Assignment 3

Thomas Soupionis

2025-03-11

Load Packages

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Download Data

```
dow_data <- read_csv("../data/official_dowjones_data.csv")
```

```
## Rows: 33336 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (1): Close
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
spx_data <- read_csv("../data/official_spx_data.csv")
```

```
## Rows: 17116 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (1): Close
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
nasdaq_data <- read_csv("../data/official_nasdaq_data.csv")
```

```
## Rows: 13635 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (1): Close
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

List all the columns/variables in the data

```
names(spx_data)
```

```
## [1] "Date"  "Close"
```

```
names(dow_data)
```

```
## [1] "Date"  "Close"
```

```
names(nasdaq_data)
```

```
## [1] "Date"  "Close"
```

Describing the Variables

Dependent Variable

The dependent variable is the Return of the index. This variable measures the percentage change in the index value over a specified period. It represents the outcome of interest: whether an investment in the index fund yields a positive return over various time horizons.

Independent Variable(s)

The key independent variable is the Time Horizon for the investment. This variable indicates the duration (e.g., 2 years, 5 years, 10 years) over which the return is calculated. It allows us to explore how the likelihood of a positive return changes with the length of the investment period. Additional independent variables might include market conditions, volatility measures, or economic indicators if available.

Renaming Both Column for Each Dataset

```
dow_data <- rename(dow_data,
                   Investment_Date = Date,
                   Closing_Price = Close)
nasdaq_data <- rename(nasdaq_data,
                   Investment_Date = Date,
                   Closing_Price = Close)
spx_data <- rename(spx_data,
                   Investment_Date = Date,
                   Closing_Price = Close)
```

Verify Column Name Changes

```
names(dow_data)
```

```
## [1] "Investment_Date" "Closing_Price"
```

```
names(nasdaq_data)
```

```
## [1] "Investment_Date" "Closing_Price"
```

```
names(spx_data)
```

```
## [1] "Investment_Date" "Closing_Price"
```