STA 380, Part 2: Exercises 1

Valerie Roth

Probability Practice

Part A

First, I compute the probability that a user said “yes” given that they were a Truthful Clicker.

P(Y|RC) + P(Y|TC) = .65

.15 + P(Y|TC) = .65

P(Y|TC) = .5

Then, I compute the probability that a user said “no” given that they were a Truthful Clicker.

P(N|RC) + P(N|TC) = .35

.15 + P(N|TC) = .35

P(N|TC) = .2

I can check that I did not make a mistake by applying the law of total probability. The sum of all possible outcomes should be one.

Total Probability:

P(Y|TC) = .5

P(N|TC) = .2

P(Y|RC) = .15

P(N|RC) = .15

Total: 1

To calculate the fraction of Truthful Clickers who answered “yes,” I divide the probability that a Truthful Clicker responds “yes” by the sum of the probabilities of every response a Truthful Clicker could give (this is “yes” and “no”).

P(Y|TC)/(P(Y|TC) + P(N|TC)) = .5/(.5+.2) = 5/7

I find that the fraction of people who are Truthful Clickers who answered “yes” is 5/7.

Part B

From the problem description, we know that:

P(tests positive|has disease) = 0.993

P(tests positive|doesn’t have disease) = 0.0001

P(has disease) = 0.000025

P(doesn't have disease) = 0.999975

P(tests positive) = P(tests positive|has disease)*P(has disease) + P(tests positive|doesn't have disease)*P(doesn't have disease) = 0.993*0.000025 + 0.0001*0.999975 = 0.0001248225

We want to find P(has disease|tests positive). To do this we can use Bayes' Rule.

With Bayes' Rule, this is equivalent to (P(tests positive|has disease)*P(has disease))/P(tests positive).

(0.993 * 0.000025)/0.0001248225 = 0.1988824130265

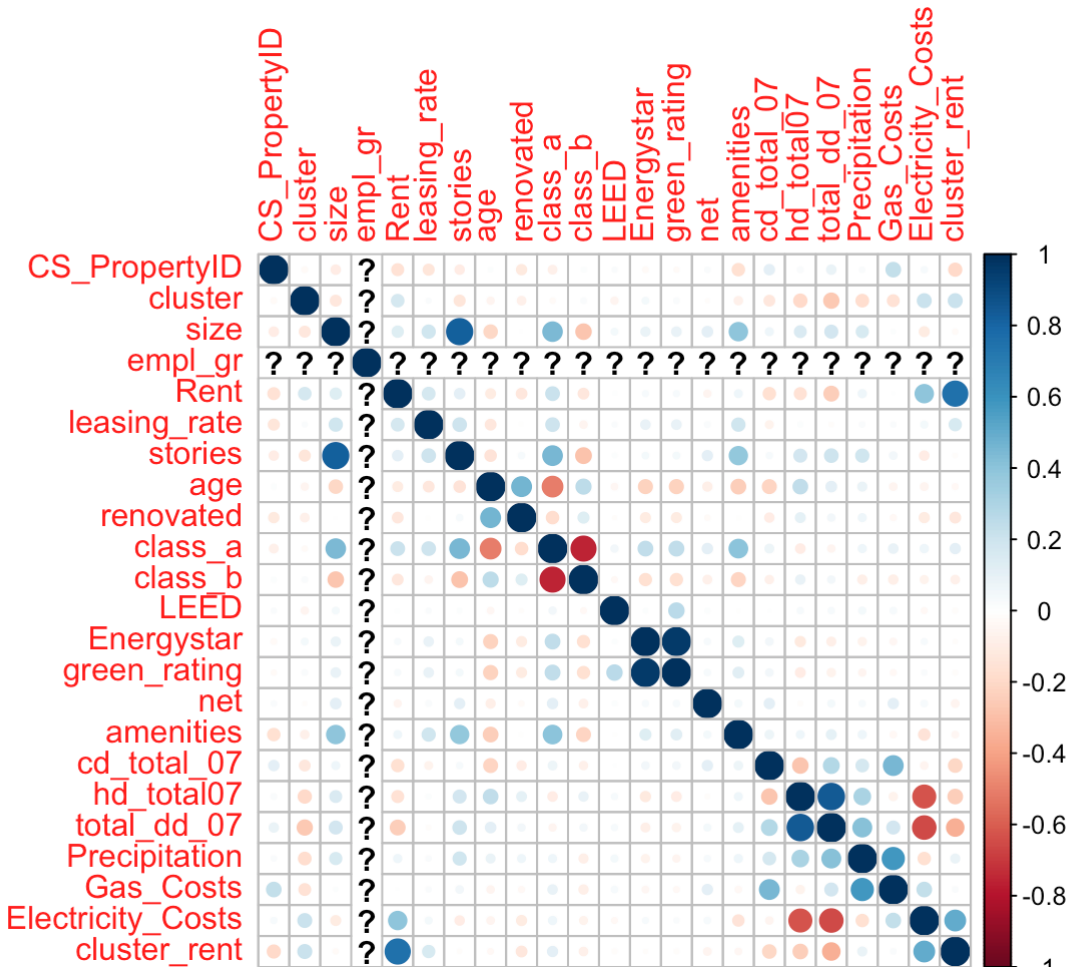
If someone tests positive, there is about a 19.9% chance that they have the disease. In light of this calculation, having a universal testing policy for this disease does not make sense. If someone tests positive, it is still unlikely that they actually have the disease.

Exploratory Analysis: Green Buildings

First, I did a simple linear regression using only green_rating as a predictor. I got -0.6632 as a coefficient for green_rating. The fact that this coefficient is negative is interesting because it means that given a building where everything else is equal, adding a green rating to that building may actually make the rent cheaper. The rent per square foot without the green_rating is given by the coefficient: $27.55.

```
##
## Call:
## lm(formula = cluster_rent ~ green_rating, data = greenbuildings)
##
## Coefficients:
##  (Intercept)  green_rating
##      27.5548       -0.6632
```

Of course, the Excel guru found that having a green building increases rent. How can this be? To explain, I will show the correlation between green_rating and all of the other predictors.



Looking at this matrix, we see that green_rating is positively correlated with size, leasing_rate, class_a, LEED, Energystar and amenities. These things could be considered "perks," and therefore may be driving the increase in rent for green_buildings more than the certification itself.

Furthermore, green_rating is negatively correlated with age, renovated, class_b, and hd_total07. This means that green buildings are likely newer, not in need of renovations as much as non-energy star buildings, probably mostly "A" class, and do not need to be heated as much.

This could certainly explain why green buildings are positively correlated with cluster rent.

The guru's error was essentially that the correlation between green_rating and cluster_rent was causation. He should have realized that there could be (and were) confounding variables like some of the ones listed above are in his analysis.

Bootstrapping

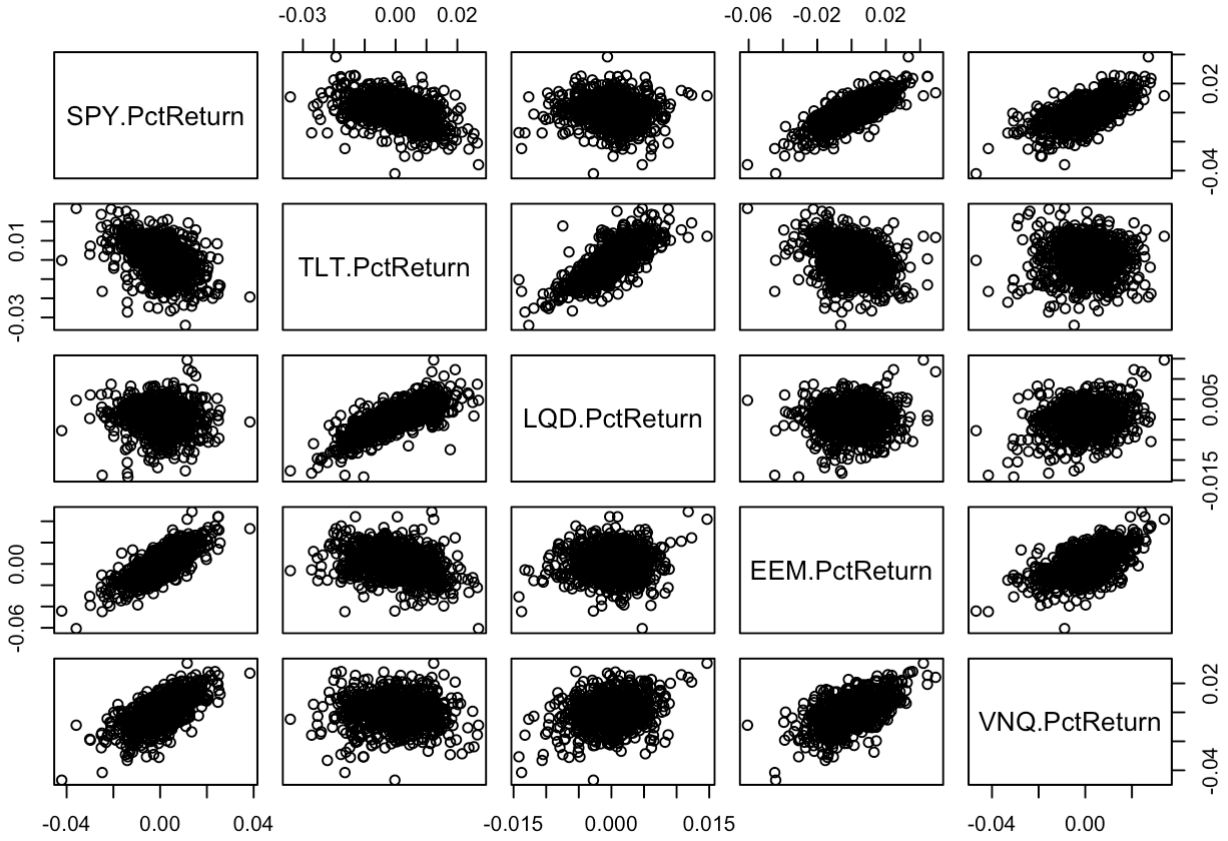
First, I performed the even split. After 20 days, my original $100000 would have turned into the amount below.

```
## [1] 102424.1
```

When correlations between stocks are positive, there is more risk and reward associated with that stock. Below, we see that: * SPY and TLT are negatively correlated (less risk/reward)

- SPY and LQD are not correlated (not particularly risky or safe)
- SPY and EEM are positively correlated (more risk/reward)

- SPY and VNQ are positively correlated (more risk/reward)
- TLT and LQD are positively correlated (more risk/reward)
- TLT and EEM are negatively correlated (less risk/reward)
- TLT and VNQ are not correlated (not particularly risky or safe)
- LQD and EEM are not correlated (not particularly risky or safe)
- LQD and VNQ are not correlated (not particularly risky or safe)
- EEM and VNQ are positively correlated (more risk/reward)



My safe portfolio includes EEM, SPY and TLT, with stock in each in approximately equal amounts. To check for safety, I look at the following three pairs:

- EEM and SPY are positively correlated (sharply)
- EEM and TLT are negatively correlated (not sharply)
- SPY and TLT are negatively correlated (not sharply)

This means that when EEM is doing well, SPY and TLT are not doing well (but each to a lesser extent than EEM). This balances out well. When SPY and TLT are doing well, EEM is not doing well. This creates a more balanced portfolio with less risk and less reward.

The results of my safe portfolio after 20 days is below:

```
## [1] 101882
```

My aggressive portfolio includes SPY, EEM and VNQ, with stock in each in approximately equal amounts. To check for potential risk/reward, I look at the following three pairs:

- SPY and EEM are positively correlated
- SPY and VNQ are positively correlated
- EEM and VNQ are positively correlated

This means that when these stocks are doing well, there could be a great deal of reward. When they are doing poorly, your entire portfolio is doing poorly and this causes risk.

The results of my aggressive portfolio after 20 days is below:

```
## [1] 104588.6
```

To determine which portfolio allocation is right for the user, one would have to ask themselves if they were willing to risk more of their money to potentially make more or if they simply want a safer allocation.

Market segmentation

I chose to use k-means clustering to see if I could divide the users into different groups. This would be useful, because if NutrientWater understood its market groups better, it could target them more specifically.

I chose 7 centers for my clusters. I also set iter.max to 30, which is much higher than the default of 10 so it would converge. Below, you can see a table representing how many users ended up in each cluster.

```
##
##    1    2    3    4    5    6    7
##  706 3583  517  371  620 1280  805
```

If you want to learn more about a particular cluster, you can look at the results below. For example, cluster 1 does not involve tv_film, fashion, or music to a significant degree but it does involve religion, school and parenting heavily. These users may be driven to buy NutrientWater because of values that they hold and want to share with their children. They are likely not motivated to buy this beverage because of trends.

```
##       chatter current_events       travel photo_sharing uncategorized
## 1 -0.13305794     0.11995353 -0.094827851   -0.09226335   -0.07866067
## 2 -0.35103166    -0.18181238 -0.233368720   -0.41192945   -0.17951679
## 3 -0.03461507     0.19347438 -0.035076430    1.21071899    0.47884662
## 4 -0.07785709    -0.07911350  0.002315739   -0.02714102    0.02413614
## 5 -0.05677598     0.09751640  1.907723116   -0.15529755   -0.06558583
## 6  1.20395250     0.34872012 -0.116495457    0.86179578    0.27579333
## 7 -0.13340595    -0.01335466 -0.140733151   -0.10137312    0.16133246
##       tv_film sports_fandom   politics        food      family
## 1 -0.01504775     2.0769610 -0.2069072  1.84307412  1.49840781
## 2 -0.17731159    -0.3003324 -0.2736223 -0.36156299 -0.28911892
## 3 -0.05170092    -0.1880640 -0.1102192 -0.18138122  0.03822276
## 4  0.23061689    -0.1177227 -0.1534691 -0.07661393  0.21063648
## 5  0.02180205     0.2005665  2.4839418  0.03865890  0.02620727
## 6  0.50867148    -0.1758277 -0.1266679 -0.20934649 -0.06259993
## 7 -0.09629210    -0.1846348 -0.1708365  0.44777977 -0.06955371
##   home_and_garden       music        news online_gaming    shopping
## 1      0.17381119  0.03635352 -0.06749503   -0.07190773 -0.03202231
## 2     -0.19760703 -0.21888964 -0.24131748   -0.23379324 -0.37371868
## 3      0.12289889  0.52546206 -0.05364115   -0.03380206  0.21674768
## 4      0.11645173  0.01567800 -0.16583779    3.49684154 -0.12137953
## 5      0.09234485 -0.05347076  1.99355308   -0.13532045 -0.11628184
## 6      0.23068683  0.38571610 -0.17184108   -0.16665995  1.09773410
## 7      0.15657144  0.02555385 -0.01800747   -0.11699513 -0.04768930
##   health_nutrition college_uni sports_playing    cooking         eco
## 1      -0.14254352 -0.12365986    0.109973540 -0.1072649  0.17355010
## 2      -0.31933758 -0.26008238   -0.265837094 -0.3355420 -0.27469728
## 3      -0.04614486 -0.02876235    0.175565380  2.7795498  0.01542295
## 4      -0.17352198  3.27874746    2.174160058 -0.1275756 -0.04575346
## 5      -0.20687617 -0.09334938   -0.013707012 -0.2172043  0.09506949
## 6      -0.23788953  0.03361753    0.003085161 -0.2363667  0.29560852
## 7       2.19356229 -0.20809936   -0.022333358  0.4043421  0.53837621
##     computers    business    outdoors      crafts  automotive         art
## 1  0.07776430  0.09244835 -0.07092743  0.72274611  0.15771336  0.08803269
## 2 -0.26635982 -0.26019053 -0.31159115 -0.28984992 -0.23011784 -0.19302284
## 3  0.05646131  0.22958842  0.04695316  0.11674106  0.02542762  0.10358586
## 4 -0.05211971 -0.05848460 -0.12562526  0.09112787  0.06929160  0.31277182
## 5  1.67870864  0.31352736  0.10100956  0.10300708  1.04541815 -0.07659420
## 6 -0.06346592  0.42341467 -0.20012219  0.23977526  0.09422695  0.40330453
## 7 -0.08689618  0.04178504  1.71722800  0.07867479 -0.11733718 -0.01103584
##      religion     beauty   parenting       dating      school
## 1  2.26548658  0.3180412  2.15221661  0.009641859  1.68194819
## 2 -0.29964931 -0.2794432 -0.31957840 -0.194015523 -0.32052929
## 3 -0.11649436  2.5467901 -0.07573347  0.021145791  0.15154697
## 4 -0.16103014 -0.2184321 -0.13485827  0.007971062 -0.21281027
## 5 -0.01107864 -0.1751411  0.04020059  0.190277103 -0.04611395
## 6 -0.21180137 -0.1427258 -0.18576199  0.333171561  0.09556118
## 7 -0.15881462 -0.2082830 -0.08991047  0.161527237 -0.16412896
##   personal_fitness     fashion small_business         spam        adult
## 1      -0.09022908  0.01194431     0.09061815 -0.009661822  0.009570746
## 2      -0.33668879 -0.29868205    -0.21134526 -0.007317100 -0.007096562
## 3      -0.02564339  2.64974037     0.19749025 -0.031240493  0.025045780
## 4      -0.18986622 -0.05856808     0.13613144  0.051762770 -0.005400287
```

```
## 5      -0.19613097 -0.18602087     0.21779904 -0.019591343 -0.094332927
## 6      -0.20834806 -0.05922801     0.40252621  0.025493515  0.038232686
## 7       2.16402781 -0.11838153    -0.13615219  0.011802107  0.021457702
```

Alternatively, cluster 6 is the most influenced by tv_film. It is also influenced by photo_sharing, chatter, and shopping. This cluster of twitter users is likely much more tech-savvy and likely to buy vitamin water because it is trendy.

That said, there are many more people in cluster 1 than cluster 6 according to the first table. Therefore, to increase sales, it may be wise to spend more time marketing to the first group.