QUESTION1

a)

```
library(readr)
SPY <- read_csv("C:/Users/Joey Zhao/Desktop/SPY.csv")
View(SPY)
library(readr)
F_ <- read_csv("C:/Users/Joey Zhao/Desktop/F .csv")
View(F_)
head(df)
```

```
# A tibble: 6 x 7
       Date  Open  High   Low Close `Adj Close`   Volume
     <date> <dbl> <dbl> <dbl> <dbl>       <dbl>    <int>
1 2015-01-02 15.59 15.65 15.18 15.36    13.22856 24777900
2 2015-01-05 15.12 15.13 14.69 14.76    12.71181 44079700
3 2015-01-06 14.88 14.90 14.38 14.62    12.59124 32981600
4 2015-01-07 14.78 15.09 14.77 15.04    12.95296 26065300
5 2015-01-08 15.40 15.48 15.23 15.42    13.28023 33943400
6 2015-01-09 15.46 15.47 15.06 15.21    13.09937 23381300
```

```
head(dspy)
```

```
# A tibble: 6 x 7
       Date   Open   High    Low  Close `Adj Close`    Volum
e
     <date>  <dbl>  <dbl>  <dbl>  <dbl>       <dbl>     <in
t>
1 2015-01-02 206.38 206.88 204.18 205.43    194.2749 121465
900
2 2015-01-05 204.17 204.37 201.35 201.72    190.7664 169632
600
3 2015-01-06 202.09 202.72 198.86 199.82    188.9696 209151
400
4 2015-01-07 201.42 202.72 200.88 202.31    191.3244 125346
700
5 2015-01-08 204.01 206.16 203.99 205.90    194.7194 147217
800
6 2015-01-09 206.40 206.42 203.51 204.25    193.1590 150812
300
```

```
tail(df)
```

```
# A tibble: 6 x 7
       Date  Open  High   Low Close `Adj Close`   Volume
```

```
     <date> <dbl> <dbl> <dbl> <dbl>      <dbl>    <int>
1 2016-12-22 12.63 12.64 12.40 12.40   11.88837 27821100
2 2016-12-23 12.43 12.46 12.36 12.46   11.94590 15578200
3 2016-12-27 12.43 12.51 12.36 12.39   11.87879 19467400
4 2016-12-28 12.37 12.45 12.22 12.25   11.74456 26875400
5 2016-12-29 12.25 12.31 12.22 12.23   11.72539 19714400
6 2016-12-30 12.24 12.28 12.08 12.13   11.62952 27405700
```

tail(dspy)

```
# A tibble: 6 x 7
      Date   Open   High    Low  Close `Adj Close`   Volum
e
     <date>  <dbl>  <dbl>  <dbl>  <dbl>      <dbl>    <in
t>
1 2016-12-22 225.60 225.74 224.92 225.38   222.2119  56219
100
2 2016-12-23 225.43 225.72 225.21 225.71   222.5373  36251
400
3 2016-12-27 226.02 226.73 226.00 226.27   223.0894  42672
500
4 2016-12-28 226.57 226.59 224.27 224.40   221.2457  64095
000
5 2016-12-29 224.48 224.89 223.84 224.35   221.1964  47719
500
6 2016-12-30 224.73 224.83 222.73 223.53   220.3879 108998
300
```

b)

```
Fprice=df$`Adj Close`
SPYprice=dspy$`Adj Close`
n=length(Fprice) # also the rows of SPY
Fret=Fprice[2:n]/Fprice[1:(n-1)]
SPYret=SPYprice[2:n]/SPYprice[1:(n-1)]
head(Fret)
```

```
[1] 0.9609374 0.9905150 1.0287276 1.0252660 0.9863814 1.00
06576
```

tail(Fret)

```
[1] 0.9810125 1.0048388 0.9943820 0.9887005 0.9983672 0.99
18235
```

head(SPYret)

```
[1] 0.9819403 0.9905811 1.0124611 1.0177452 0.9919864 0.99
21664
```

tail(SPYret)

```
[1] 0.9982726 1.0014642 1.0024811 0.9917354 0.9997773 0.99
63449
```

c)

```
d1=data.frame(SPYret,Fret)
m1=lm(Fret~SPYret)
m1
summary(m1)
```

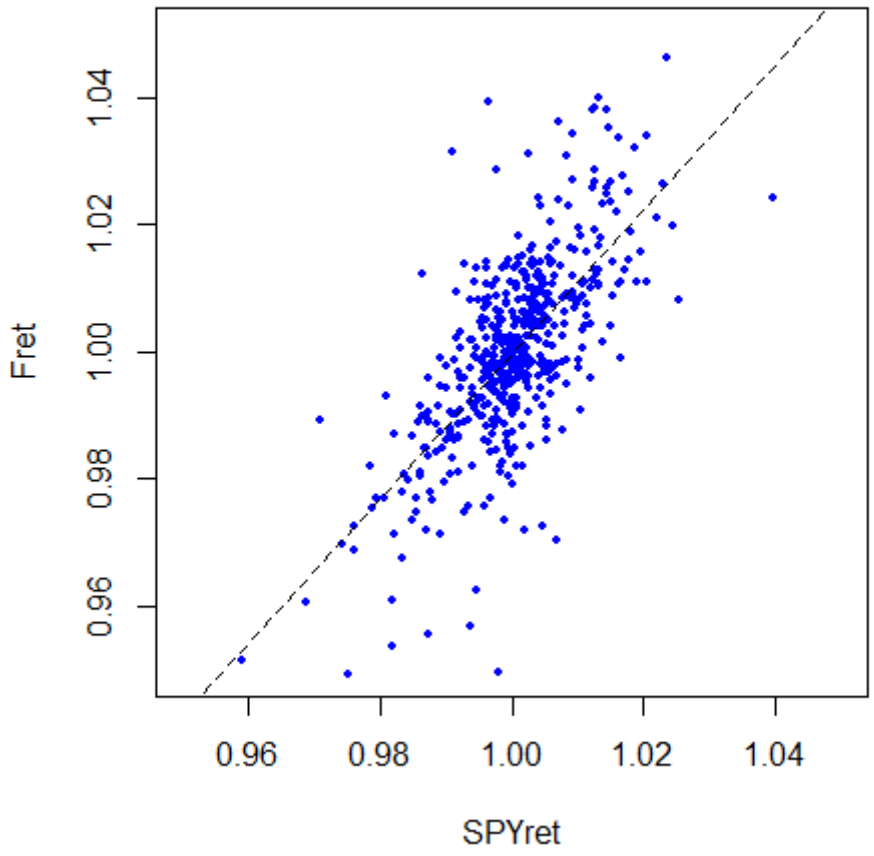
```
Call:
lm(formula = Fret ~ SPYret)

Residuals:
     Min       1Q   Median       3Q      Max
-0.082488 -0.005454  0.000067  0.005830  0.043943

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13631    0.05558  -2.453   0.0145 *
SPYret       1.13584    0.05556  20.444   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.01123 on 501 degrees of freedom
Multiple R-squared:  0.4548,    Adjusted R-squared:  0.453
7
F-statistic:   418 on 1 and 501 DF,  p-value: < 2.2e-16
```

The beta of Ford Motor Co. is 1.13584

Ford's R^2 is 0.4548, which means this model explains 45.48% of the daily returns of Ford variability

d)

```
plot(Fret~SPYret,d1,pch=19,cex=0.6,xlim=c(0.95,1.05),ylim=c(0.95,1.05),col="blue")
abline(m1,lty=2)
```



e)

```
which.max(residuals(m1))
```

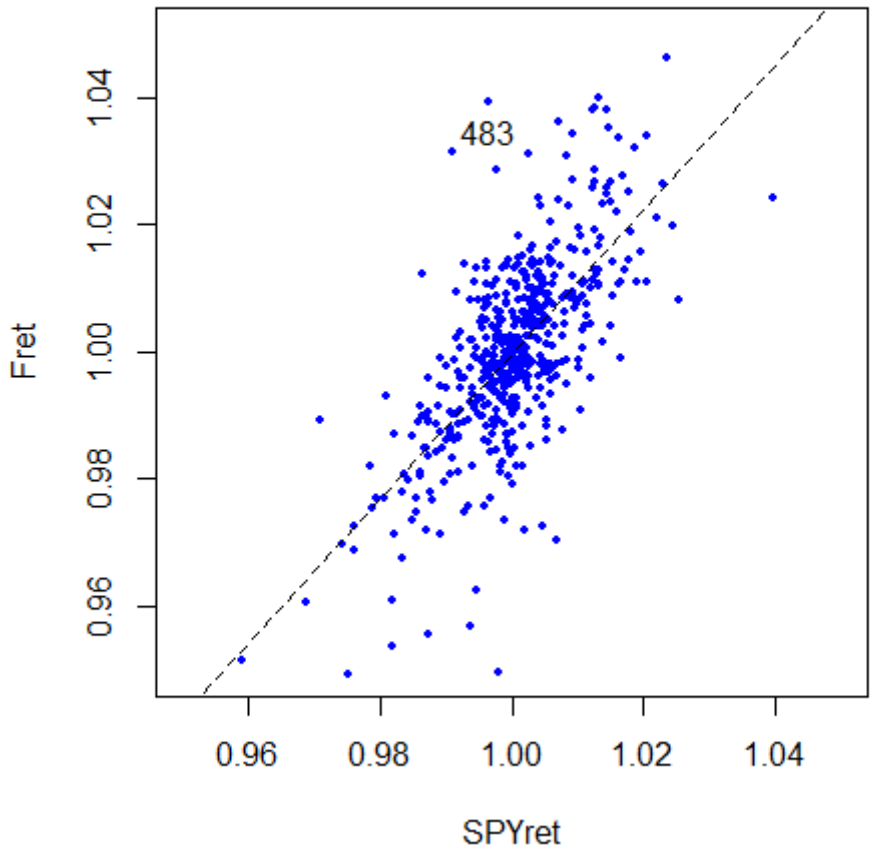
```
483
```

```
identify(Fret~SPYret)
```

or we can use this function

```
text(Fret~SPYret,d1,labels=ifelse(rownames(d1)==483,rownames(d1),""))
```

the result is the same



QUESTION 2

a)

i.

```
insurance <- read_csv("C:/Users/Joey Zhao/Desktop/insurance.csv")
View(insurance)
library(MASS)
d0=insurance
head(d0)
m1=lm(Longevity~.,d0)
m2=stepAIC(m1)
m1
```

```
Call:
lm(formula = Longevity ~ ., data = d0)

Coefficients:
(Intercept)      Mother      Father    Gmothers    Gfathe
rs      Smoker
   23.56735     0.30612     0.30301     0.03161     0.0777
9    -3.71899
```

m2

```
Call:
lm(formula = Longevity ~ Mother + Father + Smoker, data = d
0)

Coefficients:
(Intercept)      Mother      Father      Smoker
   27.2278      0.3344      0.3238     -3.7377
```

new1=data.frame(Mother=75,Father=65,Smoker=1)

predict(m2,new1,interval="conf")

```
       fit      lwr      upr
1 69.61732 68.99821 70.23644
```

ii.

n=nrow(d0)

n/2

set.seed(2)

train=sample(1:n,50)

m1=lm(Longevity~.,d0[train,])

m2=lm(Longevity~Mother+Father+Smoker,d0[train,])

y=d0$Longevity[-train]

yhat1=predict(m1,d0[-train,])

yhat2=predict(m2,d0[-train,])

MSPE1= mean((y-yhat1)^2)

MSPE2= mean((y-yhat2)^2)

sqrt(MSPE1)

sqrt(MSPE2)

```
> sqrt(MSPE1)
[1] 2.587677
> sqrt(MSPE2)
[1] 2.519974
```

The sqrt(MSPE) of m2 < The sqrt(MSPE) of m1

Model m2 is better than m1

QUESIOTN 3

a)

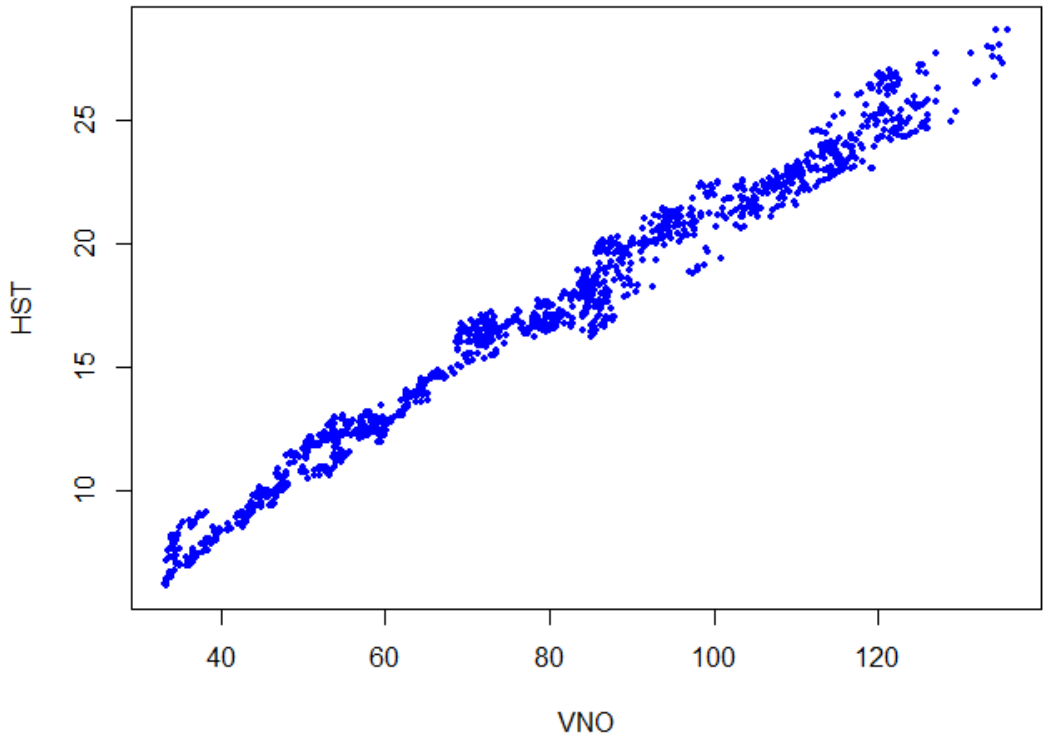
```
dim(prices)
C=cor(prices)
View(C)
dim(C)
which(C==max(C) , arr.ind = T)
for(i in 1:452)
  C[i,i]=0
which(C==max(C),arr.ind = T)
C[428,205]
names[428,]
names[205,]
```

```
> names[428,]
# A tibble: 1 x 3
  Ticker    Sector              Name
   <chr>     <chr>             <chr>
1    VNO Financials Vornado Realty Trust
> names[205,]
# A tibble: 1 x 3
  Ticker    Sector              Name
   <chr>     <chr>             <chr>
1    HST Financials Host Hotels & Resorts
```

The largest correlation is [1] 0.9901256

b)

```
x=prices[,205]
y=prices[,428]
x=as.list(x)
y=as.list(y)
newd=data.frame(y,x)
newd
plot(newd,pch=19,cex=0.6,col="blue")
```



c)

d2=names
d2$Sector
table(d2$Sector)

```
    Consumer Discretionary          Consumer Staples
                        70                        35
                        37                        74
                Health Care               Industrials     Inf
ormation Technology                       Materials
                        46                        59
                        64                        29
Telecommunications Services                 Utilities
                         6                        32
```

There are 46 health care companies in the full dataset

d)

d2f=subset(d2,d2$Sector=="Financials")

```
d2f$Ticker
rows=d2f$`c(1:452)`
rows
colnames(d1)
d1f=d1[,rows]
dim(d1f)
Cf=cor(d1f)
dim(Cf)
which(Cf==max(Cf) , arr.ind = T)
for(i in 1:74)
  Cf[i,i]=0
which(Cf==max(Cf),arr.ind = T)
```

```
    row col
VNO  71  32
HST  32  71
```

```
d2f[71,]
d2f[32,]
```

```
> d2f[71,]
    Ticker    Sector              Name c(1:452)
428    VNO Financials Vornado Realty Trust     428
> d2f[32,]
    Ticker    Sector               Name c(1:452)
205    HST Financials Host Hotels & Resorts     205
```