Trend in the percentage of the Scottish population prescribed drugs for anxiety, depression or psychosis

Fionnuala Cousins 52319277

19 November, 2023, 21:48

```
# tidyverse includes dplyr and ggplot2 so I don't need to load them separately
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(here)
```

```
## here() starts at C:/Users/Fionnuala/OneDrive - University of Aberdeen/PU5063 Intro to HDS/Assessment
```

```
library(viridis)
```

```
## Loading required package: viridisLite
```

Step 0: Define the question

What are the regional trends for the number of people prescribed drugs for anxiety, depression and psychosis in Scotland over the last ten years? What might these mean for employers' allocation of support resources? The next sections follow the Health Data Science Workflow to address these questions.

Step 1: Data Acquisition

The data was downloaded from the Scottish Public Health Observatory https://scotland.shinyapps.io/ScotPHO_profiles_tool/ on 05/11/23 for the item "population prescribed drugs for anxiety/depression/psychosis" for all available years (2010-2021 inclusive) and all 14 health boards in Scotland, as well as an overall Scotland dataset. The downloaded file was called timetrend_data.csv

```r
#reading in the data and checking its columns:
adp_data <- read_csv(here("Inputs/timetrend_data.csv"))
```

```
## Rows: 180 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): indicator, area_name, area_code, area_type, period, definition, dat...
## dbl (5): year, numerator, measure, lower_confidence_interval, upper_confiden...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
glimpse(adp_data)
```

```
## Rows: 180
## Columns: 12
## $ indicator                <chr> "Population prescribed drugs for anxiety/dep~
## $ area_name                <chr> "Scotland", "NHS Ayrshire & Arran", "NHS Bor~
## $ area_code                <chr> "S00000001", "S08000015", "S08000016", "S080~
## $ area_type                <chr> "Scotland", "Health board", "Health board", ~
## $ year                     <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 20~
## $ period                   <chr> "2010/11 financial year", "2010/11 financial~
## $ numerator                <dbl> 787040, 60822, 17226, 22280, 55334, 43976, 7~
## $ measure                  <dbl> 14.96, 16.31, 15.15, 14.75, 15.26, 14.86, 12~
## $ lower_confidence_interval <dbl> 14.93, 16.20, 14.94, 14.57, 15.14, 14.73, 12~
## $ upper_confidence_interval <dbl> 14.99, 16.43, 15.36, 14.92, 15.38, 14.98, 12~
## $ definition               <chr> "Percentage", "Percentage", "Percentage", "P~
## $ data_source              <chr> "Public Health Scotland (Prescribing Informa~
```

Step 2: Prepare/ clean data

Is it tidy?

Does each variable form a column?

Yes, for example, Year is a variable so has a single column rather than a column for each year.

Does each observation form a row?

Yes. While there are multiple columns, they are each just clarifying the observed value rather than additional observations themselves (for example, confidence intervals).

Is each cell a single value?

Yes, for example, there are no commas separating values. The period column data is in the format "YYYY/YYYY financial year" but this is still a single value because financial years always span two sequential calendar years.

```r
unique_columns <- c("area_code", "year")
duplicates <- adp_data[duplicated(adp_data[, unique_columns]) | duplicated
                       (adp_data[, unique_columns], fromLast = TRUE), ]
print(duplicates)
```

```
## # A tibble: 0 x 12
## # i 12 variables: indicator <chr>, area_name <chr>, area_code <chr>,
## #   area_type <chr>, year <dbl>, period <chr>, numerator <dbl>, measure <dbl>,
## #   lower_confidence_interval <dbl>, upper_confidence_interval <dbl>,
## #   definition <chr>, data_source <chr>
```

```r
# This chunk is for selecting and renaming columns and removing the NHS prefix.

clean_data <- adp_data %>%
  select('area_name','year','numerator', 'measure') %>%
  rename(no_prescriptions = 'numerator', NHS = 'area_name',
         percentage_pop = 'measure') %>%
  mutate(NHS = sub("^NHS ","", NHS)) %>%

  # the below line is necessary to calculate the national population for
# recalculating percentages later
  mutate(board_population = no_prescriptions / percentage_pop * 100)
glimpse(clean_data)
```

```
## Rows: 180
## Columns: 5
## $ NHS              <chr> "Scotland", "Ayrshire & Arran", "Borders", "Dumfries ~
## $ year             <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010,~
## $ no_prescriptions <dbl> 787040, 60822, 17226, 22280, 55334, 43976, 70337, 187~
## $ percentage_pop   <dbl> 14.96, 16.31, 15.15, 14.75, 15.26, 14.86, 12.45, 16.6~
## $ board_population <dbl> 5260962.57, 372912.32, 113702.97, 151050.85, 362608.1~
```

Step 3: Analyse

14 Health Boards are too many to plot in the same visualisation; the audience would be overwhelmed. The boards need to be merged into a small number of groups. This will also mean that their values need to recalculated so each group can be understood as a percentage of the population of Scotland.

```r
#This chunk groups the boards into regions
group_data <- clean_data %>%
  mutate(Region = case_when(
    NHS %in% c("Ayrshire & Arran" , "Borders" , "Dumfries & Galloway")
    ~ "Borders",
    NHS %in% c("Fife" , "Forth Valley" , "Greater Glasgow & Clyde" ,
               "Lanarkshire" , "Lothian")
    ~ "Central Belt",
    NHS %in% c("Grampian" , "Tayside")
    ~ "North East",
    NHS %in% c("Highland" , "Western Isles" , "Orkney" , "Shetland")
    ~ "Highlands & Islands",
    NHS %in% c("Scotland") ~ "Scotland"))
glimpse(group_data)
```

```
## Rows: 180
## Columns: 6
## $ NHS              <chr> "Scotland", "Ayrshire & Arran", "Borders", "Dumfries ~
## $ year             <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010,~
## $ no_prescriptions <dbl> 787040, 60822, 17226, 22280, 55334, 43976, 70337, 187~
## $ percentage_pop   <dbl> 14.96, 16.31, 15.15, 14.75, 15.26, 14.86, 12.45, 16.6~
## $ board_population <dbl> 5260962.57, 372912.32, 113702.97, 151050.85, 362608.1~
## $ Region           <chr> "Scotland", "Borders", "Borders", "Borders", "Central~
```

```r
# The below is all one chunk because it would not work in separate chunks.
# This was explained by ChatGPT (2023)

# This code was difficult to arrive at and could be reworked.

# The below creates a total number of prescriptions per region
regional_data <- group_data %>%
  group_by(Region, year) %>%
  summarise(board_population = sum(board_population), regional_prescriptions =
              sum(no_prescriptions))
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

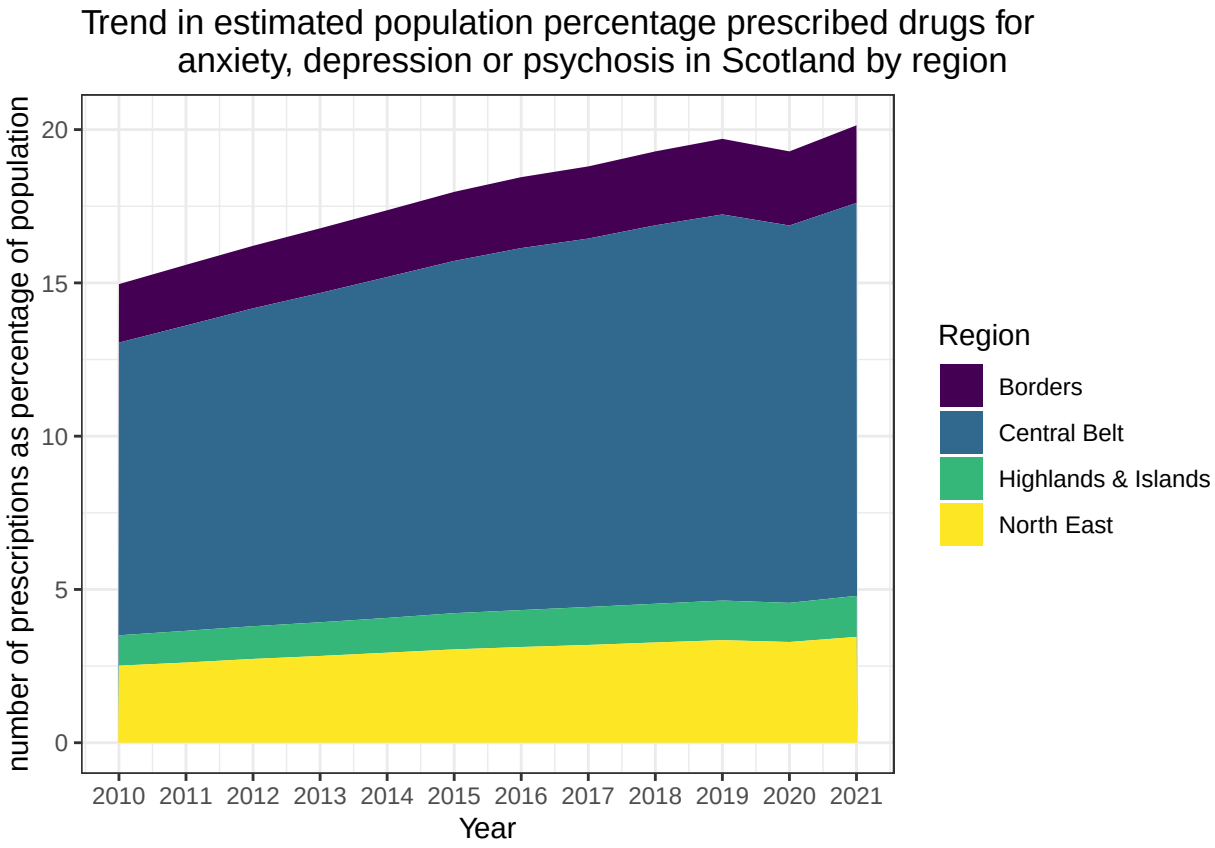
```r
# The below creates a tibble for the annual calculated Scottish population
scotland_population <- regional_data %>% filter(Region == "Scotland") %>%
  rename(whole_population = board_population, scotland_prescriptions =
           regional_prescriptions)

# The below creates each regional number of prescriptions as a percentage of the
#scottish population for that year.
plot_data <- regional_data %>%
  left_join(scotland_population, by = 'year') %>%
  mutate(percentage_scot_pop = regional_prescriptions / whole_population * 100)
glimpse(plot_data)
```

```
## Rows: 60
## Columns: 8
## $ Region.x              <chr> "Borders", "Borders", "Borders", "Borders", "Bo~
## $ year                  <dbl> 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017,~
## $ board_population      <dbl> 637666.1, 639032.5, 637888.7, 636439.7, 635120.~
## $ regional_prescriptions <dbl> 100328, 104934, 108404, 112404, 116487, 121085,~
## $ Region.y              <chr> "Scotland", "Scotland", "Scotland", "Scotland",~
## $ whole_population      <dbl> 5260963, 5298679, 5314503, 5328123, 5347916, 53~
## $ scotland_prescriptions <dbl> 787040, 826064, 861481, 894059, 928933, 965638,~
## $ percentage_scot_pop   <dbl> 1.9070274, 1.9803805, 2.0397767, 2.1096361, 2.1~
```

Step 4: Communication

The final step is to generate the visualisation. This graph shows the increase in prescriptions for anxiety, depression and psychosis across Scotland from 2010-2021, broken down into four broad regions of Scotland.

```r
plot_data %>%
  filter(Region.x != "Scotland") %>%
  ggplot(aes(x = year, y = percentage_scot_pop, fill = Region.x)) +
  geom_area()+
  labs(fill="Region")+
  xlab("Year")+
  ylab("number of prescriptions as percentage of population")+
  ggtitle("Trend in estimated population percentage prescribed drugs for
          anxiety, depression or psychosis in Scotland by region")+
  scale_fill_viridis(discrete=TRUE)+
  scale_x_continuous(breaks = unique(plot_data$year))+
  theme_bw()
```



References

I acknowledge the use of ChatGTP (chat.openai.com) to troubleshoot coding issues on this assignment.

I entered the following prompt on 7 November 2023

> I am using ggplot to plot a stacked area chart. the plot is showing only select years' as ticks on the x-axis. how do i get it to label all the years? This is my ggplot code options(scipen = 999) summed_data %>% ggplot(aes(x = year, y = total_people, fill = Region)) + geom_area()+ xlab("Year")+ ylab("Estimated number of people")+

> ggtitle("Trend in estimated population prescribed drugs for anxiety, depression or psychosis by Scottish region")+ scale_fill_viridis(discrete=TRUE)+ theme_bw()

The content generated from this prompt was used to add all year labels to the axis.

I entered the following prompt on 19 November 2023:

> In the following sequence of coding, why am I getting an error that "object 'regional_prescriptions' not found" at the end, when that object is used earlier in the code? "'{r} #This chunk creates a total number of prescriptions per region regional_data <- group_data %>% group_by(Region, year) %>% summarise(board_population = sum(board_population), regional_prescriptions = sum(no_prescriptions)) view(regional_data)

```
>>#This chunk creates a tibble for the annual calcuated Scottish population
scotland_population <- regional_data %>% filter(Region == "Scotland") %>%
  rename(whole_population = board_population)
view(scotland_population)
```

> plot_data <- regional_data %>% left_join(scotland_population, by = 'year') %>% mutate(percentage_scot_pop = regional_prescriptions / whole_population * 100) view(plot_data) Provide information on the ethical considerations when using generative artificial intelligence tools

The content generated from this prompt was used to fix errors in the code. The code was then further editeed.

PU5063 Report Discussion

This discussion document accompanies the PDF Report "Trend in the percentage of the Scottish population prescribed drugs for anxiety, depression or psychosis".

Target Audience and Key Message

The target audience of the report is employers in Scotland. Adopting the perspective of employers, a trend is evident in the estimated population prescribed drugs for anxiety, depression or psychosis. Small or regional employers looking at a health board may see the increase over the last ten years as subtle and presenting too weak a case for increasing their allocation of staff support resources. However, viewed at the country level, the gentle increases in each area accumulate to a clear national increase of about a third since 2010 (about 15% in 2010 to above 20% in 2021). This presents a much stronger case for employers to increase the allocation of support resources and this is the key message of the report. Employers operating in single regions should review their data against not just the regional but also the national picture, and employers operating across multiple areas may want to ensure that their monitoring is standardised across branches so any need for targeted resources is identified early.

This data might also be of interest to HR consultancy and service provider companies, as it is common for larger employers to contract suppliers for the provision of some types of staff support resources.

Why is this important and how can data science help address it?

Scotland's employment rate for people aged 16-64 was estimated as 75.2% in 2022 (Scottish Government, 2022), in a population of 5.4 million (National Records of Scotland, 2023). 17.7 million work days were lost to sickness absence in Scotland in 2022 (Office for National Statistics, 2023). The sickness absence rate across the UK had been trending downwards for over 20 years until 2020, when it rose sharply (though, of course, COVID-19 was a contributing factor) (ibid). Further, the proportion of people who are not working and not seeking work is increasing in Scotland and is above the UK rate (Scottish Government, 2022). Providing support so those in work so they are healthy enough to stay there will be important in countering this trend.

As part of human resource management, employers typically provide various resources to their permanent staff, including health support, and within that, mental health resources. Employers, particularly private sector employers, need strong business cases in order to justify allocating money that could be spent elsewhere or kept as profit. The mildness of the regional trends may not make for a strong enough case, particularly for employers who are limited to a single region. These employers may not be aware of the wider trend because employers typically only have access to their own HR (and other) data, constraining the size of their dataset and limiting their ability to differentiate statistically significant changes from randomness. The availability of anonymised, aggregated data at regional and national levels allows for individual employers to review their data against a much larger dataset. However, this doesn't mean that they have the resources to do this, including time and competence. Data science and data scientists can help them with this by sourcing,

interpreting and visualising this wider data. Further, by sharing the code used to do this, any HR staff interested in advancing their own analytics skills can use the code as a learning resource. Data scientists might do this as individuals but also might do it from within stakeholder organisations, such as Scottish Enterprise, Public Health Scotland, Skills Development Scotland or the Chartered Institute of Personnel and Development.

Data used to produce this visualisation

The data used to produce the visualisation is from the Scottish Public Health Observatory (PHO). It is the last 12 years of available data on the estimated percentage of the population prescribed medication for anxiety, depression or psychosis. The data is reported at both council area and health board level. No datapoints are missing and the dataset covers more than ten years, enough to illustrate the clear trend in increasing prescriptions.

There are a small number of limitations with this data, none of which are sufficient to invalidate its use. It does not include population totals per health board. The data is estimated and the estimation method is not shared. It is aggregated across anxiety, depression and psychosis. These are different conditions and warrant different approaches from employers, but the prevalence of each is unclear. Further, it is unclear if the data is from counts of diagnoses of these conditions, or counts of prescriptions of drugs commonly (but not exclusively) prescribed for these conditions. As the data shows its original source as the National Prescription Information System, it is more likely to be the latter. So, it does not include people who are undiagnosed, people who are diagnosed but not medicated, or people with other mental health conditions. Also, it may include people who have been prescribed these medications for other conditions, such as bipolar disorder. However, all of these groups are still likely to benefit from mental health support from their employers, so, the visualisation may have wider beneficiaries.

So, employers should see this data as one of many sources to take into account as they consider further investing in their staff mental health support strategies.

Strengths and limitations of the approach used for the visualisation

The primary strength of the approach is its simplicity. Four categories with different area colours makes good use of pre-attentive attributes; the viewer grasps the main message before they are conscious of it. This uses the viridis package to ensure that colours are not incompatible with colourblindness.

A limitation of the approach is that there is no pre-existing classification of the Health Boards (or council areas) into a smaller number of bigger regions. So, for example, while it is likely that people in Scotland would assume that NHS Grampian is classified under North East, they may be less likely to assume that NHS Ayrshire and Arran is classified under Borders. This could be addressed in an interactive visualisation with a tooltip that lists which NHS Boards are in which region. However, as the output is a PDF, this is not possible here.

The code was made more complicated by the need to recalculate summed regional numbers of prescriptions as percentages of the Scottish population. If they had been left as counts, the visualisation would have been ambiguous; was the prevalence of these prescriptions really increasing

or was it just that the population was increasing? The code used for this calculation was challenging to develop and is likely to be optimizable with more experience.

References

National Records of Scotland (2023). *Scotland's census first results*. Available from: https://www.nrscotland.gov.uk/news/2023/scotland%E2%80%99s-census-first-results [Last accessed 04/11/2023]

Office for National Statistics (2023). Sickness absence in the UK labour market: 2022. Available from: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/labourproductivity/articles/sicknessabsenceinthelabourmarket/2022 (Accessed 19/11/23)

Scottish Government (2022). Labour market trends: September 2022. Available from: https://www.gov.scot/publications/labour-market-trends-september-2022/ (Accessed 18/11/23)