HW1

Tyler Nicholas

August 2, 2016

Question 1

There is a survey upon entering a website that asks a yes or no question. We are given that the probablility that it is a Random Clicker (RC) is 0.3. This implies that P(TC) or Truthful clicker is equal to 0.7. We are also given that the results show P(Y) or probability of Yes is equal to 0.65 and thus P(N) or probability of No is equal to 0.35.

We wish to discover The probability of a yes, given that it was a truthful clicker or P(Y|TC). So we set up the formula using the Law of Total Probability:

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

We input our known values to obtain:

$0.65 = (0.5)(0.3) + P(Y|TC)(0.7)$

We then solve for P(Y|TC) and obtain $P(Y|TC) = 0.7143$

So we can conclude based on the given data that 71.43% of the Truthful clickers selected yes to the survey question.

Question 2

To begin we will name the variables: TP = Test Positive, TN = Test Negative, D = Have Disease, ND = Not Having Disease

We are given that the sensitivity is 0.993 or $P(TP|D) = 0.993$. We also are given the specificity is 0.9999 or $P(TN|ND) = 0.9999$. The last piece of information is that $P(D) = .000025$.

We wish to find the probability that you have the disease, given that you test positive. We say this as P(D|TP). So we begin with this formula:

$P(D|TP) = P(TP,D)/P(TP) = (P(D)P(TP|D)) \ / \ P(TP)$

Then by the law of total probability:

$(P(D)P(TP|D)) \ / \ P(TP) = (P(D)P(TP|D)) \ / \ (P(TP|D)P(D) + P(TP|ND)P(ND))$

We then input the known data to obtain:

$P(D|TP) = (0.000025)(0.993) \ / \ ((0.993)(0.000025) + (0.007)(0.999975))$

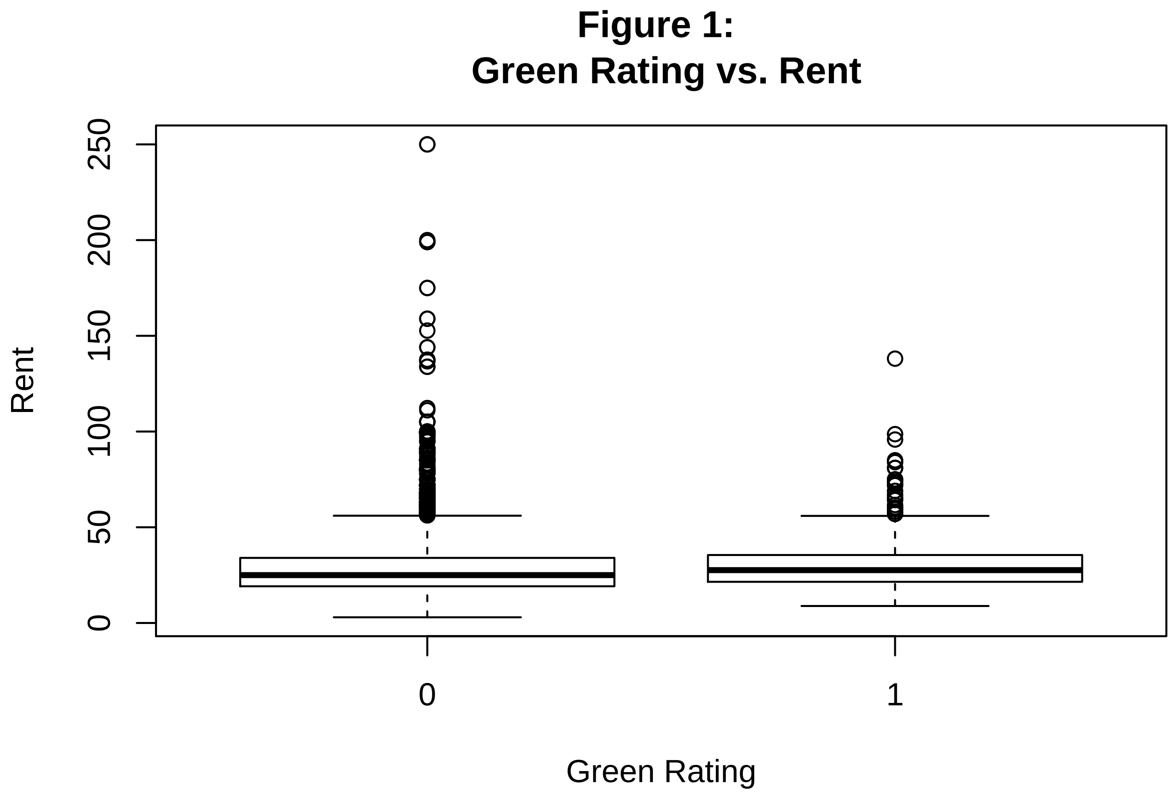
Thus $P(D|TP) = .003533984$

This result tells us that if you have a positive test, the chances that you actually have the disease is only 0.3533984%. This means that ~99.7% of people who get a positive test do not in fact have the disease. We can infer from this data that it is not a good test. There will be far more false positives than actual positives if the test is implemented universally.

Exploratory Analysis: Green Buildings

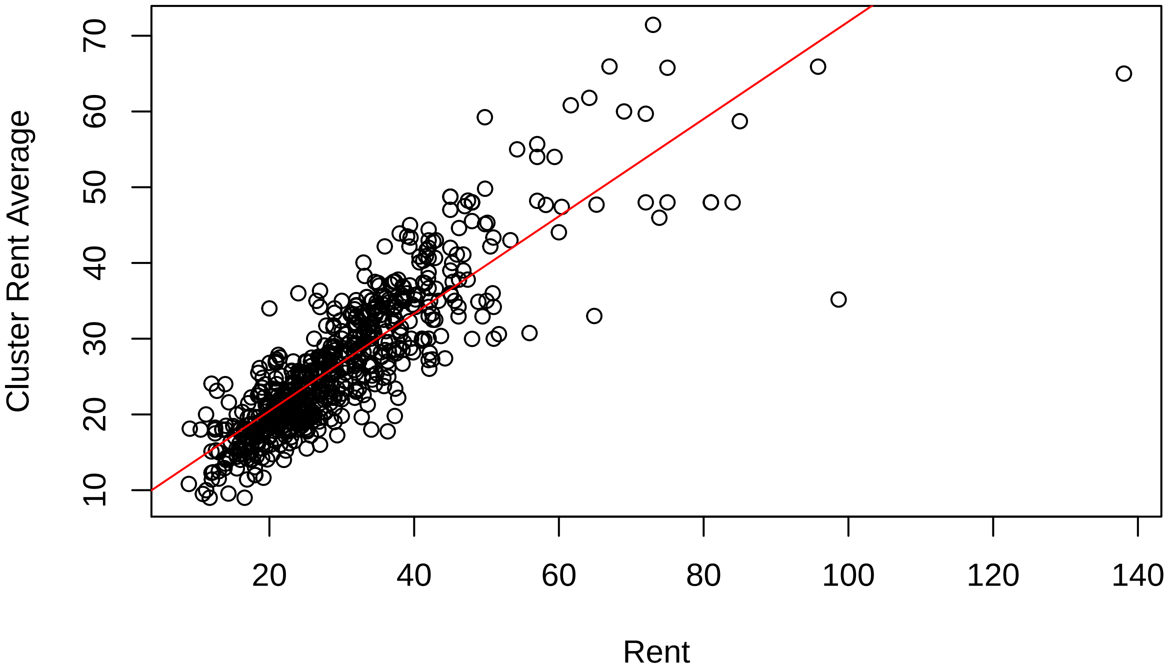
The first issue with the staffers report is that he left out data that could be meaningful. Buildings with a low occupancy rate could signify the value of buildings in the area. For our analysis, we will leave those buildings in our data.

The main claim that was made by the staffer was that buildings that are rated green have a higher median rent value than buildings that are not green. We can see in Figure 1 that the medians appear to be very similar. In fact we can see that the medians are within one standard deviation of each other and the difference is therefore not statistically significant.



To further support the above assertion, we plot the Green Rent vs. the average rent in each cluster in Figure 2. The best fit regression line has a slope of .64 which would imply that green buildings on average have a higher rent than others in their cluster.

Figure 2:
Green Rent vs. Cluster Rent



Now we create a dataframe that only includes green properties that have lower rent than the average rent for their cluster. We see here that 164 clusters contain green buildings with rent lower than average for the cluster. So even though green buildings on average have higher rent than other buildings in their cluster, roughly 25% of the time, the green buildings have rent lower than average for their cluster. Due to this fact, we need to know in which cluster the developer is building the building before we can state whether being green will have a positive or negative effect.

When we view the correlation table, there is not a strong correlation between rent and green_rating. The correlation is only .0326. The other report was ignoring all other variables that can be factors in rent price such as the class of the building or the size of the building.

In summary we cannot say that they should buy the building. We need to know which cluster represents the neighborhood that the new building is in, then we can see if there is a premium for Green Ratings in that neighborhood. Overall, there does not appear to be a significant difference in prices of Green Rated buildings, so we can reccomend not buying a Green building since it has a 5% premium.

Bootstrapping

To view the risk/return properties of each of these classes, we can view Figures 3-7 and see the variance associated with each stock.

Figure 3:
Close Prices of SPY Over Time

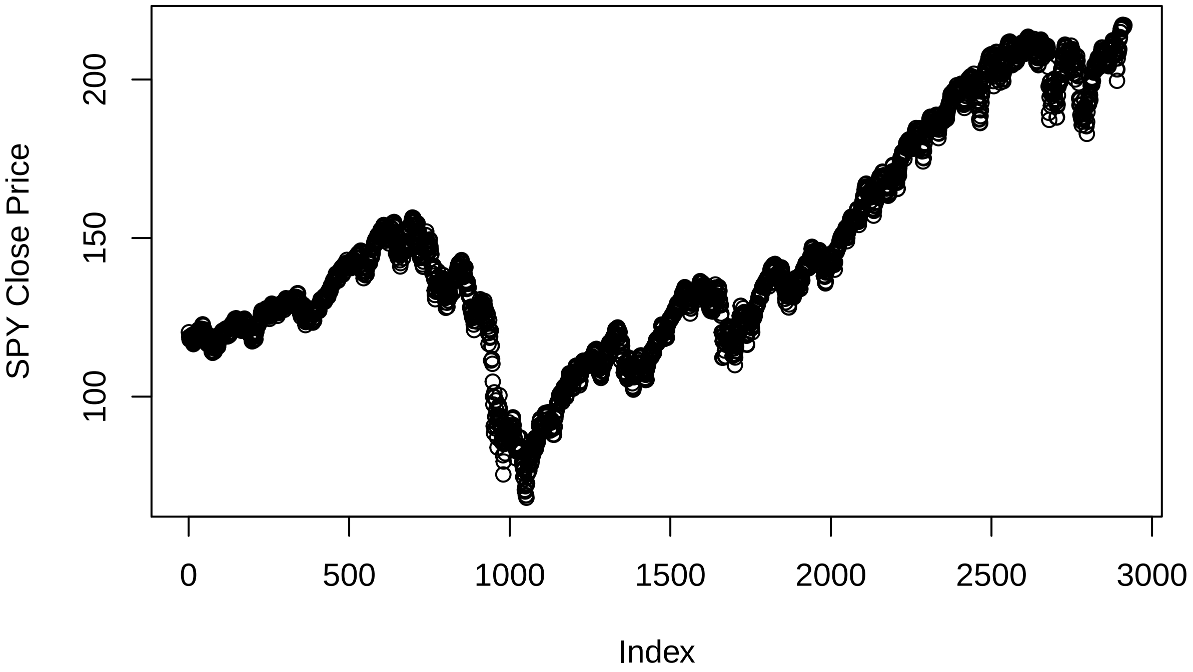


Figure 4:
Close Prices of TLT Over Time

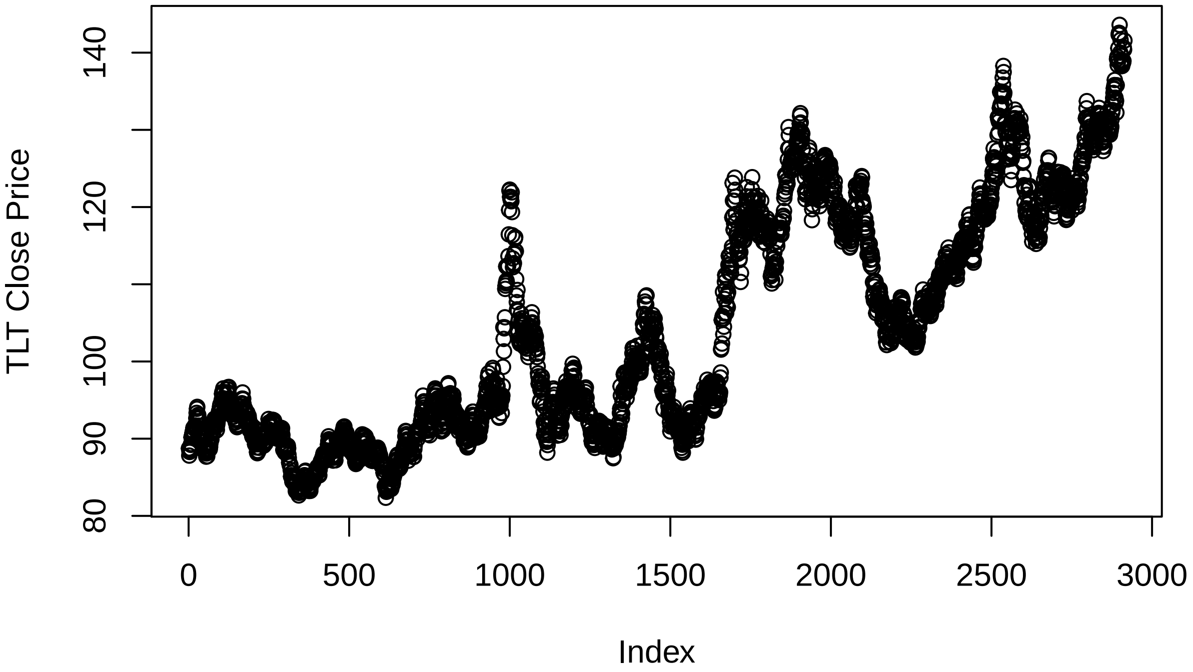


Figure 5:
Close Prices of LQD Over Time

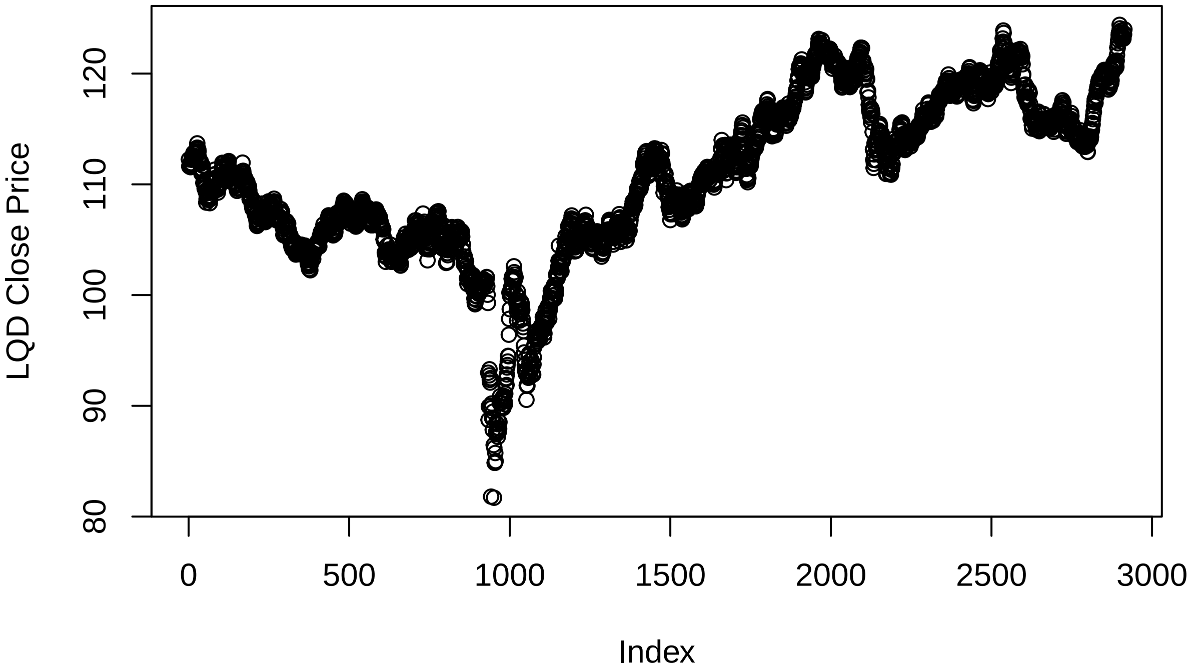


Figure 6:
Close Prices of EEM Over Time

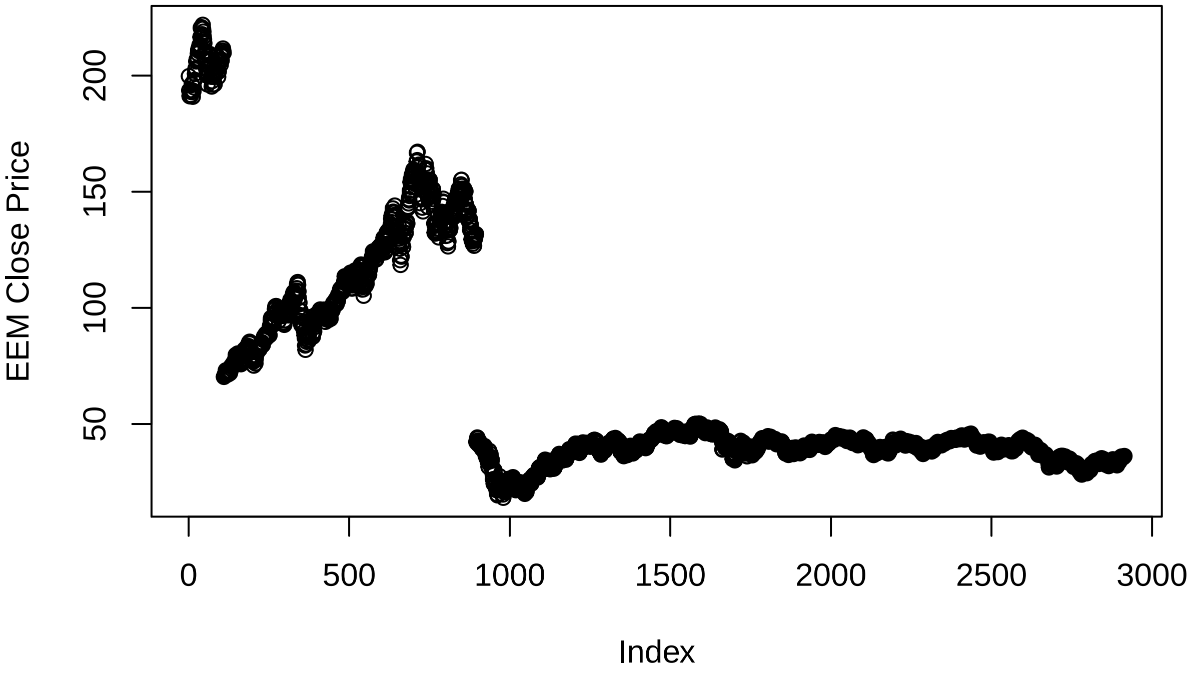
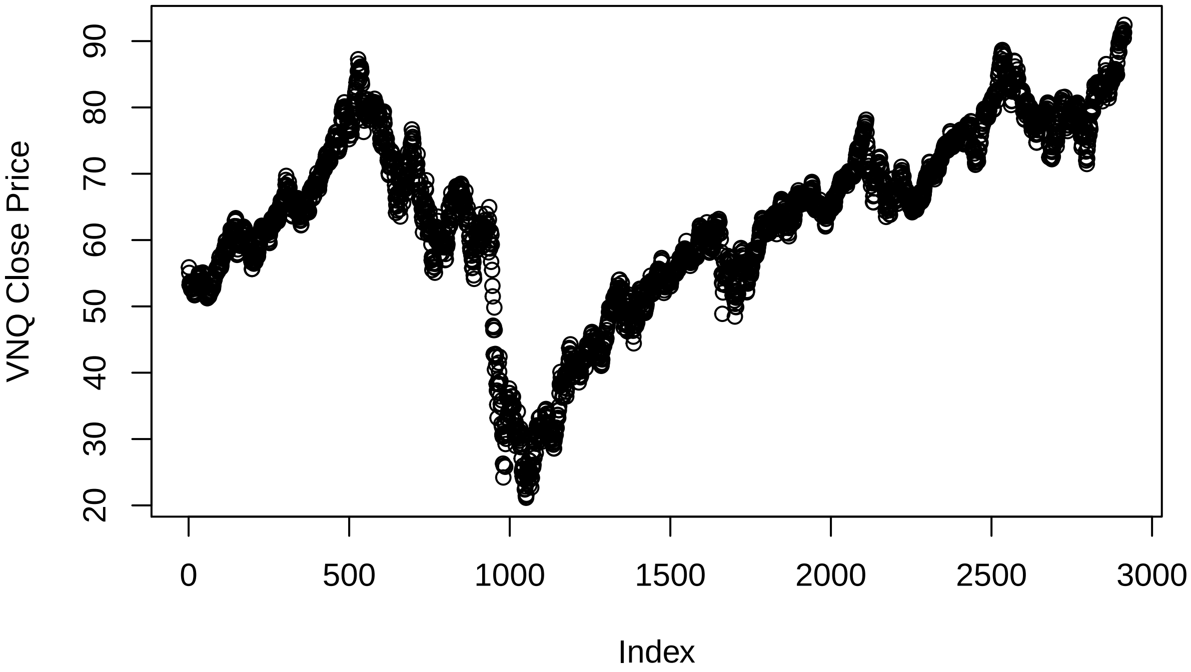


Figure 7:
Close Prices of VNQ Over Time

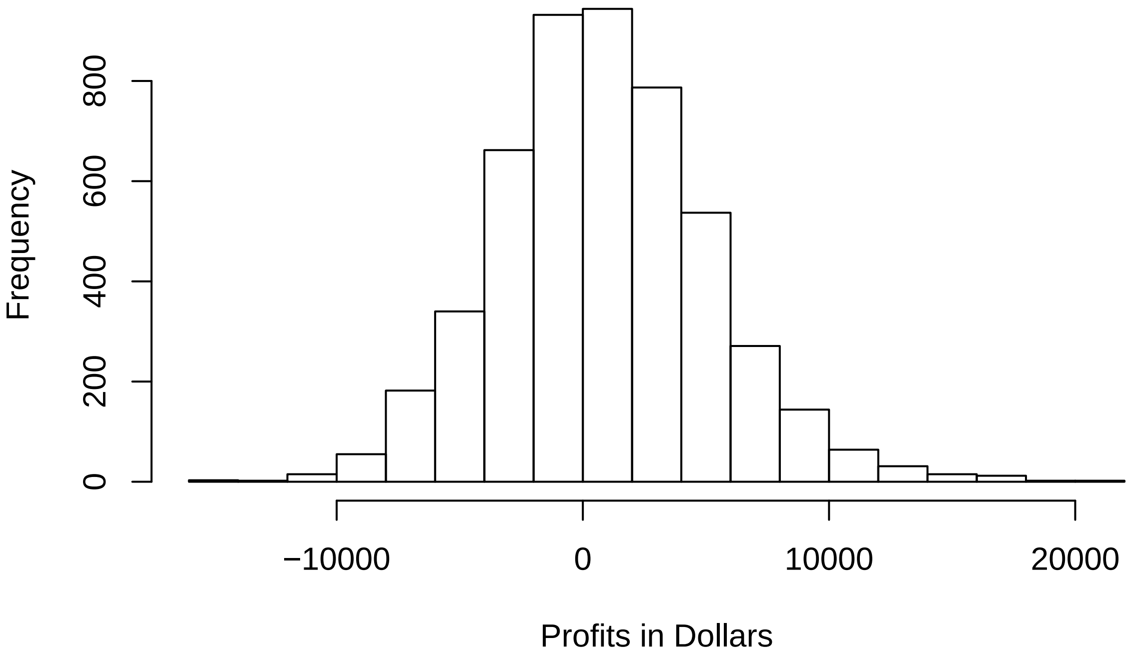


We can see from the plot and the standard deviations that we calculated on the closes for each ETF that the SPY and EEM are higher risk/return stocks. While the TLT,LQD and VNQ are all safer stocks. We will use this knowledge to determine our 3 different portfolios.

For our agressive portfolio we will take 50% each of our high risk stocks, SPY and EEM. We chose this as both were stocks with high variance over time. For our safe portfolio, we use the 3 low risk ETF's that we identified, TLT, LQD, and VNQ. Since LQD had by far the lowest variance, we will use 60% LQD and 20% each of TLT and VNQ. This portfolio will be safer as we have lower variances on the close prices for each of the 3 ETF's chosen.

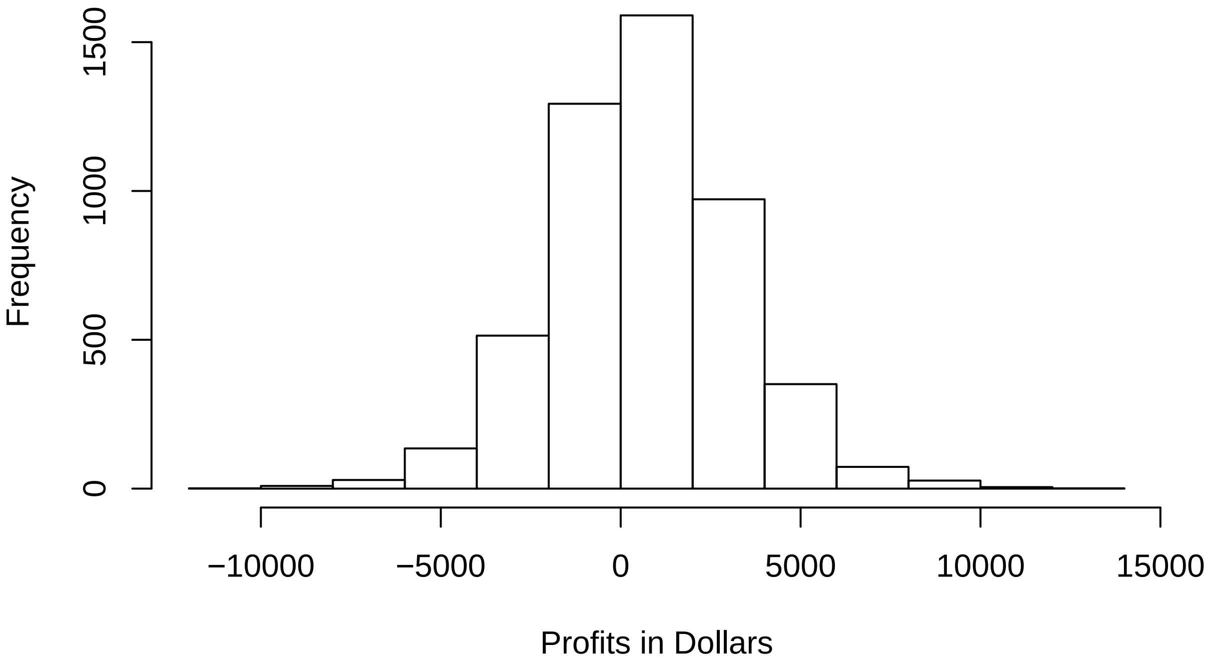
We run the bootstrapping method on our first portfolio where we have an even split of all stocks. We will then calculate the 5% value at risk. For our even split portfolio, the 5% value at risk is -5,710.371. Figure 8 shows the returns for 5000 20-day periods found by the bootstrap.

Figure 8:
Histogram of Returns for Even Split Portfolio



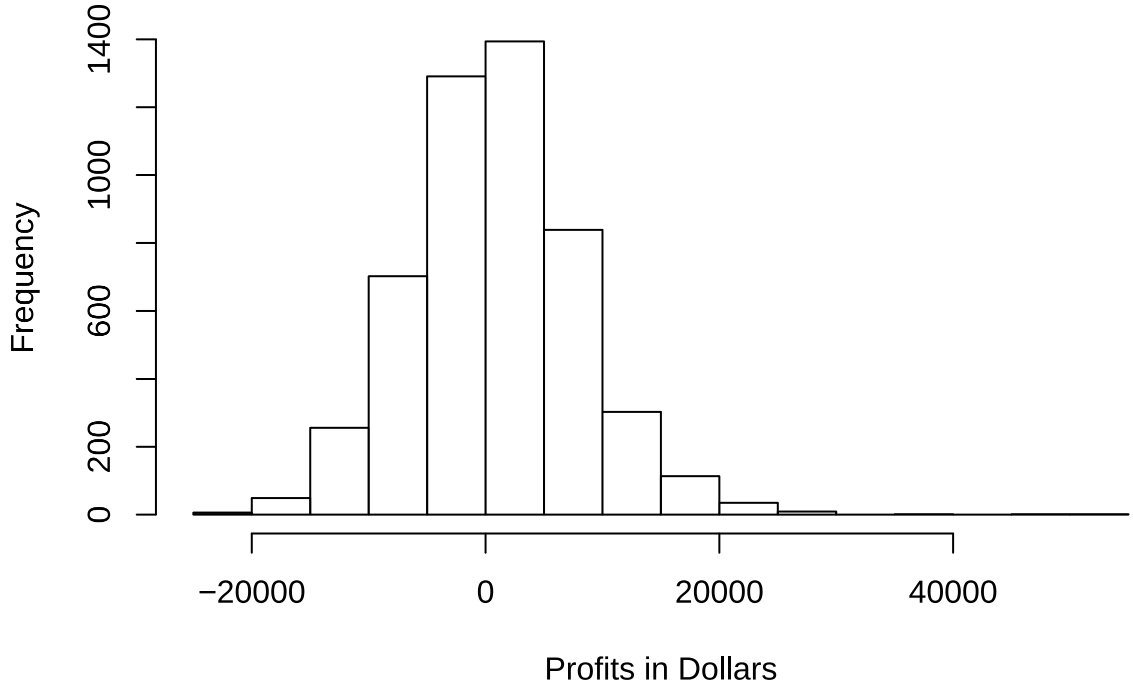
Now we run it with our safer split portfolio. Here we use 60% LQD, 20% VNQ, and 20% TLT. We chose these by looking at the standard deviations and choosing the stocks with the least deviations. Here the 5% value at risk is -3,507.24. Figure 9 shows the returns for 5000 20-day periods found by the bootstrap.

Figure 9:
Histogram of Returns for Safe Portfolio



Now we create our aggressive portfolio. We allocate 50% each to our two riskier stocks, namely SPY and EEM. We found they were higher risk by seeing that they had much larger standard deviations in their close prices than the other ETF's we analyzed. Here the 5% value at risk calculated is -10,030.47. Figure 10 shows the returns for 5000 20-day periods found by the bootstrap.

Figure 10:
Histogram of Returns for Riskier Portfolio



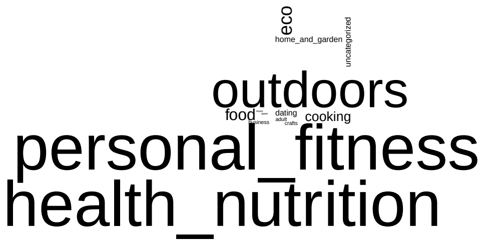
Viewing the results from each portfolio and the 5% value at risk for each portfolio, we can see there are options for multiple different kinds of investors. Viewing Figures 8, 9, and 10, we find that on average we generally do make money over a 20 day period for each portfolio. For those who are risk averse, the safe portfolio is recommended. On average you will make less money than with the other portfolios but you also have much less risk in how much you may lose.

For those who are risk takers, the riskier portfolio will give higher returns on average than the other two portfolios but will also introduce you to a higher risk, as it has the highest 5% value at risk of any of the portfolios.

For those looking to diversify, we have an even split portfolio that maximizes profits while minimizing risk. This tradeoff will leave you with returns in between the other two portfolios while also giving you a 5% value at risk between the other two portfolios.

Market Segmentation

We run a kmeans algorithm with 8 clusters to find interesting trends in the twitter followers of NutrientH20. The 2 largest clusters seem to be pulling in mostly chatter, adult, and spam messages. The next largest cluster of 799 people represents the largest cluster of meaningful tweets. The three most relevant topics from this cluster are Personal Fitness, Health Nutrition and Outdoors. This can be seen in the wordcloud below.



Cluster 1: none Cluster 2: fashion, beauty, cooking Cluster 2 size: 510 Cluster 3: online_gaming, college_uni, sports_playing Cluster 3 size: 376 Cluster 4: school, religion, parenting, sports_fandom Cluster 4 size: 349 Cluster 5: travel, politics, news, computers Cluster 5 size: 516 Cluster 6: personal_fitness, outdoors, health_nutrition Cluster 6 size: 799 Cluster 7: none Cluster 8: none

Now we run through a principal components analysis to see if we get similar results.

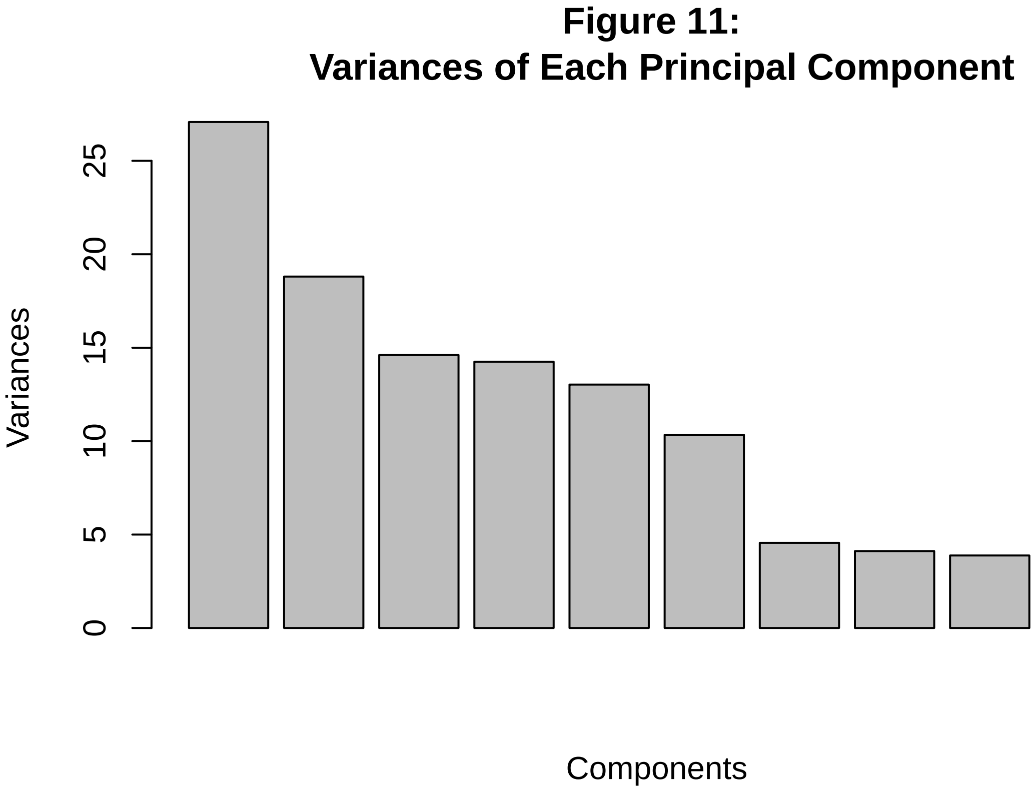


Figure 11 shows us the variance accounted for in each principal component. We can now look at the categories that are most important in each component. The first principle component is giving us most of the spam and adult content that is created by bots. We head to the second principal component to see what its most important categories are. We see that the most important categories in the second principal component are Personal Fitness, Health Nutrition and Outdoors.

These Findings agree with our kmeans analysis. From both of these methods, we have determined that the most important customers for NutrientH20 are interested in Personal Fitness, Health Nutrition and Outdoors. This market segment is the largest for the company and should be the one that is catered to by NutrientH20.