Credit Exploratory Data Analysis: Case Study – Assignment

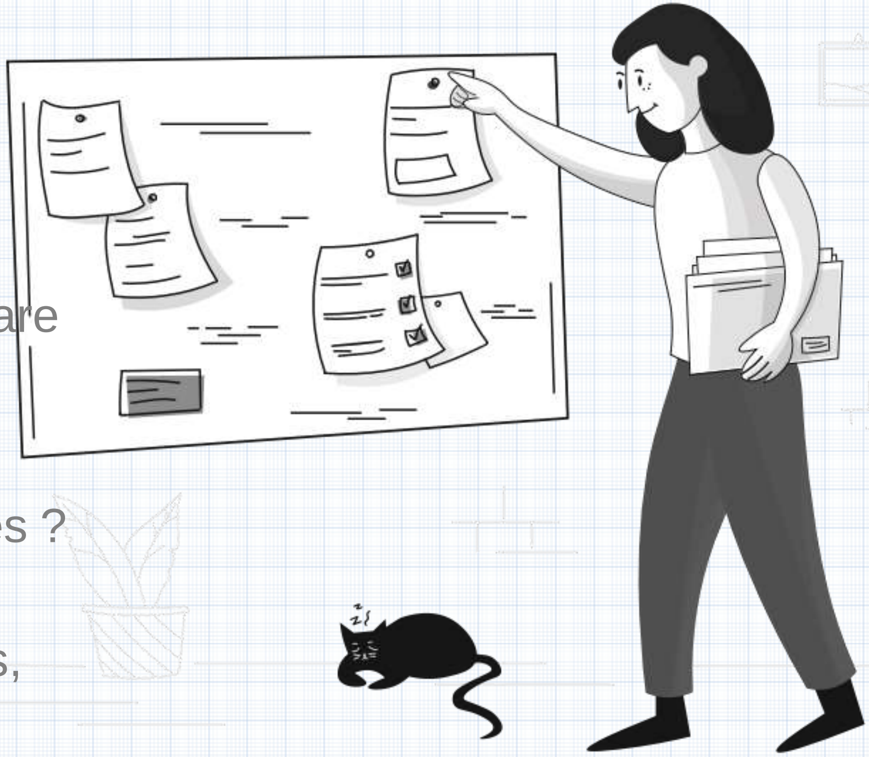
International
Institute of Information
Technology Bangalore

Amrit Pritam Sangramsingh
31 October 2023

THINGS I DO !

UNDERSTAND DATA:

- How many Rows and Columns and columns are there?
- What are the Data-types?
- How many columns are having numeric values ? Positive numbers / Negative Numbers
- re the data in correct format? (ie: Days, Hours, Months, Years. etc.)

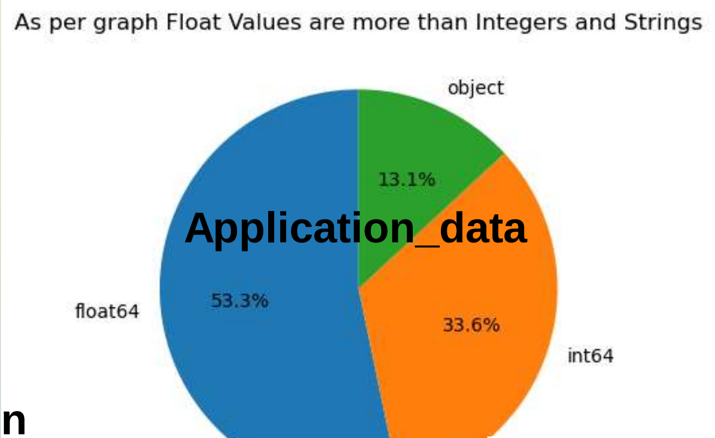


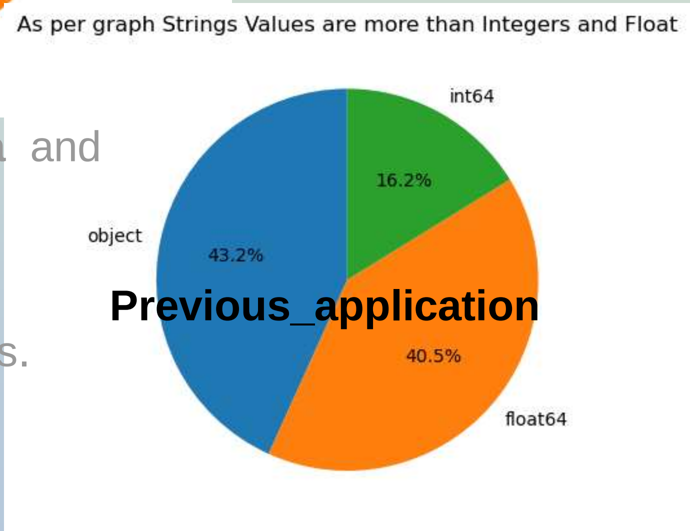
Two Data Files

Application_data

and

Previous_application





- Summary about data file read & understanding:
- There are: **3,07,511** Rows & **122** Colum s In **Application_data** and **37** columns and **1670214** rows in **Previous_application**.
- There columns having negative, positive and null values
- Days are not in proper format
- 3 types of datatype s available: Integers, Float values, Strings.

Steps Followed

Identify Anomalies

- Remove columns with more than 50% null values

Handling Missing Values

- Mean, median, mode
- Create new category
- In-place with 0

Standardize Values

- Data Scaling
  - Remove -Ve (use abs())

Uni-variate Analysis

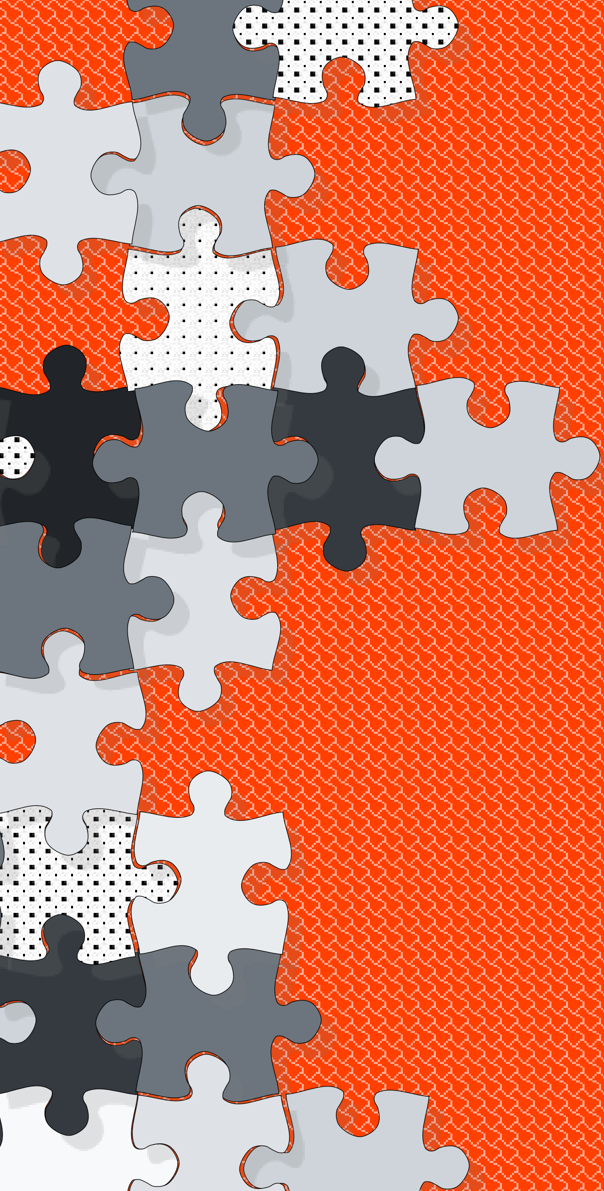
- Manage Imbalance Data

Bivariate Analysis

- Define Objectives
- Visualization
- Statistical Analysis

Handle Outliers

- Outlier Removal
- Data Transformation



Application Data

“

### States that: **Application_Data.csv**:

- There are: 307511 Rows & 52 Colums.
- 3 types of datatypes available:
    - Integers,
    - Float values,
    - Strings.
- Removed unwanted columns and other columns.
- We have worked on the negative values and converted them into positive values in some of columns.
- We have converted days and hour in proper format.

”

Previous Application Data

“In approved category, consumer loan has largest no of applicants.
There seem to be no canceled loans in cash loan category than consumer loan.
More cash loans have been refused than consumer loans.
The bank has more repeaters in all approved, refused, unused, canceled categories
POS transactions seem to be consumer loans and similar to point more cash loans have been refused than POS.
In approved category, consumer loan has largest no of applicants.
There seem to be no canceled loans in cash loan category than consumer loan.
More cash loans have been refused than consumer loans.
The bank has more repeaters in all approved, refused, unused, canceled categories
POS transactions seem to be consumer loans and similar to point 2 - more cash loans have been refused than POS.”

Key Steps

Identify Anomalies

Start by identifying and understanding the types of anomalies in the data. This include missing values, outliers, duplicates, inconsistent formats, and errors.

Data Transformation

Normalize or scale data as needed. Common transformations include standardization, log transformations, or feature scaling.

Maintain Data Quality

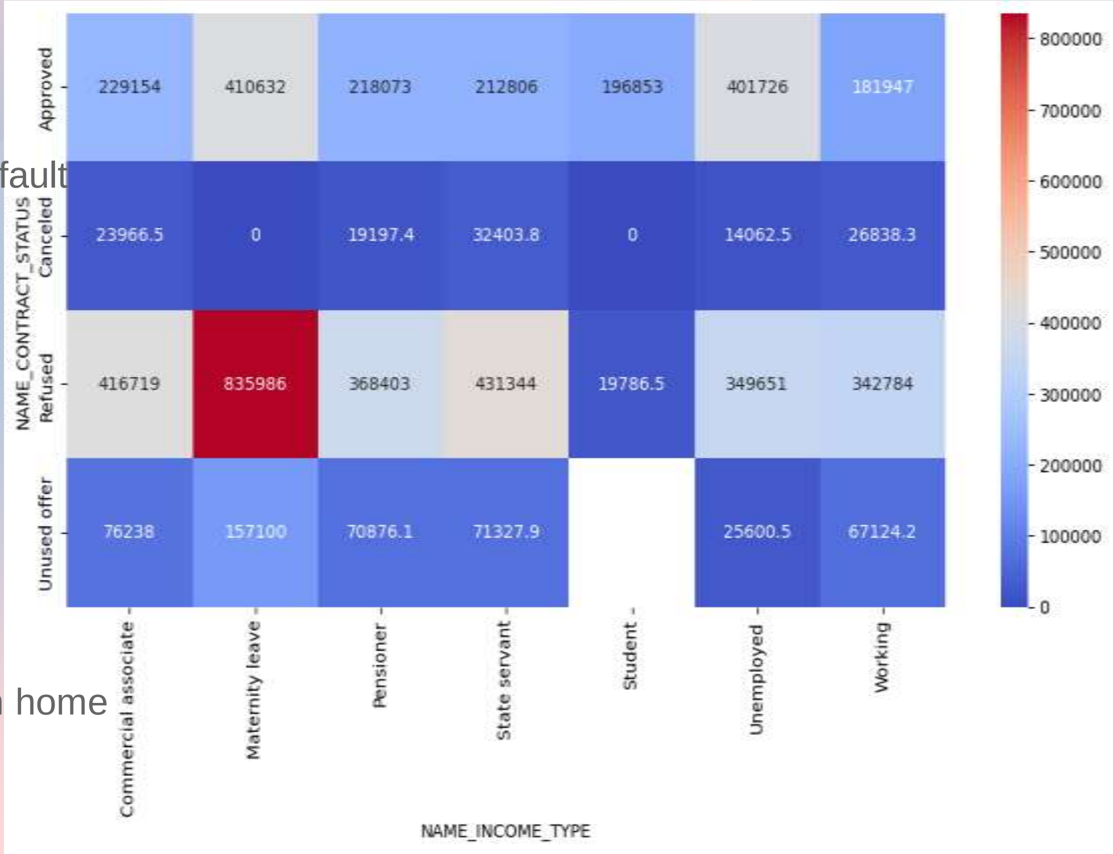
Regularly monitor and maintain data quality, especially when dealing with ongoing data streams or databases.

Case Summary

Default cases in Approved Applications:
All the below variables were established in analysis of Application dataframe as leading to default.
Checked these against the approved application and default cases and it proves to be correct

High Chance To Be Default :

- 'AMT_INCOME_RANGE' - 1L-2L
- 'AGE_GROUP' - 30-35, followed by 35-40
- 'CODE_GENDER' - Female
- 'NAME_INCOME_TYPE' - Working
- 'OCCUPATION_TYPE' - Laborers 23%
- 'ORGANIZATION_TYPE' - Business type 3
- 'OWN_CAR_flag' - 31% don't have car
- 'OWN_REALTY_flag' - 70% don't have own home


NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
NAME_INCOME_TYPE