Exploratory Data Analysis

AS

5/21/2021

1 Introduction

Before visualising the data a check was performed to ensure all local_authority population entries in the combined population/death tibble (df_deaths_pop) contained a value greater than 0.

2 EDA deaths

The individual variables year, age, cause of death, locality and sex of deaths were explored using visualisations. In addition, combination of variables were also investigated. In order to achieve this summary measures for each variable and variable combinations were obtained and stored in R objects (see 02-EDA.Rmd file for code).

As there were 347 localities investigated in this dataset, only 10 with the most frequency of deaths for both males and females were displayed in a plot. An R object was created to store this data (see 02-EDA.Rmd file). Table 1 displays this data.

Overall Birmingham contained the largest frequency of deaths for both males and females. For later analysis of death rates, Birmingham was used as the reference region that other local_authorities were compared against.

Individual plots were constructed (see 02-EDA.Rmd for code) and saved in R objects. They are later used in this report using the R package "patchwork".

3 Plots

Figure 1 A demonstrated that increased age was associated with increased deaths. This would be expected. Very few deaths occur before the age group 40-44. Also noted in this figure was the small decline in deaths across the older age groups from years 2012 to 2014. When stratifying by sex (figure 1 B) it was interesting to note that the peak age group of male deaths was 80-84, after which there was a decline. This may be due to a reduced male population at these age groups, and therefore less males would have been likely to die.

Figure 2 A demonstrated around twice the number of deaths due to cardiovascular disease than coronary heart disease, and there were fewer stroke deaths than coronary heart disease deaths. The number of cardiovascular disease deaths in England in 2012 was slightly under 132,000. There was a slight decrease in deaths for all groups from 2012 to 2014. Figure 2 B demonstrated an interesting feature. Males were likely to suffer from greater coronary heart disease deaths than females. However, females were more likely to die of stroke than males.

Figure 3 A demonstrated the top 10 localities where deaths are highest in England. Birmingham experienced the highest number of deaths by a considerable margin compared to the locality with the next highest frequency of deaths, Leeds. Apart from Cornwall (3rd) and Wiltshire (7th), all the remaining eight localities were located in the Midlands and north of England. Stratifying these figures according to sex (figure 3 B),

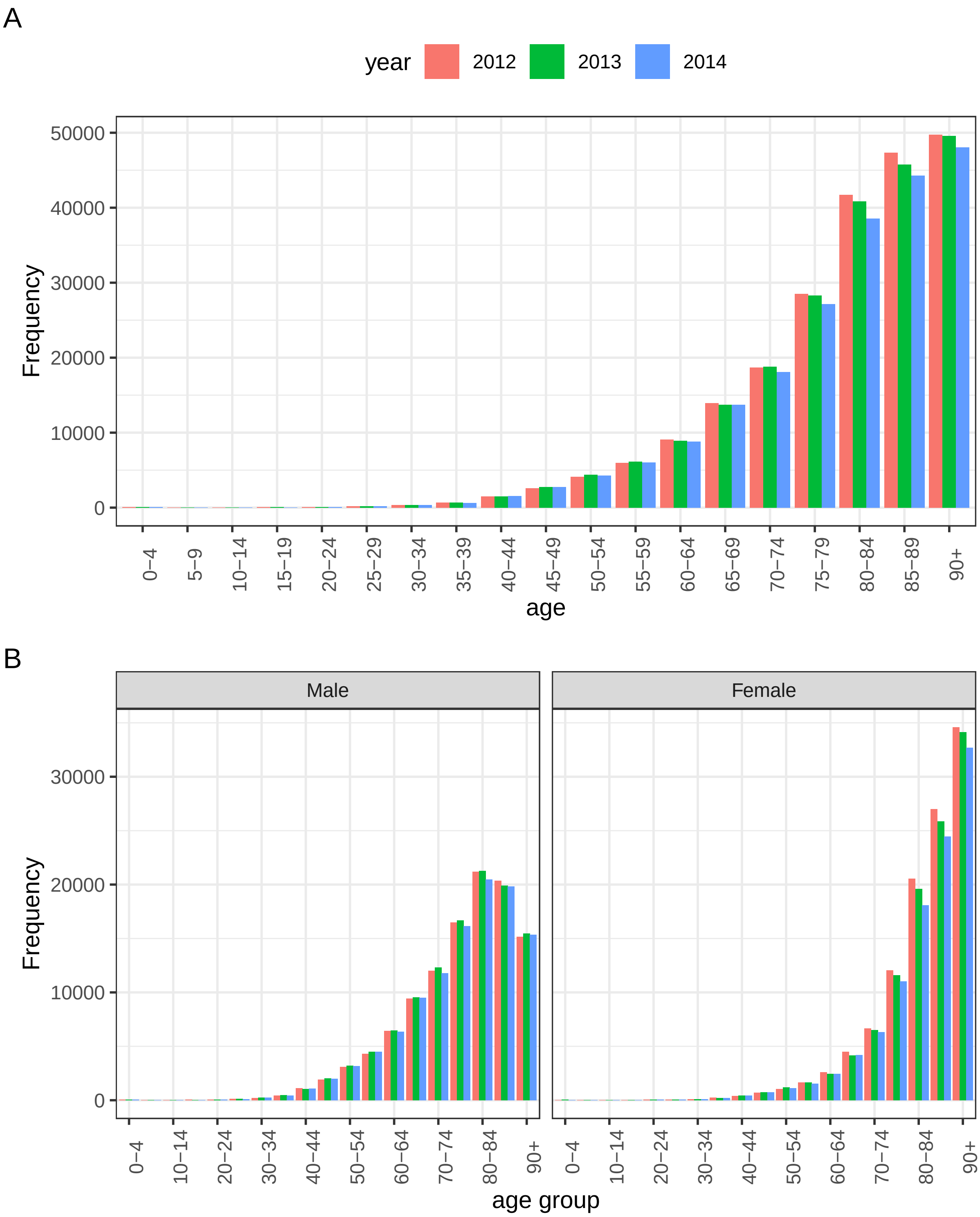

Figure 1: The effect of age and sex on frequency of death

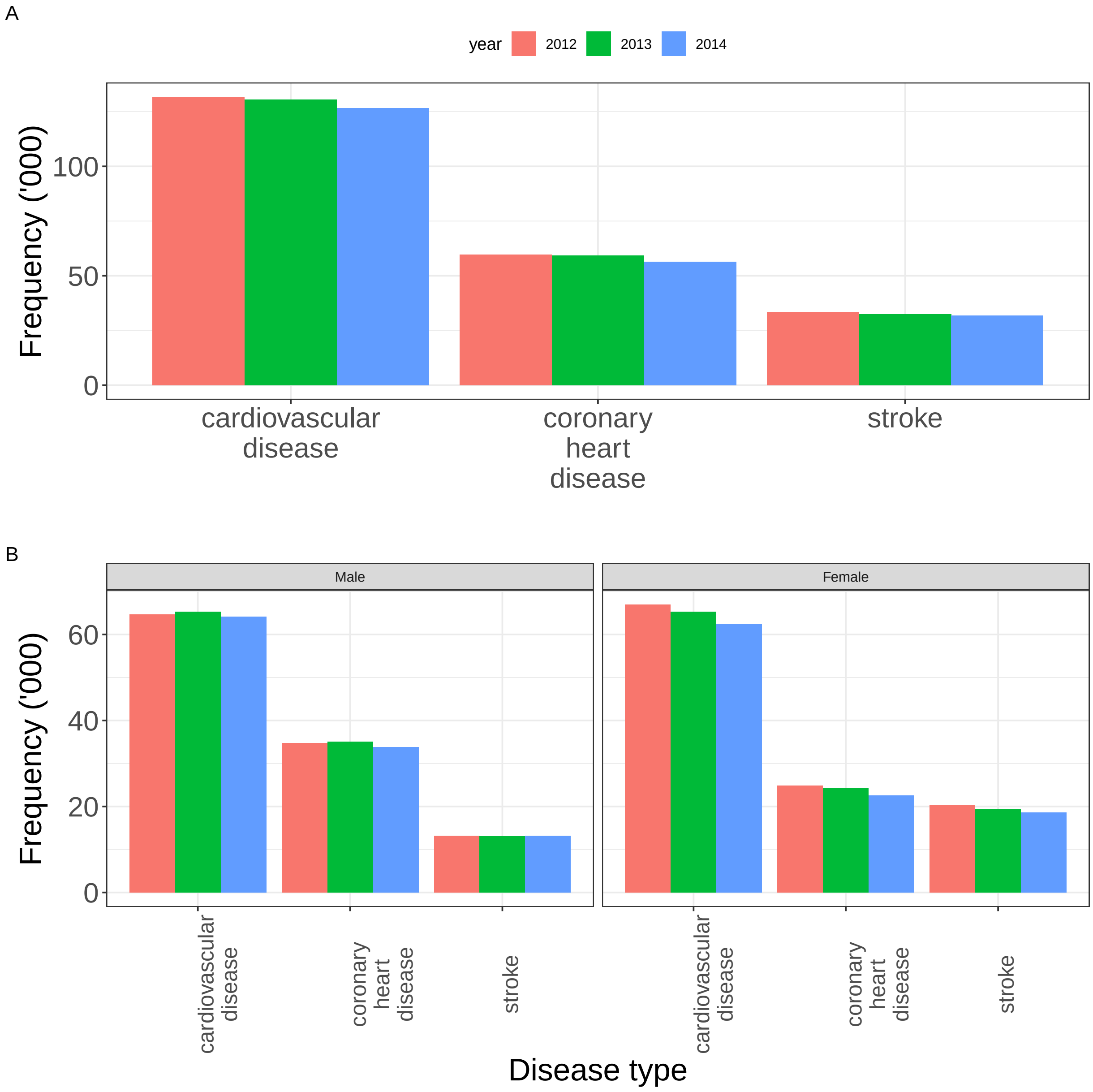

Figure 2: The effect of disease type and sex on frequency of deaths in England

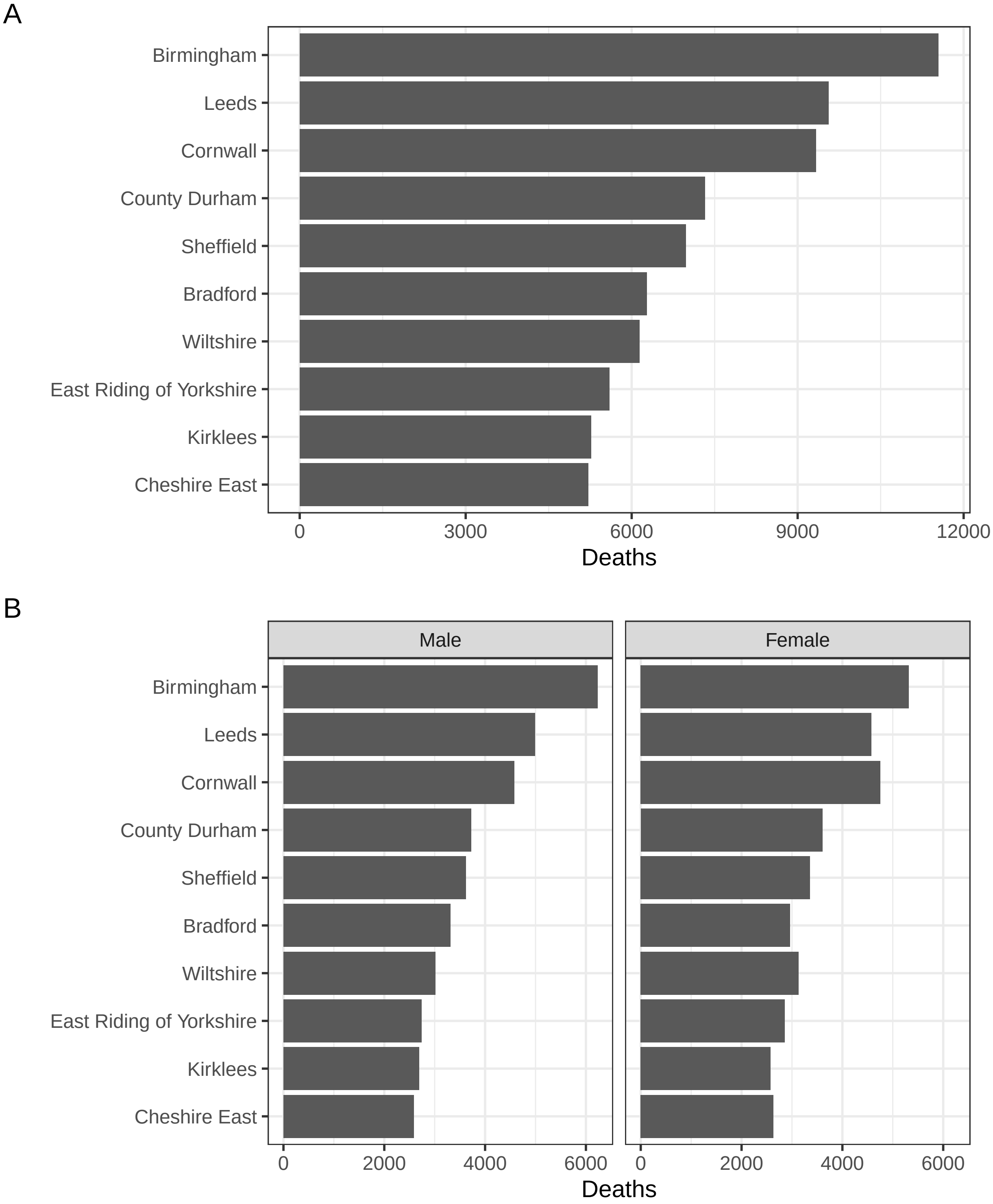

Figure 3: Local authorities with the top 10 frequencies of deaths in England

Table 1: The top ten Local Authorities with largest overall deaths due to cardiovascular disease, coronary artery disease or stroke, stratified by sex.

local_authority	sex	freq
Birmingham	Male	6229
Birmingham	Female	5317
Leeds	Male	4986
Cornwall	Female	4754
Cornwall	Male	4576
Leeds	Female	4574
County Durham	Male	3720
Sheffield	Male	3619
County Durham	Female	3605
Sheffield	Female	3356
Bradford	Male	3309
Wiltshire	Female	3132
Wiltshire	Male	3009
Bradford	Female	2961
East Riding of Yorkshire	Female	2858
East Riding of Yorkshire	Male	2740
Kirklees	Male	2693
Cheshire East	Female	2632
Cheshire East	Male	2582
Kirklees	Female	2576

more females died of the three types of diseases in Cornwall, Wiltshire and East Riding of Yorkshire. In the seven other regions more males died than females.

According to figure 4 there were greater female deaths Birmingham due to cardiovascular disease in the 85-89 and 90+ age group. Otherwise males deaths were greater for age groups between 40-44 and 80-84. Lower number of male deaths for coronary artery disease were recorded in the 90+ age group, while for age groups between 40-44 and 85-90 it appeared more male deaths occured compared to femailes. Likewise, there appears an increase in coronary heart disease in females in the 90+ age group. Interestingly, more females are likely to die of stroke compared to males for age groups 75-79 and above.

One reason for the earlier age_group deaths of males in the industrial cities could be due to the historic work environments.

Figure 5 indicates changes in deaths per year for the local_authorities with the highest frequencies of deaths. Frequencies in Birmingham, Leeds , Cheshire East and Sheffield appear to have decreased year on year. In Durham, Cornwall and Bradford, deaths appeared to increase in 2013 but then decreased in 2014. For East Riding of Yorkshire, Kirkless and Wiltshire, death frequencies appeared stable over the three years.

3.1 Raincloud plots

Cedric Scherer has created new type of plot that encompasses useful information from a boxplot and combined it with a density and jitter plot. This "raincloud" plot incorporates the distribution of data seen in the density plot, but also the key summary statistics seen in the box plot and jitter plot.

It can be seen from figure @6 that data for each of the three years appears was similarly distributed. The data for each distribution was positively skewed ie. most localities had frequency values closer to 0 than the maximum value. It was therefore worthwhile log transforming the data.

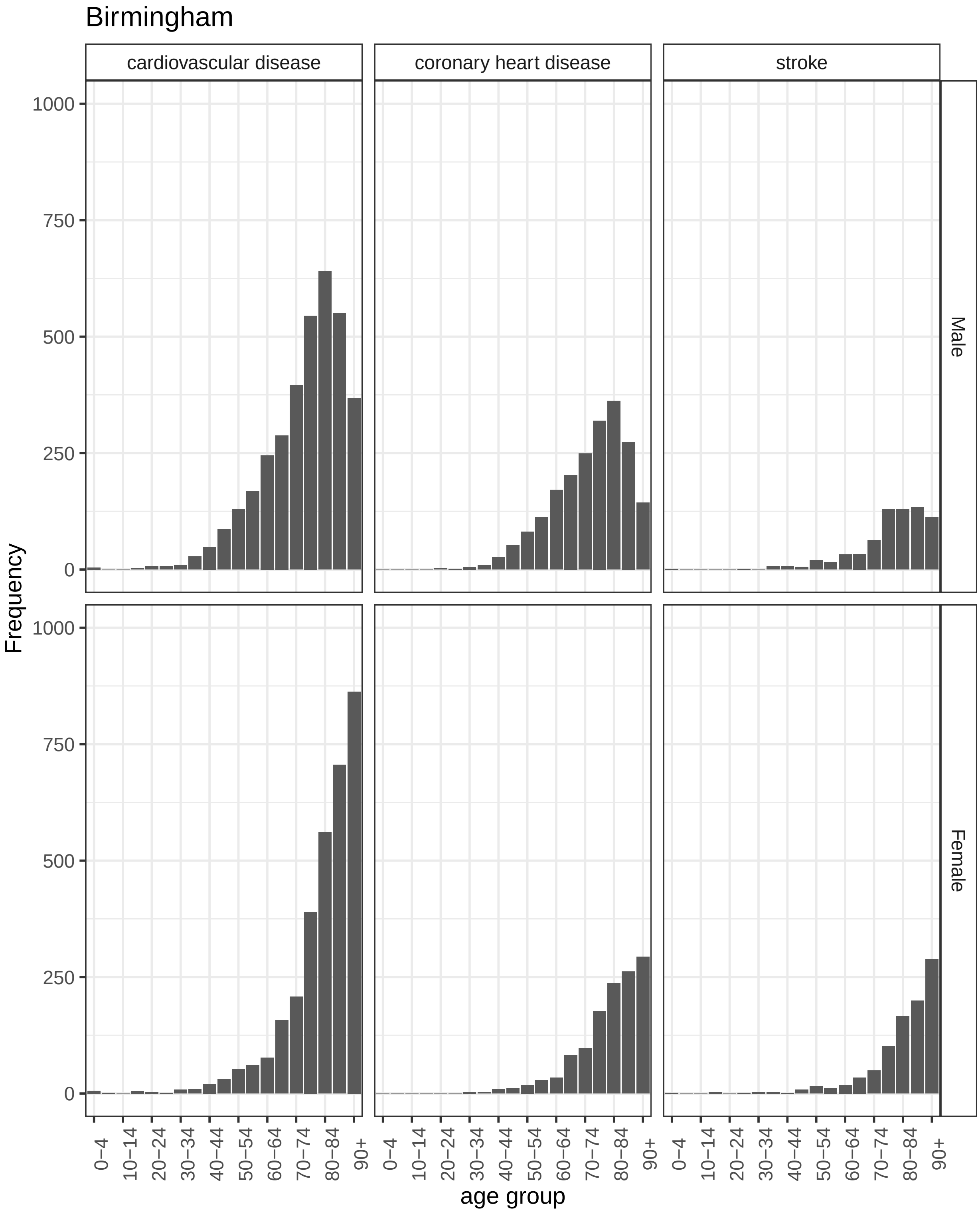

Figure 4: Deaths in Birmingham for each sex caused by the three types of diseases

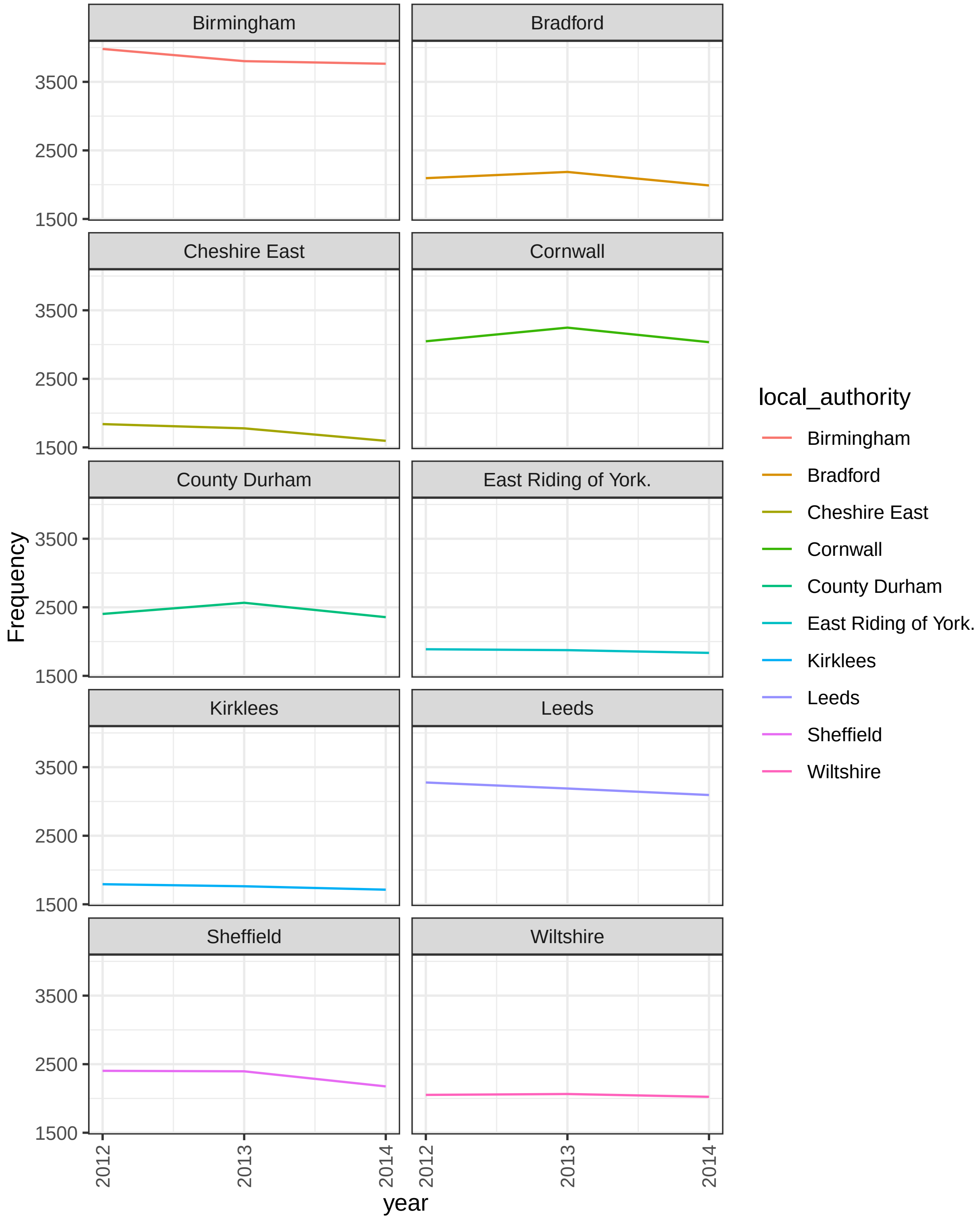

Figure 5: Changes in frequency of deaths in localities with highest amounts of deaths

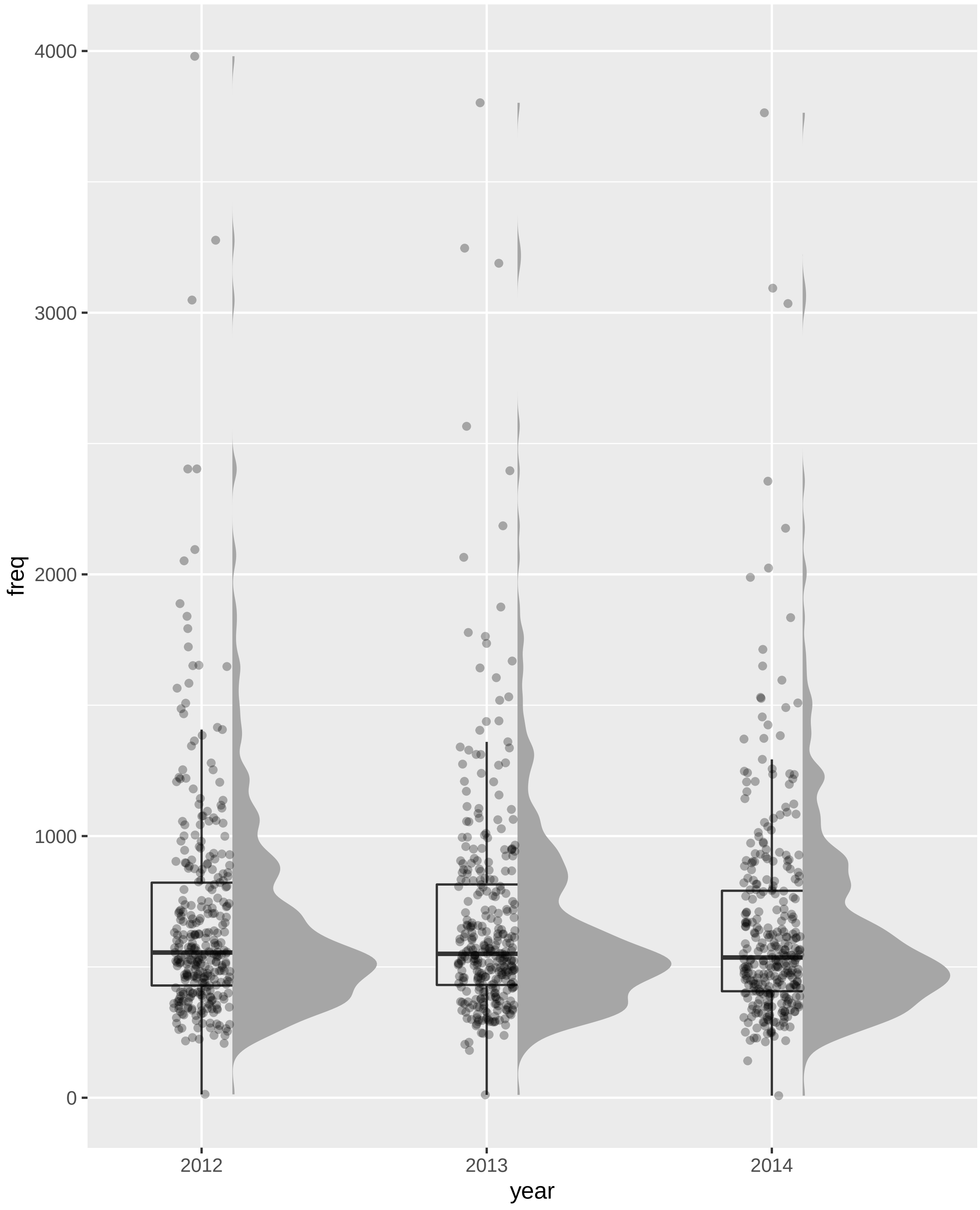

Figure 6: Raincloud plots for the distibution of deaths in localities for the three years 2012-14. All distributions are positively skewed.

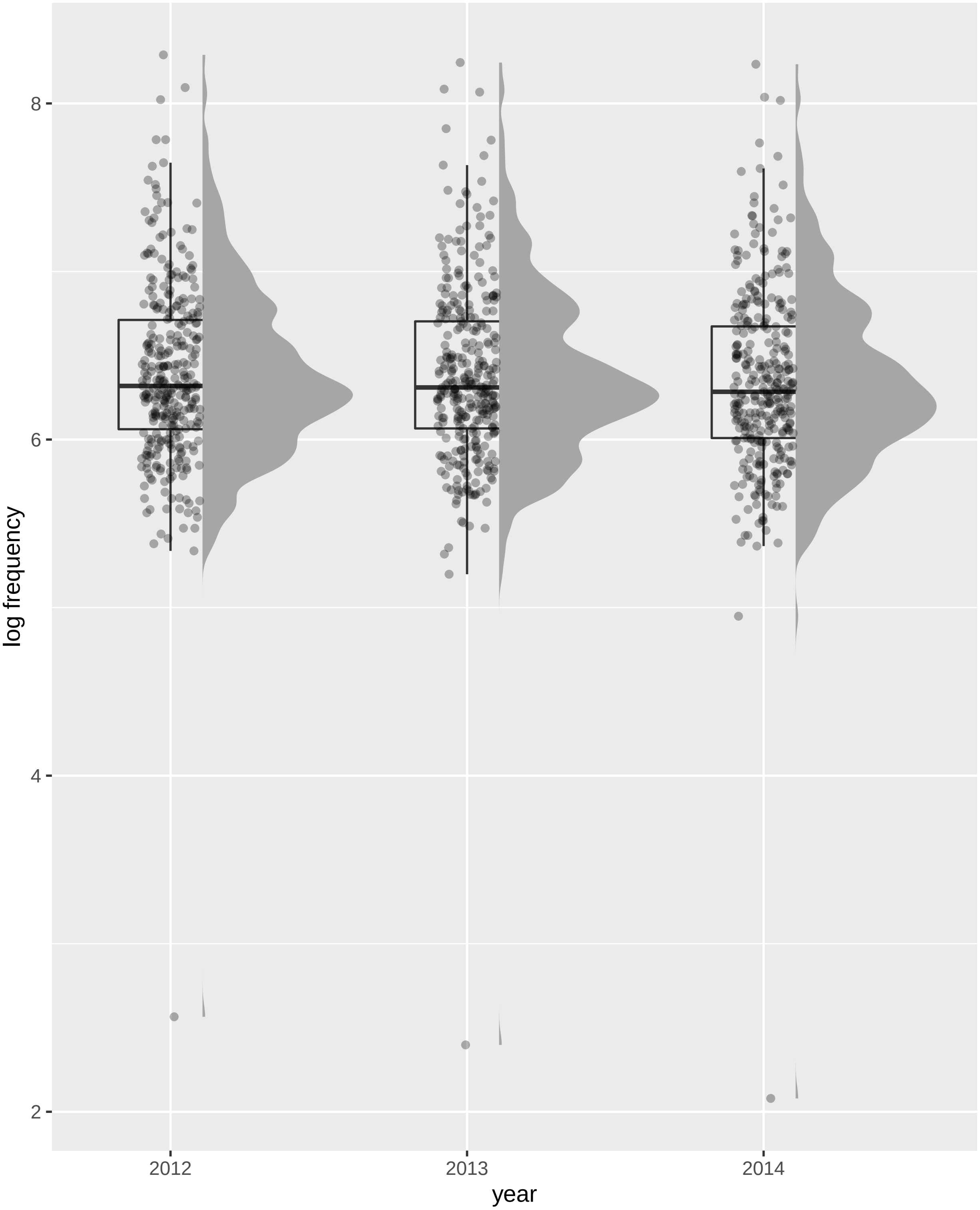

Figure 7: Raincloud plots for the distibution of log deaths in localities for the three years 2012-14. The distributions better resemble gaussian compared to the non-log deaths distributions.

The data in figure @7 was log-transformed. The distributions resembled gaussian envelopes. Apart from outliers notes with large log frequency values, there was an outlier with a log frequency value close to 0. This outlier represented data from the Isles of Scilly. It was noted that many values for the Isles of Scilly were not recorded (NA). Perhaps the low log frequency from this region was due to missing data. For this reason, the Isles of Scilly data was dropped from this analysis.

In figure @8 removal of the Isle of Scilly datapoints resulted in distributions which resembled gaussian envelopes, although still slightly positively skewed. This would be in keeping with poission (count) data, where normally the output variable is log transformed in order to fit a linear model.

Looking at the relationship between age group and type of death (figure 9), it was seen that deaths increased in an exponential manner for age groups in both cardiovascular disease and stroke. However, it was noticeable that there was a slight decrease in deaths for coronary heart disease in the age group 90+. When stratifying for sex, it was seen in figure 10 that deaths due to cardiovascular disease and stroke was much greater in females than males at older age group. This would likely be due to the longer life-span of females. More males than females died of coronary heart disease at age groups younger than 90+, possibly due to poorer lifestyle. At the age_group 90+ many more women died than men, likely due to more women living at this age.

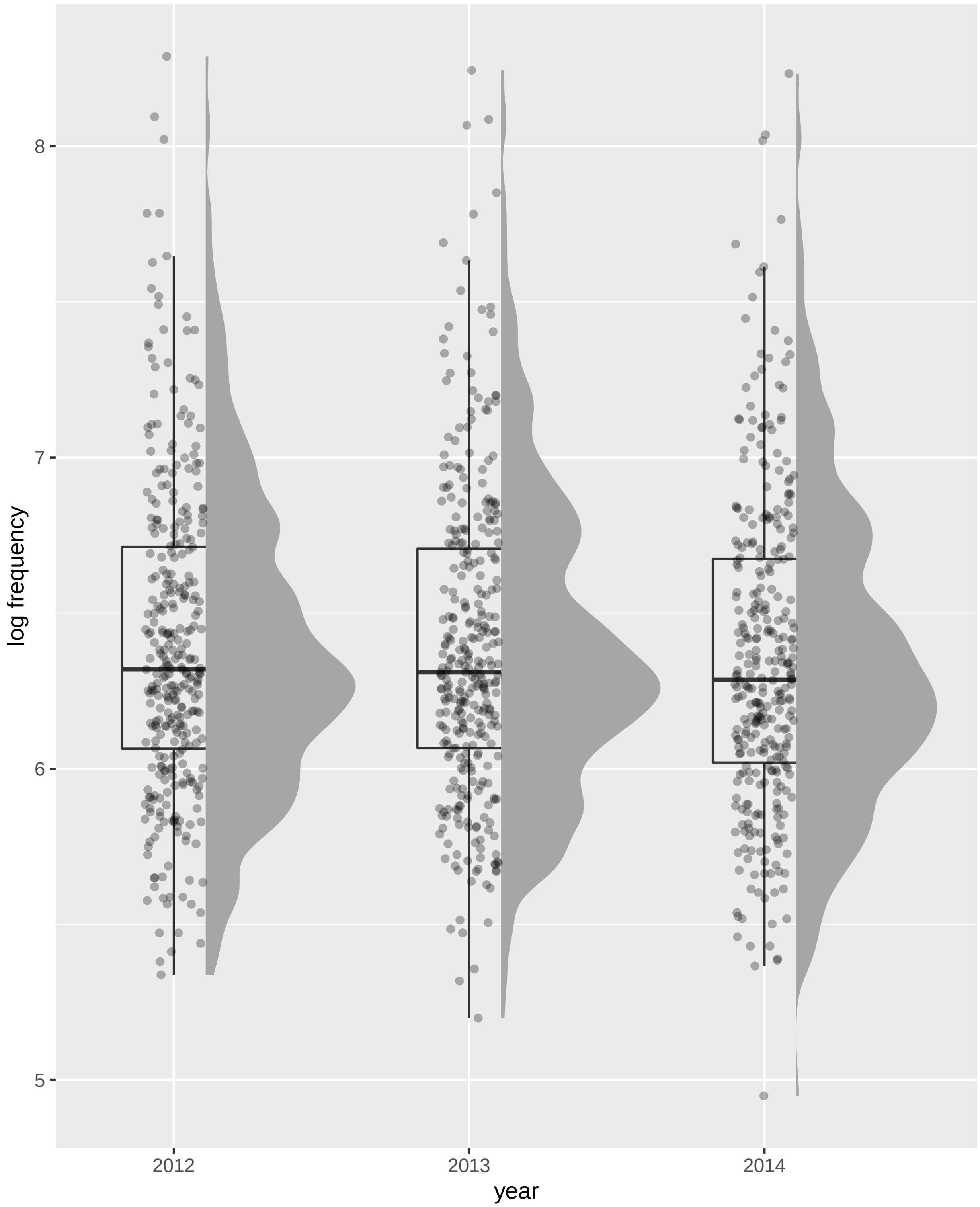

Figure 8: Raincloud plots for the three years 2012-14. The frequencies have been log transformed and the Isle of Scilly datapoints have been dropped. The distributions resemble gaussian envelopes.

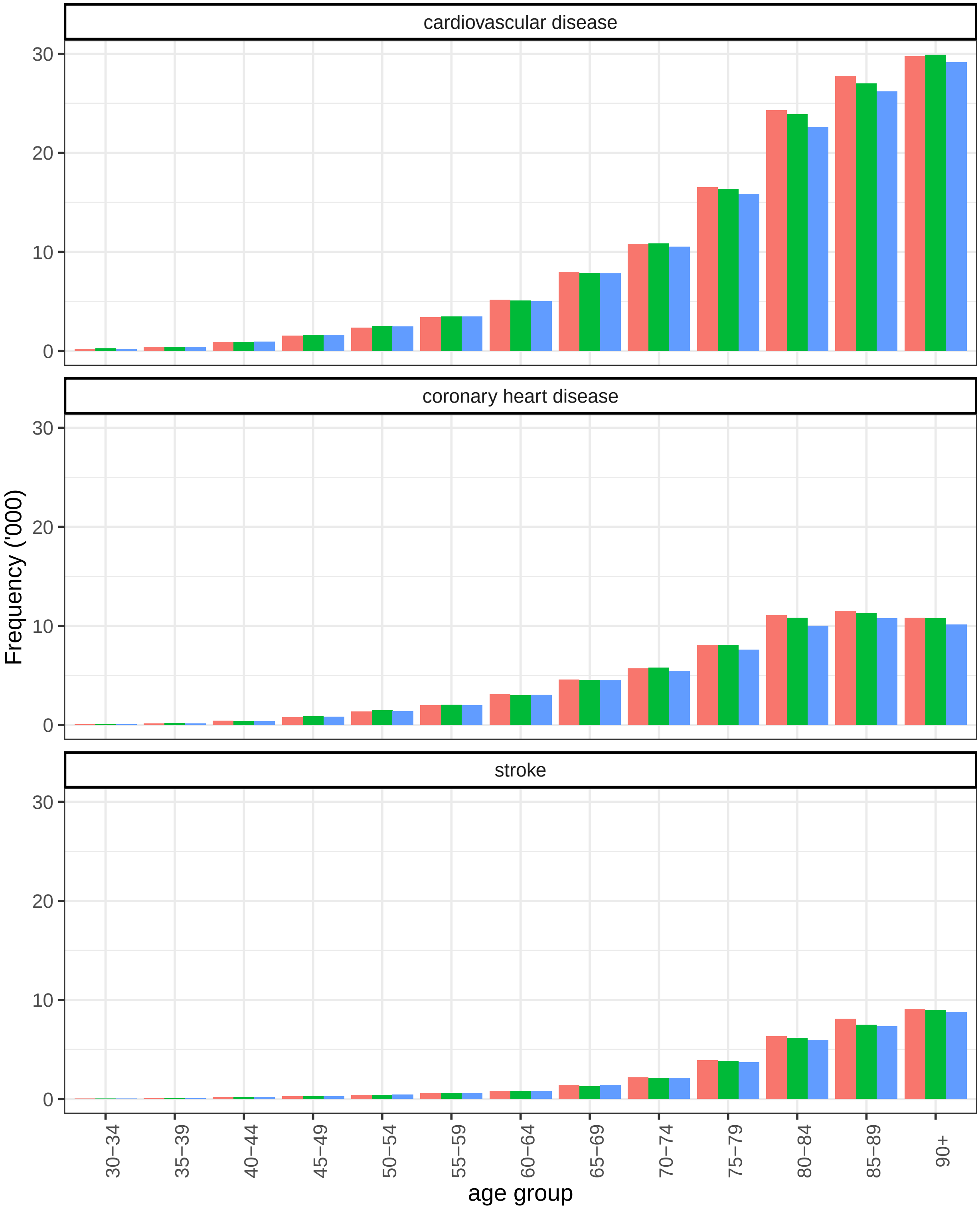

Figure 9: Deaths due to the three types of diseases investigated in England for age groups 30-34 and greater

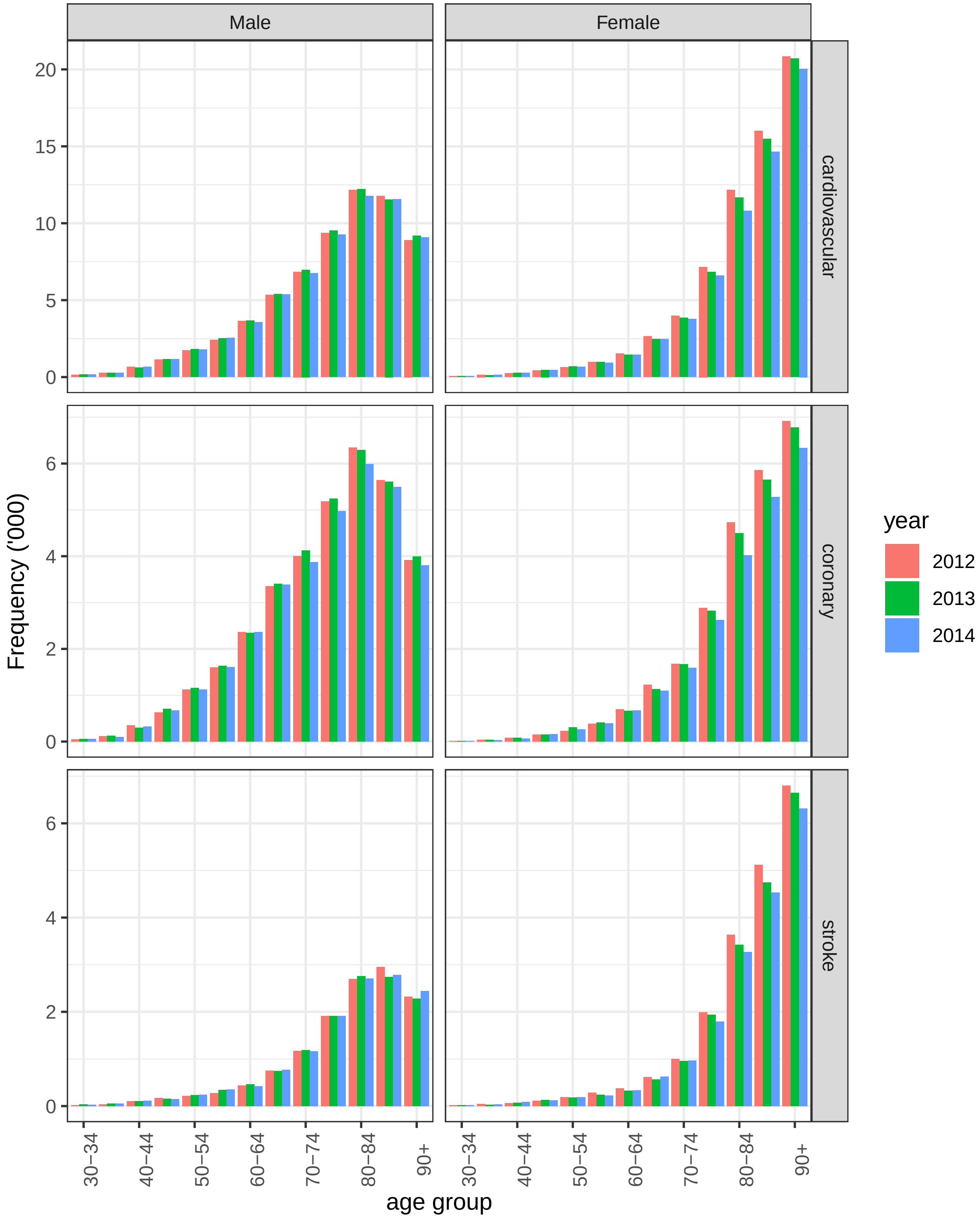

Figure 10: Deaths due to disease type between sexes in England