TheOld

Hyeseon Seo

2023-09-21

1. Very old people: Supercentenarians are people who have lived to more than 110 years of age. We suspect many of these people are in fact younger than 110 and are claiming either fraudulently or accidentally that they are older than they in fact are. In particular, we are interested in uncovering statistical evidence of fraud by looking at birthdays and whether the rate of supercentenarians dropped with the introduction of birth registration in the US.

library("tidyr")  
library("ggplot2")  
old = read.delim("/Users/woojeongkim/Desktop/23 Fall/S690/Old people/supercentenarians.txt")  
names(old)

## [1] "Nr" "Name..generally.known.as."  
## [3] "Born" "Died"   
## [5] "Years" "Days"   
## [7] "Days.Lived" "Rank.at.Death"   
## [9] "R" "S"   
## [11] "Birthplace" "Residence.Place.of.death"   
## [13] "Validated.By" "Validated.By2"   
## [15] "Validated.By3" "Validated.By4"   
## [17] "Validated.By5" "Application.Date"   
## [19] "Acceptance.Date" "Photo."

old = separate(old, col = Born, into = c("Born.date", "Born.yeaer"), sep = ",")  
  
Bdate = as.data.frame( table(old$Born.date))  
colnames(Bdate)= c("date","freq")  
  
length(old$Born.date) *# 1739*

## [1] 1739

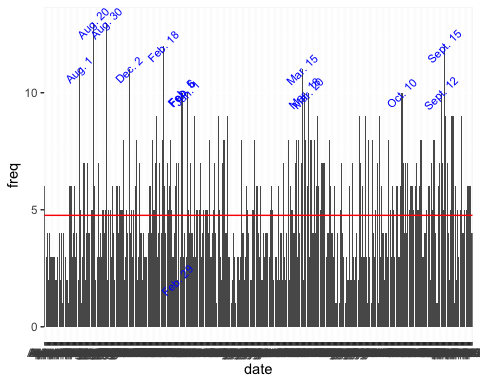
length(unique(old$Born.date)) *#361*

## [1] 361

*# we have approximately 365.24day*   
*# The probability of birthdate ( other than Feb.29) = 1/365.25*  
***## the probability of Feb.29th birthdate = 1/365.25\*4***  
  
exp = 1739/365.2 *# n \* p where n = 1739, p = 1/365.24*  
exp\_Feb29 = 1739/(365.25\*4)  
exp\_Feb29

## [1] 1.190281

ggplot(Bdate, aes(x = date, y = freq, label =date )) + geom\_bar(stat = "identity") +   
 geom\_hline(yintercept = exp, color = "red" ) +   
 geom\_text(data = subset(Bdate, freq > 9 ), size = 3, angle = 45, color = "blue") +  
 geom\_text(data = subset(Bdate, date == "Feb. 29" ), size = 3, angle = 45, color = "blue")



subset(Bdate, freq >9)

## date freq  
## 30 Aug. 1 11  
## 42 Aug. 20 13  
## 53 Aug. 30 13  
## 72 Dec. 2 11  
## 101 Feb. 18 12  
## 116 Feb. 5 10  
## 117 Feb. 6 10  
## 121 Jan. 1 10  
## 218 Mar. 15 11  
## 220 Mar. 18 10  
## 223 Mar. 20 10  
## 302 Oct. 10 10  
## 335 Sept. 12 10  
## 338 Sept. 15 12

subset(Bdate, freq <2)

## date freq  
## 16 Apr. 24 1  
## 21 Apr. 29 1  
## 32 Aug. 11 1  
## 146 Jan. 4 1  
## 158 July 15 1  
## 162 July 2 1  
## 246 May 13 1  
## 247 May 14 1  
## 249 May 16 1  
## 254 May 20 1  
## 273 Nov. 11 1  
## 276 Nov. 14 1  
## 282 Nov. 2 1  
## 331 Oct. 9 1  
## 332 Sept. 1 1  
## 346 Sept. 22 1

which(Bdate$date == "Feb. 29") *# 113*

## [1] 113

Bdate [ 113,] *#freq = 2*

## date freq  
## 113 Feb. 29 2

unique(sort(old$Born.yeaer)) *#1788 - 1903*

## [1] " 1788" " 1792" " 1807" " 1814" " 1815" " 1817" " 1820" " 1842" " 1849"  
## [10] " 1851" " 1852" " 1855" " 1856" " 1857" " 1858" " 1860" " 1861" " 1862"  
## [19] " 1863" " 1864" " 1866" " 1867" " 1868" " 1869" " 1870" " 1871" " 1872"  
## [28] " 1873" " 1874" " 1875" " 1876" " 1877" " 1878" " 1879" " 1880" " 1881"  
## [37] " 1882" " 1883" " 1884" " 1885" " 1886" " 1887" " 1888" " 1889" " 1890"  
## [46] " 1891" " 1892" " 1893" " 1894" " 1895" " 1896" " 1897" " 1898" " 1899"  
## [55] " 1900" " 1901" " 1902" " 1903"

*# exclude Feb.29, since this date has different expected value*  
Bdate1 = subset(Bdate, Bdate$date != "Feb. 29")  
chisq.test( Bdate1$freq)

##   
## Chi-squared test for given probabilities  
##   
## data: Bdate1$freq  
## X-squared = 409.94, df = 359, p-value = 0.03269

*# p-value( = 0.03) is small : this is not uniformly distributed.*  
*# Feb. 29 : observation = 2, expected = 1.190281*

# nation and state separation

library("dplyr")

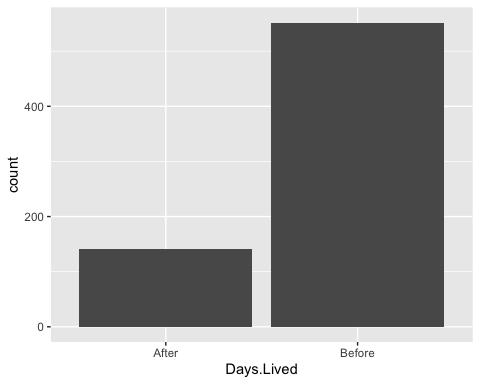
##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

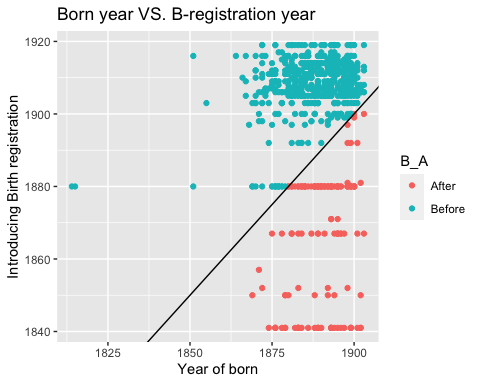
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

*# separate nation and state*   
df = old  
df = df[grep("USA",old$Birthplace),]  
df = df[-c(315,372,562,695),]  
df = df %>% separate\_wider\_delim(Birthplace,delim = " (",names = c("Country","State"))  
df$State<-gsub(")","",as.character(df$State))  
USA <- df[,c(2,3,4,8, 13)]  
USA$Reg <- USA$State  
  
*# recoding the registration year and the state accordingly*  
USA$Reg <- recode(USA$State,   
 "Alabama" = "1908", "Arizona"="1909","Arkansas"="1914",  
 "California"="1905","Colorado"="1907",  
 "Connecticut" = "1897","Delaware"="1881",   
 "District of Columbia"="1871", "Florida"="1899",   
 "Georgia"="1919", "Idaho"="1911","Illinois"="1916",  
 "Indiana"= "1907", "Iowa"="1880", "Kansas"="1911",   
 "Kentucky"="1911", "Louisiana"="1914", "Maine"="1892",  
 "Maryland" = "1898","Massachusetts" = "1841", "Michigan"="1867",  
 "Minnesota" = "1900", "Mississippi" ="1912", "Missouri" = "1910",  
 "Montana" ="1907","Nebraska" = "1905","Nevada"="1911",   
 "New Hampshire"="1850","New Jersey" = "1841","New Mexico" = "1919",  
 "New York" = "1880", "North Carolina"="1913",   
 "North Dakota"="1908", "Ohio" = "1909", "Oklahoma"= "1908",  
 "Oklahoma"= "1908", "Oklahoma Territory" = "1908",  
 "Oregon" = "1903", "Pennsylvania"="1906","Rhode Island" = "1852",  
 "South Carolina" = "1915", "South Dakota" = "1905",   
 "Tennessee"= "1914", "Texas" = "1903", "Utah" = "1905",   
 "Vermont" = "1857", "Virginia" = "1912", "Washington" = "1907",  
 "West Virginia"="1917", "Wisconsin" = "1907", "Wyoming" = "1909",  
 "Alaska" = "1913", "Hawaii" = "1896", "Puerto Rico" = "1931",  
 "Virgin Islands" = " 1919 " )

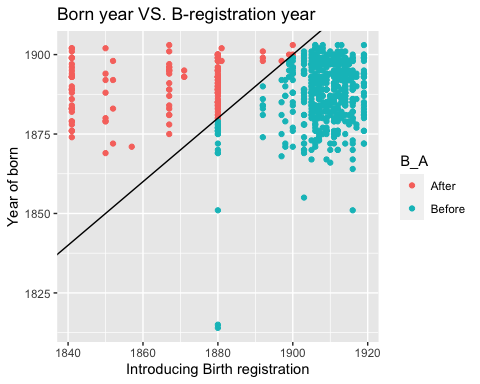
*# Categorizing before/after born year*  
USA$Born.yeaer = as.integer(USA$Born.yeaer)  
USA$Reg = as.integer(USA$Reg)  
USA = mutate(USA, B\_A = if\_else(Born.yeaer < Reg, "Before", "After" ))  
ggplot(USA, aes(x = Days.Lived )) + geom\_bar(aes(B\_A))



*# Born year vs B-registration year*  
ggplot(USA, aes(y= Reg, x = Born.yeaer, group\_by = B\_A, color = B\_A))+ geom\_point() +   
 geom\_abline(intercept = 0, slope = 1) +   
 labs( title = "Born year VS. B-registration year") +   
 ylab("Introducing Birth registration" ) +  
 xlab("Year of born")



ggplot(USA, aes(x= Reg, y = Born.yeaer, group\_by = B\_A, color = B\_A))+ geom\_point() +   
 geom\_abline(intercept = 0, slope = 1) +   
 labs( title = "Born year VS. B-registration year") +   
 xlab("Introducing Birth registration" ) +  
 ylab("Year of born")



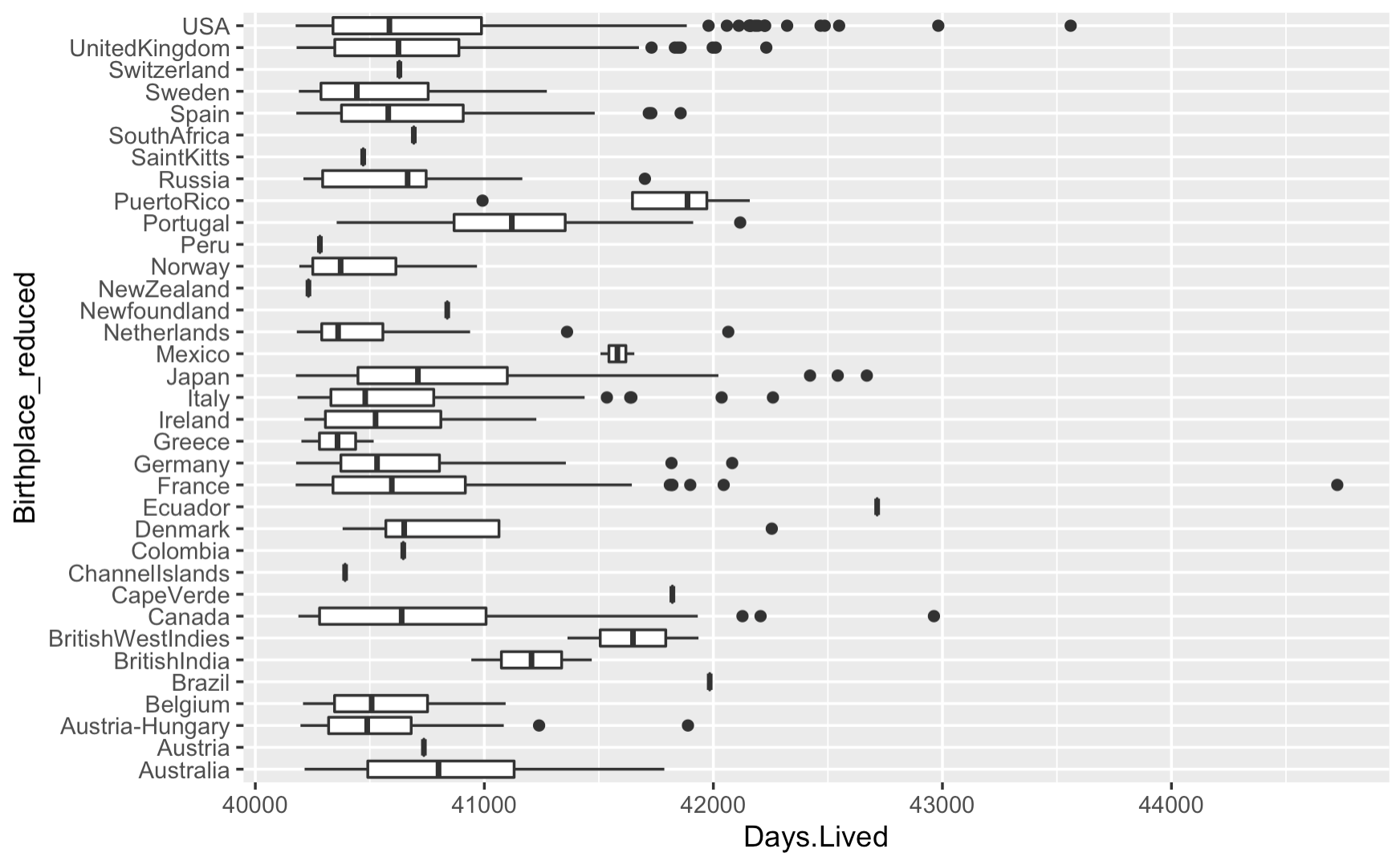
sort(unique(USA$Born.yeaer)) *# born year from 1814 to 1903 : registered in 2003*

## [1] 1814 1815 1851 1855 1864 1866 1867 1868 1869 1870 1871 1872 1873 1874 1875  
## [16] 1876 1877 1878 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890  
## [31] 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1903

sort(unique(USA$Reg))

EDA for contradictory claim on assumption

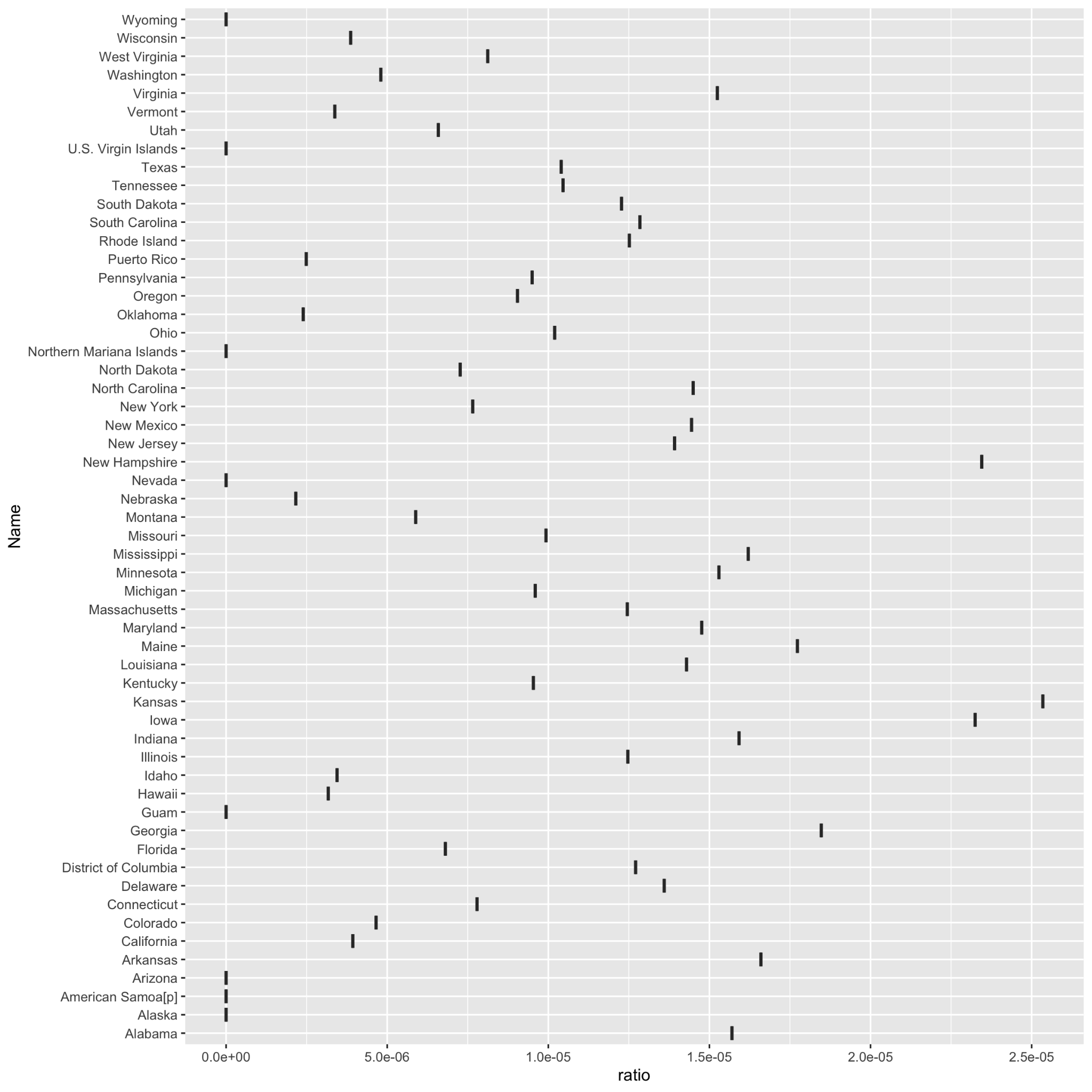
EDA1+Assumption contradiction: Box plots on the columns ‘Birthplace\_reduced vs. Days.Lived’



We should consider the total number of the population for each region, so we narrow the range of observing SC to the US only.

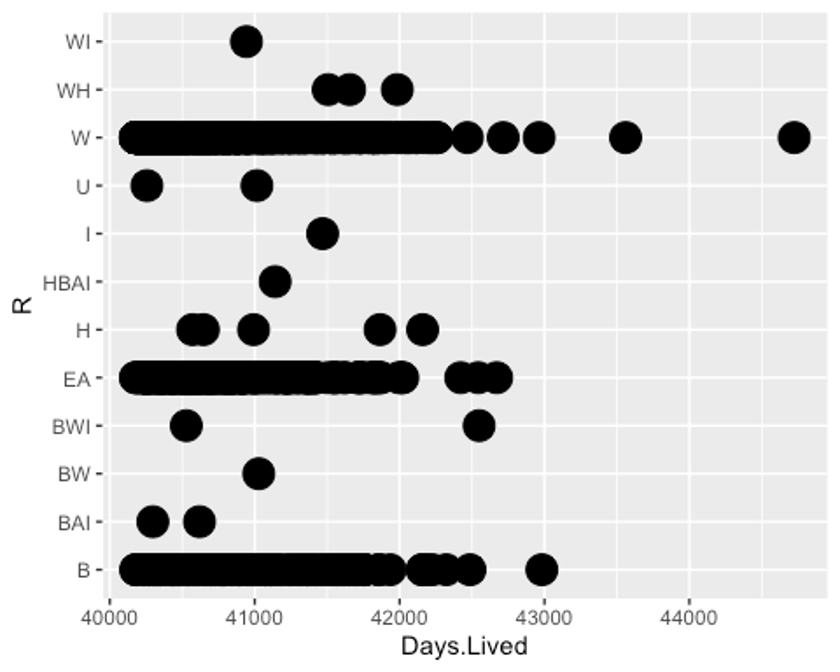
For each state, therefore, we have the number of SC people in the below graph which is shown after rescaling them as dividing by the total number of the population in each state.

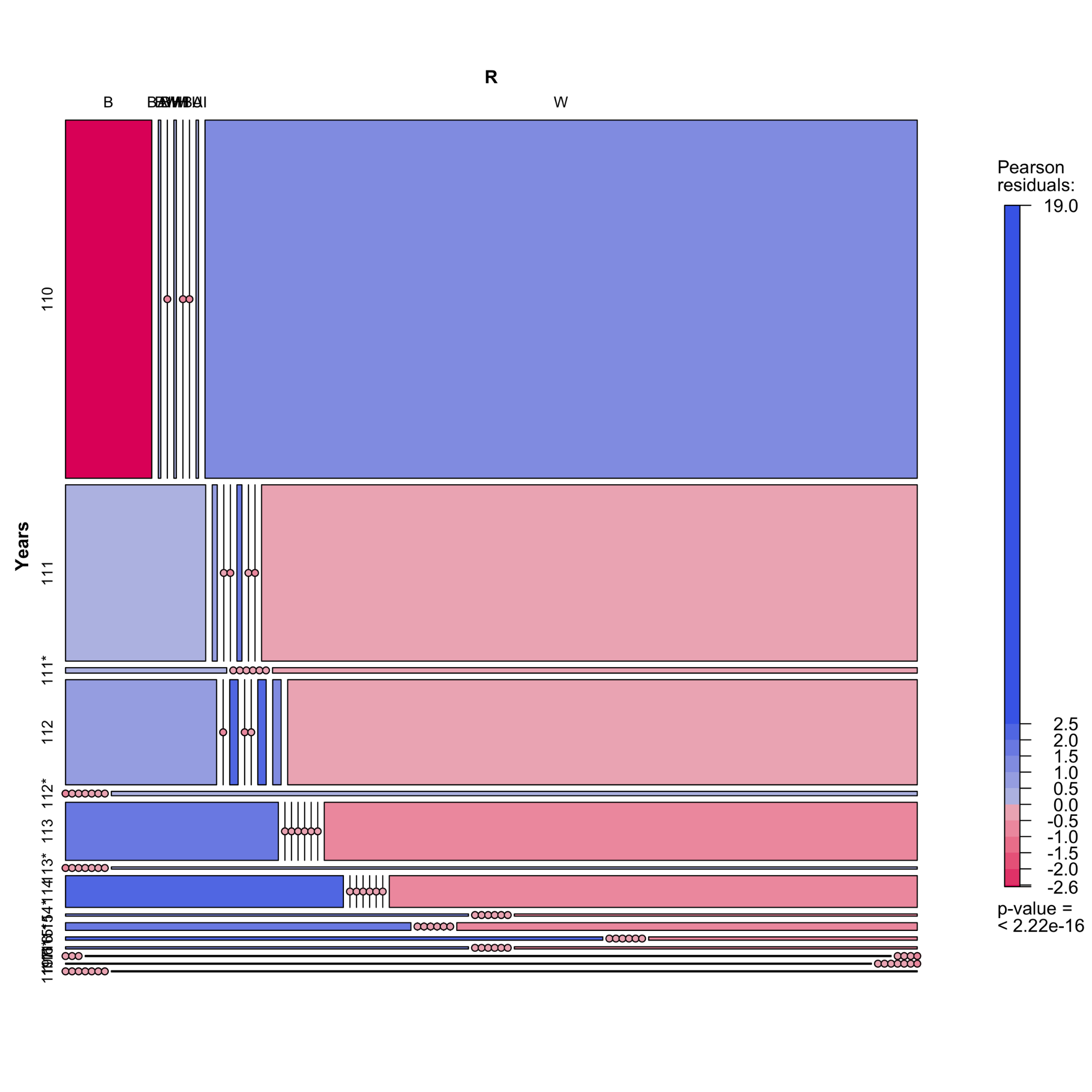
So, we consider the assumption to suspect fake recording of the number of SC. When we assume that the given condition is fair for generating SCs in each state except sex ratio and race ratio which are given in our data, It is very natural that every state has a uniform ratio of SCs according to each of their total population. But it was not like the graph below. Now we can suspect several states for the record of fake SC people, such as New Hampshir, Iowa and Kansas. So we need to observe the specific factor for these state’s high ratio of SC number with given column race ratio and sex ratio.



So, we are now see the relation between race and Days.Lived variable.

EDA2+Dependency between the columns Year and R:





For the second mosaic graph, it is clear that the people in race W and people in race B have the biggest amount for each SC’s year. Also, through the Years range on the vertical direction axis, each race category width does not shape a straight line which is shown with fixed width. Therefore, the race and Years are dependent as the result of the visualized graph above. This indicates that high ratio of the race of W in a state can imply a higher amount for the SCs than other states. This also implies that the most fake is done by the states having a high ratio of W, since the fake record is over a half in the graph for introducing birth registration.