EDA : Bank Loan Risk Analysis

Chaitanya Prabhune• 01.07.2024

Overview

This case study aims to identify the key factors driving loan defaults. Specifically, we want to pinpoint variables that serve as strong indicators of default risk. The insights gained will guide the company in making informed decisions, such as adjusting loan amounts, denying loans to riskier applicants, and offering higher interest rates. Additionally, it will help prevent the rejection of applications from creditworthy individuals.

Data Cleaning for Application data

Handling Missing Value

- Ignore columns having missing value more than 45%
- OWN_CAR_AGE is missing because users might not own cars
- Fill missing occupation types from income type
- Use mean value for AMT_ANNUITY, EXT_SOURCE_3, EXT_SOURCE_2 and mode for CODE_GENDER, NAME_TYPE_SUITE, AMT_REQ_CREDIT_BUREAU_x and remove missing rows for AMT_GOODS_PRICE

Data Cleaning for Application data

Remove non required columns

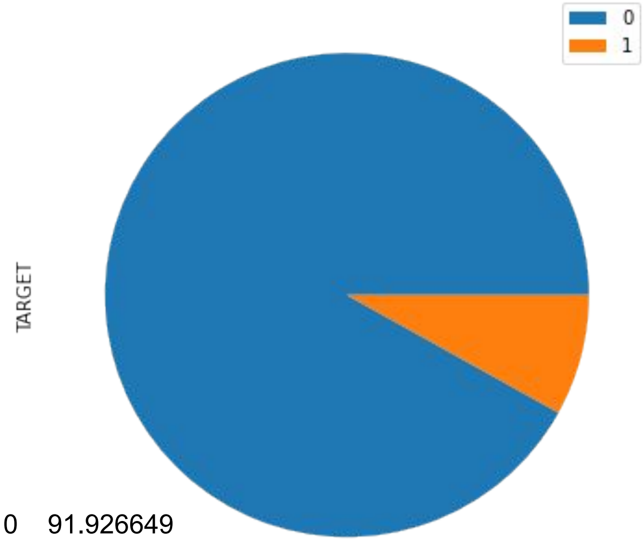
- Remove Flag based columns few others like DAYS_REGISTRATION , WEEKDAY_APPR_PROCESS_START, HOUR_APPR_PROCESS_START, LIVE_REGION_NOT_WORK_REGION, REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY,
- Remove columns like LIVE_CITY_NOT_WORK_CITY, DAYS_LAST_PHONE_CHANGE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, NAME_TYPE_SUITE

Data Cleaning for Application data

Handling Data Errors

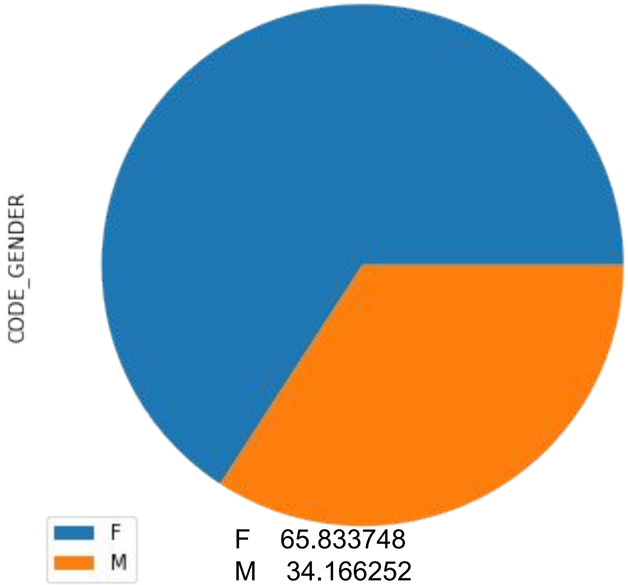
- Convert Data of DAYS_BIRTH to absolute value and convert to AGE
- Create range of data like for AGE for better analysis
- Fix outliers for AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, AGE

Data Analysis for Application data



```
0    91.926649
1     8.073351
Name: TARGET
```

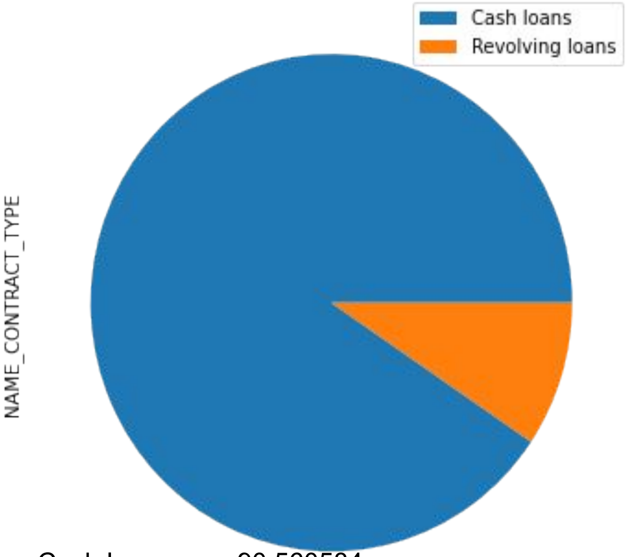
91% are non Defaulters



```
F    65.833748
M    34.166252
Name: CODE_GENDER
```

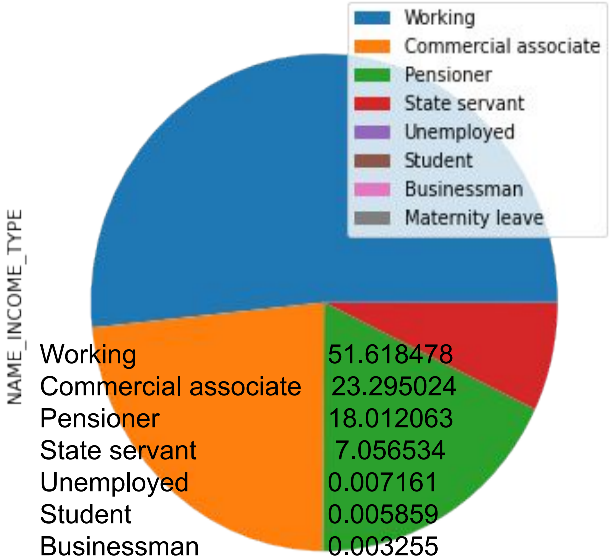
Female applicants are more

Data Analysis for Application data



```
Cash loans         90.560584
Revolving loans     9.439416
Name: NAME_CONTRACT_TYPE
```

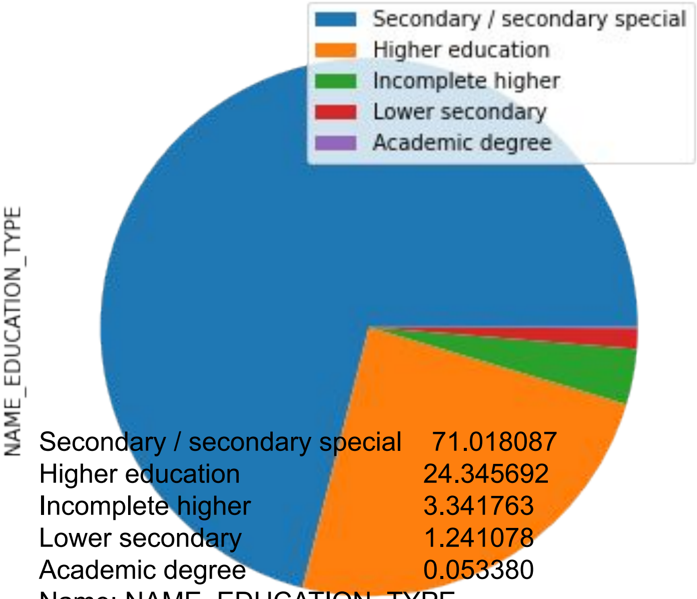
Cash loans are more



```
Working                 51.618478
Commercial associate    23.295024
Pensioner               18.012063
State servant            7.056534
Unemployed              0.007161
Student                 0.005859
Businessman             0.003255
Maternity leave         0.001627
Name: NAME_INCOME_TYPE
```

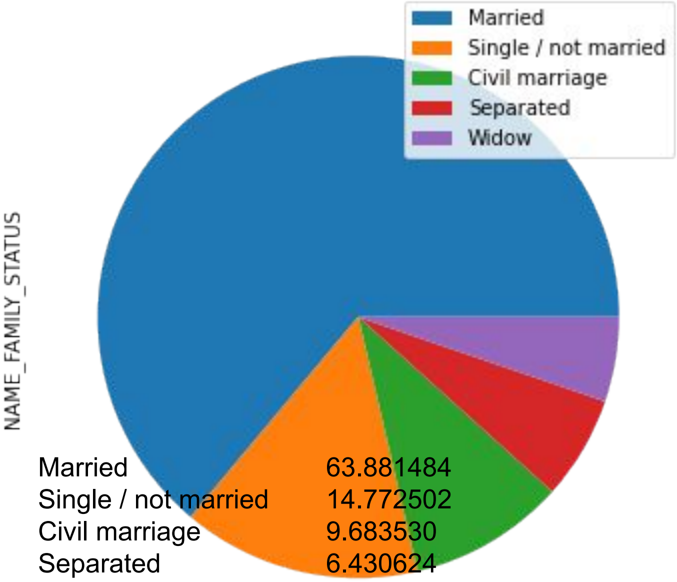
50% are working professional

Data Analysis for Application data



```
Secondary / secondary special    71.018087
Higher education                24.345692
Incomplete higher               3.341763
Lower secondary                 1.241078
Academic degree                 0.053380
Name: NAME_EDUCATION_TYPE
```

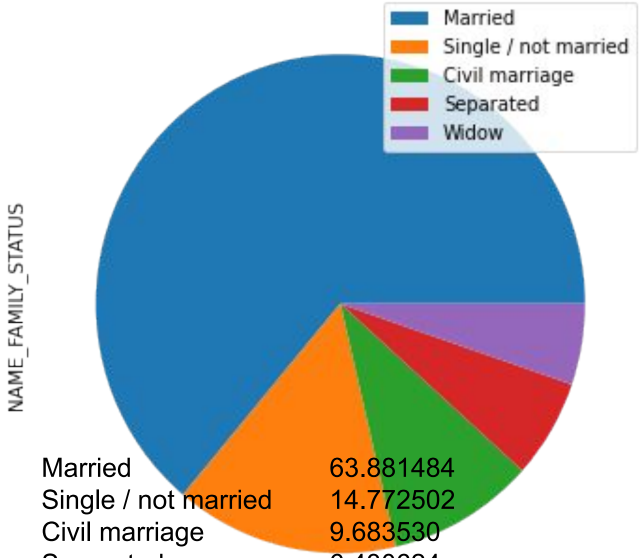
Secondary and secondary special applicants are more



```
Married                63.881484
Single / not married   14.772502
Civil marriage         9.683530
Separated              6.430624
Widow                  5.231860
Name: NAME_FAMILY_STATUS
```

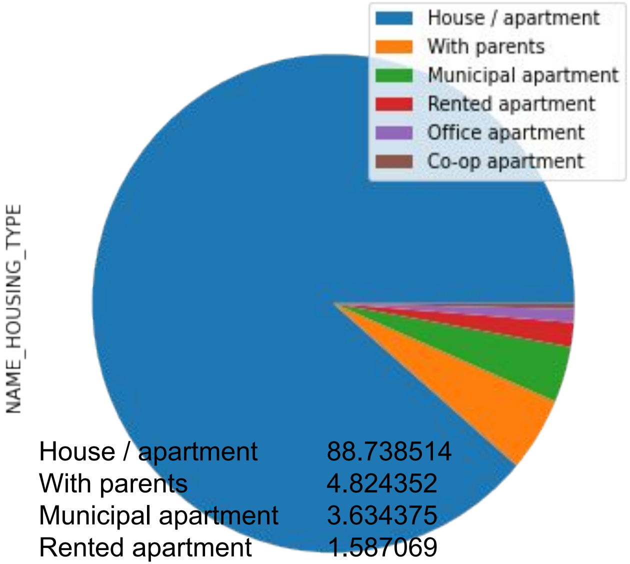
Married applicants are more

Data Analysis for Application data



```
Married                 63.881484
Single / not married    14.772502
Civil marriage           9.683530
Separated                6.430624
Widow                    5.231860
Name: NAME_FAMILY_STATUS
```

Married applicants are more



```
House / apartment      88.738514
With parents            4.824352
Municipal apartment     3.634375
Rented apartment        1.587069
Office apartment        0.851146
Co-op apartment         0.364544
Name: NAME_HOUSING_TYPE
```

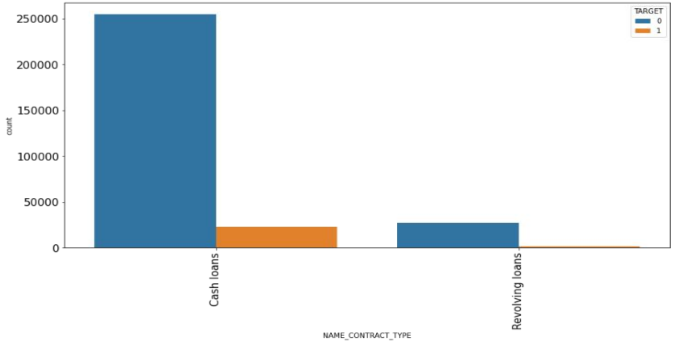
House or apartment holders are more

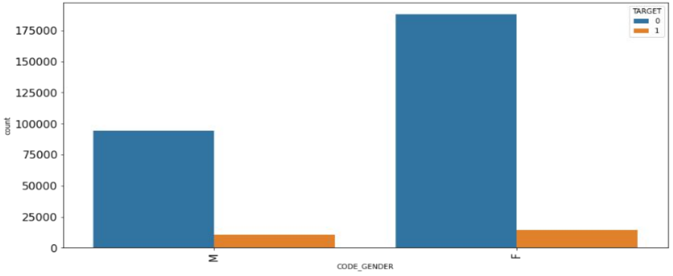
Data Cleaning for Application data

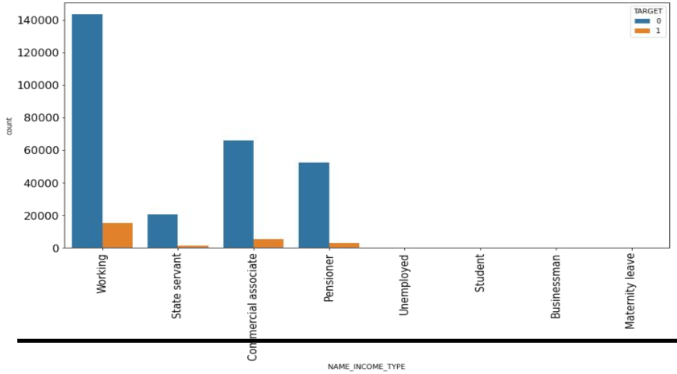
Data Inference

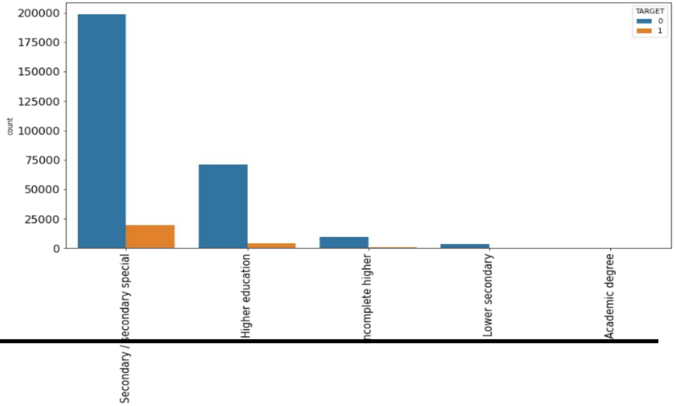
- Defaulters: Only 8% of applicants are defaulters.
- Gender Distribution: Female applicants outnumber male applicants.
- Loan Types: Cash loans significantly outnumber revolving loans.
- Occupation Types: Around 50% of loan applications come from working professionals. Other major groups include commercial associates and pensioners.
- Education Levels: Most applicants have completed higher education or above.
- Marital Status: A significant chunk of applicants are married.
- Housing Situation: Most applicants own their own house or apartment.

Univariate Analysis

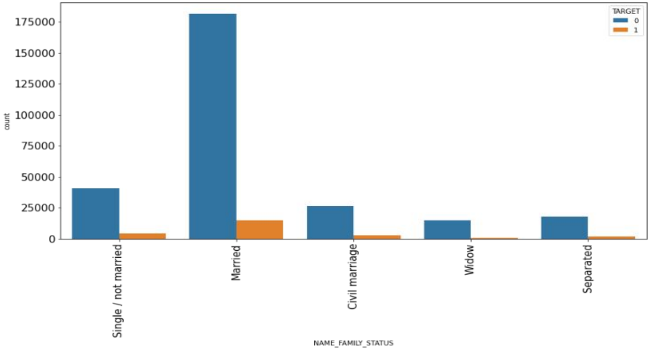


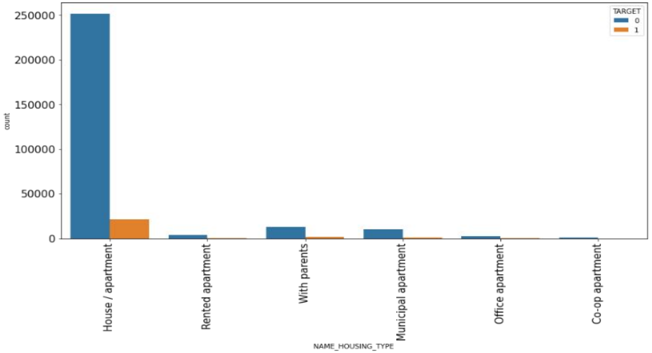


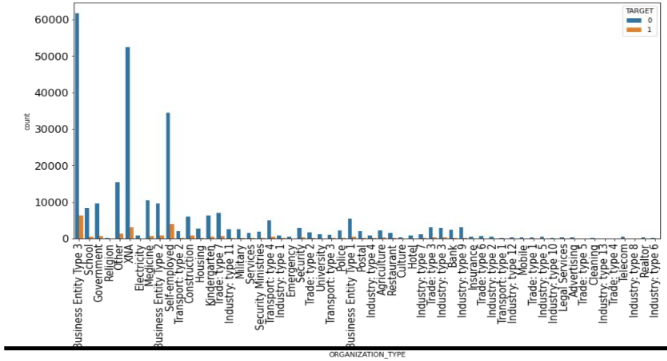


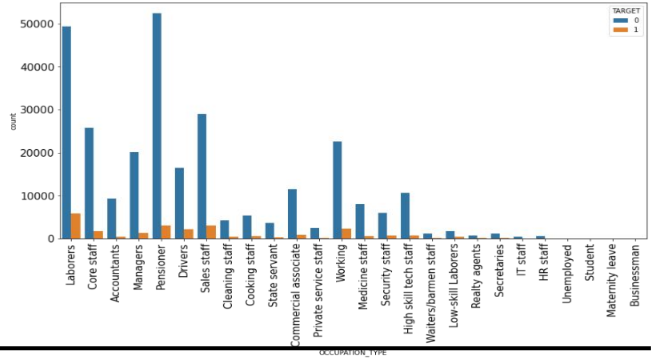


Univariate Analysis







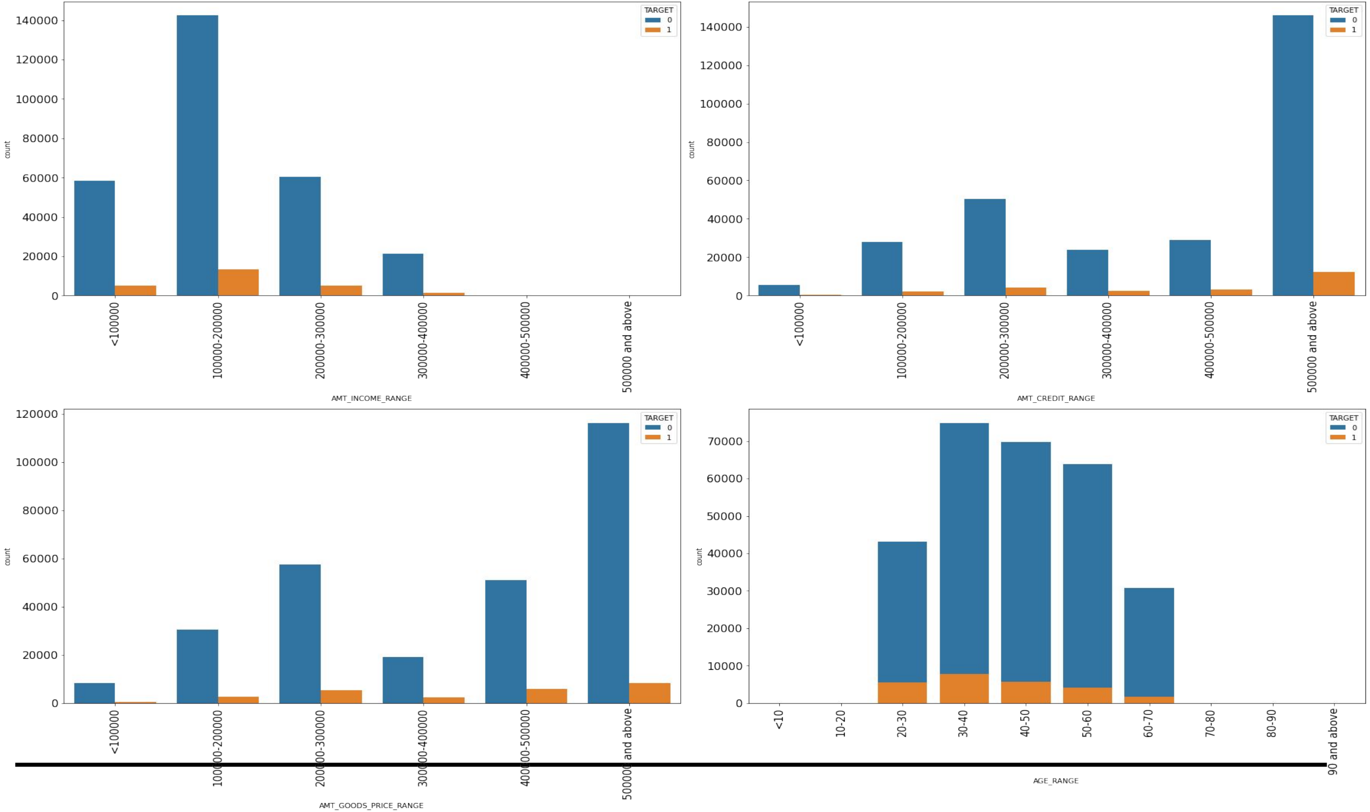


Univariate Analysis

Data Inference

- Cash Loans and Defaulters: We can infer that cash loans have more defaulters, as people tend to take more cash loans.
- Gender Comparison: The number of male and female defaulters is roughly the same. However, the defaulter ratio among females is smaller.
- Occupation and Defaulters: Working professionals constitute a larger proportion of defaulters, likely due to their higher representation among applicants.
- Education Levels and Defaulters: Secondary and higher secondary education levels are associated with a higher default rate.
- Marital Status & Defaulters: Both married applicants and defaulters are more prevalent.
- Housing Situation & Defaulters: House and apartment owners have a higher default rate.
- Business Entry Types and Defaulters: Business Entry Type 3 has more defaulters than any other occupation category.
- Labor Occupation and Defaulters: Laborers exhibit a comparatively higher default rate.

Univariate Analysis

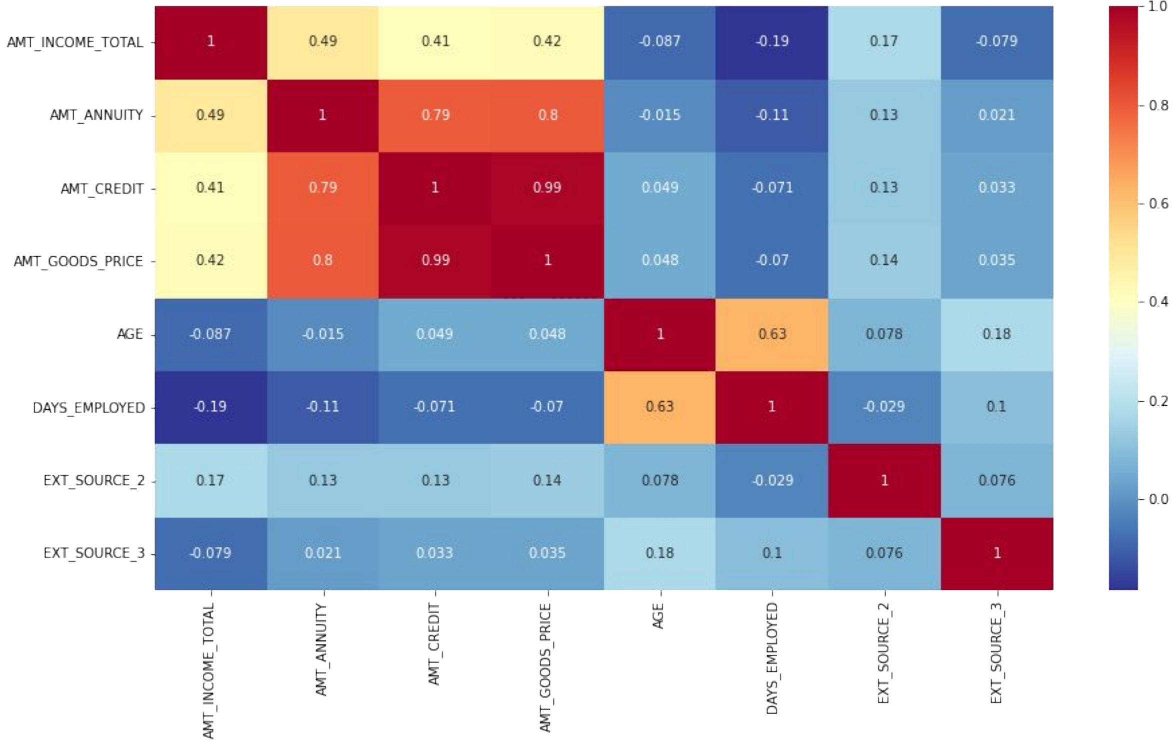


Univariate Analysis

Data Inference

- People with incomes between 10,000 and 20,000 are more likely to default.
- Credit Range and Defaulters:Individuals with credit ranges exceeding 50,000 are more prone to default.
- Goods Price Range and Defaulters:The average defaulter count across all goods price ranges is nearly the same.
- Age and Defaults:People aged 30-40 have a higher default rate, but their representation among applicants is also substantial.

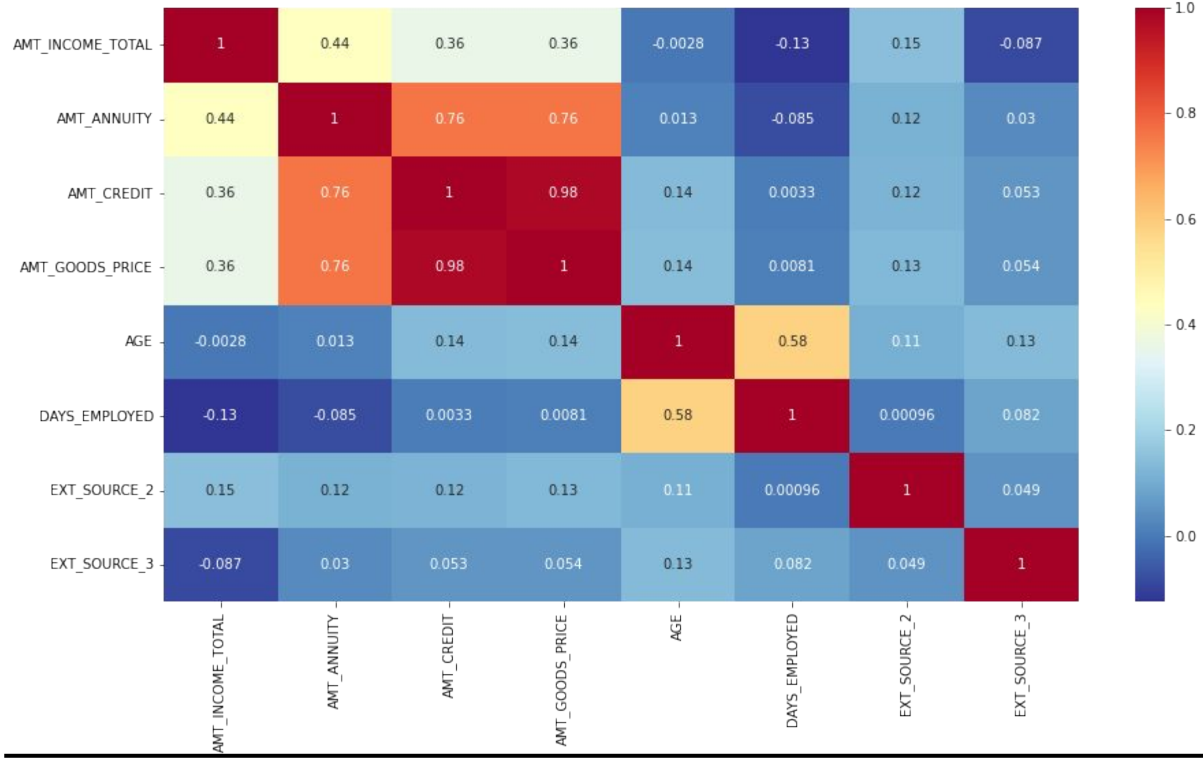
Multivariate Analysis



	AMT_INCOME_TOTAL	AMT_ANNUITY	AMT_CREDIT	AMT_GOODS_PRICE	AGE	DAYS_EMPLOYED	EXT_SOURCE_2	EXT_SOURCE_3
AMT_INCOME_TOTAL	1	0.49	0.41	0.42	-0.087	-0.19	0.17	-0.079
AMT_ANNUITY	0.49	1	0.79	0.8	-0.015	-0.11	0.13	0.021
AMT_CREDIT	0.41	0.79	1	0.99	0.049	-0.071	0.13	0.033
AMT_GOODS_PRICE	0.42	0.8	0.99	1	0.048	-0.07	0.14	0.035
AGE	-0.087	-0.015	0.049	0.048	1	0.63	0.078	0.18
DAYS_EMPLOYED	-0.19	-0.11	-0.071	-0.07	0.63	1	-0.029	0.1
EXT_SOURCE_2	0.17	0.13	0.13	0.14	0.078	-0.029	1	0.076
EXT_SOURCE_3	-0.079	0.021	0.033	0.035	0.18	0.1	0.076	1

Non Defaulters:
We can say there is high correlation between Amount credited with loan annuity, good price. Age and date of employment have some relations

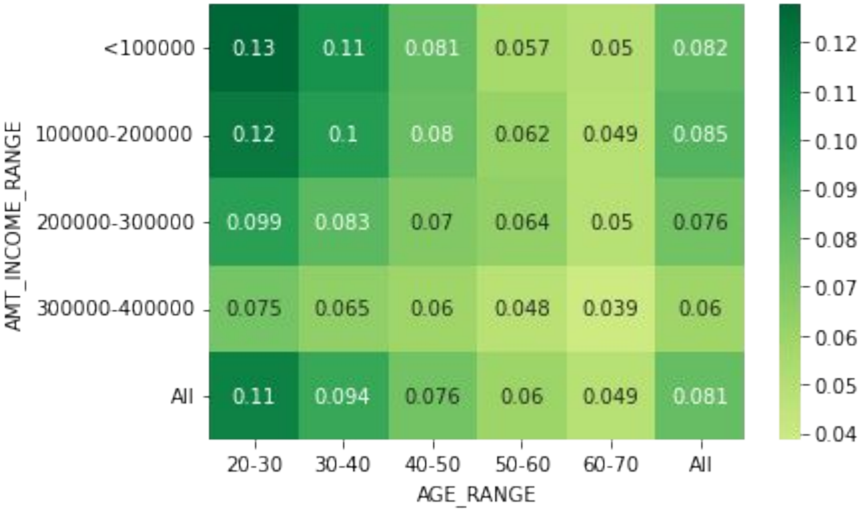
Multivariate Analysis



	AMT_INCOME_TOTAL	AMT_ANNUITY	AMT_CREDIT	AMT_GOODS_PRICE	AGE	DAYS_EMPLOYED	EXT_SOURCE_2	EXT_SOURCE_3
AMT_INCOME_TOTAL	1	0.44	0.36	0.36	-0.0028	-0.13	0.15	-0.087
AMT_ANNUITY	0.44	1	0.76	0.76	0.013	-0.085	0.12	0.03
AMT_CREDIT	0.36	0.76	1	0.98	0.14	0.0033	0.12	0.053
AMT_GOODS_PRICE	0.36	0.76	0.98	1	0.14	0.0081	0.13	0.054
AGE	-0.0028	0.013	0.14	0.14	1	0.58	0.11	0.13
DAYS_EMPLOYED	-0.13	-0.085	0.0033	0.0081	0.58	1	0.00096	0.082
EXT_SOURCE_2	0.15	0.12	0.12	0.13	0.11	0.00096	1	0.049
EXT_SOURCE_3	-0.087	0.03	0.053	0.054	0.13	0.082	0.049	1

<u>Defaulters:</u>
The credit amount is related to annuity and goods price. Repayers have a higher correlation with days employed (0.63), whereas here it is 0.58. There is a drop in correlation between total income and the amount of credit in defaulters compared to repayers.

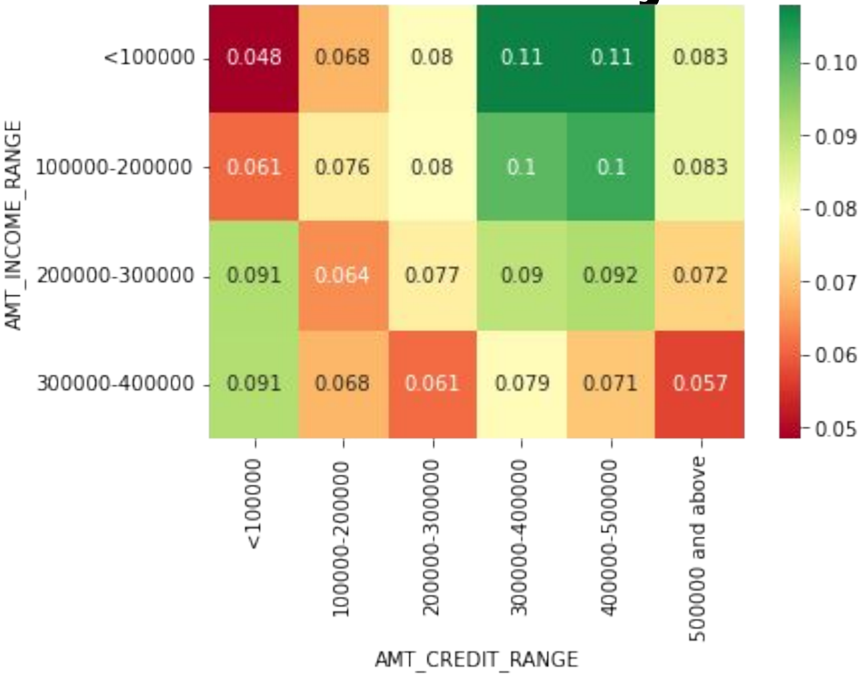
Multivariate Analysis



AMT_INCOME_RANGE	20-30	30-40	40-50	50-60	60-70	All
<100000	0.13	0.11	0.081	0.057	0.05	0.082
100000-200000	0.12	0.1	0.08	0.062	0.049	0.085
200000-300000	0.099	0.083	0.07	0.064	0.05	0.076
300000-400000	0.075	0.065	0.06	0.048	0.039	0.06
All	0.11	0.094	0.076	0.06	0.049	0.081

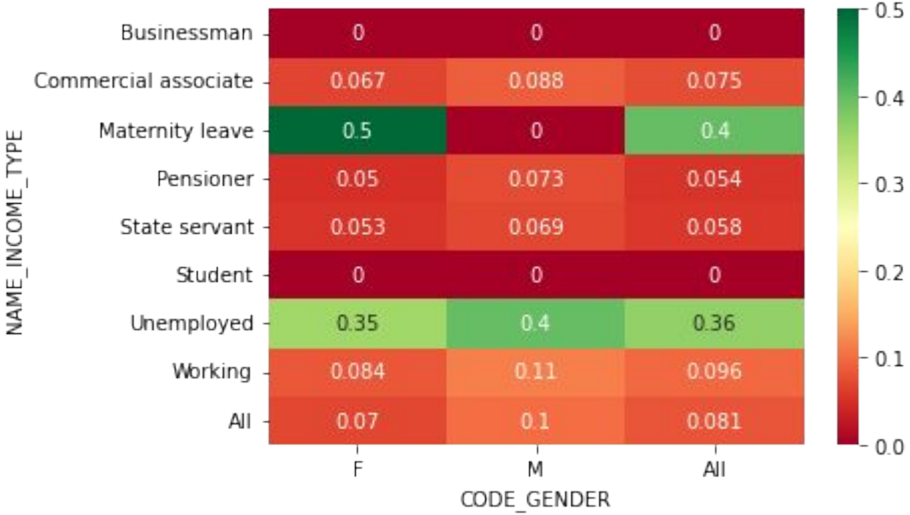
People with income less than 200000 and age 20-30 are more of defaulters

Multivariate Analysis



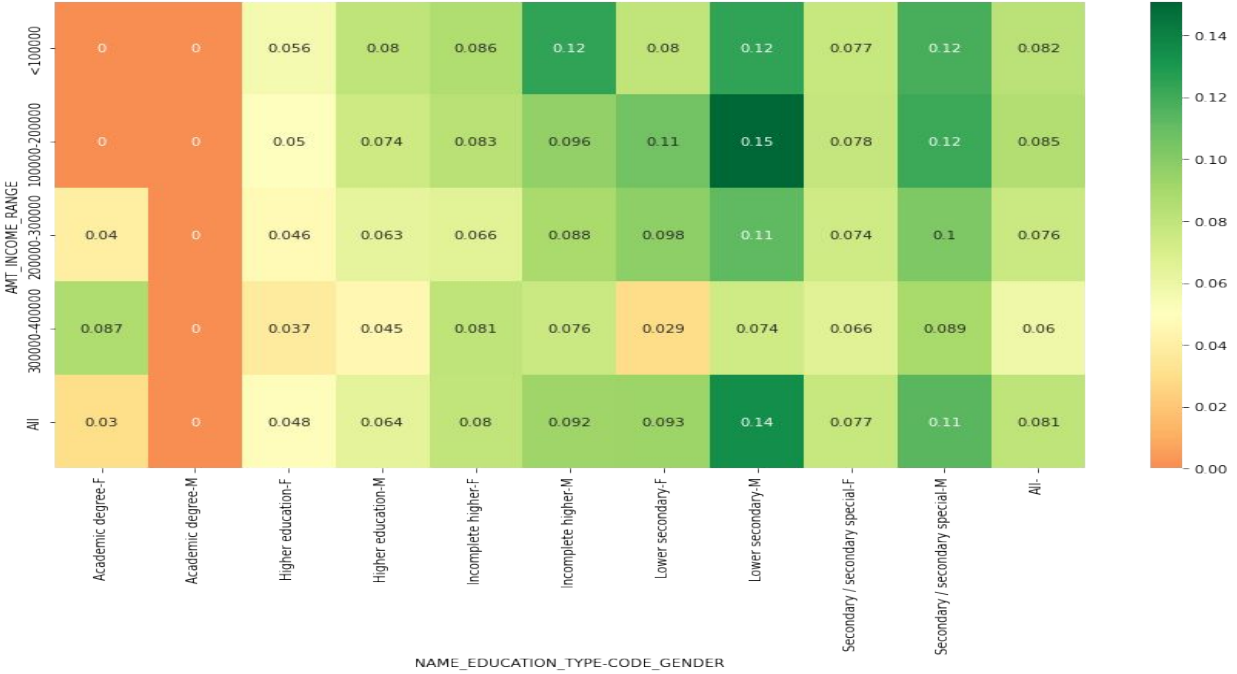
If credit provided is between 3 to 5 lakh and income is less than 2 lac then defaulters ration is high

Multivariate Analysis



Unemployed male are more defaulters than female

Multivariate Analysis



Income range Vs Education

Multivariate Analysis

Data Inference

- The defaulter ratio is higher among unemployed individuals on maternity leave compared to others.
- Unemployed males have a higher default rate than females.
- Males with cash loans are more likely to default.
- Unemployed males with secondary or secondary special education are more prone to defaulting.
- Females who are unemployed and have secondary or secondary special education are also more likely to default.
- Males with any family status fall into the defaulter category.
- Individuals living in retired apartments or with parents have a higher tendency to become defaulters.

Data Cleaning for Prev App data

Handling Missing Value

- Ignore columns having missing value more than 40%

- Use mean value for AMT_ANNUITY, CNT_PAYMENT and remove missing rows having null value for all AMT_GOODS_PRICE, AMT_ANNUITY, CNT_PAYMENT columns.

Data Cleaning for Prev App data

Remove non required columns

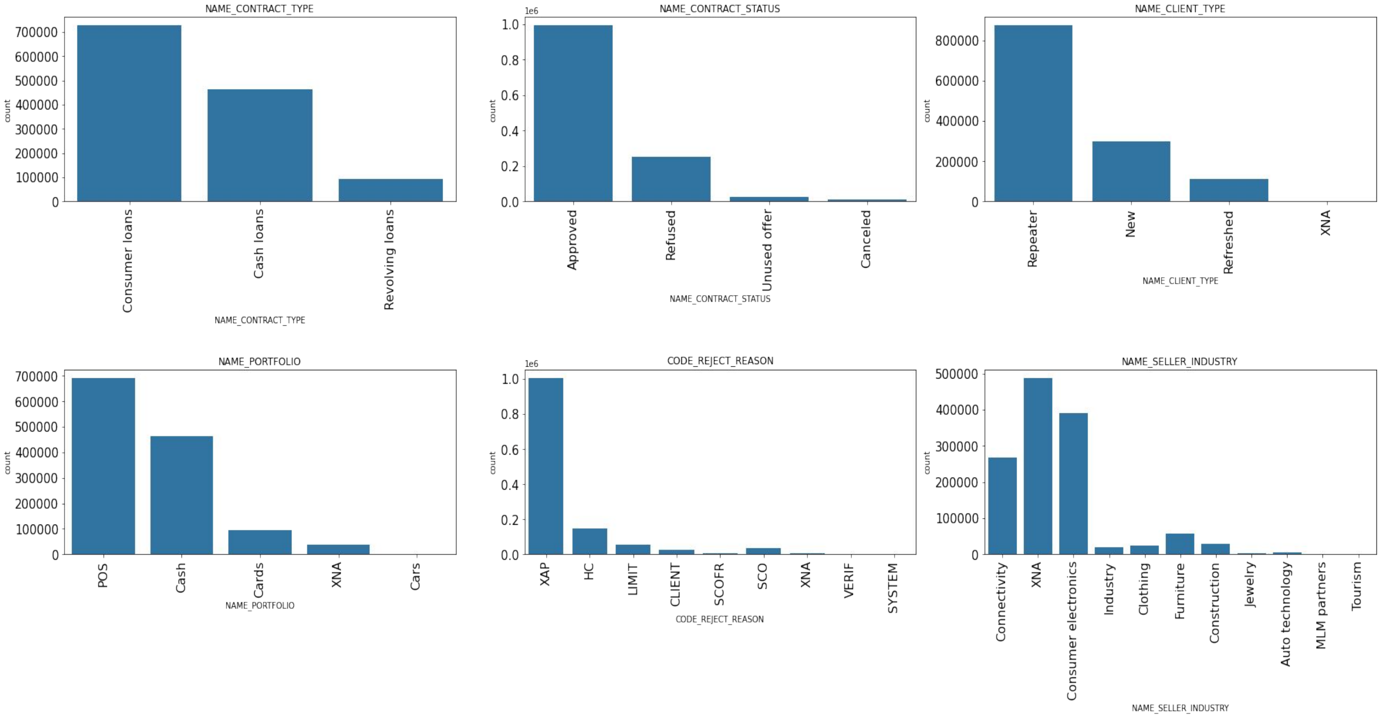
- Remove columns like FLAG_LAST_APPL_PER_CONTRACT, NFLAG_LAST_APPL_IN_DAY , WEEKDAY_APPR_PROCESS_START , HOUR_APPR_PROCESS_START

Data Cleaning for Prev App data

Handling Data Errors

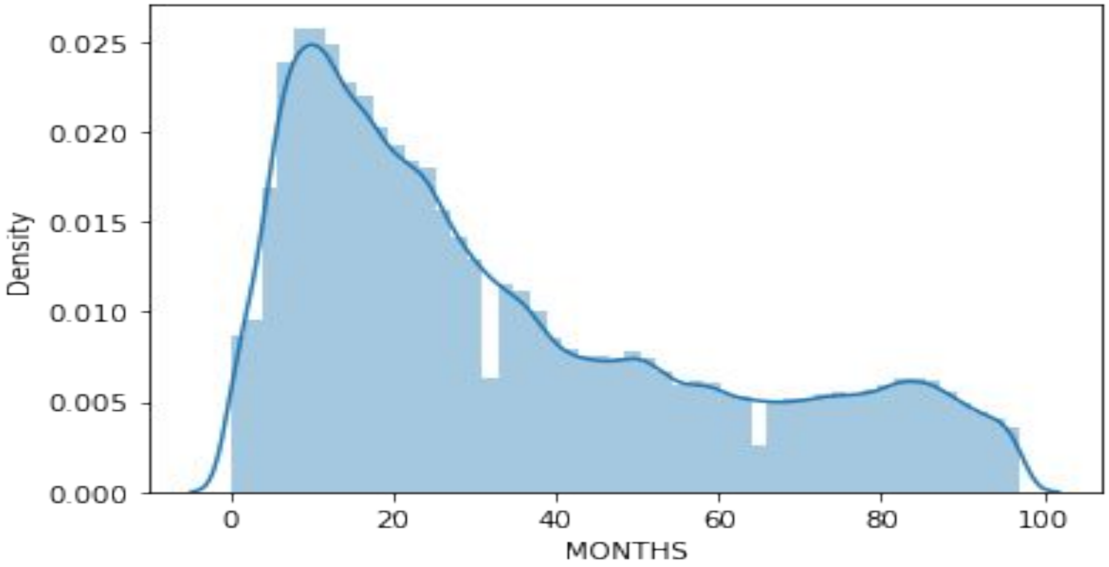
- Convert Data of DAYS_DECISION to absolute value and convert to MONTHS
- Fix outliers for AMT_APPLICATION, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, CNT_PAYMENT, DAYS_DECISION

Data Analysis for Prev App data



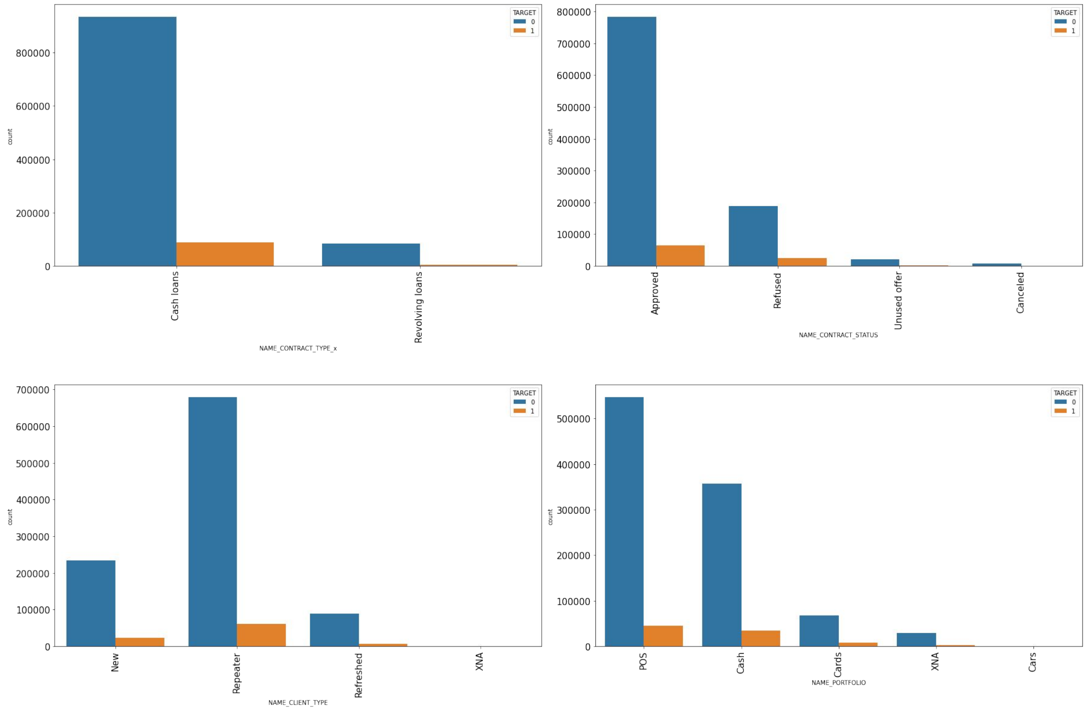
* Consumer loans are more in applications (55%)
* Almost 77% loans are approved
* Repeater % is very high
* POS loans are 53% followed by cash 33$
* XAP reason is almost 78%

Data Analysis for Prev App data

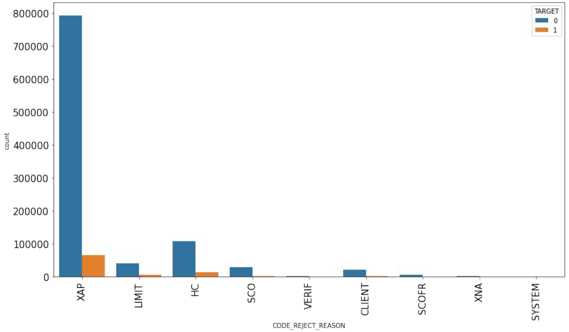


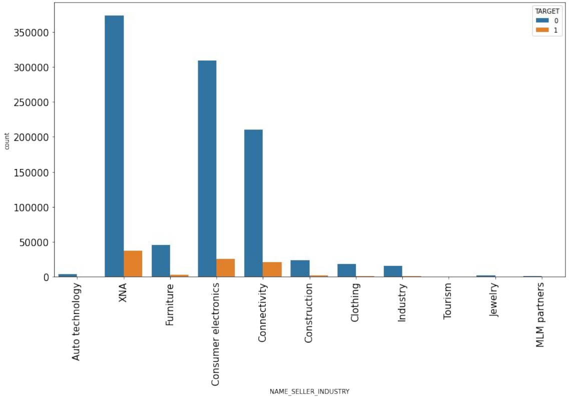
Decision Duration

Analysis of combined data



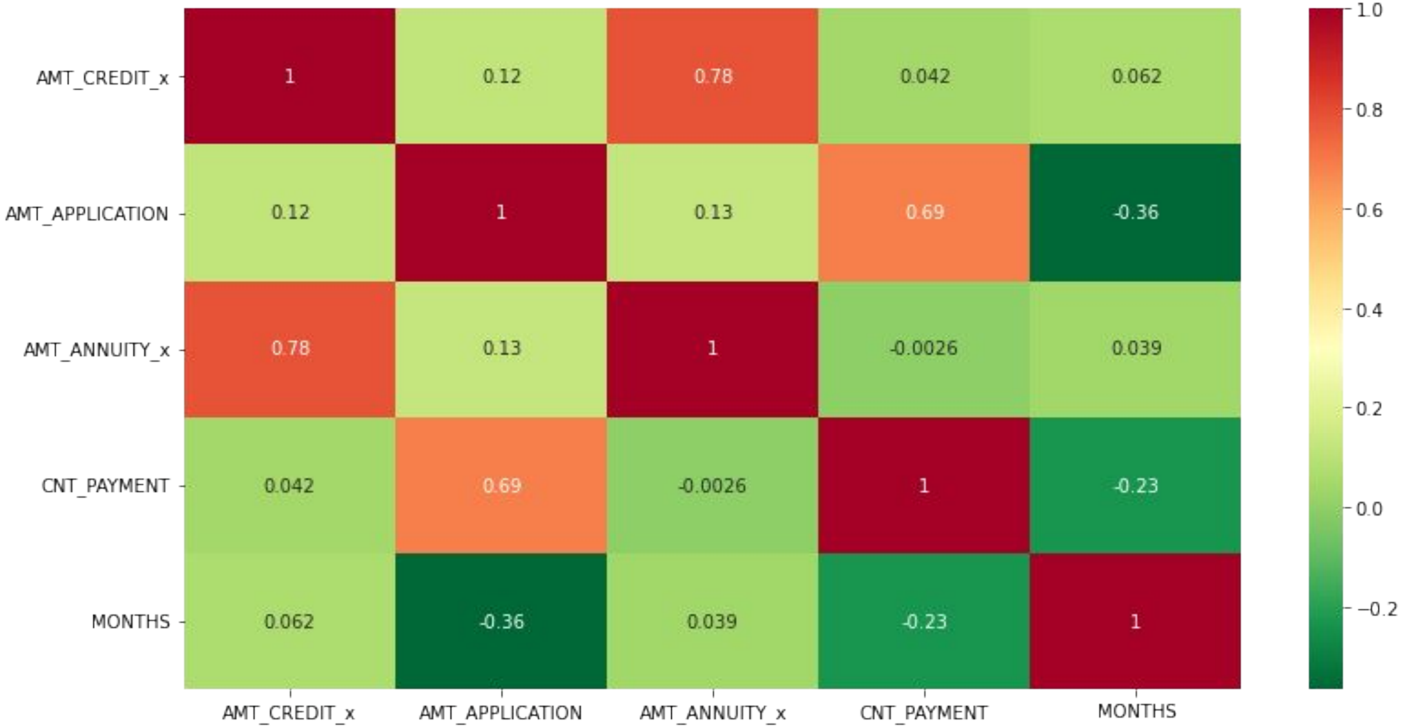
Analysis of combined data





* Repeater loans are too high and its approval percentage is also more
* Consumer electronics and connectivity loan percentage and also approval is high

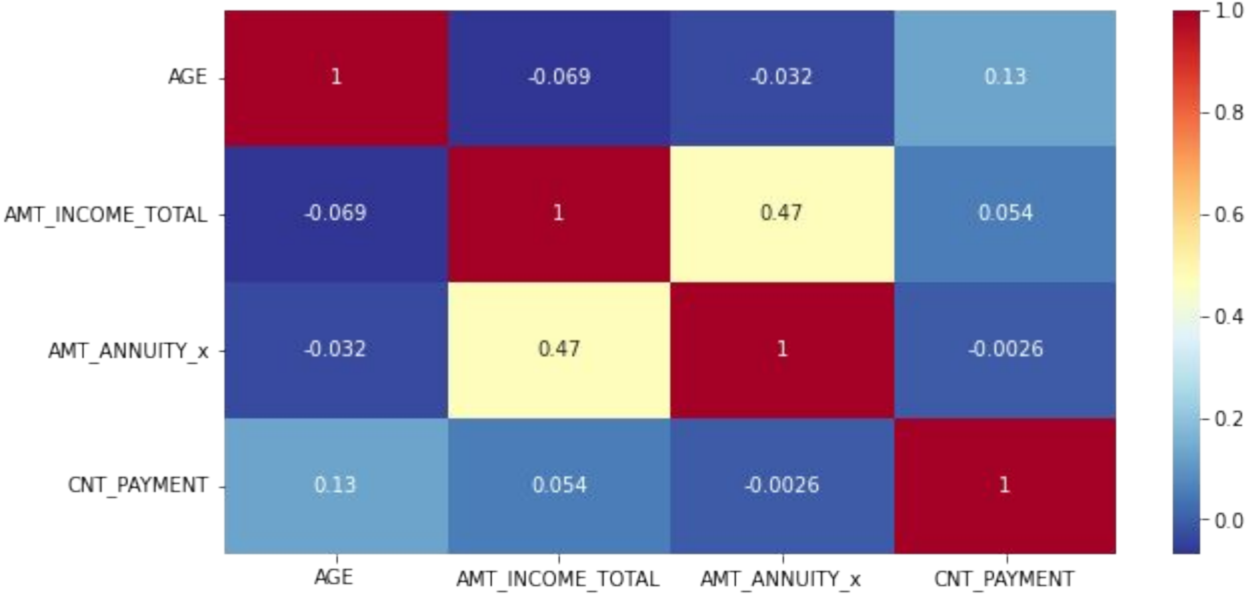
Analysis of combined data



	AMT_CREDIT_x	AMT_APPLICATION	AMT_ANNUITY_x	CNT_PAYMENT	MONTHS
AMT_CREDIT_x	1	0.12	0.78	0.042	0.062
AMT_APPLICATION	0.12	1	0.13	0.69	-0.36
AMT_ANNUITY_x	0.78	0.13	1	-0.0026	0.039
CNT_PAYMENT	0.042	0.69	-0.0026	1	-0.23
MONTHS	0.062	-0.36	0.039	-0.23	1

AMT_ANNUITY and AMT_CREDIT are correlated
CNT_PAYMENT and AMT_APPLICATION are correlated

Analysis of combined data



Correlations for Defaulters

Conclusion

1. Education Level: Secondary or secondary special education graduates struggle with loan payments.
2. Contract Type and Gender:
   a. Female borrowers with consumer loans have a higher default rate.
   b. Female gender is less likely to face payment difficulties compared to males.
   c. Recommendation: Approve more loans for females.
3. Credit Amount: Lower credit amount borrowers and very high credit amount borrowers are at higher risk of defaulting.
4. Marital Status:
   a. Married individuals struggle more with loan payments than single or separated people.
   b. More approved applicants are married.
   c. Suggestion: Consider loan approval for single or divorced individuals.
5. Applicant Type: Repeater applicants have both a high chance of non-defaulting and a high chance of defaulting compared to new applicants.

Thank You!!!