1 Robinhood Dataset

1.1 Description of the Dataset

This data is retrieved from https://robintrack.net/, the creator retrieved data from the official Robinhood API.

The dataset contains the **number** of Robinhood users holding at least one share of 8,221 securities.

> *This is a key feature of the dataset, we cannot track the holdings of individual investors, nor we know the amount of money or stocks invested in a certain security, we also cannot know the buy/sell flows on stocks, although we can proxy it with the change in holders.*

The available data spans from February 5, 2018, to August 13, 2020, covering 818 days (data is available also for non-trading days and some days are missing). Although the data was originally recorded hourly, I aggregated it to a daily frequency by computing the average number of holders per day.

Handling Missing Values The original dataset contains missing values for 3,331 securities, primarily in the earlier periods. To ensure consistency in the comparison we adopt the same method in [Fedyk, 2024]

1.1.1 Distribution of Key Features (Log-Transformed)

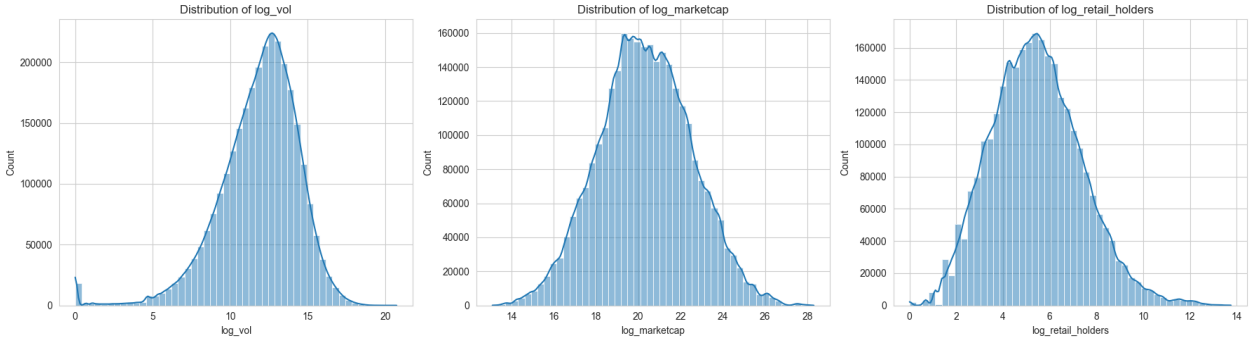
The distributions of trading volume, market capitalization, and retail holders were initially highly skewed, with a few extreme values dominating the dataset. To address this, I applied a logarithmic transformation: $x' = \log(1 + x)$.

This transformation reduces the impact of outliers, enhances interpretability by making the data more symmetric, and facilitates comparisons between stocks of different sizes.

The key observations after applying the transformation are:

- **Trading Volume:** The distribution appears approximately normal, centered around a peak, with a slight left tail. While most stocks have relatively low trading volume, a few highly traded stocks, such as large-cap or meme stocks, exist but no longer dominate the distribution.

- **Market Capitalization:** The transformed market capitalization data exhibits a bell-shaped curve, suggesting a more balanced spread across small, mid, and large-cap stocks. However, some large-cap stocks remain in the extreme right tail, indicating that a few companies, such as Apple and Microsoft, are significantly larger than the majority.
- **Retail Holders:** The number of retail holders follows a roughly log-normal distribution, confirming that a small number of stocks attract massive retail participation while most remain relatively unpopular. The left tail suggests that many stocks have very few retail holders, reinforcing the notion that retail trading is concentrated in a subset of securities.



1.2 Comparing the Portfolios

1.2.1 Different Approaches

[Fedyk, 2024] and [Welch, 2022] use the same approach to build the performance of the Robinhood crowd (or "reference index"): they build daily weights and then apply the weights from the precious day to returns of stocks. They directly build portfolio returns.

I tried, on the other hand, to build a index, applying weights to prices rather than returns.

To build a representative portfolio of the average Robinhood investor, it is necessary to retrieve the prices of the securities. A capitalization-weighted approach can be used, multiplying the price of each security by the number of users who hold it. This approach assumes that all Robinhood users hold a similar number of shares for a given ticker, or that the distribution of shares held per user follows a normal distribution.

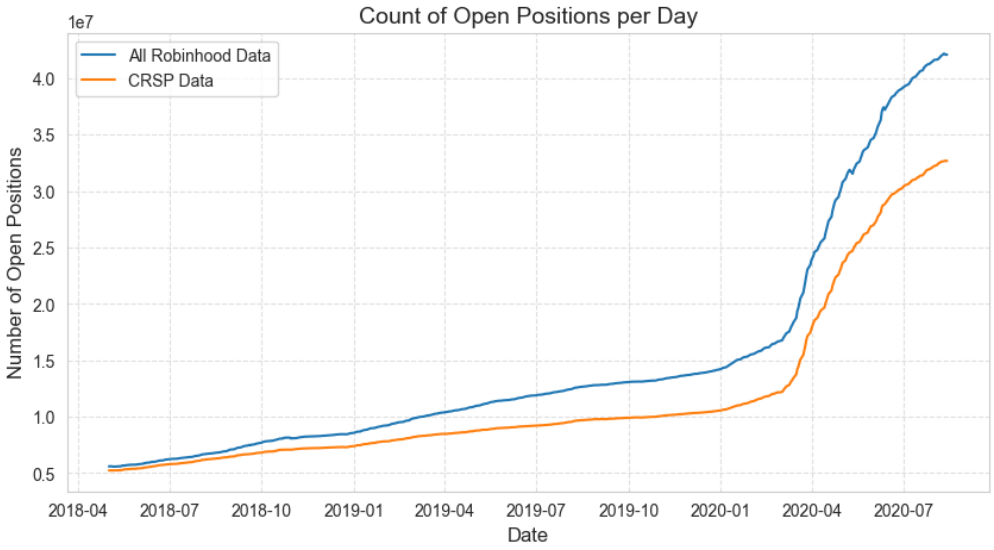
Over the years covered in the dataset, Robinhood has gained a significant number of users. Data on active users is available on Statista[1], though only on a yearly basis. Comparing the

[1] https://www.statista.com/statistics/822176/number-of-users-robinhood/

Statista figures with Robinhood's reported numbers for 2023 suggests that the active user count corresponds to December 31 of each year. This data could later be used to normalize the number of users and build a reference portfolio.

The total number of open positions can be computed as the sum of all investors who hold at least one security in each asset, effectively a row-wise sum of the dataset.

Market data for all securities was retrieved from the CRSP[2] database, accessed via WRDS. However, only 8,099 securities are available in CRSP, as it focuses exclusively on American assets. The difference in open positions between the full dataset and the CRSP subset is minimal. If, instead, all securities with missing values are dropped, leaving only 5,221 securities, the gap widens.



The graph illustrates the count of open positions per day on Robinhood from April 2018 to mid-2020, showing a steady increase over time, with a sharp acceleration in early 2020. This surge aligns with the onset of the COVID-19 pandemic, which likely drove a significant influx of new retail investors seeking market opportunities amid economic uncertainty and stimulus checks.

1.2.2 Retail Investors Prefer "Famous" Stocks

The majority of the securities are common shares, representing about 57.9%. ETFs represent about 23.7% and other funds are the 9.2% of the dataset. Other structured investments, REITs, and ADRs cover the remaining part.

[2]The Center for Research in Security Prices, based at the University of Chicago, provides high-quality historical market data widely used in finance research and investment analysis.

Analysing the securities by market capitalisation about 82.9% is represented by stocks and 9.6% by ETFs. If we look at the "Retail Market Cap" (i.e. number of positions times price), 89.2% of securities are stocks and 5.8% are ETFs.

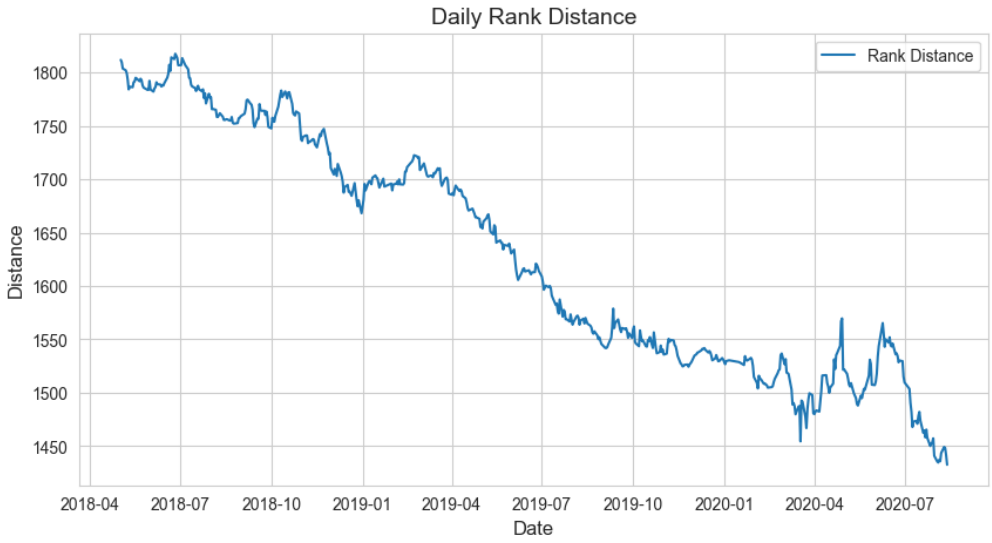
Looking at the securities Robinhood users prefer holding, ranked by "Retail Market Cap", investors prefer holding smaller cap stock. A qualitative analysis shows "famous" stocks, such as Tesla, Starbucks, and Nvidia to name a few, to appear among the most popularly owned.

1.2.3 Possible Measures of Divergence

Rank Distance To describe the preference of retail investors for smaller cap stock I propose the following measure:

$$d_R = \sum_{i=1}^{N} \frac{R_i^{\text{Mkt}} - R_i^{\text{RH}}}{R_i^{\text{RH}}}$$

Where R_i^{Mkt} is the rank of the i^{th} security by market cap, and R_i^{RH} is the rank by retail market cap. The normalization by R_i^{RH} reduces the impact of small-cap stocks with minor ranking differences.



The plotted Daily Rank Distance suggests a clear downward trend from early 2018 to mid-2020, indicating that the ranking of stocks by retail market cap has become increasingly aligned with the ranking by total market cap. Initially, the distance is above 1800, gradually declining towards 1450. This implies that retail investors, who originally exhibited a stronger preference for smaller-cap stocks, have progressively shifted towards stocks that are more representative of the broader market.

Between 2018 and 2019, the decline is relatively steady, reflecting a gradual change in retail investment behavior. However, the trend accelerates in 2019 and 2020, suggesting a more pronounced shift. The beginning of 2020 shows increased volatility, with occasional upward spikes, which could be attributed to market disruptions, possibly linked to the COVID-19 crash and the subsequent retail trading boom. The rapid expansion of retail investing during this period, fueled by stimulus checks and zero-commission trading, may have led to temporary deviations, but the overall trend continues downward.

A sustained decrease in rank distance suggests that retail investors have moved closer to institutional preferences, potentially increasing their exposure to large-cap stocks or index-tracking assets. If this trend persists, it would indicate a continued assimilation of retail behavior into the broader market structure. Conversely, a reversal in this pattern could signal renewed speculative activity or a shift back to small-cap stocks.

Appendix A Tables

Table 1: Descriptive Statistics for Daily and Rolling Returns

	count	mean	std	min	25%	50%	75%	max
rh_portfolio	564	0.000719	0.018809	-0.132368	-0.006164	0.001141	0.009484	0.072851
mc	564	0.000396	0.015470	-0.125496	-0.003944	0.001012	0.006481	0.086673
VOO	564	0.000438	0.015806	-0.124870	-0.003874	0.000942	0.006632	0.091087
VT	564	0.000184	0.015092	-0.123763	-0.004568	0.000842	0.005926	0.087470
rh_portfolio_1_return	564	0.000719	0.018809	-0.132368	-0.006164	0.001141	0.009484	0.072851
mc_1_return	564	0.000396	0.015470	-0.125496	-0.003944	0.001012	0.006481	0.086673
VOO_1_return	564	0.000438	0.015806	-0.124870	-0.003874	0.000942	0.006632	0.091087
VT_1_return	564	0.000184	0.015092	-0.123763	-0.004568	0.000842	0.005926	0.087470
rh_portfolio_5_return	564	0.003309	0.041768	-0.224427	-0.012162	0.004643	0.022078	0.153755
mc_5_return	564	0.001940	0.031094	-0.207508	-0.008577	0.004961	0.016049	0.151511
VOO_5_return	564	0.002152	0.030933	-0.204425	-0.009377	0.005838	0.016464	0.162820
VT_5_return	564	0.000864	0.030673	-0.214262	-0.010938	0.003977	0.014857	0.151788
rh_portfolio_30_return	564	0.015173	0.133421	-0.482722	-0.041325	0.020205	0.051931	0.408751
mc_30_return	564	0.009890	0.080443	-0.401515	-0.015223	0.025072	0.046527	0.246006
VOO_30_return	564	0.011115	0.078031	-0.401950	-0.011124	0.028635	0.046654	0.252864
VT_30_return	564	0.003681	0.079260	-0.406688	-0.020644	0.018305	0.039820	0.224464
rh_portfolio_60_return	564	0.010727	0.177731	-0.377518	-0.066831	0.001284	0.070271	0.641470
mc_60_return	564	0.016554	0.099118	-0.356261	-0.013618	0.029848	0.062050	0.338109
VOO_60_return	564	0.019496	0.095149	-0.355392	-0.009450	0.033666	0.066970	0.337947
VT_60_return	564	0.004301	0.101282	-0.385941	-0.024094	0.012486	0.057122	0.328680
rh_portfolio_120_return	564	-0.014462	0.114741	-0.310504	-0.098157	-0.014206	0.053267	0.370780
mc_120_return	564	0.017214	0.075478	-0.307022	-0.031108	0.030972	0.070401	0.218409
VOO_120_return	564	0.024238	0.074347	-0.302877	-0.027697	0.034991	0.076471	0.231377
VT_120_return	564	-0.006474	0.077296	-0.333294	-0.056443	0.008282	0.042280	0.186378
rh_portfolio_564_return	564	-0.011075	0.123340	-0.382891	-0.055196	-0.014086	0.045442	0.405724
mc_564_return	564	0.071788	0.069591	-0.200798	0.033823	0.070623	0.106613	0.224508
VOO_564_return	564	0.094275	0.071760	-0.168586	0.051783	0.090626	0.134336	0.251507
VT_564_return	564	0.001882	0.063696	-0.301083	-0.024162	0.014363	0.032343	0.121970

Table 2: Descriptive Statistics for 1-Day and 5-Day Returns, Covering the Whole Period
Note: Positive returns indicate the percentage of days in which the log returns were greater than zero.

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**positive returns**
rh_portfolio_1_return	564	0.000719	0.018809	-0.132368	-0.006164	0.001141	0.009484	0.072851	0.553191
mc_1_return	564	0.000396	0.015470	-0.125496	-0.003944	0.001012	0.006481	0.086673	0.558511
VOO_1_return	564	0.000438	0.015806	-0.124870	-0.003874	0.000942	0.006632	0.091087	0.563830
VT_1_return	564	0.000184	0.015092	-0.123763	-0.004568	0.000842	0.005926	0.087470	0.547872
rh_portfolio_5_return	560	0.003281	0.041909	-0.224427	-0.012379	0.004643	0.022098	0.153755	0.598214
mc_5_return	560	0.001913	0.031198	-0.207508	-0.008632	0.004961	0.016300	0.151511	0.630357
VOO_5_return	560	0.002128	0.031036	-0.204425	-0.009442	0.005838	0.016494	0.162820	0.635714
VT_5_return	560	0.000839	0.030779	-0.214262	-0.011044	0.003977	0.014911	0.151788	0.583929

Table 3: Descriptive Statistics for 1-Day and 5-Day Returns, up to February 3rd 2020
Note: Positive returns indicate the percentage of days in which the log returns were greater than zero.

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**positive returns**
rh_portfolio_1_return	430	0.000115	0.013490	-0.050597	-0.005461	0.000809	0.007377	0.068808	0.537209
mc_1_return	430	0.000419	0.008745	-0.032113	-0.003126	0.000804	0.005285	0.045916	0.553488
VOO_1_return	430	0.000485	0.008928	-0.032828	-0.003066	0.000757	0.005096	0.049350	0.558140
VT_1_return	430	0.000198	0.008361	-0.031068	-0.003794	0.000716	0.004853	0.036545	0.546512
rh_portfolio_5_return	426	0.000259	0.026549	-0.105948	-0.013623	0.002922	0.014899	0.088194	0.570423
mc_5_return	426	0.002091	0.019395	-0.075729	-0.008188	0.004110	0.014121	0.063052	0.624413
VOO_5_return	426	0.002442	0.019790	-0.081061	-0.008308	0.004981	0.014449	0.067072	0.636150
VT_5_return	426	0.001031	0.018612	-0.066412	-0.010824	0.002804	0.013208	0.060003	0.565728

References

[Fedyk, 2024] Fedyk, V. (2024). This time is different: Investing in the age of robinhood. Working paper, Arizona State University. Version dated November 2024. Check author's website or SSRN for latest version.

[Welch, 2022] Welch, I. (2022). The wisdom of the robinhood crowd. *Journal of Finance*, 77(3):1489–1527.