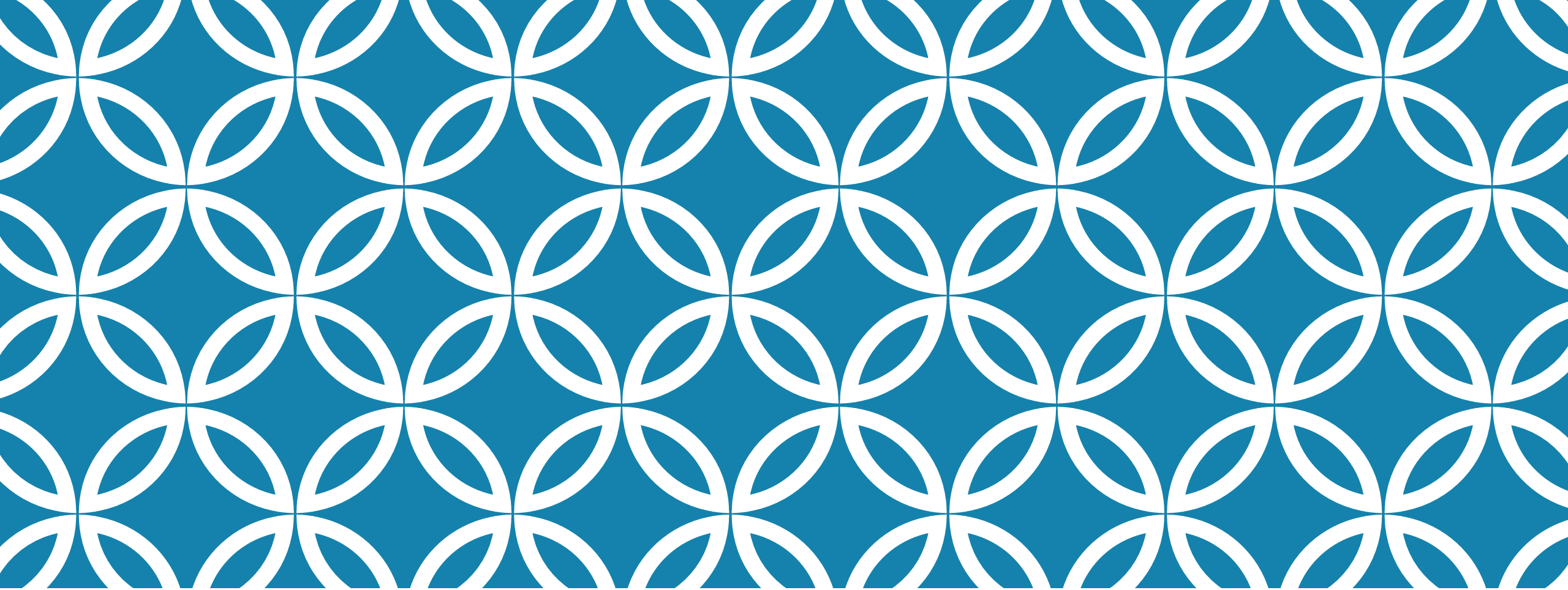


EDA CASE STUDY REPORT

Pradeep Miriyala
Chandra Sekhar Reddy Kandimalla

AGENDA

Data Overview

Data Cleaning – Drop, Impute, Convert

Data Analysis – Univariate & Bivariate

Merge and redo Analysis

Conclusion

DATA OVERVIEW

Application Data

Application data set (application_data.csv) contains data about current data.

It consists of around 3 Lakh rows and 122 columns.

Previous Application Data

Previous application data set (previous_application.csv) consists of history of loans taken by customer.

The data size is around 16Lakh rows and 37 columns.

DATA CLEANING – DROP

Both application data and previous application data contain missing values. There are columns who have missed values for more than 40% of whole data set size. For this reason, it is suggested to drop columns with such high missing values.

	Null_Percentage
COMMONAREA_MEDI	69.872297
COMMONAREA_AVG	69.872297
COMMONAREA_MODE	69.872297
NONLIVINGAPARTMENTS_MODE	69.432963
NONLIVINGAPARTMENTS_AVG	69.432963
NONLIVINGAPARTMENTS_MEDI	69.432963
FONDKAPREMONT_MODE	68.386172
LIVINGAPARTMENTS_MODE	68.354953
LIVINGAPARTMENTS_AVG	68.354953
LIVINGAPARTMENTS_MEDI	68.354953
FLOORSMIN_AVG	67.848630
FLOORSMIN_MODE	67.848630
FLOORSMIN_MEDI	67.848630
YEARS_BUILD_MEDI	66.497784
YEARS_BUILD_MODE	66.497784
YEARS_BUILD_AVG	66.497784
OWN_CAR_AGE	65.990810
LANDAREA_MEDI	59.376738
LANDAREA_MODE	59.376738
LANDAREA_AVG	59.376738
BASEMENTAREA_MEDI	58.515956
BASEMENTAREA_AVG	58.515956
BASEMENTAREA_MODE	58.515956
EXT_SOURCE_1	56.381073
NONLIVINGAREA_MODE	55.179164
NONLIVINGAREA_AVG	55.179164
NONLIVINGAREA_MEDI	55.179164
ELEVATORS_MEDI	53.295980
ELEVATORS_AVG	53.295980
ELEVATORS_MODE	53.295980
WALLSMATERIAL_MODE	50.840783
APARTMENTS_MEDI	50.749729
APARTMENTS_AVG	50.749729
APARTMENTS_MODE	50.749729
ENTRANCES_MEDI	50.348768
ENTRANCES_AVG	50.348768
ENTRANCES_MODE	50.348768
LIVINGAREA_AVG	50.193326
LIVINGAREA_MODE	50.193326
LIVINGAREA_MEDI	50.193326
HOUSETYPE_MODE	50.176091
FLOORSMAX_MODE	49.760822
FLOORSMAX_MEDI	49.760822
FLOORSMAX_AVG	49.760822
YEARS_BEGINEXPLUATATION_MODE	48.781019
YEARS_BEGINEXPLUATATION_MEDI	48.781019
YEARS_BEGINEXPLUATATION_AVG	48.781019

From Application Data

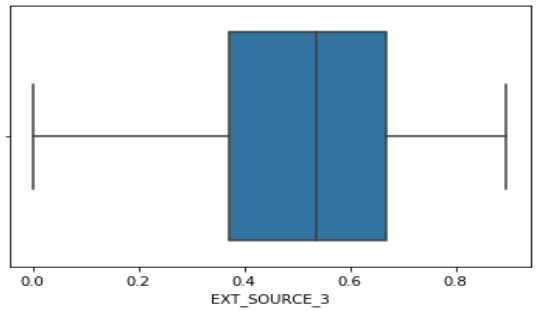
	Null_Percentage
RATE_INTEREST_PRIVILEGED	99.643698
RATE_INTEREST_PRIMARY	99.643698
AMT_DOWN_PAYMENT	53.636480
RATE_DOWN_PAYMENT	53.636480
NAME_TYPE_SUITE	49.119754
NFLAG_INSURED_ON_APPROVAL	40.298129
DAYS_TERMINATION	40.298129
DAYS_LAST_DUE	40.298129
DAYS_LAST_DUE_1ST_VERSION	40.298129
DAYS_FIRST_DUE	40.298129
DAYS_FIRST_DRAWING	40.298129

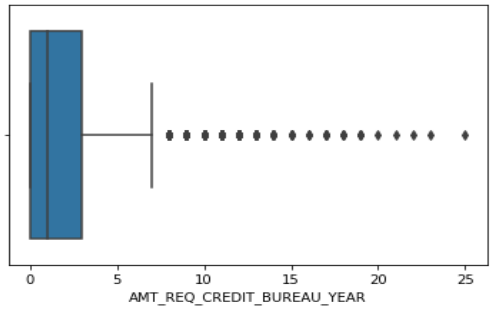
From Previous Application Data

DATA CLEANING - IMPUTE

Some columns have missing values in range of 10-30%. For these columns, missing values will be imputed.

Column	Impute decision
OCCUPATION_TYPE	This is a categorical type with different occupations. Create a new category with name “No Occupation”
EXT_SOURCE_3	This is numerical continuous variable with a normalized score for each customer. Impute missing values with 0.
AMT_REQ_CREDIT_BUREAU_*	The 6 columns are number of enquiries made for customer in past year to hour. Impute missing values with 0.

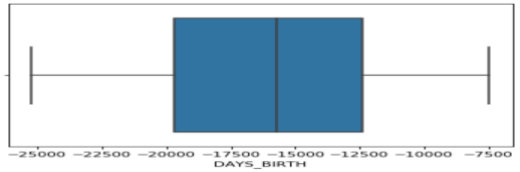


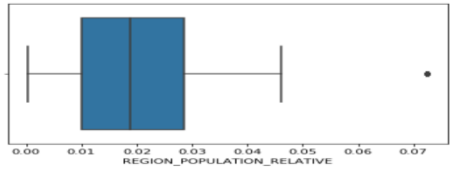


DATA CLEANING – CONVERT

Some columns are stored with types that can be converted either to numeric or category. Partially this is mainly due to fact that when a numeric column is missing value (nan), entire column will be treated as floating although all values are integers.

Column	Type conversion and cleaning
DAYS_*	The columns related to DAYS_* represent number of days since customers birth (or age), since customer is employed etc., Some of these columns are stored as negative numbers. These values must be converted to positives.
AMT_INCOME_TOTAL, AMT_CREDIT	Both these columns store income range, credit taken from bank. Although absolute numbers are useful in exact payments, for analysis, we can simply bin them to get idea of income range, credit ranges.
REGION_POPULATION_RELATIVE, EXT_SOURCE_3, EXT_SOURCE_2	These columns represent normalized scores related to population density and score given by external source to customer. For this reason, these can be binned to individual categories.





DATA ANALYSIS

The column “TARGET” is filled with either 0 or 1. 0 meaning customer has no difficulty in payments and 1 meaning customer has difficulty in payments.

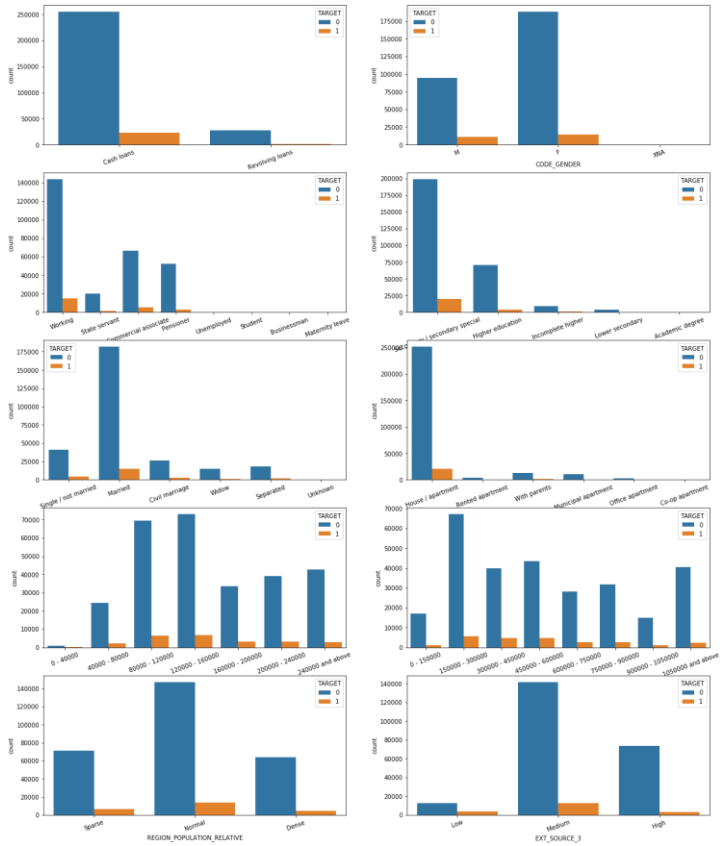
The imbalance percentages of both values are:

- Customer having no difficulty in payments : 91.9%
- Customer having difficulty in payments : 8.1%

For simplicity in further analysis, data will be split based on this column in to two individual data sets.

DATA ANALYSIS - UNIVARIATE ANALYSIS

FOR CATEGORICAL VARIABLES



Based on visual inspection of plots, below class of customers have more difficulties with payments

1. Cash loans
2. Female customers
3. Working professionals
4. Customers with SSC Education qualification
5. Married customers (This could just be causation)
6. Customers living in "House/apartment“ (This could just be causation)
7. Income range of 80000 to 160000
8. Credit range of 150000 to 600000
9. Customers in cities with nominal population density
10. Customer whose rating is given as "Medium" by external source 3

DATA CORRELATION – COLUMN TO COLUMN

Target = 0 (No difficulty in payment)

	Column1	Column2	Correlation
864	FLAG_EMP_PHONE	DAYS_EMPLOYED	1.00
1982	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
494	AMT_GOODS_PRICE	AMT_CREDIT	0.99
1301	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95
1163	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
1549	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86
2044	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
1735	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83
495	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
433	AMT_ANNUITY	AMT_CREDIT	0.77

Target = 1 (Difficulty in payment)

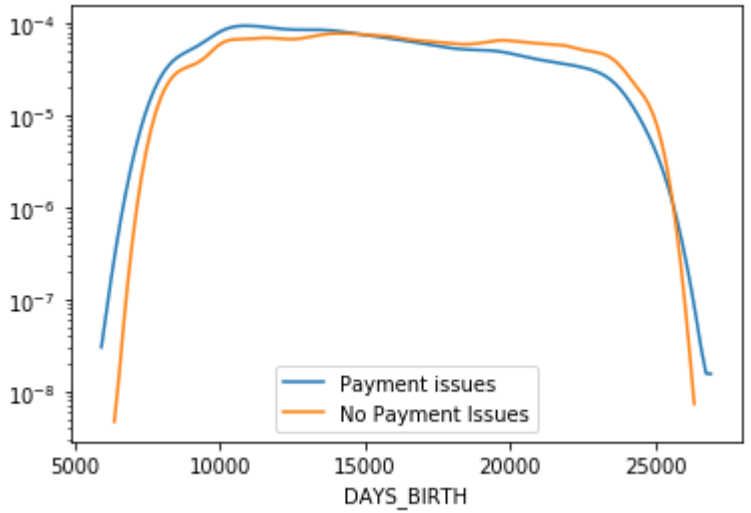
	Column1	Column2	Correlation
864	FLAG_EMP_PHONE	DAYS_EMPLOYED	1.00
1982	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
494	AMT_GOODS_PRICE	AMT_CREDIT	0.98
1301	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96
1163	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
2044	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
1549	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.85
1735	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.78
433	AMT_ANNUITY	AMT_CREDIT	0.75
495	AMT_GOODS_PRICE	AMT_ANNUITY	0.75

We can see that columns are matching for top 5 although correlation values do not match exactly.

DATA ANALYSIS - UNIVARIATE ANALYSIS

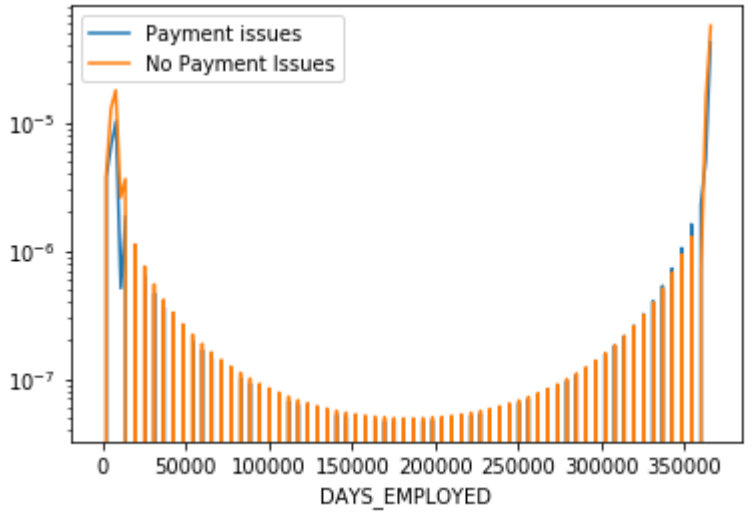
FOR CATEGORICAL VARIABLES

DAYS_BIRTH (AGE)



1. From plot, we can see that young to middle aged people (<15000 days = 41 years), customers usually tend to have more difficulties than people in above middle aged people.
2. While the trend again reverses for old age people (~69 years)

DAYS_EMPLOYED

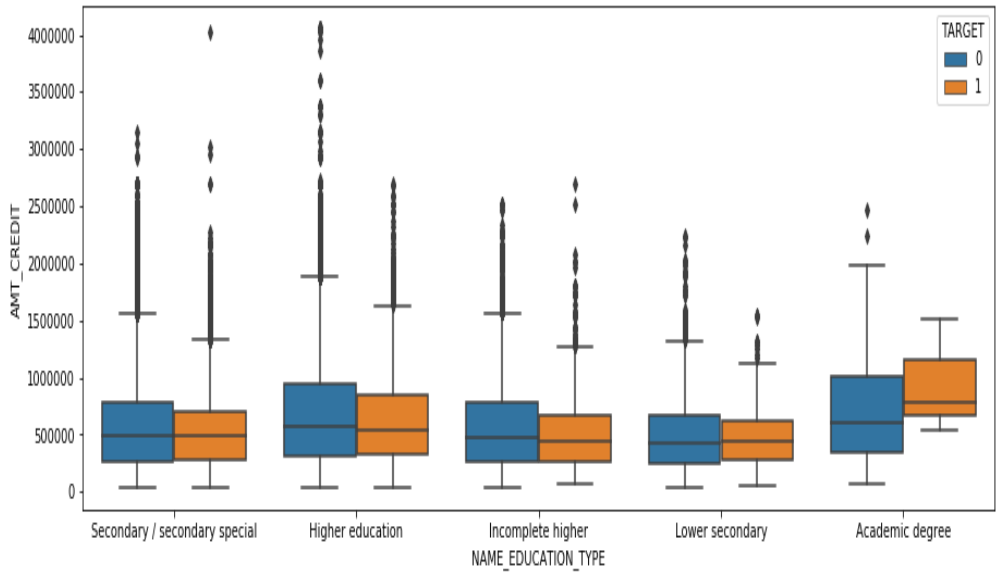


1. The plot suggests for people whose duration of employment is low, there are more issues with payment difficulties.

DATA ANALYSIS - BIVARIATE

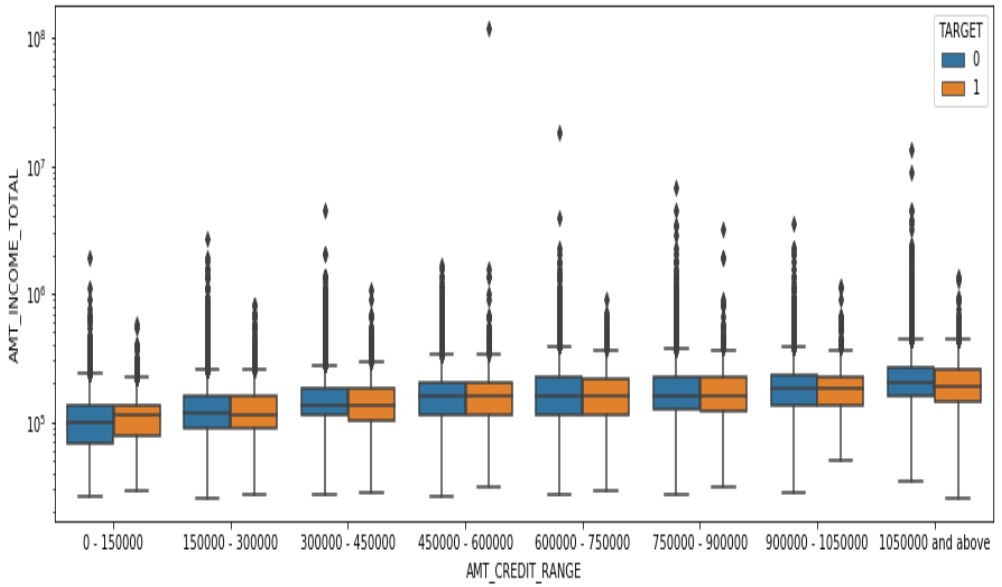
FOR NUMERICAL VARIABLES

EDUCATION, CREDIT_AMOUNT



1. From above plot we can see that on average, people with Academic degree have taken high credit and as credit amount increases, the risk of defaulting (TARGET=1) is also higher

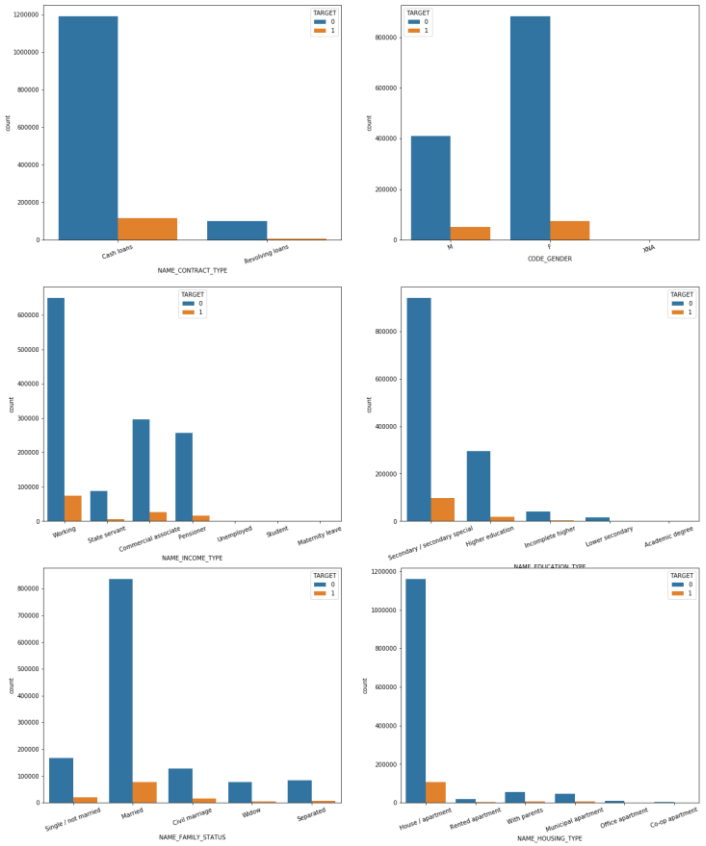
INCOME and CREDIT AMOUNT



1. For people with low income range (0 to 1.5Lakh), on average there are high payment difficulties.
2. For people with high income range (10.5Lakh and above), on average there are less payment difficulties.

DATA ANALYSIS - UNIVARIATE ANALYSIS

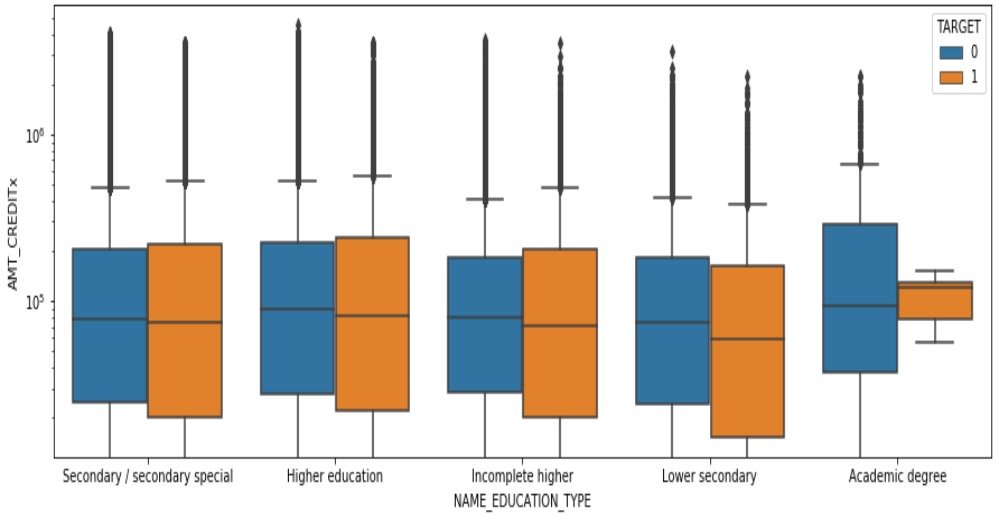
ON MERGED DATA SET



Comparing the plot pattern with plots generated earlier (slide 8), the previous analogy still holds good.

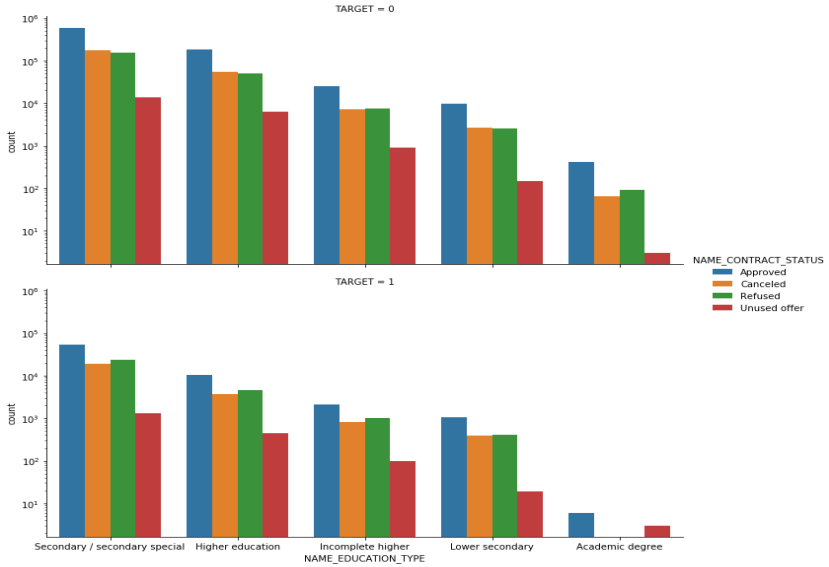
DATA ANALYSIS - BIVARIATE ANALYSIS

Education and Credit Amount



1. Looking at history, people with other than "Academic degree" consistently had issue with payments

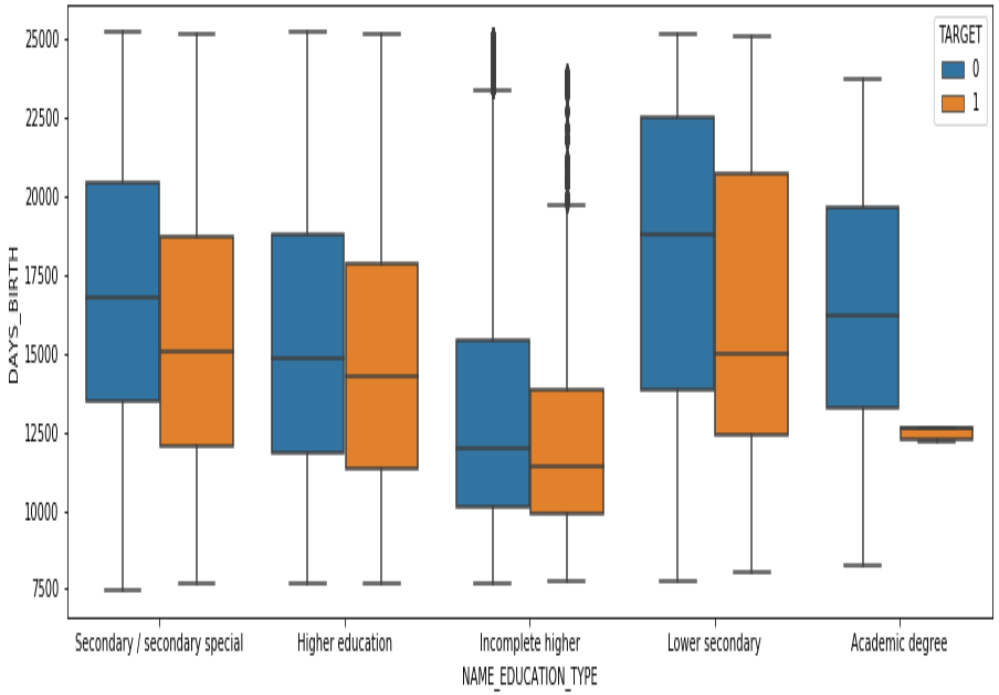
Education and Contract Status



1. Consistently people with secondary education are largest customers in all categories
2. When there are payment difficulties, people with academic degree either have approved loan or unused offer.

DATA ANALYSIS - BIVARIATE ANALYSIS

EDUCATION and DAYS of BIRTH (Age)



1. In general, people at younger age tend to have more payment difficulties compared to people with older age.
2. When we look at people with "academic degree", people at young age have payment difficulties isolating the age range. We can conclude that for people with "academic degree" , the young age is where more defaults happen.
3. The means of defaulter's age across every education type is consistently lower compared to people who do not default.

CONCLUSIONS

Female customers consistently have payment difficulties

Customers with Secondary / Secondary special education have high payment difficulties

Cash loans have high payment difficulties, however there is large quantum of cash loans as well when compared to revolving loans.

People living in "House/apartment" have high payment difficulties. But this is just a causation since people have to live somewhere.

"Married" people have high payment difficulties.

"Young people" with Academic degree often tend to default