Deliverable 1 — Dataset Selection and Initial Exploration Purpose: Build a solid foundation for your project by carefully choosing a dataset, understanding it deeply, and creating a clear plan for the modelling stages ahead. Requirements

1. Dataset Choice o Select one dataset challenge from the Kaggle pool (see below). o Explain why your team chose this dataset: interest, feasibility, data type, potential for learning, etc.

2. Data Dictionary o Document the features (input variables) and target variable. o Describe each feature in your own words using the Kaggle dataset description or metadata from other sources. o Include units (if applicable) and note categorical vs. numerical variables.

3. Exploratory Data Analysis (EDA) o Summarize dataset size and structure. o Provide descriptive statistics for numerical features. o Show distributions and visualizations of key variables. o Check for missing values, outliers, or anomalies. o Explore correlations or relationships among features. o Check for class balance (classification) or target distribution (regression).

4. Challenges and Strategies to Identify potential difficulties based on the EDA carried out in step 3 (e.g., missing values, outliers, high dimensionality, multicollinearity, imbalance, noise, etc.). o Propose and apply potential strategies to address them (e.g., imputation, scaling, feature selection, etc.).

5. Individual Contributions of each group member must contribute significantly. o Clearly document contributions (e.g., who wrote the EDA code, who prepared visualizations, who drafted the plan, etc.).

Dataset Choice to Select one dataset challenge from the Kaggle pool (see below). o Explain why your team chose this dataset: interest, feasibility, data type, potential for learning, etc.

For this group assessment, we chose the Home Credit Default Risk dataset for modeling and analysis. One of the main reasons for our choice was the dataset's richness and complexity. It includes a primary dataset containing the target variable, Default, along with numerous features that can help predict it. Additionally, there are several auxiliary datasets, such as historical credit data from CBA, which adds realism by requiring us to consider both current and past data—introducing a data integration challenge.

Many of the features are continuous numerical variables, which may require standardization or scaling, particularly if we use imputation techniques like KNN. While the dataset does not have missing values—which might seem advantageous—it also limits our ability to apply advanced imputation methods, such as Iterative Regression, which could have allowed us to estimate more realistic values for incomplete data. However, since not all features are important for such techniques, their absence may not significantly impact accuracy.

Given the large number of variables, feature selection will be critical. Failure to identify the most relevant ones could degrade model performance. Furthermore, high correlation among features suggests potential multicollinearity, which we must address to improve model interpretability and robustness.

Lastly, the topic itself has strong real-world relevance. Most people eventually take out credit, whether through credit cards or mortgages, making default prediction highly valuable for financial institutions. Banks rely on such models to assess credit risk and ensure timely repayments, making this project not only technically engaging but also practically meaningful.

CSV Files and Data Files Given
https://www.kaggle.com/competitions/home-credit-default-risk/data?select=previous_application.csv

CSV Files:

1) Application{train|test.csv} —> Already splits in Train and Test ds just need to create Validation ds
2) POS_CASH_balance.csv → df1
3) Bureau.csv → df2
4) Bureau_balance.csv → df3
5) Credit_card_balance.csv → df4
6) Installments_payments.csv
7) Previous_application.csv
8) HomeCredit_columns_description.csv

[Assuming Data Integration has been done but can check out other features and see if relevant]

Dataset Description

- **Application_{train|test}.csv → Okay so this is used for train and test**
  - This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
  - Static data for all applications. One row represents one loan in our data sample.
- **bureau.csv**
  - All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
  - For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
- **bureau_balance.csv**
  - Monthly balances of previous credits in Credit Bureau.
  - This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.
- **POS_CASH_balance.csv**
  - Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
  - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.
- **credit_card_balance.csv**
  - Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
  - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.
- **previous_application.csv**
  - All previous applications for Home Credit loans of clients who have loans in our sample.
  - There is one row for each previous application related to loans in our data sample.
- **installments_payments.csv**
  - Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
  - There is a) one row for every payment that was made plus b) one row each for missed payment.
  - One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.
- **HomeCredit_columns_description.csv**
  - This file contains descriptions for the columns in the various data files.

Data Dictionary Document the features (input variables) and target variable. o Describe each feature in your own words using the Kaggle dataset description or metadata from other sources. Include units (if applicable) and note categorical vs. numerical variables.

Following quick sorting of variables based on code that looks at df[FEATURE].value_counts().nunique() and if >= 20 we put as Categorical else Numerical and if the df[FEATURE].dtype == Object then we put as Continuous given that it has < 20 unique values

```
for x in df_train.columns:
  unique = len(df_train[x].unique())
  type = df_train[x].dtype
  if(unique >= 20 and type != object):
     print(f"{x} is Continuous Numerical Variable\n")

  elif(unique < 20 and type != object):
     print(f"{x} is Discrete Numerical Variable\n")
  else:
     print(f"{x} is Categorical Variable\n")
```

Variable Type	Variables
Continuous Numerical [Can be Categorical Var]	SK_ID_CURR, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, REGION_POPULATION_RELATIVE, DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, OWN_CAR_AGE, HOUR_APPR_PROCESS_START, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3, APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG, YEARS_BUILD_AVG, COMMONAREA_AVG, ELEVATORS_AVG, ENTRANCES_AVG, FLOORSMAX_AVG, FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG, LIVINGAREA_AVG, NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, APARTMENTS_MODE, BASEMENTAREA_MODE, YEARS_BEGINEXPLUATATION_MODE, YEARS_BUILD_MODE, COMMONAREA_MODE, ELEVATORS_MODE, ENTRANCES_MODE, FLOORSMAX_MODE, FLOORSMIN_MODE, LANDAREA_MODE, LIVINGAPARTMENTS_MODE, LIVINGAREA_MODE, NONLIVINGAPARTMENTS_MODE, NONLIVINGAREA_MODE, APARTMENTS_MEDI, BASEMENTAREA_MEDI, YEARS_BEGINEXPLUATATION_MEDI, YEARS_BUILD_MEDI, COMMONAREA_MEDI, ELEVATORS_MEDI, ENTRANCES_MEDI, FLOORSMAX_MEDI, FLOORSMIN_MEDI, LANDAREA_MEDI, LIVINGAPARTMENTS_MEDI, LIVINGAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, NONLIVINGAREA_MEDI,

	TOTALAREA_MODE, OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DAYS_LAST_PHONE_CHANGE, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_YEAR
Discrete Numerical	TARGET, CNT_CHILDREN, CNT_FAM_MEMBERS, REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY, REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION, REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY, LIVE_CITY_NOT_WORK_CITY, FLAG_MOBIL, FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, FLAG_EMAIL, DEF_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, FLAG_DOCUMENT_2, FLAG_DOCUMENT_3, FLAG_DOCUMENT_4, FLAG_DOCUMENT_5, FLAG_DOCUMENT_6, FLAG_DOCUMENT_7, FLAG_DOCUMENT_8, FLAG_DOCUMENT_9, FLAG_DOCUMENT_10, FLAG_DOCUMENT_11, FLAG_DOCUMENT_12, FLAG_DOCUMENT_13, FLAG_DOCUMENT_14, FLAG_DOCUMENT_15, FLAG_DOCUMENT_16, FLAG_DOCUMENT_17, FLAG_DOCUMENT_18, FLAG_DOCUMENT_19, FLAG_DOCUMENT_20, FLAG_DOCUMENT_21, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_QRT
Categorical Variables	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, OCCUPATION_TYPE, WEEKDAY_APPR_PROCESS_START, ORGANIZATION_TYPE, FONDKAPREMONT_MODE, HOUSETYPE_MODE, WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE

Description of Variables(Based on Context)

Note: If a variable is continuous and appears to be categorical, we will reclassify it here. Conversely, if a variable has no numerical interpretation, we will treat it as categorical. Assume all variables are random. Also, all flags have the same meaning, represented as either 0 or 1.

Target and ID

Feature	Type	Notes
SK_ID_CURR	Categorical	Unique ID for loan application.
TARGET	Categorical	1 = client defaulted (late payment > X days), 0 = no default. This is the target variable.

Personal Information

Feature	Type	Notes / Considerations
CODE_GENDER	Categorical	Encode (e.g., one-hot or label).
FLAG_OWN_CAR	Categorical	Binary 0/1.
FLAG_OWN_REALTY	Categorical	Binary 0/1.
CNT_CHILDREN	Categorical	The number of children could also be treated as numeric.
DAYS_BIRTH	Discrete Numeric	Negative days since birth → convert to age in years for modeling.
DAYS_EMPLOYED	Discrete Numeric	Negative days since the start of employment → can create features like EMPLOYED_YEARS. Watch for anomalies (e.g., very large positive values indicate missing).
FLAG_MOBIL, FLAG_EMP_PHONE, FLAG_HOME_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, FLAG_EMAIL	Categorical	Binary indicators of contact methods.

Financial Information

Feature	Type	Notes / Considerations
AMT_INCOME_TOTAL	Continuous	Client income.
AMT_CREDIT	Continuous	Total loan amount requested.
AMT_ANNUITY	Continuous	Loan installment.
AMT_GOODS_PRICE	Continuous	Price of goods/service purchased.
NAME_CONTRACT_TYPE	Categorical	Cash vs. revolving loan.
NAME_INCOME_TYPE	Categorical	Income source: working, business, maternity, etc.
OCCUPATION_TYPE	Categorical	Job title.

Housing Information

Feature	Type	Notes
NAME_EDUCATION_TYPE	Categorical	Education level.
NAME_FAMILY_STATUS	Categorical	Married, single, etc.
NAME_HOUSING_TYPE	Categorical	Own house, rent, living with parents.
CNT_FAMILY_MEMBER	Categorical	Number of family members.

Regional Information

Feature	Type	Notes
REGION_POPULATION_RELATIVE	Continuous	Relative population size.
REGION_RATING_CLIENT	Categorical (1-3)	Region rating.
REGION_RATING_CLIENT_W_CITY	Categorical (1-3)	City rating.
Address mismatch flags (REG_REGION_NOT_LIVE_REGION, etc.)	Categorical	0/1 or 1/2; binary features indicating address/work region mismatch.

Application Info

Feature	Type	Notes
WEEKDAY_APPR_PROCESS_START	Categorical	Day of week loan applied.
HOUR_APPR_PROCESS_START	Categorical	Hour of day loan applied.

Credit Risk

Feature	Type	Notes
`EXT_SOURCE_1`, `EXT_SOURCE_2`, `EXT_SOURCE_3`	Continuous	Normalized credit scores, external risk indicators.
Bureau-related features (`AMT_REQ_CREDIT_BUREAU_*`)	Categorical	Count of inquiries by different periods (hour/day/week/month/quarter/year). More inquiries usually indicate higher risk.
`STATUS` (from bureau.csv)	Categorical	Loan repayment status (active, closed, DPD0-30, etc.).

Real Estate

Feature	Type	Notes
`APARTMENTS_AVG`, `BASEMENTAREA_AVG`, `YEARS_BEGINEXPLUATATION_AVG`, etc.	Continuous	Average values across properties.
`APARTMENTS_MODE`, `BASEMENTAREA_MODE`, ...	Continuous	Mode values across properties.
`APARTMENTS_MEDI`, `BASEMENTAREA_MEDI`, ...	Continuous	Median values across properties.

Documents

Feature	Type	Notes
FLAG_DOCUMENT_2 → FLAG_DOCUMENT_21	Categorical	Binary flags if client provided documents. Can sum them to create TOTAL_DOCUMENTS_PROVIDED feature.

Social Circle Defaults

Feature	Type	Notes
OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE	Discrete	Number of overdue payments in social circle.
DEF_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE	Categorical	Defaults in social circle (binary/ordinal).

New Categorized Variables/Features:

Continuous Numerical Variables

- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- REGION_POPULATION_RELATIVE
- EXT_SOURCE_1
- EXT_SOURCE_2
- EXT_SOURCE_3
- APARTMENTS_AVG
- BASEMENTAREA_AVG
- YEARS_BEGINEXPLUATATION_AVG
- YEARS_BUILD_AVG
- COMMONAREA_AVG
- ELEVATORS_AVG
- ENTRANCES_AVG
- FLOORSMAX_AVG
- FLOORSMIN_AVG
- LANDAREA_AVG
- LIVINGAPARTMENTS_AVG
- LIVINGAREA_AVG

- NONLIVINGAPARTMENTS_AVG
- NONLIVINGAREA_AVG
- APARTMENTS_MODE
- BASEMENTAREA_MODE
- YEARS_BEGINEXPLUATATION_MODE
- YEARS_BUILD_MODE
- COMMONAREA_MODE
- ELEVATORS_MODE
- ENTRANCES_MODE
- FLOORSMAX_MODE
- FLOORSMIN_MODE
- LANDAREA_MODE
- LIVINGAPARTMENTS_MODE
- LIVINGAREA_MODE
- NONLIVINGAPARTMENTS_MODE
- NONLIVINGAREA_MODE
- APARTMENTS_MEDI
- BASEMENTAREA_MEDI
- YEARS_BEGINEXPLUATATION_MEDI
- YEARS_BUILD_MEDI
- COMMONAREA_MEDI
- ELEVATORS_MEDI
- ENTRANCES_MEDI
- FLOORSMAX_MEDI
- FLOORSMIN_MEDI
- LANDAREA_MEDI
- LIVINGAPARTMENTS_MEDI
- LIVINGAREA_MEDI
- NONLIVINGAPARTMENTS_MEDI
- NONLIVINGAREA_MEDI
- TOTALAREA_MODE

Discrete Numerical Variables

- DAYS_BIRTH
- DAYS_EMPLOYED
- DAYS_REGISTRATION
- DAYS_ID_PUBLISH
- OWN_CAR_AGE
- OBS_30_CNT_SOCIAL_CIRCLE
- OBS_60_CNT_SOCIAL_CIRCLE

Categorical Variables

- SK_ID_CURR
- TARGET
- NAME_CONTRACT_TYPE
- CODE_GENDER
- FLAG_OWN_CAR
- FLAG_OWN_REALTY
- CNT_CHILDREN (usually integer, but treated categorical here)
- NAME_TYPE_SUITE
- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- NAME_FAMILY_STATUS
- NAME_HOUSING_TYPE
- FLAG_MOBIL
- FLAG_EMP_PHONE
- FLAG_HOME_PHONE
- FLAG_CONT_MOBILE
- FLAG_PHONE

- FLAG_EMAIL
- OCCUPATION_TYPE
- CNT_FAMILY_MEMBER
- REGION_RATING_CLIENT
- REGION_RATING_CLIENT_W_CITY
- WEEKDAY_APPR_PROCESS_START
- HOUR_APPR_PROCESS_START
- REG_REGION_NOT_LIVE_REGION
- REG_REGION_NOT_WORK_REGION
- LIVE_REGION_NOT_WORK_REGION
- REG_CITY_NOT_LIVE_CITY
- REG_CITY_NOT_WORK_CITY
- ORGANIZATION_TYPE
- FONDKAPREMONT_MODE
- HOUSETYPE_MODE
- WALLSMATERIAL_MODE
- EMERGENCYSTATE_MODE
- DEF_30_CNT_SOCIAL_CIRCLE
- DEF_60_CNT_SOCIAL_CIRCLE
- DAYS_LAST_PHONE_CHANGE (although numerical, looks like a flag or encoded category)
- FLAG_DOCUMENT_2 through FLAG_DOCUMENT_21
- AMT_REQ_CREDIT_BUREAU_HOUR
- AMT_REQ_CREDIT_BUREAU_DAY
- AMT_REQ_CREDIT_BUREAU_WEEK
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_REQ_CREDIT_BUREAU_YEAR
- STATUS

EDA[Exploratory Data Analysis

Based on the context of the problem, our target variable is 1 if the client has had a late payment greater than X days on any installment; otherwise, it is 0. The goal is to predict default risk, or the likelihood of a late payment. Since the target variable is binary, it is appropriate to use a **binary classification model**.

As our objective is to classify or predict whether a client will default, we should focus on variables that describe the client's ability to repay. These may include **historical credit performance**, **financial data**, and **indicators of stability**, such as housing situation and family status. Additional factors may include **demographic risks** (e.g., age and gender), **social influences** (e.g., the client's social circle), and **application timing or documentation flags** that might be relevant.

For further insights, see the accompanying Jupyter Notebook (.ipynb) file for **visualizations and exploratory data analysis (EDA)** performed on this dataset.

Breakdown of Each Feature:

Feature Type	Description	Importance Insight
ID Features	SK_ID_CURR, SK_ID_PREV, SK_BUREAU_ID	IDs - identifiers, usually dropped or used for merging.
Demographic & Personal Info	CODE_GENDER, CNT_CHILDREN, FLAG_OWN_CAR, FLAG_OWN_REALTY, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, OCCUPATION_TYPE, REGION info, etc.	Important, capture socio-economic status & risk patterns.
Financial Features	AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, AMT_CREDIT_SUM, AMT_CREDIT_SUM_DEBT, AMT_CREDIT_SUM_OVERDUE	Crucial predictors of repayment capacity & debt burden.
Credit Behavior Features	DAYS_CREDIT, CREDIT_DAY_OVERDUE, STATUS (bureau status), CNT_CREDIT_PROLONG, MONTHS_BALANCE, AMT_REQ_CREDIT_BUREAU_	Highly informative of past and current credit behavior and risk.

Time Features	DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE	Age and stability indicators that often correlate with risk.
Property & Housing	APARTMENTS_, *BASEMENTAREA_*, FLOORSMAX_, *YEARS_BUILD_*, FONDKAPREMONT_MODE, HOUSETYPE_MODE, WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE	Indirect proxies for wealth and living conditions.
Social Circle & Phone Flags	OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, FLAG_MOBIL, FLAG_EMP_PHONE, FLAG_PHONE, FLAG_EMAIL	Social risk factors & communication channels presence.
Documents Flags	FLAG_DOCUMENT_2 -FLAG_DOCUMENT_21	Completeness of documentation; might hint at reliability.
Loan Application Process	WEEKDAY_APPR_PROCESS_START, HOUR_APPR_PROCESS_START, NAME_TYPE_SUITE, NAME_INCOME_TYPE	Contextual application timing info that might reveal fraud or urgency.

Most Important Features to Visualize based on context:

AMT_INCOME_TOTAL
- AMT_CREDT
- AMT_ANNUITY
- DAYS_BIRTH(AGE OF CLIENT)
- DAYS_EMPLOYED(HOW LONG DO THEY HAVE JOB)
- STATUS(CB loan status)
- NAME_INCOME_TYPE(Source of Income)
- EXT_SOURCE_1 - EXT_SOURCE_3
- AMT_REQ_CREDIT_BUREA_MON(Total amount of inquiries by CB in past month)
- OBS_30_CNT_SOCIAL_CIRCLE
- DEF_30_CNT_SOCIAL_CIRCLE
- OCCUPATION_TYPE
- FLAG_OWN_REALTY
- FLAG_OWN_CAR