```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt  # Data visualization library
import seaborn as sns  # Data visualization library for creating informative and attractive statistical graphics

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

```
/kaggle/input/supermarket-sales-data/annex1.csv
/kaggle/input/supermarket-sales-data/annex3.csv
/kaggle/input/supermarket-sales-data/annex2.csv
/kaggle/input/supermarket-sales-data/annex4.csv
```

Project Summary

Welcome to the Sectional Project derived from the main KaggleX Mentorship Final Project Report, titled "Simplifying Data Science: A Comprehensive Retail Business EDA Project Using a Chinese Supermarket Sales Dataset."

In this comprehensive project, I have deconstructed the primary project into six distinct EDA sections, each finely tuned to focus on a specific aspect of exploratory data analysis. The rationale behind this approach is to improve accessibility, allowing you, the reader, to delve into the areas that intrigue you the most. While the complete project remains available for those who prefer to explore it in a holistic environment, these individual sections offer a more specialized perspective on different facets of the data analysis project.

My project goes beyond the traditional scope of conducting an EDA, which is a standard process for experienced Data Analysts. The central idea here is to illustrate a comprehensive understanding of the process, specifically tailored to individuals with limited programming skills. Within these sections, you will discover a wealth of information, insights, and invaluable learning experiences. Each section is designed to stand independently while contributing to the overarching goal of demystifying data science.

Additionally, this project draws learnings from the activities within each section to identify activities that are iterable for any retail business. This contribution aims to develop a comprehensive strategic approach to conducting an EDA on Retail Businesses.

Whether you are a non-technical professional, a seasoned data analyst, a business development expert, or simply someone with a passion for data science, I have endeavored to dissect the process and activities to provide written summaries of key points, thought processes, rationale, and the necessity for each activity. This ensures that, regardless of your background, you will find valuable insights and inspiration throughout these pages.

The primary objective of this project is to bridge the knowledge gap for individuals with limited programming skills, offering step-by-step explanations and strategic insights into the realm of Exploratory Data Analysis.

The sections are structured as follows:

Section 1: Retail Business EDA - Inspect, Prepare, Process Data

Click here to go to Section 1 Project Notebook

Section 2: Retail Business EDA using SQL: Item Category Data (Current Notebook)

Click here to go to Section 2 Project Notebook

Section 3: Retail Business EDA using SQL: Loss Rate Percentage Data

Click here to go to Section 3 Project Notebook

Section 4: Retail Business EDA using SQL: Wholesale Data

Click here to go to Section 4 Project Notebook

Section 5: Retail Business EDA using SQL: Transactions Data

Click here to go to Section 5 Project Notebook

Section 6: Retail Business EDA: SQL Joins - 4 Sales Datasets

Click here to go to Section 6 Project Notebook

I encourage you to explore these sections in an order that best aligns with your interests, needs, and level of expertise. I hope that the project inspires and equips you with the knowledge and enthusiasm to propel your journey in the world of data analytics and international development.

Thank you for embarking on this data-driven adventure with me. I wish you a rewarding and insightful journey through the sectional project.

Reagan R. Ocan
Email: ocanronald@gmail.com
LinkedIn: https://linkedin.com/in/reagan-r-ocan/

About the Author

```
# Query Cell: Data Import from Multiple Files
# Summary:
# In this query cell, four datasets—'veg_category_df,' 'veg_txn_df,' 'veg_whsle_df,' and 'loss_rate_df'—are imported from separate CSV files.
# These datasets contain information related to supermarket sales and associated data, which will be used for further analysis and exploratory data analysis (EDA) tasks.
# Importing these datasets is the initial step in preparing the data for analysis, and they will be explored, cleaned, and examined in subsequent steps to extract insights and make data-driven decisions.

veg_category_df = pd.read_csv(r'/kaggle/input/supermarket-sales-data/annex1.csv')
veg_txn_df = pd.read_csv(r'/kaggle/input/supermarket-sales-data/annex2.csv')
```

```
veg_whsle_df = pd.read_csv(r'/kaggle/input/supermarket-sales-data/annex3.csv')
loss_rate_df = pd.read_csv(r'/kaggle/input/supermarket-sales-data/annex4.csv')
```

Overview of Data Inspection, Preparation, and Processing

We have to start by inspecting and processing our data to facilitate a smooth EDA. Since I have already gone through this process in Section 1, I will keep this data processing section short and sweet. If you are interested in the details of the thought process behind the data processing section, please refer to section 1 of the project.

To group the data processing queries, I have put all of them in one cell with very little comments, so you can skip over to the main project if you have already viewed the project in section 1:

In [3]:

```
# Section 2: Data Inspection, Preparation, and Processing
# Key Section Activities

# Activity 1: Inspect the first few rows of a dataframe called `veg_category_df` to gain an overview of its columns and values.
veg_category_df.head()

# Activity 2: Conduct an Exploratory Data Analysis (EDA) on a dataframe to examine the data types and column names.
veg_category_df.dtypes

# Activity 3: rename the columns in place
new_column_names = ['item_code', 'item_name', 'category_code', 'category_name']
veg_category_df.rename(columns=dict(zip(veg_category_df.columns, new_column_names)), inplace=True)

# Activity 4: Check dataset columns and values to gain overview
veg_category_df.head()
```

Out[3]:

	item_code	item_name	category_code	category_name
0	102900005115168	Niushou Shengcai	1011010101	Flower/Leaf Vegetables
1	102900005115199	Sichuan Red Cedar	1011010101	Flower/Leaf Vegetables
2	102900005115625	Local Xiaomao Cabbage	1011010101	Flower/Leaf Vegetables
3	102900005115748	White Caitai	1011010101	Flower/Leaf Vegetables
4	102900005115762	Amaranth	1011010101	Flower/Leaf Vegetables

Notebook Overview

Exploring the Category Dataset - Unveiling Insights through EDA Queries

NB: This project is Section 2 of a 6 Section project. Please refer to the project summary above for details on the complete project

Introduction

In this section, we embark on a journey to unravel the secrets hidden within the "Category Dataset." Through a series of Exploratory Data Analysis (EDA) queries, we aim to shed light on the intricacies of this dataset, unlocking its potential for decision-making in areas such as inventory management, marketing strategies, and category-specific analyses.

Before we dive into the queries, it's essential to recognize the significance of EDA. Exploratory Data Analysis is not just about crunching numbers; it's about unraveling stories and insights hidden within the

data. The "Category Dataset" is more than just rows and columns; it holds the key to understanding the trends, patterns, and relationships that can shape our strategic decisions.

The Power of Queries

Each query we present in this section is like a magnifying glass, focusing on specific aspects of the dataset. Think of them as our investigative tools, enabling us to unveil insights that may remain hidden in the raw data. These queries serve as gateways to better comprehending the dataset's structure and composition.

Inventory Management

Understanding the composition of the category dataset is vital for efficient inventory management. It allows us to determine which categories have high or low stock levels, identify fast-moving products, and plan for restocking accordingly. By leveraging EDA queries, we can ensure that our inventory is optimized to meet demand while minimizing overstock or stockouts.

Marketing Strategies

In the world of marketing, knowledge is power. EDA queries help us understand customer preferences, buying behavior, and the performance of different product categories. Armed with this knowledge, we can tailor marketing campaigns to target specific customer segments, improve product recommendations, and ultimately drive sales and customer engagement.

Category-Specific Analyses

Each category within the dataset has its unique characteristics, and EDA queries allow us to explore these distinctions. Whether it's pricing trends, customer reviews, or product attributes, a category-specific analysis can help us fine-tune our strategies and offer a personalized experience to customers within each category.

In [4]:

```
# Database Integration for Data Storage
```

In this phase, we incorporate the data from various DataFrames into a SQLite database using SQLAlchemy for effective data management and retrieval.

```
# Import SQLAlchemy and Create a SQLite Engine
# import sqlalchemy and create a sqlite engine
from sqlalchemy import create_engine
from sqlalchemy import text
engine = create_engine('sqlite://', echo=False)
con=engine.connect()

# Store DataFrames in SQLite Tables
veg_category_df.to_sql("veg_cat", con=engine)
```

Query 1: Inspect the initial rows and columns of 'veg_cat' dataframe and gain an overview of the dataset structure.

Before conducting an in-depth exploratory data analysis (EDA) on the 'veg_cat' table, it's beneficial to Query the entire dataframe to inspect the initial rows and columns. This initial query aids in gaining an overview of the dataset's structure and serves as a foundational step in developing an analytical roadmap for the specific dataframe.

```
# Define query 1: Query the entire dataframe to inspect and gain an overview of the dataset structure.
query = text(
'''SELECT * FROM veg_cat ''')
df_sql = pd.read_sql_query(query, con)
# Display the First Few Rows of the 'veg_cat' Table
df_sql.head()
```

	index	item_code	item_name	category_code	category_name
0	0	102900005115168	Niushou Shengcai	1011010101	Flower/Leaf Vegetables
1	1	102900005115199	Sichuan Red Cedar	1011010101	Flower/Leaf Vegetables
2	2	102900005115625	Local Xiaomao Cabbage	1011010101	Flower/Leaf Vegetables
3	3	102900005115748	White Caitai	1011010101	Flower/Leaf Vegetables
4	4	102900005115762	Amaranth	1011010101	Flower/Leaf Vegetables

Query 2: Total Number of Items

- **Objective**: Count the total number of distinct items within the dataset.
- **Rationale**: Understanding the variety of products and identifying potential gaps or outliers is crucial for comprehending the dataset.
- **Significance**: This information is fundamental for decision-making, especially when assessing inventory management, stock replenishment, and understanding the scale of the product range.

```
# Define Query 2: Total Number of Items Total Number of Items
query =text(
'''select count(distinct[item_code]) as dist_item_code from veg_cat''')
df_sql = pd.read_sql_query(query,con)
df_sql.head()
```

	dist_item_code

0	251

Query 3: Total Number of Categories

- **Objective**: Determine the total number of distinct categories within the dataset.
- **Rationale**: Grasping the number of unique categories is essential for effective EDA. It assists in organizing and categorizing products, which is valuable for structuring inventory.
- **Significance**: Knowing the total number of categories is necessary when creating marketing strategies, managing category-specific promotions, and analyzing sales trends by product groups.

```
# Define Query 3: Total Number of Categories
query =text('''select count(distinct[category_code]) as dist_cat_code from veg_cat''')
df_sql = pd.read_sql_query(query,con)
df_sql.head()
```

	dist_cat_code
0	6

Query 4: List of Distinct Categories

- **Objective**: Extract a list of unique category names.
- **Rationale**: Providing an overview of the different product categories aids in understanding how items are grouped and labeled, which is crucial for meaningful analysis.
- **Significance**: Having a list of distinct categories is essential for segmenting data for category-specific analysis, reporting, or product classification.

```
# Define Query 4: List of Distinct Categories

query =text('''select distinct[category_name] as dist_cat from veg_cat''')
df_sql = pd.read_sql_query(query,con)
df_sql.head()
```

	dist_cat
0	Flower/Leaf Vegetables
1	Cabbage
2	Aquatic Tuberous Vegetables
3	Solanum
4	Capsicum

Query 5: Total Items per Category

- **Objective**: Calculate the number of items within each category.
- **Rationale**: Understanding the distribution of items across categories is important in EDA. It provides insights into the popularity of categories and identifies those with a wide or narrow product range.
- **Significance**: Knowing the total items per category is essential for decision-making when planning inventory allocation, category-specific marketing, and assessing category performance.

For visualizing query 5:

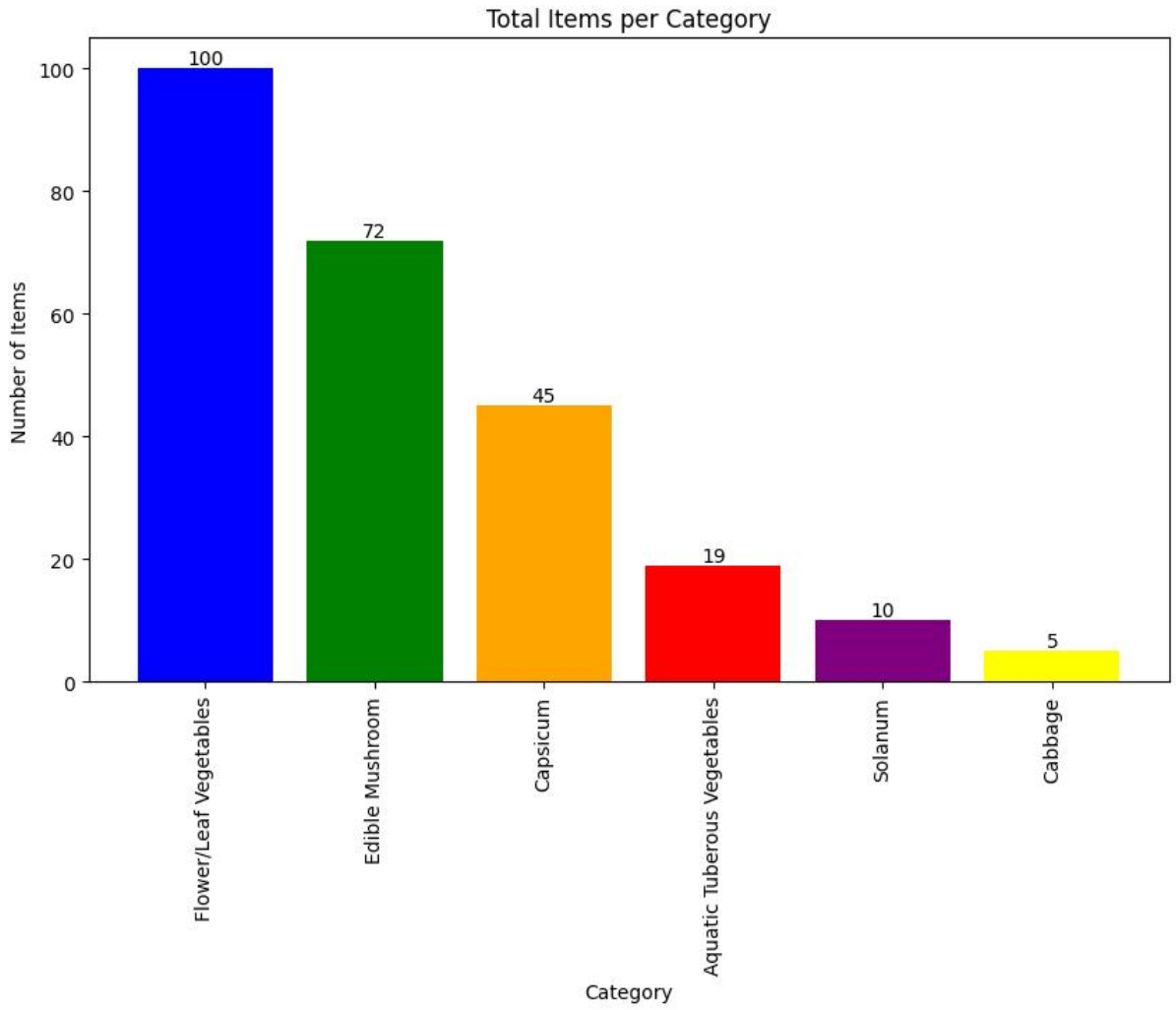
- The bar chart provides a visual representation of the total number of items in each category, allowing you to easily identify categories with a higher or lower number of items.

- The x-axis represents the category names, while the y-axis represents the number of items. The height of each bar corresponds to the number of items in the respective category.
- I created a color palette using a list of colors. Each bar in the bar chart will be assigned a different color from this palette.
- Additionally, I used the plt.text() function to annotate each bar with the corresponding number of items.

In [9]:

```
# Define Query 5: Total Items per Category
query =text('''select [category_name], Count(*) as no_of_items from veg_cat Group by [category_name] ''')
df_sql = pd.read_sql_query(query,con)
df_sql.head()

import matplotlib.pyplot as plt

# Assuming you have already executed the query and stored the results in df_sql
# Sort the dataframe by the number of items in descending order
df_sql_sorted = df_sql.sort_values('no_of_items', ascending=False)

# Create a color palette for the bars
colors = ['blue', 'green', 'orange', 'red', 'purple', 'yellow']

# Create the bar chart
plt.figure(figsize=(10, 6))
bars = plt.bar(df_sql_sorted['category_name'], df_sql_sorted['no_of_items'], color=colors)

# Annotate each bar with the number of items
for bar in bars:
    height = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, height, height, ha='center', va='bottom')

plt.xlabel('Category')
plt.ylabel('Number of Items')
plt.title('Total Items per Category')
```

```
# Rotate the x-axis labels for better readability
plt.xticks(rotation=90)

# Display the plot
plt.show()
```



Query 6: Categories with Lowest Frequency of Items

- **Objective**: Identify the category with the lowest frequency of items.
- **Rationale**: Finding categories with low item frequency highlights potential areas of improvement. It may indicate underdeveloped categories that require attention or a different approach.

- **Significance**: Identifying categories with the lowest frequency is necessary when allocating resources for category expansion, inventory replenishment, or marketing efforts in underrepresented segments.

In [10]:

```
# Define Query 6: Categories with Lowest Frequency of Items
query =text('''select [category_name], COUNT(*) from veg_cat Group by [category_name] Order by
COUNT(*)
Limit 1''')
df_sql = pd.read_sql_query(query,con)
df_sql.head()
```

	category_name	COUNT(*)
0	Cabbage	5

Query 7: Categories with Highest Frequency of Items

- **Objective**: Determine the category with the highest frequency of items.
- **Rationale**: Pinpointing categories with the highest item frequency helps identify strong product segments. This information can be used to prioritize marketing or sales strategies.
- **Significance**: Identifying categories with the highest frequency is necessary for decision-making when allocating marketing resources, optimizing product listings, and focusing on areas of business strength.

In [11]:

```
# Define Query 7: to get categories with Highest Frequency of Items:
query =text('''select [category_name], COUNT(*) from veg_cat Group by [category_name]
Order by COUNT(*) DESC
Limit 1''')
df_sql = pd.read_sql_query(query,con)
df_sql.head()
```

	category_name	COUNT(*)
0	Flower/Leaf Vegetables	100

Strategic Insights from Section 2:

Exploring the Category Dataset - Unveiling Insights through EDA Queries

In order to create a strategic and iterative approach for Exploratory Data Analysis (EDA) on the "Category Dataset," we will consolidate the queries executed in this project into key activities that form the process. These activities help understand the dataset's structure and make informed decisions regarding inventory management, marketing strategies, and category-specific analyses.

Activity 1: Data Overview and Statistics

- **Description:** This activity involves gaining an initial understanding of the dataset's characteristics, such as the total number of items and categories.
- **Rationale:** Gaining an overview of the dataset is a foundational step in any data analysis process.
- **Necessity:** Essential for assessing inventory management and product categorization.
- **Queries in this project:** Queries like "Total Number of Items" (Query 1) and "Total Number of Categories" (Query 2) exemplify this activity

Activity 2: Categorization and Classification

- **Description:** This activity focuses on extracting and organizing category-related information.
- **Rationale:** Helps in structuring data for category-specific analysis, marketing strategies, and inventory management.

- **Necessity:** Fundamental for analyzing sales trends, organizing promotions, and assessing category performance.
- **Queries in this project:** Queries like "List of Distinct Categories" (Query 3) and "Total Items per Category" (Query 4) exemplify this activity.

Activity 3: Identify Category Insights

- **Description:** In this activity, you explore categories with both low and high item frequencies.
- **Rationale:** Identifies categories that may require attention or those with strong product segments.
- **Necessity:** Necessary for optimizing marketing, focusing on business strengths, and allocating resources effectively.
- **Queries in this project:** Queries like "Categories with Lowest Frequency of Items" (Query 5) and "Categories with Highest Frequency of Items" (Query 6) exemplify this activity.

By following this strategic approach, you can systematically explore the "Category Dataset" and make data-informed decisions related to inventory management, marketing strategies, and category-specific analyses.

Section 2 Next Steps

Exploring the Category Dataset - Unveiling Insights through EDA Queries

Building on the valuable insights gained from our EDA queries in Section 2, I warmly invite you to join me in the next phase of our enlightening journey. Together, we can take our project to new heights, expand its impact, and make a lasting contribution to the field of retail data analytics. Here are the suggested next steps, encompassing both traditional data analysis and advanced machine learning and deep learning approaches:

1. Further Query Development:

Let's continue our journey of discovery by developing more EDA queries for the "Category Dataset." We'll explore additional facets of the data, such as seasonality, sales trends over time, and customer behavior within specific categories. These queries will unlock deeper insights into our dataset.

2. PowerBI Dashboard Development:

In this step, we'll create interactive PowerBI dashboards that bring our insights to life. These dashboards will offer a user-friendly interface for stakeholders to explore the data, make informed decisions, and track key performance indicators related to inventory management and marketing strategies.

3. Machine Learning for Predictive Analytics:

Imagine the potential of implementing machine learning models to predict inventory demand. By analyzing historical data and considering factors like seasonality, customer behavior, and category-specific trends, we can forecast future inventory needs with precision. This translates to optimized stock levels and reduced carrying costs.

4. Recommendation Systems:

Our journey takes us to the development of recommendation systems using collaborative filtering or content-based filtering. These systems will suggest products to customers based on their past purchases and preferences, creating a seamless shopping experience and driving sales.

5. Deep Learning for Image Analysis:

Should our dataset contain images of products, we'll embark on a fascinating exploration of deep learning techniques, such as convolutional neural networks (CNNs), to analyze product images. This opens doors to automated product categorization, quality control, and image-based recommendation systems.

6. Collaborative Projects and Data Sharing:

Collaboration is at the heart of our journey. We'll collaborate with fellow data analysts, data scientists, and retail businesses to share knowledge and insights. Sharing our project's findings, methodologies, and code contributes to a wider community of practice, fostering collective learning and growth.

7. Education and Outreach:

Our journey leads us to educate individuals, particularly those with limited programming skills. We'll conduct workshops, webinars, and seminars to share the principles of data analysis and the practical use of our project's findings. Together, we'll inspire more people to explore the world of data analytics and data-driven decision-making.

8. Feedback and Continuous Improvement:

We'll actively seek feedback from users and stakeholders to enhance our project. Suggestions and insights will guide improvements in our EDA queries, machine learning models, and data visualization tools.

9. Documentation and Knowledge Sharing:

Comprehensive documentation will be at the core of our journey. We'll maintain thorough records of our project's development and findings. Our insights, experiences, and best practices will be shared through articles, blogs, and open-source contributions, benefiting the wider data science community.

10. Customer Segmentation:

We'll utilize machine learning techniques to segment customers based on their purchase behavior, preferences, and interactions with different product categories. This segmentation will empower personalized marketing strategies, product recommendations, and customer engagement efforts. However, it's important to note that additional information on customers in the dataset will be needed to realize this step fully.

11. Sentiment Analysis:

The next phase of our journey involves implementing sentiment analysis on customer reviews and feedback related to different categories. This analysis will provide insights into customer satisfaction, identify areas for improvement, and guide the crafting of marketing messages and product descriptions that resonate with our customers.

Together, we'll continue to advance our mission of simplifying data science and making data-driven decision-making accessible to a diverse audience. Our project's impact in the retail business domain will grow, and this journey promises to be both enriching and impactful.

I hope you join me on our next exploration of possibilities!

Conclusion

Exploring the Category Dataset - Unveiling Insights through EDA Queries

In the journey through Section 2, "Exploring the Category Dataset - Unveiling Insights through EDA Queries," we embarked on a data-driven adventure to unlock the secrets held within the "Category Dataset." This section was designed to be a strategic guide for understanding the dataset's structure, with a focus on enabling data-informed decisions in the realms of inventory management, marketing strategies, and category-specific analyses.

Our exploration was guided by the principles of Exploratory Data Analysis (EDA), a process that transcends mere number-crunching, diving into the realm of stories and insights hidden within the data. We recognized that the "Category Dataset" is not a mere collection of rows and columns; it's a treasure trove of trends, patterns, and relationships that can shape our strategic decisions.

The power of queries became evident as we used them as investigative tools, each query acting as a magnifying glass to focus on specific aspects of the dataset. These queries served as gateways to better comprehend the dataset's structure and composition.

We delved into the significance of our findings for inventory management, marketing strategies, and category-specific analyses. Through EDA, we unveiled insights that would aid in optimizing stock levels, tailoring marketing campaigns, and offering personalized experiences to our customers based on their preferences and buying behavior.

As we conclude this section, it's important to emphasize that our exploration of the "Category Dataset" is just the beginning. The journey into data science is a continuous evolution. The insights gained from these queries are not static numbers; they are dynamic building blocks for informed decision-making.

The next steps are promising, as we move on to further query development, advanced machine learning, and deep learning applications. We will develop interactive PowerBI dashboards, collaborate with others, educate and inspire individuals from diverse backgrounds, and continue to improve and document our journey.

Our project's goal remains unchanged: to bridge the knowledge gap for individuals with limited programming skills, offering step-by-step explanations and strategic insights into the realm of Exploratory Data Analysis.

Thank you for joining us on this data-driven adventure. We hope that this section has provided you with valuable insights and inspiration for your own journey in the world of data analytics and international development. Stay tuned for the exciting chapters ahead, as we continue to simplify data science and make it more accessible to all.

Reagan R. Ocan
Email: ocanronald@gmail.com
LinkedIn: https://linkedin.com/in/reagan-r-ocan/

Our data-driven journey continues in the next section, where we explore the "Loss Rate Percentage" dataset.

Collaboration and Engagement:

If you're inspired by this journey and have insights to share, we invite you to fork this notebook and contribute your perspectives. Engage with us through the comments section to provide feedback, share opinions, and offer valuable advice. Together, we can enhance this notebook and expand its horizons, demystifying data science and making it more accessible to all.

Let's Work Together:

Beyond contributing to this notebook, we're open to collaborating on data projects. If you have a project in mind or need data analysis support, don't hesitate to reach out. Whether it's a collaborative effort or data-driven solutions for your business, we're here to assist. Contact us through the provided email or LinkedIn for inquiries and discussions.