Project 1: HO2 Analysis

Alex Hyman, Matt LaFlair, Robin Kim, Sasha Singh

7/13/2018

Introduction

The supply chain manager have notified us that our company has significant exposure to heating oil prices, specifically New York Harbor. The price of the heating oil affects the variable cost of multiple products, and can therefore affect our bottomline. When heating oil is volatile, so are our earnings. The purpose of this project is to explore the New York Harbors dataset and gain a better understanding of how the heating oil prices behave.

Part 1: HO2 data preparation and exploration

Data for the price of New York Harbors is provided by the turing.manhattan.edu website. Data exploration will be conducted with R. Using the url function, the dataset of the prices for New York Harbors from 1986-2016 will be loaded into R. The following R block will read the data, and ensure that the DATE column is not read as a factor. Data from the dataset will also be previewed with the head function and any pieces of missing or partial data will be removed with the na.omit function. The na.omit function will get rid of any rows that have NA as a value in either column. Finally, the structure of the data frame is viewed to ensure the data is set up as we wish.

The R skills that are required in this block include:

- Reading files
- Previewing Data
- Cleaning Data

```
# Getting the data from online at https://turing.manhattan.edu/~wfoote01/finalytics/data/
# Reading the nyhh02.csv file
# Setting strings as factors = FALSE so our dates remain as string, not factors
HO2 <- read.csv(url("https://turing.manhattan.edu/~wfoote01/finalytics/data/nyhh02.csv"),
                stringsAsFactors = F, header = T)
# Previewing the data
head(HO2)
```

```
##       DATE DHOILNYH
## 1 6/2/1986    0.402
## 2 6/3/1986    0.393
## 3 6/4/1986    0.378
## 4 6/5/1986    0.390
## 5 6/6/1986    0.385
## 6 6/9/1986    0.373
```

```
# Setting the new HO2 data frame to not include missing data
HO2 <- na.omit(HO2)
# looking at the structure of the data frames
str(HO2) # Two columns, DATE is a character, DHOILNYH is a numeric column
```

```
## 'data.frame':    7697 obs. of  2 variables:
##  $ DATE    : chr  "6/2/1986" "6/3/1986" "6/4/1986" "6/5/1986" ...
```

```
##  $ DHOILNYH: num  0.402 0.393 0.378 0.39 0.385 0.373 0.365 0.389 0.394 0.398 ...
```

Question 1 - Nature of HO2 Returns

To evaluate the nature of returns, it is necessary to create visual graphics. Plots of data tend to be simpler to process than viewing the data by itself. To model the daily changes in prices, the log difference between consecutive data points will provide the percent change in price. To model the volatility of heating oil prices, the absolute value of the percent change will be taken. Finally, percent changes in price will eventually need to be separated in order to evaluate the behavior of the price when the percent returns is negative versus positive. Each instance in the data frame will be labeled as "up", "down", or "same" based on whether returns were positive, negative, or remained unchanged.

The R skills that are required in this block include:

- Data manipulation
- Vector Creation
- Date Formatting
- Indexing on conditionals
- Manipulating vectors
- Data Frame Creation

```
# Calculating difference of logs returns to calculate approximate percent change
return <- as.numeric(diff(log(HO2$DHOILNYH))) * 100
# Getting the absolute value of the percent changes to get magnitude
size <- as.numeric(abs(return))
# creating a character vector that will hold information on whether the price went up, doum, or stayed
direction <- character(length(return))
direction[return > 0] <- "up"
direction[return < 0] <- "down"
direction[return == 0] <- "same"
# Converting the dates in HO2 to actual dates using the as.Date function. Deleting the first index of d
date <- as.Date(HO2$DATE[-1], "%m/%d/%Y")
# Creating a vector of the prices. Deleting the first index so we can have same lengths in vectors.
price <- HO2$DHOILNYH[-1]
# Creating a data frame for HO2 that contains the returns percent change, date, direction, magnitude of
HO2.df <- na.omit(data.frame(
                          Date = date,
                          Price = price,
                          Return = return,
                          Size = size,
                          Direction = direction
                           ))
# Looking at the structure of the data frame. The Dates should be included as dates and the direction s
str(HO2.df)
```

```
## 'data.frame':    7696 obs. of  5 variables:
##  $ Date     : Date, format: "1986-06-03" "1986-06-04" ...
##  $ Price    : num  0.393 0.378 0.39 0.385 0.373 0.365 0.389 0.394 0.398 0.379 ...
##  $ Return   : num  -2.26 -3.89 3.13 -1.29 -3.17 ...
##  $ Size     : num  2.26 3.89 3.13 1.29 3.17 ...
##  $ Direction: Factor w/ 3 levels "down","same",..: 1 1 3 1 1 1 3 3 3 1 ...
```

Now that all of the data has been cleaned and stored in a manner that is function for analysis, we will plot the returns of HO2 over time via the ggplot2 library and the ggplot function. The aesthetic function inside the ggplot function will set x-axis of the plot to the date and set y-axis to the return. The geom_line function

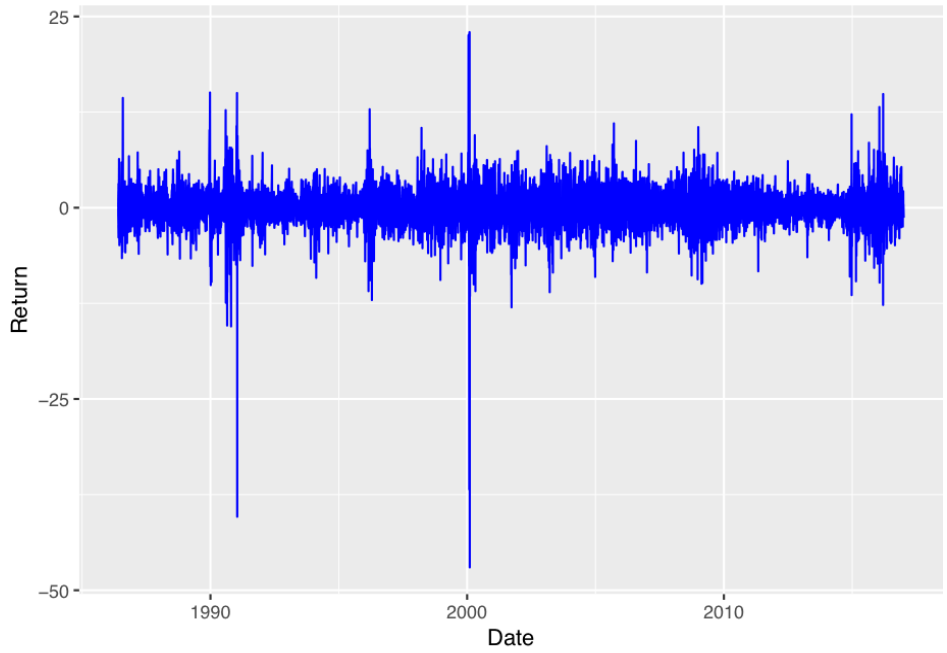

Figure 1: HO2 Returns

will be added to the ggplot to draw the line plot, and the colour argument inside the geom_line function will be set to blue to display the plot as a blue line plot.

The R skills that are required in this block include:

- Plotting with ggplot
- Using geom_line function
- Displaying plots

```
# Loading the ggplot2 library for graphics
library(ggplot2)
# Setting the variable p to the plot fo the returns over time
# The ggplot function initializes the plot object, HO2.df is the data for the plot
# aes function is setting the asthetic for the plot including x as the Date and y as the returns
# The group = 1 argument sets all the data in the same group to ensure all data are plotted
# The geom_line function is creating the line plot on the ggplot object with the colour set
# to blue to display a blue line
p <- ggplot(HO2.df, aes(x = Date, y = Return, group = 1)) + geom_line(colour = 'blue')
# the p is printing out the ggplot object to actually see the plot
p
```

A first glance analysis of plot of heating oil returns in Figure 1 suggests that there is a large amount of volatility clustering, specifically in the early 1990's and the early 2000's. There are also some less extreme clusters in the mid-1980's, mid-1990's, late 2000's, and in mid 2010's. The plot of HO2 returns also suggests that when HO2 returns are extremely volatile, there is a more dramatic downward spike in prices than an upward spike in the prices.

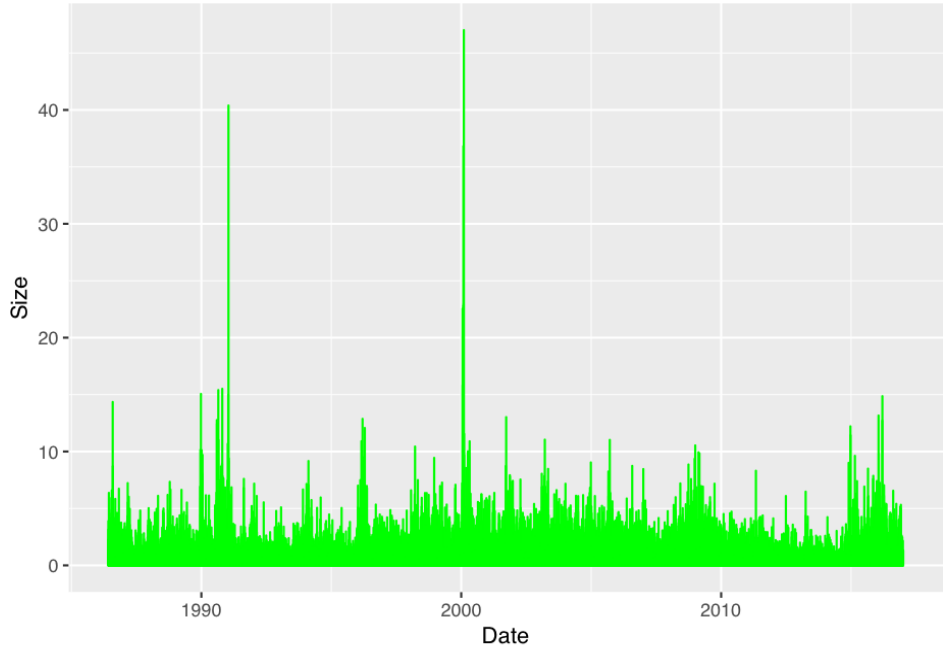

Figure 2: Absolute value of HO2 Returns

Next, we will create a bar plot of the absolute value of the returns to visualize the volatility of heating oil. This will be done using the ggplot function and the geom_bar function. The aesthetic function inside the ggplot function will set x-axis of the plot to the date and set y-axis to the absolute value of returns. The plot is displayed as a bar chart, which can be done by adding the the geom_bar function to the ggplot. Because we are not interested in count data, the stat argument within the geom_bar tool is set to identity to show the actual size of the absolute value of rate of returns over time.

The R skills that are required in this block include:

- Plotting with ggplot
- Using geom_line function
- Displaying plots

```
# Setting the variable p to the plot fo the returns over time
# The ggplot function initializes the plot object, HO2.df is the data for the plot
# aes function is setting the asthetic for the plot including x as the Date and y as the size
# The group = 1 argument sets all the data in the same group to ensure all data are plotted.
# The geom_bar function is used show the volatility in returns as a bar plot over time. The stat
# argument is set to identity so the bar is the same size as the returns, and is not counting the
# number of entries
p <- ggplot(HO2.df, aes(x = Date, y = Size, group = 1)) +
    geom_bar(stat = "identity", colour = "green")
p # Displaying the plot
```

The bar chart of the absolute value of returns displayed in Figure 2 supports our initial analysis that the returns for heating oil prices are extremely volatile and clustered. On multiple occasions there seems to

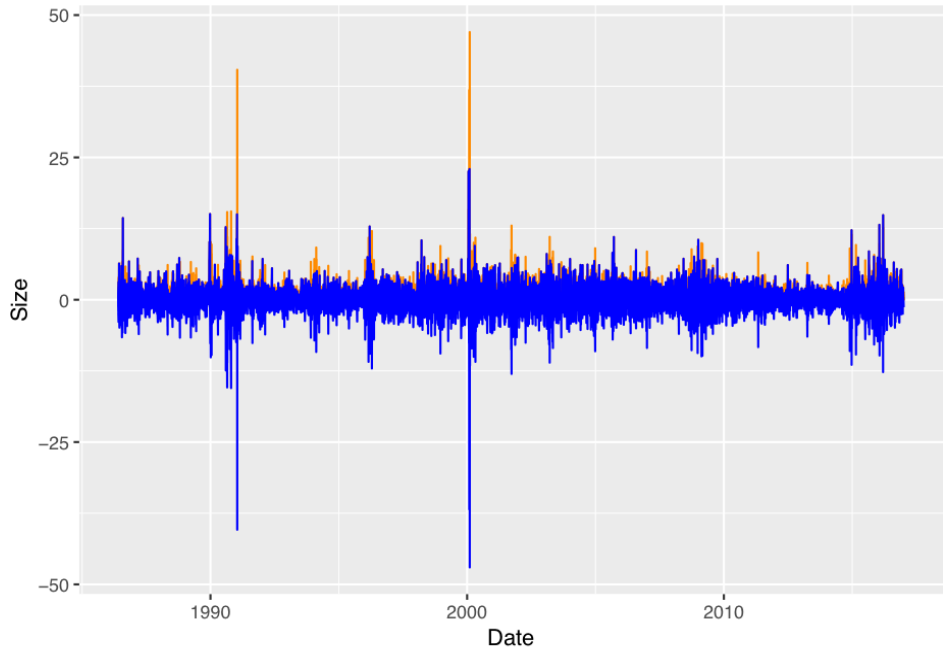

Figure 3: Overlay of HO2 Returns and Absolute value of HO2 Returns

consecutive days with a 10% or greater change in price. While this plot does not provide context of which returns are positive or negative, it does suggest that extreme changes in price are soon followed by other extreme changes in price in either direction.

Finally, we will create a plot that contains both the absolute value of the returns in an orange color and the actual returns in blue. This is can be accomplished by first adding the geom_bar function to ggplot placing it on the ggplot object first, and then adding the geom_line plot after, essentially placing the line plot on top of the bar chart.

The R skills that are required in this block include:

- Creating figures with ggplot
- Stacking plots to create overlays
- Displaying plots

```
# Plotting the orange size of returns geom_bar plot first (on bottom) and then overlaying the blue geom
# Creating the ggplot object, setting the aesthetic X to the data and the size as the y
p <- ggplot(HO2.df, aes(x = Date, y = Size)) +
    geom_bar(stat = 'identity', colour = 'darkorange') +
    # Drawing the line plot and making it the blue color
    geom_line(data = HO2.df, aes(x = Date, y = Return), colour = 'blue')
p
```

The plots provided in Figure 3 give us further insight on the behavior of negative and positive returns. When the orange bar is visible, the returns were negative for that day. With this in mind, we can see that Figure 3 confirms our suspicion that extreme prices changes in HO2 prices can be followed by another extreme change

in price, but in an either positive or negative direction. Figure 3 also informs us that when there are extreme changes in the price of HO2, when the returns are negative, they tend to be slightly greater than the positive returns in that particular volatility cluster. While this has generally been the trend, the most recent volatility cluster in the 2015-2016 time range has had a higher absolute value for positive returns than negative returns.

Question 2 - Data Moments

To further understand the nature of the HO2 returns, some descriptive statistics of the heating oil returns will be calculated. These descriptive statistics will begin our understanding of the distribution of heating oil returns. The statistics that will be calculated on the HO2 returns include:

- Mean
- Median
- Standard Deviation
- Kurtosis
- Skewness

The mean and median of the HO2 returns will provide us an understanding of the center of the distribution, or what we can expect for the long-term. The standard distribution will give us a numerical value to give some context to the actual volatility of the heating oil returns. The kurtosis of the distribution will provide an understanding of how large the tails of the distribution, or the commonality/magnitude of extreme events. A kurtosis greater than 3 would indicate that extreme events are more common or greater in magnitude in this distribution than a normal distribution. Finally, the skewness of the distribution will indicate which side of the distribution are the extreme events more likely to occur, or which side of the distribution has higher magnitude extreme events. If the skewness of the distribution is positive, then there are more smaller decreases to the price, and some extreme increases in price. If the skewness of the distribution is negative, then there are more smaller increases to the price, and some extreme decreases in price.

The following block of R code will create a function that will calculate each of these data moments for any provided set of data and return them in a data frame. A function is being used in this instance so the same calculations can be repeated multiple times without having to re-write the code. The R skills that are required in this block include:

- Creating function
- Basic statistics
- Loading packages
- Creating data frames
- Indexing columns of data frames
- Using knitr to display tables

```r
# Creating a function to calculate all the analysis data listed above
data_moments <- function(data) {
    # Need the moments library for skewness and kurtosis
    library(moments)
    #Calculate mean
    mean.r <- mean(data)
    # Calculate standard deviation
    sd.r <- sd(data)
    # Calculate median
    median.r <- median(data)
    # Calculate skewness (shifted lef or right)
    skewness.r <- skewness(data)
    # Calculte the kurtosis (peakedness)
    kurtosis.r <- kurtosis(data)
    # Creating a data frame with all the moments
    result <- data.frame(Mean = mean.r, 'Standard Deviation' = sd.r,
```

```
                        Median = median.r, Skewness = skewness.r,
                        Kurtosis = kurtosis.r
                        )
    # Returning the data frame
    return(result)
}
# Using the function we created on the returns
answer <- data_moments(HO2.df$Return)
# Rounding our data frame to only four decimal points
answer <- round(answer, 4)
# Creating table for display
knitr::kable(answer, caption = 'Data Moments of Heating Oil Returns')
```

Table 1: Data Moments of Heating Oil Returns

Mean	Standard.Deviation	Median	Skewness	Kurtosis
0.0179	2.5236	0	-1.4353	38.2595

The data moments provided in Table 1 shows that the returns of heating oil are centered around approximately zero, which is about what we would expect (median = 0, mean = 0.0179). Because the mean is slightly greater than the median, initial analysis would assume that the distribution could be positively skewed, but because there is a miniscule difference between the two measures of centrality, the skewness will provide a better picture of the actual nature of the skew. The distribution also has a standard deviation of 2.52%, which means that approximately 95% of the daily changes to the price of heating oil is between -5% and 5%. A standard deviation this large confirms our initial analysis from the previous plots that the prices for heating oil is quite volatile. The kurtosis of the heating oil returns is an extremely high value of 38.26, suggesting that there are significantly more events along the tails of the distribution of heating oil returns compared to a normal distribution and that the magnitude of these events is quite large. The skewness of the heating oil returns is also a very large value of -1.44. This skewness means that there is a significant negative skewness to the price of heating oil returns, meaning that there are either more extreme events on the negative side of the distribution than the positive side of the distribution or that the extreme negative returns are larger in magnitude than the positive returns.

Because the kurtosis and skewness statistics are extreme numbers, trying to estimate earnings could be quite difficult; however, because the distribution is negatively skewed, some of these extreme changes in price could be to our benefit, as this would give us reason to believe that when the prices are very volatile, the extreme changes in price are in our favor (price going down). Due to this heavy negative tail, it could be beneficial to alter our supply chain strategy and when we see an extreme decrease in the price of heating oil, either order in bulk for multiple shipments in the future instead of on an as-needed basis.

Question 3 - Pivot on Direction

To gain a better understanding of how both the positive and negative returns behave, we will create a summary table that provides the average returns, standard deviation of returns, 5th percentile, 95th percentile, and the percentage of instances for each of the three different returns categories (“up”, “down”, “same”). Logically, the returns that are the same should have an average of zero, standard deviation of zero, and both quantiles equal to zero. The pivot table for when returns are negative or positive will provide even more context into how the returns are distributed and give some more insight in how the tails of the distribution are behaving. The means provided in the pivot table will give an average for negative and positive returns. The standard deviation along with the 5% and 95% quantile will provide the middle 90% interval for the range of returns. Finally, the percentage of positive, negative, and same distribution will provide an idea of how likely it is for price to remain increase, decrease, or remain the same.

This block of R code will first provide counts of instances the price increased, decreased, and remained the same. It will then group the data frame by the "Direction" factor and calculate the average, standard deviation, 5th and 95th quantile, and percentage of instances using the summarise function in the dplyr package. The R skills that are required in this block include:

- Using the table function to get count data
- Pivot tables
- Descriptive statistics with R
- Formatting tables for displays

```
#Set results to "asis" from karl broman from kbroman.org
#Count number of results less than 0 (equal to zero is false)
table(HO2.df$Return < 0)
```

FALSE TRUE 4039 3657

```
#Count number of results greater than 0 (equal to zero is false)
table(HO2.df$Return > 0)
```

FALSE TRUE 3936 3760

```
#Count number of instances in each category
table(HO2.df$Direction)
```

down same up 3657 279 3760

```
#Counts number of returns that ended where they finished
table(HO2.df$Return == 0)
```

FALSE TRUE 7417 279

```
#Import the dplyr library for pivot table analysis
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Starting the pivot table by separtating by direction
pivot.table <- group_by(HO2.df, Direction)
# Making all columns visible
options(dplyr.width = Inf)
# Getting the number of entries so percantage can be calculated
HO2.count <- length(HO2.df$Return)
#Creating the pivot table based on the grouped direction. This data includes average return, standard d
pivot.table <- summarise(pivot.table,
                         return.avg = round(mean(Return), 4),
                         return.sd = round(sd(Return), 4),
                         quantile.5 = round(quantile(Return, 0.05), 4),
                         quantile.95 = round(quantile(Return, 0.95), 4),
                         percent = round(length(Return) / HO2.count, 4) * 100
                         )
#Displaying the pivot table
```

```
knitr::kable(pivot.table, digits = 2, caption = 'HO2 Returns Pivot Table by Direction')
```

Table 2: HO2 Returns Pivot Table by Direction

Direction	return.avg	return.sd	quantile.5	quantile.95	percent
down	-1.77	1.99	-4.78	-0.19	47.52
same	0.00	0.00	0.00	0.00	3.63
up	1.76	1.75	0.18	4.82	48.86

```
#Importing the xtable library
library(xtable)
#Initializing the caption for another table
HO2.caption <- "Heating Oil No. 2: 1986 - 2016"
#Printing the pivot table in LaTeX
print(xtable(t(pivot.table), digits = 2, caption = HO2.caption, align = rep("r", 4), table.placement = '
```

% latex table generated in R 3.5.1 by xtable 1.8-2 package % Mon Jul 30 12:29:22 2018

Table 3: Heating Oil No. 2: 1986 - 2016

	1	2	3
Direction	down	same	up
return.avg	-1.7718	0.0000	1.7598
return.sd	1.9862	0.0000	1.7460
quantile.5	-4.7761	0.0000	0.1817
quantile.95	-0.1894	0.0000	4.8203
percent	47.52	3.63	48.86

```
#Printing the details of all the returns
print(xtable(answer, caption = "Total Distribution of HO2 Returns", caption.placement = "top"), digits =
```

% latex table generated in R 3.5.1 by xtable 1.8-2 package % Mon Jul 30 12:29:22 2018

Table 4: Total Distribution of HO2 Returns

	Mean	Standard.Deviation	Median	Skewness	Kurtosis
1	0.02	2.52	0.00	-1.44	38.26

Table 2 and Table 3 provide a pivot table for the heating oil returns with categories “up”, “same”, and “down” indicating the direction of the price change. Because the returns in the “same” category have values of zero for their statistics, they will be excluded from this discussion.

Tables 2 and 3 show that the average returns when the price has increased is 1.76% while the average returns when the price has decreased is -1.77%, showing a fairly similar central tendency. The difference between the two categories becomes more apparent when looking at the standard deviation of returns. Negative returns have a standard deviation of 1.99% while positive returns have a standard deviation of 1.75%. This shows that the negative returns have a greater spread than positive returns, even to the point where it would be impossible for a daily negative return to be greater than one standard deviation from the mean (negative returns must be less than 0).

It becomes even more apparent that the extremes of the negative returns are more significant than the extremes of the positive returns when observing the comparable quantiles in conjunction with the standard deviation. The middle 90% of positive returns has a larger spread in values than the middle 90% of negative returns (0.1817 to 4.8203 vs. -0.1894 to -4.776). While the spread of the middle 90% is larger for positive returns than negative returns, knowing that the standard deviation for negative returns is greater than the

positive returns must mean that there are more values extreme in magnitude at the tail of the negative returns distribution than are at the tail of the positive returns distribution.

Finally, we note that 48.86% of days result in heating oil increasing in price versus the 47.52% of days that result in a decrease in the price of heating oil (3.63% have no price change). This would signify that there is approximately an equal chance the price of heating oil will increase as there is the price will decrease. With the mean of returns centered around zero, and our knowledge of the central limit theorem, if heating oil prices are particularly high on a given day, given that the need for heating oil is not urgent, we can expect the price of oil to return to a manageable price in the near future.

Part 2: HO2 Distribution

Question 1. Differences in Shape

To further understand how the positive returns behave differently than the negative returns, it can be helpful to create a cumulative distribution function for both the negative and positive returns. The cumulative distribution will provide display what percentage of returns are below a given value for both the negative and positive returns. A cumulative distribution function can also be created for all the returns to provide a whole picture on the shape of the distribution. The cumulative distribution function for "same" returns was not included in this section because it would be a vertical line at X = 0, which would clutter the plot. The R skills that are required in this block include:

- Using the quantile function
- Manipulating text with the paste function
- Plotting with ggplot
- Creating cumulative distribution function for different factors
- Creating vertical lines
- Adding annotations to ggplot objects
- Using the or statement

```
#Setting the tolerance to 5%
HO2.tol.pct <- 0.95
#Figuring out what the 95th percentile of returns would be
HO2.tol <- quantile(HO2.df$Return, HO2.tol.pct)
#Creating the label with the past function that says where our tolerable limit is
HO2.tol.label <- paste("Tolerable Rate = ", round(HO2.tol, 2), sep = "")
#Plotting the cumulative density function for negative, positive, and same returns. Same returns should
ggplot(HO2.df[HO2.df$Direction == "up" | HO2.df$Direction == "down",], aes(Return, colour = Direction))
    # starting the plot with returns as our data and a different density function for each direction
    stat_ecdf(size = 0.75) + # plotting the density functions in blue
    geom_vline(xintercept = HO2.tol, # drawing a vertical redline at the tolerance limit
        colour = "red", size = 1.5) +
    annotate("text", x = HO2.tol + 10, # adding the label to the plot
        y = 0.75, label = HO2.tol.label,
        colour = "darkred") + theme(legend.position = "top")
```

```
# Creating a plot that shows overall cumulative distribution
ggplot(HO2.df, aes(Return)) + stat_ecdf(colour = "blue") + geom_vline(xintercept = HO2.tol, colour = "r
```

The plot provided in Figure 4 shows the differences in shape between the positive and negative returns. The tolerable risk at the 95% quantile is also provided on the plot in a dark red vertical line, indicating where the cost could be prohibitive for the company to move forward.

The cumulative distribution function for the negative returns seem to become noticeable around the -11% returns and begins to rapidly increase in slope around the -5% returns. The slope continues to sharply

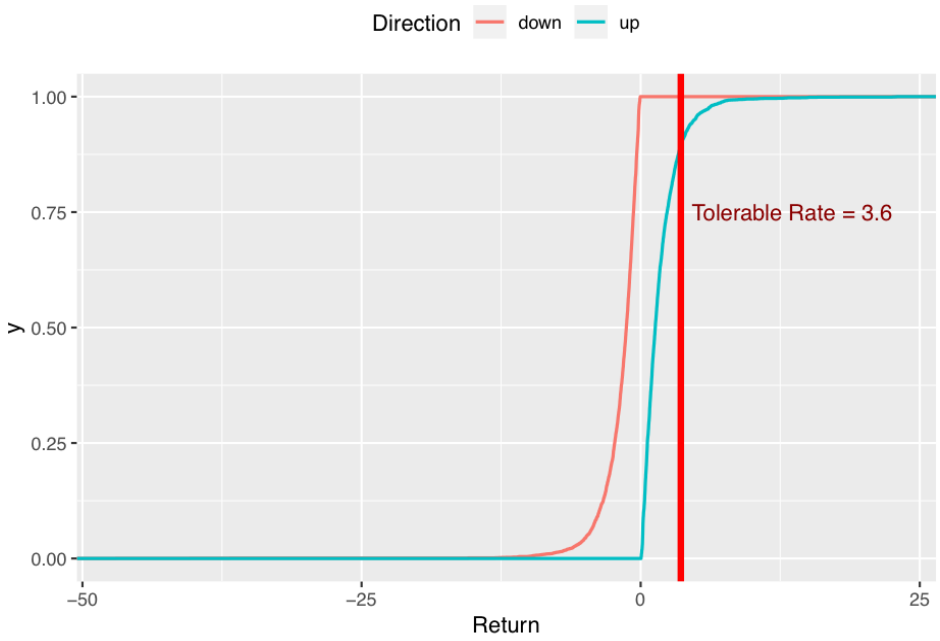

Figure 4: Cumulative Distribution Function by Direction

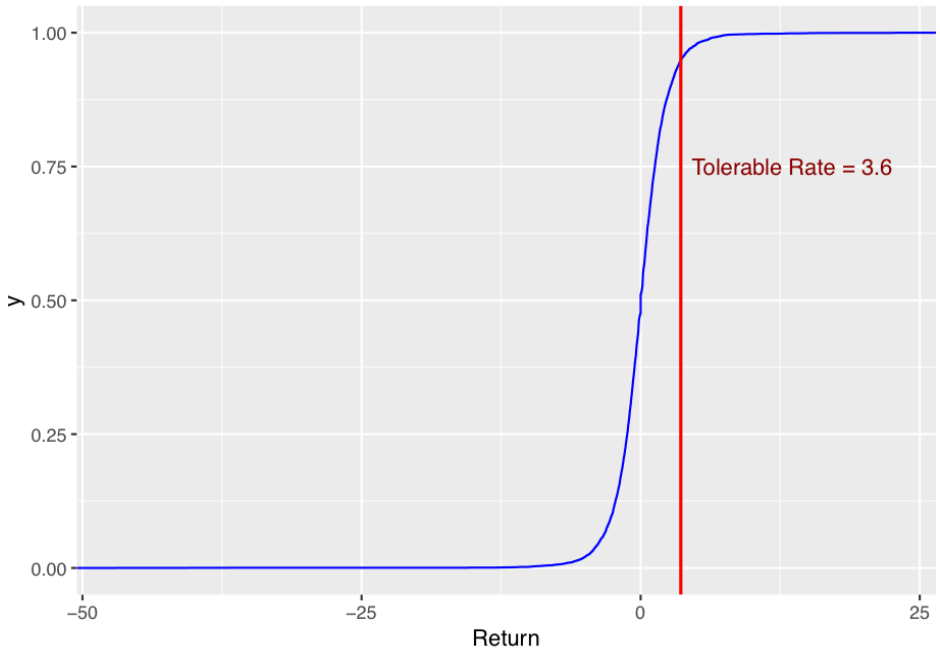

Figure 5: Total Returns Cumulative Distribution

increase until all the negative values have been accounted for at 0. Because the mean of the entire distribution is near 0, analysis of the positive returns should be conducted starting at -inf and working towards 0, with the cumulative distribution being 1 - y. With this in mind, positive returns begin to separate from 0 around 14% with the slope beginning to increase rapidly around 5%.

One key difference in the nature of the positive returns versus the negative returns is that elbow in the negative returns is more gradual than that of the positive returns. This behavior is consistent the fact that there was a greater standard deviation for the negative returns than positive returns.

Additionally, Figure 5 is a cumulative distribution function for all returns. This plot is more consistent with what is expected for a fairly symmetrical plot that is centered around 0. The tolerable rate is also provided to give greater context to the 95% quantile. It is also noteworthy that the leftmost x-limit begins around -50 compared to the rightmost x-limit of 25. This is because the extreme events for negative returns were greater in magnitude than the positive returns.

Using the plots and our tolerable rate of returns, we can look at recent data (due to inflation) and find the mean price of heating oil for that period. If the current price is greater than 3.6% of the mean price, we should attempt to wait for a decrease in price within 3.6% of the mean, unless there is a dire need for heating oil.

Question 2. Repeatable Data Analysis

To reduce the amount to time spent coding for future analysis of other variable costs (or another HO2 vendor), we can create a function that can produce the same data frame, pivot table, and formatted pivot table given that we have data formatted in the same manner that the New York Harbors price data was formatted. The R skills that are required in this block include:

- Reading a file
- Writing functions
- Creating vectors
- Mathematical operations on vectors
- Vector indexing
- Date formatting
- Data frame creating
- Pivot tables with dplyr
- Determining quantiles
- Creating LaTeX tables with xtable
- Creating lists

```r
#First we are going to create a function that will automatically create the pivot table, xtable
#and data frame for future analysis. This function has default arguments of the NYH file and
#default caption of Heating oil No.2: 1986- 2016
HO2_movement <- function(file = "data/nyhh02.csv",
                         caption = "Heating Oil No. 2: 1986-2016"){
    # The file is extracted from the manhattan site. It is also pasted to create a single
    # string
    file <- paste("https://turing.manhattan.edu/~wfoote01/finalytics/", file, sep = "")
    # Reading the csv file and keeping the strings as string
    HO2 <- read.csv(url(file), header = TRUE, stringsAsFactors = FALSE)
    # Deleting the missing data
    HO2 <- na.omit(HO2)
    # using the log difference as percent change of returns
    return <- as.numeric(diff(log(HO2$DHOILNYH))) * 100
    # Calculating the absolute value of the returns to measure volitility
    size <- abs(return)
    # Creating an empty character vector the length of return
```

```
    direction <- character(length(return))
    # Where percent change is greater than 0, put up in the character vector
    direction[return > 0] <- "up"
    # Where percent change is less than 0, put down in the character vector
    direction[return < 0] <- "down"
    # Where percent change is 0, put same in the character vector
    direction[return == 0] <- "same"
    # Converting the dates that were imported as string into the type date and deleting teh
    # first index so it has same length as teh other columns
    date <- as.Date(HO2$DATE[-1], "%m/%d/%Y")
    # Converting the price to a number and deleting the first index because needs to be the
    # same length of the other columns
    price <- as.numeric(HO2$DHOILNYH[-1])
    # Creating the data frame that has the data, price, return, size (absolute value of
    # return), and  direction of returns
    HO2.df <- data.frame(Date = date, Price = price, Return = return,
                         Size = size, Direction = direction)
    # importing the library to create the pivot tables
    require(dplyr)
    # Creating a grouped table that is grouped by the direction of returns
    pivot.table <- group_by(HO2.df, Direction)
    # Making all columns visible
    options(dplyr.width = Inf)
    # Seeing how many entries are in the table so a percentage can be gathered for each
    # category
    HO2.count <- length(HO2.df$Return)
    # Creating the pivot table that returns the average return, return standard deviation, 5%
    # quantile, 95% quantile, and the proportion of entries for positive,negative, and no          # re
    pivot.table <- summarise(pivot.table, return.avg = mean(Return),
                             return.sd = sd(Return),
                             quantile.5 = quantile(Return, 0.05), quantile.95 =
                              quantile(Return, 0.95),
                             percent = length(Return) / HO2.count * 100)
    require(xtable)
    # Creating a nice LaTeX table of the transposed pivot table using the caption argument of
    # the function
    pivot.xtable <- xtable(t(pivot.table), # transposing the pivot table
                           digits = 2, caption = caption,
                           # Limiting to 2 digits, and wring the
                           #given caption
                           # allignging the columns to the right
                           align = rep("r", 4), table.placement = "V")
    # Creating a list to contain all the woutputs of the function
    output.list <- list(table = pivot.table,
                        xtable = pivot.xtable, df = HO2.df)
    # Returning the output list
    return(output.list)
    }
```

After creating a function, it is important test the function to confirm that is performs how it is supposed to. The following block of R code will run the function and extract the pivot table from the returned list. The pivot table will be displayed with the knitr tool. Writing functions for data analysis can save significant time in data processing and cleaning step.

```r
# We are now going to test to see if the function works, and knit the table to look nice
knitr::kable(HO2_movement(file = "data/nyhh02.csv")$table,
             # knit table and extract the pivot table "table" from the output
    digits = 2) # Set the table to display only two digits
```

Direction	return.avg	return.sd	quantile.5	quantile.95	percent
down	-1.77	1.99	-4.78	-0.19	47.52
same	0.00	0.00	0.00	0.00	3.63
up	1.76	1.75	0.18	4.82	48.86

Question 3. Distribution Modeling

While the returns of heating oil prices are random, they are likely to be random within a particular distribution. Finding the distribution that fits these returns would be helpful in modeling price scenarios of a significant variable cost. In turn, these distributions would help determine a better pricing model for our product.

The following block of R code will import the MASS library for statistical analysis, and attempt to find the distribution of negative returns, positive returns, and all returns. We will use the gamma distribution in an attempt to model both positive and negative returns. In the future, we could also see how the returns would fit a lognormal distribution. We will also attempt to fit the distribution of negative returns with both the Student's t-distribution and the gamma distribution. Finally, we will look try and fit a Student's t-distribution to all of the returns to see what the shape of all returns looks like.

The R skills that are required in this block include:

- Using the fitdistr function to estimate a distribution
- Indexing data based on a factor
- Indexing data frame columns

```r
# Importing the statistical package MASS
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Using our function to convert the csv file into usable data and extracting the data frame from the li
HO2.data <- HO2_movement(caption = "HO2 NYH")$df
# Looking at the structure of the data frame
# Date is as typr date, Price, Return, and Size are type numeric, and the Direction is a factor, "up",
str(HO2.data)
```

```
## 'data.frame':    7696 obs. of  5 variables:
##  $ Date     : Date, format: "1986-06-03" "1986-06-04" ...
##  $ Price    : num  0.393 0.378 0.39 0.385 0.373 0.365 0.389 0.394 0.398 0.379 ...
##  $ Return   : num  -2.26 -3.89 3.13 -1.29 -3.17 ...
##  $ Size     : num  2.26 3.89 3.13 1.29 3.17 ...
##  $ Direction: Factor w/ 3 levels "down","same",..: 1 1 3 1 1 1 3 3 3 1 ...
```

```r
# Now we will model the distribution of returns when the returns are positive
# Using the fitdistr function in the MASS package, we will retrieve the estimated shape and scale for f
fit.gamma.up <- fitdistr(HO2.data[HO2.data$Direction == "up", "Return"], "gamma") # The rows with the d
```

```
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
```

```
# Displaying the shape and rate (inverse scale) of the estimated fitted gamma distribution
fit.gamma.up
```

```
##      shape         rate
##   1.30753665   0.74299635
##  (0.02716171) (0.01872184)
```

```
# lognormal fit for positive returns
fit.lnorm.up <- fitdistr(HO2.data[HO2.data$Direction == "up", "Return"], "lognormal")
# lognormal fit for negative returns
fit.lnorm.down <- fitdistr(-HO2.data[HO2.data$Direction == "down", "Return"], "lognormal")
# Not going to model the returns that are the same because all have a value of 0
# Model the negative returns with the t-distribution
# Using fitdistr from MASS, filtering instances that only have the "down" direction
fit.t.down <- fitdistr(HO2.data[HO2.data$Direction == "down", "Return"], "t", hessian = T)
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in log(s): NaNs produced
```

```
## Warning in log(s): NaNs produced
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in log(s): NaNs produced
```

```
## Warning in log(s): NaNs produced
```

```
## Warning in log(s): NaNs produced
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
```

```
## Warning in log(s): NaNs produced
```

```
# Looking at the mean, standard deviation, and degrees of freedom in the t-distribution of negative ret
fit.t.down
```

```
##         m             s            df
##   -1.30565487    0.91307703    2.50894659
##  ( 0.02170850) ( 0.02061868) ( 0.12442996)
```

```
# Fitting a gamma distribution to the negative returns
# Must make the negative returns positive by multiplying by -1 or use the size column of direction == '
fit.gamma.down <- fitdistr(-HO2.data[HO2.data$Direction == "down", "Return"], "gamma", hessian = TRUE)
# Looking at the shape and the rate (inverse scale) of the distribution for negative returns
fit.gamma.down
```

```
##      shape         rate
##   1.31056202   0.73969342
##  (0.02761041) (0.01889467)
```

```
# I want to fit a t-distribution for all returns
t.dist.all <- fitdistr(HO2.data$Return, "t", hessian = T)
```

```
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
## Warning in log(s): NaNs produced
## Warning in dt((x - m)/s, df, log = TRUE): NaNs produced
## Warning in log(s): NaNs produced
## Warning in log(s): NaNs produced
```

```
t.dist.all
```

```
##        m             s            df
##   0.02423709   1.66640769   3.72264703
##  (0.02265778) (0.02407289) (0.16718933)
```

The first thing that is noticeable about the gamma distributions is that the gamma distribution for positive returns is nearly identical to the gamma distribution of negative returns. This would mean that the analysis on positive returns could also be true of the negative returns. Using these distribution parameters, we could conduct some monte carlo simulations to find out if the prices were to increase by a significant amount, how bad could it be. I we decided to run the analysis with a lognormal distribution instead, we could account for even more extreme events.

The t distribution that was fit for negative returns can help model the fat end of the gamma distribution to provide insight of what we can expect for a typical day of negative returns. Finally, the fit for the t-distribution for all return data can be used to project the returns for a typical day. This is because we figure that returns are somewhat normally distributed around the mean on a given day.