Formula 1 Project Milestone

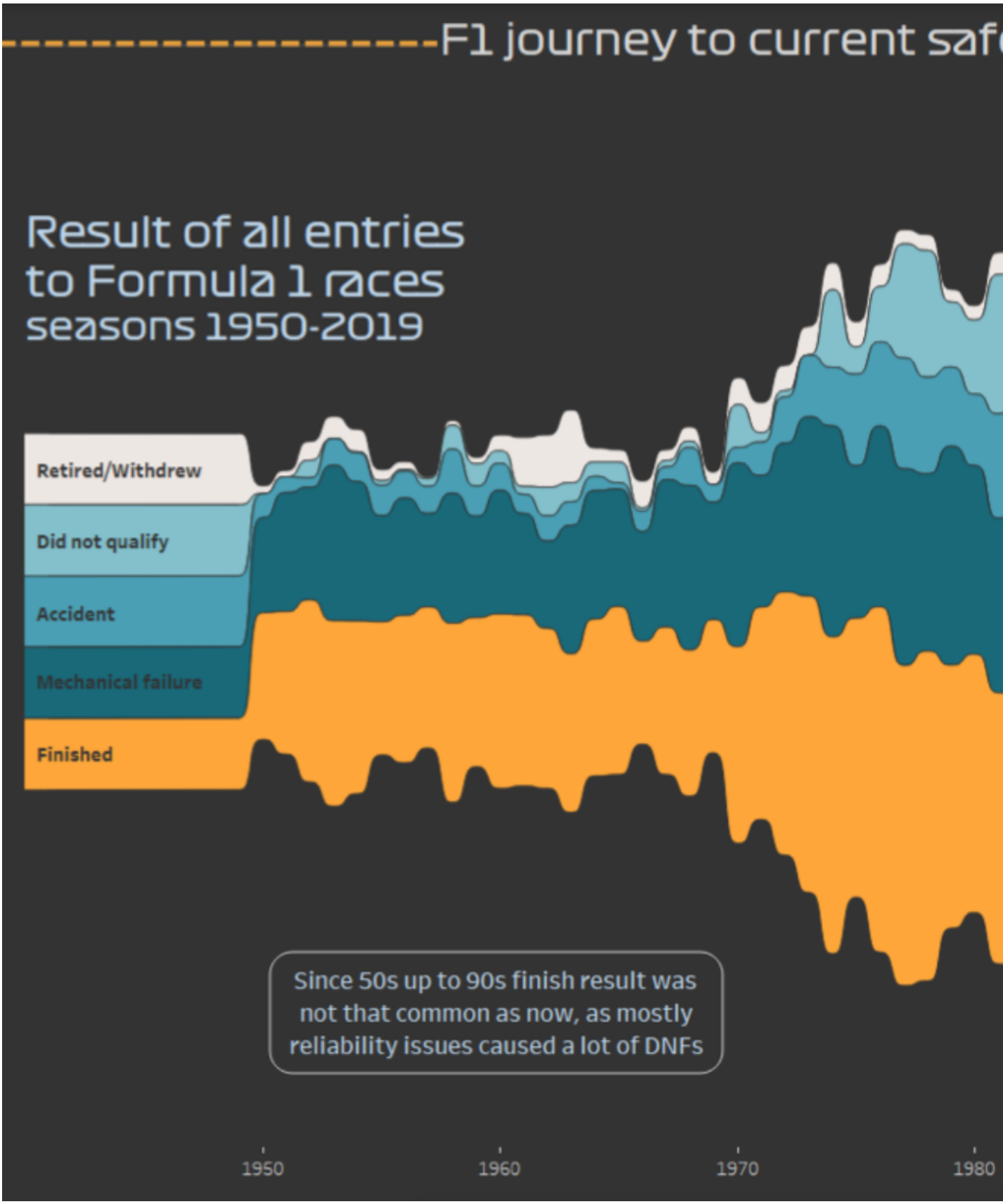
Brea Koenes

10/28/2020

Overview

"F1 journey to current safety standards was not an easy one" is the claim the original graph made, which I am replicating in my visualization. I am working with numerical and categorical data. Specifically, the variables are the year, result of all entries, and the sum of the entries' results. The graph will show the count of the statuses of the entries by year, categorized by 5 status results.

Visualization



Dataset

The original visualization was taken from [this article] (https://www.reddit.com/r/formula1/comments/j4fdcr/is_formula_1_still_drive_to_survive_take_a_look/). The dataset came from (http://ergast.com/mrd/db/#csv). The data was compiled by Ergast Developer API, who provides a historical record of motor racing data for non-commercial purposes. The data is free for me to use. Additionally, the data is raw and has not undergone much manipulation.

```
races <- read_csv("data/races.csv")
```

```
## Parsed with column specification:
## cols(
##   raceId = col_double(),
##   year = col_double(),
##   round = col_double(),
##   circuitId = col_double(),
##   name = col_character(),
##   date = col_date(format = ""),
##   time = col_character(),
##   url = col_character()
## )
```

```
status <- read_csv("data/status.csv")
```

```
## Parsed with column specification:
## cols(
##   statusId = col_double(),
##   status = col_character()
## )
```

```
results <- read_csv("data/results.csv", col_types = cols(number = col_character()))
```

I used 3 csv files from the raw datasets. Each row represents an entry. There are 1035 rows in "races," 136 rows in status, and 24840 rows in results. The data is diverse, but mainly numeric, regressional data.

Wrangling

I need to join the 3 datasets together using their keys. Specifically, I need the year and the statuses in one dataset.

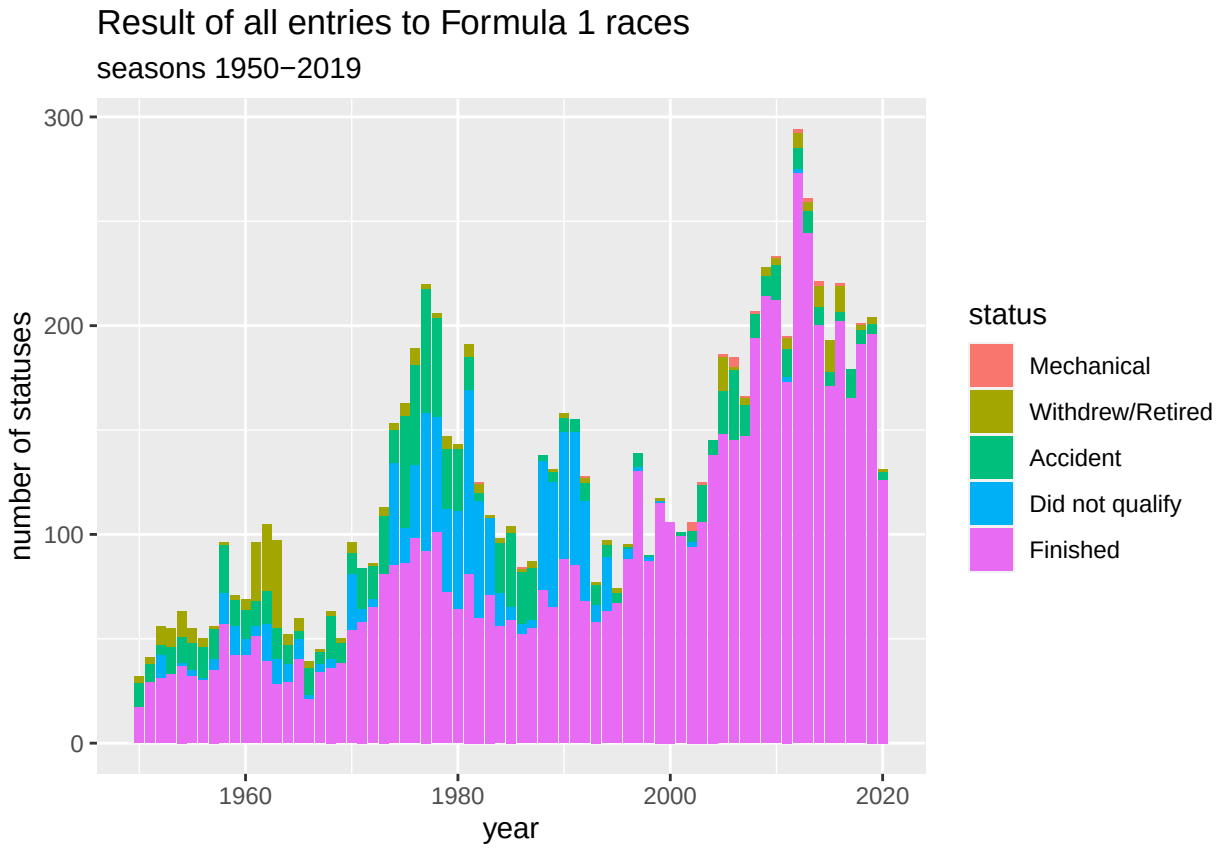
```
joined_results <- left_join(status, results, by='statusId') %>%
                  left_join(races, results, by='raceId')
```

Replication

```
joined_results %>%
  mutate(status = if_else((status == "Withdrew") |
                          (status == "Retired"),
                          "Withdrew/Retired",
                          status)) %>%                  # Mutate new a new status category
  filter(status == "Withdrew/Retired" |
         status == "Accident" |
         status == "Did not qualify" |
         status == "Mechanical" |
```

```
        status == "Finished") %>%                          # Filter 5 categories needed
  group_by(status, year) %>%
  summarise(status_sum = n()) %>%
  ggplot(aes(x = year,
             y = status_sum,
             fill = reorder(status, status_sum))) +     # Create stacked bar graph
  geom_col() +
  labs(x = "year",
       y = "number of statuses",
       fill = "status",
       title = "Result of all entries to Formula 1 races",
       subtitle = "seasons 1950-2019")                  # Give label
```

```
## `summarise()` regrouping output by 'status' (override with `.groups` argument)
```



Analysis

The graph above demonstrates the count of the statuses of the entries by year, categorized by 5 status results.

A difficulty I encountered was putting the cumulative count of the entries by status on the right side of the graph (similar to the original visualization). I did not overcome this challenge. However, a challenge that I did overcome was turning the graph from a group of separate columns into a stacked bar plot.

Alternative designs

Alternatively, I could make different design choices when I visualize the Formula 1 data.

First, the original data chose to display the five status results of mechanical, withdrew/retired, accident, did not qualify, and finished. An alternative choice to demonstrate the results of the races could be finished, mechanical, and accident. This alternative choice is superior to the original becuase it better suits the original claim. The original claim is that "F1 journey to current safety standards was not an easy one." Withdrew/retired and non-qualifying entries have nothing to do with entries that ended in unsafe ways. They manipulate the data to make the data compared to the safe "Finished" variable look larger (in an attempt to prove their point). This redesign can support the new claim that Formula 1 accidents and mechanical failures have decreased over the years.

Second, the original data chose to not visually break up the stacked bars by year. Instead, the number of statuses over the years is smoothed together. Alternatively, I would keep white space that breaks the data up by year. It supports their claim due to its demonstration of results over the course of years. This redesign can also support the new claim that Formula 1 accidents and mechanical failures have decreased over the years.

Summary