

report5

Group

August 9, 2018

Probability practice

Part A

$$P(RC) = 0.3$$

$$P(TC) = 1 - 0.3 = 0.7$$

$$P(Yes) = 0.65$$

$$P(No) = 0.35$$

$$P(Yes \text{ and } RC) = P(Yes|RC) \times P(RC) = 0.5 * 0.3 = 0.15$$

$$P(No \text{ and } RC) = P(RC) - P(Yes \text{ and } RC) = 0.3 - 0.15 = 0.15$$

$$P(Yes|TC) = \frac{P(Yes \text{ and } TC)}{P(TC)} = \frac{P(Yes) - P(Yes \text{ and } RC)}{0.7} = \frac{0.65 - 0.15}{0.7} = 0.714$$

Part B

$$P(Pos|D) = 0.993$$

$$P(Neg|D') = 0.9999$$

$$P(Pos|D') = 1 - P(N|D') = 0.0001$$

$$P(D) = 0.000025$$

$$P(D') = 1 - P(D) = 0.999975$$

$$P(D|Pos) = \frac{P(D \text{ and } Pos)}{P(Pos)} = \frac{P(D \text{ and } Pos)}{P(D \text{ and } Pos) + P(D' \text{ and } Pos)} = \frac{P(Pos|D) \times P(D)}{P(Pos|D) \times P(D) + P(Pos|D') \times P(D')} = \frac{0.993 \times 0.000025}{0.993 \times 0.000025 + 0.0001 \times 0.999975}$$

Problems in implementing a universal testing policy:

1. The probability of having the disease when getting positive result is actually not high. The result might cause unnecessary panic;
2. All assumed probabilities need to be very accurate which might be hard to achieve in reality.

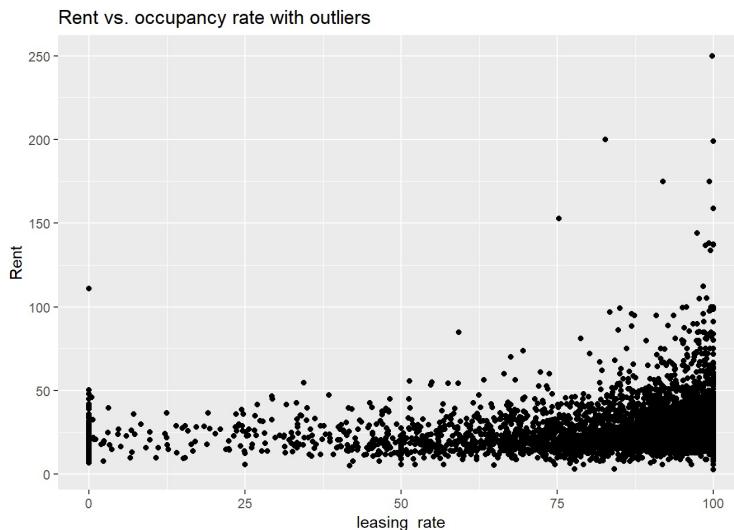
Exploratory analysis: green buildings

```
library(ggplot2)
dat=read.csv("greenbuildings.csv")
```

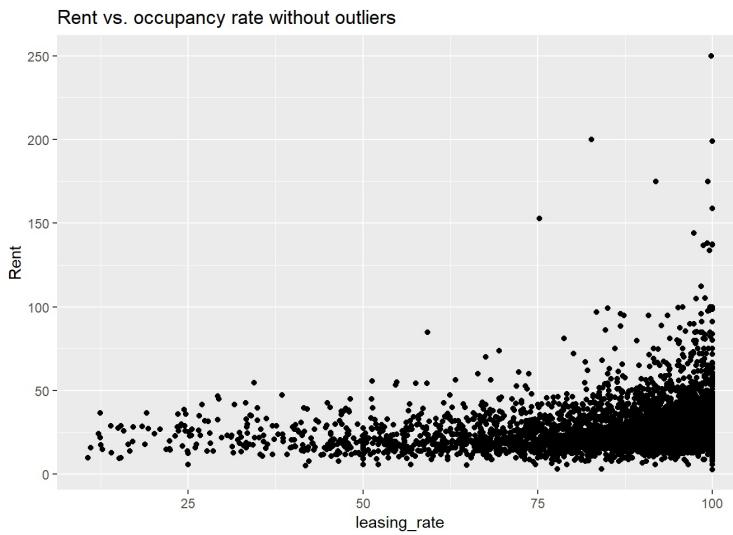
1. Is it reasonable to remove low occupied outliers?

"I decided to remove these buildings from consideration, on the theory that these buildings might have something weird going on with them, and could potentially distort the analysis."

```
par(mfrow=c(1,2))
greenbuilding = dat[,-c(12,13)]
greenbuilding$green_rating = factor(greenbuilding$green_rating)
greenbuilding$cluster = factor(greenbuilding$cluster)
ggplot(greenbuilding, aes(x=leasing_rate, y=Rent)) + geom_point() + ggtitle("Rent vs. occupancy rate with outliers")
```



```
mask = which(greenbuilding$leasing_rate > 10)
greenbuilding = greenbuilding[mask,]
ggplot(greenbuilding, aes(x=leasing_rate, y=Rent)) + geom_point() + ggtitle("Rent vs. occupancy rate without outliers")
```

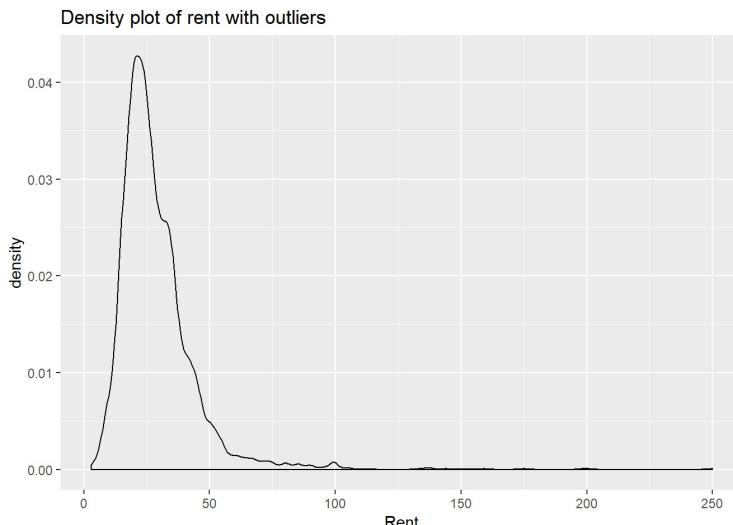


From the scatterplot we can see that there is a specific pattern when x is small. Therefore, it is correct to remove the outliers.

2. Should we use median or mean?

"I used the median rather than the mean, because there were still some outliers in the data, and the median is a lot more robust to outliers."

```
ggplot(greenbuilding,aes(x=Rent))+geom_density() + ggtitle("Density plot of rent with outliers")
```



```
mask2 = which(greenbuilding$Rent > 75)
paste("The number of outliers is",length(mask2))
```

```
## [1] "The number of outliers is 112"
```

```
outliers = greenbuilding[mask2,]
remain = greenbuilding[-mask2,]
```

We decided to use median or mean mainly based on the distribution and size of the dataset. From the first plot, we could see that the number of outliers is not small and the density plot is right-skewed. Therefore, it might be better to use median to measure the central tendency of the dataset. However, without the outliers the distribution of the dataset is quiet normally distributed which might be caused by its large size. So let's try to dig into these outliers to check whether it is reasonable to remove them. We could also see that if we are able to remove them, the distribution is much more normally distributed.

```
library(gridExtra)

paste("The mean of rent of outliers is ",mean(outliers$Rent)," , while the mean of rent of all buildings is ",mean(greenbuilding$Rent))

## [1] "The mean of rent of outliers is 103.209285714286 , while the mean of rent of all buildings is 28.5858
458132569"

paste("The mean of stories of outliers is ",mean(outliers$stories)," , while the mean of stories of all building
s is ",mean(greenbuilding$stories))

## [1] "The mean of stories of outliers is 29.2946428571429 , while the mean of stories of all buildings is 1
3.8299257715848"
```

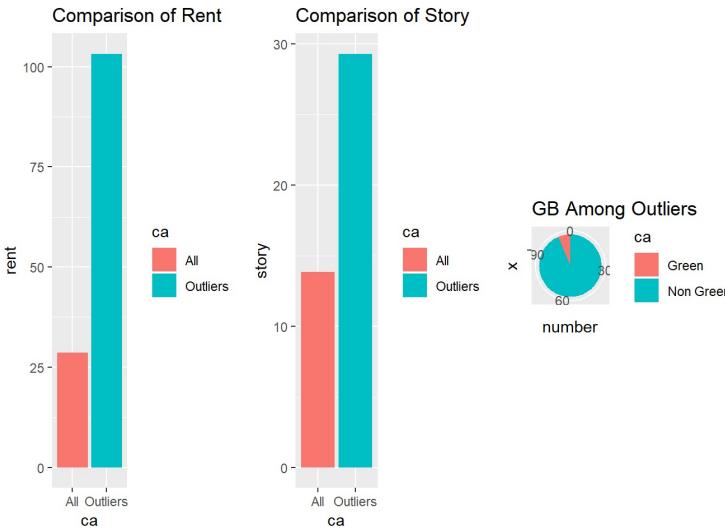
```
paste("The number of green buildings among outliers is ",sum(outliers$green_rating=="1"),", while the number of outliers is ",dim(outliers)[1],". The fraction is ",sum(outliers$green_rating=="1")/dim(outliers)[1])
```

```
## [1] "The number of green buildings among outliers is 7 , while the number of outliers is 112 . The fraction is 0.0625"
```

```
rent_set = data.frame(ca=c("Outliers","All"),
                      rent=c(mean(outliers$Rent),mean(greenbuilding$Rent)))
p1=ggplot(data=rent_set,aes(x=ca,y=rent,fill=ca))+geom_bar(stat="identity")+ggtitle("Comparison of Rent")

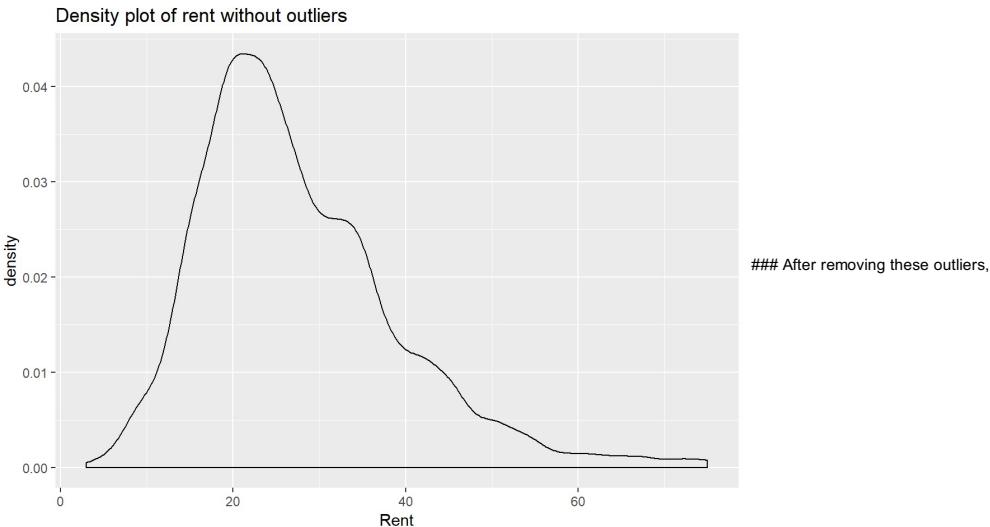
story_set = data.frame(ca=c("Outliers","All"),
                       story=c(mean(outliers$stories),mean(greenbuilding$stories)))
p2=ggplot(data=story_set,aes(x=ca,y=story,fill=ca))+geom_bar(stat="identity")+ggtitle("Comparison of Story")

number_set = data.frame(ca=c("Green","Non Green"),
                        number=c(sum(outliers$green_rating=="1"),dim(outliers)[1]-sum(outliers$green_rating=="1")))
p3=ggplot(number_set,aes(x="",y=number,fill=ca))+geom_bar(width = 1, stat = "identity")+coord_polar("y", start=0)+ggtitle("GB Among Outliers")
grid.arrange(p1,p2,p3, ncol=3)
```



From the graphs, we can see that first, most of these outliers have significantly higher rent than the average level. And then most of them have specifically high stories level which might be the reason of extremely high rent. Considering we are estimating for a 15-story building. These outliers might be less valuable for analysis. In addition, among these outliers, there exist few green buildings, which means we are not able to compare green and non-green building among these outliers. In conclusion, we might not miss insight even if we remove these outliers.

```
ggplot(remain,aes(x=Rent))+geom_density()+ggtitle("Density plot of rent without outliers")
```



the distribution of rent approximates to normal distribution and the data size is still large enough. In this case, we are able to use "mean" to estimate central tendency in later analysis.

3. Should we just compare the mean rent of buildings from all areas?

"The median market rent in the non-green buildings was \$25 per square foot per year, while the median market rent in the green buildings was \$27.60 per square foot per year"

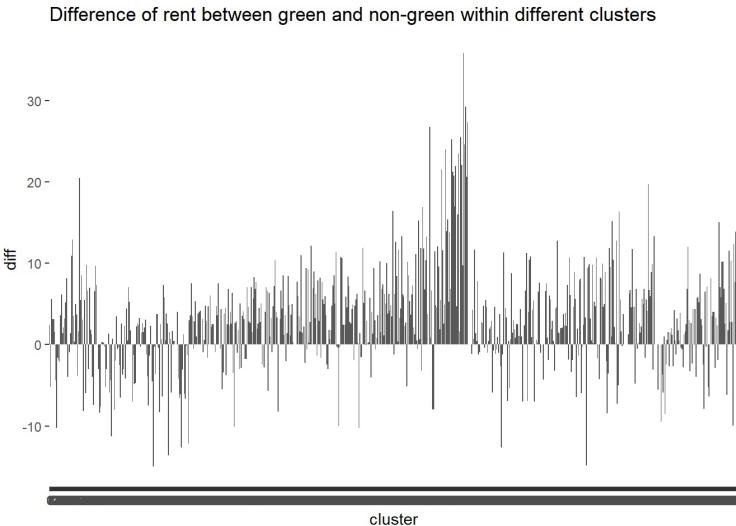
```

library(dplyr)

mean_rent_green = remain %>% group_by(cluster) %>%
  summarise(diff=mean(Rent[green_rating==1])-mean(Rent[green_rating==0]))

ggplot(data=mean_rent_green, aes(x=cluster, y=diff)) +
  geom_bar(stat="identity", position=position_dodge())+ggtitle("Difference of rent between green and non-green w
ithin different clusters")

```



From the result we can see that the difference of median price of green and non-green buildings vary a lot. Therefore, we need to find out which cluster this building belongs to and estimate the price again based on its cluster.

4. Future Estimation

"The median market rent in the non-green buildings was \$25 per square foot per year, while the median market rent in the green buildings was \$27.60 per square foot per year: about \$2.60 more per square foot."

Instead of revenue, we might be more interested in the future profit, which means we should also consider about the cost and the occupancy rate.

```

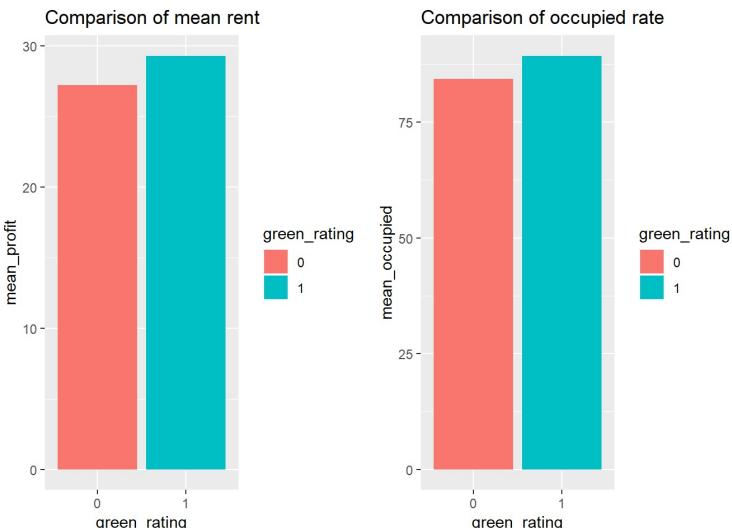
remain["profit"] = remain["Rent"] - remain["Gas_Costs"] - remain["Electricity_Costs"]
remain = remain[,-c(19,20)]

profit_set = remain %>% group_by(green_rating) %>% summarise(mean_profit=mean(profit),mean_occupied=mean(lesasing_rate))
p1=ggplot(data=profit_set,aes(x=green_rating,y=mean_profit,fill=green_rating))+geom_bar(stat="identity")+ggtitle("Comparison of mean rent")
p2=ggplot(data=profit_set,aes(x=green_rating,y=mean_occupied,fill=green_rating))+geom_bar(stat="identity")+ggtitle("Comparison of occupied rate")
profit_set

## # A tibble: 2 × 3
##   green_rating   mean_profit   mean_occupied
##   <dbl>          <dbl>           <dbl>
## 1 0              27.3            84.3
## 2 1              29.3            89.3

grid.arrange(p1,p2, ncol=2)

```



Based on the results above, a normal building might gain an additional

$250000 * (29.3163 * 0.8935 - 27.2548 * 0.8429) = 80506$ dollars of extra revenue per year if we build the green building.

Therefore, we might be able to recuperate the extra \$5 million cost in around 6.2 years. Since the building will be able to earn rents for more than 30 years, "going green" is probably a good investment.

Bootstrapping

Setting up the dataset including these five ETFs from 2007-01-01

```
library(mosaic)
library(quantmod)
library(foreach)

#Import the needed stocks data from 2007-01-01
stocks_list = c("SPY","TLT","LQD","EEM","VNZQ")
getSymbols(stocks_list,from="2007-01-01")
```

```
## [1] "SPY" "TLT" "LQD" "EEM" "VNZQ"
```

```
SPYa = adjustOHLC(SPY)
TLTa = adjustOHLC(TLT)
LQDa = adjustOHLC(LQD)
EEMA = adjustOHLC(EEM)
VNZQa = adjustOHLC(VNZQ)
```

Estimate the risk/return properties of the five major asset classes. Here we just show the process of handling the SPY because the processes of other ETFs are the same. We will just store the result and display them together in the latter part.

```
set.seed("666666")

all_returns = C1C1(SPYa)

all_returns = as.matrix(na.omit(all_returns))

initial_wealth = 100000
simSPY = foreach(i=1:5000, .combine='rbind') %do% {
  total_wealth = initial_wealth
  holdings = total_wealth
  n_days = 20
  wealthtracker = rep(0, n_days)
  for(today in 1:n_days) {
    return.today = resample(all_returns,1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
  }
  wealthtracker
}

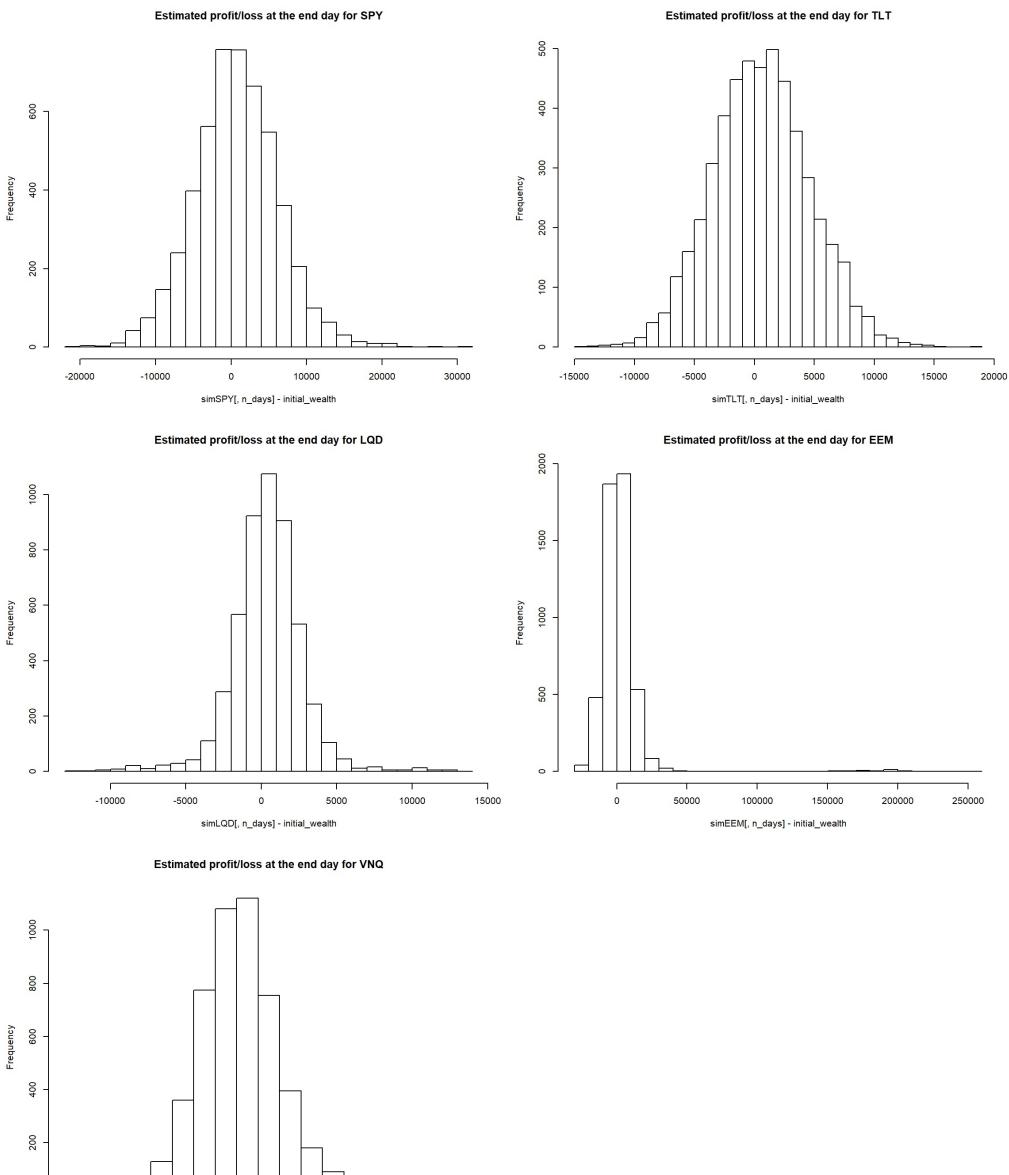
# Calculate 5% value at risk
SPYVar=quantile(simSPY[,n_days], 0.05) - initial_wealth
SPYprofit = quantile(simSPY[,n_days], 0.90) - initial_wealth
SPYmean=mean(simSPY[,n_days])

t <- matrix(c(SPYVar,SPYprofit,TLTVar,TLTprofit,LQDVar,LQDprofit,EEMVar,EEMprofit,VNZQVar,VNZQprofit),ncol=2,byrow=TRUE)
colnames(t) <- c("5% Value at Risk","The top 10% potential profit")
rownames(t) <- c("SPY","TLT","LQD","EEM","VNZQ")
t <- as.table(t)
t

##      5% Value at Risk The top 10% potential profit
## SPY      -8388.289      7527.351
## TLT      -5999.653      5939.240
## LQD      -3028.365      2872.259
## EEM     -13558.980      11992.560
## VNZQ    -13729.160      12246.862

par(mfrow=c(3,2))
hist(simSPY[,n_days]- initial_wealth, breaks=30,main="Estimated profit/loss at the end day for SPY")
hist(simTLT[,n_days]- initial_wealth, breaks=30,main="Estimated profit/loss at the end day for TLT")
hist(simLQD[,n_days]- initial_wealth, breaks=30,main="Estimated profit/loss at the end day for LQD")
hist(simEEM[,n_days]- initial_wealth, breaks=30,main="Estimated profit/loss at the end day for EEM")
hist(simVNZQ[,n_days]- initial_wealth, breaks=30,main="Estimated profit/loss at the end day for VNZQ")

par(mfrow=c(3,3))
```



```

x=1:n_days
plot(x,simSPY[1,],xlab="Time",ylab="Estimated prices",type = "l",ylim=c(80000,160000),main="Estimated 5000 tren
ds for SPY")
for(i in 2:5000){
  lines(x,simSPY[i,],col=i)
}

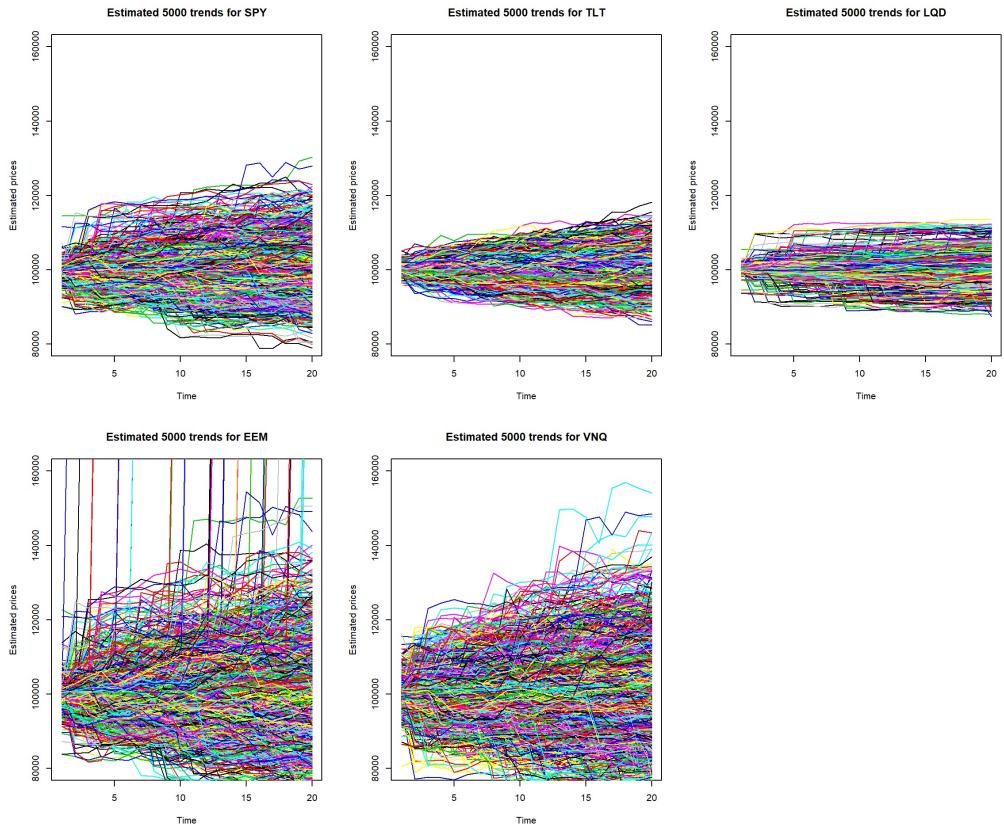
plot(x,simTLT[1,],xlab="Time",ylab="Estimated prices",type = "l",ylim=c(80000,160000),main="Estimated 5000 tren
ds for TLT")
for(i in 2:5000){
  lines(x,simTLT[i,],col=i)
}

plot(x,simLQD[1,],xlab="Time",ylab="Estimated prices",type = "l",ylim=c(80000,160000),main="Estimated 5000 tren
ds for LQD")
for(i in 2:5000){
  lines(x,simLQD[i,],col=i)
}

plot(x,simEEM[1,],xlab="Time",ylab="Estimated prices",type = "l",ylim=c(80000,160000),main="Estimated 5000 tren
ds for EEM")
for(i in 2:5000){
  lines(x,simEEM[i,],col=i)
}

plot(x,simVNO[1,],xlab="Time",ylab="Estimated prices",type = "l",ylim=c(80000,160000),main="Estimated 5000 tren
ds for VNO")
for(i in 2:5000){
  lines(x,simVNO[i,],col=i)
}

```



Observing the distribution of estimated prices and potential trends for these five ETFs, we could conclude that the EEM and VNQ are much more risky than other three ETFs. Among these three, LQD is the safest one with the smalleset variance, followed by TLT. And SPY's variance is the median of these five ETFs.

The even split

```

set.seed("666666")

all_returns = cbind( C1C1(SPYa),
                     C1C1(TLTA),
                     C1C1(LQDa),
                     C1C1(EEMa),
                     C1C1(VNQa) )

all_returns = as.matrix(na.omit(all_returns))

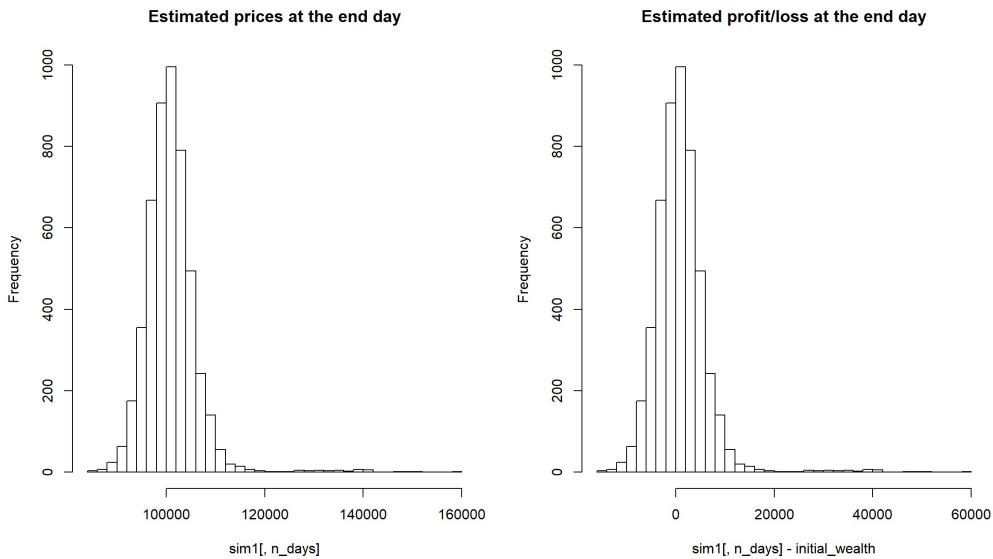
initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  total_wealth = initial_wealth
  weights = c(0.2, 0.2, 0.2, 0.2, 0.2)
  holdings = weights * total_wealth
  n_days = 20
  wealthtracker = rep(0, n_days)
  for(today in 1:n_days) {
    return.today = resample(all_returns, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
  }
  wealthtracker
}

par(mfrow=c(1,2))

hist(sim1[,n_days], 30, main="Estimated prices at the end day")

#profit/loss
hist(sim1[,n_days]- initial_wealth, breaks=30,main="Estimated profit/loss at the end day")

```



```

# Calculate 5% value at risk
evenmean=mean(sim1[,n_days])
evenVar=quantile(sim1[,n_days], 0.05) - initial_wealth
paste("The 5% value at risk is ",evenVar)

```

```
## [1] "The 5% value at risk is -6167.09355643528"
```

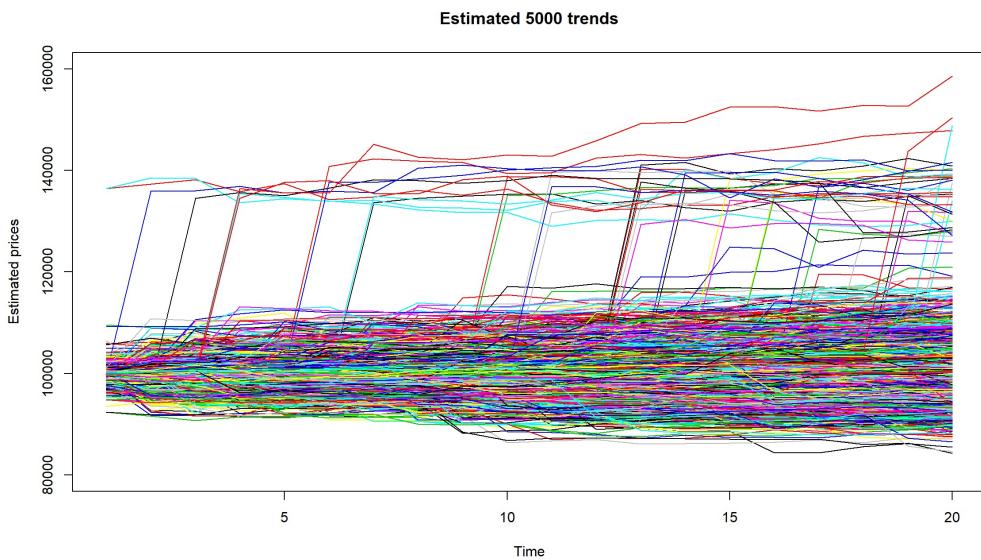
```

evenprofit = quantile(sim1[,n_days], 0.90) - initial_wealth

par(mfrow=c(1,1))

x=1:n_days
plot(x,sim1[,],xlab="Time",ylab="Estimated prices",type = "l",ylim=c(80000,160000),main="Estimated 5000 trends ")
for(i in 2:5000){
  lines(x,sim1[i,],col=i)
}

```



From the graphs we can see that the even portfolio's distribution is similar to the distribution of SPY, the median of these five ETFs. And it has more outliers with high prices, which might be caused by the investment in EEM and VNO.

The safer choice

From the results, we can see that SPY, TLT and LQD, especially LQD, are relatively safer choices. Therefore, we allocate 0.2, 0.3, 0.5 to SPY, TLT and LQD

```
set.seed("666666")

all_returns = cbind( ClCl(SPYa),
                     ClCl(TLTa),
                     ClCl(LQDa))

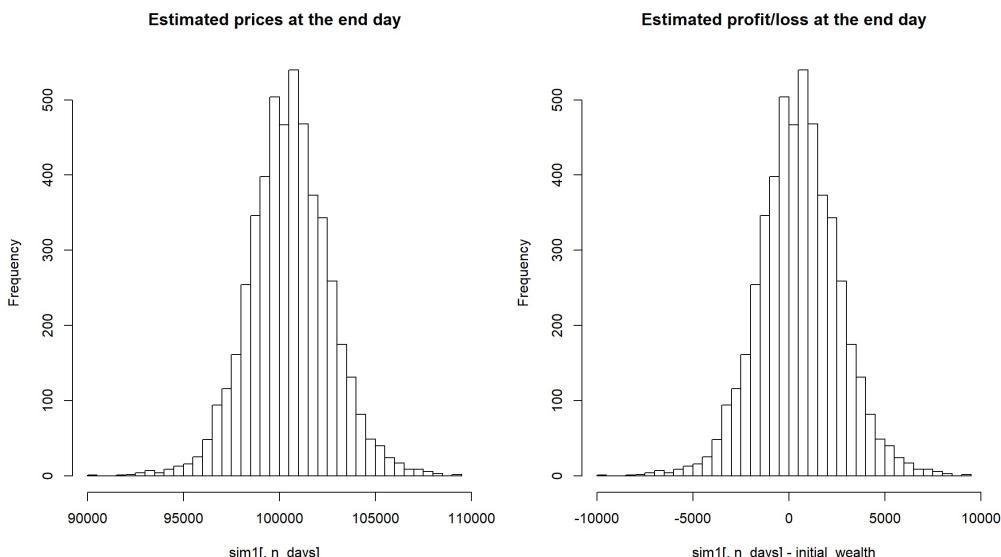
all_returns = as.matrix(na.omit(all_returns))

initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  total_wealth = initial_wealth
  weights = c(0.2, 0.3, 0.5)
  holdings = weights * total_wealth
  n_days = 20
  wealthtracker = rep(0, n_days)
  for(today in 1:n_days) {
    return.today = resample(all_returns, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
  }
  wealthtracker
}

par(mfrow=c(1,2))

hist(sim1[,n_days], 30, main="Estimated prices at the end day")

#profit/loss
hist(sim1[,n_days]- initial_wealth, breaks=30,main="Estimated profit/loss at the end day")
```



```
# Calculate 5% value at risk
safermean=mean(sim1[,n_days])
saferVar=quantile(sim1[,n_days], 0.05) - initial_wealth
paste("The 5% value at risk is ",saferVar)
```

```

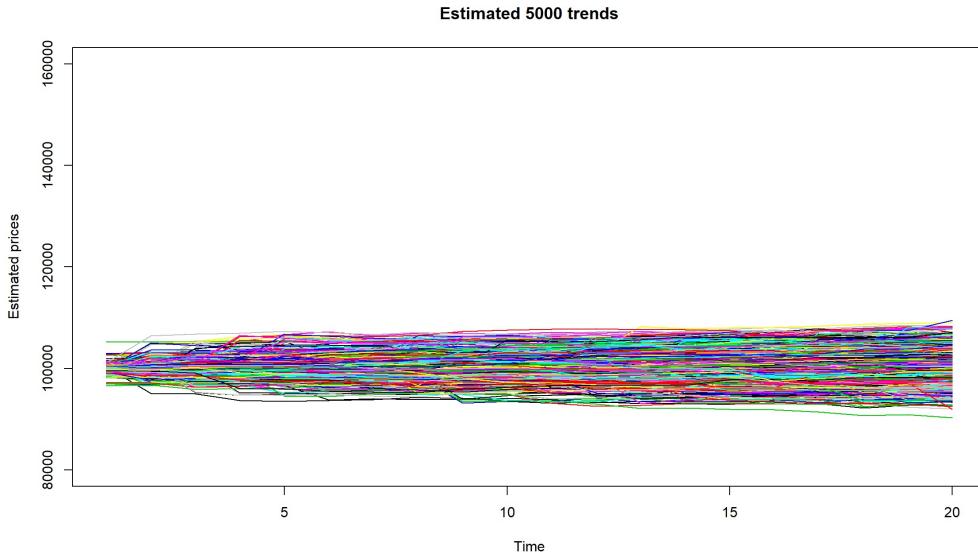
## [1] "The 5% value at risk is -2870.62687328462"

saferprofit = quantile(sim1[,n_days], 0.90) - initial_wealth

par(mfrow=c(1,1))

x=1:n_days
plot(x,sim1[,1],xlab="Time",ylab="Estimated prices",type = "l",ylim=c(80000,160000),main="Estimated 5000 trends")
for(i in 2:5000){
  lines(x,sim1[i,],col=i)
}

```



From the graphs, we can see the the safer portfolio has much smaller variance even than TLT, just a bit larger than LQD. And it has less outliers with low or high prices than LQD. The probability of significant loss is small but the probability of high return is also small.

The more aggressive choice

From the result, we can see that EEM and VNQ are more risky choices, but we still want to keep the SPY to control the risk within acceptable range. Therefore, we allocate 0.4,0.3,0.3 to SPY,EEM and VNQ.

```

set.seed("666666")

all_returns = cbind( C1C1(SPYa),
                     C1C1(EEMa),
                     C1C1(VNQa) )

all_returns = as.matrix(na.omit(all_returns))

initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  total_wealth = initial_wealth
  weights = c(0.4,0.3,0.3)
  holdings = weights * total_wealth
  n_days = 20
  wealthtracker = rep(0, n_days)
  for(today in 1:n_days) {
    return.today = resample(all_returns, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
  }
  wealthtracker
}

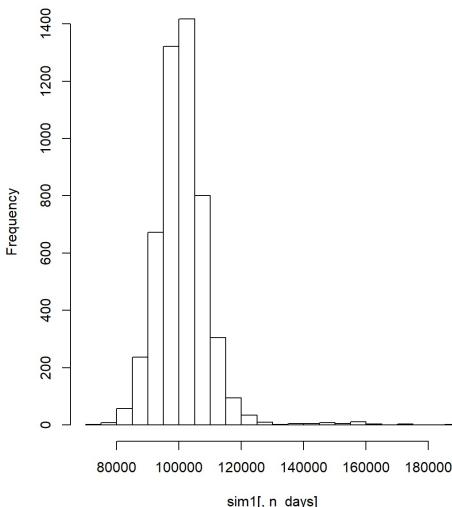
par(mfrow=c(1,2))

hist(sim1[,n_days], 30, main="Estimated prices at the end day")

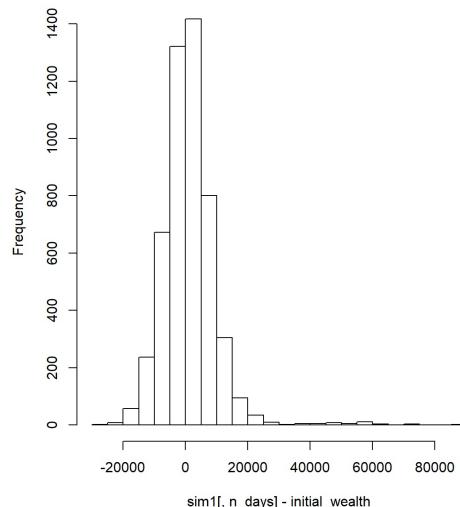
#profit/loss
hist(sim1[,n_days]- initial_wealth, breaks=30,main="Estimated profit/loss at the end day")

```

Estimated prices at the end day



Estimated profit/loss at the end day



```
# Calculate 5% value at risk
aggmean=mean(sim1[,n_days])
aggVar=quantile(sim1[,n_days], 0.05) - initial_wealth
paste("The 5% value at risk is ",aggVar)

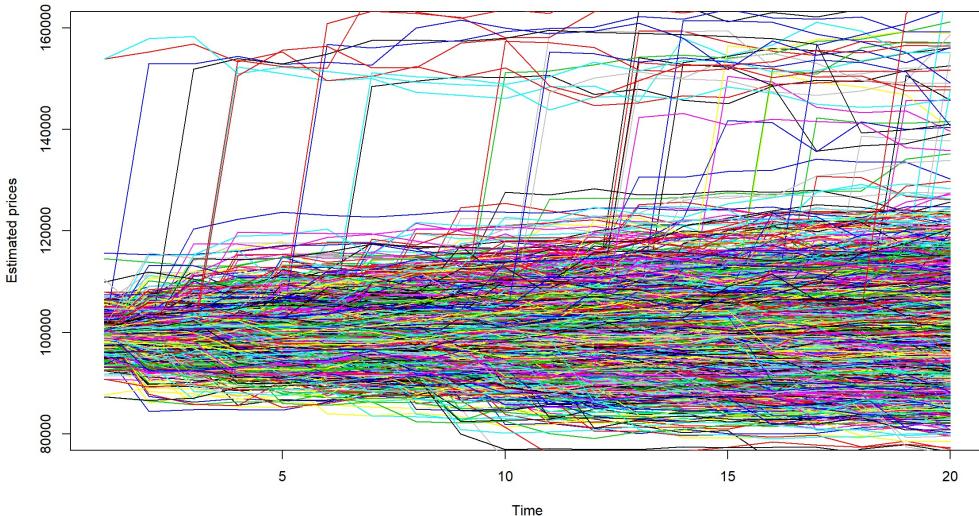
## [1] "The 5% value at risk is -10659.5986819229"

aggprofit = quantile(sim1[,n_days], 0.90) - initial_wealth

par(mfrow=c(1,1))

x=1:n_days
plot(x,sim1[,1],xlab="Time",ylab="Estimated prices",type = "l",ylim=c(80000,160000),main="Estimated 5000 trends")
for(i in 2:5000){
  lines(x,sim1[i,],col=i)
}
```

Estimated 5000 trends



From the graphs, we can see that this portfolio has much larger variance. It has chance to gain significant high return, but it is also more risky than other portfolios.

Comparison of all portfolios for decision making

```
t <- matrix(c(SPYVar,SPYmean,SPYprofit,TLTVar,TLTmean,TLTprofit,LQDVar,LQDmean,LQDprofit,EEMVar,EEMmean,EEMprofit,VNQVar,VNQmean,VNQprofit,evenVar,evenmean,evenprofit,saferVar,safermean,saferprofit,aggVar,aggmean,aggprofit),ncol=3,byrow=TRUE)
colnames(t) <- c("5% Value at Risk","Expected return","The top 10% potential profit")
rownames(t) <- c("SPY","TLT","LQD","EEM","VNQ","Even","Safer","Aggressive")
t <- as.table(t)
t
```

	5% Value at Risk	Expected return	The top 10% potential profit
## SPY	-8388.289	100759.214	7527.351
## TLT	-5999.653	100597.012	5939.240
## LQD	-3028.365	100392.989	2872.259
## EEM	-13558.980	102036.121	11992.560
## VNQ	-13729.160	100750.798	12246.862
## Even	-6167.094	100907.227	6099.977
## Safer	-2870.627	100527.441	3113.661
## Aggressive	-10659.599	101139.761	9825.424