# Data exploration and preparation

**Shaima Alharbi**

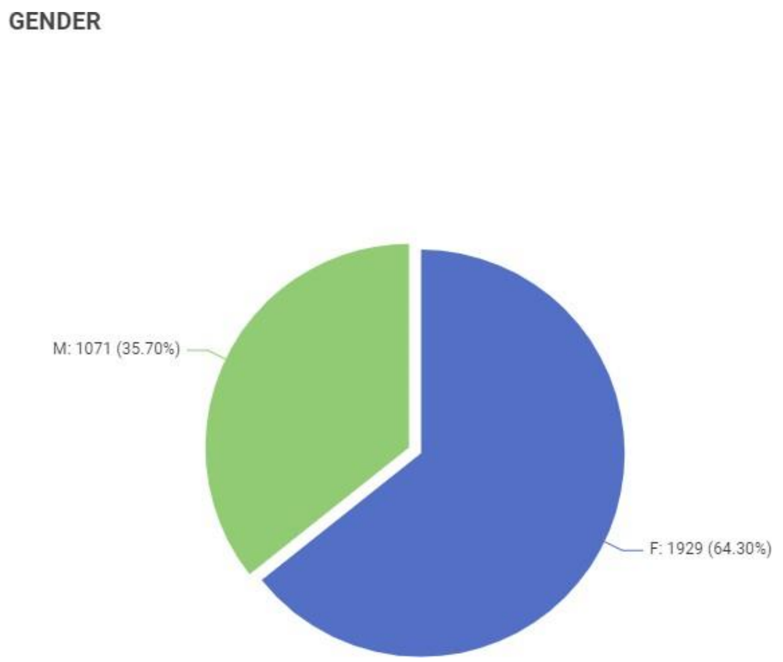
## 1A. Initial data exploration

### 1.1 SK_ID_CURR : Nominal

Sorting these values serves no purpose as they function as unique identification numbers for individual Customers across 3,000 records in the dataset.

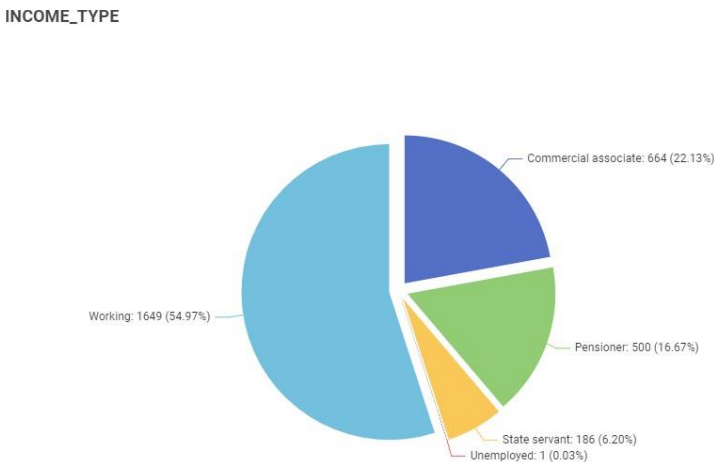
### 1.2 GENDER_TYPE : Nominal

There are 3,000 customer records, with gender distribution split Specifically approximately 64.30% of the total records are female (1929 records), while around 35.70%
are male (1071 records).



## 1.3 INCOME_TYPE : Nominal

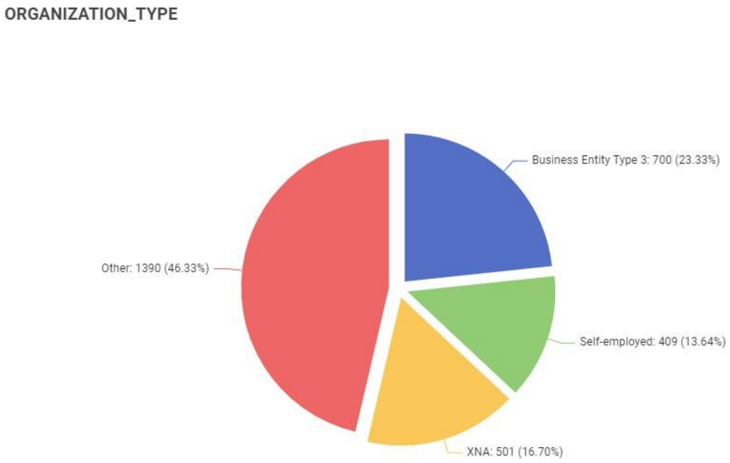
The working class, which makes up the majority of employees (54.97%), is the largest category. Commercial associates, who account for almost 20% of the workforce (22.13%), are the second largest group. At 16.67% and 6.20%, respectively, the workforce percentages for retirees and state employees are lower. In the sample, there is just one unemployed worker, which amounts to just 0.03% of the total workforce.
The fact that the working class makes up the majority of the workforce shows that they are essential to the operation of the business. Pensioners and business associates are both significant employment categories, although they serve different purposes inside the company. The business partners almost certainly contribute to the company's revenue generation.



## 1.4 ORGANIZATION_TYPE : Nominal

The largest group of workers (46.33% of the total employment) falls under the "Other" category. The business entity type 3 is the second-largest group, accounting for more than 20% (23.33%) of the workforce. At 13.64% and 16.70%, respectively, the self-employed and xna workers make up lesser percentages of the workforce.
The fact that the "Other" category has the most workers indicates that it is crucial to the economy. Although they serve diverse functions in the economy, the business entity type 3 and xna employees are nevertheless significant groupings of workers.

The "Other" category may be the biggest for the following reasons:
There are a variety of various organizations kinds that it might contain, including government agencies, non-profits, and educational institutions.

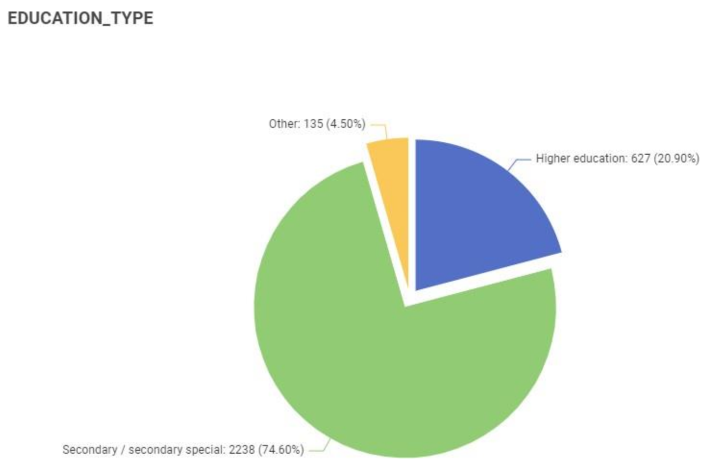


### 1.5 EDUCATION_TYPE : Nominal

Employees with secondary or special education make up the largest category (74.60%), followed by those with higher education (20.90%). With 4.50 percent of the total workforce, the "Other" category is the smallest.
The workforce appears to be generally well-educated based on the fact that the majority of employees have a secondary or special education. The fact that a sizable fraction of employees have a higher education, however, underscores the significance of this fact and implies that the economy has a strong need for trained personnel.

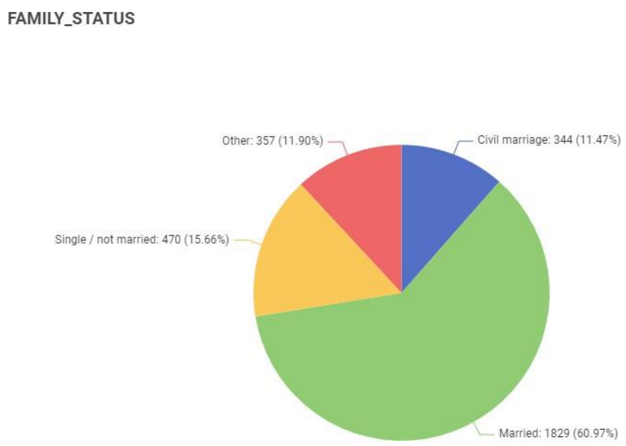
The secondary/special education group may be the largest for the following reasons:

- It could consist of a diverse variety of academic credentials, including high school degrees, occupational certificates, and trade apprenticeships.
- Employees in fields like manufacturing, construction, and retail that don't call for a college degree may be included.



### 1.6 FAMILY_STATUS : Nominal

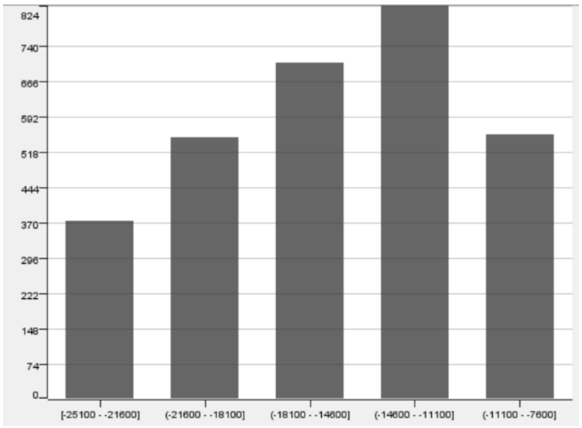
Married is the most common type, followed by single or never married, civil unions, and others.
This indicates that most of the company's employees are married. However, a sizable portion of employees are single or unmarried, while a lesser portion are involved in civil unions or other kinds of family arrangements.
The demands of the company's employees can be better understood with the help of this information. For instance, the business could want to provide extra benefits and assistance to workers who are single parents or who have other family obligations.



### 1.7 CNT_CHILDREN : Ratio

The graph demonstrates that as applicants have aged, they have had fewer children. 824 applications in the 25–34 age range have 1-4 children, which is the greatest percentage of applicants with children. Only 206 applicants in the 75+ age category have children, which is the lowest percentage of applicants with children.

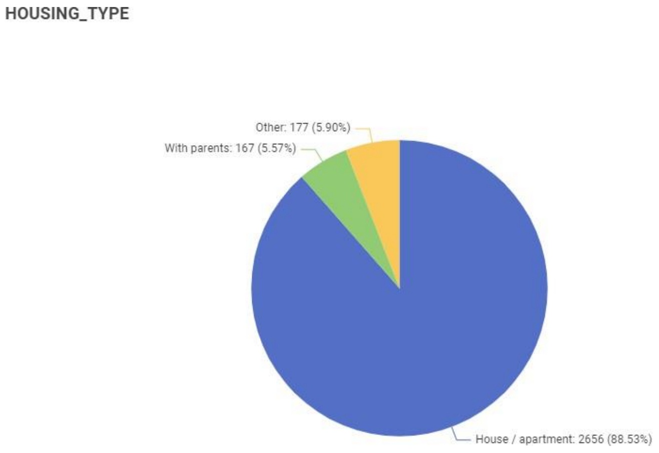
| | |
|---|---|
| Min | 0 |
| Max | 4 |
| Mean | 0.44800000000000056 |
| Median | 0 |
| Standard deviation | 0.7381130922690585 |
| Variance | 0.5448109369789917 |



### 1.8 HOUSING_TYPE : Nominal

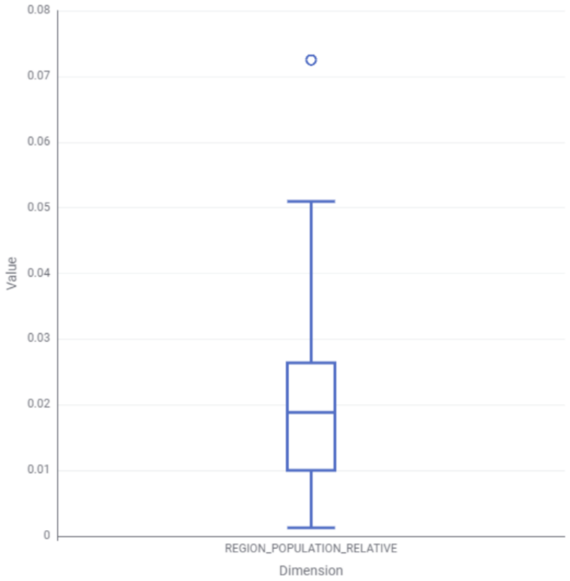
A company's 3000 employees' housing preferences are depicted in a pie chart. According to the pie chart, the vast majority of employees (88.53%) reside in homes or apartments. Following this are 5.90% of employees who live in other housing types, like townhomes or condos, and 5.57% of employees who live with their parents.

This information can be used by the business to decide where to put its facilities and offices. For instance, if the majority of employees live in homes and apartments, the business might wish to put its offices in places that are easily accessible.



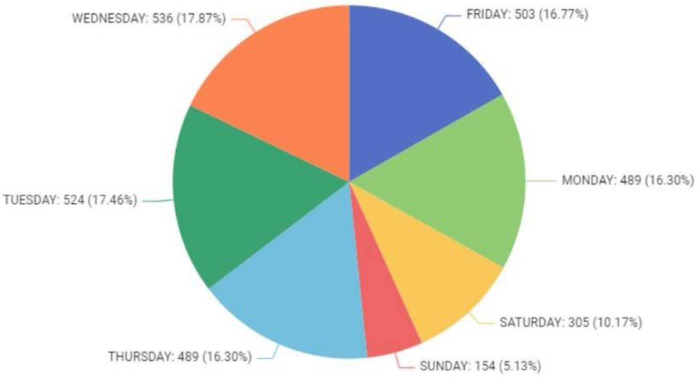
## 1.9 REGION_POPULATION_RELATIVE : Ratio

The box plot reveals that most clients reside in areas with a normalized population of between 0.04 and 0.06 per square mile. Some consumers reside in areas with a normalized population above 0.06, but they are the exception.



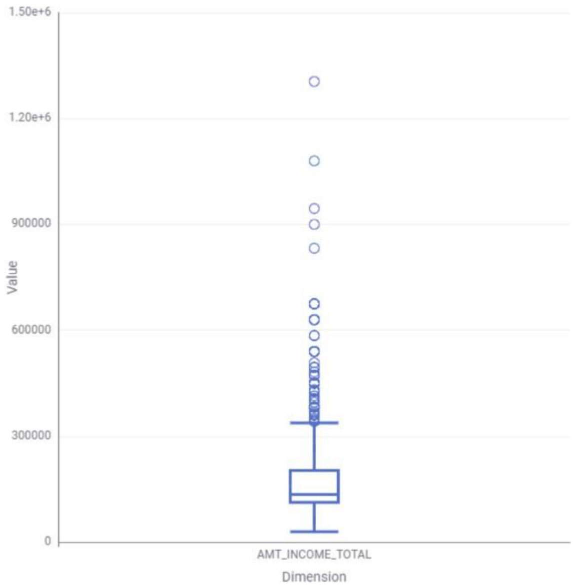
## 1.10 WEEK_DAY_APPR_PROCESS_START : NOMINAL

The graph reveals that Wednesday, Friday, and Tuesday are the three days of the week when people most frequently request for loans. Saturday and Sunday are the days of the week when people request for loans the least.



### 1.11 AMT_INCOME_TOTAL : Ratio

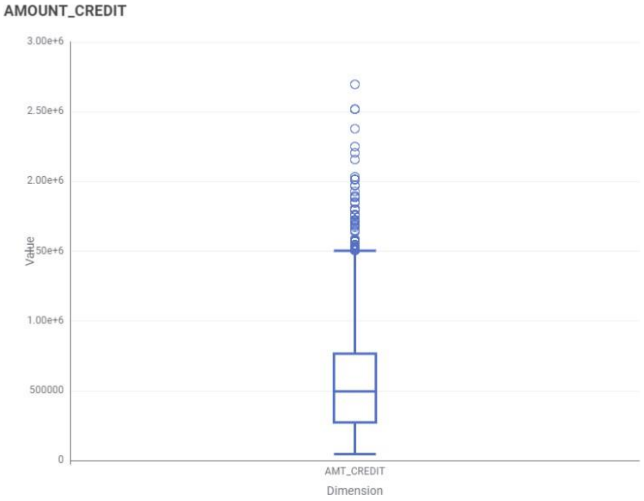
The scatter plot demonstrates a significant positive relationship between total income and population.



### 1.12 AMOUNT_CREDIT : Ratio

The distribution of credit for 3,000 employees of a company is shown by the box and whisker plot. The box represents the middle 50% of the data, and the horizontal line inside the box indicates the median (the middle number). 90% of the data is contained inside the whiskers, which reach the 5th and 95th percentiles. Outliers are any data points that lie outside of the whiskers.
The company can use this chart to Recognize workers with high or low credit scores. The use of this data for risk assessment or marketing campaign targeting is possible.

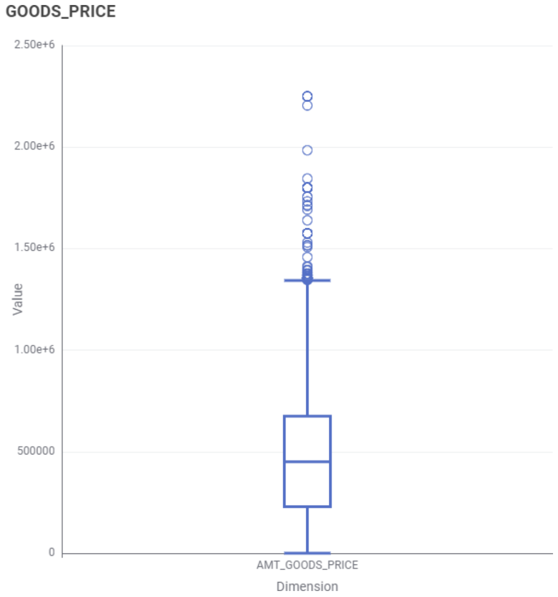
| | |
|---|---|
| Min | 45000.0 |
| Max | 2695500.0 |
| Mean | 569824.012499999 |
| Median | 495000.0 |
| Standard deviation | 373319.8913441097 |
| Variance | 1.393677412731779E11 |



### 1.13 GOODS_PRICE : Ratio

The box and whisker plot you have shown illustrates how the cost of the things an applicant purchases is distributed. The box represents the middle 50% of the data, and the horizontal line inside the box indicates the median (the middle number). 90% of the data is contained inside the whiskers, which reach the 5th and 95th percentiles. Outliers are any data points that lie outside of the whiskers.

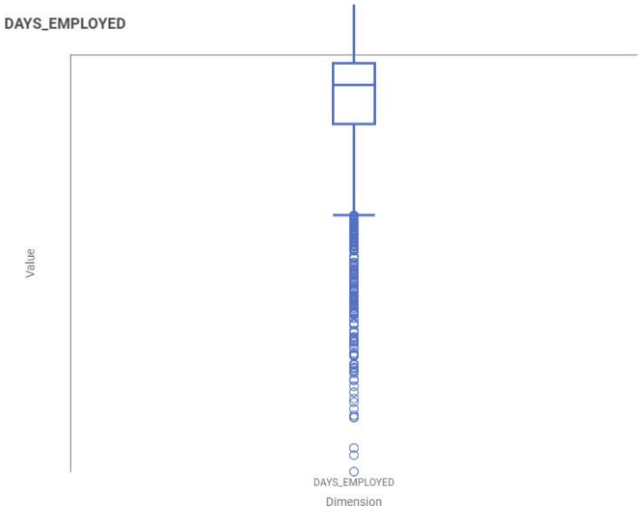
| | |
|---|---|
| Min | 0 |
| Max | 2250000.0 |
| Mean | 505787.60700000037 |
| Median | 450000.0 |
| Standard deviation | 339865.22017428797 |
| Variance | 1.1550836788411723E11 |



### 1.14 DAYS_EMPLOYED : Ratio

The distribution of the number of days that each applicant's present employment began prior to the application is shown in the graph. The number of days till application is displayed on the x-axis, while the number of applicants is displayed on the y-axis. The box represents the middle 50% of the data, and the horizontal line inside the box indicates the median (the middle number). 90% of the data is contained inside the whiskers, which reach the 5th and 95th percentiles. Outliers are any data points that lie outside of the whiskers.

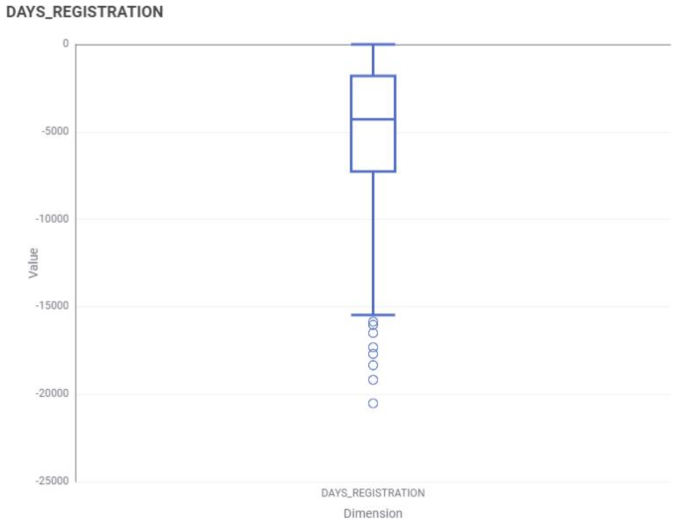
| | |
|---|---|
| Min | -14318 |
| Max | 365243 |
| Mean | 59248.91833333334 |
| Median | -1040.0 |
| Standard deviation | 137045.90207754137 |
| Variance | 1.878157927624706E10 |



### 1.15 DAYS_REGISTRATION: Ratio

The box and whisker plot you offered displays the distribution of the amount of days a client altered their registration before the application. The y-axis displays the number of clients, while the x-axis displays the days till application. The box represents the middle 50% of the data, and the horizontal line inside the box indicates the median (the middle number). 90% of the data is contained inside the whiskers, which reach the 5th and 95th percentiles. Outliers are any data points that lie outside of the whiskers.

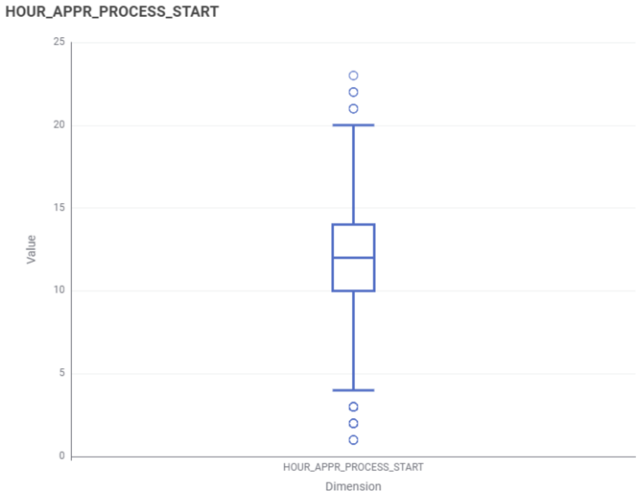
| | |
|---|---|
| Min | -20514.0 |
| Max | 0 |
| Mean | -4815.106000000001 |
| Median | -4285.0 |
| Standard deviation | 3528.7138805303503 |
| Variance | 1.2451821650647562E7 |



### 1.16 HOUR_APPR_PROCESS_START : Ratio

The figure displays the distribution of the hour that loan applications were submitted by clients. The hour of the day is indicated on the x-axis, while the number of clients is indicated on the y-axis. The box represents the middle 50% of the data, and the horizontal line inside the box indicates the median (the middle number). 90% of the data is contained inside the whiskers, which reach the 5th and 95th percentiles. Outliers are any data points that lie outside of the whiskers.

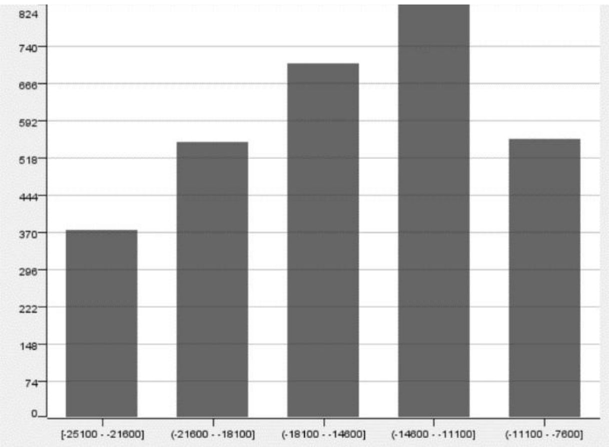
| | |
|---|---|
| Min | 1 |
| Max | 23 |
| Mean | 11.871666666666705 |
| Median | 12 |
| Standard deviation | 3.2965870812630405 |
| Variance | 10.867486384350372 |



### 1.17 AMT_ANNUITY : Ratio

The graph demonstrates how an applicant's average annuity amount declines as their income rises. The applicants with earnings less than 7,600 receive the highest average annuity amount, which is 222. The applicants with salaries of at least 25,100 receive an average annuity of 824, which is the lowest amount.

| | |
|---|---|
| Min | 45000.0 |
| Max | 2695500.0 |
| Mean | 569824.012499999 |
| Median | 495000.0 |
| Standard deviation | 373319.8913441097 |
| Variance | 1.393677412731779E11 |

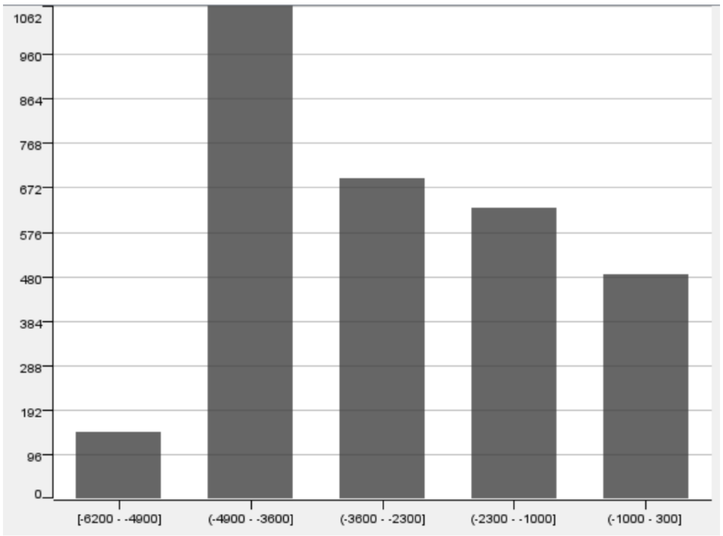


### 1.18 DAYS_ID_PUBLISH: Ratio

100 days make up the average number of days. Accordingly, half of the clients amended their identity document more than 100 days before to the application, and the other half did so fewer than 100 days prior.

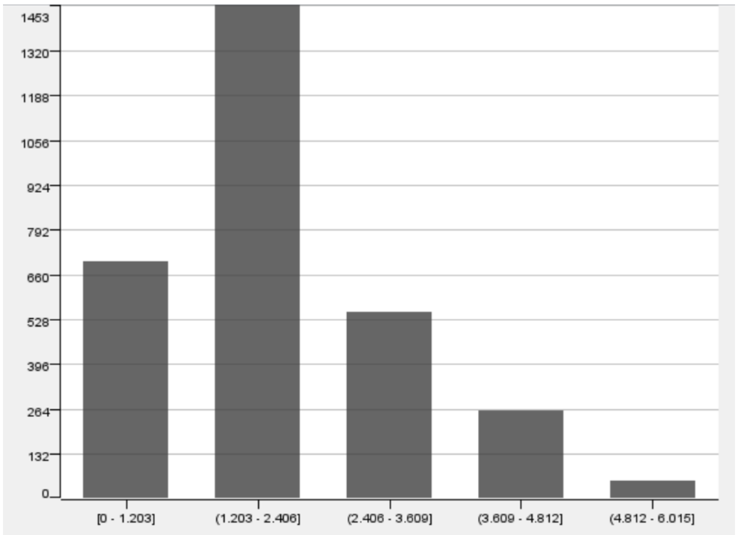
Between 0 and 400 days make up the 95th and 5th percentiles, respectively. This indicates that 90% of the applicants had their identity documents altered during the last 400 days before applying.

There are a small number of outliers, or clients, who updated their identity document less than 0 days or more than 400 days prior to the application.



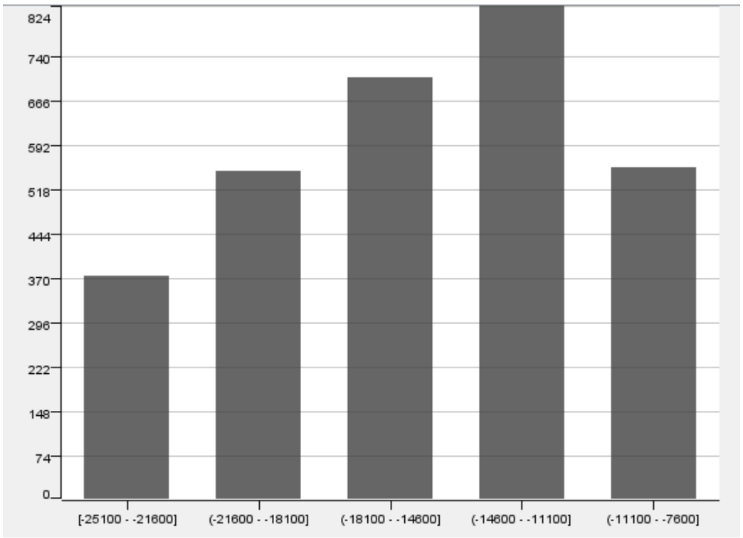
### 1.19 CNT_FAM_MEMBERS: Ratio

The number of clients is shown on the y-axis, while the number of family members is shown on the x-axis. The box represents the middle 50% of the data, and the horizontal line inside the box indicates the median (the middle number). 90% of the data is contained inside the whiskers, which reach the 5th and 95th percentiles. Outliers are any data points that lie outside of the whiskers.



### 1.20 DAYS_BIRTH : Ratio

The chart shows that the average client age is 7407 days at the time of application. This is equivalent to 20.3 years. The chart also shows that there is a wide range of client ages, with some clients being as young as 148 days old and others being as old as 25100 days old.

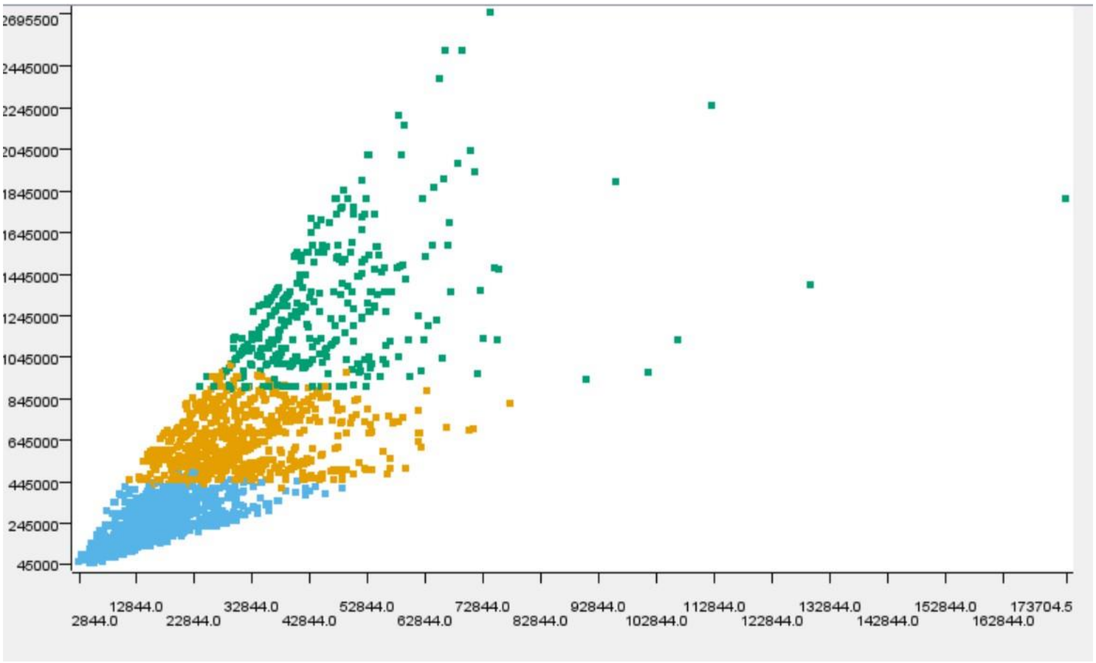


## 3. Clustering

The scatter plot's clustering of the applicant's total borrowing capacity and total income amount can be explained by two different things, including:

Creditworthiness: Generally speaking, applicants with greater incomes are viewed as having better credit, which indicates that they are more likely to be able to repay their debts. Lenders often are more inclined to lend money to candidates with higher earnings as a result.

Debt-to-income ratio: When determining how much money to lend an applicant, lenders also take into account their debt-to-income ratio. The applicant's monthly income is divided by the applicant's total monthly debt payments to get the applicant's debt-to-income ratio. In general, lenders view applicants with lower debt-to-income ratios as more creditworthy since they have more money available to pay back their loans.



## 1B- Data preprocessing :

### 1. Use the following binning techniques to smooth the values of the following two attributes

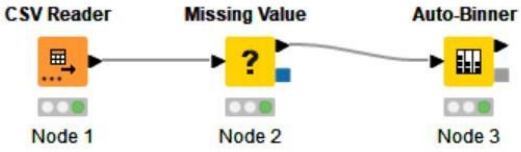
- DAYS_EMPLOYED

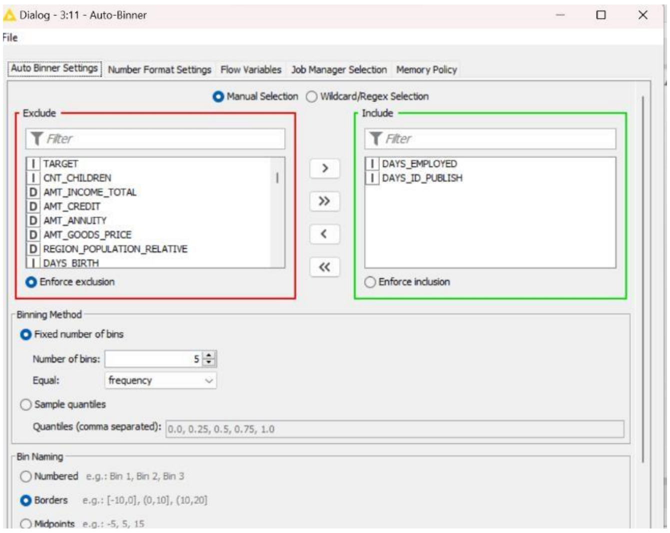
- DAYS_ID_PUBLISH

To begin, I load the data into the file reader node and link it to the missing value, and

auto-binner nodes.

Next, I chose the attribute in the auto-binner node's setup, set the default bins number of 5 for it, and chose the two types of binning which are equi-width and frequency.




Dialog - 3:11 - Auto-Binner

File

Auto Binner Settings | Number Format Settings | Flow Variables | Job Manager Selection | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection

Exclude: I TARGET, I CNT_CHILDREN, D AMT_INCOME_TOTAL, D AMT_CREDIT, D AMT_ANNUITY, D AMT_GOODS_PRICE, D REGION_POPULATION_RELATIVE, I DAYS_BIRTH

● Enforce exclusion

Include: I DAYS_EMPLOYED, I DAYS_ID_PUBLISH

○ Enforce inclusion

Binning Method

● Fixed number of bins

Number of bins: 5

Equal: frequency

○ Sample quantiles

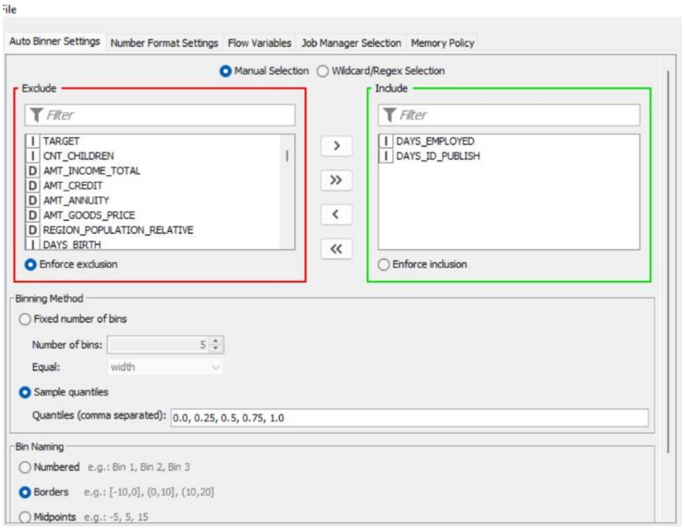
Quantiles (comma separated): 0.0, 0.25, 0.5, 0.75, 1.0

Bin Naming

○ Numbered e.g.: Bin 1, Bin 2, Bin 3

● Borders e.g.: [-10,0], (0,10], (10,20]

○ Midpoints e.g.: -5, 5, 15


File

Auto Binner Settings | Number Format Settings | Flow Variables | Job Manager Selection | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection

Exclude: I TARGET, I CNT_CHILDREN, D AMT_INCOME_TOTAL, D AMT_CREDIT, D AMT_ANNUITY, D AMT_GOODS_PRICE, D REGION_POPULATION_RELATIVE, I DAYS_BIRTH

● Enforce exclusion

Include: I DAYS_EMPLOYED, I DAYS_ID_PUBLISH

○ Enforce inclusion

Binning Method

○ Fixed number of bins

Number of bins: 5

Equal: width

● Sample quantiles

Quantiles (comma separated): 0.0, 0.25, 0.5, 0.75, 1.0

Bin Naming

○ Numbered e.g.: Bin 1, Bin 2, Bin 3

● Borders e.g.: [-10,0], (0,10], (10,20]

○ Midpoints e.g.: -5, 5, 15

## 2. Use the following techniques to normalise the following attribute

- AMT_INCOME_TOTAL

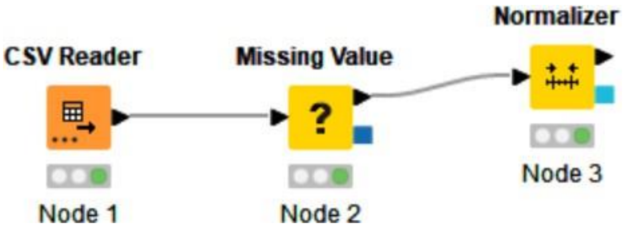
Normalization via min-max

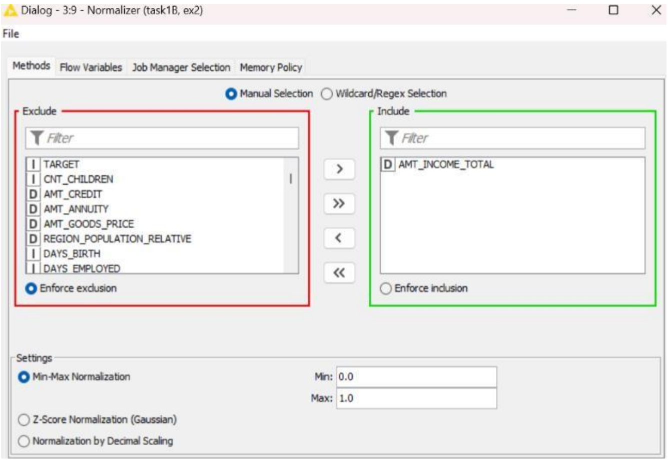
1- To begin, I load the data into the file reader node and link it to the missing value, and normalizer nodes.

2- I chose the AMT_INCOME_TOTAL attribute in the normalizer nodes, set min-max normalization.

Normalization via z-score

3- For implementing z-score normalization I chose the AMT_INCOME_TOTAL attribute in the normalizer node, and set z-score normalization.




Dialog - 3:9 - Normalizer (task1B, ex2)

File

Methods | Flow Variables | Job Manager Selection | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection

Exclude: Filter

- I TARGET
- I CNT_CHILDREN
- D AMT_CREDIT
- D AMT_ANNUITY
- D AMT_GOODS_PRICE
- D REGION_POPULATION_RELATIVE
- I DAYS_BIRTH
- I DAYS_EMPLOYED

● Enforce exclusion

Include: Filter

- D AMT_INCOME_TOTAL

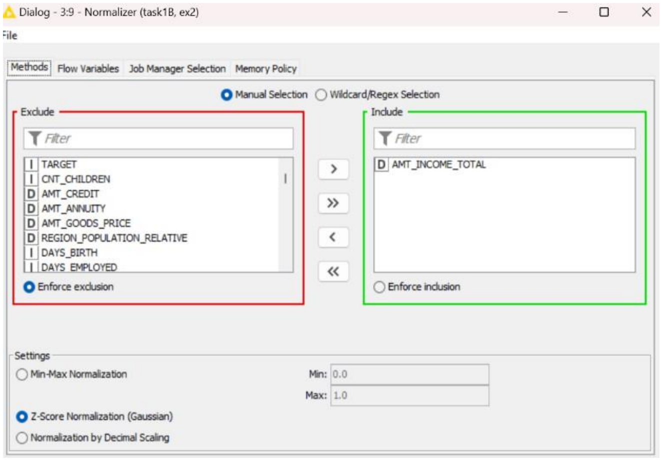
○ Enforce inclusion

Settings

● Min-Max Normalization Min: 0.0 Max: 1.0

○ Z-Score Normalization (Gaussian)

○ Normalization by Decimal Scaling


Dialog - 3:9 - Normalizer (task1B, ex2)

File

Methods | Flow Variables | Job Manager Selection | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection

Exclude: Filter

- I TARGET
- I CNT_CHILDREN
- D AMT_CREDIT
- D AMT_ANNUITY
- D AMT_GOODS_PRICE
- D REGION_POPULATION_RELATIVE
- I DAYS_BIRTH
- I DAYS_EMPLOYED

● Enforce exclusion

Include: Filter

- D AMT_INCOME_TOTAL

○ Enforce inclusion

Settings

○ Min-Max Normalization Min: 0.0 Max: 1.0

● Z-Score Normalization (Gaussian)

○ Normalization by Decimal Scaling

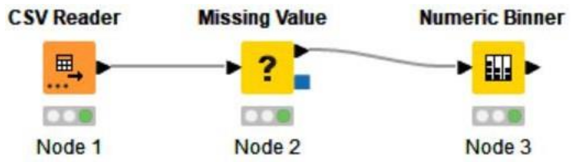
### 3. Discretise the DAYS_BIRTH attribute into the following categories:

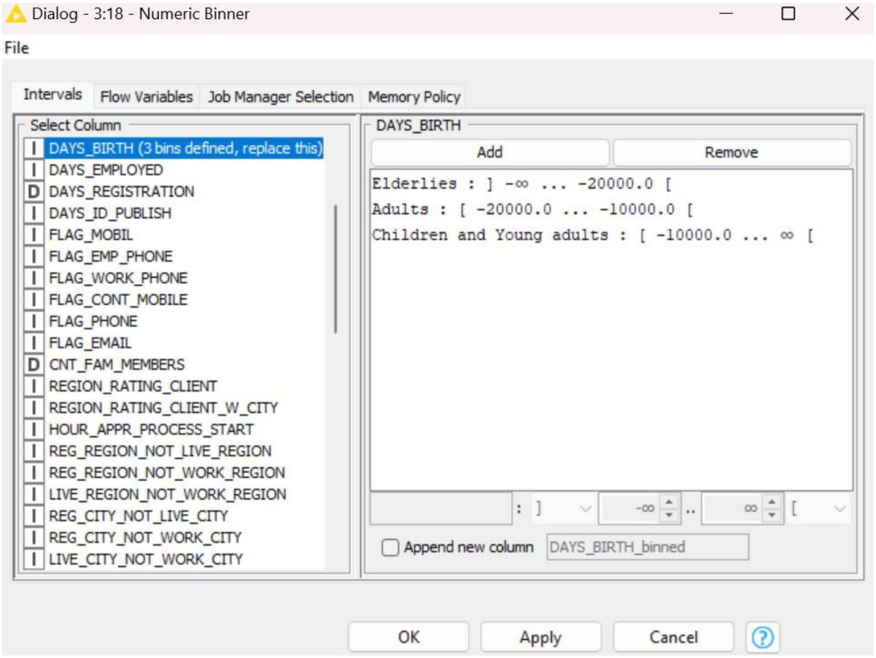
- (0 – 10,000-) Children and Young adults

- (10,000 - — 20,000-) Adults

- (20,000- — 30,000-) Elderlies

1- To begin, I load the data into the file reader node and link it to the missing value, and discretize the Numeric binner nodes as Elderlies : [ -∞ ... -20000.0 ], Adults : [ -20000.0 ... -10000.0 ], Children and Young adults : [ -10000.0 ... ∞ ]





Select Column

- I DAYS_BIRTH (3 bins defined, replace this)
- I DAYS_EMPLOYED
- D DAYS_REGISTRATION
- I DAYS_ID_PUBLISH
- I FLAG_MOBIL
- I FLAG_EMP_PHONE
- I FLAG_WORK_PHONE
- I FLAG_CONT_MOBILE
- I FLAG_PHONE
- I FLAG_EMAIL
- D CNT_FAM_MEMBERS
- I REGION_RATING_CLIENT
- I REGION_RATING_CLIENT_W_CITY
- I HOUR_APPR_PROCESS_START
- I REG_REGION_NOT_LIVE_REGION
- I REG_REGION_NOT_WORK_REGION
- I LIVE_REGION_NOT_WORK_REGION
- I REG_CITY_NOT_LIVE_CITY
- I REG_CITY_NOT_WORK_CITY
- I LIVE_CITY_NOT_WORK_CITY

DAYS_BIRTH

```
Elderlies : ] -∞ ... -20000.0 [
Adults : [ -20000.0 ... -10000.0 [
Children and Young adults : [ -10000.0 ... ∞ [
```

: ] -∞ .. ∞ [

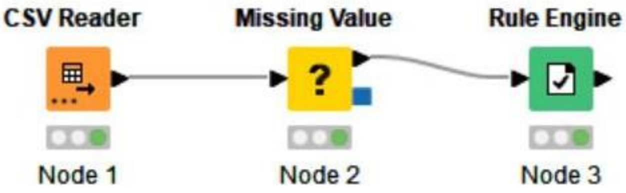
□ Append new column DAYS_BIRTH_binned

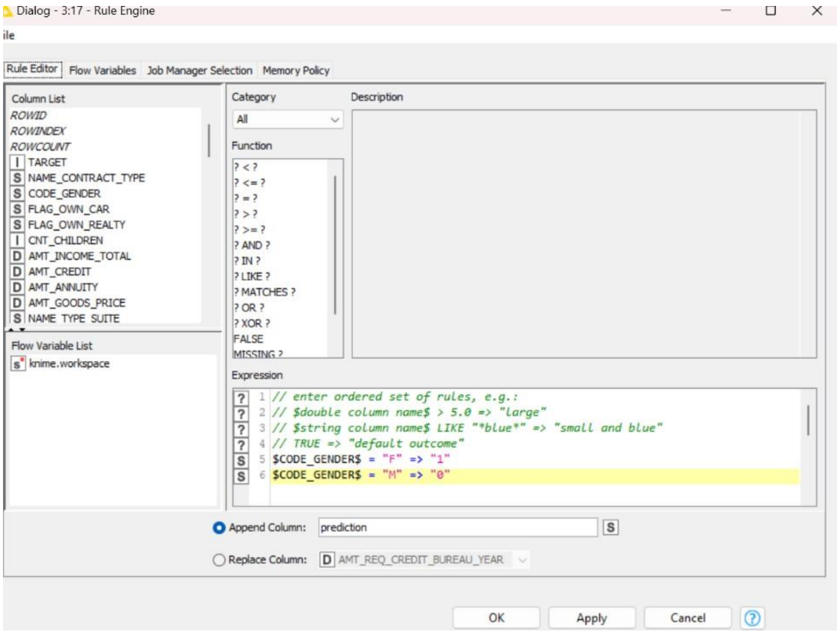
### 4. Binarise the CODE_GENDER variable [with values "0" or "1"]

1- To begin, I load the data into the file reader node, link it to the missing value, and

Rule Engine nodes.

2- I have written a code identifying that the values represented as women (F) as number = 1 and men (M) as number = 0.





```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => "large"
// $string column name$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
$CODE_GENDER$ = "F" => "1"
$CODE_GENDER$ = "M" => "0"
```

● Append Column: prediction

○ Replace Column: AMT_REQ_CREDIT_BUREAU_YEAR

## 1C. Summary

To improve its suitability for machine learning modeling, the data in this assessment was smoothed, normalized, and discretized. And Because lenders view applicants with larger salaries as being more creditworthy, they typically have greater borrowing capacity. This is due to the fact that lenders think applicants with greater wages will be better able to pay back their obligations. Additionally, applicants with higher salaries might possess a greater variety of assets that might be pledged as security for a loan.

Lenders utilize the debt-to-income ratio (DTI) to determine a candidate's creditworthiness, which is another factor contributing to the clustering. By dividing a person's monthly debt payments by their monthly income, the DTI is determined. Because they have more money available to pay down their debts, applicants with lower DTIs are viewed as being more creditworthy.

Therefore, the fact that lenders utilize income and DTI to evaluate an applicant's creditworthiness can account for the clustering of the applicant's total borrowing capacity and total income amount. A candidate is considered to be more creditworthy and has a higher borrowing capacity if they have a higher income and lower DTI.