# AP2 — Formula 1 Driver Clustering Report

**Murmani Akhaladze**

TarnNished

## 1. Model and Approach

In this project I used **K-Means clustering** to group Formula 1 drivers based on their performance metrics. The idea was to treat every driver as a data point in a multidimensional space, where each dimension represents a performance indicator such as win rate, podium rate, pole rate, fastest lap rate, and points per race entry.

Before applying the model, I standardized the data using `StandardScaler` from scikit-learn, because features had very different scales (for example points can be in thousands, while win rate is less than 0.5). Scaling converts everything into comparable vector magnitudes using the Euclidean norm. Then, K-Means minimizes the total squared distance (L2 norm) between each driver vector and its cluster center.

The math behind it relates to vector and matrix norms: the algorithm repeatedly computes distances between driver vectors and centroid vectors, essentially applying matrix norm operations until convergence.

## 2. Experiment

The dataset came from Kaggle and includes over 800 drivers from all Formula 1 history. I selected quantitative columns:

- **Win_Rate** = race wins / starts
- **Podium_Rate** = podium finishes / starts
- **FastLap_Rate** = fastest laps / starts
- **Pole_Rate** = pole positions / starts
- **Points_Per_Entry** = total points / race entries

I tested cluster numbers ($k = 2$ to 8) using a Streamlit slider and found that around $k = 4$ gave the most meaningful results. The clusters were clearly visible in both 2D and 3D scatter plots, and the radar charts helped compare their average performance profiles.

The app was built using **Streamlit** and **Plotly**, allowing interactive exploration where users can change cluster numbers and instantly see how drivers move between groups.

The experiment showed that statistical data alone can meaningfully group drivers into categories. For example:

- One cluster contained top-tier champions like Schumacher and Hamilton (frequent winners).
- Another cluster grouped mid-level drivers with some podiums.
- The last cluster represented drivers who barely scored points.

K-Means worked effectively because Formula 1 performance data naturally separates by scale. The normalization step ensured fair distance comparison and accurate clustering.

Overall, this project helped me understand the connection between clustering, vector norms, and matrix operations. I also enjoyed that it can be visualized in 3D, which makes explaining the results to others much easier.