STAT380_Exer1

Anthony Garino, Anqi Huang, Olivia Hong, Yun Guo

Summer 2016

Probability practice

Variables:

RC = Random clicker; TC = truthful clicker; Y = answered yes

Probabilities:

P(RC) = 0.30

P(TC) = 0.70

P(Y) = 0.65

P(Y|RC) = 0.50

Used a function and set P(Y|TC) = X and used a function to solve for X. The equation was the Law of Total Probability P(Y) = P(Y|TC)P(TC) + P(Y|RC)P(RC)

0.65 = P(Y|TC)(0.70) + (0.50)(0.30)

$P(Y|TC) = 0.714$

Fraction of truthful clickers who answered 'yes' - 0.714 = 5/7

Part B.

Variables:

W = test positive

PNND = test negative

PD = has disease

PND = does not have disease

Probabilities:

P(P|D) = 0.993

P(N|ND) = 0.9999

P(D) = 0.000025

P(P|ND) = 1-P(N|ND) = .0001

Using Bayes Rules

P(P) = P(P|D)P(D) + P(P|ND)P(ND)

P(P) = (0.993)(0.000025) + (.0001)(0.999975)

P(P) = 0.00002482 + 0.0001 = 0.00012482 P(P) = Z Z = 0.000124815

$P(D|P) = \frac{P(D)*P(P|D)}{P(P)}$

$P(D|P) = \frac{0.000025*0.993}{.00012482}$

$P(D|P) = 0.1988944$

The probability that someone has the disease given that he/she tested positive is almost 20%. This is a problem because testing positive and not having the disease is much more likely than actually having the disease.

Exploratory analysis: green buildings

First, in the conclusion of the developer's on staff stats guru, he found that there are $2.6 differences for the medium in the market per square food per year between green and non-green buildings. Without any supportive evidence, he attributes all the difference to being "green", so he concludes that the rent will be $2.6 higher per square foot if it is a green building. And all the rest of the pay-off calculations are based on this spurious assupmtion.

So there may be some other factors that have big influence and they all together cause a higher market rent, and being a "green building" can be just part of the reason for the rent difference.

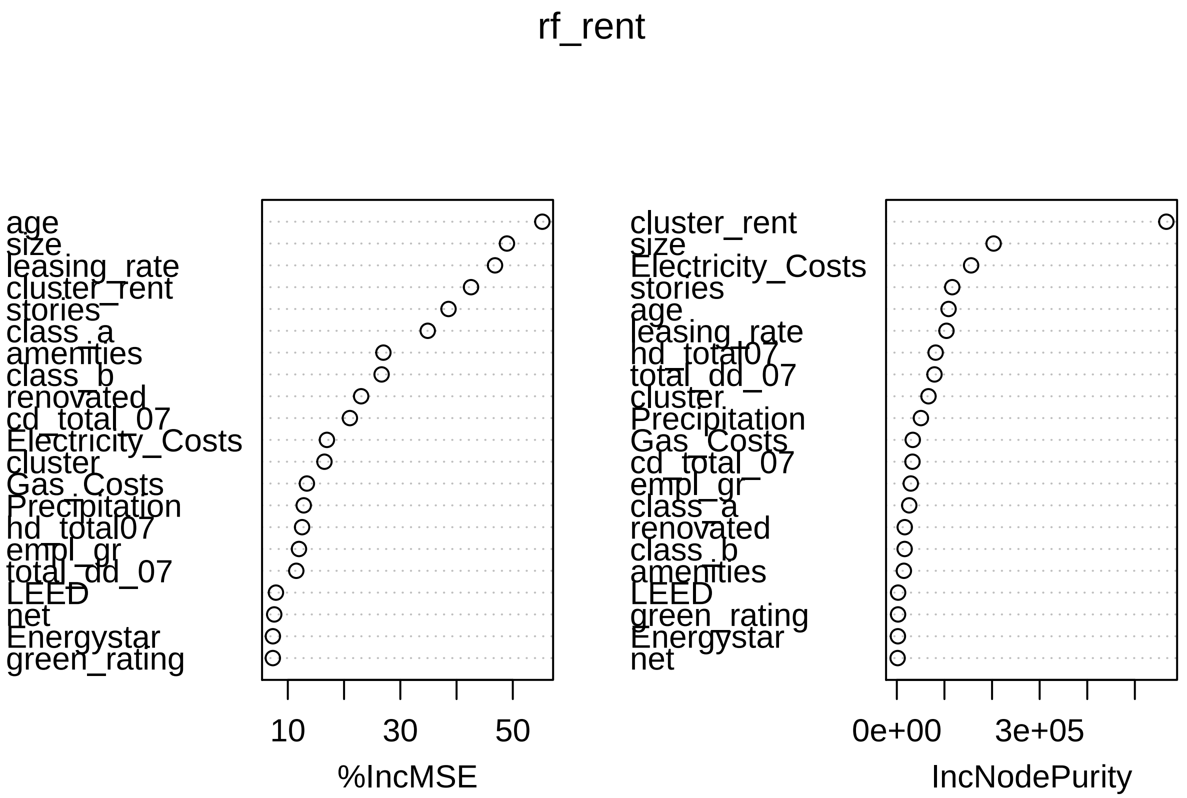
To find out whether certain factors are more important in deciding the rent per square foot per year, we first try random forest model. Because the Property ID, on common sense, is assigned by estate officers and it shouldn't have any influence on rent, we didn't include it in our predictors.

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
## [1] "Variable Importance\n"
```

```
##                      %IncMSE IncNodePurity
## cluster            16.510832     66457.761
## size               48.957949    203332.072
## empl_gr            11.981896     29270.602
## leasing_rate       46.839248    104246.959
## stories            38.569202    116389.417
## age                55.247140    108634.010
## renovated          23.039761     16641.777
## class_a            34.877969     26046.876
## class_b            26.672740     16401.471
## LEED                7.887412      2656.405
## Energystar          7.343951      2241.070
## green_rating        7.336426      2540.435
## net                 7.582124      1650.461
## amenities          26.989707     14772.467
## cd_total_07        21.034389     32850.015
## hd_total07         12.536314     81500.198
## total_dd_07        11.503805     79047.819
## Precipitation      12.829526     50592.002
## Gas_Costs          13.389150     33478.603
## Electricity_Costs  16.952469    156035.616
## cluster_rent       42.570127    566089.372
```



Looking at the results and plot of feature importance, we will find that the "age", "size" and "leasing_rate" is actually the factors that have the biggest decision power of the rent per square feet. And "green_rating", "LEED" and "Energystar" are least important features. Here, we can conclude that whether the building is "green" or not isn't the reason or the main reason which caused $2.6 higher in rent per square foot per year.

And the results shown above can still explain the phenomenon that "green buildings" often have higher rent. Because, in practical, usually green buildings, as a pretty new and contemporary concept that just put into use recent years, are younger than others thus have smaller age and bigger size, the newer and bigger sized buildings are supposed to lead to a higher rent. So we need to note that, it is the age, size and leasing rate that contribute most to a higher rent for a green building, not being "green".

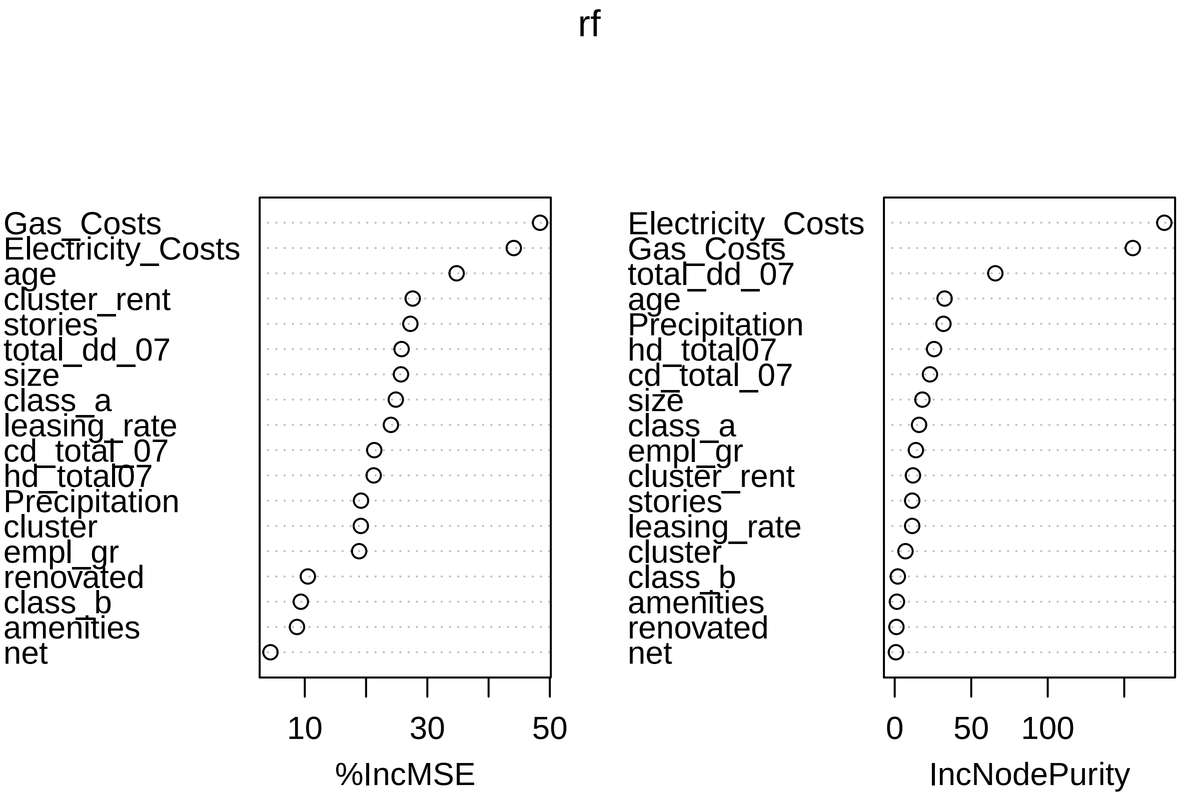
Then in order to see if there any confounding variables for the relationship between rent and green status, we ran the random forest model again and set the "green_rating" as the target.

Because we know that "Energystar" and "LEED" are the factors that evaluate the green rating, besides removing "Property ID", we also removed those two features to avoid noise.

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
## [1] "Variable Importance\n"
```

```
##                     %IncMSE IncNodePurity
## cluster           19.145861     7.0224710
## size              25.697680    18.0895523
## empl_gr           18.869300    13.8383100
## leasing_rate      24.052997    11.4229330
## stories            27.233398    11.4379394
## age                34.782131    32.5867572
## renovated          10.498560     1.1094654
## class_a            24.855305    15.9089316
## class_b             9.352955     2.0187529
## net                 4.392168     0.7863716
## amenities           8.723475     1.4346874
## cd_total_07        21.336267    23.0332656
## hd_total07         21.242572    25.7231138
## total_dd_07        25.798900    65.7495719
## Precipitation      19.173736    31.8460524
## Gas_Costs          48.399176   155.6048538
## Electricity_Costs  44.111427   176.1843201
## cluster_rent       27.614056    11.9106258
```



We can find that, from the results and plot, gas costs and electricity costs are the two most important factors that contribute to a building's green rating. That is to say, the gas and electricity costs can greatly influence whether the building should obtain a green certificate.

Now we look back to the previous plot which shows the feature importance for "rent". Gas costs and electricity costs are also two factors that have much more importance than "green status" when deciding the rent per square foot. Thus, we can have a reasonable guess that a higher rent and being "green" are both the consequence of higher gas and electricity costs, that is to say, higher rent isn't come from a high green rate.

In conclusion, the recommendation in the question is unacceptable. Even if the new building that they are working on isn't going to be green, the rent will be somehow higher than the median market rent square per feet because it's new and most likely it will have bigger size than the older buildings. And to be honest, the extra 5% premium for green certification may not lead to sizeable difference in rent or only a very small difference.

In order to figure out how the green status will have impact on the rent, a multi-regression model may help. Because from random forest, green_rating is already the least important factors for rent, we include almost all variables in the regression model and only remove the Property ID column.

```
##
## Call:
## lm(formula = Rent ~ ., data = green)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.723  -3.564  -0.516   2.489 174.033
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -8.277e+00  1.018e+00  -8.130 4.96e-16 ***
## cluster            7.697e-04  2.839e-04   2.712 0.006709 **
## size               6.756e-06  6.561e-07  10.296  < 2e-16 ***
## empl_gr            6.153e-02  1.693e-02   3.635 0.000280 ***
## leasing_rate       8.710e-03  5.318e-03   1.638 0.101480
## stories           -3.537e-02  1.617e-02  -2.188 0.028725 *
## age               -1.220e-02  4.715e-03  -2.587 0.009710 **
## renovated         -2.042e-01  2.565e-01  -0.796 0.426143
## class_a            2.860e+00  4.377e-01   6.535 6.76e-11 ***
## class_b            1.169e+00  3.427e-01   3.412 0.000648 ***
## LEED               1.863e+00  3.583e+00   0.520 0.603063
## Energystar        -2.683e-01  3.818e+00  -0.070 0.943974
## green_rating       7.508e-01  3.839e+00   0.196 0.844961
## net               -2.574e+00  5.930e-01  -4.341 1.43e-05 ***
## amenities          6.169e-01  2.503e-01   2.465 0.013740 *
## cd_total_07       -1.179e-04  1.464e-04  -0.805 0.420768
## hd_total07         5.474e-04  8.951e-05   6.115 1.01e-09 ***
## total_dd_07               NA         NA      NA       NA
## Precipitation      4.426e-02  1.597e-02   2.771 0.005608 **
## Gas_Costs         -3.233e+02  7.650e+01  -4.227 2.40e-05 ***
## Electricity_Costs  1.893e+02  2.493e+01   7.592 3.51e-14 ***
## cluster_rent       1.005e+00  1.408e-02  71.372  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.414 on 7799 degrees of freedom
## Multiple R-squared:  0.6124, Adjusted R-squared:  0.6114
## F-statistic: 616.2 on 20 and 7799 DF,  p-value: < 2.2e-16
```

So from the result of the multi-regression, it is easy to find that the coefficient for green_rating is only ~$0.7, not $2.6 mentioned in the question. That is to say, holding everything else constant, with the green certificate, the rent will only be $0.7 higher than a non-green building.

If the building has 250,000 square foot, then they will generate 250,000 * 0.7 = $175,000 of extra revenue per year if they build the green building. And because there are around 5% premium for a green certificate, which is 5% * 100 million = $5,000,000, they will need 5,000,000 / 175,000 = 28.57 years to recuperate these costs. However, we still need to note that, from the multi-regression, the green_rating is rejected at all statistically significant level, which means people would better not conclude any impact of green status on rent.

```
## [1] 415
```

And looking at the leasing rate of all the green building, 415 out of 685 (60.59%) have more than 90% leasing rate. So we believe they are reasonable to assume that the leasing rate for the new building can exceed 90%. So when the leasing rate is around 90%, they probably need more than 30 years to cover the premium on green certificate.

Instead of only considering green building can lead to a higher rent, we will recommend the developer to look at the gas and electricity costs of the building location and see if there is any need to get a green certificate. From the previous analysis, these costs are actually the main points when decide whether to be “green”.

```
## [1] 3424
```

Then let’s take a look of the INDIRECT impact of green status on rent.

For non-green buildings, 3424 out of 7209 (47.5%) has leasing rate more than 90%, which is lower than the 60.59% for green buildings. So as the question has described, there are some benefit of the leasing rate if the building has a green certificate. From the previous analysis, leasing rate is a very important factor that can contribute to rent.

```
## [1] 433
```

433 / 685 = 63.2%

```
## [1] 3350
```

3350 / 7209 = 46.5%

So the same thing happens to “empl.gr”: the year-on-year growth rate in employment in the building’s geographic region. The building with green certificate is more likely to have a higher year to year growth rate in employment, which should be desired by most of the companies and thus increase the rent accordingly.

Combining both the direct and indirect impact “green rating” on the rent, the actual time to cover the 5% of premium for green certificate should be shorter than 30 years as if we only consider the direct influence on green but still much longer than 8 years as the recommendation in question.

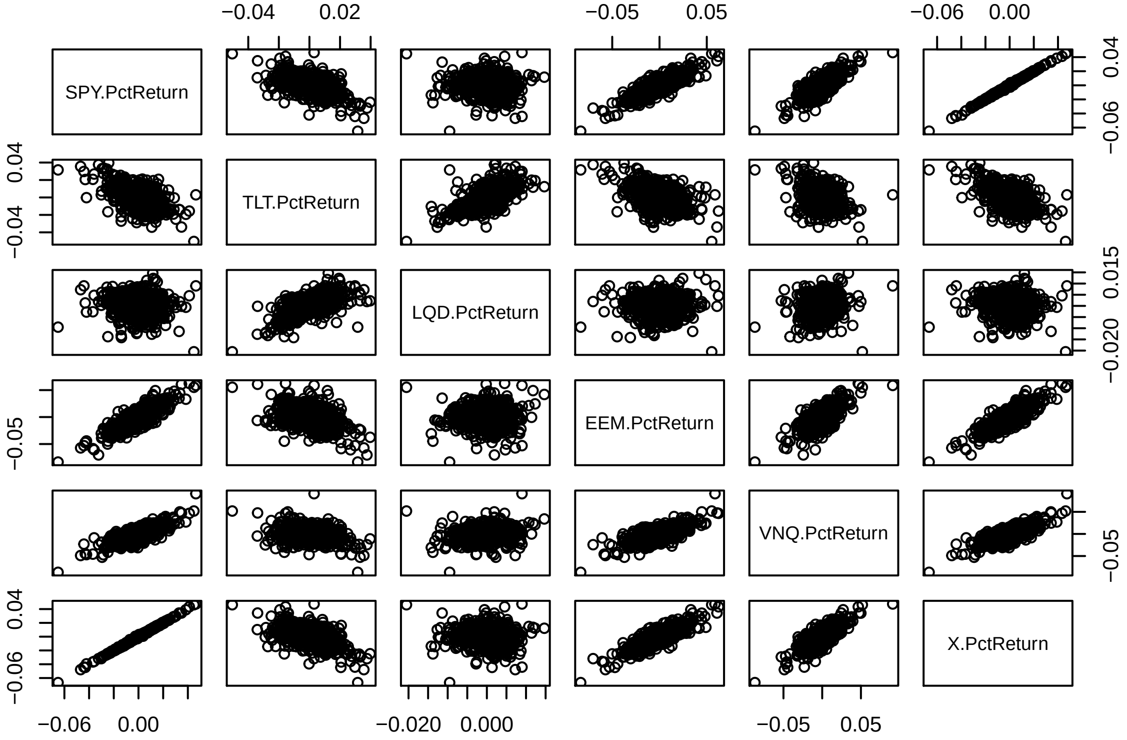
Bootstrapping

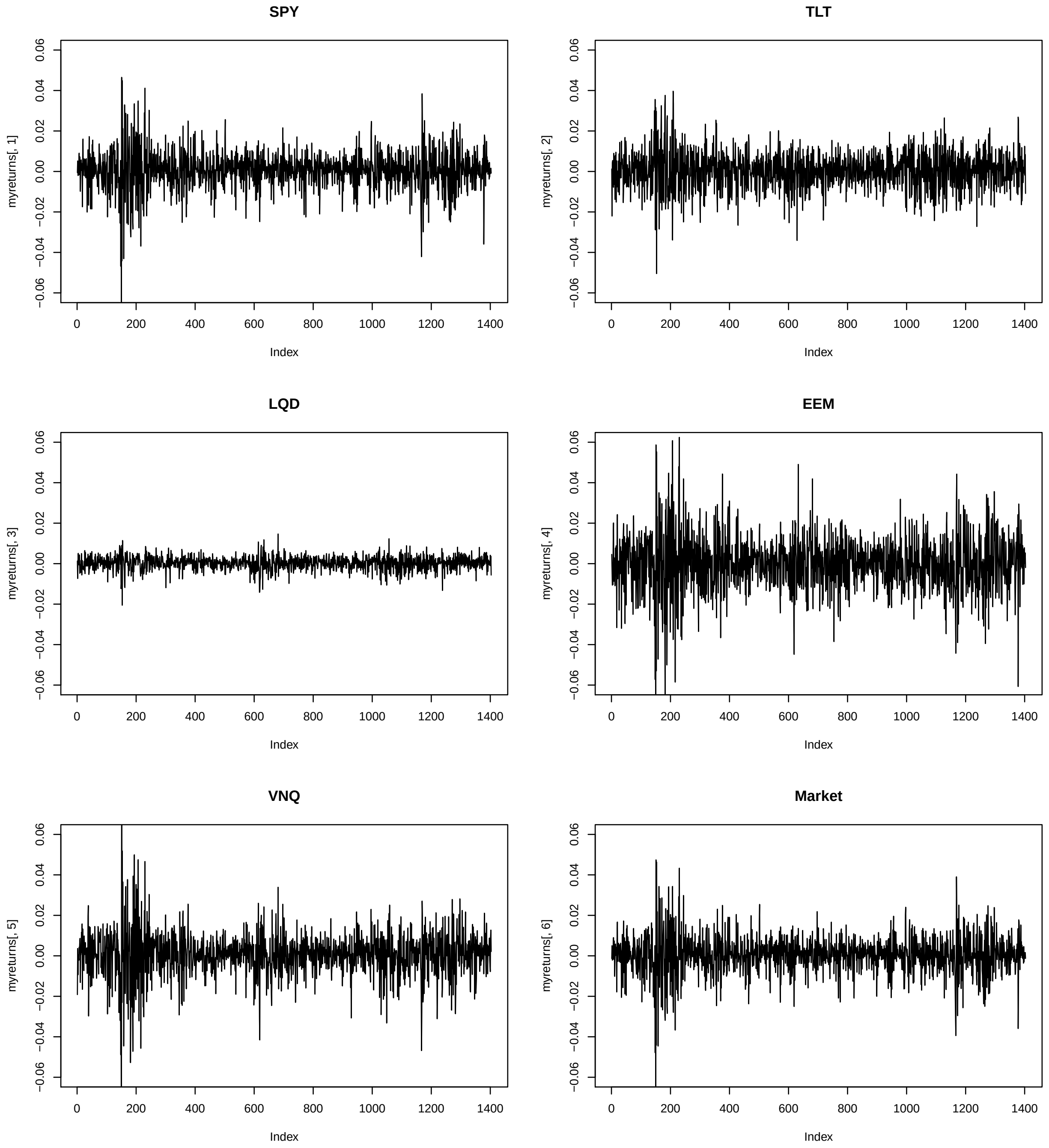
Given five asset classes, we can create portfolios to explore different combinations of assets based on their return and risk. We will look at the returns for the time period between January 2011 and August 2016, so a span of more than 5 years. The five classes are:

- US domestic equities (SPY: the S&P 500 stock index)
- US Treasury bonds (TLT)
- Investment-grade corporate bonds (LQD)
- Emerging-market equities (EEM)
- Real estate (VNQ)

The S&P 500 will represent the market.

Exploring the data





From the pairwise plot, there are some positive correlations, such as those between TLT and LQD, SPY and EEM, SPY and VNQ. So the porfolio will have more ups and downs as a result. We can also look at the returns over time. The graphs of returns show that they all roughly center around 0, with EEM having a much larger range of ups and downs than ETFs like LQD.

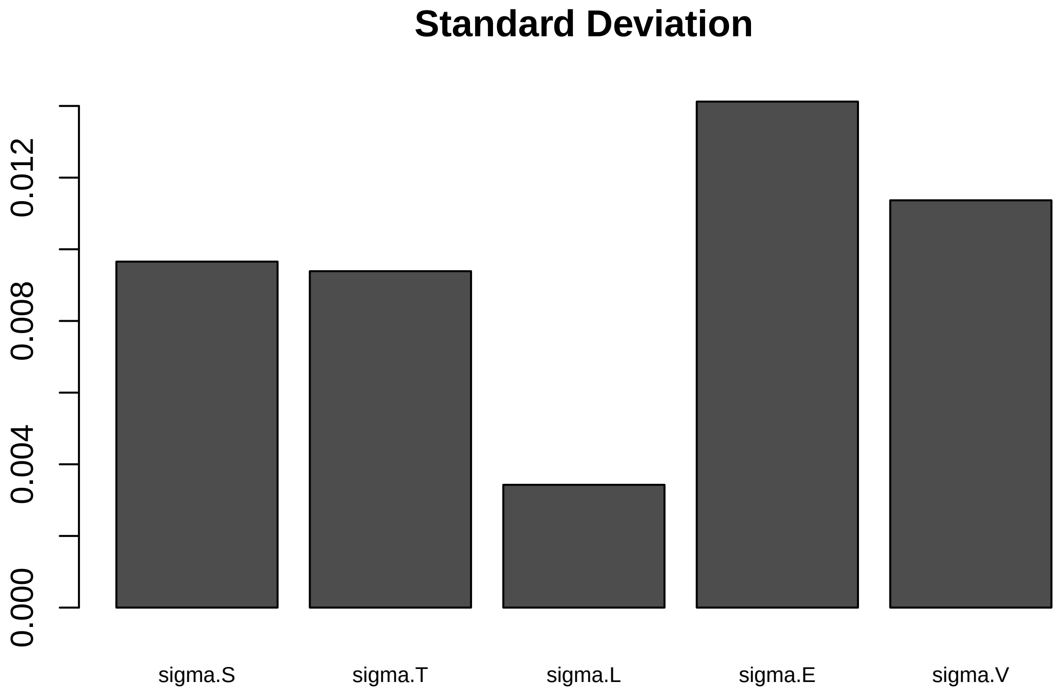
Next, we calculate the beta from the CAPM model to measure the volatility of an asset compared to the whole market. From this we see that all classes except for EEM have a beta less than 1. This means that they are less volatile than the market. Taking the standard deviation of the returns also indicates something similar; EEM has the highest SD while LQD has the lowest.

```
## SYP,  8.339909e-05 0.9923535
## TLT,  0.0006698393 -0.5205087

## LQD,  0.0002578101 -0.04262795

## EEM,  -0.0005446485 1.216288

## VNQ,  0.0001886721 0.8908399
```



For a safe portfolio, we can choose mostly index-tracking funds, U.S. treasury bonds (that may protect a portfolio against a stock market crash/recession but may have lower yields), and low-volatility ETFs. One option is SPY, TLT, and LQD (chosen for its low volatility). We can use a higher weighting on SPY and TLT (e.g. 40% each), with 20% for LQD.

For an aggressive portfolio, we can choose EEM which has a beta greater than 1, and SPY, with 50% each.

Bootstrap resampling

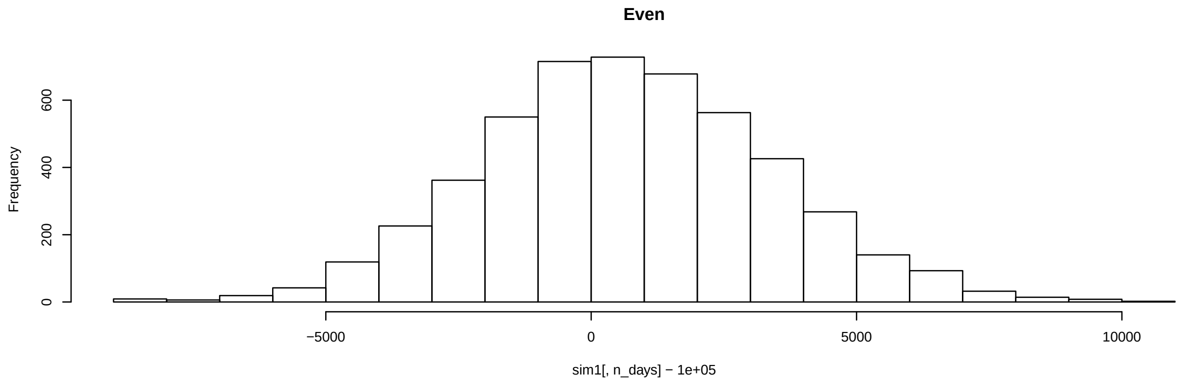
We use the bootstrap to estimate the 4-week value at risk for each portfolio at the 5% level, using 5000 Monte Carlo simulations.

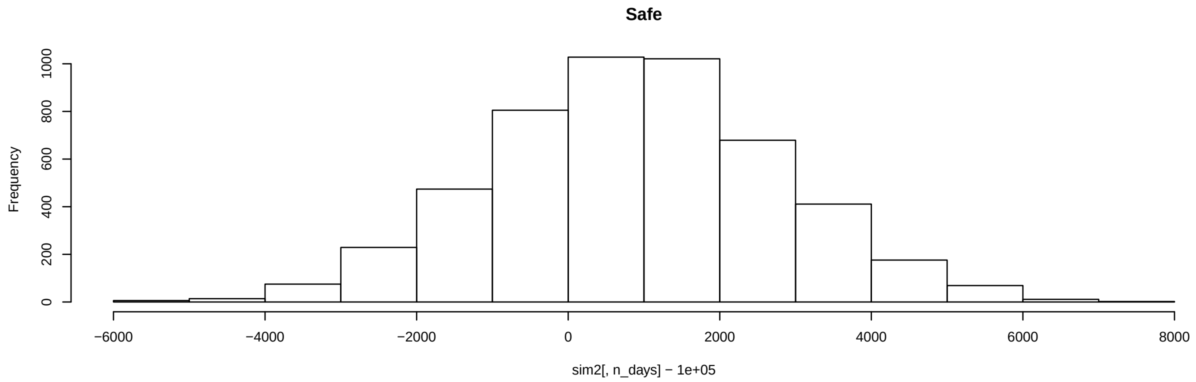
The value at risk for each is as follows:

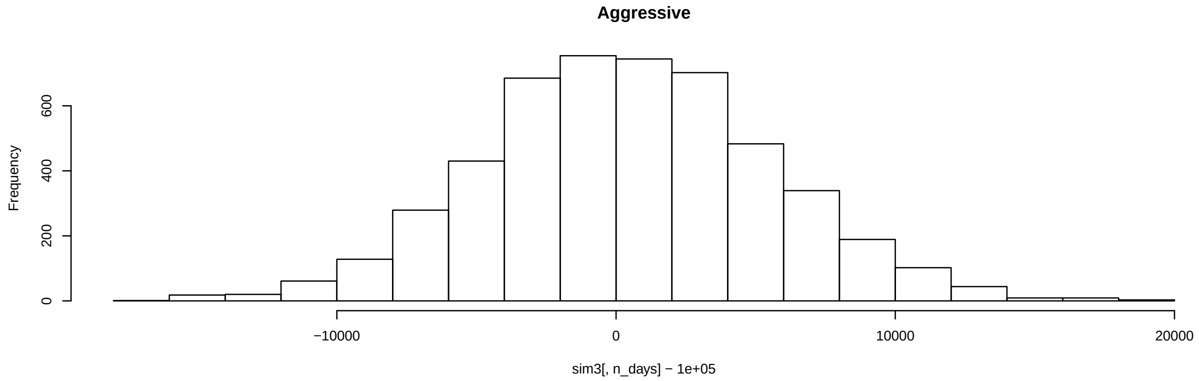
- Even split: -3697.652
- Safe: -2248.11
- Aggressive: -7814.795

So the aggressive portfolio could have the largest loss from the initial investment.

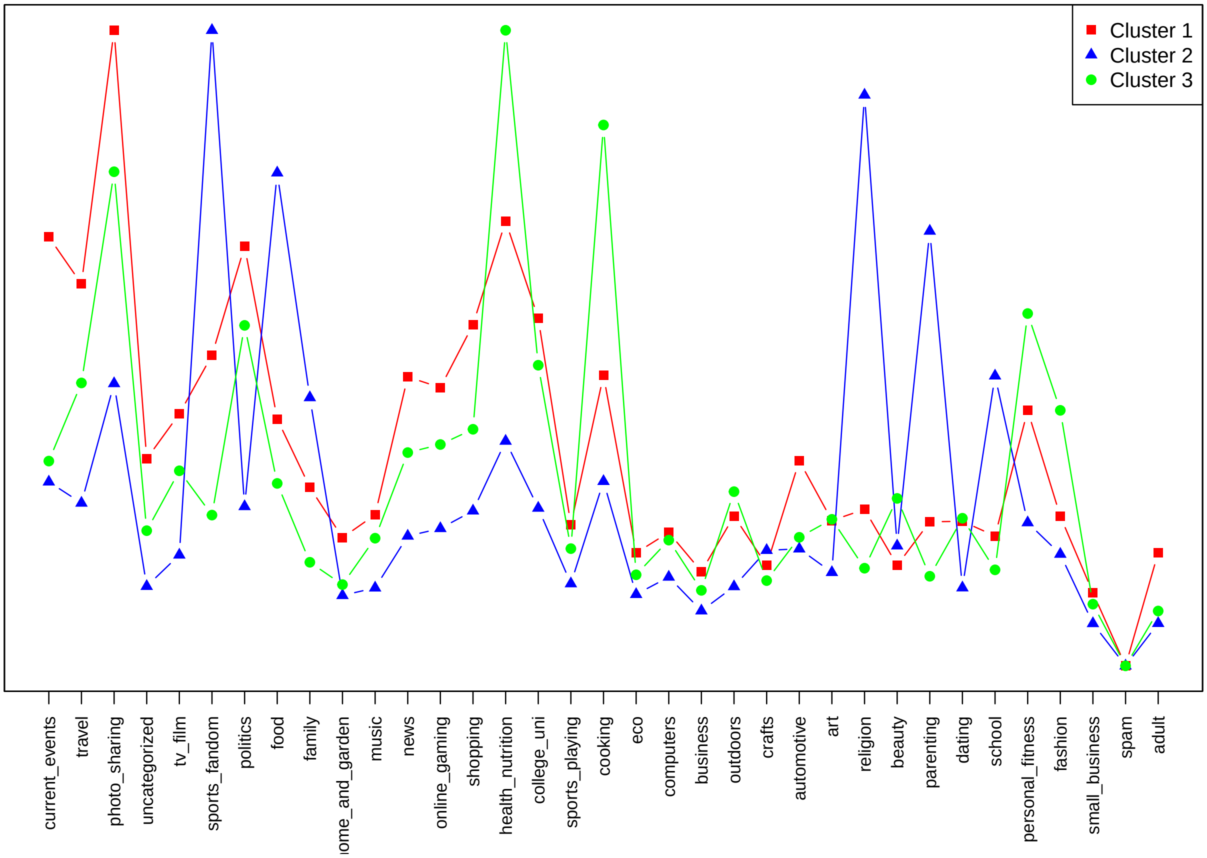
We can also look at the histogram for the profit and loss distribution in 4 weeks. From this, the aggressive portfolio has the largest spread. It could result in a large yield, but generally it would be better to use a less volatile method. The even and split portfolios both have slightly more weight on the profit end with similar levels of uncertainty. The conservative investor may want to choose one of these two methods.







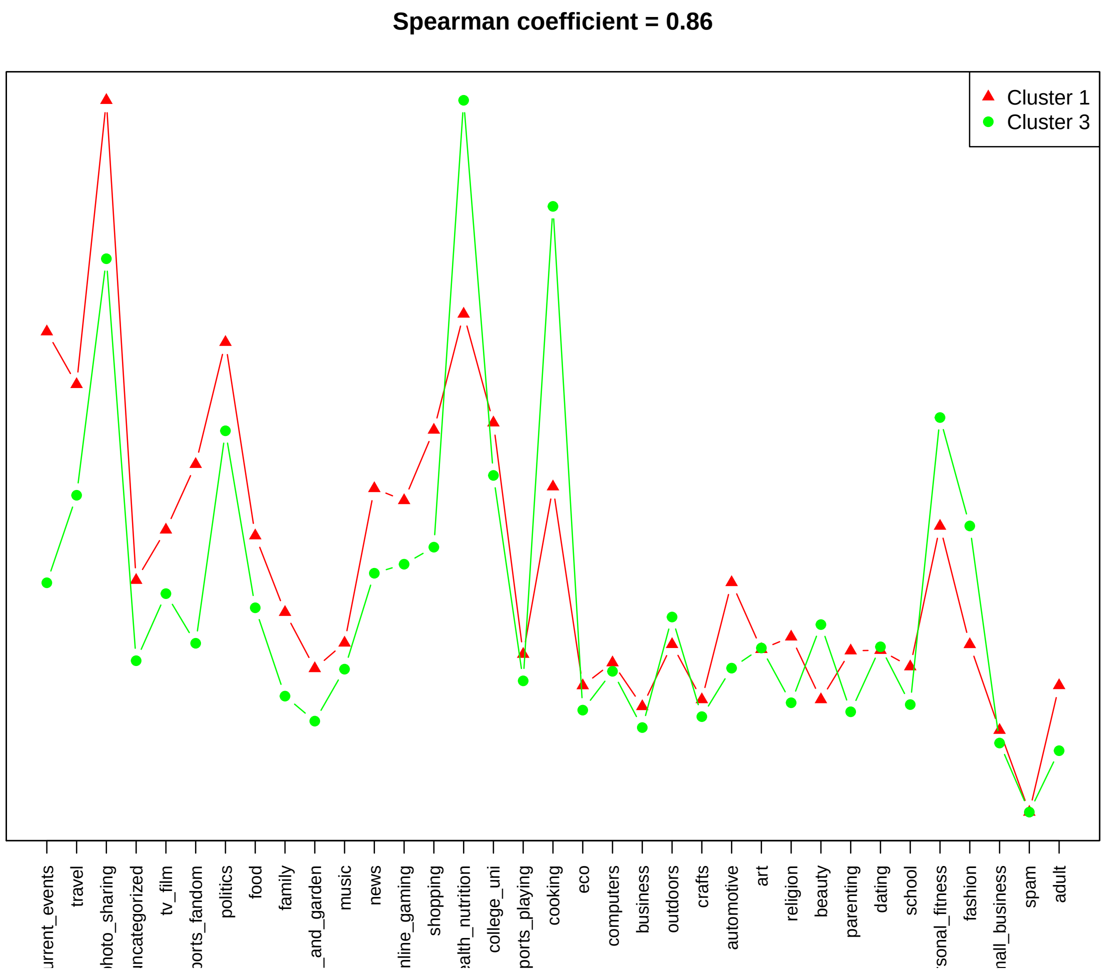
Market Segmentation



Obviously, there are a few 'crest' topics within each cluster. Cluster 2 contains keywords such as "sports_fandom", "food", "religion"... This is reminiscent of a male clientele.

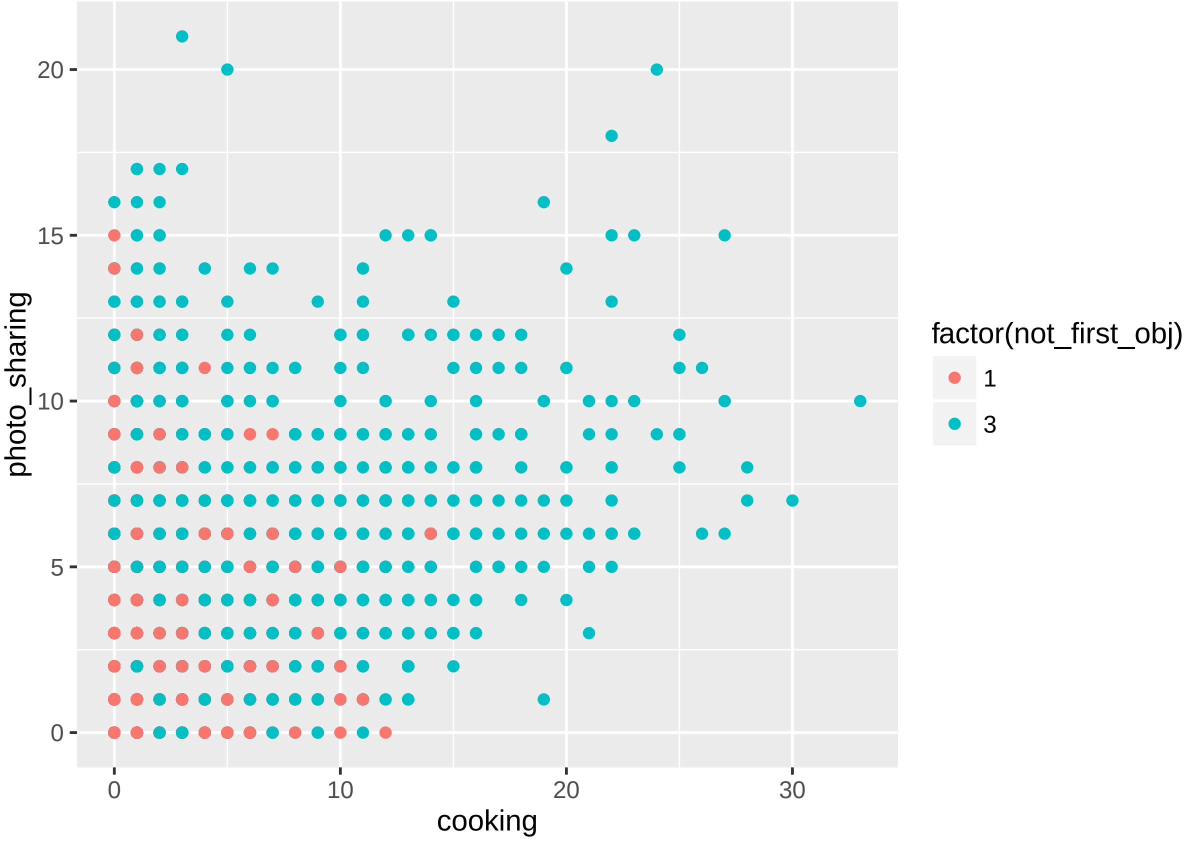
However, the remaining two groups (Clusters 1 and 3) show a high level of consistency in terms of term frequency. They are concerned about homogeneous subjects such as "photo_sharing", "health_nutrition", "cooking" - a reasonable assumption can be made here that they actually form a single distinctive consumer base - females.

Next we try to see if Clusters 1 and 3 can be collapsed into one.



The Spearman correlation between Cluster 1 and 3 term frequency is ~0.86, indicating a high inter-relatedness. And since most of the keywords are home/beauty-related, we can perhaps treat them as one cluster.

Next we try to see if Clusters 1 and 3 can be collapsed into one.



Finally, when we plot the top topic for both cluster 1 and 3, we can see that cluster 3 is high on the cooking dimension, while cluster 1 tends to concentrate on the photo_sharing (beauty) aspect (quite low on cooking). This likely points to the possibility that cluster 1 represents younger portion of the female customer body.

In conclusion, we have discovered 1 main market segmentation and and 2 niches: Cluster 1 consists primarily of younger-aged females (5041 people); Cluster 2 consists primarily of males (822 people); Cluster 3 consists primarily of middle-aged females (2019 people).

It would serve NutrientH20 well to tap into the consumer cluster corresponding to its targeted market segment strategy.