STA 380, Part 2: Exercises 1

Brooks Beckelman, Zack Bilderback, Davis Townsend

Probability Practice

Part A.

This problem can be solved using the Law of Total Probability, which states that the probability that an event will happen is the sum of the probabilities of all of the different ways that the event can happen. In this context, this means that the probability of a "yes" answer to the survey is equal to the probability of a "yes" answer from a random clicker plus the probability of a "yes" answer from a random clicker. This can be modeled by the following equation:

$$P(Yes) = P(Yes|RandomClicker) * P(RandomClicker) + P(Yes|TruthfulClicker) * P(TruthfulClicker)$$

This equation can be rearranged to solve for the probability of a "yes" answer given that the user is a truthful clicker:

$$P(Yes|TruthfulClicker) = \frac{P(Yes) - P(Yes|RandomClicker) * P(RandomClicker)}{P(True)}$$

This equation is solved using the code below.

```
prob_yes = 0.65
prob_randClicker = 0.30
prob_trueClicker = 1 - prob_randClicker
prob_yes_givenRand = 0.50

prob_yes_givenTrue = (prob_yes - (prob_yes_givenRand*prob_randClicker)) / prob_trueClicker
prob_yes_givenTrue
```

```
## [1] 0.7142857
```

The estimated fraction of truthful clickers who answered "yes" is about 0.714, or 71.4%.

Part B.

This problem can be solved using both the Law of Total Probability and Bayes' Rule. The Law of Total Probability is described in Part A. Bayes' Rule is a method that can be used to flip the conditional probability of two events. In this case, we must obtain the probability of having the disease given a positive test from the probability of receiving a positive test given that you have the disease. The Bayes' Rule is modeled in the following equation:

$$P(Disease|Positive) = \frac{P(Disease) * P(Positive|Disease)}{P(Positive)}$$

Everything in the above equation is given except for the overall probability of a positive test: P(Positive). This can be obtained using the Law of Total Probability:

$$P(Positive) = P(Positive|Disease) * P(Disease) + P(Positive|NoDisease) * P(NoDisease)$$

The code below solves for the probability that someone has the disease given that they tested positive.

```
prob_disease = 0.000025
prob_noDisease = 1 - prob_disease
prob_pos_givenDisease = 0.993
prob_pos_givenNoDisease = 1 - 0.9999

prob_pos = prob_pos_givenDisease*prob_disease + prob_pos_givenNoDisease*prob_noDisease

prob_disease_givenPos = (prob_disease*prob_pos_givenDisease) / prob_pos
prob_disease_givenPos
```

```
## [1] 0.1988824
```

Given that a person tests positve, there is only about a 20% chance that they actually have the disease. Therefore, this test should probably not be implemented in a universal testing policy for the disease since you are far more likely to receive a false positive than a true one.

Exploratory Analysis: Green Buildings

Let's begin by checking the accuracy of the findings of the on-staff stats guru.

```
# Filter out buildings that have less than 10% of available space occupied
buildings_lr10 = buildings[buildings$leasing_rate>10,]
attach(buildings)

# Separate data set into green buildings and non-green buildings
green_buildings_lr10 = buildings_lr10[buildings_lr10$green_rating==1,]
nongreen_buildings_lr10 = buildings_lr10[buildings_lr10$green_rating==0,]

# Find medians for each
median_green = median(green_buildings_lr10$Rent)
median_nongreen = median(nongreen_buildings_lr10$Rent)
median_green
```

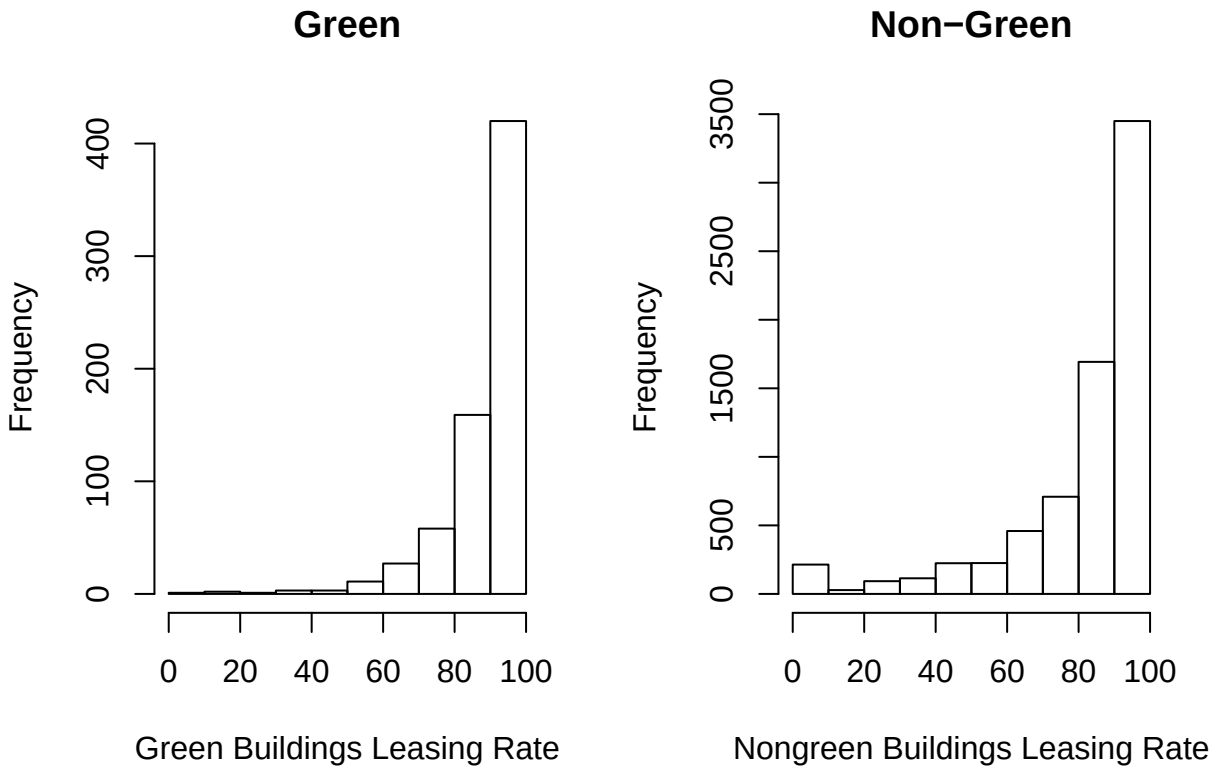
```
## [1] 27.6
```

```
median_nongreen
```

```
## [1] 25.03
```

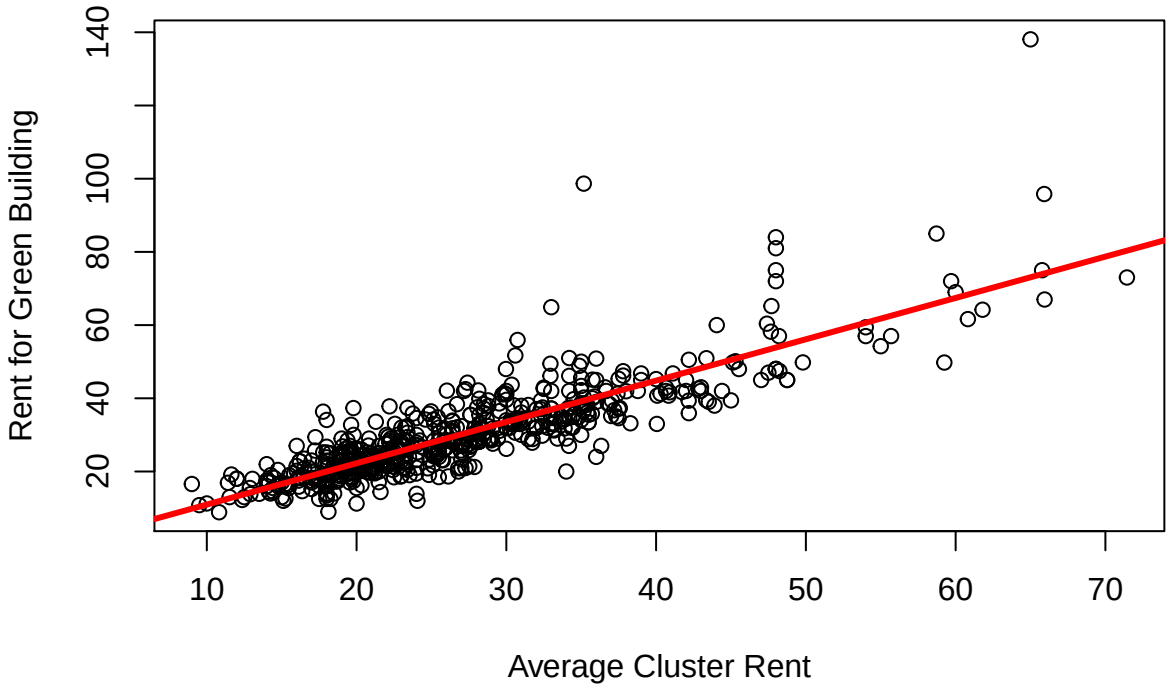
It appears that the median market rents that were found by the guru are correct if you remove buildings that have less than 10% of the available space occupied. We concluded that it is reasonable to remove the buildings that have less than 10% leasing rate because those are likely not extremely relevant to the problem at hand, and, as the guru indicated, could potentially distort the analysis.

In fact, we think that it may be beneficial to eliminate a greater number of properties from our analysis based on leasing rate. Let's look at the leasing rate distributions for green and non-green buildings.

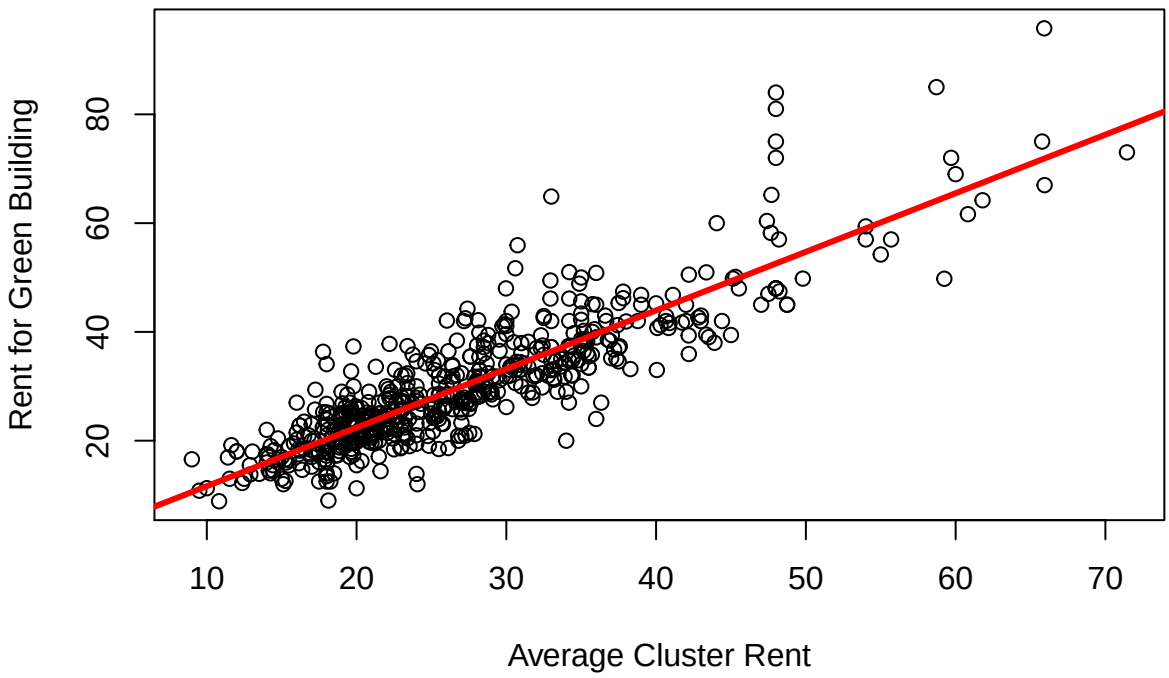


It appears that the leasing rates for both green and non-green buildings are largely skewed toward higher leasing rates. The majority of buildings in both cases have leasing rates greater than 80%. Therefore, going forward we will only examine those buildings to eliminate any properties that may have unique circumstances that are leading to low leasing rates.

Even though the guru reported accurate results, this does not mean his/her analysis is not flawed. For one, he/she takes the median rent of all green buildings and the median rent of all non-green buildings without considering any other factors that may contribute to the rent of a building. One way to mitigate this is to compare rents for green buildings and non-green buildings within each cluster. Each cluster contains one green building and all non-green buildings within a quarter-mile radius of the green building. Therefore, comparing rents within a cluster should account for factors associated with location that may influence rent. If we perform a linear regression comparing the rent of a green building to the average rent in that building's cluster, we should get a better idea of the premium for rent in green buildings.



It appears that there are a couple of outliers that we should get rid of to obtain a more accurate relationship.

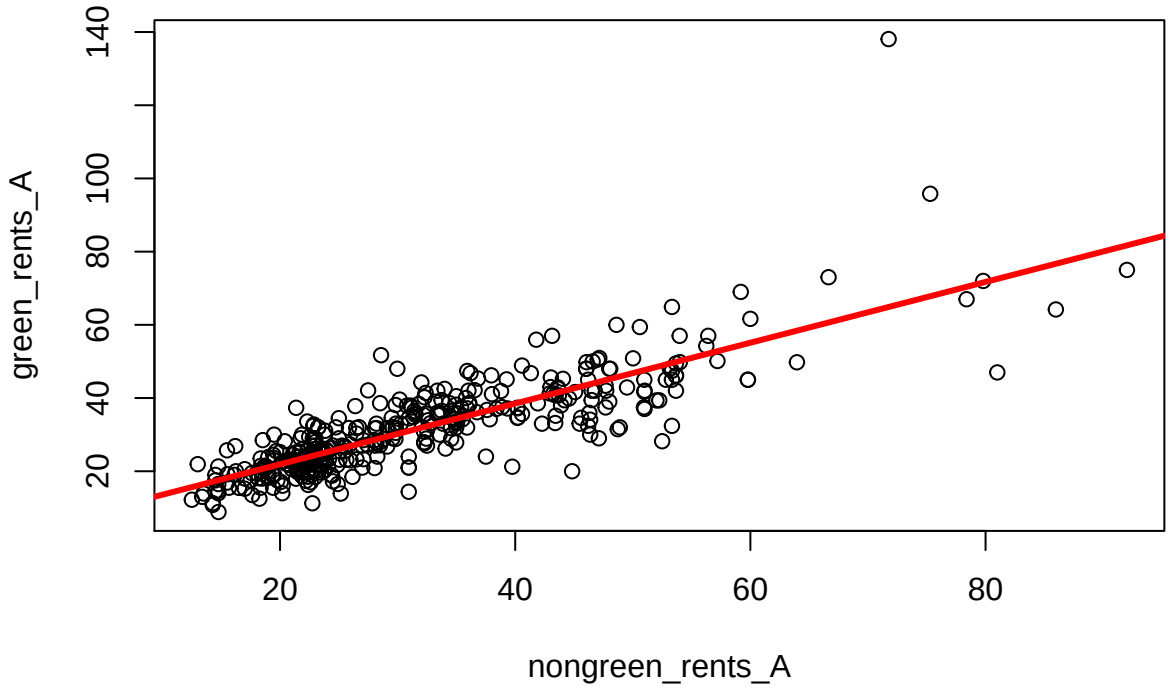



```
##
## Call:
## lm(formula = Rent ~ cluster_rent, data = green_buildings_noOut)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4933  -3.3699  -0.6662   2.7645  31.4263
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.86976    0.71201   1.222    0.222
## cluster_rent  1.07716    0.02461  43.763   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.855 on 575 degrees of freedom
## Multiple R-squared:  0.7691, Adjusted R-squared:  0.7687
## F-statistic:  1915 on 1 and 575 DF,  p-value: < 2.2e-16
```

The coefficient relating average rent in a cluster to rent of the green building in that cluster is about 1.07716. This means that the linear model estimates that rent for a green building is, on average, 1.07716 times more than for a non-green building in the same area.

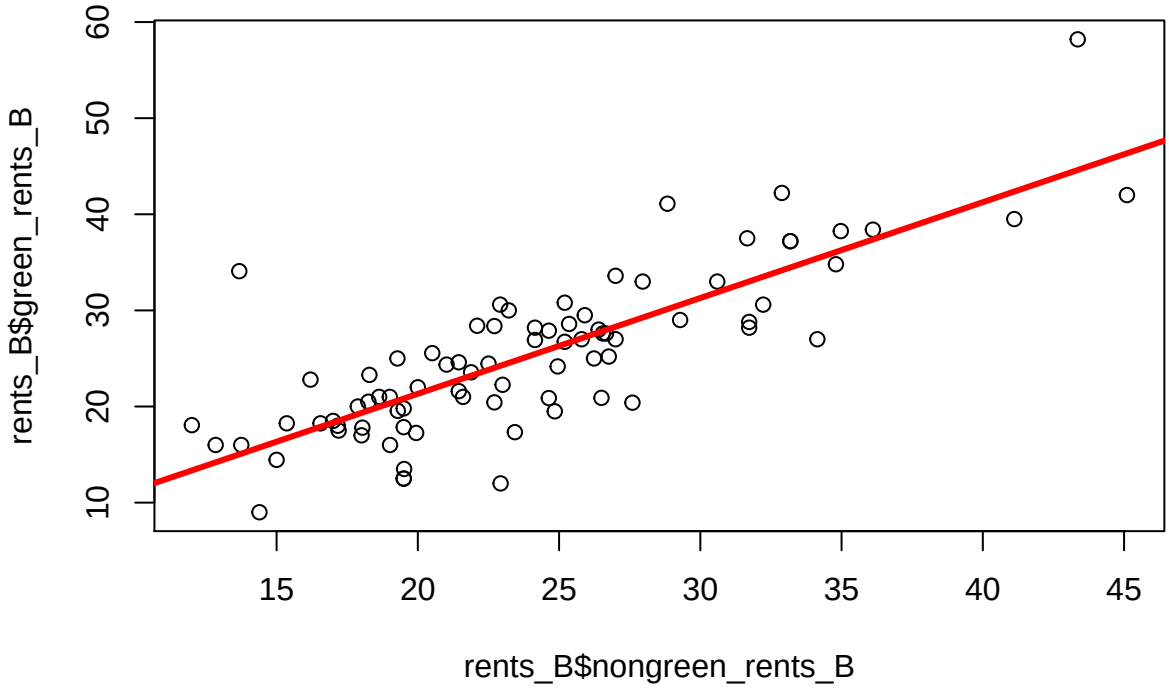
However, there are still other factors to consider. Each building in the data set is given a class A, B, or C rating. Class A rated buildings are considered to be the highest-quality properties in an area, followed by class B, and class C properties are the least desirable. In order to get an accurate idea of the premium for

green buildings, we should compare those buildings only to other buildings within the same class. Below, we compare green buildings to nongreen buildings within the same cluster and the same class using linear regressions.




```
##
## Call:
## lm(formula = green_rents_A ~ nongreen_rents_A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.548  -3.520  -0.306   3.298  73.198
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.28150    1.01601   5.198 3.34e-07 ***
## nongreen_rents_A  0.83046    0.02928  28.361  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.469 on 368 degrees of freedom
##   (98 observations deleted due to missingness)
## Multiple R-squared:  0.6861, Adjusted R-squared:  0.6852
## F-statistic: 804.3 on 1 and 368 DF,  p-value: < 2.2e-16
```

For class A properties, it appears that there is no premium for green buildings. In fact, with a coefficient of about 0.83, it seems that rents are typically higher in non-green buildings.




```
##
## Call:
## lm(formula = green_rents_B ~ nongreen_rents_B, data = rents_B)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2416  -2.7785   0.2628   2.2225  19.0581
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.38683    2.01722   0.687    0.494
## nongreen_rents_B  0.99672    0.08096  12.311   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.966 on 80 degrees of freedom
##   (21 observations deleted due to missingness)
## Multiple R-squared:  0.6545, Adjusted R-squared:  0.6502
## F-statistic: 151.6 on 1 and 80 DF,  p-value: < 2.2e-16
```

For class B properties, the model produces a coefficient of about 1 for the relationship between green and non-green properties. This indicates that there is no premium for green buildings among class B properties either.

Also, looking at the plots, it is apparent that there are far more green properties with a class A rating than a class B rating. This could potentially explain the premium that we thought we saw from our previous

regression. By only comparing green properties to other buildings in their cluster regardless of class, we were likely comparing class A green buildings to cluster rent means that were watered down by class B buildings.

Given these new findings, we are forced to disagree with the stats guru and conclude that it is likely not a good financial move to build the green building.

Bootstrapping

In order to construct our portfolios, the first step is to obtain the performance data for the exchange-traded funds that we are considering. We chose to pull daily data from September 2004 to the present because that is the first year that yahoo has data for VNQ, one of the funds that we are considering, and because we felt this time frame would give us both good and bad runs of stock-market performance. The most recent data is shown below.

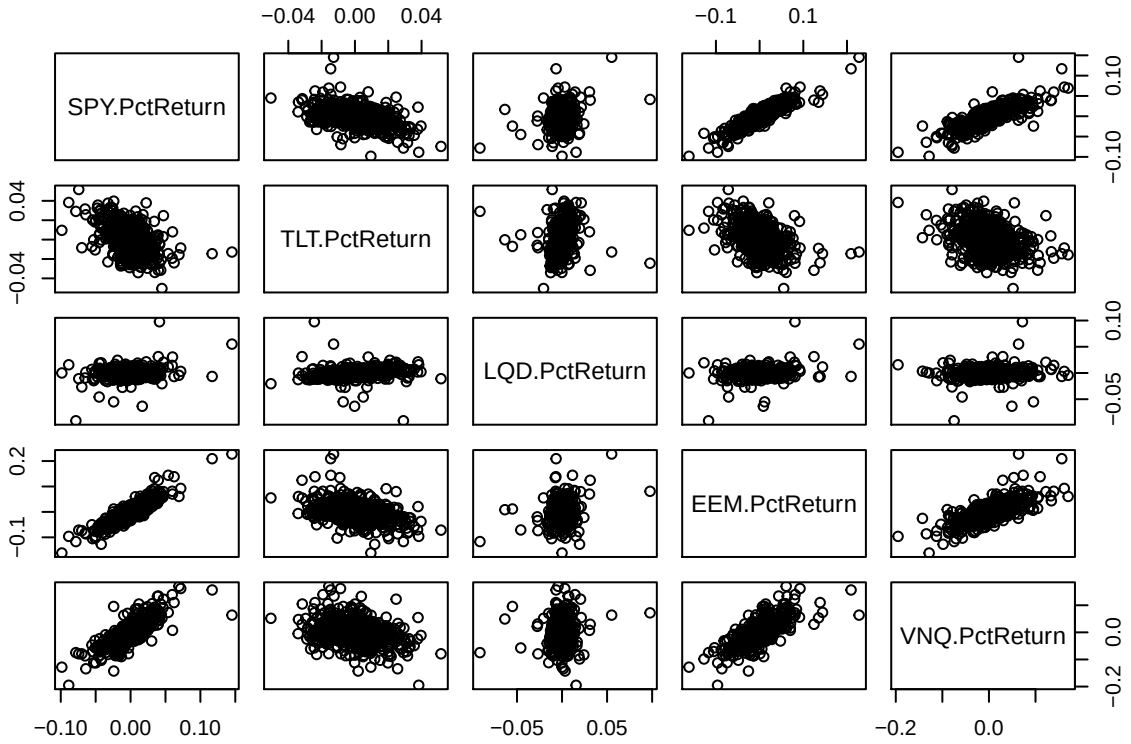
```
## GMT
##            SPY.Open SPY.High SPY.Low SPY.Close SPY.Volume SPY.Adj.Close
## 2016-07-26   216.53   217.17  215.76    216.75   70080500        216.75
## 2016-07-27   217.19   217.27  215.62    216.52   84083900        216.52
## 2016-07-28   216.29   217.11  215.75    216.77   65035700        216.77
## 2016-07-29   216.46   217.54  216.13    217.12   79519400        217.12
## 2016-08-01   217.19   217.65  216.41    216.94   73311400        216.94
## 2016-08-02   216.65   216.83  214.57    215.55   92295500        215.55
##            TLT.Open TLT.High TLT.Low TLT.Close TLT.Volume TLT.Adj.Close
## 2016-07-26   139.52   139.52  138.46    138.92    5243400      138.6668
## 2016-07-27   139.55   140.65  139.40    140.65    8508100      140.3937
## 2016-07-28   139.80   140.84  139.71    140.39    7336100      140.1341
## 2016-07-29   140.45   141.68  140.22    141.56    8766700      141.3020
## 2016-08-01   139.78   140.50  139.61    139.77    9809300      139.7700
## 2016-08-02   137.83   139.26  137.50    138.33   10051200      138.3300
##            LQD.Open LQD.High LQD.Low LQD.Close LQD.Volume LQD.Adj.Close
## 2016-07-26   123.56   123.56  122.97    123.15    4929300      122.8292
## 2016-07-27   123.39   123.74  123.20    123.66    5976800      123.3379
## 2016-07-28   123.39   123.62  123.25    123.54    5276100      123.2182
## 2016-07-29   123.50   124.04  123.50    123.99    4226100      123.6670
## 2016-08-01   123.32   123.44  122.95    122.97    4827700      122.9700
## 2016-08-02   122.43   122.67  122.11    122.28    4526700      122.2800
##            EEM.Open EEM.High EEM.Low EEM.Close EEM.Volume EEM.Adj.Close
## 2016-07-26    35.81    35.96   35.76     35.89   63593800         35.89
## 2016-07-27    35.95    36.12   35.65     36.01   67888600         36.01
## 2016-07-28    35.95    36.02   35.79     36.02   45590000         36.02
## 2016-07-29    36.03    36.25   35.87     36.21   85265700         36.21
## 2016-08-01    36.30    36.36   36.09     36.14   57794000         36.14
## 2016-08-02    36.08    36.17   35.62     35.88   81854600         35.88
##            VNQ.Open VNQ.High VNQ.Low VNQ.Close VNQ.Volume VNQ.Adj.Close
## 2016-07-26    91.69    91.97   90.94     91.16    2460200         91.16
## 2016-07-27    91.14    91.14   89.70     90.45    3935200         90.45
## 2016-07-28    90.31    91.65   90.10     91.29    2853100         91.29
## 2016-07-29    91.54    92.92   91.20     92.45    3502600         92.45
## 2016-08-01    92.33    92.75   92.26     92.65    3335100         92.65
## 2016-08-02    92.38    92.54   91.00     91.17    3877100         91.17
```

The dataframe provides several daily values for each stock: opening price, high price, low price, closing price, volume, and adjusted closing price. We can use this data to create a new dataframe with the percent return

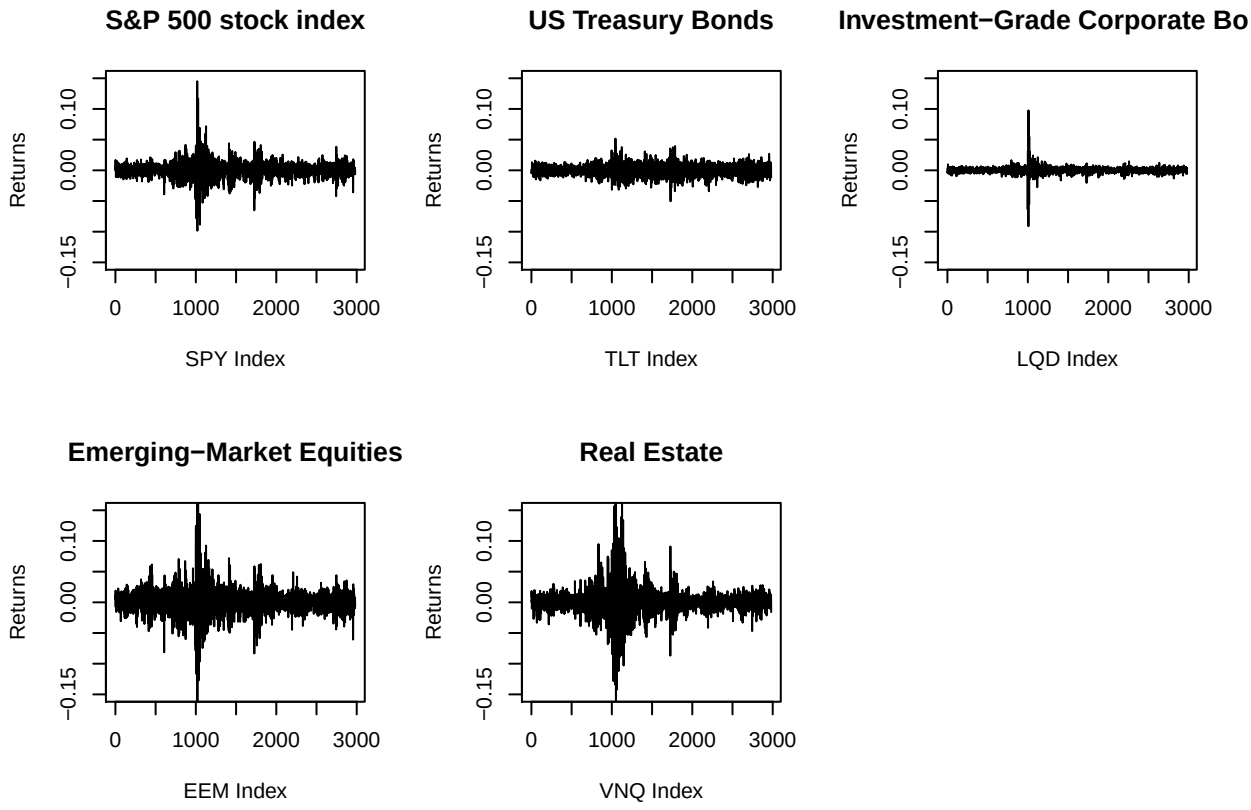
on each stock for each day in our time range.

```
##            SPY.PctReturn TLT.PctReturn LQD.PctReturn EEM.PctReturn
## 2016-07-26  0.0004616017   0.001586164 -0.0009734371  0.0078629035
## 2016-07-27 -0.0010611119   0.012453182  0.0041413119  0.0033435220
## 2016-07-28  0.0011546277  -0.001848524 -0.0009704238  0.0002777562
## 2016-07-29  0.0016145730   0.008333915  0.0036425145  0.0052748195
## 2016-08-01 -0.0008290024  -0.010841998 -0.0056361034 -0.0019331677
## 2016-08-02 -0.0064072969  -0.010302654 -0.0056111409 -0.0071941895
##            VNQ.PctReturn
## 2016-07-26  -0.004912095
## 2016-07-27  -0.007788580
## 2016-07-28   0.009286943
## 2016-07-29   0.012706715
## 2016-08-01   0.002163386
## 2016-08-02  -0.015974139
```

Let's look at a pairwise scatter plot of the returns for each stock to see how they may be related to one another.



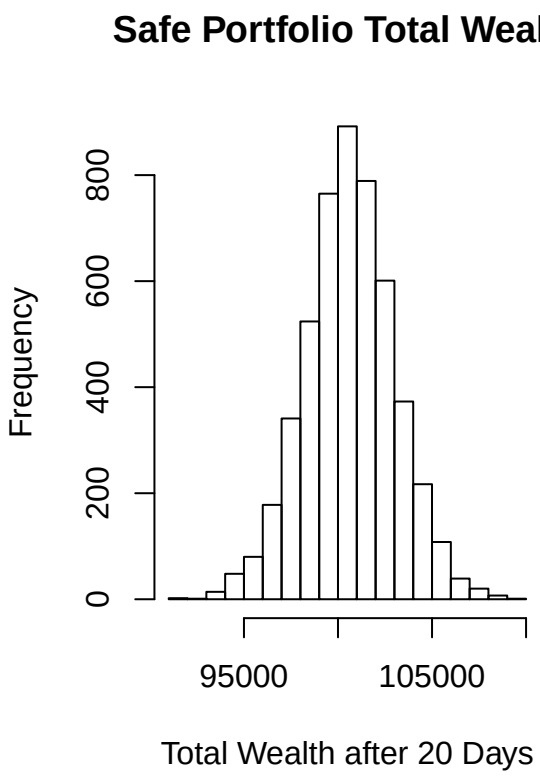
It looks like the strongest positive correlations are between US domestic equities (SPY), emerging-market equities (EEM), and real estate (VNQ). This could be interepreted to mean that these three funds are more volatile, changing more as a result of overall market fluctuations. US treasury bonds (TLT) and investment-grade corporate bonds (LQD) do not seem to have as strong of correlations with the other funds, indicating that perhaps these funds are safer, withstanding fluctuations in the overall market. Let's take a closer look by plotting the daily returns for each stock over time.

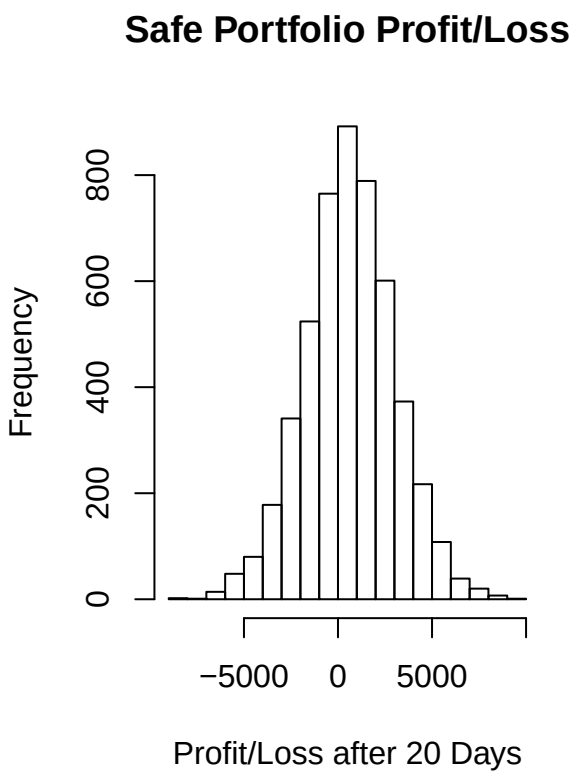


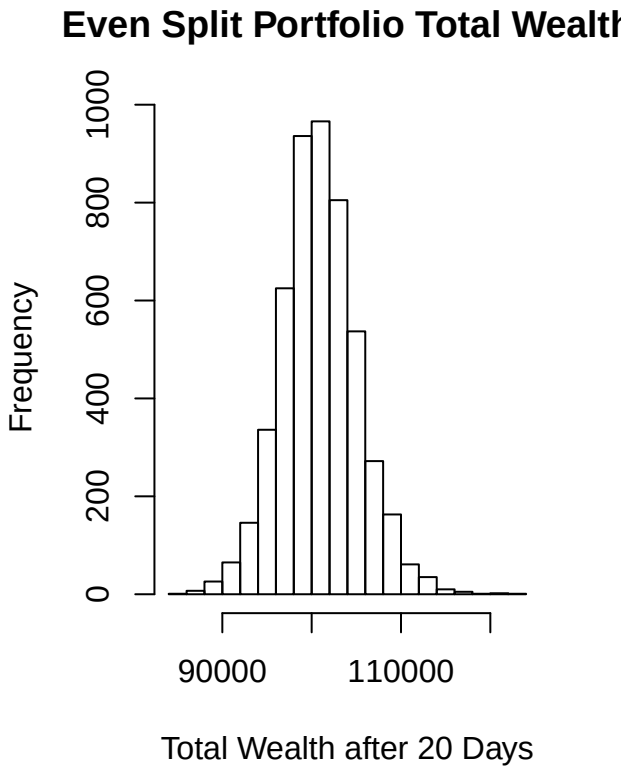
From the above plots, we can get a sense for the volatility of each fund from the range of returns that we see. As expected, US treasury bonds and investment-grade corporate bonds appear to be the least volatile, while the remaining three funds have consistently higher fluctuations in stock price. Emerging-market equities and real estate look to clearly be the most volatile, slightly more so than the the S&P 500. This leads us to the conclusion that US treasury bonds and investment-grade corporate bonds are safer investments with less risk but also less potential for high returns. Emerging-market equities and real estate are riskier funds with more risk and more potential for high returns. The S&P 500 lies somewhere in the middle.

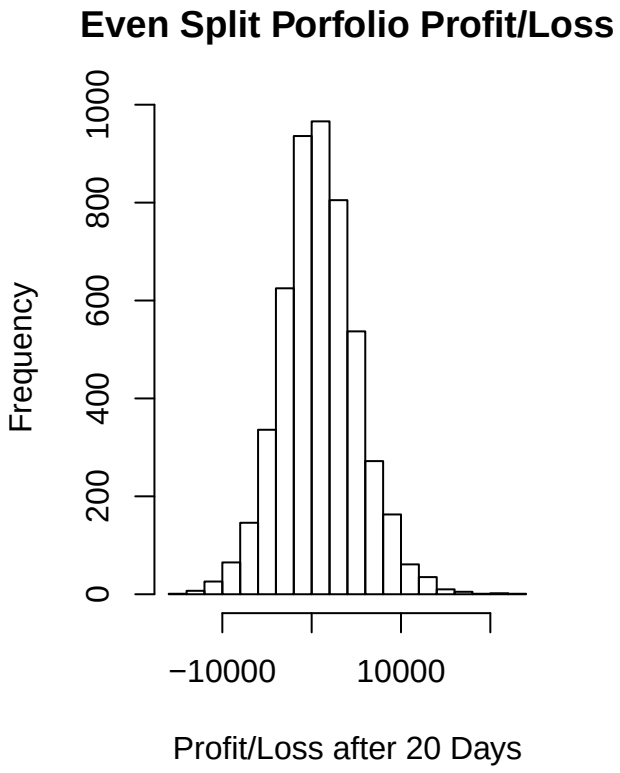
Now let's create 3 different portfolios to compare: one "safe" portfolio, one "aggressive" portfolio, and one porfolio with our assets evenly distributed between the 5 funds. For the "safe" portfolio, we will allocate 45% of our assets to US treasury bonds, 45% to investment-grade corporate bonds, and 10% to the S&P 500. For our "aggressive" portfolio, we will distribute 50% to emerging market equities and 50% to real estate. Each portfolio will contain $100,000 total.

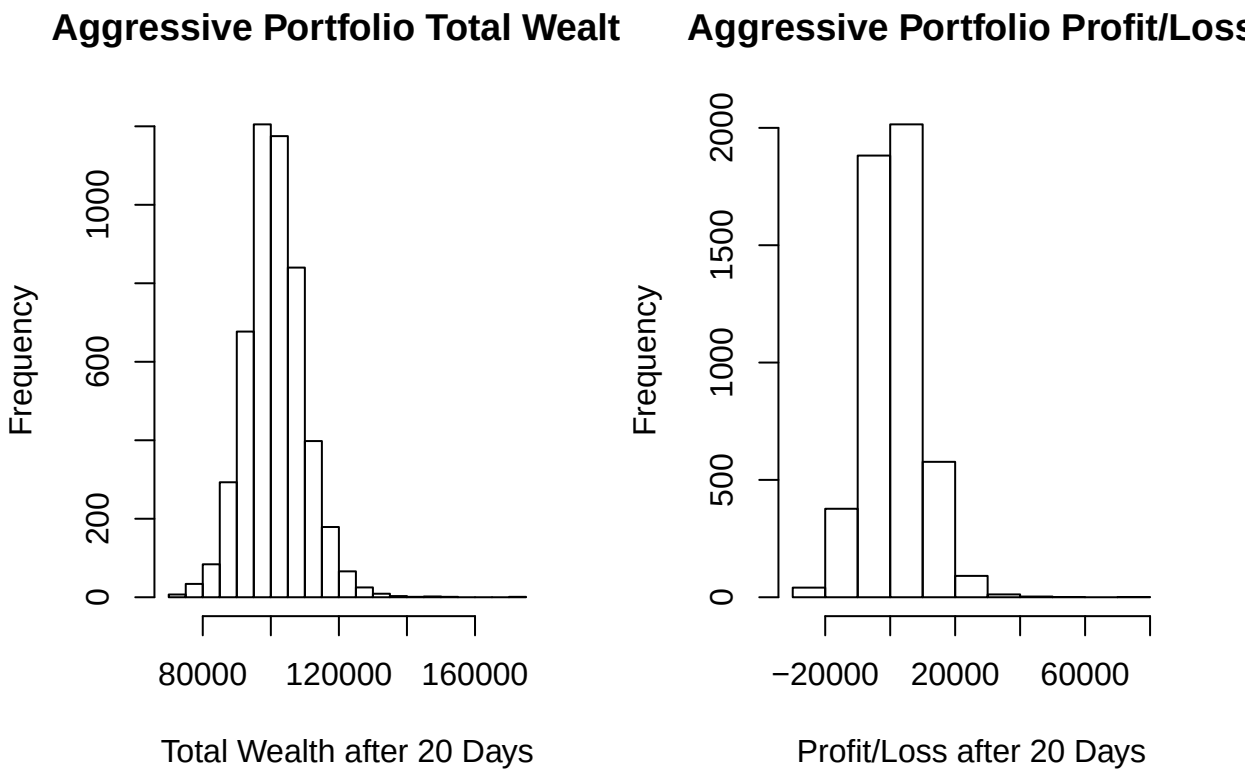
For each portfolio, we will project returns over a 4-week (20-day) period 5,000 times. Each time, we will calculate the value of the portfolio's holdings after the 4-week period and the profit or loss. The histograms below represent the distributions of each calculation over the 5,000 simulations.


Safe Portfolio Total Wealth
Frequency
0
200
400
600
800
95000
105000
Total Wealth after 20 Days


Safe Portfolio Profit/Loss
Frequency
0
200
400
600
800
−5000
0
5000
Profit/Loss after 20 Days


Even Split Portfolio Total Wealth
Frequency
0
200
400
600
800
1000
90000
110000
Total Wealth after 20 Days


Even Split Porfolio Profit/Loss
Frequency
0
200
400
600
800
1000
−10000
10000
Profit/Loss after 20 Days



For each portfolio, the majority of simulations resulted in a profit around 0 with slightly more yielding positive results than negative. However, as one would expect, the range of outcomes varies greatly. The range for the safe portfolio is only about $10,000, varying from a $5,000 loss to a $5,000 profit. For the evenly split portfolio, the range approximately doubles with a maximum profit and loss of about $10,000. The range approximately doubles again for the aggressive portfolio, which ranges from a $20,000 loss to a $20,000 gain. The best choice of the three portfolios depends on the preferences of the investor.

Market Segmentation

In preparing the data, We only took out the ID column in case we could find interesting results about how many bots were tweeting.

There's a large portion of users that fall under chatter, which seems to suggest that a lot of users are'nt talking about any coherent thing that they can market towards. Aside from that, a lot of their users seem to be interested in health, nutrition, personal fitness, cooking, and sharing photos.

We create a new category called healthy_lifestyle that aggregates the correlated interests of health, nutrition, and personal fitness.

You could consider people interested in health, nutrition, and personal fitness to be a market segment under the umbrella of wanting to live a healthy lifestyle. When we rerun the model with this new market segment, we clearly see that most of NutrientH20's users are interested in a healthy lifestyle, and this should be the company's main demographic for marketing efforts.