**Data Analytics Capstone Final Report:**

Active Management: Identifying Opportunities for Use

Curtis Schrack, Conor Keating, Jordan Sabol & Pat O’Reilly

5/9/2023

**1. Purpose**

The main goal of this project is to identify opportunities for active management to perform passive management. Active management is when an individual decides whether to buy, sell, or hold a fund, while passive management focuses on leaving a fund to grow over time to get the 10% market return. The group used metrics such as cross-sectional standard deviation and pairwise correlations of daily stock returns to determine when a fund might be in a state of high volatility. These high volatility points mark areas active management can provide the ability to recognize high profit or loss. We plan to test the hypothesis that moments where there is a low pairwise correlation (when stocks in the fund do not relate to each other) and a high cross-sectional standard deviation (when value is above an acceptable change in gain or loss). From this, we hope to determine day-by-day areas of opportunity.

**2. Background**

The group's primary goal was to build a program to help recognize active management opportunity periods. Pat O'Reilly, of the Federated Hermes Business Support Team, described the expectations of the project as a program that can pull financial data, perform pairwise correlation and standard deviation calculations, and display these in an easy-to-read graph. After the first few meetings with Pat and Dr. Petrovich, adding a user interface (UI) was the best way to present the project. This addition led to the project being built using Python. Python was chosen based on a large amount of support for libraries that a programmer can use to implement the calculation alongside a programmer-friendly UI environment.

We started by first developing a way to perform the calculations. Pat provided the group with a CSV file in which we pulled the data from the file into our programming environment. The group worked on a few ways to get the standard deviations and correlations based on a day-by-day returns of tickers. It was decided to put the information in a data frame and use a loop to get the information one day at a time. Initially, the group found that the calculation process took too long. After some experimentation, the group took a fifteen-minute process down to under two minutes by removing two loops.

The next step was adding graphing. The graphing took about a week to build the initial design. The graph was an ever-changing part of the project, with it being manipulated on a weekly basis to fit the needs of new additions. The first design contained the correlations, standard deviations over time, and the medians for both metrics. Once the areas of opportunity were determined later in the project, we added these areas by displaying them with shaded color areas between the two median lines. A few minor visual tweaks were made to the graph, such as line color changes to improve the viewing experience.

The third step was creating an easy-to-build system to add funds to the program. This process was where the group reached their biggest hurdle. The first attempt was to pull data directly from an investor insight product called Morningstar. Morningstar was a program that Federated Hermes used for evaluating funds. Morningstar had recently added a Python Jupyter Lab environment called Analytics Lab. The group initially built the project in Analytics Lab but found multiple issues. The first is due to Analytics Lab being less than a year old, there was limited documentation on how to use the system. Once the group found out how to pull data correctly, there was another hurdle where a limited amount of data was allowed to be pulled in one day. This limitation led to a one-month delay, and the group experimented with other options until the limit was eventually lifted. During that time, however, the group created a separate project using yfinance, which pulls stock information from Yahoo Finance.

The next step was creating the areas of opportunity. This process was done by writing a loop that would evaluate each day and see if the correlation was lower than the correlation median and if the standard deviation was above the standard deviation median. If both criteria were met, the program would flag that day as an area of opportunity. The longest part of this process was getting the areas of opportunity to show up nicely on the graph. The process took about one week to implement on the Morningstar and Yahoo Finance programs.

The last step was the implementation of AI. The Yahoo! Finance program used a Tkinter extension called CustomTkinter. This library had a lot of documentation, making implementation relatively manageable though time-consuming. This implementation continued until the project's end, with about half the program being developed incorporating a UI environment. The analytics lab UI was much more difficult. Analytics lab restricts a large number of libraries making UI development difficult. The group found that the combination of the restrictions and the small amount of documentation led to no viable UI being implemented into Analytics Lab.

**3. Results**

In the early stages of our template development, the goal was to ensure that Pat's calculations in his shared Excel file could be accomplished using Python code. Before a preliminary template was built, Cross-Sectional Standard Deviation and Pairwise Correlation calculations were tested in Python. Data were copied from Pat’s original Excel file (Pairwise+Correlations+v7 (version 1).xlsx) into our .csv file to be calculated. Pat used a specific timeframe (Day to Day Return 2021-11-01 through 2021-12-13) for his CSSD & Correlation Matrix.

**3.1 Cross Sectional Standard Deviation Test**

Using this extracted data, we first calculated the CSSD of all Tickers for the 31-day window (excludes weekends):

Here is a snapshot of the CSSD of 5 days, along with summary statistics of our findings:

|  |
| --- |
| **count 30** |
| **mean 2.51** |
| **min 1.87** |
| **25% 2.17** |
| **50% 2.45** |
| **75% 2.73** |
| **max 3.61** |

A screenshot of a computer

Description automatically generated with low confidence

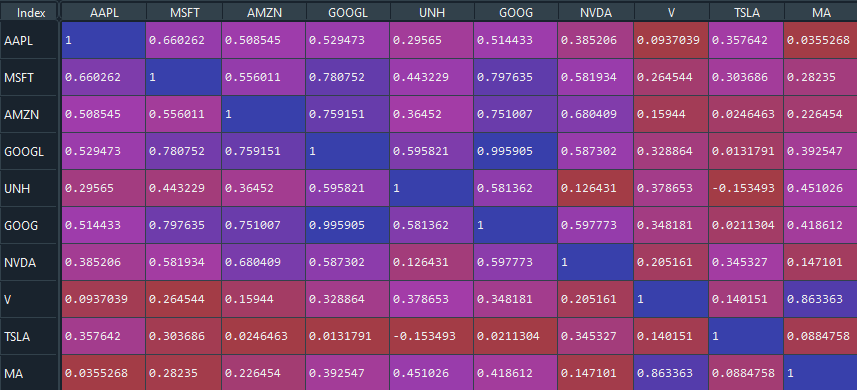
Chart, line chart

Description automatically generated**Over the course of 31 days, the average CSSD between stock returns was 2.51. The minimum CSSD was 1.87 and the max was 3.61.**

Additionally, a basic line graph of CSSD over time was plotted to ensure it can be used in the final template graph of CSSD & Correlation.

**3.2 Pairwise Correlation Test**

The second calculation to identify opportunities for active management is Pairwise Correlation. Again, before a full template was made, correlation calculations in Python were completed and matched to the results in Pat’s Excel file. At a preliminary level, the correlation calculations were done on the first 10 Tickers in our dataset to ensure the Python code worked correctly.



The results of our correlation calculation tests can be seen in the matrix above. The correlation values fall between 0 and 1, with 0 being no correlation between daily stock returns and 1 being a perfect correlation.

**The CSSD and the Pairwise Correlation Tests were then scaled out to be calculated using all tickers over a 264-day period. Preliminary results were successfully obtained using data extracted from Pat’s Excel file under the “Data” tab in our CSV (Russell 1000 Growth Ticker Returns 1/3/22-1/11/2023). A CSSD line graph over this period was plotted along with a horizontal median Standard Deviation of the CSSD results (a mock of Pat’s graph without correlation data overlayed).**

A picture containing text, antenna

Description automatically generated

**3.3 Building a Template**

Using the same data extracted from the “Data” tab of Pat’s file into our CSV file, a full template using a “for loop” was built to calculate both the correlation medians & CSSD of the entire data frame. The completed template can input the number of days (rolling window) within a data set to look at and outputs the two metrics needed to create a line graph over time. In our case, Pat wanted us to look at 126 days (6-month working days), so this was the integer for our rolling period value.

In the graph below, both the Cross-Sectional Standard Deviations & Pairwise Correlations of the Russell 1000 Growth are plotted with the median values of both metrics as horizontal lines. **The graph is identical to the graph produced by Pat in his Excel file, meaning our preliminary results, produced by an efficient Python code template, concur with Pat’s calculations. Using this graph, opportunities for Active Management could be identified, with the “areas of opportunity” being where both metrics fall between the median lines.**

Chart, line chart, histogram

Description automatically generated

**3.5 Yahoo! Program**

During the waiting period to hear from Morningstar about extending our data limit, the group worked on the project using yfinance. To collect the data, additional code was written to add tickers in a fund. Then yfinance would provide the ability to pull data based on a date range. After the ticker data was pulled, the closing date from the previous day would be used to determine the percentage of change between days to get the daily return.

The code for calculating a correlation and standard deviation was able to be added from the previous code developed at the beginning of the semester. Then add the graphing creation code, which combines effortlessly with the already created code. The one addition added was the graph would now save to the file location of the Python script. This saved image of the graph was for easier viewing.

Chart, histogram

Description automatically generatedText

Description automatically generatedWhile the data limit for Morningstar was still being worked on, the Yahoo group began the implementation of the areas of opportunity. This implementation was done by determining intervals where the correlation was below the median correlation median and standard deviation median. The shaded area was added to the graph to show the time intervals where these active management areas were. **The date ranges where the S&P 500 was in an area where active management should be considered. Graph showing the areas of active management corresponding ton the table on the left in light green shaded areas.**

Graphical user interface, text, application, chat or text message

Description automatically generatedThe final addition to the Yahoo! program was the implementation of UI. The original goal was to have the whole program have a UI. However, due to time constraints, that became limited. The UI that got implemented was a popup in the top right corner of the user screen asking for the fund, date range, number of rolling days, and smoothing strength.

**Example of UI popup and what user needs to input to create an Active Management graph.**

**3.4 Morningstar Analytics Lab Program**

After receiving permission from a Morningstar representative, our data query limit was extended temporarily to 5 million data points daily. With this extension, we could move forward in the Morningstar Analytics Lab environment and build a near-identical program to the Yahoo! program.

The code for calculations of cross-sectional standard deviation & pairwise correlations and the time-series line graphing remained the same within Analytics Lab. However, the way the data were retrieved was different. Morningstar Analytics Lab uses a unique identifier, “investment\_id”, for each securities in its database. Unlike the Yahoo! program, which pulls individual ticker’s open and closing prices for each day, we can pull daily return data directly for any fund/index within the Morningstar database.

A user can change the “investment\_id” for any fund/index they would like to produce a graph by hard coding it in the program or copying and pasting it from Morningstar directly. The program will automatically pull in four quarters of holding dates + the holdings in a selected index. Data's start and end date can also be changed to show as much historical data as Morningstar will allow. Like the previous program, the rolling window can be changed to any user's desired integer (currently set at 126). Graphs can be customized at the end of the program to user preferences. Comments can be found within the notebook for easy execution of code by any user.

Graphical user interface, chart

Description automatically generated

**Here is an example of the final graph produced by the Morningstar Analytics Lab notebook. The time-series graph displays the active management opportunities for the S&P 500 Growth TR USD from the past 10 years.**

**3.6.1 Validation: Active Management vs. Excess Return Graphs**

With the ability to create graphs for any fund or index found in both the Yahoo! and Morningstar databases, a validation of our main hypothesis that high standard deviations and low correlations are good predictors of opportunities for active management was necessary. The main validation technique was to align our programs’ graphs for sector-specific active managed funds on top of the Excess Return graphs for these same securities. Pat provided us with seven different Excess Return graphs in different sectors, including Consumer Cyclical, Consumer Defensive, Energy, Financials, Health, Real Estate, and Utilities.

We then used our programs to produce the necessary Active Management Opportunity Identifier graphs for all 7 Sectors, matching the exact dates as the Excess Return Graphs. We created a shared document with Pat of our produced graphs aligned with the Excess Return graphs.

**Chart

Description automatically generatedFindings: 1.) Real EstateChart

Description automatically generated**

This image is a good example of where our green areas of opportunities resulted in a positive excess return. In other words, the US Real Estate Active Fund outperformed its Morningstar Real Estate TR benchmark during our predicted area of opportunity, where the correlation falls below the median, and CSSD is above (orange box).

**Chart

Description automatically generated**

**Findings: 2.) Financial**

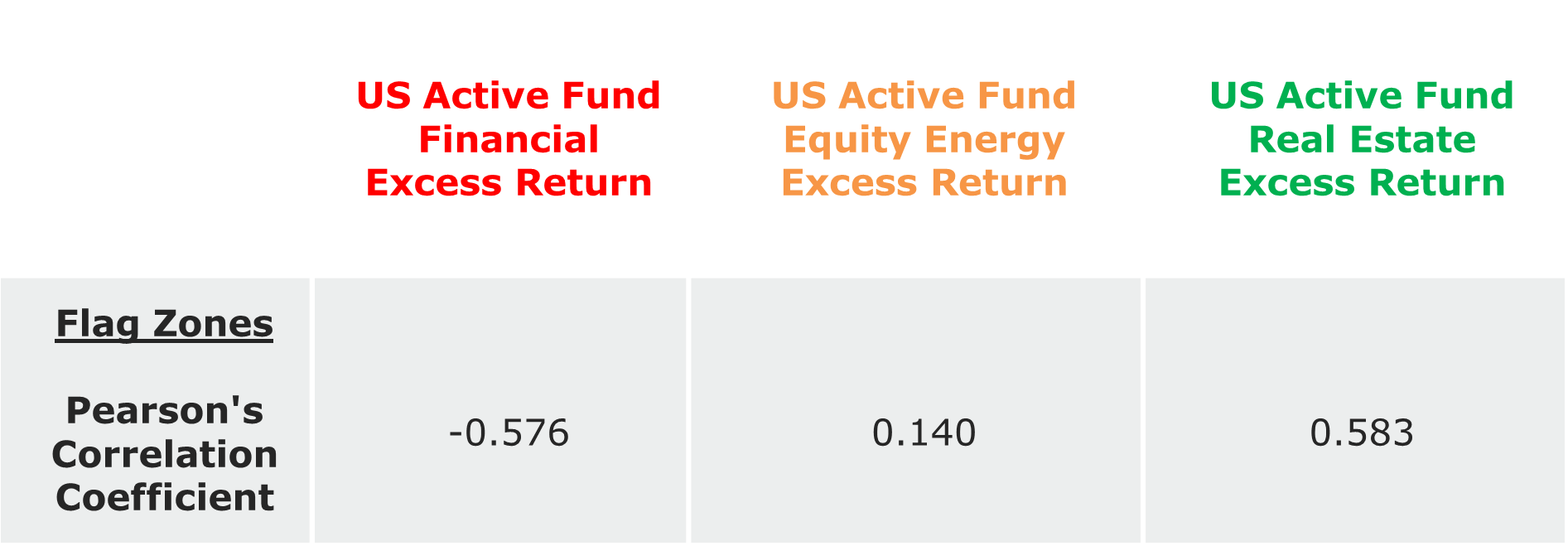
**Chart

Description automatically generated**

This is a poor example of where our produced graph for the US Financial Active Managed Fund does not translate to the correct excess return. In this case, our program did not trigger an area of opportunity due to the high correlation (above the median), and the standard deviation is low/just above the median. However, this period had a very high excess return (black box).

**3.6.2 Validation: Correlation Between Opportunity Zones & Excess Return**

Taking the validation a step further, the second technique used to attempt to validate our hypothesis was to find the correlation between the opportunity zones & excess returns. If we could put a numeric value, i.e., a correlation coefficient, specifically Pearson’s, we could evaluate the strength of the relationship between these two variables.

****

Using the same correlation Python function used in the program for daily returns, we could also produce a correlation coefficient to see the relationship between our “flag zones” and excess returns for the Financial, Energy, and Real Estate sectors. Daily Excess Return data for the Financial, Energy, and Real Estate sectors was provided to us by Pat.

**We can conclude that our flag zones do a poor job of predicting positive excess returns in the financial sectors, as the correlation is negative (-0.576). Our zones do a better job in the Energy sector, with a 0.140 correlation. The Real Estate sector is our best sector in predicting excess returns with our flag zones. These Pearson coefficients match our findings in the validation graphs above.**

**4. Implications**

Although all graphs for the indexes requested to be produced, e.g., S&P 500, Russell 1000, Russell 1000 Growth, Russell 1000 Value, etc., are still left up for interpretation by Federated Hermes themselves, we have built a solid foundation for the company. The programs created for this Capstone project will be handed off to Federated for their use and can help them draw more conclusions in the future.

From the graphs we were able to produce, specifically in the seven sectors we were asked to look at and compare to excess return data, we can draw some early conclusions about our hypothesis. From some preliminary graph interpretations and strength of relationships calculations, active management performs differently in different sectors. We cannot adequately confirm or deny our original hypothesis that low pairwise correlation and high cross-sectional standard deviation are opportunities for active management or that active management will always produce a positive excess return under these conditions.