# STA380, Part2: Exercises 1

Nimish Amlathe, Hitesh Prabhu, Stuti Madaan

August 8, 2016

## Probability Practice

### Part A

**Here's a question a friend of mine was asked when he interviewed at Google.**

**Visitors to your website are asked to answer a single survey question before they get access to the content on the page. Among all of the users, there are two categories: Random Clicker (RC), and Truthful Clicker (TC). There are two possible answers to the survey: yes and no. Random clickers would click either one with equal probability. You are also giving the information that the expected fraction of random clickers is 0.3.**

**After a trial period, you get the following survey results: 65% said Yes and 35% said No.**

**What fraction of people who are truthful clickers answered yes?**

**Solution:**

$Prob(RC) = 0.3$

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

$Prob(Yes|RC) = Prob(No|RC) = 0.5$

$Prob(Yes) = 0.65$

$Prob(Yes|TC) = ?$

We solved this problem by using the concept of Total Sum of Probilities which states that:

$Prob(Yes) = Prob(Yes|RC) * Prob(RC) + Prob(Yes|TC) * Prob(TC)$

$0.65 = 0.5 * 0.3 + Prob(Yes|TC) * 0.7$

$0.65 - 0.15 = Prob(Yes|TC) * 0.7$

$0.5 = Prob(Yes|TC) * 0.7$

$Prob(Yes|TC) = 0.5/0.7 = 0.71$

We know that the probability of a Random clicker is 0.3. Also, we know that the probability of Yes/No given a Random Clicker is 0.5. From this, we can derive the overall probaility of a Yes/No from a Random Clicker = 0.5 * 0.3 = 0.15.

Thus, the fraction of 'Yes'es from a Truthful speaker = 0.5/0.7 = 0.7142857

## Part B

**Imagine a medical test for a disease with the following two attributes:**

- **The sensitivity is about 0.993. That is, if someone has the disease, there is a probability of 0.993 that they will test positive.**
- **The specificity is about 0.9999. This means that if someone doesn't have the disease, there is probability of 0.9999 that they will test negative.**

**In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it (or 0.000025 as a decimal probability).**

**Suppose someone tests positive. What is the probability that they have the disease? In light of this calculation, do you envision any problems in implementing a universal testing policy for the disease?**

**Solution:**

$Prob(Positive|Disease) = 0.993$

$Prob(Negative|DoesnothaveDisease) = 0.9999$

$Prob(Disease) = 0.000025$

$$Prob(Disease|Postive) = (P(Positive|Disease) * P(Disease))/(P(Positive|Disease) * P(Disease) + P(Positive|NoDisease) * P(DoesNotHaveDisease))$$

$$= (0.993 * 0.000025)/(0.993 * 0.000025 + (1 - 0.9999) * (1 - 0.000025))$$

$$= 0.1988824$$

If a test's result is positive, the probability that there is a disease is very less i.e.approximately 0.2. This implies that there are a lot of false positives from this test. If this is implemented as a universal testing policy, a lot of people will be falsely informed of having the disease when they don't. In medical world, this would be a blunder.

## Exploratory analysis: green buildings

### Loading data and loading

```
greenBuildings <- read.csv("files/greenbuildings.csv")

summary(greenBuildings[which(greenBuildings$green_rating == 0
```

```
                                & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.43   25.03   28.44   34.18  250.00
```

```
summary(greenBuildings[which(greenBuildings$green_rating == 1
                                & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.50   27.60   30.03   35.54  138.10
```

### 1. Looking at the relationship between the variables *EnergyStar*, *LEED* and *green_rating*

The reason we decided to investigate this first is because

EPA's ENERGY STAR identifies the nation's most energy-efficient commercial buildings and industrial plants. Through ENERGY STAR, EPA offers 1 – 100 *ENERGY STAR* scores that rate buildings against their peers. To earn the ENERGY STAR, a fully operational facility must earn an ENERGY STAR score of 75 or higher, meaning that it performs in the top 25 percent of similar facilities nationwide for energy efficiency.

LEED is a green building rating system administered by the private non-profit U.S. Green Building Council. LEED addresses several environmental attributes in addition to energy efficiency, such as materials, waste, and water. To earn LEED certification, a building does not always need to meet the rigorous energy performance level required to earn EPA's ENERGY STAR.

While LEED can help organizations achieve a wide range of sustainability goals, ENERGY STAR certification is the only way to ensure superior energy performance. For this reason, the two programs can work very well together.

LEED is frowned upon by many owners and investors, however, because it can be incredibly expensive to become certified. Many owners, developers and investors pass on LEED certification because the additional cost of commissioning, paperwork and professional fees seems daunting and unnecessary. In fact, LEED and Energy Star are complimentary to each other. Buildings may be both LEED certified and Energy Star rated, and LEED requires Energy Star as part of its EB (Existing Building) rating system.

```
##                         LEED    0    1
## green_rating Energystar
## 0            0               7209    0
##              1                  0    0
## 1            0                  0   47
##              1                631    7
```

There are only 47 buildings which are LEED certified amongst the 685 green buildings. The rest are Energy-star rated buildings. Now let's look at the median rents of these subgroups:

```r
# Checking medians of buildings which are amongst the categories in the x-tab above
summary(greenBuildings[which(greenBuildings$Energystar == 0
                            & greenBuildings$LEED == 0
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.43   25.03   28.44   34.18  250.00
```

```r
summary(greenBuildings[which(greenBuildings$Energystar == 1
                            & greenBuildings$LEED == 0
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.55   28.12   30.06   35.79  138.10
```

```r
summary(greenBuildings[which(greenBuildings$Energystar == 0
                            & greenBuildings$LEED == 1
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   21.50   24.36   29.21   31.02   98.65
```

```r
summary(greenBuildings[which(greenBuildings$Energystar == 1
                            & greenBuildings$LEED == 1
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.50   20.50   24.00   32.99   38.22   72.00
```

The median rents of the LEED-rated buildings was found to be around 28.12, but those of LEED-only certified (which form the majority of the greeen buildings) was found to be 24. Clearly, this is our first confounding variable. Not *all* green buildings demand higher rent. In fact, LEED-only certified buildings have a lower median than non-green buildings!

### 2. *net* rent builings and *green_rating*

It is quite important not to compare rents amongst those buildings which include utilities and those who don't separately. Let's take a look at the relative distribution of utilities-incuded buildings and vice-versa.

```
##              net    0    1
## green_rating
## 0               6974  235
## 1                646   39
```

Out of 685 green buildings, 646 buildings have net = 0. That is, utilities are included as part of their rent. This is a significant proprotion of green buildings (94 percent). Let's take a closer look at the medians.

```
summary(greenBuildings[which(greenBuildings$net == 1
                            & greenBuildings$green_rating == 0
                            & greenBuildings$leasing_rate > 10)
,'Rent'],na.rm = T)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.19   19.10   21.96   24.32   24.91   82.43

summary(greenBuildings[which(greenBuildings$net == 0
                            & greenBuildings$green_rating == 0
                            & greenBuildings$leasing_rate > 10)
,'Rent'],na.rm = T)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.50   25.34   28.59   34.20  250.00

summary(greenBuildings[which(greenBuildings$net == 0
                            & greenBuildings$green_rating == 1
                            & greenBuildings$leasing_rate > 10)
,'Rent'],na.rm = T)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.69   28.20   30.37   35.99  138.10

summary(greenBuildings[which(greenBuildings$net == 1
                            & greenBuildings$green_rating == 1
                            & greenBuildings$leasing_rate > 10)
,'Rent'],na.rm = T)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.27   19.88   22.29   24.39   26.76   50.53
```

Amongst green buildings, rent amongst those including utilities is 28.2 while those without utilities included is 22.29. This clearly shows that we cannot judge median rents of green buildings without first accounting for the fact that 94 percent of them include utilities (even though a similar number of non-green buildings also have net = 0).

### 2. Relationship Classes *A* and *B* builings and *green_rating*

People tend to willing to pay more rent for buildings with higher quality. Hence it goes without saying that Class A buildings will demand more rent than a similar Class B building. Looking at their distribution:

```
##                       class_b    0    1
## green_rating class_a
## 0            0                1103 3495
##              1                2611    0
```

```
## 1            0                  7  132
##              1                546    0
```

```
##                          class_b    0    1
## Energystar LEED class_a
## 0          0    0              1103 3495
##                 1              2611    0
##            1    0                 1   14
##                 1                32    0
## 1          0    0                 6  117
##                 1               508    0
##            1    0                 0    1
##                 1                 6    0
```

546 out of 685 green buildings are class A buildings. This is a a lot higher than the porportion of Class A buildings (81 percent) amongst the rest which stands at 2611 (36.3 percent). Let's take a closer look at the medians.

```r
summary(greenBuildings[which(greenBuildings$class_a == 1
                            & greenBuildings$green_rating == 0
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   21.50   28.20   32.64   38.00  250.00
```

```r
summary(greenBuildings[which(greenBuildings$class_a == 0
                            & greenBuildings$green_rating == 0
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   18.00   23.65   25.98   31.80  200.00
```

```r
summary(greenBuildings[which(greenBuildings$class_a == 0
                            & greenBuildings$green_rating == 1
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   19.51   25.68   26.23   31.43   98.65
```

```r
summary(greenBuildings[which(greenBuildings$class_a == 1
                            & greenBuildings$green_rating == 1
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   22.07   28.44   30.99   36.59  138.10
```

```r
summary(greenBuildings[which(greenBuildings$class_b == 1
                            & greenBuildings$green_rating == 0
```

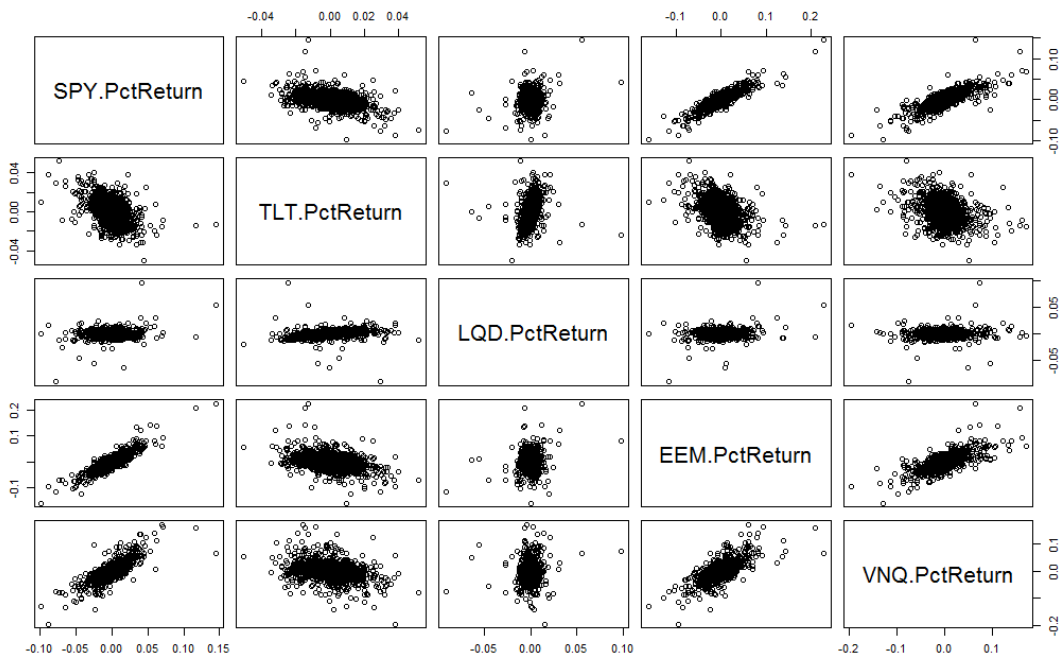
```
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   18.09   24.00   26.52   32.50  199.00

summary(greenBuildings[which(greenBuildings$class_b == 0
                            & greenBuildings$green_rating == 0
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.07   20.13   25.85   30.26   35.21  250.00

summary(greenBuildings[which(greenBuildings$class_b == 0
                            & greenBuildings$green_rating == 1
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   22.03   28.44   30.95   36.52  138.10

summary(greenBuildings[which(greenBuildings$class_b == 1
                            & greenBuildings$green_rating == 1
                            & greenBuildings$leasing_rate > 10) ,'Rent'],
na.rm = T)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   19.52   25.20   26.12   30.60   98.65
```

Amongst green buildings, rent amongst those which are Class A buildings is 28.44 while the same amongst Class B or C is 25.68. This clearly shows that we cannot judge median rents of green buildings without first accounting for the fact that 81 percent of them class A buildings, and this could be the only reason that their rent is higher.

### Conclusion

We conclude by having idenified 4 variables that have to be controlled for before comparing green and non-green buildings: EnergyStar, LEED, net and class_a. Each of these 4 variables could by itself raise or lower rent by almost 5 dollars per square foot.

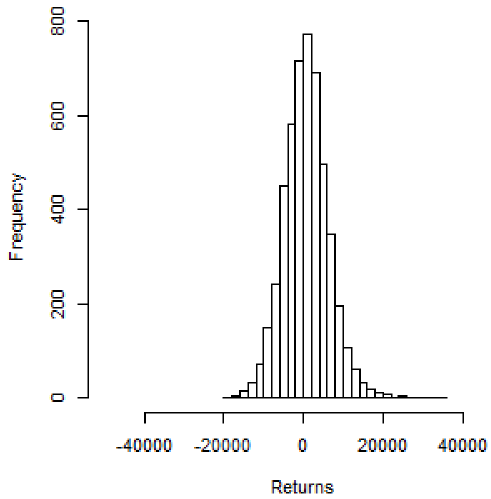
## Bootstrapping



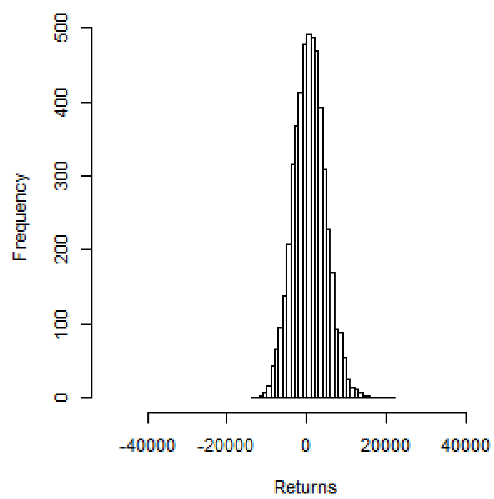

The correlation graph between stocks gives us an overview of how the stocks are related:

- A safer stock would be the one which does not vary much due to changes in other stocks and has positive returns. In this case, LQD is one such stock with close to zero correlation with all other stocks except TLT. LQD has a negative correlation of 0.42 with TLT, which is again very less.
- Also, SPY, EEM and VNQ are positively correlated with each other. This makes them the riskier stocks since they are sensitive to variations in other stocks. The Exact order of riskiness can e determined buu the means and standard deviations of individual stocks
- TLT is negatively correlated with SPY, EEM and VNQ but it is safer than these 3 stocks since the correlation is weaker. The correlation plots for TLT are loose and spread out.

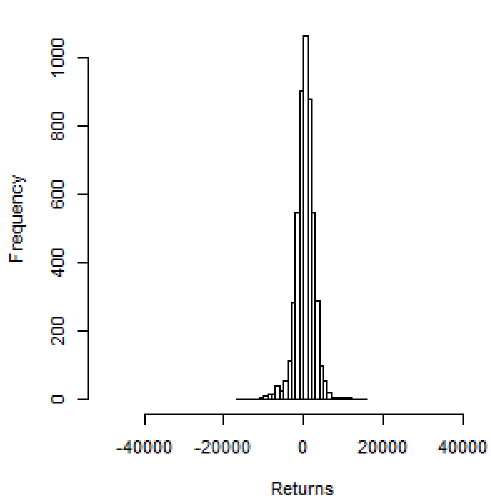
**Histogram for SPY.PctReturn**



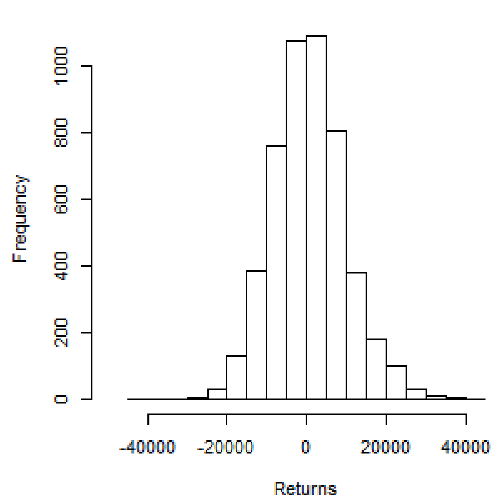

**Histogram for TLT.PctReturn**



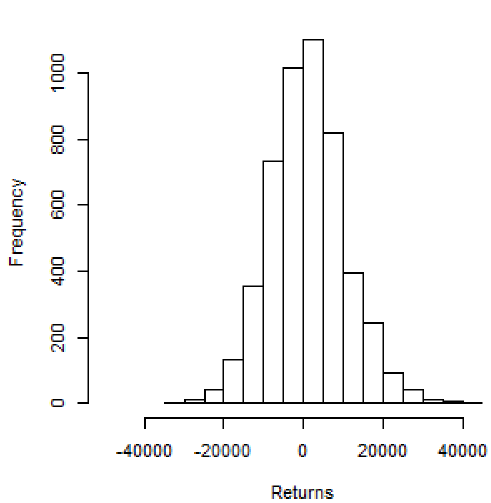

**Histogram for LQD.PctReturn**




**Histogram for EEM.PctReturn**




**Histogram for VNQ.PctReturn**

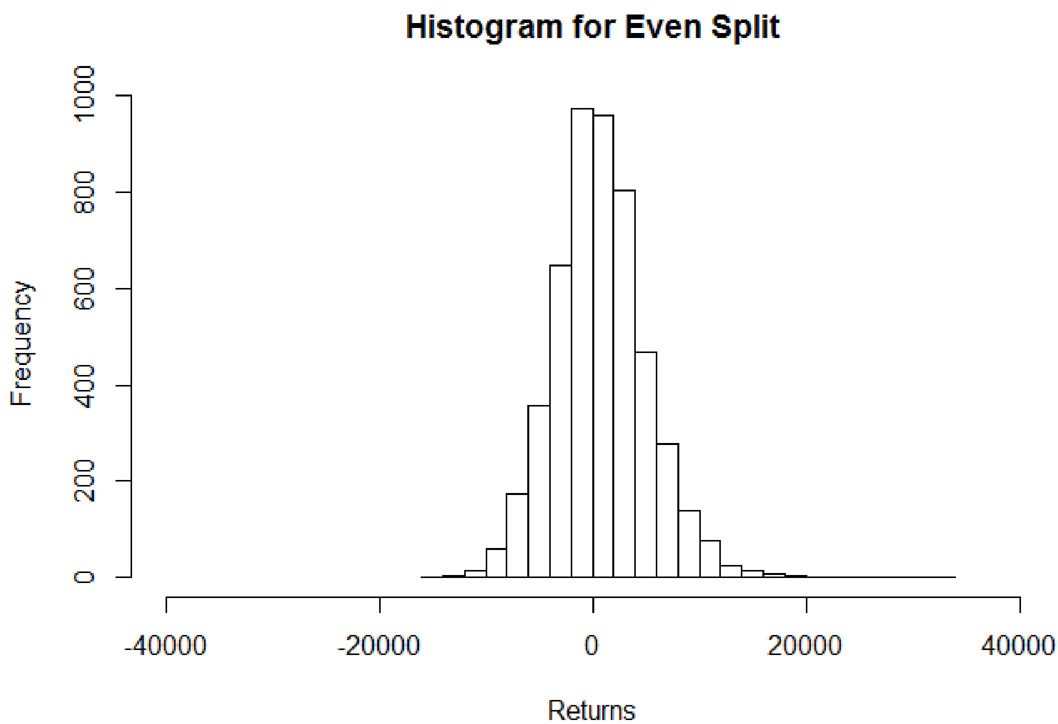


The individual 5% Value at Risk for the 5 stocks in the order of SPY, TLT, LQD, EEM and VNQ are: -8113.3881793, -5840.7310754, -3212.4730924, -1.367477110^{4}, -1.37367610^{4}

The individual means for the 5 stocks in the order of SPY, TLT, LQD, EEM and VNQ are: 1.00714210^{5}, 1.007344110^{5}, 1.004080210^{5}, 1.00799210^{5}, 1.011642810^{5}

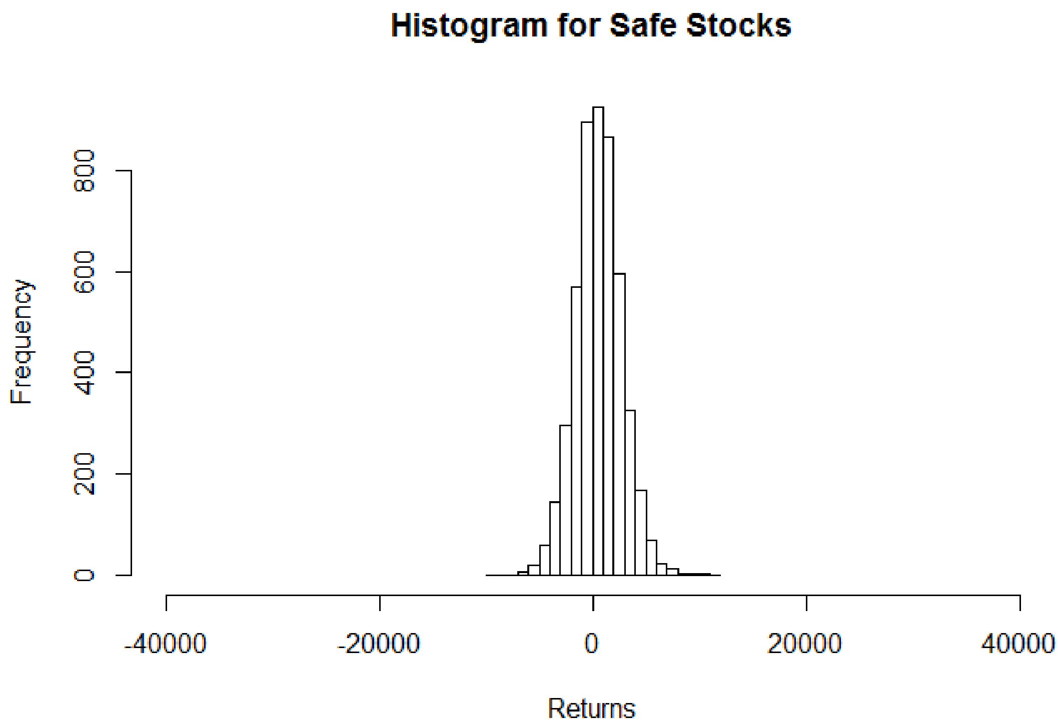
The individual standard deviations for the 5 stocks in the order of SPY, TLT, LQD, EEM and VNQ are: 5553.2480367, 4040.2782479, 2399.7497675, 9196.8423902, 9468.5590593

**Even split: 20% of the assets in each of the five ETFs above.**




The 5% Value at Risk for Even Split = -6049.0732004

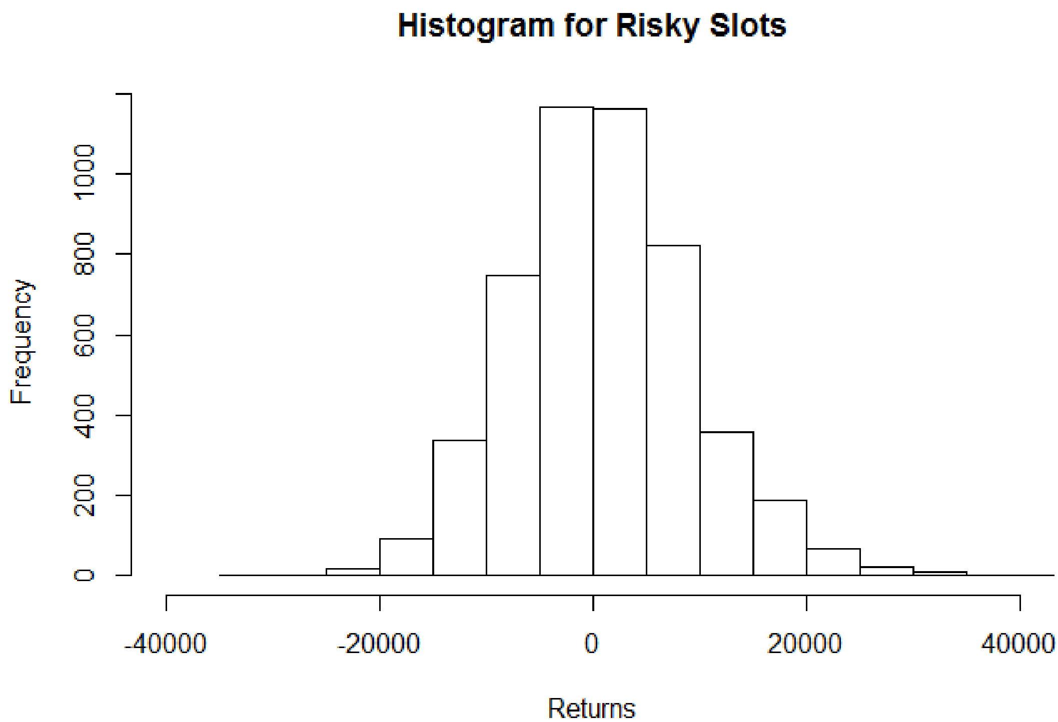
## Safe Investment




We have established that the Assets : SPY, TLT and LQD are the safer stocks by looking at their individual histograms. The proportion that we have taken for the investment is 0.1, 0.3, 0.6 , respectively.

Using this proportion, the 5% value at risk comes out to be -2941.9844668

### Aggressive Investment




## Conclusion

We have established that the Assets : EEM and VNQ are the riskier stocks. The proportion that we have taken for the investment is 0.5, 0.5, respectively.

Using this proportion, the 5% value at risk comes out to be -1.231489110^{4}

# Market segmentation

### Inital Set-up and Loading the Data:

```
library(flexclust)

## Loading required package: grid

## Loading required package: modeltools

## Loading required package: stats4

library(ggplot2)
library(reshape2)
library(corrplot)
library(corrgram)

social_mkt = read.csv("files/social_marketing.csv")
```

Initial analysis of data suggested that 'spam' and 'adult' just noise and don't add any value to the analysis.

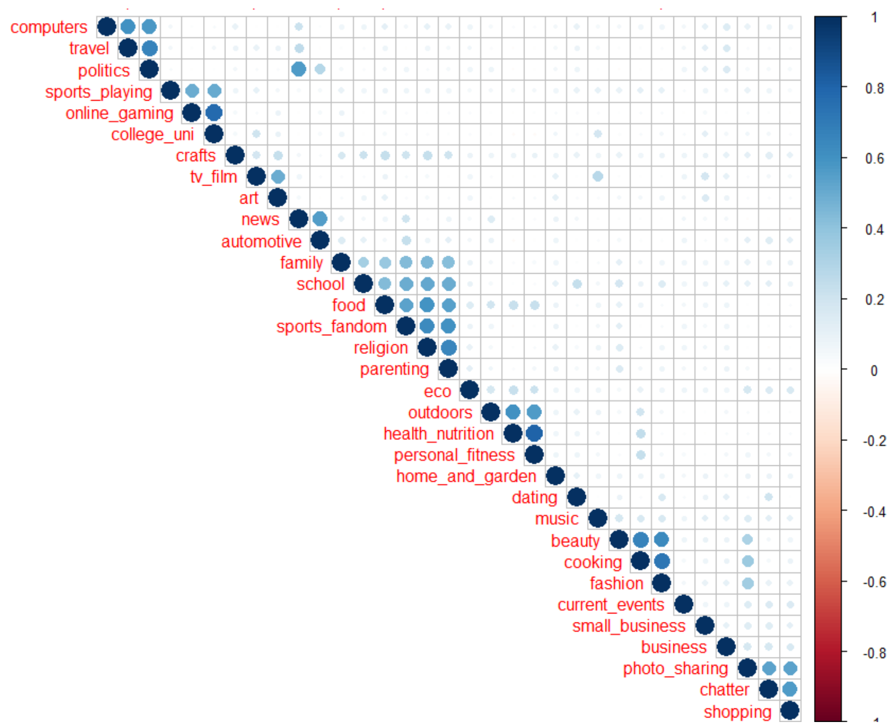
```
sm_wo_junk = social_mkt[,-c(36,37)]
sm_wo_id = sm_wo_junk[,-1]
```

Since annotators used both *uncategorized* and *chatter* for tweets that didn't fit into any of the categories, we decided to club both of these together. '

```
sm_wo_id$chatter = sm_wo_id$uncategorized + sm_wo_id$chatter
sm_wo_id = sm_wo_id[,-5]
```

In order to get started with cluster formation and analysis, we began by exploring the correlations.

```
correlation_matrix = cor(sm_wo_id)
corrplot(correlation_matrix, type="upper", order="hclust")
```




```
sm_wo_id_scaled <- scale(sm_wo_id, center=TRUE, scale=TRUE)

# Creating pair-wise correlations ordered by max correlation
zdf <- as.data.frame(as.table(cor(sm_wo_id)))
zdf_2 <- subset(zdf, (abs(Freq) > 0.4 & Var1 != Var2))
zdf_2[order(zdf_2$Freq,decreasing = T), ]

##                  Var1             Var2      Freq
## 493   personal_fitness health_nutrition 0.8099024
## 1005 health_nutrition personal_fitness 0.8099024
## 412        college_uni     online_gaming 0.7728393
## 508      online_gaming       college_uni 0.7728393
```

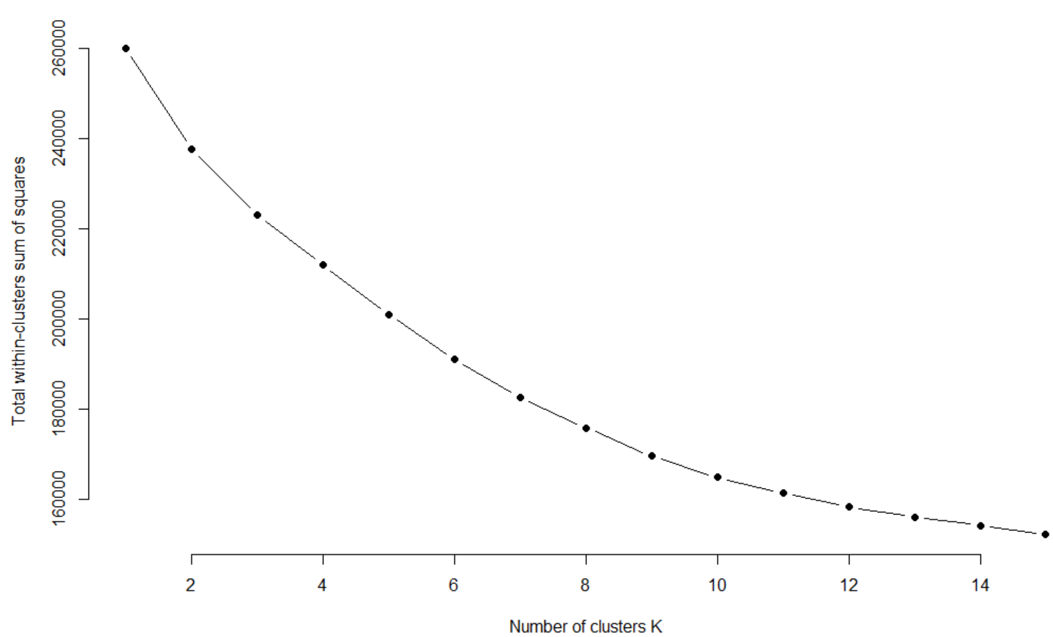
```
## 593            fashion          cooking 0.7214027
## 1041           cooking          fashion 0.7214027
## 588             beauty          cooking 0.6642389
## 876            cooking           beauty 0.6642389
## 73            politics           travel 0.6602100
## 201             travel         politics 0.6602100
## 853          parenting         religion 0.6555973
## 917           religion        parenting 0.6555973
## 191           religion    sports_fandom 0.6379748
## 831      sports_fandom         religion 0.6379748
## 890            fashion           beauty 0.6349739
## 1050            beauty          fashion 0.6349739
## 484           outdoors health_nutrition 0.6082254
## 708   health_nutrition         outdoors 0.6082254
## 193          parenting    sports_fandom 0.6077181
## 897      sports_fandom        parenting 0.6077181
## 86           computers           travel 0.6029349
## 630             travel        computers 0.6029349
## 257           religion             food 0.5913181
## 833               food         religion 0.5913181
## 218          computers         politics 0.5721506
## 634           politics        computers 0.5721506
## 14            shopping          chatter 0.5686992
## 430            chatter         shopping 0.5686992
## 724   personal_fitness         outdoors 0.5677903
## 1012          outdoors personal_fitness 0.5677903
## 210               news         politics 0.5618422
## 370           politics             news 0.5618422
## 387         automotive             news 0.5554175
## 771               news       automotive 0.5554175
## 259          parenting             food 0.5449481
## 899               food        parenting 0.5449481
## 113           shopping    photo_sharing 0.5356210
## 433      photo_sharing         shopping 0.5356210
## 4        photo_sharing          chatter 0.5342643
## 100            chatter    photo_sharing 0.5342643
## 173               food    sports_fandom 0.5326384
## 237      sports_fandom             food 0.5326384
## 855             school         religion 0.5162180
## 983           religion           school 0.5162180
## 512     sports_playing      college_uni 0.5063748
## 544        college_uni   sports_playing 0.5063748
## 921             school        parenting 0.4996164
## 985          parenting           school 0.4996164
## 157                art          tv_film 0.4987718
## 797            tv_film              art 0.4987718
## 195             school    sports_fandom 0.4931062
## 963      sports_fandom           school 0.4931062
## 413     sports_playing    online_gaming 0.4912993
## 541      online_gaming   sports_playing 0.4912993
```

```
## 290          religion          family 0.4527685
## 834            family        religion 0.4527685
## 174            family   sports_fandom 0.4378104
## 270     sports_fandom          family 0.4378104
## 261            school            food 0.4324039
## 965              food          school 0.4324039
## 292         parenting          family 0.4205780
## 900            family       parenting 0.4205780
```

Looking at the correlation plot, its seems that there are 5-7 combinations of correlated variables. People with majority of tweets in these categories can be clustered together.

In order to get an optimal number of clusters, we implemented the *Elbow* method which gave us an optimum cluster number.

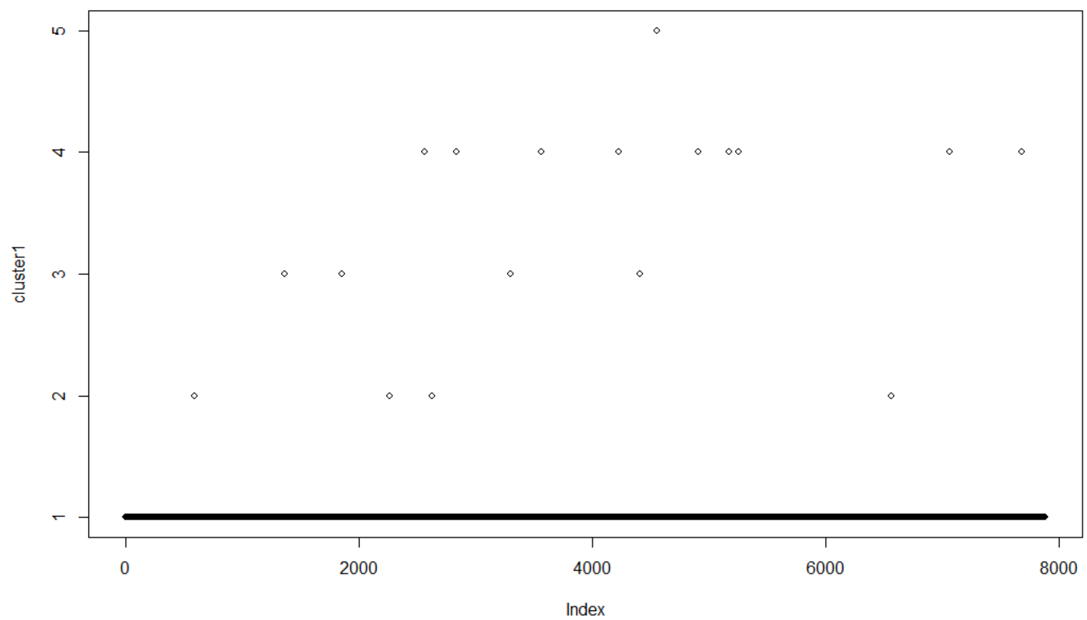
```r
set.seed(1)
# Compute and plot wss for k = 2 to k = 15
k.max <- 15 # Maximal number of clusters
data <- sm_wo_id_scaled
wss <- sapply(1:k.max,
        function(k){kmeans(data, k, nstart=10 )$tot.withinss})
plot(1:k.max, wss,
       type="b", pch = 19, frame = FALSE,
       xlab="Number of clusters K",
       ylab="Total within-clusters sum of squares")
```



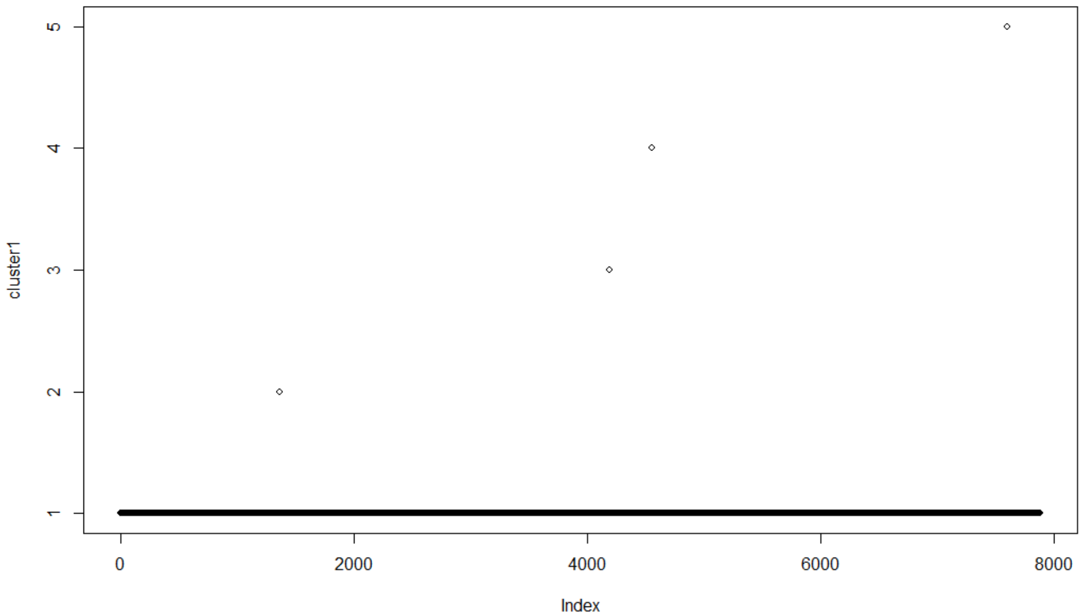

The optimal number of clusters as per the algorithm is 5-7. Now, we evaluate different clustering methods to get the clusters.

## Heirarchical Clustering

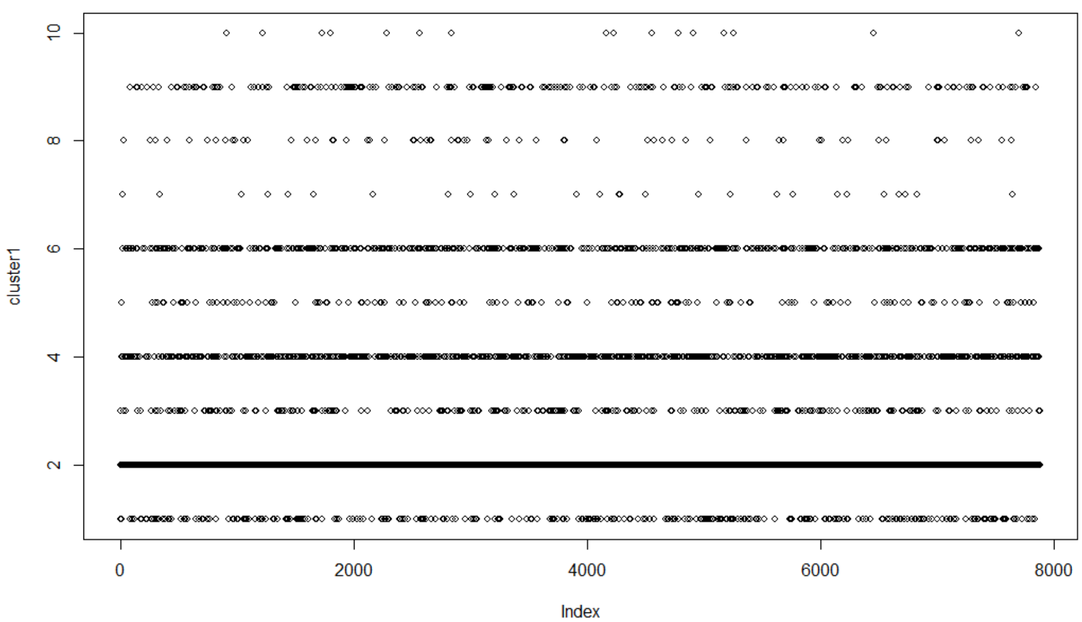
```
##Heirarchical Clustering ##
par(mfrow=c(1,1))
# Form a pairwise distance matrix using the dist function
distance_matrix = dist(sm_wo_id_scaled, method='euclidean')

# Now run hierarchical clustering
hier_p = hclust(distance_matrix, method='average')
cluster1 = cutree(hier_p, k=5)

# Plot the dendrogram
plot(cluster1, cex=0.8)
```



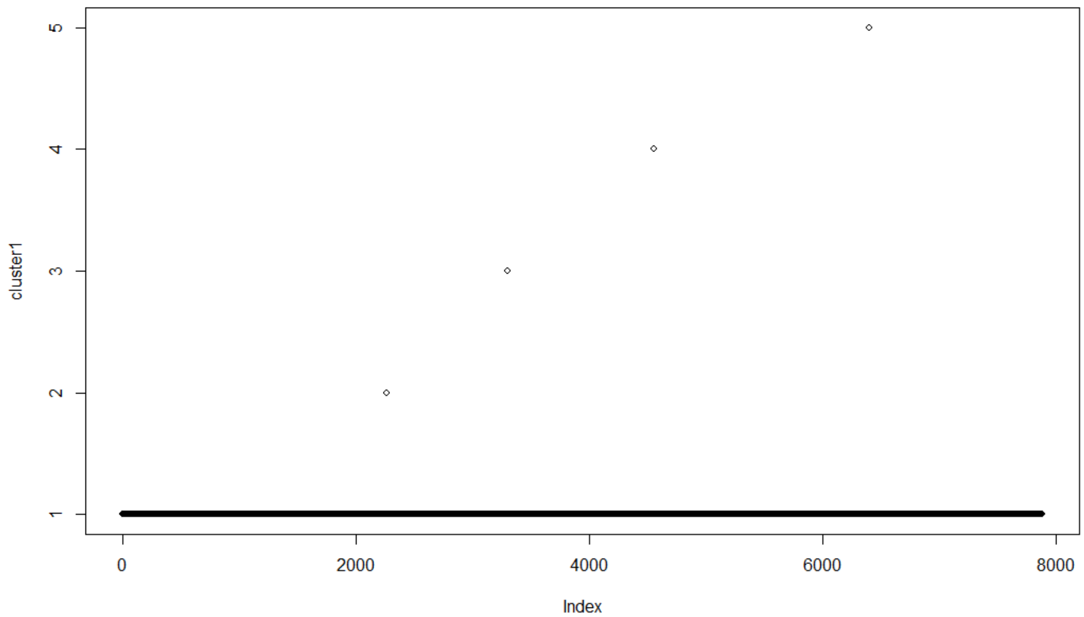

```
# Now run hierarchical clustering
hier_p = hclust(distance_matrix, method='centroid')
cluster1 = cutree(hier_p, k=5)

# Plot the dendrogram
plot(cluster1, cex=0.8)
```




```
# Now run hierarchical clustering
hier_p = hclust(distance_matrix, method='complete')
cluster1 = cutree(hier_p, k=10)

# Plot the dendrogram
plot(cluster1, cex=0.8)
```



```r
# Now run hierarchical clustering
hier_p = hclust(distance_matrix, method='single')
cluster1 = cutree(hier_p, k=5)

# Plot the dendrogram
plot(cluster1, cex=0.8)
```

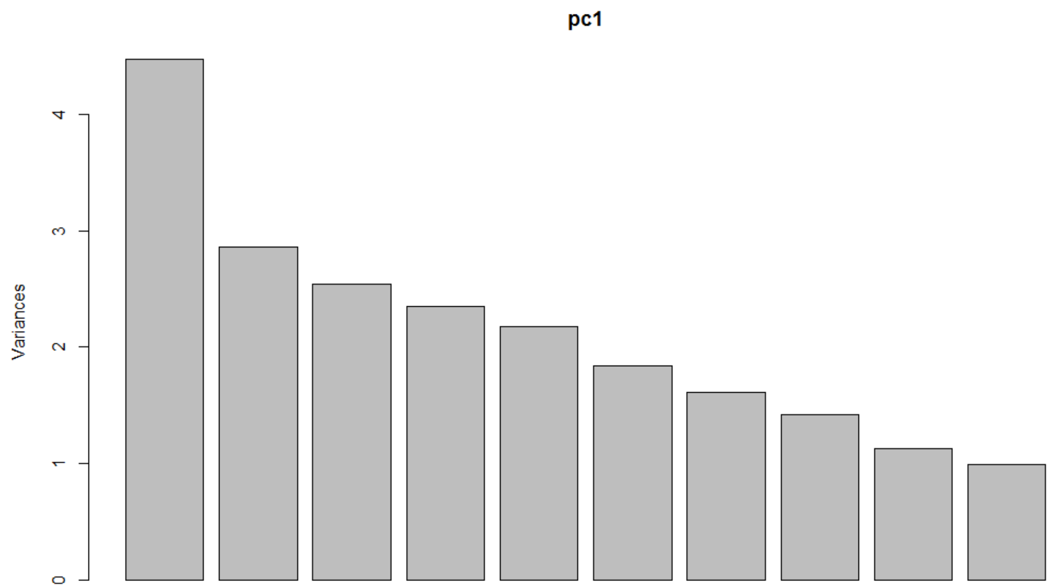



```r
ind =which(cluster1 == 3)

sm_wo_id_node2 = sm_wo_id[ind,]
```

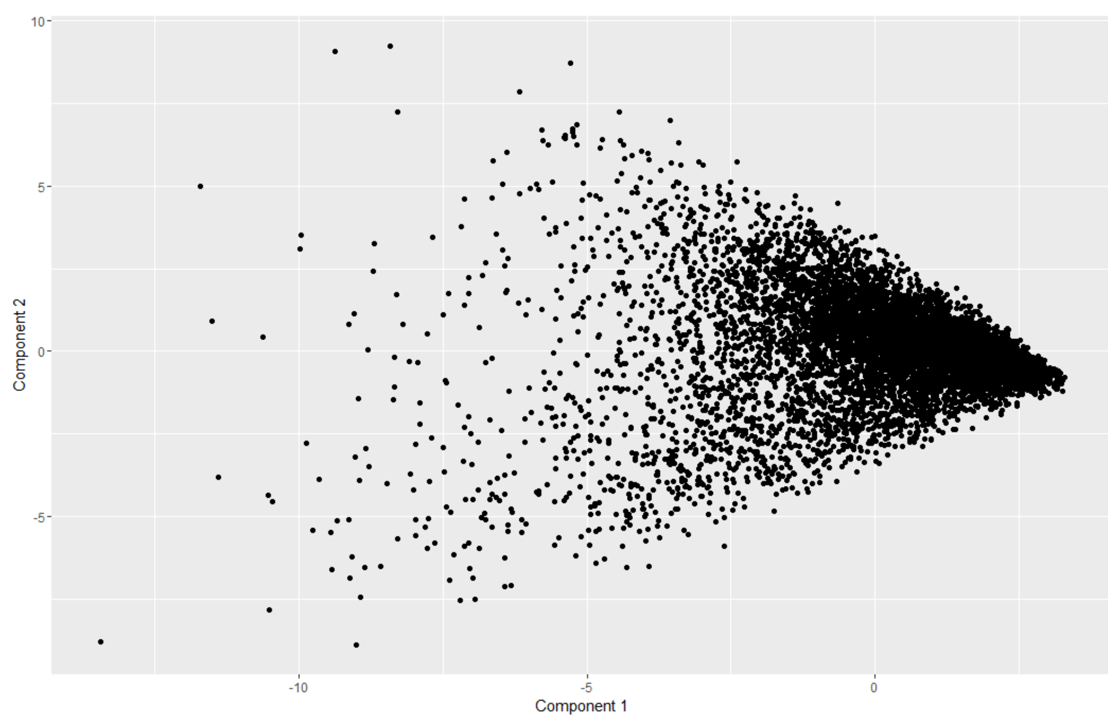
As we can see in Hierarchical clustering, the depth of the tree is very large. We tried to cut tree to 5 Clusters, but the results are hard to interpret and even after we cut the trees at lesser levels, we couldn't derive meaning from the different hierarchies.

### PCA

```r
## PCA ##
Z = sm_wo_id
Z_normalized = scale(Z, scale=T, center=T)
pc1 = prcomp(as.matrix(Z), scale.=TRUE)
plot(pc1)
```




```
loadings = pc1$rotation
scores = pc1$x
qplot(scores[,1], scores[,2], xlab='Component 1', ylab='Component 2')
```




```
o1 = order(loadings[,1])
colnames(Z)[head(o1,25)]
```

```
##  [1] "religion"         "food"             "parenting"      
##  [4] "sports_fandom"    "school"           "family"         
##  [7] "beauty"           "crafts"           "cooking"        
## [10] "fashion"          "photo_sharing"    "eco"            
## [13] "computers"        "chatter"          "outdoors"       
## [16] "personal_fitness" "business"         "shopping"       
## [19] "automotive"       "politics"         "sports_playing" 
## [22] "news"             "health_nutrition" "music"          
## [25] "small_business"
```

```r
colnames(Z)[tail(o1,25)]
```

```
##  [1] "cooking"          "fashion"          "photo_sharing"   
##  [4] "eco"              "computers"        "chatter"         
##  [7] "outdoors"         "personal_fitness" "business"        
## [10] "shopping"         "automotive"       "politics"        
## [13] "sports_playing"   "news"             "health_nutrition"
## [16] "music"            "small_business"   "travel"          
## [19] "home_and_garden"  "dating"           "current_events"  
## [22] "art"              "tv_film"          "college_uni"     
## [25] "online_gaming"
```

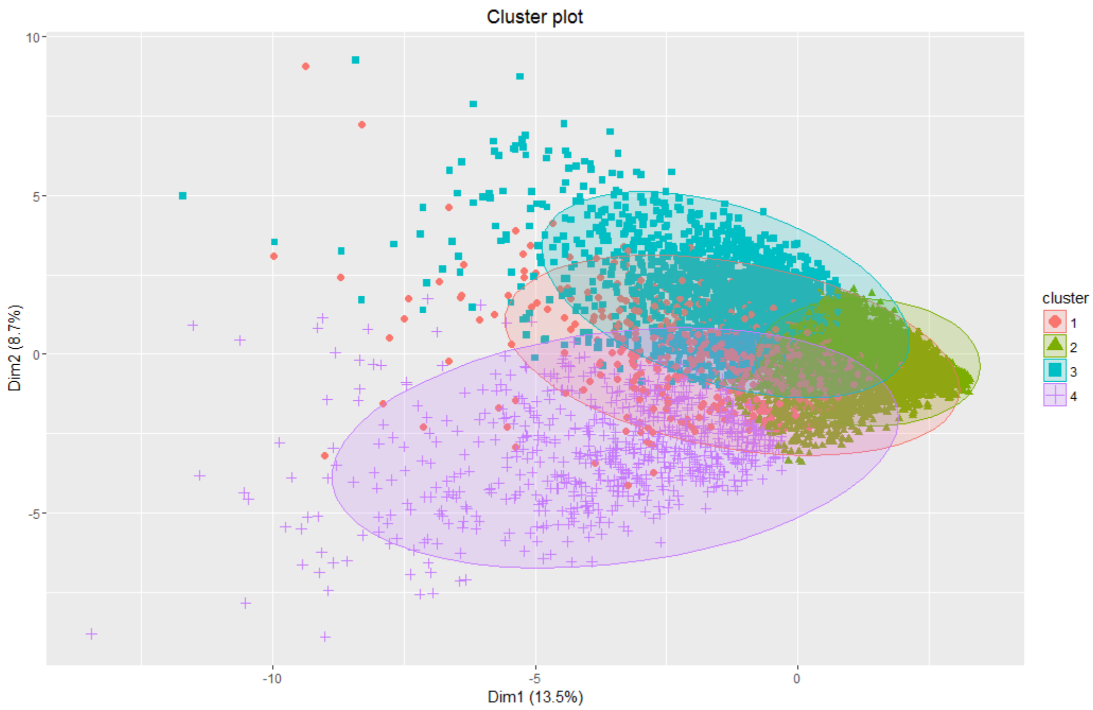
The Principal Component analysis gave us almost 6 of significant components. Substantial information could not be extracted from 4-6 components to infer the cluster composition, or to understand why certain set of groups scored high on one or more of the PCs.

### K-means

```r
## try kmeans
library(factoextra)
library(cluster)
library(NbClust)

sm_wo_id_scaled <- scale(sm_wo_id, center=TRUE, scale=TRUE)
```

We used K-means to get clusters varying between 4-7 and understand cluster composition.

```r
# K-means clustering with 4
set.seed(1)
km.res <- kmeans(sm_wo_id_scaled, 4, nstart = 25)

# Visualize k-means clusters
fviz_cluster(km.res, data = sm_wo_id_scaled, geom = "point",
             stand = FALSE, frame.type = "norm")
```

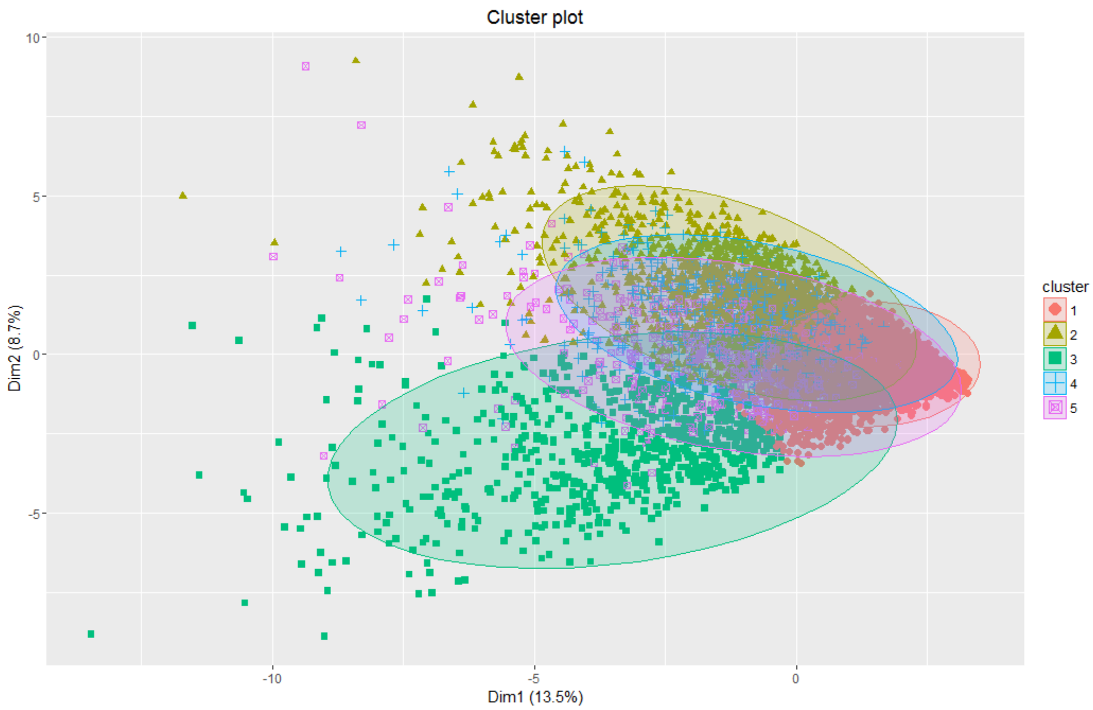



```
clusters_pars = km.res$centers
transposed = t(clusters_pars)
cluster_1 = transposed[which(abs(transposed[,1])>=0.5),1]
cluster_2 = transposed[which(abs(transposed[,2])>=0.5),2]
cluster_3 = transposed[which(abs(transposed[,3])>=0.5),3]
cluster_1

##     travel   politics       news  computers automotive
##   1.763153   2.369479   1.930022   1.546314   1.118160

cluster_2

## numeric(0)

cluster_3

##          chatter     photo_sharing          shopping health_nutrition
##        0.6362693         0.8200723         0.6006349        0.6698805
##          cooking          outdoors            beauty personal_fitness
##        0.8862218         0.5394830         0.6573305        0.6821269
##          fashion
##        0.7696686
```

**Trying with 5 clusters:**

```
# K-means clustering with 5
set.seed(1)
km.res <- kmeans(sm_wo_id_scaled, 5, nstart = 25)
```

```r
# Visualize k-means clusters
fviz_cluster(km.res, data = sm_wo_id_scaled, geom = "point",
             stand = FALSE, frame.type = "norm")
```




```r
clusters_pars = km.res$centers
transposed = t(clusters_pars)
cluster_1 = transposed[which(abs(transposed[,1])>=0.5),1]
cluster_2 = transposed[which(abs(transposed[,2])>=0.5),2]
cluster_3 = transposed[which(abs(transposed[,3])>=0.5),3]
cluster_4 = transposed[which(abs(transposed[,4])>=0.5),4]
cluster_5 = transposed[which(abs(transposed[,5])>=0.5),5]

cluster_1

## numeric(0)

cluster_2

##       chatter photo_sharing         music      shopping   college_uni
##     0.8043252     1.0644587     0.5610928     0.7707609     0.5933295
##       cooking        beauty       fashion
##     0.8638999     0.8555066     0.9585076

cluster_3

## sports_fandom          food        family        crafts      religion
##     2.0495745     1.8226930     1.4947438     0.7068447     2.2515579
##     parenting        school
##     2.1181399     1.6614164
```

```
cluster_4

## health_nutrition              eco         outdoors personal_fitness
##        2.1591362        0.5169836        1.6686893        2.1239862

cluster_5

##     travel   politics       news  computers automotive
##   1.842251   2.429340   1.935524   1.637643   1.077207
```

The cluster composition we got from five clusters makes sense intuitively and is interpretable. In order to visualize these clusters and understand their prominent characteristics, we used word cloud.

```
# A word cloud
par(mfrow=c(2,2))
library(wordcloud)

## Loading required package: RColorBrewer

for (i in 2:5) {
wordcloud(colnames(sm_wo_id_scaled), km.res$centers[i,], min.freq=0,
max.words=100, scale=c(4,.4))
}
```

## Conclusion

NutrientH20 can use the clusters that form above to get a better idea about their customers. A simple aggregation of personalities would be to call them *active on social media and young*, *Community and family-minded*, *fit and community minded* and *gadget-savvy and well-travelled*.

This can be used to better target that population of this customers with relevant advertisements so as to pique more interest in their products.