# 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_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?

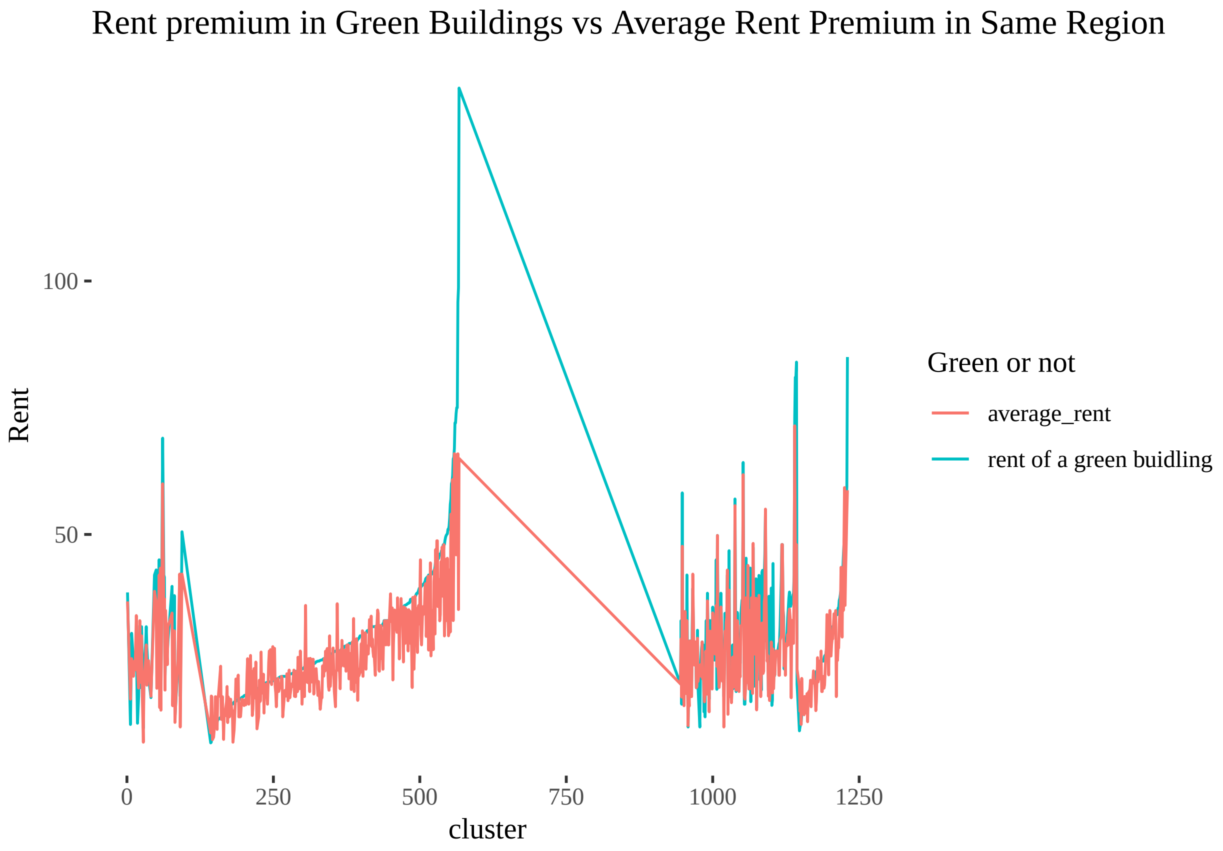
The people with no disease have a very high weight. Even though that the specificity is high, the high probabitliy of people with no disease still makes it a large number compared to the True positive rate times the probability of having the disease. Thus, it is hard to test the effectiveness of medical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

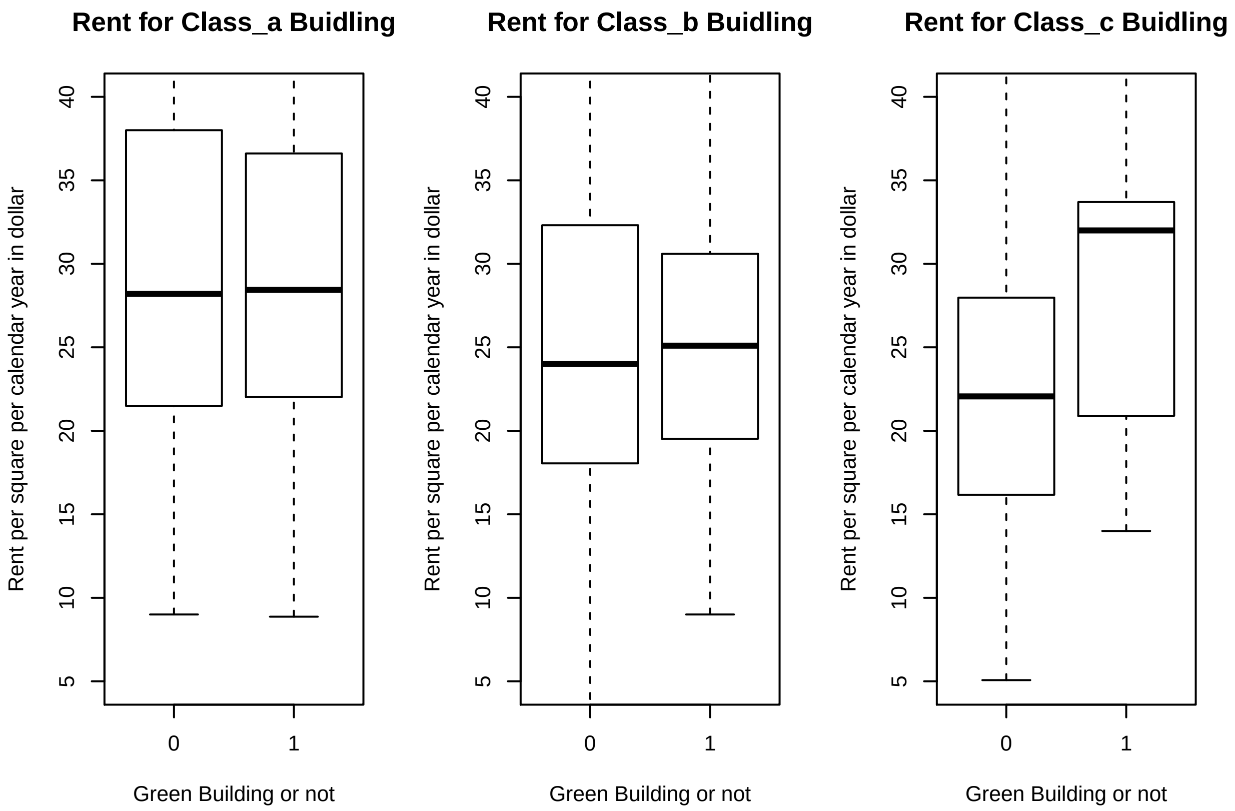
```
## [1] 0.1988824
```

## Exploratory analysis: green buildings

An Austin real-estate developer is interested in building a new 15-story mixed-use building on East Cesar Chaves, 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 enocomicly worth to invest in a green building rather than an ordinary building.



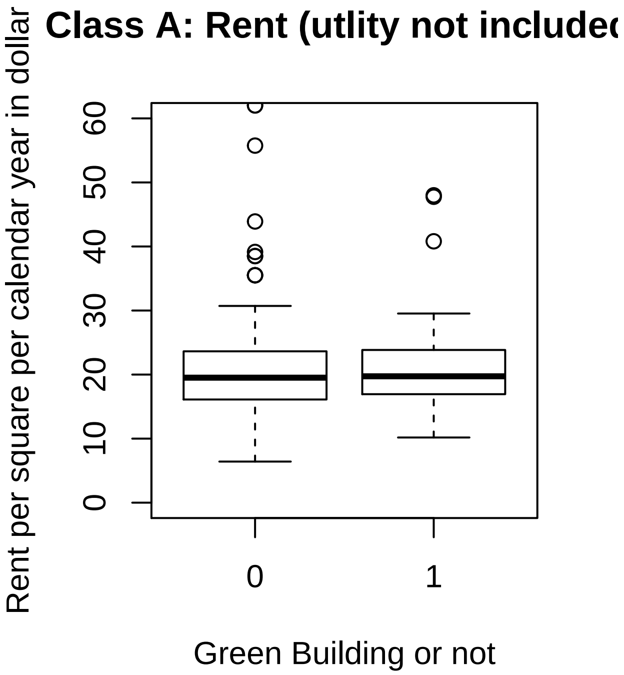
We know that a green building is associated with social responsibility and ecological awareness, and might therefore command a premium from potential tenants. We can see this relationship through the plot above: rent rate of green buidlings are higher than average price of the same region most of the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

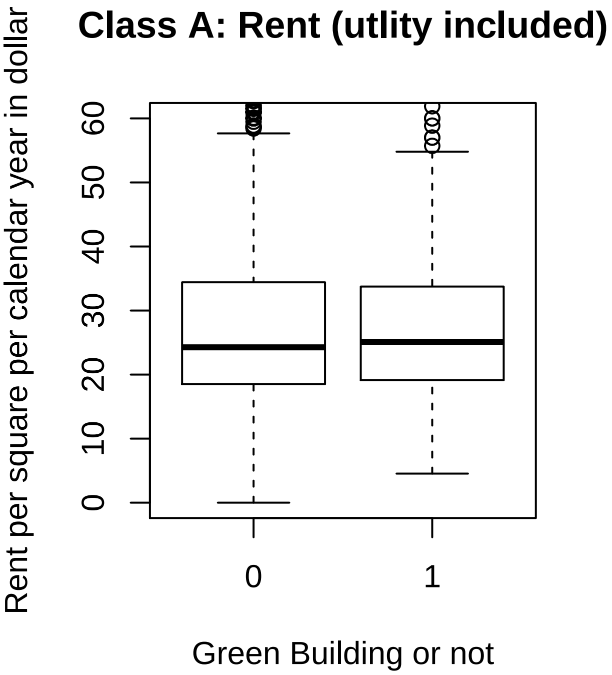


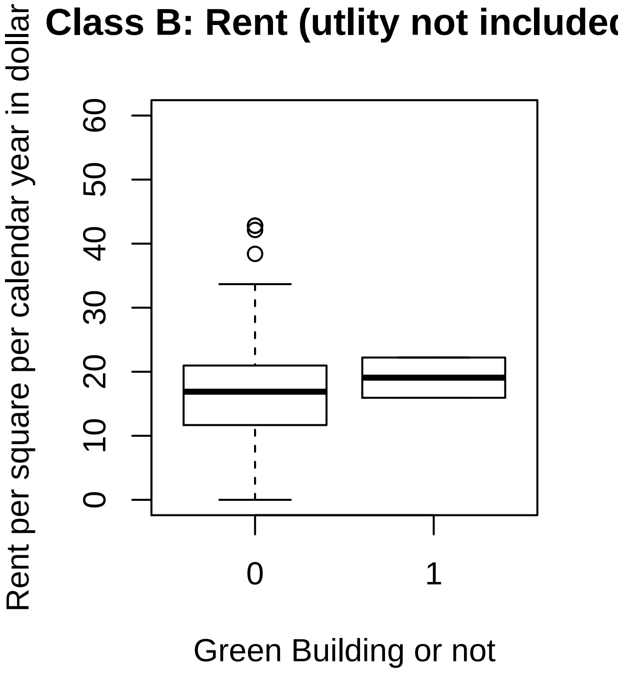
By looking at the rent under three classes of buildings: class a, class b, and class c, the rent premium decreases as the class of building decreases. It is surprising that a class c green building has the highest rate. However, a high rent premium is usually associated with low occupancy rate. Thus,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

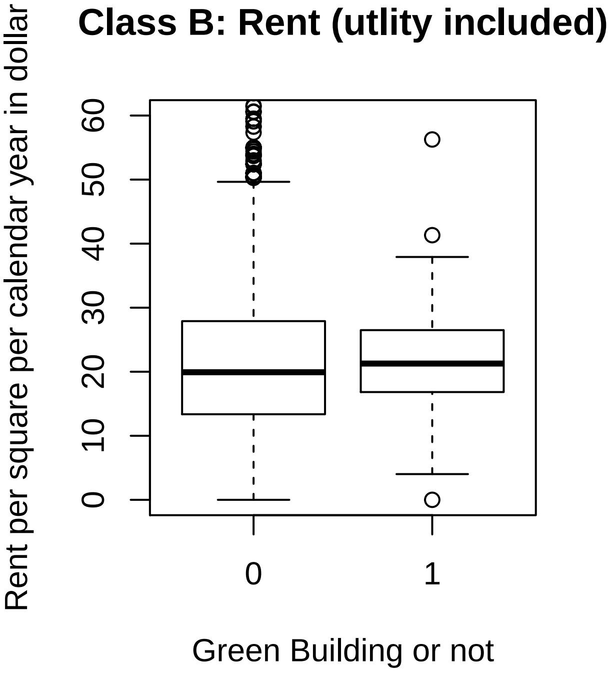
```
data['rent_income_per_square'] = data['Rent'] * data['leasing_rate'] / 100
```

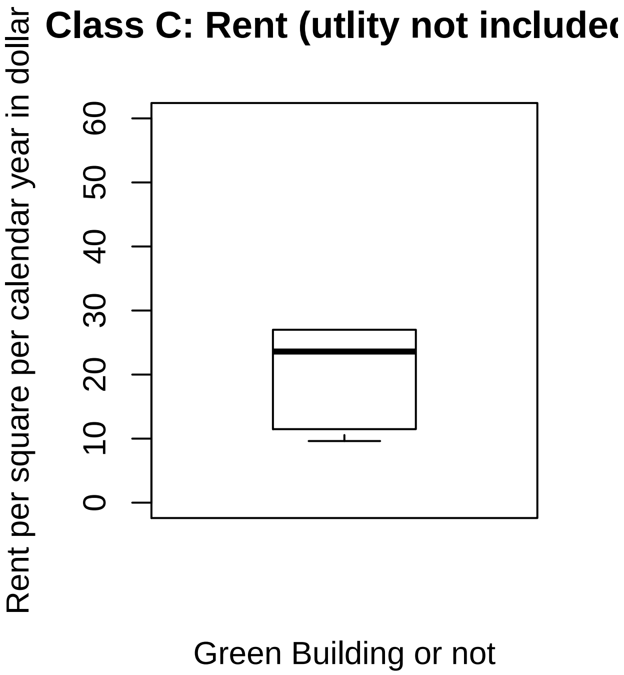
Because some rent premiums do not include utility bills, I separated the data into two groups by the varaible 'net' and compared the rent income of a green building with the rent income of an ordinary building under three classes. The median rent income per square of a green building is much higher than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

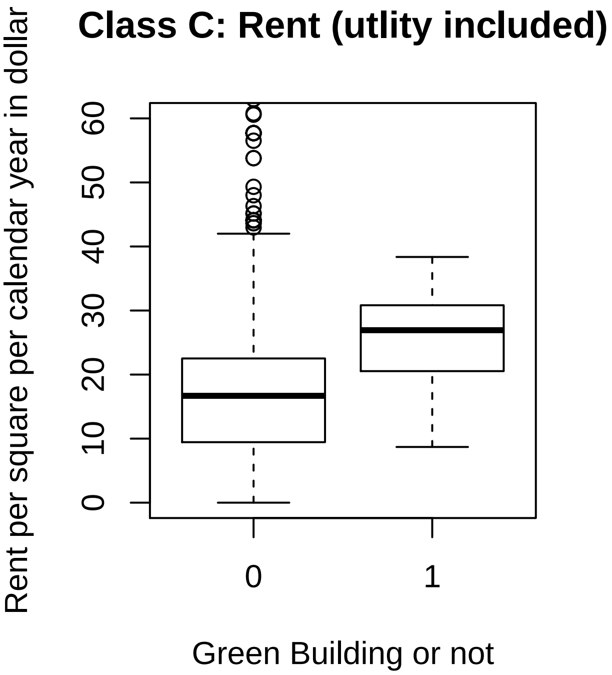

Class A: Rent (utlity not includec
Rent per square per calendar year in dollar
0
10
20
30
40
50
60
0
1
Green Building or not


Class A: Rent (utlity included)
Rent per square per calendar year in dollar
0
10
20
30
40
50
60
0
1
Green Building or not

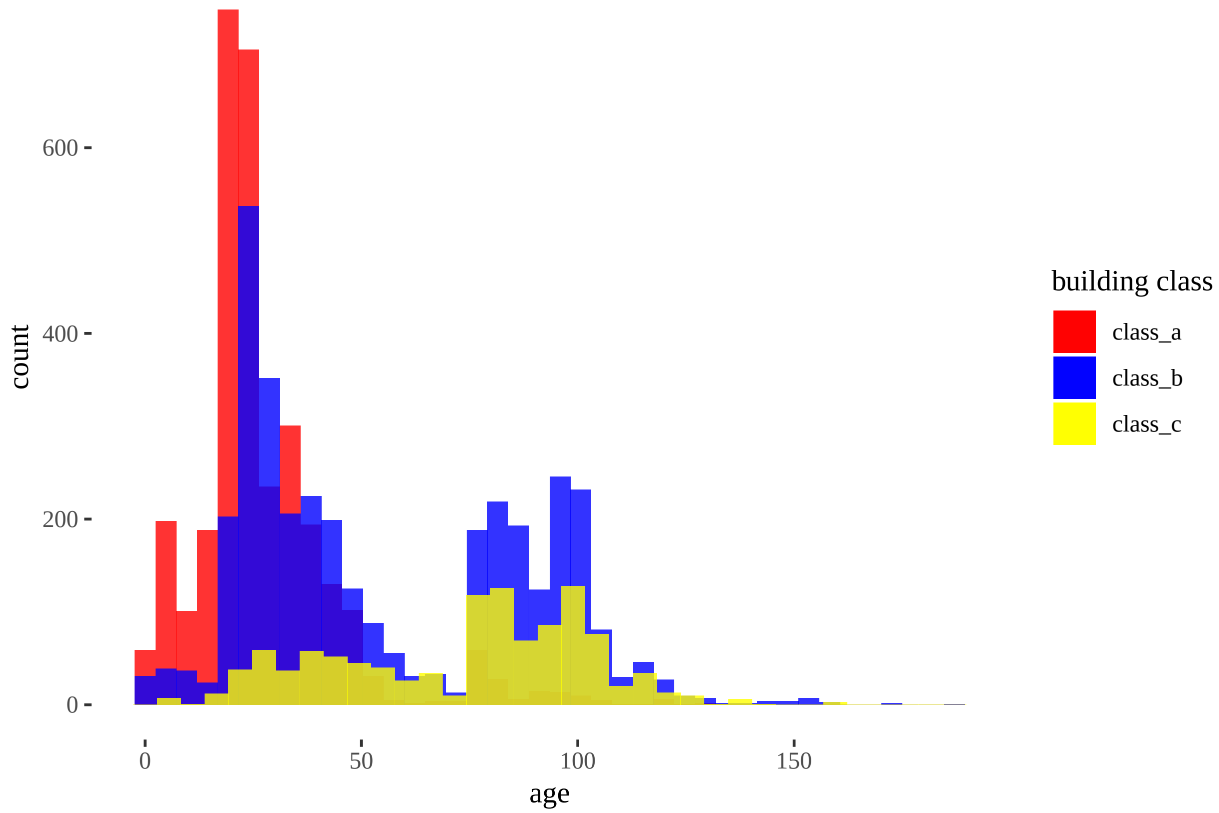

Class B: Rent (utlity not includec
Rent per square per calendar year in dollar
0
10
20
30
40
50
60
0
1
Green Building or not


Class B: Rent (utlity included)
Rent per square per calendar year in dollar
0
10
20
30
40
50
60
0
1
Green Building or not





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 $4.9325 \times 10^6$. If the developer decides to build an ordinary building, the total income from tenants per calendar year would be $4.87872 \times 10^6$. The time to recover the total cost would be 21.2873796 years for green building but 20.4971796 years for non-green buidling 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 assocaited with green buildings. Green buildings would have a longer buidling 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 biuld 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 domesitc 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 aggresive 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] 0.012452649 0.009158813 0.005221217 0.040254408 0.021163780
```

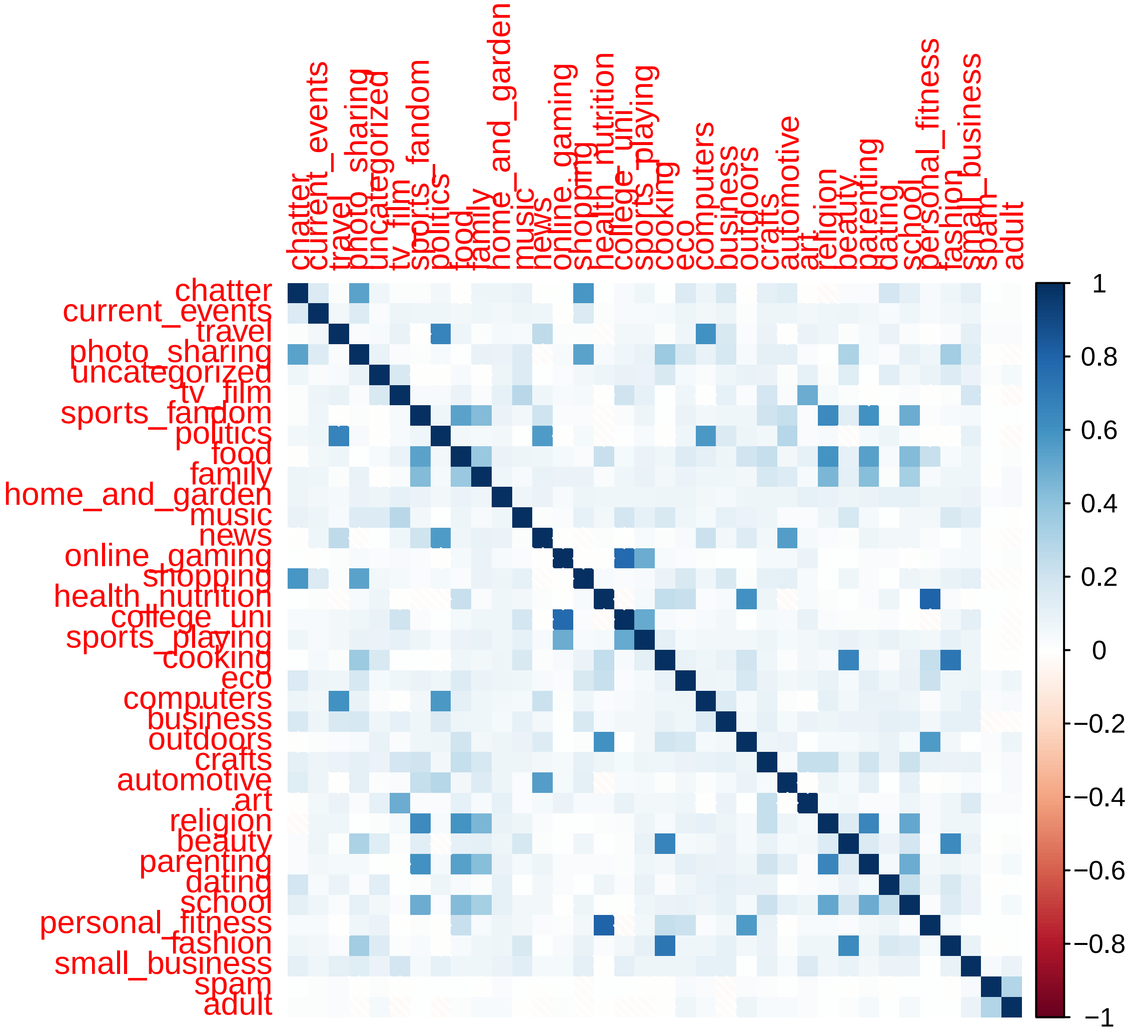
| portfolio | 5% VaR | Mean increase in wealth |
|---|---:|---:|
| the even split | -6206.510 | 995.889 |
| safer split | -3134.351 | 596.517 |
| aggressive split | -11718.592 | 1388.084 |

# Market Segmentation

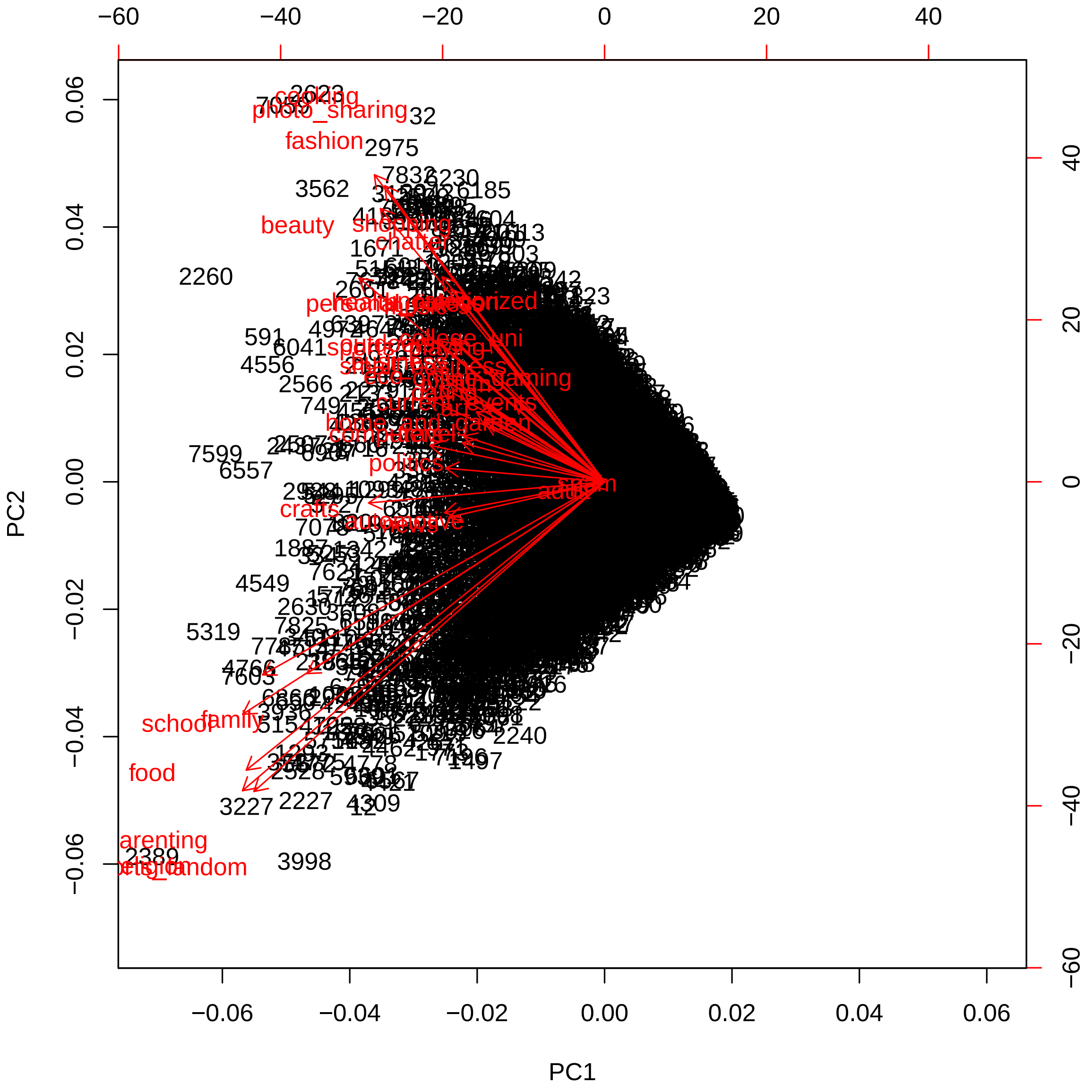
There are two clear market segments that appear to stand out in NutrientH20's social-media audience. One group of audience are associated with labels such as photo_sharing, fashion, beauty, cooking, shopping, and chatter; Another group of audience are associated with labels such like school, family, food, parenting, sports_fandom. This conclusion is drawn based on a correlation plot, PCA analysis, and K means.

First, we know from the correlation plot below that some labels are correlated. For example, plotics and chatter; college_uni and online_gaming; personal_fitness and health_nutrition.

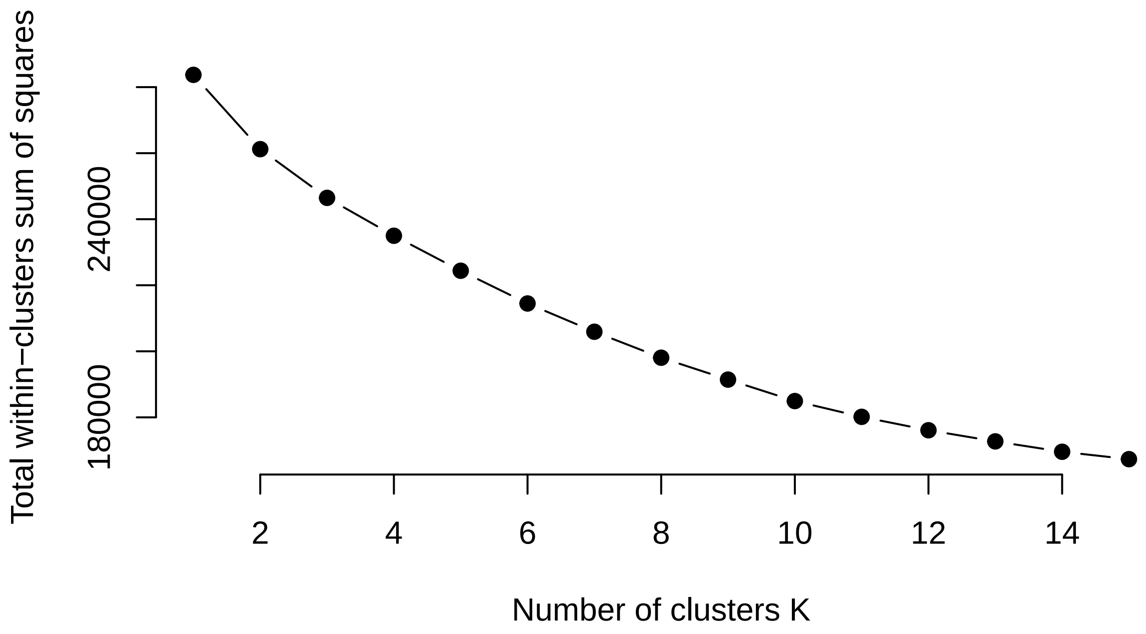
```
## corrplot 0.84 loaded
```



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.

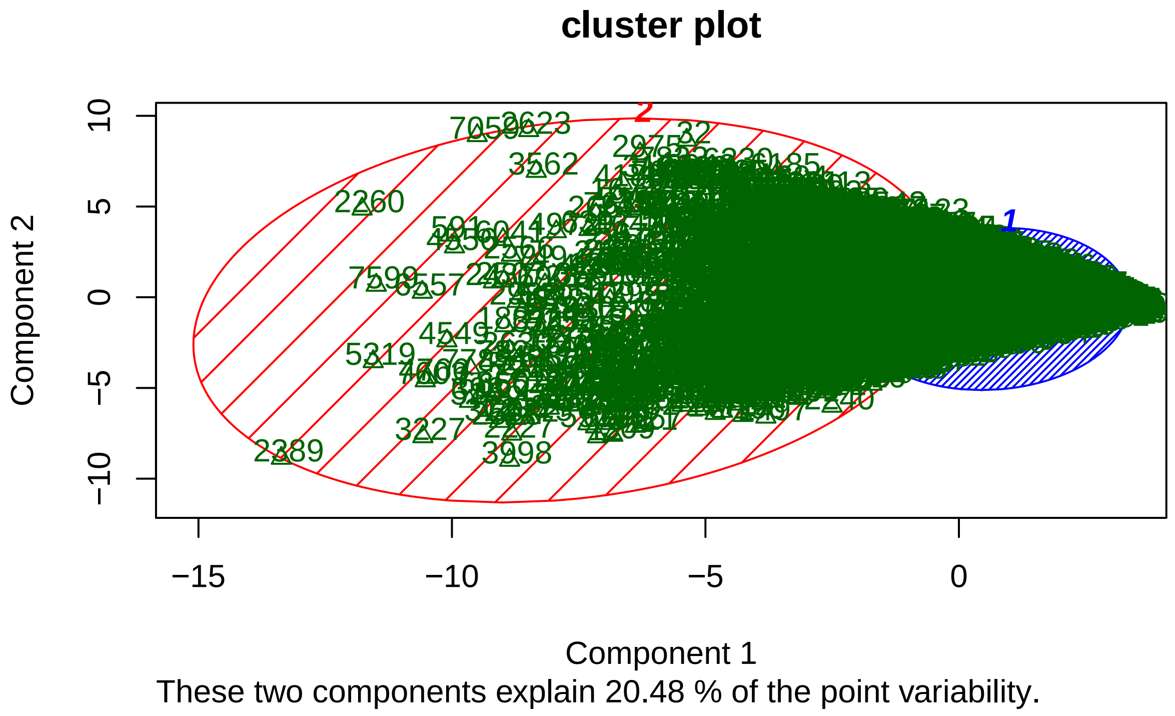


I also did k-means clustering. After I scaled and centered the data, I use the elbow method to select the optimal cluster size, which would be two.

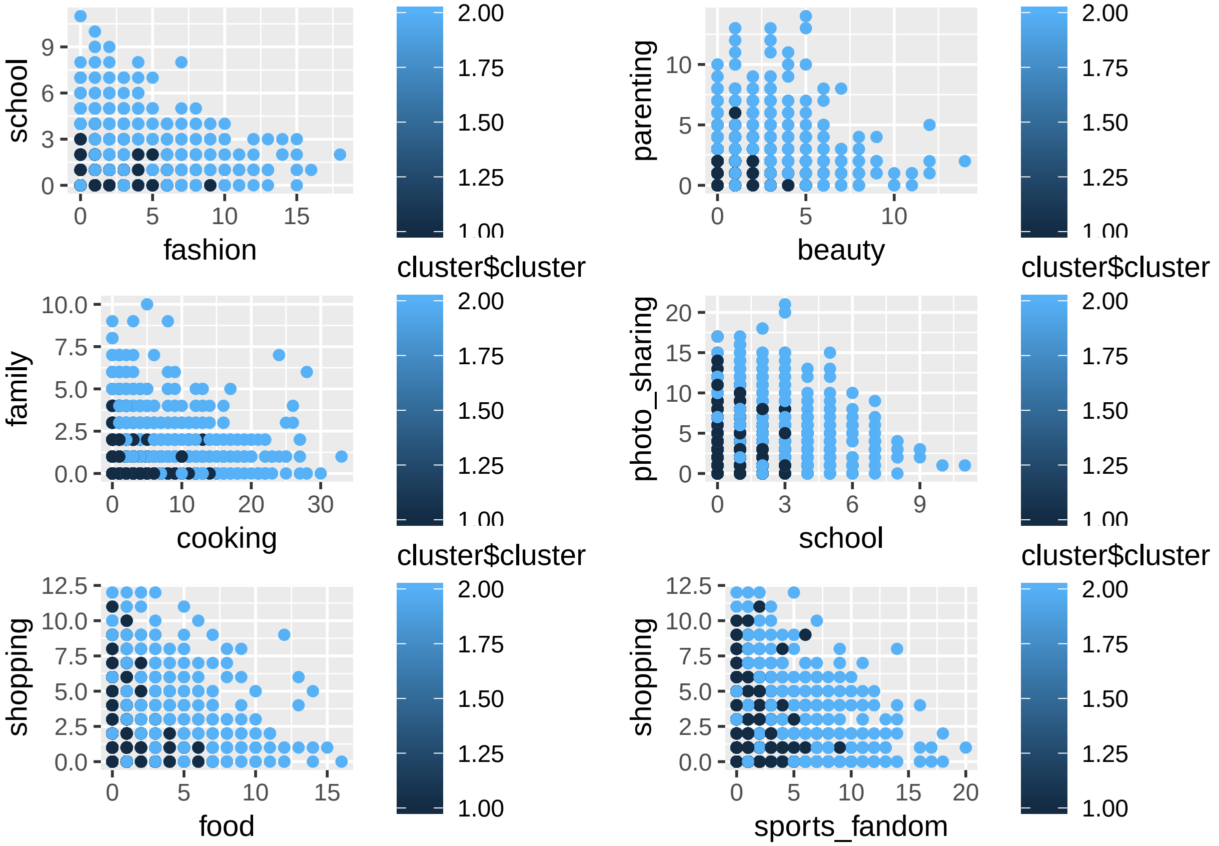


```
cluster = kmeans(market_scaled, centers=2, nstart=20)
```

The cluster plot is shown as below.



In the end, I validated my findings with the following plots. There are clear two clusters in these plots, which proves my findings in the beginning. One group of audience are associated with labels such as photo_sharing, fashion, beauty, cooking, shopping, and chatter; Another group of audience are associated with labels such like school, family, food, parenting, sports_fandom, and dating.


school
fashion
cluster$cluster
parenting
beauty
cluster$cluster
family
cooking
cluster$cluster
photo_sharing
school
cluster$cluster
shopping
food
shopping
sports_fandom