



CREDIT EDA ASSIGNMENT

By Arunabha Roy



INTRODUCTION

For credit risk analysis, what is EDA?

In the context of risk management, exploratory data analysis (EDA) refers to the systematic analysis of risk data with the goal of identifying and summarizing their key features in text-based (tabular) or visual presentations.

Here we will be encountering two data frames and merging them to extract outcomes on different levels(matrices) and giving conclusion at end .



Analysis Pathway :-

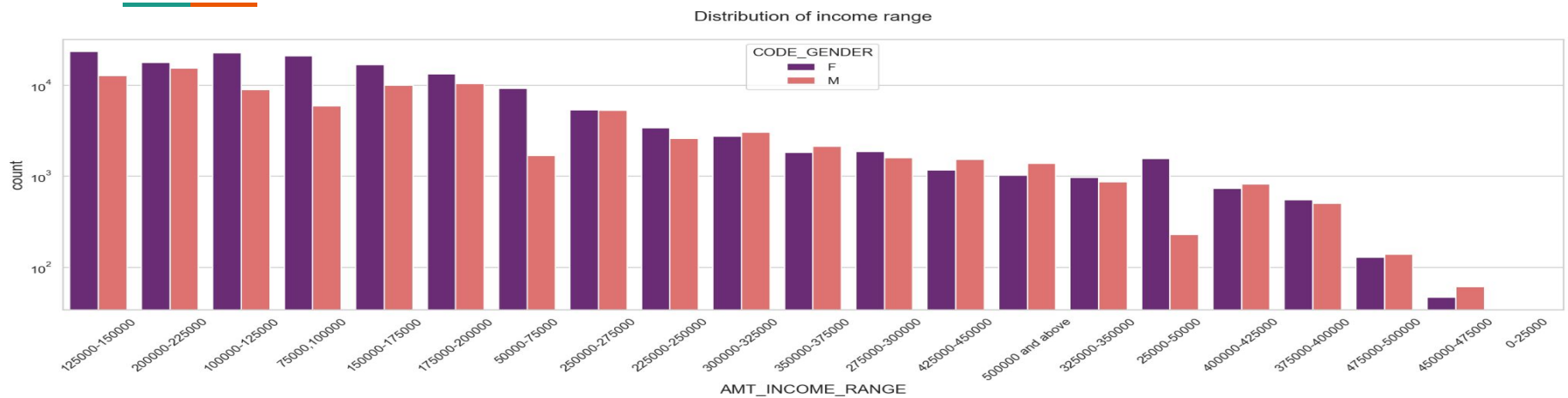
1. Understanding and Sourcing of data
2. Check for data (quality, imbalance, univariate & bivariate etc)
3. Merging the data
4. Analysing them accordingly.
5. Conclusions

In the python notebook file , the analysis started with previous application tallying with the current application data analysis.



Univariate Analysis of Continuous data in Application Data

Distribution of Income Range



Concluding remarks based on the graph above.

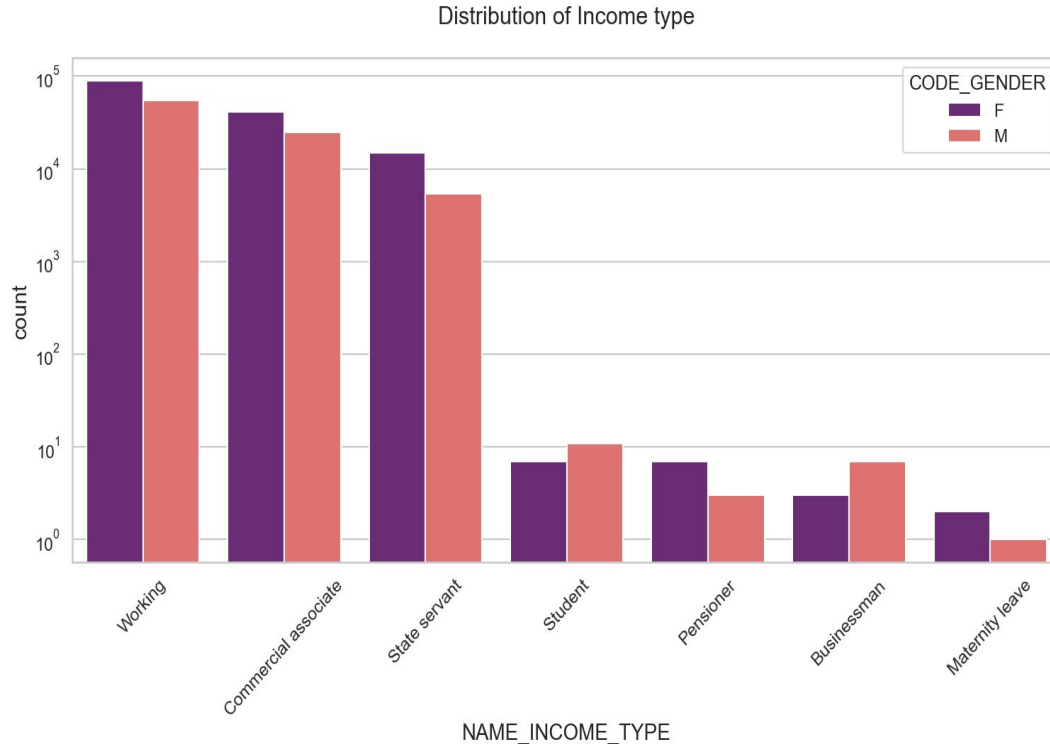
1. More women than men are counted.

2 There are additional credits available for incomes between 100,000 and 200,000.

3 According to this graph, women have more credit for that range than men do.

4 For income levels of 400000 and beyond, very few count.

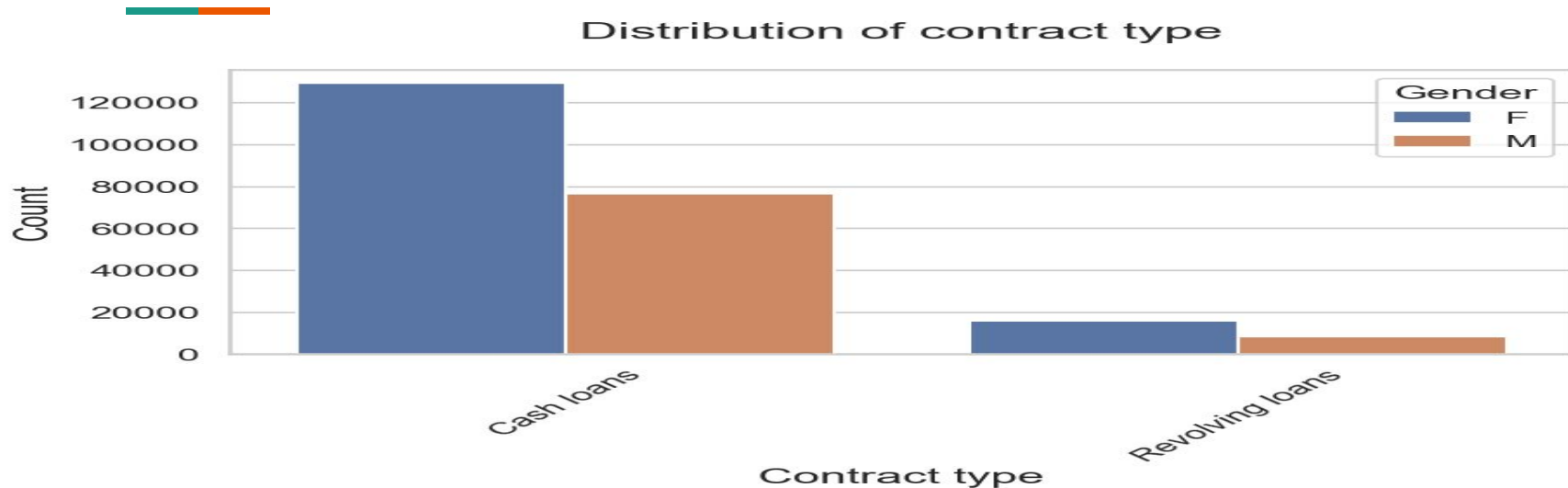
Distribution of Income Range



Based on the information provided in the graph:

1. The number of credits is larger for income types "working," "commercial associate," and "state servant."
2. Fewer credits are available for the income types "student," "pensioner," "businessman," and "maternity leave."
3. Due to this, females have more credits than males.

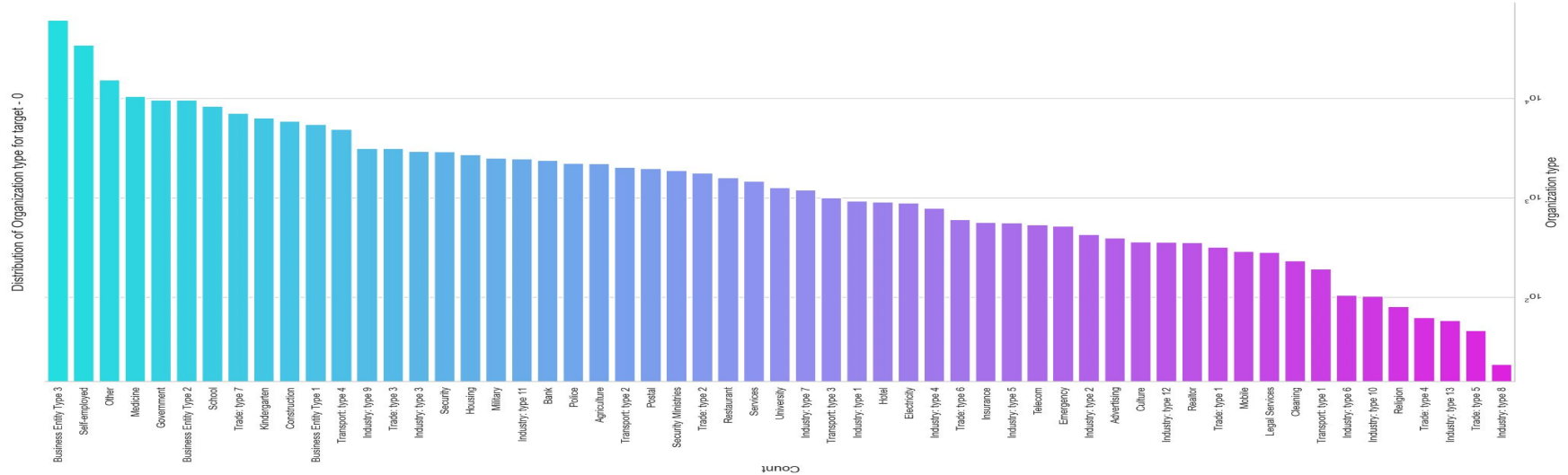
Distribution of Contract Type



Based on the information depicted in the graph:

1. Cash loans, as a contract type, exhibit a higher number of credits compared to revolving loans.
2. Additionally, it is evident from the graph that women tend to dominate in credit applications within this particular context.

Distribution Organization Wise



Based on the information depicted in the graph:

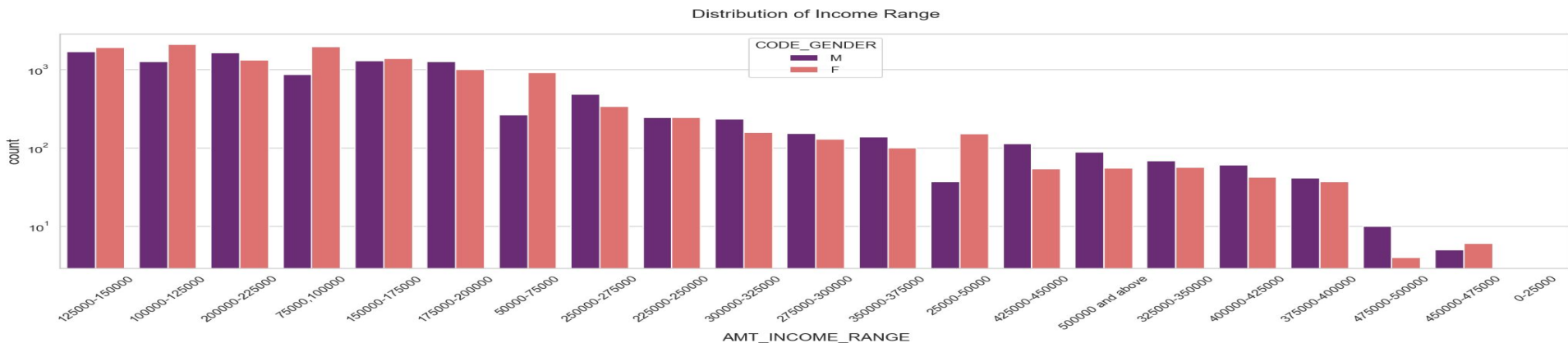
1. The majority of credit applicants belong to organization types such as "Business entity Type 3," "Self employed," "Other," "Medicine," and "Government."

2. In contrast, there is a comparatively lower number of customers associated with industry types 8, 6, 10, 5, religion and 4 in terms of credit applications.



Categorical Univariate Analysis for Target 1

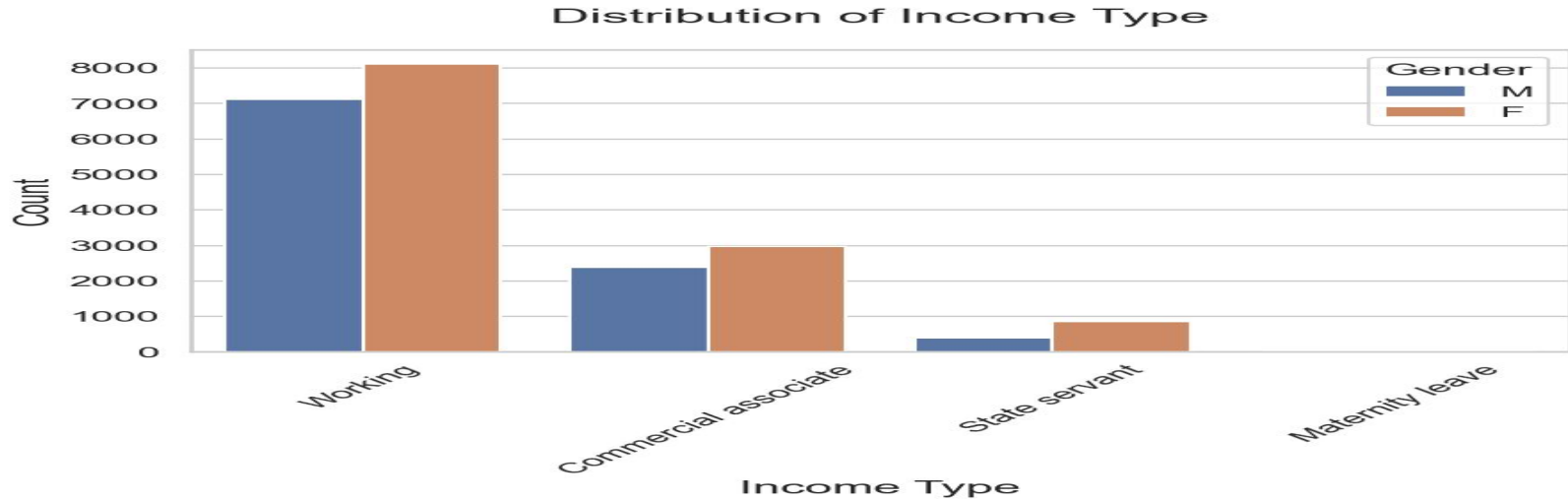
Distribution of Income Range



Based on the information depicted in the graph:

1. Male numbers outnumber female counts.(13/20)
2. There are additional credits available for incomes between 100,000 and 200,000.
3. This graph demonstrates that men have more credits in that range than women do.
4. For income levels of 400,000 and beyond, very few count.

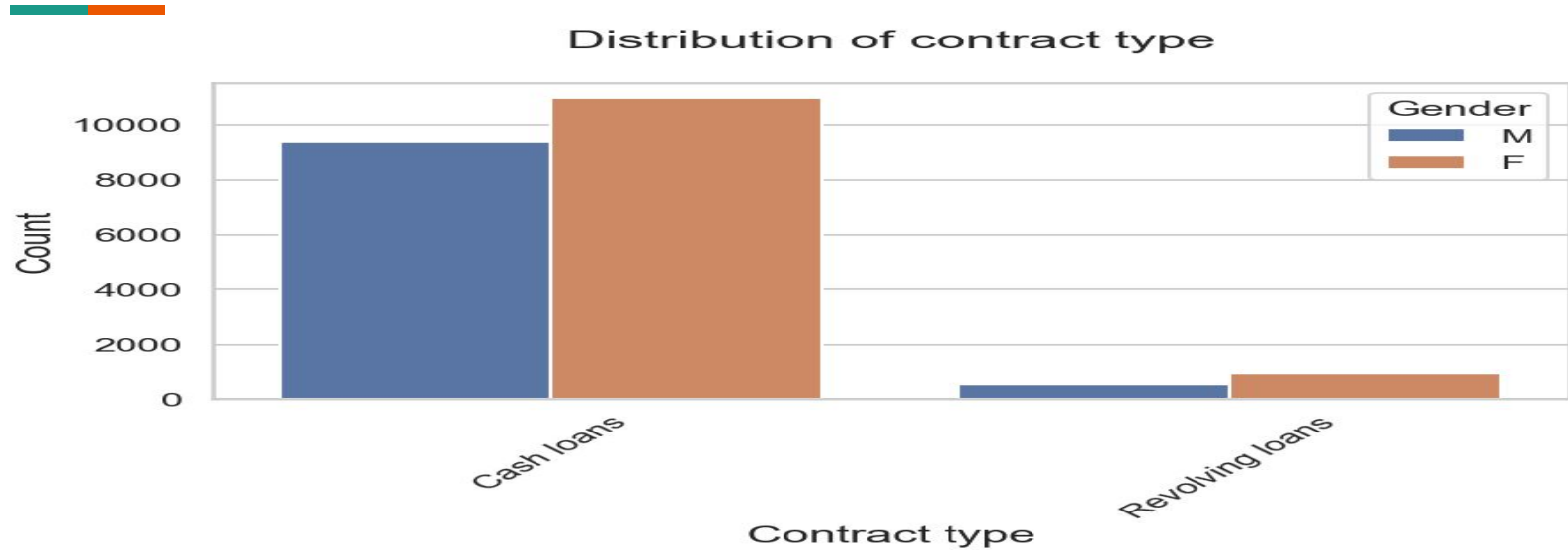
Distribution of Income Type



Based on the information depicted in the graph:

1. Income types such as "Working," "Business associate," and "State employee" have a notably higher number of credits compared to other income types, such as "Maternity leave."
2. The graph suggests that, due to the larger number of credits associated with these income types, women tend to have more credits than men in these categories.
3. Conversely, the "Maternity leave" income category exhibits fewer credits in comparison to other income types.
4. Notably, there are no instances of late payments (type 1) associated with income types "Student," "Pensioner," and "Businessman," indicating a positive payment behavior within these categories.

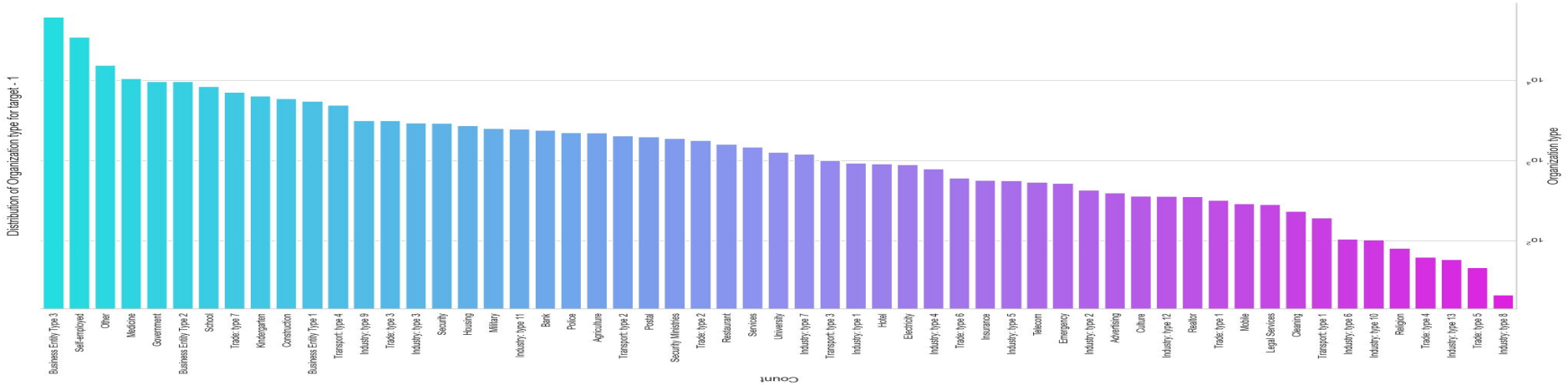
Distribution of Contract Type



Based on the information presented in the graph:

1. Cash loans, as a contract type, clearly have a higher number of credits compared to revolving loans.
2. The graph suggests that there is a greater likelihood of women seeking credit as compared to men.
3. In the context of type 1 loans, there appear to be only female applicants for revolving loans, indicating a gender-specific pattern in this loan category.

Distribution of Organization type for target - 1



Based on the information presented in the rotated graph:

1. The majority of credit applicants belong to organization types such as "Business entity Type 3," "Self employed," "Other," "Medicine," and "Government."
2. In contrast, there are fewer customers associated with trade types 5, 4, and 8, industry type 6, and the religion category in terms of credit applications.
3. The distribution of organization types appears to be consistent with type 0, indicating a similar pattern across both types.



Relating Alpha & Beta - Correlation of Target 0 & Target 1

```
In [85]: # Relating to alpha
```

```
alpha
```

```
Out[85]:
```

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED
CNT_CHILDREN	1.000000	-0.021950	-0.023652	-0.010795	-0.030579	0.266534	
AMT_INCOME_TOTAL	-0.021950	1.000000	0.403876	0.472204	0.110074	-0.054666	
AMT_CREDIT	-0.023652	0.403876	1.000000	0.826689	0.060706	-0.169030	
AMT_ANNUITY	-0.010795	0.472204	0.826689	1.000000	0.064328	-0.100287	
REGION_POPULATION_RELATIVE	-0.030579	0.110074	0.060706	0.064328	1.000000	-0.041663	
DAYS_BIRTH	0.266534	-0.054666	-0.169030	-0.100287	-0.041663	1.000000	
DAYS_EMPLOYED	0.030948	-0.060868	-0.104251	-0.074643	0.000900	0.307787	
DAYS_REGISTRATION	0.155518	0.040559	-0.015318	0.010712	-0.042400	0.265449	
DAYS_ID_PUBLISH	-0.119164	-0.036702	-0.038197	-0.027354	-0.010299	0.083331	
HOURLY_APPR_PROCESS_START	-0.030162	0.073503	0.036923	0.032953	0.133213	0.051299	
REG_REGION_NOT_LIVE_REGION	-0.022813	0.077634	0.015118	0.033435	-0.025292	0.058627	

Alpha & Beta - Plotting Heat Maps for Both

```
In [86]: #Relating to beta
```

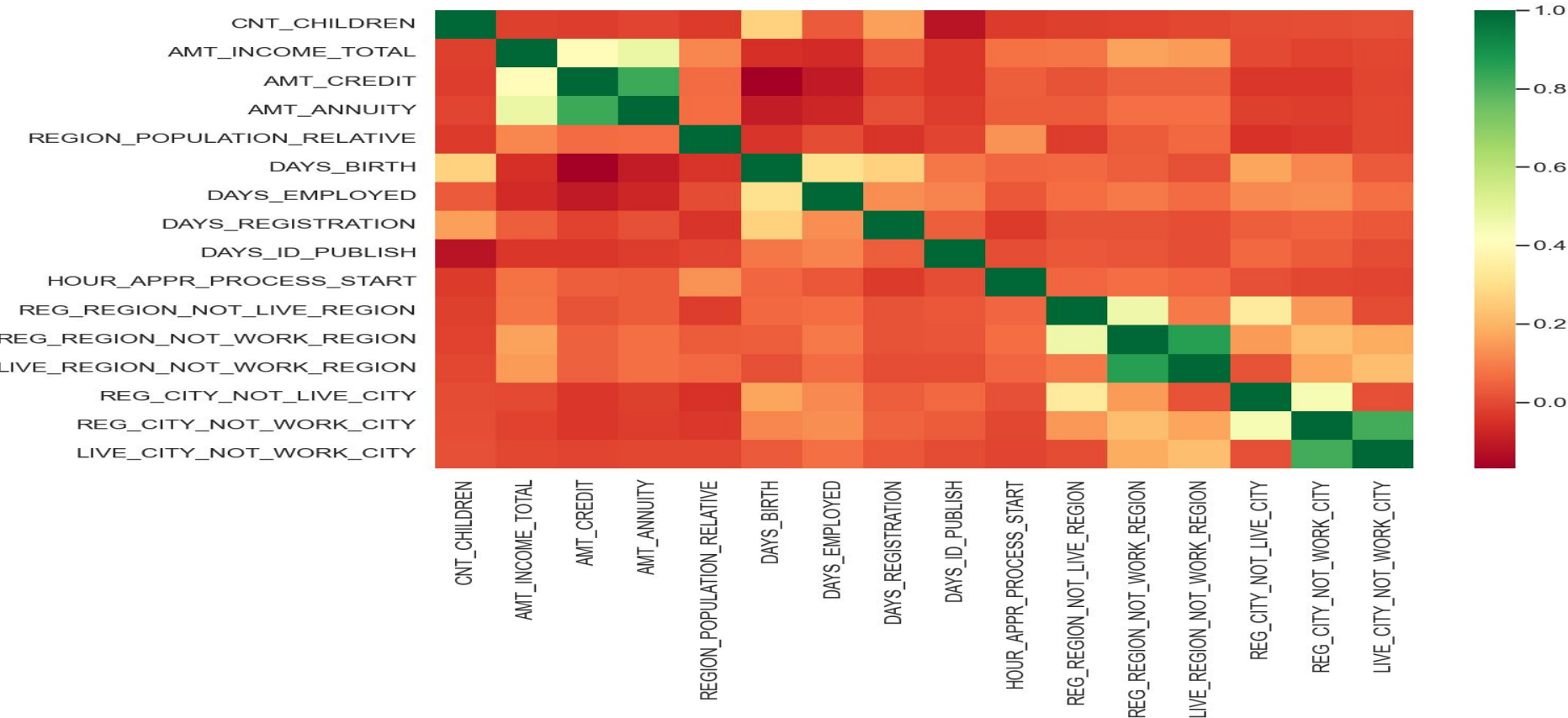
```
beta
```

```
Out[86]:
```

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED
CNT_CHILDREN	1.000000	-0.039123	0.000427	0.015133	-0.029682	0.175025	
AMT_INCOME_TOTAL	-0.039123	1.000000	0.364559	0.428947	0.058005	-0.103026	
AMT_CREDIT	0.000427	0.364559	1.000000	0.812093	0.043545	-0.200718	
AMT_ANNUITY	0.015133	0.428947	0.812093	1.000000	0.028666	-0.100200	
REGION_POPULATION_RELATIVE	-0.029682	0.058005	0.043545	0.028666	1.000000	-0.044444	
DAYS_BIRTH	0.175025	-0.103026	-0.200718	-0.100200	-0.044444	1.000000	
DAYS_EMPLOYED	0.006823	-0.053798	-0.107605	-0.060193	-0.015246	0.256870	
DAYS_REGISTRATION	0.110854	0.011378	-0.021973	0.019762	-0.033490	0.192350	
DAYS_ID_PUBLISH	-0.091042	-0.051113	-0.065143	-0.044128	-0.017779	0.146246	
HOURLY_APPR_PROCESS_START	-0.040338	0.078779	0.024616	0.021129	0.109400	0.041994	
REG_REGION_NOT_LIVE_REGION	-0.035213	0.075615	0.015043	0.029646	-0.032702	0.046320	
REG_REGION_NOT_WORK_REGION	-0.040853	0.156374	0.032536	0.060363	-0.008160	0.022208	
LIVE_REGION_NOT_WORK_REGION	-0.027993	0.145982	0.034861	0.059724	0.012602	0.000356	
REG_CITY_NOT_LIVE_CITY	-0.016072	-0.003813	-0.030974	-0.011744	-0.057239	0.145884	
REG_CITY_NOT_WORK_CITY	-0.005444	-0.006241	-0.032882	-0.015938	-0.044761	0.096181	
LIVE_CITY_NOT_WORK_CITY	0.009557	0.004230	-0.012465	-0.003012	-0.014753	0.009633	

Plotting Alpha - (Correlation Matrix for Target 0)

Correlation Matrix for Target 0



Observations Alpha - (Correlation Matrix for Target 0)

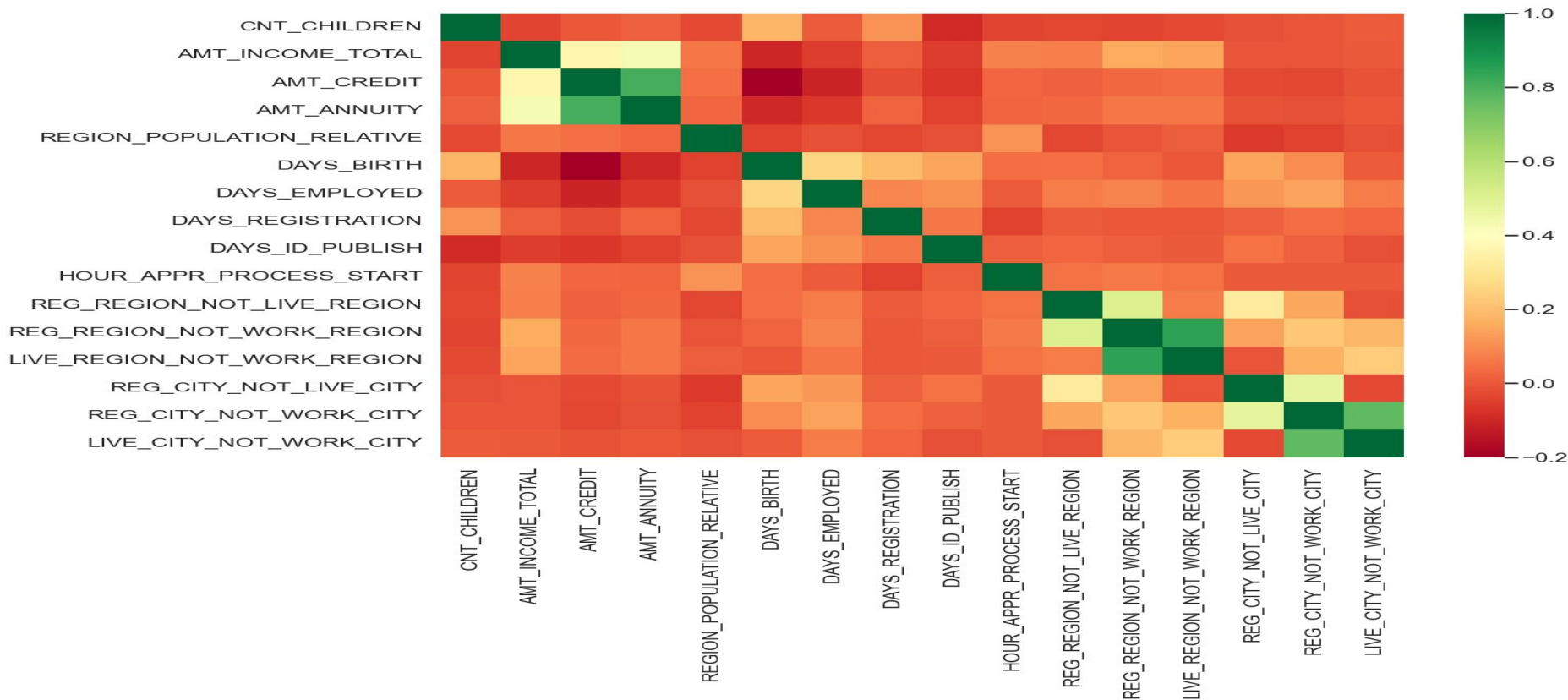


From the correlation heatmap presented above, several noteworthy observations can be made:

1. ****Inverse Relationship Between Credit Amount and Date of Birth:**** It is evident that credit amount tends to be higher for clients with a lower age, and conversely lower for older clients.
2. ****Inverse Relationship Between Credit Amount and Number of Children:**** The data suggests that credit amount is higher for clients with fewer children, while it tends to decrease as the number of children increases.
3. ****Inverse Relationship Between Income and Number of Children:**** There appears to be an inverse correlation between income and the number of children a client has. Clients with fewer children tend to have higher incomes.
4. ****Fewer Children in Densely Populated Areas:**** The data indicates that clients with fewer children are more likely to reside in densely populated areas.
5. ****Higher Credit Amounts in Densely Populated Areas:**** Densely populated areas are associated with higher credit amounts.
6. ****Higher Income in Densely Populated Areas:**** Clients residing in densely populated areas tend to have higher incomes.

Plotting Beta - (Correlation Matrix for Target 1)

Correlation Matrix for Target 1



Observations Beta - (Correlation Matrix for Target 1)



From the correlation heatmap presented above, several noteworthy observations can be made:

1. Permanent Address vs. Contact Address:

- Fewer children are born at the client's permanent address compared to their contact address.
- Conversely, more children are born at the client's contact address than at their permanent address.

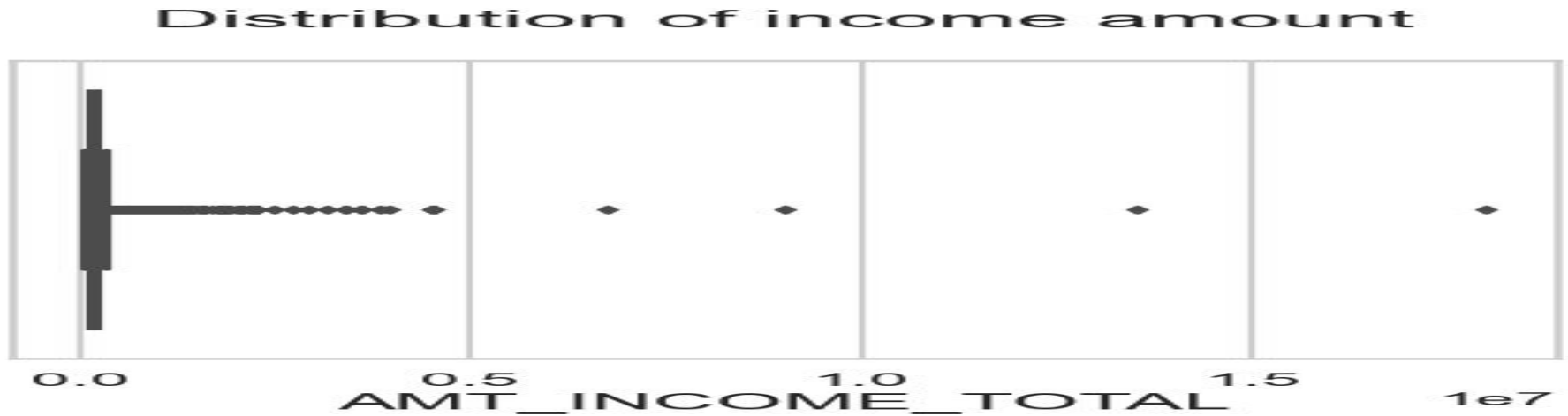
2. Permanent Address vs. Work Address:

- Fewer children are born to clients whose permanent addresses differ from their work addresses.
- Conversely, more children are born to clients whose permanent addresses match their work addresses.



Univariate analysis for variables - Target 0 (Box-Plot)

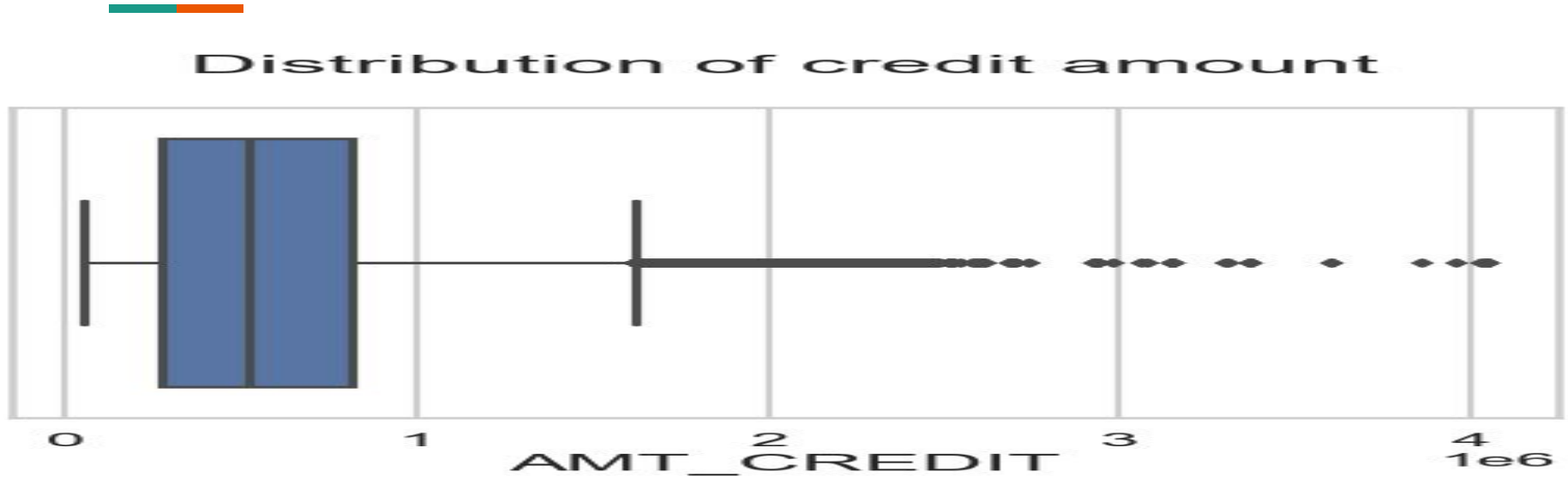
Income Amount Plot



Income Amount plot describes:-

1. Some income outliers have been identified.
2. In terms of income, the third quartile is quite low.

Credit Amount Plot

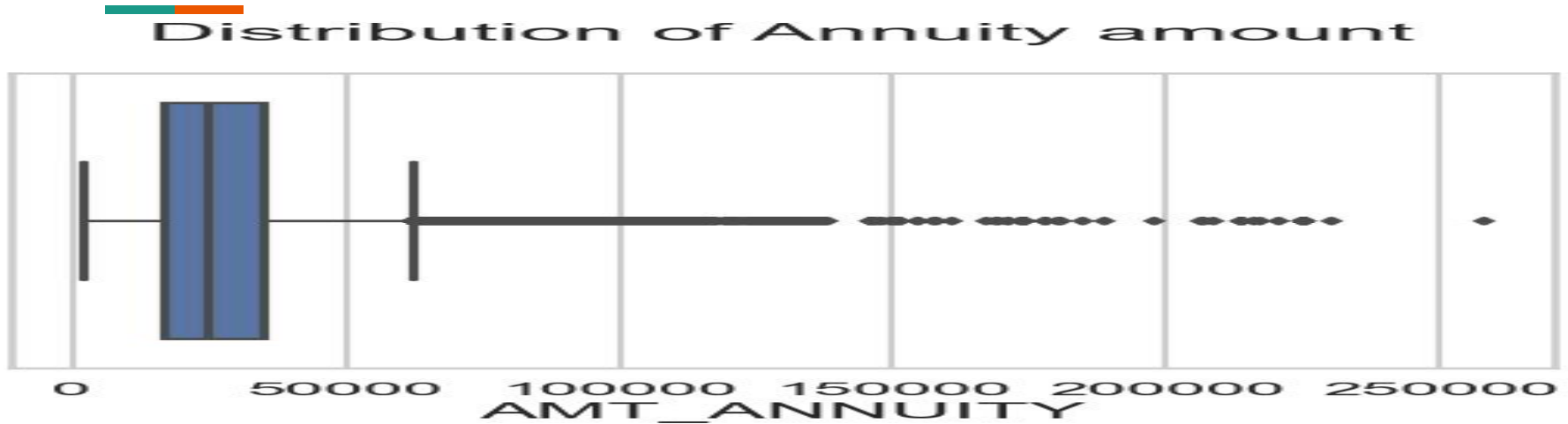


Credit Amount plot describes:-

1. Some credit amount anomalies are discovered.

2. For credit amount, the first quartile is greater than the third quartile, indicating that the majority of clients' credits are in the first quartile.

Annuity Amount Plot



Annuity Amount plot describes:-

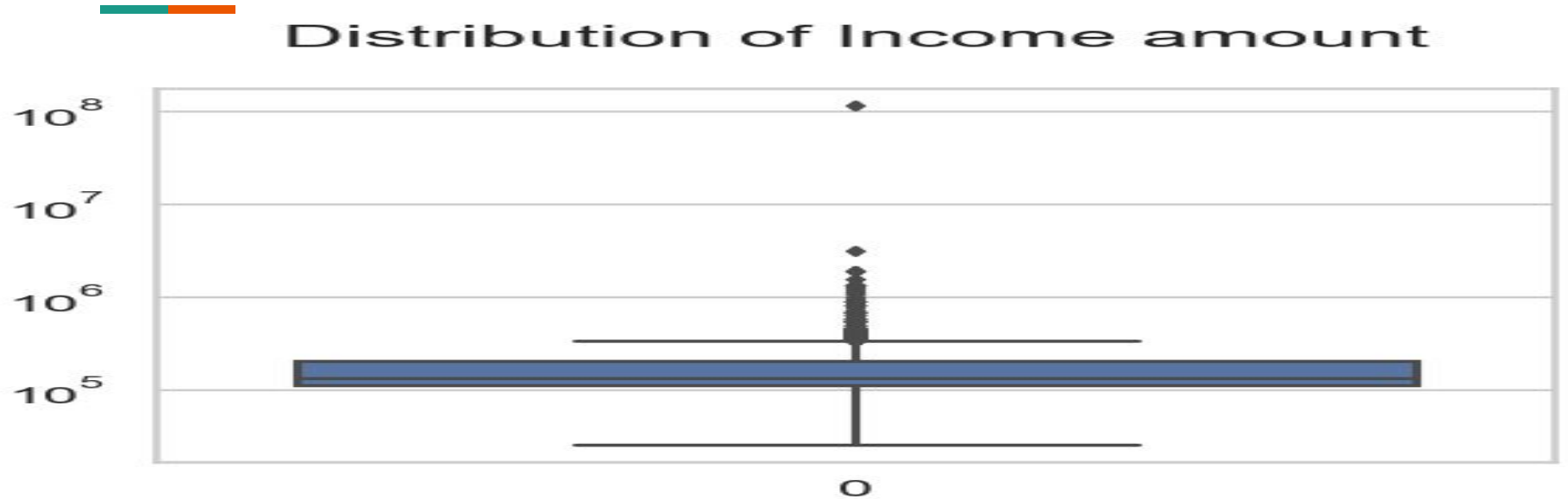
1. There are certain anomalies in annuity amounts.

2. The first quartile has a higher annuity amount than the third quartile, implying that the majority of annuity clients are in the first quartile.



Univariate analysis for variables - Target 1 (Box-Plot)

Income Amount

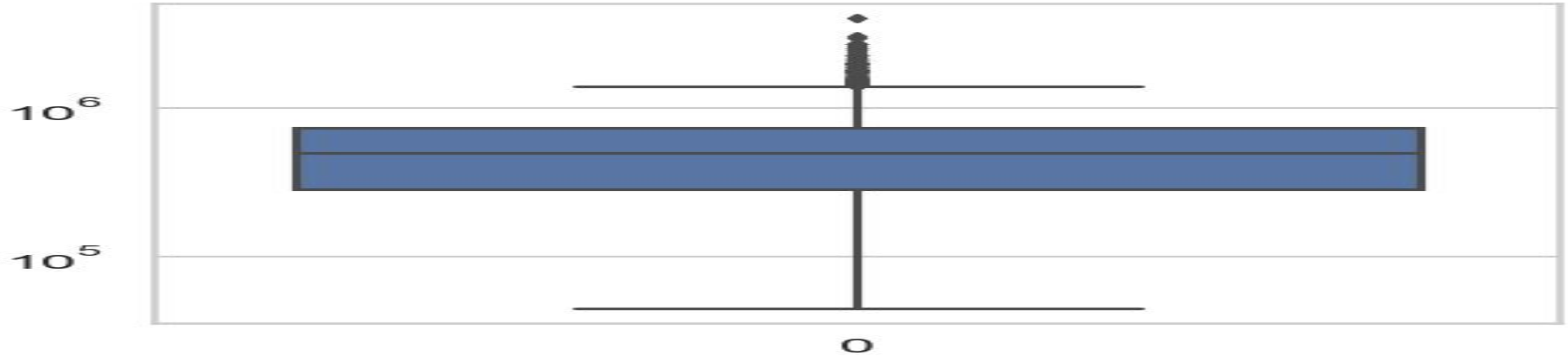


Income Amount

1. Income amounts have some outliers.
2. In terms of income, the third quartile is quite little.

Credit Income

Distribution of Credit amount

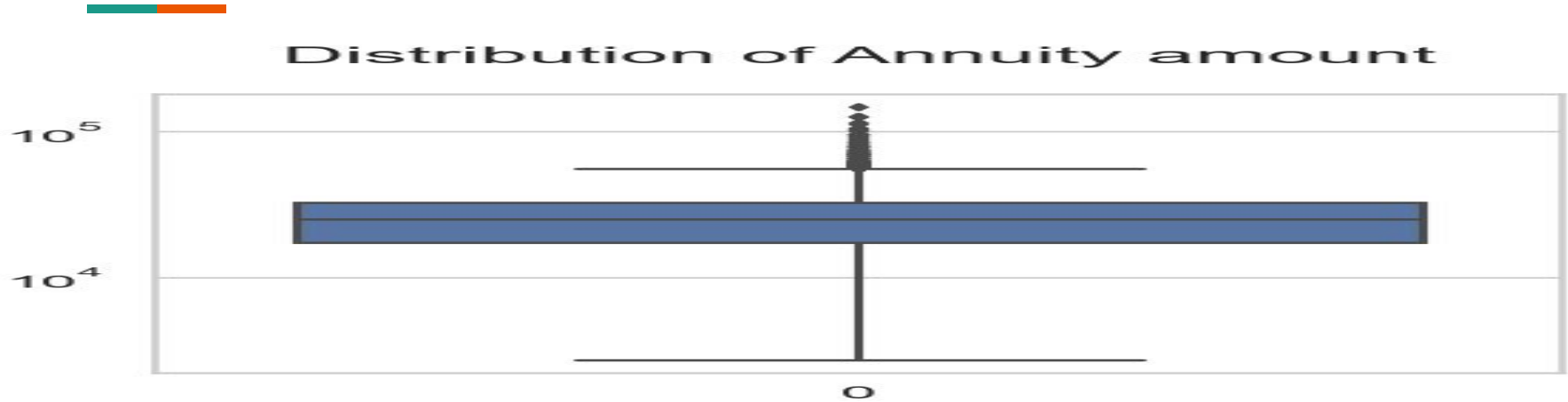


Credit Income

1. Some credit amount anomalies are discovered.

2. For credit amount, the first quartile is greater than the third quartile, indicating that the majority of clients' credits are in the first quartile.

Annuity Income



Annuity Income

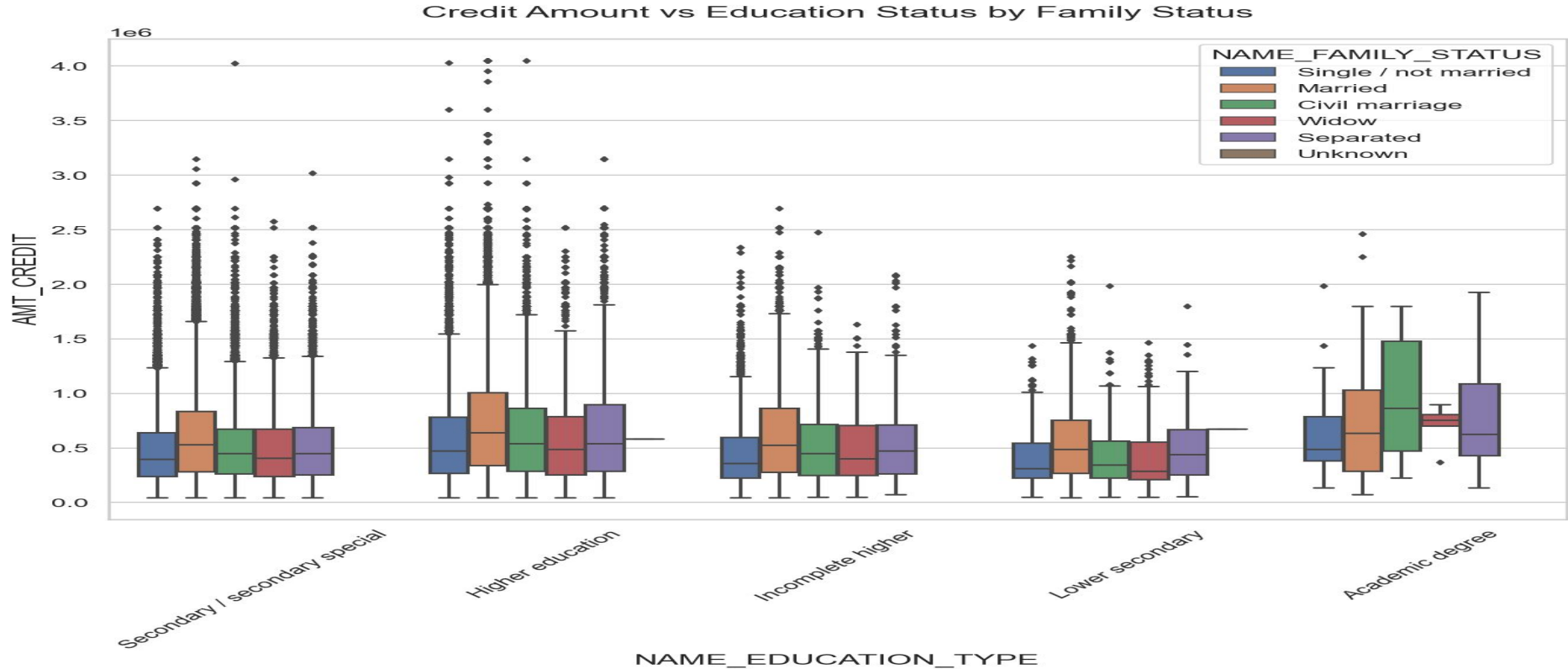
1. Some outliers are noticed in annuity amount.

2. The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.



Bivariate analysis for numerical variables(Target 0)

Credit Amount vs Education Status by Family Status





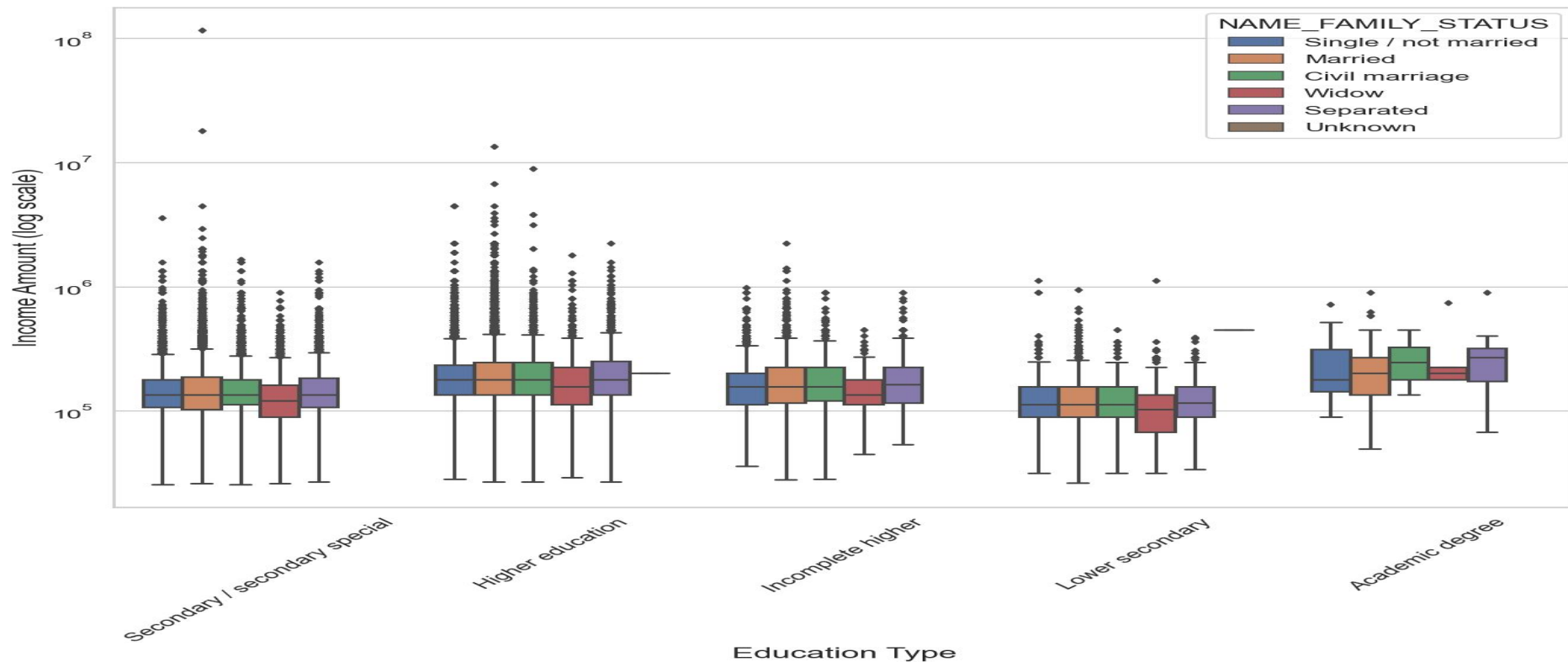
SUMMARY of Plot(Credit Amount vs Education Status by Family Status)

OUTLINE

According to the above box plot, family statuses of 'civil marriage,' 'marriage,' and 'separated' of academic degree education have a larger number of credits than others. Furthermore, greater education of family status of marriage, 'single,' and 'civil marriage' has more outliers. Civil marriage for Academic degree has the majority of credits in the third quartile.

Income Amount vs Education Status

Income Amount vs Education Status





SUMMARY of Plot(Income Amount vs Education Status)

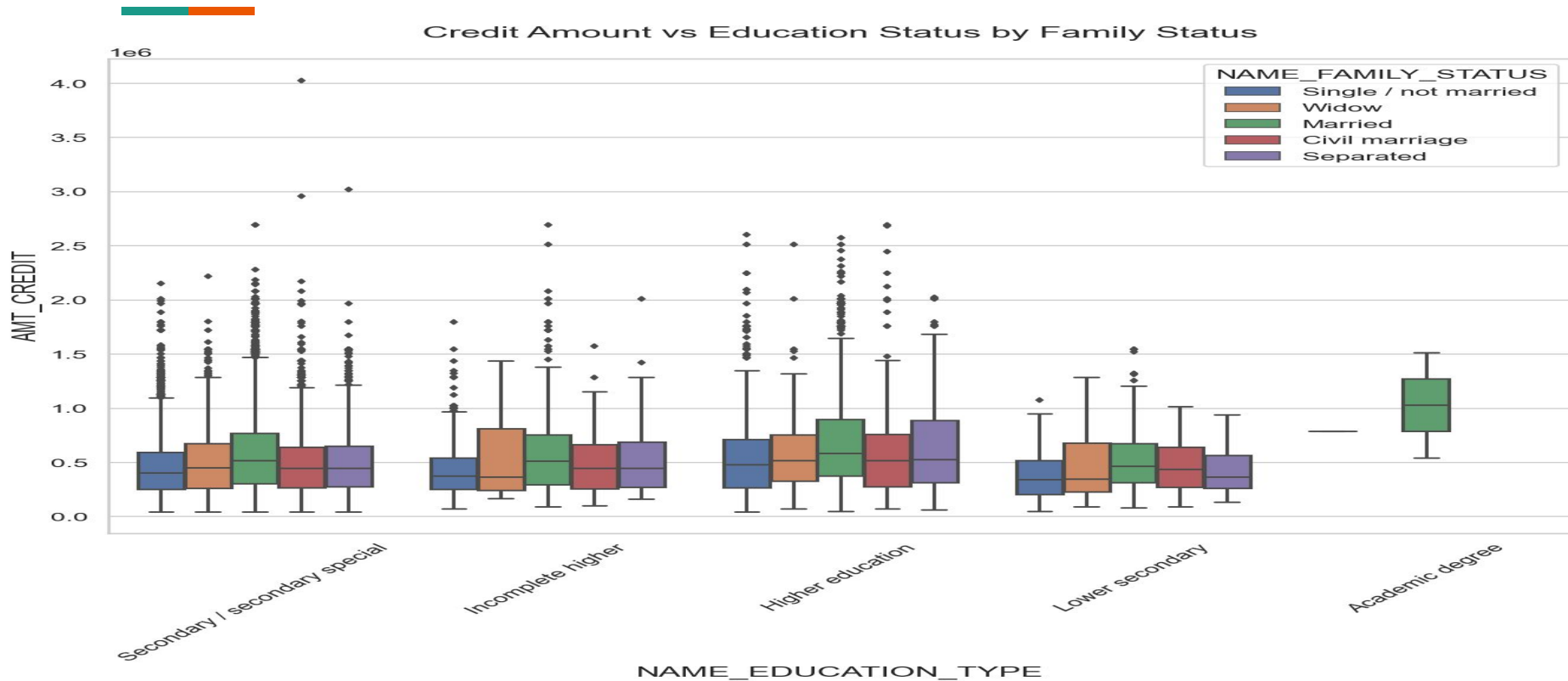
OUTLINE

Based on the above boxplot, the income amount for Education type 'Higher education' is basically equivalent with family status. It does have a lot of outliers. Academic degrees are less outlier, but their compensation is slightly greater than higher education. Lower secondary civil marriage family status has lower income than others.



Bivariate analysis for numerical variables(Target 1)

Credit Amount vs Education Status by Family Status



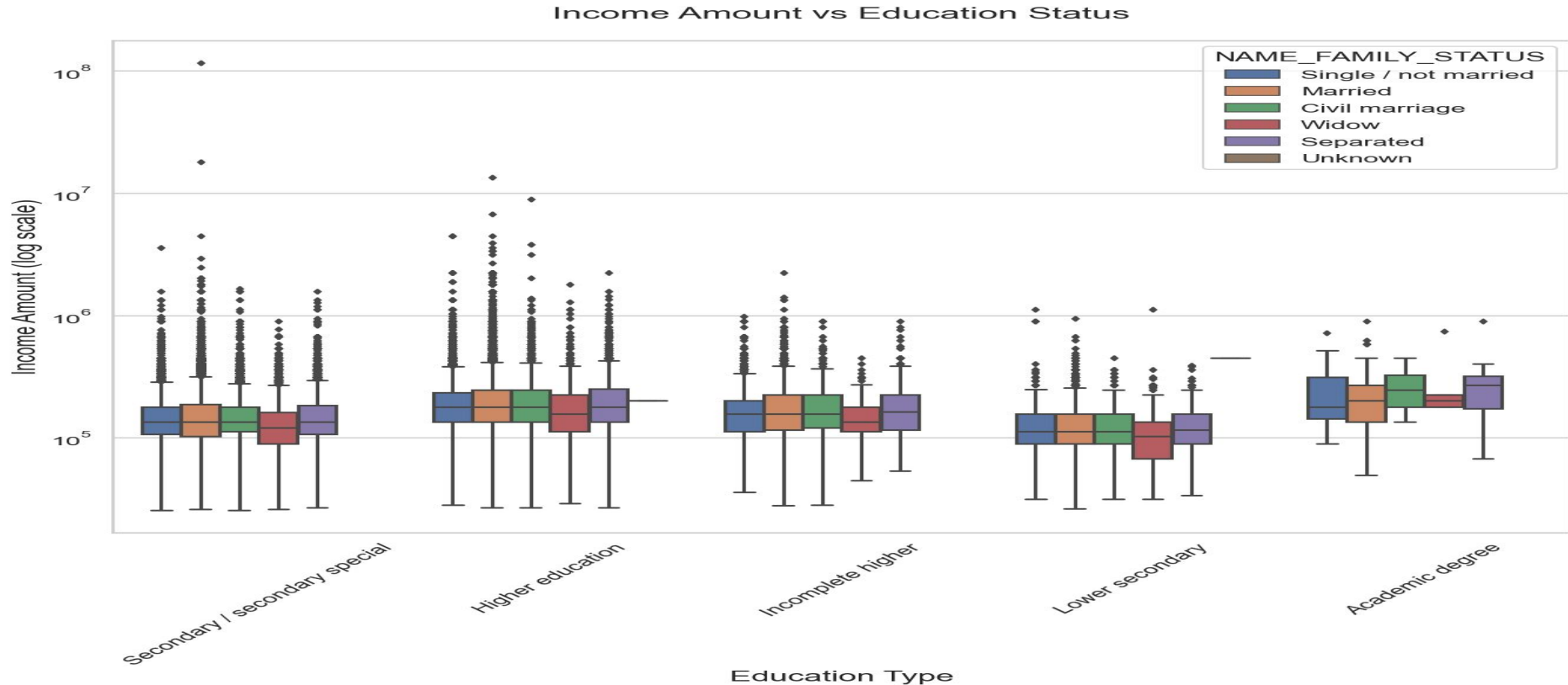


SUMMARY of Plot(Credit Amount vs Education Status by Family Status)

OUTLINE

Identical to Target 0. According to the above box plot, family statuses of 'civil marriage,' 'marriage,' and 'separated' of academic degree education have a larger number of credits than others. The majority of the outliers are from the Education types 'Higher education' and 'Secondary'. Civil marriage for Academic degree has the majority of credits in the third quartile.

Income Amount vs Education Status





SUMMARY of Plot(Income Amount vs Education Status)

OUTLINE

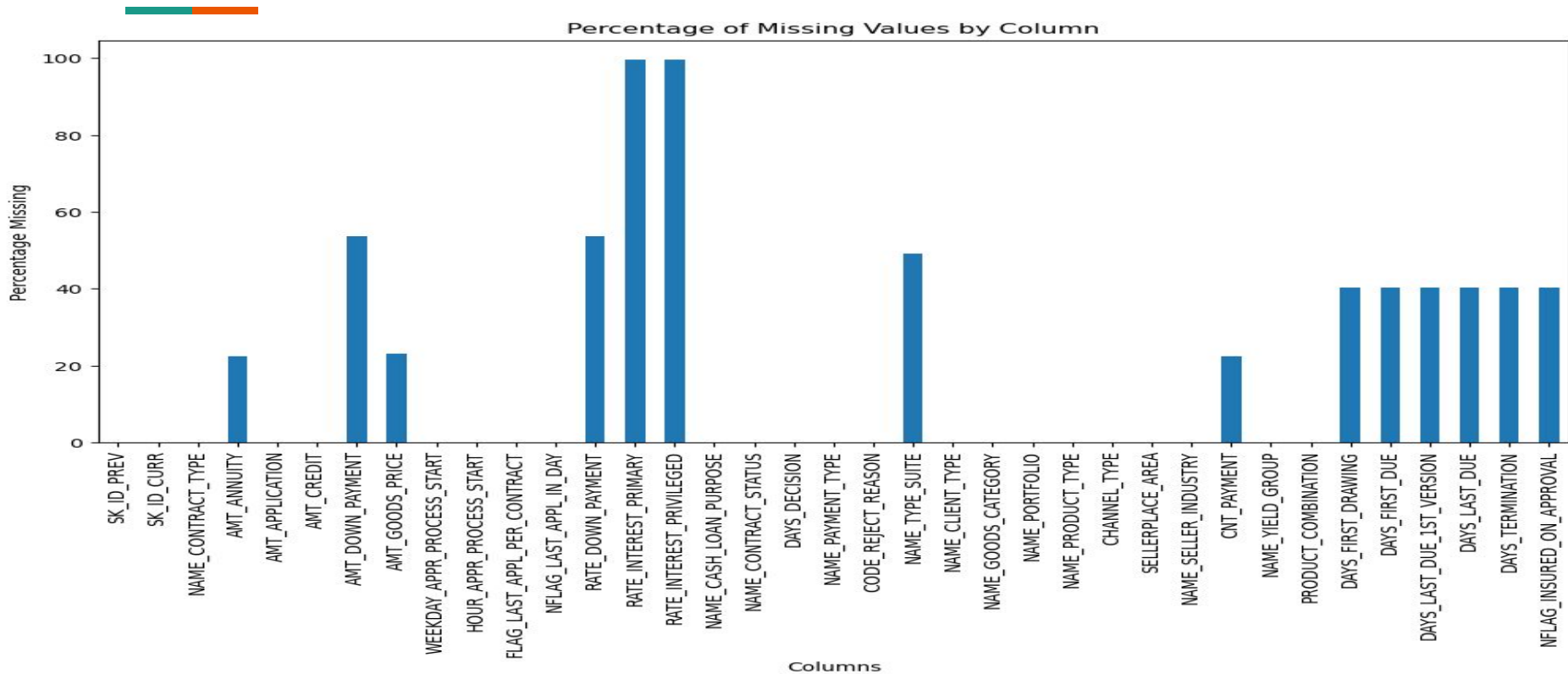
Similar to Target0, the income amount from the above boxplot for Education type 'Higher education' is basically equivalent with family status. Academic degrees are less outlier, but their compensation is slightly greater than higher education. Lower secondary students earn less than others.



Previous Application Analysis

For our analysis, we have now used the second data set, the "Previous Application Data Set." This includes details on prior loans made to the client. It includes information about whether the prior application was accepted, rejected, canceled, or not used. There are 37 columns and 1670214 rows in this data collection.

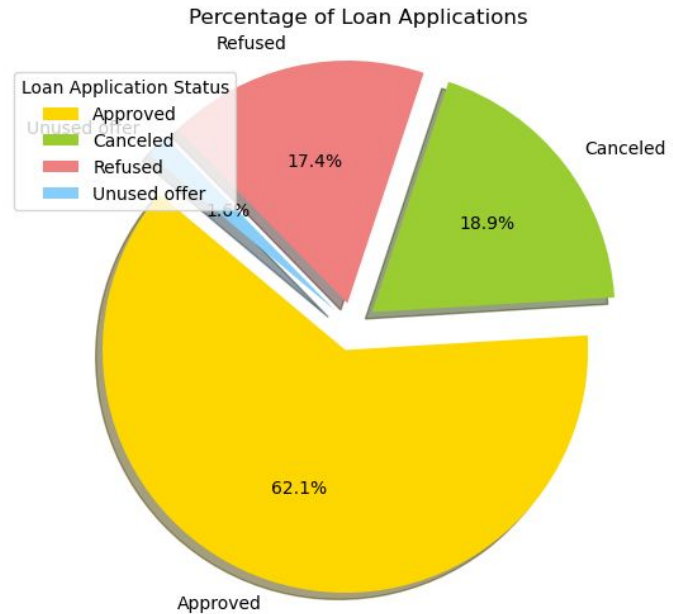
Percentage of Missing Values by column



Contract Status

	NAME_CONTRACT_STATUS	Percentage_of_Values
0	Approved	62.07
1	Canceled	18.94
2	Refused	17.40
3	Unused offer	1.58

Observation - Approved is high

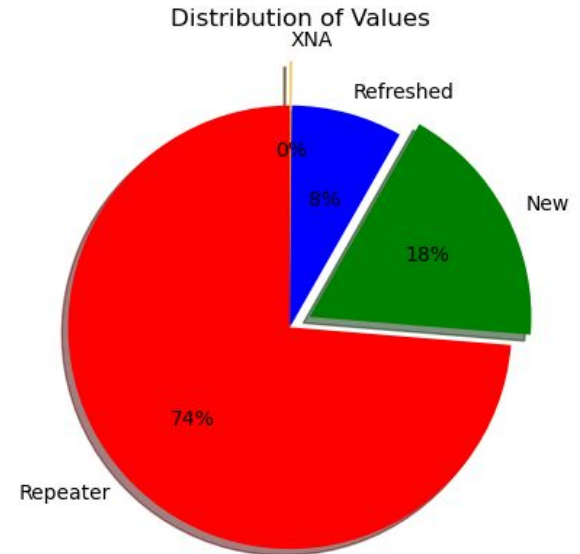


Client Type



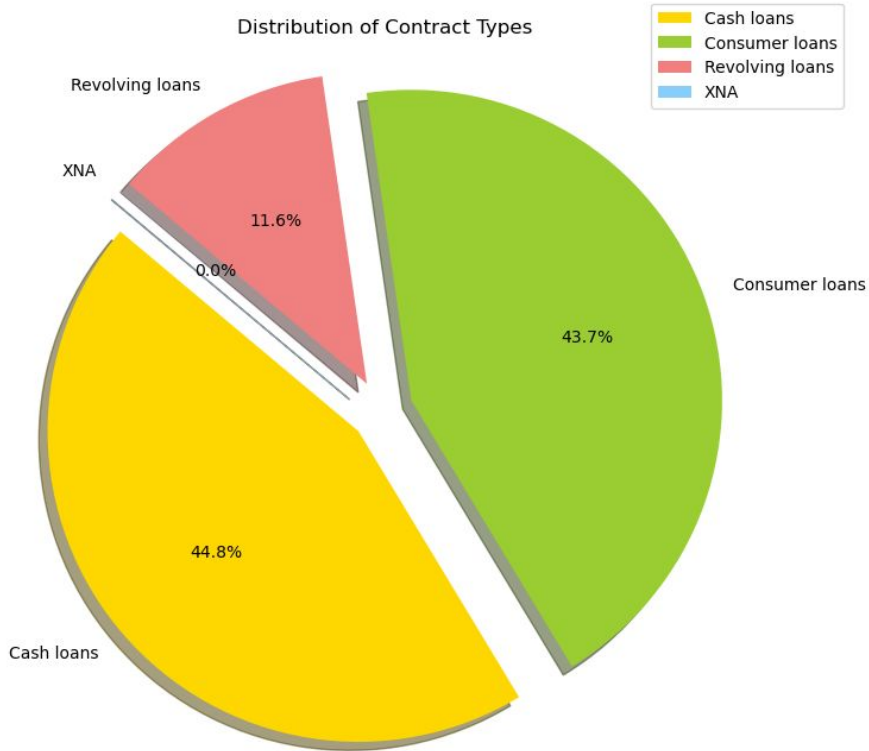
	NAME_CLIENT_TYPE	Percentage_of_Values
0	Repeater	73.72
1	New	18.04
2	Refreshed	8.12
3	XNA	0.12

Observation - Repeater is at peak comparatively



Contract Type

Distribution of Contract Types

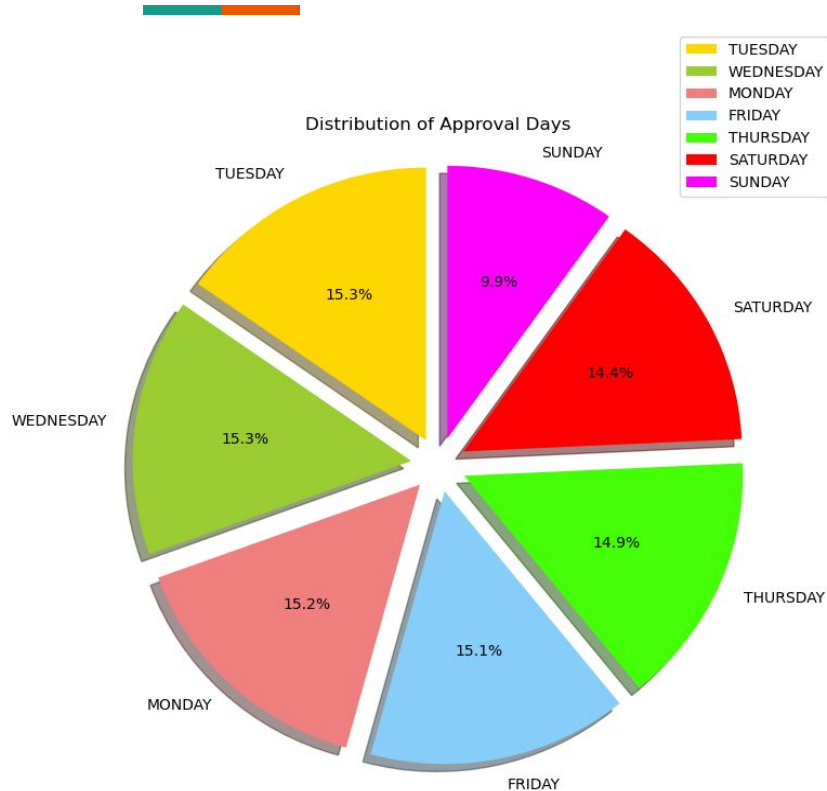


NAME_CONTRACT_TYPE Percentage_of_Values

0	Cash loans	44.76
1	Consumer loans	43.66
2	Revolving loans	11.57
3	XNA	0.02

Observation - Cash Loans are higher followed by Consumer loans

Based on Days of Approval

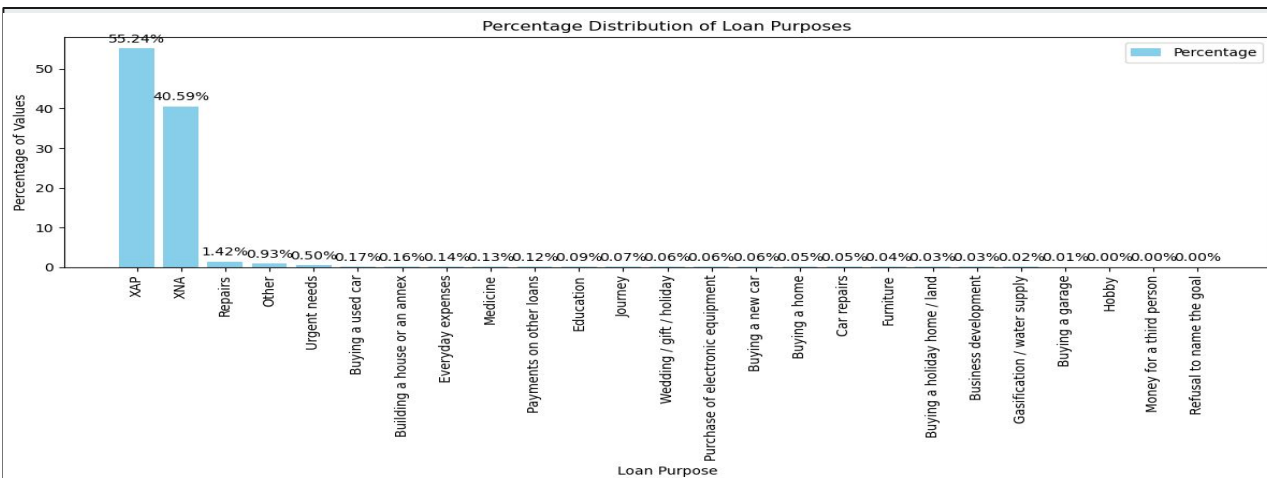


	Weekday	Percentage
0	TUESDAY	15.27
1	WEDNESDAY	15.27
2	MONDAY	15.18
3	FRIDAY	15.09
4	THURSDAY	14.91
5	SATURDAY	14.41
6	SUNDAY	9.86

Observation - Tuesday number of applicant .

Weekdays>Weekends

Distribution of Loan



**Observation - Most loan is in XAP
(record not found)**

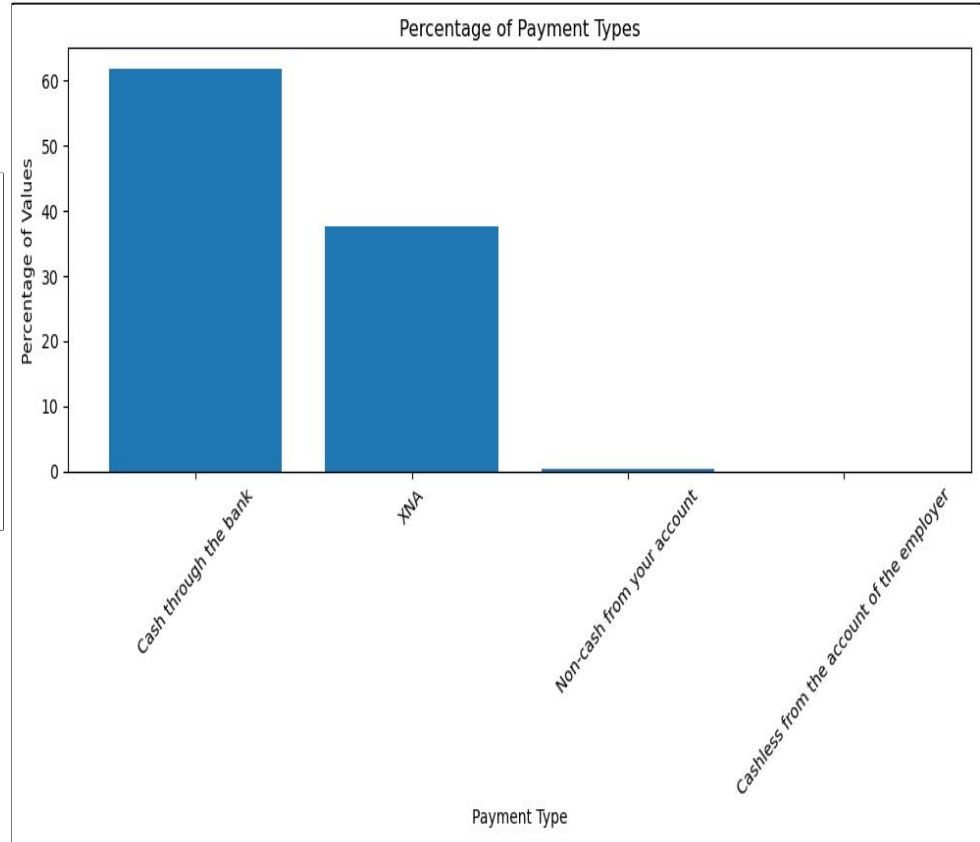
	Loan_Purpose	Percentage
0	XAP	55.24
1	XNA	40.59
2	Repairs	1.42
3	Other	0.93
4	Urgent needs	0.50
5	Buying a used car	0.17
6	Building a house or an annex	0.16
7	Everyday expenses	0.14
8	Medicine	0.13
9	Payments on other loans	0.12
10	Education	0.09
11	Journey	0.07
12	Wedding / gift / holiday	0.06
13	Purchase of electronic equipment	0.06
14	Buying a new car	0.06
15	Buying a home	0.05
16	Car repairs	0.05
17	Furniture	0.04
18	Buying a holiday home / land	0.03
19	Business development	0.03
20	Gasification / water supply	0.02
21	Buying a garage	0.01
22	Hobby	0.00
23	Money for a third person	0.00
24	Refusal to name the goal	0.00

Payment Type - Source Plot

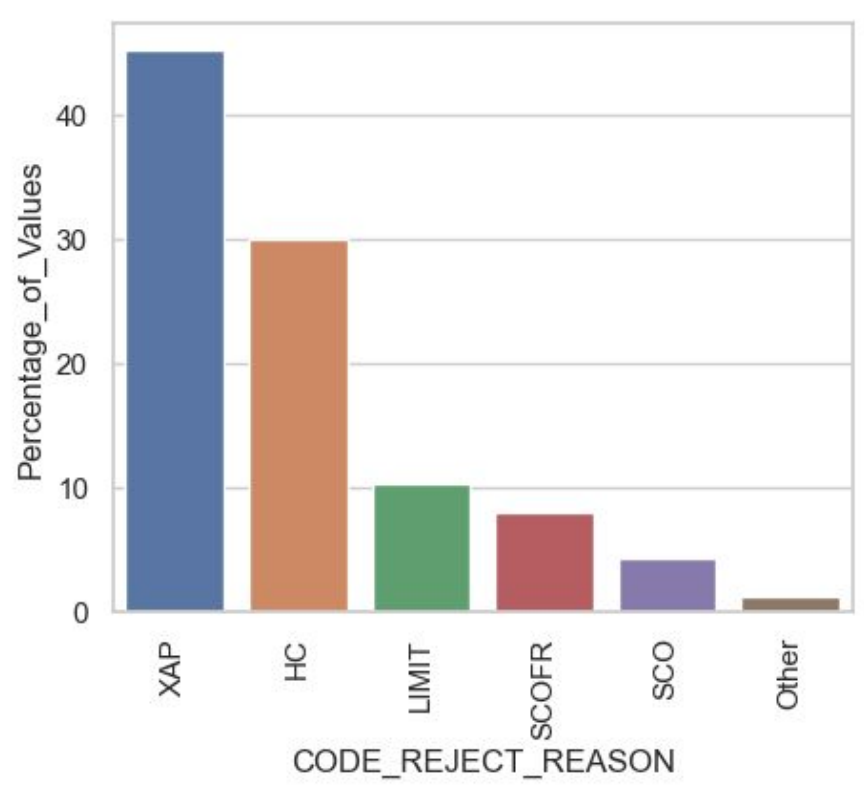


	NAME_PAYMENT_TYPE	Percentage_of_Values
0	Cash through the bank	61.88
1	XNA	37.56
2	Non-cash from your account	0.49
3	Cashless from the account of the employer	0.06

Observation - Cash was most preferable.



Loan Rejection Reasons

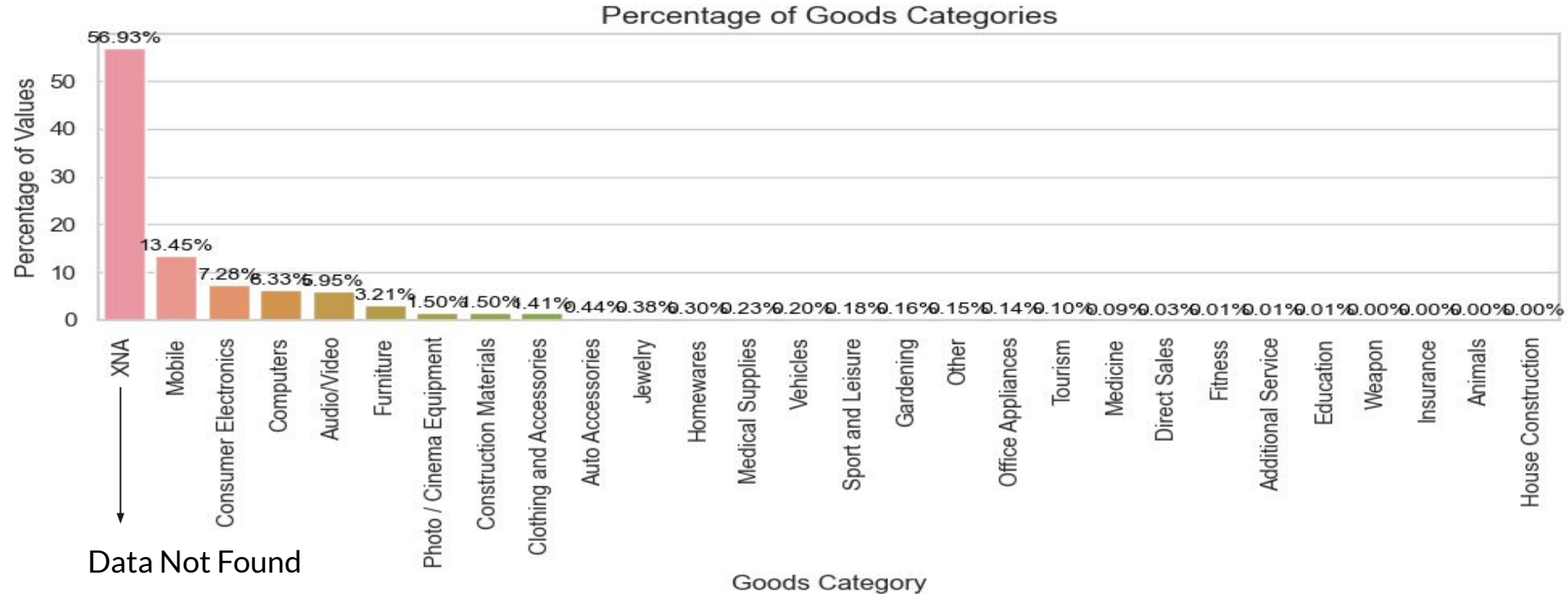


CODE_REJECT_REASON	Percentage_of_Values
XAP	45.25%
HC	30.10%
LIMIT	10.25%
SCOFR	8.00%
SCO	4.20%
Other	1.20%

Observation - Due to XAP loans rejection was high followed by HC.

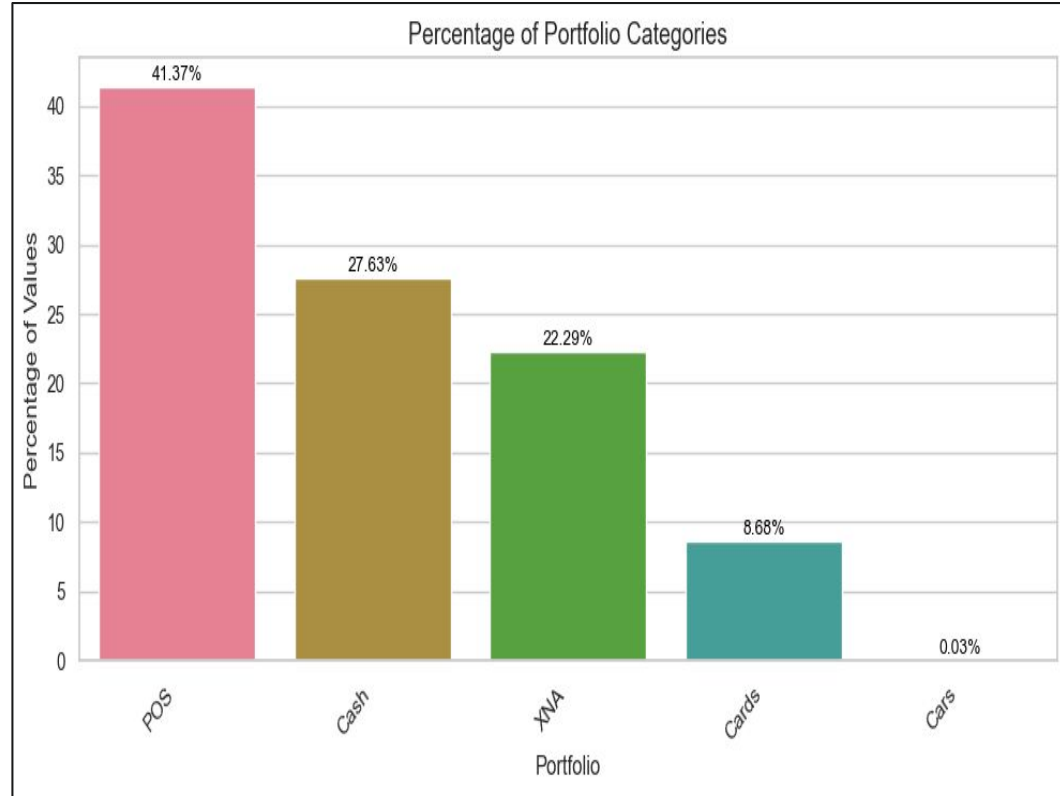
Customer Order Variation

Observation - Mobile and Electronics was most applied.



Percentage of Portfolio Categories (Name_Portfolio)

	NAME_PORTFOLIO	Percentage_of_Values
0	POS	41.37
1	Cash	27.63
2	XNA	22.29
3	Cards	8.68
4	Cars	0.03



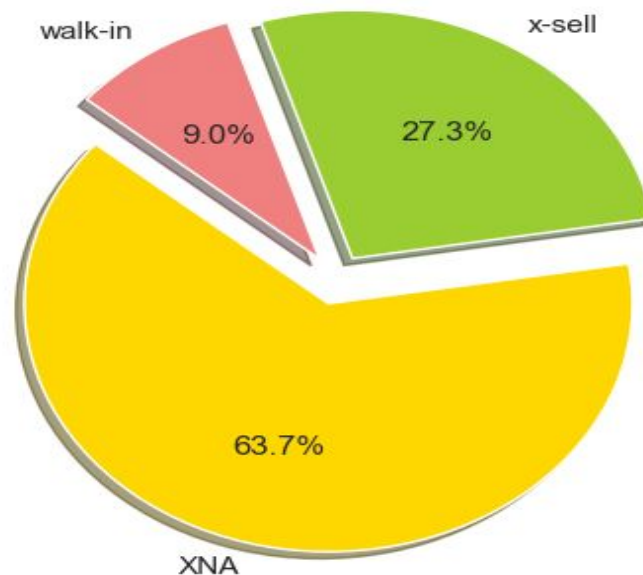
Observation - POS got the highest application followed by cash

Product_Type



	NAME_PRODUCT_TYPE	Percentage_of_Values
0	XNA	63.68
1	x-sell	27.32
2	walk-in	9.00

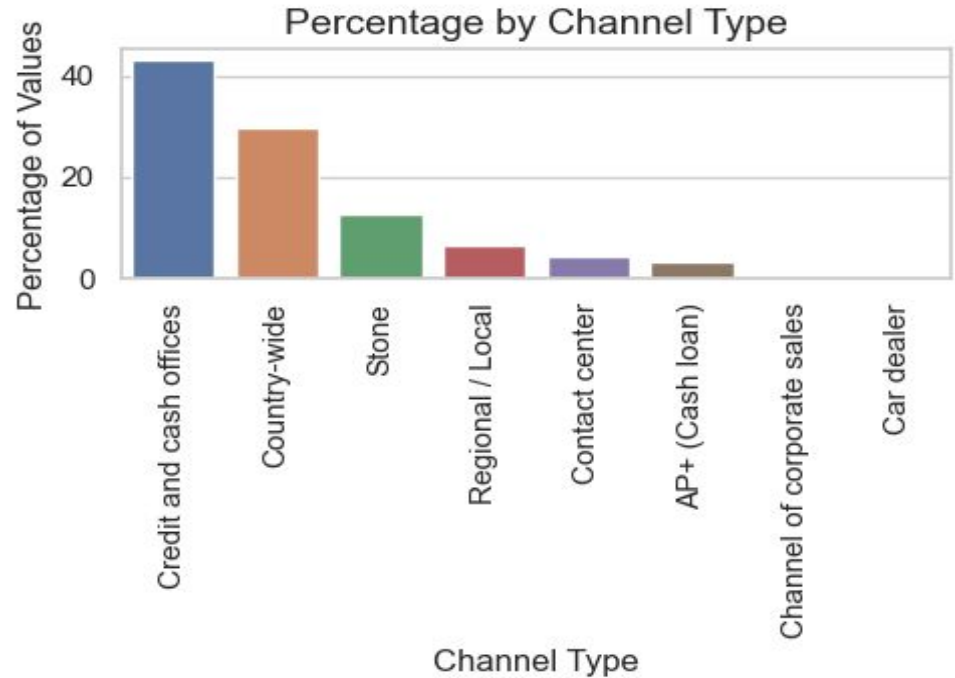
Observation - X-Sell > Walk-in



Observation - Credit and Cash
Offices are on the peak max
number of Clients

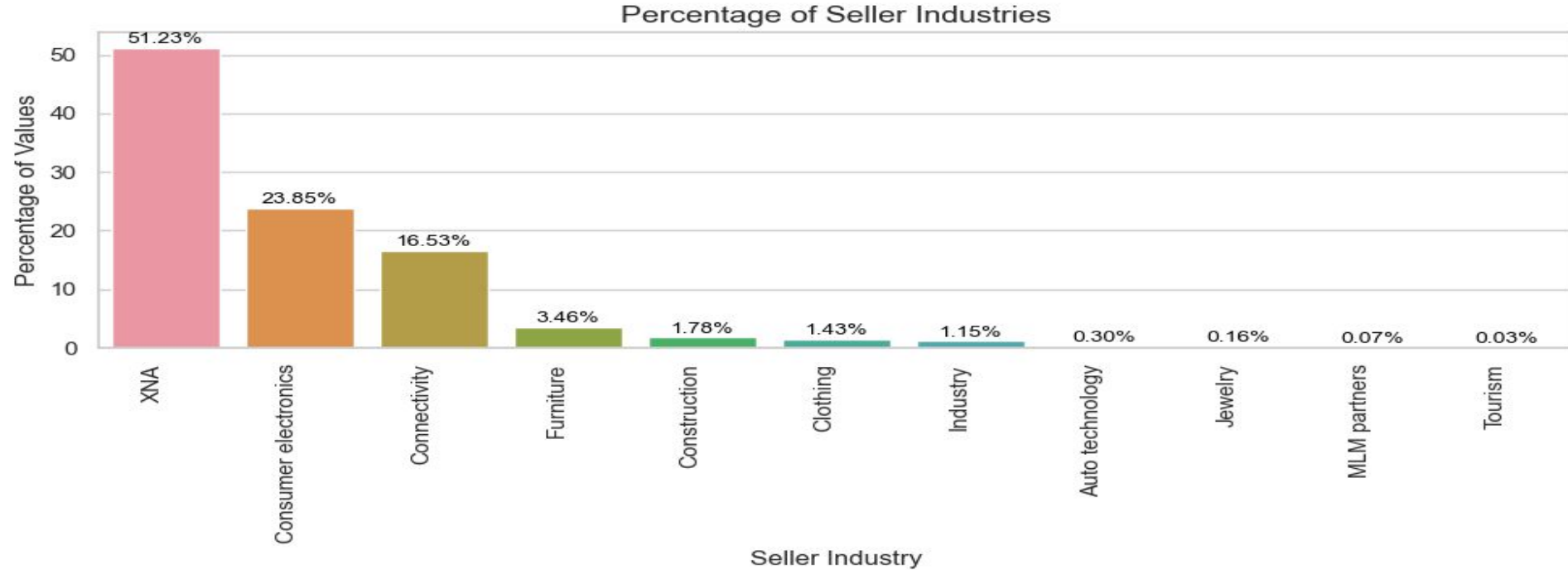
Channel Type Bifurcation

	CHANNEL_TYPE	Percentage_of_Values
0	Credit and cash offices	43.11
1	Country-wide	29.62
2	Stone	12.70
3	Regional / Local	6.50
4	Contact center	4.27
5	AP+ (Cash loan)	3.42
6	Channel of corporate sales	0.37
7	Car dealer	0.03



Name_Seller_Industry(%)

Observation - Consumer Electronics are the highest number of sellers

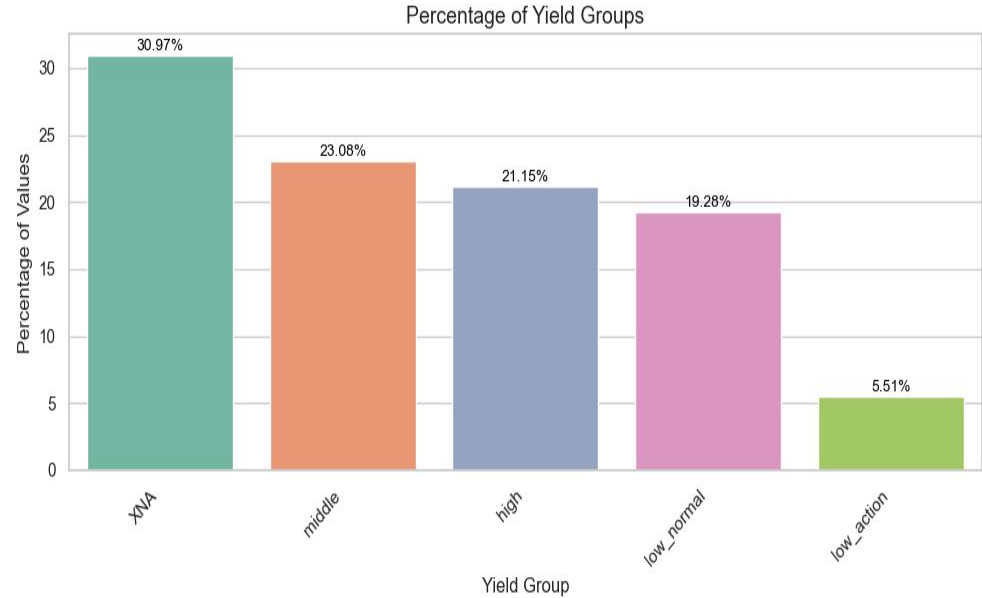


NAME_YIELD_GROUPS



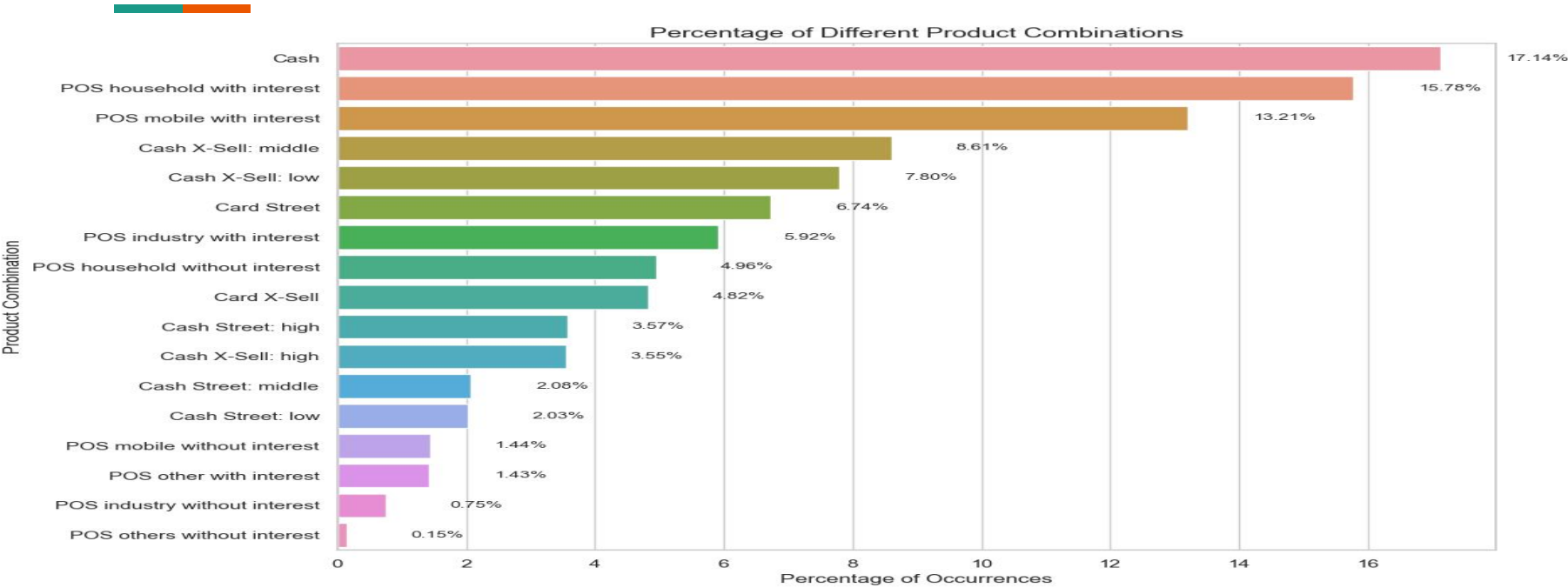
	NAME_YIELD_GROUP	Percentage_of_Values
0	XNA	30.97
1	middle	23.08
2	high	21.15
3	low_normal	19.28
4	low_action	5.51

Observation - Middle is the centre of attention.

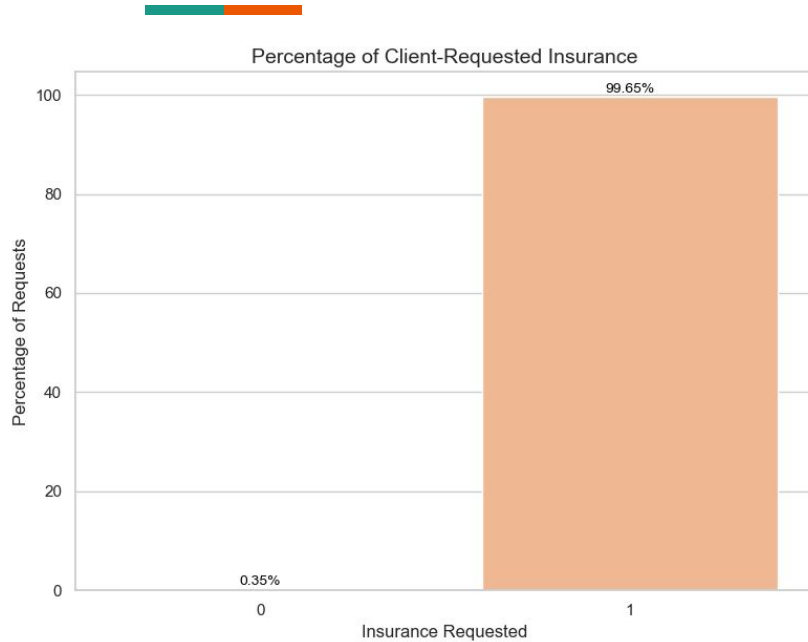


Observation - Cash is highest followed by POS household with interest.

Product Combo



Insurance Percentage Scale (Requested)



Observation: For many clients, this is their final application of the day.

Correlation of Heat Maps _ previous data

INPUT CODE

CORRELATION_HEATMAP

```
In [46]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Assuming you have defined the 'previous_data' Dataframe

# Calculate the correlation matrix
Correlation = previous_data.corr()

# Create a correlation heatmap
plt.figure(figsize=(12, 8))
sns.set(style="white") # Set background style

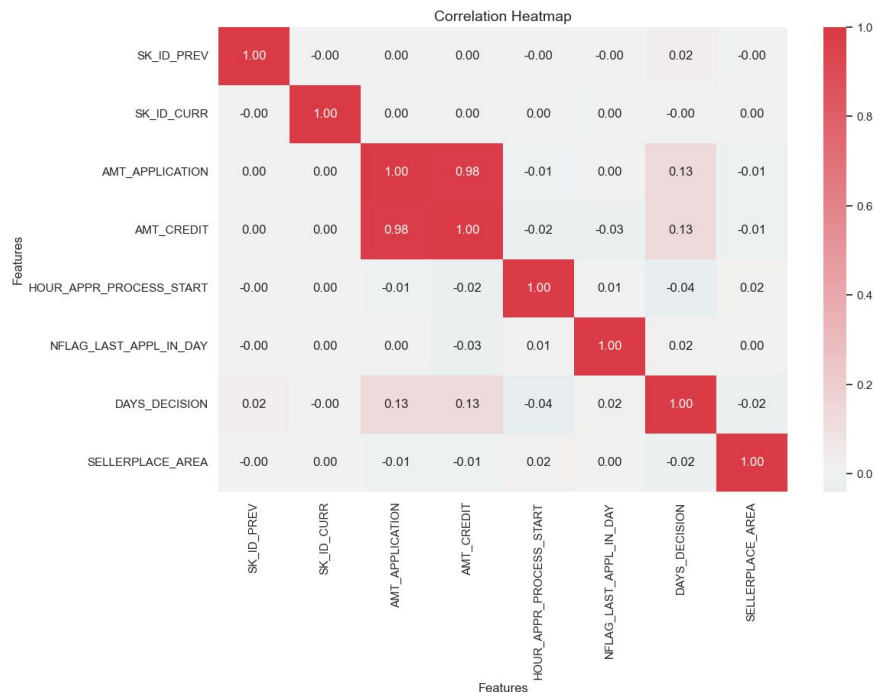
# Customize the colormap for better visibility
cmap = sns.diverging_palette(220, 10, as_cmap=True) # Use 'as_cmap' instead of 'as_cmap'

# Create the heatmap
heatmap = sns.heatmap(Correlation, annot=True, fmt=".2f", cmap=cmap, center=0)

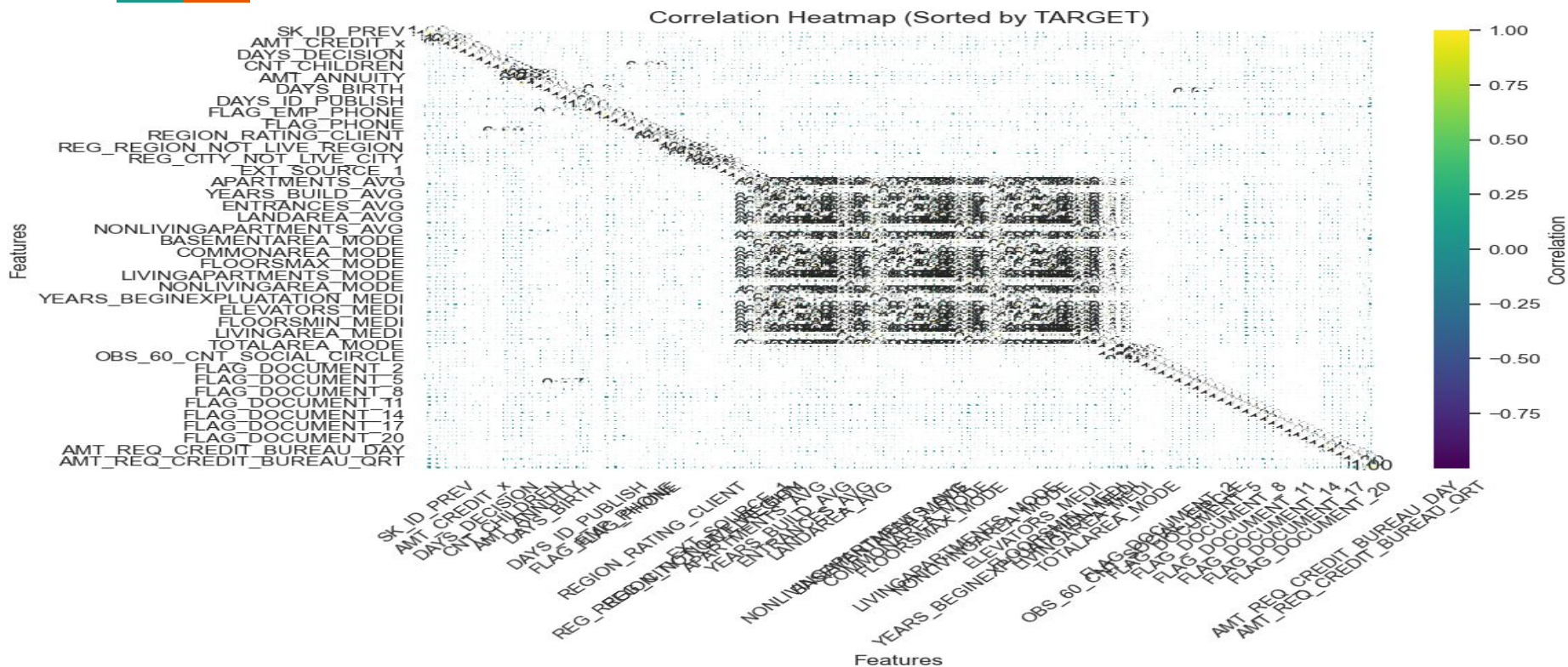
# Add Labels and title
heatmap.set_xlabel("Features", fontsize=12)
heatmap.set_ylabel("Features", fontsize=12)
heatmap.set_title("Correlation Heatmap", fontsize=14)

plt.show()
```


OUTPUT PLOT



Merging Previous and Current plot:-



Correlating here's the result:-



```
In [53]: Correlation.head(6)["TARGET"][1:]
```

```
Out[53]: SK_ID_CURR          -0.001246  
         AMT_APPLICATION    -0.005583  
         AMT_CREDIT_x       -0.002350  
         HOUR_APPR_PROCESS_START_x -0.027809  
         NFLAG_LAST_APPL_IN_DAY -0.002887  
         Name: TARGET, dtype: float64
```

```
In [54]: Correlation.tail(6)["TARGET"][1:]
```

```
Out[54]: AMT_REQ_CREDIT_BUREAU_DAY    0.005025  
         AMT_REQ_CREDIT_BUREAU_WEEK  0.001149  
         AMT_REQ_CREDIT_BUREAU_MON   -0.012606  
         AMT_REQ_CREDIT_BUREAU_QRT   -0.001526  
         AMT_REQ_CREDIT_BUREAU_YEAR  0.016432  
         Name: TARGET, dtype: float64
```

Conclusion



1. For effective payments, banks should concentrate more on contract types such as "student," "pensioner," and "businessman" with dwelling types other than "coop apartments."
2. Since working income has the highest percentage of failed payments, banks should pay less attention to this income category.
3. The number of timely payments that fail for the loan purpose "Repair" is greater as well.
4. As they have the fewest failed payments, try to attract as many consumers from the dwelling type "With Parents" as you can.