Applied A.I. Solutions
Foundations of Data Management

Lab Exercises 1

Group-10 members

1. Goyal, Vinayak
2. Bhasgauri, Harshal Shashikant
3. Sebastian, Arun
4. ., Himani
5. Singh, Satyajeet
6. Trongkitroongruang, Kajhonprom
7. Cheng, Qianfan

When analyzing data sheets, Pandas Framework is used in the process of creating this document. All orders, personnel, and returns sheets are scrutinized for missing values, duplicated data, completeness, and discrepancies. The code used to complete this task is included in Appendix I.

1. **Analysis:** data analysis of **Sample Superstore** spreadsheet:

Data Overview:

1) Orders

- Number of Entries: 9994
- Number of Columns: 21

Columns Information:

➢ 1. Row ID 2. Order ID

➢ 3. Order Date 4. Ship Date

➢ 5. Ship Mode 6. Customer ID

➢ 7. Customer Name 8. Segment

➢ 9. Country/Region 10. City

➢ 11. State 12. Postal Code

➢ 13. Region 14. Product ID

➢ 15. Category 16. Sub-Category

- 17. Product Name
- 18. Sales
- 19. Quantity
- 20. Discount
- 21. Profit

Data Quality Analysis:

- Duplicate Rows: 0
- Missing Values:
  - Postal Code: 11 missing entries, if no further data will be provided to fill in these missing values, these missing values will be filled in according to its ‘State’ randomly.

Analysis Summary:

- The dataset is quite clean with no duplicate entries.
- There are some missing values in the 'Postal Code' column that might require attention depending on the use case.
- **Inconsistencies:** 'Sales' and 'Quantity' are unclear because it does not directly state whether sales value includes all quantities, hence further assumptions are made based on this.
- **Redundancies:** This dataset has no redundancies.
- **Detailed Analysis:** Dive deeper into the columns like 'Sales', 'Quantity', 'Discount', and 'Profit' to understand the data distribution and there are no possible errors or outliers.

2) People

- Number of Entries: 4
- Number of Columns: 2

Columns Information:

- 1. Regional Manager
- 2. Region

Data Quality Analysis:

- Duplicate Rows: 0
- Missing Values: None

Analysis Summary:

- The second spreadsheet contains information about Regional Managers and the regions in which they work.

- There are no missing or duplicate values, indicating good data quality for this small dataset.
- When we combine data from this sheet with data from the Orders sheet, we can use only one column to reduce redundant information.

3) Returns

- Number of Entries: 800
- Number of Columns: 2

Columns Information:

- 1. Returned 2. Order ID

Data Quality Analysis:

- Duplicate Rows: 504
- Missing Values: None

Analysis Summary:

- The third spreadsheet contains information about orders, including whether or not they were returned.
- There are **no missing values**, but there are a large number of **duplicate** rows (504), which could be deliberate (if several products per order can be returned) or could necessitate further research and cleaning.

Summary of All Datasets:

- **Sample Superstore Data:** A comprehensive dataset with details about orders, customers, and financials.
- **People Sheet:** A mapping between regional managers and their respective regions.
- **Returns Sheet:** Information about orders that were returned, though it contains many duplicate rows that might need further examination.
- The "**Order ID**" appears to be a common link between "Orders" and "Returns," which could provide information about returns. Similarly, the "**Region**" data in "Orders" might be related to the regional manager data in 'People'.

2. Target Audience

- **Operational Reports:**

  Target Audience: Operations Team, Regional Managers, Customer Service Team

  Intended Use:

  - **Monitor and Control:** Track sales, returns, and customer interactions to identify issues and opportunities in real-time.
  - **Performance Improvement:** Identify areas/products where returns are high, or sales are low and need improvements.
  - **Achievable:** Minimize returns, optimize stock levels, improve customer service, and enhance operational efficiency.
  - **Reduce cost:** Minimize costs by checking the location and planning to send to save money.
  - **Return Problems:** analyze the return product issue and resolve it.
  - **Delivery Days:** Analysts monitor inventory costs from day to day.
  - **Discount:** Analysts discount the reasons for price reductions and the impact on profit loss.

- **Executive Reports**

  Target Audience: Executives, Strategic Planners, Marketing Team

  Intended Use:

  - **Decision Making:** Utilize data to strategize marketing efforts, manage resources, and plan future initiatives.
  - **Research Analysis:** Analyze trends, customer behavior, and sales performance for informed business strategy development.
  - **Achievable:** Make informed strategic decisions, identify market trends, optimize marketing efforts, and enhance overall business strategy.
  - **Pareto Principle:** Focus on the top 20% of customers using the Pareto Principle (80/20) in both profit and loss terms to improve profit and reduce loss.

3. Context and Additional Assumptions.

- The data across all sheets is accurate and up to date.

- The trends and patterns in the historical data are representative and can be utilized for future planning.
- The "Sample Superstore" data represents the entirety of the sales, not a subset.
- 'People' regional managers are in charge of all sales and returns in their particular territories.
- By combining 'Returns' and 'Orders' data, analysts can follow product returns and identify problems.
- Make a bar graph to compare each product and region.
- The term 'Sales' in the sheet refers to the overall sales of the order.
- One 'Sub-Category' will not belong to multiple 'Category's.

4. **Operational and Executive Reports**:
   - **Operational Reports**
     - **Information Displayed:** Sales, Returns, Customer Details, Regional Data, Distribution of Delivery Days
   - **KPIs:**
     - **Sales Performance**: $Sales\ KPIs = \frac{Total\ Sales}{Total\ Orders}$
     - **Sales Target Attainment:** $= \frac{Sales\ for\ the\ current\ period}{Sales\ target\ (Assume)} \times 100\%$
     - **Return Rate:** $Return\ Rate = \frac{Total\ Returns}{Total\ Orders} \times 100\%$

   - **Executive Reports**
   - **Information Displayed**: Sales Summary, Return Overview, Financials

- (Profit, Loss), Customer segment ratio with Delivery days.
- **KPIs:**
  - **Cost of Goods Sold (COGS):** $= Sales - Gross\ Profit$
  - **Profit Margin:** $= \frac{Profit}{Sales} \times 100\%$
  - **Sales Growth:** $= \frac{Current\ peroid\ sales - Sales\ during\ past\ period}{Sales\ during\ period}$
  - **Customer Lifetime Value:** Assumed to be calculated using purchase history and retention rates (requires further data and analysis).
  - **Retention Rate:**

$$= \left[\frac{(Number\ of\ customers\ at\ the\ end\ of\ time\ period - New\ customer\ added)}{Number\ of\ customers\ at\ begining\ of\ the\ time\ period}\right]$$

5. Design empty templates

Information Display													
Region	State	City	Item	Sales Overview			Returns		Regional Performance		Discount	Discount	Gross
				Total Sales	Units Sold	Average Sales Per Order	Total Returns	Return Rate	Sales per region	Returns per region	Applied	Cap	Sale
				($)	(Items)	($)	(Items)	(%)	(%)	(%)	(%)	(%)	($)

KPIs		
Sales Performance	Sales Target Attainment	Return Rate

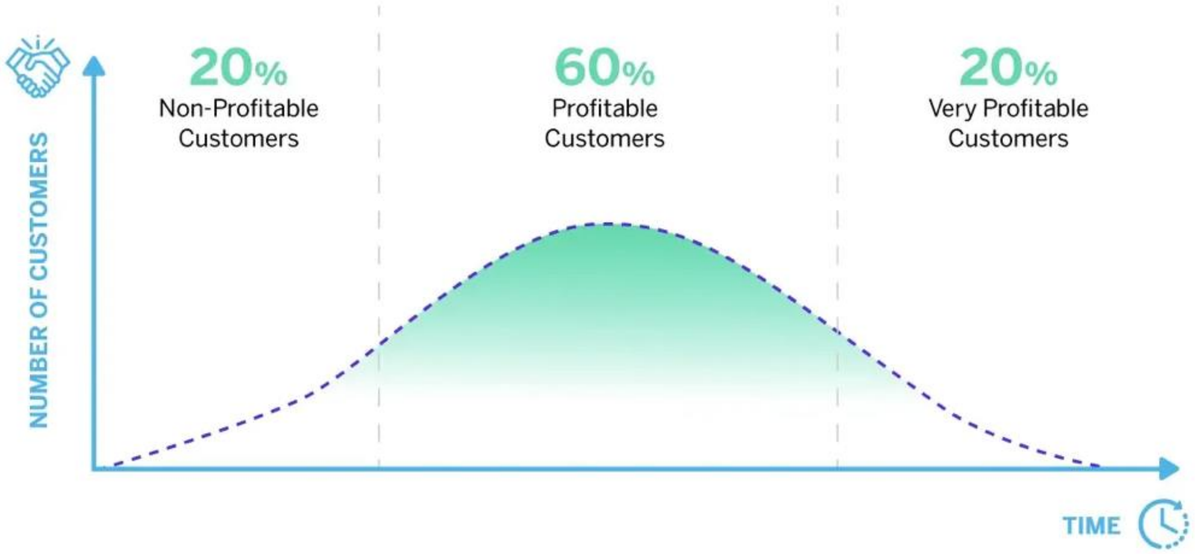
Executive Report Template

Province	Information Display								
	Financial Overview			Sales Summary			Market Trends		Strategic Insights
	Total Profit	Profit Margin	Losses due to Returns	Total Sales	Sales per Segment	Sales per region	Popular product	Emerging customer preferences	Data-driven recommendation and observations

Province	Top Product		Top Customer	
	Profit	Loss	Profit	Loss

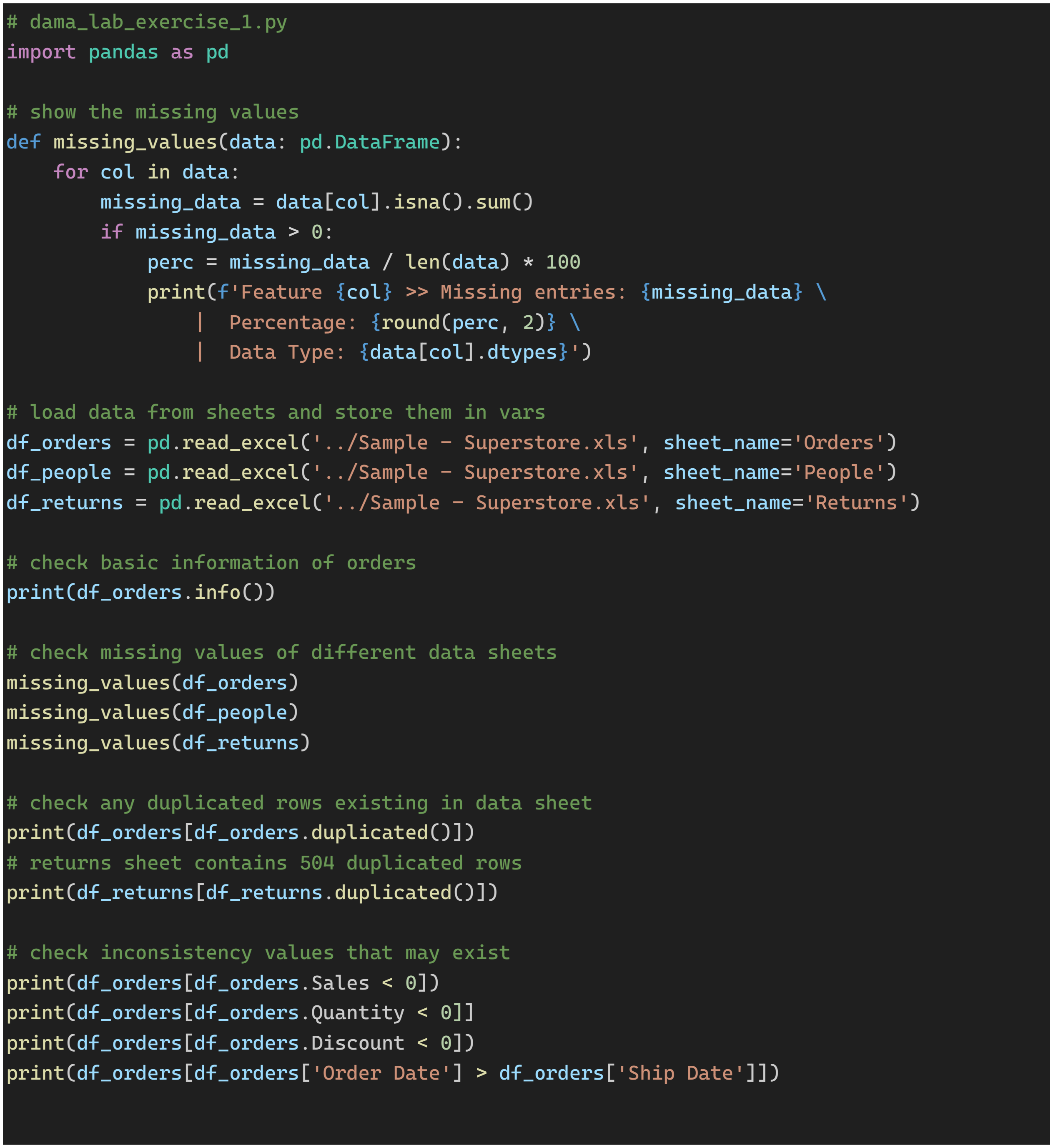
qualtrics.XM

Customer Lifetime Value is the net profit contribution of the customer to the firm over time



KPIs											
Province	2021	2020	2019	2018	Sales Growth		Profit Margin		COGS		Retention Rate
					AVG	Growth (2021)	AVG	Growth (2021)	AVG	Growth (2021)	
	($)	($)	($)	($)	(%)	(%)	(%)	(%)	(%)	(%)	(Numbers)

Appendix I



```python
# dama_lab_exercise_1.py
import pandas as pd

# show the missing values
def missing_values(data: pd.DataFrame):
    for col in data:
        missing_data = data[col].isna().sum()
        if missing_data > 0:
            perc = missing_data / len(data) * 100
            print(f'Feature {col} >> Missing entries: {missing_data} \
                |  Percentage: {round(perc, 2)} \
                |  Data Type: {data[col].dtypes}')

# load data from sheets and store them in vars
df_orders = pd.read_excel('../Sample - Superstore.xls', sheet_name='Orders')
df_people = pd.read_excel('../Sample - Superstore.xls', sheet_name='People')
df_returns = pd.read_excel('../Sample - Superstore.xls', sheet_name='Returns')

# check basic information of orders
print(df_orders.info())

# check missing values of different data sheets
missing_values(df_orders)
missing_values(df_people)
missing_values(df_returns)

# check any duplicated rows existing in data sheet
print(df_orders[df_orders.duplicated()])
# returns sheet contains 504 duplicated rows
print(df_returns[df_returns.duplicated()])

# check inconsistency values that may exist
print(df_orders[df_orders.Sales < 0])
print(df_orders[df_orders.Quantity < 0]]
print(df_orders[df_orders.Discount < 0])
print(df_orders[df_orders['Order Date'] > df_orders['Ship Date']])
```