BOSTON UNIVERSITY

Financial Anomaly Detection Analysis

ASSETTE

Team A2: Chun Zhou, Ji Qi, Jennifer Horita, Arpit Jain, Sri Amruta

Introduction

Business Context:

- Assette LLC. is a cloud based service that allows asset managers to create and provide automated client and sales reports and communications.
- In order to provide the most reliable and accurate data engine, the following Assette project wants to explore and develop tools that can identify anomalous data.

Objectives:

- Identify and visualize anomalies for financial client data using methodologies in Statistics (Moving Average, Exponential Smoothing) and Machine Learning (K-Means)
- We also wish to implement forecasting techniques to identify if future data points could be anomalous.

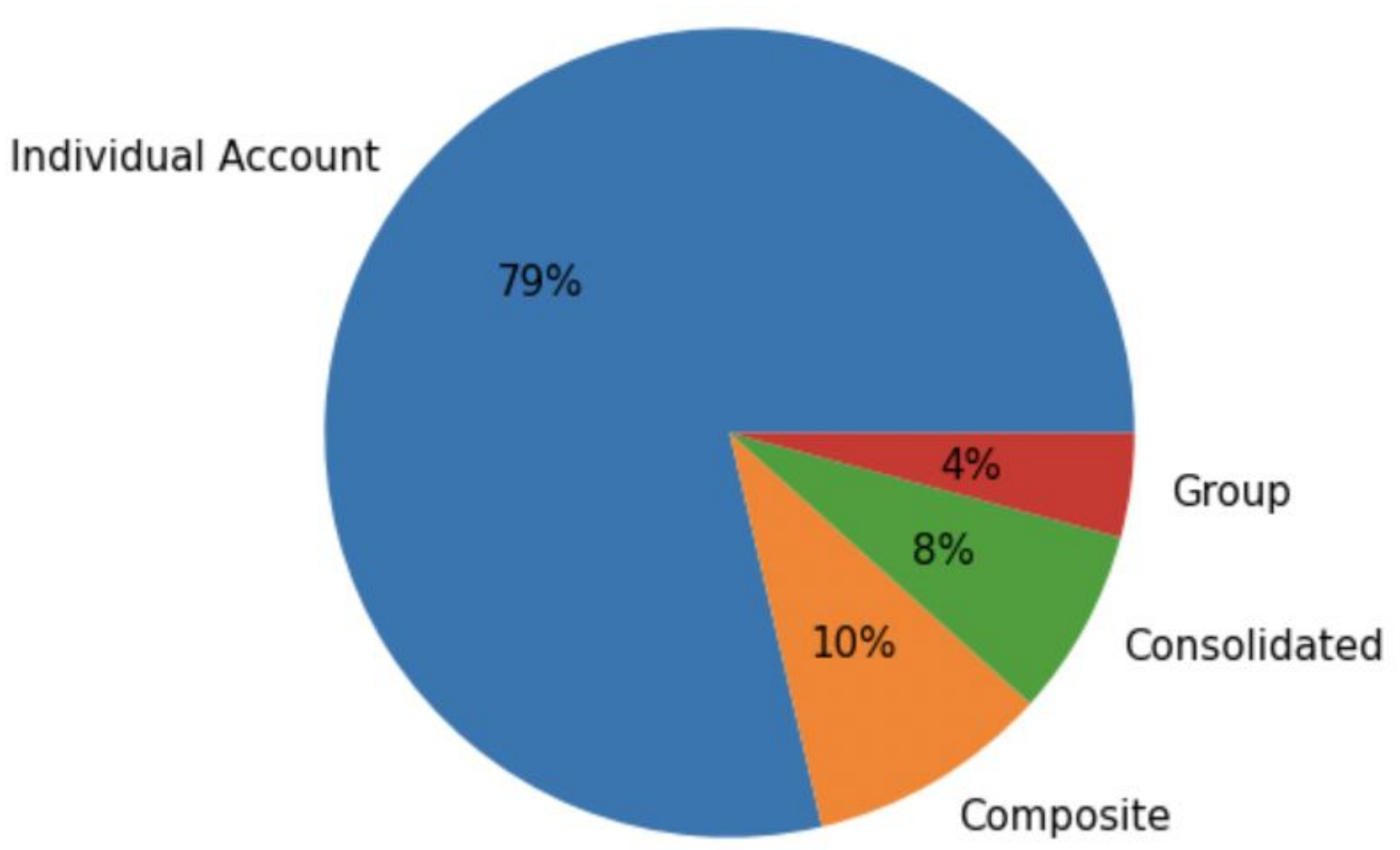
Business Impacts

- Anomalous data can largely affect how an application or engine produces predictive portfolios and reports.
- It is vital that the data used to train models accurately reflects the performance of similar accounts to those that are trying to be created.
- We hope to test and develop an application that can help identify anomalous data early.

Data Cleaning and Analysis

- **Datasets:**
  - Accounts:
    - Listed anonymized Account IDs as well their financial performances
    - 600 unique accounts, 31 product ID, 7 account types and 4 account categories
  - Benchmarks:
    - Corresponds to different accounts as well and index which we could compare the account data to.
    - Ex: S&P 500, Russell 1000 Index, etc..

Figure 1.Distribution of Account Categories



Methodology

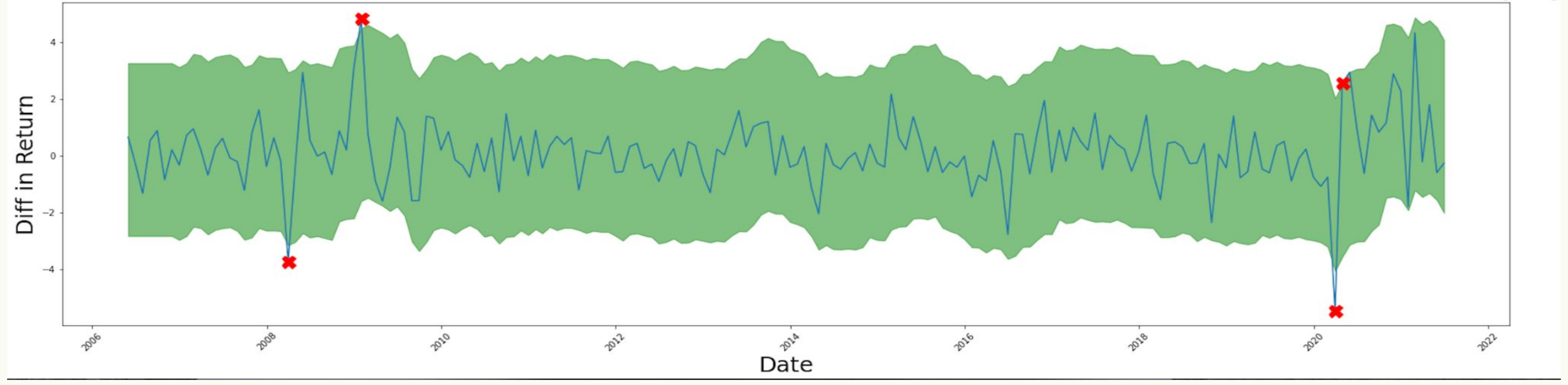
Initially we used Benchmark History dataset to build a baseline model to help detect the outliers for 244 unique Benchmark IDs. The baseline model was basically trained in Simple Moving Average approach. For our second phase, we firstly utilized the Differences between the Accounts Return and the Benchmark Return to determine if there are any anomalies for any account. Moreover, by exploring multiple methodologies such as Simple Moving Average, Exponential Smoothing and K-means, we evaluated and compared model performances by efficiency and the total number outliers identified through each approach. In the following, we will select an account '912' and demonstrate different outlier detection methods and results.

- **Simple Moving Average:**
  - The simple moving average (SMA) is a fast way to capture the pattern in a time series
  - It is the average of the past N data points
  - Best N is 7 with MSE_min equal to 1.60
  - 99% Confidence Interval

$$F_t = \frac{\left(A_{t-1} + A_{t-2} + A_{t-3} + \ldots + A_{t-N}\right)}{N}$$

- N = total number of periods in the average
- A_{t-1} = observation for period t-1
- Point forecast for period t: F_t

Figure 4.
The Return Difference Between "Account 912" & "Benchmark Russell 1000 Value" from 2006-05-31 to 2021-06-30(MA)

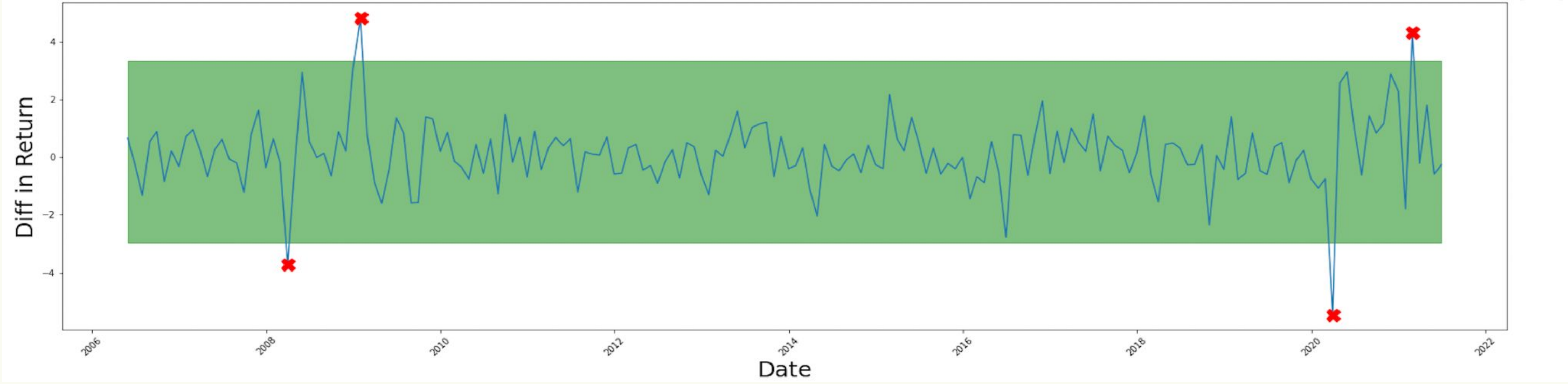


- **Exponential Smoothing:**
  - Recent data are given relatively more weight in forecasting than the older data
  - Include all past observations
  - Best alpha is 0 with MSE_min equal to 1.44
  - 99% Confidence Interval

$$F_{t+1} = \alpha\, y_t + (1-\alpha)\, F_t$$

- F_{t+1}= forecast for the next period
- y_t = time series value in time t
- α = smoothing constant

Figure 5.
The Return Difference Between Account "912" and Benchmark "Russell 1000 Value" from 2006-05-31 to 2021-06-30(ES)



- **K-means:**
  - K-means is a widely used unsupervised clustering algorithm. It creates 'K' clusters of data points. Data instances that fall outside of these groups could potentially be identified as anomalies.
  - According to Figure 3, we notice that the Elbow Curve levels off after 10 clusters, indicating that the addition of more clusters do not explain much more of the variance in our relevant variable.
  - Calculate the distance between each point and the corresponding nearest centroid. The biggest distances are considered as anomaly.

Figure 6.
The Return Difference Between Account "912" and Benchmark "Russell 1000 Value" from 2006-05-31 to 2021-06-30(Kmeans)

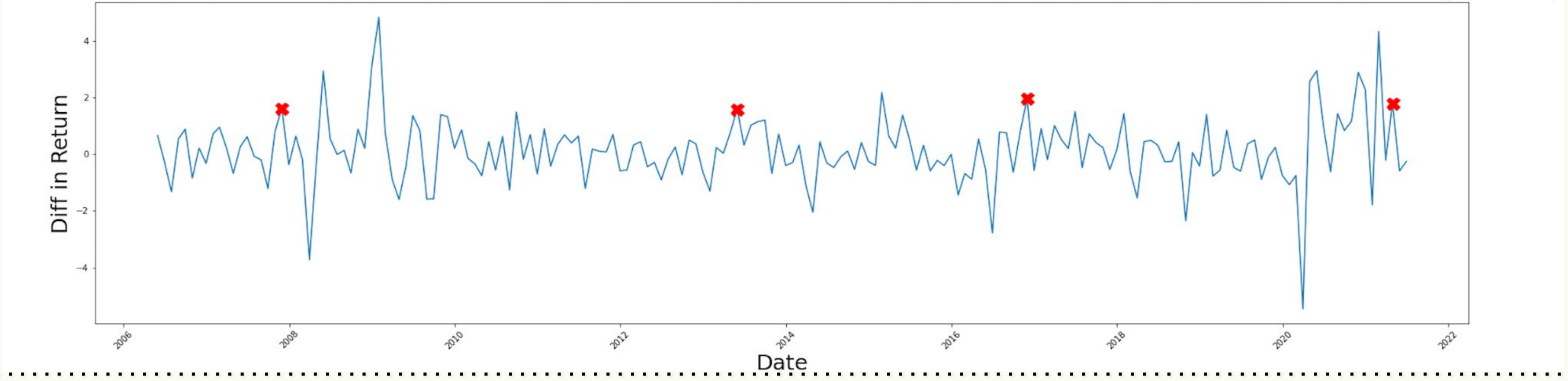


Figure 2: Top 5 Benchmark IDs Used for Accounts

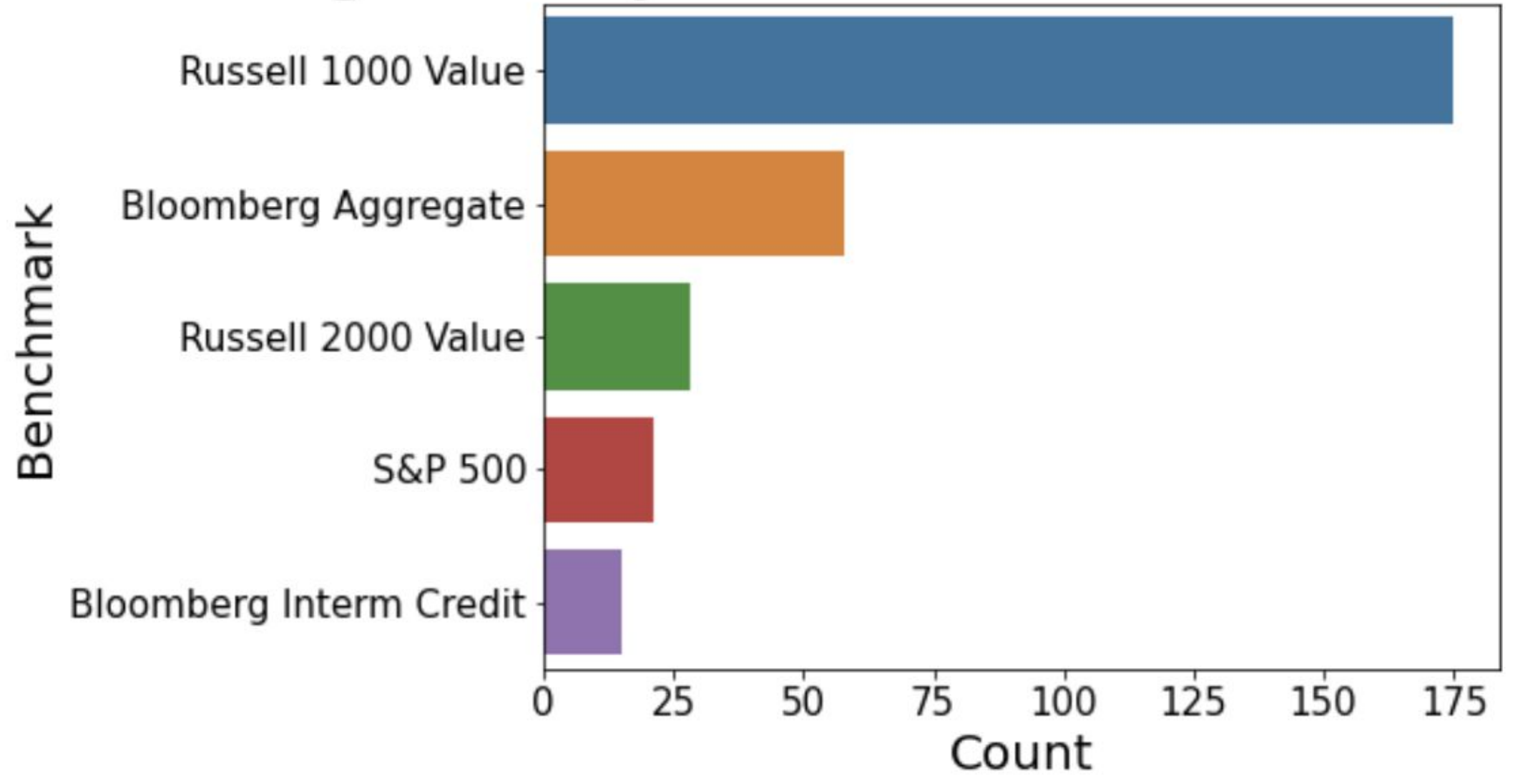
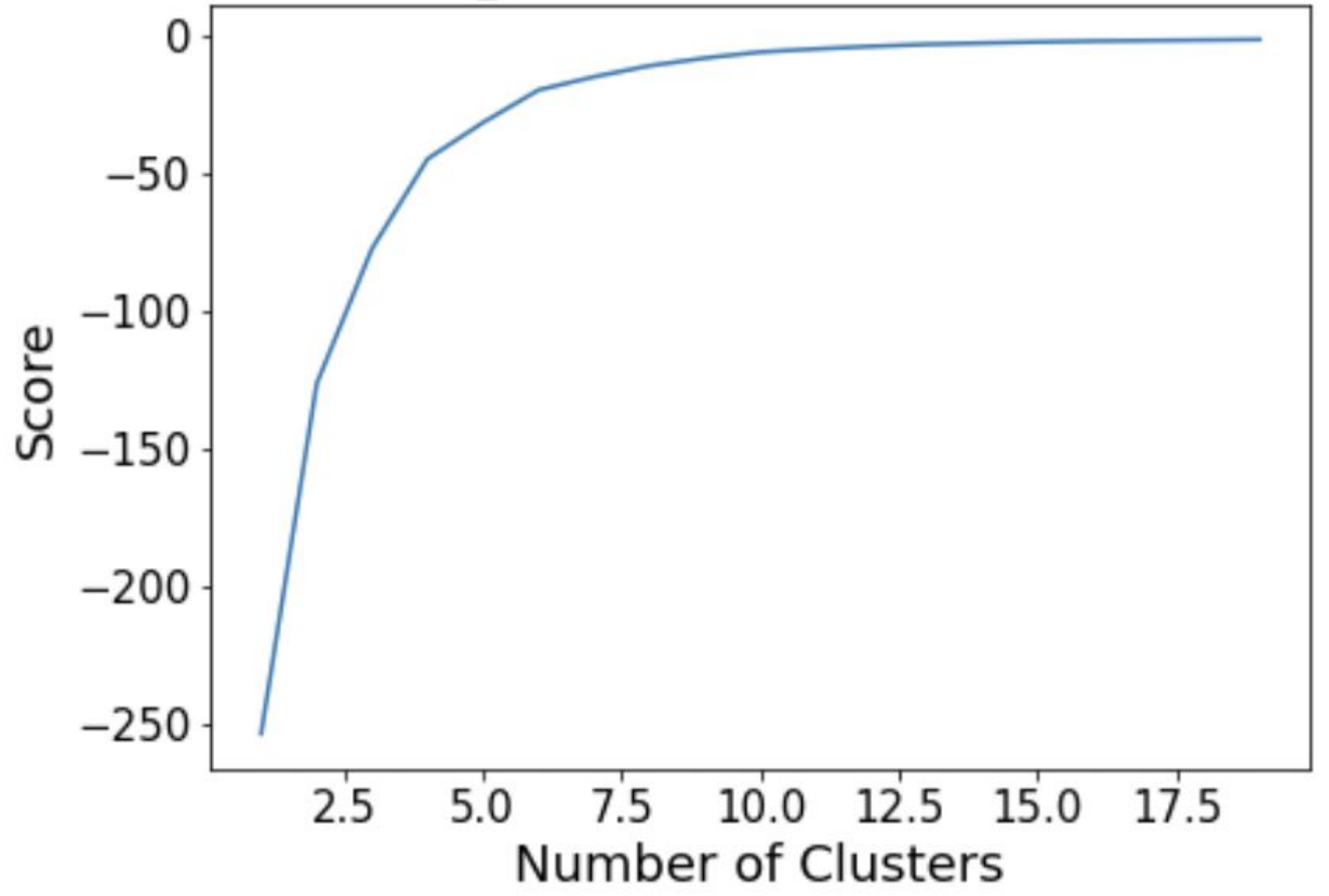


Figure 3: Elbow Curve



Results

- For all 244 benchmarks, there are 19768 out of 314888 data points (6.28%) detected as outliers using 3-year chunks
- 22193 out of 314888 (7.05%) are identified as anomalies by the chunk of 5 years.
- The whole outliers tables and graphs are generated less than 2 mins using one-line code depended on several functions in Python.
- In conclusion, Simple Moving Average method is the most time efficient, and could eliminate any extreme data appeared after N periods. Exponential Smoothing approach include all history data instances, so it is much more suitable for smaller datasets. According to Figure 4 and 5, Above 2 methods generate almost the similar outliers using 99% confidence interval.
- For K-means, it is more helpful for identifying the local anomalies within the longer time series data compared to the other two methods. However, it takes 10 times longer to conduct outlier detections.
- Based on the 'Return Difference' and 'Date', we can see that the outliers detected by all three methods make sense, since the 'Date' matches some important event happened before, such as the Stock market crash in 2008 and outbreak of COVID in early 2022 with the corresponding large negative return differences.

Challenges and Next Steps:

- Require more time to deeply understand the financial concepts behind variables and relationships among hundreds of datasets.
- Need to encapsulate the written functions and classes for future sponsor's use.
- Select the most suitable anomaly detection methods applied on the future datasets.
- Take a comprehensive look into the causes of anomalies to help clients adjust investment strategies .

Acknowledgements: We would like to thank Assette LLC and our Advisor Maneesha Premaratne for their data contribution to the project, our faculty advisor Jialin Li, as well as the Boston University Questrom School of Business for the opportunity to conduct this capstone project.