EDA CASE STUDY

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).\

Data Sets:

1. 'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Solution Approach

1. ***Data Cleaning:***

a.) Imported the 2 data sets “application data.csv” & “previous application.csv” in python notebook as a dataframe and named them as “App_Data” &”Prev_Data”.

b.) Removed unnecessary columns from application and previos data

c.) Rename some of the column names for better understanding.

d.) Check the percentage of missing field or values in all columns in both the data frame.

e.) Few of the columns had more than 50% values missing and were not much important from Analysis perspective so dropped the columns.

f.) Imputed the missing or NA values in the columns with respective values either 1,0,median , mode , others

g.) Calculate XNA(Not available) and XAP(Not applicable values. Drop or impute based on the percentage

h.) Calculated new colums from the existing columns. For example “Age” from “Days of Birth” column , “Income_Perc” columns from “AMT_INCOME_TOTAL” and “AMT_ANNUITY”

i.) After calculating new columns , drop existing unnecessary columns.

After cleaning the data we have left with below important columns:

```
App_Data.columns
```

```
Index(['SK_ID_CURR', 'TARGET', 'Loan_Type', 'Gender', 'Car', 'Realty',
       'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_GOODS_PRICE',
       'Income_Type', 'Education', 'Family_Status', 'House_Status',
       'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'Mobile_Reachable',
       'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT',
       'REGION_RATING_CLIENT_W_CITY', 'LIVE_CITY_NOT_WORK_CITY',
       'ORGANIZATION_TYPE', 'DAYS_LAST_PHONE_CHANGE', 'Income_perc', 'Age',
       'Count_Contact', 'Score_Ext', 'DEF_SOCIAL_CIRCLE', 'Total_Docs',
       'Total_Cr_Enq'],
      dtype='object')
```

```
Prev_App.columns
```

```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'Loan_Type', 'AMT_ANNUITY',
       'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'Loan_Status',
       'DAYS_DECISION', 'Client_Type', 'Portfolio', 'CHANNEL_TYPE',
       'Payment_Term', 'Yield_group', 'DAYS_FIRST_DRAWING', 'DAYS_FIRST_DUE',
       'DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE', 'DAYS_TERMINATION',
       'Insured_On_Approval'],
      dtype='object')
```

---

2) Data Analysis

As in the problem statement two types of Clients/Customer were given:

a.) The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample

b.) All other cases: All other cases when the payment is paid on time.

Split the "App_Data" dataframe in two dataframe - "**default_cust" (Target=1),**

and **other_cust(Target=0).**

After separating the dataframe in two part , we have left with below dataframes:

	SK_ID_CURR	TARGET	Loan_Type	Gender	Car	Realty	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	...	LIVE_CITY_I
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	351000	...	
26	100031	1	Cash loans	F	N	Y	0	112500.0	979992.0	702000	...	
40	100047	1	Cash loans	M	N	Y	0	202500.0	1193580.0	855000	...	
42	100049	1	Cash loans	F	N	N	0	135000.0	288873.0	238500	...	
81	100096	1	Cash loans	F	N	Y	0	81000.0	252000.0	252000	...	
94	100112	1	Cash loans	M	Y	Y	0	315000.0	953460.0	900000	...	
110	100130	1	Cash loans	F	N	Y	1	157500.0	723996.0	585000	...	
138	100160	1	Cash loans	M	N	Y	0	292500.0	675000.0	675000	...	
154	100181	1	Cash loans	F	N	Y	0	157500.0	245619.0	166500	...	
163	100192	1	Cash loans	F	N	N	0	111915.0	225000.0	225000	...	
180	100209	1	Revolving loans	M	N	Y	3	180000.0	540000.0	540000	...	
184	100214	1	Cash loans	F	N	Y	1	202500.0	436032.0	360000	...	

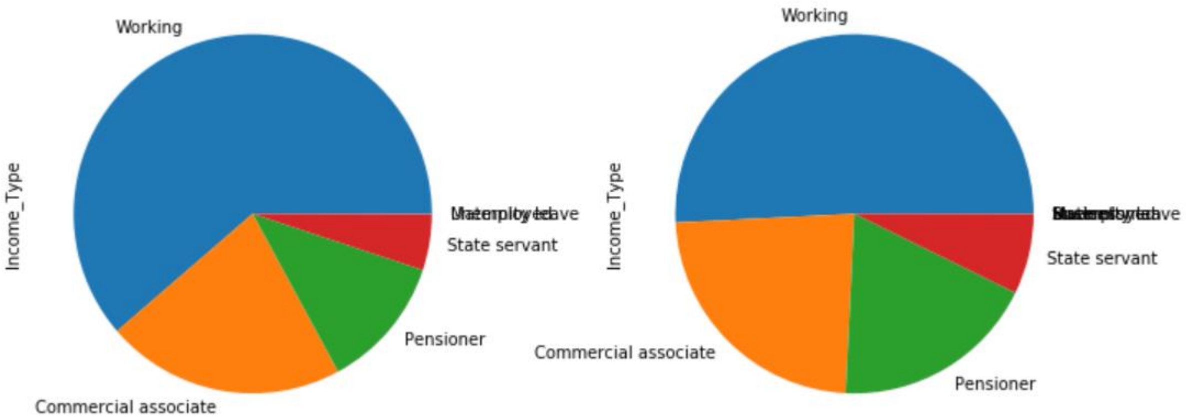
	SK_ID_CURR	TARGET	Loan_Type	Gender	Car	Realty	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	...	LIVE_CIT
1	100003	0	Cash loans	F	N	N	0	270000.000	1293502.5	1.1295e+06	...	
2	100004	0	Revolving loans	M	Y	Y	0	67500.000	135000.0	135000	...	
3	100006	0	Cash loans	F	N	Y	0	135000.000	312682.5	297000	...	
4	100007	0	Cash loans	M	N	Y	0	121500.000	513000.0	513000	...	
5	100008	0	Cash loans	M	N	Y	0	99000.000	490495.5	454500	...	
6	100009	0	Cash loans	F	Y	Y	1	171000.000	1560726.0	1.395e+06	...	
7	100010	0	Cash loans	M	Y	Y	0	360000.000	1530000.0	1.53e+06	...	
8	100011	0	Cash loans	F	N	Y	0	112500.000	1019610.0	913500	...	
9	100012	0	Revolving loans	M	N	Y	0	135000.000	405000.0	405000	...	
10	100014	0	Cash loans	F	N	Y	1	112500.000	652500.0	652500	...	

Analysis Approach

- *Compare the relevant columns from both target dataframses to identify the significant difference and analysis points*
- *Perform u**nivariate/segmented univariate** analysis on columns.*

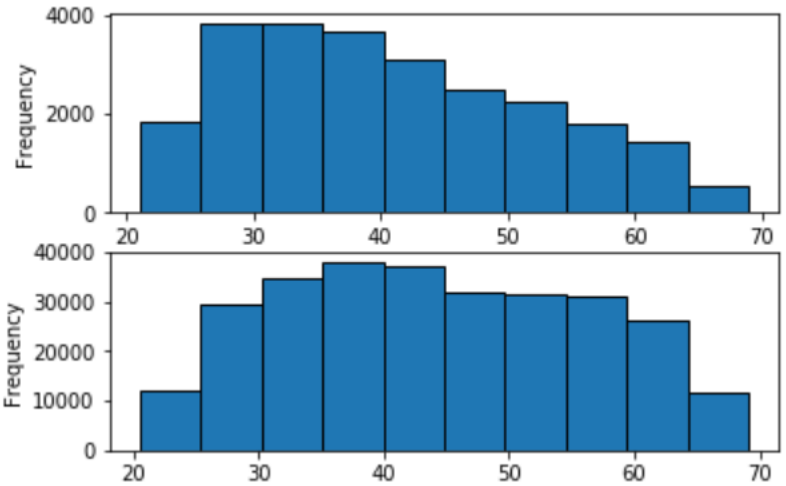
Ex-

INCOME_TYPE

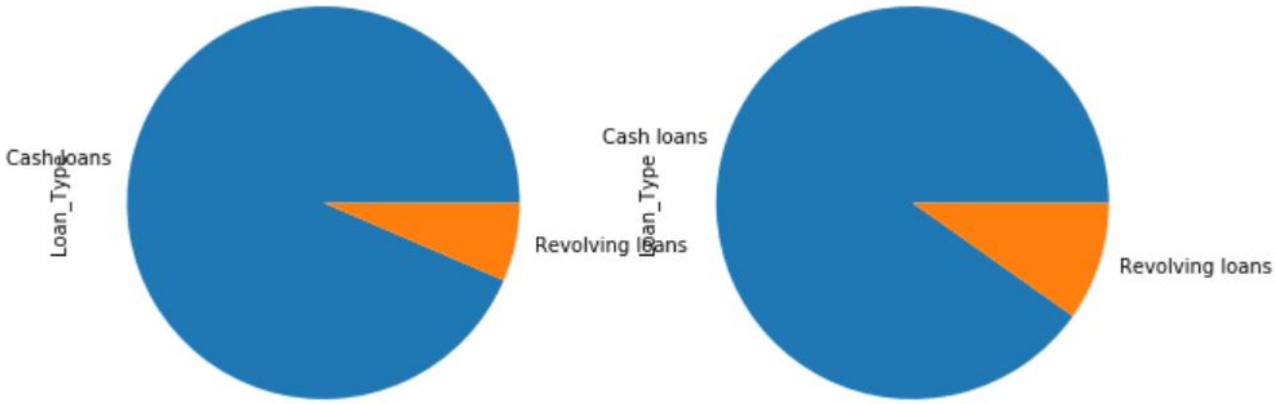


-

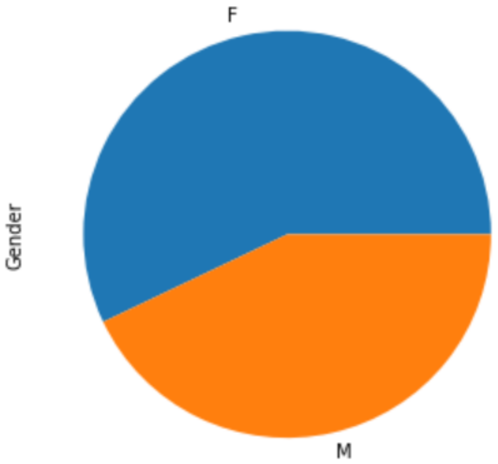
AGE

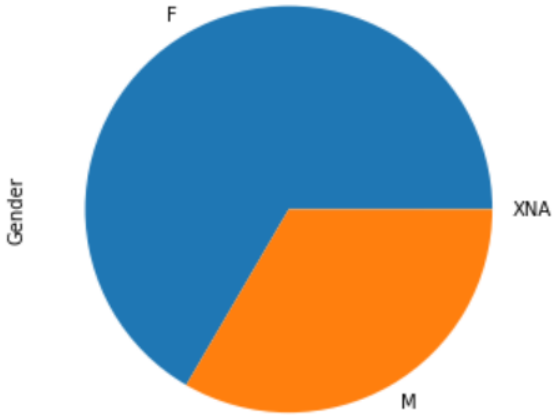


LOAN TYPE

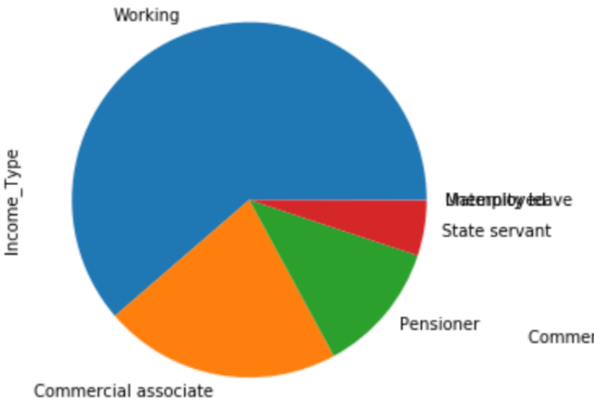


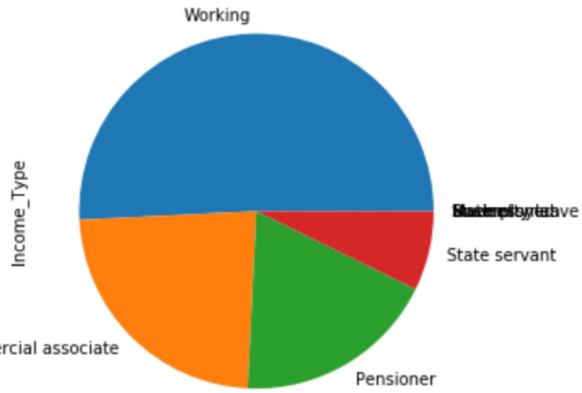
GENDER





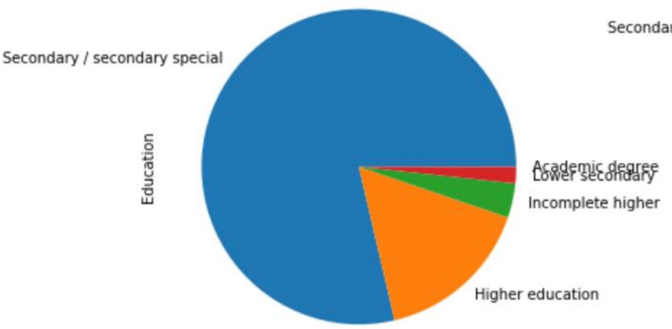
-
- ***INCOME TYPE***

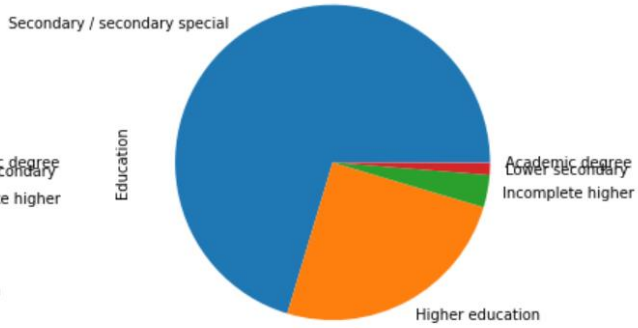




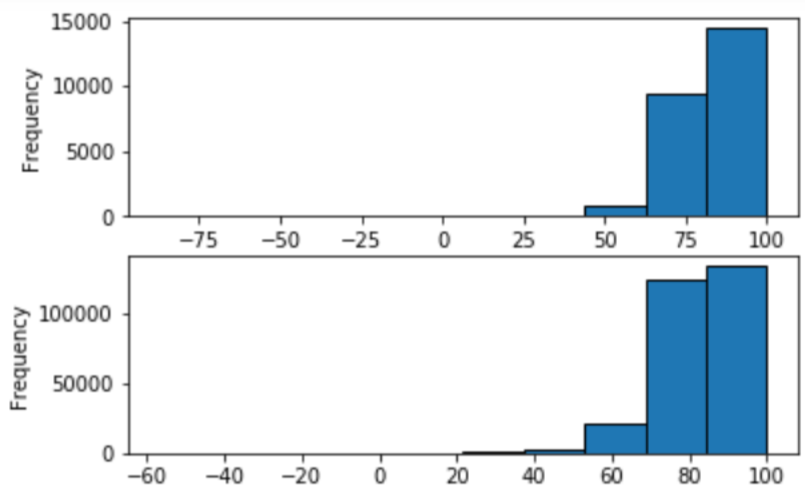
-

EDUCATION





INCOME SAVING PERCENT



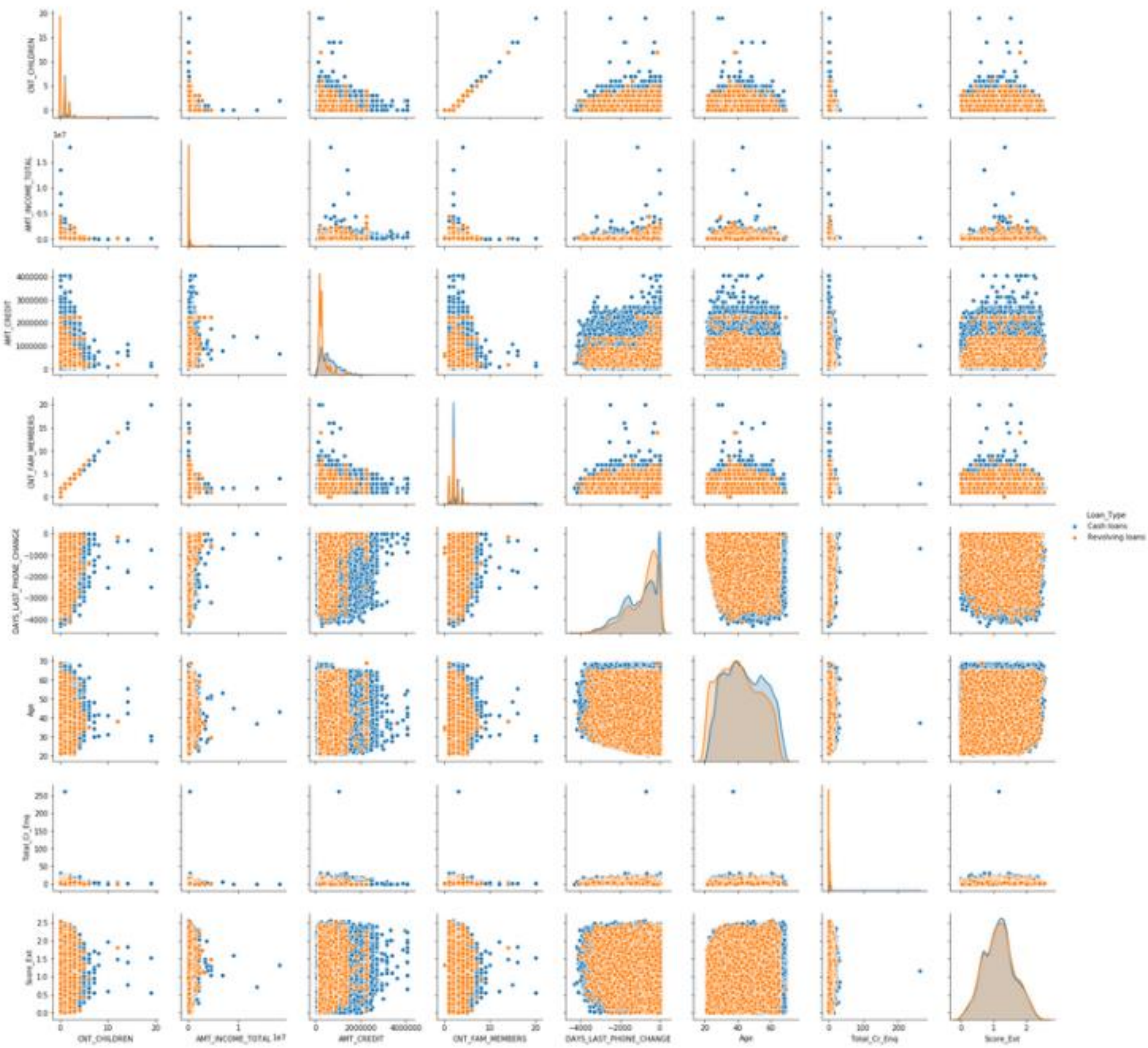
Based on univariate/segmented univariate analysis, the following are the

<u>TOP FINDINGS</u>

- ***Cash loans have more chances o default***
- ***Males are more likely to be defaulters***
- ***Working professionals are more likely to default***
- ***People with Education- Secondary/Senior Secondary Education default more***
- ***Defaulters have less income percentage as savings***
- ***Defaulters fall mostly in the age-group 25-50***

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

- *Based on the relevant columns, perform a bivariate analysis for all relevant columns –*
- ***Bivariate Analysis on Defaulters Data***

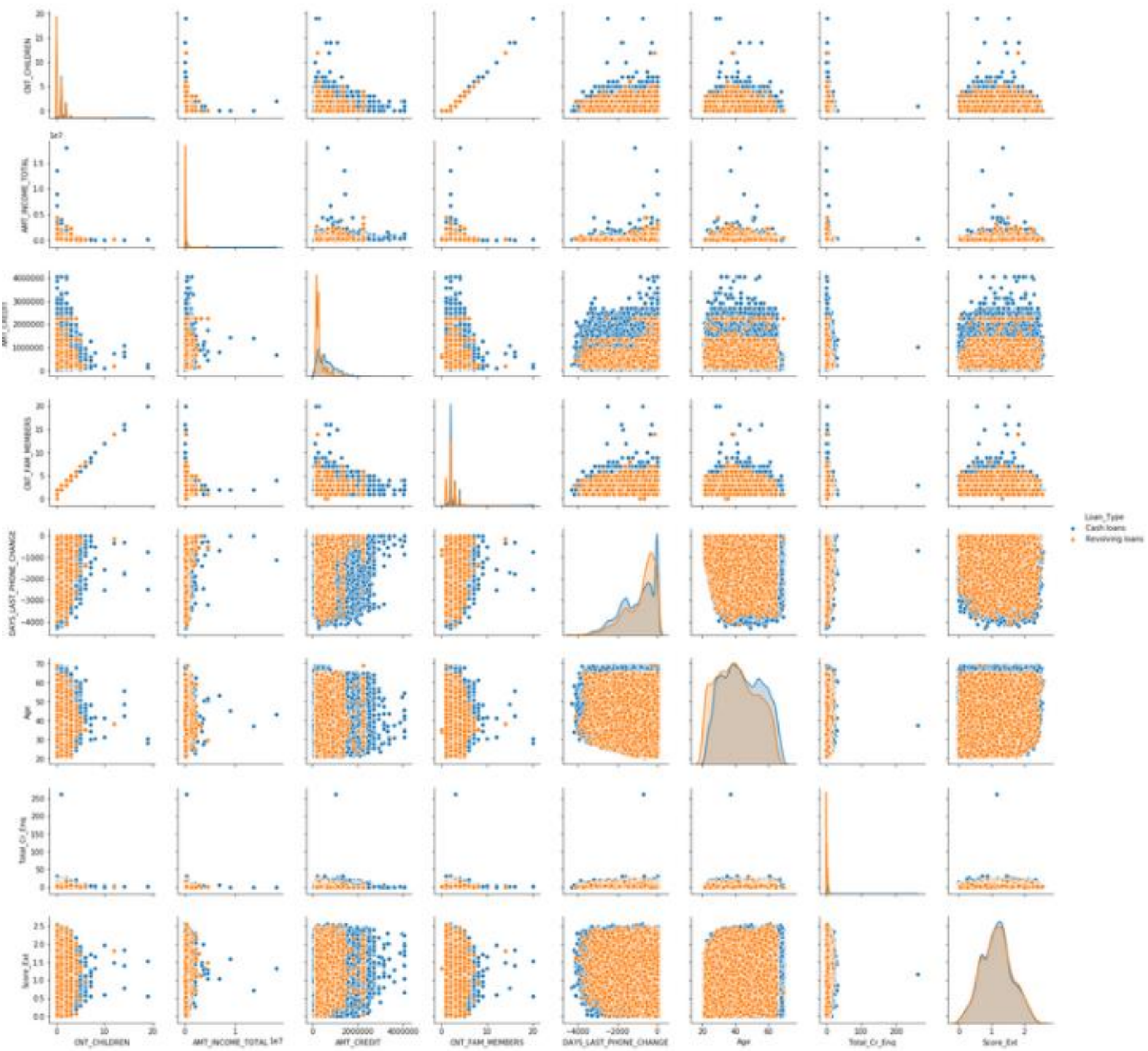


TOP 10 CORRELATION-

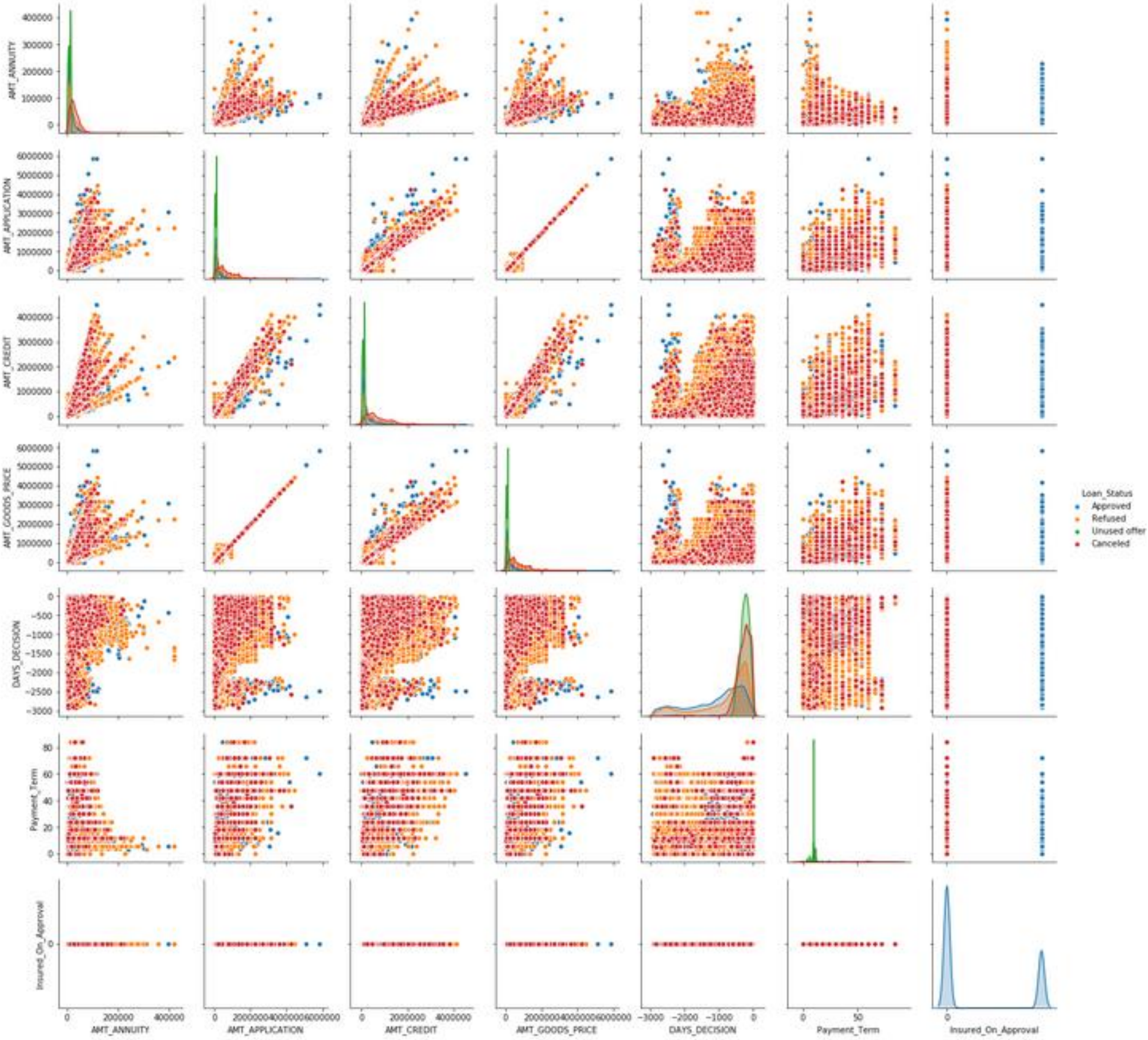
(Defaulters and Others Data)

1) *CNT_CHILDREN & CNT_FAM_MEMBERS*
2) *AMT_INCOME & DAYS_LAST_PHONE_CHANGE*
3) *AMT_CREDIT & CNT_FAM_MEMBERS*
4) *CNT_FAM_MEMBERS & CNT_FAM_MEMBERS*
5) *CNT_CHILDREN & AGE*
6) *CNT_CHILDREN & SCORE_EXT*
7) *SCORE_EXT & AMT_CREDIT*
8) *TOTAL_CR_ENQ & DAYS_LAST_PHONE_CHANGE*
9) *SCORE_EXT & AMT_CREDIT*
10) *AGE & AMT_CREDIT*

- ***Bivariate Analysis on Others Data***



- ***Bivariate Analysis on Others Data***



TOP 10 CORRELATION-

(Previous Data)

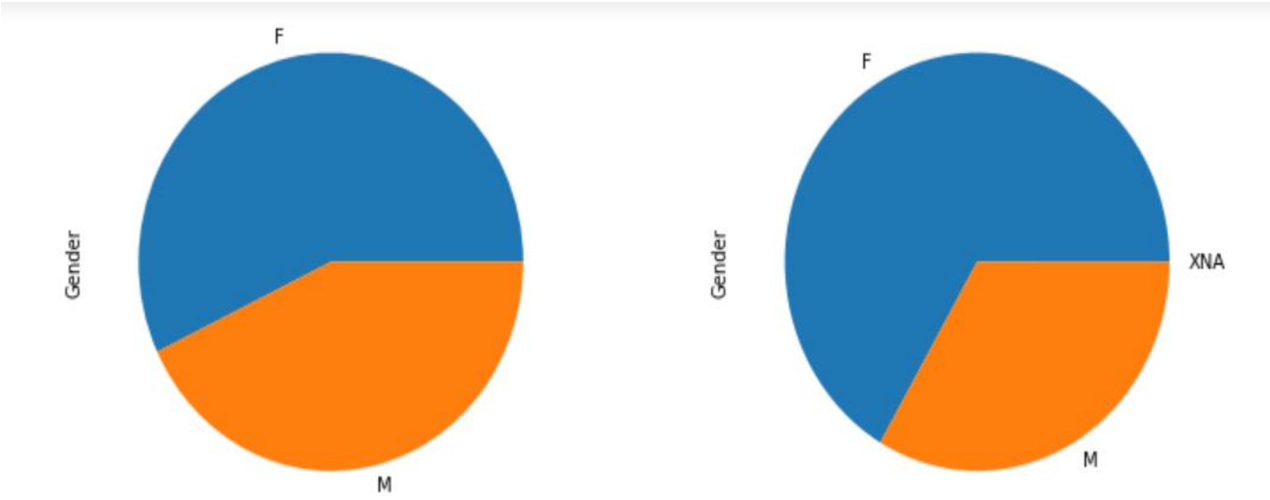
1) *AMT_ANNUITY & AMT_APPLICATION*
2) *AMT_ANNUITY & AMT_CREDIT*
3) *AMT_ANNUITY & AMT_GOODS_PRICE*
4) *AMT_APPLICATION & AMT_CREDIT*
5) *AMT_APPLICATION & AMT_GOODS_PRICE*
6) *AMT_CREDIT & AMT_GOODS_PRICE*
7) *DAYS_DECISION & AMT_APPLICATION(For Approved Loans)*
8) *DAYS_DECISION & AMT_GOODS_PRICE (For Approved Loans)*
9) *AMT_ANNUITY & DAYS_DECISION*
10) *AMT_APPLICATION & DAYS_DECISION(For Refused Loans)*

OTHER POINTS-

Univariate Analysis

Below are some of the Univariate Analysis performed in columns of both the dataframe:

a.) Gender column:

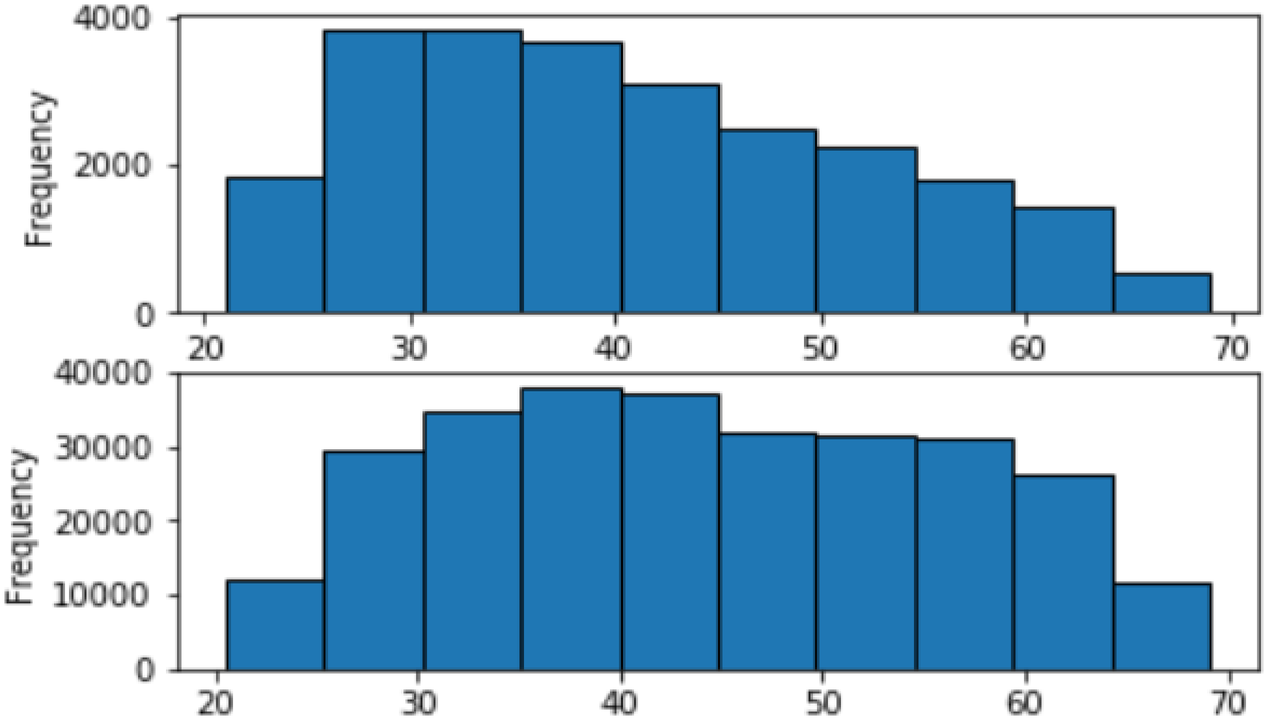


default_cust other_cust

Default_cust other_cust

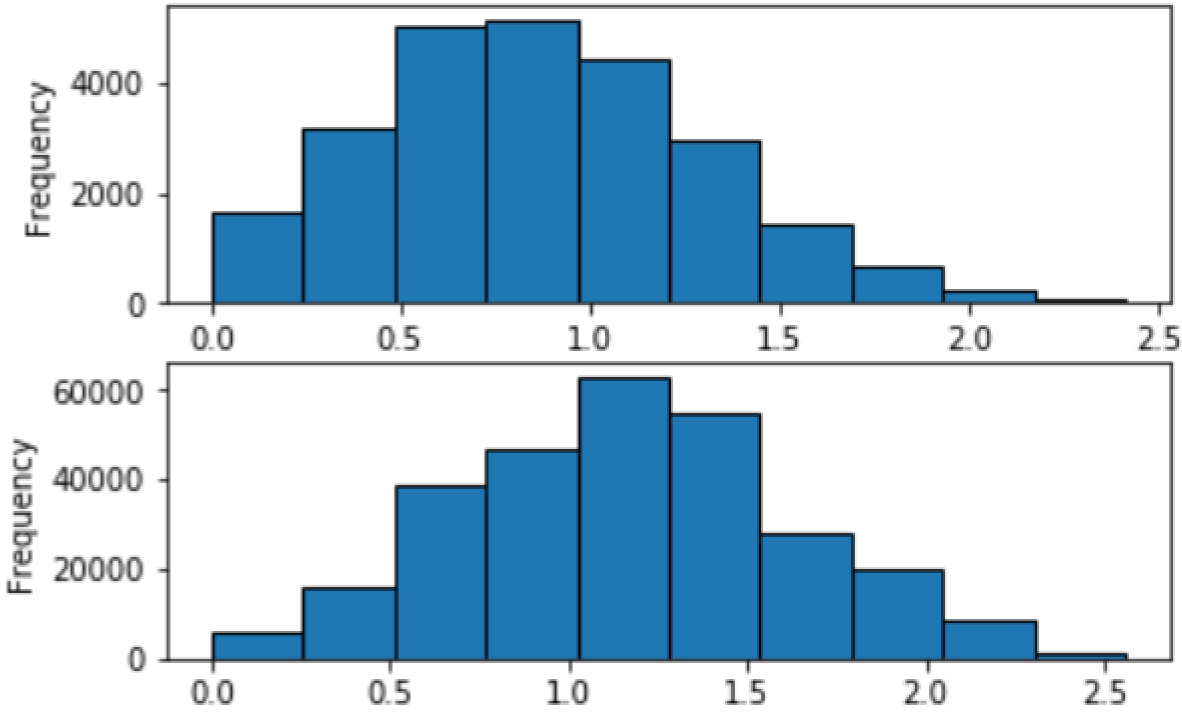
b.) Age column:

Segmented Univariate



Shows the defaulters are mostly in the age ranging from 25 – 50years

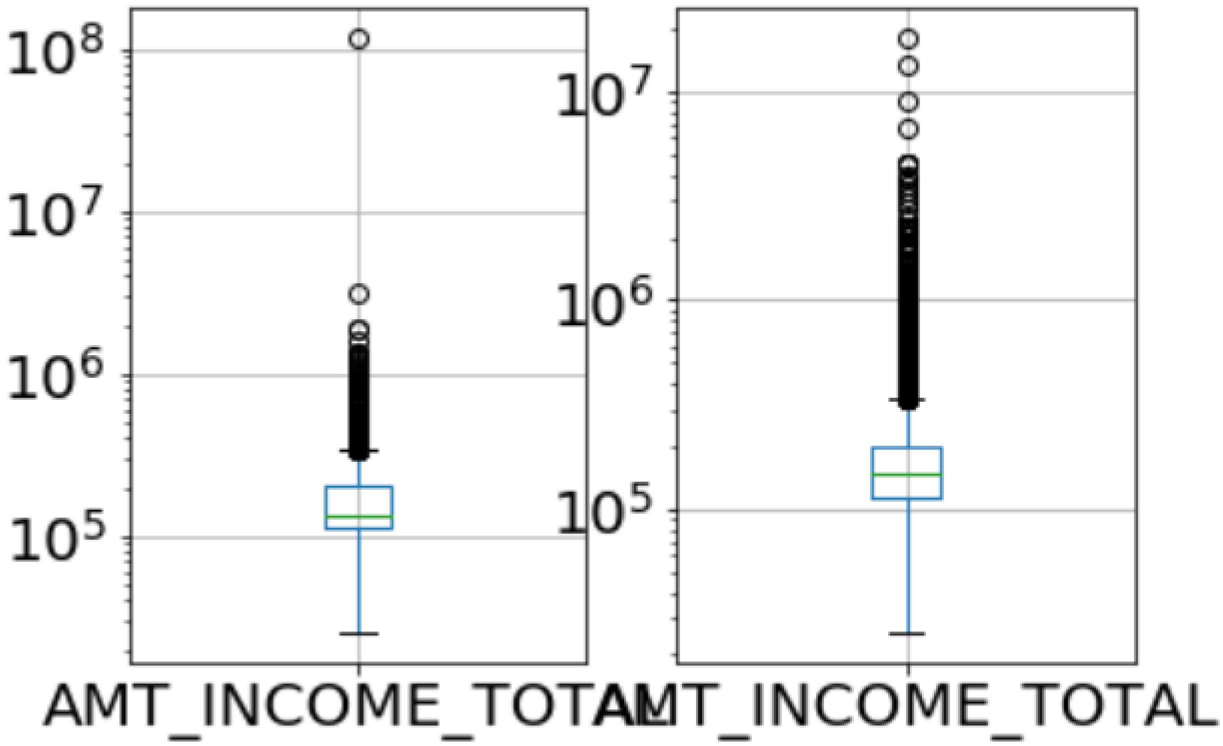
c) Score_Ext column:



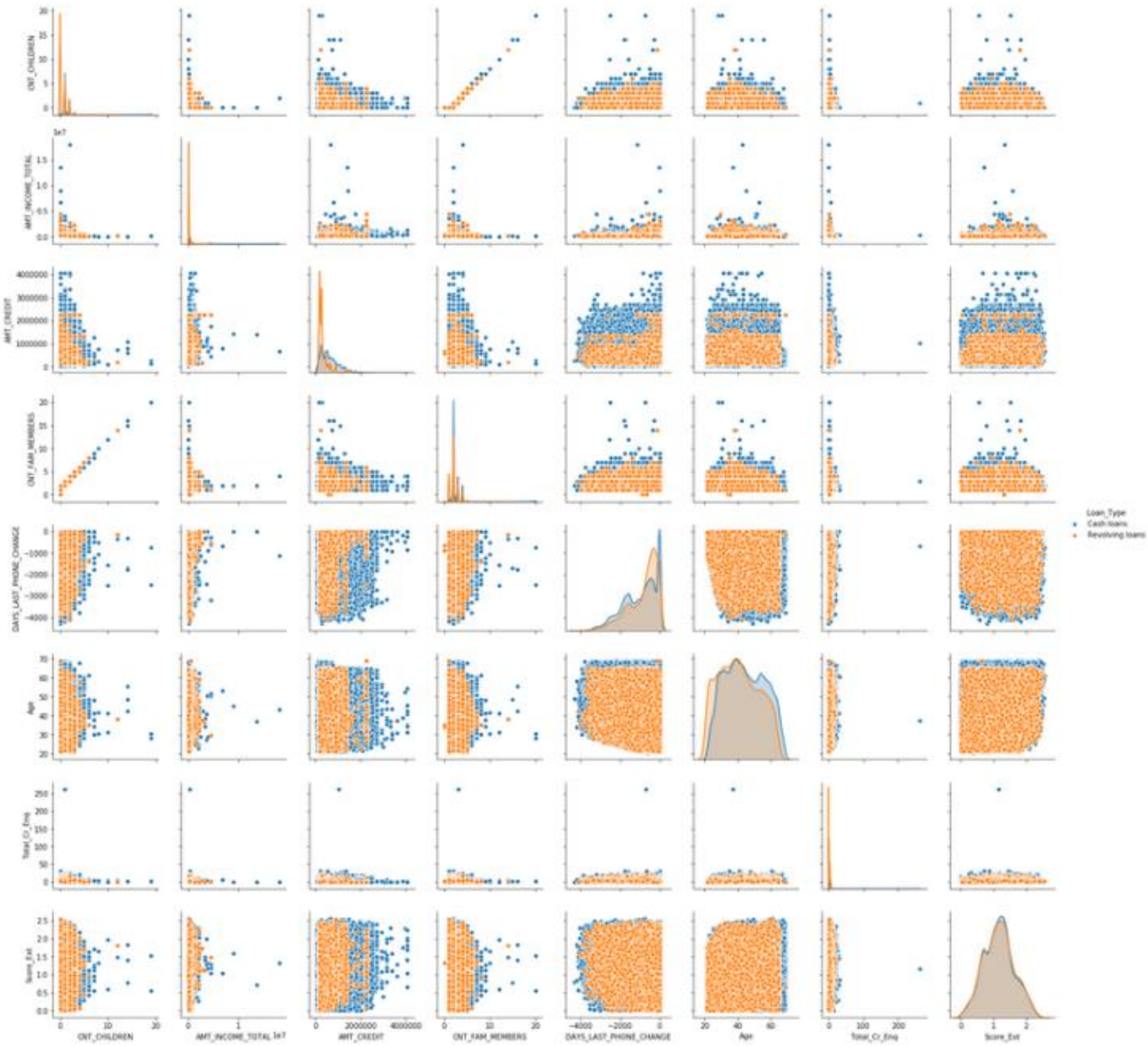
Shows most of the defaulters are having scores ranging 0.5 to 1.5 from external sources

Identifying Outliers:-

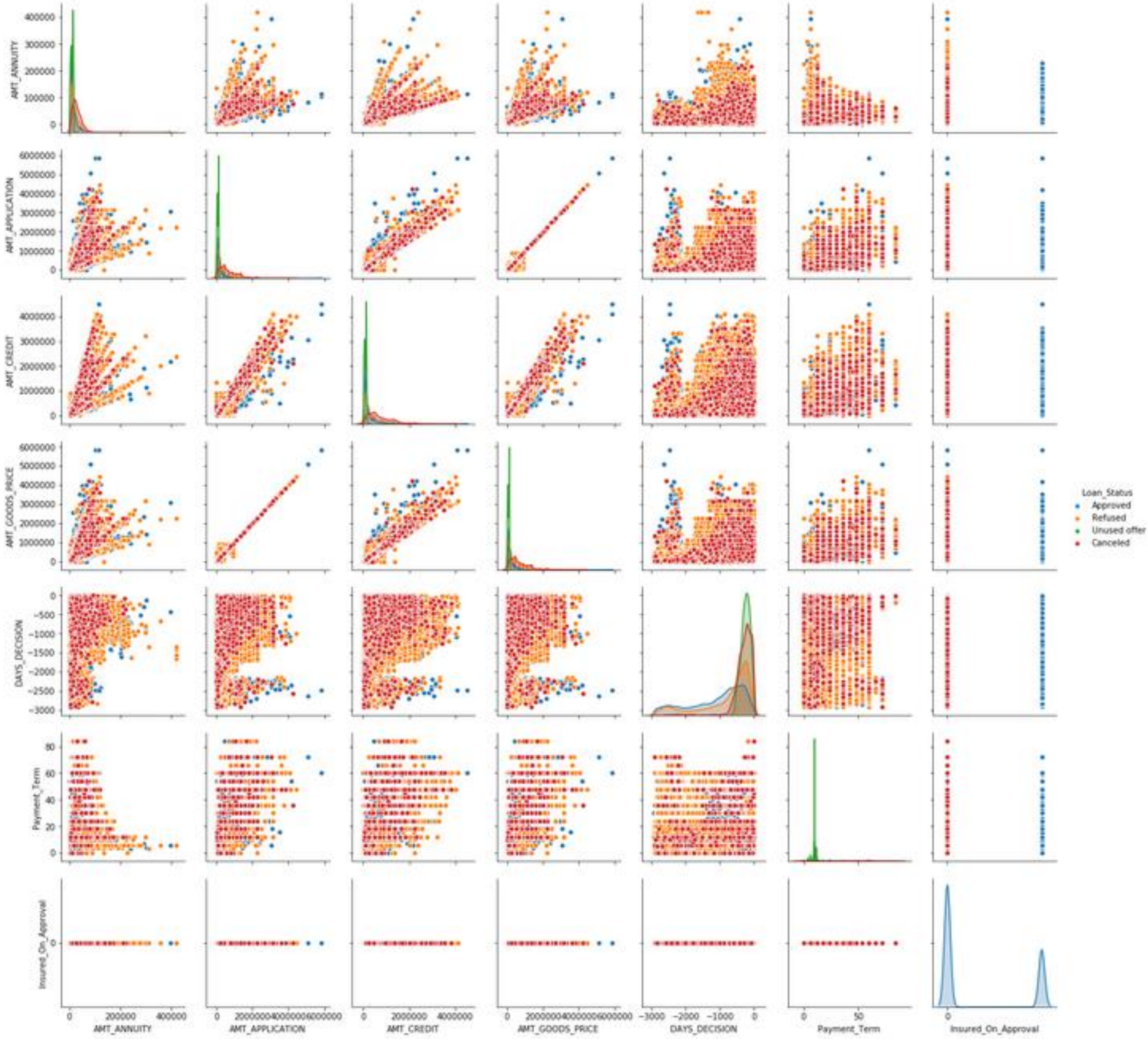
e.) AMT_INCOME_TOTAL column:



Bivariate Analysis:-


CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
CNT_FAM_MEMBERS
DAYS_LAST_PHONE_CHANGE
Age
Total_Cr_Enq
Score_Ext
Loan_Type
Cash loans
Revolving loans

--


AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_GOODS_PRICE
DAYS_DECISION
Payment_Term
Insured_On_Approval
Loan_Status
Approved
Refused
Unused offer
Canceled
