Homework 1

Caroline Nelson

August 7, 2016

Problem 1

Part A

We know:

- *P(RC)=0.3*
- *P(TC)=0.7*
- *P(Y|RC)=0.5*
- *P(N|RC)=0.5*
- *P(Y)=0.65*
- *P(N)=0.35*
- *P(Y|TC)=?*

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

0.65=0.7x+0.5(0.3)

0.5=0.7x

P(Y|TC)=5/7

About 71.4% of truthful clickers said 'yes'.

Part B

- T=Test Positive
- D=Have the disease
- T'=Test Negative
- D'= Don't have the disease

$Sensitivity = P(T|D) = 0.993$

$Specificity = P(T'|D') = 0.9999$

$P(D) = 0.000025$

P(D|T)=?

$P(T|D) = (P(T) * P(D|T))/P(D)$

$P(T) = P(T|D)P(D) + P(T|D')P(D')$

$= 0.993 * 0.000025 + (1 - 0.9999) * (1 - 0.000025)$
$= 0.0001248$

0.993= (0.0001248*x)/0.000025

$$P(D|T) = 0.198$$

This means that if you test positive for a disease, you have a 19.8% chance of actually having the disease. This does not sound like an ideal test for this specific disease, since its results are less than 20% accurate.

Problem 2:

Would investing in a green building be a good idea? Previous analysis from a consultant suggested that since rents are generally higher in green buildings, there would be a quick return on investment without paying too much more (5% premium). The charts below almost agree with this argument, but I will later discuss the confounding variables that were not considered.

Table 1:

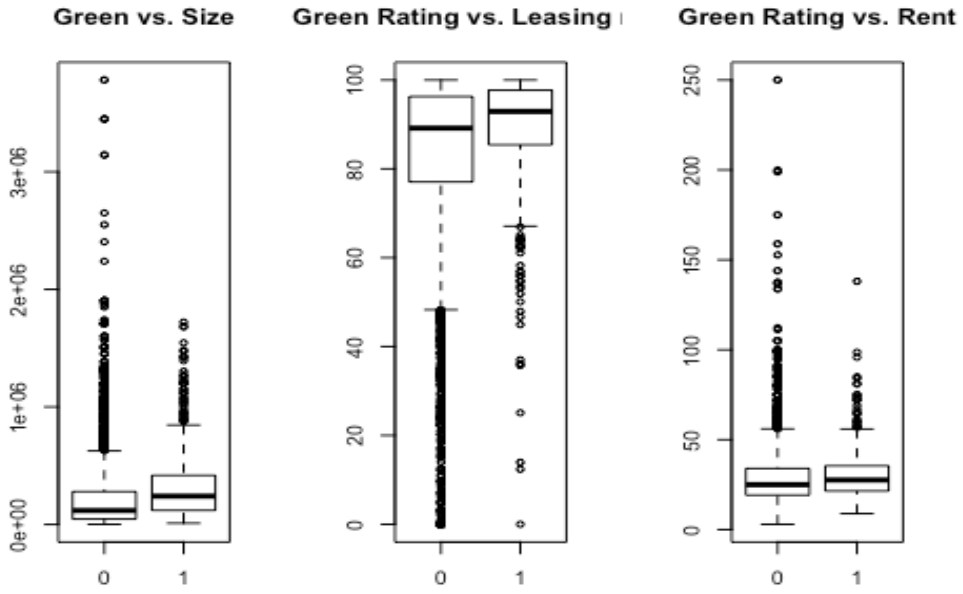


Table 1 enhance the consultant's argument that a green-rated building would bring more revenue. They show that while there is a small change in rent between green and non-green, there is a much smaller range of sizes for green buildings, so rent must overall be more expensive for green buildings, or so it appears. The plots also exhibit that the green-building is likely to be leased, which would also bring in more revenue. These arguments, however, are not big-picture arguments.

Counter-arguments

Non-green buildings have lots of outliers in size and rent costs. The statistics shown indicate that the average rent for green buildings is actually about the same as non-green buildings; however, there are lots of data points on the lower side of rent that bring the non-green buildings' median rent cost down. The median size of non-green buildings is 118,700 square feet, which is smaller than the 25th percentile of green buildings. We can conclude from this that rent costs and size counteract, so rent could be approximately the same for buidings with and without a green certification.

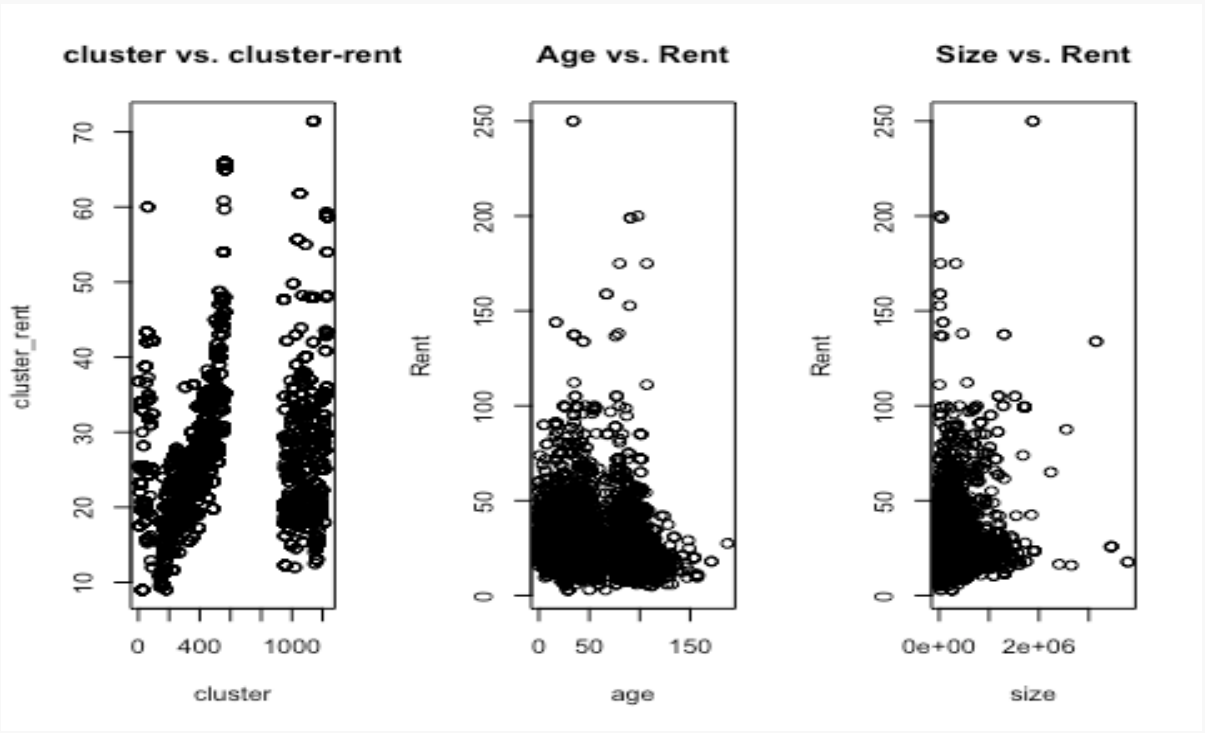


Table 2

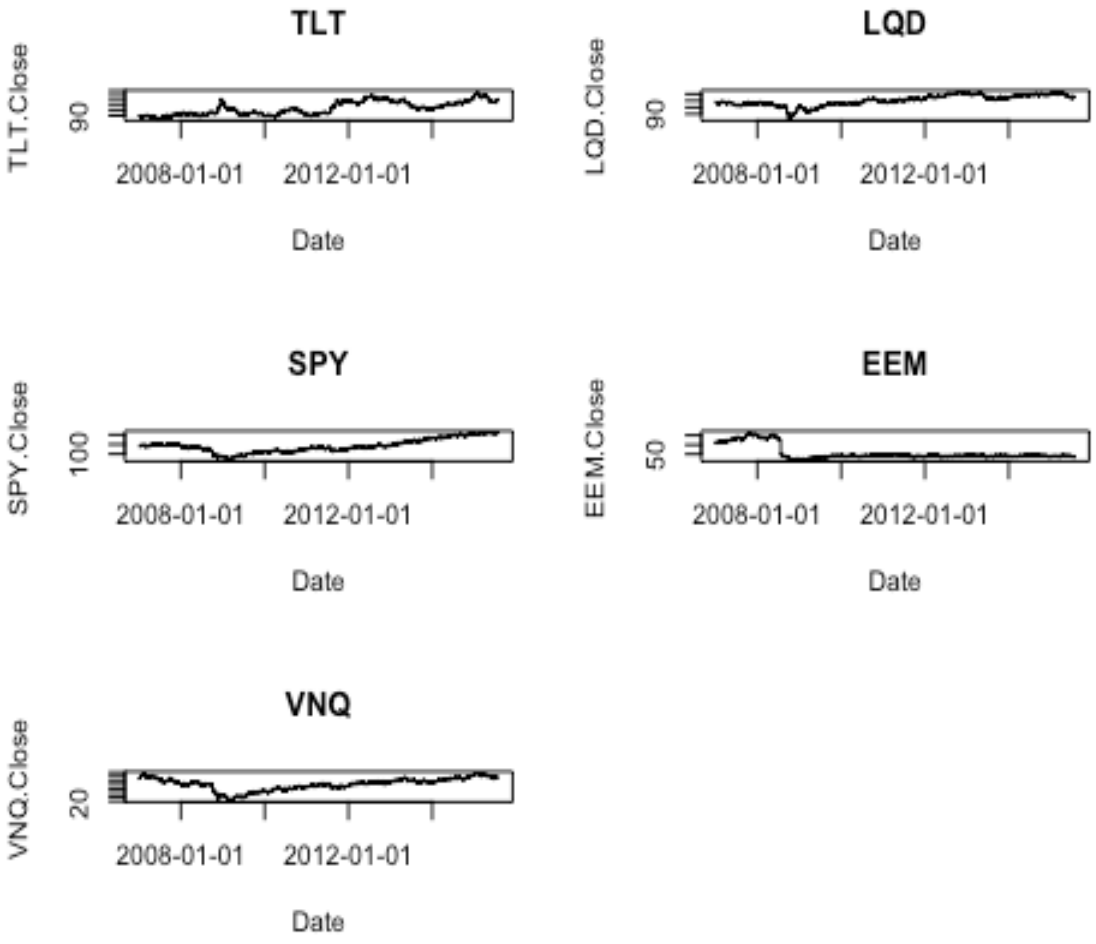
Conclusion

The initial consultant's results do not consider many factors in this problem. We see that rent and size almost counteract, but there are other factors, such as which cluster a building would be located, and those in Table 2 above. Location is very important in rent costs, as well as the factors we have looked at, and others, such as age and leasing rate. Whether or not a company should invest in a green building should take such factors into account. We saw that there is not a large difference between rent costs of a green-certified building, and non-green buildings. We also saw that the range of size of green buildings is much smaller than that of non-green buildings; this does enhance the consultant's argument, but there are other factors at play. The green buildings considered in the analysis could just be younger, newer buildings that can charge more for rent because of their age; they could also be in a cluster where rent is overall more expensive.

Problem 3:

In this problem, I constructed three different portfolios, one with all of my supposed assets divided evenly among five different stocks, another choosing only three safer stocks, and one choosing risky stocks. Using bootstrapping, we can estimate total wealth and value at risk over the course of 20 days. The plots of each stock from January 2008 to July 2015 are shown below.

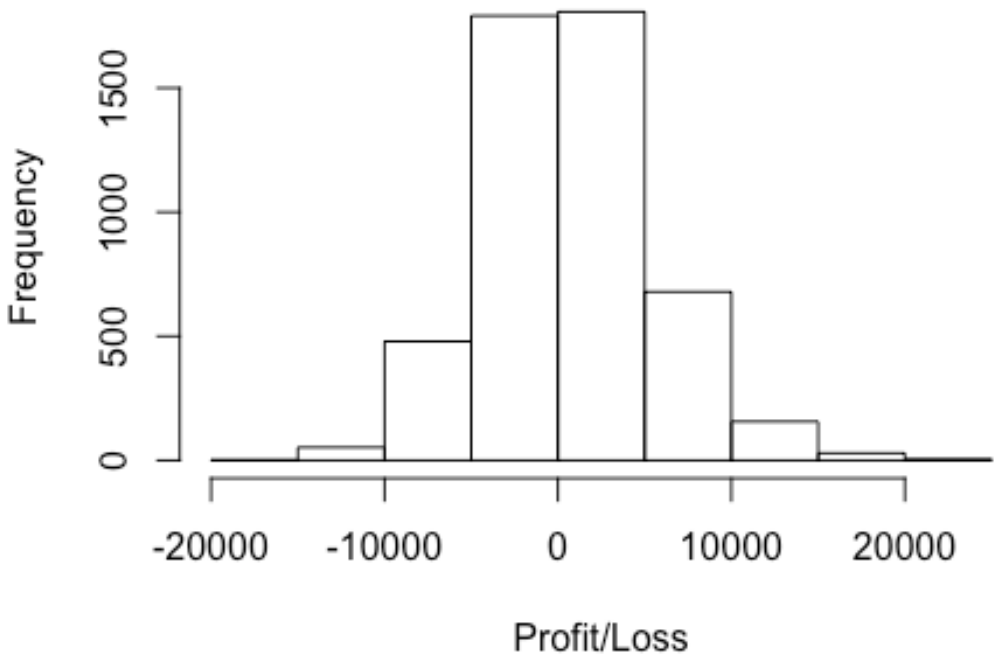
Table 3



From these graphs, we can see that:

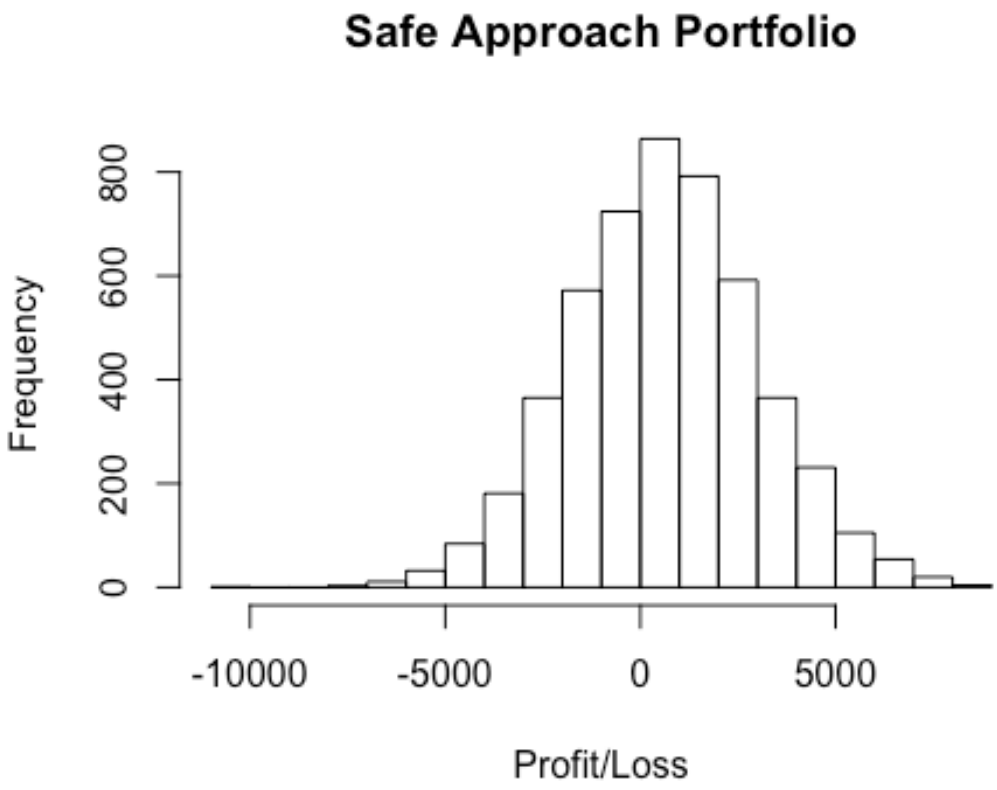
- EEM has had the most difficulty recovering from the crash in 2008.
- TLT has had lots of variation since 2012, but has somewhat recovered since 2008.
- LQD has had great success since 2008, and seems to be growing steadily.
- SPY has also steadily grown since the stock market crash, and is still growing.
- VNQ is slowly recovering, and seems to have recovered from the 2008 crash.

Evenly-split Approach



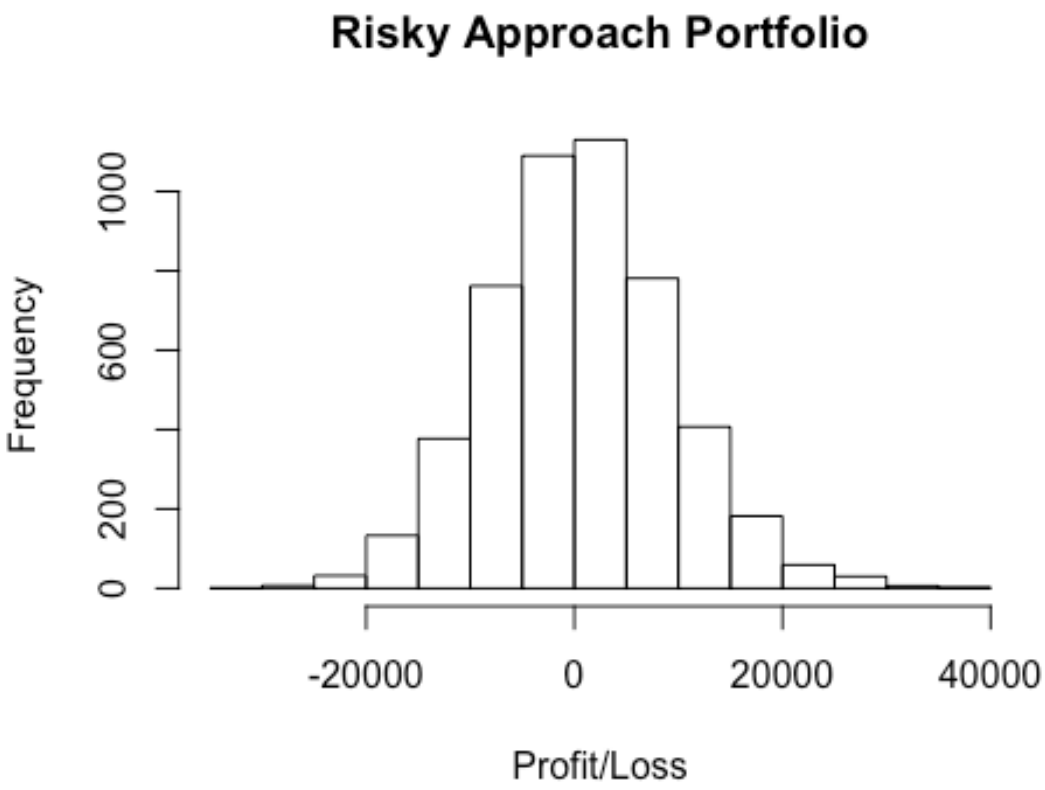
The evenly-split approach gives a total wealth of $105,016.70, and a value at risk of $6,967.83, as seen in this graph. There is more at risk when one takes on the variance of five different stocks.

Safe Approach



This safe portfolio took weights of 0.5, 0.25, and 0.25 to ETF's TLT, LQD, and SPY respectively. $3269.253 is the value at risk at the 5% level; this is very low, as expected. The total wealth decreased to $97,876.13 with this portfolio.

Risky Approach



The riskier portfolio looked at OIH and XLE (oil and energy stocks, respectively). Oil and energy stocks have fluctuated heavily the past couple of years, so they are a risky investment. Weights of 0.5, 0.25, and 0.25 were given to OIH, SPY and XLE respectively. The value at risk at the 5% level was $13,467.67. This was expected; however the total wealth was the highest of the three portfolios, at $107,572. Though the safe approach had the lowest value at risk, the risky portfolio gave the highest returns. Taking risks can sometimes have rewards.

Problem 4

In this problem, we are looking at a data set with each of Nutrient H2O's followers, and the number of tweets per user that fall into 36 different categories. In order to find clusters that stand out as the most common, or the categories that are most commonly tweeted together, K-means is a good option.

This word cloud gives all possible tweet subjects, most of which being chatter, which, being such a generic term, is not surprising.

The top ten terms of each cluster are as follows:

```
##       1                2               3
##  [1,] "cooking"        "religion"      "chatter"
##  [2,] "fashion"        "parenting"     "shopping"
##  [3,] "college_uni"    "sports_fandom" "photo_sharing"
##  [4,] "online_gaming"  "food"          "current_events"
##  [5,] "beauty"         "school"        "tv_film"
##  [6,] "sports_playing" "family"        "uncategorized"
##  [7,] "music"          "crafts"        "art"
##  [8,] "photo_sharing"  "beauty"        "business"
##  [9,] "uncategorized"  "spam"          "small_business"
## [10,] "spam"           "eco"           "home_and_garden"
##       4                  5
##  [1,] "health_nutrition" "politics"
##  [2,] "personal_fitness" "news"
```

```
##  [3,] "outdoors"         "travel"
##  [4,] "cooking"          "automotive"
##  [5,] "food"             "computers"
##  [6,] "eco"              "sports_fandom"
##  [7,] "dating"           "outdoors"
##  [8,] "spam"             "business"
##  [9,] "adult"            "spam"
## [10,] "home_and_garden"  "dating"
```

We can see from these clusters, there is a specific market segment that would potentially tweet about the subjects in that cluster. A market segment, in this case, is a group with similar interests, and therefore similar tweets. There is some overlap among clusters, but they also have some distinct characteristics that line up with certain demographics of society.

Cluster 1 is likely college students looking for easy recipes, constantly sharing photos, and interested in latest trends in music and fashion. Cluster 2 can be parents, in general; parents are concerned with raising their kids the best they can, looking to religion for help. They are also concerned with food and school for their kids, and dads are normally big sports fans. Cluster 3 could be moms who stay up to date on latest fashions, current events, tv shows and movies. They constantly share photos of their kids, and advertise for their family business. Cluster 4 could represent health conscious, active people who enjoy the outdoors. Cluster 5 could be men that keep up with the latest cars and computers, as well as politics and news; they like to travel and keep up with their college football team after graduating.

Conclusion:

From these clusters, Nutrient H2O's followers seem to be of similar demographics, either new parents or young adults, who are interested in current trends, dating or parenting, and personal fitness.