NOVA IMS Information Management School

MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Business Intelligence

Ride4All

Mara, Mesquita, number: 20241039
Martim, António, number: 20240601
Victor, Pita, number: 20240596

Group 18
NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

March, 2025

INDEX

1.Introduction *1*

2. Business Needs *2*

3. Identification of Business Questions *2*

4. Data Sources *4*

4.1 Dataset Structure 4

Table 2: Datasets 5

4.2 Discovery Process 5

5. Data Warehouse: Data Modelling Methodology *6*

5.1 Overview of the Methodology 6

5.2 Identifying the Business Process 6

5.3 Defining the Granularity 6

Table 3: Fact Table 7

5.4 Dimensional Design 7

5.5. Star Schema Model 8

6. 2nd Delivery Refinements *9*

7. ETL Process *9*

7.1. SQL Scripts 10

7.2. Loading the Data into the Data Warehouse 10

8. Notebooks *17*

8.1. Notebook 1 17

8.2. Notebook 2 19

9. Semantic model *20*

9.1. Integration of Previous Steps 20

9.2. Description of the Semantic Model 20

10. Powe bi report *26*

10.1. General Story Telling Approach 26

10.2. Visual Approach and Structure 26

10.3. Client Analysis 27

10.3. Driver Analysis 28

10.4. Car Analysis 29

10.5. Operational Analysis 30

10.6. Efficency Analysis 31

11. Conclusion *32*

1.INTRODUCTION

In today's fast-paced urban environments, ride-sharing services have become an essential component of modern transportation. Ride4All is a company committed to providing convenient, efficient, and cost-effective mobility solutions by bridging the gap between passengers and drivers through an intuitive mobile application. Operating across multiple cities, Ride4All enables users to request rides on demand, ensuring timely and accessible transportation while offering drivers a flexible earning opportunity.

Ride4All operates in multiple cities, facilitating thousands of rides daily. However, the company currently relies on fragmented reporting and manual data aggregation, limiting real-time visibility into business performance. As the company continues to expand, decision-making challenges arise due to the increasing volume of data generated daily from ride transactions, driver activity, and customer behavior. This lack of structured analysis impacts the company's ability to track ride performance trends, optimize ride allocation, assess driver efficiency, and personalize customer engagement.

To address these challenges, Ride4All seeks to implement a Business Intelligence (BI) system that enables data-driven decision-making. A BI-driven framework will allow for real-time performance tracking, predictive insights, and operational optimizations, ultimately improving business efficiency, customer experience, and scalability.

2. BUSINESS NEEDS

With the growing demand for ride-sharing services, Ride4All faces increasing complexity in managing its operations efficiently. The company generates substantial data daily, covering ride details, customer demographics, driver activity, and location-based demand patterns. However, without an effective analytical framework, extracting actionable insights from this data remains a challenge. Several key issues have been identified:

- **Trips per Month, Quarter, and Yearly (Accumulated to Date):** The company lacks a structured system to analyze rides over different time frames, making it difficult to identify demand fluctuations and seasonal patterns.
- **Number of Rides and Average Distance of Rides:** Understanding ride frequency and distance is crucial for assessing operational efficiency and setting competitive pricing strategies.
- **Comparison of Current Year vs. Previous Year:** Ride4All needs to evaluate growth trends and market dynamics by comparing performance across different periods.
- **Building Client Profiles:** There is no detailed profiling of customers, leading to missed opportunities for personalized engagement and service improvements.
- **Ride Details (at the Trip Level):** Tracking individual ride transactions enables more granular insights into operational performance and user behavior.
- **Aggregate Rides by Driver, Pickup Point, and Destination Point:** Understanding driver performance and location-based demand is essential for balancing supply and optimizing route planning.

Addressing these needs requires a structured approach to data analysis, which is where Business Intelligence plays a crucial role. By integrating a BI-driven system, Ride4All can transform raw data into meaningful insights.

3. IDENTIFICATION OF BUSINESS QUESTIONS

After analyzing Ride4All's operational challenges and data requirements, we have identified five key categories to guide the BI implementation: Ride Performance & Trends, Customer Insights, Driver & Vehicle Performance, Operational Optimization, and Location Analysis. These categories encapsulate the essential aspects of Ride4All's business model and will serve as the foundation for Power BI reports. The following business questions address critical areas for improvement and will support data-driven decision-making:

Category	Business Question	Why It Matters
Ride Performance & Trends	What is the total number of rides each month, quarter, and year?	Identifies seasonal patterns and demand fluctuations.
	What is the percentage growth or decline in total rides year over year?	Provides insight into business growth and market trends.
	What is the average ride duration by day, week, and month?	Helps optimize ride pricing and driver allocation.
	What is the total revenue each month, and what is its growth trend?	Assesses financial health and revenue trends.
	How does ride demand fluctuate by hour and day of the week?	Enables better resource allocation and surge pricing strategies.
Customer Insights	What percentage of customers are repeat users, and how often do they ride?	Measures customer loyalty and retention.
	What are the peak demand hours in each district?	Helps improve ride availability and service planning.
	How does ride frequency and spending vary by age and gender?	Supports targeted marketing and personalized offers.
	What are the most frequently traveled routes for different customer segments?	Helps optimize routes and pricing.
	What is the average fare for each customer profile?	Assesses affordability and revenue potential.
Driver & Vehicle Performance	Who are the top drivers based on total rides and earnings?	Recognizes high-performing drivers and incentivizes efficiency.
	What is the average number of rides per driver each day and month?	Ensures fair workload distribution and driver productivity.
	What is the average ride distance for each driver?	Helps optimize driver efficiency and route planning.

	What are the top car brands/models based on total rides?	Aids in vehicle selection and fleet optimization.
	What is the average fuel consumption per ride for each vehicle?	Helps reduce operational costs and enhance sustainability.
Operational Optimization	What are the busiest pick-up and drop-off locations?	Improves ride distribution and reduces wait times.
	What are the most profitable routes?	Enhances revenue generation and cost efficiency.
	What is the average idle time for drivers?	Helps minimize downtime and improve driver utilization.
	How can ride allocation be improved to reduce customer wait times?	Enhances customer satisfaction and service reliability.

Tabela 1 – Businnes questions

4. DATA SOURCES

The data for this project was provided by **Ride4All**, a ride-sharing service, spanning the period from **2020 to 2023**. The dataset is refreshed daily, ensuring it reflects the most up-to-date ride and driver information. Provided in CSV format, the data offers a wide range of insights, from ride statistics to customer demographics.

4.1 Dataset Structure

The dataset consists of the following files:

Dataset	Description	Key Columns
Rides.csv	Contains transaction-level ride details such as timestamps, duration, driver, pickup/drop-off locations, fuel consumption, distance traveled, and fare amounts.	Ride ID, StartDatetime, EndDatetime, Duration, Amount, Pickup Location, Dropoff Location
Cars and Drivers A.csv	Contains data for **female drivers**, including car plates, brands, manufacturers, driver names, and birthdates.	Car ID, Car Plate, Brand, Manufacturer, Driver Name, Birthdate, Gender

Cars and Drivers B.csv	Contains data for **male drivers**, with similar details as in the A file.	Car ID, Car Plate, Brand, Manufacturer, Driver Name, Birthdate, Gender
Customers.csv	Contains customer demographic data, including IDs, names, genders, and birthdates.	Customer ID, Name, Gender, Date of Birth
Locations.csv	Provides geographical information related to ride locations, with location IDs that can be mapped to ride data.	Location ID, Location Name, Country

Table 2: Datasets

4.2 Discovery Process

In exploring the datasets, the following key insights were discovered:

Rides.csv:

The Amount column consistently shows a value of **3.6** across multiple ride transactions, suggesting it could represent a **minimum fare** or **cancellation fee**. This common value warrants confirmation from **Ride4All** to clarify its significance. The **maximum ride duration** recorded is **6060 seconds** (approximately 1 hour and 41 minutes), which is a valuable data point for understanding the extreme ride durations and identifying potential outliers in ride time.

Cars and Drivers A.csv & B.csv

A gender disparity in ride frequency was observed, with **male drivers** in Cars and Drivers B.csv having more rides than **female drivers** in Cars and Drivers A.csv. This indicates that male drivers tend to be more active or available for rides than female drivers, reflecting potential operational or scheduling patterns.

Customers.csv:

The age range of customers is predominantly between **22 and 45 years**, indicating that **younger to middle-aged adults** are the primary users of the service.**Male customers** outnumber **female customers** in the dataset, reflecting a similar trend to the driver gender disparity**.**Gender label inconsistencies were found in the dataset, such as non-standard entries like **H** and **S**, which complicate the analysis of gender distribution and require data cleaning.

Locations.csv:

The dataset provides information on **pickup and drop-off locations**, with certain locations appearing more frequently than others, indicating high-demand areas for the ride-sharing service.**Urban centers** and **tourist areas** were identified as common pickup/drop-off points, highlighting the need for optimized driver allocations in these high-traffic regions.The **country** column is uniform (only **Portugal**), so no additional geographic variation was observed beyond location-level data.

5. DATA WAREHOUSE: DATA MODELLING METHODOLOGY

5.1 Overview of the Methodology

For this project, we adopted the **Kimball methodology** to design the Dimensional Model. Kimball's approach is widely recognized for its simplicity, scalability, and ability to handle large datasets efficiently. This methodology allows for the creation of a centralized data warehouse, which integrates and stores data from multiple operational systems into a structured model.

Kimball's method relies on the **star schema design**, consisting of a **fact table** surrounded by **dimension tables**. This design is particularly useful for reporting and analytical purposes, as it simplifies querying and ensures efficient data access.

5.2 Identifying the Business Process

The core business process represented in the data warehouse is ride management, as the focus of Ride4All revolves around the tracking of ride transactions. The fact table is designed to capture data at the level of individual ride transactions, which are central to the company's operations.

5.3 Defining the Granularity

For this project, we defined the granularity at the **individual ride transaction level**. Each record in the fact table represents a unique ride, regardless of whether the same customer takes multiple rides on the same day. This provides flexibility in reporting and enables both high-level aggregates and detailed analysis.

The following key attributes will be captured for each ride transaction in the **fact table**:

Attribute	***Description***
Client	Customer ID
Driver	Driver ID

Attribute	*Description*
Car	Car ID
Pickup Location	Location ID for pickup
Dropoff Location	Location ID for drop-off
Start Date and Time	Timestamp for when the ride started
End Date and Time	Timestamp for when the ride ended
Currency	Currency used for the fare
Fare Amount	Total fare for the ride
Distance Traveled	Total distance traveled in kilometers
Fuel Consumption	Amount of fuel used during the ride
Ride Duration	Total duration of the ride in seconds

Table 3: Fact Table

5.4 Dimensional Design

To support effective reporting and analysis, we created several **dimension tables** surrounding the **fact table**. These dimensions provide descriptive context for each of the key entities involved in a ride transaction.

The following **dimension tables** were defined:

Dimension	*Key Attributes*
Customer Dimension	Customer ID, Name, Gender, Age, Registration Date
Driver Dimension	Driver ID, Name, Gender, Age, Driver's Car Information
Car Dimension	Car ID, Car Model, Manufacturer, Year of Manufacture
Location Dimension	Location ID, Location Name, Country
Time Dimension	Day, Week, Month, Year, Time of Day

Table 4: Dimension

Each **dimension table** contains descriptive attributes, which allow users to group and filter the data by specific categories. For instance, the **Customer Dimension** provides detailed demographic information about customers, enabling segmentation based on age, gender, or other factors. The **Time Dimension** allows for

time-based aggregation, making it easier to analyze trends over daily, weekly, or monthly periods.

5.5. Star Schema Model

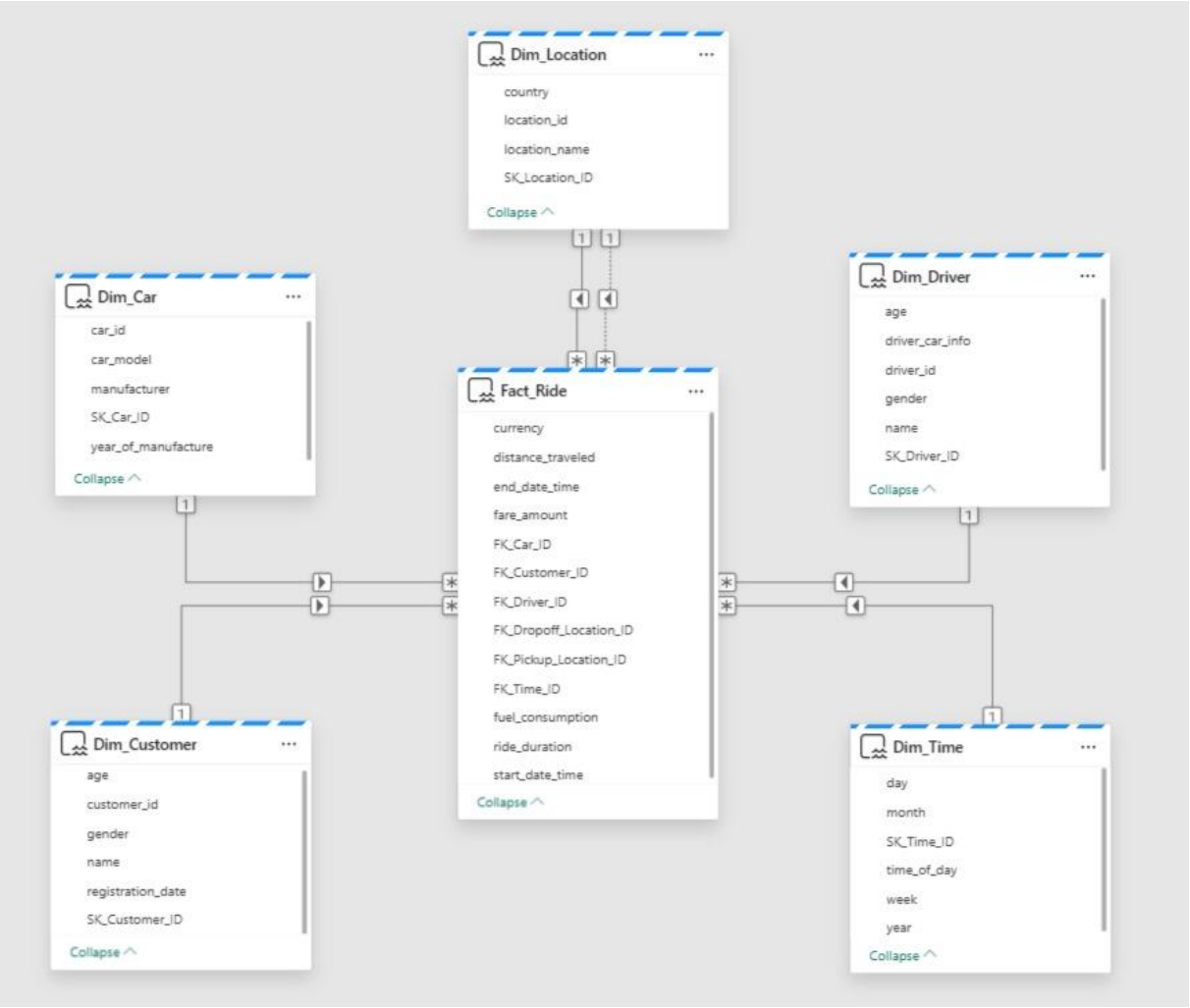


Figure 1: Star Schema

6. 2ND DELIVERY REFINEMENTS

For our 2nd delivery some changes were made in contrast with the previous delivery, most of which were made to the star schema we originally had:

A new dimension table was added named Dim Currency. Initially, there was indeed a currency attribute in our fact table, but we later realized that there was no currency table. Therefore, we decided that it was important to create a specific dimension table for currencies.

We split our old Dim_Time table in two new dimension tables, Dim_Time and Dim_Date, basically separating time information (hours, minutes, seconds) from date information (day, month, year) in our dataset. A few more attributes were created as well for each of those new tables individually.

Within each dimension and fact tables there were small adjustments made that will be mentioned down the line.

7. ETL PROCESS

This next stage of our project revolves around the selection of appropriate data sources for extraction, transforming data and loading it into the Data Warehouse (ETL).

For this step of we used various CSV files sourced by the company and established a Source Lakehouse named ***Project_LH*** that we used to upload the data:

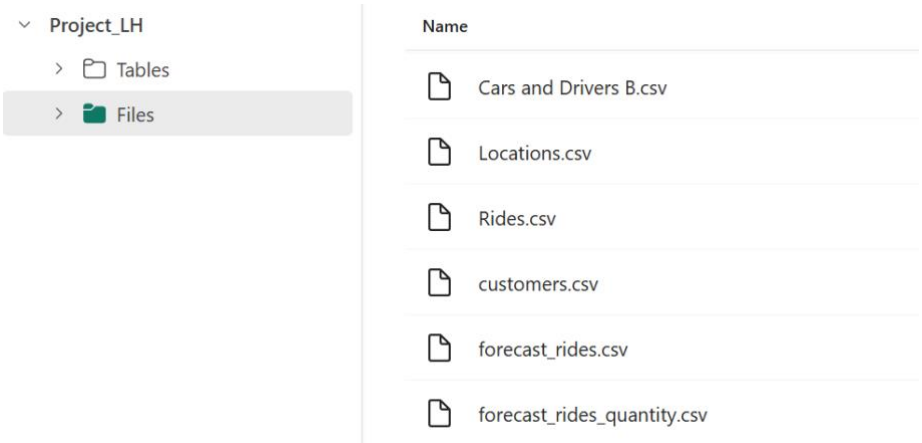


Figure 2: Lakehouse Project_LH

7.1. SQL Scripts

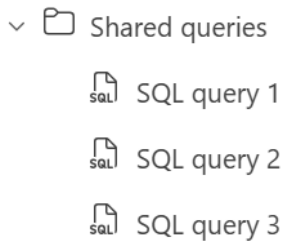


Figure 3: SQL queries in the Data Warehouse

To create and load all our dimensions and fact tables we used ***SQL query 1***. However, when we were populating the *Dim_Location* table, we spotted an error in a few attributes, specifically in *District* and *Country* which contained a maximum string length of 25 characters. After discussing among the team, we decided to alter that number to 50, as it could potentially lead to constraints further ahead in the project, using ***SQL query 2*** to apply the change. We decided to create this additional query because we had already populated the remaining dimension tables, with no mistakes detected, and didn't want to lose that progress, therefore the existence of that query 2. In ***SQL query 3*** the old *Dim_Location* table was deleted to avoid confusion and redundancy.

7.2. Loading the Data into the Data Warehouse

1. **Dataflow of the Dim_Client table**:

We began by importing the customers.csv file to create the dataflow corresponding to the client's table. During the data exploration process that we did earlier, we noticed that the gender column contained four distinct values: 'F', 'S', 'M', and 'H'. To simplify the analysis and improve clarity, we decided to group 'F' and 'S' as Female, and 'M' and 'H' as Male. This allowed us to reduce the number of categories and made the data more consistent and easier to interpret.

To address the possibility of repeated clients, we thought it would be more effective to apply a grouping operation based on B*K_Client, Client_Name, Client_Gender,* and *Client_Birthdate*. This approach enabled us to keep all the associated records for each client while representing them only once in the main table. We chose this method over simply deleting duplicates, as it ensures no information is lost. By storing related data in a nested GroupedData column, we preserved the full detail of the dataset and left open the possibility for future aggregations or deeper analysis.

2. **Dataflow of the Dim_Drivers table**:

For this process, we worked with two different data sources: Cars and Drivers A.csv and Cars and Drivers B.csv, both containing information about drivers. Our objective was to consolidate all available data into a single, unified table.

Once the merge was completed, we removed columns that were not relevant in the context of driver information, such as brand, manufacturer, and year, since they referred to car data and did not make sense in this context.

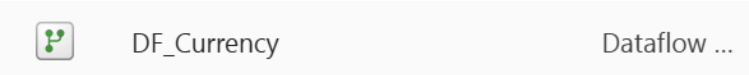
Additionally, we standardized the values in the gender column — where different letters were originally used to represent gender — by replacing them with the full terms “Female” and “Male”.

3. **Dataflow of the Dim_Car table**:

The process was fairly straightforward — we began by removing all columns that were unrelated to the car itself, such as those referring to drivers, in order to keep the table focused and relevant. We also added an index column to serve as a unique identifier. Since later stages of the project require us to map both a Business Key (BK) and a Surrogate Key (SK) for each car, we decided to include the index to ensure we

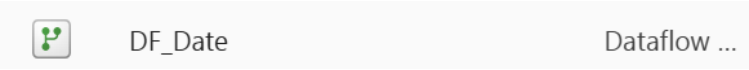
had a consistent and sequential key available to assign either role as needed during data warehouse integration.

4. **Dataflow of the Dim_Currency table**:



For the currency dataset, we started by keeping only the relevant column containing the currency codes. We then added an index column to serve as a unique identifier, following the same reasoning applied in other tables — to support the creation of a business key or surrogate key if needed later. Additionally, we created a new column called Currency Nam**e**, where we mapped the codes to their full names (e.g., “EUR” to “Euro”). Although all entries in our data were in euros, we included logic for other currencies like USD and GBP, and a fallback value (“Other”) to ensure that if future payments are made in different currencies, the transformation won’t result in errors or missing values.

5. **Dataflow of the Dim_Date table**:



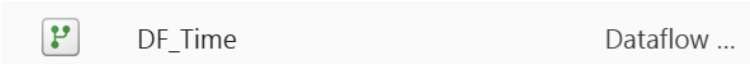
The Dim_Date that we created and mentioned in the changes we made for the 2nd delivery was, unlike the previous dimension tables in our project, built entirely from scratch. We chose to generate it using Power Query's List.Dates() function, which allowed us to define a complete and continuous range of dates from January 1st, 2020 to December 31st, 2023. This approach ensured that no dates would be missing.

After generating the date list, we added standard calendar attributes such as year, month, day, quarter, and week.

One column we gave particular attention to was Is_Special, which flags dates like New Year’s Day (01/01) and Christmas (25/12).

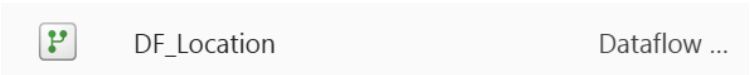
Finally, we created the SK_Date column to serve as a surrogate key in the YYYYMMDD format.

6. **Dataflow of the Dim_Time table**:



For the creation of Dim_Time, we chose to generate the time dimension from scratch using Power Query's *List.Times()* function. This allowed us to create a complete list of all minutes in a 24-hour day (1,440 rows), ensuring no time values were missing, a key point that led us to prefer this method over extracting times from transactional data, which might be incomplete. After converting the list into a table, we extracted the hour and minute components, and created a Time Key in HHMM format for readability and matching with fact tables. In parallel, we added a numeric SK_Time column to serve as a surrogate key.

7. **Dataflow of the Dim_Location table**:



The Dim_Location table was fairly straightforward to prepare. We started with the provided CSV file and made a few basic transformations to clean and organize the data. These included renaming some columns for clarity, adding an index to serve as a Surrogate Key (SK_Location), and reordering the columns to improve structure. Since the dataset was already well-structured and complete, no complex operations were needed beyond these foundational adjustments.

8. **Dataflow of the Fact table**:

As the final step in the table population process, we focused on building and preparing the fact table. We began by splitting the DateTime columns (*StartDatetime* and *EndDatetime*) from the original Rides.csv file into separate Date and Time components. This decision was necessary because we are using independent

dimension tables (*dim_date* and *dim_time*) to store and manage each of these attributes.

We also added new calculated columns such as *Duration_Per_Km* and *Gas_Amount_Ratio*. These were derived from existing fields and reflect metrics we considered relevant to support the business questions previously identified in the "Identification of Business Questions" section. Our goal was to enrich the fact table with indicators that could later be used in analytics and decision-making.

During the data exploration phase, we noticed that the *Kms* column contained extremely high and clearly unrealistic values. Based on this observation, we decided to apply a cap of 10,000 kilometers and remove any rows that exceeded this threshold. We considered this a necessary step to reduce the impact of outliers.

Once the data was cleaned and all relevant metrics were created, we proceeded to the merging stage. In this step, we replaced the business keys (BKs) from the fact table with the corresponding surrogate keys (SKs) from the dimension tables. For each dimension-related attribute, we performed a left outer join using the BK as the matching field and extracted the corresponding SK to be stored as a foreign key (FK) in the fact table.

Although we were able to extract most of the required SKs successfully, we encountered a critical issue when merging with the *dim_currency*, *dim_date*, and *dim_time* tables. Specifically, we were unable to retrieve the surrogate keys from these dimensions due to a data format mismatch or join condition problem. This error blocked us from fully completing the mapping of the fact table.

As of the date of this delivery, we have not yet resolved this issue. Before moving to the next phase of the project, it will be essential to address this problem and ensure the fact table is fully integrated with all dimension tables.

9. **PL_Ride4all_Load Pipeline**:

Regarding our workflow structure, we created a single pipeline, ***PL_Ride4all_Load***, which served as the heart of our solution, executing the ETL process for each dimension and fact table, and subsequently loading the data into the Data Warehouse.

SQL script activities to clear the current data in the Dimensions and Fact Tables:

Figure 3: Script Activity

Script activities grant a clean slate with every loading. This approach ensures that future loadings with additional entries don't introduce duplicate records, thereby maintaining data integrity.

Wait activities:

Figure 4: Wait Activity

Wait activities serve as placeholders. Their inclusion is crucial to ensure the coordination/synchronizing of steps within the workflows. We placed two Wait activities in our pipeline; the first, ***Wait1*** one is located right after the script activities, that remove all the rows from the original fact and dimension tables while preserving the table structure. This one allows the system to complete the clearing process before moving on to load the data into the dimension tables. As for the second, ***Wait2,*** is placed after the ETL processes, and its purpose is to ensure that all the earlier dimension tables are fully loaded before the pipeline moves on and acts as a synchronization step.

All together:

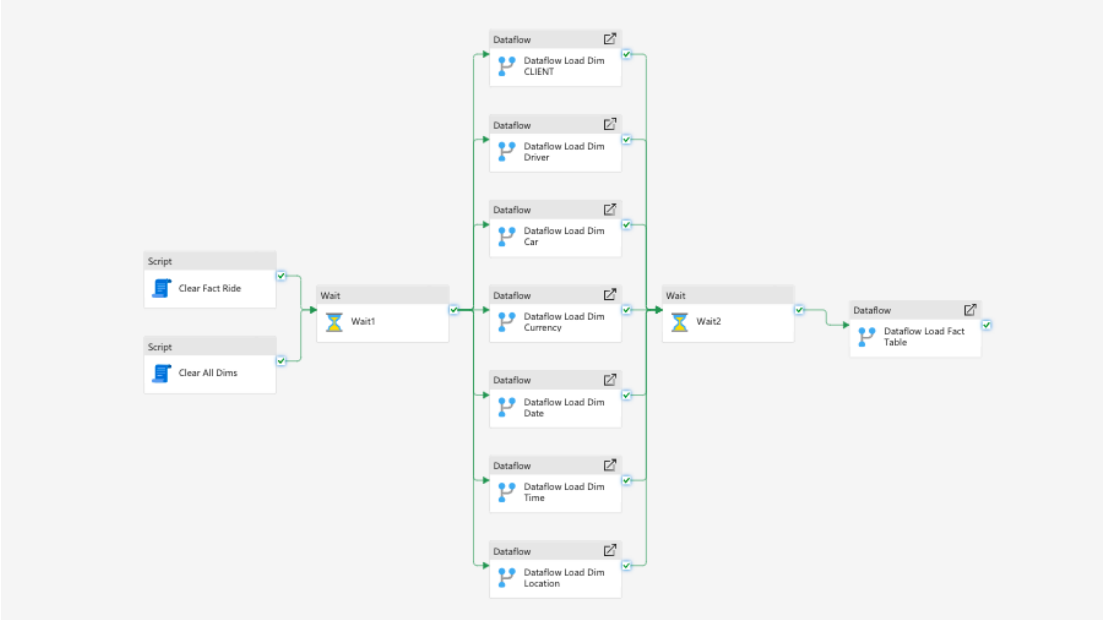


Figure 5: *PL_Ride4all_Load*

We successfully established connections and data loading flows for all the dimension tables. However, due to the unresolved error encountered during the extraction of surrogate keys in the Fact table we could not establish that connection.

8. NOTEBOOKS

For our analysis, we made use of two separate notebooks made in Microsoft Fabric, each serving different purposes. In both cases, we believe that they contribute valuable insights, helping in a business decision-making process and better understand demand patterns for Ride4All.

8.1. Notebook 1

The first one, *Notebook 1*, we utilized to build our number of rides forecast using time series analysis. By using ARIMA and SARIMAX models we were able to predict future ride counts based on historical data. The notebook performs the following:

1. **Preprocessing:**

   To better understand the ride data for analysis, we converted *StartDatetime* column from string to datetime format.

   - A new column named *date* was created by extracting the date part from *StartDatetime* and used to group the data by date, to be used for the calculation of the number of rides per day.

   - The dataset is reindexed to ensure all dates in the range are present (even days with zero rides). Missing ride data for those days is filled with “0”.

2. **Plot Evolution of Daily Rides:**

   - In this step we created a graph to observe how the number of rides per day has changed over the entire period, highlighting trends, spikes, and dips in rides activity.

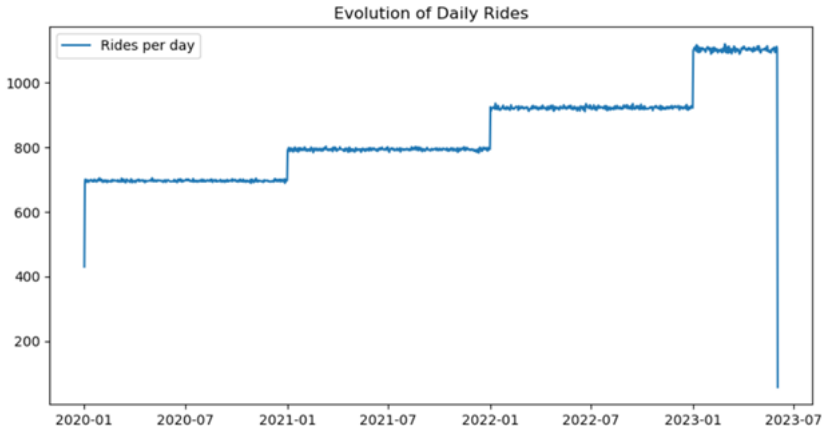


Figure 5: Evolution of Daily Rides Plot

3. Model Training and Evaluation:

- Primarily focusing on training the ARIMA and SARIMAX, which are then evaluated by using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) and logged using MLflow to track the performance metrics for future comparisons and to helps us choose the appropriate model.

4. Forecasting and Results:

- Finally, the best-performing model was loaded (we decided to go with SARIMAX) and used it to forecast the number of rides for the next 60 days based on the last date in the dataset.

- We generated a graph to compare the actual rides for the last 120 days with the forecasted rides for the next 60 days:

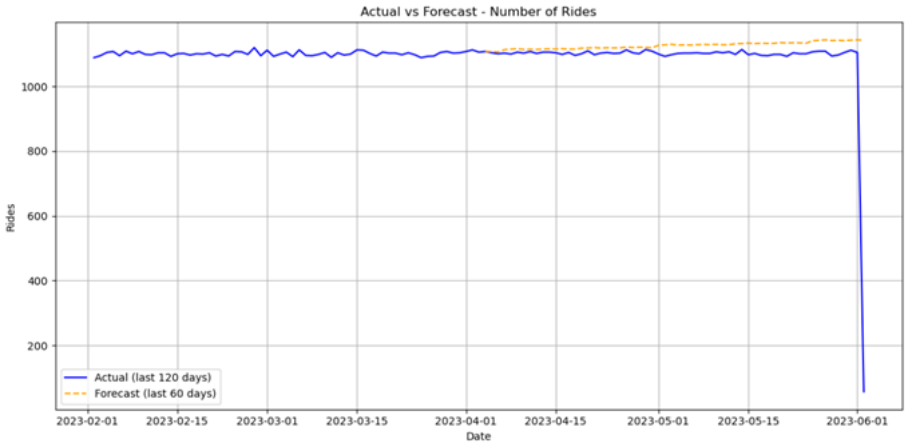


Figure 6: Actual vs Forecast – Number of rides Plot

8.2. Notebook 2

To operationalize the forecasting logic, we created Notebook 2, which loads the *SARIMAX* model previously trained and registered in *Notebook 1* and then uses it to forecast the number of rides for the next seven days. This notebook reads the trained model from the MLflow registry, performs the forecast, and writes the results into a csv file stored in the Lakehouse. Additionally, the training and forecasting process was tracked and saved under an ML Experiment named ride_forecasting, and the model was registered as *Sarimax_ML* within the Fabric environment.

9. SEMANTIC MODEL

9.1. Integration of Previous Steps

Before diving into the construction of the semantic model, it's important to highlight a key challenge faced by our group. At the previous delivery, we encountered issues publishing the fact table. To avoid delays, we decided to begin our semantic model and initial visualizations using only the available dimension tables. This approach allowed us to move forward, for time efficiency reasons, while continuing to investigate and resolve the problem in parallel. Once the fact table was successfully published, we integrated it into the model, enhancing the analytical depth and enabling us to create more complete and insightful dashboards in Power BI.

Additionally, in the previous delivery, we had already developed a forecasting model and exported the predictions to a .csv file stored in the Lakehouse, named *forecast_rides.csv*. However, at that stage, we had not yet created a proper table or dataflow to load it into the semantic model. This was later addressed, and the forecast data was integrated into the model as a table named *Forecast Rides*, making it available for future use in visualization.

9.2. Description of the Semantic Model

We did not build our model entirely from scratch. Instead, we leveraged the existing dataflows available in our Fabric warehouse, which included both the dimension and fact tables. From there, we established the necessary relationships, always ensuring a many-to-one connection, where the “many” side is the fact table and the “one” side is the corresponding dimension**.**

To simplify and organize the model, we renamed the tables using a consistent prefix (D for dimensions and F for the fact table) and ensured column names were intuitive and clean. At the same time, we hid technical columns , such as surrogate, to avoid clutter and make the report-building process more user-friendly.

- **D Driver**

In the *D Driver* dimension, as with the other tables in the model, technical fields were hidden to simplify the development environment and facilitate the report-building process in Power BI. Throughout the analysis, and according to the needs identified in the visualizations, several DAX measures were created to extract more relevant and actionable insights. These measures enabled us to calculate metrics such as the

average driver age, the percentage of young drivers, and the number of drivers under the age of 45, among other useful indicators for profiling the driver population.

Moreover, to ensure a clean and professional presentation, the data format of all relevant measures was reviewed and adjusted,both for existing fields and newly created DAX measures. The table below summarizes the main measures developed for the D Driver dimension, along with their appropriate data formats:

Measure Name	Data Format
% Young Drivers	Percentage
Avg Driver Age	Decimal (1 decimal place)
Drivers Under 45	Whole Number
Drivers with Car Assigned	Whole Number
Total Drivers	Whole Number
Total Unique Cars	Whole Number

Table 5 – DAX Measures D Driver

- **D Client**

During the exploration of the client data, we identified specific measures that could help us profile our customer base. Therefore, we created the following DAX measures, ensuring each was properly formatted to enhance clarity in visualizations:

Measure Name	Data Format
% Young Clients	Percentage
Avg Client Age	Decimal (1 decimal place)
Total Clients	Whole Number

Table 6 – DAX Measure D Client

The *Client Age Group* column, derived during data preparation, was also included in the semantic model to enable meaningful aggregations and filtering by age segments.

- **D Date**

In the case of the *D Date* table, we opted for a more refined configuration, as this dimension plays a central role in time-based analysis across the report. Our group decided to hide most of the technical or redundant columns such as various versions of month, quarter, semester, and weekday labels (e.g., Month_Name_Short, Quarter_Short, Weekday_Type, etc.), these were mostly different textual formats that did not add analytical value for our specific visual needs.

To enhance analytical capabilities, we developed calculated DAX measures that were time intelligent that allowed more dynamic filtering and conditional logic throughout the report. These measures included:

Measure Name	Data Format
Current Year	Whole Number
Is Current Year	True/False

Table 7 – DAX Measures D Date

We also created a Year Hierarchy to simplify drill-down capabilities in our Power BI visualizations. This decision was essential to allow users to analyze trends across multiple time granularities (from year down to day) without having to build separate visuals for each level.

Year Hierarchy: Year → Semester → Quarter → Month → Day

- **D Car**

In the case of the *D Car table*, our group adopted a similar methodology as with other dimensions by hidding all technical or irrelevant fields

Moreover, we structured a hierarchy named Manufacturer Hierarchy to support more insightful drill-down analyses in Power BI visualizations, particularly useful for brand-related insights:

Manufacturer → Brand → Year of the Car

For the Dax we created:

DAX Measure Name	Data Format
Car Age	Whole number
Cars by Brand	Whole number
Newest Car Year	Whole number
Oldest Car Year	Whole number
Total Cars	Whole number

Table 8 – DAX Measures D Car

These measures helped us understand patterns such as fleet composition and age distribution across different manufacturers and brands, ultimately contributing to the Car Analysis and Efficiency Analysis pages in our report.

- **D Location**

For the *D Location* table we hid the technical keys and to facilitate geographic drill-downs in Power BI, we created a custom hierarchy named Country Hierarchy, composed of the following levels:

Country → City → District

- **D Time**

Recognizing the importance of time-of-day analysis in ride-related insights, we created a hierarchy titled Period of Day Hierarchy:

Period of Day → Hour → Minute

This hierarchy proved particularly useful in evaluating demand fluctuations across different times of the day and supports intuitive filtering in visuals.

- **D Currency**

Regarding the *D Currency* table, our group decided to hide all fields, including the surrogate key and currency descriptors. Since currency variations did not add meaningful differentiation or value to our insights or visualizations, we opted not to include this dimension in the analytical layer.

- **F Rides**

As the central component of our star schema, the F Rides fact table aggregates the core operational data of the entire system. Given its high volume and level of granularity, our group made several key decisions to improve clarity and analytical utility.

We created a substantial number of DAX measures directly within this table to support our business questions and drive advanced insights. These measures cover monetary indicators (like revenue per ride or per km), resource efficiency (gas consumption per km), and operational performance (average duration, distance, and total rides).

Where appropriate, we formatted each measure for clarity and consistency (e.g., using decimal formatting for duration, currency for amount-based metrics, percentage for ratios). The table below summarizes all DAX measures and their corresponding formats:

DAX Measure	Format
Amount Actual	Currency (123 €)
Amount Target	Currency (123 €)
Avg Amount per Ride	Currency (123 €)
Amount per City	Currency (123 €)
Avg Gas per Km	Decimal Number
Avg Km by Period	Decimal Number
Avg Ride Duration	Decimal Number
Amount per Km	Currency (123 €)
Duration per Km	Decimal Number
Gas Consumption per Km	Decimal Number
Total Rides	Whole Number

Table 9 – DAX Measures F Rides

10. POWE BI REPORT

10.1. General Story Telling Approach

With the semantic model fully prepared and refined, our group moved on to the development of the Power BI report, the core deliverable to visually communicate the business insights.

Rather than simply presenting disconnected metrics, we adopted a storytelling-driven approach. Our goal was to construct a coherent and logical narrative that allows users to navigate through the data in a structured, business-oriented manner, while answering key operational and strategic questions.

To achieve this, we divided the report into five thematic pages, each targeting a specific perspective of the business ecosystem:

1. **Client Analysis** – Who are our clients?
2. **Driver Analysis** – Who are our drivers?
3. **Car Analysis** – What is the composition and age of our fleet?
4. **Operational Analysis** – How are we performing in terms of revenue and ride volume?
5. **Efficiency Analysis** – How efficiently are we using our resources (e.g., fuel, distance, periods of day)?

10.2. Visual Approach and Structure

When designing our Power BI report, we aimed for a clean, consistent, and user-friendly interface that could support both readability and insight extraction.

We decided to use a blue-based color palette throughout the report, ensuring visual coherence and helping the user stay focused on the analytical content. Within charts and comparisons, we applied slightly different shades of blue to separate categories without introducing unnecessary distraction.

Titles across all pages were centered, written in uppercase, and bold, giving immediate clarity to each page's analytical focus. We placed these titles at the top to define the subject before presenting any data.

For the most relevant figures, we created highlight cards and chose to place them horizontally, most of the time, at the top of each page. These cards were formatted

with drop shadows and soft borders, creating slight elevation and helping them stand out without being intrusive.

Charts and tables were selected based on their ability to present comparisons, trends, or distributions effectively. For example, we chose bar charts for comparing categories like gender or brands, and line charts for time series like forecasted rides. In some pages, tables were used when detailed views were more appropriate.

We also incorporated hierarchies (e.g., in date, time, or location) to support drill-downs directly in visuals, allowing users to move from overview to detail within the same context.

To support interactivity, we included slicers aligned on the right side of each page, styled with light frames and shadow effects to separate them visually from the main content area. These slicers enable optional exploration without interfering with the default storytelling flow.

Each page follows a top-to-bottom reading flow, starting with summary figures, followed by analytical visuals, and finishing with tools for deeper exploration. The five pages are structured to follow a logical narrative: beginning with user profiles (clients, drivers), continuing with asset analysis (cars), and finishing with operational and efficiency evaluations.

10.3. Client Analysis

This page in the Power BI presents a demographic overview of the platform's client base, with a focus on age and gender segmentation.

Business Questions Addressed

- **How many clients are registered on the platform?**

  The platform currently has 10,000 clients.

- **What is the average age of the client base?**
  The average client age is 34.67 years.

- **What proportion of clients are considered young (under 30 years old)?**
  A total of 40.0% of clients fall into the “young” segment.

- **What is the gender distribution among clients?**
  The client base is composed of 56.14% male and 43.86% female users.

- **What is the average age by gender?**
  - Male clients: 35.67 years
  - Female clients: 33.39 years

10.3. Driver Analysis

This page focuses on profiling the platform’s drivers, aiming to provide insights into their age distribution, gender, and volume.

Business Questions Addressed

- **How many drivers are registered on the platform?**
  The platform has a total of 1,000 drivers**.**

- **What is the average age of the drivers?**
  The average driver age is 39.84 years.

- **What proportion of drivers are considered young (under 40 years old)?**
  A total of 58.3**%** of drivers fall into the “young” segment.

- **What is the gender distribution among drivers?**
  The driver base is composed of 65.4% male and 34.6% female drivers.

- **What is the average age by gender?**
  - Male drivers: 39.1 years
  - Female drivers**:** 36.3 years

10.4. Car Analysis

This page provides a structural overview of the platform's fleet, with a focus on vehicle distribution by brand and model year.

Business Questions Addressed

- **How many cars are registered in the platform's fleet?**
  There are a total of 500 cars in the system.

- **What is the most recent car manufacturing year in the fleet?**
  The newest car was manufactured in 2022.

- **What is the oldest car manufacturing year in the fleet?**
  The oldest car in the fleet is from 2017.

- **Which brands are most represented in the fleet?**
  The most dominant brands are **SEAT** and **Peugeot**. Both brands appear multiple times in the top 5 most common car types in the fleet.

10.5. Operational Analysis

This page explores core operational indicators of the ride-sharing service, focusing on revenue trends, volume of rides, and time-based performance breakdowns.

Business Questions Addressed

- **How many rides have been completed on the platform?**
  The platform has registered a total of 1,000,000 rides.

- **What is the average amount charged per ride?**
  The average fare per ride is 10.27€.

- **What is the total actual revenue generated?**
  The total actual revenue is 10.77€ million.

- **How does actual performance compare to revenue targets over time (KPI)?**
  A visual comparison between actual and target revenue shows that in 2020, the platform fell short of the 3 million € target, achieving 2.65 million €. However, in the two subsequent years, revenue grew faster than the 10% target benchmark. In 2022, actual revenue reached 3.44 million €, surpassing expectations. In 2023, however, revenue dropped to 1.72 million €, while the target for that year was 3.993 million €, a 56.6% decrease compared to the previous year.

- **Which period of the day generates the highest revenue?**
  Revenues are relatively balanced across the day, with the afternoon period showing slightly higher values, and night being the lowest.

- **Which countries or regions generate the most revenue?**
  The only country represented in the data is Portugal, with the Aveiro district recording the highest volume of rides and revenue.

10.6. Efficency Analysis

This final page evaluates performance efficiency based on fuel consumption and travel distance across different brands and time periods.

Business Questions Addressed

- **Which car brands consume the most fuel?**

  A pie chart reveals that SEAT Ibiza accounts for the highest total gas consumption across the fleet, while the Audi A4 is the most fuel-efficient among the observed brands.

- **What is the total distance traveled by period of the day?**

  A line chart shows that while the number of rides remains relatively constant across periods, the distance traveled is significantly higher during the daytime. In contrast, night-time trips tend to be shorter.

- **What is the average distance of a ride?**

  The average ride distance across the dataset is 52.04 kilometers, providing a useful baseline for evaluating operational patterns and vehicle usage.

11. CONCLUSION

Throughout the analysis of RIDE4ALL, we uncovered a story of a young, dynamic platform serving a growing number of users and navigating the complexities of operational growth. The data revealed a strong connection with a younger customer base, suggesting that the service is well-positioned among newer generations of urban users. At the same time, we noticed a workforce of experienced drivers supporting this demand, creating a balance between digital innovation and human infrastructure.

The growth trajectory, however, was not linear. While the platform showed promising revenue gains in the early years, even surpassing targets at certain points, 2023 brought a noticeable decline. This sudden drop broke the trend and highlighted the importance of continuously aligning expectations with market behavior. It also suggested that previous growth may not be guaranteed going forward — a reality that invites further investigation.

In the background of these patterns, the fleet played a central role. Concentrated around a few key brands and models, the vehicle pool revealed signs of efficiency and consistency, but also raised questions about over-dependence on specific types. Ride behavior remained steady across the day, yet subtle variations such as longer trips during daylight and shorter ones at night , gave us insight into how the service is used, when and by whom.

This project was our first experience using Microsoft Fabric and Power BI. Navigating new tools while managing a complex dataset presented challenges, but also valuable learning opportunities. We managed to build a coherent model, develop a structured report, and extract actionable insights. Looking back, there is still room to improve, particularly in refining the visual design and in finding better ways to address the business questions we couldn’t fully explore. Nonetheless, this work laid a strong foundation for more advanced analysis in the future.