STA 380 Homework 1

Liuxuan (Kelly) Yu and Lindsay Tober

August 8, 2016

STA 380 Homework 1

Probability practice

(A) What fraction of people who are truthful clickers answered yes to an online survey?

If people visiting a website are asked to answer a single survey question before they get access to the content on the page, what fraction of those who were truthful clickers actually answered yes? This question invokes the logic of Bayes Rule:

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

Applied to this question, we are looking to find: given that someone is a truthful clicker (A), what is the probability that they answered yes (B)? Based on the setup of only two categories, Random Clicker (RC) and Truthful Clicker (TC), and only two possible answers to the survey: yes and no, conditional probability is easily applied. Survey results find that 65% of respondent said Yes whereas 35% said No. Paired with the assumptions that (1) random clickers would click either Yes or No with equal probability and (2) the expected fraction of random clickers is 0.3:

$$P(YES|TC) = \frac{P(YES) * P(TC|YES)}{P(TC)}$$

The setup informs us that:

1. $P(TC) = 1 - P(RC) = 1 - 0.3 = 0.7$
2. $P(YES) = 0.65$
3. $P(YES|RC) = P(NO|RC) = 0.5$

To solve for $P(TC|YES)$, using the given information that $P(YES|RC) = 0.5$, and $P(YES|RC) = \frac{P(YES) * P(RC|YES)}{P(RC)}$:

$$P(RC|YES) = \frac{P(YES|RC) * P(RC)}{P(YES)} = \frac{0.5 * 0.5}{0.65} = 0.385$$

$$P(TC|YES) = 1 - P(RC|YES) = 1 - 0.385 = 0.615$$

Therefore, using the information for $P(YES)$, $P(TC|YES)$, and $P(TC)$:

$$P(YES|TC) = \frac{P(YES) * P(TC|YES)}{P(TC)} = \frac{0.65 * 0.615}{0.7} = 0.57$$

We can then conclude that 0.57, or 57% of of the truthful clickers (TC) actually answered yes.

(B) Supposing someone tests positive for a medical disease, what is the probability that they have the disease? In light of this, do you envision any problems in implementing a universal testing policy for the disease?

Given someone tests positive for a medical disease, what is the probability that they actually have the disease? A number of the test results may be false positives, or false negatives, so we must apply conditional probability via Bayes Rule.

The provided statistics on the medical tests show:

1. $P(+|D) = 0.993$
2. $P(-|ND) = 0.9999$
3. $P(D) = 0.000025$

where + / - represent positive / negative test results and D / ND represent Disease / No Disease.

In order to calculate the likelihood of disease given a postivie test result, or $P(D|+)$, we can apply Bayes Rule again:

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

With the provided statistics on medical tests, we can calculate the percentage of positive test results using the Law of Total Probability:

$$P(+) = P(+|D) * P(D) \ + \ P(+|ND) * P(ND) = 0.993 * 0.000025 \ + \ (1 - 0.9999) * (1 - 0.000025) = 0.000125$$

With the final component for applying Bayes Rule, we can calculate:

$$P(D|+) = \frac{P(D) \ * \ P(+|D)}{P(+)} = \frac{0.000025 \ * \ 0.993}{0.000125} = 0.199$$

This result means that if someone tests positive for the given medical disease, the probability that they have the disease is 19.9%. In light of this, any application of a universal testing policy may create problems given only approximately 1 in 5 people who test positive actually have the disease. While 19.9% is significant in the fact that it improves the likelihood from 0.0025%, it is still very low and would likely generate unneccesary concern for the other 80.1% of people who test positive but do not have the disease. A more accurate test, if available, would be a better option to predict the likelihod of actually having the disease given a positive result.

Exploratory analysis: green buildings

An Austin real-estate developer is interested in the possible economic impact of “going green” in her latest project: a new 15-story mixed-use building on East Cesar Chavez, just across I-35 from downtown, where baseline construction costs are estimated at $100 million with a 5% expected premium for green certification (equivalent to an additional $5 million). She is looking to understand if moving forward with the green building will be a wise investement decision from an economic perspective.

The real-estate developer had a staff member with Excel experience perfom an initial cost / benefit analysis using historical data on other green buildings, where he concluded that the new project would be a good investment decision requiring 7 to 9 years to break even. However, his analysis had some concerns:

- He excluded a handful of buildings in the historical data set with very low occupancy rates (less than 10% of available space occupied) that he believed could potentially distort the analysis, based on an untested theory that these buildings might have low occupancy rates due to something abnormal with the buildings themselves. Removing these low occupancy data points may have given an invalid boost to the expected occupancy rate of the new building.
- He looked at the green buildings and non-green buildings separately, comparing median market rent in the non-green buildings ($25/sq.ft. per year) with median market rent in the green buildings ($27.60/sq.ft. per year) to reach a difference of about $2.60/sq.ft. per year for green buildings. He used the median rather than the mean due to outliers in the dataset, as the median is a lot more robust to outliers. While the use of the median was not incredibly concerning, the calculation of difference in $2.60/sq.ft. per year (which he then applied to the square footage of the new building to understand incremental revenue) assumes a linear relationship between rent and building size that does not necessarily exist. In addition, it does not account for other confounding variables that may influence building rent.
- He assumed an occupancy rate of 100% for the initial break even calculation, which he did not tie to actual occupancy rates, and his caveat for a flat occupance rate of 90% is still unsubstantiated by the actual occupancy rates in the data.

Given the inherent gaps in his analysis the model likely has accuracy issues, and the real-estate developer was right to look for an additional opinion. We do not agree with the conclusions of the on-staff analyst and would like to perform additional analysis to better inform the real-estate developer in making her investment decision.

First, we see that the data has a small set of ‘green’ buildings, as only 685 out of 7894 buildings in the data set were rated green. Of these green buildings, the split between Energystar and LEED is significant. Only 7 of the 685 green buildings are both Energystar and LEED ceritified, with an additional 47 being only LEED certified and 631 being only Energystar certified. Given the small proportion of the data set however, we will include all ‘green’ rated buildings in the analysis.

Distribution of rent/sq.ft. per year varies minimally between buildings with and without green ratings, leading to the $2.60/sq.ft. per year difference that the analyst determined. Particularly with outliers excluded, there is a very slight difference between rent for the two ratings:

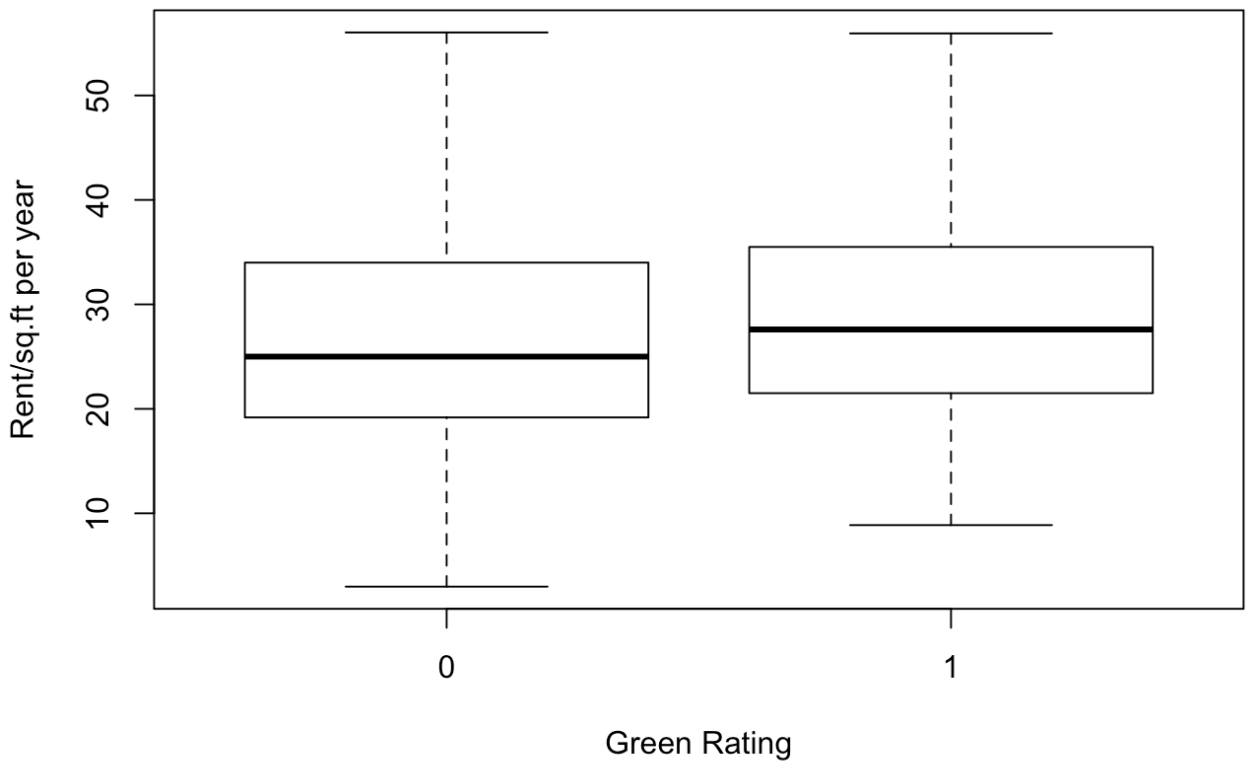


Figure 1: Boxplot of the rent per square foot per calendar year for buildings with (1) or without (0) a green rating.

While the boxplot shows some minor differences in rent between the two groups of buildings, the correlation is quite low between rent and green rating regardless of the correlation methodology used. Direct correlation of 0.0326659 was only slightly different when using Spearman's rho (0.0620797) and Kendall's tau (0.0508199). This suggests that the green status might not have strong relationship with rent price.

Table 1: Summary statistics for leasing rates of buildings with and without a green rating.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Green Buildings	0	85.4	92.92	89.28	97.7	100
Non-Green Buildings	0	77.05	89.17	81.97	96.28	100

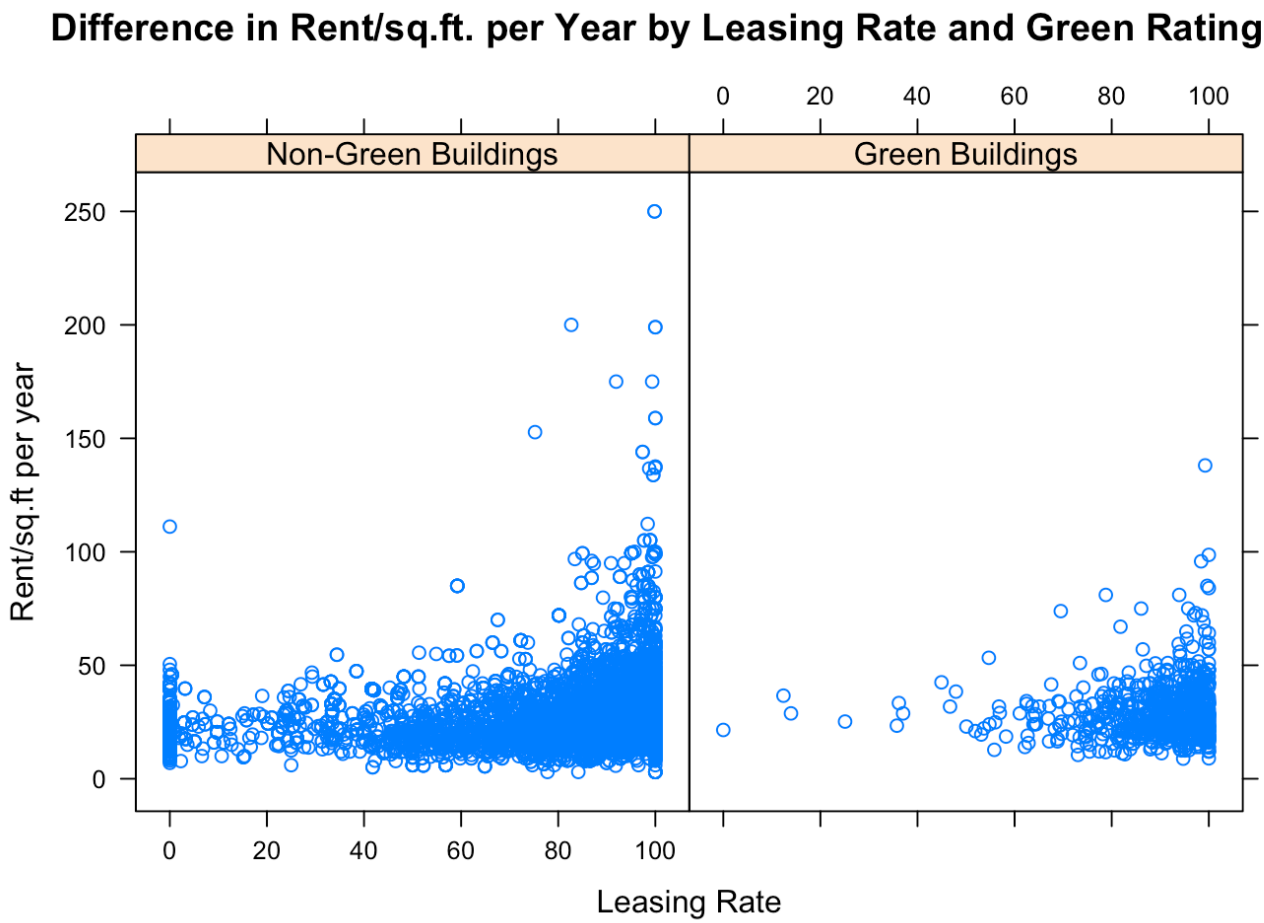


Figure 2: Lattice plot of the rent per square foot per calendar year by leasing rate (occupancy rate), for buildings with or without a green rating.

Considering the leasing outlier that the model removes, there are about 215 buildings out of the total 7,894 buildings that had less than 10% of available space occupied. From the lattice plot we can see that it is significant that in buildings with green ratings, there are fewer low leasing rate buildings. This may be an indicator that the leasing rate is a good index to judge whether a building is green or not given that all of the buildings with lower rates are not green-rated. Therefore, it is very inaccurate to just delete them all from the model.

From Summary of the leasing rate, we find that the green house leasing rate has 75% higher than 85.4, median is 92.9,however the non green one's median is only 89 and 75% higher than 77.05, so there is significantly difference about the leasing rate.

A linear regression of the rent and green rating shows that for buildings which have a green rating, the rent would be 1.75x higher than the non-green buildings. The confidence intervals show that prices for green buildings are distinguishable from price of non-green buildings. However, the R^2 value is very low in this model, nearly 0.001. Therefore, only using green rating in the linear model does not sufficiently explain the price difference between green and non-green buildings.

By using stepwise function of various attributes in the green building data set, we find that additional variables that are important when fitting a linear regression model of the real price of house on the variables that the stepwise function chooses.

Using regFoward Function to complete forward regression on real price with multiple variables, the green rating variable was not included as an indicator. Therefore, it is difficult to establish a clear relationship about whether green rating will influnece the whole model. If we do include it in to replace the LEED, it only shows about 0.6623 premium for green

buildings per square foot with a p_value that is not statistically significant, so we might use the green building rate as a good predictor of the building price.

Another potential improvement for further analysis would be to predict using a `net contract' basis where the renter is charged for additional utility costs (e..g, gas costs and electricity costs). This could help the real-estate agent to improve profitiability of her project.

Bootstrapping

(1) Characterize the risk/return properties of the five major asset classes

Supposing there is a notional $100,000 to invest in a type of portfolios, we would like to characterize the risk/return properties of the five major asset classes to inform the investment decision. The asset classes under consideration include:

- 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)

We will take SPY to just represent the market risk and return. Comparing the coefficient of the stock with the market coeficient 1, we could find that the TLT LQD are very safe and with the negative relationship with market,because their beta are very low, only -0.3 and -0.07, so even the market increases or decreases significantly, they likely won't change much. However, EEM and VQN are more risky because their coefficients run higher, to about 0.76 and 0.64 in a positive relationship with the market.

(2) Outline your choice of the "safe" and "aggressive" portfolios

For the safe portfolio, we chose to use the combination of the three safe ETF(TLT,LQD,EEM) and choose the percentile as (0.6,0.1,0.3) because their coefficients are in the opposite direction so they can hedge the risk,so the new coefficients of the safe portfolio would be 0.4*coef(lm_TLT)+0.4*coef(lm_LQD)+0.2*coef(lm_VNQ)=0.6*-0.38+0.1-*0.03+0.3*0.74=-0.015, which means almost no risk. For all ETF, the coefficient would be (0,0.6,0.1,0, 0.3)

For a risky portfolio, we use EEM, VNQ and SPY itself to construct a portfolio with a high positive correlation with the market, without any negative-market-related ETF in this portfolio. We used the weight of as(0.5,0.3,0.2), so for all ETf the weight would be (0.2,0,0,0.5,0.3).

(3) Use bootstrap resampling to estimate the 4-week (20 trading day) value at risk of each of your three portfolios at the 5% level.

Even Split Portfolio

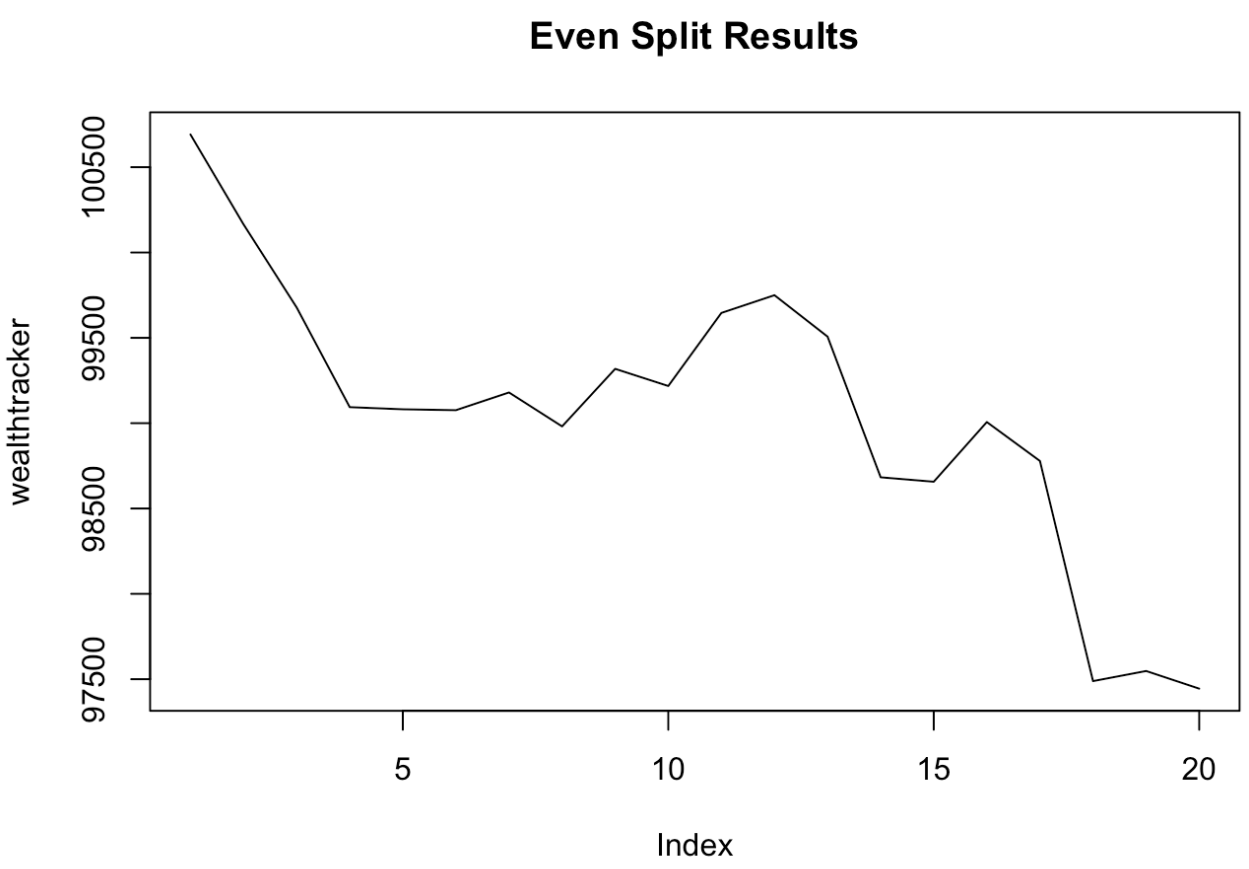


Figure 3: Line plot of total wealth by day for even split portfolio.

For the the even split portfolio, the estimate 4-week value at risk at the 5% level is about -3295.801.(The first 20 days is 98175.47).

Safe Portfolio

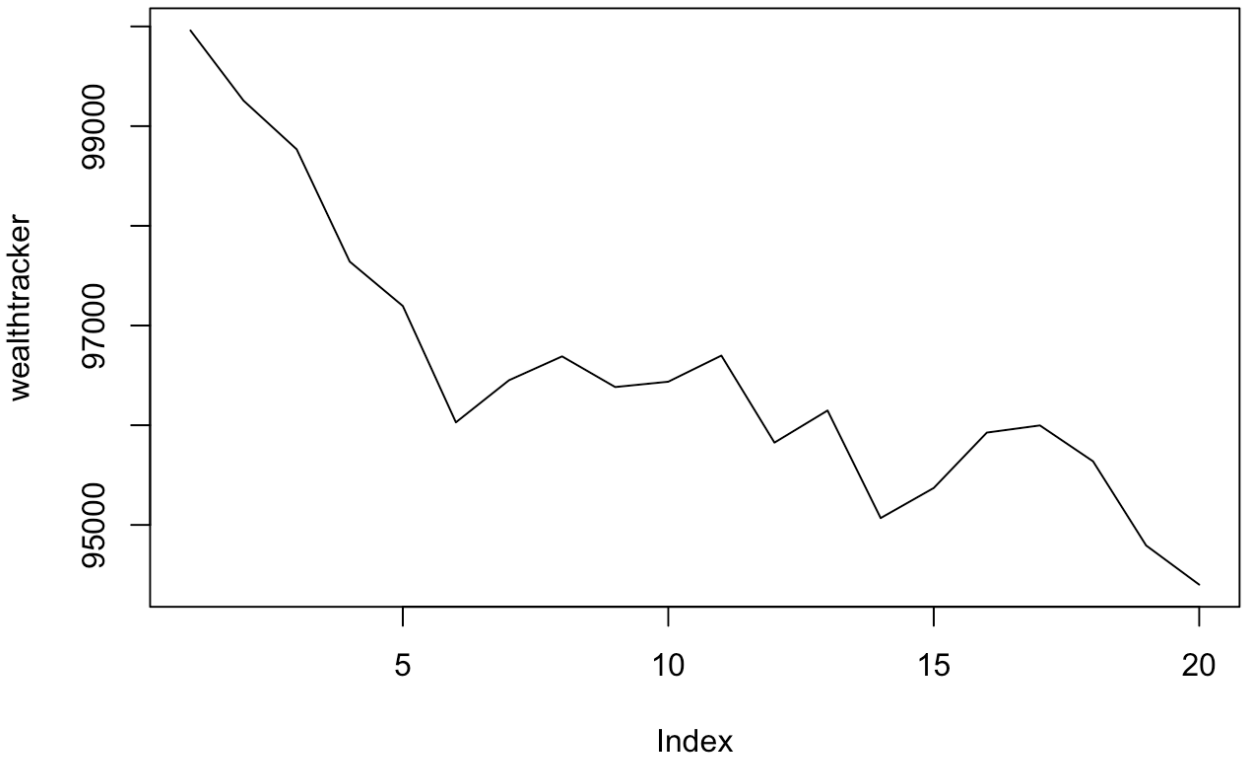


Figure 4: Line plot of total wealth by day for safe portfolio.

For the safe portfolio, we used the combination of the three safe ETF(TLT,LQD,EEM) and chose the percentile as (0.6,0.1,0.3), because their coeficient are in the opposite direction so they can hedge the risk. The new coefficients of the safe portfolio would be 0.4*coef(lm_TLT)*+*0.4*coef(lm_LQD)+0.2*coef(lm_VNQ)=0.6*-0.38+0.1-*0.03*+*0.3*0.74=-0.015, which means almost no risk. The estimated 4-week value at risk of the safe potfolio at the 5% level is about -2746.392. (The first 20 days is 101970.9).

Risky Portfolio

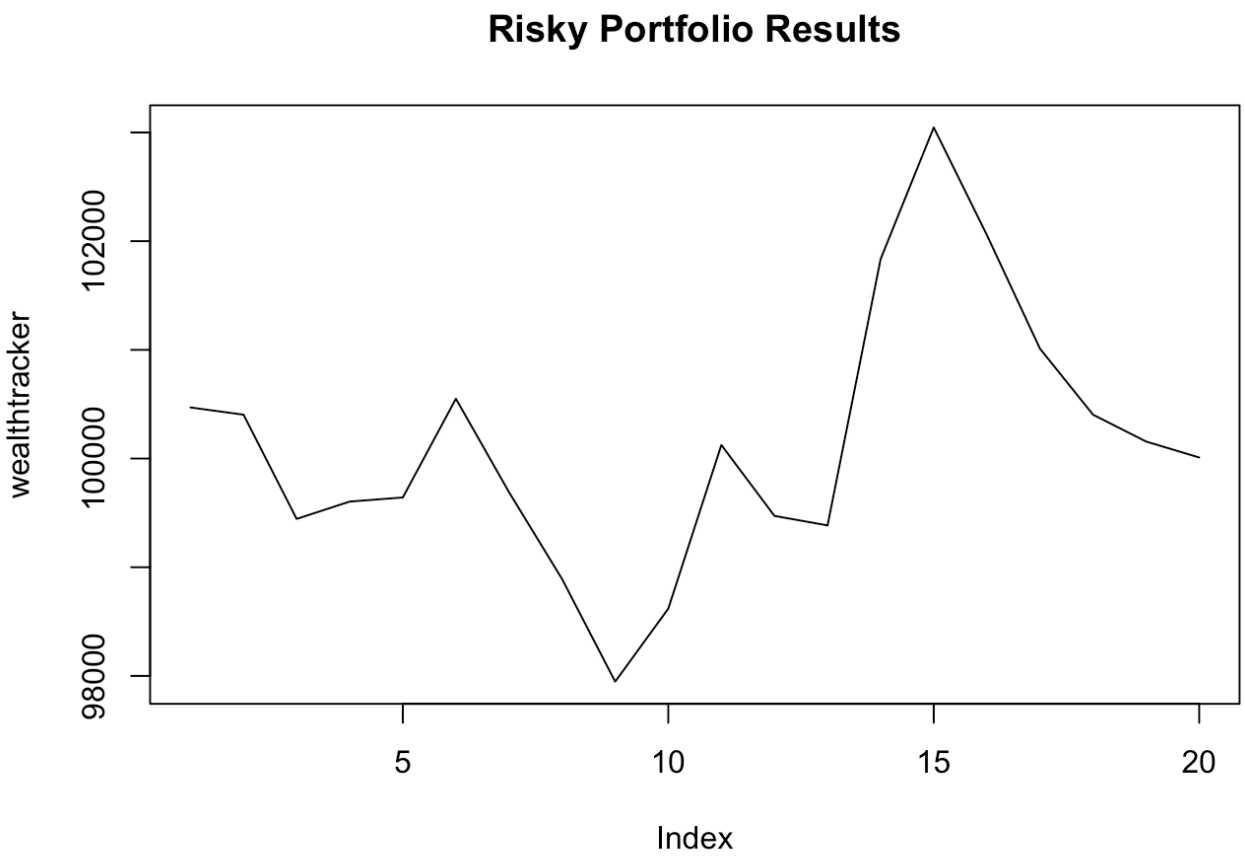


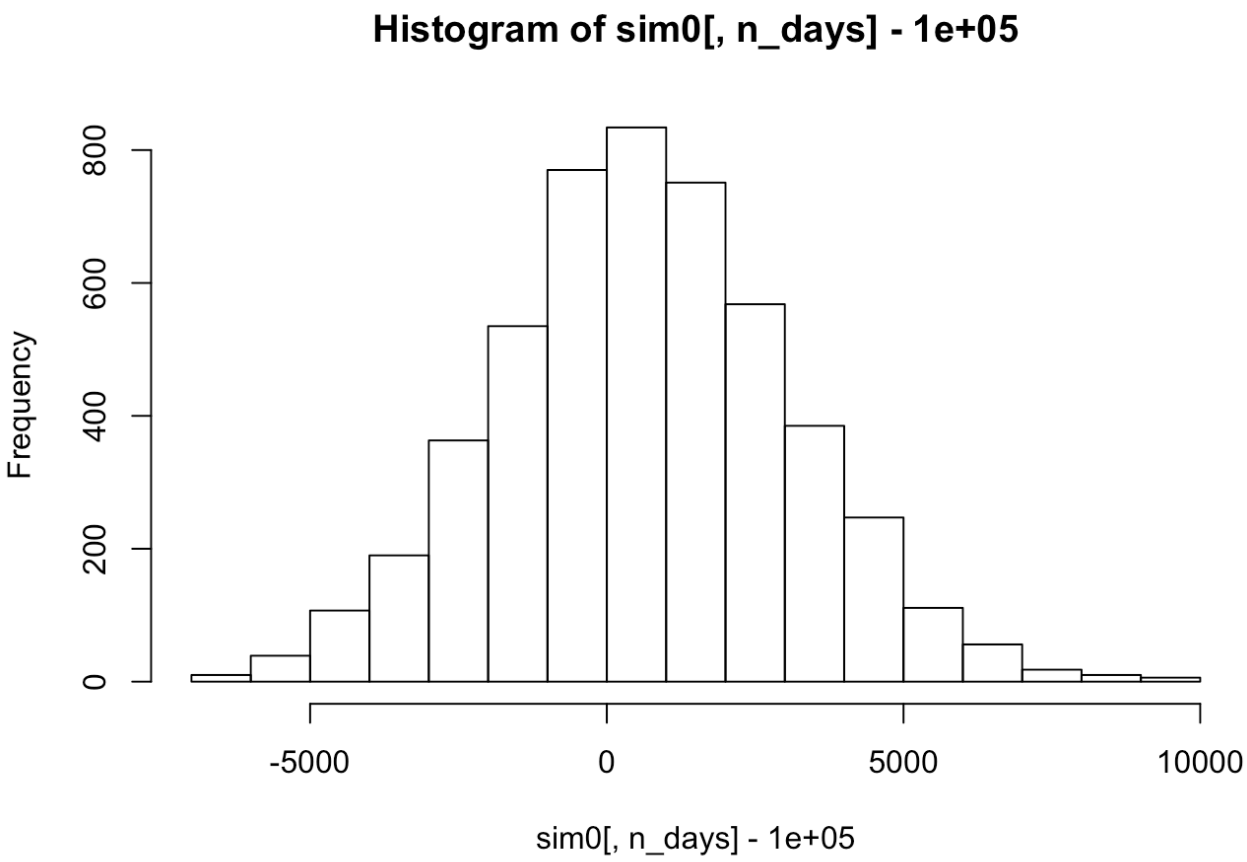
Figure 5: Line plot of total wealth by day for risky portfolio.

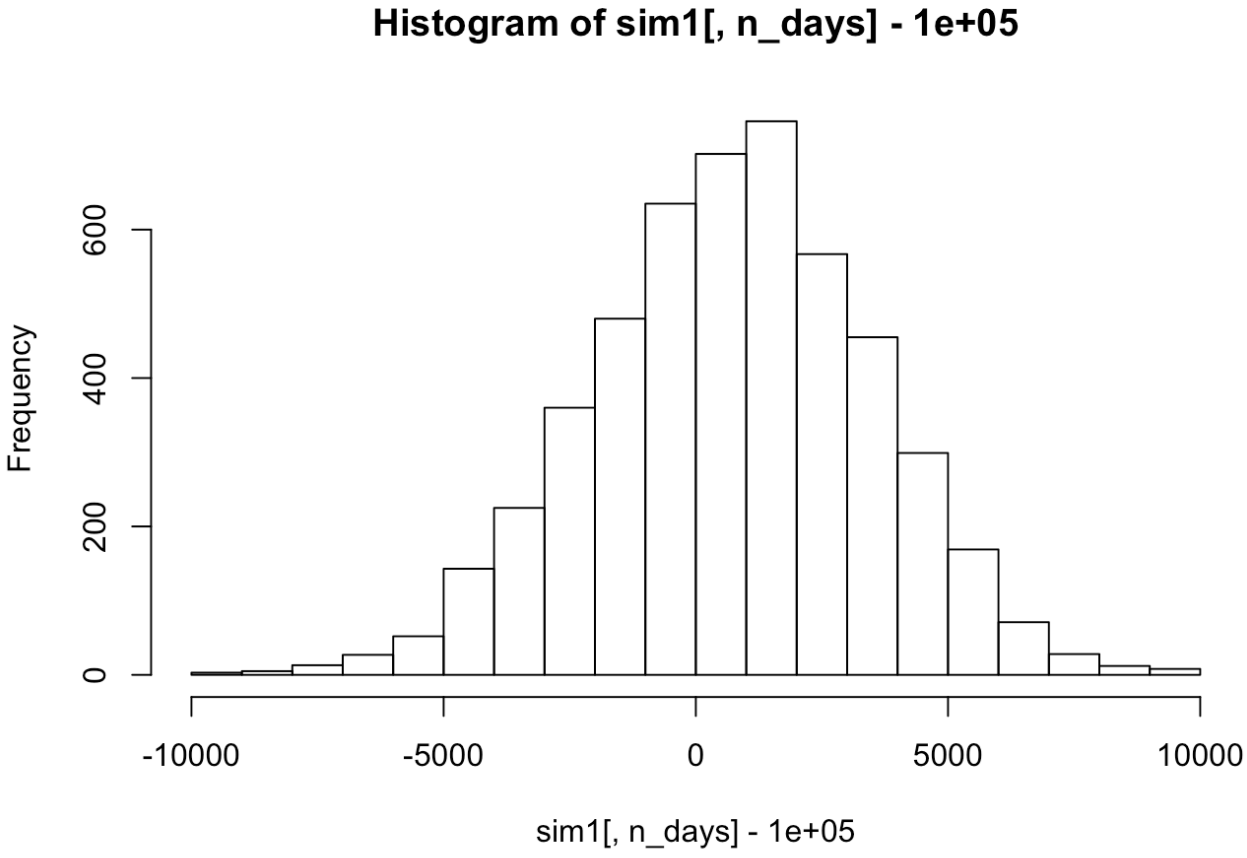
The estimated 4-week value at risk of the risky potfolio at the 5% level is about -6501.389, which is a big loss (The wealth after first 20 days is 93590.03).

(4) Compare the results for each portfolio

Based on previous modeling, the following results were obtained on portfolio variance:

- For evenly split:VaR is -3295.801
- For safe portfolio:VaR is -2746.392
- For risky portfolio:VaR is -6501.389


Histogram of sim0[, n_days] - 1e+05
Frequency
0
200
400
600
800
-5000
0
5000
10000
sim0[, n_days] - 1e+05


Histogram of sim1[, n_days] - 1e+05
Frequency
0
200
400
600
-10000
-5000
0
5000
10000
sim1[, n_days] - 1e+05

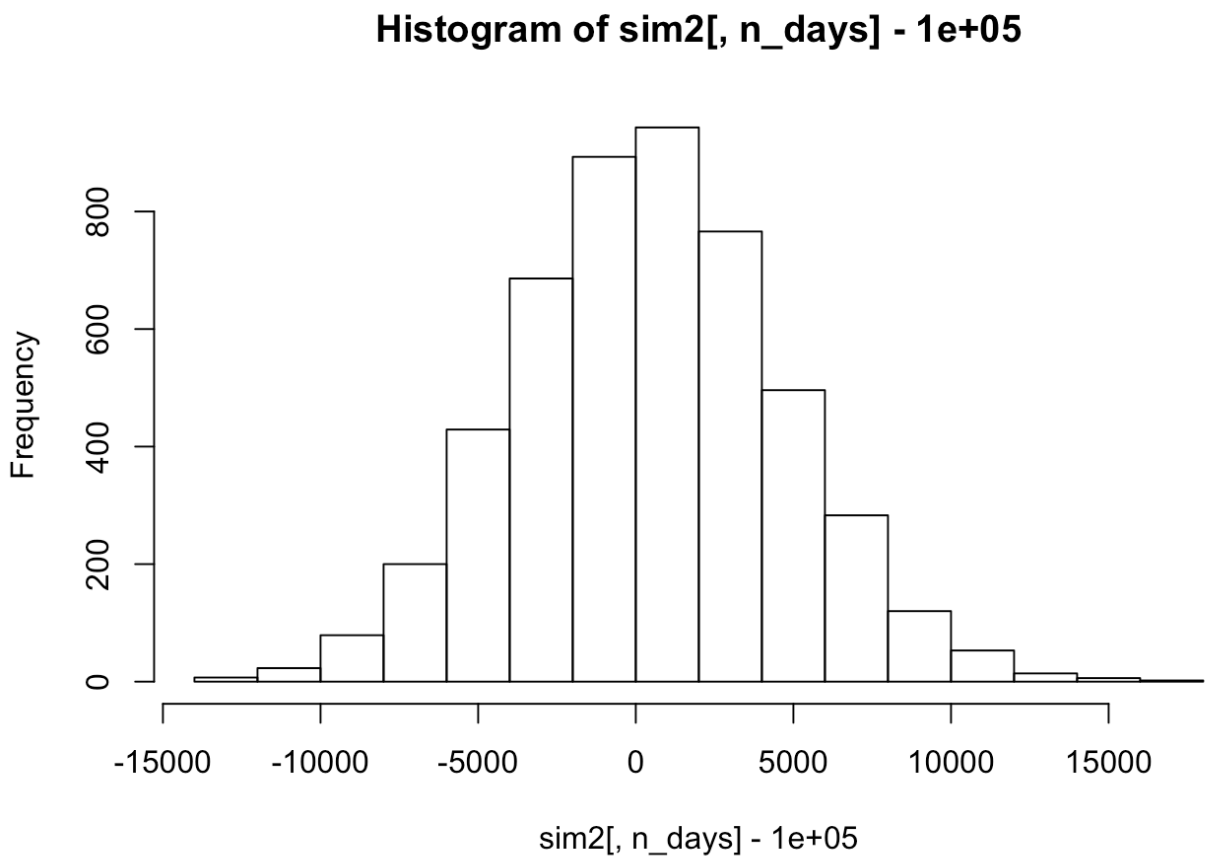


Figure 6: Histograms showing distribution of loss and returns.

This summary supports choosing the safe portfolio because the loss in this portfolio is less than other two. However, that might be because the datapoints we chose are more likely to have bad returns and economic depression so that it doesn't have good return on market. Regardless, the safe portfolio results in less risk.

Market segmentation

A market-research study was completed based on tweets of followers of NutrientH20 Twitter account, where they had an advertising firm analyze tweets of its followers over a seven day period in June 2014 so as to to explain its social-media audence in order to better hone its audience. Initial categorization was completed on the tweets to assign them different labels. For example, a hypothetical post such as "I'm really excited to see grandpa go wreck shop in his geriatic soccer league this Sunday!" might be categorized as both "family" and "sports."

Using K-means clustering, we segmented NutrientH20's Twitter followers into 5 clusters and gathered the top categories for each cluster:

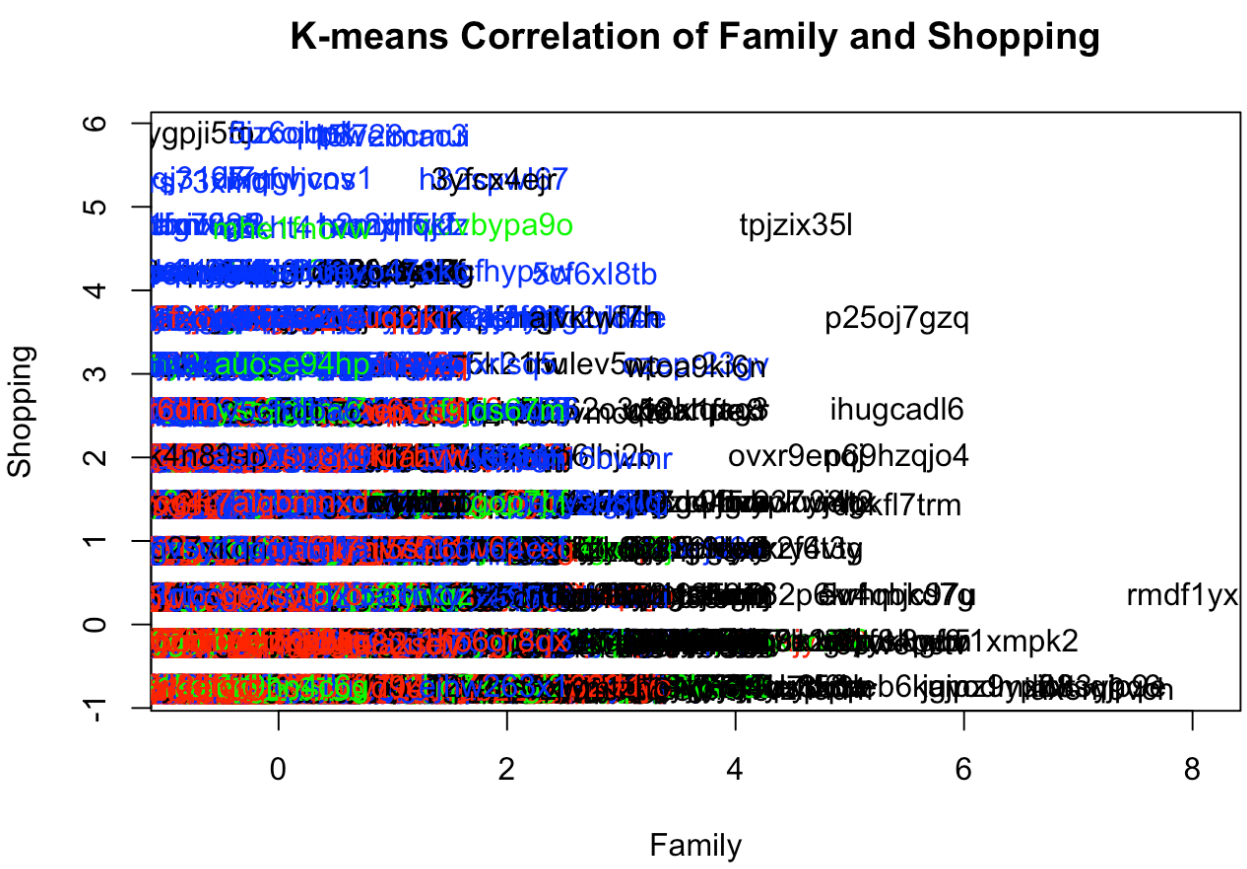


Figure 7: Plot of K-means clusters for K=5 by twitter follower indicator, against Family and Shopping.

From the K-means summary, we can see five defined groups:

- Interested in dating, fashion, beauty
- Interested in online gaming and playing sports
- Interested in artistic areas (e.g., film, art, music)
- Interested in culture (e.g., photo sharing, shopping, news)
- Interested in health nutrition and personal fitness

NutrientH20 should consider delivering targeted ads to these different groups of people to improve sales. In addition, NutrientH20 should review their ad campaigns to ensure that all of these different clusters are addressed with the scope of their campaigns.