# Project Scope

This project aims to detect and reduce e-commerce delivery fraud at Walmart by analysing delivery data from Central Florida. The goal is to identify patterns indicating items marked as delivered but not received by customers. Findings will help create a model to minimize theft and fraud across all U.S. regions.

# Contextualization

Walmart is the largest retail chain in the United States, generating an average of US$ 1.6 billion in daily revenue. In terms of in-store sales, Walmart averages US$ 17,000 per second, US$ 1.1 million per minute, and US$ 68 million per hour.

Its online platform and retail sales contribute a large portion of its profits. Walmart is also the largest food retailer in the U.S., with over US$ 264 billion in food sales last year.

According to a recent Lending Tree survey, self-checkout theft is a real issue, and individuals who engage in it are likely to repeat the behavior. Walmart has faced growing retail theft losses — estimated at US$ 3 billion in 2021, US$ 6.1 billion in 2022, and US$ 6.5 billion in 2023 — showing an increase of US$ 400 million in theft-related losses over the last year.

As a data scientist working in Walmart’s e-commerce division, you’ve identified that the largest proportional increase in theft occurred in online purchases where customers report missing items from their orders. Of the total theft growth from 2022 to 2023, 53% came from e-commerce transactions.

You were assigned to lead a project to reduce fraud and theft in Walmart’s e-commerce deliveries. It was identified that these thefts primarily occur during the delivery process.

The main goal of the project is to identify possible frauds in deliveries made by Walmart in Central Florida, which will serve as a model for other U.S. regions if successful. The focus is on analyzing delivery data to detect patterns and anomalies that could indicate that items reported as delivered were not actually received by customers. Walmart has received numerous consumer complaints about incomplete deliveries, and through data analysis, aims to determine whether the issue stems from the delivery drivers or other causes.

# Project Guideline

* **First Week (Week 43, 2025):** During the first week, it will be done an overall analysis for the project. We’ll define goals, analyse the databases to detect possible anomalies, brainstorm possible hypotheses, structure the project and do the first SQL queries.
* **Second Week (Week 44, 2025):** The second week will be dedicated to *data preparation*, where we will be analysing our databases and doing important data manipulations using *SQL*.
* **Third Week (Week 45, 2025):** We will dedicate this week for the *EDA (Exploratory Data Analysis)*, where we will export our data to *Jupyter Notebook* and perform data cleaning, initial data analysis, data segmentation and data visualization to be displayed in the final printed report.
* **Fourth and Fifth Week (Week 46 - 47, 2025):** Those two weeks will be dedicated to the end of the project. Here, we’ll design and build a *Power BI* dashboard. Other feature we’ll deliver is a *PowerPoint* presentation with all the important insights we found.

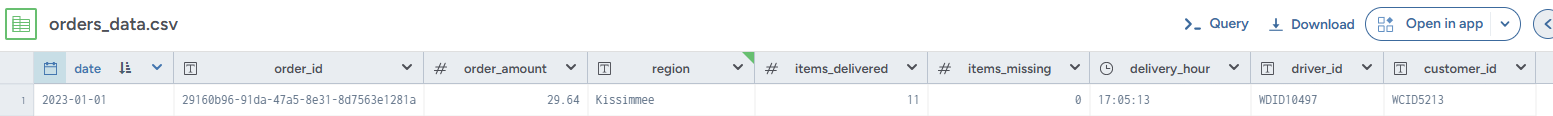
# 1. First Week (Week 43, 2025)

## First Insights

While analysing the data in data.world, we could check some cues of how to identify fraudulent drivers and costumers and some insights we can have about the missing packages.

1. **Repetitive costumers/drivers:** It’s not uncommon to have a missing product in the order, but what is unusual is having missing products in many orders. The first pattern we’ll check is people that had more than 3 orders with missing products in the year of 2023. This will indicate a high chance of costumer fraud. The same will be done with the drivers. We’ll do it by joining the datasets “orders\_data” with “customers\_data” and A screenshot of a computer

   AI-generated content may be incorrect.A close-up of a computer screen

   AI-generated content may be incorrect.“drivers\_data”.
2. A screenshot of a computer

   AI-generated content may be incorrect.A screenshot of a grocery list

   AI-generated content may be incorrect.**Data cleaning**: In the products dataset, we can see a very strange pattern on some products. A high frequency of food products are being categorized as “supermarket”, which does not match the logic of the segmentation. In the same dataset, we can check that others food products are being characterized in a more specific way, having been labelled such as “Snacks”, “Bakery” and “Beverages”. The same happen with products categorized as “Electronics”. We’ll use ChatGPT to remake the segmentation of the “Supermarket” products following the logic of the others ones. Since the product’s dataset has only 314 rows, this will be viable with GPT-5.0.
3. **Possible Segmentations:** As mentioned before, we can use the correct segmentation of the products to understand some patterns of lost items, such as which ones are the bringing more loss to Walmart. Other segmentations we can make is by “region” and “delivery hour”. They can bring useful insights to the company, revealing areas we need to be more careful with the delivery and the hours of the day where we have more packages with lost items. The “age” segmentation can also be insightful revealing the predisposition of a costumer to be fraudulent one. Uniting those segments, we can create a “persona” of a high chance of fraudulent person.

## Conclusions

After analysing the data we got to a few important conclusions that we must approach in this work. To make a good recommendation to the Walmart, we’ll answer the questions:

1. Who is the most probable persona of committing a fraud? How can we identify them? Where are they located with more concentration? Who are bringing more prejudice to Walmart: Frauds committed by customers or frauds committed by drivers? Who is more probable to commit a crime? How can Walmart reduce their prejudice with fraudulent people?

**Action:** Build an entire section for constructing the best representation of a fraudulent customer/driver. Make correlative comparisons with the data provided. Based on historical data, segment clients and drivers on least probable of being fraudulent to most probable of being fraudulent. We can perform a cluster chart to identify potential outliers.

1. How much did Walmart lose with all those frauds? Which were the locations with the most loss? Which are the products that are giving more prejudice? Which are the most crucial hours of the day, where most products are lost? How can Walmart prevent those losses?

**Action:** Build a section of our dashboard to display all the important insights found. We can split this dashboard into two parts for a more harmonic layout.

# 2. Second Week (Week 44, 2025)

During the second week of the project, we dedicated our time to perform data extraction from a database simulation. We uploaded the .csv documents, given from our client, and performed multiple SQL DQL (Data query language) commands for retrieving and joining the data. Although the existence of DDL (Data Definition Language), DML (Data Manipulation Language) and DCL (Data Control Language) is well-known, we chose to only extract the data from our databases and perform manipulation using python.

The first step was to create a centralized database with the customers and drivers datasets and join them with the orders dataset. We performed this action with the following query:

A screenshot of a computer program

AI-generated content may be incorrect.

With the missing orders with followed the same logic, joining the dataset *missing\_items\_data* with the *products\_data*. Since we had 3 columns with missing items in the the *missing\_items\_data* dataset, we decided to unpivot the data transforming the 3 columns of missing items into one. After this process, we joined the data with with the products\_data, to obtain a completer and more centralized dataset. We performed this action with the following query:

A screenshot of a computer code

AI-generated content may be incorrect.

After those two main queries, we decided to unify all of the information into a single dataset, making it easier to analyse it later with python. The final query was:

A screenshot of a computer program

AI-generated content may be incorrect.