# upGrad

## Statistics and EDA - Module 09 – Credit EDA Case Study

*Author*: Anish Mahapatra, Karthik  Premanand

*Email*: anishmahapatra01@gmail.com, karthikprem26@gmail.com

**Business Objectives**

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
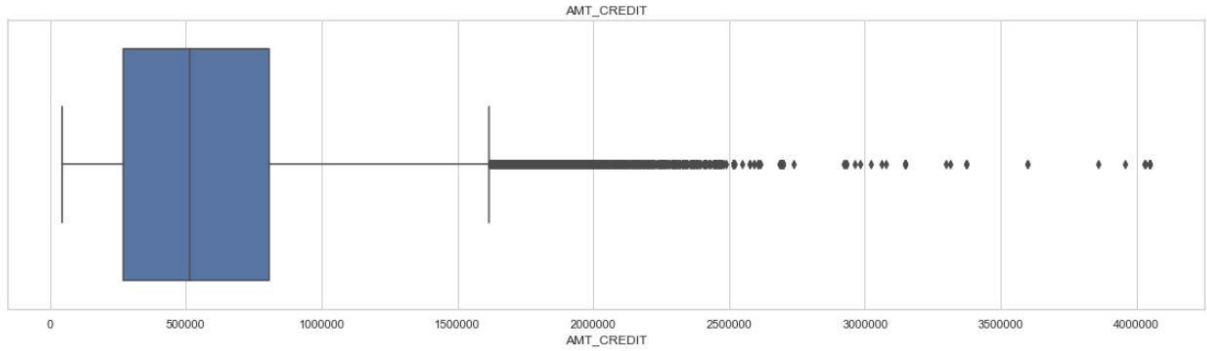
The conclusion from the graphs can be found as a part of the image in the interest of time. Kindly consider the same.

## Outlier Treatment Analysis

**Note:** The Boxplots below have been plotted with the standard whiskers of 1.5 x (IQR)
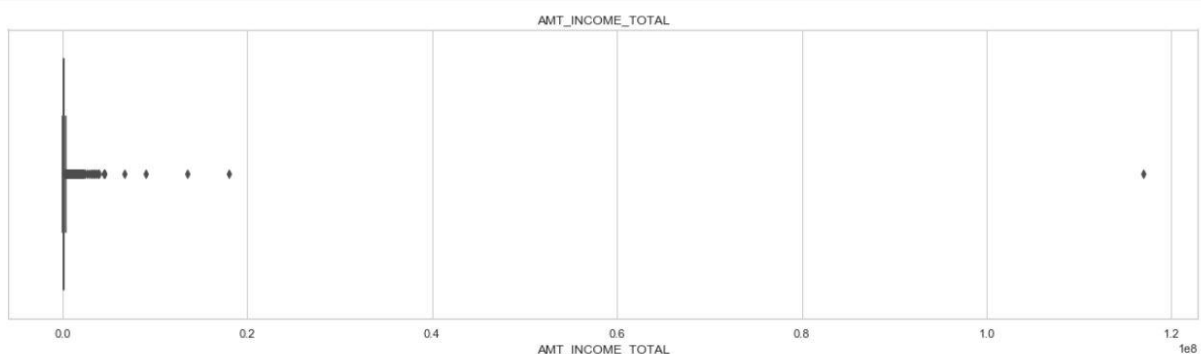


**Findings:** From the above plot of the income of the client above, we notice that quite a few people earn more than the average.

AMT_CREDIT

**Findings:** From the above plot, we see that most people get a loan between 3 to 8 lakhs and a maximum of about 7 lakhs. however, we notice that a considerable amount of people also get loans of upto 40 lakhs
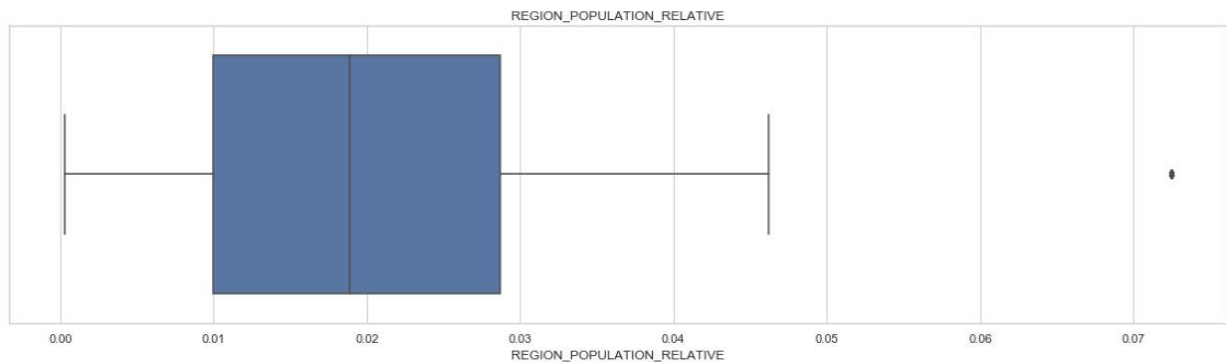


AMT_ANNUITY
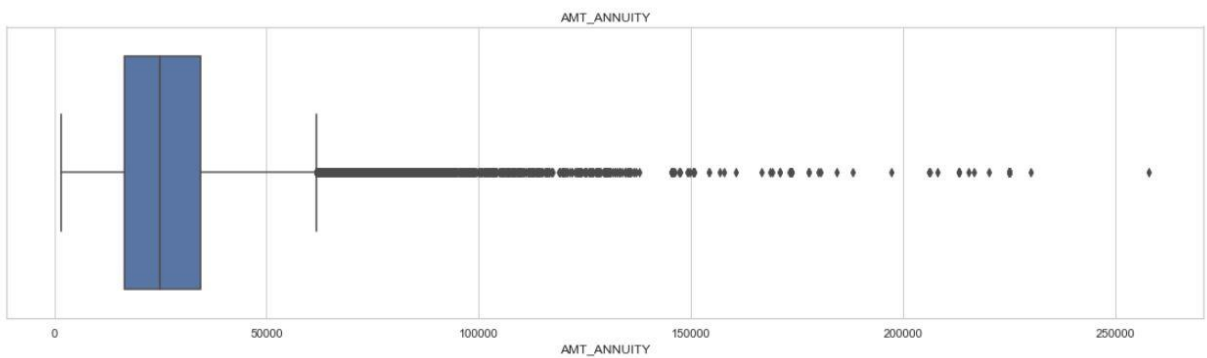
**Findings:** Due to annuity, loans, we see that history suggest that people hold around 65,000 worth of annuity. However, we see that this can go up significantly upto 20,00,000 and above
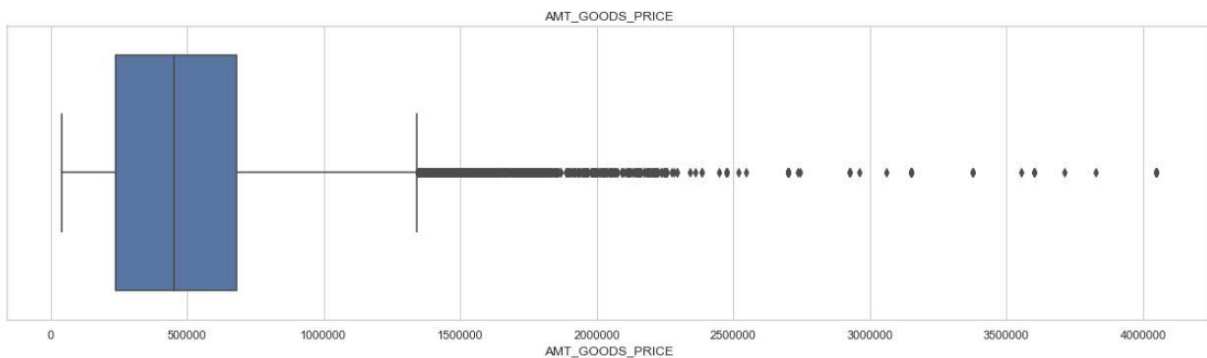


AMT_GOODS_PRICE

**Findings:** AMT_GOODS_PRICE is defined as: For consumer loans it is the price of the goods for which the loan is given.

We notice that it follows trends similar to AMT_CREDIT with most of the spread lying within 13 lakhs and can go up to 40 lakhs.

REGION_POPULATION_RELATIVE



| | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 |
REGION_POPULATION_RELATIVE

**Findings:** REGION_POPULATION_RELATIVE - Normalized population of region where client lives (higher number means the client lives in more populated region)

We notice that one client lives in a highly populated region, whereas the others mostly sat in regions that have 0.01 to 0.03 (approx) normalized population density

DAYS_BIRTH
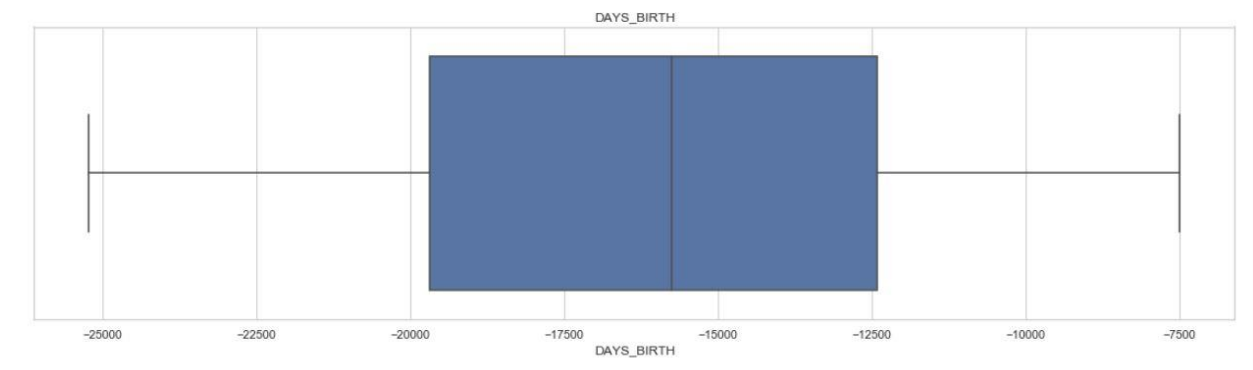


**Findings:** Based on the above plot, we notice that majority of the client are between 34 to 54 years of age with the youngest being 20 and the oldest being around 68
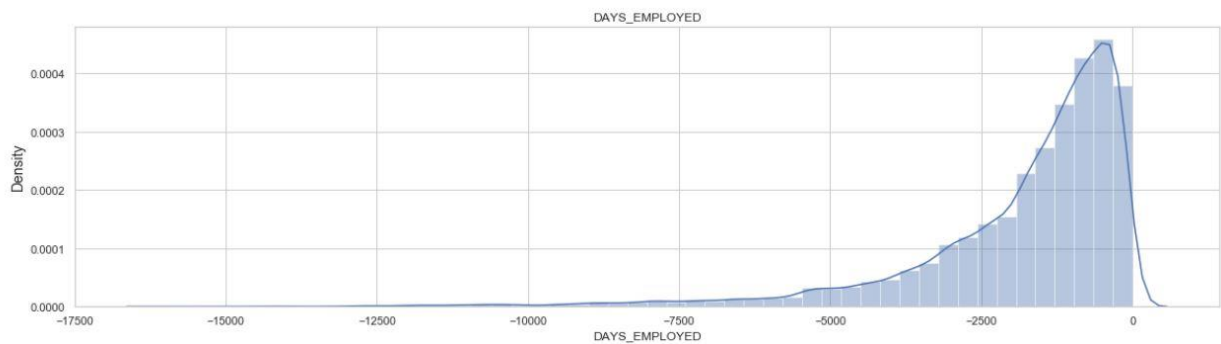
Univariate Analysis on data where 'TARGET' = 1

*Analysis of Client Income*
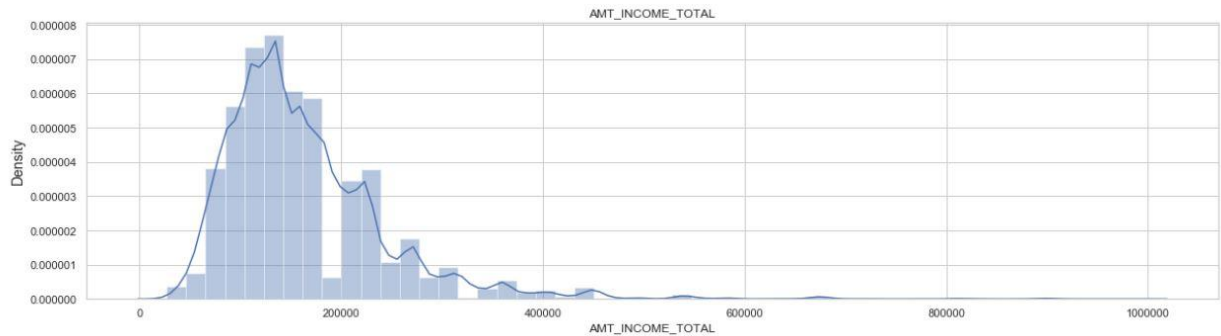
AMT_INCOME_TOTAL

**Findings:** Here we can observe that the highest frequency of income amount for clients is between Rs. 100000 and Rs. 200000

*Analysis of Days of Employment of the client*



DAYS_EMPLOYED

**Findings:** Here we can observe that the highest frequency of days that the clients have been employed is between 625 and 937 days i.e between 1.7 and 2.5 years
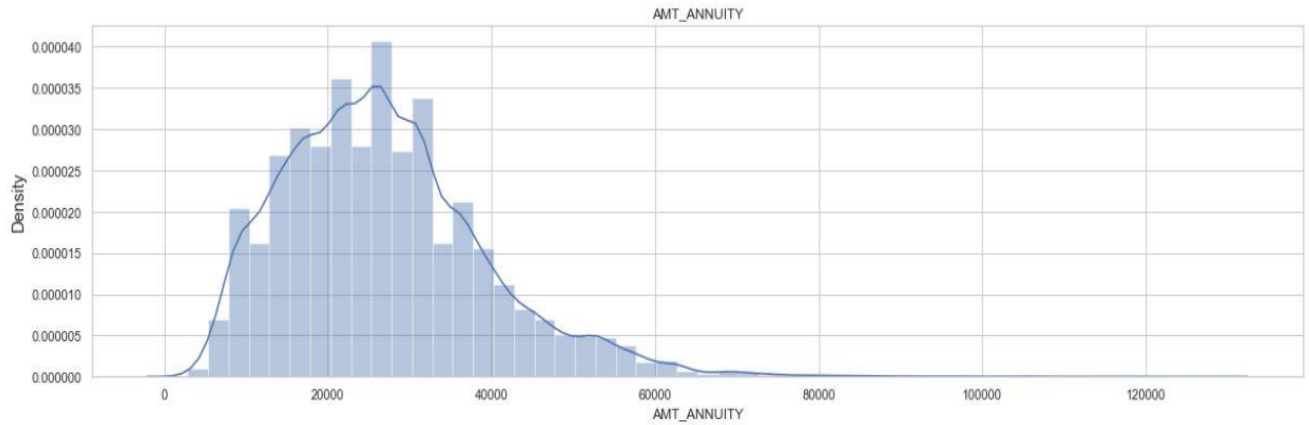
*Analysis of Credit Amount of Loan*
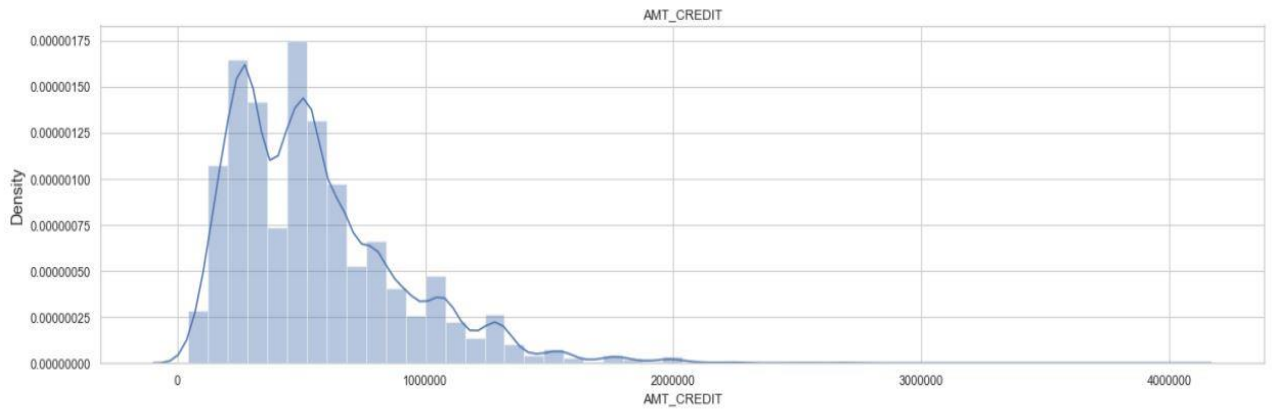
**Findings:** Here we can observe that the highest frequency density of credit is approximately Rs 500000 - Rs 600000

*Analysis of Price of Loan Goods*
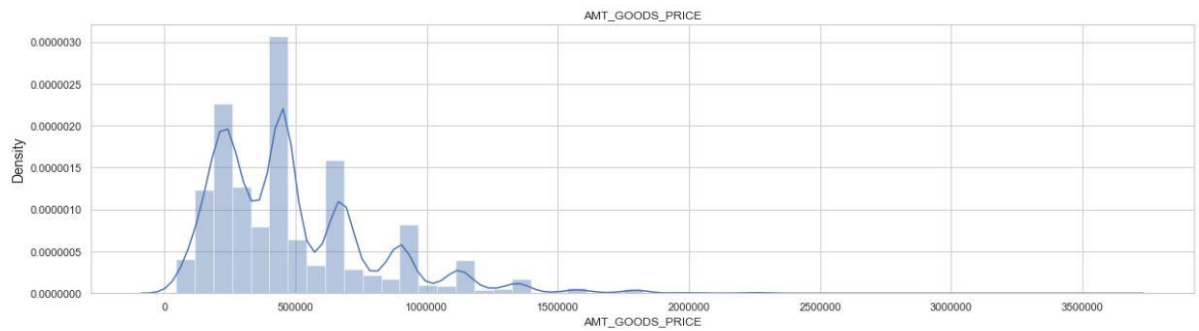


**Findings:** Here we can observe that the highest frequency density of price of the goods is approximately Rs 4,70,000
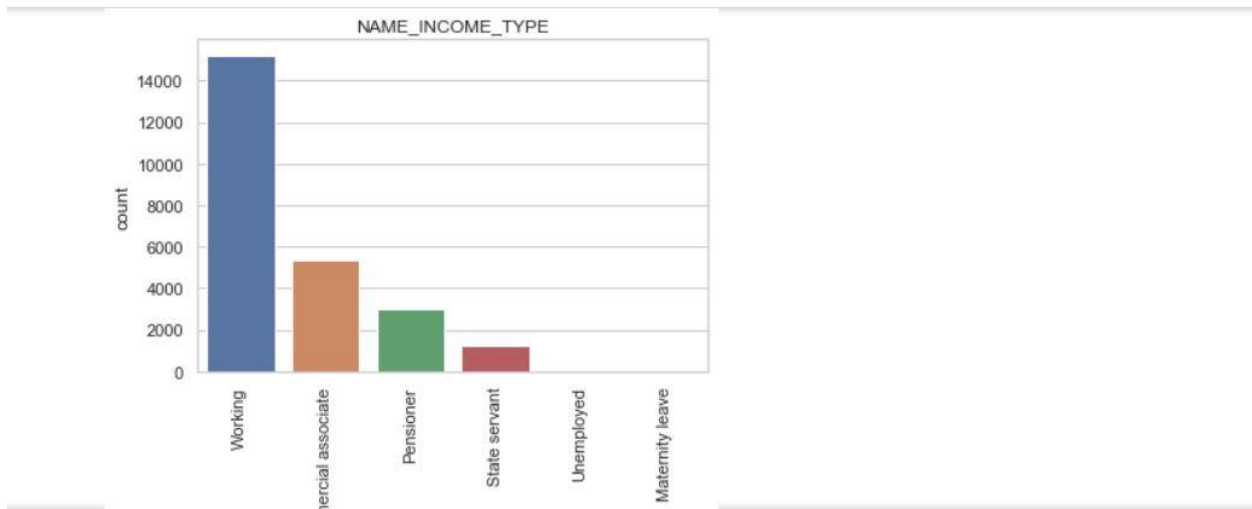
*Analysis of Loan Annuity*

AMT_ANNUITY

**Findings:** Here we can observe that there is relatively even distribution of loan annuity for the clients with approx Rs 28000- Rs 30000 having the highest density
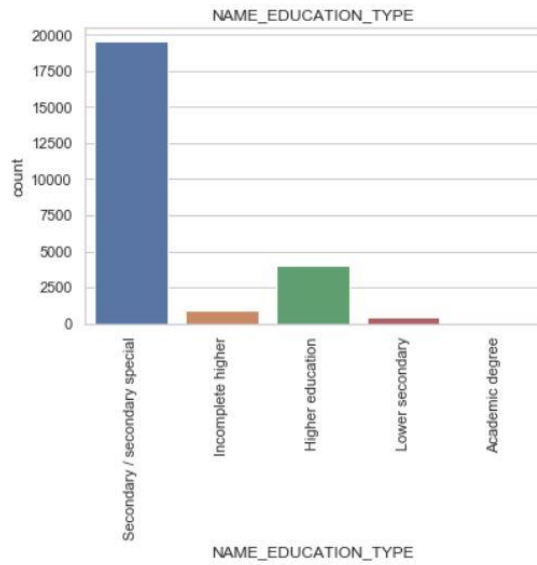
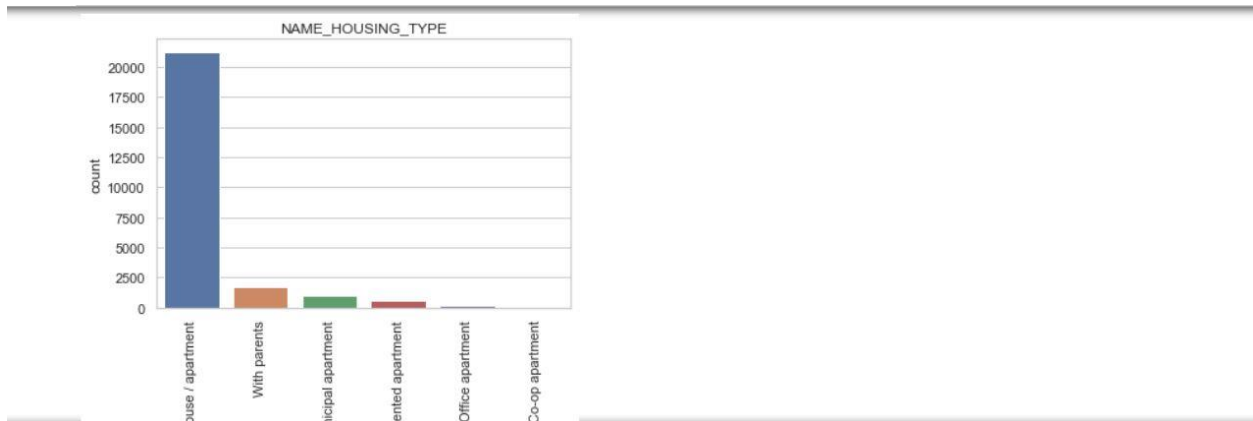## Univariate Analysis for categorical variables

*Analysis of Client Income Type*



NAME_INCOME_TYPE

**Findings:** Here we can observe that most of the clients are working followed by Commercial associates
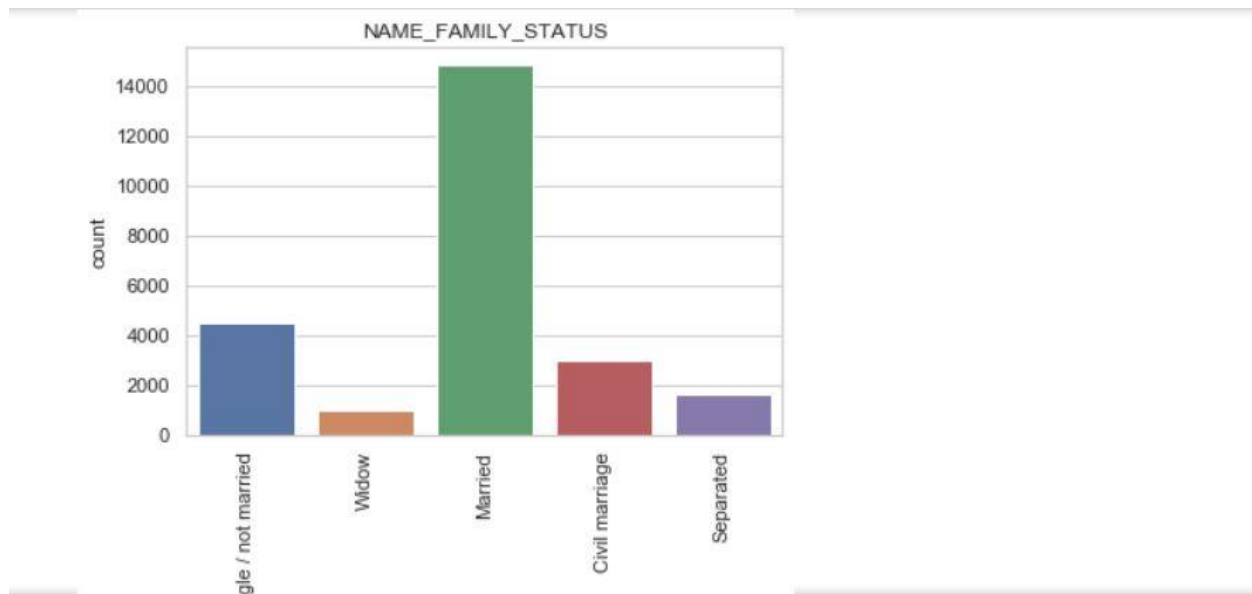
*Analysis of Client Education Level*

**Findings:** Here we can observe that most of the clients completed Secondary / Secondary special followed by Higher Education

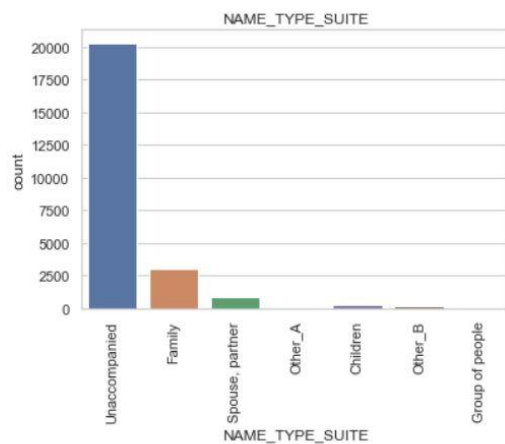*Analysis of Client Housing Type*



**Findings:** Here we can observe that most of the clients stay in House/Apartments of their own followed by that of clients living with their parents

*Analysis of Client Family Status*

**Findings:** Here we can observe that most of the clients are married followed by single/not married status
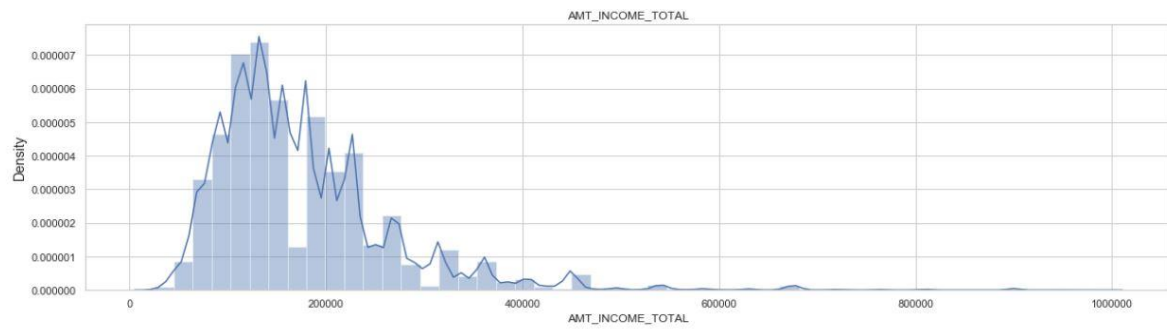
*Analysis of Client Accompaniment status*



Findings: Here we can observe that most of the clients are unaccompanied when applying for loan followed by accompaniment by family
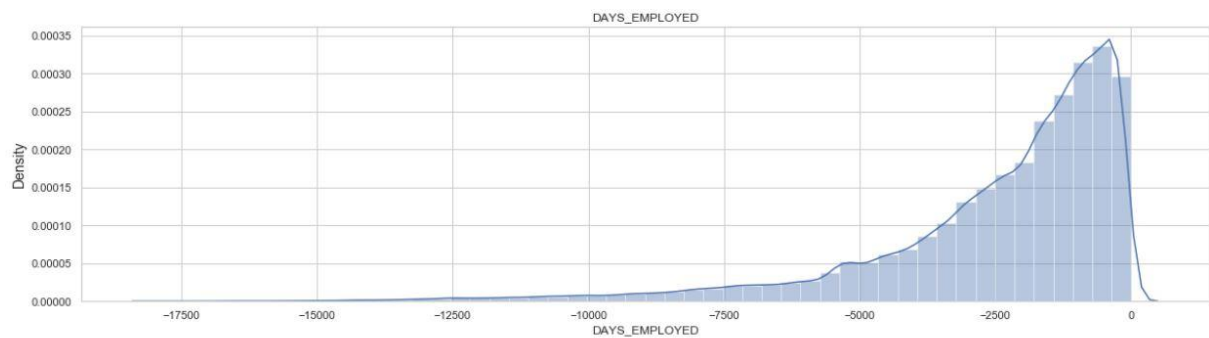
## Univariate Analysis where 'TARGET' = 0

*Analysis of Client Income*

AMT_INCOME_TOTAL

**Findings:** Here we can observe that the highest frequency of income amount for clients is between Rs. 100000 and Rs. 200000

*Analysis of Days of Employment of the client*
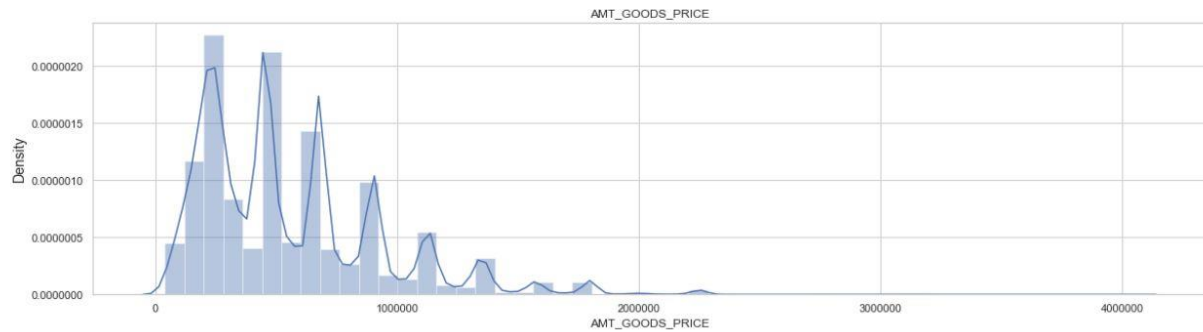


DAYS_EMPLOYED

**Findings:** Here we can observe that the highest frequency of days that the clients have been employed is between 715 and 1040 days i.e between 1.9 and 2.8 years
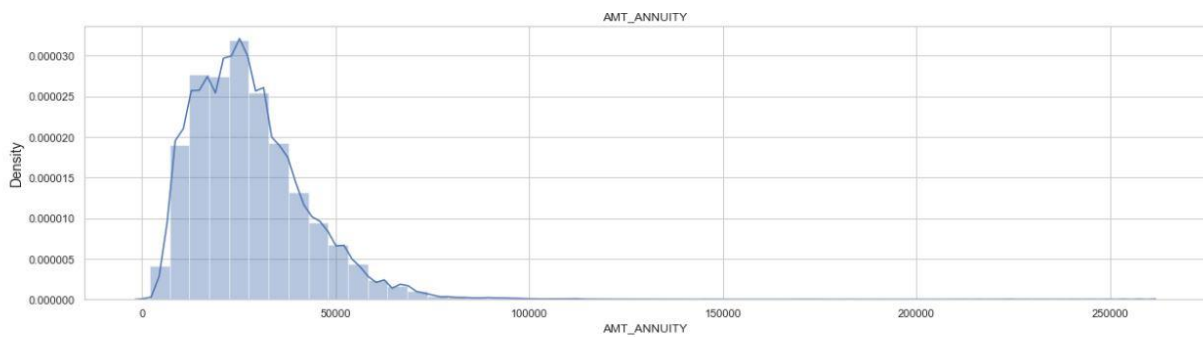
*Analysis of Credit Amount of Loan*
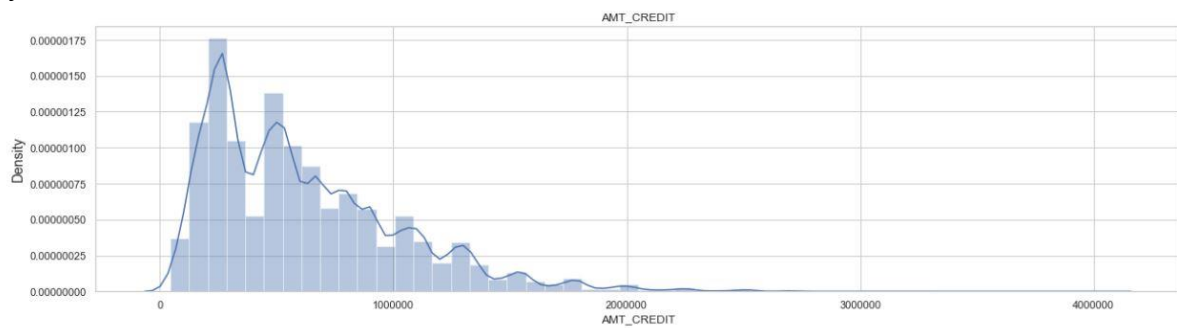


AMT_CREDIT

**Findings:** Here we can observe that the highest frequency density of credit is approximately Rs 200000 - Rs 350000

*Analysis of Price of Loan Goods*



**Findings:** Here we can observe that the highest frequency density of price of the goods is between Rs 3,30,000 and Rs 4,20,000
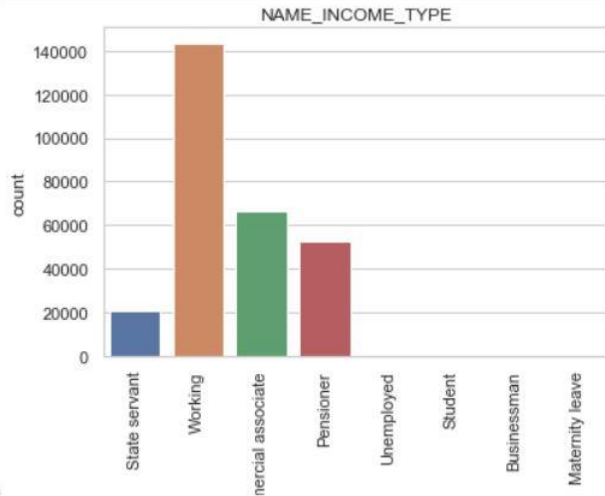
*Analysis of Loan Annuity*



**Findings:** Here we can observe that there is relatively even distribution of loan annuity for the clients with approx Rs 30000- Rs 35000 having the highest density

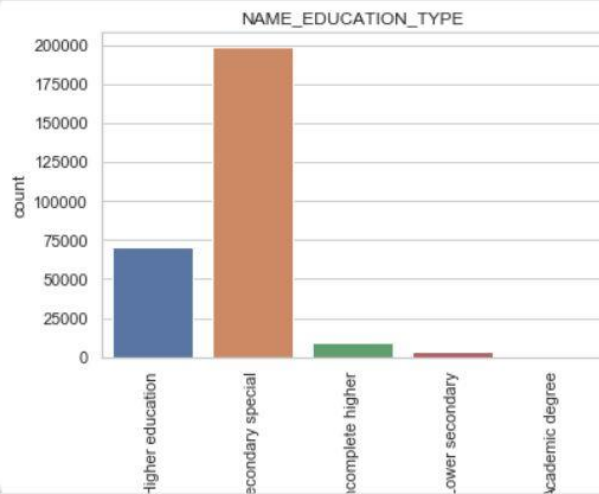## Univariate Analysis for categorical variables
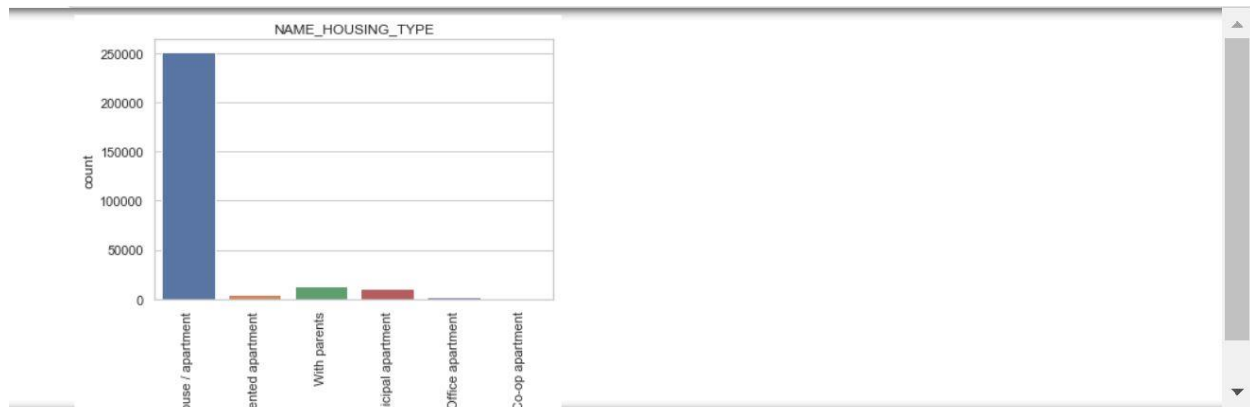
*Analysis of Client Income Type*

**Findings:** Here we can observe that most of the clients are working followed by Commercial associates and Pensioners

*Analysis of Client Education Level*



**Findings:** Here we can observe that most of the clients completed Secondary / Secondary special followed by Higher Education

*Analysis of Client Housing Type*

**Findings:** Here we can observe that most of the clients stay in House/Apartments of their own followed by that of clients living with their parents and Municipal Apartments

*Analysis of Client Family Status*



**Findings:** Here we can observe that most of the clients are married followed by single/not married status
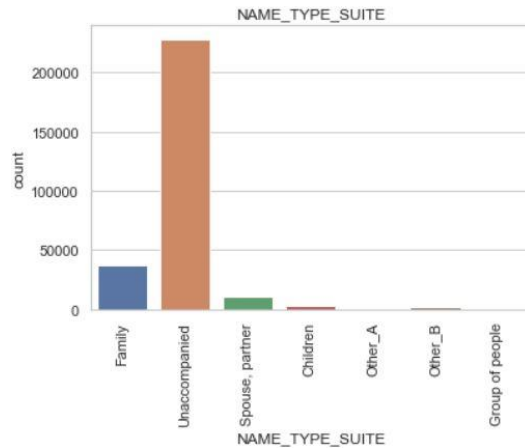
*Analysis of Client Accompaniment status*

**Findings:** Here we can observe that most of the clients are unaccompanied when applying for loan followed by accompaniment by family

# Bivariate Analysis where 'TARGET' = 1

*Analysis of Education Type versus Loan Annuity*
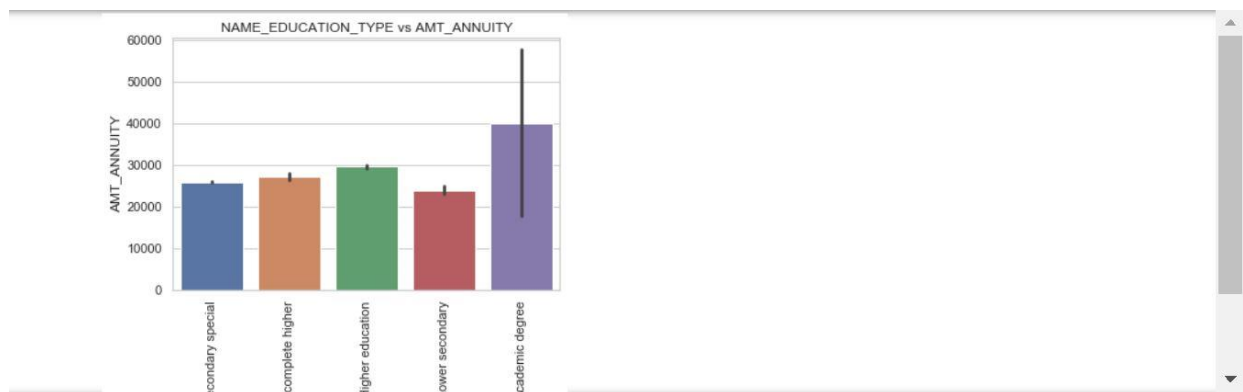


**Findings:** Here we can observe that most of the clients with Academic Degree has the highest mean amount of loan annuity followed by clients that completed Higher education

*Analysis of Family Status versus Loan Credit Amount*

NAME_FAMILY_STATUS vs AMT_CREDIT

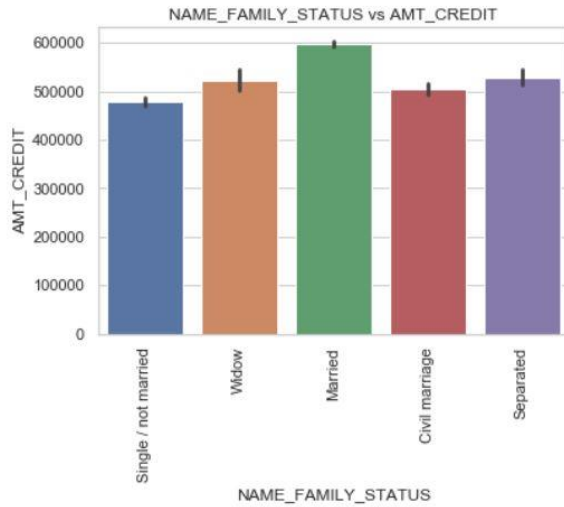**Findings:** Here we can observe that most of the clients that are married have the highest credit amount followed by Widowed clients

*Analysis of Income type versus Loan Credit Amount*



NAME_INCOME_TYPE vs AMT_CREDIT

**Findings:** Here we can observe that most of the clients that are on maternity leave have the highest credit amount followed by Unemployed and State servant clients

*Analysis of Housing status versus Loan Annuity*

NAME_INCOME_TYPE vs AMT_CREDIT

**Findings:** Here we can observe that most of the clients that are on maternity leave have the highest credit amount followed by Unemployed and State servant clients

*Analysis of Days employed versus Loan Credit Amount*



DAYS_BIRTH vs AMT_CREDIT

**Findings:** Here we can observe that most of the clients that are old has the highest credit loan amount followed by senior citizens

# Bivariate Analysis where 'TARGET' = 0

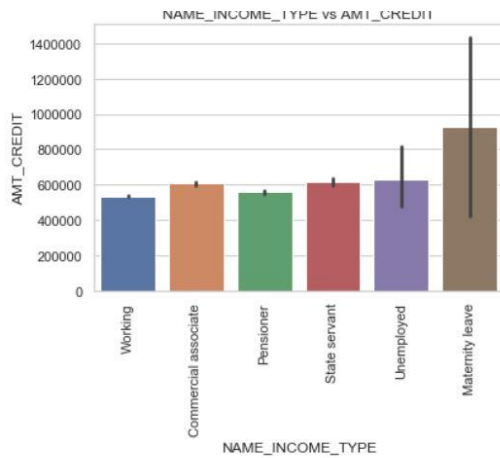*Analysis of Education Type versus Loan Annuity*

**Findings:** Here we can observe that most of the clients with Academic Degree has the highest mean amount of loan annuity followed by clients that completed Higher education
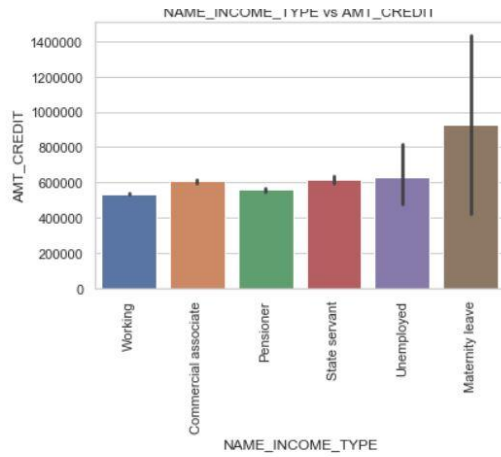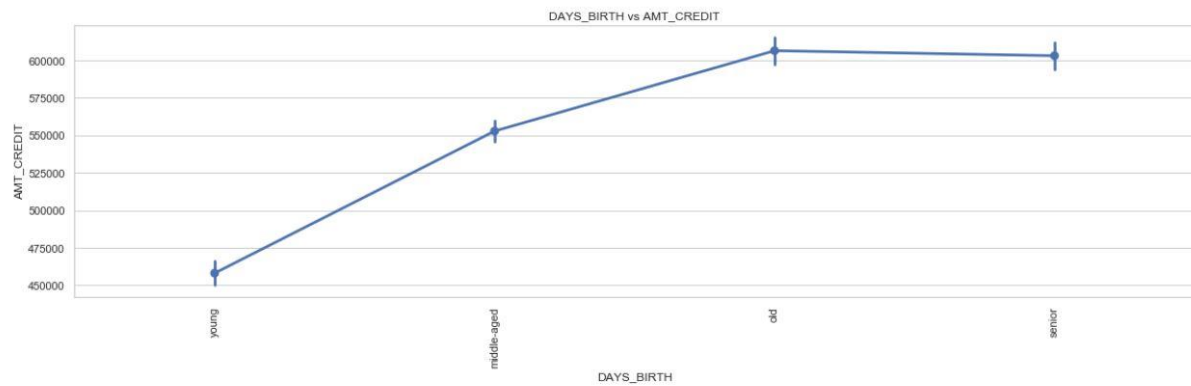
*Analysis of Family Status versus Loan Credit Amount*
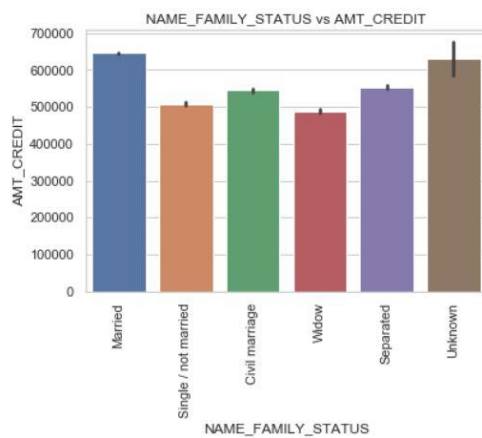


**Findings:** Here we can observe that most of the clients that are married have the highest credit amount followed by clients that are unwilling to share their family status

*Analysis of Income type versus Loan Credit Amount*

NAME_INCOME_TYPE vs AMT_CREDIT
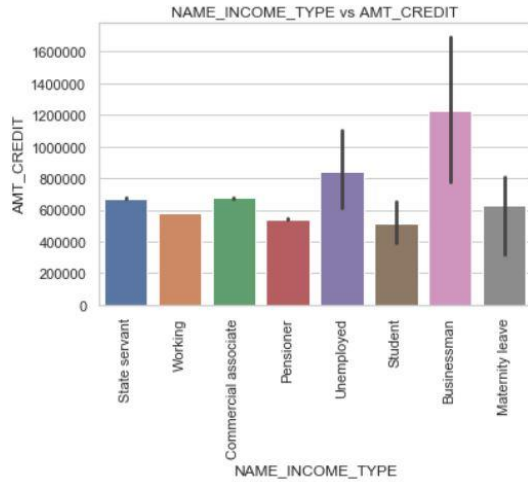
**Findings:** Here we can observe that most of the clients that are Businessmen have the highest credit amount followed by Unemployed and clients on maternity leave
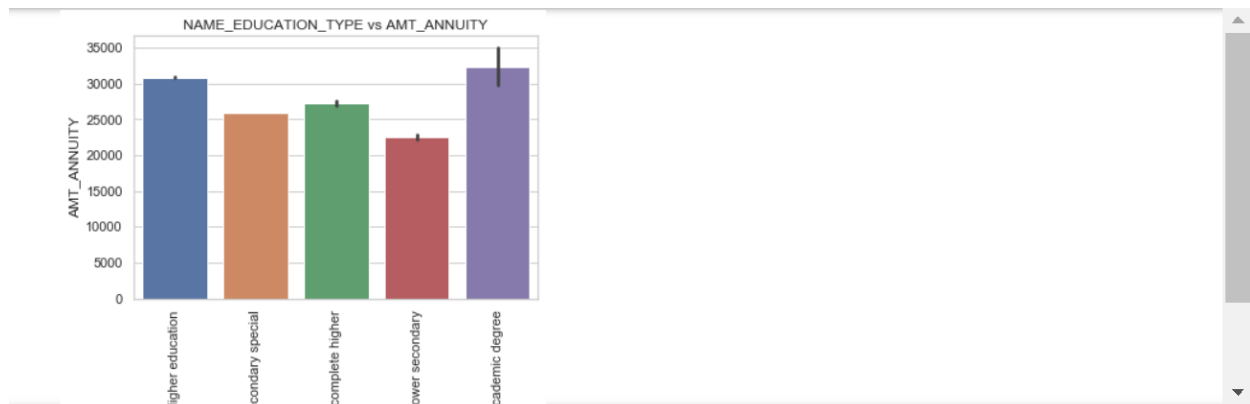
## Analysis of Housing status versus Loan Annuity
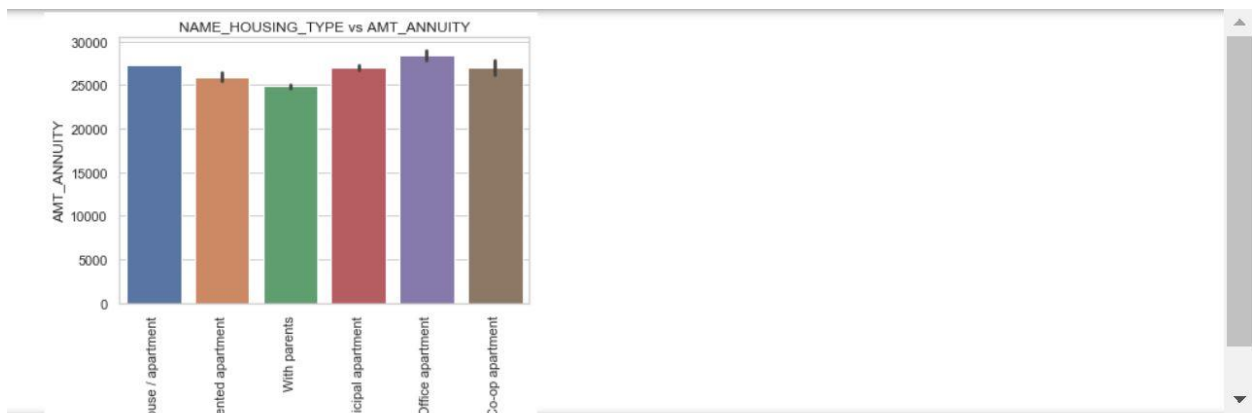


NAME_HOUSING_TYPE vs AMT_ANNUITY

**Findings:** Here we can observe that most of the clients that live in office provided apartment have the highest loaan annuity followed by Co-op apartment and Municipal apartment

## Analysis of Days employed versus Loan Credit Amount
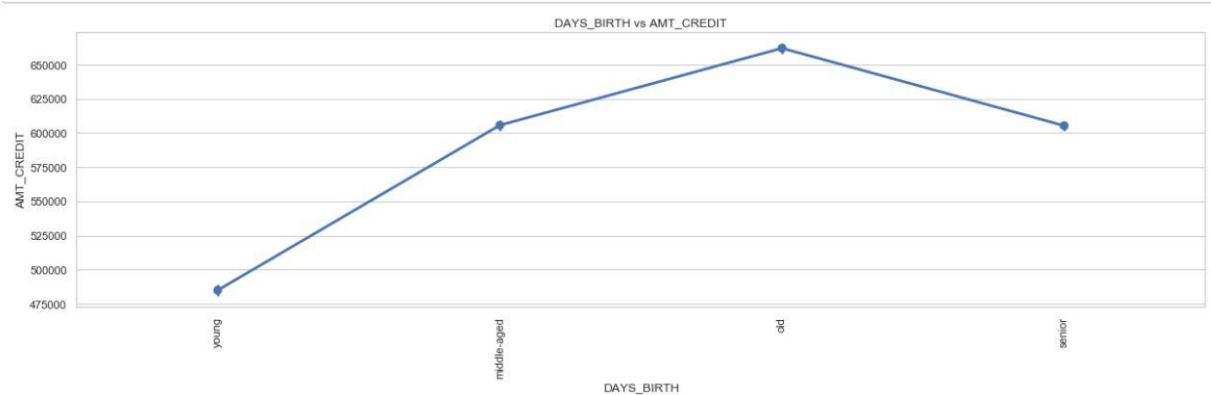
**Findings:** Here we can observe that most of the clients that are old has the highest credit loan amount followed by senior citizens

# EDA of Previous Application

Univariate Analysis on previous data where 'TARGET' = 1

*Univariate Analysis for numerical variables*

*Analysis of Credit asked for the previous application*
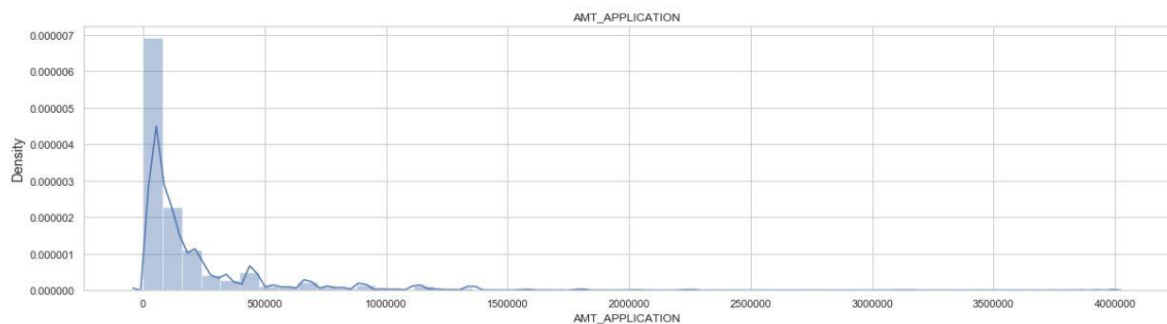


**Findings:** Here we can observe that the highest frequency of income amount for clients is between Rs. 83000 and Rs. 100000

*Analysis of Number of children for the client*

CNT_CHILDREN

**Findings:** Here we can observe that the highest frequency of children for the client is Zero followed by one

## *Analysis of Credit Amount of Loan for the previous application*
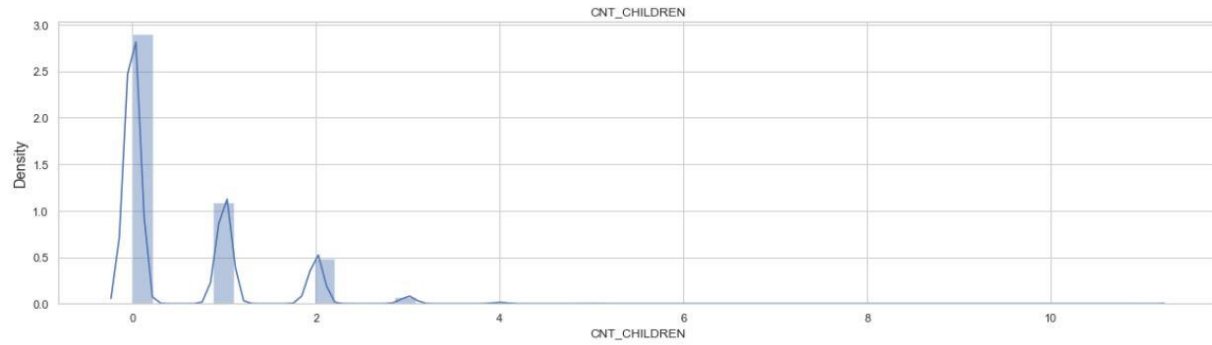


AMT_CREDIT_x

**Findings:** Here we can observe that the highest frequency density of credit is approximately Rs 70000 - Rs 140000

## *Analysis of Income of the previous application*



AMT_INCOME_TOTAL

**Findings:** Here we can observe that the highest frequency of income for the previous application is between Rs.140000 and Rs.160000

*Analysis of Goods price of previous application*
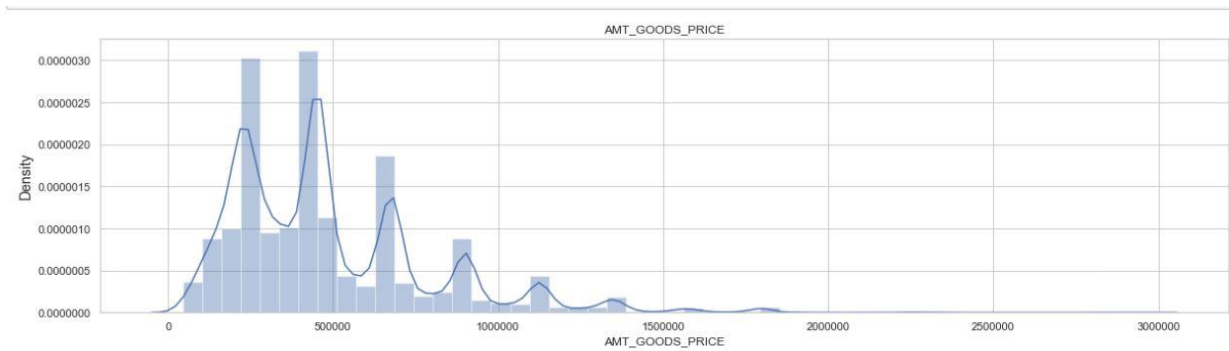


AMT_GOODS_PRICE

**Findings:** Here we can observe that highest frequency of Goods price is between Rs.390000 and Rs. 450000

*Univariate Analysis for categorical variables*



NAME_INCOME_TYPE

**Findings:** Here we can observe that most of the clients are working followed by Commercial associates

*Analysis of Client Education Level*

**Findings:** Here we can observe that most of the clients completed Secondary / Secondary special followed by Higher Education

*Analysis of Client Housing Type*



**Findings:** Here we can observe that most of the clients stay in House/Apartments of their own followed by that of clients living with their parents

*Analysis of Client Family Status*

NAME_FAMILY_STATUS

**Findings:** Here we can observe that most of the clients are married followed by single/not married status and those who followed Civil marriage

*Analysis of Client Accompaniment status*



NAME_TYPE_SUITE

**Findings:** Here we can observe that most of the clients are unaccompanied when applying for loan followed by accompaniment by family

## Univariate Analysis on previous data where 'TARGET' = 0

*Univariate Analysis for numerical variables*

*Analysis of Credit asked for the previous application*
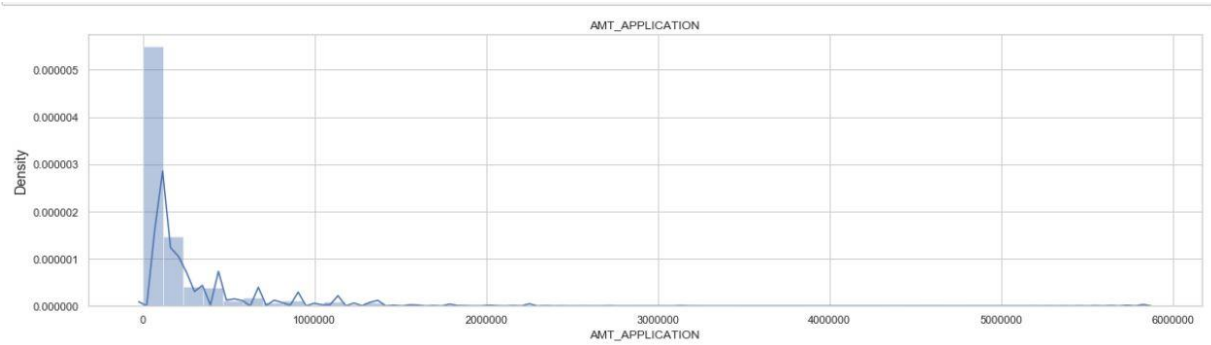
**Findings:** Here we can observe that the highest frequency of income amount for clients is between Rs. 111000 and Rs. 200000

## Analysis of Number of children for the client
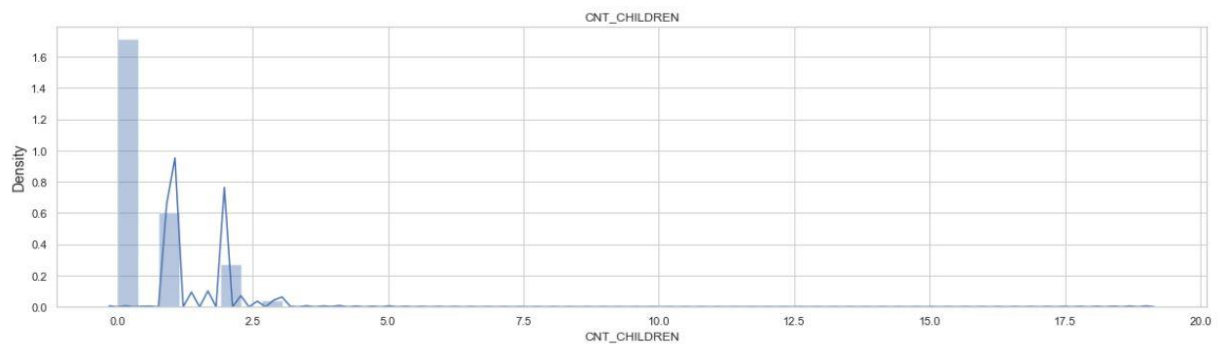


**Findings:** Here we can observe that the highest frequency of children for the client is Zero followed by one

## Analysis of Credit Amount of Loan for the previous application



**Findings:** Here we can observe that the highest frequency density of credit is approximately Rs 70000 - Rs 140000

*Analysis of Income of the previous application*
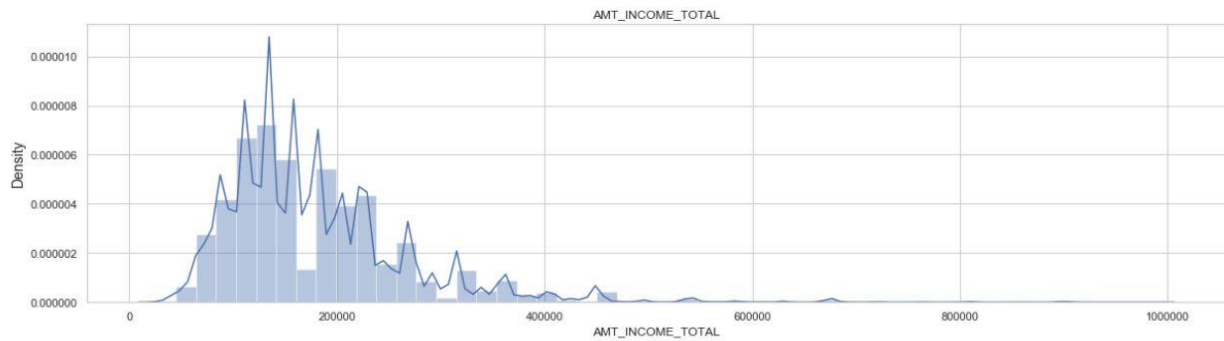


**Findings:** Here we can observe that the highest frequency of income for the previous application is between Rs.140000 and Rs.160000

*Analysis of Goods price of previous application*
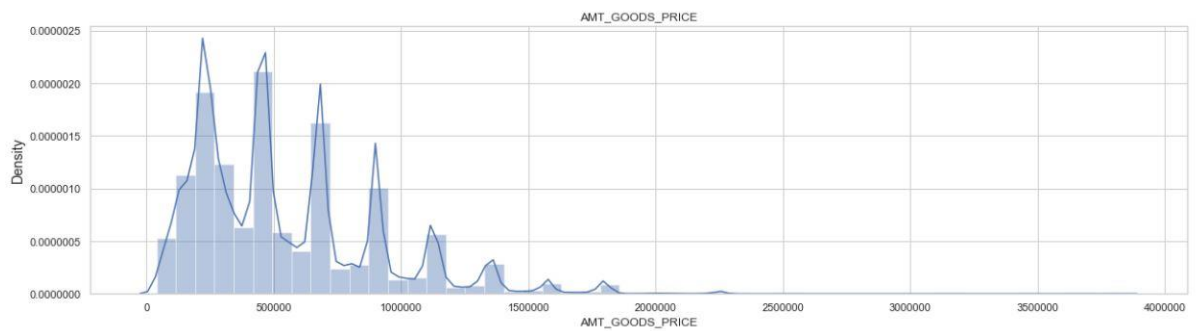


**Findings:** Here we can observe that highest frequency of Goods price is between Rs.450000 and Rs.500000

*Univariate Analysis for categorical variables*

*Analysis of Client Income Type*

NAME_INCOME_TYPE

**Findings:** Here we can observe that most of the clients are working followed by Commercial associates and pensioners

*Analysis of Client Education Level*



NAME_EDUCATION_TYPE

**Findings:** Here we can observe that most of the clients completed Secondary / Secondary special followed by Higher Education

*Analysis of Client Housing Type*

NAME_HOUSING_TYPE

**Findings:** Here we can observe that most of the clients stay in House/Apartments of their own followed by that of clients living with their parents

## Analysis of Client Family Status



NAME_FAMILY_STATUS

**Findings:** Here we can observe that most of the clients are married followed by single/not married status and those who followed Civil marriage

## Analysis of Client Accompaniment status

**Findings:** Here we can observe that most of the clients are unaccompanied when applying for loan followed by accompaniment by family

## Bivariate Analysis for previous application where 'TARGET' = 1

*Analysis of Education Type versus Loan Annunity for previous application*
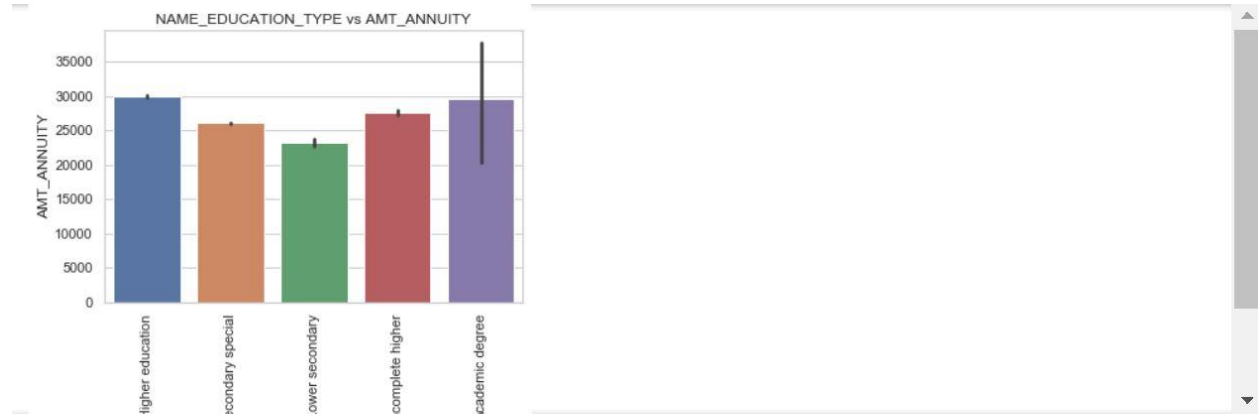


**Findings:** Here we can observe that most of the clients with Academic Degree has the highest mean amount of loan annuity followed by clients that completed Higher education

*Analysis of Family Status versus Loan Credit Amount for previous application*

**Findings:** Here we can observe that most of the clients that are married have the highest credit amount followed by Separated and Civil marriage

## Analysis of Income type versus Loan Credit Amount



**Findings:** Here we can observe that most of the clients that are on maternity leave have the highest credit amount followed by Unemployed and State servant clients

## Analysis of Housing status versus Loan Annuity

NAME_HOUSING_TYPE vs AMT_ANNUITY

**Findings:** Here we can observe that most of the clients that live in office provided apartment have the highest loaan annuity followed by Co-op apartment and Rented apartment

*Analysis of Days employed versus Loan Credit Amount*



DAYS_BIRTH vs AMT_CREDIT_x

**Findings:** Here we can observe that most of the clients that are Senior citizens has the highest credit loan amount followed by people of old category

## Bivariate Analysis for previous application where 'TARGET' = 0

*Analysis of Education Type versus Loan Annuity*

Findings: Here we can observe that most of the clients with Academic Degree has the highest mean amount of loan annuity followed by clients that completed Higher education

*Analysis of Family Status versus Loan Credit Amount*
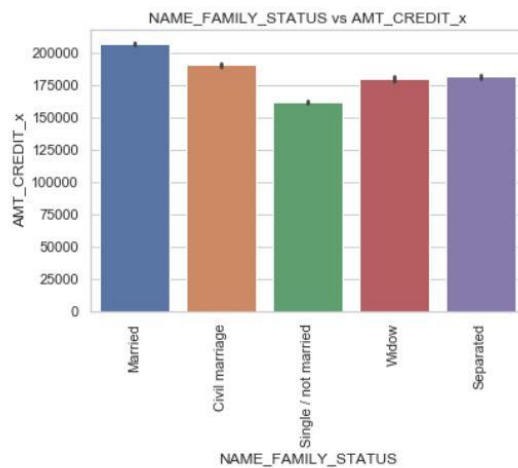


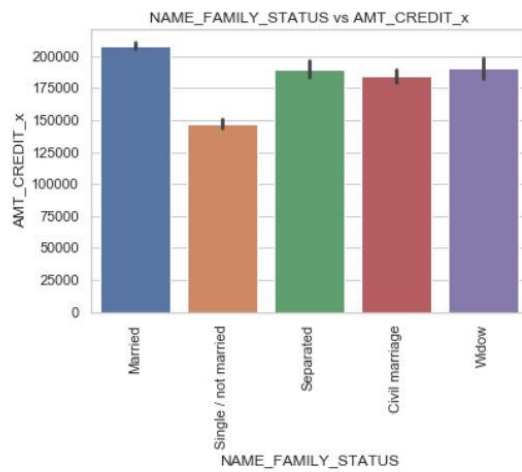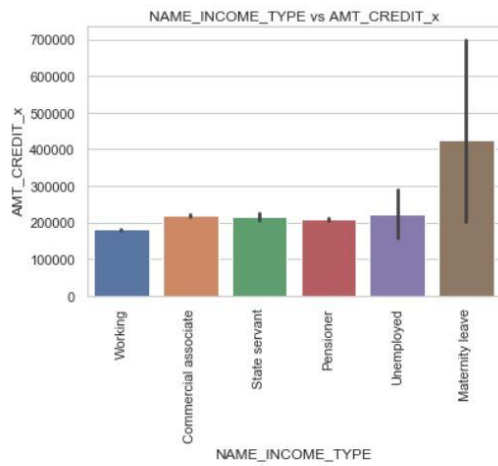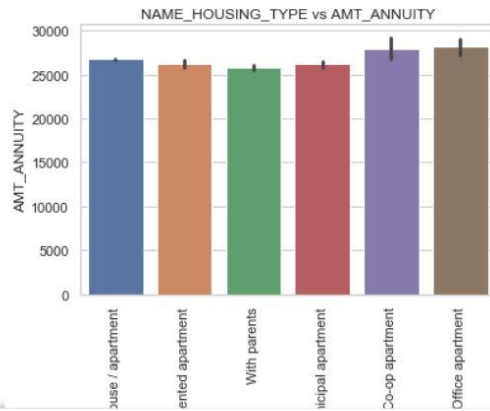Findings: Here we can observe that most of the clients that are married have the highest credit amount followed by Separated and widowed clients

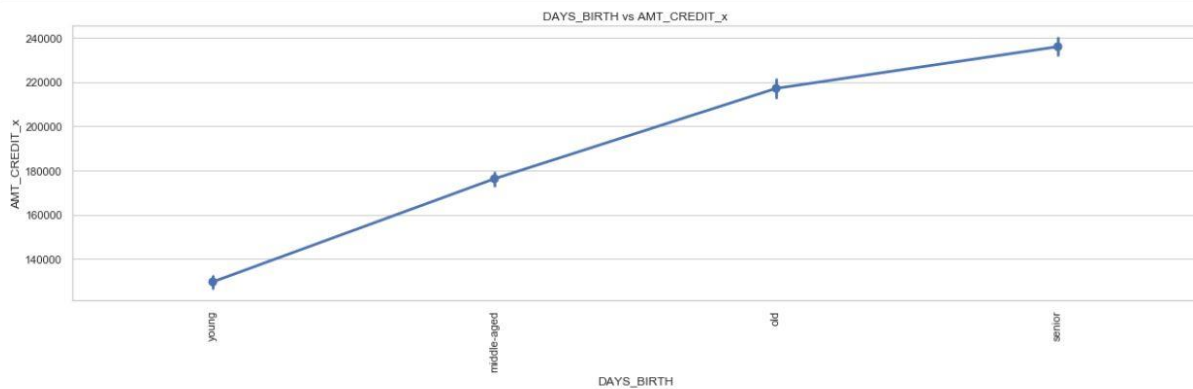*Analysis of Income type versus Loan Credit Amount*

NAME_INCOME_TYPE vs AMT_CREDIT_x
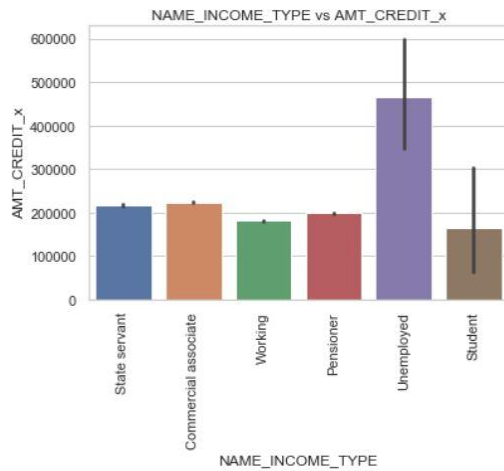
**Findings:** Here we can observe that most of the clients that are unemployed have the highest credit amount followed by Commercial associates and State servant clients
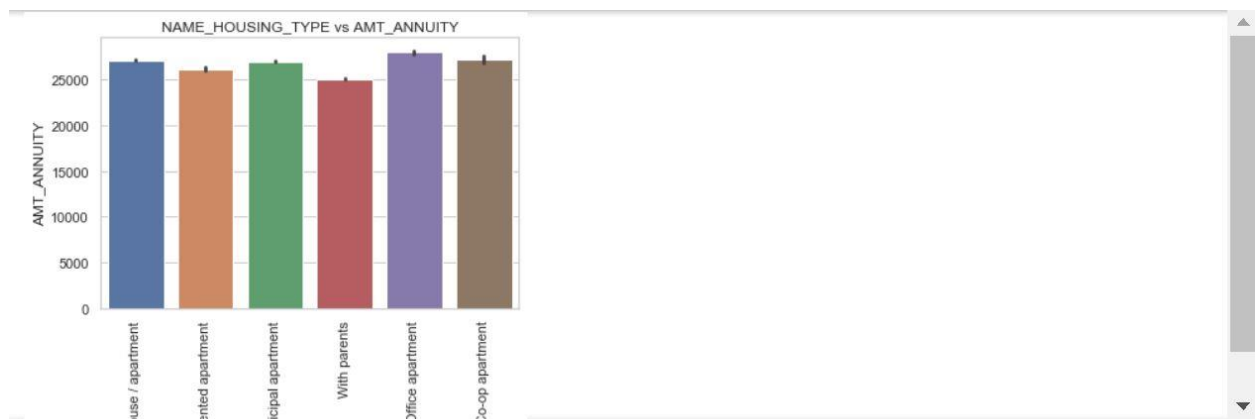
## Analysis of Housing status versus Loan Annuity



NAME_HOUSING_TYPE vs AMT_ANNUITY

**Findings:** Here we can observe that most of the clients that live in office provided apartment have the highest loan annuity followed by Co-op apartment and Rented apartment. This is fairly distributed among all the sectors
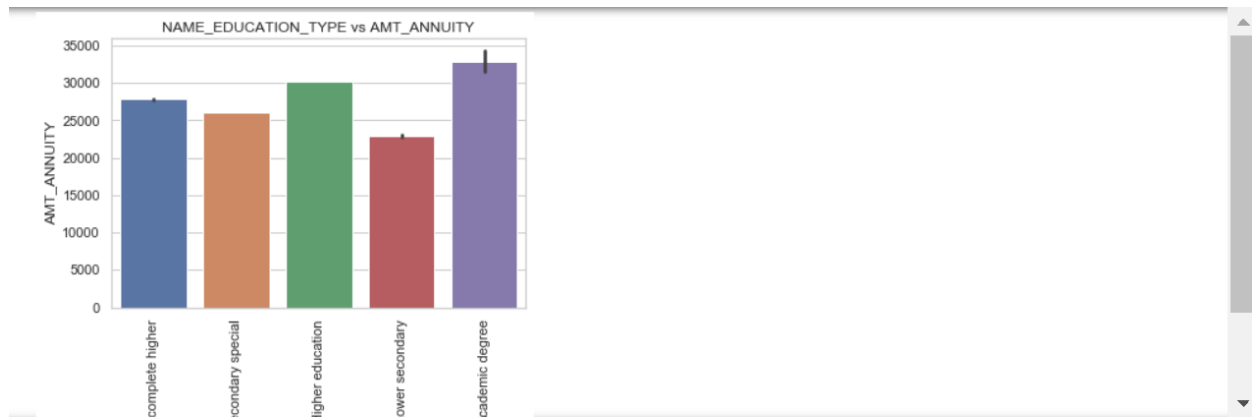
## Analysis of Days employed versus Loan Credit Amount
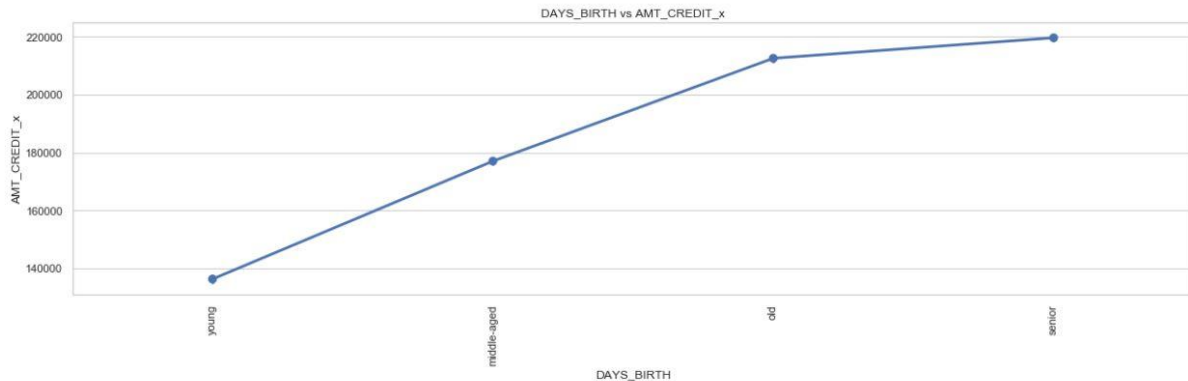
DAYS_BIRTH vs AMT_CREDIT_x

**Findings:** Here we can observe that most of the clients that are Senior citizens has the highest credit loan amount followed by people of old category

# Inferences From Exploratory Data Analysis

**1.** Overwhelming majority of credit applicants with risk involved are young working clients that earn 1-2 lakhs per month and have experience of 1.5 to 2 years, risk analysis should target this age group

**2.** Majority of the applicants request for Credit between 5-6 lakhs for goods prices worth approximately 5 lakhs. These applicants are likely to default

**3.** Most of the applicants that leads to credit default have loan annuity between Rs.25000-30000.

**4.** Defaulters are applicants that completely higher secondary education, own a house and are married. We can conclude that experience of owning capital can still lead to loan defaults

**5.** Applicants that are unaccompanied during application process overwhelming shows the possibility of defaulting. Risk analysis should be more focused among this group

**6.** Applicants that are likely to pay the loan amount earn comparatively the same amount as those who default but these applicants are likely to have more job experience

**7.** Applicants that pay the loans on time apply for lesser credit loan between Rs.200000 - Rs 350000 compared to the defaulters

**8.** Credit loan amount of applicants in other cases are more inline with the price of the goods than that of defaulters

**9.** Majority of applicants in non- defaulters group are educated, married and owns a house or an apartment

**10.** Among the applicants having difficulty paying the loan amounts, a significant portion of those on maternity leaves have high loan annuity amounts

**11.** Applicants that complete an academic degree and high annuity amount are more likely to default. This maybe due to pre-existing loans that the applicant may have paid off in the past.

**12.** Senior Citizens and applicants of old-age group have higher loan amount and are more likely to default

**13.** Married applicants generally have the highest load credit amount and are more likely to belong to non-defaulters group

**14.** A large number of applicants of young age group have lower credit amount and show promising signs of being a non-defaulter. By targeting these sectors through marketing, higher revenue can be generated

**15.** Comparing the previous application and current application data, most of the applicants apply for higher amount of loan in the current period and thus more likely to default

**16.** There is no significant jump in the salary of defaulters between the current and previous application although they submit for relatively higher amount of loan. Applicants that apply for more loan without salary jump is a red flag

**17.** Applicants with one or more children are more likely to pay the loan than applicants with no children. Defaulters also tend to not report the number of family members

**18.** Comparing previous and current application, we can infer that the unemployed customers that recently procure a job are more likely to pay the loans on time

**19.** Applicants that move from Rented apartments to their own house/apartments between applications are more likely to be a non-defaulter

**20.** There is no different in trend for defaulters that belong to senior or old age group between current and previous applications

# Conclusion

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. We aim to understand the driving factors behind loan defaults for different types of applicants.
Analysis was conducted on the prospective clients based on the their current as well as previous application data. Recommendations and insights were provided for the same.