AEM 2850 / AEM 5850 group-project

Overview

In this project, we will use tools from the tidyverse to wrangle and visualize data for firms in the S&P 500. In practice, if you were to use R for investment research in industry, you would want to use additional packages that provide specialized functions to process and visualize financial data. We do not have time to go into the details of these packages in this course, but this project should give you a taste of how R can be used in financial modeling.

The primary data for this project are in `sp500.RData`, which contains two data frames. `sp500_companies` contains information on the companies in the S&P 500, where each observation corresponds to a company. `sp500_prices` contains daily price records form 2017-2021 for companies in the S&P 500. Each observation corresponds to a symbol-day. For each observation, there are five prices (`open`, `high`, `low`, `close`, and `adjusted`) as well as the trading volume (`volume`).

Logistics and Expectations

- Work in the groups posted on canvas. Here are some tips for collaborating:
  - Start by agreeing on a way to manage your shared code (e.g., a shared project on Posit Cloud, local files synced through Dropbox, etc.). We set up separate workspaces for each group with projects that allow collaborative editing, so this will probably be the best approach
  - Leave notes to yourself and others in your source code(s)
  - Set up a timeline with regular meetings/check-ins
- Use Quarto to create a PDF document that summarizes your work, presents visualizations, and discusses takeaways. Submit **both** a compiled PDF and a .zip of your project directory that is self-contained, with all: the .qmd source file, .RProj, data files we provided, any other scripts or data you used, etc. We should be able to open your project and render your document with one click
- Do not use any packages other than those we use in class or are specifically referenced below
- Utilize TAs for Parts 1-3.3. If you need help with 3.4, please make a good faith effort first and then come to Prof. Gerarden's office hours with questions. You are on your own for Part 4.
- **You will be graded on both the content and its presentation.** Some tips:
  - **Please be concise** when writing text to summarize your work and discuss takeaways
  - **Make each graphic clear and comprehensible (even on its own):** follow best practices from class and readings when making visuals
  - **Do not include your source code in the memo** unless you think it is crucial for communicating your methods or findings. You can avoid printing the code in all the chunks by adding this code chunk at the start of the .qmd file (but below the yaml header):

```
knitr::opts_chunk$set(echo = FALSE)
```

- **Follow Quarto etiquette rules and style** that we introduced throughout the semester in labs: don't output extraneous messages or warnings, use nice formatting for tables, and adjust the dimensions for your figures as needed (see lab solutions for examples)
- Each group will submit a single deliverable and receive a single grade, but I reserve the right to penalize students who do not contribute to their group

Part 1: AAPL

Load `sp500.RData` and use `sp500_prices` to:

1. Make a plot of AAPL's adjusted stock price over time.
2. Create a candlestick chart for AAPL over the course of 2021. For simplicity, you may just plot a vertical line between the open and close prices each day, and omit the high and low prices. Use logical color-coding for whether the price change was positive or negative each day.
3. Tidy the price data and use it to compute AAPL's returns for each type of price from the previous day to the current day: $(p_t - p_{t-1})/p_{t-1}$. Express them as percentage points (i.e., ranging from -100 to 100, not -1 to 1). Use the resulting data frame to plot the returns for each day over 2017-2021 using a bar chart, with each type of price as a separate facet. Arrange the facets in two rows, with open and close in the first row, then high and low in the second row. Omit adjusted prices. Use logical color-coding for whether the return was positive or negative each day.

Part 2: The S&P 500

Use adjusted prices for Parts 2 and 3.

1. Compute the cumulative returns of each sector in the S&P 500, using the variable `weight` in `sp500_companies` to weight companies within each sector. Present the results in a nicely formatted table.[1]

Part 3: Our Class Portfolio

For this part, we will analyze the class portfolio based on survey responses at the beginning of the semester. The company names are listed in `our-companies.csv` along with the count of students who listed each company.

1. Load the data from `our-companies.csv`. Use it to filter `sp500_prices` to include just the firms in `our-companies`. Are any of our companies missing from the S&P 500? If so, what are their tickers?

For the remaining questions that refer to our portfolio, only use the companies from our class portfolio that are in the S&P 500.

2. Create a figure that summarizes the **cumulative** returns for: each individual stock in our portfolio, the portfolio as a whole, and the S&P 500. For our portfolio's weights, use the survey results summarized in the variable `n`. For the S&P 500, use the variable `weight` in `sp500_companies`. Use aesthetics like fill and/or color to compare and contrast different results, and organize them in a manner that facilitates comprehension.
3. Make a plot that shows the distribution of **monthly** returns for each stock in our portfolio over the period 2017-2021. Use adjusted prices from the first day of each month to compute monthly returns.
4. Use a linear regression model to estimate the capital asset pricing model using **daily** returns for each company in our portfolio:

$$r_{it} = \alpha_i + \beta_i m_t + \epsilon_{it}$$

[1] Some starter information on formatting tables in R Markdown is available here.

where r_{it} is the return for company i on day t, and m_t is the return of the S&P 500 index as a whole on day t. Work with total rather than excess returns for simplicity.[2] After you estimate the model, use the point estimates to make a visualization of stock performance for each of our companies. Try to create a visual that allows you to compare stocks in terms their αs and βs at the same time. Find a creative way to label the companies (e.g., using ticker symbols on the figure itself). Discuss the results. What are the pros and cons of your visualization? How could you use the information to inform your investment strategy? Does your visualization convey information about uncertainty in the αs and βs? If not, discuss how you could do so (conceptually, without worrying about implementation in R). *Please try to complete this question on your own, without any help from the TAs. If you put in a good faith effort and still need help, come to Prof. Gerarden's office hours.*

Part 4: Reflection

Note: You must complete this question without any help from the TAs.

1. Step back and discuss your work in Parts 1-3. Are all of the figures you produced useful to guide investment decisions? Which do you think is most useful? Which do you think is least useful? Explain your thinking.

Part 5: AEM 5850 Groups Only

Note: You may consult TAs for help with Part 5.

Quarto supports many different output formats beyond the PDF documents we have used for assignments in this class. For example, one can produce presentations in HTML, PPTX, or PDF formats.

1. Create at least one extra visualization based on the data used in this project.
2. Use Quarto to produce a PowerPoint that summarizes your key findings from this project, including your extra visualization(s). Please submit your rendered .pptx file as well as the .qmd script used to produce it. If you would like to do any manual post-processing, that is permissible, but please submit the final version as a second file in addition to rendered version.

[2] If you want an extra challenge, use excess returns instead. One way to do this would be to use 3-month T-bills to compute the risk-free rate of return. To get data for this, you could use the R package `quandl` to download 3-month T-bill prices (original data source: FRED).

Grading breakdown

Component	Points
1.1	3
1.2	4
1.3	4
2.1	5
3.1	2
3.2	5
3.3	5
3.4	10
4.1	5
Code submitted	1
Code renders document	3
Overall presentation	5
Total - AEM 2850	**52**
5.1	5
5.2	10
Total - AEM 5850	**67**