Generative AI for Automated Business Reports: Insights from the Olist E-Commerce Dataset

Sacred Heart University
Final Project Proposal

Presented by
Eliza Paul Ganta
gantae@mail.sacredheart.edu

Introduction

Business reporting is one of the most critical functions in organizations. Executives rely on regular reports to track business health, understand customer behavior, identify operational bottlenecks, and make strategic decisions. Traditionally, these reports are prepared by analysts who must extract and clean raw data, merge multiple datasets, compute performance metrics, visualize results, and finally draft written reports summarizing the findings. While effective, this process has limitations:

- **Time-consuming:** Manual preparation of reports can take days or weeks, creating delays in decision-making.
- **Resource-intensive:** Analysts spend significant time on repetitive reporting tasks rather than strategic analysis.
- **Error-prone:** Manual compilation of metrics and reports increases the risk of inconsistencies and errors.

The emergence of **Generative AI**, powered by Large Language Models (LLMs) like GPT-4 and GPT-5, provides an opportunity to transform reporting. Instead of relying on manual text creation, LLMs can convert structured data into clear, natural language narratives. This allows businesses to automate executive summaries and provide decision-makers with timely, actionable insights.

This project focuses on building a proof-of-concept AI-powered reporting system using the **Olist Brazilian E-Commerce Dataset**, which captures real-world e-commerce operations across multiple dimensions such as orders, customers, deliveries, and reviews. The project demonstrates how AI can serve as a "junior business analyst" by automatically generating reports that summarize business performance, highlight trends, and provide recommendations.

Problem Statement

Companies such as Olist handle thousands of daily transactions across diverse categories and geographies. Executives and managers require frequent reports on:

- **Revenue and Sales Trends** – to monitor growth and profitability.
- **Customer Retention and Churn** – to evaluate loyalty and long-term business value.
- **Delivery and Logistics Performance** – to ensure operational efficiency.
- **Customer Sentiment** – to understand satisfaction and areas of concern.

Currently, preparing such reports involves:

1. **Cleaning Data:** Removing duplicates, handling missing values, standardizing timestamps.
2. **Merging Data Tables:** Integrating customer, order, payment, and review data.
3. **Computing KPIs:** Summarizing metrics such as monthly revenue or churn rates.
4. **Creating Dashboards:** Visualizing data for interpretation.
5. **Writing Reports:** Drafting executive summaries for stakeholders.

This manual process is inefficient, especially for fast-moving e-commerce companies where insights need to be near real-time. As a result, decision-making often lags behind data availability.

The proposed project addresses this problem by developing an **end-to-end reporting pipeline** where Generative AI automatically transforms raw data and KPIs into natural language reports, reducing manual workload and accelerating decision-making.

Objectives

The project's objectives are structured around the full analytics pipeline:

1. **Data Collection & Preparation**
    - Acquire the Olist dataset from Kaggle.
    - Clean and merge multiple tables into an analytical schema suitable for reporting.
2. **KPI Derivation**
    - Define and compute critical KPIs such as:
        - Monthly revenue and growth rates.
        - Average order value (AOV).
        - Customer churn and retention.
        - Average delivery time and % of late deliveries.
        - Review sentiment distribution.
3. **Generative AI Integration**
    - Convert KPI outputs into structured "fact sheets."
    - Use LLMs to automatically generate business reports, including executive summaries and recommendations.
4. **Visualization & Comparison**
    - Build dashboards to visualize KPIs.
    - Compare AI-generated reports with dashboards and human-written summaries.
5. **Demonstration & Validation**
    - Deliver a working prototype that combines dashboards with automated natural language reports.
    - Evaluate the clarity, accuracy, and usefulness of AI-generated reports.

Dataset

The project will use the **Olist Brazilian E-Commerce Dataset**, publicly available on Kaggle. It contains data from a large Brazilian e-commerce marketplace across multiple linked tables:

- **Customers:** Unique customer IDs, state, and location.
- **Orders:** Order IDs, purchase and delivery timestamps, order status.
- **Order Items:** Product categories, quantities, prices, freight values.
- **Payments:** Payment type (credit card, boleto, voucher), installment information, and transaction amounts.
- **Products:** Product categories and characteristics.
- **Reviews:** Customer ratings and written reviews.
- **Geolocation:** Customer and seller locations by state and city.

This dataset is comprehensive, spanning financial, operational, and customer experience metrics. It is ideal for demonstrating the value of AI-driven reporting, as it covers the **end-to-end e-commerce cycle**: order placement → payment → delivery → customer feedback.

Methodology

The project will follow the **CRISP-DM framework** with detailed steps:

1. Data Understanding and Preparation

- Load raw CSVs into Python using Pandas.
- Handle missing values and remove inconsistencies.
- Standardize date/time fields to calculate delivery times and monthly revenue.
- Merge tables into an analytical schema (star schema format).
- Create derived features such as:
    - **Revenue per order** (price + freight).
    - **Late delivery flag** (delivery > promised date).
    - **Repeat purchase indicator** (customer order frequency).

2. Exploratory Data Analysis (EDA)

- Summarize distributions of key variables (order volume, payment methods, delivery times).
- Identify geographic patterns (e.g., highest sales by region).
- Explore customer review distributions (average ratings, positive vs negative review trends).

3. KPI Computation

Define business metrics across dimensions:

- **Financial KPIs**
    - Monthly revenue and growth rate.
    - Average order value.
    - Payment method distribution.
- **Customer KPIs**
    - Retention rate: proportion of repeat customers.
    - Churn rate: customers who do not reorder in 90 days.
- **Logistics KPIs**
    - Average delivery time.
    - % late deliveries.
    - Freight cost contribution to revenue.
- **Sentiment KPIs**
    - Average review score.
    - Positive vs negative review ratio.
    - Word cloud of most frequent terms in reviews.

4. Visualization

Use **Tableau** and **Python visualizations (Matplotlib, Plotly)** to present insights. Dashboards will include:

- Revenue and order trends (monthly).
- Geographic sales distribution (by Brazilian states).
- Category performance breakdown.
- Delivery and logistics performance metrics.
- Sentiment analysis of customer reviews.

5. Generative AI Integration

- Transform KPI outputs into structured JSON “fact sheets” for each month.
- Use **LangChain + GPT API** to prompt LLMs to generate:
    - **Executive Summary:** concise overview of company performance.
    - **Key Highlights:** financial, customer, and operational insights.
    - **Recommendations:** actions based on observed trends.

Example AI-generated summary output:

"In March 2018, Olist generated R$3.2M in revenue with a 5% growth compared to February. However, late deliveries rose to 15%, primarily in the Southeast region. Customer churn increased to 12%, with negative reviews mentioning delays. Recommendation: Improve logistics in high-delay regions and incentivize repeat purchases with targeted campaigns."

6. Evaluation

- Compare AI-generated reports against:
  - Tableau dashboards.
  - Manually written reports.
- Criteria:
  - **Accuracy:** Are numerical values consistent with KPI calculations?
  - **Clarity:** Are insights communicated clearly for non-technical users?
  - **Usefulness:** Do reports include actionable recommendations?

Tools & Technologies

- **Python (Pandas, NumPy, Plotly, Matplotlib):** Data preparation, KPI computation, exploratory analysis.
- **Tableau:** Interactive dashboards and visual reporting.
- **LangChain + GPT API:** AI-driven report generation.
- **Streamlit (optional):** Web app to combine dashboards and reports.
- **Kaggle Olist Dataset:** Primary dataset source.

Deliverables

1. **Final Report (10–15 pages)** – documenting methodology, findings, and AI integration.
2. **Jupyter Notebook** – reproducible code for data cleaning, KPI analysis, and AI pipeline.
3. **Dashboards** – Tableau or Python dashboards for interactive visualizations.
4. **AI Report Generator** – automated reporting pipeline.
5. **Sample AI Reports** – monthly business reports generated by the system.
6. **PowerPoint Presentation** – summarizing the project's approach, results, and real-world impact.

Timeline (5 Weeks)

- **Week 1:** Acquire dataset, data cleaning, schema building.
- **Week 2:** KPI derivation and exploratory data analysis.
- **Week 3:** Dashboard development in Tableau/Python.
- **Week 4:** AI integration with LangChain + GPT.
- **Week 5:** Report writing, evaluation, and final presentation preparation.

Real-World Impact

The project demonstrates how Generative AI can reshape reporting processes in businesses:

1. **Efficiency Gains:** Reports generated in minutes instead of days.
2. **Consistency & Accuracy:** Reduced risk of human error.

3. **Enhanced Decision-Making:** Executives receive concise, AI-generated summaries with actionable recommendations.
4. **Scalability:** Approach can be applied across industries (retail, finance, healthcare).
5. **Future Applications:** Integration with real-time data streams to generate live reports.

For students, this project showcases the integration of **data analytics, visualization, and AI-driven natural language generation** in a single workflow—demonstrating the type of end-to-end pipeline that is highly relevant to industry applications.

References

- Kaggle: Olist Brazilian E-Commerce Dataset.
- McKinsey & Company (2024). *Embracing Generative AI in Business Reporting.*
- LangChain Documentation.
- OpenAI API Documentation (2025).
- Tableau Whitepapers on Business Analytics.
