Assignment 4

Thomas Soupionis

2025-04-21

Load Packages:

```
# tidyverse for data wrangling and ggplot2 graphics, lubridate for date parsing
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
```

```
library(lubridate)
library(ggplot2)
```

Import Data:

```
# Read daily closing data and add an "Index" identifier
# ymd() and mdy() convert character dates to Date objects

dow_data    <- read_csv("../data/official_dowjones_data.csv") %>%
  mutate(
    Date = ymd(Date),
    Index = "Dow Jones")
```

```
## 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.

## Warning: There was 1 warning in `mutate()`.
## i In argument: `Date = ymd(Date)`.
## Caused by warning:
## !  32433 failed to parse.
```

```
spx_data    <- read_csv("../data/official_spx_data.csv") %>%
  mutate(
    Date = mdy(Date),
    Index = "S&P 500")
```

```
## 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") %>%
  mutate(Date = mdy(Date),
         Index = "NASDAQ")
```

```
## 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.
```

Calculate Holding-Period Returns

```
# Function to compute percent return over 2-, 5-, and 10-year lags
# 504 trading days is about 2 years, 1260 is about 5 years, 2520 is about 10 years
calculate_returns <- function(df) {
  df %>%
    arrange(Date) %>%
    mutate(
      Return_2yr  = (Close / lag(Close, 504)  - 1) * 100,
```

```
      Return_5yr  = (Close / lag(Close, 1260) - 1) * 100,
      Return_10yr = (Close / lag(Close, 2520) - 1) * 100
    )
}

# Apply the function to each index series
dow_ret    <- calculate_returns(dow_data)
spx_ret    <- calculate_returns(spx_data)
nasdaq_ret <- calculate_returns(nasdaq_data)

# Combine all into one long data frame for plotting
all_returns <- bind_rows(dow_ret, spx_ret, nasdaq_ret) %>%
  select(Date, Index, starts_with("Return")) %>%
  pivot_longer(
    cols      = starts_with("Return"),
    names_to  = "Horizon",
    values_to = "Return_pct"
  ) %>%
  mutate(
    Horizon = recode(Horizon,
      Return_2yr  = "2 Years",
      Return_5yr  = "5 Years",
      Return_10yr = "10 Years"
    ),
    # set factor levels for correct order
    Horizon = factor(Horizon, levels = c("2 Years", "5 Years", "10 Years"))
  ) %>%
  filter(!is.na(Return_pct))
```

Summary Statistics

```
# Calculate mean, median, standard deviation, and positive/negative return percentages per horizon
avg_returns <- all_returns %>%
  group_by(Horizon) %>%
  summarize(
    Mean_Return   = mean(Return_pct),
    Median_Return = median(Return_pct),
    SD_Return     = sd(Return_pct),
    Pct_Positive  = mean(Return_pct > 0) * 100,
    Pct_Negative  = mean(Return_pct < 0) * 100
  )

# Display results in a table
knitr::kable(avg_returns, digits = 2,
             caption = "Summary: Mean, Median, SD, and % Positive vs Negative Returns by Horizon")
```

Table 1: Summary: Mean, Median, SD, and % Positive vs Negative Returns by Horizon

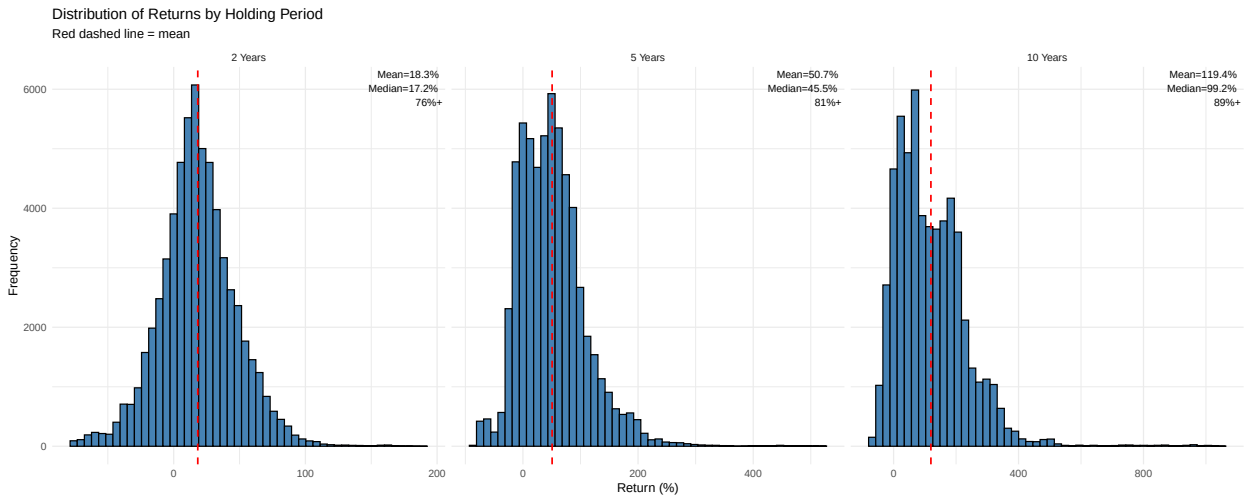
Horizon	Mean_Return	Median_Return	SD_Return	Pct_Positive	Pct_Negative
2 Years	18.28	17.20	29.03	76.28	23.69
5 Years	50.69	45.46	58.35	80.65	19.34
10 Years	119.43	99.20	114.21	89.14	10.85

Faceted Histograms with Mean Lines

```
# Precompute stats for annotating each facet
stats <- all_returns %>%
  group_by(Horizon) %>%
  summarize(
    Mean_Return   = mean(Return_pct),
    Median_Return = median(Return_pct),
    Pct_Positive  = mean(Return_pct > 0) * 100
  )

# Plot histogram with red line at mean and corner text summary
ggplot(all_returns, aes(x = Return_pct)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "black") +
  geom_vline(data = stats, aes(xintercept = Mean_Return),
             color = "red", linetype = "dashed", size = 0.7) +
  geom_text(data = stats,
            aes(
              x = Inf, y = Inf,
              label = sprintf("Mean=%.1f%%\nMedian=%.1f%%\n%.0f%%+",
                              Mean_Return, Median_Return, Pct_Positive)
            ),
            hjust = 1.1, vjust = 1.1, size = 3) +
  facet_wrap(~ Horizon, scales = "free_x") +
  labs(
    title    = "Distribution of Returns by Holding Period",
    subtitle = "Red dashed line = mean",
    x        = "Return (%)",
    y        = "Frequency"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

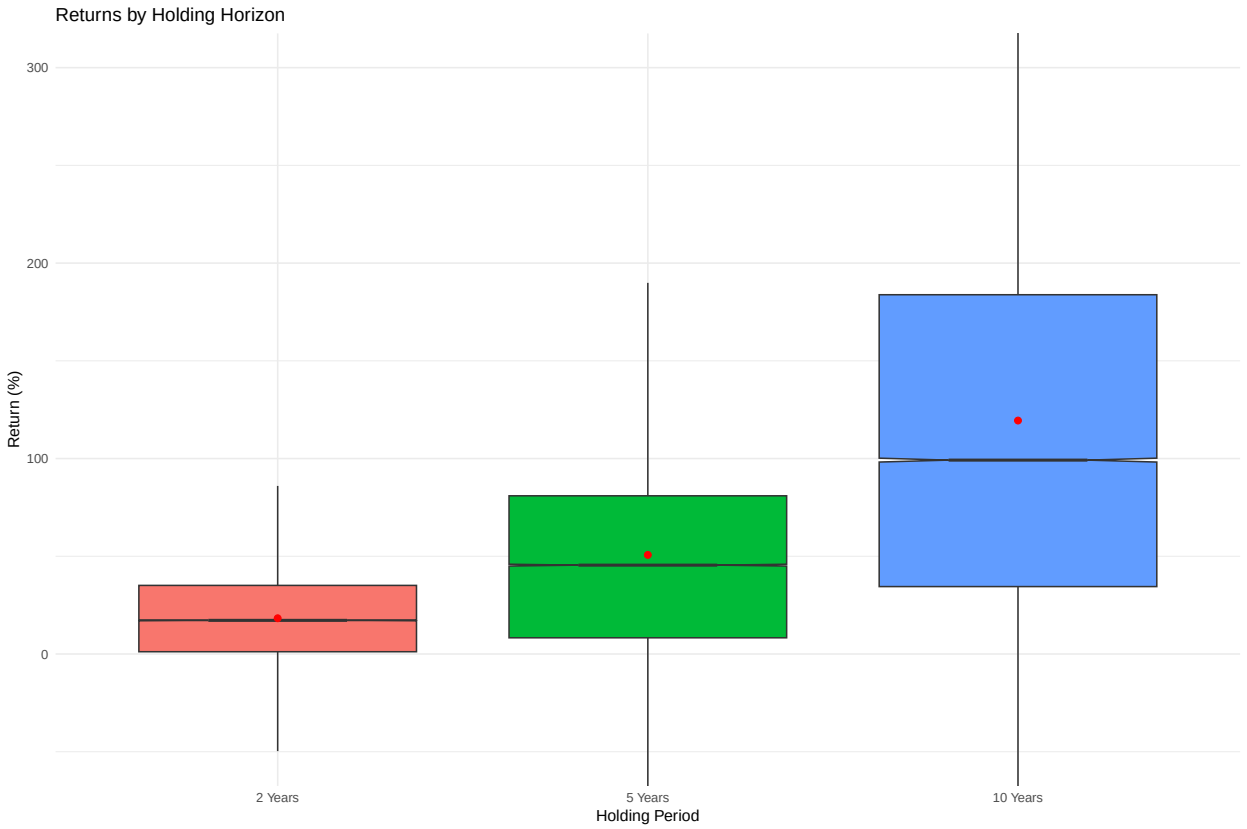


Description of Visualized Distribution

Over 2-year horizons, the histogram is wide and roughly symmetric around zero, showing many negative outcomes.
At 5 years, the bulk shifts rightward and the left tail shrinks, indicating fewer losses.
By 10 years, nearly all returns are positive and the distribution tightens, reinforcing that longer investments historically yield more consistent gains.

Boxplots Comparison

```
# Boxplot:
#   • notch = approximate 95% CI around the median (median ±1.58·IQR/sqrt(n)
#     -> non-overlapping notches suggest significantly different medians
#   • red dot = mean
#   • y-axis capped to focus on the main distribution
ggplot(all_returns, aes(x = Horizon, y = Return_pct, fill = Horizon)) +
  geom_boxplot(notch = TRUE, outlier.shape = NA) +
  stat_summary(fun = mean, geom = "point",
               shape = 20, size = 3, color = "red") +
  coord_cartesian(ylim = c(-50, 300)) +
  labs(
    title = "Returns by Holding Horizon",
    x     = "Holding Period",
    y     = "Return (%)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



Description of the Visualized Relationship

Looking at both the summary table and the boxplot, you can see a clear trend as the holding period lengthens:

2-Year Horizon

- Mean = 18.3 %, Median = 17.2 %, SD = 29.0 %
- 76.3 % of 2-year periods were positive, 23.7 % negative.
- In the boxplot the notch (approximate 95 % CI around the median) straddles zero, the box is wide, and the whiskers extend well into negative territory—showing high volatility and a non-trivial chance of losses over any given two-year window.

5-Year Horizon

- Mean = 50.7 %, Median = 45.5 %, SD = 58.4 %
- 80.7 % positive, 19.3 % negative.
- Here the notch sits entirely above zero and the box narrows compared to the 2-year plot. While variability remains (SD still > median), most five-year spans generate solid gains and the risk of a loss drops.

10-Year Horizon

- Mean = 119.4 %, Median = 99.2 %, SD = 114.2 %

- 89.1 % positive, 10.9 % negative.
- The ten-year box is tall but narrow, with a tight notch well above zero and only a few outliers below. This reflects very consistent—and large—positive returns over a decade, with downside risk greatly diminished.

Takeaway: As the horizon extends, not only do both the average and median returns grow, but the share of losing periods shrinks and the middle 50 % of outcomes becomes more tightly clustered well above zero. This strongly supports the argument that long-term investing in broad U.S. equity indexes historically offers higher and more reliable gains compared to shorter holding periodS or not investing at all.