Credit data ETL ingestion

Outline

1. Context
2. Data overview / Data problems
3. ETL Flows design
4. Data transformation
5. Data validation
6. Data processing engine
7. Testing / Logging / Orchestration
8. Insight

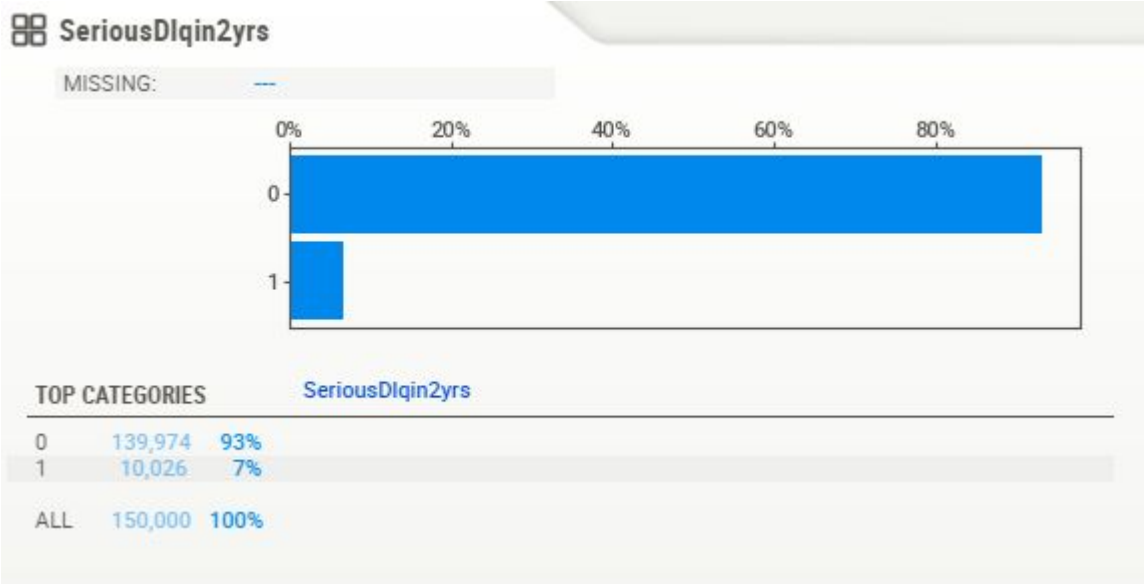
Context

- Write ETL pipeline for credit data source.
- Data artifacts have to be in centralized place, easy to access.
- Main stakeholders: Data Scientist and Data Analyst.

SeriousDlqin2yrs	Binary	Person experienced 90 days past due delinquency or worse.
RevolvingUtilizationOfUnsecuredLines	Float	Total balance on credit cards and personal lines of credit except real estate and installment debt (e.g. car loans) divided by the sum of credit limits.
age	Integer	Age of borrower in years.
NumberOfTime30-59DaysPastDueNotWorse	Integer	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.
DebtRatio	Float	Monthly debt payments, alimony, and living costs divided by monthly gross income.
MonthlyIncome	Float	Monthly income.
NumberOfOpenCreditLinesAndLoans	Integer	Number of open loans (e.g. car loan, mortgage) and lines of credit (e.g. credit cards).
NumberOfTimes90DaysLate	Integer	Number of times borrower has been 90 days or more past due.
NumberRealEstateLoansOrLines	Integer	Number of mortgage and real estate loans including home equity lines of credit.
NumberOfTime60-89DaysPastDueNotWorse	Integer	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.
NumberOfDependents	Integer	Number of dependents in family excluding applicant (spouse, children, etc...).

	0	1	2	3	4	5	
SeriousDlqin2yrs	1.000000	0.000000	0.000000	0.00000	0.000000	0.000000	
RevolvingUtilizationOfUnsecuredLines	0.766127	0.957151	0.658180	0.23381	0.907239	0.213179	
age	45.000000	40.000000	38.000000	30.00000	49.000000	74.000000	5
NumberOfTime30-59DaysPastDueNotWorse	2.000000	0.000000	1.000000	0.00000	1.000000	0.000000	
DebtRatio	0.802982	0.121876	0.085113	0.03605	0.024926	0.375607	57
MonthlyIncome	9120.000000	2600.000000	3042.000000	3300.00000	63588.000000	3500.000000	
NumberOfOpenCreditLinesAndLoans	13.000000	4.000000	2.000000	5.00000	7.000000	3.000000	
NumberOfTimes90DaysLate	0.000000	0.000000	1.000000	0.00000	0.000000	0.000000	
NumberRealEstateLoansOrLines	6.000000	0.000000	0.000000	0.00000	1.000000	1.000000	
NumberOfTime60-89DaysPastDueNotWorse	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	
NumberOfDependents	2.000000	1.000000	0.000000	0.00000	0.000000	1.000000	

Some first rows of out dataset. All columns are numeric, quite clean.



Data size: 150.000
Very high imbalance in label distribution

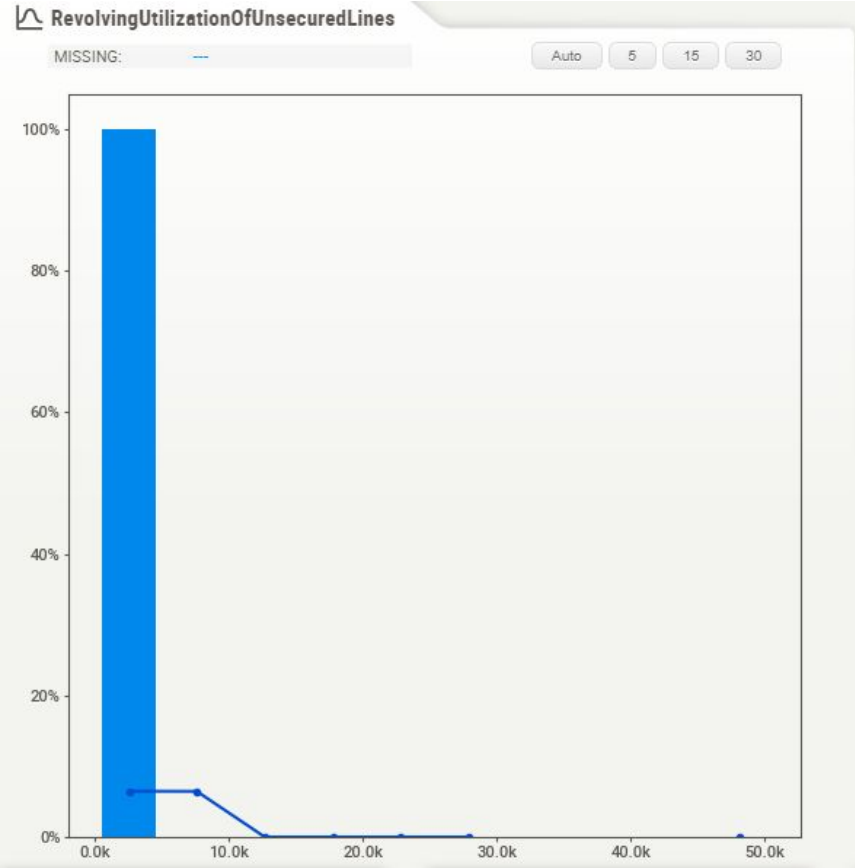
Data problems

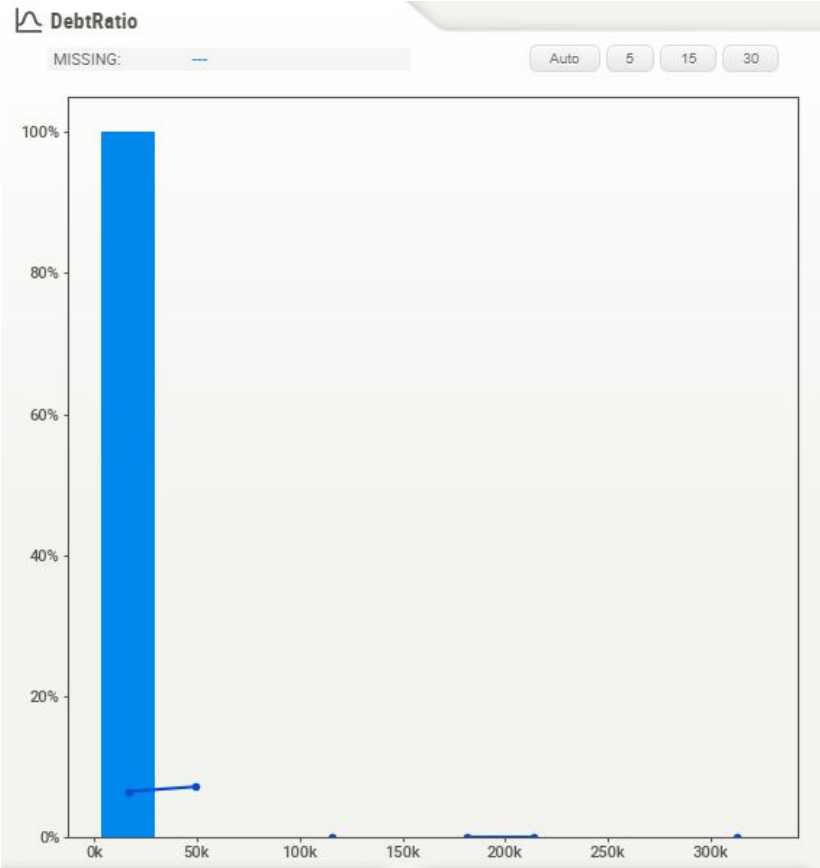
- Column names are camelCase: not so pythonic, and contain special character “-”
- Many data duplicated: since the data is personal credit information, so it’s strange if we have duplicated rows here.
- Data inconsistent:
  - How can MonthlyIncome = 0.0 ?
  - Special values 96, 98 in days past due columns, should be treated as NULL ?
- Some columns have outliers:
  - Unrealistic high value of RevolvingUtilizationOfUnsecuredLines, MonthlyIncome, DebtRatio
  - Keeping these outliers does help improve model performance and may affect data analysis result.
  - Since number of outliers are small, consider to remove, or cap at realistic values.

	137102	83552	139345	27724	54653	69656	49931	82655	89633	101499
SeriousDlqin2yrs	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RevolvingUtilizationOfUnsecuredLines	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
age	21.0	21.0	21.0	21.0	22.0	22.0	22.0	22.0	22.0	22.0
NumberOfTime30-59DaysPastDueNotWorse	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DebtRatio	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MonthlyIncome	0.0	820.0	820.0	NaN	0.0	1.0	820.0	820.0	820.0	820.0
NumberOfOpenCreditLinesAndLoans	1.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0
NumberOfTimes90DaysLate	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NumberRealEstateLoansOrLines	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NumberOfTime60-89DaysPastDueNotWorse	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NumberOfDependents	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

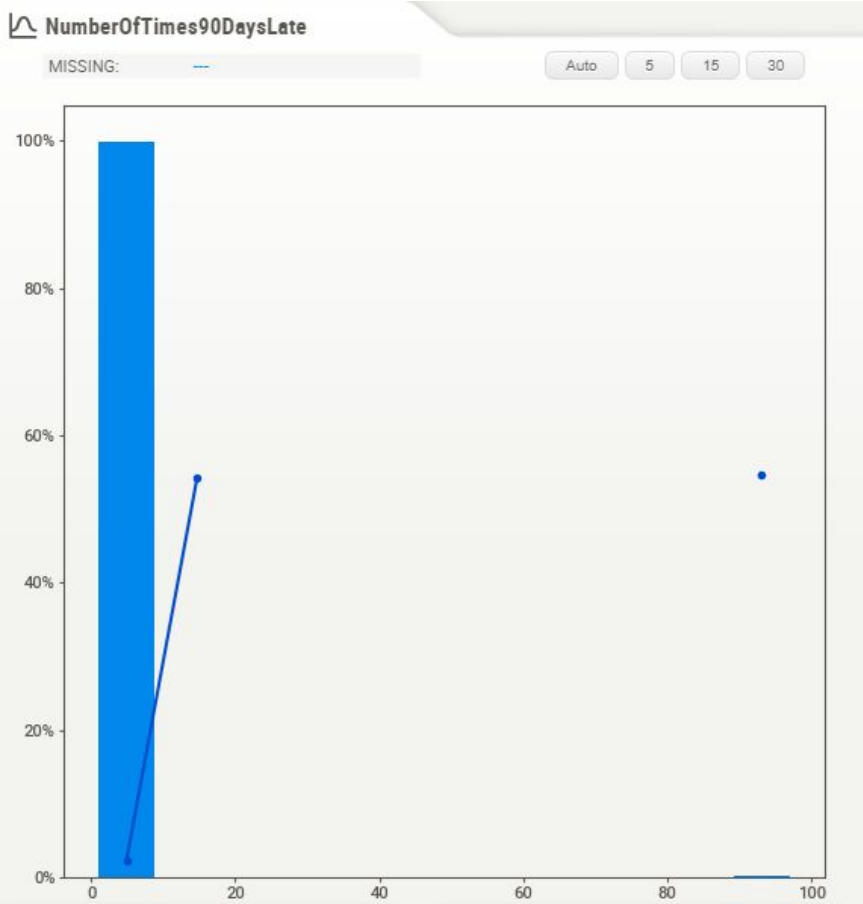
11 rows × 609 columns

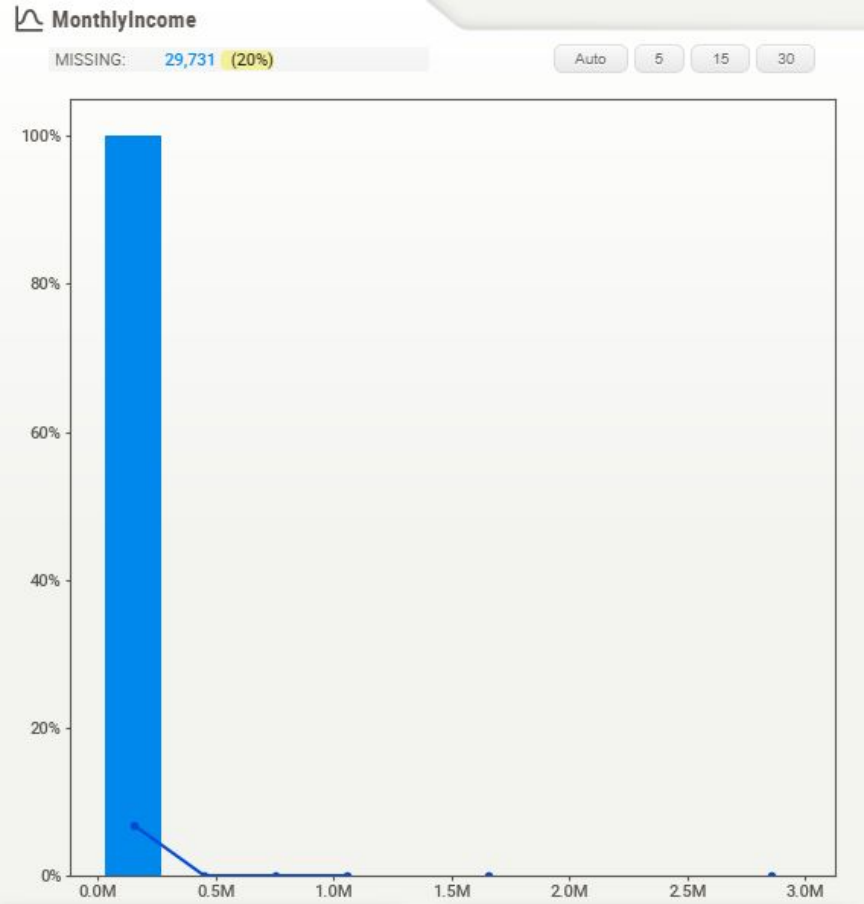
Duplicated rows in our dataset



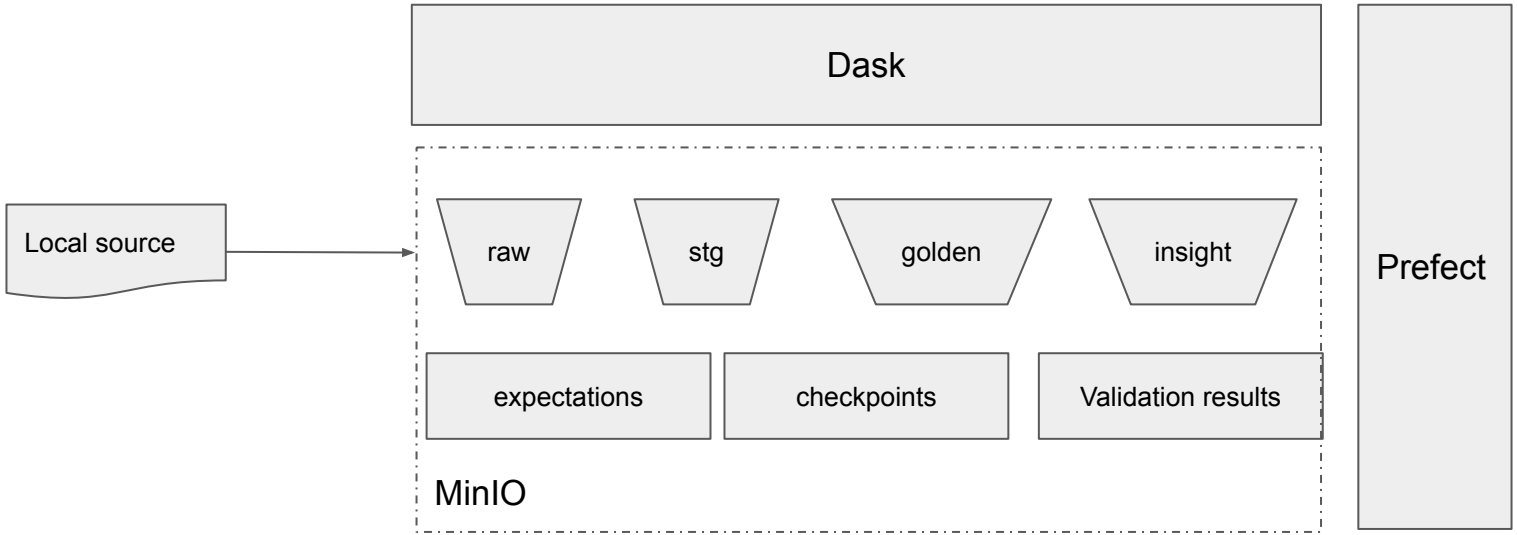


Outlier in our dataset





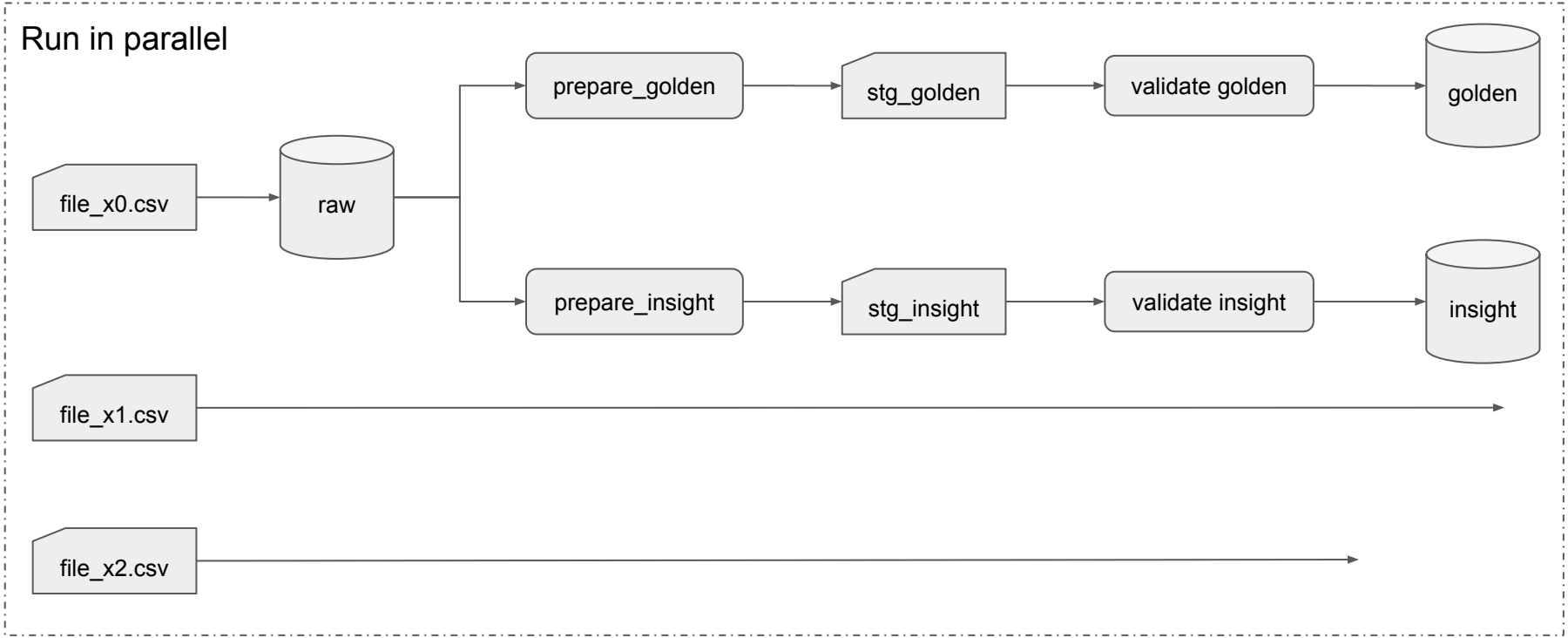
Outlier in our dataset



Architecture:

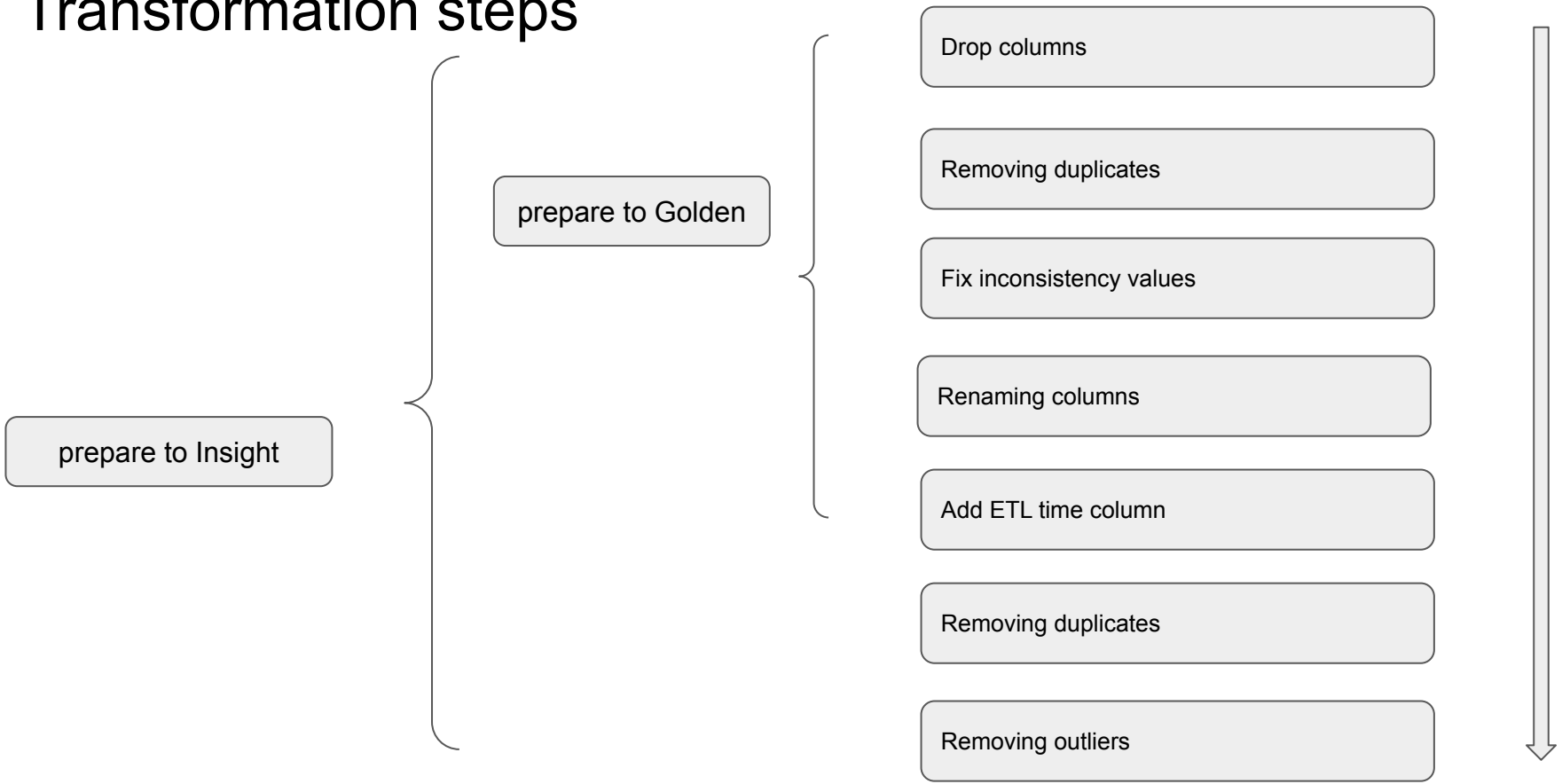
- *MinIO as centralized storage*
- *Dask as data processing engine*
- *Great Expectation as data validation tool*
- *Prefect: orchestration / scheduler*

ETL flow



Designed ETL flow

Transformation steps



Data validation

- Raw-zone rules:
  - Data size > 0
  - Data columns matching
- Golden-zone rules:
  - Data columns matching
  - etl_time must be added
  - Target label must not be Null
- Insight-zone rules:
  - Data columns matching
  - Target label must not be Null
  - Column values expected to be in range [min - max]
- Validation results are stored in S3 so we can re-check if something wrong.

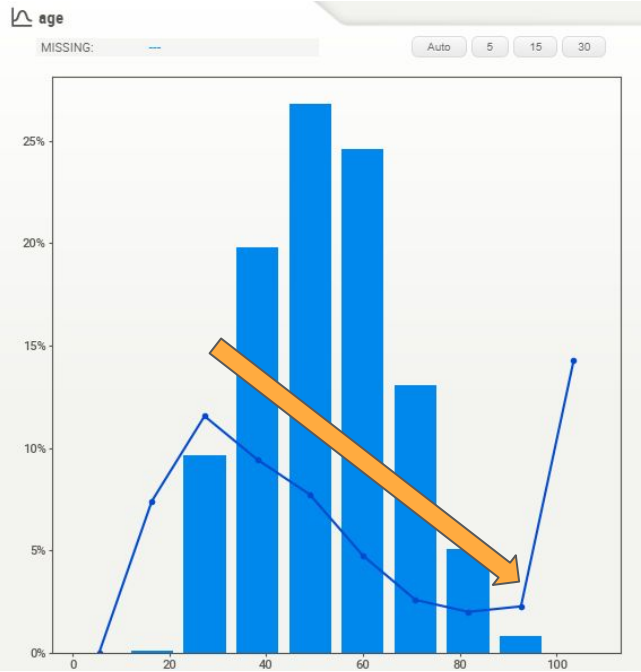
Data Processing

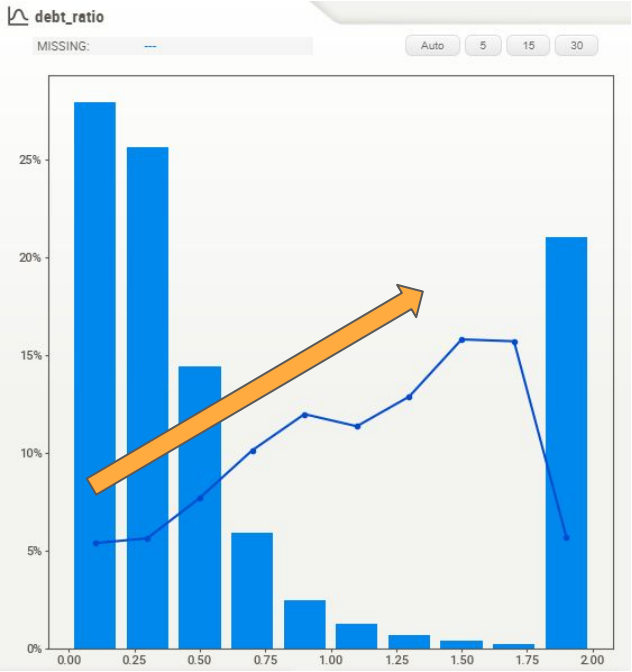
- Use Dask as scalable data processing engine
  - Parallel calculation across processes or threads
  - Handle large files by partitioning
- In production, should be connected to a Dask distributed cluster

Unit test / Logging / Orchestration

- Use built in Prefect logging
- Use pytest for Unit tests
- Use DaskTaskRunner to run parallel tasks
- Use interval config to setup flow run on daily basis.

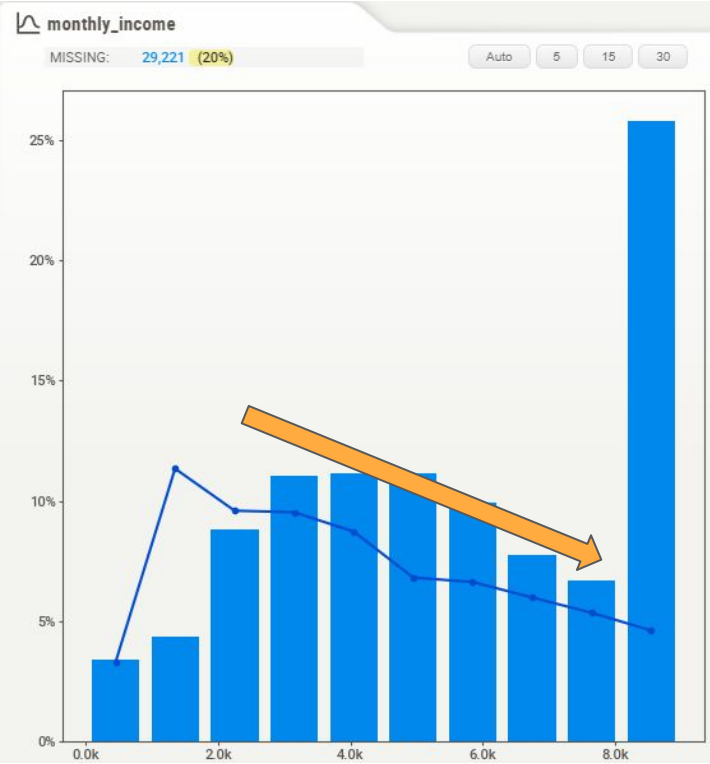
Data visualization after cleaning

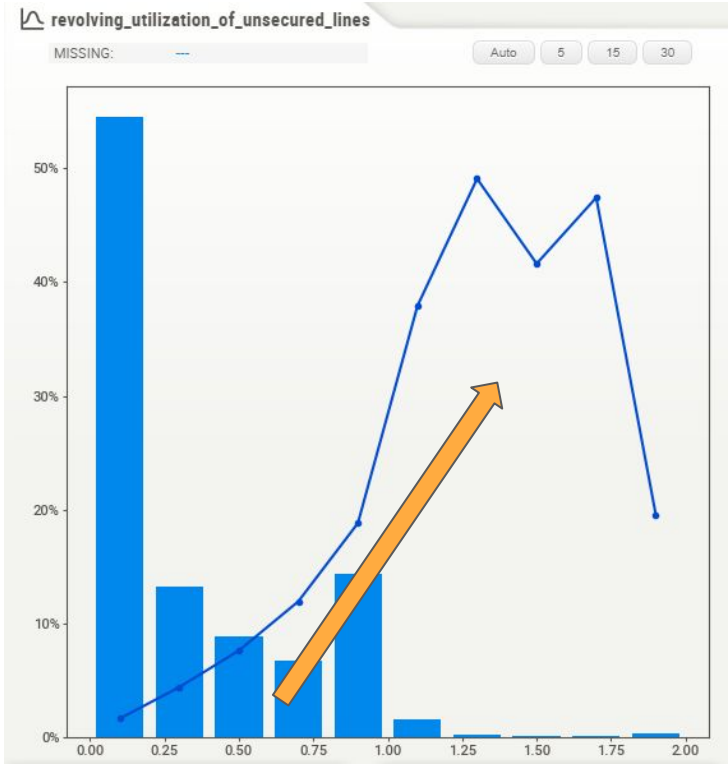


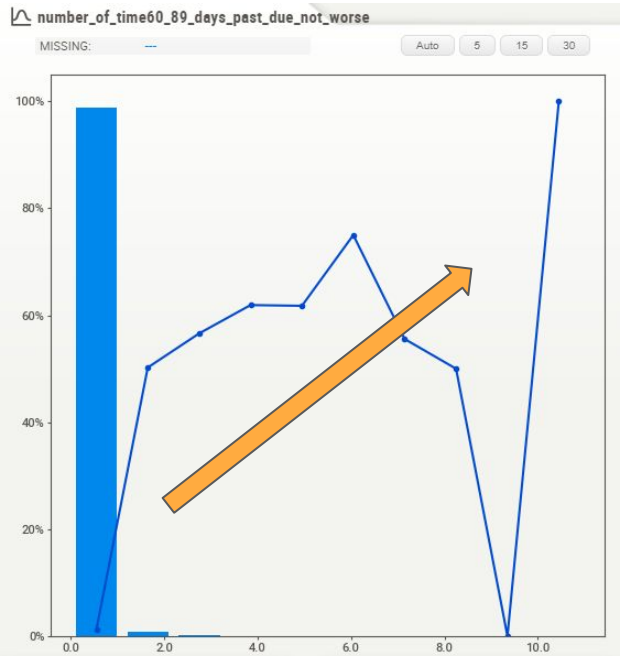


Line: probability of being delinquent in 2 years (“Y”)

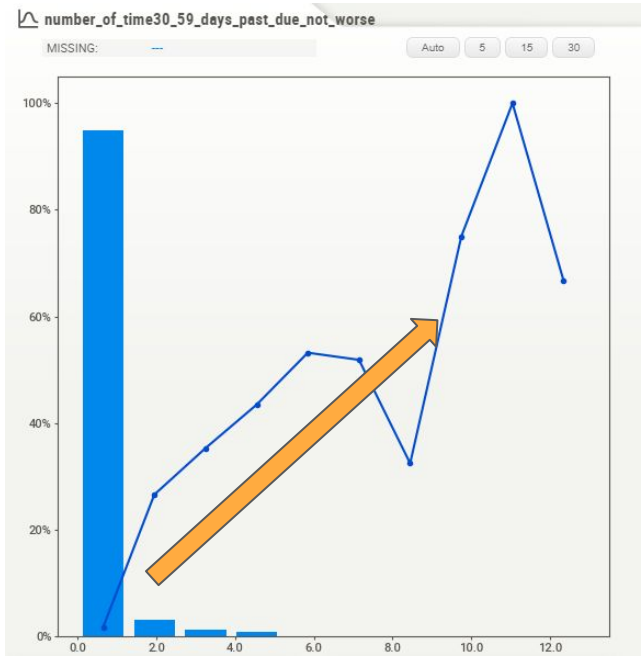
Column: distribution of according variable


monthly_income
MISSING: 29,221 (20%)
Auto
5
15
30
25%
20%
15%
10%
5%
0%
0.0k
2.0k
4.0k
6.0k
8.0k

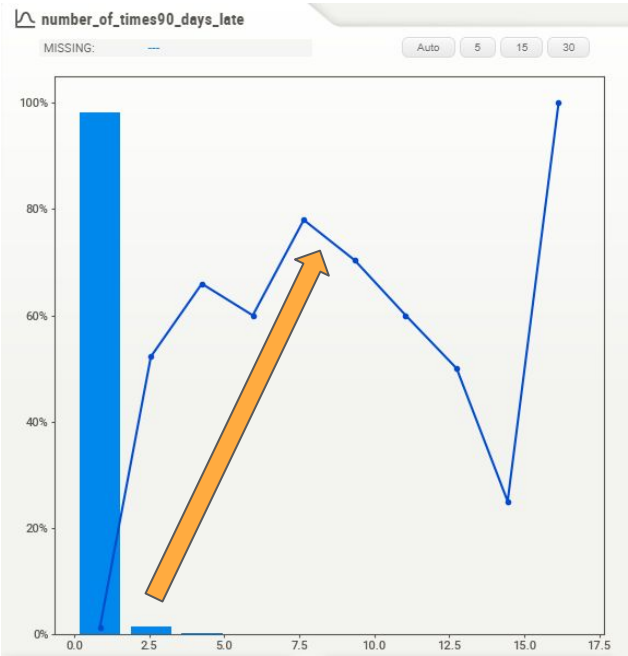

revolving_utilization_of_unsecured_lines
MISSING: ---
Auto
5
15
30
50%
40%
30%
20%
10%
0%
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


number_of_time60_89_days_past_due_not_worse
MISSING:

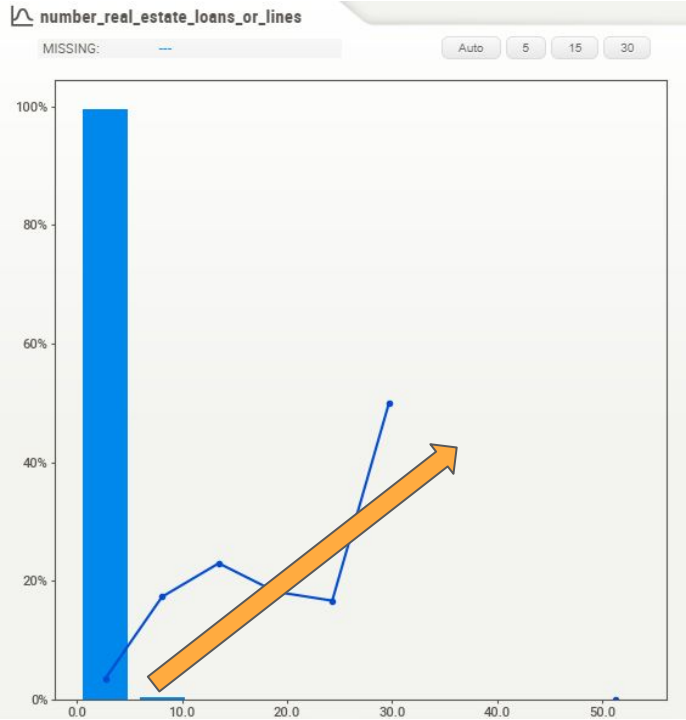
Auto
5
15
30
100%
80%
60%
40%
20%
0%
0.0
2.0
4.0
6.0
8.0
10.0


number_of_time30_59_days_past_due_not_worse
MISSING:

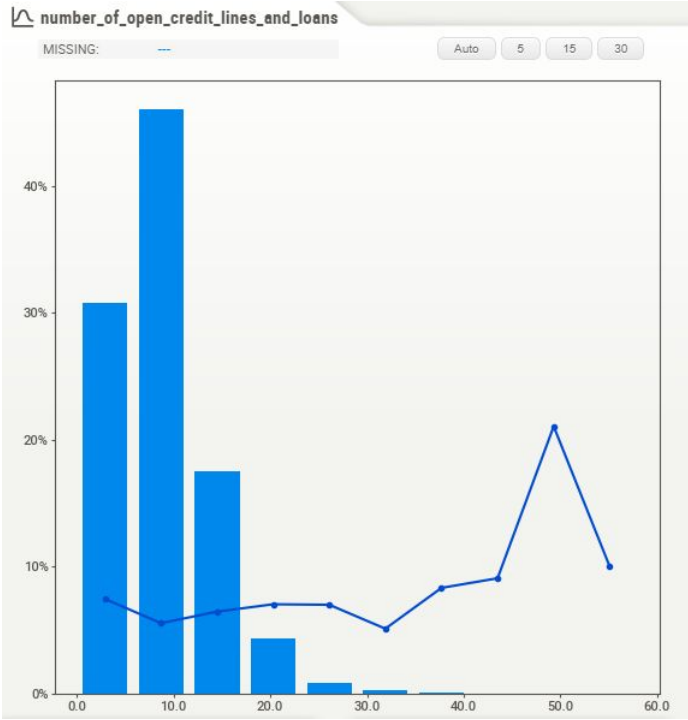
Auto
5
15
30
100%
80%
60%
40%
20%
0%
0.0
2.0
4.0
6.0
8.0
10.0
12.0


number_of_times90_days_late
MISSING:

Auto
5
15
30
100%
80%
60%
40%
20%
0%
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


number_real_estate_loans_or_lines
MISSING:

Auto
5
15
30
100%
80%
60%
40%
20%
0%
0.0
10.0
20.0
30.0
40.0
50.0


number_of_open_credit_lines_and_loans
MISSING:

Auto
5
15
30
40%
30%
20%
10%
0%
0.0
10.0
20.0
30.0
40.0
50.0
60.0

Insight

- There are some trend in data, some columns are highly correlated with target variable like: age, DebtRatio, MonthlyIncome, RevovingUtilization
- Number of loans and Days past due columns are also good features to regress probability of target variable.
- Number of dependencies seems does not affect much to target variable.