

# Math 789 Project

## Group 1

Cassie Kang, Junghwa Jang, Larry Jones, Suhyeon Lee, Lisa Wang



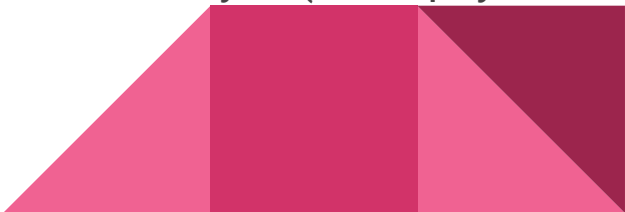
# Technical Project:

## Home Credit Default Risk Prediction

# Project Overview

Securing a home loan is a pivotal aspect of many individuals' lives, especially for those with limited or nonexistent credit histories. Moreover, home credit acts as a vital revenue stream for commercial banks. Accurately forecasting the risk of home credit defaults is therefore essential. Doing so not only ensures a positive loan experience for the underserved population but also safeguards banks, allowing them to extend home credit offerings to an even broader clientele.

Through our project, we aim to develop the most effective prediction model by examining various machine-learning-based models. Additionally, we will identify key covariates that significantly influence the model's performance. The training data contains labels, and our objective is to teach a model to predict these labels based on the given features (supervised learning). This is a classification task where the label is binary: 0 (will repay the loan on time) and 1 (will face difficulty repaying the loan).



# Research Questions

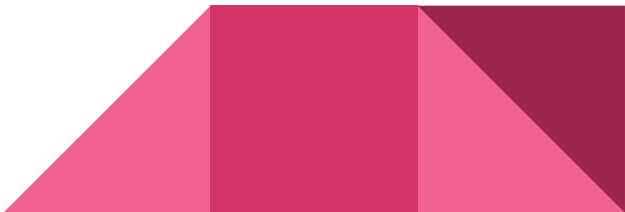
- 1: Which machine-learning model best predicts home credit defaults?
- 2: Which covariates are most influential in prediction accuracy?
- 3: What are the broader implications of this predictive insights for the financial strategies and risk management practices of commercial banks?



# Dataset

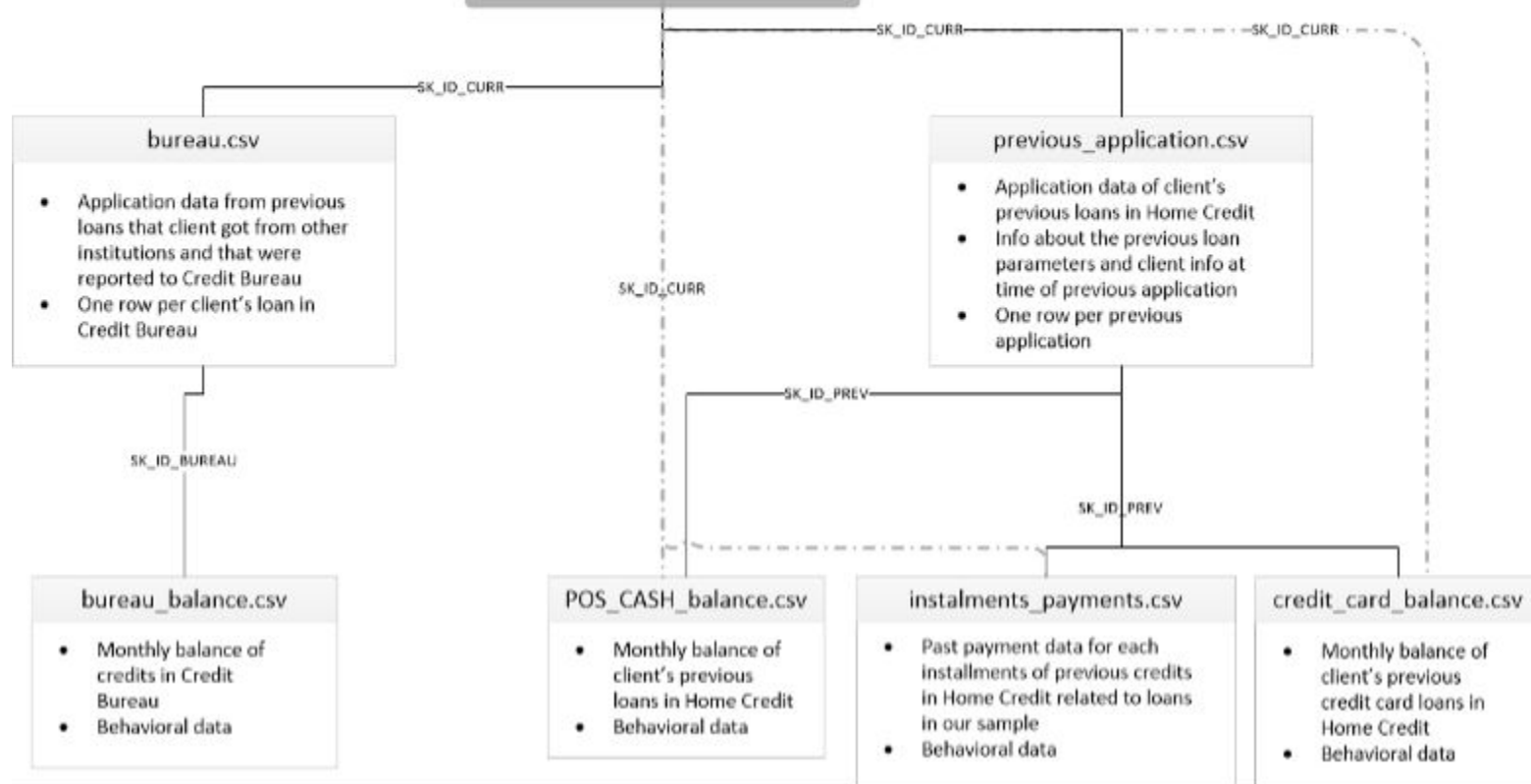
- From Kaggle competition (Home Credit Default Risk)

Anna Montoya, inversion, Kirill Odintsov, Martin Kotek. (2018). Home Credit Default Risk. Kaggle. <https://kaggle.com/competitions/home-credit-default-risk>

- The dataset we'll be working with is named `application\_{train|test}.csv`. This is the primary training and testing dataset containing information about each loan application at Home Credit. Each loan is represented by a row and can be identified using the `SK\_ID\_CURR`. The training data includes the `TARGET` variable where 0 indicates the loan was repaid, and 1 indicates it was not. The data is split into two files: Train (which includes the `TARGET`) and Test (which does not include the `TARGET`). The size of the application train data is (307511, 122).
- 

### application\_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time



# Dataset Structure

```
In [3]: application_train.head()
```

```
Out[3]:  
shape: (5, 122)
```

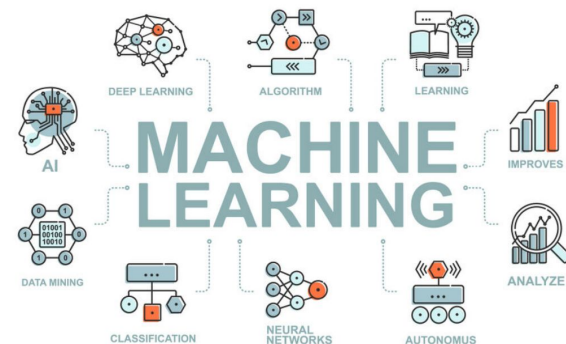
SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUIT
i64	i64	str	str	str	str	i64	f64	f64	f64
100002	1	"Cash loans"	"M"	"N"	"Y"	0	202500.0	406597.5	24700.5
100003	0	"Cash loans"	"F"	"N"	"N"	0	270000.0	1293502.5	35698.5
100004	0	"Revolving loan..."	"M"	"Y"	"Y"	0	67500.0	135000.0	6750.0
100006	0	"Cash loans"	"F"	"N"	"Y"	0	135000.0	312682.5	29686.5
100007	0	"Cash loans"	"M"	"N"	"Y"	0	121500.0	513000.0	21865.5

# Methodology

- Data preprocessing and EDA
- Models to consider:
  - Support Vector Machine(SVM)
  - XGBoost
  - Random Forest (using K-Fold for cross-validation to avoid overfitting)
  - Logistic Regression
  - LDA/QDA
- Feature engineering: perform feature selection using techniques such as
  - PCA
  - remove collinear features
  - remove features with greater than a threshold percentage of missing values
  - keep only the most relevant features using feature importance from a model


<https://www.kaggle.com/code/willkoehrsen/introduction-to-feature-selection>

- Evaluations:
  - metric (ROC AUC): for each SK\_ID\_CURR in the test set, we will predict a probability for the TARGET variable.
  - do model comparison, identify which predictor is more important

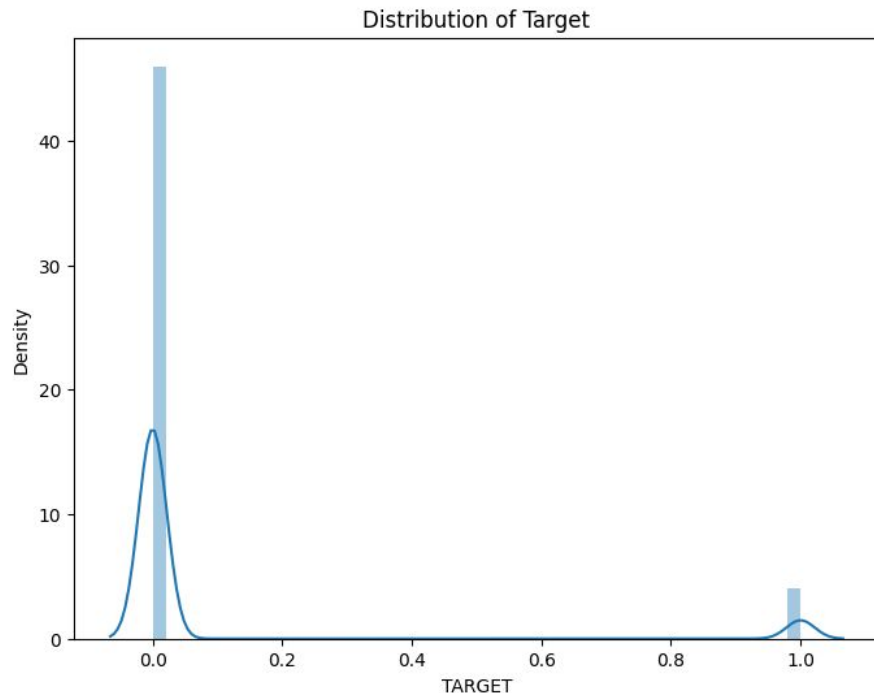




# Progress Update

1. We initiated the project by creating a collaborative notebook on Kaggle, allowing each team member to contribute online. The notebook is publicly accessible at:  
<https://www.kaggle.com/code/lisa0910/fin789-group1>
  2. Successfully loaded the dataset without encountering any loading issues.
  3. Conducted a comprehensive Exploratory Data Analysis (EDA) on the dataset to gain insights and understand its characteristics.
  4. Performed principal component analysis (PCA) on the variables to explore dimensionality reduction and feature importance.
  5. Based on the findings from our EDA, it became evident that additional data cleaning and preprocessing are necessary to enhance the dataset's quality and suitability for further analysis.
- 

# Exploratory Data Analysis: Dependent Variable

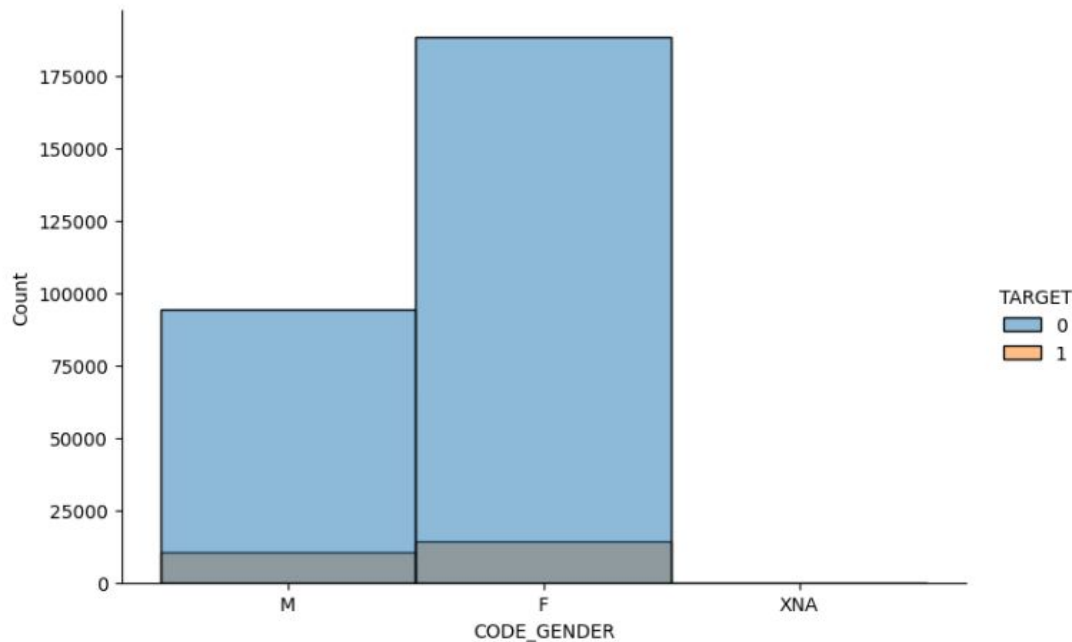


The dependent variable is “Target,” which represents whether the client has repaid on time. If yes, it is coded as 0, otherwise, if one has failed to repay for more than a certain number of times, coded as 1.

Our data set is unbalanced data, which has a majority of diligent clients, and a small number of non-diligent clients.

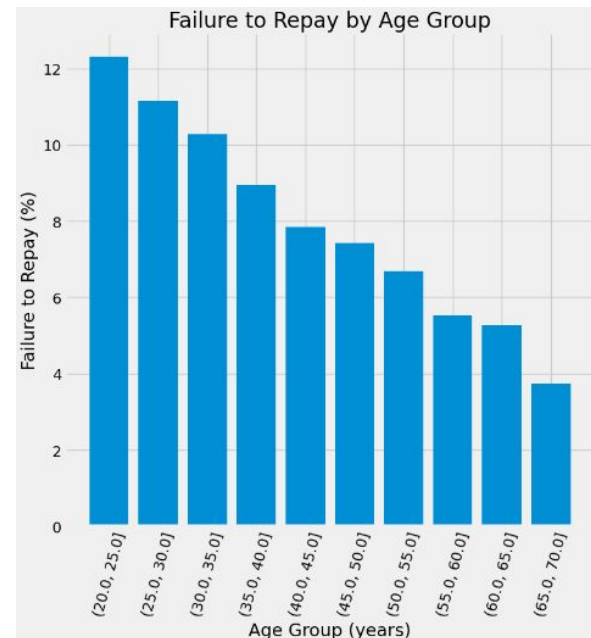
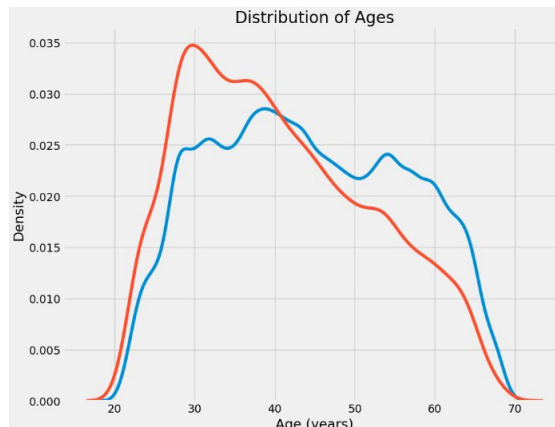
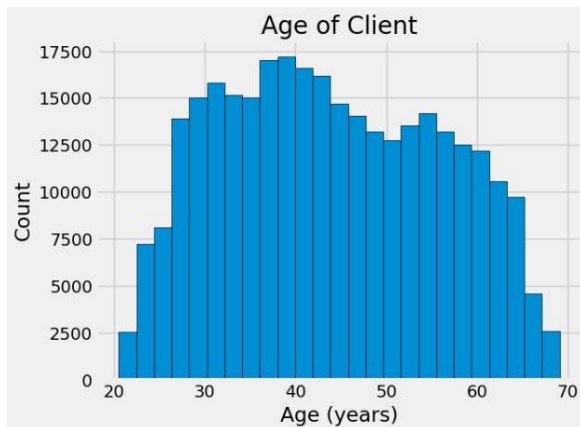
This distribution should be considered in further analysis.

# EDA: Gender distribution



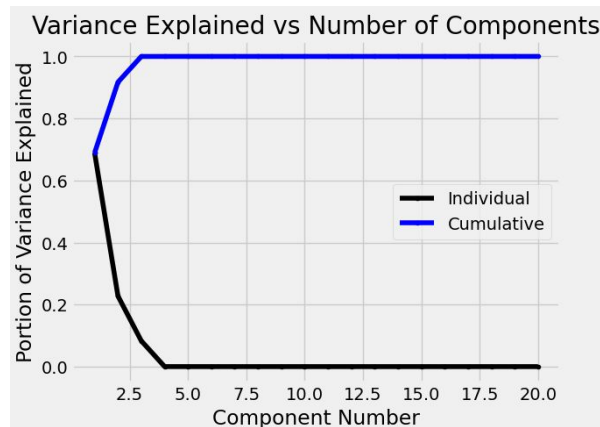
Overall, there are more female clients than male clients, and the proportion of non-diligent clients is higher in male. We will include this variable as a distinctive independent variable to check its importance as a covariate as well as for the later interpretation.

# EDA: Age

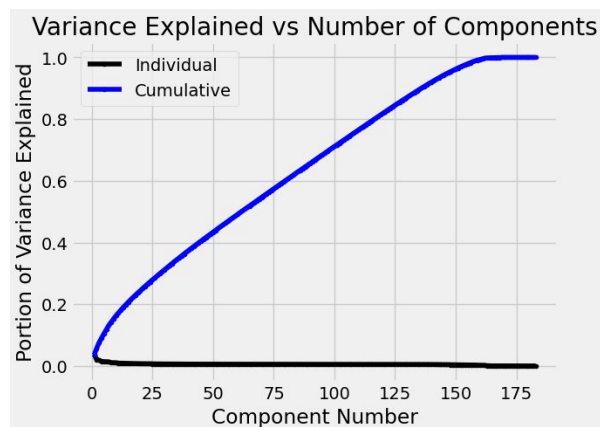


Most of the clients are in their 30s to 50s. Our initial analysis shows that the proportion non-diligent clients is high in the age 20s and 30s group, and interestingly, the rate of failure to repay is inversely related with the age. Clients in their 20s show the highest rate of failure to repay, and this decreases as the age of clients increases.

# Principal Component Analysis



Unnormalized data



Normalized data

We conducted an initial round of dimension reduction using principal component analysis. The result was drastically different when using unnormalized data vs normalized data. With unnormalized data, with only 4 principal components, over 99% of the variance could be explained, whereas with normalized data, we need more than 100 components to explain 80% of the variance. It is clear that we need to normalize the data since features are on vastly different scales. We will experiment with different predictive models and determine how many components are appropriate to select in order to achieve optimal model performance.

# Next Step for Technical Project

1. **Missing Value Analysis & Treatment:** We will meticulously assess the dataset for missing values and develop a strategy to handle them. This may include imputation techniques, data interpolation, or considering the removal of variables or data points with excessive missing values.
2. **Further EDA (Exploratory Data Analysis):** We will continue to explore the dataset in greater detail. This includes in-depth visualizations, statistical analyses, and feature engineering to uncover patterns, relationships, and potential outliers within the data.
3. **Prediction Modeling:** Our next objective is to build predictive models. We will select and evaluate various machine learning algorithms to identify the most suitable model for the dataset. This involves tasks such as data splitting, model training, hyperparameter tuning, and performance evaluation using appropriate metrics.
4. **Regression Modeling for Interpretation:** In addition to prediction models, we may opt for regression modeling. Regression analysis can help us understand the relationships between variables, identify significant predictors, and interpret the impact of each predictor on the target variable. This interpretation can provide valuable insights into the factors that influence the outcome of interest.




# Non-Technical Project:

## Evaluating Credit Risk in J.P. Morgan Chase

# Project Overview


We're reviewing the credit risk at JP Morgan Chase, a widely utilized bank among students. Given its substantial size and dual focus on commercial and investment activities, examining it is particularly intriguing. This project aims to provide practical experience in risk management and investment banking, allowing participants to actively engage with real-world scenarios and challenges.

By exploring JP Morgan Chase's credit risk from various angles, such as balance sheet and cash flow sheets, we offer students a valuable opportunity to acquire and apply relevant skills. Additionally, this project serves as a platform to identify and explore career opportunities in risk management and investment banking, shedding light on the diverse roles available and the skills and qualifications required in the field.



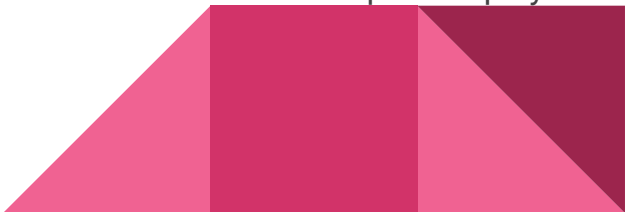


# Research Questions

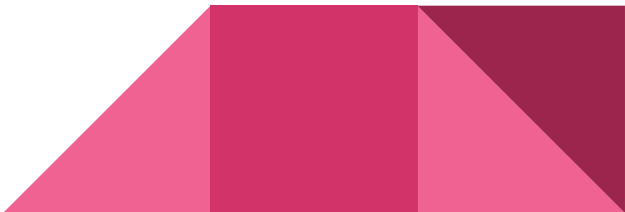
1. What is the credit risk profile of JP Morgan Chase, and how does it impact the financial stability of the institution?
  2. How does JP Morgan Chase's credit risk exposure vary across different financial statements, such as the balance sheet and cash inflow sheets?
  3. What risk management strategies and practices does JP Morgan Chase employ to mitigate credit risk, and how effective are these strategies in ensuring financial stability?
  4. What job opportunities exist within the field of risk management at JP Morgan Chase, and what are the key qualifications and skills required for these roles?
  5. What specific job opportunities are available for investment banking analysts, and what are the key qualifications and skills required for these roles?
- 

# Approaches

Our methodology for addressing the credit risk associated with J.P. Morgan Chase adopts a structured and multifaceted strategy designed to provide a holistic perspective and derive insights from various information sources:

1. **Balance Sheet Examination:** We will conduct a rigorous examination of J.P. Morgan Chase's balance sheet. This analysis will encompass data spanning the first three quarters of 2023, as well as information from the past three years. Our primary focus will be on discerning the intricate structure of the bank's assets and liabilities.
  2. **Academic Literature Review:** We will embark on a scholarly exploration by studying research papers in the field of bank risk analysis. This academic approach will equip us with a deeper understanding of the methodologies employed to assess credit risks within banking institutions.
  3. **Media and News Analysis:** A meticulous examination of press releases from J.P. Morgan Chase and other pertinent news publications will be undertaken. We will scrutinize these sources to identify noteworthy events and subsequently establish correlations with observed trends in the data.
  4. **Quantitative Roles Investigation:** We will conduct an in-depth study of quantitative roles within J.P. Morgan Chase's risk department. This will serve as a means to gain valuable insights into the quantitative tools and techniques employed in their risk management practices.
- 

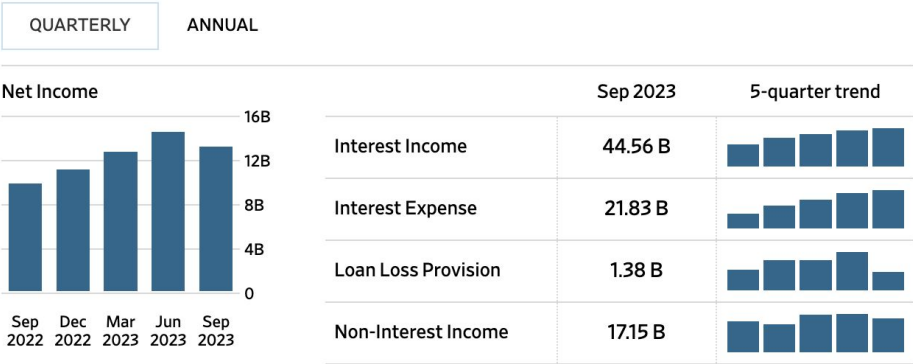
# Progress Update

1. We have successfully gathered J.P. Morgan Chase's balance sheets for the first three quarters of 2023, as well as for the years 2022, 2021, and 2020, directly from the official J.P. Morgan Chase website.
  2. In our research endeavors, we have identified a scholarly paper that specializes in bank risk analysis. We intend to leverage the methodology outlined in this study as a foundational framework for our research. The paper in question can be accessed via this link: <https://www.nber.org/papers/w21334>
  3. We've discovered an invaluable resource within J.P. Morgan's official website—specifically, the investment section that provides relevant information for investors. This information will be instrumental in enhancing our understanding of the bank's operations and financial outlook. You can access this resource here: <https://www.jpmorganchase.com/ir/news>
  4. To facilitate our comprehensive job analysis, we've identified J.P. Morgan Chase's dedicated career page, which specifically caters to risk management-related positions. Our investigation into the bank's risk management practices will commence from this portal: <https://careers.jpmorgan.com/us/en/our-businesses/risk>
- 

# Balance Sheet

- Both the interest income and interest expenses growth over the past several quarters. And the loan loss provision dropped suddenly last quarter.
- Notably, there was a significant decrease in the growth of Cash and Due from Banks in the same period. We intend to conduct a thorough investigation to uncover the underlying causes of this substantial decline and gain a deeper understanding of the events that transpired

## Income Statement JPMorgan Chase & Co. →



### Assets

Fiscal year is January-December. All values USD Millions.

	30-Sep-2023	30-Jun-2023	31-Mar-2023	31-Dec-2022	30-Sep-2022	5-qtr trend
Total Cash & Due from Banks	24,921	50,364	45,298	53,097	50,454	
Cash & Due from Banks Growth	-50.52%	11.18%	-14.69%	5.24%	-	

# Cash Flow

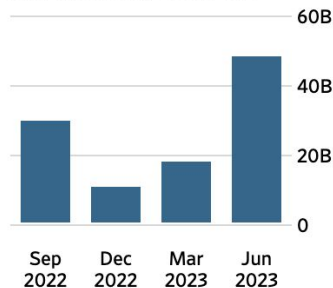
We have observed fluctuations in the net operating cash flow over both recent quarters and previous years. Our objective is to comprehend the underlying reasons for these variations and evaluate their potential impact on the overall financial stability of the bank

## Cash Flow JPMorgan Chase & Co. →

QUARTERLY

ANNUAL

### Net Operating Cash Flow



Jun 2023

5-quarter trend

Net Financing Cash Flow

-49.91 B

Free Cash Flow

+47.66 B

Cash Flow Per Share

-

Free Cash Flow Per Share

-

# Investment Banking Analyst

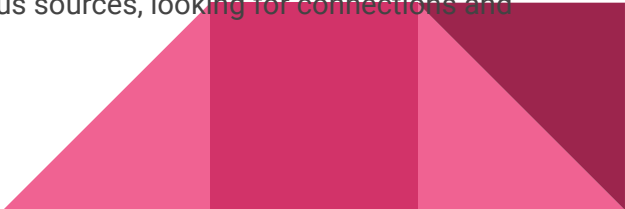
- Key responsibilities:
  - Transaction execution: support due diligence, document preparation, negotiation etc and liaising with relevant counterpart at the client/other advisers (accounting, legal, etc)
  - Working with J.P. Morgan product and sector teams
  - Building and using complex financial models, completing valuation and analytical exercises
  - Drafting presentation materials, management presentations, Board materials, Memorandums, and other presentation materials for use in M&A and capital markets transactions or strategic client dialogue Keeping abreast of key market developments and establishing knowledge of the target client base and relevant market dynamics
- Qualifications:
  - Bachelor's degree in Finance, Economics, Business Administration, or a related field
  - Solid knowledge of corporate finance and accounting, or understanding of corporate charters, bylaws, and governance practices
  - Ability to comfortably interact with clients and colleagues in a professional and mature manner
  - Outstanding ethics, integrity, and judgment
  - Intellectual curiosity, strong work ethic, and desire to learn

Sample of interview people with experience in finance jobs:

# Quantitative Developer (QD)

- Progress: interview completed
  - Spoke to a Quantitative Developer at Two Sigma
    - Questions asked include:
      - Can you provide an overview of your day-to-day responsibilities as a QD?
      - What educational background and qualifications are essential for success in this role?
      - What technical skills are crucial for quantitative development?
      - How important is it to have a deep understanding of financial markets and trading strategies in this role?
      - What are the most significant challenges you face in your daily work as a Quantitative Developer?
      - What strategies do you employ to stay competitive in the dynamic and highly competitive field?
      - What is the typical career progression for someone starting as a Quantitative Developer in this industry?
    - Also researched current Quantitative Development job posting to see if (minimum) requirements are consistent with what was specified by the interviewee, and if not, how they have changed.

# Future work

1. Data Validation and Preprocessing: Review the collected balance sheet data to ensure its accuracy and consistency. Prepare the data for analysis by cleaning, formatting, and structuring it appropriately.
  2. Methodology Implementation: Begin implementing the methodology outlined in the identified paper on bank risk analysis as the basis for our research. Customize it to suit our specific objectives and dataset.
  3. Continuous Data Monitoring: Regularly monitor J.P. Morgan Chase's website for any updates to their balance sheets and investor-related information for the remainder of 2023. This ensures we have the most current data for our analysis.
  4. Investor Information Analysis: Dive into the information available in the investment section of J.P. Morgan's website to extract valuable insights and trends relevant to our research objectives.
  5. Job Analysis and Interviews: Initiate the job analysis process using both the risk management-related and investment banking analyst job postings on J.P. Morgan Chase's career page. Delve into the roles, responsibilities, and qualifications for a comprehensive grasp of their risk management and investment banking practices. As part of the interview preparation, focus on behavioral questions, technical inquiries, case study evaluations, and consider conducting mock interviews to ensure readiness.
  6. Synthesize Findings: Begin synthesizing the information and data collected from various sources, looking for connections and insights that will contribute to our overall analysis.
- 





# Timeline & Task Division

# Timeline and Task Division

	Technical	Non-Technical
10/23 - 11/1	EDA, data visualization (Junghwa), Principal component analysis (Suhyeon)	Gather data for JP Morgan's balance sheet, job descriptions (Lisa, Cassie, Larry)
11/2-11/8	Further EDA (Cassie), Feature Engineering (Suhyeon)	Balance Sheet Ratio Analysis (Lisa), more job descriptions (Suhyeon)
11/9-11/15	Logistic / LDA/QDA / SVM (Junghwa)	Balance Sheet Trend Analysis, Interview people who have work experience in the financial industry (Cassie, Junghwa, Larry)
11/16-11/21	XGBoost / Random Forest (Suhyeon)	Balance Sheet Comparison (Lisa, Larry)
11/17-12/5	Final Consolidation of Reports (All group members)	Final Consolidation of Reports (All group members)
12/6	Final Presentation (All group members)	

# Division of Tasks (Completed)

## **Technical**

Junghwa: Exploratory data analysis

Suhyeon: Principal component analysis (dimension reduction)

## **Non-technical**

Lisa: Balance sheet analysis

Cassie & Larry: JP Morgan entry-level job description analysis & Initial interview

