Homework 2(Part 2)

Question 1: (A and B)

PlantGrowth is a dataset in R that contains crop weights of a control group and two treatment groups.

In [25]:
```
#Code to Get Data

library(datasets)

data(PlantGrowth)

library(tidyverse)
```

In [26]:
```
PlantGrowth_df <- PlantGrowth
```

In [27]:
```
PlantGrowth_df
```

weight	group
4.17	ctrl
5.58	ctrl
5.18	ctrl
6.11	ctrl
4.50	ctrl
4.61	ctrl
5.17	ctrl
4.53	ctrl
5.33	ctrl
5.14	ctrl
4.81	trt1
4.17	trt1
4.41	trt1
3.59	trt1
5.87	trt1
3.83	trt1
6.03	trt1
4.89	trt1
4.32	trt1
4.69	trt1
6.31	trt2
5.12	trt2
5.54	trt2
5.50	trt2
5.37	trt2
5.29	trt2
4.92	trt2
6.15	trt2
5.80	trt2
5.26	trt2

(i) Create two separate datasets, one with datapoints of treatment 1 group along with control group and other with datapoints of treatment 2 group with the control group.

In [28]:
```
PlantGrowth_df1 <- PlantGrowth_df[PlantGrowth_df$group %in% c("trt1", "ctrl"), ]
```

In [44]:
```
PlantGrowth_df1
```

weight	group
4.17	ctrl
5.58	ctrl
5.18	ctrl
6.11	ctrl
4.50	ctrl
4.61	ctrl
5.17	ctrl
4.53	ctrl
5.33	ctrl
5.14	ctrl
4.81	trt1
4.17	trt1
4.41	trt1
3.59	trt1
5.87	trt1
3.83	trt1
6.03	trt1
4.89	trt1
4.32	trt1
4.69	trt1

In [29]:
```
PlantGrowth_df2 <- PlantGrowth_df[PlantGrowth_df$group %in% c("trt2", "ctrl"), ]
```

In [43]:
```
PlantGrowth_df2
```

	weight	group
1	4.17	ctrl
2	5.58	ctrl
3	5.18	ctrl
4	6.11	ctrl
5	4.50	ctrl
6	4.61	ctrl
7	5.17	ctrl
8	4.53	ctrl
9	5.33	ctrl
10	5.14	ctrl
21	6.31	trt2
22	5.12	trt2
23	5.54	trt2
24	5.50	trt2
25	5.37	trt2
26	5.29	trt2
27	4.92	trt2
28	6.15	trt2
29	5.80	trt2
30	5.26	trt2

A)

Now compute the difference estimator for treatment 1 and treatment 2 datasets that were created, in comparison with the control group?

In [30]:
```
reg_all <- lm(weight ~  group, data = PlantGrowth_df1)
summary(reg_all)
```

```
Call:
lm(formula = weight ~ group, data = PlantGrowth_df1)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0710 -0.4938  0.0685  0.2462  1.3690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.0320     0.2202  22.850 9.55e-15 ***
grouptrt1    -0.3710     0.3114  -1.191    0.249    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6964 on 18 degrees of freedom
Multiple R-squared:  0.07308,	Adjusted R-squared:  0.02158 
F-statistic: 1.419 on 1 and 18 DF,  p-value: 0.249
```

For control group average weight in intercept value as we can see above ~5

For treatment 1 group avg weight is ~(5-0.3) = 4.7

The value of "difference estimator" b1 is -0.37. So this is the weight on average that is added to a crop's weights if the group was selected as a treatment 1 to a control, if everything else was constant.

In [31]:
```
reg_all <- lm(weight ~  group, data = PlantGrowth_df2)
summary(reg_all)
```

```
Call:
lm(formula = weight ~ group, data = PlantGrowth_df2)

Residuals:
   Min     1Q Median     3Q    Max 
-0.862 -0.410 -0.006  0.280  1.078 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.0320     0.1637  30.742   <2e-16 ***
grouptrt2     0.4940     0.2315   2.134   0.0469 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5176 on 18 degrees of freedom
Multiple R-squared:  0.2019,	Adjusted R-squared:  0.1576 
F-statistic: 4.554 on 1 and 18 DF,  p-value: 0.04685
```

For control group average weight in intercept value as we can see above ~5

For treatment 2 group avg weight is ~(5+0.49) = 5.49

The value of "difference estimator" b1 is +0.49. So this is the weight on average that is added to a crop's weights if the group was selected as a treatment 2 to a control, if everything else was constant.

In [ ]:

B)

From the PlantGrowth dataset what is the average crop weight of the control group, treatment 1 group, and treatment 2 group, comment on which group has the highest average?

In [37]:
```
PlantGrowth_df_Control <- PlantGrowth_df[PlantGrowth_df$group %in% c("ctrl"), ]
```

In [38]:
```
PlantGrowth_df_trt1 <- PlantGrowth_df[PlantGrowth_df$group %in% c("trt1"), ]
```

In [39]:
```
PlantGrowth_df_trt2 <- PlantGrowth_df[PlantGrowth_df$group %in% c("trt2"), ]
```

In [47]:
```
print(paste("Average crop weight of the control group is: ", mean(PlantGrowth_df_Control$weight)))
print(paste("Average crop weight of the treatment 1 group is: ", mean(PlantGrowth_df_trt1$weight)))
print(paste("Average crop weight of the treatment 2 group is: ", mean(PlantGrowth_df_trt2$weight)))
```

```
[1] "Average crop weight of the control group is:  5.032"
[1] "Average crop weight of the treatment 1 group is:  4.661"
[1] "Average crop weight of the treatment 2 group is:  5.526"
```

Treatment 2 Group has the highest average crop weight.

In [ ]:

Question 1: (C, D and E)

The Minimum Wage Law protects the right of workers to get a minimum wage. Consider a scenario where the law of minimum wage was changed just in the state of New Jersey (i.e., law has not been changed in other states). We want to use the data from company XYZ to observe the difference in hours worked by full time employees in New Jersey before and after the law was changed.

Note: The variable 'State' indicates the citizenship of the worker, i.e., State = "New Jersey" indicates the worker is from NJ else the worker is not from NJ (is from Philadelphia).

Note: The variable fte contains the number of hours worked by an employee.

Note: The variable d indicates whether or not the data was collected before or after the law changed, i.e. d = 1 indicates the data was collected after the law was changed, and d = 0 indicates the data was collected before the law was changed.

In [50]:
```
library("readxl")
```

In [52]:
```
Min_wage <- read_csv("Min_Wage.xls")
```

```
Parsed with column specification:
cols(
  d = col_double(),
  d_nj = col_double(),
  fte = col_double(),
  bk = col_double(),
  kfc = col_double(),
  roys = col_double(),
  wendys = col_double(),
  co_owned = col_double(),
  centralj = col_double(),
  southj = col_double(),
  pa1 = col_double(),
  pa2 = col_double(),
  demp = col_double(),
  State = col_character()
)
```

In [54]:
```
head(Min_wage)
dim(Min_wage)
```

d	d_nj	fte	bk	kfc	roys	wendys	co_owned	centralj	southj	pa1	pa2	demp	State
0	0	15.00	1	0	0	0	0	1	0	0	0	12.00	New Jersey
0	0	15.00	1	0	0	0	0	1	0	0	0	6.50	New Jersey
0	0	24.00	0	0	1	0	0	1	0	0	0	-1.00	New Jersey
0	0	19.25	0	0	1	0	1	0	0	0	0	2.25	New Jersey
0	0	21.50	1	0	0	0	0	0	0	0	0	13.00	New Jersey
0	0	9.50	0	1	0	0	0	0	0	0	0	1.00	New Jersey

768 14

In [55]:
```
table(Min_wage$d) # To understand the split of data pre and post the law change
```

```
  0   1 
384 384 
```

In [56]:
```
table(Min_wage$State) # To understand the split of data of the employees
```

```
  New Jersey Philadelphia 
         618          150 
```

C)

In the above problem, classify the workers into four groups and assign the corresponding group with the group title (A,B,C and D) (i.e., control group before change to the group A etc.). where the group titles are as follows:

	Before	After
Control	A	C
Treated	B	D

In [57]:
```
table(Min_wage$State,Min_wage$d)
```

```
                 0   1
  New Jersey   309 309
  Philadelphia  75  75
```

d = 1 indicates the data was collected after the law was changed, and d = 0 indicates the data was collected before the law was changed.

New Jersey is Treated and Philadelphia is Control

So,

A -> 75; B -> 309; C -> 75; D -> 309

In [ ]:

D)

To estimate the difference in difference we need four averages for the above categorized groups i.e., control group before change, control group after change, treatment group before change and treatment group after change. Compute the following

(i) Calculate the mean of the 'fte' variable for each of the four groups in R and print them

(ii) Using these averages estimate the value of the difference in difference

CG_BC <- control group before change

CG_AC <- control group after change

TG_BC <- treatment group before change

TG_AC <- treatment group after change

In [61]:
```
CG_BC <- filter(Min_wage, Min_wage$State == 'Philadelphia' & Min_wage$d == 0)
CG_AC <- filter(Min_wage, Min_wage$State == 'Philadelphia' & Min_wage$d == 1)
TG_BC <- filter(Min_wage, Min_wage$State == 'New Jersey' & Min_wage$d == 0)
TG_AC <- filter(Min_wage, Min_wage$State == 'New Jersey' & Min_wage$d == 1)
```

In [62]:
```
Mean_CG_BC <- mean(CG_BC$fte)
Mean_CG_AC <- mean(CG_AC$fte)
Mean_TG_BC <- mean(TG_BC$fte)
Mean_TG_AC <- mean(TG_AC$fte)
```

In [67]:
```
print(paste("Mean of the ‘fte’ variable for control group before change: ", round(Mean_CG_BC,3)))
print(paste("Mean of the ‘fte’ variable for control group after change: ", round(Mean_CG_AC,3)))
print(paste("Mean of the ‘fte’ variable for treatment group before change: ", round(Mean_TG_BC,3)))
print(paste("Mean of the ‘fte’ variable for treatment group after change: ", round(Mean_TG_AC,3)))
```

```
[1] "Mean of the ‘fte’ variable for control group before change:  23.38"
[1] "Mean of the ‘fte’ variable for control group after change:  21.097"
[1] "Mean of the ‘fte’ variable for treatment group before change:  20.431"
[1] "Mean of the ‘fte’ variable for treatment group after change:  20.897"
```

In [ ]:

The Difference in Difference (D-in-D) estimater is calculated by:

" (Mean_TG_AC - Mean_TG_BC) - (Mean_CG_AC - Mean_CG_BC) " = 2.75

In [68]:
```
(Mean_TG_AC - Mean_TG_BC) - (Mean_CG_AC - Mean_CG_BC)
```

2.75

In [ ]:
```
```

E)

Estimate the DID (Difference in Difference) using regression model.

In [77]:
```
reg_all <- lm(fte ~ State + d, data = Min_wage)
summary(reg_all)
```

```
Call:
lm(formula = fte ~ State + d, data = Min_wage)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.203  -6.699  -1.203   4.415  64.301 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       20.69914    0.51449  40.232   <2e-16 ***
StatePhiladelphia  1.57442    0.86659   1.817   0.0696 .  
d                 -0.07044    0.68710  -0.103   0.9184    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.521 on 765 degrees of freedom
Multiple R-squared:  0.00431,	Adjusted R-squared:  0.001707 
F-statistic: 1.656 on 2 and 765 DF,  p-value: 0.1917
```

This indicates that the number of hours worked by an employee (in the state of New Jersey) is ~20 hrs. Secondly, the number of hours worked by an employee (in the state of Philadelphia) is an addition of 1.6 hrd, that comes to ~21.6 hrs.

So, the people in the State of New Jersey work ~1.6 hrs less than people in the State of Philadelphia.

In [ ]:
```
```

Question 2: (From A to I)

For the following questions, use the dataset Berkshire.csv with the following variables: Berkshire

· Column (1): Date, Calendar Date

· Column (2): BRKret, Berkshire Hathaway’s monthly return

· Column (3): MKT, the return on the aggregate stock market

· Column (4): RF, the risk free rate of return

You may/may not need the following dependencies:

“PerformanceAnalytics” package

“lubridate” package

Round all answers to the nearest hundredth.

In [82]:
```
library(PerformanceAnalytics)
library(xts)
library(lubridate)
```

In [207]:
```
Berkshire_df <- read.csv("Berkshire.xls")
```

In [208]:
```
head(Berkshire_df)
dim(Berkshire_df)
```

Date	BrkRet	MKT	RF
11/30/1976	0.1544	0.0076	0.0040
12/31/1976	0.1465	0.0605	0.0040
1/31/1977	0.0000	-0.0369	0.0036
2/28/1977	0.0000	-0.0160	0.0035
3/31/1977	0.0778	-0.0099	0.0038
4/30/1977	-0.0103	0.0053	0.0038

500 4

In [209]:
```
Berkshire_df$Date<-mdy(Berkshire_df$Date)
```

In [ ]:

A)

Find the standard deviation of Berkshire Hathaway over the sample period

In [210]:
```
print(paste("Standard deviation of Berkshire Hathaway (BrkRet): ", round(sd(Berkshire_df$BrkRet),2)))
```

```
[1] "Standard deviation of Berkshire Hathaway (BrkRet):  0.07"
```

In [ ]:

B)

Find Berkshire Hathaway’s average return over the sample period?

In [211]:
```
print(paste("Average return (MKT): ", round(mean(Berkshire_df$MKT),2)))
```

```
[1] "Average return (MKT):  0.01"
```

In [ ]:

C)

By what percentage per month on average has Berkshire Hathaway outperformed the market?

In [212]:
```
#create an xts dataset
All.dat <- xts(Berkshire_df[,-1],order.by = Berkshire_df$Date,)
```

In [213]:
```
table.Stats(All.dat$BrkRet)
```

	BrkRet
Observations	500.0000
NAs	0.0000
Minimum	-0.2174
Quartile 1	-0.0162
Median	0.0122
Arithmetic Mean	0.0190
Geometric Mean	0.0168
Quartile 3	0.0476
Maximum	0.3548
SE Mean	0.0030
LCL Mean (0.95)	0.0131
UCL Mean (0.95)	0.0249
Variance	0.0046
Stdev	0.0675
Skewness	0.6987
Kurtosis	2.8198

In [214]:
```
table.Stats(All.dat$MKT)
```

	MKT
Observations	500.0000
NAs	0.0000
Minimum	-0.2264
Quartile 1	-0.0156
Median	0.0136
Arithmetic Mean	0.0102
Geometric Mean	0.0092
Quartile 3	0.0389
Maximum	0.1289
SE Mean	0.0019
LCL Mean (0.95)	0.0063
UCL Mean (0.95)	0.0140
Variance	0.0019
Stdev	0.0436
Skewness	-0.7281
Kurtosis	2.3473

In [215]:
```
0.0190 - 0.0102
```

0.0088

So, on average permonth Berkshire Hathaway outperformed the market by 0.0088

In [ ]:
```

```

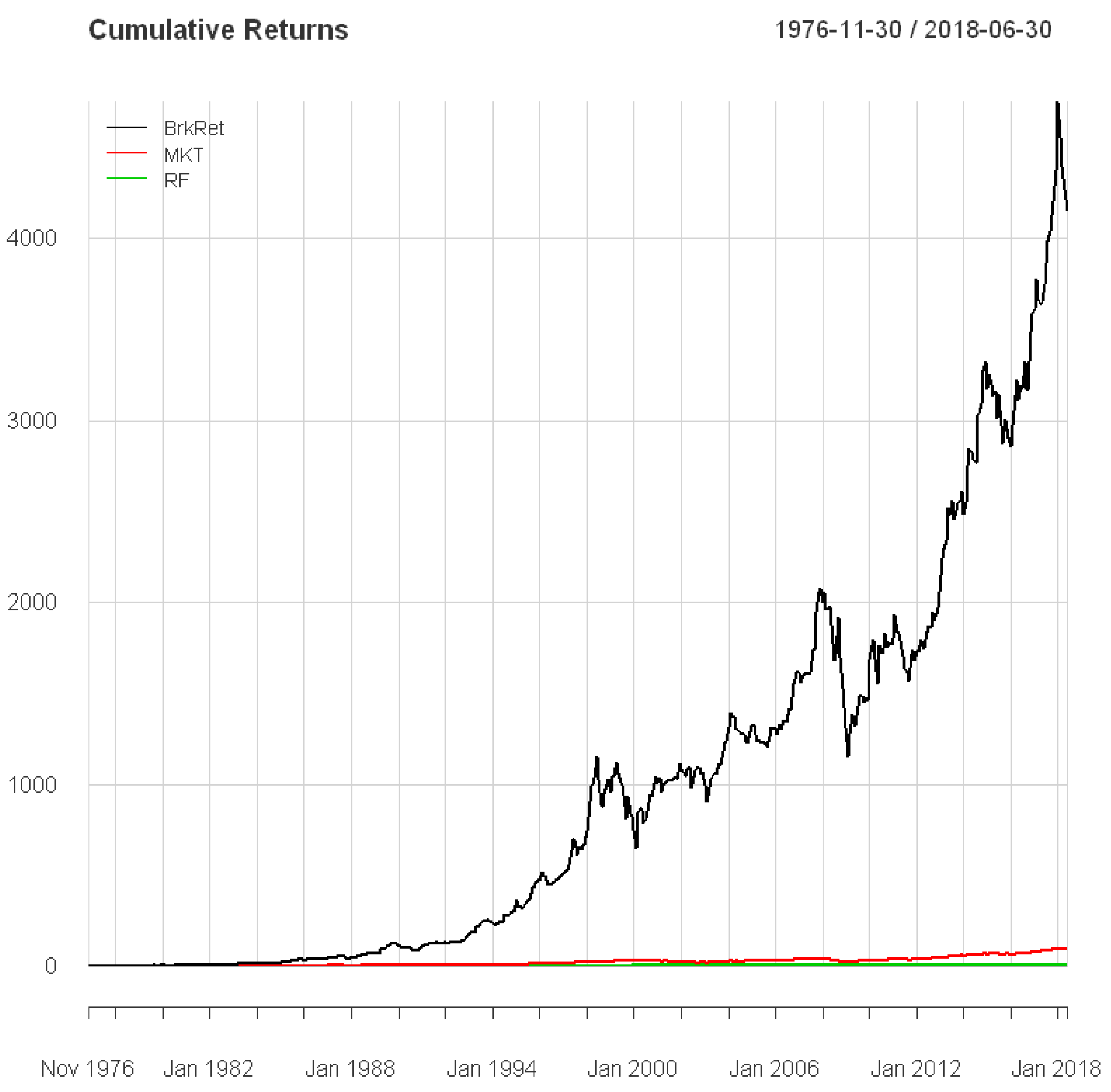
D)

$10,000 invested in Berkshire Hathaway at the start of the sample period would have grown to ____ by the end of the sample period

In [218]:
```
Return.cumulative(All.dat$BrkRet,geometric = TRUE)
```

	BrkRet
Cumulative Return	4143.99

In [219]:
```
chart.CumReturns(All.dat, wealth.index =FALSE, geometric = TRUE, legend.loc = "topleft",main="Cumulat
```



In [166]:
```
Berkshire_df$BrkRet[Berkshire_df$BrkRet == 0.1544] <- 10000
```

In [216]:
```
All.dat<-xts(Berkshire_df[,-1],order.by=Berkshire_df$Date,)
```

In [168]:
```
Return.cumulative(All.dat, geometric =TRUE)
```

	BrkRet	MKT	RF
Cumulative Return	35909598	96.22392	5.359804

So, If 10,000 Dollars invested in Berkshire Hathaway at the start of the sample period would have grown to 35,909,598 Dollars by the end of the sample period.

If we consider the graph and if we consider the inetial value was 0 Dollars and it rose to 4143.99 Dollars, then if we start our investment at 10,000 Dollars then it will end up at approximately 41,439,900 Dollars

In [ ]:

```

```

E)

Plot the cumulative return of Berkshire and Market across all years and include a legend. Describe your observation.

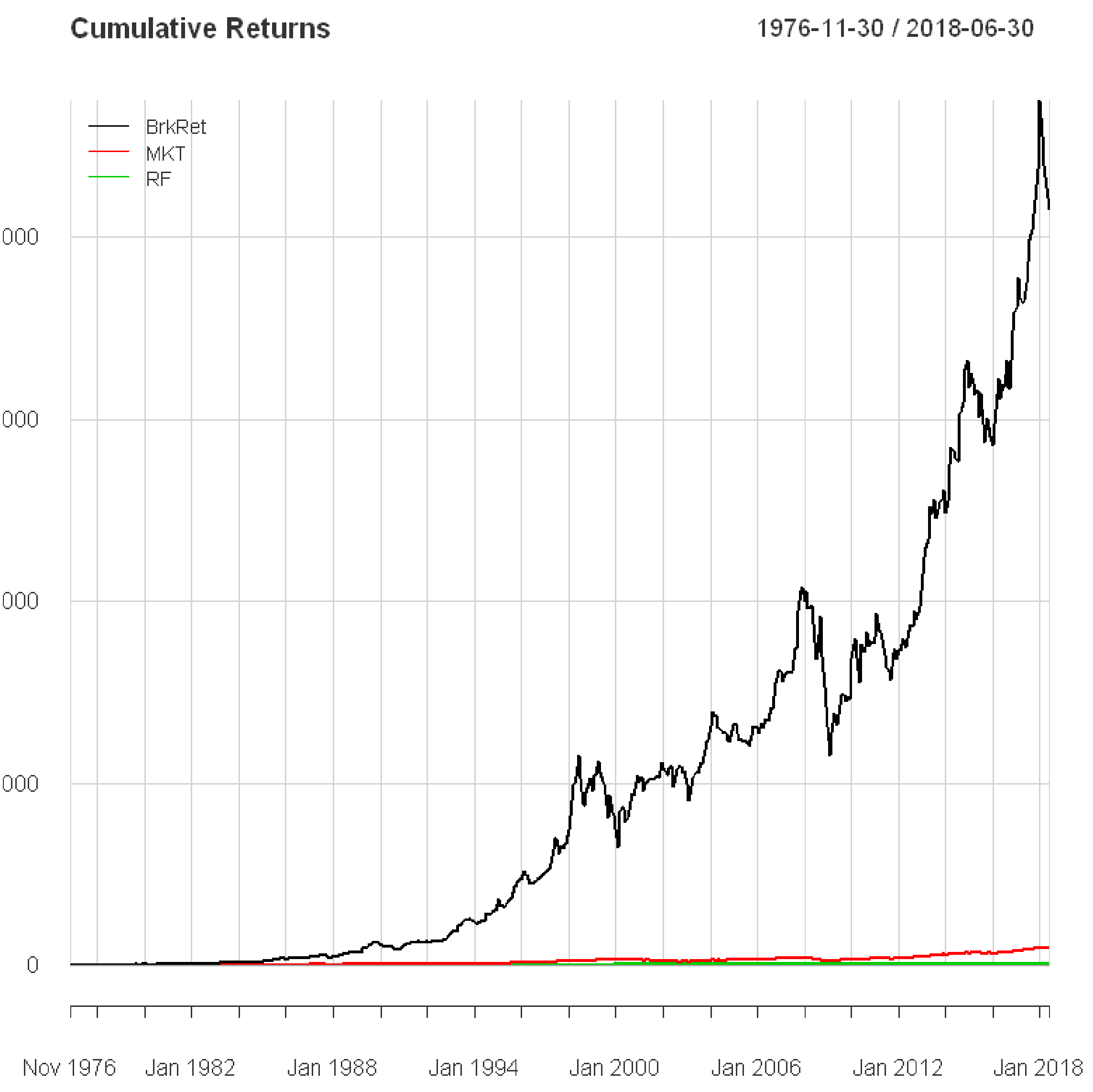
In [221]:

```
# Berkshire_df$Date<-mdy(Berkshire_df$Date)
All.dat<-xts(Berkshire_df[,-1],order.by=Berkshire_df[,1],)
```

In [222]:

```
Return.cumulative(All.dat, geometric =TRUE)
chart.CumReturns(All.dat, wealth.index =FALSE, geometric = TRUE, legend.loc = "topleft",main="Cumulat
```

	BrkRet	MKT	RF
Cumulative Return	4143.99	96.22392	5.359804



If you look at the data supporting this chart, the cumulative return for our fund is 4143.99, the benchmark return is only 96.22392.

And by way of comparison, the risk free rate or the rate of return on a treasury bond is only 5.359804.

So our fund has significantly, significantly outperformed this benchmark.

In [ ]:

```

```

F)

What is Berkshire Hathaway’s monthly Sharpe ratio?

In [223]:
```
SharpeRatio(All.dat$BrkRet,All.dat$RF)
```

	BrkRet
StdDev Sharpe (Rf=0.4%, p=95%):	0.2262115
VaR Sharpe (Rf=0.4%, p=95%):	0.2060944
ES Sharpe (Rf=0.4%, p=95%):	0.1728184

The output shows that the Sharpe ratio for the fund is 0.22. An higher ratio of Sharpe indicates higher reward per unit risk.

G)

What is the Sharpe Ratio for the market index? Comparing this value to Berkshire Hathaway’s Sharpe ratio, which one is higher and what does that mean?

In [224]:
```
SharpeRatio(All.dat$MKT,All.dat$RF)
```

	MKT
StdDev Sharpe (Rf=0.4%, p=95%):	0.14794528
VaR Sharpe (Rf=0.4%, p=95%):	0.09486128
ES Sharpe (Rf=0.4%, p=95%):	0.05677763

Our fund (BrkRet) has had really good performance, its Sharpe ratio of 0.22 is much higher than the Sharpe ratio of the benchmark index (MKT), which has Sharpe ratio of 0.14. It means that our fund has done really good compared to market index.

In [ ]:
```

```

H)

What is Berkshire Hathaway’s estimated beta?

In [230]:
```
TreynorRatio(All.dat$BrkRet,All.dat$MKT,All.dat$RF)
```

0.244858117256507

The Treynor ratio is also a reward to risk ratio, while here the risk is measured relative to beta. And so the interpretation is very similar to Sharpe ratio, a higher Treynor ratio indicates higher reward per unit risk.

In [ ]:
```

```

I)

On a monthly basis, what is Jensen’s alpha for Berkshire Hathaway?

In [203]:
```
All.dat<-transform(All.dat,MktExcess=MKT-RF,FundExcess=BrkRet-RF)
```

In [204]:
```
Alpha=lm(FundExcess~MktExcess,data=All.dat)
summary(Alpha)
```

```
Call:
lm(formula = FundExcess ~ MktExcess, data = All.dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.17263 -0.03475 -0.00688  0.02608  0.33062 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.010829   0.002724   3.976 8.05e-05 ***
MktExcess   0.689755   0.061777  11.165  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06025 on 498 degrees of freedom
Multiple R-squared:  0.2002,	Adjusted R-squared:  0.1986 
F-statistic: 124.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

The Alpha value(Intercept) is +0.0108 and it is statistically significant, the fund has outperformed.

The other thing we can take a look at is R squared or adjusted R squared and that's ~0.20. This tells us a pretty high fraction of this funds return **can not** be explained by the overall market.

In [ ]:
```
```