Henry Gillis Intro to Data Science HW #7

Contents

Intro/Data set Description

This data set is about all drivers that have competed in Formula 1, it includes all races up to the 2023 Bahrain Grand Prix. The data that is included:

- Driver Name
- Driver Nationality
- Seasons competed
- Number of championships
- Number of race entries
- Number of race starts
- Number of pole positions
- Number of race wins
- Number of podiums
- Number of fastest laps
- Number of points

This data was sourced from: https://www.kaggle.com/datasets/dubradave/formula-1-drivers-dataset

```
#loading all necessary data and libraries
library(tidyverse)

df <- read_csv("F1DriversDataset.csv")
```

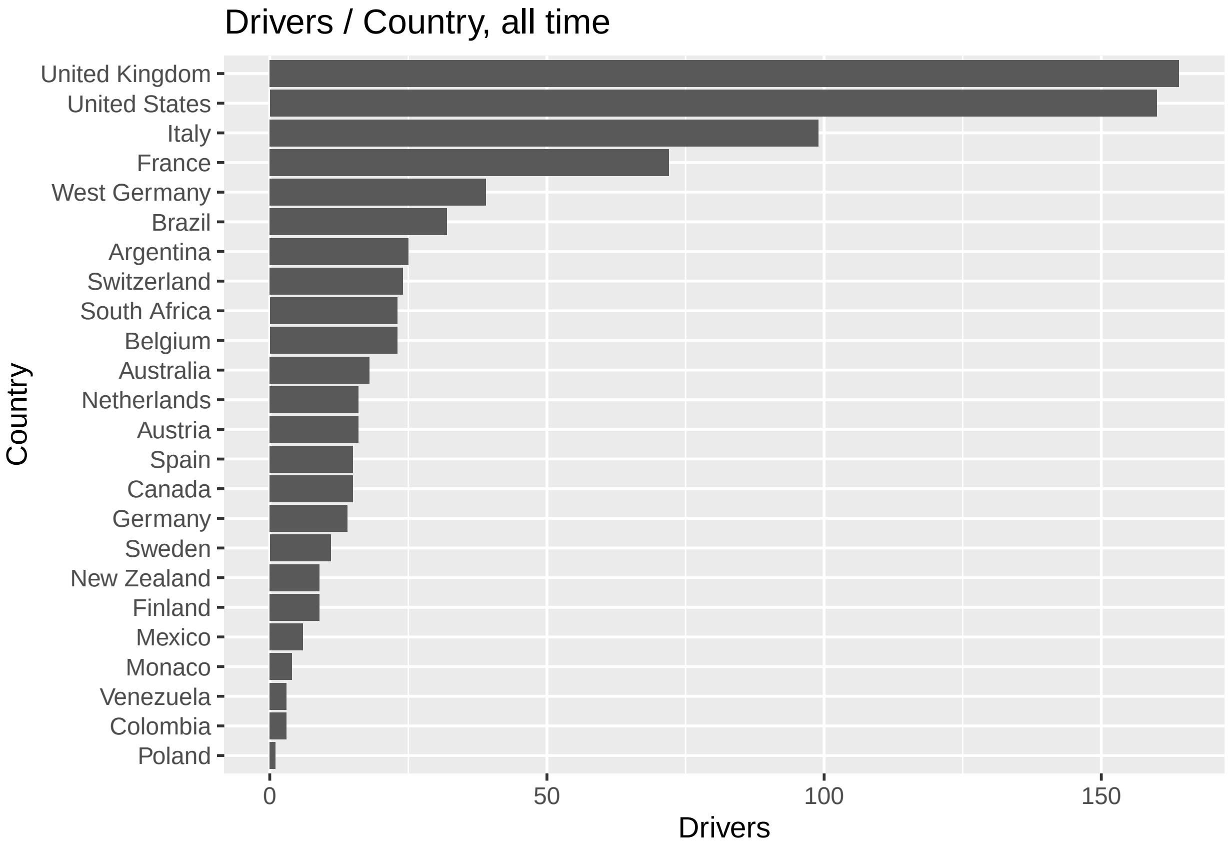
```
country_count <-  df %>%
  group_by(Nationality) %>%
    summarise(
     Drivers = n(),
     Winners = sum(Race_Wins > 0),
      )

country_count <- country_count %>%
        mutate(winning_driver_percent = (country_count$Winners / country_count$Drivers) * 100) %>%
  arrange(desc(winning_driver_percent))


#filter out countries with 0 race winners
filtered_country_count <- filter(country_count, winning_driver_percent > 0)

df
```

```
## # A tibble: 868 x 22
##    Driver            Nationality  Seasons Championships Race_Entries Race_Starts
##    <chr>             <chr>        <chr>           <dbl>        <dbl>       <dbl>
##  1 Carlo Abate       Italy        [1962,~             0            3           0
##  2 George Abecassis  United King~ [1951,~             0            2           2
##  3 Kenny Acheson     United King~ [1983,~             0           10           3
##  4 Andrea de Adamich Italy        [1968,~             0           36          30
##  5 Philippe Adams    Belgium      [1994]              0            2           2
##  6 Walt Ader         United Stat~ [1950]              0            1           1
##  7 Kurt Adolff       West Germany [1953]              0            1           1
##  8 Fred Agabashian   United Stat~ [1950,~             0            9           8
##  9 Kurt Ahrens Jr.   West Germany [1966,~             0            4           4
## 10 Jack Aitken       United King~ [2020]              0            1           1
## # i 858 more rows
## # i 16 more variables: Pole_Positions <dbl>, Race_Wins <dbl>, Podiums <dbl>,
## #   Fastest_Laps <dbl>, Points <dbl>, Active <lgl>, ‘Championship Years‘ <chr>,
## #   Decade <dbl>, Pole_Rate <dbl>, Start_Rate <dbl>, Win_Rate <dbl>,
## #   Podium_Rate <dbl>, FastLap_Rate <dbl>, Points_Per_Entry <dbl>,
## #   Years_Active <dbl>, Champion <lgl>
```

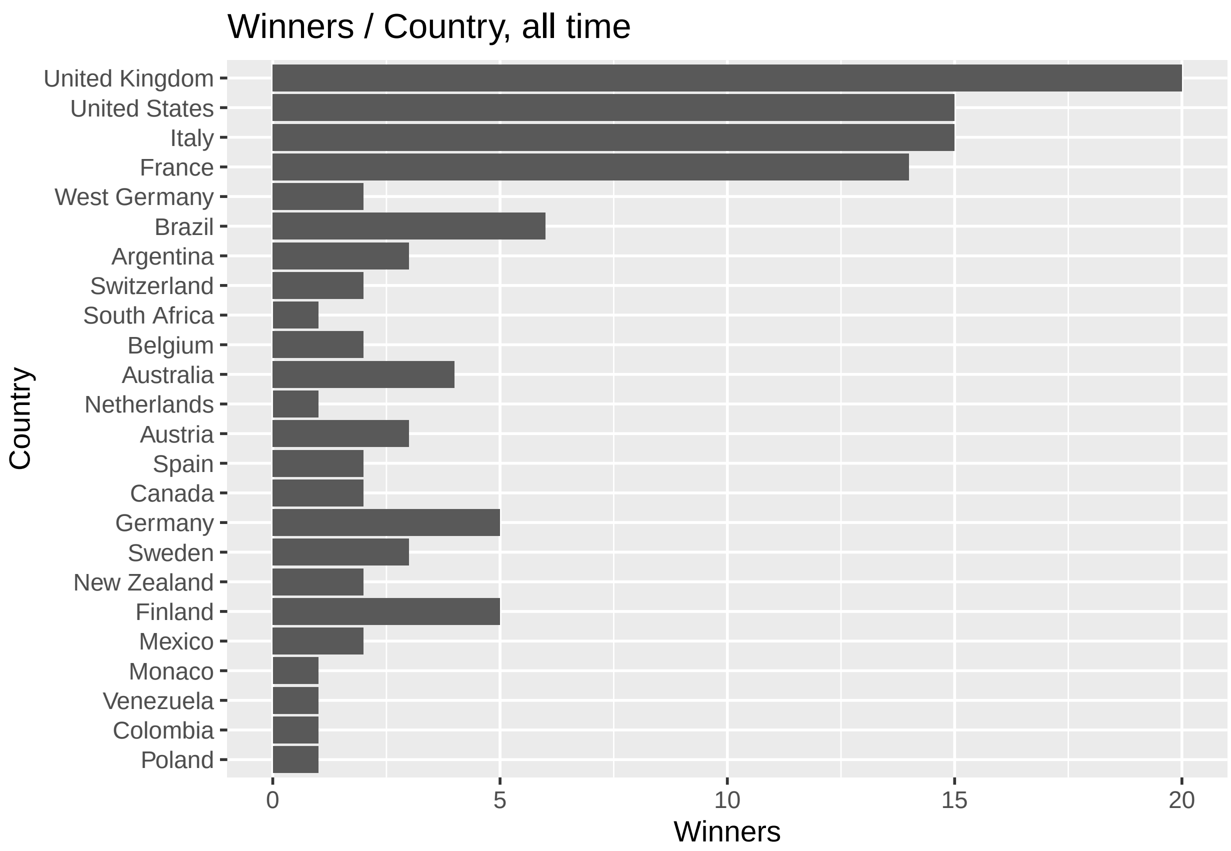
Question #1

How many drivers and race winners has each country had?

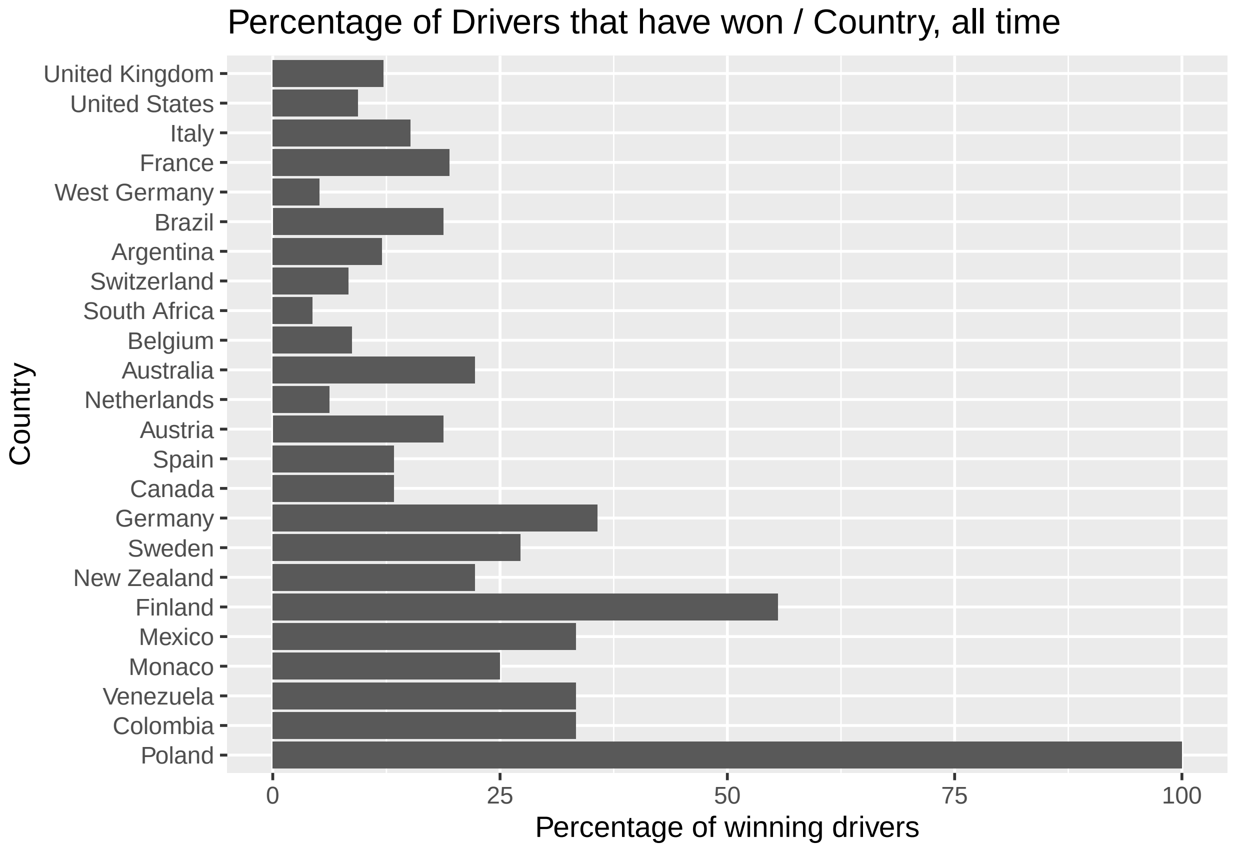
```
ggplot(data=filtered_country_count, aes(x= Drivers, y=reorder(Nationality, Drivers))) +
  geom_bar(stat="identity") +
  labs(title = "Drivers / Country, all time", x = "Drivers", y = "Country")
```



```
ggplot(data=filtered_country_count, aes(x= Winners, y=reorder(Nationality, Drivers))) +
  geom_bar(stat="identity") +
  labs(title = "Winners / Country, all time", x = "Winners", y = "Country")
```



```
ggplot(data=filtered_country_count, aes(x= winning_driver_percent, y=reorder(Nationality, Drivers))) +
  geom_bar(stat="identity") +
  labs(title = "Percentage of Drivers that have won / Country, all time", x = "Percentage of winning dri
```

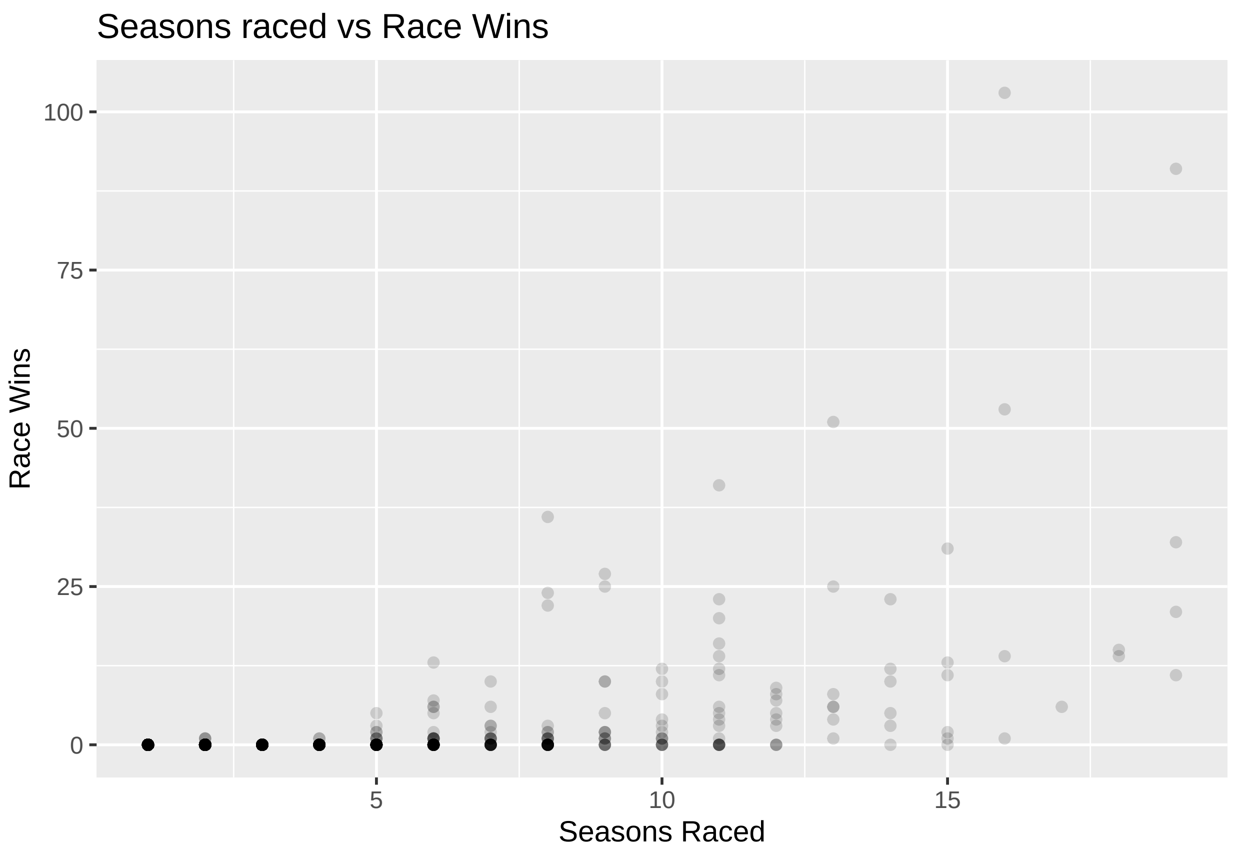


Something interesting that we can see from these graphs, is that there is a trend for countries with less drivers, those drivers seem to be more likely to be a race winner. My guess for this would be that if a driver from that country is going to make it to F1, they will need to be a better talent, since F1 is not as popular their, for them to even make it to F1 they already really need to be a good driver. As opposed to somewhere like the United Kingdom where karting is more common, and it wouldn't be quite as hard to be noticed by teams.

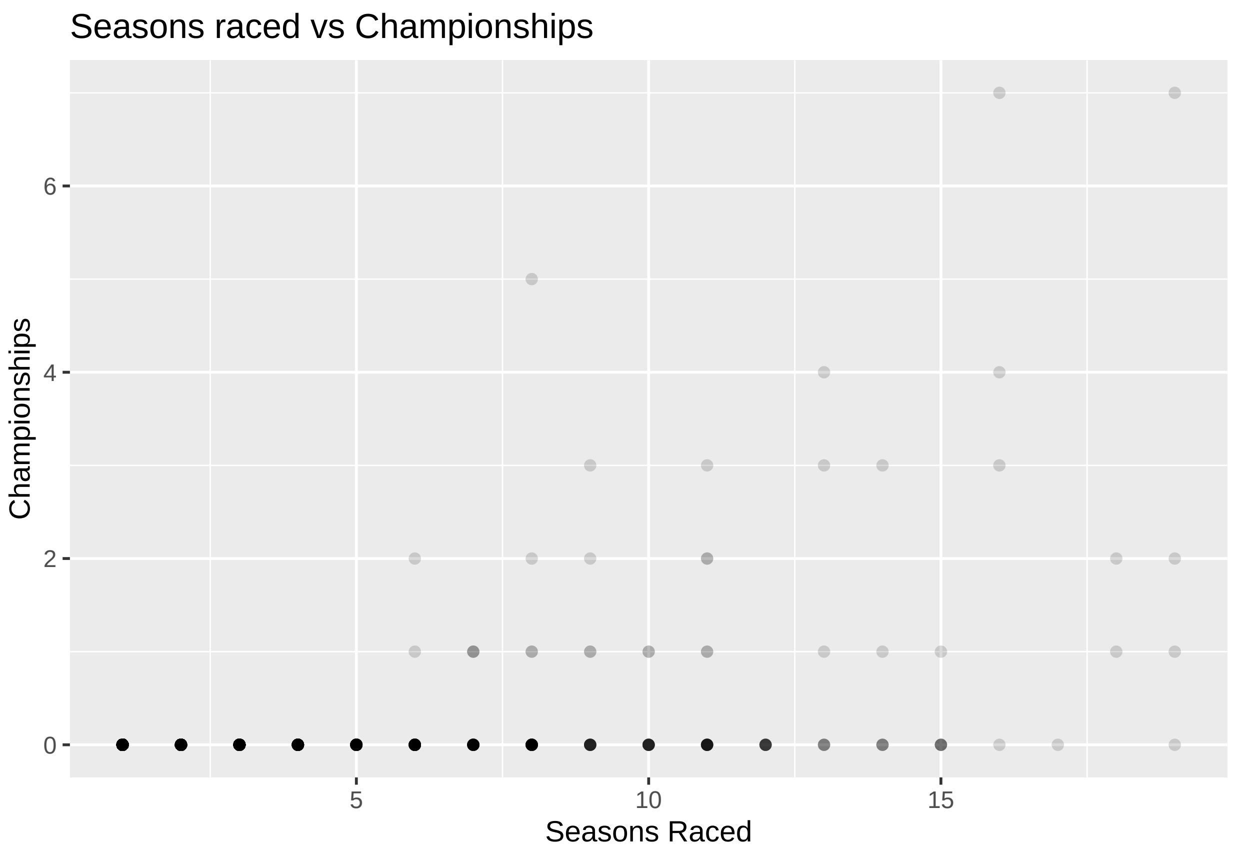
Question #2

Is there is any correlation between seasons raced and "success"?

```
ggplot(data=df, aes(x= Years_Active, y= Race_Wins)) +
  geom_point(stat="identity", color = "black", alpha = .15) +
  labs(title = "Seasons raced vs Race Wins", x = "Seasons Raced", y = "Race Wins")
```



```r
ggplot(data=df, aes(x= Years_Active, y= Championships)) +
  geom_point(stat="identity", color = "black", alpha = .15) +
  labs(title = "Seasons raced vs Championships", x = "Seasons Raced", y = "Championships")
```



While it does appear that the more season raced, the more drivers win races and championships, something that I found to be interesting was that no one who retired with less than 6 seasons has ever won a Championship, and the trend of winning more races as you drive more seasons doesn't really start up until we look at drivers who retired with at least 5 years of racing, which does make a bit of sense as drivers who were unsuccessful would have been fired by their team, and not picked up by another, leading to an early retirement with no race wins.

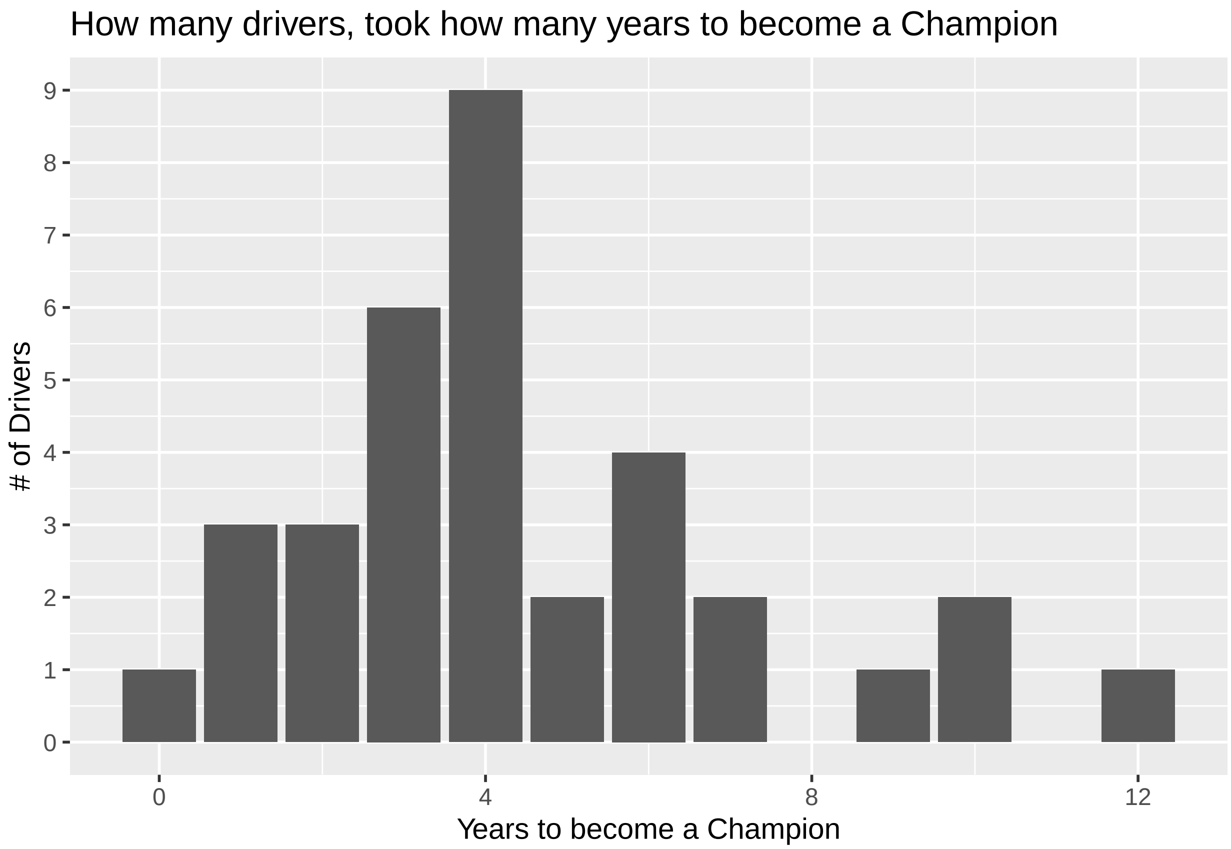
Question #3

Is there a trend in how long it takes drivers to win a championship?

```
#filter the whole list of drivers down to just those that have won a championship
champions <- filter(df, Champion == TRUE)

#add a row to champions df, that extracts the year they won their first championship from Championship
champions <- champions %>%
  mutate(first_championship = as.numeric(substr(champions$`Championship Years`, 2,5)))

#add a row to champions df, that extracts the year they debuted from Championship Years, and turn it in
champions <- champions %>%
  mutate(first_season = as.numeric(substr(champions$Seasons, 2,5)))

#add a new row where we subtract their debut season from their first championship season to see how man
```

```
champions <- champions %>%
  mutate(years_to_champ = first_championship - first_season)

#group by years taken to become a champ
champions_groupped <-  champions %>%
  group_by(years_to_champ) %>%
    summarise(
     Drivers = n()
      )
```

```
ggplot(data=champions_groupped, aes(x=years_to_champ, y=Drivers)) +
  geom_bar(stat="identity") +
  labs(title = "How many drivers, took how many years to become a Champion", x = "Years to become a Char
  scale_y_continuous(breaks = c(0,1,2,3,4,5,6,7,8,9,10))
```



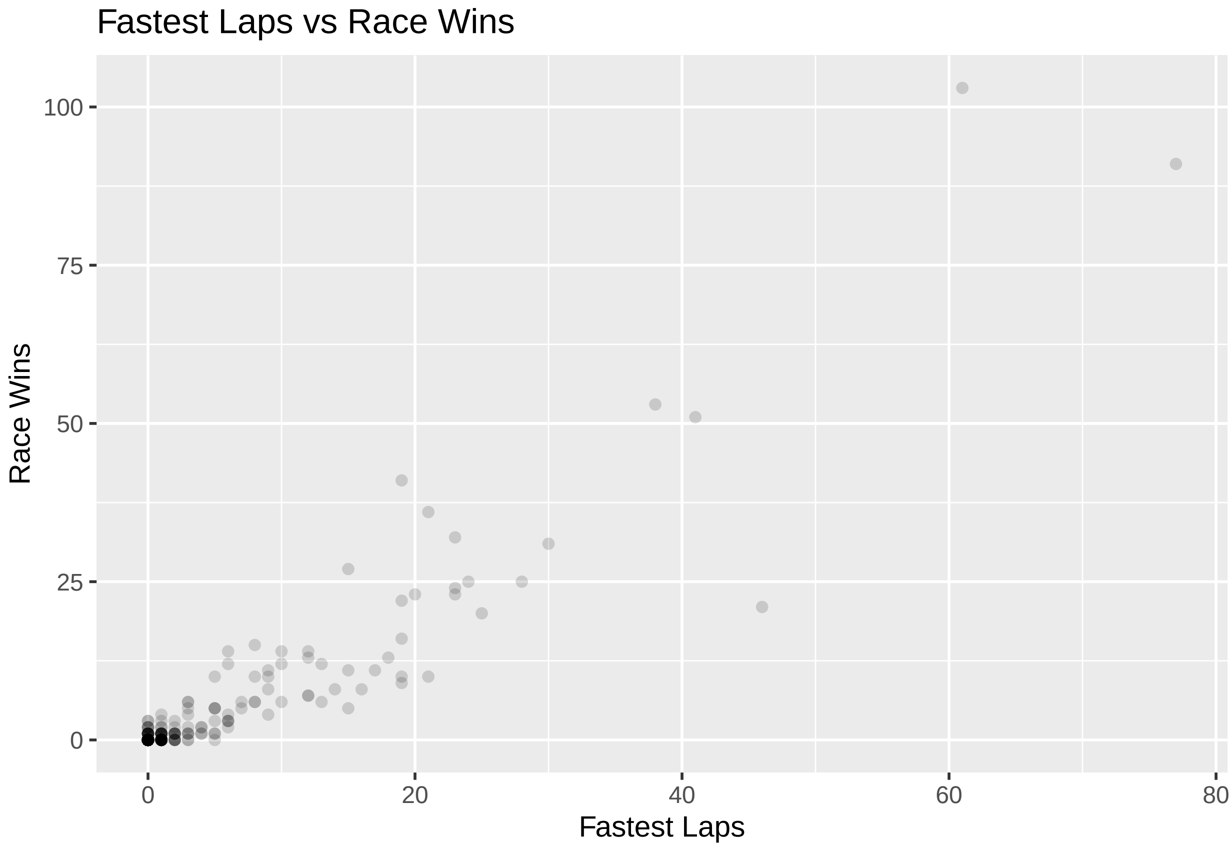
Looking at this graph, there does seem to be a trend, with there being a bell curve, centered on 4 years taken to become a champion, with the number of people taking longer falling off, and the number of people taking less time building up to that 4 years.

Question #4

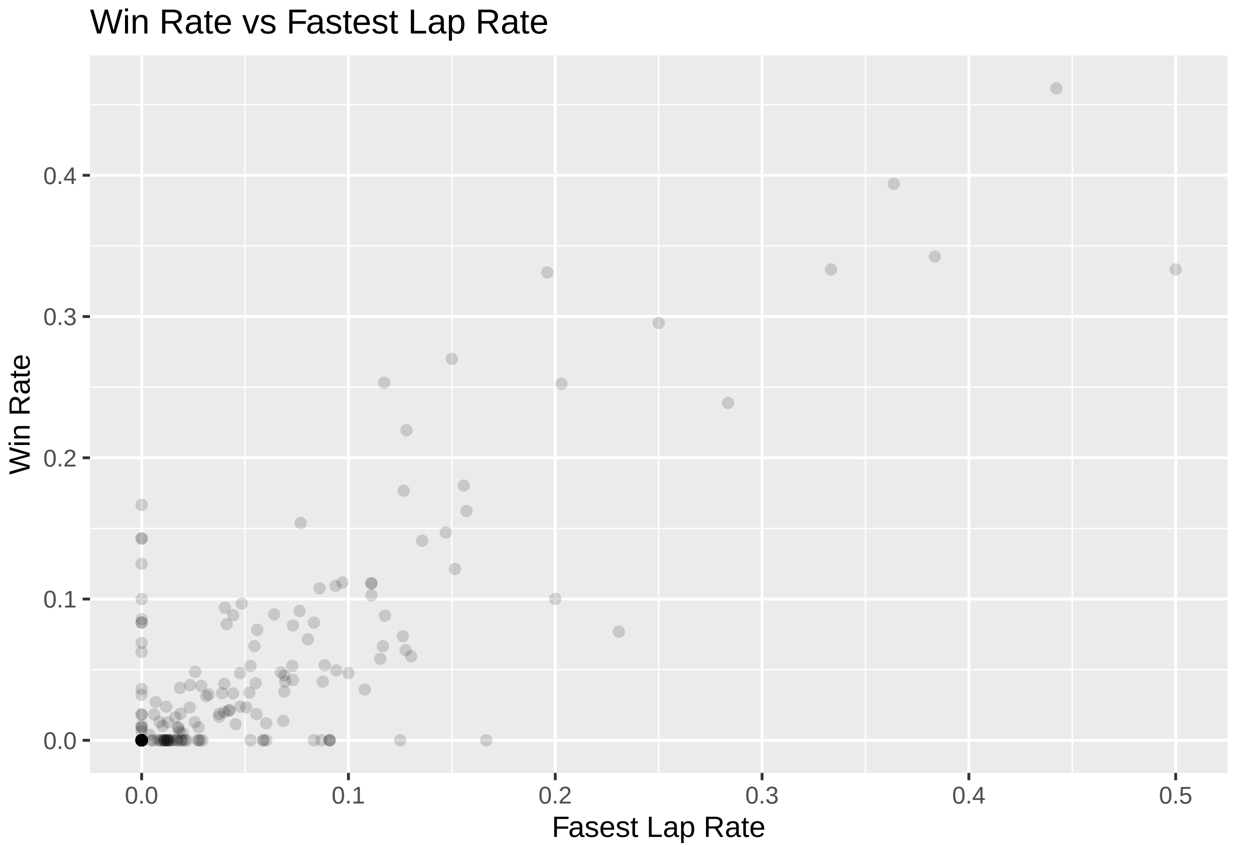
if their is any correlation between fastest laps and race wins.

```
fastest_winner <- df %>%
  select(Race_Wins, Fastest_Laps, Win_Rate, FastLap_Rate)
```

```
ggplot(data=fastest_winner, aes(x= Fastest_Laps, y= Race_Wins)) +
  geom_point(stat="identity", color = "black", alpha = .15) +
  labs(title = "Fastest Laps vs Race Wins", x = "Fastest Laps", y = "Race Wins")
```



```
ggplot(data=fastest_winner, aes(x= FastLap_Rate, y= Win_Rate)) +
  geom_point(stat="identity", color = "black", alpha = .15) +
  labs(title = "Win Rate vs Fastest Lap Rate", x = "Fasest Lap Rate", y = "Win Rate")
```



There is definitely a correlation between fastest laps and wins. But what i find most interesting is that there is a decent number of drivers that are along either the x or y axis, meaning that they either have some race wins without getting fastest laps or vice versa. This only happens within a square starting at the origin up to (.2, .2) which makes me think that those drivers got their win or fastest lap through a fluke/luck rather than skill.

Conclusion

Some analysis that I would like to be able to do in the future is to look at who is that number one all time driver. This would be done by looking at drivers head to head and seeing who comes out better, making some kind of ELO system to find the best driver. This could not be done with this data set however since it does not go into detail of each race, simply the cumulative results. What we could do with this set is look to see if there is a change in how important fastest laps are as time goes on. Since in the early seasons of F1, car reliability was non existent, having a fast car might not bbe the best if it couldn't finish races.