Exercises 1

Vincent Chou, Nikita Lakhotia, Fan Liu, Hillary Regan

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?

Answer:

Using the Law of Total Probability ("mixture rule"):

$$P(Y) = P(Y|RC) * P(RC) + P(Y|TC) * P(TC)$$

where:

- P(Y) = .65, the probability a person said Yes
- P(Y|RC) = .5 , the probability a person said Yes given they randomly clicked
- P(RC) = .3 the probability a person randomly clicked
- P(Y|TC) = ?, the probability a person said Yes given they truthfully clicked
- P(TC) = .7, the probability a person truthfully clicked

Therefore, using the formula for the Law of Total Probability:

$$.65 = (.5) * (.3) + P(Y|TC) * (.7)$$

Solving for P(Y|TC) gives:

```
ProbabilityTruthYes = (.65-(.5*.3))/.7
```

$$P(Y|TC) = 0.7142857$$

The fraction of people who are truthful clickers answered yes is 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?

Answer:

Using Bayes' Rule:

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

A: person has the disease

B: test is positive

- P(A|B) = ?, the probability of having the disease given the test was positive
- P(A) = .000025, the probability of having the disease
- P(B|A) = .993, the probability that the test is positive given the person has the disease
- P(B) = ?, the probability the test is positive.

P(B) can be found by using the Law of Total Probability ("mixture rule"):

$$P(B) = P(B|A) * P(A) + P(B|notA) * P(notA)$$

where P(B|notA) = (1-.9999), the probability that the test is positive given the person does not have the disease

$$P(B) = (.993) * (.000025) + (1 - 0.9999) * (1 - .000025)$$

```
ProbabilityTestPositive = (.993)*(.000025) + (1-0.9999)*(1-.000025)
ProbabilityDiseasePositive = (.000025)*(.993)/ProbabilityTestPositive
```

$$P(B) = 1.248225 \times 10^{-4}$$

so,

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)} = \frac{(.000025)(.993)}{1.248225 \times 10^{-4}} = 0.1988824$$

The probability that a person has the disease given the test is positive is 0.1988824. Since the chance of this is small, I do envision problems in implementing a universal testing policy for the disease. This is not a good enough test for the disease.

Exploratory Analysis: Green Buildings

Overview:

Buildings can be certified-green by two organizations, Energystar and LEED. We looked at the effects that this certification can have on rent prices. This is imperative because there are higher upfront costs to creating a green-certified building. If consumers are not willing to pay a higher rent for these buildings, there is no economic reason to build a green-certified building. A previous analysis of this data found that there is a premium on green-certified buildings.

Data and Model:

Our dataset contains green-certified residential buildings and the buildings around them (to control for the economic value of the locations of these buildings) as well as numerous other variables that may affect rent (local weather, building size, etc). We used a linear regression model to try to isolate the effects of having a green certified building because the coefficients in a linear regression model show the effects of variables holding all others constant.

Results:

```
##  CS_PropertyID        cluster            size            empl_gr
##  Min.   :      1   Min.   :   1.0   Min.   :   1624   Min.   :-24.950
##  1st Qu.: 157452   1st Qu.: 272.0   1st Qu.:  50891   1st Qu.:  1.740
##  Median : 313253   Median : 476.0   Median : 128838   Median :  1.970
##  Mean   : 453003   Mean   : 588.6   Mean   : 234638   Mean   :  3.207
##  3rd Qu.: 441188   3rd Qu.:1044.0   3rd Qu.: 294212   3rd Qu.:  2.380
##  Max.   :6208103   Max.   :1230.0   Max.   :3781045   Max.   : 67.780
##                                                       NA's   :74
##       Rent         leasing_rate       stories            age
##  Min.   :  2.98   Min.   :  0.00   Min.   :  1.00   Min.   :  0.00
##  1st Qu.: 19.50   1st Qu.: 77.85   1st Qu.:  4.00   1st Qu.: 23.00
##  Median : 25.16   Median : 89.53   Median : 10.00   Median : 34.00
##  Mean   : 28.42   Mean   : 82.61   Mean   : 13.58   Mean   : 47.24
##  3rd Qu.: 34.18   3rd Qu.: 96.44   3rd Qu.: 19.00   3rd Qu.: 79.00
##  Max.   :250.00   Max.   :100.00   Max.   :110.00   Max.   :187.00
##
##    renovated         class_a          class_b           LEED
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000000
##  Mean   :0.3795   Mean   :0.3999   Mean   :0.4595   Mean   :0.006841
```

```
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000
##
##    Energystar       green_rating          net            amenities
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :0.08082   Mean   :0.08677   Mean   :0.03471   Mean   :0.5266
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##   cd_total_07     hd_total07    total_dd_07   Precipitation
##  Min.   :  39   Min.   :   0   Min.   :2103   Min.   :10.46
##  1st Qu.: 684   1st Qu.:1419   1st Qu.:2869   1st Qu.:22.71
##  Median : 966   Median :2739   Median :4979   Median :23.16
##  Mean   :1229   Mean   :3432   Mean   :4661   Mean   :31.08
##  3rd Qu.:1620   3rd Qu.:4796   3rd Qu.:6413   3rd Qu.:43.89
##  Max.   :5240   Max.   :7200   Max.   :8244   Max.   :58.02
##
##    Gas_Costs        Electricity_Costs  cluster_rent
##  Min.   :0.009487   Min.   :0.01780   Min.   : 9.00
##  1st Qu.:0.010296   1st Qu.:0.02330   1st Qu.:20.00
##  Median :0.010296   Median :0.03274   Median :25.14
##  Mean   :0.011336   Mean   :0.03096   Mean   :27.50
##  3rd Qu.:0.011816   3rd Qu.:0.03781   3rd Qu.:34.00
##  Max.   :0.028914   Max.   :0.06280   Max.   :71.44
##
```

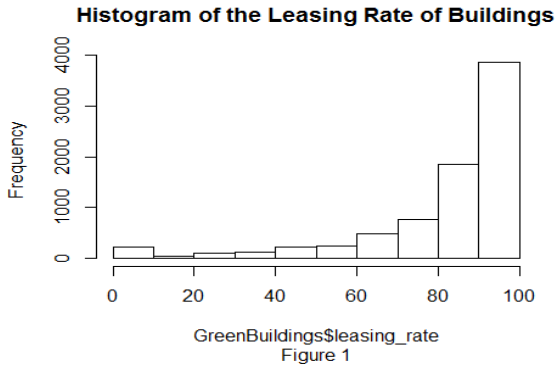


Figure 1

The "stats guru" was correct when he said "a handful of the buildings in the data set had very low occupancy rates" which is shown by the histogram above. Yet, I do not agree with removing these buildings from consideration. There is not a guarantee that the new building will have a high occupancy rate.

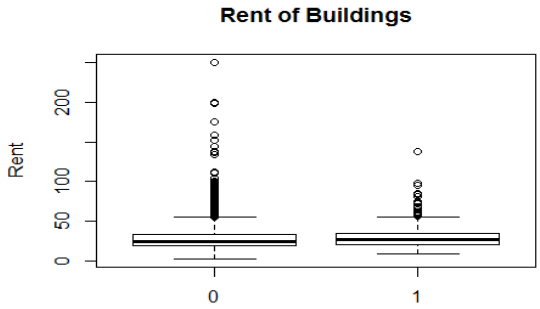


Figure 2

On the x-axis, 0 represents the buildings that are not green. 1 represents buildings that have a green rating. The median rent is very similar for the two types of buildings, as is the first and third quartiles, but the rent for a green building is slightly higher. There are more outliers for non-green buildings.

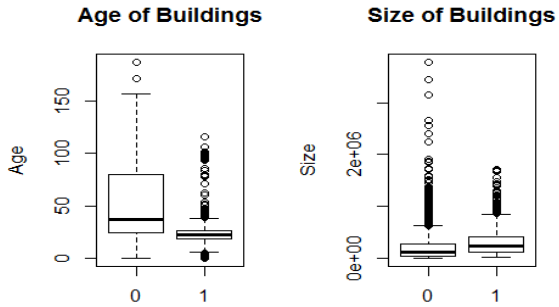


Figure 3 Figure 4

On average, green buildings tend to be newer buildings and bigger buildings compared to non-green buildings.

```
##
## Call:
## lm(formula = Rent ~ ., data = GreenBuildings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.753  -3.581  -0.526   2.491 173.916
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -8.315e+00  1.018e+00  -8.167 3.67e-16 ***
## CS_PropertyID      2.959e-07  1.574e-07   1.879 0.060241 .
## cluster            7.532e-04  2.840e-04   2.653 0.008006 **
## size               6.741e-06  6.561e-07  10.276  < 2e-16 ***
## empl_gr            6.450e-02  1.700e-02   3.794 0.000149 ***
## leasing_rate       9.454e-03  5.332e-03   1.773 0.076247 .
## stories           -3.472e-02  1.617e-02  -2.147 0.031823 *
## age               -1.249e-02  4.717e-03  -2.649 0.008096 **
```

```
## renovated          -1.425e-01  2.586e-01  -0.551 0.581681
## class_a             2.872e+00  4.377e-01   6.563 5.63e-11 ***
## class_b             1.186e+00  3.427e-01   3.462 0.000539 ***
## LEED                1.877e+00  3.582e+00   0.524 0.600318
## Energystar         -2.127e-01  3.818e+00  -0.056 0.955572
## green_rating        6.969e-01  3.839e+00   0.182 0.855929
## net                -2.559e+00  5.929e-01  -4.316 1.61e-05 ***
## amenities           6.703e-01  2.519e-01   2.661 0.007802 **
## cd_total_07        -1.248e-04  1.464e-04  -0.852 0.394005
## hd_total07          5.354e-04  8.972e-05   5.967 2.52e-09 ***
## total_dd_07                NA         NA      NA       NA
## Precipitation       4.830e-02  1.611e-02   2.997 0.002735 **
## Gas_Costs          -3.559e+02  7.842e+01  -4.538 5.76e-06 ***
## Electricity_Costs   1.886e+02  2.493e+01   7.563 4.38e-14 ***
## cluster_rent        1.008e+00  1.421e-02  70.949  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.413 on 7798 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.6126, Adjusted R-squared:  0.6116
## F-statistic: 587.2 on 21 and 7798 DF,  p-value: < 2.2e-16
```

Conclusions:

We could not find evidence that having a green-certified building has a significant increase in rent prices. The "Excel Guru" that performed the previous study did not account for confounders. As shown in **Figures 3 & 4**, green buildings tend to be newer buildings and bigger buildings which our model showed to be actual significant variables. This would account for the difference in medians that the "Excel Guru" identified. Our final decision would have to be to save on costs and not become green-certified.

Bootstrapping

Overview:

There is a notional $100,000 to invest in assets: US domestic equities (SPY: the S&P 500 stock index), US Treasury bonds (TLT), Investment-grade corporate bonds (LQD), Emerging-market equities (EEM), and Real estate (VNQ). We consider three portfolios, the even split, safe and aggressive portfolios.

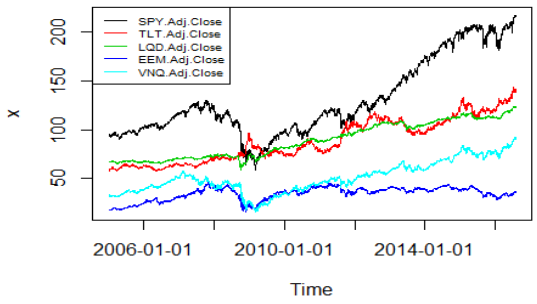
Data and Model:

Data from daily reports of exchange-traded funds were chosen over a ten year period, including the 2008, the year the housing market crashed, to ensure that the data included both good runs and bad runs of stock-market performance. Three portfolios were considered. The first being the even split which entails 20% of the assets in each of the five ETFs above. Second, a portfolio safer than the even split, comprising investments in 'LQD','EEM','VNQ'. The allocation was 75% in 'LQD', 12.5% in 'EEM', and 12.5% in 'VNQ.'

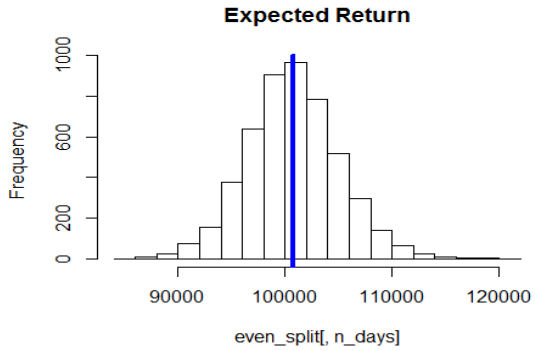
These 3 were chosen based on figure X, which shows the historical adjusted closing price for each of the five ETFs, noticing the previous trends and the stability of prices for these investments over the years. 75% was invested in 'LQD' due to the fact that 'LQD' was less affected during 2008 financial crisis, so it is a safer choice than 'EEM' and 'VNQ'. The third portfolio considered was more aggressive. It comprised of investments of 90% in 'SPY' and 10% in 'TLT'. This portfolio included exchange-traded funds that previously had higher returns, but showed a higher risk of loss in Figure X. Note that even though 'TLT' was not affected by the housing market crash at all, there were several sharp drops over the years. Bootstrap resampling was used to estimate the 4-week (20 trading day) expected return (mean) and value at risk of each of the three portfolios at the 5% level.

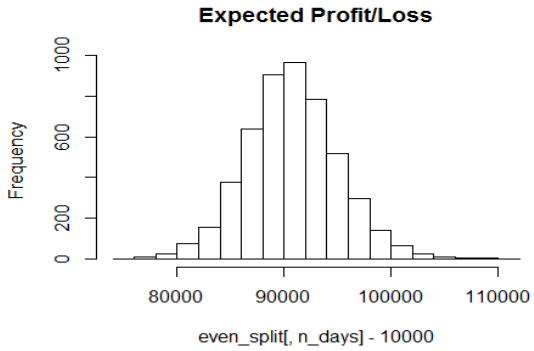
Results:

```
## GMT
##            SPY.Open SPY.High SPY.Low SPY.Close SPY.Volume SPY.Adj.Close
## 2005-01-03   121.56   121.76  119.90    120.30   55748000      95.29625
## 2005-01-04   120.46   120.54  118.44    118.83   69167600      94.13178
## 2005-01-05   118.74   119.25  118.00    118.01   65667300      93.48222
## 2005-01-06   118.44   119.15  118.26    118.61   47814700      93.95751
## 2005-01-07   118.97   119.23  118.13    118.44   55847700      93.82284
## 2005-01-10   118.34   119.46  118.34    119.00   56563300      94.26645
##            TLT.Open TLT.High TLT.Low TLT.Close TLT.Volume TLT.Adj.Close
## 2005-01-03    88.18    88.84   88.16     88.74    1168000      58.23352
## 2005-01-04    88.72    88.75   87.81     87.81    1935400      57.62323
## 2005-01-05    87.99    88.55   87.94     88.28    1094100      57.93165
## 2005-01-06    88.29    88.54   88.22     88.34    1057400      57.97103
## 2005-01-07    88.76    88.87   88.35     88.54     738700      58.10227
## 2005-01-10    88.64    88.72   88.47     88.68     379400      58.19415
##            LQD.Open LQD.High LQD.Low LQD.Close LQD.Volume LQD.Adj.Close
## 2005-01-03   111.71   112.25  111.50    112.25    1497800      66.89484
## 2005-01-04   112.27   112.29  111.50    111.62      90200      66.51940
## 2005-01-05   111.65   111.91  111.46    111.71     120700      66.57303
## 2005-01-06   111.55   111.95  111.55    111.79      43900      66.62071
## 2005-01-07   111.91   111.92  111.43    111.74      68300      66.59091
## 2005-01-10   111.76   111.76  111.40    111.55      73200      66.47768
##            EEM.Open EEM.High EEM.Low EEM.Close EEM.Volume EEM.Adj.Close
## 2005-01-03   201.70   202.45  199.38    199.75    4275000      18.03730
## 2005-01-04   199.25   199.35  193.60    193.60    4205700      17.48196
## 2005-01-05   193.40   193.77  191.20    191.23    3006900      17.26795
## 2005-01-06   191.85   192.12  190.13    191.10    2268000      17.25621
## 2005-01-07   192.40   192.76  190.50    191.47    4920300      17.28962
## 2005-01-10   192.60   193.61  191.65    191.71    2007000      17.31129
##            VNQ.Open VNQ.High VNQ.Low VNQ.Close VNQ.Volume VNQ.Adj.Close
## 2005-01-03    56.75    56.75   55.50     55.89      31900      33.22696
## 2005-01-04    55.89    56.28   55.05     55.05      52500      32.72758
## 2005-01-05    55.17    55.17   52.83     53.22      77300      31.63963
## 2005-01-06    53.15    53.82   53.15     53.63      42200      31.88338
## 2005-01-07    53.87    53.88   53.27     53.51      24200      31.81204
## 2005-01-10    53.45    53.65   53.20     53.34      12500      31.71097
```




```
##               [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## result.1  99078.66  99647.28  99576.80  99623.08  99650.12  98454.72
## result.2 100367.76 100146.30 100760.49  98703.85  99213.36  99226.98
## result.3  99723.67  98061.44  98425.48  97509.35  98749.28  98895.30
## result.4 100381.95 100764.70 100651.34 102067.06 102379.58 102299.27
## result.5 100795.12 100375.56  98449.34  98746.13  98427.49  97963.57
## result.6  98797.46  98652.76  99213.33 100319.41  99953.37  99562.57
##               [,7]      [,8]      [,9]     [,10]     [,11]     [,12]
## result.1  97887.49  96886.71  97284.32  97531.59  97637.00  97736.82
## result.2  99744.38  99413.06  99202.97  96406.46  95751.34  95890.80
## result.3  98512.60  97849.83  99056.07  99242.39  95952.12  95908.09
## result.4 102194.23 102587.46 103057.48 103380.40 103772.60 102845.14
## result.5  98354.23  97792.51  97450.83  97854.42  97650.41  97564.83
## result.6  99545.87  99686.27  99715.09  99128.53  99343.95  99994.04
##              [,13]     [,14]     [,15]     [,16]     [,17]     [,18]
## result.1  97726.42  97804.32  96456.42  96614.48  96969.86  96874.55
## result.2  95626.91  95525.76  96340.24  96731.23  96926.39  97006.37
## result.3  94268.53  95106.09  94855.65  93634.61  95037.27  96346.56
## result.4 103176.67 102947.74 103510.94 103916.90 103680.85 103707.57
## result.5  97893.98  97430.87  98034.12  97547.42  97649.35  97557.63
## result.6 100313.96 100290.41 100791.65 100671.28 100148.19  99493.60
##              [,19]     [,20]
## result.1  96329.20  96031.57
## result.2  96950.31  97591.99
## result.3  96656.86  96452.88
## result.4 103313.28 103781.08
## result.5  97033.00  96539.52
## result.6  98987.55  99137.62
```



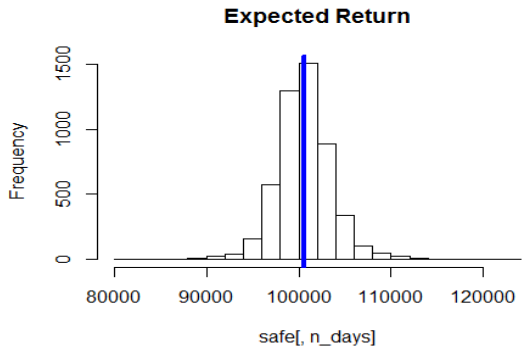


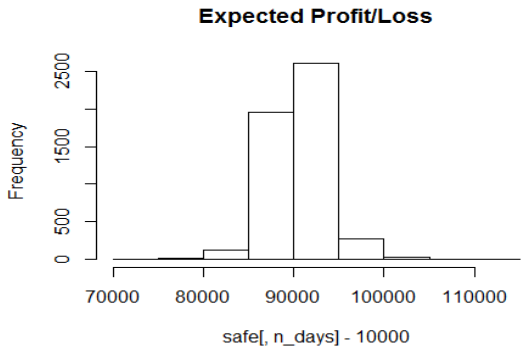

```
##            SPY.PctReturn TLT.PctReturn LQD.PctReturn EEM.PctReturn
## 2005-01-04  -0.012219463 -0.0104800469 -0.0056124508 -0.0307883143
## 2005-01-05  -0.006900613  0.0053524770  0.0008062761 -0.0122416480
## 2005-01-06   0.005084304  0.0006796284  0.0007161459 -0.0006800459
## 2005-01-07  -0.001433254  0.0022640275 -0.0004472934  0.0019361727
## 2005-01-10   0.004728113  0.0015811945 -0.0017003223  0.0012535267
## 2005-01-11  -0.006890755  0.0058637513  0.0023307372 -0.0018777916
##            VNQ.PctReturn
## 2005-01-04  -0.015029512
## 2005-01-05  -0.033242486
## 2005-01-06   0.007703883
## 2005-01-07  -0.002237592
## 2005-01-10  -0.003176942
## 2005-01-11  -0.010123752

##            LQD.PctReturn EEM.PctReturn VNQ.PctReturn
## 2005-01-04 -0.0056124508 -0.0307883143  -0.015029512
## 2005-01-05  0.0008062761 -0.0122416480  -0.033242486
## 2005-01-06  0.0007161459 -0.0006800459   0.007703883
## 2005-01-07 -0.0004472934  0.0019361727  -0.002237592
## 2005-01-10 -0.0017003223  0.0012535267  -0.003176942
## 2005-01-11  0.0023307372 -0.0018777916  -0.010123752

##               [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## result.1  99907.12  99881.66  99733.72  99756.86  99667.68  99671.02
## result.2 100196.16 100377.51 100426.18 100503.37 100144.52 100433.26
## result.3  99632.04  99594.12  99564.27  99793.22  99930.62  99656.00
## result.4 100318.22 100324.66  99808.64 100401.04  99674.32  99939.58
## result.5 100446.78 100947.16 101088.77 100533.52 100661.53 100556.47
## result.6 100201.75 100384.54 100238.28 100866.65 100907.73 100828.72
##               [,7]     [,8]     [,9]    [,10]    [,11]     [,12]     [,13]
## result.1  99979.97 100981.6 100885.4 100117.0 100080.8 100220.53 100538.00
## result.2 100665.46 100739.5 101178.3 101753.9 101390.0 101062.64 101249.60
## result.3  99531.55 100048.7 100026.6 100485.3 100602.1 100594.62 101198.09
## result.4 100187.97 100143.2 100078.7 100064.7 100010.6 100361.54 100223.52
## result.5 100462.17 100630.9 100687.2 100767.6 100742.7  99932.48  99022.91
## result.6 101184.53 100988.3 101176.3 100566.7 100144.3 100324.22 100108.85
##               [,14]    [,15]     [,16]     [,17]     [,18]     [,19]
```

```
## result.1 100172.69 100470.6 100412.68 100518.25 100700.50 100975.75
## result.2 100669.15 100987.9 101031.33 101161.63 100919.89 100840.55
## result.3 101494.75 101575.5 101709.16 101749.26 101827.75 102011.62
## result.4  99896.73 100127.0 100272.27 100219.90 100270.45 101611.10
## result.5  97314.35  97982.8  97766.61  97716.01  97614.47  98580.84
## result.6 100675.24 100706.9 100908.81 100491.74 100548.32 100674.96
##              [,20]
## result.1 101152.52
## result.2 100859.57
## result.3 102787.08
## result.4 101477.10
## result.5  99110.92
## result.6 101604.77
```

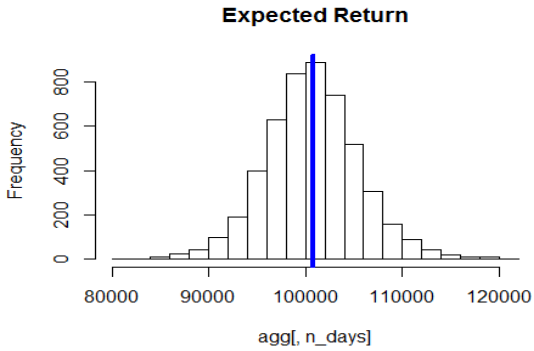


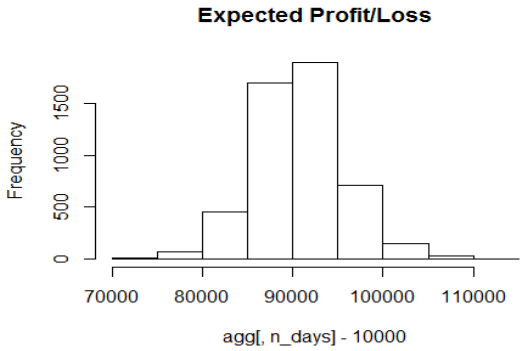



```
##            SPY.PctReturn TLT.PctReturn
## 2005-01-04  -0.012219463 -0.0104800469
## 2005-01-05  -0.006900613  0.0053524770
## 2005-01-06   0.005084304  0.0006796284
## 2005-01-07  -0.001433254  0.0022640275
## 2005-01-10   0.004728113  0.0015811945
## 2005-01-11  -0.006890755  0.0058637513

##              [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## result.1  99713.96 100344.03 100488.86 100644.24 101042.06 101153.76
## result.2  99037.31  99330.04  99647.73 100430.29 101285.62 101176.69
## result.3 100357.81 100912.37 100455.38  98993.38 101763.23 101062.03
## result.4 100184.75  99878.99 102685.44 102744.17 102513.40 102264.59
## result.5 100791.45 100909.57 100472.67 101152.29 101818.51 102976.32
## result.6 100317.19 100053.57  99028.16 100170.64  99082.59  98264.33
##              [,7]      [,8]      [,9]    [,10]     [,11]     [,12]
## result.1 101414.53 101994.68 102611.39 102408.2 102659.76 104371.03
## result.2 100785.86 100739.96 100986.20 100777.1 101088.22 101122.76
## result.3 101676.91 101541.04 101567.30 102931.8 102717.85 101534.23
## result.4 102437.18 100815.19  99821.85 101213.2 101115.71 104626.69
## result.5 104667.75 104664.54 105754.06 106219.9 106828.47 105522.56
## result.6  98746.77  98259.96  98034.02  96046.5  95847.18  94760.99
##             [,13]     [,14]     [,15]     [,16]     [,17]     [,18]
## result.1 104493.95 104530.97 102980.16 103499.12 103429.80 103476.31
```

```
## result.2 100541.84 101488.57 102119.23 103050.90 103644.52 104608.75
## result.3 101188.89 100998.17  99995.93 100084.99 100591.12  98329.19
## result.4 104959.44 106111.78 106599.15 106877.37 106597.50 105229.41
## result.5 104678.06 104662.47 104606.29 103663.83 104152.14 105506.82
## result.6  95231.89  94887.85  95265.47  95212.93  95960.31  95820.51
##              [,19]     [,20]
## result.1 104015.45 103343.92
## result.2 106340.00 105761.41
## result.3  99648.28  97978.83
## result.4 106763.87 105962.56
## result.5 105272.95 105272.00
## result.6  95117.76  95234.95
```





The even split portfolio's expected return (mean) on the 20th trading day is $100769 which produces a profit of $769 and the 5% Value at Risk is $83874, that is, a loss of $16125. The safe portfolio's expected return (mean) is $100609 which produces a profit of $609 and the 5% Value at Risk is $86130, that is, a loss of $13869. The aggressive portfolio's expected return (mean) is $100741 which produces a profit of $741 and the 5% Value at Risk is $83084, that is, a loss of $16915.

If the "safer" portfolio is chosen, less profit would be made, but less risk undertook. If the more aggressive portfolio is chosen, surprisingly, a lower return than that from the even split (but higher than the "safer" choice) is expected; yet, there would be a higher risk of loss. Investors can make intelligent decisions using the results above to choose which of the three options to invest in.

Conclusions:

Based on the analysis, a person who is reluctant to take on a risk should choose the safer choice. In contrast, a person who are willing to take higher risks to achieve above-average returns should go with more aggressive portfolio in theory. The person making this type of decision should weigh all the factors involved in the risk and assess these risks against the probabilities of different outcomes. Unfortunately, the results produced show the aggressive option gives a lower expected return than the even split; therefore, it is not worth the risk. A risk-neutral individual will choose the assets with the highest possible gains or returns without taking into account possible outcomes, so the even split would be the optimal choice in this case.

Marketing

Overview:

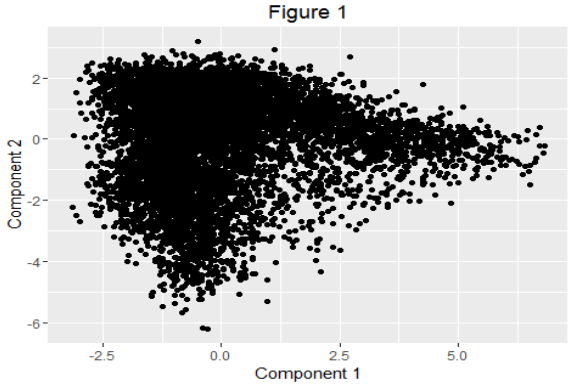
We looked to categorize followers of NutrientH2O on Twitter to help NutrientH2O understand their customers, particularly their primary interests (as shown by the topics they tend tweet about). This would help NutrientH2O with targeted marketing and their ad campaigns focus on their primary audience and make them more aware of the types of customer they're lacking in and can expand to.

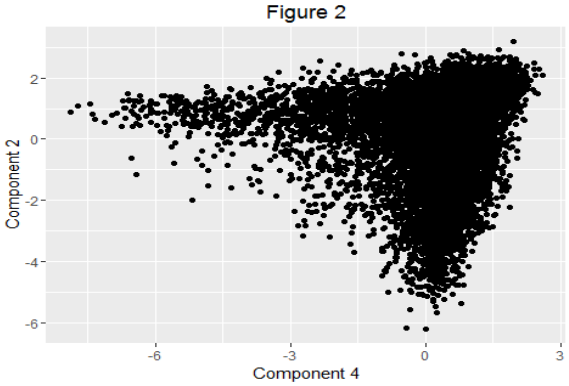
Data and Models:

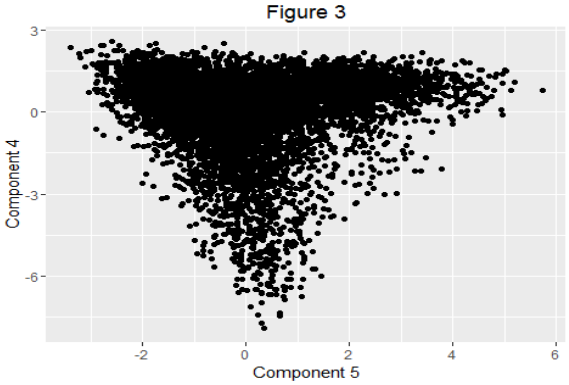
This dataset contained the categorized tweets of users who follow NutrientH2O on Twitter. We expect a little noise in this data since they were categorized by Amazon Mechanical Turk users who are not infallible (and often are bots themselves). Because we have a wide variety of categories, and we really do not understand what types of groups we're going to find, we used PCA, an unsupervised learning method.

```
## Loading required package: RColorBrewer

## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     1.69908 1.61882 1.54302 1.46058 1.40975 1.27620
## Proportion of Variance 0.08019 0.07279 0.06614 0.05926 0.05521 0.04524
## Cumulative Proportion  0.08019 0.15299 0.21912 0.27838 0.33358 0.37883
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     1.20384 1.14303 1.09172 1.05517 1.02939 0.99739
## Proportion of Variance 0.04026 0.03629 0.03311 0.03093 0.02943 0.02763
## Cumulative Proportion  0.41908 0.45537 0.48848 0.51941 0.54884 0.57648
##                           PC13    PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.98693 0.98152 0.97315 0.95282 0.94539 0.92387
## Proportion of Variance 0.02706 0.02676 0.02631 0.02522 0.02483 0.02371
## Cumulative Proportion  0.60353 0.63029 0.65660 0.68182 0.70664 0.73035
##                           PC19    PC20    PC21    PC22    PC23    PC24
## Standard deviation     0.8960 0.87279 0.85324 0.82936 0.82067 0.80427
## Proportion of Variance 0.0223 0.02116 0.02022 0.01911 0.01871 0.01797
## Cumulative Proportion  0.7527 0.77381 0.79404 0.81314 0.83185 0.84982
##                           PC25    PC26    PC27    PC28    PC29    PC30
## Standard deviation     0.78788 0.77386 0.76896 0.76030 0.74885 0.73396
## Proportion of Variance 0.01724 0.01663 0.01643 0.01606 0.01558 0.01496
## Cumulative Proportion  0.86706 0.88370 0.90012 0.91618 0.93176 0.94672
##                           PC31    PC32    PC33   PC34    PC35      PC36
## Standard deviation     0.70016 0.64719 0.61999 0.5659 0.55169 3.812e-15
## Proportion of Variance 0.01362 0.01163 0.01068 0.0089 0.00845 0.000e+00
## Cumulative Proportion  0.96034 0.97197 0.98265 0.9916 1.00000 1.000e+00
```








```
## [1] "school"        "food"          "parenting"     "sports_fandom"
## [5] "religion"

## [1] "health_nutrition" "personal_fitness" "outdoors"
## [4] "cooking"          "fashion"
```

Results:

Using PCA, 33 principal components were identified. Each component accounted for less than 9% of the variation, so this is clearly a very varied group of individuals. **Figure 1, 2, and 3** show several plots of the components. The highly positive and negative variable weights in each component are shown, which is used as the way of identifying the topics the customers were Tweeting about. For example, in **Figure 1**, there is a concentration of customers at positive PC2 and negative PC1. These points were associated with these variables: school, food, parenting, sports fandom, and religion as well as health nutrition, personal fitness, outdoors, cooking, and fashion.

Conclusion:

The followers of NutrientH2O are extremely varied, covering most of the ranges of all of our principal component models. While NutrientH2O has a very wide base, there seemed to be strong clusters of followers who tweeted about food, sports, fitness, religion, and the outdoors (using principal components 1 and 2). This demographic looks strongly like the health conscious Millennials. This is supported when looking at other principal components (2 and 4). This group is similar to the first but has additional interests in video games, playing sports, and university. Not only do these seem to be college students/Millennials, but the presence of religion and sports may point towards the Southeast and Southwest US which are generally more religious.

Looking into the areas without followers that NutrientH2O can expand into, the empty areas consistently seemed to represent potential customers interested in automotives, politics, and the news. These topics tend to interest and older demographic which is intuitive since the primary demographic of NutrientH2O's followers appeared to be younger.