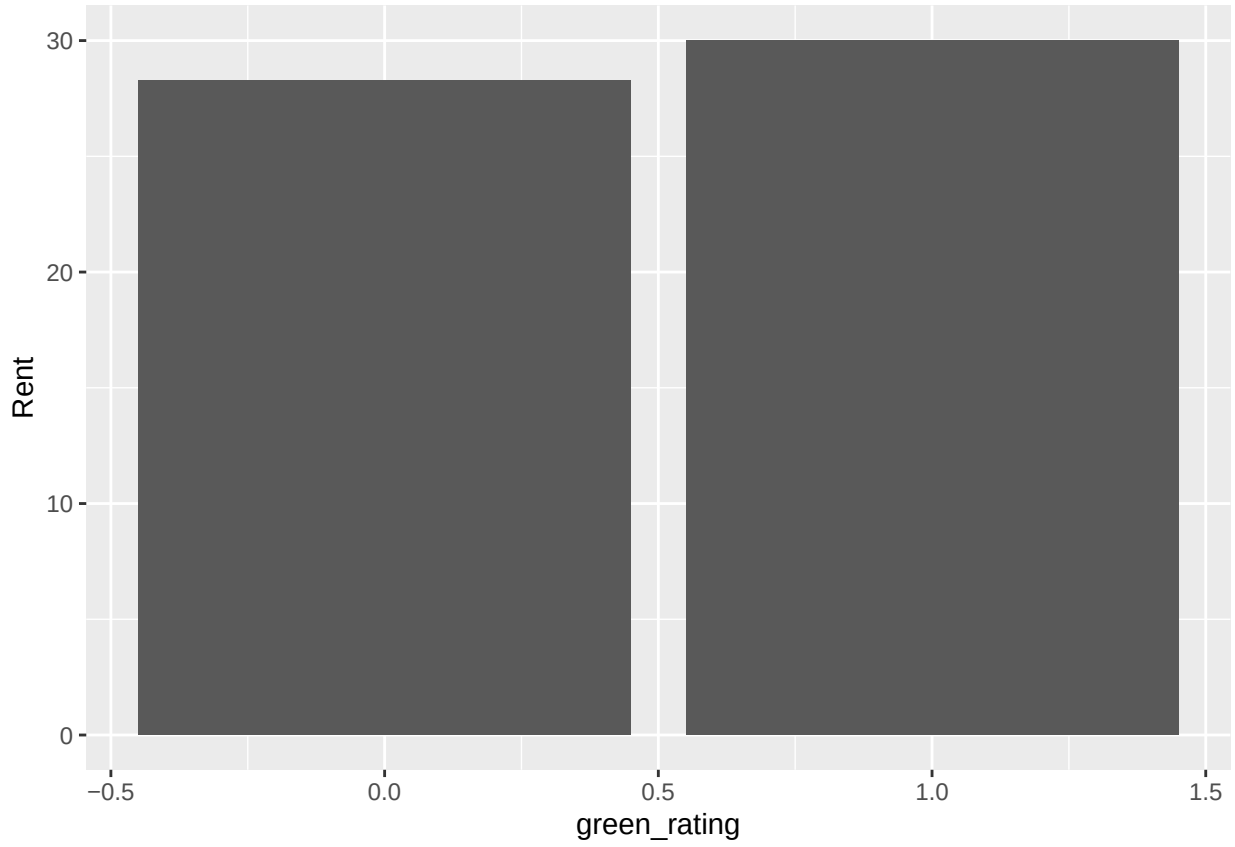
STA380 Exercises

8/16/2021

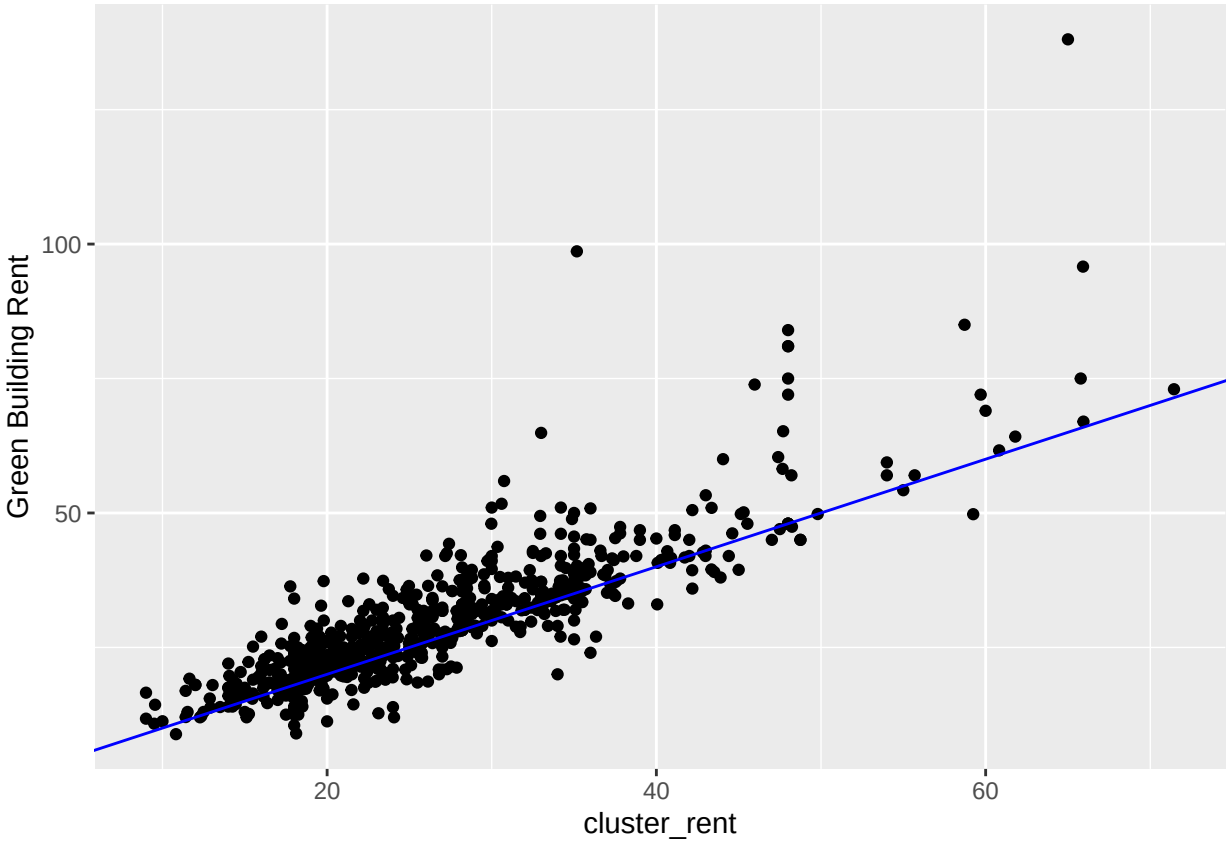
github link

1. Visual story telling part 1: green buildings



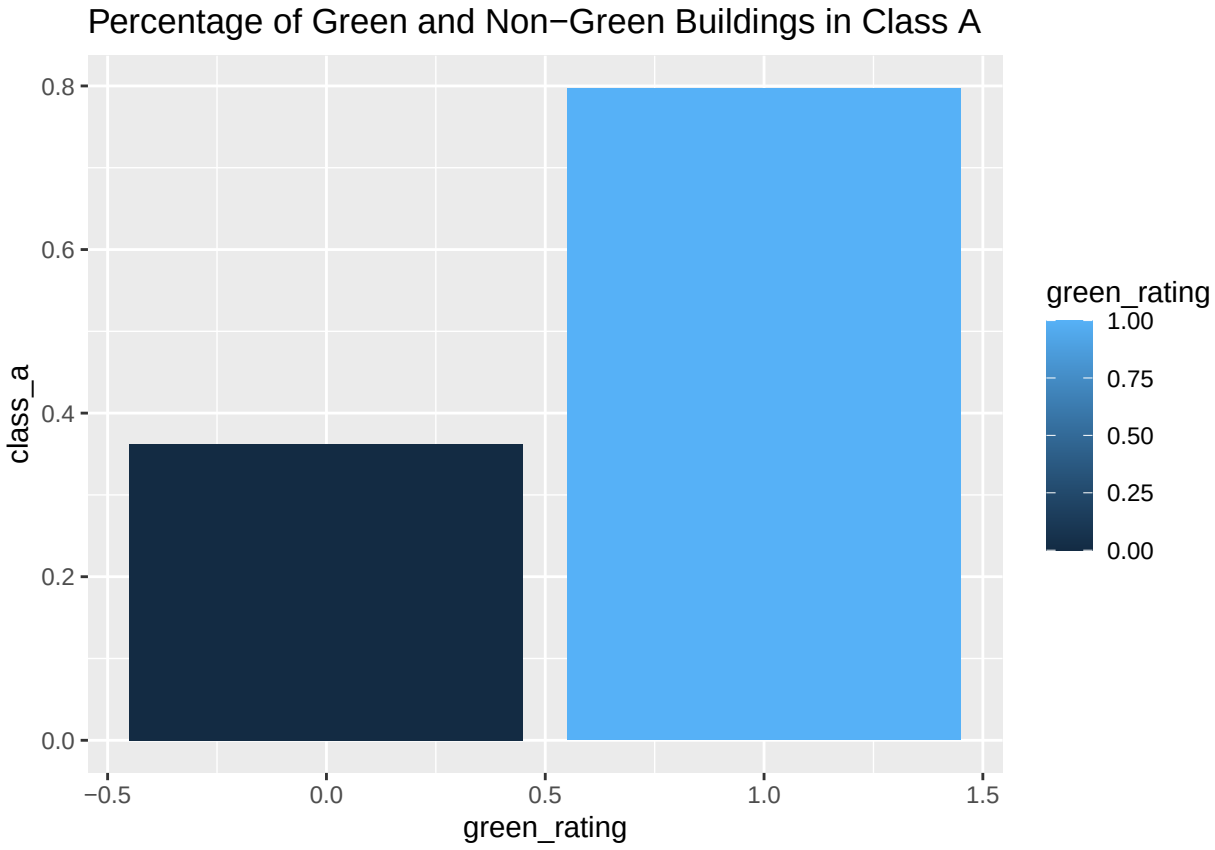
It's possible green buildings are usually built in more expensive areas, and are not more expensive themselves.

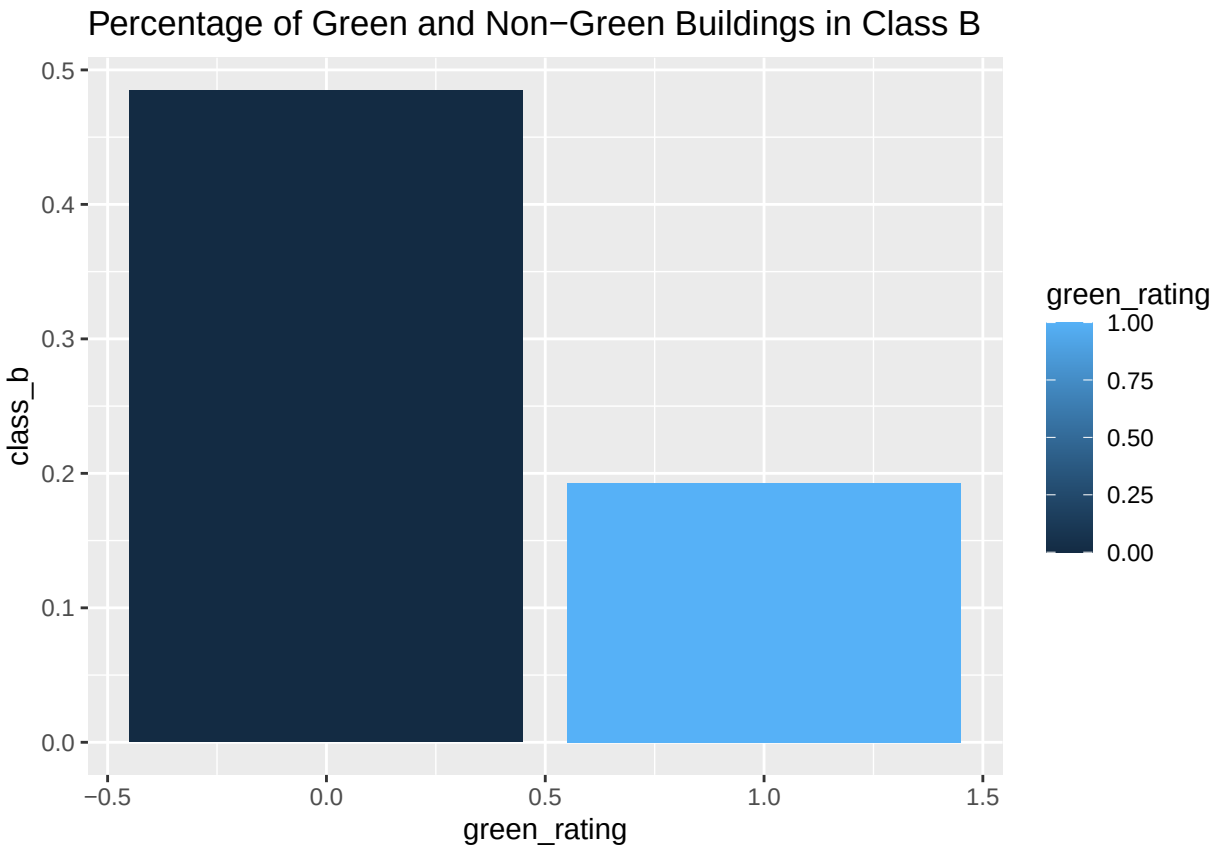
But when we graph it, we see green buildings are more expensive than their counterparts in the same cluster. If green buildings cost the same as their cluster, we would expect them to all be along that blue line, but we see most are above it.

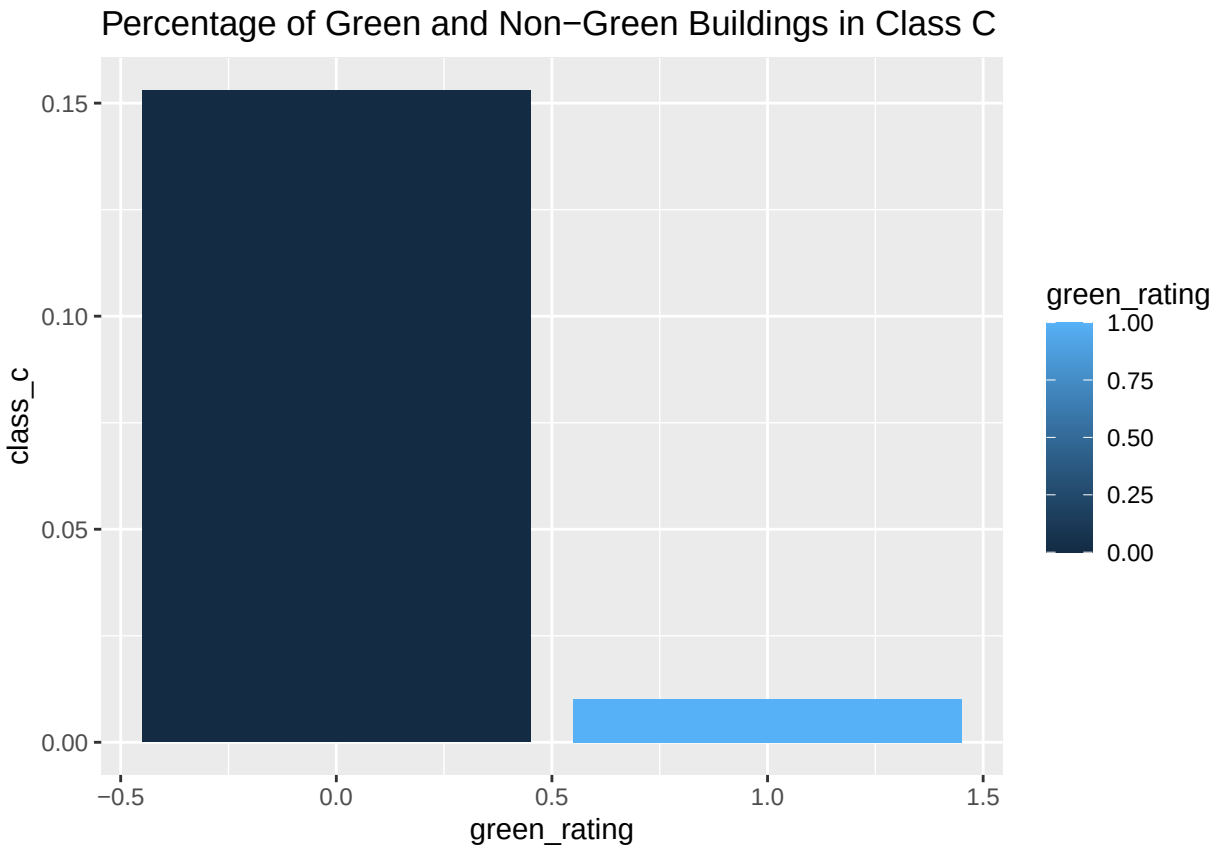


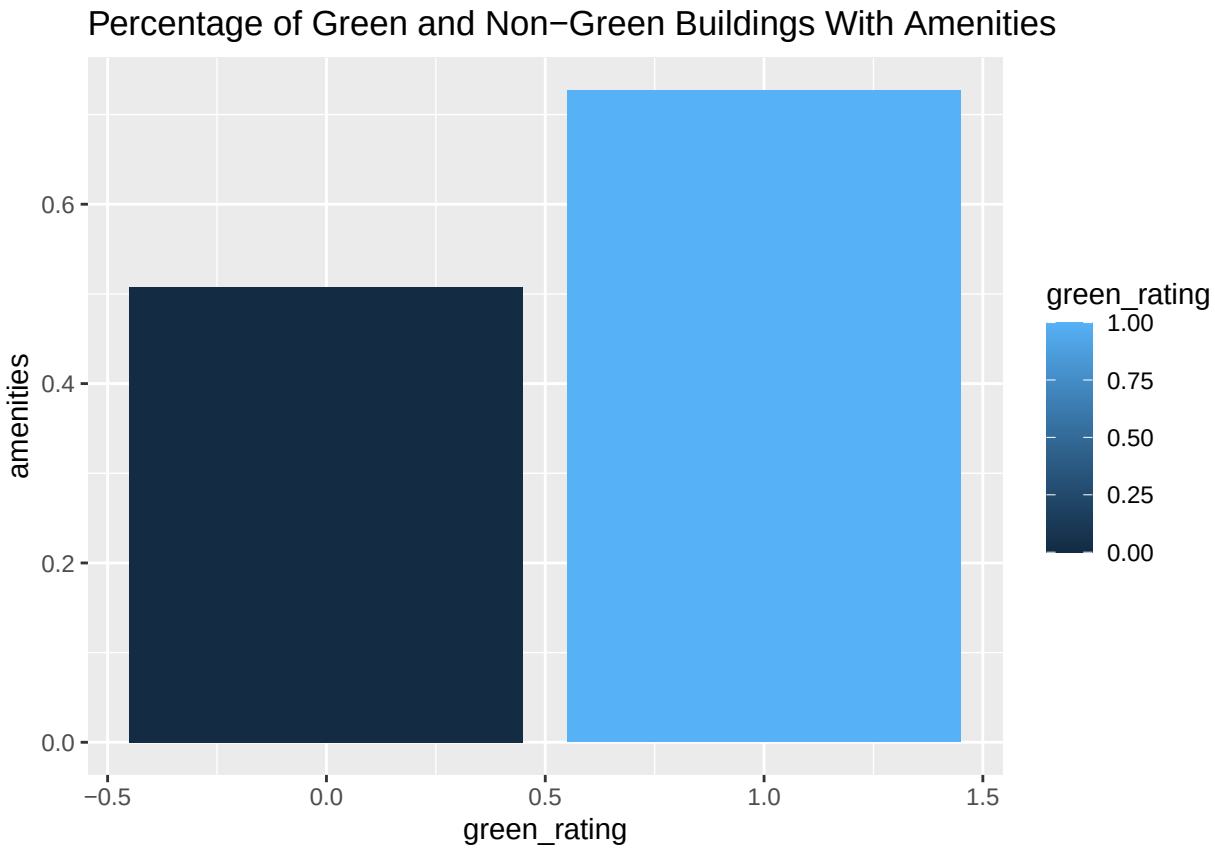
Are green buildings usually nicer than non-green buildings?

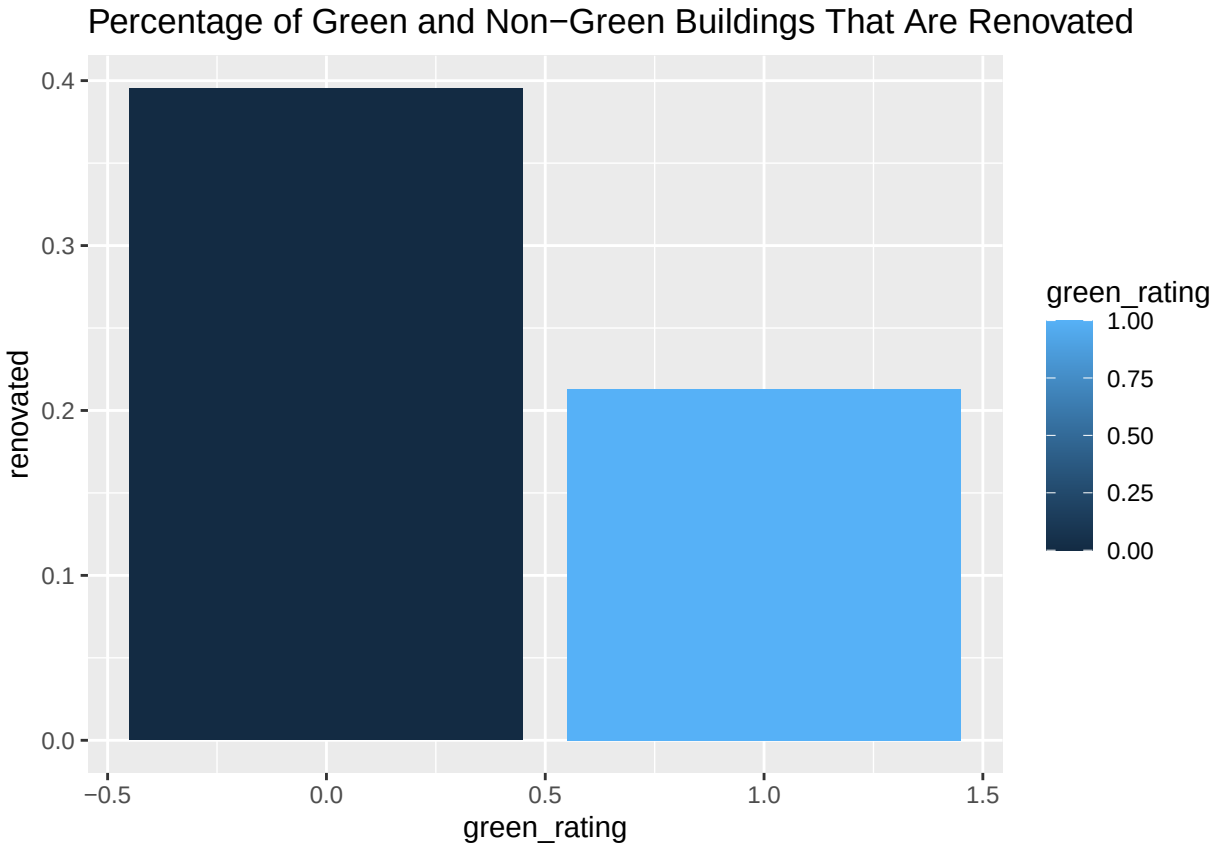
When we graph it, we see green buildings are disproportionately in the highest class of buildings class_a, and non-green buildings are more likely to be class_b or class_c. They are also more likely to have renovated, and are usually newer rather than renovated. Additionally, we can see that nicer buildings cost more to rent, as you'd expect.

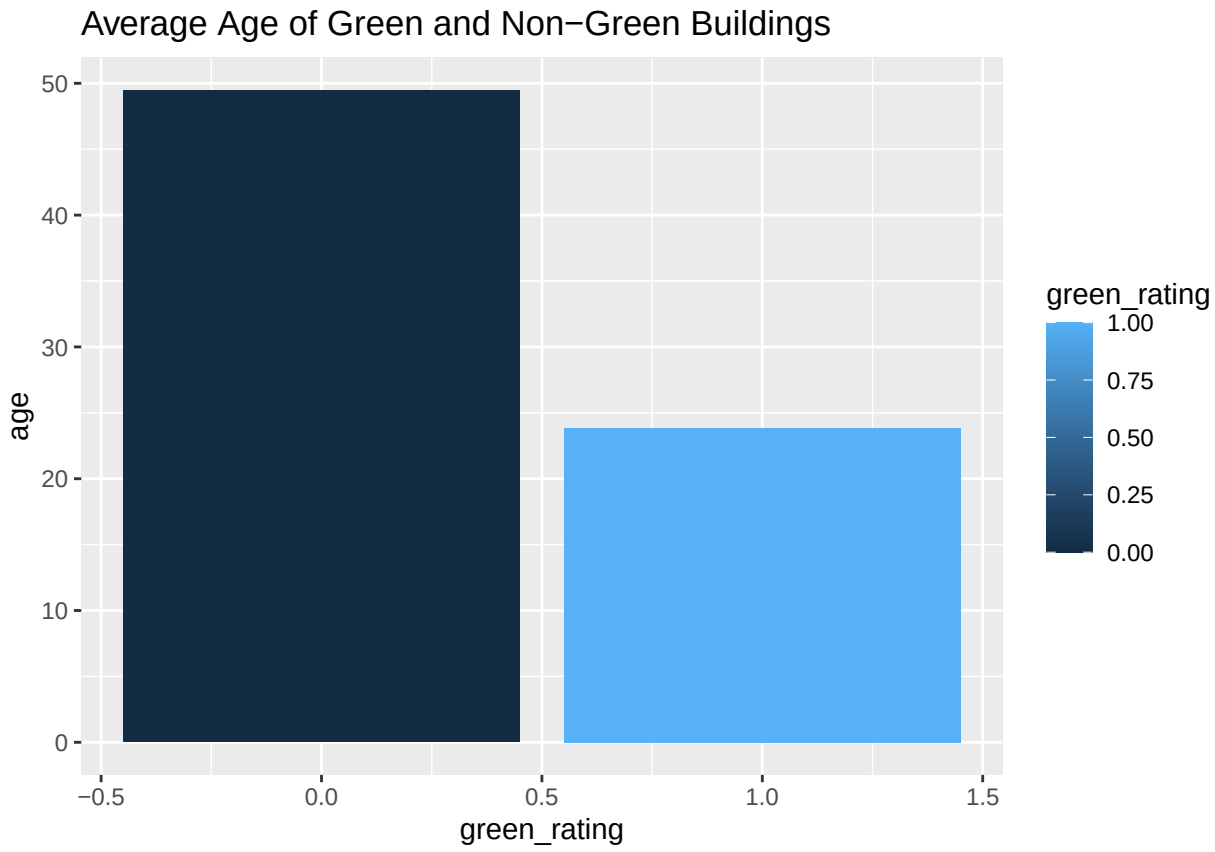

Percentage of Green and Non−Green Buildings in Class A
class_a
0.0
0.2
0.4
0.6
0.8
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00

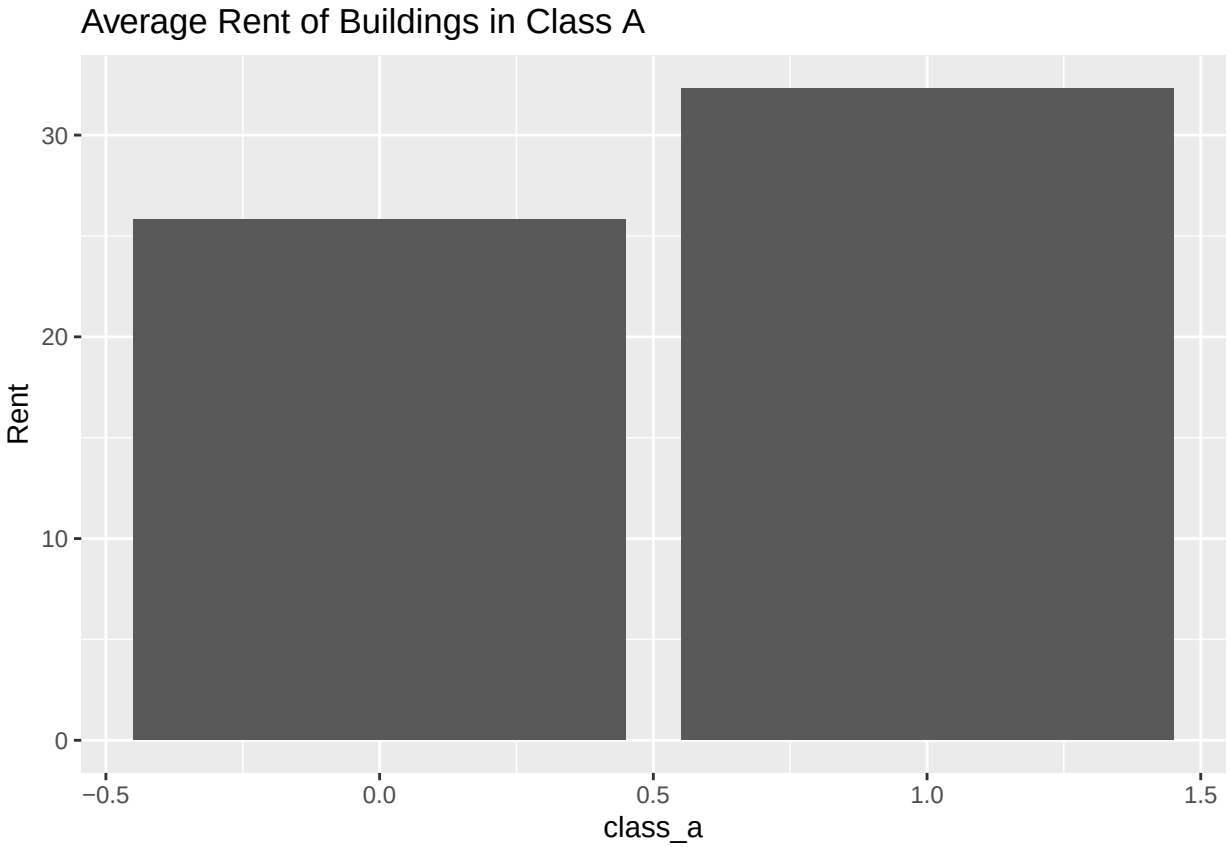

Percentage of Green and Non−Green Buildings in Class B
class_b
0.5
0.4
0.3
0.2
0.1
0.0
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00

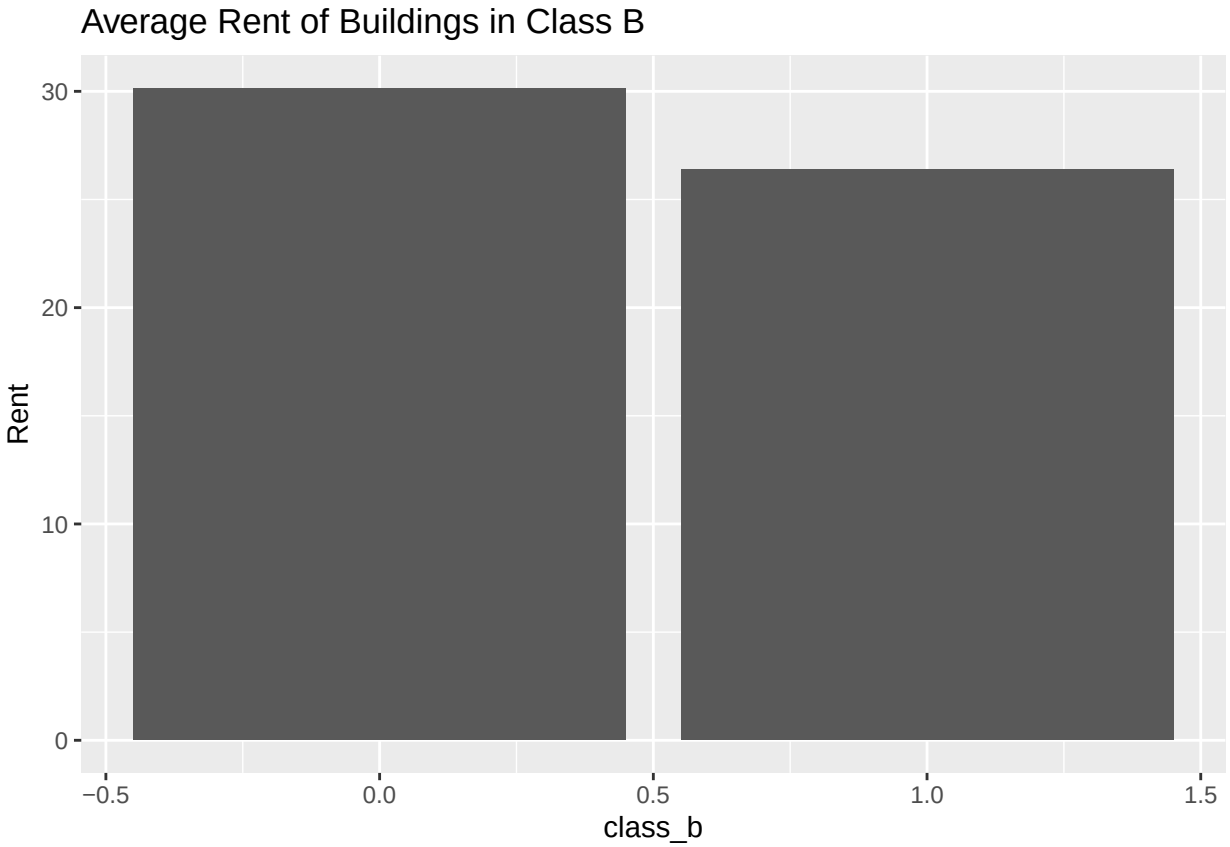

Percentage of Green and Non−Green Buildings in Class C
0.15
0.10
0.05
0.00
class_c
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00

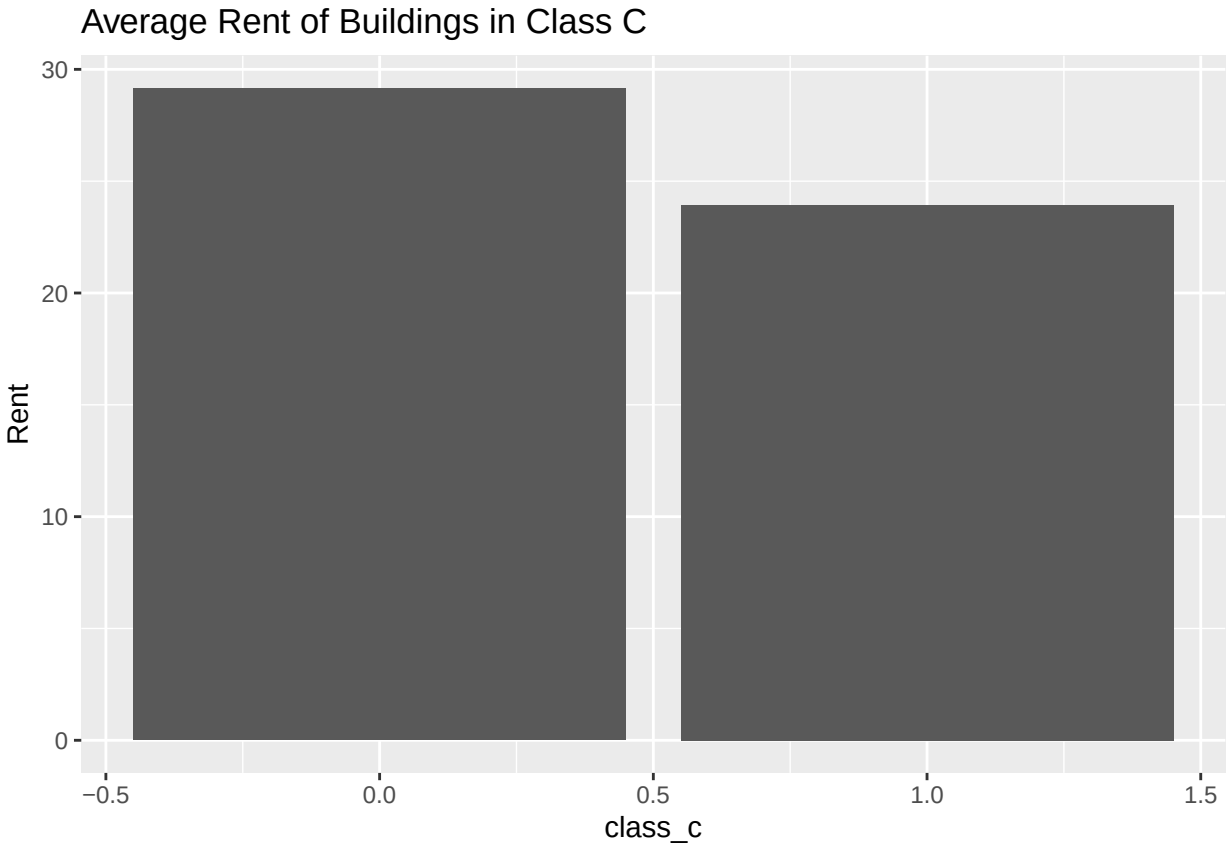

Percentage of Green and Non−Green Buildings With Amenities
amenities
0.6
0.4
0.2
0.0
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00

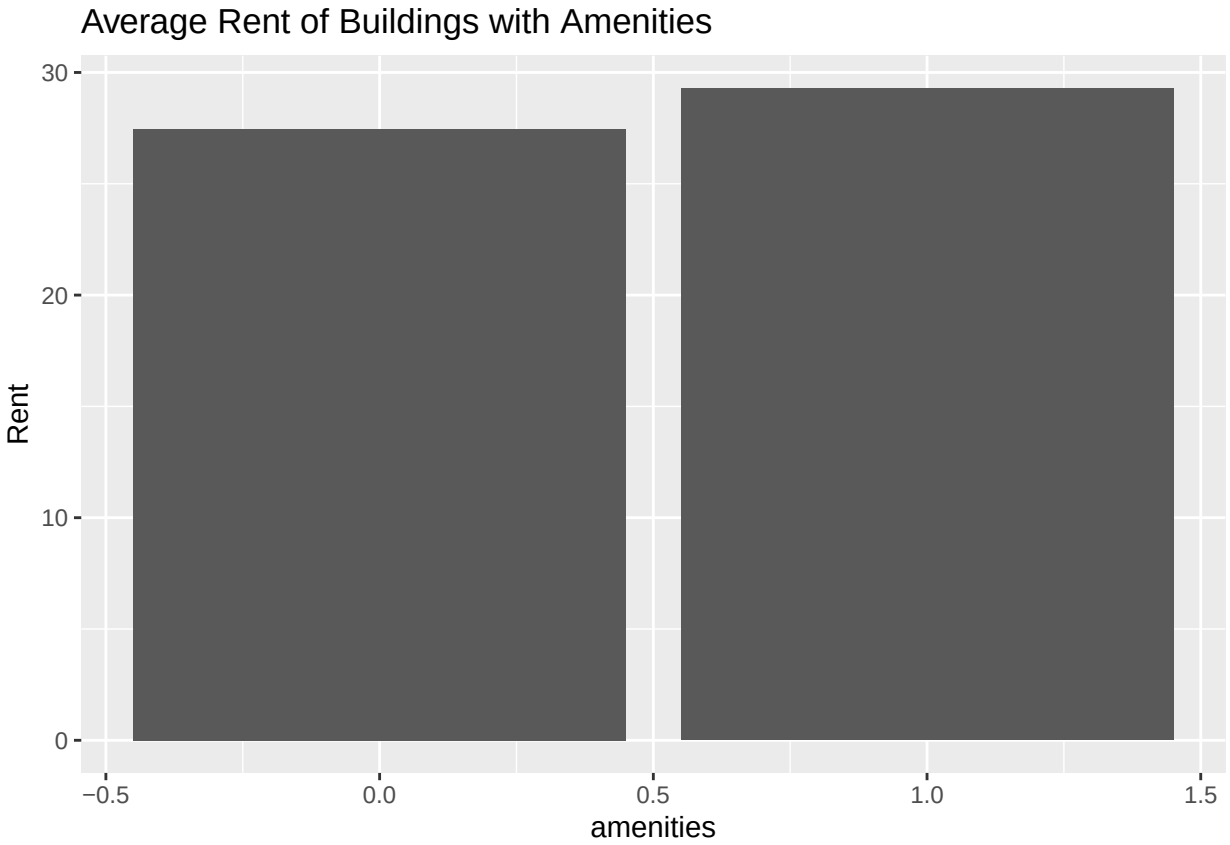

Percentage of Green and Non−Green Buildings That Are Renovated
renovated
0.4
0.3
0.2
0.1
0.0
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00

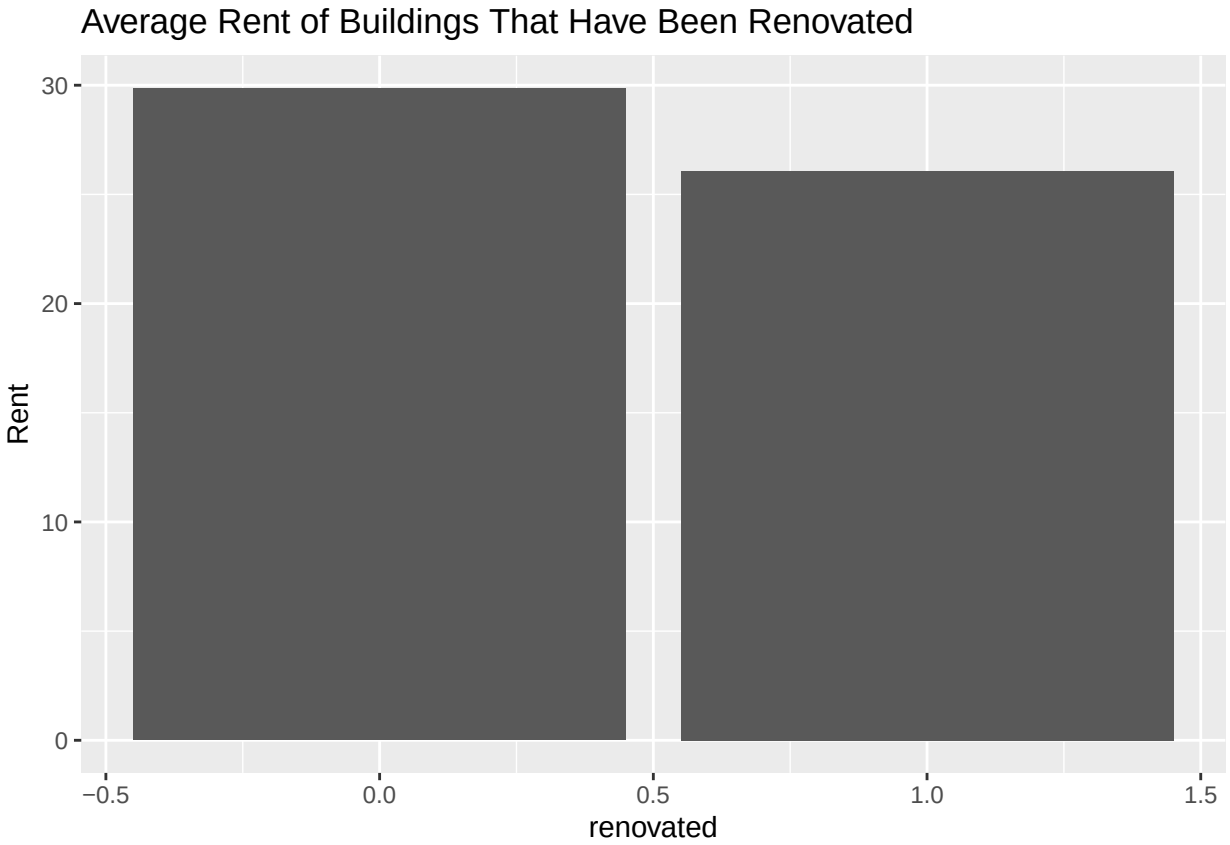

Average Age of Green and Non−Green Buildings
50
40
30
20
10
0
age
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00

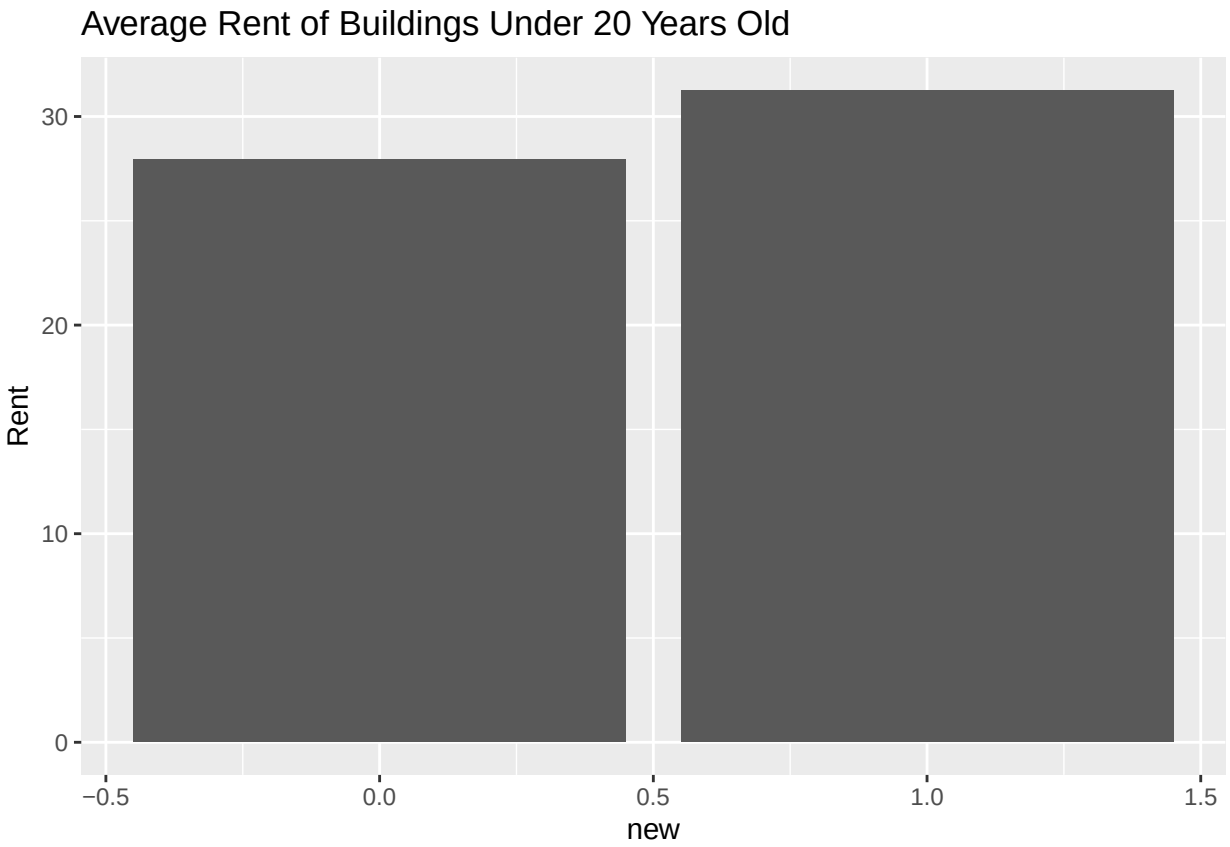

Average Rent of Buildings in Class A
Rent
30
20
10
0
−0.5
0.0
0.5
1.0
1.5
class_a


Average Rent of Buildings in Class B
Rent
30
20
10
0
−0.5
0.0
0.5
1.0
1.5
class_b


Average Rent of Buildings in Class C
Rent
30
20
10
0
−0.5
0.0
0.5
1.0
1.5
class_c

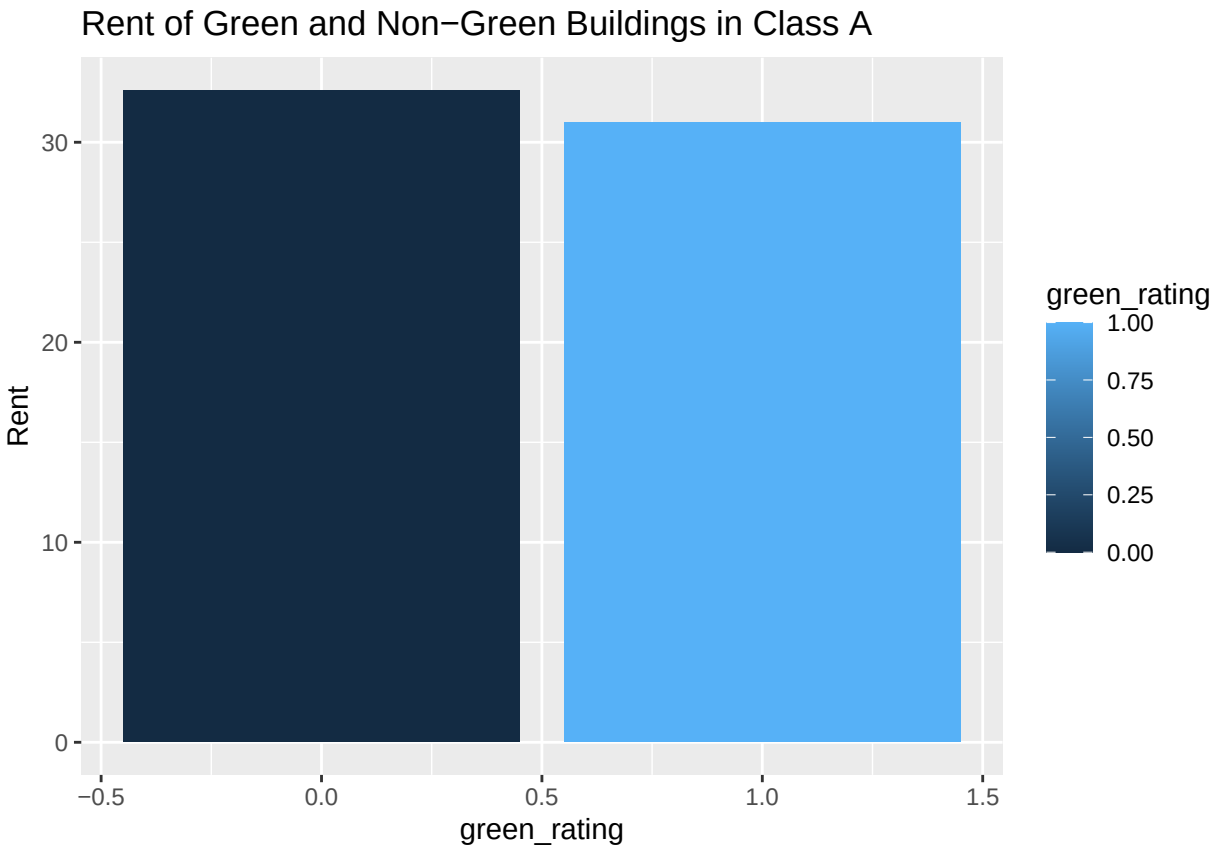

Average Rent of Buildings with Amenities
Rent
30
20
10
0
−0.5
0.0
0.5
1.0
1.5
amenities

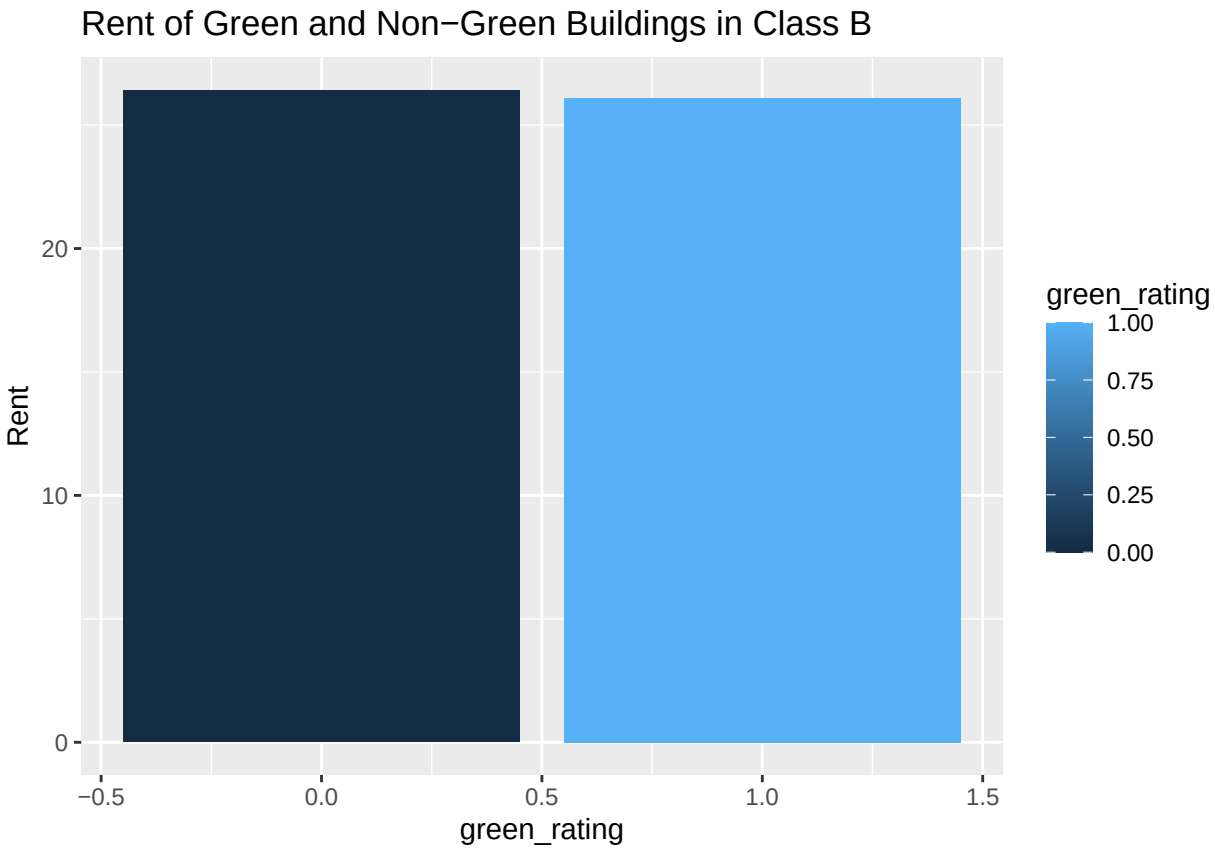

Average Rent of Buildings That Have Been Renovated
Rent
30
20
10
0
−0.5
0.0
0.5
1.0
1.5
renovated

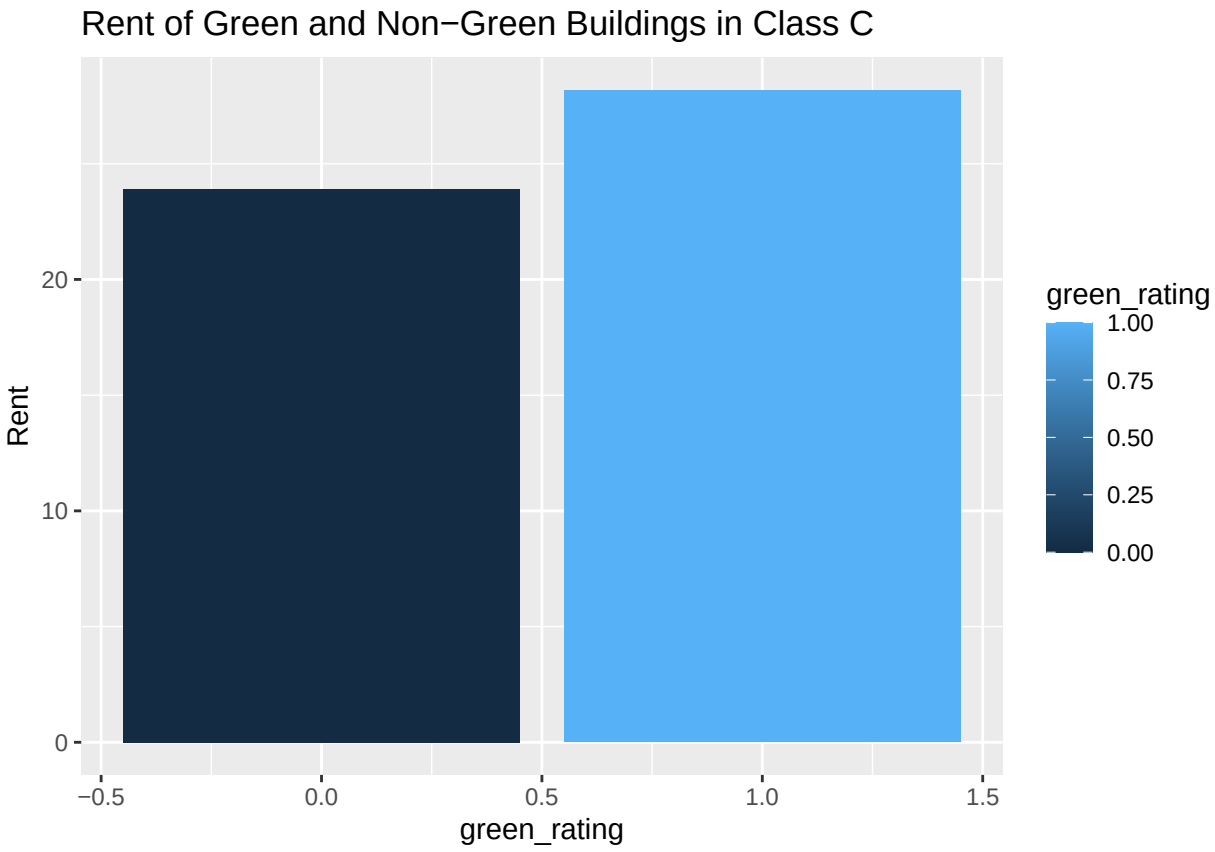


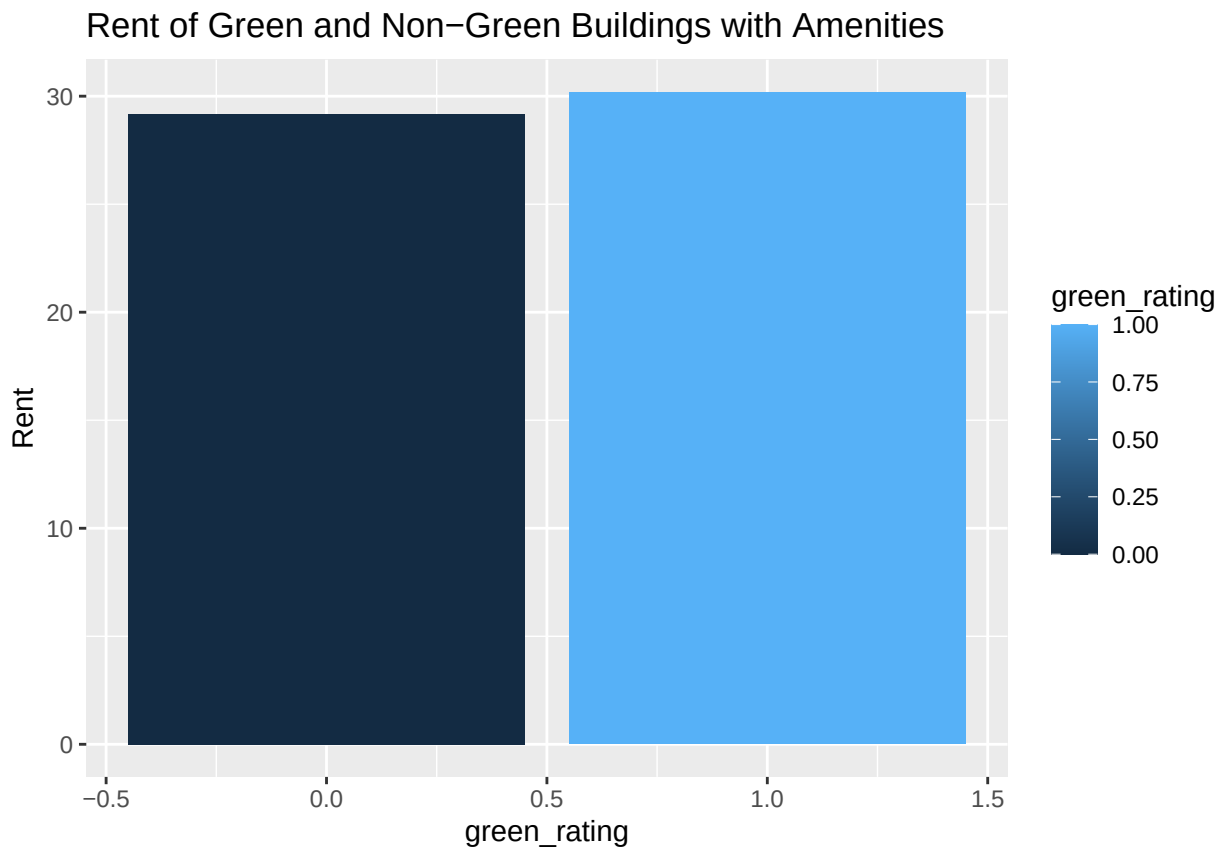
At this stage, however, we don't know if the buildings cost more because they're more likely to be green or if they're just higher class.

So let's compare prices between green buildings and non-green buildings while holding niceness constant.


Rent of Green and Non−Green Buildings in Class A
Rent
30
20
10
0
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00


Rent of Green and Non−Green Buildings in Class B
Rent
20
10
0
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00


Rent of Green and Non−Green Buildings in Class C
Rent
20
10
0
−0.5
0.0
0.5
1.0
1.5
green_rating
green_rating
1.00
0.75
0.50
0.25
0.00



Here we can see green buildings cost about as much as their non-green counterparts when you hold niceness constant. The only exception in class_c buildings, and considering how few green class_c buildings there are it's dangerous to draw too many conclusions from that data.

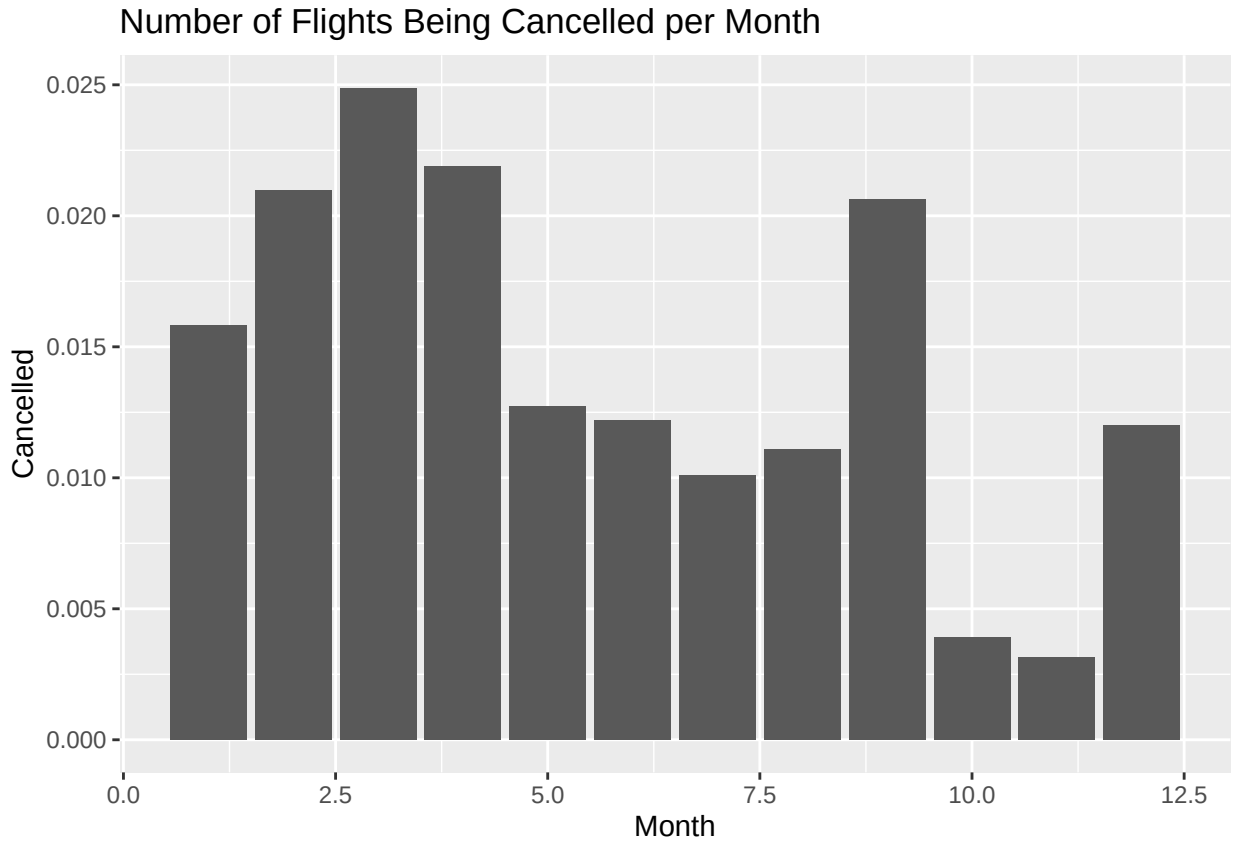
From this we can conclude that you are unlikely to make significantly more money from rent by building a green building than a non-green one.

There are still reasons to go green: it's good PR, they tend to last longer, it helps the environment, you might even save some money on utilities. But don't expect to make more rent money by going green.

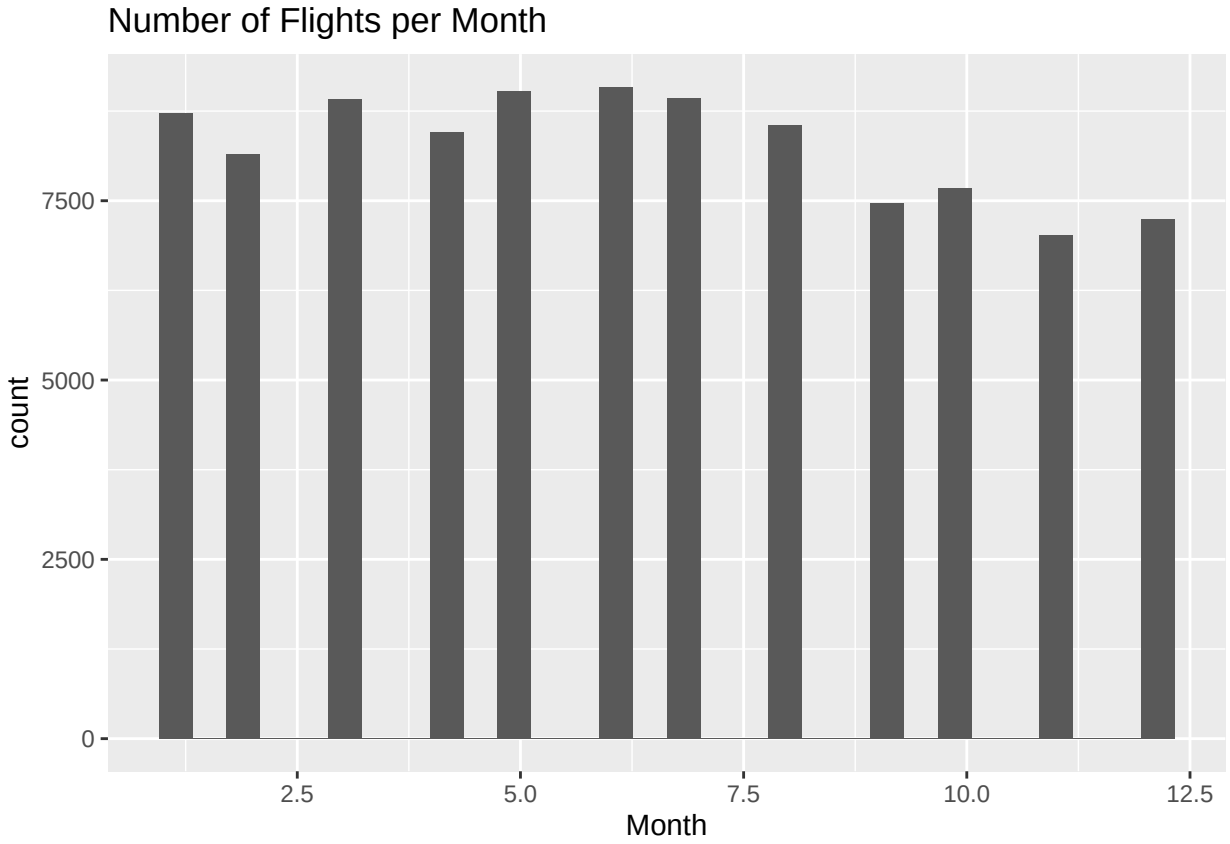
2. Visual story telling part 2: flights at ABIA

We can see a general trend in cancelled flights in Austin - a flight is most likely to be cancelled in the Spring and least likely to get cancelled in the fall, with a more or less smooth sine curve between those two extremes, with one notable exception. September has almost as many cancellations as April.

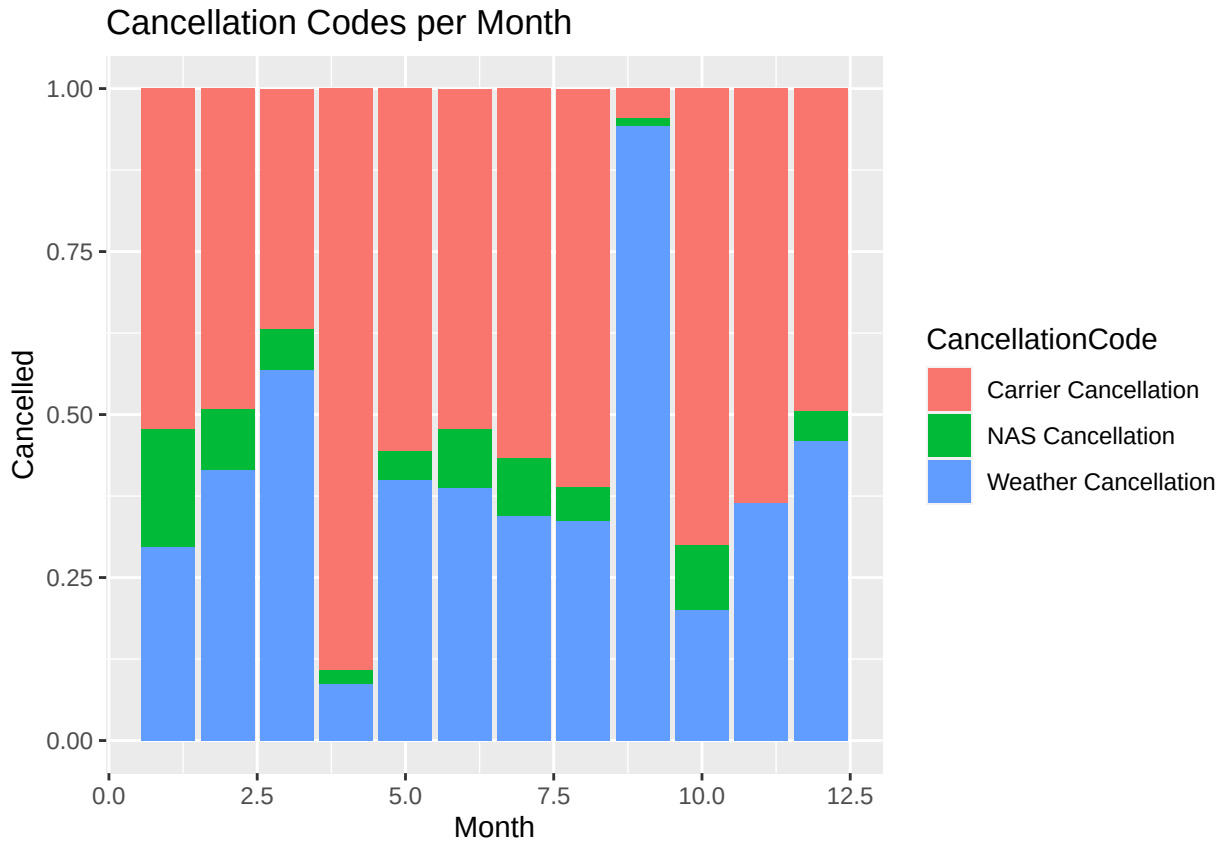
What's causing this?



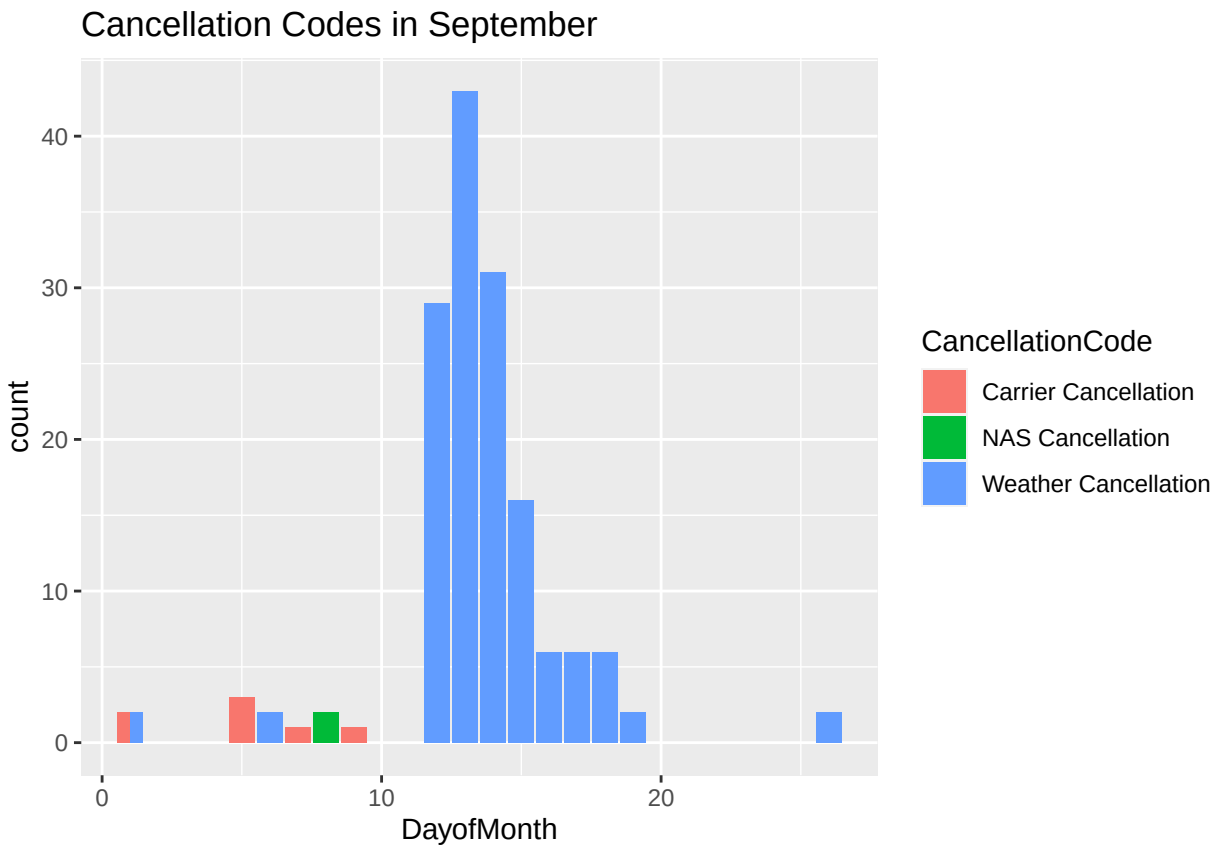
The number of flights each month is relatively constant, so its not that September is just a particularly busy month.



By looking at the cancellation codes, we can see that September flights are much more likely to get cancelled for weather reasons than any other month. The most likely cause of this is Hurricane Ike, which hit Galveston on September 13th, 2008.



We can confirm this by graphing cancelled flights in September over time. Below, you can see a sharp spike around the 13th, right when Ike hit.



3. Portfolio Modeling

We want to create three diverse portfolios that are not only diverse within the portfolios themselves, but also diverse compared to the other portfolios.

(a) Portfolio 1:————————————————————————————

Portfolio 1 contains 3 categories: High Volatility, Low Volatility, and Diverse:

High Volatility:

Technology Equities ETFs: These ETFs offer exposure to stocks within the technology sector. Note that tech stocks tend to carry a bit more volatility than other sectors, as they have a higher risk/reward profile.

- VGT - Vanguard Information Technology ETF
- SOXX - iShares Semiconductor ETF

Low Volatility:

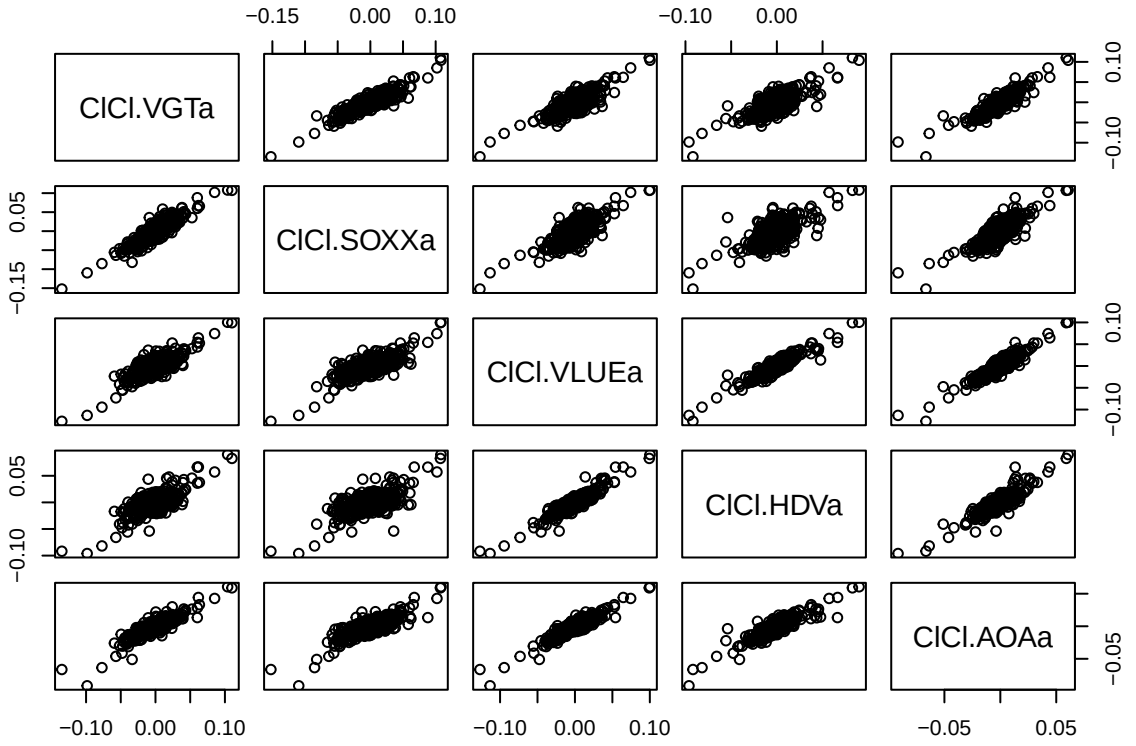
Large Cap Value Equities ETFs: These ETFs offer exposure to domestic large cap size securities deemed to possess value characteristics. These types of securities are generally in stable industries with low to moderate growth prospects and trade at relative low price-to-earnings ratios. As such, value equities tend to be more appealing to income-focused investors rather than those who are entirely interested in capital appreciation.

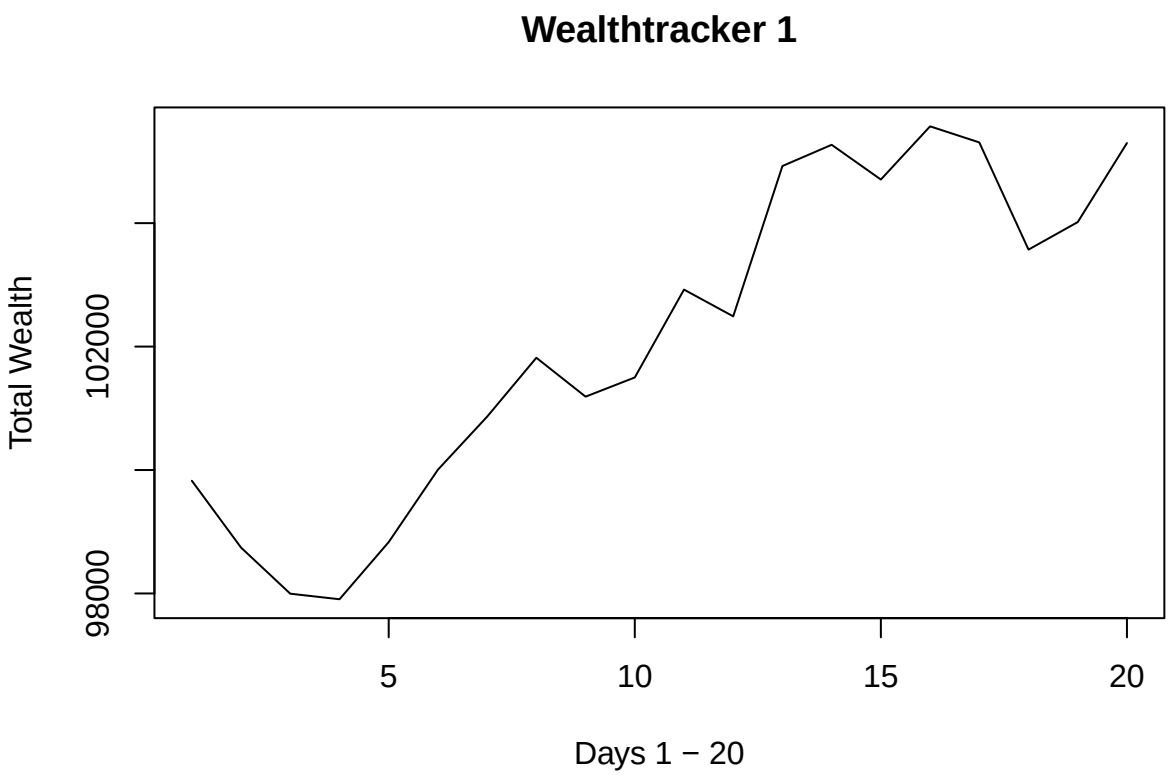
- VLUE - iShares MSCI USA Value Factor ETF
- HDV - iShares Core High Dividend ETF

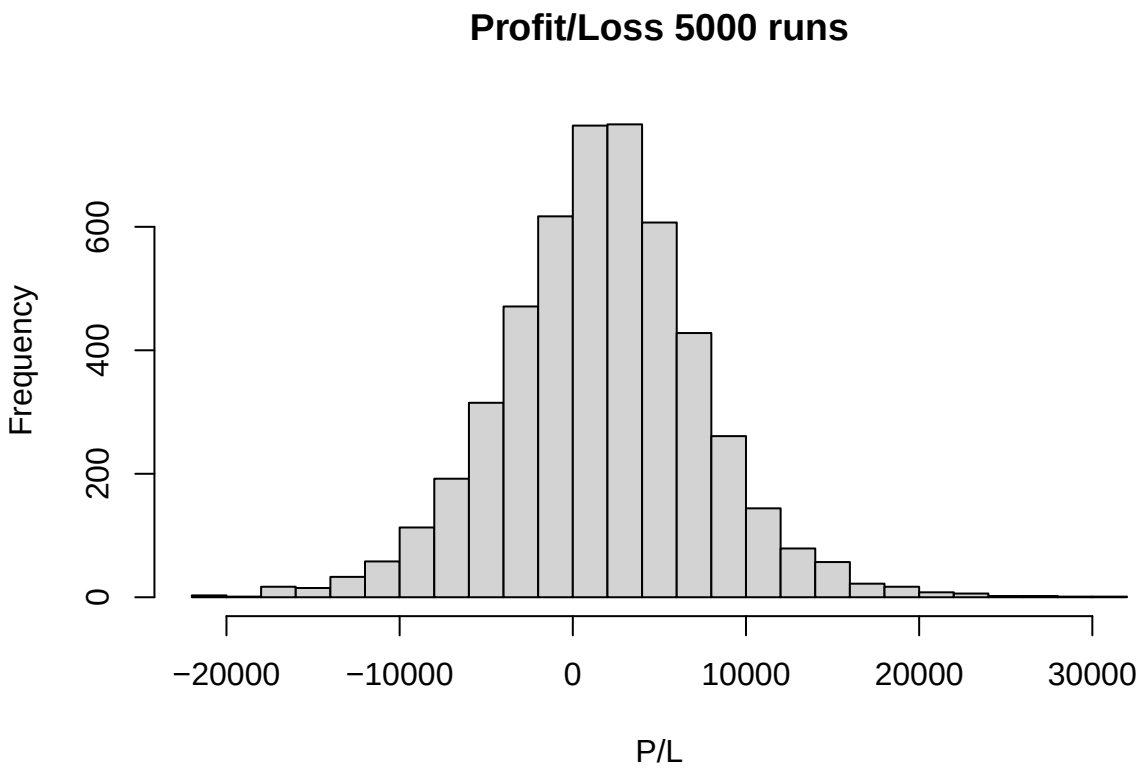
Diverse:

Diversified Portfolios ETFs: These ETFs offer investors exposure to multiple asset classes through a single ticker. These funds vary in investment objectives and risk/return profiles, but typically invest in a mix of equities and fixed income securities. Some diversified portfolio ETFs also offer exposure to commodity and currency exposure as well.

- AOA - iShares Core Aggressive Allocation ETF




Wealthtracker 1
Total Wealth
98000
102000
5
10
15
20
Days 1 − 20




```
## [1] "The mean Profit for Portfolio 1 is:"

## [1] 101674.6

## [1] "The mean Loss for Portfolio 1 is:"

## [1] 1674.645

## [1] "The VaR for Portfolio 1 is:"

##        5%
## -7845.694
```

Portfolio 1 produces the lowest VaR of all portfolios

(b) Portfolio 2:————————————————————————————

Portfolio 2 contains 3 categories: Large Market Cap, Alternative W/ Risk, and Global:

Large Market Cap:

Large Cap Growth Equities ETFs: These ETFs invest in growth company stocks that are believed to have a large market capitalization size, generally with a market capitalization of $10 billion or more.

- SPY - SPDR S&P 500 ETF Trust
- XLC - Communication Services Select Sector SPDR Fund

Alternative W/ Risk:

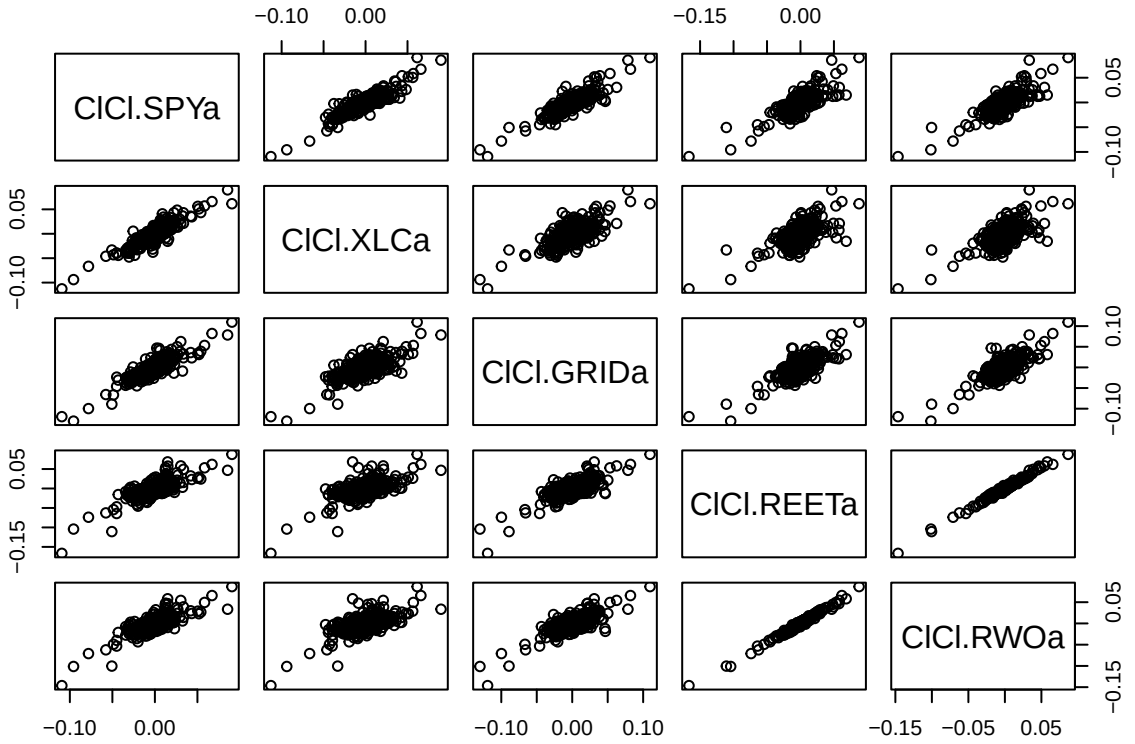
Alternative Energy Equities ETFs: These ETFs invest in alternative energy companies. The most popular and most common industry in this category is solar energy, although wind, hydroelectric, and geothermal energies are also represented here.

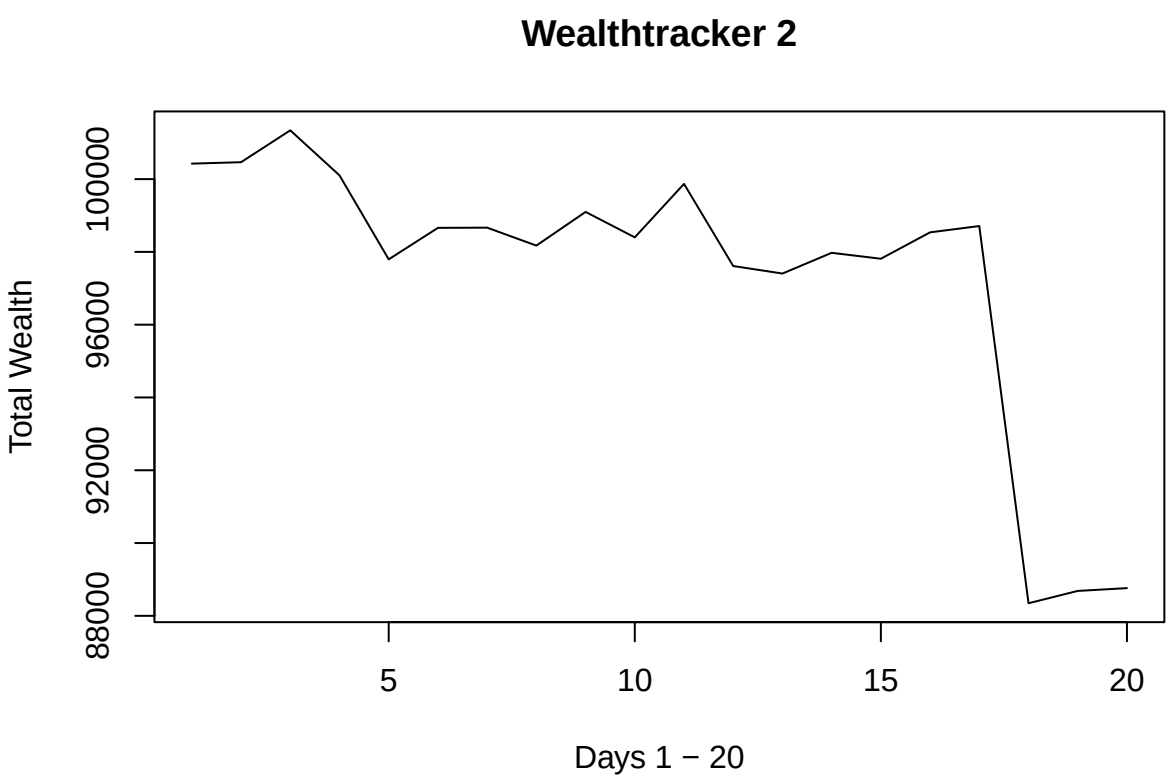
- GRID - First Trust Nasdaq Clean Edge Smart GRID Infrastructure Index

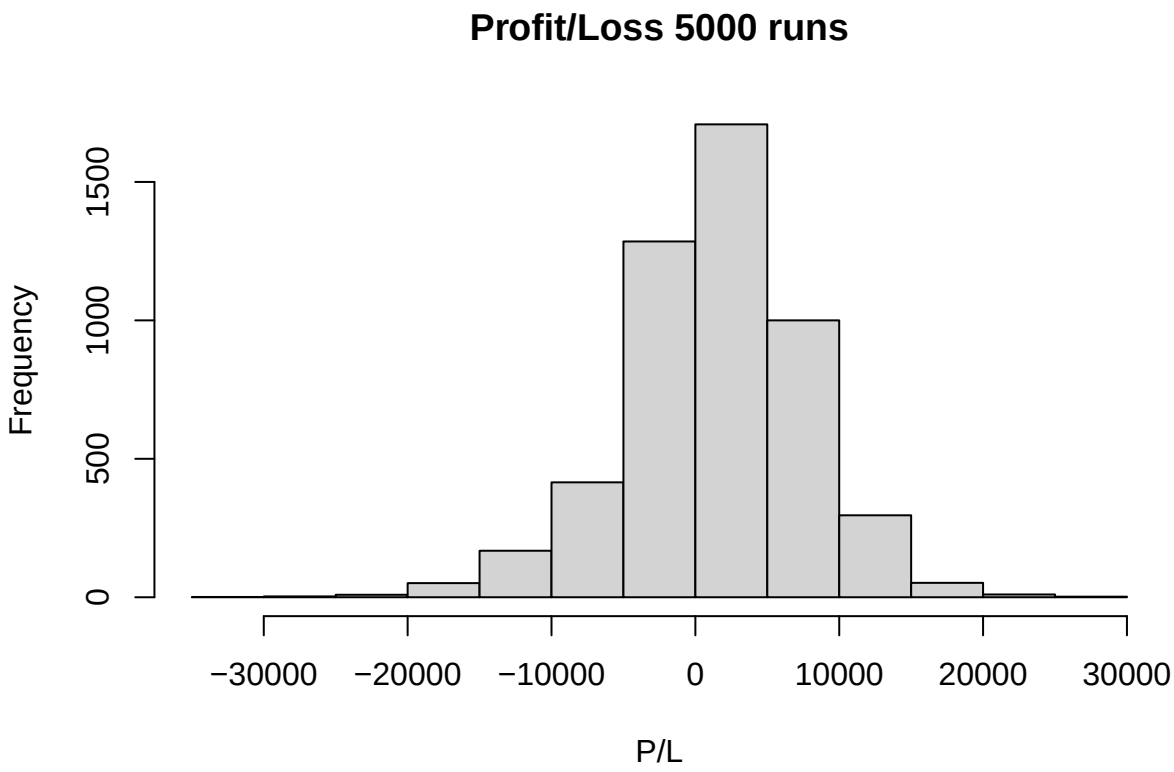
Global:

Global Real Estate ETFs: These ETFs invest in real estate companies from all over the world. These ETFs can offer broad exposure to the industry, or can target specific subsectors such as residential property. In addition, some of these funds focus on the global ex-U.S. market, while others target a specific region or country.

- REET - iShares Global REIT ETF
- RWO - SPDR Dow Jones Global Real Estate ETF




Wealthtracker 2
Total Wealth
88000
92000
96000
100000
5
10
15
20
Days 1 − 20




```
## [1] "The mean Profit for Portfolio 2 is:"

## [1] 101399.2

## [1] "The mean Loss for Portfolio 2 is:"

## [1] 1399.212

## [1] "The VaR for Portfolio 2 is:"

##       5%
## -9644.88
```

Portfolio 2 produces a slightly higher VaR than Portfolio 1.

(c) Portfolio 3:————————————————————————————————

Portfolio 3 contains 2 categories: Defensive and Complex:

Defensive:

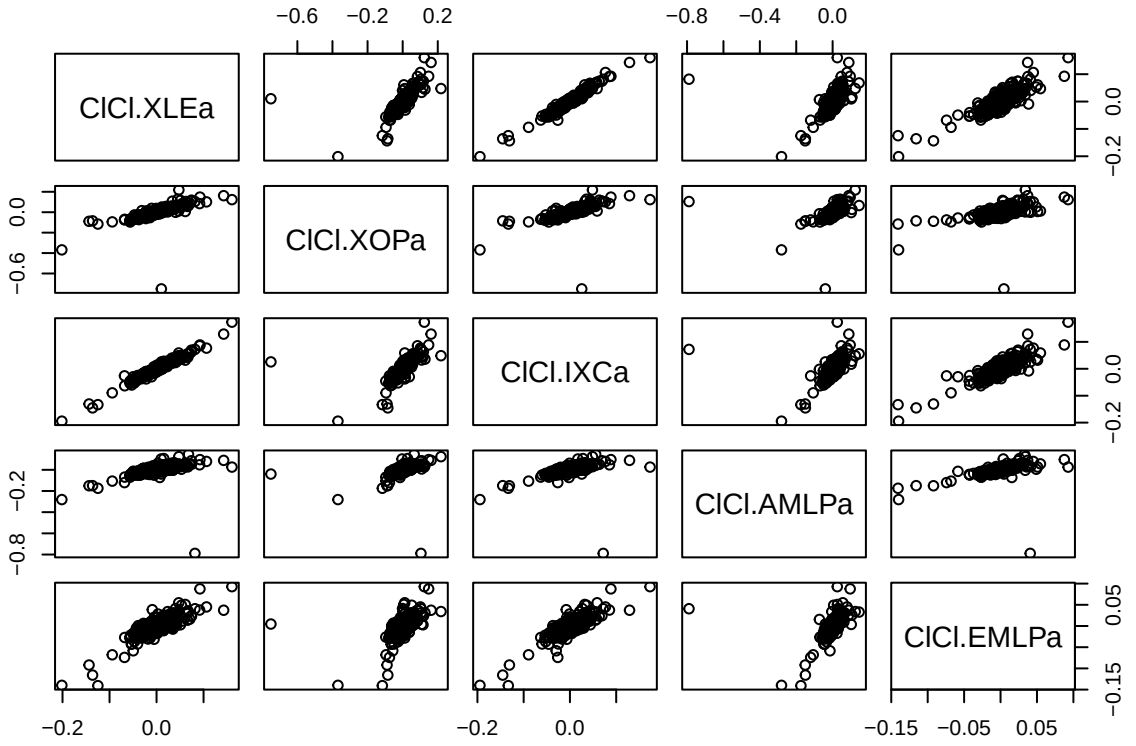
Energy Equities ETFs: An Inverse ETF (also known as a Short ETF or Bear ETF) is an ETF that is profitable for an investor during a market decline. These ETFs help investors defend against potential market downfalls, and can be used as a hedge against long positions during periods of market weakness.

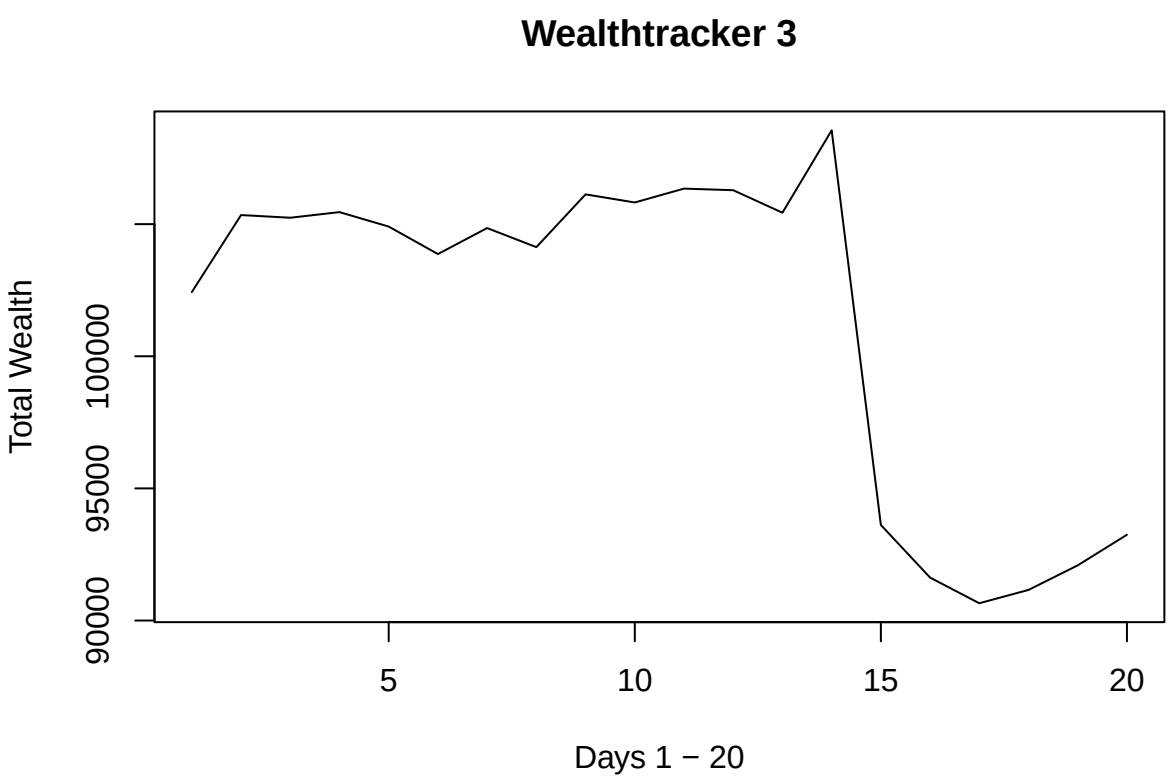
- XLE - Energy Select Sector SPDR Fund
- XOP - SPDR S&P Oil & Gas Exploration & Production ETF
- IXC - iShares Global Energy ETF

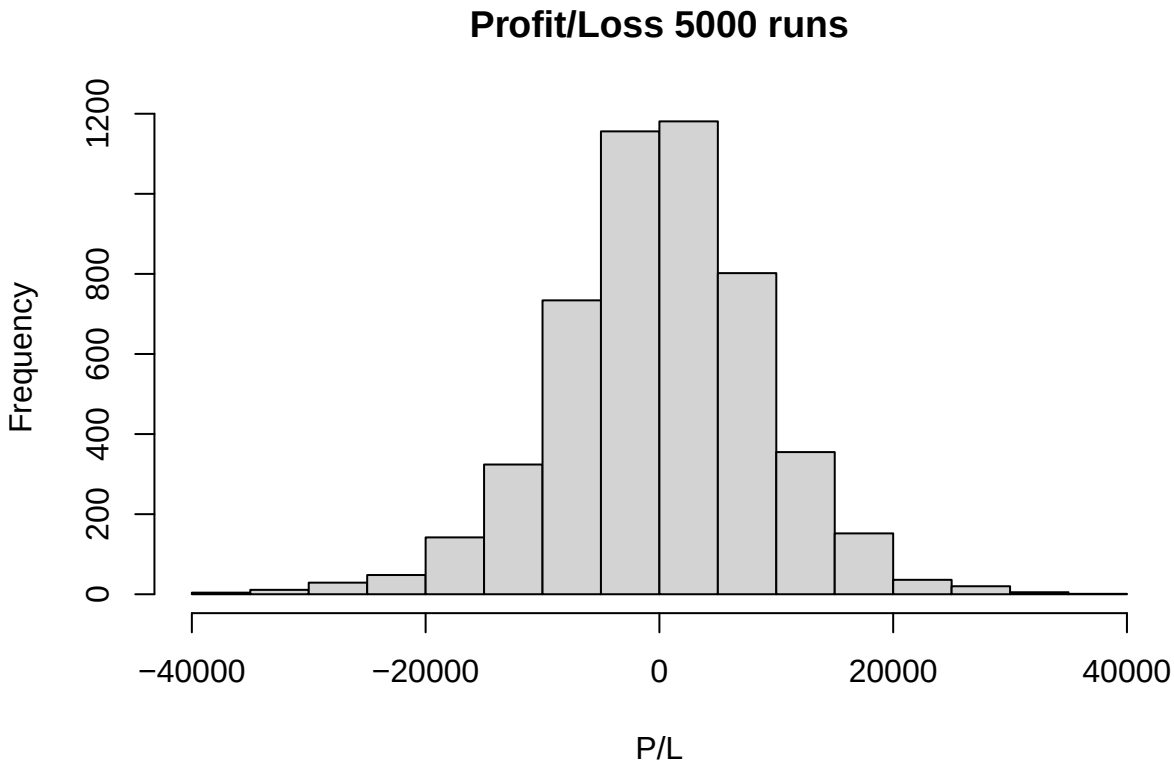
Complex:

MLPs ETFs: These ETFs invest in Master Limited Partnerships (MLPs). These companies are generally involved in the transportation, storage, and processing of energy commodities such as oil, natural gas, refined products, and natural gas liquids (NGLs). Funds in this category tend to have attractive dividend payouts.

- AMLP - Alerian MLP ETF
- EMLP - First Trust North American Energy Infrastructure Fund




Wealthtracker 3
Total Wealth
90000
95000
100000
5
10
15
20
Days 1 − 20




```
## [1] "The mean Profit for Portfolio 3 is:"

## [1] 100083.9

## [1] "The mean Loss for Portfolio 3 is:"

## [1] 83.94862

## [1] "The VaR for Portfolio 3 is:"

##        5%
## -14473.91
```

Portfolio 3 produces the highest VaR.

Summary:

It was a bit surprising to discover that Portfolio 3 produced the highest VaR considering one category was defensive. The MLP ETFs can be difficult to diversify, but that space is growing as their investment strategies are becoming easier to understand. Some owners of MPL ETFS are seeing high dividend payouts, but the VaR of this portfolio is about double compared to Portfolio 1, and the mean profit returns a loss.

This analysis would lead us to invest with Portfolio 1. Not only does it have the lowest VaR, it also returns the highest profit mean, with a mean loss that is always positive. The tech equities are considered somewhat risky because of that ever changing environment, but for now the reward profile looks to pay off.

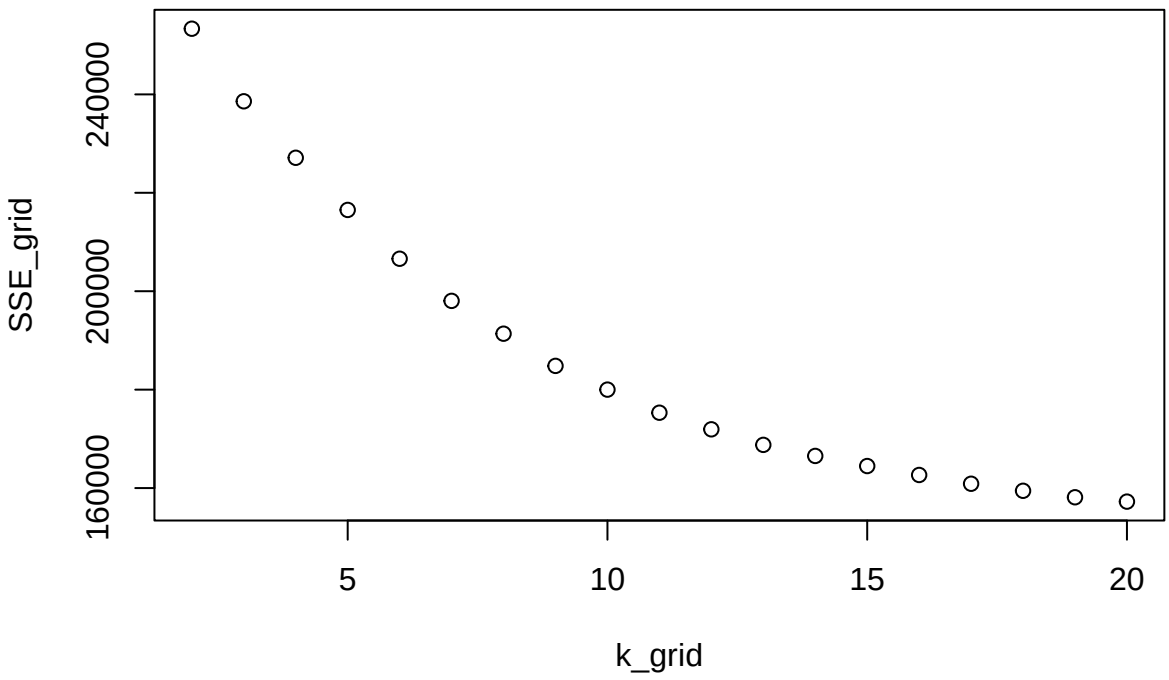
Based off of this analysis, we would build a portfolio that is income focused, diverse, and includes some risk but is balanced out by value oriented low to moderate growth prospects.

4. Market Segmentation

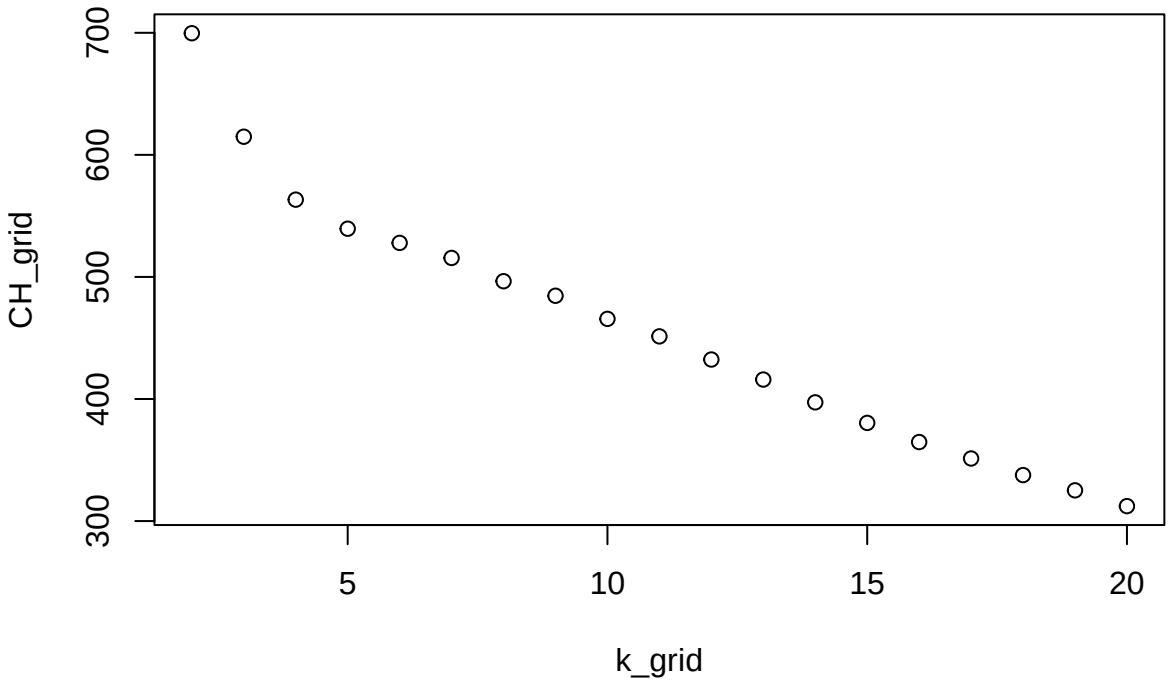
We will use clustering to analyze the social marketing data provided by NutrientH20. The goal is to place the social media audience into small groups so that ads can be targeted more specifically. Because of the limited amount of information it provides, we will leave out the spam column for this analysis. Listed below are the 36 features that we will be working with:

```
##  [1] "chatter"          "current_events"   "travel"           "photo_sharing"
##  [5] "uncategorized"    "tv_film"          "sports_fandom"    "politics"
##  [9] "food"             "family"           "home_and_garden"  "music"
## [13] "news"             "online_gaming"    "shopping"         "health_nutrition"
## [17] "college_uni"      "sports_playing"   "cooking"          "eco"
## [21] "computers"        "business"         "outdoors"         "crafts"
## [25] "automotive"       "art"              "religion"         "beauty"
## [29] "parenting"        "dating"           "school"           "personal_fitness"
## [33] "fashion"          "small_business"   "adult"
```

To begin the clustering process, we look for the best value for k, or number of clusters. First we will look at an elbow plot:



The elbow plot is not giving a definitive enough value for k, so we will try plotting the CH index:



Now we see that the best value for k should be somewhere around 4 or 5. After trying 3, 4, 5, and 6, 5 produces the best results and captures the best separation of features. Now we will look at the top 5 features within each cluster:

The Cluster Top 5 Features breakdown is as follows:

Cluster 1:

- **"chatter", "photo_sharing", "current_events", "college_uni", "health_nutrition"**

Cluster 2:

- **"chatter", "photo_sharing", "cooking", "college_uni", "fashion"**

Cluster 3:

- **"politics", "travel", "news", "chatter", "computers"**

Cluster 4:

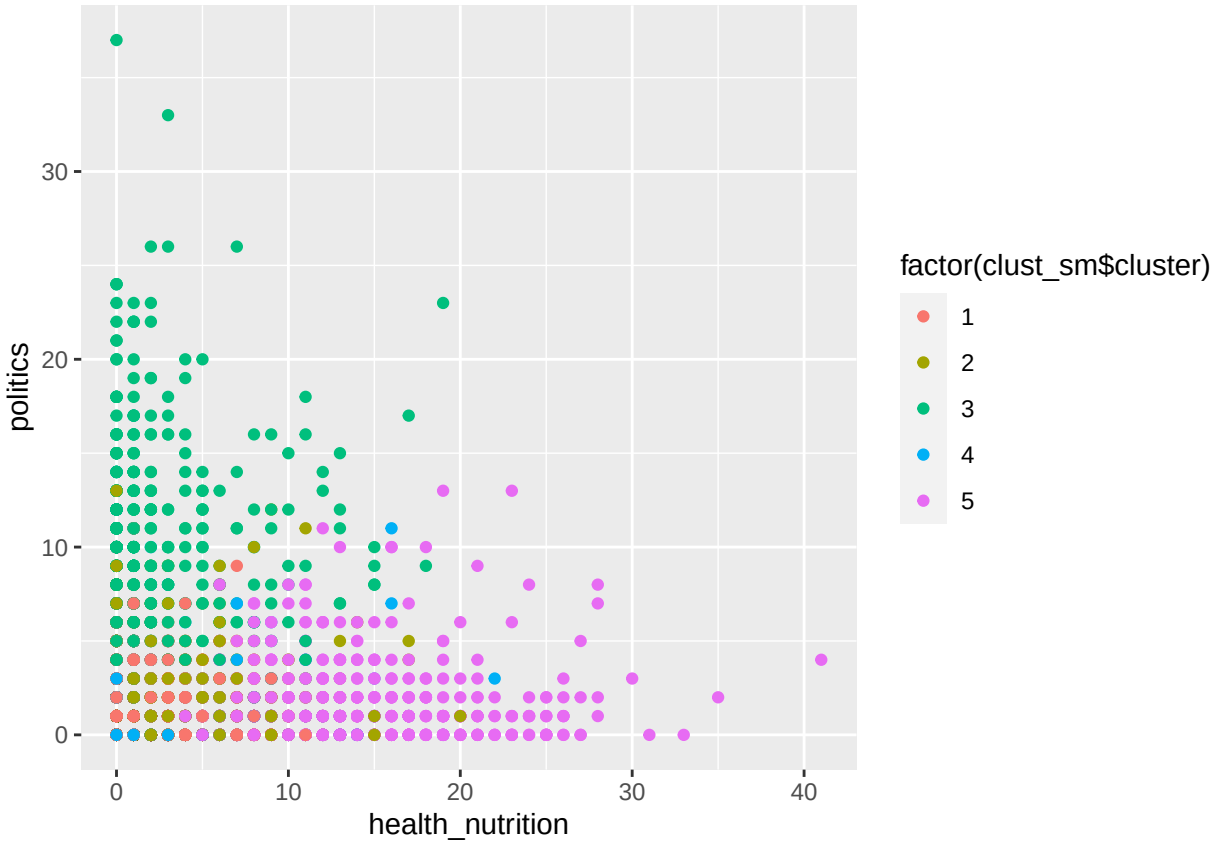
- **"sports_fandom", "religion", "food", "parenting", "chatter"**

Cluster 5:

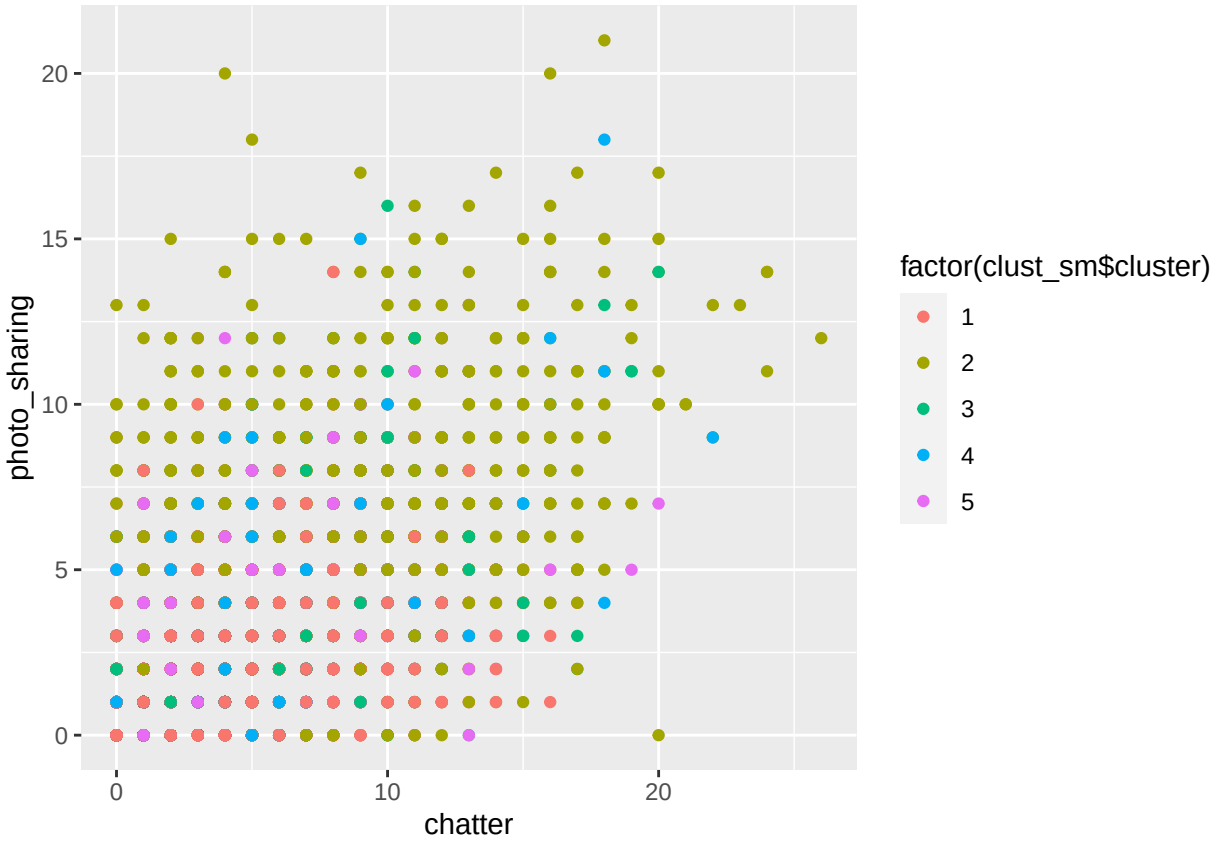
- **"health_nutrition", "personal_fitness", "chatter", "cooking", "outdoors"**

We now examine each cluster looking for further evidence to narrow our findings:

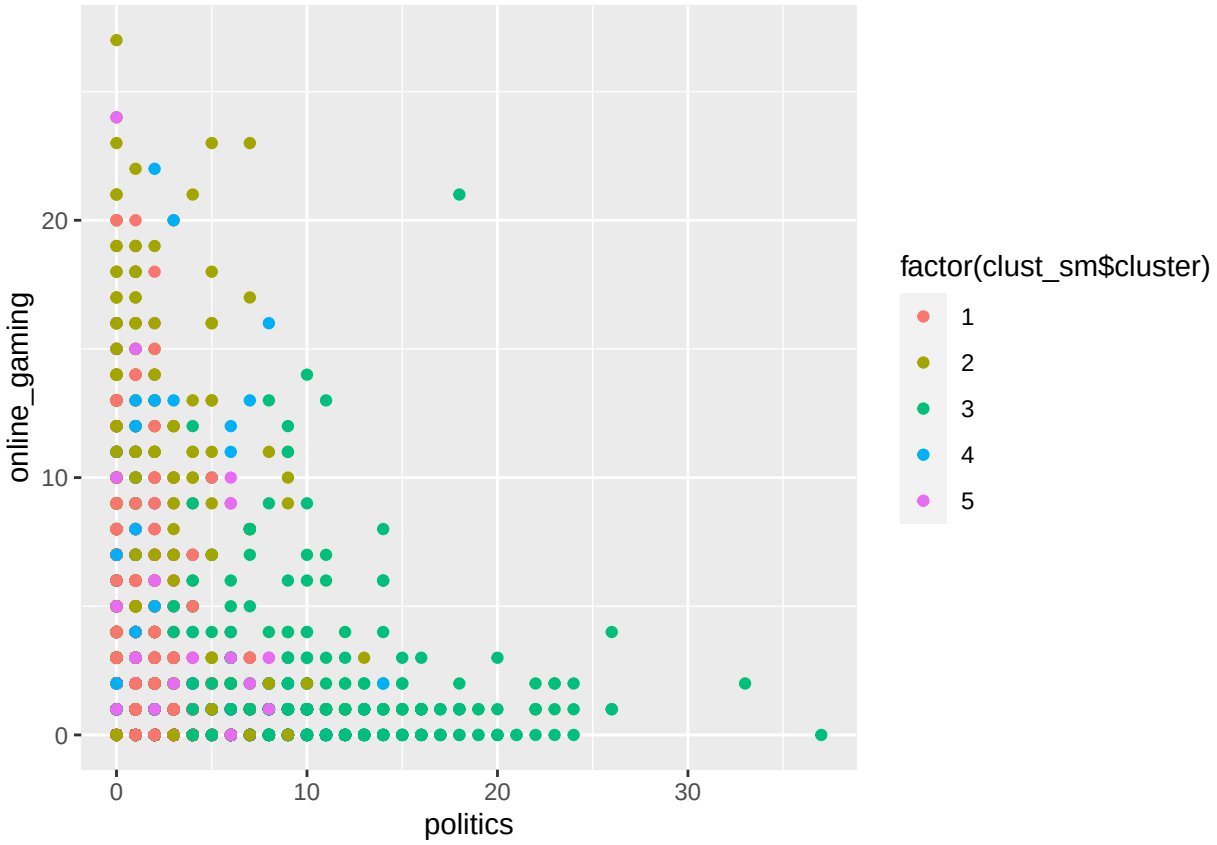
- **health_nutrition and politics scored the highest out of any cluster and as you can see below the two features are separated rather nicely. This clearly tells us that there are customers who would respond better to one of these categories, but not both.**



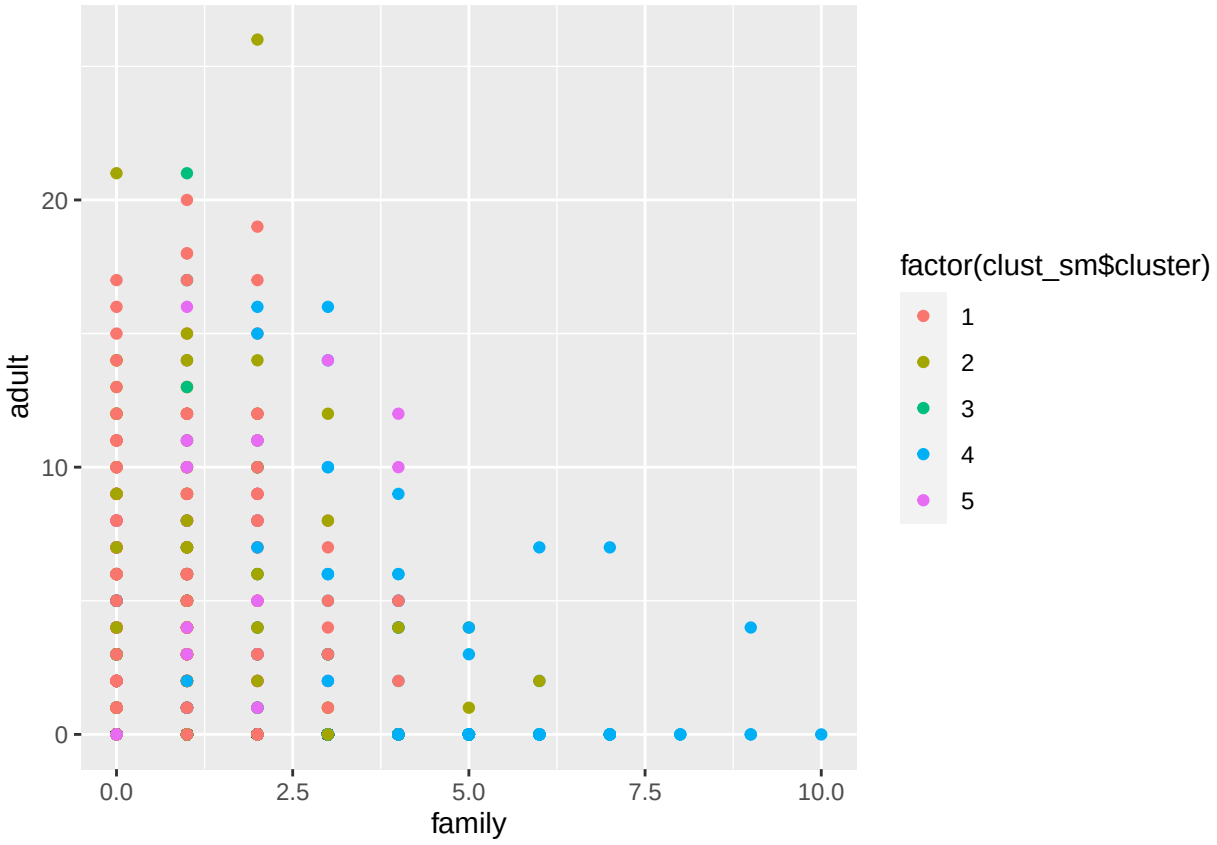
- **The plot below reveals that perhaps tweets that share photos are incorrectly classified as chatter. The chatter feature contains a lot of information so more precise classification could produce valuable results. For now, we suggest targeting chatter tweeters with photo_sharing content/ads.**

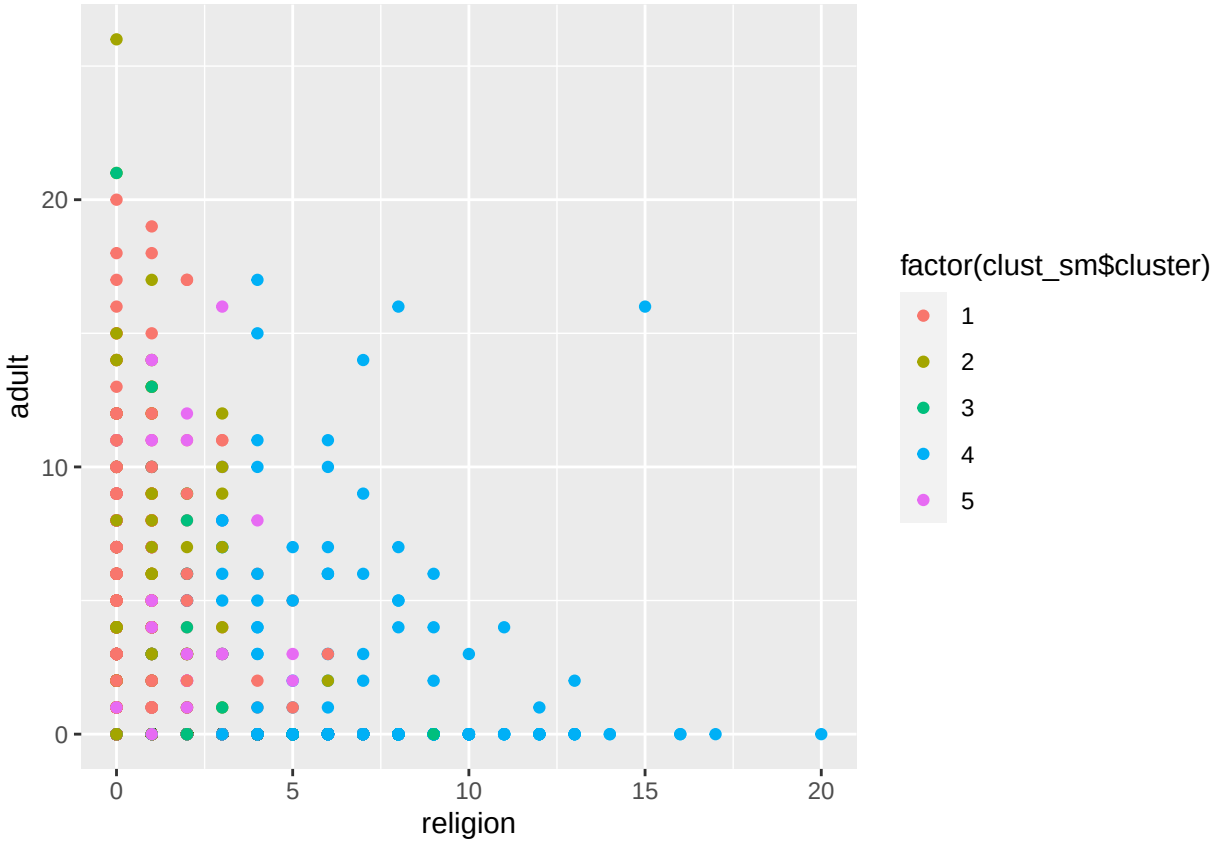


- **In the plot below we see that gamers don't tweet about politics much, telling us that the two categories shouldn't be mixed when targeting consumers**

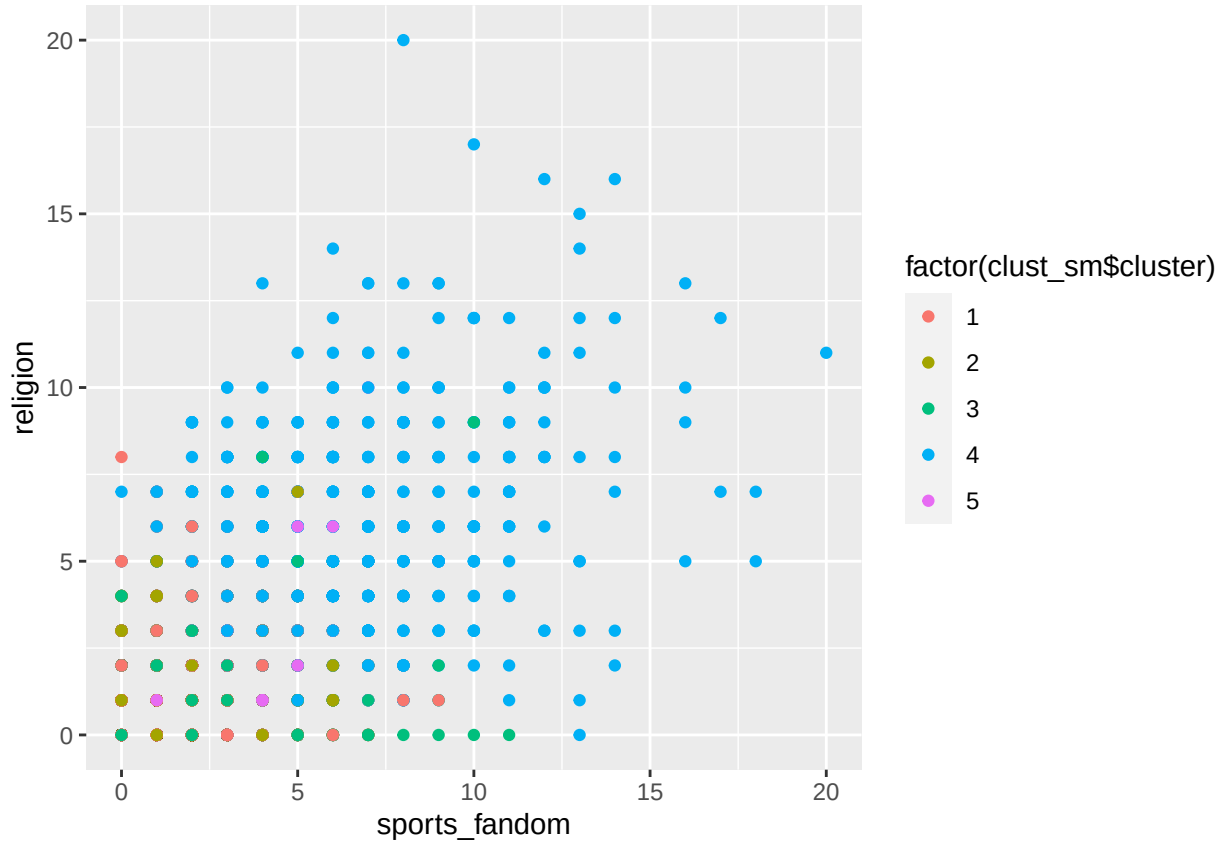


- **In both of the plots below, we see that users who tweet about family and religion typically don't tweet about adult content. Hoever, there are some interesting outliers, and it's concerning that family doesn't converge totally from the adult feature. These features should not mix well, and must always be considered because one misplaced ad could deter a lot of business.**




adult
religion
0
10
20
0
5
10
15
20
factor(clust_sm$cluster)
1
2
3
4
5

- **Below we see that targeting tweeters who are interested in both sports and religion may prove profitable.**



Summary:

This analysis is just beginning to tap into what could be discovered with this data set. We believe the biggest takeway is to more accurately classify every single tweet rather than just dropping a lot of them into a general category like chatter or uncategorized. There is a lot of potentially valuable information lost if the categorization process is not done with complete precision. That being said, we were able to find 5 distinct clusters allowing for more targeted ads and information gathering.

5. Author attribution

Each story is processed to remove apostrophes, convert to lowercase, change non-alphanumeric characters to spaces, and so on. It is then tokenized into a list of words, and added to a list of lists of tokens, called all_stories. We keep track of the author names and whether that story was in the train or test set separately.

all_stories is used to create a Corpus object, which we remove all useless stopwords from (as defined by the tm package). We then convert the Corpus object to a Document Term Matrix, which keeps track of how often each word appeared in each story across the entire corpus. We remove all sparse terms here (terms that appear in less than 5% of stories). The TF-IDF weights of each word is now calculated and used to build the dataframe we'll be working with.

We remove all columns that are all zeroes (they would cause issues with the PCA calculations), then run a PCA analysis on our dataframe. From looking at the summary we can see that with 158 columns we can encode 50% of the original information.

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.89956 3.15195 2.73843 2.72803 2.57864 2.43487 2.30372
## Proportion of Variance 0.01839 0.01201 0.00907 0.00900 0.00804 0.00717 0.00642
## Cumulative Proportion  0.01839 0.03040 0.03947 0.04847 0.05651 0.06368 0.07009
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     2.26246 2.22367 2.17086 2.09289 2.06275 2.03457 1.98372
## Proportion of Variance 0.00619 0.00598 0.00570 0.00530 0.00515 0.00501 0.00476
## Cumulative Proportion  0.07628 0.08226 0.08796 0.09326 0.09840 0.10341 0.10817
##                           PC15    PC16    PC17    PC18   PC19    PC20    PC21
## Standard deviation     1.93904 1.85748 1.85612 1.81339 1.7968 1.77671 1.75677
## Proportion of Variance 0.00455 0.00417 0.00417 0.00398 0.0039 0.00382 0.00373
## Cumulative Proportion  0.11271 0.11688 0.12105 0.12503 0.1289 0.13275 0.13648
##                           PC22    PC23    PC24    PC25    PC26    PC27   PC28
## Standard deviation     1.7484 1.71925 1.70721 1.69195 1.68455 1.66373 1.6510
## Proportion of Variance 0.0037 0.00357 0.00352 0.00346 0.00343 0.00335 0.0033
## Cumulative Proportion  0.1402 0.14375 0.14727 0.15074 0.15417 0.15751 0.1608
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     1.62859 1.61931 1.60892 1.59122 1.58016 1.57789 1.56984
## Proportion of Variance 0.00321 0.00317 0.00313 0.00306 0.00302 0.00301 0.00298
## Cumulative Proportion  0.16402 0.16719 0.17032 0.17338 0.17640 0.17941 0.18239
##                           PC36    PC37    PC38    PC39    PC40   PC41    PC42
## Standard deviation     1.5492 1.54624 1.54247 1.53326 1.52858 1.5218 1.50718
## Proportion of Variance 0.0029 0.00289 0.00288 0.00284 0.00283 0.0028 0.00275
## Cumulative Proportion  0.1853 0.18818 0.19106 0.19390 0.19673 0.1995 0.20227
##                           PC43   PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     1.50207 1.4941 1.49256 1.48776 1.48192 1.47892 1.47231
## Proportion of Variance 0.00273 0.0027 0.00269 0.00268 0.00266 0.00264 0.00262
## Cumulative Proportion  0.20500 0.2077 0.21040 0.21307 0.21573 0.21837 0.22099
##                           PC50    PC51    PC52    PC53    PC54   PC55    PC56
## Standard deviation     1.47089 1.46307 1.45542 1.44656 1.44152 1.4389 1.43424
## Proportion of Variance 0.00262 0.00259 0.00256 0.00253 0.00251 0.0025 0.00249
## Cumulative Proportion  0.22361 0.22620 0.22876 0.23129 0.23380 0.2363 0.23879
##                           PC57    PC58    PC59    PC60    PC61    PC62    PC63
## Standard deviation     1.43323 1.42806 1.42387 1.42016 1.41586 1.41448 1.40625
## Proportion of Variance 0.00248 0.00247 0.00245 0.00244 0.00242 0.00242 0.00239
## Cumulative Proportion  0.24128 0.24374 0.24619 0.24863 0.25106 0.25348 0.25587
##                           PC64    PC65    PC66    PC67    PC68    PC69    PC70
## Standard deviation     1.40401 1.39568 1.39254 1.39163 1.38817 1.38552 1.38455
## Proportion of Variance 0.00238 0.00236 0.00234 0.00234 0.00233 0.00232 0.00232
## Cumulative Proportion  0.25825 0.26061 0.26295 0.26529 0.26762 0.26995 0.27226
##                           PC71    PC72    PC73    PC74    PC75    PC76    PC77
## Standard deviation     1.3792 1.37574 1.37411 1.37009 1.36712 1.36403 1.35482
## Proportion of Variance 0.0023 0.00229 0.00228 0.00227 0.00226 0.00225 0.00222
## Cumulative Proportion  0.2746 0.27685 0.27913 0.28140 0.28366 0.28591 0.28813
##                           PC78    PC79   PC80    PC81    PC82    PC83    PC84
## Standard deviation     1.35393 1.35204 1.3485 1.34572 1.34107 1.33942 1.33541
## Proportion of Variance 0.00222 0.00221 0.0022 0.00219 0.00217 0.00217 0.00216
## Cumulative Proportion  0.29035 0.29256 0.2948 0.29695 0.29912 0.30129 0.30345
##                           PC85    PC86    PC87    PC88    PC89    PC90   PC91
```

```
## Standard deviation     1.33273 1.33103 1.32771 1.32496 1.32233 1.32019 1.3184
## Proportion of Variance 0.00215 0.00214 0.00213 0.00212 0.00211 0.00211 0.0021
## Cumulative Proportion  0.30560 0.30774 0.30987 0.31199 0.31411 0.31622 0.3183
##                           PC92    PC93    PC94    PC95    PC96    PC97    PC98
## Standard deviation     1.31625 1.31477 1.30964 1.30768 1.30601 1.30477 1.30283
## Proportion of Variance 0.00209 0.00209 0.00207 0.00207 0.00206 0.00206 0.00205
## Cumulative Proportion  0.32041 0.32250 0.32458 0.32664 0.32871 0.33077 0.33282
##                           PC99   PC100   PC101   PC102   PC103   PC104  PC105
## Standard deviation     1.29973 1.29819 1.29303 1.29039 1.28894 1.28836 1.2868
## Proportion of Variance 0.00204 0.00204 0.00202 0.00201 0.00201 0.00201 0.0020
## Cumulative Proportion  0.33486 0.33690 0.33892 0.34093 0.34294 0.34495 0.3469
##                          PC106   PC107   PC108   PC109   PC110   PC111   PC112
## Standard deviation     1.28329 1.27927 1.27516 1.27201 1.27070 1.26910 1.26591
## Proportion of Variance 0.00199 0.00198 0.00197 0.00196 0.00195 0.00195 0.00194
## Cumulative Proportion  0.34894 0.35092 0.35289 0.35485 0.35680 0.35875 0.36068
##                          PC113   PC114   PC115   PC116   PC117   PC118  PC119
## Standard deviation     1.26491 1.26438 1.26263 1.26004 1.25754 1.25580 1.2539
## Proportion of Variance 0.00193 0.00193 0.00193 0.00192 0.00191 0.00191 0.0019
## Cumulative Proportion  0.36262 0.36455 0.36648 0.36840 0.37031 0.37222 0.3741
##                          PC120   PC121   PC122   PC123   PC124   PC125   PC126
## Standard deviation     1.25142 1.24938 1.24497 1.24345 1.24165 1.24143 1.23785
## Proportion of Variance 0.00189 0.00189 0.00187 0.00187 0.00186 0.00186 0.00185
## Cumulative Proportion  0.37601 0.37790 0.37977 0.38164 0.38351 0.38537 0.38722
##                          PC127   PC128   PC129   PC130   PC131   PC132   PC133
## Standard deviation     1.23461 1.23390 1.23293 1.23079 1.22989 1.22667 1.22612
## Proportion of Variance 0.00184 0.00184 0.00184 0.00183 0.00183 0.00182 0.00182
## Cumulative Proportion  0.38907 0.39091 0.39275 0.39458 0.39641 0.39823 0.40004
##                          PC134   PC135  PC136  PC137  PC138   PC139   PC140
## Standard deviation     1.22484 1.22269 1.2212 1.2202 1.2184 1.21677 1.21481
## Proportion of Variance 0.00181 0.00181 0.0018 0.0018 0.0018 0.00179 0.00178
## Cumulative Proportion  0.40186 0.40367 0.4055 0.4073 0.4091 0.41085 0.41264
##                          PC141   PC142   PC143   PC144   PC145   PC146   PC147
## Standard deviation     1.21421 1.21209 1.20889 1.20777 1.20351 1.20263 1.20206
## Proportion of Variance 0.00178 0.00178 0.00177 0.00176 0.00175 0.00175 0.00175
## Cumulative Proportion  0.41442 0.41620 0.41797 0.41973 0.42148 0.42323 0.42498
##                          PC148   PC149   PC150   PC151   PC152   PC153   PC154
## Standard deviation     1.20057 1.19838 1.19651 1.19589 1.19438 1.19005 1.18954
## Proportion of Variance 0.00174 0.00174 0.00173 0.00173 0.00172 0.00171 0.00171
## Cumulative Proportion  0.42672 0.42846 0.43019 0.43192 0.43364 0.43535 0.43707
##                          PC155   PC156  PC157   PC158   PC159   PC160   PC161
## Standard deviation     1.18916 1.18763 1.1849 1.18296 1.18220 1.18008 1.17912
## Proportion of Variance 0.00171 0.00171 0.0017 0.00169 0.00169 0.00168 0.00168
## Cumulative Proportion  0.43878 0.44048 0.4422 0.44387 0.44556 0.44724 0.44893
##                          PC162   PC163   PC164   PC165   PC166   PC167   PC168
## Standard deviation     1.17774 1.17598 1.17499 1.17296 1.17127 1.17109 1.16977
## Proportion of Variance 0.00168 0.00167 0.00167 0.00166 0.00166 0.00166 0.00165
## Cumulative Proportion  0.45060 0.45228 0.45394 0.45561 0.45727 0.45893 0.46058
##                          PC169   PC170   PC171   PC172   PC173   PC174   PC175
## Standard deviation     1.16597 1.16477 1.16338 1.16158 1.16000 1.15968 1.15762
## Proportion of Variance 0.00164 0.00164 0.00164 0.00163 0.00163 0.00163 0.00162
## Cumulative Proportion  0.46222 0.46386 0.46550 0.46713 0.46876 0.47039 0.47201
##                          PC176   PC177  PC178  PC179  PC180   PC181   PC182
## Standard deviation     1.15640 1.15400 1.1517 1.1506 1.1498 1.14617 1.14567
## Proportion of Variance 0.00162 0.00161 0.0016 0.0016 0.0016 0.00159 0.00159
```

```
## Cumulative Proportion  0.47362 0.47523 0.4768 0.4784 0.4800 0.48163 0.48321
##                           PC183   PC184   PC185   PC186   PC187   PC188   PC189
## Standard deviation     1.14490 1.14360 1.14182 1.13943 1.13749 1.13712 1.13586
## Proportion of Variance 0.00159 0.00158 0.00158 0.00157 0.00156 0.00156 0.00156
## Cumulative Proportion  0.48480 0.48638 0.48796 0.48953 0.49109 0.49265 0.49421
##                           PC190   PC191   PC192   PC193   PC194   PC195   PC196
## Standard deviation     1.13433 1.13392 1.13222 1.13079 1.12964 1.12801 1.12717
## Proportion of Variance 0.00156 0.00155 0.00155 0.00155 0.00154 0.00154 0.00154
## Cumulative Proportion  0.49577 0.49732 0.49887 0.50042 0.50196 0.50350 0.50504
##                           PC197   PC198   PC199   PC200   PC201   PC202   PC203
## Standard deviation     1.12547 1.12339 1.12143 1.12064 1.11984 1.11871 1.11744
## Proportion of Variance 0.00153 0.00153 0.00152 0.00152 0.00152 0.00151 0.00151
## Cumulative Proportion  0.50657 0.50810 0.50962 0.51113 0.51265 0.51416 0.51567
##                           PC204  PC205  PC206   PC207   PC208   PC209   PC210
## Standard deviation     1.11654 1.1150 1.1139 1.11152 1.11091 1.11009 1.10747
## Proportion of Variance 0.00151 0.0015 0.0015 0.00149 0.00149 0.00149 0.00148
## Cumulative Proportion  0.51718 0.5187 0.5202 0.52168 0.52317 0.52466 0.52614
##                           PC211   PC212   PC213   PC214   PC215   PC216   PC217
## Standard deviation     1.10617 1.10435 1.10282 1.10030 1.10026 1.09925 1.09857
## Proportion of Variance 0.00148 0.00147 0.00147 0.00146 0.00146 0.00146 0.00146
## Cumulative Proportion  0.52762 0.52910 0.53057 0.53203 0.53350 0.53496 0.53642
##                           PC218   PC219   PC220   PC221   PC222   PC223   PC224
## Standard deviation     1.09734 1.09534 1.09431 1.09327 1.09182 1.09057 1.08916
## Proportion of Variance 0.00146 0.00145 0.00145 0.00145 0.00144 0.00144 0.00143
## Cumulative Proportion  0.53787 0.53932 0.54077 0.54222 0.54366 0.54510 0.54653
##                           PC225   PC226   PC227   PC228   PC229   PC230   PC231
## Standard deviation     1.08801 1.08673 1.08546 1.08436 1.08402 1.08252 1.08132
## Proportion of Variance 0.00143 0.00143 0.00142 0.00142 0.00142 0.00142 0.00141
## Cumulative Proportion  0.54796 0.54939 0.55082 0.55224 0.55366 0.55508 0.55649
##                           PC232   PC233  PC234  PC235  PC236   PC237   PC238
## Standard deviation     1.08006 1.07810 1.0766 1.0759 1.0742 1.07308 1.07086
## Proportion of Variance 0.00141 0.00141 0.0014 0.0014 0.0014 0.00139 0.00139
## Cumulative Proportion  0.55790 0.55931 0.5607 0.5621 0.5635 0.56489 0.56628
##                           PC239   PC240   PC241   PC242   PC243   PC244   PC245
## Standard deviation     1.06965 1.06936 1.06820 1.06586 1.06487 1.06394 1.06201
## Proportion of Variance 0.00138 0.00138 0.00138 0.00137 0.00137 0.00137 0.00136
## Cumulative Proportion  0.56767 0.56905 0.57043 0.57180 0.57317 0.57454 0.57590
##                           PC246   PC247   PC248   PC249   PC250   PC251   PC252
## Standard deviation     1.06051 1.05927 1.05881 1.05722 1.05620 1.05524 1.05380
## Proportion of Variance 0.00136 0.00136 0.00136 0.00135 0.00135 0.00135 0.00134
## Cumulative Proportion  0.57726 0.57862 0.57998 0.58133 0.58268 0.58402 0.58537
##                           PC253   PC254   PC255   PC256   PC257   PC258   PC259
## Standard deviation     1.05225 1.05155 1.05051 1.04894 1.04665 1.04599 1.04515
## Proportion of Variance 0.00134 0.00134 0.00133 0.00133 0.00132 0.00132 0.00132
## Cumulative Proportion  0.58671 0.58804 0.58938 0.59071 0.59203 0.59336 0.59468
##                           PC260   PC261   PC262   PC263   PC264  PC265  PC266
## Standard deviation     1.04331 1.04189 1.04110 1.04006 1.03949 1.0382 1.0367
## Proportion of Variance 0.00132 0.00131 0.00131 0.00131 0.00131 0.0013 0.0013
## Cumulative Proportion  0.59599 0.59731 0.59862 0.59992 0.60123 0.6025 0.6038
##                           PC267   PC268   PC269   PC270   PC271   PC272   PC273
## Standard deviation     1.03461 1.03377 1.03343 1.03118 1.03007 1.02851 1.02768
## Proportion of Variance 0.00129 0.00129 0.00129 0.00129 0.00128 0.00128 0.00128
## Cumulative Proportion  0.60513 0.60642 0.60771 0.60900 0.61028 0.61156 0.61284
##                           PC274   PC275   PC276   PC277   PC278   PC279   PC280
```

```
## Standard deviation     1.02683 1.02504 1.02470 1.02437 1.02296 1.01985 1.01911
## Proportion of Variance 0.00127 0.00127 0.00127 0.00127 0.00127 0.00126 0.00126
## Cumulative Proportion  0.61411 0.61538 0.61665 0.61792 0.61919 0.62044 0.62170
##                           PC281   PC282   PC283   PC284   PC285   PC286   PC287
## Standard deviation     1.01752 1.01639 1.01569 1.01469 1.01410 1.01265 1.01158
## Proportion of Variance 0.00125 0.00125 0.00125 0.00124 0.00124 0.00124 0.00124
## Cumulative Proportion  0.62295 0.62420 0.62545 0.62669 0.62794 0.62918 0.63041
##                           PC288   PC289   PC290   PC291   PC292   PC293   PC294
## Standard deviation     1.01124 1.00996 1.00824 1.00779 1.00666 1.00627 1.00390
## Proportion of Variance 0.00124 0.00123 0.00123 0.00123 0.00123 0.00122 0.00122
## Cumulative Proportion  0.63165 0.63288 0.63411 0.63534 0.63657 0.63779 0.63901
##                           PC295   PC296   PC297   PC298  PC299  PC300  PC301
## Standard deviation     1.00234 1.00124 1.00005 0.99928 0.9979 0.9973 0.9955
## Proportion of Variance 0.00121 0.00121 0.00121 0.00121 0.0012 0.0012 0.0012
## Cumulative Proportion  0.64022 0.64144 0.64265 0.64385 0.6451 0.6463 0.6475
##                          PC302  PC303   PC304   PC305   PC306   PC307   PC308
## Standard deviation     0.9952 0.9948 0.99218 0.99088 0.99068 0.98982 0.98891
## Proportion of Variance 0.0012 0.0012 0.00119 0.00119 0.00119 0.00118 0.00118
## Cumulative Proportion  0.6487 0.6499 0.65104 0.65223 0.65342 0.65460 0.65578
##                           PC309   PC310   PC311   PC312   PC313   PC314   PC315
## Standard deviation     0.98806 0.98734 0.98554 0.98472 0.98334 0.98252 0.98146
## Proportion of Variance 0.00118 0.00118 0.00117 0.00117 0.00117 0.00117 0.00116
## Cumulative Proportion  0.65696 0.65814 0.65932 0.66049 0.66166 0.66283 0.66399
##                           PC316   PC317   PC318   PC319   PC320   PC321   PC322
## Standard deviation     0.98051 0.97942 0.97810 0.97735 0.97506 0.97422 0.97225
## Proportion of Variance 0.00116 0.00116 0.00116 0.00116 0.00115 0.00115 0.00114
## Cumulative Proportion  0.66515 0.66631 0.66747 0.66863 0.66977 0.67092 0.67207
##                           PC323   PC324   PC325   PC326   PC327   PC328   PC329
## Standard deviation     0.97161 0.97100 0.97011 0.96903 0.96809 0.96669 0.96500
## Proportion of Variance 0.00114 0.00114 0.00114 0.00114 0.00113 0.00113 0.00113
## Cumulative Proportion  0.67321 0.67435 0.67549 0.67662 0.67775 0.67888 0.68001
##                           PC330   PC331   PC332   PC333   PC334   PC335   PC336
## Standard deviation     0.96376 0.96371 0.96290 0.96157 0.96120 0.95984 0.95891
## Proportion of Variance 0.00112 0.00112 0.00112 0.00112 0.00112 0.00111 0.00111
## Cumulative Proportion  0.68113 0.68226 0.68338 0.68450 0.68561 0.68673 0.68784
##                           PC337   PC338   PC339  PC340  PC341  PC342  PC343
## Standard deviation     0.95743 0.95722 0.95673 0.9556 0.9538 0.9531 0.9517
## Proportion of Variance 0.00111 0.00111 0.00111 0.0011 0.0011 0.0011 0.0011
## Cumulative Proportion  0.68895 0.69005 0.69116 0.6923 0.6934 0.6945 0.6956
##                           PC344   PC345   PC346   PC347   PC348   PC349   PC350
## Standard deviation     0.95082 0.95052 0.94940 0.94654 0.94526 0.94448 0.94338
## Proportion of Variance 0.00109 0.00109 0.00109 0.00108 0.00108 0.00108 0.00108
## Cumulative Proportion  0.69665 0.69775 0.69883 0.69992 0.70100 0.70208 0.70315
##                           PC351   PC352   PC353   PC354   PC355   PC356   PC357
## Standard deviation     0.94190 0.94106 0.93950 0.93862 0.93712 0.93542 0.93473
## Proportion of Variance 0.00107 0.00107 0.00107 0.00107 0.00106 0.00106 0.00106
## Cumulative Proportion  0.70423 0.70530 0.70636 0.70743 0.70849 0.70955 0.71061
##                           PC358   PC359   PC360   PC361   PC362   PC363   PC364
## Standard deviation     0.93386 0.93286 0.93180 0.93046 0.92983 0.92887 0.92714
## Proportion of Variance 0.00105 0.00105 0.00105 0.00105 0.00105 0.00104 0.00104
## Cumulative Proportion  0.71166 0.71271 0.71376 0.71481 0.71586 0.71690 0.71794
##                           PC365   PC366   PC367   PC368   PC369   PC370   PC371
## Standard deviation     0.92699 0.92672 0.92529 0.92439 0.92358 0.92228 0.92142
## Proportion of Variance 0.00104 0.00104 0.00104 0.00103 0.00103 0.00103 0.00103
```

```
## Cumulative Proportion  0.71898 0.72002 0.72105 0.72208 0.72312 0.72414 0.72517
##                           PC372   PC373   PC374   PC375   PC376   PC377   PC378
## Standard deviation     0.92076 0.91970 0.91878 0.91765 0.91637 0.91515 0.91477
## Proportion of Variance 0.00103 0.00102 0.00102 0.00102 0.00102 0.00101 0.00101
## Cumulative Proportion  0.72620 0.72722 0.72824 0.72926 0.73027 0.73129 0.73230
##                           PC379   PC380  PC381  PC382  PC383  PC384  PC385
## Standard deviation     0.91331 0.91247 0.9110 0.9106 0.9095 0.9082 0.9071
## Proportion of Variance 0.00101 0.00101 0.0010 0.0010 0.0010 0.0010 0.0010
## Cumulative Proportion  0.73331 0.73431 0.7353 0.7363 0.7373 0.7383 0.7393
##                           PC386   PC387   PC388   PC389   PC390   PC391   PC392
## Standard deviation     0.90603 0.90534 0.90343 0.90154 0.90052 0.89974 0.89907
## Proportion of Variance 0.00099 0.00099 0.00099 0.00098 0.00098 0.00098 0.00098
## Cumulative Proportion  0.74030 0.74130 0.74228 0.74327 0.74425 0.74522 0.74620
##                           PC393   PC394   PC395   PC396   PC397   PC398   PC399
## Standard deviation     0.89862 0.89766 0.89675 0.89583 0.89430 0.89420 0.89287
## Proportion of Variance 0.00098 0.00097 0.00097 0.00097 0.00097 0.00097 0.00096
## Cumulative Proportion  0.74718 0.74815 0.74913 0.75010 0.75106 0.75203 0.75299
##                           PC400   PC401   PC402   PC403   PC404   PC405   PC406
## Standard deviation     0.89247 0.89083 0.88935 0.88889 0.88867 0.88742 0.88605
## Proportion of Variance 0.00096 0.00096 0.00096 0.00096 0.00095 0.00095 0.00095
## Cumulative Proportion  0.75396 0.75492 0.75587 0.75683 0.75778 0.75874 0.75968
##                           PC407   PC408   PC409   PC410   PC411   PC412   PC413
## Standard deviation     0.88455 0.88373 0.88292 0.88117 0.88101 0.87986 0.87915
## Proportion of Variance 0.00095 0.00094 0.00094 0.00094 0.00094 0.00094 0.00093
## Cumulative Proportion  0.76063 0.76158 0.76252 0.76346 0.76440 0.76533 0.76627
##                           PC414   PC415   PC416   PC417   PC418   PC419   PC420
## Standard deviation     0.87710 0.87644 0.87510 0.87433 0.87379 0.87229 0.87207
## Proportion of Variance 0.00093 0.00093 0.00093 0.00092 0.00092 0.00092 0.00092
## Cumulative Proportion  0.76720 0.76812 0.76905 0.76998 0.77090 0.77182 0.77274
##                           PC421   PC422   PC423   PC424   PC425   PC426  PC427
## Standard deviation     0.87054 0.87022 0.86830 0.86699 0.86605 0.86582 0.8638
## Proportion of Variance 0.00092 0.00092 0.00091 0.00091 0.00091 0.00091 0.0009
## Cumulative Proportion  0.77365 0.77457 0.77548 0.77639 0.77730 0.77820 0.7791
##                           PC428  PC429  PC430   PC431   PC432   PC433   PC434
## Standard deviation     0.8633 0.8626 0.8610 0.85986 0.85897 0.85812 0.85751
## Proportion of Variance 0.0009 0.0009 0.0009 0.00089 0.00089 0.00089 0.00089
## Cumulative Proportion  0.7800 0.7809 0.7818 0.78270 0.78359 0.78448 0.78537
##                           PC435   PC436   PC437   PC438   PC439   PC440   PC441
## Standard deviation     0.85651 0.85500 0.85378 0.85209 0.85118 0.84970 0.84882
## Proportion of Variance 0.00089 0.00088 0.00088 0.00088 0.00088 0.00087 0.00087
## Cumulative Proportion  0.78626 0.78714 0.78802 0.78890 0.78978 0.79065 0.79152
##                           PC442   PC443   PC444   PC445   PC446   PC447   PC448
## Standard deviation     0.84868 0.84753 0.84603 0.84568 0.84489 0.84376 0.84219
## Proportion of Variance 0.00087 0.00087 0.00087 0.00086 0.00086 0.00086 0.00086
## Cumulative Proportion  0.79239 0.79326 0.79413 0.79499 0.79585 0.79671 0.79757
##                           PC449   PC450   PC451   PC452   PC453   PC454   PC455
## Standard deviation     0.84127 0.84098 0.83988 0.83870 0.83783 0.83707 0.83633
## Proportion of Variance 0.00086 0.00086 0.00085 0.00085 0.00085 0.00085 0.00085
## Cumulative Proportion  0.79843 0.79928 0.80014 0.80099 0.80183 0.80268 0.80353
##                           PC456   PC457   PC458   PC459   PC460   PC461   PC462
## Standard deviation     0.83565 0.83533 0.83411 0.83294 0.83283 0.83174 0.83111
## Proportion of Variance 0.00084 0.00084 0.00084 0.00084 0.00084 0.00084 0.00084
## Cumulative Proportion  0.80437 0.80522 0.80606 0.80690 0.80774 0.80857 0.80941
##                           PC463   PC464   PC465   PC466   PC467   PC468   PC469
```

```
## Standard deviation     0.83043 0.82932 0.82830 0.82735 0.82621 0.82490 0.82354
## Proportion of Variance 0.00083 0.00083 0.00083 0.00083 0.00083 0.00082 0.00082
## Cumulative Proportion  0.81024 0.81107 0.81190 0.81273 0.81356 0.81438 0.81520
##                           PC470   PC471   PC472   PC473   PC474   PC475   PC476
## Standard deviation     0.82284 0.82108 0.82103 0.81942 0.81816 0.81723 0.81705
## Proportion of Variance 0.00082 0.00082 0.00082 0.00081 0.00081 0.00081 0.00081
## Cumulative Proportion  0.81602 0.81683 0.81765 0.81846 0.81927 0.82008 0.82088
##                          PC477  PC478  PC479  PC480  PC481  PC482   PC483
## Standard deviation     0.8154 0.8152 0.8137 0.8132 0.8125 0.8118 0.80939
## Proportion of Variance 0.0008 0.0008 0.0008 0.0008 0.0008 0.0008 0.00079
## Cumulative Proportion  0.8217 0.8225 0.8233 0.8241 0.8249 0.8257 0.82648
##                           PC484   PC485   PC486   PC487   PC488   PC489   PC490
## Standard deviation     0.80784 0.80706 0.80688 0.80617 0.80454 0.80393 0.80377
## Proportion of Variance 0.00079 0.00079 0.00079 0.00079 0.00078 0.00078 0.00078
## Cumulative Proportion  0.82727 0.82806 0.82884 0.82963 0.83041 0.83119 0.83197
##                           PC491   PC492   PC493   PC494   PC495   PC496   PC497
## Standard deviation     0.80234 0.80166 0.80045 0.79862 0.79816 0.79769 0.79673
## Proportion of Variance 0.00078 0.00078 0.00077 0.00077 0.00077 0.00077 0.00077
## Cumulative Proportion  0.83275 0.83353 0.83430 0.83508 0.83585 0.83661 0.83738
##                           PC498   PC499   PC500   PC501   PC502   PC503   PC504
## Standard deviation     0.79560 0.79402 0.79320 0.79239 0.79199 0.79134 0.79054
## Proportion of Variance 0.00077 0.00076 0.00076 0.00076 0.00076 0.00076 0.00076
## Cumulative Proportion  0.83815 0.83891 0.83967 0.84043 0.84119 0.84195 0.84270
##                           PC505   PC506   PC507   PC508   PC509   PC510   PC511
## Standard deviation     0.78911 0.78795 0.78660 0.78542 0.78490 0.78443 0.78350
## Proportion of Variance 0.00075 0.00075 0.00075 0.00075 0.00074 0.00074 0.00074
## Cumulative Proportion  0.84345 0.84421 0.84495 0.84570 0.84644 0.84719 0.84793
##                           PC512   PC513   PC514   PC515   PC516   PC517   PC518
## Standard deviation     0.78253 0.78178 0.78079 0.77995 0.77942 0.77802 0.77647
## Proportion of Variance 0.00074 0.00074 0.00074 0.00074 0.00073 0.00073 0.00073
## Cumulative Proportion  0.84867 0.84941 0.85015 0.85088 0.85162 0.85235 0.85308
##                           PC519   PC520   PC521   PC522   PC523   PC524   PC525
## Standard deviation     0.77599 0.77579 0.77420 0.77321 0.77263 0.77150 0.77106
## Proportion of Variance 0.00073 0.00073 0.00072 0.00072 0.00072 0.00072 0.00072
## Cumulative Proportion  0.85381 0.85453 0.85526 0.85598 0.85670 0.85742 0.85814
##                           PC526   PC527   PC528   PC529   PC530   PC531   PC532
## Standard deviation     0.76992 0.76930 0.76884 0.76701 0.76566 0.76519 0.76447
## Proportion of Variance 0.00072 0.00072 0.00071 0.00071 0.00071 0.00071 0.00071
## Cumulative Proportion  0.85886 0.85957 0.86029 0.86100 0.86171 0.86242 0.86312
##                          PC533  PC534  PC535  PC536  PC537  PC538  PC539   PC540
## Standard deviation     0.7636 0.7624 0.7621 0.7602 0.7596 0.7588 0.7584 0.75691
## Proportion of Variance 0.0007 0.0007 0.0007 0.0007 0.0007 0.0007 0.0007 0.00069
## Cumulative Proportion  0.8638 0.8645 0.8652 0.8659 0.8666 0.8673 0.8680 0.86872
##                           PC541   PC542   PC543   PC544   PC545   PC546   PC547
## Standard deviation     0.75658 0.75606 0.75577 0.75521 0.75334 0.75255 0.75250
## Proportion of Variance 0.00069 0.00069 0.00069 0.00069 0.00069 0.00068 0.00068
## Cumulative Proportion  0.86941 0.87010 0.87079 0.87148 0.87217 0.87285 0.87353
##                           PC548   PC549   PC550   PC551   PC552   PC553   PC554
## Standard deviation     0.75096 0.75006 0.74907 0.74848 0.74709 0.74666 0.74460
## Proportion of Variance 0.00068 0.00068 0.00068 0.00068 0.00067 0.00067 0.00067
## Cumulative Proportion  0.87422 0.87490 0.87558 0.87625 0.87693 0.87760 0.87827
##                           PC555   PC556   PC557   PC558   PC559   PC560   PC561
## Standard deviation     0.74406 0.74279 0.74106 0.73942 0.73843 0.73802 0.73782
## Proportion of Variance 0.00067 0.00067 0.00066 0.00066 0.00066 0.00066 0.00066
```

```
## Cumulative Proportion  0.87894 0.87961 0.88027 0.88093 0.88159 0.88225 0.88291
##                           PC562   PC563   PC564   PC565   PC566   PC567   PC568
## Standard deviation     0.73577 0.73488 0.73435 0.73360 0.73250 0.73226 0.73149
## Proportion of Variance 0.00065 0.00065 0.00065 0.00065 0.00065 0.00065 0.00065
## Cumulative Proportion  0.88356 0.88422 0.88487 0.88552 0.88617 0.88682 0.88746
##                           PC569   PC570   PC571   PC572   PC573   PC574   PC575
## Standard deviation     0.73052 0.72892 0.72765 0.72660 0.72589 0.72471 0.72454
## Proportion of Variance 0.00065 0.00064 0.00064 0.00064 0.00064 0.00064 0.00063
## Cumulative Proportion  0.88811 0.88875 0.88939 0.89003 0.89067 0.89130 0.89194
##                           PC576   PC577   PC578   PC579   PC580   PC581   PC582
## Standard deviation     0.72350 0.72232 0.72109 0.72058 0.71937 0.71853 0.71797
## Proportion of Variance 0.00063 0.00063 0.00063 0.00063 0.00063 0.00062 0.00062
## Cumulative Proportion  0.89257 0.89320 0.89383 0.89446 0.89508 0.89571 0.89633
##                           PC583   PC584   PC585   PC586   PC587   PC588   PC589
## Standard deviation     0.71588 0.71487 0.71434 0.71358 0.71269 0.71148 0.71081
## Proportion of Variance 0.00062 0.00062 0.00062 0.00062 0.00061 0.00061 0.00061
## Cumulative Proportion  0.89695 0.89757 0.89819 0.89880 0.89942 0.90003 0.90064
##                           PC590   PC591   PC592   PC593  PC594  PC595  PC596
## Standard deviation     0.70981 0.70900 0.70831 0.70736 0.7063 0.7059 0.7040
## Proportion of Variance 0.00061 0.00061 0.00061 0.00061 0.0006 0.0006 0.0006
## Cumulative Proportion  0.90125 0.90186 0.90246 0.90307 0.9037 0.9043 0.9049
##                          PC597  PC598  PC599   PC600   PC601   PC602   PC603
## Standard deviation     0.7039 0.7032 0.7016 0.70093 0.69953 0.69878 0.69802
## Proportion of Variance 0.0006 0.0006 0.0006 0.00059 0.00059 0.00059 0.00059
## Cumulative Proportion  0.9055 0.9061 0.9067 0.90726 0.90785 0.90844 0.90903
##                           PC604   PC605   PC606   PC607   PC608   PC609   PC610
## Standard deviation     0.69787 0.69586 0.69451 0.69384 0.69304 0.69148 0.69072
## Proportion of Variance 0.00059 0.00059 0.00058 0.00058 0.00058 0.00058 0.00058
## Cumulative Proportion  0.90962 0.91021 0.91079 0.91137 0.91195 0.91253 0.91311
##                           PC611   PC612   PC613   PC614   PC615   PC616   PC617
## Standard deviation     0.68999 0.68967 0.68826 0.68628 0.68566 0.68457 0.68333
## Proportion of Variance 0.00058 0.00058 0.00057 0.00057 0.00057 0.00057 0.00056
## Cumulative Proportion  0.91368 0.91426 0.91483 0.91540 0.91597 0.91654 0.91710
##                           PC618   PC619   PC620   PC621   PC622   PC623   PC624
## Standard deviation     0.68290 0.68060 0.68020 0.67985 0.67920 0.67811 0.67752
## Proportion of Variance 0.00056 0.00056 0.00056 0.00056 0.00056 0.00056 0.00056
## Cumulative Proportion  0.91766 0.91822 0.91878 0.91934 0.91990 0.92046 0.92101
##                           PC625   PC626   PC627   PC628   PC629   PC630   PC631
## Standard deviation     0.67620 0.67516 0.67366 0.67359 0.67239 0.67208 0.67148
## Proportion of Variance 0.00055 0.00055 0.00055 0.00055 0.00055 0.00055 0.00055
## Cumulative Proportion  0.92156 0.92212 0.92266 0.92321 0.92376 0.92431 0.92485
##                           PC632   PC633   PC634   PC635   PC636   PC637   PC638
## Standard deviation     0.66997 0.66867 0.66703 0.66620 0.66579 0.66552 0.66402
## Proportion of Variance 0.00054 0.00054 0.00054 0.00054 0.00054 0.00054 0.00053
## Cumulative Proportion  0.92539 0.92593 0.92647 0.92701 0.92754 0.92808 0.92861
##                           PC639   PC640   PC641   PC642   PC643   PC644   PC645
## Standard deviation     0.66332 0.66230 0.66159 0.66061 0.65988 0.65880 0.65783
## Proportion of Variance 0.00053 0.00053 0.00053 0.00053 0.00053 0.00052 0.00052
## Cumulative Proportion  0.92915 0.92968 0.93021 0.93073 0.93126 0.93178 0.93231
##                           PC646   PC647   PC648   PC649   PC650   PC651   PC652
## Standard deviation     0.65703 0.65646 0.65555 0.65495 0.65380 0.65258 0.65126
## Proportion of Variance 0.00052 0.00052 0.00052 0.00052 0.00052 0.00051 0.00051
## Cumulative Proportion  0.93283 0.93335 0.93387 0.93439 0.93491 0.93542 0.93593
##                           PC653   PC654   PC655   PC656   PC657  PC658  PC659
```

```
## Standard deviation     0.65010 0.64931 0.64884 0.64758 0.64641 0.6455 0.6447
## Proportion of Variance 0.00051 0.00051 0.00051 0.00051 0.00051 0.0005 0.0005
## Cumulative Proportion  0.93644 0.93695 0.93746 0.93797 0.93848 0.9390 0.9395
##                          PC660  PC661  PC662   PC663   PC664   PC665   PC666
## Standard deviation     0.6428 0.6421 0.6409 0.63878 0.63862 0.63780 0.63736
## Proportion of Variance 0.0005 0.0005 0.0005 0.00049 0.00049 0.00049 0.00049
## Cumulative Proportion  0.9400 0.9405 0.9410 0.94147 0.94196 0.94246 0.94295
##                           PC667   PC668   PC669   PC670   PC671   PC672   PC673
## Standard deviation     0.63546 0.63402 0.63303 0.63182 0.63155 0.63121 0.62996
## Proportion of Variance 0.00049 0.00049 0.00048 0.00048 0.00048 0.00048 0.00048
## Cumulative Proportion  0.94343 0.94392 0.94441 0.94489 0.94537 0.94585 0.94633
##                           PC674   PC675   PC676   PC677   PC678   PC679   PC680
## Standard deviation     0.62921 0.62821 0.62675 0.62631 0.62454 0.62368 0.62293
## Proportion of Variance 0.00048 0.00048 0.00047 0.00047 0.00047 0.00047 0.00047
## Cumulative Proportion  0.94681 0.94729 0.94776 0.94824 0.94871 0.94918 0.94965
##                           PC681   PC682   PC683   PC684   PC685   PC686   PC687
## Standard deviation     0.62114 0.62084 0.62044 0.61872 0.61817 0.61681 0.61484
## Proportion of Variance 0.00047 0.00047 0.00047 0.00046 0.00046 0.00046 0.00046
## Cumulative Proportion  0.95012 0.95058 0.95105 0.95151 0.95197 0.95243 0.95289
##                           PC688   PC689   PC690   PC691   PC692   PC693   PC694
## Standard deviation     0.61445 0.61323 0.61212 0.61083 0.60984 0.60893 0.60729
## Proportion of Variance 0.00046 0.00045 0.00045 0.00045 0.00045 0.00045 0.00045
## Cumulative Proportion  0.95335 0.95380 0.95425 0.95470 0.95515 0.95560 0.95605
##                           PC695   PC696   PC697   PC698   PC699   PC700   PC701
## Standard deviation     0.60697 0.60665 0.60496 0.60386 0.60311 0.60240 0.60092
## Proportion of Variance 0.00045 0.00045 0.00044 0.00044 0.00044 0.00044 0.00044
## Cumulative Proportion  0.95649 0.95694 0.95738 0.95782 0.95826 0.95870 0.95914
##                           PC702   PC703   PC704   PC705   PC706   PC707   PC708
## Standard deviation     0.59900 0.59831 0.59759 0.59680 0.59492 0.59346 0.59331
## Proportion of Variance 0.00043 0.00043 0.00043 0.00043 0.00043 0.00043 0.00043
## Cumulative Proportion  0.95957 0.96000 0.96044 0.96087 0.96129 0.96172 0.96215
##                           PC709   PC710   PC711   PC712   PC713   PC714   PC715
## Standard deviation     0.59208 0.59183 0.58988 0.58831 0.58771 0.58729 0.58547
## Proportion of Variance 0.00042 0.00042 0.00042 0.00042 0.00042 0.00042 0.00041
## Cumulative Proportion  0.96257 0.96299 0.96341 0.96383 0.96425 0.96467 0.96508
##                           PC716   PC717   PC718   PC719   PC720   PC721  PC722
## Standard deviation     0.58438 0.58292 0.58230 0.58189 0.58077 0.57967 0.5786
## Proportion of Variance 0.00041 0.00041 0.00041 0.00041 0.00041 0.00041 0.0004
## Cumulative Proportion  0.96550 0.96591 0.96632 0.96673 0.96713 0.96754 0.9679
##                          PC723  PC724  PC725  PC726  PC727  PC728  PC729   PC730
## Standard deviation     0.5777 0.5769 0.5759 0.5738 0.5731 0.5720 0.5716 0.56887
## Proportion of Variance 0.0004 0.0004 0.0004 0.0004 0.0004 0.0004 0.0004 0.00039
## Cumulative Proportion  0.9684 0.9688 0.9691 0.9696 0.9699 0.9703 0.9707 0.97113
##                           PC731   PC732   PC733   PC734   PC735   PC736   PC737
## Standard deviation     0.56843 0.56743 0.56616 0.56580 0.56496 0.56342 0.56118
## Proportion of Variance 0.00039 0.00039 0.00039 0.00039 0.00039 0.00038 0.00038
## Cumulative Proportion  0.97152 0.97191 0.97230 0.97268 0.97307 0.97345 0.97383
##                           PC738   PC739   PC740   PC741   PC742   PC743   PC744
## Standard deviation     0.56061 0.55822 0.55788 0.55607 0.55486 0.55368 0.55279
## Proportion of Variance 0.00038 0.00038 0.00038 0.00037 0.00037 0.00037 0.00037
## Cumulative Proportion  0.97421 0.97459 0.97497 0.97534 0.97571 0.97608 0.97645
##                           PC745   PC746   PC747   PC748   PC749   PC750   PC751
## Standard deviation     0.55115 0.55100 0.54999 0.54752 0.54676 0.54618 0.54474
## Proportion of Variance 0.00037 0.00037 0.00037 0.00036 0.00036 0.00036 0.00036
```

```
## Cumulative Proportion  0.97682 0.97719 0.97755 0.97792 0.97828 0.97864 0.97900
##                          PC752   PC753   PC754   PC755   PC756   PC757   PC758
## Standard deviation     0.54360 0.54296 0.54239 0.54160 0.54001 0.53715 0.53574
## Proportion of Variance 0.00036 0.00036 0.00036 0.00035 0.00035 0.00035 0.00035
## Cumulative Proportion  0.97935 0.97971 0.98007 0.98042 0.98077 0.98112 0.98147
##                          PC759   PC760   PC761   PC762   PC763   PC764   PC765
## Standard deviation     0.53540 0.53401 0.53207 0.53170 0.52929 0.52845 0.52736
## Proportion of Variance 0.00035 0.00034 0.00034 0.00034 0.00034 0.00034 0.00034
## Cumulative Proportion  0.98182 0.98216 0.98250 0.98285 0.98318 0.98352 0.98386
##                          PC766   PC767   PC768   PC769   PC770   PC771   PC772
## Standard deviation     0.52559 0.52394 0.52377 0.52064 0.51915 0.51701 0.51664
## Proportion of Variance 0.00033 0.00033 0.00033 0.00033 0.00033 0.00032 0.00032
## Cumulative Proportion  0.98419 0.98452 0.98486 0.98518 0.98551 0.98583 0.98616
##                          PC773   PC774   PC775   PC776   PC777   PC778   PC779
## Standard deviation     0.51616 0.51448 0.51357 0.51116 0.50984 0.50869 0.50786
## Proportion of Variance 0.00032 0.00032 0.00032 0.00032 0.00031 0.00031 0.00031
## Cumulative Proportion  0.98648 0.98680 0.98712 0.98743 0.98775 0.98806 0.98837
##                          PC780   PC781  PC782  PC783  PC784  PC785  PC786
## Standard deviation     0.50644 0.50530 0.5015 0.5001 0.4995 0.4962 0.4960
## Proportion of Variance 0.00031 0.00031 0.0003 0.0003 0.0003 0.0003 0.0003
## Cumulative Proportion  0.98868 0.98899 0.9893 0.9896 0.9899 0.9902 0.9905
##                          PC787   PC788   PC789   PC790   PC791   PC792   PC793
## Standard deviation     0.4958 0.49308 0.49083 0.48829 0.48776 0.48655 0.48489
## Proportion of Variance 0.0003 0.00029 0.00029 0.00029 0.00029 0.00029 0.00028
## Cumulative Proportion  0.9908 0.99109 0.99138 0.99166 0.99195 0.99224 0.99252
##                          PC794   PC795   PC796   PC797   PC798   PC799   PC800
## Standard deviation     0.48118 0.47875 0.47722 0.47498 0.47302 0.47218 0.47037
## Proportion of Variance 0.00028 0.00028 0.00028 0.00027 0.00027 0.00027 0.00027
## Cumulative Proportion  0.99280 0.99308 0.99336 0.99363 0.99390 0.99417 0.99444
##                          PC801   PC802   PC803   PC804   PC805   PC806   PC807
## Standard deviation     0.46914 0.46664 0.46451 0.46249 0.45923 0.45802 0.45349
## Proportion of Variance 0.00027 0.00026 0.00026 0.00026 0.00026 0.00025 0.00025
## Cumulative Proportion  0.99470 0.99497 0.99523 0.99549 0.99574 0.99599 0.99624
##                          PC808   PC809   PC810   PC811   PC812   PC813   PC814
## Standard deviation     0.45041 0.44887 0.44746 0.44444 0.44245 0.44011 0.43529
## Proportion of Variance 0.00025 0.00024 0.00024 0.00024 0.00024 0.00023 0.00023
## Cumulative Proportion  0.99649 0.99673 0.99697 0.99721 0.99745 0.99768 0.99791
##                          PC815   PC816   PC817   PC818  PC819   PC820   PC821
## Standard deviation     0.43055 0.42272 0.41888 0.41199 0.4078 0.39543 0.38897
## Proportion of Variance 0.00022 0.00022 0.00021 0.00021 0.0002 0.00019 0.00018
## Cumulative Proportion  0.99814 0.99835 0.99856 0.99877 0.9990 0.99916 0.99934
##                          PC822   PC823   PC824   PC825   PC826   PC827
## Standard deviation     0.37748 0.37261 0.37134 0.29806 0.17815 0.05832
## Proportion of Variance 0.00017 0.00017 0.00017 0.00011 0.00004 0.00000
## Cumulative Proportion  0.99952 0.99968 0.99985 0.99996 1.00000 1.00000
```

For our model we decided to go with a random forest. Maybe certain authors use words that no other author does - for example a sports columnist will use the word 'umpire' frequently but a non-sports columnist would never use it. A decision tree is perfect for using such clear-cut divisions in the data to separate classes, and a random forest might be able to handle complexity of the dataset better than a simple decision tree.

We also tried training our model with several different datasets: the raw TF-IDF matrix, 10 PCA columns, 158 PCA columns, and all PCA columns, to see what did the best and how long it took.

Unsurprisingly, datasets with more features are more accurate but are much slower. Interestingly, the all

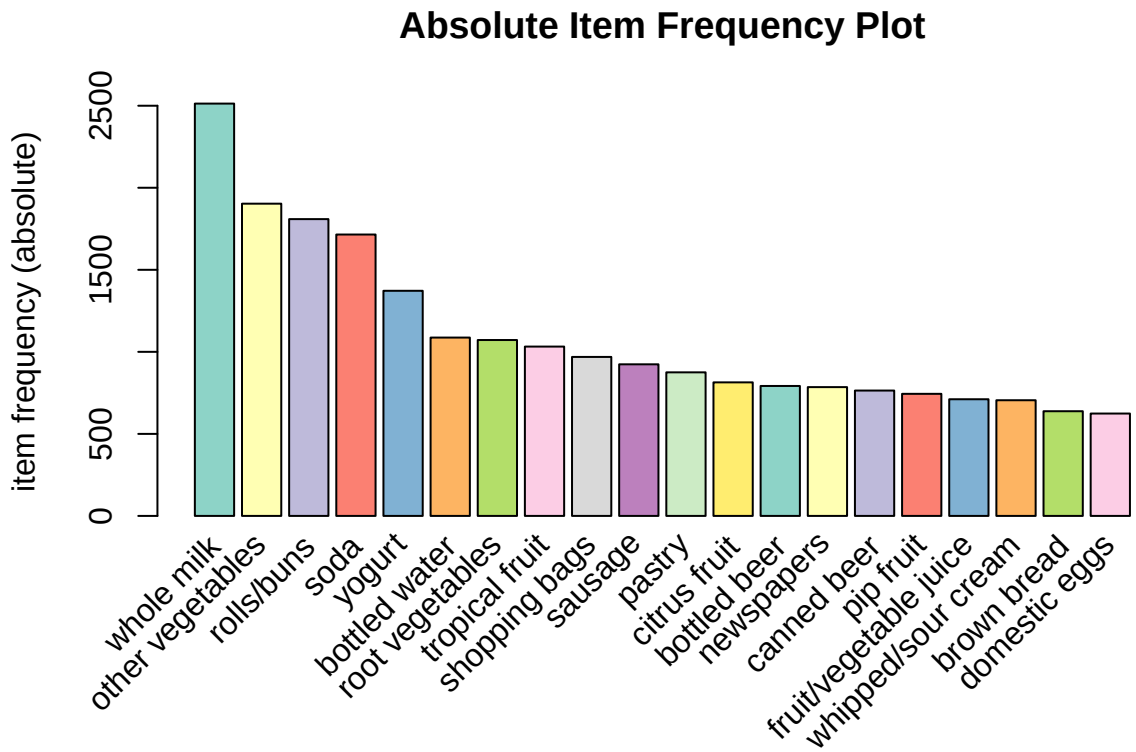
PCA columns dataset is outperformed by both the TF-IDF matrix and the 158 PCA columns dataset. Likely it encodes a lot of useless information in it's later columns that only confuse the model.

We can see that using 158 PCA columns has almost as good an accuracy as the TF-IDF matrix but takes less than half the time to train. What model is best ultimately depends on what the use case is. If accuracy is the most important, use the TF-IDF matrix without PCA reduction. If you care how long it takes, use some PCA columns, with how many you use depending on how long you can afford for the algorithm to run. Just don't use too many, or you'll start getting worse accuracies.

```
## [1] 62.04
## [1] 39.28
## [1] 56.68
## [1] 51.12
```

6. Association Rule Mining

We first read in a large list of baskets as individual transactions and plot the top 20 most frequently occurring items:



Now that the data is in transaction form, we can use the A-Priori Algorithm to create rules. Listed below are the first ten rules:

```
##      lhs                          rhs                support     confidence
## [1]  {honey}                   => {whole milk}       0.001118454 0.7333333
```

```
## [2]  {cereals}                    => {whole milk}       0.003660397 0.6428571
## [3]  {rice}                       => {whole milk}       0.004677173 0.6133333
## [4]  {liver loaf,yogurt}          => {whole milk}       0.001016777 0.6666667
## [5]  {curd cheese,tropical fruit} => {other vegetables} 0.001016777 0.6666667
## [6]  {curd cheese,rolls/buns}     => {whole milk}       0.001016777 0.6250000
## [7]  {cleaner,other vegetables}   => {whole milk}       0.001016777 0.6250000
## [8]  {liquor,red/blush wine}      => {bottled beer}     0.001931876 0.9047619
## [9]  {butter,jam}                 => {whole milk}       0.001016777 0.8333333
## [10] {jam,root vegetables}        => {whole milk}       0.001321810 0.6842105
##      coverage    lift      count
## [1]  0.001525165  2.870009 11
## [2]  0.005693950  2.515917 36
## [3]  0.007625826  2.400371 46
## [4]  0.001525165  2.609099 10
## [5]  0.001525165  3.445437 10
## [6]  0.001626843  2.446031 10
## [7]  0.001626843  2.446031 10
## [8]  0.002135231 11.235269 19
## [9]  0.001220132  3.261374 10
## [10] 0.001931876  2.677760 13
```

Now we analyze a couple of specific rules. Our examples here attempt to predict what customers buy before and after buying soda and shopping bags. Below you can see a few of these predictions:

```
##     lhs       rhs          support    confidence coverage  lift      count
## [1] {soda} => {rolls/buns} 0.03833249 0.2198251  0.1743772 1.1951242 377
## [2] {soda} => {whole milk} 0.04006101 0.2297376  0.1743772 0.8991124 394
```

```
##     lhs                                     rhs    support     confidence
## [1] {coffee,misc. beverages}             => {soda} 0.001016777 0.7692308
## [2] {bottled beer,bottled water,sausage} => {soda} 0.001118454 0.7333333
##     coverage    lift     count
## [1] 0.001321810 4.411303 10
## [2] 0.001525165 4.205442 11
```

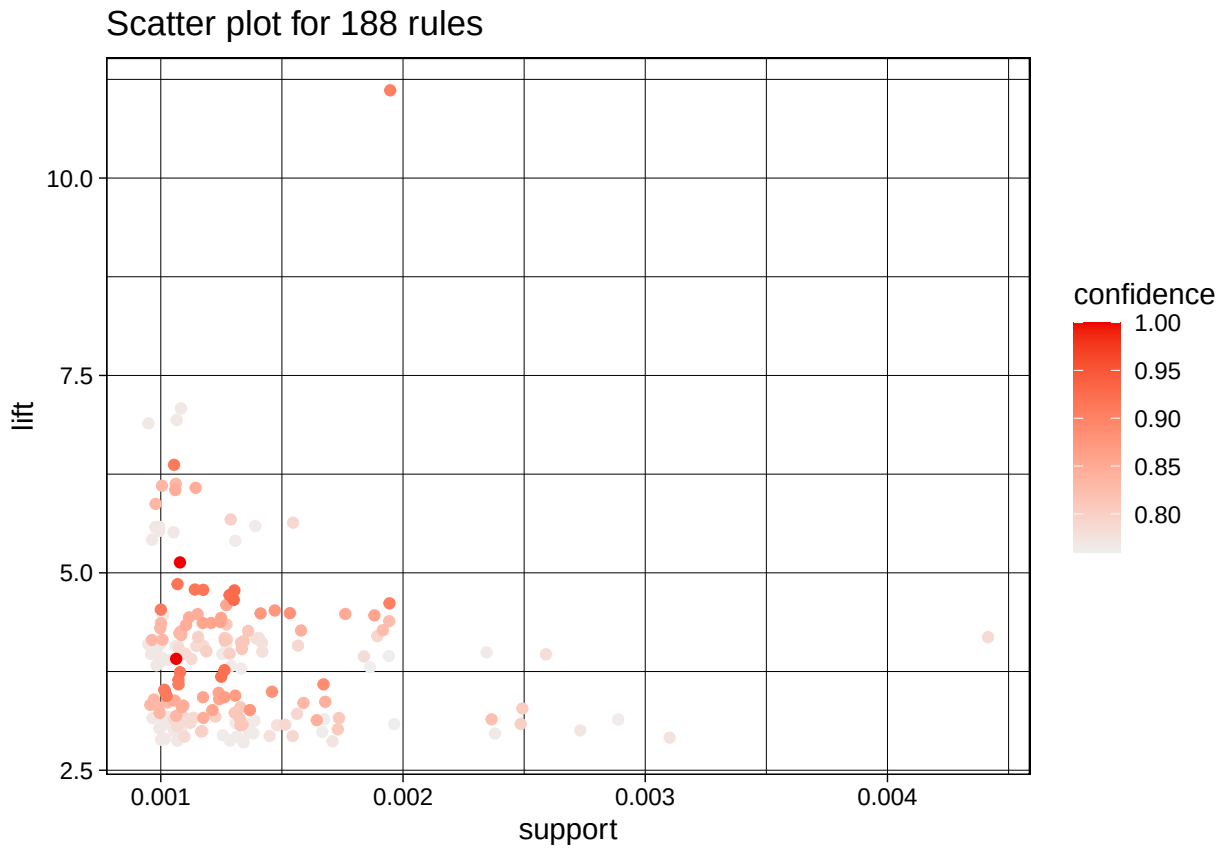
```
##     lhs                rhs                      support    confidence coverage
## [1] {shopping bags} => {canned beer}           0.01138790 0.1155831  0.09852567
## [2] {shopping bags} => {fruit/vegetable juice} 0.01067616 0.1083591  0.09852567
## [3] {shopping bags} => {pastry}                0.01189629 0.1207430  0.09852567
## [4] {shopping bags} => {sausage}               0.01565836 0.1589267  0.09852567
## [5] {shopping bags} => {bottled water}         0.01098119 0.1114551  0.09852567
##     lift     count
## [1] 1.487905 112
## [2] 1.498892 105
## [3] 1.357152 117
## [4] 1.691606 154
## [5] 1.008428 108
```

```
##     lhs                         rhs                 support confidence    coverage     lift count
## [1] {fruit/vegetable juice,
##      red/blush wine}        => {shopping bags} 0.001016777  0.5555556 0.001830198 5.638688    10
## [2] {brown bread,
```

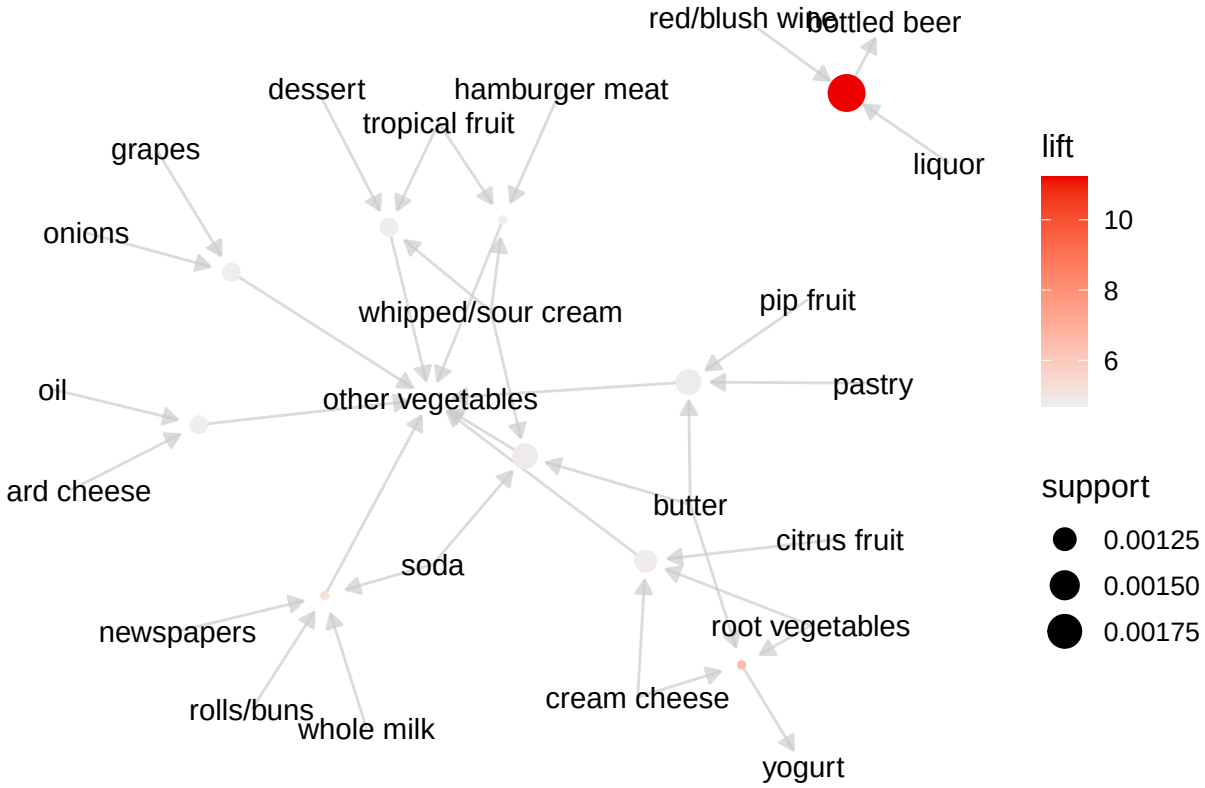
```
##      salty snack}          => {shopping bags} 0.001016777  0.4166667 0.002440264 4.229016    10
## [3] {canned beer,
##      domestic eggs}        => {shopping bags} 0.001423488  0.4000000 0.003558719 4.059856    14
## [4] {canned beer,
##      sausage}              => {shopping bags} 0.002643620  0.4193548 0.006304016 4.256300    26
## [5] {canned beer,
##      sausage,
##      soda}                 => {shopping bags} 0.001016777  0.4166667 0.002440264 4.229016    10
```

Now we will visualize the rules below

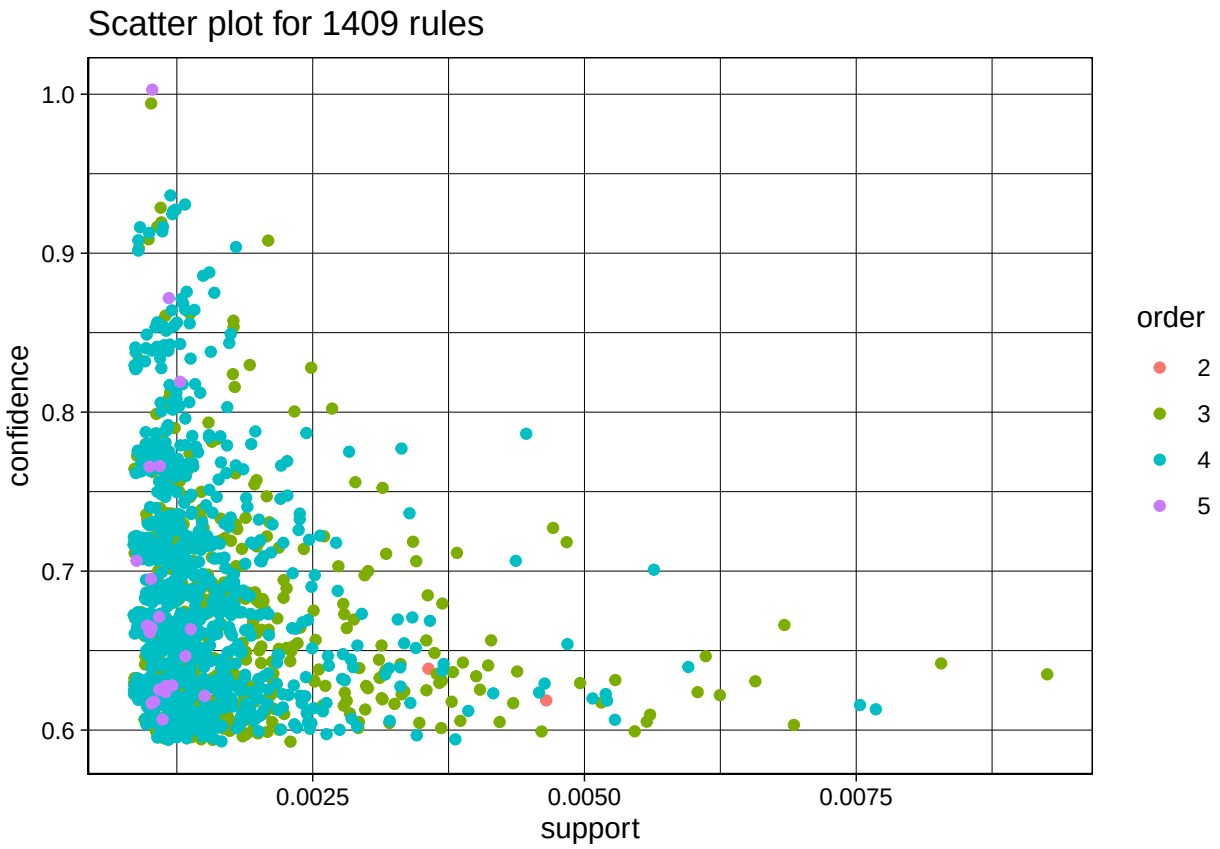
Scatter plot with conditional subset: Rules with confidence values above 0.75 are plotted here. The lift, or "interestingness" measures here indicate that these rules have a complementary effect. If an item is purchased, another item is likely to be purchased with it, not instead of it. There is a clear outlier here. This could be something like milk and cereal, which is often purchased together. All of these points have high confidence, but don't come into play very often with this data set.



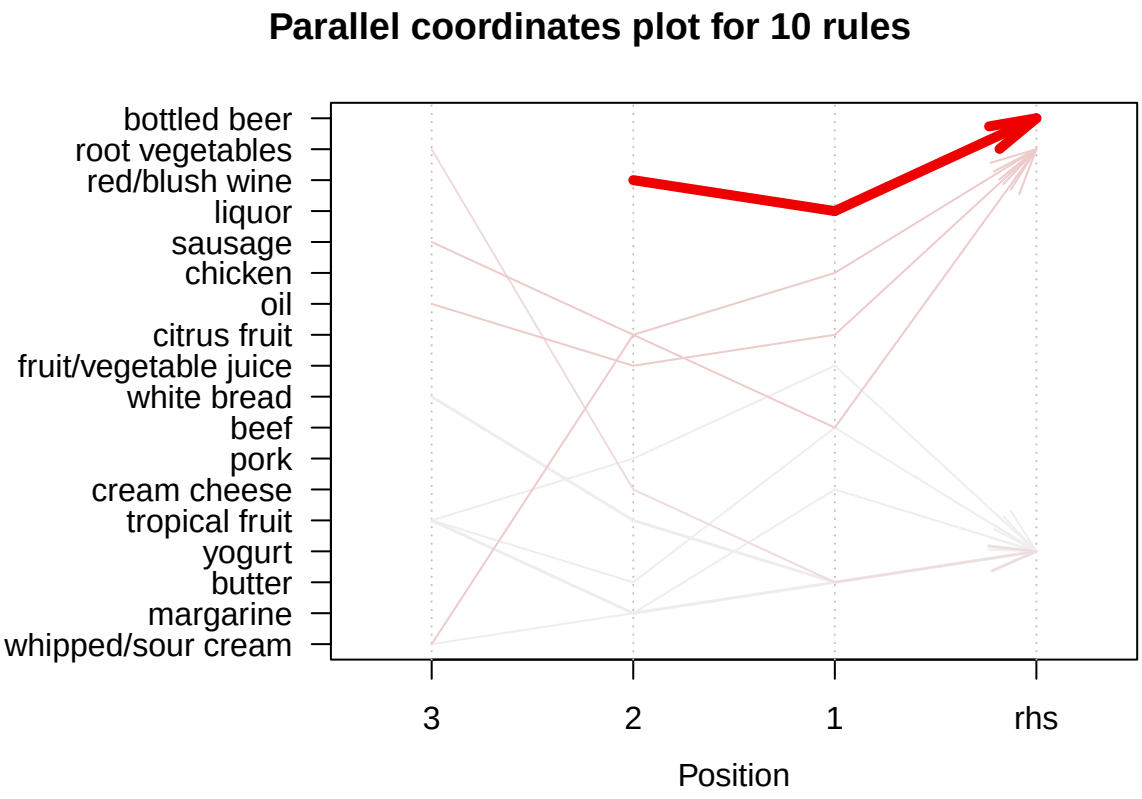
Graph-based visualization: Here we see some interesting relationships. Clearly different types of alcohol are bought together.



Two-key plot: The support for order 3 varies widely here.



Parallel coordinate plot: Here we can see that yogurt and root vegetables purchases are a likely occurrence after buying at least one of a number of items.



Summary:

Vegetables and whole milk are a major centerpiece in most grocery shopping transactions. Wine, liquor, and beer are bought together often. If customers buy beer, they're also likely to buy bags, perhaps to keep their plans under wraps. We can also predict with confidence that if customers buy coffee or other miscellaneous beverages, they will also buy soda. There are plenty of other rules that were created by the A-Priori Algorithm, and one could hone quite an advertising strategy based on this analysis alone.