Report 8: standardizing age/birthyear + quick analysis

A casual report

Ricky Heinrich for Mélanie Méthot

2023-09-25

Contents

Cleaning **1**

Tables 2

After first cleaning **3**

Quick Analysis **4**

General Distribution 4

By State 6

Through time 10

Cleaning

There are 2200 cases that have something recorded in the 'age' column, which is 71.8%. Unfortunately, as is standard is most data collection projects, the data is not clean; some cases have birth years, others an age, some have descriptors such as "aged" or "young", others have multiple values, like "1895 or 1900" and "26 or 30 years old". There are 208 cases that contain letters in the 'age' column, so not just numbers like we would like.

For multiple values, I could just pick one arbitrarily, which may not be accurate but should still be within the same period in one's life; there isn't much difference between someone who is 26 or 30.

I've pulled out the years out of the 'age' column (cases were there were 4 consecutive digits, presumed to be a year), and derived the age by subtracting the presumed birthyear from the main date (indictment). I also extracted ages, assuming there is no bigamist with an age in the triple digits, by extracting 2 consecutive digits. In the cases where a "birthyear" and an "age" were present, I compared the values to see if they agreed.

There were 65 where they didn't. Most of the time, the age recorded vs the one derived from the birthyear were within a few years of each other, so can probably safely be taken as muddle record keeping in primary sources. The following table shows the times were there was a difference of more than 3 years and where the age_derived was less than 0. 'Date' is the main date associated with the case (indictment date), 'age' is the original recorded column, 'birthyear' is the extracted birthyear from 'age', 'age_years' is the extracted age from 'age', and 'age_derived' is the difference between the 'Date' and the 'birthyear'.

Some cases, like row 6 where the entry in age was '12 Dec 1915', my code took the birthyear to be 1915 (as expected), and took the 'age' to be '12', when in this case the two digits represent the day of birth. (A

quick look seem to reveal that this is the only case where the birthdate was recorded in this manner, so we shouldn't have to worry about this mistake happening elsewhere.) In row 12, it looks like the date recorded in the 'age' column corresponds presumably to the date of the news article: can assume the transcriber wanted to emphasize the bigamist was 40 in 1888. In row 10, a date just two years after the indictmnet date is recorded, with no way to extract a reasonable age. In row 13, a birthdate in 1981 was recorded; it is possible it is meant to be 1881 (I took a quick look in the dossier for this case and didn't find anything relating to birthdate. I didn't look through the 15 links). In row 1 and 4, the differences between the age recorded and the derived year is 13 and 20 years; it may have to do with the indictment date? Maybe the age recorded is the age of when a wedding occurred. That's a case where it is unclear.

I'm fixing the 'obvious' mistakes, but including this part in the report still (even though it's not much use to you) to emphasize how consistent recording really helps. The second table shows the cases where the recorded age and the derived age differ by more than three years, after fixing the obvious problems.

Tables

	state	Date	age	birthyear	age_years	age_derived
1	NSW	1892-02-01	1855, 24	1855	24	37
2	NSW	1906-11-27	1866 36 (Link 1) 45 (Link 12)	1866	36	40
3	NSW	1913-11-26	1876, 29	1876	29	37
4	NSW	1916-05-25	28/ young man (4) / 1868	1868	28	48
5	NSW	1920-06-01	22/23 in 1920	1920	22	0
6	NSW	1920-09-01	19-20 in 1920	1920	19	0
7	NSW	1943-06-01	12 Dec 1915	1915	12	28
8	NSW	1945-02-05	32, 1917	1917	32	28
9	NSW	1947-06-06	40 (1912)	1912	40	35
10	Queensland	1883-09-17	JAN.7 1885	1885	NA	-2
11	Queensland	1887-11-28	40/ July 21 1888	1888	40	-1
12	Queensland	1890-05-26	38 / SEPT. 11 1890	1890	38	0
13	Queensland	1919-03-05	03.12.1981	1981	NA	-62
14	South_Australia	1877-03-05	34; 1835-1917	1835	34	42
15	South_Australia	1919-01-01	56 (1867)	1867	56	52
16	South_Australia	1927-02-28	47(51 in 1927)	1927	51	0
17	Tasmania	1892-08-11	25 (1862)	1862	25	30
18	Tasmania	1951-10-02	35 (1920)	1920	35	31
19	Western_Australia	1903-03-09	1867 OR 68	1867	68	36
20	Western_Australia	1933-12-05	40 (1888)	1888	40	45
21	Western_Australia	1934-12-04	37 (1890)	1890	37	44

Table after fixing:

	state	Date	age	birthyear	age_years	age_derived
1	NSW	1892-02-01	1855, 24	1855	24	37
2	NSW	1906-11-27	1866 36 (Link 1) 45 (Link 12)	1866	36	40
3	NSW	1913-11-26	1876, 29	1876	29	37
4	NSW	1916-05-25	28/ young man (4) / 1868	1868	28	48
5	NSW	1945-02-05	32, 1917	1917	32	28
6	NSW	1947-06-06	40 (1912)	1912	40	35
7	Queensland	1883-09-17	JAN.7 1885	1885	NA	-2
8	Queensland	1919-03-05	03.12.1981	1981	NA	-62
9	South_Australia	1877-03-05	34; 1835-1917	1835	34	42
10	South_Australia	1919-01-01	56 (1867)	1867	56	52
11	Tasmania	1892-08-11	25 (1862)	1862	25	30
12	Tasmania	1951-10-02	35 (1920)	1920	35	31
13	Western_Australia	1933-12-05	40 (1888)	1888	40	45
14	Western_Australia	1934-12-04	37 (1890)	1890	37	44

After first cleaning

There are 2165 cases that have a 'clean' age (either extracted directly from the recorded 'age' or derived from extracted birthyear), which is 70.7%. The following table shows the cases where something was originally recorded in the age column, but where an age couldn't be extracted by my code. We see that for most cases, it's because there are only words like "young" or "middle-aged" with no numerical value, but in a few it's because the value recorded was ambiguous in a way I didn't deal with in my code: "30s", "50+", and "34*". For these cases, I will subjectively assign"35", "50", and "34".

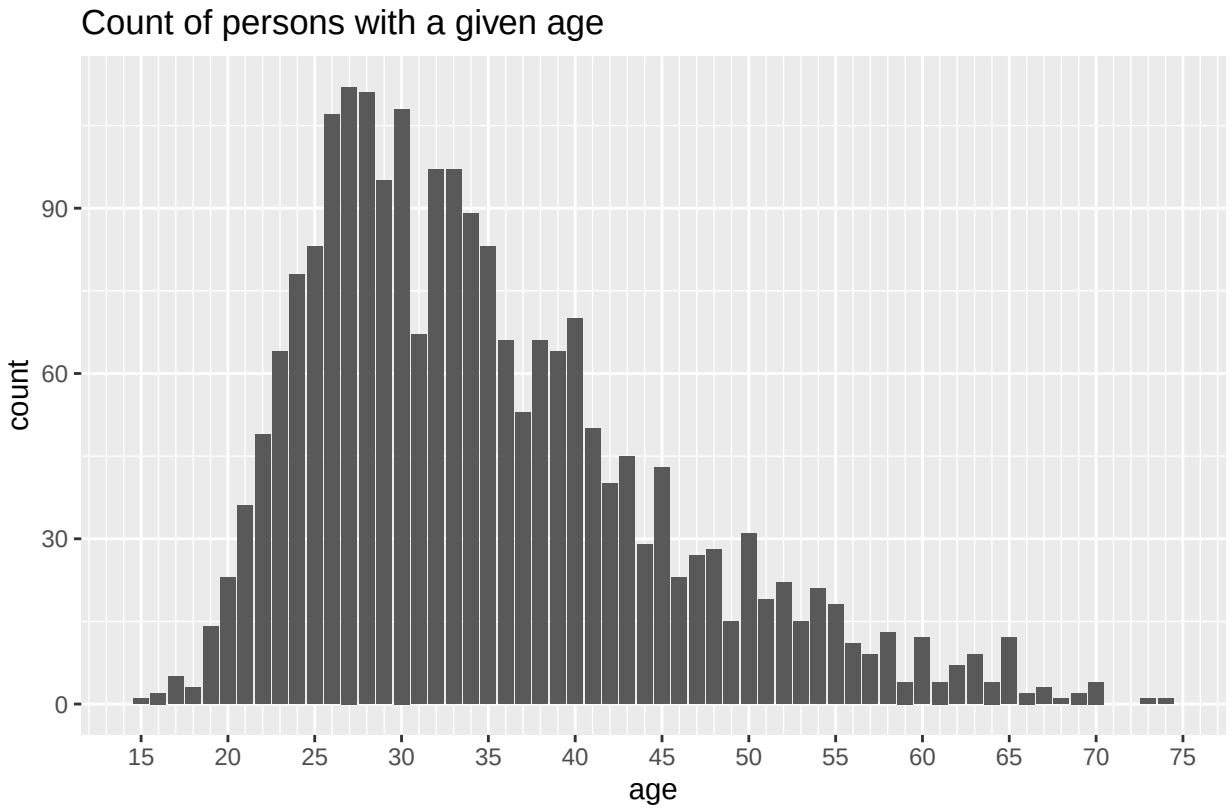
	state	Date	age
1	NSW	1836-12-13	”aged”
2	NSW	1863-05-01	ELDERLY
3	NSW	1914-01-31	Young
4	NSW	1916-10-09	"looked beyong middle age" (8)
5	NSW	1917-01-02	young man (2)
6	NSW	1931-06-22	30s
7	NSW	1932-02-11	Young man
8	Queensland	1867-04-22	young
9	Queensland	1883-09-17	JAN.7 1885
10	Queensland	1884-01-12	middle age
11	Queensland	1912-01-31	"An elderly man"
12	Queensland	1919-03-05	03.12.1981
13	Queensland	1919-07-11	in his twenties
14	Queensland	1919-08-22	young woman
15	Queensland	1920-12-09	repeatedly described as beign a young woman
16	Queensland	1926-10-01	middle-aged
17	South_Australia	1866-05-17	middle aged
18	Tasmania	1873-08-26	50+
19	Tasmania	1926-02-11	34*
20	Tasmania	1932-09-20	young
21	Western_Australia	1874-10-07	middle age
22	Western_Australia	1897-04-07	middle aged OR elderly
23	Western_Australia	1897-09-14	middle aged
24	Western_Australia	1938-02-01	young
25	Victoria	1875-07-14	middle age
26	Victoria	1876-05-15	middle age
27	Victoria	1879-07-28	elderly
28	Victoria	1882-06-01	young
29	Victoria	1885-07-06	middle age
30	Victoria	1893-02-16	elderly
31	Victoria	1899-03-17	middle age
32	Victoria	1905-09-01	middle age
33	Victoria	1919-11-28	middle age
34	Victoria	1920-10-26	middle age
35	Victoria	1921-02-18	middle age

Quick Analysis

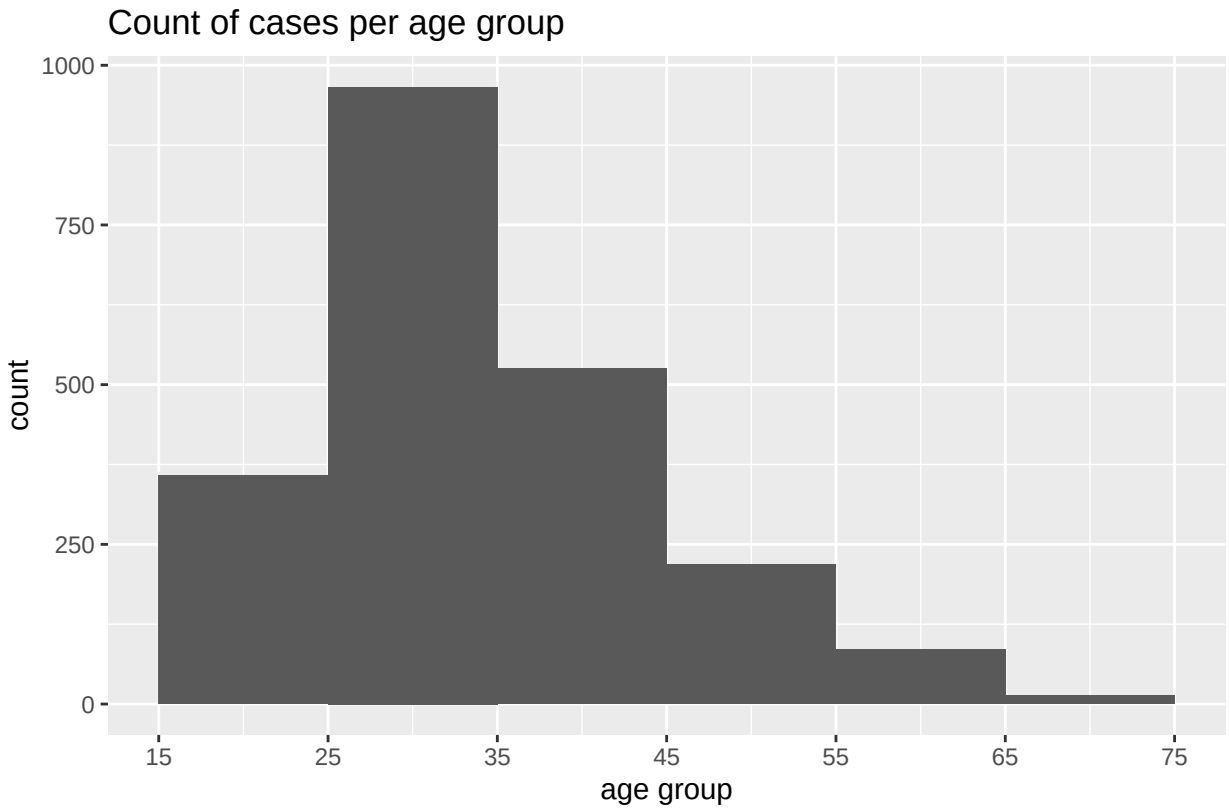
For the cases where the age recorded and the age derived disagreed, I just used the age recorded.

General Distribution

A quick distribution plot follows. We see that the ages range from 15 to 74. The distribution is not uniform (not every age has the same count of cases), but rather there are more cases with ages 25-35. We see an uncharacteristic drop at 31 years, where there seems to be at least 80 counts for every year between 25 and 35, but at 31, there are only 67.

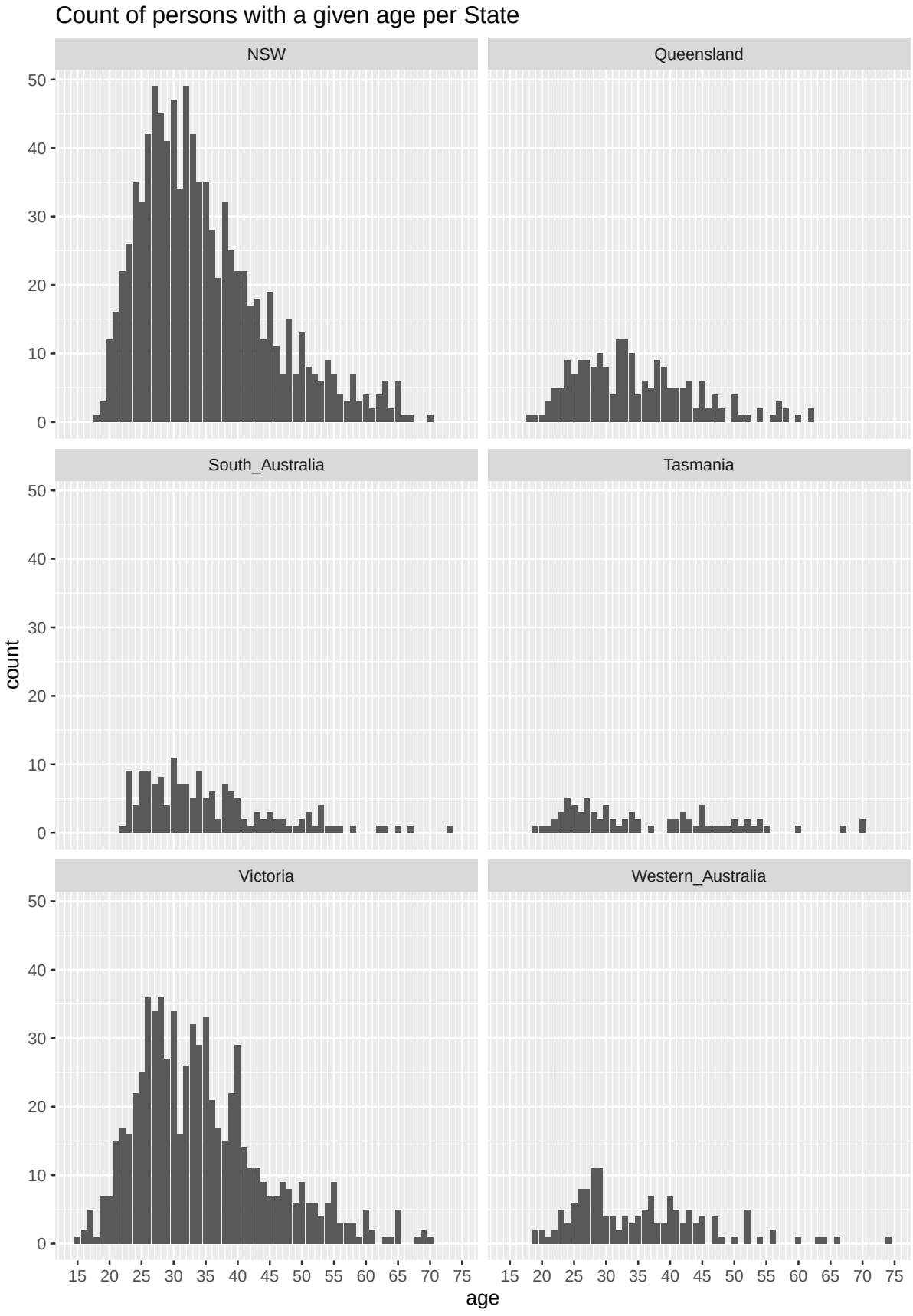


In the following plot we see the counts, split in groups of 10 years (15-24, 25-34, etc). We see that the most occurring category is the 25-34 year olds, with nearly 1000 cases falling in this category. The least occurring is 65-74, with only 26 cases.

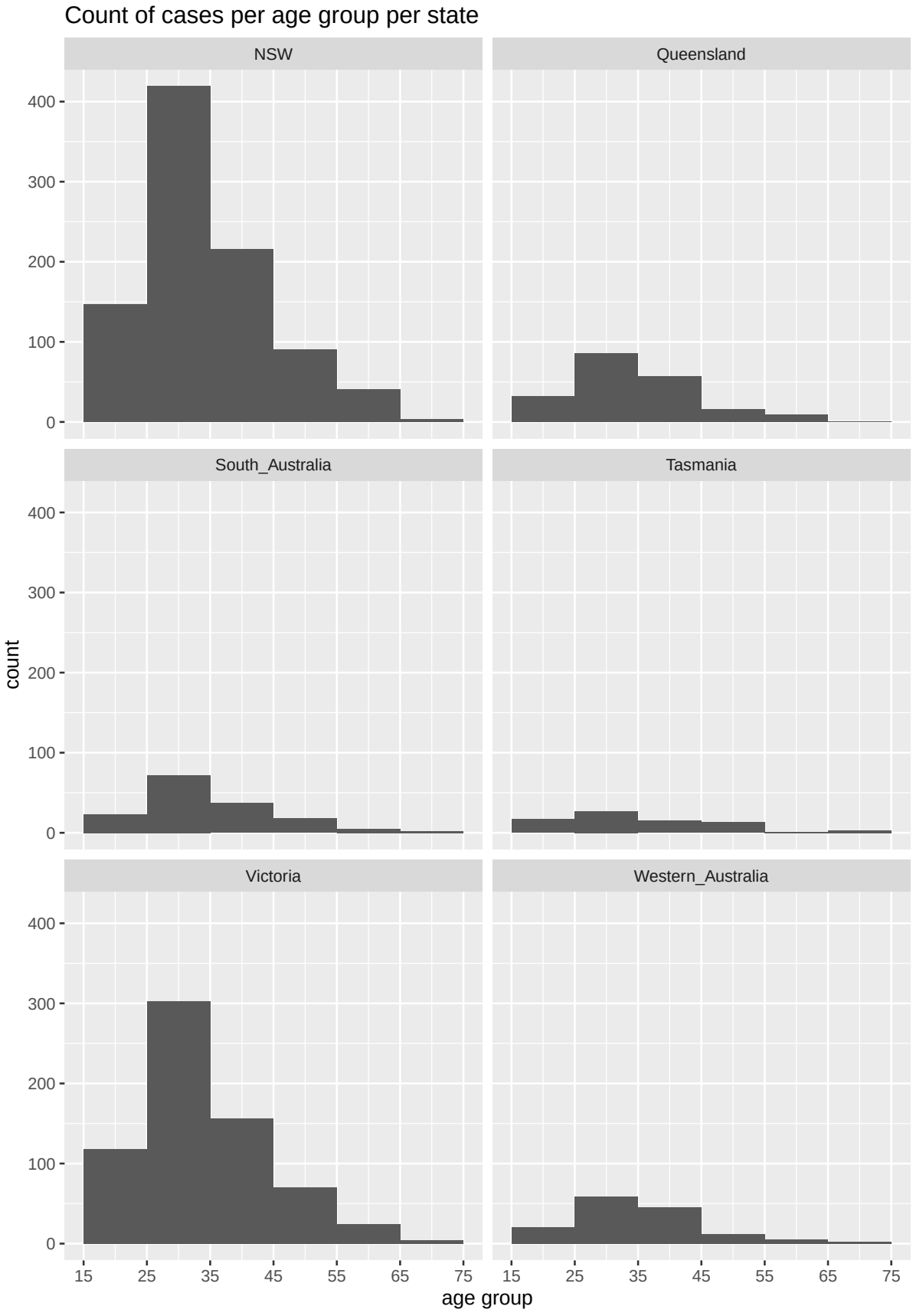


By State

We see that the trends are somewhat similar for each state; obviously, many more cases in NSW overall. In Tasmania, the distribution looks more uniform: no age has over 5 cases, and only 5 cases over the age of 55. South Australia is the only state with no cases under the age of 23.

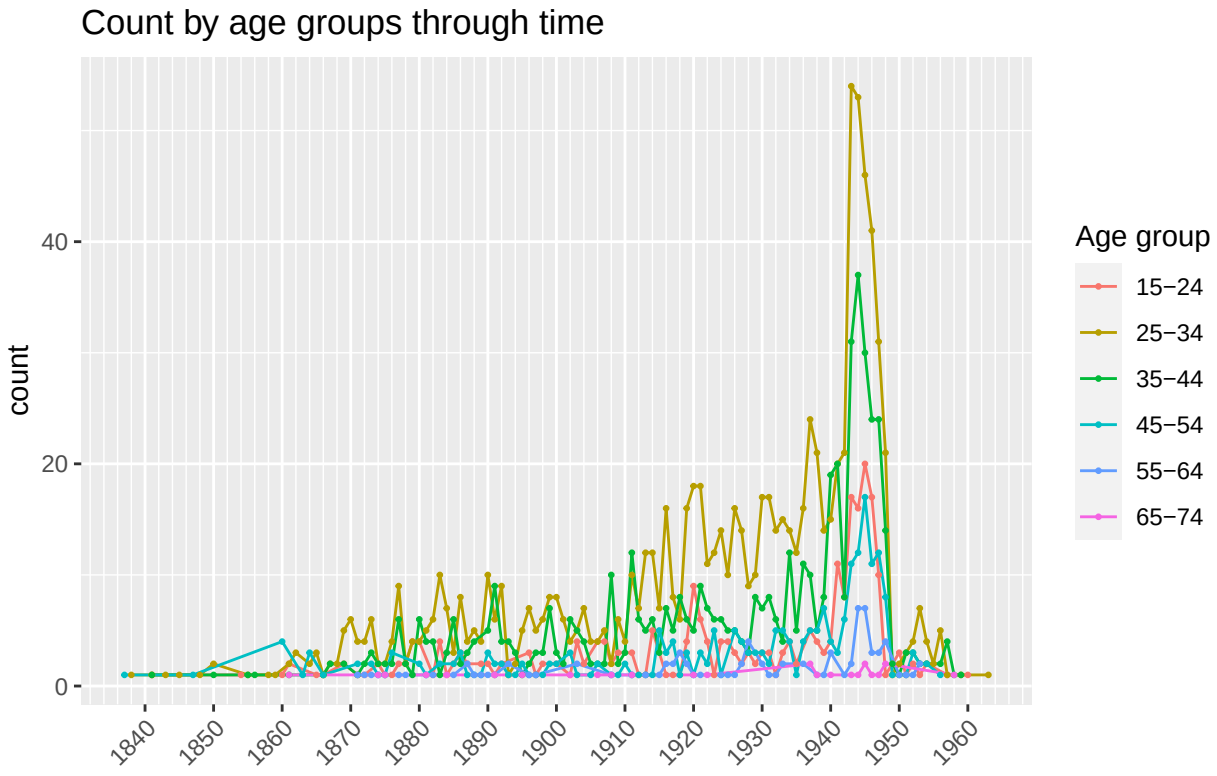

Count of persons with a given age per State
NSW
Queensland
South_Australia
Tasmania
Victoria
Western_Australia
count
age
0
10
20
30
40
50
15
20
25
30
35
40
45
50
55
60
65
70
75

In the following plot, we see that all states have the most cases in the age group of 25-34, although in Western Australia and Tasmania, other age groups are following not far behind in count. NSW and Victoria have the biggest count difference between their largest age group and the next largest (35-44 year olds in both states / all states except Tasmania).

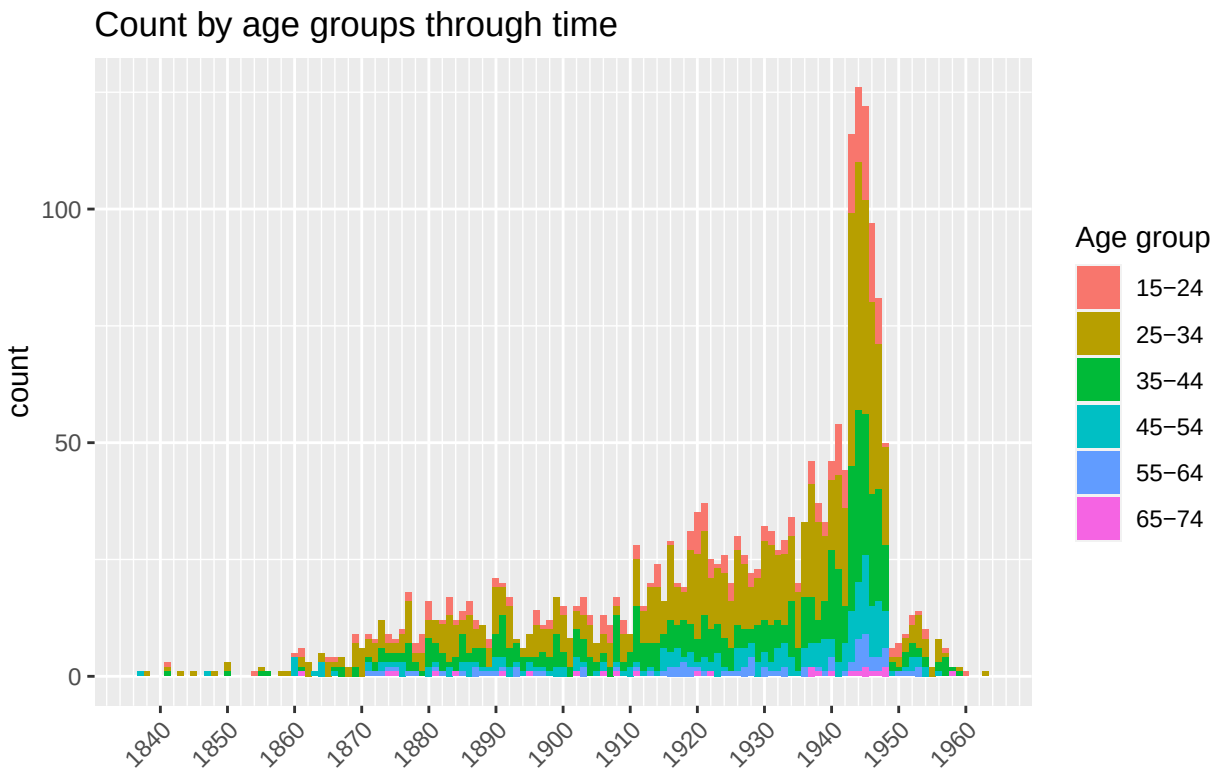

Count of cases per age group per state
NSW
Queensland
South_Australia
Tasmania
Victoria
Western_Australia
count
0
100
200
300
400
15
25
35
45
55
65
75
age group

Through time

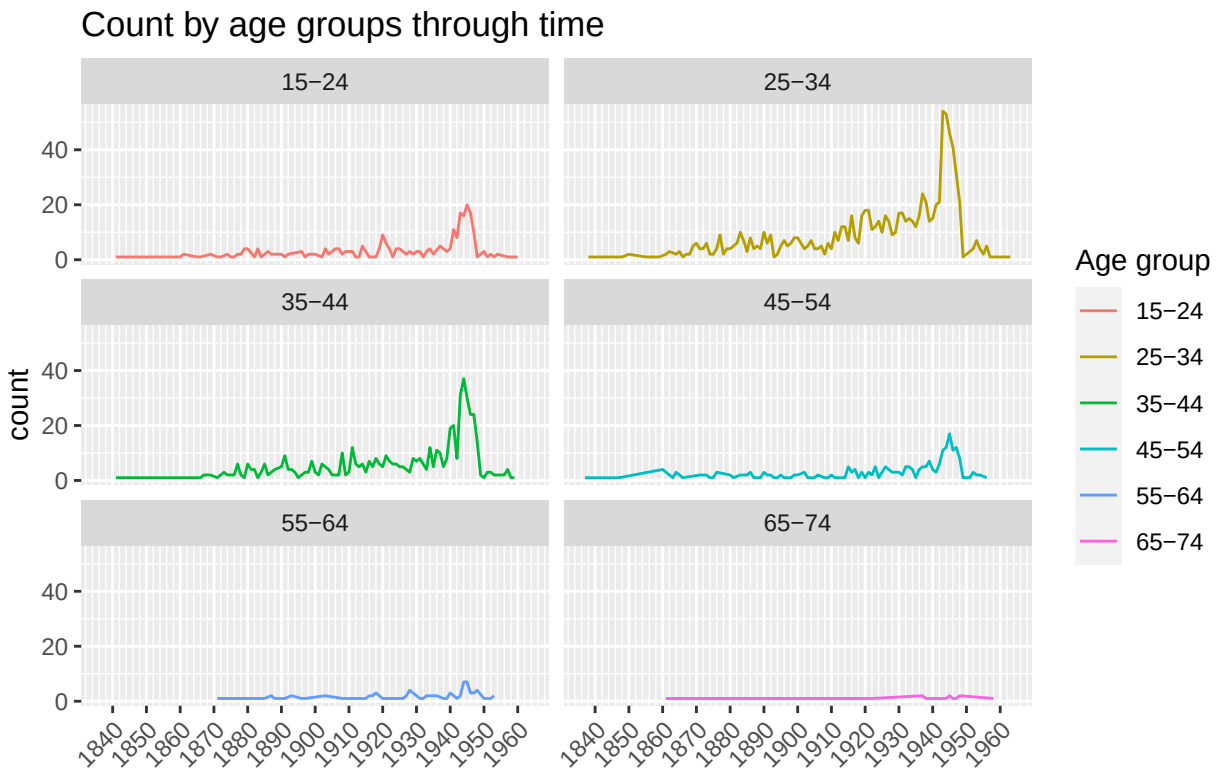
We see that for the most years, the age group 25-34 occurs most frequently, with the exception some years like 1908 and 1911 where there were more ages between 35-44. You see with this line and point chart how the line connects each point across the years, giving the impression that in the years from 1850-1860 there were more 45-54 year olds, when in fact, there were none in the years that don't have a point on the line. You can see more clearly that that is the case in the following bar plot. The line and point plot is somewhat easier to compare between the groups directly, but may be misinterpreted so. Its hard to compare 6 different groups on one chart as any plot will get messy.



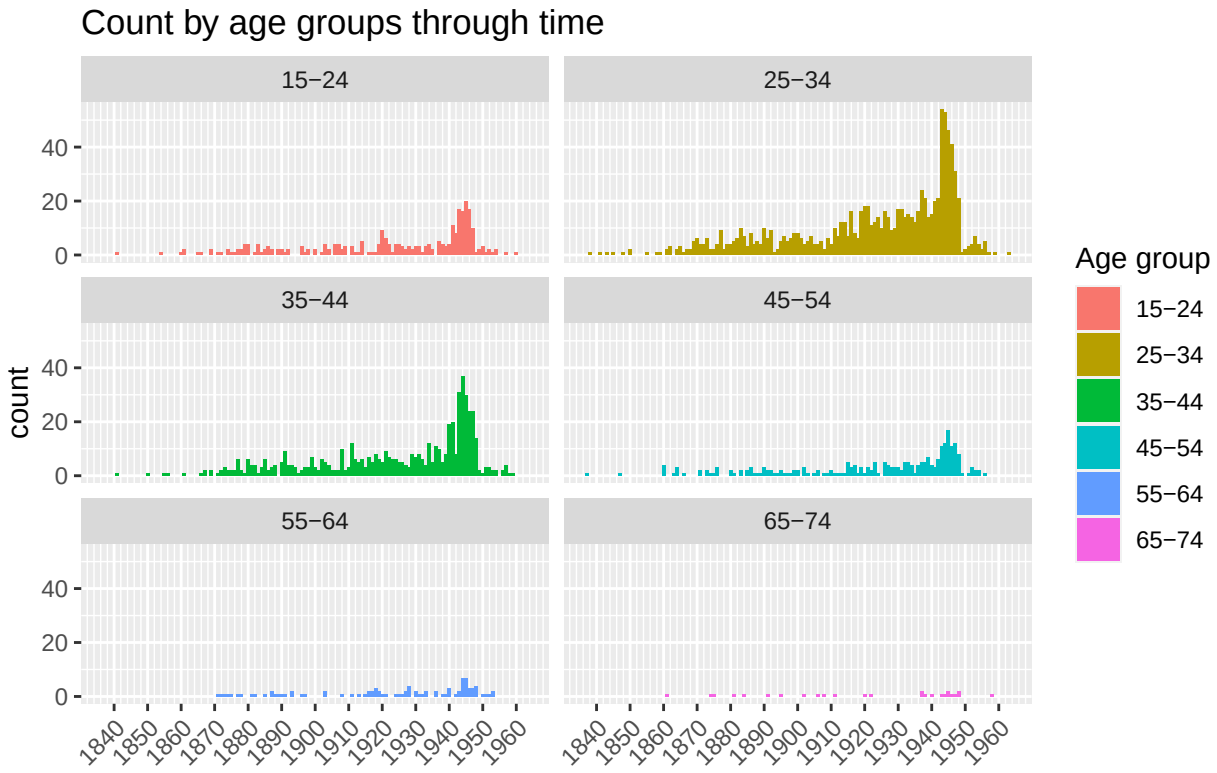
In this bar plot, you can see the total counts through the years, and the bar split in color according to age group proportions. You can see how a given year is split by age group, but it is hard to compare year after year the trends in a specific group.



The next plot splits the line plot of each group into their own plot, so that it is less messy; you can see the trend for each group, but it is harder to compare directly which group has more in a given year.



In this barplot, you can see how cases of 65-74 years old are rare, and are the only age group that didn't really increase much in numbers during WWII. You can clearly see that there were no cases with 55-64 year olds before 1871, and cases with 65-74 year olds were sparse through the years.


Count by age groups through time
15–24
25–34
35–44
45–54
55–64
65–74
count
0
20
40
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
Age group
15–24
25–34
35–44
45–54
55–64
65–74