**Formula 1 Driver Analyser**A visual analysis tool for inspecting F1 drivers’ performances  
  
Luca Sannino 1542194  
Sapienza University of Rome

*Abstract*

**In this relation I describe a web application that uses the D3.js library to create a visual environment where a user can choose a Formula 1 driver (from 1994 to present day), so that they can learn more about their performance and place in the sport’s history. This is achieved by interacting with visualizations that focus on the aspects of a driver’s career that are of particular interest, like points achieved, results against their teammates, DNFs and consistency of their placements. Such visualizations, that are all synchronized between each other, include bar graph, lollipop charts, box plots and scatter plots on multidimensional data achieved through PCA.**

**1. Introduction**

At any given time, about only two dozens of people in the entire world can say to be Formula 1 drivers. Given the lack of seats, not only they need to drive as best as they can to climb up the standings, but they also need to prove that they are not stealing the place of far more skilled drivers.

As such, their performance has always been a topic of study and discussion, both internally in the teams themselves who are always looking for the next best hire, and externally across passionate fans, including people looking to make some quick money through bets. And since their performance is based on stats, it means it is a topic of interest for Visual Analytics as well.

This environment is then meant to help users (like those described above) gain a better knowledge of the drivers they are interested in, by presenting interactive visualizations that try to answer questions such as:

● Is the driver on an upward trajectory, or are his best days clearly behind him?  
● How does his performance compare against the ones of his teammates across his career?  
● Does he gets DNF’d often? If so, it’s because he is a hothead who constantly gets involved in crashes/collisions, or it’s due to unlucky reliability issues?

● How constant he is in his placements? In which constructor did he gain the better results? In which circuits?

● How does he fit in the history of the sport itself, in terms of achievements reached?

This relation will go in detail in the related works, the dataset that made all of this possible, the visualizations themselves, how the user can interact with them, and will conclude with some final thoughts about its limitations.

**2. Related Work**

This environment is partially inspired by the work presented by Lampprecht et al. [1].  
While their visualization was “race-focused”,  
meaning that a user could learn about the drivers’ performances in relation to a given selected race, this work is instead meant to offer a “driver-focused” prospective, meaning that users get to choose drivers in order to gain a better knowledge about their performances across the entire stretch of their careers. The work proposed here is then meant as companion piece, since both tools together can now allow analyses that can be both focused on more narrow details, or that can look at the bigger picture.

**3. The Dataset**

The dataset at the basis of this project was found on Kaggle, courtesy of Rohan Rao [2]: it consists of a series of .csv files that cover a lot of interesting stats since the beginnings of Formula 1 until present day. In particular, the ones that were important for this environment were the files that dealt with the results of races and qualifying sessions, which also contained information about points obtained and the reason of a DNF, if there was any.

However, the dataset was fractured and had to be integrated before it could be effectively used: in order to do so, I used Talend Open Studio, which is a tool that allows the user to implement GAV mappings [Figure 1] in a intuitive way.

At the end of the process [Figure 2] I had obtained the following files: results.csv (that deals with race results), qualifying.csv (that deals with qualifying results) and standings.cvs (that deals with the drivers’ standings after each race). I’ll now describe just results.csv, since the other two file have only small differences in their structure:

● **year**: Season in which the race was set;

● **round**: Round of the race in the season;

● **grandprix**: Official name of the race;

● **circuit**: Name of the circuit were the race was set in;

● **surname**: Surname of the driver;

● **name**: Name of the driver;

● **constructor**: Constructor the driver was   
driving for in that race;

● **position**: Position achieved at the finish line;

● **points**: Points achieved with the position;

● **status**: A comment that explain the reason of DNF, if there was any.

**Source Schemas**

results(…, raceId, driverId, constructorId, …, position, …, points, …, statusId)  
races(raceId, year, round, circuitId, name, ..)  
circuits(circuitId, …, name, …)  
driver(driverId, …, forename, surname, …)  
constructors(constructordId, …, name, …)  
status(statusId, status)

**Global Schema**

results(year, round, grandprix, circuit, surname, name, constructor, position, points, status)

**GAV Mapping**

{ (a, b, c, d, e, f, g, h, i, l) |

results(…, x1, x2, x3, …, h, .., I, …, x4),  
races(x1, a, b, x5, c, …),  
circuits(x5, …, d),

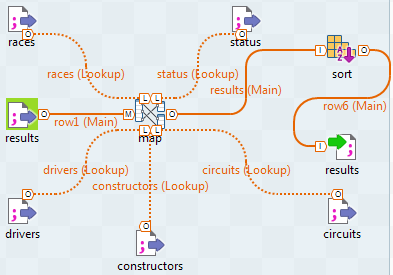
driver(x2, …, e, f, …),

constructors(x3, .., g, …),

status(x4, l)

} => results(a, b, c, d, e, f, g, h, i, l)

*Figure 1: GAV mapping for the file results.csv*

**

*Figure 2: Implementing GAV mapping in Talend in order to create results.csv*