**BDAT1007 – Group Final Project**

**Technical Design Document (TDD)**

Professor: Mr. Robin Huang

Student: Namood-e-Sehar

Mikhail Mishenin

Qin Ren

Fangi Chen

Behnaz Selehipourshirazi

Oluwadamilare Shiji

Dec 09, 2023

**Table of Contents**

Executive Summary-------------------------------------------------------------------------P1

Data Size & Scalability----------------------------------------------------------------------P1

Part 1: Technique Structure Designs----------------------------------------------------P1

Part 2: Implementation----------------------------------------------------------------------P1

Part 3: Implementation----------------------------------------------------------------------P1

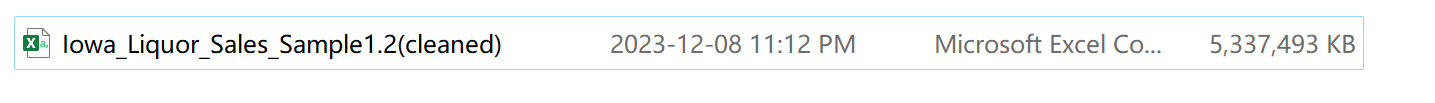
Part 4: Power BI - Data Visualization---------------------------------------------------P1  
  
Conclusion—----------------------------------------------------------------------------------P1

**Executive Summary:**

The purpose of this document is to outline the technical framework and methodologies used in developing comprehensive data mining solutions for our clients. Our technology stack includes Python for data processing analytics, and advanced machine learning models, for database management, and Power BI for data visualization and business intelligence reporting. This document provides a detailed description of the toolsets, coding languages, data operators, data models, data flow, and pipeline design.

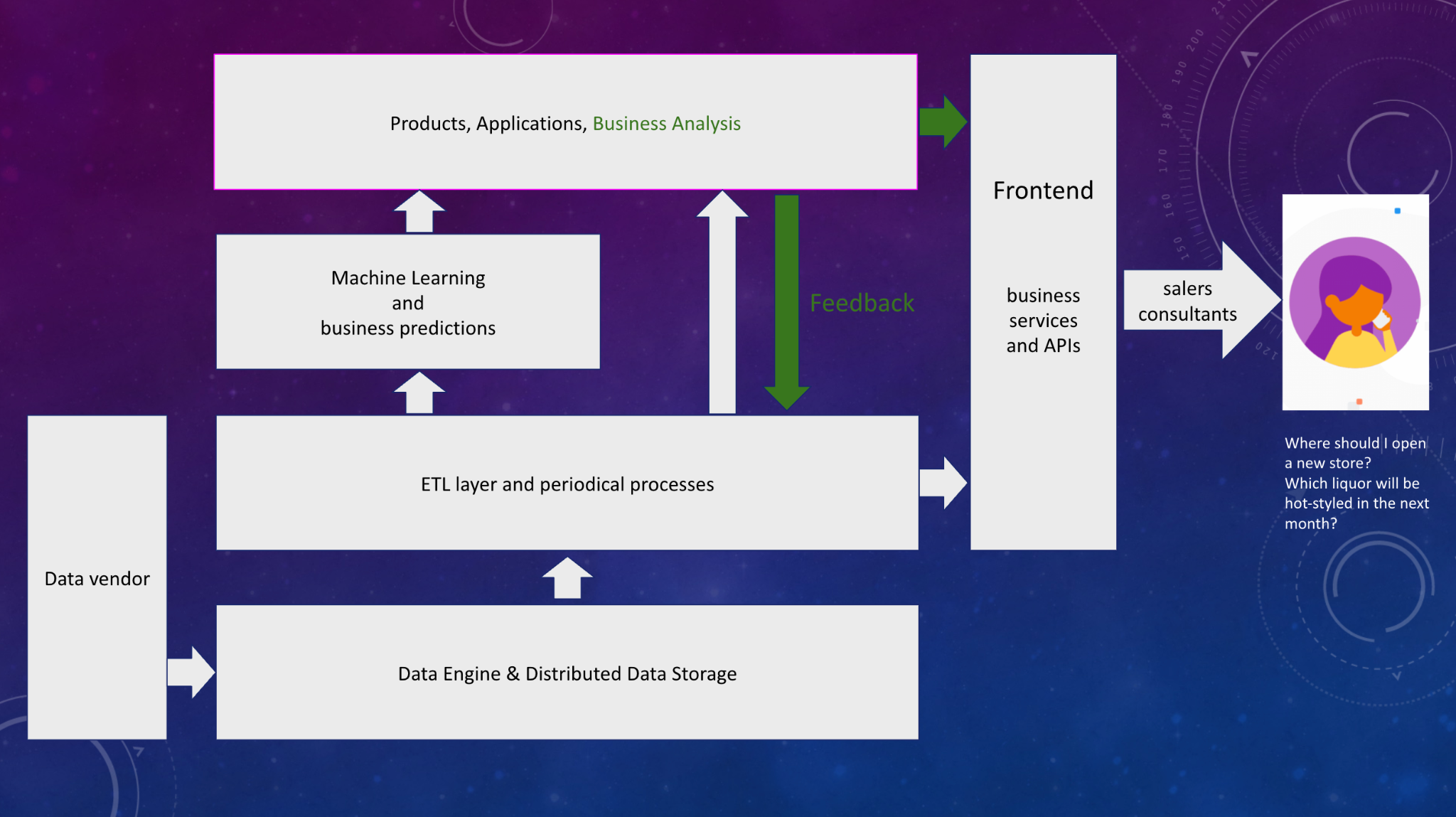
Our aim is to harness the capabilities of these diverse yet integrated technologies to transform complex data sets into actionable insights. By doing so, we empower our clients to achieve significant cost savings, market expansion, and increased profitability. Through a strategic combination of these tools, we deliver sophisticated, end-to-end data mining solutions tailored to meet the unique needs and challenges of our clients, ranging from startups to Fortune Global 500 entities.  
  
**Data Size & Scalability**

* The dataset originates from a CSV file named 'Iowa\_Liquor\_Sales.csv'.
* The dataset sizes: 5G.



* We implement efficient data storage and processing mechanisms to handle large-scale data analytics.

**Part 1: Technique Structure Designs**

****

**Data Vendor:**

This represents the external sources from which data is obtained. Data vendors can be third-party services that provide datasets for use in various applications.

**Data Storage:**

This is where the data is stored after being collected from the data vendors. It implies the use of a distributed system for storage, which is scalable and can handle large volumes of data.

**ETL:**

ETL stands for Extract, Transform, Load. This layer is responsible for taking data from the distributed data storage, transforming it into a suitable format, and loading it into a data warehouse or analytical model. Periodical processes include batch jobs for data processing that run on a schedule.

**Machine Learning:**

At the Machine Learning level, we deploy sophisticated algorithms to process transformed data, generating vital business predictions. This innovative feature transcends traditional reliance on data analysts for business decision-making and advisory support. Empowered by cutting-edge artificial intelligence algorithms, we now garner insights that bolster our business choices. Our data analysts, who traditionally operated at this stratum, are now liberated from the minutiae of data crunching, redirecting their expertise towards more abstract business iterations and the integration of data analytics.

This newfound capability expands our business horizons, enabling us to address our clients' core concerns more comprehensively and expediently. Backed by machine learning, we tackle problems with enhanced confidence, expending less time and yielding more focused and confident decision-making.

**Products, Applications, Business Analysis:**

The pinnacle of our architecture, this layer fuses the intelligence gleaned from machine learning models into our suite of products and applications. It's the crucible where business analysis is transformed, converting raw data into actionable insights. With the newfound agility provided by machine learning, data analysts can channel their acumen into evolving our offerings and refining the alchemy of data into strategic business directives.

**Frontend, Business Services and APIs:**

The frontend serves as the interactive canvas where users engage with our system. It elegantly showcases processed data, analytics, and business intelligence, all rendered in an intuitive format. This interface is now more robust than ever, enriched by deeper machine learning insights, providing users with a powerful tool to navigate and interpret complex data effortlessly.

This layer is the powerhouse of our frontend, offering a spectrum of functionalities through APIs. These services are the cogs in our machinery, enabling seamless interactivity and integration with our backend intelligent systems. With APIs fueled by machine learning insights, we offer a more dynamic, responsive, and personalized user experience.

**Sales, Consultants, Customers:**

They are the ultimate beneficiaries of our architectural ecosystem, leveraging the intelligence served up by the frontend to refine their sales tactics and strategic decisions. Our advancements in machine learning translate into enriched consultation, sharper sales strategies, and ultimately, a heightened customer experience that is both insightful and efficient.

**Part 2: Implementation - Data Preprocessing and Exploratory Data Analysis**

In this phase, we have developed a prototype that encapsulates the core components of our Technical Structure, comprising:

* Data Extraction, Transformation, and Loading (ETL)
* Machine Learning
* Business Analysis

This prototype forms the backbone of our system, allowing us to streamline data workflows, apply analytical models for predictive insights, and conduct in-depth business analysis. It serves as the foundation for further expansion into distributed data storage solutions and interactive frontend interfaces.

**Here is a brief summary of our prototype implementation.**

**Toolset and Coding Language(s)**

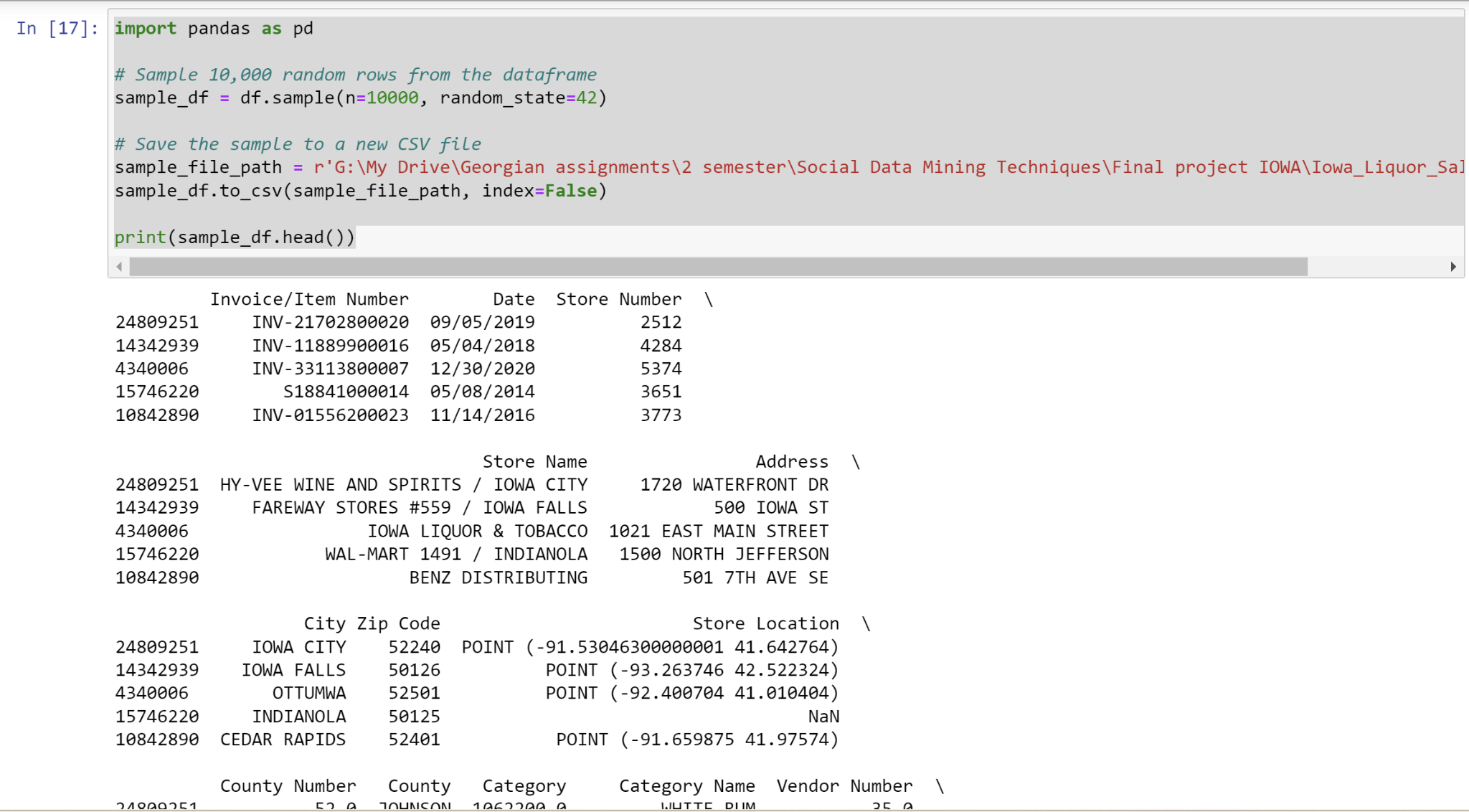
* **Python**: Our primary programming language, chosen for its robust libraries and community support.
* **Pandas Library**: Used for efficient data manipulation and analysis within Python.
* **Jupyter Notebook**: An interactive computing environment that enables us to create and share documents that contain live code, equations, visualizations, and narrative text.

**Data Operators & Models**

* **Data Operators**:
  + **read\_csv**: To import and read data from CSV files.
  + **dtype**: To ensure data types are correctly identified for accurate processing.
  + **describe**: For summarizing statistical data and understanding distributions.
  + **isnull**: To check and sum up missing values in the dataset.
  + **duplicated**: To identify and quantify duplicate entries.
* **Data Models**:
  + Descriptive statistical models to summarize data features.
  + Predictive models to forecast future trends based on historical data.
  + Data sampling techniques to generate representative subsets for analysis.

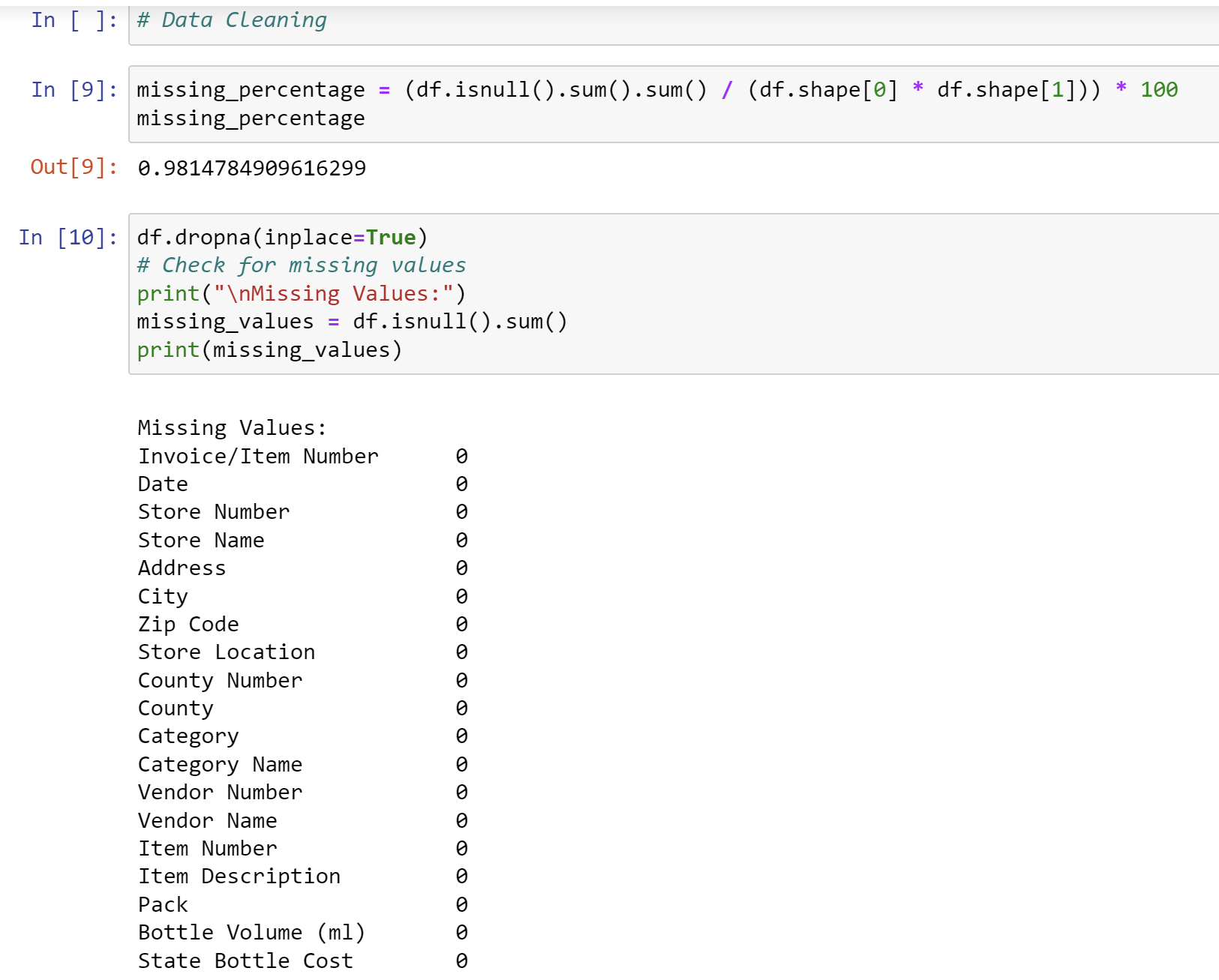
**Data Flow & Pipeline**

1. **Data Ingestion**: Securely importing large datasets from the sources.

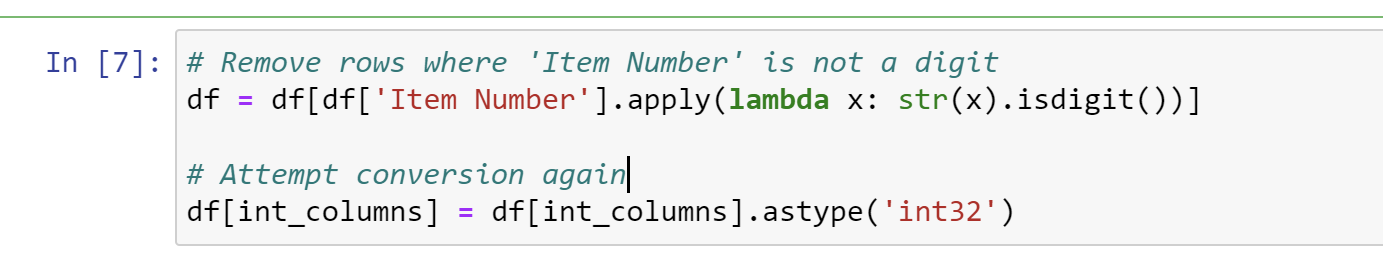


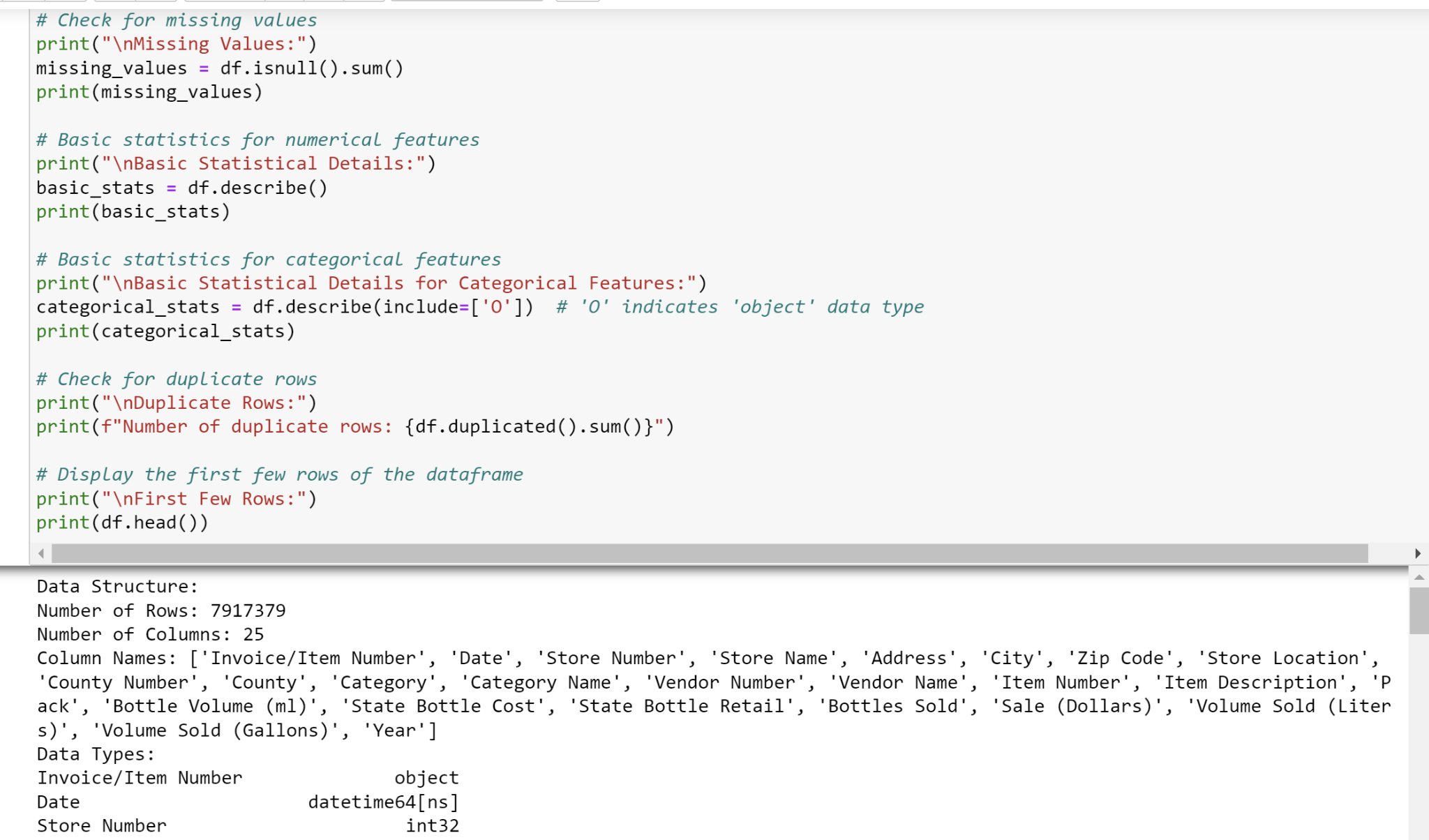
1. **Data Cleaning**: Identifying and rectifying missing values, duplicate data, and inconsistencies.

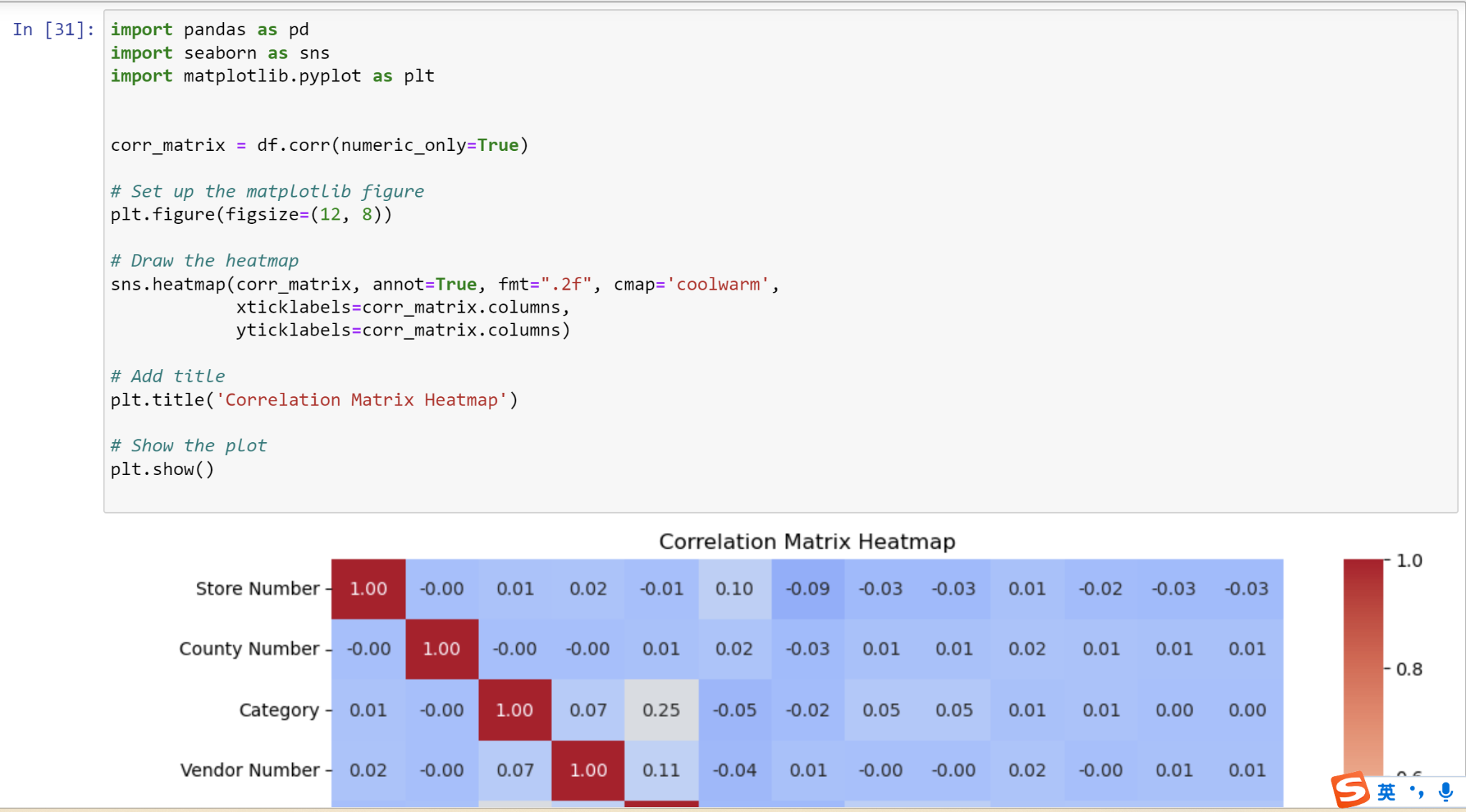
Any rows with missing values will be removed.



Remove invalid rows:



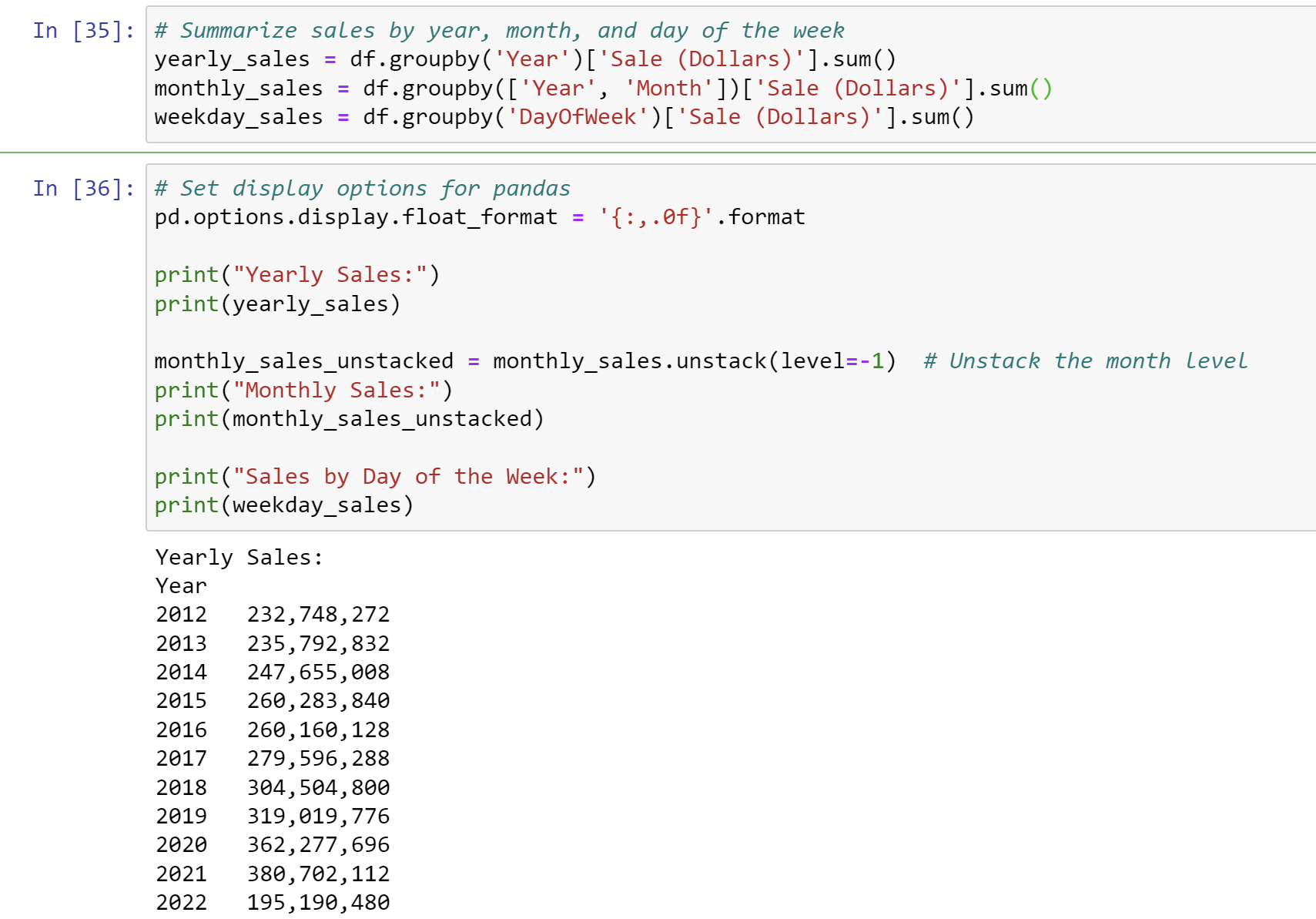
Check duplications:  


1. **Data Exploration**:  
   Calculating correlation matrix is a key part of exploratory data analysis and understanding feature relations:  
   
2. **Data Transformation**: Normalizing and transforming data to fit the needs of predictive models

Data type unifying:

****

Creating Time Series for future use:



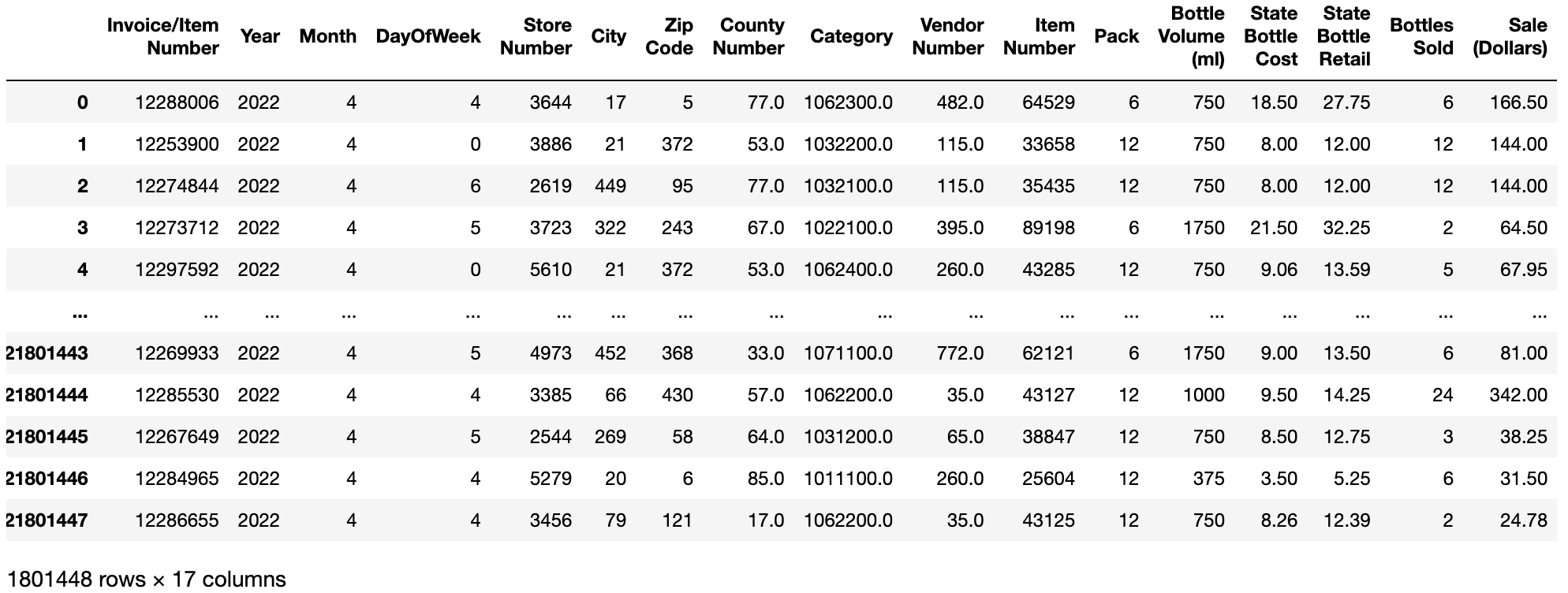
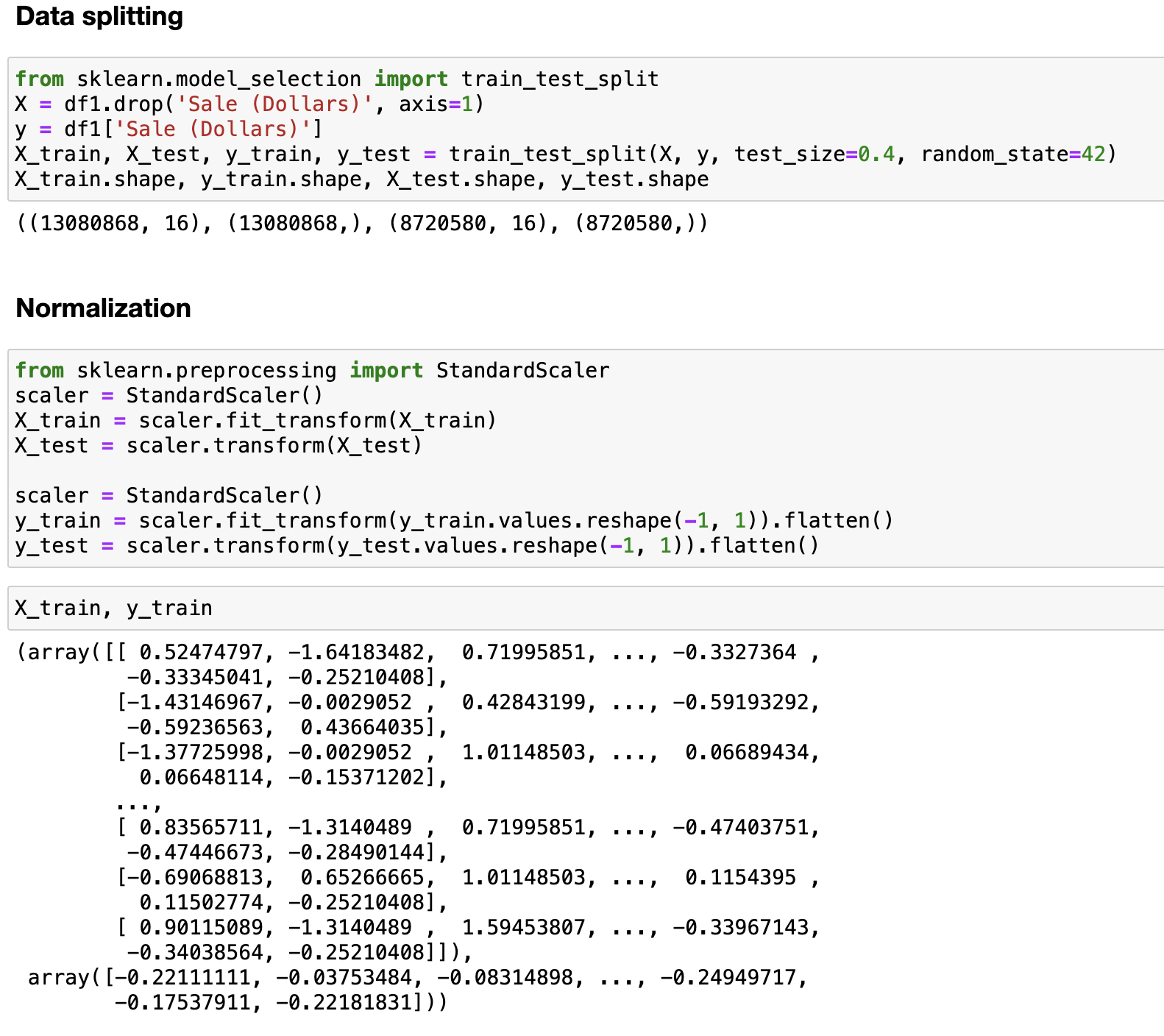
1. **Model Development**: Creating and training models to uncover patterns and make business predictions

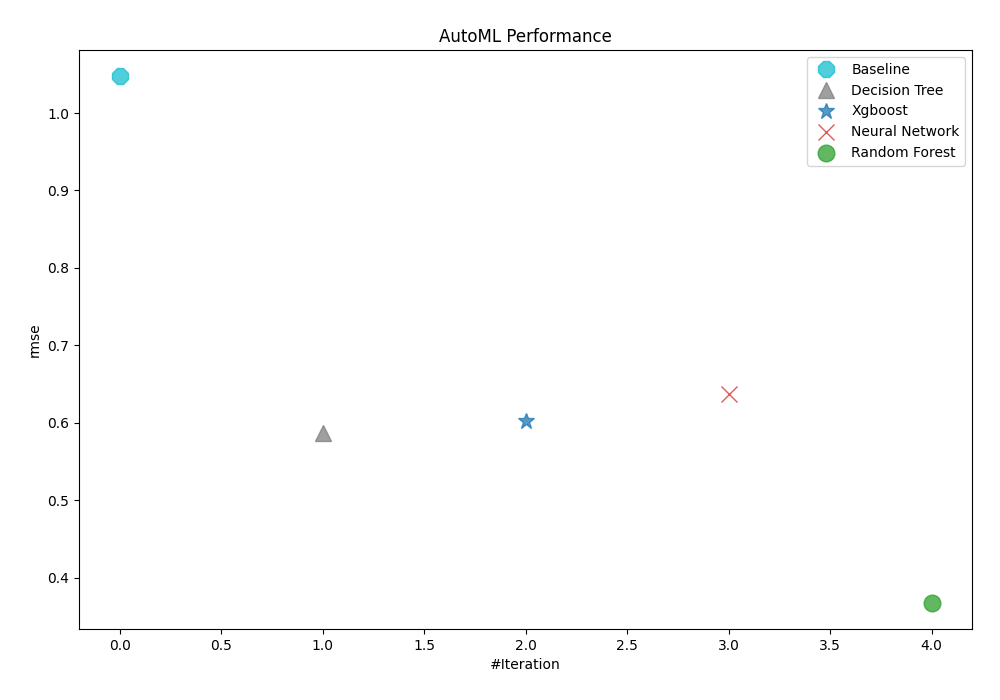
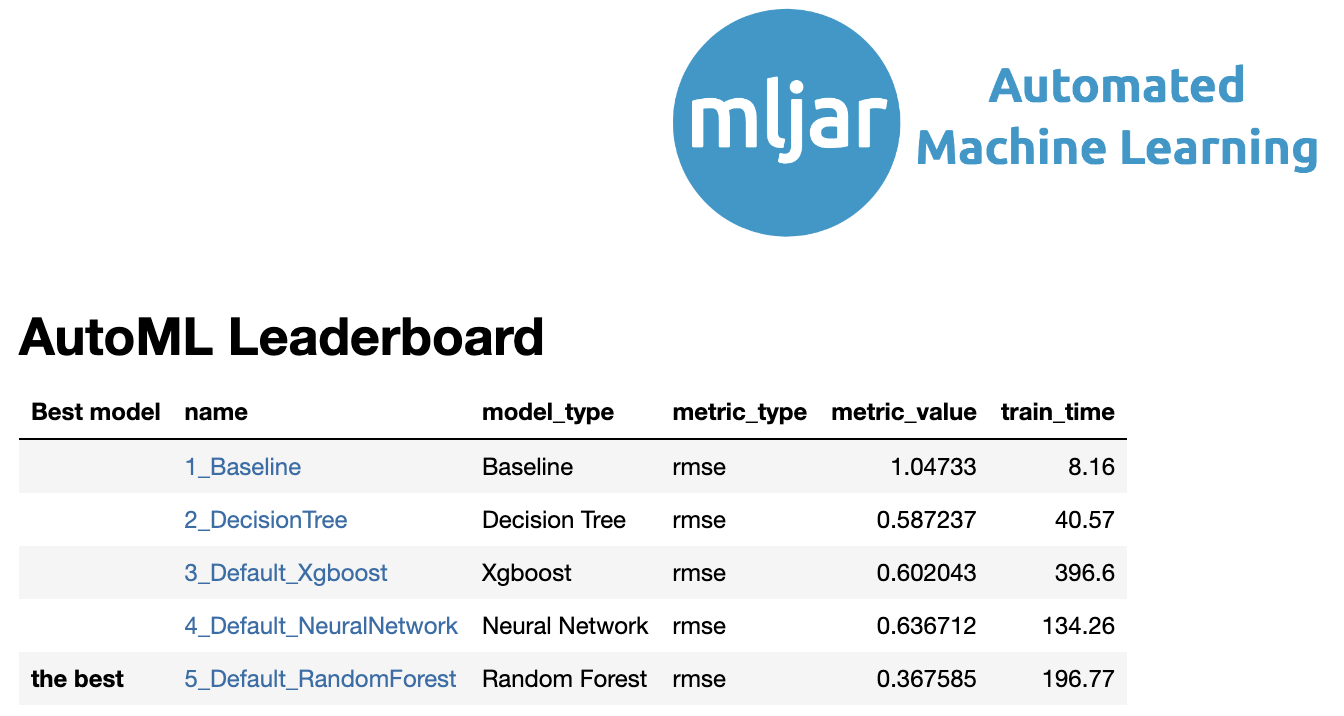
This part will be explained at the next phrase, **Auto Machine Learning**.

1. **Insight Generation**: Translating model outputs into clear insights and actionable recommendations.

Details will be kindly expanded in our presentation. In our presentation, we'll be taking a closer look at 'Insight Generation' and how to transform complex model outputs into clear, actionable insights, all through the lens of a seasoned sales expert.

**Auto Machine Learning**:

1. Data preparation (feature selection, feature encoding, normalization, ):
2. Normalization and data splitting:
3. Model training, model selection and reports:



1. Model training reports:

Model Evaluation Metrics:

----------------------------

Mean Absolute Error (MAE): 0.0994

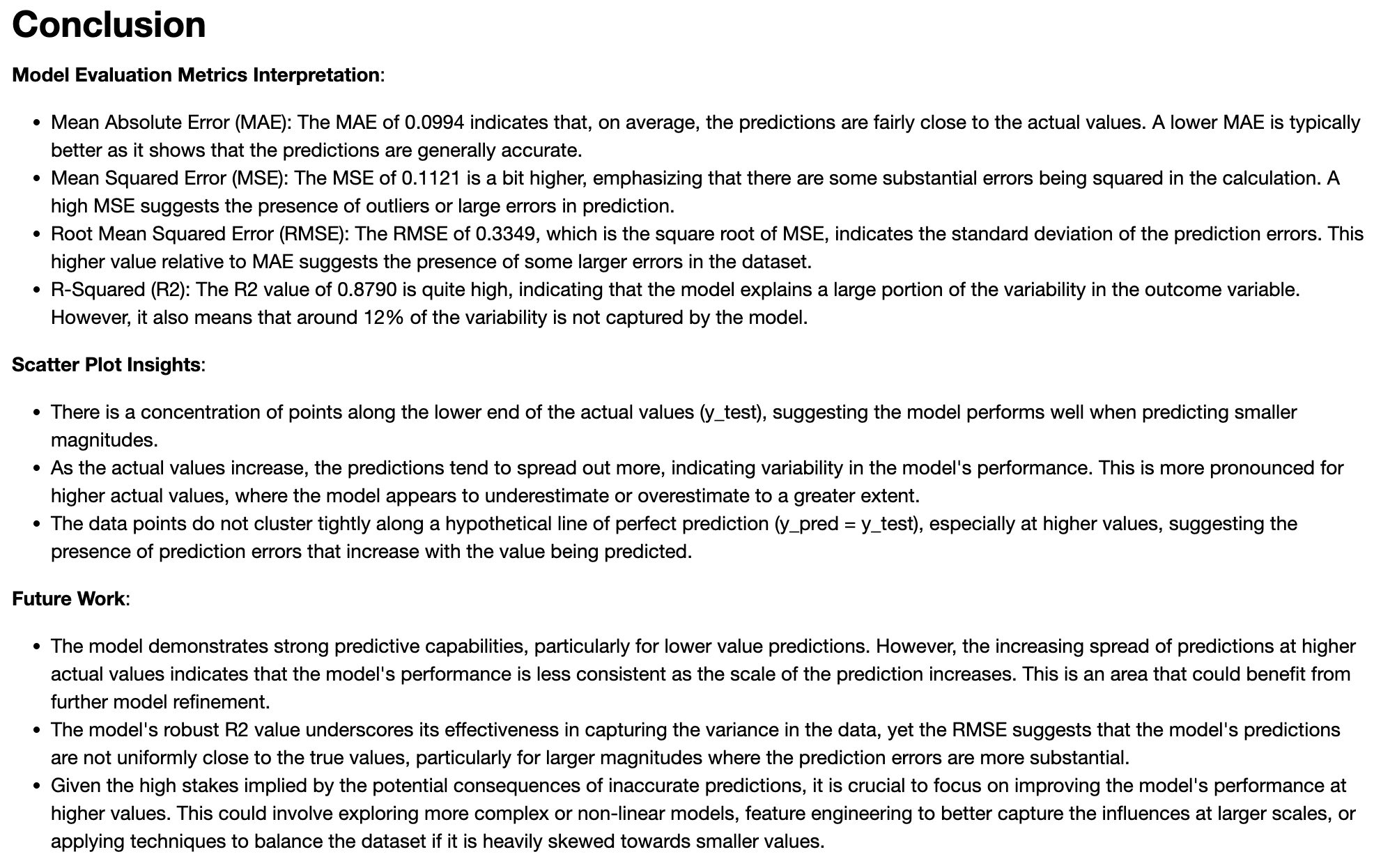
Mean Squared Error (MSE): 0.1121

Root Mean Squared Error (RMSE): 0.3349

R-Squared (R2): 0.8790

----------------------------

1. Conclusions:



**Part 3: Power BI - Data visualization (Consulting Service)**

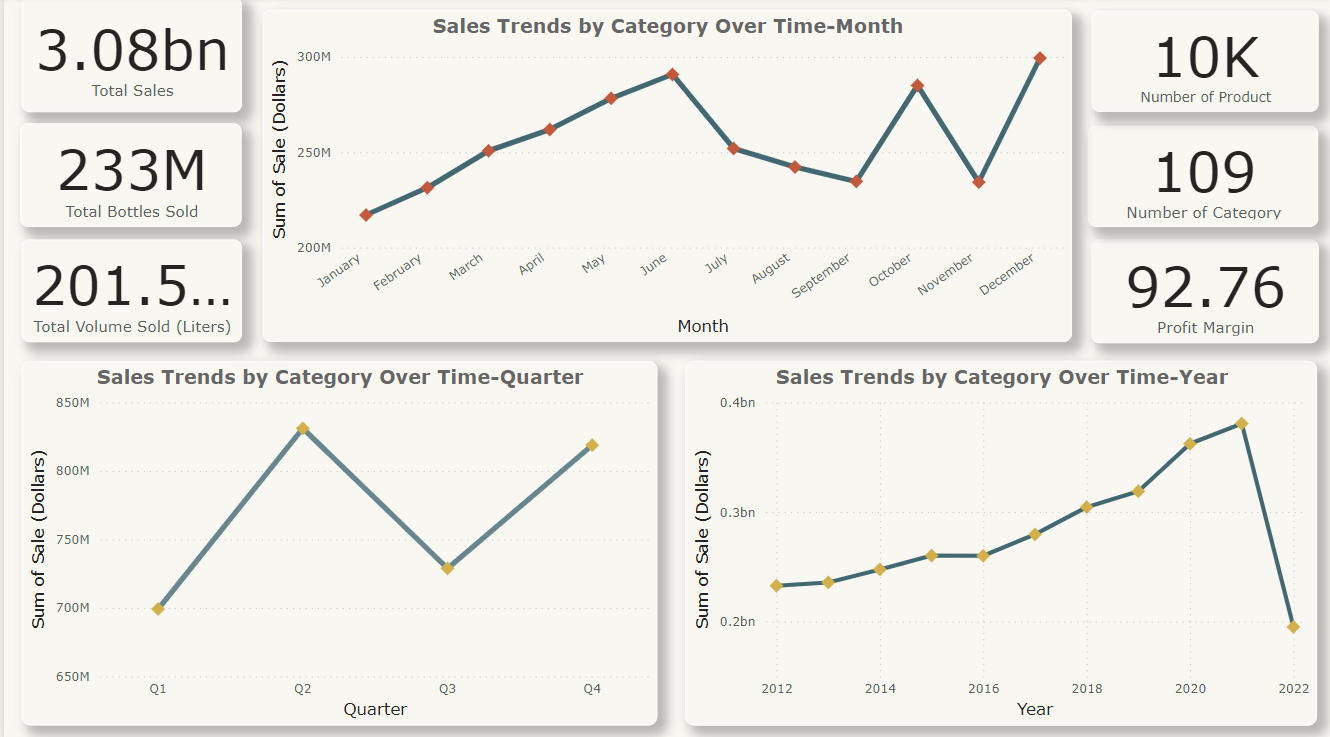
In this section, we utilize the power of Power BI to create interactive visualization charts. Our diverse clientele, which includes vendors, retail stores, and breweries, benefits immensely from these visualizations. By transforming their raw data into visually appealing and easy-to-understand charts, we empower them to gain deeper insights into their operations and make data-driven decisions. Our goal is to make data analysis an integral part of their strategy and daily operations.

• Data Models: Relationship between tables

• Dashboard Layout:

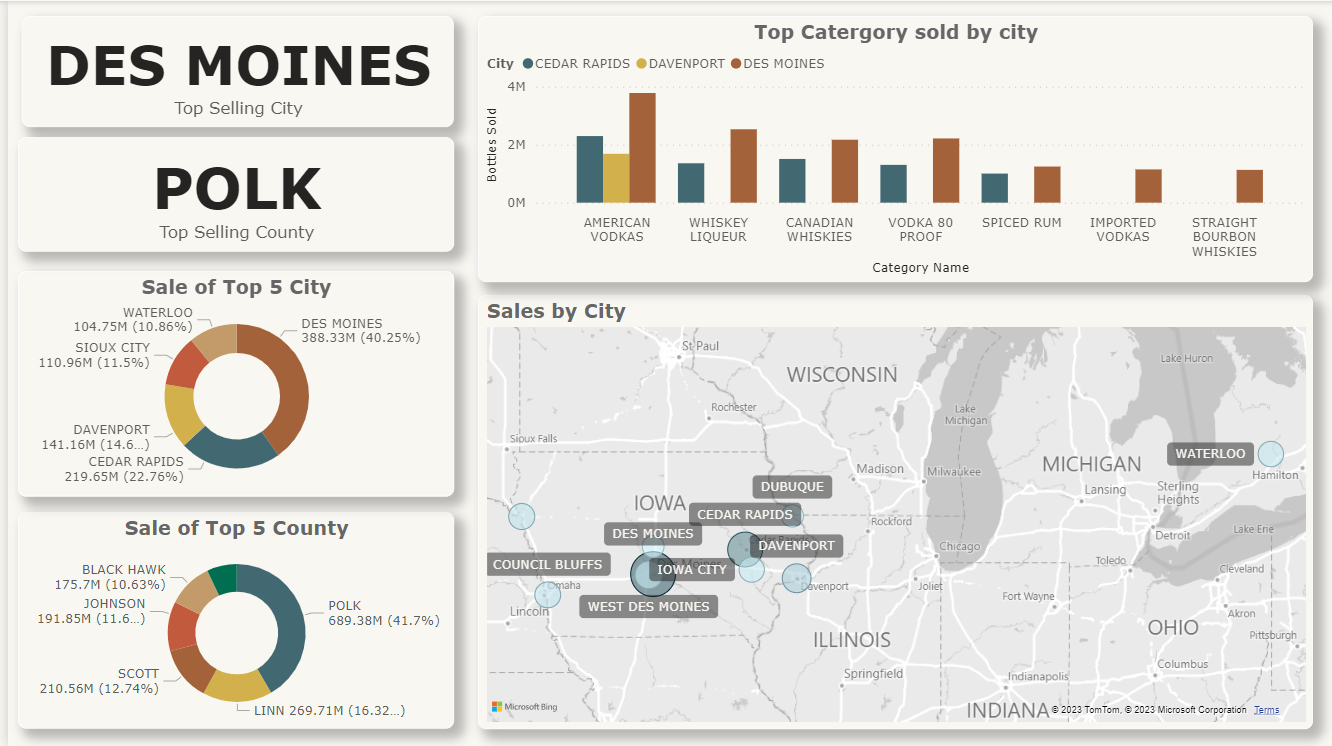
This dashboard has a total number of three pages. The key visualizations in this project are:

The total sales, total number of alcohol bottles sold over time

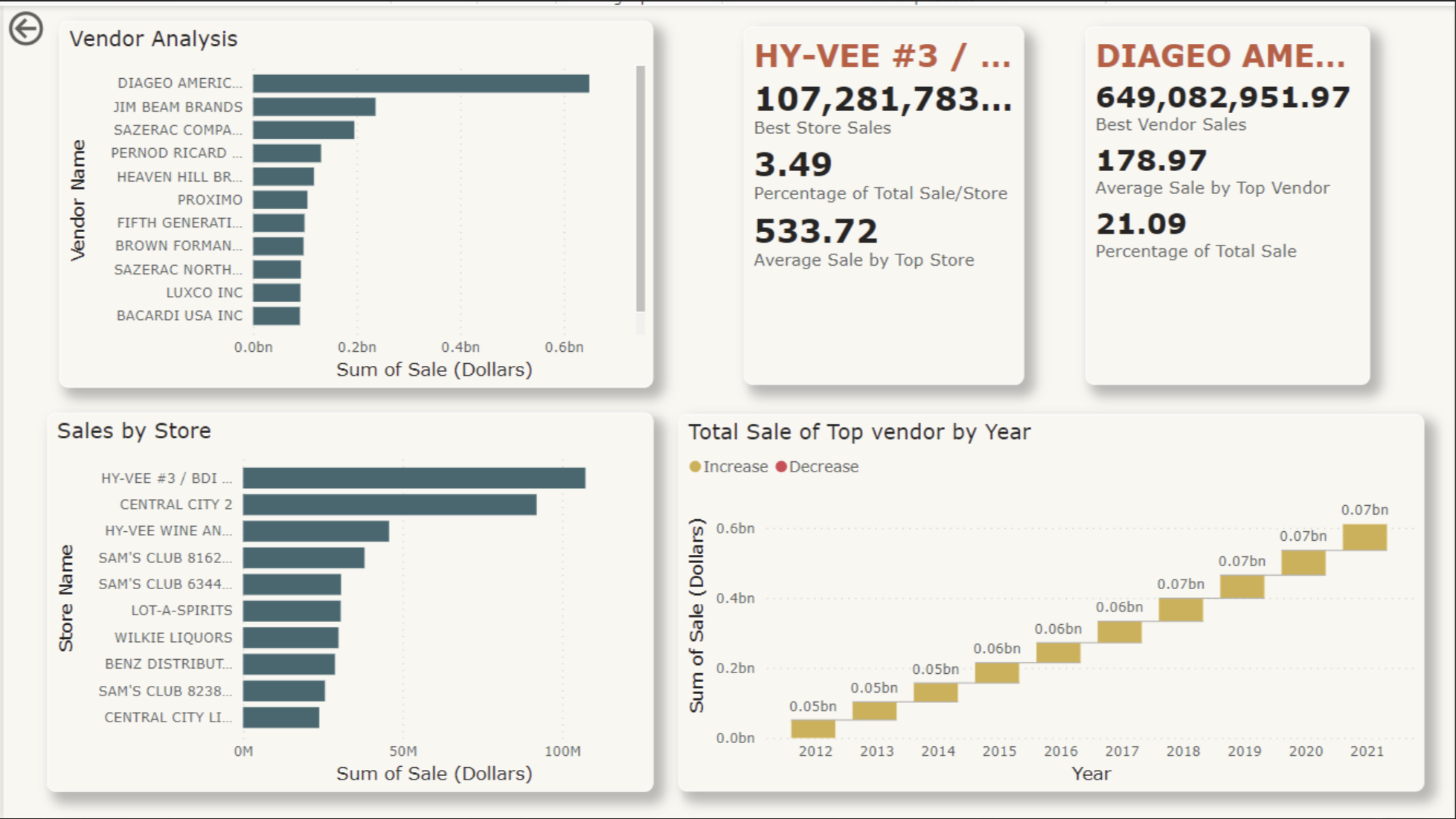


The chart illustrates Iowa’s liquor sales data, highlighting the sales trends by category over time and the number of product categories. It shows the total sales in billions of dollars and the profit margin in percentage. The chart also depicts sales distribution across different Iowa cities, suggesting consumption patterns and brand preferences.

The top 5 cities, top 5 countries with the highest number of bottles sold



The chart visualizes liquor sales data, highlighting Des Moines as the city with the highest sales and Polk as the corresponding county, suggesting a high volume of liquor purchases there. It also indicates the most popular liquor categories and showcases sales distribution across different Iowa cities, which could suggest consumption patterns and brand preferences.  
  
Vendor Analysis, Sales by store Analysis



The chart presents a detailed sales analysis of different vendors within a year. It highlights the vendor analysis, sales by sum of sale (dollars), and the total sale of the top vendor by year. This data provides insights into the performance and market dominance of various vendors, aiding in strategic decision-making.

**Conclusion**

This TDD has outlined the approach our team takes to leverage data mining techniques for the benefit of our clients. With our expertise in analyzing liquor sales data, we are poised to help retailers and investors make informed decisions that positively impact their bottom line.

**Value Proposition for Clients**

* **Retailers**: We provide precise demand forecasting to optimize inventory levels, reducing overstocking or understocking issues.
* **Business Insights**: Transforming raw sales data into strategic insights that uncover new opportunities and streamline operations.
* **Data-Driven Decision Making**: Shifting from intuition-based to analytics-based strategies for competitive advantage.
* **Cost Savings & Profit Growth**: Identifying inefficiencies and market trends to save costs and expand profit margins.