Data-driven Insight into Formula 1 Racing: EDA and Visualization using Python and Tableau

*Note: All data used in this project is sourced from publicly available
Formula 1 datasets on the Ergast API, including driver statistics, race results, and constructor data.

Bin (Smars) Hu - 0843954
Trent University
smarshu@trentu.ca

***Abstract*—This project investigates historical Formula 1 racing data through exploratory data analysis (EDA) and visualization. The core objective is to uncover performance trends, team dominance, and nationality patterns using Tableau for interactive dashboards and Python (Matplotlib, Seaborn) for statistical visualizations. While a cloud-native data pipeline on Azure (Delta Lake, Databricks) was implemented to support data quality and automation, the analytical focus remains on insight generation. Key findings include a strong correlation between total points and championship rankings, evidence of grid position influencing final results, and regional dominance in driver contributions. The combined use of Tableau's BI capabilities and Python's flexibility enables both intuitive exploration and deeper hypothesis testing, demonstrating the power of visual analytics in sports data contexts. This approach is scalable and can be extended to other sports domains.**

***Keywords*—Formula 1 Racing; Exploratory Data Analysis; Data Visualization; Tableau; Python; Matplotlib; Seaborn; Databricks; Azure; Interactive Dashboard; Multivariate Visualization**

I. Introduction

In recent years, data-driven methodologies have become increasingly prevalent in the analysis of professional sports, enabling analysts to uncover deep patterns, trends, and strategic insights from historical datasets. Formula 1 (F1), one of the most data-intensive [1] and globally followed motorsports, offers a rich and structured dataset that spans decades of races, drivers, teams, and countries. The inherently competitive and quantifiable nature of F1 racing—defined by measurable variables such as lap times, grid positions, finishing positions, and championship points—makes it an ideal domain for exploratory data analysis (EDA) and data visualization.

This project focuses on leveraging the principles of **EDA and data visualization** to analyze and interpret Formula 1 historical data. Our aim is not to build predictive models or deploy machine learning algorithms, but rather to investigate and communicate the underlying patterns, relationships, and trends that emerge from the data through visual exploration. Key research interests include understanding how starting positions affect race outcomes, identifying consistent high-performing drivers, evaluating the influence of fastest laps on total points, and comparing national-level statistics in terms of driver counts and accumulated scores.

To support this visual and statistical investigation, we employed two complementary tools: **Tableau** and **Python**. Tableau is used to create a rich, **interactive BI dashboard** [2], allowing users to dynamically explore metrics such as cumulative team scores, driver performance over time, and country-based aggregations through bar charts, line graphs, and geographic maps. Meanwhile, **Python**—specifically the Matplotlib and Seaborn libraries [3] —is employed for more granular, statistical visualizations. This includes bubble plots, heatmaps, and pairwise plots used to analyze multidimensional relationships such as grid position versus final rank or total points versus championship titles.

While the core of the project centers on EDA and visualization, a **supporting data processing pipeline** was implemented to ensure clean, well-structured, and reliable data sources for analysis. To achieve this, we designed and deployed a cloud-native **data architecture using Azure**. The pipeline uses Azure Data Lake Storage Gen2 for storing raw and structured data, Azure Databricks for distributed ETL tasks powered by PySpark and SparkSQL, and follows the medallion architecture (bronze-silver-gold layers) to refine data through multiple stages [4]. Though this component is not the primary focus, it plays a critical role in preparing high-quality, analysis-ready data for the EDA layer.

In summary, this project exemplifies how modern data visualization techniques—backed by reliable data engineering—can be used to generate insights into Formula 1 performance data. The objective is to provide a comprehensive yet intuitive exploration of the sport through visual storytelling, focusing on team and driver dynamics, race outcomes, and cross-national patterns, all powered by an enterprise-grade but lightweight data infrastructure.

II. Previous Work

A. EDA in Sports Analytics

The application of exploratory data analysis (EDA) has become an integral part of modern sports analytics. Numerous studies have applied EDA techniques to uncover patterns in athletic performance, team strategy, and competition outcomes [5]. In the context of Formula 1, prior analyses have explored relationships between **starting grid positions and final results** [6] revealing how qualifying performance can influence race outcomes. Other investigations have focused on **year-over-year championship trends**, helping identify periods of team dominance or shifts in competitive balance. Additionally, **driver consistency metrics** [7] such as average finishing position, standard deviation, and podium frequency—have been used to quantify performance stability over multiple seasons.

Beyond Formula 1, EDA has been widely used in analyzing other sports such as basketball, soccer, and tennis. In these domains, researchers have used data visualization and statistical summaries to study **player efficiency, match dynamics, and win-loss patterns**. These approaches have proven effective in transforming raw performance data into actionable insights for coaches, analysts, and fans alike. They form the analytical foundation that this project builds upon in the context of Formula 1.

B. *Visualization Tools in EDA: Tableau and Python*

Among the tools commonly used for EDA, **Tableau** and **Python** are particularly prominent. Tableau's drag-and-drop interface and dashboarding capabilities make it ideal for presenting business intelligence in an interactive and digestible format. It supports various visualization types—from bar charts to heatmaps to geographic maps—and is widely adopted in industries ranging from finance to sports. In the context of sports analytics, Tableau has been effectively used to build dashboards that track **team performance, player statistics, and seasonal comparisons.**

Python, on the other hand, offers greater flexibility and statistical depth. Libraries such as **Matplotlib** and **Seaborn**, enable analysts to generate highly customized visualizations, perform variable correlation analysis, and explore complex multidimensional relationships. In sports analytics, Python has been employed to detect **outliers, trends, and clusters** in performance data, often enabling more nuanced exploration than GUI-based tools alone.

C. *Positioning This Project*

While previous works have either focused on single-tool analysis or isolated datasets, **this project distinguishes itself by combining multi-tool EDA practices with a robust cloud-native data pipeline**. The backend infrastructure follows a **Lakehouse architecture**—commonly used in modern enterprise environments [8]—to ensure reliable data ingestion, transformation, and delivery to the analytical layer. Although the engineering component is not the focus of this paper, it plays a supporting role in ensuring the consistency and freshness of data used in the visual exploration.

In summary, this project draws upon established EDA methodologies in sports analytics, leverages widely accepted tools like Tableau and Python, and integrates them into a scalable data workflow—offering both analytical richness and practical reproducibility that enhances its real-world applicability.

III. METHODOLOGY

To uncover meaningful insights from Formula 1 historical data, a structured methodology was adopted that integrates scalable data engineering with multi-layered data analysis. This approach ensures data consistency, quality, and analytical depth throughout the project. The methodology consists of two main components: first, the construction of a cloud-native data processing pipeline based on the medallion architecture to prepare clean, analysis-ready datasets; and second, the application of advanced visualization and exploratory techniques using Tableau and Python to extract patterns and relationships across drivers, teams, and countries. The following subsections outline each component in detail.

A. *Data Preparation: A Medallion Architecture Pipeline*

To support reliable and high-quality exploratory data analysis, we designed a modular data processing pipeline based on Azure's cloud-native ecosystem. Figure 1 below illustrates the end-to-end data flow from raw ingestion to visualization-ready datasets.

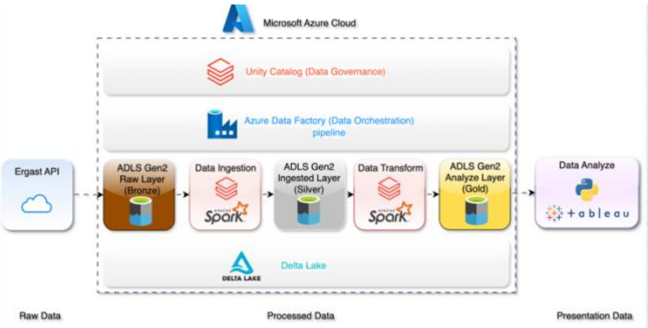

Microsoft Azure Cloud
Unity Catalog (Data Governance)
Azure Data Factory (Data Orchestration) pipeline
Ergast API
ADLS Gen2 Raw Layer (Bronze)
Data Ingestion
Spark
ADLS Gen2 Ingested Layer (Silver)
Data Transform
Spark
ADLS Gen2 Analyze Layer (Gold)
Data Analyze
+ a b | e a u
Delta Lake
DELTA LAKE
Raw Data
Processed Data
Presentation Data

Fig. 1. Tech Architecture. (*Microsoft Azure Cloud Native Pipeline*)

The pipeline begins with the ingestion of raw Formula 1 datasets—sourced from the Ergast API. We could download raw data as JSON or CSV files—into the Bronze layer of Azure Data Lake Storage Gen2 (ADLS Gen2). This layer stores the unprocessed data in its native format. Using Apache Spark on Azure Databricks, data is then refined and moved to the Silver layer, where schema alignment, and initial cleaning are performed. This intermediate stage prepares the data for more structured and analytical use.

Next, in the Gold layer, the data is further aggregated and modeled into presentation-friendly formats. This includes calculated fields such as total points, which are directly relevant for visual analytics. This layer serves as the foundation for downstream analysis tools.

The entire process is governed using Unity Catalog for data access control and metadata management, while Azure Data Factory orchestrates the ETL workflows automatically, ensuring that new data can be processed and refreshed regularly.

Ultimately, the Gold layer feeds directly into visualization tools like Tableau and Python-based analysis notebooks. This architecture ensures that the visual layer operates on clean, consistent, and up-to-date data, which is critical for producing trustworthy EDA results.

In summary, this architecture provides a structured data flow from ingestion to insight, enabling seamless integration between cloud-scale data processing and interactive data visualization..

B. *Data Analysis & Visualization*

The data analysis and visualization phase of this project consists of two complementary components: (1) the construction of an interactive Business Intelligence (BI) dashboard using Tableau, and (2) exploratory data analysis (EDA) using Python. This dual approach enables both high-level pattern discovery and deeper statistical insights from the Formula 1 dataset.

1) *Tableau Dashboard Design*

Tableau was chosen as the primary BI tool to deliver intuitive, user-friendly visualizations for aggregated, two-dimensional analyses. A key advantage of Tableau is its seamless integration with Azure Databricks, which allows direct connectivity to the curated data in the Gold layer and supports live updates. This ensures that the dashboard reflects the most recent transformations performed in the data pipeline.

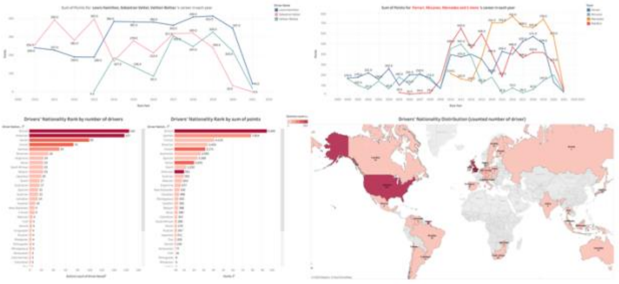


Fig. 2. Tableau Dashboard Screenshot (Partial)

The dashboard design centers around three core analytical dimensions in Formula 1: Constructors (Teams), Drivers, and Nationalities, supported by metrics such as total points, championship wins, and performance over time. The primary chart types employed include:

- **Bar charts** to show cumulative points and championship counts for teams, drivers, and countries;
- **Line charts** to visualize longitudinal performance trends of top drivers or constructors;
- **Maps** with geographical overlays to display the spatial distribution of driver nationalities and comparative rankings across countries;
- **Combined charts** to capture temporal championship dynamics (e.g., shifts in dominance across decades).

The Tableau dashboard is powered by the **f1_presentation** database from the Gold layer. Auto-refresh mechanisms were configured to ensure that dashboard visuals are synchronized with backend data updates. All dashboards are fully interactive—users can filter by century, year range, country, or driver name, enabling a flexible exploration of F1 history across different lenses.

2) Python-Based Statistical Exploration

To complement Tableau's aggregated visual summaries, we employed Python to conduct more detailed, multivariate EDA. The analysis was implemented in Jupyter Notebook, leveraging pandas for data wrangling and matplotlib/seaborn for statistical visualization.

Key analytical workflows include:

- **Grid position vs. final race result:** scatterplots and heatmaps were used to investigate the correlation between starting position and final ranking;
- **Fastest lap time vs. total points:** assessing whether race pace significantly contributes to season-long success;
- **Driver consistency tracking:** time-series line plots revealed the point trajectory of top drivers over multiple seasons;
- **Country- and constructor-level dominance:** bar charts and geographic maps summarized championship frequency and point accumulation by country and team;
- **Multivariate relationship analysis:** bubble plots were constructed to explore interactions between variables such as wins, total points, and championship rank.

Throughout the process, careful attention was paid to data preprocessing, feature selection, and transformation logic to ensure that visualizations were both accurate and meaningful. The full Python codebase was documented and exported in HTML format for transparency and reproducibility.

Overall, the Tableau dashboard offered a comprehensive and interactive overview for end users, while the Python-based analysis provided depth for more technical and research-oriented questions. Both components are grounded in the structured outputs of the Gold layer, ensuring consistency and analytical integrity across all visualizations.

IV. Results

This section presents key analytical findings derived from both Tableau dashboards and Python-based visualizations. By exploring multi-dimensional patterns in driver performance, team dominance, and nationality trends, we uncover hidden insights from historical Formula 1 data through a comprehensive exploratory data analysis (EDA) approach.

A. EDA based on the Tableau BI Dashboard

The interactive Tableau BI dashboard is public and could be accessed by the link: (https://public.tableau.com/app/profile/smars.hu/viz/eda_visualization_on_formula1_racing/Dashboard1)

1) Team and Driver Performance Across Eras

A series of bar charts present the cumulative point totals for Formula 1 teams and drivers, segmented by historical periods including all-time, the 21st century, 2000–2010, and 2011–2021. From a team perspective, Ferrari consistently dominates across all eras, with Mercedes and McLaren also emerging as persistent high performers. Notably, Mercedes exhibits a sharp rise in dominance post-2010, reflecting the hybrid engine era where its engineering superiority significantly influenced race outcomes.

In the driver dimension, Lewis Hamilton ranks highest in all-time and 2011–2021 segments, followed by Sebastian Vettel and Fernando Alonso. A closer examination reveals the evolution of driver dominance over time: Michael Schumacher's peak in the early 2000s is later replaced by Hamilton's prolonged ascendancy. This temporal segmentation enables EDA to identify era-specific shifts in competitive balance, reinforcing the importance of period-based slicing in longitudinal sports data analysis.

2) Nationality-Based Aggregations

Several visualizations explore the nationality dimension through three lenses: driver count, cumulative points, and championship counts. The United Kingdom leads across all categories, contributing the highest number of drivers (165), the most cumulative points (9,449), and the most championship titles (22), underscoring its historical centrality in the F1 ecosystem. Germany, Brazil, and Finland follow, with Germany demonstrating strong point accumulation despite fewer contributing drivers, suggesting greater average performance per driver.

The spatial heatmaps further emphasize Europe's dominance, particularly Western and Northern Europe. However, non-European contributions from countries such as Brazil, the United States, and Australia indicate broader global engagement. These findings support the hypothesis that while F1 has a global footprint, its competitive core remains concentrated in specific geographic regions—a key pattern discoverable only through multi-dimensional EDA.

3) Performance Trends by Year

Line charts comparing yearly point totals for top drivers (e.g., Lewis Hamilton, Sebastian Vettel, Valtteri Bottas) and constructors (Ferrari, McLaren, Mercedes, Red Bull) provide a dynamic view of performance evolution. These visualizations enable temporal trend discovery, a key strength of EDA.

Hamilton's trajectory from 2009 to 2021 displays remarkable consistency, with point totals often exceeding 350 per season—highlighting his dominance and team stability. Vettel’s career, by contrast, features a sharp peak during 2010–2013, corresponding to his Red Bull championship run, followed by a noticeable decline. Bottas’s emergence post-2014 reflects Mercedes’ strategic rotation of supporting drivers during their peak.

On the team side, Mercedes exhibits a steep rise in point accumulation from 2014 onwards, reflecting their hybrid engine era supremacy. Ferrari shows a more fluctuating performance, with occasional recoveries but lacking the sustained dominance of previous decades. This section demonstrates how year-over-year comparisons reveal shifts in team strategy, technological advantage, and driver synergy—insights only accessible through interactive longitudinal EDA.

4) Constructor and Championship Correlations

The bar charts examining the number of championships by constructors over different eras (all-time, 20th century, 21st century) provide critical context for evaluating engineering consistency and institutional success. Ferrari leads overall with 21 constructors’ titles, yet Mercedes emerges as the dominant force in the 21st century with 8 titles, closely followed by Red Bull and Ferrari.

This distinction between all-time versus recent success illustrates the concept of temporal performance clusters—where certain teams experience golden eras due to innovation cycles, regulation changes, or exceptional driver-engineer pairings. The ability to correlate these with historical context (e.g., rule changes in 2014) exemplifies how EDA supports interpretive storytelling rather than static reporting.

Moreover, constructor dominance often aligns with driver championships, as seen with Mercedes and Hamilton or Red Bull and Vettel, emphasizing the interdependence between mechanical and human performance in motorsport success.

B. Statistical Exploration with Python: Patterns Behind Performance

In addition to the Tableau dashboards, Python was used to conduct fine-grained statistical analysis using libraries such as Matplotlib and Seaborn. This allowed deeper examination of race dynamics, driver consistency, and long-term performance evolution. The Python-based EDA offers interpretability through visual exploration, pattern recognition, and hypothesis generation.

1) Grid Start vs Final Position

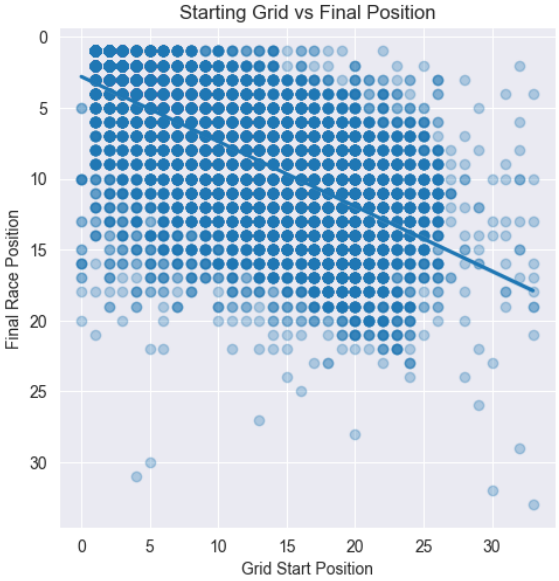


Fig. 3. linear regression plot between grid position and final race position

The linear regression plot between grid position and final race position shows a clear positive correlation, with most data points concentrated within positions 1–25. Drivers starting at the front of the grid tend to finish higher, as reflected by a downward-sloping trend (inverted y-axis). However, a cluster of outliers—drivers who started poorly but finished strongly—indicates that race-day conditions, driver skill, or strategic interventions can sometimes overturn initial disadvantages. This reinforces a core EDA insight: while correlation provides general rules, outliers often reveal exceptional narratives.

2) Fastest Lap vs Points Earned

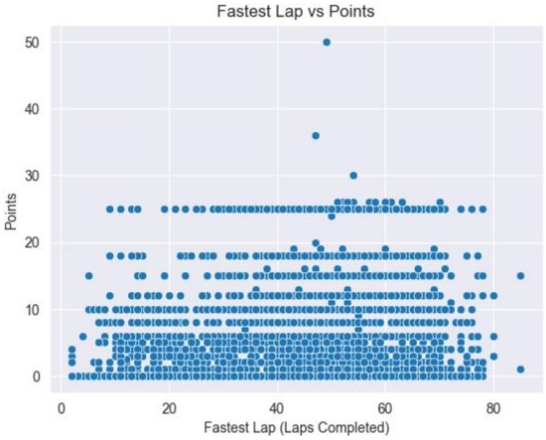


Fig. 4. Scatterplot of faster lap timing and total race points

Contrary to intuition, the scatterplot of fastest lap timing versus total race points reveals no discernible trend. The lack of linear or nonlinear pattern suggests that setting a fastest lap—regardless of when it occurs—does not significantly impact total point accumulation. This underlines an important EDA discovery: not all measurable race variables hold predictive value for outcomes, highlighting the need to empirically validate assumptions.

3) Driver Consistency and Ranking Distribution

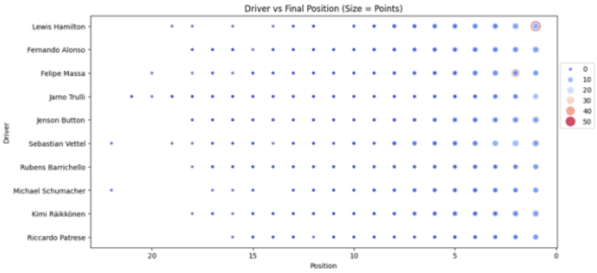
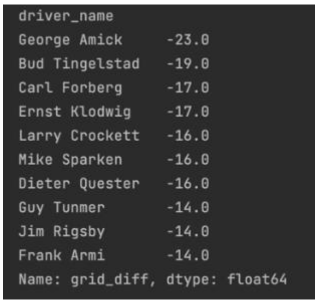


Fig. 5. Bubble plots of drivers and race positions and points

Bubble plots mapping drivers to their race positions and points reveal performance clustering. Lewis Hamilton's dominance is visually evident, with large bubbles clustered in positions 1–3, indicating frequent high finishes and point accumulations. In contrast, drivers like Fernando Alonso and Felipe Massa display broader spreads, representing varied performance across seasons. This visualization aids in identifying driver archetypes—dominant, consistent, or volatile—an essential outcome of multi-variable EDA.



```
driver_name
George Amick      -23.0
Bud Tingelstad    -19.0
Carl Forberg      -17.0
Ernst Klodwig     -17.0
Larry Crockett    -16.0
Mike Sparken      -16.0
Dieter Quester    -16.0
Guy Tunmer        -14.0
Jim Rigsby        -14.0
Frank Armi        -14.0
Name: grid_diff, dtype: float64
```

Fig. 6. Underperforming drivers analysis

A separate grid differential analysis further identifies underperformers by comparing average positions gained or lost during races. For example, George Amick exhibits an average loss of 23 positions—one of the steepest declines—indicating systematic underperformance. Such diagnostics offer a quantitative basis for flagging struggling drivers and raise hypotheses about team support, vehicle reliability, or driver skill.

4) *Points vs Ranks: Multivariate Analysis*

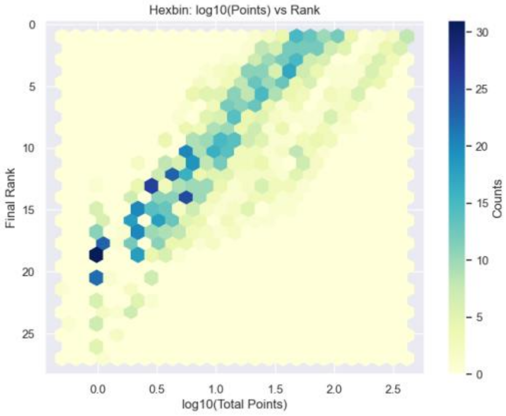


Fig. 7. Hexbin plots of points and rank

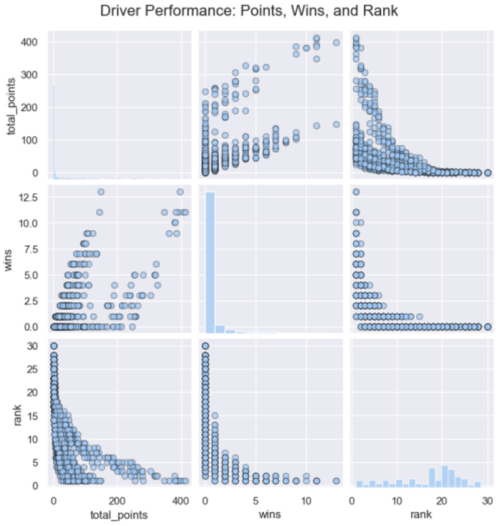


Fig. 8. Pair Plots of Points, Wins and Rank

Through hexbin and pair plots, relationships between rank, wins, and total points are explored. Log-transformed plots highlight a nonlinear concentration of mid-field drivers around low point ranges (10–30), while top-ranked drivers cluster around 100+ points. Despite strong correlation between points and wins, anomalies exist: several drivers secure high rankings with few wins, indicating that consistent top finishes, even without victories, can ensure competitive standing. This nuanced result emphasizes the role of consistency over occasional brilliance.

5) *Champions' Points Over Time*

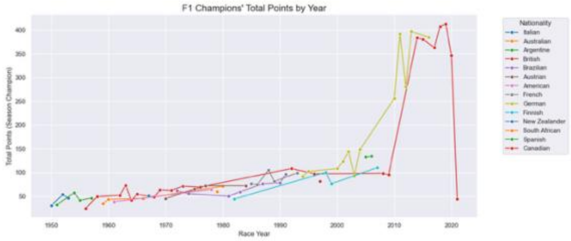


Fig. 9. Time-seris line chart of champions' total points by year

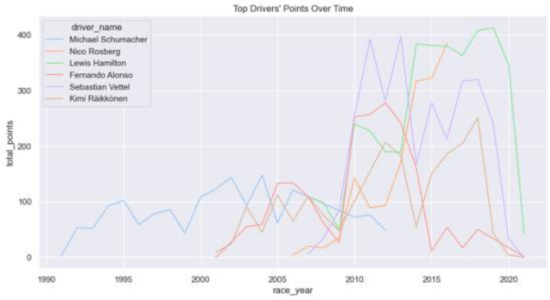


Fig. 10. Time-seris line chart of Champions' Points

A time-series line chart of season champions reveals a rising trend in total points across decades, with notable surges post-2010. This suggests changes in race formats or scoring systems. Nationality breakdowns show eras of dominance—Germany and Finland in the 2000s, followed by a sustained British lead post-2014, led by Hamilton. Anomalies such as low-point champions in recent years reflect rule modifications or shortened seasons, providing temporal context to quantitative fluctuations.

6) *Constructor-Level Analysis*

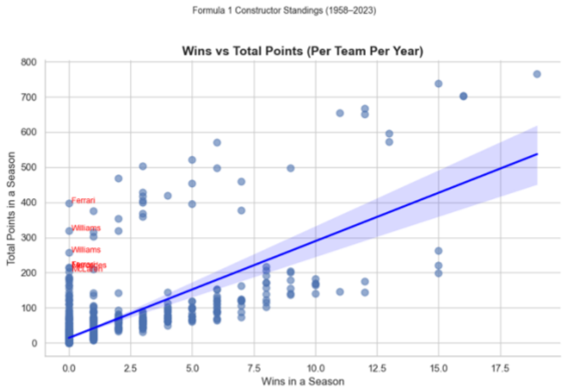


Fig. 11. Linear regression plot of wins and total points (Per Team Per Year)

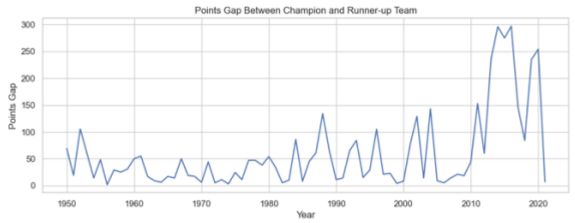


Fig. 12. Time-series line chart of point gap between champion and runner-up team

Team-based plots echo driver-level insights. A linear model of wins versus points shows a strong positive correlation, though teams like Ferrari and Williams occasionally accumulated over 200 points in winless seasons—indicating race consistency and the benefit of reliable finish rates. Additionally, the points gap between first and second constructors reveals shifting competitiveness. While the pre-2000 era was tightly contested, the hybrid era (2014–2016) saw unprecedented gaps exceeding 250 points, reflecting Mercedes' domination. Recent seasons indicate rebalance, aligning with Red Bull's resurgence.

7) *Joint Distribution of Position and Points*

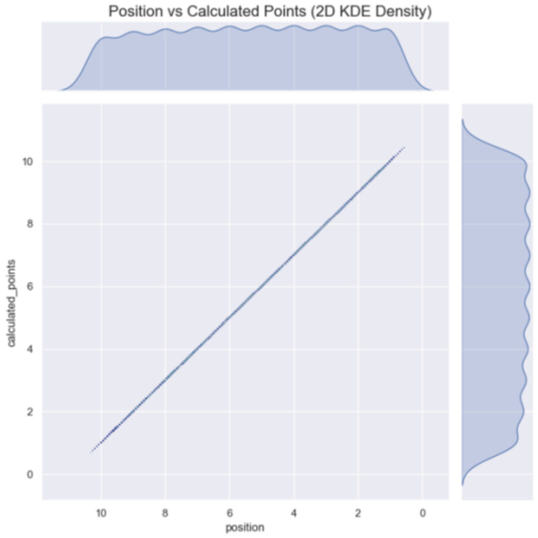


Fig. 13. 2D KDE plot of position and calculated points

Finally, a 2D KDE plot of position versus calculated points confirms the stability of F1's point distribution system. The dense diagonal ridge in the kernel density plot demonstrates that better finishing positions consistently yield higher points, validating the sport's reward structure.

Conclusions

This project demonstrates the value of combining modern cloud-native data pipelines with dual-perspective visual analytics to uncover meaningful insights in Formula 1 racing. By leveraging Azure-based medallion architecture, we ensured reliable, scalable data preparation that empowered efficient exploratory data analysis (EDA). The integration of Tableau and Python proved especially effective: Tableau enabled interactive, real-time dashboards for comparative and geographical trends, while Python offered fine-grained statistical exploration using Matplotlib and Seaborn.

Our findings confirmed several hypotheses, such as the strong correlation between grid start and final position, and the concentrated dominance of certain drivers and constructors across eras. The analysis also revealed nuanced insights—for example, fastest laps do not necessarily lead to higher points, and consistent podium finishes often outweigh occasional wins.

Despite these contributions, the project is limited to **structured historical data**, lacking telemetry, environmental, or real-time data streams. Future work could incorporate live sensor data using Kafka and Spark Streaming, or enable user interaction through web platforms such as Streamlit or Power BI Embedded. Moreover, the framework developed here is transferable to other sports domains (e.g., NBA, football), where performance trends, player dynamics, and team analytics are equally data-rich and analytically promising.

References

[1] P Akin, L. (2024). Data-Driven Vehicle Performance Optimization for Formula Student Racing (Doctoral dissertation, Politecnico di Torino). https://webthesis.biblio.polito.it/33043/

[2] Pala, S. K. (2017). Advance Analytics for Reporting and Creating Dashboards with Tools like SSIS, Visual Analytics and Tableau. International Journal of Open Publication and Exploration, 5(2), 3006-2853. https://www.researchgate.net/publication/378679002_Advance_Analytics_for_Reporting_and_Creating_Dashboards_with_Tools_like_SSIS_Visual_Analytics_and_Tableau

[3] Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of open source software, 6(60), 3021. https://joss.theoj.org/papers/10.21105/joss.03021

[4] Wiselka, M. (2024). Development of Modern Data Platform using Medallion Architecture. https://www.theseus.fi/handle/10024/869487

[5] Gudmundsson, J., & Horton, M. (2017). Spatio-Temporal Analysis of Team Sports - A Survey. ACM Computing Surveys, 50(2), 1-34. https://doi.org/10.1145/3054132

[6] Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017, August). Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1605-1613). https://dl.acm.org/doi/abs/10.1145/3097983.3098051

[7] Phillips, A. J. (2014). Uncovering Formula One driver performances from 1950 to 2013 by adjusting for team and competition effects. Journal of Quantitative Analysis in Sports, 10(2), 261-278. https://www.degruyterbrill.com/document/doi/10.1515/jqas-2013-0031/html

[8] O'Connell, R., & Allen, R. (2020). The Lakehouse Paradigm: Building the Next Generation of Data Platforms.