**Project info**

Project title: Expected Returns Factor Creation and Feature Engineering

Project short title (30 characters): Factor Analysis

URL of project idea page: <https://github.com/rstats-gsoc/gsoc2024/wiki/ExpectedReturns-Factor-Creation-and-Feature-Engineering>

**Bio of Contributor**  
My name is Yi. I am currently a junior at St. Lawrence University, majoring in Mathematics. My expertise is primarily in pricing and risk management. At school, I've enrolled in Computer Science courses like Data Structure, Economics courses like Econometrics, and Statistics courses like Applied Regression. The project "Expected Returns in R" is more than just a coding project; it demands interdisciplinary knowledge to grasp the economic and financial implications of factor analysis. My work in finance and data science has provided me with both theoretical and programming skills, ensuring the successful completion of the project.

Born and raised in China, I moved to a small country called Eswatini in Southern Africa for high school. There, I spent two years studying with peers from over 40 nationalities and dedicating myself to a community project where I fundraised study materials for a refugee camp. This experience has made me collaborative, well-adapted to various situations and eager to contribute to non-profit open-source projects. Hence, I am confident in my ability to navigate through difficulties and adapt to necessary changes for the completion of the project.

**Technical Knowledge**

I am a junior majoring in Mathematics. The courses that I have done that can potential help me in this project:

* Data Structure
* Technique of Computer Science
* Applied Regression Analysis
* Probability
* Econometrics
* Investments

I have done projects during internships and research that require me to program and perform statistical analysis. For example:

* Developed econometric models in R to analyze stock price responses to 100 S&P 600 small-cap companies' financial announcements, assessing market efficiency and optimal investment timings based on news
* Aggregated monthly data of bank’s balance sheets into a database to monitor and visualize 5-year trends of market value and key risk metrics using Python
* Researched how the tier, cost, and location of undergraduate education impact the future earnings of low-income students in STATA

There is the link to my resume: [GSoC/YEYUXI YI Resume May 2025.pdf at main · YYINUS/GSoC (github.com)](https://github.com/YYINUS/GSoC/blob/main/YEYUXI%20YI%20Resume%20May%202025.pdf)

**Contact Information**

Contributor name: YEYUXI YI

Contributor postal address: 23 Romoda Drive, Canton, New York, 13617

Telephone(s): +1 3156033649

Email(s): yyi21@stlawu.edu / yeyuyi@waterforduwc.net

Other communications channels: WhatsApp (same as phone number)

**Contributor affiliation**

Institution: St. Lawrence University

Program: Bachelor of Science

Stage of completion: Junior

Contact to verify: edharcourt@stlawu.edu

Schedule Conflicts: There is no conflicting schedule. Contributor at GSoC will be a full-time role for me.

**Mentors**

Evaluating mentor name and email:

* Justin M. Shea, Executive Director Finance Honors track & Assistant Professor
* [justin.shea@gmail.com](mailto:justin.shea@gmail.com)

Co-mentor name(s) and email(s):

* Look for a co-mentor

I reached out to Professor Justin and LinkedIn on March 25th. Professor Justin kindly replied to me and invited me to an R meeting on March 27th. I had a brief chat with Professor Justin over the questions I had for the R project and proposal writing. I have sent my first draft to Professor Justin for revision on March 30th .

**Coding Plan and Methods**  
To successfully replicate and create factor constructor functions for all the research papers listed on the idea page, I will adhere to the following procedure:

Step 1: Make an annotated paper and codes.

Understanding the methodology and approaches used in the research paper is critically important. Hence, writing an annotated paper and summarizing the methodologies and findings in my own words are essential to fully comprehend the paper and ensure the successful construction of factors. There will be an example of an annotated research paper from the idea page (I will add the link here over this weekend). I will also write annotations for each function, object, and variable in the package to fully understand how those codes lead to the results. An example of annotated codes from the R Package Expected Returns will be added (I will add the link here over this weekend).

Step 2: Communicate with Mentors

After completing the annotations and understanding the materials myself, I shall communicate with my mentors to ensure that I have correctly understood the concepts and procedures by showing the annotated work I have done.

Step 3: Functions Identification

I will list any functions I can create based on my understanding of the papers and code documentation. This can include create functions to import real life data from financial database, perform linear/log transformations, clean anomalies/outliners, filter empty/undesired data, calculate ratios, perform regression analysis, and compute advance functions like Fama French Three Factor Model for the portfolio.

Step 4: Functions Creation and Testing

I will start creating functions based on the desired output. During the creation stage, I will perform regular tests, such as those for extreme values, missing or zero values, non-normally distributed values, and other tests that might be specific to a function, to ensure that the function will work under stress.

Step 5: Documentation

I will create documentation for the functions created above. The documentation will conclude with references to the end of a research paper and corresponding R codes.

**Potential Obstacles During the stage 1 to 5:**  
Step 1: Annotate Papers and Code

During this stage, I may encounter situations where I do not understand a specific econometric or finance concept/process. In such cases, I will seek support through Google, ChatGPT, or my mentors.

Step 2: Communicate with Mentors

My mentors are willing to generously spend their time to guide and teach me. However, sometimes they may be busy with their businesses or families. If I cannot reach my mentor, I will leave them a note and proceed with the plan.

Step 3: Function Identification

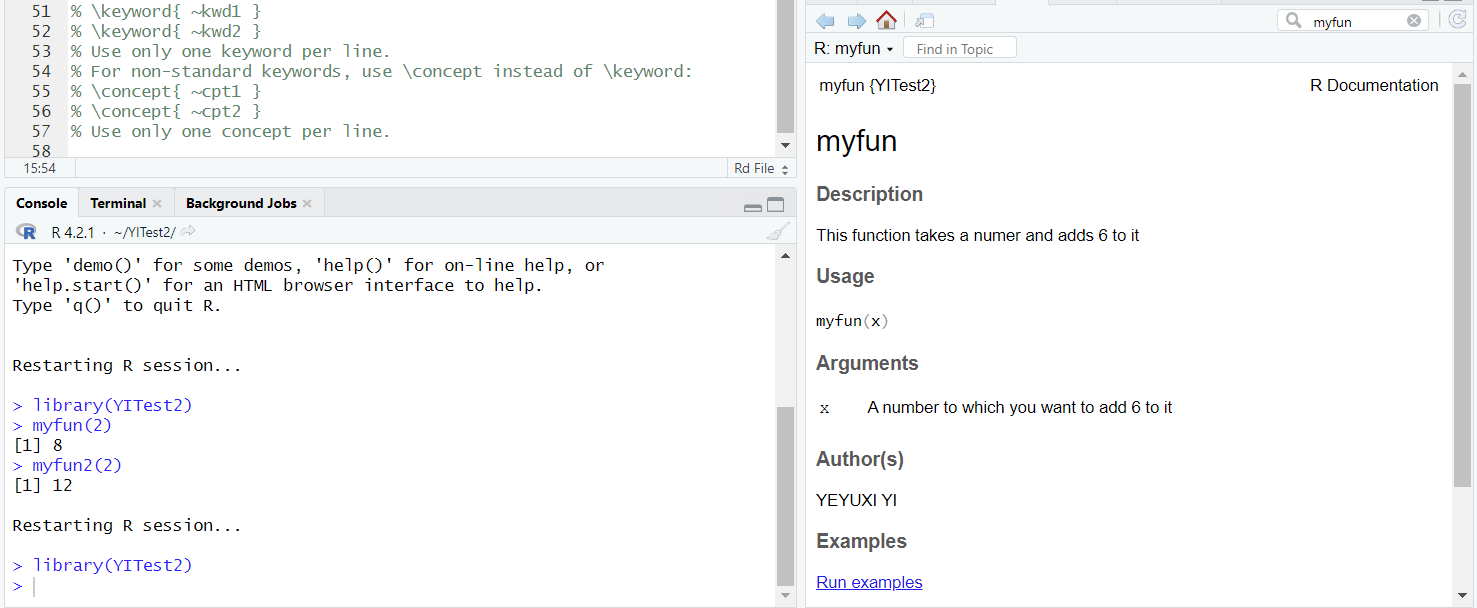
The success of this stage largely depends on how well I understand and annotate the materials in Stage 1. The thoroughness of Stage 1 will ensure ease in identifying necessary functions. If there are additional functions that could be beneficial, I will discuss including them with my mentors.

Step 4: Function Creation and Testing

When creating functions to handle raw data from financial databases, we might encounter missing or zero values that make log transformations and division computations fail. We can either drop these values or estimate them. For values that are not normally distributed, we might need to apply transformations or identify outliers. For merging datasets lacking unique identifiers, we might need to adjust the data or create new columns. This stage requires mastery of models and functions. Therefore, for the first two projects, I plan to spend a longer time on Stage 4, from one week to two weeks. With sufficient experience and guidance from mentors, I will become more proficient in creating and testing functions.

Step 5: Documentation

I have never created a package before, but I have done some self-learning to get a taste of it. The screenshot is a simple package created in R for testing purposes before I reached out to my mentors. I believe that once I document the first two projects, I will become proficient in creating documentation.

****

**Timeline**

I plan to work 40hours per week during the official coding period. However, if time is needed to complete the milestone on time, I am willing to work overtime to ensure the completion. The completion of step 5 should be a milestone on the timeline.

Weeks between May 1 - 26

* Bonding with my mentors
* Reading the book [Expected Returns: An Investors Guide to Harvesting Market Rewards](https://www.wiley.com/en-us/Expected+Returns%3A+An+Investor%27s+Guide+to+Harvesting+Market+Rewards-p-9781119990727) by [Antti Ilmanen](https://www.aqr.com/About-Us/OurFirm/Antti-Ilmanen)
* Learning how to make a detailed documentation.
* Starting to make annotated papers and codes

Week May 27

* Perform step 1 to 4 for Ari Levine, Yao Hua Ooi, Matthew P. Richardson, Caroline Sasseville (2016). Commodities for the Long Run

Week June 3

* Perform step 4 to 5

Week June 10

* Perform step 1 to 4 for Hou, Kewei and Mo, Haitao and Xue, Chen and Zhang, Lu (2016). Which Factors?

Week June 17

* Perform step 4 to 5

Week June 24

* Perform step 1 to 5 for **Value-oriented equity selection**, chapter 12 of the book [Expected Returns: An Investors Guide to Harvesting Market Rewards](https://www.wiley.com/en-us/Expected+Returns%3A+An+Investor%27s+Guide+to+Harvesting+Market+Rewards-p-9781119990727) by [Antti Ilmanen](https://www.aqr.com/About-Us/OurFirm/Antti-Ilmanen)
* Evaluate whether after two projects, I am proficient enough to go through all steps of development in a week and make the change of plan accordingly

Week July 1

* Perform step 1 to 5 for Asness, Clifford and Frazzini, Andrea (2012). The devil in HML's details

Week July 8

* Perform step 1 to 5 for Asness, Clifford S. and Moskowitz, Tobias J. and Pedersen, Lasse Heje (2013). Value and momentum everywhere
* Prepare midterm evaluation

Week July 15

* Perform step 1 to 5 for **Commodity Momentum and trend following**, Chapter 14 of the book.
* Submit midterm evaluation.

Week July 22

* Perform step 1 to 5 for Moskowitz, Tobias J and Ooi, Yao Hua and Pedersen, Lasse Heje (2012). Time Series Momentum

Week July 29

* Perform step 1 to 5 for Balts, Kosowski (2012). Demystifying Time-Series Momentum Strategies: Volatility Estimators, Trading Rules and Pairwise Correlations

Week Aug 5

* Perform step 1 to 5 for Balts, Kosowski (2013). Momentum Strategies in Futures Marketsand Trend-Following Funds

Week Aug 12

* Perform step 1 to 5 for a paper that both mentor and I find interesting to replicate and create functions for

Week Aug 19

* Examine documentation and run tests for the package. Fix all the bugs identified during the step

Week Aug 26

* An extra week for any change during the plan or mentors and I think there is something worth to improve and work on

**Management of Coding project**

Codes should be at least committed twice in a week. In any week that codes are not committed twice indicates a problem. For all the codes that are tested, I shall write a document to record what I test and how I test.

**Test**

Tests for this project are listed below.

1. Pre-req: Please show an R project you've completed.
2. Easy:  Downloading and building the Expected Returns and Factor Analytics packages locally. List any build errors or issues you encounter on install and see if you can work through those and get the package to build.
3. Intermediate: Check the files in the vignettes directory and find one that doesn't build and identify bugs.
4. Harder: Reflect on the steps above. How do you interpret the statistical estimates of the vignettes that are working for you? In addition, was there any repetitious code in the vignette that may be written as a function for future use? If so please include it as an example.

Below are two links for my response to all test problems.

Response to 1/3/4: <https://github.com/YYINUS/GSoC/blob/main/GSoC.md>

Response to 3 (Instead of checking only one file, I went through all of them and identified bugs): <https://github.com/YYINUS/GSoC/blob/main/Potential%20Bugs.md>