Big Data Analytics – project presentation

24 May 2023  
Group: Big Data Big Dreams

* Ariq Bintang (22-605-901)
* Luca Gewehr (22-620-967)
* Marco Hafid (22-620-546)

*Note to self... The stages between data collection and data analysis typically include the following steps:*

1. *Data Validation: This step ensures that the data collected is accurate, complete, and meets the specified standards.*
2. *Data Integration: If data is gathered from multiple sources, this step involves combining the data into a consistent format.*
3. *Data Cleaning or Data Scrubbing: This involves checking the data for errors and correcting or removing any inconsistencies, outliers, or inaccuracies.*
4. *Data Transformation: In this step, data is converted into a suitable format for analysis. This might involve tasks such as normalization or aggregation.*
5. *Data Storage: This involves deciding where and how to store the data, such as in a database or a data warehouse. It might also involve considerations about data security and access.*
6. *Data Preparation: This final step before analysis involves selecting the specific data that will be used for analysis and possibly creating derived variables or metrics. This step makes the data ready for analysis.*
7. *Exploratory Data Analysis (EDA): An initial exploration of the data, often using visualization techniques, to understand the properties of the data, check for patterns, or formulate hypotheses.*

*The process of generating new columns or features from existing ones is indeed a part of feature engineering. Feature engineering involves creating new features or modifying existing ones to improve the performance of machine learning algorithms.*

*In the stages you listed, feature engineering could fall under both Data Transformation and Data Preparation:*

1. *Data Transformation: This stage may involve some feature engineering when the existing data is converted or transformed into a new format. This can include processes like normalization, scaling, or creating interaction features (combinations of existing features).*
2. *Data Preparation: This stage often involves a good deal of feature engineering. Here, the selected data may be further refined and prepared for analysis. This could include creating derived variables, which is essentially feature engineering.*

*So, in your case, generating thousands of additional columns that are combinations of the 26 initial columns would be a part of feature engineering, and this process would occur primarily during the Data Transformation and Data Preparation stages.*

**Introduction**

This project aims to determine the optimal quantitative investment strategies considering various investment parameters. We explore a range of sub-questions, from defining the relevant investment parameters to validation of the statistical reliability of the optimal strategy.

The research employs a collection of data from several sources, including Bloomberg Terminal, World Bank, and Swiss National Bank. The data includes price data of selected indices and currency pairs, Swiss inflation data, CHF money market rates, and spot interest rates on Swiss Confederation bond issues. The analysis involves data cleaning, manipulation, and use of machine learning algorithms to derive optimal investment strategies.

Quantitative investment strategies have been the subject of extensive research in financial economics. These strategies employ mathematical frameworks to identify profitable investment opportunities, based on certain parameters. This study seeks to identify the optimal quantitative investment strategies for various possible combinations of investment parameters. The data used in this study comes from diverse sources, including the Bloomberg Terminal, the World Bank, and the Swiss National Bank (SNB). The research comprises a multi-faceted methodology, including data collection and preparation, data analysis, and results validation.

The study revolves around defining and identifying the most effective quantitative investment strategies given a variety of investment parameters. We examine a spectrum of elements, from discerning the pertinent investment parameters to verifying the statistical credibility of the identified strategy. For our analysis, we utilize a robust dataset drawn from various sources such as Bloomberg Terminal, World Bank, and Swiss National Bank. The data encompasses a multitude of features such as price data for selected indices and currency pairs, Swiss inflation data, CHF money market rates, and spot interest rates on Swiss Confederation bond issues. Our methodology includes an extensive process of data cleaning, manipulation, and the application of machine learning algorithms to uncover the optimal investment strategies.

**Research Question**

The main research question is: "What are the optimal, quantitative investment strategies for different possible combinations of investment parameters?"

The main research question is: What are the optimal, quantitative investment strategies for different possible combinations of investment parameters?

Our main research question is, "What are the optimal, quantitative investment strategies for different possible combinations of investment parameters?" This broad question is supplemented by various sub-questions which intend to provide a comprehensive investigation of various factors influencing the optimal investment strategy, considering parameters such as desired investment objectives, risk constraints, ESG criteria, asset class and geographic restrictions, and more.

Our pivotal research question is, "What are the optimal, quantitative investment strategies for various potential combinations of investment parameters?" To fully address this central query, we incorporate multiple sub-questions that delve into numerous aspects influencing the choice of optimal investment strategy. These aspects include, but are not limited to, investment objectives, risk tolerance, ESG criteria, asset class considerations, geographic restrictions, etc.

**Data Source(s)**

Data is collected from four sources:

1. Bloomberg Terminal: Provides the price data of selected indices and currency pairs.
2. World Bank: Offers Swiss inflation data (CPI in %).
3. Swiss National Bank: Offers CHF money market rates.
4. Swiss National Bank: Provides spot interest rates on Swiss Confederation bond issues.

Data for this research was sourced from multiple platforms:

1. "Bloomberg Terminal spreadsheet builder.xlsx" - Price data of selected indices and currency pairs.
2. "API\_FP.CPI.TOTL.ZG\_DS2\_en\_excel\_v2\_5454868.xls" (source: <https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG>) - Swiss inflation data (CPI in %).
3. "snb-chart-data-rendeidglfzch-en-all-20230502\_1430.xlsx" and "snb-chart-data-zimomach-en-all-20230502\_1430.xlsx" (source: <https://data.snb.ch/en/topics/ziredev/chart/zimomach>) - CHF money market rates and spot interest rates on Swiss confederation bond issues.

Our research leverages data from multiple sources, including:

1. "Bloomberg Terminal spreadsheet builder.xlsx" from Bloomberg Terminal, providing price data of selected indices and currency pairs.
2. "API\_FP.CPI.TOTL.ZG\_DS2\_en\_excel\_v2\_5454868.xls" from [World Bank Data](https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG), offering Swiss inflation data (CPI in %).
3. "snb-chart-data-rendeidglfzch-en-all-20230502\_1430.xlsx" and "snb-chart-data-zimomach-en-all-20230502\_1430.xlsx" from [Swiss National Bank Data](https://data.snb.ch/en/topics/ziredev/chart/rendeidglfzch), supplying CHF money market rates and spot interest rates on Swiss Confederation bond issues respectively.

Our investigation relies on a comprehensive set of data derived from multiple sources, as follows:

1. "Bloomberg Terminal spreadsheet builder.xlsx" provided by Bloomberg Terminal, offering the price data of selected indices and currency pairs.
2. "API\_FP.CPI.TOTL.ZG\_DS2\_en\_excel\_v2\_5454868.xls" retrieved from World Bank Data, supplying Swiss inflation data (CPI in %).
3. "snb-chart-data-rendeidglfzch-en-all-20230502\_1430.xlsx" and "snb-chart-data-zimomach-en-all-20230502\_1430.xlsx" sourced from Swiss National Bank Data, contributing CHF money market rates and spot interest rates on Swiss Confederation bond issues respectively.

**Summary of Methods and Results**

The data, collected from multiple sources, undergoes cleaning and manipulation in R to ensure consistency across different dataframes. The transformed data is then used to calculate returns in different scenarios. Machine learning algorithms, specifically, random forest and LASSO, are applied to determine the optimal investment strategy. The results and their interpretation will be discussed in the following sections.

Methods employed in this study involve data collection, cleaning, analysis, and validation. The data were collected from the mentioned sources and transformed into the required format using R. Challenges in data handling and computational efficiency were addressed through effective data management and usage of efficient R functions and libraries. The details of the analysis, results, and interpretation are yet to be determined, as the research is still ongoing.

Our methods involve a complex process of data collection, cleaning, preparation, and analysis, executed in R. We load the data from different documents into R, aligning dates across different datasets, transforming data formats, and performing various operations to derive meaningful metrics and insights. The results derived from this study are then used to identify the optimal investment strategies. The details of the results will be discussed in the 'Results' section.

Our approach incorporates a complex procedure encompassing data collection, cleaning, preparation, and analysis, conducted in R. Data from different documents is imported into R, where we ensure alignment of dates across datasets, convert data formats, and perform an array of operations to extract meaningful metrics and insights. These insights are then employed to pinpoint optimal investment strategies. Further specifics regarding these results will be addressed in the forthcoming 'Results' section.

**Data Collection and Data Storage**

Data collection was a significant challenge given the diversity of sources and the different formats in which the data was presented. We used various libraries and functions in R to load data from Excel and CSV files, convert data types, and extract relevant portions of the data. For example, the relevant data didn't always start on the first tab, first row, and first column, requiring careful extraction.

The raw data is stored in their original formats (Excel and CSV) to preserve data integrity. We use a consistent directory structure and naming conventions to keep track of different versions of the data.

For computational efficiency and ease of management, we stored the manipulated dataframes in R's native format (.RData) and used separate dataframes for each source.

r

# Code snippet for data collection

source("code/data\_collection.R")

Data collection involved a systematic approach of gathering relevant and reliable financial data from the above-mentioned sources. The data encompassed a broad spectrum, including price data, inflation data, and interest rates. The main challenge faced in this stage was obtaining consistent and complete historical data. This was solved by carefully choosing the indices for data extraction and removing the ones with insufficient historical data.

The raw data was stored in data frames in R. This structure was selected due to its compatibility with R's extensive data manipulation and analysis capabilities. The data frames were named descriptively for easy reference and traceability in subsequent stages of the research.

Refer to source("code/data\_collection.R") for the data collection code.

Our data collection procedure involved gathering data from various sources such as Bloomberg Terminal, World Bank Data, and Swiss National Bank Data. One major challenge was to ensure that the data was consistent across different securities, given that the data varied in format and structure across different sources. We addressed this by choosing indices with the longest and most consistent history across our data sources, and then standardizing the data formats across all datasets.

The raw data is stored in Excel files which are loaded into R for analysis. This method of storage was chosen for its simplicity and easy accessibility, as well as for the ability to manually review and verify the data before and after the analysis.

r

# Example of data loading from Excel to R

library(readxl)

bloomberg\_data <- read\_excel("Bloomberg Terminal spreadsheet builder.xlsx")

worldbank\_data <- read\_excel("API\_FP.CPI.TOTL.ZG\_DS2\_en\_excel\_v2\_5454868.xls")

snb\_data1 <- read\_excel("snb-chart-data-rendeidglfzch-en-all-20230502\_1430.xlsx")

snb\_data2 <- read\_excel("snb-chart-data-zimomach-en-all-20230502\_1430.xlsx")

Collecting data was a significant task as it required dealing with a wide variety of sources, each with different data formats. We used specific libraries and functions in R to load data from Excel and CSV files, convert data types, and select the necessary parts of the data. In some cases, the data we needed did not start from the first tab, row, or column, which required careful extraction.

We kept the raw data in its original formats (Excel and CSV) to ensure the data remained accurate. We used clear and consistent file and folder names to keep track of different versions of the data.

For efficiency and easier management, we stored the processed dataframes in R's native format (.RData), using separate dataframes for each source. You can see our data collection process by running source("code/data\_collection.R") in R.

**Data Cleaning and Preparation**

Data cleaning was carried out with a focus on consistency, accuracy, and efficiency. Challenges included aligning dates across different sources, handling missing values, and standardizing inflation values. Bottlenecks included slow computation due to the high number of columns and the large size of data.

To improve data cleaning, we leveraged dplyr's efficient data manipulation functions and purrr's map functions to apply changes across multiple dataframes. We reduced the number of columns, which increased computational performance by calculating possible combinations between columns.

r

# Code snippet for data cleaning and preparation

source("code/data\_preparation.R")

The raw data collected underwent rigorous cleaning and preparation processes to ensure its usability in the subsequent analysis. Data was sorted, filtered, and aligned by dates for consistency. The column names were modified for comprehensibility and efficiency. Data cleaning also involved transforming inflation values into percentages and removing certain indices to improve computational performance.

R and its libraries like 'dplyr' and 'tidyverse' were extensively used for data cleaning. The challenge here was dealing with missing values and potential outliers. In such cases, appropriate statistical techniques such as imputation and robust statistics were employed.

Refer to source("code/data\_preparation.R") for the data preparation code.

The process of cleaning and preparing the data involved numerous steps, including sorting, filtering, and aligning the dates between the data from different files; changing the column names to more understandable names; and transforming inflation values into percentages. One major bottleneck was computational efficiency, as dealing with large numbers of columns could greatly decrease performance. We addressed this by removing columns for some selected indices that only had recent data, thereby reducing the number of columns and thus increasing computational efficiency.

r

# Example of data preparation in R

# Assuming 'data' is your dataframe

library(dplyr)

data <- data %>%

rename(price = old\_column\_name) %>% # Rename column

mutate(date = as.Date(date, format = "%m/%d/%Y")) %>% # Convert date column to Date format

filter(date >= as.Date("1990-01-01")) %>% # Filter data starting from 1990

arrange(date) # Sort data by date

# Transform the inflation values to percentages

worldbank\_data$inflation <- worldbank\_data$inflation / 100

# Cleaning the data was an important step that focused on making the data accurate and consistent across all sources. We faced challenges such as making sure dates matched across different sources, dealing with missing values, and standardizing inflation values. The large size of the data and the high number of columns sometimes slowed down computations.

# To improve the data cleaning process, we used the dplyr library in R for efficient data manipulation and purrr's map functions to apply changes across multiple dataframes. We also increased computational speed by reducing the number of columns through calculating possible combinations. The data cleaning and preparation steps can be seen by running source("code/data\_preparation.R") in R.

# Data Analysis and Data Visualization

Our analysis process included calculation of return series in CHF nominal, real, and excess terms. We faced challenges in implementing this due to the large volume of data and the need to perform computations over many combinations of securities.

We used data.table for its efficient handling of large data sets and ggplot2 for visualization. The reason behind using these tools is their capability to handle large volumes of data effectively.

r

# Code snippet for data analysis and visualization

source("code/data\_analysis.R")

Data analysis was performed using various statistical and machine learning techniques. The methodology for data analysis is yet to be detailed, and so are the specific techniques and their justifications. However, they would likely include methods for estimating and comparing portfolio return series, and measures such as maximum drawdown and conditional VaR. The large volume of data would likely necessitate efficient computational techniques and possibly parallel processing.

For visualization, R libraries like 'ggplot2' would be used to present the results graphically.

Refer to source("code/data\_analysis.R") for the data analysis code.

The analysis included finding correlations between various factors (such as inflation and different indices), implementing quantitative investment strategies, backtesting these strategies, and evaluating their performance.

One significant challenge was to balance risk and return in the investment strategies. This was addressed by applying Modern Portfolio Theory (MPT), which aims to optimize the trade-off between risk and return by combining different assets in a portfolio.

r

# Example of analysis in R

# Assuming 'data' contains your time series data and 'returns' contains the calculated returns

library(PerformanceAnalytics)

correlation\_matrix <- cor(returns) # Calculate correlation matrix

chart.Correlation(returns, histogram=TRUE, pch=19) # Plot the correlation matrix

In our analysis, we calculated the return series in CHF in nominal, real, and excess terms. The high volume of data and the need to perform calculations for many combinations of securities posed challenges.

We used the data.table library in R for efficient handling of large datasets. For visualization, we used the ggplot2 library, which is effective for dealing with large volumes of data. For more details on our data analysis and visualization, please refer to our script by running source("code/data\_analysis.R") in R.

**Results**

Our results provide an optimal investment strategy under different combinations of investment parameters. The detailed results will be discussed in a separate document along with tables and figures that illustrate our findings.

r

source("code/result\_generation.R")

The results section is pending completion, given the research is still ongoing. Upon completion, the main findings would be summarized and supported by exhibits (tables or figures), each accompanied by explanatory notes. These exhibits would be generated using R's data visualization libraries.

The results obtained from the analysis, including the optimal asset allocation for different investment parameters and the performance of the proposed investment strategies, will be discussed in the following sections.

Please note that this is a simplified example of the processes involved, and the actual analysis would likely involve more complex steps and techniques.

Our analysis resulted in an optimal investment strategy tailored to different investment parameters. Detailed results will be shared in an accompanying document which will present our findings along with supporting tables and figures. For a glimpse into how we produced these results, refer to our script via source("code/result\_generation.R") in R.

# Scaling and Cloud Deployment

If the project were to be scaled up significantly, we would leverage cloud-based resources for increased computational power and storage capacity. Services such as Amazon S3 could be used for data storage, while Amazon EC2 instances would provide the required computational power. These services offer scalability and are cost-effective. For executing machine learning algorithms on large data sets, we would use Amazon SageMaker, which provides a complete set of tools to build, train, and deploy machine learning models at scale.

Please note that the information provided above and the related code snippets are placeholders and may need modification based on the actual project implementation. Further, the choice of data storage, machine learning models, and cloud solutions would depend on the specific needs and constraints of the project.

To scale up the data pipeline with significantly more data, the use of cloud-based solutions would be considered. Amazon Web Services (AWS), with its wide array of data processing and analysis services, would be a potential choice. The specific AWS services to be used would depend on the exact requirements of the data pipeline, including the need for high-performance computing, machine learning, data storage, and so on. These cloud solutions would offer scalable resources to efficiently handle the increased data volume, ensuring the continuation of the study without compromising performance.

If this project were to expand significantly in scale, we would utilize cloud-based solutions to handle the increased computational demand and storage needs. We might use services like Amazon S3 for data storage and Amazon EC2 for computational power. These services are scalable and cost-effective. If the project requires executing machine learning algorithms on large datasets, Amazon SageMaker would be an ideal choice as it offers a complete set of tools for developing, training, and deploying machine learning models on a large scale.

**Interpretation**

The interpretation section would elaborate on the findings from the data analysis. It would explain the meaning of the results, tying them back to the initial research question: What are the optimal, quantitative investment strategies for different possible combinations of investment parameters?

In this section, the performance of the proposed investment strategies would be compared and contrasted, examining factors such as return, risk, stability over time, and sensitivity to changes in the investment parameters. This analysis would provide insight into the strengths and weaknesses of the different strategies and would indicate which ones might be most suitable for different investment goals and contexts.

To enhance the understanding of the results, this section might also include a discussion of the economic and financial theories or phenomena that underlie the observed patterns in the data. This could include topics such as market efficiency, behavioral finance, and the impact of macroeconomic factors on asset prices.

The interpretation of the results would involve understanding the implications of the identified correlations, the optimal asset allocations, and the performance of the proposed investment strategies. This would also involve considering the limitations of the analysis and the potential areas for further research.

In this section, we would interpret our findings from the data analysis, relating them back to the initial research question: What are the optimal quantitative investment strategies for different possible combinations of investment parameters? Here, we would assess and compare the performance of the proposed investment strategies, considering elements such as returns, risk, stability over time, and how sensitive they are to changes in investment parameters.

**Limitations and Further Research**

Every research study has its limitations and potential areas for further exploration. The following are a few potential limitations and avenues for further research in this study:

**1. Data limitations:** The data used in this study could have limitations such as missing data points, outliers, or inconsistencies. While these issues would be addressed as much as possible during data cleaning, some residual effects might remain. Furthermore, the scope of the data could limit the generalizability of the results. For instance, if the data mainly covers certain regions or periods, the strategies might not perform as well under different circumstances.

**2. Methodological limitations:** The methods used to analyze the data and construct investment strategies could also have certain limitations. For example, they might make assumptions about the distribution of asset returns or the relationships between variables that do not fully hold in reality. Furthermore, the strategies might rely on certain parameters that need to be estimated from the data, introducing the potential for estimation error.

**3. Computational limitations:** The computation required for data analysis and strategy construction could become a bottleneck, especially as the volume of data increases. While parallel processing and cloud computing solutions could be employed to mitigate this issue, they might introduce additional complexities and potential sources of error.

As for further research, this could include extending the scope of the data to cover more regions, periods, or types of assets, exploring alternative methods for strategy construction, or investigating the impact of various other investment parameters. Additionally, more research could be done on the practical aspects of implementing these strategies, such as transaction costs, regulatory considerations, and investor behavior.

One limitation of this study is the assumption of stable correlations over time, which may not hold in reality. Future research could involve exploring dynamic correlations and their impact on investment strategies. Another potential area for further research could be to include more diverse asset classes and to investigate the effect of active versus passive investment strategies.

Every study has its limitations and potential for further research. Here are a few specific to our study:

1. Data limitations: The data used could have missing values, outliers, or inconsistencies that might affect the analysis. Even though we've tried to address these issues during data cleaning, some effects might still be present.
2. Methodological limitations: The methods used to analyze the data and construct investment strategies might have limitations. They could assume certain things about asset returns or relationships between variables that might not be entirely accurate.
3. Computational limitations: The computation required for data analysis and strategy construction might become a bottleneck as the volume of data increases.

Suggestions for further work:

1. Introduction
2. Expand the scope of research to incorporate more data sources for a broader perspective.
3. Introduce additional investment parameters to increase the versatility of the strategies derived.
4. Apply different machine learning models to compare results and enhance the reliability of the optimal strategy.
5. Research question:
6. Explore the influence of macroeconomic factors on the optimal investment strategy.
7. Evaluate the impact of investor behavior and market sentiments on the choice of strategy.
8. Investigate the role of emerging technologies and alternative investments in shaping investment strategies.
9. Data sources:
10. Incorporate data from additional sources to enhance the robustness of analysis.
11. Consider real-time data analysis to account for rapid market changes and shifts.
12. Employ third-party data validation to ensure the accuracy and credibility of the data sources used.
13. Summary of methods and results
14. Use advanced machine learning algorithms to enhance the efficiency and accuracy of the data analysis process.
15. Integrate a feedback mechanism to continuously update the investment strategy based on evolving market conditions.
16. Consider conducting sensitivity analysis to understand the robustness of the derived strategies to changes in various parameters.
17. Data collection, storage, cleaning, preparation, analysis and visualization:
18. Develop a robust data validation mechanism that can verify the integrity and completeness of the collected data from multiple sources.
19. Implement a more sophisticated data cleaning and preparation process using advanced techniques like machine learning-based imputation for missing values.
20. To improve computational efficiency, consider utilizing parallel processing or distributed computing techniques, particularly during data analysis.
21. Expand the data visualization process by incorporating interactive visualization tools like Shiny in R.
22. Consider implementing additional statistical and machine learning algorithms in the analysis process to gain more profound insights and potentially enhance the prediction of optimal investment strategies.
23. Results
24. To validate our results, we could consider a sensitivity analysis, examining how small changes in the input parameters might influence the optimal investment strategy.
25. We should keep an eye on evolving market conditions as these might warrant a modification of the optimal investment strategy.
26. Interpretation
27. Engage financial experts in the interpretation of results to ensure we consider all possible financial phenomena and factors affecting the investment strategies.
28. Contextualize the findings within the broader economic landscape, including current market trends, to ensure the strategies are applicable and valuable.
29. Limitations and further research:
30. Additional research could cover more regions, periods, or types of assets, explore alternative methods for strategy construction, or investigate the impact of other investment parameters.
31. Future research might also consider practical aspects of implementing these strategies, such as transaction costs, regulatory considerations, and investor behavior.