# Business Intelligence Project Report - ***Analysis of UFO sightings***

**Trainees:** Ana Carneiro, Cidália Castro, Johab Santos
**Numbers:** 8068787, 5579374, 7018655
**UFCD:** 10804 - Business Intelligence
**Delivery Date:** March 28, 2025

Work carried out within the scope of the IEFP Business Intelligence course, with the aim of applying data analysis and visualization techniques that allow us to explore and detect patterns in UFO sightings.

# Dedication

We dedicate this project to everyone who believes in the power of knowledge, persistence and teamwork.

To those who do not give up in the face of challenges and who overcome themselves at each new stage. To ourselves: Cidália, Ana Carolina and Johab for their commitment, resilience and desire to evolve, learn and always do better.

This work is the result of joint effort, but also of individual passion for growing professionally.

# Acknowledgements

This project is the result of true teamwork, developed in collaboration with colleagues Cidália, Johab and Ana Carolina. Each person contributed ideas, skills and effort, and this made all the difference in the final result.

We would like to give special thanks to the trainers who accompanied us throughout the training:

- Ricardo Pinto Novo, for guidance in the areas of data ingestion and visualization;
- Dionísio Creoulo, for sharing solid knowledge in data management and storage, SQL and standardization;
- Ana Sofia Teixeira, for the practical and in-depth approach to transformation, modeling and advanced data analysis;
- Rúben Folha, for his contribution to the storytelling component, essential for communicating results with clarity and impact.

Finally, a special thank you to IEFP, for providing this learning opportunity, for the resources made available and for the commitment to excellent professional qualification.

# Epigraph

*"Data can be the key to understanding the unknown, but only when we are willing to look beyond the conventional."*

Jacques Vallee

# Summary

Throughout the ages, human beings have always observed the sky in search of answers to the mysteries of the universe. This ancestral curiosity is still alive, fueling questions about what is beyond our understanding. It was with this spirit of investigation that this Business Intelligence project was developed, centered on the analysis of sightings of Unidentified Flying Objects (UFOs), based on a dataset from the Kaggle platform.

The work covered all stages of the BI cycle: from collecting, cleaning and normalizing data on a relational basis in MySQL to creating analytical dashboards in Power BI. Techniques for data modeling, outlier treatment, column formatting and variable categorization were applied. Due to technical limitations in connecting to the SQL server, the process was adapted by importing CSV files.

More than visualizing data, the project sought to interpret and extract meaning from information: understanding what the data communicate, identifying patterns, crossing variables and raising hypotheses. This approach allowed us to understand complex contexts in an analytical and quantitative way, draw relationships and, who knows, get us a little closer to the truth… after all, is it really out there?

**Keywords**: Business Intelligence, UFO, UFO, Data Visualization, SQL, Power BI, Data Analysis

# Abstract

Throughout the ages, humans have looked to the skies in search of answers about the universe that surrounds them. This ancestral curiosity continues to inspire questions about what lies beyond our understanding. With that investigative spirit, this Business Intelligence project was developed to analyze Unidentified Flying Object (UFO) sightings, using a dataset from the Kaggle platform.

The project followed all stages of the BI lifecycle: from data collection, cleaning, and normalization in a relational SQL Server database, to the construction of analytical dashboards in Power BI. Techniques such as data modeling, outlier treatment, column formatting, and variable categorization were applied. Due to technical limitations in the direct connection to the SQL server, the team adapted the workflow using CSV file imports.

More than just data visualization, this project aimed to extract meaning from information — identifying patterns, crossing variables, and forming hypotheses. This approach enabled the team to analyze complex contexts quantitatively and uncover relationships that may bring us closer to an answer. After all… is the truth really out there?

**Keywords**: Business Intelligence, UFO, Data Analysis, Data Visualization, Power BI, SQL

# Index

# Index of Tables and Figures

# 1. Introduction

Human beings have always been fascinated by the unknown. Among the many mysteries that surround us, sightings of Unidentified Flying Objects (UFOs) continue to arouse curiosity, speculation and scientific interest. The analysis of sighting records is, therefore, an opportunity not only to explore patterns of behavior and geographic distribution, but also to apply modern data processing and visualization methodologies.

In this context, this Business Intelligence project appears as an academic and technical challenge, aiming to apply the complete BI cycle to a real and complex dataset. Through the use of tools such as Excel, MySQL and Power BI, we sought to structure and analyze a database of UFO sightings, collected on the Kaggle platform.

This introduction marks the starting point for an analytical journey that aims to go beyond simple visualization, with the aim of interpreting data, identifying patterns and launching hypotheses on one of today's most enigmatic and thought-provoking topics.

# 2. Objectives

## 2.1 General Objective

Develop a complete Business Intelligence project applied to the analysis of sightings of Unidentified Flying Objects (UFOs), using a real dataset from the Kaggle platform, with the aim of applying, integrating and consolidating skills in data collection, transformation, modeling, analysis and visualization.

## 2.2 Specific Objectives

- Explore and understand the structure of a real and semi-structured dataset;
- Clean and normalize data in a relational database created in SQL Server;
- Apply good data modeling practices, ensuring integrity and efficiency in relationships between tables;
- Import, transform and prepare data in Power BI, correcting formats and treating outliers;
- Create calculated columns and DAX measures for advanced analytics;
- Develop interactive dashboards that allow identifying geographic, temporal and behavioral patterns in sightings;
- Stimulate analytical and critical reasoning when faced with complex data, promoting evidence-based decision making.

# 3. Data collection and processing

The first stage of the project consisted of selecting and understanding the dataset to be worked on. The data source chosen was the platform Kaggle, known for the diversity and quality of its public datasets. The research was guided by thematic interest (UFO sightings) and the potential complexity of the available data, which would allow us to apply all phases of the Business Intelligence cycle.

The selected dataset contained two main files: complete.csv and scrubbed.csv. We chose to use scrubbed.csv, as it presented a cleaner version, without incomplete or invalid records, making it easier to start processing and modeling.

## 3.1 Pre-Analysis in Excel

Before creating the database, we carried out an initial exploratory analysis in Microsoft Excel, with the following objectives:

- **Duplicate removal**: We identify and eliminate repeating rows based on column combinations such as date, location and description.
- **Creating a unique identifier (ID)**: We assign an ID to each sighting to facilitate indexing and later data normalization.
- **Evaluation of available columns**: We identified fields that had a high rate of null values, inconsistent names, or inappropriate formats.

## 3.2 Problems detected in the original dataset

During this phase, several problems that required intervention were identified:

- Column names with spaces and special characters, which makes them difficult to use in SQL and Power BI. Examples: Duration (seconds) and Date Posted;
- The DateTime column had to be renamed as it conflicted with the SQL reserved word;

- Inconsistent formats in date and duration columns, with ambiguous entries or in text format;
- Fields with low relevance or low quality, such as:
    - Posted: date of publication of the sighting, with no direct analytical use;
    - City and Country: with many null values, choosing to keep the longitude and latitude and later renaming the column based on this more specific data.

This first analysis made it possible to establish the foundations for the creation of the relational database, ensuring that only relevant data with minimum quality conditions were maintained for the following phases of the project.

# 4. Database Creation and Data Insertion

After preliminary processing of the data in Excel, we began building the relational database using MySQL Workbench, with the aim of organizing the information in a structured way, facilitating the normalization process and allowing efficient queries.

## 4.1 Database Creation

A database was created with the name area52, in allusion to the well-known “Area 51”, reinforcing the project’s identity and its connection to the theme of UFOs. The creation of the database was done through the MySQL Workbench graphical environment, using SQL commands and tools available on the platform.

## 4.2 Data Insertion

The main table with the raw data was imported directly from the previously processed file area52.csv using the Table Import Wizard, an integrated MySQL Workbench tool that allows you to quickly and intuitively import CSV files into already created tables.

# 5. Database Normalization

After importing the raw data into MySQL Workbench, we proceeded to the normalization, com the objective of ensuring an efficient database structure, without redundancies, and which would allow for more robust analyzes in Power BI.

Standardization followed classical principles until 3rd Normal Form, ensuring that each table represented a distinct entity, with well-defined relationships and referential integrity maintained through primary and foreign keys.

From the original area52 table, three main entities were identified and justified the separation into their own tables:

- **Location**: latitude e longitude
- **Shape**: type of object sighted
- **Duration (duration)**: duration of the sighting

## 5.1 Assigning IDs, Creating Tables and Inserting Data

The following steps were performed:

1. Addition of ID columns to the area52 table, as indicated in *figure 1*
    a. The location_id, shape_id and duration_id columns were created;
    b. We use the ROW_NUMBER() function in subqueries to generate unique IDs for each distinct value.

```sql
-- Criar a coluna `location_id`
ALTER TABLE area52 ADD COLUMN location_id INT;

-- Atribuir IDs únicos a localizações com base em latitude e longitude
UPDATE area52 a
JOIN (
    SELECT latitude, longitude,
           ROW_NUMBER() OVER (ORDER BY latitude, longitude) AS location_id
    FROM (SELECT DISTINCT latitude, longitude FROM area52) AS unique_locations
) temp ON a.latitude = temp.latitude AND a.longitude = temp.longitude
SET a.location_id = temp.location_id;
```

*Figure 1: SQL script for creating columns and assigning unique IDs*

2. Location_id assignment, as shown in *figure 1*
    a. Unique IDs were assigned based on the latitude + longitude combination.
3. shape_id assignment
    a. Unique IDs were assigned to each distinct shape.
4. duration_id assignment
    a. The durationhoursmin column has been dropped, and IDs have been assigned based on durationseconds only.
5. Creation of Standardized Tables, as we can see in *figure 2*
    a. Three standard tables were created:
        i. locations (id, latitude, longitude)
        ii. shapes (id, shape_name)
        iii. durations (id, durationseconds)\

```sql
-- Criar a tabela `locations`
CREATE TABLE locations (
    id INT PRIMARY KEY AUTO_INCREMENT,
    latitude DOUBLE NOT NULL,
    longitude DOUBLE NOT NULL,
    UNIQUE(latitude, longitude)  -- Evita duplicação
);

INSERT INTO locations (latitude, longitude)
SELECT DISTINCT latitude, longitude FROM area52;
```

*Figure 2: SQL script for table creation and data insertion*

6. Creation of the Final sightings table
    a. The main sightings table was then created, consisting of the following fields:
        i. id_sighting (PK)
        ii. datetime_sighting (text format)
        iii. location_id (FK)
        iv. shape_id (FK)
        v. duration_id (FK)
        vi. comments (texto)
7. Data Migration
    a. The data was transferred from the original area52 table to the new structure based on INSERT INTO and SELECT, as indicated in *figure 2*, respecting the links between tables. Finally, the validation migration, with a query to check whether there were null values in the connection fields (location_id, shape_id, duration_id);
    b. After validation, the area52 temporary table was deleted, completing the transition to the normalized relational structure.

# 6. Data Import, Cleansing and Transformation

With the relational database finalized, we moved on to the ETL (Extract, Transform, Load) phase in Power BI, where the data was prepared for visualization and analysis.

## 6.1 Importing Data into Power BI

The initial approach was to establish a direct connection between Power BI and the MySQL database through the native connector. The connection was successful and allowed you to import the necessary tables into the Power Query environment.

However, as we performed more complex transformations or relevant changes in Power Query, the connection to the database became unstable, resulting in timeout errors and lost connections to the server.

This repeated problem compromised the continuity of model development in Power BI and prevented the fluid workflow necessary at this stage of the project.

Given the technical constraint, we made the decision to export all MySQL normalized tables to CSV files. This alternative guaranteed full control over the data and allowed:

- Import the files directly into Power BI;
- Work locally without depending on the stability of the connection with the server;
- Accelerate the transformation and modeling process in Power BI.

## 6.2 Relationship between tables

Before starting any transformation or cleaning, it was ensured that the relationships between the tables were correctly defined in the Power BI data model, replicating the relational structure created in MySQL. These connections were based on foreign keys:

- sightings[location_id] → locations[id] - [N - 1]
- sightings[shape_id] → shapes[id] - [N - 1]
- sightings[duration_id] → durations[id] - [N - 1]

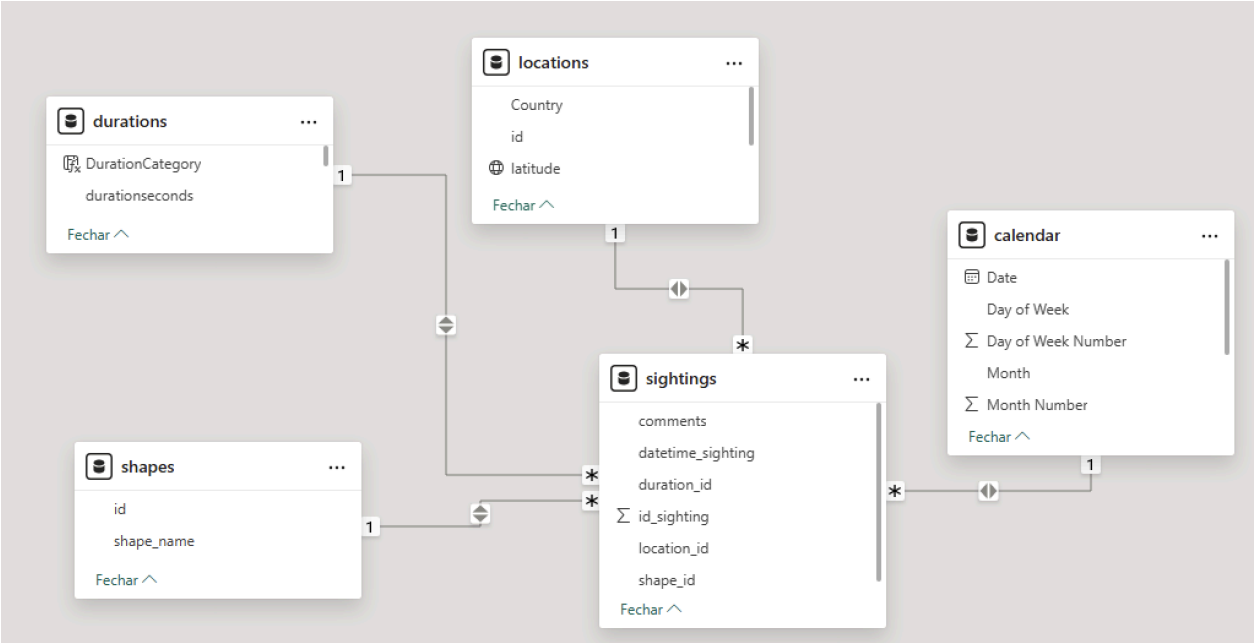



*Figure 3: Relationships between tables in Power BI*

## 6.3 Transformation and Cleaning

Within Power Query, several data cleaning, enrichment and preparation operations were carried out, namely:

- **Data type conversion**: dates, numbers and texts have been adjusted to the correct types;

- **Handling null or inconsistent values**: incomplete or invalid lines were removed;
- **Textual cleaning**: excess spaces and special characters were removed and text fields such as comments and shape_name were standardized;
- **Geographic identification**: to enrich the spatial analysis, a request was made via API based on the coordinates (latitude and longitude) to obtain the name of the corresponding location, through a personalized column, as per *figure 4*;

```
let
    Latitude = [latitude],
    Longitude = [longitude],
    URL = "https://nominatim.openstreetmap.org/reverse?format=json&lat=" & Text.From(Latitude)
& "&lon=" & Text.From(Longitude),
    JSON = try Json.Document(Web.Contents(URL)) otherwise null,  // Evita erro caso a API não
retorne resposta
    Address = try JSON[address] otherwise null,  // Se não houver "address", retorna null
    City = try Address[city] otherwise try Address[town] otherwise try Address[village] otherwise
try Address[municipality] otherwise "Unknown city",
    Country = try Address[country] otherwise "Unknown country",
    Cidade_Pais = if City = "Unknown city" and Country = "Unknown country" then "Unknown
location" else City & ", " & Country
in
    Cidade_Pais
```

*Figure 4: M language for connecting to the API*

These transformations ensured that the data was clean, consistent, and ready for analysis.

## 6.4 Treatment of Outliers

One of the most critical aspects of data preparation was the treatment of outliers in the durationseconds column, which presented extremely high values and outside the expected standard for a sighting.

To this end, a statistical approach based on the Interquartile Range (IQR) was used, but implemented directly in DAX, with the following steps:

1. Calculation of the 1st Quartile (Q1) and 3rd Quartile (Q3) for durationseconds, with DAX measurements, as per *figure 5*:

```
Q1_Duration = PERCENTILEX.INC(durations, durations[durationseconds], 0.25)

Q3_Duration = PERCENTILEX.INC(durations, durations[durationseconds], 0.75)
```

*Figure 5: DAX measurements for calculating the 1st and 3rd quartile*

2. Creation of a new column with a logical filter that identified whether or not a record was an outlier, as per *figure 6*:

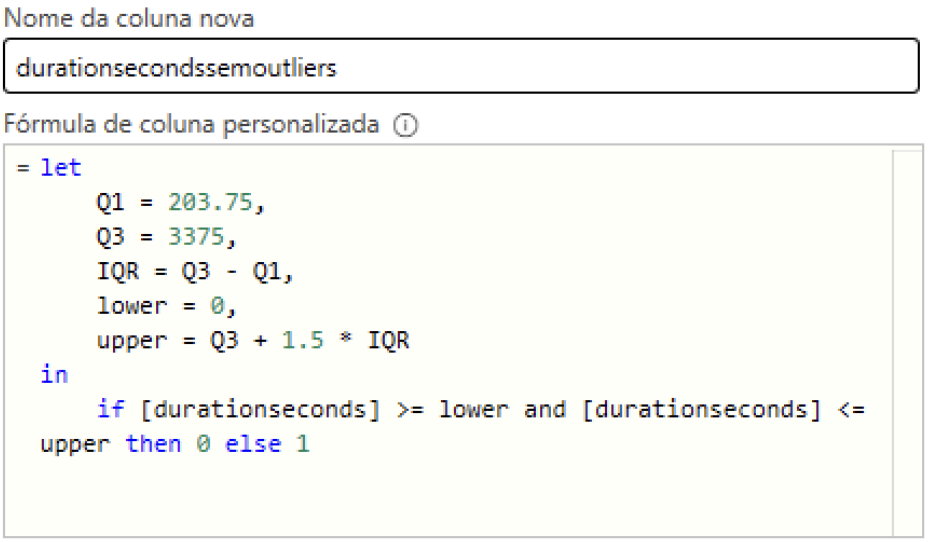



*Figure 6: M language for creating custom columns and identifying outliers*

After creating this custom column, we filter the 0 values, thus removing outliers.

## 6.5 Complementary processes outside of Power Query

After the ETL phase, other operations were carried out in the modeling area and using DAX:

- Categorization of the duration of sightings based on the durationseconds column, creating groups such as Very Short, Short, Long and Very Long, as per *figure 7*;

```
DurationCategory =
SWITCH(TRUE(),
    Durations[durationseconds] <= 60, "Very Short",
    Durations[durationseconds] <= 300, "Short",
    Durations[durationseconds] <= 1800, "Medium",
    Durations[durationseconds] <= 7200, "Long",
    "Very Long"
)
```

*Figure 7: Creating a column in DAX for duration categorization*

- Creation of a calendar table, as per *figure 8*, for detailed temporal analysis, identifying seasonal patterns and trends over time. This calendar was related to the sightings table across the field *date*, as previously demonstrated (*figure 3)*;

```
calendar =
ADDCOLUMNS (
    CALENDAR (DATE(1949,1,1), DATE(2015,12,31)),
    "Year", YEAR([Date]),
    "Month", FORMAT([Date], "MMMM"),
    "Month Number", MONTH([Date]),
    "Day of Week", FORMAT([Date], "dddd"),
    "Day of Week Number", WEEKDAY([Date]),
    "Quarter", "Q" & FORMAT([Date], "Q"),
    "Year-Month", FORMAT([Date], "YYYY-MM")
)
```

*Figure 8: Creating a table in DAX for calendar*

- Creation of a DAX measure to calculate the average sighting duration in a readable format (minutes and seconds), avoiding confusing representations such as 1.5K, as per *figure 9*;

```
AverageDurationFormatted =
VAR TotalSeconds = AVERAGE(Durations[durationseconds])
VAR Minutes = INT(TotalSeconds / 60)
VAR Seconds = ROUND(MOD(TotalSeconds, 60), 0)
RETURN
Minutes & " min " & Seconds & " sec"
```

*Figure 9: DAX measure for calculating duration average in readable format*

# 7. Dashboard Construction

With the data prepared, cleaned and organized in a solid relational model, we began building visualizations in Power BI, with the aim of exploring and communicating relevant insights in a clear, interactive and visually attractive way.

Various types of visualizations were developed, each with different purposes, from the presentation of global indicators to the geographic, temporal and categorical analysis of sightings.

## 7.1 Indicator Cards (KPI)

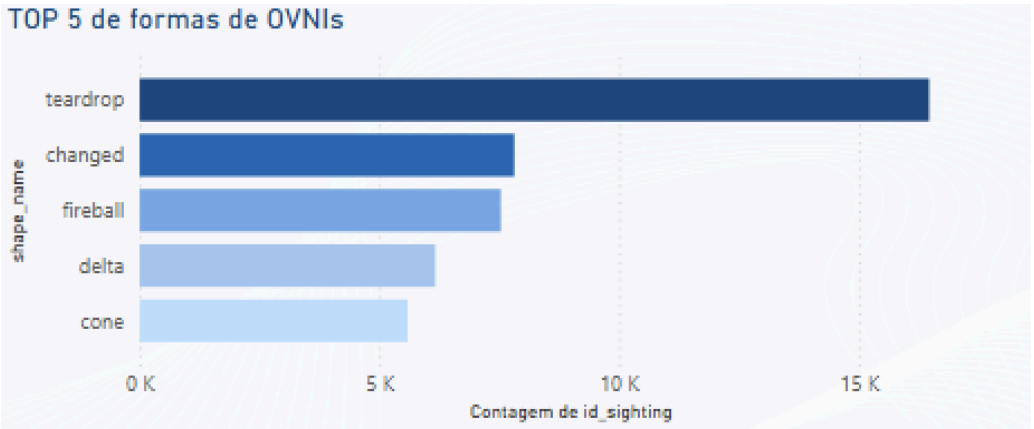
They were used simple cards to highlight high-level metrics:

- **Total Sightings**: Global count of valid records.
- **Average Duration of Sightings**: Created with DAX measure that converts the average of seconds to a readable format in minutes and seconds, avoiding the use of values such as 1.5K, which made interpretation difficult, as per *figure 9*.

## 7.2 Top 5 Ways

A horizontal bar chart was created to show the five most frequently reported formats, as *figure 10*. This visualization allowed us to understand which types of objects were most observed.

*Figure 10: TOP 5 forms of UFOs*



## 7.3 Top 5 Countries

The column chart was constructed to highlight the five countries with the highest number of records, helping to identify geographic areas with the highest incidence of sightings, as per *figure 11*.


*Figure 11: TOP 5 Countries with the Most Sightings*

## 7.4 Temporal Evolution

Created based on extraction of the year from the sighting date, this graph showed the annual evolution of the number of records, allowing us to observe the temporal trend, as *figure 12*.

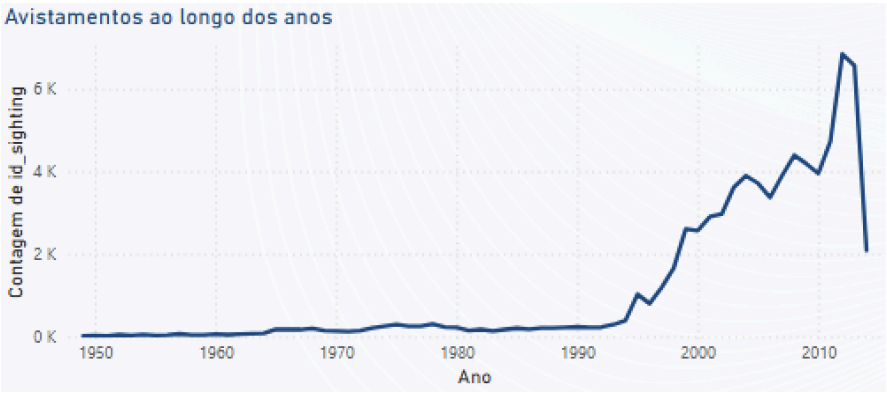



*Figure 12: Evolution of sightings over time*

## 7.5 Geographic Map

The map was built based on latitude and longitude fields, enabling geospatial visualization of sightings. Each point represents a sighting, with the possibility of interaction and cross-filtering with other views, according to *figure 13*.


*Figure 13: Sightings by location*

## 7.6 Word Cloud

Based on the comments column, previously cleaned in Power Query, a word cloud which highlighted the most used words in descriptions of sightings, as *figure 14*.

This visualization brought a qualitative layer to the analysis, highlighting recurring terms such as “light”, “disk”, “moving”, among others.


*Figura 14: Word Cloud*

## 7.7 Sightings by Duration Category

Pie chart representing the distribution of sightings by duration category (Very Short, Short, Long, Very Long), previously created with DAX. This visualization allows you to understand the prevalence of each type of sighting based on the duration, according to *figure 15*.

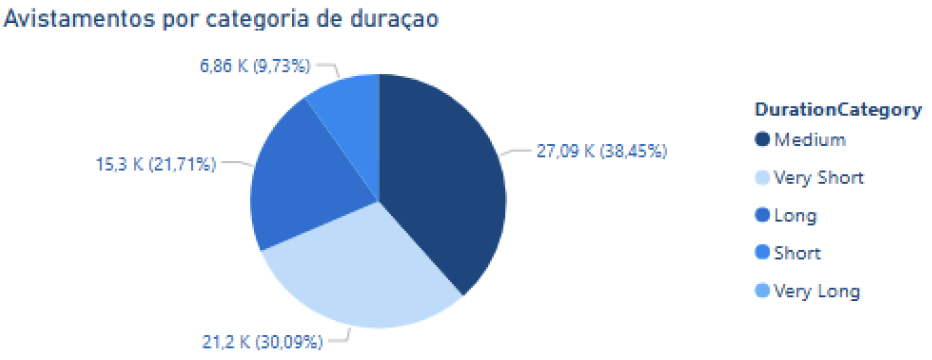



*Figure 15: Sightings by duration category*

## 7.8 Sightings by Day of the Week

Using the day of the week of the calendar created in DAX, a bar graph was constructed that highlights What days do most sightings occur?. This analysis allows you to identify possible behavioral or environmental patterns, as *figure 16*.


*Figure 16: Sightings by day of the week*

## 7.9 Interactive Segmenters

Segmenters have been added to allow the user to dynamically filter data based on:

- Year of sighting
- Country
- Shape type
- Duration category

These filters offer greater flexibility to analysis and improve navigation between different report views. However, due to the length of the data, filters were applied to the segmenters, so that only the 6 years with the most sightings, 5 countries with the highest number of sightings and, likewise, the 5 most sighted forms appear.

## 8. Dashboard - Overview


*Figure 17: Dashboard - Overview*

## 9. Dashboard - Exploration and Detail

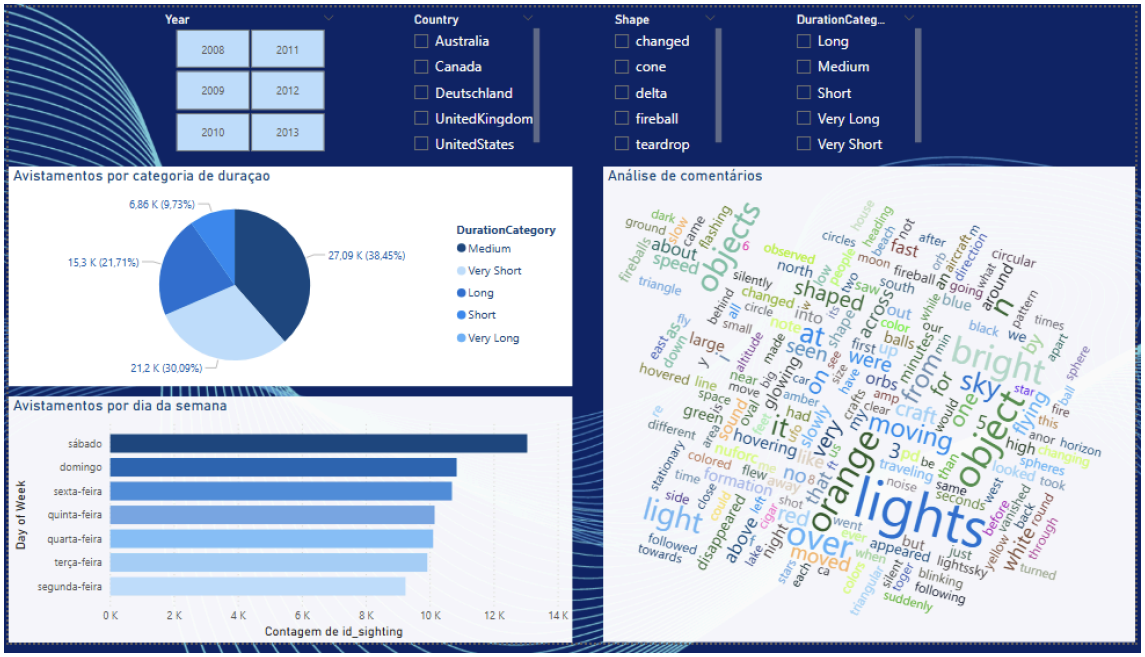



*Figure 18: Dashboard - Exploration and Detail*

# Conclusion

The development of this Business Intelligence project made it possible to apply, in a practical way, the entire data processing and analysis cycle, from collection, cleaning and modeling in SQL to the construction of analytical dashboards in Power BI.

On an analytical level, the data revealed relevant insights into UFO sightings. The **average duration of sightings** was from **21 minutes and 35 seconds**, and the total number of valid records exceeded **74 thousand**.

The geographic map showed a clear concentration of sightings in the USA, being also the countries with the most occurrences Canada, the United Kingdom, Australia and Germany.

With regard to the most seen forms, the categories **teardrop**, **changed**, **fireball**, **delta** and **cone**. The temporal analysis revealed a sharp increase in sightings from **1994**, reaching a peak between **2008 and 2012**, followed by a notable drop to **2014**.

In terms of behavioral patterns, it was observed that the majority of sightings are classified as **average duration**, and that the **Saturdays** and **Sundays** These are the days with the highest incidence of reports, which may be related to more free time, closer observation or social phenomena.

In addition to the results of the analysis, this project was a challenging exercise in adapting and solving real problems, namely the instability of the connection between Power BI and MySQL, which required the reformulation of the technical approach. The alternative found allowed the project to continue and reinforced the importance of flexibility and resilience in a BI context.

We conclude that, more than seeking definitive answers, the true value of data analysis lies in **exploring patterns, raising hypotheses and supporting decisions** evidence-based. And yes… maybe the truth is out there.

We analyze the unknown, face real challenges, and emerge stronger. We are Area52 and we are ready for the next puzzle!

# Bibliography and Webography


- Dataset UFO sightings
- Manual DAX Microsoft
- Manual MySQL

# Attachment list

Annex A - CSV file - mother table with first data cleansing

Annex B - MySQL Script - data creation, insertion and normalization

Appendix C - CSV files - extracted from database

Anexo D - Dashboard Power BI