Project presented by: Cynthia Abissegue.

Project Name: Analysis of Supermarket sales in the United States in 2019.

Year: 2023

Time Taken: 3days

TABLE OF CONTENT

1. INTRODUCTION

This report presents an in-detail analysis of Supermarket sales collected across varies states in the United States from January 2019 to December 2019. The goal of this analysis is to uncover patterns, trends, and insights related to sales, customer behaviour and popular products.

During the course of this report, we shall traverse various stages of the data analysis process such as data collection, data cleaning, data visualization and data comprehension. In doing so, we shall provide answers to important questions such as the state receiving the highest number of orders, the product highest in demand and so on. Additionally, we shall be providing answers to inconsistency in the data and providing solutions and recommendations.

2. Data Sources

The data used throughout the course of this analysis was obtained from a nationwide Supermarket chain using an ERP (Enterprise Resource planning) Software. The data was presented in an csv format with each file containing data from each month of the year 2019. That is to say our data base consisted of 12 csv data sets.

In our database, the only data type represented was object.

We therefore concatenated our 12 data sets into one big database since they all had the same originality (Sales).

3. Exploratory Data Analysis

Exploratory data analysis is the first step in our study, which is crucial for understanding the database structure, composition and spotting early patterns. Through data cleaning, descriptive statistics and data visualization, we provide the groundwork for our upcoming studies.

3.1 Data Overview and Cleaning

a. Duplicated rows

Duplicated rows may occur because of various reasons such as data entry errors, merging data, repetitive user actions.

We had a total of 1162 duplicated rows.

To handle duplicated rows, we dropped them entirely.

b. Handling missing values

Missing values are a common occurrence in real-world datasets and can arise due to various reasons such as data entry errors, system glitches, absence of information or even data manipulation from personnel.

Initially we had a total of 545 missing values across each column. That is to say the same number of missing columns on each column.

After removal of duplicated rows, we obtained just 1 missing or NaN value on each column. We observed that all the missing values were located on the same single row.

To handle the missing values, we dropped the entire row.

In conclusion, a lot of the missing values was due to the fact that a lot of rows were duplicated.

3.2 Descriptive Statistics

Now that we are done with the data cleaning, our data base is ready for analysis. This section contains a variety of analysis designed to glean insights and patterns within the supermarket sales.

a. **Identifying the month with highest amount of sales**

Step 1: Generating a month column.

Our database contains a column called “Order Date” which represents data presented in the format MM/DD/YY hr:mm.
We then decided to split this format and store the month for each row in a new column titled “Month”

Step 2: Calculating the total sale recorded for each product

Given that our database has data type object, we first had to convert the “Quantity Ordered” and “Price Each” columns to numeric data types. Subsequently, the “Total Sale” column was created in which we performed multiplication of the 2 columns mentioned above per row.

Step 3: Grouping the sum of Total Sales by month

Finally, by using the groupby() and sum() functions, we identified the month with the highest cumulative sales.

b. **Identifying the City with the most sales recorded**

Step 1: Generating a city column

Our database features a “Purchase Address” column which portrays address in the form “Address, City, State”. We then proceeded by splitting this and obtained a new City column for each row.

Step 2: Grouping the sum of Total Sales by City

Using the groupby() and sum() functions, we obtained an array of Total Sales grouped by City.

c. <u>Identifying the City where the most orders were placed</u>

Here we just had to calculate the sum of the "Quantity Ordered" column and group by "City" using the groupby() and sum() functions.

d. <u>Hour of the day during which the highest number of sales was recorded</u>

Step 1: Generating an hour column
Like we saw in (a), we had to split the data from the "Order Date" column to obtain a "Time" column which we further split to obtain an "Hour" column.

Step 2: Grouping the sum of Total Sales by Hour.
Using the sum() and groupby() functions, we were able to obtain an array containing the sum of Total of Sales grouped by Hour.

e. <u>Product which bought in the most revenue</u>

Using the groupby() and sum() functions we calculated the sum of Total Sales grouped by "Product" column.

f. <u>Quantity of each product sold</u>

This result was obtained by grouping the sum of the "Quantity Ordered" column by "Product" using groupby() and sum() functions.

g. <u>Mean unit price of each product across all the states in the US</u>

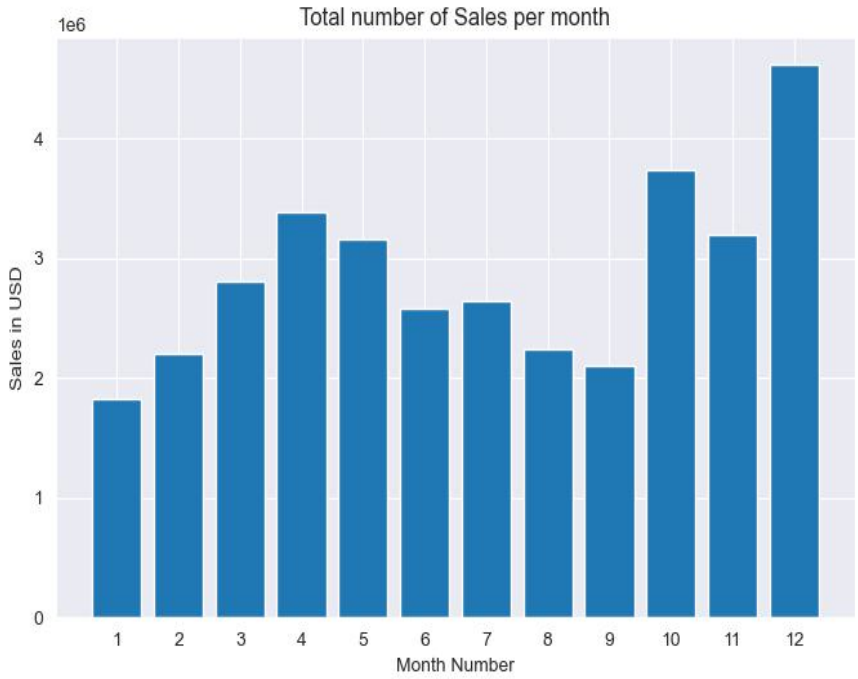
In this part of our data analysis, we grouped the mean of the column "Price Each" by Product which led to us obtaining the desired result.

3.3 Data Visualization

Building on the foundation created by data cleaning and analysis, we now dive into Data Visualization. Through data visualisation, we make our findings perceptible, bringing to light patterns and insights refined in earlier stages. These graphics simplify complex data and reveal trends that improve our understanding of the database.
Throughout the course of this stage, we shall be using bar charts due the comparative insights, categorical representation and visual impact they allow.

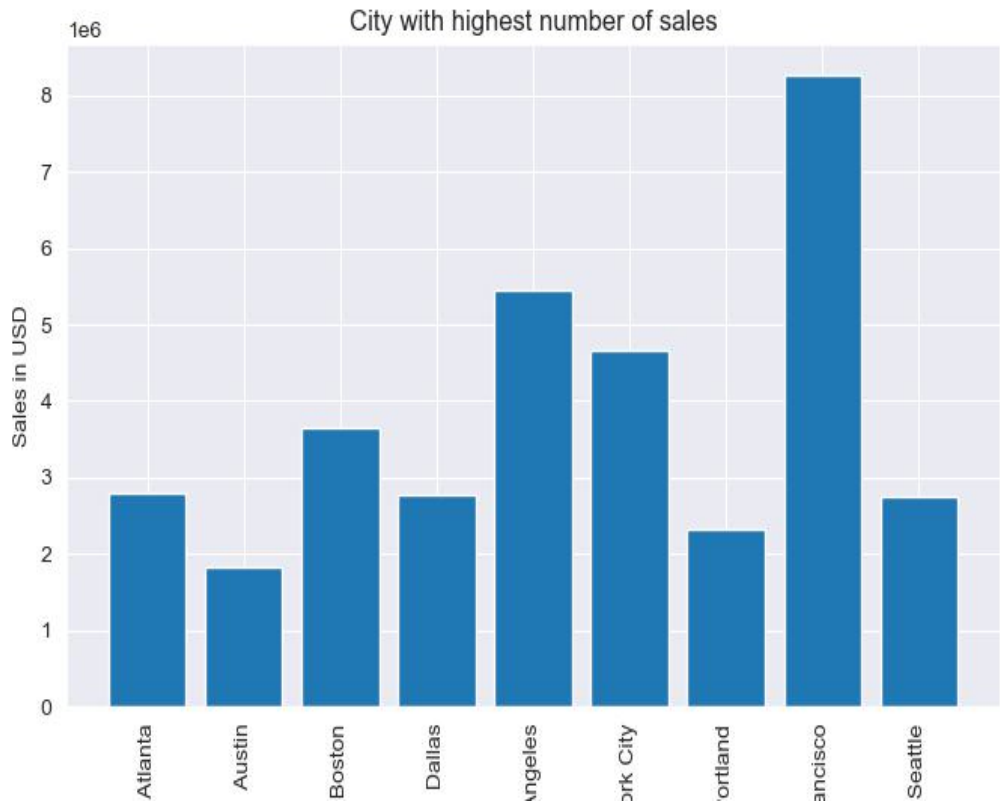
a. <u>Month with the highest number of sales throughout the year</u>

We have a graph of month plotted vs the sales in USD. Each bar corresponds to a month and its height depicts its total sale.



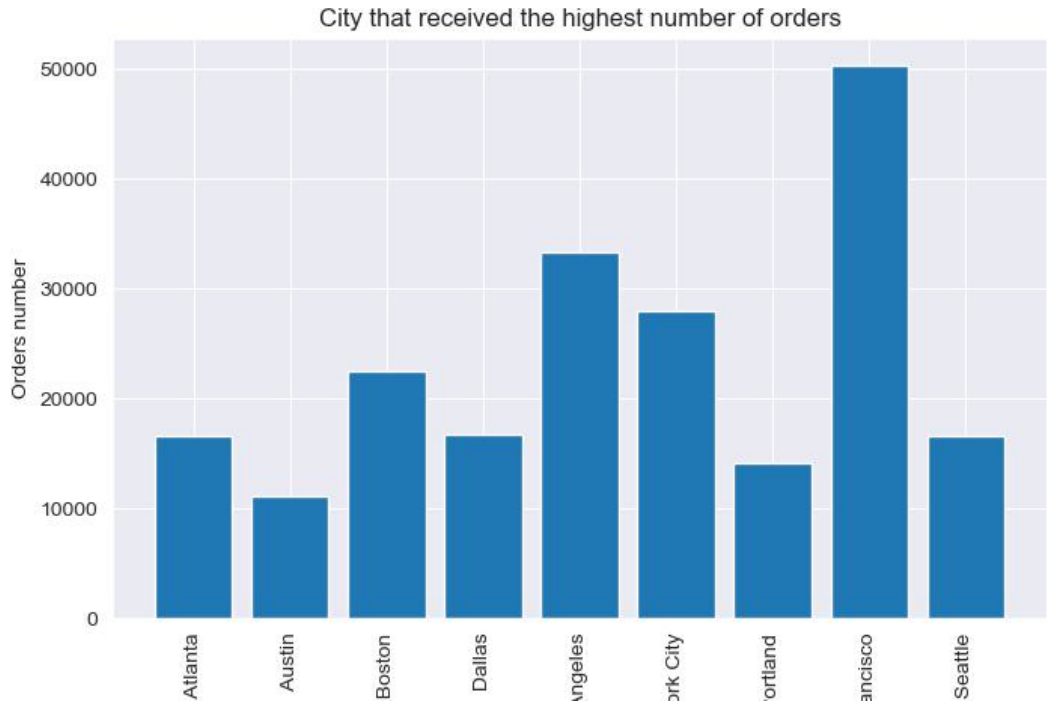
b. <u>City which recorded the most sales</u>

We generated a graph showing the correlation between City and total sales in USD with each bar corresponding to a specific City and its height illustrates the corresponding total sales.



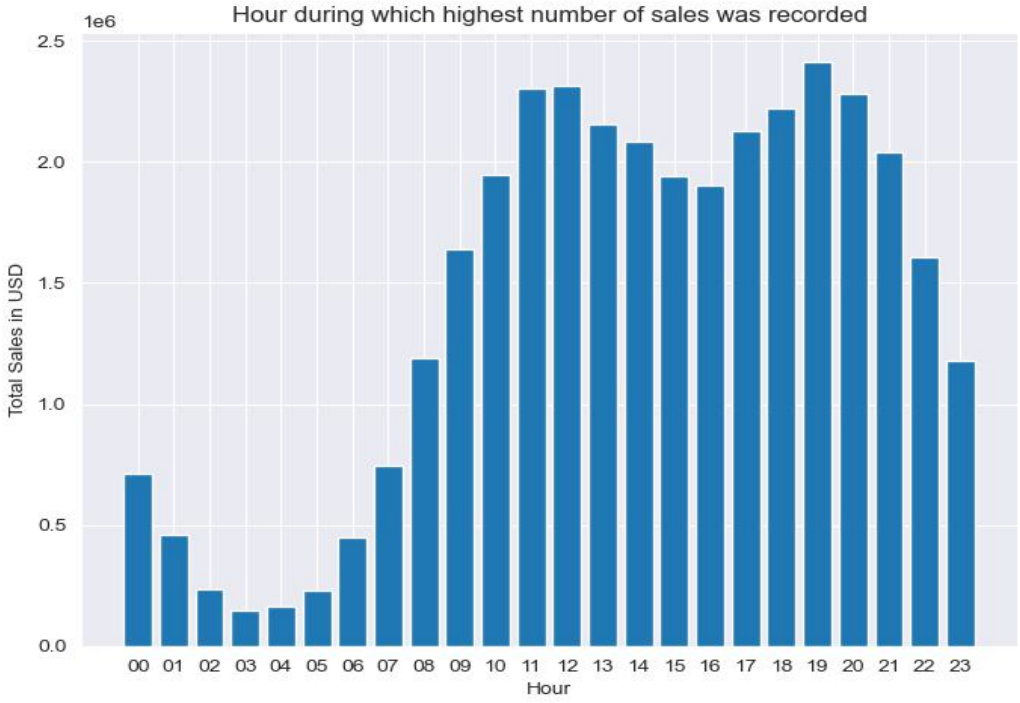
c. <u>City in which most orders were placed</u>

This illustration represents the City in terms of orders placed.
Vertical bars represent individual cities and their lengths represent orders per city**.**



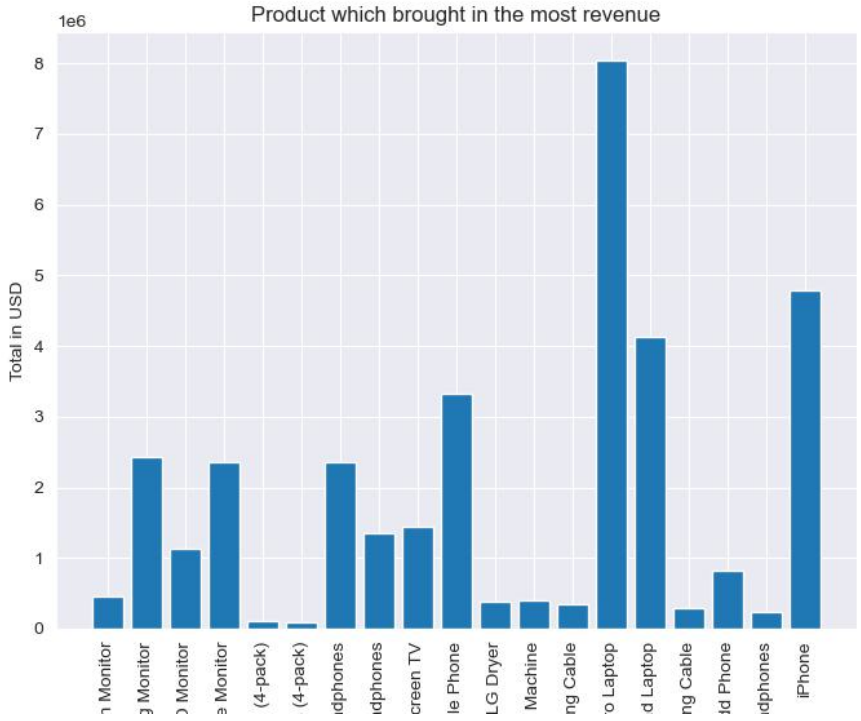
d. <u>Hour during which highest number of sales was recorded</u>

Here we present a graphical representation plotting hours vs total sales in USD.



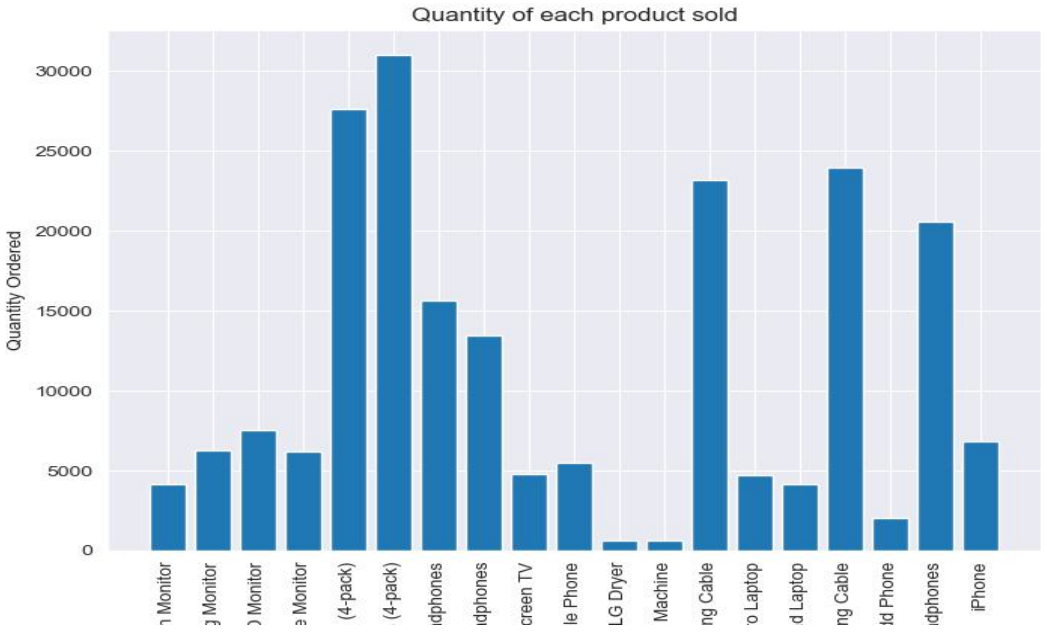
e. Product which brought in the most revenue

This graph displays the product names plotted against total sales in USD.



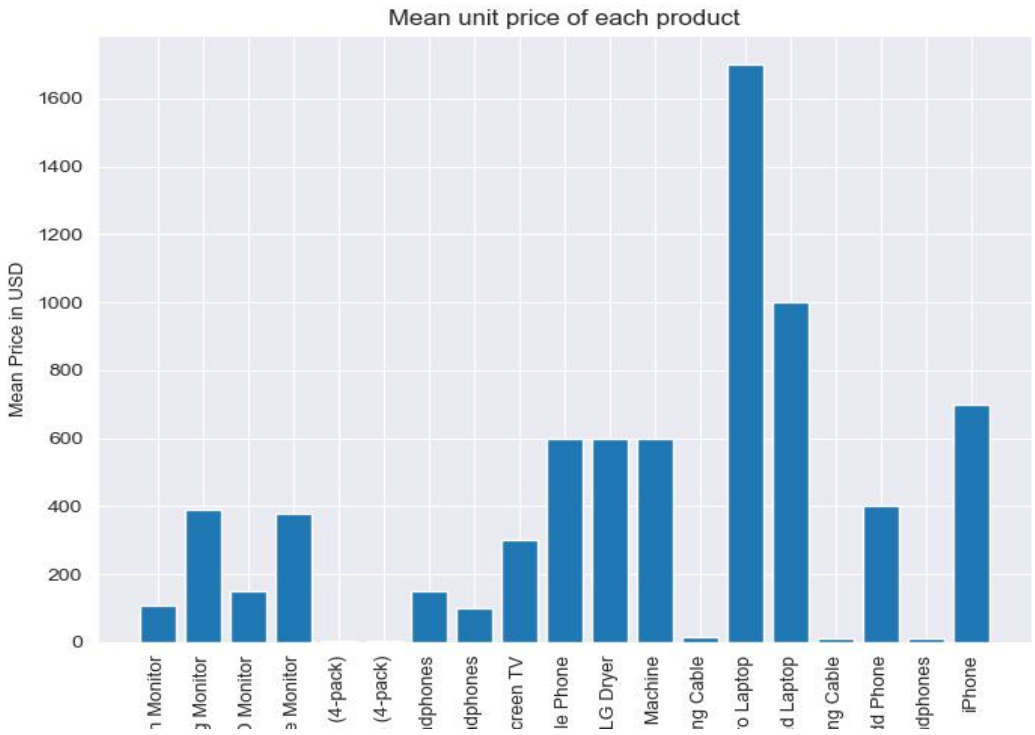
f. Quantity of each product sold

Here, we have an illustration of product names against quantity ordered.



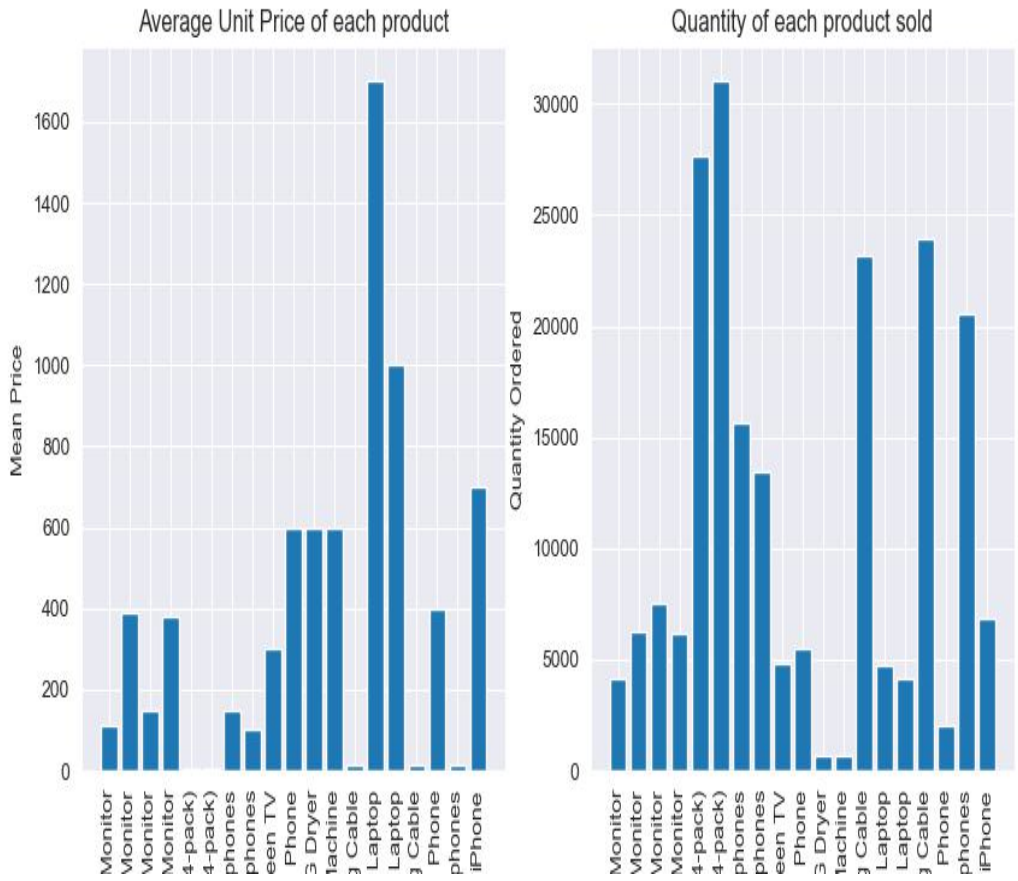
g. Mean unit price of each product across the country.

A visual depiction of the meant unit price of each product plotted against the product name.

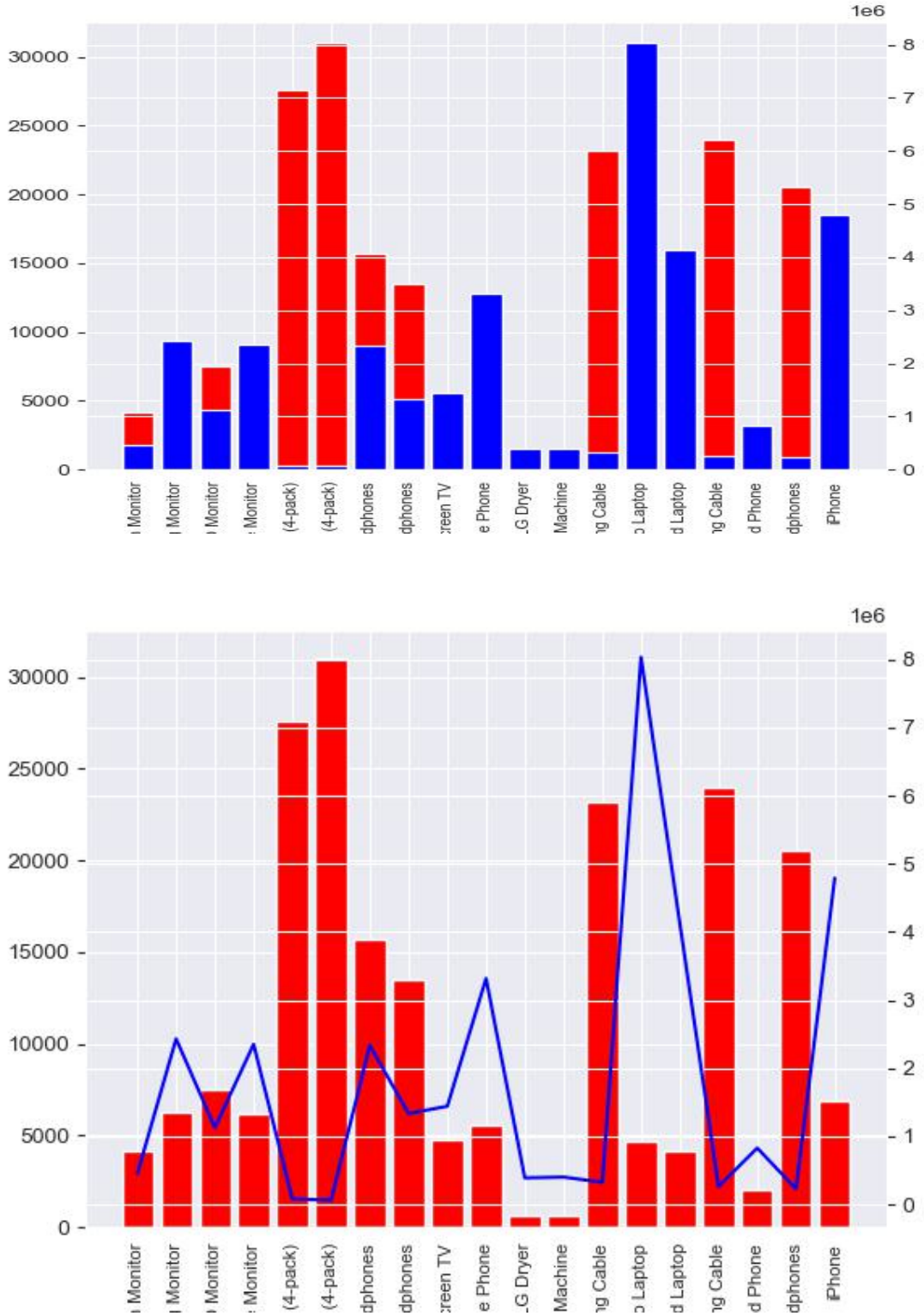


h. <u>Visualization of mean unit price of each product and number of products sold on the same plot.</u>

This graph depicts 2 subplots. We showcased this graphic in 3 different ways to ease understanding and interpretation.



Here blue represents the mean price of each product and red represents the quantity sold or ordered plotted against product name and quantity ordered.



4. Data Analysis

This analysis focused on discovering the relationship between sales, quantity, unit price and revenue.

4.1 Insights and Observations

- December was the month in which the highest sales were recorded
- Amongst all the cities, San Francisco was the city who saw the most commercial success, having the highest number of sales in the year 2019.
- San Francisco also was the city that received the highest number of orders.
- Austin was the city that received the least number of orders.
- Austin generated the least amount of revenue.
- 7pm was the hour during which the highest amount of revenue was generated.
- MacBook pro laptop was the product that garnered the most revenue amongst all the other products.
- AAA Batteries (4 pack) was the product which received the highest number of orders.
- The lowest-priced product was the AAA Batteries (4 pack) while the most costly product was the MacBook pro laptop.
- The AAA Batteries (4 Pack) received 31,017 orders meanwhile the MacBook received 47,28 orders.
- The MacBook Pro laptop received the 6th lowest orders despite being the most expensive product, hinting at the fascinating dynamics of product popularity.
- The LG Dryer and Washing machine were the least sold products.
- A general observation was less costly products received more orders for the most part. Indicating an indirectly proportional relationship between pricing and demand.

4.2 Interpretations and Implications

- **December Sales Increase:** The surge of sales in December may be attributed to various factors such as: holiday season, holiday promotions, end of year promotions and many others. This insight can help with future target marketing strategies.
- **San Francisco's Economic prowess:** The city's dominance in sales and orders prompts an evaluation of the standard of living, minimum wage, consumer demographics and business strategies.
- **Time and Revenue:** 7pm being the time during which highest sales are recorded leaves room for a lot of questions about consumer behaviour. Is this time during which most consumers are on their way home back from home? Answering the question "Why 7pm?" could lead to optimized marketing campaigns.

- **Product Orders:** The AAA Batteries (4 Pack) being the most ordered product goes to show it's undeniably utility or essentiality meanwhile the MacBook's relative moderate orders gives insights to its unique appeal.
- **Price-Order relationship:** Further exploration of the relationship between pricing and demand could uncover consumer preferences and explain the poor performance in sales for some products which could give insights on the measures to be put in place to improve sales.
- The MacBook pro laptop bringing in the most revenue is justified by the fact that it's the most high-priced product and more importantly the relatively moderate demand for it.
- Despite being the most ordered product, the AAA Batteries brought in the least revenue. This contrast is justified by its affordability as it's the lowest priced product.

4.3 Conclusion

Our analysis of the data has revealed a range of insights and observations that explore the essence of customer behaviour, product dynamics and market trends. Here as some key takeaways:

- Timing matters: It is important to understand when and how certain activities can be organized so as to ensure maximum visibility and participation.
- Consumer choices: Understanding consumer choices and consumer demographic explain our price-demand relationship which empowers business decisions.
- Market Representation: San Francisco's performance in sales and orders serves as a case custody for customer behaviour and economic evaluation, offering the room for market expansion plans.

5. Conclusion

In conclusion, after diving deep into data cleaning, exploratory data analysis and data analysis so as to uncover insights, make sense of our data, provide solutions and justify certain occurrences, we've gotten answers guiding us towards informed action and strategic decision-making.

The discoveries from this in-depth analysis serve as guidance for understanding customer choices and trends and also market representation. Through graphs and charts, we've been able to uncover insights.

We've seen the power of raw data and reaffirmed the importance of data analysis.

6. Recommendations

- In our data cleaning phase, we found some missing values, this could be a call for action. Interviewing employees and tightening up security can be a step forward.
- Capitalizing on December sales would be key to seeing business success. Activities such as targeted campaigns, limited promotions and holiday-themed products would attract the public
- Target driving advertisements, promotions and engaging content at 7pm will drive revenue and ensure engagements.
- There's room for pricing adjustments in order to balance the price-demand relationship. Understanding customer demographic, customer preferences and economic situations would be key.