STA 380 - Exercise 1

Dallas Griffin, Estevan Gonzalez, Sean Kessel

August 8, 2016

Probablity Practice

Part A - Truthful Clickers

$$P = \frac{.65 - (.3/2)}{1 - .3}$$

```
## By the Law of Total Probability:

## The fraction of truthful yes responces is  0.7143
```

Part B - Medical Testing

$$P(HaveDisease|Test+) = \frac{P(HaveDisease) * P(Test + |HaveDisease)}{P(Test+)}$$

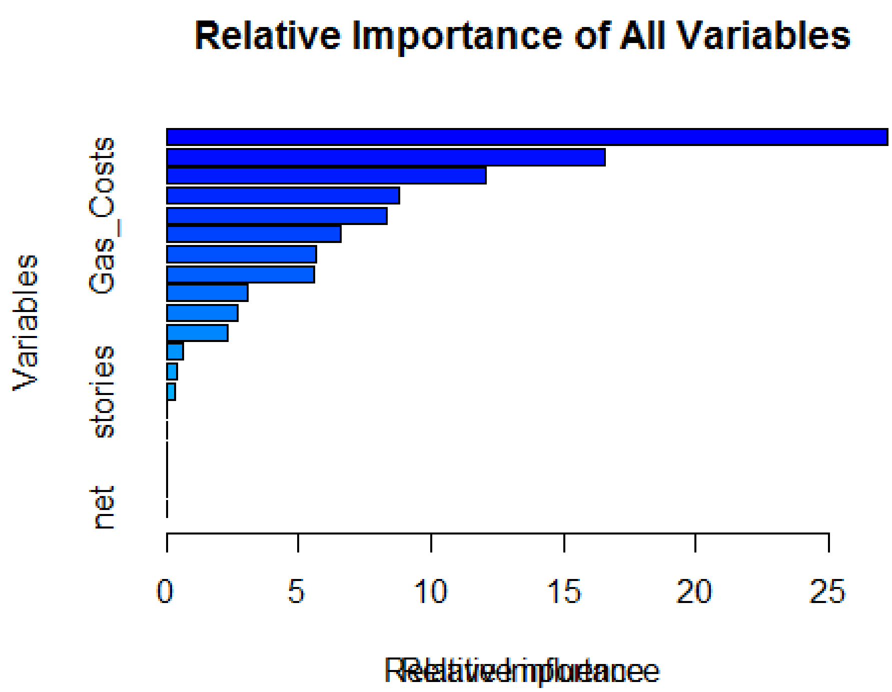
$$P(Test+) = P(Test + |HaveDisease) * P(HaveDisease) + P(Test + |DoNotHaveDisease) * P(DoNotHaveDisease)$$

```
## According to Bayes' Law:
##  The probability of the patient having the disease given a positive test 
is: 2.48e-05
##  Therefor I would not recommend implementing a universal testing policy.
```

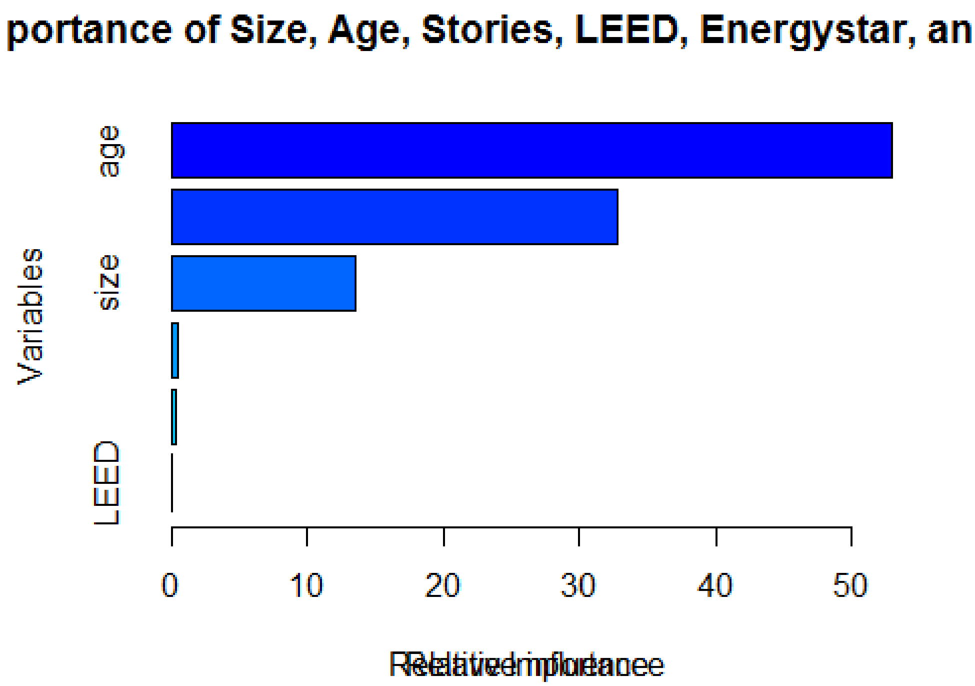
Exploratory Analysis: Green Buildings

```
## Loading required package: survival

## Loading required package: splines

## Loading required package: parallel

## Loaded gbm 2.1.1

##                                   var    rel.inf
## Electricity_Costs Electricity_Costs 27.2126630
## cluster                       cluster 16.5494701
## total_dd_07               total_dd_07 12.0083581
## empl_gr                       empl_gr  8.8094434
## Gas_Costs                   Gas_Costs  8.3383463
## Precipitation           Precipitation  6.5994642
## leasing_rate             leasing_rate  5.6772572
## cd_total_07               cd_total_07  5.5446126
## class_a                       class_a  3.0437545
```

```
## age                                       age  2.6815338
## size                                     size  2.2653595
## hd_total07                         hd_total07  0.6103942
## amenities                           amenities  0.3755243
## stories                               stories  0.2838189
## renovated                           renovated  0.0000000
## class_b                               class_b  0.0000000
## LEED                                     LEED  0.0000000
## Energystar                         Energystar  0.0000000
## green_rating                     green_rating  0.0000000
## net                                       net  0.0000000
```




```
##                           var    rel.inf
## age                       age 53.0036273
## stories               stories 32.7945842
## size                     size 13.4664915
## green_rating green_rating  0.4394801
## Energystar      Energystar  0.2958169
## LEED                     LEED  0.0000000
```




```
## Rent per square foot, with all energy certifications:  28.7672644698202
## Rent per square foot, with no energy certifications:  29.0887191543914
## These rents are nearly identical, and given that all previous models have
shown the insignificance of the green rating, the small variation in the two
predictions is most likely due to randomness in the data.
```

To begin, the boost model shows the relative importance of LEED, Energystar, and the green rating to be nearly zero in determining the rent per square foot.

Then, another boost model was run, this time factoring in only variables that are known going into construction of the building (age, stories, class_a, green rating, EnergyStar, and Leed). Once again, the energy efficient ratings were not significant in determining rent.

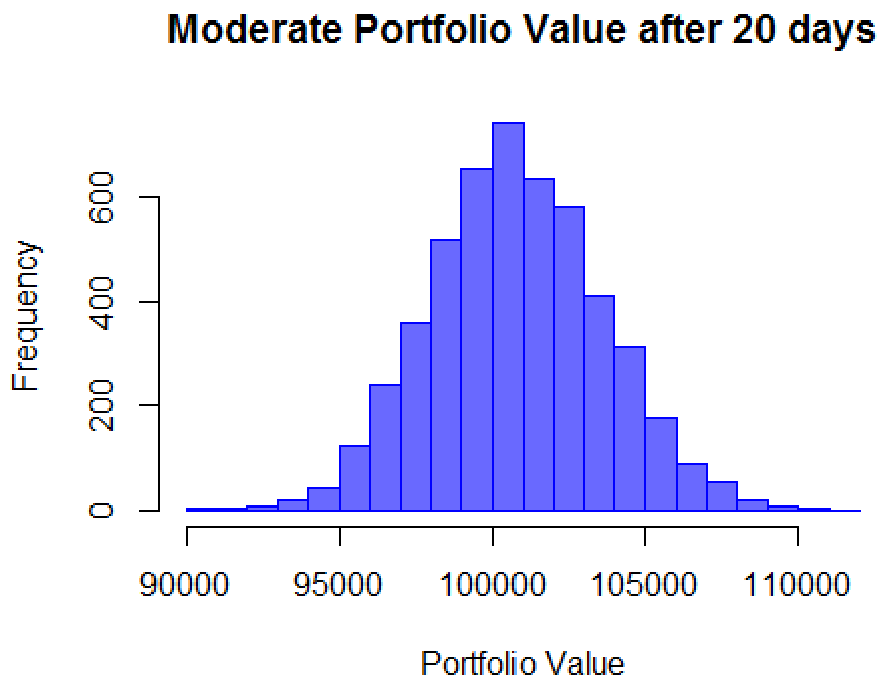
Finally, using the model that only factors in known variables, and plugged in the variables to predict for the model described in the assignment in two different set ups and compared them side-by-side. One set up had a green building, the other was a non-green building. The model predicts very similar rents per square foot, and as such it would not be profitable to add 5% to the building costs for green certification.

The "stats guru" made an error in simply comparing the average rent with green certification and without. As shown below, newer buildings tend to be energy certified, and newer buildings also tend to have a higher rent per square foot. This and many other factors, such as class_a and size are highly correlated with green certification, to an extent with which one can reject the null hypothesis, and these factors are also associated with higher rent. These confounding variables are the reason why the stats guru's logic is misguided.

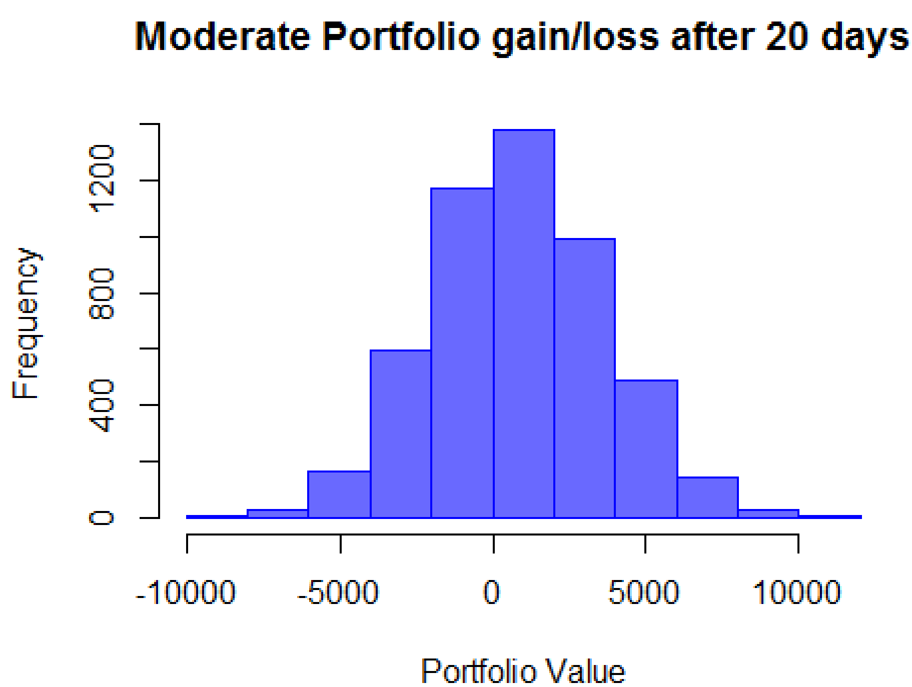
```
##
## Call:
## lm(formula = green_rating ~ . - CS_PropertyID - cluster_rent -
##     LEED - Energystar - Rent, data = data.frame(scale(train_gBuild)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2537 -0.4840 -0.1860  0.0188  3.8250
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.0001455  0.0170874  -0.009  0.99321
## cluster             0.0308647  0.0178108   1.733  0.08321 .
## size                0.0861440  0.0307703   2.800  0.00515 **
## empl_gr            -0.0059867  0.0187641  -0.319  0.74971
## leasing_rate        0.0755625  0.0184503   4.095 4.32e-05 ***
## stories            -0.1575115  0.0306775  -5.134 3.00e-07 ***
## age                -0.0746970  0.0242169  -3.084  0.00206 **
## renovated          -0.0520145  0.0195854  -2.656  0.00795 **
## class_a             0.2351328  0.0340214   6.911 5.80e-12 ***
## class_b             0.0229442  0.0266745   0.860  0.38977
## net                 0.0346994  0.0175969   1.972  0.04871 *
## amenities          -0.0273181  0.0200002  -1.366  0.17207
## cd_total_07         0.0753560  0.0278165   2.709  0.00678 **
## hd_total07         -0.0078814  0.0327506  -0.241  0.80984
## total_dd_07                NA         NA      NA       NA
## Precipitation      -0.0002367  0.0267779  -0.009  0.99295
## Gas_Costs          -0.1055707  0.0381899  -2.764  0.00574 **
## Electricity_Costs -0.0028595  0.0385261  -0.074  0.94084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9536 on 3109 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.09719,    Adjusted R-squared:  0.09254
## F-statistic: 20.92 on 16 and 3109 DF,  p-value: < 2.2e-16
```

Even Split

This section calculates the relative risk for a moderate portfolio. The portfolio represents an even split among 5 asset classes: *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 graph above is a histogram of the 20-day portfolio value after being randomly calculated 5000 times.



```
## $breaks
##  [1] -10000  -8000  -6000  -4000  -2000      0   2000   4000   6000   8000
## [11]  10000  12000
##
## $counts
##  [1]    4   26  163  599 1173 1379  992  488  142   29    5
##
## $density
##  [1] 0.0000004 0.0000026 0.0000163 0.0000599 0.0001173 0.0001379 0.0000992
##  [8] 0.0000488 0.0000142 0.0000029 0.0000005
##
## $mids
##  [1] -9000 -7000 -5000 -3000 -1000  1000  3000  5000  7000  9000 11000
##
## $xname
## [1] "sim1[, n_days] - 1e+05"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

Here is the same graph after subtracting the original investment amount, leaving a histogram of gains and losses.

```
##        5%
## -3740.174
```

The value at risk at 5% for this equaly-rated protfolio is -$3,886.

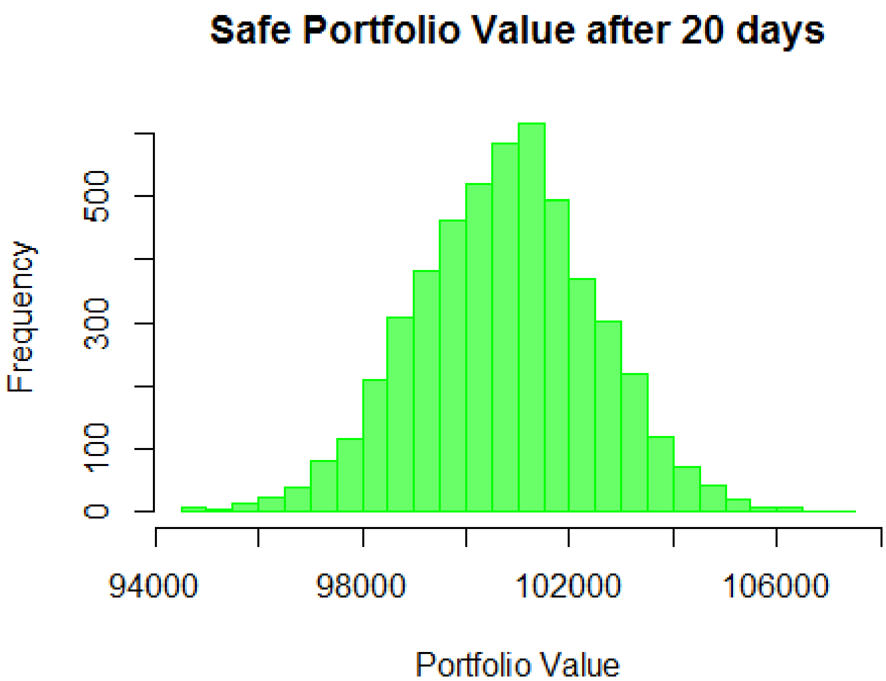
Safe Split

This portfolio is a safe one, comprised of safe assest that offer low risk, at the expense of less reward.

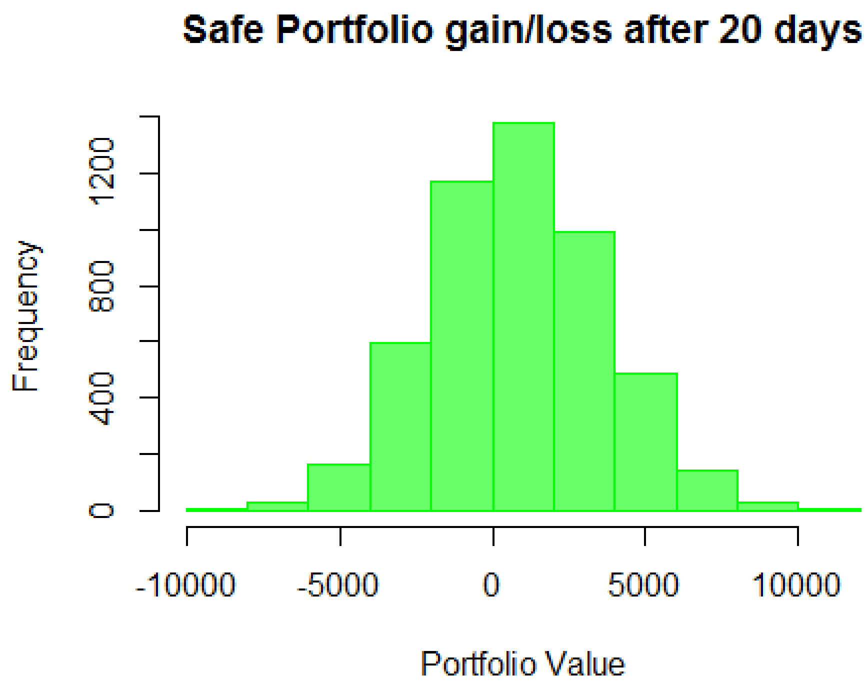
```
sigma_SPY

## [1] 0.009913704

sigma_TLT

## [1] 0.009504

sigma_LQD

## [1] 0.003489003

sigma_EEM

## [1] 0.01443178
```

```
sigma_VNQ

## [1] 0.01239193
```

Based on the stocks' standard deviations (volatility), SPY, TLT and LQD are the "safest" stocks. Weights of .4(LQD), .3(TLT), and .3(SPY) will be used.



The graph above is a histogram of the 20-day portfolio value after being randomly calculated 5000 times.




```
## $breaks
##  [1] -10000  -8000  -6000  -4000  -2000      0   2000   4000   6000   8000
## [11]  10000  12000
##
## $counts
##  [1]    4   26  163  599 1173 1379  992  488  142   29    5
##
## $density
##  [1] 0.0000004 0.0000026 0.0000163 0.0000599 0.0001173 0.0001379 0.0000992
##  [8] 0.0000488 0.0000142 0.0000029 0.0000005
##
## $mids
##  [1] -9000 -7000 -5000 -3000 -1000  1000  3000  5000  7000  9000 11000
##
## $xname
## [1] "sim1[, n_days] - 1e+05"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

Here is the same graph after subtracting the original investment amount, leaving a histogram of gains and losses.

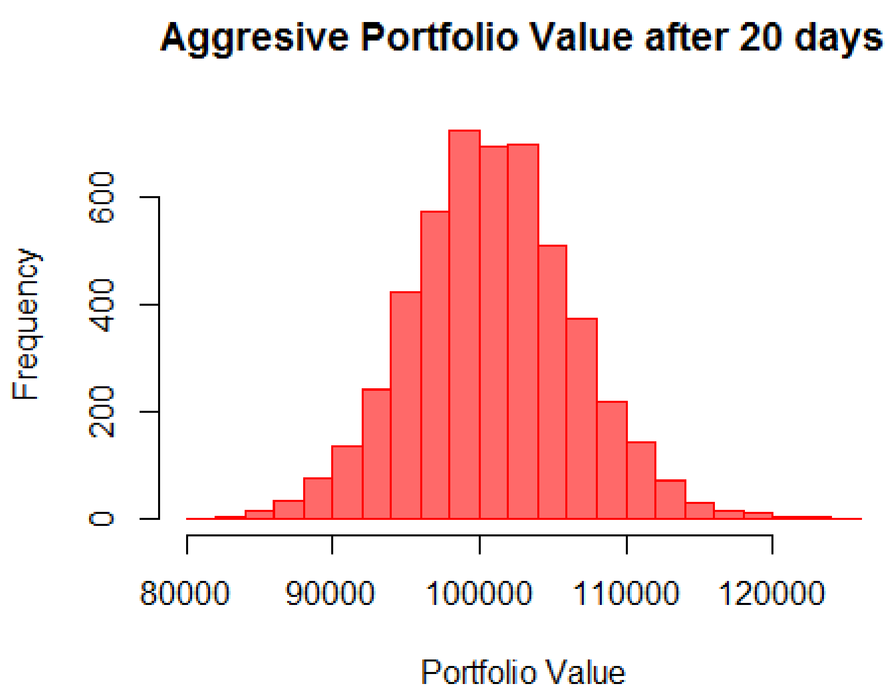
```
##        5%
## -3740.174
```

he value at risk at 5% for this "safe" portfolio is -$1,986, which is expected due to its low risk.

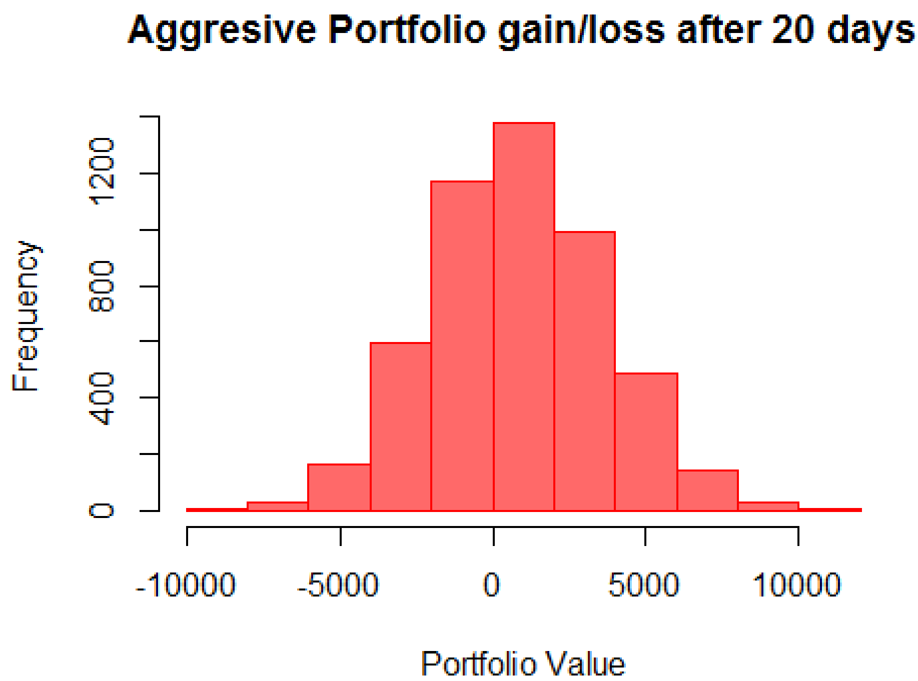
Aggresive Split

This portfolio is a risky porfolio, comprised of assests that reward high risk with high returns.

Based on the stocks' standard deviations (volatility), EEM and VNQ are the "aggresive" stocks. Weights of .5(EEM) and .5(VNQ) will be used.



The graph above is a histogram of the 20-day portfolio value after being randomly calculated 5000 times.

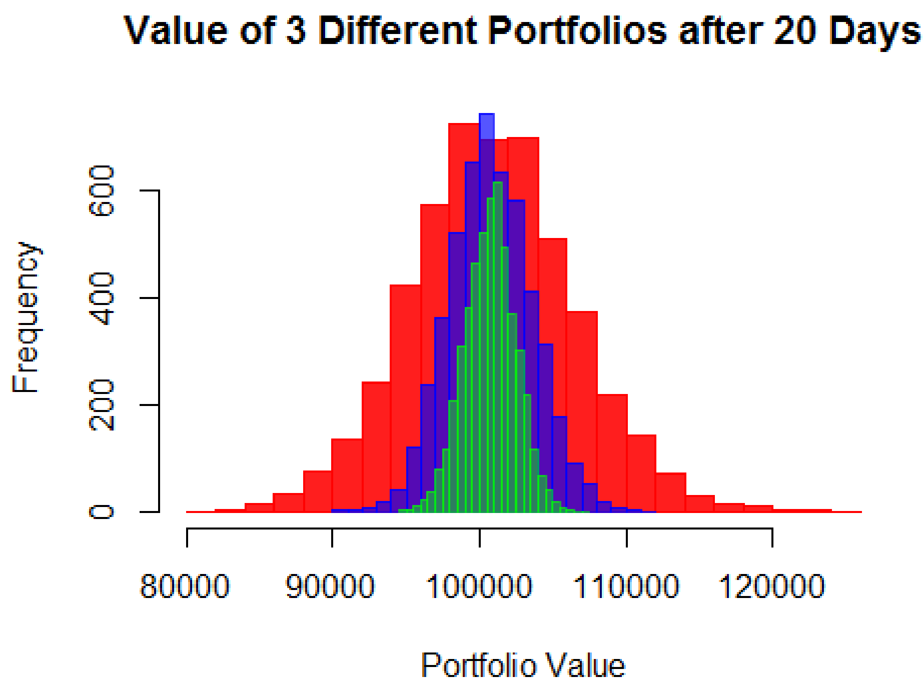



```
## $breaks
##  [1] -10000  -8000  -6000  -4000  -2000      0   2000   4000   6000   8000
## [11]  10000  12000
##
## $counts
##  [1]    4   26  163  599 1173 1379  992  488  142   29    5
##
## $density
##  [1] 0.0000004 0.0000026 0.0000163 0.0000599 0.0001173 0.0001379 0.0000992
##  [8] 0.0000488 0.0000142 0.0000029 0.0000005
##
## $mids
##  [1] -9000 -7000 -5000 -3000 -1000  1000  3000  5000  7000  9000 11000
##
## $xname
## [1] "sim1[, n_days] - 1e+05"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

Here is the same graph after subtracting the original investment amount, leaving a histogram of gains and losses.

```
##        5%
## -3740.174
```

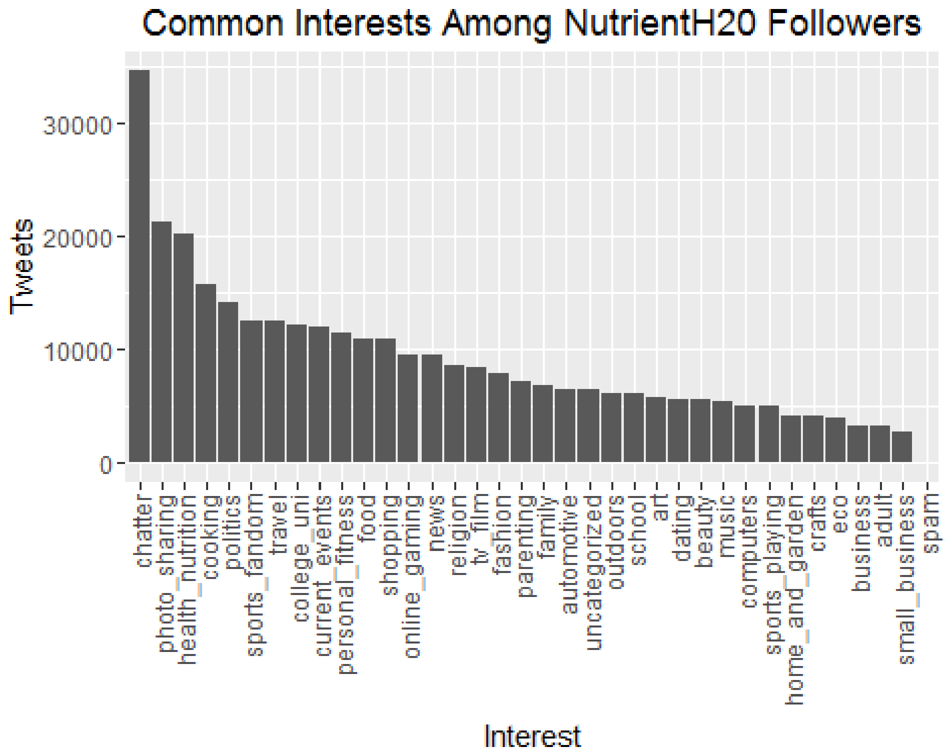
The value at risk at 5% for this "aggressive" portfolio is -$8,057, which is expected due to its high risk.



The graph above combines the 3 histograms into 1. Notice that the low risk portfolio (green) has a less spread and peak. On the other hand, the high risk portfolio (red) has a higher spread and peak. The moderately risky portfolio (blue) falls inbetween the two.

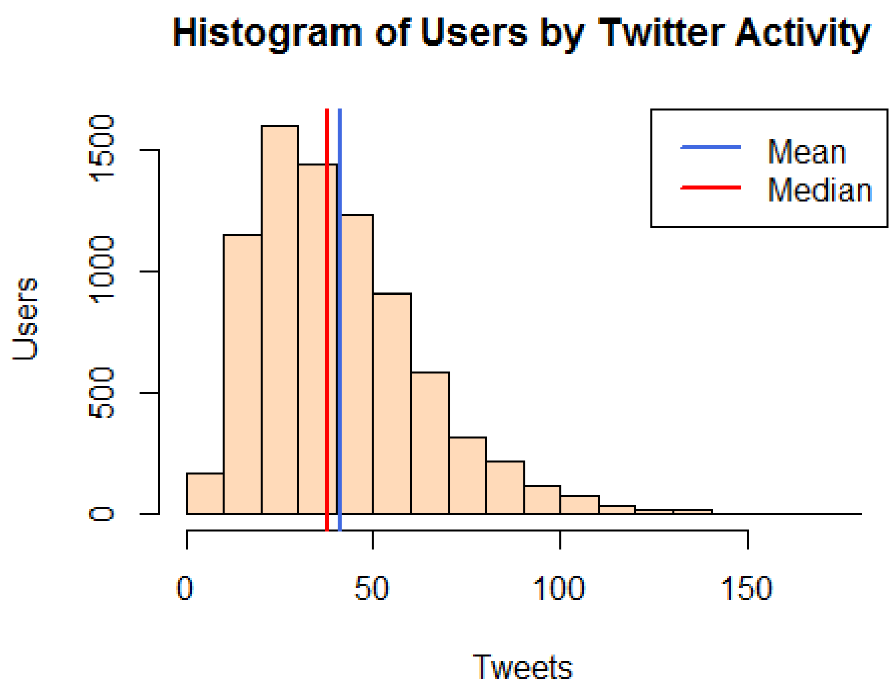
Market Segmentation Report - NutrientH20 Twitter Followers

To better understand NutrientH20's Twitter followers, we first examined what they were talking about most in the figure below.



By far the most frequently discussed topic was the rather amgibuous "chatter". Photo sharing and health & nutrition were a close second, with the latter perhaps revealing a little about the "anonymous" consumer brand's product offering and target market. However, without data on a larger segment of Twitter followers, it is difficult to draw too many conclusions about NutrientH20's followers as a whole.

Next, we looked to understand in the figure below the distribution of NutrientH20's follower activity (as defined by number of tweets).

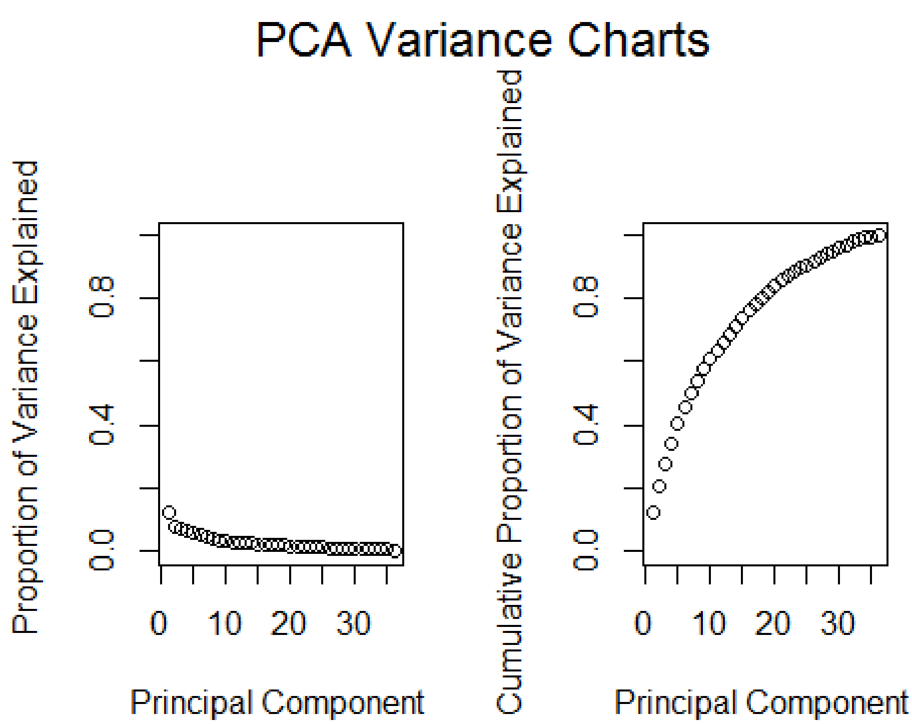


NOTE: Tweet volumes don't account for tweets categorized in multiple interests

Finally, we attempted to segment the followers using Principle Compoments Analysis. See table below for the rotation of the first two principle components and the chart below for the variance explained by each sequential principle component.

```
## Rotation of First Two Principle Components

##                          PC1          PC2
## chatter          -0.12599239  0.197225501
## current_events   -0.09723669  0.064036499
## travel           -0.11664903  0.039947269
## photo_sharing    -0.18027952  0.303077634
## uncategorized    -0.09443507  0.146498856
## tv_film          -0.09745666  0.079352508
## sports_fandom    -0.28773177 -0.316923635
## politics         -0.13026617  0.013939964
## food             -0.29690952 -0.237808675
## family           -0.24426866 -0.196253208
## home_and_garden  -0.11576501  0.046803486
## music            -0.12408921  0.144259544
## news             -0.12764328 -0.036198891
## online_gaming    -0.07388979  0.083591578
## shopping         -0.13299500  0.209852847
## health_nutrition -0.12420109  0.146577761
```

```
## college_uni      -0.09415672  0.115959664
## sports_playing   -0.13021653  0.108595355
## cooking          -0.18880850  0.314287972
## eco              -0.14533561  0.085321972
## computers        -0.14333124  0.037334899
## business         -0.13501004  0.098782574
## outdoors         -0.14260424  0.113581774
## crafts           -0.19362762 -0.021623185
## automotive       -0.13132522 -0.031564108
## art              -0.09794933  0.060347094
## religion         -0.29709999 -0.316152778
## beauty           -0.20151836  0.208609941
## parenting        -0.29400412 -0.295082234
## dating           -0.10515646  0.071535239
## school           -0.28063791 -0.197572367
## personal_fitness -0.13750109  0.144611756
## fashion          -0.18388185  0.279799725
## small_business   -0.11904181  0.094048059
## spam             -0.01146092 -0.004551609
## adult            -0.02673097 -0.006918154
```



Unfortunately the variance explained by each of the Principle Components increases fairly linearly, so it is difficult to reduce the dimensions of the data without sacrificing a substantial amount of information. However, it is possible to make a few educated guesses about the first two principle componets. Due to the fact all coefficients are negative, the first seems to capture the fairly high number of less active users (see histrogram in Fig Y) who do not post much about anything. The second, with high coefficients in chatter(.2), photo sharing (.3), and fashion (.27) and low coefficients in sports fandom (- .3) and parenting (-.3) is most likely capturing young women. While PCA gives a good start towards segmenting the population, further analysis with a combination of more sophisticated tools is necessary due to the high dimensionality and fairly low correlation of the dataset.