

**Agile DataOps**

Year 2 (2024/25)

**SCHOOL OF INFOCOMM TECHNOLOGY**

Diploma in Data Science

**ASSIGNMENT** **2**

|  |  |
| --- | --- |
| **Team Name:** | Team 2 |
| **Students’ Name/ID:** | Aden Chong Hao An, S10258369C  Fariha Tasnim, S10256073A  Jovan Yi Lee De, S10258267D  Keaven Toni Woo, S10258273A  Nabihah Fadilah, S10258311F |

# Table of Contents

[Table of Contents 1](#_Toc150117482)

[1. Executive Summary 2](#_Toc1962918342)

[a. Project Objectives 2](#_Toc970558759)

[b. Solution Overview 3](#_Toc310977336)

[c. Key Results 5](#_Toc334501554)

[2. Project Overview 6](#_Toc2131152444)

[a. Company Background 6](#_Toc1859293159)

[b. Current Challenges 7](#_Toc634550638)

[c. Proposed Solution 8](#_Toc426100391)

[i. Cloud Platform Migration 8](#_Toc1738977480)

[ii. OLAP System 8](#_Toc1686120498)

[iii. DataOps Pipeline Benefits 9](#_Toc1090495175)

[3. DataOps Pipeline and Cloud Migration 10](#_Toc291450819)

[a. Cloud Platform Selection 10](#_Toc681895666)

[b. Dimensional Data Warehouse 11](#_Toc13866970)

[i. ELT Process (EL) 11](#_Toc1988679454)

[c. Data Migration Process 14](#_Toc422922822)

[i. OLTP to OLAP 14](#_Toc361139155)

[ii. ELT Process (T) 15](#_Toc946647551)

[DRIVER\_TABLE 15](#_Toc1283825526)

[GRID\_STATS 16](#_Toc1462401779)

[RACE\_TABLE 18](#_Toc1790136746)

[TEAM\_STATS 20](#_Toc162009206)

[Primary and Foreign Keys 21](#_Toc263577795)

[Dynamic Tables 23](#_Toc1483956551)

[iii. Ensuring Data Accuracy, Consistency, and Security 24](#_Toc760184289)

[d. SnowSQL 25](#_Toc220365020)

[4. Role-Based Access Control (RBAC) 25](#_Toc1364523535)

[a. Overview of RBAC 26](#_Toc1802128394)

[b. Analyst Role 26](#_Toc344956001)

[c. Developer Role 26](#_Toc1438543620)

[d. CI Role 26](#_Toc1225688440)

[e. Summary of RBAC 27](#_Toc343711674)

[5. Continuous Integration/Continuous Deployment (CI/CD) 27](#_Toc137941088)

[a. Overview of CI/CD 28](#_Toc1069659981)

[b. Automation Tools 28](#_Toc2035825194)

[c. Version Control 28](#_Toc1835800372)

[d. Testing and Monitoring 28](#_Toc1818048365)

[e. Summary of CI/CD 29](#_Toc2014940279)

[6. Dashboard Insights 30](#_Toc1964266935)

[a. Driver Table Dashboard 31](#_Toc2010395906)

[i. Dashboard Overview 31](#_Toc1848164885)

[ii. Dashboard Key Metrics 33](#_Toc1808447989)

[iii. Dashboard Business Value 33](#_Toc22807254)

[iv. Dashboard Unique Insights 34](#_Toc141379888)

[b. Grid Performance & Sprint Trends 35](#_Toc418304120)

[i. Dashboard Overview 35](#_Toc293443761)

[ii. Dashboard Key Metrics 36](#_Toc494767435)

[iii. Dashboard Business Value 41](#_Toc1018047700)

[iv. Dashboard Unique Insights 41](#_Toc1531300423)

[c. Race Performance & Strategy 42](#_Toc1292762345)

[i. Dashboard Overview 42](#_Toc1737030578)

[ii. Dashboard Key Metrics 43](#_Toc1683557811)

[iii. Dashboard Business Value 46](#_Toc553288388)

[iv. Dashboard Unique Insights 46](#_Toc1082481462)

[d. Team Table Dashboard 47](#_Toc1039817753)

[i. Dashboard Overview 47](#_Toc1889110892)

[ii. Dashboard Key Metrics 47](#_Toc1217949118)

[iii. Dashboard Business Value 49](#_Toc1178299132)

[iv. Dashboard Unique Insights 50](#_Toc276096282)

[e. Summary of F1 Statistics 51](#_Toc1935009109)

[i. Dashboard Overview 52](#_Toc93214592)

[ii. Dashboard Key Metrics 52](#_Toc349423603)

[iii. Dashboard Business Value 57](#_Toc1305756783)

[iv. Dashboard Unique Insights 58](#_Toc2003730393)

[7. Conclusion 59](#_Toc362577051)

[a. Final Results 60](#_Toc632933451)

[b. Business Impact 61](#_Toc2138081350)

[c. Closing Thoughts 62](#_Toc1671855114)

[8. Appendix 62](#_Toc734143095)

[a. Challenges 62](#_Toc1793627335)

[i. Technical Challenges 62](#_Toc1427242987)

[ii. Team Dynamics 64](#_Toc1285731186)

[iii. Solutions 65](#_Toc563223156)

[b. Reflection and Lessons Learned 66](#_Toc971252927)

[i. Agile Methodology 67](#_Toc191658230)

[Sprint Structure and Timeline 67](#_Toc1992074576)

[Sprint Components and Discussions 67](#_Toc744662209)

[Impact of Agile Methodology 68](#_Toc182207921)

[ii. Personal Growth 68](#_Toc568254053)

[Aden 69](#_Toc262275475)

[Fariha 69](#_Toc1885011582)

[Jovan 70](#_Toc285392283)

[Keaven 70](#_Toc615795657)

[Nabihah 71](#_Toc148718137)

[iii. Team Growth 71](#_Toc577342739)

[Aden 71](#_Toc133485946)

[Fariha 72](#_Toc1569722265)

[Jovan 72](#_Toc1256825421)

[Nabihah 73](#_Toc2139455921)

[Keaven 73](#_Toc2044974198)

[iv. Future Applications 73](#_Toc1460244016)

[9. References 75](#_Toc1135950864)

[a. Content 76](#_Toc1808507740)

[b. Images 76](#_Toc1217877446)

# 1. Executive Summary

## a. Project Objectives

Formula 1 (F1) is an internationally renowned organisation that hosts annual events which attract millions of fans and participants worldwide. With the sport’s growing popularity and the increasing volume of data generated from these events, F1 is facing significant challenges in managing its on-site databases.

The primary pain points stem from the limitations of the current physical database system. These include:

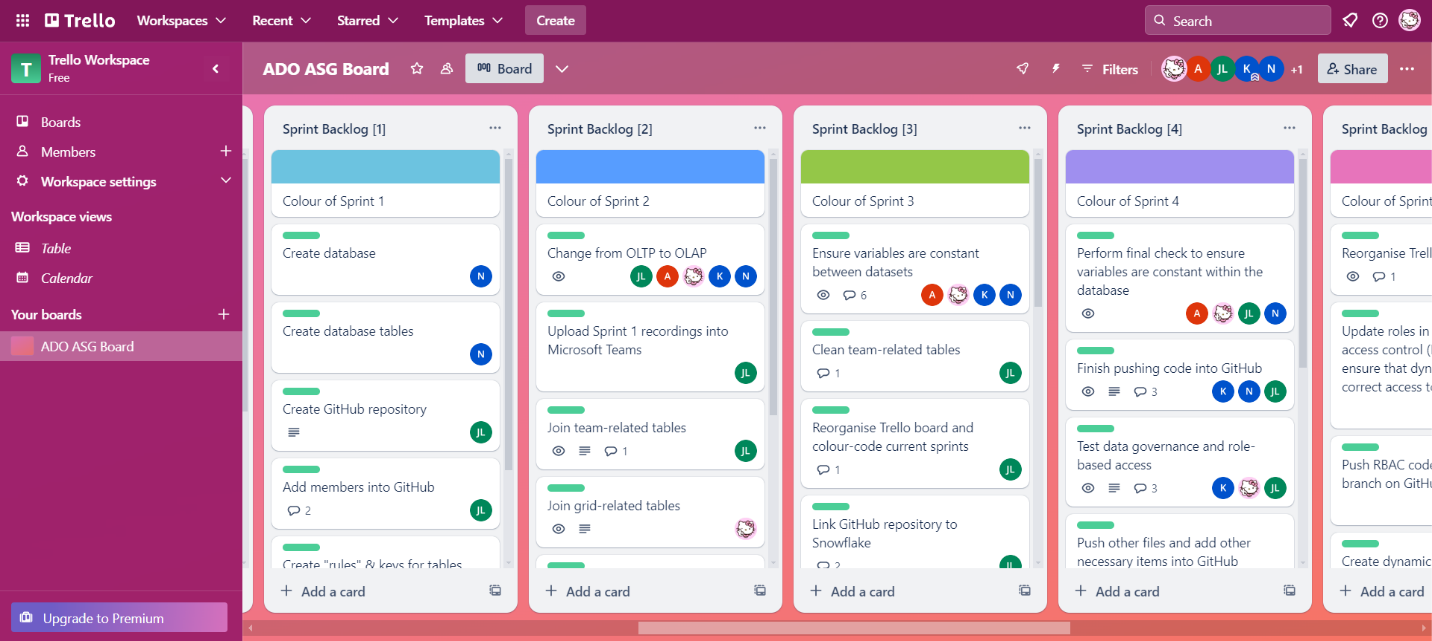
* **Scalability Issues:** The on-site databases struggle to keep up with the growing volume of data, requiring time-intensive and manual processes to expand capacity.
* **Inefficiency:** Maintaining the physical infrastructure demands substantial manpower and financial resources, which is neither sustainable nor cost-effective.
* **Lack of Flexibility:** The physical nature of the database restricts remote accessibility, which poses challenges given the global nature of F1's operations.

To address these issues, the project aims to modernise F1’s data architecture by migrating its database from an on-premises system to a cloud-based platform. This migration is expected to achieve the following management goals:

* **Improve Scalability:** Enable the database to seamlessly handle growing data demands without manual intervention.
* **Increase Efficiency:** Reduce operational costs and the manpower required to maintain and update the system.
* **Enhance Flexibility:** Provide employees with the ability to access and update the database from any location, aligning with the international scope of F1 events.
* **Strengthen Analytics Capabilities:** Equip the data team with advanced tools and improved data accessibility to deliver deeper business insights and drive data-driven decision-making.

By addressing these challenges and meeting these objectives, this project ensures that F1 remains agile and competitive in an increasingly data-driven industry.

## b. Solution Overview

*Screenshot of our team’s Trello Board*

The migration process was guided by the principles of Agile and DataOps, which provided a structured and collaborative framework to ensure the project’s success.

Agile methodologies, particularly through the use of scrums, played a crucial role in keeping the project on track. The team used Trello to document progress in detail, ensuring that every task was accounted for and regularly updated. This granular visibility allowed us to identify potential issues early, make necessary adjustments to our code and methods, and prevent errors from accumulating. By addressing these minor issues proactively, we ensured the overall efficiency and smooth execution of the project.

DataOps was central to implementing a robust and scalable solution. The team leveraged GitHub to collaboratively manage and commit changes to the Snowflake-based codebase. This fostered open communication and teamwork, allowing us to refine our pipelines iteratively. Within Snowflake, we designed data pipelines that streamlined the migration process and ensured the database would remain scalable and adaptable for future needs.

Data governance was another critical component of the solution. Using DataOps practices, we established strict access controls and security measures to safeguard the migrated database. Only authorised team members were granted access, minimising risks and ensuring the integrity of the pipelines.

By combining Agile practices with DataOps principles, the team successfully migrated the database from a physical on-premises server to a cloud-based platform. This structured approach not only ensured a smooth transition but also laid the groundwork for enhanced dashboard visualisation.

## c. Key Results

All in all, the migration of F1's database from on-premises to the cloud has yielded significant improvements for the company and its data team. These outcomes are as follows:

1. **Enhanced Data Architecture**

The shift to a cloud-based system has revolutionised the way F1 can manage its data. The new architecture supports scalability, allowing F1 to adapt seamlessly to the growing volume of data generated by its annual events and increasing fan base. This scalability eliminates the need for manual interventions to expand storage capacity, enabling F1 to efficiently handle its dynamic and ever-expanding data requirements.

1. **Accelerated Reporting Speed**

With data now housed in a cloud environment, the reporting process has become significantly faster. This improvement stems from the optimised pipelines designed during the migration, which streamline data processing and ensure timely availability of critical insights. F1’s data team can now generate real-time reports, enabling quicker decision-making and more responsive operations during high-pressure events such as the Grand Prix.

1. **Improved Business Insights via Dashboards**

The integration of advanced dashboards has empowered F1’s stakeholders to access actionable insights effortlessly. These dashboards, powered by the migrated data, provide intuitive visualisations that facilitate deeper analysis of performance metrics and operational efficiency. With improved accessibility and clarity, the dashboards have posed themselves as an invaluable tool for both strategic planning and day-to-day management.

1. **Operational Efficiency for the Data Team**

The cloud-based system has simplified data governance and accessibility. The data team can now collaborate more effectively, as they no longer face the limitations of a physical database. Features such as centralised access controls and automated workflows have reduced the time and effort required for routine maintenance and updates, allowing the team to focus on more impactful tasks, such as predictive analytics and innovation.

1. **Cost Optimisation**

By transitioning to the cloud, F1 has reduced its dependency on expensive physical infrastructure. The cloud solution not only minimises operational costs but also offers pay-as-you-go pricing, ensuring cost efficiency in the long term. These savings can be redirected towards enhancing other areas of the business, such as fan engagement and technology development.

Through this project, F1 has experienced a transformative improvement in its data management capabilities. The benefits realised position the company for future growth while enabling its data team to achieve greater productivity and innovation.

# 2. Project Overview

## a. Company Background

F1 is the highest class of international single-seater motorsport, sanctioned by the Fédération Internationale de L'Automobile (FIA). Since the inception of the FIA Formula One World Championship in 1950, F1 has grown into a global sporting and commercial enterprise, with races held across multiple continents and a vast international audience.

As a data-intensive organisation, F1 generates and processes large volumes of data across various domains, including race performance, vehicle telemetry, regulatory compliance, and commercial operations. This data is critical for optimising decision-making, enhancing competition integrity, and improving operational efficiency. Given the increasing scale and complexity of its data landscape, F1 continuously seeks to modernise its data infrastructure to support seamless data management, accessibility, and analytics.

This project aligns with F1’s broader digital transformation efforts by focusing on the migration of its legacy on-premise database to a cloud-based solution, enabling more efficient data processing, improved scalability, and enhanced accessibility for key stakeholders.

## b. Current Challenges

* **Data Integrity and Driver Performance Analysis**

One of the primary challenges F1 faces when analysing driver performance is the inability to directly capture or quantify individual driver skills. Current data typically focuses on time-based metrics, such as lap times, pit stops, and race completion times, but does not include subjective performance indicators such as steering ability or overall driving technique. As a result, it becomes difficult to isolate and analyse the impact of the driver's skill apart from the car's performance. This limitation complicates direct comparisons between drivers, especially when evaluating their potential in varied race conditions. While timing data remains a key performance indicator, it provides an indirect measure of driver performance that is not as nuanced as data that could capture skill-based metrics.

* **Data Overload and Complexity**

F1’s vast and varied data streams, including telemetry data from cars, sensor readings, GPS data, and video analysis, present significant challenges. The sheer volume and complexity of this data often result in difficulties in processing, interpreting, and using the information effectively. The unstructured nature of some of these data types can make it challenging to extract meaningful insights without advanced data management techniques. The large number of variables involved further complicates the extraction of accurate, timely insights, which can lead to slower decision-making.

* **Scalability and Performance Issues**

The growing volume of data being generated, especially during high-speed, real-time events like races and practice sessions, places significant strain on the existing on-premise database systems. As the amount of data increases, there are challenges in maintaining quick response times and low latency when running analysis or generating reports. These system limitations result in delays, particularly when dealing with large data sets, which can hinder the ability of race strategists to make real-time, data-driven decisions during races.

* **Data Accuracy and Inconsistencies**

The complexity of data, coupled with the possibility of inaccuracies and inconsistencies, further complicates the analytics process. Inaccurate data points or inconsistent reporting methods across different races and sessions can lead to skewed results in visualisations and reports. For example, variations in race conditions, such as course layout or weather, impact driver performance and times, making it difficult to directly compare performance metrics across events. Additionally, the absence of standardised methods for handling these variations can result in misleading insights, necessitating additional data cleaning and correction efforts to ensure reliability.

These challenges highlight the limitations of F1's current data infrastructure, which struggles to manage the scale, complexity, and accuracy of data, ultimately impacting the speed and effectiveness of decision-making processes.

## c. Proposed Solution

### i. Cloud Platform Migration

The proposed solution involves migrating F1’s existing on-premise database to a cloud-based platform, specifically Snowflake. This migration will enhance scalability, flexibility, and data management capabilities. The first step is to upload F1’s datasets from external sources (such as Kaggle) into Snowflake. Snowflake’s cloud infrastructure offers dynamic scalability, allowing the system to handle large volumes of data with ease. Once the data is transferred, Snowflake’s powerful features will be used to clean, transform, and optimise the data, making it more accessible and usable for analytics. The transformed data will then be integrated with Power BI to create actionable visualisations, providing valuable insights for F1’s data analysts and decision-makers.

### ii. OLAP System

Upon reviewing the F1 datasets, it became clear that the data was structured for an OLTP (Online Transaction Processing) system rather than an OLAP (Online Analytical Processing) system. OLTP systems are designed for transactional data handling, whereas OLAP systems are optimised for complex queries and data analysis, making them essential for generating insights. Since the data will primarily be used for analytics to inform strategic decision-making for F1, it is crucial to convert the dataset to an OLAP-compatible format. This transition will occur once the data is successfully migrated to Snowflake, allowing us to optimise the database for faster querying and in-depth reporting. The shift to OLAP will ensure that the data is structured in a way that facilitates efficient analytical processing and supports advanced reporting and visualisation tasks.

### iii. DataOps Pipeline Benefits

* **Scalability**

The DataOps pipeline, built on Snowflake’s cloud-based architecture, ensures that the data infrastructure can scale with F1’s growing data needs. By organising data into individual tables, which are then joined and related, the system can accommodate a diverse set of data sources. This structure enables the team to quickly generate reports and visualisations from an array of datasets, improving accessibility and efficiency in data analysis.

* **Performance**

Cloud-based platforms like Snowflake offer significant performance advantages over physical servers. Unlike on-premise systems, which are constrained by hardware limitations and geographical proximity, cloud servers leverage high-speed internet connectivity and can scale resources dynamically based on demand. This flexibility allows for rapid expansion as the dataset grows, reducing the time needed for maintenance and infrastructure upgrades. Additionally, cloud systems tend to provide faster data retrieval and processing, enabling real-time reporting and analytics for the F1 team.

* **Data Security & Governance**

Cloud platforms provide robust security protocols, with regular updates and automatic compliance with global standards. This ensures that the data is consistently protected against breaches and other security threats. In contrast, maintaining in-house servers requires ongoing monitoring and manual intervention by IT staff to safeguard data. Snowflake’s cloud infrastructure follows a shared security model, ensuring that both data protection and compliance are handled efficiently. This removes the burden of managing security from F1’s internal teams, providing peace of mind and enhanced data integrity.

# 3. DataOps Pipeline and Cloud Migration

## a. Cloud Platform Selection

For this project, we selected Snowflake as the cloud platform to host F1’s migrated database. Snowflake has rapidly emerged as a leading solution in the cloud database industry, offering cutting-edge capabilities that align with the needs of modern organisations transitioning from on-premises systems. Its adoption by companies across various industries demonstrates its effectiveness in addressing scalability, performance, and data governance challenges.

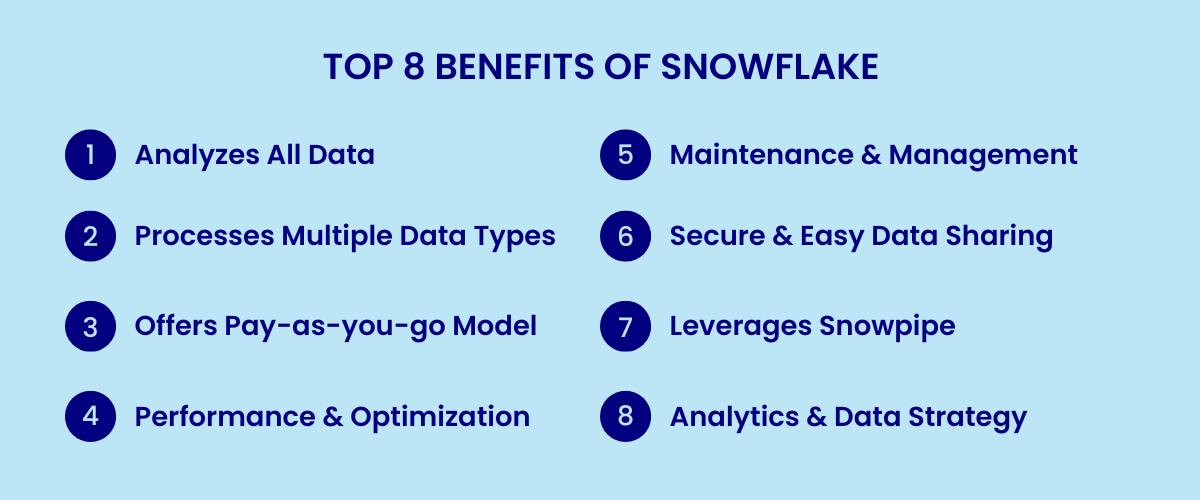
One of Snowflake’s most significant advantages is its scalability. Unlike traditional on-site databases, Snowflake’s cloud infrastructure enables seamless expansion to accommodate the growing volume of F1’s data. This ensures that the database remains agile and can scale in line with the increasing complexity of Formula 1’s operations.

In terms of performance, Snowflake supports Continuous Integration and Continuous Development (CI/CD), facilitating collaboration among F1’s data engineers. Through integration with GitHub, multiple engineers can work simultaneously on improving the database, a capability unavailable in the on-premises system, which restricted access to a single user at a time. This feature accelerates development cycles and enhances overall efficiency.

Snowflake also excels in data governance. Its robust role-based access control system allows for precise permission settings and role inheritance, ensuring that sensitive data is only accessible to authorised personnel. This enhances the security and integrity of the database, reducing the risk of accidental deletions or unauthorised access.

Additionally, Snowflake’s commitment to continuous innovation ensures that the platform evolves to meet the changing needs of its users. New features and enhancements are regularly introduced, which will likely provide further benefits to F1’s data storage and operational requirements in the future.

Given these advantages, Snowflake is an ideal choice for Formula 1’s transition to a cloud-based database. It provides the scalability, performance, and governance needed to support F1’s dynamic operations while ensuring that the database remains future-ready and secure.

*Figure 3.1 Image of top 8 benefits of Snowflake*

## b. Dimensional Data Warehouse

### i. ELT Process (EL)

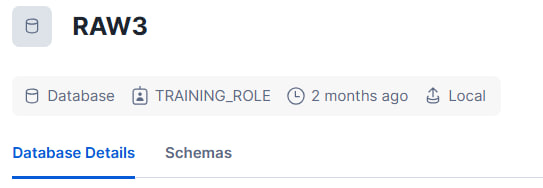
The Extract and Load (EL) component of the ELT process focused on migrating the on-premise F1 database to the cloud environment in Snowflake. The following steps were carried out to set up a solid foundation for the data migration:

1. **Warehouse Setup**

A new warehouse was created in Snowflake to provide the necessary compute resources for the migration and subsequent data processing tasks. This warehouse ensures efficient handling of the data extraction and loading processes.

1. **Database Creation**

A dedicated database named RAW\_3 was created in Snowflake to organise and store the raw data extracted from the F1 database.



*Screenshot of created database*

1. **Schema Definition**

Within RAW\_3, a schema named F1\_3 was created to logically structure the data and improve accessibility for downstream transformations.



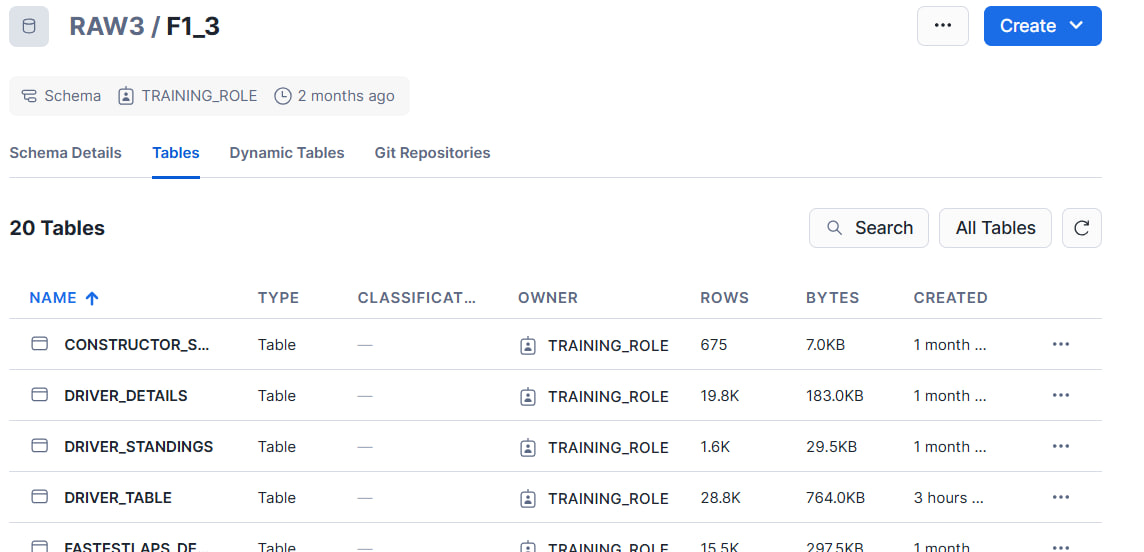
*Screenshot of defined schema*

1. **Table Creation**

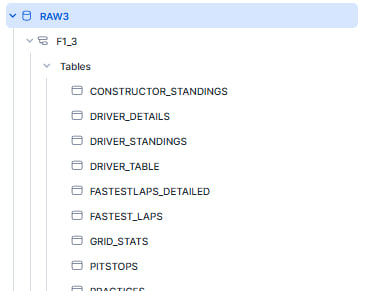
All 14 tables from the F1 database were recreated in Snowflake under the F1\_3 schema. Each table's structure was defined to match the original source data but prepared for further refinements during the transformation (T) phase. Key data types, such as STRING and DATE, were configured based on the requirements of the extracted data.

The tables included in this process were:

* 1. driver\_details.csv
  2. driver\_standings.csv
  3. pitstops.csv
  4. practices.csv
  5. sprint\_results.csv
  6. sprint\_grid.csv
  7. starting\_grids.csv
  8. fastest\_laps.csv
  9. fastestlaps\_detailed.csv
  10. qualifyings.csv
  11. race\_details.csv
  12. race\_summaries.csv
  13. constructor\_standings.csv
  14. team\_details.csv

*Screenshot of created tables*

The EL phase concluded with the successful migration of the table structures and raw data into the Snowflake environment, enabling the team to proceed with the transformation (T) phase for further data processing and refinement.



*Screenshot of result of EL process*

## c. Data Migration Process

This section outlines the data migration process. It details the steps and methodologies utilised, as well as the rationale behind key decisions made during the process.

### i. OLTP to OLAP

To transition from an OLTP to an OLAP system, we consolidated the 14 original tables into four primary tables: DRIVER\_TABLE, GRID\_STATS, RACE\_TABLE, and TEAM\_STATS. We categorised the original tables based on their relevant entities, allowing us to define clear relationships between the data. By identifying the key tables to join, we applied left joins to ensure comprehensive data retention without loss. After the tables were merged, we undertook a thorough data cleaning process to reduce the number of null values, enhancing the quality of the dataset for analytical purposes. This approach ensured the data was structured for optimal performance in an OLAP system, facilitating more effective reporting and analysis.



*Screenshot of tables from the GitHub repository*

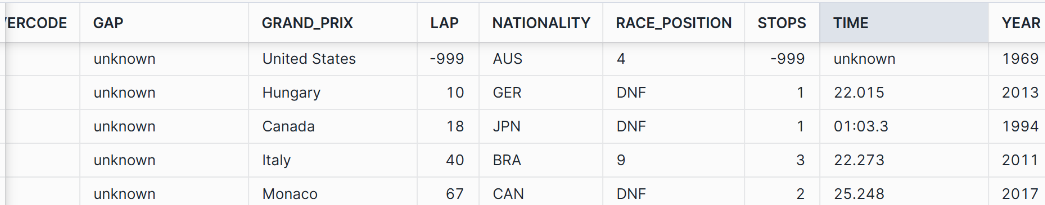
### ii. ELT Process (T)

#### DRIVER\_TABLE

The DRIVER\_TABLE aggregates information related to drivers, including data on cars, Grand Prix events, laps, and times. The various data sources were joined on relevant columns, such as year, car, driver, Grand Prix, and driver code, to ensure data consistency and completeness. This careful join process prevented both data overloading and omission. The DRIVER\_TABLE incorporates data from several source tables, including driver\_details, driver\_standings, pitstops, practices, and sprint\_results.

Upon merging the data, it was observed that several columns contained null values. To handle these nulls, different strategies were employed depending on the data type:

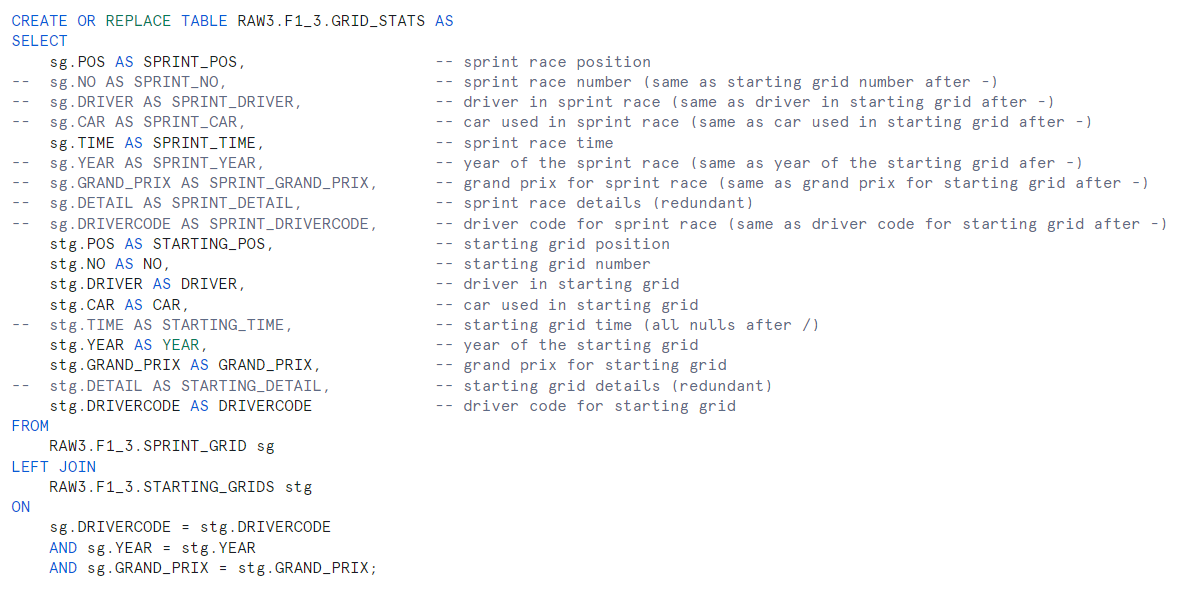
* **Categorical Data:** Null values were imputed with the value "unknown" to preserve data integrity.
* **Numerical Data:** Null values were imputed with a placeholder value of "-999" to differentiate them from actual data.

*Screenshot of inconsistent values in time*

This approach was implemented to avoid imputation with mean or median values, which could introduce inaccuracies or confusion, particularly in time-related data. For instance, inconsistencies in the time column were addressed by marking unknown or missing values with these extreme placeholders. This ensures that when the data is visualised in Power BI, these outliers can be easily filtered out, allowing for more accurate insights and avoiding misleading trends.

#### GRID\_STATS

The GRID\_STATS table was created by joining the sprint\_grid and starting\_grids tables using a LEFT JOIN on shared columns. This approach ensured the inclusion of all relevant grid data while avoiding data loss from unmatched rows.

*Screenshot of joining of sprint\_grid and starting\_grids to create GRID\_STATS*

The data cleaning process was carried out in the following manner:

1. **Removal of Duplicate Columns**

After the initial join, several columns from the sprint\_grid table (SPRINT\_NO, SPRINT\_DRIVER, SPRINT\_CAR, SPRINT\_YEAR, SPRINT\_GRAND\_PRIX, and SPRINT\_DRIVERCODE) were identified as duplicates of corresponding columns in STARTING\_GRIDS (NO, DRIVER, CAR, YEAR, GRAND\_PRIX, and DRIVER\_CODE). These were removed to streamline the dataset.

1. **Removal of Irrelevant Columns**

The SPRINT\_DETAIL and STARTING\_DETAIL columns were excluded as they did not provide any meaningful insights for the analysis.

The STARTING\_TIME column was removed due to its containing only NULL values after the join.

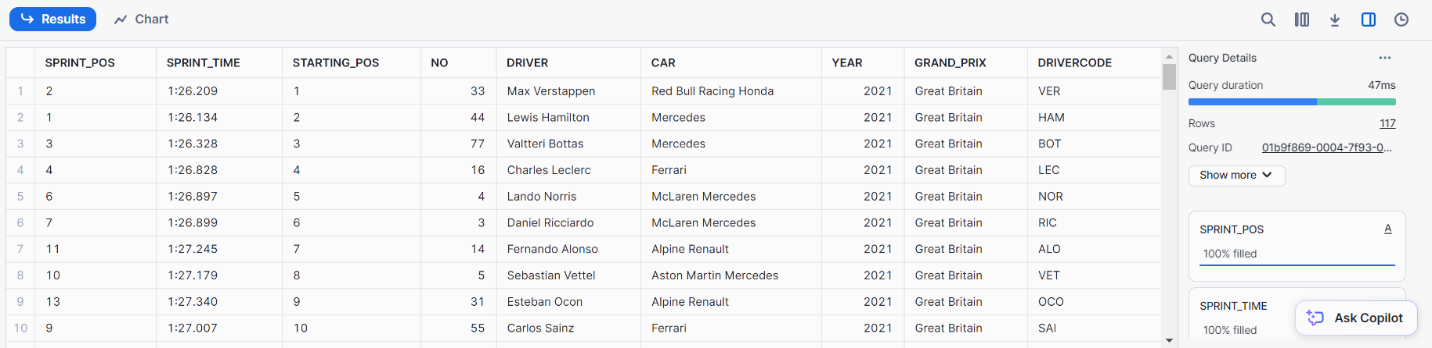
1. **Row Filtering**

Rows with NULL values in key columns from STARTING\_GRIDS (e.g., STARTING\_POS, NO, DRIVER, CAR, YEAR, GRAND\_PRIX, DRIVERCODE) were removed. This step eliminated incomplete records, ensuring data integrity.

1. **Handling NULL Values in SPRINT\_TIME**

To manage NULL values in the SPRINT\_TIME column, the COALESCE function was used to replace NULL values with 'N/A'. This ensured consistency and readability in the data, particularly for visualisation purposes.

The final GRID\_STATS table is a clean, consolidated dataset containing the essential grid and starting grid information. By eliminating redundancies, irrelevant data, and NULL values, the table has been optimised for further analysis. This process ensured the data remains accurate, reliable, and ready for actionable insights.

*Screenshot of a snippet of the final GRID\_STATS*

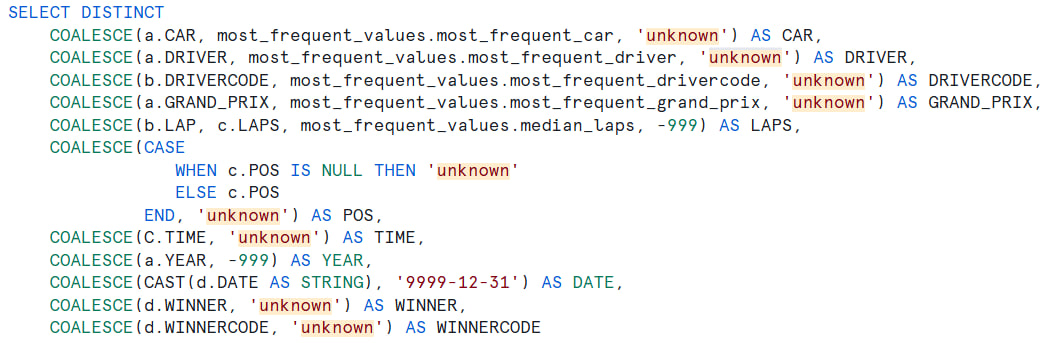
#### RACE\_TABLE

The RACE\_TABLE was created to consolidate key race-related data from multiple sources, ensuring a comprehensive and clean dataset for analysis. This table combined data from five original tables: fastest\_laps, fastestlaps\_detailed, qualifyings, race\_details, and race\_summaries. The process included:

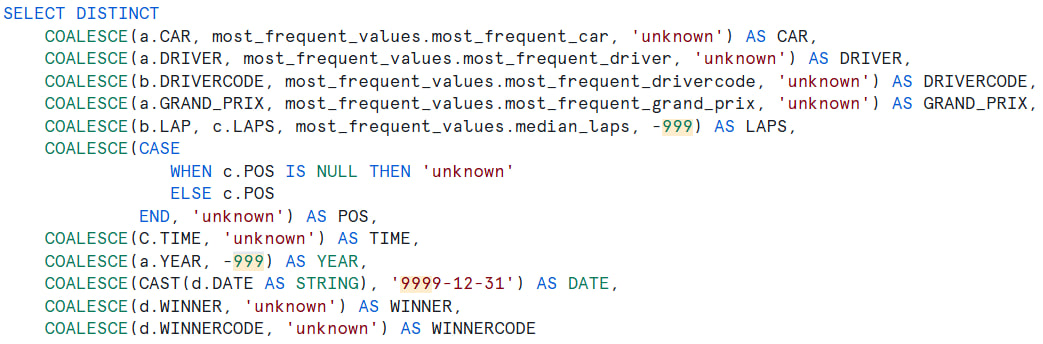
1. **Joining Tables**

The tables were merged using LEFT JOIN operations on key columns, such as DRIVER, YEAR, CAR, and GRAND\_PRIX. This ensured all relevant information was captured, even when some data sources had incomplete records.

1. **Handling Missing Data**
2. Categorical data (e.g., CAR, DRIVER, GRAND\_PRIX) was filled with the most frequently occurring values where possible or replaced with 'unknown' if necessary.

 *Screenshot of handling of missing categorical data*

1. Numeric data (e.g., LAPS, TIME) was imputed using median values or assigned -999 as placeholders for easy identification of missing values.
2. Missing dates were set to 9999-12-31 to clearly distinguish them from valid dates.



*Screenshot of handling of missing numeric data and dates*

1. **Data Cleaning**

Inconsistent time formats (e.g., values containing “:”) were standardised by converting them into seconds.



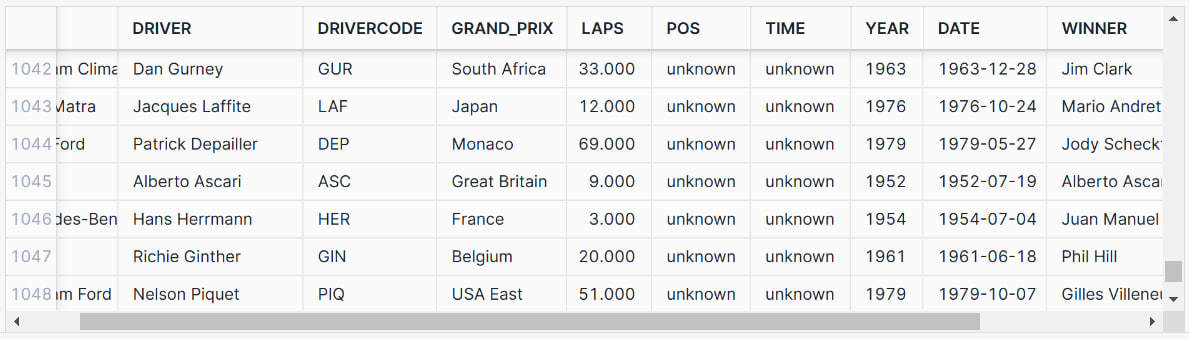
*Screenshot of time format standardisation*

Logical checks were applied to ensure the integrity of combined data, such as verifying column consistency and removing duplicates.

1. **Validation**

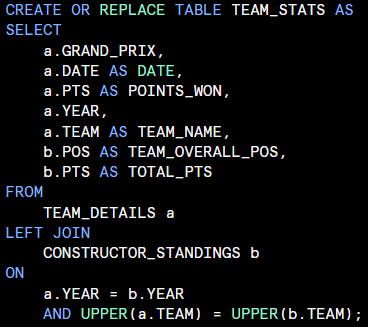
The final table was verified to confirm all expected columns and rows were present and correctly formatted. Constraints were checked using Snowflake's metadata to ensure they were implemented as planned.

By transforming and integrating these datasets, the Race Table serves as a clean, structured source for analysing race-specific statistics, ensuring data reliability and consistency for further use.

*Screenshot of a snippet of the final RACE\_TABLE*

#### TEAM\_STATS

The TEAM\_STATS table was created by joining the constructor\_standings table with the team\_details table using a left join. To ensure consistent value mapping between the tables, the UPPER() function was applied to the TEAM column during the join process. This standardisation helped address any inconsistencies in the team name formatting.



*Screenshot of usage of UPPER() in the joins to create TEAM\_STATS*

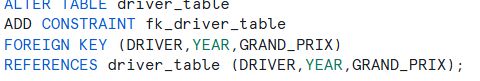
After the join, the table contained 12 null values in the TEAM\_OVERALL\_POS and TOTAL\_PTS columns, all of which belonged to the same year and team. These null values were updated based on the original dataset, ensuring that the data remained accurate and complete. This step was essential for maintaining data integrity and supporting reliable analysis in subsequent stages of the project.

*Screenshot of the final TEAM\_STATS*

#### Primary and Foreign Keys



* Primary Key
  + The primary key was established with the driver, year and grand prix columns. This was made so that the table can be identified with a key that uniquely identifies each row of data.



* Foreign Key
  + For the foreign key, the key has been made the same to the primary key. This foreign key ensures that there is an established relationship between the driver tables to the rest of the tables.



* Checking columns
  + With this code, we can use the tab that pops up below up, to check for the values that are in the table itself. We can press on each column to see the percentages of null values in each column.



* Primary Key
  + The primary key for the team tables has been defined with the columns team name, year and grand prix. This is to ensure that each of the rows of the data has been defined uniquely.
* Foreign Key
  + The foreign key has been defined the same as the primary key so that when linking to other tables, there can be an established relationship. However, this table is by itself since it contains data that do not have relationship with the other tables



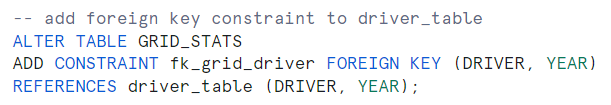
* Data Checking
  + This code checks for the data values in the columns of the table. When running the code, we can see the pop-up tab in the front that checks for the nulls values and the overall preview of the table.

As for GRID\_STATS and RACE\_TABLE, primary and foreign keys were defined as such:

* **Primary Keys**

In both GRID\_STATS and RACE\_TABLE, a PRIMARY KEY constraint was applied on the combination of YEAR, GRAND\_PRIX, and DRIVERCODE. This combination uniquely identifies each record in the tables and ensures no duplication of data.

* **Foreign Keys**
* For GRID\_STATS, a foreign key relationship was established by linking DRIVER and YEAR columns to the DRIVER\_TABLE using fk\_grid\_driver. This ensures that the grid table refers to valid driver records.

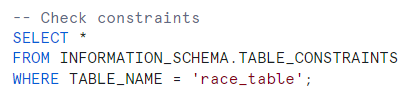
*Screenshot of adding of foreign key constraint in GRID\_STATS*

* For RACE\_TABLE, a similar foreign key relationship was established by linking DRIVER and YEAR columns to the DRIVER\_TABLE via fk\_race\_driver. This ensures data consistency and integrity across the related tables.
* **Data Consistency**

By setting up these foreign key relationships, we ensured referential integrity across the tables, meaning that entries in both GRID\_STATS and RACE\_TABLE only reference valid records from the DRIVER\_TABLE and TEAM\_STATS. This prevents orphan records and maintains the accuracy of relationships between data entries.

* **Validation**

To confirm the successful addition of primary and foreign keys, queries were executed on INFORMATION\_SCHEMA.TABLE\_CONSTRAINTS, validating the presence and correctness of the defined constraints.



*Screenshot of check for constraints in RACE\_TABLE*

#### Dynamic Tables

Dynamic tables are a new type of table in Snowflake. They offer significant advantages by automatically updating whenever the underlying data changes, whether through insertion, updates, or deletions. These tables are created and maintained using straightforward SQL statements, making them not only easy to implement but also simple to manage over time.

The implementation of dynamic tables took place in Sprint 5. In the preceding sprints (2 and 3), the team consolidated the data into four aforementioned main tables. Using these consolidated tables, four corresponding dynamic tables were created — one for each of the primary data tables.

The key feature of dynamic tables is their ability to check for and remove duplicate entries when new data is inserted. If a new entry is identified as a duplicate, it is automatically removed to ensure data consistency.

These dynamic tables play a crucial role in connecting to Microsoft Power BI. Since the data is frequently updated in real-time, any changes to the dynamic tables automatically trigger updates in the Power BI dashboards. This capability streamlines real-time data processing, allowing for faster and more accurate dissemination of information. For F1, this ensures that dashboards are refreshed promptly, which in turn accelerates decision-making processes and enhances overall productivity across the company.

### iii. Ensuring Data Accuracy, Consistency, and Security

|  |  |
| --- | --- |
| **Accuracy** | Data was validated post-migration, with nulls handled systematically (e.g., 'unknown' for categorical, -999 for numeric) to prevent imputing values that were assumed. Formats like time values were standardised for uniformity. |
| **Consistency** | Schema alignment, referential integrity via foreign keys, and de-duplication ensured consistency across datasets and joins. |
| **Security** | Role-based access, encrypted transfers, and audit logs safeguarded data throughout the migration process. |

## d. SnowSQL

SnowSQL served as the primary tool for interacting with Snowflake, providing an efficient and seamless interface for executing SQL queries directly within the data warehouse. This command-line utility enabled the team to manage Snowflake’s data infrastructure, including the creation of databases, schemas, and tables, all from a unified environment. By using SnowSQL, there was no need to switch between multiple tools, ensuring a streamlined workflow for executing tasks and enabling real-time updates to the data warehouse.

Furthermore, SnowSQL facilitated the smooth migration, transformation, and validation of data. Its ability to automatically reflect changes in the data warehouse reduced the complexity of these processes and ensured data integrity throughout the project. This integration helped to optimise the management of the data pipeline and contributed to a more efficient and cohesive data migration process.

# 4. Role-Based Access Control (RBAC)

## a. Overview of RBAC

This section details the implementation of RBAC within the database system. RBAC is used to manage access to data by assigning specific roles to employees based on their responsibilities, ensuring that each role has an appropriate level of access to the database. In this system, three roles are defined: Analyst, Developer, and CI (Continuous Integration). The role hierarchy is structured to progressively allow more access, with each role’s permissions building on the previous one.

## b. Analyst Role

The Analyst role is the most restricted role, with limited access rights. Analysts can only view dynamic tables within the database. When connecting to Power BI using the Analyst role, they will only have access to the dynamic tables for reporting and analysis purposes. This role ensures that users can analyse the data without altering any underlying structures.

## c. Developer Role

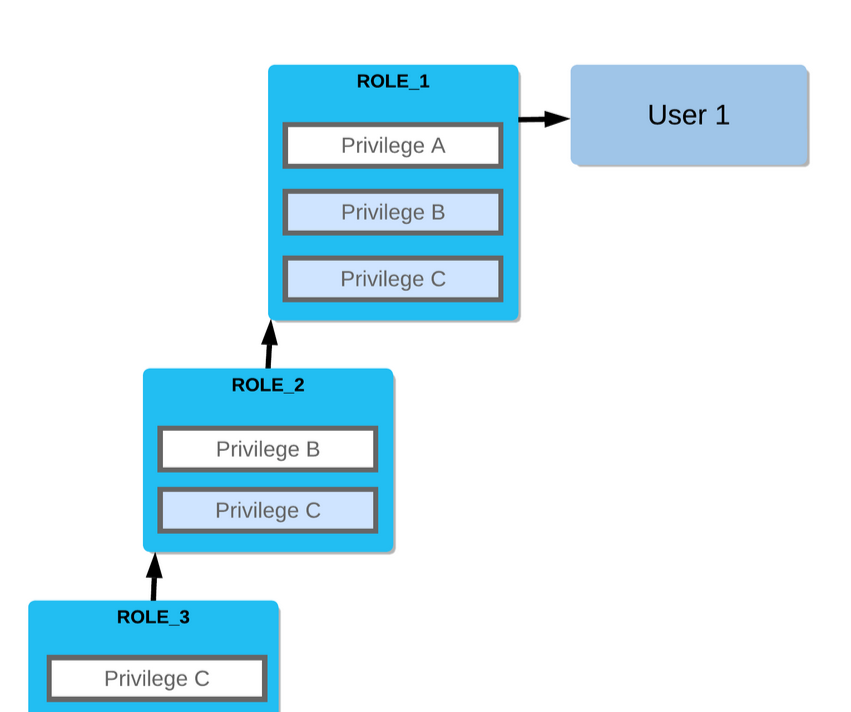
The Developer role includes all privileges of the Analyst role but extends access to the original and consolidated tables, in addition to the dynamic tables. Developers also have the ability to insert new data into these tables when necessary. This role provides broader access, enabling developers to make necessary updates and enhancements to the data without compromising security or control.

## d. CI Role

The CI role inherits all the privileges of both the Analyst and Developer roles. However, the CI role also grants the ability to perform update and delete operations on the database. These capabilities are carefully restricted, as modifying existing data can be risky. The CI role, being the highest level of access before the account administrator, is designed to ensure that any changes to the data are made cautiously, reducing the risk of data corruption or loss.

## e. Summary of RBAC

RBAC is a critical component of the database security framework. By implementing a structured hierarchy of roles, it helps safeguard data confidentiality, minimise the risk of data leakage, and ensure that employees only have access to the information necessary for their tasks. This control also mitigates the potential for errors or malicious actions that could affect the integrity of the database. The RBAC structure provides a clear delineation of responsibilities and access levels, aligning with best practices for data security in any organisation. The following diagram illustrates the RBAC structure employed in this project.



*Figure 4.1 Image of RBAC role hierarchy*

# 5. Continuous Integration/Continuous Deployment (CI/CD)

## a. Overview of CI/CD

This section provides an overview of the CI/CD pipeline used in this project. CI/CD aims to streamline the software development lifecycle by automating code integration, testing, deployment, and monitoring. The CI/CD pipeline consists of three main components: automation tools, version control, and testing and monitoring. These components work together to simplify and accelerate the process of code pushing and pulling, ensuring faster and more reliable deployments.

## b. Automation Tools

Snowflake’s Snowsight was used to facilitate the CI/CD process by automating database schema updates and code deployment. A remote Git repository was integrated with Snowflake, allowing for seamless synchronisation of files between the Snowflake environment and the repository. This setup enables efficient management of database schema updates, reducing manual intervention and ensuring that all updates are consistent and up-to-date.

## c. Version Control

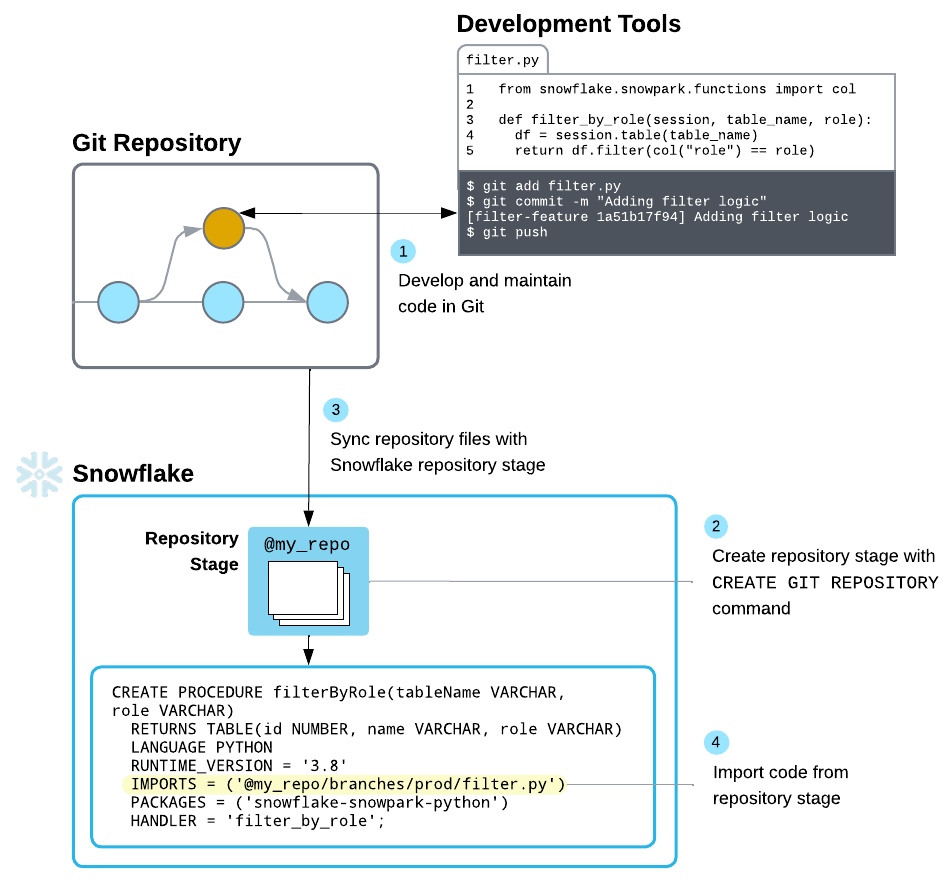
GitHub was employed as the version control system to manage and track changes to the codebase. The integration between Snowflake and GitHub simplified collaboration, as team members could easily push and pull code changes between the two platforms. A development branch was created to serve as the primary branch for all contributions, ensuring that no changes were directly committed to the main branch. This setup enabled easy rollback of changes if needed, and ensured consistency across the codebase as all team members worked from the same branch.

## d. Testing and Monitoring

Before committing code to the development branch, all changes were thoroughly tested. Snowflake provides a built-in check to ensure that the most recent version of the code is pulled from the remote repository before any edits are made. Additionally, one team member was designated to merge branches and monitor the merging process, providing an extra layer of oversight and ensuring that the code was tested again before being merged into the main branch. This workflow ensured that only tested, validated code was deployed, minimising the risk of errors.

## e. Summary of CI/CD

The CI/CD pipeline is a crucial component in optimising development tasks such as code integration, testing, deployment, and monitoring. By automating these tasks, the pipeline allows database engineers and developers to focus on more strategic tasks, improving overall efficiency and reducing the time spent on manual processes. The CI/CD approach ensures faster, smoother deployments and helps maintain code quality throughout the development process.

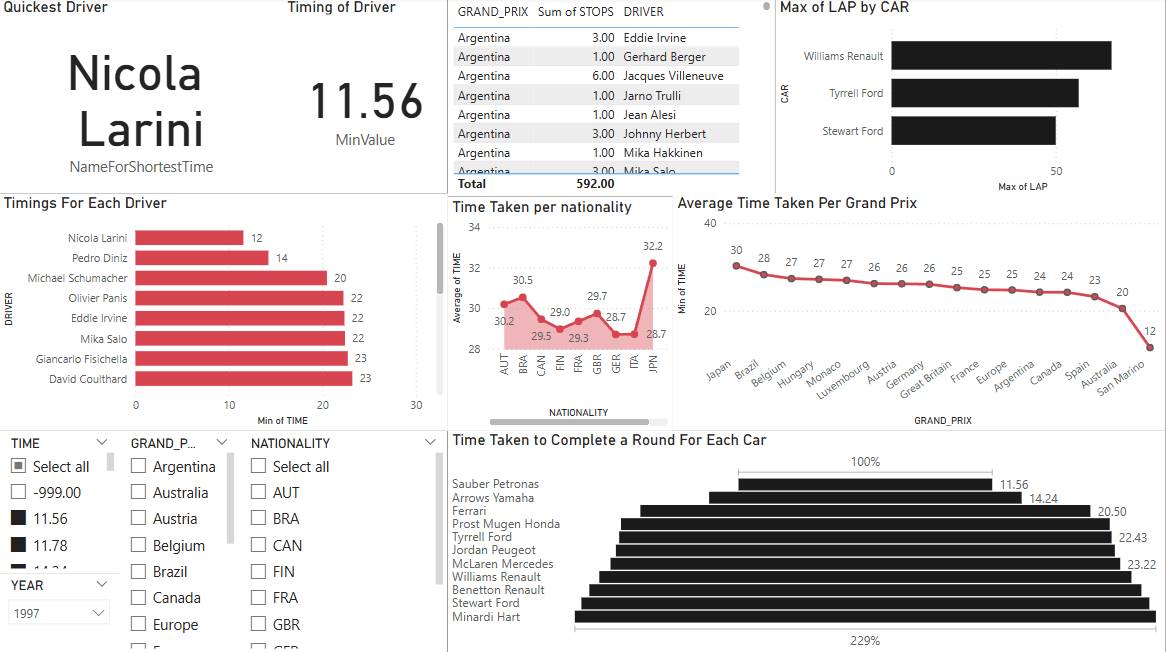


*Figure 5.1 Image of how Snowflake works with Git repositories*

# 6. Dashboard Insights

## a. Driver Table Dashboard

### i. Dashboard Overview

*Screenshot of Driver Table Dashboard*

The Driver Table Dashboard provides a comprehensive overview of key driver performance metrics, allowing users to analyse data filtered by year, Grand Prix, and time. Below is a breakdown of the dashboard components and their functions:

1. **Filters (Bottom Left)**

The filters allow users to narrow down the data based on specific parameters such as year, Grand Prix, and time values. The time filter is primarily used to remove extreme values that were imputed due to missing data. The Grand Prix and year filters help users focus on specific events and time periods, improving the clarity and relevance of the data.

1. **Driver Performance Card (Top Left)**

The card at the top left highlights the drivers with the fastest lap times based on the applied filters. This provides a quick snapshot of the highest-performing drivers, helping users identify key trends.

1. **Driver Timing Card (Middle right of the page)**

Positioned beside the driver performance card, this card displays the specific timing data for the selected driver, offering insights into individual performance.

1. **Average Time Line Chart (Middle right of the page)**

The purpose of this visual was to show what the average of the time taken for each of the grand prix so that it sorts of sets a benchmark for the drivers. The filters can be used to select which grand prix that the users want to see, alternatively clicking on the markers itself.

1. **Laps Completed Horizontal Bar Chart (Top Right)**

This horizontal bar chart displays the number of laps completed by each car, providing a visual comparison of the drivers' race participation and consistency. The graph has been filtered to the top 3 drivers so that it would be easier to see what the top 3 leading drivers are in this situation.

1. **Circuit Time Vertical Bar Chart (Bottom Right)**

The funnel bar chart is to show the time taken for each of the drivers to complete each lap. This can give insights as to how the driver is performing individually as well as how they are performing compared to other drivers.

1. **Driver Timings Horizontal Bar Chart (Left middle of the page)**

The horizontal bar chart on the right shows the individual lap timings for each driver, offering a detailed comparison of driver performance across races.

Together, these components provide a detailed, interactive view of driver performance, enabling users to easily explore and analyse data for better decision-making.

1. **Time taken to complete a lap per nationality**

This visualisation is just to show which country has the best racers and their nationalities. Some countries may produce strong drivers, which can suggest strong drivers in that nation in the future as well.

### ii. Dashboard Key Metrics

* Most of the visualisations in the dashboard leverage the timing column, providing a clear understanding of how drivers and their vehicles perform across different races and conditions.
* By showcasing the quickest drivers for each filter, the dashboard enables data analysts to draw meaningful insights, such as identifying the top-performing cars and assessing which drivers demonstrate better vehicle control. These insights help analysts pinpoint key patterns and trends that can inform strategies or decisions.
* Through this dashboard, data analysts can gain a deeper understanding of both driver performance and vehicle efficiency. The use of timing data in various visualisations provides a comprehensive view of how different factors contribute to race outcomes, enabling more informed decision-making and strategic planning.

### iii. Dashboard Business Value

The dashboard enhances the ability to analyse and track the performance of F1 drivers through intuitive visuals and filter options. By providing clear insights into driver performance across various filter values, it offers valuable information for both the F1 team and the wider community.

* **Performance Analysis**

The dashboard’s use of timing data, detailing the time taken by each driver to complete the Grand Prix and laps, allows for comprehensive performance analysis. By analysing these timings, teams and analysts can identify areas where drivers excel or face challenges, enabling the development of more effective race strategies and performance improvements. Patterns and weaknesses that emerge under specific conditions can be addressed to optimise future performance.

* **Audience Engagement**

The dashboard also opens up valuable data to F1 fans, providing them with a direct view of their favourite drivers’ race timings. It empowers fans to track historical performance and compare drivers’ achievements, increasing engagement with the sport. Fans can better understand the intricacies of races and become more involved in driver progress.

* **Driver Accountability**

By sharing performance data with drivers, teams can hold drivers accountable for their performance. The dashboard provides drivers with a clear overview of their own data, enabling them to pinpoint areas for improvement, such as specific corners or aspects of road management. This self-awareness is essential for drivers to refine their techniques, make adjustments, and ultimately enhance their race performance.

### iv. Dashboard Unique Insights

The dashboard offers several unique insights that can significantly impact both race strategies and driver performance.

* **Average Time Per Grand Prix**

The line chart displaying the average time taken for cars to complete the Grand Prix offers teams a clear indication of the typical performance for each race location. This helps in forecasting potential outcomes for future races at the same circuit, providing teams with a benchmark to guide their preparations. It also serves as a performance metric to assess if the teams are improving their lap times over time.

* **Minimum Timing Per Lap**

The dashboard also highlights the fastest lap times recorded in each Grand Prix. This gives teams and drivers a benchmark of peak performance, showing the best achievable time for a specific track. Drivers can use this as a target to aim for, fostering a competitive spirit to improve their own lap times and strive for faster performances, potentially helping them to secure better positions in future races.

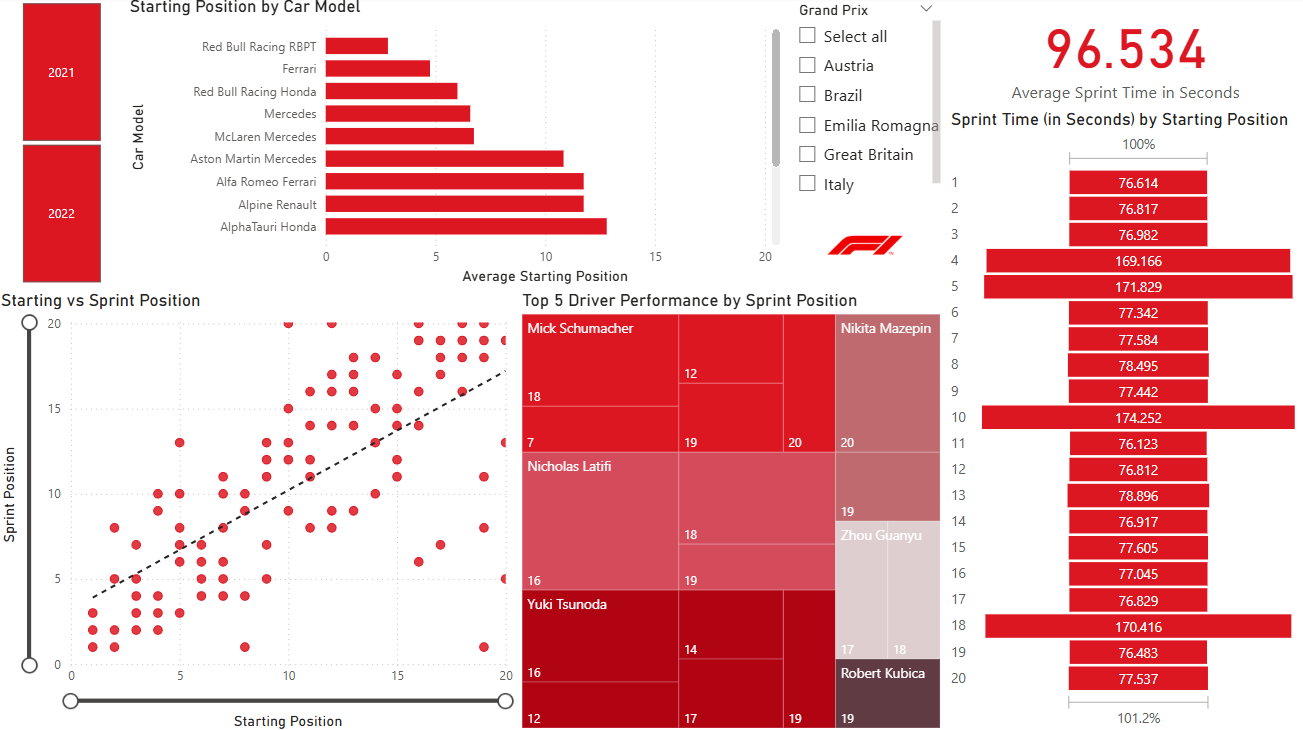
* **Competitive Insights**

The card located at the top left corner of the dashboard displays the driver who took first place in the race. This feature adds a layer of competitiveness to the dashboard, motivating drivers to improve their times in order to claim the top spot. The visual prominence of this card encourages a sense of achievement for the top performers and drives others to enhance their timing, pushing them to strive for better results in subsequent races.

Together, these unique insights provide F1 teams, drivers, and fans with valuable performance data, fostering both strategic improvements and competition.

## b. Grid Performance & Sprint Trends

### i. Dashboard Overview

*Screenshot of Grid Performance & Sprint Trends dashboard*

This dashboard is designed to support F1 Sports Data Scientists (SDS) in analysing the relationships between starting positions, sprint positions, and recorded times across different races and seasons. By uncovering performance trends, the dashboard provides data-driven insights into how various factors influence race outcomes.

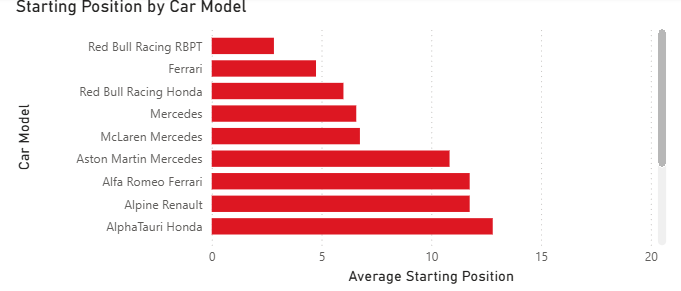
It features five key visuals that offer a comprehensive view of car model trends, sprint dynamics, and driver performance:

* **Clustered Bar Chart:** Starting Position by Car Model.
* **Scatter Plot:** Starting vs Sprint Position.
* **Treemap:** Top 5 Driver Performance by Sprint Position.
* **Card:** Average Sprint Time in Seconds.
* **Funnel Chart:** Sprint Time (in seconds) by Starting Position.

To enhance usability, slicers for Year and Grand Prix allow for custom filtering, enabling deeper exploration of performance patterns across different contexts.

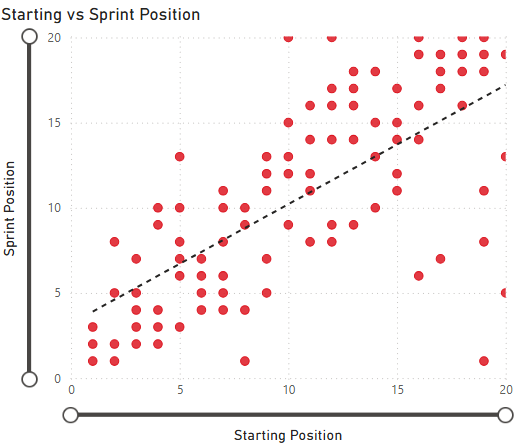
### ii. Dashboard Key Metrics

1. **Starting Position by Car Model**

*Screenshot of clustered bar chart on average starting position by car model*

This chart shows the average starting position for different car models, providing insights into how different teams perform in qualifying.

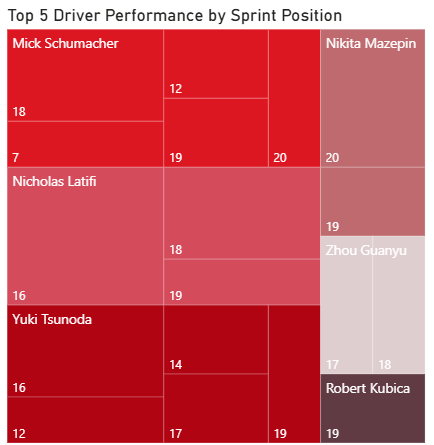
1. **Starting vs Sprint Position**



*Screenshot of scatter plot on starting position against sprint position*

This visual highlights how starting positions relate to sprint positions, allowing strategists to assess the impact of a driver’s starting position on their sprint performance.

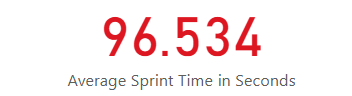
1. **Top 5 Driver Performance by Sprint Position**



*Screenshot of treemap on the top 5 drivers’ sprint positions and respective counts*

The treemap visualises the performance of the top 5 drivers based on their sprint positions, revealing which drivers consistently perform well during sprints.

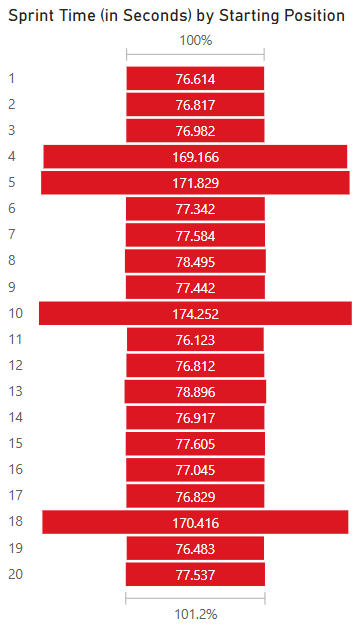
1. **Average Sprint Time in Seconds**



*Screenshot of card on average sprint time in seconds*

This card displays the overall average sprint time in seconds, summarising the data for quick analysis.

1. **Average Sprint Time by Starting Position**



*Screenshot of funnel chart on average sprint time (in seconds) by starting position*

This visual illustrates the average sprint time based on the starting position, offering insight into how starting grid placement affects the efficiency of sprint performances.

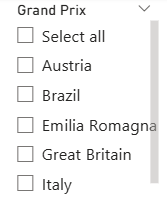
1. **Year Slicer**



*Screenshot of slicer on year*

This slicer allows users to filter the data by year, making it easy to compare trends across multiple seasons.

1. **Grand Prix Slicer**



*Screenshot of slicer on Grand Prix*

This slicer enables filtering by the Grand Prix event, providing flexibility to analyse data for specific races.

### iii. Dashboard Business Value

The dashboard plays a crucial role in meeting the analytical needs of F1’s data team, particularly their SDS. By consolidating key metrics into a single, interactive platform, it enables data-driven analysis of starting positions, sprint positions, and recorded times across different seasons and races. These insights help uncover performance patterns, providing valuable information to teams, broadcasters, and analysts.

The slicers allow for quick comparisons between different years and Grand Prix events, enabling deeper exploration of trends. Additionally, the integration of various visualisations — including the scatter plot, clustered bar chart, and funnel chart — offers a comprehensive view of driver performance, car model efficiency, and sprint time dynamics.

By leveraging this dashboard, SDS can provide fact-based insights that contribute to the continuous refinement of F1 performance analytics, competition fairness, and strategic decision-making at an organisational level.

### iv. Dashboard Unique Insights

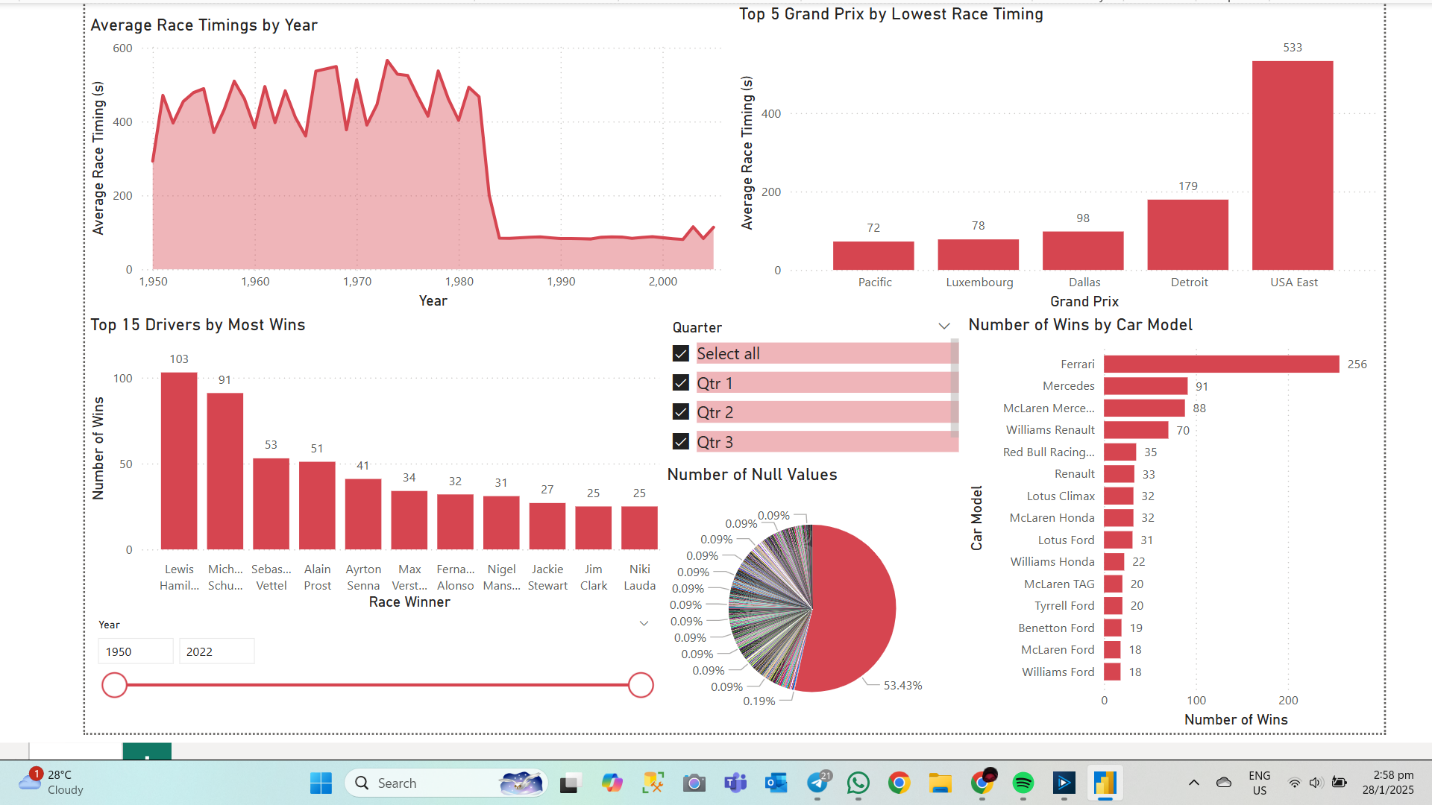
The Grid Performance & Sprint Trends dashboard has uncovered several unique insights that can contribute to F1’s data-driven evaluations of grid positioning and sprint performance:

* **Impact of Starting Position on Sprint Performance:** The scatter plot with a trend line reveals a clear correlation between starting and sprint positions. While lower grid positions generally lead to better sprint results, certain cases show drivers outperforming expectations from higher positions. This suggests that driver skill, race strategy, and track conditions play key roles beyond just starting placement.
* **Car Model Performance Trends:** The clustered bar chart highlights how different car models tend to perform in qualifying, offering insights into which teams consistently start in stronger positions. This data can help F1 analysts assess vehicle development trends and competitiveness across seasons.
* **Sprint Time Efficiency:** The funnel chart shows that drivers starting in front generally achieve lower sprint times, reinforcing the advantage of a strong qualifying position. However, some exceptions suggest that certain drivers or teams may optimise sprint phases differently, providing room for deeper analysis.
* **Top Driver Performances:** The treemap, filtered for the top five drivers by sprint position, helps identify consistent high performers in sprint races. This insight can support historical comparisons, driver evaluations, and predictive analytics for future sprints.

These insights provide F1’s SDS with an analytical foundation for evaluating grid positioning trends, sprint efficiencies, and car model impacts. By leveraging this data, F1 can refine benchmarking strategies, track development progress, and generate insights that inform future performance evaluations across seasons.

## c. Race Performance & Strategy

### i. Dashboard Overview

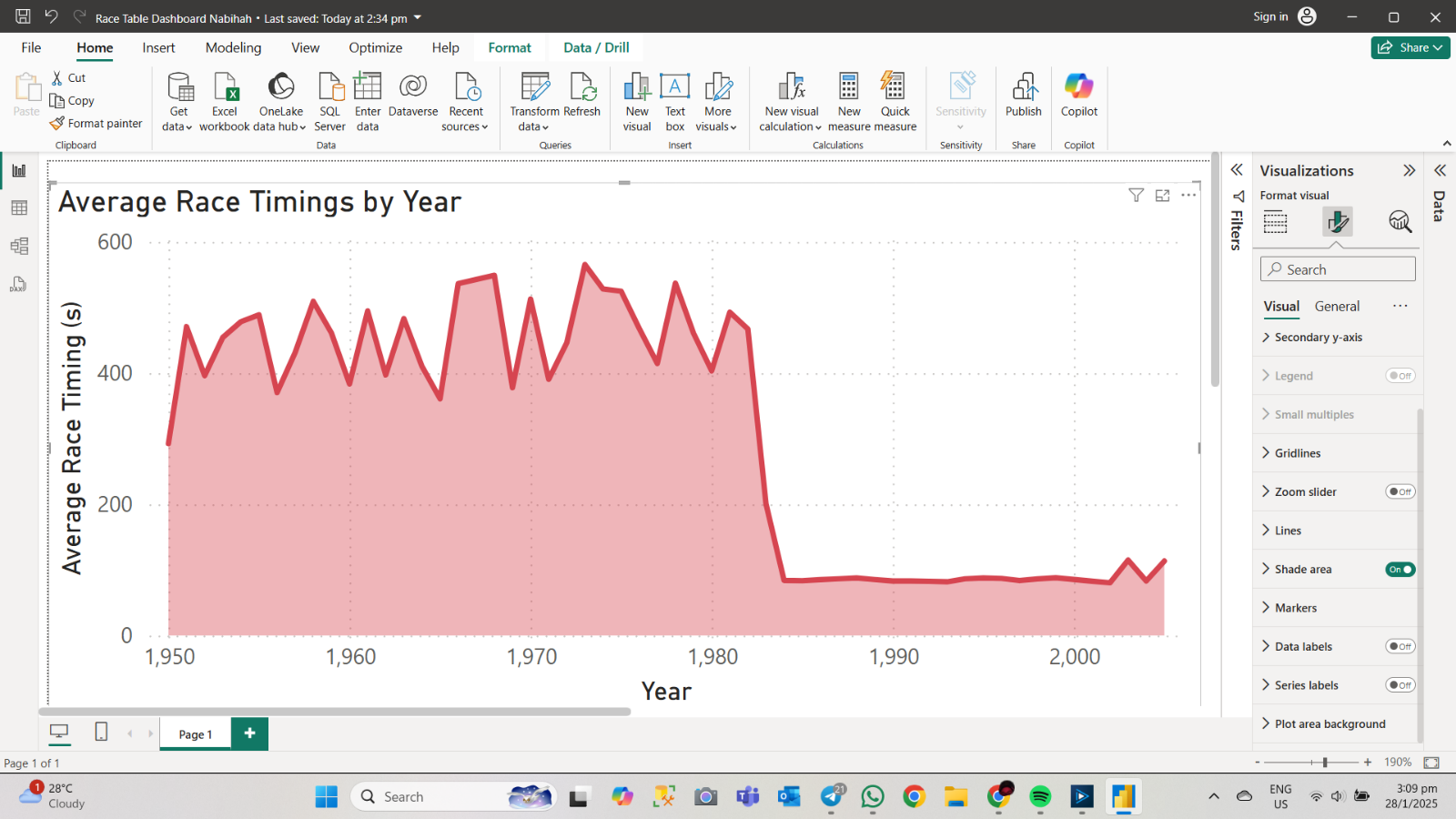


*Screenshot of Race Table Dashboard*

Designed for Motorsport Strategists, this dashboard provides a data-driven analysis of historical race performances, aligning with the user story. It enables the identification of trends in race timings, evaluation of driver achievements, and assessment of car model impact on race outcomes. By exploring key metrics over time, strategists can optimize race strategies and enhance team performance. Interactive filters, including a quarter slicer and a year range slider (1950-2022), allow for tailored insights into specific seasons and competitors.

### ii. Dashboard Key Metrics

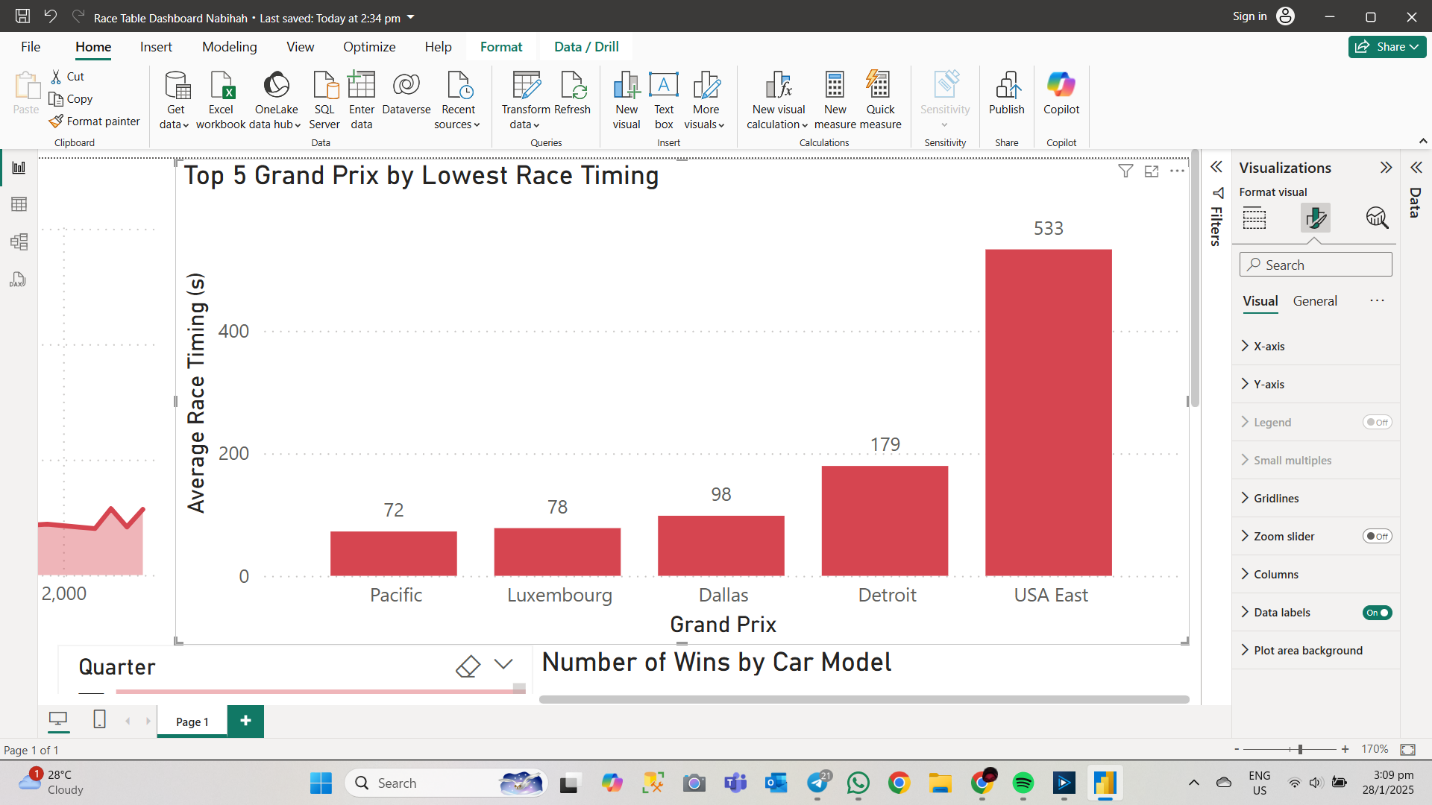
1. **Average Race Timings by Year**



*Screenshot of area chart on average race timings by year*

* 1. Displays fluctuations in race durations from 1950 to 2022.
  2. Key metric: Average race timing (seconds) over the years.
  3. Insight: Significant drop in average race times post-1980, indicating possible regulatory or technological advancements.

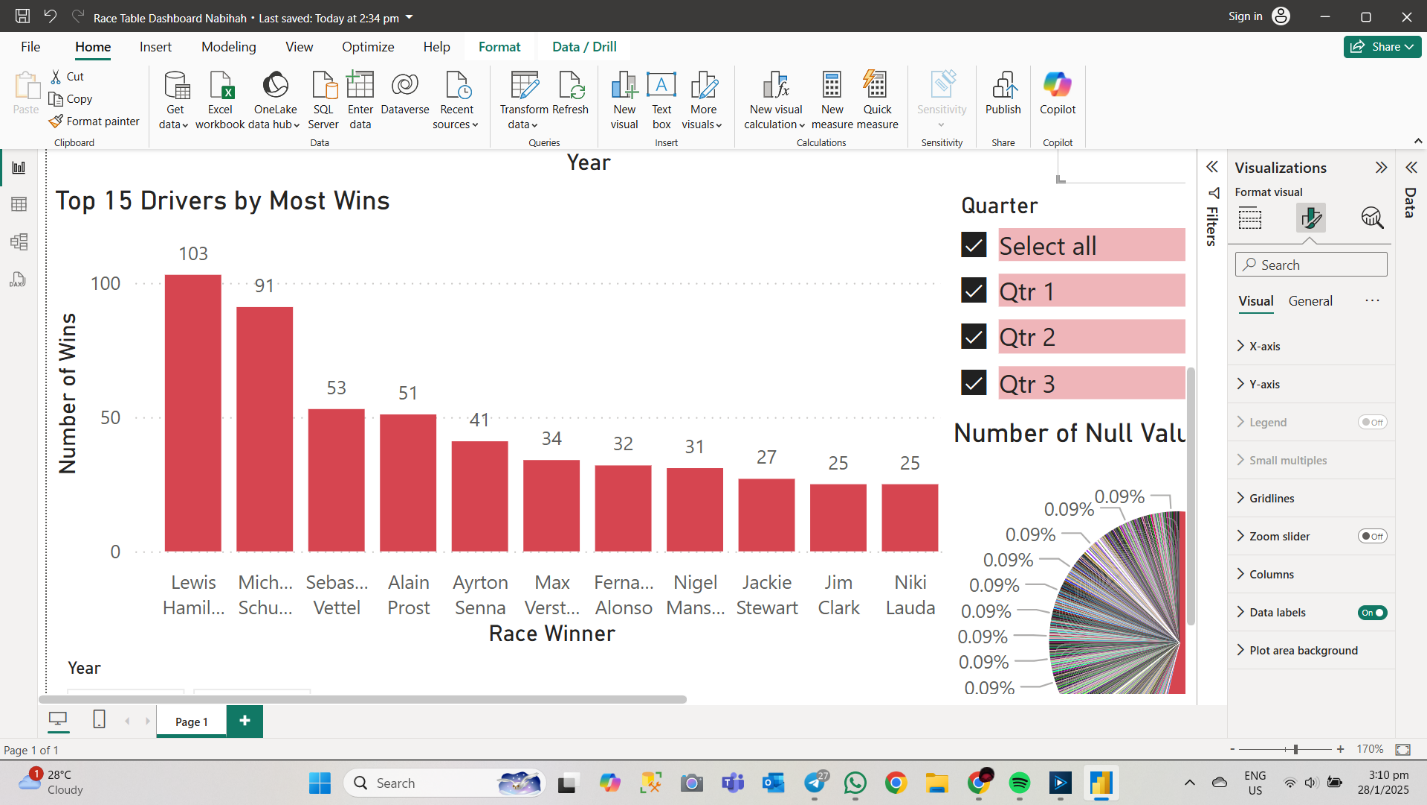
1. **Top 5 Grand Prix by Lowest Race Timing**



*Screenshot of clustered column chart on top 5 Grand Prix by lowest race timing*

* 1. Highlights the Grand Prix events with the shortest average race times.
  2. Key metric: Average race timing for each Grand Prix.
  3. Insight: USA East had the lowest recorded race time, significantly shorter than others, suggesting track characteristics or specific race conditions played a role.

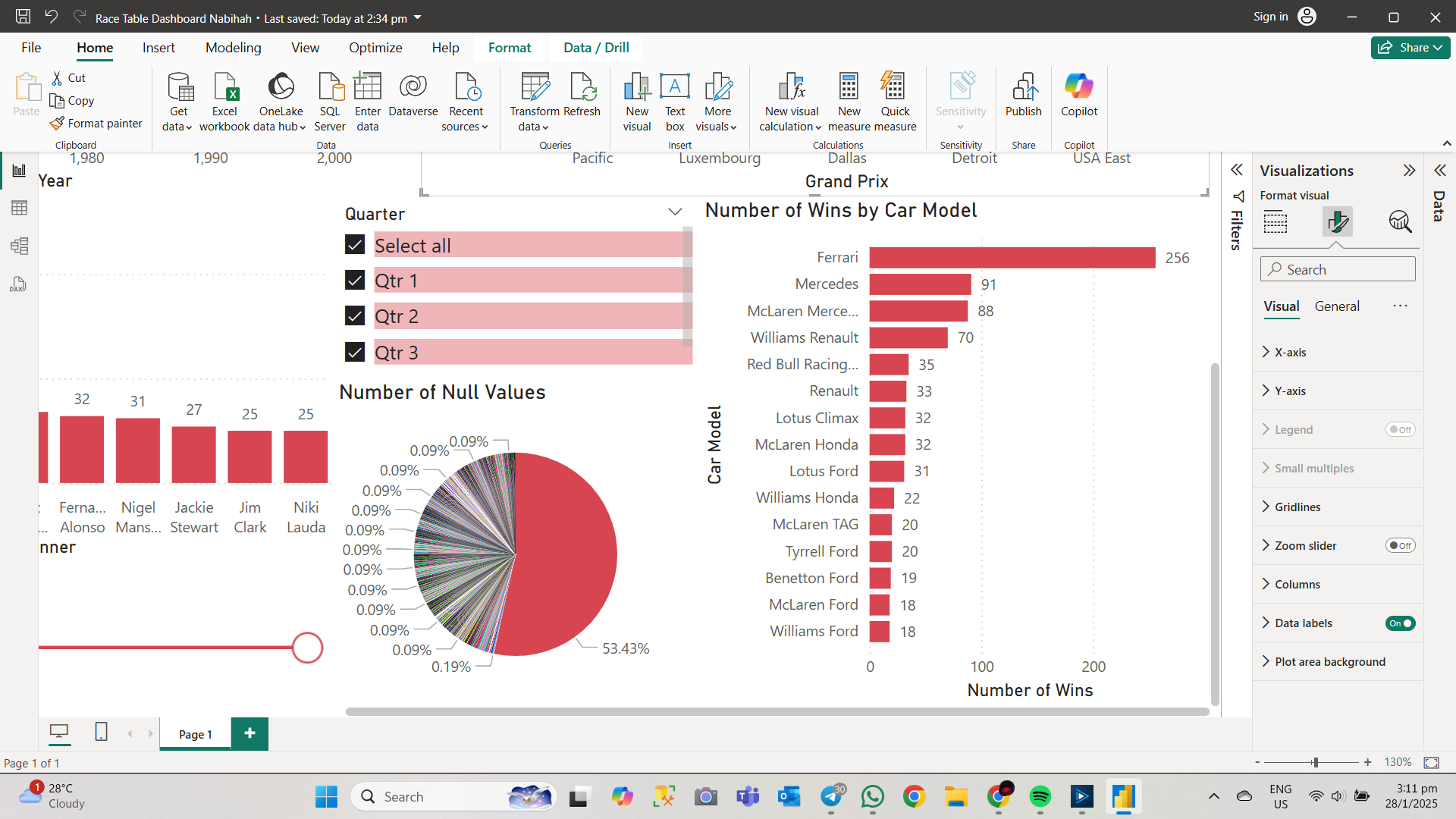
1. **Top 15 Drivers by Most Wins**



*Screenshot of clustered column chart on top 15 drivers by most wins*

* 1. Showcases the most successful drivers in terms of race victories.
  2. Key metric: Number of wins per driver.
  3. Insight: Lewis Hamilton and Michael Schumacher dominate with 103 and 91 wins, respectively, highlighting their prolonged success.

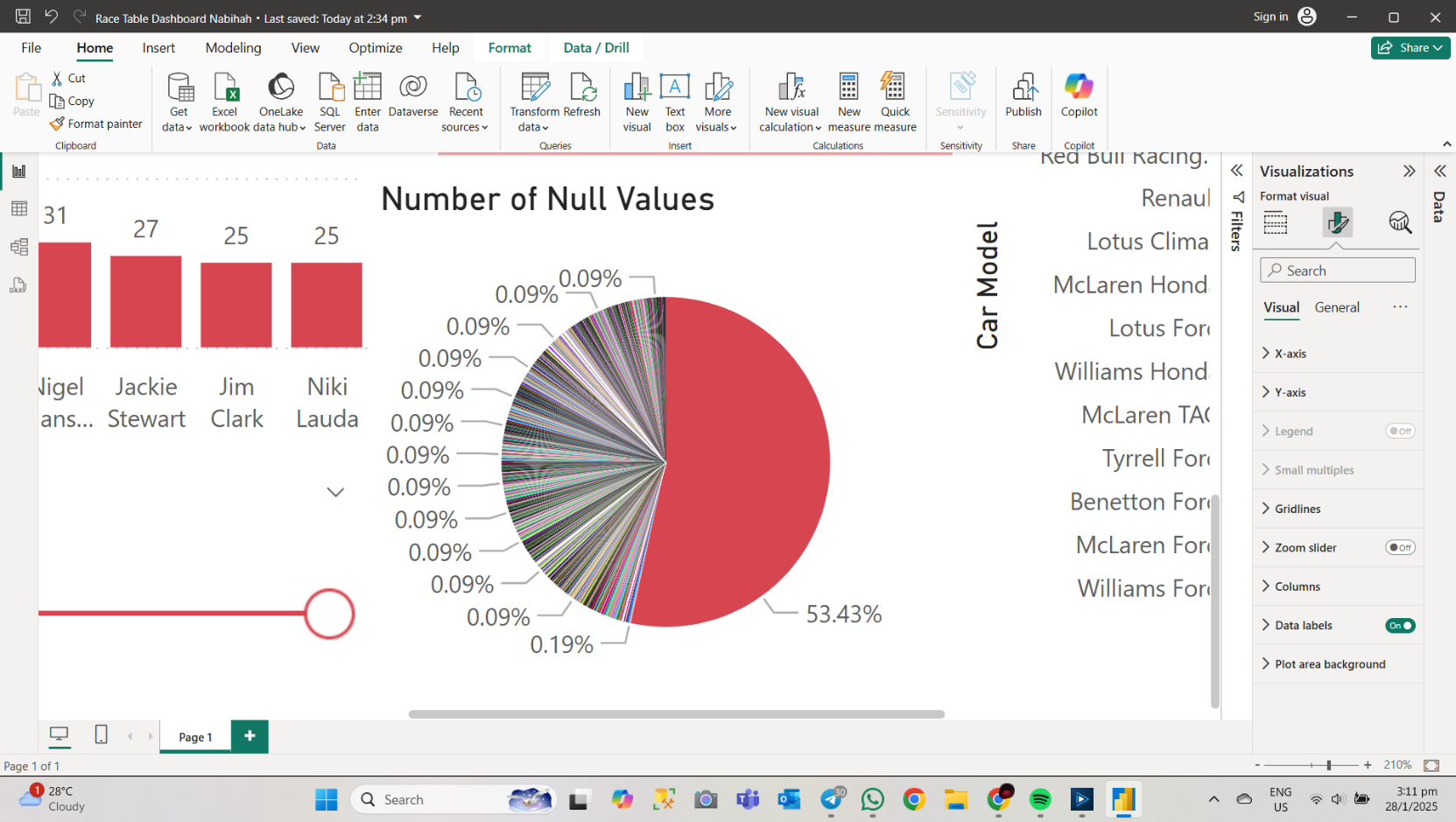
1. **Number of Wins by Car Model**



*Screenshot of clustered bar chart on number of wins by car model*

* 1. Compares the performance of various car manufacturers based on total race wins.
  2. Key metric: Number of wins by car model.
  3. Insight: Ferrari leads with 256 wins, far surpassing competitors, showing its dominance in race history.

1. **Number of Null Values**



*Screenshot of pie chart on number of null values in RACE\_TABLE*

* 1. Provides an overview of missing data in the dataset.
  2. Key metric: Percentage of null values across attributes.
  3. Insight: A significant portion (53.43%) of the dataset consists of null values, which may impact the completeness and reliability of race performance analysis. Understanding where these gaps occur is essential for accurate trend evaluation.

### iii. Dashboard Business Value

This dashboard equips Motorsport Strategists with actionable insights to optimize race strategies. By analysing historical trends, car manufacturers can assess model performance, race organizers can refine regulations, and teams can benchmark performance for future races. Understanding race timings, driver success, and car efficiency enables data-driven decisions to enhance competitive edge. It helps in:

* **Strategic Decision-Making:** Car manufacturers can leverage insights into car model performance and driver success to inform future investments and racing strategies.
* **Performance Benchmarking:** Race organisers can identify patterns in race timings and track efficiency, enabling improvements to event organisation and race regulations.
* **Historical Analysis:** By analysing past trends in race times and winners, stakeholders can better predict future outcomes and optimise race conditions for more competitive events.

### iv. Dashboard Unique Insights

**Significant Presence of Null Values**

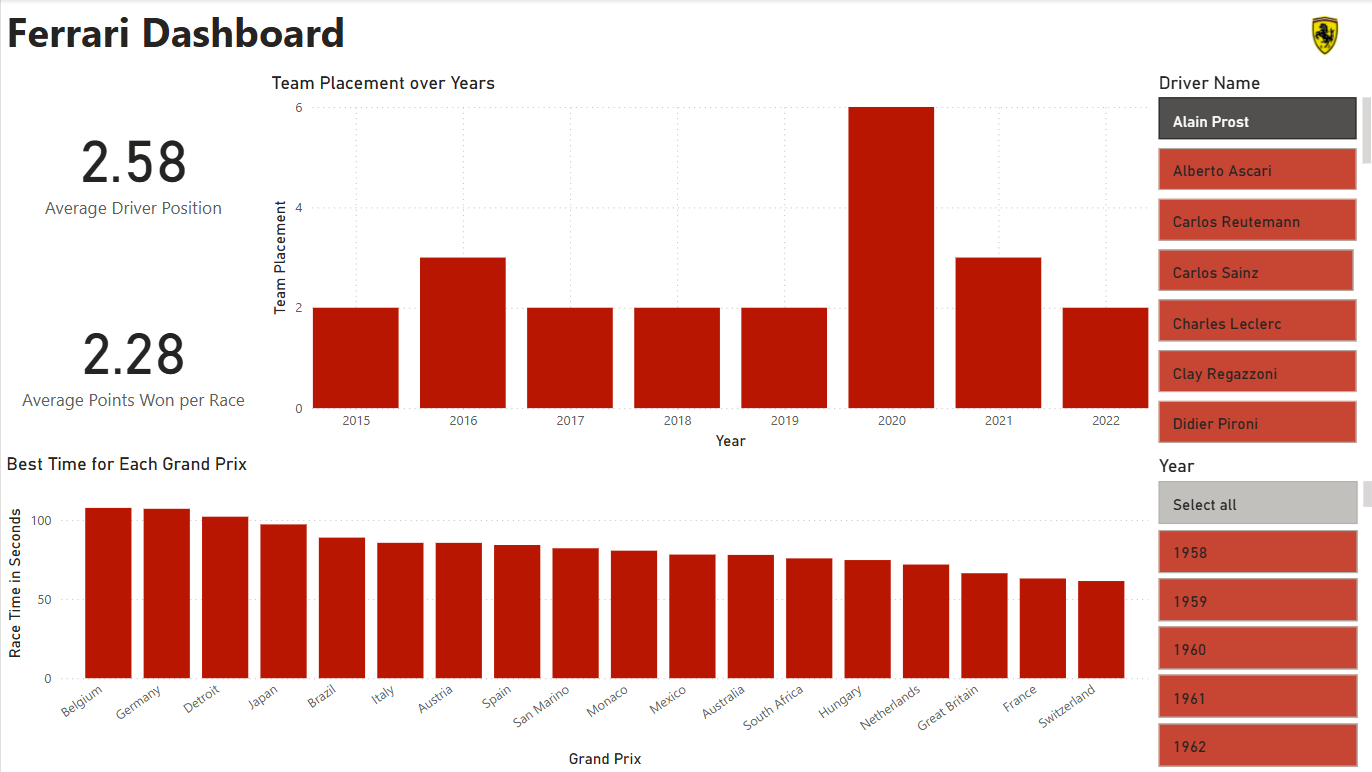
* The null value chart reveals that a substantial portion (53.43%) of the dataset contains missing values, which could impact the accuracy of race performance analysis.
* The high null percentage suggests potential gaps in historical data collection, inconsistent reporting standards over different racing eras, or missing information for specific variables (e.g., race timings, driver details).
* If the missing data is concentrated in specific time periods or attributes, this could affect trend analysis and decision-making. Further investigation into the cause of these null values is crucial for data reliability.

**Drastic Drop in Race Timings Post-1980**

* While technological advancements are an obvious factor, the sudden and sustained drop suggests that other external influences, such as regulatory changes, race format alterations, or fuel/pit stop rule modifications, played a significant role.
* This pattern could indicate a major shift in racing strategy, such as shorter track lengths, improved aerodynamics, or optimised race planning by teams.
* Investigating this further could help understand how external factors beyond just car performance contribute to faster races, providing valuable insights for race organisers and engineers.

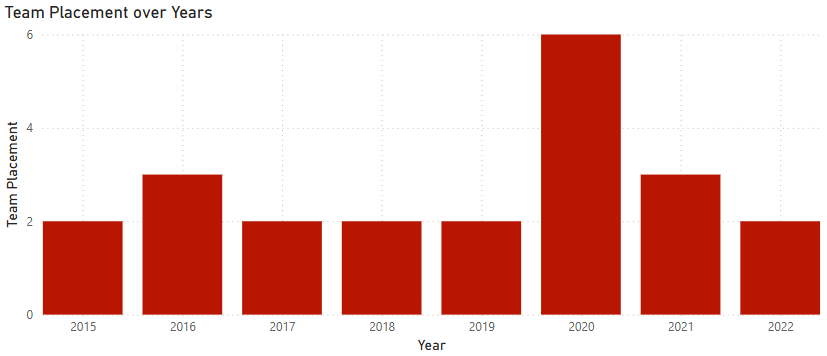
## d. Team Table Dashboard

### i. Dashboard Overview

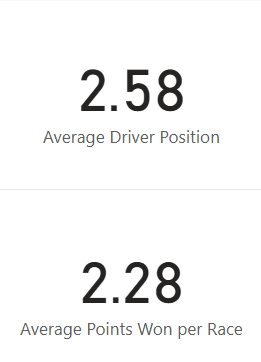
*Screenshot of Ferrari Dashboard*

The dashboard shows how a team can leverage the data to find insights on their overall team metrics and individual player metrics. Focusing on one team, Ferrari, they can use these key metrics and filters to see which players they want to keep or drop and compare their best times for each race to help drivers improve their own racing times.

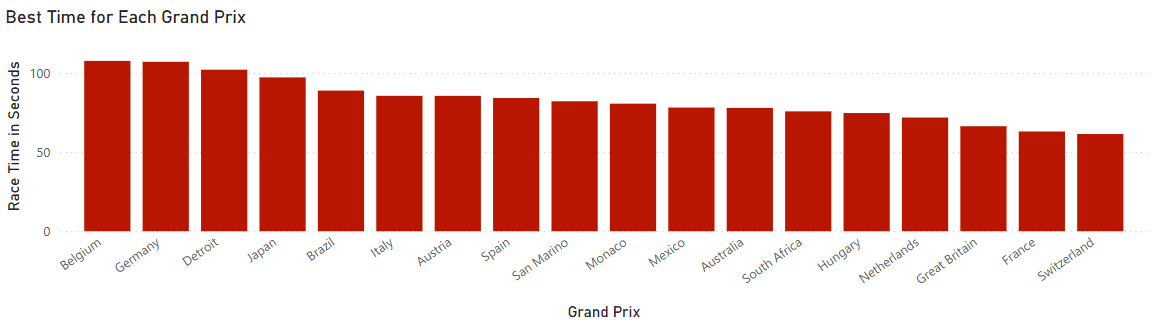
### ii. Dashboard Key Metrics

*Screenshot of Team Placement over Years*

Looking at the overall team placements in the last 8 years, we can see that Ferrari in recent years has been doing extremely well, constantly placing between 2nd and 3rd place, with the outlier year being 2020 where they placed 6th. Since the Ferrari players are doing extremely well, rather than finding new players and having to train, I recommend extending their contracts and finding ways to improve their times on the track.



*Screenshot of Average Driver Position and Average Points Won per Race*

*Screenshot of Best Time for Each Grand Prix*

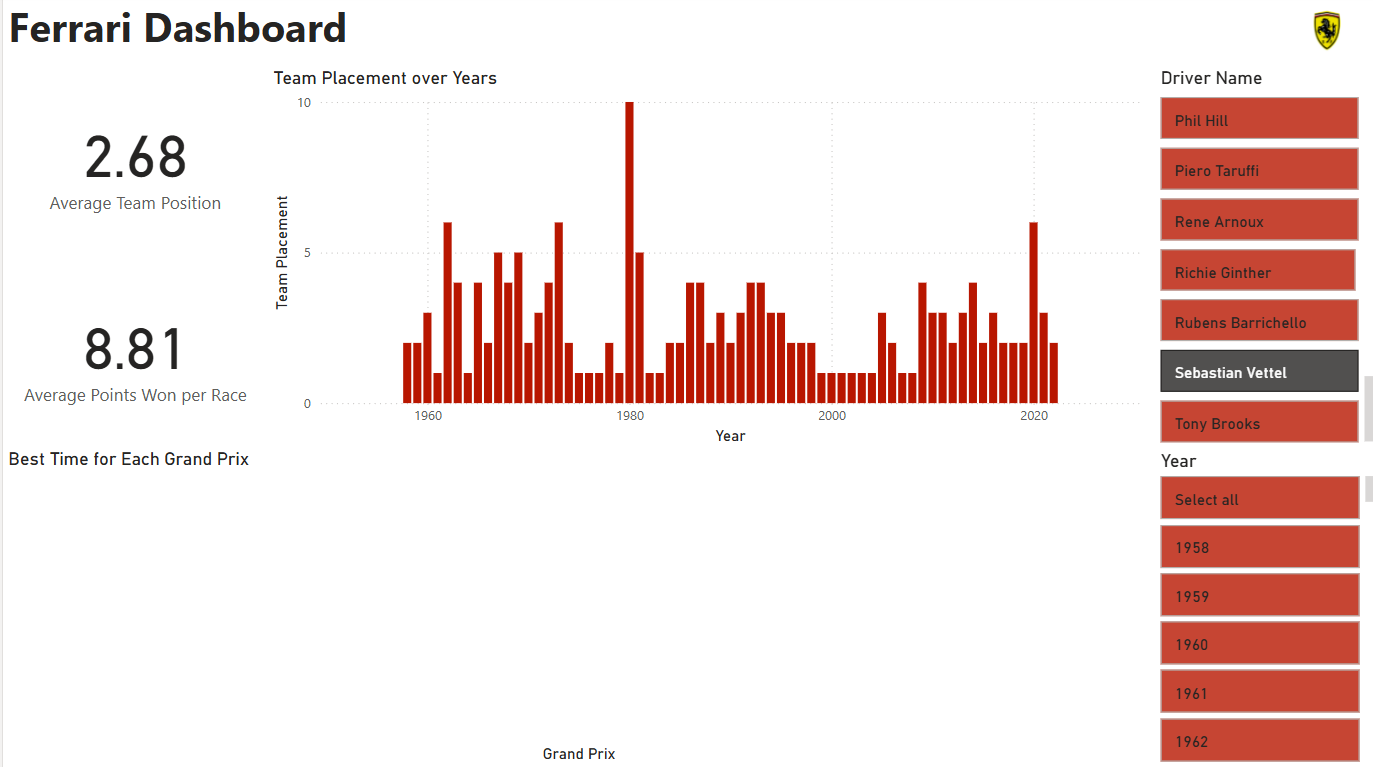
Filtering for the driver, Alain Prost, we can see that he is an excellent player, allowing the team to place with an average positioning between 2nd and 3rd and winning them an average of 2.28 points per game. This was significant as he played during 1980 to 1993, where the allocation of points was much lower. Looking at the best times for each Grand Prix, we can see which Grand Prix could be improved, which could have helped Alain Prost focus and train on those tracks more often.

### iii. Dashboard Business Value

This dashboard delivers valuable insights into the Ferrari team and their drivers' performance over decades, enabling stakeholders (e.g., Ferrari’s management, the racers and sponsors) to understand key trends and areas of improvement.

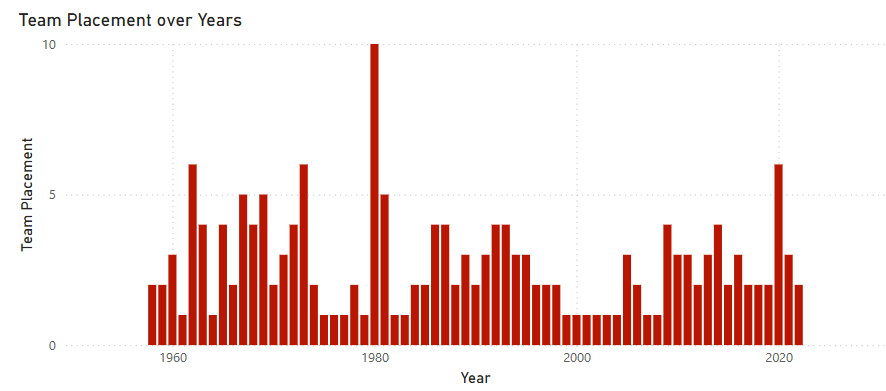
* **Strategic Decision-Making:** Ferrari’s management can choose to drop and find or extend players contract depending on their performance on the track. This allows them to have more time to find better drivers and make offers to them before other teams.
* **Performance Optimisation:** Drivers can see which times can be improved based on their best times and allocate my time into practicing the tracks they struggle with.
* **Historical Analysis:** By analysing past trends in the team statistics, stakeholders can better predict future outcomes and choose to continue to invest in the team or pull out.

### iv. Dashboard Unique Insights

*Screenshot of Ferrari Dashboard when filtered to a newer player*

**Significant Presence of Null Values**

There is a significant presence of null values for the recent players, where their best times were not recorded. For example, Sebastain Vettel, a racer who played from 2007 to 2022, does not have any of his best times recorded despite being a well-known great racer.



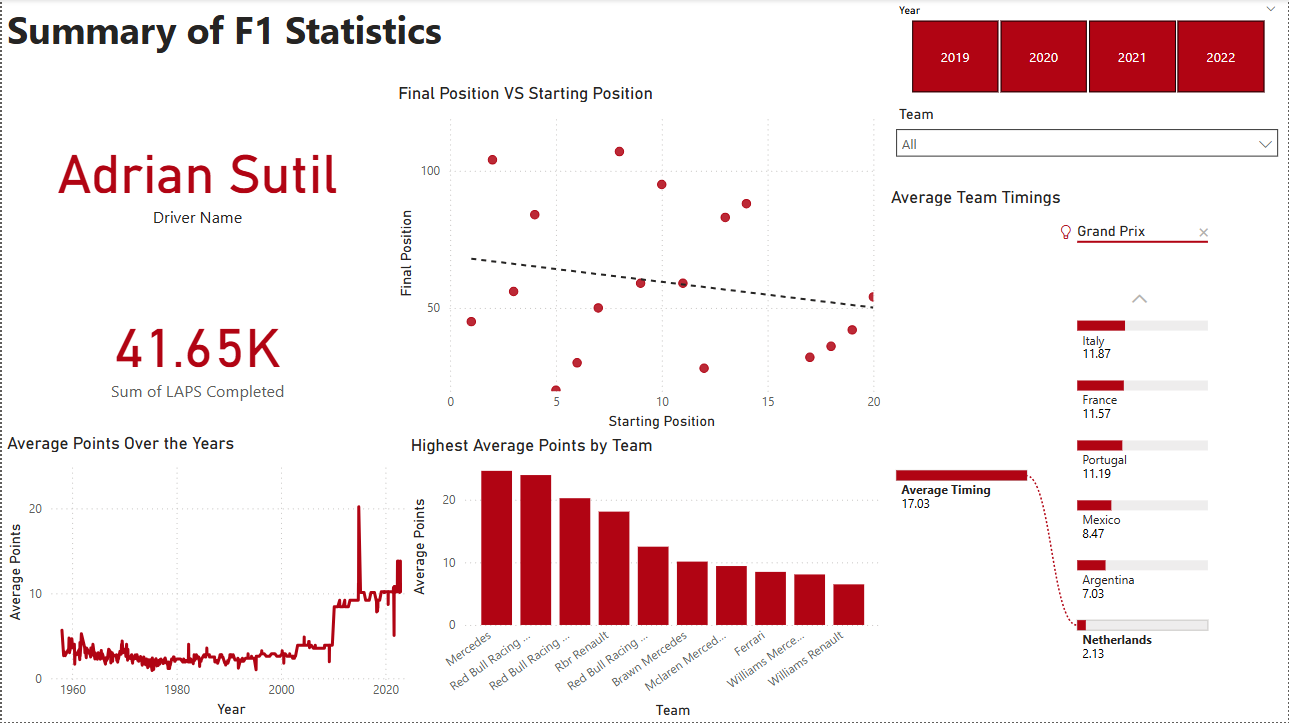
*Screenshot of Team Placement over Years*

**Historical Data**

Looking at the overall team placement over the years, Ferrari has been doing extremely well, usually placing between 1st to 5th with a few years being the exception. This shows that Ferrari is an extremely strong team with relatively good consistency in their placements.

## e. Summary of F1 Statistics

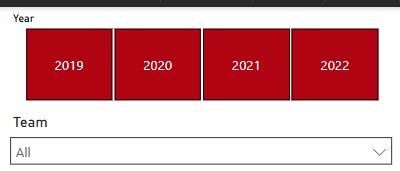
### i. Dashboard Overview

*Screenshot of Dashboard*

The dashboard above is to show, a general overview of the statistics for Formula 1 races in general. This dashboard encompasses the usage of all the datasets in the database. This dashboard is meant to show a general overview of statistics catered for teams in Formula 1 and management. This is to show general trends and patterns for the F1 sphere.

### ii. Dashboard Key Metrics

**Slicers Used**

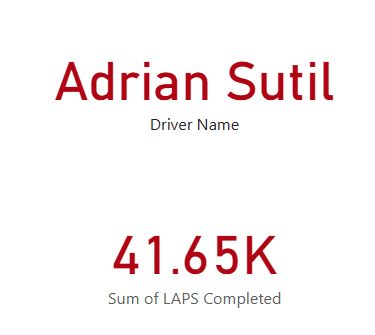


*Screenshot of Slicer*

The image shows two slicers that were used for the: the year slicers and the team slicer. The year slicer only shows the last 4 years for the from the current year. This is to aid in convince. Since it shows the 4 most recent years.

As for the team slicer, it is a drop-down slicer to show the teams that have participated in the races.

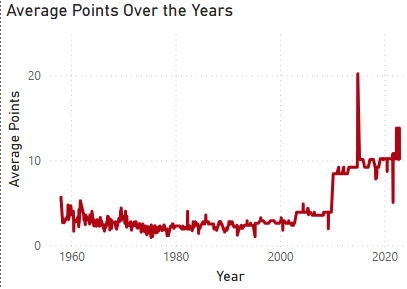
**Cards used**



*Screenshot of Card*

The card visuals here would show the driver that represented the team for that year and the total number of laps the driver. This to show the most experienced driver in the team.

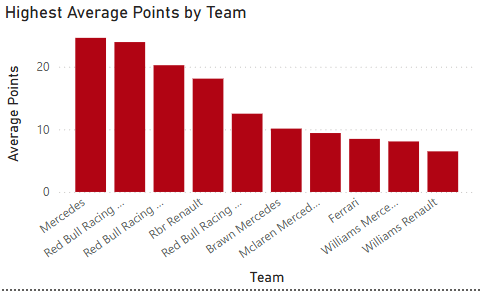
**Average Points over the Years**



*Screenshot of Line Chart*

The chart here is to show the average points per year. This have teams and team’s management to see the trends of the point scoring system. This can help team’s change their competition strategies while management are aware of the performances of the overall F1 sphere or the F1 drivers and pit crew.

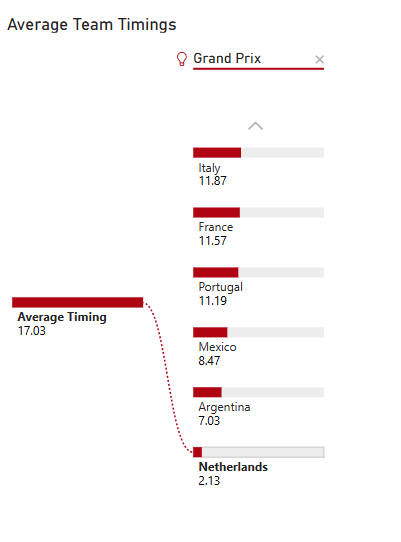
**Highest Average Point by Team**



*Screenshot of Bar Chart*

This bar chart is to show the highest average points by the teams. This chart allows for the management to show which teams are the best performing and which teams are needing improvement.

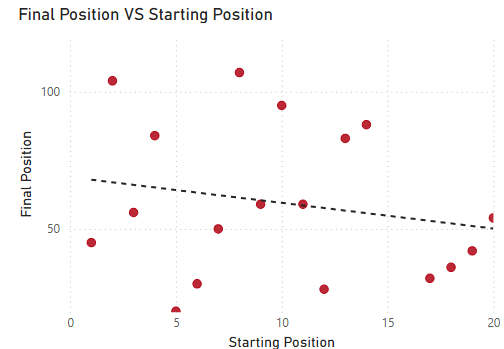
**Average Grand Prix timings**



*Screenshot of Decomposition Tree*

This decomposition tree is to show the average timings of each grand prix tracks. This is to help team benchmark timing when the driver does simulations when training. This can help drivers improve their skill on the track as well as the pit crew coordination for the races by help reduce timings for on track repairs.

**Final Position Vs Starting Position**



*Screenshot of Scatter Plot*

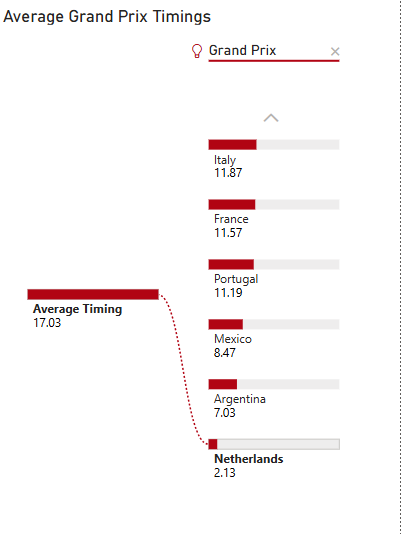
This chart is meant to show the correlation between the starting position and final position of the drivers. This is meant to show the correlation between how the starting position would affect the outcome of the driver’s position. Since final position affects the points of the team.

### iii. Dashboard Business Value

This dashboard provides a high business value for driver and teams and their management.

* Driver
  + Provides performance analysis for the drivers for them to improve and create strategies
  + Better performance leads to better brand deals since drivers would get better recognition
* Teams
  + The performance analysis of this dashboard aid in teams to create competitive strategies
  + The changes in competitive strategies would attract more spectators and viewership since viewers enjoy the competition.
* Management
  + The performance analysis of this dashboard would aid helping management changing staff members.
  + If drivers are not performing management would use the aid of this dashboard to help change drivers like team members in the pit crew.

### iv. Dashboard Unique Insights



*Screenshot of Decomposition Tree*

The decomposition tree would allow for the teams to be able to see the average timings for each track in each grand prix. Since F1 takes places in many different countries. The tracks will vary according to the country. During the times when F1 driver are not competing, simulations are used to help drivers practice improve their timings. However, some tracks are not record properly. The Italy has two different timings since the same track has changed the names, like other countries.

However, the usefulness of this chart gives a much deeper insight into the race statistics compared to the other charts. Since timings are essentially the cornerstone when it comes to the scoring system in the Formula 1 sphere.

# 7. Conclusion

## a. Final Results

The primary goal of this project was to modernise F1’s data architecture by migrating the company’s legacy on-premises database to a cloud-based platform. F1's original database system was faced with several pain points, including scalability issues, inefficiency in maintaining physical infrastructure, and a lack of flexibility in accessing and managing data remotely.

Project Objectives

* **Improve Scalability:** The on-site database struggled to handle the growing volume of data generated by F1’s events, requiring time-consuming manual intervention to expand capacity.
* **Increase Efficiency:** The operational costs and manpower required to manage the physical database system were unsustainable, which prompted the need for a more cost-effective solution.
* **Enhance Flexibility:** The limitations of the physical infrastructure made remote access difficult, which posed challenges to F1’s global operations and required a more flexible system.
* **Strengthen Analytics Capabilities:** With the increasing volume of data, there was a clear need for advanced tools that could offer deeper insights and support data-driven decision-making.

How the Project Addressed These Pain Points

1. **Cloud Migration for Scalability:** By migrating to Snowflake, a cloud-based platform, F1 now has access to virtually unlimited storage and compute resources, allowing the database to scale efficiently in line with the growing data requirements. This migration eliminated the need for time-intensive manual capacity expansions, enabling F1 to manage data with greater ease.
2. **Improved Operational Efficiency:** The shift to Snowflake has dramatically reduced the operational costs associated with maintaining on-premises infrastructure. The cloud environment is self-managing, reducing the manpower required for system updates, hardware maintenance, and overall management. This has led to long-term cost savings and greater operational efficiency.
3. **Increased Flexibility with Cloud Access:** The cloud-based system provides F1’s global team with 24/7 access to the database from any location, aligning with the international scope of F1 events. Employees no longer face the limitations of being tied to on-site infrastructure, allowing for quicker decision-making and collaboration across regions.
4. **Strengthened Analytics and Decision-Making:** With Snowflake’s advanced data warehousing capabilities, F1 can now leverage enhanced analytics tools to gain deeper insights into data. This increased accessibility and processing power have enabled the team to perform more complex analyses, empowering them to make faster, data-driven decisions.

## b. Business Impact

The migration to a cloud-based database has delivered substantial advantages to F1. Transitioning to Snowflake as the hosting platform has been instrumental in addressing F1’s data challenges and positioning the organisation for continued success in a data-driven industry.

Key benefits of this migration include enhanced data accessibility and significantly faster reporting times. The previous on-site database system restricted access to one user at a time and often required several seconds or minutes for data retrieval. Snowflake eliminates these issues by enabling simultaneous multi-user access and reducing retrieval times to just milliseconds, fostering improved collaboration and operational efficiency.

Snowflake’s use of SQL as the primary programming language ensures a seamless transition for employees accustomed to MySQL. This compatibility minimises the need for extensive retraining while reducing disruptions to ongoing operations. Additionally, Snowflake’s commitment to regular updates and innovations ensures that the platform remains cutting-edge, providing long-term productivity benefits to F1.

The platform’s direct integration with Power BI further enhances decision-making capabilities. This integration simplifies dashboard creation and accelerates data sharing, enabling stakeholders to quickly access critical insights and make informed decisions. This accessibility ensures F1’s operations remain agile and responsive to changing demands.

In summary, the migration to Snowflake has significantly improved F1’s data accessibility, reporting speeds, and decision-making capabilities. These advancements strengthen F1’s competitive edge, boost organisational productivity, and lay the foundation for sustained growth and innovation.

## c. Closing Thoughts

This project marks a significant milestone in the company’s transition to cloud-based database solutions. The successful implementation of Snowflake demonstrates the feasibility of migrating from on-premises databases to a scalable, cloud-based architecture. Given the increasing need for efficient data storage and management, this project serves as a foundation for further advancements in cloud data solutions, including leveraging Snowflake’s data warehouses and data lakes for future expansion.

In the short term, the success of this project is evident in its seamless integration with existing workflows. Since the database implementation is entirely SQL-based—from table creation to database maintenance—employees can transition to the new system with minimal disruption. This reduces training costs and ensures continuity in database operations.

Looking ahead, the ease of use provided by Snowflake, combined with the company’s current stage in digital transformation, will help delay any immediate need for further database transitions. The scalability and flexibility of Snowflake position the company well for future data growth, ensuring that its cloud infrastructure can evolve alongside its needs.

# 8. Appendix

## a. Challenges

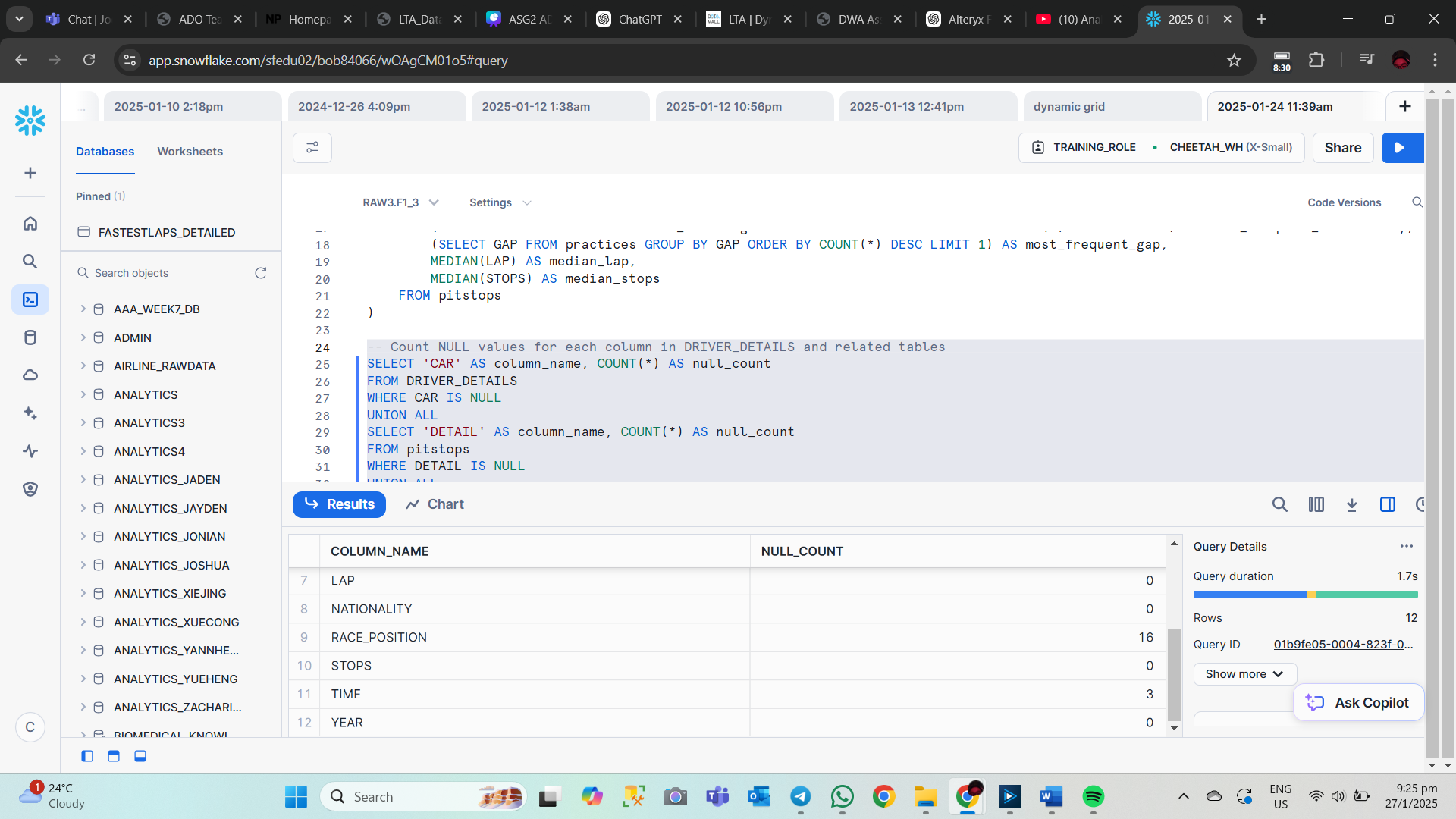
### i. Technical Challenges

During the implementation of this project, several technical challenges were encountered, particularly concerning data quality and GitHub integration.

* **Data Quality Issues**

One of the primary data quality challenges was the misalignment of information within datasets. Certain datasets contained incorrect or misplaced data, making it difficult to establish accurate relationships between tables. For instance, the driver code, which logically should have been stored in the driver\_details table, was instead found in the driver\_standings table. This inconsistency required adjustments in table joins and data mapping to ensure accurate integration.

Another critical issue was the prevalence of null values across multiple numerical and categorical columns, such as RACE\_POSITION and TIME. These missing values introduced inconsistencies, disrupted table joins, and affected the accuracy of statistical analyses. Without proper handling, they could have led to misleading insights.



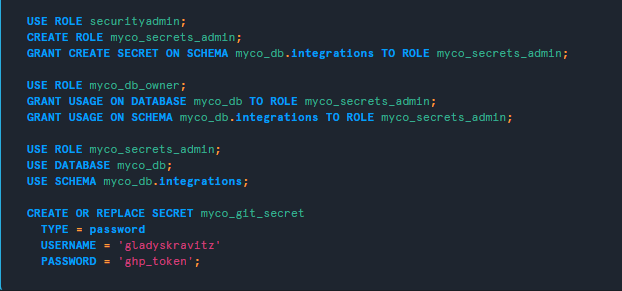
*Screenshot of count of null values*

* **GitHub Integration Issues**

The primary challenge with GitHub integration was the complexity and lack of clarity in the provided documentation. The instructions were difficult to follow, with unnecessary redundancies that added confusion rather than facilitating a smooth integration process.

One notable inefficiency in the documentation was the inclusion of multiple steps that were not essential for the actual integration. For instance, the documentation suggested creating a new role for GitHub integration, despite an existing role already having the necessary permissions. This redundancy led to unnecessary steps, complicating the implementation process.

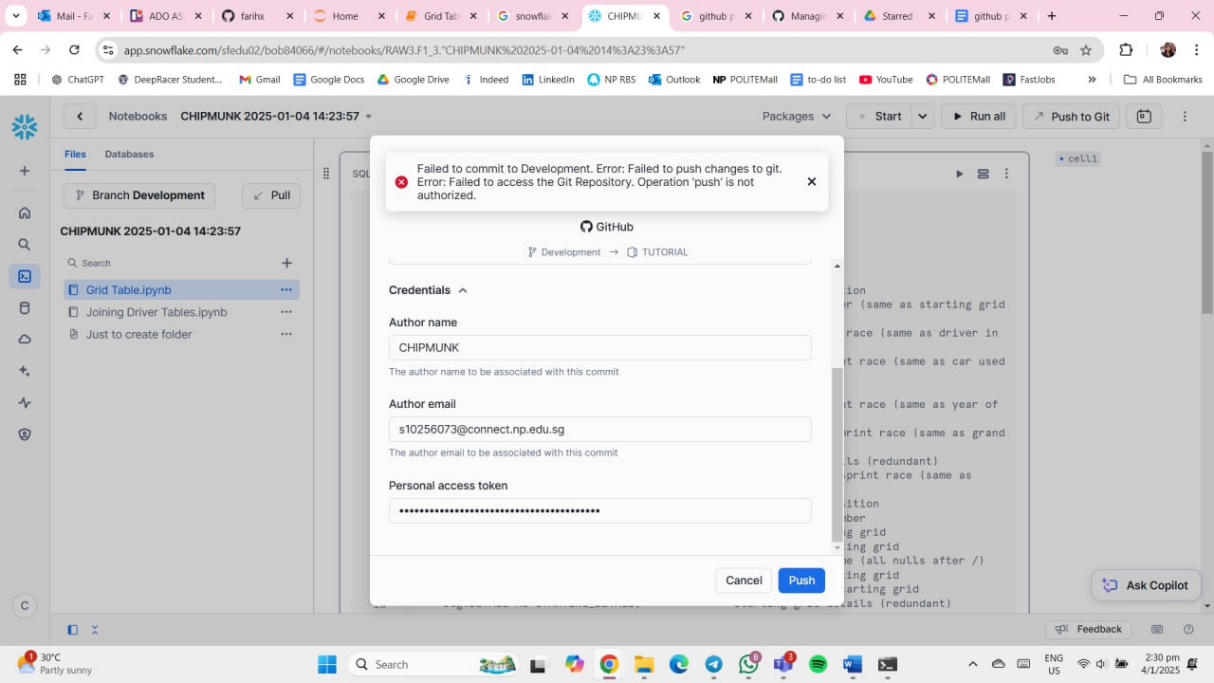
Additionally, the documentation provided multiple blocks of code for setting up authentication, but only the final block was truly required for creating a secret key for authentication. The presence of these extra steps resulted in wasted effort and unnecessary troubleshooting before identifying the minimal and effective approach.



*Figure 8.1 Image of redundant code creating a secret for GitHub integration*

Another issue encountered during GitHub integration was that re-running the integration code in Snowflake would cause the previously linked remote repository to become unlinked. This resulted in disruptions to the development workflow, as the established connection to the repository was lost each time the integration script was executed.

For example, when a notebook was created to link to one of the .ipynb files in the GitHub repository, executing the integration code caused Snowflake to unlink the repository. It also removed the previously granted role access for reading and writing to the repository. This meant that permissions had to be manually reassigned within Snowflake after each execution, adding unnecessary overhead to the process.



*Screenshot of a team member facing issues due to removal of role access from re-running the code*

### ii. Team Dynamics

Team dynamics played a crucial role in the success of this project. Effective collaboration and communication were key to ensuring smooth execution across all sprints. The primary communication platforms used were:

* **Telegram** – For daily text-based communication and quick updates.
* **Microsoft Teams** – For virtual meetings, discussions, and formal team check-ins.
* **Trello** – For task management, sprint planning, and tracking contributions.

This section outlines the team’s collaboration strategies, challenges faced in earlier sprints, and how those issues were resolved to enhance overall efficiency.

Sprint 1 & 2: Challenges and Resolutions

1. **Scheduling Daily Standups and Sprint Reviews**

In the initial stages of the project, scheduling daily standups, sprint reviews, and retrospectives was challenging due to conflicting individual timetables. This led to inconsistencies in discussions and delays in aligning progress.

1. **Workload Allocation & Task Tracking**

During Sprint 1, workload distribution was unclear due to inconsistent updates on Trello. Team members hesitated to start tasks, unsure if they were already assigned or completed.

Sprints 3, 4, & 5: Improved Team Cohesion

By implementing structured meeting schedules and improving workload visibility, team efficiency improved in later sprints. The team established a steady workflow, allowing for seamless collaboration and consistent progress. These optimisations ensured that each sprint was executed effectively, enabling the team to work as a cohesive unit.

### iii. Solutions

To address the various technical and organisational challenges encountered throughout the project, several strategies were implemented. These solutions improved data quality, optimised GitHub integration, and enhanced team coordination.

Data Quality Improvements

* **Categorical Null Values:** Imputed with "Unknown" to maintain data integrity while ensuring missing values did not disrupt analysis.
* **Numerical Null Values:** Assigned an extreme placeholder value (-999) to allow for easy filtering in Power BI without skewing calculations.

GitHub Integration Optimisation

* **Unclear Documentation:** External guides and alternative documentation were referenced to supplement the official Snowflake documentation. This allowed for a clearer understanding of the integration process.
* **Testing GitHub Integration:** Multiple test runs were conducted using different configurations to validate the integration before full implementation.
* **Preventing Unintended Repository Unlinking:** Access to execute the integration code was restricted, allowing only the account admin role to rerun the integration setup. This prevented accidental disruptions to repository connections.

Team Coordination Enhancements

* **Meeting Schedule Fixation:** Standups, sprint reviews, and retrospectives were scheduled at predetermined times, allowing team members to plan around these fixed sessions.
* **Task Allocation Improvements:**
  + From Sprint 2 onwards, Trello was updated consistently, with every task assigned to a specific team member.
  + Updates on task completion were communicated via Telegram message logs to ensure visibility and prevent duplication of work.

These solutions collectively enhanced project execution, ensuring smooth collaboration, reliable data management, and an optimised workflow.

## b. Reflection and Lessons Learned

### i. Agile Methodology

For this project, the team followed the Scrum framework within the Agile methodology, ensuring an iterative, flexible, and collaborative approach to managing tasks and deliverables.

#### Sprint Structure and Timeline

The project was divided into five sprints:

* Sprints 1, 2, 4, and 5 lasted approximately two weeks each.
* Sprint 3 was allocated three weeks due to its complexity and workload.

Each sprint followed a structured workflow:

1. Sprint Planning (Beginning of Sprint)
2. Daily Stand-ups (Throughout Sprint)
3. Sprint Review & Retrospective (End of Sprint)

Each sprint had an allocated Scrum Master to facilitate Agile processes and a Product Owner responsible for managing the product backlog.

#### Sprint Components and Discussions

1. **Sprint Planning**

At the start of each sprint, the team conducted a planning session to set clear expectations and define goals:

* **Sprint Goal** – The overall objective of the sprint, agreed upon by the team.
* **Definition of Done** – A standard that defines when a sprint is considered successfully completed.
* **Sprint Backlog** – A dynamic list of tasks required to be completed by the end of the sprint (modifiable by the team).
* **Product Backlog** – Managed by the Product Owner, who moves completed tasks from the Sprint Backlog into the Product Backlog.

1. **Daily Stand-ups**

To maintain transparency and coordination, the team conducted daily stand-ups to discuss:

* **Completed Tasks** – Items from the Sprint Backlog that were finished.
* **Challenges & Help Requests** – Team members flagged tasks where assistance was needed.
* **Upcoming Tasks & Plans** – Members shared their next steps and ensured alignment.

1. **Sprint Review & Retrospective**

At the end of each sprint, a structured session was held to reflect on progress and improve future sprints:

* What was done well – Acknowledging successful aspects of the sprint.
* Areas for Improvement – Identifying blockers, challenges, and strategies to prevent similar issues in future sprints.
* Lessons learned & carry-forward practices – Highlighting key takeaways to reinforce in the next sprint.
* Team dynamics discussion – Addressing communication, collaboration, and resolving any disputes to ensure optimal team productivity.

#### Impact of Agile Methodology

The team consistently followed this iterative cycle, ensuring:

* Clarity and alignment from sprint planning, allowing for fair task distribution.
* Transparency and collaboration through daily stand-ups, keeping all team members updated.
* Continuous improvement via sprint retrospectives, refining processes and preventing recurring issues.
* Stronger team dynamics, ensuring optimal efficiency and synergy throughout the project.

By adhering to Scrum principles, the team maximised productivity, minimised roadblocks, and successfully delivered a structured, collaborative, and agile-driven project.

### ii. Personal Growth

#### Aden

This project has been a huge learning experience for me, not just in terms of technical skills but also in personal growth. Over the past 2–3 months, I’ve learned how to apply Agile methodology in a real-world setting, which has helped me become more adaptive, organised, and mindful of my time. Before this, I didn’t really track my work in a structured way, but following the Scrum framework has made me more aware of how I manage my workload. I know that as life gets busier, these skills will only become more important, and I’m glad I got the chance to develop them now.

Another key takeaway from this assignment was learning that progress isn’t a one-time thing — it’s a continuous process. Because we held Scrum meetings every week, we didn’t just complete a task and move on. Instead, we consistently checked in, reviewed our work, and improved on it where necessary. If mistakes were made, they were caught early, and we had the chance to correct them before they snowballed into bigger issues. This experience really reinforced the importance of consistent effort and ongoing improvement. Looking back, this project didn’t just help me grow in a classroom setting; it also shaped the way I work and think, both as an individual and as part of a team.

#### Fariha

This project has been the longest and most extensive I’ve worked on during my time at Ngee Ann Polytechnic, and it has been an incredibly fulfilling experience. More than just a technical challenge, it pushed me to grow in how I approach teamwork, leadership, and long-term project management.

At the start, stepping into the role of Scrum Master for the first sprint was a challenge. Everything was still new, and I often felt like I wasn’t handling things as effectively as I could have. However, as the project progressed, I adapted, learned from my experiences, and became more confident in managing the workflow. Now, leading the final sprint as Product Owner, I can see how much I’ve grown. This time, I feel more prepared, more capable of making decisions, and better at guiding the team towards our goals.

Beyond leadership, this project reinforced the importance of long-term planning, commitment, and teamwork. Working within a structured Agile framework made me more mindful of pacing, balancing multiple tasks, and ensuring clear communication with my teammates. I’ve come to appreciate how valuable it is to lean on each other’s strengths and to navigate challenges as a team.

Looking back, these past three months have been a major milestone in my journey as a Data Science student. I’ve not only sharpened my technical abilities but also gained confidence in my leadership and collaboration skills — things that will undoubtedly stay with me far beyond this project.

#### Jovan

This project felt like a great endeavour, with many challenges yet enjoyable moments. It has been a great pleasure to work with such compatible teammates and has been extremely fulfilling experience. It pushed me to not only step up in terms of leadership, but also changed my perspective on how a team should work together.

When we started the project, I found myself stuck in a very different environment. I barely knew anything about databases or the Agile framework, and the things I knew were too basic to be applicable. However, as time passed, I learned more from my teammates, watching them lead the team, receiving their guidance and feedback on issues I had and continued to push forward.

When it came to Sprint 3 where I was the Product Owner, I was nervous. I distinctively remember my hands trembling as before we started sprint planning. I felt the weight on my shoulders and knew I needed to step forward and lead my team towards my sprint goal. However, I knew that my teammates also had this mutual feeling since they had to also step up when it came to their sprints. Hence, I showed a brave front and lead the team. As the days passed by in the sprint, the more work we accomplished, the more confident I was as leader that we were going to finish what we needed to and reach the sprint goal. And at the end of that sprint, looking back, we finished the CI/CD and Role based access processes, a huge step forward led by me, showing me how much I have grown and that I could lead.

This project has opened my eyes to a framework that I never taught would work, where everyone takes their own initiative to lead the team. It also boosted my confidence in leading a team, something I tend to push away from doing or avoid all together. All in all, I am appreciative of the team I worked with and would like to thank them for not only being supportive teammates, but also ones who taught me that I could lead.

#### Keaven

For me, the biggest takeaway from this project wasn’t just the technical skills of using Snowflake or Power BI, but rather the implementation of the Agile framework within a real-world project.

Before this project, I often felt like I was just going through the motions in group work. I wanted to be a significant contributor to the team, but I always felt like I fell short. However, this project gave me the opportunity not just to contribute ideas but also to step up and lead when needed.

During Sprint 3, I took on the role of Scrum Master, where I moderated daily stand-ups, facilitated discussions, and ensured that meetings were scheduled at suitable times. A major challenge was coordinating meetings while most of the team was overseas on holiday, but I managed to ensure full participation despite the logistical difficulties.

In Sprint 4, I was the Product Owner, and midway through the sprint, I noticed that the team was spending more time debugging and troubleshooting than focusing on our actual sprint goal — connecting to Power BI. Recognising this issue, I held a discussion with the team, proposed a pivot in our sprint objectives, and adjusted the definition of done to better align with our immediate needs. This experience taught me the importance of being adaptable, making quick decisions, and understanding the needs of the team while balancing my own responsibilities as a contributor.

This project has had a meaningful impact on me — not just as a team member, but also as a leader. It reinforced my ability to step up, communicate effectively, and support my team, whether in a leadership role or as an active participant.

#### Nabihah

This project taught me how to collaborate more effectively, especially by trusting my teammates to take ownership of their respective sections. I gained a deeper understanding of how SQL can be applied to data science scenarios, moving beyond simple database management to more advanced tasks like cleaning and transforming data for analysis. Additionally, I enhanced my ability to adapt to new tools like Snowflake, which streamlined the data migration and transformation process, and reinforced my problem-solving skills in tackling data quality challenges.

### iii. Team Growth

#### Aden

Reflecting on our team’s growth, I am proud to say that we adapted quickly, overcame challenges, and ultimately became a strong, well-coordinated team. At the start of the assignment, finding a common time for our scrum meetings was difficult due to our conflicting schedules and initial unfamiliarity with the Agile framework. This challenge became even more prominent during the holidays, as many of us were overseas and occupied with personal commitments.

However, after a month of working together, we established a better rhythm, aligning our schedules and conducting daily scrums more consistently. This allowed us to follow the Agile framework more effectively and improve our overall workflow. Our ability to adapt to each other’s availability and working styles played a crucial role in resolving these early difficulties, ultimately strengthening our teamwork and efficiency.

#### Fariha

At the start of the project, our team faced the challenge of aligning our individual strengths with the project’s requirements. As we progressed, we developed a deeper understanding of each other’s capabilities, which allowed us to delegate tasks more effectively. This shift in role management ensured that everyone could focus on areas where they could contribute the most value, improving our overall efficiency.

As the timeline extended and the project’s complexity increased, we also learned to adapt to changing priorities and unanticipated challenges. Whether it was troubleshooting technical issues or refining our approach to data cleaning and integration, we grew more resilient as a team. This growth has not only strengthened our collective problem-solving skills but has also built a foundation of trust and camaraderie that will benefit us in future collaborations.

#### Jovan

From the beginning, our team faced many challenges. One of the biggest challenges was that none of us took cloud databases, only sharing the common module of databases, which only teaches the basics of SQL. Hence, we found ourselves not understanding how to code certain SQL process. Another challenge was that we had were the many schedule conflicts and unfamiliarity with the agile framework. In the first sprint, we often found it difficult to set a meeting time, something we brought up in the sprint retrospective.

However, over time, we started to not only get familiar with the agile network but also the cohesion of our team, quickly adapting to each other strengths and weaknesses. We then started to assign our work based on our strengths and communicating our issues with one another to ensure problems will be fixed as soon as possible. We coordinated our task together, making sure that our work can come together as a whole.

By the end of this project, our team has significantly improved how we addressed issues, collaborate on work and communicate effectively. No one was afraid to speak out to address issues, afraid to help when needed nor afraid to suggest new ideas. We truly came together as a team, and I am proud to say that this was the best team I had on any project I took in Ngee Ann Polytechnic.

#### Nabihah

Throughout the project, our team significantly improved in both collaboration and communication. Over time, we became more comfortable addressing challenges and seeking assistance when needed, creating a supportive environment where no one hesitated to ask for help. As we grew more familiar with each other's working styles, we adapted to complement one another’s strengths and stepped in to offer support even without being asked. This mutual understanding and proactive mindset allowed us to work more efficiently and cohesively, ultimately enhancing our overall productivity and teamwork.

#### Keaven

Within just three months, the team experienced significant growth, both in coordination and teamwork. From the start, we shared a common goal and were aligned in our approach, but early challenges — such as scheduling conflicts and workload distribution — tested our ability to collaborate effectively. However, we quickly adapted, implementing solutions that improved our workflow and communication.

By the later sprints, our coordination had strengthened, and we developed a strong sense of trust in each other’s abilities. This allowed us to work more efficiently and support one another when challenges arose. Collaboration is an essential skill in any professional setting, and this project has reinforced its importance. Working with this team has not only strengthened our bonds but also built a foundation of mutual trust and respect. I am truly grateful to have had the opportunity to work alongside such independent yet strong leaders.

### iv. Future Applications

The learnings from this project have equipped our team with valuable insights and skills that can be applied to future DataOps projects and real-world scenarios. These lessons will undoubtedly enhance our approach to managing data, collaborating effectively, and leveraging technology in diverse industries. Key takeaways include:

* **Scalable Cloud-Based Architectures**

Designing and implementing scalable, cloud-based solutions with Snowflake has provided us with a blueprint for handling large datasets efficiently. This can be applied to industries such as healthcare, retail, and logistics, where data scalability and accessibility are critical.

* **Proficiency in Modern Data Tools**

Our experience with Snowflake, Power BI, and SQL has strengthened our technical capabilities. These skills are transferable to future projects requiring data integration, migration, and visualisation. For example, Snowflake’s flexibility and Power BI’s visualisation features can streamline workflows across various data-intensive sectors.

* **Agile Collaboration and Communication**

Adopting agile methodologies and refining our communication strategies have improved our team’s efficiency and role management. These practices can be directly applied to future collaborative projects, ensuring clearer goals, faster execution, and better problem-solving.

* **Attention to Detail in Data Migration**

Working on large-scale data migration tasks has highlighted the importance of precision and thoroughness. These skills will be crucial for ensuring data integrity and consistency in any future projects involving complex datasets.

* **Future-Oriented Insights**

The direct integration of Snowflake with analytics platforms like Power BI demonstrated how modern tools can simplify decision-making processes. This insight encourages the use of integrated platforms for creating actionable insights in real-time.

As a team, we now have a strong foundation in DataOps principles and an adaptable skill set to address data challenges in any environment. These lessons will enable us to drive continuous improvement and innovation in future data-driven projects.

# 9. References

## a. Content

* **Section 1c:** Rahaman, M. M. (2024). *Opportunities and Challenges of Data Migration in Cloud*. *American Scientific Research Journal for Engineering, Technology, and Sciences*, 48(1), 73–89. Retrieved from <https://asrjetsjournal.org/index.php/American_Scientific_Journal/article/download/4450/1573>
* **Section 2b:** Catapult. (2024, February 13). *How Data Analysis Transforms F1 Race Performance.* Retrieved from <https://www.catapult.com/blog/f1-data-analysis-transforming-performance>
* **Section 2ciii:** Shane Jordan. (2024, July 26). *The Pros and Cons of Cloud vs. In-House Servers*. SysGen. Retrieved from <https://sysgen.ca/cloud-vs-in-house-servers/>
* **Section 3a:** Allam, K., Ankam, M., & Nalmala, M. (2023). Cloud Data Warehousing: How Snowflake Is Transforming Big Data Management. *International Journal of Computer Engineering and Technology*, 14(3), 156–162. Retrieved from <https://iaeme.com/Home/article_id/IJCET_14_03_016>

## b. Images

* **Figure 3.1:** OneClick IT Consultancy. (2023). [Image of top 8 benefits of Snowflake]. In *Top 8 Benefits of Snowflake Data Cloud Platform.* Retrieved from <https://www.oneclickitsolution.com/blog/benefits-of-snowflake>
* **Figure 4.1:** Snowflake. (n.d.). In *Overview of Access Control.* [Image of overview of role-based access]. Retrieved from <https://docs.snowflake.com/en/user-guide/security-access-control-overview>
* **Figure 5.1:** Snowflake. (n.d.). In *Using a Git repository in Snowflake*. [Image of how Snowflake works with a Git repository]. Retrieved from <https://docs.snowflake.com/en/developer-guide/git/git-overview>
* **Figure 8.1:** Snowflake. (n.d.). [Image of redundant code creating a secret for GitHub integration]. In *Setting up Snowflake to use Git.* Retrieved from <https://docs.snowflake.com/en/developer-guide/git/git-setting-up>