Credit Risk Analysis

-Papri Boyra

Introduction

- Problem Statement: Analyze financial and demographic attributes of loan applicants.
- Objective: Gain insights into factors influencing loan default rates.
- Goal: Develop strategies to mitigate risks associated with lending.
- Approach: Conduct a thorough EDA to identify key trends and patterns.

Problem Statement

  - Lending companies face challenges in approving loans due to insufficient credit history.
  - Need to identify patterns in data to minimize the risk of loan defaults.
  - Two key risks:
    - Rejecting creditworthy applicants results in lost business.
    - Approving loans to defaulters leads to financial loss.
- Briefly explain: The need to use EDA to understand the driving factors behind loan default.

Business Objectives

- Identify patterns indicating clients' difficulty in repaying loans.
- Use insights to take actions such as:
  - Denying loans to high-risk applicants.
  - Adjusting loan amounts.
  - Lending at higher interest rates for risky applicants.
- Ensure that creditworthy applicants are not rejected.
- Understand the driving factors behind loan defaults.

Objectives

Importing necessary Modules:

Import the modules necessary for Data Manipulation and Visualization.

Reading dataset:

Read the dataset containing loan applicant information.

Task 1 - Exploring the Dataset:

Understand the Structure and various datatypes of the attributes within the dataset.

Task 2 - Missing value analysis:

Identify and analyze missing values in the dataset.

Task 3 - Analysing categorical and numerical columns:

Analyze categorical and numerical columns to understand the statistical properties and relationships within the dataset.

Task 4 - Univariate Analysis:

Conduct univariate analysis to explore the distribution and characteristics of individual variables.

Task 5 - Outliers:

identify and analyze outliers within the dataset to understand their impact on the analysis.

Task 6 - Merging Datasets:

Identify and merge different Datasets for further analysis.

Task 7 - Bivariate analysis:

Conduct bivariate analysis to explore relationships between different variables and their impact on loan default rates.

Data Understanding

- `application_data.csv`:
  - Contains client information at the time of application.
  - Indicates whether a client has payment difficulties.
- `previous_application.csv`:
  - Contains information about clients' previous loan applications.
  - Indicates if previous applications were Approved, Cancelled, Refused, or Unused offer.
- `columns_description.csv`:
  - Provides a data dictionary explaining the meaning of variables.

Import Libraries

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- import seaborn as sns
- sns.set(color_codes=True)

Imports necessary libraries for data analysis and visualization.

Suppress Warnings

- import warnings
- warnings.filterwarnings('ignore')

Filters out warning messages to keep the output clean.

Overall Approach

- **Data Cleaning:**
  - **Handling missing values by removing columns with excessive NaN values.**
  - **Converting days to years for time-related features.**
  - **Outlier detection and handling in quantitative columns.**
- **Data Imbalance:**
  - **Analysis of imbalance in the target variable (payment difficulty).**
- **Exploratory Data Analysis:**
  - **Univariate analysis (distributions of single variables).**
  - **Segmented univariate analysis (distributions within groups).**
  - **Bivariate analysis (relationships between two variables).**
  - **Correlation analysis (identifying strongly correlated variables)**

Data Exploration

  - Dataset contains loan applicant information.
  - Inspected dimensions, data types, and records.
  - Observed the number of rows and columns.
- Visual: A simplified table showing key columns from `application_train.head()`.

Missing Value Analysis

  - Identified columns with high missing value percentages.
  - Removed columns with > 50% missing values.
  - Remaining missing values in `OCCUPATION_TYPE`, `EXT_SOURCE_3` etc, will be analyzed further.
- Visual: A bar chart showing missing value percentages for the top 5-10 columns.

As we can observe, there are lot of columns with missing values. There are some columns which has missing values around or more than 50%. Other columns has significantly less missing value. Also, the columns for which has missing values are around or more than 50% are mostly either mean, median or mode. So, there is no way one can replace these missing data. So, we will not consider these columns for analysis. We will consider other columns for analysis. Let's analyse the other columns.

Data Cleaning - Outlier Treatment

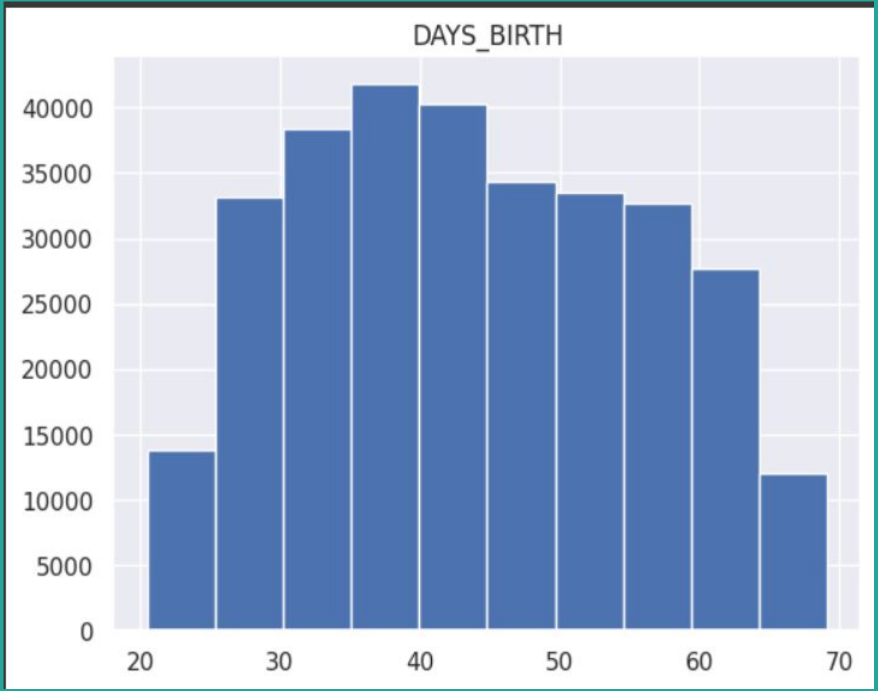
- Box plots used to visualize outliers in numerical columns such as 'AMT_INCOME_TOTAL', 'AMT_CREDIT', etc.
- Outliers were capped using the interquartile range (IQR) method.
- Visualizations before and after cleaning were done to show the impact of outlier removal

Data Imbalance

- The target variable 'TARGET' is imbalanced with a high number of non-defaulting clients.
- The imbalance ratio will be highlighted.
- Bar plots showing the imbalance in counts and percentages of target values (0 and 1)
- Mention the Imbalance Ratio

Univariate Analysis - Age

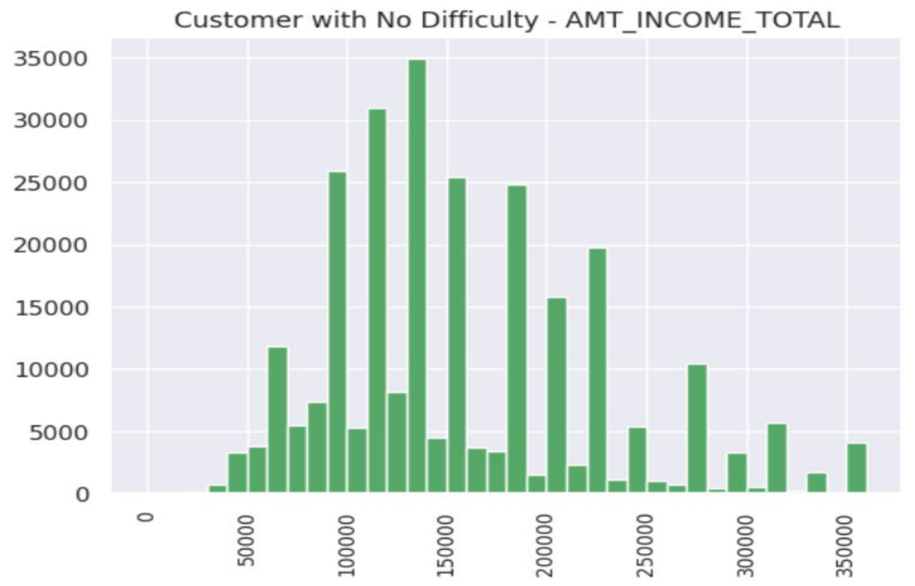
- Average age of applicants around 44 years
- Briefly mention the range of the age

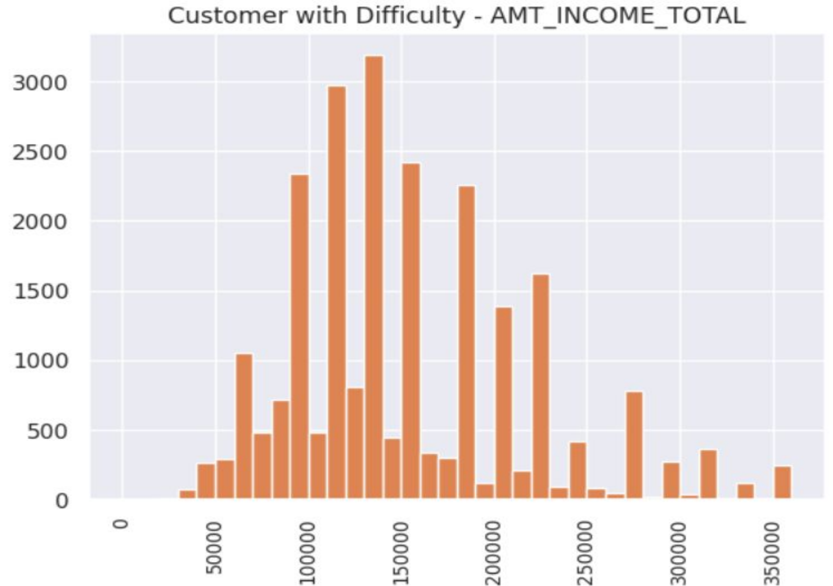


Univariate Analysis - Income and Credit

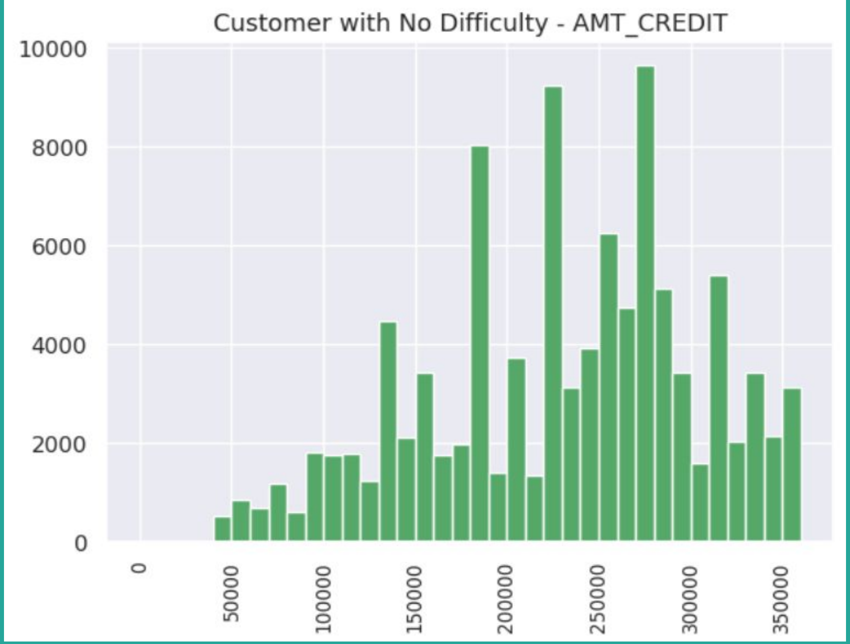
- Distributions of income and credit amounts
- Briefly mention the range of income and credit

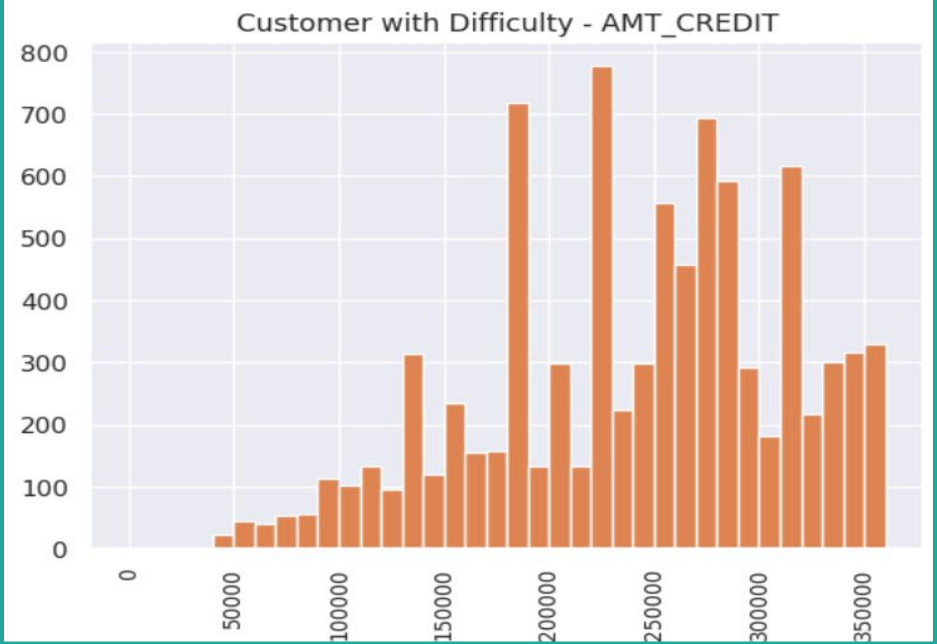
UNIVARIATE ANALYSIS - AMT_INCOME_TOTAL





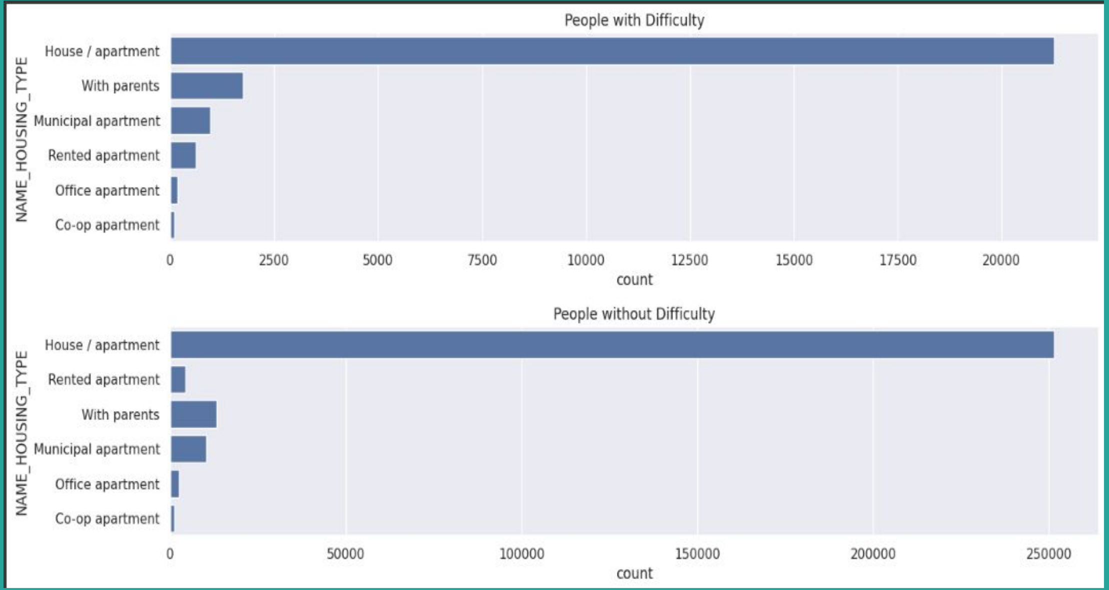
UNIVARIATE ANALYSIS - AMT_CREDIT





Segmented Univariate Analysis

- **Visualizations:**
  - **Count plots of 'NAME_HOUSING_TYPE', 'NAME_FAMILY_STATUS', 'NAME_EDUCATION_TYPE', 'NAME_INCOME_TYPE', 'ORGANIZATION_TYPE' for both 'TARGET = 0' and 'TARGET = 1' separately.**



- **Highlight the difference between the two target categories based on each categorical variable above.**

Categorical Data Insights

- Examined relationships between categorical features and loan defaults.
- Higher default rates noted in specific categories (gender, income, education).
- Significant default rate differences observed between subgroups.

Univariate Analysis (Numerical)

- Analyzed numerical feature distributions.
- Visualized with histograms and distribution plots.
- Identified key numerical columns that show difference in distributions between target 0 and 1.

Remaining Missing Values

- Calculated missing value percentages for remaining columns.
- Identified columns like `OCCUPATION_TYPE`, `EXT_SOURCE_3`.

Feature Engineering (Dates)

- Transformed `DAYS_BIRTH`, `DAYS_REGISTRATION`, `DAYS_ID_PUBLISH` to years.
- Used `-round()/365` method.

Data Balance

- Calculated ratio of `TARGET=0` to `TARGET=1`.
- Observed significant class imbalance.
- Ratio = `(train['TARGET'] == 0).sum() / (train['TARGET'] == 1).sum()`

Categorical Analysis Setup

- Created subsets: `train_0` (`TARGET=0`) and `train_1` (`TARGET=1`).
- Defined `plotting` function for pie, count, and bar plots

Univariate Analysis (Categorical)

One slide for each of the following:

- NAME_CONTRACT_TYPE
- CODE_GENDER
- FLAG_OWN_CAR
- FLAG_OWN_REALTY
- NAME_TYPE_SUITE
- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- NAME_FAMILY_STATUS
- NAME_HOUSING_TYPE

Categorical Insights

- Gender, income type, education, family status impact loan defaults.
- Males have higher default rate.
- Pensioners show lower default rates.
- Secondary education has higher default rates compared to higher education.
- Married individuals have lower default rates than single and civil marriage.

Numerical Analysis Setup

- Selected numerical columns.
- Prepared to perform univariate and outlier analysis.

Outlier Analysis (Scatter)

One slide each for:

- CNT_CHILDREN
- AMT_INCOME_TOTAL
- FLAG_MOBIL
- OBS_30_CNT_SOCIAL_CIRCLE
- DEF_30_CNT_SOCIAL_CIRCLE
- OBS_60_CNT_SOCIAL_CIRCLE

Outlier Insights

- Observed potential outliers in CNT_CHILDREN, AMT_INCOME_TOTAL.
- No outliers were removed, they are kept for further analysis

Numerical to Categorical Conversion

- Converted AMT_ANNUITY to categories: low, medium, high, very high.
- Using a defined function with conditionals.

Univariate Analysis (Numerical, Histograms)

- Content: Include one slide for each of the following columns (numerical), including histograms and distribution plots, along with key observations.
  - AMT_CREDIT
  - AMT_ANNUITY
  - AMT_GOODS_PRICE
  - DAYS_BIRTH
  - DAYS_EMPLOYED
  - DAYS_REGISTRATION
  - DAYS_ID_PUBLISH
  - HOURS_APPR_PROCESS_START
  - EXT_SOURCE_2
  - EXT_SOURCE_3
  - AMT_REQ_CREDIT_BUREAU_YEAR
- Visual: Histograms and distribution plots separated by TARGET

Key Numerical Insights

- Identified columns that exhibit differences in distribution for TARGET=0 and TARGET=1.
  - The columns are AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, DAYS_BIRTH, HOURS_APPR_PROCESS_START, EXT_SOURCE_2, EXT_SOURCE_3, and AMT_REQ_CREDIT_BUREAU_YEAR.

Previous Application Data Insights

- **Noted duplicate `SK_ID_CURR` values, indicating multiple loans per person.**
- **`SK_ID_PREV` is unique.**

Merging Datasets

- **Merged train and previous applications using `SK_ID_CURR`.**
- **Duplicated `SK_ID_CURR` values were kept to study repeat borrowers.**

Merging Data Frames: Train and Previous Application Based on SK_ID_PREV

After merging both data frames using the SK_ID_PREV column as the key, the resulting dataframe will also contain duplicate SK_ID_PREV values. This duplication is not an issue, as our objective is to explore patterns, including cases where a lender has previously taken out a loan more than once. Retaining these duplicates allows us to analyze the data comprehensively and identify any recurring trends or behaviors among borrowers with multiple loan histories.

Bivariate Analysis

- **Visualizations:**
  - **Joint plot of 'AMT_CREDIT' vs 'AMT_INCOME_TOTAL'**
  - **Joint plot of 'DAYS_BIRTH' vs 'AMT_CREDIT'**
  - **Joint plot of 'DAYS_EMPLOYED' vs 'AMT_CREDIT'**
  - **Joint plot of 'CNT_CHILDREN' vs 'AMT_CREDIT'**

- **Explanation**: Analyzing relationships between variables provides insights into how different attributes together can affect loan defaults.

Bivariate Analysis

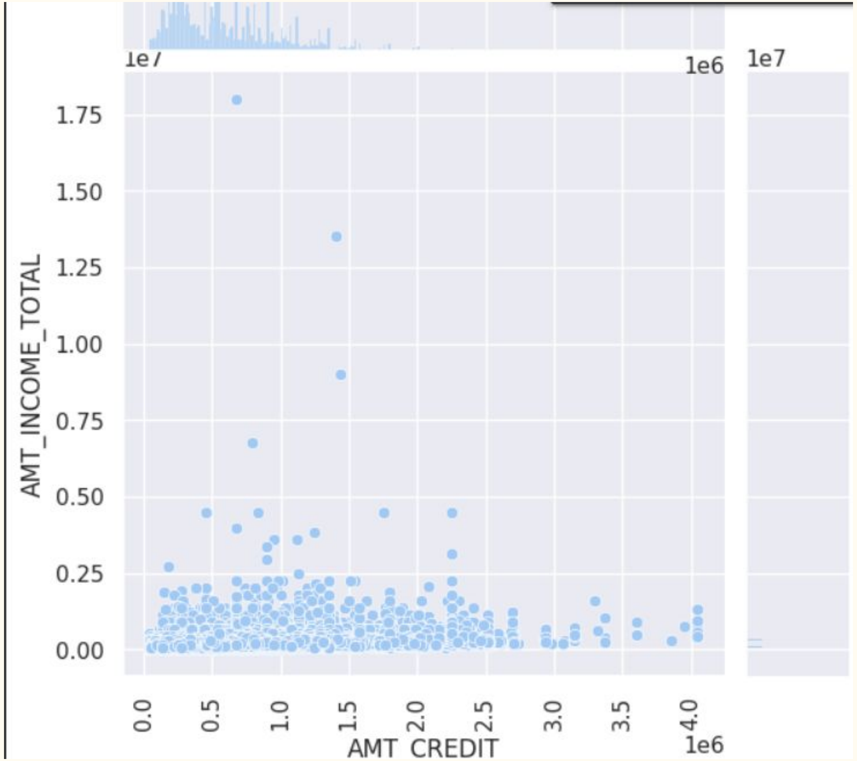
- Title: Bivariate Analysis - Key Findings

    - People with 'Secondary special' education tend to apply for more loans, which are also often approved.
    - Married individuals are more likely to repay their loans compared to single individuals based on the loan approval status.

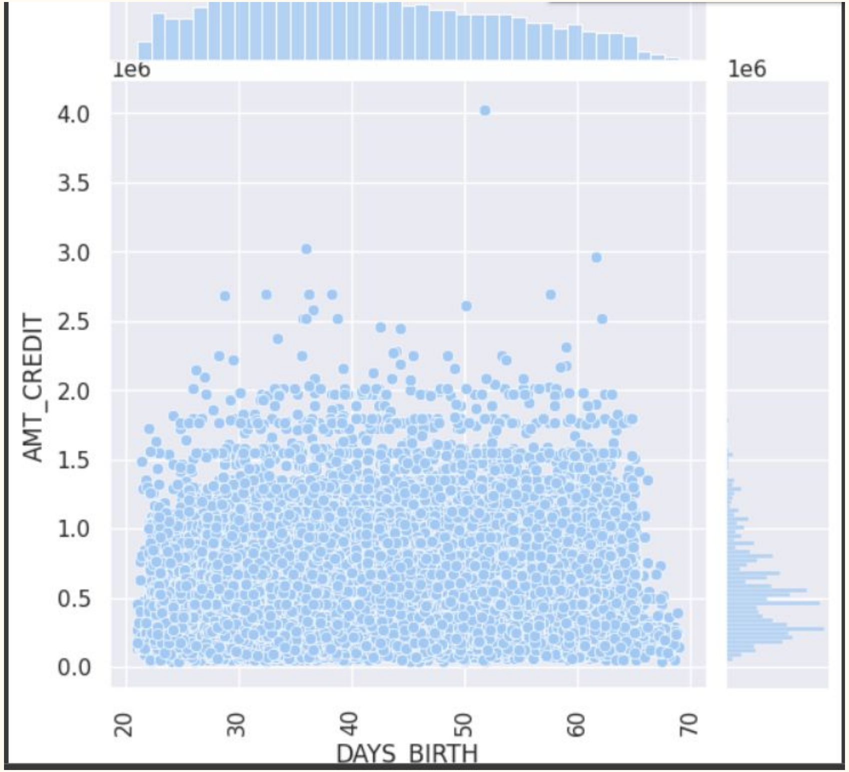
- Include one slide for each of the bivariate analysis and show plots with the key observations:
    - NAME_EDUCATION_TYPE and NAME_CONTRACT_STATUS

    - NAME_FAMILY_STATUS and NAME_CONTRACT_STATUS

Bivariate Analysis

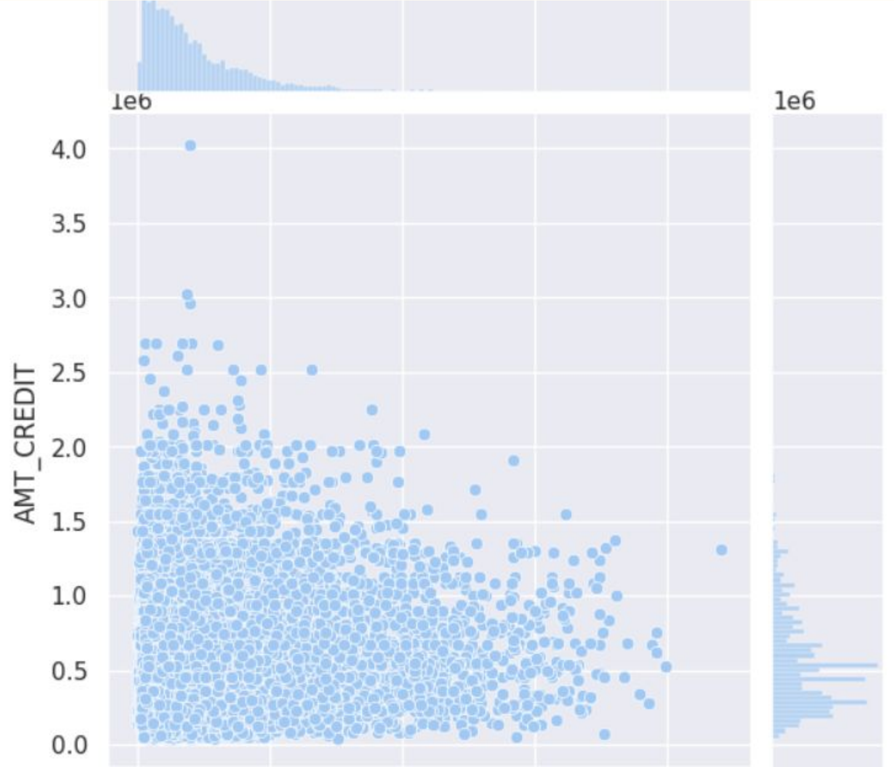


- Joint plot of 'AMT_CREDIT' vs 'AMT_INCOME_TOTAL'

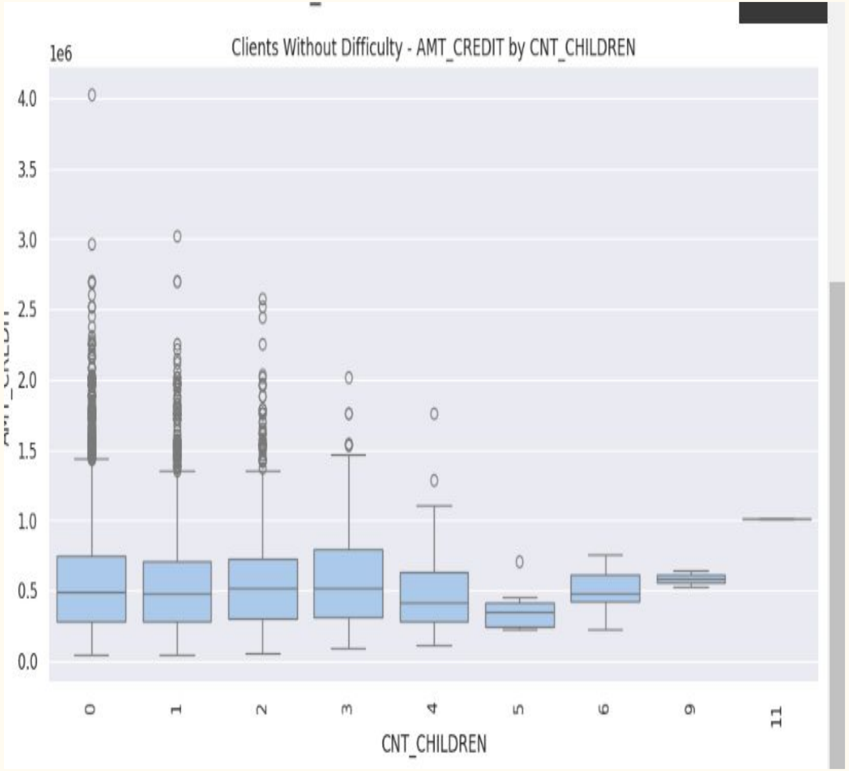


- Joint plot of 'DAYS_BIRTH' vs 'AMT_CREDIT'

Bivariate Analysis



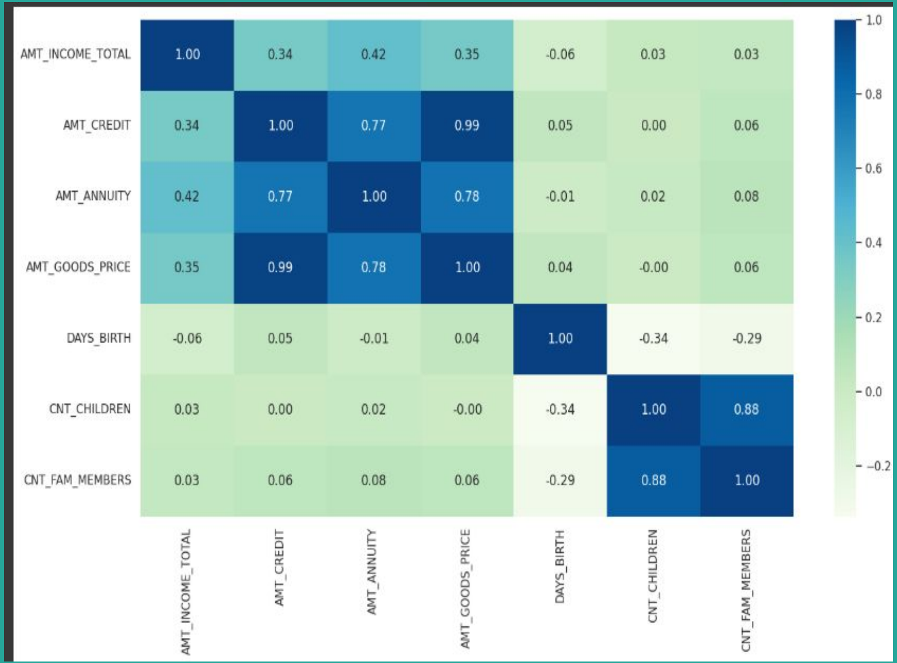
Joint plot of 'DAYS_EMPLOYED' vs 'AMT_CREDIT'



Clients Without Difficulty - AMT CREDIT by CNT CHILDREN

Correlation Analysis

- **Visualizations:** Heatmaps of correlation matrix for 'TARGET = 0' and 'TARGET = 1' dataframes separately
- **Bullet Points:**
  - Identify the top correlations between quantitative variables for each target.



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	CNT_CHILDREN	CNT_FAM_MEMBERS
AMT_INCOME_TOTAL	1.00	0.34	0.42	0.35	-0.06	0.03	0.03
AMT_CREDIT	0.34	1.00	0.77	0.99	0.05	0.00	0.06
AMT_ANNUITY	0.42	0.77	1.00	0.78	-0.01	0.02	0.08
AMT_GOODS_PRICE	0.35	0.99	0.78	1.00	0.04	-0.00	0.06
DAYS_BIRTH	-0.06	0.05	-0.01	0.04	1.00	-0.34	-0.29
CNT_CHILDREN	0.03	0.00	0.02	-0.00	-0.34	1.00	0.88
CNT_FAM_MEMBERS	0.03	0.06	0.08	0.06	-0.29	0.88	1.00

Conclusion

Title: Project Summary and Conclusion

- Successfully explored the dataset, identified key factors influencing loan defaults.
- Found gender, income type, education, marital status, and occupation to be significant predictors of loan repayment.
- Outliers were identified and flagged for further analysis.
- Merged datasets and performed bivariate analysis to get deeper insights.

Future Steps

- Develop predictive models.
- Refine loan approval criteria.
- Further explore other relevant features

Q&A

- **Open floor for question and answer.**

Thank you