data_preprocessing

Data Preprocessing

```
library(dplyr)
library(knitr)
library(tidyverse)
library(RSQLite)
library(jsonlite)
library(quantmod)
library(PerformanceAnalytics)
```

Reading In All Relevant Datasets

```
# reading in treasury rates from 2010-2020
treasury <- read.csv("treasury_rate_2010-2020.csv")
head(treasury)
```

```
##        Date X1.Mo X2.Mo X3.Mo X6.Mo X1.Yr X2.Yr X3.Yr X5.Yr X7.Yr X10.Yr X20.Yr
## 1  1/4/2010  0.05   N/A  0.08  0.18  0.45  1.09  1.66  2.65  3.36   3.85    4.6
## 2  1/5/2010  0.03   N/A  0.07  0.17  0.41  1.01  1.57  2.56  3.28   3.77   4.54
## 3  1/6/2010  0.03   N/A  0.06  0.15   0.4  1.01   1.6   2.6  3.33   3.85   4.63
## 4  1/7/2010  0.02   N/A  0.05  0.16   0.4  1.03  1.62  2.62  3.33   3.85   4.62
## 5  1/8/2010  0.02   N/A  0.05  0.15  0.37  0.96  1.56  2.57  3.31   3.83   4.61
## 6 1/11/2010  0.01   N/A  0.04  0.13  0.35  0.95  1.55  2.58  3.32   3.85   4.64
##   X30.Yr
## 1   4.65
## 2   4.59
## 3    4.7
## 4   4.69
## 5    4.7
## 6   4.74
```

```
# using RSQLite to read in Microsoft option prices data from a db browser
dcon <- dbConnect(SQLite(), dbname = "msft_option.db")
query1 <- paste0("
SELECT *
FROM option_prices;
")

res <- dbSendQuery(conn = dcon, query1)
df2 <- dbFetch(res, -1)
dbClearResult(res)
head(df2)
```

```
##    secid     date symbol symbol_flag   exdate last_date cp_flag strike_price
## 1 107525 20100104 MQF.AB           0 20100116  20091217       C        10000
## 2 107525 20100104 MQF.AC           0 20100116  20100104       C        15000
## 3 107525 20100104 MQF.AD           0 20100116  20100104       C        20000
## 4 107525 20100104 MQF.AE           0 20100116  20090617       C         5000
## 5 107525 20100104 MQF.AS           0 20100116  20091214       C        19000
## 6 107525 20100104 MQF.AU           0 20100116  20100104       C        21000
##   best_bid best_offer volume open_interest impl_volatility     vega optionid
## 1     20.9      21.05      0            66        2.745595 0.071172 45161118
## 2     15.9      16.05     15           455        1.808707 0.103056 45889463
## 3     10.9      11.05    856         15551        1.141079 0.150936 45620071
## 4     25.9      26.05      0             0            <NA>     <NA> 46053521
## 5     11.9      12.05      0            77        1.260992 0.139267 45344313
## 6      9.9      10.05     10           537        1.026391 0.164313 45420619
##   cfadj am_settlement contract_size ss_flag forward_price expiry_indicator root
## 1     1             0           100       0          <NA>             <NA>  MQF
## 2     1             0           100       0          <NA>             <NA>  MQF
## 3     1             0           100       0          <NA>             <NA>  MQF
## 4     1             0           100       0          <NA>             <NA>  MQF
## 5     1             0           100       0          <NA>             <NA>  MQF
## 6     1             0           100       0          <NA>             <NA>  MQF
##   suffix    cusip ticker  sic index_flag exchange_d class issue_type
## 1     AB 59491810   MSFT 7372          0          6  <NA>          0
## 2     AC 59491810   MSFT 7372          0          6  <NA>          0
## 3     AD 59491810   MSFT 7372          0          6  <NA>          0
## 4     AE 59491810   MSFT 7372          0          6  <NA>          0
## 5     AS 59491810   MSFT 7372          0          6  <NA>          0
## 6     AU 59491810   MSFT 7372          0          6  <NA>          0
##   industry_group                issuer div_convention exercise_style
## 1           <NA> MICROSOFT CORPORATION           <NA>              A
## 2           <NA> MICROSOFT CORPORATION           <NA>              A
## 3           <NA> MICROSOFT CORPORATION           <NA>              A
## 4           <NA> MICROSOFT CORPORATION           <NA>              A
## 5           <NA> MICROSOFT CORPORATION           <NA>              A
## 6           <NA> MICROSOFT CORPORATION           <NA>              A
##   am_set_flag
## 1        <NA>
## 2        <NA>
## 3        <NA>
## 4        <NA>
## 5        <NA>
## 6        <NA>
```

```
# using quantmod to get closing prices for Microsoft
getSymbols(Symbols = "MSFT", from = "2010-01-01", to = "2020-01-01")
```

```
## [1] "MSFT"
```

```
daily_closing_prices <- Cl(MSFT)
daily_closing_prices <- as.data.frame(daily_closing_prices)
head(daily_closing_prices)
```

```
##            MSFT.Close
## 2010-01-04      30.95
## 2010-01-05      30.96
## 2010-01-06      30.77
## 2010-01-07      30.45
## 2010-01-08      30.66
## 2010-01-11      30.27
```

```
#write.csv(daily_closing_prices, "C:/Users/robin/Desktop/RStudio/daily_closing_prices.csv")
```

Daily Closing Prices: Estimating Sigma

```
# reading in closing prices .csv file
df <- read.csv("daily_closing_prices.csv")

#df$date <- as.character(df$date)
#df$date <- sub("(.{4})(.{2})(.{2})", "\\1-\\2-\\3", df$date)

colnames(df) <- c("date", "closing_price")
df$date <- as.Date(df$date)
str(df$date)
```

```
##  Date[1:2516], format: "2010-01-04" "2010-01-05" "2010-01-06" "2010-01-07" "2010-01-08" ...
```

```
# assuming that historical volatility from the previous 20 trading days (approximately one tradi
ng month) is representative of the volatility over the life of the option
estimate_sigma <- function (x){
  diff1 <- diff(x)
  denominator <- x[1:nrow(as.data.frame(x))-1]
  sd(diff1 / denominator)
}

rownames(df) <- df[,1]
df[,1] <- NULL

df$sigma_20 <- apply.rolling(df, width=20, FUN="estimate_sigma")
to_cbind <- df$sigma_20
rownames(to_cbind) <- NULL
df <- cbind(df, to_cbind)
colnames(df) <- c("closing_price", "sigma_20_to_erase", "sigma_20")
df[,2] <- NULL

df$date <- rownames(df)

# reordering columns
date_sigma <- df[, c("date", "closing_price", "sigma_20")]
date_sigma$date <- as.Date(date_sigma$date)
tail(date_sigma)
```

```
##                  date closing_price    sigma_20
## 2019-12-23 2019-12-23        157.41 0.006497259
## 2019-12-24 2019-12-24        157.38 0.006470788
## 2019-12-26 2019-12-26        158.67 0.006633161
## 2019-12-27 2019-12-27        158.96 0.006320637
## 2019-12-30 2019-12-30        157.59 0.005909566
## 2019-12-31 2019-12-31        157.70 0.005837897
```

Options Data: Getting Time Differences

```
# changing dates into a date class
df2$date <- as.character(df2$date)
df2$date <- sub("(.{4})(.{2})(.{2})", "\\1-\\2-\\3", df2$date)
df2$date <- as.Date(df2$date)

df2$exdate <- as.character(df2$exdate)
df2$exdate <- sub("(.{4})(.{2})(.{2})", "\\1-\\2-\\3", df2$exdate)
df2$exdate <- as.Date(df2$exdate)

# creating a new column that shows the number of days between the date the option was purchased
 and the expiration date
df2$date_ndiff <- df2$exdate - df2$date
head(df2$date_ndiff)
```

```
## Time differences in days
## [1] 12 12 12 12 12 12
```

Options Data/Treasury: Matching Treasury Yield

```
# process of matching the yield on the US Treasury instrument having maturity closest to the tim
e until expiration of each option, a widely accepted options trading practice
options_df <- df2
options_df$date_ndiff <- as.numeric(options_df$date_ndiff)

# deleting column for 2 months because all its values are NA
treasury[,3] <- NULL
treasury$Date <- as.Date(treasury$Date, format="%m/%d/%Y")
colnames(treasury) <- c("date", "X1mo",
                        "X3mo", "X6mo",
                        "X1yr", "X2yr",
                        "X3yr", "X5yr",
                        "X7yr", "X10yr",
                        "X20yr", "X30yr")

option_df_with_all_tr <- merge(x=options_df, y=treasury, by="date", all.x=TRUE)
head(option_df_with_all_tr)
```

```
##         date  secid symbol symbol_flag     exdate last_date cp_flag
## 1 2010-01-04 107525 MSQ.SG           0 2010-07-17  20100104       P
## 2 2010-01-04 107525 VMF.MW           0 2011-01-22  20100104       P
## 3 2010-01-04 107525 MSQ.SF           0 2010-07-17  20100104       P
## 4 2010-01-04 107525 MSQ.NM           0 2010-02-20  20100104       P
## 5 2010-01-04 107525 MSQ.PA           0 2010-04-17  20100104       P
## 6 2010-01-04 107525 VMF.MJ           0 2011-01-22  20100104       P
##   strike_price best_bid best_offer volume open_interest impl_volatility
## 1        35000     5.00       5.05    147           195        0.242593
## 2        17500     0.39       0.43     20         42392        0.412329
## 3        30000     1.98       2.01     14          6915        0.267571
## 4        34000     3.30       3.35     10           108        0.232962
## 5        26000     0.30       0.32     14          5153        0.296587
## 6        50000    19.30      19.50     10           410        0.258252
##       vega optionid cfadj am_settlement contract_size ss_flag forward_price
## 1  7.20556 46547393     1             0           100       0          <NA>
## 2 3.808464 33753897     1             0           100       0          <NA>
## 3 8.683207 46398372     1             0           100       0          <NA>
## 4 2.270448 46715858     1             0           100       0          <NA>
## 5 3.286701 46855925     1             0           100       0          <NA>
## 6  2.76301 33812784     1             0           100       0          <NA>
##   expiry_indicator root suffix    cusip ticker  sic index_flag exchange_d class
## 1             <NA>  MSQ     SG 59491810   MSFT 7372          0          6  <NA>
## 2             <NA>  VMF     MW 59491810   MSFT 7372          0          6  <NA>
## 3             <NA>  MSQ     SF 59491810   MSFT 7372          0          6  <NA>
## 4             <NA>  MSQ     NM 59491810   MSFT 7372          0          6  <NA>
## 5             <NA>  MSQ     PA 59491810   MSFT 7372          0          6  <NA>
## 6             <NA>  VMF     MJ 59491810   MSFT 7372          0          6  <NA>
##   issue_type industry_group                issuer div_convention exercise_style
## 1          0           <NA> MICROSOFT CORPORATION           <NA>              A
## 2          0           <NA> MICROSOFT CORPORATION           <NA>              A
## 3          0           <NA> MICROSOFT CORPORATION           <NA>              A
## 4          0           <NA> MICROSOFT CORPORATION           <NA>              A
## 5          0           <NA> MICROSOFT CORPORATION           <NA>              A
## 6          0           <NA> MICROSOFT CORPORATION           <NA>              A
##   am_set_flag date_ndiff X1mo X3mo X6mo X1yr X2yr X3yr X5yr X7yr X10yr X20yr
## 1        <NA>        194 0.05 0.08 0.18 0.45 1.09 1.66 2.65 3.36  3.85   4.6
## 2        <NA>        383 0.05 0.08 0.18 0.45 1.09 1.66 2.65 3.36  3.85   4.6
## 3        <NA>        194 0.05 0.08 0.18 0.45 1.09 1.66 2.65 3.36  3.85   4.6
## 4        <NA>         47 0.05 0.08 0.18 0.45 1.09 1.66 2.65 3.36  3.85   4.6
## 5        <NA>        103 0.05 0.08 0.18 0.45 1.09 1.66 2.65 3.36  3.85   4.6
## 6        <NA>        383 0.05 0.08 0.18 0.45 1.09 1.66 2.65 3.36  3.85   4.6
##   X30yr
## 1  4.65
## 2  4.65
## 3  4.65
## 4  4.65
## 5  4.65
## 6  4.65
```

```
option_df_with_all_tr$treasury_rate <- ifelse(option_df_with_all_tr$date_ndiff <= 45, option_df_
with_all_tr$X1mo, ifelse(option_df_with_all_tr$date_ndiff <= 135, option_df_with_all_tr$X3mo, if
else(option_df_with_all_tr$date_ndiff <= 270, option_df_with_all_tr$X6mo, ifelse(option_df_with_
all_tr$date_ndiff <= 547, option_df_with_all_tr$X1yr, ifelse(option_df_with_all_tr$date_ndiff <=
912, option_df_with_all_tr$X2yr, ifelse(option_df_with_all_tr$date_ndiff <= 1460, option_df_with
_all_tr$X3yr, ifelse(option_df_with_all_tr$date_ndiff <= 2190, option_df_with_all_tr$X5yr, ifels
e(option_df_with_all_tr$date_ndiff <= 3102, option_df_with_all_tr$X7yr, ifelse(option_df_with_al
l_tr$date_ndiff <= 3975, option_df_with_all_tr$X10yr, ifelse(option_df_with_all_tr$date_ndiff <=
6625, option_df_with_all_tr$X20yr, ifelse(option_df_with_all_tr$date_ndiff > 6625, option_df_wit
h_all_tr$X30yr, NA))))))))))

# selecting the columns of interest
options_df_pre_final <- option_df_with_all_tr[, c("date", "exdate", "cp_flag",
                                                  "strike_price", "best_bid", "best_offer",
                                                  "volume", "open_interest", "impl_volatility",
                                                  "date_ndiff", "treasury_rate")]
head(options_df_pre_final)
```

```
##         date     exdate cp_flag strike_price best_bid best_offer volume
## 1 2010-01-04 2010-07-17       P        35000     5.00       5.05    147
## 2 2010-01-04 2011-01-22       P        17500     0.39       0.43     20
## 3 2010-01-04 2010-07-17       P        30000     1.98       2.01     14
## 4 2010-01-04 2010-02-20       P        34000     3.30       3.35     10
## 5 2010-01-04 2010-04-17       P        26000     0.30       0.32     14
## 6 2010-01-04 2011-01-22       P        50000    19.30      19.50     10
##   open_interest impl_volatility date_ndiff treasury_rate
## 1           195        0.242593        194          0.18
## 2         42392        0.412329        383          0.45
## 3          6915        0.267571        194          0.18
## 4           108        0.232962         47          0.08
## 5          5153        0.296587        103          0.08
## 6           410        0.258252        383          0.45
```

```
# merging closing prices and sigma_20 on date
options_df_final <- merge(x=options_df_pre_final, y=date_sigma, by="date", all.x = TRUE)
head(options_df_final)
```

```
##         date     exdate cp_flag strike_price best_bid best_offer volume
## 1 2010-01-04 2010-07-17       P        35000     5.00       5.05    147
## 2 2010-01-04 2011-01-22       P        17500     0.39       0.43     20
## 3 2010-01-04 2010-07-17       P        30000     1.98       2.01     14
## 4 2010-01-04 2010-02-20       P        34000     3.30       3.35     10
## 5 2010-01-04 2010-04-17       P        26000     0.30       0.32     14
## 6 2010-01-04 2011-01-22       P        50000    19.30      19.50     10
##   open_interest impl_volatility date_ndiff treasury_rate closing_price sigma_20
## 1           195        0.242593        194          0.18         30.95       NA
## 2         42392        0.412329        383          0.45         30.95       NA
## 3          6915        0.267571        194          0.18         30.95       NA
## 4           108        0.232962         47          0.08         30.95       NA
## 5          5153        0.296587        103          0.08         30.95       NA
## 6           410        0.258252        383          0.45         30.95       NA
```

```
nrow(options_df_final)
```

```
## [1] 1770527
```

```
# deleting all rows with incomplete data
options_df_final <- na.omit(options_df_final)

nrow(options_df_final)
```

```
## [1] 1489016
```

```
#write.csv(options_df_final, "C:/Users/robin/Desktop/RStudio/msft_final_df2.csv")
```