STA 380-Part 2: Exercises 1

Shuyuan Sun,Mengying Yu,Cuiting Zhong,Lining Jiang

August 5, 2018

Probability practice

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?

Given:
$P(yes) = 0.65$
$P(yes|RC) = 0.5$
$P(RC) = 0.3$
$P(TC) = 0.7$

And:
$P(yes) = P(yes, RC) + P(yes, TC) = P(yes|RC) * P(RC) + P(yes|TC) * P(TC)$
$0.65 = 0.3 * 0.5 + P(yes|TC) * 0.7$

Therefore:
$P(yes|TC) = (0.65 - 0.3 * 0.5)/0.7 = 0.714$
Among those people who are truthful clickers, 71.4% answered yes.

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?

Name:
$'yes' =' actuallyhasdisease'$
$'no' =' actuallynodiease'$
$'+' =' testpositive'$
$'-' =' testnegative'$

Given:
$sensitivity = P(+|yes) = 0.993$
$specificity = P(-|no) = 0.9999$
i.e. $P(+|no) = 0.0001$
$P(yes) = 0.000025$

And:
$P(+) = P(+, yes) + P(+, no) = P(+|yes) * P(yes) + P(+|no) * P(no) = 0.993 * 0.000025 + 0.0001 * (1 - 0.000025) = 0.0001248$

Therefore:
$P(yes|+) = P(yes) * P(+|yes)/P(+) = 0.000025 * 0.993/0.0001248 = 0.1989$
Suppose someone tests positive. the probability that they have the disease is only 19.89%.
This test should not be implemented universally because the probabilities of having this disease or not are extremely imbalanced. Only having high sensitivity and specificity is not enough, and this will cause a huge amount of people who actually don't have the disease to spend unnecessarily on treatments.

Exploratory analysis: green buildings

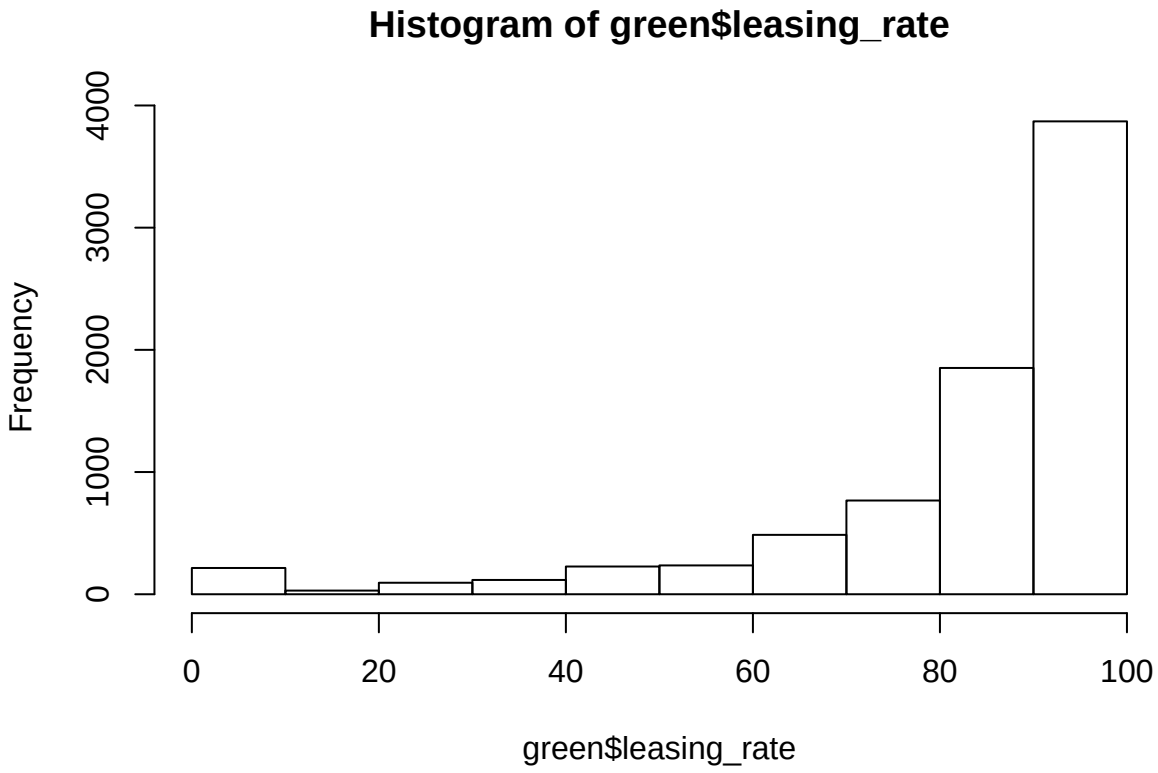
Do you agree with the conclusions of her on-staff stats guru? If so, point to evidence supporting his case. If not, explain specifically where and why the analysis goes wrong, and how it can be improved. (For example, do you see the possibility of confounding variables for the relationship between rent and green status?)

Note: this is intended mainly as an exercise in visual and numerical story-telling. Tell your story primarily in plots, and while you can run a regression model if you want, that's not the goal here. Keep it concise.

1. Discussion about guru's data cleaning

Read in data and divided into two groups (leasing rate<=10%/>10%)

Histogram of leasing rate and fivestats of the two groups



First from this histogram we can see that there are relatively more buildings with leasing rate below 10%. There may be some reasons causing these buildings to have lower occupancy. So we examined the summaries if the two groups and found some variables with different features in different leasing rate groups. For example, below are the fivestats and corresponding boxplots for the two groups:

```
## [1] "Age"
```

```
##   green_copy$occupancy10 min Q1 median Q3 max    mean       sd    n
## 1                   <10%   0 28     57 85 118 54.4186 32.67443  215
## 2                  >=10%   0 23     34 79 187 47.0431 32.15998 7679
##   missing
## 1       0
## 2       0
```

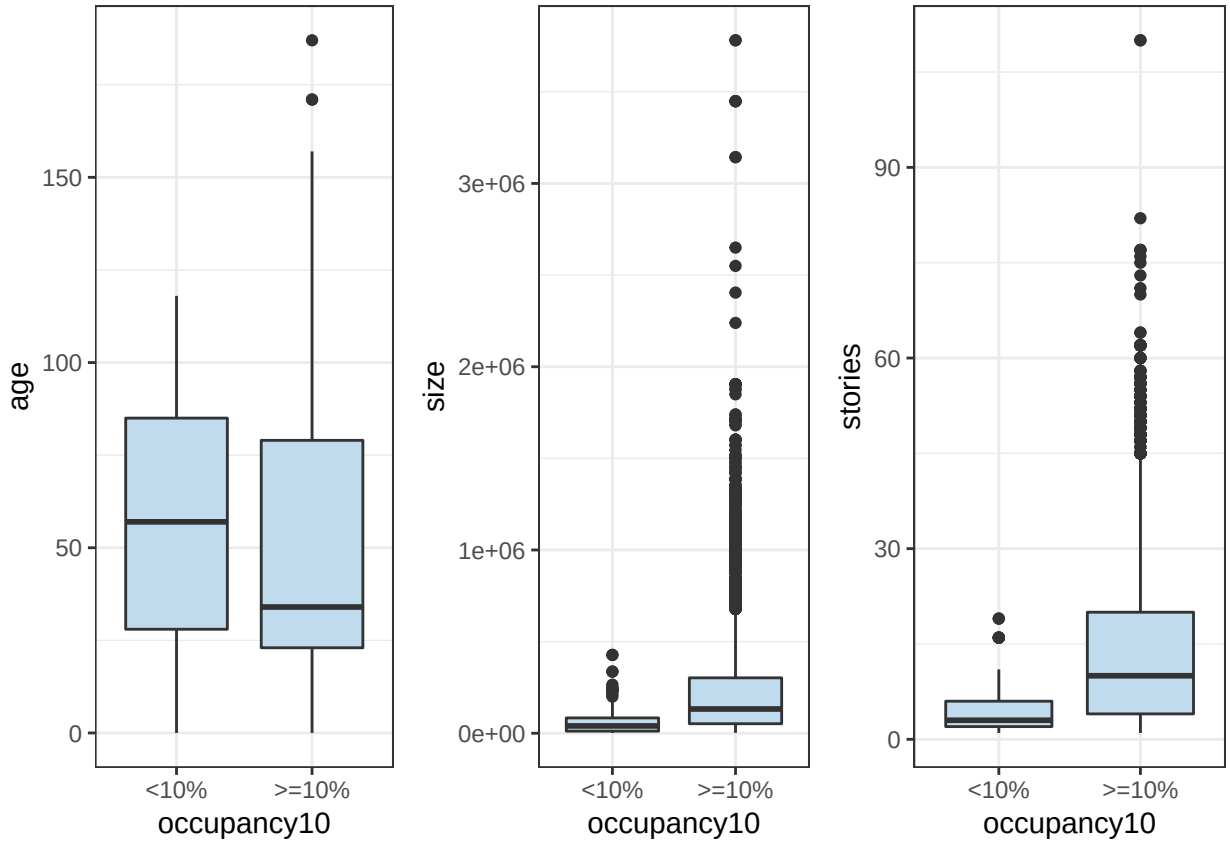
```
## [1] "Size"
```

```
##   green_copy$occupancy10  min    Q1 median       Q3     max      mean
## 1                   <10% 1624 11661  40000  83769.5  427383  62209.12
## 2                  >=10% 2378 52000 132417 302375.0 3781045 239465.48
##          sd    n missing
## 1  74652.03  215       0
## 2 299989.78 7679       0
```

```
## [1] "Stories"
```

```
##   green_copy$occupancy10 min Q1 median Q3 max      mean       sd    n
## 1                   <10%   1  2      3  6  19  4.818605  3.78507  215
## 2                  >=10%   1  4     10 20 110 13.829926 12.35268 7679
##   missing
```

```
## 1        0
## 2        0
```




```
## [1] "Fraction of green buildings in group with leasing rate<=10% is"
## [1] 0.004651163
## [1] "Fraction of green buildings in group with leasing rate>10% is"
## [1] 0.0890741
```

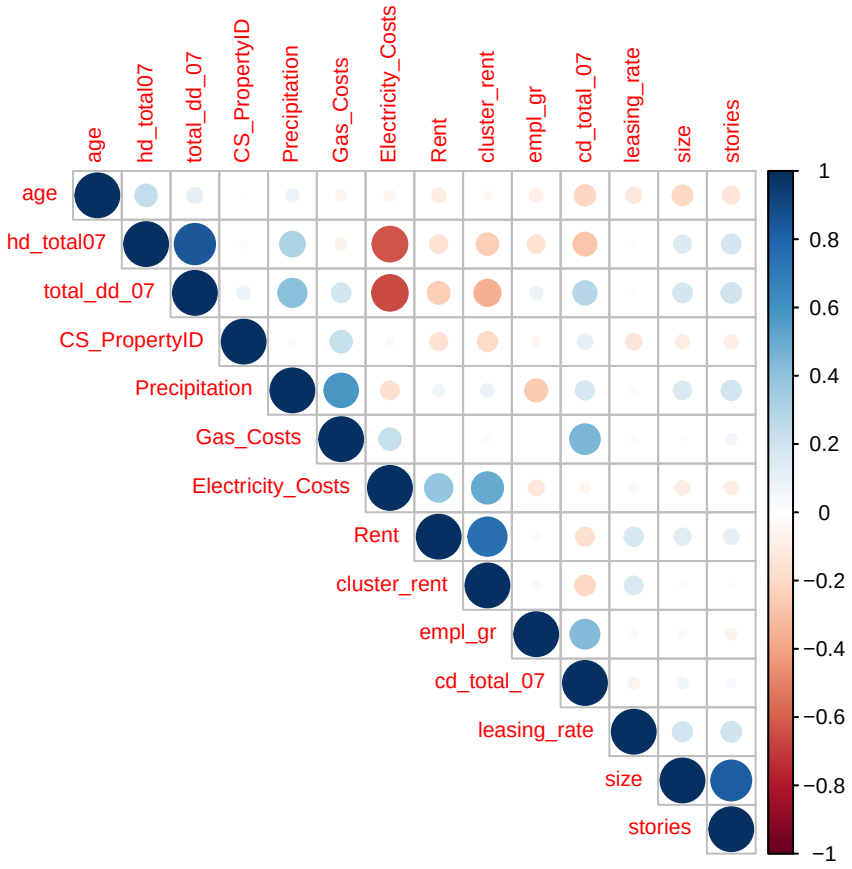
Based on the plots above, we see that buildings which have lower leasing rates are much likely to be old building, and their size and stories are low. We also noticed that the fraction of green buildings in group<=10% is only 0.47%. However it is 8.9% in group>10%. If we include these data in our dataset, they may cause a bias for our analysis. So, we removed the ovbservations which leasing rate is lower than 10%.

2. Discussion about rent and green status

According to the guru, calculating the extra revenue by multiplying the square feet with extra median rent figure means that the green buildings all have a rent of 27.60 USD, and the non-green ones all are $25 per square foot per year. However, not to mention using median as a substitute, the rent itself is decided not only by the green-status. There are other confounding variables influencing the rent and the green status.
"Confounding variables are an outside influence that change the effect of a dependent and independent variable. Confounding variables can ruin an experiment and produce useless results. They suggest that there are correlations when there really are not."
We should find and study all the confounding variables to obtain more accurate results. Exploratory analysis could help us find the relationships between variables.

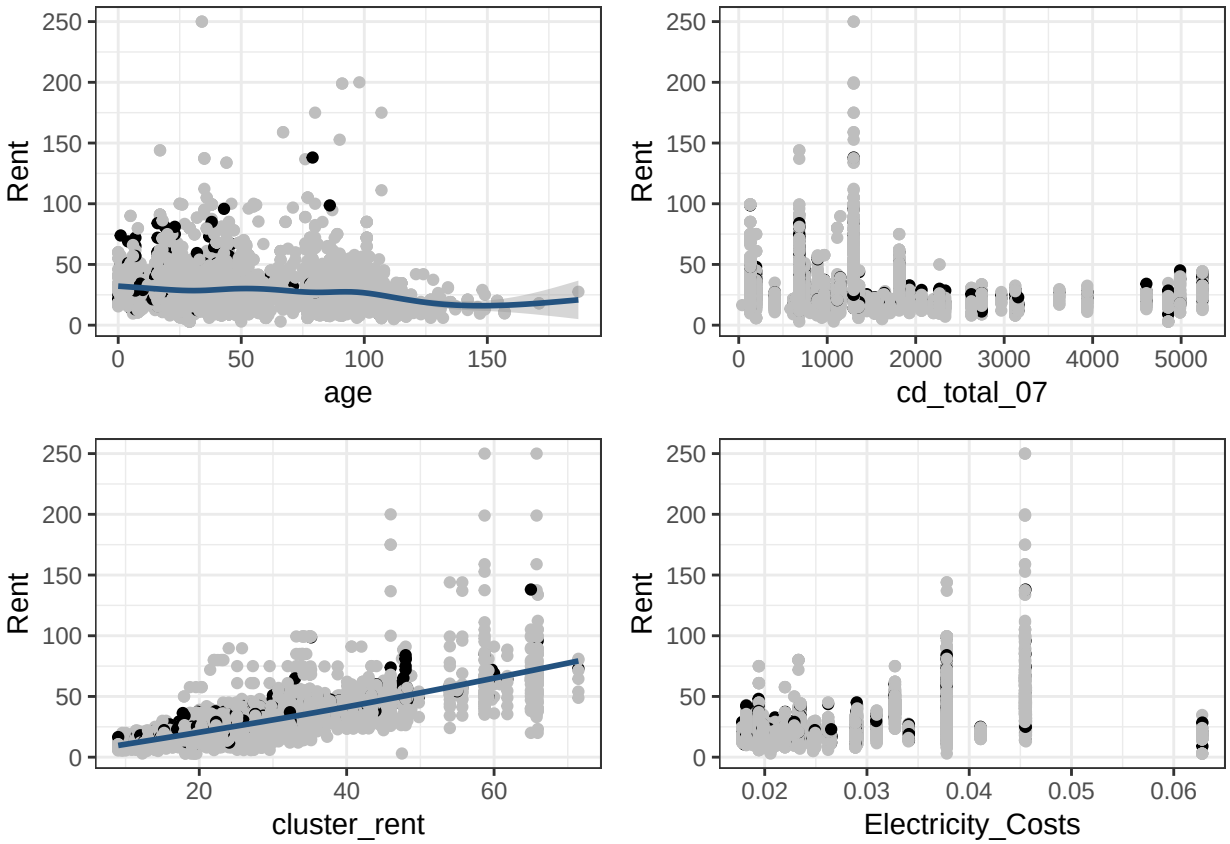
Correlation plot

First let's look at the correlations between numeric variables.

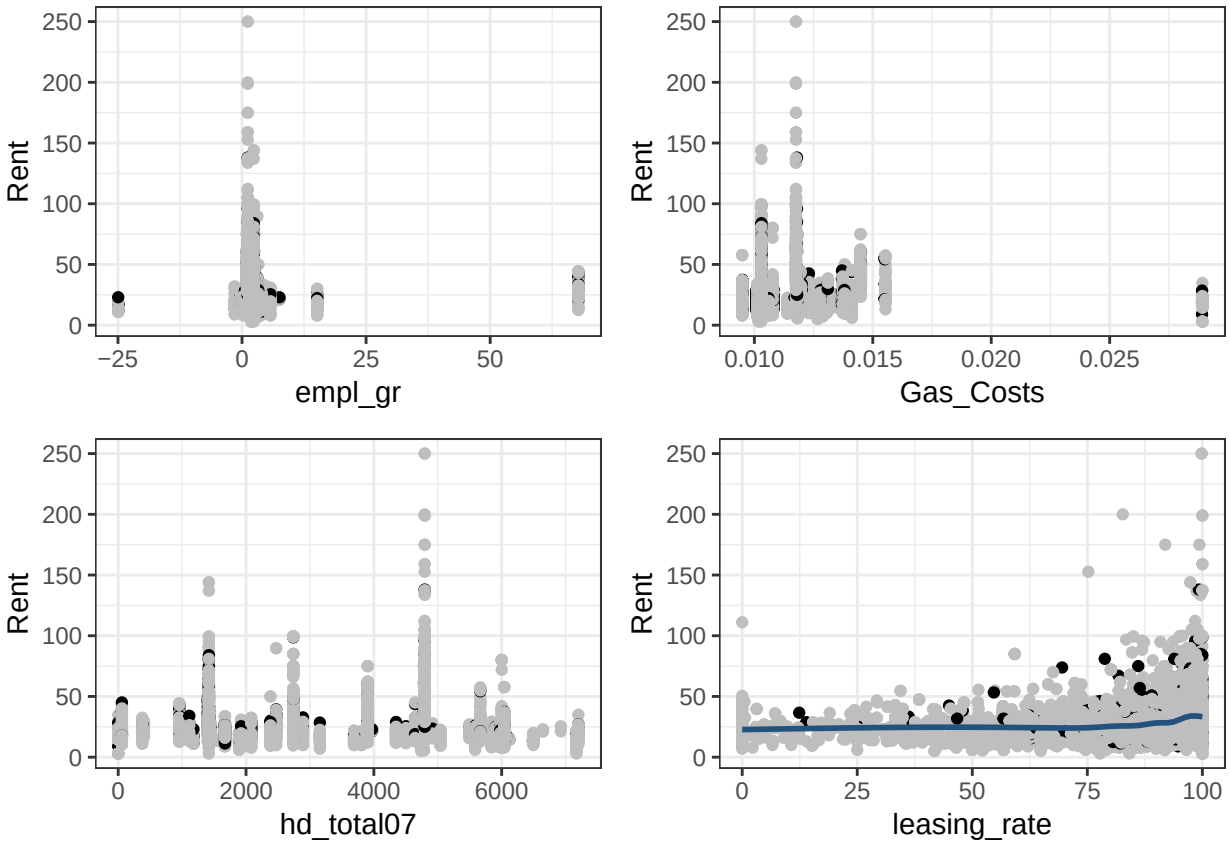


From the correlation plot we can see roughly that there are many variables correlated with the target–Rent.

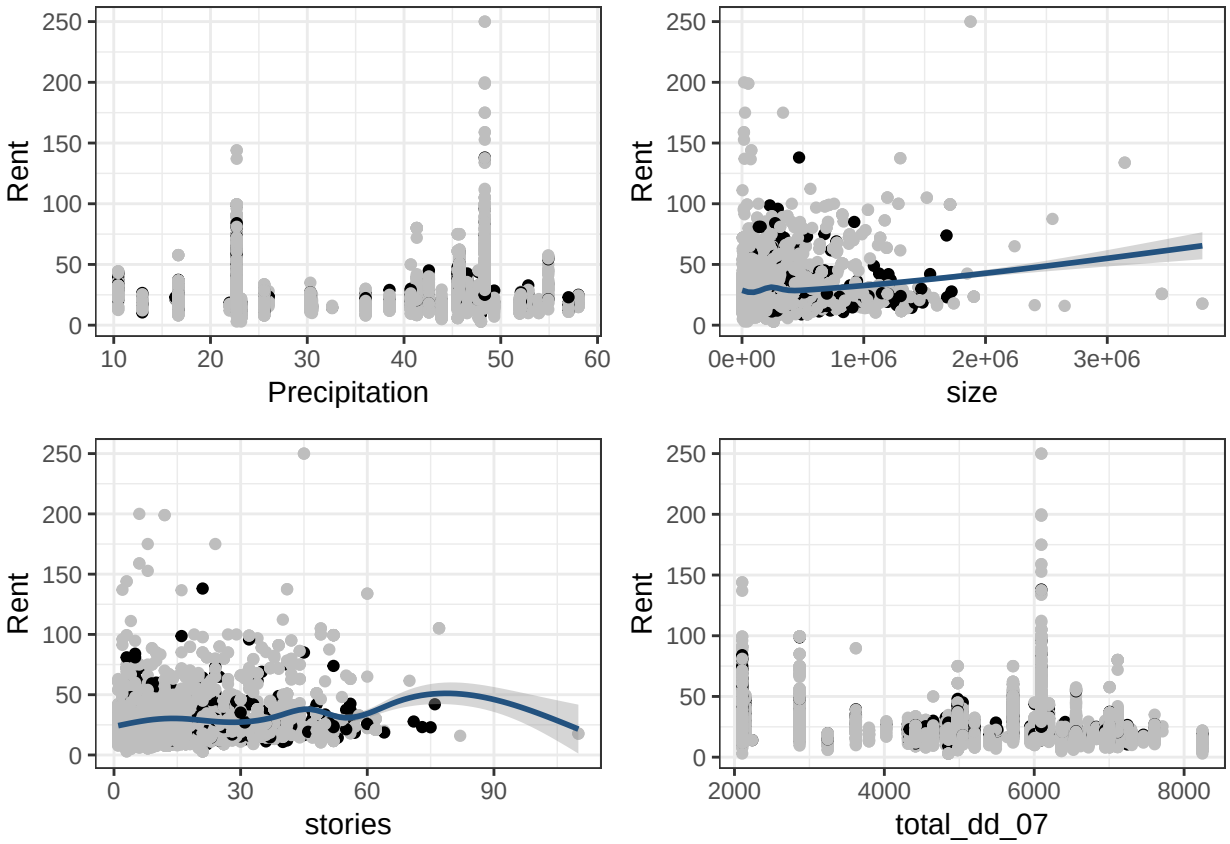
Numerical variables



(Black dots represent those with a green rating. Grey dots are those not.) As seen from the scatterplot, those buildings with a longer age tend to have a lower rent, which is fairly obvious. Those with fewer cooling degree days tend to have a higher rent maybe because fewer cooling degree days means the temperature is not extreme and requires less cooling. Because the cluster rent is calculated by individual rents in the cluster, so there is a linear relationship. Moreover, most of the buildings are not on a "net contract" basis, so with utility costs included, the higher the electricity costs, the higher the rent.



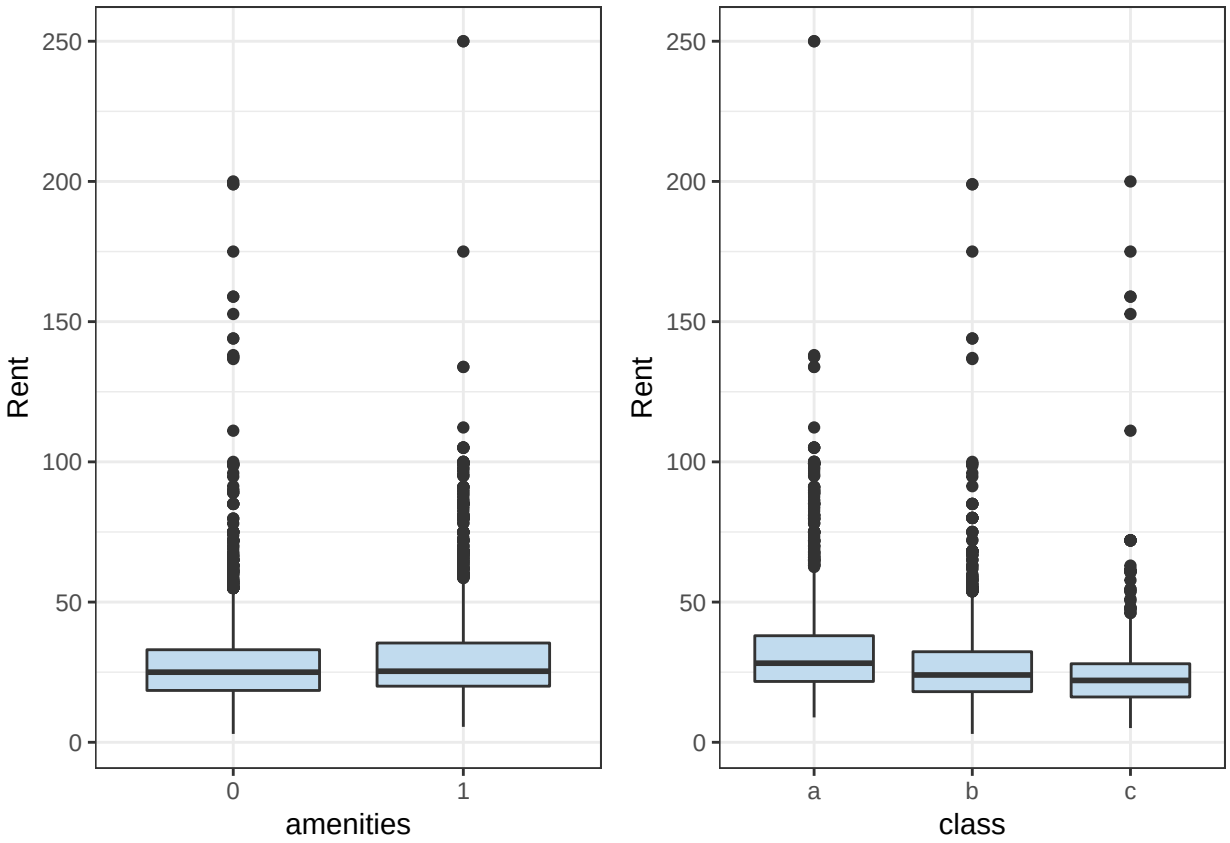
Accordingly, gas costs are similar to electricity costs and heating degree days to cooling degree days. While the leasing rate(occupancy) increases, the rent will also increase.



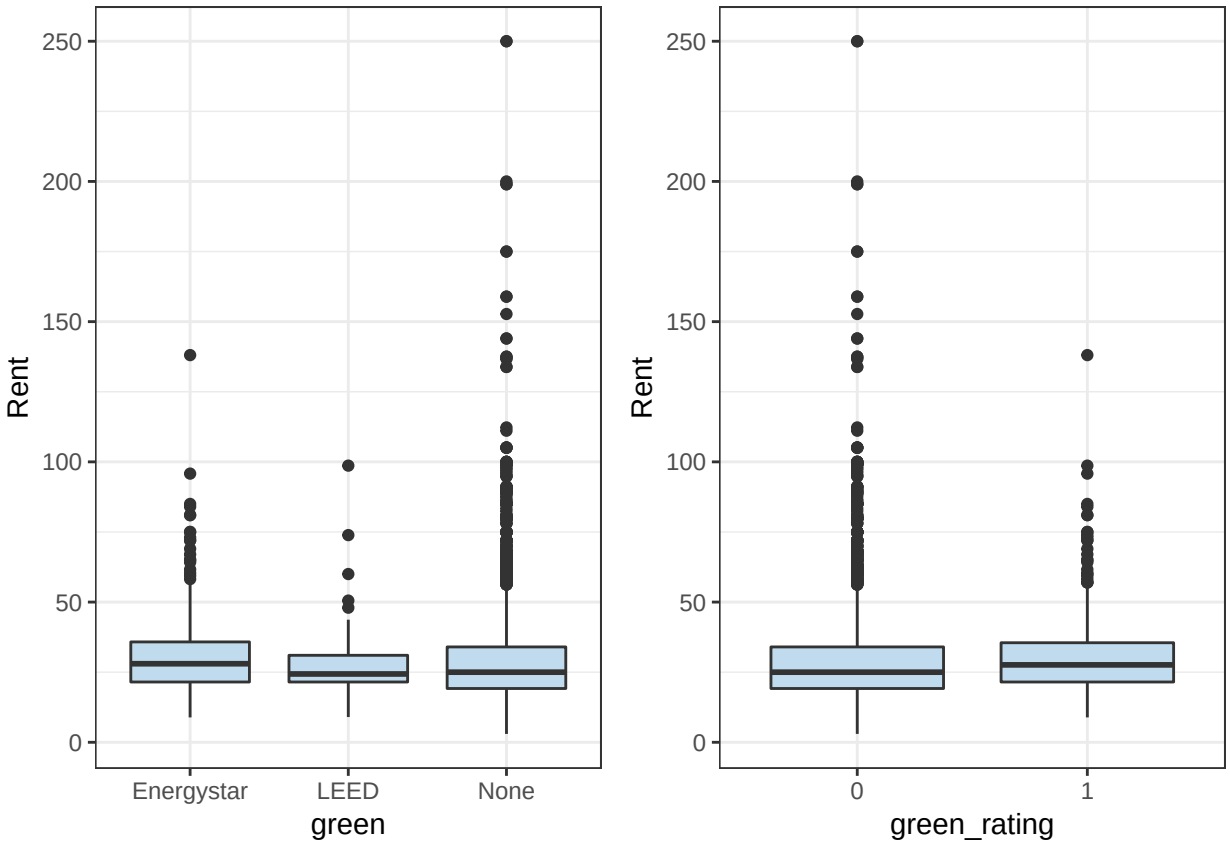
As seen, when the size and stories of a building increase, the rent will usually increase.

Categorical variables

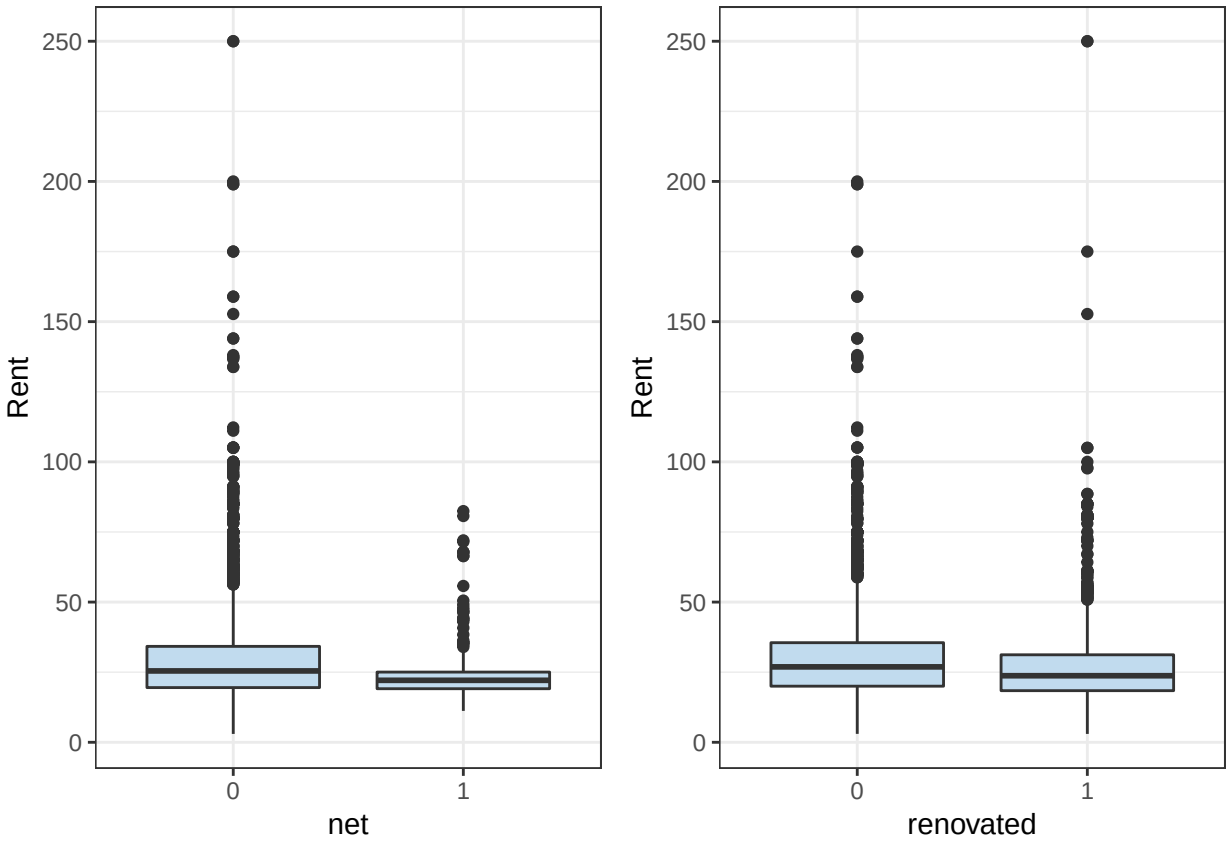
Then let's examine categorical variables.



It seems that whether at least one of the amenities is available has not much impact on the rent. But from the right plot we can see that higher-quality properties(like Class A) will charge higher rents.



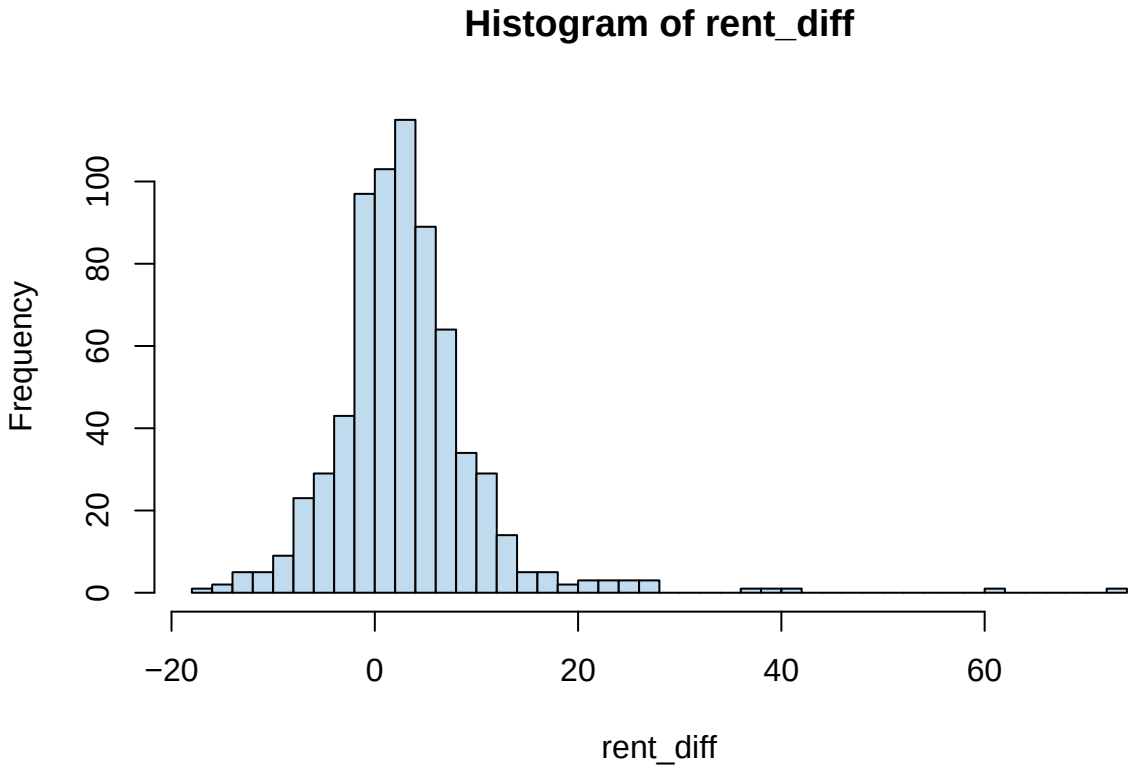
From the right plot, those buildings with a green rating will have higher rents. But plot on the left indicates that it is mainly the "Energystar" rating that raises the rent.



With a net contract basis, the rent spread is relatively small, and the rent on average is lower(obviously because no utility cost is included). Those buildings undergone substantial renovations charge a lower rent. Maybe this is because the fact of going through renovations indicates a poorer quality or a longer age of the building.

To sum up, there are many other variables influencing on our target Rent. Thus, only taking into account the green status is not enough for estimating the rent of a building.

3. The rent difference between green buildings and non-green buildings

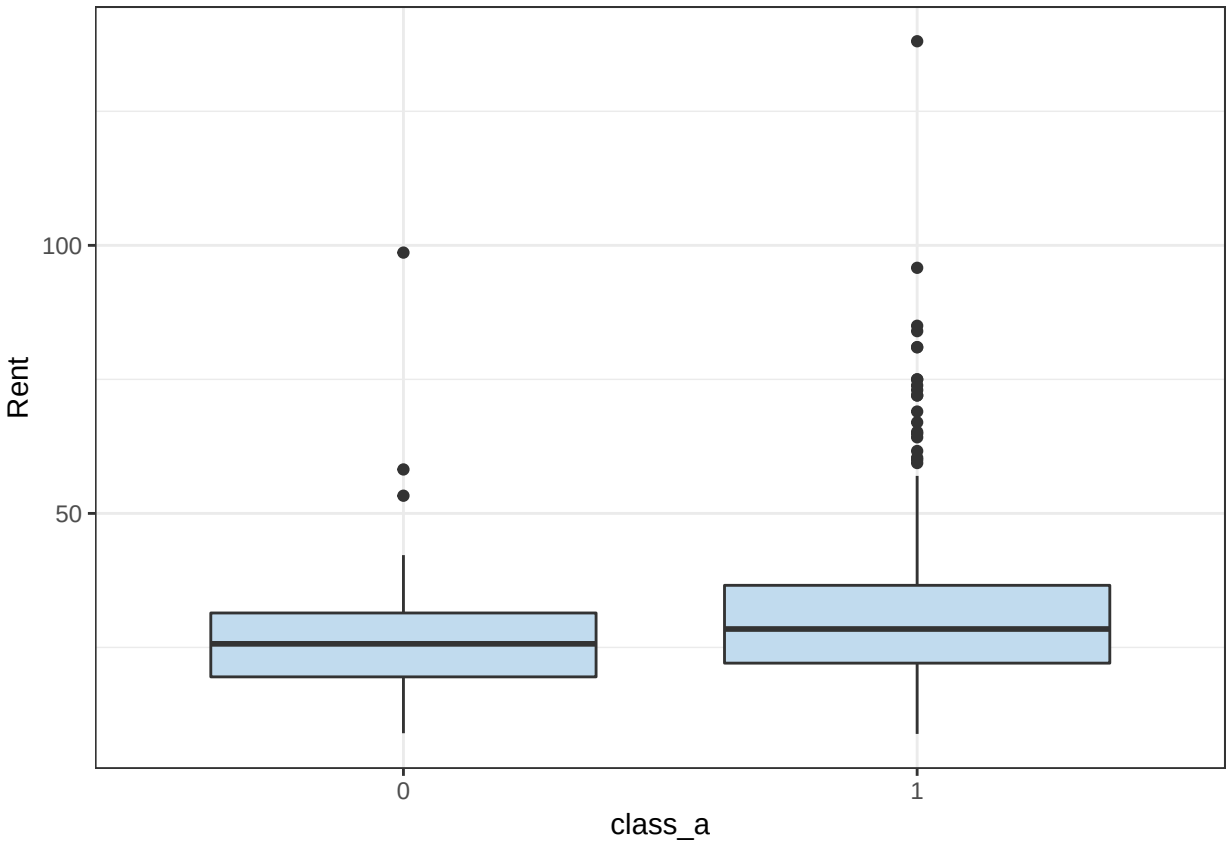


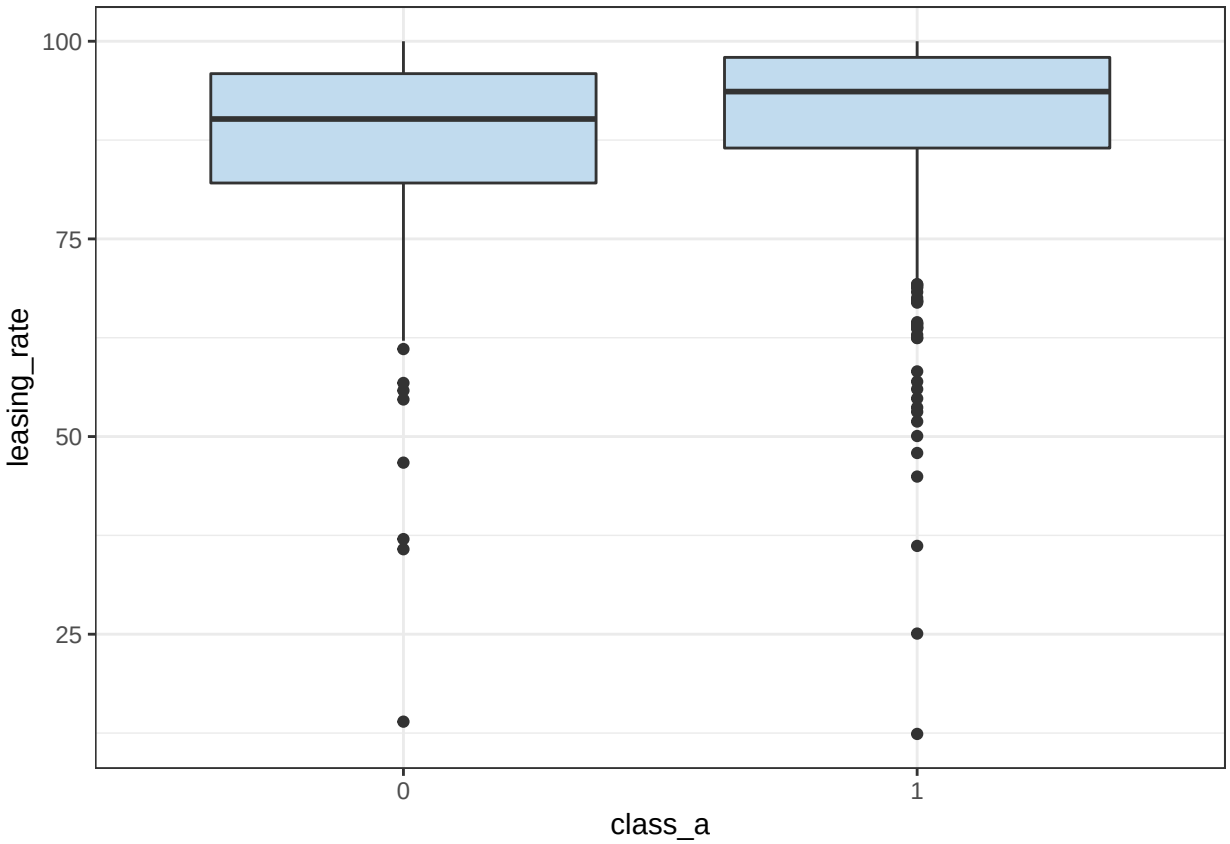

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -16.820  -0.250   2.500   3.091   5.930  72.237
```

```
## [1] "confidence interval"
```

```
## [1] 2.54239 3.64026
```

Since the dataset was seperated in small clusters and some variables such as the growth rate in employment and precipitation are same, it is more convincible to use the difference of rent between green buildings and non-green buildings which are in a same cluster to decide the type of our building. We calculated the difference of rent in each cluster and then found that the mean is about 3.1, the 95% confident interval is (2.54, 3.64). It is a strong evidence to show that green building has a higher rent price than non-green building and we can assume that we can earn 3.1dollars more per square foot per calendar year.


Rent
100
50
0
1
class_a



Based on the plots we showed above, we found that people who prefer to green buildings much likely to rent the green buildings which are Class A buildings, i.e. the highest-quality properties in a given market. Moreoever, the rent price of green buildings which are Class A buildings is also higher than other green buildings.

4. Other aspects to consider

The rent should be estimated using all known information. For example, "a **new**, **15-story**, **mixed-use** building on East Cesar Chavez, just across I-35 from **downtown**". Instead of only considering the green status, all the information could be plugged into a regression model to estimate the rent.
Moreover, to calculate the economic value of a project, we should also take into consderation the value of money. What the guru says *"Based on the extra revenue we would make, we would recuperate these costs in $5000000/650000 = 7.7 years. Even if our occupancy rate were only 90%, we would still recuperate the costs in a little over 8 years"* does not include value of money in his calculation.
Typically, we calculate the net present value(NPV) to evaluate a project. For example, estimated extra rent is 250000*3.1(mean)=775000 annually(which will be discounted by the discount factor), extra investment is 5000000; if we want this project to begin to generate profit in 8 years, the IRR should be 5.05%.In rhe other words, if the interest rate is lower than 5.05%, we can invest in a green building if we want to get profit in 8 years.

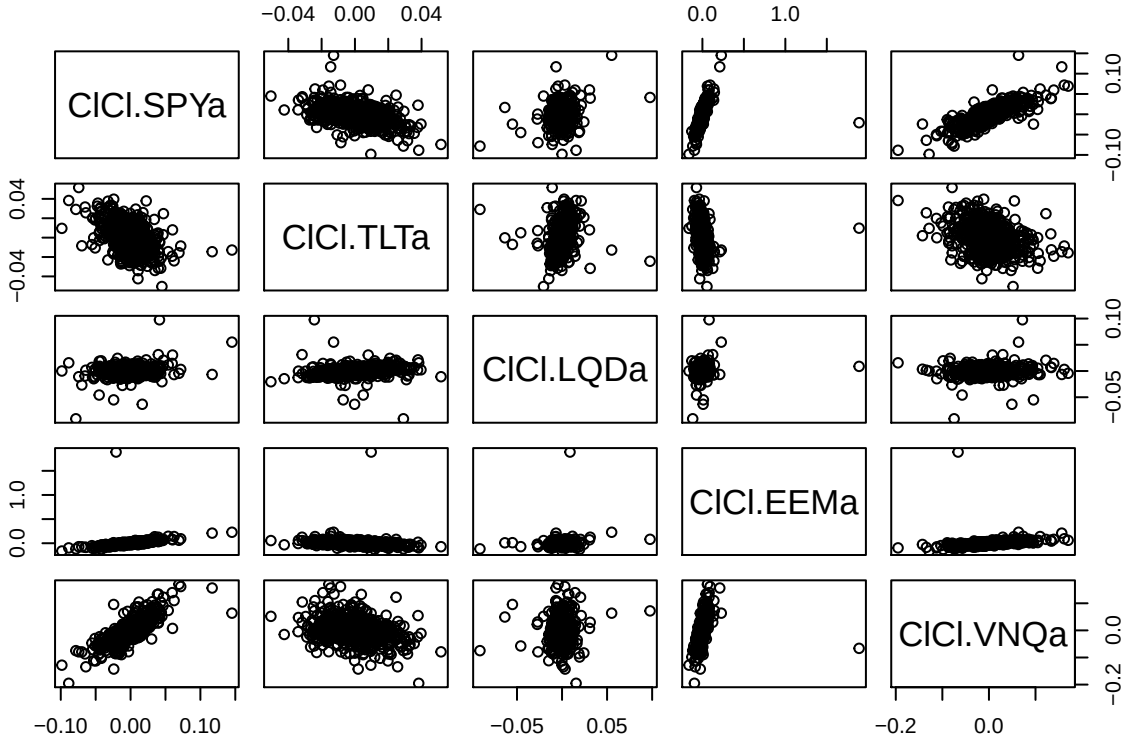
5. Conclusion

The green building has a higher leasing rate and rent than non-green buildings, about 3.1dollars more per square foot per calendar year. People who prefer to green buildings also prefer to Class A buildings. However, the main goal for us is earing. So we also need to consider the interest rate. If we decide to build a green

building and want to make profit on 8th year, the interest rate must be lower than 5.05%. Otherwise, it is much better to build a non-green building.

Boostrapping

Setup:




```
##             ClCl.SPYa  ClCl.TLTa  ClCl.LQDa   ClCl.EEMa   ClCl.VNQa
## ClCl.SPYa  1.0000000 -0.4362148 0.10133468  0.40676425  0.76813129
## ClCl.TLTa -0.4362148  1.0000000 0.43237319 -0.16758098 -0.25332123
## ClCl.LQDa  0.1013347  0.4323732 1.00000000  0.08784764  0.07156075
## ClCl.EEMa  0.4067643 -0.1675810 0.08784764  1.00000000  0.29228612
## ClCl.VNQa  0.7681313 -0.2533212 0.07156075  0.29228612  1.00000000
```

From the pairwise plots and the correlation matrix, we find that:
1) relatively strong negative correlations: SPYa & TLTa, TLTa & VNQa;
2) relatively strong positive correlations: SPYa & VNQa, TLTa & LQDa; SPYa & EEMa.

```
##    ClCl.SPYa            ClCl.TLTa            ClCl.LQDa
##  Min.   :-0.0984477   Min.   :-0.0504495   Min.   :-0.0911111
##  1st Qu.:-0.0038636   1st Qu.:-0.0051997   1st Qu.:-0.0019083
##  Median : 0.0006589   Median : 0.0005596   Median : 0.0004165
##  Mean   : 0.0003981   Mean   : 0.0002788   Mean   : 0.0002095
##  3rd Qu.: 0.0056254   3rd Qu.: 0.0057014   3rd Qu.: 0.0024660
##  Max.   : 0.1451977   Max.   : 0.0516616   Max.   : 0.0976772
```

```
##    ClCl.EEMa            ClCl.VNQa
##  Min.   :-0.1616620   Min.   :-0.1951372
##  1st Qu.:-0.0085338   1st Qu.:-0.0069050
##  Median : 0.0008056   Median : 0.0006695
##  Mean   : 0.0009814   Mean   : 0.0004157
##  3rd Qu.: 0.0091897   3rd Qu.: 0.0077793
##  Max.   : 1.8891250   Max.   : 0.1700654
```

```
##   ClCl.SPYa   ClCl.TLTa   ClCl.LQDa   ClCl.EEMa   ClCl.VNQa
## 0.012436800 0.009148468 0.005213764 0.040200400 0.021136720
```

It seems that among these five asset classes, SPY, TLT and LQD are the safe ones while EEM and VNQ are the aggressive ones.

Even split portfolio:

First, let's try to peoform an even split portfolio.

```
## [1] 100982.3
```

```
##        5%
## -6206.433
```

The average profit of the even split portfolio is 863.4 and the corresponding 5% value at risk is 6417.

Safe split portfolio:

Let's try several different safer portfolios. For each portfolio, we just invest SPY, TLT and LQD.
####(1)safe split (0.3, 0.4, 0.3, 0, 0):
The strategy for this split is that, since TLT has the lowest risk for losing much money among these three safe assets, I would like to assign 0.4 to TLT and evenly assign 0.3 to each of the other two.

```
## [1] 100626.3
```

```
##        5%
## -2931.861
```

The average profit of this portfolio is 547.3 and the corresponding 5% value at risk is 2948.733. Comparing to the even split, it has lower profit on average, and meanwhile has lower corresponding 5% value at risk.

(2)safe split (0.2, 0.4, 0.4, 0, 0):

The strategy for this split is that, since 1) TLT is the safest one, and 2) TLT's return has negative correlation with SPY's return but postive correlation with LQD's return, I'd like to assign 0.2, 0.4, 0.4 to SPY, TLT, LQD, respectively.

```
## [1] 100580.5
```

```
##       5%
## -3020.04
```

The average profit of this portfolio is 507.2 and the corresponding 5% value at risk is 2945.589. Comparing to the former two portfolios, this one is the safest one with lowest profit and risk.

Aggressive split portfolio:

Now, we would go for aggresive investment. For each portfolio, we just consider to invest EEM and VNQ.
####(1)Aggressive split (0, 0, 0, 0.5,0.5):
The strategy for this split is just to invest these two aggresive assets evenly.

```
## [1] 101511
```

```
##        5%
## -12641.51
```

The average profit of this portfolio is 1330.1 and the corresponding 5% value at risk is 12601.31. As expected, this portfolio is likely to produce much higher profit. However, the risk also increases dramaticaly.

(2)Aggressive split (0, 0, 0, 0.8,0.2):

Comparing the performances of EEM and VNQ, EEM is the more aggressive one. This time, we take a risk for higher profit. Let's assgin 0.8 of wealth for EEM and the left 0.2 for VNQ.

```
## [1] 101894.6
```

```
##        5%
## -12881.11
```

The average profit of this portfolio is 1679.2 and the corresponding 5% value at risk is 12876.22.
Again, this portfolio could may more profit than the previous one. And making tradeoff between profit and risk, I think this portfolio actually outperform the even aggresive split portfolio.

(3)Aggressive split (0, 0, 0, 0.9,0.1):

Let's try an even more aggresive one.

```
## [1] 102022.3
```

```
##        5%
## -13086.61
```

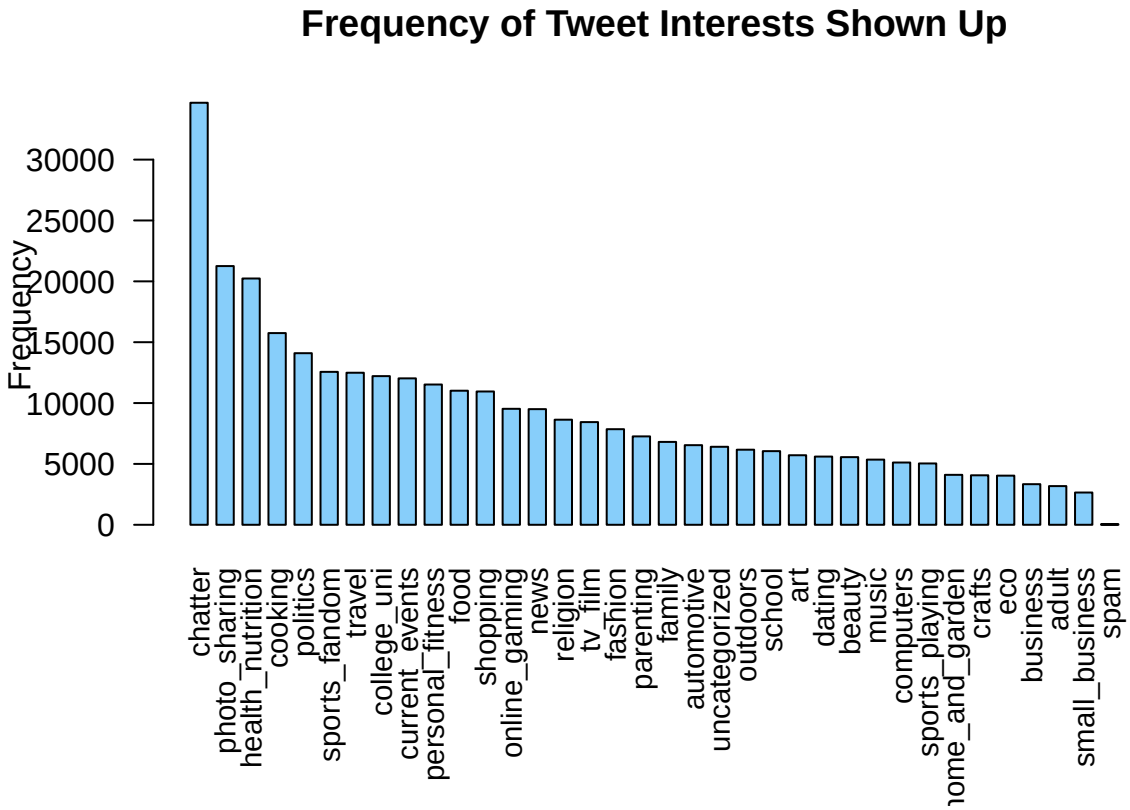
The average profit of this portfolio is 1794 and the corresponding 5% value at risk is 13040.19. Same conclusion as before. Higher profit, higher risk!
#Market segmentation
###Data pre-processing

```
## [1] 7882   36
```

```
##  [1] "chatter"          "current_events"   "travel"
##  [4] "photo_sharing"    "uncategorized"    "tv_film"
##  [7] "sports_fandom"    "politics"         "food"
## [10] "family"           "home_and_garden"  "music"
## [13] "news"             "online_gaming"    "shopping"
## [16] "health_nutrition" "college_uni"      "sports_playing"
## [19] "cooking"          "eco"              "computers"
## [22] "business"         "outdoors"         "crafts"
## [25] "automotive"       "art"              "religion"
## [28] "beauty"           "parenting"        "dating"
## [31] "school"           "personal_fitness" "fashion"
## [34] "small_business"   "spam"             "adult"
```

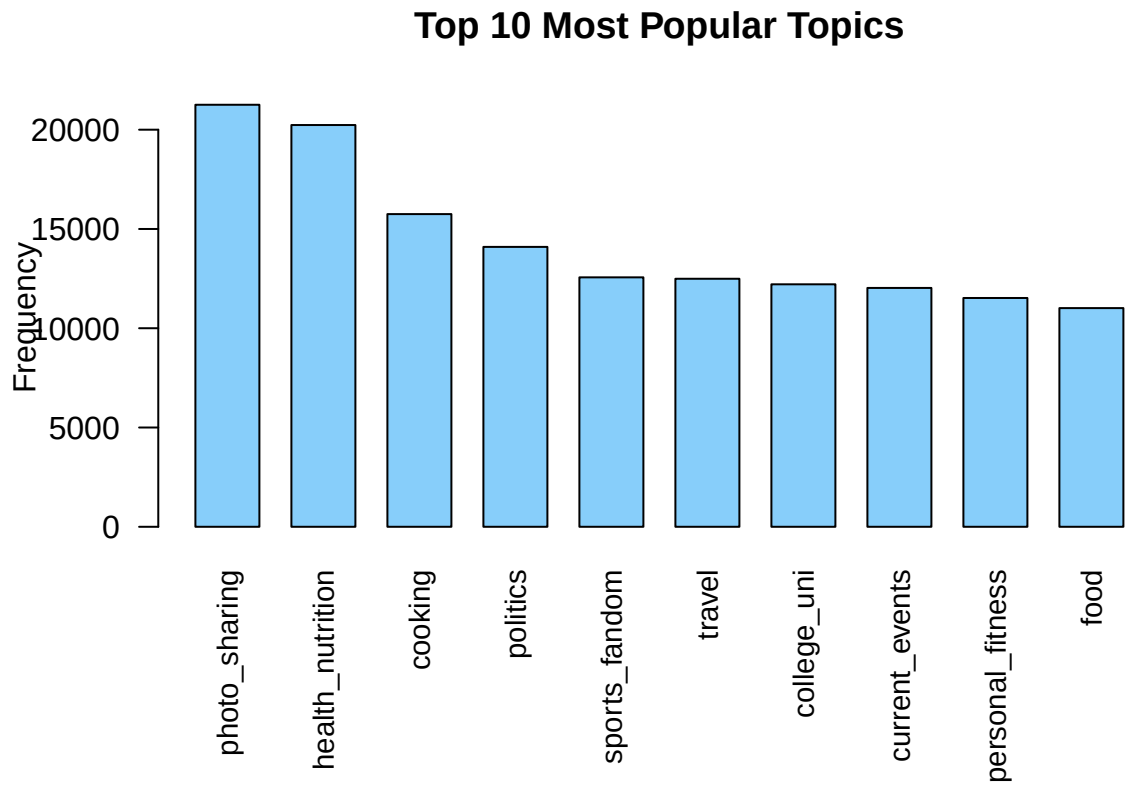
From the data pre-processing, we can see that we have 7882 twitter participates with 36 different interests categories.

Trends of interests in 7 days



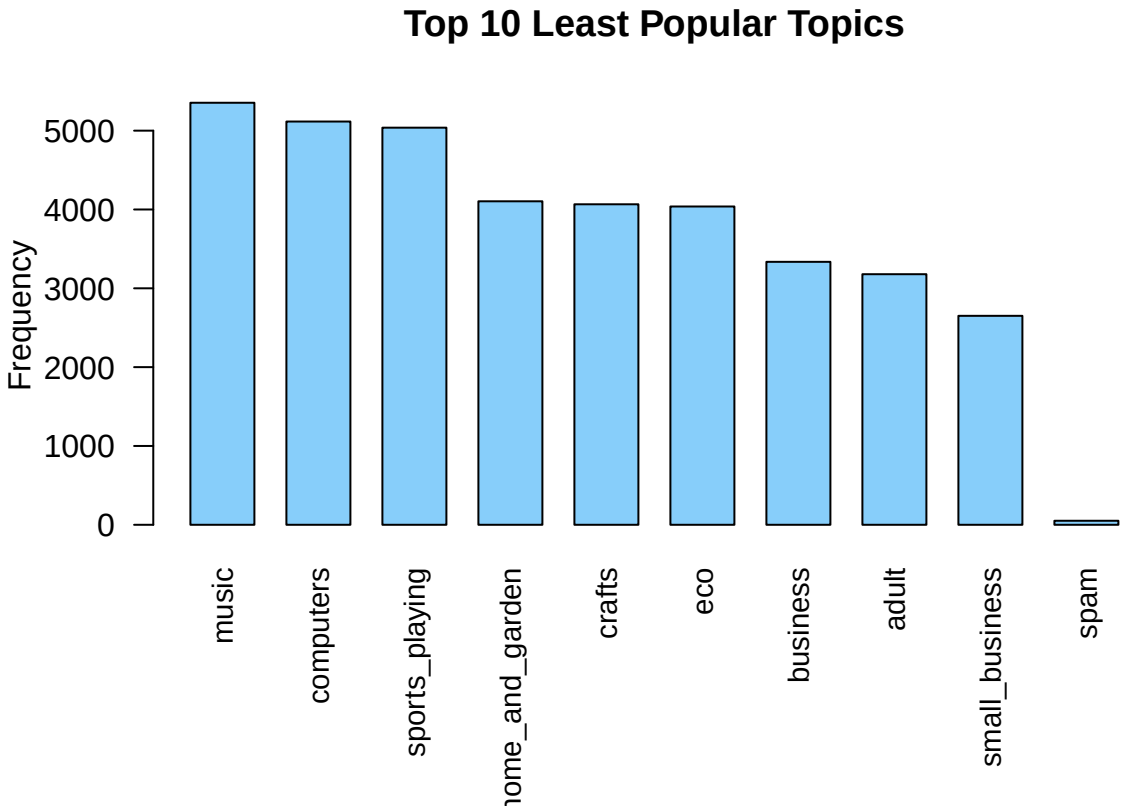
From the bar plot we can easily see the trends in 7 days, the most popular topics as well as the least. One thing we might notice is that "chatter" is the top most frequent category, but it could due to lots of annotators used this category for posts that didn't fit at all into any of the listed interest categories.

The top 10 most popular interests



This plot shows the most frequent topics shown up in our 7-day data sample, except chatter, which are uncategorized. We suggest NutrientH20 post more tweets related to these topics.

The top 10 least popular interests

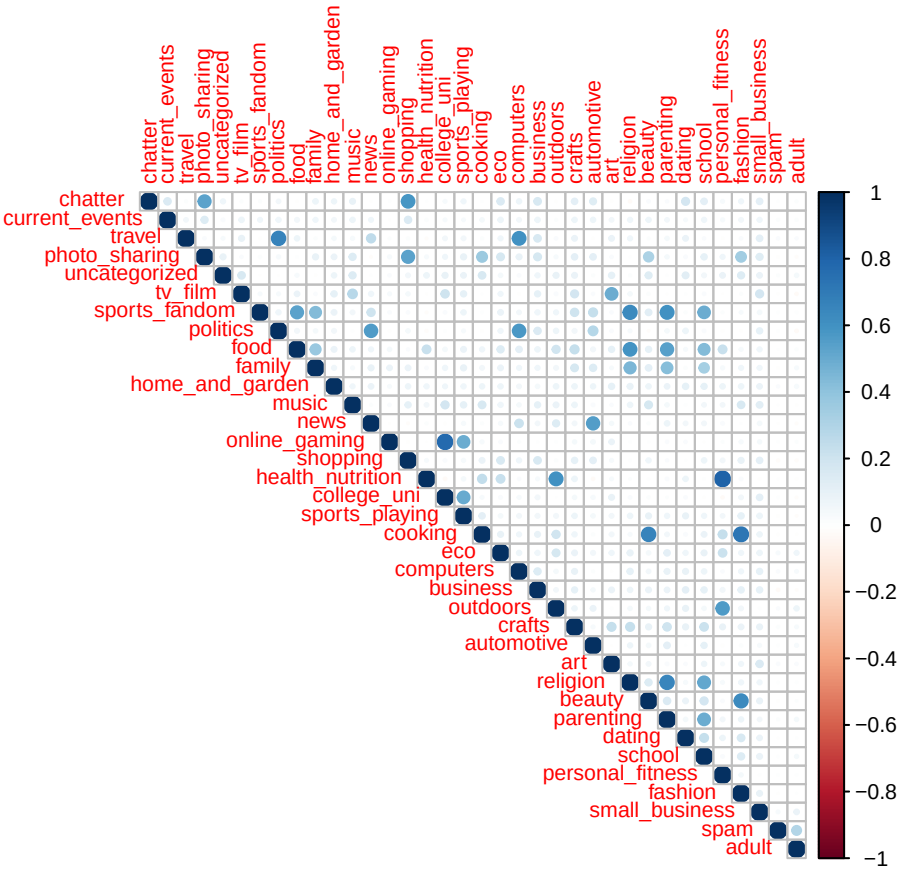


Similarly, this plot shows the least frequent topics shown up in our 7-day data sample. We suggest NutrientH20 post less tweets related to these topics.

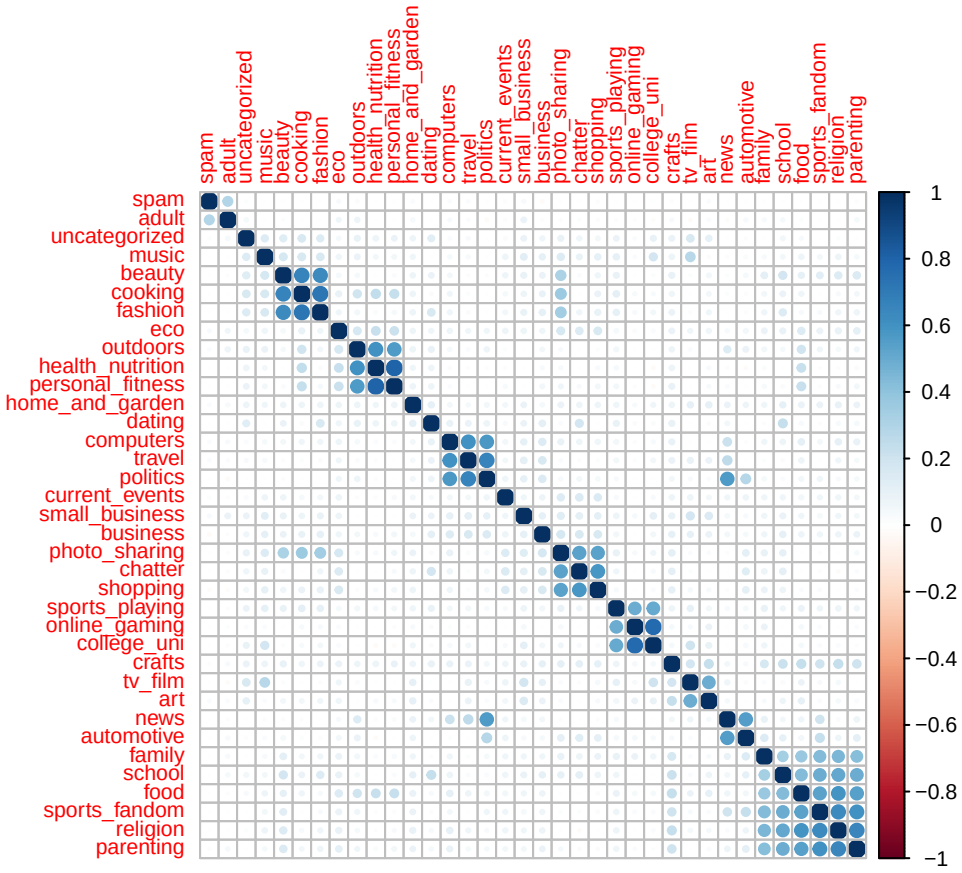
Correlation between interests

```
## Compare row 27  and column  29 with corr  0.656
##   Means:  0.128 vs 0.088 so flagging column 27
## Compare row 19  and column  28 with corr  0.664
##   Means:  0.116 vs 0.086 so flagging column 19
## Compare row 8  and column  3 with corr  0.66
##   Means:  0.1 vs 0.085 so flagging column 8
## Compare row 32  and column  16 with corr  0.81
##   Means:  0.089 vs 0.084 so flagging column 32
## Compare row 17  and column  14 with corr  0.773
##   Means:  0.089 vs 0.085 so flagging column 17
## All correlations <= 0.65
```

```
## [1] 27 19  8 32 17
```



From the correlation plots, we can see serval highly corelated pairs such as “personal fitness” and “health nutrition”, “college university” and “online gaming”, “cooking” and “fashion”.



From the correlation matrix, there are some obvious relatively high correlations within some feature groups.

Group 1: beauty, cooking, fashion

Group 2: outdoors, health nutrition, personal fitness

Group 3: computers, travel, politics

Group 4: photo sharing, chatter, shopping

Group 5: sports playing, online gaming, college university

Group 6: tv film, art

Group 7: news, automotive

Group 8: family, school, food, sports fandom, religion, parenting

This result seems reasonable to our intuition. Take Group 5 as an example. The users who have more college university related posts tend to have more posts on sports playing topics and online gaming topics, since these users are likely to be college students.

factor analysis

```
##
## Call:
## factanal(x = N20, factors = 10, rotation = "varimax")
##
## Uniquenesses:
##          chatter   current_events           travel    photo_sharing
```

```
##            0.339            0.944            0.297            0.380
##    uncategorized         tv_film    sports_fandom         politics
##            0.911            0.120            0.354            0.200
##             food           family  home_and_garden            music
##            0.451            0.687            0.943            0.860
##             news    online_gaming         shopping health_nutrition
##            0.263            0.244            0.440            0.119
##      college_uni   sports_playing          cooking              eco
##            0.174            0.658            0.164            0.877
##        computers         business         outdoors           crafts
##            0.462            0.894            0.545            0.849
##       automotive              art         religion           beauty
##            0.472            0.715            0.288            0.396
##        parenting           dating           school personal_fitness
##            0.388            0.235            0.536            0.253
##          fashion   small_business             spam            adult
##            0.306            0.898            0.899            0.138
##
## Loadings:
##                  Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## chatter                                           0.795
## current_events                                    0.193
## travel                                    0.830                   0.105
## photo_sharing                     0.338           0.706
## uncategorized                     0.146                           0.196
## tv_film                                                           0.932
## sports_fandom     0.777
## politics                                  0.789
## food              0.695   0.224
## family            0.536
## home_and_garden                                                   0.124
## music                             0.152           0.123           0.293
## news                                      0.288
## online_gaming                                             0.869
## shopping                                          0.744
## health_nutrition          0.936
## college_uni                                               0.890   0.178
## sports_playing                                            0.557
## cooking                   0.224   0.878
## eco               0.117   0.232                   0.205
## computers                                 0.720
## business                                  0.171   0.207           0.118
## outdoors                  0.654
## crafts            0.270                           0.124           0.212
## automotive        0.126                           0.144
## art                                                               0.531
## religion          0.837
## beauty            0.149           0.754
## parenting         0.779
## dating
## school            0.627                           0.105
## personal_fitness          0.860
## fashion                           0.811           0.108
## small_business                            0.113   0.147           0.215
```

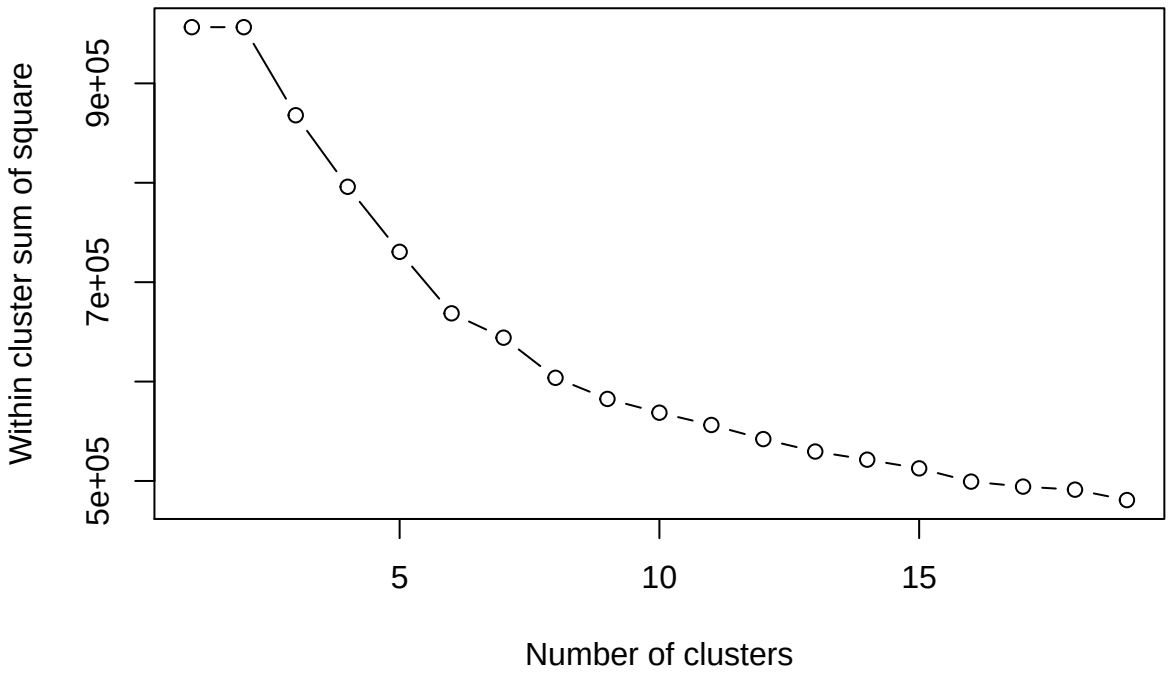
```
## spam
## adult
##                  Factor8 Factor9 Factor10
## chatter                           0.159
## current_events
## travel
## photo_sharing
## uncategorized                     0.115
## tv_film
## sports_fandom     0.200
## politics          0.415
## food
## family
## home_and_garden
## music
## news              0.801
## online_gaming
## shopping
## health_nutrition
## college_uni
## sports_playing
## cooking
## eco
## computers
## business
## outdoors          0.113
## crafts
## automotive        0.698
## art
## religion
## beauty
## parenting
## dating                            0.861
## school                            0.217
## personal_fitness
## fashion                           0.116
## small_business
## spam                      0.317
## adult                     0.928
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## SS loadings      3.258   2.232   2.214   1.986   1.957   1.896   1.499
## Proportion Var   0.091   0.062   0.062   0.055   0.054   0.053   0.042
## Cumulative Var   0.091   0.153   0.214   0.269   0.324   0.376   0.418
##                Factor8 Factor9 Factor10
## SS loadings      1.380   0.989    0.889
## Proportion Var   0.038   0.027    0.025
## Cumulative Var   0.456   0.484    0.508
##
## Test of the hypothesis that 10 factors are sufficient.
## The chi square statistic is 1459.59 on 315 degrees of freedom.
## The p-value is 1.9e-146
```

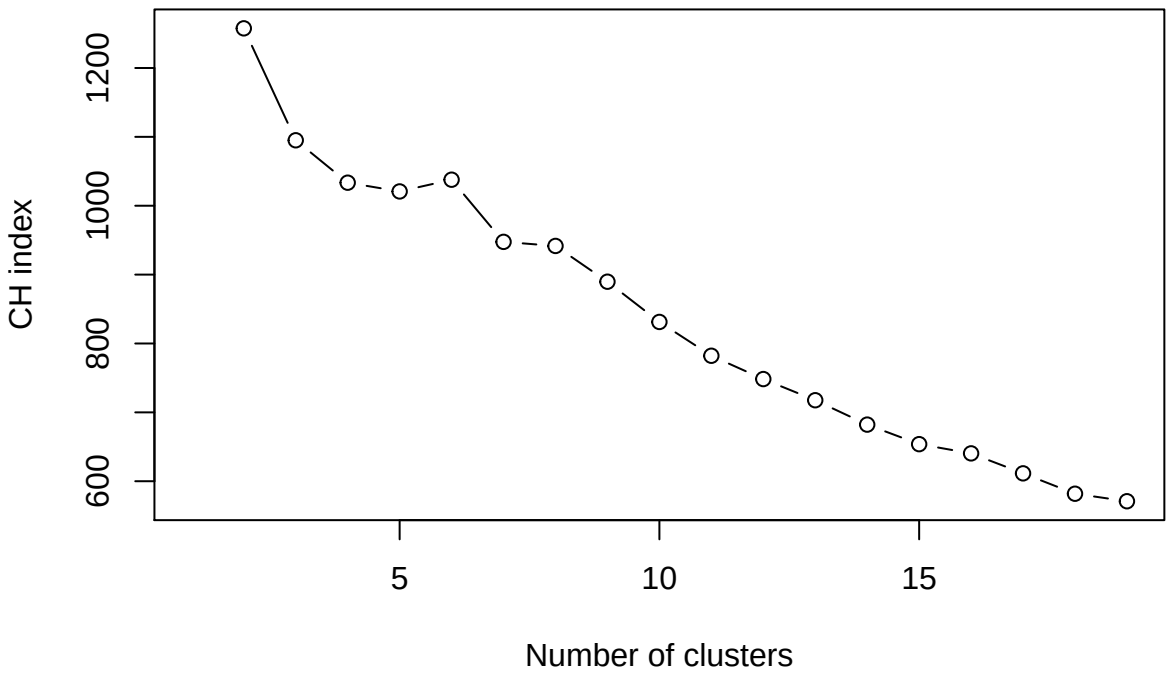
The first ten factors can only account for 50.8% of the variance. But still, we can obtain some useful

information from factor analysis. For example, those users with high score of factor1 are likely to be parents, because they are interested in sports_fandom, food, family, religion, parenting, school topics. As for factor2 which is mainly composed by health_nutrition, outdoors and personal_fitness, users with high score of factor2 are those who care about personal health. We can calculate the scores of these factors for each users, and thus to detech which consumer groups they might be.

Any clusters?

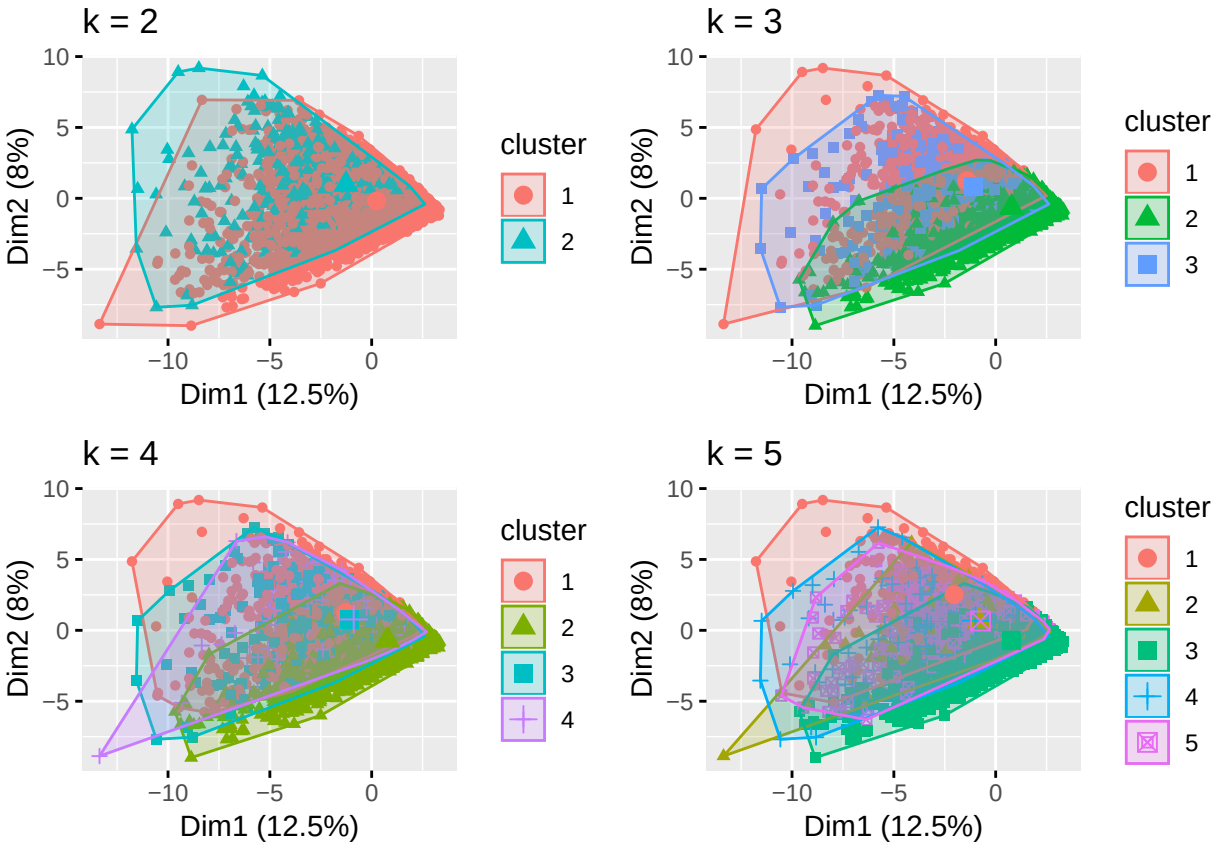
K-means clustering

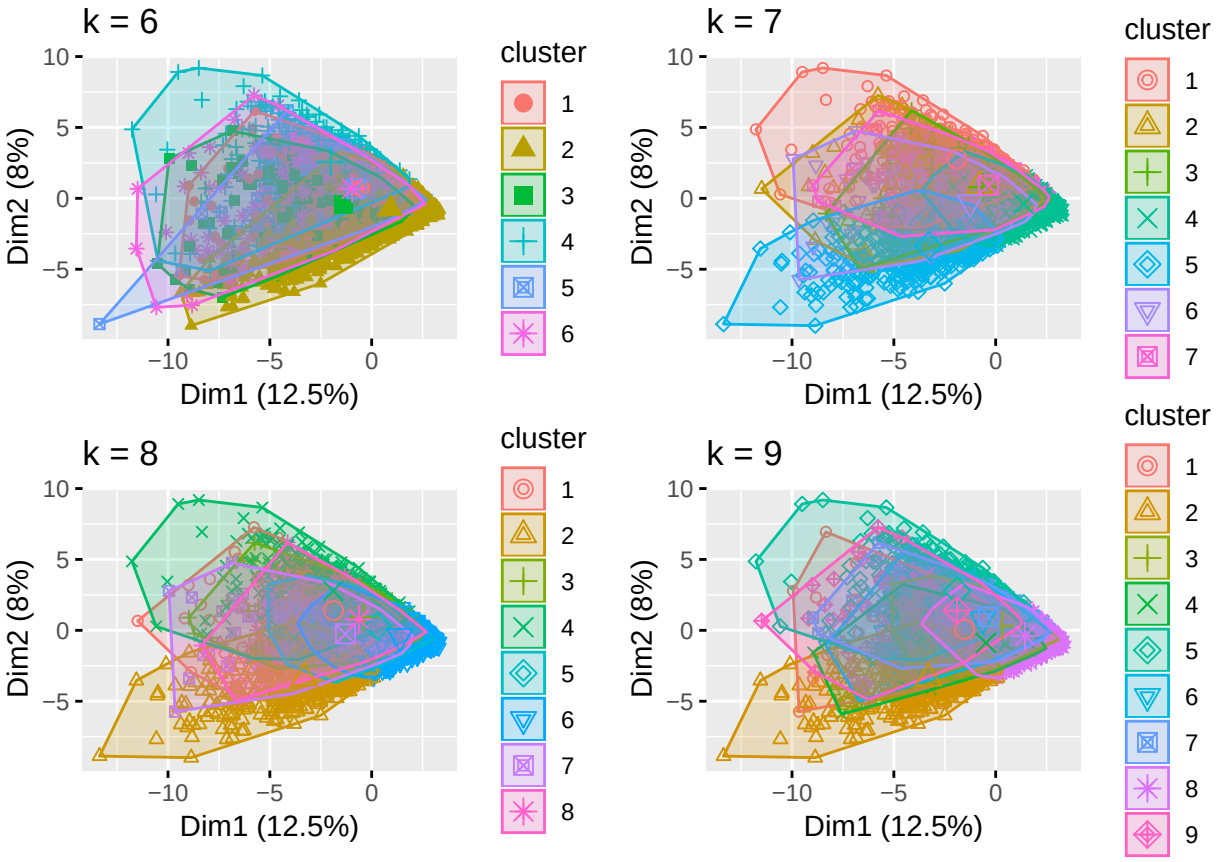




It's hard to choose number of clusters.

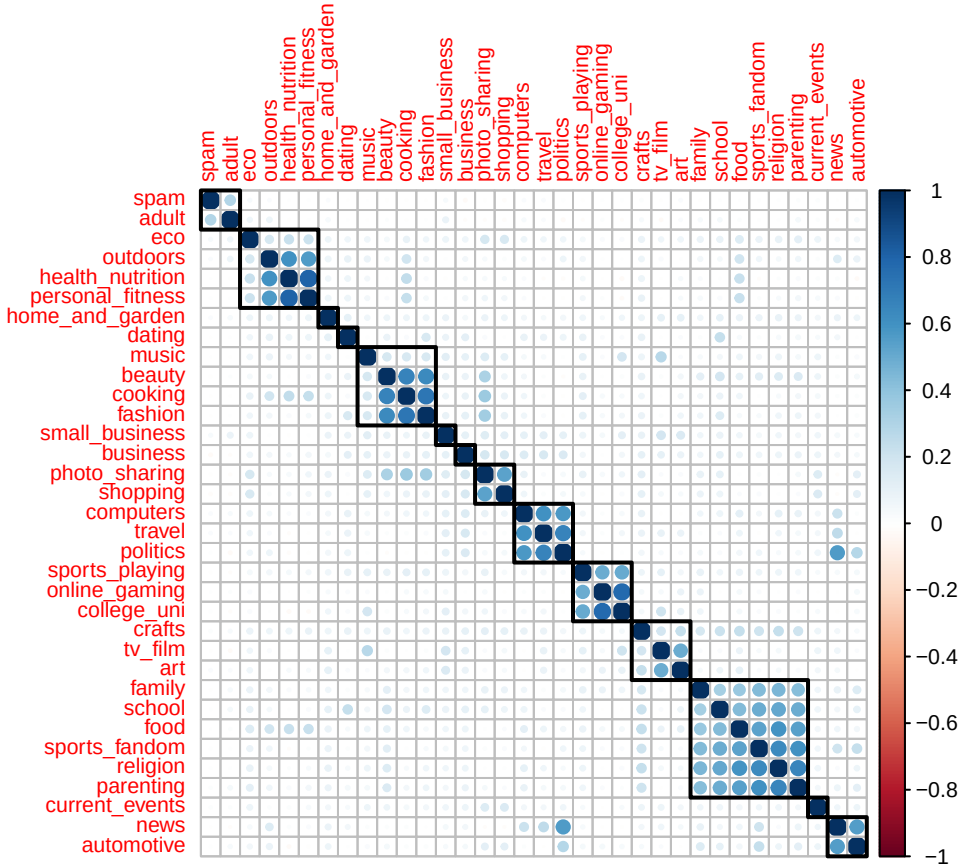
Plots to compare


k = 2
k = 3
k = 4
k = 5
Dim1 (12.5%)
Dim2 (8%)
cluster
1
2
3
4
5



We can view our results after we try different k value. Since we have more than two variables, the "fviz_cluster" function will perform principal component analysis for us and plot the data points based on the first two principal components which could explain the majority of the variance.

Conclusion



Among all categories, *photo_sharing* is the most popular category and *small busienss* is the least popular category, exclude *spam*.

Based on our analysis, we seperate the market into 14 group by using correlation analysis. For example, * music, beauty, cooking, fashion* can be put into one group and it is reasonable. Since a person who loves music or fashion most likely to be a person who pay attention to the quantity of life, so he or she may also interests in cooking and beauty. We also tried to use cluster but which is not a good choice for this dataset.