Credit Risk Assessment –EDA

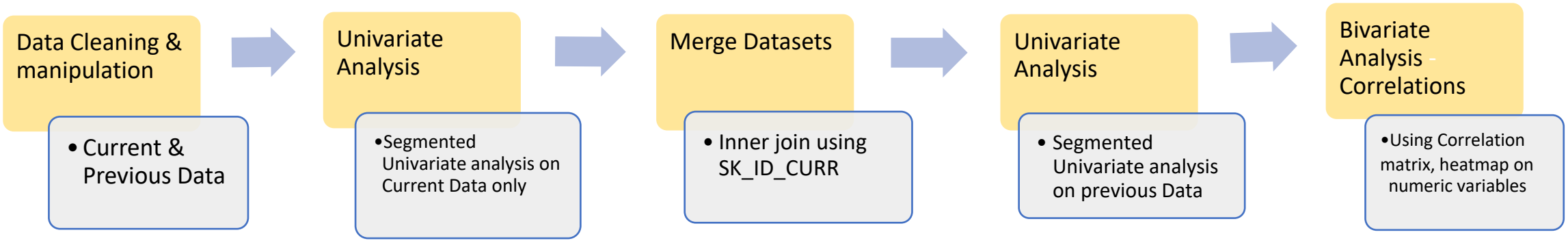
Abhik Gupta

Agenda

- Problem Statement & EDA Framework
- Data Cleaning and Manipulation
- Data Imbalance
- Univariate Analysis
  - Segmented Univariate Analysis
  - Categorical/Numerical Analysis
- Bivariate/Multivariate Analysis
  - Correlation Analysis
- Executive Summary and Recommendations

Problem Statement

Identifying the underlying factors resulting in loan defaulters and incorporating this knowledge into the Portfolio Risk Management of the company to reduce the defaulters



Data Cleaning and Manipulation (Current Data)

Assessments and dealings

49
- Columns with >40% of values missing

8
- Columns with 10-40% of values missing

Binning Numerical Columns to create a new categorical column namely :

- AMT_INCOME_RANGE
- AMT_CREDIT_RANGE
- AGE_GROUP
- EMPLOYMENT_YEAR

Box Plot to identify outliers

- AMT_ANNUITY, AMT_CREDIT
- AMT_GOODS_PRICE,CNT_CHILDREN
- AMT_INCOME_TOTAL
- DAYS_BIRTH
- DAYS_EMPLOYED

Insights and Actions

- Drop columns with missing values greater than 40%
- Converting negative days to positive days of DATES Columns.
- Nearly half of all loan applicants earn between **100k** and **200k** and nearly **90%** of the loan applicants have an income below **300K**
- Approximately **16%** of loan applicants took loans exceeding **1 million**
- There are **31%** of loan applicants who are over **50 years** of age and over **55%** who are **over 40 years of age**.
- Nearly **80%** of loan applicants have **less than 10 years** of work experience, and more than **55%** have work experience **within 0-5 years**
- Using **mode()[0],** impute the categorical variable **'NAME TYPE SUITE**,' which has a lower null percentage **(0.42 %)**, with the most frequent category
- Impute categorical variable **'OCCUPATION_TYPE'** which has higher null percentage**(31.35%)** with a new category as **'Unknown**'
- Impute **numerical variables** with the **median** on safer side
- Some outliers exist for the attributes AMT ANNUITY, AMT CREDIT, AMT GOODS PRICE, and CNT CHILDREN**(Not Conclusive**).
- Numerous outliers in AMT INCOME TOTAL indicate that only a **small fraction** of loan applicants have **high incomes relative** to the other applicants.
- The fact that DAY'S BIRTH has no outliers indicates that **the data is consistent**
- DAYS EMPLOYED has outlier values of **350000(days),** which is almost 958 years, is impossible, and necessitates the **conclusion that this entry is erroneous.**

Data Cleaning and Manipulation (Previous Data)

Assessments and dealings

11
- Columns with >40% of values missing

3
- Columns with 10-40% of values missing

Converting Categorical columns from Object to categorical

DAYS_DECISION_GROUP
NAME_CONTRACT_STATUS
NAME_CLIENT_TYPE
CODE_REJECT_REASON

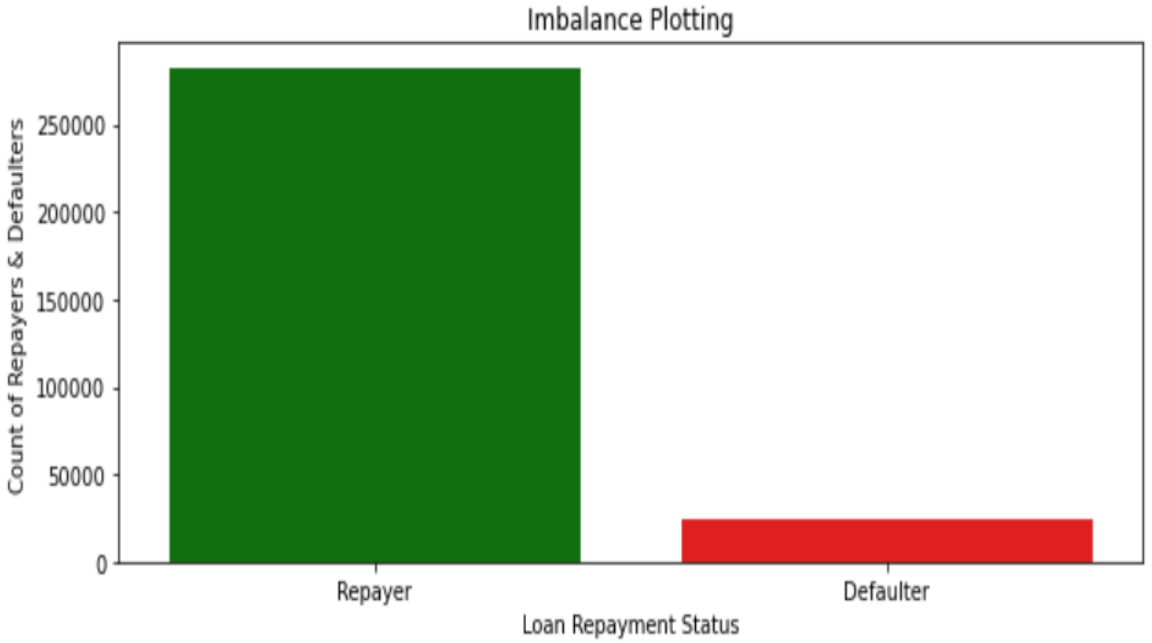
Box Plot to identify outliers

- AMT_ANNUITY, AMT_CREDIT , SELLERPLACE_AREA
- AMT_GOODS_PRICE, AMT_APPLICATION
- CNT_PAYMENT
- SK_ID_CURR
- DAYS_DECISION

Insights and Actions

- Drop columns with missing values greater than 40%
- Converting negative days to positive days of DAYS_DECISION
- Nearly **37%** of loan applicants requested for a new loan between **0 and 400 days** after the last loan decision.
- As most fields either have missing data or trash entries in them, remove rows related with applications that were cancelled.
- Imputing AMT_ANNUITY with median as mean would not be right approach
- Impute AMT_GOODS_PRICE with mode and CNT_PAYMENT with 0 as the NAME_CONTRACT_STATUS for these indicate that most of these loans were not started
- AMT GOODS PRICE, APPLICATION, CREDIT, ANNUITY, and SELLERPLACE AREA have a great deal of anomalies.
- Few outlier values exist for CNT PAYMENT
- Given that SK ID CURR is an ID column and has no outliers, and DAYS DECISION has a small number of outliers, these prior applications' judgments were likely made a long time ago.

Data Imbalance



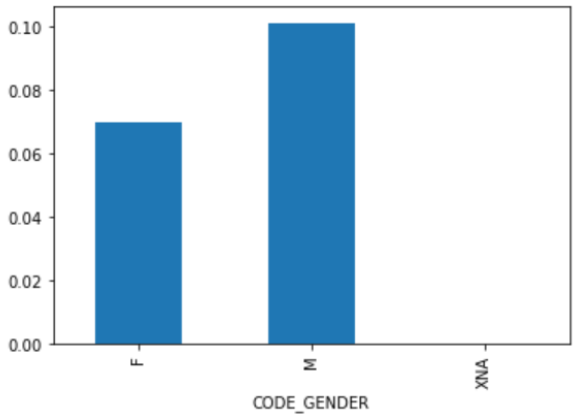
11.39	• Imbalance Ratio in %
91.93	• Non-Defaulters in %
8.07	• Defaulters in %

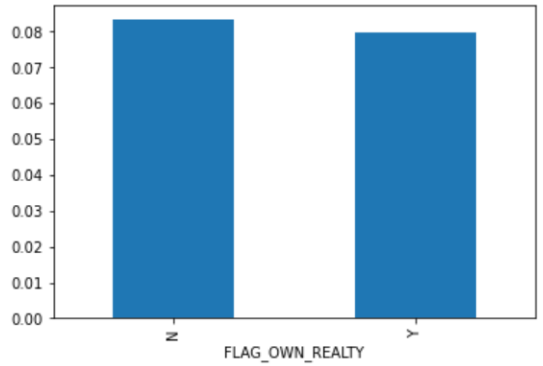
Interpretations & Approach

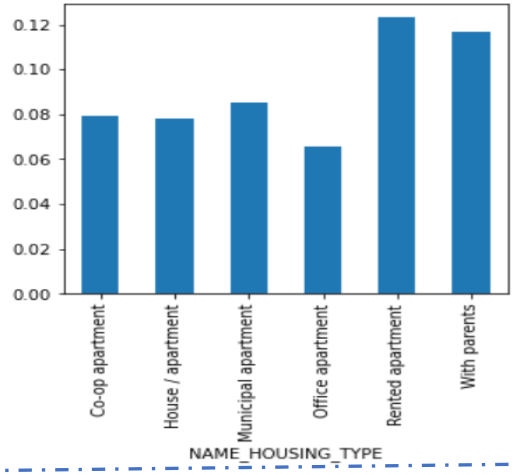
- Interpretations
  - Heavy Imbalance exists for the Target Variable.
  - **91.93%** of the data are **non-defaulters.**
  - Also, the data is significantly skewed in favor of cash loans **(90.48 %)**

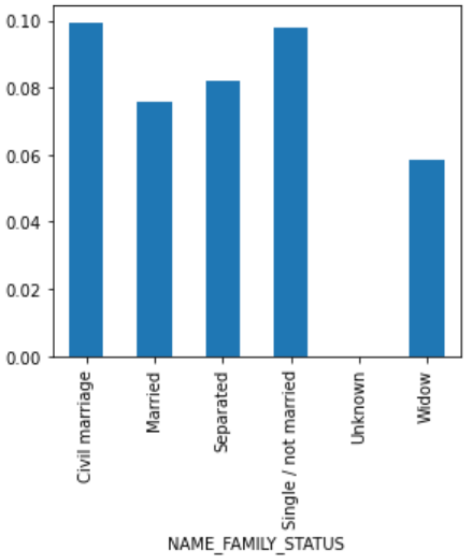
- Approach
  - To facilitate further analysis, segment the data.
  - Analyze data considering percentage as the basis

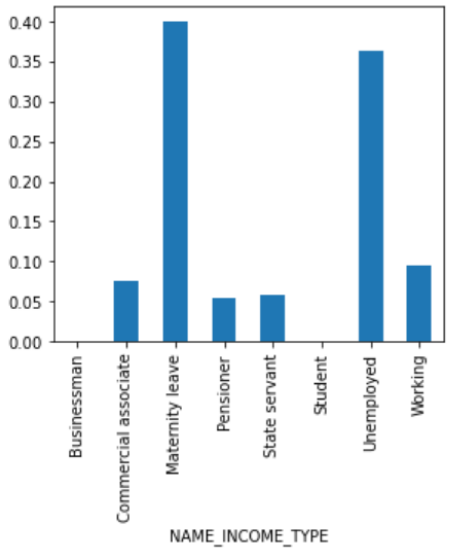
Segmented Univariate Analysis

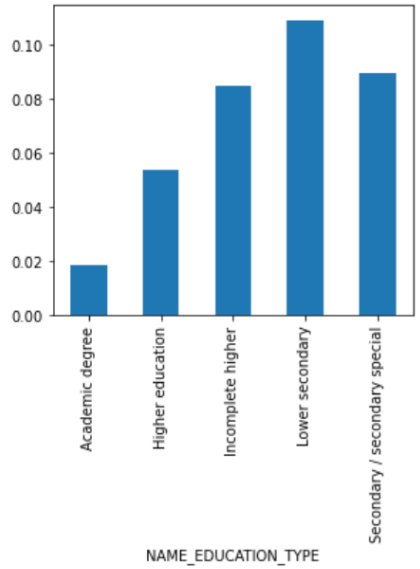








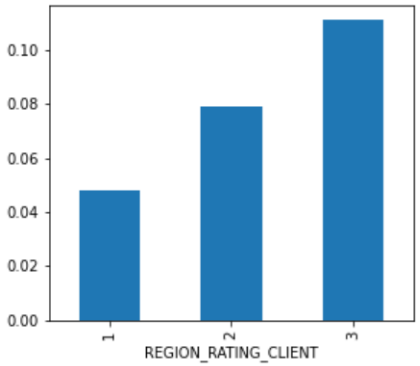


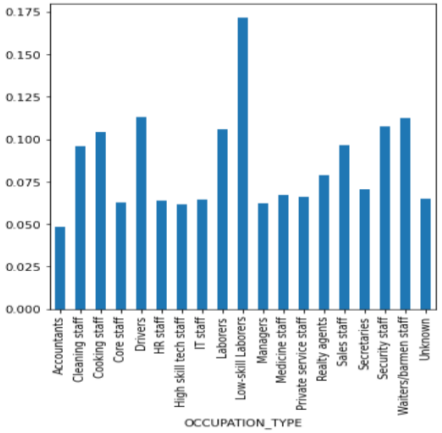


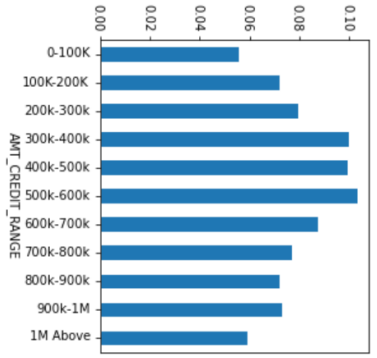
Key Takeaways

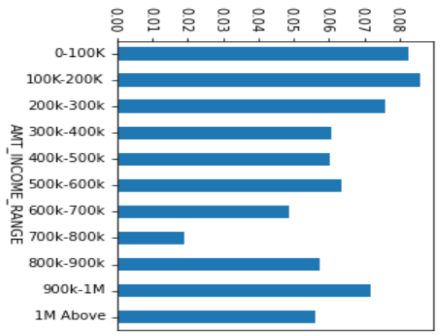
- The proportion of female clients is almost nearly twice that of male customers. According to the percentage of defaulted credits, **men are more likely than women** to **default on their loans (10%).** (7 %)
- More than double as many clients own real estate as do not. But the default rate for both categories is **almost the same (8%).** Therefore, there **is no correlation between owning a reality** and making a **loan default**.
- People who live with their **parents (11.5%)** and in **rented flats (>12%)** are more likely to **default** and people living in **office apartments** have **lowest default rate**
- **Civil marriage** and **Single/not married** Category have the highest proportion **of not repaying a loan (10%)**, with **Widow** having the **lowest rate** (exception being Unknown).
- **Maternity leave** applicants had a nearly **40% rate** of **defaulting** on loans, followed by **unemployed applicants** (37%) and **Student** and **Businessmen**, though less in numbers do not have any default record. Thus ,these two category are safest for providing loan.
- The **Lower Secondary** category has the **highest rate of loan defaults**, although being rare **(11 %**). The default rate for those with **academic degrees** is **less than 2%.**

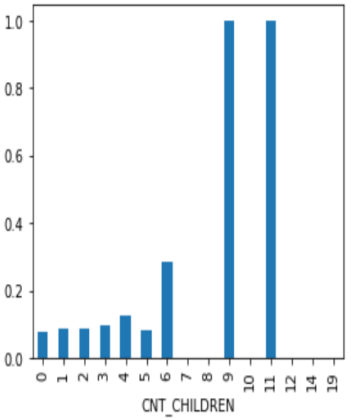
Numerical Columns like Age, Amt Credit etc. Transformed into Categorical for Analysis

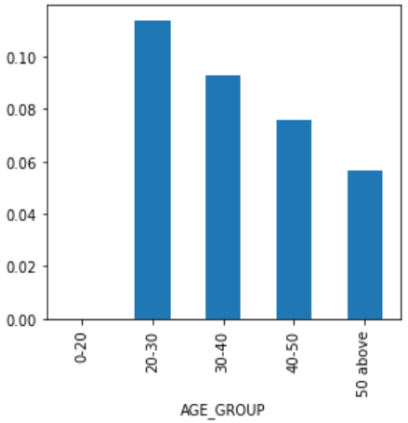


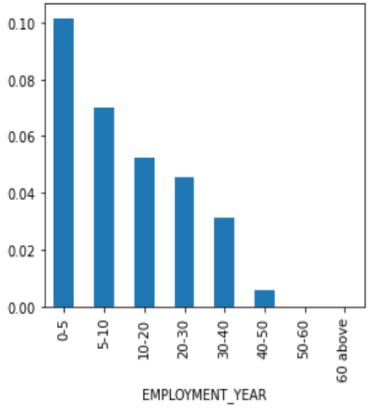


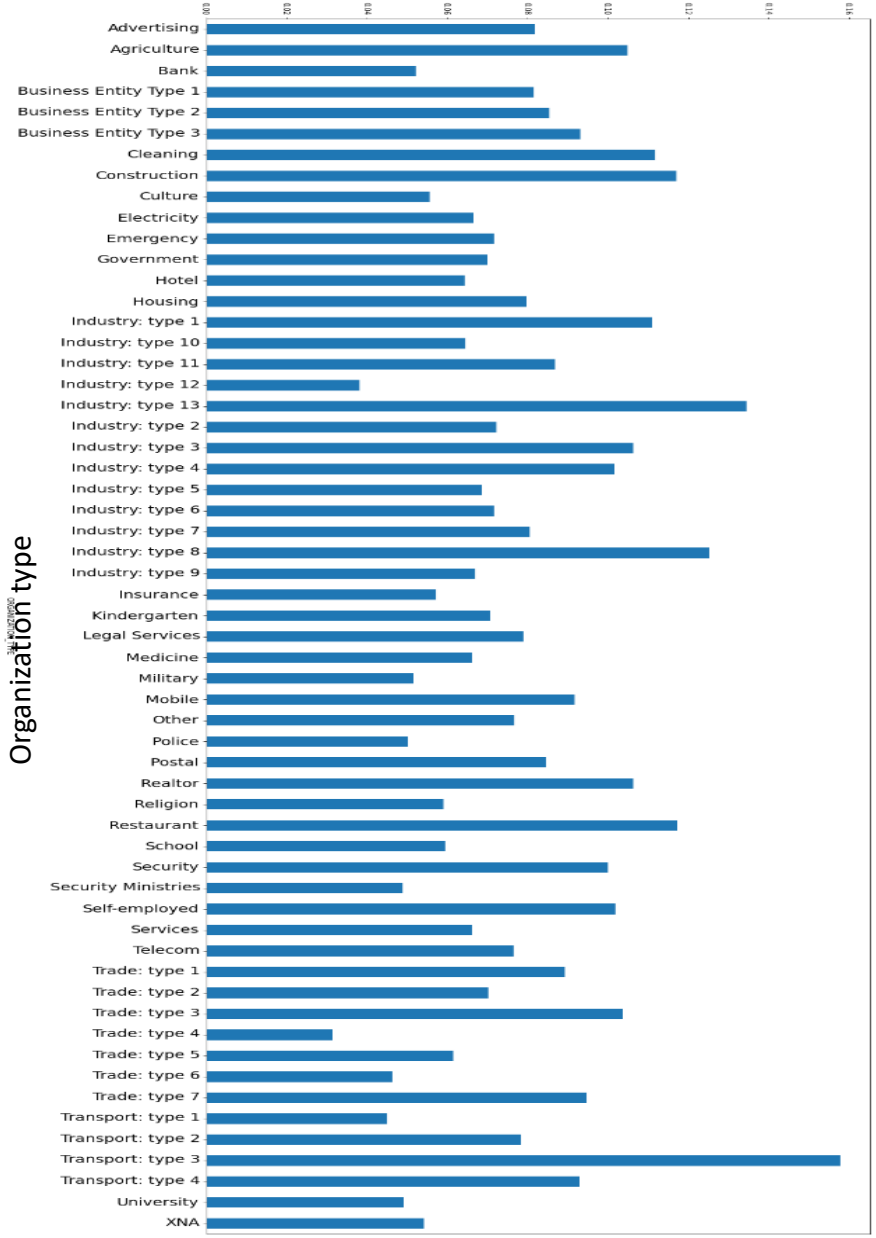










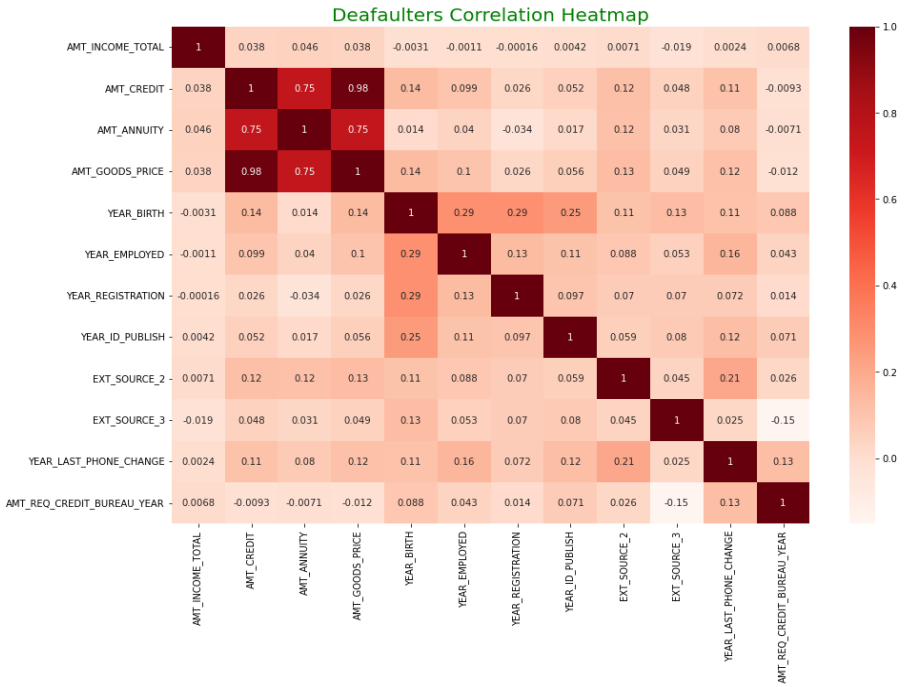


Key Takeaways from previous slide

- Region **Rating 1** applicants have the lowest likelihood **of defaulting** and Region **Rating 3** applicants have the **greatest default** rate (**11%),** making it safer to approve loans for them.
- **Low-skilled laborer's** (over **17%**) had the largest percentage **of unpaid loans**, followed by **drivers**, **waiters/bartenders**, security personnel, labourers, and kitchen personnel.
- Loans between **300K-600K** are more frequently **defaulted** upon by borrowers.
- Applications with incomes of **less than 300K** are **more likely to default** than applications with incomes of **more than 700K** are.
- People **over 50** have a **low probability o**f **defaulting**, but those between the ages of **20 and 40 have a larger probability**.
- People with **40+ years of experience** have a default rate **of less than 1%,** which is steadily declining as employment year **increases**.
- Clients with **more than 4 children** default at a **very high rate**, with kid counts of **9 and 11** exhibiting **100%** default rate.
- Transport: **type 3 (16%),** Industry**: type 13 (13.5%),** Industry: **type 8 (12.5%),** and **Restaurant** have the greatest percentage of **loans that have not been repaid** (less than 12 %). Since the **default rate** for **self-employed** people is **relatively high**, they should **either not be allowed for loans** or be given loans with **higher interest rates to reduce the risk of** default.
  - It is evident that the categories of organizations listed below have **fewer defaulters**, making them safer to lend to :
    - Industry **Type 12**
    - Trade **Type 4**

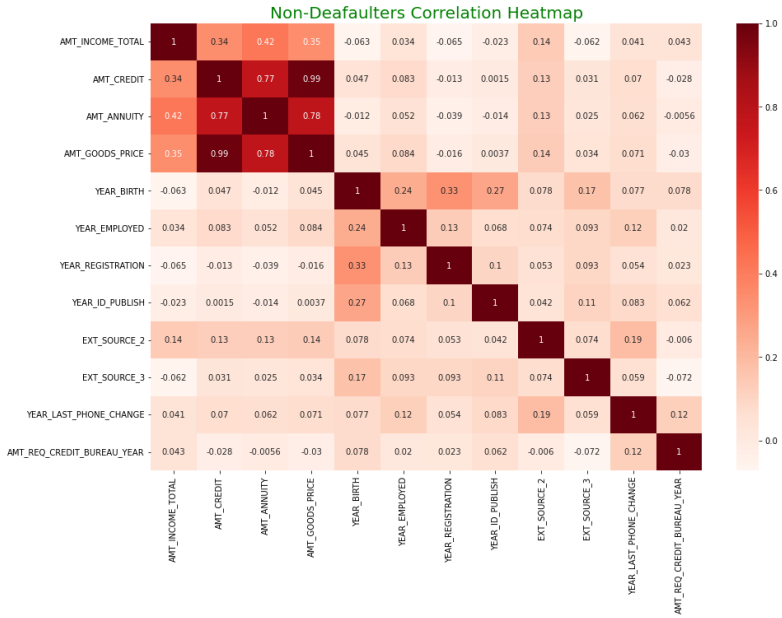
Bivariate Analysis - Correlation using heatmaps

Defaulters Heatmap



Columns Name	Columns Name	Correlation(r)
AMT_CREDIT	AMT_GOODS_PRICE	0.98
AMT_GOODS_PRICE	AMT_ANNUITY	0.75
AMT_CREDIT	AMT_ANNUITY	0.75
YEAR_BIRTH	YEAR_REGISTRATION	0.29
YEAR_EMPLOYED	YEAR_BIRTH	0.29
YEAR_BIRTH	YEAR_ID_PUBLISH	0.25
YEAR_LAST_PHONE_CHANGE	EXT_SOURCE_2	0.21
YEAR_EMPLOYED	YEAR_LAST_PHONE_CHANGE	0.16
EXT_SOURCE_3	AMT_REQ_CREDIT_BUREAU_YEAR	0.15
YEAR_BIRTH	AMT_GOODS_PRICE	0.14

Non-Defaulters Heatmap



Columns Name	Columns Name	Correlation(r)
AMT_CREDIT	AMT_GOODS_PRICE	0.98
AMT_GOODS_PRICE	AMT_ANNUITY	0.75
AMT_CREDIT	AMT_ANNUITY	0.75
YEAR_BIRTH	YEAR_REGISTRATION	0.29
YEAR_EMPLOYED	YEAR_BIRTH	0.29
YEAR_BIRTH	YEAR_ID_PUBLISH	0.25
YEAR_LAST_PHONE_CHANGE	EXT_SOURCE_2	0.21
YEAR_EMPLOYED	YEAR_LAST_PHONE_CHANGE	0.16
EXT_SOURCE_3	AMT_REQ_CREDIT_BUREAU_YEAR	0.15
YEAR_BIRTH	AMT_GOODS_PRICE	0.14

Key Takeaways

Observations

- Heatmaps for **defaulters** and **non-defaulters** are displayed separately.
- **Top 3** correlations are **consistent** and strong for **defaulters** and **non-defaulters**.
- **Mild** correlation exists for the **next 3** items in the non defaulter's dataset
- Weak correlation exists for the next 3 items in the defaulter's dataset

Insights

- The customer's credit request is directly proportional to the price of the goods they intend to purchase, and thus the associated annuity.
- Non-defaulters have a slightly higher correlation **(Annuity-Credit)**, indicating **that Annuity is higher for lower credit** when **default occurs**.
- The non-defaulters dataset shows that a person's prospects of repaying the loan are better **if their income is proportionally higher** to their **annuity , credit**.

Executive Summary and Recommendations

Executive Summary	Recommendations
➢ Significant loans are reportedly given to self-employed, less educated, and labourers who live in sparsely populated areas. These loans frequently default, on average.	➢ To reduce non-performing assets, have a better mix of educated people living in urban and semi-urban areas.
➢ Women make up a higher proportion of loans and have a lower defaulting rate.	➢ To encourage greater uptake among them, loans may be made available at a discounted rate.
➢ The rate of default is typically higher in younger persons.	➢ Perform additional due diligence procedures, such as examining recent employment changes, cell number changes, and identification documents.
➢ When people leave their registered city, there appears to be a higher default.	➢ Before approving such loans, additional due diligence checks, such as additional support documents and collateral as security, can be undertaken.
➢ 90% of applications have a total income of less than 300K and a high probability of defaulting.	➢ They may be offered a loan with a higher interest rate than other income groups.
➢ The default rate for clients with 4 to 8 children is extremely high.	➢ Their loans should have higher interest rates.
➢ 90% of the clients who were previously dropped have really repaid the loan.	➢ The bank may be able to assess and negotiate conditions with these customers who are repaying their loans in the future, so broadening its commercial opportunities, if it keeps track of the cancellation's reasons.
➢ 88 percent of the customers whose loan requests were previously denied by the bank have now turned to repaying customers.	➢ By stating the reason for the denial, the business impact may be minimized, and these clients could be contacted about additional loans.