FINAL PROJECT

FINANCIAL ENVIRONMENT DEVELOPMENT

Marcela Caetano

PROJECT GOAL

- Build a chatbot-driven financial environment to assist users in managing their finances effectively.
- Promote healthy financial habits by offering personalized, actionable guidance aligned with individual goals.
- Leverage insights from two financial books (The Richest Man in Babylon and The Intelligent Investor) to provide practical strategies for sustained wealth-building and financial success.

CHALLENGES & SOLUTIONS

Challenges

Determine the best interface for the financial environment, focusing on individual user experience, the overall user base, or the analysis of usage patterns.

Solutions:

Build a dashboard to analyze patterns across all users and focus the chatbot experience on personalized user engagement and needs.

DATASET OVERVIEW

- Generated a synthetic dataset using the Faker library.
- The Dataset contains 3.000 rows and 16 columns and includes user information, income, expenses, and financial allocations such as savings and investments.

METHODOLOGY

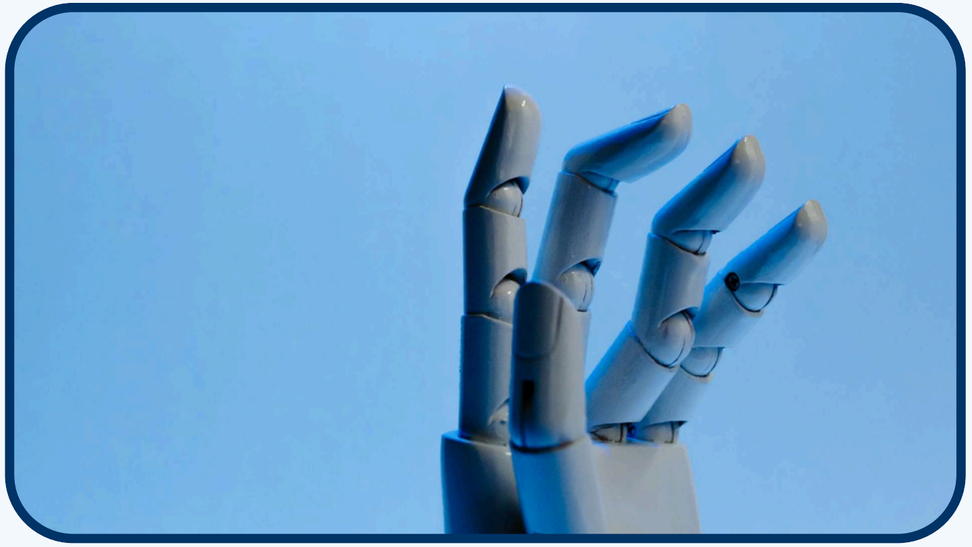
ETL

- Generate synthetic data for 250 users over 12 months.
- Load raw data into the MySQL database and retrieve it using an SQL query.
- Create a function to execute the SQL query and fetch the data.
- Explore the dataset by checking for nulls, duplicates, and data types.
- Clean the data by standardizing column names.

METHODOLOGY

EDA

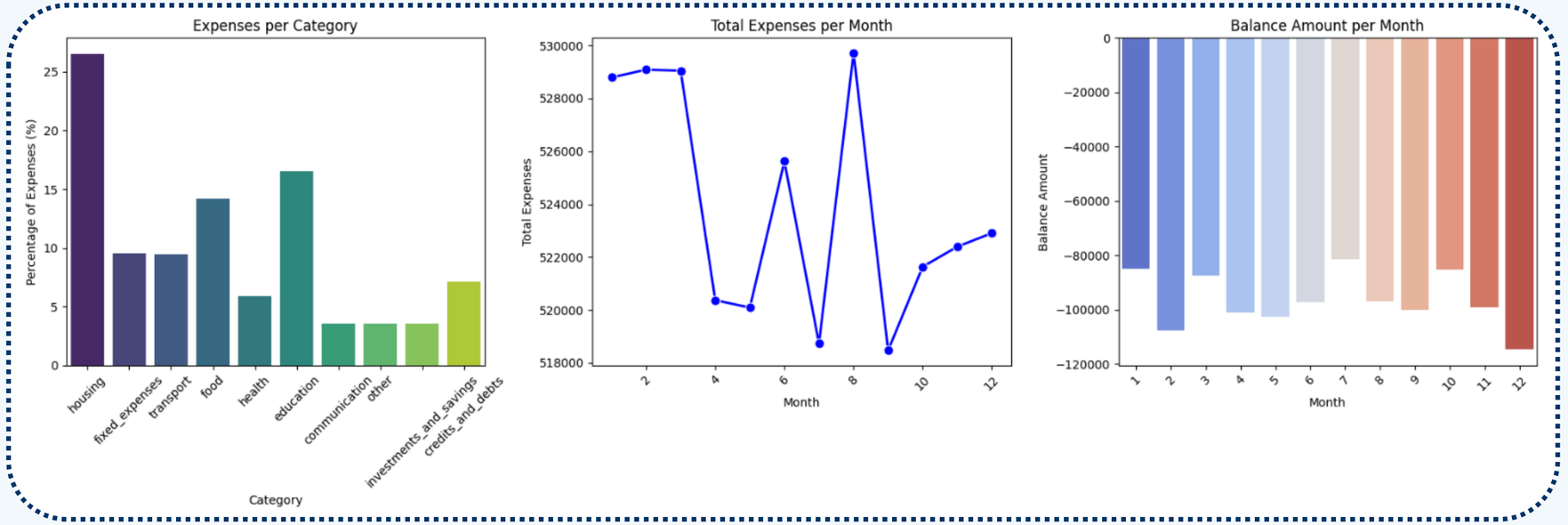
- Define a method to calculate key statistics like the average, median, and range of the data.
- Visualize the distribution of data with graphs for each numerical column.



METHODOLOGY

EDA

- Define a function to calculate and visualize expense categories, total monthly expenses, and balance data.
- Create and display three side-by-side charts: a bar chart for category contributions, a line chart for monthly expenses, and a bar chart for monthly balance.

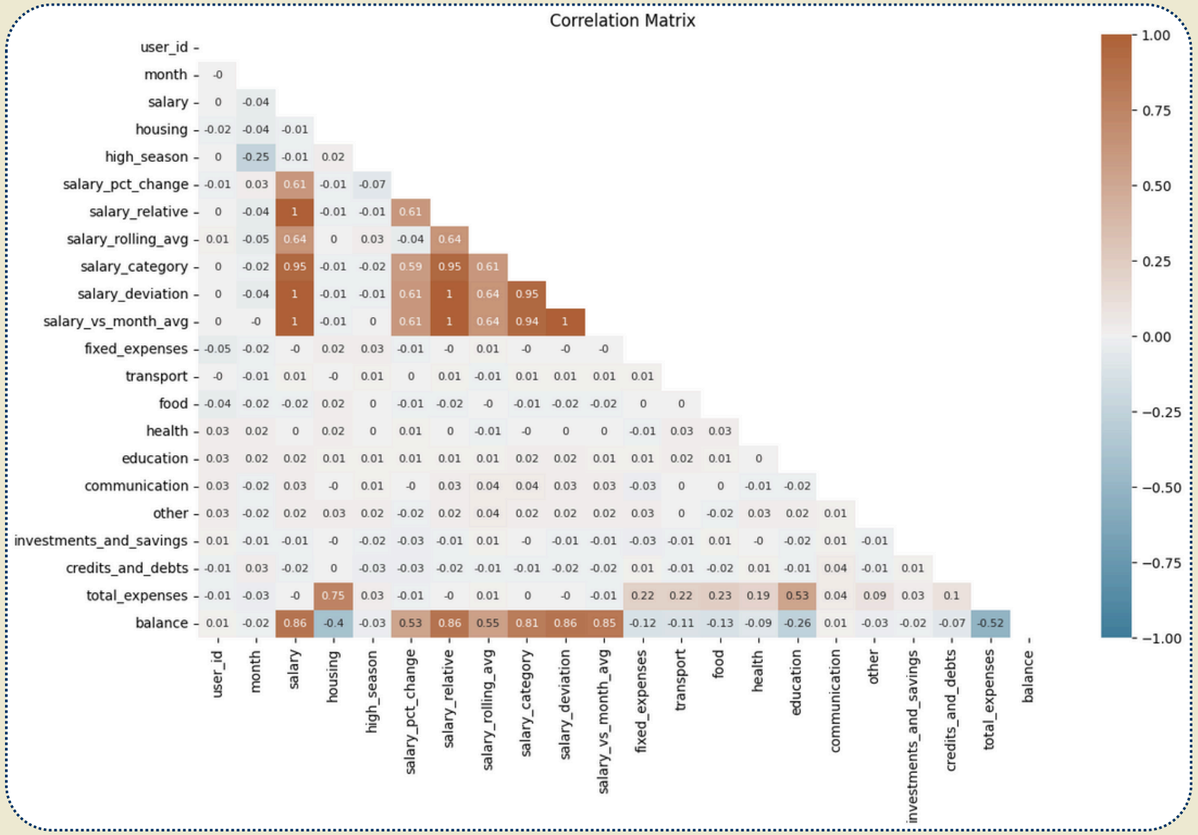


DATA PREPROCESSING &
FEATURE ENGINEERING

- Drop the column name and perform data engineering by creating new features.

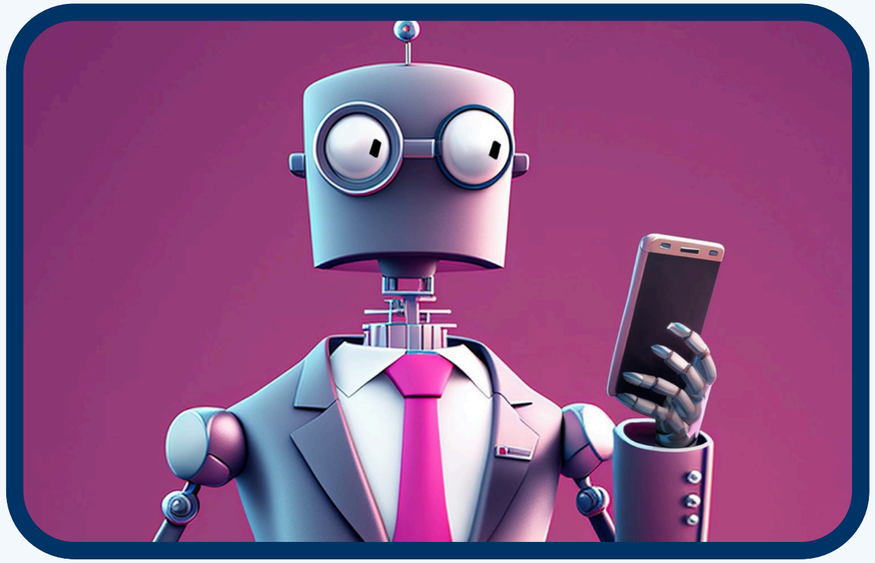
CORRELATION MATRIX

- Remove features with no correlation to the target and those showing high multicollinearity with the target or between each other.
- Keep only the relevant columns that contribute meaningfully to the model's predictions.



DATA MODELLING

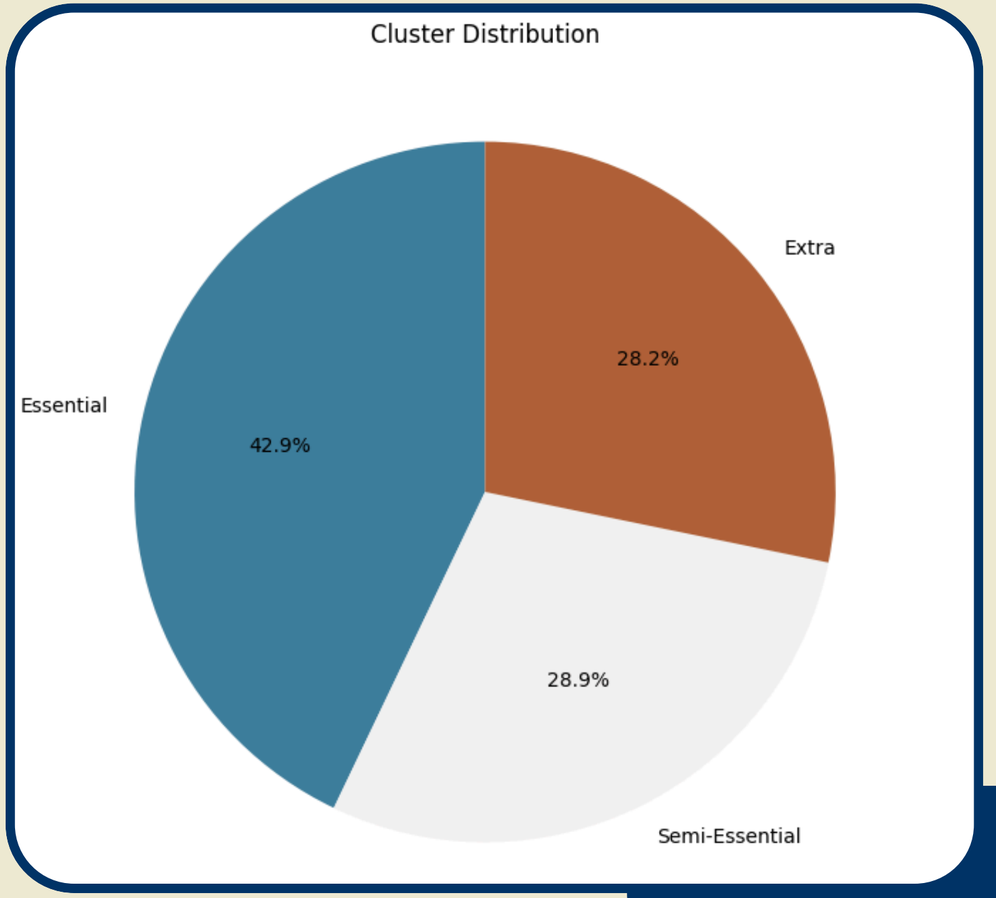
- Performed a 70/30 train-test split to divide the dataset into training and testing sets.
- Selected total expenses and balance as the target variables for prediction.
- Used K-Means for clustering to identify spending patterns.
- Applied Isolation Forest for anomaly detection to detect unusual financial behaviors.
- Implemented XGBoost and Random Forest to predict total expenses and balance based on user financial data.



MODEL SELECTION
K-MEANS

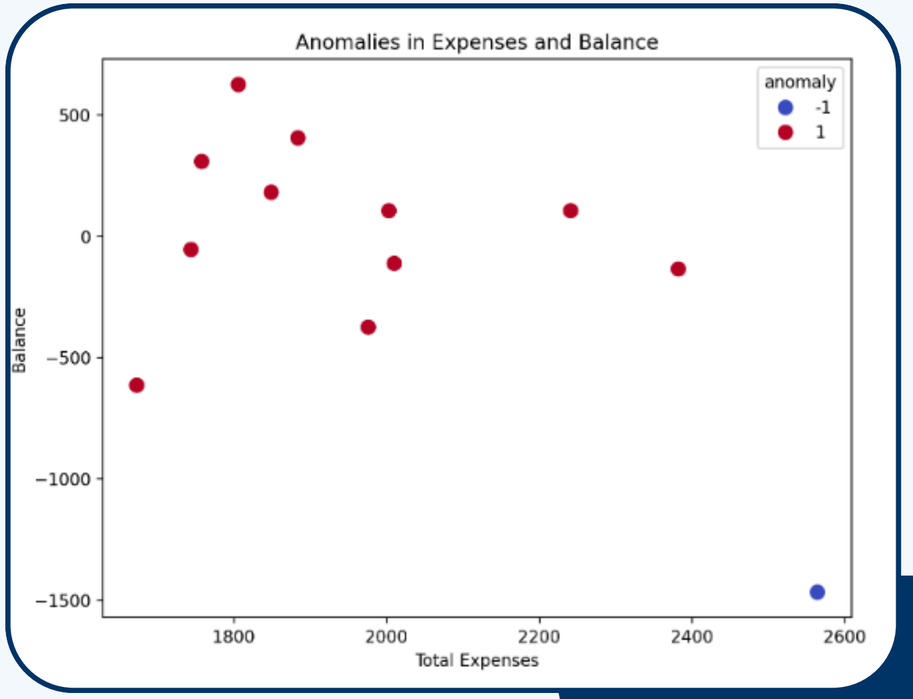
- Used KMeans to group users into categories like Essential, Semi-Essential and Extra based on their expenses.
- Displayed a pie chart to show how users are distributed across clusters.
- Created reports that compare each user's expenses to their cluster's average.

```
User 10003 (Month 5) belongs to the 'Semi-Essential' cluster.
The user's expenses are more aligned with the following categories:
housing: Essential
fixed_expenses: Essential
transport: Less Important
food: Less Important
health: Essential
education: Less Important
communication: Essential
other: Essential
investments_and_savings: Essential
credits_and_debts: Less Important
total_expenses: Essential
balance: Essential
```



MODEL SELECTION
ISOLATION FOREST

- Implemented the Isolation Forest model to identify anomalies in the dataset, categorizing data points as -1 (anomaly) and 1 (normal).
- Trained the model using 100 estimators, with the assumption that 10% of the data is expected to be anomalous.
- Counted the anomalies and normal data points, and add a new column to indicate the anomaly status.



MODEL SELECTION
XGBOOST OR RANDOM FOREST

Model 1 ❌

Random Forest Results

	Target	MSE	RMSE	MAE	R^2
0	total_expenses	4171.586778	64.587822	51.623822	0.947891
1	balance	9879.933226	99.397853	78.744767	0.964334

Model 2 ✅

XGBoost Results

	Target	MSE	RMSE	MAE	R^2
0	total_expenses	2653.932861	51.516336	40.507618	0.966848
1	balance	5705.843262	75.537032	60.140591	0.979402

STREAMLIT FINANCIAL ENVIROMENT

- Users can review finances, track expenses, and download reports with highlighted anomalies by month.
- The bot provides financial advice based on two financial books, tailored to user needs.
- Streamlit Interface & Dashboard: Built on Streamlit, offering an intuitive interface and dashboard for visualizing financial data and insights.

FINBOT
CHATBOT DEVELOPMENT

Driven Points

- Chatbot Development: Build AI-driven chatbot using OpenAI and NLP for context-aware financial advice.
- Data Retrieval: Retrieve relevant financial documents from SQL database for accurate responses.
- NLP & Prompt Engineering: Created a promped to provide clear and concise financial advice, with practical solutions based on the financial books, explaining key concepts by using analogies.
- Deployment & Integration: Deploy on Streamlit, integrate Chroma DB for fast document retrieval.
- Maintenance & Improvements: Monitor and improve model parameters, enhance clustering model and dashboard filters for better analysis.

CONCLUSIONS & NEXT STEPS

Conclusions

- The average user balance is generally very low, while total spending exhibits a high range and variance.
- The clustering model reveals that users with higher salaries tend to have better financial balances than others. However, in some cases, the model struggles to accurately capture group patterns.
- The Isolation Forest model is performing well in anomaly detection, and XGBoost, the chosen model, achieves over 90% accuracy in predicting financial outcomes.

Next Steps

- Fine Tuning the K-means model to improve the silhuete score metric, and exploring evalution on the Isolation Forest Model.
- Improve the Streamlit Dashboard to capture more hiden patterns.

THANK YOU