Explore and Summarize Data Analysis using R

Analysis of Female labor force.

An analysis was performed to understand the distribution of female employment among the three distinct employment categories, across countries and to study the change in distribution over the years.The female labor employment data is collected from www.Gapminder.org.

The desired data which is used for this analysis ,is obtained from three different files .

The three files are

- Female salaried workers.This file describes the percentage of all female labor that earns a salary from working.
- Female self-employed .This file describes the percentage of all female labor that works as self-employed workers.
- Female family workers.This file describes the percentage of all female labor that works as contributing family workers.

The structure of salaried worker is shown below.

```
## 'data.frame':    11 obs. of  29 variables:
##  $ countries: Factor w/ 160 levels "","Algeria","American Samoa",..: 8 14 27 52 71
73 75 104 134 151 ...
##  $ 1980     : Factor w/ 39 levels "","[Add other fields as required]",..: 11 8 14
10 7 5 4 13 6 17 ...
##  $ 1981     : Factor w/ 28 levels "","[Download csv]  Not available yet!",..: 13 1
0 17 12 9 7 5 15 8 20 ...
##  $ 1982     : Factor w/ 20 levels "","39.90000153",..: 11 7 14 9 6 4 2 13 5 17 ...
##  $ 1983     : num  87.3 81.8 91.8 86 73.3 ...
##  $ 1984     : num  87.8 81.8 91.6 86.4 72.7 ...
##  $ 1985     : num  87.3 81.9 90.9 86.8 73.6 ...
##  $ 1986     : num  86.8 82 92 87.2 73.4 ...
##  $ 1987     : num  87.2 82.1 90.1 87.5 73.9 ...
##  $ 1988     : num  87.1 82.5 90 87.8 74.4 ...
##  $ 1989     : num  87.8 82.4 90.2 88.4 75.2 ...
##  $ 1990     : num  87.8 82.5 89.9 89 75.9 ...
##  $ 1991     : num  87.4 82.3 89.8 89.6 76.2 ...
##  $ 1992     : num  87.2 82.3 89.5 90.2 76.4 ...
##  $ 1993     : num  86.9 81.9 88.8 90.8 75.7 ...
##  $ 1994     : num  87.4 82.1 88.3 91.2 75.8 ...
##  $ 1995     : num  87.9 82.1 88.3 91.7 76.2 ...
##  $ 1996     : num  88.1 82.4 87.5 92.1 76.4 ...
##  $ 1997     : num  87.8 82.9 86.7 92.5 76.8 ...
##  $ 1998     : num  88.6 83.5 86.7 92.7 77 ...
##  $ 1999     : num  88.9 84.1 87.1 92.9 77.4 ...
##  $ 2000     : num  89.3 NA 87.7 93.1 78 ...
##  $ 2001     : num  89.7 NA 88.8 93.3 78.4 ...
##  $ 2002     : num  89.6 NA 88.6 93.2 78.8 ...
##  $ 2003     : num  90 NA 88.7 91.9 78.9 ...
##  $ 2004     : num  90.2 NA 88.8 92.4 77.8 ...
##  $ 2005     : num  90.2 87.7 88.6 92.5 79.4 ...
##  $ 2006     : num  90.7 88.2 88.7 92.3 79.5 ...
##  $ 2007     : num  91 NA 88.6 92.7 80 ...
```

Each data file includes data from 153 countries all over the world from the year 1980 to 2007.

The countries Australia, Belgium, Canada, Finland, France, Italy, Japan, Korea,Rep., Norway, United Kingdom and United States are being considered for analysing the data, as they have almost complete data for the years 1980- 2007 in all three categories .

Each data file needs some sort of processing to proceed with our analysis. The column names needs to be changed and the datavalues corresponding to the year 1980,1981 and 1982 has to be converted from factor . The structure of self_employed is shown below.

```
## 'data.frame':    11 obs. of  29 variables:
##  $ countries: Factor w/ 160 levels "","Algeria","American Samoa",..: 8 14 27 52 71
73 75 104 134 151 ...
##  $ 1980     : Factor w/ 39 levels "","[Add other fields as required]",..: 4 17 16
6 8 5 9 12 7 10 ...
##  $ 1981     : Factor w/ 26 levels "","[Download csv]  Not available yet!",..: 5 18
16 7 9 6 11 12 8 14 ...
##  $ 1982     : Factor w/ 20 levels "","12.19999981",..: 2 17 15 5 7 3 9 11 6 13 ...
##  $ 1983     : num  12.1 9.9 6.4 14 16.3 ...
##  $ 1984     : num  11.8 9.9 6.7 13.6 16.2 ...
##  $ 1985     : num  12.1 9.9 7.3 13.2 15.9 ...
##  $ 1986     : num  11.9 9.8 6.5 12.8 16.4 ...
##  $ 1987     : num  11.4 9.7 8.5 12.5 16.4 ...
##  $ 1988     : num  11.5 9.6 8.9 12.2 16.4 ...
##  $ 1989     : num  11 9.6 8.8 11.6 16.8 ...
##  $ 1990     : num  10.9 9.5 9.2 11 16.6 ...
##  $ 1991     : num  11.4 9.6 9.3 10.4 16.5 ...
##  $ 1992     : num  11.5 9.6 9.7 9.8 16.5 ...
##  $ 1993     : num  11.7 9.8 10.3 9.2 15.8 ...
##  $ 1994     : num  11.2 10 11 8.8 16.2 ...
##  $ 1995     : num  10.8 10 11 8.3 16.4 ...
##  $ 1996     : num  10.7 10 11.8 7.9 16.5 ...
##  $ 1997     : num  11 10 12.5 7.5 16.4 ...
##  $ 1998     : num  10.4 9.7 12.6 7.3 16.4 ...
##  $ 1999     : num  9.9 9.5 12.4 7.1 16.6 ...
##  $ 2000     : num  9.6 NA 11.9 6.9 16.1 ...
##  $ 2001     : num  9.6 NA 10.9 6.7 15.6 ...
##  $ 2002     : num  9.8 NA 11.1 6.8 15.3 ...
##  $ 2003     : num  9.5 NA 11 6.1 15.3 ...
##  $ 2004     : num  9.3 NA 11 5.8 15.2 ...
##  $ 2005     : num  9.3 8.9 11.2 5.9 14.7 ...
##  $ 2006     : num  8.9 8.9 11.1 6.2 14.6 ...
##  $ 2007     : num  8.7 NA 11.2 6.3 14.4 ...
```

The first five rows of self_employed are displayed.

```
##    countries 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991
## 7  Australia 12.7 12.2 12.2 12.1 11.8 12.1 11.9 11.4 11.5 11.0 10.9 11.4
## 13   Belgium  9.4  9.6  9.7  9.9  9.9  9.9  9.8  9.7  9.6  9.6  9.5  9.6
## 26    Canada  6.0  5.6  6.0  6.4  6.7  7.3  6.5  8.5  8.9  8.8  9.2  9.3
## 50    France 14.6 14.5 14.3 14.0 13.6 13.2 12.8 12.5 12.2 11.6 11.0 10.4
## 68     Italy 16.0 15.7 15.8 16.3 16.2 15.9 16.4 16.4 16.4 16.8 16.6 16.5
##    1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005
## 7  11.5 11.7 11.2 10.8 10.7 11.0 10.4  9.9  9.6  9.6  9.8  9.5  9.3  9.3
## 13  9.6  9.8 10.0 10.0 10.0 10.0  9.7  9.5   NA   NA   NA   NA   NA  8.9
## 26  9.7 10.3 11.0 11.0 11.8 12.5 12.6 12.4 11.9 10.9 11.1 11.0 11.0 11.2
## 50  9.8  9.2  8.8  8.3  7.9  7.5  7.3  7.1  6.9  6.7  6.8  6.1  5.8  5.9
## 68 16.5 15.8 16.2 16.4 16.5 16.4 16.4 16.6 16.1 15.6 15.3 15.3 15.2 14.7
##    2006 2007
## 7   8.9  8.7
## 13  8.9   NA
## 26 11.1 11.2
## 50  6.2  6.3
## 68 14.6 14.4
```

The structure of service employ is shown below.

```
## [1] "/Users/ambilurama/Documents/Nano/own_interest/salaried_social_self"
```

```
## 'data.frame':    11 obs. of  29 variables:
##  $ countries: Factor w/ 160 levels "","Algeria","American Samoa",..: 8 14 27 52 71
73 75 104 134 151 ...
##  $ 1980     : Factor w/ 38 levels "","[Add other fields as required]",..: 4 16 10
3 7 11 12 14 9 1 ...
##  $ 1981     : Factor w/ 27 levels "","[Download csv]  Not available yet!",..: 6 19
12 5 10 13 15 17 11 1 ...
##  $ 1982     : Factor w/ 19 levels "","0","0.600000024",..: 3 16 8 2 6 9 12 14 7 1
...
##  $ 1983     : num  0.6 8.3 1.8 0 10.5 ...
##  $ 1984     : num  0.4 8.4 1.7 0 11.1 ...
##  $ 1985     : num  0.6 8.3 1.8 0 10.5 ...
##  $ 1986     : num  1.3 8.3 1.5 0 10.2 ...
##  $ 1987     : num  1.4 8.2 1.4 0 9.7 ...
##  $ 1988     : num  1.4 8 1.1 0 9.1 ...
##  $ 1989     : num  1.2 8.1 1 0 8 ...
##  $ 1990     : num  1.2 8 0.9 0 7.5 ...
##  $ 1991     : num  1.2 8.1 0.9 0 7.3 ...
##  $ 1992     : num  1.4 8.1 0.8 0 7.1 ...
##  $ 1993     : num  1.4 8.3 0.9 0 8.4 ...
##  $ 1994     : num  1.4 7.9 0.7 0 8 ...
##  $ 1995     : num  1.3 7.8 0.7 0 7.4 ...
##  $ 1996     : num  1.2 7.6 0.7 0 7.1 ...
##  $ 1997     : num  1.3 7.2 0.7 0 6.8 ...
##  $ 1998     : num  1 6.7 0.7 0 6.6 ...
##  $ 1999     : num  1.2 6.4 0.5 0 6 ...
##  $ 2000     : num  1.1 NA 0.4 0 5.9 ...
##  $ 2001     : num  0.7 NA 0.3 0 6 ...
##  $ 2002     : num  0.6 NA 0.3 0 5.9 ...
##  $ 2003     : num  0.5 NA 0.3 1.9 5.8 ...
##  $ 2004     : num  0.5 NA 0.3 1.7 3.7 ...
##  $ 2005     : num  0.4 3.4 0.2 1.6 2.8 ...
##  $ 2006     : num  0.4 2.9 0.2 1.4 2.7 ...
##  $ 2007     : num  0.4 NA 0.2 1 2.6 ...
```

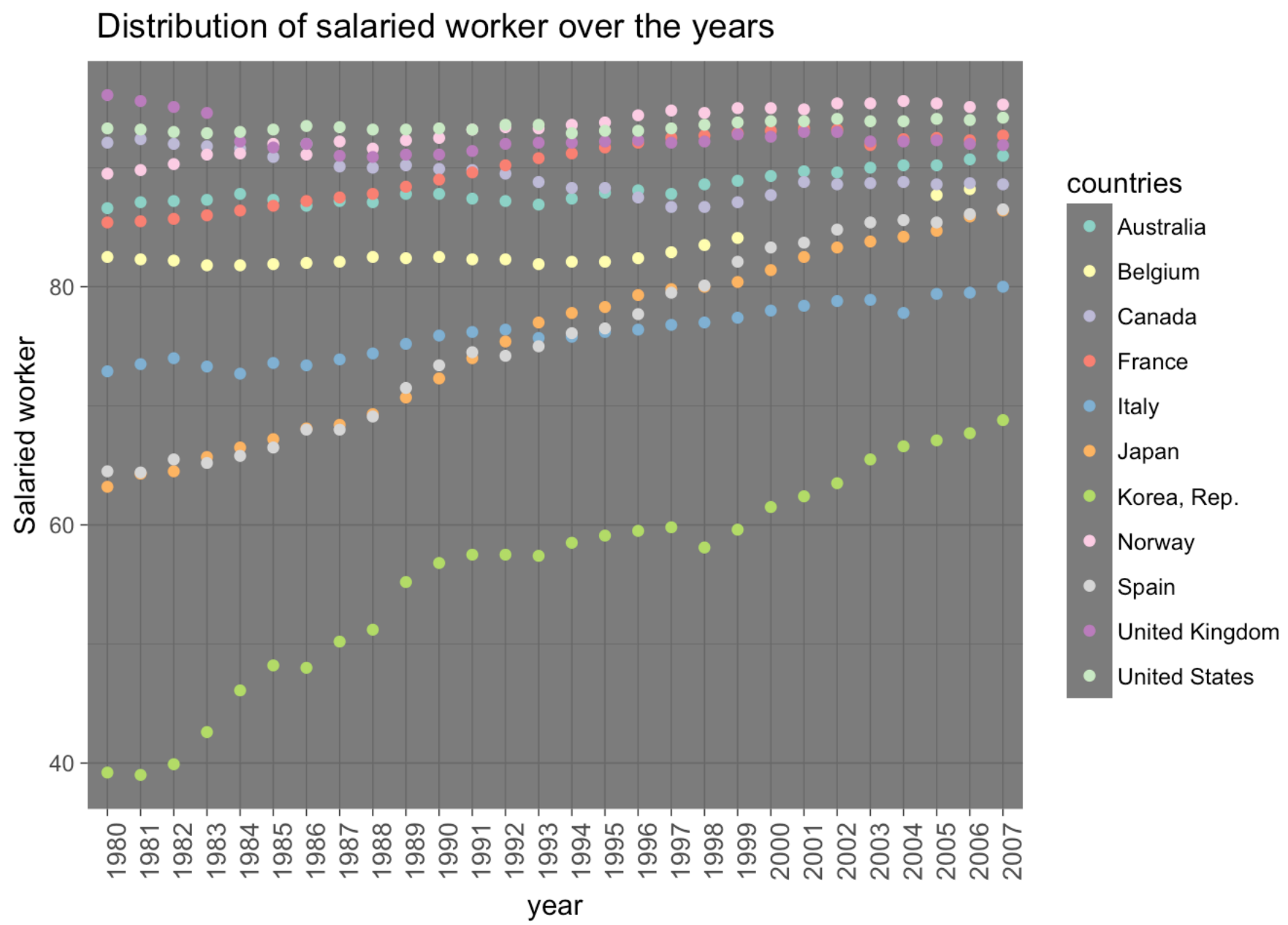
And then the data is converted to the required tidy form using melt.The three different files are combined together and thus becoming our desired dataset .

A new variable called “Year category” has been created . This variable segregates each year into one of the categories 1980-1985, 1985-1990 , 1990-1995 ,1995-2000 and 2000-2007.The same data is analysed with the year category variable as well.

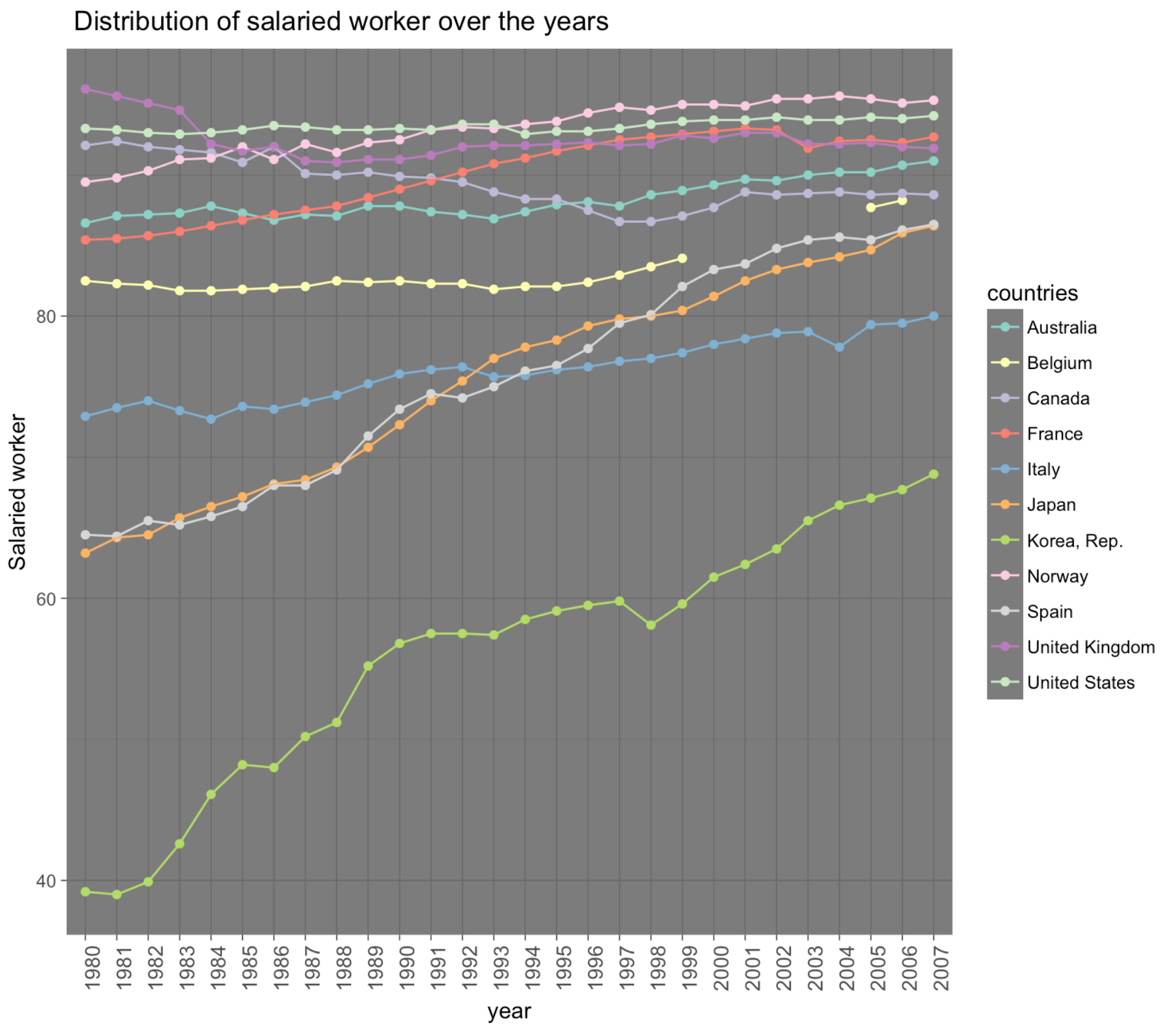
The first five rows of our desired dataset are displayed .

```
##   year countries salaried_worker self_employed family_workers    year_cat
## 1 1980 Australia            86.6          12.7             0.6 [1980,1985]
## 2 1980   Belgium            82.5           9.4             8.0 [1980,1985]
## 3 1980    Canada            92.1           6.0             2.0 [1980,1985]
## 4 1980    France            85.4          14.6             0.0 [1980,1985]
## 5 1980     Italy            72.9          16.0            11.1 [1980,1985]
```

salaried worker plots

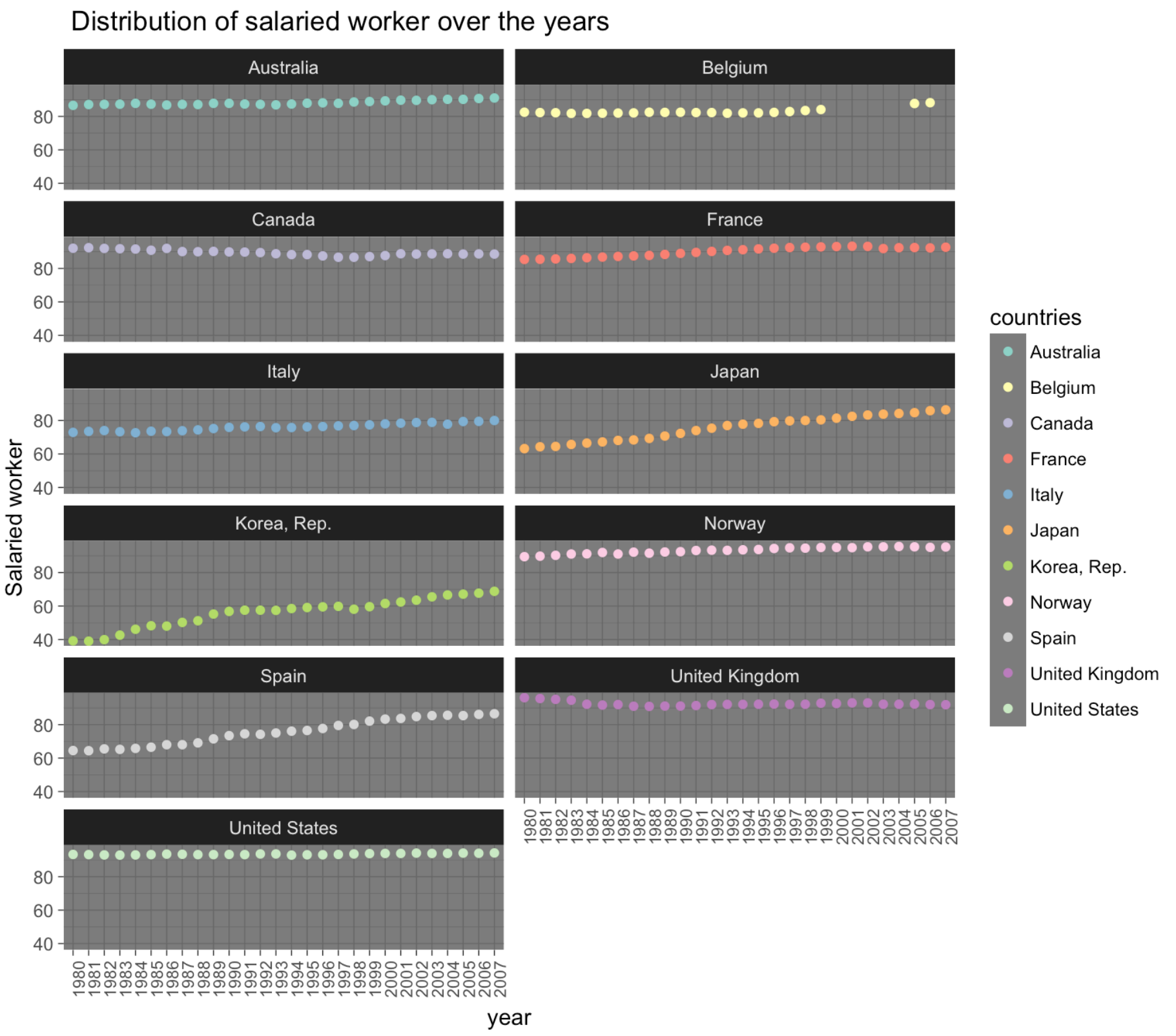


This scatterplot shows the distribution of salaried_worker vs year for all the countries .Each country is represented by a unique colour.As you can see, the salaried worker values above 80 are crowded and some of the data points are overlapping with other points. To see the patterns and datapoints clearly,it would be better to connect datapoints of a country with a line.



In this plot ,the slopes are better visible and could be easily compared with one another.

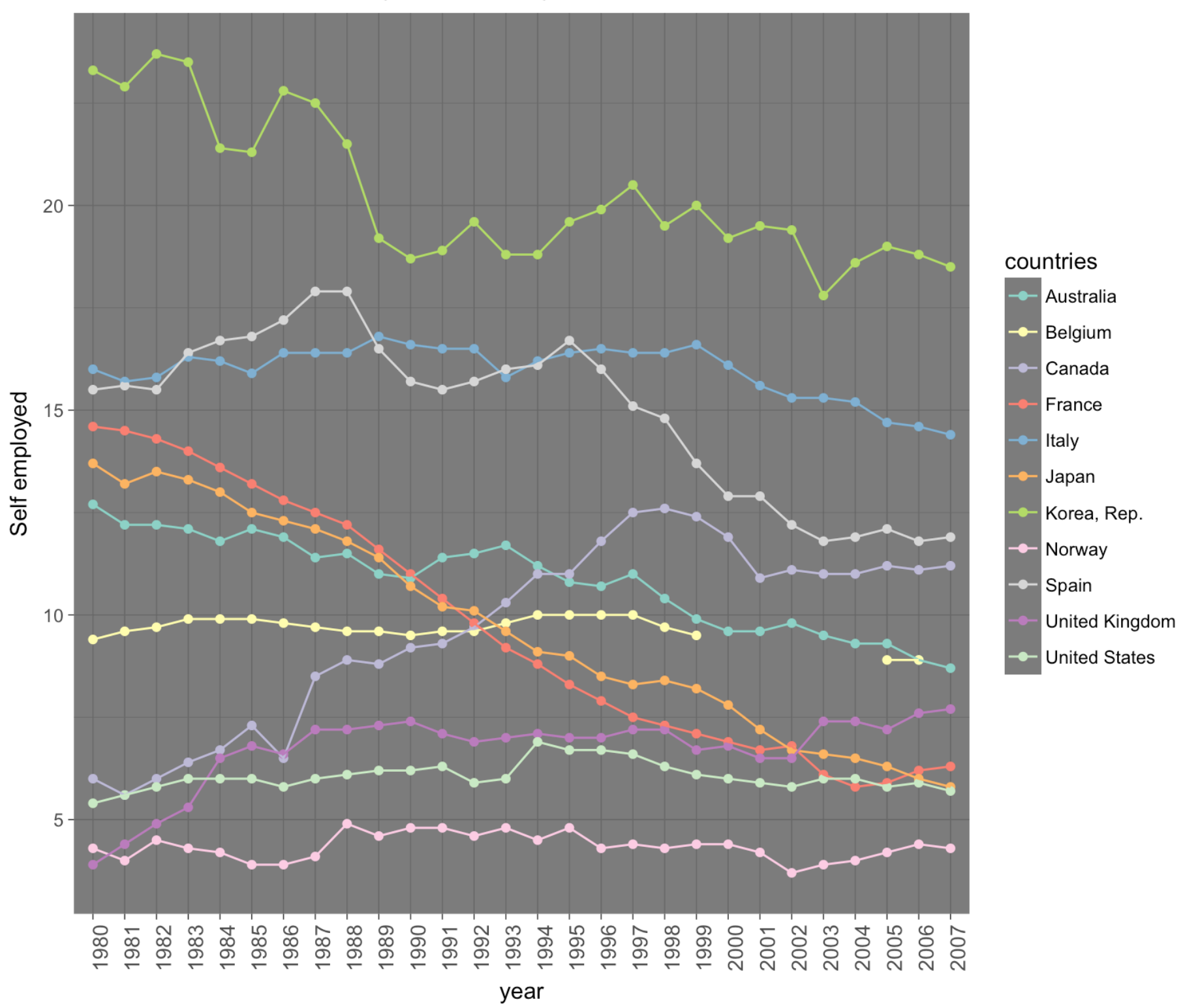
- Korea,Rep. , spain and Japan have a high slopes showing the high rise of salaried workers over the years.
- The broken line connection on Belgium indicates the missing datapoints for the corresponding years.
- United kingdom and canada show a slight down trend.
- Except United kingdom ,United states and canada ,all the other countries show a rising trend for salaried worker.
- Norway which was fourth top in the salaried worker values in 1980 has become the top in 2007 .



In this plot, the patterns of each country are clearly visible. This helps in understanding the trend.

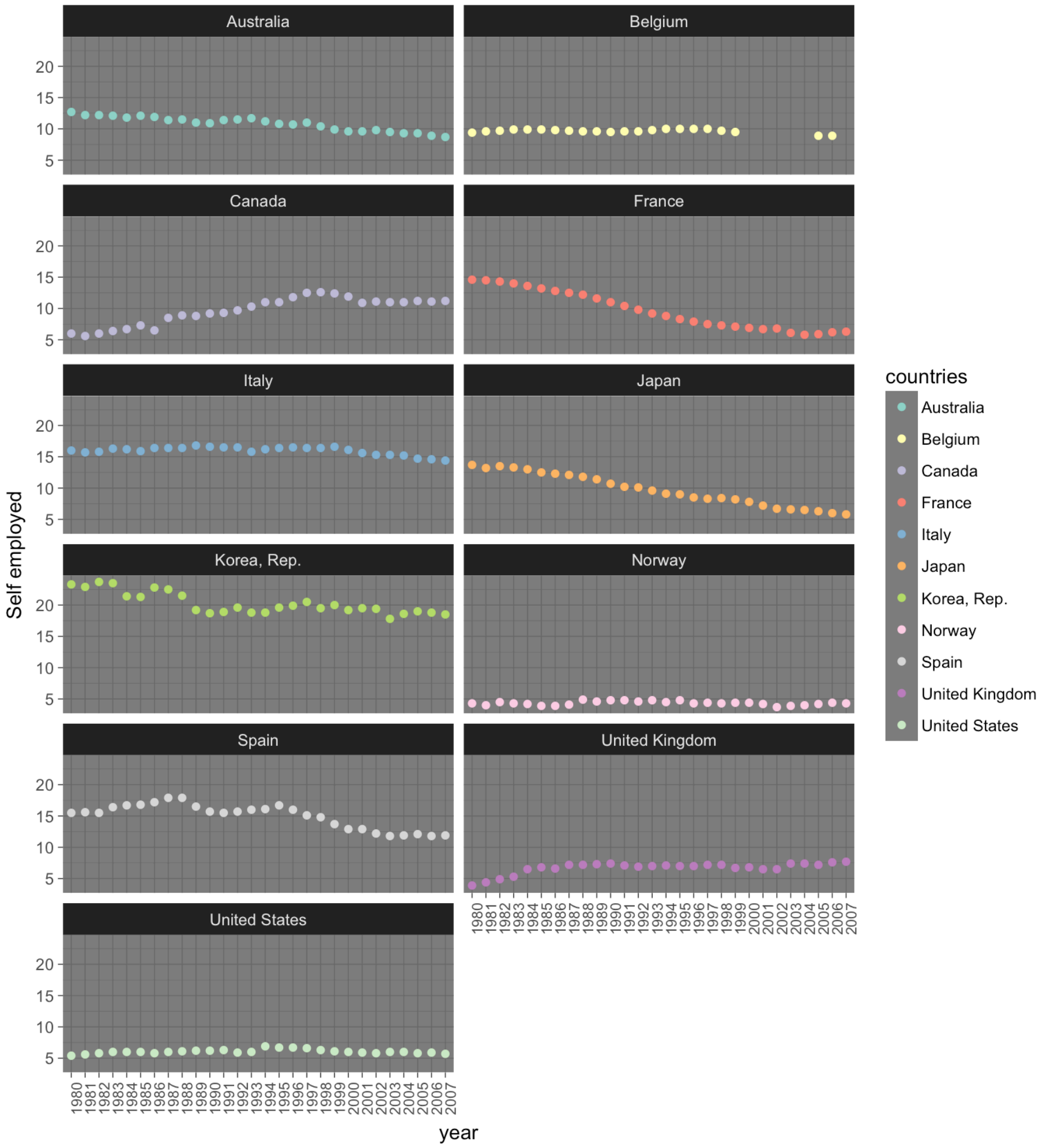
- Belgium has some missing datapoints for certain years.
- United kingdom and canada show a slight down trend.
- United states maintains the nearly same percentage of salaried workers every year.
- Except United kingdom ,United states and canada ,all the other countries show a rising trend for salaried worker.
- Spain ,Korea Rep. and Japan shows the high rise of salaried workers over the years.

Self- employed plots



This plot shows the distribution of self employed vs year for all the countries. It shows

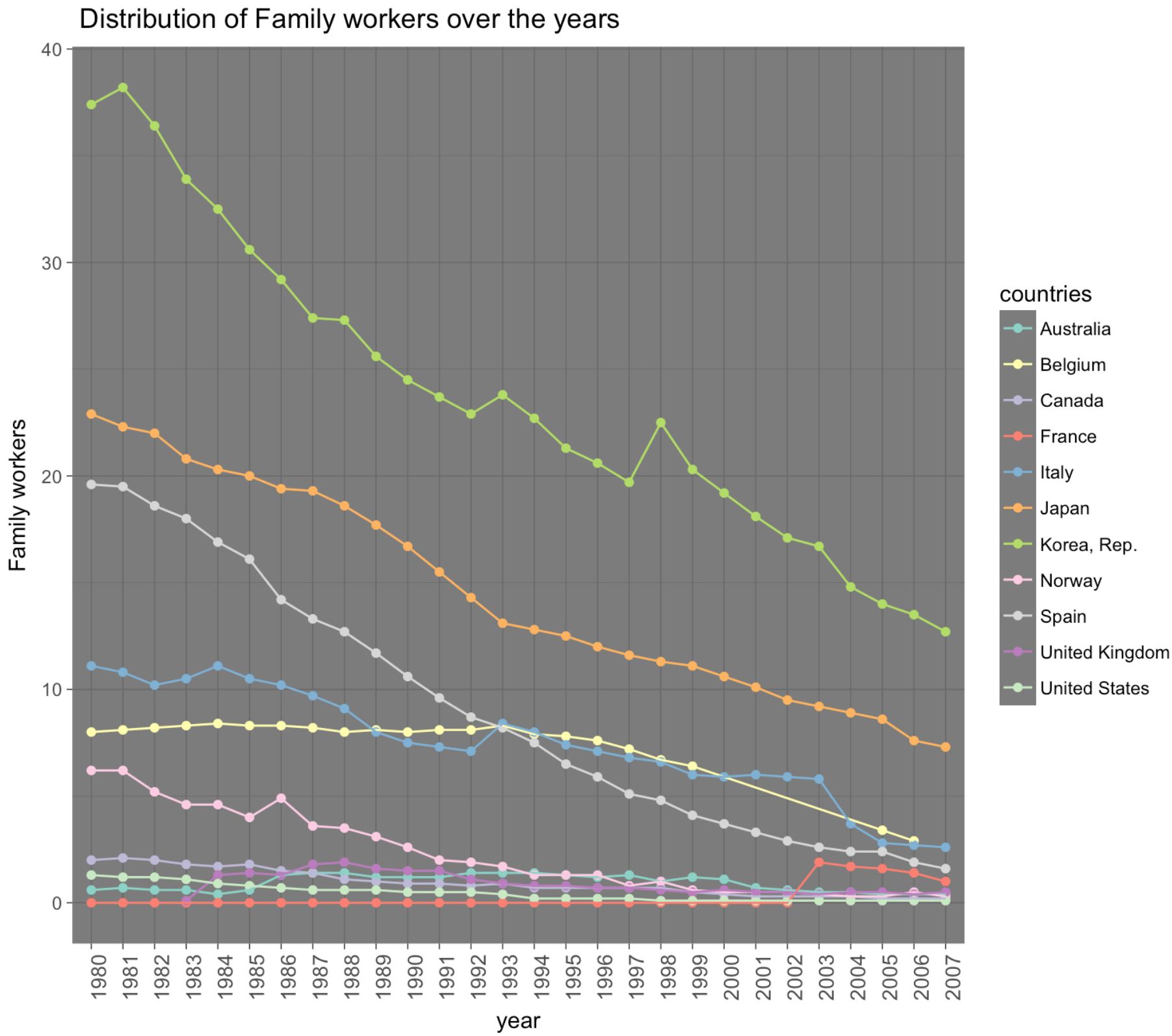
- France and Japan show a deep drop than others.
- United kingdom and canada show a clear rising trend.
- Korea Rep. stays on top all these years, eventhough the self employed values drops over the years.
- Norway has the least self employed values over the years.



In this plot, the patterns of each country are clearly visible. This helps in understanding the trend.

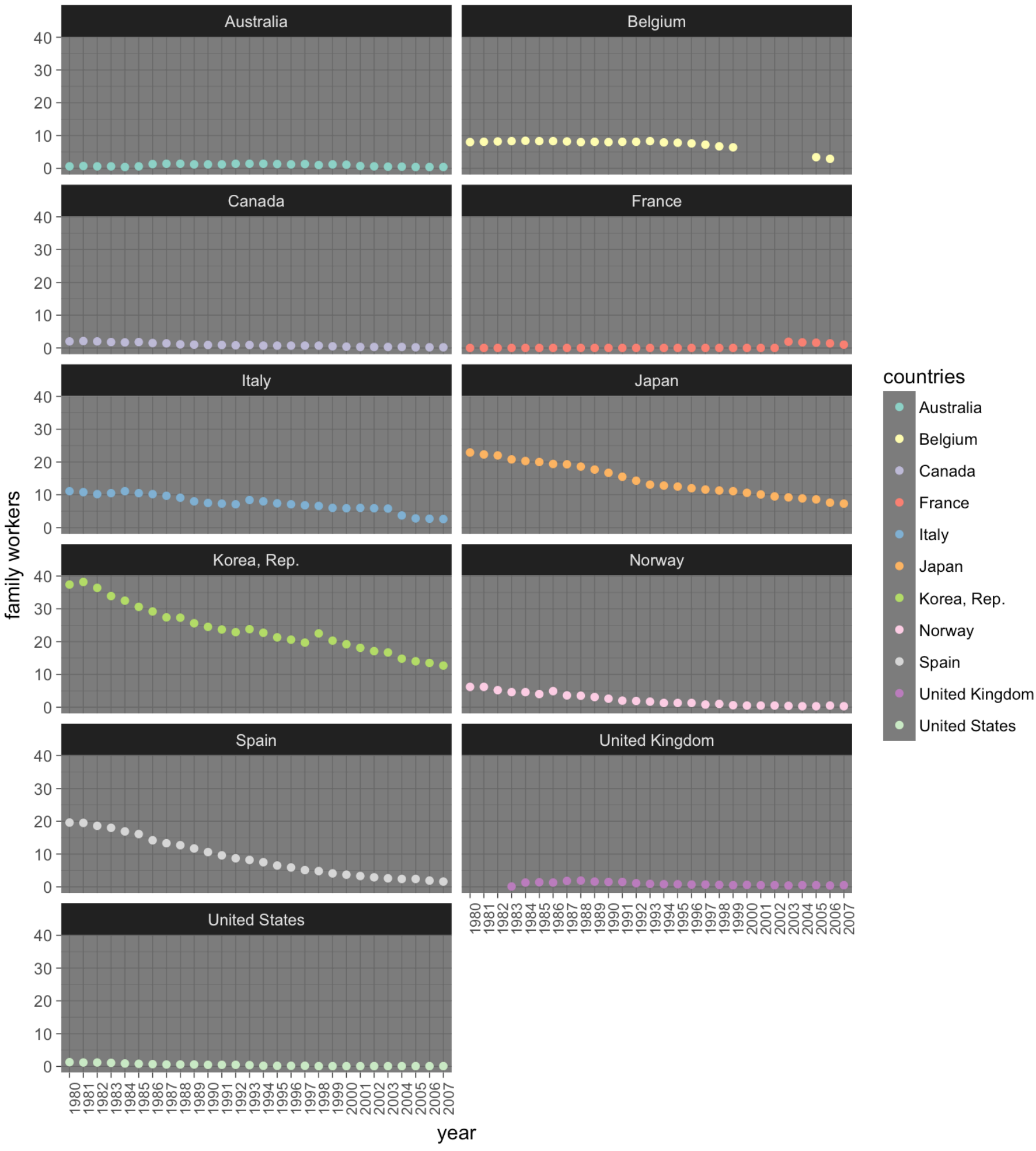
- Belgium has missing datapoints for certain years.
- United states and Norway maintain nearly same values over the years.
- United kingdom and canada show a clear rising trend.canada show a better rising trend than United kingdom.

Family workers plots



This plot shows the distribution of family workers vs year for all the countries. It shows

- Korea Rep. stays on top for all the years eventhough its values drop over the years
- Japan stays on second top for all the years.
- France has a value of 0.0 until the year 2002 and raised to 1.9 showing downtrend thereafter.



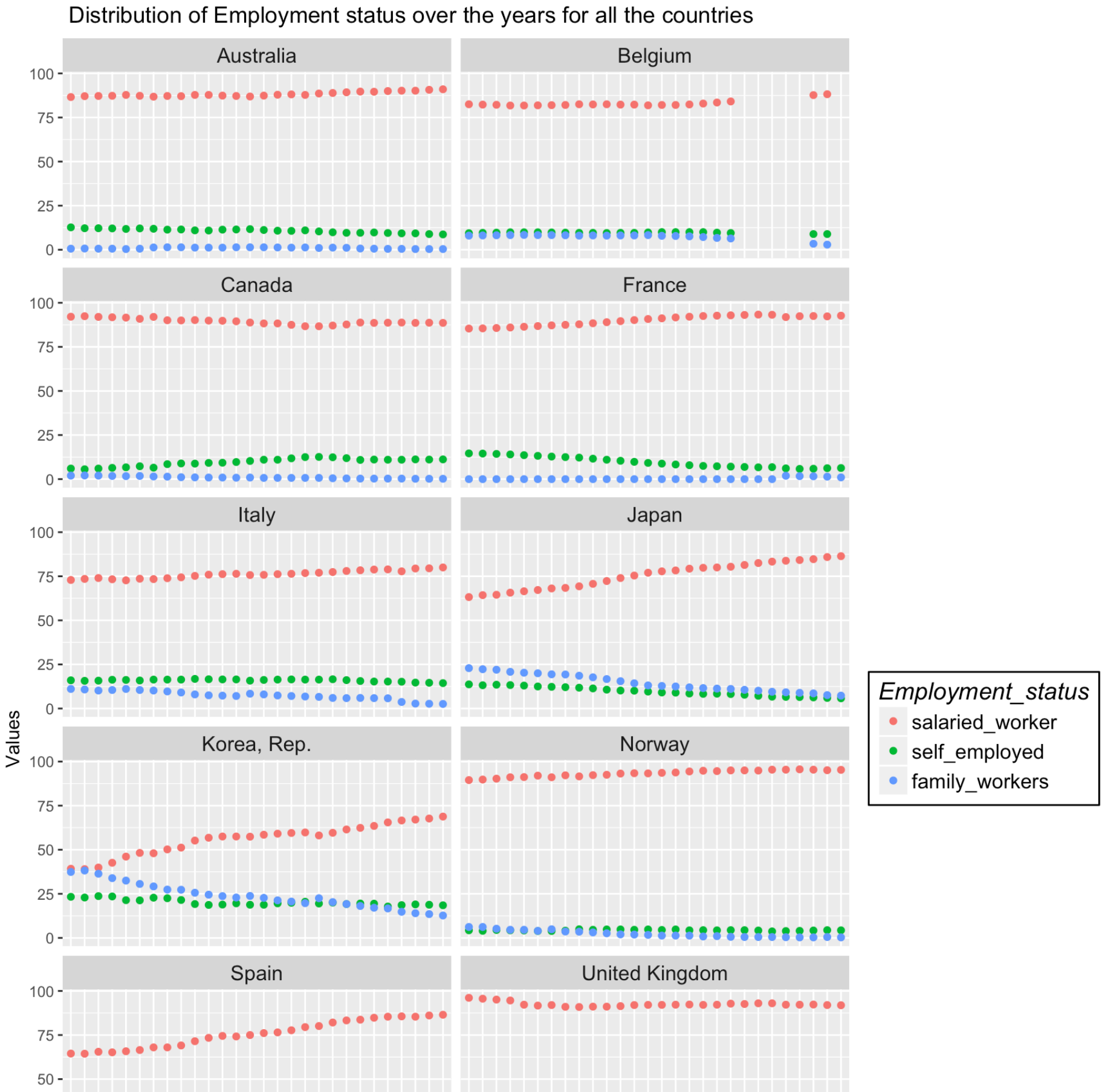
In this plot, the patterns of each country are clearly visible. This helps in understanding the trend.

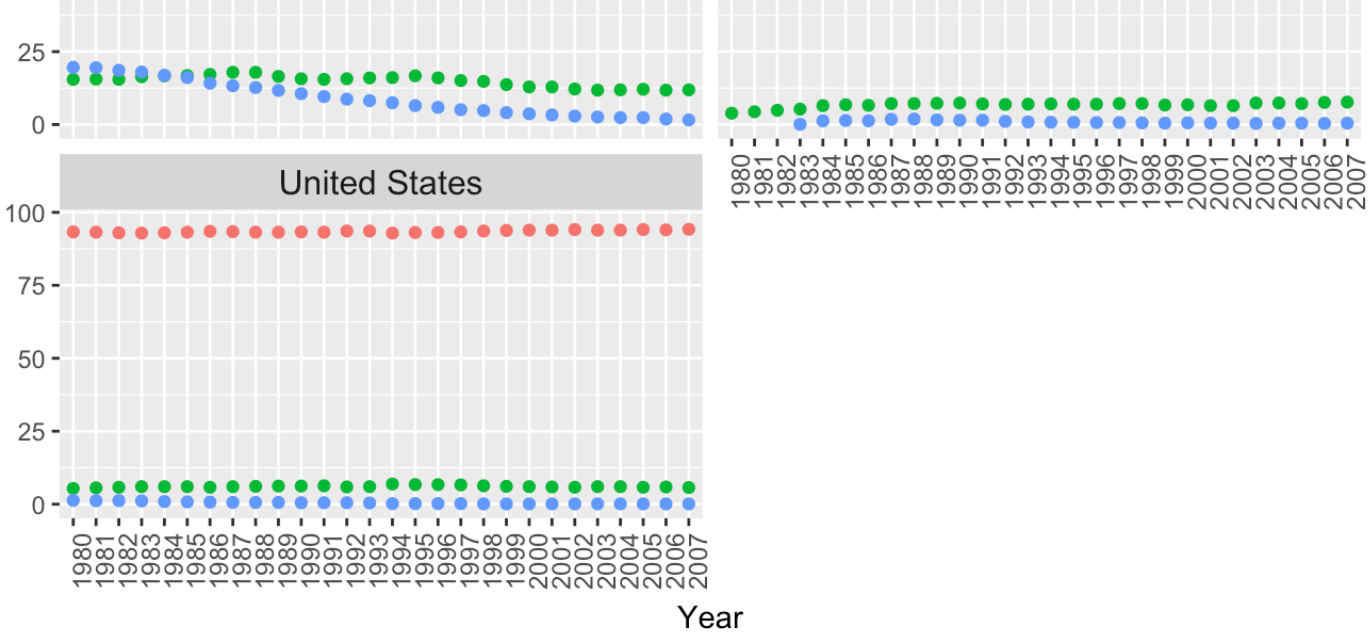
- All the countries except france show a drastic downtrend from 1980 to 2007.
- France show a downtrend from 2003 to 2007.
- Australia,canada,united states ,france and united kingdom have values less than 2.

Multivariate plots

The data needs to be converted to long format to analyse the data through all categories of employment status. The first five rows are displayed from the new dataset.

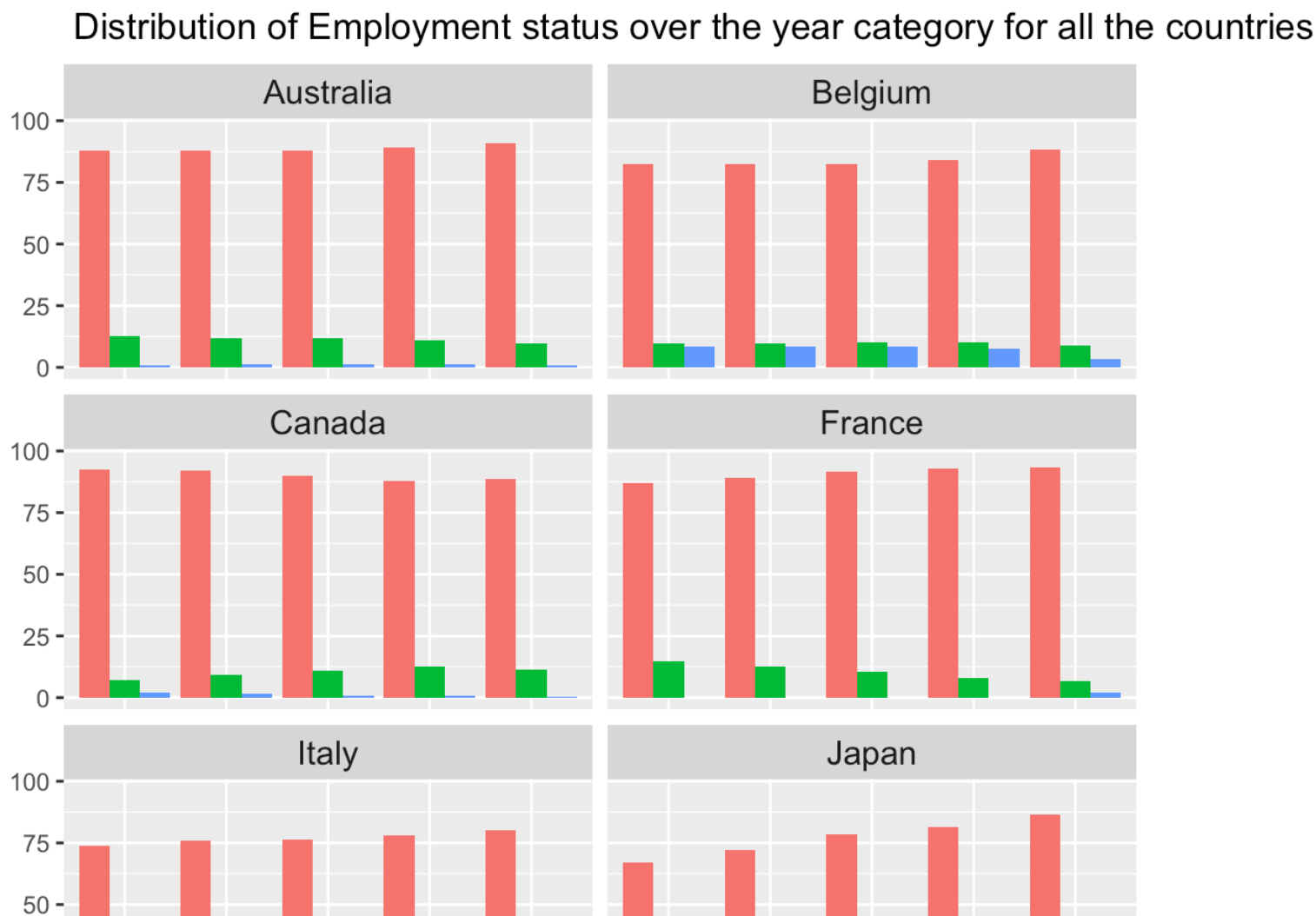
```
##   year countries    year_cat Employment_status values
## 1 1980 Australia [1980,1985]   salaried_worker   86.6
## 2 1980   Belgium [1980,1985]   salaried_worker   82.5
## 3 1980    Canada [1980,1985]   salaried_worker   92.1
## 4 1980    France [1980,1985]   salaried_worker   85.4
## 5 1980     Italy [1980,1985]   salaried_worker   72.9
```

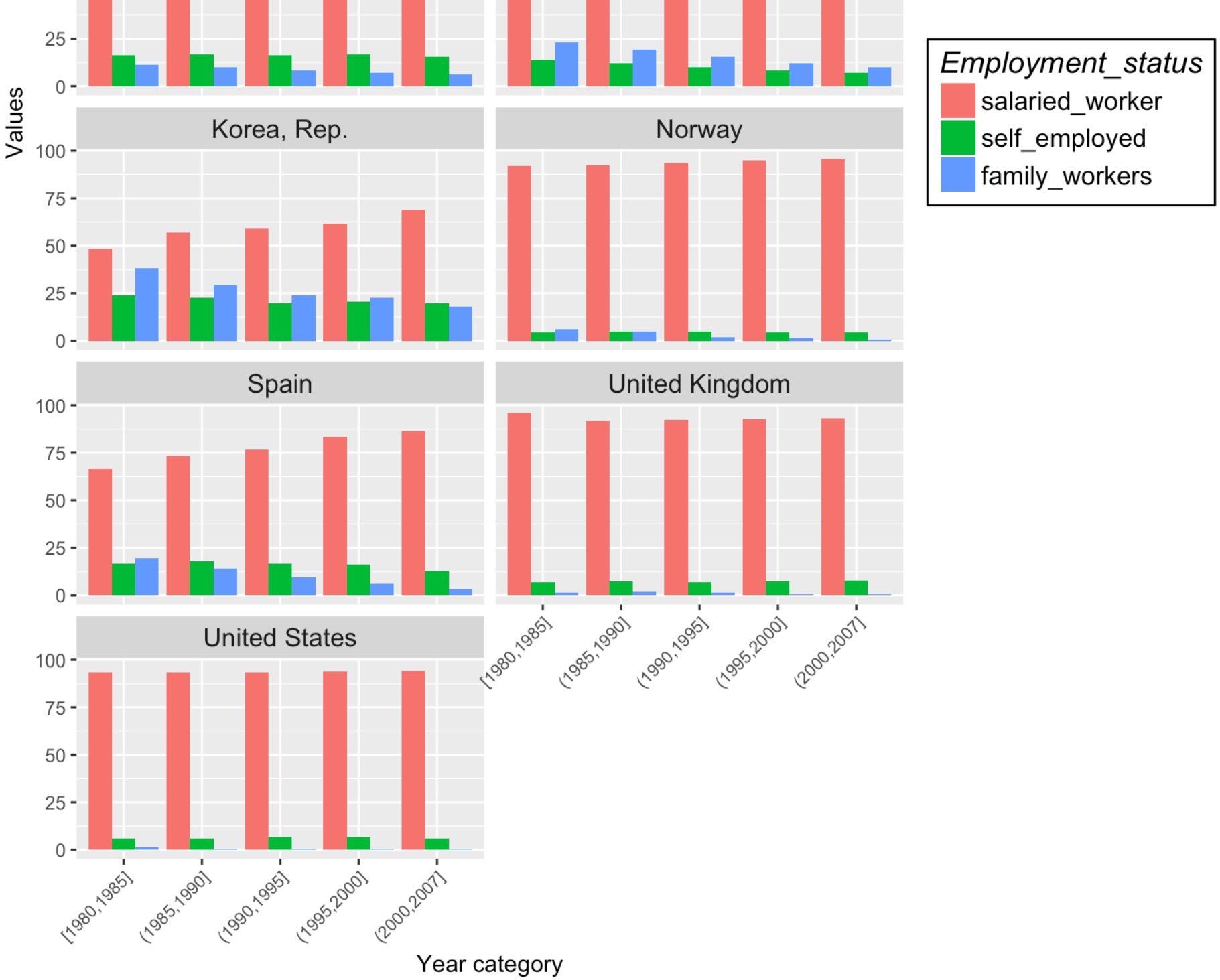




This plot shows the distribution of Employment status over the years for all the countries.There are few trends which are very common for all the countries.

- Family workers values are lower than or overlapping with Self employed values for all the countries except Japan and korea, Rep.
- Salaried worker values are higher than self-employed and family workers for all the countries.
- Japan has lower self-employed values than family workers values for all the years.
- family workers and Self employed values stay under 25 for all the countries except Korea, Rep whose family workers values are over 25.
- Belgium has a overlapping values for self-employed and family_workers
- The plot also shows some missing datapoints.





This bar plot shows the distribution of Employment status over the year category for all the countries.

- Salaried_workers values are high for all the countries.
- France has a value of zero for the years 1980-2000.
- Family_workers values show a downtrend in 2007.