EDA Assignment

Bank Loan Risk Analysis

by
Bipin Joseph Odattil

Available Data Sheets

1. application_data.csv

Contains information about clients at the time of application

~307K rows, 122 columns

2. previous_application.csv

Contains information about the client's previous loan applications

~1670K rows, 37 columns

3. columns_description.csv

Contains the description for columns in above two sheets.

Note: We will not be loading this sheet as a python dataframe as this is a reference data sheet

application_data.csv

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	F
100002	1	Cash loans	M	N	
100003	0	Cash loans	F	N	
100004	0	Revolving loans	M	Y	
100006	0	Cash loans	F	N	
100007	0	Cash loans	M	N	
100008	0	Cash loans	M	N	
100009	0	Cash loans	F	Y	
100010	0	Cash loans	M	Y	
100011	0	Cash loans	F	N	
100012	0	Revolving loans	M	N	

previous_application.csv

SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICAT
2030495	271877	Consumer loans	1730.430	171
2802425	108129	Cash loans	25188.615	6075
2523466	122040	Cash loans	15060.735	1125
2819243	176158	Cash loans	47041.335	4500
1784265	202054	Cash loans	31924.395	3375
1383531	199383	Cash loans	23703.930	3150
2315218	175704	Cash loans	NaN	
1656711	296299	Cash loans	NaN	
2367563	342292	Cash loans	NaN	

Data Understanding- Columns

application_data.csv

Data Types of Columns

- There are 122 columns and its not easy to go through the data types for each of the columns.

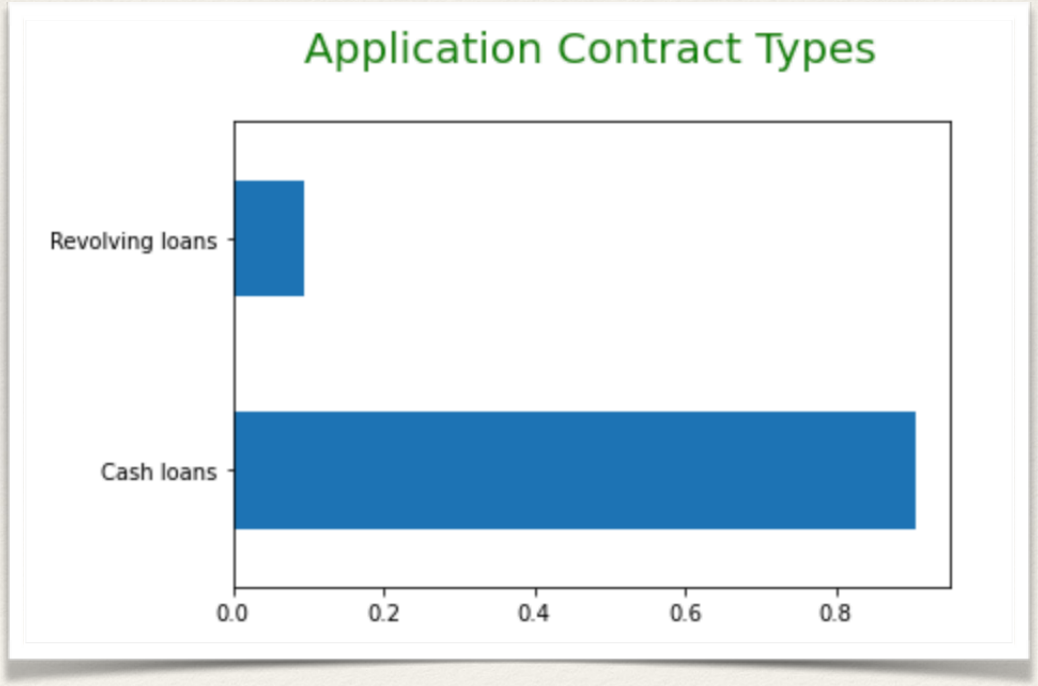
So, lets try and do this one at a time on what we believe are the important columns

- Key points:
  - SK_ID_CURR is the id of the loan application
  - TARGET (1/0)
    - 1 means the applicant had difficulty paying loan installment
    - 0 means the applicant had no difficulty paying the installments

```
SK_ID_CURR                      int64
TARGET                          int64
NAME_CONTRACT_TYPE             object
CODE_GENDER                    object
FLAG_OWN_CAR                   object
                                 ...
AMT_REQ_CREDIT_BUREAU_DAY     float64
AMT_REQ_CREDIT_BUREAU_WEEK    float64
AMT_REQ_CREDIT_BUREAU_MON     float64
AMT_REQ_CREDIT_BUREAU_QRT     float64
AMT_REQ_CREDIT_BUREAU_YEAR    float64
Length: 122, dtype: object
```

Contract Type of the Loan

- There are two types of loan applications
- Out of the 307K applications, more than 90% of the loan applications are Cash Loans



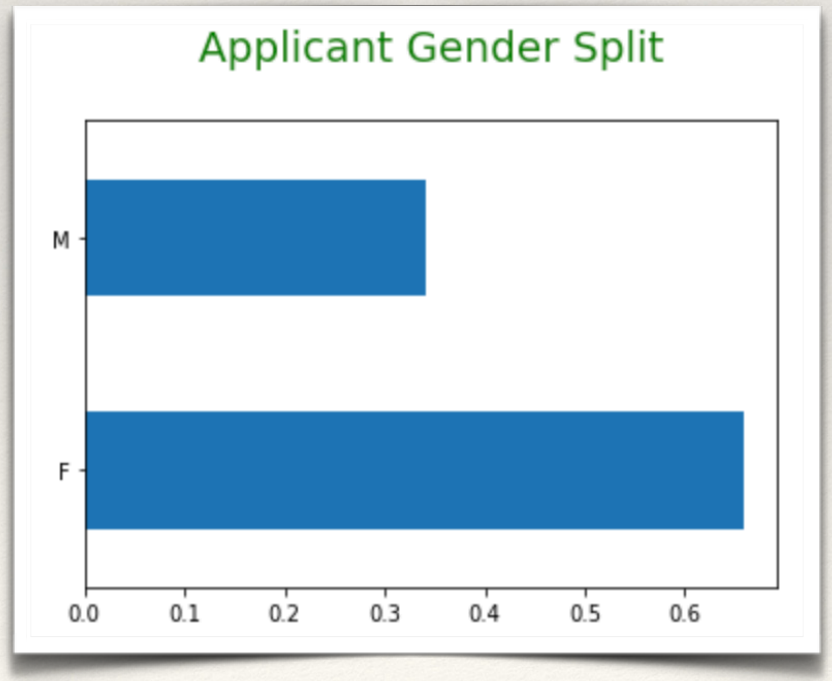
Gender of Loan Applicants

- There are 4 unknown values(XNA) in the gender column
- Since these are only 4 records among 307K applicants, we will remove these records and check the proportion

```
F          202448
M          105059
XNA             4
Name: CODE_GENDER,
```

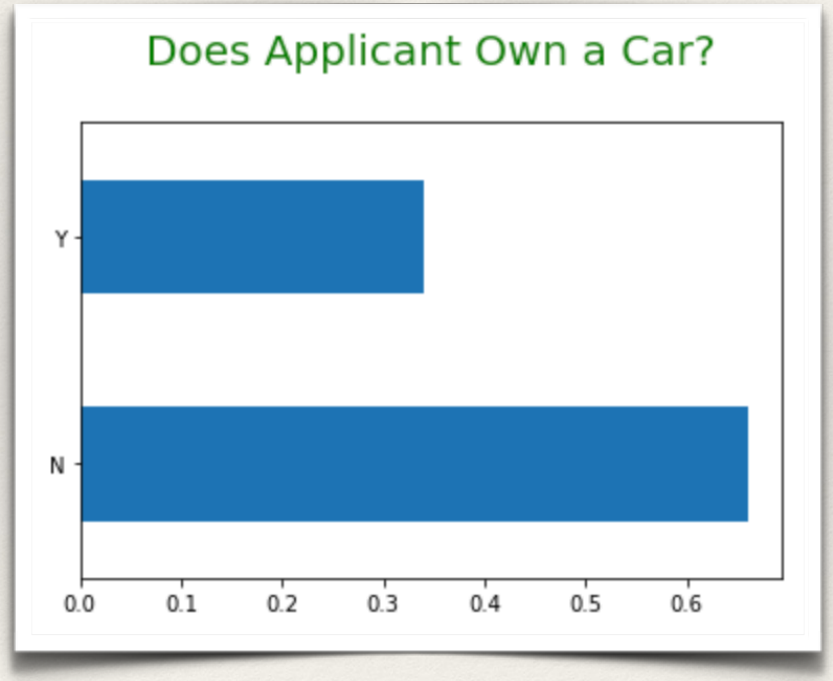
Proportion of applicant's gender after removing unknown records

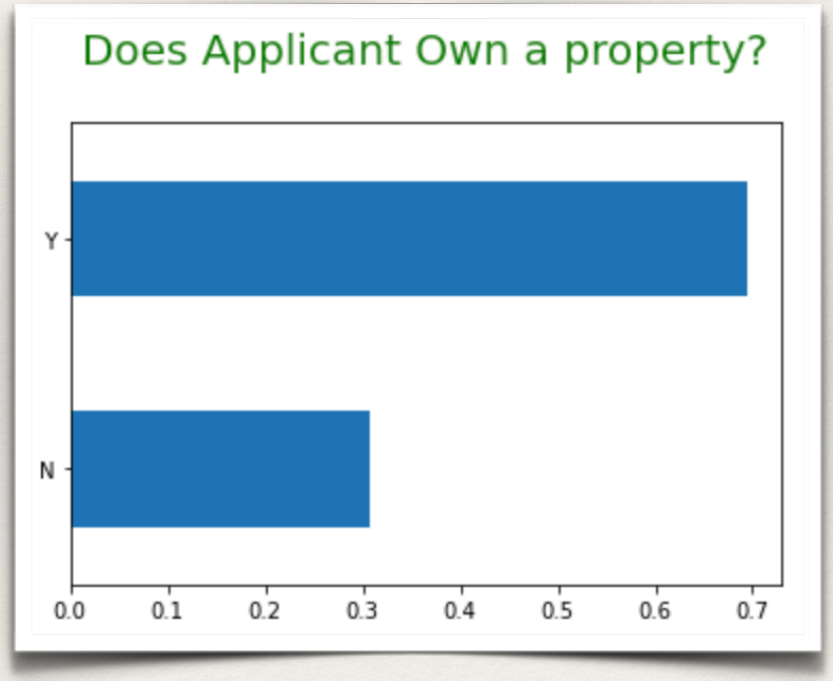
Almost 70% of all applicants are Female



Ownership of car or property

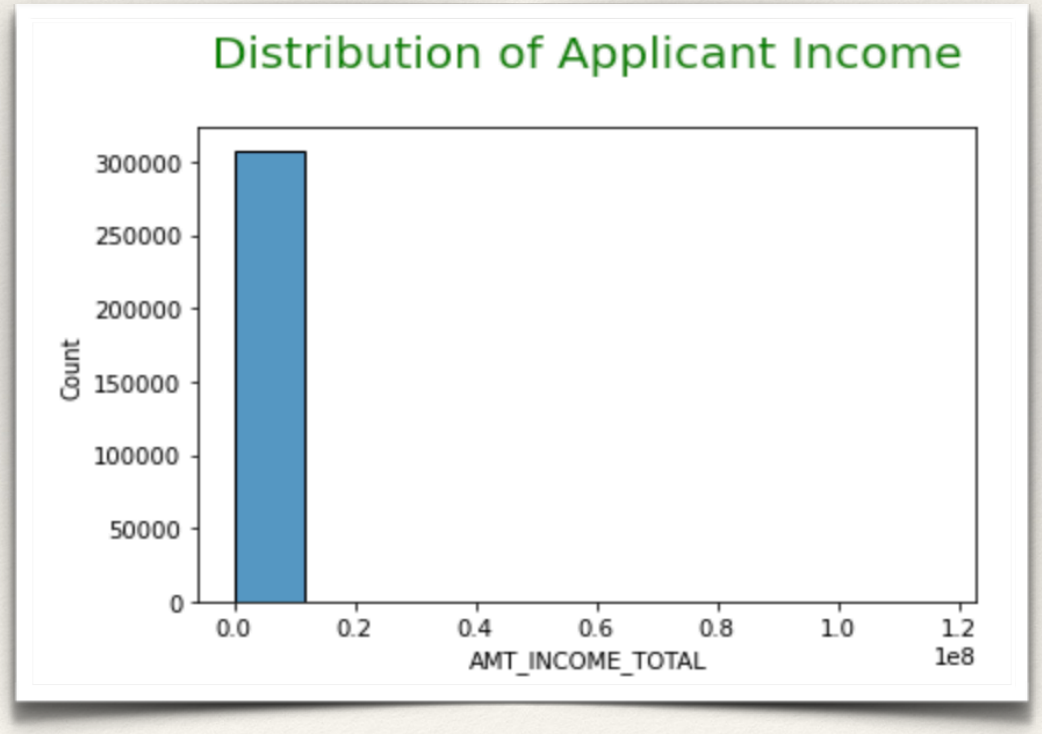
Most(70%) applicants own a property, but not a car





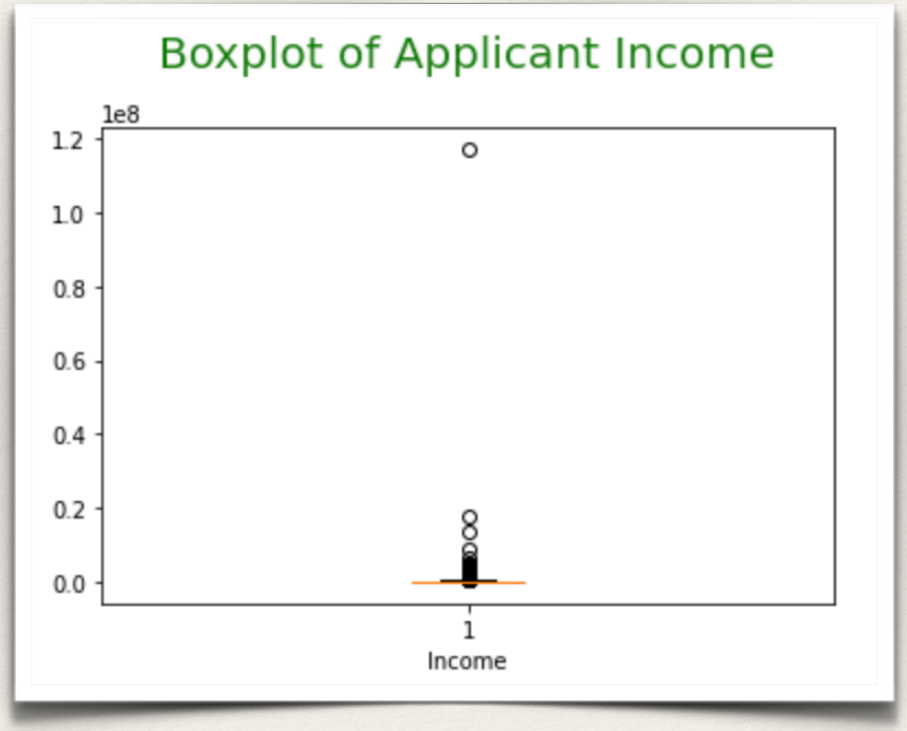
Applicant Income

- The distribution is heavily imbalanced. Most of the values have come in a single bin
- There might be few high valued outliers that need to be addressed



Applicant Income Boxplot

We can clearly see the outlier value in the box plot.
The value is much higher than all the other values.



Let us have a detailed look at the column Total Income and try to address this

Detailed look at Income

- The max value is around 1000 times that of the 75th percentile value
- Mean is around 160,000
- Median is around 140,000

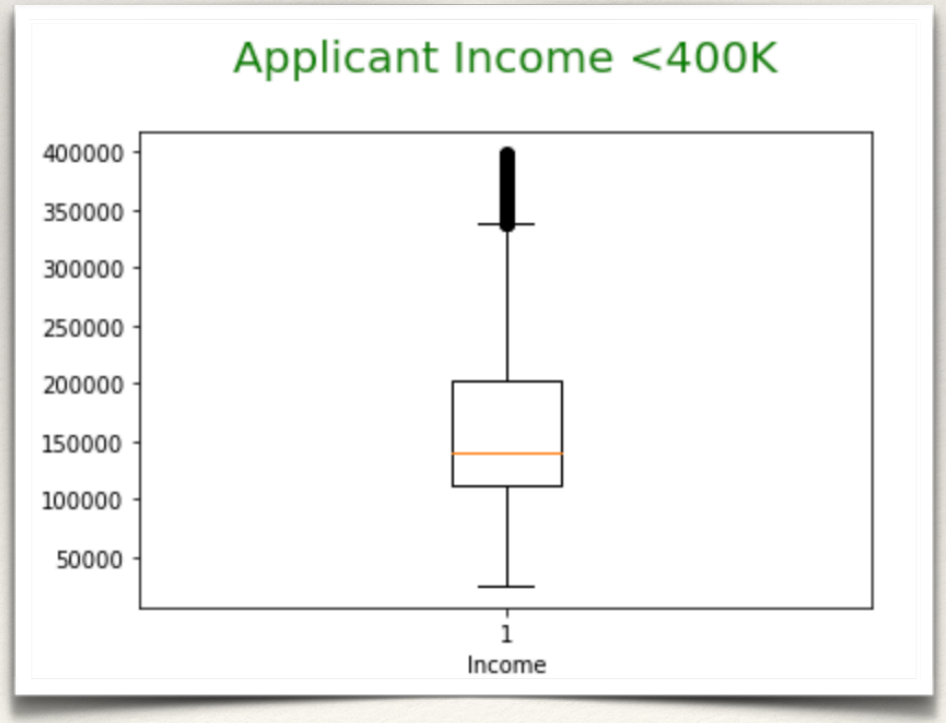
```
count      3.075070e+05
mean       1.687977e+05
std        2.371246e+05
min        2.565000e+04
25%        1.125000e+05
50%        1.471500e+05
75%        2.025000e+05
max        1.170000e+08
Name: AMT_INCOME_TOTAL,
```

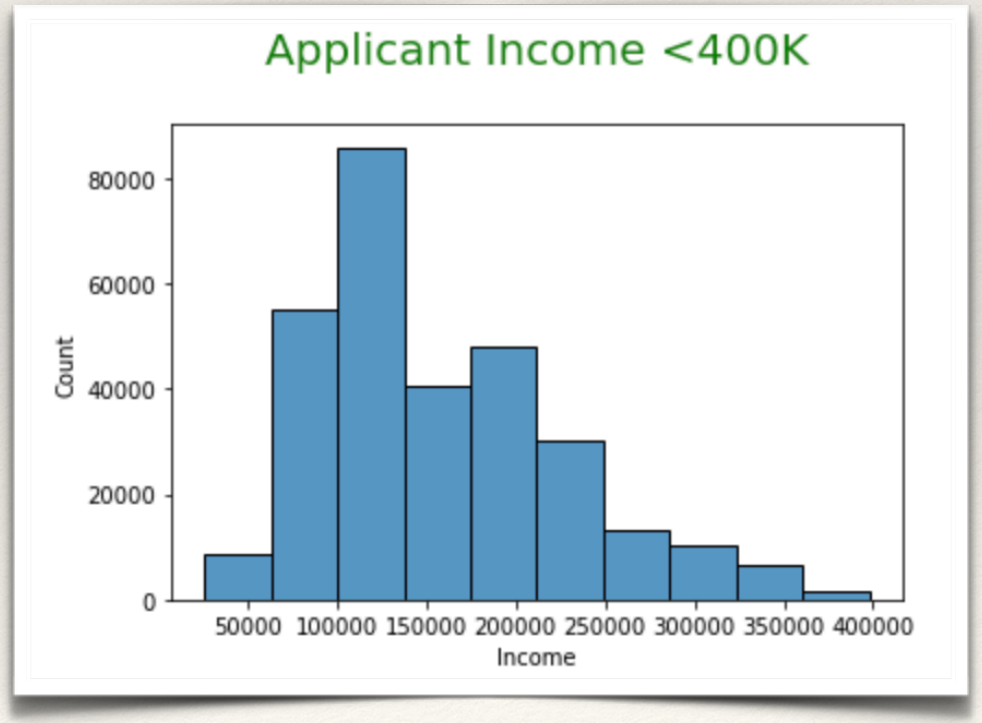
- These are the 10 applicants with most income
- Here itself you can see the difference in amount

	SK_ID_CURR	TARGET	AMT_INCOME_TOTAL
12840	114967	1	117000000.0
203693	336147	0	18000090.0
246858	385674	0	13500000.0
77768	190160	0	9000000.0
131127	252084	0	6750000.0
187833	317748	0	4500000.0
103006	219563	0	4500000.0
287463	432980	0	4500000.0
204564	337151	0	4500000.0
181698	310601	0	3950059.5

Handling outliers in Income

- We are taking only applicants with income less than 400K for this analysis
- 400K is higher than 97th percentile of the data.
- Thus, we will be getting data set which does not have any outliers
- Most of the applicants lie in the 100K to 150K Income bin

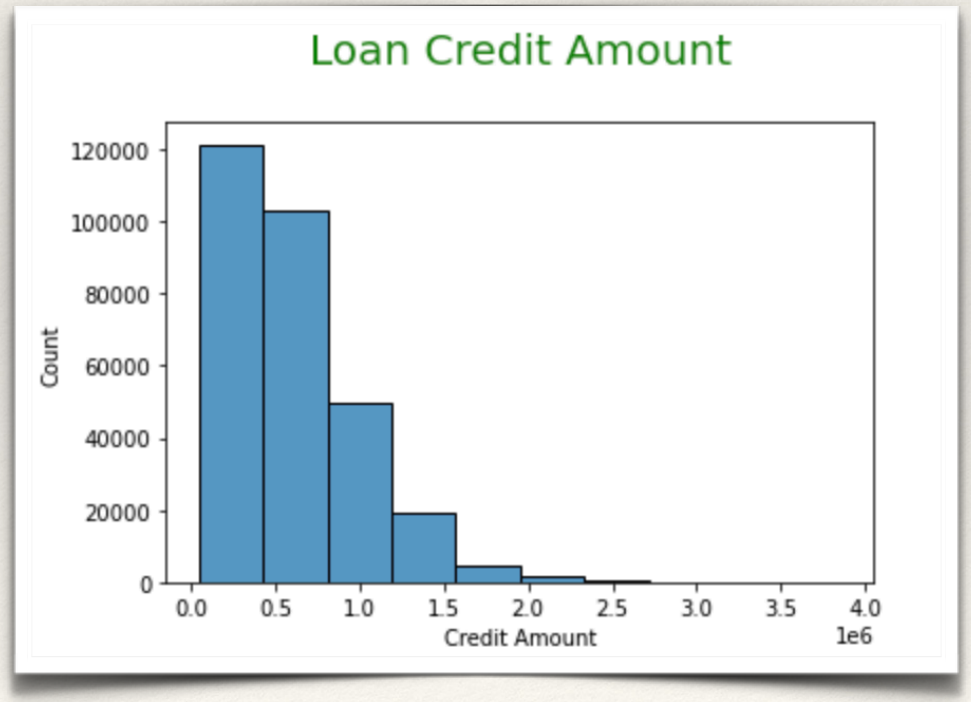




Loan Credit Amount

- Most loan applications are for less than 1Million
- Largest application is for around 3.8 Million

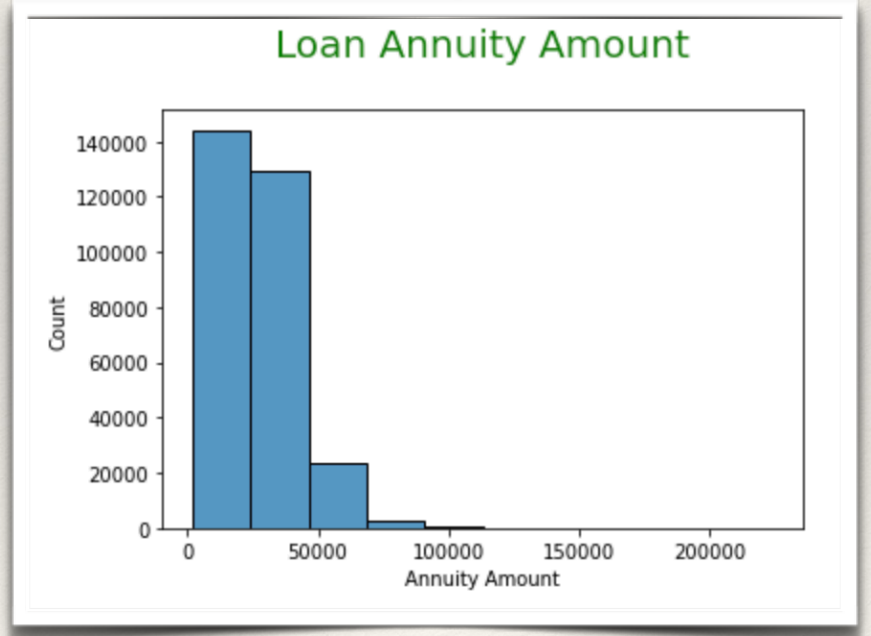
```
count      2.994440e+05
mean       5.870204e+05
std        3.905561e+05
min        4.500000e+04
25%        2.700000e+05
50%        5.084955e+05
75%        8.086500e+05
max        3.860019e+06
Name: AMT_CREDIT, dtyp
```



Loan Annuity Amount

- Most annuities are less than 50,000
- Largest annuity is 225,000
- Trend is similar to credit amount
- It also has 12 null value records which were removed

```
count     299432.000000
mean       26565.519584
std        13714.129700
min         1615.500000
25%        16371.000000
50%        24588.000000
75%        33814.125000
max       225000.000000
Name: AMT_ANNUITY, dtyp
```



Age of Applicant

- Applicant's age is given in days
- But all are negative values
- May be they were calculated in reference to day of application

```
count     307507.000000
mean      -16037.027271
std         4363.982424
min       -25229.000000
25%       -19682.000000
50%       -15750.000000
75%       -12413.000000
max        -7489.000000
Name: DAYS_BIRTH, dtype:
```

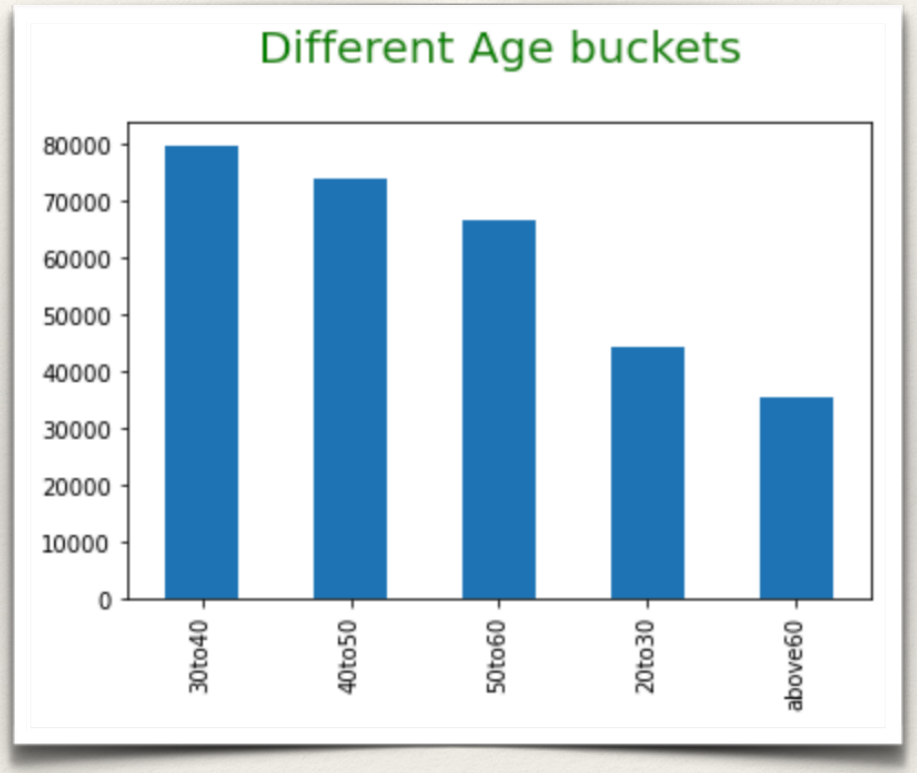
After changing these to positive values and converting to years for easier calculation

```
count    299444.000000
mean         43.955718
std          12.009309
min          20.520000
25%          33.950000
50%          43.170000
75%          54.020000
max          69.040000
Name: AGE, dtype: float64
```

- Minimum age is 20 years
- Maximum is 69 years
- Mean age is 43 years

Age into buckets

- Most applicants are in 30 to 40 age bin
- Followed by 40 to 50 and 50 to 60
- Above 60 years age are the least



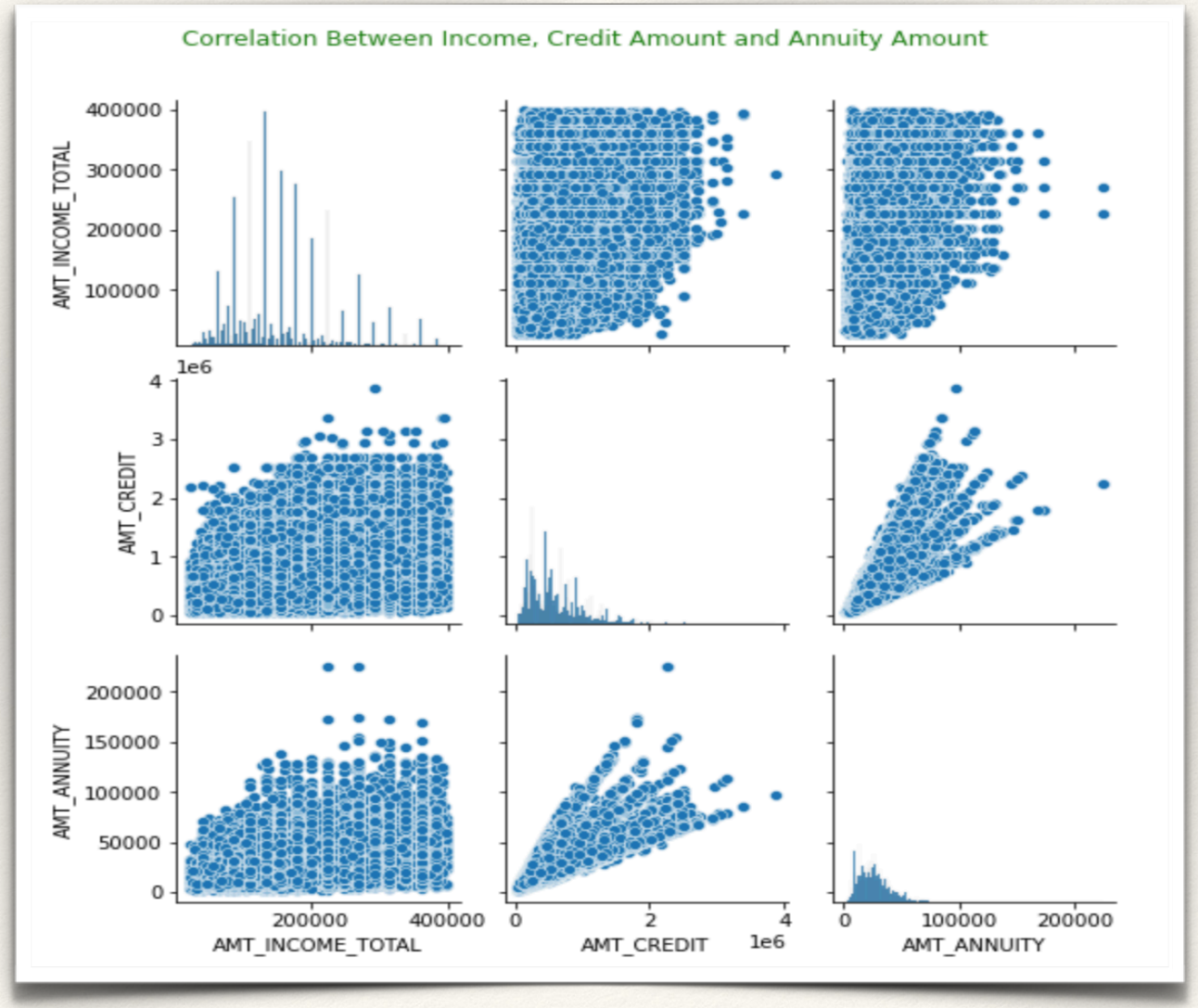
Applicants with Annuity > Income

There are 33 applicants whose total income is less than the annuity of the loan
High risk category

	SK_ID_CURR	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
678	100784	54000.0	1885500.0	74191.5
20727	124157	25650.0	2173500.0	48118.5
25802	129999	46044.0	1024740.0	52452.0
35791	141454	45000.0	2215224.0	58567.5
49185	156942	33750.0	835380.0	36927.0
58379	167672	54000.0	1724688.0	60079.5
64338	174612	54000.0	1574532.0	64260.0
68803	179800	31500.0	755190.0	32125.5
95599	210988	76500.0	2013840.0	82152.0
112431	230424	40500.0	472500.0	44635.5
115872	234376	67500.0	765000.0	72634.5
120680	239922	56250.0	790830.0	57676.5
122480	242003	54000.0	497520.0	59175.0

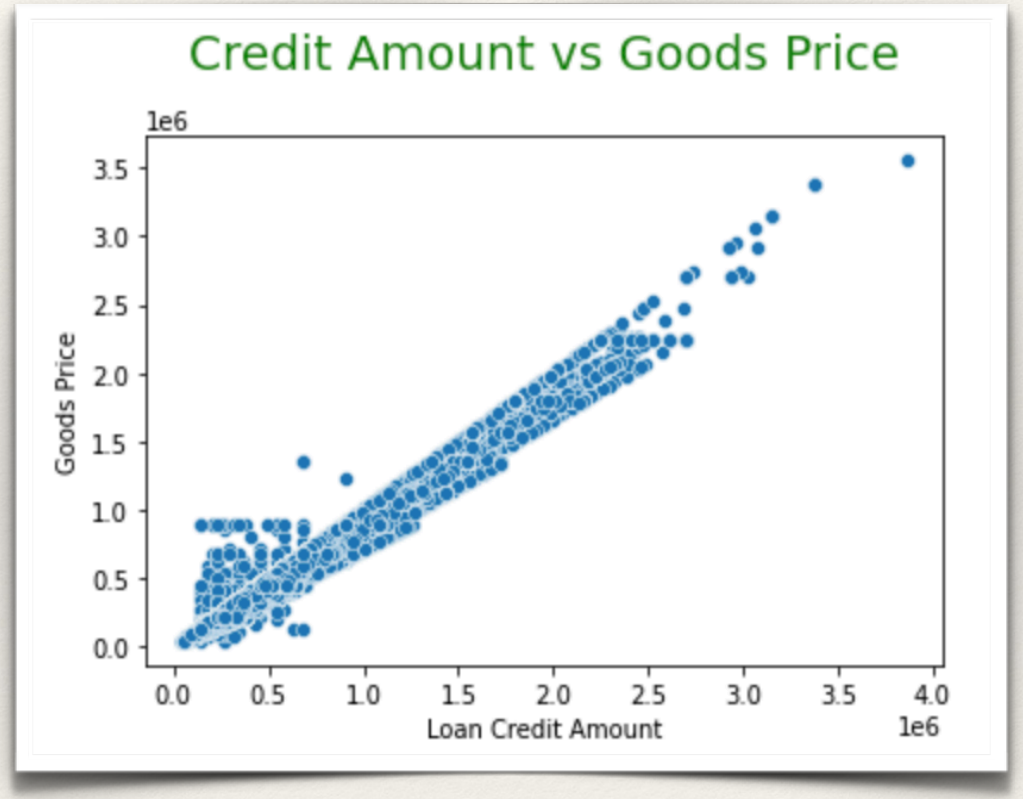
Income, Credit Amount and Annuity Amount Correlation

- As expected, credit amount and annuity are highly correlated
- But, income has no correlation with credit amount and annuity



Credit Amount and Price of goods

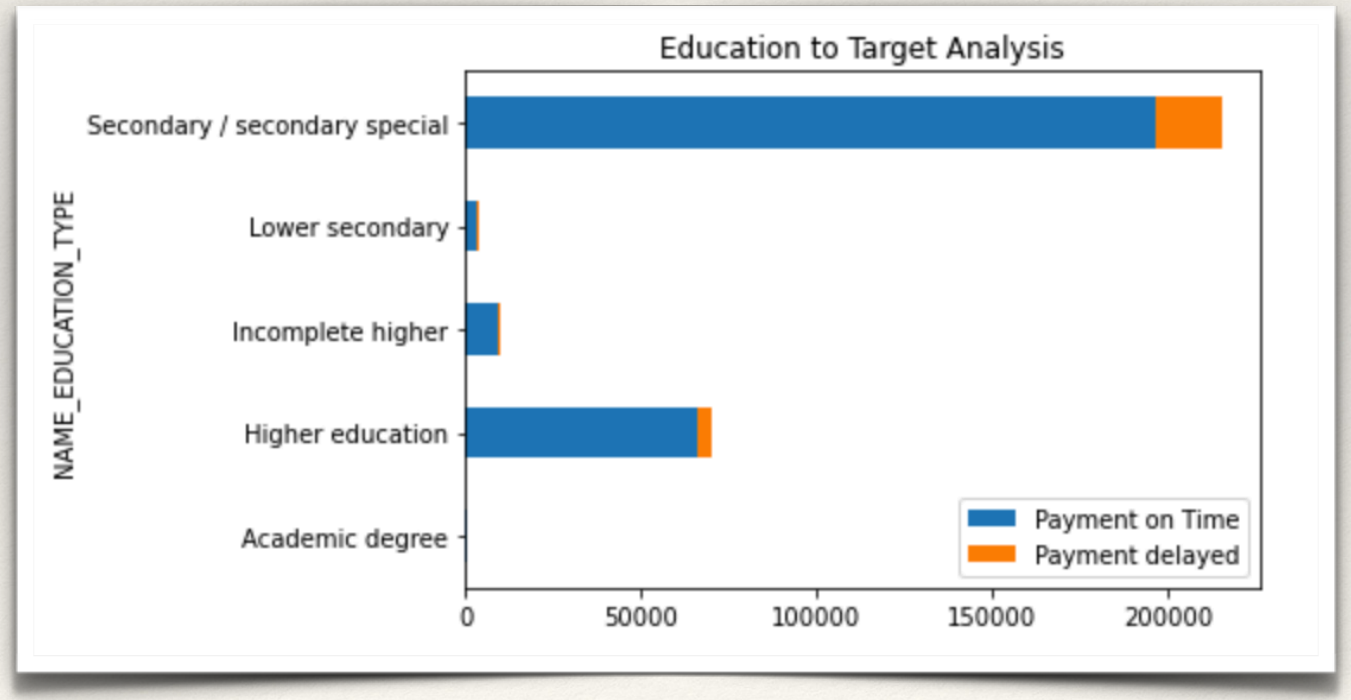
- As expected, credit amount and goods price are highly correlated
- But for lesser values, we see there are some credit amount that is not enough for buying the goods
- And in few cases, credit amount is more than the price of the goods



Education to Target Analysis

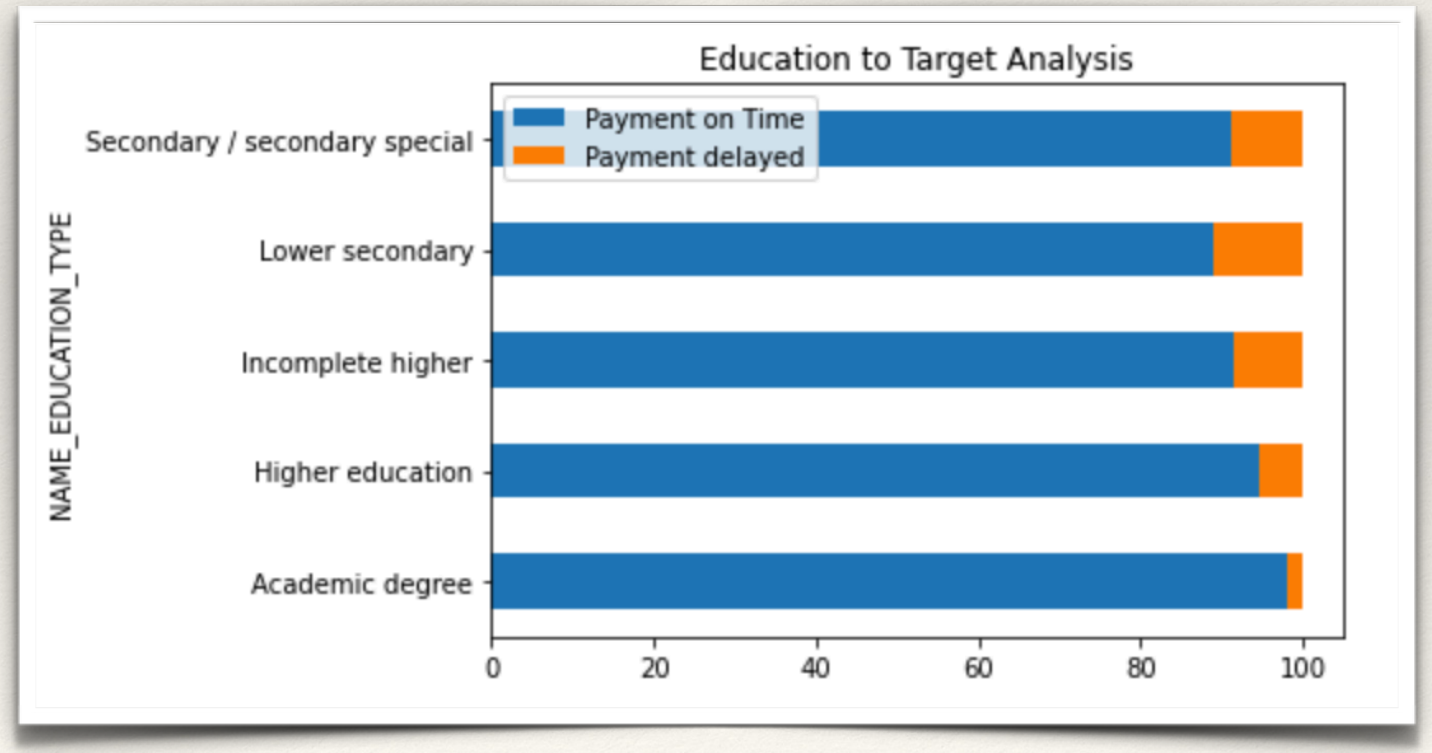
This shows the count of applicants who had / did not have difficulty paying installments with respect to their education

But, since the counts in each category are different, we cannot see the percentage of late payments.



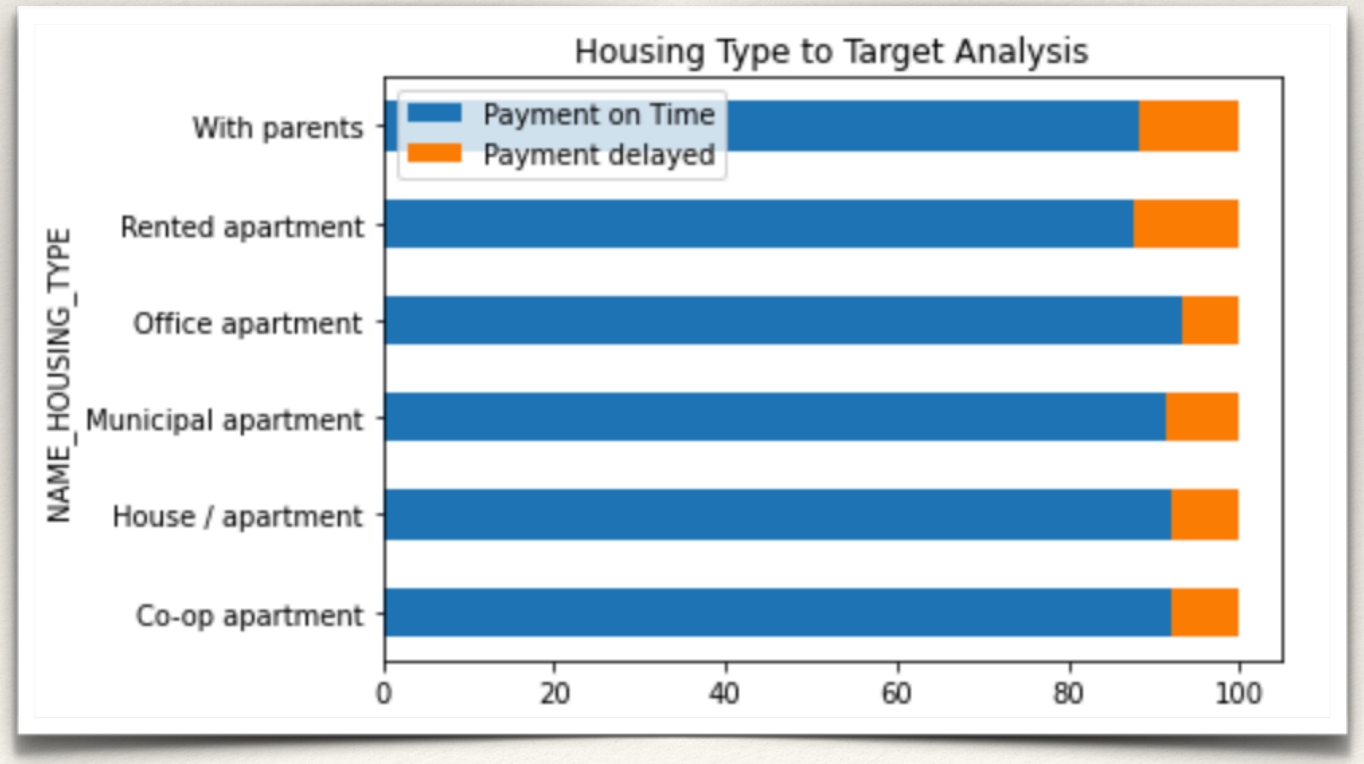
So, lets normalize the same graph

- The applicants with Lower secondary education has the highest percentage of late payments
- Followed by Secondary / Secondary Special and Incomplete higher education



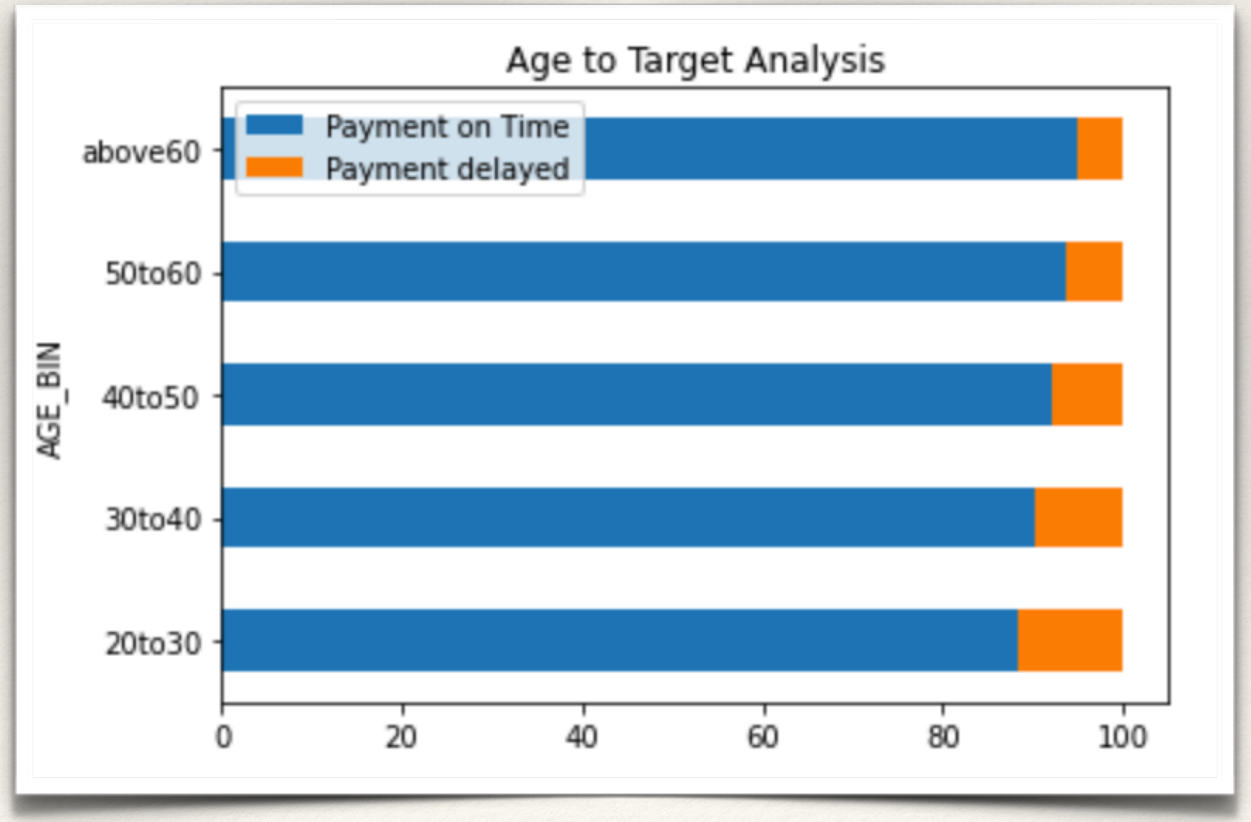
Housing to Target Analysis

- The applicants who are staying with their parents or at rented apartments have had most difficulty paying installments



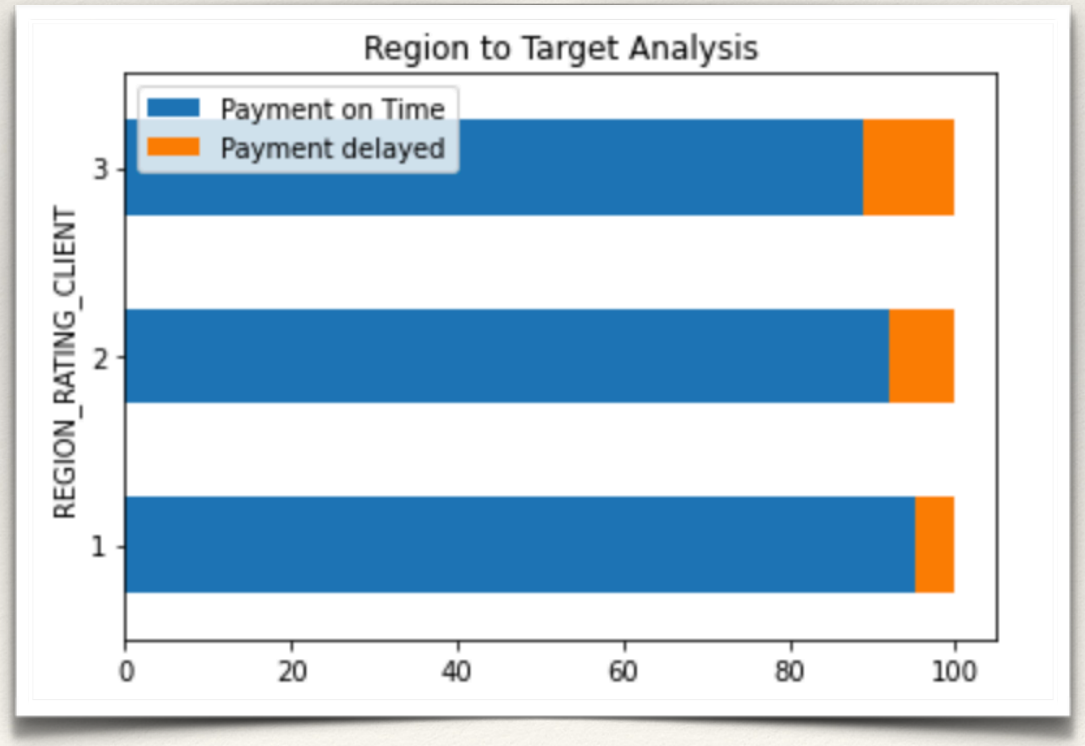
Age to Target Analysis

- As the applicants' age increases, more chances of paying installments on time.
- Applicants between age 20 to 30 have the most percentage with payment difficulty



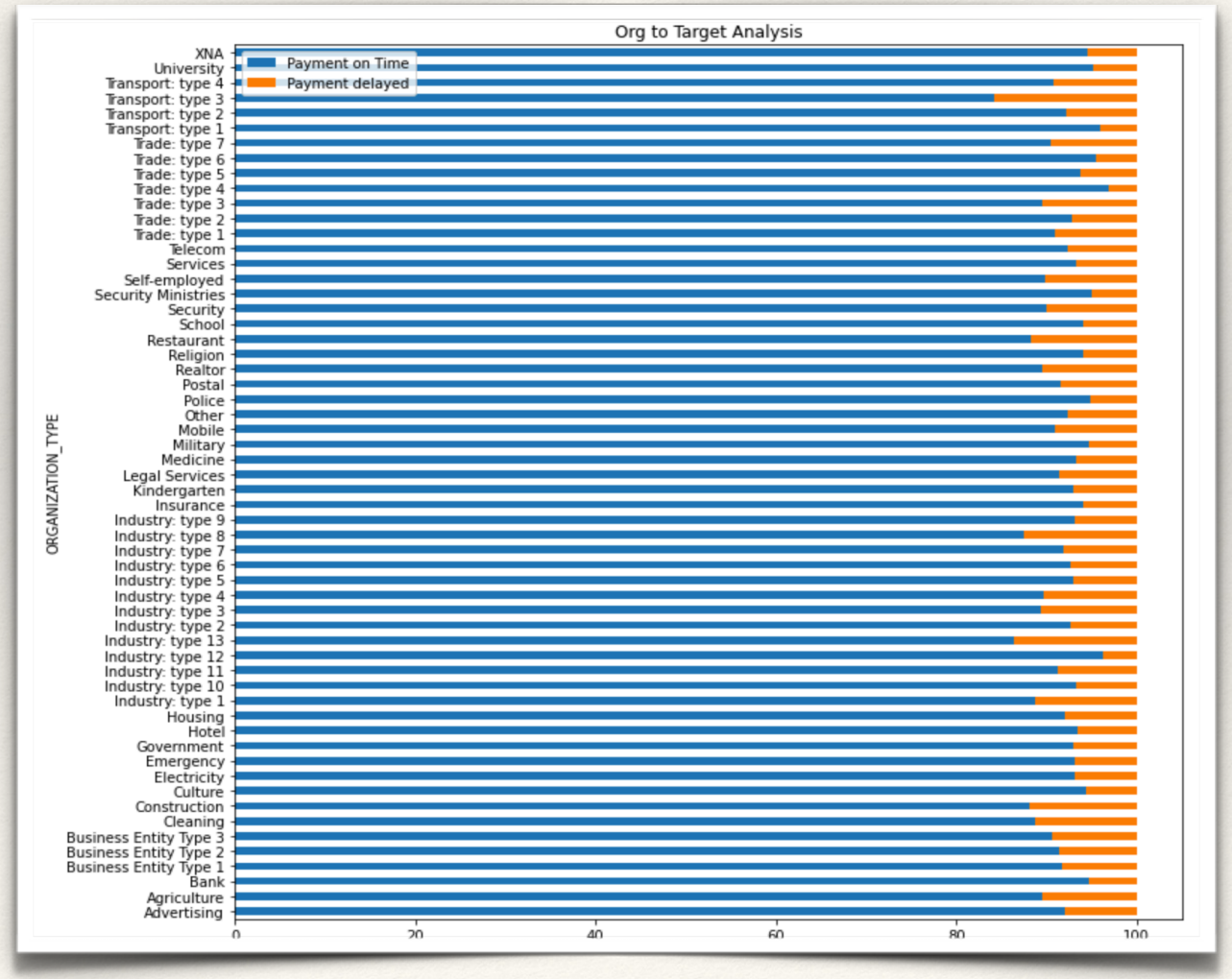
Region to Target Analysis

- Applicants from Region with rating 3 have more percentage with payment difficulty



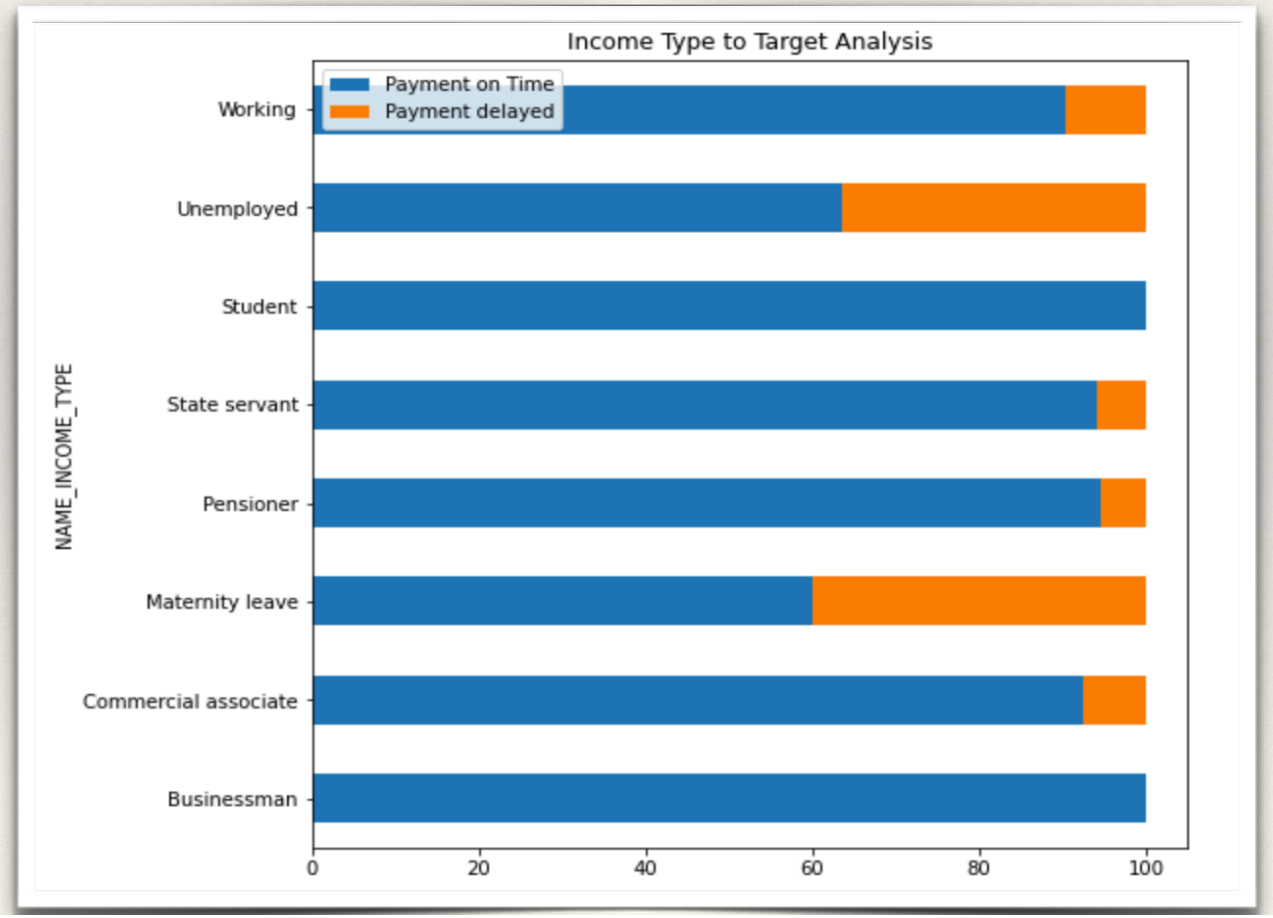
Working Org to Target Analysis

- Working organization of applicant and Target does not seem to have a clear pattern.



Income Type to Target Analysis

- Applicants who are unemployed and who are on maternity leave have had major payment difficulties
- Students and Businessman seem to have had no difficulty paying back the loan

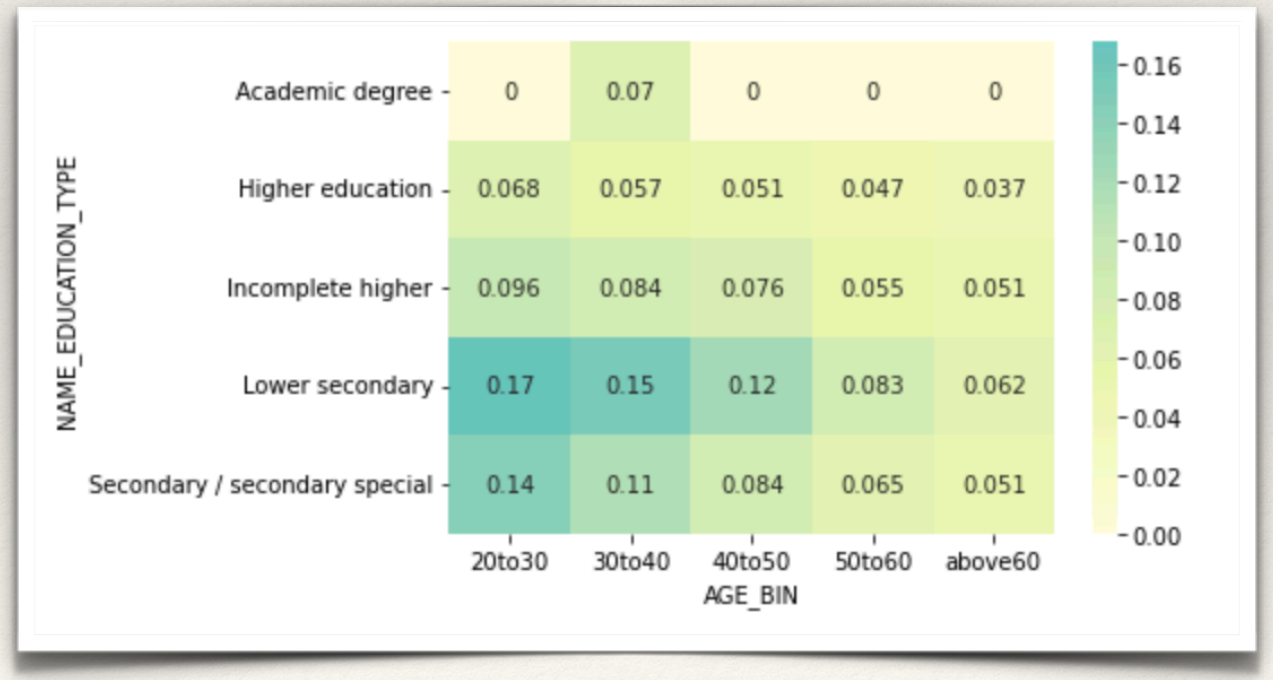


Income Type and Age to Target

- All the applicants who are on Maternity Leave who had payment difficulty are in 30 to 40 age group
- Those who are unemployed and of age above 60 also have high rate of payment difficulty

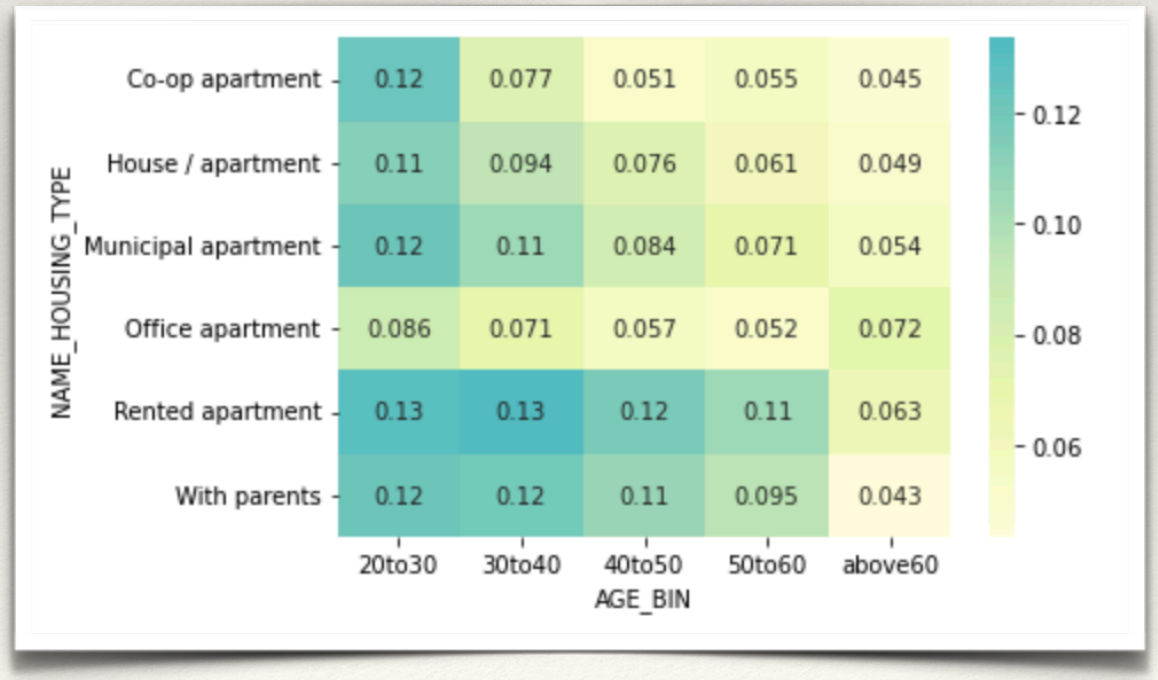
Age and Education against Target

- Younger people who have Lower Secondary, Secondary / Secondary Special education have had more difficulty paying installments
- Payment difficulties tend to reduce with higher education and if applicant is older



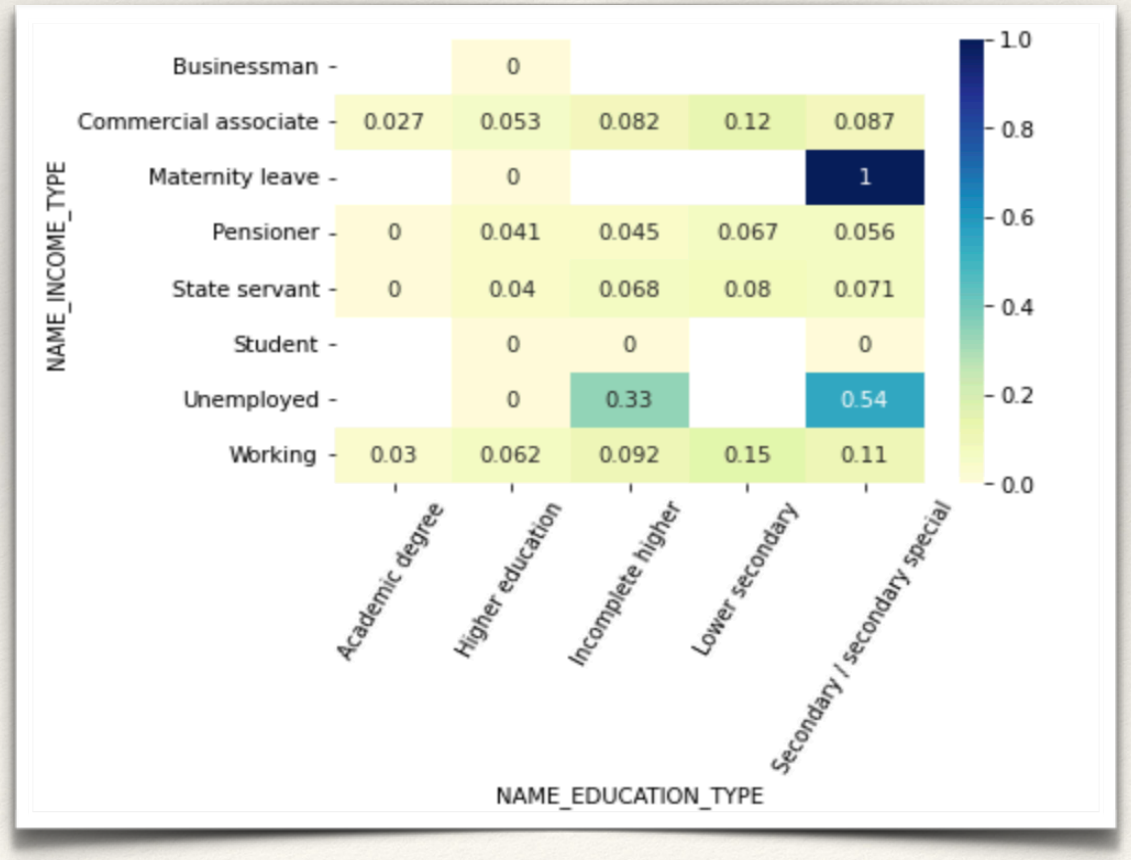
Housing Type and Age to Target

- Again, age seems to be the influencing factor, with younger applicants missing more payments
- People staying with parents or in a rented home are more likely to miss payments



Income Type and Education against Target

- All applicants on maternity leave who missed payments had Secondary education
- Applicants who are unemployed with Secondary or incomplete higher education also had considerable missed payments



Previous Application data

Data Understanding and Correction

Previous Loan Application Amount

- There are records in data where amount of loan in application was 0.
- There are 392K such records where amount of loan is 0.
- Maximum amount in previous applications is close to 7 Million

```
count      1.670214e+06
mean       1.752339e+05
std        2.927798e+05
min        0.000000e+00
25%        1.872000e+04
50%        7.104600e+04
75%        1.803600e+05
max        6.905160e+06
Name: AMT_APPLICATION,
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_APPLICATION	AMT_CREDIT	AMT_
6	2315218	175704	Cash loans	0.0	0.0	
7	1656711	296299	Cash loans	0.0	0.0	
8	2367563	342292	Cash loans	0.0	0.0	
9	2579447	334349	Cash loans	0.0	0.0	
22	1172842	302212	Cash loans	0.0	0.0	
...	...	...	...	...	...	
1670186	1433968	272570	Cash loans	0.0	0.0	
1670187	1971628	435554	Cash loans	0.0	0.0	
1670190	2381880	175920	Cash loans	0.0	0.0	
1670192	2101747	339383	Revolving loans	0.0	67500.0	
1670193	1120445	179433	Revolving loans	0.0	0.0	

392402 rows × 6 columns

Correcting Data set

- Removed the records where loan amount was 0.
- Mean has changed from 175K to 229K
- Median changed from 71K to 100K

Before removal of 0 loans

```
count      1.670214e+06
mean       1.752339e+05
std        2.927798e+05
min        0.000000e+00
25%        1.872000e+04
50%        7.104600e+04
75%        1.803600e+05
max        6.905160e+06
Name: AMT_APPLICATION,
```

After removal of 0 loans

```
count      1.277812e+06
mean       2.290463e+05
std        3.157818e+05
min        3.456000e+03
25%        5.174550e+04
50%        1.125000e+05
75%        2.377811e+05
max        6.905160e+06
Name: AMT_APPLICATION,
```

Multiple applications for a loan

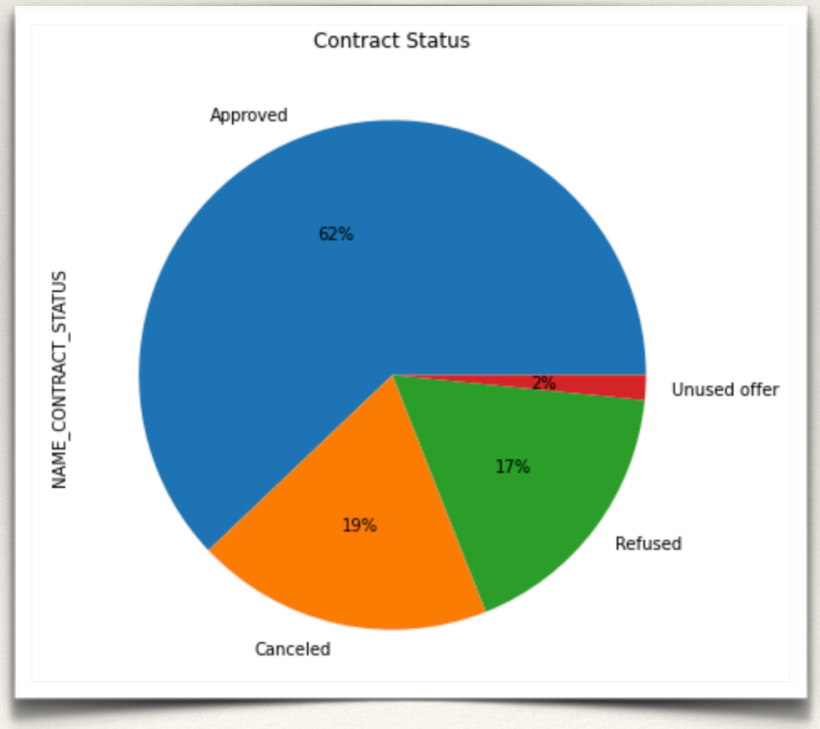
- FLAG_LAST_APPL_PER_CONTRACT shows whether that was the last application for that contract(Last application will be marked as 'Y')
- There are 8475 application which are not the last one.
- All of the 8475 have 'Refused' status
- These data can be removed

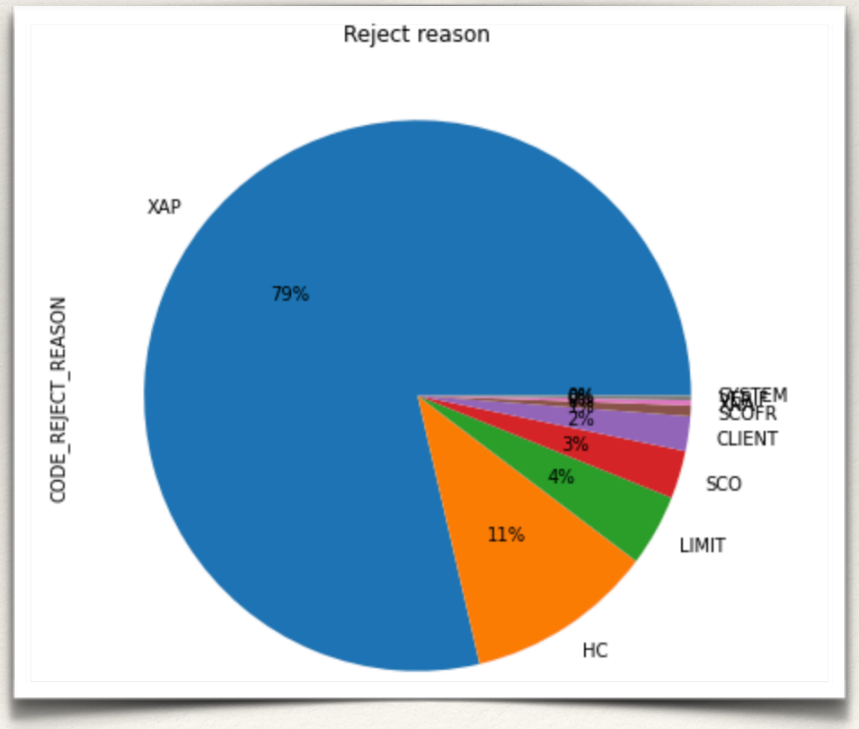
	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE
570	1889112	289922	Revolving loans	12375.0	247500.0	247500.0	NaN	247500.0
352125	2783143	289922	Revolving loans	12375.0	247500.0	247500.0	NaN	247500.0
1445923	2042469	289922	Revolving loans	12375.0	247500.0	247500.0	NaN	247500.0
1510066	2298361	289922	Revolving loans	12375.0	247500.0	247500.0	NaN	247500.0

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE
482	2319125	218182	Revolving loans	22500.0	337500.0	900000.0	NaN	337500.0
128414	1146460	218182	Revolving loans	22500.0	337500.0	675000.0	NaN	337500.0
263349	1527145	218182	Revolving loans	22500.0	337500.0	450000.0	NaN	337500.0

Total Applications

- Total applications after removal is now 1.27 Million
- Total number of unique applicants in the data is only 337K which 1/4th of total
- Below pie chart shows status of these previous applications
- Reject Reason is not giving much clarity, we will look into this in detail





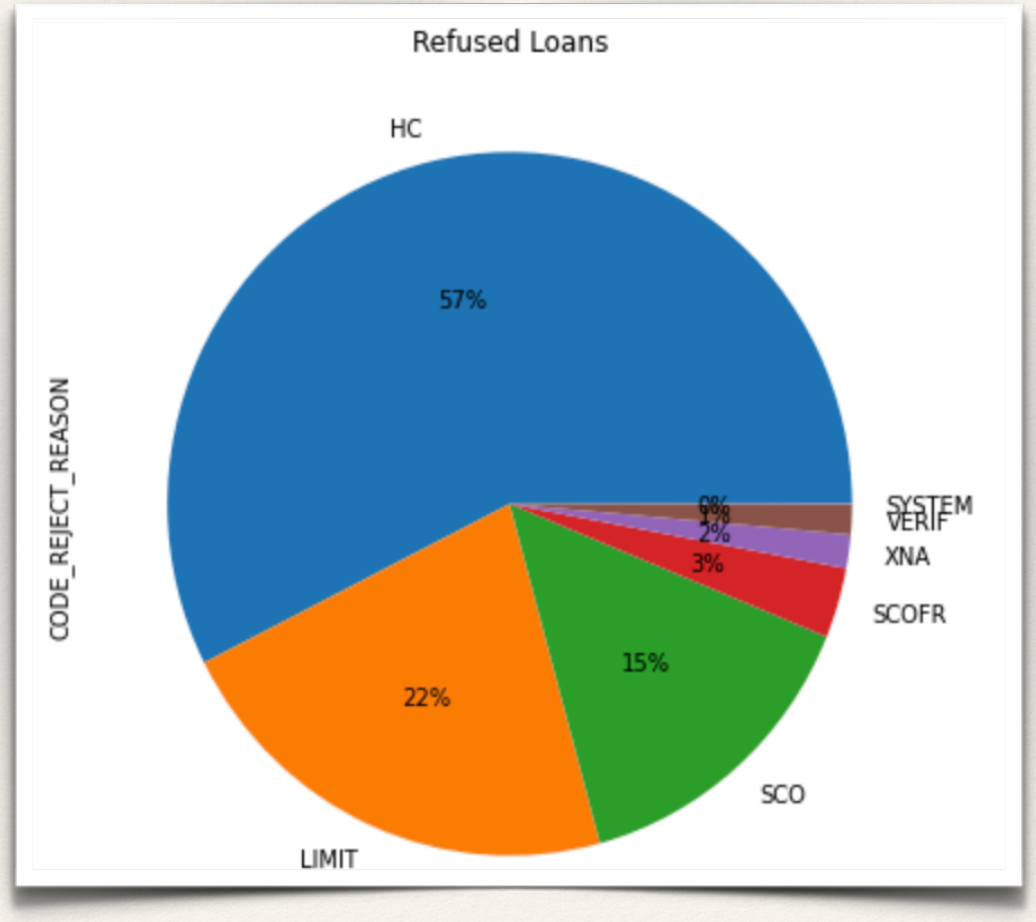
Understanding Reject Reason

- Approved mostly(almost 100%) has the reason 'XAP' which seems to mean approved
- We will make all the approved loans to reject reason 'XAP' for easier analysis
- Canceled loans also have only one reject reason 'XAP'.
- Unused Offer has reason 'Client', which means these were unused due to client
- All other reject reasons are different categories for Refused loans

```
NAME_CONTRACT_STATUS  CODE_REJECT_REASON
Approved              XAP                 991469
                      XNA                      8
Canceled              XAP                 10741
Refused               HC                  141459
                      LIMIT               53682
                      SCO                 35892
                      SCOFR               7965
                      SYSTEM                53
                      VERIF               3376
                      XNA                 3789
Unused offer          CLIENT              25942
Name: CODE_REJECT_REASON, dtype: int64
```

Refused Loans and reject reason

These are the various reasons for refused loans



Conclusion

We have identified these high risk categories or mix of categories from applicant data

- **Education of Applicant**
  - Applicants with Secondary Education have a 5% to 10% more chance for default
- **Housing Type**
  - Applicants who are living in rented apartments or with parents tend to have difficulty paying installments
- **Age**
  - Those who are from age 20 to 30 have the highest payment difficulty among all ages.
  - The risk of non payment tends to decrease with increase in age.
- **Region**
  - Applicants from region rated 3 have had more difficulty in paying than other regions
- **Income Type**
  - Applicants who are unemployed have had a very large percentage of defaulters.
  - Those who are on maternity leave also tend to miss their payment more often. But we have less data for this category.

Conclusion

- Unemployed applicants with only secondary education and those who are unemployed and above 60 are high risk categories
- Few applicants have income less than the loan annuity amount.