Customer Data Management and Analysis using SSMS, Python, and Azure

By: Abdelrhman, Ziad, Kareem, Mona, Mohamed, and Ehab
24/10/2024

Project Idea:

Problem Statement: Managing and analyzing customer data effectively to gain insights and support business decisions.

Solution Overview: Built a robust data engineering pipeline incorporating data warehousing, SQL Server database management, and Python for data analysis. Developed a dashboard for visual insights and a predictive model for customer purchasing behavior.

Unique Value Proposition: Integrates multiple technologies (SSMS, Python, Power BI, Azure) to streamline data management, enhance data accessibility, and provide actionable insights.

Project Wireframe:

The dashboard in Power BI offers an intuitive interface for analyzing customer data trends, sales, and demographic patterns.

End Users + Features:

Primary Users:

- Business Analysts and Decision Makers in Retail/Customer Service sectors.

Key Features:

- **Data Management**: Centralized SQL database for efficient data storage and retrieval.
- **Data Visualization**: Power BI dashboard for visual insights into customer trends and sales performance.
- **Predictive Modeling**: Decision Tree Classifier to predict potential purchases.

Value for End Users:

- Simplifies access to crucial data, improves decision-making, and helps anticipate customer needs based on purchasing behavior.

Data Structure:

Database Architecture: Relational database using SQL Server, incorporating key entities such as Customer, Product, Store, Transactions, and Territories.

Data Relationships: Designed with a schema that facilitates efficient querying and data analysis.

Data Flow:

- Data stored in SQL Server, transformed via SSIS, accessed by Python for model training and analysis.
- CSV files were used initially for data merging and model development, with features selected and engineered for model accuracy.

Programming Languages, Frameworks, and Cloud Services:

Programming Languages:

- **Python**: Used for data processing and modeling (Decision Tree model).
- **SQL**: Used for data queries and database management.

Frameworks and Libraries:

- **SQLAlchemy**: Connects SSMS to Python for seamless data flow.
- **Scikit-learn**, Matplotlib, and Seaborn: Libraries for data analysis, visualization, and machine learning.

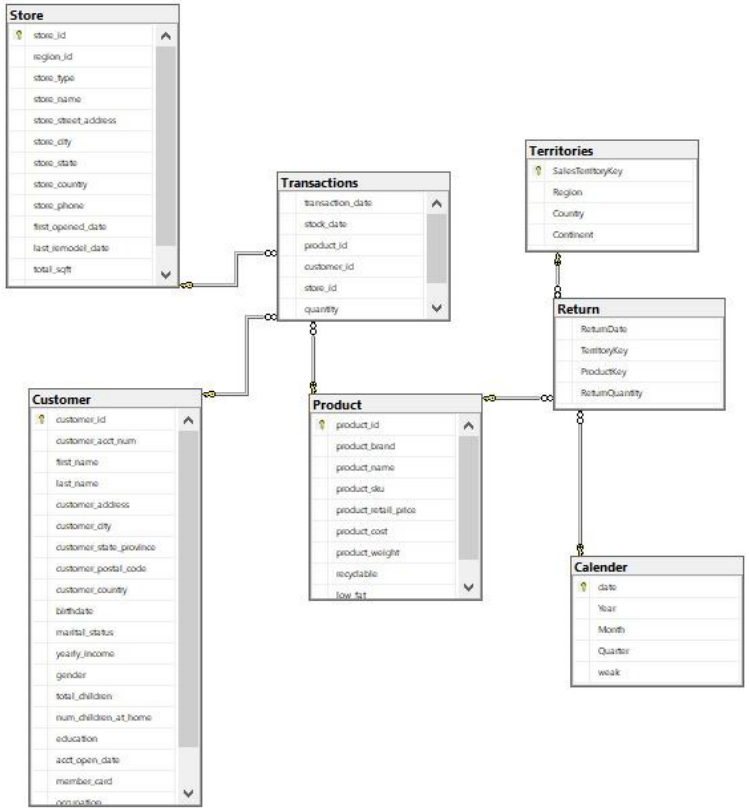
Cloud Services:

- **Azure Data Factory**: ETL tool used for data ingestion from on-prem SQL Server to Azure.
- **Azure Data Lake Gen2**: Cloud storage solution for all project data.
- **Azure Databricks**: Used for data transformation and preprocessing.
- **Azure Synapse Analytics**: Centralized analytics service to load transformed data.
- **Power BI**: Visualization tool connected to Azure Synapse for report creation.

Deliverables:

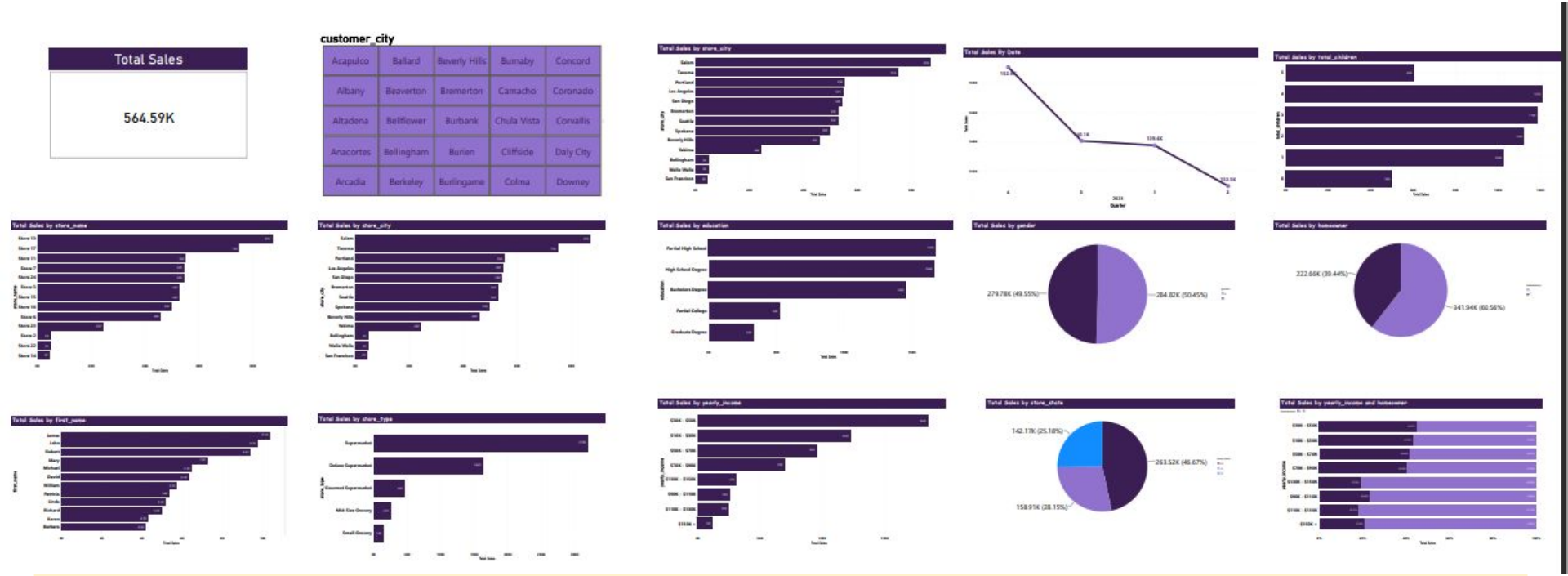
Reports and Documentation

- **Technical Specifications**: Detailed documentation on the project's technical setup, including database schema, data pipeline, and ETL process.
- **Data Exploration Report**: Insights from data exploration, covering key findings, data distributions, and patterns.



Dashboard

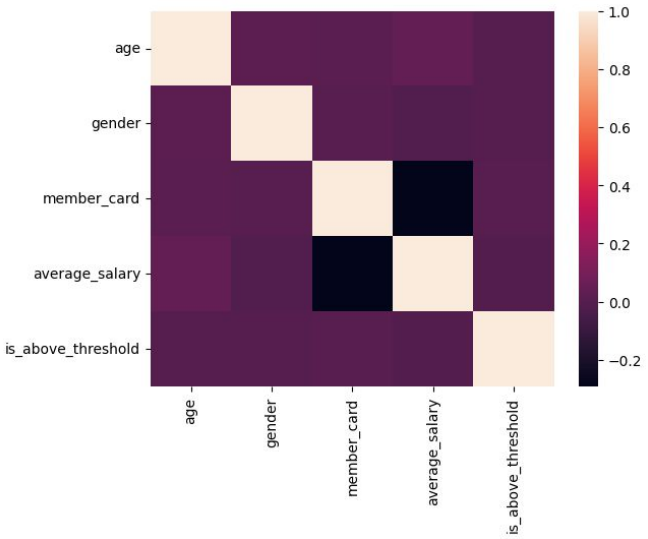
- **Power BI Dashboard**: Pulls data from Azure Synapse Analytics to create visual reports for business decisions.



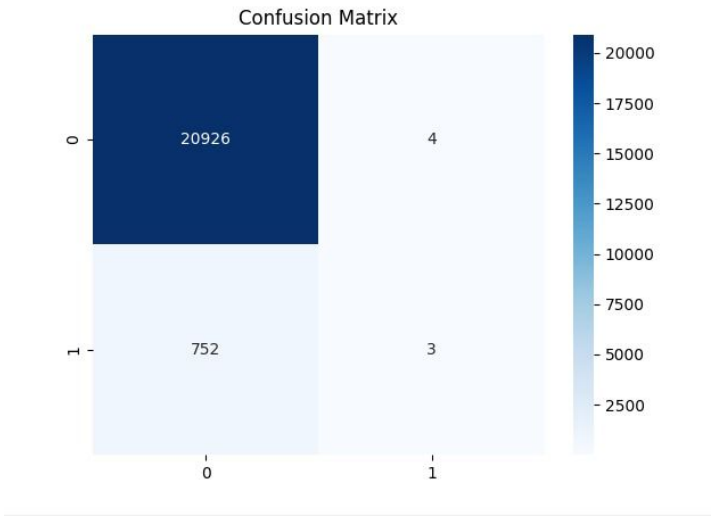
Model

- **Decision Tree Classification Model**: A model built in Python to predict customer purchase behavior, providing actionable insights on customer segmentation and potential purchasing trends.

Correlation matrix



Confusion matrix



Codebase and Documentation

- **Source Code Repository**: Organized code for all project stages, including data ingestion, transformation, model training, and visualization, with comments and documentation to ensure clarity.

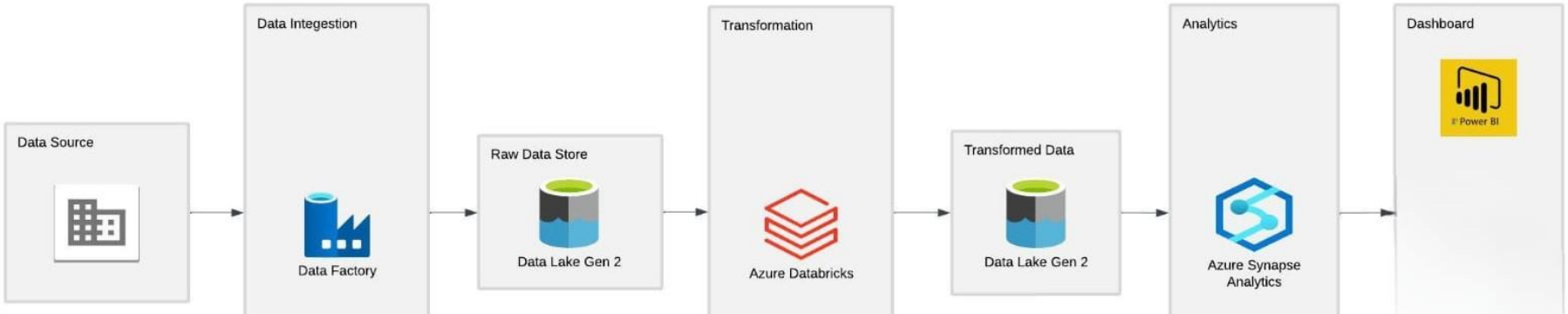
```
# Split data into dependent/independent variables
X = merged_df.iloc[:, :-1].values
y = merged_df.iloc[:, -1].values

# Split data into test/train set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state =
True)

# Correlation matrix
merged_df.corr()
sns.heatmap(merged_df.corr())
.
.
.
# Predict purchase with Age(45) and Salary(97000)
print(classifier.predict(sc.transform([[60, 1, 2, 100000]])))
```

Codebase and Cloud Infrastructure

- **Source Code Repository**: All project code, including ETL scripts from Azure Data Factory, Databricks notebooks, and Power BI reports.
- **Cloud Architecture Documentation**: Diagrams and descriptions of Azure services used, including Data Factory, Data Lake, Synapse Analytics, and Power BI, to ensure an end-to-end data solution.



List of our Team Members:

- Abdelrhman Faheem
- Ziad Badr
- Kareem Ragab
- Mohamed Ali
- Mona Yousef
- Ehab