GDA of *happy* (from **GGally**)

Background A small selection of variables related to happiness from the General Social Survey (GSS).

Aims How do age, health, marital status and other factors influence people's opinions on how happy they are? Has this changed over the years?

Source Smith, Tom W., Peter V. Marsden, Michael Hout, Jibum Kim. General Social Surveys, 1972-2006

Structure 51020 observations on 10 variables (2 discrete variables, 6 factors, 1 weight variable, 1 ID)

The distribution of how happy respondents said they were is shown in Figure 1.

```r
data(happy, package="GGally")
ggplot(happy, aes(happy)) + geom_bar() + labs(x=NULL, y=NULL)
```

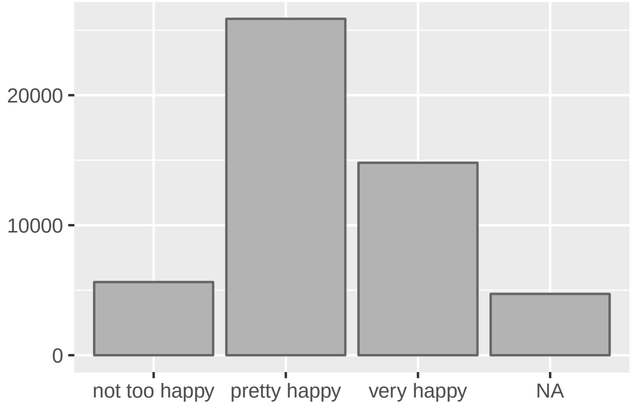

Fig 1: Most people said they were 'pretty happy'. There were a non-negligible number of NAs (missing values).

The distributions of two of the other variables are shown in Figure 2.

```r
library(dplyr)
ggplot(happy, aes(health)) + geom_bar() + labs(x=NULL, y=NULL)
ggplot(na.omit(happy %>% select(marital)), aes(marital)) + geom_bar() + labs(x=NULL, y=NULL)
```

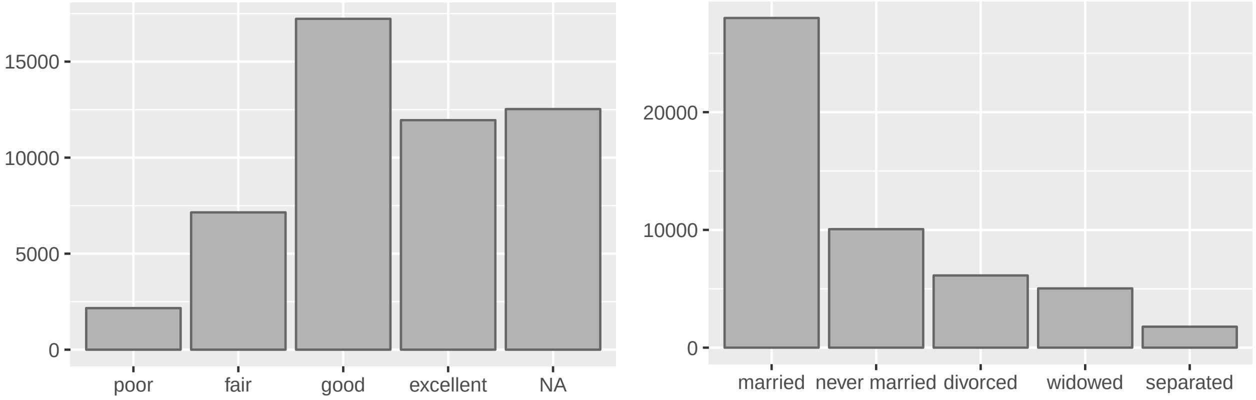

Fig 2 Most people in the survey regarded themselves as being in good or excellent health (although there were a surprisingly large number of NAs). The majority of respondents were married. For this variable there were only a few NAs, so they were left out of the plot to make room for the labels.

Figure 3 shows how the distribution for happy changes over time in two different ways for two slightly different versions of the data. For the plot on the left, a factor version of year is used to produce the stacked barcharts. The horizontal axis then gives the sequential order of the surveys. In the plot on the right, the missing values have been excluded and the proportions of each answer are plotted over time.

```
library(dplyr)
happy <- happy %>% mutate(yearF=factor(year))
ggplot(happy, aes(yearF, fill=happy)) + geom_bar(position="fill") +
  labs(x=NULL, y=NULL, fill="") +
  theme(legend.position="bottom", axis.text.x=element_text(angle=90))
h3 <- happy %>% filter(!is.na(happy)) %>% group_by(year, happy) %>%
  summarise(freq=n()) %>% group_by(year) %>% mutate(gsum=sum(freq), shar=freq/gsum)
ggplot(h3, aes(year, shar)) + geom_line(aes(group=happy, colour=happy)) + ylim(0, 0.6) +
  labs(x=NULL, y=NULL) + geom_point(aes(colour=happy)) +
  theme(legend.position="bottom", legend.title=element_blank())
```

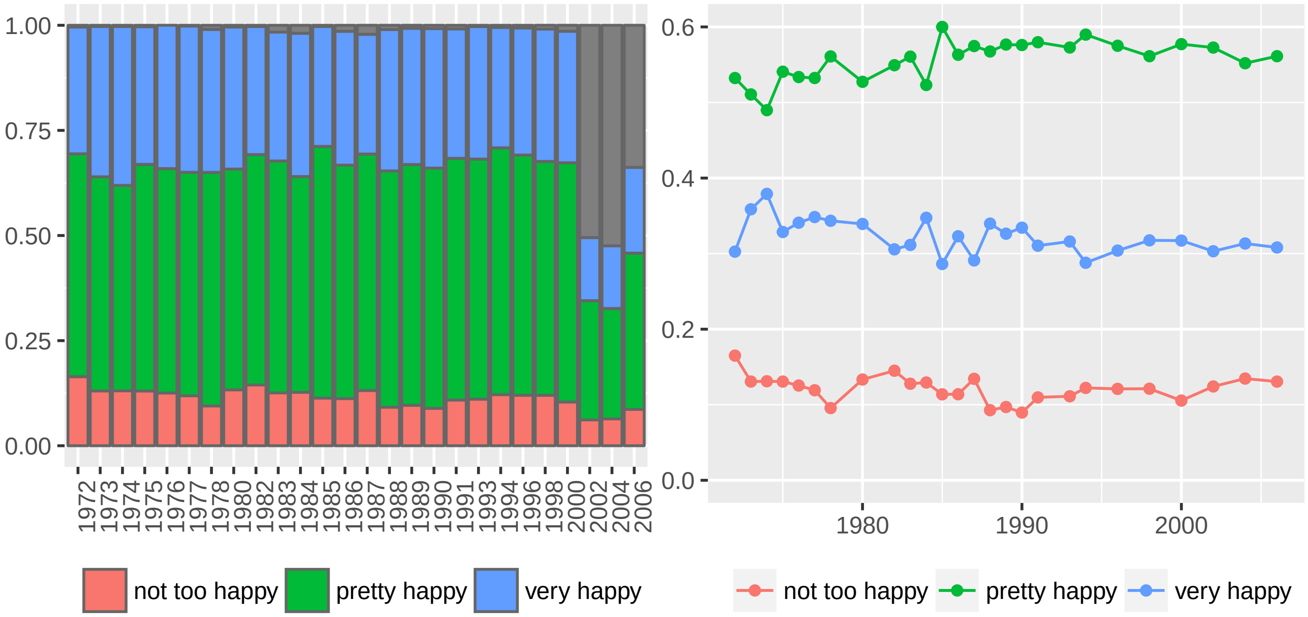

Fig 3: The left plot shows that most of the missing values for the ‘happy’ question arose in the three most recent surveys. Excluding them, as in the plot on the right, shows that the proportions giving the three answers have remained fairly constant over time.

Checking the codebook on the web, which includes some summary tables, reveals that almost all these missings were coded ‘non-applicable’ and that no cases were coded ‘non-applicable’ for these questions in other years. No further explanation was given.

Since NAs are evident for both ‘happy’ and ‘health’ variables, it is worth looking at NAs for the dataset as a whole. Figure 4 shows how many and what patterns of missings there were.

```r
library(extracat)
visna(happy, sort="b")
visna(happy, sort="b", fr=0.98, fc=0.98)
```

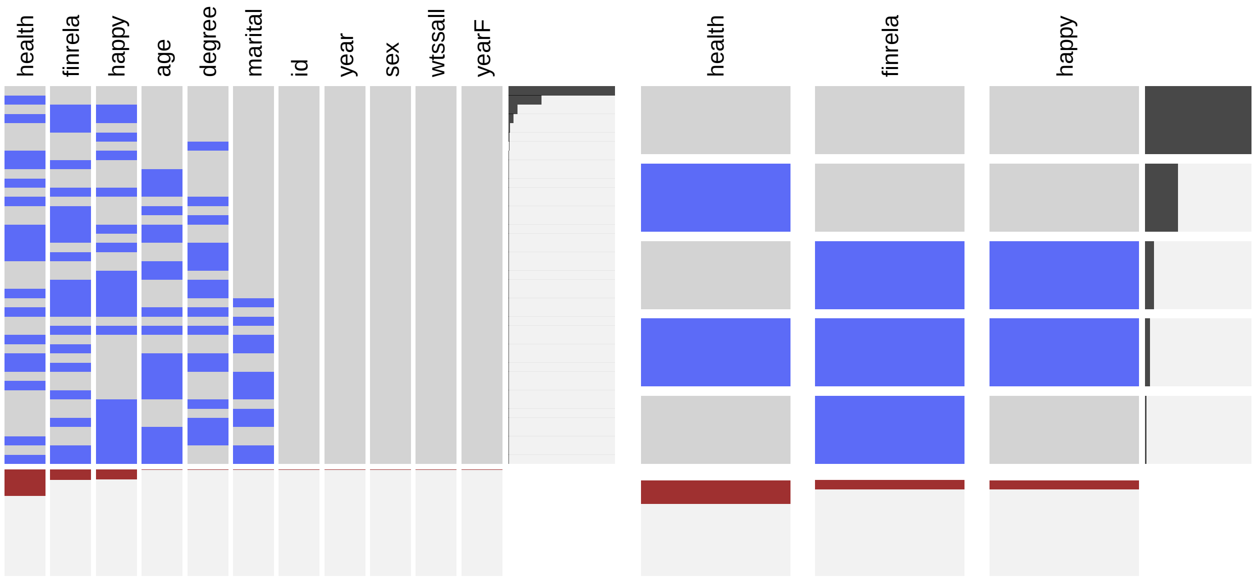

Fig 4: The plot on the left shows that the health variable had most NAs, followed by finrela (financial status) and happy. Restricting the plot to the rows and columns covering at least 98% of the data gives the plot on the right showing that the majority of cases had no missings, that health was often missing alone, and that finrela and happy were often missing together (the non-applicables mentioned above).

Surveys carried out each year for a number of years do not always include the same questions. Figure 5 shows the distribution of answers to the health question over the years.

```r
ggplot(happy, aes(yearF, fill=health)) + geom_bar(position="fill") +
  labs(x=NULL, y=NULL, fill="") + theme(legend.position="bottom")
```

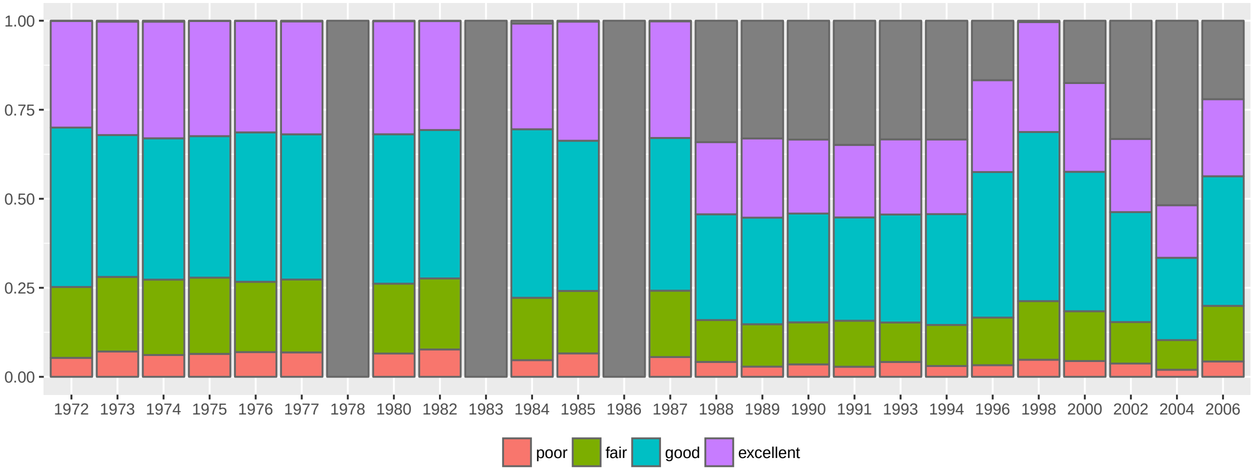

Fig 5: The health question was not asked in three of the surveys in the 1980s. Otherwise there were hardly NAs in the earlier surveys. In later surveys there were high proportions of NAs, often over 25%, once as high as 50%, and yet once just about 0%.

The distributions of age and sex are fairly consistent over the 26 surveys, probably by design. Two of the other variables, marital status and degree changed quite strikingly.

```r
hm <- happy %>% group_by(year, marital) %>% summarise(freq=n())
hm <- hm %>% group_by(year) %>% mutate(gsum=sum(freq), shar=freq/gsum)
ggplot(na.omit(hm), aes(year, shar)) + geom_line(aes(group=marital, colour=marital)) +
  ylim(0, 0.75) + labs(x=NULL, y=NULL, fill="") + theme(legend.position="left")
hd <- happy %>% group_by(year, degree) %>% summarise(freq=n())
hd <- hd %>% group_by(year) %>% mutate(gsum=sum(freq), shar=freq/gsum)
ggplot(na.omit(hd), aes(year, shar)) + geom_line(aes(group=degree, colour=degree)) +
  ylim(0, 0.6) + labs(x=NULL, y=NULL, fill="") + theme(legend.position="right")
```

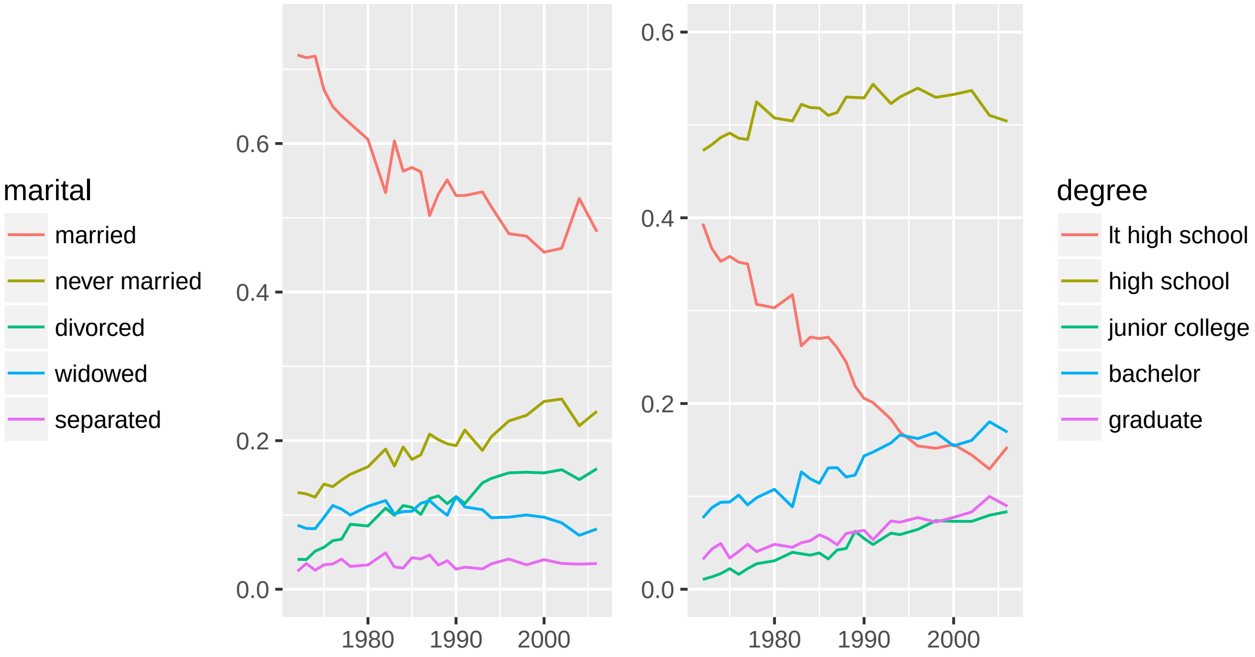

Fig 6: The proportion married in the surveys fell from over 70% to just under 50%. The never married and divorced groups increased accordingly. The proportion reporting education as 'less than high school' decreased from 40% to around 15%, while the other four groups all rose. In the most recent surveys included in this dataset the proportion with only high school education fell slightly.

There were a number of other features that might be expected: more women than men were in the dataset and the proportion of women increased with age above 65; widows and widowers were older; respondents with higher degrees said they were financially better off, males were better off than women, the middle-aged were better off than young or old. All these features and more just emphasise that the possible explanatory variables are associated.

Both health and financial status are important explanatory variables and Figure 7 shows their effect on happy. Cases missing on any of the three variables have been excluded, just over 31%.

```r
h7 <- na.omit(happy %>% select(health, happy, finrela))
ggplot(h7, aes(health, fill=happy)) + geom_bar(position="fill", width=1) +
  facet_grid(~finrela) + theme(legend.position="bottom", legend.title=element_blank()) +
  labs(x=NULL, y=NULL) + ggtitle("Happiness by health and financial status")
```

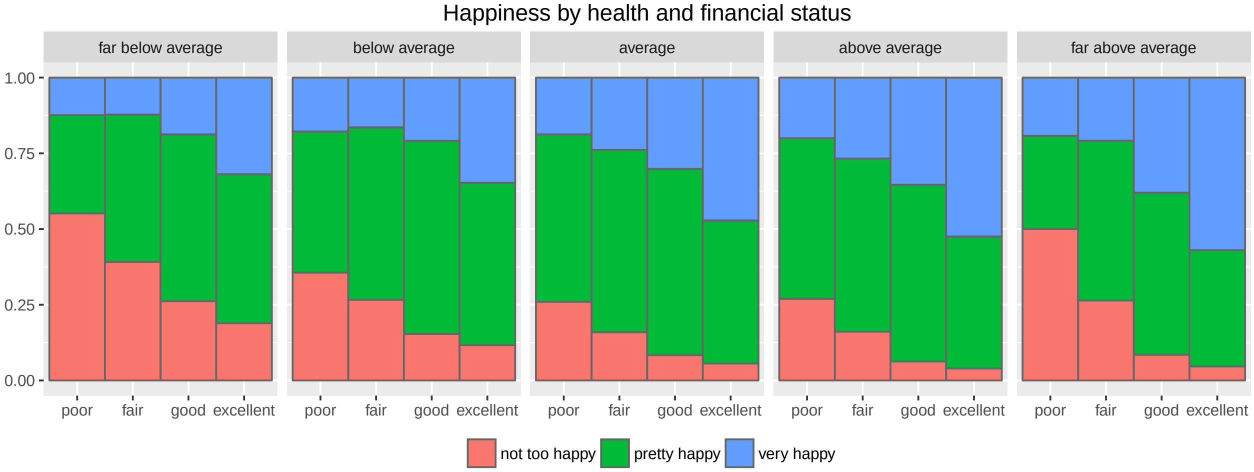

Fig 7: Happiness increased with health for all categories of financial status. The proportion of 'not too happy' declined and the proportion of 'very happy' increased (although not so much for 'far below average' finrela). Happiness also increased with financial status, but that is not so clearcut: notice the higher levels of 'not too happy' in the 'far above average' group for health being 'poor' or 'fair'. Not too much should be made of this, as despite the overall size of the dataset (51020 cases) there were very few in those groups.

Sex and marital status could also be relevant.

```r
h9 <- na.omit(happy %>% select(sex, happy, marital))
ggplot(h9, aes(marital, fill=happy)) + geom_bar(position="fill", width=1) +
  facet_grid(~sex) + theme(legend.position="bottom", legend.title=element_blank()) +
  labs(x=NULL, y=NULL) + ggtitle("Happiness by sex and marital status")
```

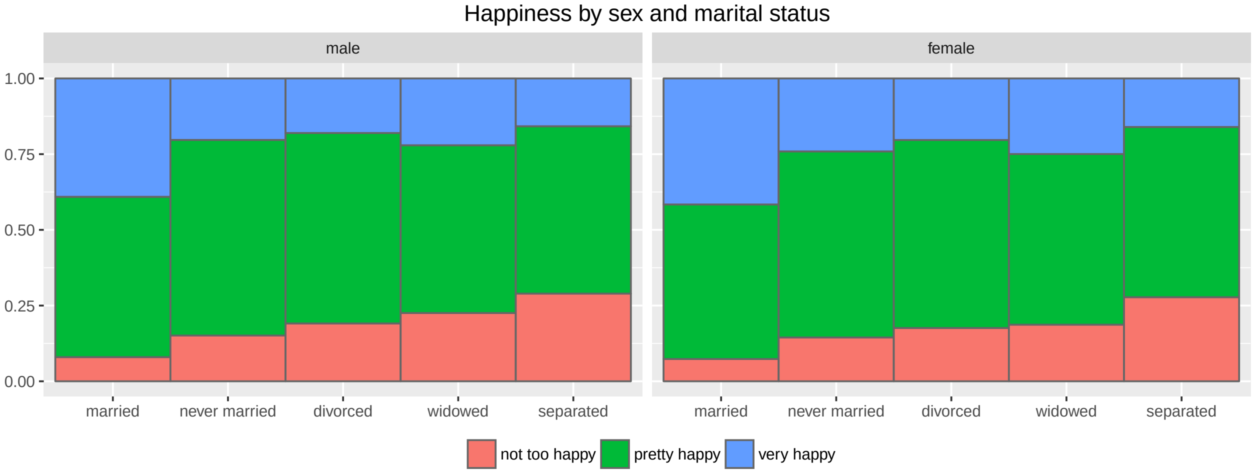

Fig 8: The married group were happiest and the separated group least happy. It is striking how similar the patterns for men and women were.

The dataset was used in an article by Hadley Wickham and Heike Hofmann "Product plots" (IEEE Trans Vis Comput Graph. 2011 Dec;17(12):2223-30).