ingestion-prototype

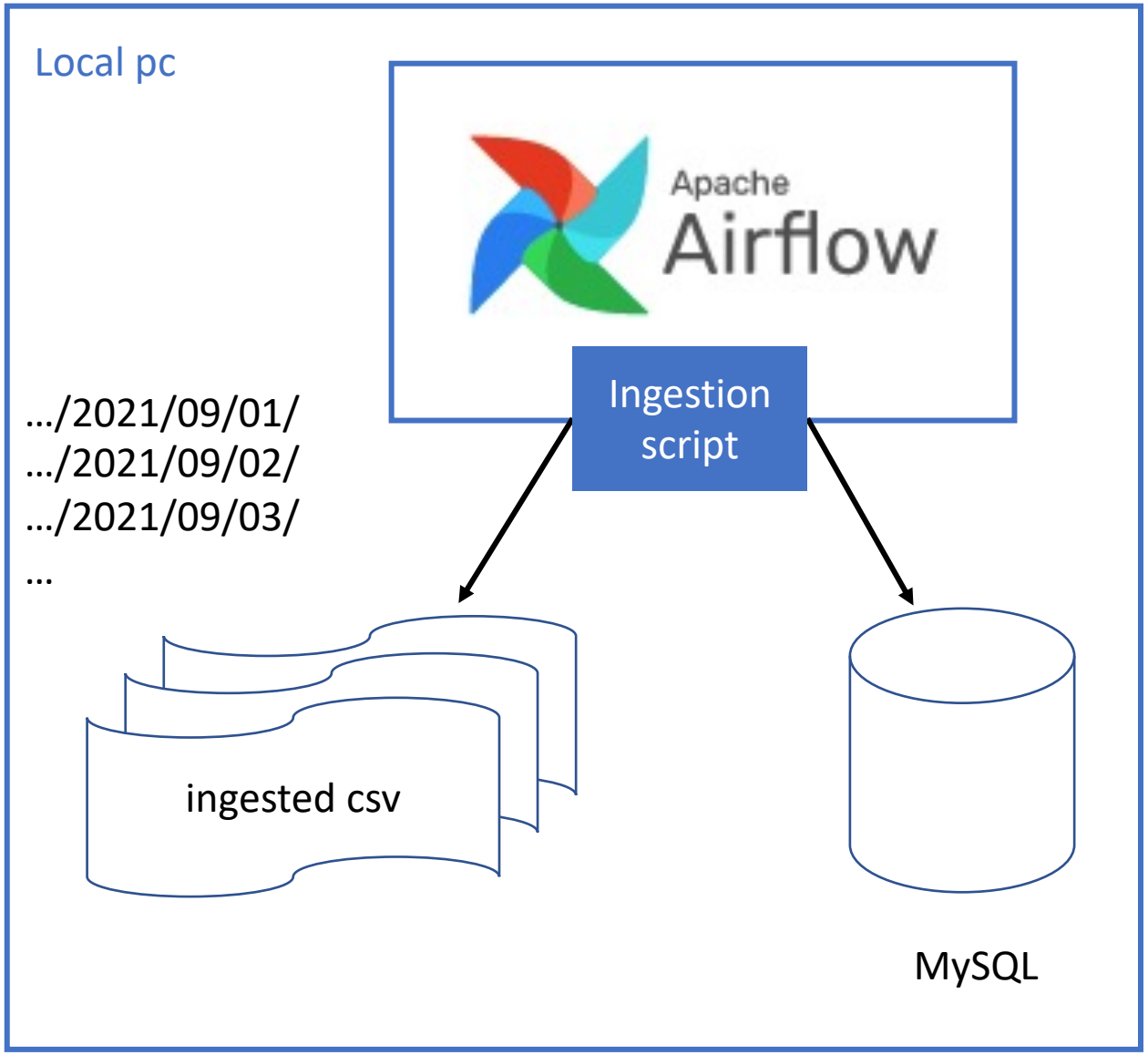
muqtafiakhmad

Overview

- Data Ingestion.
- Exploratory data analysis.

Data Ingestion

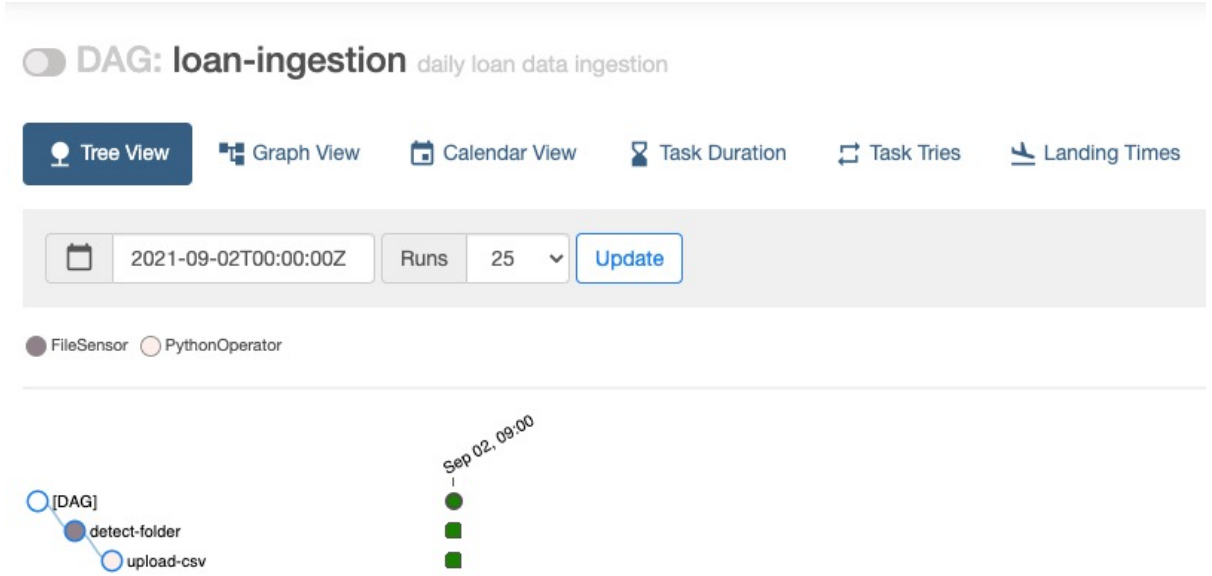
Overview



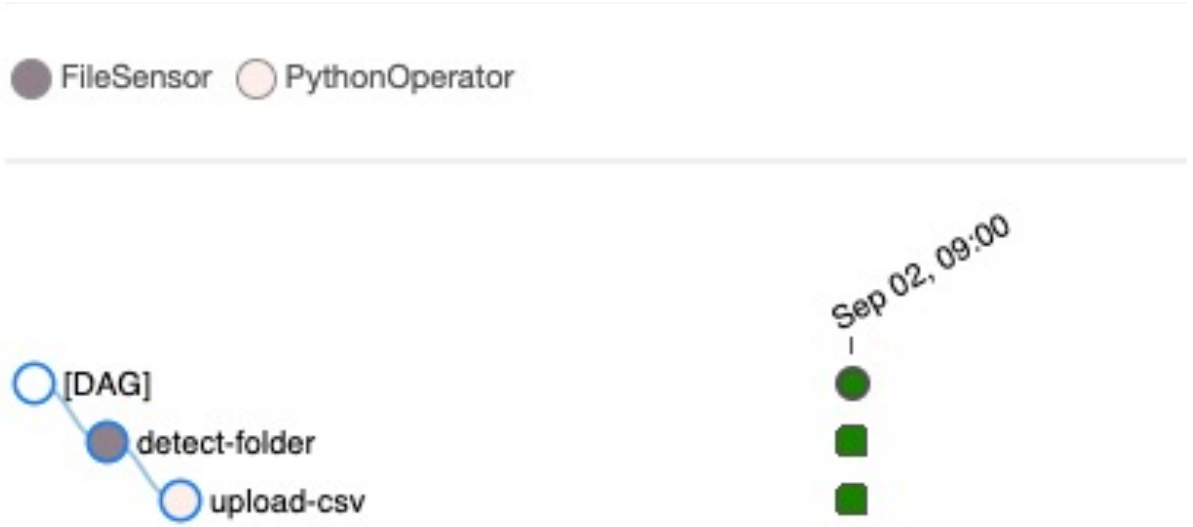
About Airflow: https://airflow.apache.org/

Key features:

- Definition of data processing as DAG.
- Can rerun past jobs.



Ingestion script



- DAG with two operators.
  - detect-folder.
    - FileSensor.
    - Detect whether folder containing ingest target is ready.
    - Expected format …/YYYY/MM/DD/.
  - upload-csv.
    - PythonOperator.
    - Parse csv and then upload to MySQL.

Ingestion script

```
with DAG(
    'loan-ingestion',
    default_args=default_args,
    description='daily loan data ingestion',
    schedule_interval=timedelta(days=1),
    # set September 2nd as start date
    start_date=datetime(2021,9,2),
    tags=['daily', 'credit', 'loan'],
) as dag:
    # construct folder path
    date_str = "{{ (execution_date-macros.timedelta(days=1)).strftime('%Y/%m/%d') }}"
    folder_path = '/Users/muqtafiakhmad/Desktop/credit/credit-data-ingestion/sample/'+date_str

    file_sensor = FileSensor(
        task_id='detect-folder',
        poke_interval=30,
        filepath=folder_path)
```

We put the sample data at 2021/09/01, assume that ingestion is triggered everyday to ingest yesterday's data

Folder path derived from previous date

Ingestion script (cont'd)

```
def upload_csv(folder_path, insert_date): ...

upload_csv_operator = PythonOperator(
    task_id='upload-csv',
    python_callable=upload_csv,
    op_kwargs={
        'folder_path': '/Users/muqtafiakhmad/Desktop/credit/credit-data-ingestion/sample/'+date_str,
        'insert_date': "{{ (execution_date-macros.timedelta(days=1)).strftime('%Y/%m/%d').replace('/', '-') }}"
    },
)

# define DAG
file_sensor >> upload_csv_operator
```

Parameters passed to the invoked python function

Corners cut and possible improvements

- user=password=root
  - Need to use proper data loader account for ingestion.
- Ingested file, Airflow, and target DB are deployed in a single host.
  - In production environment, its possible that data are stored in a Hadoop cluster.
  - Airflow already provide operators to submit Spark jobs.

```
# construct connection to mysql
connection_string = 'mysql+mysqlconnector://{user}:{password}@{host}/{dbname}'.format(
    user='root',
    password='root',
    host='localhost',
    dbname='credit'
)

from sqlalchemy import create_engine

engine = create_engine(connection_string)
```

Exploratory data analysis

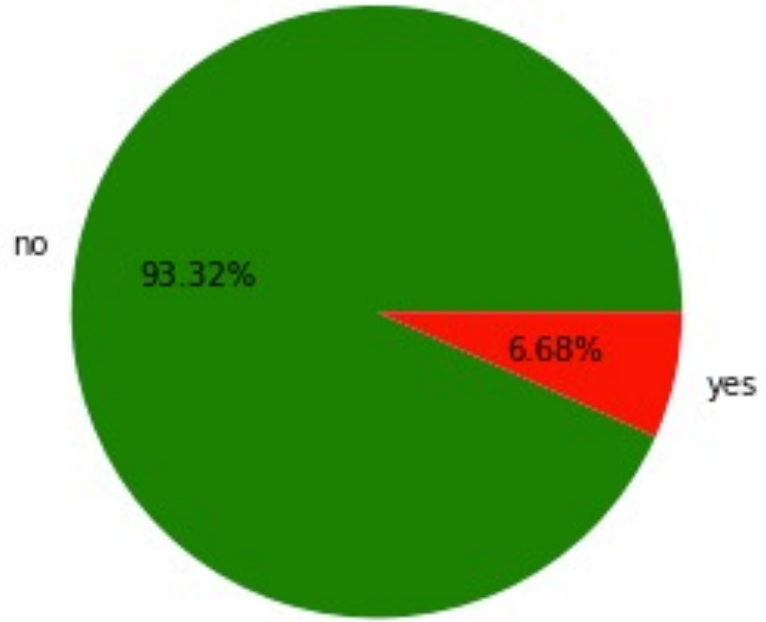
Data overview: dataset for classification problem

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

Target variable, assume
0 = No
1 = Yes

SeriousDlqin2yrs distribution is skewed to no issue case

Label	Count	Size
no	139974	0.9332
yes	10026	0.0668



Short summary

	SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	Age	NumberOfTime30To59DaysPastDueNotWorse	DebtRatio	MonthlyIncome
count	150000.000000	150000.000000	150000.000000	150000.000000	150000.000000	1.202690e+05
mean	0.066840	6.048438	52.295207	0.421033	353.005076	6.670221e+03
std	0.249746	249.755371	14.771866	4.192781	2037.818523	1.438467e+04
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00
25%	0.000000	0.029867	41.000000	0.000000	0.175074	3.400000e+03
50%	0.000000	0.154181	52.000000	0.000000	0.366507	5.400000e+03
75%	0.000000	0.559046	63.000000	0.000000	0.868254	8.249000e+03
max	1.000000	50708.000000	109.000000	98.000000	329664.000000	3.008750e+06

Contains empty values

Std > mean

0 values

Short summary

	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60To89DaysPastDueNotWorse
count	150000.000000	150000.000000	150000.000000	150000.000000
mean	8.452760	0.265973	1.018240	0.240387
std	5.145951	4.169304	1.129771	4.155179
min	0.000000	0.000000	0.000000	0.000000
25%	5.000000	0.000000	0.000000	0.000000
50%	8.000000	0.000000	1.000000	0.000000
75%	11.000000	0.000000	2.000000	0.000000
max	58.000000	98.000000	54.000000	98.000000

	NumberOfDependents
count	146076.000000
mean	0.757222
std	1.115086
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	20.000000

Contains empty values

Std > mean

0 values

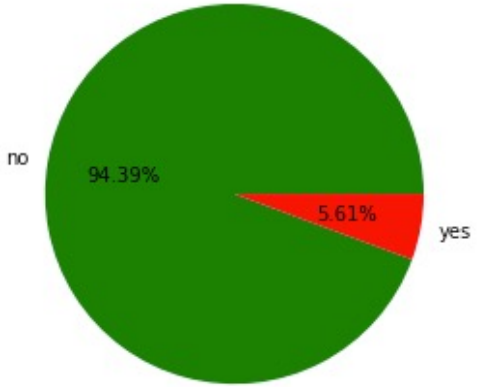
Closer look at empty values

```
Total count  29731
Total pct dataset  19.8207
```

MonthlyIncomeNan	NumberOfDependentsNan	count	pct_dataset
True	False	25807	17.2047
	True	3924	2.6160

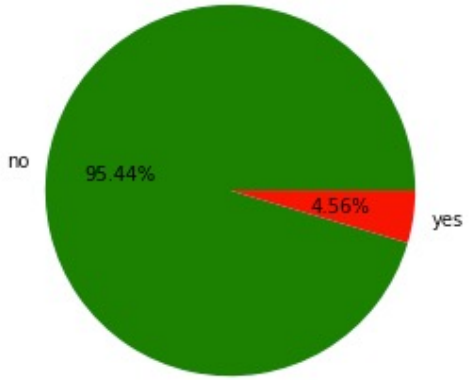
MonthlyIncome empty

```
Label  Count    Size
   no  28062  0.9439
  yes   1669  0.0561
```



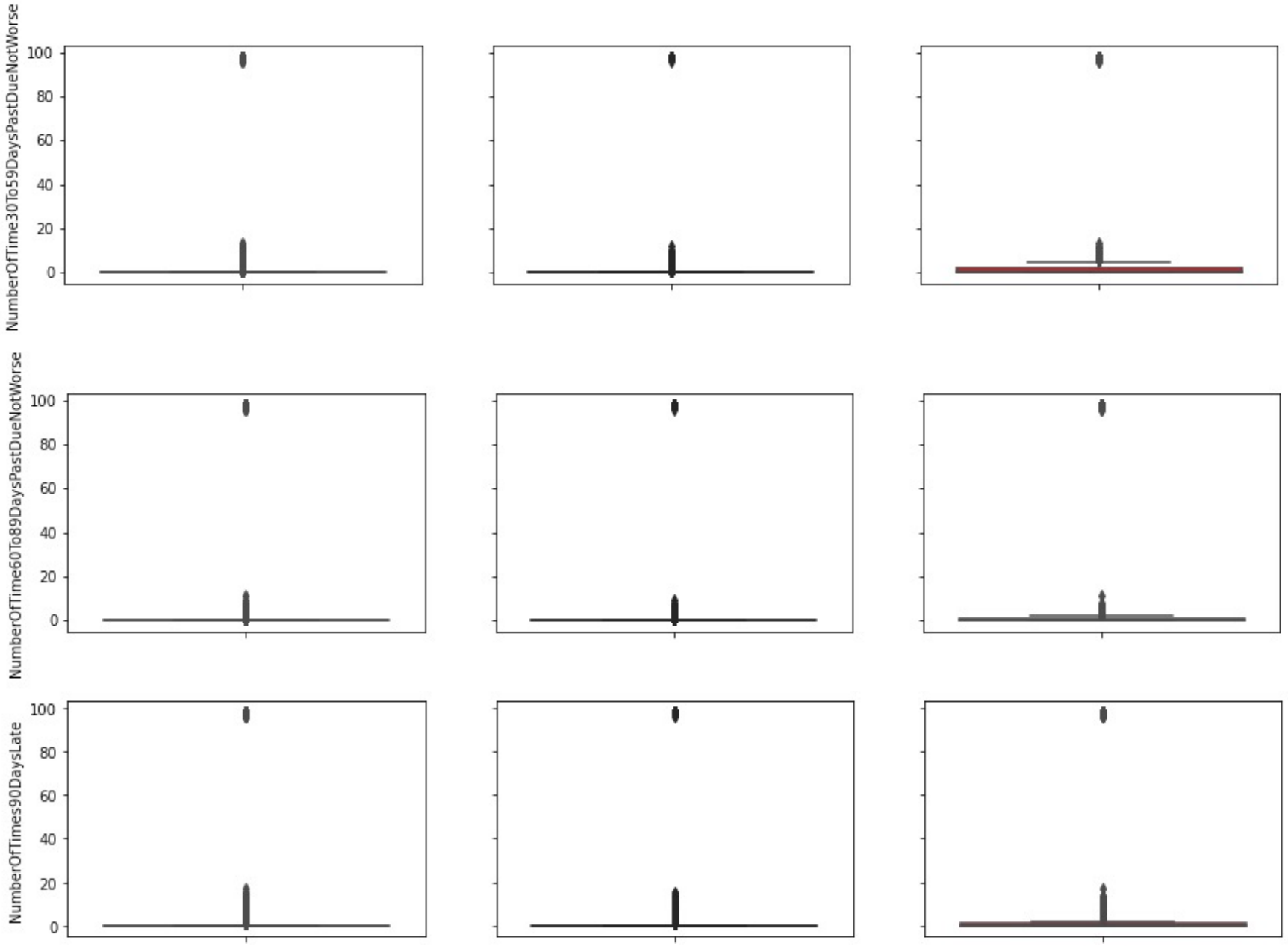
NumberOfDependents empty

```
Label  Count    Size
   no   3745  0.9544
  yes    179  0.0456
```



Spotting Outliers

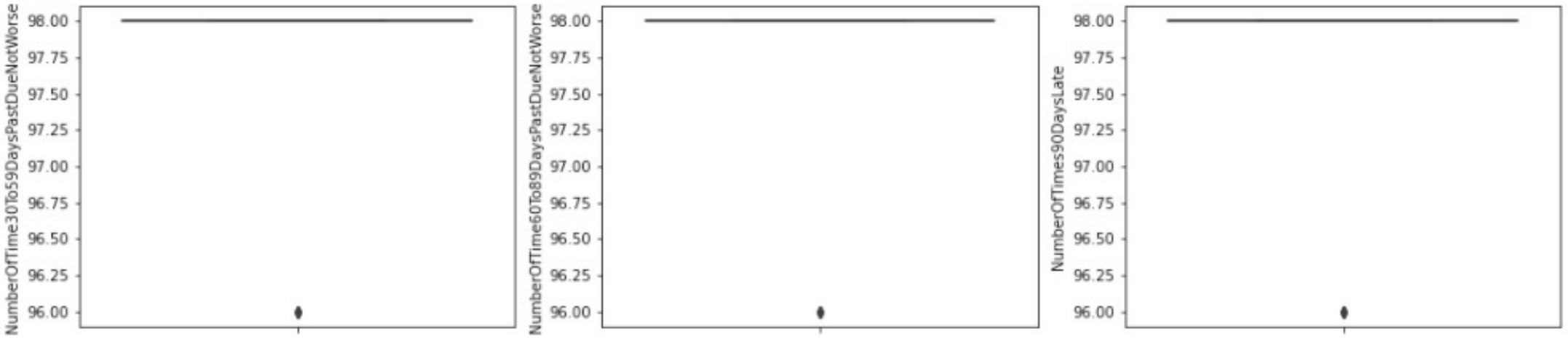
Majority is < 20



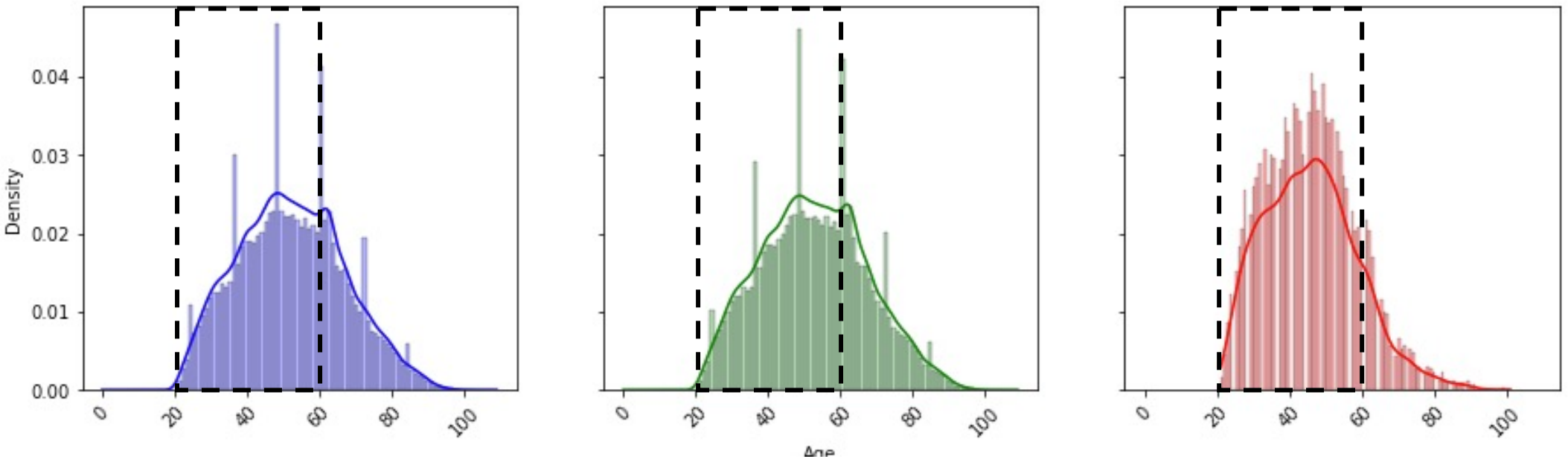
Correlation among outliers

(NumberOfTime30To59DaysPastDueNotWorse > 40) or (NumberOfTimes90DaysLate > 40) or (NumberOfTime60To89DaysPastDueNotWorse > 40)

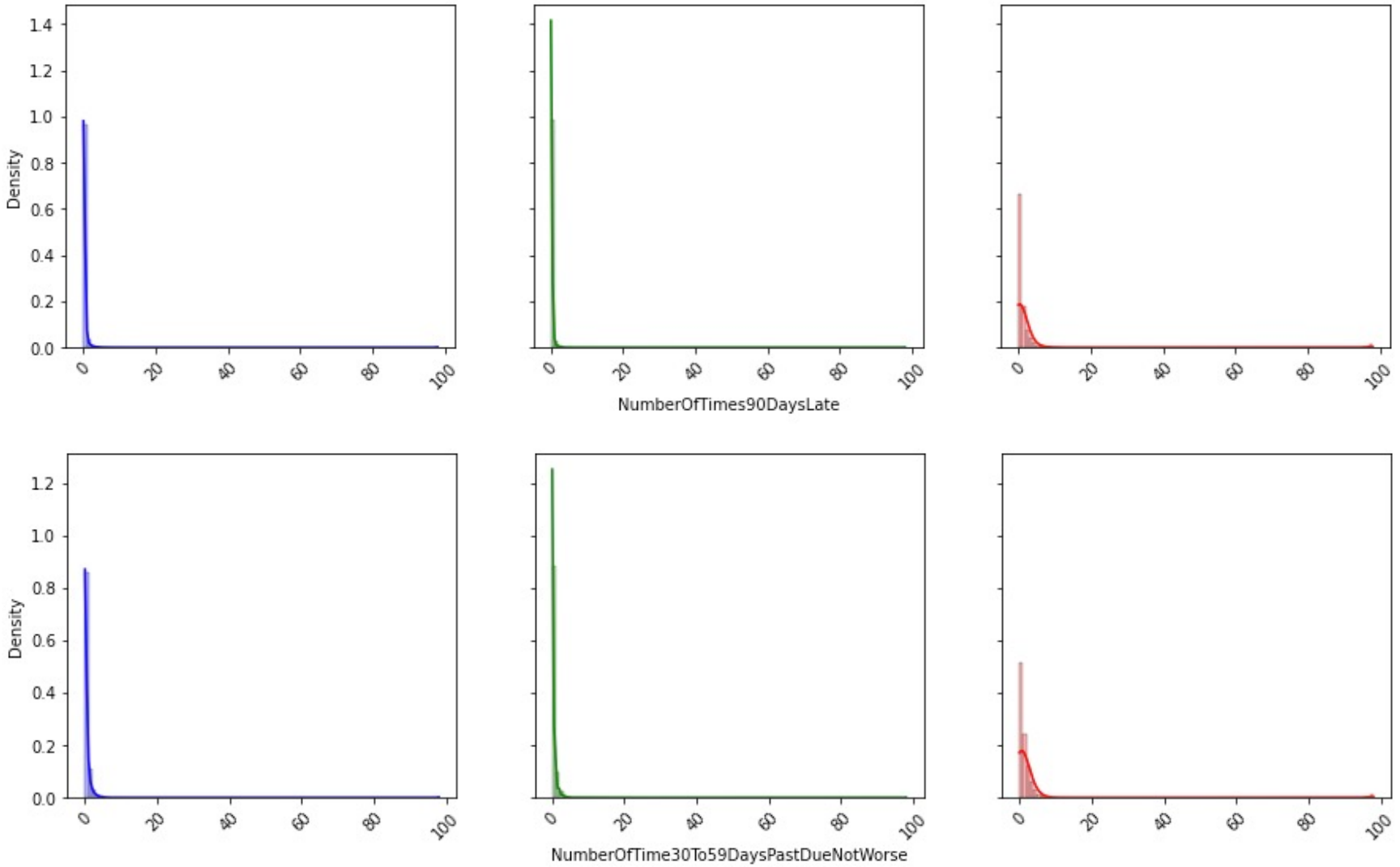
```
Total count  269
Total pct dataset  0.17933
```



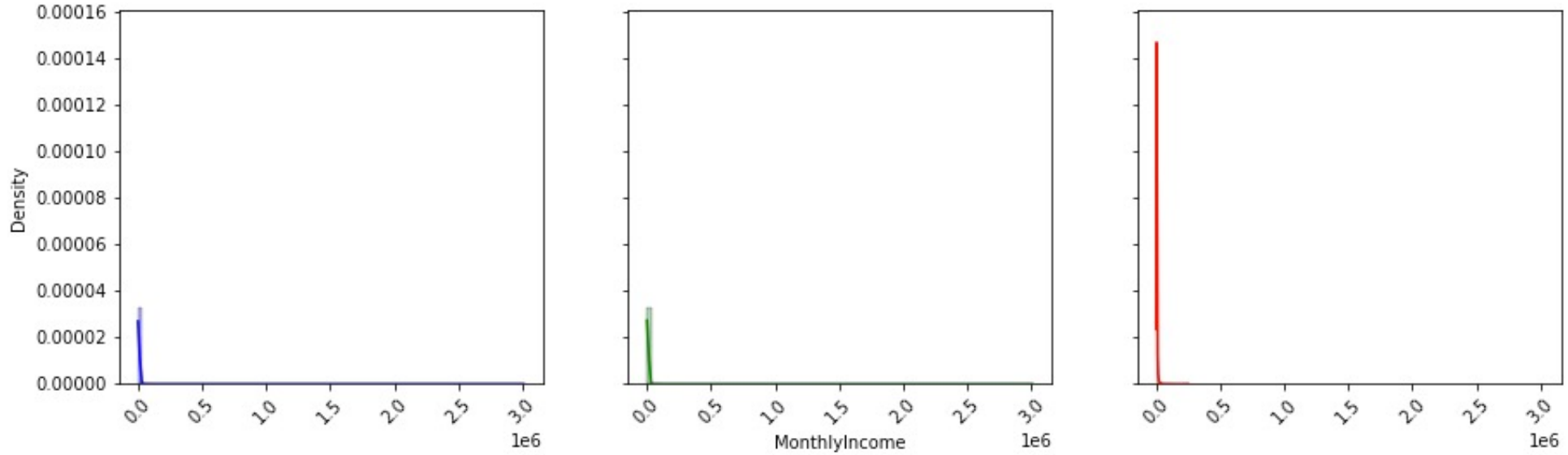
Problematic cases more centered on 20<age<60



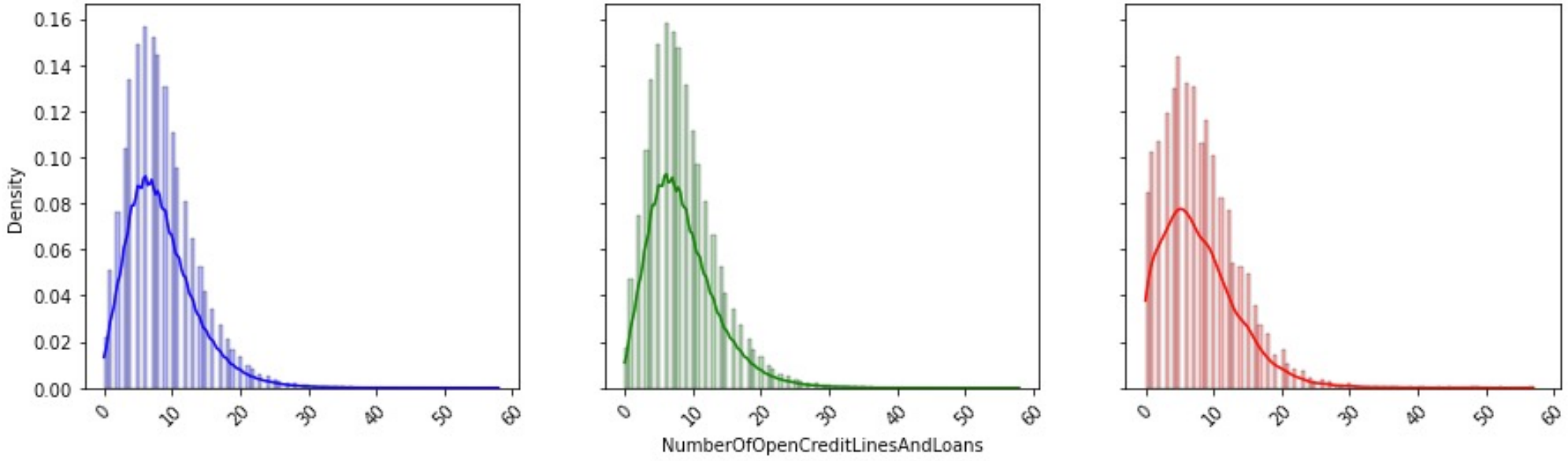
Problematic cases has more history of past dues



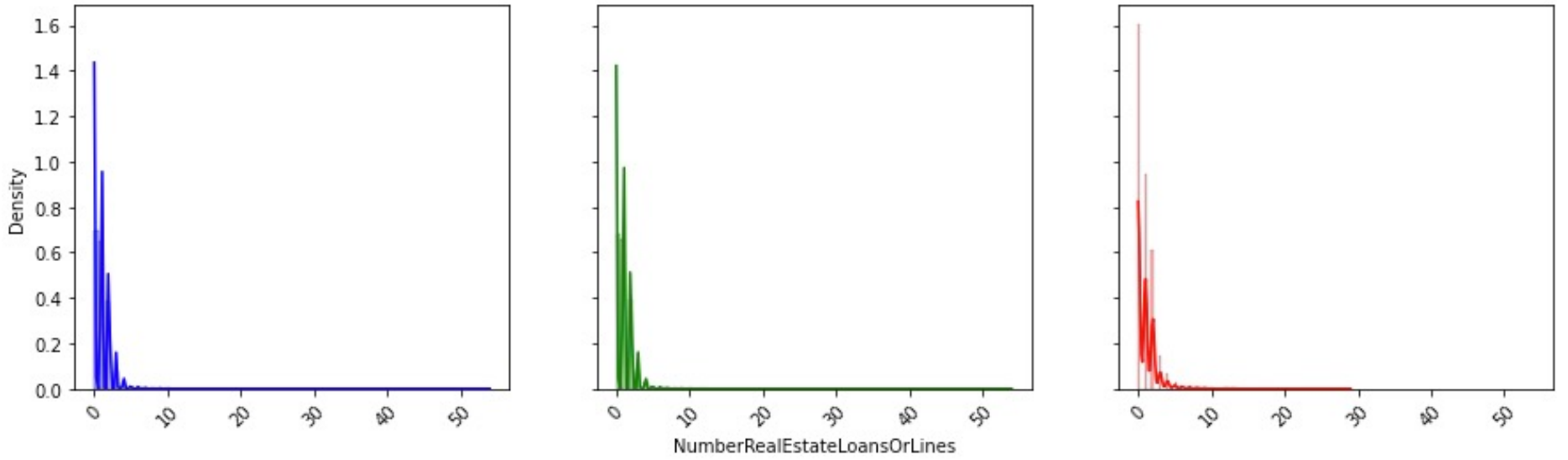
Problematic cases has lower monthly income



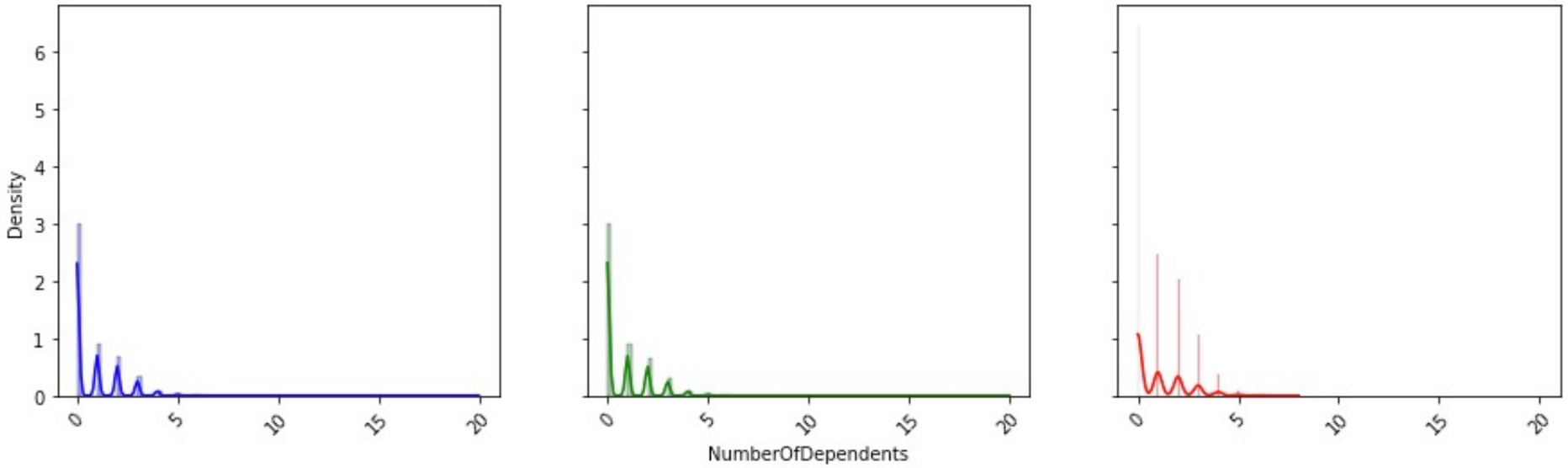
Problematic cases has more sizable number of people with 0 open credit lines and loans

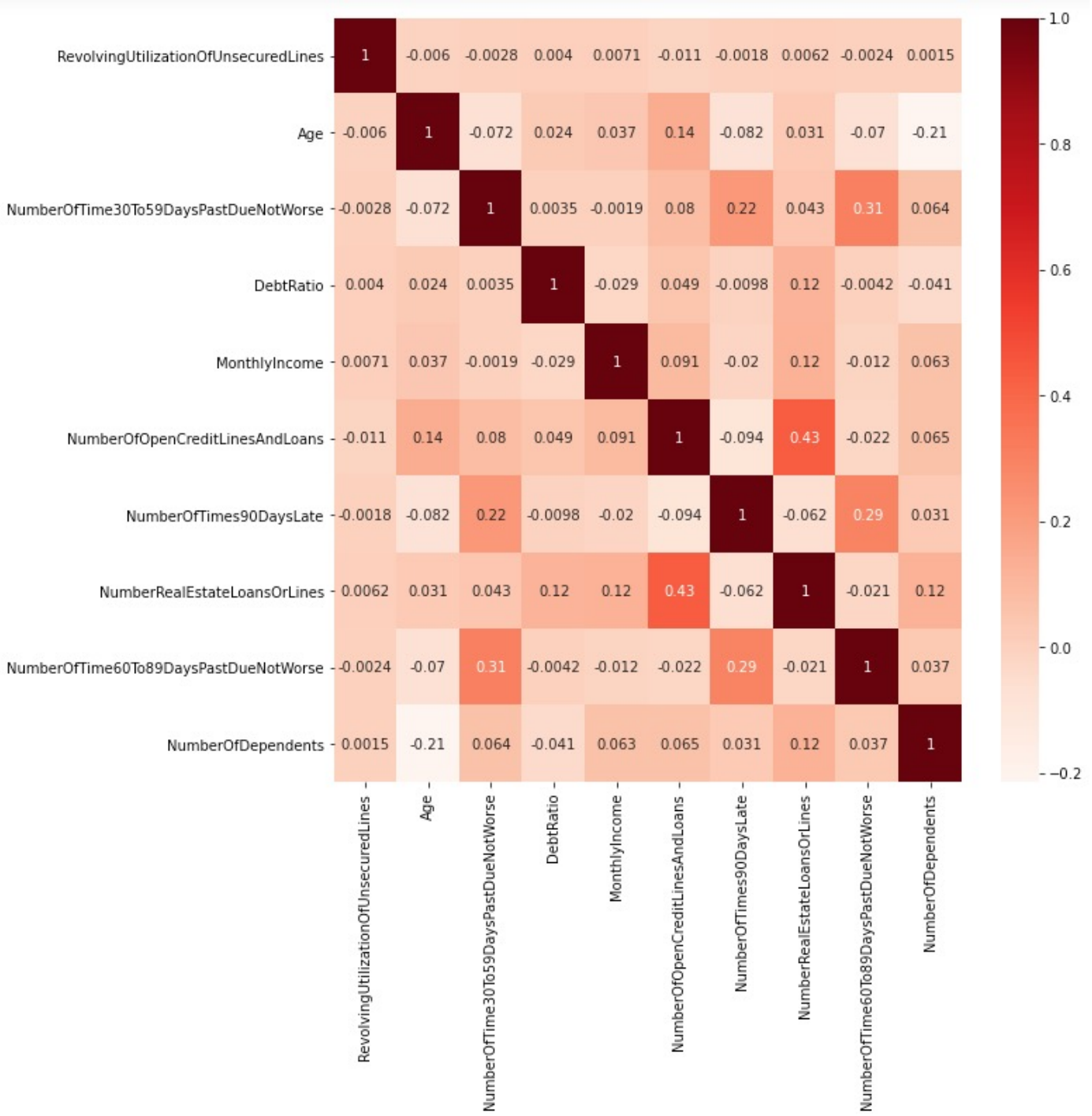


Problematic cases has lower number of real estate loans or lines



Problematic cases has lower number of dependents



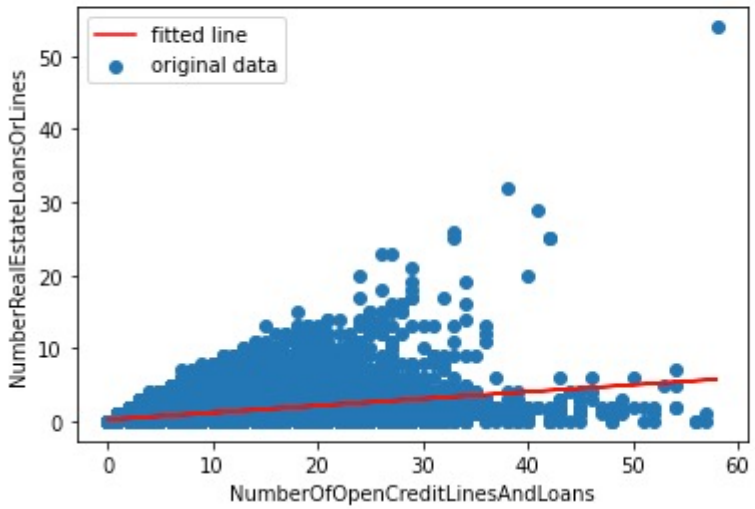


Correlation matrix (outliers excluded)

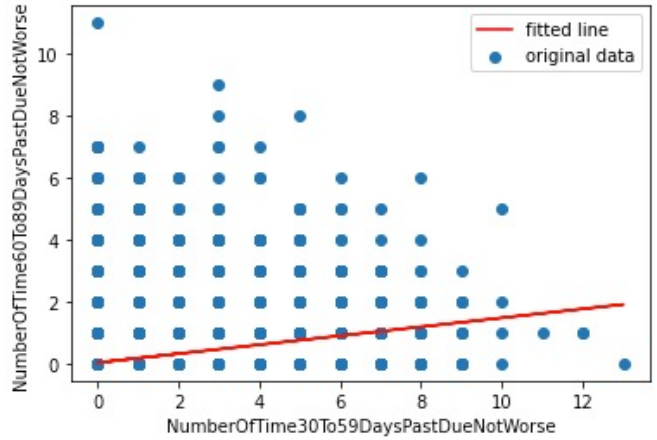
Top (+) correlation pairs:
['NumberOfOpenCreditLinesAndLoans',
'NumberRealEstateLoansOrLines']
['NumberOfTime30To59DaysPastDueNotWorse',
'NumberOfTime60To89DaysPastDueNotWorse']
['NumberOfTimes90DaysLate',
'NumberOfTime60To89DaysPastDueNotWorse']
['NumberOfTime30To59DaysPastDueNotWorse',
'NumberOfTimes90DaysLate']

Top (-) correlation pairs:
['Age','NumberOfDependents']

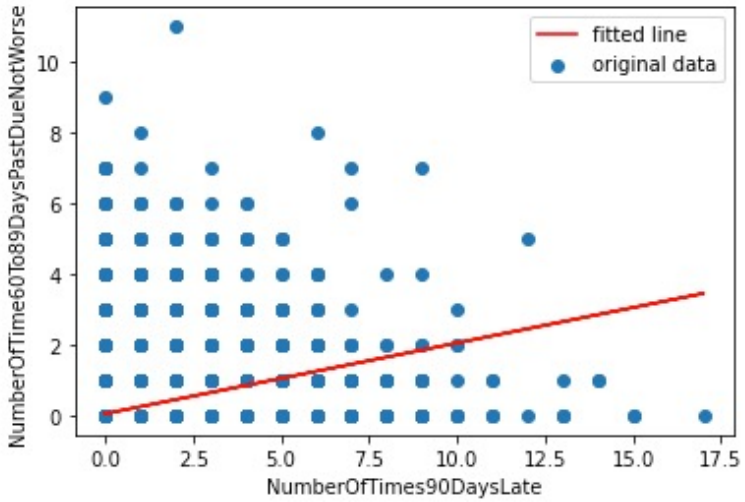
['NumberOfOpenCreditLinesAndLoans', 'NumberRealEstateLoansOrLines']
fitted line y=(0.09515076350729364)x+(0.21433908235395682)
r square: 0.18719853689339763



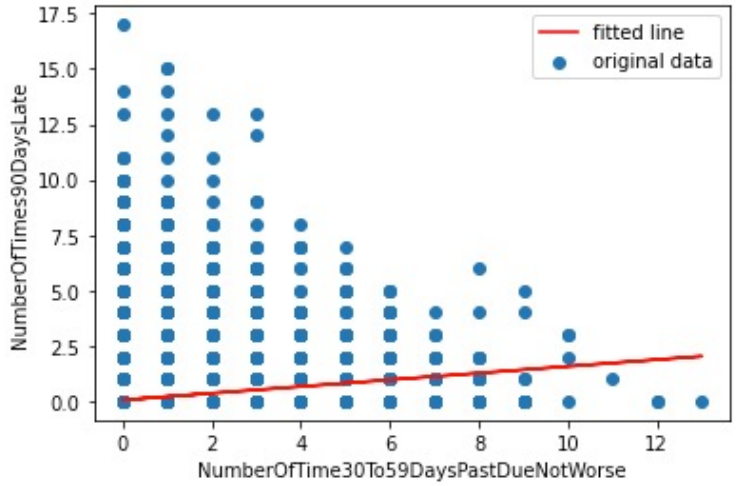
['NumberOfTime30To59DaysPastDueNotWorse', 'NumberOfTime60To89DaysPastDueNotWorse']
fitted line y=(0.1446897467875431)x+(0.02925902618013672)
r square: 0.093560356286794



['NumberOfTimes90DaysLate', 'NumberOfTime60To89DaysPastDueNotWorse']
fitted line y=(0.20030213510503217)x+(0.046704475907710784)
r square: 0.08681166287703788

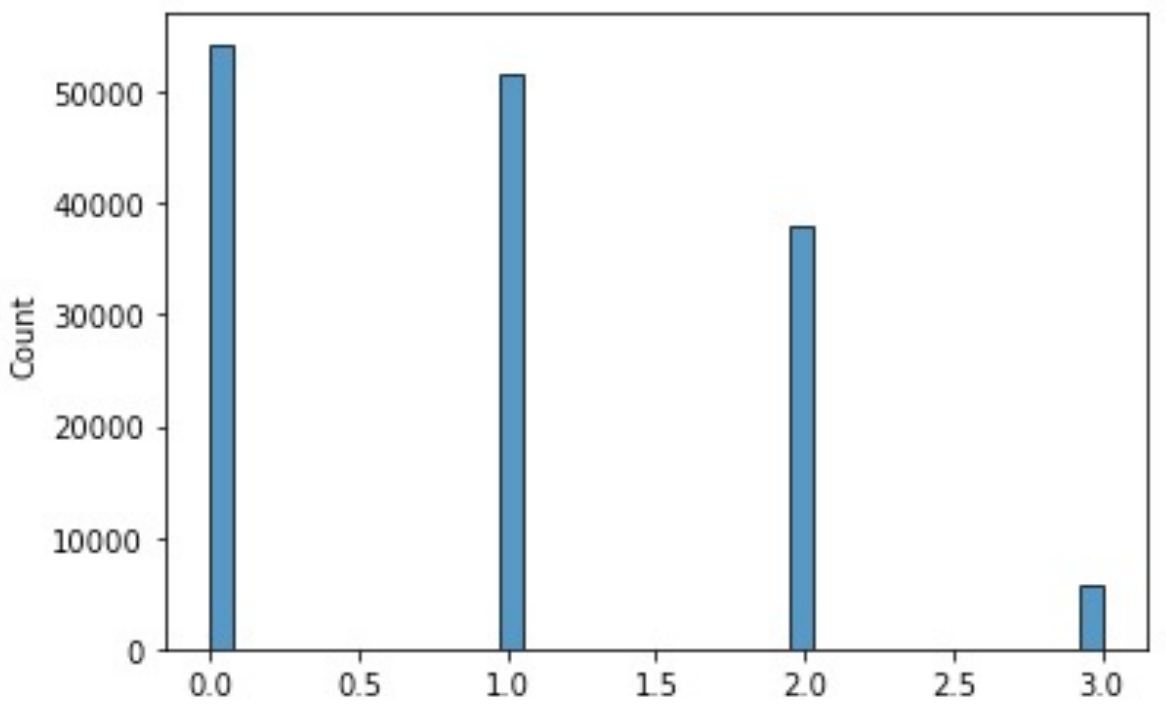


['NumberOfTime30To59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate']
fitted line y=(0.15179016619062546)x+(0.053146419336586356)
r square: 0.04758791845970719



Dataset clustering

- KMeans clustering.
- Steps
  - Exclude outliers.
  - Input income (median) and number of dependent (mode).
  - Fit to Kmeans with N=4.



Distribution by cluster 0..3.

Cluster profiles

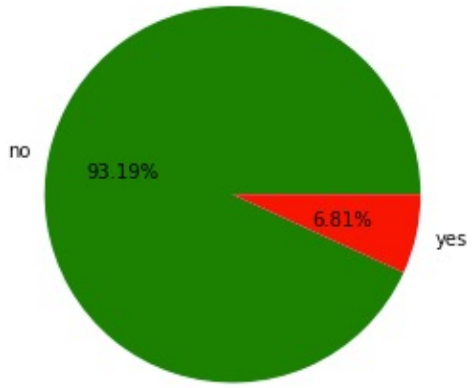
Cluster 0
- Highest revolving utilization of unsecured lines.
- Age 39 (lowest)
- Low monthly income.

Cluster 1
- Lowest number of dependents.
- Age 66 (highest).
- Slightly higher monthly income.

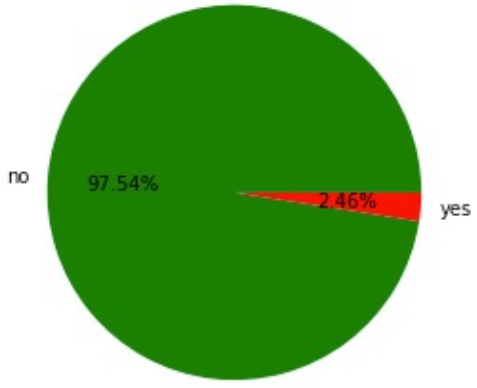
Cluster 2
- Lowest revolving utilization of unsecured lines.
- Age 53.
- Highest monthly income, debt ratio, number of open credits and loans.

Cluster 3
- Age 46.
- Low monthly income.
- Highest past dues from history (30-59 days, 60-89 days, over 90 days).

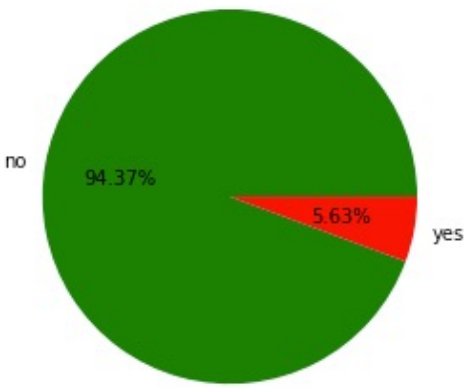
Label	Count	Size
no	50636	0.9319
yes	3698	0.0681



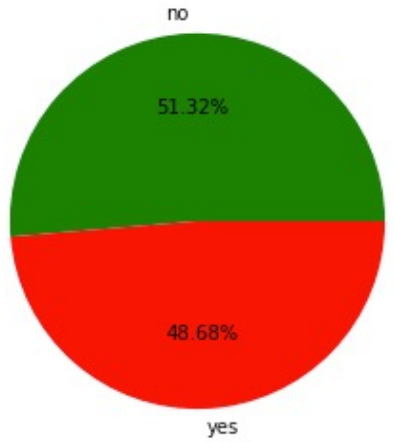
Label	Count	Size
no	50488	0.9754
yes	1274	0.0246



Label	Count	Size
no	35809	0.9437
yes	2138	0.0563



Label	Count	Size
no	2919	0.5132
yes	2769	0.4868



Discussions

- Data is skewed to no issue case (93.3%).
- May need to handle empty values on monthly income and number of dependents.
  - Empty monthly income 19.8%.
  - Empty number of dependents 2.6% (subset of empty monthly income).
- There are outliers with very high value on NumberOfTime30To59DaysPastDueNotWorse, NumberOfTime60To89DaysPastDueNotWorse, NumberOfTimes90DaysLate.
  - Outliers are correlated.
  - 0.18% of dataset.

Discussions

Compared to no-issue cases, problematic loan cases distribution has the following properties:

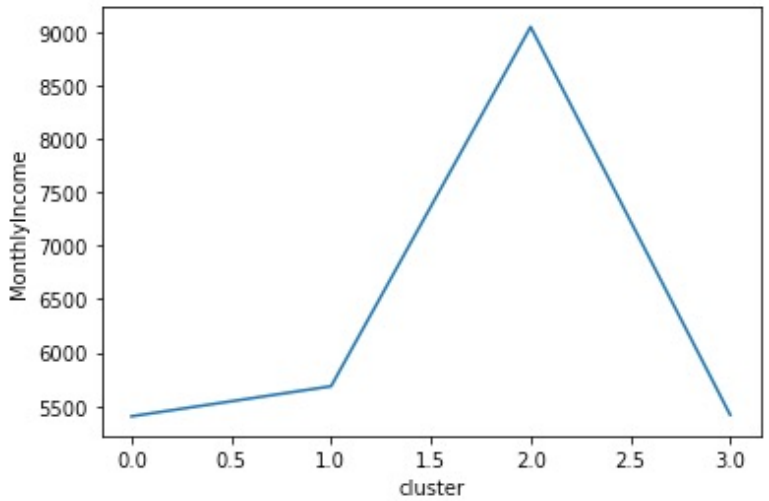
- Higher
  - Concentration on 20 < age < 60.
  - Number of past dues on 30-59 days, 60-89 days, > 90 days. Moreover, these three indicators are strongly correlated.
- Lower
  - Monthly income.
  - Number of credit loans and lines.
  - Number of real estate loans and lines.

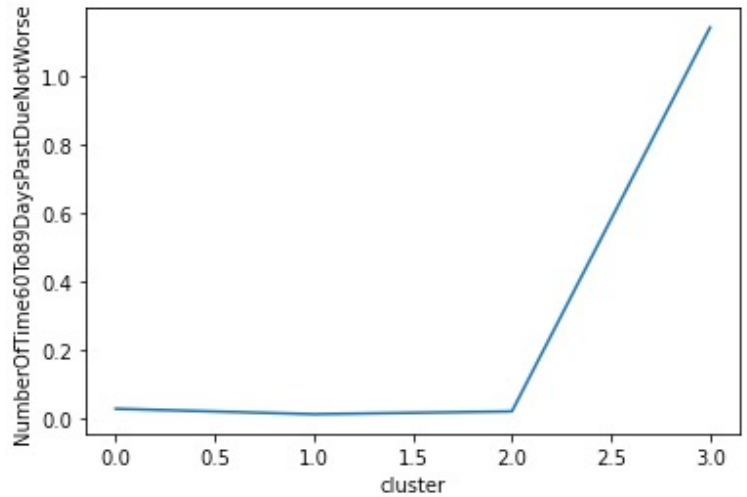
Discussions

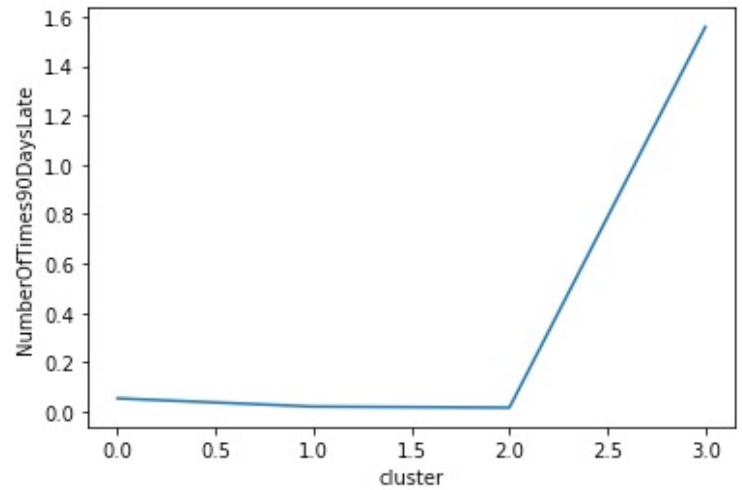
- Initial result of clustering may be beneficial for risk analysis/loan policy.
  - Cluster = stereotype = customer segment.
- Each segment may have different risk (lets say cluster 0 is the typical applicant).
  - Cluster 1: retirees.
    - Retirement age, no dependent.
    - Lowest risk.
  - Cluster 2: high income, high consumption.
    - Still in productive age, high income but also high debt ratio.
    - Slightly lower risk.
  - Cluster 3: repeat offenders.
    - Still in productive age, has history of past dues.
    - High risk.

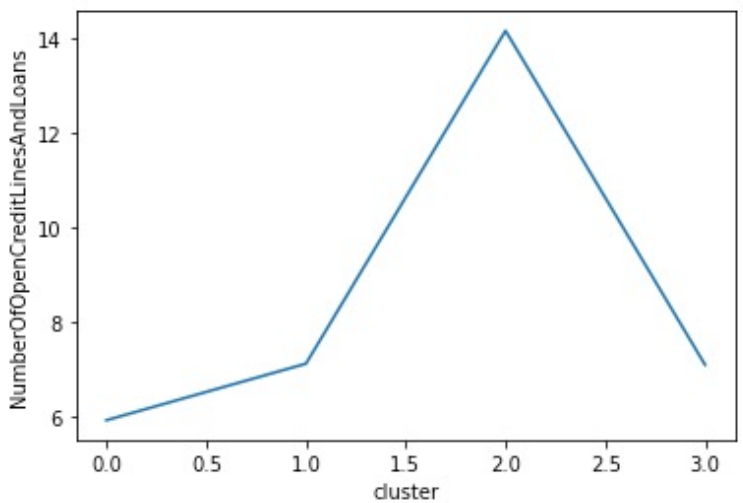
Thank you

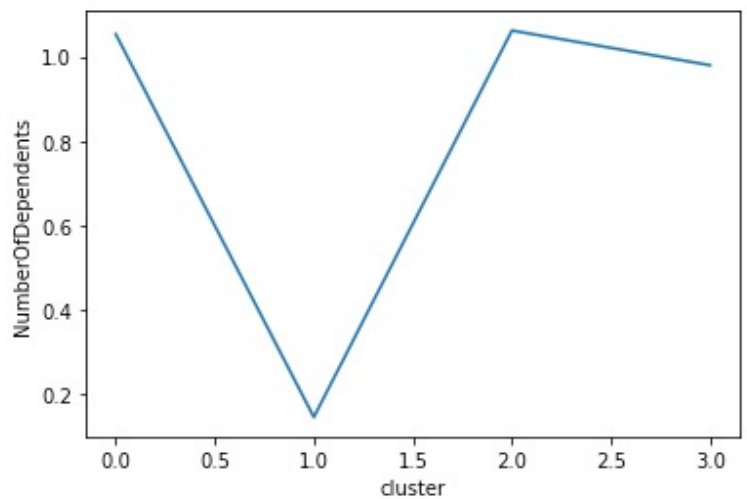
Cluster center profile

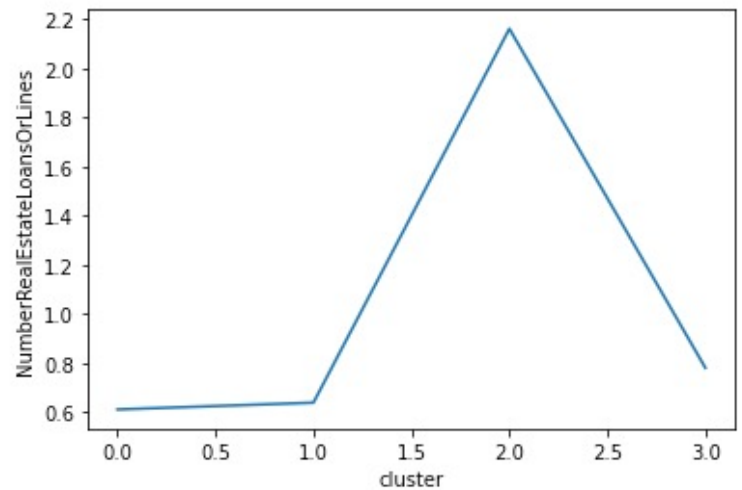












Cluster center profile