Hybrid Analytics of Formula 1 Radio Messages Across Circuits and Teams

Efstratios Demertzoglou | TH20580

Hellenic Mediterranean University

June 7, 2025

Introduction and Objective

Formula 1 (F1) racing stands at the intersection of advanced engineering and high-stakes strategy, generating vast amounts of both structured and unstructured data. While structured race data has long been the focus of motorsport analytics, radio messages between teams and drivers offer a novel perspective on team strategy, crisis response, and communication under pressure. This work integrates structured F1 data from PostgreSQL with unstructured radio message data from MongoDB, enabling hybrid analytics. The objective is to provide cross-domain insights into communication patterns and strategic behaviors in F1 that are not possible using traditional, single-source analytics.

Methodology

Data Sources:

- **Structured data:** Race information, results, constructors, pit stops, and circuits from a PostgreSQL database (based on the open F1 dataset).
- **Unstructured data:** Raw and tagged radio messages in MongoDB (`F1_Message_Context.message_context`), including fields such as `race_id`, `driver_id`, `lap`, `message_text`, `message_type`, and `tags`.

Custom Python scripts using `pandas`, `pymongo`, and `psycopg2` perform ETL and analytics. Key relational mappings such as driver-to-team per race and race-to-circuit enable rich cross-database queries. All scripts and synthetic data generators are available at: `https://github.com/StratosDns/F1_PROJECT`.

Message Generation and Event Integrity:
To ensure analytical integrity and realistic event correlation, synthetic radio messages are programmatically generated based on actual race event data. For example, pit stop messages are created only if a corresponding pit stop event exists in the structured database, and retirements or technical warnings are linked to actual race retirements. This process preserves causal relationships and enables accurate validation of communication patterns against race events, preventing the inclusion of implausible or orphaned messages.

Analytical Pipeline: Queries and analytics include:

1. Matching "box" or pit-related radio messages to actual pit stop events.
2. Extracting messages preceding mechanical retirements.
3. Aggregating message volumes by circuit.
4. Analyzing message type frequencies per team.

5. Determining, for each circuit, both the total number of messages and the most communicative team.

The pipeline is designed for extensibility and can integrate further data sources.

Results and Discussion

- **High-density Circuits:** In synthetic test data, circuits such as *Monza* and *Silverstone* showed the highest radio message volumes, likely reflecting their strategic complexity.
- **Team Communication Patterns:** Teams like *Mercedes* and *Red Bull* dominated radio traffic at certain venues, possibly indicating aggressive strategies.
- **Message Type Distributions:** "Box" and pit-related messages were most frequent among leading teams, while technical failure messages varied by team and circuit.
- **Retirement Communications:** Analysis revealed a typical escalation: technical warnings followed by succinct "stop the car" directives.
- **Visualization:** Tableau dashboards allow interactive exploration by team, circuit, message type, and year.

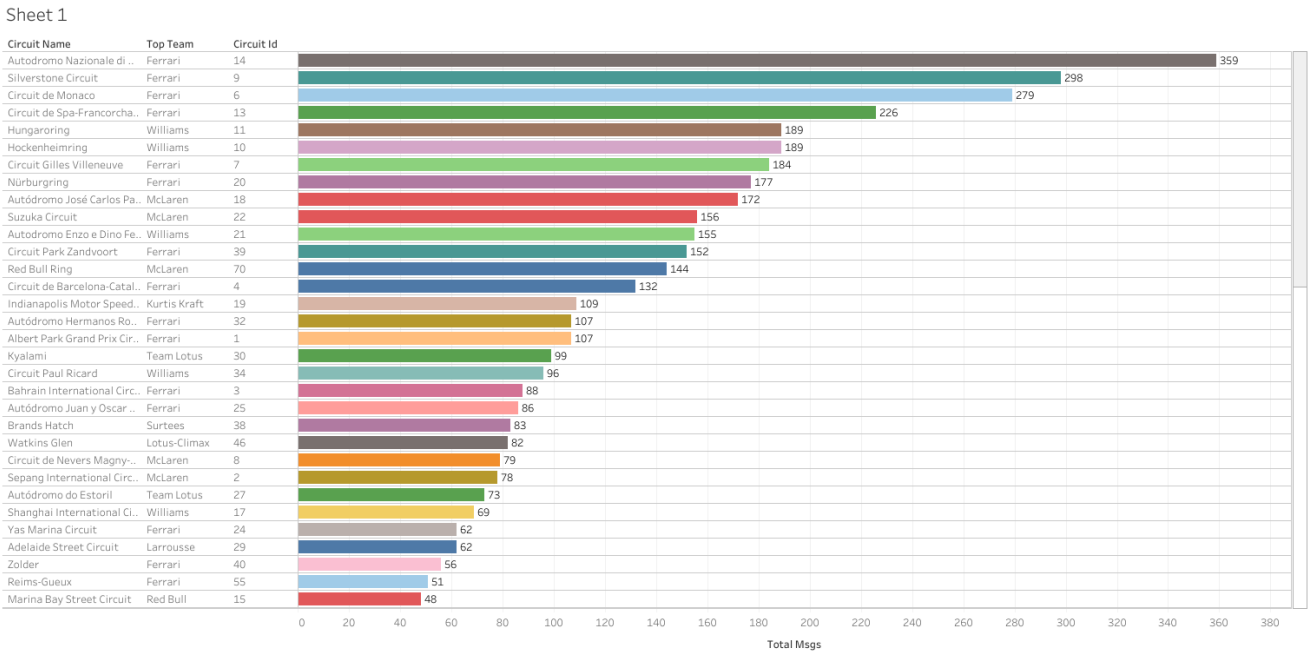

Figure 1: Example visualization: Radio message volume by circuit (synthetic data).

Conclusion

This work demonstrates the value of hybrid analytics in motorsport, integrating structured and unstructured data for advanced, context-rich insights. The modular pipeline supports rapid querying, scalable aggregation, and reproducible analytics.

References

- Kaggle F1 Dataset: `https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-19`
- Project Repository: `https://github.com/StratosDns/F1_PROJECT`