Research in Finance Homework 1

Chi-yu Cheng

Contents

1 Introduction

In this article, we investigate the weekly reports issued by the Commodity Futures Trading Commission (CFTC) using statistical and plot packages in R. We are interested in certain dynamics of the VIX futures and options market. Specifically we investigate the market participation on settlement days, and the correlation between the VIX index and the number of positions of various types of traders. Here is a brief description of the structure of this article.

1.1 Road map

We load the CFTC data set in Section 2.1 and combine it with the settlement days in Section 2.3. Readers who are not familiar with the mechanisms of the VIX market may consult Section 2.2.

After processing and cleaning, we calculate in Section 2.6 the average and the standard deviation across three columns of the data set, and compare their counterparts of non-settlement days. We see higher average and standard deviation on settlement days. We then provide histograms and time series plots in Section 2.7 that are in line with this observation.

In Section 3, we investigate more dynamics of the VIX market. In Section 3.1, we discuss how non-commercial spreads, defined in Section 2.2, are distributed through plots and statistical tests. In Section 3.2, we inquire whether there are correlations between the VIX index and the market participation. We will see that while the positions of non-commercial participants are negatively correlated to the VIX index, the largest four participants are not severely affected.

1.2 Relevant data sources

Here is the list of the sources of the data used in this article:

- The CFTC weekly report data at Nasdaq: https://data.nasdaq.com/databases/CFTC#anchor-futures-and-options-metrics-open-interest-and-trader-count-qdl-fon-
- The CFTC explanatory note: https://www.cftc.gov/MarketReports/CommitmentsofTraders/ExplanatoryNotes/index.htm
- The settle dates: https://www.macroption.com/vix-expiration-calendar/#history
- The VIX history: https://www.cboe.com/tradable_products/vix/vix_historical_data/

2 Problem 1

2.1 (a) Loading the data

We use our own API key to load the metrics *Open Interest and Trader Count*, and *Concentration Ratios* as *com* and *conc* respectively.

```
com <- Quandl.datatable('QDL/LFON',contract_code='1170E1',type="FO_L_ALL")
conc <- Quandl.datatable('QDL/FCR',contract_code='1170E1',type="FO_L_ALL_CR")
```

We now briefly walk through the structures of the data sets.

2.2 (b) Understanding the data

According to Nasdaq, the two data sets just loaded provide information on commitment of traders and concentration ratios. Here is a breakdown of the information of the codes that retrieved our data sets:

- 1170E1: Contract code referring to VIX futures and options market.
- com: The data set of Open Interest and Trader Count.
- conc: The data set of Concentration Ratios.

The VIX index measures how much people are willing to pay to buy or sell the S&P 500. Higher prices indicating greater uncertainty (Investopedia). Trading in the VIX futures started in 2004 and trading in VIX options started in 2006. One contract is on $1,000$ times the index.

For example, suppose a trader buys an April futures contract on the VIX when the futures price is 18.5 (corresponding to a 30-day S&P 500 volatility of 18.5%) and closes out the contract when the futures price is 19.3. The trader makes a gain of $800 (C.Hull 2017).

Let's take a look at the columns of our merged data set.

```
merged <- merge(com, conc, by = "date")
colnames(merged)
```

```
##  [1] "date"                    "contract_code.x"
##  [3] "type.x"                  "market_participation"
##  [5] "non_commercial_longs"    "non_commercial_shorts"
##  [7] "non_commercial_spreads"  "commercial_longs"
##  [9] "commercial_shorts"       "total_reportable_longs"
## [11] "total_reportable_shorts" "non_reportable_longs"
## [13] "non_reportable_shorts"   "contract_code.y"
## [15] "type.y"                  "largest_4_longs_gross"
## [17] "largest_4_shorts_gross"  "largest_8_longs_gross"
## [19] "largest_8_shorts_gross"  "largest_4_longs_net"
## [21] "largest_4_shorts_net"    "largest_8_longs_net"
## [23] "largest_8_shorts_net"
```

The columns are defined at Nasdaq. To further understand them better, one can look at CFTC's explanatory note. Here is a summary of some important information:

- Open interest is the total of all futures and/or option contracts entered into and not yet offset by a transaction, by delivery, by exercise, etc. Open interest held or controlled by a trader is referred to as that trader's position.
- The spread measures the extent to which each non-commercial trader holds equal short and long positions. For example, if a trader holds 500 long and 700 short positions, then the data set would record 500 spreads and 200 shorts from the trader.
- Reportable positions record traders that hold positions above specific reporting levels set by CFTC regulations.
- The concentration ratios show the percentages of open interest held by the largest four and eight reportable traders, without regard to whether they are classified as commercial or non-commercial.
- The largest four (resp. eight) positions account for the largest four (resp. eight) traders, represented as the percentage of total participation. The net position is obtained by offsetting each trader's equal long and short positions.

2.3 (c) Joining different data sets

We now join our data sets with the expiration dates. The original file provided is not up-to-date. We manually copied the dates listed on this site.

```
settledates <- read.csv(path_) %>%
    mutate(date = as.Date(Date) -1)%>% #Shift the expiration date as instructed
    subset(select = -c(Date)) # mutate adds a new column,
#drop Date to match the date column from the other data sets
# Keep non-settlement data by including all
```

```r
mergednew <- merge(settledates, merged, by = "date", all = TRUE) %>%
    arrange(date) # dates in the ascending order
```

2.4 (d) Processing

We discovered that there is always trading data on non-settlement days. This can be confirmed by the following code:

```r
# There are always trading on non-settlement days
mergednew[which(is.na(mergednew["VIX.Settle.Day"])),] %>%
    subset(select = -c(VIX.Settle.Day)) %>% # avoid counting na values from settle date
    is.na() %>%
    sum()
```

```
## [1] 0
```

However, on a settlement day, there can be either some or none trading data. The following codes confirm this:

```r
# There are no trading data on some settlement days
filter(mergednew, !is.na(VIX.Settle.Day)) %>%
    subset(select = c(non_commercial_longs)) %>%
    is.na() %>%
    sum()
```

```
## [1] 14
```

```r
# There are trading data on some settlement days
# Need to wrap !is.na() in curly braces in the pipeline for some reason
mergednew[which(!is.na(mergednew["VIX.Settle.Day"])),] %>%
    subset(select = c(non_commercial_longs)) %>%
    {!is.na(.)} %>%
    sum()
```

```
## [1] 215
```

2.5 (e) Cleaning

Instead of leaving NA on non-settlement days, we overwrite NA values with 0. This makes further analysis more convenient. For example, if one wants to compare trading behavior on settlement versus non-settlement days.

```r
# filter out earlier trading data and replace NA by zero for VIX.Settle.Day
mergedclean <- filter(mergednew, date > as.Date("2006-08-28")) %>%
    mutate(VIX.Settle.Day = ifelse(is.na(VIX.Settle.Day),0,1))
```

2.6 (f) Uncovering patterns in data

We will calculate the weekly differences of the market participation column, and the relative changes of the non-commercial spreads column. Namely, if M_n is the number of non-commercial spreads in week n, we calculate the quantity

$$\frac{M_n}{M_{n-1}} - 1 = \frac{M_n - M_{n-1}}{M_{n-1}}.$$

We then aggregate the mean and the variance for the market participation, the non-commercial spreads, together with the reportable shorts columns.

```
# Calculate weekly differences of market participation
market_participation_df <- mergedclean[c("VIX.Settle.Day",
                                         "date",
                                         "market_participation")] %>%
  mutate(difference = market_participation - lag(market_participation))
# Calculate relative changes of non commercial spreads
non_commercial_spreads_df <- mergedclean[c("VIX.Settle.Day",
                                           "date",
                                           "non_commercial_spreads")] %>%
  mutate(rel_change = non_commercial_spreads / lag(non_commercial_spreads) - 1) %>%
  # replace infinite values
  mutate(rel_change = ifelse(is.infinite(rel_change), NA, rel_change))
# aggregation
mergedclean[c("VIX.Settle.Day",
              "market_participation",
              "non_commercial_spreads",
              "total_reportable_shorts")] %>%
  group_by(VIX.Settle.Day) %>%
  summarize(# mp = market participation
            mpmean = mean(market_participation, na.rm = TRUE),
            mpstd = sd(market_participation, na.rm = TRUE),
            # ncs = non-commercial-spreads
            ncsmean = mean(non_commercial_spreads, na.rm = TRUE),
            ncstd = sd(non_commercial_spreads, na.rm = TRUE),
            # trs = total reportable shorts
            trsmean = mean(total_reportable_shorts, na.rm = TRUE),
            trsstd = sd(total_reportable_shorts, na.rm = TRUE))
```

```
## # A tibble: 2 x 7
##   VIX.Settle.Day  mpmean   mpstd ncsmean  ncstd trsmean  trsstd
##            <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>
## 1              0 302233. 154972.  67303. 45324. 278429. 145828.
## 2              1 318486. 164804.  73121. 49032. 293743. 155561.
```

We see that the average and the standard deviation are higher across the three columns chosen. Let us also apply the two-sample t-test to see if the difference is statistically significant.

```
# Test if market participation have the same mean
# for settlement and non-settlement days
t.test(filter(mergedclean, VIX.Settle.Day == 0)$market_participation,
       filter(mergedclean, VIX.Settle.Day == 1)$market_participation) %>%
  tidy()
```

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low conf.high
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl>
## 1  -16253.   302233.   318486.     -1.27   0.204      325.  -41360.     8854.
## # i 2 more variables: method <chr>, alternative <chr>
```

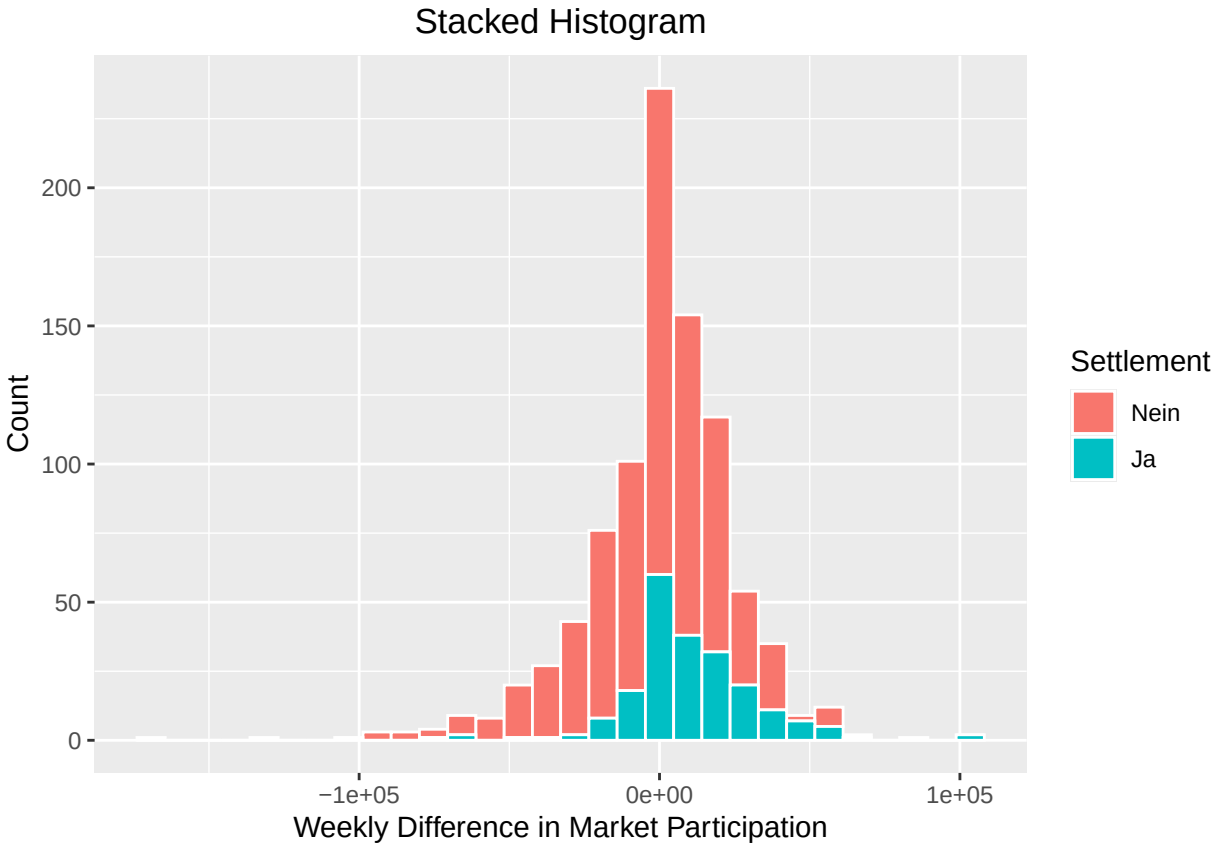
Since $p > 0.05$, the difference in average is not statistically significant. However, in the next section, we will display graphs that are in line with the higher mean values on settlement days.

2.7 (g) Analysis and plot

We demonstrate that trading activities tend to peak around settlement days by histograms and time series plots.

2.7.1 Market participation

```
# Histogram for total participation
ggplot(market_participation_df, aes(x = difference)) +
  theme(plot.title = element_text(hjust=0.5))+
  geom_histogram(aes(fill = factor(VIX.Settle.Day)), color = "white")+
# modify legend title
  scale_fill_discrete(name = 'Settlement',
                      labels = c('Nein', 'Ja'))+
  labs(
      title = "Stacked Histogram",
      x = "Weekly Difference in Market Participation",
      y = "Count"
      )
```



Here are the observations:

1. The red block is almost always longer than the cyan one in each bin. This is not surprising as there are more non-settlement days in the data.
2. The cyan blocks skew more to the right. This means an increase in participation is more likely to occur on a settlement day.

The second observation may be due to the phenomenon that many traders exercise their options or futures on the settlement day. We can confirm this by the following time series plot (truncated up to 2009-12-15 for better visualization):

```
# label settlement days with red vertical lines
vertical_lines <- filter(market_participation_df, date <= as.Date('2009-12-15'),
                         VIX.Settle.Day == 1)$date
```

```
ggplot(data = filter(market_participation_df, date <= as.Date('2009-12-15')))+
    theme(plot.title = element_text(hjust = 0.5))+
    geom_line(mapping = aes(x = date, y = market_participation), color = 'blue')+
    geom_vline(xintercept = vertical_lines, color = 'red')+
    labs(title = "Market Participation History", x = "Date", y = NULL)
```

We see that market participation tends to peak around settlement days, confirming our inference that traders tend to be more active on the settlement days. We next try out the column non-commercial spreads.

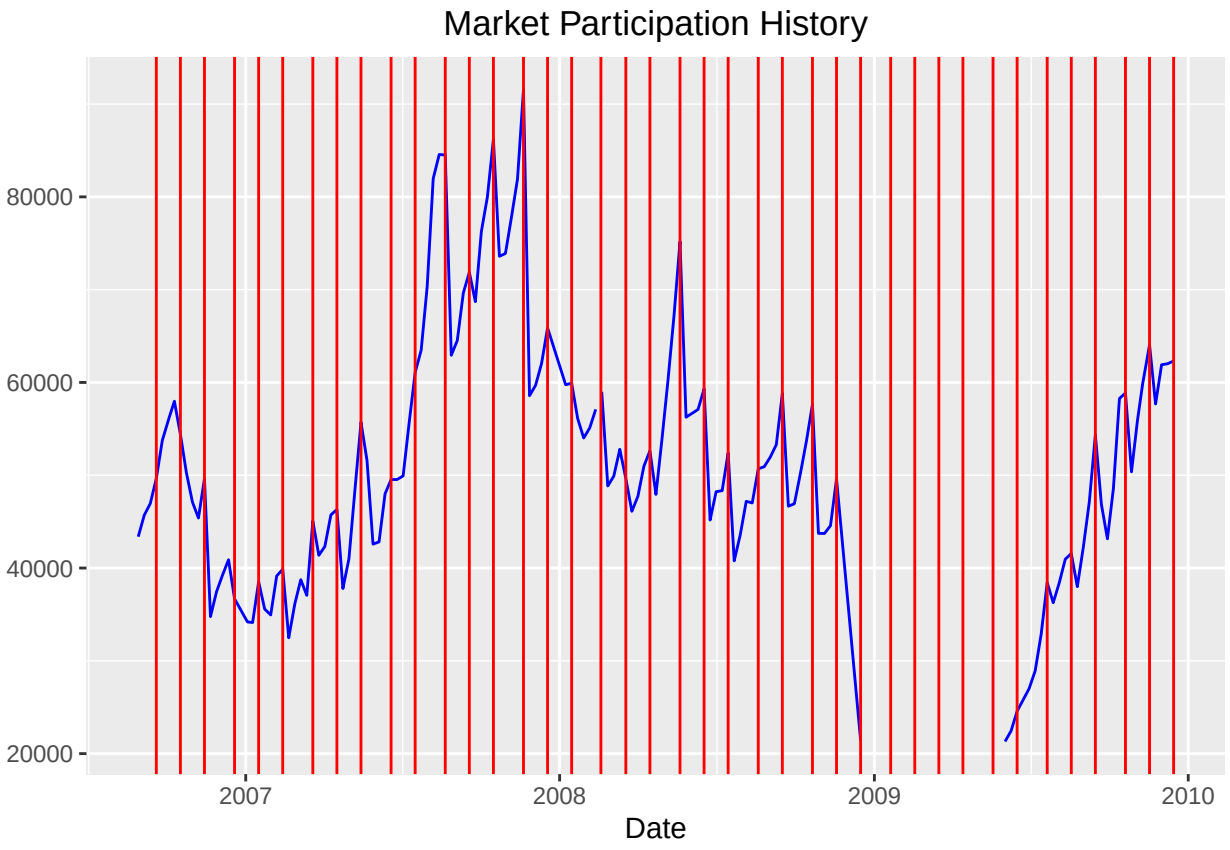
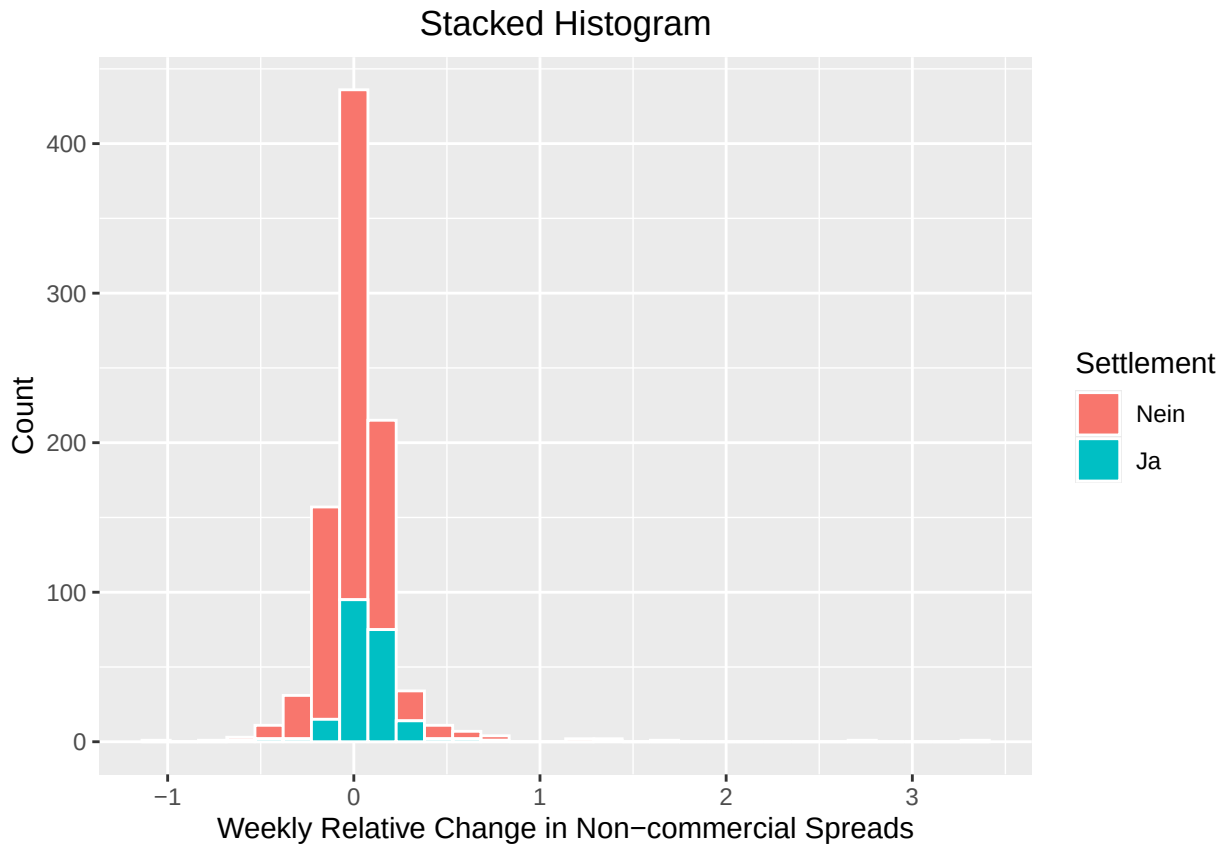


Figure 1: Time Series of Market Participations

2.7.2 Non commercial spreads

Here we visualize the relative change in non commercial spreads.

```
# Histogram for non commercial spreads
ggplot(non_commercial_spreads_df, aes(x = rel_change)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_histogram(aes(fill = factor(VIX.Settle.Day)), color = "white") +
  scale_fill_discrete(name = 'Settlement',
                      labels = c('Nein', 'Ja')) +
  labs(
      title = "Stacked Histogram",
      x = "Weekly Relative Change in Non-commercial Spreads",
      y = "Count"
      )
```



Here both the red and the cyan blocks skew to the right. In the next time series plot, similar to what we have seen in market participation, we see that the spreads tend to peak around settlement days. These are in line with the higher average we calculated earlier in Section 2.6.

```
ggplot(data = filter(non_commercial_spreads_df, date <= as.Date('2009-12-15')))+
    theme(plot.title = element_text(hjust = 0.5))+
    geom_line(mapping = aes(x = date, y = non_commercial_spreads), color = 'blue')+
    geom_vline(xintercept = vertical_lines, color = 'red')+
    labs(title = "Non-commercial Spreads History", x = "Date", y = NULL)
```

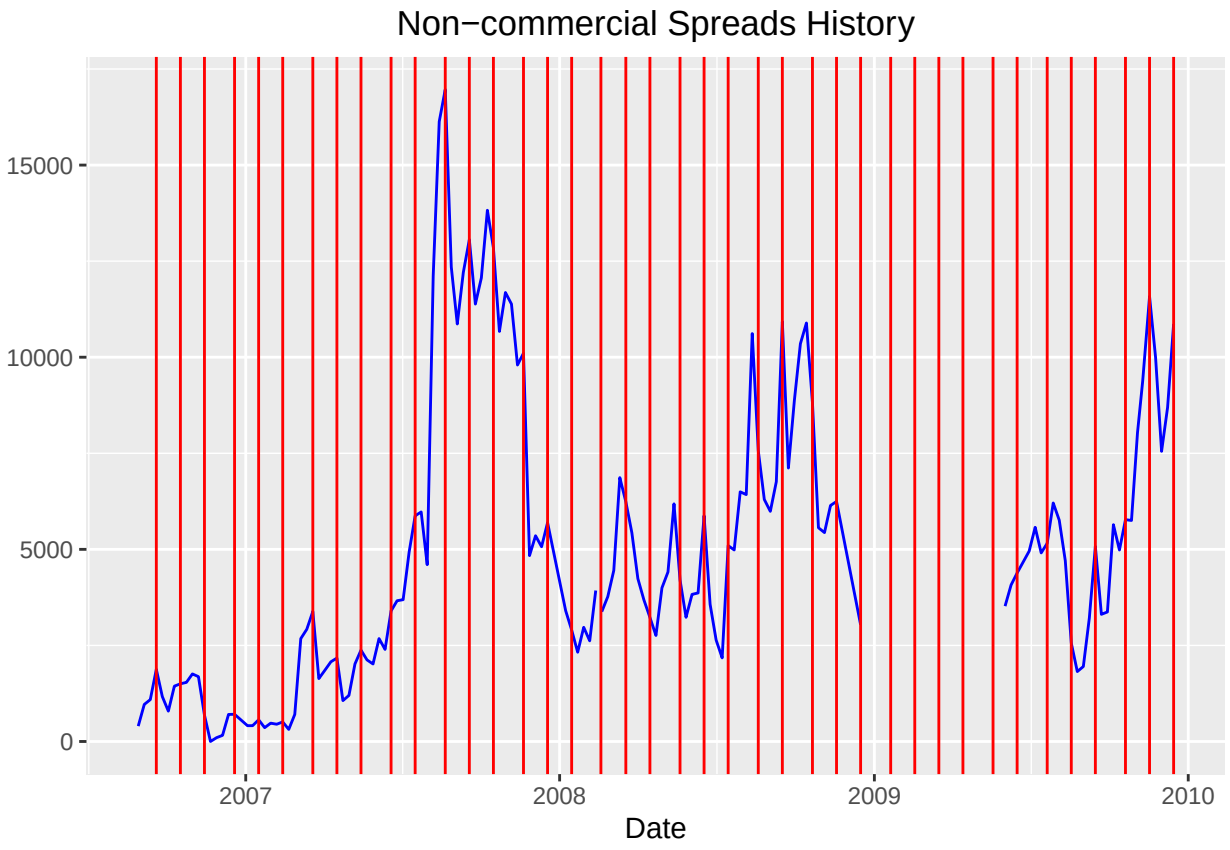


Figure 2: Time Series of Non-commercial Spreads

3 Problem 2

Here we are interested in the behavior of non commercial spreads. Specifically we examine the following two aspects:

1. Section 3.1: How close is the relative change of non commercial spreads to a normal distribution?
2. Section 3.2: How is non commercial spreads related to the VIX index? How does it compare to major players (the largest 4) on the market?

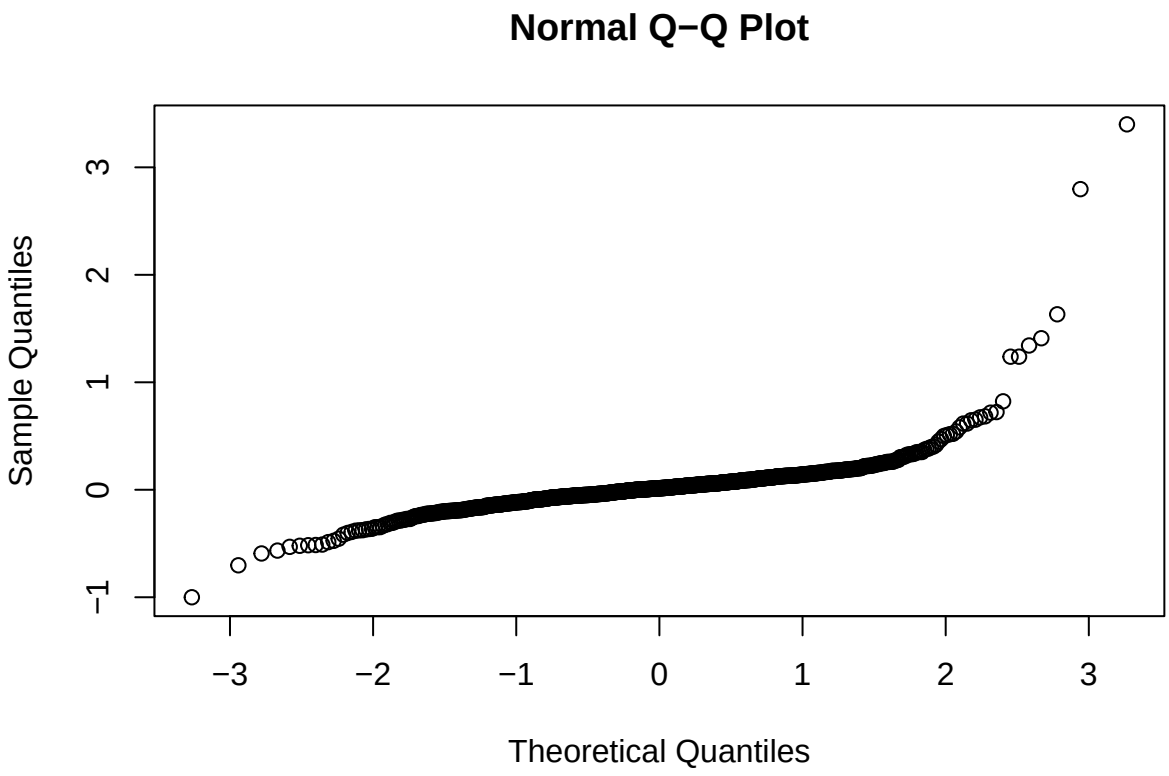
The first topic can be useful if one wants to model non-commercial spreads. The second topic comes from our curiosity to see if and how the major participants affect the market dynamics. Nevertheless, due to the scope of this article, we will not answer these questions in full. Still, we will see that

1. The entire data set of non-commercial spreads fails the normality test, while a middle portion that is approximately a third of it passes the test.
2. While the VIX index is negatively correlated to non-commercial spread, it is slightly positively correlated to the weight of major participants.

3.1 Normality test

We begin by plotting the sample quantile of non commercial spreads against the theoretical quantile of a normal distribution.

```
qqnorm(non_commercial_spreads_df$rel_change)
```



We see that it potentially resembles a normal distribution, except that there are some outliers on the tails. Let us apply the Shapiro-Wilk test.

```
shapiro.test(non_commercial_spreads_df$non_commercial_spreads)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  non_commercial_spreads_df$non_commercial_spreads
## W = 0.94731, p-value < 2.2e-16
```

Even though the W-value is high, the p-value is extremely low. Therefore, the entire data set fails the Shapiro test. Nevertheless, the Shapiro-Wilk test is sensitive to outliers. The following codes show that the test result depends heavily on the sample chosen:

```
##
##  Shapiro-Wilk normality test
##
## data:  non_commercial_spreads_df$rel_change[1:199]
## W = 0.76034, p-value = 2.913e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  non_commercial_spreads_df$rel_change[200:500]
## W = 0.99284, p-value = 0.1627
```

```
##
##  Shapiro-Wilk normality test
##
## data:  non_commercial_spreads_df$rel_change[501:800]
## W = 0.98699, p-value = 0.009057
```

We manually check which data corresponds to which date, and arrive at the following summary:

Period	Sample Size	W	p-value
08/29/2006 - 11/30/2010	199	0.760	2.913e-16
12/07/2010 - 08/30/2016	301	0.993	0.163
09/06/2016 - 05/17/2022	300	0.987	0.009

Table 1: Summary of normality test on different periods

The middle chunk does pass the Shapiro-Wilk test, but the other two chunks fail.

The middle chunk records the data from the December of 2010 to the August of 2016. The relative change during this period can be modeled as a normal distribution.

Also recall that the first chunk (from 2006 to 2010) among the three is the one that deviates from the normal distribution the most. Let us plot the time series to get some more insights.

```
# Time series plot for relative change
ggplot(non_commercial_spreads_df) +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_line(aes(x = date, y = rel_change)) +
  labs(x = "Date",  y = NULL)
```

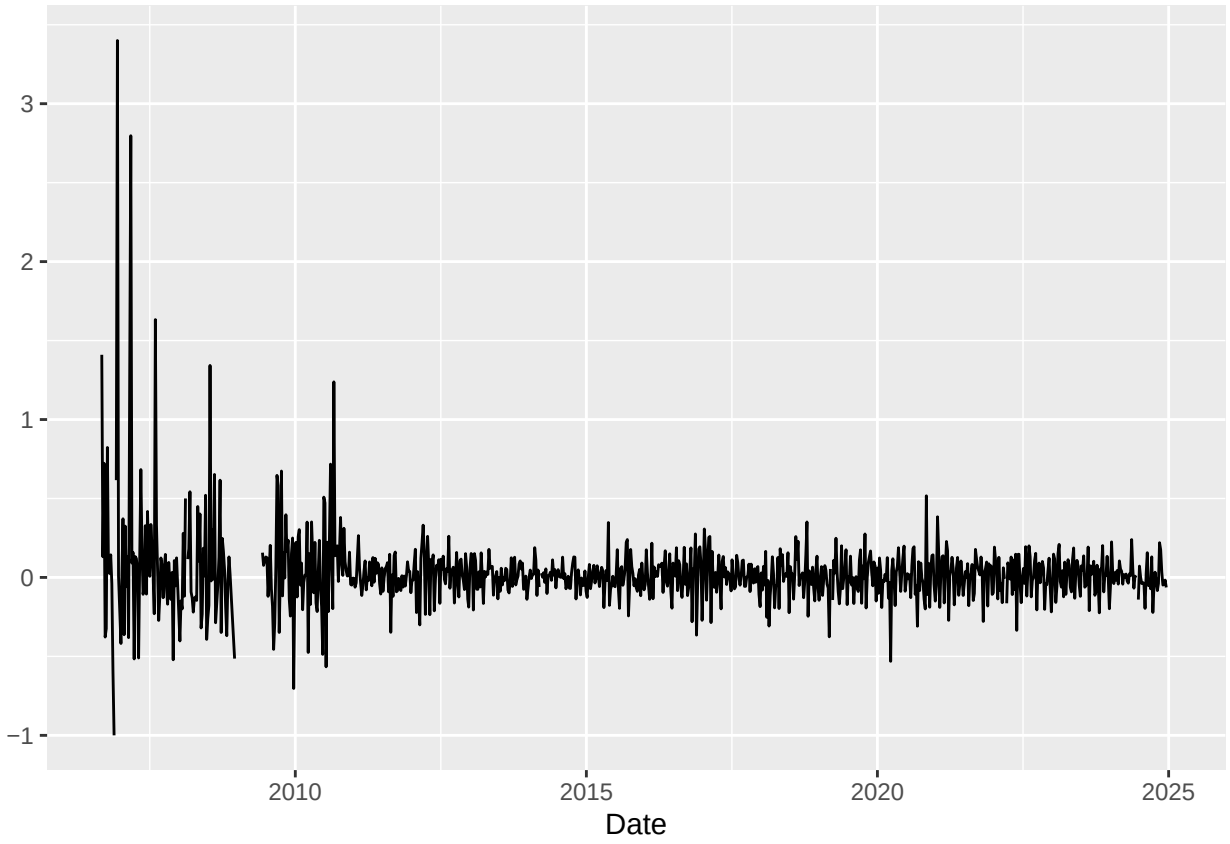


Figure 3: Time Series of Relative Change in Non-commercial Spreads

We see that

- The relative change becomes more stable after the year 2010.

In the following time series plot, we will also see that

- The spreads from 2006 to 2010 are small compared to the rest of the period, making the relative change more sensitive.

```
# Time series plot for spreads
ggplot(non_commercial_spreads_df) +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_line(aes(x = date, y = non_commercial_spreads)) +
  labs(x = "Date", y = NULL)
```

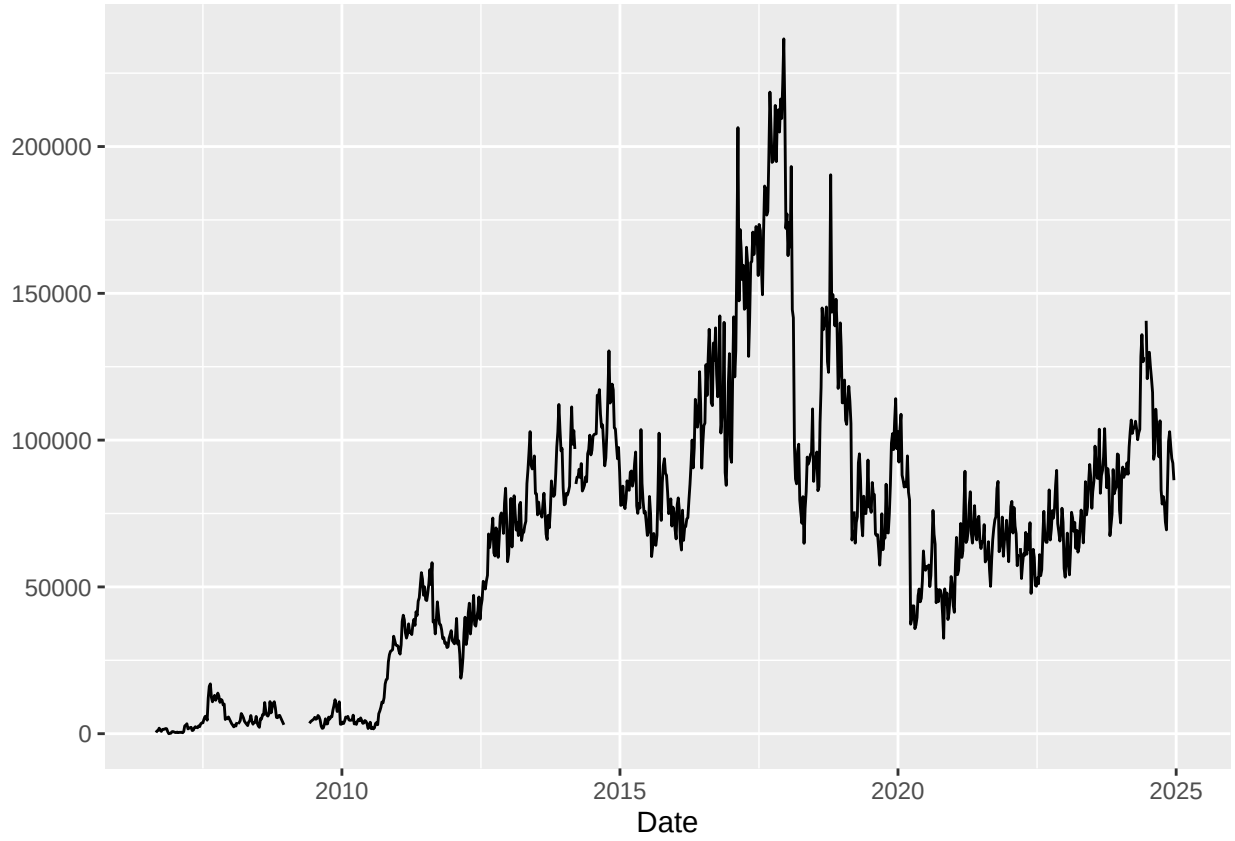


Figure 4: Time Series of Non-commercial Spreads

3.2 The effect of the VIX index

Here we investigate the correlation between the VIX index and non-commercial spread. We will see that they are negatively correlated, while the major participants are not negatively affected by the VIX index.

```r
# import, process the VIX index data
VIX_history <- read.csv(path_) %>%
  mutate(date = as.Date(DATE, format = "%m/%d/%Y")) %>%
  select(-DATE) %>%
  filter(date >= as.Date("2006-08-29"))
```

The VIX index is published daily at the Chicago Board Options Exchange, but the CFTC publishes data weekly. We therefore take the weekly average of the VIX index. The first step is to bin the VIX index by the dates from the *mergedclean* data set.

The VIX history contains the high, low, open and close values each day. We chose the daily high VIX index to take the average. There is no special consideration, as choosing others does not result in too much of a difference.

```r
# function to bin VIX_history
cut_dates <- function(date_list, breaks) {
  return <- cut(date_list, breaks = breaks,
                labels = breaks[-1],
                include.lowest = TRUE, right = TRUE)
}
# bin VIX_history dates
VIX_history$date_bin <- cut_dates(
  VIX_history$date, mergedclean$date
)
# aggregate (high) VIX weekly average
VIX_avg <- VIX_history %>%
  group_by(date_bin) %>%
  summarise(weekly_avg = mean(HIGH))
```

We are now ready to analyze the effect of the VIX index on the market. For better layout, we move the plot to the next page.

```
ggplot() +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(
      x = "VIX Index (averaged between CFTC report dates)",
      y = "Non-commercial Spreads"
      ) +
  geom_point(aes(x = VIX_avg$weekly_avg,
                 # Technicality: Since the first date has no prior week history,
                 # we cannot take VIX average for it
                 # Use -1 to remove the first entry
                 y = non_commercial_spreads_df$non_commercial_spreads[-1]))
```

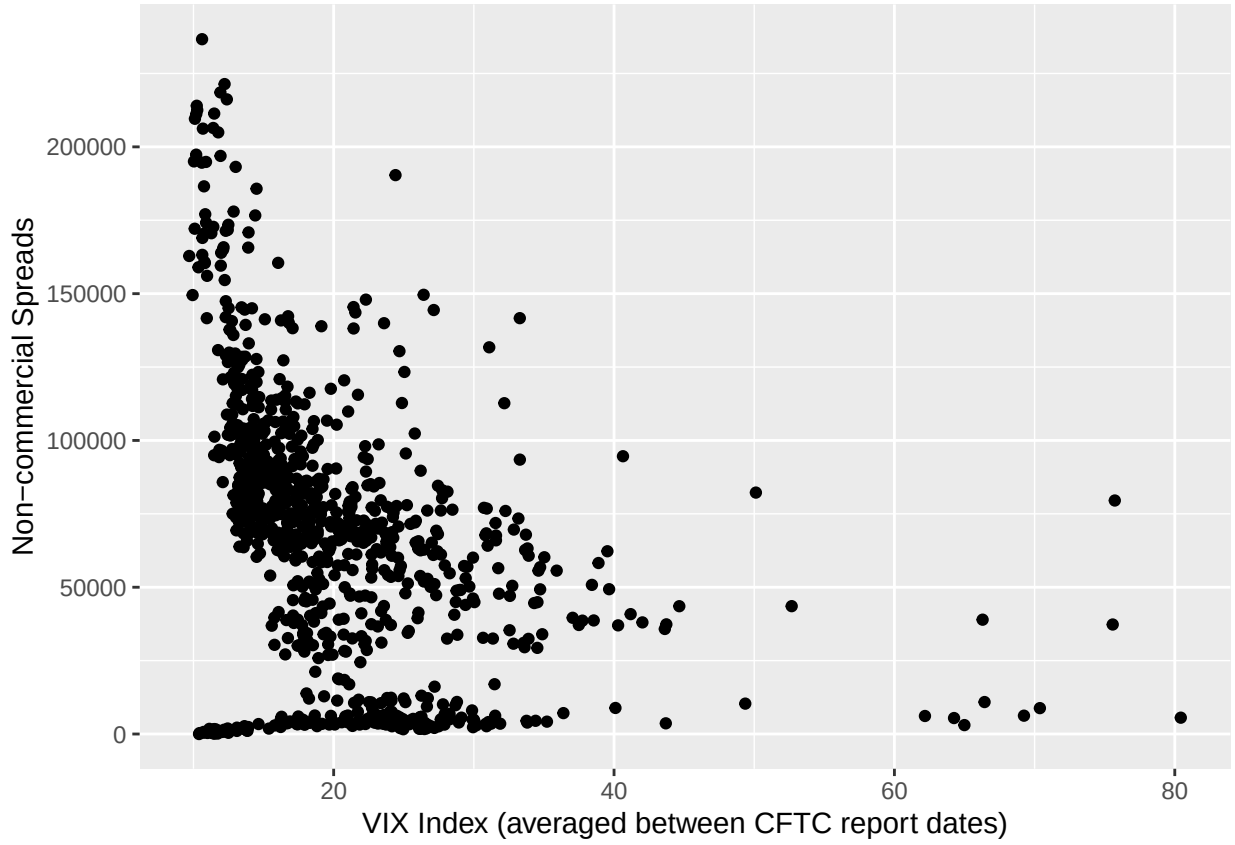


Figure 5: Scatter Plot of VIX Index Against Non-commercial Spreads

There seems to be a negative correlation between the VIX index and non-commercial spread. Let us apply the t-test for correlation.

```
cor.test(VIX_avg$weekly_avg, non_commercial_spreads_df$non_commercial_spreads[-1])
```

```
##
##  Pearson's product-moment correlation
##
## data:  VIX_avg$weekly_avg and non_commercial_spreads_df$non_commercial_spreads[-1]
## t = -13.725, df = 923, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4638532 -0.3567122
## sample estimates:
##       cor
## -0.4117044
```

We see that there is indeed a negative correlation between the VIX index and non-commercial spread. In fact, the same phenomenon holds across long and short positions for non-commercial participants. The following table summarizes the results of the t-tests applied to all positions of non-commercial participants:

	Non-commercial		
Position	Long	Short	Spread
t-statistic	-14.019	-14.545	-13.725
p-value	$< 2.2e$-16	$< 2.2e$-16	$< 2.2e$-16
Correlation	-0.419	-0.432	-0.412

Table 2: Summary of VIX effect on non-commercial positions

We have seen that when the VIX index increases, the number of non-commercial positions tends to decline. What about the major market participants? In the following, we plot the percentage of the largest 4 gross long positions in the market against the VIX index, and apply the t-test. We then summarize the results of the t-test applied to all positions of the largest 4 participants.

It is important to recall the comment on the CFTC website: A reportable trader with relatively large, balanced long and short positions in a single market, may be among the four and eight largest traders in both the gross long and gross short categories, but will probably not be included among the four and eight largest traders on a net basis.

For better layout we move the scatter plot to the next page .

```
ggplot() +
  labs(
    x = "VIX Index (averaged between CFTC report dates)",
    y = "Largest 4 Longs Gross (in %)"
      ) +
  geom_point(aes(x = VIX_avg$weekly_avg,
                 y = mergedclean$largest_4_longs_gross[-1]))
```

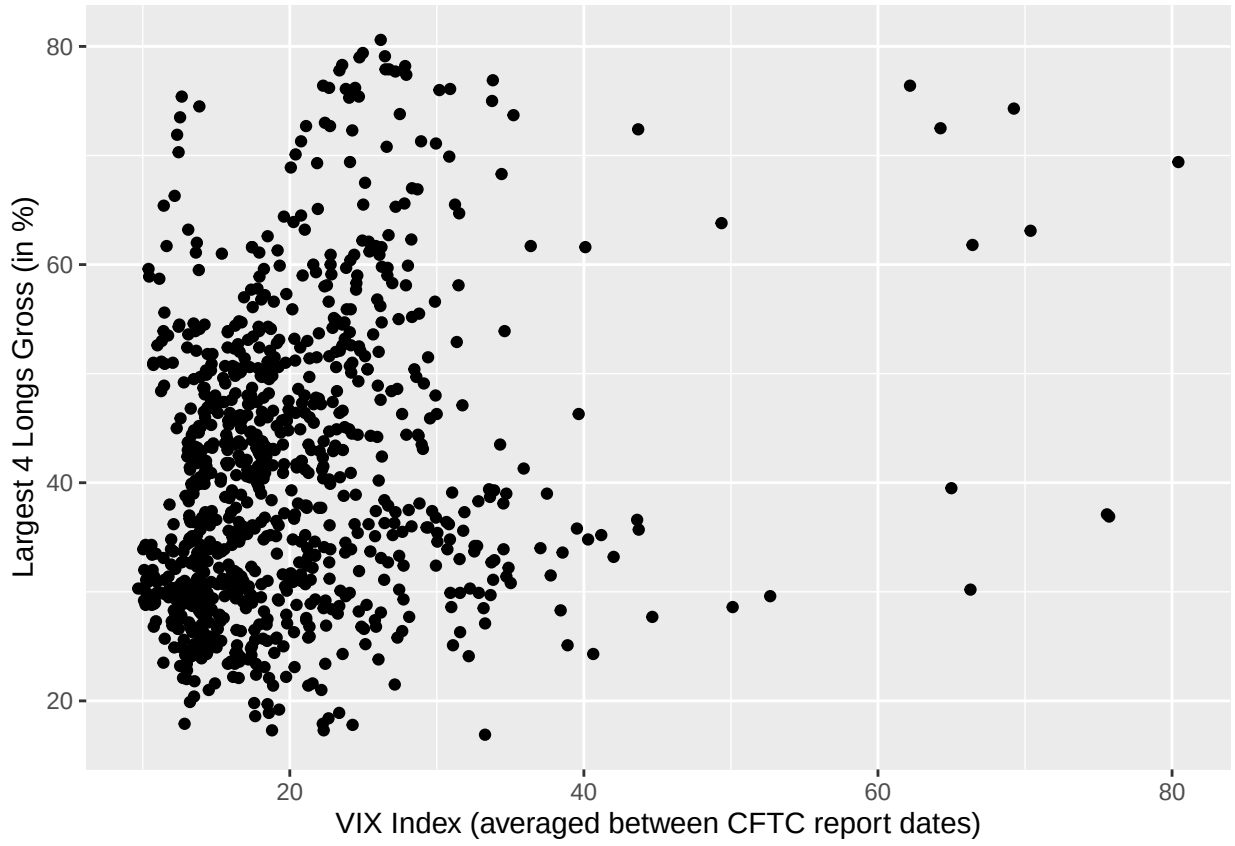


Figure 6: VIX Index Against Largest 4 Longs Gross

The scatter plot does not display prominent patterns. Let us apply the t-test as before:

```
cor.test(VIX_avg$weekly_avg, mergedclean$largest_4_longs_gross[-1])
```

```
##
##  Pearson's product-moment correlation
##
## data:  VIX_avg$weekly_avg and mergedclean$largest_4_longs_gross[-1]
## t = 7.1416, df = 923, p-value = 1.869e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1668340 0.2890269
## sample estimates:
##       cor
## 0.2288316
```

The test shows that there is a positive correlation between the VIX index and the largest 4 long positions in the gross category. Here is a table summarizing the results of the test applied to all positions of the largest 4 participants:

Position	Largest 4 Longs		Largest 4 Shorts	
	Net	Gross	Net	Gross
t-statistic	6.032	7.142	3.408	3.904
p-value	2.345e-09	1.869e-12	6.823e-04	1.016e-04
Correlation	0.195	0.229	0.111	0.127

Table 3: Summary of VIX effect on the largest 4 positions

We see that there are positive correlations between the VIX index and the weights of the largest 4 participants, where the correlation is more prominent in the gross category across long and short positions. We now summarize and interpret the results accumulated so far.

3.3 Summary and interpretations

In Section 2, we investigated the market behavior on the settlement days. Specifically we looked at overall market participation and non-commercial spreads. Figure 1 and Figure 2 suggest that they tend to peak around settlement days. This may be due to the fact the VIX options are European options, meaning that VIX options traders can only exercise on settlement days.

We have also seen in Section 2.6 that on average there is more trading on the settlement days. Nevertheless, the difference is not major, as the data fail the two sample t-test. This may be due to insufficient data, and that there are also VIX futures at play, where a trader can close out their contracts early for their own interests.

In Section 3.1, we have seen that the entire data set of non-commercial spreads does not fit a normal distribution. This is due to the following:

- The Shapiro-Wilk test is sensitive to outliers, so a large sample may fail the test.
- The market has different behavior across different periods. We have seen in Figure 4 that the participation was low before 2010, making the relative change more sensitive (Figure 3). This may be due to the fact that the VIX option was a new commodity (not introduced until 2006).

Nevertheless, the plot of the quantiles of non-commercial spread suggests that it should not deviate too much from a normal distribution. Indeed, the spreads from the December of 2010 to the August of 2016, which consist of about a third of the data set, can be modeled as a normal distribution (Table 1).

Depending on the purpose and practicality, one needs to determine if it is reasonable to model the relative change as a normal distribution. With further investigation one also needs to design a criterion that determines which points in the data set should be counted as outliers.

In Section 3.2, we have seen in Figure 5 a negative correlation between the VIX index and the non-commercial spread. In fact, the same phenomenon holds across all positions of non-commercial traders (Table 2). This means when there is more volatility, non-commercial traders tend to decrease their positions. It is natural also to investigate the behavior of the major participants.

In Figure 6 and in Table 3, we have seen slight positive correlations between the weight of the largest 4 positions with the VIX index. This means when the market is more volatile, major participants are not affected much. It is not very clear if the largest 4 traders actually increase their positions. For example, if the non-commercial traders actually decrease their participation, the largest 4 players weigh more without having to increase their participation.

We have also seen that the correlation between the VIX index and the weight of the largest 4 positions is more prominent in the gross category. This may be due to the fact that the weights of both longs and shorts in the gross category increase, leading to a less significant growth in the weights of the net category during volatile periods. Also, as was mentioned, the largest 4 participants in the gross category need not be the same in the net category. These two different categories may represent different types of traders.

References

C.Hull, John. 2017. *Options, Futures, and Other Derivatives.* Pearson.