5206 Midterm

Wayne Lee

3/4/2021

Important notes:

- The dataset is quite large so you should NOT try to upload these to Ed.
  - You should have `tidyverse` and other packages used in class installed on your local computer.
- You may want to close your browsers if you're running into memory issues. If you have old variables from your previous projects, you may want to remove them with `rm()` first.
- All questions must be asked in a PRIVATE manner so your classmates will not see your posts on Ed.
- Wayne will be answering questions on Zoom at
  - 3/12, (class time) 10-11am, 12-12:40pm

  Then on Ed at
  - 3/12, 5-5:30pm, 10-10:30pm
  - 3/13, 7-7:30am, 8-8:30am
- You can return a .R, .Rmd, or a .pdf file on CANVAS. You must have all of your code and answers written in a single file. For non-code questions, please use `#` to create comments if you're turning in a `.R` file.
- Hardcoding is when you manually enter values that are not based on the data or problem. This is discouraged because its validity may change if the dataset is changed.
- If a question asks you to **calculate** something, make sure you have a variable assigned with the corresponding output.
- If a question asks you to **report** something, make sure you have a `print()` statement that prints out a message with the necessary information.
- We will take off points if you print out an entire data frame, list, or any data that occupies more than 20 lines (text wrapping will count against your line total) for the sake of understanding the data. The only exception is if the problem asks you to list out all the possible examples and this rule is violated.

Linking Financial Dynamics to the News

In this exam, we will be linking major stock market movements to specific news articles to automatically label historical events.

As a proxy to the overall stock market, we will use the adjusted daily prices of the Vanguard Index Fund VOO. The prices and the adjustments are done by Alpha Vantage. In general, high prices correspond to positive outlooks for the economy and vice versa.

Similar to our in-class exercises, we will use the NYTimes archives out of convenience to obtain our news sources.

Question 0: Honor Code

We, the students of Columbia University, hereby pledge to value the integrity of our ideas and the ideas of others by honestly presenting our work, respecting authorship, and striving not simply for answers but for understanding in the pursuit of our common scholastic goals. In this way, we seek to build an academic community governed by our collective efforts, diligence, and Code of Honor.

I affirm that I will not plagiarize, use unauthorized materials, or give or receive illegitimate help on assignments, papers, or examinations. I will also uphold equity and honesty in the evaluation of my work and the work of others. I do so to sustain a community built around this Code of Honor.

Will you follow the honor code?

- To answer this, please create a variable called `i_will_follow_the_honor_code` and assign the appropriate TRUE/FALSE value to it.
- You must also create a character variable named `UNI` that contains your UNI.
- WARNING: You will receive **0 points** for the midterm if you do not get this problem correct.

```
i_will_follow_the_honor_code <-
UNI <- ""
```

Q1.0 Wrangle the VOO data

For this problem, you'll need the file `alpha_vantage_voo_ts_daily.csv`.

- How many days are covered from the oldest to the newest record? Please show the code that does this calculation. You answer should be an integer and should change automatically if I change the dataset, e.g. `x<- 1; paste("Q1.0 answer is", x, "days")`.
- How many rows and columns are in the CSV file (2 integers)?
- Please create a new column called `percent_change` that contains the percent change of the `adjusted close` value (this is the adjusted prices) from the **working day before** (working days are M-F excluding national holidays), i.e. if a holiday closed the market, we will ignore that day.

Q1.1 Wrangle the VOO data

There is not a unique solution to this problem. If you cannot solve Q1.0, please use `5206_midterm_Q1.0_backup.csv`.

- Please report the average and standard deviation for the percentage change in the entire dataset, please report at least 4 significant digits.
- Please create a boolean (logical) column called `perc_change_outlier` indicating whether that day's percentage change in the adjusted closing price was an outlier.
  - We will define outlier events as events that occur less than 0.5% of the time.
  - You must have at least 5 records that satisfy the definition of an outlier.
  - If you decide to hardcode certain values, please show the code with comments that led you to the value. No penalty will be given if this is documented well.
  - Your definition should include both positive and negative changes.
- Using code, please show that your new column satisfies our definition and requirements of an outlier, i.e. if a human ran your code, there would be a readable statement based on your data that would inform us whether the definition and requirements are met. Hint: `paste()` allows you to compose messages.
- Please report the dates associated with the outliers and assign these to a variable called `outlier_dates`.

Q2.0 Wrangle the News Data

In the file `nytimes_archive.json`, you'll find a file that contains the metadata for various NYTimes articles. For Q2.0, please aviod using a for-loop.

WARNING: you may want to do this **last** given this is the most computationally time-consuming problem.

- Please create a data frame where each row corresponds to an article where the columns are:
  - `pub_date`: the published date of the article in "YYYY-MM-DD" format, e.g. "2014-01-01". This should be a character variable.
  - `sample_text`: the leading paragraph concatenated with the main headline of the article, e.g. if the leading paragraph is "hello" and the main headline is "world", you should have "hello world" here.
  - `related2fin`: A boolean indicating whether any of the lowered-cased keywords in the article has the substring "econ" or "stock" in it. Please use regular expression to achieve this task.
  - `news_desk`: the `news_desk` value for that article

Q2.1

If you cannot solve Q2.0, you can use the dataset `5206_midterm_Q2.0_backup.csv`.

- Please report the number of rows and columns for the data frame you'll be working with.
- Please calculate, for each day and each news desk type, the proportion of articles that have `related2fin=TRUE`.
  - Please wrangle this so the rows correspond to different days where the columns correspond to different `news_desk` values. Your data frame should have as many rows as unique dates and the number of columns should be at least as many unique values in `news_desk`.
  - Please make sure you have one column that contains the dates.
  - Please make a call about dealing with missing values here.
- Please report the dimension of this final data frame (or tibble) and report the first 3 records as well.

Q3 Data mining

Most of the outlier events are in March 2020, the month that many states in the US started to implement shutdowns. If this is not true for you, please re-examine your definition for an outlier.

To identify possible causes, please:

- Isolate articles in March 2020 where `related2fin` is `TRUE` then split the articles between those from days where `perc_change_outlier` is TRUE vs FALSE. We will call the first group "outlier articles" and the second group "non-outlier articles".
- Calculate the proportion of each keyword among all outlier articles, i.e. if 3% of the outlier articles have the keyword "Small Business", then you should have 0.03 matched to "Small Business".
  - You can assume that keywords are not repeated within the same article.
  - You can use whatever data type makes the most sense to you.
    - For the exam, please assume keywords are case-sensitive and any different spelling would be considered a different keyword.
- Please report the top 3 keywords among outlier articles according to the proportion. Please include the proportions in this report.
- Calculate the proportion of each keyword among the non-outlier articles.
- Please report the top 3 keywords among outlier articles according to the proportion. Please include the proportions in this report.
- Calculate the relative proportion of keywords between outlier vs non-outliers (i.e. outlier proportions divided by non-outlier proportions).
  - Please calculate this using 2 approaches:
    - one with a for-loop
    - one without a for-loop
  - Please handle the situation where keywords may not be shared between the two groups.
- Please verify your 2 approaches produce the same answer.
  - hints:
    - you may need to sort your output.
    - `all()`

- Please report the top 3 keywords according to the relative frequency. Please include the relative frequencies in your report.

Side comment (NO WORK REQUIRED): you should be able to article what we're doing here.