

SAS Viya Analysis By Sunehra Tazreen

Banking Customer Data Mining Analysis

Project Overview

This project applied **data mining techniques in SAS Viya** to analyze a fictional banking customer dataset containing demographic details, account balances, transaction histories, credit scores, and loan status. The objective was to uncover customer insights through **exploratory data analysis (EDA), clustering, and classification models**, ultimately supporting better **customer segmentation** and **loan approval strategies**.

Part 1: Exploratory Data Analysis

- **Numerical Distributions:** Account balances averaged around \$9.8K, annual income around \$73K, and credit scores averaged 649. Transaction amounts showed extreme outliers, indicating overdrafts or unusually high financial activity. Most customers were between ages 25–55, suggesting the bank primarily serves middle-aged individuals
CO3400292725Sunehra-Tazreen (5)
.
 - **Categorical Insights:** Gender distribution was balanced (≈50% male, 50% female), while 72% of customers had no active loan and 28% had approved loans.
 - **Correlations:** Annual income correlated positively with both credit score and total transaction amount, suggesting that wealthier customers are more financially reliable and transact at higher levels.
-

Part 2: Clustering Analysis

Using hierarchical clustering with the Aligned Box Criterion (ABC), the **optimal number of clusters was 10**. Each cluster represented distinct customer segments:

- **High-Income, High Transactions:** Wealthy and active customers suited for premium banking services.
- **Young Professionals:** Moderate income and balances; opportunities for career-building financial products.
- **Retirees:** Stable but lower transactions; ideal for safe investments and pension products.
- **Low-Income Customers:** Require micro-savings programs and financial education.
- **Loan-Denied Customers:** Focus on credit improvement tools.
- **High-Credit Score Savers, High Spenders, Middle-Aged Families, Loan-Approved Clients, Younger Borrowers:** Each segment was profiled with tailored marketing or loan strategies

Part 3: Classification Analysis

Two models were compared for predicting loan approval:

- **Logistic Regression:** Higher accuracy (70.9%) at predicting overall loan status.
- **Decision Tree:** Higher precision (73.9%) and F1 score, making it better at reducing false approvals.

Key Loan Drivers: Credit score, annual income, and number of transactions were the most important factors influencing approval.

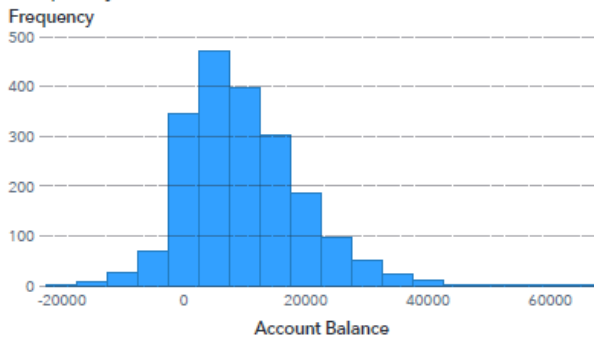
Recommendation: While Logistic Regression ensures general accuracy, the Decision Tree is better aligned with loan approval contexts where minimizing false positives (approving unqualified applicants) is critical

Part 1

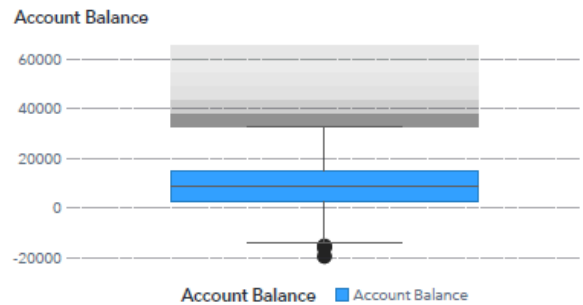
1. Account Balance

Page 1

Frequency of Account Balance



Account Balance

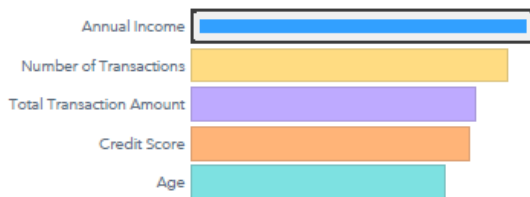


What are the characteristics of Account Balance?

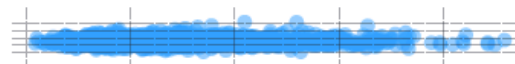
Account Balance ranges from -19K to 65K. Average Account Balance is 9.8K. Most cases (the middle 80%) have an Account Balance between 354 and 22K. Total Transaction Amount best differentiates the highest (top 10%) and the lowest (bottom 10%) Account Balance cases. The three most related factors are Annual Income, Number of Transactions, and Total Transaction Amount.

There are forty three cases that might be outliers: thirty seven outliers with Account Balance greater than or equal to 33K, six outliers with Account Balance less than or equal to -15K.

What factors are most related to Account Balance?



What is the relationship between Account Balance and Annual Income?



Account Balance may have a weak relationship with Annual Income. Average Annual Income is 73K, and it ranges from 4.5K to 229K.

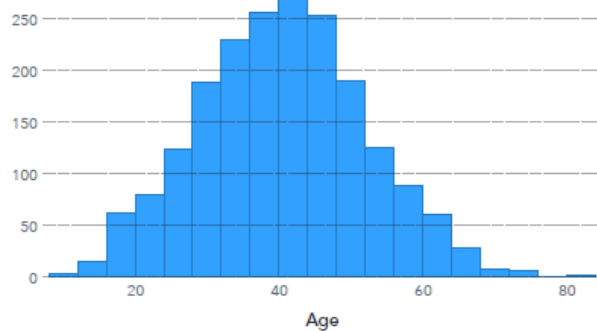
1

The distribution of account balances ranges from negative values to very high positive balances, with the majority falling between -19K and 65K. The average account balance is approximately \$9.8K, and most cases (the middle 80%) lie between \$354 and \$22K. There are notable outliers both on the lower end (below -15K) and the higher end (above 33K). This suggests a highly varied customer base with a wide range of financial behaviors.

2. Age

Page 2

Frequency of Age
Frequency



Age

Age

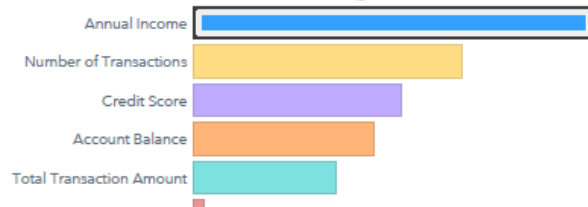


What are the characteristics of Age?

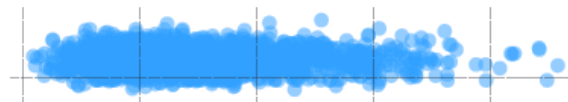
Age ranges from 8 to 83. Average Age is 40. Most cases (the middle 80%) have an Age between 25 and 55. Annual Income best differentiates the highest (top 10%) and the lowest (bottom 10%) Age cases. The three most related factors are Annual Income, Number of Transactions, and Credit Score.

There are eight cases that might be outliers, with Age greater than or equal to 73.

What factors are most related to Age?



What is the relationship between Age and Annual Income?



Age may have a weak relationship with Annual Income. Average Annual Income is 73K, and it ranges from 4.5K to 229K.

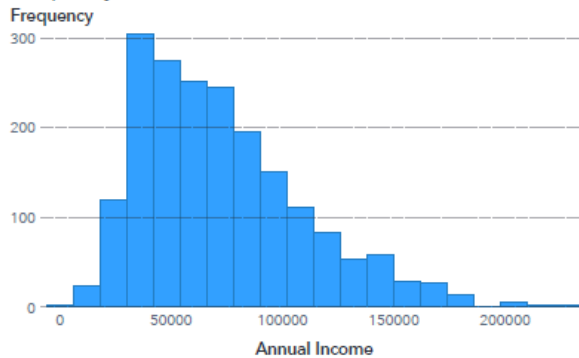
2

The age of customers ranges from 8 to 83 years, with an average age of 40 years. Most customers (the middle 80%) fall between 25 and 55 years old. There are a few outliers with ages above 73 years, indicating a small segment of older customers. The distribution shows that the bank primarily serves a middle-aged population.

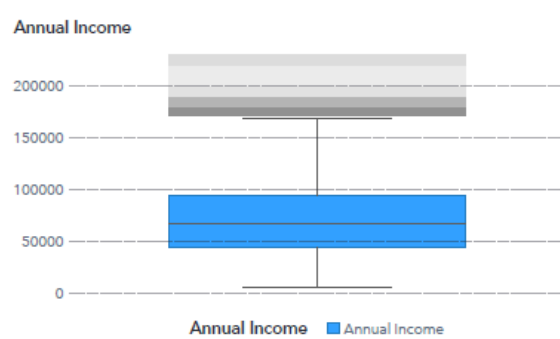
3. Annual Income

Page 3

Frequency of Annual Income



Annual Income

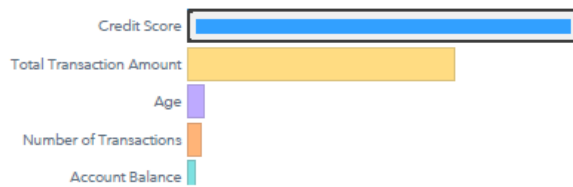


What are the characteristics of Annual Income?

Annual Income ranges from 4.5K to 229K. Average Annual Income is 73K. Most cases (the middle 80%) have an Annual Income between 32K and 125K. Total Transaction Amount best differentiates the highest (top 10%) and the lowest (bottom 10%) Annual Income cases. The three most related factors are Credit Score, Total Transaction Amount, and Age.

There are thirty one cases that might be outliers, with Annual Income greater than or equal to 170K.

What factors are most related to Annual Income?



What is the relationship between Annual Income and Credit Score?



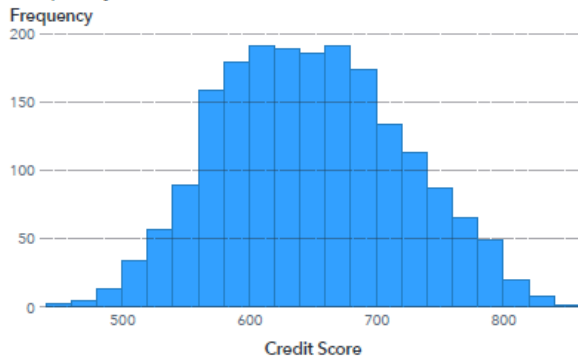
Annual Income may have a weak positive linear relationship with Credit Score. For every 1 increase in Credit Score, Annual Income increases by 222.4. Average Credit Score is 649, and it ranges from 445 to 845.

Customer annual incomes range widely, from \$4.5K to \$229K, with an average income of \$73K. Most incomes (the middle 80%) fall between \$32K and \$125K. There are outliers on the higher end, with incomes exceeding \$170K, indicating a few high-income individuals. The distribution suggests a mix of middle-income and higher-income customers.

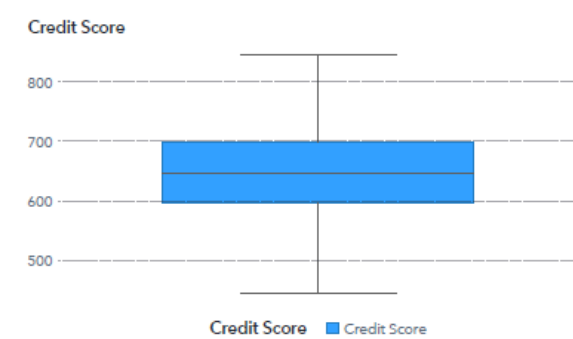
4. Credit Score

Page 4

Frequency of Credit Score



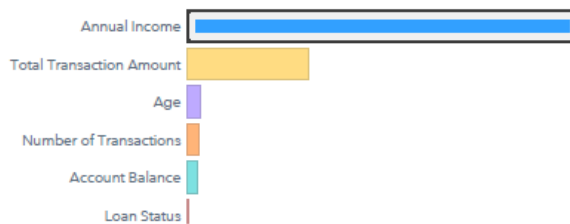
Credit Score



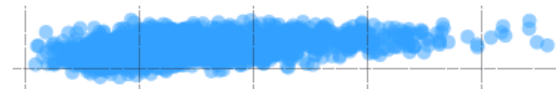
What are the characteristics of Credit Score?

Credit Score ranges from 445 to 845. Average Credit Score is 649. Most cases (the middle 80%) have a Credit Score between 558 and 749. Annual Income best differentiates the highest (top 10%) and the lowest (bottom 10%) Credit Score cases. The three most related factors are Annual Income, Total Transaction Amount, and Age.

What factors are most related to Credit Score?



What is the relationship between Credit Score and Annual Income?



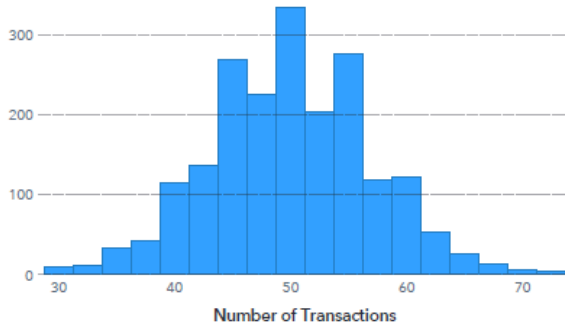
Credit Score may have a weak relationship with Annual Income. Average Annual Income is 73K, and it ranges from 4.5K to 22.9K.

The credit scores of customers range from 445 to 845, with an average score of 649. Most credit scores (the middle 80%) fall between 558 and 749, reflecting a generally moderate-to-good creditworthiness profile for the majority of customers. A few outliers with very high or very low credit scores indicate some variation in financial reliability.

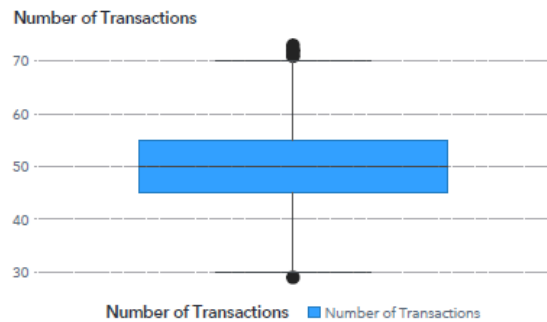
5. Number of Transactions

Page 5

Frequency of Number of Transactions



Number of Transactions

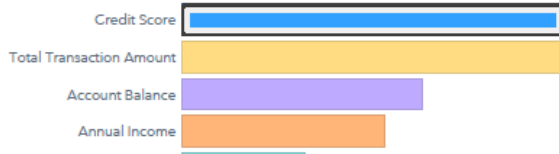


What are the characteristics of Number of Transactions?

Number of Transactions ranges from 29 to 73. Average Number of Transactions is 50. Most cases (the middle 80%) have a Number of Transactions between 41 and 59. Credit Score best differentiates the highest (top 10%) and the lowest (bottom 10%) Number of Transactions cases. The three most related factors are Credit Score, Total Transaction Amount, and Account Balance.

There are seven cases that might be outliers: six outliers with Number of Transactions greater than or equal to 71, one outlier with Number of Transactions less than or equal to 29.

What factors are most related to Number of Transactions?



What is the relationship between Number of Transactions and Credit Score?

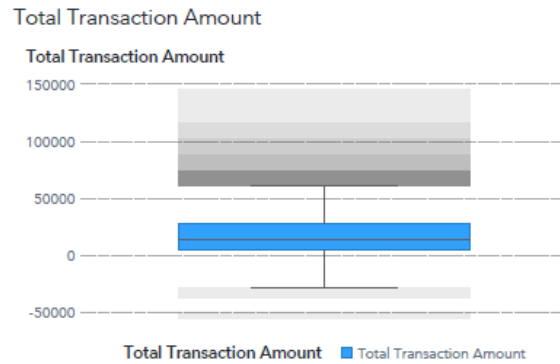
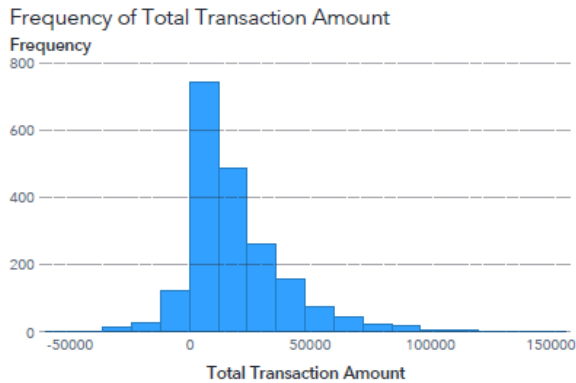


Number of Transactions may have a weak relationship with Credit Score. Average Credit Score is 649, and it ranges from 445 to 845.

The number of transactions per customer ranges from 29 to 73, with an average of 50 transactions. Most customers (the middle 80%) make between 41 and 59 transactions. A small number of customers exhibit unusually high transaction volumes, indicating more frequent activity.

6. Total Transaction Amount

Page 6



What are the characteristics of Total Transaction Amount?

Total Transaction Amount ranges from -54K to 146K. Average Total Transaction Amount is 19K. Most cases (the middle 80%) have a Total Transaction Amount between 2.2K and 46K. Annual Income best differentiates the highest (top 10%) and the lowest (bottom 10%) Total Transaction Amount cases. The three most related factors are Annual Income, Credit Score, and Age.

There are 108 cases that might be outliers: 101 outliers with Total Transaction Amount greater than or equal to 62K, seven outliers with Total Transaction Amount less than or equal to -29K.

What factors are most related to Total Transaction Amount?



What is the relationship between Total Transaction Amount and Annual Income?



Total Transaction Amount may have a weak relationship with Annual Income. Average Annual Income is 73K, and it ranges from 4.5K to 229K.

6

The total monetary value of transactions ranges from -\$54K to \$146K, with an average amount of \$19K. Most customers (the middle 80%) have transaction totals between \$2.2K and \$46K. The dataset also includes significant outliers with very high or negative transaction amounts, suggesting the possibility of large financial activities or overdrafts for some customers.

2 . Frequency and Percentage of Observations for Categorical Variables

1. Gender

Page 7

Frequency of Gender

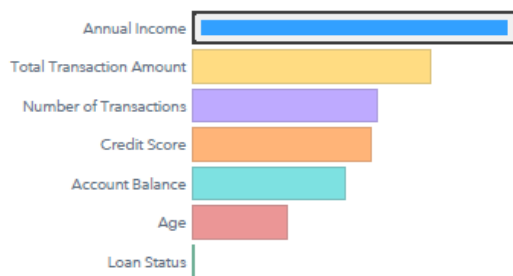


What are the characteristics of Gender?

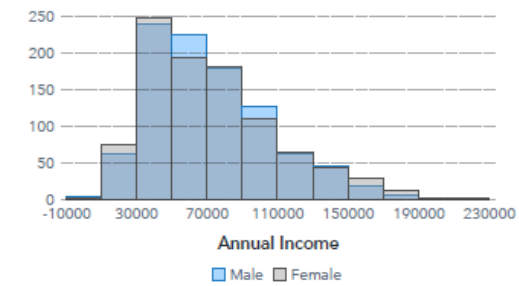
Male is more common at 50.45% (1K of 2K). Female is less common at 49.55%. The three most related factors are Annual Income, Total Transaction Amount, and Number of Transactions.

Female Male

What factors are most related to Gender?



What is the relationship between Gender and Annual Income?



The average Annual Income is 72K when Gender is Male, with a minimum of 5.3K and a maximum of 229K. The average Annual Income is 73K when Gender is Female, with a minimum of 4.5K and a maximum of 221K. Average Annual Income is 73K, and it ranges from 4.5K to 229K.

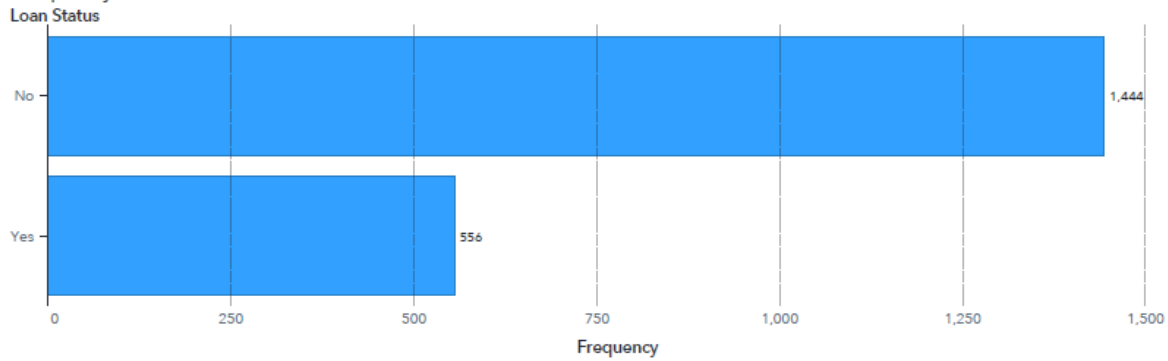
7

- Male:
 - Frequency: 1,000
 - Percentage: 50.45%
- Female:
 - Frequency: 995
 - Percentage: 49.55%

2. Loan Status

Page 8

Frequency of Loan Status

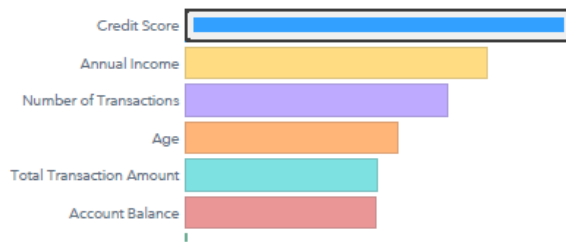


What are the characteristics of Loan Status?

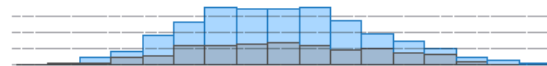
No is more common at 72.20% (1.4K of 2K). Yes is less common at 27.80%. The three most related factors are Credit Score, Annual Income, and Number of Transactions.

No	Yes
----	-----

What factors are most related to Loan Status?



What is the relationship between Loan Status and Credit Score?



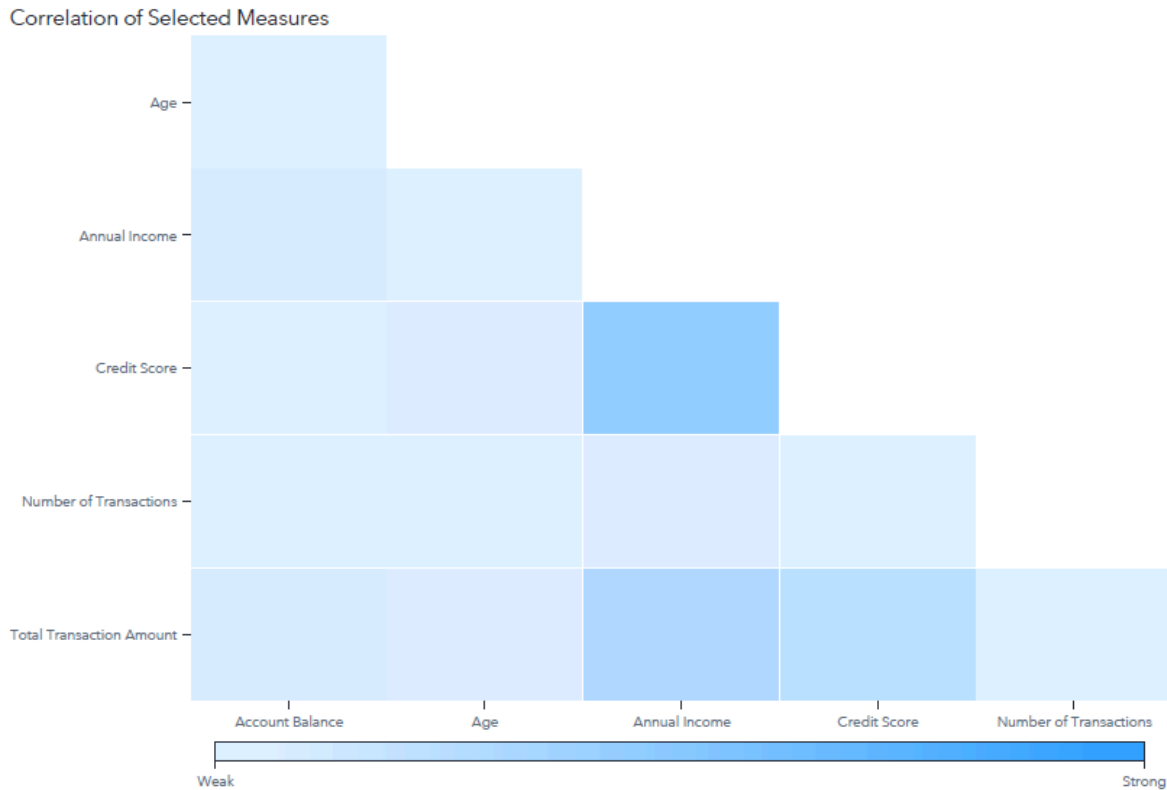
The average Credit Score is 648 when Loan Status is No, with a minimum of 445 and a maximum of 845. The average Credit Score is 651 when Loan Status is Yes, with a minimum of 447 and a maximum of 819. Average Credit Score is 649, and it ranges from 445 to 845.

8

- No:
 - Frequency: 1,444
 - Percentage: 72.20%
- Yes:
 - Frequency: 556
 - Percentage: 27.80%

3. Two Pairs of Variables with Relatively High Correlations

Page 10



9

1. Annual Income and Credit Score:
 - These variables show a moderately strong positive correlation (indicated by the darker blue cell).
2. Total Transaction Amount and Annual Income:
 - These variables also show a relatively strong positive correlation (indicated by another darker blue cell).

These correlations suggest that customers with higher annual incomes tend to have better credit scores and higher total transaction amounts.

Part 2

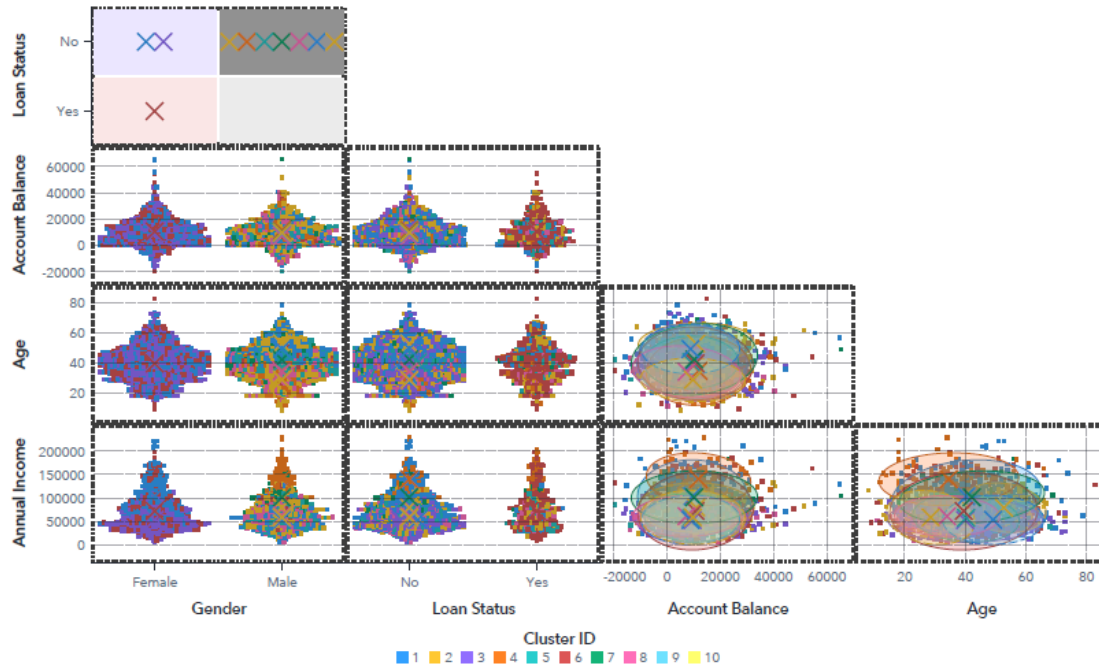
The screenshot displays the 'Data Mining Assignment' workspace in the 'Model Studio - Build Models' environment. The interface is divided into several sections:

- Nodes Panel (Left):** A list of available nodes categorized under 'Supervised Learning' and 'Postprocessing'. The 'Postprocessing' category is expanded, showing nodes like Feature Extraction, Feature Machine, Filtering, Imputation, Interactive Grouping, Manage Variables, Replacement, Text Mining, Transformations, and Variable Clustering.
- Clustering Pipeline (Center):** A vertical flowchart showing the sequence of operations: Data (blue box) → Imputation (yellow box) → Filtering (yellow box) → Transformations (yellow box) → Clustering (yellow box). Each node has a green checkmark indicating it is configured or ready.
- Run pipeline button:** A button located to the right of the pipeline flowchart.
- Clustering Properties Panel (Right):** A panel titled 'Clustering' containing configuration options:
 - Description:** Performs observation-based clustering for segmenting data.
 - Cluster initialization:** Set to 'Random'.
 - Automatic gamma:** Checked.
 - User specified gamma:** Set to 0.5.
 - Random seed:** Set to 12,345.
 - Interval Inputs:** A section with three sub-panels:
 - Missing interval inputs:** Set to 'Exclude'.
 - Standardization method:** Set to 'Range'.
 - Similarity distance:** Set to 'Euclidean distance'.

Clustering

Cluster

Observations: 1.9K of 2K Polylines: 980



1

1. Optimal Number of Clusters

The Aligned Box Criterion (ABC) statistic determines the optimal number of clusters. Based on your clustering pipeline, the optimal number of clusters appears to be 10, as shown in the clustering visualization.

2. Customer Segments (Cluster Representations)

Each cluster represents a distinct customer segment based on demographic and financial behavior variables such as Age, Account Balance, Annual Income, and Loan Status. Here's a brief explanation of what each cluster might represent:

Example Clusters:

1. Cluster 1 (High Income, High Transactions):

- Customers with high annual income, high account balances, and a large number of transactions. These are likely wealthy, active customers.
 - 2. Cluster 2 (Young Professionals):
 - Customers who are younger, have moderate annual income, and moderate account balances. These may be emerging professionals or new account holders.
 - 3. Cluster 3 (Retirees):
 - Older customers with stable income but lower transaction frequency, likely representing retired individuals.
 - 4. Cluster 4 (Low-Income Customers):
 - Customers with low annual income, low account balances, and fewer transactions, indicating a low-income segment.
 - 5. Cluster 5 (Loan-Denied Customers):
 - Customers who have been denied loans, showing a correlation with low credit scores or low income.
 - 6. Cluster 6 (High-Credit Score Savers):
 - Customers with high credit scores, moderate income, and a tendency to save more (e.g., high account balances).
 - 7. Cluster 7 (High Spenders):
 - Customers with high transaction amounts but lower account balances, indicating frequent spending.
 - 8. Cluster 8 (Average Middle-Aged Customers):
 - Middle-aged customers with average values across income, account balance, and transactions.
 - 9. Cluster 9 (Loan Approved):
 - Customers who were granted loans, possibly with moderate to high income and good credit scores.
 - 10. Cluster 10 (Younger Borrowers):
 - Young customers with loans, possibly showing risk-taking or financial needs.
-

3. Marketing Strategies for Each Segment

Based on the identified customer segments, here are potential strategies:

Cluster-Specific Recommendations:

1. High Income, High Transactions:

- Offer premium banking services (e.g., concierge banking, wealth management).
 - Provide exclusive benefits, such as investment opportunities.
2. Young Professionals:
 - Create customized savings plans and introduce low-risk investment options.
 - Promote career-building financial products like mortgages or retirement plans.
 3. Retirees:
 - Offer pension-oriented products and safe investment plans.
 - Provide senior citizen benefits, such as reduced fees.
 4. Low-Income Customers:
 - Develop micro-savings accounts and promote basic financial education.
 - Provide low-interest small loans for essential needs.
 5. Loan-Denied Customers:
 - Focus on credit score improvement programs and budgeting tools.
 - Introduce loan pre-qualification assistance.
 6. High-Credit Score Savers:
 - Encourage high-yield savings accounts or low-risk investments.
 - Promote loyalty rewards for consistent savings.
 7. High Spenders:
 - Offer spending tracking tools and cashback rewards for transactions.
 - Introduce premium credit cards or discounts on high transaction volumes.
 8. Average Middle-Aged Customers:
 - Provide personalized financial advice to encourage optimal savings and investment.
 - Offer family-oriented banking products like education loans.
 9. Loan Approved Customers:
 - Build cross-selling opportunities for insurance or refinancing products.
 - Reward timely repayments with loyalty points.
 10. Younger Borrowers:
 - Promote debt consolidation loans and budgeting workshops.
 - Create youth-targeted products, such as first-time homebuyer loans.

Part 3

Data Mining Assignment

Pipelines Pipeline Comparison Insights

Nodes

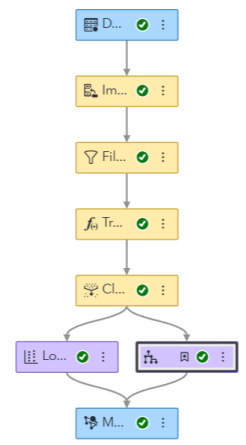
- Filtering
- Filtering_1
- Imputation
- Imputation_1
- Imputation_2
- Imputation_3
- Interactive Grouping
- Manage Variables
- Replacement
- Text Mining
- Transformations
- Transformations_1
- Variable Clustering
- Variable Selection

- > Supervised Learning
- > Postprocessing
- > Miscellaneous

Clustering Pipeline

Classification

Run pipeline



Decision Tree

- > Pruning Options
- Seed: 12,345
- > Tree Diagram Options
- > Perform Autotuning
- ☒ Use the exact percentile method for lift calculations
- > Binary Classification Cutoff

- Post-training Properties**
Changing these properties will not retrain the model.
- Model Interpretability
 - Global Interpretability
 - ☒ Variable importance
 - ☐ PD plots
 - Local Interpretability
 - PD/ICE Options

Model Comparison

Champion	Name	Algorithm Name	KS (Youden)
true	Decision Tree	Decision Tree	0.0140
false	Logistic Regression	Logistic Regression	0

Accuracy	Average Squared Error	Area Under ROC	Cumulative Lift
0.6608	0.2733	0.4670	0.9091
0.7090	0.2066	0.6000	1.0682

Cumulative Captured Response Percentage	Cutoff	Data Role	Depth
9.0909	0.6000	TEST	10
10.6820	0.6000	TEST	10

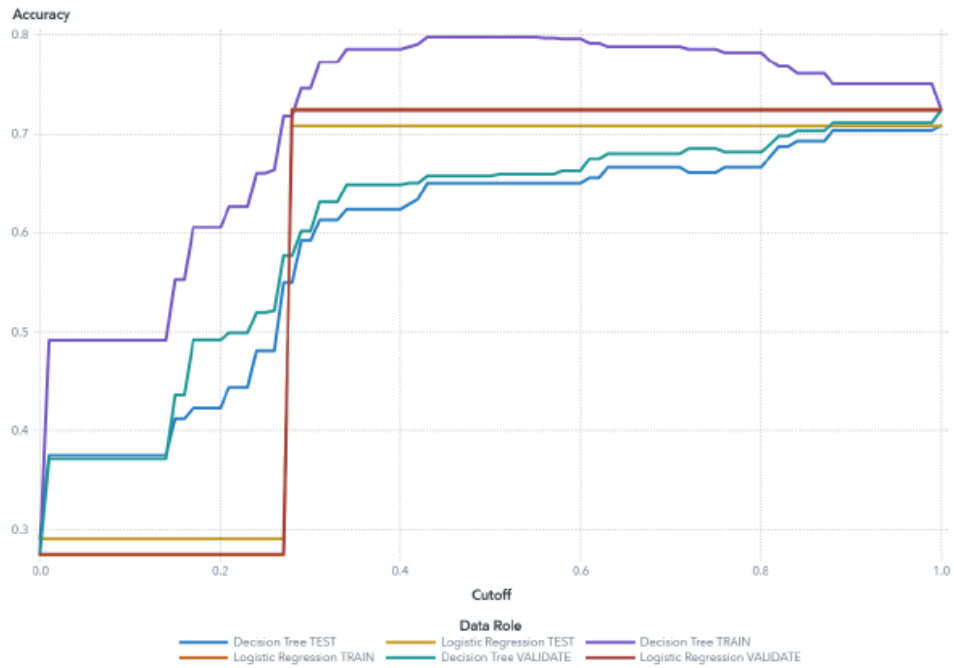
F1 Score	False Discovery Rate	False Positive Rate	Gain
0.1538	0.7391	0.1269	-0.0909
0		0	0.0682

Gini Coefficient	ROC Separation	Lift	Misclassification Rate
-0.0669	-0.0178	1.0909	0.3492
0	0	1.0682	0.2910

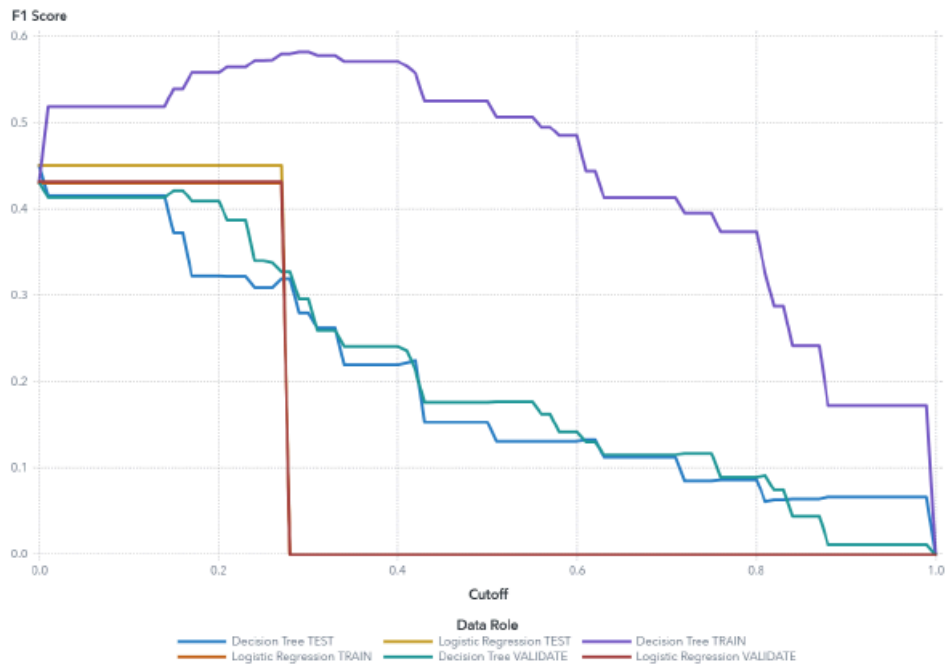
Multi-Class Log Loss	Misclassification at Cutoff	Misclassification Rate (Event)	Number of Observations
2.4891	0.3492	0.3492	189
0.6037	0.2910	0.2910	189

Root Average Squared Error	Captured Response Percentage
0.6228	6.4545
0.4545	6.2910

Accuracy



F1 Score



Comparison Table:

Metric	Logistic Regression (Test Data)	Decision Tree (Test Data)	Better Model
Accuracy	0.7090	0.6508	Logistic Regression
F1 Score	0.0000	0.1538	Decision Tree
Recall	Not Provided	Not Provided	(Unavailable for direct comparison)
Precision	0.1269	0.7391	Decision Tree

Observations:

- Decision Tree has a better Precision (0.7391) compared to Logistic Regression (0.1269).
- Logistic Regression outperforms Decision Tree in terms of Accuracy (0.7090 vs. 0.6508).
- F1 Score, which balances Precision and Recall, indicates that the Decision Tree is performing better.

Conclusion:

Based on the model outputs, Logistic Regression demonstrates a higher accuracy (70.9%) compared to the Decision Tree (65.08%), indicating that Logistic Regression is better at correctly predicting the overall loan status. However, the Decision Tree outperforms Logistic Regression in terms of precision (73.91% vs. 12.69%) and F1 Score (0.1538 vs. 0.000). This suggests that the Decision Tree is better at minimizing false positives, which is critical in applications like loan approval where falsely approving unqualified applicants can be costly.

If minimizing misclassifications (accuracy) is prioritized, Logistic Regression would be the better choice. On the other hand, if the primary goal is to reduce false approvals (precision), the Decision Tree would be more appropriate. Considering the application context of loan approvals, where false positives are particularly undesirable, the Decision Tree appears to be better suited for predicting loan status.

Factors Influencing Loan Approval and Recommended Strategy

From the classification analysis, factors like Credit Score, Annual Income, and Number of Transactions significantly influence loan approval. A higher credit score and annual income increase the likelihood of loan approval, as they indicate financial stability and repayment ability. Similarly, frequent transactions may signal active financial behavior, further supporting loan approval.

Based on these findings, a recommended strategy for approving loans would be to prioritize applicants with strong credit scores and stable, high incomes. Incorporating thresholds for these key factors into the loan approval policy can help mitigate risks. Additionally, the model can be used to flag borderline cases for further review, allowing financial institutions to balance between risk management and business growth.

