CASE STUDY UNIT 8

MODELING RUNNERS' AGE IN THE CHERRY BLOSSOM RACE

Rajni Goyal, James Hosker, Damon Resnick

March 8th 2018

MSDS 7333: Quantifying the World (Thursday, 8:30 PM)

Outline

- Question #10 for the 10K Cherry Blossom (CB) race data
  - *From Chapter 2 of the text book: "Data Science in R, A Case Studies Approach to Computational Reasoning and Problem Solving" by Deborah Nolan and Duncan Temple Lang, CRC Press*
  - *Cherry Blossom Race Data website: http://www.cherryblossom.org/*
- Data Cleaning for Men and Women
- Q-Q Plot of Age for Men and Women
- Density Plots for relative Distribution of Age across Years for Men and Women
- Boxplot of Age of Each Year for Men and Women
- Violin Plot of Each Year for Men and Women
- Histogram of Age across Year for Men and Women

Question # 10

- We have seen that the 1999 runners were typically older than the 2012 runners (*using the 10K CB race data*).
  - *Compare the age distribution of the runners across all 14 years of the races.*
  - *Use quantile–quantile plots, boxplots, and density curves to make your comparisons.*
  - *How do the distributions change over the years?*
  - *Was it a gradual change?*

Data Set Summary

- Annual results spanning from 1999 to 2012 (14 years)
- Main Fields are - Place, Number, Name, Age, Hometown, Gun time, Net time and Time (Gun time is the official time)
- Some of the results include Pace and Div/Tot as well
- Men

```
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
3196 3017 3623 3724 3948 4156 4327 5237 5276 5906 6651 6911 7011 7193
```

```
> summary(cbMen)
      year          sex                name               home                age            runTime
 Min.   :1999   Length:70072       Length:70072       Length:70072       Min.   : 0.00   Min.   :  1.50
 1st Qu.:2004   Class :character   Class :character   Class :character   1st Qu.:30.00   1st Qu.: 77.23
 Median :2007   Mode  :character   Mode  :character   Mode  :character   Median :37.00   Median : 86.53
 Mean   :2007                                                           Mean   :38.61   Mean   : 87.48
 3rd Qu.:2010                                                           3rd Qu.:46.00   3rd Qu.: 96.60
 Max.   :2012                                                           Max.   :89.00   Max.   :175.60
                                                                        NA's   :23      NA's   :2
```

```
      age
 Min.   : 0.00
 1st Qu.:30.00
 Median :37.00
 Mean   :38.61
 3rd Qu.:46.00
 Max.   :89.00
 NA's   :23
```

Men

- Women

```
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
2356 2167 2973 3335 3544 3899 4336 5437 5692 6397 8325 8855 9030 9730
```

```
> summary(cbwomen)
      year          sex                name               home                age            runTime
 Min.   :1999   Length:75974       Length:75974       Length:75974       Min.   : 0.00   Min.   : 51.73
 1st Qu.:2005   Class :character   Class :character   Class :character   1st Qu.:27.00   1st Qu.: 88.53
 Median :2008   Mode  :character   Mode  :character   Mode  :character   Median :32.00   Median : 97.33
 Mean   :2007                                                           Mean   :33.85   Mean   : 98.09
 3rd Qu.:2010                                                           3rd Qu.:39.00   3rd Qu.:106.78
 Max.   :2012                                                           Max.   :87.00   Max.   :177.52
                                                                        NA's   :23      NA's   :2
```

```
      age
 Min.   : 0.00
 1st Qu.:27.00
 Median :32.00
 Mean   :33.85
 3rd Qu.:39.00
 Max.   :87.00
 NA's   :23
```

Women

Data Cleaning for Men

- 23 entries were identified and removed with null values for age.
  - *These records are unable to be visualized for age distributions.*

	year <int>	sex <chr>	name <chr>	home <chr>	age <dbl>	runTime <dbl>
1999.1083	1999	M	Rob Faye	Vienna VA	NA	78.25000
2001.3562	2001	M	F OPEN guideline *	er USATF Age-Group	NA	NA
2002.1227	2002	M	William STEIGEL	Washington DC	NA	77.70000
2002.2491	2002	M	Dave BOYER	Washington DC	NA	91.31667
2002.3724	2002	M	TF OPEN guideline *	der USATF Age-Group	NA	NA
2005.267	2005	M	John Marquart	Madison WI	NA	66.01667
2005.542	2005	M	Runner Dx Iii X Viii	Rockville MD	NA	69.76667
2005.925	2005	M	Runner Cb Vii Ix V	Rockville MD	NA	77.55000
2005.1158	2005	M	Runner Xx	Rockville MD	NA	78.80000
2005.1791	2005	M	Joseph Ferguson	Richmond VA	NA	86.68333
2005.2129	2005	M	Runner Bc Iii Ii O	Rockville MD	NA	87.00000
2005.3464	2005	M	Ronald Henry	Arlington VA	NA	97.16667
2005.3874	2005	M	Unidentified Runner Xx	Rockville MD	NA	103.75000
2005.3893	2005	M	Runner Cf Vii Iii Iv	Rockville MD	NA	112.61667
2005.4200	2005	M	Runner Ch Iii Viii Iii	Rockville MD	NA	116.38333
2007.213	2007	M	Jim Catella	Washington DC	NA	64.18333
2007.2657	2007	M	Monty Hoffman	Potomac MD	NA	86.33333
2007.2904	2007	M	Unknown		NA	87.86667
2010.27	2010	M	Dan Kahn	Houston TX	NA	52.60000
2010.329	2010	M	Bart Forsyth	Arlington VA	NA	65.45000
2010.1180	2010	M	Aaron Griggs		NA	74.98333
2010.5741	2010	M	Mike Hutchinson	Washibgton DC	NA	104.08333
2012.7192	2012	M	Joseph White	Forestville MD	NA	148.96667

- 10 observations were identified with an age less than 10 and were removed.
  - *Although young child can compete in the 10K run, there is a separate CB race (see website[1]) on a different day for young children between the ages of 4 to 10.*

	year <int>	sex <chr>	name <chr>	home <chr>	age <dbl>	runTime <dbl>
2001.1377	2001	M	Steve PINKOS	Washington DC	0	80.11667
2001.3003	2001	M	Jeff LAKE	Clarksville MD	0	99.43333
2001.3052	2001	M	Greg RHODE	Washington DC	0	97.23333
2002.2163	2002	M	Arlon WILBER	Durham NC	4	88.93333
2002.3282	2002	M	Nicholas RUGH	Arlington VA	1	103.75000
2003.1337	2003	M	John Riedel	Annapolis MD	2	79.35000
2003.2376	2003	M	Robert Anderson	Washington DC	0	88.80000
2007.4723	2007	M	Angelo Morelli	Huntingdon PA	9	105.71667
2011.5669	2011	M	Jake Ravitch	Bethesda MD	8	102.50000
2012.5221	2012	M	Jake Ravitch	Bethesda MD	9	96.60000

1. See http://www.cherryblossom.org/theraces/kidsrun.php

Data Cleaning for Women

- 23 entries were identified and removed with null values for age.
  - *These records are unable to be visualized for age distributions.*

	year <int>	sex <chr>	name <chr>	home <chr>	age <dbl>	runTime <dbl>
1999.3	1999	F	Lidiya Grigoryeva	Russia	NA	53.66667
1999.8	1999	F	Gladys Asiba	Kenya	NA	54.83333
1999.17	1999	F	Connie Buckwalter	Lancaster PA	NA	59.60000
1999.2175	1999	F	Ann Reid	Bethesda MD	NA	113.05000
2001.2973	2001	F	F OPEN guideline *	er USATF Age-Group	NA	NA
2002.270	2002	F	Unknown RUNNER	Washington DC	NA	78.98333
2002.1281	2002	F	Melissa AKEY	Washington DC	NA	94.20000
2002.2184	2002	F	Yvonne BONNER	Alexandria VA	NA	102.13333
2002.3261	2002	F	Unnamed Athlete	Unknown	NA	124.91667
2002.3335	2002	F	TF OPEN guideline *	der USATF Age-Group	NA	NA
2005.151	2005	F	Ashley Griffin	Washington DC	NA	75.06667
2005.159	2005	F	Lindsay Vogtsberger	Arlington VA	NA	68.81667
2005.742	2005	F	Angelica Jimenez	Washington DC	NA	81.41667
2005.1455	2005	F	Runner Xxxii	Rockville MD	NA	87.30000
2005.2454	2005	F	Michelle Hinman		NA	99.21667
2005.2496	2005	F	Xandra Brandon	Washington DC	NA	99.01667
2005.3241	2005	F	Michelle Merola	Washington DC	NA	105.25000
2005.3950	2005	F	Nancy Samko	Coraopolis PA	NA	117.60000

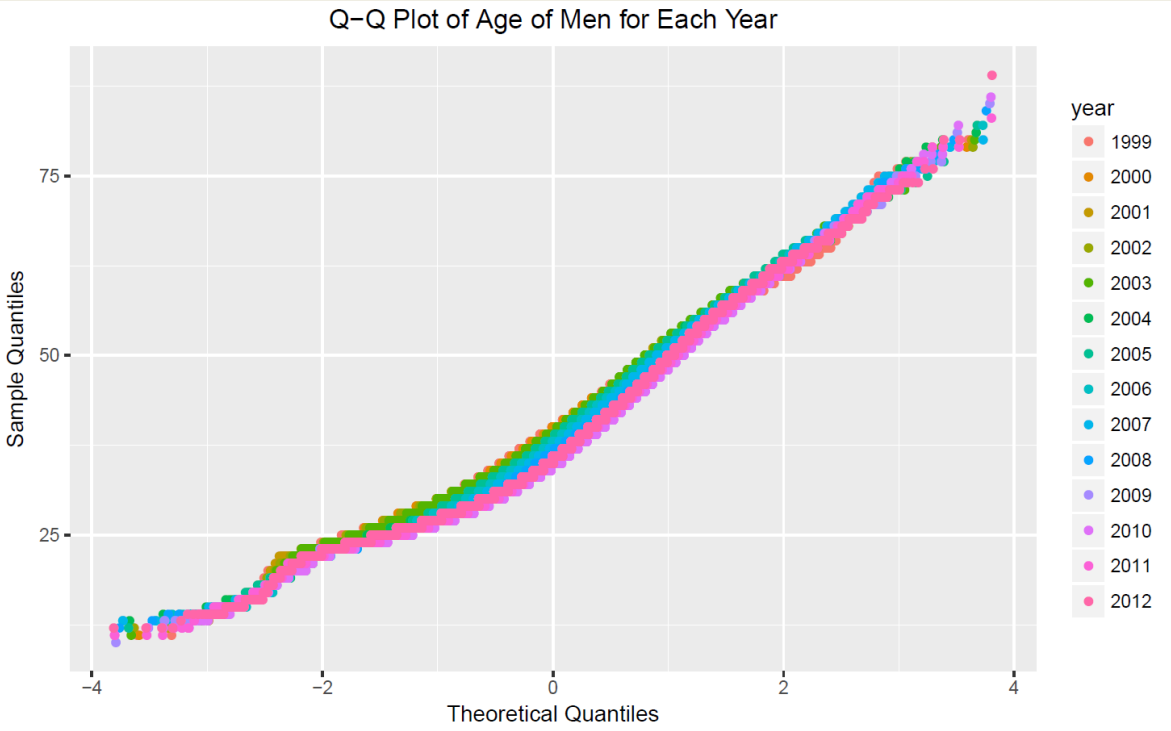
- 2 observations were identified with an age less than 10 and were removed.
  - *Although young child can compete in the 10K run, there is a separate CB race (walk/run, see website[1]) on a different day for young children between the ages of 4 to 10.*

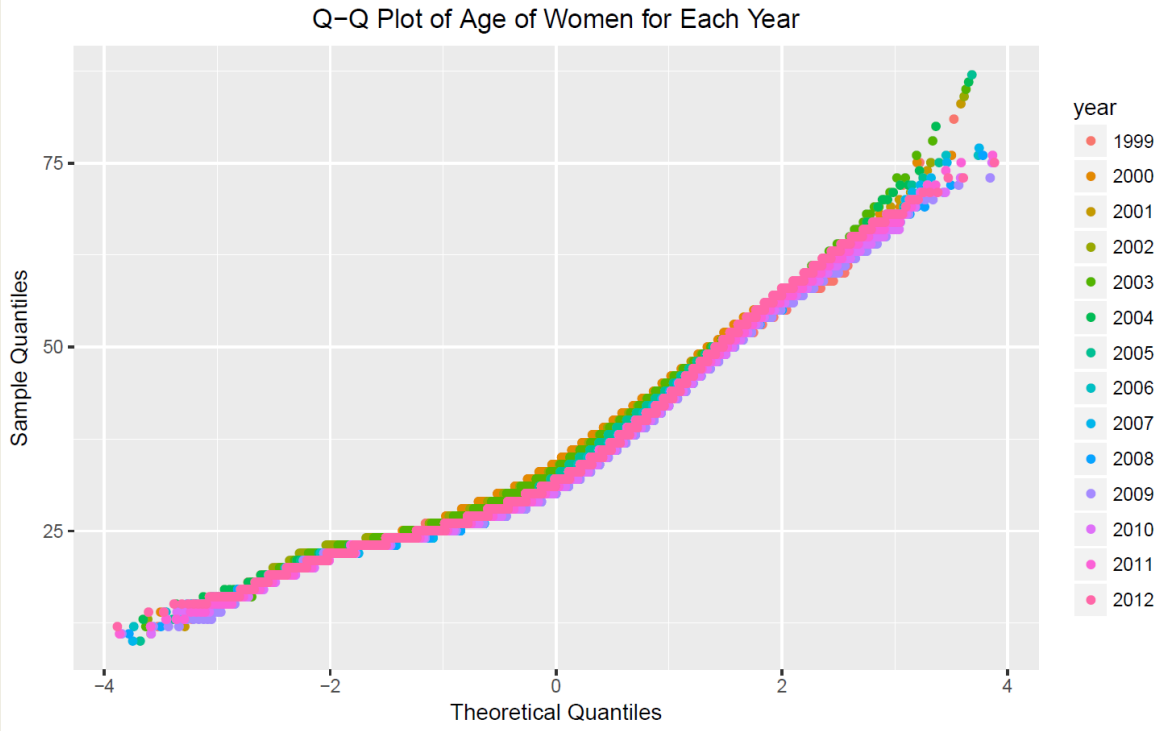
	year <int>	sex <chr>	name <chr>	home <chr>	age <dbl>	runTime <dbl>
2001.2611	2001	F	Loretta CUCE	Alexandria VA	0	113.6333
2009.6624	2009	F	Sydney Garrett	Newport News VA	7	109.8667

1. See http://www.cherryblossom.org/theraces/kidsrun.php

Q-Q Plots for Age for Men and Women

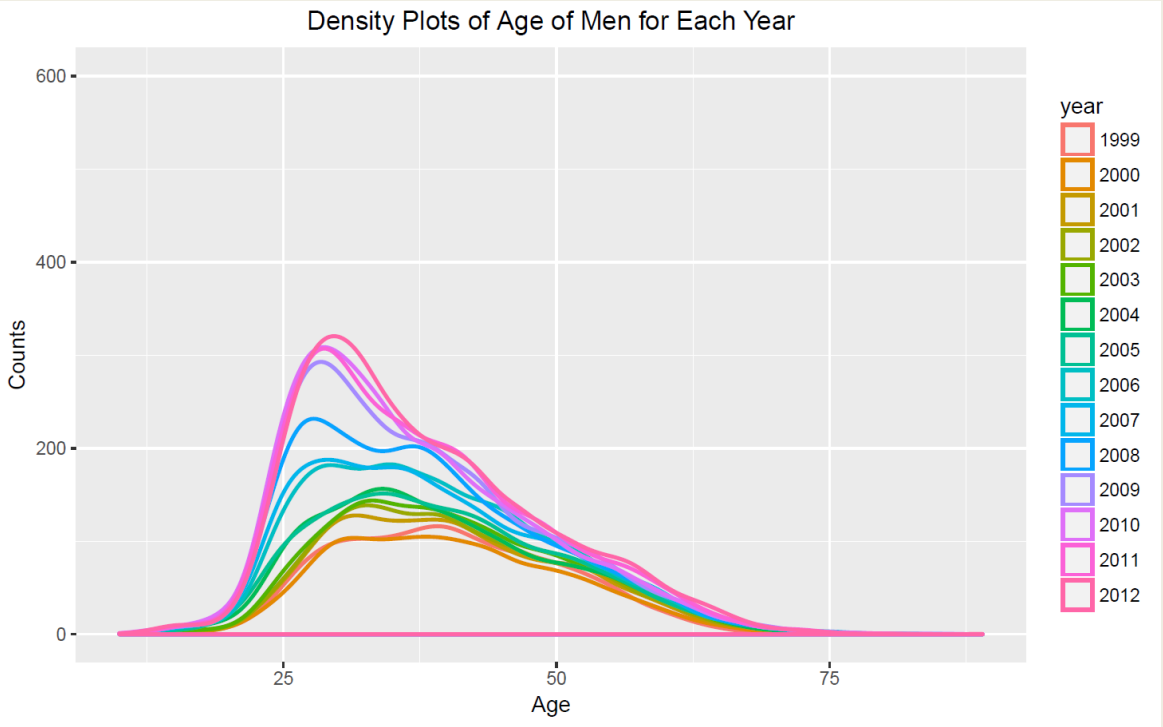
- For the most part, age follows a fairly normal distribution with some non-normal behavior at the right tails (quantile 4).
- More normality from 1999 to 2004 but a bend in the curve (less normality) starting in 2008 to 2010 that becomes more pronounced by 2012.

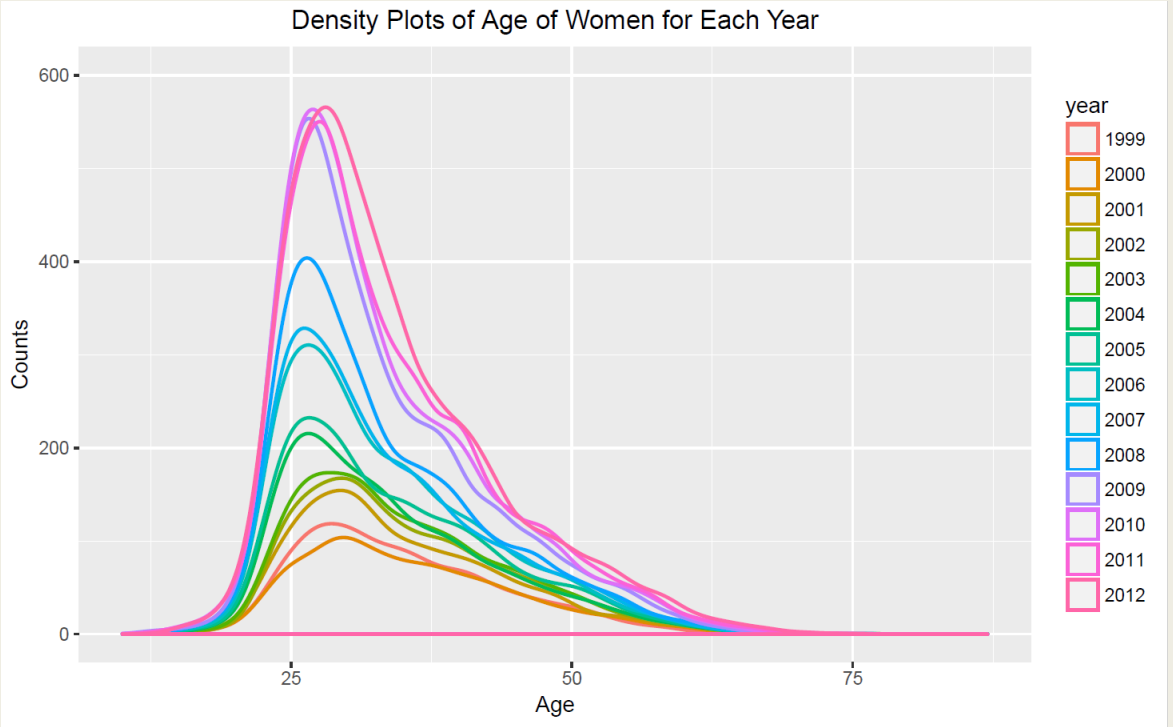




Distribution of Age across Years for Men & Women

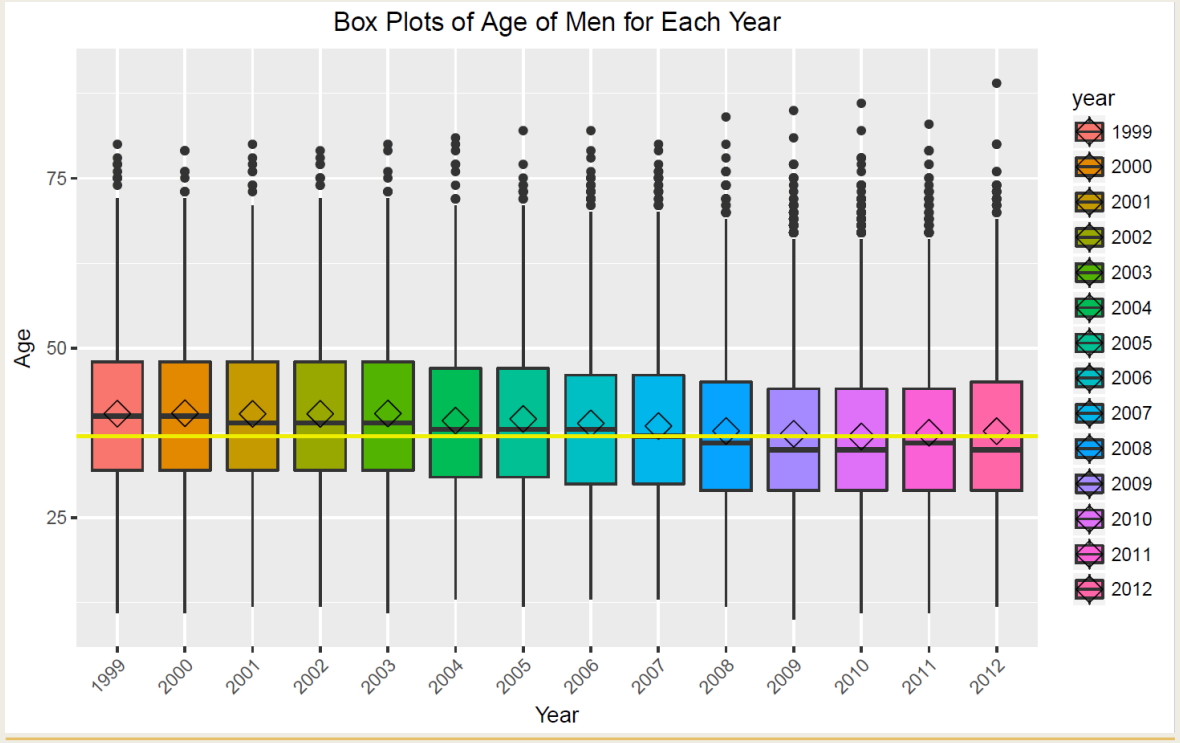
- From 1999-2012, there has been a gradual decrease in the mean age (younger entrants) for men and women in the CB race
- One way to explain this is that more younger runners entered the race in recent years while the number of older runners increases more gradually

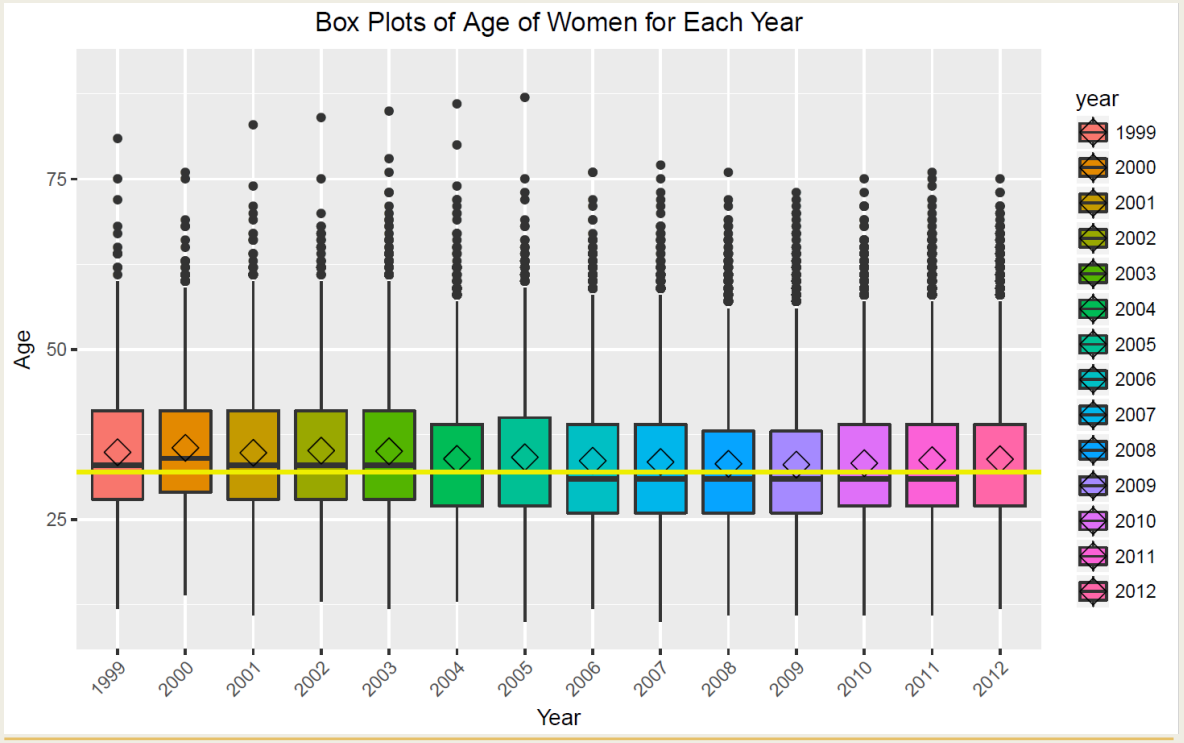




Boxplots for Age by Year for Men & Women

- Gradual Decrease in Mean and Median Age for Both Genders over the Years
  - More pronounced decrease for men than women as shown by comparing to the median of age over all years, the yellow line.
  - Most apparent trend was that median age gradually decreases over time.



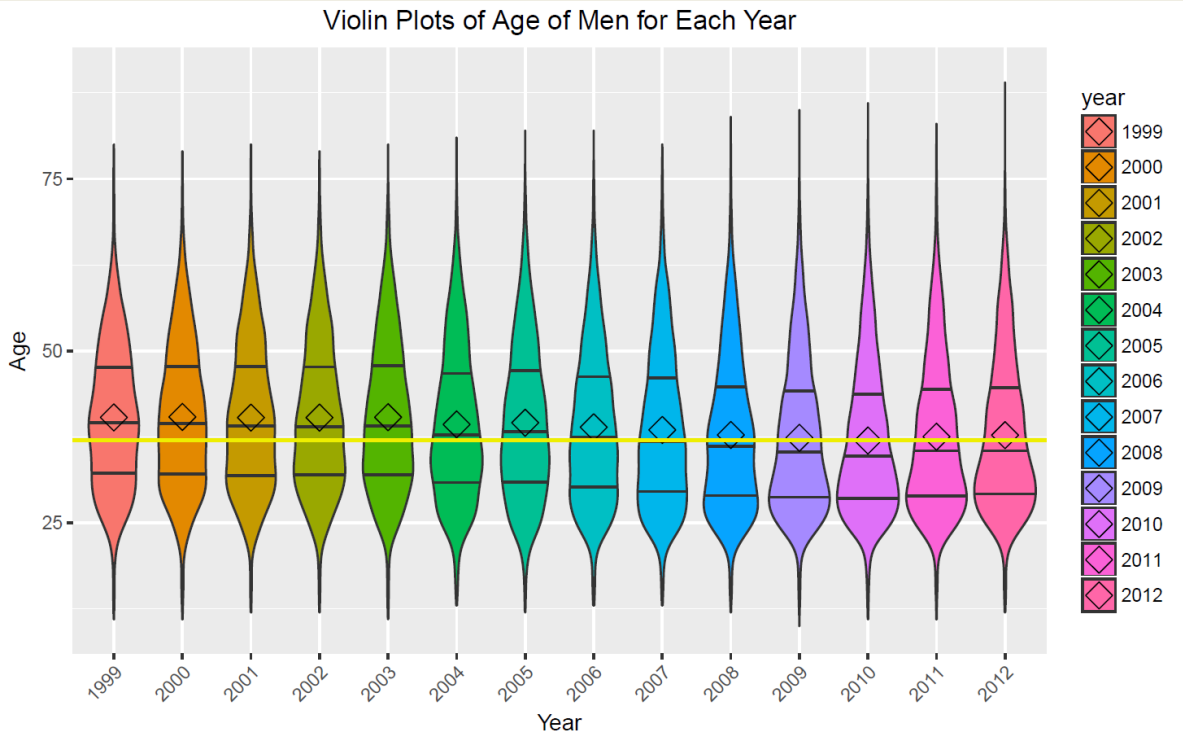


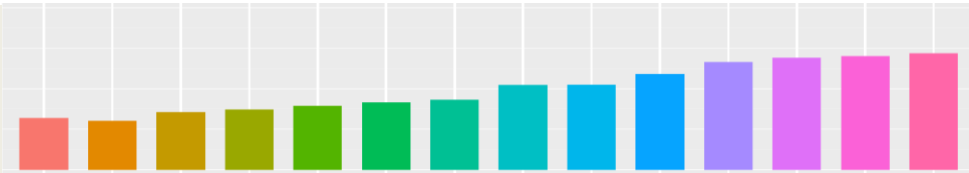
Men	1999	All years	2012
Mean	40.34	38.61	37.75
Median	40	37	35

Women	1999	All years	2012
Mean	34.9	33.85	33.88
Median	33	32	32

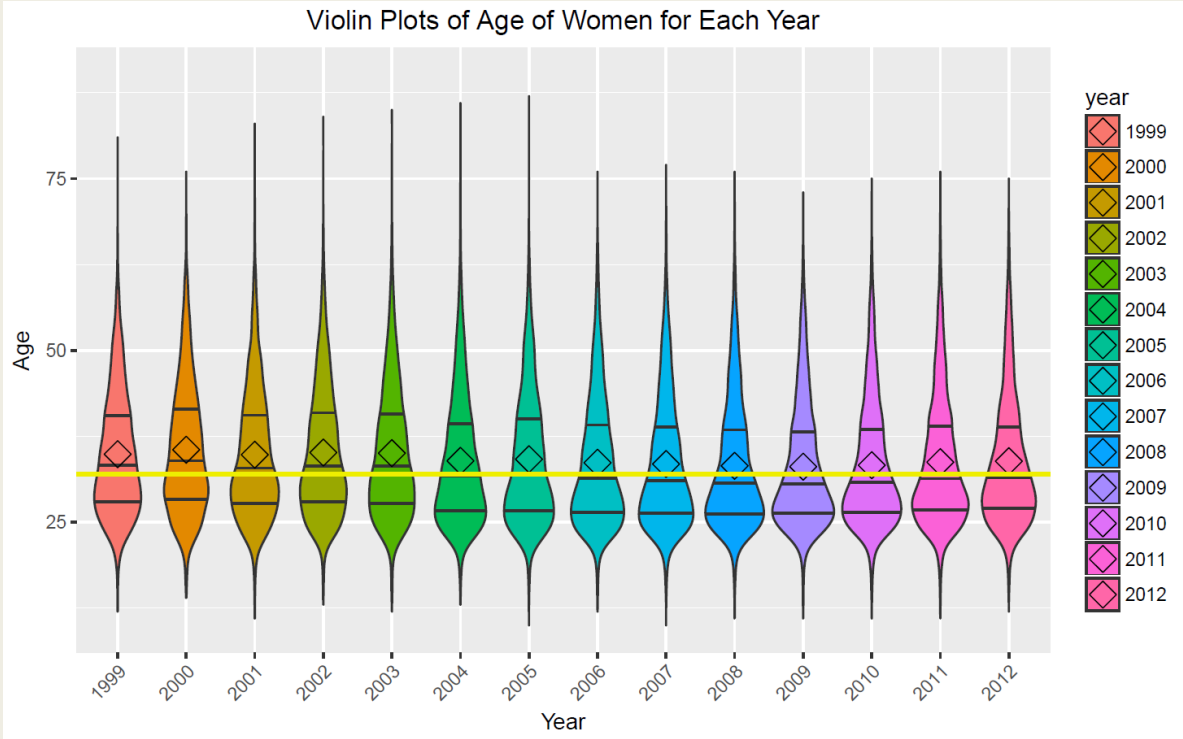
Violin Plots for Age by Year for Men and Women

- Gradual decrease in mean and median age, less pronounced for women
- Starts as bimodal distributions for men to right skewed in age
- More and more younger women





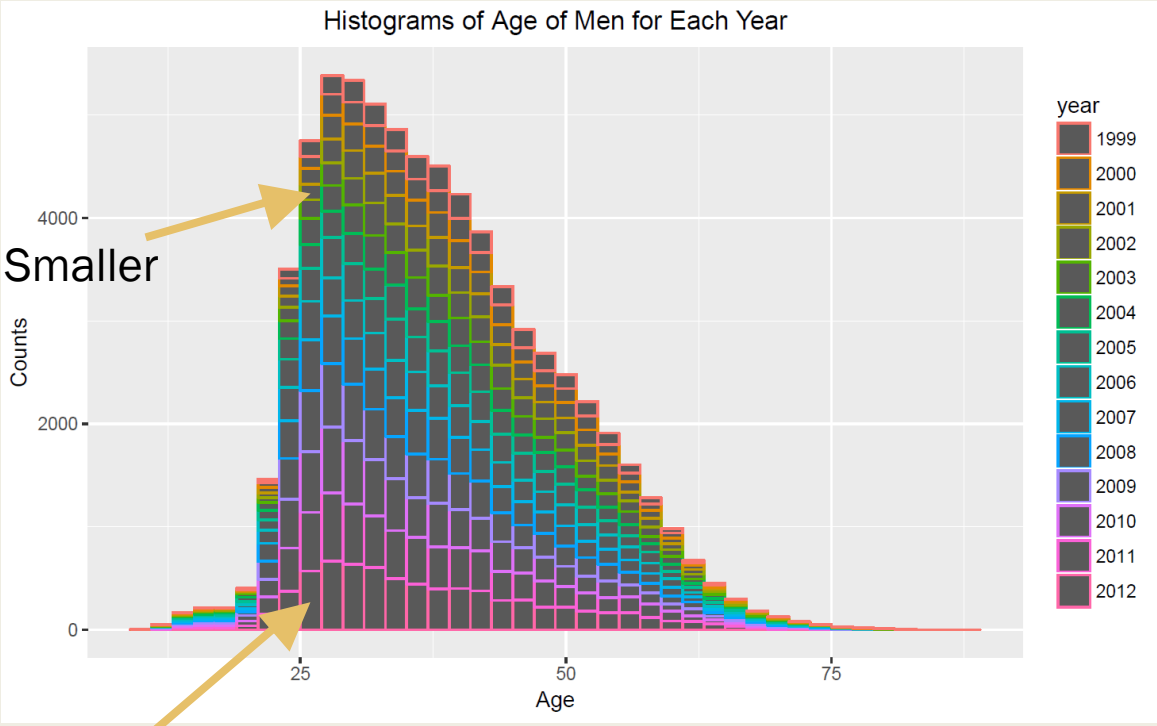
Total Entries of 3189 to 7191

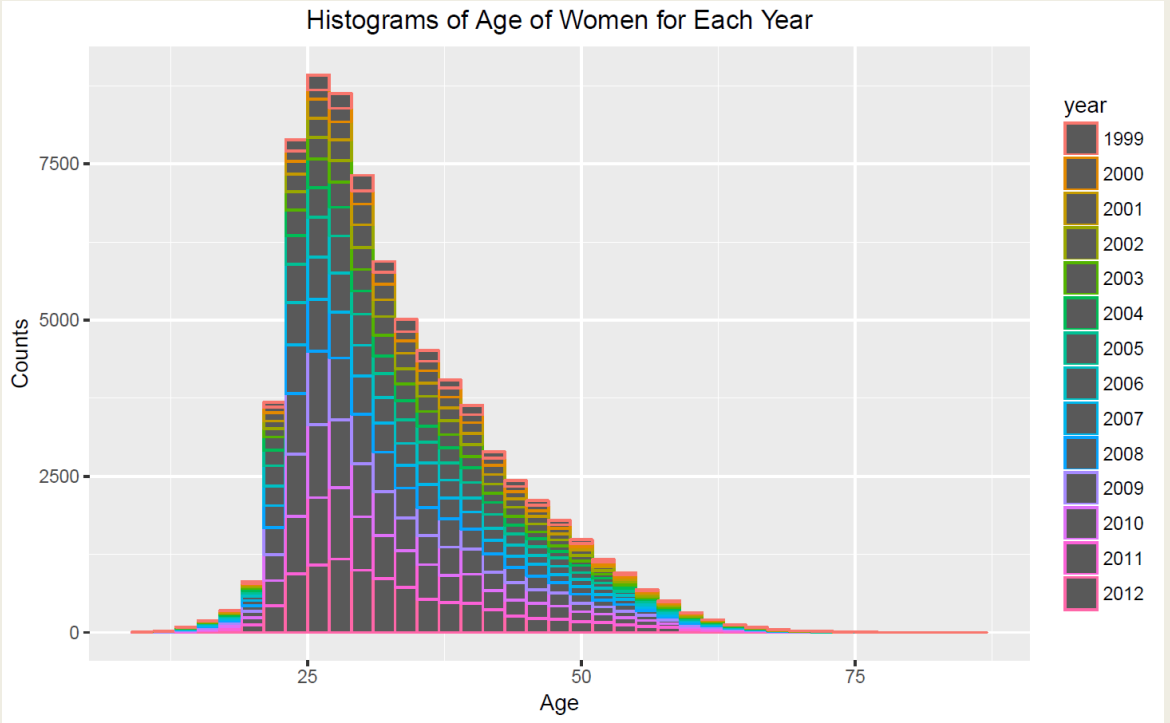


Total Entries of 2352 to 9730

Histograms of Age for Men & Women

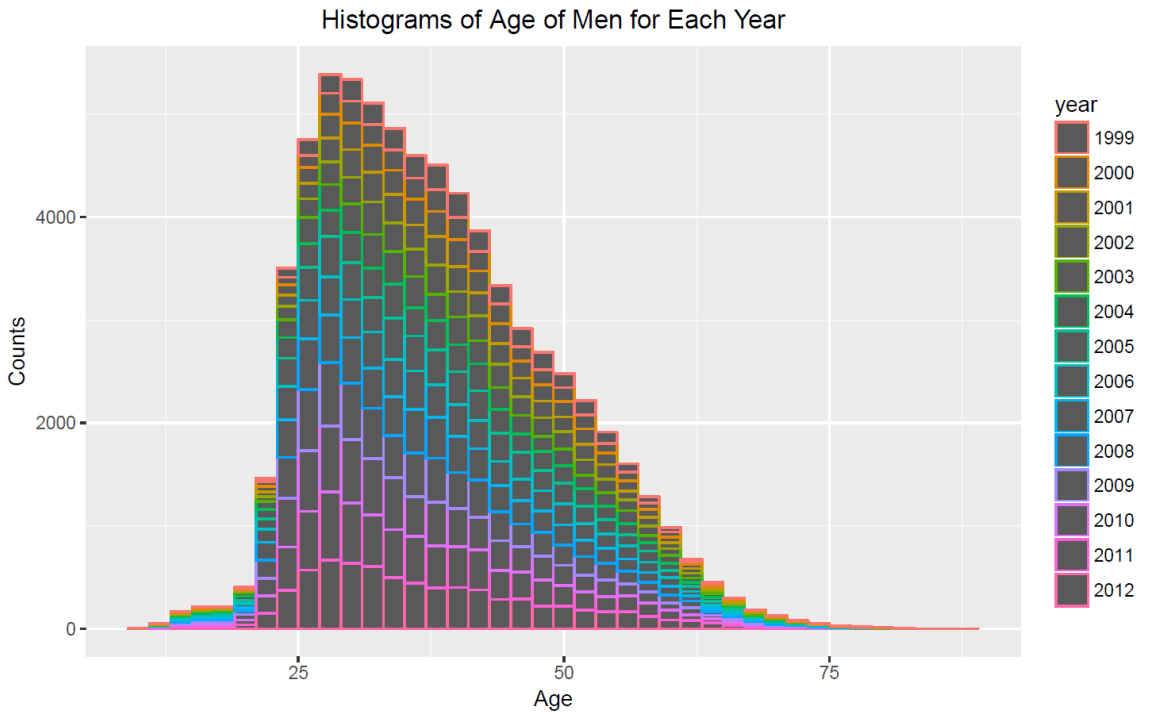
- The histogram of total ages is right skewed - longer tail as age increases
- If you look closely you can see the boxes for the younger runners get much bigger with year while the older runner's boxes get just a little bigger.

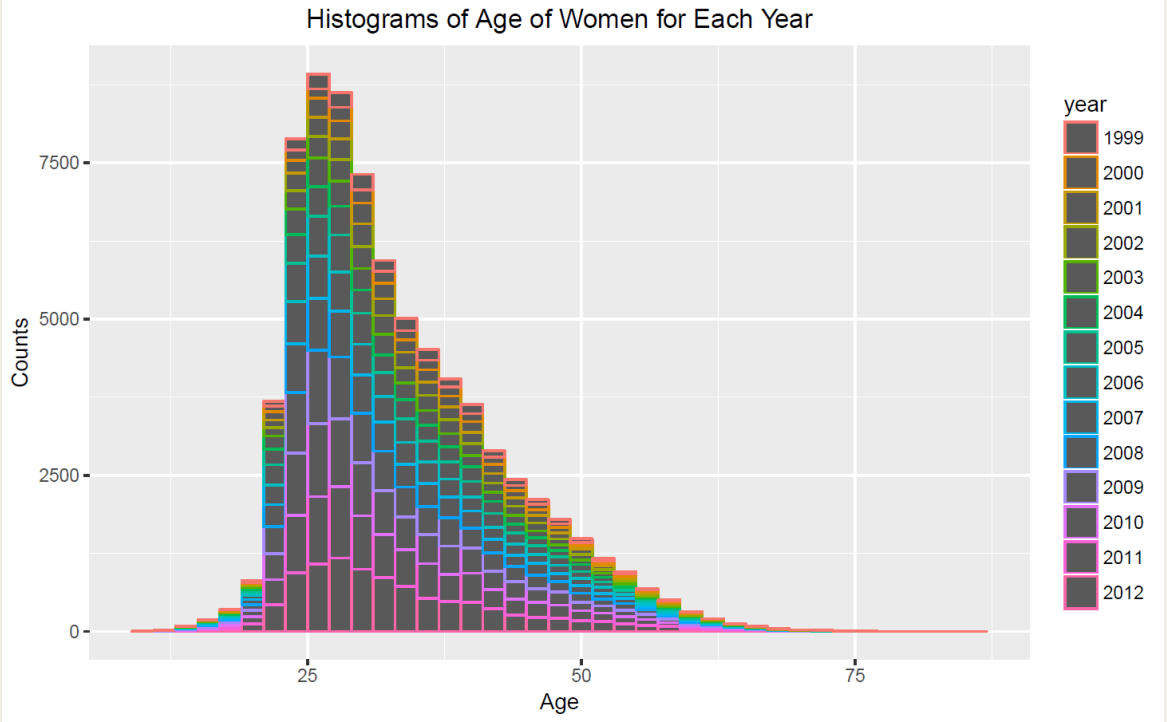




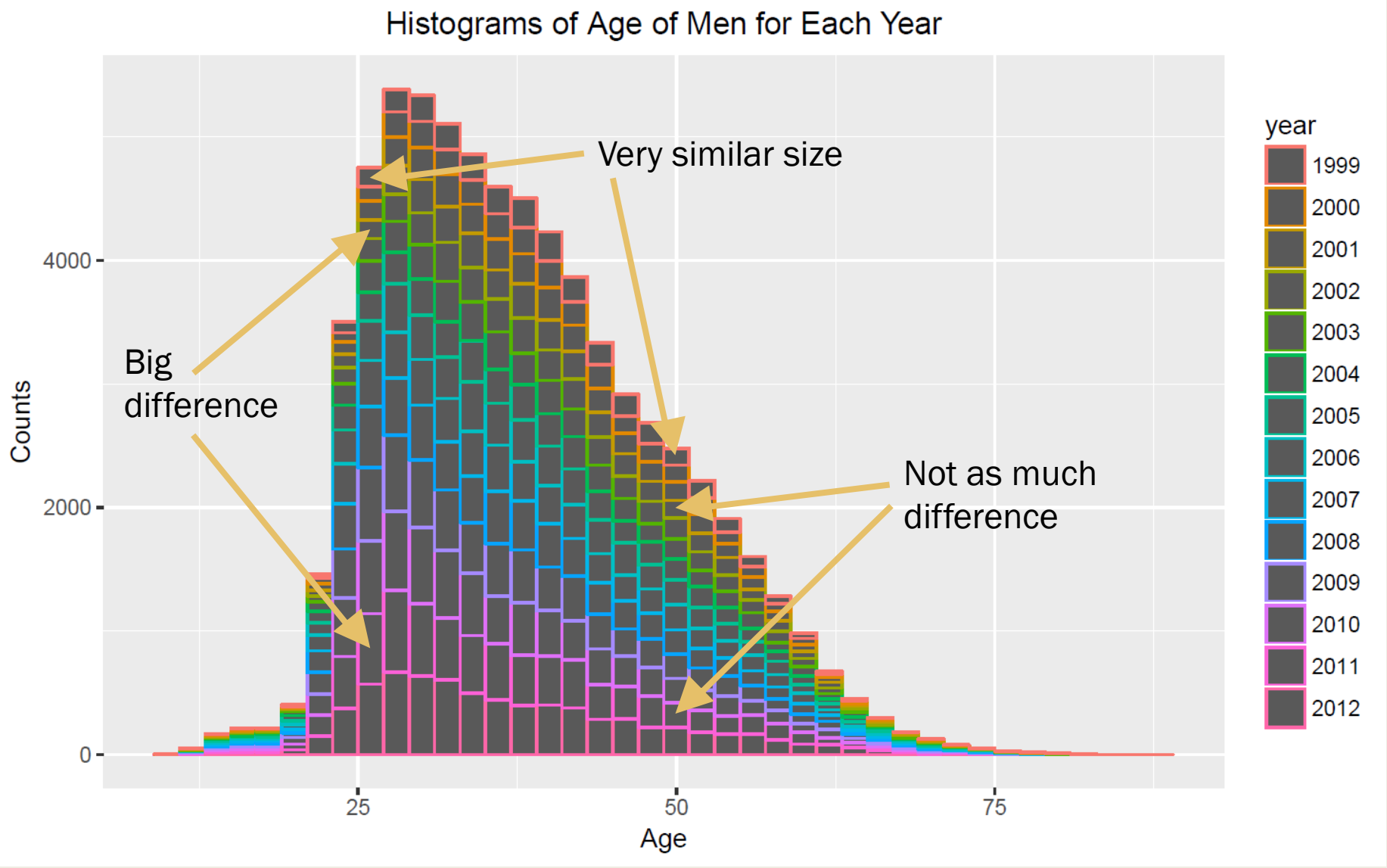
Histograms of Age for Men & Women

- The histogram of total ages is right skewed - longer tail as age increases
- If you look closely you can see the boxes for the younger runners get much bigger with year while the older runner's boxes get just a little bigger.

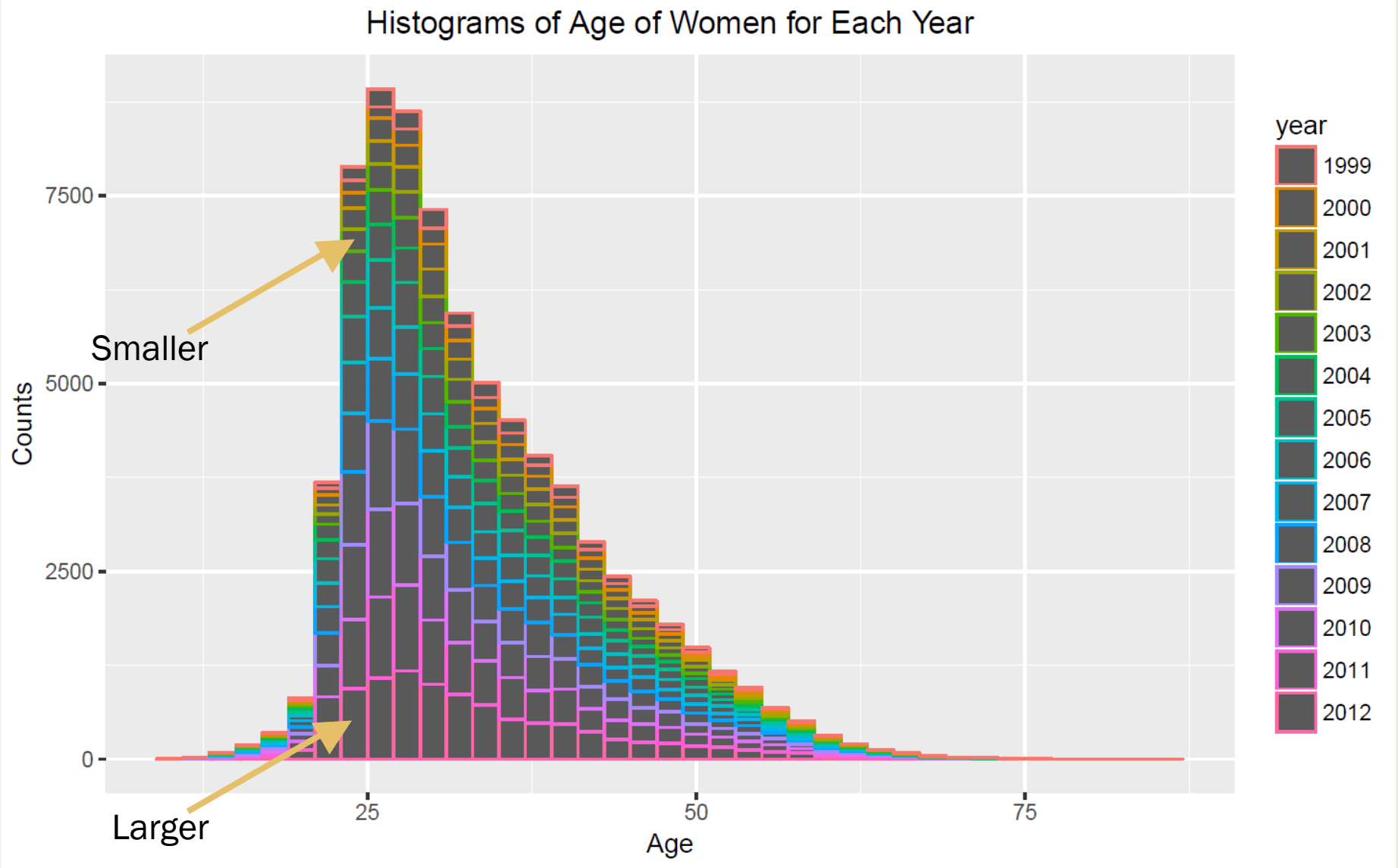




Histogram of Age Count for Men



Histogram of Age Count for Women



Summary

- The total number of entrants in the race more than doubles for the men and more than quadruples for the women.
- Both of the distributions in men's and women's ages shift to younger ages in more recent years.
- The bimodal distribution for men becomes more single modal and right skewed.
- These trends are the result of a gradual increase in entrants for the younger ages being larger than the gradual increase in the older ages.

"Data Science in R, A Case Studies Approach to Computational Reasoning and Problem Solving" by Deborah Nolan and Duncan Temple Lang, CRC Press