### Repository Name

`ecommerce-analytics-solution`

### Directory Structure and Files

1. \*\*Create the repository structure:\*\*

```plaintext

ecommerce-analytics-solution/

│

├── data/

│ ├── raw/

│ │ ├── sales\_data.csv

│ │ ├── orders\_data.csv

│ │ ├── customer\_data.csv

│ │ └── churn\_data.csv

│ └── processed/

│ ├── sales\_data\_cleaned.csv

│ ├── orders\_data\_cleaned.csv

│ ├── customer\_data\_cleaned.csv

│ └── churn\_data\_cleaned.csv

│

├── etl/

│ ├── extract.py

│ ├── transform.py

│ └── load.py

│

├── analysis/

│ ├── sales\_analysis.ipynb

│ ├── orders\_analysis.ipynb

│ ├── customer\_analysis.ipynb

│ └── churn\_analysis.ipynb

│

├── dashboards/

│ ├── sales\_dashboard.twb

│ ├── orders\_dashboard.twb

│ ├── customer\_dashboard.twb

│ └── churn\_dashboard.twb

│

├── docs/

│ ├── requirements.md

│ ├── design.md

│ └── user\_guide.md

│

├── scripts/

│ ├── generate\_sample\_data.py

│ ├── start\_etl.sh

│ └── setup\_environment.sh

│

├── tests/

│ ├── test\_extract.py

│ ├── test\_transform.py

│ └── test\_load.py

│

├── .gitignore

├── README.md

└── requirements.txt

```

2. \*\*README.md\*\*

```markdown

# E-commerce Analytics Solution

This repository contains the code and documentation for providing a complete analytical solution for an e-commerce website, including customer churn prediction. The project includes building an enterprise data warehouse, performing data analysis, creating dashboards, and deploying machine learning models for churn prediction.

## Project Description

We worked with a startup e-commerce company to build a centralized system for managing and tracking their data. Previously, the data was stored in multiple formats such as Excel, CSV, and flat files. We provided an end-to-end solution from building an ETL pipeline to creating interactive dashboards and implementing customer churn prediction models.

## Data Analyst Role

As a Data Analyst, my responsibilities included:

- Understanding client requirements

- Preparing technical documentation

- Building an enterprise data warehouse

- Creating an ETL pipeline using SSIS

- Data preprocessing and analysis using Python

- Building dashboards in Tableau/PowerBI

- Implementing and deploying machine learning models for churn prediction

## Data Sources

- Sales data

- Orders data

- Finance data

- Customer data

- Products data

- Churn data

## Tools Used

- SSIS for ETL

- SQL Server for data storage

- Python for data preprocessing, analysis, and machine learning

- Tableau/PowerBI for dashboard creation

## Directory Structure

- `data/`: Contains raw and processed data files.

- `etl/`: Contains scripts for extracting, transforming, and loading data.

- `analysis/`: Contains Jupyter notebooks for data analysis and machine learning.

- `dashboards/`: Contains Tableau/PowerBI dashboard files.

- `docs/`: Contains project documentation files.

- `scripts/`: Contains utility scripts for data generation and environment setup.

- `tests/`: Contains unit tests for ETL processes.

- `README.md`: Project documentation.

- `requirements.txt`: List of Python dependencies.

## Setup

1. Clone the repository:

```bash

git clone https://github.com/yourusername/ecommerce-analytics-solution.git

```

2. Install the required Python packages:

```bash

pip install -r requirements.txt

```

3. Run the environment setup script:

```bash

./scripts/setup\_environment.sh

```

4. Generate sample data (optional):

```bash

python scripts/generate\_sample\_data.py

```

5. Update the ETL scripts in the `etl/` directory to connect to your data sources.

6. Run the ETL scripts to load data into the SQL Server:

```bash

./scripts/start\_etl.sh

```

7. Use the Jupyter notebooks in the `analysis/` directory to perform data analysis and machine learning.

8. Open the Tableau/PowerBI files in the `dashboards/` directory to view the interactive dashboards.

## Usage

- Run the ETL scripts to extract, transform, and load data into the SQL Server.

- Use the Jupyter notebooks for data analysis and machine learning.

- Open the Tableau/PowerBI dashboard files to interact with the visualizations.

## License

This project is licensed under the MIT License.

```

3. \*\*data/raw/sales\_data.csv\*\*

```csv

order\_id,product\_id,customer\_id,quantity,price,date

1,101,1001,2,20,2023-01-01

2,102,1002,1,10,2023-01-02

3,103,1003,5,50,2023-01-03

4,104,1004,3,30,2023-01-04

5,105,1005,4,40,2023-01-05

```

4. \*\*data/raw/orders\_data.csv\*\*

```csv

order\_id,customer\_id,order\_date,status

1,1001,2023-01-01,shipped

2,1002,2023-01-02,pending

3,1003,2023-01-03,delivered

4,1004,2023-01-04,cancelled

5,1005,2023-01-05,shipped

```

5. \*\*data/raw/customer\_data.csv\*\*

```csv

customer\_id,name,email,phone

1001,John Doe,john.doe@example.com,1234567890

1002,Jane Smith,jane.smith@example.com,0987654321

1003,Emily Davis,emily.davis@example.com,1122334455

1004,Michael Brown,michael.brown@example.com,2233445566

1005,Jessica Wilson,jessica.wilson@example.com,3344556677

```

6. \*\*data/raw/churn\_data.csv\*\*

```csv

customer\_id,active\_days,last\_purchase\_date,total\_spent,churn

1001,180,2023-06-01,1000,0

1002,90,2023-07-15,500,1

1003,120,2023-06-30,750,0

1004,45,2023-08-05,300,1

1005,200,2023-05-20,1100,0

```

7. \*\*etl/extract.py\*\*

```python

import pandas as pd

def extract\_data():

sales\_data = pd.read\_csv('../data/raw/sales\_data.csv')

orders\_data = pd.read\_csv('../data/raw/orders\_data.csv')

customer\_data = pd.read\_csv('../data/raw/customer\_data.csv')

churn\_data = pd.read\_csv('../data/raw/churn\_data.csv')

return sales\_data, orders\_data, customer\_data, churn\_data

if \_\_name\_\_ == "\_\_main\_\_":

sales\_data, orders\_data, customer\_data, churn\_data = extract\_data()

print(sales\_data.head())

print(orders\_data.head())

print(customer\_data.head())

print(churn\_data.head())

```

8. \*\*etl/transform.py\*\*

```python

import pandas as pd

def transform\_data(sales\_data, orders\_data, customer\_data, churn\_data):

# Example transformation: Remove rows with missing values

sales\_data\_cleaned = sales\_data.dropna()

orders\_data\_cleaned = orders\_data.dropna()

customer\_data\_cleaned = customer\_data.dropna()

churn\_data\_cleaned = churn\_data.dropna()

# Save cleaned data

sales\_data\_cleaned.to\_csv('../data/processed/sales\_data\_cleaned.csv', index=False)

orders\_data\_cleaned.to\_csv('../data/processed/orders\_data\_cleaned.csv', index=False)

customer\_data\_cleaned.to\_csv('../data/processed/customer\_data\_cleaned.csv', index=False)

churn\_data\_cleaned.to\_csv('../data/processed/churn\_data\_cleaned.csv', index=False)

if \_\_name\_\_ == "\_\_main\_\_":

sales\_data = pd.read\_csv('../data/raw/sales\_data.csv')

orders\_data = pd.read\_csv('../data/raw/orders\_data.csv')

customer\_data = pd.read\_csv('../data/raw/customer\_data.csv')

churn\_data = pd.read\_csv('../data/raw/churn\_data.csv')

transform\_data(sales\_data, orders\_data, customer\_data, churn\_data)

```

9. \*\*etl/load.py\*\*

```python

import pyodbc

import pandas as pd

def load\_data():

conn = pyodbc.connect('DRIVER={SQL Server};SERVER=your\_server\_name;DATABASE=your\_database\_name;UID=your\_username;PWD=your\_password')

cursor = conn.cursor()

# Load sales data

sales\_data = pd.read\_csv('../data/processed/sales\_data\_cleaned.csv')

for index, row in sales\_data.iterrows():

cursor.execute("INSERT INTO Sales (order\_id, product\_id, customer\_id, quantity, price, date) VALUES (?, ?, ?, ?, ?, ?)",

row['order\_id'], row['product\_id'], row['customer\_id'], row['quantity'], row['price'], row['date'])

# Load orders data

orders\_data = pd.read\_csv('../data/processed/orders\_data\_cleaned.csv')

for index, row in orders\_data.iterrows():

cursor.execute("INSERT INTO Orders (order\_id, customer\_id, order\_date, status) VALUES (?, ?, ?, ?)",

row['order\_id'], row['customer\_id'], row['order\_date'], row['status'])

# Load customer data

customer\_data = pd.read\_csv('../data/processed/customer\_data\_cleaned.csv')

for index, row in customer\_data.iterrows():

cursor.execute("INSERT INTO Customers (customer\_id, name, email, phone) VALUES (?, ?, ?, ?)",

row['customer\_id'], row['name'], row['email'], row['phone'])

# Load churn data

churn\_data = pd.read\_csv('../data/processed/churn\_data\_cleaned.csv')

for index, row in churn\_data.iterrows():

cursor.execute("INSERT INTO Churn (customer\_id, active\_days, last\_purchase\_date, total\_spent, churn) VALUES (?, ?, ?, ?, ?)",

row['customer\_id'], row['active\_days'], row['last\_purchase\_date'], row['total\_spent'], row['churn'])

conn.commit()

conn.close()

if \_\_name\_\_ == "\_\_main\_\_":

load\_data()

```

10. \*\*analysis/sales\_analysis.ipynb\*\*

```python

import pandas as pd

import matplotlib.pyplot as plt

# Load cleaned sales data

sales\_data = pd.read\_csv('../data/processed/sales\_data\_cleaned.csv')

# Perform analysis

sales\_summary = sales\_data.groupby('product\_id').agg({'quantity': 'sum', 'price': 'sum'}).reset\_index()

# Plot sales summary

plt.figure(figsize=(10, 5))

plt.bar(sales\_summary['product\_id'], sales\_summary['price'])

plt.xlabel('Product ID')

plt.ylabel('Total Sales')

plt.title('Total Sales by Product')

plt.show()

```

11. \*\*analysis/orders\_analysis.ipynb\*\*

```python

import pandas as pd

import matplotlib.pyplot as plt

# Load cleaned orders data

orders\_data = pd.read\_csv('../data/processed/orders\_data\_cleaned.csv')

# Perform analysis

order\_status\_summary = orders\_data['status'].value\_counts().reset\_index()

order\_status\_summary.columns = ['status', 'count']

# Plot order status summary

plt.figure(figsize=(10, 5))

plt.bar(order\_status\_summary['status'], order\_status\_summary['count'])

plt.xlabel('Order Status')

plt.ylabel('Count')

plt.title('Order Status Distribution')

plt.show()

```

12. \*\*analysis/customer\_analysis.ipynb\*\*

```python

import pandas as pd

import matplotlib.pyplot as plt

# Load cleaned customer data

customer\_data = pd.read\_csv('../data/processed/customer\_data\_cleaned.csv')

# Perform analysis

customer\_summary = customer\_data['customer\_id'].value\_counts().reset\_index()

customer\_summary.columns = ['customer\_id', 'count']

# Plot customer summary

plt.figure(figsize=(10, 5))

plt.bar(customer\_summary['customer\_id'], customer\_summary['count'])

plt.xlabel('Customer ID')

plt.ylabel('Count')

plt.title('Customer Distribution')

plt.show()

```

13. \*\*analysis/churn\_analysis.ipynb\*\*

```python

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

# Load cleaned churn data

churn\_data = pd.read\_csv('../data/processed/churn\_data\_cleaned.csv')

# Feature engineering

X = churn\_data[['active\_days', 'last\_purchase\_date', 'total\_spent']]

y = churn\_data['churn']

# Split the data

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Train a Random Forest model

model = RandomForestClassifier(n\_estimators=100, random\_state=42)

model.fit(X\_train, y\_train)

# Make predictions

y\_pred = model.predict(X\_test)

# Evaluate the model

accuracy = accuracy\_score(y\_test, y\_pred)

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

class\_report = classification\_report(y\_test, y\_pred)

print(f'Accuracy: {accuracy}')

print('Confusion Matrix:')

print(conf\_matrix)

print('Classification Report:')

print(class\_report)

```

14. \*\*requirements.txt\*\*

```

pandas

numpy

matplotlib

plotly

sqlalchemy

pyodbc

scikit-learn

jupyter

```

15. \*\*scripts/generate\_sample\_data.py\*\*

```python

import pandas as pd

import numpy as np

# Generate sample sales data

sales\_data = pd.DataFrame({

'order\_id': range(1, 101),

'product\_id': np.random.randint(100, 200, size=100),

'customer\_id': np.random.randint(1000, 1100, size=100),

'quantity': np.random.randint(1, 10, size=100),

'price': np.random.randint(10, 100, size=100),

'date': pd.date\_range(start='2023-01-01', periods=100, freq='D')

})

# Generate sample orders data

orders\_data = pd.DataFrame({

'order\_id': range(1, 101),

'customer\_id': np.random.randint(1000, 1100, size=100),

'order\_date': pd.date\_range(start='2023-01-01', periods=100, freq='D'),

'status': np.random.choice(['shipped', 'pending', 'delivered', 'cancelled'], size=100)

})

# Generate sample customer data

customer\_data = pd.DataFrame({

'customer\_id': range(1000, 1100),

'name': [f'Customer {i}' for i in range(1000, 1100)],

'email': [f'customer{i}@example.com' for i in range(1000, 1100)],

'phone': [f'123-456-789{i % 10}' for i in range(1000, 1100)]

})

# Generate sample churn data

churn\_data = pd.DataFrame({

'customer\_id': range(1000, 1100),

'active\_days': np.random.randint(30, 200, size=100),

'last\_purchase\_date': pd.date\_range(start='2023-01-01', periods=100, freq='D'),

'total\_spent': np.random.randint(100, 2000, size=100),

'churn': np.random.choice([0, 1], size=100)

})

# Save to CSV

sales\_data.to\_csv('../data/raw/sales\_data.csv', index=False)

orders\_data.to\_csv('../data/raw/orders\_data.csv', index=False)

customer\_data.to\_csv('../data/raw/customer\_data.csv', index=False)

churn\_data.to\_csv('../data/raw/churn\_data.csv', index=False)

print("Sample data generated and saved.")

```

16. \*\*scripts/start\_etl.sh\*\*

```bash

#!/bin/bash

echo "Starting ETL process..."

python ../etl/extract.py

python ../etl/transform.py

python ../etl/load.py

echo "ETL process completed."

```

17. \*\*scripts/setup\_environment.sh\*\*

```bash

#!/bin/bash

echo "Setting up the environment..."

# Install Python dependencies

pip install -r ../requirements.txt

# Additional setup commands can go here

echo "Environment setup completed."

```

18. \*\*tests/test\_extract.py\*\*

```python

import unittest

import pandas as pd

from etl.extract import extract\_data

class TestExtract(unittest.TestCase):

def test\_extract\_data(self):

sales\_data, orders\_data, customer\_data, churn\_data = extract\_data()

self.assertIsInstance(sales\_data, pd.DataFrame)

self.assertIsInstance(orders\_data, pd.DataFrame)

self.assertIsInstance(customer\_data, pd.DataFrame)

self.assertIsInstance(churn\_data, pd.DataFrame)

if \_\_name\_\_ == '\_\_main\_\_':

unittest.main()

```

19. \*\*tests/test\_transform.py\*\*

```python

import unittest

import pandas as pd

from etl.transform import transform\_data

class TestTransform(unittest.TestCase):

def test\_transform\_data(self):

sales\_data = pd.read\_csv('../data/raw/sales\_data.csv')

orders\_data = pd.read\_csv('../data/raw/orders\_data.csv')

customer\_data = pd.read\_csv('../data/raw/customer\_data.csv')

churn\_data = pd.read\_csv('../data/raw/churn\_data.csv')

transform\_data(sales\_data, orders\_data, customer\_data, churn\_data)

sales\_data\_cleaned = pd.read\_csv('../data/processed/sales\_data\_cleaned.csv')

orders\_data\_cleaned = pd.read\_csv('../data/processed/orders\_data\_cleaned.csv')

customer\_data\_cleaned = pd.read\_csv('../data/processed/customer\_data\_cleaned.csv')

churn\_data\_cleaned = pd.read\_csv('../data/processed/churn\_data\_cleaned.csv')

self.assertFalse(sales\_data\_cleaned.empty)

self.assertFalse(orders\_data\_cleaned.empty)

self.assertFalse(customer\_data\_cleaned.empty)

self.assertFalse(churn\_data\_cleaned.empty)

if \_\_name\_\_ == '\_\_main\_\_':

unittest.main()

```

20. \*\*tests/test\_load.py\*\*

```python

import unittest

import pyodbc

from etl.load import load\_data

class TestLoad(unittest.TestCase):

def test\_load\_data(self):

# Connect to the database

conn = pyodbc.connect('DRIVER={SQL Server};SERVER=your

\_server\_name;DATABASE=your\_database\_name;UID=your\_username;PWD=your\_password')

cursor = conn.cursor()

# Load data

load\_data()

# Check if data is loaded into the database

cursor.execute("SELECT COUNT(\*) FROM Sales")

sales\_count = cursor.fetchone()[0]

cursor.execute("SELECT COUNT(\*) FROM Orders")

orders\_count = cursor.fetchone()[0]

cursor.execute("SELECT COUNT(\*) FROM Customers")

customers\_count = cursor.fetchone()[0]

cursor.execute("SELECT COUNT(\*) FROM Churn")

churn\_count = cursor.fetchone()[0]

self.assertGreater(sales\_count, 0)

self.assertGreater(orders\_count, 0)

self.assertGreater(customers\_count, 0)

self.assertGreater(churn\_count, 0)

conn.close()

if \_\_name\_\_ == '\_\_main\_\_':

unittest.main()

```

21. \*\*.gitignore\*\*

```

# Byte-compiled / optimized / DLL files

\_\_pycache\_\_/

\*.py[cod]

\*$py.class

# C extensions

\*.so

# Packages

\*.egg

\*.egg-info/

dist/

build/

eggs/

parts/

bin/

var/

sdist/

develop-eggs/

.installed.cfg

lib/

lib64/

\_\_pypackages\_\_/

# Installer logs

pip-log.txt

pip-delete-this-directory.txt

# Unit test / coverage reports

htmlcov/

.tox/

.nox/

.coverage

coverage.\*

.cache

nosetests.xml

coverage.xml

\*.cover

.hypothesis/

.pytest\_cache/

# Jupyter Notebook

.ipynb\_checkpoints

# Environments

.env

.venv

env/

venv/

ENV/

env.bak/

venv.bak/

# PyCharm

.idea/

# VSCode

.vscode/

```

22. \*\*docs/requirements.md\*\*

```markdown

# Project Requirements

This document outlines the requirements for the E-commerce Analytics Solution project.

## Functional Requirements

1. \*\*Data Extraction\*\*

- Extract data from multiple sources (CSV, Excel, databases).

- Handle different data formats and ensure data consistency.

2. \*\*Data Transformation\*\*

- Clean and preprocess data.

- Handle missing values, duplicates, and data anomalies.

- Perform necessary data transformations.

3. \*\*Data Loading\*\*

- Load data into a centralized SQL Server database.

- Ensure data integrity during the loading process.

4. \*\*Data Analysis\*\*

- Perform data analysis using Python.

- Generate reports and visualizations.

5. \*\*Dashboards\*\*

- Create interactive dashboards using Tableau/PowerBI.

- Provide insights into sales, orders, and customer data.

6. \*\*Churn Prediction\*\*

- Implement machine learning models to predict customer churn.

- Evaluate and monitor model performance.

## Non-Functional Requirements

1. \*\*Performance\*\*

- Ensure the ETL process is efficient and can handle large volumes of data.

- Optimize database queries for fast data retrieval.

2. \*\*Scalability\*\*

- Design the solution to accommodate future growth in data volume and complexity.

3. \*\*Security\*\*

- Implement data security measures to protect sensitive information.

- Ensure secure access to the database and dashboards.

4. \*\*Usability\*\*

- Provide user-friendly interfaces for data analysis and dashboard viewing.

- Ensure documentation is clear and comprehensive.

## Tools and Technologies

- SSIS for ETL

- SQL Server for data storage

- Python for data preprocessing, analysis, and machine learning

- Tableau/PowerBI for dashboard creation

```

23. \*\*docs/design.md\*\*

```markdown

# Design Document

This document outlines the design of the E-commerce Analytics Solution project.

## Architecture

The project follows a typical ETL (Extract, Transform, Load) architecture, where data is extracted from various sources, transformed to ensure consistency and quality, and then loaded into a centralized database for analysis and reporting.

## Data Flow

1. \*\*Data Extraction\*\*

- Extract data from CSV files (sales\_data.csv, orders\_data.csv, customer\_data.csv, churn\_data.csv).

- Use Python scripts for extraction.

2. \*\*Data Transformation\*\*

- Clean and preprocess the extracted data.

- Handle missing values, duplicates, and anomalies.

- Use Python scripts for transformation.

3. \*\*Data Loading\*\*

- Load the cleaned data into a SQL Server database.

- Use Python scripts with pyodbc for loading.

4. \*\*Data Analysis\*\*

- Perform data analysis using Jupyter notebooks and Python libraries (pandas, matplotlib).

- Generate insights and visualizations.

5. \*\*Dashboards\*\*

- Create interactive dashboards using Tableau/PowerBI.

- Provide visual insights into sales, orders, customer data, and churn prediction.

6. \*\*Churn Prediction\*\*

- Implement machine learning models to predict customer churn.

- Use Python libraries such as scikit-learn for model training and evaluation.

## Database Schema

### Sales Table

| Column Name | Data Type | Description |

|-------------|-----------|-------------|

| order\_id | INT | Unique identifier for each order |

| product\_id | INT | Unique identifier for each product |

| customer\_id | INT | Unique identifier for each customer |

| quantity | INT | Quantity of the product ordered |

| price | FLOAT | Price of the product |

| date | DATE | Date of the order |

### Orders Table

| Column Name | Data Type | Description |

|-------------|-----------|-------------|

| order\_id | INT | Unique identifier for each order |

| customer\_id | INT | Unique identifier for each customer |

| order\_date | DATE | Date of the order |

| status | VARCHAR | Status of the order (shipped, pending, delivered, cancelled) |

### Customers Table

| Column Name | Data Type | Description |

|-------------|-----------|-------------|

| customer\_id | INT | Unique identifier for each customer |

| name | VARCHAR | Name of the customer |

| email | VARCHAR | Email of the customer |

| phone | VARCHAR | Phone number of the customer |

### Churn Table

| Column Name | Data Type | Description |

|---------------------|-----------|--------------------------------------------|

| customer\_id | INT | Unique identifier for each customer |

| active\_days | INT | Number of active days |

| last\_purchase\_date | DATE | Date of the last purchase |

| total\_spent | FLOAT | Total amount spent by the customer |

| churn | INT | Churn status (0 = not churned, 1 = churned)|

## ETL Process

1. \*\*Extract\*\*

- Use `extract.py` to read data from CSV files and load them into pandas DataFrames.

2. \*\*Transform\*\*

- Use `transform.py` to clean and preprocess the data.

- Handle missing values, duplicates, and anomalies.

- Save the cleaned data into processed CSV files.

3. \*\*Load\*\*

- Use `load.py` to load the cleaned data into the SQL Server database.

- Insert data into Sales, Orders, Customers, and Churn tables.

## Dashboards

1. \*\*Sales Dashboard\*\*

- Visualize total sales, sales by product, and sales trends over time.

2. \*\*Orders Dashboard\*\*

- Visualize order status distribution, orders by date, and customer order history.

3. \*\*Customer Dashboard\*\*

- Visualize customer distribution, top customers, and customer order patterns.

4. \*\*Churn Dashboard\*\*

- Visualize churn rate, factors affecting churn, and churn predictions.

## Scripts

1. \*\*generate\_sample\_data.py\*\*

- Generates sample data for testing and development.

2. \*\*start\_etl.sh\*\*

- Starts the ETL process by running the extract, transform, and load scripts.

3. \*\*setup\_environment.sh\*\*

- Sets up the development environment by installing required dependencies.

## Testing

1. \*\*test\_extract.py\*\*

- Unit tests for the data extraction process.

2. \*\*test\_transform.py\*\*

- Unit tests for the data transformation process.

3. \*\*test\_load.py\*\*

- Unit tests for the data loading process.

```

24. \*\*docs/user\_guide.md\*\*

```markdown

# User Guide

This document provides instructions on how to use the E-commerce Analytics Solution project.

## Prerequisites

- Python 3.x installed on your system

- SQL Server installed and running

- Tableau/PowerBI installed for dashboard viewing

## Setup

1. Clone the repository:

```bash

git clone https://github.com/yourusername/ecommerce-analytics-solution.git

```

2. Navigate to the repository directory:

```bash

cd ecommerce-analytics-solution

```

3. Install the required Python packages:

```bash

pip install -r requirements.txt

```

4. Run the environment setup script:

```bash

./scripts/setup\_environment.sh

```

5. Generate sample data (optional):

```bash

python scripts/generate\_sample\_data.py

```

## Running the ETL Process

1. Update the ETL scripts in the `etl/` directory to connect to your data sources and database.

2. Run the ETL process:

```bash

./scripts/start\_etl.sh

```

This will extract data from the raw data files, transform and clean the data, and load it into the SQL Server database.

## Data Analysis

1. Open the Jupyter notebooks in the `analysis/` directory:

```bash

jupyter notebook analysis/sales\_analysis.ipynb

jupyter notebook analysis/orders\_analysis.ipynb

jupyter notebook analysis/customer\_analysis.ipynb

jupyter notebook analysis/churn\_analysis.ipynb

```

2

. Perform data analysis, machine learning, and generate reports and visualizations.

## Dashboards

1. Open the Tableau/PowerBI files in the `dashboards/` directory:

- sales\_dashboard.twb

- orders\_dashboard.twb

- customer\_dashboard.twb

- churn\_dashboard.twb

2. Interact with the dashboards to explore insights into sales, orders, customer data, and churn predictions``