Super Market Sales Analysis
Data Wrangling Report

Omar Awad
Ekrami Ahmed
Kenzy Walid
Alaa Eid

Supervision

Eng Yousef Ezz Eldeen

Contents

1 Introduction

Real-world data is often messy and unstructured. To handle this, I will utilize Python and its libraries to collect data from various sources in multiple formats, evaluate its quality and Tidiness, and then clean it. This entire process is referred to as data wrangling. My data wrangling steps will be documented in a Jupyter Notebook, which is included in the project folder, and will be demonstrated through analysis and visualizations using Python (and its associated libraries) as well as Power Bi.The dataset I will be wrangling (and analyzing and visualizing) is the historical sales data of a supermarket company, recorded across three different branches over a span of three months. This dataset is ideal for predictive data analytics as it provides insights into sales trends, customer behavior, and branch performance. The dataset's context is aligned with the increasing growth of supermarkets in highly populated cities, where market competition is fierce. Let's explore the data and gain valuable insights through our analysis.

2 Data Gathering

2.1 File in hand 'Capstone Data - Supermarket Sales'

The historical sales data of the supermarket is provided as a file, which you can imagine as being readily available for analysis. To access this dataset, it can be manually downloaded using the following link: supermarket_sales_data.csv.

3 Data Assessing

3.1 Quality Issues

1. **Completeness:** Data should be complete, meaning that all necessary information is present and accounted for.
2. **Validity:** Data must be valid, ensuring it meets the required format and constraints.
3. **Accuracy:** Data should be accurate, reflecting the correct values and information.
4. **Consistency:** Data must be consistent, with no conflicting information across different sources or records.

Founded Issues

- Duplicates
- Missing values in Tax and Total columns
- Inconsistent values in Quantity column (typo) have negative values
- Inconsistent data type in Unit price column because of inconsistent values

- Inconsistent value (97) in Rating column
- Inconsistent value in Time column (8 - 30 PM)
- Inconsistent values in Customer type column (typo) should write Member instead of Memberr
- Inconsistent values in Customer column (-)

3.2 Tidiness Issues

1. **Each variable forms a column and contains values:** (Yangon, Naypyitaw, Mandalay) types are columns.
2. **Each observation forms a row**
3. **Each type of observational unit forms a table**

4 Data Cleaning

4.1 Fixing Quality Issues

1. **Duplicates:** We removed them.
2. **Missing values in Tax and Total columns:** we calculated them from the function

   Tax 5% = Unit price * Quantity * 0.05

   Total = Tax 5% + Unit price * Quantity
3. **Inconsistent values in Quantity column (typo) have negative values:** we made them positive
4. **Inconsistent data type in Unit price column because of inconsistent values:** We removed the "USD" From the Text and change the data type to folat
5. **Inconsistent value (97) in Rating column:** we changed it to 9.7
6. **Inconsistent value in Time column (8 - 30 PM):** we changed it to 20: 30
7. **Inconsistent values in Customer type column (typo) Memberr :** we changed Memberr values to Member
8. **Inconsistent values in Customer column (-):** Thier count is (27) entries, equivalent to 2.7% of all data ,so we choose to drop them

4.2 Fixing tidiness Issues

Yangon, Naypyitaw, Mandalay) types are columns : we have two ways to fix this problem by mergering the the (Yangon, Naypyitaw, Mandalay) columns City or by branche Branch "A" is "Yangon" City and Branch "B" is Mandalay City and Branch "C" is Naypyitaw City

5 Data Storing

After gathering, Assessing and cleaning data, We need to store the data into a .CSV file to make it easy for access. Which you can find in the project folder.

6 Data Visualization

As part of the comprehensive data exploration, we will delve into several critical areas to better understand the dynamics of sales trends and customer behaviors. These Areas include

- Sales and Revenue Analysis
- Product Performance Analysis
- Marketing Effectiveness Analysis