Data Exploration

2023-11-30

3DS : Cherry Blossom

Lyuda Bekwinknoll, Meghana Cyanam, Theresa Marie Duenas, Kevin Kiser

With our data visualization we are determining the association between age and fitness based on running data from the Cherry Blossom Ten-mile Run held in Washington DC from 1973 to 2020.

Loading and Cleaning Data

Describing our Data

Variable Names	Data Type	Variable Descriptions
Year	Integer	Year the race was held.
Name	Character	First and last name of runner.
Age	Integer	Age of runner at time of race. He
Time	Time/Numeric	Time in hr:min:sec format to run 10 miles.
Division	Character	Groupings based on age and gender.
pos_by_sex	Integer	
total_by_sex	Integer	
Sex	Character	Gender of runner.
PRCP	Numeric	
TMAX	Integer	Temperature maximum for the race day
TMIN	Integer	Temperature minimum for the race day

Dataset Overview:

In the original data set we have 347402 rows and 17 columns. After cleaning the data set we ended up with 339934 rows and 11 columns. 7468 rows of data were omitted from the data we used because they had missing values for the time and/or age variables. Below is the description of the variables and data we excluded for our data analysis/visualization:

What was excluded	Reason for exclusion
Hometown	
Distance	
Date	
pos_by_div	
total_by_division	
Pace	

What was excluded	Reason for exclusion
Year 1977?	

Summary Statistics:

Year, Age, Time, Sex main variables to focus on.

Checklist for this section:

summary stats: mean, median, mode, range, sd, percentiles, distributions by sex variable, etc.

mention how many women and how many men in each year and overall

```
summary.data.frame(df)
```

```
##       Year          Name                Age             Time
##  Min.   :1974   Length:339214      Min.   : 8.0   Min.   :00:43:20
##  1st Qu.:2001   Class :character   1st Qu.:29.0   1st Qu.:01:19:35
##  Median :2009   Mode  :character   Median :35.0   Median :01:30:50
##  Mean   :2006                      Mean   :36.6   Mean   :01:31:25
##  3rd Qu.:2015                      3rd Qu.:43.0   3rd Qu.:01:42:22
##  Max.   :2019                      Max.   :87.0   Max.   :02:20:00
##
##    Division           pos_by_sex     total_by_sex       Sex
##  Length:339214      Min.   :    1   Min.   :   27   Length:339214
##  Class :character   1st Qu.: 1109   1st Qu.: 3513   Class :character
##  Mode  :character   Median : 2445   Median : 6792   Mode  :character
##                     Mean   : 3134   Mean   : 6298
##                     3rd Qu.: 4739   3rd Qu.: 9030
##                     Max.   :11042   Max.   :11042
##                     NA's   :6       NA's   :6
##       PRCP             TMAX           TMIN
##  Min.   :0.0000   Min.   :44.0   Min.   :32.00
##  1st Qu.:0.0000   1st Qu.:56.0   1st Qu.:39.00
##  Median :0.0000   Median :64.0   Median :43.00
##  Mean   :0.0538   Mean   :63.3   Mean   :43.11
##  3rd Qu.:0.0500   3rd Qu.:70.0   3rd Qu.:47.00
##  Max.   :0.9300   Max.   :84.0   Max.   :58.00
##
```

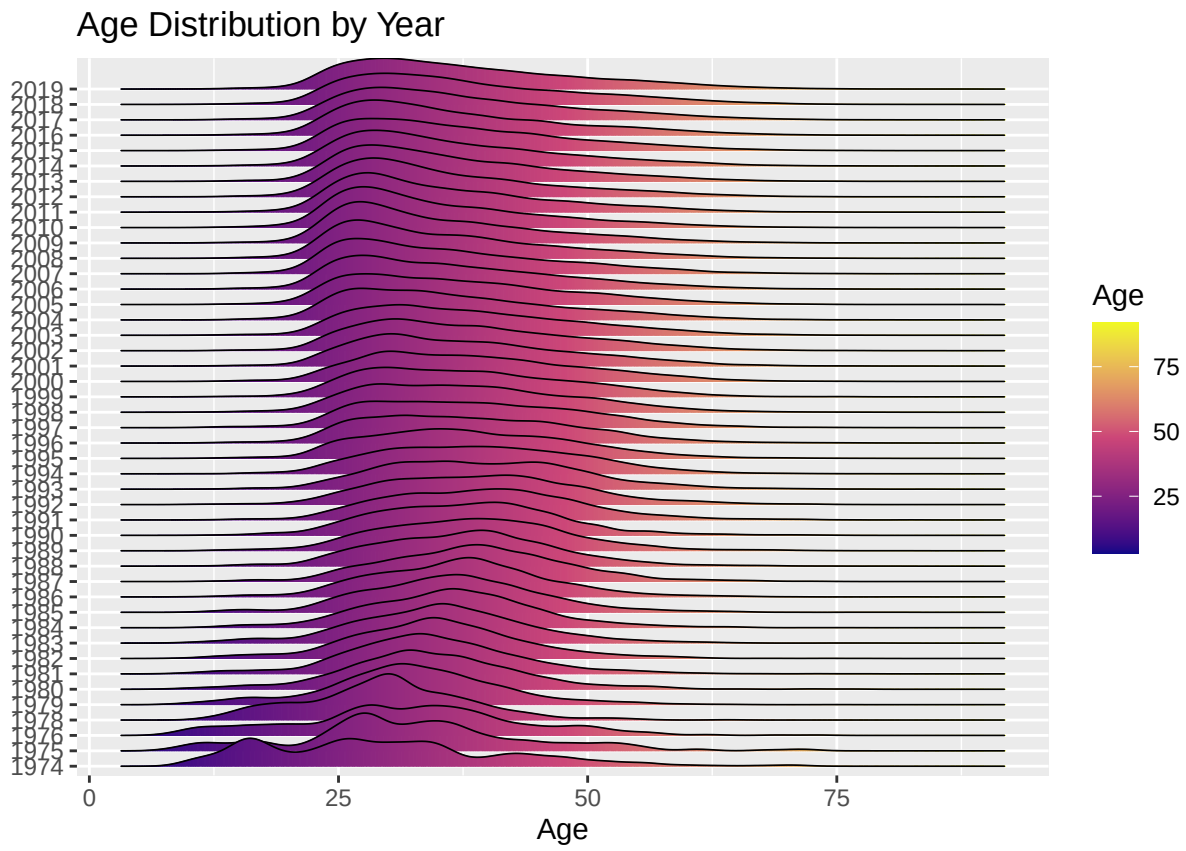
These were helping me evaluate the data cleaning, we can fix or replace them later

```
plot_age_dist <- ggplot(df, aes(x = Age, y = as.factor(Year))) +
   geom_density_ridges_gradient(
     aes(fill = ..x..), scale = 3, size = 0.3
   ) +
   scale_fill_gradientn(
     colours = c("#0D0887FF", "#CC4678FF", "#F0F921FF"),
     name = "Age"
   ) +
   labs(title = 'Age Distribution by Year', y="")

plot_age_dist
```

```
## Warning: The dot-dot notation (‘..x..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(x)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

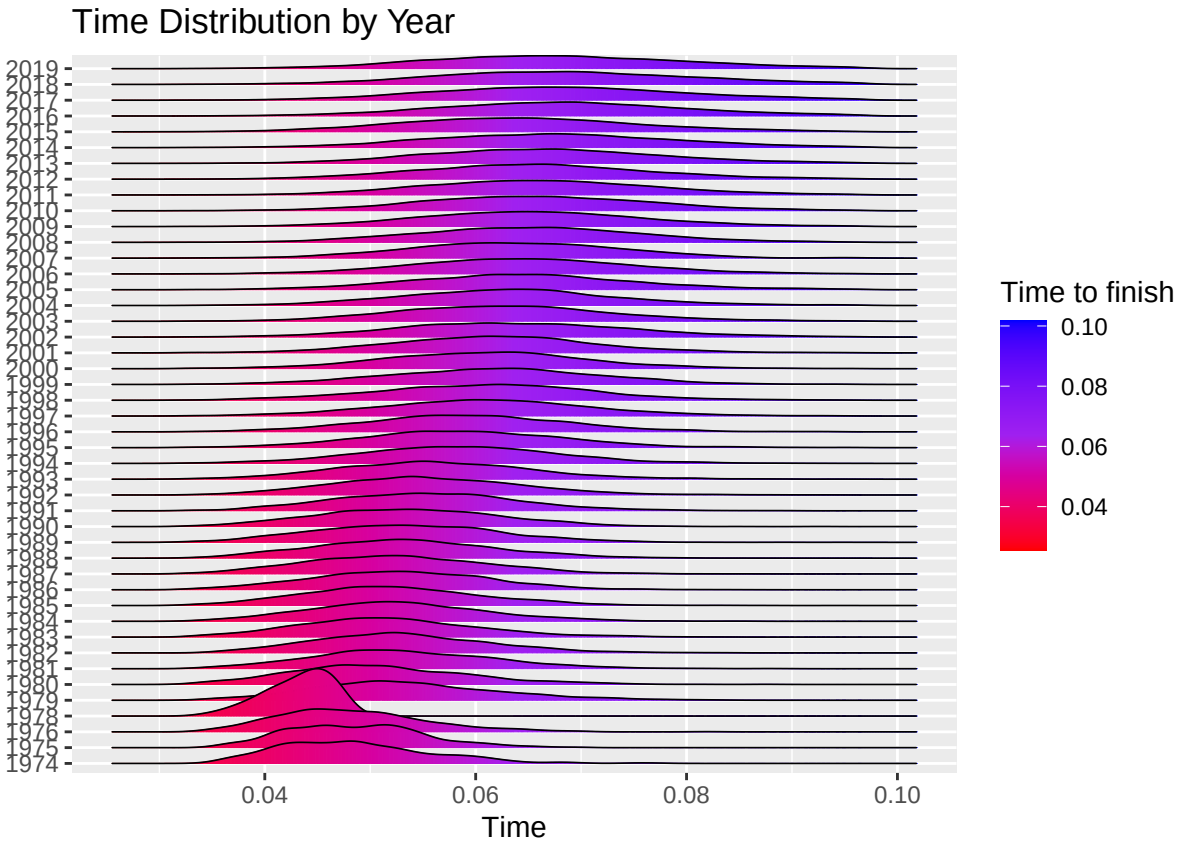
```
## Picking joint bandwidth of 1.6
```




```
 # Plotting density ridgeline plot for Time by Year
plot_time_dist <-  ggplot(df, aes(x = Time, y = as.factor(Year))) +
   geom_density_ridges_gradient(
     aes(fill = ..x..), scale = 3, size = 0.3
   ) +
   scale_fill_gradientn(
     colours = c("red", "purple", "blue"),
     name = "Time to finish"
   ) +
   labs(title = 'Time Distribution by Year', y = "")
plot_time_dist
```

```
## Don’t know how to automatically pick scale for object of type <times>.
## Defaulting to continuous.
```

```
## Picking joint bandwidth of 0.00152
```


Time Distribution by Year
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999
1998
1997
1996
1995
1994
1993
1992
1991
1990
1989
1988
1987
1986
1985
1984
1983
1982
1981
1980
1979
1978
1976
1975
1974
0.04
0.06
0.08
0.10
Time
Time to finish
0.10
0.08
0.06
0.04