STA 380, Part 2: Exercises 1

Marissa Hausman

August 7, 2016

I. Probability Practice

Part A

1. Overview

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?

2. Data and Model

Variables:

- TC = Truthful Clicker
- RC = Random Clicker
- Y = answered 'yes'

Equation:

- P(Y) = P(Y|TC) x P(TC) + P(Y|RC) x P(RC)

Values:

- P(Y) = .65
- P(TC) = .7
- P(Y|RC) = .5
- P(RC) = .3

3. Results

P(Y|TC) = (.65-(.5*.3))/.7 = .7142857

4. Conclusion

71.43% of people who are truthful clickers answered yes.

Part B

1. Overview

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?

2. Data and Model

Variables:

- D = has the disease
- TP = test positive

Equations:

- P(TP) = (P(TP|D) x P(D)) + (P(TP|not D) x P(not D))
- P(D|TP) = P(D) x P(TP|D)/P(TP)

Values:

- P(TP|D) = .993
- P(not TP|not D) = .9999
- P(D) = .000025

3. Results

P(TP) = .993 x .000025 + .0001 x .9975 = .000124575

P(D|TP) = .000025 x .993 / .000124575 = .1992775

4. Conclusion

The probability that a person has the disease given that he/she tests positive is only 19.93%. With such a large proportion of false positives, I would anticipate serious difficulty in implementing a universal testing policy for the disease.

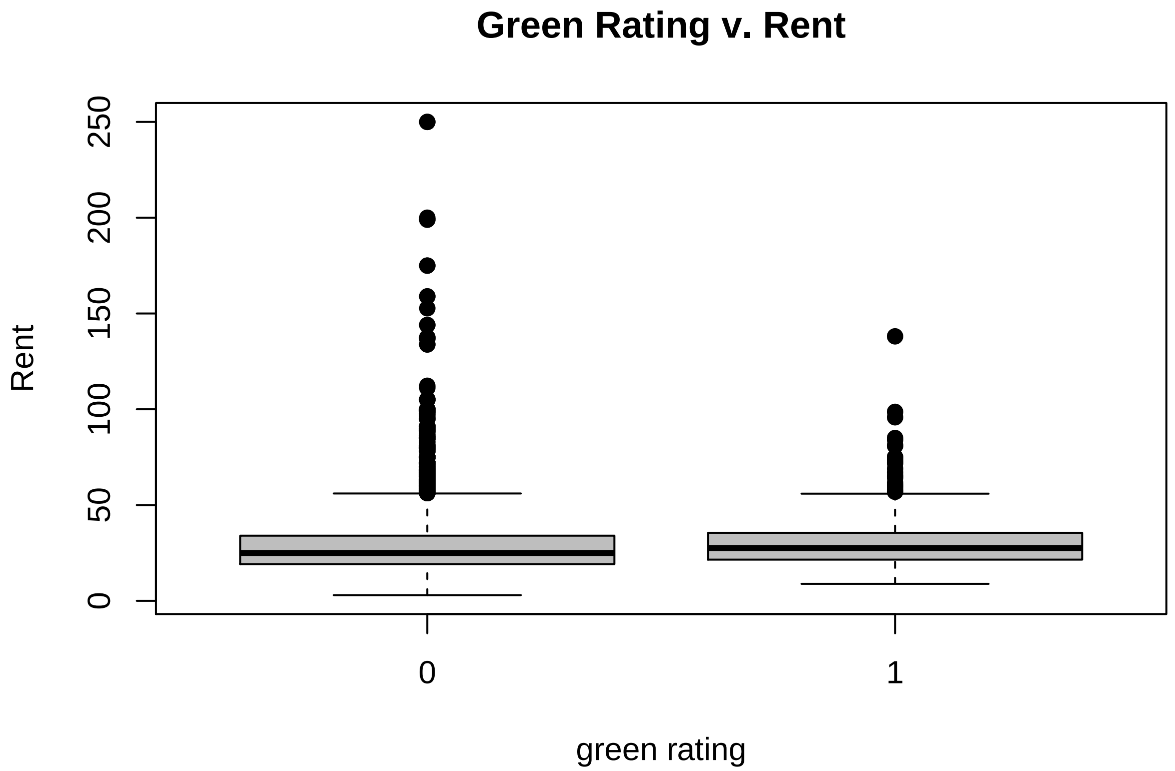
II. Exploratory Analysis: Green Buildings

1. Overview

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. Will investing in a green building be worth it, from an economic perspective? A stats guru assessed the problem and concluded, "It seems like a good financial move to build the green building." Do you agree with the conclusions of her on-staff stats guru? If so, point to evidence supporting his case. If not, explain specifically where and why the analysis goes wrong, and how it can be improved.

2. Data and Model

In plotting green rating vs. Rent, we find that the distributions of Rents for green rated buildings and non-green rated buildings are similar. Immediately the stats guru's conclusion is called into question.



The guru excluded low-occupancy buildings from the dataset. How would his results have changed if these buildings were not excluded? The model created here will not exclude buildings with low occupancy.

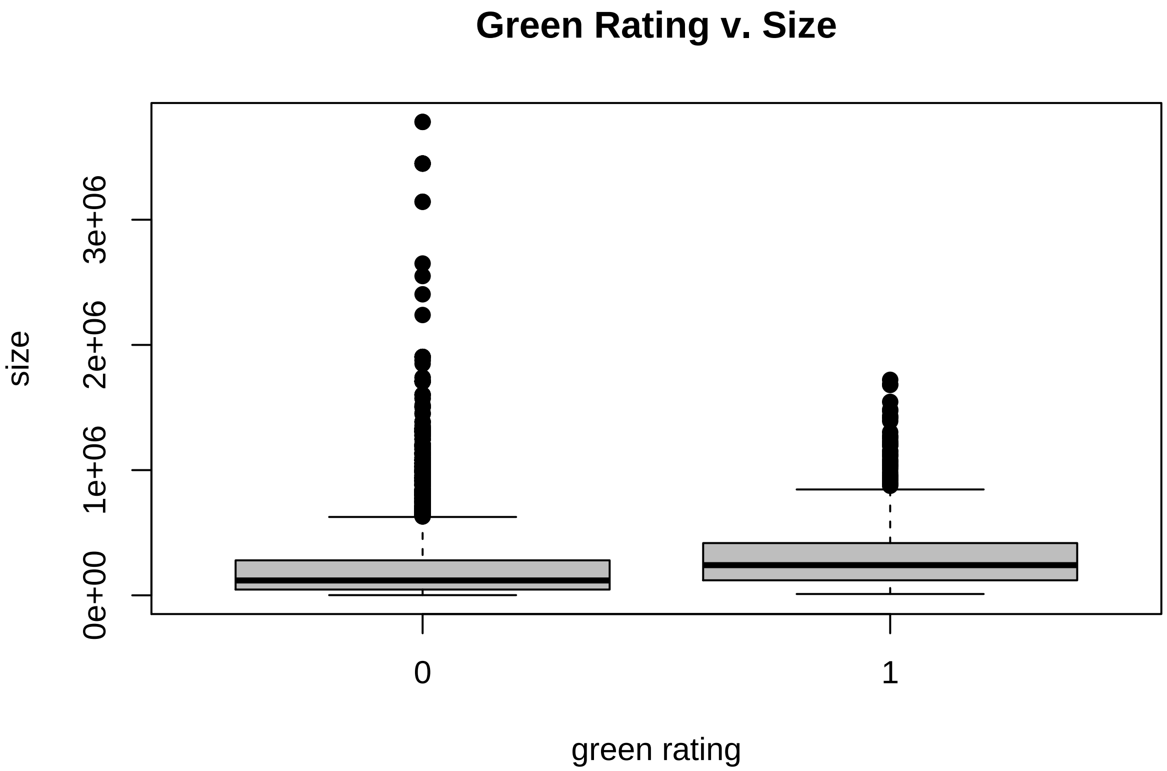
Most importantly, the model created here will explore the possibility of confounding variables, i.e. variables that could explain the relationship between green rating and Rent aside from green rating itself.

3. Results

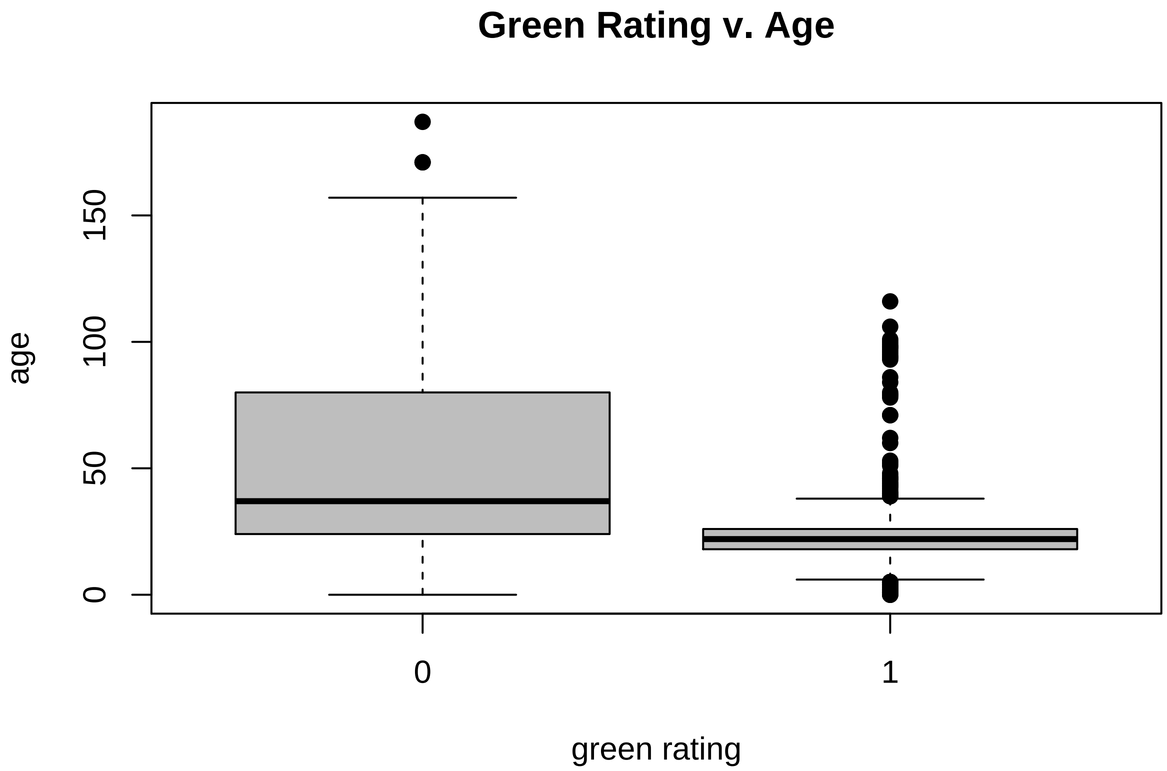
The Austin real estate developer stated, "The median market rent in the non-green buildings was $25 per square foot per year, while the median market rent in the green buildings was $27.60 per square foot per year: about $2.60 more per square foot" after removing buildings in the dataset with very low occupancy rates. Without removing buildings with low occupancy rates, we find that the median Rents remain $25 and $27.60 for non-green and green buildings, respectively, which is consistent with the developer's findings. With this in mind, there is a need to examine potential confounding variables.

Confounding variables:

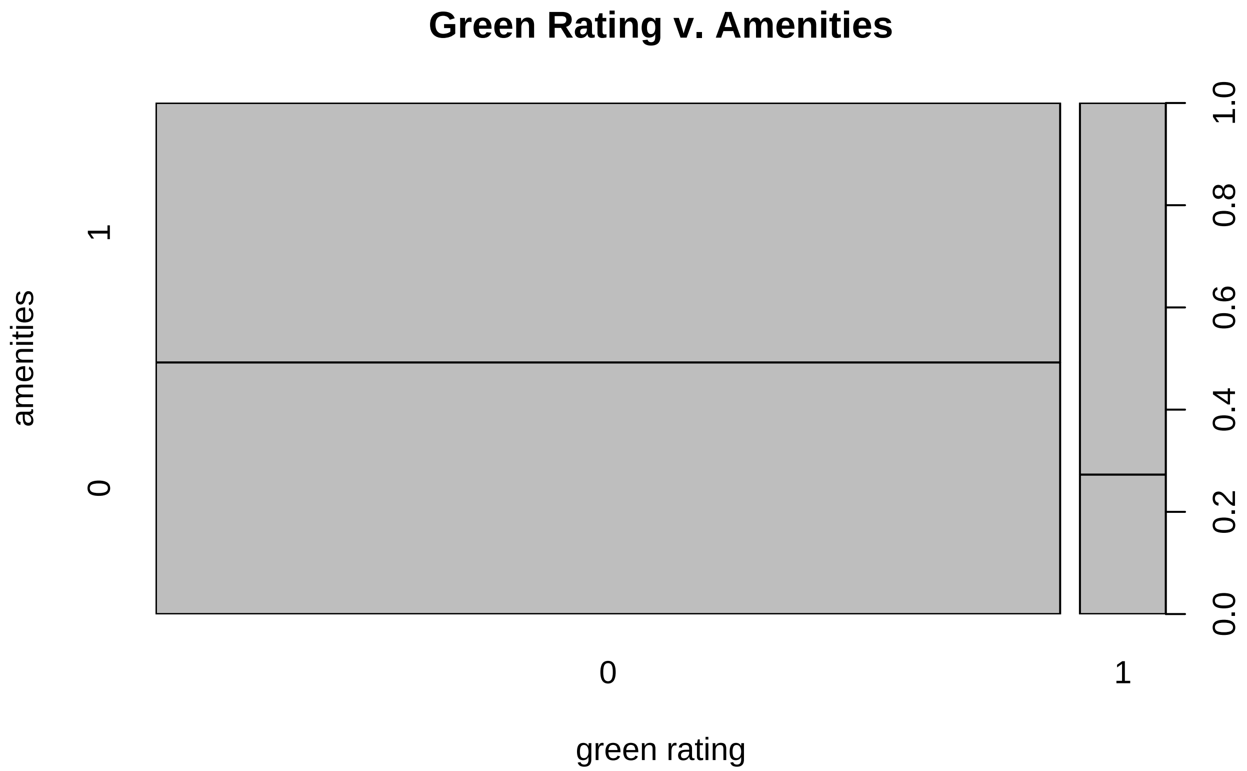
Green-rated buildings tend to be larger than non-green rated buildings.



Green-rated buildings tend to be significantly newer than non-green rated buildings.



Green-rated buildings very often have more amenities than non-green rated buildings.



4. Conclusion

The conclusions of the stats guru are not well-grounded. His analysis does not consider the effect of confounding variables on the relationship between whether a building is green rated and the Rent of the building. The size, age, and amenities of green buildings all point to higher Rents for green rated buildings, meaning green buildings may not be more expensive simply because they are green rated but because they have certain better qualities than non-green rated buildings. The developer could improve his analysis by controlling for these confounding variables and finding the difference in Rents between non-green and green rated buildings all else equal.

III. Bootstrapping

1. Overview

Consider the following five asset classes, together with the ticker symbol for an exchange-traded fund that represents each class:

- 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)

Consider three portfolios:

- the even split: 20% of your assets in each of the five ETFs above.
- something that seems safer than the even split

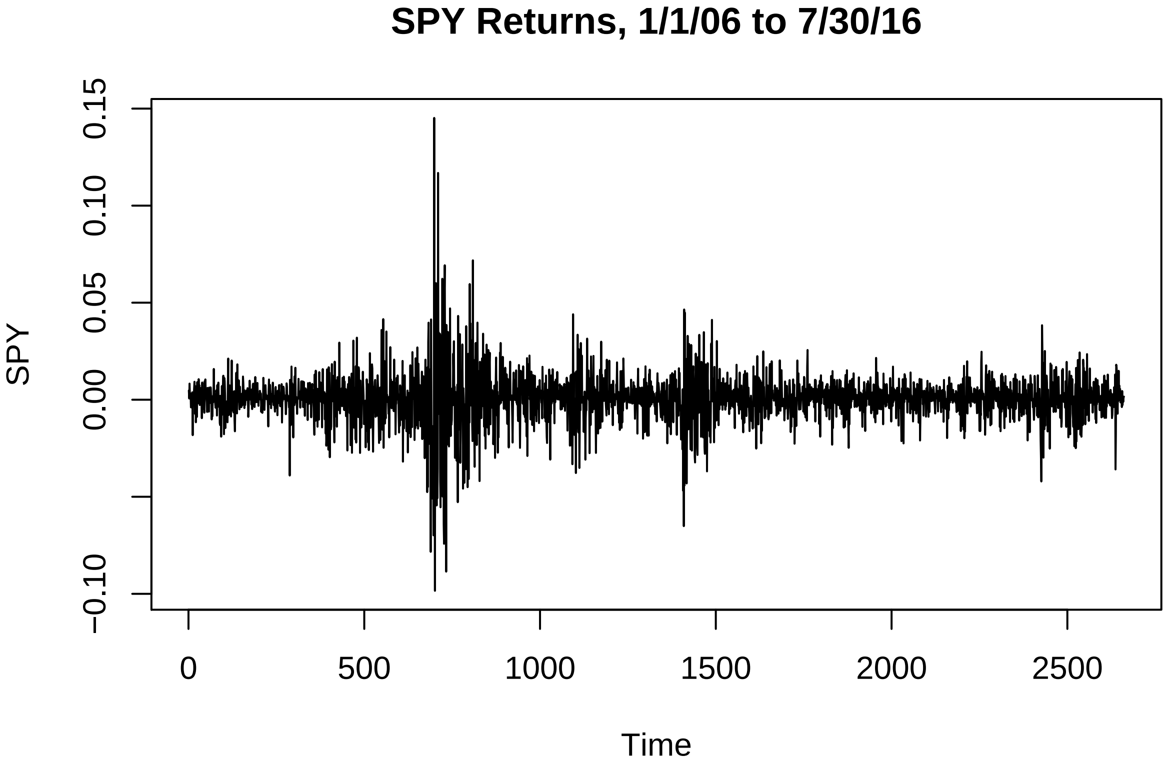
- something more aggressive

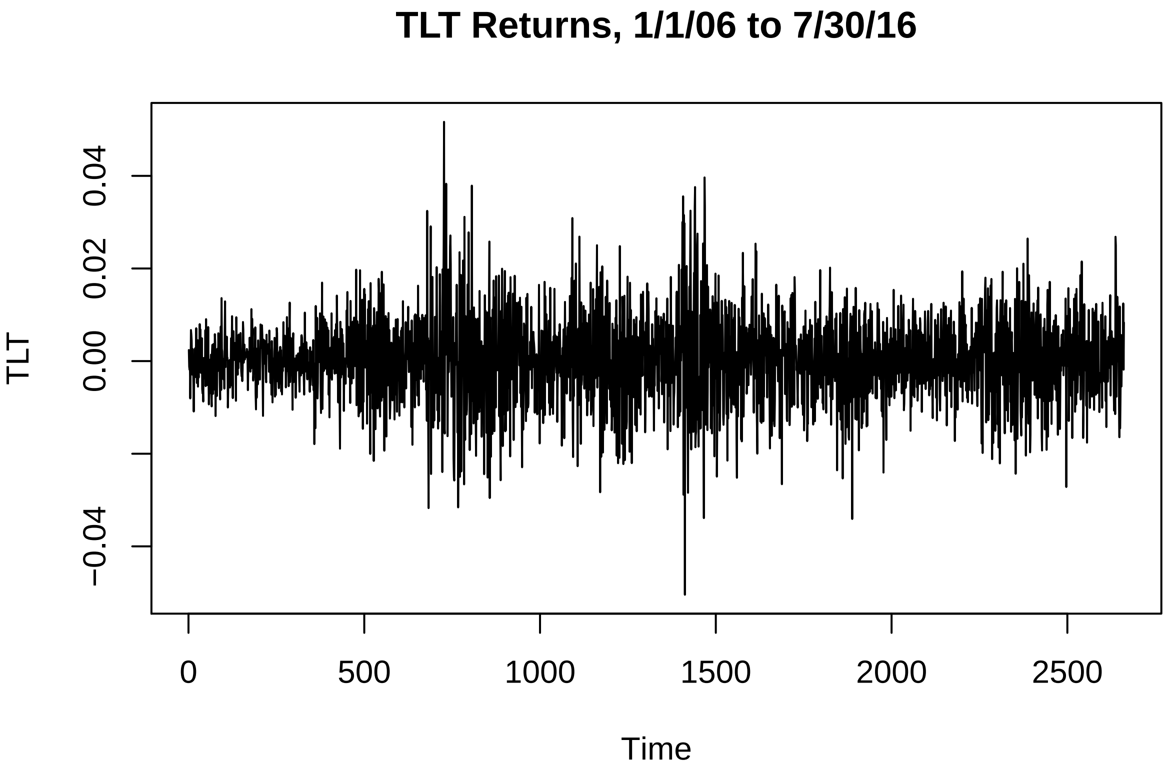
Write a brief report that:

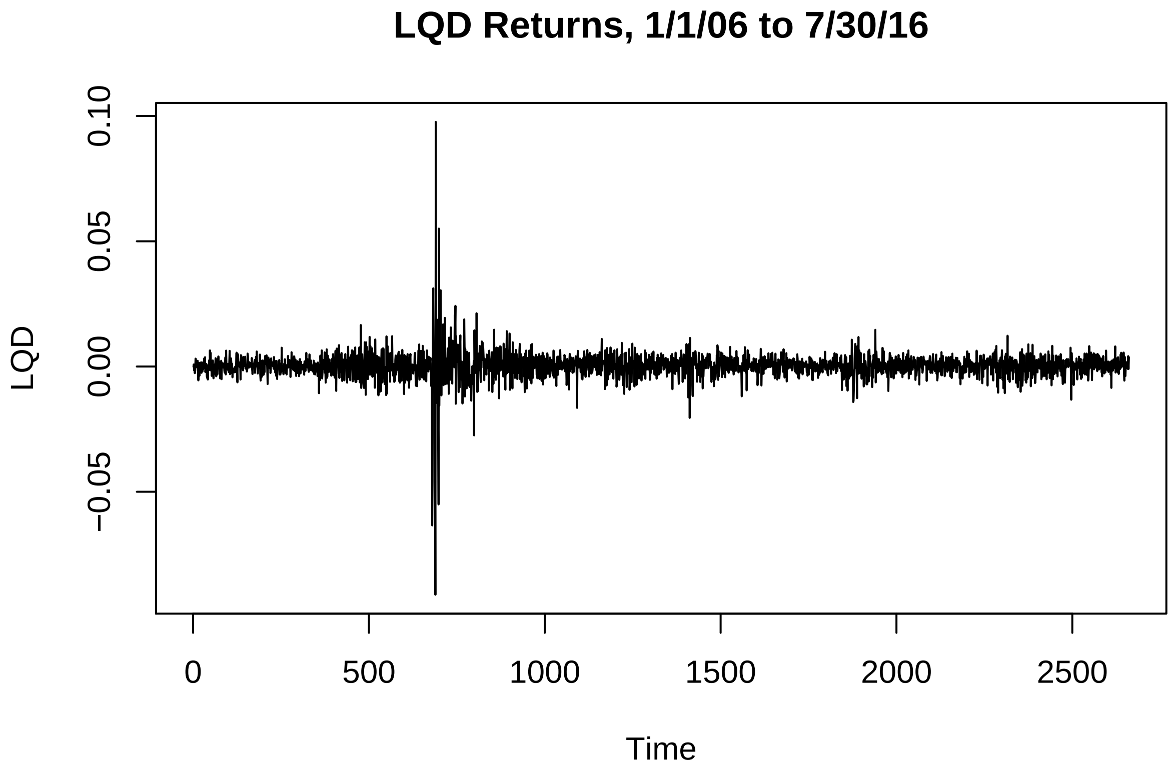
- marshals appropriate evidence to characterize the risk/return properties of the five major asset classes listed above.
- outlines your choice of the “safe” and “aggressive” portfolios.
- uses bootstrap resampling to estimate the 4-week (20 trading day) value at risk of each of your three portfolios at the 5% level.
- compares the results for each portfolio in a way that would allow the reader to make an intelligent decision among the three options.

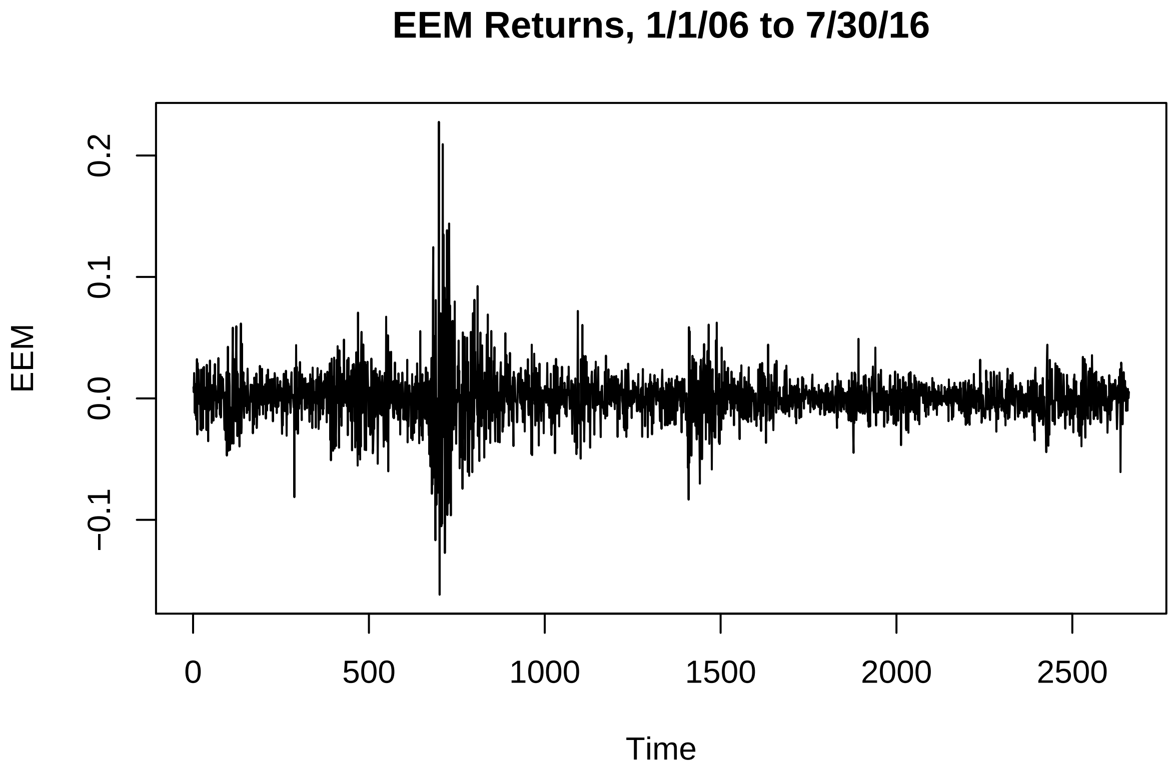
2. Data and Model

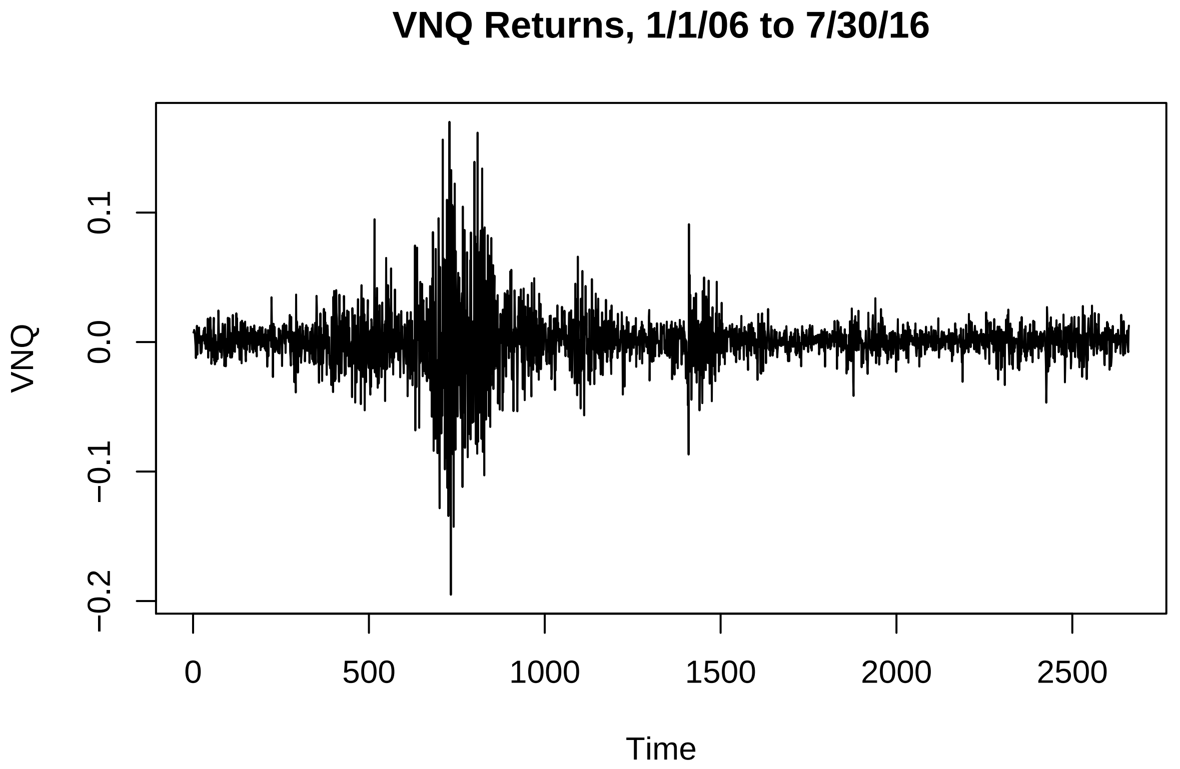
The following plots show the returns over time for each asset.




TLT Returns, 1/1/06 to 7/30/16
TLT
0.04
0.02
0.00
−0.04
0
500
1000
1500
2000
2500
Time


LQD Returns, 1/1/06 to 7/30/16
LQD
0.10
0.05
0.00
−0.05
0
500
1000
1500
2000
2500
Time


EEM Returns, 1/1/06 to 7/30/16
EEM
0.2
0.1
0.0
−0.1
0
500
1000
1500
2000
2500
Time



The risk/return profile of each asset will be determined using the mean and standard deviation of returns on that asset over the past 10 years, as well as the asset's Beta. The mean and standard deviation will be used to calculate the confidence interval of the asset's returns. Assets with low Betas and confidence intervals that hover around 0 are considered "safe" assets and will comprise the safe portfolio. Assets with high Betas and confidence intervals that stray from 0 are considered "aggressive" assets and will comprise the aggressive portfolio.

3. Results

The following table summarizes the risk/return characteristics of each asset, as well as the risk rating assigned to each asset based on those characteristics:

Asset	Mean	Standard Deviation	Confidence Interval	Beta	Risk Rating
SPY	.000366	.0129	-.0125 to .0133	1.000	3
TLT	.000346	.0093	-.0090 to .0096	-.320	2
LQD	.000242	.0054	-.0052 to .0056	.045	1
EEM	.000358	.0211	-.0207 to .0215	1.437	4
VNQ	.000573	.0220	-.0214 to .0226	1.333	5

List of safest to riskiest investments:

- 1. LQD. Lowest Beta, mean, and standard deviation. Low risk, low return.
- 2. TLT. Second lowest Beta, mean, and standard deviation.
- 3. SPY. Benchmark for returns.

- 4. VNQ. Similar confidence interval to EEM, but lower Beta.
- 5. EEM. Highest Beta.

Choice of the "safe" and "aggressive" portfolios:

Safe portfolio: LQD is significantly safer than the other investment options, so the "safe" portfolio will consist of 60% LQD. The remainder of the safe portfolio will consist of 30% TLT and 10% SPY. The potential return on this portfolio is not particularly high, but the risk of loss is also low.

Aggressive portfolio: EEM and VNQ are the riskiest assets with great potential for returns in comparison to the other investment options, so the "aggressive" portfolio will consist of 50% EEM and 50% VNQ. The potential return on this portfolio is high, as is the risk of loss.

5% values at risk:

- Even split portoflio: -6498.119
- Safe portfolio: -3228.854
- Aggressive portfolio: -12726.53

4. Conclusion

There is a 5% chance of losing $6,498.12 over the course of 4 weeks, given normal market conditions, with the even split portfolio. The same goes for the safe and aggressive portfolios, with chances of losing $3,228.85 and $12,726.53, respectively, over the course of 4 weeks. An investor can decide among the three portfolio options depending on his/her risk preferences. If he/she is confident in upcoming market conditions, he/she may be willing to take on the aggressive portfolio. If he/she is wary of upcoming market conditions, he/she may choose the safe portfolio.

IV. Market Segmentation

1. Overview

Consider the data in social_marketing.csv. This was data collected in the course of a market-research study using followers of the Twitter account of a large consumer brand that shall remain nameless—let's call it "NutrientH20" just to have a label. The goal here was for NutrientH20 to understand its social-media audience a little bit better, so that it could hone its messaging a little more sharply.

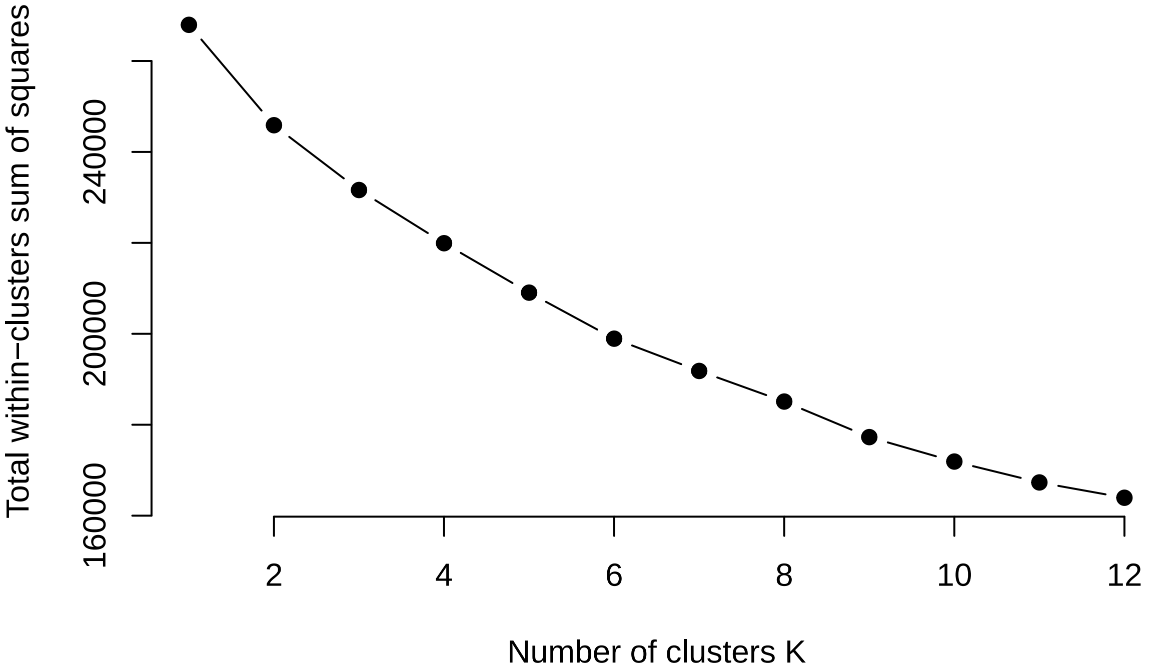
Your task to is analyze this data as you see fit, and to prepare a report for NutrientH20 that identifies any interesting market segments that appear to stand out in their social-media audience. You have complete freedom in deciding how to pre-process the data and how to define "market segment." (Is it a group of correlated interests? A cluster? A latent factor? Etc.) Just use the data to come up with some interesting, well-supported insights about the audience.

2. Data and Model

K-means clustering will be used to identify market segments based on common interests. The 'chatter' and 'uncategorized' columns will not be included in the determination of market segments because they are not useful in identifying common interests. The elbow method will be used to attempt to find the optimal number of clusters.

3. Results

In testing k=2 to k=12 clusters with nstarts=10 and plotting Number of clusters K vs. Total within clusters sum of squares (tot.withinss), we find that there is no clearly defined optimal number of clusters. For simplicity, we will use k=5 clusters.



The following table describes the size and interests of each of the 5 clusters, as well as the label assigned to each cluster based on those interests:

Cluster	Size	Common Phrases	Label
1	185	adult, spam, small business, outdoors, eco, parenting, automotive, home and garden, art, crafts	Mature
2	1540	cooking, health nutrition, personal fitness, fashion, beauty, outdoors, photo sharing, music, sports playing, eco	Fit
3	4673	spam, online gaming, current events, tv film, college uni, adult, art, shopping, dating, home and garden	Intellectual
4	707	politics, news, travel, computers, automotive, business, dating, sports fandom, small business, outdoors	Worldly
5	777	religion, parenting, sports fandom, food, school, family, crafts, beauty, eco, home and garden	Community-Minded

Phrases prevalent in multiple clusters:

- Eco occurs in clusters 1, 2, and 5.
- Outdoors occurs in clusters 1, 2, and 4.
- Home and garden occurs in clusters 1, 3, and 5.

4. Conclusion

NutrientH20 can use the groups outlined above (Mature, Fit, Intellectual, Worldly, and Community-Minded) to better target its social media audience with messages that are pertinent to particular market segments. In addition, NutrientH20 should keep in mind that the majority of its audience is interested in eco, outdoors, and home and garden related topics.