## Problem Definition and Project Planning

### Problem Definition:

Our objective is to construct a predictive model to forecast sales for thousands of product families sold at Favorita stores in Ecuador. The training data encompasses dates, store and product information, promotional status, and sales numbers. Supplementary data includes store metadata, oil prices, holidays and events, which may influence sales patterns.

The challenge lies in analyzing these datasets to predict future sales, thereby enabling better inventory management and sales strategies for Favorita stores. Accurate predictions can help optimize stock levels, reduce waste, and improve customer satisfaction by ensuring product availability

### Project Planning

* ***Understanding the Objective:***

We clearly defined the problem in the problem definition. Our focus will be to develop the most accurate **submission.csv** file for the competition.

* ***Data Explanation:  
  Data Split:  
  The dataset is divided into several files:***
  + **train.csv**: Contains historical sales data for building machine learning models. This includes the target sales numbers.
  + **test.csv**: Used to evaluate model performance on unseen data. Sales numbers are to be predicted for the dates in this file.
  + **sample\_submission.csv**: A sample file to demonstrate the correct submission format.
  + **stores.csv**: Store metadata including city, state, type, and cluster.
  + **oil.csv**: Daily oil prices, providing economic context affecting sales.
  + **holidays\_events.csv**: Information about holidays and events that might influence sales.  
    ***Target Variable:****sales:* The total sales for a product family at a particular store on a given date.

***Key Features:****store\_nbr:* Identifies the store where products are sold.  
*family:* Identifies the type of product sold.  
*onpromotion:* Indicates the total number of items in a product family that were being promoted at a store on a given date.  
*date:* The specific date of the sales record.  
Additional contextual features from supplementary files include:  
*city, state, type, cluster:* Store metadata from stores.csv.  
*dcoilwtico:* Daily oil prices from oil.csv.  
*holiday\_type, transferred:* Holiday and event details from holidays\_events.csv.

* ***Import Libraries:*** *We will import the necessary libraries for data manipulation, analysis, and modeling.*
* ***Load and Check Data:*** *We will load the datasets and perform initial checks to understand their structure and contents.*
* ***Understand Dataset:*** *We will gain insights into the dataset by examining its characteristics and statistical properties.*
* ***Variable Description:*** *We will review the variables to understand their types (categorical or numerical) and relevance to the problem.*
* ***Exploratory Data Analysis (EDA):*** *We will conduct exploratory data analysis to uncover patterns and relationships within the data. Visualization will be a key tool in this step.*
* ***Basic Data Analysis:*** *We will perform basic data analysis to delve deeper into the dataset and extract meaningful insights.*
* ***Data Cleaning:*** *We will clean the data by addressing duplicates, missing values, outliers, and ensuring data consistency. This is a crucial step for any data science project.*
* ***Feature Engineering:*** *We will create new features and transform existing ones to improve the predictive power of our model. This step allows for creativity and domain-specific knowledge application.*
* ***Modeling:*** *We will select and train machine learning models that best fit our data. The goal is to find a model that provides accurate sales predictions.*

# Exploratory Data Analysis (EDA)

We can understand data deeper in Exploratory Data Analysis (EDA). In this step, we will do exploratory data analysis. We will focus on univariate variable analysis. We will do some visualization according to our data. We will separate data to categorical and numerical variables. Firstly we should look our categorical and numerical variables:

**Categorical Variables:** family, holiday\_type, locale, locale\_name, description, transferred, city, state, store\_type, day\_of\_week

**Numerical Variables:**

* **Discrete:** id, store\_nbr, cluster, year, month, week, quarter
* **Continuous:** sales, onpromotion, dcoilwtico, transactions

**DateTime Variable::** date

sales is our target variable

# Basic Data Analysis

We will do basic data analysis. Basic data analysis serves as the foundation for understanding and extracting valuable insights from raw data.

* Store Type - Sales
* Family (Product Category) - Sales
* Holiday Type - Sales
* Sales Distribution by Promotion Status
* Oil Price vs Sales Correlation
* Monthly Sales Trends
* Day of Week Analysis
* Store Cluster Performance
* City-Level Analysis

# Data Cleaning

In data cleaning, we have the most important step for the data science lifecycle. In a lot of projects, this step is %80 of the work. We will give importance to that data cleaning step. We will apply that steps:

* **Outlier Detection**We will focus on the IQR test for outlier detection.
* **Missing Values**We will find and fill missing values.

## Outlier Detection

We will do outlier detection. We have some outlier values in the dataset. If you pass this step, you can't have a good score in your model. You shouldn't pass!

You can find a lot of methods for outlier detection like IQR, z-score etc. . We will use the IQR test in this dataset.

Business-validated outliers detected: 46,748 (1.53% of total data)

**Decision: Retain Outliers** After careful analysis, we will preserve these identified outliers because they represent legitimate business scenarios, have statistical significance, provide value for robust modeling, and serve as natural indicators for exceptional events.