Zero-Day Close: Automating Financial Accounting

Mikhail Sinitcyn, Efe Erhan

Abstract

This report serves as a proof-of-concept to a rapidly developing state-of-the-art paradigm in financial accounting: the Zero-Day Close, leveraging enterprise systems and financial pipelines to achieve state-of-the-art automation in bookkeeping and financial reporting.

1 Introduction

1.1 Intro to Accounting

Accounting, commonly referred to as "the language of business," is a system for collecting and analyzing financial data, creating financial statements, and interpreting the findings to make informed business decisions. The process of preparing financial statements is called "the close", traditionally a labor-intensive process that could take weeks to complete.

1.2 The Accounting Equation and Double-Entry Bookkeeping

The accounting equation forms the foundation of double-entry bookkeeping, a system pivotal in modern accounting. This fundamental equation necessitates that at any given time, a company's assets (what it owns) are financed by its liabilities (what it owes) and equity (the owner's claim after debts are paid). Mathematically, it is represented as: Assets = Liabilities + Equity. This equation ensures that every financial transaction is balanced, reflecting the dual impact on the financial statements.

1.3 Implementing the Zero-Day Close

"A zero-day close—also known as a continuous or touchless close—will let organizations close the books quickly and have access to up-to-date information. And while the goal is zero days, it's the actual process improvements that, with each one, are truly advancing the finance function..." - Workday

1.3.1 Assumption and Exemptions

- The company is assumed to be operating within the United States of America and accounting in USD.
- Although publicly traded companies are required by the SEC to report their financial performance quarterly, this report covers an annual account period, typical for private firms which have more flexibility in their reporting frequency

- Dynamic accounting factors such as asset depreciation, taxes, Cost of Goods and Services (COGS) are not addressed in this report because they are significantly beyond the scope of this course

2 Data Acquisition

2.1 The Data

Typically, a firm handles financial data from multiple sources such as an ERP (Enterprise Resource Planning) system and retail payment portals. For the purpose of this report, the given company has three sources of data: financial and managerial accounting (both extracted from ERP but handled separately), and online retail data from a payment portal. The accounting datasets were sourced from Kaggle, and the third was sourced from the UC Irvine ML Repository. They are individually described in further detail in subsequent sections.

2.1.1 Financial Accounting

This dataset contains financial transactions in double-entry bookkeeping format, each transaction recorded as a debit and a credit entry. Each entry corresponds to one of nine unique accounts. Additionally, each entry has a category (limited to Asset, Expense, Liability, and Revenue), transaction type, and payment method. This data is used to construct the Income Statement and Balance Sheet, both essential financial documents for evaluating a company's performance.

2.1.2 Managerial Accounting

Unlike the financial accounting dataset focused on external reporting, this dataset delves deeper into the operational aspects of the firm, primarily informing internal decision-making. Four teams across four projects, distributed evenly amongst four US cities. Each entry corresponds to one of four unique accounts and one of nine unique categories and four transaction types

2.1.3 Online Retail

This dataset comprises rows of transaction information that are keyed by an invoice number "InvoiceNo" and an item identifier "StockCode". Per each invoice number, there are fields "Quantity", "Price", "CustomerID", and "Country". A timestamp column "InvoiceDate" is also included. Order cancellations are denoted by a 'C' prepended to the invoice number, and returns can be identified by a negative quantity field.

Missing from this dataset is information on payment method. When converting to double-entry form, all transactions are assumed to be paid by credit card, not cash. Due to sales being done online, this is a valid assumption, and most good ERP services would record that information.

This dataset has over 500,000 rows of invoice information. For ease of use and for preserving CPU longevity, this was downsized to 100,000 using the Pandas DataFrame.sample() function. Additionally, attempts to manipulate the data into normality have been fruitless.

2.1.4 Assumptions and exemptions

The firm is assumed to be a privately owned company, thus simplifying the owners' equity calculations and foregoing the need to abide by SEC reporting regulations. Additional dynamic financial factors such as taxes, income taxes, inflation, and asset appreciation/depreciation are exempt from the data and report.

The financial and managerial accounting datasets along with the online retail dataset were transformed to comply with double-entry bookkeeping. In doing so, certain assumptions had been made and are described in the ETL notebooks.

3 Explore-Transform-Load (ETL)

3.1 Extract

Financial Accounting and Managerial Accounting were sourced from Kaggle. Online Retail dataset was sourced from UC Irvine Machine Learning Repository.

3.2 Transform

None of the three datasets were compliant with double-entry bookkeeping. Inside of the ETL folder are the three notebooks that transformed the datasets into the correct format. These scripts convert each transaction into two entries - a debit entry and a credit entry - to balance the two sides of the accounting equation. Each entry was classified according to the transaction type, account, and category. These scripts leverage accounting protocols and accomplish the transformation task using rather simple if-else statements with no need for machine learning at this stage.

3.3 Load

Given that the three datasets are small enough to comfortably work with locally - except online retail dataset which was resampled to a smaller sample size - the data is stored and loaded from the data folder in the repository. After the successful close of each accounting period, this data would typically be offloaded to a data lake for long term storage.

4 Data Analysis

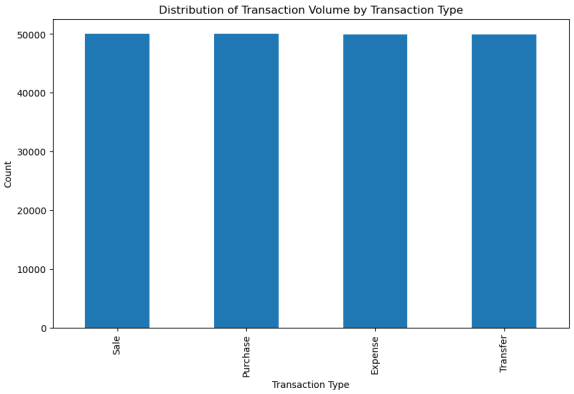
Once the data is ensured to be double-entry compliant, it is analyzed via charts and statistical tests to draw as many insights as possible. The data is ultimately used for the purpose of creating financial statements to describe and forecast the performance of the company, thus making this the most crucial step for decision making.

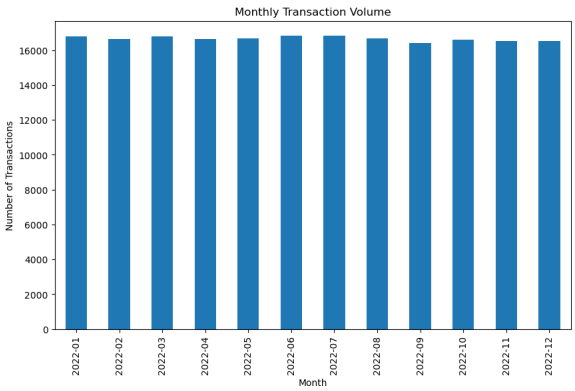
It is important to note that the data used in this report is fictitious and procedurally generated. Thus, distributions of data are unrealistic, resulting in partially unrealistic observations of the firm's yearly performance. Much further information is needed for prescriptive statistics. Regardless, the robustness of the pipeline and the value in this report remain unimpacted.

4.1 Financial Accounting

Transactions

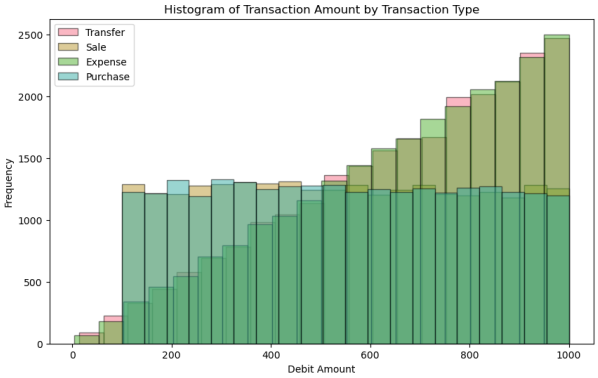
First, we observe the transactions as a whole, considering transaction volume and count, both monthly and annually.





We find that there is no difference in transaction volume month-to-month *(p=0.995)*. We find that there is no difference in the transaction by transaction type annually *(p=0.95)*.

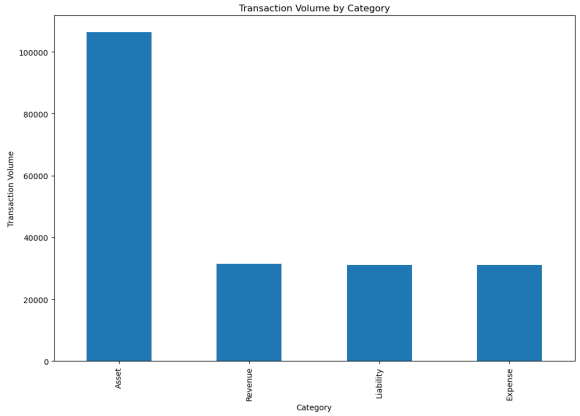
Having observed that transaction volume does not vary by transaction type, we analyze further and find that there is a difference in the distribution of transaction amounts by transaction type *(p=0)*.

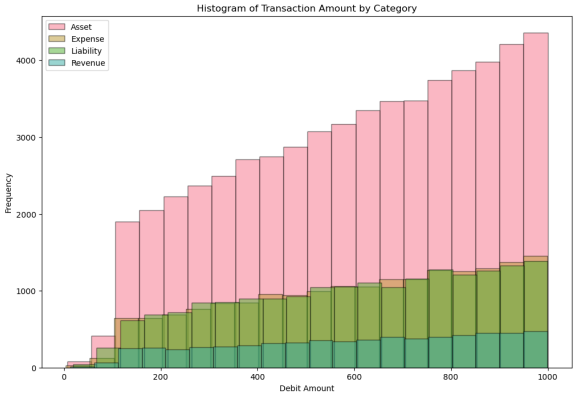


Performing a post-hoc Tukey's test, no pair of transaction types observe the same distribution.

Category

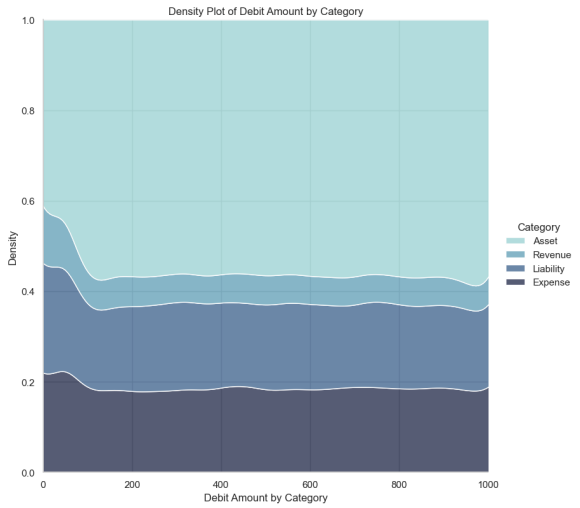
Per the barplot below, transaction volume varies by category, with assets observing a significantly higher frequency.

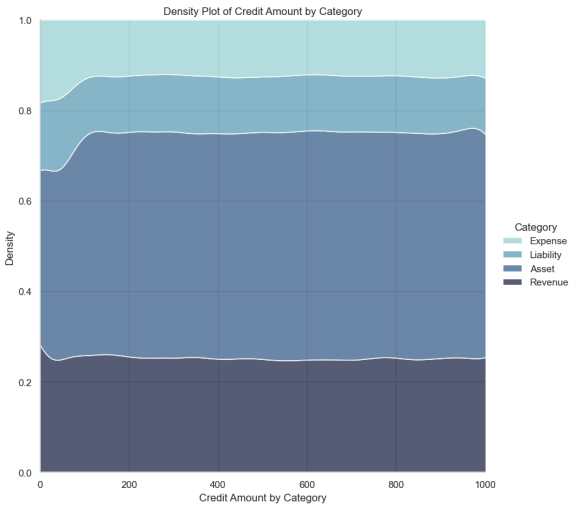




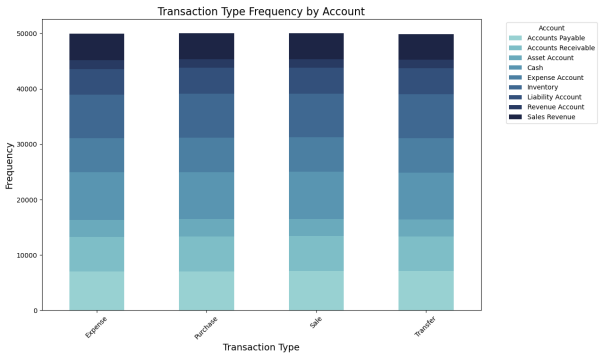
Plotting the distribution of transaction amount by category we find that there is no difference in the distribution of transaction amounts by category *(p=0.051)*, although with such a low p-value one may exercise best judgment and consider the difference statistically significant.

We further visualize the distribution of transaction amounts between categories via density plots, corresponding to debit and credit transactions. These density plots make the relationship between category and frequency of transaction amount more apparent.



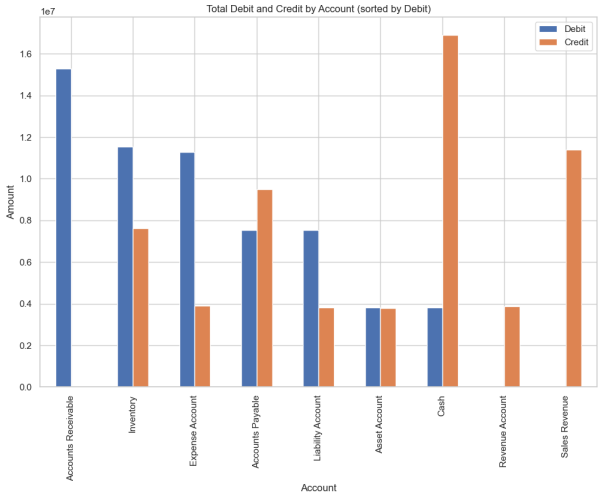


Account We plot the volume of transactions per transaction type and color the account distribution therein.



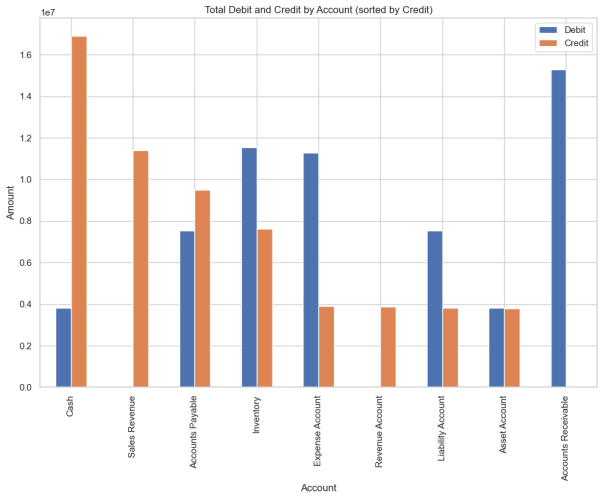
After conducting a chi-square test we find that there is no relationship between account and transaction type frequency *(p=0.99)*.

We rank the accounts by the sum of debits in the barchart below.



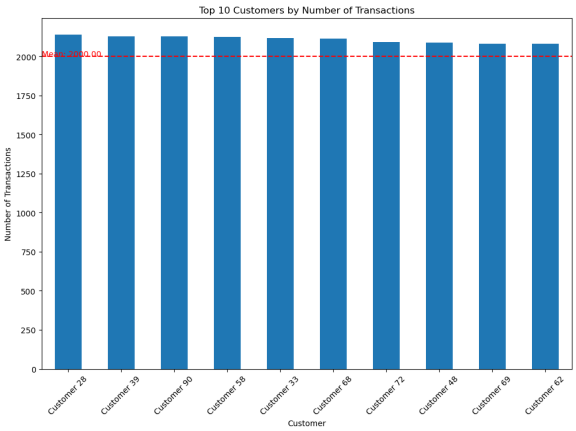
High debits in Accounts Receivable indicate robust sales on credit, likely a positive sign of business growth, but also requires further monitoring for analysis of liquidity and future collection (not covered in this report). Secondly, significant debits in Inventory and Expense accounts indicate substantial operational activities and cost management, while Accounts Payable being among the top debited accounts may reflect a strategic approach to managing cash flow by leveraging credit terms with suppliers.

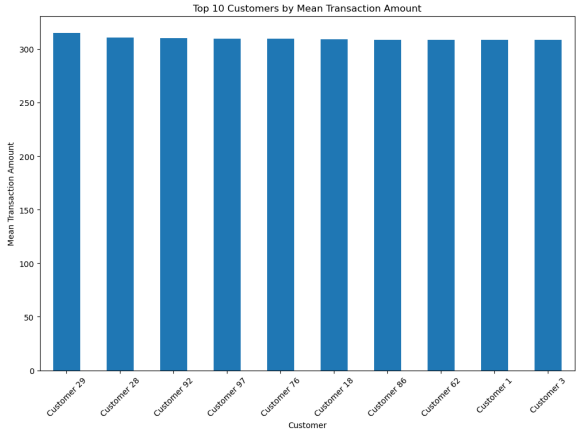
We also rank accounts by the sum of credits in the barchart below.



High credits in Cash and Sales Revenue suggest strong revenue generation and effective cash management, ultimately indicating healthy business operations.

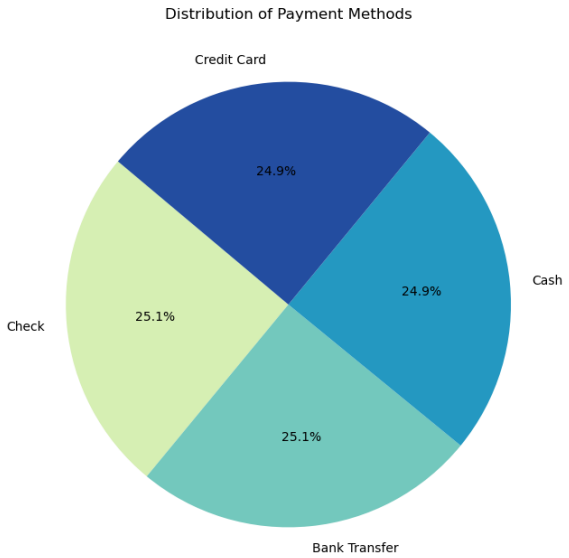
Customer Over the accounting period, the firm has had transactions with exactly 100 unique customer accounts.

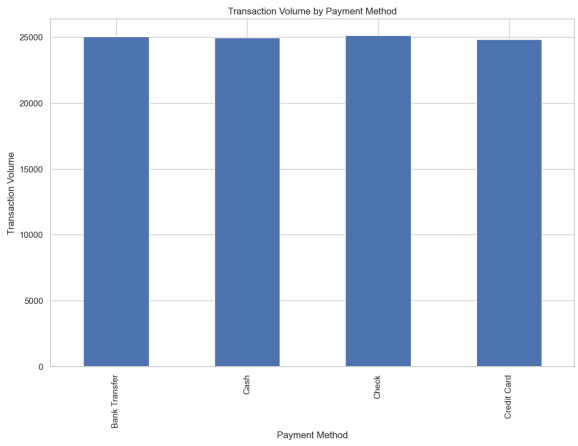




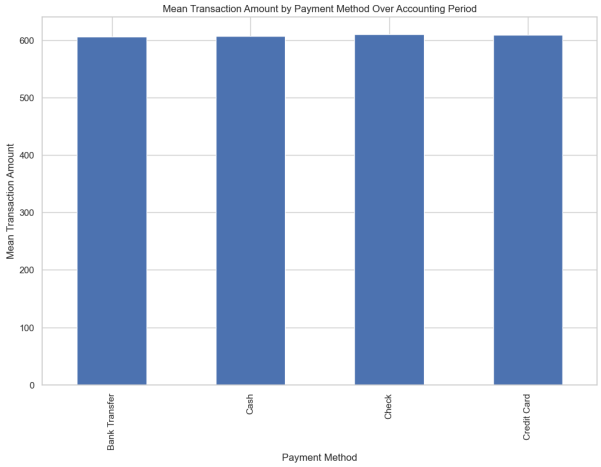
Plotting the customer accounts by transaction volume and mean transaction amount, we find that there is no difference in either between customers *(p=1.0, p=0.99)*

Payment Methods Over the accounting period, the firm has conducted transactions in 4 unique payment methods.





The pie chart indicates that the distribution of payment methods over the accounting period is even, as confirmed by the barplot and its corresponding ANOVA test (p=0.23). A chi-square test confirms that, at a 5 % significance level, there is no relationship between month and payment method *(p=0.51)*.



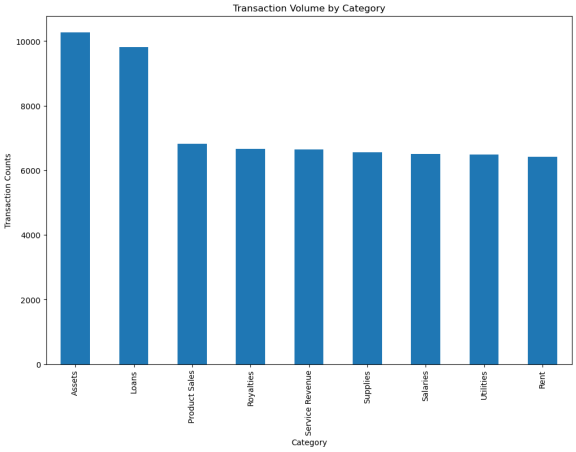
Per the barplot above and its corresponding ANOVA test, there is no relationship between payment method and mean transaction amount. We also find that there is no relationship between transaction type and payment method *(p=0.21)*

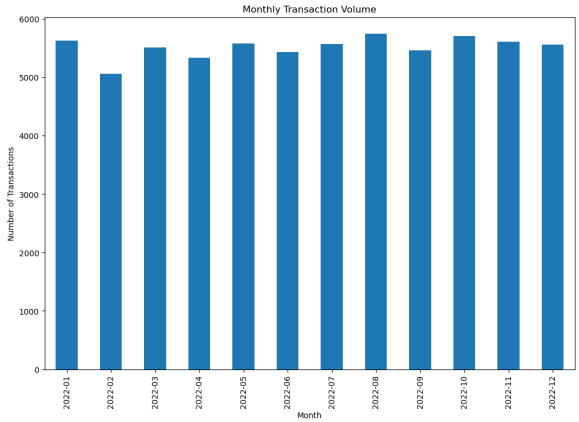
Are sales increasing? Are expenses growing faster than revenue?

Fitting a linear regression model over the account period, we find no trend in revenue *(p=0.79)* or expenses *(p=0.85)*.

4.2 Managerial Accounting

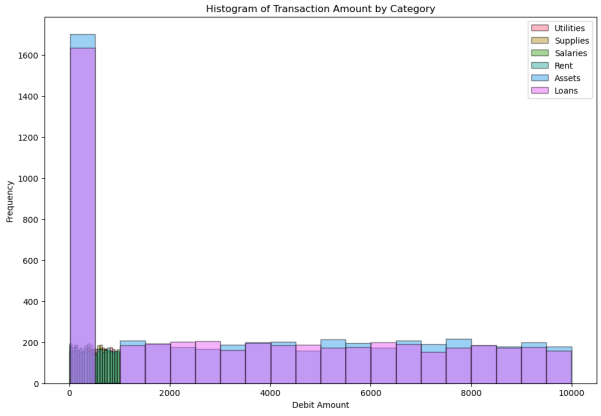
Transactions First, we observe the transactions as a whole, considering monthly and annual transaction volume.





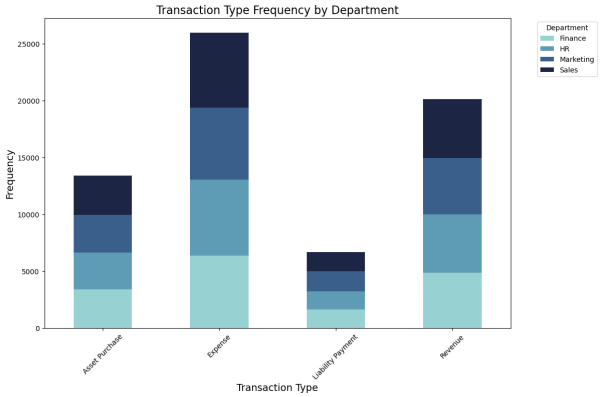
Assets and Loans have the higher concentration of transaction volume, but overall volume does not vary on a monthly basis within the accounting period *(p=0.97)*.

Category



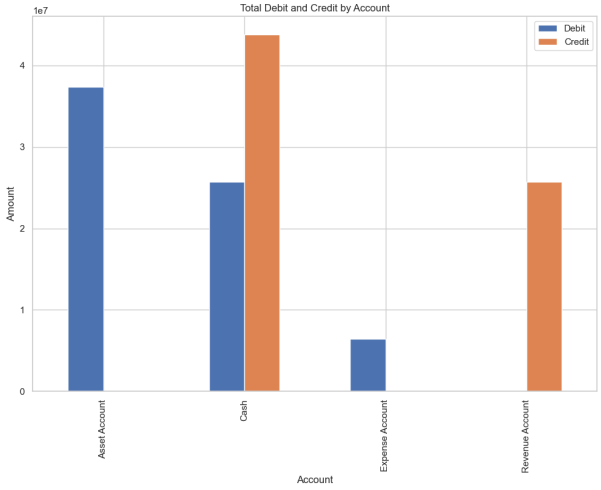
We find that there is a significant difference in the distribution transaction amounts by category *(p=0.0)*. Post-Hoc Tukey's HSD Test results are available in the corresponding notebook.

Department



Separating transactions by transaction type and department, we observe that transaction type volume varies by department, although the inner transaction types remain evenly distributed *(p=0.001)*.

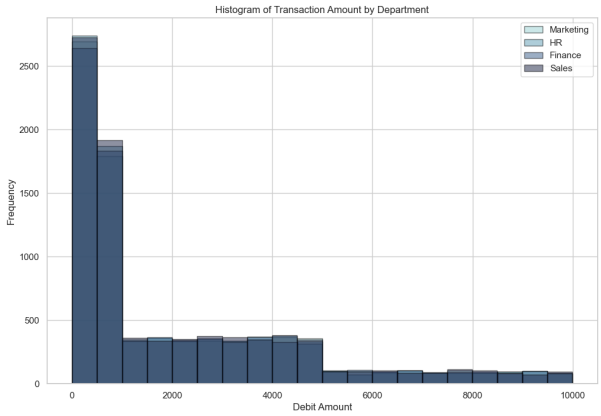
Account



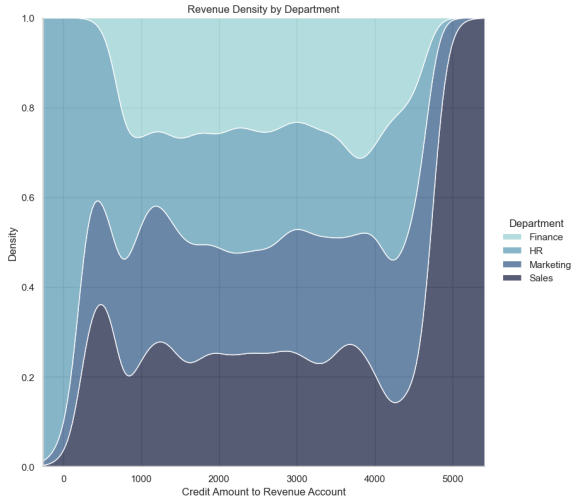
Having ranked the accounts by sum debit amounts, we observe that significant debits in the asset account indicate significant investment and asset acquisition, suggesting expansion and growth of business operations. As a result, substantial debits in the cash account reflecting high cash outflows due to purchasing inventory, paying expenses, and investing in assets. This pattern indicates active financial management geared towards growth and operational enhancement over the accounting period.

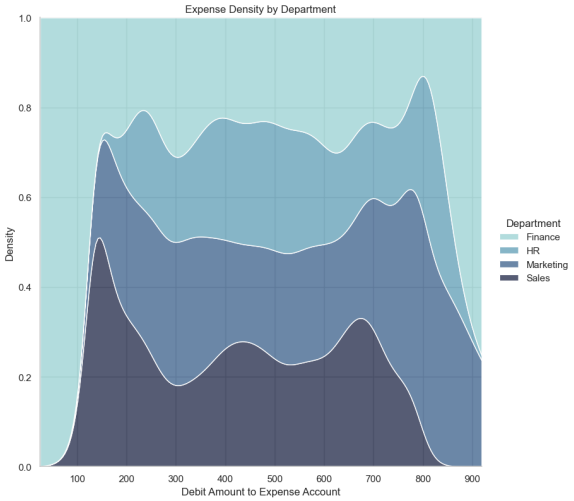
The credit amount in the cash and revenue accounts indicate strong financial performance and robust cash inflows as a result of successful financial activities over the accounting period. This pattern of credit transactions signifies effective revenue generation and cash management, pointing towards a healthy financial state as the company is successfully converting its operations into cash inflows, outweighing the cash outflows.

Department Transaction amounts are equally distributed between departments *(p=0.54)*



Analyzing revenue and expenses by department, we observe that the sales department has the highest ratio of high credits to the revenue account and the lowest ratio of high debits to the expense account.





These insights indicate that the sales department is the primary driver of the company's revenue as a highly effective source of income without incurring proportional expenses, leading to favorable profit margin.

4.3 Online Retail

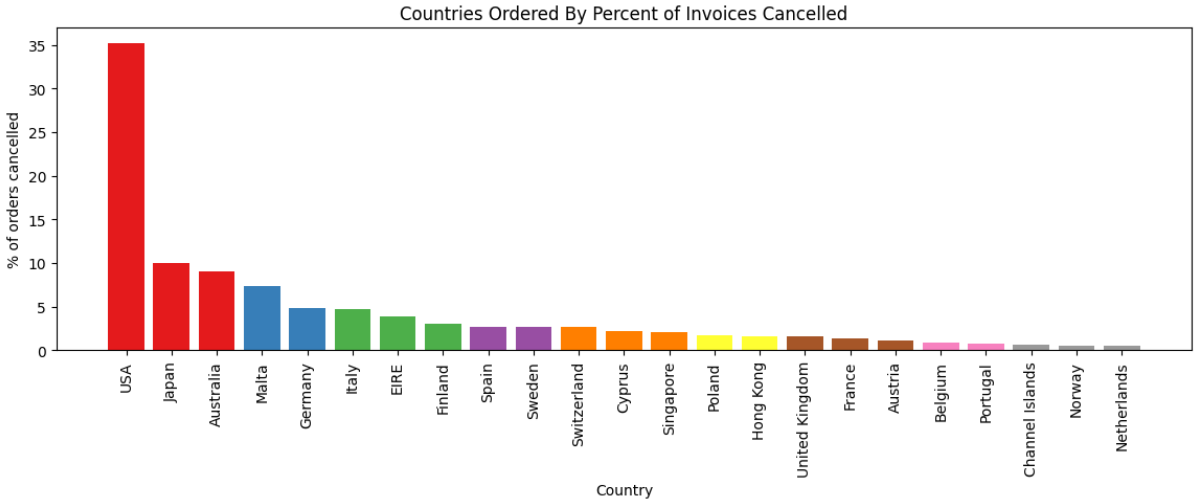
Most EDA done for Online Retail was based on questions one might ask to make business decisions, such as regional warehouse inventory, satisfaction, and unit pricing. Due to the sheer number of options in each column, categorical analysis was difficult to accomplish. Furthermore, the data was not normally distributed under any sort of aggregation or transformation.

Correlation between Quantity and Unit Price: *corr = -0.01915* This statistic checks if higher-priced items are bought in lower quantities. A slight negative value implies that yes, as unit price increased, quantity decreased. This coefficient is incredibly small, however. More information could offer more clarity, such as another Customer or Contracts table that outlines who exactly the Customer is, or recurring wholesale payments.

Correlation between Quantity and Unit Price - No UK: *corr = -0.0256* Same principle as above. Since the company in question is UK-based, and because the GBP is valued higher than most currencies, it is good to analyze the effects of price on quantity sold. Though the coefficient is greater than before, in the direction we'd assume it would go, it is still too small to reach any strong conclusions.

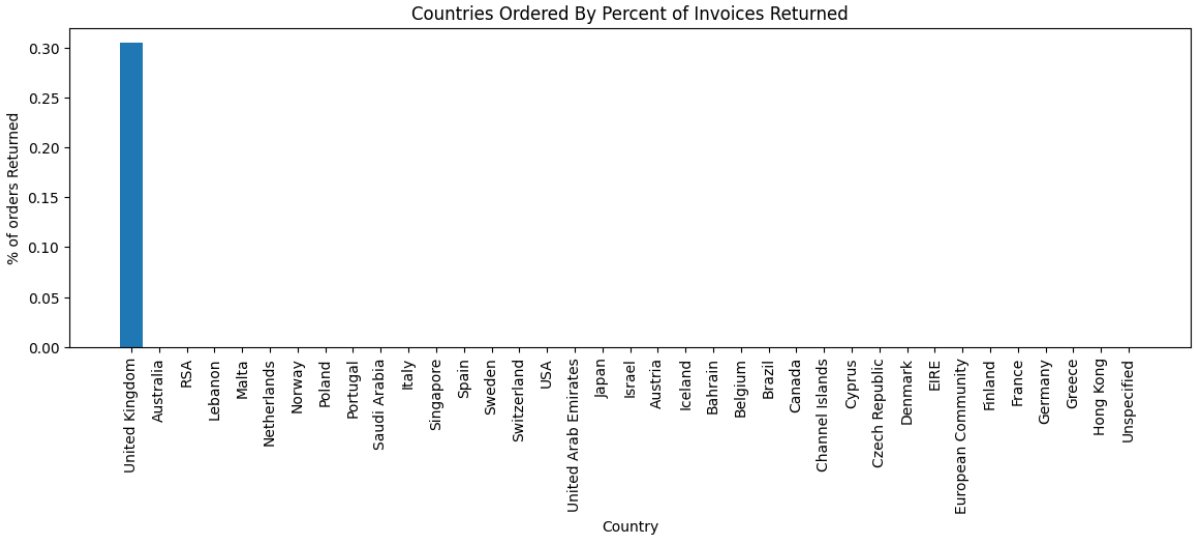
Is Unit Price a factor in cancellation rates: *t-statistic = 3.144, p-value = 0.0017* Comparing the means of the unit price in cancelled orders vs fulfilled orders, a positive statistic shows that the mean unit price of cancelled orders is higher than the mean unit price of fulfilled orders *(p=0.0017)*.

During the time period in our dataset, did some countries cancel more orders than others?



To answer this, we aggregated data by Country, InvoiceNo, and compared each group's percent of returned orders in comparison with all orders. The USA is shown to have canceled around 35% of all of their orders in 2022. Further analysis could be conducted with information such as customer satisfaction survey results or shipping delays.

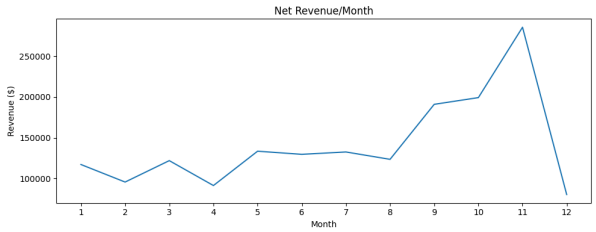
During the time period in our dataset, did some countries return more orders than others?



While this plot is not extremely detailed, it demonstrates the following points:

- Returns are only accepted domestically
- 30% of orders from the UK were returned.
- This does not consider quantity of items returned, but the count of InvoiceNo

Net income over time - aggregated by month - 2022



Item that, on average, was bought in the highest quantity per invoice

	StockCode	Description	Quantity
94986	21785	RAIN PONCHO	407

Weather reports in the Channel Island of Jersey from the month leading up to the invoice show an abundance of rain. Further knowledge of the Customer ID could reveal key information in understanding the item's popularity.

Item that, on average, was bought in the smallest quantity per invoice There were 24695 items that were bought, on average, in quantities of 1. A cursory look over showed no discernible pattern in product description, and there are not any components that could provide more insight as to why.

5 Machine Learning

The integration of machine learning capabilities in the accounting process enhances accuracy, efficiency, and strategic decision-making. One of the most impactful use cases of machine learning in this context is anomaly detection. Anomalies can range from simple data entry errors to complex fraud schemes, and by leveraging sophisticated algorithms, anomaly detection systems continuously scan and analyze transaction data, flagging any deviations from established patterns. These flagged transactions are then subjected to in-depth analysis by accountants in the loop.

We have trained various classification models on the financial accounting and managerial accounting data for classifying various column values. These classification models can be leveraged for anomaly detection. In this implementation, we create a dataframe of transactions for which the classification model is unable to assign a clear class with high confidence. If the highest probability for every class within a transaction is below a certain threshold, it is automatically flagged as an unusual transaction. Such transactions are flagged as potentially anomalous for further investigation.For further insights, view the ML notebooks in the repository.

Managerial Accounting Due to the unrealistic distributions in the datasets, the classification models are unable to detect insightful patterns. As a result, department, project and category classifiers all have 25% accuracy rates across the board, reflecting the observations in the EDA step. Given a realistic dataset, these classification models would be significantly more accurate and provide business value. Account and transaction type classifiers have 100% accuracy because they follow well-defined accounting principles per the ETL step. Clustering proved unsuccessful due to the unrealistic patterns in the data.

Financial Accounting Following the same process from the managerial accounting ML notebook, this dataset has observed moderate results in account and category classification, but requires realistic data for the other values. Clustering proved unsuccessful due to the unrealistic patterns in the data.

Online Retail Again, following the same process, the dataset was applied to different classification and clustering models. Here, the large size of the dataset was helpful. The Random Forest classifier effectively predicted Countries. Unfortunately, because most CustomerIDs only appeared around 1-3 times (with the exception of some, which most likely buy for resale), classification based on this feature was infeasible. Clustering was similarly difficult, due to the dataset having many distinct customer or product categories that appear few times. Anomaly detection could only be done with Country predictions, since stratification based on StockCode or Customer ID requires filtering out all entries that occur less than twice, which reduces our dataset enough to invalidate the whole point.

6 Automated Financial Reporting

Having ensured that the data is compliant with double-entry bookkeeping, flagged and analyzed unusual transactions, and gained insights regarding the company's financial performance over the accounting period, the final step is to close the accounts.

Balance Sheet We automate the balance sheet by leveraging the double-entry scripts from the ETL step and grouping transactions by account into the corresponding side of the accounting equation.

```
Assets:
              Account      Balance
1  Accounts Receivable  16831489.89
2        Asset Account 112424631.55
3                 Cash -67947371.23
5            Inventory   3935363.86

Liabilities:
             Account     Balance
0   Accounts Payable  1953886.10
6  Liability Account -3717127.19

Equity:
           Account      Balance
4  Expense Account -27082806.42
7  Revenue Account  81136930.02
8    Sales Revenue  12953231.56
```

The resulting balance sheet satisfies the accounting equation as all accounts are in their right place and transactions are balanced. Note the negative sign on the cash account as profits are credited to the cash account, and revenue in online retail is credited to revenue.

Income Statement Confirming the insights in the EDA step, the company has high operating profit margins resulting in high overall net income.

Category	Amount
Financial Total Revenue	11464357.68
Financial Total Expenses	3644634.44
Financial Net Income	7819723.24
Managerial Net Income	0.00
Online Retail Revenue	1559176.95
Overall Net Income	24487892.31

7 Conclusion

In this report we have utilized enterprise systems to organize data and gain deeper insights into the company's performance over the accounting period. By leveraging these insights we have implemented machine learning models to detect patterns unobservable by human accountants, streamlined the process and automated the close.