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:

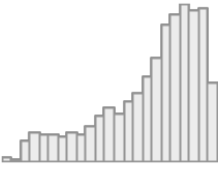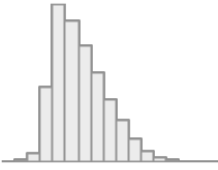summary stats: mean, median, mode, range, sd, percentiles, distributions by sex variable, etc.

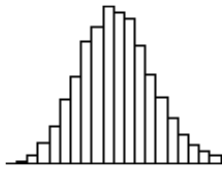mention how many women and how many men in each year and overall
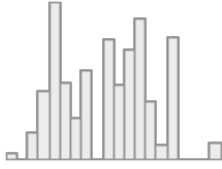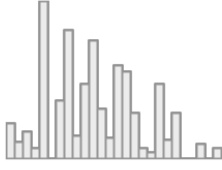
```
df %>% select(Year, Age, Time, Sex, PRCP, TMAX, TMIN) %>%
dfSummary(
        varnumbers = FALSE,
        plain.ascii  = FALSE,
        style        = "grid",
        graph.magnif = 0.75,
        valid.col    = FALSE,
        tmp.img.dir  = "/tmp")
```

Data Frame Summary

df   Dimensions: 339214 x 7
Duplicates: 11759

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Year [integer]	Mean (sd) : 2006.4 (10.5) min < med < max: 1974 < 2009 < 2019 IQR (CV) : 14 (0)	45 distinct values		0 (0.0%)
Age [integer]	Mean (sd) : 36.6 (10.3) min < med < max: 8 < 35 < 87 IQR (CV) : 14 (0.3)	80 distinct values		0 (0.0%)

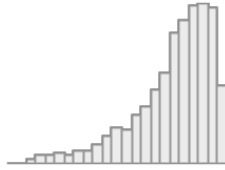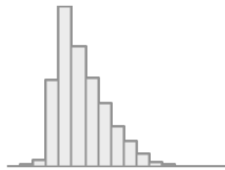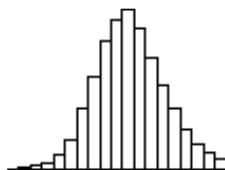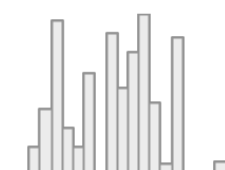| Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Time [times]	Mean (sd) : 00:00:00 (0) min < med < max: 00:00:00 < 00:00:00 < 00:00:00 IQR (CV) : 0 (00:00:00)	5649 distinct values		0 (0.0%)
Sex [character]	1. F 2. M	166755 (49.2%) 172459 (50.8%)		0 (0.0%)
PRCP [numeric]	Mean (sd) : 0.1 (0.1) min < med < max: 0 < 0 < 0.9 IQR (CV) : 0 (2.4)	17 distinct values		0 (0.0%)
TMAX [integer]	Mean (sd) : 63.3 (8.3) min < med < max: 44 < 64 < 84 IQR (CV) : 14 (0.1)	24 distinct values		0 (0.0%)
TMIN [integer]	Mean (sd) : 43.1 (5.6) min < med < max: 32 < 43 < 58 IQR (CV) : 8 (0.1)	23 distinct values		0 (0.0%)

```r
df %>% select(Year, Age, Time, Sex, PRCP, TMAX, TMIN) %>% group_by(Sex) %>%
dfSummary(
        varnumbers = FALSE,
        plain.ascii  = FALSE,
        style        = "grid",
        graph.magnif = 0.75,
        valid.col    = FALSE,
        tmp.img.dir  = "/tmp")
```
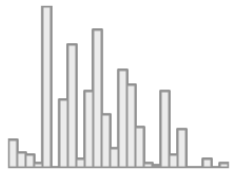
Data Frame Summary
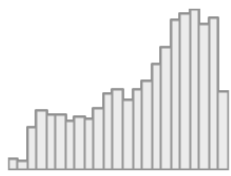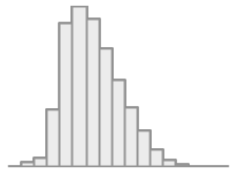
df   Group: Sex = F
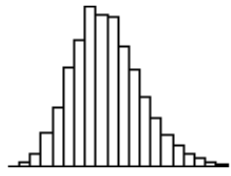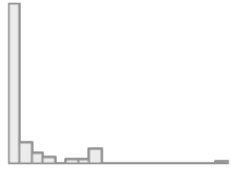**Dimensions:** 166755 x 7
Duplicates: 7844

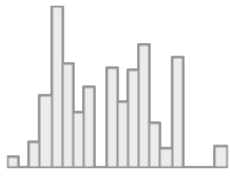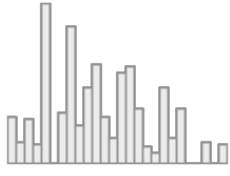| Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Year [integer]	Mean (sd) : 2009.2 (8.3) min < med < max: 1974 < 2011 < 2019 IQR (CV) : 10 (0)	45 distinct values		0 (0.0%)
Age [integer]	Mean (sd) : 34.6 (9.5) min < med < max: 8 < 32 < 87 IQR (CV) : 13 (0.3)	80 distinct values		0 (0.0%)
Time [times]	Mean (sd) : 00:00:00 (0) min < med < max: 00:00:00 < 00:00:00 < 00:00:00 IQR (CV) : 0 (00:00:00)	5175 distinct values		0 (0.0%)
PRCP [numeric]	Mean (sd) : 0 (0.1) min < med < max: 0 < 0 < 0.9 IQR (CV) : 0 (2.6)	17 distinct values		0 (0.0%)
TMAX [integer]	Mean (sd) : 64 (8) min < med < max: 44 < 66 < 84 IQR (CV) : 14 (0.1)	24 distinct values		0 (0.0%)

4

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
TMIN [integer]	Mean (sd) : 43 (5.2) min < med < max: 32 < 43 < 58 IQR (CV) : 8 (0.1)	23 distinct values		0 (0.0%)

Group: Sex = M
Dimensions: 172459 x 7
Duplicates: 3915

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Year [integer]	Mean (sd) : 2003.7 (11.6) min < med < max: 1974 < 2007 < 2019 IQR (CV) : 18 (0)	45 distinct values		0 (0.0%)
Age [integer]	Mean (sd) : 38.5 (10.7) min < med < max: 8 < 37 < 87 IQR (CV) : 15 (0.3)	80 distinct values		0 (0.0%)
Time [times]	Mean (sd) : 00:00:00 (0) min < med < max: 00:00:00 < 00:00:00 < 00:00:00 IQR (CV) : 0 (00:00:00)	5556 distinct values		0 (0.0%)
PRCP [numeric]	Mean (sd) : 0.1 (0.1) min < med < max: 0 < 0 < 0.9 IQR (CV) : 0.1 (2.2)	17 distinct values		0 (0.0%)

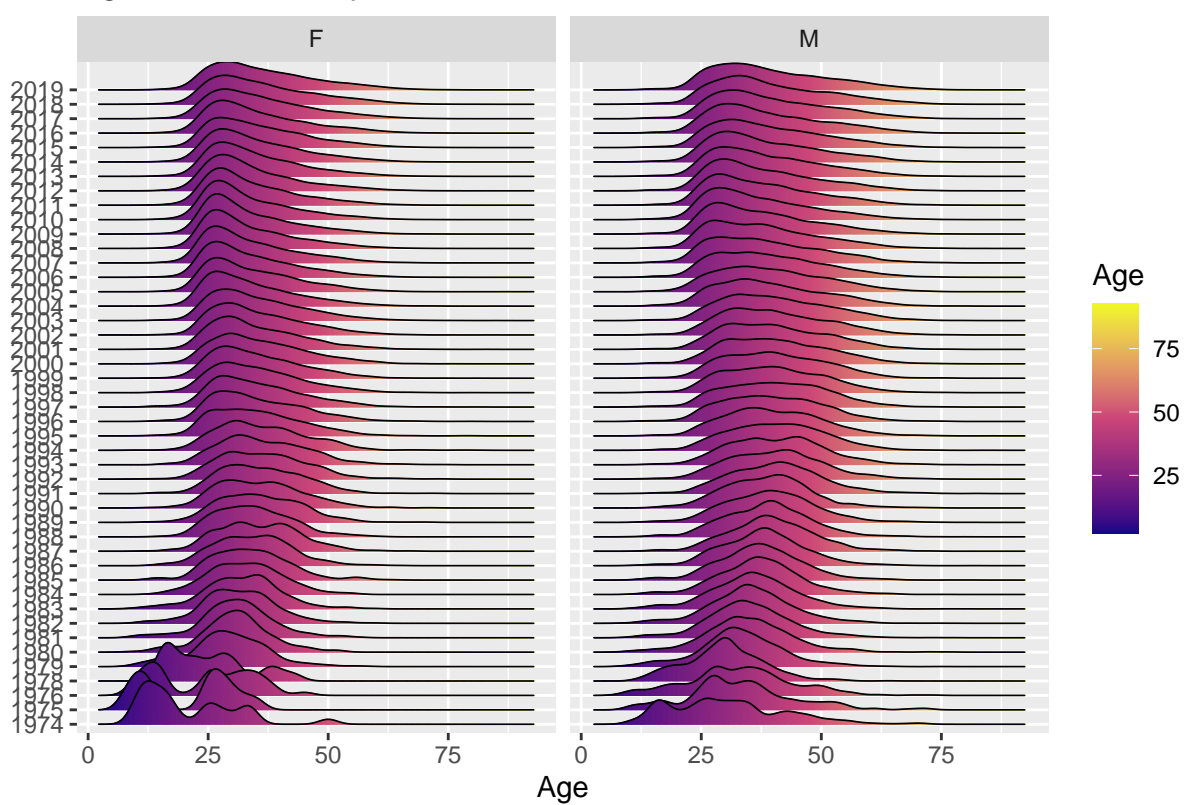| Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
TMAX [integer]	Mean (sd) : 62.7 (8.5) min < med < max: 44 < 63 < 84 IQR (CV) : 14 (0.1)	24 distinct values		0 (0.0%)
TMIN [integer]	Mean (sd) : 43.2 (6) min < med < max: 32 < 43 < 58 IQR (CV) : 8 (0.1)	23 distinct values		0 (0.0%)

```
#view(dfSummary(df))
```

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

```
## Picking joint bandwidth of 1.93
```

```
## Picking joint bandwidth of 1.78
```

Age Distribution by Year


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

## Picking joint bandwidth of 0.00185

## Picking joint bandwidth of 0.00159
```

Time Distribution by Year