Predictive Analytics Exercise 1

Vishwa Bhuta

August 6, 2016

Probability Practice

Part A

Since users can either be an RC or a TC(these are mututally exclusive), and they can answer either yes or no(also mututally exclusive), we can say that to the survey results are:

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

P(Yes)=0.65
P(RC)=0.3
P(Y|RC)=0.5
P(TC)=0.7
Therefore, $P(Y|TC) = \frac{5}{7}$

Part B

To me, there were two ways to think about this problem to get to the same solution: first, as a confusion matrix problem, and second as a Bayes' rule problem. Both result in the same answer.

```
                        Actual
                  Positive | Negative
             -----------------------------
  Pred Pos   True Pos(TP)  | False Pos(FP)
  Pred Neg   False Neg(FN) | True Neg(TN)
             -----------------------------
             Actual Pos    | Actual Neg
```

We are looking for $\frac{TP}{TP+FP}$

$$Sensitivity = \frac{TP}{(TP+FN)} = 0.993$$

$$Specificity = \frac{TN}{(TN+FP)} = 0.9999$$

$$ActualPositive = TP + FN = 0.000025$$

$$ActualNegative = TN + FP = 1 - 0.000025 = 0.999975$$

Therefore $TP = (TP + FN) * 0.993 = 0.000024825$
Therefore $TN = (TN + FP) * 0.9999 = 0.999875$
Therefore $FP = ActualNegative - TN = 0.999975 - 0.999875 = 0.0001$
Therefore

$$\frac{TP}{TP+FP} = \frac{0.000024825}{0.000024825 + 0.0001} = 0.19888$$

Bayes Rule:

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

Where D+ means you have the disease and T+ means you tested positive. We know that $P(D+) = 0.000025$, and that $P(T+|D+) = 0.993$. We also know that $P(T+) = (P(T+|D+) * P(D+)) + (P(T+|D-) * P(D-))$ and that $P(D-) = 1 - 0.000025 = 0.999975$. The only piece we don't know, therefore, is $P(T+|D-)$.

From conditional probability rules, we can say that $P(T+|D-) = \frac{P(T+andD-)}{P(D-)}$. The probability of $P(T+andD-)$ is $P(D-) - P(T-andD-)$, since if you don't have the disease you can either test positive or negative. We do know that $P(T-|D-) = 0.9999$, and that $P(T-|D-) = \frac{P(T-andD-)}{P(D-)}$. So if $0.9999 = \frac{P(T-andD-)}{0.999975}$, $P(T-andD-) = 0.999875$, and $P(T+andD-) = 0.999975 - 0.999875 = 0.0001$. Finally, this means that $P(T+|D-) = \frac{0.0001}{0.999975}$

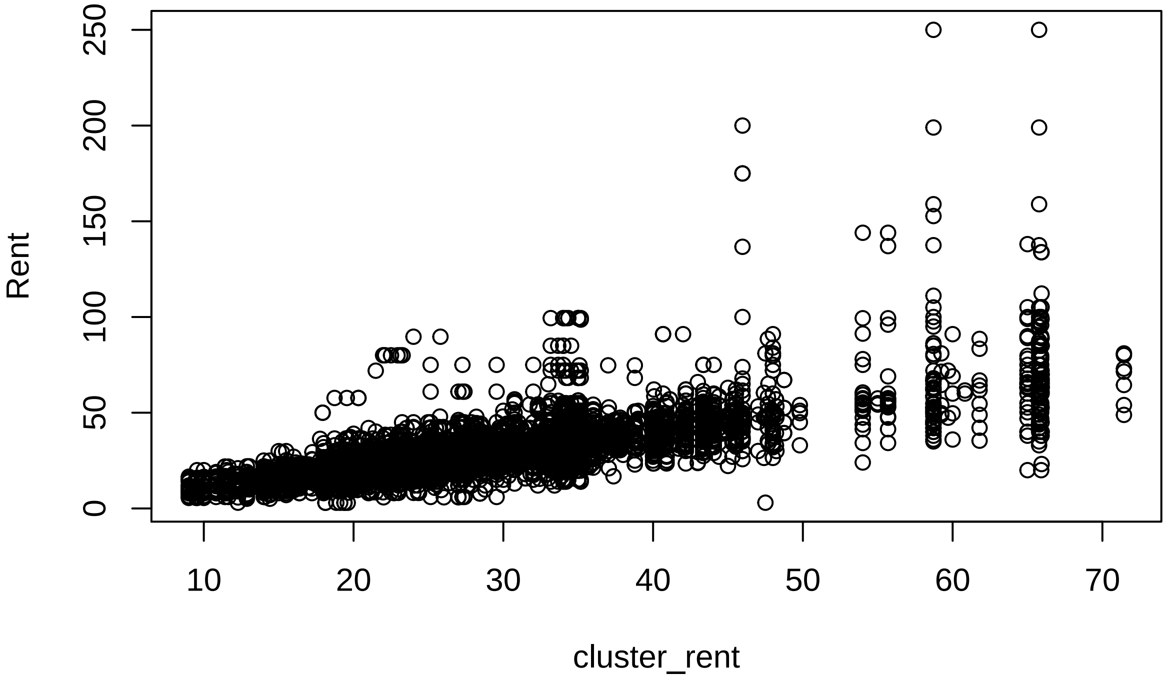
$$P(D+|T+) = \frac{P(D+) * P(T+|D+)}{P(T+)}$$

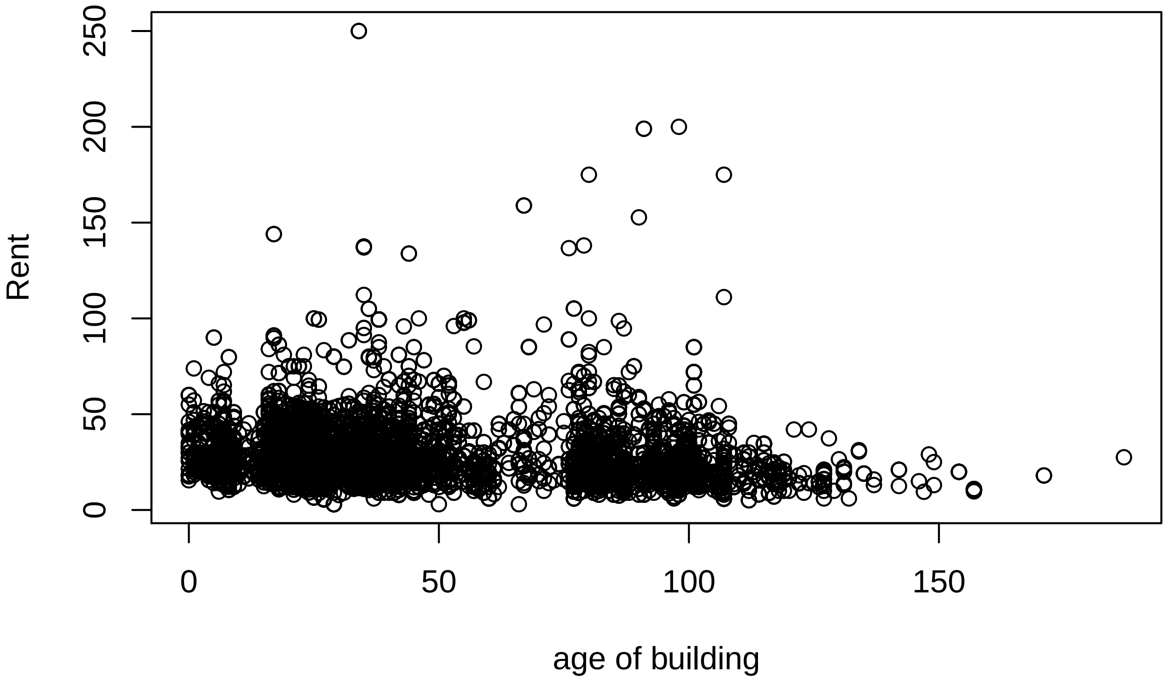
$$P(D+|T+) = \frac{0.000025 * 0.993}{(0.993 * 0.000025) + (0.0001 * 0.999975)} = 0.19888$$

Either way you get to the solution, if you can only be around 20% certain that someone who has tested positive for this test has the disease, the incentive for people to get tested is low, because they cannot trust the test results. This test could cost people a lot of money if they test positive and are among the 80% that does not have the disease, because they'll have to pay for unnecessary tests adn such. This could lead to many angry people if a universal testing policy implemented.

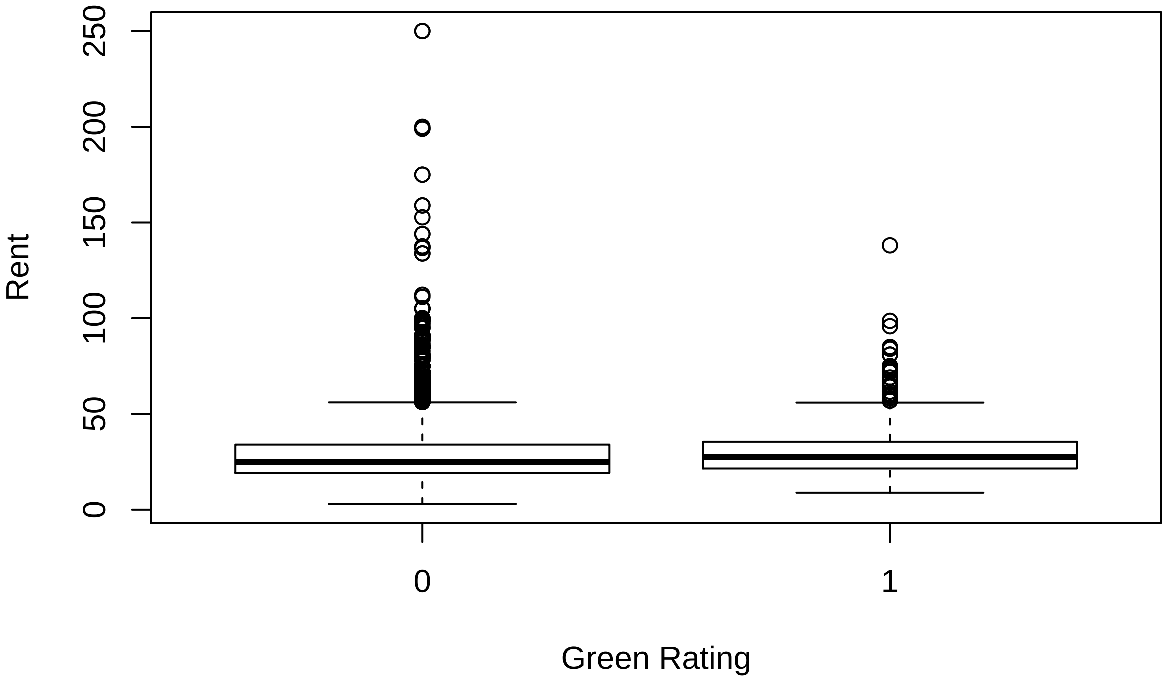
Exploratory Analysis: Green Buildings

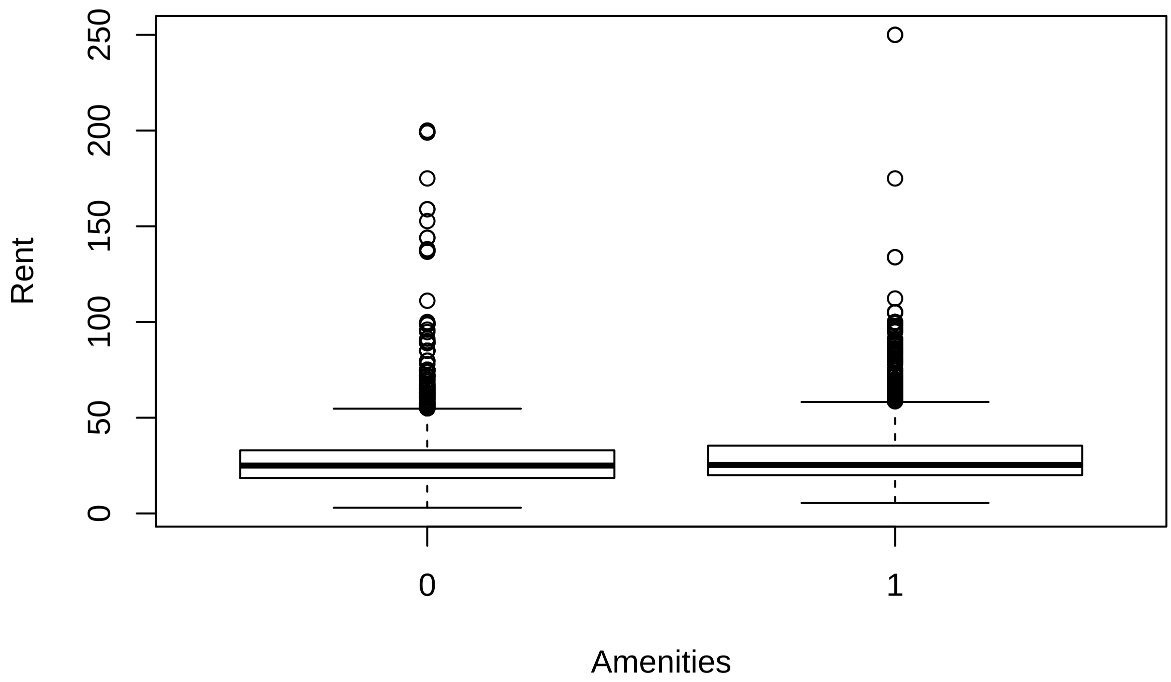
To first try and understand the data, I ran a correlations between all the numerical variables and rent to get first impressions on whether there were any linear relationships. Right off the bat, you can see that cluster rent is 76% correlated with rent, suggesting that buildings within a quarter mile of each other are likely to have similar rents, despite green rating. This brought in the potential confounding variable of a neighborhood. The median rent in green buildings could be higher because green buildings on average could be in more expensive neighborhoods. Other factors that seemed to impact rent were electricity and the total number of degree days as well as age.


Rent
0
50
100
150
200
250
10
20
30
40
50
60
70
cluster_rent

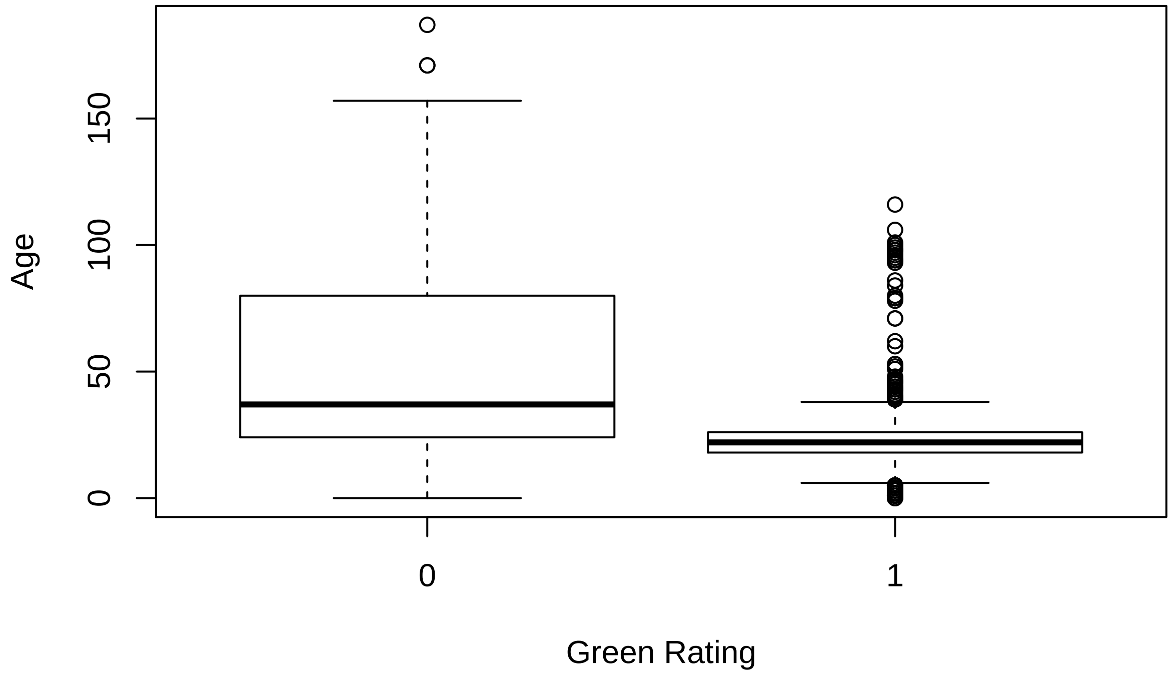

Rent
0
50
100
150
200
250
0
50
100
150
age of building

Next I ran box plots with rent and the categorical variables to find that green_rating, amenities or a specific energy rating don't seem to affect rent significantly. This finding conflicts with the stats guru's finding. I think he goes wrong because by taking the median, he is ignoring some of the variance in the data and is therefore increasing his error.

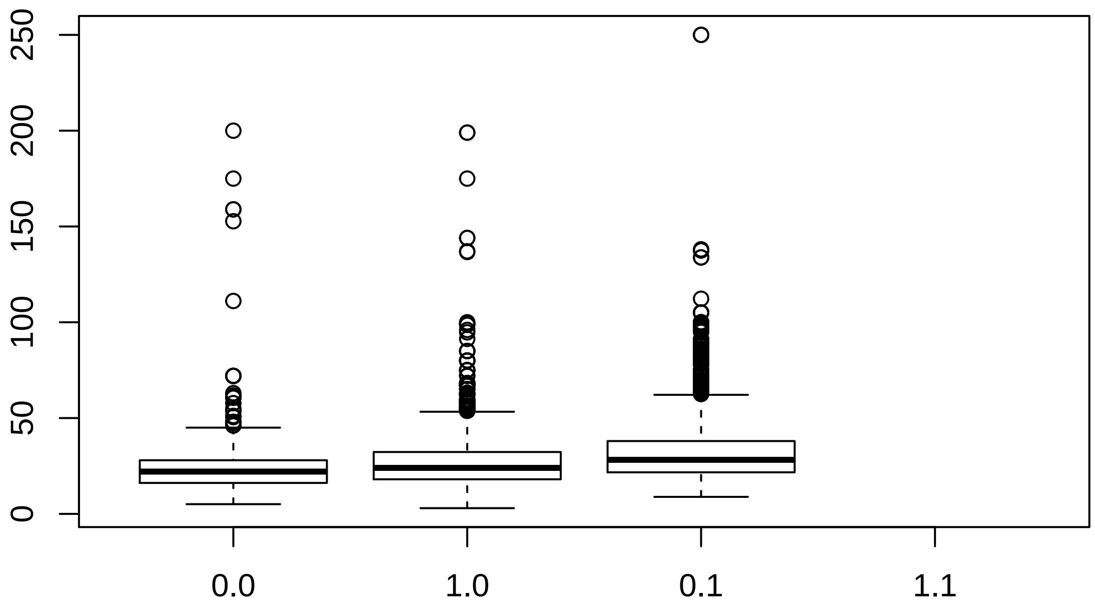




Since age and neighborhood seemed related to rent, I wanted to see if there were differences in the ages and the building quality between green and non-gren buildings, and I found that there is.



```
## , , green_rating = 0
##
##        class_a
## class_b    0    1
##       0 1103 2611
##       1 3495    0
##
## , , green_rating = 1
##
##        class_a
## class_b    0    1
##       0    7  546
##       1  132    0
```



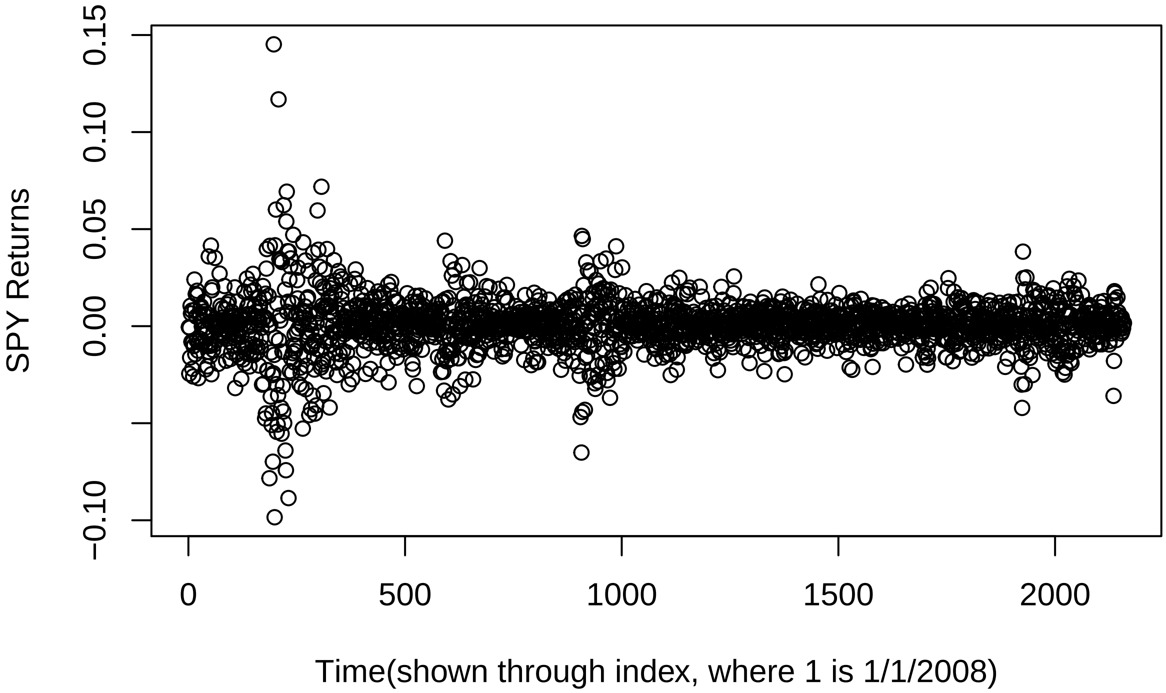
0.0 are class C, 1.0 are class B and 0.1 are class A buildings

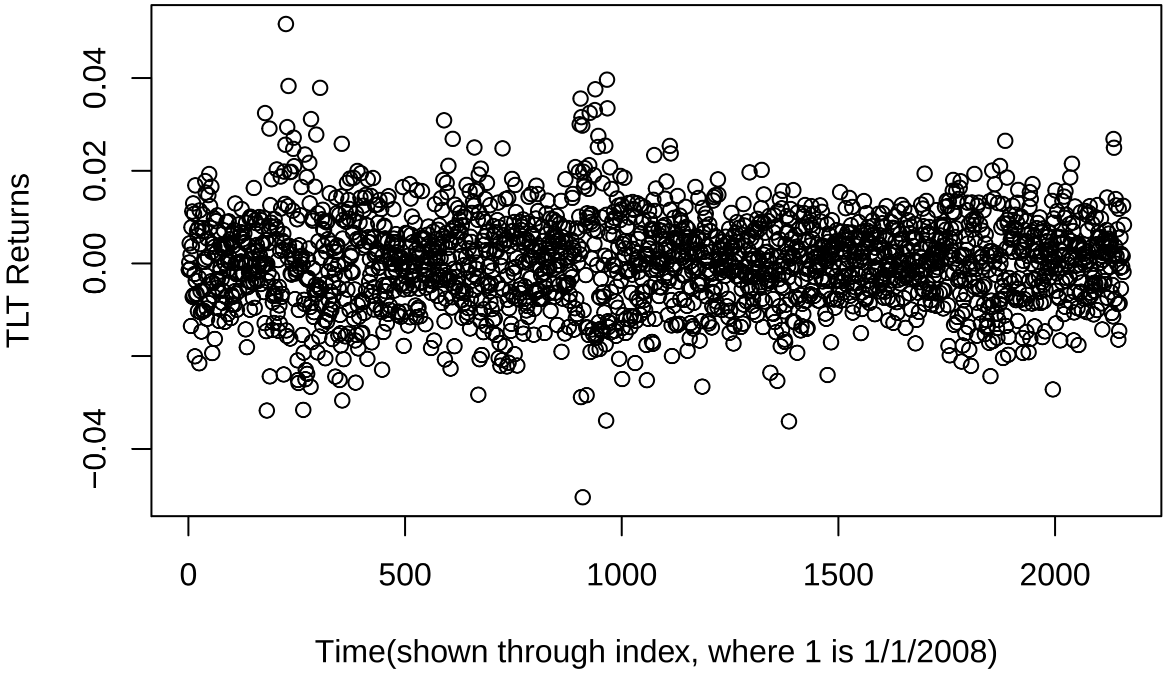
As you can see, green buildings (with some exceptions) tend to be a lot newer, and we know that newer buildings tend to bring in a slightly higher rent. Furthermore, we see from the contingency tables that most green buildings are class A, some in class B and almost none in class C. Again, there is the question of whether having a higher median rent in green buildings is simply because they are also of better quality. I plot the different classes of buildings against rent, and though the differences do not look too significant, it does seem that the mean rent is higher for Class A buildings.

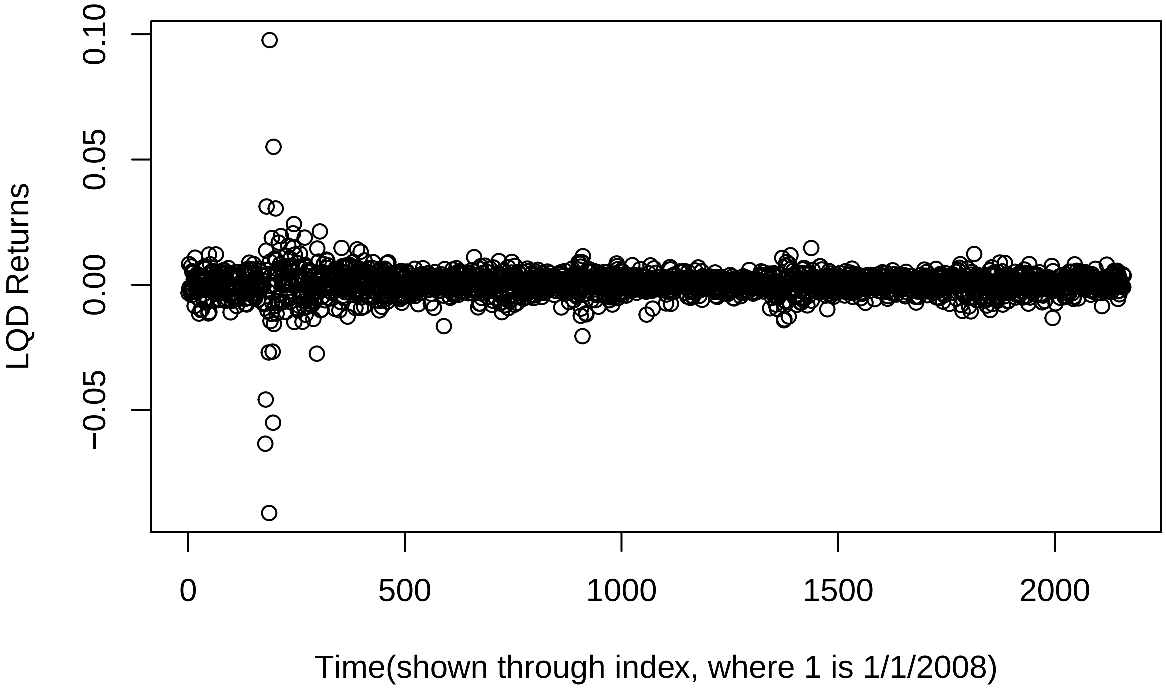
All this considered, I do not think that there is yet enough data to confidently say that building a green building will be financially worth it; it might make just as much financial sense to build a building in a nice part of town. I think there needs to be a more telling indicator of neighborhood/part of town in the data, so that we can rule out that as a confounding variable. We could then also look at buildings that are of similar age and quality and then see if there is a difference in rent in green buildings within that.

Bootstrapping

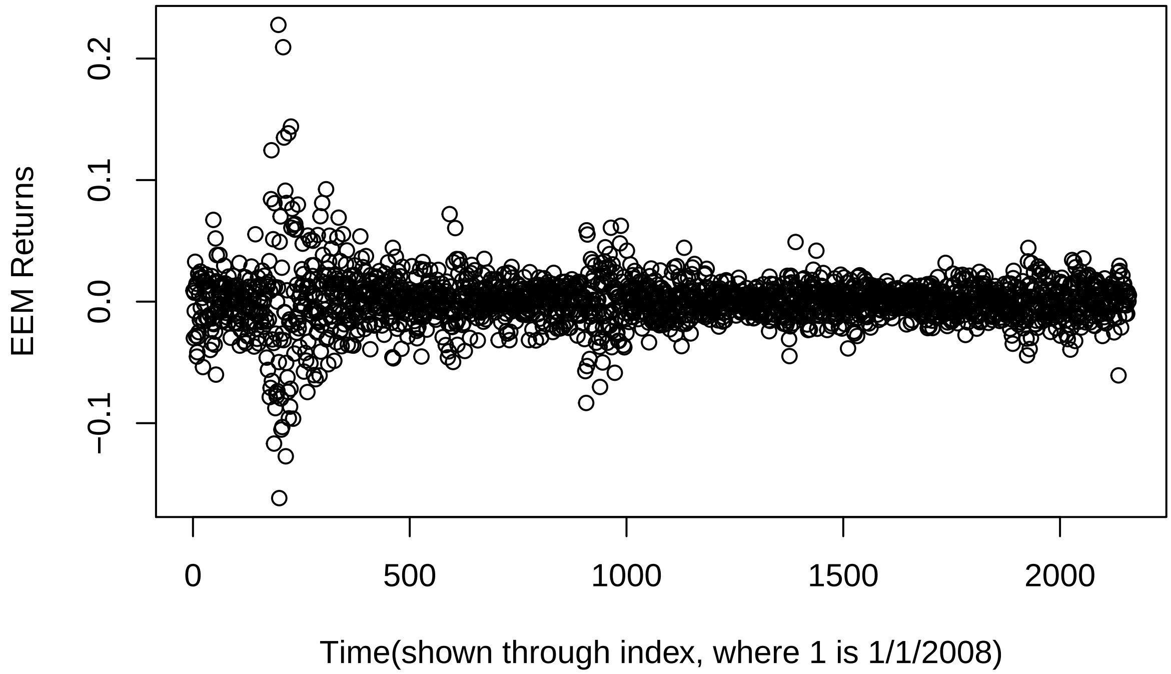
The data we are using dates back to January 2008; the data from the great recession will hopefully provide robustness in our estimates to any potential market downturns, resulting in conservative predictions.

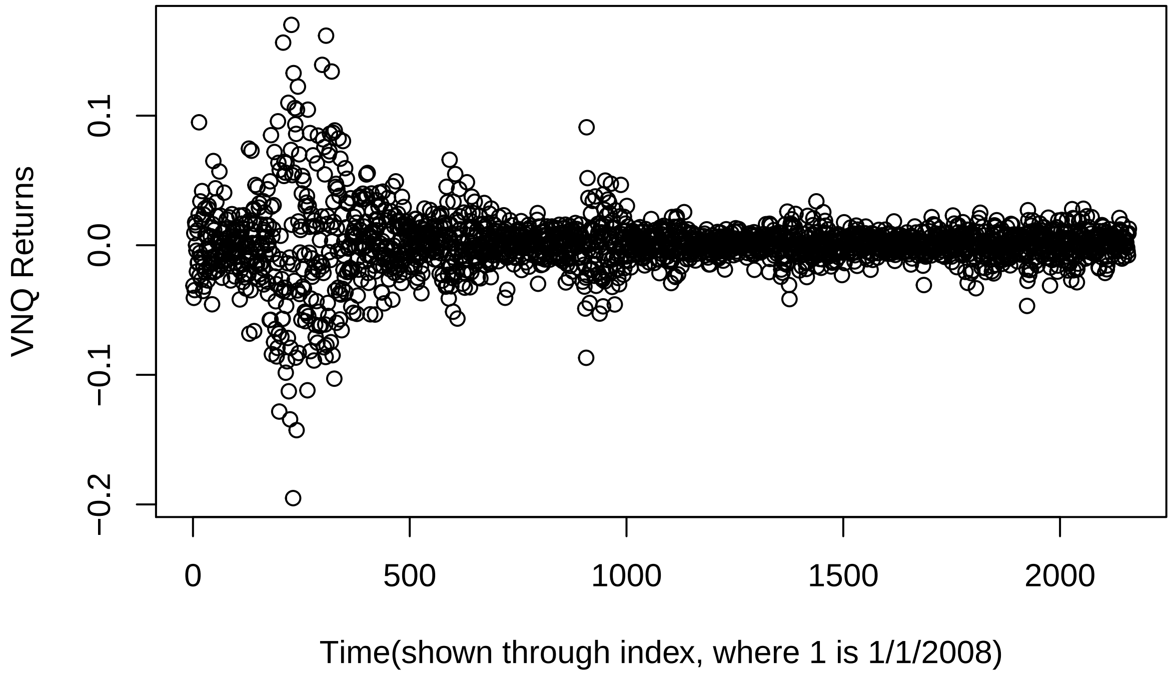

SPY Returns
0.15
0.10
0.05
0.00
−0.10
0
500
1000
1500
2000
Time(shown through index, where 1 is 1/1/2008)


TLT Returns
0.04
0.02
0.00
−0.04
0
500
1000
1500
2000
Time(shown through index, where 1 is 1/1/2008)



LQD Returns
0.10
0.05
0.00
−0.05
0
500
1000
1500
2000
Time(shown through index, where 1 is 1/1/2008)


EEM Returns
0.2
0.1
0.0
−0.1
0
500
1000
1500
2000
Time(shown through index, where 1 is 1/1/2008)



From the plots, we get a sense of the returns for each ETF classified so that we can assess their risk/reward properties. SPY represents the domestic market overall, which fluctuates with the market cycles of booms and busts. Treasury bonds have some of the most stable returns in the long run because as you can see from the plot, there doesn't seem to be any correlation between time and returns, which fluctuate within the + or - 0.02 range. Corporate bonds appear to be very steady investments with an overage 0 return and few fluctuations. EEM and VNQ both seem to fluctuate the most during the market downturns in 08 and 09, so they seem the most risky.

The even split portfolio 4 week value at risk at the 5% level is $6,687.

For the "safe" portfolio, I removed the real estate ETF (VNQ). I then assigned the remaining four assets the following weights:

- SPY - 0.3
- TLT - 0.4
- LQD - 0.2
- EEM - 0.1

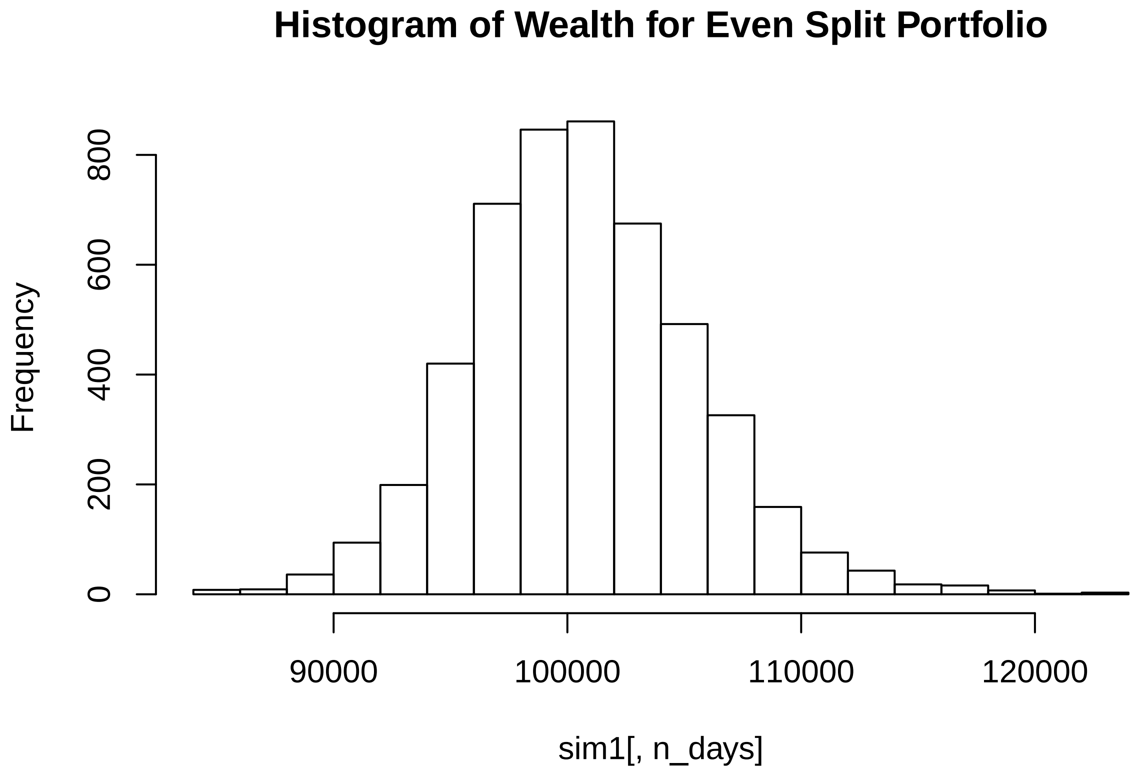
These were assigned based on how stable the returns for each type of asset are.

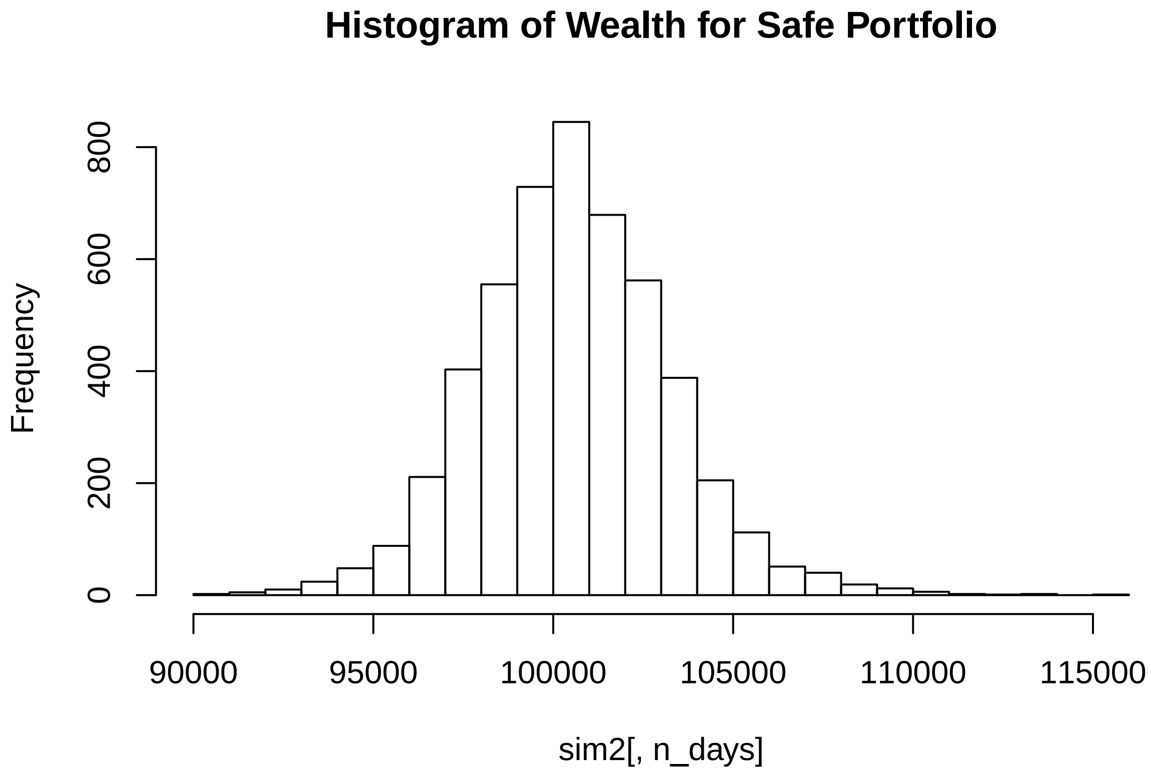
The 4 week value at risk at the 5% level for the safe portfolio is $3,560.

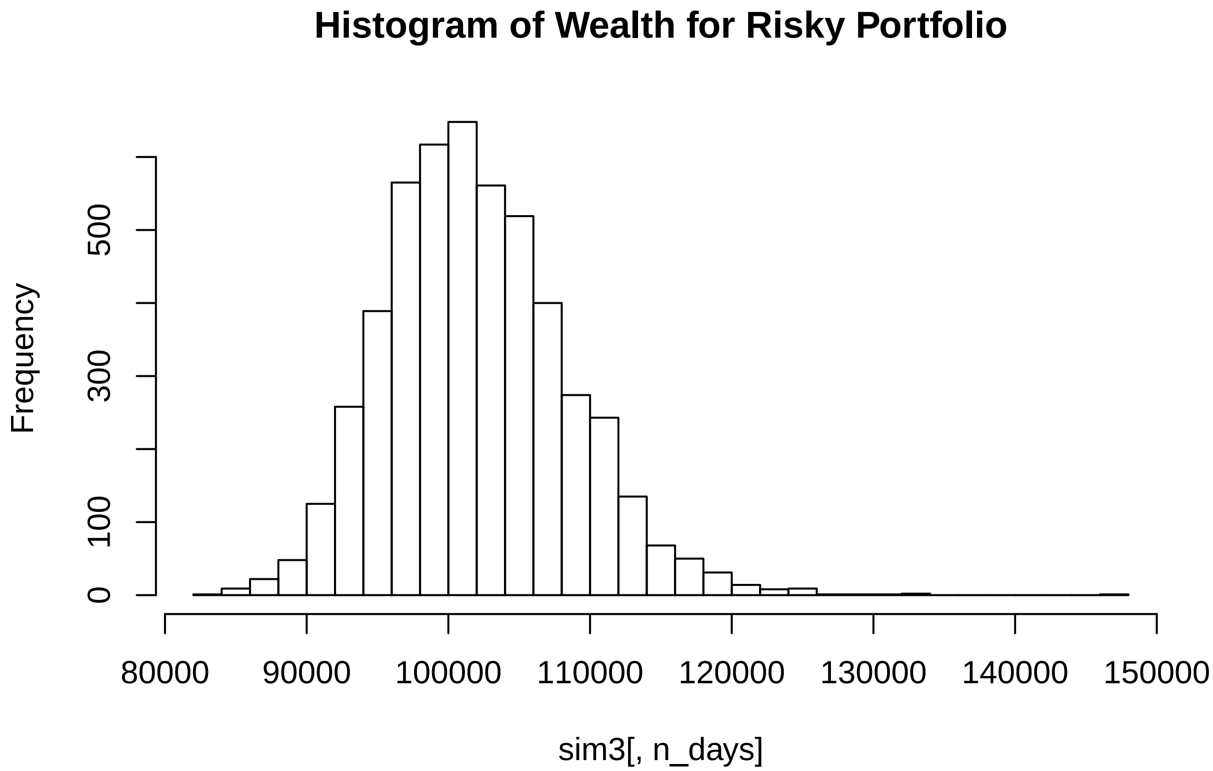
For the "aggressive" portfolio, I used 2 leveraged ETFs, with the treasury bonds as a 'stabilizer'. SSO is the ProShares Ultra S&P 500 Fund, which seeks investments that give twice the return of the S&P 500. Therefore the SSO should follow the trends of the S&P 500. The DTO, on the other hand, is the PowerShares DB Crude Oil Double Short ETN, which essentially earns returns when crude oil prices go down (ie. the

ETF seeks to short the market). I hoped that having these risky ETFs be making money sometimes in the opposite direction would help keep returns high.

The 4 week value at risk at the 5% level for the risky portfolio is $7,559.




Histogram of Wealth for Safe Portfolio
Frequency
0
200
400
600
800
90000
95000
100000
105000
110000
115000
sim2[, n_days]



As you can see from the histograms, the riskiness of the portfolio does impact the amount of wealth that can be expected at the end of the 4 week period. All the portfolios are skewed to be profitable (the even split seems to mostly break even), but the riskier portolio has the most to gain and the most to lose (although gains seem more likely from our simulations).

Market Segmentation

In approaching this problem, I decided to use Principal Component Analysis(PCA) to attempt market segmentation. Since PCA attempts to find the linear combinations of variables that best capture the variance in the data, I thought it would be a good choice in finding the few new variables(components) that would each contain parts of the original topics users tweet about. The ratio of topics contained within each new component could then be used to say something meaningful about the users.

I did not include the spam and adult categories in the analysis, as these not only do not contribute to our market segmentation purpose, but are also discouraged by Twitter. I did, however, leave in the ‘uncategorized’ label. While this label, by definition, doesn’t help us understand the user’s tweets, it does say something about the complexity of the issue we are dealing with, and I did not want to remove that. The analysis shows that 8 components explain about 78% of the variance in the original data. After 9 components(around 1/4 of our original number of variables), the variance explained by each subsequent component drops to ~1%, so 8 seems like a good stopping point.

There were some patterns in the components that could help NutritionH2O segment the market:

- Component 1:
  - High positive scores in ‘Health & nutrition’, ‘cooking’ and ‘personal fitness’ suggests that users

that have high scores in PC1 really care about their health.

  - Close to 0 on topics such as ‘college’, ‘school’, ‘parenting’ and ‘dating’, suggesting that users scoring high in PC1 might be in their mid-20s, a time when they are out of school but not quite starting families.

- Component 2:

  - High negative scores in ‘chatter’, ‘photo-sharing’, ‘cooking’, but positive scores in ‘health & nutrition’ and ‘personal fitness’. This suggests that users with negative score in PC2 tend to use twitter mostly to share photos or tweet somewhat randomly (ie. about their days).

  - The higher the user’s score in PC2, the more they seem to care about fitness and nutrition, similar to PC1, but they care substantially less about chatter and photo-sharing.

  - In the biplot below, you can see that there are many users falling into the quadrant of having high positive scores in PC1 and PC2. These are users that tweet a lot about nutrition but may not necessarily be tweeting about random things. These users could make good candidates for a marketing campaign for NutritionH2O.

- Component 3:

  - High negative scores in travel, religion, politics, news, and college. Users with negative scores in PC3, therefore, could be your politically active millenials.

- Component 4:

  - High positive scores in college, cooking, online gaming. Users with high positive scores in PC4 are likely to be college-aged and spend a lot of time online.

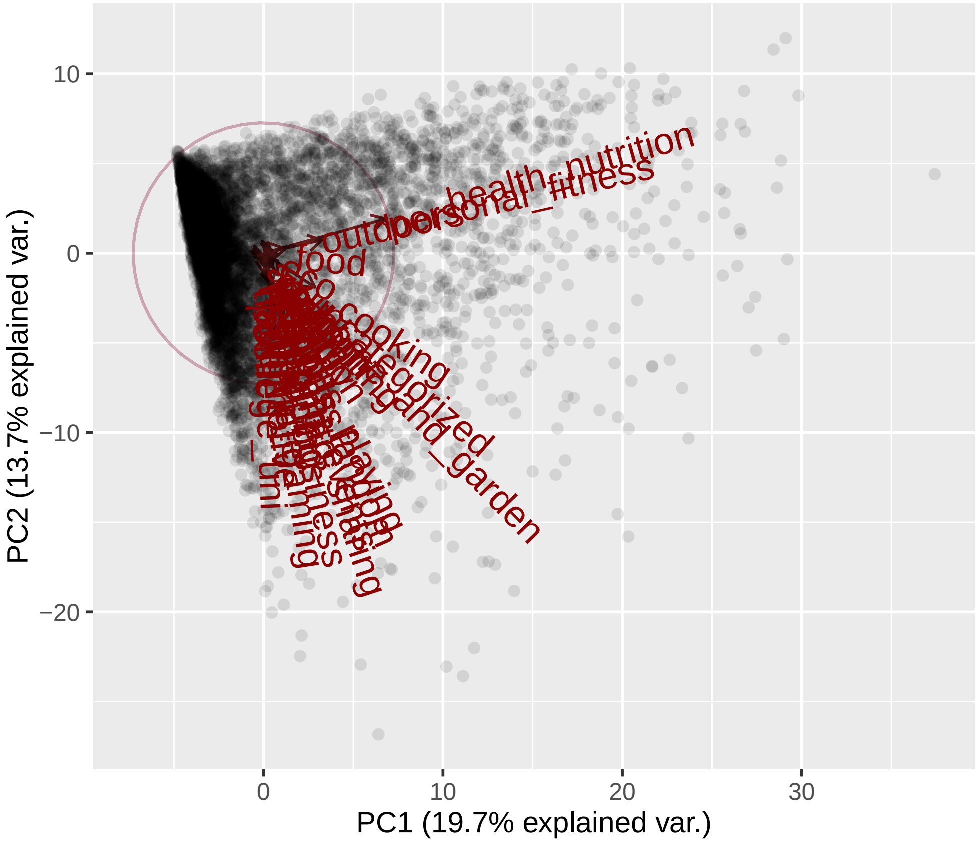
- Component 5: Not super interpretable

- Component 6:

  - High negative scores in parenting, religion, family, food and sports fandom. Users with negative scores on PC6 are likely to be parents, perhaps with more traditional family values.

- Component 7 and 8:

  - High positive scores in travel, tv/film, music, news and art. Users with positive scores on PC7 or 8 are artistic and media buffs, so perhaps their attention could be captured with beautiful graphics.


10
0
−10
−20
0
10
20
30
PC1 (19.7% explained var.)
PC2 (13.7% explained var.)
health_nutrition
personal_fitness
food
cooking
and_garden