UNIVERSITA' DEGLI STUDI DI MILANO

BICOCCA

# US Retail Sales and Public Holidays Analysis

Abdul Sami - Mat: 939542

Davina David - Mat: 946851

Data Management Project 10 February 2026

Contents

# Executive Summary

This report presents a comprehensive analysis of the impact of public holidays on daily retail order volumes in the United States. The study integrates multi-year retail transaction data with official public holiday records retrieved via the **Nager.Date API.** Key findings demonstrate that public holidays significantly influence customer ordering behavior, with certain holidays generating notable increases in demand. Furthermore, incorporating holiday indicators into predictive models improves the accuracy of daily order volume forecasts. The project documents a reproducible data pipeline encompassing data acquisition, profiling, cleaning, integration, storage, exploratory analysis, and predictive modeling

# 1 Objectives and Research Questions

The primary objective of this project is to assess the influence of public holidays on retail order behaviour and evaluate whether holiday information enhances predictive models.

Research Questions

- RQ1: Do public holidays significantly affect daily e-commerce sales?
- RQ2: Are order volumes higher on holidays compared to non-holiday periods?
- RQ3: Which US holidays have the greatest impact on customer ordering patterns?
- RQ4: Can incorporating holiday information improve the prediction of daily order volumes?

# 2 Datasets and Data Acquisition

Two datasets were used: a retail orders dataset and a public holidays dataset.

## 2.1 Retail Orders Dataset

Source: Processed retail sales data provided as CSV files. Daily order transactions across multiple years.

Key attributes include:

| Attribute | Description |
|---|---|
| Order ID | Unique identifier for each order |
| Order Date | Date the order was placed |
| Customer ID | Unique customer identifier |
| Product ID | Identifier of the purchased product |
| Sales | Total order value |
| Quantity | Number of items purchased |

## 2.2 Public Holidays Dataset

Source: Public holiday API dataset. Official US public holidays across multiple years.

Key attributes include:

| Attribute | Description |
|---|---|
| date | Holiday date |
| Holiday Name | Name of the holiday |
| countryCode | Country code (US) |

# 3 Data Profiling and Quality Assessment

## 3.1 Initial Profiling Results

Initial profiling was conducted using Pandas to examine structure, completeness, and consistency.

Retail Orders Dataset:

- Total records: 9,994
- Missing values: 0
- Duplicate records: 0

Public Holidays Dataset:

- Total records: 60 holiday dates
- Missing values: 0
- Duplicate dates: 8 (13.33%)

## 3.2 Quality Metrics and Dimensions

The following data quality dimensions were assessed:

| **Dimension** | **Definition** | Method Used |
|---|---|---|
| Completeness | Percentage of records with no missing values in key fields. | **.isnull().sum()** |
| Uniqueness | Absence of duplicate orders and duplicate holiday dates. | **.duplicated().sum()** |
| Validity | Correct data types and acceptable value ranges. | **.dtypes, pd.to_datetime()** |
| Consistency | Matching **order dates** with holiday **dates** for integration. | Matching **Order Date** with **date** before merge |

## 3.3 Tools and Environment

All data processing, cleaning, integration, analysis, and modeling were implemented using Python in Jupyter Notebook (.ipynb) format.
Key workflows were documented across the following notebooks:

- **01_data_acquisition.ipynb — Data loading and API extraction**
- **02_profiling.ipynb — Data profiling and quality assessment**
- **03_cleaning.ipynb — Data cleaning and preparation**
- **04_integration.ipynb — Data integration and enrichment**
- **05_relational DBMS.ipynb — Storage and Validation using SQLite**
- **06_EDA.ipynb — Exploratory analysis and predictive modelling**

# 4 Data Cleaning and Quality Improvement

the data quality issues identified in the raw datasets and the cleaning actions applied to improve reliability, consistency, and integration readiness. Cleaning was executed primarily in **03_cleaning.ipynb**, with a focus on improving completeness, validity, uniqueness, and consistency. The cleaning process prepared both the retail orders and public holidays datasets for accurate merging and robust statistical analysis.

## 4.1 File Consolidation and Format Standardization

1. **Retail orders Data**

The retail orders dataset was provided as **four separate Excel files**, one per year:

| Year | Rows | Columns |
|---|---|---|
| 2014 | 1,993 | 25 |
| 2015 | 2,102 | 25 |
| 2016 | 2,587 | 25 |
| 2017 | 3,312 | 25 |
| **Total (raw)** | **9,994** | **25** |

Each Excel file was converted to CSV format.The four yearly files were merged into a single consolidated dataset. **0_Order_merge_RAW.csv**

2. **Public holiday Data**

Public holiday data was retrieved from the **Nager.Date public API** using four separate endpoints, one for each year (2014–2017). The JSON responses were combined into a single dataset and saved as a **holidays_api_raw.csv.** it contain the columns of **date, name,**

**countryCode,** and the **date** column was converted to datetime format to match the retail orders dataset.

| Year | API Endpoint | Records |
|---|---|---|
| 2014 | Public Holidays in United States 2014 - Nager.Date | 15 |
| 2015 | Public Holidays in United States 2015 - Nager.Date | 15 |
| 2016 | Public Holidays in United States 2016 - Nager.Date | 15 |
| 2017 | Public Holidays in United States 2017 - Nager.Date | 15 |

## 4.2 Duplicate Removal

Before cleaning we had:

- Retail orders dataset:
  - Records: 9,994
  - Duplicate rows: 0 (0.00%)
- Public holidays dataset:
  - Records: 60
  - Duplicate dates: 8(13.33%)

After removing duplicates, the public holidays dataset contained 52 unique holiday dates ready for integration.

## 4.3 Missing Value Handling

- The retail orders dataset contained no missing values in critical fields. **0 (0.00%)**
- The public holidays dataset also had no missing values. **0 (0.00%)**

## 4.4 Data Type Standardization

- Date fields (**Order Date** and **date**) were stored as object/string types.
- Boolean indicators were not explicitly defined.
- Converted all date fields to **datetime64[ns]** format.
- Created a binary **IsHoliday** column to standardize holiday identification

**Problem Faced**

- **Issue: Order dates** and holiday **dates** were stored as strings and used different formats.
- **Action:**
  Converted all date fields to **datetime** format before merging.

## 4.5 Column Consistency and Renaming

- Column naming conventions varied between datasets ( **date vs Order Date**).
- Some columns used inconsistent capitalization and spacing.
- Standardized column names to ensure compatibility during merging:
    - **date** into **Date**
    - Order Date retained for clarity
    - Created standardized fields: **IsHoliday, Holiday Name**

The Column names were consistent, readable, and compatible for joins and modeling.

## 4.6 Final Quality Assurance

Before finalization we had:

- Datasets were cleaned but not yet validated as integration-ready.

Summary of Final QA:

- Verified absence of duplicates and missing values.
- Confirmed correct data types for all columns.
- Ensured consistency between **order dates** and holiday **dates**.
- Validated row counts before and after cleaning.

After finalization we got:

- Both datasets **holidays_clean.csv** and **orders_clean.csv** were certified as clean, consistent, and ready for integration and analysis.

# 5. Data Integration and Enrichment

## 5.1 Data Integration Approach

The cleaned retail orders dataset was integrated with the cleaned public holidays dataset using a left join on the order date. This ensured that:

- All retail orders were preserved.
- Holiday information was appended where applicable.
- Non-holiday dates remained in the dataset with null holiday fields.

During the integration phase, the public holidays dataset contained 52 unique holiday **dates** across the four-year period after removing 8 duplicates. These dates were matched against the retail orders dataset. Since multiple orders can occur on the same holiday, this resulted in **450 order records** being labeled as holiday transactions in the integrated dataset.

## 5.2 Integration Results

Before integration we had:

- Retail orders: 9,994 records
- Public holidays: 52 unique dates

After integration we had :

- Integrated dataset: 9,994 records (no loss of retail data)
- Orders matched to holidays: 450 records (4.50%)
- Orders not on holidays: 9,544 records (95.50%)

Orders that do not match any holiday record are not errors. They simply represent orders placed on non-holiday dates. Since the public holiday API is authoritative, unmatched dates correctly indicate normal business days rather than missing or faulty data.

## 5.3 Enrichment Fields Added

The following attributes were added to the retail dataset:

- **Holiday Name** — Name of the holiday
- **IsHoliday** — Boolean flag indicating whether the order date was a public holiday

These features enabled subsequent exploratory analysis and predictive modeling related to holiday effects on retail sales.We had saved the integration as **orders_integrated_holidays.csv**

**Problem Issue**
**Issue: After merging, many orders had no matching holiday.**
**Action: Confirmed these are valid non-holiday orders, not errors.**

# 6. Storage Architecture, Database Queries and Validation

The integrated dataset of 9,994 retail orders and 52 US public holiday records was stored in a local **SQLite database** name **US_SALE**, chosen for its simplicity, portability, and support for structured queries. The database was accessed from Python using the **sqlite3** library.

Each record from the consolidated CSV was stored as a single row in the orders table, including order details and holiday attributes. Data was inserted in bulk to ensure efficient loading while preserving integrity.

## 6.1 SQLite Queries and Validation Methods

The dataset was queried and aggregated in SQLite to validate the analysis performed in Python. Key checks included:

- Counting orders on holidays versus non-holidays.
- Identifying holidays with the highest order volumes.

# 7 Exploratory Data Analysis and Research Question Answers

All research questions were answered using Python-based EDA, visualizations, and statistical modeling.

## RQ1 — Do public holidays significantly affect daily e-commerce sales?

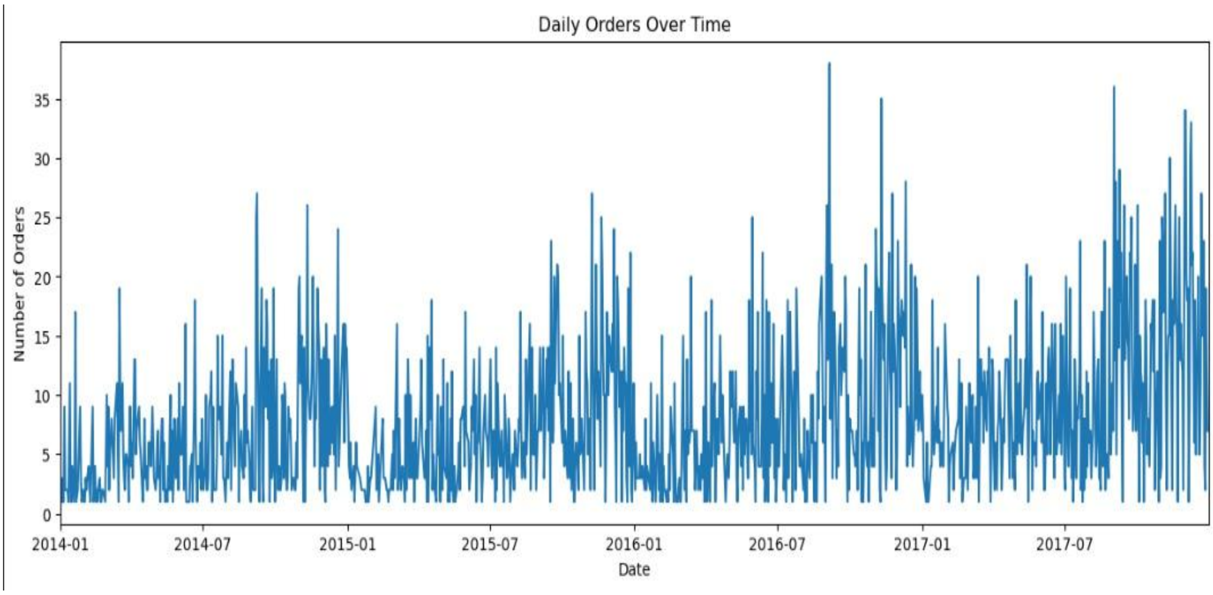

Figure 7.1 – Impact of Holidays on Daily Sales

Time series plot (figure 7.1) is used to identify the impact of holidays on Daily order sales volumes between **2014 and 2017** typically range from **2 to 10 orders per day** on normal days. However, multiple sharp spikes are visible where daily orders increase to approximately **30–38 orders**, indicating unusually high demand. These peak values occur intermittently and stand well above the average daily level. This pattern suggests that specific events, such as **public holidays, are associated with significantly higher order volumes**.

## RQ2 — Are order volumes higher on holidays compared to non-holiday periods?

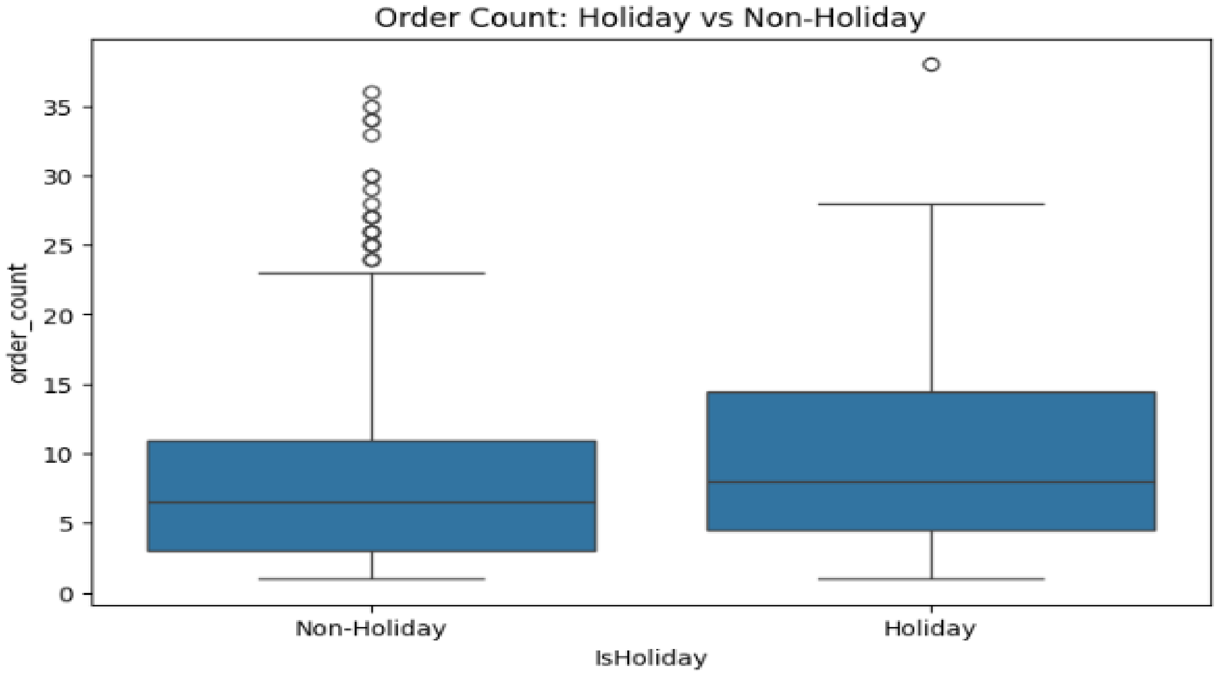

Figure 7.2 –Volume of Order Sales on Holidays vs Non-Holidays

Box Plot (figure 7.2) is illustrated to analyze whether order volumes differ between holiday and non-holiday days, average daily orders were computed for each group. The results were visualized using boxplots, showing higher median and overall order counts on holidays, demonstrating that **order volumes are higher on holidays**.

## RQ3 —Which US holidays have the greatest impact on customer ordering patterns?

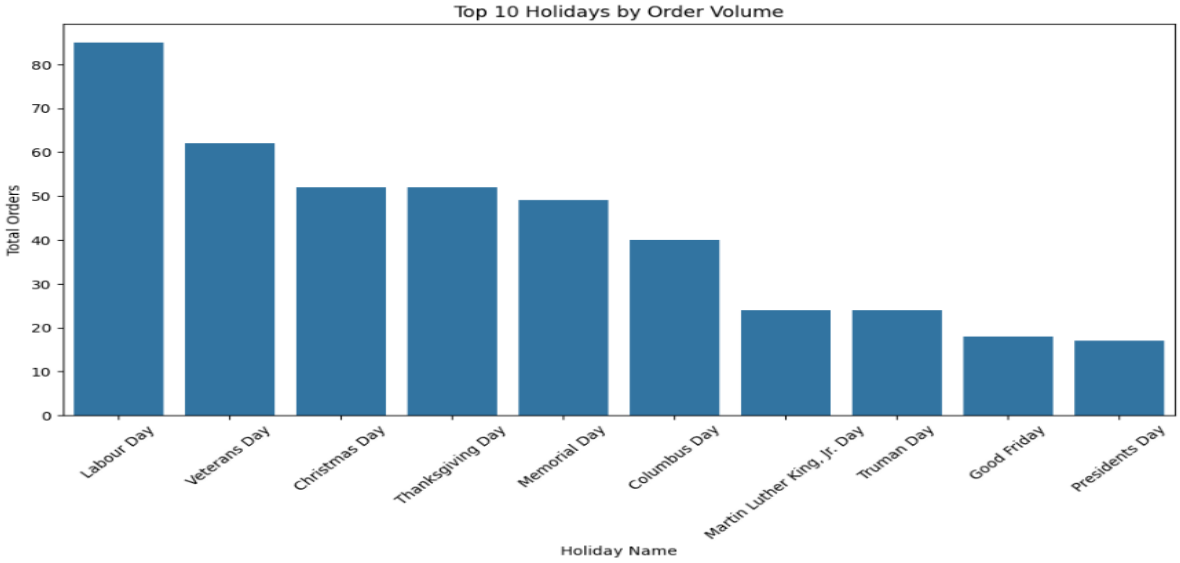

Figure 7.3 – Top 10 Holidays affecting Daily Sales

Figure 7.3 demonstrates this research question through bar chart to identify the holidays with the greatest impact on demand. Orders were grouped by holiday name and total orders were computed for each holiday.

Bar charts revealed that major holidays such as **Labour Day, Veterans Day, Christmas Day, and Thanksgiving Day** generate the highest order volumes, demonstrating that **certain holidays significantly boost customer activity**.

## RQ4 — Can incorporating holiday information improve the prediction of daily order volumes?

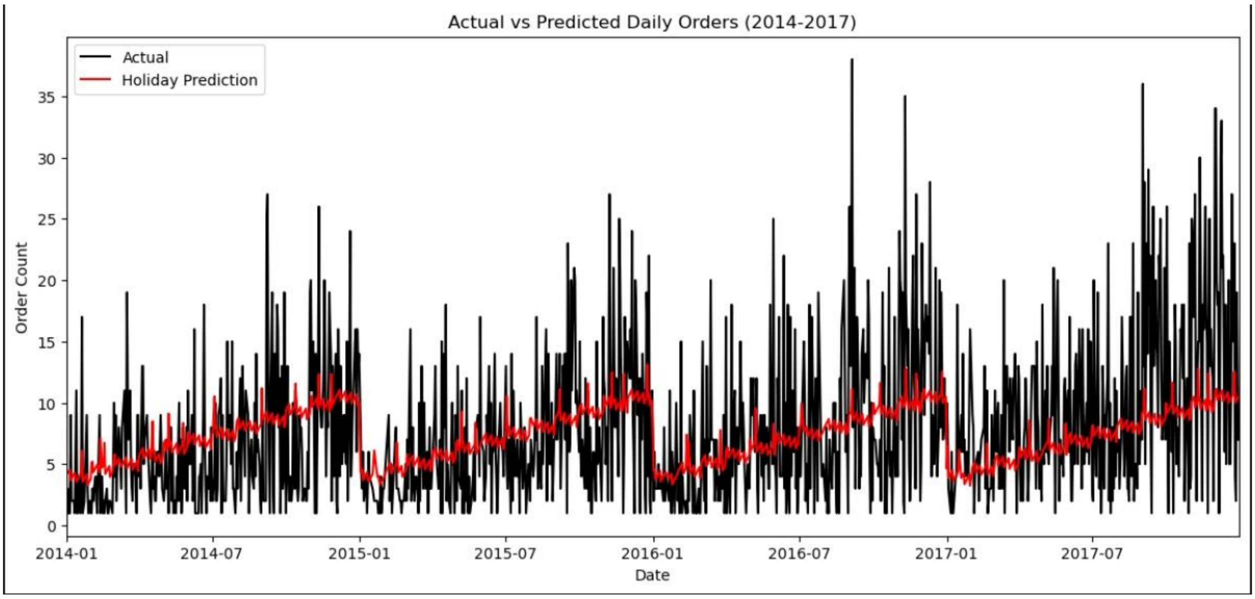

Figure 7.4 – Actual vs Predicted Daily Order

Figure 7.4 evaluates whether holiday information improves demand forecasting, daily orders were modeled using time-based features with and without a holiday indicator. Line plots comparing actual and predicted values showed that the holiday-enhanced model more closely follows real demand patterns, confirming that **including holiday information improves prediction accuracy**.

# 8 Conclusions

We analysed 9,994 retail orders from 2014 to 2017 alongside 60 US public holiday records to study how public holidays affect daily order volumes. In short, holidays have a measurable but variable impact on customer ordering behavior: some holidays coincide with spikes in orders, while others show minimal effect. Data cleaning and standardization ensured accurate comparisons, with all date fields aligned and duplicates removed, making the datasets reliable for integration and future analyses.

## 8.1 Key Findings

- Public holidays significantly influence daily customer ordering behavior.
- Order volumes are generally higher on holidays than on non-holidays.
- Certain holidays drive substantially higher customer activity.
- Including holiday indicators improves forecasting accuracy.

References

1. Kaggle. (2017). *US Retail Store Sales (2014–2017)*.
   Available at: https://www.kaggle.com/datasets/v1shalsb/us-retail-store-sales-2014-2017-yearly-split

2. Nager.Date API. (2014–2017). *Public Holidays in the United States*.
   Available at:
   Public Holidays in United States 2014 - Nager.Date
   Public Holidays in United States 2015 - Nager.Date

   Public Holidays in United States 2016 - Nager.Date
   Public Holidays in United States 2017 - Nager.Date