Data Engineer – Data Fellowship 12 IYKRA

Assignment 4

Alya Mutiara Firdausyi

Instruction

Scenario

You work as a data analytics engineer for a Ritz Jager Bank, analyzing customer behavior and predicting churn. The dataset provides information on existing credit card customers, including demographic data, financial behavior, and interaction history. Your goal is to build a predictive model to identify customers likely to churn (Attrition_Flag = "Attrited Customer") based on their attributes and behavior. This model will help the bank retain valuable customers and reduce customer churn.

Dataset Link

https://drive.google.com/file/d/1uD1MQWducAAu1-gEJhllRZ0wV-zd6-fC/view?usp=sharing

Answer the following questions

1. Create the data quality report of the dataset you received. What is the issue? How to solve the issue?
2. Solve the data quality issue on the dataset.
3. Create the exploratory data analysis of the dataset.
4. Explain 5 univariate and bivariate analyses that you found interesting.
5. Your company requests that you create a predictive analytics model based on the above problem. Demonstrate the process and show the model performance result

---

Data Quality Report: Ritz Jager Bank Customer Churn Dataset

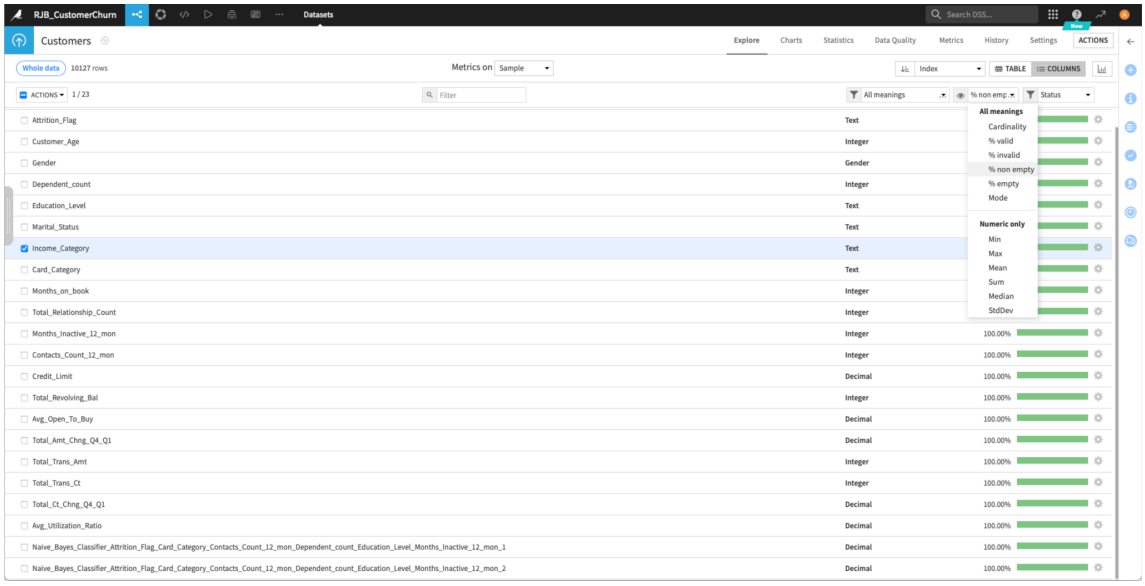
Introduction

This report assesses the quality of the Ritz Jager Bank customer dataset delivered to the data analytics engineer team. The goal is to identify potential data quality issues that would hinder the model's accuracy and effectiveness in predicting customer churn.

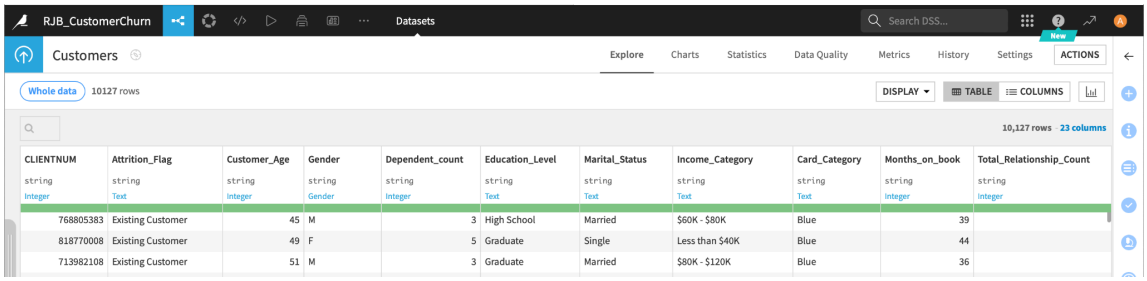
Data Quality Evaluation

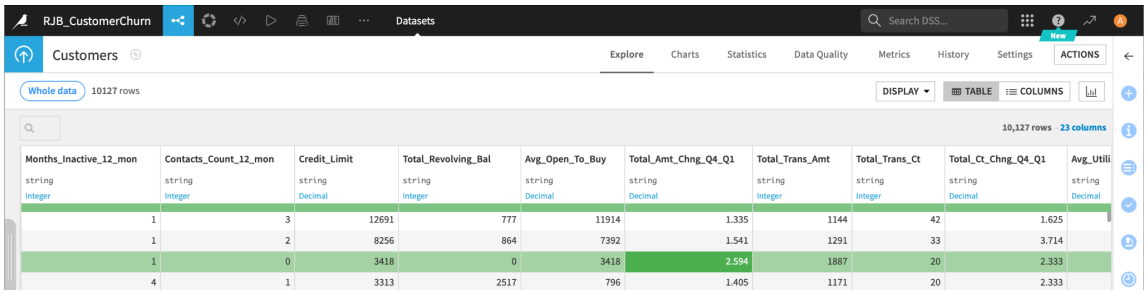
The following dimensions were evaluated to assess the data quality within the dataset.

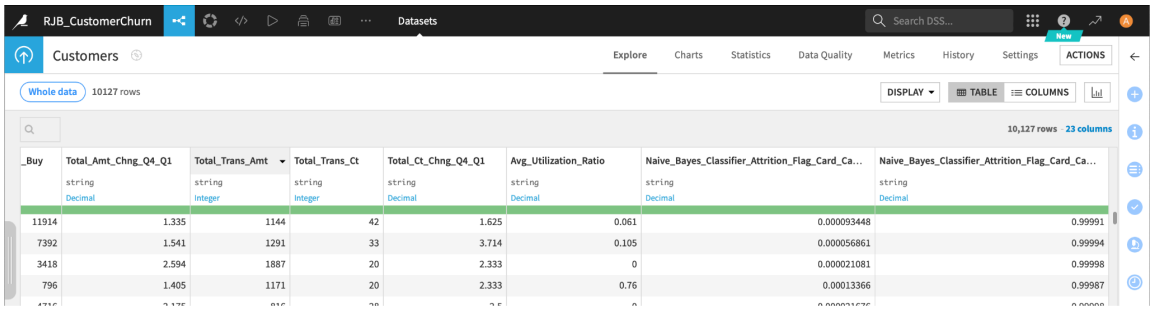
1. **Completeness**: All attributes appear to have values and there is no missing value from the dataset.



2. **Consistency**: It seems we need to justify some column types to better fit the table schema.

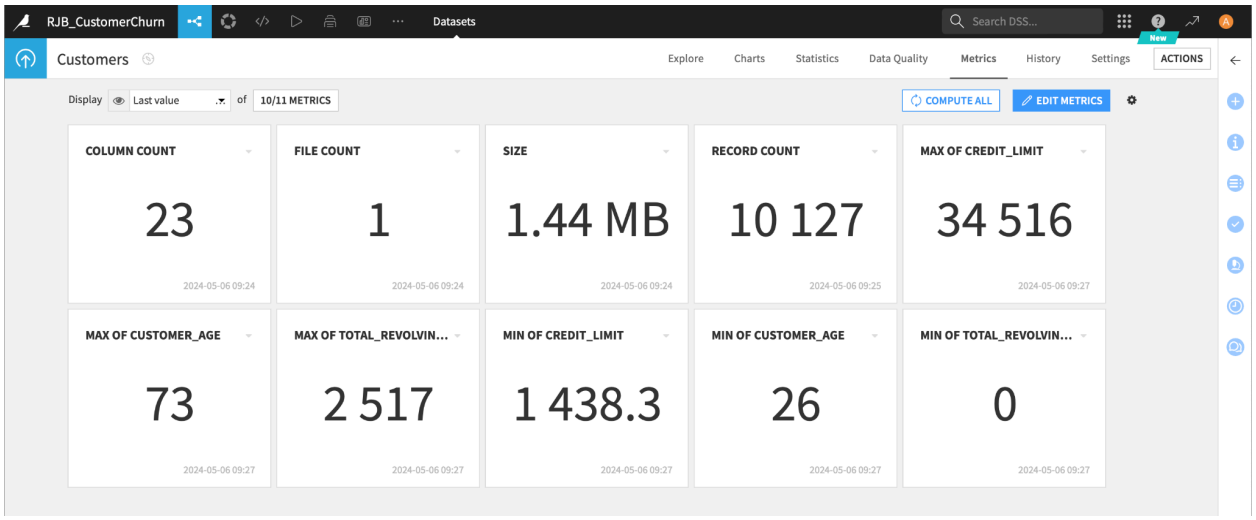






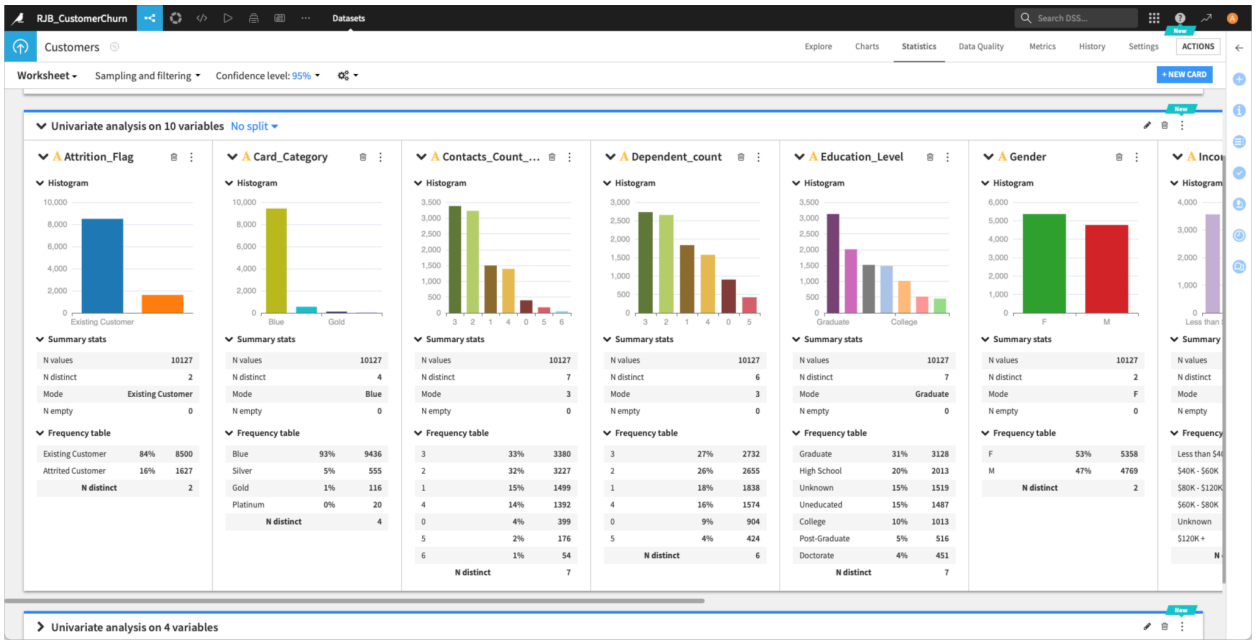
3. Validity:

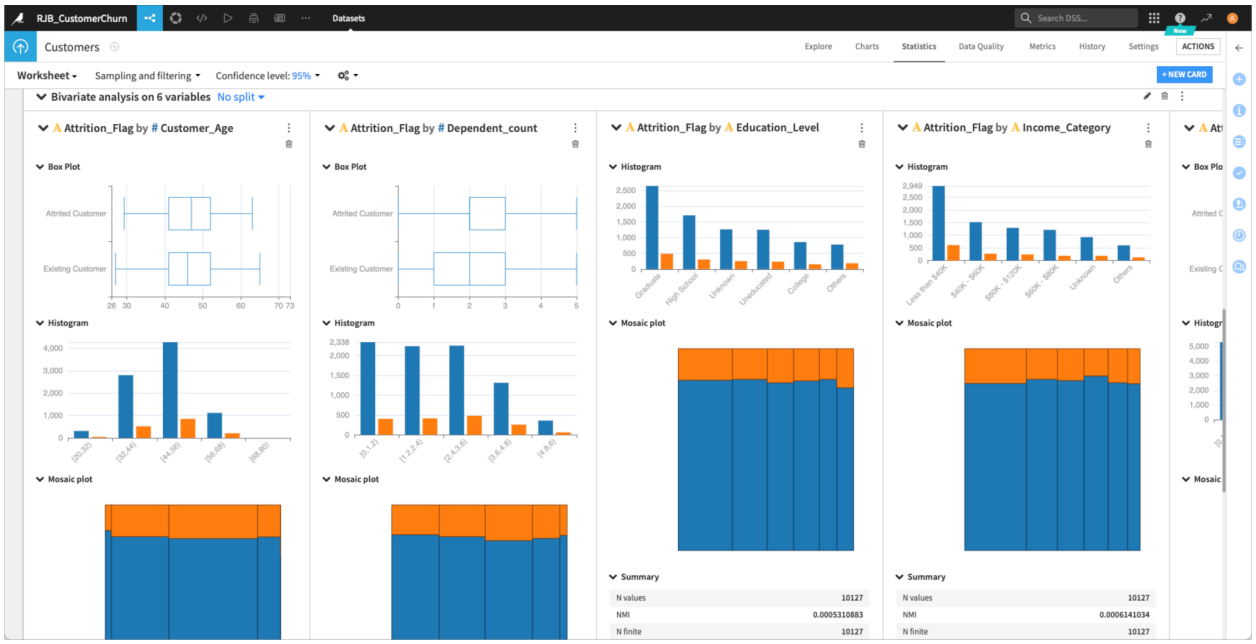
Metadata



Exploratory Data Analysis

Attritied Customer Analysis


RJB_CustomerChurn
Datasets
Search DSS...
Customers
Explore
Charts
Statistics
Data Quality
Metrics
History
Settings
ACTIONS
Worksheet
Sampling and filtering
Confidence level: 95%
+ NEW CARD
Univariate analysis on 10 variables
No split
Attrition_Flag
Card_Category
Contacts_Count_...
Dependent_count
Education_Level
Gender
Histogram
Summary stats
Frequency table
N values 10127
N distinct 2
Mode Existing Customer
N empty 0
Existing Customer 84% 8500
Attrited Customer 16% 1627
N distinct 2
N values 10127
N distinct 4
Mode Blue
N empty 0
Blue 93% 9436
Silver 5% 555
Gold 1% 116
Platinum 0% 20
N distinct 4
N values 10127
N distinct 7
Mode 3
N empty 0
3 33% 3380
2 32% 3227
1 15% 1499
4 14% 1392
0 4% 399
5 2% 176
6 1% 54
N distinct 7
N values 10127
N distinct 6
Mode 3
N empty 0
3 27% 2732
2 26% 2655
1 18% 1838
4 16% 1574
0 9% 904
5 4% 424
N distinct 6
N values 10127
N distinct 7
Mode Graduate
N empty 0
Graduate 31% 3128
High School 20% 2013
Unknown 15% 1519
Uneducated 15% 1487
College 10% 1013
Post-Graduate 5% 516
Doctorate 4% 451
N distinct 7
N values 10127
N distinct 2
Mode F
N empty 0
F 53% 5358
M 47% 4769
N distinct 2
Univariate analysis on 4 variables


RJB_CustomerChurn
Datasets
Search DSS...
Customers
Explore
Charts
Statistics
Data Quality
Metrics
History
Settings
ACTIONS
Worksheet
Sampling and filtering
Confidence level: 95%
+ NEW CARD
Bivariate analysis on 6 variables
No split
Attrition_Flag by # Customer_Age
Attrition_Flag by # Dependent_count
Attrition_Flag by Education_Level
Attrition_Flag by Income_Category
Box Plot
Histogram
Mosaic plot
Attrited Customer
Existing Customer
Summary
N values 10127
NMI 0.0005310883
N finite 10127
N values 10127
NMI 0.0006141034
N finite 10127

Predictive Analytics Model