

STA 380, Part 2: Exercises 1

Kemei Zhuo

August 2, 2018

Part A.

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?

```
rc_yes = 0.5
rc_no = 0.5
rc_weight = 0.3
tc_weight = 1 - rc_weight
tc_yes = (0.65 - rc_weight*rc_yes) / tc_weight
tc_no = 1 - tc_yes

## [1] 0.7142857
```

Part B.

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?

The people with no disease have a very high weight. Even though that the specificity is high, the high probability of people with no disease still makes it a large number compared to the True positive rate times the probability of having the disease. Thus, it is hard to test the effectiveness of medical test.

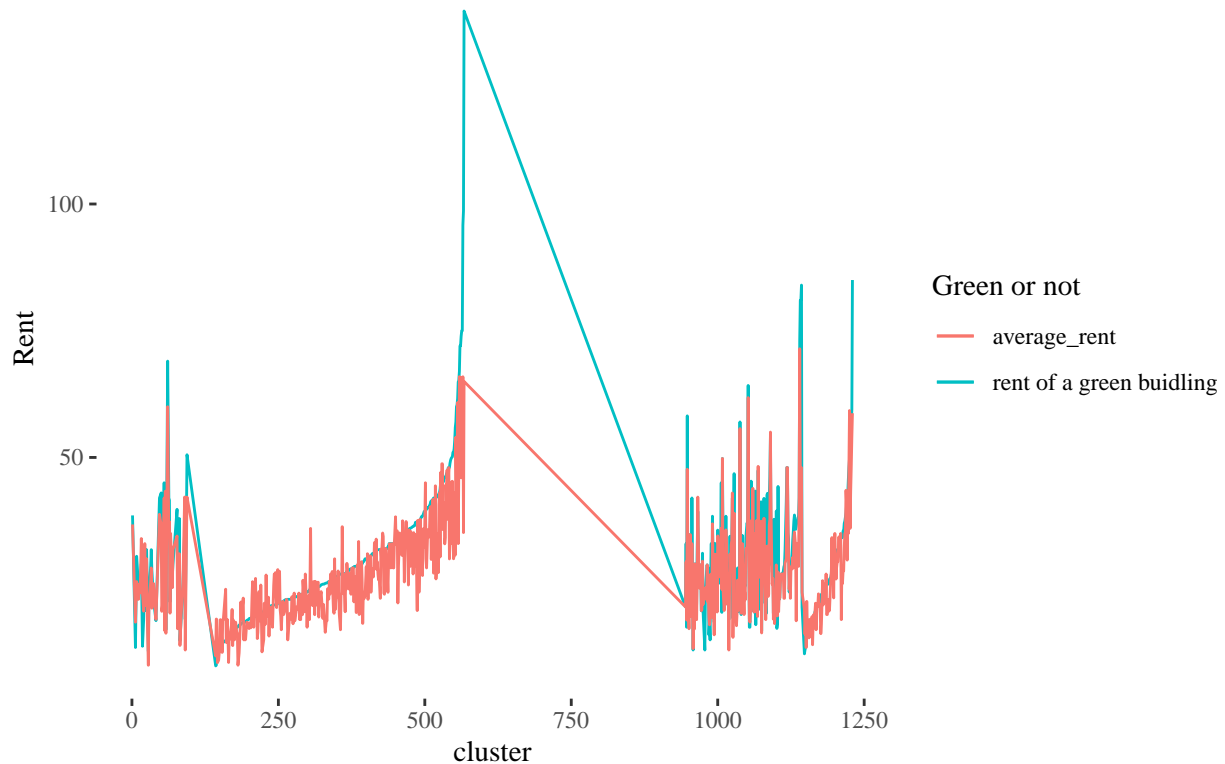
```
True_positive = 0.993
True_negative = 0.9999
False_positive = 1 - True_negative
Disease_proba = 0.000025
Positive_percentage = True_positive * Disease_proba + False_positive * (1-Disease_proba)
True_proba = True_positive * Disease_proba / Positive_percentage

## [1] 0.1988824
```

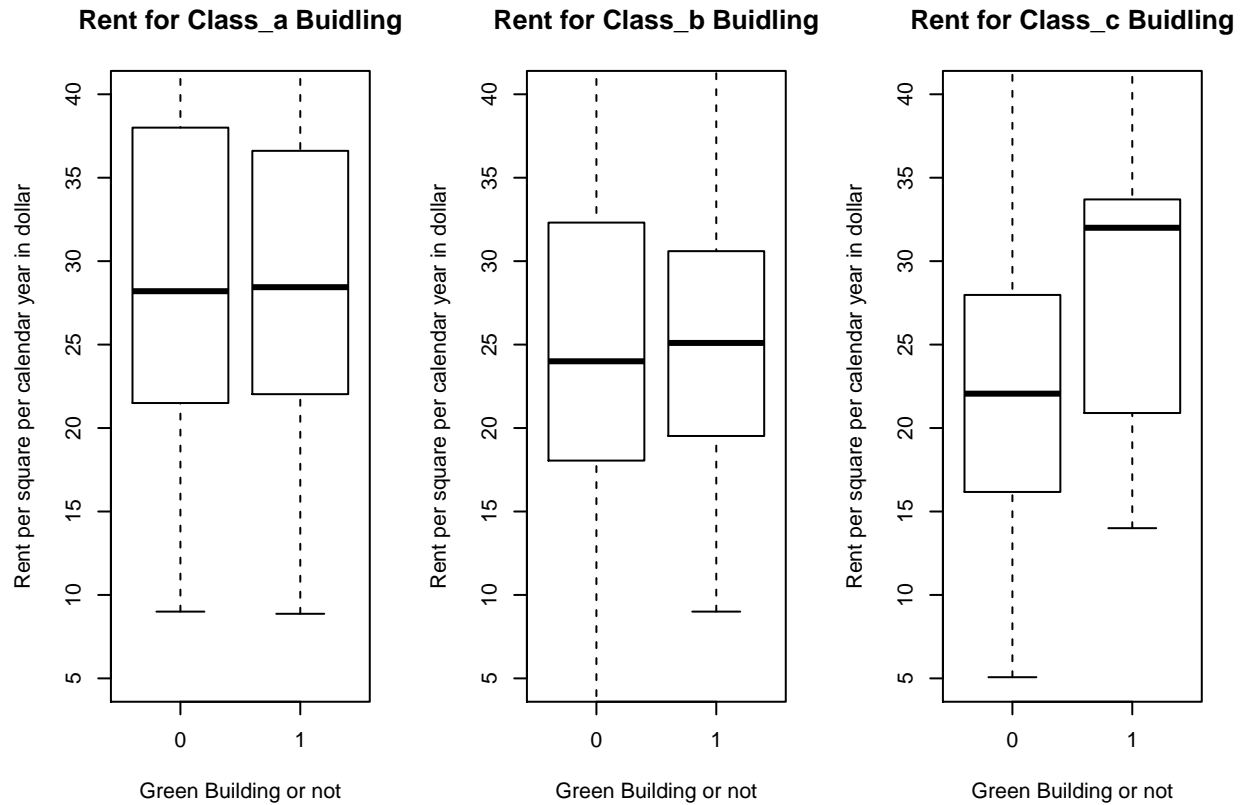
Exploratory analysis: green buildings

An Austin real-estate developer is interested in building a new 15-story mixed-use building on East Cesar Chavez, just across I-35 from downtown. The baseline construction costs are \$100 million, with a 5% expected premium for green certification. She wants to know that whether it would be economically worth to invest in a green building rather than an ordinary building.

Rent premium in Green Buildings vs Average Rent Premium in Same Region



We know that a green building is associated with social responsibility and ecological awareness, and might therefore command a premium from potential tenants. We can see this relationship through the plot above: rent rate of green buildings are higher than average price of the same region most of the time.

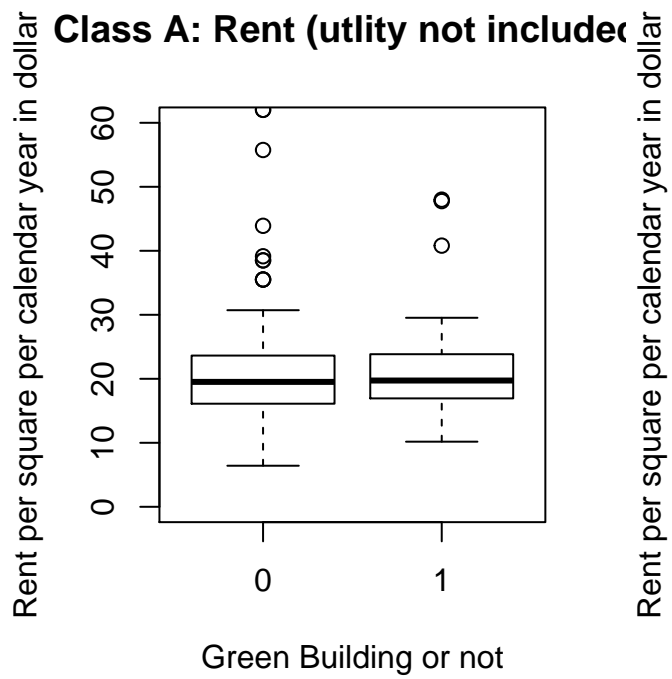


By looking at the rent under three classes of buildings: class a, class b, and class c, the rent premium decreases as the class of building decreases. It is surprising that a class c green building has the highest rate. However, a high rent premium is usually associated with low occupancy rate. Thus, we take both rent and leasing rate into consideration.

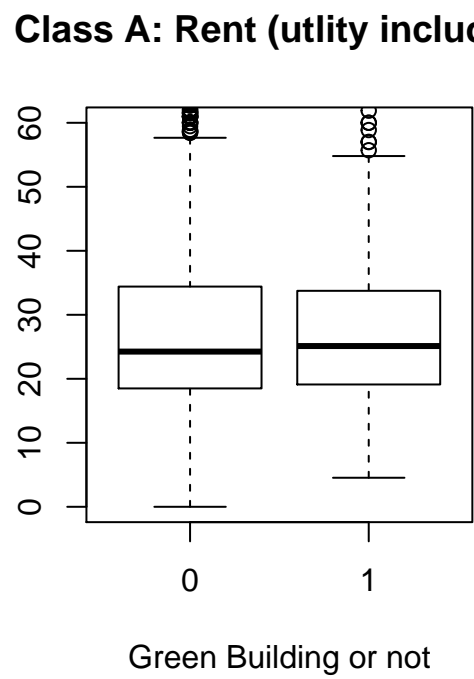
```
data['rent_income_per_square'] = data['Rent'] * data['leasing_rate'] / 100
```

Because some rent premiums do not include utility bills, I separated the data into two groups by the variable 'net' and compared the rent income of a green building with the rent income of an ordinary building under three classes. The median rent income per square of a green building is much higher than an ordinary building.

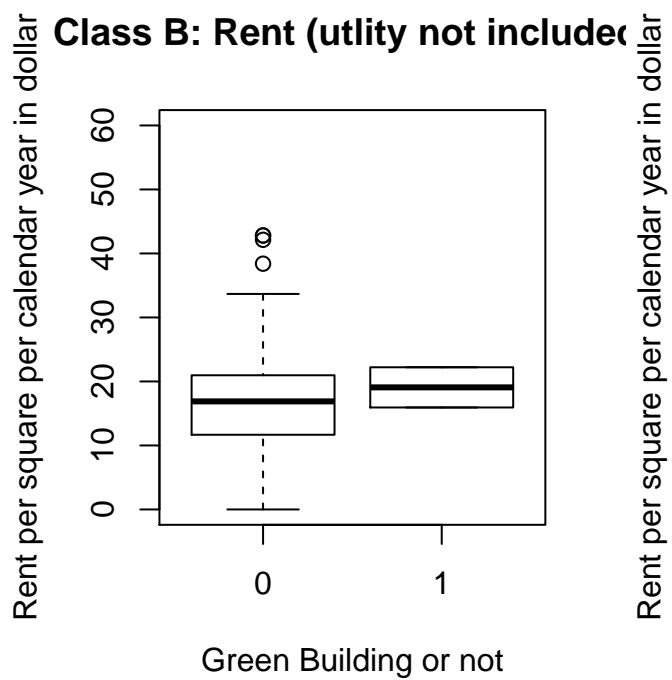
Class A: Rent (utlity not included)



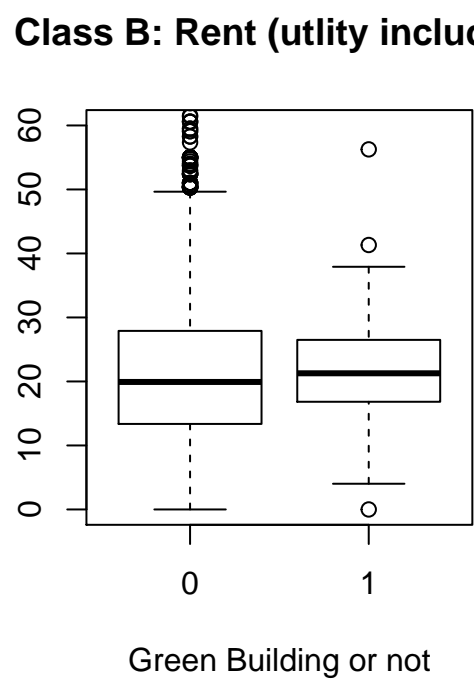
Class A: Rent (utlity included)

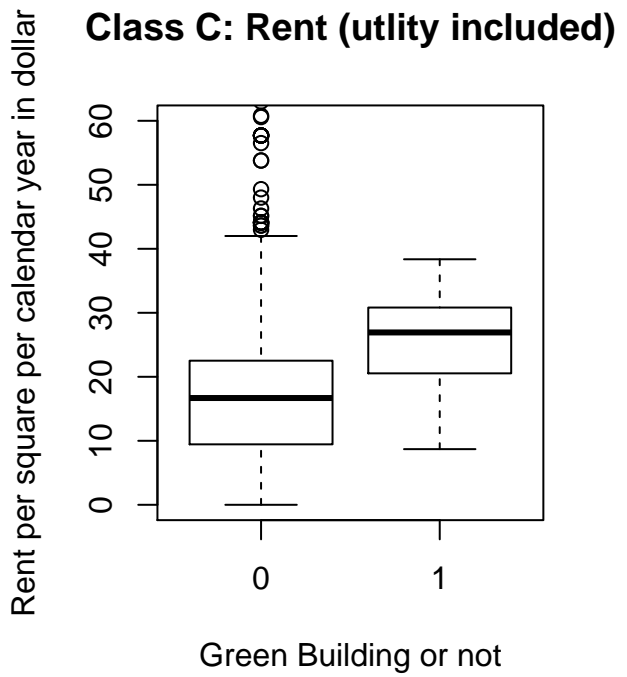
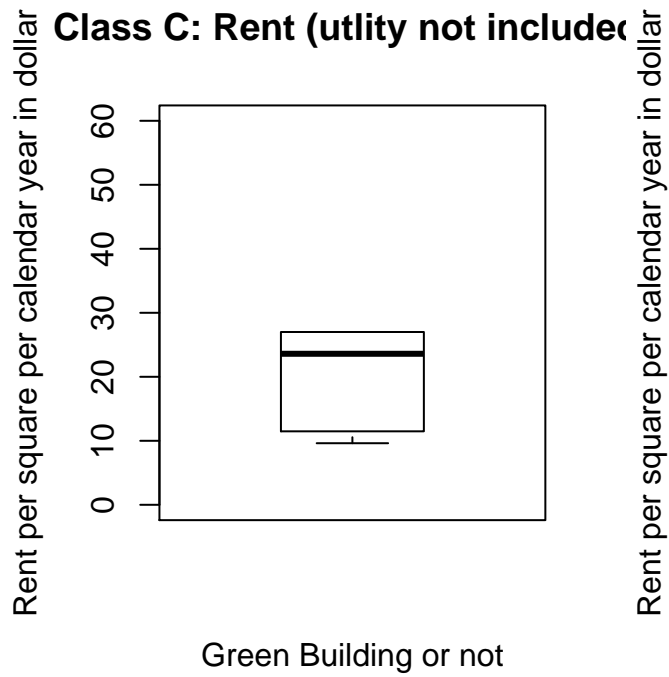


Class B: Rent (utlity not included)

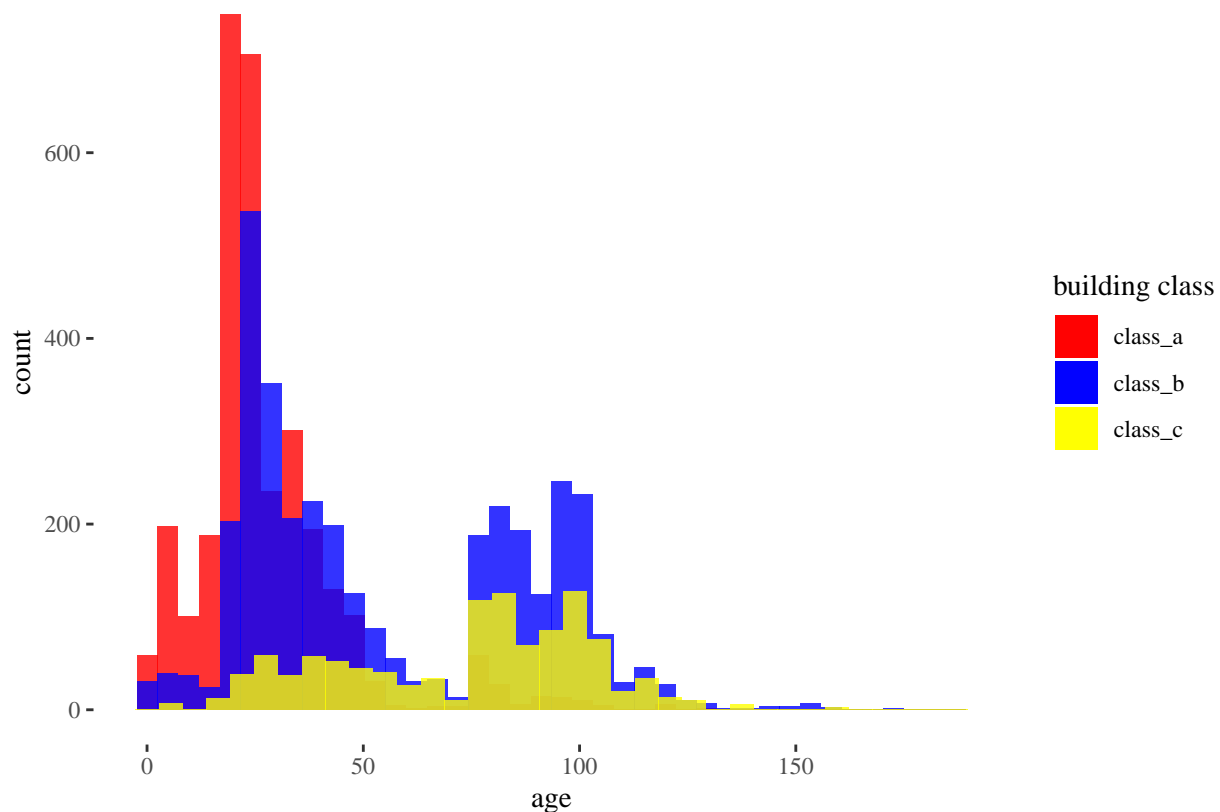


Class B: Rent (utlity included)





We don't know which class our building will be classified into for sure, but we know that class of building is associated with building age. Assuming that the developer wants to build a good building and our building will be new, our building will be likely to be classified into class a.



The majority of buildings include utility bill in rent premium, I use its medium as our base rent income per square. Because the building has 250,000 square feet, the total income from tenants per calendar year would be 4.9325×10^6 . If the developer decides to build an ordinary building, the total income from tenants per calendar year would be 4.87872×10^6 . The time to recover the total cost would be 21.2873796 years for green building but 20.4971796 years for non-green building if we don't consider the interest rate or inflation rate. Thus, building an ordinary building saves 9 month to recover the cost. Because we used the rent income of non-net contract, a part our rent income is used to pay utility bills. A green building would have low utility costs so we pay less each year. There are other benefits associated with green buildings. Green buildings would have a longer building life with lower energy risk, and higher employee productivity. As you can see from the plot, a class c green building also have a high rent premium. Thus, by looking far into the future, the developer should build a green building.

Bootstrapping

```
## 'getSymbols' currently uses auto.assign=TRUE by default, but will
## use auto.assign=FALSE in 0.5-0. You will still be able to use
## 'loadSymbols' to automatically load data. getOption("getSymbols.env")
## and getOption("getSymbols.auto.assign") will still be checked for
## alternate defaults.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.warning4.0"=FALSE). See ?getSymbols for details.
##
## WARNING: There have been significant changes to Yahoo Finance data.
```

```
## Please see the Warning section of '?getSymbols.yahoo' for details.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.yahoo.warning"=FALSE).
## [1] "SPY" "TLT" "LQD" "EEM" "VNQ"
```

According to the standard deviation of daily returns below, we can see that the order of risk: Investment-grade corporate bonds < US Treasury bonds < US domestic equities < Real Estate < Emerging-market equities. My choice of a safer portfolios contains 30% of US domestic equities, 50% US Treasury bonds, and 20% of investment-grade corporate bonds. My choice of an aggressive portfolio contains 25% of US domestic equities, 50% of emerging-market equities, and 25% of real estate. The 5% VaR for the even split portfolio is -6062.159 ; the 5% VaR for the safe portfolio is -3182.673; the 5% VaR for the aggressive portfolio is -11356.04. The mean wealth increase for the even split portfolio is 957.7837; The mean wealth increase for the safe portfolio is 551.642; The mean wealth increase for the aggressive portfolio is 1190.229. Thus, we can see that rewards are associated with higher risk.

```
# the standard deviation
c(sd1, sd2, sd3, sd4, sd5)
```

```
## [1] 0.012450525 0.009157254 0.005220425 0.040247536 0.021163391
```

portfolio	5% VaR	Mean increase in wealth
the even split	-6403.877	872.721
safer split	-3151.561	644.712
aggressive split	-11576.890	1292.196

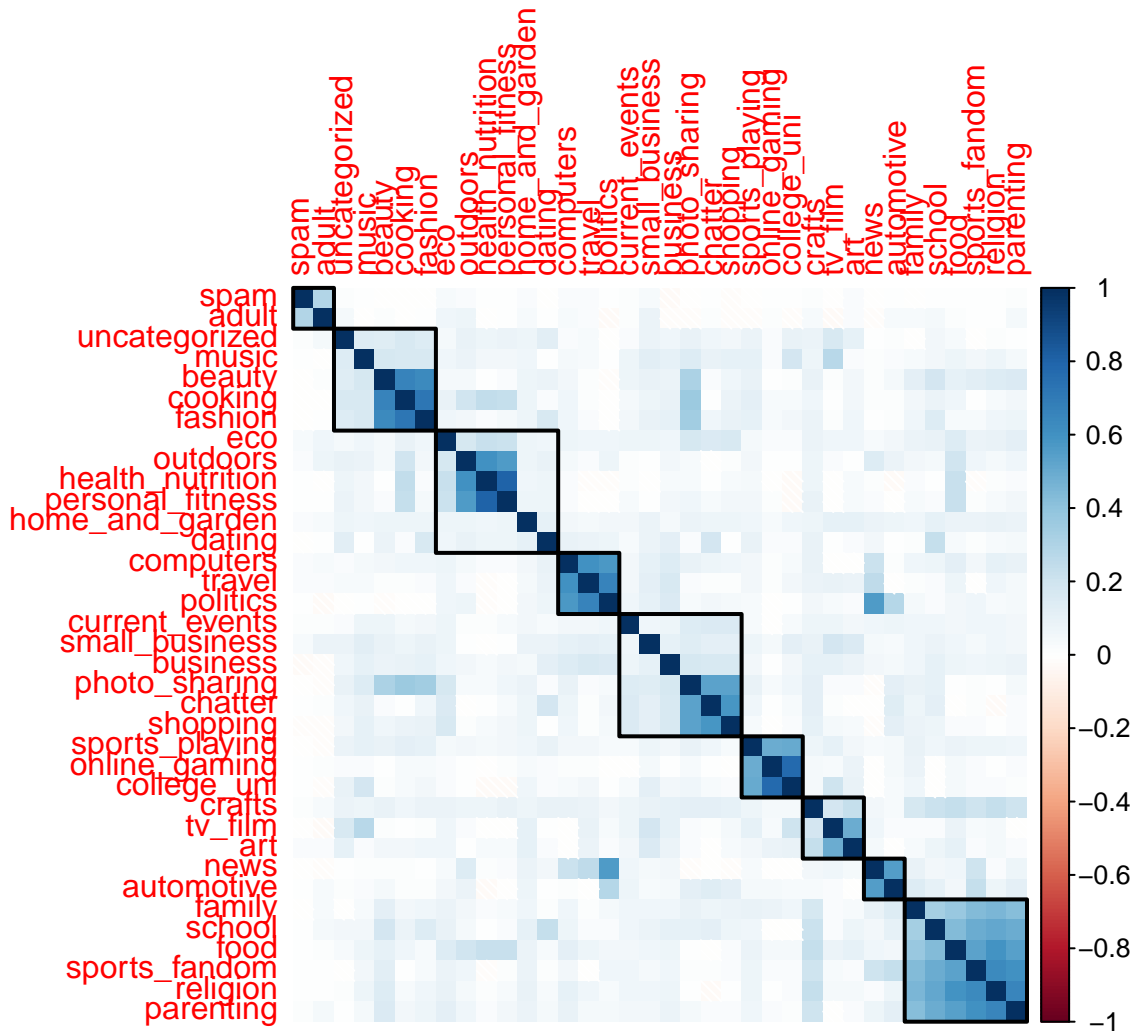
Market Segmentation

There are 9 clear market segments that appear to stand out in NutrientH20's social-media audience. The characteristics of each group is summarized in the following table.

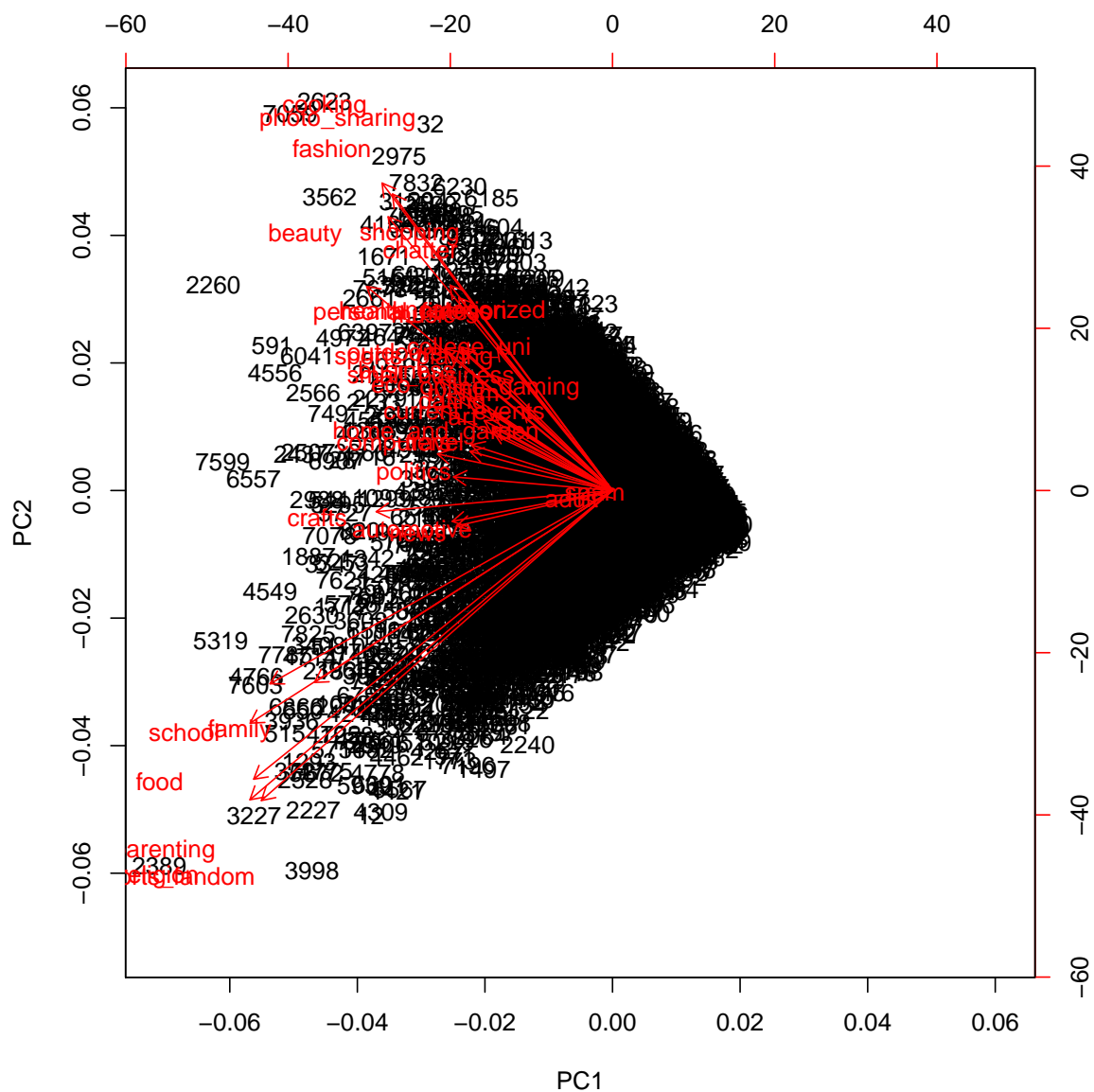
name	more talked about	less mentioned
first component	spam, adult, online_gaming	sports_fandom, food, parenting, religion
second component	chatter, photo_sharing, shopping, fasion	food, parenting, religion, sports_fandom
third component	health_nutrition, cooking, personal_fitness, beauty	travel, politics, news, computers, automotive
fourth component	college_uni, online_gaming, sports_playing, photo_sharing	health_nutrition, outdoors, politics, personal_fitness
fifth component	college_uni, online_gaming, sports_playing, tv_film	photo_sharing, shopping, beauty, chatter
sixth component	chatter, shopping, photo_sharing, eco	cooking, travel, fasion, beauty
seventh component	tv_film, art, crafts, small_business	college_uni, sports_playing, automotive, online_gaming
eighth component	travel, computer, dating, business	art, tv_film, news, automotive
nineth component	tv_film, business, music, health_nutrition	eco, small_business, spam, adult

I first explored the vairable correlation using correlation plot. I circled out 9 clusters here because I used 9 cluster in my later PCA analysis.

```
## corrplot 0.84 loaded
```



From PCA analysis, we see that photo_sharing, fashion, beauty, cooking, shopping, chatter are closely associated with each other at one end; school, family, food, parenting, sports_fandom, and dating are associated with each other at the opposite direction. The crafts and the politics are at the middle. I choose 9 components, which would explain 57.595% of variance.



Then I summarized the cluster characteristics based on the the following model summary

```
summary(pc1)
```

```
## Importance of components%s:
##               PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation    2.1186 1.69824 1.59388 1.53457 1.48027 1.36885
## Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087 0.05205
## Cumulative Proportion 0.1247 0.20479 0.27536 0.34077 0.40164 0.45369
##               PC7      PC8      PC9      PC10     PC11     PC12
## Standard deviation    1.28577 1.19277 1.15127 1.06930 1.00566 0.96785
## Proportion of Variance 0.04592 0.03952 0.03682 0.03176 0.02809 0.02602
## Cumulative Proportion 0.49961 0.53913 0.57595 0.60771 0.63580 0.66182
##               PC13     PC14     PC15     PC16     PC17     PC18
## Standard deviation    0.96131 0.94405 0.93297 0.91698 0.9020 0.85869
```

## Proportion of Variance	0.02567	0.02476	0.02418	0.02336	0.0226	0.02048
## Cumulative Proportion	0.68749	0.71225	0.73643	0.75979	0.7824	0.80287
##	PC19	PC20	PC21	PC22	PC23	PC24
## Standard deviation	0.83466	0.80544	0.75311	0.69632	0.68558	0.65317
## Proportion of Variance	0.01935	0.01802	0.01575	0.01347	0.01306	0.01185
## Cumulative Proportion	0.82222	0.84024	0.85599	0.86946	0.88252	0.89437
##	PC25	PC26	PC27	PC28	PC29	PC30
## Standard deviation	0.64881	0.63756	0.63626	0.61513	0.60167	0.59424
## Proportion of Variance	0.01169	0.01129	0.01125	0.01051	0.01006	0.00981
## Cumulative Proportion	0.90606	0.91735	0.92860	0.93911	0.94917	0.95898
##	PC31	PC32	PC33	PC34	PC35	PC36
## Standard deviation	0.58683	0.5498	0.48442	0.47576	0.43757	0.42165
## Proportion of Variance	0.00957	0.0084	0.00652	0.00629	0.00532	0.00494
## Cumulative Proportion	0.96854	0.9769	0.98346	0.98974	0.99506	1.00000