Formula 1 Racing Data Integrity Analysis

Brea Koenes

01/04/2020

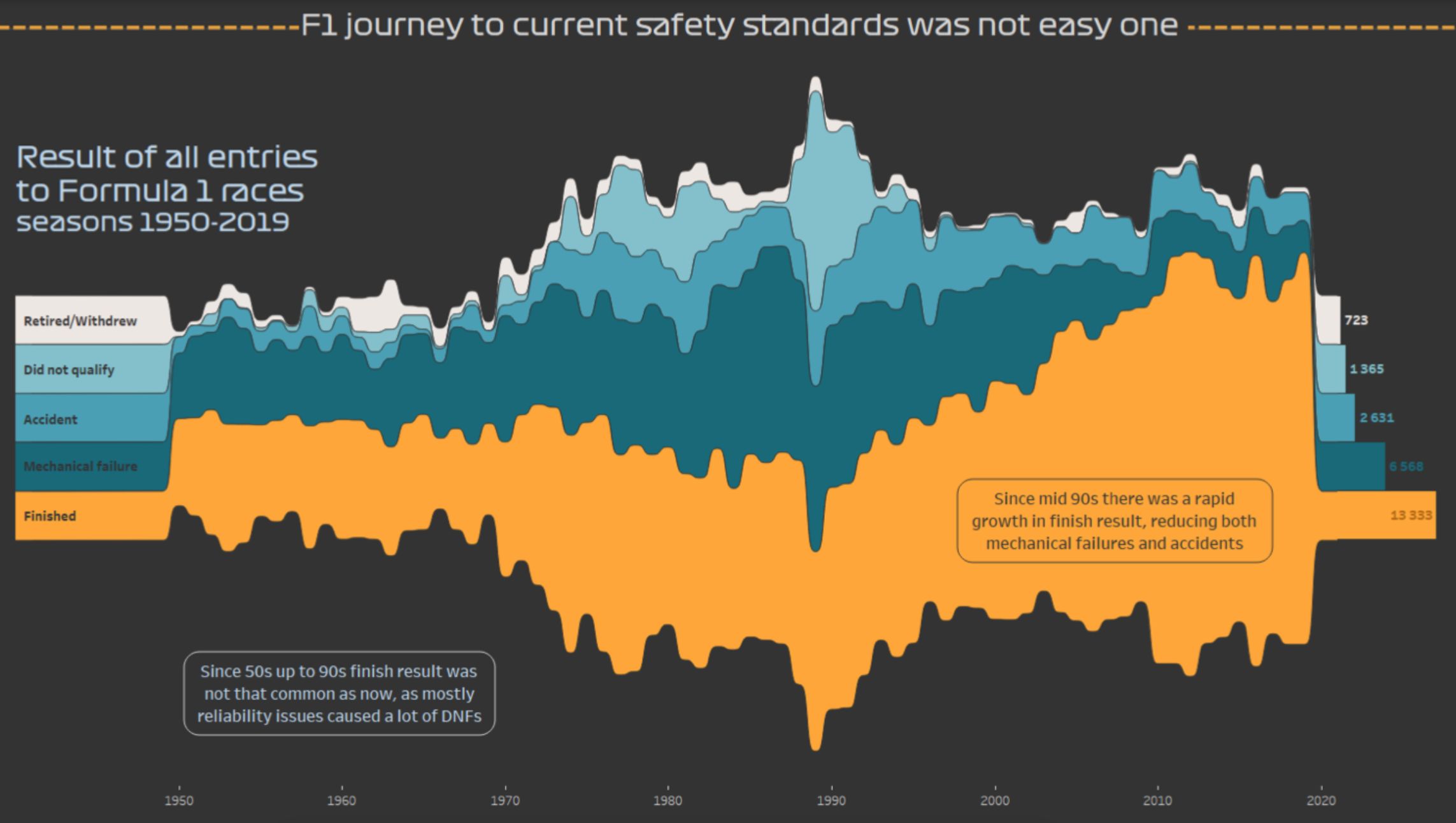
## Overview

This project is built upon the data science principle of honesty. I began by finding a trending article about Formula 1 races. I found the raw dataset corresponding to the article and recreated their analysis based on their claim—then corrected their wrangling and visualization.

“F1 journey to current safety standards was not an easy one” is the claim their graph made. They worked with numerical and categorical data. Specifically, the variables are the year, result of all entries, and the sum of the entries’ results. The graph also shows the count of the statuses of the entries by year, categorized by 5 status results.

My recreation of their analysis demonstrates that the amount of racing accidents is relatively unchanging throughout the years, unlike the claim their graph displays.

## The original visualization from a popular article



## 

## 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 manipulation.

races <- read\_csv("data/races.csv")

##   
## ── 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")

##   
## ── 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. Specifically, the data that I use is the status (categorical), sum of statuses (continuous), and years (continuous).

## Wrangling

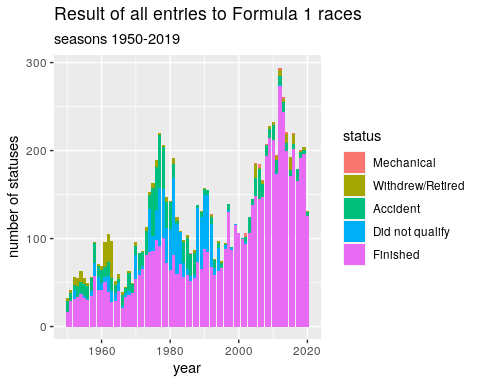
I joined the 3 datasets together using their keys. I did so due to needing the year and the statuses in one dataset.

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



## Analysis

The graph above demonstrates the count of the statuses of the entries by year, categorized by 5 status results. Specifically, the number of statuses controls the height of the graph, while the years control the width. The statuses divide the data by color in the stacked bar chart.

This is a basic replication of the original graph using their same varibles and wrangling.

## 

## Alternative designs in recreating their graph

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 only using the variables finished, mechanical, and accident. This alternative choice is superior to the original because it better suits the original claim 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 choice supports a more unbiased graph.

Second, the original visualization chose to not make each year’s number of statuses distinct from each other. Instead, the number of statuses over the years is smoothed together. Alternatively, I would make each year’s number of statuses visually distinct by using a scatter plot. It supports the original claim better due to its emphasis on seeing the results over of years.

## 

## Recreation using alternative designs

status\_data <- joined\_results %>%  
 mutate(status = if\_else((status == "Accident") |  
 (status == "Mechanical"),  
 "Accident/Mechanical",  
 status)) %>% # Mutate new a new category  
 filter(status == "Accident/Mechanical" |  
 status == "Finished") %>% # Filter 2 categories  
 group\_by(status, year) %>%   
 summarise(status\_sum = n())

## `summarise()` has grouped output by 'status'. You can override using the `.groups` argument.

last\_counts <- status\_data %>%  
 group\_by(status) %>%  
 slice\_tail(n = 5) %>%   
 summarise(year = 2020, status\_sum = mean(status\_sum)) # Summarize last 5 years of statuses for the legend  
   
status\_data %>%   
 ggplot(aes(x = year,   
 y = status\_sum,   
 color = reorder(status, status\_sum))) + # Create scatter plot  
 geom\_point() +  
 geom\_smooth() +  
 geom\_text(data = last\_counts,   
 aes(label = status),  
 direction = "y",  
 nudge\_x = -5,  
 nudge\_y = -20) + # Reformat legend  
 theme(legend.position = "none") +  
 labs(x = "year",   
 y = "number of status entries",   
 title = "Finished vs. Failure Results of Formula 1 Race Entries",   
 subtitle = "from 1950-2019") # Give label

Chart, line chart

Description automatically generated

## Summary

After recreating their graph with wrangling and a visual that more accurately explored their claim, my understanding of the original claim has changed. The original claim is that the “F1 journey to current safety standards was not an easy one.” After plotting the data with variables only relevant to the claim, I found that the rough start of Formula 1 races their claim implied was not as bad as the original graph displayed. Yes, there were more accidents and failures in races in the beginning years. However, without withdrawn, retired, and non qualifying entries included in the graph, Formula 1 did not have as rough of a start as it appeared.

I was faithful to the original graph’s purpose of displaying entry’s statuses over time. The variables are the same, but I adjusted which statuses I included. I also changed the type of graph from a stacked bar chart to a scatter chart to better understand the trends of the statuses. Furthermore, I chose to keep the number of finished races to have in reference to the accidents/mechanical failures.

I would say that my alternative design was very successful. It is easy to interpret, unlike the previous graph, and the graph is not manipulated to provoke an exaggerated conclusion. Specifically, I combined accidents and mechanical issues to make the distinction between failures and finished entries easier to see.

A follow-up question is how has the percentage of failures compared to finished entries change over time (instead of comparing the sums)? Also, are entries that have mechanical failures that get fixed during the race counted as finished and mechanical failure? Or, if they finish with a mechanical failure during the race are they counted only as finished?

These questions will be explored in a future analysis, along with a predictive model for the ratio between accidents and finished races in 2022.