Exercises 1

Danchen Zhao, Shan Qin, Candice Zuo

August 9, 2018

Probability practice

Part A

Let Right be the event that an user is a right clicker.

Let Random be the event that an user is a random clicker.

Let Yes be the event that the user click “yes”.

Let No be the event that the user click “no”.

Solve for P(Yes|Right)

P(Yes|Random) = 0.5

P(No|Random) = 0.5

P(Random) = 0.3

P(Right) = 1 - P(Random) = 1 - 0.3 = 0.7

P(Yes|Right) = x

P(Yes) = 0.65

P(Yes|Random)P(Random) + P(Yes|Right)P(Right) = 0.65

P(Yes|Right) = 5/7

Part B

Let Positive be the event that the test result is positive.

Let Negative be the event that the test result is positive.

Let Disease be the event that a person has this disease.

Let Health be the event that a person does not this disease.

Solve for P(Disease|Positive)

P(Positive|Disease) = 0.993

P(Negative|Disease) = 1 - P(Positive|Disease) = 0.007

P(Negative|Health) = 0.9999

P(Positive|Health) = 0.0001

P(Disease) = 0.000025

P(Health) = 0.999975

P(Disease|Positive) = (P(Positive|Disease)P(Disease))/(P(Positive|Disease)P(Disease) + P(Positive|Health)P(Health))

P(Disease|Positive) = 0.1996783

Exploratory analysis: green buildings

After analysis, we disagree with the guru. The guru failed to consider the possibility that the positive correlation between green rating and rents are brought by confounding variables but not green rating itself. Analyses are as follow.

First, we cleaned the data to drop the building with less than 10% occupancy.

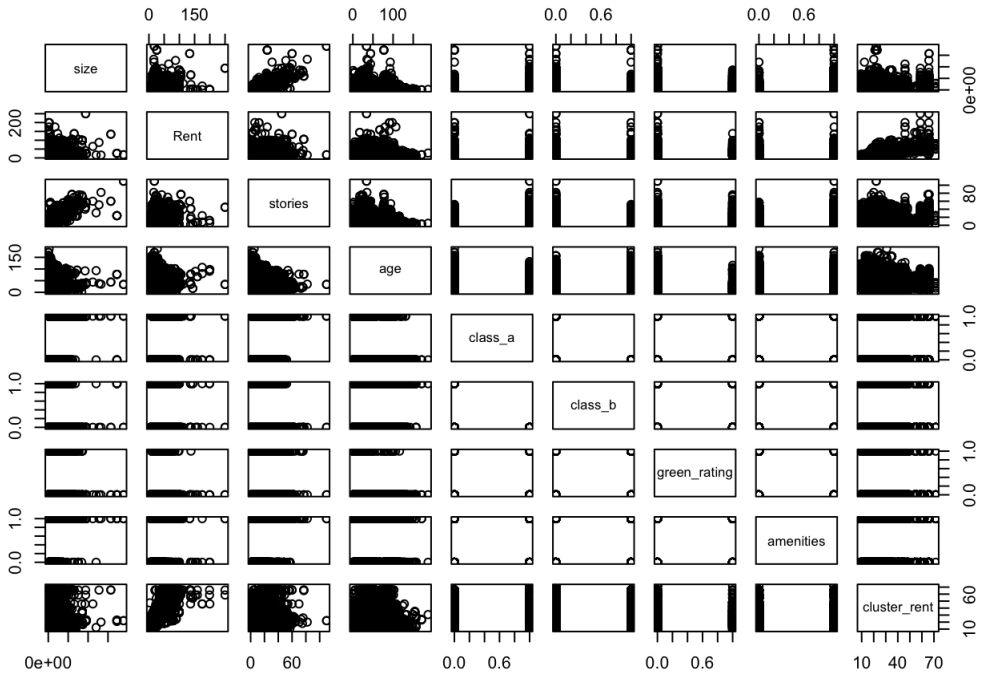
We believe this can help the analysis because these buildings may not used to rent and their prices may not corresponding to the market price.

We also created a new column “class” that contains all three classes to help the further analysis.

After examine all the correlation plots between variables, we find some of them have relative obvious correlation with rent and green_rating.

We build pairs plots for these variables and check their correlations.

```
pairs(green_table_clean_leasing_addclass[, c(3,5,7,8,10,11,14,16,23)])
```




```
cor(green_table_clean_leasing_addclass[, c(3,5,7,8,10,11,14,16,23)])
```

```
##                    size        Rent     stories         age    class_a
## size          1.00000000  0.13325900  0.82575774 -0.20296328  0.4406940
## Rent          0.13325900  1.00000000  0.11109820 -0.10133577  0.2067169
## stories       0.82575774  0.11109820  1.00000000 -0.14528492  0.4492976
## age          -0.20296328 -0.10133577 -0.14528492  1.00000000 -0.5068233
## class_a       0.44069399  0.20671694  0.44929758 -0.50682329  1.0000000
## class_b      -0.27804657 -0.12656366 -0.28828291  0.26963761 -0.7645472
## green_rating  0.09017177  0.02980230  0.03839392 -0.22523781  0.2481154
## amenities     0.38650765  0.05286757  0.37917699 -0.24289767  0.3983044
## cluster_rent -0.02876952  0.75821345 -0.02012161 -0.02923463  0.1063749
##                   class_b green_rating   amenities cluster_rent
## size         -0.27804657   0.09017177  0.38650765  -0.02876952
## Rent         -0.12656366   0.02980230  0.05286757   0.75821345
## stories      -0.28828291   0.03839392  0.37917699  -0.02012161
## age           0.26963761  -0.22523781 -0.24289767  -0.02923463
## class_a      -0.76454715   0.24811536  0.39830438   0.10637487
## class_b       1.00000000  -0.16764157 -0.22734609  -0.07696384
## green_rating -0.16764157   1.00000000  0.11924036  -0.02043843
## amenities    -0.22734609   0.11924036  1.00000000  -0.04885971
## cluster_rent -0.07696384  -0.02043843 -0.04885971   1.00000000
```

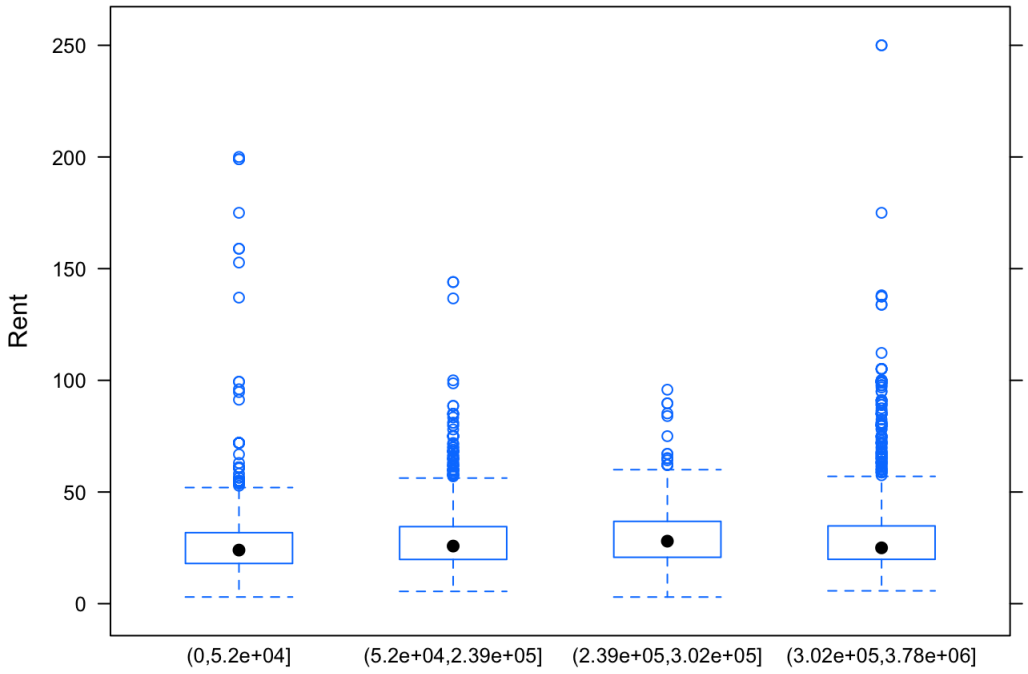
From this summary, we could find that buildings with green rating tend to have more amenities, larger size, more class a and smaller age. Meanwhile, rents have positive correlation with size, amenities and class a, and negative correlation with age. It is possible that confounding variables cause the relative higher rent of the buildings with green rating.

We split the data into green building and non-green building.

```
green_only= subset(green_table_clean_leasing_addclass, green_rating==1)
non_green= subset(green_table_clean_leasing_addclass, green_rating==0)
```

Focus on size, we notice both rent and green building have positive correlation with size, so we divided size into four part to see how rent change in different size range.

```
green_table_clean_leasing_addclass$sizecategory = cut(green_table_clean_leasing_addclass$size, breaks=c(0,52000,239465,302375,3781045))
bwplot(Rent ~ sizecategory, data=green_table_clean_leasing_addclass)
```



Median rent of green and non-green buildings for each size range.

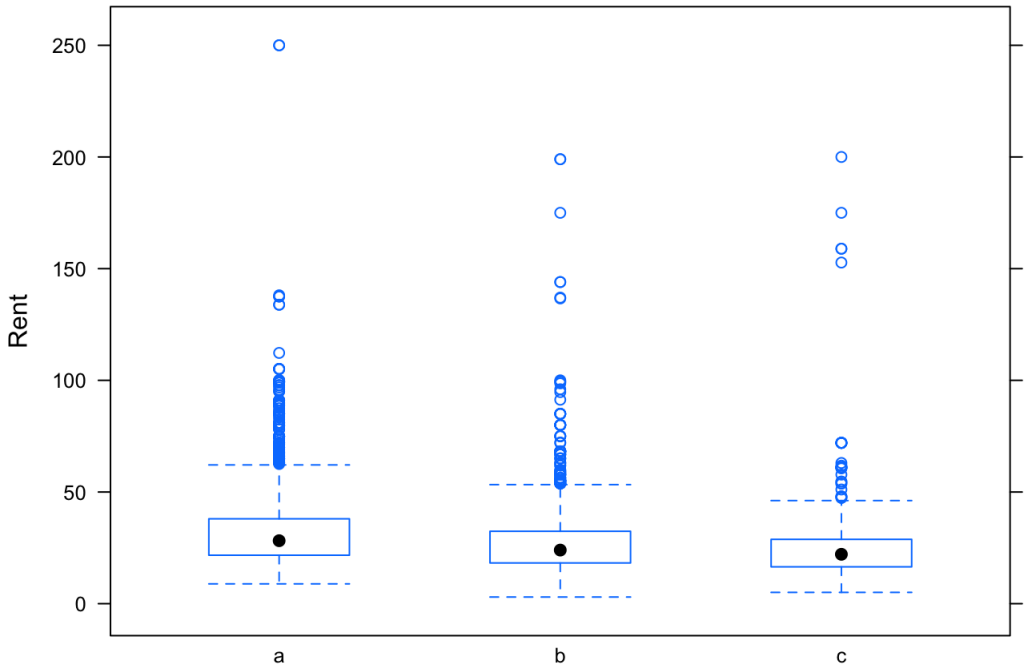
size	Green	Non_Green
00%-25%	27.6	24
25%-50%	28.47	25.5
50%-75%	30	27.95
75%-100%	25.97	24.91

From the data, we find the rent of green building increases relatively bigger when the size is small (<75%).

When the size is large, the increase caused by green building become smaller.

Then we look at class.

```
bwplot(Rent ~ class, data=green_table_clean_leasing_addclass)
```




```
table1=xtabs(~green_rating + class, data=green_table_clean_leasing_addclass)
prop.table(table1, margin=1)
```

```
##              class
## green_rating          a          b          c
##            0 0.37012152 0.48477484 0.14510365
##            1 0.79824561 0.19152047 0.01023392
```

Median rent of green and non-green buildings for each class.

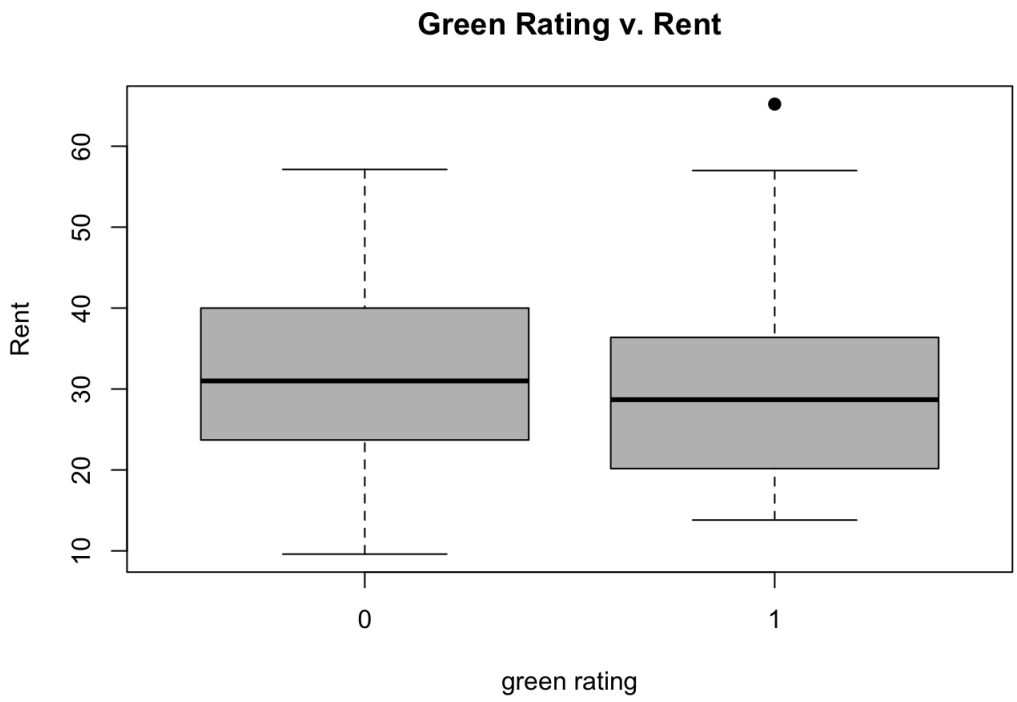
class	Green	Non_Green
a	28.44	28.2
b	25.2	24
c	32	22.11

From the data, we can see green building almost has no rent advantage when the class is a. And from the percentage table, green building has much more class a building comparing to non-green.

There is possibility that the green building has higher median rent because they have more class a building.

To test our theory, we control the variables – age, size, amenities and class, to check the median rent of green and non-green building.

```
green_table_controlled = subset(green_table_clean_leasing_addclass, age < 10 & size > 52000 & size < 302375
& amenities == 1 & class == 'a')
plot(factor(green_table_controlled$green_rating), green_table_controlled$Rent, pch=19, col='grey', xlab='gre
en rating', ylab='Rent', main='Green Rating v. Rent')
```



As a result, the green building actually has a lower rent range under the control.

Therefore, we can say guru's statement is wrong. The guru doesn't consider the effect of other variables on the rent. The higher median rent green building has may be caused by other features shared by most green building.

Bootstrapping

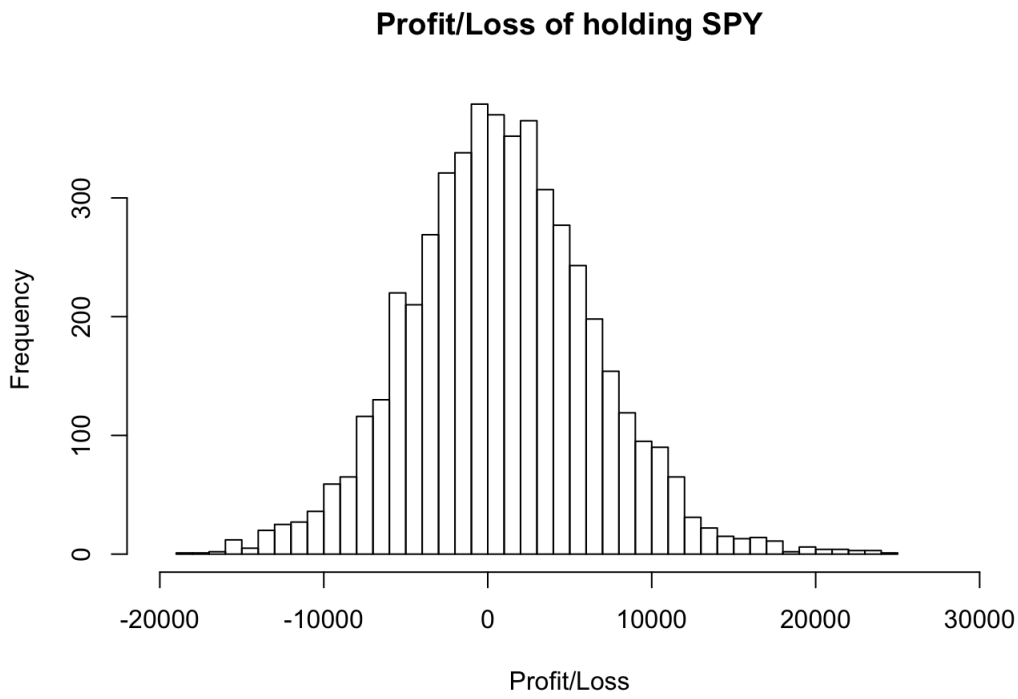
```
## [1] "SPY" "TLT" "LQD" "EEM" "VNQ"
```

```
##                ClCl.SPYa    ClCl.TLTa     ClCl.LQDa    ClCl.EEMa
## 2007-01-03            NA           NA             NA           NA
## 2007-01-04  0.0021221123  0.006063328  0.0075152938 -0.013809353
## 2007-01-05 -0.0079763183 -0.004352668 -0.0006526807 -0.029238205
## 2007-01-08  0.0046250821  0.001793566 -0.0002798843  0.007257535
## 2007-01-09 -0.0008498831  0.000000000  0.0001866169 -0.022336235
## 2007-01-10  0.0033315799 -0.004475797 -0.0013063264 -0.002303160
##               ClCl.VNQa
## 2007-01-03           NA
## 2007-01-04  0.001296655
## 2007-01-05 -0.018518518
## 2007-01-08  0.001451392
## 2007-01-09  0.012648208
## 2007-01-10  0.012880523
```

Analyzing the risk/return properties of the five major asset classes:

```
## the expected return for SPY is
```

```
## [1] 100859.9
```

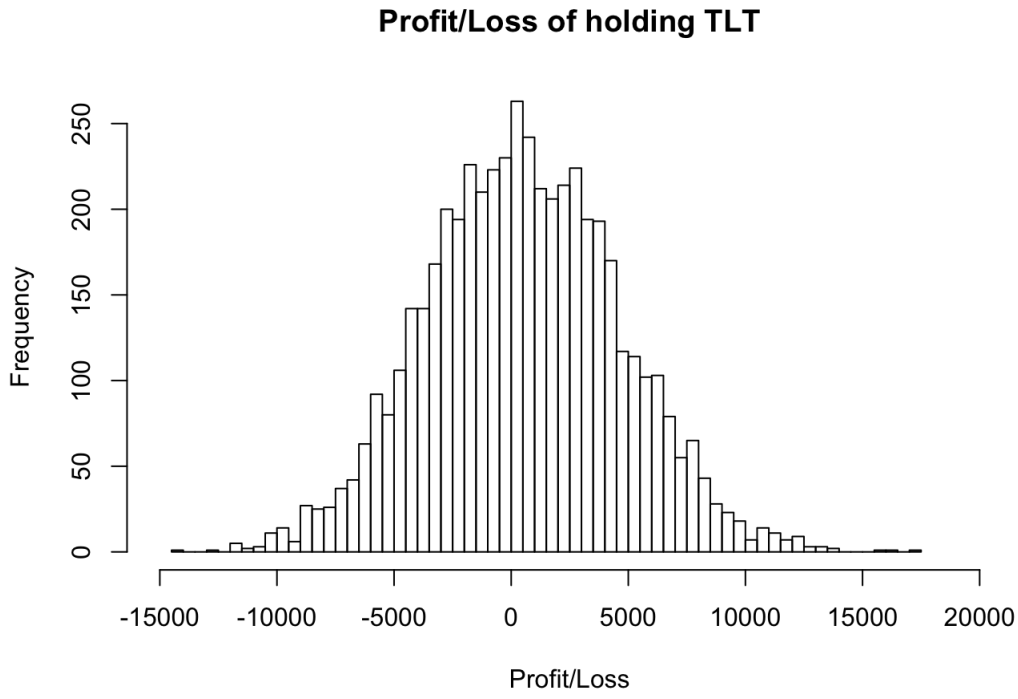



```
## the 5% value at risk for SPY is
```

```
## [1] -8058.488
```

```
## the expected return for TLT is
```

```
## [1] 100501.6
```

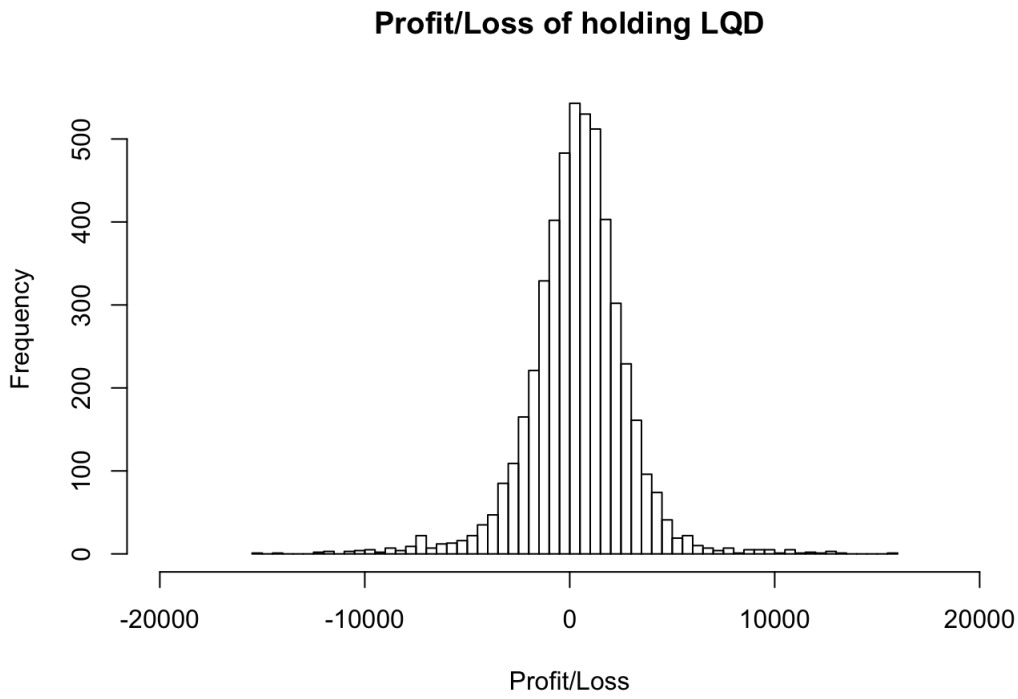



```
## the 5% value at risk for TLT is
```

```
## [1] -6115.421
```

```
## the expected return for LQD is
```

```
## [1] 100382
```

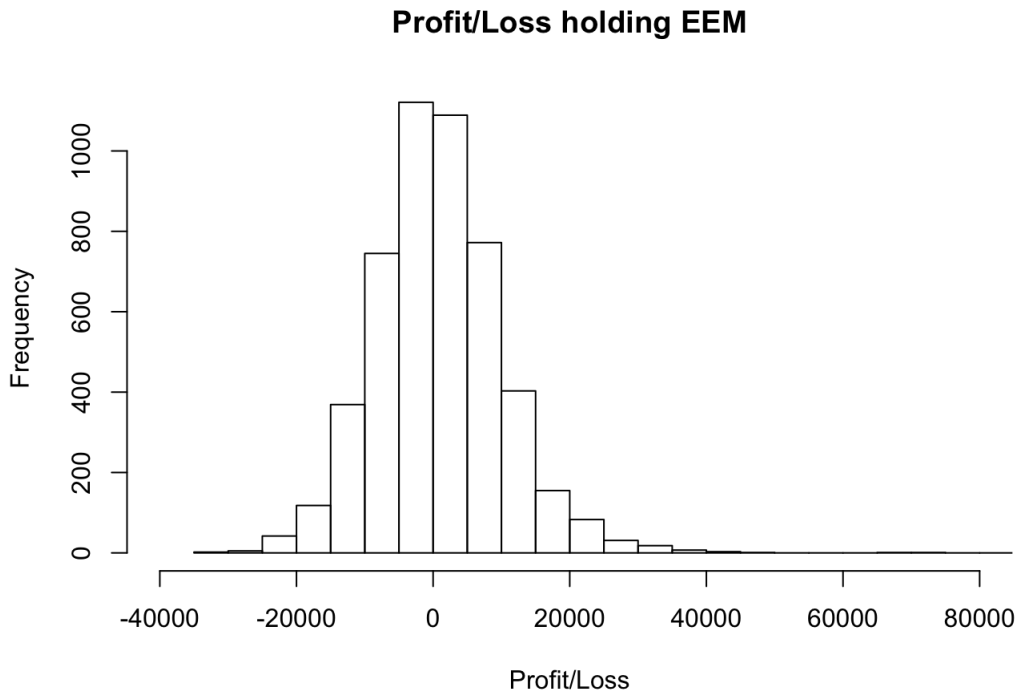



```
## the 5% value at risk for LQD is
```

```
## [1] -3236.454
```

```
## the expected return for EEM is
```

```
## [1] 102110.8
```

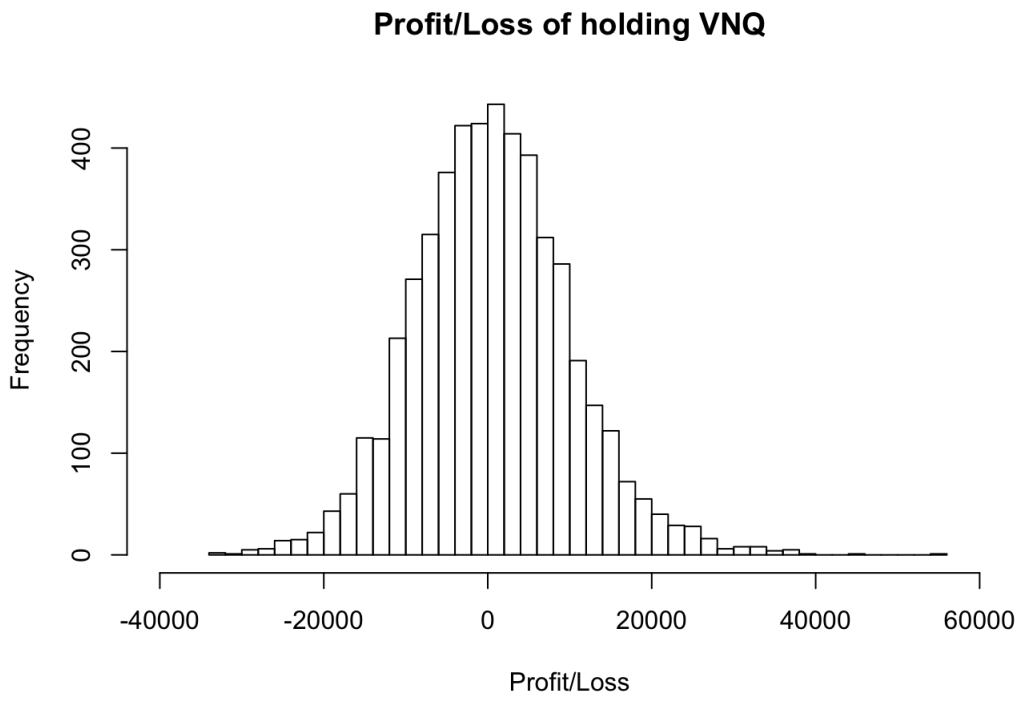



```
## the 5% value at risk for EEM is
```

```
## [1] -13302.82
```

```
## the expected return for VNQ is
```

```
## [1] 100638.2
```




```
## the 5% value at risk for VNQ is
```

```
## [1] -14605.19
```

Suppose we invest $100,000 in each of the five classes and hold those investments for 4 weeks, we can simulate the distributions of profit and loss using bootstrap resampling.

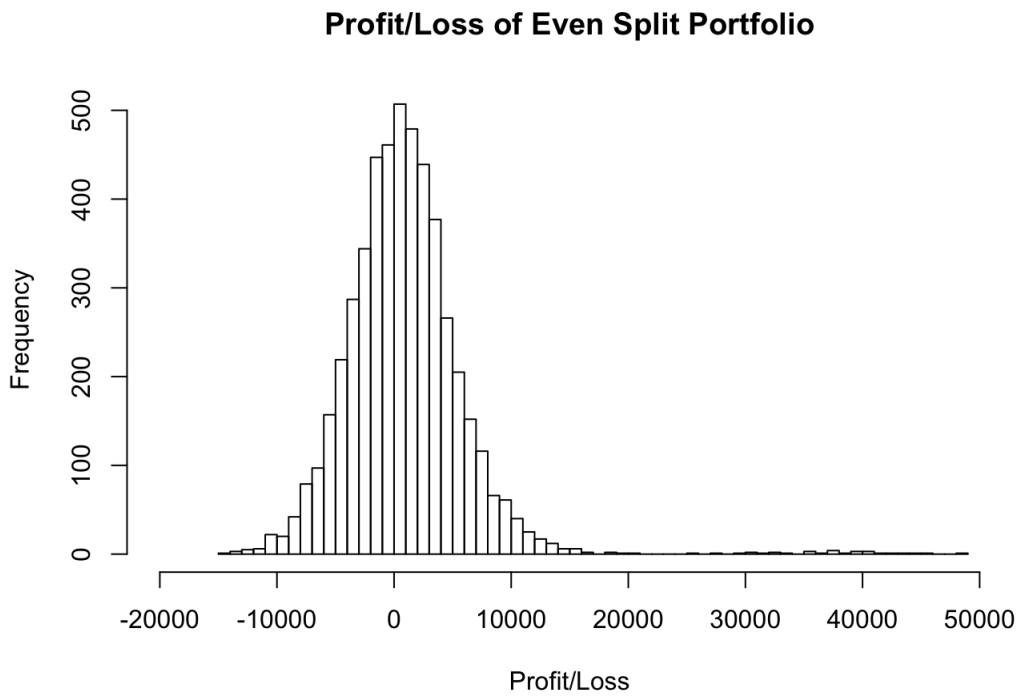
From the simulations, we ranked the assets by return and risks as following (with return raking of 1 being the highest return class and risk ranking of 1 being the lowest risk class):

Asset	expected return	Value at Risk	return ranking	risk ranking
SPY	$100859.9	-$8058.488	2	3
TLT	$100501.6	-$6115.421	4	2
LQD	$100382.0	-$3236.454	5	1
EEM	$102110.8	-$13302.82	1	4
VNQ	$100626.4	-$14643.62	3	5

For example, after investing $100,000 for 4 weeks in SPY, we expect to gain $739.3 on average. There is also a 5% chance of losing $8242.07 or more. Among the five classes, SPY has the 2nd highest return on average and 3rd highest risk.

```
## the expected return on the even split portfolio is
```

```
## [1] 100878.8
```




```
## the 5% value at risk for the even split portfolio is
```

```
## [1] -6140.826
```

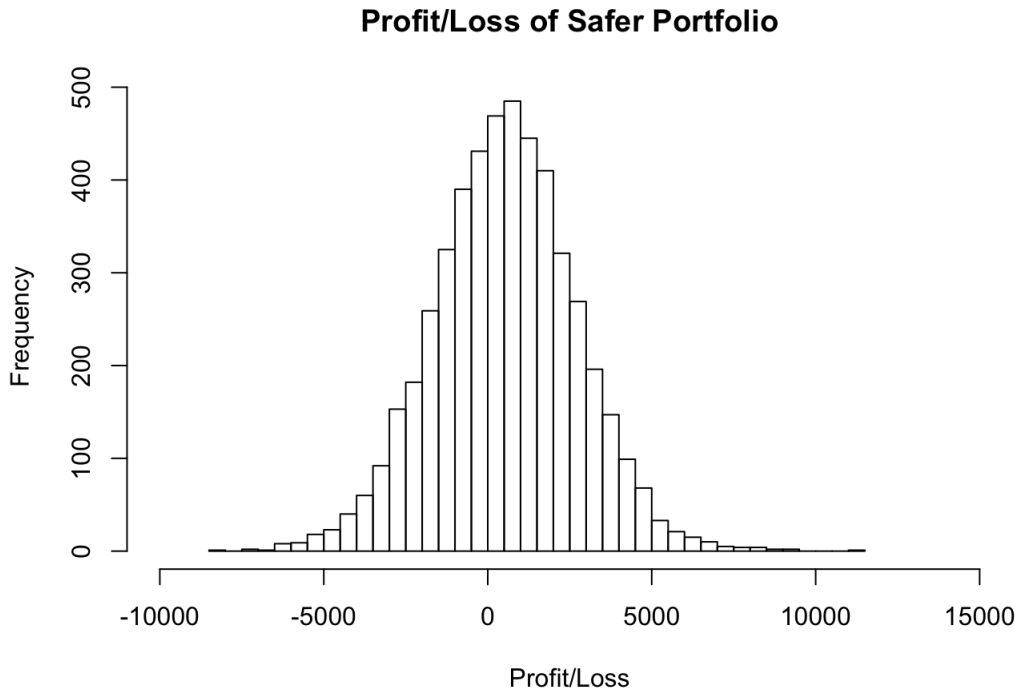
```
#Safer Portfolio: we invested 20% of the $100,000 in SPY, 40% in TLT, and 40% in LQD. The top 3 classes that has the lowest risk.
initial_wealth = 100000
set.seed(1)
sim2 = foreach(i=1:5000, .combine='rbind') %do% {
    total_wealth = initial_wealth
    weights = c(0.2, 0.4, 0.4, 0, 0)
    holdings = 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)
        holdings = weights * total_wealth
        wealthtracker[today] = total_wealth
    }
    wealthtracker
}

cat("the expected return on safe portfolio is")
```

```
## the expected return on safe portfolio is
```

```
# Profit/loss
mean(sim2[,n_days])
```

```
## [1] 100530.7
```

```
hist(sim2[,n_days]- initial_wealth, breaks=50, xlim = c(-10000, 15000), xlab = 'Profit/Loss', ylab = 'Frequency', main = "Profit/Loss of Safer Portfolio")
```




```
cat("the 5% value for the safe portfolio is")
```

```
## the 5% value for the safe portfolio is
```

```
# Calculate 5% value at risk
SAFE_risk = (quantile(sim2[,n_days], 0.05) - initial_wealth)
unname(SAFE_risk)
```

```
## [1] -3011.383
```

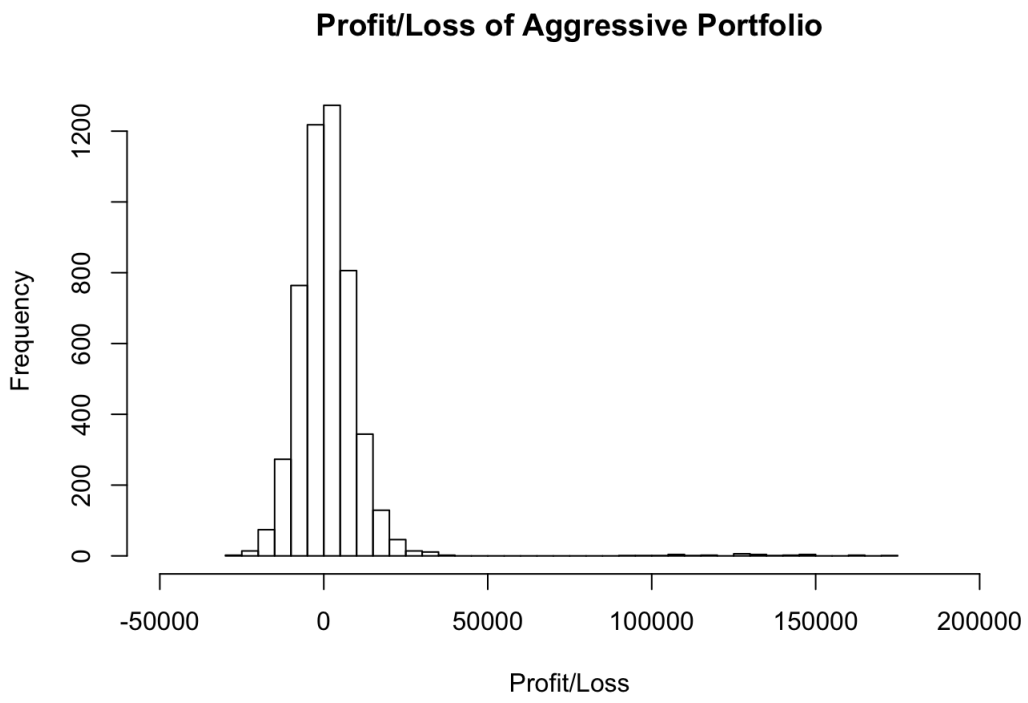
```
#Aggresive Portfolio: we invested 30% of the $100,000 in SPY and 70% in EEM, since they are the top 2 classe
s with highest return on average.
initial_wealth = 100000
set.seed(1)
sim3 = foreach(i=1:5000, .combine='rbind') %do% {
    total_wealth = initial_wealth
    weights = c(0.3, 0, 0, 0.7, 0)
    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)
        holdings = weights * total_wealth
        wealthtracker[today] = total_wealth
    }
    wealthtracker
}

cat("the expected return on the aggressive portfolio is")
```

```
## the expected return on the aggressive portfolio is
```

```
# Profit/loss
mean(sim3[,n_days])
```

```
## [1] 101555.5
```

```
hist(sim3[,n_days]- initial_wealth, breaks=50, xlim = c(-50000,200000),xlab = 'Profit/Loss', ylab = 'Frequency', main = "Profit/Loss of Aggressive Portfolio")
```




```
cat("the 5% value at risk for the aggressive portfolio is")
```

```
## the 5% value at risk for the aggressive portfolio is
```

```
# Calculate 5% value at risk
AGG_risk = quantile(sim3[,n_days], 0.05) - initial_wealth
unname(AGG_risk)
```

```
## [1] -11494.4
```

Portfolio Summary:

Even-Split: 20% of $100,000 is invested in each of the five assets

Safer : Invested in the top 3 classes that has the lowest risk (20% in SPY, 40% in TLT, and 40% in LQD)

Aggressive : Invested in the top 2 classes that has the highest return (30% in SPY and 70% in EEM)

Compare the 3 Portfolios:

Portfolios	expected return	Value at Risk	return ranking	risk ranking
Even-Split	$100875.8	-$6140.826	2	2
Safer	$100530.7	-$3011.383	3	1
Aggressive	$101555.5	-$11494.4	1	3

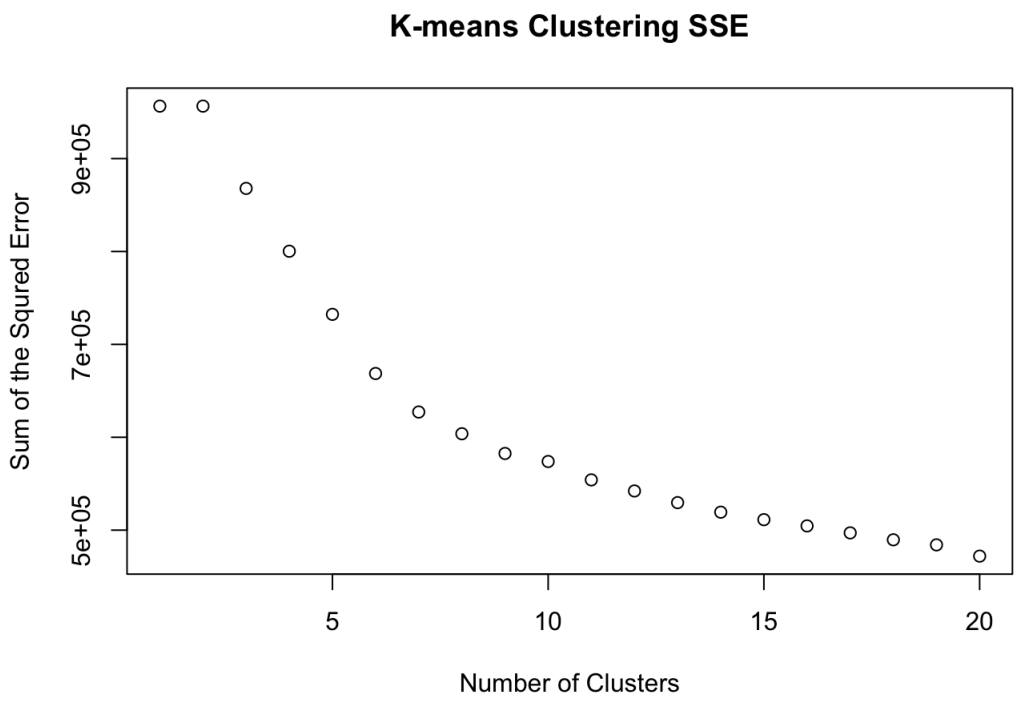
Given normal market conditions, even-split portfolio is expected to gain $977.1 on average. Even-split portfolio also has a 5% chance of losing $6360.44.

In conclusion, return and risk are inversely correlated.

The investor can choose among the three options depending on his or her risk tolerance and judgement on the upcoming market. For example, if the investor wants high return, can tolerate high risk, and is confident in the upcoming market, the aggressive portfolio is the best option.

Market segmentation

```
library(LICORS)
consumer_table = read.csv("social_marketing.csv", header = TRUE, sep = ",", row.names = 1)


Z = scale(consumer_table)
betweenss_table = rep(0, 20)
set.seed(1)
for (i in 1:20){
  kmeansPP_consumer = kmeanspp(consumer_table, k=i)
  betweenss_table[i] = kmeansPP_consumer$tot.withinss
}

plot(betweenss_table, xlab='Number of Clusters', ylab='Sum of the Squred Error', main='K-means Clustering SS
E')
```



We think keeping “adult”" and “chatter”" category inside the dataset could help us detect protentail spam accounts. Using the elbow method, we determine the optimal cluster number.

We choose 7 according to the plot by the elbow rule.

```
set.seed(1)
kmeansPP_consumer = kmeanspp(consumer_table, 7, nstart = 10)

print(apply(kmeansPP_consumer$centers,1,function(x) colnames(consumer_table)[order(x, decreasing=TRUE)[1:10]
]))
```

```
##       1                  2                3
##  [1,] "sports_fandom"    "politics"       "health_nutrition"
##  [2,] "religion"         "travel"         "personal_fitness"
##  [3,] "food"             "news"           "chatter"
##  [4,] "parenting"        "chatter"        "cooking"
##  [5,] "chatter"          "computers"      "outdoors"
##  [6,] "school"           "photo_sharing"  "photo_sharing"
##  [7,] "family"           "automotive"     "food"
##  [8,] "photo_sharing"    "sports_fandom"  "current_events"
##  [9,] "current_events"   "current_events" "shopping"
## [10,] "health_nutrition" "food"           "politics"
##       4                  5                  6
##  [1,] "chatter"          "chatter"          "cooking"
##  [2,] "photo_sharing"    "photo_sharing"    "photo_sharing"
##  [3,] "current_events"   "shopping"         "fashion"
##  [4,] "travel"           "current_events"   "chatter"
##  [5,] "tv_film"          "politics"         "beauty"
##  [6,] "politics"         "sports_fandom"    "health_nutrition"
##  [7,] "sports_fandom"    "health_nutrition" "shopping"
##  [8,] "health_nutrition" "college_uni"      "current_events"
##  [9,] "college_uni"      "travel"           "travel"
## [10,] "cooking"          "dating"           "college_uni"
##       7
##  [1,] "college_uni"
##  [2,] "online_gaming"
##  [3,] "chatter"
##  [4,] "photo_sharing"
##  [5,] "sports_playing"
##  [6,] "health_nutrition"
##  [7,] "cooking"
##  [8,] "travel"
##  [9,] "sports_fandom"
## [10,] "tv_film"
```

```
print(kmeansPP_consumer$size)
```

```
## [1]  607  557  933 3686 1209  493  397
```

Cluster	Size	Characteristics	Label
1	1209	sports_fandom, religion, food, parenting, chatter, school, family, photo_sharing, current_events, health_nutrition	Community-Minded
2	493	politics, travel, news, chatter, computers, photo_sharing, automotive, sports_fandom, current_event, food	Wordy
3	557	health_nutrition, personal_fitness, chatter, cooking, outdoors, photo_sharing, food, current_events,shopping,politics	Healthy & Fit
4	397	chatter, photo_sharing, current_events, travel, tv_film, politics, sports_fandom, health_nutrition, college_uni,cooking	Social
5	933	chatter, online_gaming, chatter, current_events, politics, sports_fandom, health_nutrition, college_uni, travel, dating	In-Door
6	607	cooking, photo_sharing, fashion, chatter,beauty, health_nutrition, shopping, current_events, travel, college_uni	Healthy & Social
7	3686	college_uni, online_gaming, chatter, photo_sharing, health_nutrition, cooking, travel, sports_fandom, tv_film	Young-adults

The first cluster enjoys daily lives. People in this cluster likes to cook, share photos, and follow fashion.

The second cluster is composed of people who care about social lives.

The third cluster cares about the world and loves to travel.

The fourth cluster prefers in-door activities.

The fifth cluster tends to be college students who love internet and social medias.

The sixth cluster values sports, religion, food, and family.

The seventh cluster cares about health and fitness.

```
consumer_scaled <- scale(consumer_table, center=TRUE, scale=TRUE)

consumer_distance_matrix = dist(consumer_scaled, method='euclidean')
hier_consumer = hclust(consumer_distance_matrix, method='average')


H_cluster = cutree(hier_consumer, k = 7)
summary(factor(H_cluster))
```

```
##    1    2    3    4    5    6    7
## 7864   10    2    2    1    2    1
```

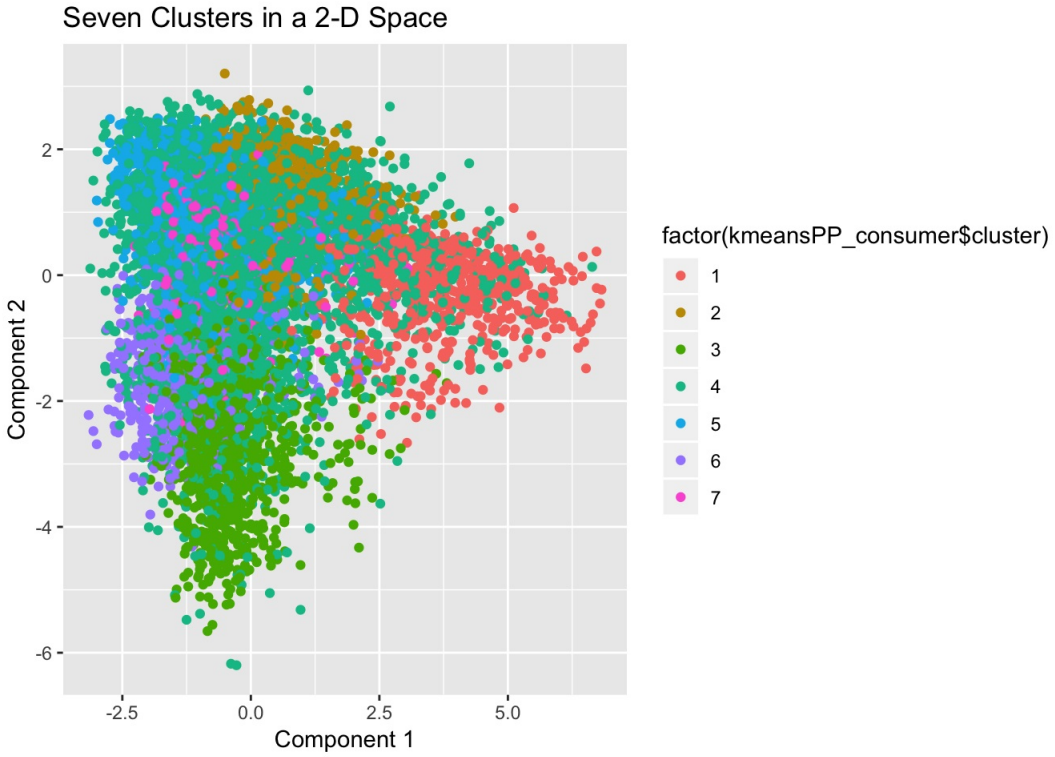
The hierarchical clustering give us a strange result. It seems like most of clusters are composed of outliers.

```
consumer2_scaled <- scale(consumer_table[which(H_cluster == 1),], center=TRUE, scale=TRUE)

consumer2_distance_matrix = dist(consumer2_scaled, method='euclidean')
hier2_consumer = hclust(consumer2_distance_matrix, method='average')


H_cluster2 = cutree(hier2_consumer, k = 7)
summary(factor(H_cluster2))
```

```
##    1    2    3    4    5    6    7
## 7796   47    2   11    5    2    1
```

Another hierarchical cluster within cluster 1 does not solve this problem.

```
Z = consumer_table/rowSums(consumer_table)
set.seed(1)
pc2 = prcomp(Z, scale=TRUE, rank=2)
loadings = pc2$rotation
scores = pc2$x

qplot(scores[,1], scores[,2], color=factor(kmeansPP_consumer$cluster), xlab='Component 1', ylab='Component 2
', main = 'Seven Clusters in a 2-D Space')
```



Cluster	Common Phrases	Labels
1	religion, sports fandom, parenting, food, school, family, news, automotive, crafts, politics	Community-minded

Cluster	Common Phrases	Labels
2	chatter, politics, travel, shopping, automotive, current_events, photo_sharing, news, computers, tv_film	Intellectual
3	politics, news, travel, outdoors, health_nutrition, personal_fitness, computers, automotive, adult, tv_film	Worldy & Healthy
4	photo_sharing, shopping, chatter, politics, news, automotive, computers, cooking, beauty, fashion	Social
5	fashion, beauty, cooking, politics, news, travel, computers, automotive, dating, college_uni	Trendy
6	online_gaming, automotive, sports_playing, college_uni, photo_sharing, news, shopping, family, chatter, sports_fandom	Sporty
7	automotive, news, tv_film, art, music, sports_fandom, home_and_garden, outdoors, current_events, beauty	Artsy & Sporty

The first cluster tends to be religious, who love sports, and who are parents and community-minded.

The second cluster values intellectual enjoyments.

The third cluster cares about both the world and themselves.

The fourth cluster loves social activities.

The fifth cluster follows fashion and beauty, and cares about the world around them.

The sixth cluster tends to be college student who like sports and games.

The seventh cluster loves automotive and arts.

From k-means and PCA, we found some common groups: people with liberal thoughts(enjoy intelligent activities and care about health), people who care about personal images, and traditional family people who pay attention to cars, parenting and schools, people value fitness, and people who like indoor activities. There is an interesting findings: photo-sharing is common among different clusters. We think it is because the whole data set is gained from the social media, where people like to share photos.