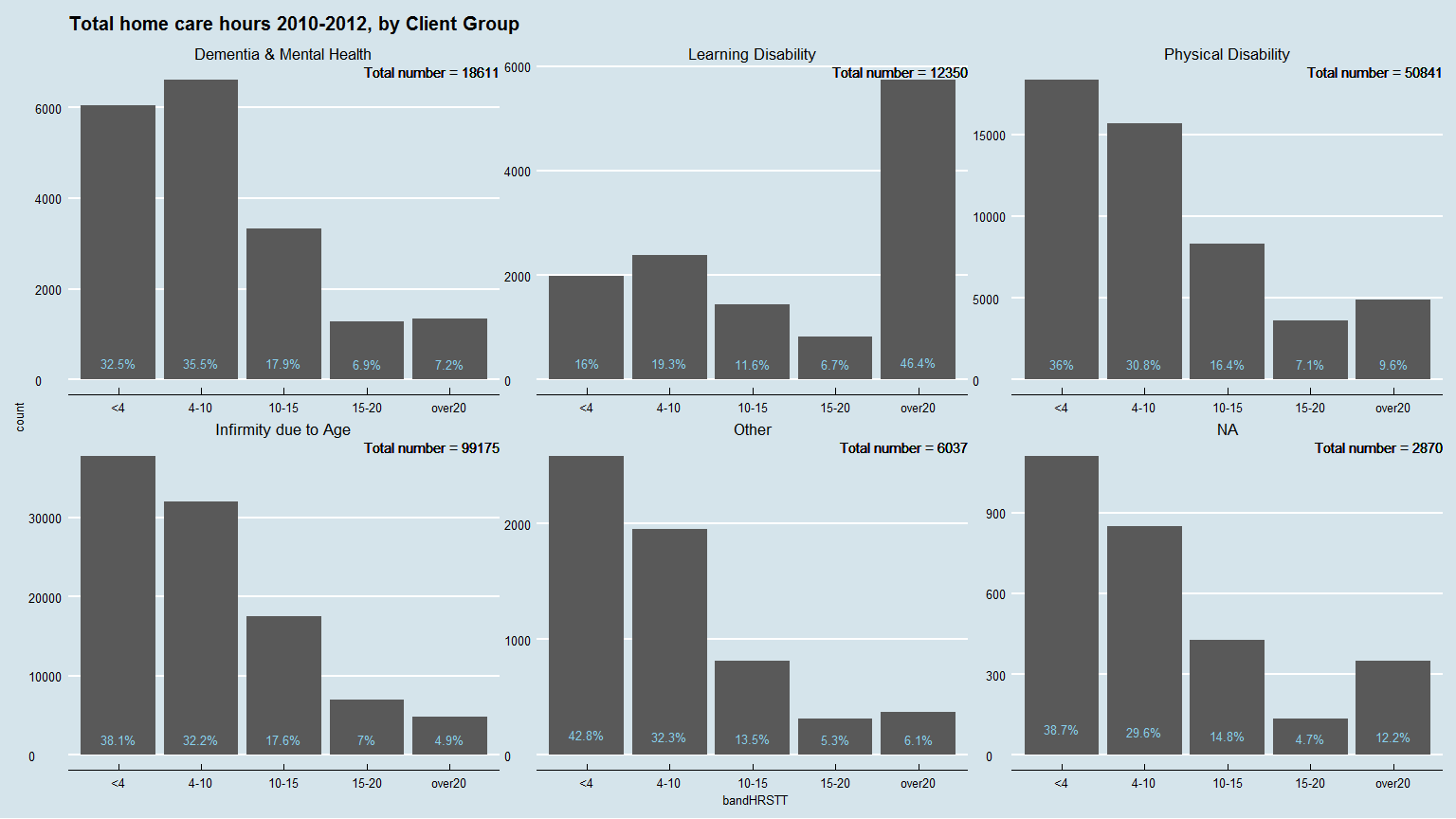
So we can see some pretty wide variations here. To make comparison a little easier I'll stack the data.



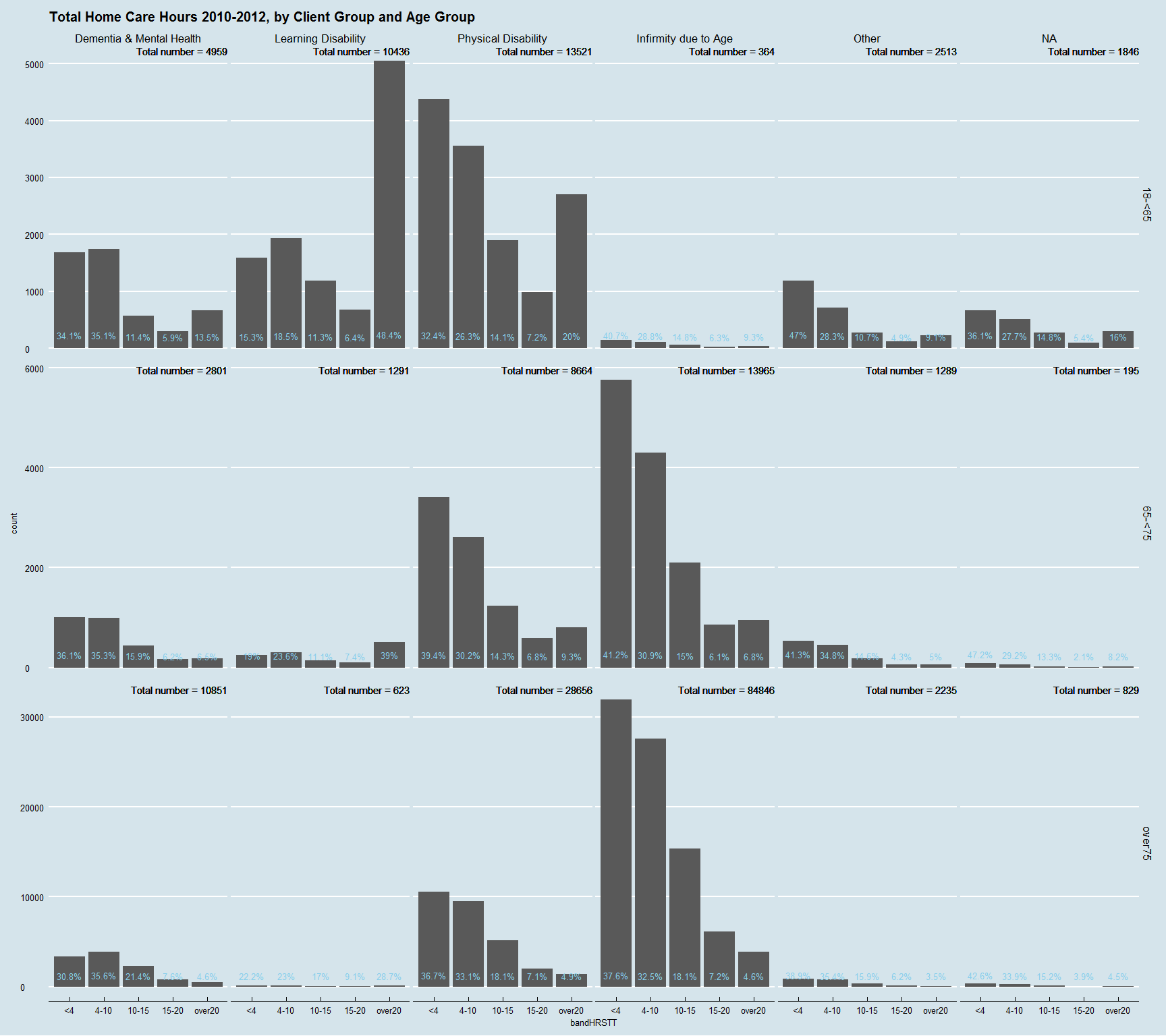
So Angus seems like a bit of an outlier with almost 70% of its clients receiving less than 4hrs of Home care. There doesn't seem to be any obvious pattern to me - rural/urban or rich/poor LAs seem to be distributed randomly. Big variations in the provision of care - does this affect outcomes??

### Distribution by ClientGRP

I know I said the ClientGRP variable is not completed very well, but I'll plot Total Home care hours by ClientGRP anyway - just for a look....



And again, this time with Age group as an added facet.

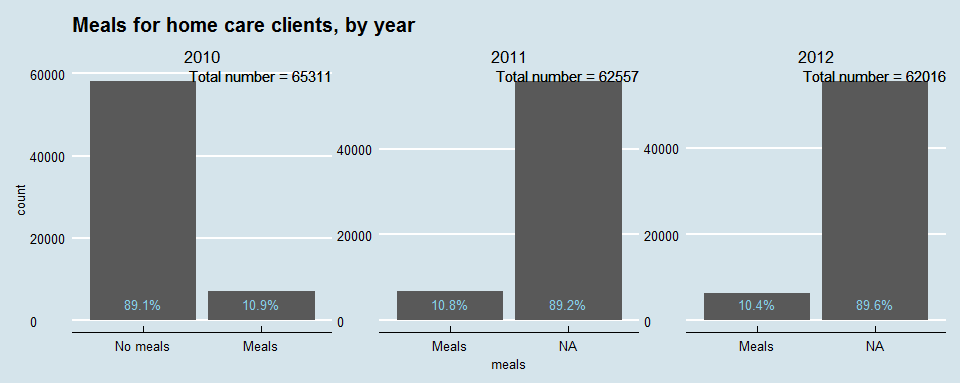


## Meals

As I have mentioned before the meals data is pretty poor. The Social Care Survey reports acknowledge as much saying something along the lines of, "...meals data as proved difficult for Local Authorities to capture."

What is noticeable from the data is that it is completed for 2010 but not 2011 or 2012. I'm going to visulaise quickly to get an idea of how well 2010 is completed.

### National Distribution

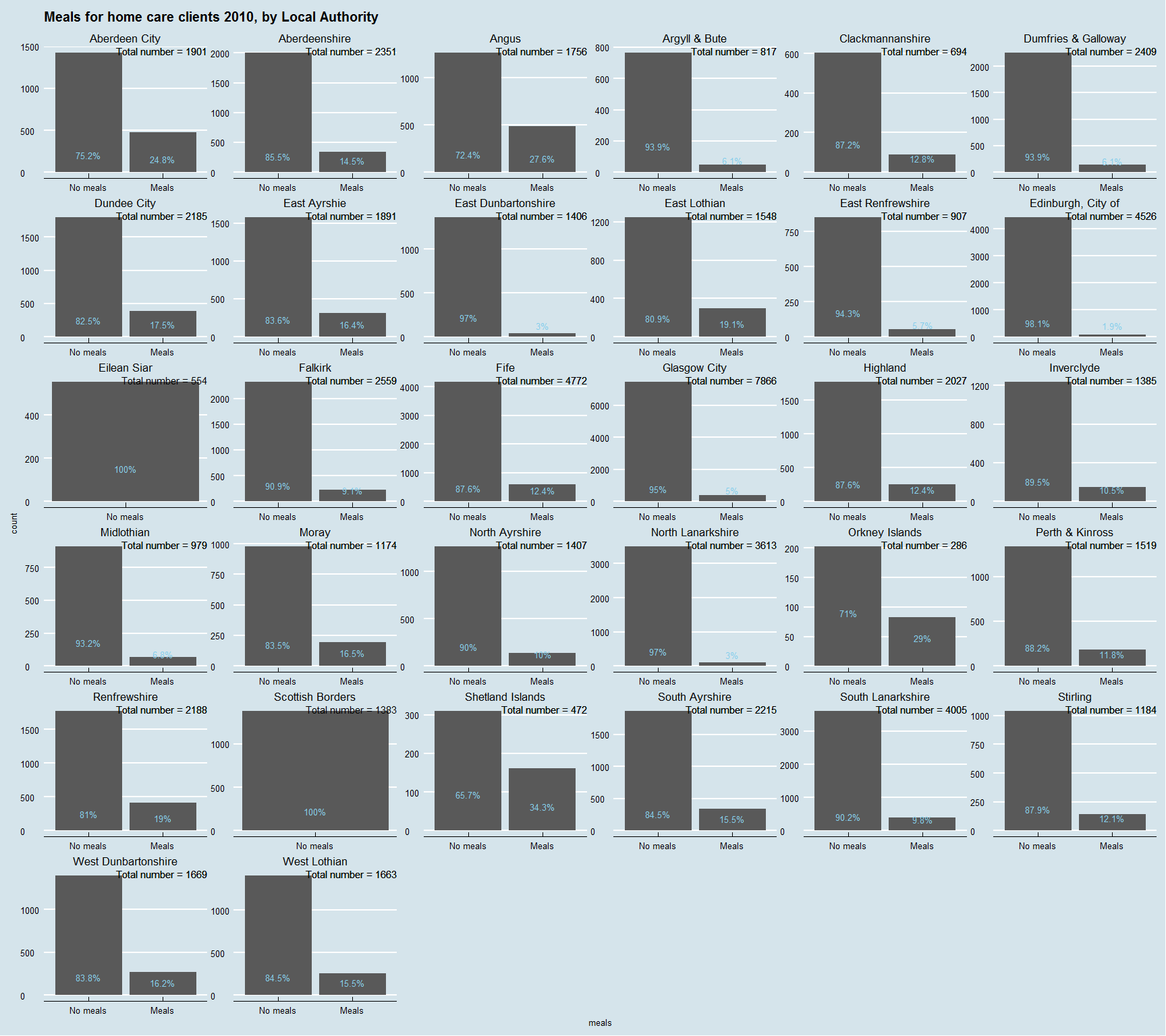


Ok. Interesting. Is the data NA (i.e. missing) or is it "No meals" as in 2010? Difficult to know. Do we really think only 10% of Home care clients get meals delievered? I really don't know. As alluded to earlier, I think we cannot rely on this variable for any meaningful analysis.

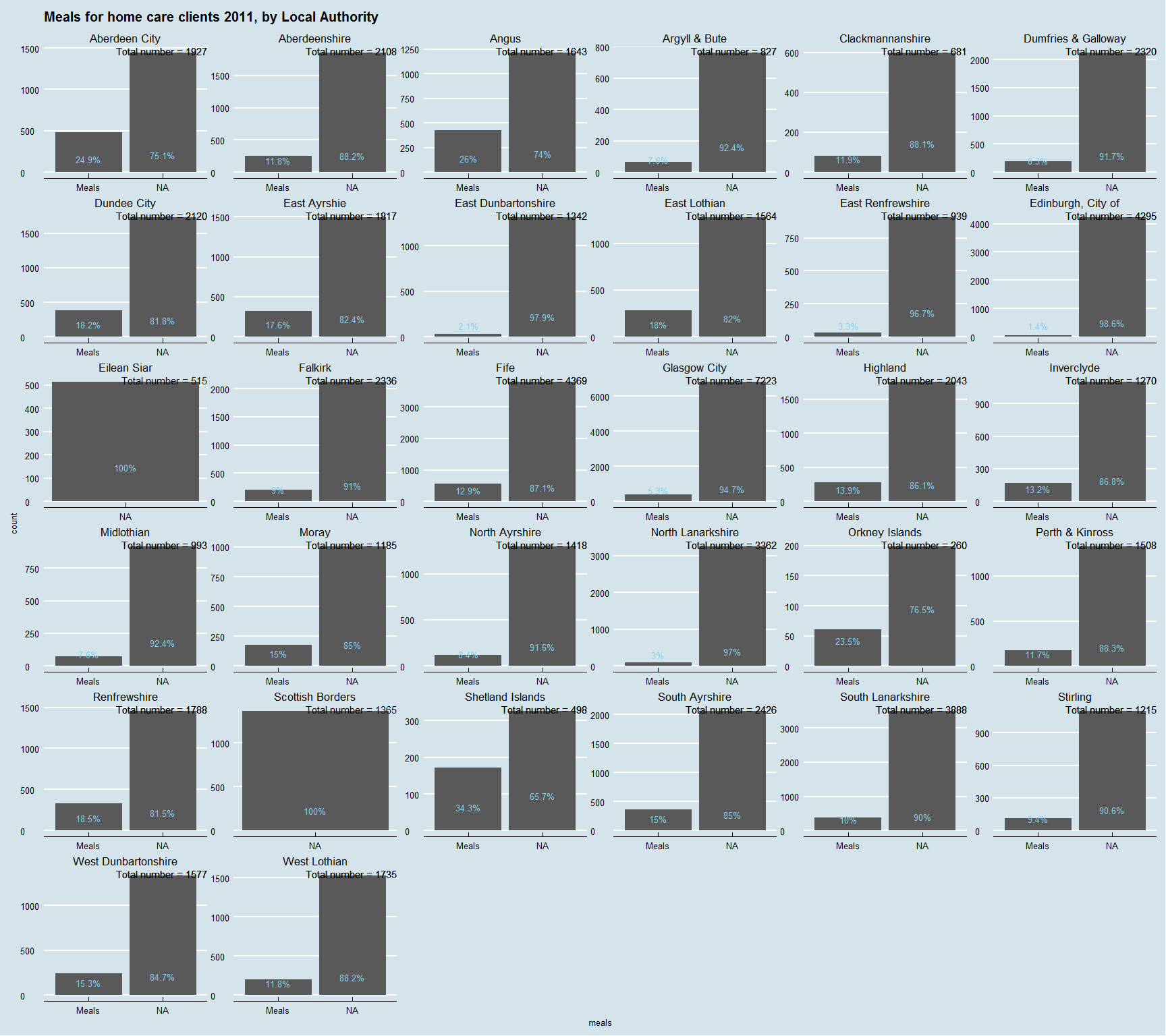
### Distribution by Local Authority

Before I drop it, I'd like to see if the data is collected really well in a small number of LAs.I'll do 3 seperate facet plots for each years worth of data.

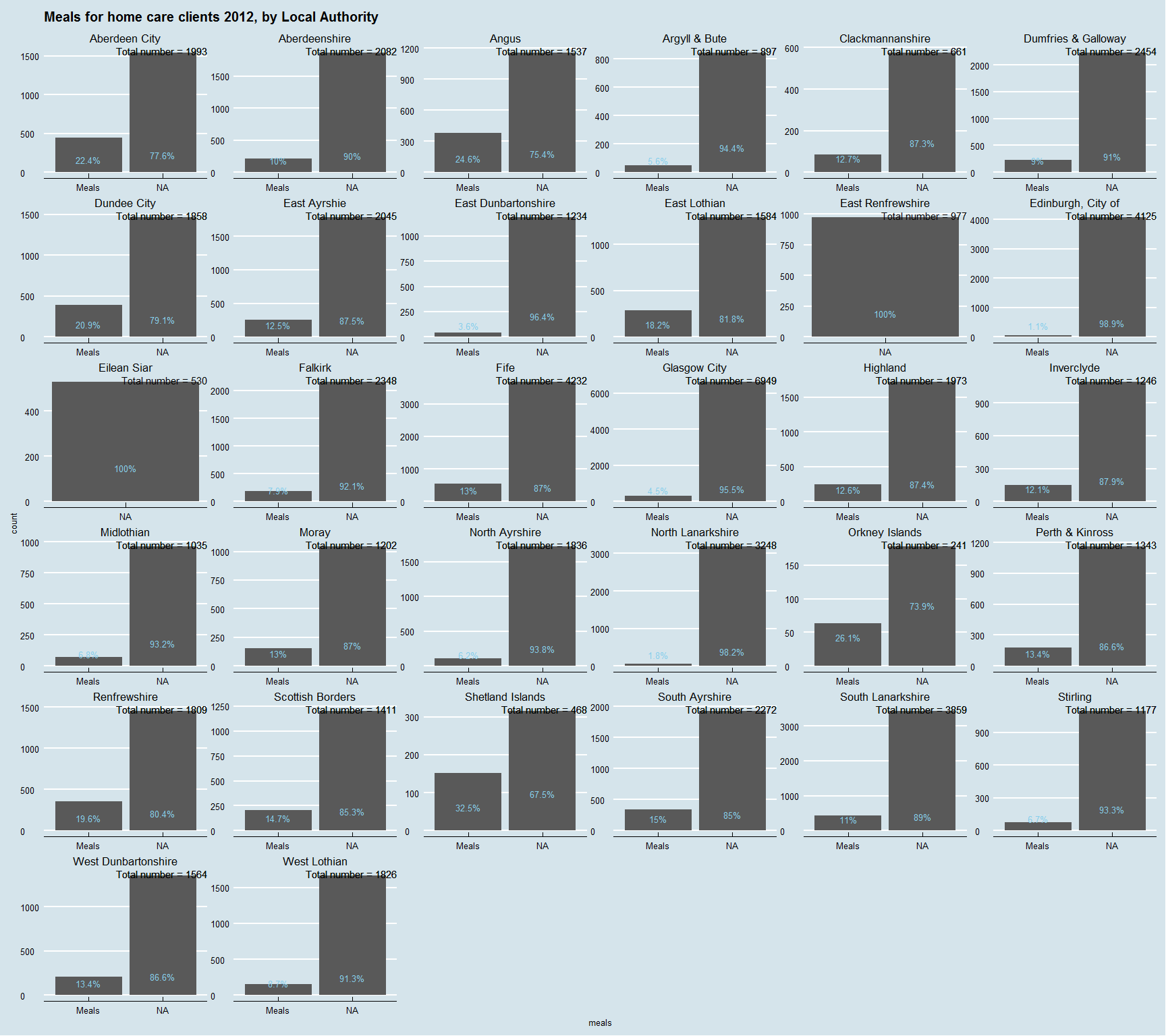
#### 2010



#### 2011



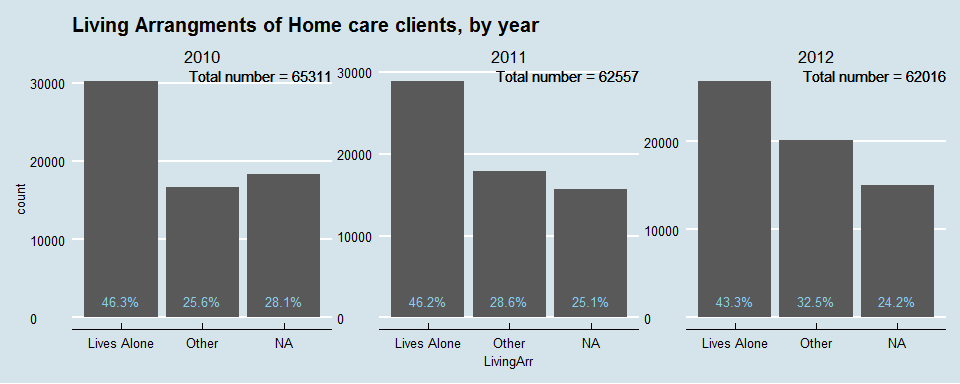
#### 2012



So again, illustrates the data is pretty unreliable. Although, proportions seem to be fairly consistant across LAs, I suspect there is a lot of undereporting. I also think there is no clear distinction betwen missing data and "No meals". Eilean Siar is reported as 100% No Meals in 2010 and 100% NA in the following 2 years - I know that people get meals in the Western Isles. The range of those getting meals is 1.4% in Edinburgh to 35% in Shetland - I find that difficult to believe.

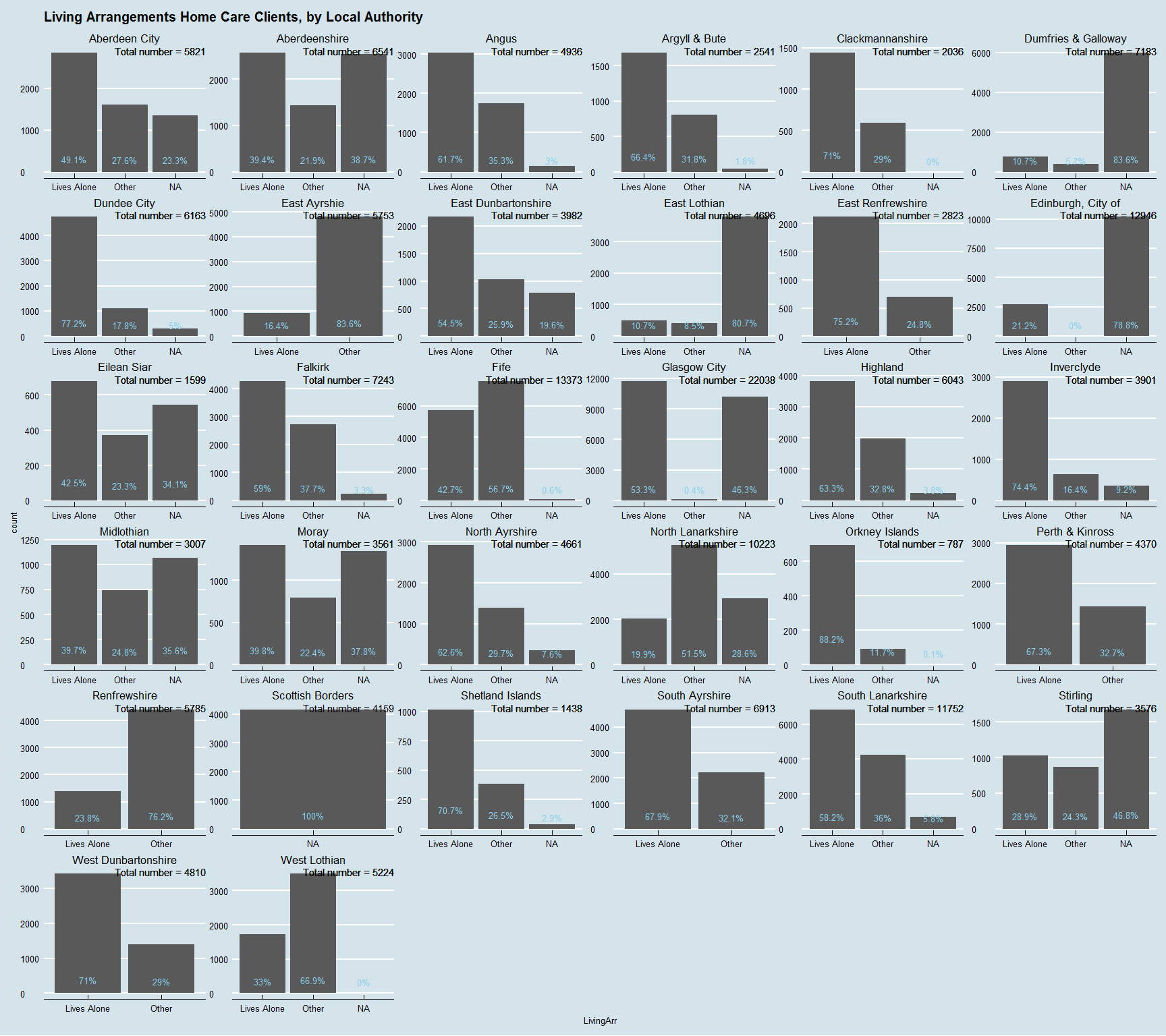
## Living Arrangements

### National Distribution



So a decent amount of missing data here - can we identify if it is missing from 1 particular place?

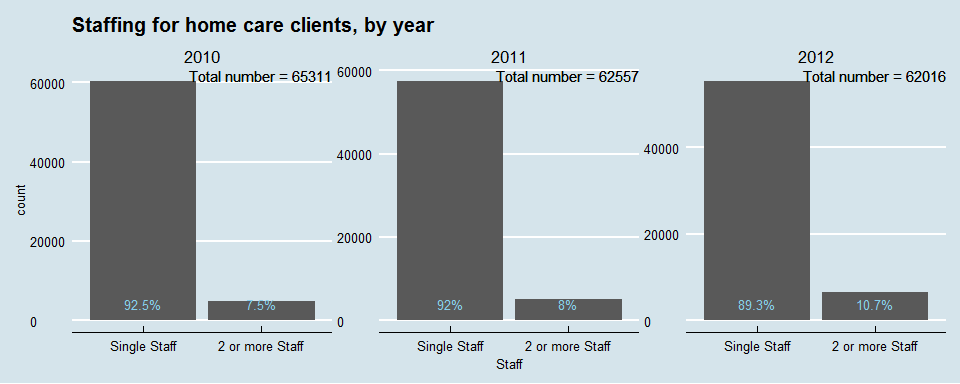
### Distribution by Local Authority



So, another varible that is shows a wide variety of completeness. 18 out of the 32 LAs complete this variable with less than 10% missing data. In the others, the missingness varies from 19.6% in East Dunbartonshire to 100% in the Borders. I can't see any meaningful, comparisons being drawn with this data due to quality.

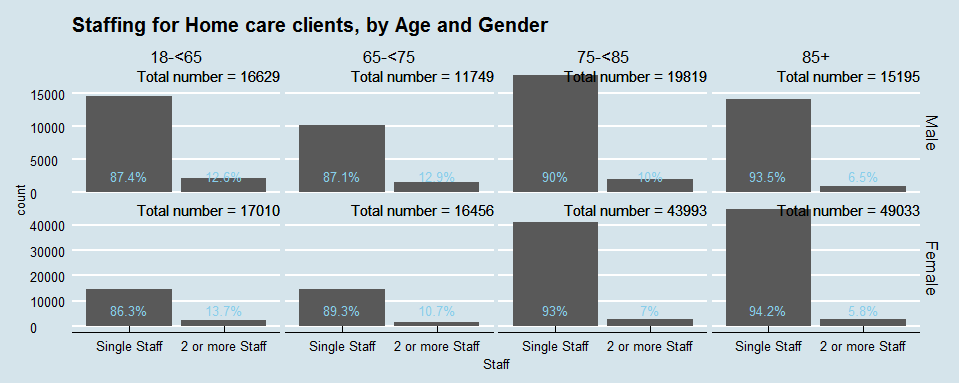
## Staff

### National distribution

This variable measure whether 2 or more staff are required for a client. It may be a good proxy variable for need - extra staff generally required for very immobile cleints. 

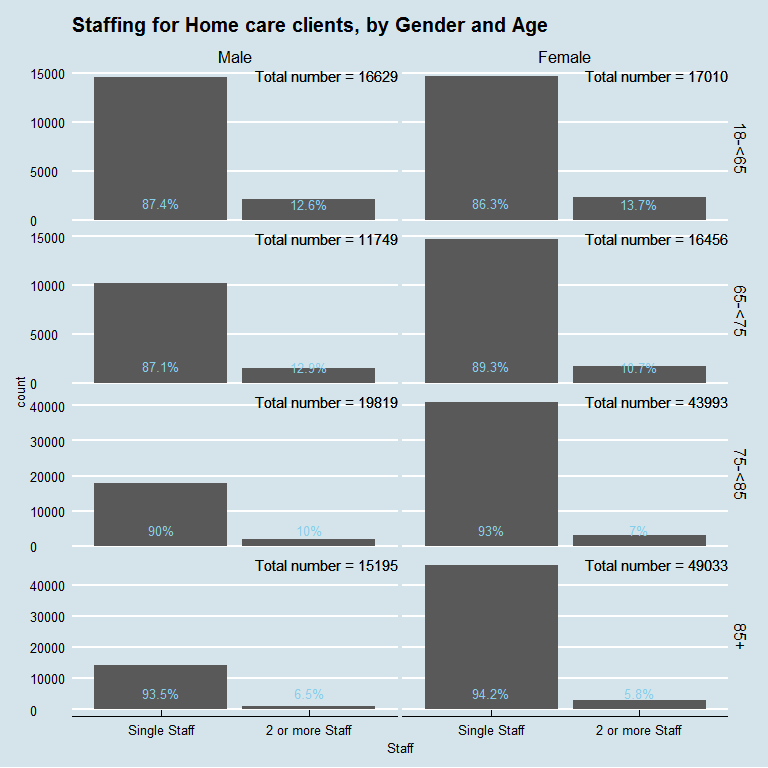
So a small, but increasing, proportion of patients require 2 or more staff. Who are they?

### Distribution by Age and Gender



So the y-axis are fixed to gender which means it is fairly easy to compare across age bands within genders - i.e. the absolute numbers of multistaffed clients is fairly similar for males across ages (slight decrease in 85+), whereas slight increase for 75-85 age group for females although this represents a smaller proportion of overall clients in this group. In general - under 75s have slightly higher proportions of multi-staff.

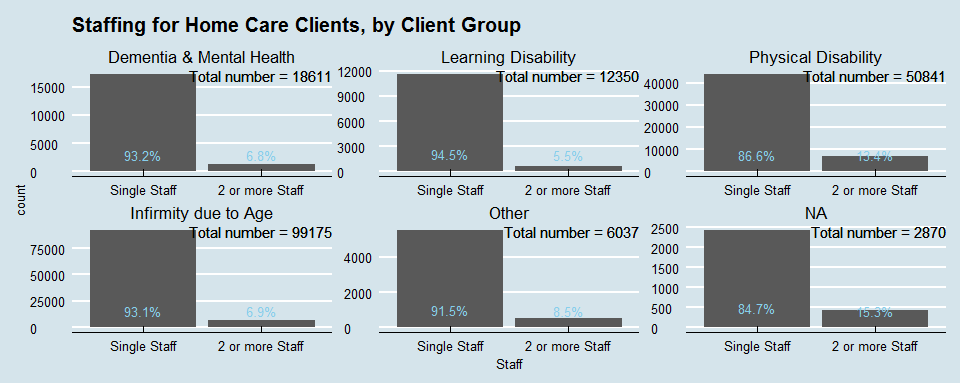
To make comparison of absolute numbers across genders I will flip the chart - same data, different view



So here it is easier to see proportions of multistaff are fairly similar across sexes at different ages and also demonstrates the big difference in numbers receiving home care betwen sexes.

### Distribution by Client Group

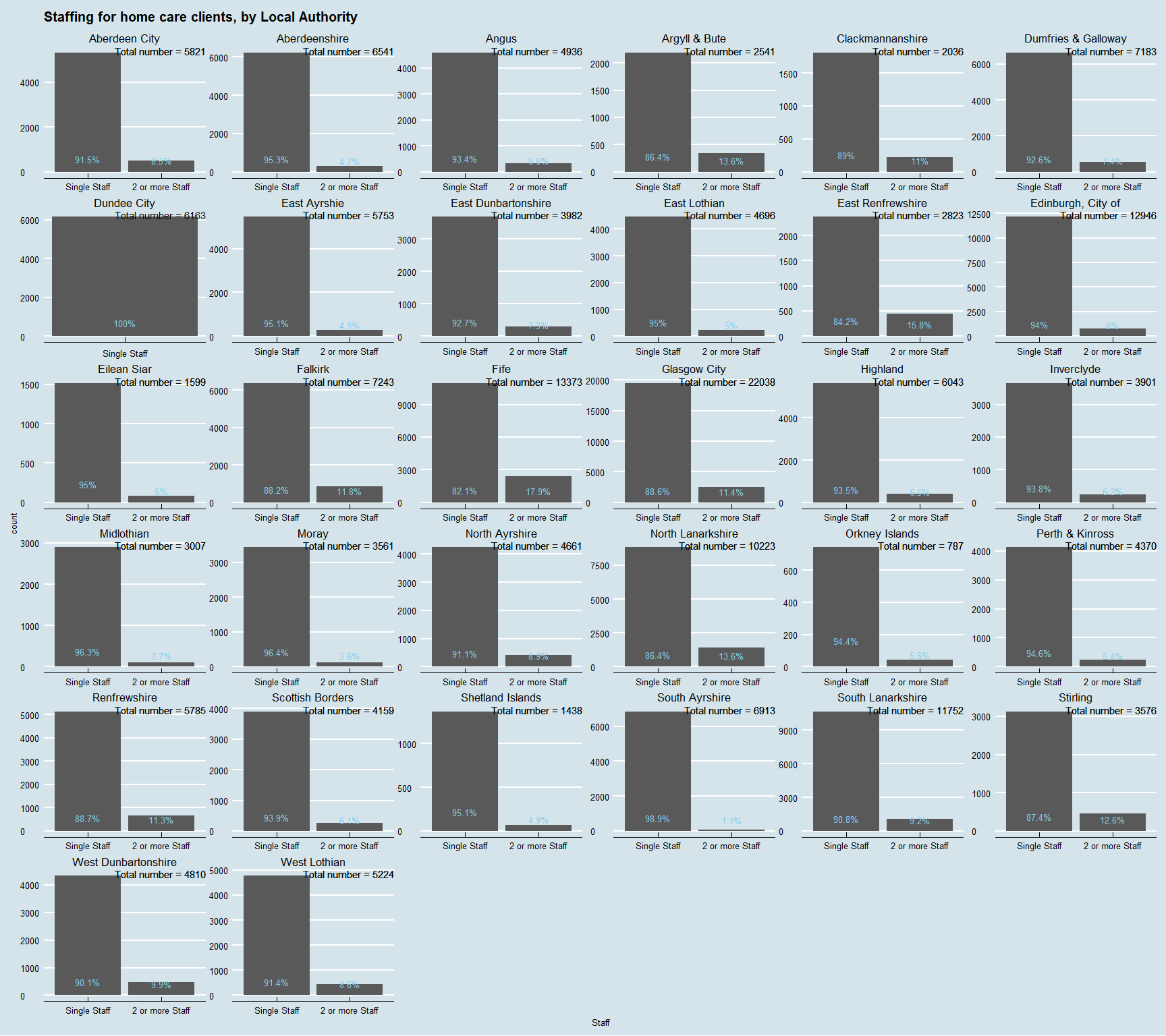
Here, I am mainly interested to see if Learning Disability is associated with multi-staffing (with all the caveats re Client Group data quality)



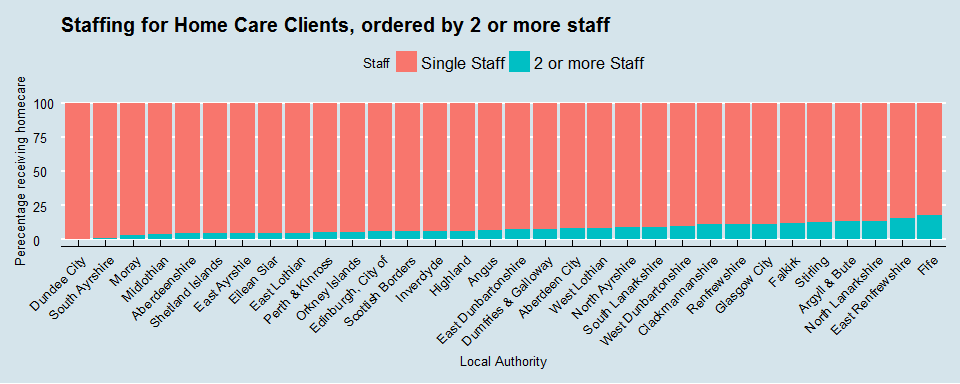
The answer to that question being no. Interestingly of the small proportion of missing data for Client Group - 15% are multi-staffed.

### Distribution by Local Authority

Any LAs with significantly higher proportions of multi-staffed clients?



There actuall is a bit of varition that might be worth looking at on a stacked chart. Dundee City has only single staffed clients??? I'd be wary that this is true.



Multi-staffing ranges from 1.1% in South Ayrshire to 17.9% in Fife which is quite a large range. I can't see any obvious pattern.

## Housing type

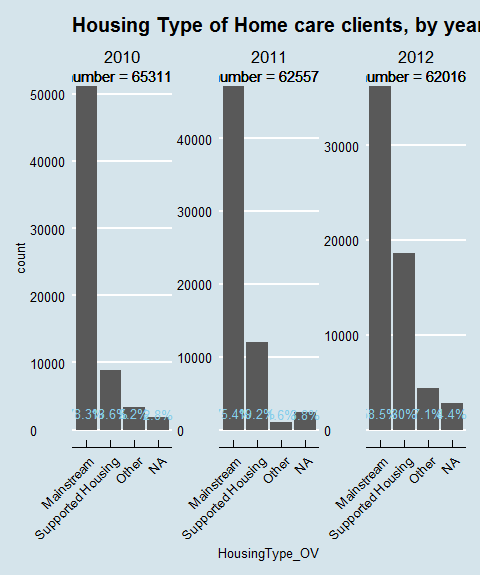
It should be noted that the "OV" part of the Housing type variable name means "Optional variable". I suspect many will not have taken up the option!!!

The metadata published with the open data gives us a little more detail on the categories the variable uses:-

* Mainstream Housing
* This is a private home (either owned/mortgaged or rented) which has not been adapted for special needs in any way.
* Supported Housing
* Special housing: premises that have been adapted to meet the need of people with particular needs, e.g. wheelchair access.
* Amenity housing: a group of premises with special modifications for particular needs but not supported by a warden. + Sheltered housing: self-contained premises linked to a warden who provides specialist support to tenants. + Supported accommodation: A home where external support is put in place to help the tenants live as independently as possible. • Other

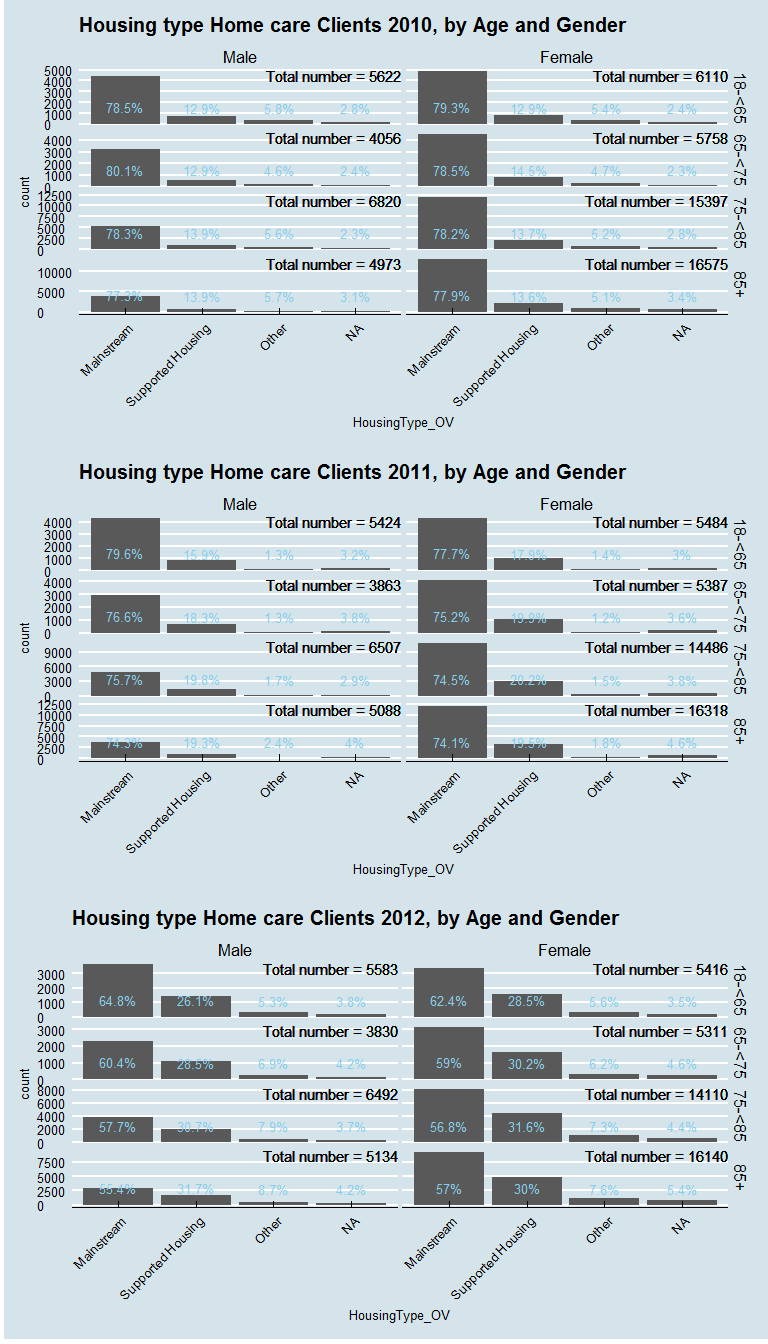
Absolutely no information on what "Other" means!!

### National distribution



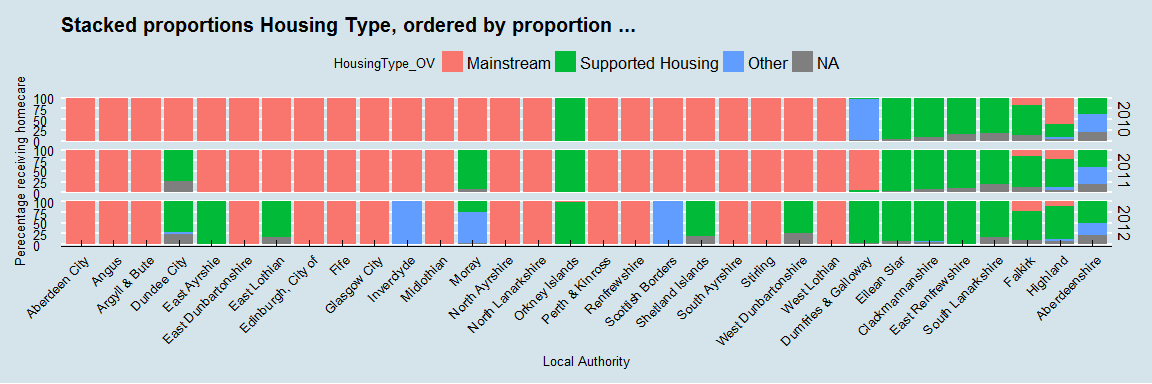
Hmmm. A really big increase in the proprotion of Supported Housing across the years. Does this reflect better data quality or changes to reporting criteria? I can't imagine there was a sudden increase in the amount of supported accomodation available. I'll need to do all sub-analyses by year. Queries over quality here again...

### Distribution by Age and Gender



Similar distributions across age and gender. We can see the gradual increase in proportions of Supported housing is across all ages and both sexes.

### Distribution by LA



We see here that any variation previously noted in the distribution of Housing Type is enitrely down to reporting practices. In 2010, the majority of LAs reported 100% mainstream housing. A minority reported almost 100% Supported Accom.

This continued in 2011 with a few LAs flipping from entirely Mainstream to entirely Supported.

2012 saw an increase in LAs flipping from Mainstream to Supported and also a couple deciding everybody should be in "other".

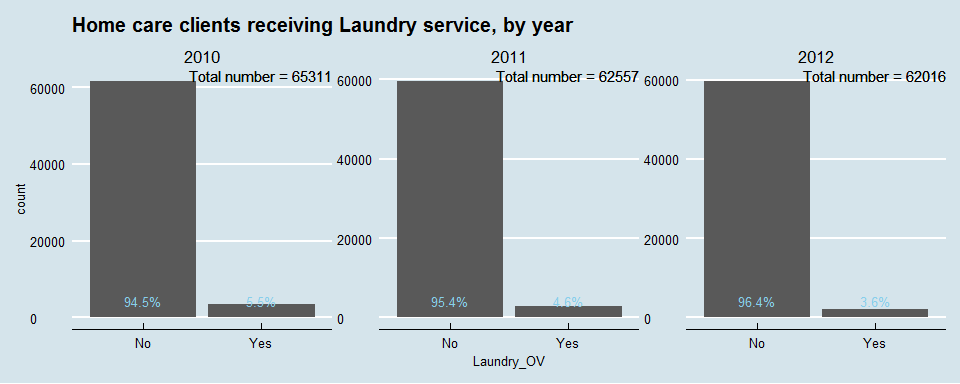
My own personal favourite is Dumfries and Galloway who reported almost 100% "Other" in 2010, almost 100% "Mainstream" in 2011, and almost 100% "Supported Housing" in 2012.

This variable is useless.

## Laundry

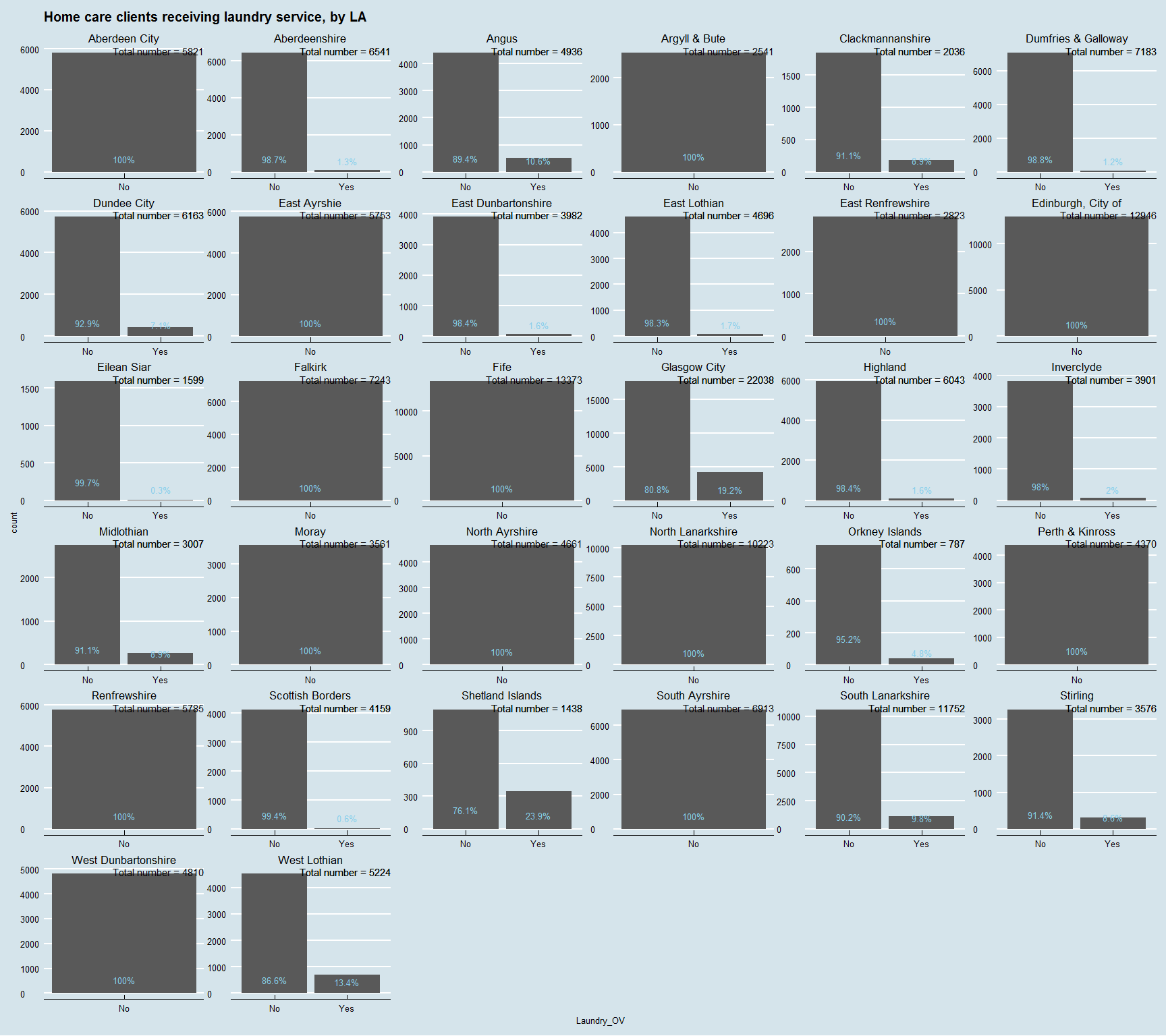
I'm not holding out big hopes here - this is another optional variable.

### National Distribution



### Distribution by Local Authority

I think I'll jump straight to LA distribution to see how well it is reported.

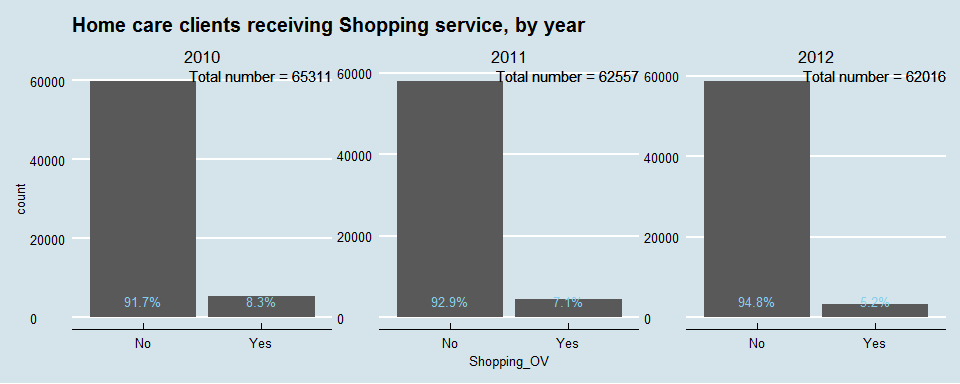


So 14 LAs either don't provide Laundry services or don't report on it. It could be either. I suppose the question is: do we think the 5% of home care clients receiving laundry service have different outcomes from those that don't?

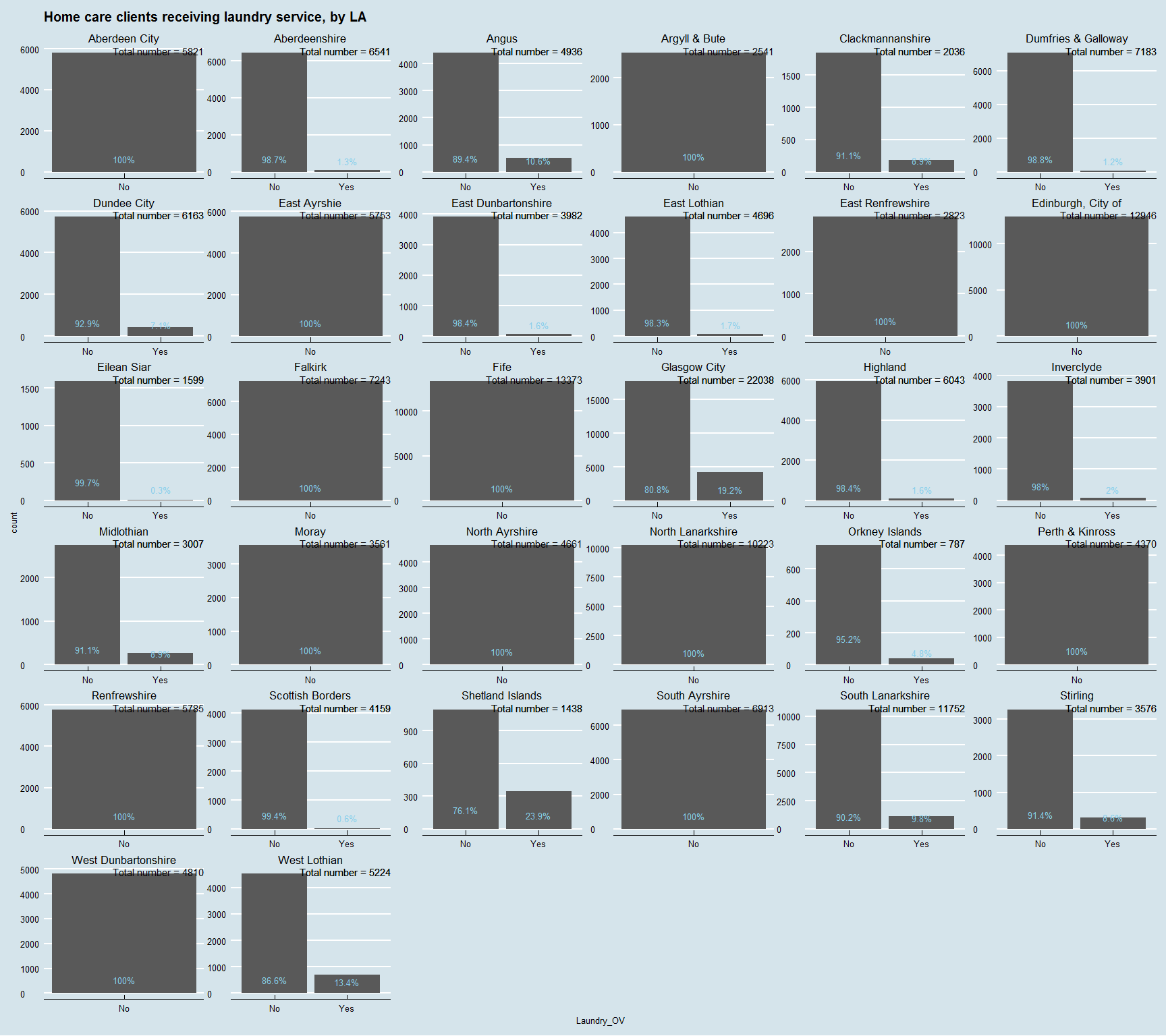
I'm not going to do any further breakdown analyses here but may use this variable to derive a new "other services" variable. I want to look at the other optional variables first.

## Shopping

### National Distribution

 I suspect we may have the same story as Laundry here...

### Distribution by Local Authority



Exactly the same LAs returning data as for Laundry.

## Housing Support

Metadata description:-

Housing support services help people to live as independently as possible in the community. They can either be provided in your own home or in accommodation such as sheltered housing or a hostel for homeless people.

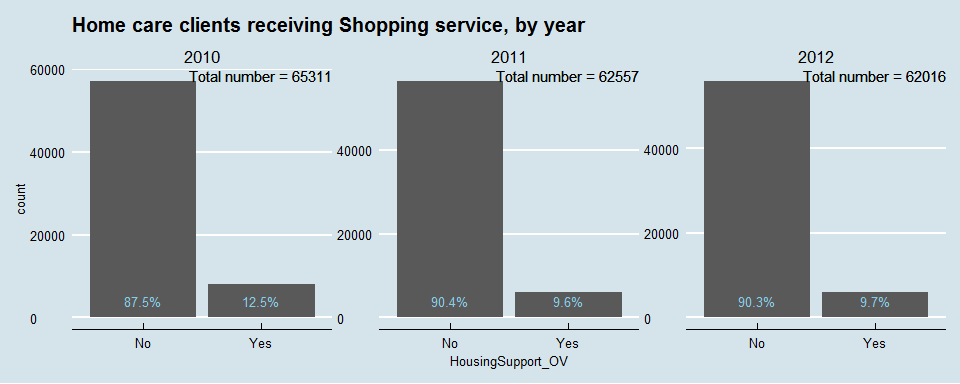
Housing support services help people manage their home in different ways. These include assistance to claim welfare benefits, fill in forms, manage a household budget, keep safe and secure, get help from other specialist services, obtain furniture and furnishings, and help with shopping and housework. The type of support that is provided will aim to meet the specific needs of an individual person.

Housing support services are mainly provided by local authorities, housing associations and voluntary sector organisations. They help a wide range of people to live independently in the community, by providing practical support and advice. People who may benefit from housing support services include.

* older people.
* homeless people.
* people with physical or learning disabilities.
* people with mental health problems.
* people with drug or alcohol problems.
* people with HIV or AIDS.
* care leavers and.
* women escaping domestic violence

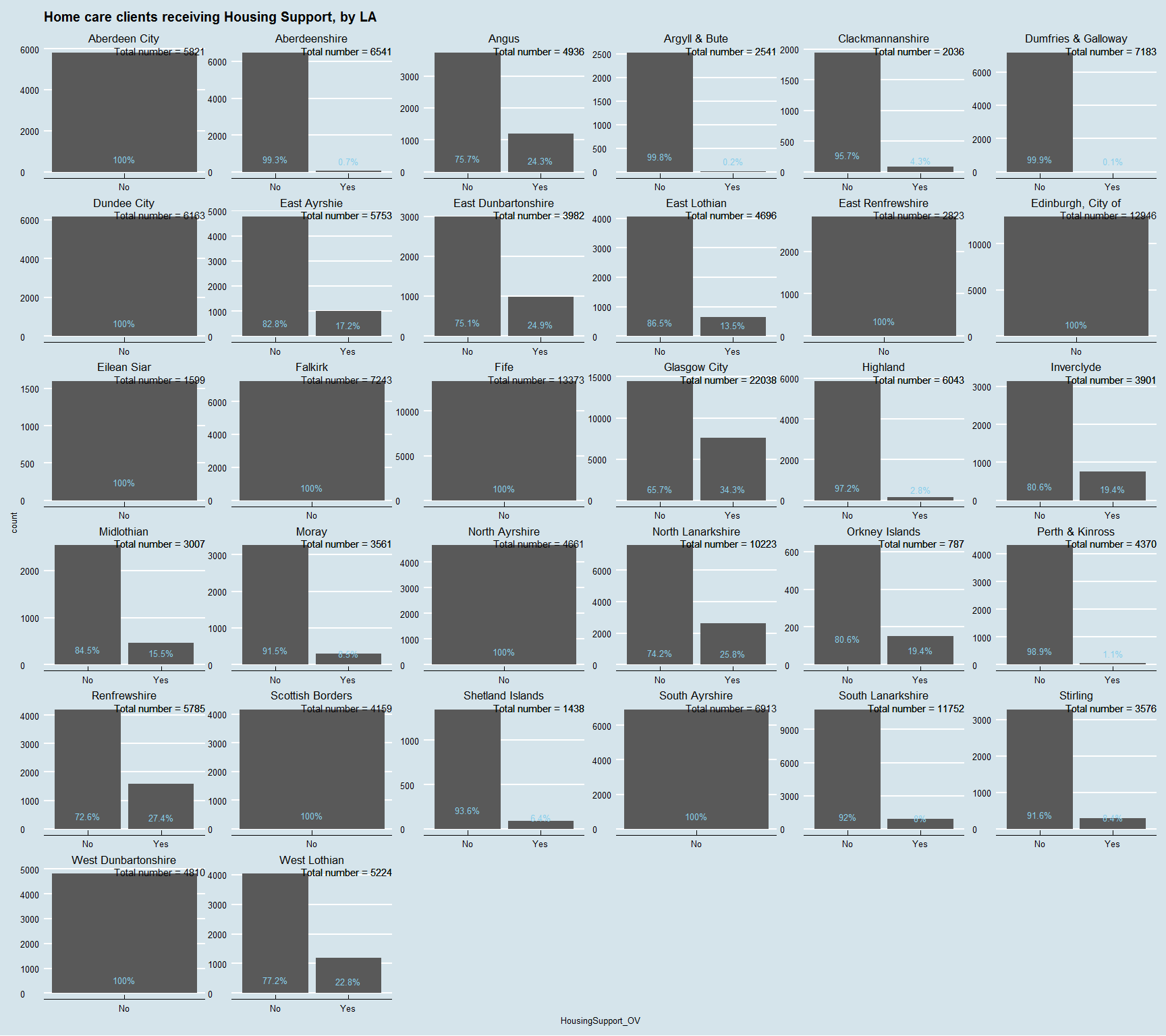
There are many different kinds of services. Examples include home adaptations for disabled people; visiting support to help with housework and shopping; resettlement support; and community alarms. Services can be provided in someone's own home or within temporary accommodation such as homeless hostels and refuges.

### National Distribution



Higher proportions of this variable.

### Distribution by Local Authority.



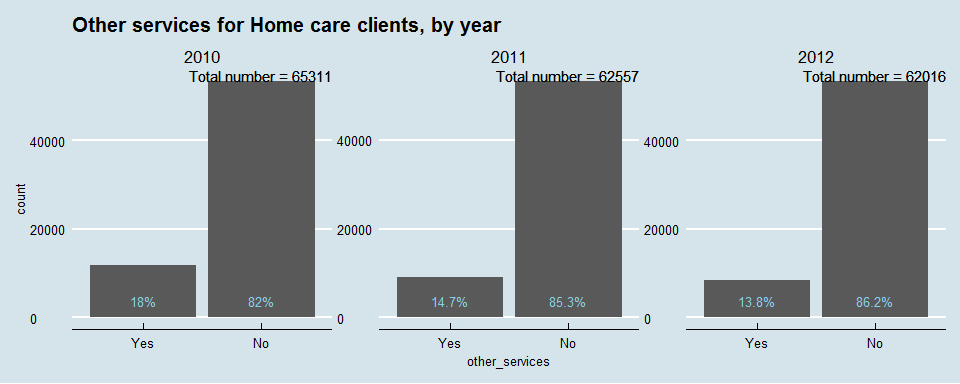
Slightly more LAs providing/reporting on this service.

## Other services.

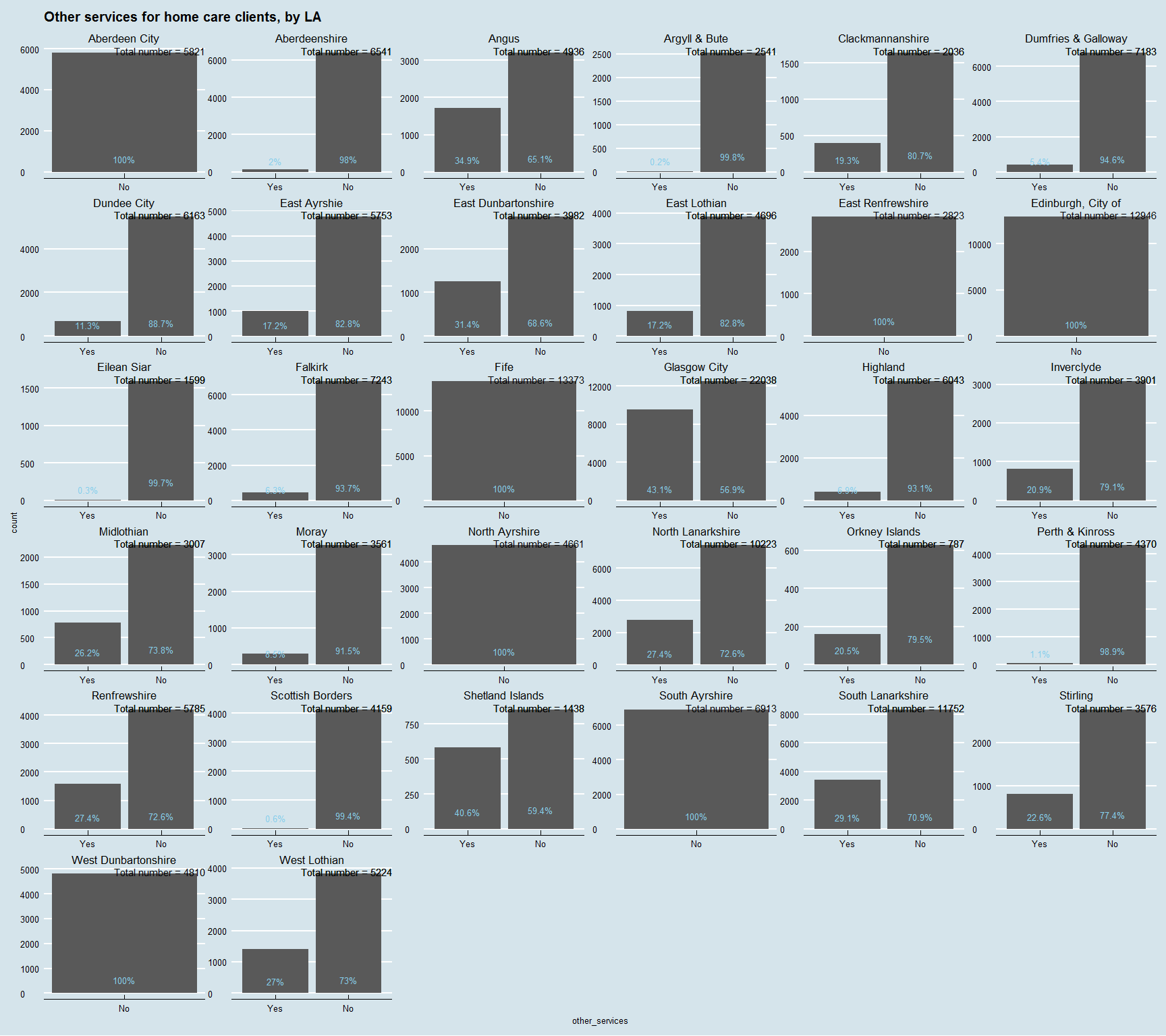
I am going to derive a new variable "Other Services" by identifying those that receive *any* of the following - Laundry, Shopping, Housing Support. I'll then describe this variable as above by age, gender etc.

### Create variable

### National Distribution



### Distribution by Local Authority

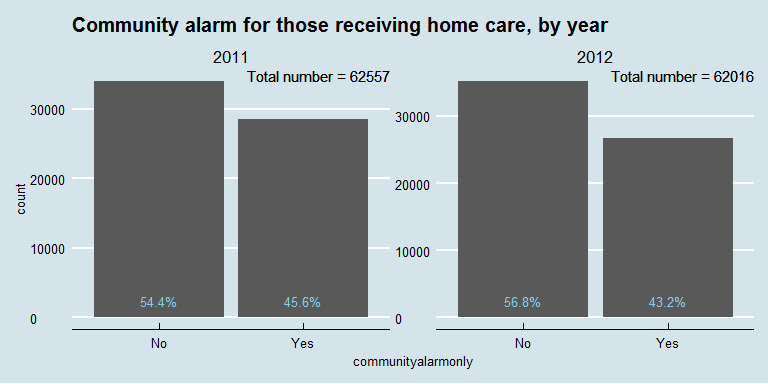


I had wondered if this derived variable may have been of some use. Ideally it would be nice to compare outcomes for those that do or do not receive extra services. Having said that the quality of the data again is questionable here. The variations are so large. Is that due to real variations in provision of care - or is it due to how well this data is reported. Given these are optional variables I don't think they can be trusted.

## Community Alarm only

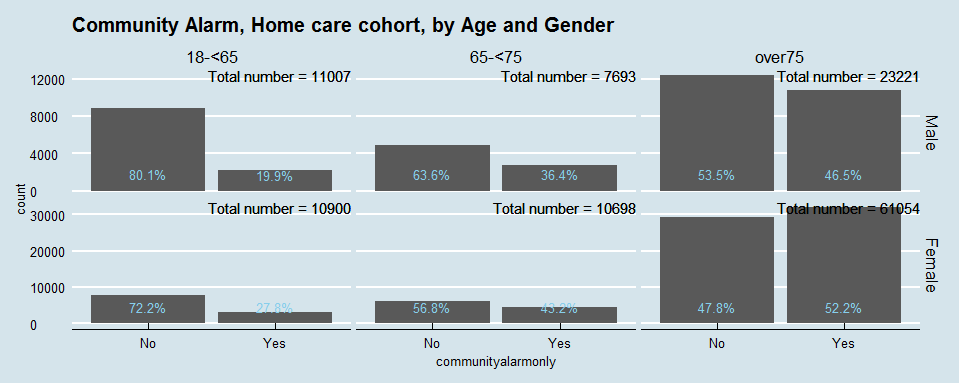
2011 and 2012 report data on 3 telecare variables. I will look at these now. 1st of all I'll create a subset dataframe omitting 2010.

### National Distribution



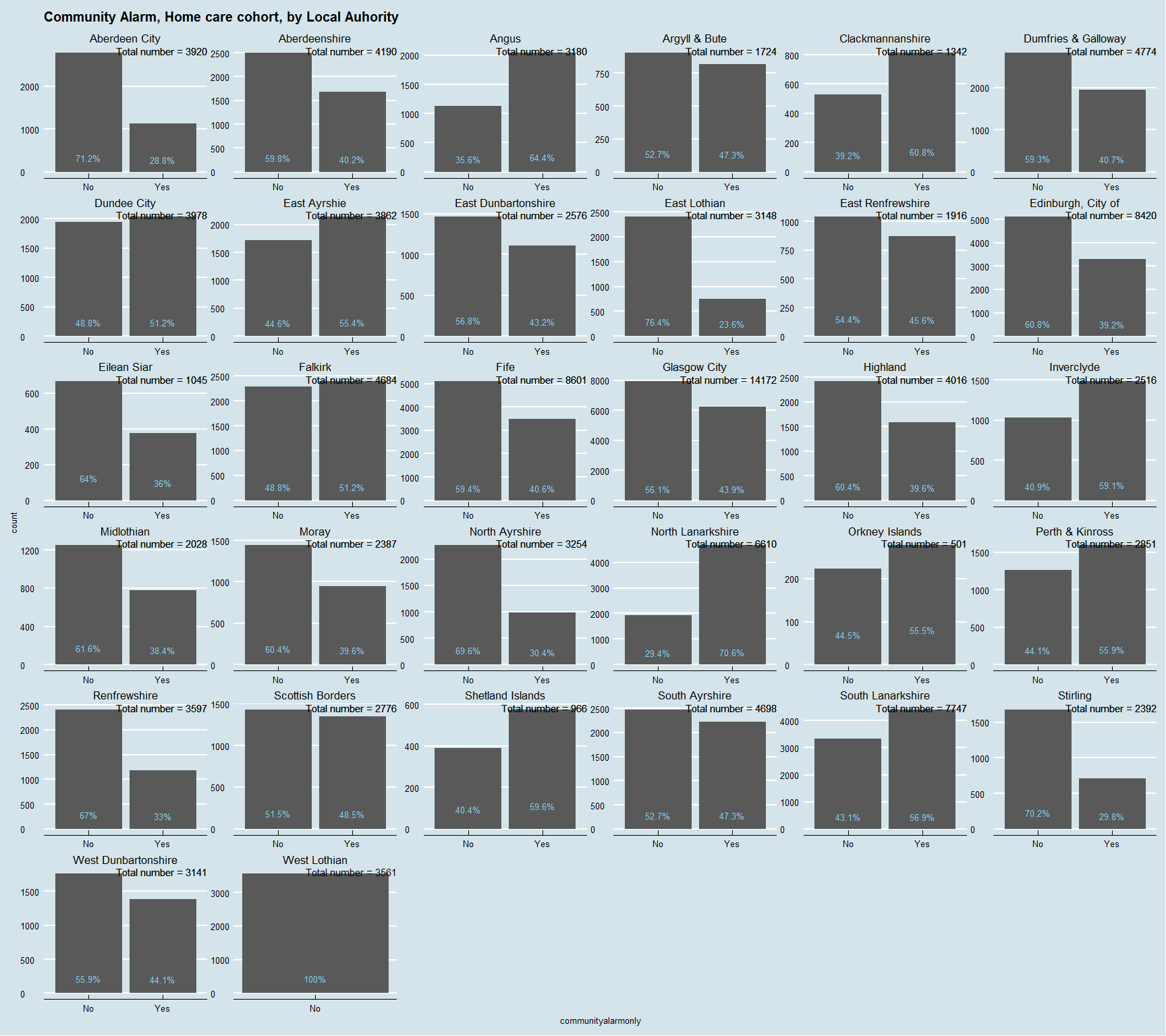
So a large proportion of clients receiving home care also have a community alarm.

### Distribution by Age and Gender

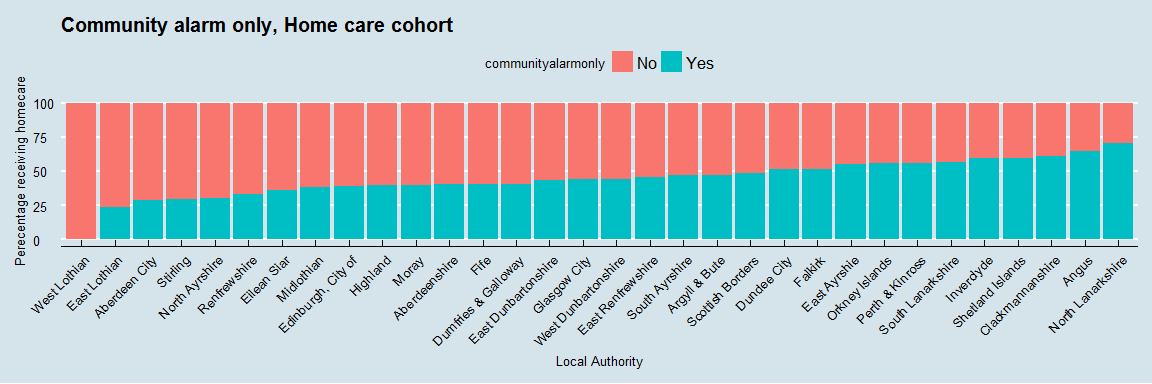


Females more likely to have a community alarm across all age bands.

### Distribution by Local Authority



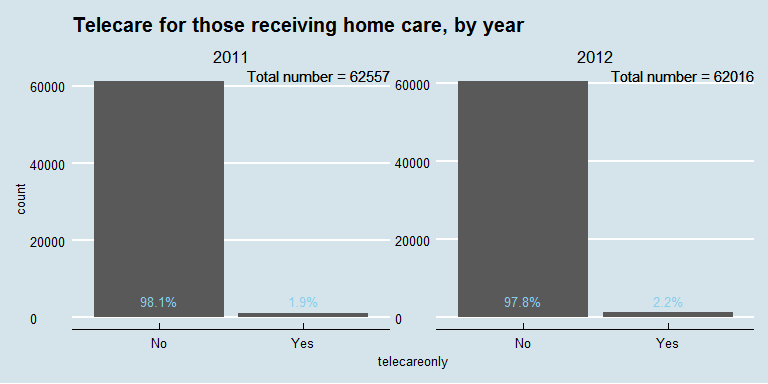
Big differences across LAs in who does/does not provide community alarm for their home care clients. This could be a useful distinction for outcomes.

I'll stack and order 

## Telecare only

The metadata does not give a description of exactly what telecare is. Clearly it is something different to a community alarm requiring seperate data collection.

### National Distribution



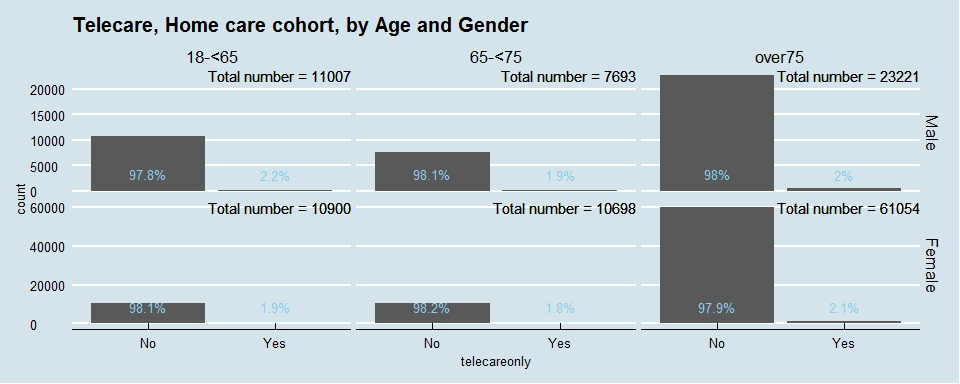
So only a fraction of the number receiving community alarm. How many exactly?

## [1] 2544

Unlikely to have power to be used as a comparator on its own - but could be lumped with commuity alarm and community alarm and telecare(which we haven't looked at yet).

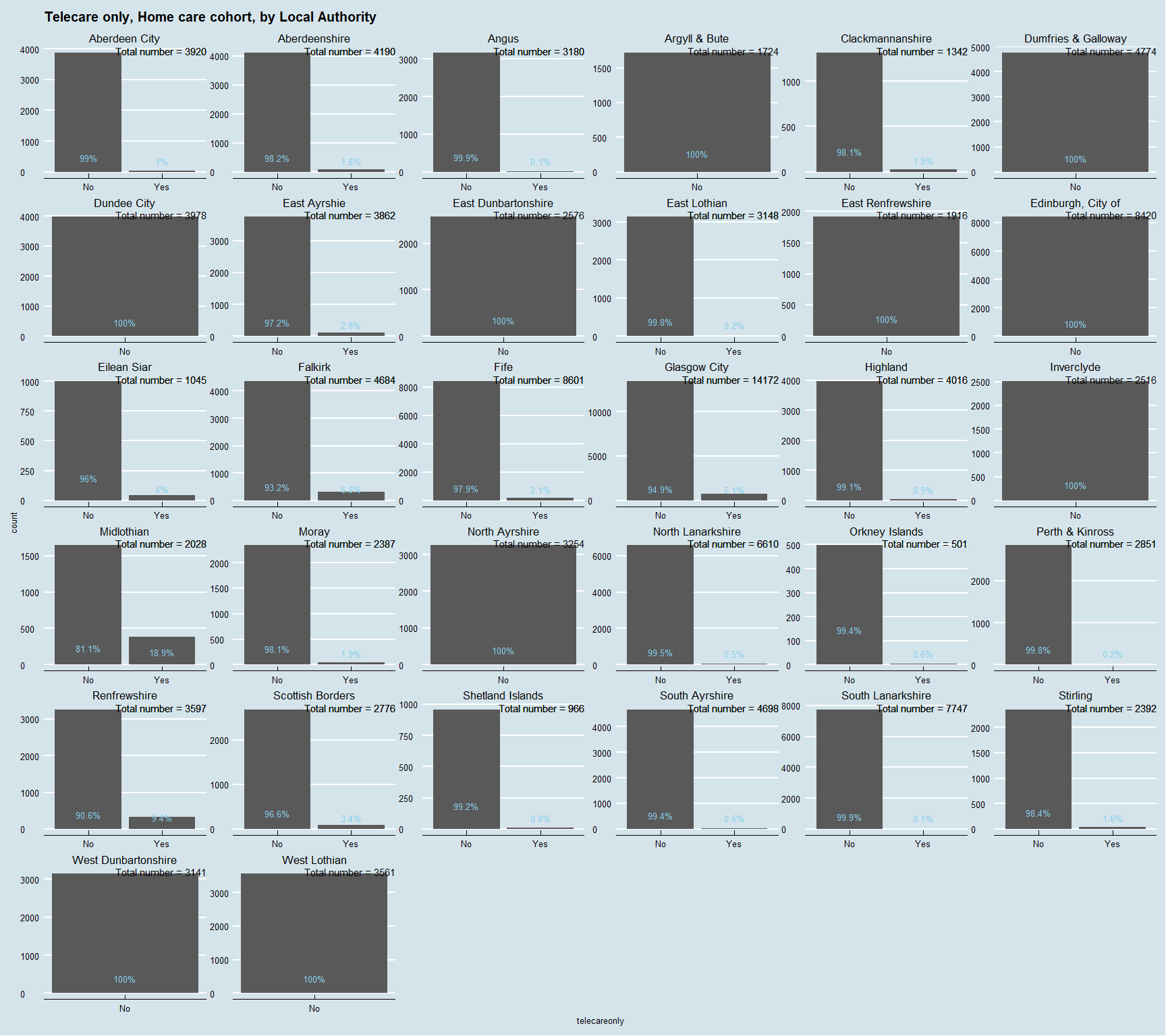
Quick look at distributions anyhow.

### Distribution by Age and Gender



Similar distrinution across age and gender.

### Distribution by Local Authority

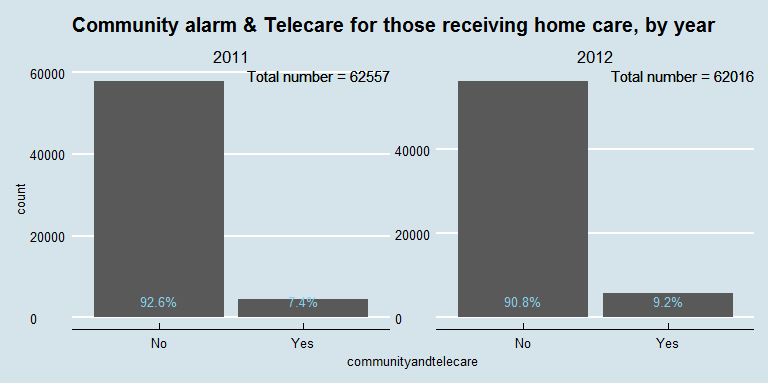


So some LAs provide no telecare. Distribution fairly evenly spread among those that do with the notable exception of Midlothian. Recording error or policy initiative? If the latter then may be worth sub-group analysis although we are talking about less than 400 individuals.

## Community alarm AND telecare

The final telecare variable is for thoise that receive home care, have a community alarm AND alos have a telecare service. Expecting small numbers.

### National Distribution

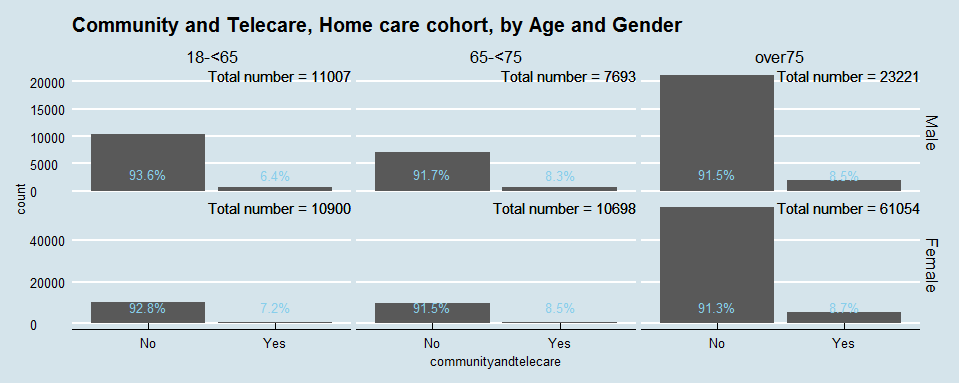


Ha! that's me told - higher numbers than in the telecare only cohort!

## [1] 10347

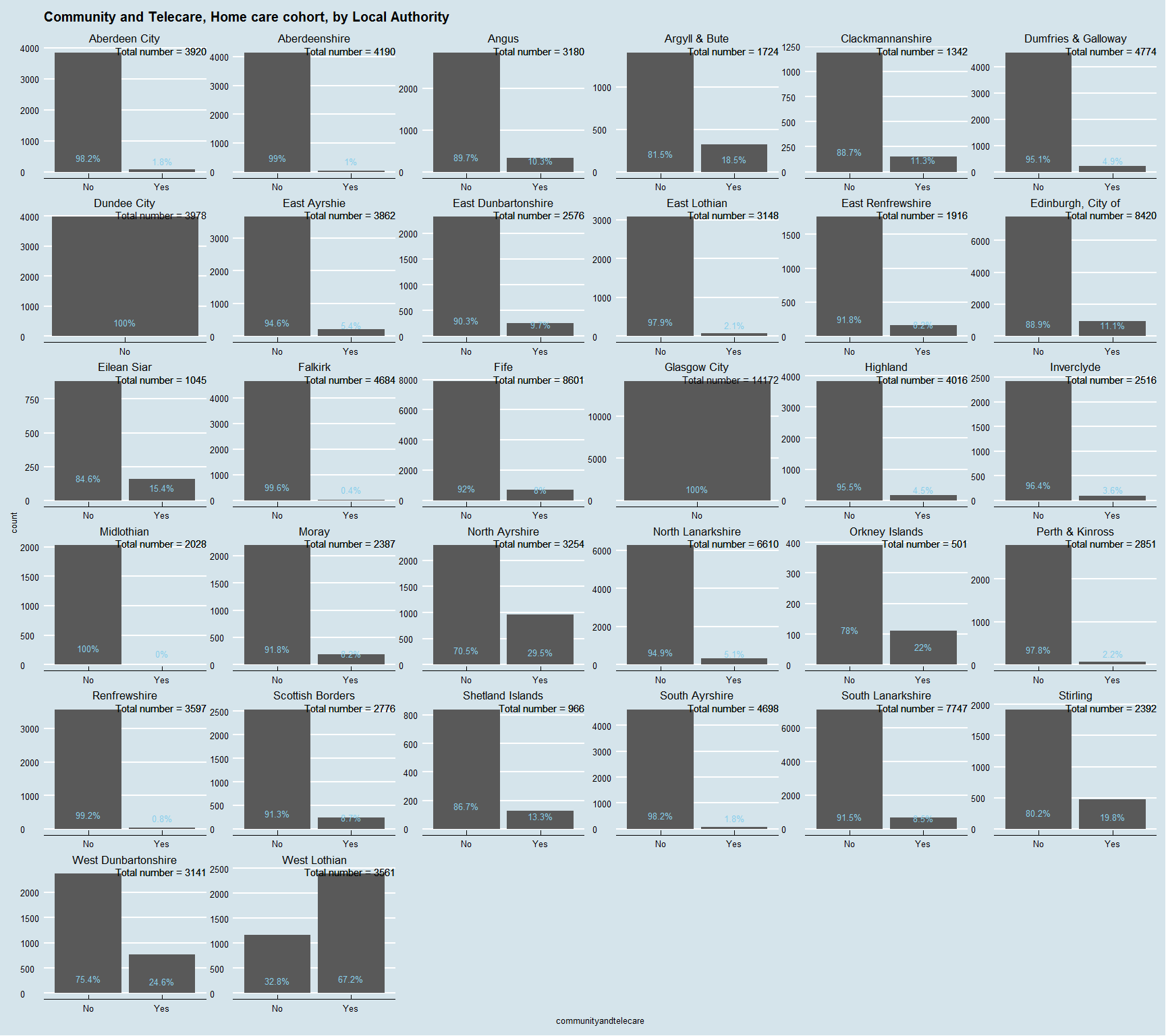
5 times as much!

### Distribution by Age and Gender



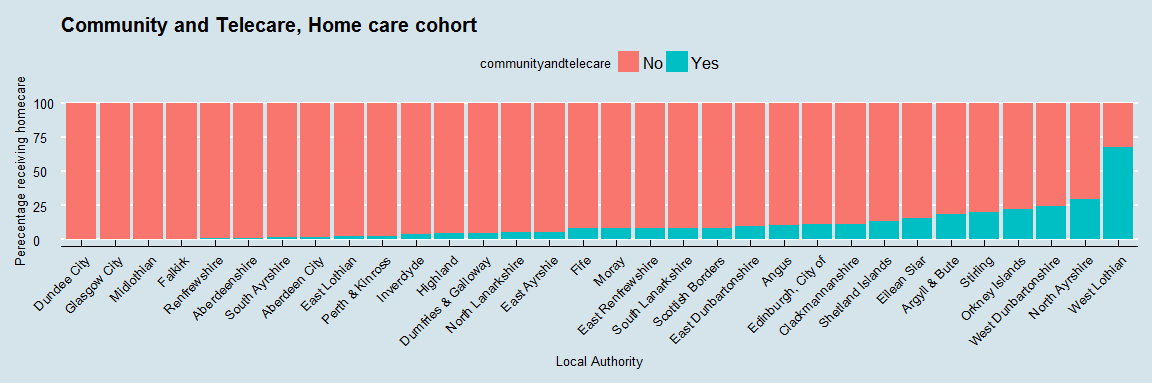
Again, similar distributions across age and gender.

### Distribution by Local Authority



Hmm, really wide variations again - which leads me to worry about quality. Does the provision of this type of service really vary from 0% in some LAs to 67.2% in others??

I'll stack it for easier comparison.



# Session Info

## Session info --------------------------------------------------------------

## setting value   
## version R version 3.3.2 (2016-10-31)  
## system x86\_64, mingw32   
## ui RTerm   
## language (EN)   
## collate English\_United Kingdom.1252   
## tz Europe/London   
## date 2017-03-02

## Packages ------------------------------------------------------------------

## package \* version date source   
## assertthat 0.1 2013-12-06 CRAN (R 3.3.2)  
## backports 1.0.5 2017-01-18 CRAN (R 3.3.2)  
## codetools 0.2-15 2016-10-05 CRAN (R 3.3.2)  
## colorspace 1.3-2 2016-12-14 CRAN (R 3.3.2)  
## cowplot \* 0.7.0 2016-10-28 CRAN (R 3.3.2)  
## DBI 0.5-1 2016-09-10 CRAN (R 3.3.2)  
## devtools 1.12.0 2016-06-24 CRAN (R 3.3.2)  
## digest 0.6.12 2017-01-27 CRAN (R 3.3.2)  
## dplyr \* 0.5.0 2016-06-24 CRAN (R 3.3.2)  
## evaluate 0.10 2016-10-11 CRAN (R 3.3.2)  
## forcats \* 0.2.0 2017-01-23 CRAN (R 3.3.2)  
## ggplot2 \* 2.2.1 2016-12-30 CRAN (R 3.3.2)  
## ggthemes \* 3.3.0 2016-11-24 CRAN (R 3.3.2)  
## gtable 0.2.0 2016-02-26 CRAN (R 3.3.2)  
## htmltools 0.3.5 2016-03-21 CRAN (R 3.3.2)  
## knitr 1.15.1 2016-11-22 CRAN (R 3.3.2)  
## labeling 0.3 2014-08-23 CRAN (R 3.3.2)  
## lazyeval 0.2.0 2016-06-12 CRAN (R 3.3.2)  
## magrittr 1.5 2014-11-22 CRAN (R 3.3.2)  
## memoise 1.0.0 2016-01-29 CRAN (R 3.3.2)  
## munsell 0.4.3 2016-02-13 CRAN (R 3.3.2)  
## plyr 1.8.4 2016-06-08 CRAN (R 3.3.2)  
## R6 2.2.0 2016-10-05 CRAN (R 3.3.2)  
## Rcpp 0.12.9 2017-01-14 CRAN (R 3.3.2)  
## reshape2 1.4.2 2016-10-22 CRAN (R 3.3.2)  
## rmarkdown 1.3 2016-12-21 CRAN (R 3.3.2)  
## rprojroot 1.2 2017-01-16 CRAN (R 3.3.2)  
## scales 0.4.1 2016-11-09 CRAN (R 3.3.2)  
## stringi 1.1.2 2016-10-01 CRAN (R 3.3.2)  
## stringr 1.1.0 2016-08-19 CRAN (R 3.3.2)  
## tibble 1.2 2016-08-26 CRAN (R 3.3.2)  
## withr 1.0.2 2016-06-20 CRAN (R 3.3.2)  
## yaml 2.1.14 2016-11-12 CRAN (R 3.3.2)