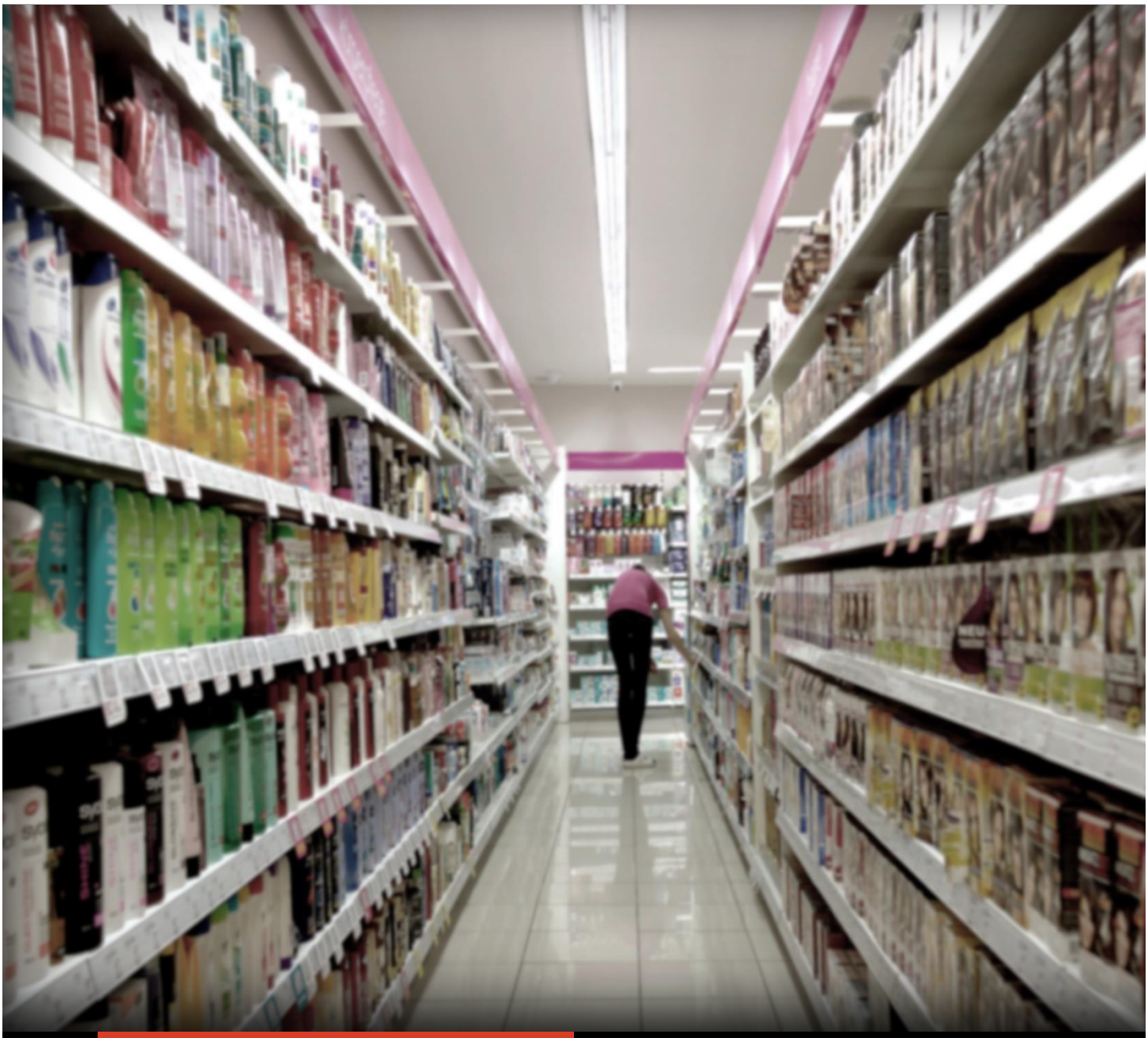


WRANGLE REPORT

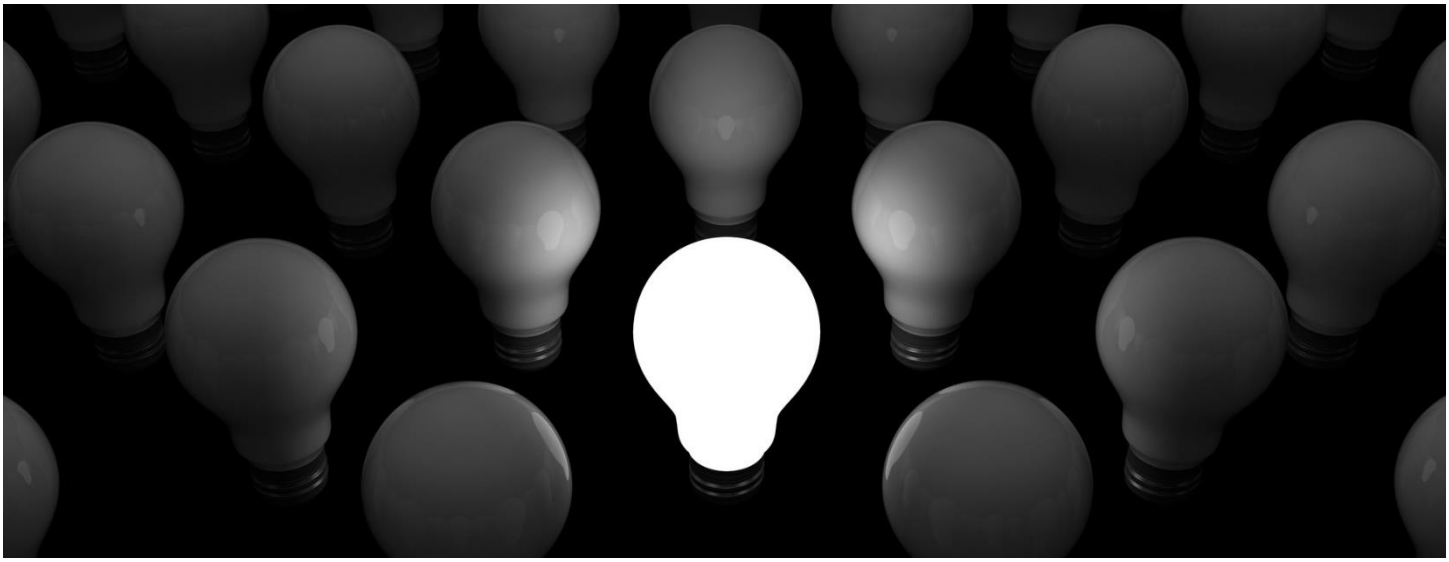
3 MONTHS IN REVIEW

OMAR@GMAIL.COM

SUPERMARKET SALES

INTRODUCTION

In real-world scenarios, data is often requires extensive preprocessing. Utilizing Python and its robust libraries, I will collect data. The next step involves thoroughly assessing the data's quality and structure, followed by a meticulous cleaning process, known as data wrangling. These efforts will be systematically documented in a Jupyter Notebook within the project folder. Furthermore, I will present the wrangled data through comprehensive analyses and visualizations using Python (and its libraries) and POWER BI.



General idea :
The dataset I will be wrangling, analyzing, and visualizing comprises sales data from a supermarket, collected over a span of three months. This dataset includes detailed records of transactions, product categories, sales figures, and customer information. The objective is to clean and preprocess this data to ensure accuracy and consistency, followed by a thorough analysis to uncover sales trends, customer behavior patterns, and product performance. This analysis will provide valuable insights into the supermarket's operations and support data-driven decision-making.

Invoice ID	Branch	Yangon	Naypyitaw	Mandalay	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	Rating
750-67-8428	A	1	0	0	Normal	Male	Health and beauty	74.69	7	26.1415	NaN	1/5/2019	13:08	Ewallet	9.1
226-31-3081	C	0	1	0	Normal	Male	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	9.6
631-41-3108	A	1	0	0	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	7.4
123-19-1176	A	1	0	0	Normal	Male	Health and beauty	58.22	8	NaN	489.0480	1/27/2019	8 - 30 PM	Ewallet	8.4
373-73-7910	A	1	0	0	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	5.3

PROJECT INDEX

Data gather:

- **File load into the notebook**

Data assessing:

- **Explore the data**
- **Identify data quality issues and tidiness issues**

Data cleaning:

- **Fixing each issue that have been Identified**

Data saving:

- **Saving clean version of data**

Data visualization:

- **General analysis**
- **Customer analysis**
- **Product analysis**
- **Payment analysis**
- **Time series analysis**

Data gather:

- **File load into the notebook**

The dataset, 'supermarket sales.csv,' contains detailed records of transactions. Please download this file manually using the following link: GitHub

Data assessing:

- **Explore the data**

After gathering each of the above pieces of data, we need to assess them visually and programmatically for quality and tidiness issues.

```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Invoice ID     1006 non-null   object
 1   Branch         1006 non-null   object
 2   Yangon         1006 non-null   int64
 3   Naypyitaw      1006 non-null   int64
 4   Mandalay       1006 non-null   int64
 5   Customer type  1006 non-null   object
 6   Gender         1006 non-null   object
 7   Product line   1006 non-null   object
 8   Unit price     1006 non-null   object
 9   Quantity       1006 non-null   int64
 10  Tax 5%         997 non-null    float64
 11  Total          1003 non-null   float64
 12  Date           1006 non-null   object
 13  Time           1006 non-null   object
 14  Payment        1006 non-null   object
 15  Rating         1006 non-null   float64
dtypes: float64(3), int64(4), object(9)
```

```
Unique values for column 'Branch':['A' 'C' 'B']
----------
----------
Unique values for column 'Yangon':[1 0]
----------
----------
Unique values for column 'Naypyitaw':[0 1]
----------
----------
Unique values for column 'Mandalay':[0 1]
----------
----------
Unique values for column 'Customer type':['Normal' '-' 'Member' 'Memberr']
----------
----------
Unique values for column 'Gender':['Male' 'Female']
----------
----------
Unique values for column 'Product line':['Health and beauty' 'Electronic accessories' 'Home and lifestyle'
 'Sports and travel' 'Food and beverages' 'Fashion accessories']
----------
```

- **Identify data quality issues and tidiness issues**

data quality issues:

- Tax and Total columns with **missing values**.
- Incorrect data type: Unit Price column stored as object instead of numeric.
- Inconsistent values: Customer Type column with inconsistent values. quantity have negative values
- Mixed units: Unit Price column with "USD" unit.
- Inconsistent format: Time column with 12-hour and 24-hour formats.
- outliers: one outlier in rating column '97' probably meant 9.7 and there is some in total and I n tax column but its not far from threshold so i will not drop it

data tidiness issues:

- city variable separated into 3 columns
- but each city has only one branch so its not big issue

Data cleaning:

- Solving each one of the data issues

Data Visualization: 

After that, we now have a clean, tidy and stored data set. We can now use our visuals to extract some insights from the data, I have made two types of insights. Let's have a look at them.

1. **General analysis**
   Some insights about data

2. **Customer analysis**
   Some insights about customer gender, customer type and other aspects

3. **Product analysis**
   Some insights about product line, unit price and quantity and other aspects

4. **Payment analysis**
   Some insights about payment method, rating and total and other aspects

5. **Time series analysis**

   In here I analyze a lot of aspects with respect to time