Task: Outline a detailed plan to optimize the data mart. Your plan should include the following elements:

1. Performance Analysis	**Methods and Tools** 1. **Data Profiling** o Used Python's pandas library to load and review data from Excel files. This helped identify the structure of the datasets, pinpoint missing values, and detect redundancy. o Example: A summary of missing values was generated to highlight data quality concerns. 2. **Database Inspection** o Leveraged SQLAlchemy to create and review a SQLite database, enabling the identification of unused or redundant columns and helping refine the database schema. 3. **Query Performance Benchmarking** o Ran sample queries to measure response times, particularly during peak usage hours. o Used database-specific tools to analyze query execution plans and identify bottlenecks in performance. **Bottlenecks Identified** • **Null Values**: Several columns contain missing values, impacting data completeness and the reliability of analyses. Handling these gaps during query execution increases processing time and decreases performance. • **Format of Values**: Inconsistent data formats, such as varying text cases in categorical fields or mixed data types in numerical columns, make queries inefficient and can lead to inaccurate results.

	• **Redundant Columns**: Initial analysis revealed multiple columns with duplicate or unnecessary data. These inflate the storage size, slow down query performance, and complicate schema design. • **ETL Frequency**: The current ETL process runs every three hours, which does not satisfy stakeholders' need for near real-time data updates. This delay affects decision-making processes that depend on up-to-date information.
2. Optimization Strategy	**Improving Query Performance** 1. **Schema Refinement** o Removed redundant columns to streamline data storage. o Normalized tables to reduce duplication and improve data access efficiency. 2. **Indexing** o Added indexes to fields commonly used in queries, significantly reducing lookup times. 3. **Partitioning** o Organized tables into partitions, such as by month or region, to make queries targeting specific data subsets faster. **ETL Optimization** 1. **Incremental Updates** o Revised the ETL process to update only changed data instead of reloading everything. o Example: Applied SQL MERGE statements to efficiently manage data changes. 2. **Parallel Processing** o Implemented parallel processing pipelines using tools like Apache Airflow or AWS Glue, reducing ETL runtime. 3. **Optimized Transformation Logic**

	◦ Focused on lightweight transformations during data ingestion and reserved complex calculations for downstream analysis.
3. Data Governance and Quality	**Ensuring Data Quality** • Addressed missing values based on column types: ◦ **Categorical Data**: Filled gaps with "Unknown" to maintain interpretability. ◦ **Numerical Data**: Used the mean value to fill missing entries, ensuring analytical reliability. ◦ **Other Types**: Defaulted missing values to 0 for completeness. • Removed duplicate records to ensure data accuracy and prevent inflated storage. **Governance Plan** 1. **Metadata Management** ◦ Created a data catalog to document table structures, column purposes, and relationships across the data mart. 2. **Access Control** ◦ Defined roles for users, ensuring that data access aligns with individual needs while safeguarding sensitive information. 3. **Version Control** ◦ Tracked schema changes using a version control system, enabling accountability and simplifying rollback when needed.
4. Monitoring and Maintenance	**Monitoring Setup** • Standardized column names (e.g., replaced spaces with underscores and converted text to lowercase) to improve query consistency and reduce potential errors.

	• Ensured data cleaning produced structured and predictable formats, facilitating smooth operation of monitoring tools. **Ongoing Maintenance** • Regularly reviewed the schema and ETL workflows to adapt to changing business needs. • Applied software patches and updates to keep the system secure and efficient. • Leveraged cloud-based auto-scaling to handle increased demand during peak hours while minimizing costs during off-peak times

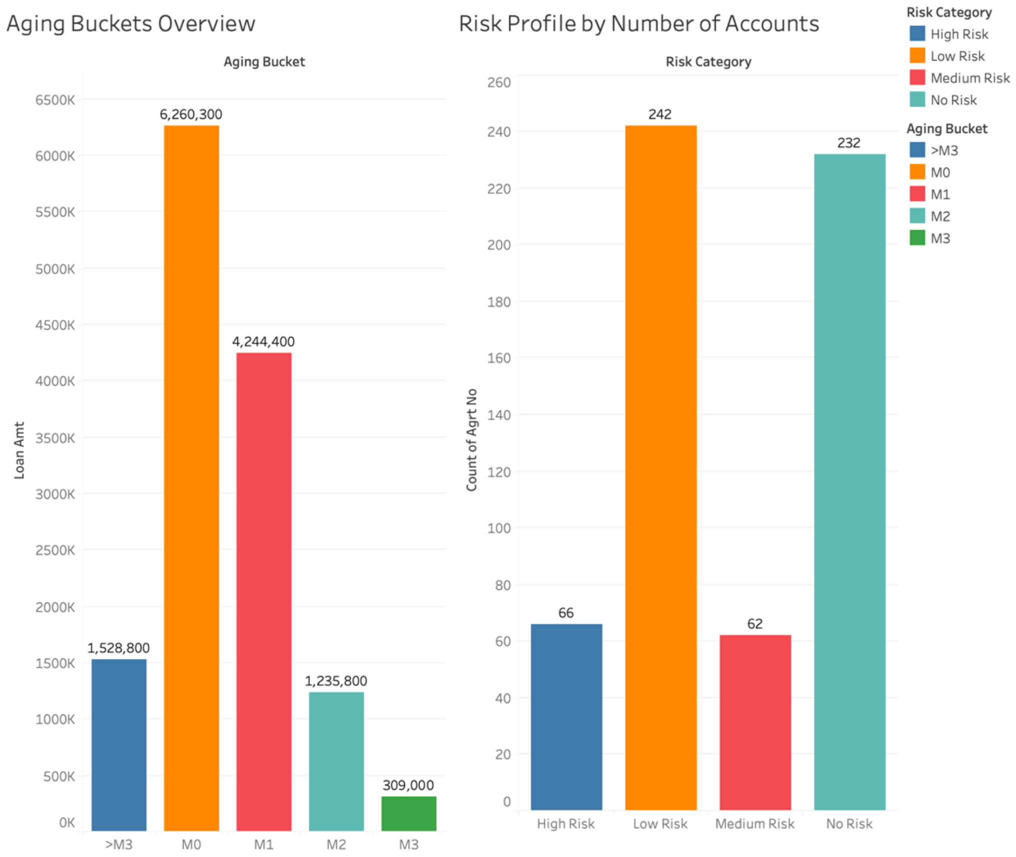
Scenario: You are providing within 3 data sets – Genie Dataset 1,Genie Dataset 3 and Genie Dataset 4

Task 1: Convert both excel sheet into SQL database for analysis purpose. [Hint: Using python to conduct data processing and insert processed data into SQL database]. Candidate required to submit result via GitHub or documentation form for assessment purposes.

Kindly refer to the Task 1.ipynb file and the genie_data.db database uploaded on GitHub.

https://github.com/YongTeng/Genie-Financial-Services-Malaysia-/tree/main

Task 2: Generate hp aging bucket by number of account and general ledger balance snapshot based on Genie Dataset 1 and Genie Dataset 4. Aging buckets refer to category M0 : 0 days, category M1 : 1 – 30 days, category M2 : 31– 60 days, category M3 : 61 – 90 days, category > M3 : > 90 days. Aging is defined based on days different between reporting day and contract commencement date. Candidate could provide the result in python format or using any BI Tools.


Aging Buckets Overview

Risk Profile by Number of Accounts

Task 3: Based on HP OS, what is company risk profile and illustrate current account performance summary ie. Current active loan holder paid installment in term of month and installment period position.

Company Risk Profile and Account Performance

Aging Buckets and Risk Profile Analysis

The loans are grouped into aging buckets based on their overdue status:

- **M0 (0 days overdue)**: This category represents the largest segment, with loans totaling approximately 6.26 million. These loans are current, indicating strong cash flow and minimal risk.
- **M1 (1–30 days overdue)**: Loans in this group total 4.24 million. They represent accounts with minor repayment delays, classified as medium risk.
- **M2 (31–60 days overdue)**: This group, totaling 1.24 million, reflects a smaller but noticeable portion of delayed repayments.
- **M3 (61–90 days overdue)**: This is the smallest overdue category, with loans totaling 309,000. These accounts are being closely monitored due to higher repayment risks.
- **>M3 (Over 90 days overdue)**: Loans in this category amount to 1.52 million, reflecting high-risk accounts that may need intervention or recovery efforts.

Risk Categorization by Number of Accounts

The “Risk Profile by Number of Accounts” chart provides insights into the company’s risk exposure:

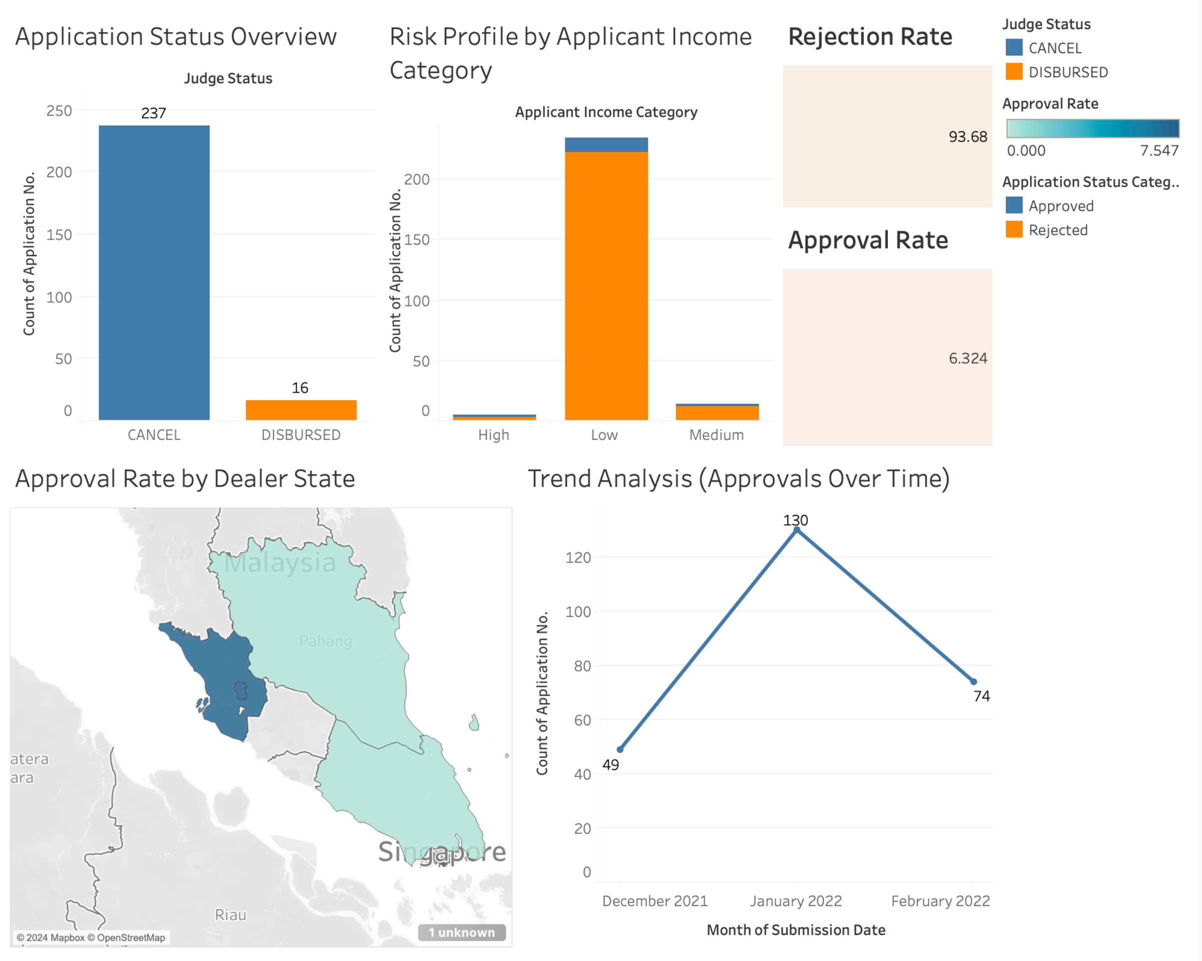
- **Low Risk**: The majority of loans—242 accounts—fall into this category, reflecting efficient loan management and minimal overdue issues.
- **Medium Risk**: A total of 62 accounts show manageable but concerning delays, mainly in the M1 category.
- **High Risk**: There are 66 accounts in the high-risk category with overdue periods exceeding 90 days (>M3). These require focused recovery efforts.
- **No Risk**: Another 232 accounts are fully paid or current, contributing to the portfolio's overall stability.

Conclusion

The company’s risk profile appears well-controlled, with most active loans categorized as low or no risk. While there are overdue accounts across M1 to >M3, these are manageable, and only a small percentage fall into the high-risk category. Overall, the company demonstrates strong repayment performance, with most borrowers adhering to installment schedules. Strategic monitoring and timely action for high-risk accounts can further strengthen financial stability and improve portfolio health.

Scenario: You are providing Genie Dataset 2. Can you provide your insights and recommendations inline with the following questions

Question 1: What do you think company underwriting risk profile? Please elaborate your thought and how we could improve from current situation.



Underwriting Risk Profile and Recommendations

Underwriting Risk Overview

The company's current underwriting process shows significant challenges, with a high rejection rate of 93.68% and a low approval rate of just 6.32%. Out of 253 applications reviewed, only 16 were approved, while 237 were canceled. This indicates a conservative or overly stringent approach that may be hindering the number of loans being disbursed. The high number of cancellations suggests that either the rejection criteria are too rigid, or there are inefficiencies in the application or approval process.

Income Category Risk Analysis

A closer look at rejected applications reveals that a significant portion comes from the Low-Income category, while applicants in the Medium- and High-Income groups make up only a small percentage. This trend suggests that the underwriting process may be disproportionately restrictive toward lower-income borrowers, likely due to concerns about their ability to repay. While this approach may reduce risk, it also limits the company's market reach and loan distribution potential.

Approval Trends and Dealer State Analysis

Approval rates have fluctuated over time, with a peak of 130 approvals in January 2022, followed by a sharp decline to 74 in February 2022. This inconsistency could point to changes in underwriting standards or seasonal variations in application quality. Additionally, analysis by dealer state reveals disparities in approval rates across different regions, highlighting areas where the company could focus on improving performance.

Approval Rate by Dealer State

The approval rate analysis by dealer state reveals notable regional differences in performance. Kuala Lumpur and Selangor demonstrate significantly higher approval rates compared to Johor and Pahang. This suggests that applications originating from dealers in Kuala Lumpur and Selangor are more likely to meet underwriting criteria or possess stronger applicant profiles. Conversely, the lower approval rates in Johor and Pahang may indicate potential issues such as higher risk applicants, incomplete documentation, or inconsistencies in the application process within these regions.

Recommendations

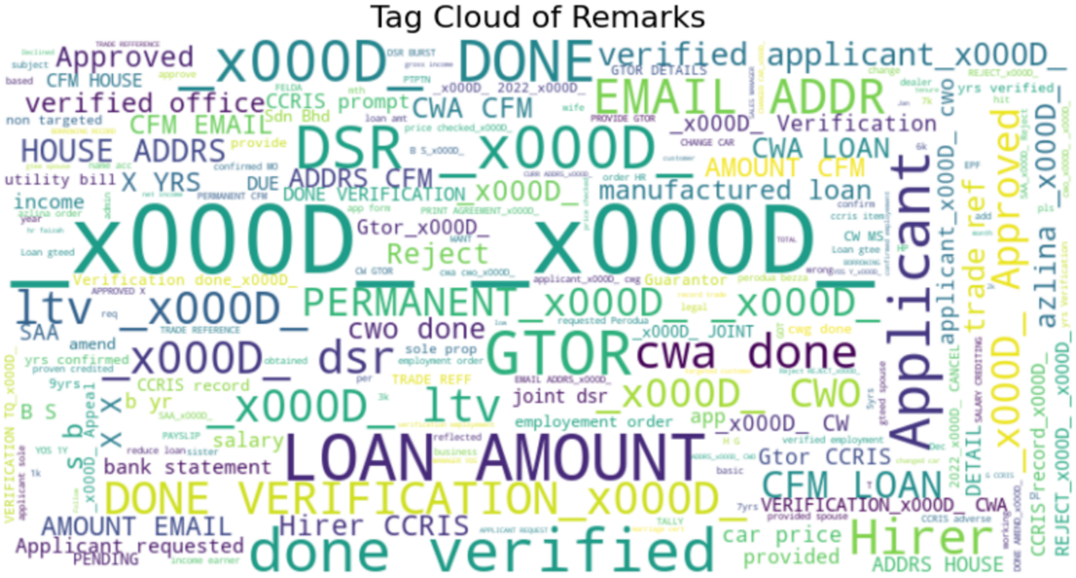
To improve approval rates and better support lower-income applicants, it's important to take a closer look at the underwriting process. This involves reassessing the risk assessment framework and adopting alternative ways to evaluate creditworthiness. For example, instead of relying solely on traditional credit scores, factors like consistent employment history or regular payment of utilities could provide a more complete picture of an applicant’s financial reliability.

In Malaysia, understanding income brackets like the B40 and B20 groups is vital. The B40 group earns a median income of RM3,440, with an upper range of RM5,249, while the B20 subset earns between RM2,501 and RM3,169. These figures highlight the need for more inclusive lending criteria that cater specifically to these groups.

Another way to improve access to loans is by enhancing risk-based pricing. By introducing tiered loan structures or flexible pricing models, lenders can offer interest rates and repayment terms tailored to the applicant's risk profile. This approach ensures that lower-income borrowers have access to loans that are affordable and manageable, while still balancing the lender’s risk.

Finally, lenders can analyze data from canceled and rejected applications to identify patterns or common reasons for rejection. This can help simplify the application process, address unnecessary roadblocks, and reduce cancellation rates. By making these changes, the underwriting process can become not only more effective but also more inclusive, giving a fair chance to applicants from lower-income backgrounds.

```
Number of positive sentiment remarks: 63
Number of negative sentiment remarks: 49
Number of neutral sentiment remarks: 141
```



Sentiment Analysis of Car Loan Judgments: Key Insights and Influencing Factors When looking at the sentiment analysis and word cloud of remarks, it's clear that several factors play a role in determining car loan outcomes, such as whether the loan is approved or canceled. Positive terms like "Approved," "verified," "DONE," and "PERMANENT" suggest that loan approval often depends on thorough verification processes, stable employment, and reliable income. This implies that applicants who provide the necessary documentation—such as proof of income, employment verification, and residential details—are more likely to have their loans approved.

On the other hand, negative terms like "Reject," "pending," and repeated mentions of "x000D" may point to incomplete verifications or unresolved issues, which could lead to loan cancellations. Terms like "DSR" (Debt Service Ratio) and "CCRIS" indicate that poor financial ratios or negative credit records could also contribute to rejections. Ultimately, the decision seems to hinge on verifying the applicant's financial stability and ensuring they meet the lender's requirements.

Appendix

New Calculated Field Added in Tableau for Visualization

Dataset 2	
1. Approval Rate	(COUNT(IF [Judge Status] = "DISBURSED" THEN 1 END) / COUNT([Application No.]))*100

2. Applicant Income Category	IF [G Net Salary] < 2500 THEN "Low"
	ELSEIF [G Net Salary] >= 2500 AND [G Net Salary] < 5000 THEN "Medium"
	ELSE "High"
	END

3. Application Status Category	IF [Judge Status] = "DISBURSED" THEN "Approved"
	ELSEIF [Judge Status] = "CANCEL" THEN "Rejected"
	ELSE "Pending"
	END

Dataset 4	

4. Aging Bucket	IF [Mth Due] = 0 THEN 'M0'
	ELSEIF [Mth Due] <= 1 THEN 'M1'
	ELSEIF [Mth Due] <= 2 THEN 'M2'
	ELSEIF [Mth Due] <= 3 THEN 'M3'
	ELSE '>M3'
	END

5. Risk Category	IF [Arrears] = 0 THEN 'No Risk'
	ELSEIF [Arrears] <= [Mthly Instal] THEN 'Low Risk'
	ELSEIF [Arrears] <= [Mthly Instal] * 3 THEN 'Medium Risk'
	ELSE 'High Risk'
	END