Sprints

E-Commerce Sales Data Analysis Report

Table of Contents

1. Introduction

Overview

This project focuses on analyzing an e-commerce sales dataset to extract meaningful insights, optimize business strategies, and improve customer engagement. The analysis will cover data cleaning, customer segmentation, time-series forecasting, and performance visualization.

Objective

- Clean and preprocess raw data for accurate analysis.
- Perform descriptive and exploratory data analysis (EDA) to uncover trends.
- Segment customers using RFM analysis to identify key customer groups.
- Analyze sales trends across different time periods.
- Forecast future sales using moving average techniques.
- Present insights through interactive dashboards and visualizations.

Dataset Description

The dataset includes transactional data from an online retail store. Key columns include:

Column Name	Description	Type
InvoiceNo	Unique invoice number for each transaction.	Text
StackCode	Product identifier.	Text
Description	Product description.	Text
Quantity	Number of units purchased.	Real Number
InvoiceDate	Date and time of purchase.	Date
UnitPrice	Price per unit of product.	Real Number
CustomerID	Unique customer identifier.	Real Number
Country	Country where the purchase was made	Categorical

Dataset statistics

- Number of variables: 8
- Number of observations: 541,909
- Missing cells: 136,534 (3.1%)
- Duplicate Rows: 4879 (0.9%)

Variable Types:

- Text 3
- Numeric 3
- DateTime 1
- Categorical 1

Variable Details

- **InvoiceNo:**
  - Distinct 25,900 (4.8%)
  - Missing 0
- StockCode:
  - Distinct 4,070 (0.8%)
  - Missing 0

- Description
    - Distinct 4223 (0.8%)
    - Missing 1454 (0.3%)
- Quantity

Distinct	722 (0.1%)	Negative	10624 (2%)
Missing	0	Zeros	0
Maximum	-80995	Minimum	-80995

- InvoiceDate

Distinct	23260 (4.3%)	Negative	10624 (2%)
Missing	0	Missing (%)	0
Maximum	2011-12-09 12:50:00	Minimum	2010-12-01 08:26:00

- UnitPrice

Distinct	1630 (0.3%)	Negative	2 (<0.2%)
Missing	0	Zeros	2515 (0.5%)
Maximum	38970	Minimum	-11062.06

- CustomerID

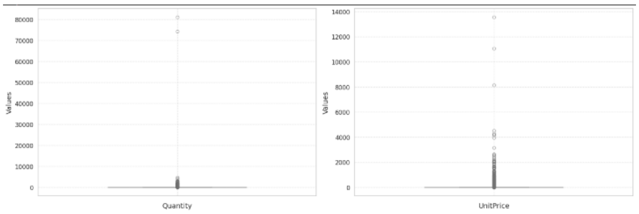
Distinct	4372 (1.1%)	Negative	0
Missing	135080 (24.9%)	Zeros	0
Maximum	18287	Minimum	12346

- Country
    - Distinct 38
    - Missing 0

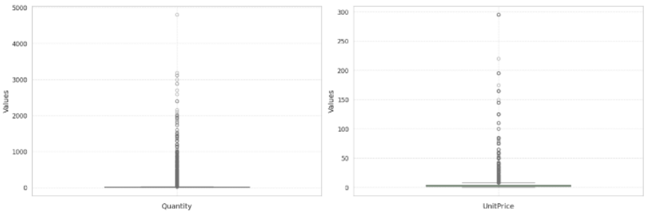
- **Data Quality Alerts**
    - Description : has 299 missing records (0.3%)
    - CustomerID : has 34,935 missing records (34.8%)
    - UnitPrice: has 1 missing records (<0.1%)
    - Country : has 1 missing records (<0.1%)
    - Country : is highly inbalanced with 92,591 (86.2%) of the entires concentrated in one country.
    - UnitPrice : is highly skewed (γ1 = 113.1442141)
    - Dataset has 849 (0.8%) duplicate rows
- **Insights from the dataset**
    - **Customer Segmentation**
        - A very large number of customers do not have a recorded CustomerID
        - The majority of sales seems to be concentrated in a few countries, indicating potential market dependency
    - **Product Pricing**
        - UnitPrice` shows significant skewness, suggesting outliers such as abnormally high or low prices that could affect sales trends
    - **Data Duplication**
        - Duplicate rows indicate possible redundancy or errors in each record, which may distort revenue and sales analysis

2. Task 0: Early Discovering and Preprocessing

- Description
  - Clean and preprocess the dataset by handling missing values, formatting data types, and removing incorrect records.
- Requirements
  - Handle missing values in CustomerID
  - Convert InvoiceDate` to DateTime type
  - Remove rows with negative `Quantity` or UnitPrice
  - Create a TotalPrice column
- Data Quality Assessment and Cleaning:
  - Duplicates
    - Identify duplicate rows
    - Drop duplicates
  - Incorrect Data
    - Identify invalid entires (e.g. negative values, impossible values) and remove them
  - Outliers
    - Detect Outliers using IQR
    - Handle Outliers
      - Cap or floor extreme values
      - Transform Data (log scaling)
      - Remove outliers if they are errors
  - Missing Values
    - Check Null values in the dataset
    - Handle missing values
      - Fill (e.g. mean, median, mode or interpolation)
      - Drop rows
      - Use prediction/Imputation model for filling
- Observations
  - All the negative records have 'C' in the begining of the InvoiceNO, so we can safely assume all the rows with negative quantities are a cancelled orders and drop them
- Dealing with Outliers



Before



After

3. Task 2: Exploratory Data Analysis (EDA)

- Description

Perform basic descriptive statistics and identify insights from the dataset. Analyze top-selling products and calculate total revenue and transactions.values. You will also check the data types of each column to understand how the data is structured.

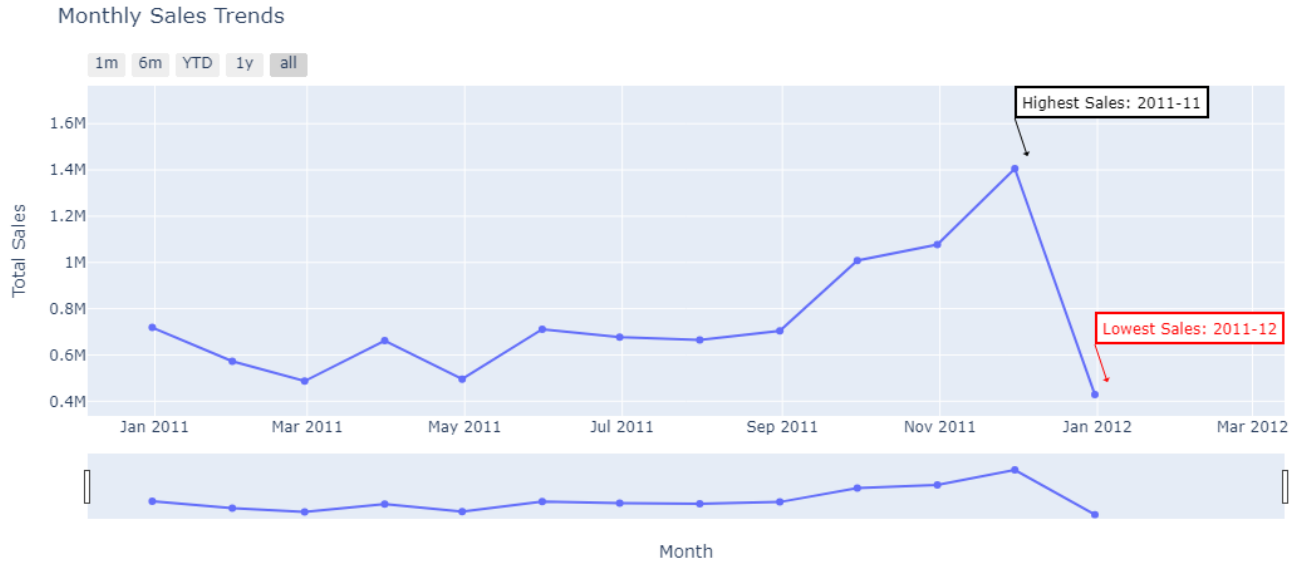
- Top 10 selling products

Description	Quantity
WORLD WAR 2 GLIDERS ASSTD DESIGNS	54,951
JUMBO BAG RED RETROSPOT	48371
WORLD WAR 2 GLIDERS ASSTD DESIGNS	37872
POPCORN HOLDER	36749
PACK OF 72 RETROSPOT CAKE CASES	36396
ASSORTED COLOUR BIRD ORNAMENT	36362
RABBIT NIGHT LIGHT	30739
MINI PAINT SET VINTAGE	26633
PACK OF 12 LONDON TISSUES	26119
PACK OF 60 PINK PAISLEY CAKE CASES	24820

- Total Revenue and number of transactions

Total Revenue	$9613760.14
number of transactions	19953

4. Task 3: Time Series Analysis



- Monthly Sales Trend Analysis: Observations and Insights

1- Overall Trend

Sales fluctuated throughout the year, with distinct peaks and troughs

2- Lowest Sales Point

The lowest sales were recorded at 2011-12-31 with Sales $0.4M

3- Highest Sales Point

The highest sales were recorded at 2011-11-30 with Sales $1.4

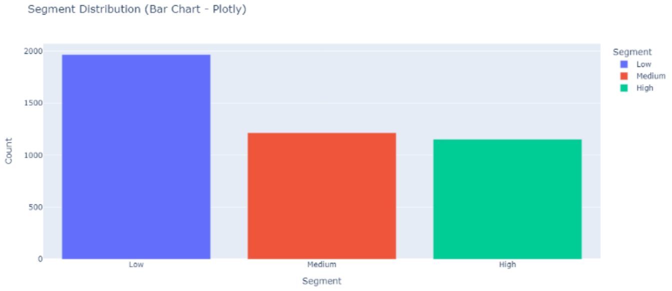
4- Stable Period

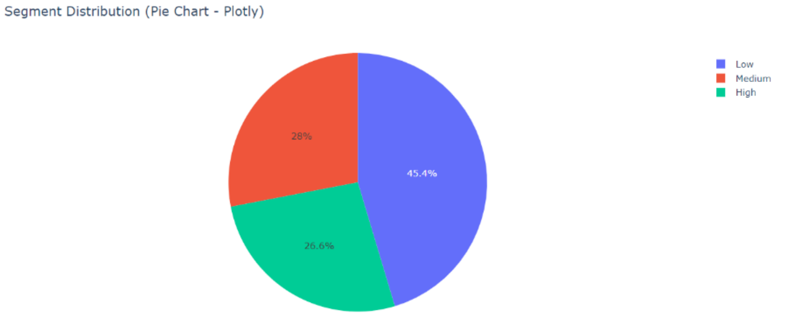
From **June 2011 to Augest 2011**, sales remained relatively stable, hovering around the **700K** to **800K** range, without significant upward or downward spikes

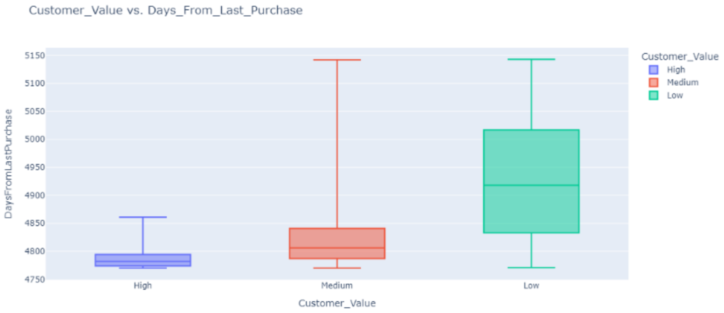
5- Significant Decline Post-peak

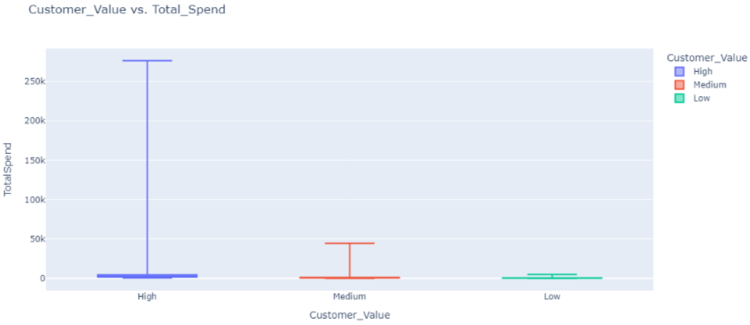
Following the peak in `November 2011`, sales dropped significantly in `January 2012`, signaling a sharp post-seasonal decline

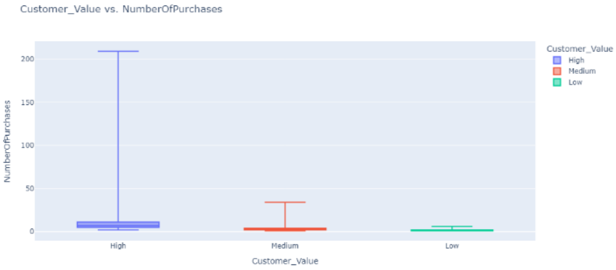
5. Task 4: RFM Analysis (Customer Segmentation)

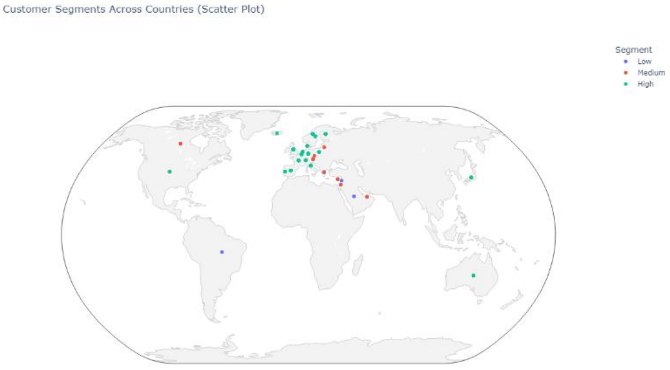


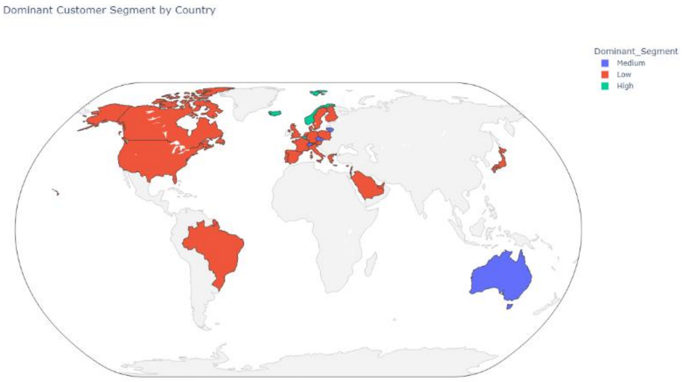






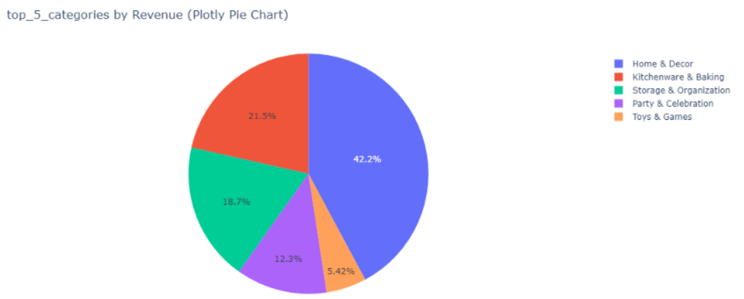


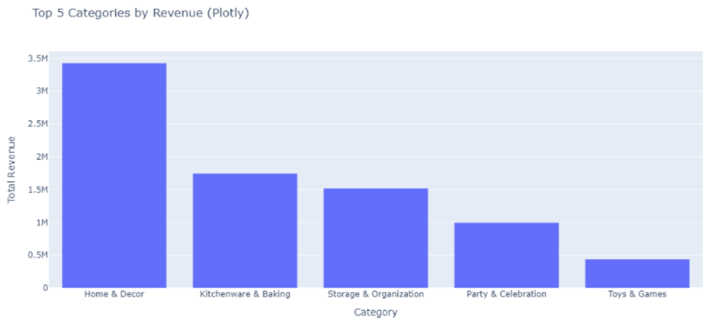




6. Task 5: Product Category Analysis

- We divided our products into 13 different categories
  - Home & Decore
  - Craft & Stationery
  - Hot Water Bottles
  - Travel & Accessories
  - Toys & Games
  - Storage & Organization
  - Christmas & Seasonal
  - Gardening & Outdoor
  - Mugs & Drinkware
  - Uncategorized

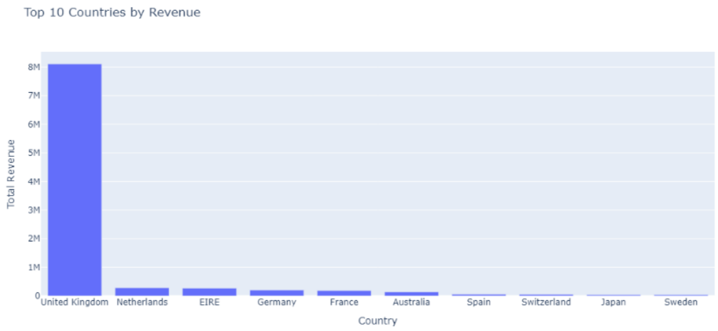




- Top 5 Categories

No	Category	Total Revenue
1	Home & Decore	3.56M
2	Kitchenware & Baking	1.9M
3	Storage & Organization	1.6M
4	Party & Celebration	1.25M
5	Travel & Accessories	0.5M

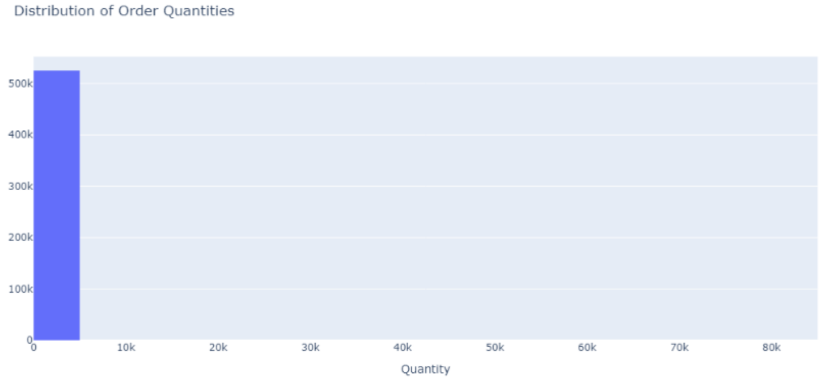
7. Task 6: Geographical Analysis



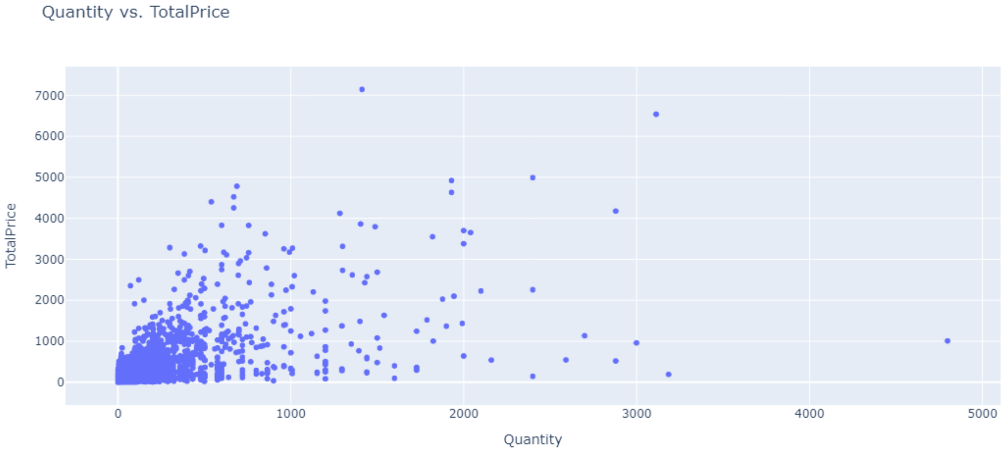
- Observations
  - **Country:** is highly unbalanced with 92,591 (86.2%) of the entries concentrated in one country which is **United Kingdom**
  - Top 3 **Countries** are United Kingdom, Netherlands, EIRE
  - Percentage **Sales** from top 3 **countries** is 89.94%

8. Task 7: Customer Behavior Analysis

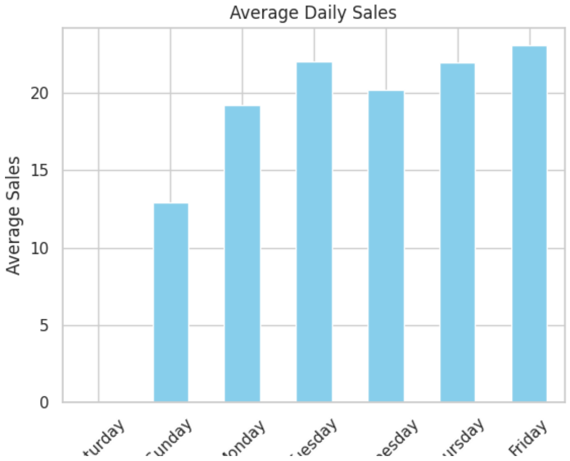
- Distribution of order quantities



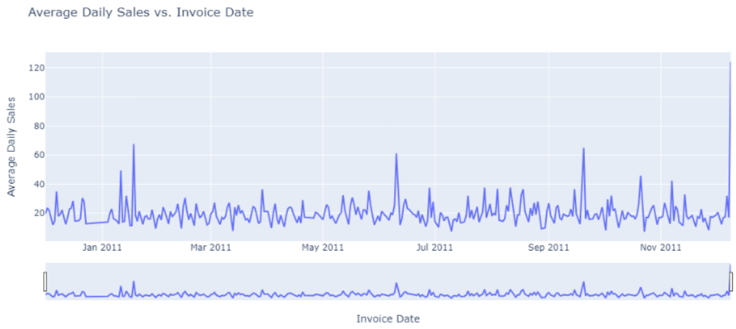
- Scatter plot of quantities vs total price



- Average Daily Sales



- Average Daily Sales vs Invoice Date



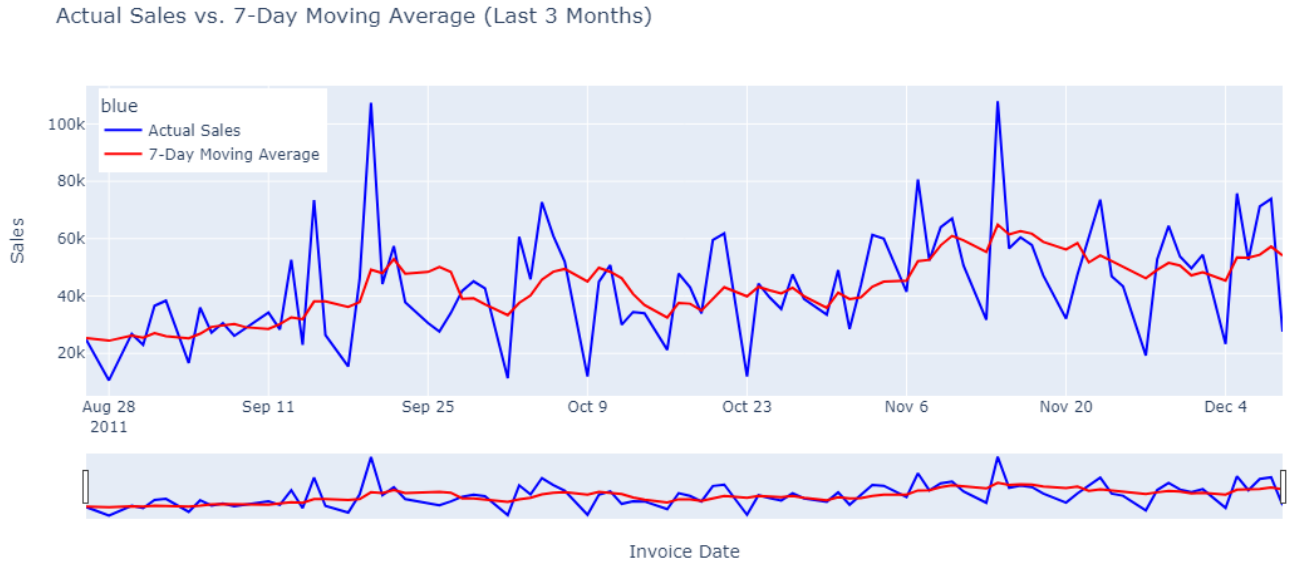
9. Task 8: Moving Average Forecast

- Approach Taken:

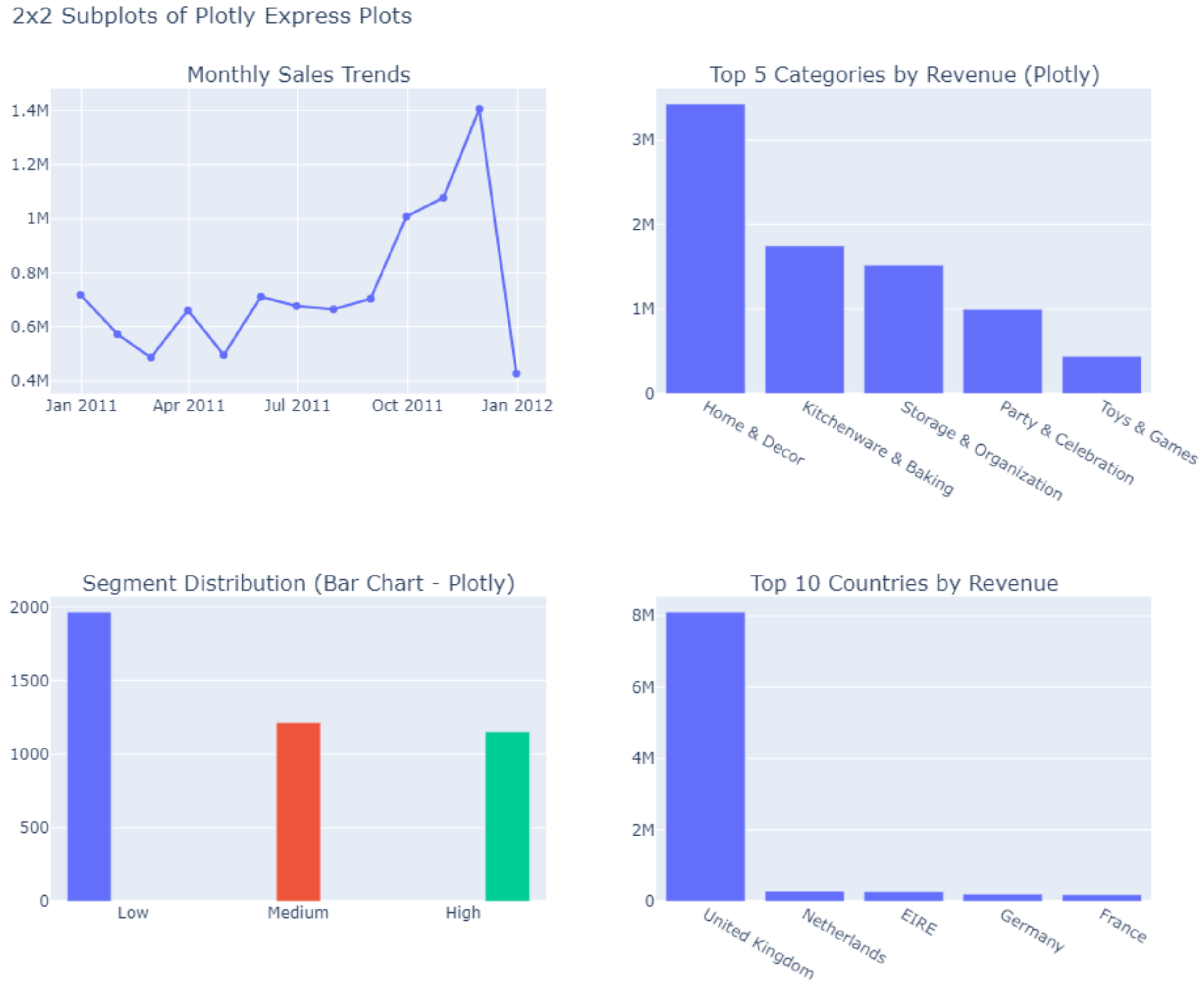
  Step 1: We grouped out dataset by **InvoiceNo.dt.date** and then we took the sum of the **TotalPrice** Column

  Step 2: We calculated a 7-day moving average of sales by using rolling(window=7).mean Methode

  Step 3: We plot the actual sales vs moving average of the last three month



10.Task 9: Summary Dashboard Creation



11.Task 10: Optimize Data Processing

- Comparing Time

The Summation of for loop operation is 0.106seconds

The Summation of vectorized operation is 0.0016 seconds

- Comparing Values

The sum of all the elements in Quantity Col using for loop is 5408986

The sum of all the elements in Quantity Col using vectorized operation is 5408986

- Summary

The vectorized operation is 67 times faster than the for-loop operation

12.Task 11: Report Generation

- Overall Revenue: $10M
- Total Transactions: 19960
- Top Category: Home & Decor
- Top-Selling Products: JUMBO BAG RED RETROSPOT
- Best Customer Segments:
- Top Selling Countries: United Kingdom
- Insights from time series analysis
  - **Overall Trend**
    Sales fluctuated throughout the year, with distinct peaks and troughs
  - **Lowest Sales Point**
    The lowest sales were recorded at 2011-12-31 with Sales $0.4M
  - **Highest Sales Point**
    The highest sales were recorded at 2011-11-30 with Sales $1.4
  - **Stable Period**
    From June 2011 to September 2011, sales remained relatively stable, hovering around the 700K to 800K range, without significant upward or downward spikes
  - **Significant Decline Post-peek**
    Following the peak in `November 2011`, sales dropped significantly in `January 2012`, signaling a sharp post-seasonal decline
- Recommendations
  - Deal with outliers
    - We should make further investigation to deal with outliers to know if they are correct data or wrong inputs
    - Take transformation for out data such as log transform or make standardization
    - Last option is to drop the outliers
  - Use NLP to detect Categories
    - NLP-based Category Extraction We then explored natural language processing (NLP) techniques to extract categories automatically from the `descriptions`. While NLP models showed promise, they required significant computational resources (e.g., GPU power) and substantial processing time. In one instance, the process was left running for over three hours without producing results, making this approach impractical for our needs at that stage.