STA 380, Part 2: Exercises

Meghna Kundur

2022-08-01

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? Hint: use the rule of total probability.

- P of Random Clickers = 0.3
  - P of RC (Yes) = 0.5
  - P of RC (No) = 0.5
- P of Truthful Clickers = 0.7 (1-0.3 = 0.7)
- P of Survey (Yes) = 0.65
  - P of Survey (No) = 0.35
- P of Truthful Clickers (Yes) = X

Formula: P of Survey (Yes) = P of Random Clickers x P of RC (Yes) + P of Truthful Clickers x P of Truthful Clickers (Yes)

Calculation:

- 0.65 = 0.3 * 0.5 + 0.7X
- 0.65 = 0.15 + 0.7X
- 0.50 = 0.7X
- 0.5/0.7 = X

Solution: The fraction of people who are truthful clickers that answered yes is 0.714.

Part B

- 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?

- P of Disease in Pop. = 0.000025
- P of Positive Given Disease = 0.993
- P of No Disease in Pop. = 0.999975 (1-0.000025)
- P of Positive Given No Disease = 0.0001 (1-0.9999)

Formula: P of Disease Given Positive = (P of Disease in Population * P of Positive Given Disease)/(P of Disease in Population * P of Positive Given Disease) + (P of No Disease in Population * P of Positive Given No Disease)

Calculation: (0.993. * 0.000025)/(0.993 * 0.000025) + (0.999975 * 0.0001)

Solution: The probability that someone who tests positive has the disease is 0.198.

Wrangling the Billboard Top 100

Part A

Make a table of the top 10 most popular songs since 1958, as measured by the total number of weeks that a song spent on the Billboard Top 100. Note that these data end in week 22 of 2021, so the most popular songs of 2021 will not have up-to-the-minute data; please send our apologies to The Weekend.

Table 1: Figure 1.A. is a table of the top 10 most popular songs since 1958, as measured by the total number of weeks that a song spent on the Billboard Top 100.

performer	song	count
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
Jason Mraz	I'm Yours	76
The Weeknd	Blinding Lights	76
LeAnn Rimes	How Do I Live	69
LMFAO Featuring Lauren Bennett & GoonRock	Party Rock Anthem	68
OneRepublic	Counting Stars	68
Adele	Rolling In The Deep	65
Jewel	Foolish Games/You Were Meant For Me	65
Carrie Underwood	Before He Cheats	64

Part B

Is the "musical diversity" of the Billboard Top 100 changing over time? Let's find out. We'll measure the musical diversity of given year as the number of unique songs that appeared in the Billboard Top 100 that year. Make a line graph that plots this measure of musical diversity over the years. The x axis should show the year, while the y axis should show the number of unique songs appearing at any position on the Billboard Top 100 chart in any week that year. For this part, please filter the data set so that it excludes the years 1958 and 2021, since we do not have complete data on either of those years. Give the figure an informative caption in which you explain what is shown in the figure and comment on any interesting trends you see.

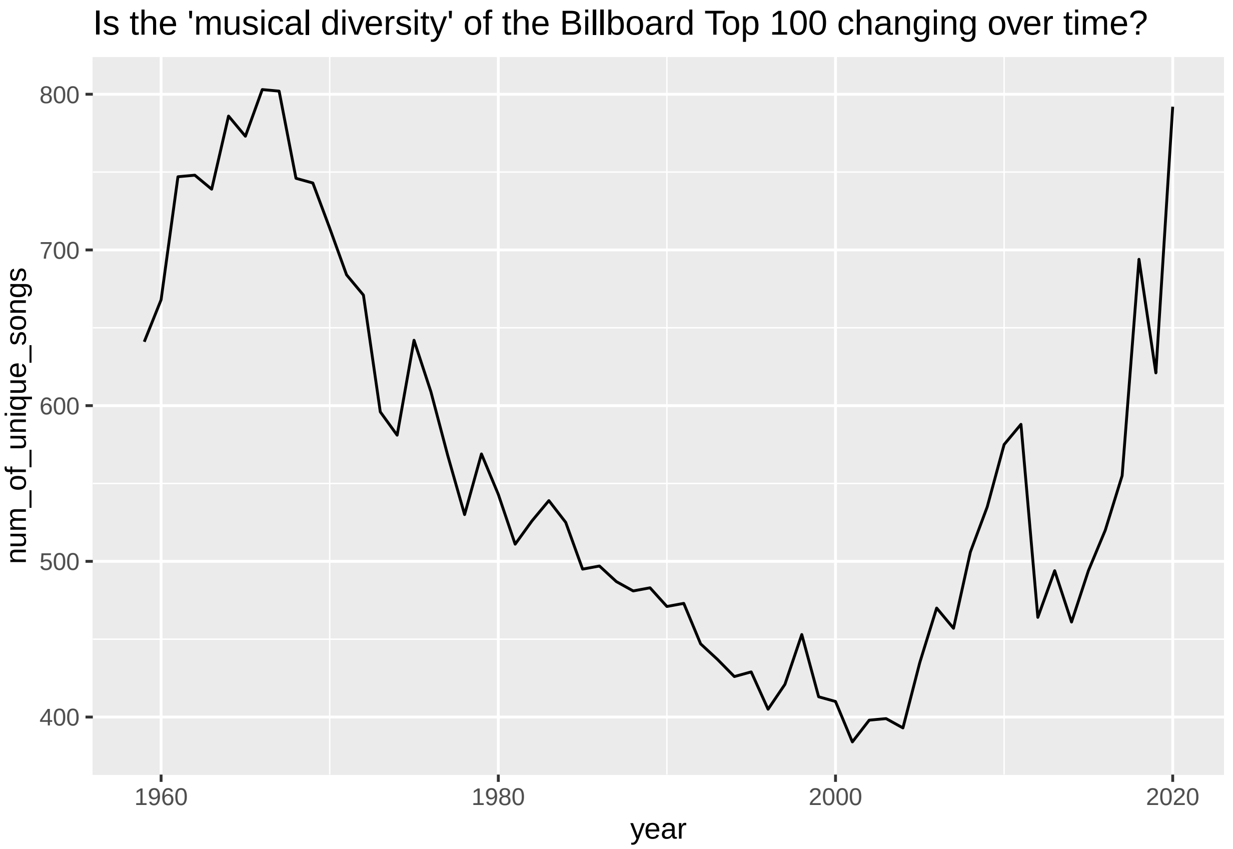


Figure 1.B. is a line graph that plots the measure of musical diversity over the years. The x axis shows the year, while the y axis shows the number of unique songs appearing at any position on the Billboard Top 100 chart in any week that year.

Part C

Let’s define a “ten-week hit” as a single song that appeared on the Billboard Top 100 for at least ten weeks. There are 19 artists in U.S. musical history since 1958 who have had at least 30 songs that were “ten-week hits.” Make a bar plot for these 19 artists, showing how many ten-week hits each one had in their musical career. Give the plot an informative caption in which you explain what is shown.

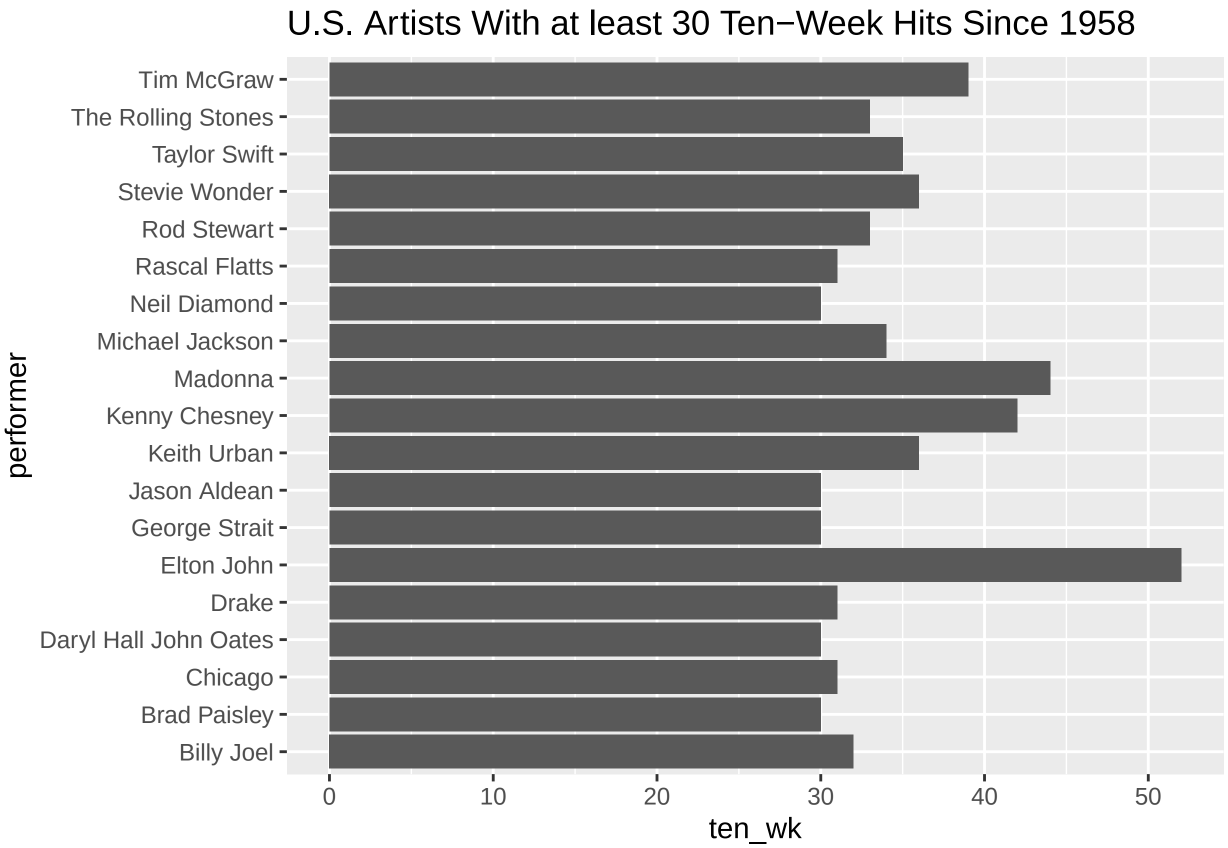


Figure 1.C. is a bar plot for 19 artists in U.S. musical history who have had at least 30 songs that were ten-week hits since 1958. The plot shows the performer on the x-axis and the number of ten-week hit songs they had.

Visual story telling part 1: green buildings

The developer has had someone on her staff, who's been described to her as a "total Excel guru from his undergrad statistics course," run some numbers on this data set and make a preliminary recommendation. 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. Do you see the possibility of confounding variables for the relationship between rent and green status? If so, provide evidence for confounding, and see if you can also make a picture that visually shows how we might "adjust" for such a confounder. Tell your story in pictures, with appropriate introductory and supporting text.

Structured the visual story telling as if it were an email just for fun! All the elements are provided, they just flow as if it were an email.

Subject: Revisiting Green Buildings

Good afternoon,

I've completed my evaluation on the economic return of your company potentially investing in a green building and "going green." Firstly, I do agree with your staff member's Excel report in that my final conclusion too is to move forward with the investment, however I received different numerical results because the previous report didn't account for confounders, or other variables that could affect return besides if a building is green or not. The Excel report also used an arbitraty occupancy rate and building size (I'm still unsure exactly where the size 250,000 sq ft originated from) and the final conclusion would be better drawn from median values of both categories. I've provided plots/tables and comments below followed by my final calculations and suggestions.

Plots:

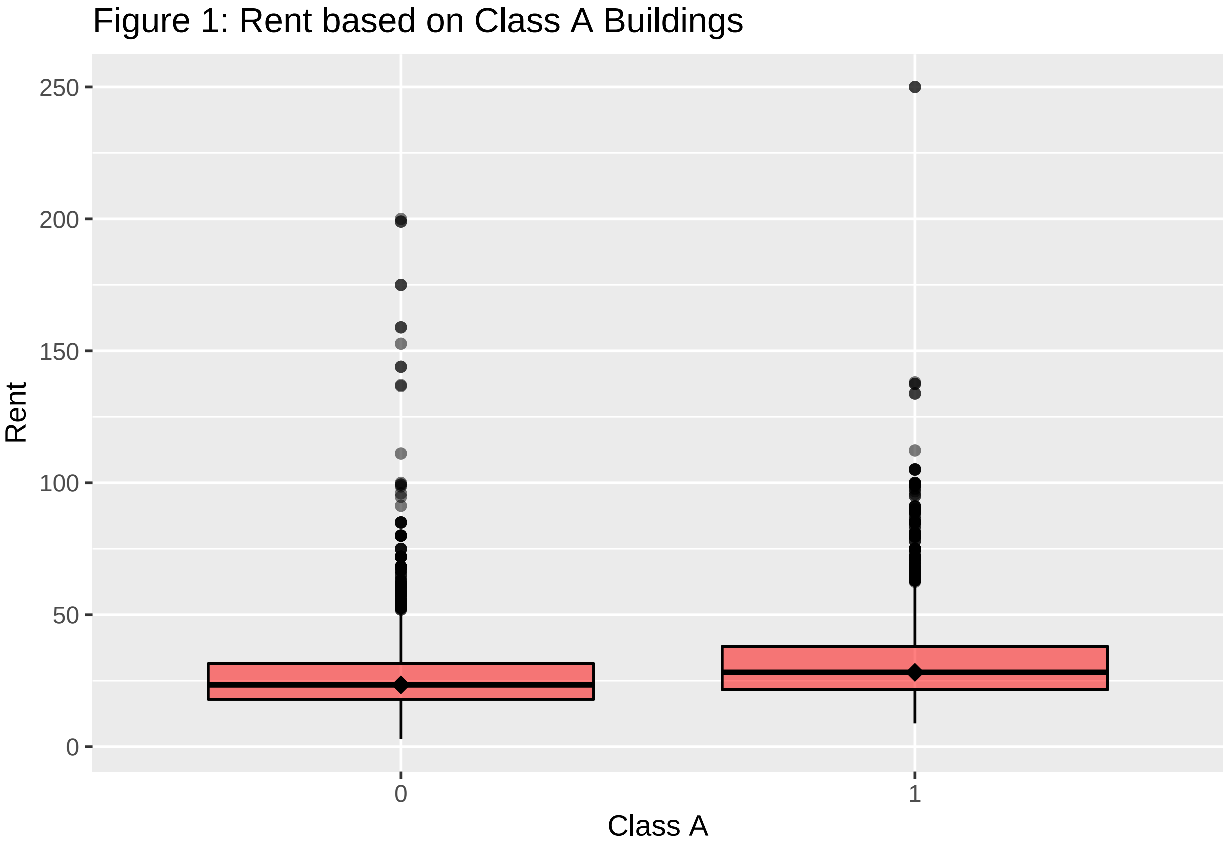


Figure 1: You may not deal with box plots often in real estate, but they are very useful in understanding the difference in distributions between groups! Here we see that Class A buildings have a greater median rent price than non-Class A buildings, which is to be expected since they are premium properties. It is still helpful however to have illustrated this relationship since we can identify building class as a potential confounder.

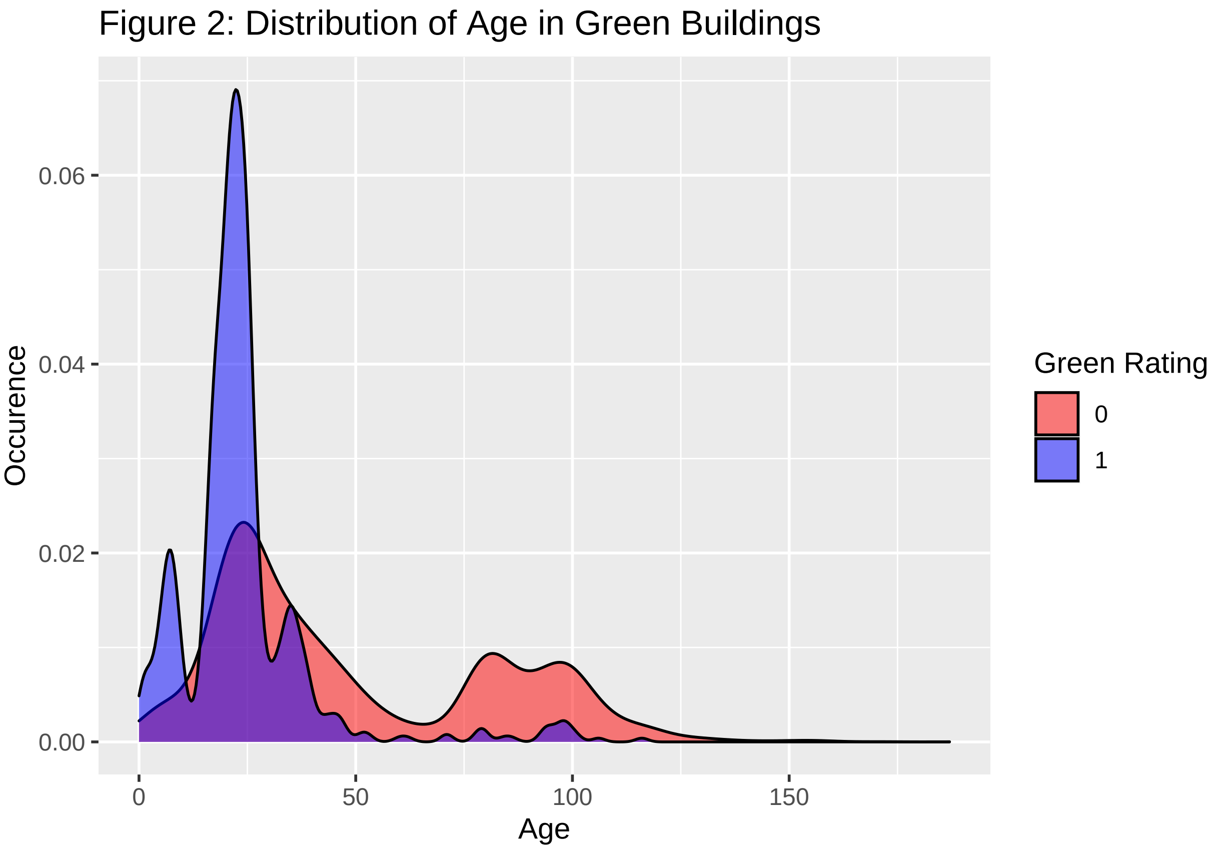


Figure 2: This plot was useful in depicting the distribution of age based on if a building is considered green or not. Here we see that green buildings are typically "younger" which is understandable as companies are increasingly "going green."

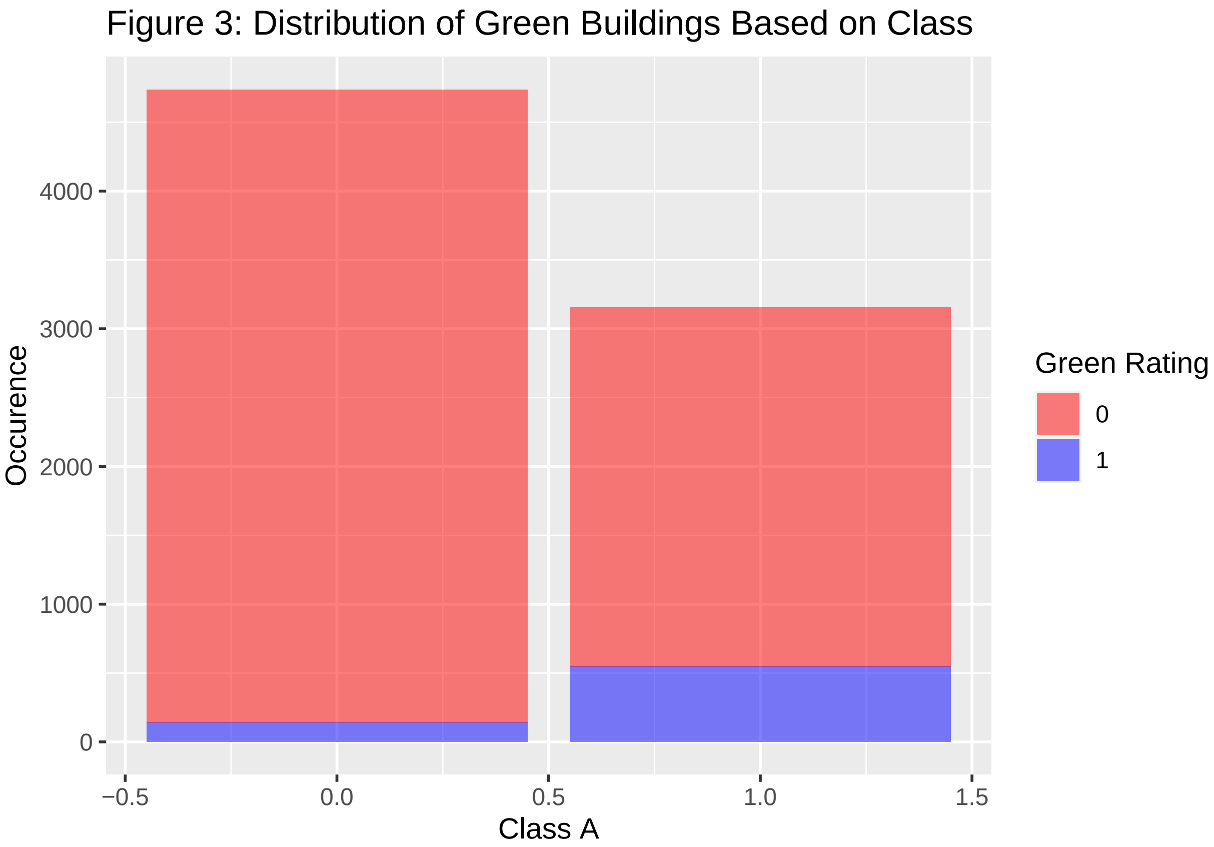


Figure 3: There is a greater proportion of green buildings that are Class A as opposed to non-Class A.

Tables:

```
##   green_rating   size
## 1            0 118696
## 2            1 241150
```

Table 1: Since the median size for a green building is 241,150 sq feet, this is the value that will be used in the final calculation.

```
##   class_a green_rating  Rent
## 1       0            0 23.43
## 2       1            0 28.20
## 3       0            1 25.55
## 4       1            1 28.44
```

Table 2: This table represents the median rent based on building class and green rating.

```
##   class_a green_rating leasing_rate
## 1       0            0        87.15
## 2       1            0        92.63
## 3       0            1        89.80
## 4       1            1        93.63
```

Table 3: This table presents the median leasing rate based on building class and green rating.

- Quick Summary before Final Calculations and Suggestions:
  - Class A buildings have a greater median rent
  - Green buildings are typically "younger"
  - More Class A buildings are considered green
  - The median size of a green building is 241,150 sq ft
  - The median rent of a Class A green building is 28.44 while the median rent for a non-Class A green building is 25.55, thus the difference is 2.89
  - The median leasing rate of a Class A green building is 93.63

Final Calculations:

```
## [1] "You can expect to recuperate your expenses in the following time frame: 7.66 years"
```

Final Suggestions:

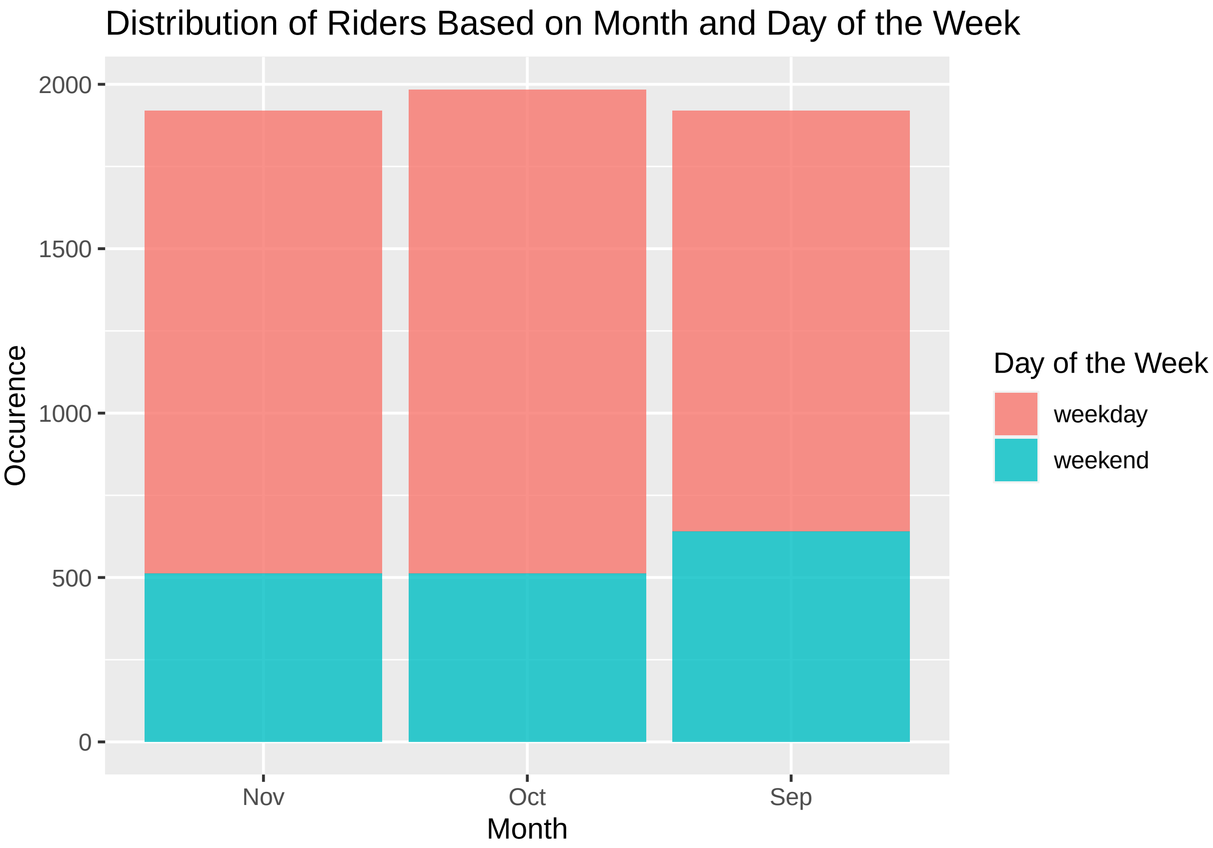
After taking into account median values for rent and occupancy rate based on class and green rating without scrubbing for low occupancy, I would suggest moving forward with the investment if you're able to secure a Class A property type as you would otherwise anticipate unsatisfactory returns.

If you have any further questions don't hestiate to reach out, I hope you've found this helpful in making your decision!

Best, Meghna Kundur

Visual story telling part 2: Capital Metro data

Your task is to create a figure, or set of related figures, that tell an interesting story about Capital Metro ridership patterns around the UT-Austin campus during the semester in question. Provide a clear annotation/caption for each figure, but the figure(s) should be more or less stand-alone, in that you shouldn't need many, many paragraphs to convey its meaning. Rather, the figure together with a concise caption should speak for itself as far as possible.



In order to tell an interesting story about Capital Metro ridership patterns around the UT-Austin campus, I evaluated the distribution of riders based on the month (September, October, and November) and the day of the week (weekend or weekday). We can observe that there were more metros utilized in October compared to November and September. We can also see that more weekend travels occurred in September compared to the other months. In general more weekday travels are undertaken, but that is to be expected because there are simply more weekdays than weekend days.

Portfolio modeling

Question:

In this problem we will create three different portfolios of exchange-traded funds, or ETFs, and use bootstrap resampling to analyze the short-term tail risk of the portfolios. We will allocate $100,000 of capital between three to ten ETFs who have at least fove years worth of information and estimate the 4-week (20 trading day) value at risk for each of the three portfolios at the 5% level.

Approach:

The three different portfolios were categorized in the following way:

- Portfolio A: Defensive Portfolio
  - DVY - iShares Select Dividend ETF
  - VIG - Vanguard Dividend Appreciation ETF
  - SPLV - Invesco S&P 500 Low Volatility ETF

- Portfolio B: Aggressive Portfolio
  - EFG - iShares MSCI EAFE Growth Index ETF
  - GOEX - Global X Gold Explorers ETF
  - IBUY - Amplify Online Retail ETF
- Portfolio C: High Risk Portfolio
  - VWO - Vanguard Emerging Markets ETF
  - GNR - SPDR S&P Global Natural Resources ETF
  - SCZ - iShares MSCI EAFE Small Cap Index

Then I imported the stocks respective to the portfolio along with their prices for the past five years and adjusted for splits using the adjustOHLC function within a for loop. Next I combined all of the close to close changes in a single matrix using the cbind function. Afterwards I created a block that included my capital of $100,000, the weight amount for each stock in the portfolio, the specified time horizon of 4 weeks, a total wealth tracker, and a recursive update for wealth (in dollar value) using samples from the previously created matrix. Each block was repeated 5,000 times to emulate a variety of futures for each simulation. Lastly I calculated the value at risk for each portfolio at the 5% level.

Results:

Portfolio A:

```
## [1] "DVY"  "VIG"  "SPLV"
```

```
##        5%
## -8208.659
```

```
##
## Average return of investement after 4 weeks 100920.8
```

$8,268.17 is the four week value at risk for *Portfolio A* at the 5% level and the average return of investment for the same time period is $100,825.10.

Portfolio B:

```
## [1] "EFG"  "GOEX" "IBUY"
```

```
##        5%
## -190568.9
```

```
##
## Average return of investement after 4 weeks 906530.7
```

$186,696.30 is the four week value at risk for *Portfolio B* at the 5% level and the average return of investment for the same time period is $908,433.20.

Portfolio C:

```
## [1] "VWO" "GNR" "SCZ"
```

```
##       5%
## 709431.1
```

```
##
## Average return of investement after 4 weeks 90324.9
```

$18,233.44 is the four week value at risk for *Portfolio C* at the 5% level and the average return of investment for the same time period is $90,463.90.

Conclusion:

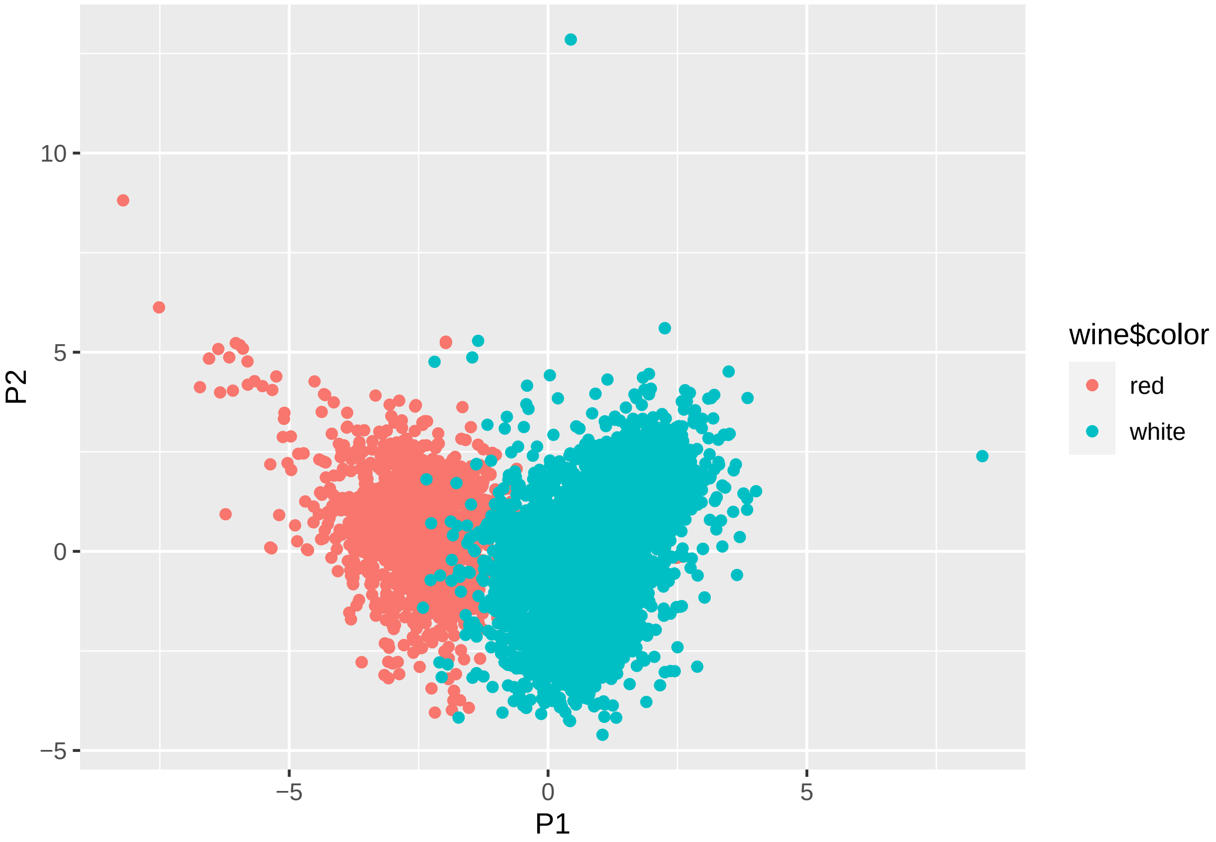
Since this was one of my first exposures to portfolio modeling, I was interested in exploring the difference in portfolio types and built models that could garner that exposure. The defensive, *Portfolio A* had a more conservative value at risk and a decent average return on investment. On the other hand, *Portfolios B and C* had greater values for both VaR and ROI which made sense given the high risk ETFs used in the portfolios. Moreover, it was interesting to see just how aggressive Portfolio B had performed since its VaR and ROI were of a dramatically different magnitude. In order to maintain some unity when modeling, each stock was weighted at 0.3 to use 90% of the initial wealth. I would be interested in experimenting with different weights for each portfolio in further analysis.

Clustering and PCA

Run both PCA and a clustering algorithm of your choice on the 11 chemical properties (or suitable transformations thereof) and summarize your results. Which dimensionality reduction technique makes more sense to you for this data? Convince yourself (and me) that your chosen method is easily capable of distinguishing the reds from the whites, using only the "unsupervised" information contained in the data on chemical properties. Does your unsupervised technique also seem capable of distinguishing the higher from the lower quality wines?

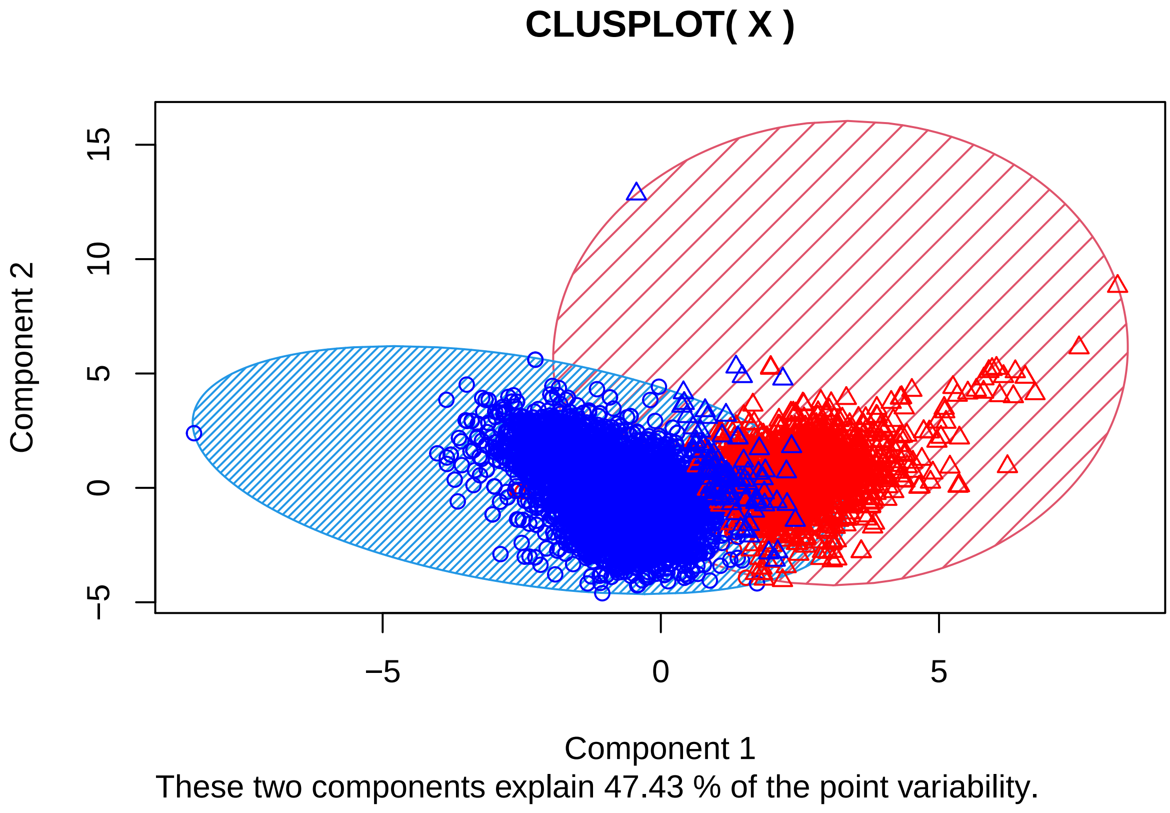
PCA Algorithim and Results:

```
##                  metric         PC1        PC2
## 1         fixed.acidity -0.25692873  0.2618431
## 2      volatile.acidity -0.39493118  0.1051983
## 3           citric.acid  0.14646061  0.1440935
## 4        residual.sugar  0.31890519  0.3425850
## 5             chlorides -0.31344994  0.2697701
## 6   free.sulfur.dioxide  0.42269137  0.1111788
## 7  total.sulfur.dioxide  0.47441968  0.1439475
## 8               density -0.09243753  0.5549205
## 9                    pH -0.20806957 -0.1529219
## 10            sulphates -0.29985192  0.1196342
## 11              alcohol -0.05892408 -0.4927275
## 12              quality  0.08747571 -0.2966009
```



Clustering Algorithim and Results:

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides
## 1    -0.2833598       -0.4002457   0.1163397      0.2036791 -0.3133076
## 2     0.8203503        1.1587448  -0.3368131     -0.5896682  0.9070517
##   free.sulfur.dioxide total.sulfur.dioxide    density         pH  sulphates
## 1           0.2875184            0.4055994 -0.2316456 -0.1918762 -0.2847052
## 2          -0.8323898           -1.1742444  0.6706333  0.5554976  0.8242456
##       alcohol     quality
## 1  0.02984993  0.09700235
## 2 -0.08641804 -0.28082994
```



Conclusion:

Although both PCA and clustering algorithms were able to distinguish red from white wines and the higher from lower quality wines, the clustering dimensionality reduction technique makes more sense for this data since it had a clearer distinction between the red and white wines and is thus easily capable of differentiating between both wines.

Market Segmentation

Your task to is analyze this data as you see fit, and to prepare a concise report for NutrientH20 that identifies any interesting market segments that appear to stand out in their social-media audience. You have complete freedom in deciding how to pre-process the data and how to define "market segment." (Is it a group of correlated interests? A cluster? A latent factor? Etc.) Just use the data to come up with some interesting, well-supported insights about the audience, and be clear about what you did.

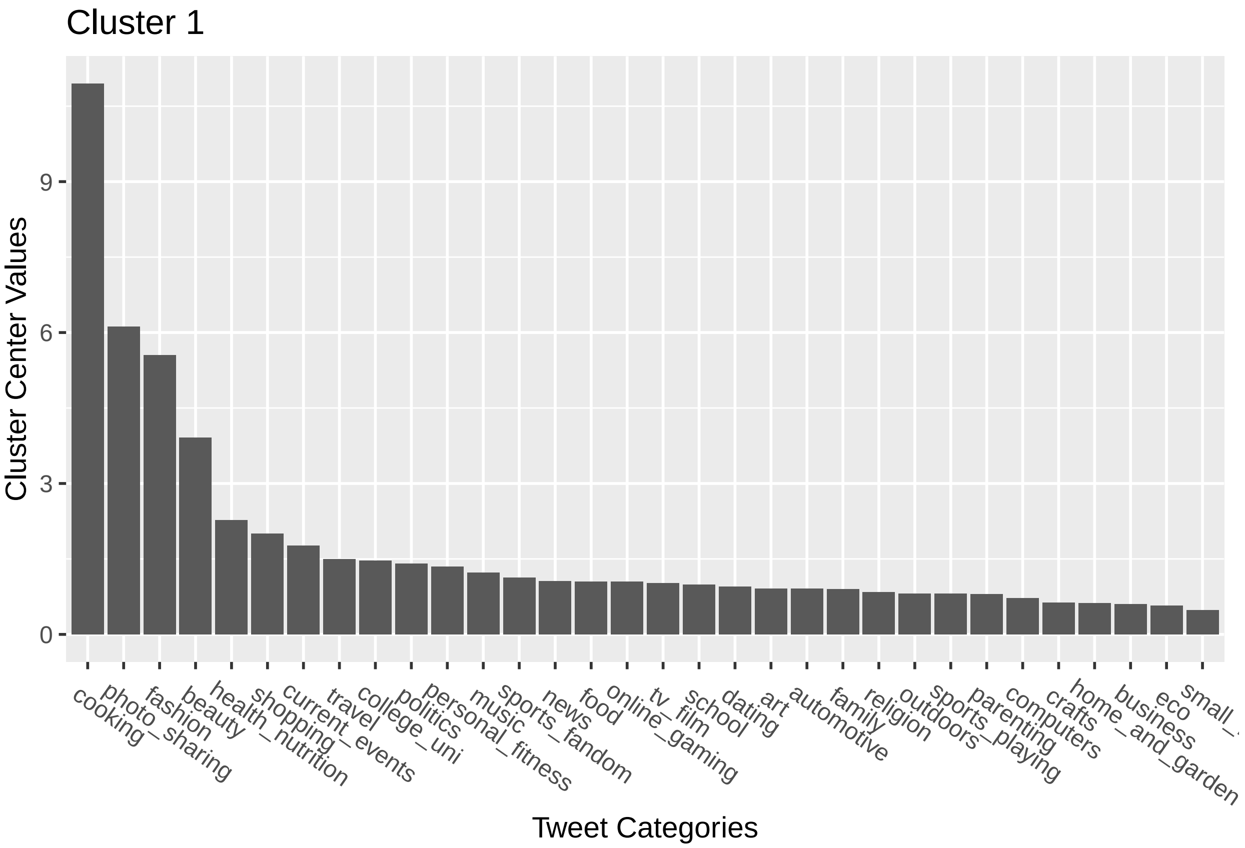
Question:

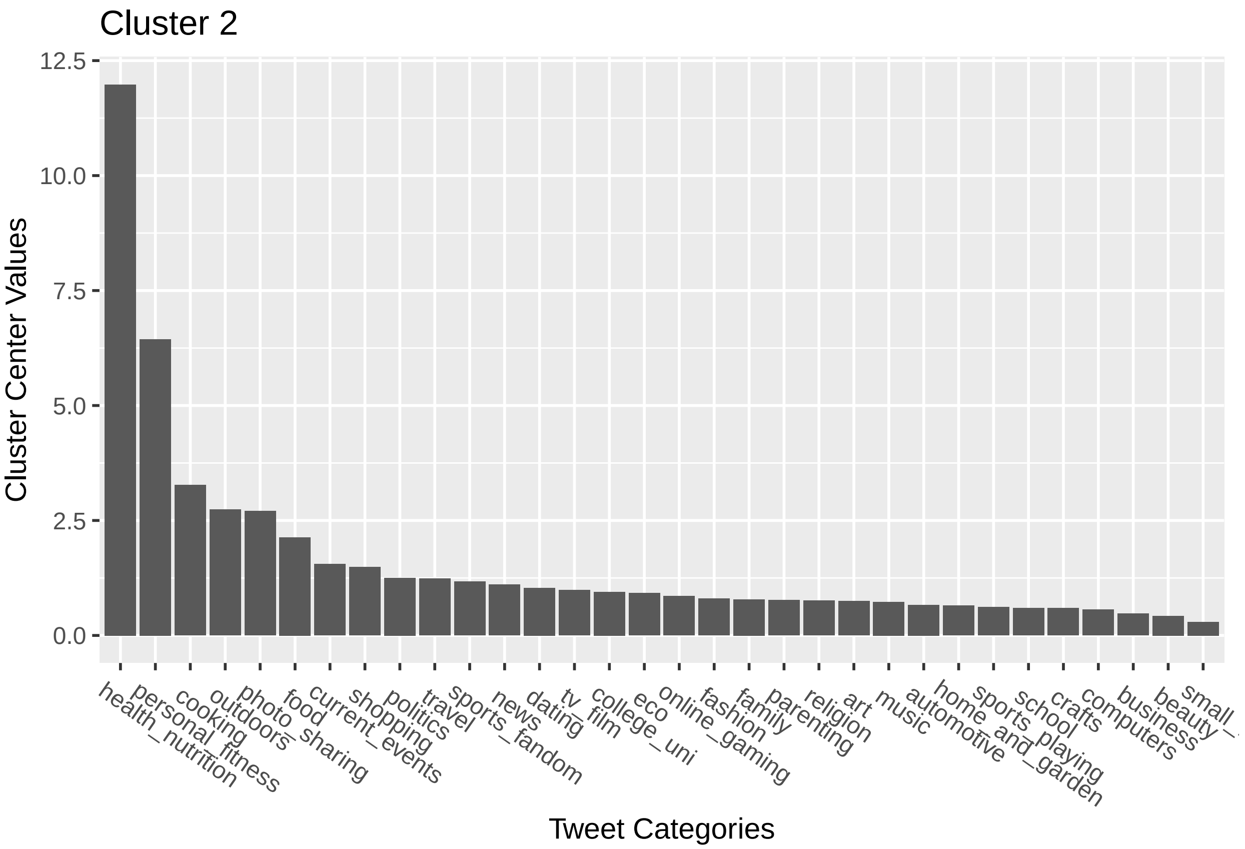
What interesting market segments appear to stand out in NutrientH20's social-media audience?

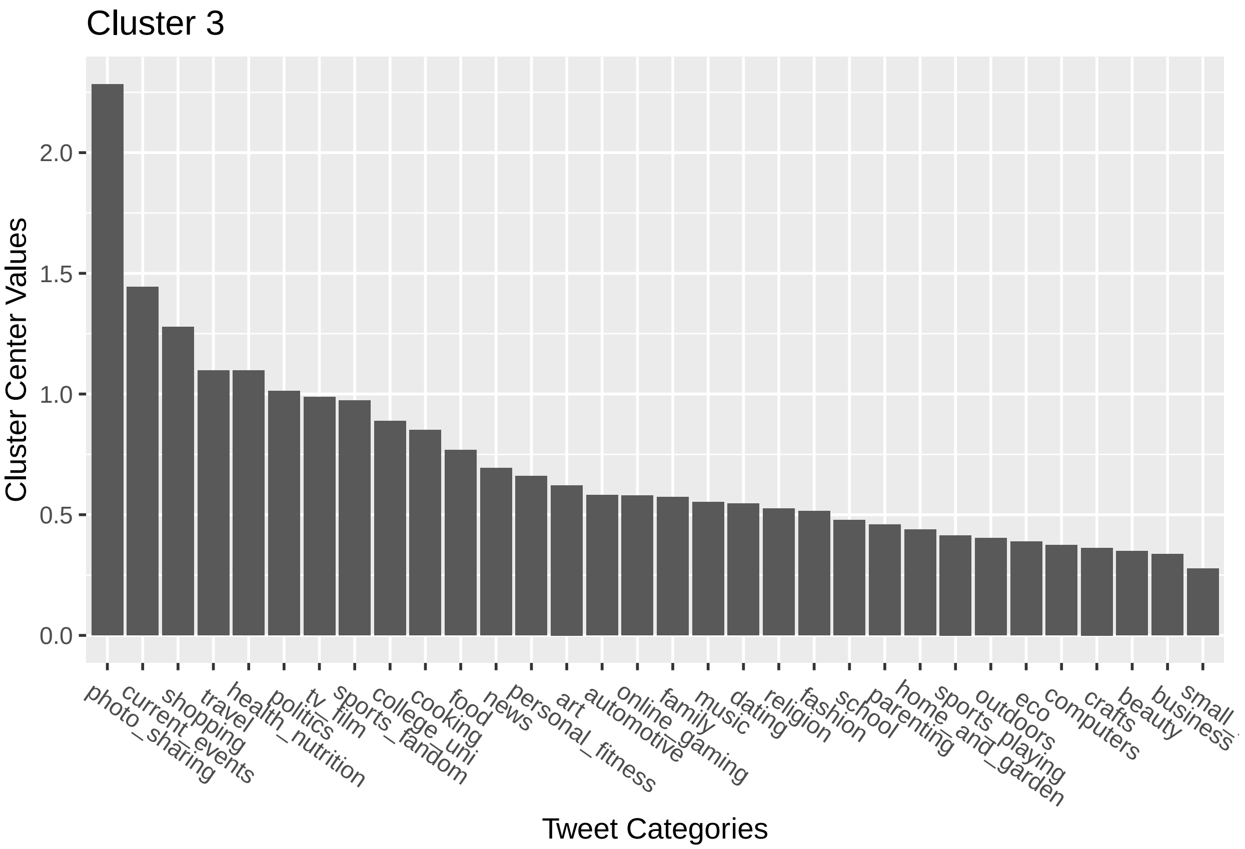
Approach:

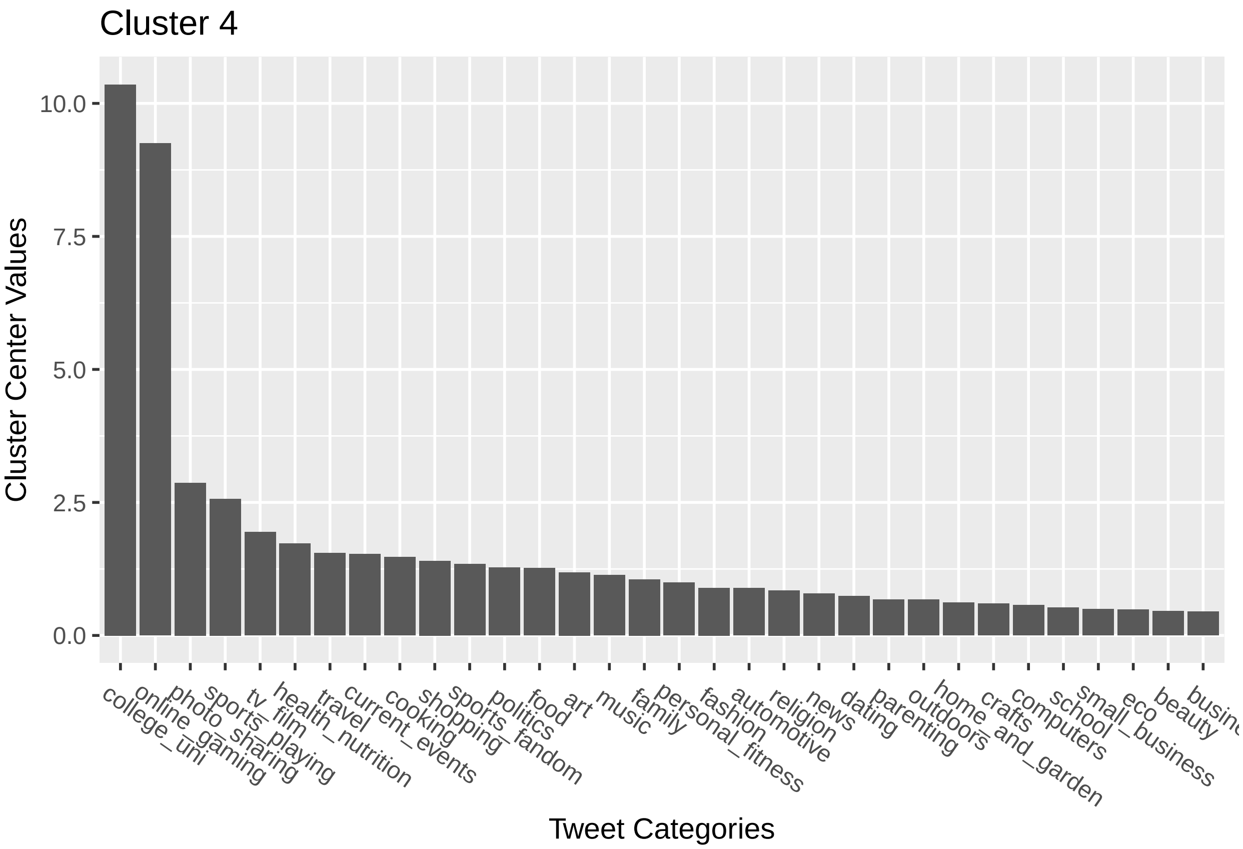
In order to identify interesting market segments, first we will preprocess the data to reduce noise and scale the data set before moving into determining the number of clusters using K means clustering methods.

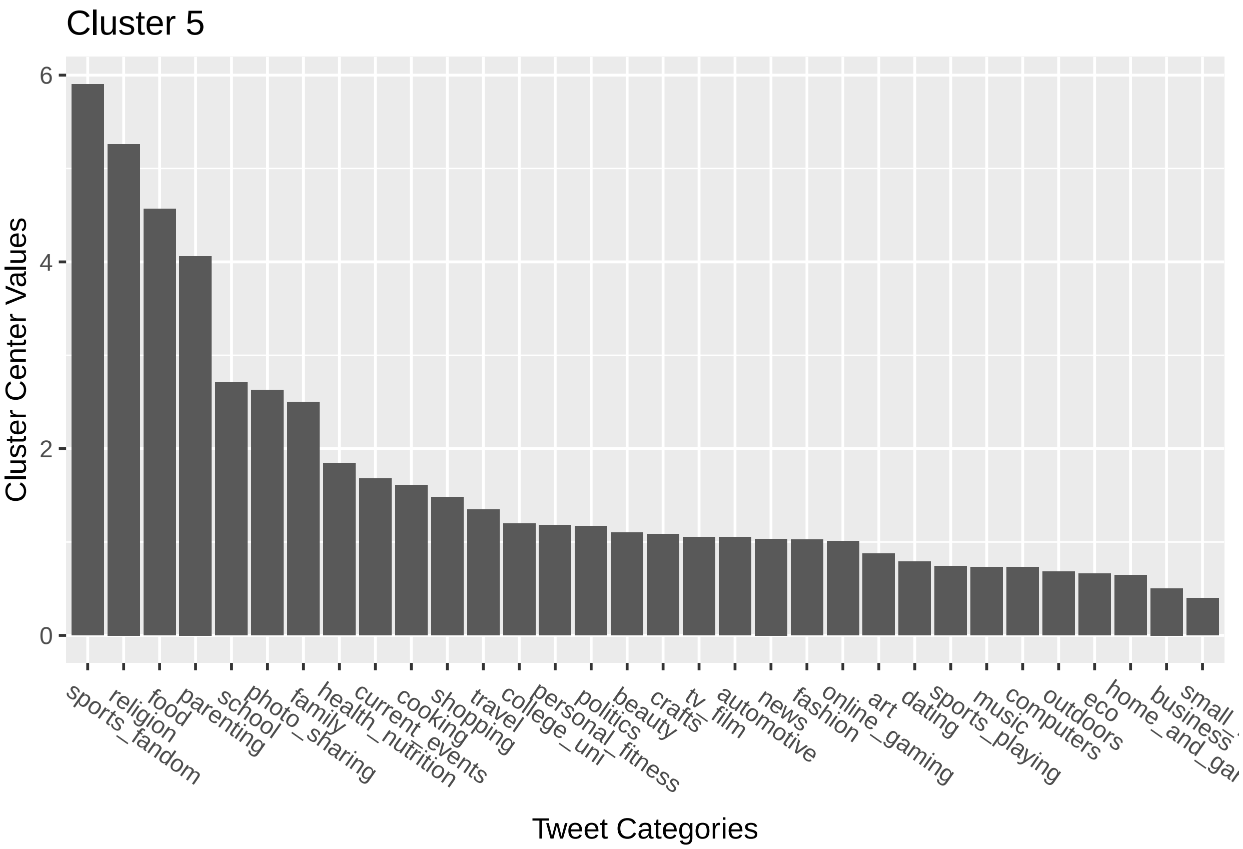
Results:

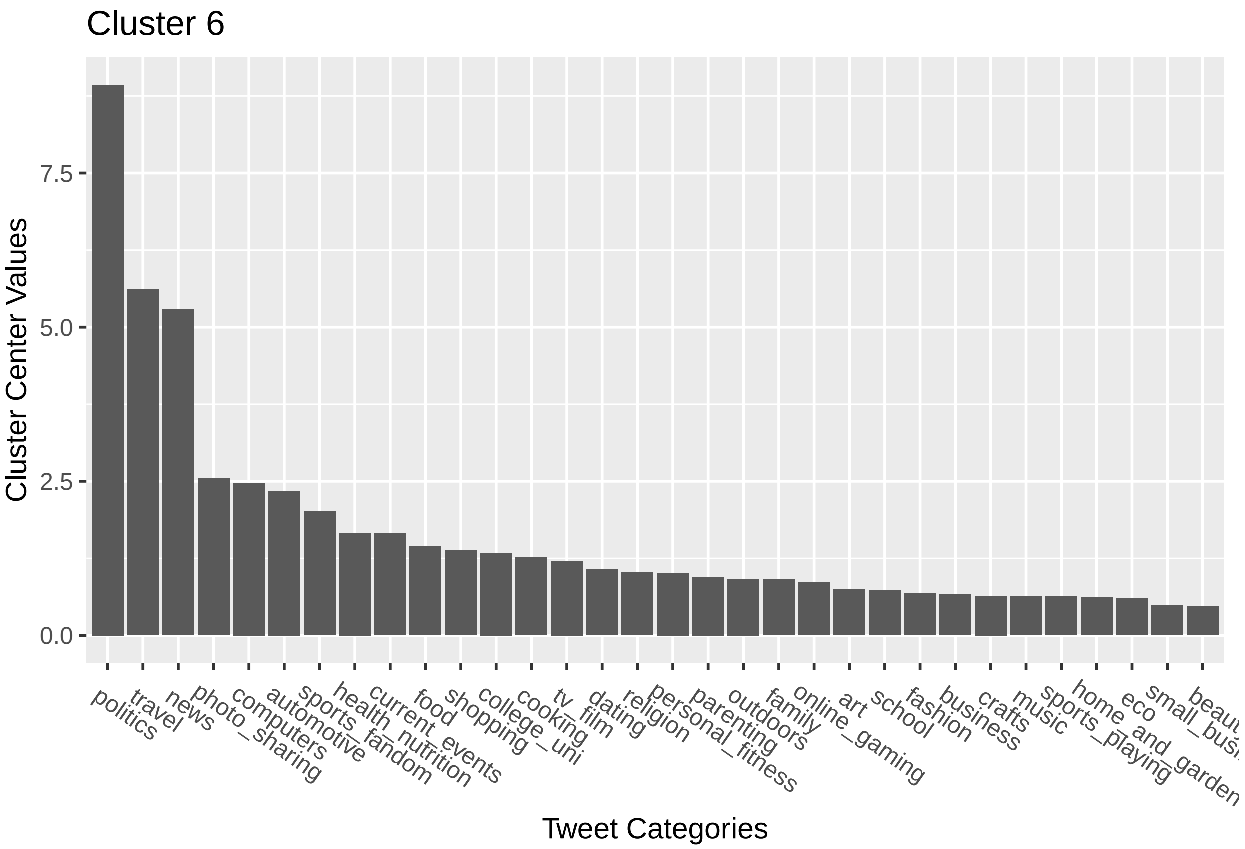

Cluster 1
Cluster Center Values
0
3
6
9
cooking
photo_sharing
fashion
beauty
health_nutrition
shopping
current_events
travel
college_uni
politics
personal_fitness
music
sports_fandom
news
food
online_gaming
tv_film
school
dating
art
automotive
family
religion
outdoors
sports_playing
parenting
computers
crafts
home_and_garden
business
eco
small_
Tweet Categories


Cluster 2
Cluster Center Values
12.5
10.0
7.5
5.0
2.5
0.0
health_nutrition
personal_fitness
cooking
outdoors
photo_sharing
food
current_events
shopping
politics
travel
sports_fandom
news
dating
tv_film
college_uni
eco
online_gaming
fashion
family
parenting
religion
art
music
automotive
home_and_garden
sports_playing
school
crafts
computers
business
beauty
small_
Tweet Categories


Cluster 3
Cluster Center Values
2.0
1.5
1.0
0.5
0.0
photo_sharing
current_events
shopping
travel
health_nutrition
politics
tv_film
sports_fandom
college_uni
cooking
food
news
personal_fitness
art
automotive
online_gaming
family
music
dating
religion
fashion
school
parenting
home_and_garden
sports_playing
outdoors
eco
computers
crafts
beauty
business
small_
Tweet Categories


Cluster 4
Cluster Center Values
10.0
7.5
5.0
2.5
0.0
college_uni
online_gaming
photo_sharing
sports_playing
tv_film
health_nutrition
travel
current_events
cooking
shopping
sports_fandom
politics
food
art
music
family
personal_fitness
fashion
automotive
religion
news
dating
parenting
outdoors
home_and_garden
crafts
computers
school
small_business
eco
beauty
busine
Tweet Categories


Cluster 5
Cluster Center Values
6
4
2
0
sports_fandom
religion
food
parenting
school
photo_sharing
family
health_nutrition
current_events
cooking
shopping
travel
college_uni
personal_fitness
politics
beauty
crafts
tv_film
automotive
news
fashion
online_gaming
art
dating
sports_playing
music
computers
outdoors
eco
home_and_gar
business
small_
Tweet Categories



* Identifying market segments by cluster:

- Cluster 1: Current Events, Photo sharing
- Cluster 2: Religion, Sports Fandom
- Cluster 3: Politics, Travel
- Cluster 4: Health/Nutrition, Personal Fitness
- Cluster 5: Cooking, Photo Sharing
- Cluster 6: College/Uni, Online Gaming

Conclusion:

Through K means clustering, six market segments were identified as outlined in the above bullet points. Here, "market segment" is defined as the the tweet categories with the largest cluster center values. There are some obvious segments such as Cluster 4 that deals with primarily health and fitness, which may be a social audience NutrientH20 would be keen on advertising to. On the other hand a segment such as Cluster 6, which may have emerged more recently given general video game demographic trends, would require further research as advertising to this audience would be more intuitive.

The Reuters Corpus

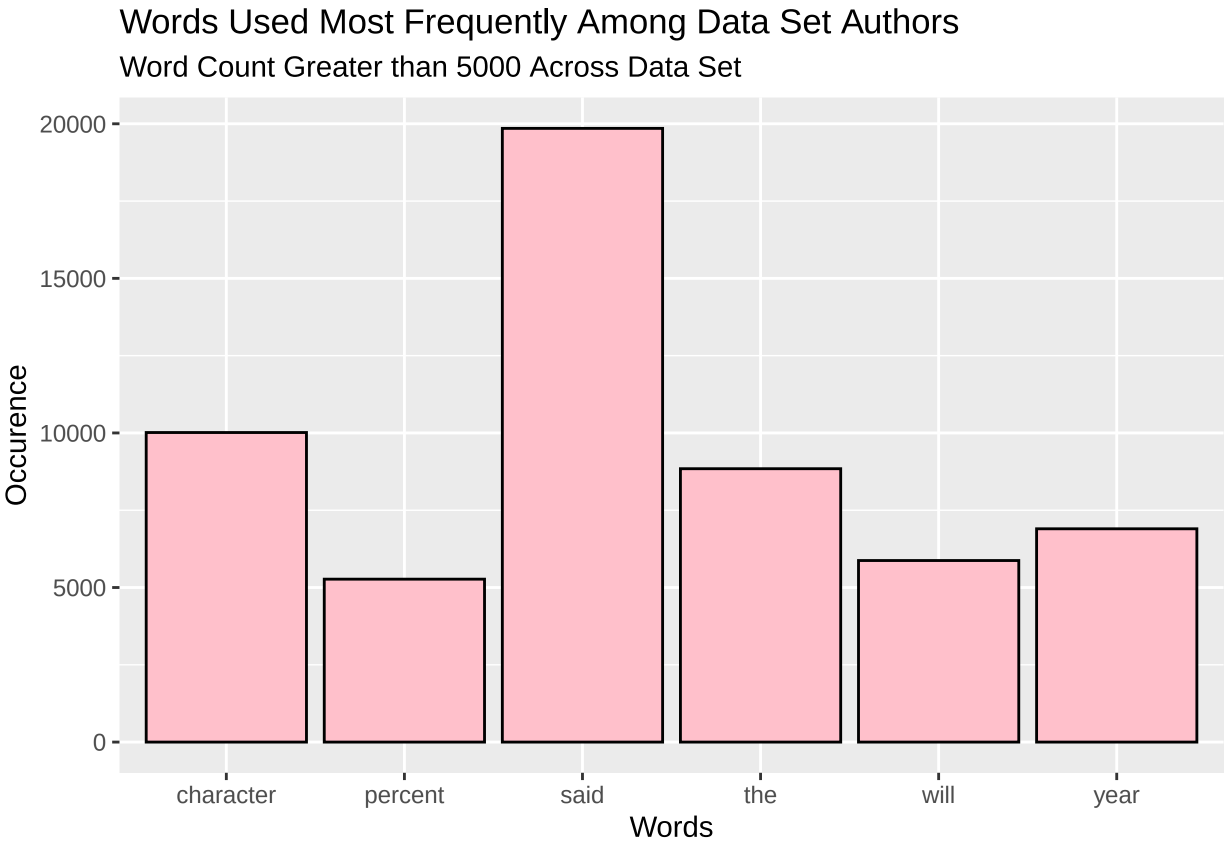
Revisit the Reuters C50 text corpus that we briefly explored in class. Your task is simple: tell an interesting story, anchored in some analytical tools we have learned in this class, using this data. Describe clearly what question you are trying to answer, what models you are using, how you pre-processed the data, and so forth. Make sure you include at least one really interesting plot (although more than one might be necessary, depending on your question and approach.)

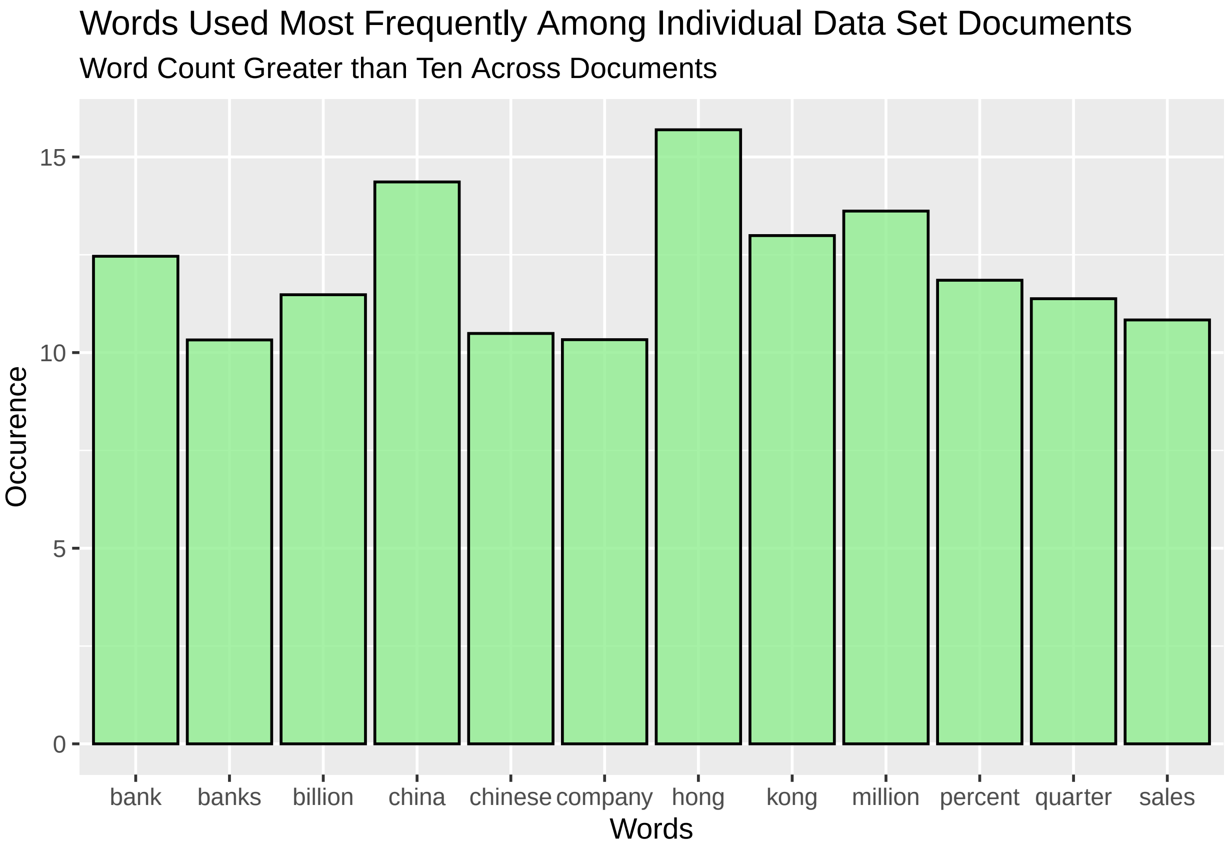
Question:

Are the most frequently used words in the data set among authors the same words as those found most commonly among individual documents?

Approach:

In order to determine the words that were used most frequently across the data set itself, the Document-Term-Matrix was utilized, and the inverse document frequency was used to determine the words that showed up most often within the documents.





### Results:

```
##          name count
##  1:      said 19851
##  2: character 10013
##  3:       the  8843
##  4:      year  6899
##  5:      will  5873
##  6:   percent  5270
##  7:   million  4848
##  8:       new  3508
##  9:    market  3215
## 10:   company  3200
```

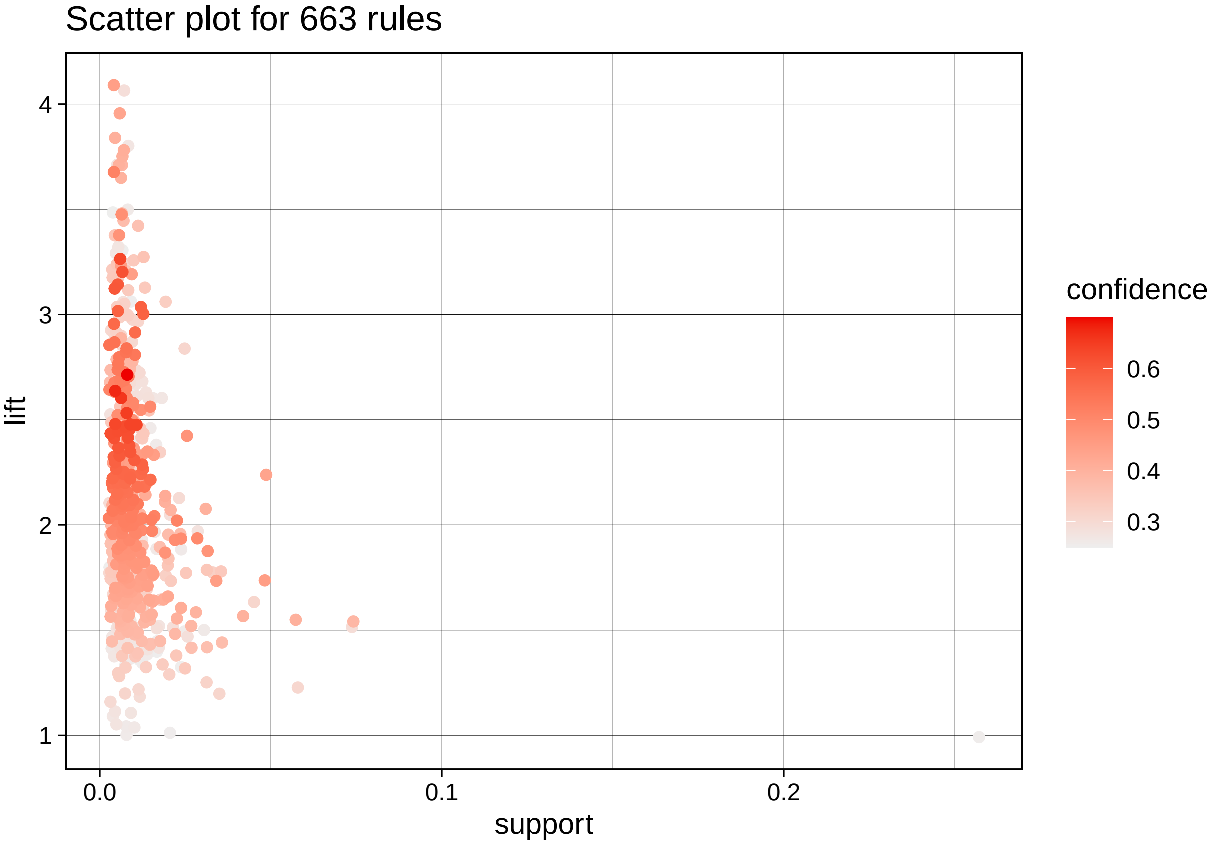
```
##        name    count
##  1:    hong 15.69451
##  2:   china 14.36147
##  3: million 13.61660
##  4:    kong 12.99058
##  5:    bank 12.46214
##  6: percent 11.84928
##  7: billion 11.47878
##  8: quarter 11.37724
##  9:   sales 10.83425
## 10: chinese 10.48939
```

Conclusion:

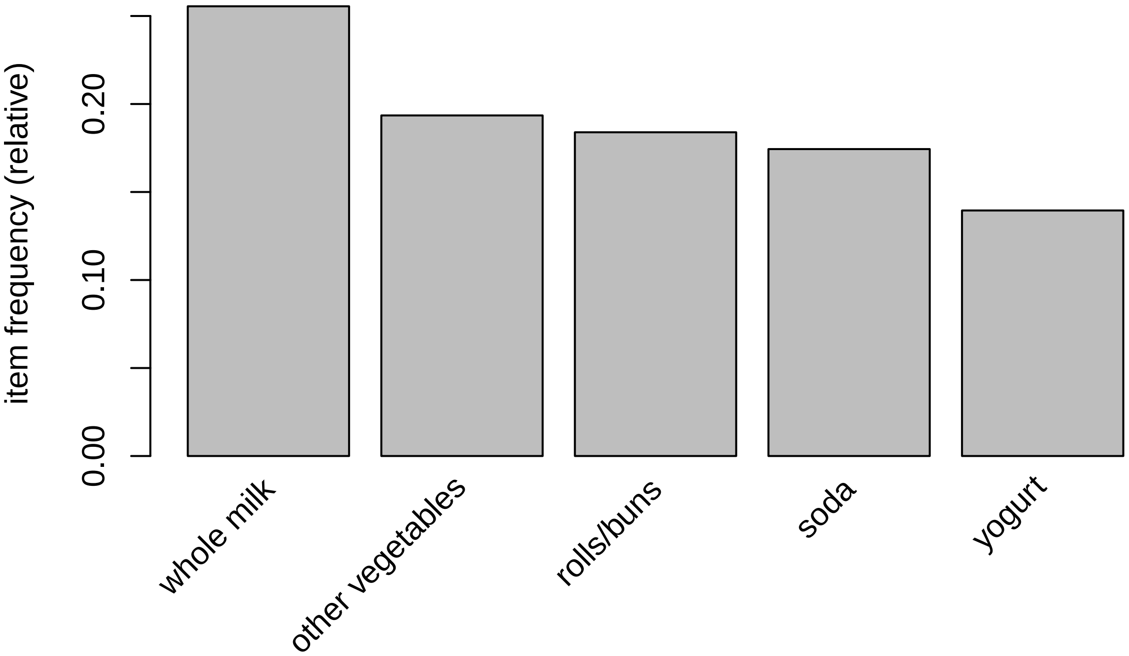
Ultimately, it is interesting to conclude that the words used most often within individual documents doesn't necessarily translate to the words used most often among the data set as whole by all of the authors. Rather general terms are frequent in the data, such as "said", "the", and "percent", but more specific words are found in each document such as locations and languages.

Association rule mining

Use the data on grocery purchases in groceries.txt and find some interesting association rules for these shopping baskets. The data file is a list of shopping baskets: one person's basket for each row, with multiple items per row separated by commas. Pick your own thresholds for lift and confidence; just be clear what these thresholds are and say why you picked them. Do your discovered item sets make sense? Present your discoveries in an interesting and visually appealing way.

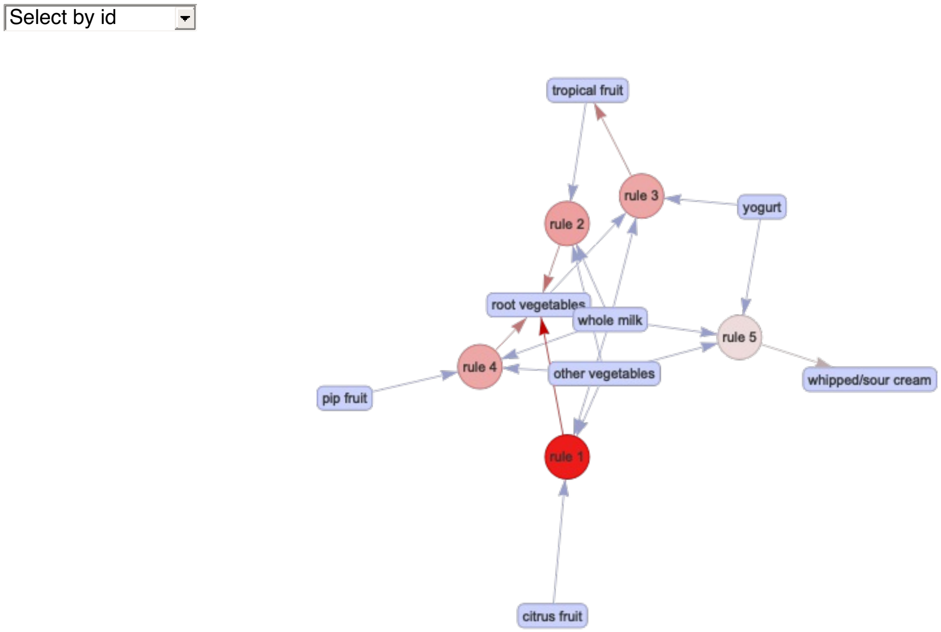


The above scatter plot is a visual guide in determining our lift and confidence values. Based on the plot, it seems that a lift of at least greater than 2.5 and a confidence level greater than 0.5 could aid in identifying association rules.



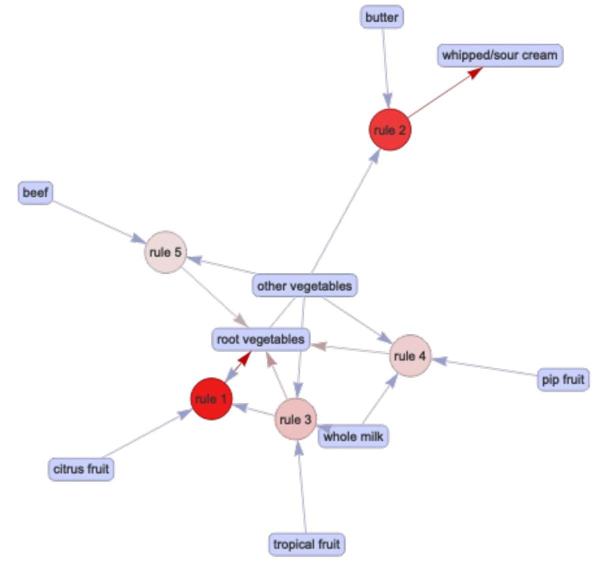
The above bar plot outlines the five grocery items that occur most frequently in the data. I will choose to focus on these variables when looking at association rules.

Plots of the top five assoication rules for the top five grocery items based on lift



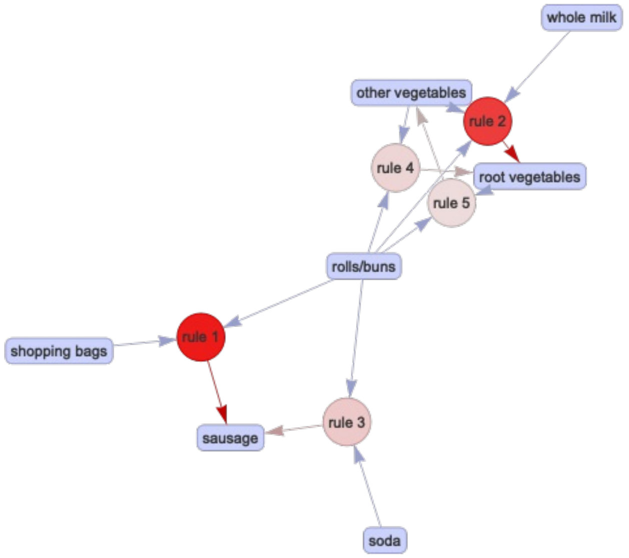
Whole milk seems to be most associated with other dairy items which is understandable given traditional grocery store layout structures as dairy products are typically placed in close distance of one another.

Select by id ▾

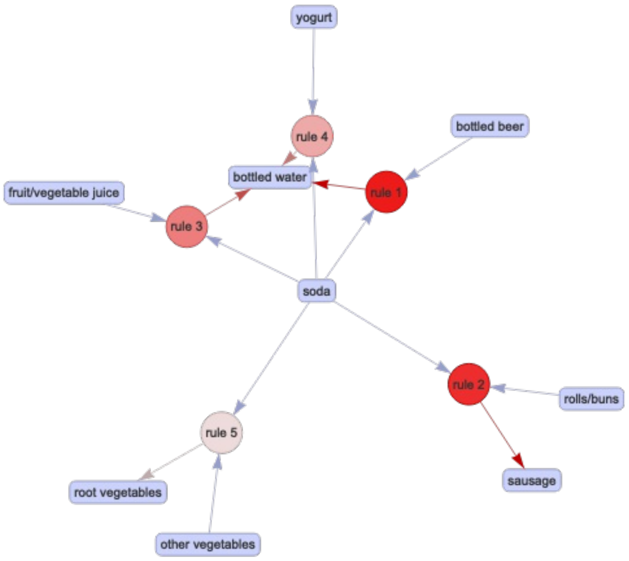


Other vegetables are mostly associated with what one may consider regular grocery items such as fruit, root vegetables, and beef.

Select by id ▾

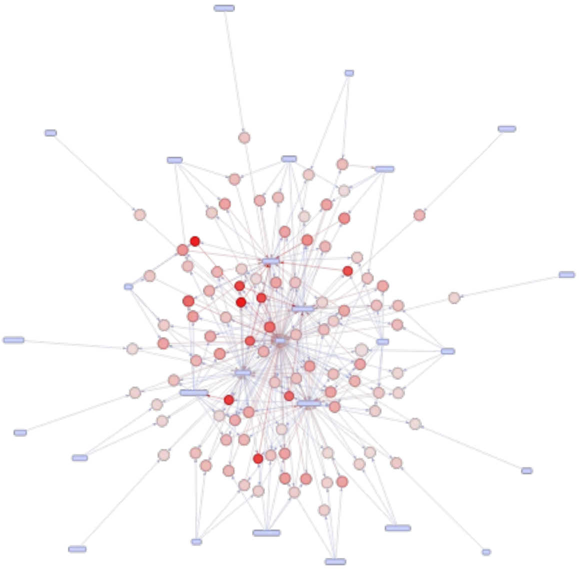


Select by id



Similar to whole milk, soda is mostly associated with other beverage products which can be attributed to the layout of grocery store.

Select by id



Yogurt is mostly associated with other dairy products such as whole milk and whipped cream along with fruits.

It seems that grocery store layout may play a role in determining what items go in a customer's grocery basket. In this way, for certain holidays or special events stores may consider placing certain grocery items together to increase purchases. For instance during hotter months stores could have a beverage display with cool drinks, since multiple beverages are being spotlighted close to one another it could act as an incentive for buyers.