High School Girls Cross Country Trends

Kristina Stern

5/15/2022

What

My project is about trends in Minnesota girls high school cross country. I want to show how, on average, the race times have gotten faster every year and that the talent pool has deepened. I also want to prove that placing well when you are younger does not guarantee success as you age.

Why

I've followed cross country and track for almost 20 years now; it is my 'fantasy football' basically. I've noticed that, over time, the performances have been consistently getting better and I wanted to prove it statistically. I've also had several heated internet forum discussions regarding girls who are deemed 'prodigies' at a young age and I am looking forward to providing evidence as to why doing well in 9th grade doesn't always translate to doing well in 12th grade.

How

In addition to trend fitting, I am going to be making use of subsets to be able to really look at how the data changes within particular groups. I will also be using a data filter to find runners who have appeared more than once in the data set.

Body

Why Cross Country

Cross country is running long distances while not on a track. Some races are done on various trails while others are done on golf courses. Minnesota hold their cross country championships on a golf course, which has remained at the same place since 1992. Boys cross country state meets started in 1943 while girls meets started in 1975. Approaches to Coaching philosophies, gear, nutrition and exposure to cross country on a bigger stage have all shifted over the years, which also contributes to the changing of running averages.

I chose cross country over track because there is more data and I chose girls cross country because girls are often overlooked when people talk about MSHSL history. Women's sports are often marginalized and not taken as seriously as men's sports, with Title IX still undergoing attacks 50 years later. I wanted to demonstrate the huge strides women (and girls) have taken over the years in the sport.

The Data

I used the data found on Raceberry Jam. This site has complete data from 1991 in a semi-usable format. There is older data out there but it isn't formatted as well and is missing various data points, such as times. I copied and pasted it into Notepad++ and used regular expressions to re-format it into something more usable. I replaced all double white space with a comma (to preserve names in one column), replaced any double commas with a single comma (sometimes more than once) and then saved it to a CSV file. The original data looked something like this -

```
1        Megan Hasz, 11               5:15.2    13:40.9    Alexandria
2        Bethany Hasz, 11             5:15.4    13:44.4    Alexandria
3        Tess Misgen, 10              5:23.3    14:13.5    Shakopee
4     1  Emma Benner, 11              5:25.8    14:13.6    Forest Lake
5        Emily Covert, 8              5:24.8    14:15.0    Minneapolis Washburn
6     2  Rachel King, 12              5:26.5    14:15.1    St. Michael-Albertvil
7     3  Anna French, 12              5:17.6    14:15.9    Wayzata
8     4  Annika Lerdall, 10           5:39.7    14:18.4    Wayzata
9     5  Emily Betz, 12               5:27.0    14:19.0    East Ridge
```

The first column is the place, the second column is the team place, followed by name, grade, average time, overall time and school. I wasn't interested in the team place or the average time so I needed to delete those. I created a macro to delete all the cells where there was a team place and shift the remaining cells in that row over. That made it easier to just delete the average time column once everything was aligned. For all the times prior to 2015 I had to convert them to their equivalent 5k times using this equation:

$$t2 = t1 * (\frac{d2}{d1})^{1.06}$$

There was some trouble with this since the formatting wasn't always consistent. Out of the 4617 lines of data, I had to change about 20 by hand. I originally added an index to the data but found I didn't need it so I removed it from the final dataset.

Using the data in R

To use the data in R, I used the readr library and then imported the data into an R table using instructions from Statology.

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Rows: 4617 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Name, School
## dbl (4): Place, Grade, Time, Year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

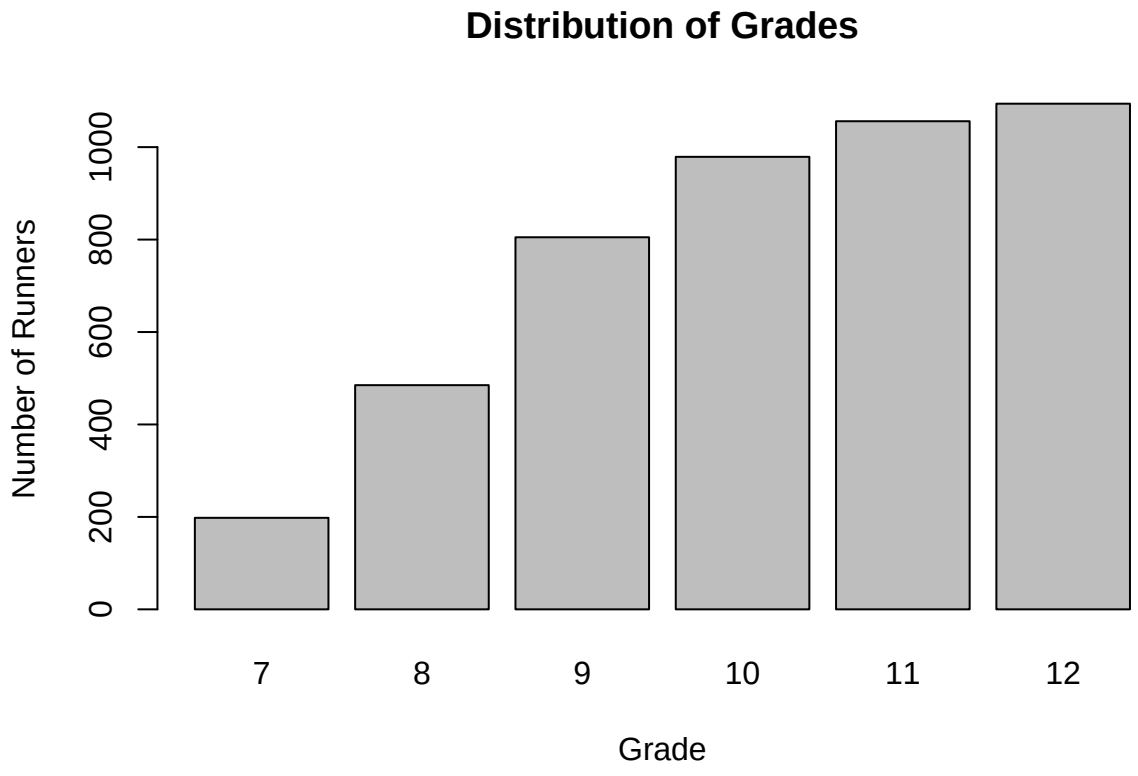
First, a check that the data imported correctly.

```
## # A tibble: 6 x 6
##   Place Name             Grade  Time School  Year
##   <dbl> <chr>            <dbl> <dbl> <chr>  <dbl>
## 1     1 Carrie Tollefson     9  17.3 Dblv    1991
## 2     2 Tina Forthmiller    10  17.9 StF     1991
## 3     3 Kara Wheeler         8  17.9 DuE     1991
## 4     4 Amy Hill             8  17.9 DuE     1991
## 5     5 Keri Zweig          11  18.1 Mtk     1991
## 6     6 Turena Johnson      11  18.1 Brn     1991
```

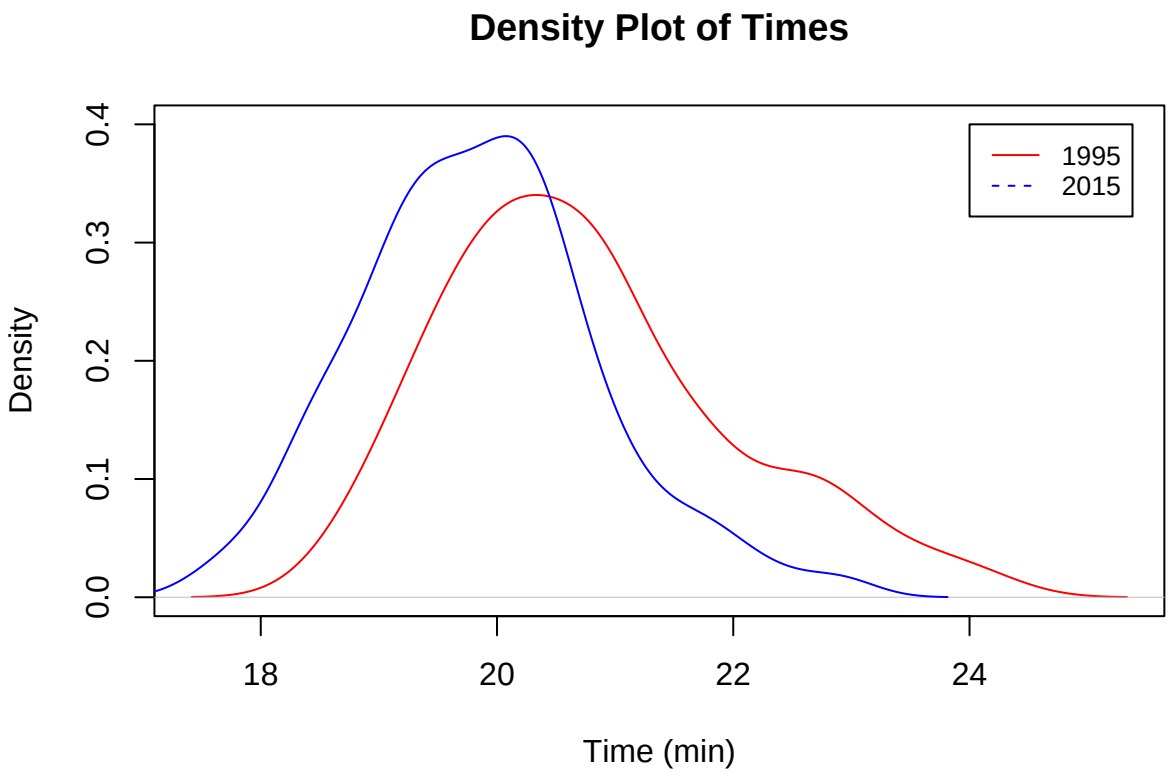
How many runners from each grade there were overall:

```
##
##    7    8    9   10   11   12
##  198  485  805  979 1056 1094
```

A plot of the above table:



I wanted to look at how times had progressed from different decades. I chose 1995 and 2015 for this example. Below is the density plots of the times of those two years.



The average time and standard deviation from 1995:

```
## [1] 20.75802
```

```
## [1] 1.221485
```

The average time and standard deviation from 2015:

```
## [1] 19.84972
```

```
## [1] 1.015177
```

This shows that not only were times faster in 2015 but the spread was much tighter.

Let's look at the data for a few specific places. I wanted to include 150th place in my plots but the race didn't have as many participants the first decade.

First place:

```
## # A tibble: 6 x 6
##   Place Name             Grade  Time School                Year
##   <dbl> <chr>            <dbl> <dbl> <chr>                <dbl>
## 1     1 Carrie Tollefson     9  17.3 Dblv                  1991
## 2     1 Lisa Aro            11  18.7 Buffalo               1992
## 3     1 Kara Wheeler        10  18.7 Duluth East           1993
```

```
## 4     1 Carrie Tollefson    12  17.8 Lac Qui Parle Val/D B  1994
## 5     1 Elaine Eggleston    11  18.7 Roseville Area         1995
## 6     1 Josie Johnson        8  18.6 Rochester John Marshl  1996
```

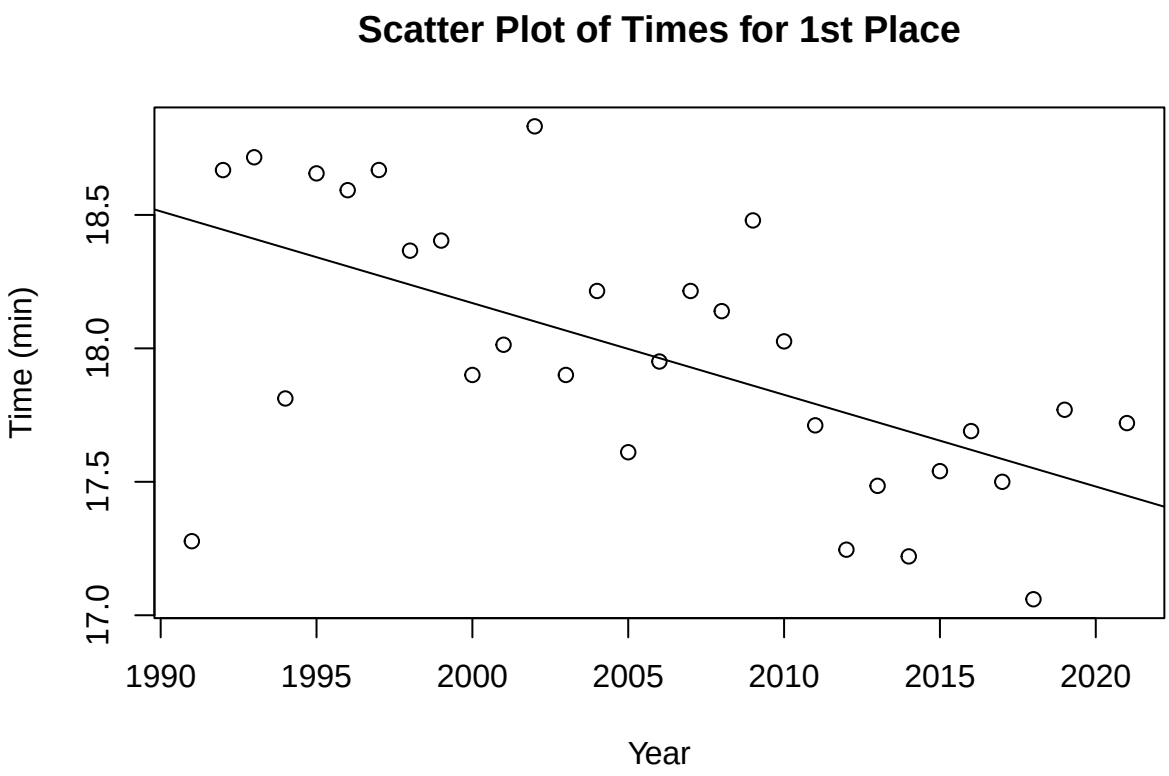
Tenth place:

```
## # A tibble: 6 x 6
##   Place Name            Grade  Time School                Year
##   <dbl> <chr>           <dbl> <dbl> <chr>                <dbl>
## 1    10 Jennifer Watson    12  18.4 Ank                   1991
## 2    10 Karen Walczak      12  19.7 Park Center           1992
## 3    10 Amy Maciasek        9  19.7 Mounds View           1993
## 4    10 Beth Rautmann       8  18.8 White Bear Lake Area  1994
## 5    10 Kelly Brinkman      9  19.3 Hutchinson            1995
## 6    10 Serena Sullivan    10  19.3 Hibbing               1996
```

One Hundredth place:

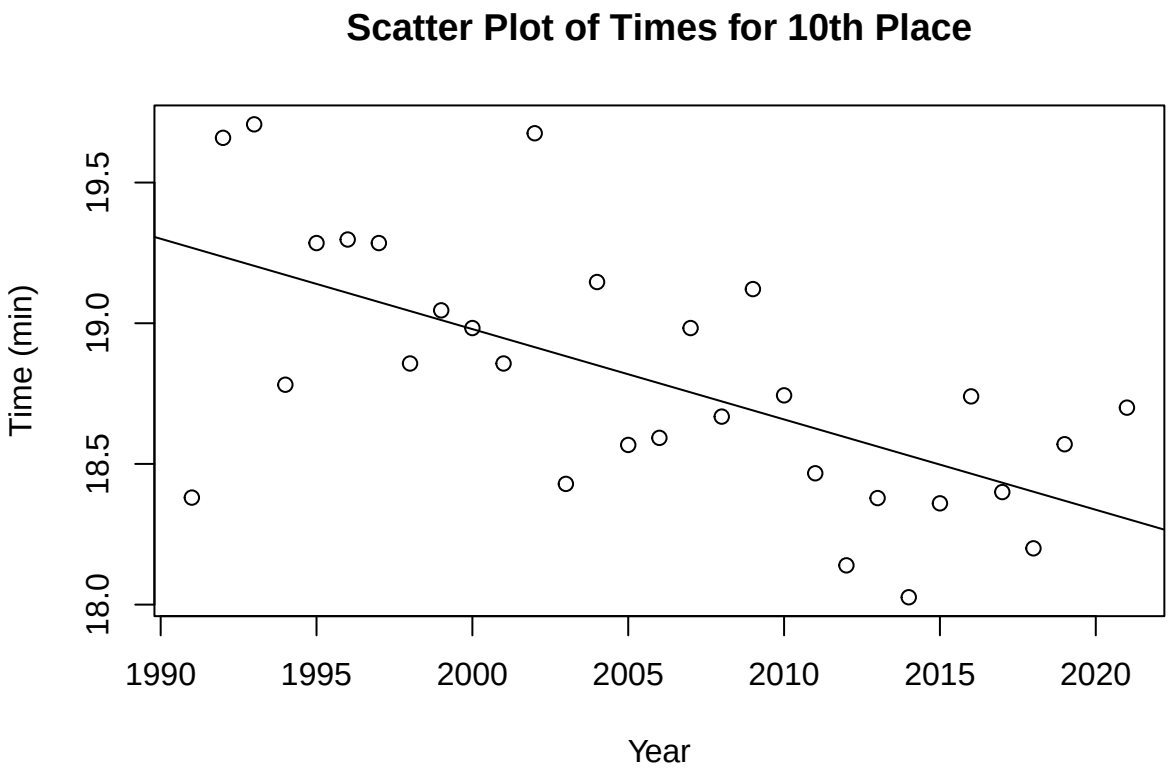
```
## # A tibble: 6 x 6
##   Place Name               Grade  Time School                Year
##   <dbl> <chr>              <dbl> <dbl> <chr>                <dbl>
## 1   100 Jordan Cushing        12  20.6 Mso                   1991
## 2   100 Kristal Drazkowski    12  22.0 Winona                1992
## 3   100 Jen Kennington        11  22.0 Duluth Central        1993
## 4   100 Amanga Lang            9  21.3 Rocori                1994
## 5   100 Sarah Fifield         11  21.6 Minneapolis South     1995
## 6   100 Alisha Boyd            8  21.4 White Bear Lake Area  1996
```

Plotting the times with a line fit and coefficients for 1st place:



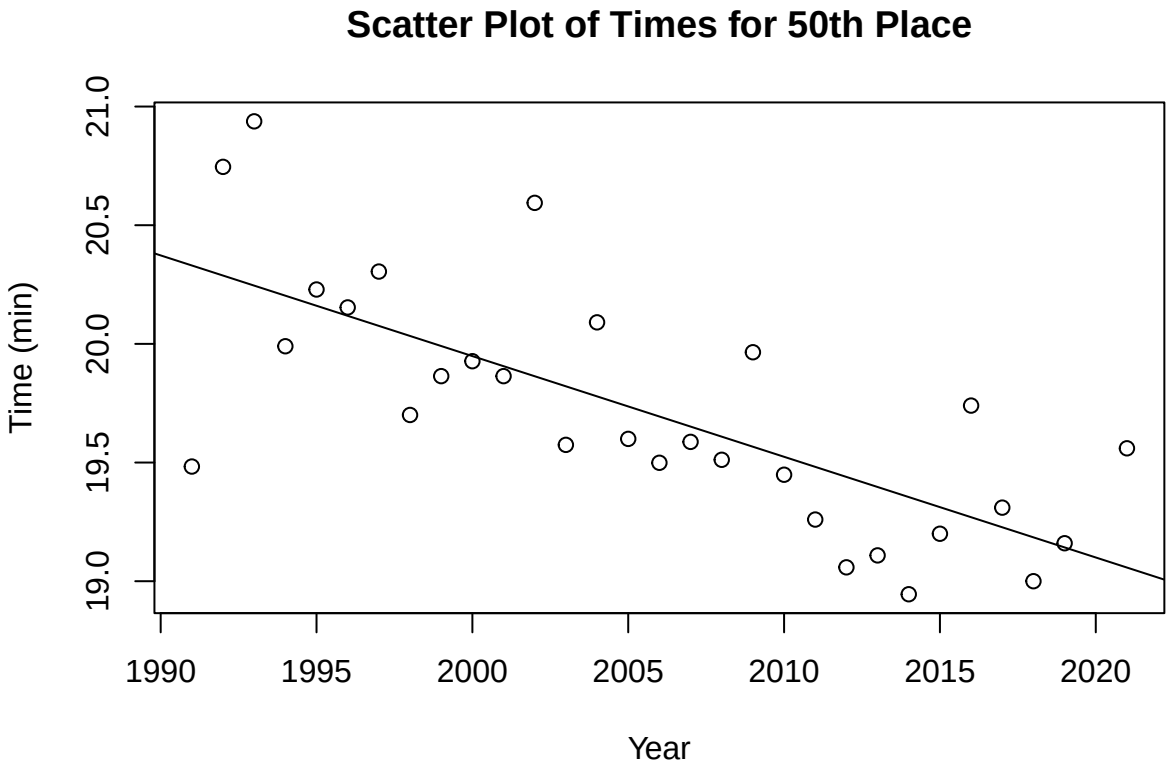

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  86.91625574 17.113507120  5.078810 2.235757e-05
## place_1$Year -0.03437327  0.008533065 -4.028245 3.895416e-04
```

Plotting the times with a line fit and coefficients for 10th place:



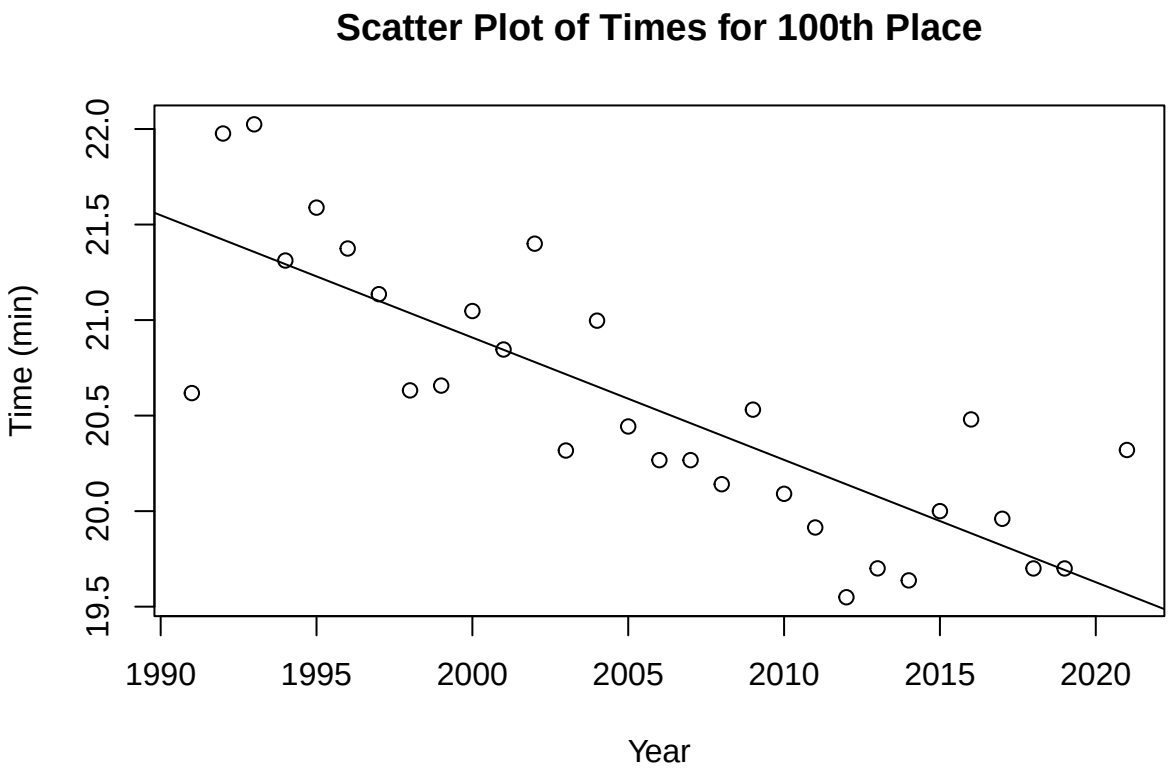

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)   83.19128489 15.046381898  5.528989 6.546892e-06
## place_10$Year -0.03210604  0.007502363 -4.279457 1.978128e-04
```

Plotting the times with a line fit and coefficients for 50th place:




```
##                   Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)   104.79318412 14.93893837  7.014768 1.251463e-07
## place_50$Year  -0.04242236  0.00744879 -5.695201 4.168772e-06
```

Plotting the times with a line fit and coefficients for 100th place:




```
##                    Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)    148.94465149 16.947296286  8.788697 1.535169e-09
## place_100$Year  -0.06401809  0.008450189 -7.575936 2.979356e-08
```

As you can see, there is noticeable difference between the trends from 1st place to 100th place.

```
-0.03437327*60
```

```
## [1] -2.062396
```

```
-0.06401809 *60
```

```
## [1] -3.841085
```

The first place times decreased, on average, of about 2 seconds per year while the 100th place times decreased at a rate of -3.8 seconds a year.

```
-2.062396*30/60
```

```
## [1] -1.031198
```

```
-3.841085*30/60
```

```
## [1] -1.920543
```

The girls in 1st place are about a minute faster and the girls in 100th place are almost 2 minutes faster compared to 30 years ago.

I realize that while a linear correlation works right now, that the actual fit is an exponential decay. At some point, the improvement in times will flatten because it is impossible for them to get to zero. However, due to the nature of the data, it is difficult at this time to find that actual fit line. In 30 more years, when there is more data towards the tail, it would be easier to have a predictive model.

Now I wanted to break up the data by each grade:

```
data_7 <- subset(xc_data, xc_data$Grade == 7)
data_8 <- subset(xc_data, xc_data$Grade == 8)
data_9 <- subset(xc_data, xc_data$Grade == 9)
data_10 <- subset(xc_data, xc_data$Grade == 10)
data_11 <- subset(xc_data, xc_data$Grade == 11)
data_12 <- subset(xc_data, xc_data$Grade == 12)
head(data_7)
```

```
## # A tibble: 6 x 6
##   Place Name             Grade  Time School       Year
##   <dbl> <chr>            <dbl> <dbl> <chr>       <dbl>
## 1    65 Molly Aberg          7  19.8 Wbl          1991
## 2    40 Mandy Dockendorf     7  20.5 Rocori       1992
## 3    95 Tracy Musech         7  21.8 Hibbing      1992
## 4    30 Kelly Brinkman       7  20.3 Hutchinson   1993
## 5    98 Megan Daymont        7  22.0 Northfield   1993
## 6    53 Kendall Wheeler      7  20.0 Duluth East  1994
```

Standard deviation of the times per grade:

```
## [1] 0.8581629
```

```
## [1] 1.028409
```

```
## [1] 1.086925
```

```
## [1] 1.172572
```

```
## [1] 1.135137
```

```
## [1] 1.181283
```

Average of the times per grade:

```
## [1] 20.18436
```

```
## [1] 20.0755
```

```
## [1] 20.1391
```

```
## [1] 20.20532
```

```
## [1] 20.17808
```

```
## [1] 20.20771
```

Averages of the place per grade:

```
## [1] 91.40404
```

```
## [1] 79.62474
```

```
## [1] 77.39255
```

```
## [1] 78.74668
```

```
## [1] 77.41951
```

```
## [1] 76.45338
```

Seeing this has re-oriented my thoughts on the younger runners a little bit. I think that some of the 7th graders are filling empty slots on smaller teams so not all of them would finish very high.

Now I wanted to find all the runners who have made at least 2 trips to the championship race:

```
## # A tibble: 3,154 x 7
## # Groups:   xc_data$Name [1,120]
##    Place Name              Grade  Time School  Year `xc_data$Name`
##    <dbl> <chr>             <dbl> <dbl> <chr>  <dbl> <chr>
##  1     1 Carrie Tollefson      9  17.3 Dblv    1991 Carrie Tollefson
##  2     2 Tina Forthmiller     10  17.9 StF     1991 Tina Forthmiller
##  3     3 Kara Wheeler          8  17.9 DuE     1991 Kara Wheeler
##  4     4 Amy Hill              8  17.9 DuE     1991 Amy Hill
##  5     5 Keri Zweig           11  18.1 Mtk     1991 Keri Zweig
##  6     6 Turena Johnson       11  18.1 Brn     1991 Turena Johnson
##  7     8 Julie Golla           9  18.3 Rjm     1991 Julie Golla
##  8    12 Stephanie Simones    10  18.4 Msw     1991 Stephanie Simones
##  9    13 Andrea Lentz         10  18.5 Wil     1991 Andrea Lentz
## 10    15 Barb Jones            9  18.6 Wbr     1991 Barb Jones
## # ... with 3,144 more rows
```

The ratio of runners who ran in the championships more than once versus overall per grade:

```
## [1] 0.7272727
```

```
## [1] 0.7175258
```

```
## [1] 0.7167702
```

```
## [1] 0.7170582
```

```
## [1] 0.6979167
```

```
## [1] 0.5904936
```

Then I wanted to find all the girls in top ten over the 30 year period:

```
## # A tibble: 271 x 7
## # Groups:   xc_data$Name [152]
##    Place Name             Grade  Time School       Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>       <dbl> <chr>
##  1     1 Carrie Tollefson     9  17.3 Dblv         1991 Carrie Tollefson
##  2     2 Tina Forthmiller    10  17.9 StF          1991 Tina Forthmiller
##  3     3 Kara Wheeler         8  17.9 DuE          1991 Kara Wheeler
##  4     4 Amy Hill             8  17.9 DuE          1991 Amy Hill
##  5     5 Keri Zweig          11  18.1 Mtk          1991 Keri Zweig
##  6     6 Turena Johnson      11  18.1 Brn          1991 Turena Johnson
##  7     8 Julie Golla          9  18.3 Rjm          1991 Julie Golla
##  8     2 Amy Hill             9  19.0 Duluth East  1992 Amy Hill
##  9     3 Missy Johnson       11  19.1 Hibbing      1992 Missy Johnson
## 10     4 Turena Johnson      12  19.2 Brainerd     1992 Turena Johnson
## # ... with 261 more rows
```

To begin to test my theory of girls not being prodigies when they run fast in 9th grade, I made a data set of all the girls who were in the top 10 in 9th and 12th grade and who had run in the championship more than once:

```
## # A tibble: 119 x 7
## # Groups:   xc_data$Name [105]
##    Place Name             Grade  Time School                Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>                <dbl> <chr>
##  1     1 Carrie Tollefson     9  17.3 Dblv                  1991 Carrie Tollefs~
##  2     8 Julie Golla          9  18.3 Rjm                   1991 Julie Golla
##  3     2 Amy Hill             9  19.0 Duluth East           1992 Amy Hill
##  4     5 Kara Wheeler         9  19.2 Duluth East           1992 Kara Wheeler
##  5     7 Casey Cherne         9  19.6 Duluth East           1993 Casey Cherne
##  6     8 Yvonne Glenn         9  19.6 Duluth Central        1993 Yvonne Glenn
##  7    10 Amy Maciasek         9  19.7 Mounds View           1993 Amy Maciasek
##  8     4 Beth Rautmann        9  18.9 White Bear Lake Area  1995 Beth Rautmann
##  9     7 Serena Sullivan      9  19.0 Hibbing               1995 Serena Sullivan
## 10    10 Kelly Brinkman       9  19.3 Hutchinson            1995 Kelly Brinkman
## # ... with 109 more rows
```

To get the real number of girls in this, I needed to find the duplicate names:

```
## # A tibble: 14 x 7
## # Groups:   xc_data$Name [14]
##    Place Name             Grade  Time School                 Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>                 <dbl> <chr>
##  1     1 Carrie Tollefson    12  17.8 Lac Qui Parle Val/D B  1994 Carrie Tollef~
##  2     3 Kara Wheeler        12  18.7 Duluth East            1995 Kara Wheeler
```

```
##  3     8 Amy Hill            12  19.1 Duluth East            1995 Amy Hill
##  4     3 Kendall Wheeler     12  18.6 Duluth East            1999 Kendall Wheel~
##  5     9 Nicole McCann       12  18.4 Owatonna               2003 Nicole McCann
##  6     1 Elizabeth Yetzer    12  17.6 Lakeville North        2005 Elizabeth Yet~
##  7     2 Jamie Piepenburg    12  17.7 Alexandria             2011 Jamie Piepenb~
##  8     7 Anna French         12  18.0 Wayzata                2014 Anna French
##  9     1 Bethany Hasz        12  17.5 Alexandria             2015 Bethany Hasz
## 10     2 Megan Hasz          12  17.6 Alexandria             2015 Megan Hasz
## 11     4 Tess Misgen         12  18.5 Shakopee               2016 Tess Misgen
## 12     1 Emily Covert        12  17.1 Minneapolis Washburn   2018 Emily Covert
## 13     2 Lauren Peterson     12  17.5 Farmington             2018 Lauren Peters~
## 14     1 Ali Weimer          12  17.7 St. Michael-Albertvil  2021 Ali Weimer
```

As you can see, of the 105 girls who fit in the previous filter, only 14 had been top 10 in 9th and 12th grade.

Now, lets do the same for 11th and 12th grade:

```
## # A tibble: 133 x 7
## # Groups:   xc_data$Name [103]
##    Place Name             Grade  Time School            Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>            <dbl> <chr>
##  1     5 Keri Zweig          11  18.1 Mtk               1991 Keri Zweig
##  2     6 Turena Johnson      11  18.1 Brn               1991 Turena Johnson
##  3     3 Missy Johnson       11  19.1 Hibbing           1992 Missy Johnson
##  4     6 Andrea Lentz        11  19.3 Willmar           1992 Andrea Lentz
##  5     4 Julie Herrmann      11  19.3 Saint Louis Park  1993 Julie Herrmann
##  6     6 Anna Gullingsrud    11  19.6 Mounds View       1993 Anna Gullingsrud
##  7     2 Kara Wheeler        11  17.8 Duluth East       1994 Kara Wheeler
##  8     5 Amy Hill            11  18.6 Duluth East       1994 Amy Hill
##  9     6 Heather Anderson    11  18.6 Osseo             1994 Heather Anderson
## 10     1 Elaine Eggleston    11  18.7 Roseville Area    1995 Elaine Eggleston
## # ... with 123 more rows
```

Finding the duplicate names:

```
## # A tibble: 30 x 7
## # Groups:   xc_data$Name [30]
##    Place Name             Grade  Time School          Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>          <dbl> <chr>
##  1     4 Turena Johnson      12  19.2 Brainerd        1992 Turena Johnson
##  2     7 Keri Zweig          12  19.6 Minnetonka      1992 Keri Zweig
##  3     3 Kara Wheeler        12  18.7 Duluth East     1995 Kara Wheeler
##  4     8 Amy Hill            12  19.1 Duluth East     1995 Amy Hill
##  5     7 Elaine Eggleston    12  19.2 Roseville Area  1996 Elaine Eggleston
##  6     3 Victoria Moses      12  18.8 Irondale        1997 Victoria Moses
##  7     3 Kendall Wheeler     12  18.6 Duluth East     1999 Kendall Wheeler
##  8     3 Lauren Burks        12  18.4 Park            2001 Lauren Burks
##  9     4 Kari Higdem         12  18.5 Willmar         2004 Kari Higdem
## 10     6 Katie Anderson      12  18.8 Blaine          2004 Katie Anderson
## # ... with 20 more rows
```

For 11th and 12th grade, 30 out of 103 girls had been in the top 10 more than once.

The ratios of that would be:

```
14/105
```

```
## [1] 0.1333333
```

```
30/103
```

```
## [1] 0.2912621
```

To add a little more data, I wanted to run the same experiment but with the top 20 finishers.

```
## # A tibble: 528 x 7
## # Groups:   xc_data$Name [292]
##    Place Name              Grade  Time School  Year `xc_data$Name`
##    <dbl> <chr>             <dbl> <dbl> <chr>  <dbl> <chr>
##  1     1 Carrie Tollefson      9  17.3 Dblv    1991 Carrie Tollefson
##  2     2 Tina Forthmiller     10  17.9 StF     1991 Tina Forthmiller
##  3     3 Kara Wheeler          8  17.9 DuE     1991 Kara Wheeler
##  4     4 Amy Hill              8  17.9 DuE     1991 Amy Hill
##  5     5 Keri Zweig           11  18.1 Mtk     1991 Keri Zweig
##  6     6 Turena Johnson       11  18.1 Brn     1991 Turena Johnson
##  7     8 Julie Golla           9  18.3 Rjm     1991 Julie Golla
##  8    12 Stephanie Simones    10  18.4 Msw     1991 Stephanie Simones
##  9    13 Andrea Lentz         10  18.5 Wil     1991 Andrea Lentz
## 10    15 Barb Jones            9  18.6 Wbr     1991 Barb Jones
## # ... with 518 more rows
```

The top 20 multi-championship 9th and 12th graders:

```
## # A tibble: 224 x 7
## # Groups:   xc_data$Name [197]
##    Place Name             Grade  Time School          Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>          <dbl> <chr>
##  1     1 Carrie Tollefson     9  17.3 Dblv            1991 Carrie Tollefson
##  2     8 Julie Golla          9  18.3 Rjm             1991 Julie Golla
##  3    15 Barb Jones           9  18.6 Wbr             1991 Barb Jones
##  4     2 Amy Hill             9  19.0 Duluth East     1992 Amy Hill
##  5     5 Kara Wheeler         9  19.2 Duluth East     1992 Kara Wheeler
##  6    14 Heather Anderson     9  19.7 Osseo           1992 Heather Anderson
##  7    15 Jenny Fiedler        9  19.7 Buffalo         1992 Jenny Fiedler
##  8     7 Casey Cherne         9  19.6 Duluth East     1993 Casey Cherne
##  9     8 Yvonne Glenn         9  19.6 Duluth Central  1993 Yvonne Glenn
## 10    10 Amy Maciasek         9  19.7 Mounds View     1993 Amy Maciasek
## # ... with 214 more rows
```

Finding duplicates

```
## # A tibble: 27 x 7
## # Groups:   xc_data$Name [27]
##    Place Name                Grade  Time School             Year `xc_data$Name`
##    <dbl> <chr>               <dbl> <dbl> <chr>             <dbl> <chr>
##  1     1 Carrie Tollefson       12  17.8 Lac Qui Parle Va~  1994 Carrie Tollef~
```

```
##  2     3 Kara Wheeler            12  18.7 Duluth East        1995 Kara Wheeler
##  3     8 Amy Hill                12  19.1 Duluth East        1995 Amy Hill
##  4     3 Kendall Wheeler         12  18.6 Duluth East        1999 Kendall Wheel~
##  5    12 Courtney Hugstad-Vaa    12  19.0 Eastview           2000 Courtney Hugs~
##  6    15 Jessica Goeden          12  19.1 Grand Rapids       2001 Jessica Goeden
##  7    18 Veronica Sackett        12  19.2 Grand Rapids       2001 Veronica Sack~
##  8    12 Courtney Dauwalter      12  19.7 Hopkins            2002 Courtney Dauw~
##  9     9 Nicole McCann           12  18.4 Owatonna           2003 Nicole McCann
## 10    13 Ladia Albertson         12  18.6 Stillwater Area    2003 Ladia Alberts~
## # ... with 17 more rows
```

For the top 20, 27 of 197 girls had been in the top 20 in 9th and 12th grade.

Again, the top 20 for 11th and 12th grade:

```
## # A tibble: 248 x 7
## # Groups:   xc_data$Name [188]
##    Place Name             Grade  Time School              Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>              <dbl> <chr>
##  1     5 Keri Zweig          11  18.1 Mtk                 1991 Keri Zweig
##  2     6 Turena Johnson      11  18.1 Brn                 1991 Turena Johnson
##  3    19 Becky Loberg        11  18.7 Buf                 1991 Becky Loberg
##  4     3 Missy Johnson       11  19.1 Hibbing             1992 Missy Johnson
##  5     6 Andrea Lentz        11  19.3 Willmar             1992 Andrea Lentz
##  6    12 Jessica Faith       11  19.7 Saint Cloud Apollo  1992 Jessica Faith
##  7    13 Jaime Miller        11  19.7 Duluth East         1992 Jaime Miller
##  8    16 Tina Forthmiller    11  19.8 Saint Francis       1992 Tina Forthmiller
##  9     4 Julie Herrmann      11  19.3 Saint Louis Park    1993 Julie Herrmann
## 10     6 Anna Gullingsrud    11  19.6 Mounds View         1993 Anna Gullingsrud
## # ... with 238 more rows
```

Finding the duplicate names:

```
## # A tibble: 60 x 7
## # Groups:   xc_data$Name [59]
##    Place Name             Grade  Time School              Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>              <dbl> <chr>
##  1    13 Jessica Goeden      11  19.1 Grand Rapids        2000 Jessica Goeden
##  2     4 Turena Johnson      12  19.2 Brainerd            1992 Turena Johnson
##  3     7 Keri Zweig          12  19.6 Minnetonka          1992 Keri Zweig
##  4    11 Missy Johnson       12  19.7 Hibbing             1993 Missy Johnson
##  5    14 Jessica Faith       12  19.8 Saint Cloud Apollo  1993 Jessica Faith
##  6     4 Amber Affeldt       12  18.6 Coon Rapids         1994 Amber Affeldt
##  7    12 Anna Gullingsrud    12  18.8 Mounds View         1994 Anna Gullingsrud
##  8    20 Carrie Palmer       12  19.2 Stillwater Area     1994 Carrie Palmer
##  9     3 Kara Wheeler        12  18.7 Duluth East         1995 Kara Wheeler
## 10     8 Amy Hill            12  19.1 Duluth East         1995 Amy Hill
## # ... with 50 more rows
```

For 11th and 12th grade, 60 out of 188 girls had been in the top 20 more than once.

The ratios of that would be:

```
27/197
```

```
## [1] 0.1370558
```

```
60/180
```

```
## [1] 0.3333333
```

The ratio between the top 10 and top 20 is very similar.

Finally, I wanted to look at 8th graders to see how many stopped appearing as they got older. This time I am using 8 and 10th grade as my two data sets.

```
## # A tibble: 82 x 7
## # Groups:   xc_data$Name [70]
##    Place Name                Grade  Time School              Year `xc_data$Name`
##    <dbl> <chr>               <dbl> <dbl> <chr>              <dbl> <chr>
##  1     3 Kara Wheeler            8  17.9 DuE                 1991 Kara Wheeler
##  2     4 Amy Hill                8  17.9 DuE                 1991 Amy Hill
##  3    10 Beth Rautmann           8  18.8 White Bear Lake A~  1994 Beth Rautmann
##  4     9 Kendall Wheeler         8  19.1 Duluth East         1995 Kendall Wheel~
##  5     1 Josie Johnson           8  18.6 Rochester John Ma~  1996 Josie Johnson
##  6     6 Kelsey Dahlgren         8  19.1 Centennial          1997 Kelsey Dahlgr~
##  7     6 Kassandra Hendricks     8  18.7 Hutchinson          1998 Kassandra Hen~
##  8     9 Jenny Trump             8  19.0 Bloomington Jeffe~  1999 Jenny Trump
##  9     5 Elizabeth Yetzer        8  18.6 Lakeville           2001 Elizabeth Yet~
## 10    10 Brenna Carey            8  18.9 Big Lake            2001 Brenna Carey
## # ... with 72 more rows
```

Filtering duplicate names:

```
## # A tibble: 12 x 7
## # Groups:   xc_data$Name [12]
##    Place Name             Grade  Time School                Year `xc_data$Name`
##    <dbl> <chr>            <dbl> <dbl> <chr>                <dbl> <chr>
##  1     1 Kara Wheeler        10  18.7 Duluth East           1993 Kara Wheeler
##  2     3 Amy Hill            10  19.3 Duluth East           1993 Amy Hill
##  3     3 Beth Rautmann       10  18.9 White Bear Lake Area  1996 Beth Rautmann
##  4     2 Elizabeth Yetzer    10  18.1 Lakeville             2003 Elizabeth Yet~
##  5     1 Laura Hughes        10  18.2 Mankato West          2007 Laura Hughes
##  6     1 Maria Hauger        10  18.0 Shakopee              2010 Maria Hauger
##  7     1 Bethany Hasz        10  17.5 Alexandria            2013 Bethany Hasz
##  8     2 Megan Hasz          10  17.9 Alexandria            2013 Megan Hasz
##  9     8 Emily Covert        10  18.6 Minneapolis Washburn  2016 Emily Covert
## 10     3 Anna Fenske         10  17.9 Farmington            2018 Anna Fenske
## 11     4 Ali Weimer          10  18   St. Michael-Albertvil 2019 Ali Weimer
## 12     6 Molly Moening       10  18.3 St Paul Highland Park 2019 Molly Moening
```

12 out of 72 girls in the top 10 where there in 8th and 10th grade.

```
12/72
```

```
## [1] 0.1666667
```

Again, a very similar ratio to the 9th and 12th grade data.

Topics From Class

RMarkdown

Honestly, I love R Markdown. It is relatively easy to use and really does a great job of presenting your data/findings in a very organized way. I still have a lot to learn about R (such as functions which I mentioned above) but I can't see myself ever presenting data in a different way now.

Github

I have heard of Github and used it to download some code but I've never actually used it before. It took some figuring out for me but I eventually got it to work and I really like it. I can definitely see myself using it more in my schoolwork. Learning the concepts will also help me with future software jobs where they may use a similar type of system.

Regression

I needed to do linear regression to find the line fits and the slope intercept of my data. I enjoy algebra and have used y=mx+b frequently in my education and career but I had never really done it with large sets of data before. It was interesting to see how it worked and I like how R does it compared to Excel. If I continue with this topic in future data analysis classes I will see if I can find a exponential decay fit for it.

Probability

I think finding the probability for girls who were good runners in different grades was actually my favorite part of this entire project. It was also a good opportunity to see aspects to R we didn't learn in class and, ultimately, it was very interesting to see the ratios in my data. Using probabilities also helped me see some gaps in my data that I normally wouldn't have thought of. For instance, there are far fewer 7th and 8th grade data points and that impacted a few of my calculations.

Normal Distribution

Coming from my draft, I had a different approach to this. Listening to feedback from my classmates and seeing what others did with their projects had me rethink of how to show the data distribution. Instead of showing just one denisty plot of all the times from over the 30 years, I thought it would be better to show two different years in order to demonstrate the change in distribution.

Conclusion

The data shows that my initial hypothesis, that times have gotten faster and the talent depth has improved, was correct. I also think there is evidence that fast girls do slow down as they age but I am not sure if there is enough data to have a definite conclusion. Cross country data has always been a passion of mine so it was fun to actually do something with everything I have observed over the last 20 years. While I have always loved playing with data and math, stats has never been my strong suit so doing this helped me a lot in understanding how to use statistics.

I learned a lot about R and re familiarized myself with some math (notably using natural logarithms when I was trying to fit a curve to the data and ended up not doing). As stated above, I really wanted to fit a line showing the decay on the scatter plots but I don't think the data was uniform enough for that. Maybe there is a way to do it with the data I have, if so I would like to learn what that is. Due to that, I don't think a predictive model would be accurate at this time due to it being based on a linear trend and not a logarithmic one.

One thing I would have liked to do is show more plots and make more data sets. I think that would tell a better story overall but I felt that was probably too much for this project at this time.

Future Improvements

In the future, one thing I'd like to do is make a function that will automatically create the subsets depending on what data you give it. I could do it in Python easily but I am not as familiar with how to write functions in R.

Sources

I made extended use of Statology for basically every question I had.

I used this site for help in setting up GitHub for my project since I wasn't able to do it correctly in class.