FAR-Trans Data Visualization Project

Introduction

For this project I will be analyzing a Financial Asset Dataset from which I will attempt to answer a few basic questions to gain a better understanding of investment behavior. This project will eventually help me gain a better understanding of Machine Learning and recommender systems for financial assets. I will attempt to answer the following questions:

1. What is the monthly distribution of transactions for each type of security (Bonds, Stocks, Mutual Funds).
2. Do the most volatile securities provide the most return on investment and who trades them (Individuals, Professionals, Institutions)?
3. Are the most active investors also the most profitable?
4. Are there any behavior changes in investments or customer risk tolerances before and after the COVID pandemic (2021)?

Data Introduction

The data set I will be using for this exploratory data analysis was published by the University of Glasgow. It was intended to provide data for financial asset recommendation systems. The dataset is comprised of several documents which provide customer information, securities such as stocks, bonds, and mutual funds. It also includes a time series of customer transactions for securities between the years 2018 and 2022. The dataset contains customer profile information (excluding personal information) such as risk tolerances, income range, customer type (individual, institution, etc.). Collectively, this dataset contains over 1-million records and over 20 different features from which we draw conclusions.

Data Pre-processing

The FAR-Trans dataset is relatively clean, however the document which contains the asset information has several null/missing values. Although, this does not pose an issue as I will not be utilizing these features for my visualizations. The original research paper from this dataset mentions that it has been pre-processed to address some of the common issues with financial data such as Sell transactions which do not have a Buy transaction preceding them at an earlier date. Assets with closing date gaps greater than 10 days have also been removed from the dataset to ensure consistency. Lastly, stock forward splits and reverse splits have been addressed by transforming the data preceding the split providing a consistent ratio across the time series. I plan on filtering the data based on several criteria as well as performing several aggregate operations to group data according to my needs.

Data Visualization

To visualize this dataset, I chose to use relatively simple methods from Seaborn and Matplotlib. I began by extracting transforming the date of each transaction into a datetime object from which I could further extract months and years to perform aggregation functions. I was able to perform a groupby operation from which I could then make a Catplot with seaborn to visualize the monthly distribution of transactions from the year 2018 to 2022 seen in Figure 1. The particular Greek bank from which this dataset was sourced from greatly benefited from the COVID pandemic as can be seen from the great increase in transactions beginning in 2020 at start of the pandemic and the following years. We can also see a massive spike in transactions in January of 2018; I will need to perform more research to discover the season for this spike.

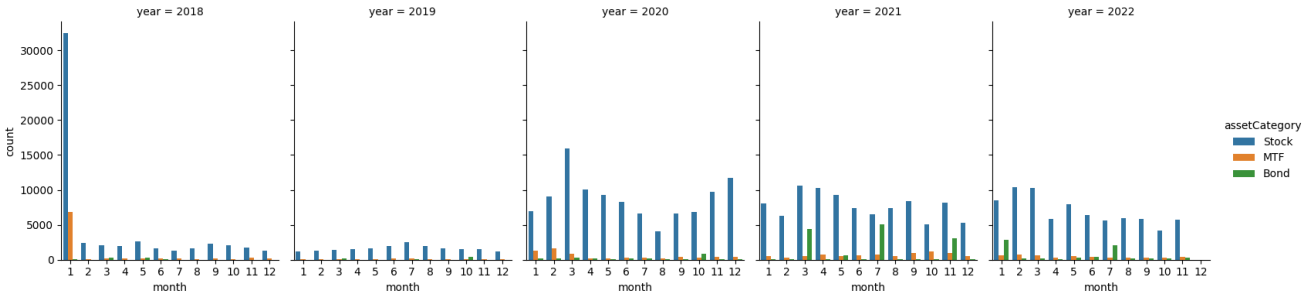

Figure 1: Bar cat-plot of monthly transactions from 2018 to 2022

I am also interested in discovering which are the most volatile securities and what types of customers trade them. I began by visualizing the number of transactions performed by customer type. Figure 2 clearly shows Mass traders, which are non-professional traders with less than 60K in investments, execute most of the trades compared to Professional traders and Legal Entities. While this visualization has not directly answered any of my questions, it will also help me narrow down the most active investors to determine if they are more profitable than less active investors.

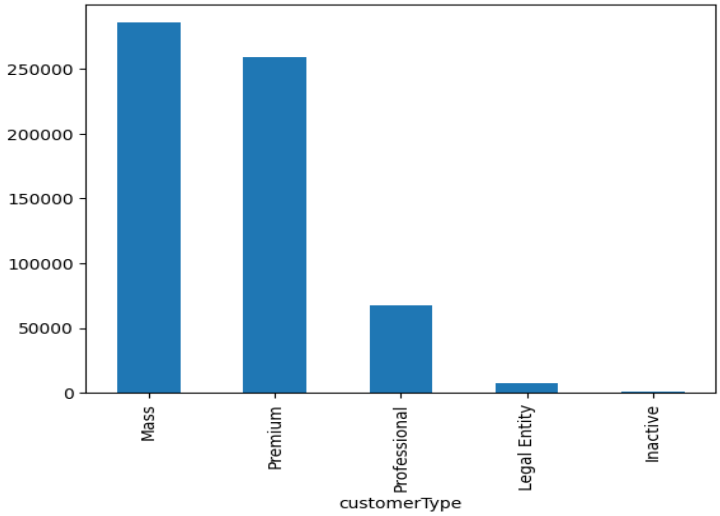

Figure 2: Bar plot of the distribution of transactions by customer types

The visualization in Figure 3 shows the top 10 traded securities ranked in decreasing order of transactions, I plan to analyze these specific securities further to discover their volatility.

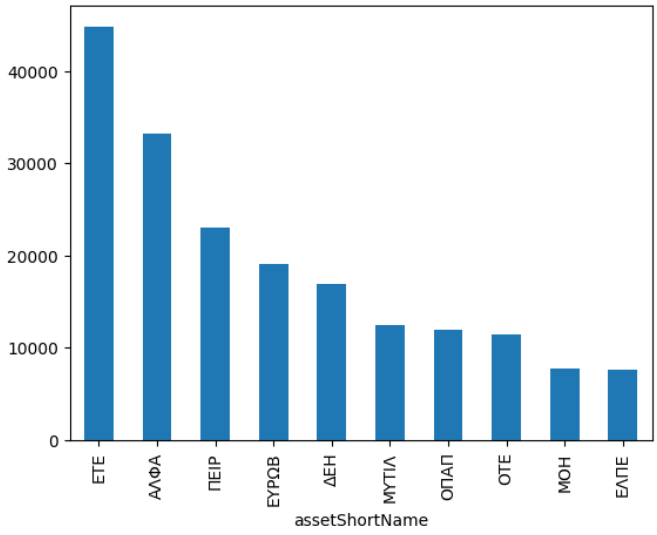

Figure 3: Top 10 securities ranked by number of transactions

The following graph in Figure 4 is a 30-day rolling volatility for the top 10 traded securities, I achieved this by computing the fractional price changes using Pandas pct_change and taking the standard deviation of a 30-day rolling window and multiplying by the square root of trading days which is around 252 days per year. As a result, the first 30-days' worth of observations did not have corresponding volatility consequently they had to be removed from the set. Figure 4 also shows a massive spike in volatility for the ETE ticker symbol, which I will have to investigate further to gain more insight and assess if these provide better return on investment than less volatile securities.

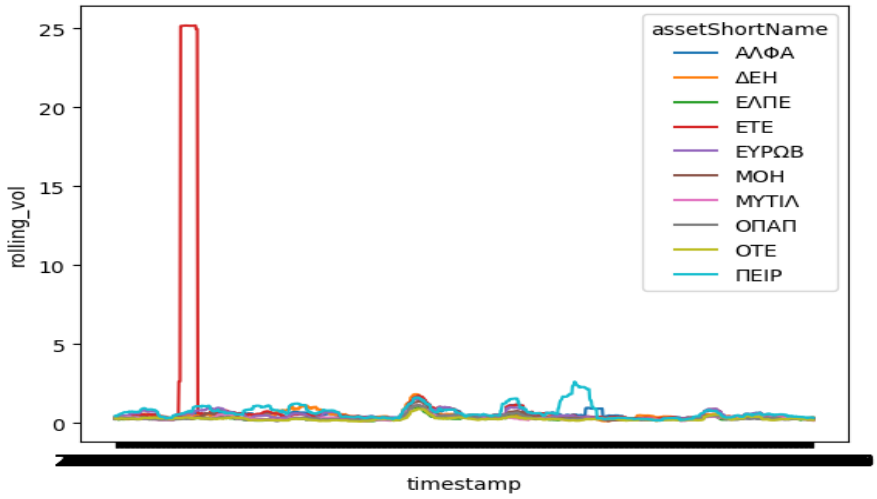

Figure 4: Rolling 30-day volatility of the top 10 traded securities

Storytelling

While I did gain some insight into the FAR-Trans dataset, I was not able to answer many of my questions yet. I aim to study more Time Series Analysis techniques so that I can continue delving into this dataset and possibly expand it into an actual recommendation system. The visualizations above helped me see the complexity of Time Series data and the importance of understanding it. Furthermore, I realized that I would need much more complex data analysis to effectively answer the questions I proposed for this project.

Impact

Since this dataset is sourced from a single financial institution, these visualizations hold no value to other areas of the world as they only represent trading behavior of a small group of investors. Thus, it could be misleading if an individual were to use these as reference for other investments. Additionally, since I am only visualizing the top 10 traded securities to assess if they are most volatile an individual may interpret this as if the remaining securities also follow the same trends.

Sources

Javier Sanz-Cruzado, Nikolaos Droukas, Richard McCreadie. FAR-Trans: An Investment Dataset for Financial Asset Recommendation. IJCAI-2024 Workshop on Recommender Systems in Finance (Fin-RecSys). Jeju, South Korea, August 2024.