[Project title. This will flow through to the header]

[Insert an abstract here. Summarize your project in 400 words or less. It should fit on this page. Put your name, email, phone number and date below the abstract.]

Jake Li  
lijake@umich.edu  
2484803953  
11/30/2022

Is this amount of data enough? Should I purchase more tickers?

Analysis complex enough? Should I analyze correlation with daily volatility?

Where to put hypothesis?

Harry Markowitz introduces the concept of *volatility* in his renoun *Portfolio Selection* paper (1952). He defines the volatility of a portfolio as the standard deviation of the returns of this portfolio. This definition of uncertainty in financial markets is very much agreed upon

Contents

[Introduction 1](#_Toc121214962)

[What goes in the introduction? 1](#_Toc121214963)

[Conventional Wisdom 1](#_Toc121214964)

[Relevance 1](#_Toc121214965)

[Conclusion 1](#_Toc121214966)

[The template 1](#_Toc121214967)

[Heading 2 example 2](#_Toc121214968)

[Heading 3 example 2](#_Toc121214969)

[Method and data 2](#_Toc121214970)

[Data Procurement 2](#_Toc121214971)

[Methodology 2](#_Toc121214972)

[Results 3](#_Toc121214973)

[Objective and Hypothesis 4](#_Toc121214974)

[Data Insights 5](#_Toc121214975)

[Statistical Significance 5](#_Toc121214976)

[Application 5](#_Toc121214977)

[Conclusion 5](#_Toc121214978)

Tables

Table 1. Big 10 Football Standings November 23, 2021 3

Figures

Figure 1. Percentage change from a year ago in the U.S. Consumer Price Index for all urban consumers: All items in U.S. City Average, seasonally-adjusted 3

# Introduction

## What goes in the introduction?

Tell the reader what your overall question is, how you went about answering the question, and what your conclusion is. Part of this section should be devoted to an explanation of why your question is interesting or important.

## Conventional Wisdom

It’s generally believed that the best time to buy an ETF is when it is most liquid, i.e. when trading volume is highest, or at the beginning and end of the trading day. The premise is that with high volume comes a narrowing of the ETF’s spread: the difference in its bid and ask prices (what buyers are willing to pay and what sellers are willing to sell for, respectively). Ideally, one buys when the spread is small and sells when the spread is large – this way, one minimizes transaction losses to the spread.

From basic economic theory, higher quantities of a good mean there are more ‘substitutes’ for the good and therefore price elasticity of the good increases; the good becomes more ‘commoditized’ and surplus decreases. In trading, this suggests that, when transacting in the market during times of higher liquidity, buyers *should* pay less of a spread ‘premium’ and sellers *should* receive less of that ‘premium’ as well.

Empirically, this seems to be the case – placing a limit order (an order to buy or sell a share for a specific price or better) at market open when liquidity is high, for example, seems to execute much more consistently than in the middle of the day when liquidity is lower.

## Relevance

The importance of this study is reflected in the implications of being on the ‘wrong side of the spread’ for large-volume institutional investors like market makers and high frequency traders (HFTs). These firms often employ delta-neutral strategies, and under those circumstances, must carefully analyze spread. Though spread is tiny compared to the underlying asset’s price, firms who trade frequently, even with market-neutral positions, can expose themselves to unnecessary losses by trading at the wrong times and overpaying for an asset.

## Conclusion

Surprisingly, the conventional wisdom is only half right. In fact, reality is almost the exact opposite – spread and volume correlate positively, meaning that, barring a few outliers, the smallest spreads are in the middle of the day when volume is lower.

## The template

Please use this template to make your report. It will help me to compare projects across students. Use the style formatting above to automatically create headings, and use “Insert Caption” under the References tab to automatically create captions for tables and figures.

* If you use the style formatting to automatically create headings you will be able to populate the table of contents.
* If you use captions you will be able to automatically create a list of tables and a list of figures. Make sure that you update the table of contents, and lists of tables and figures because they don’t automatically update. You right click on the list and select “update field” and then select the option to update the entire table.
* Have at least one table and at least one chart. How many tables and charts is best will vary depending upon the report. Exercise your judgment about presentation of tables (borders, right justification versus center justification, headings, layout etc.). Don’t let tables span one page to the next unless the table is so large is cannot fit on one page.

Your report must have the headings: Introduction, Method and data, Results, and Conclusion. The report should be only five pages.

The this template has examples of the following styles: Heading 1, Heading 2, Heading 3, and a bulleted list paragraph. It also has examples of a table and figure.

## Heading 2 example

### Heading 3 example

* Bullets example
  + Bullets example
    - Bullets example

# Method and data

## Template

In the method and data section, make a complete description of the data you analyzed, and the sources. The reader needs this information because the reader wants to know how generalizable the results are. Everyone’s project is different so interpret the description below in the context of your project.

Tell the reader how many companies you analyzed, and over what time period. Tell the reader how many days, weeks, months, or years or data you analyzed (this varies by project) and what the overall number of observations are. For example, if you had annual data, Apple 2020 and Apple 2019 represent two observations. The reader can determine whether results from 1980 are relevant for decision-making in 2021. Address whether your sample is of companies currently trading, or also includes inactive companies. Address whether other filters were applied, e.g. index membership, analyst coverage, market capitalization cut-off, sales cut-off. Address the industry coverage.

Tell the reader what your research method is. Often we write reports with separate method and data sections. But in your case, the report is only five pages, and it is challenging for you to write separate method and data sections. If is easier write the method in the same section as you use to describe the data.

## Data Procurement

Due to the high cost of granular historical intraday ticker data containing bid/ask prices, I am using a popular ETF (ticker IVE, the iShares S&P 500 Value ETF) as a case study. The results can loosely be generalized to other similar ETFs. I downloaded a dataset from Kibot for the IVE ETF with second timestamps, dates, close prices, bid and ask prices, and volumes. Kibot’s data is NBBO (National Best Bid and Offer) and is taken from multiple exchanges and ECNs, making it relatively trustworthy. The data spans 4,810 days (of which 3,316 had transactions) from 09/28/2009 to 11/28/2022 and contains 10.4 million entries.

## Methodology

My analysis involved the following steps: loading data, cleaning data, performing calculations, then performing grouping calculations and visualizations on several interval lengths.

To clean the data, I removed all entries with NA values or 0’s for the prices. These would have skewed the distribution of data significantly to the left. I also removed all entries whose timestamp fell outside of the normal 9:30AM-4:00PM trading day, as there was significantly less trading volume outside of this range. Trading ETFs after hours is generally risky and would be unrepresentative of typical trading behavior. The final step of cleaning was to remove all rows with uncharacteristically high spreads to reduce noise in the data. I arbitrarily set an upper limit of the spread being 2% of the asset’s price. In total, I removed 19,624 entries with ‘bad’ data, or 0.2% of the dataset.

To perform calculations, I used the bid and ask fields to calculate absolute spread for each entry. Since the price of the ETF changes over the years, I normalized the spread at each timestamp by converting it to a percentage of the price at that time.

I then identified several (arbitrary) interval lengths: 1-minute, 5-minute, 20-minute, 30-minute, and 1-hour blocks. I did not go below 1 minute because the data would be too granular and uneven, as large orders would cause spikes. Conversely, I did not go above 1 hour because there would not be enough blocks of time to visualize any trends, as a normal trading day only lasts 6.5 hours.

For each of these intervals, I performed a group-by operation to classify spreads into ‘bins’ of the chosen interval length, then took the mean for each ‘bin’. I repeated a similar process for volumes and normalized the bins’ sums to a percentage of the total trading day volume for readability and unit consistency, as total volumes at times of the day across years was less useful. Plots were generated from the outputs measuring spread over the course of the trading day, volume over the course of the trading day, and correlation between volume and spread.

Finally, I checked variance based on tertiary factors, such as year and month, to see if seasonality affected spreads. I performed a 2-level group-by operation on the tertiary factor and a chosen interval, like before. From this, a grouped bar chart was generated that showed changes in intraday spreads, month-by-month or year-by-year.

[insert table here for head of dataframe]

# Results

## Template

In the results section, explain your findings in the context of your research question. Some of you may elect to discuss statistical significance. But I have not emphasized statistical significance because we need to make resource allocation decisions. There is enough time to analyze one large dataset on a narrow research question, including extracting data, writing code and writing a five-page report. There is not enough time to also teach statistical inference. But we have discussed basic research methods.

* For example, one student is considering whether high or low leverage is associated with high or low price/earnings ratio, and that consideration has to be made by comparing firms in the same industry (because leverage and P/E ratios vary systematically across industries.
* A student is comparing whether enterprise value/sales ratios on profitable companies are higher today than just before the market peak of March 2000; again, this comparison is being made across industries because valuation ratios and profitability vary by industry.

In summary, your analysis is largely descriptive and you won’t necessarily measure the probability that you have observed a result by chance (the *p-*value). Some of you will include *p-*values on the basis of your prior statistical knowledge, but that is a choice that can be made by each individual student.

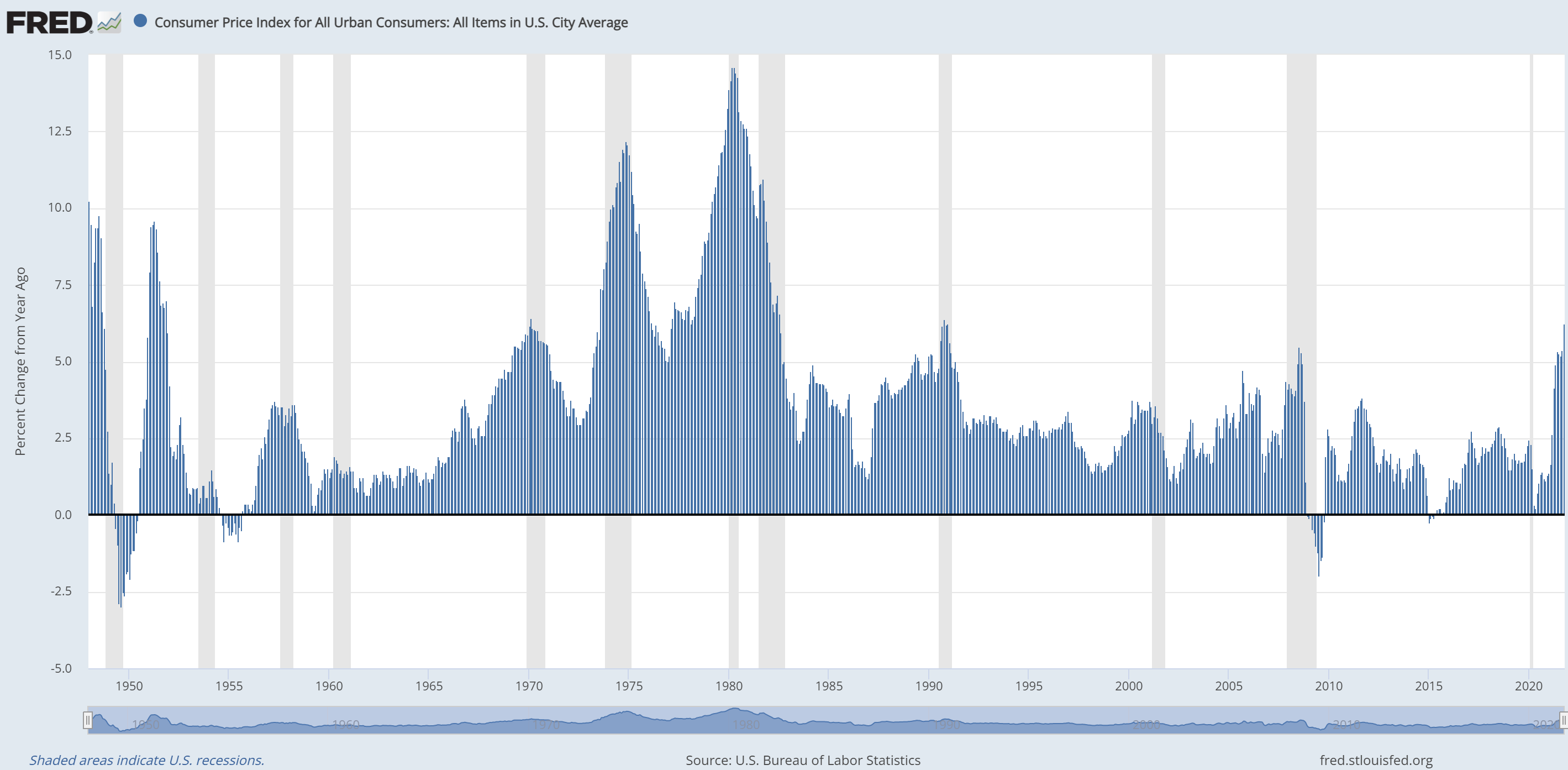
Here is an example of a table. Because I inserted a caption before making the table I can use Cross-reference under the References tab to refer to the table using above/below, table number or the table number and heading. For example, I can write “in the table ” and put in “below” using the Cross-reference button; or I can write “in ” and put in the table number using the Cross-reference button; or I can write “in ” and put in the entire heading.

Table . Big 10 Football Standings November 23, 2021

|  |  |  |
| --- | --- | --- |
| Team | Conference | Overall |
| **Big 10 East** |  |  |
| Ohio State | 8-0 | 10-1 |
| Michigan | 7-1 | 10-1 |
| Michigan State | 6-2 | 9-2 |
| Penn State | 4-4 | 7-4 |
| Maryland | 2-6 | 5-6 |
| Rutgers | 2-6 | 5-6 |
| Indiana | 0-8 | 2-9 |
| **Big 10 West** |  |  |
| Wisconsin | 6-2 | 6-3 |
| Iowa | 6-2 | 9-2 |
| Purdue | 5-3 | 7-4 |
| Minnesota | 5-3 | 7-4 |
| Illinois | 3-5 | 4-7 |
| Nebraska | 1-7 | 3-8 |
| Northwestern | 1-7 | 3-8 |

Here is an example of a figure. Using Cross-references I can refer to the figure , I can refer to , and I can refer to .

Figure . Percentage change from a year ago in the U.S. Consumer Price Index for all urban consumers: All items in U.S. City Average, seasonally-adjusted



## Objective and Hypothesis

The main objective of this study was to analyze whether conventional wisdom of buying at the start and end of a trading period (the trading day) holds true.

My hypothesis was that conventional wisdom is correct. I expect the highest volumes at market open and close; this would imply tighter spreads, and therefore I expect the market-neutral (regardless of whether the price will rise or fall) best time to transact is at market open and close.

## Data Insights

Figure XXX displays the spread (as a percent of the asset price) and volume (as a percent of the total daily volume) throughout trading hours, in ‘bins’ of XXX minutes.

As expected, volume roughly follows a ‘U’ shaped curve: the highest trading volume is indeed in the periods right after 9:30AM (market open) and right before 4PM (market close). The half hours after market open and before market close together account for over 26% of the daily volume.

Spread is highest at market open, and quickly shrinks until around 11AM, when it stabilizes. From then on, it trends downwards slowly for the remaining of the trading day. This might be due to the fact that while ETFs can be traded after hours (like stocks), lower liquidity means orders may not be filled until the next market open, creating a situation with a wide array of prices investors are willing to buy and sell at.

Interestingly, spread appears to be positively correlated with volume, especially from market open until around 2PM. Figure XXX further illustrates the spread-volume correspondence using ‘bins’ of XXX minutes. [say something about how the points which don’t fit on the line are after 2pm, illustrated by color or something] The overall correlation, R2, is XXX.

At the tail end of the trading day, volume increases dramatically whereas the spread stays relatively constant. With high liquidity and low spread, market close would be a good time to execute trades.

Under the assumption that prices follow a lognormal distribution???, I calculated the daily volatility of the ETF by calculating the natural logarithm of daily changes, grouping by month, and scaling to monthly and yearly volatilities. The results are illustrated in Figure XXX along with daily spreads. There appears to be a slight correlation between general spread levels and volatility: for example, in the year graph…

One would expect that higher volatilities would correlate with higher spreads because dramatic price swings should cause an … **Volatility usually increases during periods of rapid market decline or advancement**. At these times, the bid-ask spread is much wider because market makers want to take advantage of—and profit from—it (Investopedia)

## Statistical Significance

## Application

Jason’s example of a managed mutual fund receiving fees and goal of beating index by some %. If a fund rebalanced their portfolio fully, multiple times a year, even a fraction of a percent difference in spread can add up quickly.

DELETE

# Conclusion

Reach your conclusion in this section. Given that you only have five pages, this section will be short.