BUSINESS CHALLENGE: DATA CLEANING AND EDA

Presented by: Natalia Domínguez Jiménez

OVERVIEW

01
Problem Introduction

02
Dataset Overview

03
Project Focus

04
Data Cleaning

05
EDA (Exploratory Data Analysis)

06
Conclusions

01

PROBLEM INTRODUCTION

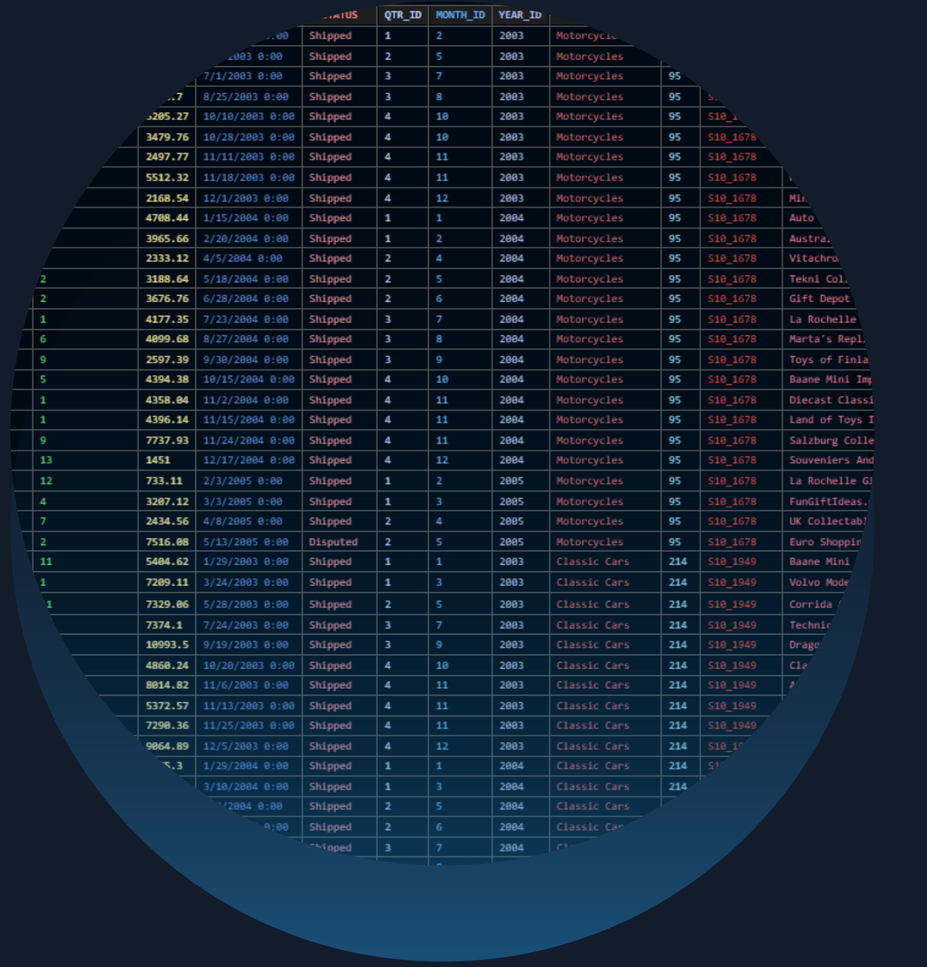
Sales performance is a critical metric for any business. Understanding the factors that influence it enables informed decision-making, strategy optimization, and ultimately, business growth.

This project aims to identify the key drivers of sales performance by analyzing a comprehensive sales dataset.

02 DATASET OVERVIEW

The dataset contains over 2,800 sales records documenting various transactions of classic and specialty vehicles.

It includes information such as product quantity and price, total sales amount, and the date of each transaction. Additionally, it provides details about customers and their geographical locations.



			QTR_ID	MONTH_ID	YEAR_ID			
		Shipped	1	2	2003	Motorcycles		
	2003 0:00	Shipped	2	5	2003	Motorcycles		
	7/1/2003 0:00	Shipped	3	7	2003	Motorcycles	95	
.7	8/25/2003 0:00	Shipped	3	8	2003	Motorcycles	95	
3285.27	10/10/2003 0:00	Shipped	4	10	2003	Motorcycles	95	S10_
3479.76	10/28/2003 0:00	Shipped	4	10	2003	Motorcycles	95	S10_1678
2497.77	11/11/2003 0:00	Shipped	4	11	2003	Motorcycles	95	S10_1678
5512.32	11/18/2003 0:00	Shipped	4	11	2003	Motorcycles	95	S10_1678
2168.54	12/1/2003 0:00	Shipped	4	12	2003	Motorcycles	95	S10_1678
4708.44	1/15/2004 0:00	Shipped	1	1	2004	Motorcycles	95	S10_1678
3965.66	2/20/2004 0:00	Shipped	1	2	2004	Motorcycles	95	S10_1678
2333.12	4/5/2004 0:00	Shipped	2	4	2004	Motorcycles	95	S10_1678
2	3188.64	5/18/2004 0:00	Shipped	2	5	2004	Motorcycles	95
2	3676.76	6/28/2004 0:00	Shipped	2	6	2004	Motorcycles	95
1	4177.35	7/23/2004 0:00	Shipped	3	7	2004	Motorcycles	95
6	4099.68	8/27/2004 0:00	Shipped	3	8	2004	Motorcycles	95
9	2597.39	9/30/2004 0:00	Shipped	3	9	2004	Motorcycles	95
5	4394.38	10/15/2004 0:00	Shipped	4	10	2004	Motorcycles	95
1	4358.04	11/2/2004 0:00	Shipped	4	11	2004	Motorcycles	95
1	4396.14	11/15/2004 0:00	Shipped	4	11	2004	Motorcycles	95
9	7737.93	11/24/2004 0:00	Shipped	4	11	2004	Motorcycles	95
13	1451	12/17/2004 0:00	Shipped	4	12	2004	Motorcycles	95
12	733.11	2/3/2005 0:00	Shipped	1	2	2005	Motorcycles	95
4	3207.12	3/3/2005 0:00	Shipped	1	3	2005	Motorcycles	95
7	2434.56	4/8/2005 0:00	Shipped	2	4	2005	Motorcycles	95
2	7516.08	5/13/2005 0:00	Disputed	2	5	2005	Motorcycles	95
11	5484.62	1/29/2003 0:00	Shipped	1	1	2003	Classic Cars	214
1	7209.11	3/24/2003 0:00	Shipped	1	3	2003	Classic Cars	214
1	7329.06	5/28/2003 0:00	Shipped	2	5	2003	Classic Cars	214
7374.1	7/24/2003 0:00	Shipped	3	7	2003	Classic Cars	214	S10_1949
10993.5	9/19/2003 0:00	Shipped	3	9	2003	Classic Cars	214	S10_1949
4860.24	10/20/2003 0:00	Shipped	4	10	2003	Classic Cars	214	S10_1949
8014.82	11/6/2003 0:00	Shipped	4	11	2003	Classic Cars	214	S10_1949
5372.57	11/13/2003 0:00	Shipped	4	11	2003	Classic Cars	214	S10_1949
7290.36	11/25/2003 0:00	Shipped	4	11	2003	Classic Cars	214	S10_1949
9064.89	12/5/2003 0:00	Shipped	4	12	2003	Classic Cars	214	S10_1
.3	1/29/2004 0:00	Shipped	1	1	2004	Classic Cars	214	S1
	3/10/2004 0:00	Shipped	1	3	2004	Classic Cars	214	
	/2004 0:00	Shipped	2	5	2004	Classic Cars		
	0:00	Shipped	2	6	2004	Classic Cars		
		Shipped	3	7	2004	Cl		

03

PROJECT FOCUS

What impact sales?

04

DATA CLEANING AND INITIAL ANALYSIS

📊 Dataset Observations

During the data cleaning and exploration phase, we identified several key points that impact the analysis and structure of the dataset:

- No Unique Key: The dataset lacks a primary key to uniquely identify each row.
- Order-to-Product Relationship: A single ORDERNUMBER can be associated with multiple PRODUCTCODE entries. This indicates that each row represents a line item within an order, not a complete order itself.
- Null Values: Several columns, such as ADDRESSLINE2, STATE, POSTALCODE, and TERRITORY, contain a significant number of null values.
- Sales Column Discrepancy: Initially, the meaning of the SALES column was not immediately clear, but it was determined to represent the multiplication of QUANTITYORDERED * PRICEEACH. Upon validation, it was discovered that some rows contained an incorrect value of 100, which was an error in data entry.

04

DATA CLEANING AND INITIAL ANALYSIS

📊 Creating a Subset Table

After performing this preliminary analysis, I decided to create a subset table to focus on the study of three primary factors: geographical, temporal, and product line.

The table was created by selecting the following columns from the original dataset:

	ORDERNUMBER	PRODUCTCODE	QUANTITYORDERED	SALES	CITY	COUNTRY	MONTH_ID	QTR_ID	PRODUCTLINE
0	10107	S10_1678	30	2871.00	NYC	USA	2	1	Motorcycles
1	10121	S10_1678	34	2765.90	Reims	France	5	2	Motorcycles
2	10134	S10_1678	41	3884.34	Paris	France	7	3	Motorcycles
3	10145	S10_1678	45	3746.70	Pasadena	USA	8	3	Motorcycles
4	10159	S10_1678	49	5205.27	San Francisco	USA	10	4	Motorcycles
...	...	...	...	...	...	...	...	...	...
2818	10350	S72_3212	20	2244.40	Madrid	Spain	12	4	Ships
2819	10373	S72_3212	29	3978.51	Oulu	Finland	1	1	Ships
2820	10386	S72_3212	43	5417.57	Madrid	Spain	3	1	Ships
2821	10397	S72_3212	34	2116.16	Toulouse	France	3	1	Ships
2822	10414	S72_3212	47	3079.44	Boston	USA	5	2	Ships

2823 rows × 9 columns

05

EDA (EXPLORATORY DATA ANALYSIS)

🔍 Exploratory Data Analysis (EDA)

Following the data cleaning phase, I conducted an exploratory data analysis to better understand the key business characteristics and the relationships between variables. I based my analysis on the three factors I identified and followed these steps:

- **Variable Types Check**: I first examined the data types of each column to ensure they were appropriate for my analysis. In this case, no changes were necessary.
- **Univariate EDA**: I analyzed individual variables to understand their distributions, central tendencies, and unique values (e.g., looking at sales distribution or the count of each product line).
- **Bivariate EDA**: Finally, I explored the relationships between pairs of variables (e.g., how sales relate to the month or city).

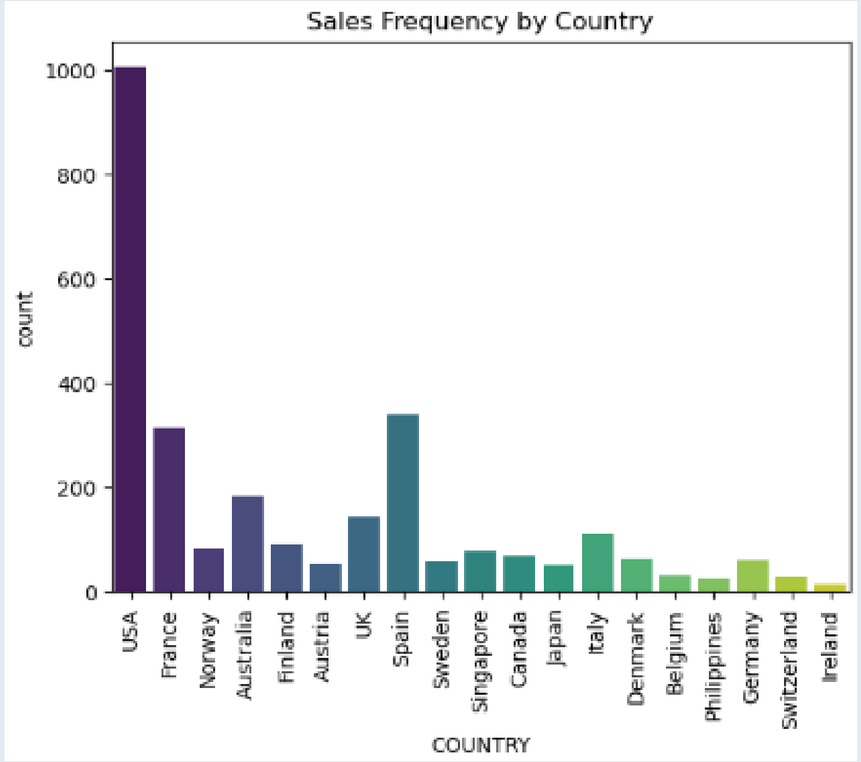
Here are the some os the findings from my EDA:

05

EDA (EXPLORATORY DATA ANALYSIS)

🗺️ Geographical Factors

- Sales by Country: I identified the United States (USA) and Spain (Spain) as the countries with the highest sales volume.

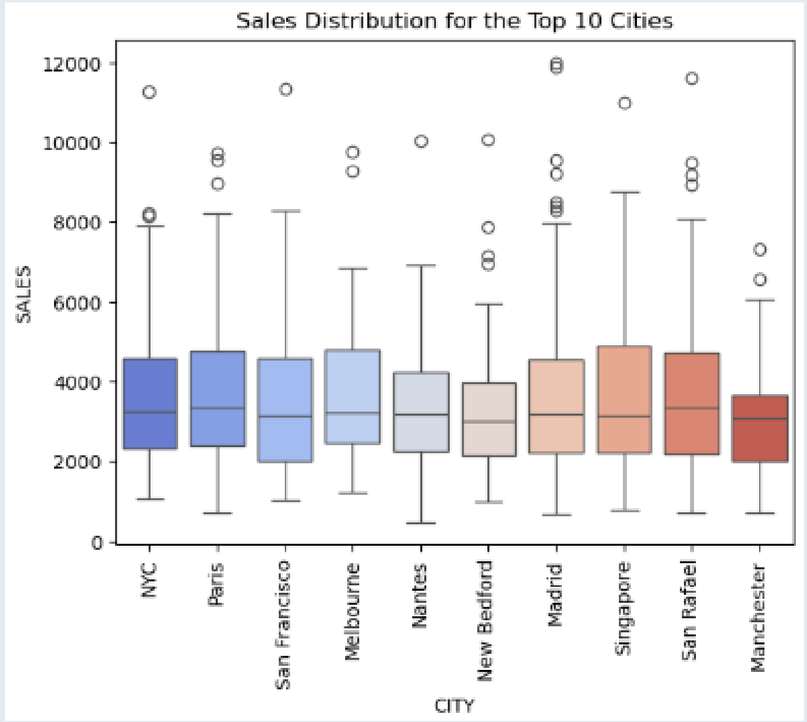


05

EDA (EXPLORATORY DATA ANALYSIS)

🗺️ Geographical Factors

- Geographical Distribution: The boxplot shows that cities like NYC, Paris, and San Francisco generally have higher median sales and a wider distribution of high-value transactions. In contrast, cities such as San Rafael and Manchester exhibit a more concentrated distribution of lower sales values, with fewer high-value outliers.

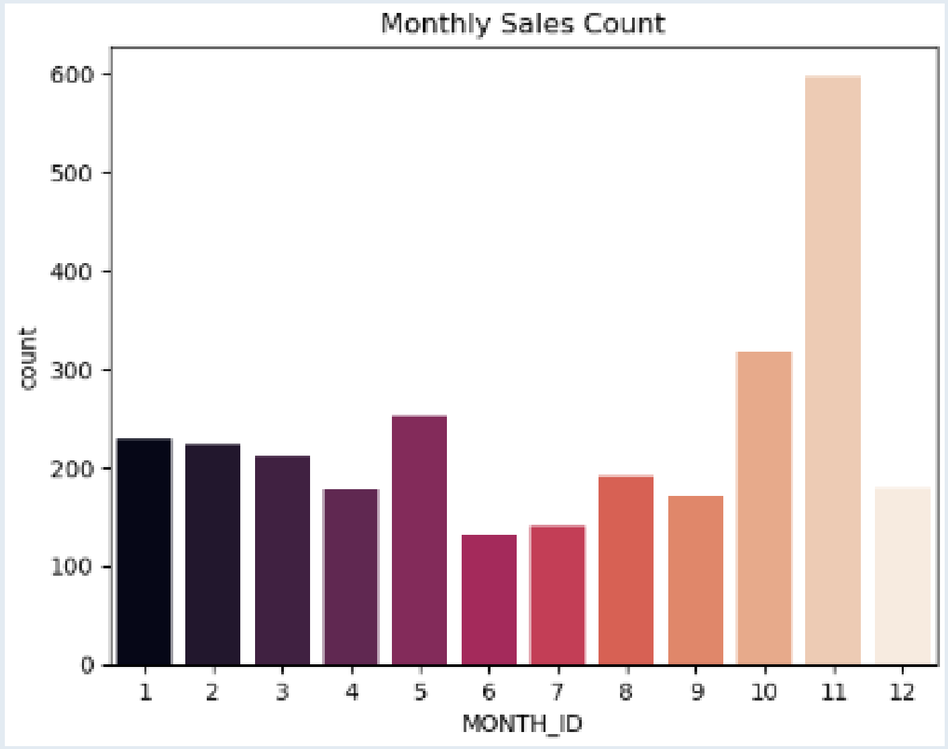


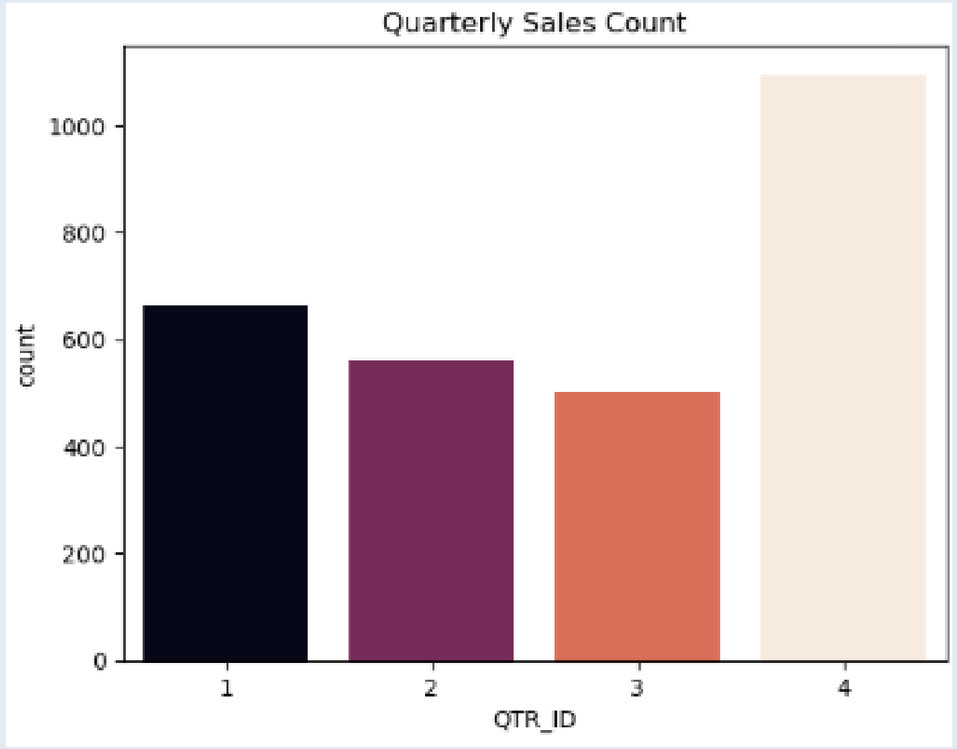
05

EDA (EXPLORATORY DATA ANALYSIS)

📈 Temporal Factors

- Annual Cyclical Pattern: I observed a clear cyclical pattern in sales. The fourth quarter (QTR_ID 4), particularly the month of November, shows a significant sales spike, indicating strong seasonal behavior.



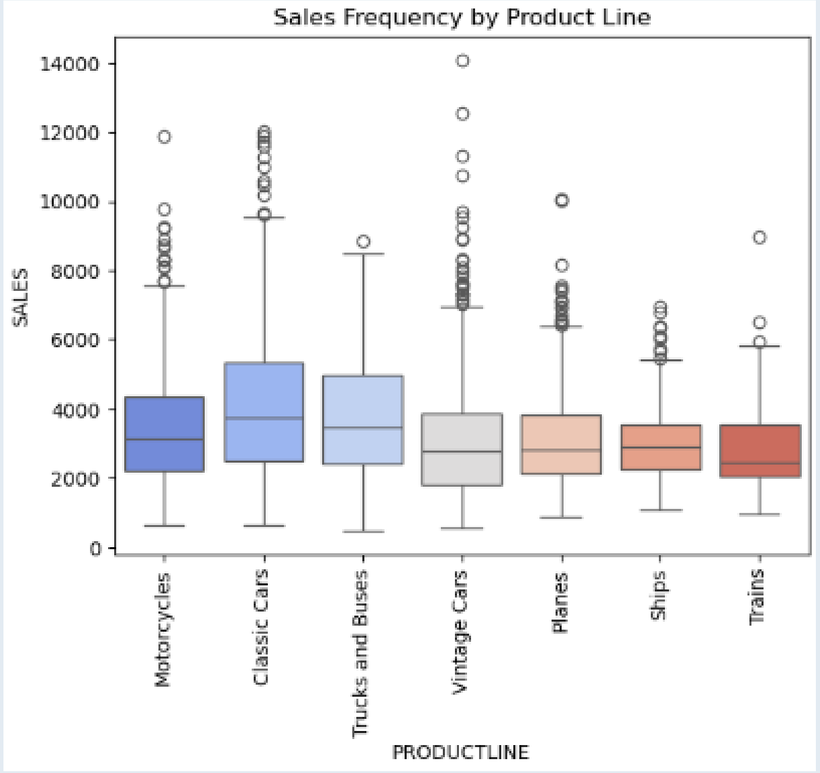


05

EDA (EXPLORATORY DATA ANALYSIS)

Product Line Factors

- Top-Selling Products: My analysis of average sales by product line reveals a clear hierarchy. Classic cars have the highest average amount of sales, whereas trains have the least average amount of sales.



06

CONCLUSIONS

This project gave me a comprehensive understanding of the sales dataset. The main challenge I faced was during the initial analysis, as the dataset lacked clear definitions and I had to determine the meaning of the SALES column.

This project gave me useful insights into sales behavior, where customers are located, and how products are performing. These results will help in creating future predictive models and making better strategic decisions.

For the complete code and a detailed breakdown of each step, please refer to my notebook.

THANK YOU