1. Background

Factor investing has enjoyed its success as one of the forefront strategies in equity investment. The paradigm started out as mainly using fundamental analysis to analyze companies, as seen in the three-factor model (Fama and French 1993), consisting of the market, size, and value factors. The model is an expansion of Eugene Fama's and Kenneth French's Capital Asset Pricing Model, explaining a stock's character by only measuring the risk and return relative to the market. Presently, factor investing has expanded its analysis with the inclusion of technical and quantitative aspects, employing factors such as momentum (Asness 1994; Carhart 1997; Jegadeesh and Titman 1993) and volatility. Therefore, factor investing utilizes the three pillars of investing: fundamental, technical, and quantitative analyses. From this standpoint, the goal of the model is to implement a portfolio that enhances diversification, drives above market-relative returns, and decreases risk. Due to different risk premium characteristics, some factors might favor certain market conditions, causing those factors to fail. But because of diversification benefits, each factor's risk will compensate each other, making the overall model effectively timeless. There are many attempts of bending the original factor model from time to time, adding and tweaking factors with the goal to accomplish a greater risk premium. For example, time series momentum (Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen 2011), is an alternative strategy to factor investing's cross sectional momentum, capturing an individual asset's momentum anomaly instead of analyzing a group of assets' momentum.

2. Problem Statement

Throughout the majority of asset management history, factor investing was executed by separating each factor in silo portfolios; one might have a portfolio for value and a separate portfolio for quality. Recent studies have illustrated that with the birth of quantamentals investing, investors have attempted to combine fundamental and technical factors under a quantitative framework. The term quantamentals is a portmanteau combining "quant"itative and fund"amental" investing; it defines a fusion of computer and mathematical models with fundamental methods' that analyzes company characteristics. The framework is still young and research on the topic is scarce, which may imply that 1. The underlying practical approach to quantamentals is idiosyncratic and confidential to each investing entity and 2. There are a plethora of approaches (no ideal way) to execute a quantamentals model. This proceeding induces an opportunity to utilize data science to explore the risk premium of equity factor investing under the quantamentals framework and to introduce a quantamentals strategy using machine learning.

3. Proposed Solution

Referring back to factor investing, the method attempts to rank assets based on different characteristics. Traditionally, investors would group assets according to a single characteristic, by selecting assets to invest by ranking by factor magnitudes. We remedy for the lack of factors interaction effect by using a K-means Clustering model. This way, we let the algorithm determine the valid clusters of different factors and we will evaluate the risk premium by backtesting on each cluster. By using a (unsupervised) clustering model, we are essentially creating a quantamentals model that groups assets and combines fundamental and technical factors without having constraint rules that might overfit in unseen data. We are also using clustering to determine unknown interactions; i.e. assets with high quality and high momentum might provide better profits than assets with high value and high quality. To optimize our K-means algorithm, we have to determine the optimal 'K' for the number of clusters. We propose two metrics as evaluation; the elbow method and the silhouette score. The silhouette score is chosen because it can be evaluated based on mathematics, while the elbow method is more feasible for visual evaluations. Therefore, the elbow method is an ambiguous and subjective procedure, lacking an efficient method to determine what actually represents an 'elbow'. *Figure 1* demonstrates this phenomenon. Details on optimization metrics will be discussed in *Section 7 : Evaluation Metrics*.

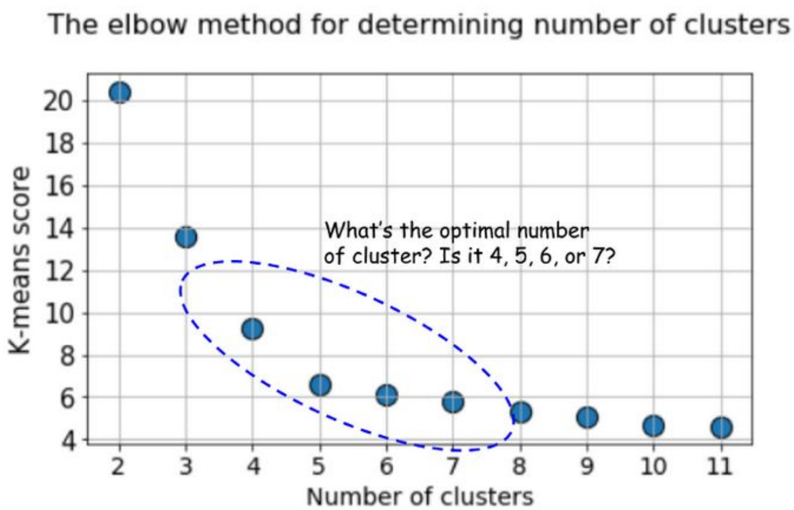

Figure 1: Choosing optimal 'K' with the elbow method is ambiguous
Image taken from 'Clustering metrics better than the elbow-method' by Tirthajyoti Sarkar

4. Data

4.1 Investment Universe
We will be basing our research on the Russell 3000 universe of stocks, containing the top 3000 stocks ranked by market capitalization. This is in conjunction with the size factor, and ensures sufficient trading volume. Our data's time window will start in 2002-01 and end in 2020-09.

4.2 Data Sources

All data sources are paid subscriptions.

Norgate Data - Norgate is a bias-free data provider that provides economic indicators, historical constituents, and end-of-day price data. Their data is survivorship bias-free, which is extremely crucial since we are basing our analysis on a universe like Russell 3000. Price data is also back-adjusted if needed, with the option of 'total return' adjustment, where price is accounted for dividends and splits. We will be keeping track of the active/delisted constituents of Russell 3000 and will be using daily price data from Norgate to backtest and create price derivatives (i.e. momentum factor) if needed. Norgate data is stored locally, and it provides software to keep an asset's price data as a CSV. However, we will be using an api to access the data.

Sharadar - Sharadar provides numerous datasets including historical fundamentals and price data. They are also survivorship bias-free. We will be utilizing Sharadar's fundamentals dataset to analyze companies' factors like quality and value. Particularly, we will be using the 'As Reported' data, since this is forward looking-bias free and contains metrics as reported on every quarter and is not a restatement of a company's report. Data is stored locally as CSVs.

4.3 Initial Concerns

Since we will be using two different data providers, one obstacle that arises is the difficulty in merging and cleaning data. Symbol names are different in each provider, therefore a dictionary reference in joining technical and fundamental features has to be created. Also, a company may fail to report their financials resulting in missing data. Because of this, we will be using trailing 12 months as reported data. Sharadar offers an option on choosing 'as reported' (AR) or 'most recent reported' (MR) data. The MR data contains restatements of company financials after the official report date. The AR data contains financial statements of the company during the official reporting period, which is suitable for backtesting.

5. Introducing the Factors of Factor Investing

In this section, we introduce the factors that were chosen to be analyzed for our investment model. Factors are broadly divided into fundamental and technical categories.

5.1 Fundamental Factors

These are the metrics that can be found in a company's quarterly balance sheet and income statement.

5.1.1 Value Factor

Value investing is perhaps the most popular approach to stock investing. One can observe its long history in the investment realm, being employed by the forefathers of value investing like Benjamin Graham and Warren Buffet. Graham's book on 'The Intelligent Investor' (1934) was considered to be the pioneer of value. Later on, Fama and French adopted the idea in an academic framework, defining value as the book-to-market ratio which is used as the value factor alongside the size and market factors in 'The Fama-French Three Factor Model'. There are a plethora of metrics that are used to define a stock's value, and deciding on one ideal factor is a subjective issue. Nevertheless, the theory behind value is that relatively cheap stocks tend to outperform relatively expensive stocks.

For our metric of choice, we will utilize the EBIT (earning before interest and taxes) to EV (enterprise value). This metric is used by Joel Greenbatt, a highly praised value investor and professor at Columbia University. The higher the EBIT/EV multiple, the better for the investor as this indicates the company has low debt levels and higher amounts of cash. Joel Greenbatt statement highlights the EBIT/EV ratio's benefits: "Using the metric allows us to put companies with different levels of debt and different tax rates on an equal footing when comparing earnings yields."

$$EBIT/EV\ Formula = \frac{Earning\ Before\ Interest\ and\ Taxes}{Enterprise\ Value}$$

5.1.2 Quality Factor

A company's quality or profitability is a more recent factor compared to value. The underlying theory is that there is a causality between the measure of high profitability and high returns. Common values that explain a company's profitability are return of equity, return on investment capital, and gross profitability.

We will use ROIC (return on investment capital) to assess a company's quality, as employed by Joel Greenbatt. ROIC is the amount of return a company makes above the average cost it pays for its debt and equity capital. It is always expressed as an annualized value, calculated on a trailing-12 month basis. This is one of the reasons why we'll be using the ART (as reported trailing 12 month) data from Sharadar.

5.2 Technical Factors

These are the derivatives that can be calculated from a stock's price.

5.2.1 Cross-Sectional Momentum

Cross-Sectional Momentum defines the causality between relatively strong historical performance and high future returns. The relatively strong performers would have a relatively higher return than weak performers. This type of momentum is also called relative momentum, since it compares the performance of stocks in a universe. The article is documented by Jagadeesh and Titman in 1993 in the paper 'Returns of Buying Winners and Selling Losers: Implications for Stock Market Efficiency'.

We define the cross-sectional momentum as the trailing-6 month return, skipping 1 month, due to the natural short term reversal effects of the market. This is a well known study in the topic of behavioral finance.

5.2.2 Low Volatility

The Low Volatility, a.k.a. Low Beta factor is the phenomenon of low volatility stocks outperforming high volatility stocks. Low risk equals higher returns. This Low Volatility factor challenges the CAPM model, where the former promotes buying low beta stocks and the latter promotes buying high beta stocks.

We will use the 100-day lookback standard deviation to define our volatility factor.

5.2.3 Time Series Momentum

Time Series Momentum, or absolute momentum the other type of momentum describes a strategy where one uses a stock's historical performance to decide on going long or going short. This is also employed by entities in the CTA industry under the term 'trend following'. The popular metrics used to calculate time series momentum is the rate of return and moving averages, with either one returning almost the same effects. Evidence of absolute momentum can be found in the paper, 'A Century of Evidence on Trend-Following Investing" by Brian Hurst, Yao Hua Ooi and Lasse H. Pedersen'.

We will use the 100 and 200-day moving average to assess the historical trend of stocks.

6. Strategy and Model Pipeline

Our pipeline can be summarized in the following steps.

6.1 Use Norgate Data to access the Russell 3000 Historical Constituents from 2002-01-01 till 2020-09-29. We will also query for time series data of all the constituents, indicating on each day whether the stock is part of the index or not.

6.2 Use the list of constituents as a basis for querying time series data from Norgate and Sharadar; stock prices and fundamental (ART) data respectively.

6.3 From the time series stock prices, we perform feature engineering to transform into derivatives: avg 200-day volume, 6-month returns (momentum), and 100-day trailing std (volatility). We use the ART data to fetch time series of ROIC (Quality), EBIT and EV (Value) metrics.

6.4 Form an all factor time series that contains information on all constituents. On each investment rebalance day (observation), we filter stocks that are part of the index, then filter the top 1000 stocks ranked by their avg 200 day volume.

6.5 Pre process data of stocks using quality, value and volatility factors.

6.6 Apply K-means clustering to cluster stocks by quality, value and volatility.

6.7 Assess each cluster's risk premium by backtesting each cluster. Rebalancing monthly.

6.8 Filter each cluster's stocks by the top 20 relative momentum values. Each cluster now contains 20 stocks.

6.9 Assess each cluster's risk premium by backtesting each cluster. Rebalancing monthly.

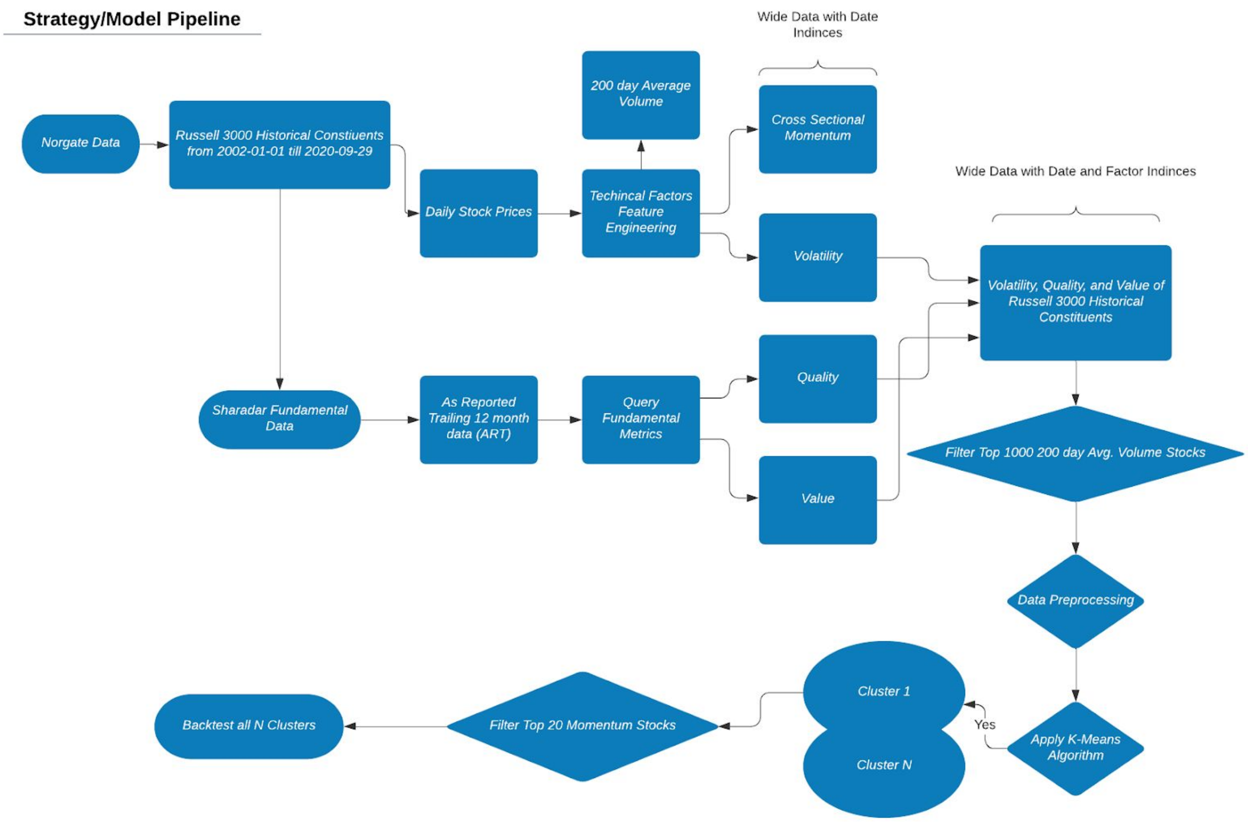

Figure 2: Strategy and Model Pipeline

7. Evaluation Metrics

7.1 Optimizing for K of clusters with Silhouette Score

In K-means clustering, one obvious problem is determining the number of K clusters that will produce an optimal solution. We utilize the silhouette score in determining the 'optimal K'. To recall, for a given element i and its cluster, the silhouette coefficient S_i is defined as

$$S_i = \frac{b_i - a_i}{max\{a_i , b_i\}};\ i\ = 1, ..., N,$$

where a_i average distance between element i and all other elements in element $i's$ cluster, and b_i is the average distance between element i and all other elements outside element $i's$

cluster. This measures compares inter-cluster and intra-cluster distances. A value of 1 means an element is clustered well, and a value of -1 is clustered badly. For a cluster partition, the clustering quality q is calculated as

$$q = \frac{mean\ (\ \{S_i\ , .. , S_n,\})}{\sqrt{var\ (\{S_i\ , .. , S_n,\})}}$$

We select K with the highest clustering quality.

7.2 Backtest metrics

We will evaluate the risk premium of our clusters by backtesting each cluster and comparing the performance with the SPY ETF (SP500). Metrics used for evaluation are the Sharpe Ratio, Cumulative Annual Growth Rate, Cumulative Returns, Annual Volatility and Maximum Drawdown.

8. Exploration and Data Wrangling

8.1 Russell 3000 Historical Constituents

Norgate provides an api to fetch the historical time series describing a series of binaries of whether the asset is a part of Russell 3000 or not. We use this to initiate our historical universe of stocks. We extract all active/delisted symbols and then make a wide constituent dataset of asset columns and binaries on each day. We then explore the number of constituents on each day.

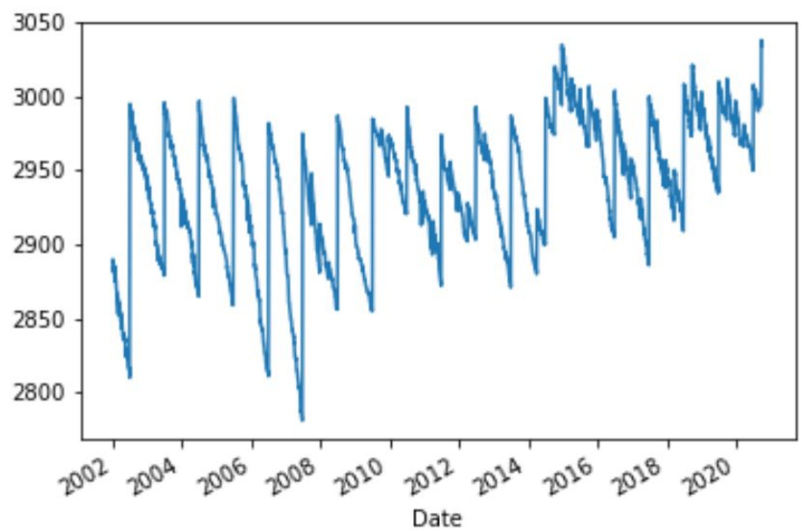

Figure 3 Russell 3000 Constituents Time Series from 2002-01-01 t 2020-09-29

Observing the constituents indicates that the number of constituents are not consistent at 3000. We can see a number of volatile shifts throughout the years and this might be due to periods of assets getting delisted and some lag time before new assets get listed. Nevertheless, for our interest, the goal is to test out risk premiums of stocks and we would have at least 2800 stocks at each point in time.

8.2 Communicating between Norgate and Sharadar's nomenclature

One challenge in handling data comes from aligning symbol names between Norgate and Shardar's symbol format. We decided that we will be basing our symbol formats on Norgate; one reason because Norgate should contain a more complete database. Some symbols data might not be available in Sharadars, and in the end we would use the intersection between Norgate and Sharadar's available symbols as our symbol universe.

Some reasons that Sharadar's fundamental data might not be complete:

- Some of the constituents might be a fund and not a company, which means they would not have value and quality metrics that we seek for
- A company can have a series of spin offs and they are named differently, but fundamentally they are the same company. For example, FWONK (Norgate) is a series C company, but Sharadar does not have fundamental data for it. Instead Sharadar has FWONA, the series A company. However, FWONA was not listed as a Russell 3000 constituent but the two companies would share the same fundamentals nevertheless. For this reason, we chose to omit these kinds of companies since we have no way of knowing the references/dictionary, at least in terms of programming logic.
- The company's data is missing

Norgate	Sharadar
AAC	AAC
AAC-200004	AAC1
AAC-199601	AAC2

Table 1: Differences of symbol formats

Norgate's symbol format is in a symbol-delisted date format. If the company is not delisted, it would have a delisted date. Company symbols can be recycled which means each symbol with the same basis name (different delisted dates), are different companies. Instead of dates, Sharadar symbol names include a number after the symbol. The greater this number, the more historic. The table above shows a symbol having been used three times, but a symbol can only be used one or two times as well. As a disclaimer, AAC-200004 and AAC-199601 existed but not AAC. We handled the above situation through a series of looping and inverting dictionaries, matching the symbols names by creating a final dictionary containing a Norgate key names and Sharadar value names. We started with 7265 assets and ended up with 6729 in the end.

The ending result is depicted in a dictionary below:

```
{
AA': 'AA',
 'AACC-201306': 'AACC',
 'AACE-200610': 'AACE',
 'AACH': 'AACH',
 'AAI-201105': 'AAI',
 'AAIIQ-200603': 'AAIIQ',
 'AAL': 'AAL',
 'AAMC': 'AAMC',
 'AAMRQ-201312': 'AAMRQ',
 'AAN-201012': 'AAN',
 'AAOI': 'AAOI',
 'AAON': 'AAON',
………………….,
}
```

8.3 Constructing our individual factors dataframe

8.3.1 Quality and Value factors

Once we have our universe of assets, we then queried for roic and ebit, and ev using the ART dimension. For null values or missing data, we forward filled the data so that the most recent metric can be used. We would not want to remove the observation entirely and instead we can approximate the value of the metric from its latest value. Another transformation we did is to convert Sharadar's quarterly time series into daily format using the daily date references from our constituent series. So on any given day, there are metrics available.

8.3.2 Momentum,Volatility and Average 200 day Volume

Using our 6729 universe, we use this to create a price series and volume series from Norgate. For each of the series, we form a wide dataframe by merging each individual assets time series together. For our momentum factor we use the 6 month return (skip 1 month) transform the price series into a new returns series. The equation below illustrates the return formula:

$$Return_{6month} = \frac{Price_{t-1} - Price_{t-6}}{Price_{t-6}}$$

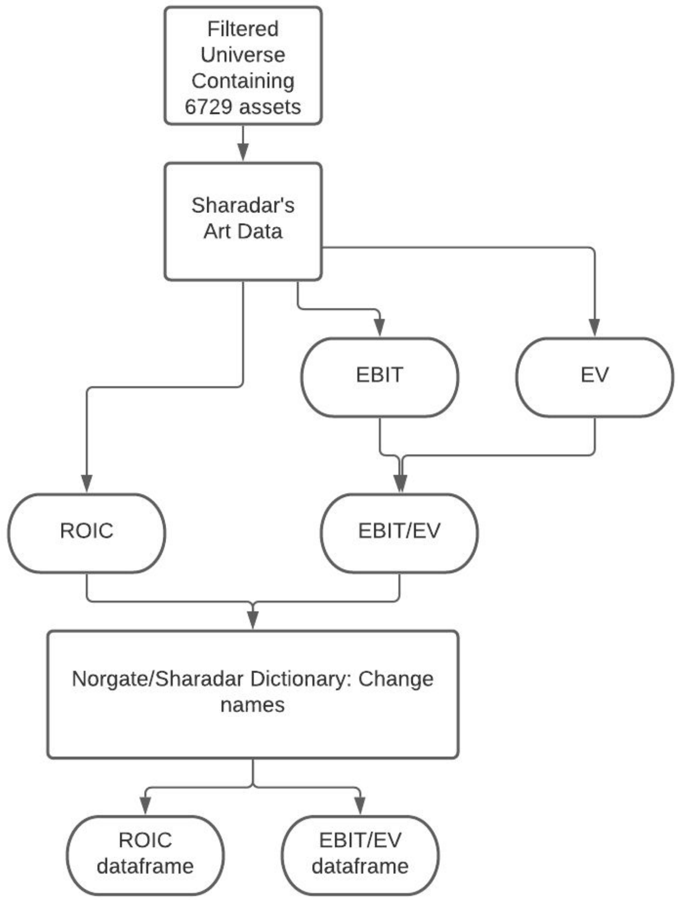

Figure 4: Creating our quality and value factors dataframe

For our volatility, we take the 100 day standard deviation of daily returns, as depicted by the equation below.

$$Volatility = Standard\ Deviation\ (dailyreturns_t, ...dailyreturns_{t-100})$$

And for the average 200 day volume,

$$Average\ 200\ day\ volume = mean\ (volume_t, ..., volume_{t-200})$$

8.4 Constructing our all factors dataframe

We finally take all our individual factors that is, volatility, momentum, ebit/ev, and roic, to form a date-factor mutlti-index wide dataframe consisting of all the factors and assets.

Date		A	AA	AACC-201306	AACE-200610	AACH	AAI-201105	AAIIQ-200603
2010-01-04	mom	0.495436	NaN	-0.207497	0.000000	NaN	-0.197731	0.000000e+00
	qua	0.023000	NaN	0.043000	0.739000	NaN	0.042000	-2.570000e-01
	val	0.010223	NaN	0.074602	0.130589	NaN	0.067231	-1.938734e+00
	vol	0.016846	NaN	0.032986	0.000000	NaN	0.036512	2.699878e-09

Table 2: A snippet of the ‘All Factors Dataframe’

Some of the values might be missing since those companies haven’t been active or present as of the time.

9. Data Preprocessing

From our daily observations of factor values belonging to all the Russell 3000 constituents, we implemented an approach that can form monthly observations so that we can later run K-means on each rebalancing day. We filter out end of month observations from the constituent time series to use as a reference indicating which day we will run K-means on. This also means we can use this to determine which assets’ data we are going to include. We then select only the top 1000 assets rank by their average 200 day volume.

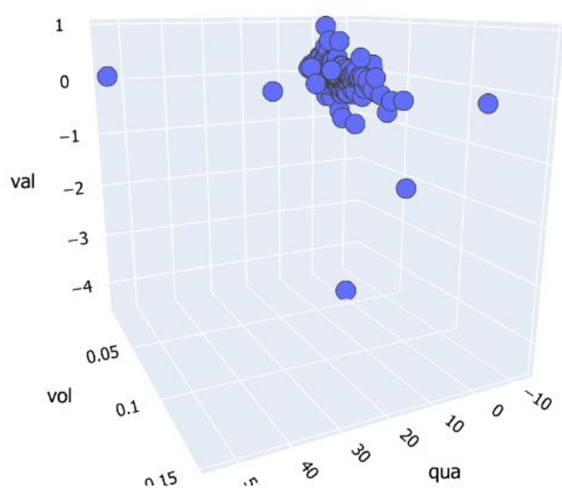

Figure 4: Plotting the stocks’ factors on 2010-05-28

Figure 4 shows our stock factors data after we filter 1000 stocks. We can see that the scaling of the factors is not quite on the same scale. There are stocks that are quite sparse, which means we have a choice of treating them as outliers. We decided to not remove them, since in terms of finance there are scenarios where companies' financial state is in an extreme condition. However, we still have to normalize the data. We transform each of our factors into percentile scores so that all factors are in range from 1-100. This is in sync with formal cross-sectional factor analysis, where we do not analyze stocks by magnitude but by their relative rankings.

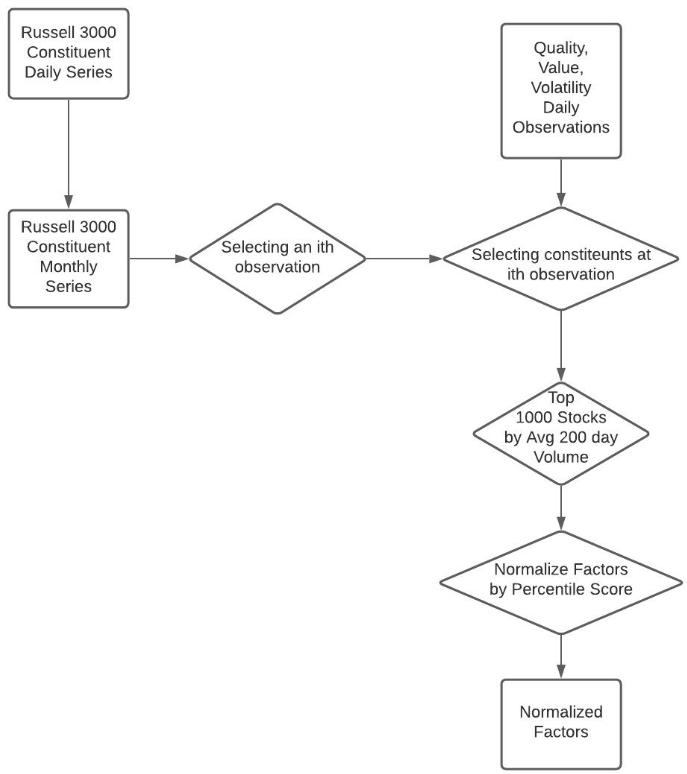

Figure 5: Data Preprocessing Pipeline

10. Clustering Factors vis K-means

10.1 Initial Solution

Our initial solution for running the K-means algorithm is to implement a double loop to find appropriate clusters. Our first loop involves finding the optimal 'K' by evaluating the q score with silhouettes. The nested loop then repeats the first loop multiple times, initializing a random centroid location. We then select the clustering with the highest q. This solves the problem with finding the right 'K' and making our results not prone to randomness. However, this induces some problems. First, we may run into a situation where we end up with different 'K' at different rebalance days, which means when backtesting, the partitions are not appropriate for comparison. Second, the cluster characteristics (by factors) can not be evaluated, because clusters can change locations. For example, when we backtest, if results indicate that cluster 0 produces better results than cluster 1, then how do we determine which factors are the key to cluster 0's success.

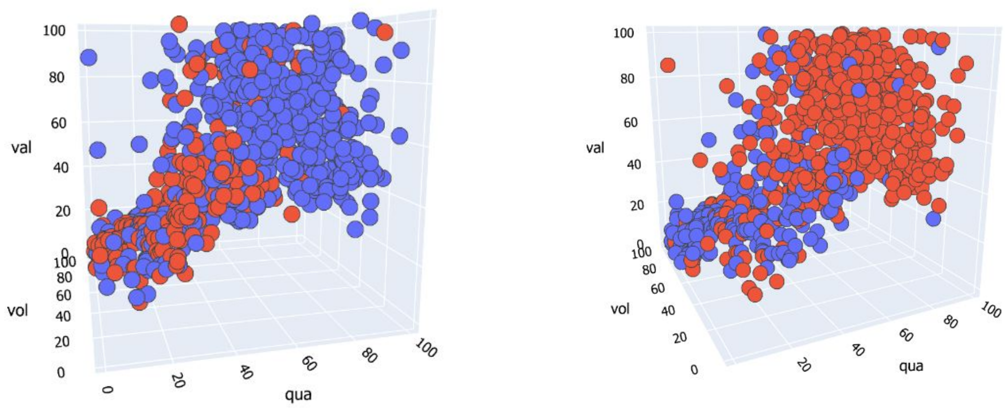

Figure 6: Two Partitions with ambiguous cluster and factor characteristics

10.2 Refinement: Fixing 'K' and Limiting Centroid Initializations

To be able to evaluate partition characteristics and to fix the number of clusters, we proposed that for three factors, we would divide our stocks into 9 clusters. These 9 clusters would define a region of their own that contains a fixed characteristic over time. Having 9 clusters would capture the essence of the factors in the three dimensional space. Characteristics of clusters and are defined below.

Cluster Index	Quality Init.	Volatility Init.	Value Init.
0	50	50	50
1	High: random.(65,85)	Low: random.(15, 35)	High: random.(65, 85)
2	High: random.(65, 85)	High :random.(65, 85)	Low: random.(15, 35)
3	High: random.(65, 85)	Low: random.(15, 35)	Low: random.(15, 35)
4	Low: random.(15, 35)	High: random.(65, 85)	Low: random.(15, 35)
5	Low: random.(15, 35)	High: random.(65, 85)	High: random.(65, 85)
6	Low: random.(15, 35)	Low: random.(15, 35)	High: random.(65, 85)
7	High: random.(65, 85)	High: random.(65, 85)	High: random.(65, 85)
8	Low: random.(15, 35)	Low: random.(15, 35)	Low: random.(15, 35)

Table 3: Cluster Characteristics and Initializations

For each of the clusters, we define a region where it captures the value of the percentiles as either the higher half or the lower half ranges. For example, cluster 1 has high quality, low volatility, and low value and cluster 8 has low quality, low volatility, and low value. This solves our problem of ambiguous cluster characteristics. Then, we initialize the centroids for each cluster to capture the high and low ranges of the factors. Since the higher half of the values are between 50-100, we randomize the values to between 65 to 85. First, this centers the centroid in each region, and second we allow for some randomness and evaluate the q score for each initialization.

We then run our refined K-means on all the monthly observations and collect partition data to be used in our backtest.

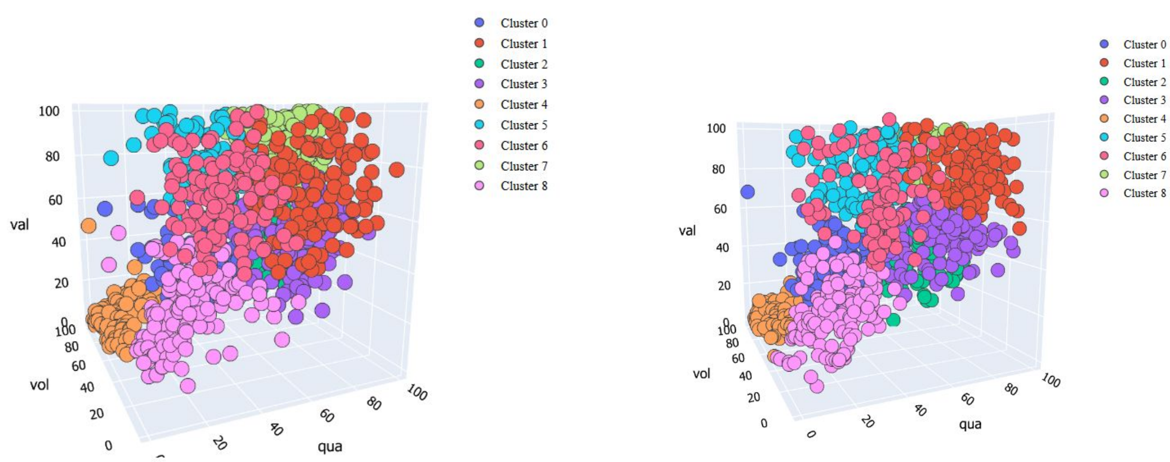

FIgure 7: Refined Clusters

Figure 7 shows our 9 clusters in three dimensional space and the regions are stable.

11. Backtesting

We now evaluate the risk premium of our clusters by examining the Sharpe Ratio, CAGR, Annual Volatility, Maximum Drawdown, and Cumulative Returns. Our simulation time window starts from 2003-01-01 till 2020-09-29. Every backtest starts with an equity of $10000.

11.1 Original Clusters (Quality, Volatility, Value)

In this simulation, we evaluate performance purely on the clusters. Each stock is weighted equally. Cluster 3 achieves the highest Sharpe Ratio with lowest maximum drawdown and a cumulative return of 840%.

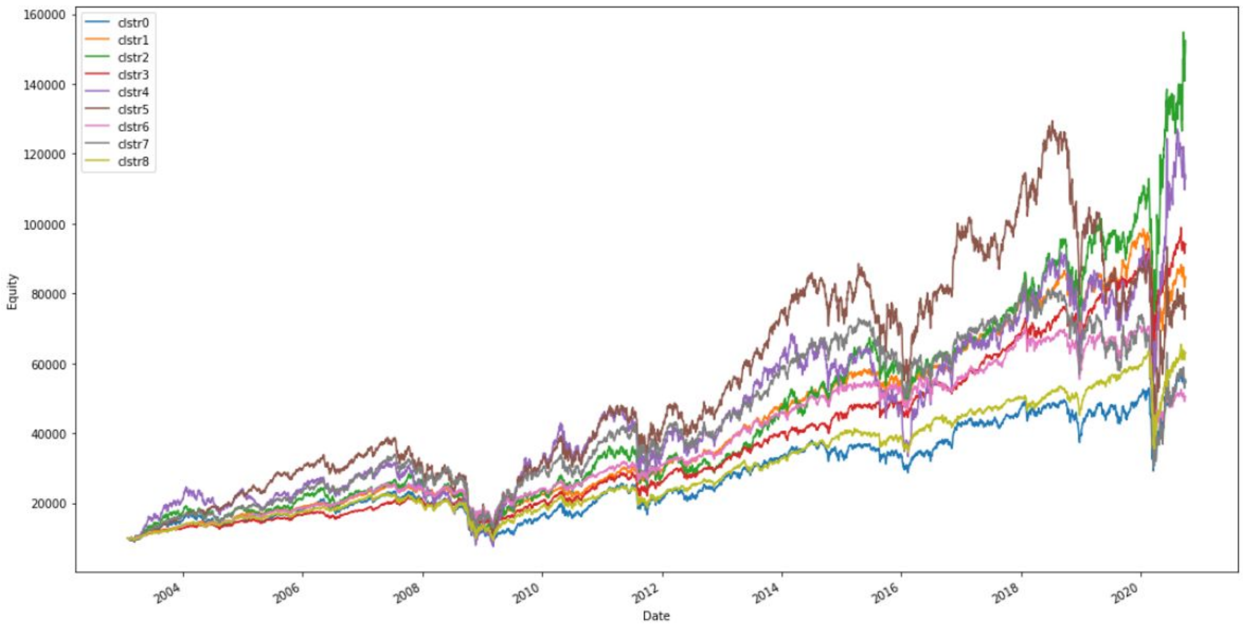

Figure 8: Quality, Volatility, and Value Simulation

	CAGR	Annual Vol	MaxDD	Cum Returns	Sharp Ratio
Cluster 0	0.100678	0.266747	-0.633272	4.442542	0.494242
Cluster 1	0.128240	0.184905	-0.502153	7.423829	0.748317
Cluster 2	0.166742	0.284003	-0.648588	14.237101	0.686077
Cluster 3	0.135287	0.177643	-0.418390	8.403094	0.805304
Cluster 4	0.146993	0.350232	-0.761316	10.270884	0.568020
Cluster 5	0.121644	0.323109	-0.728329	6.595062	0.518460
Cluster 6	0.095895	0.200962	-0.504792	4.039591	0.559585
Cluster 7	0.101014	0.283939	-0.665238	4.471927	0.482604
Cluster 8	0.109790	0.220093	-0.593879	5.295606	0.586959

Table 4: Quality,Volatility, and Value Performance

11.2 Original Clusters + Top 20 Cross Sectional Momentum

On top of each cluster, we now add relative momentum to the three clusters. We pick the top 20 stocks rank by their 6-month return.

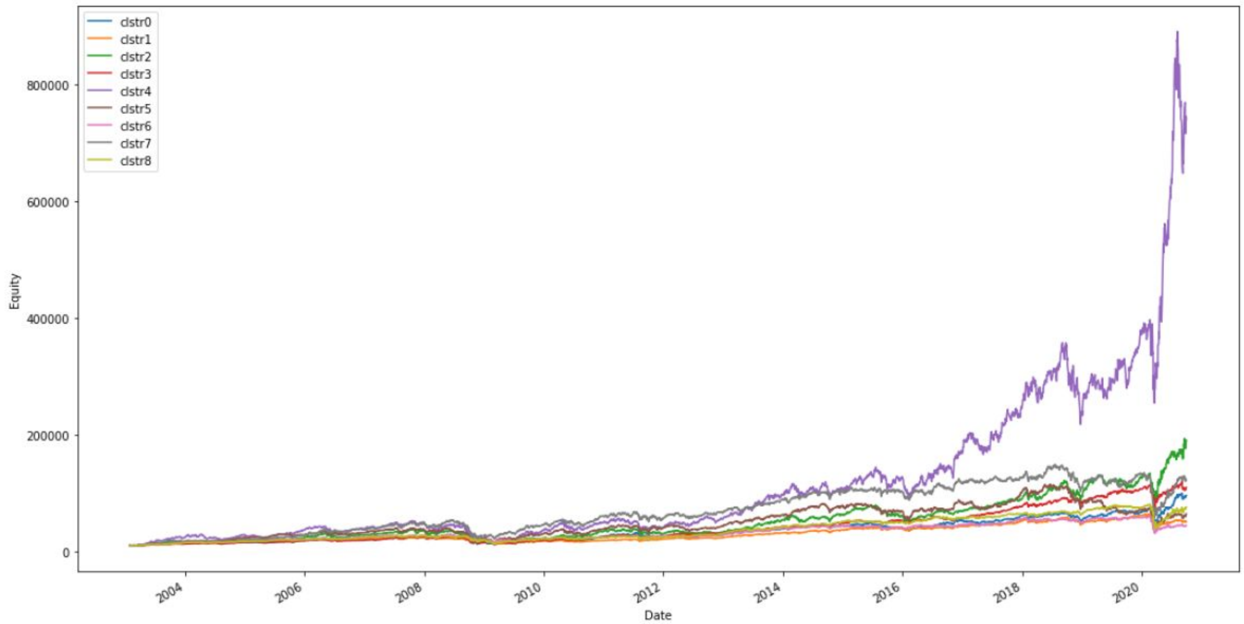

Figure 9: Quality, Volatility, and Value, Momentum Simulation

	CAGR	Annual Vol	MaxDD	Cum Returns	Sharp Ratio
Cluster 0	0.136192	0.273359	-0.602716	8.536322	0.605281
Cluster 1	0.096374	0.190065	-0.515972	4.078653	0.582565
Cluster 2	0.181449	0.303593	-0.692511	18.010015	0.702359
Cluster 3	0.144949	0.191232	-0.504599	9.921260	0.805925
Cluster 4	0.275876	0.362393	-0.780573	72.922526	0.856626
Cluster 5	0.110868	0.309367	-0.720749	5.404521	0.496327
Cluster 6	0.087427	0.201959	-0.489577	3.394367	0.518782
Cluster 7	0.152242	0.283597	-0.620344	11.217310	0.645401
Cluster 8	0.121212	0.232816	-0.554475	6.543632	0.613720

Table 5: Quality, Volatility, and Value, Momentum Performance

By adding momentum to our strategy, the Sharpe Ratio of many clusters have increased significantly.

11.3 Original Clusters + Top 20 Cross Sectional Momentum + Trend Filter

On top of the four factors, we now add a time series momentum switch using a 200 day moving average of Russell 1000. If the index is above its long term trend, then we trade and if not then we hold cash. We might consider moving into treasury bonds, but I will omit the inclusion for now. We will now demonstrate the potential of time series momentum.

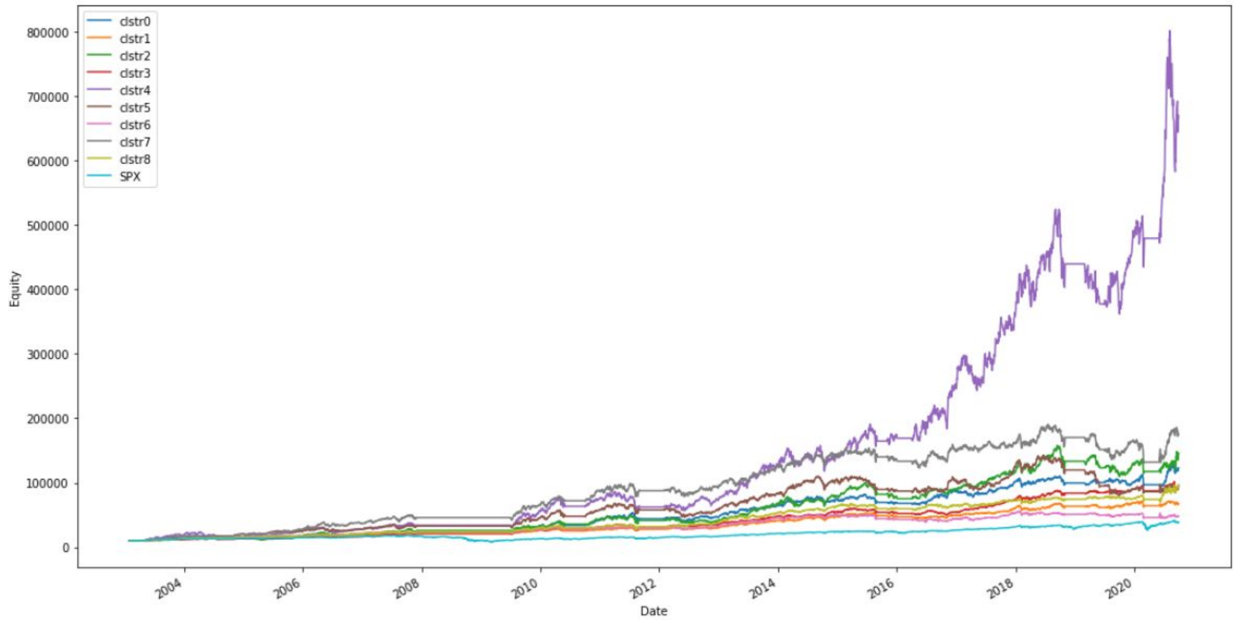

Figure 10: Four Factors + Trend Filter Simulation

	CAGR	Annual Vol	MaxDD	Cum Returns	Sharp Ratio
Cluster 0	0.152637	0.193438	-0.275911	11.291498	0.833067
Cluster 1	0.114371	0.129280	-0.176748	5.770711	0.904417
Cluster 2	0.163904	0.218937	-0.306829	13.595457	0.803834
Cluster 3	0.134831	0.140927	-0.238143	8.336560	0.970881
Cluster 4	0.268292	0.267070	-0.366910	65.534705	1.024468
Cluster 5	0.136929	0.201035	-0.438899	8.646145	0.741055
Cluster 6	0.093436	0.135489	-0.252059	3.843528	0.729944
Cluster 7	0.176142	0.196248	-0.318905	16.557021	0.926932
Cluster 8	0.136680	0.146754	-0.218047	8.608861	0.948485
SPX	0.080072	0.193017	-0.567754	2.897941	0.496847

Table 6: Four Factors + Trend Filter Performance

By adding our trend filter, the performance of our clusters have been boosted. We can see abnormal returns, an increase in Sharpe Ratio and especially a decrease in drawdowns. Cluster 4 has the highest Sharpe Ratio, but also a large drawdown and high volatility. This is due to the cluster having low quality, high volatility, and low value. Buying high volatility stocks means volatile performance and we can also deduce that low quality and value does not hold up well when markets are performing poorly. It is also worth mentioning that cluster 4 is the contrarian strategy to formal factor investing, where the latter buys high value, low volatility, and high quality. Cluster 3 and 8 performs quite well, having a drawdown of slightly over 20%, great cumulative returns and CAGR and acceptable volatility. Interestingly, cluster 8 with low volatility, low quality, and low value performs well. Cluster 1, which resembles formal factor investing, still holds up to have the highest volatility and lowest drawdown. Investing in this cluster should be most tolerable in terms of risk profile and stability.

12. Benchmark Comparison

Let's compare the performance with the SP500. As depicted, it is a no brainer that investing in SP500 would not beat any of the clusters. SP500 finished with a 2.89 Sharpe Ratio, 56% drawdown, and around 8% CAGR. It is safe to say that there are ways to invest smarter with a couple of simple rules, instead of just buying and holding on to the popular stock index.

13. Conclusion

In this paper, we have explored the impact of factors that drive unexplained returns. These factors are quality, volatility, value, cross-sectional momentum and time-series momentum. We have explored the background academic literature around traditional ways of utilizing these factors in factor investing, where most of the investment execution is done using idiosyncratic-factor portfolios.

We construct an investment model to analyze multiple drivers under the quantamentals framework by implementing a K-means algorithm that clusters assets based on quality, volatility, and value. We refined the formal K-means algorithm by explicitly assigning 9 clusters and set centroid initializations such that each cluster can be classified by factor characteristics. We then also incorporate cross-sectional momentum and time-series momentum in our investment decisions. All together, we capture the essence of quantamentals by combining fundamental and technical factors in our decision making. We evaluated how each cluster would perform and classified which combination of factor magnitudes would drive which kind of risk premium. Under a constructive rule based strategy, we demonstrated that cluster 1 (formal factor investing) is still a viable strategy today. Interestingly, cluster 3 and cluster 8 characteristics have shed light to new discoveries. All in all, our clusters have beaten the SPX benchmark by a large feat and we conclude that our strategy should work better than a purely passive investment.

For further studies, we suggest varying parameters for cross-sectional momentum and time-series momentum. For instance, we can try 3 month returns for cross-sectional and 6 month returns for our trend filter. We can consider running K-means on all the four factors instead of three and evaluate clusters directly.