MILESTONE-3

Dashboard & Reporting

1. Introduction / Project Overview

The Receipt and Invoice Digitizer is a data driven document processing and analytics system designed to automate the extraction, storage analysis and visualization of financial information from receipts and invoices. In traditional expense management systems users manually record expenses which is time consuming error prone and difficult to analyze over long periods. This project removes manual effort by using Optical CharacterRecognition Natural Language Processing & dataanalytics to transform unstructured receipt images into structured and actionable financial data.

The system allows users to upload receipt or invoice images through a web based dashboard developed using Streamlit. Once uploaded the system extracts important fields such as vendor name bill amount tax date and category. The extracted information is validated, cleaned and stored in a centralized database. Over time this data forms a personal or organizational expense repository referred to as the Receipt Vault.

Beyond basic data extraction the project focuses strongly on analytics and reporting. Users can monitor monthly spending trends, view vendor wise summaries, analyze category wise expenses and detect anomalies such as unusually high value bills. The dashboard also provides CSV and Excel export options allowing seamless integration with accounting systems and external reporting tools.

This project demonstrates practical applications of OCR NLP feature engineering descriptive analytics and dashboard visualization making it suitable for both enterprise expense tracking and individual financial management use cases.

2. Problem Statement

Manual receipt handling is inefficient, inconsistent and unreliable. Individuals and organizations frequently lose receipts, miscalculate expenses or fail to identify spending patterns due to fragmented and unstructured records. Existing solutions often focus only on storage or require extensive manual intervention for categorization validation and analysis.

Unstructured receipt formats make automated processing difficult and manual data entry introduces errors, missing values and inconsistencies. The absence of real time analytics prevents users from understanding their spending behavior effectively. There is no built in mechanism to automatically detect duplicate receipts, abnormal transactions or unusual tax values. Traditional expense systems also provide limited reporting and visualization capabilities.

Due to the lack of an intelligent end to end receipt analytics system users experience poor financial visibility, reduced control over expenses and ineffective decision making.

3. Objectives

The primary objective of the **Receipt and Invoice Digitizer (Receipt Vault Analyzer)** is to transform unstructured receipt and invoice data into structured, reliable, and actionable financial insights through automation and analytics. To achieve this overarching goal, the project is divided into the following well-defined sub-objectives:

1. Automated Receipt and Invoice Digitization

- To eliminate manual expense entry by automatically digitizing receipts and invoices using Optical Character Recognition (OCR).
- To support multiple receipt formats and layouts, ensuring adaptability to real-world variations.
- To reduce human error and data loss commonly associated with manual record keeping.

2. Accurate Data Extraction and Validation

- To extract critical financial fields such as vendor name, transaction date, bill amount, tax, and category using NLP and regex-based techniques.
- To validate extracted data by handling missing values, correcting formatting inconsistencies, and standardizing currency and date formats.
- To identify and remove duplicate receipts to maintain data integrity and reliability.

3. Centralized and Scalable Data Storage

- To design a structured and query-friendly database for storing processed receipt and transaction data.
- To create a long-term expense repository (Receipt Vault) that supports historical analysis and reporting.
- To ensure efficient data retrieval for analytics, visualization, and export operations.

4. Feature Engineering and Data Enrichment

- To generate derived attributes such as month, year, weekday, vendor frequency, and category-wise totals.
- To enrich raw transaction data to enable deeper analytical insights.
- To prepare the dataset for trend analysis, comparative analytics, and anomaly detection.

5. Analytical Insight Generation

- To perform descriptive analytics including total expenditure, average spending, highest and lowest transactions, and summary statistics.
- To analyze spending behavior over time through monthly, weekly, and seasonal trend analysis.
- To compare expenses across vendors, categories, and time periods using comparative analytics techniques.
- To apply advanced analytical methods such as Pareto analysis for identifying top cost contributors.

6. Anomaly and Outlier Detection

- To identify unusually high-value transactions and abnormal tax values using statistical methods.
- To highlight spending outliers through box plots and distribution analysis.
- To support financial audits and cost control by flagging irregular transactions.

7. Interactive Dashboard and Visualization

- To design an intuitive and interactive dashboard using Streamlit for real-time data visualization.
- To present analytical insights through purpose-driven charts such as line charts, bar charts, stacked bar charts, histograms, and box plots.
- To ensure visualizations are decision-focused, aiding users in understanding trends, distributions, and spending behavior.

8. Reporting and Data Export

- To enable users to export processed data and analytical summaries in CSV and Excel formats.
- To support integration with external accounting tools and reporting systems.
- To provide reusable reports for audits, financial reviews, and compliance needs.

9. Usability and User Experience Enhancement

- To refactor the user interface for clear navigation and ease of use.
- To make complex financial data accessible to both technical and non-technical users.
- To ensure the dashboard supports both individual and enterprise-level use cases.

10. System Sustainability and Future Scalability

- To design the system with modular architecture for future feature expansion.
- To ensure maintainability, performance efficiency, and long-term usability.
- To lay the foundation for advanced enhancements such as machine learning–based categorization, predictive analytics, and multi-source receipt ingestion.

4. Framework and Technology Stack

The user interface of the system is built using **Streamlit** which provides an interactive dashboard for upload navigation and visualization. Charts and graphs are created using Plotly to support clear data representation.

The backend processing layer is implemented in Python which handles data extraction validation and analytics. **Tesseract OCR** is used for text extraction from receipt images while OpenCV and PIL are used for image preprocessing tasks such as resizing noise removal and enhancement. Regex patterns and NLP techniques are applied to identify entities like vendor amount tax and date.
The data layer uses **SQLite3** to store receipt records, vendor details and category information. Pandas is used extensively for data transformation cleaning and analytical operations.
The analytics layer supports descriptive analytics trend analysis business analysis comparative analytics and anomaly detection to provide meaningful insights into spending behavior.
Plotly Express is a high-level Python library that allows for the quick creation of interactive and beautiful data visualizations with minimal code. It is a wrapper for the more comprehensive Plotly graph library, simplifying the API for common chart types.Advantages of Plotly Express over Matplotlib and Seaborn

The primary advantages of Plotly Express (and by extension, Plotly) over Matplotlib and Seaborn stem from its interactivity and ease of use.

Feature	Plotly Express	Matplotlib/Seaborn	Advantage
Interactivity	**Interactive** charts (zoom, pan, hover tooltips) are created by default.	**Static** images by default. Interactivity requires additional code (e.g., extensions like mpld3).	Facilitates deeper data exploration and presentation without extra development effort.
Code Conciseness	**High-level API** requires a single function call for complex charts (e.g., px.scatter(), px.histogram()).	**Low-to-medium level API** often requires multiple lines of code to set up figures, axes, and plot elements.	Faster development and easier for quick, exploratory data analysis.
Data Format	Directly accepts a **Pandas DataFrame** and column names, simplifying data mapping.	Primarily works with NumPy arrays; Seaborn accepts DataFrames but still requires more explicit data binding.	More intuitive and less error-prone for users already working with Pandas DataFrames.

Output Format	Charts can be rendered in web browsers, notebooks, and easily exported as HTML, allowing for easy sharing.	Primarily exports as static image files (PNG, JPEG, PDF).	Native support for modern web-based data visualization standards.
Theming	Comes with excellent, modern default color themes and styles.	Default aesthetics can be visually basic and often require custom styling for professional reports.	Better-looking charts out-of-the-box.

5. Analytics Module Core of Milestone

The analytics module in the Receipt Vault Analyzer is designed to convert raw receipt data into meaningful insights that help users clearly understand their spending habits. After receipt data is stored the system performs data preprocessing steps such as duplicate receipt detection, missing value handling and standardization of dates currency values and numerical fields.

Feature engineering is applied to generate additional attributes including month year weekday total bill amount vendor frequency and category wise spending totals. These engineered features enable deeper analysis and improve insight quality.

Descriptive analytics is used to compute total expenditure average spending per receipt highest and lowest transaction values and vendor wise and category wise summaries. To understand spending behavior over time the system performs trend analysis using time series techniques which track monthly, weekly and seasonal spending patterns.

Comparative analytics compares expenses across vendors categories and different time periods such as month on month comparisons. Intelligence enhancements such as Pareto analysis are applied to identify the top vendors contributing to the majority of expenses. Vendor and category ranking helps prioritize major spending sources while anomaly detection highlights unusually high value receipts or abnormal tax rates.

All analytical results are presented using visualizations such as line charts, bar charts, stacked bar charts, heatmaps and boxplots to make insights intuitive and easy to interpret.

6. Types of Charts and Their Purpose

For business oriented reporting line charts are used to represent monthly spending trends which support financial forecasting and budget planning. Stacked bar charts are used to display category wise monthly expenses enabling clear comparison of cost distribution across time periods.

For individual usage pie charts are used to show personal expense distribution across categories while boxplots are used to identify unusually high expenses and spending

outliers. Each chart is designed to directly support decision making rather than serve as a decorative visual element.

Monthly Spending (Statistics)

These charts focus on trends, comparisons, and distribution over time:

- **Line Chart:** Best for showing spending trends across months and makes increases/decreases easy to spot.
- **Bar Chart / Column Chart:** Great for comparing spending amounts month-by-month; easy to read and widely used in reports.
- **Area Chart:** Similar to a line chart but emphasizes volume of spending over time; useful when highlighting cumulative impact.
- **Stacked Bar Chart:** Shows total monthly spending plus category-wise breakdown; helpful when spending is split across types (rent, food, travel, etc.).
- **Histogram:** Shows the distribution of spending values, useful for identifying frequent spending ranges.
- **Box Plot:** Highlights median, spread, and outliers in monthly expenses; often used in analysis-heavy reports.

Vendor Statistics

These charts focus on comparison, contribution, and proportions:

- **Bar Chart:** Best for comparing spending or transactions across vendors; works well when vendors are many.
- **Pie Chart / Donut Chart:** Shows each vendor's percentage contribution; best when vendors are few (5–7 max).
- **Stacked Bar Chart:** Compares vendors across categories or time periods; useful for multi-dimensional analysis.
- **Treemap:** Excellent for visualizing vendor contribution by size; space-efficient and visually intuitive.
- **Pareto Chart:** Highlights top vendors contributing to the majority of spending (80/20 rule); great for cost optimization insights.
- **Table with Conditional Formatting:** Provides precise values with visual cues (color scales, data bars); useful for dashboards and audits.

Category Distribution

- **Proportion & Share (Who takes how much?)**
    - **Pie Chart:** Shows percentage share of each category; simple, intuitive; best for few categories (≤6).
    - **Donut Chart:** Same as pie, but cleaner and more modern; the center can show total spending.

    - **Treemap:** Great when there are many categories; the size of each block equals contribution.
- **Comparison Across Categories**
    - **Bar Chart (Horizontal preferred):** Best for comparing category values clearly; handles many categories better than pie charts.
    - **Stacked Bar Chart:** Shows category contribution within a total; useful when comparing across months or vendors.
- **Distribution & Spread (Statistical view)**
    - **Histogram:** Shows the frequency of spending amounts per category; good for identifying common ranges.
    - **Box Plot:** Highlights variation, median, and outliers per category; best for analytical or research contexts.
- **Advanced / Insight-Driven**
    - **Pareto Chart:** Identifies categories contributing to the majority of spending; useful for cost control and prioritization.
    - **Sunburst Chart:** Best for hierarchical categories (e.g., Food → Dining → Online); shows parent-child relationships clearly.

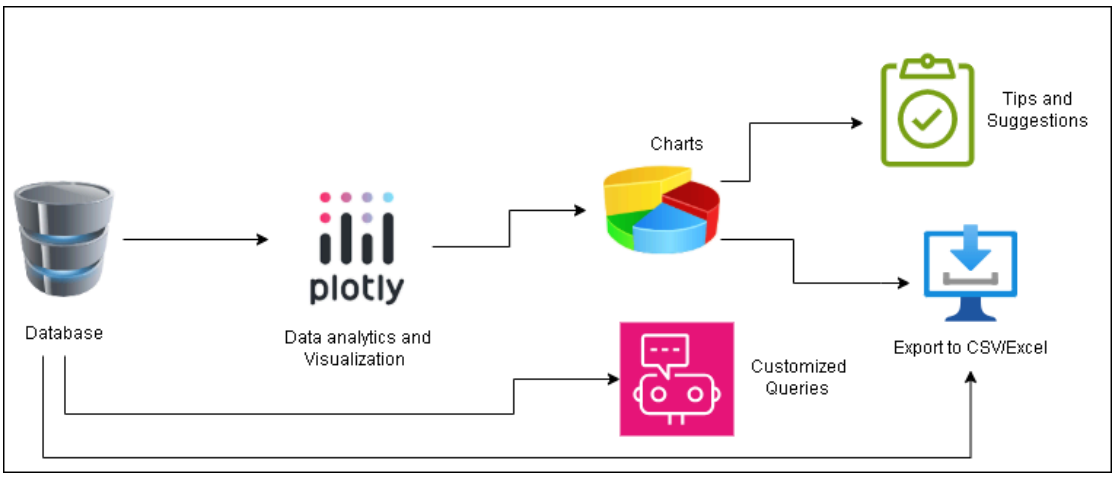
7. Dashboard & Reporting Milestone3 Deliverable

The dashboard is developed using Streamlit and provides a unified interface for receipt upload review and analytics. It displays real time extraction and validation status along with monthly spending analytics, vendor statistics and category wise summaries.

The dashboard also includes export functionality allowing users to download processed data and analytical results in CSV and Excel formats. This dashboard serves both technical validation requirements and end user usability needs.

8. System Architecture

The system follows a modular layered architecture consisting of a user interface layer implemented using Streamlit, a processing layer responsible for OCR and NLP based extraction, a storage layer for database management, an analytics layer for insight generation and visualization and an export layer for generating CSV and Excel reports.



3.1 Data Extraction Layer

The data extraction layer retrieves structured invoice and receipt data from the SQL database. This layer queries the database to fetch relevant financial records including vendor information, transaction amounts, dates, categories, and tax details. The extracted data is formatted into pandas DataFrames for efficient manipulation and analysis.

3.2 Visualization Engine

The visualization engine processes the extracted DataFrame data and generates interactive charts using plotly.express as the primary framework. This layer handles data aggregation, metric calculation, and chart configuration. The engine supports multiple chart types to represent different aspects of financial data effectively.

3.3 Forecasting Module

The user interface layer presents the generated visualizations in an organized dashboard layout. This layer handles user interactions, chart rendering, and navigation between different analytics views. The interface is designed to be intuitive and responsive, enabling users to explore their financial data seamlessly.

3.4 Tips and Suggestions Module

The LLM-powered conversational interface is designed to significantly improve how users interact with expense data. By supporting natural language queries, it allows users to ask questions in a manner that feels intuitive and human-like, eliminating the need to learn complex query syntax or navigate cumbersome menus.

This advanced interface provides instant expense insights, delivering immediate and relevant information that helps users understand their spending patterns quickly. Furthermore, it enables interactive data exploration, allowing users to drill down into details, filter results, and visualize data in dynamic ways. The overall goal of these features is to enhance user engagement and usability, making the process of expense management more efficient, accessible, and user-friendly.

3.5 Web Dashboard Module

The application will feature a user-friendly Streamlit-based interface with a clean and modern design, incorporating a responsive layout and a drag-and-drop interface for ease of use. It will also include a secure login and authentication system, along with multi-language support to ensure maximum accessibility.

9. Conclusion

The Receipt and Invoice Digitizer successfully automates expense tracking and analytics by integrating OCR NLP and data visualization into a single cohesive platform. The system significantly reduces manual effort, improves data accuracy and provides actionable insights into spending behavior.The analytics driven dashboard ensures transparency accuracy and usability making the solution suitable for individual users as well as enterprise level expense management applications.

References

1. Towards Natural Language-Based Document Image Retrieval: New Dataset and Benchmark https://arxiv.org/abs/2512.20174
2. ReceiptSense: Beyond Traditional OCR -- A Dataset for Receipt Understanding https://arxiv.org/abs/2406.04493
3. E2E Process Automation Leveraging Generative AI and IDP-Based Automation Agent: A Case Study on Corporate Expense Processing https://arxiv.org/abs/2505.20733
4. MMDocBench: Benchmarking Large Vision-Language Models for Fine-Grained Visual Document Understanding https://arxiv.org/abs/2410.21311
5. Large Language Models for Simultaneous Named Entity Extraction and Spelling Correction https://arxiv.org/abs/2403.00528
6. TransDocAnalyser: A Framework for Offline Semi-structured Handwritten Document Analysis in the Legal Domain https://arxiv.org/abs/2306.02142

7. Visual Information Extraction in the Wild: Practical Dataset and End-to-end Solution https://arxiv.org/abs/2305.07498
8. Extending TrOCR for Text Localization-Free OCR of Full-Page Scanned Receipt Images https://arxiv.org/abs/2212.05525
9. Robustness Evaluation of Transformer-based Form Field Extractors via Form Attacks https://arxiv.org/abs/2110.04413
10. LayoutReader: Pre-training of Text and Layout for Reading Order Detection https://arxiv.org/abs/2108.11591
11. End-to-End Information Extraction by Character-Level Embedding and Multi-Stage Attentional U-Net https://arxiv.org/abs/2106.00952
12. ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction https://arxiv.org/abs/2103.10213
13. TLGAN: document Text Localization using Generative Adversarial Nets https://arxiv.org/abs/2010.11547
14. Abstractive Information Extraction from Scanned Invoices (AIESI) using End-to-end Sequential Approach https://arxiv.org/abs/2009.05728
15. EATEN: Entity-aware Attention for Single Shot Visual Text Extraction https://arxiv.org/abs/1909.09380
16. Deep Learning Approach for Receipt Recognition https://arxiv.org/abs/1905.12817
17. Business Research Insights. (2025). Invoice OCR API Market Size, Share & Trends,2025-2033 https://www.businessresearchinsights.com/market-reports/invoice-ocr-api-market-115569
18. Invoice OCR in 3–5 Seconds: 2025 Benchmark of Veryfi vs. Google Cloud Vision vs. Mindee https://www.veryfi.com/ai-insights/invoice-ocr-competitors-veryfi/
19. Five Reasons Why OCR Isn't Enough https://tipalti.com/blog/five-reasons-why-ocr-isnt-enough/
20. DocuClipper. (2025). 9 Biggest OCR Limitations And How To Overcome Them. https://www.docuclipper.com/blog/ocr-limitations/