STA380 Final Excercises

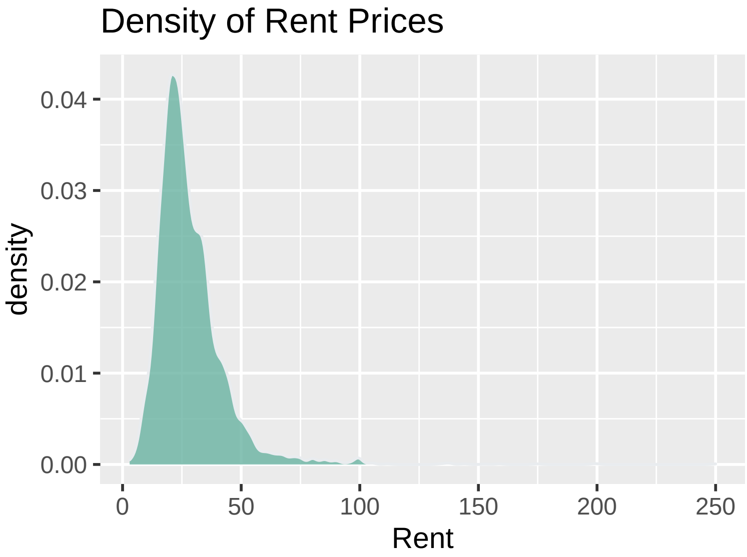
Rohan Garg, Nir Rauch, Jacob Rhymes, Rianna Patel

#Visual Story Telling 1: GreenBuildings

Regarding the analysis performed by the "stats guru," we believe that he does decent introductory exploration, but neglects a multitude of key factors including leasing rate and building quality in green versus nongreen buildings, which confounds his results.

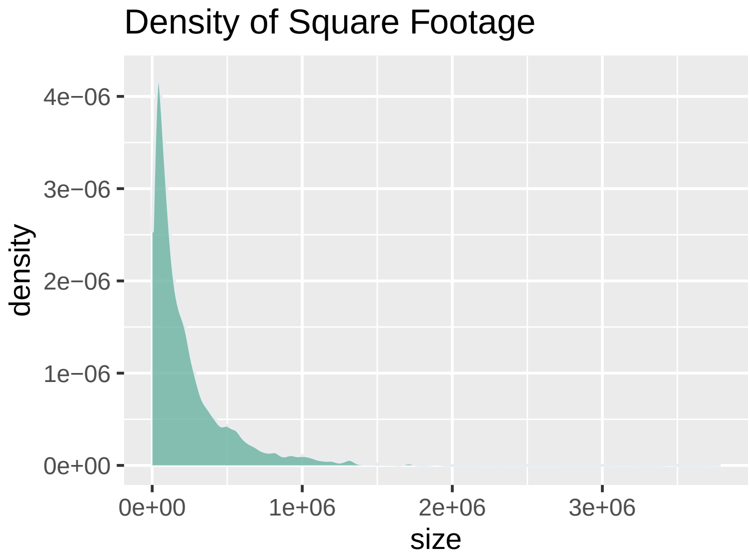
While our study does replicate some of his ideas, we provide a more complete analysis below: We began our exploratory analysis by creating density plots of some explanatory variables.

Rent Density:



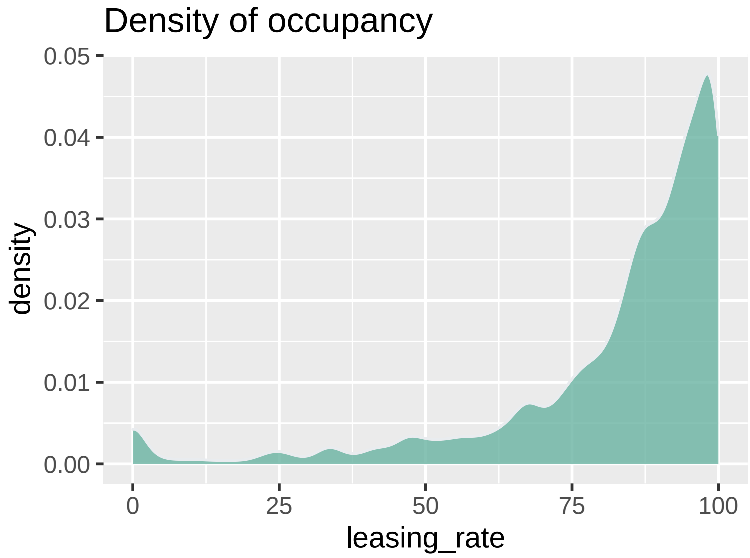
Rent density for all homes appears approxomately normal.

Square Footage Density:



Square footage density is concentrated at the lower end. This seems obvious.

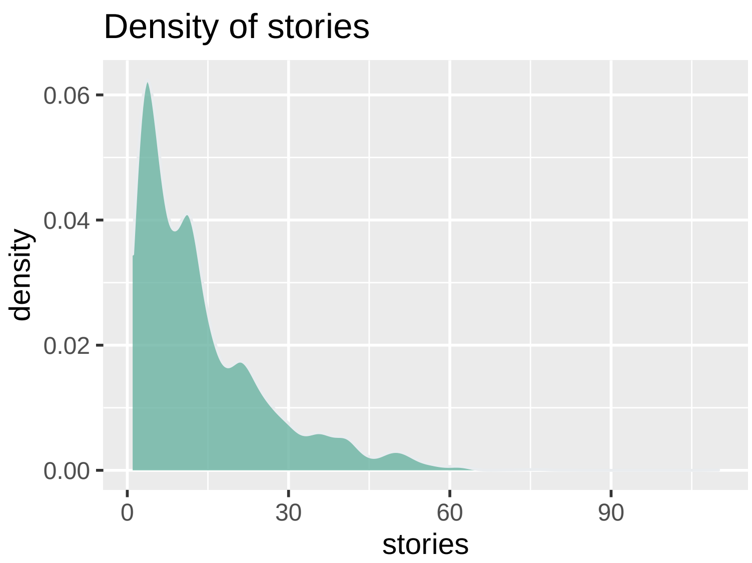
Leasing Rate Density:



Leasing rate is, for the most part, above 75%.

However, we see what the “status guru” noted: a small but noticeable amount of occupies are below 10%. We also decide to remove these, as they may be unoccupied or typos.

Stories Density:



Nothing unexpected to report here.

Next, we ran a multiple linear regression predicting our Rent variable. This is simply meant to determine variable importances, and give our exploratory analysis some additional context.

Variables that were significant at all levels are:

*Size

*Employment Growth

*Leasing Rate

*Stories

*Renovated

*Class A1

*Class B1

*Net Contract Rent

*Cooling Days

*Heating Days

*Precipitation

*Gas Costs

*Electricity Costs

Variables that were significant at the 5% level are:

*Ameneties

*Age

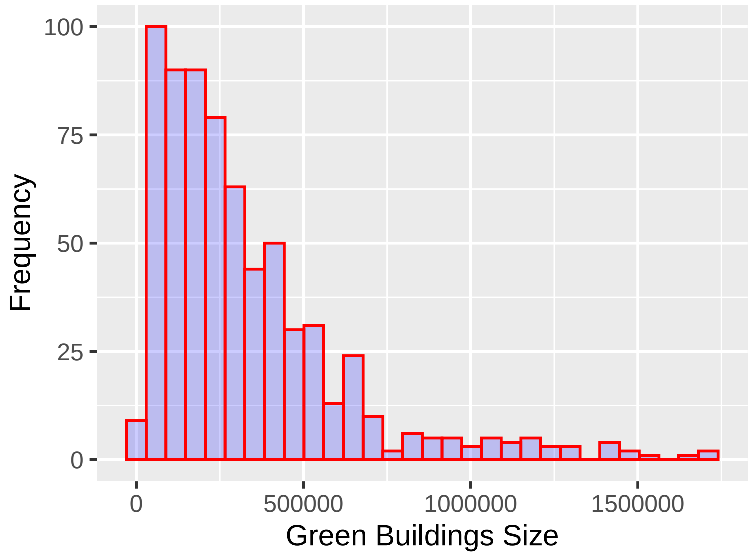
Surprisingly, LEED, Energystar, and Green Rating are all insignificant.

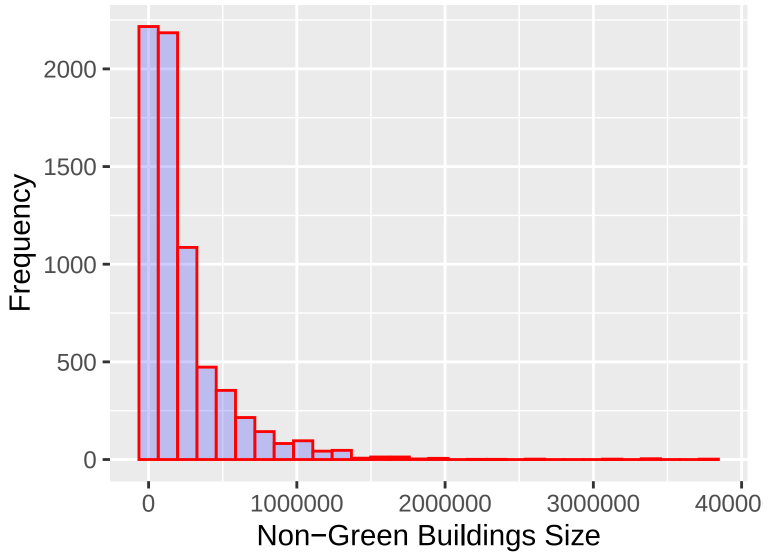
We assume that this insignificance is because there is multicolinearity between Green Rating and other important variables, like Leasing Rate, Age, Size, Renovated, Gas Costs and Electricity Costs.

Investigating this led us to the realization that, due to the clustered nature of the data, which is averaged by all buildings in the location, we cannot evaluate the Gas Costs, or Electricity Costs for Green versus Non-green buildings. However, Leasing Rate, Age, Size and Renovated are all not subject to clustering averages.

Next, we split the data into Green Buildings and Non-Green Buildings

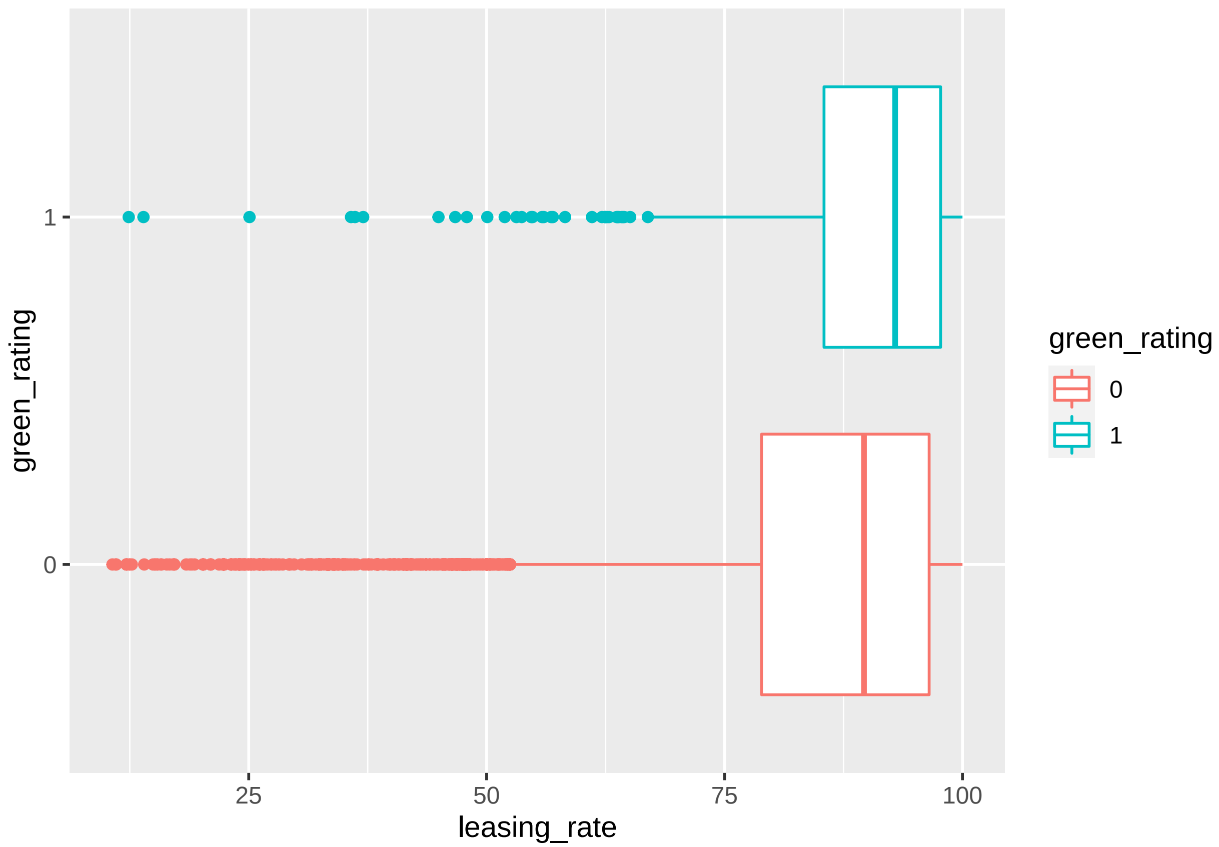
Now we compare the square footage of Green and Non-Green buildings





The building sizes are consistent.

We suspect that green buildings may have a higher average leasing rate.



Leasing rates in green buildings:

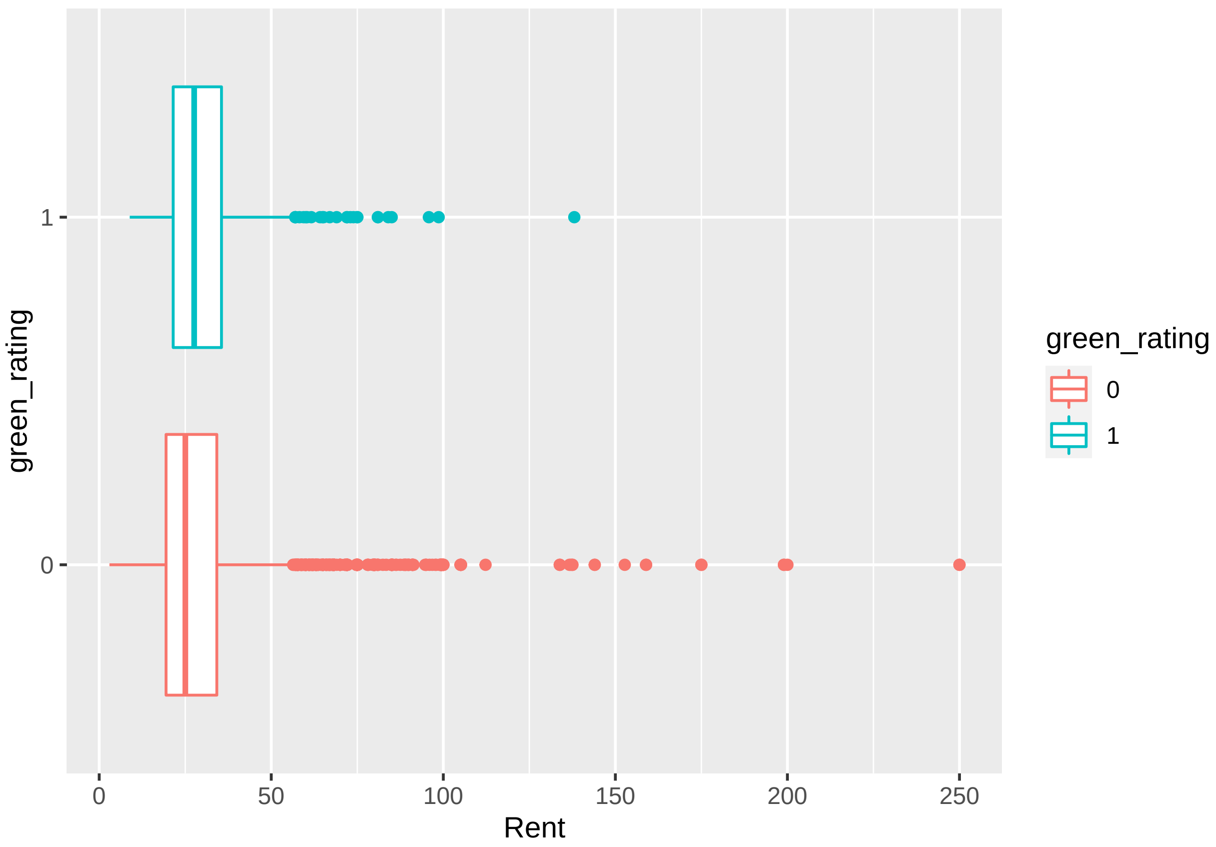
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.39   85.45   92.92   89.41   97.70  100.00
```

Leasing rates in nongreen buildings:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.68   78.89   89.65   84.44   96.50  100.00
```

Green buildings do have a higher leasing rate, by about 5%. We use the mean to evaluate this, because outliers are not a problem in evaluating leasing rate. This was overlooked by the stats guru. What an idiot.

We also suspect that green buildings have higher average rents than non-green buildings.



Rent in green buildings:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.50   27.60   30.03   35.54  138.07
```

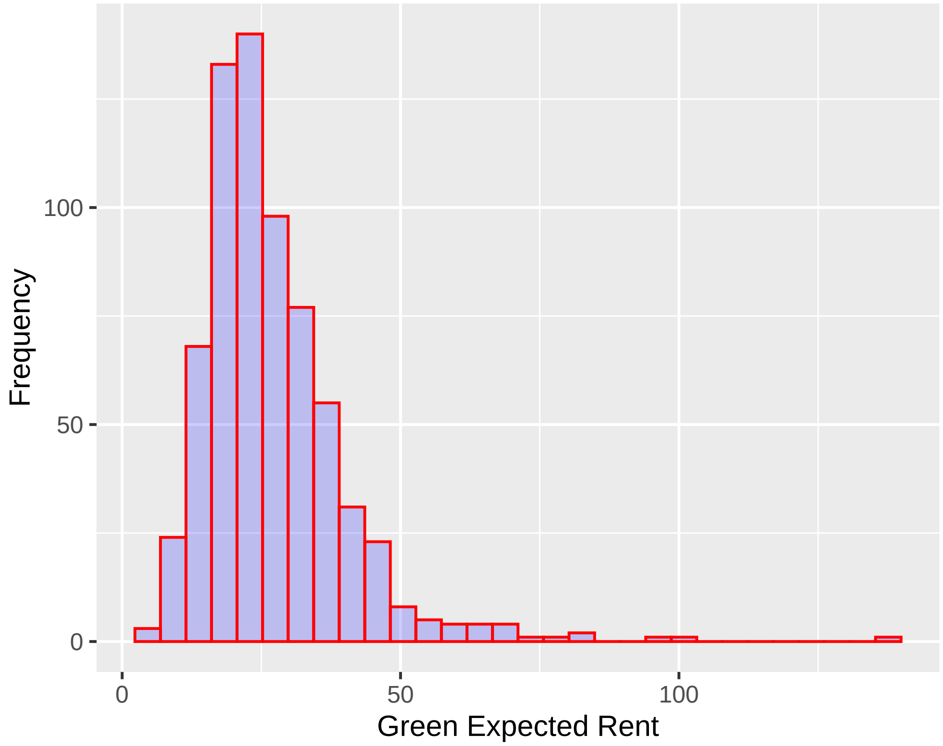
Rent in nongreen buildings:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.43   25.03   28.44   34.18  250.00
```

They do, by about $2.50, we use median here because rent is subject to outliers. This conclusion is similar to the stats guru.

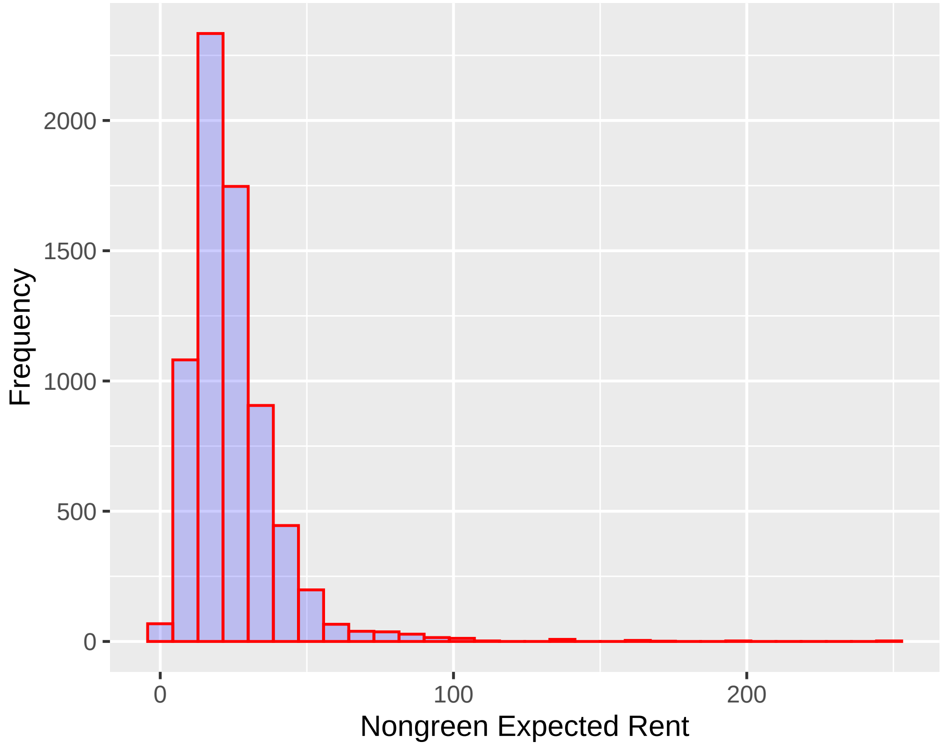
Next we create a variable called "Expected Rent," which multiplies leasing rate by rent, and determines the expected monthly rent per room.

Visualizing expected rent for green and Non-Green buildings



Expected rent in green buildings:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.015  18.777  24.434  27.019  32.676 137.034
```

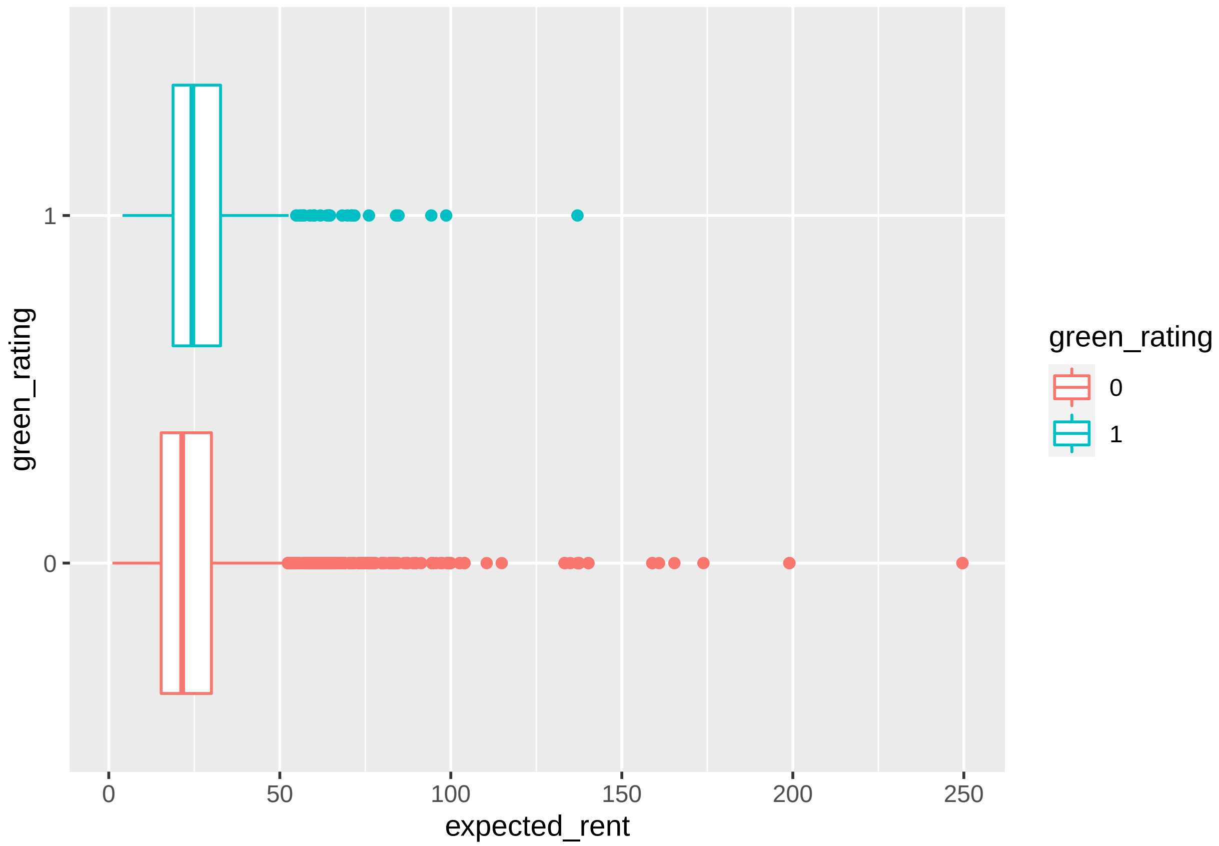


Expected rent in nongreen buildings:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    1.068   15.306   21.454   24.493   30.015  249.600
```

When comparing expected rent per room, green buildings win by about $3.

We visualize this relationship again below with a boxplot.

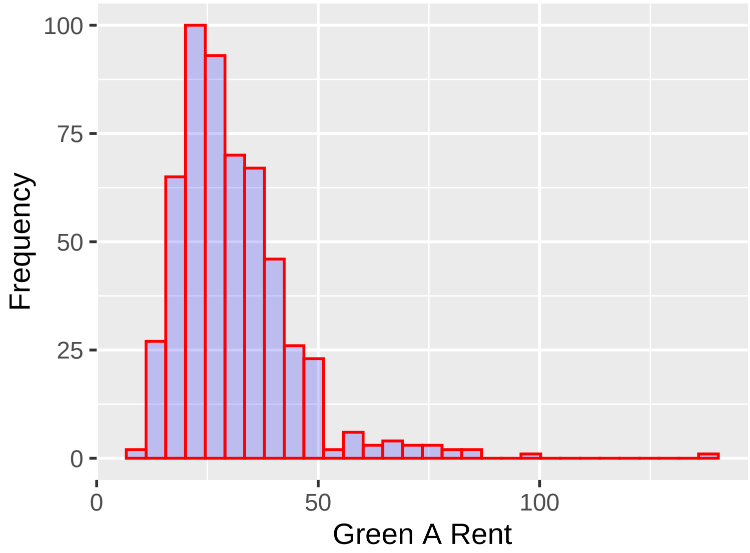


Unlike the stats guru, we are not comfortable estimating the time it would take to make the $5 million dollars back, because we do not know the room sizes, the exact location (where in Austin), the demand for green housing in Austin relative to other cities, or what the heating and cooling days may look like, among many other potential confounders. However, we can confidently say that there are higher rents **and** higher leasing rates in green buildings than in non-green buildings.

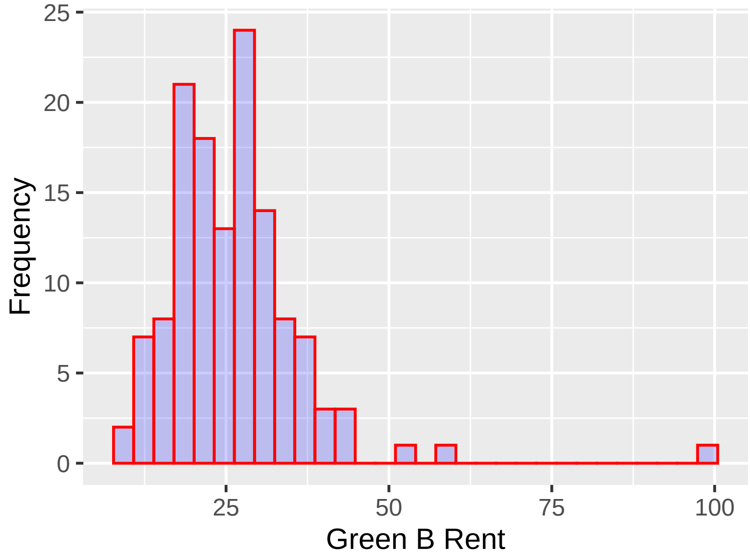
Finally, if she is going to build a green building, we want to determine wether class A or class B is more cost-efficient.

We separate our green buildings into green A and green B.

First, we look at the rent of green A versus green B buildings.



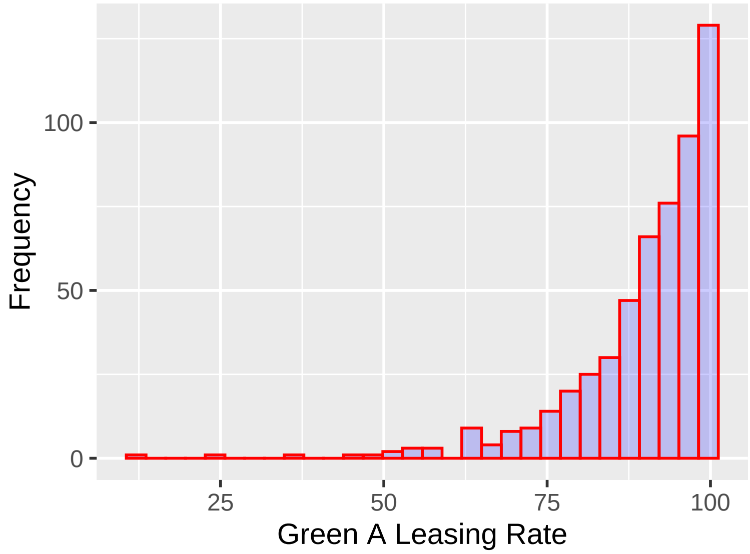
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   22.07   28.44   30.99   36.59  138.07
```



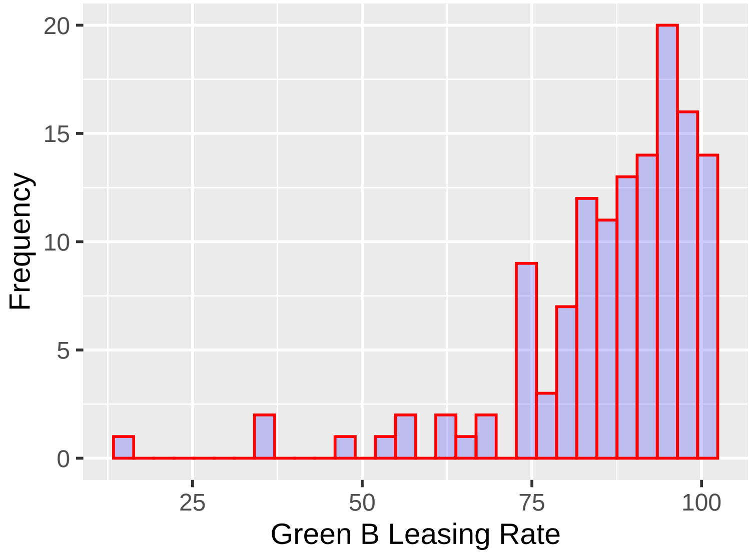
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   19.52   25.20   26.12   30.60   98.65
```

Green A has a higher Rent by about $4. This is expected.

Next, we look at the difference in leasing rate between green A and green B buildings.



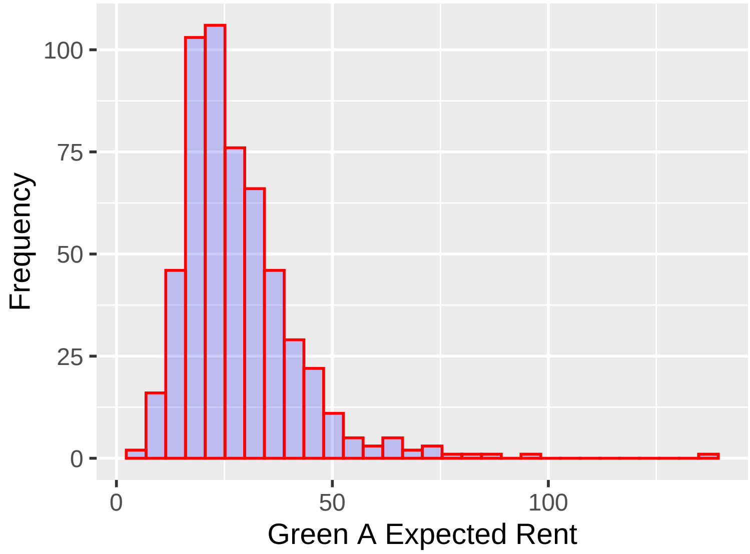
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.39   86.49   93.63   90.13   97.96  100.00
```



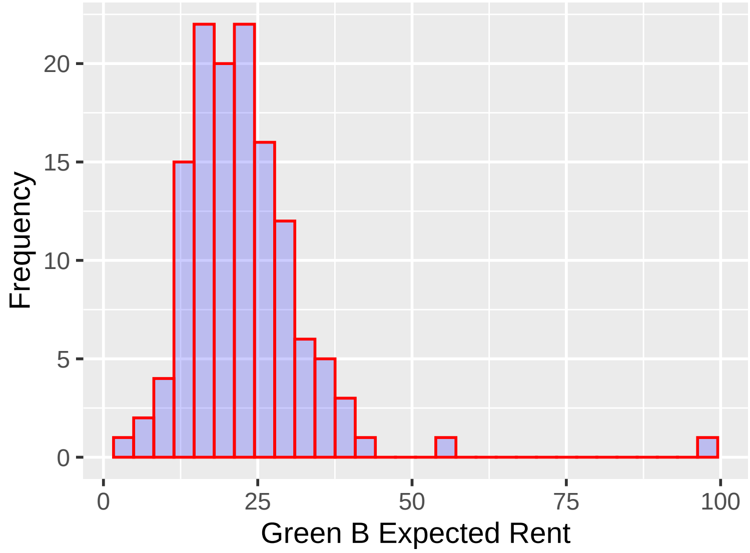
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.94   82.04   89.69   86.50   95.89  100.00
```

Green A is also about 4% more occupied. This was not expected, and leads us to conclude that she should build a nicer green building (type A).

Finally, we compare the expected rent of both building types.



```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.536  19.138  25.151  28.112  33.784 137.034
```



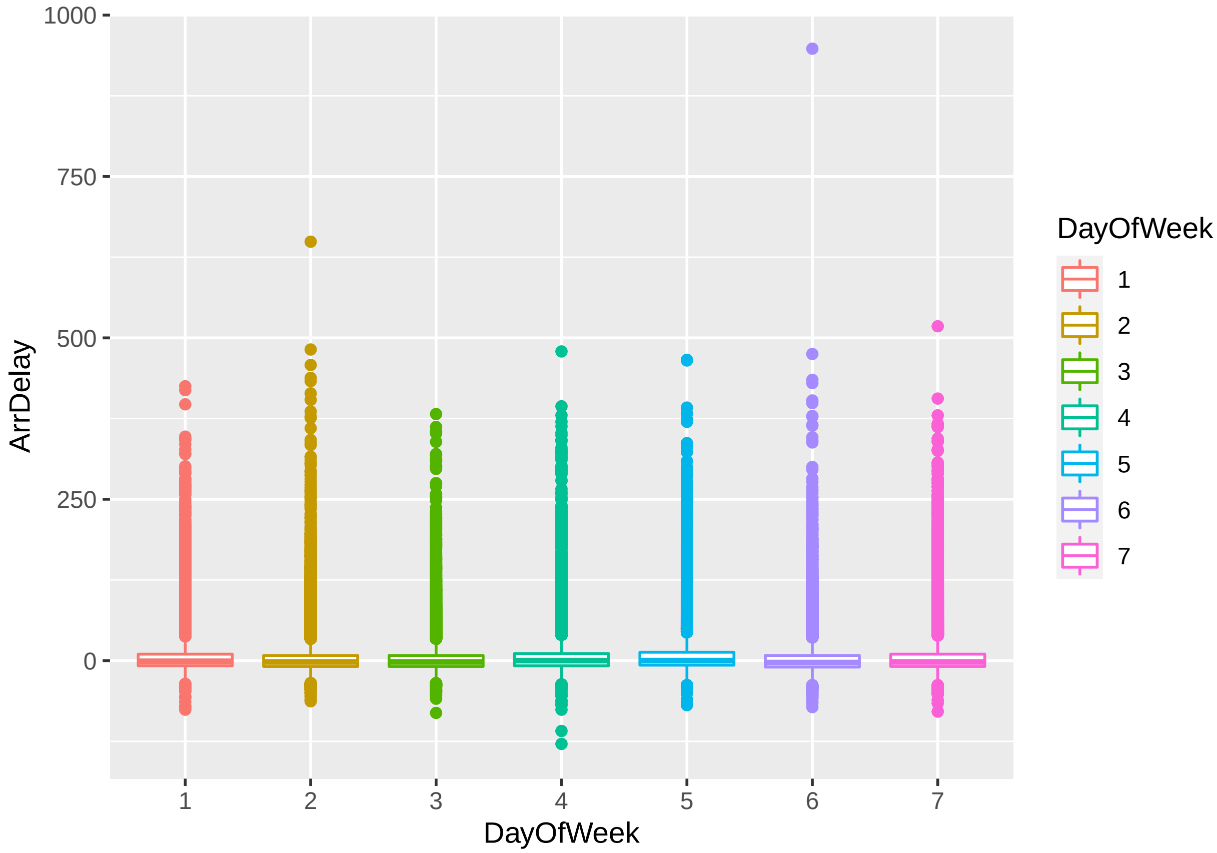
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.015  16.915  21.331  22.559  26.495  98.650
```

As anticipated, the expected rent is higher for green A buildings. However, looking at the leasing_rates is more informative here.

We conclude that a green building, on average, generates more money per-room, due to a mixture of higher rents and higher leasing rates. Additionally, green A buildings are 4% more occupied than green B buildings. More research regarding external factors is necessary in deciding to proceed with the project, but if the choice to proceed is made, she should build a green type A building.

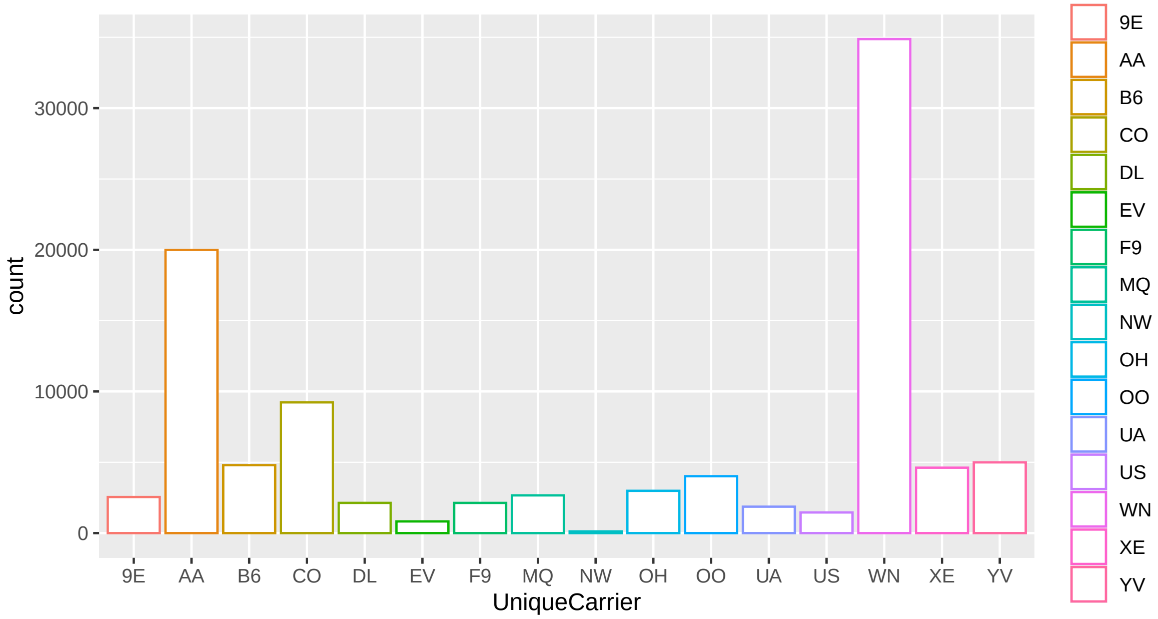
#Visual Story Telling 2: Flights at ABIA

In evaluating the flights into and out of Austin, we first looked at the delays grouped by the day of the week.



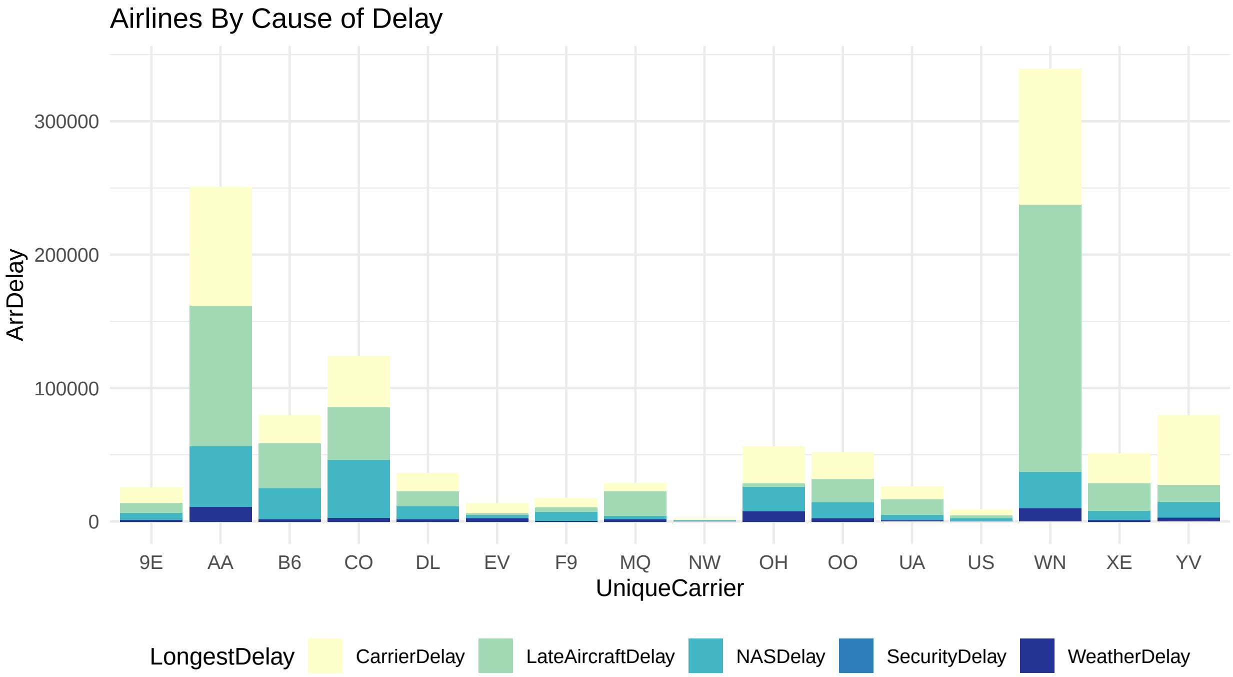
There is no obvious difference in arrival delays per day of the week.

Next, we decided to give this analysis a bit more context by evaluating the number of flights into and out of Austin by airline.



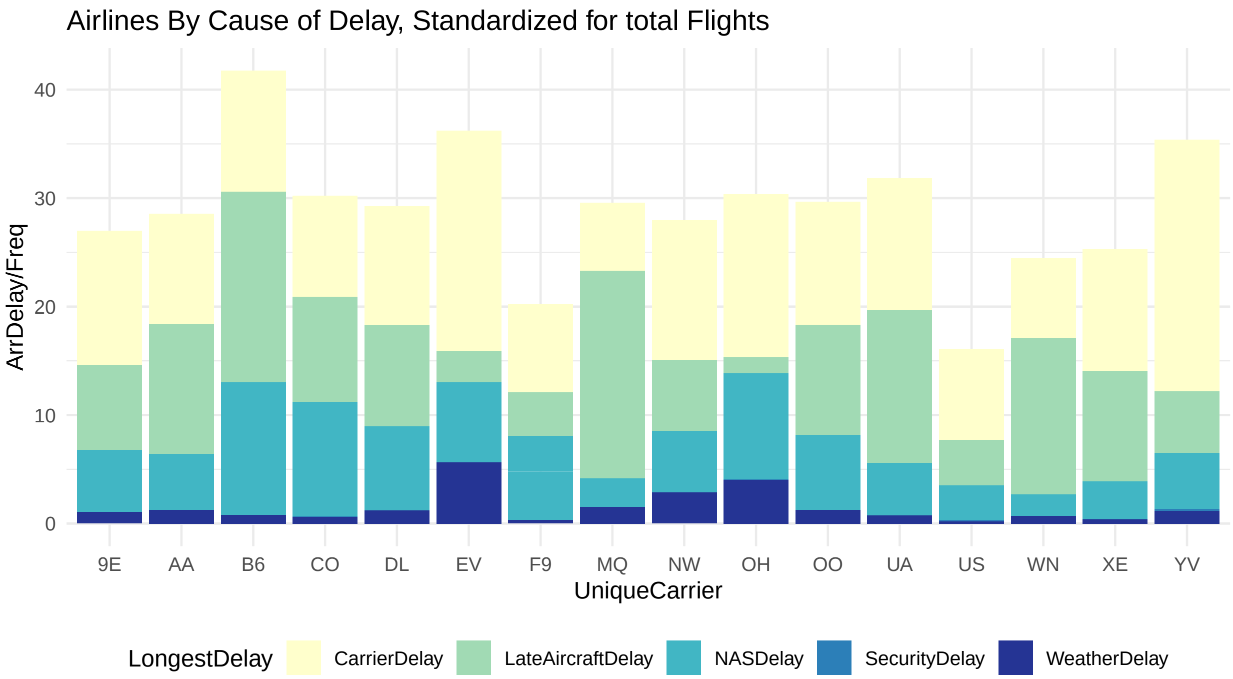
Southwest and American Airlines have the most flights in and out of Austin. More importantly, there is a good bit of variation in flight total by airline.

Next, we looked at the total delay time for each airline, shaded by the cause of the delay.



As we noted previously, Southwest and American Airlines make the most flights in and out of ABIA, and thus have the most delay. As we can see, the most frequent cause of delays is ‘Late Aircraft Delay’ indicating there may be some timeliness issues with the pilots and crews at ABIA.

Finally, we decide to standardize total delay time for number of flights. This allows us to visualize and evaluate airline efficiency without the confounding factor of flight count.



As we can see on this chart, JetBlue (IATA Code: B6) has the longest delays per flight, and Puerto Rico Airline (IATA Code: US) has the least delayed flights on average. ExpressJet (IATA Code: EV) had the most weather delays per flight of any airline followed by PSA (IATA Code: OH). To counter this, these two airlines had some of the lowest 'Late Aircraft Delay' minutes indicating that the airline is successful in employing efficient pilots and crews. Mesa Airlines (IATA Code: YV) has the longest average carrier delay indicating they may not be the best option for those on a tight schedule.

#Problem 3: Portfolio Modeling

In this Monte Carlo Simulation we were tasked with comparing 3 portfolios of ETFs and comparing their performance.

Our team selected to create **Portfolio 1** using 6 of the largest ETFs on the market, focused on Large Cap Growth Stocks. These ETFs typically hold the large tech companies like Apple, Microsoft, and Alphabet, as well as other Blue Chip firms like Berkshire Hathaway, Johnson and Johnson, and JP Morgan Chase. The often track indexes like the S&P 500 or DJI to track the market, however focus on long term growth for their holders.

Portfolio 1 = Large Cap Growth:

SPY - S&P 500 ETF Trust

VTI - Vanguard Total Stock Market ETF

VOO - Vanguard S&P 500 ETF

QQQ - Invesco QQQ Trust

DIA - Dow Jones Indus Average ETF

SCHX - Schwab US Large Cap ETF

Portfolio 2 is focused on Energy ETFs. These funds invest in energy companies, research groups, and commodities to achieve returns on the fluctuating energy market. The last year in particular has been incredibly volatile for energy companies, and our team figured it would be interesting to see the difference between a stable portfolio like Portfolio 1 and a volatile one like Portfolio 2.

Portfolio 2 = Energy (Oil Gas):

VDE - Vanguard Energy

XES - S&P oil and gas

PSCE - Small Cap Energy Fund

PXE - Oil and gas exploration

Portfolio 3 contains 5 Emerging Market Funds, intentionally not focused on a specific region or state. While a vast majority of the companies held in Portfolio 1 are American, our team wanted to investigate how successful nations deemed "emerging markets" have been, particularly with data that includes the Covid-19 market crashes. Some of the countries represented in this portfolio are: China, South Africa, South Korea, Russia, and India.

Portfolio 3 = Emerging Markets:

VWO - Vanguard Emerging Markets

SCHE - Schwab Emerging Markets

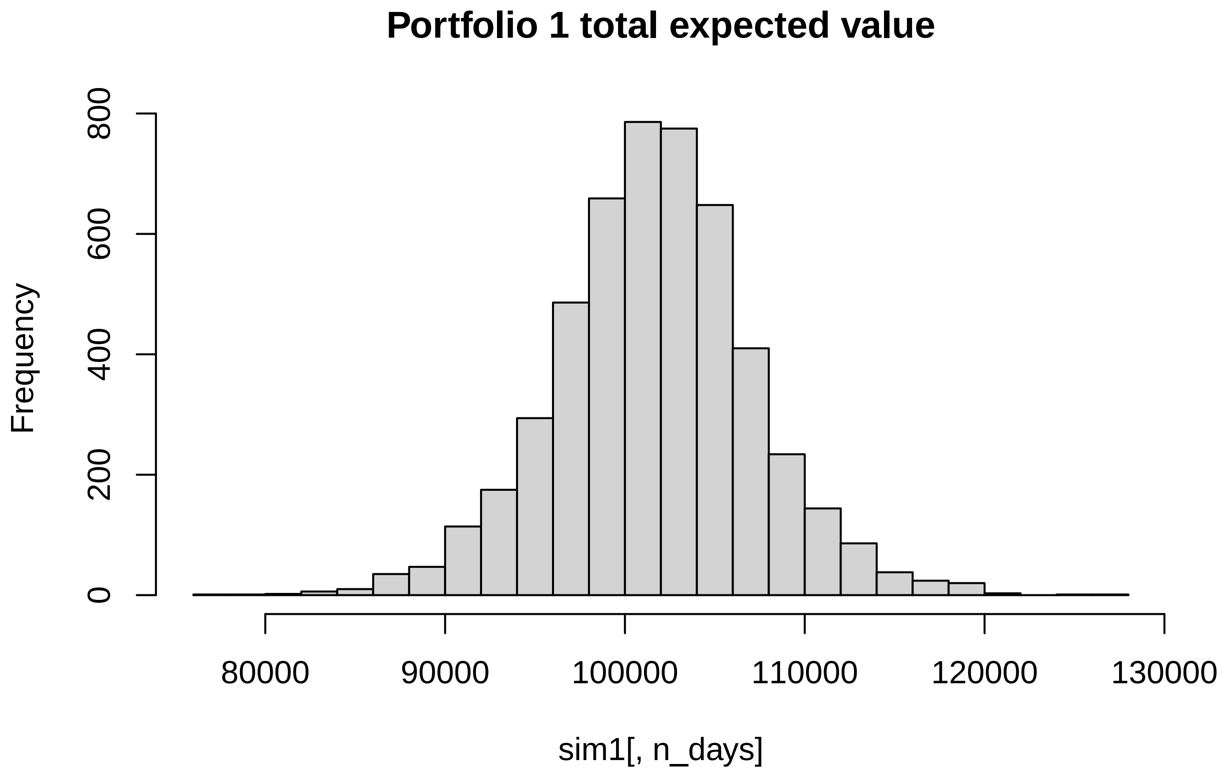
GEM - Goldman Sachs Emerging Markets

JEMA - JP Morgan Emerging Markets

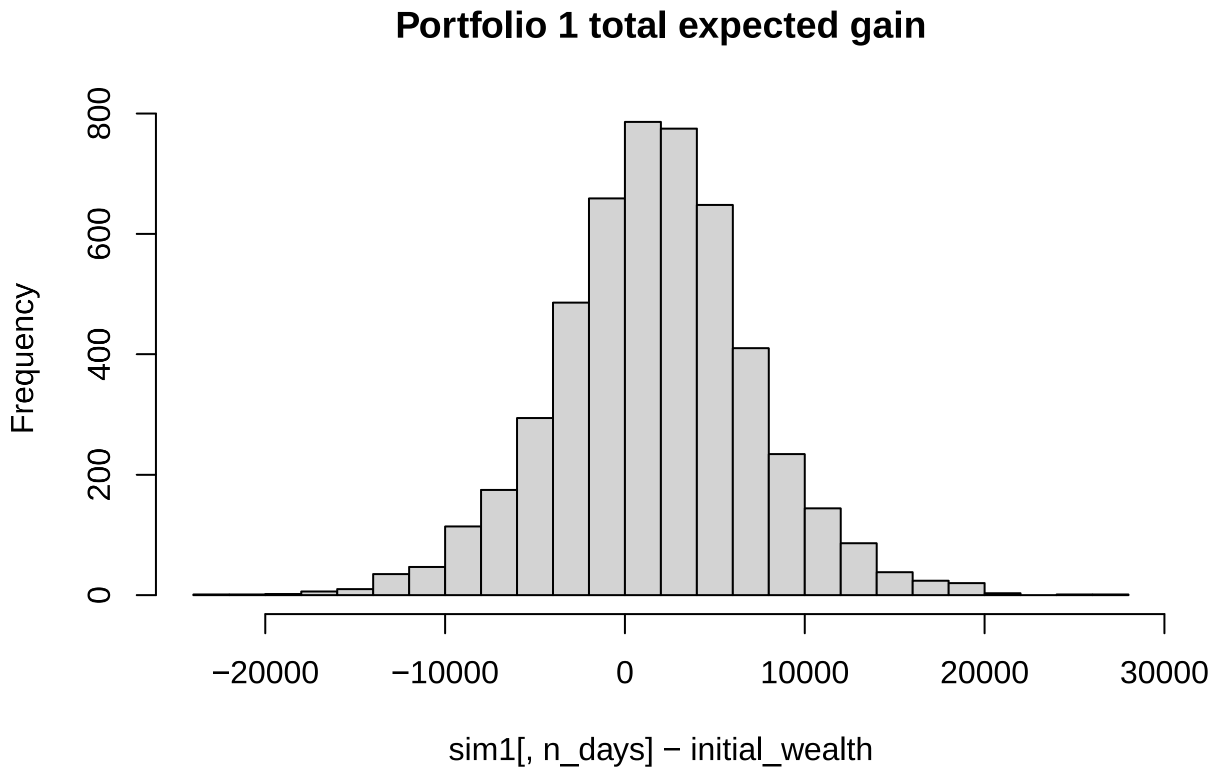
SPEM - S&P Emerging Funds

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

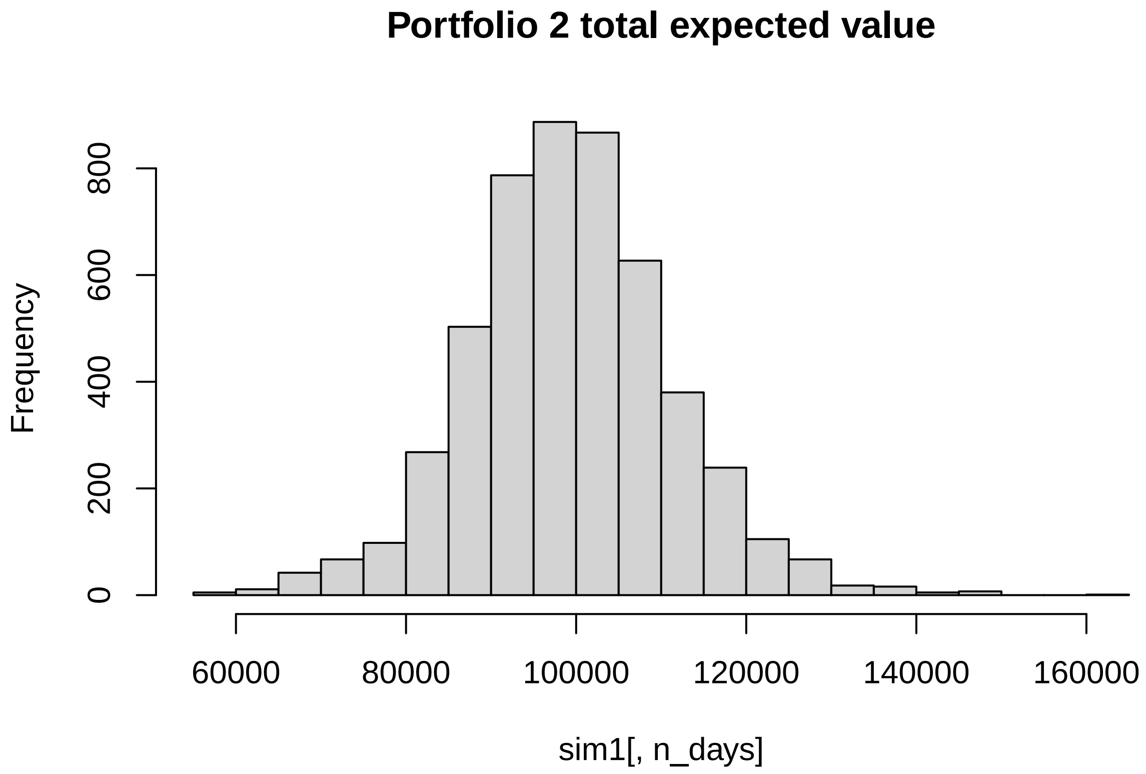
```
## pausing 1 second between requests for more than 5 symbols
## pausing 1 second between requests for more than 5 symbols
```



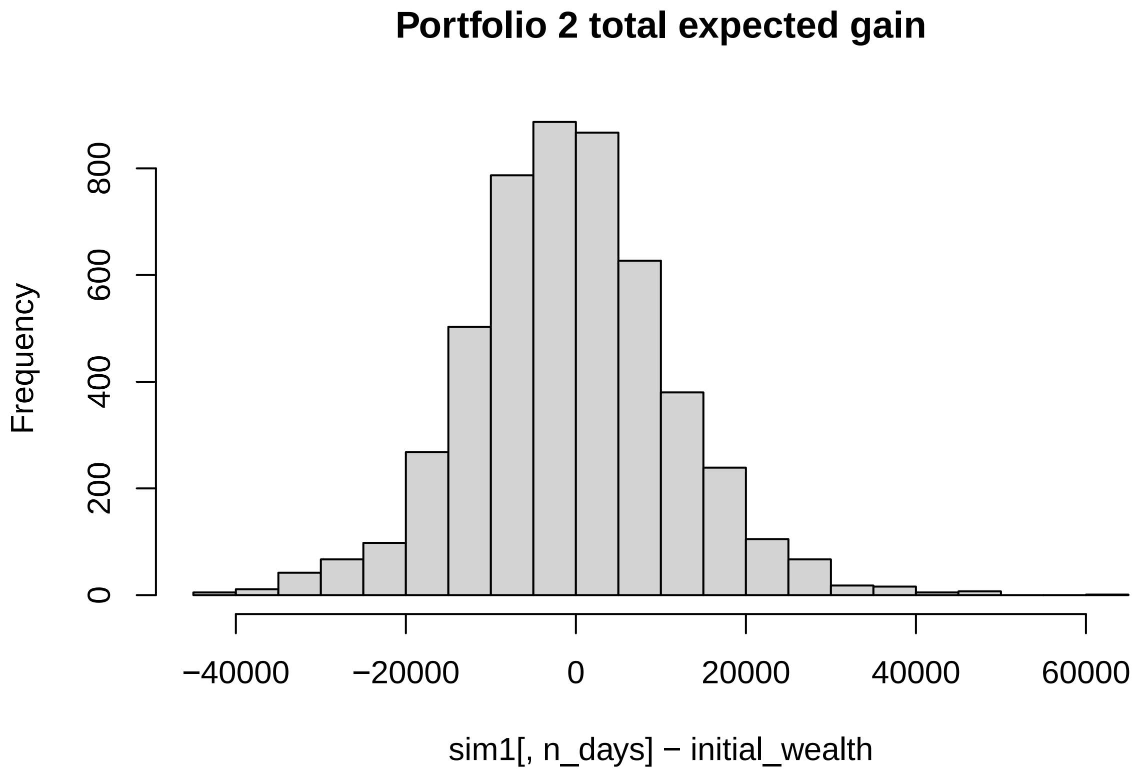
```
## [1] "The expected value is  101672.055469596"
```



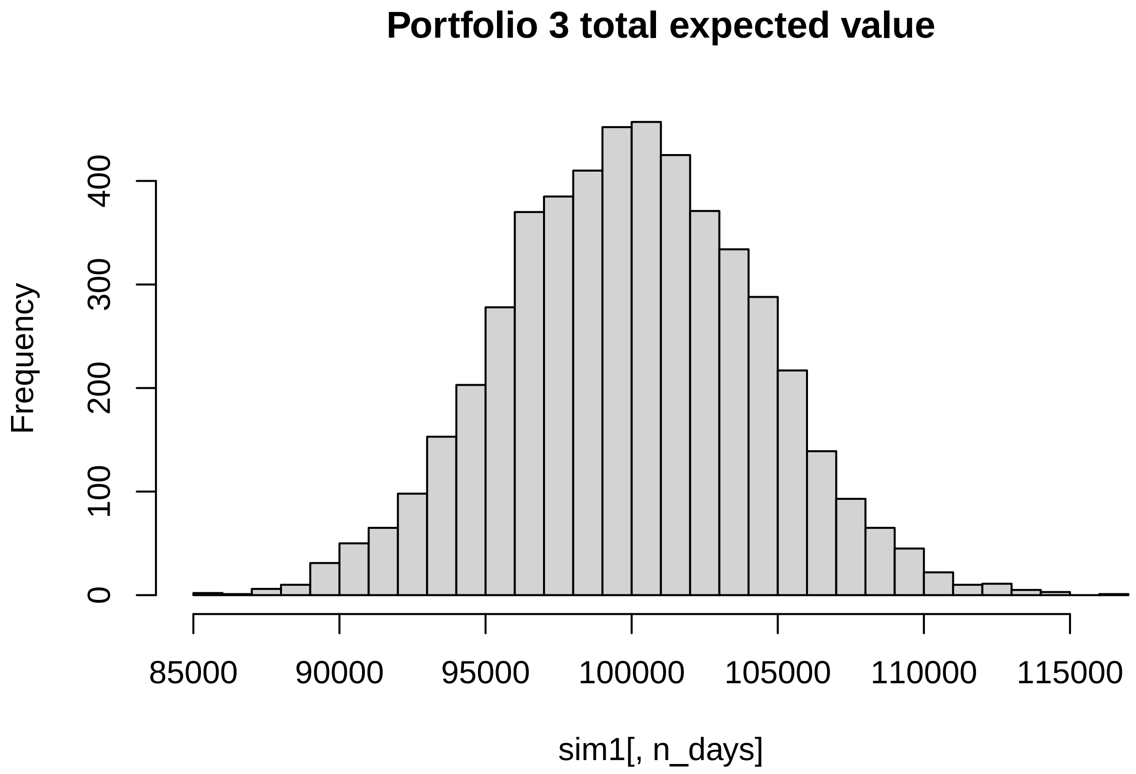
```
## [1] "The expected gain is:  1672.05546959634"
## [1] "The 5% VAR is: $7496.45413048029"
```



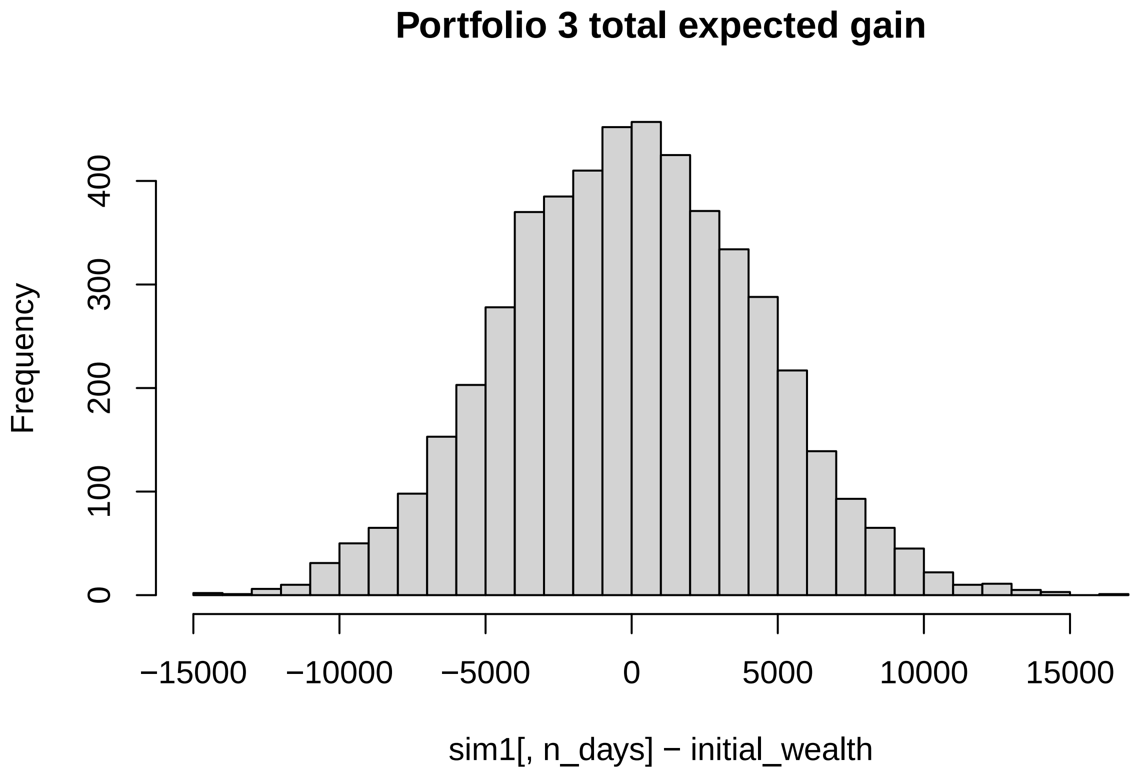
```
## [1] "The expected value is  99338.6524675106"
```



```
## [1] "The expected gain is:  -661.347532489363"
## [1] "The 5% VAR is: $19339.9916072046"
```



```
## [1] "The expected value is  99974.3780022183"
```



```
## [1] "The expected gain is:  -25.6219977817231"
## [1] "The 5% VAR is: $7153.21260026744"
```

Within the last 5 years, the stock market has faired relatively well. For the first 3.5 years, the market was considered one of the strongest ever and investors were astounded at their returns. Since March of 2019, the Covid-19 pandemic hampered the market for months and caused a crash. Since this period the market has recovered, however this variability in total market performance may yield interesting returns in this simulation.

For Portfolio 1 (Large Cap Growth funds), over a 5000 day simulation, we were able to earn $1577.157. This is approximately a 1.5% return which is modest given the time period. At the 5% risk level, this portfolio's VaR was $7068.17, indicating that in the worst 5% of cases, this portfolio lost around 7% of it's initial value. In the Histograms for Portfolio 1, we can see a data point that achieved over a $30,000 gain which is pretty spectacular for any ETF over 5000 simulations. Overall, Large Cap 'Growth' funds do not appear to have grown that consistently in the last 5 years.

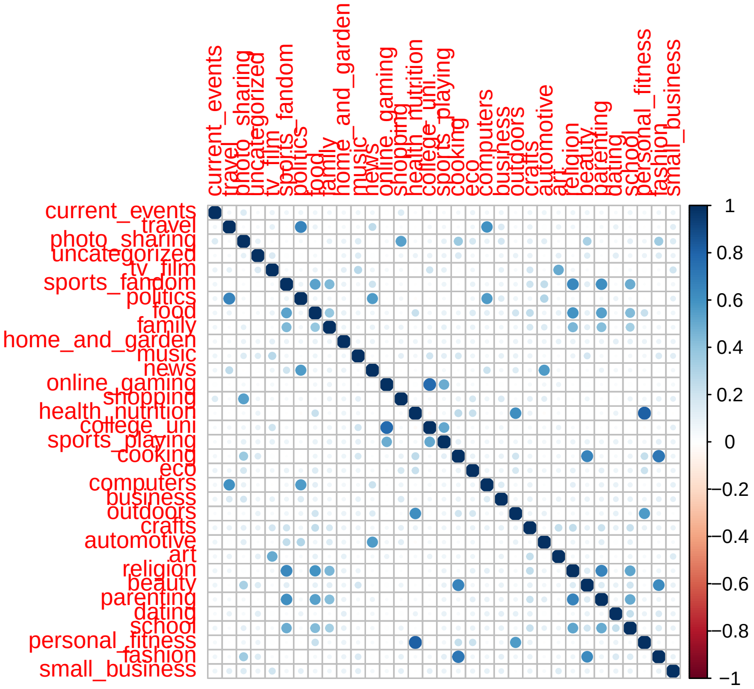
For Portfolio 2 (Energy Based Funds), the model actually predicted a $471 loss. The range of performances by Portfolio 2 is far greater than in Portfolio 1. This is due to the energy market being more volatile and feeling the impacts of market fluctuations more than stable Large Cap Stocks. At the 5% VaR level this portfolio is expected to lose $19427 which is almost 20% of the initial starting value. Some portfolios were able to gain and lose about 60% of the initial value which is incredible.

For Portfolio 3 (Foreign Market Funds), the simulation predicted essentially breaking even ($26 loss). These funds appear to be more stable than the Energy funds as the VaR is only $6900 and no porfolio made or lost more than $15000.

#Problem 4: Market Segmentation

In evaluating market segmentation for NutrientH20 through the tweets in our dataset, we first centered and scaled the data.

Next, we ran a correlation matrix on the dataset to detect patterns, both predicted and unpredicted, among the included variables.



Below are the Top-3 Highest correlated variables:

1.) It looks like **"personal_fitness" and "health_nutrition"** are very highly correlated

2.) It looks like **"college_univ" and "online_gaming"** are very highly correlated

3.) It looks like **"fashion" and "cooking"** are very highly correlated

Some other highly correlated variables include:

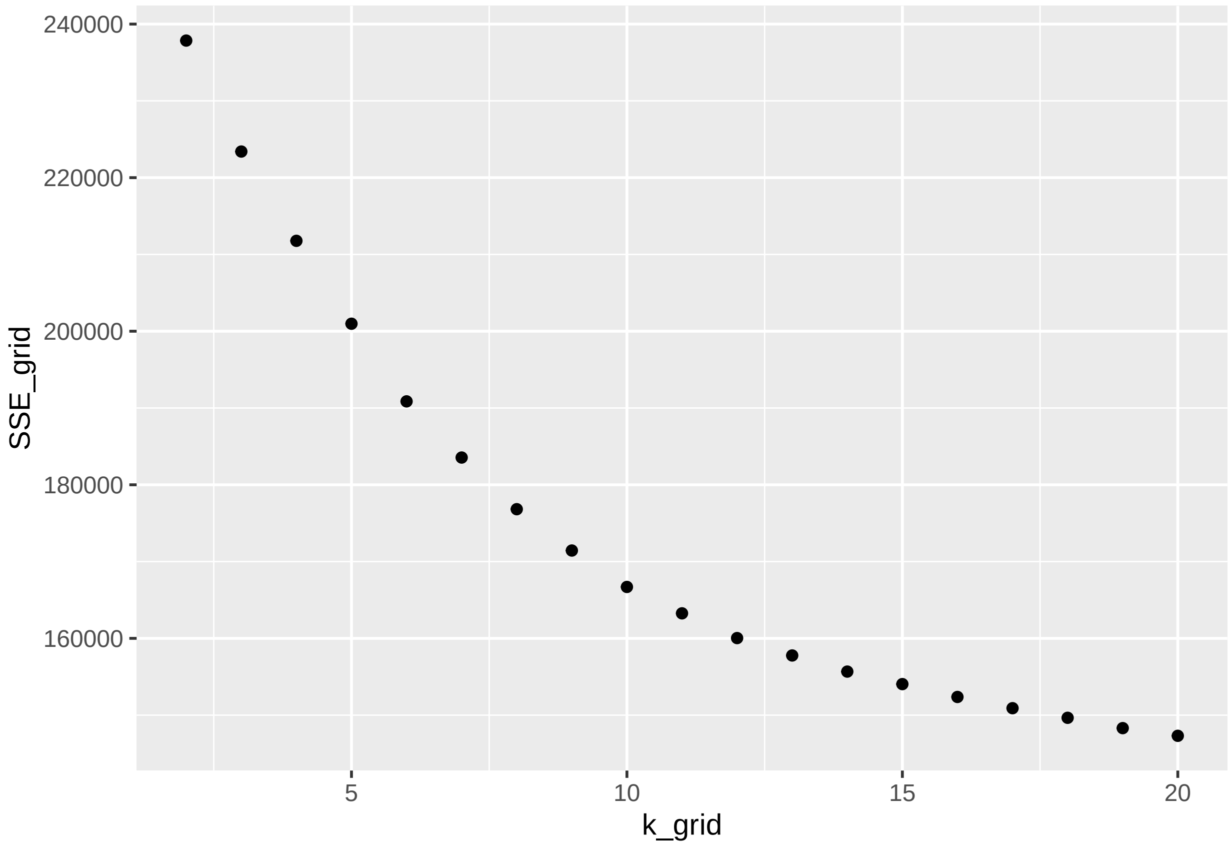
('religion' and 'sports_fandom'), ('politics' and 'travel'), and ('cooking' and 'beauty') among others

Next, we ran an elbow plot to determine the most efficient cluster value of K.

```
## Warning: did not converge in 15 iterations
```

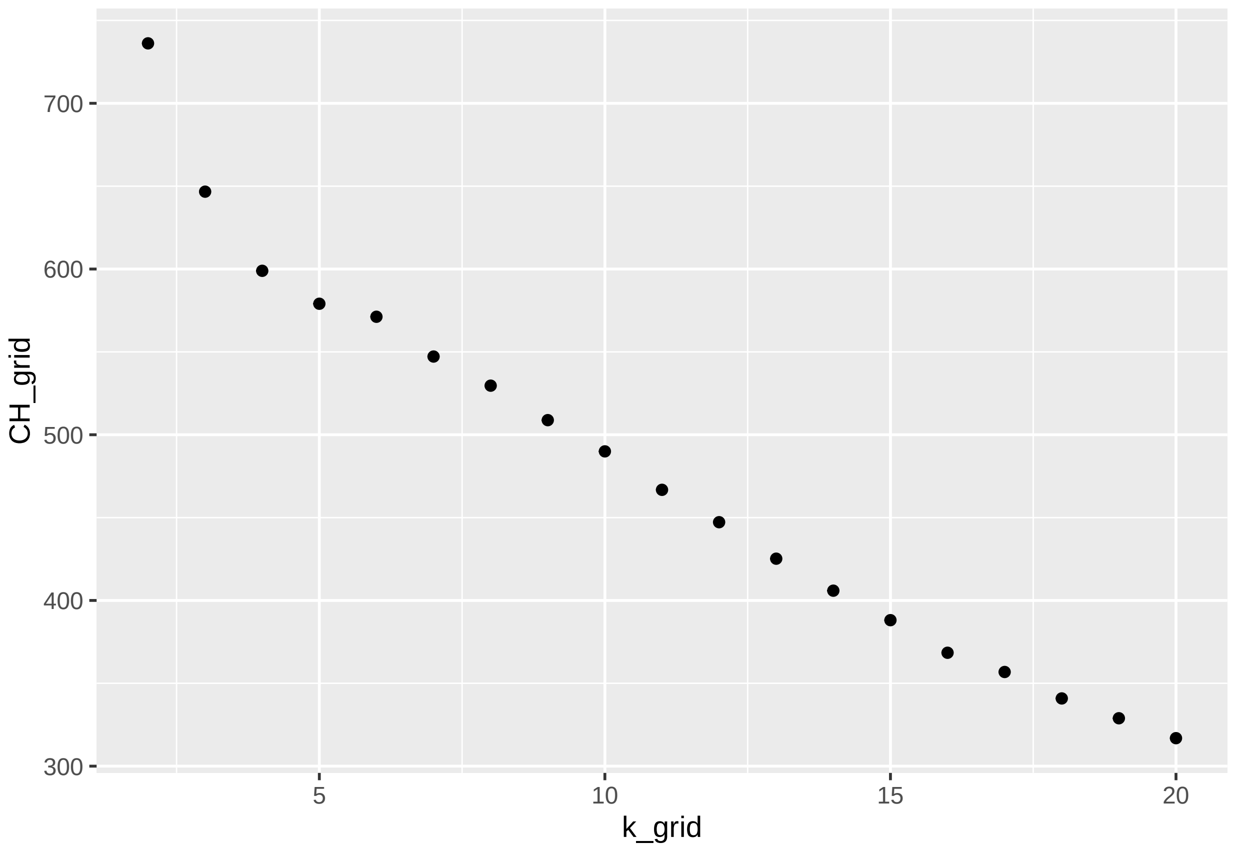
```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 394100)
```

```
##  [1]  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  [1] 237851.8 223400.8 211773.3 200976.6 190863.0 183549.6 176821.1 171437.9
##  [9] 166694.9 163259.0 160035.4 157771.2 155671.4 154043.7 152368.0 150902.9
## [17] 149653.5 148308.4 147308.7
```



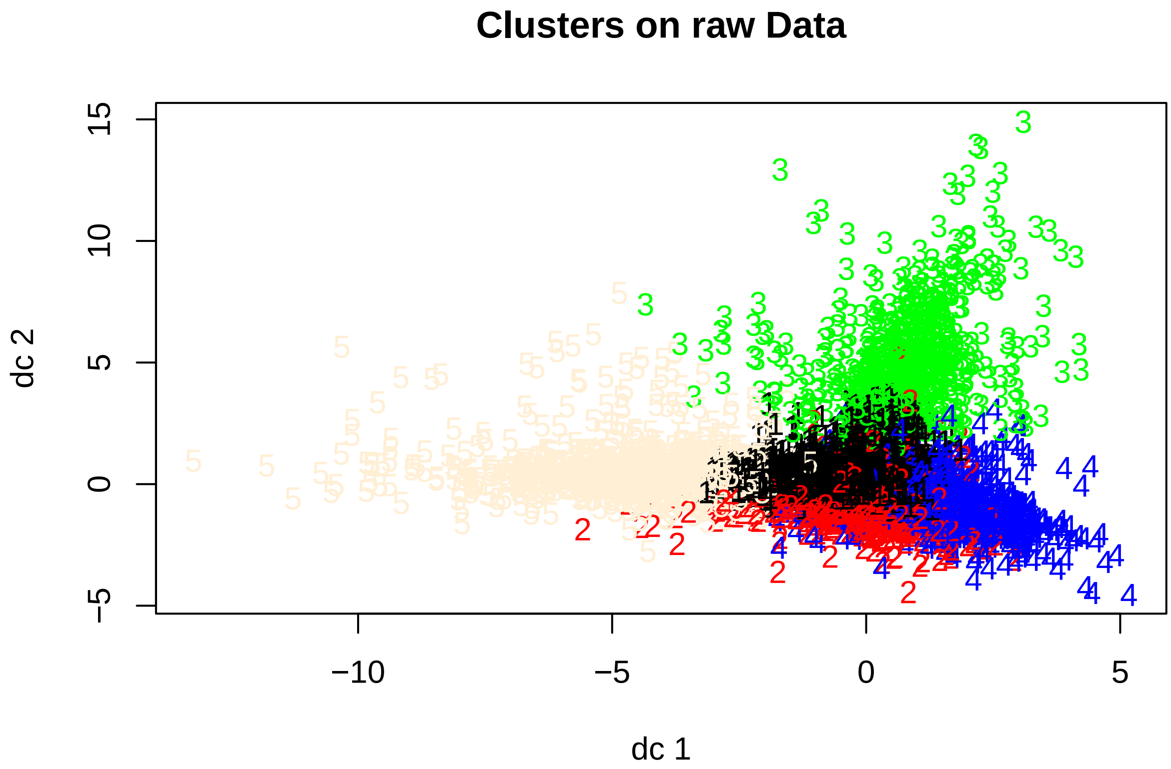
The elbow plot appears to suggest 7 to 9 clusters.

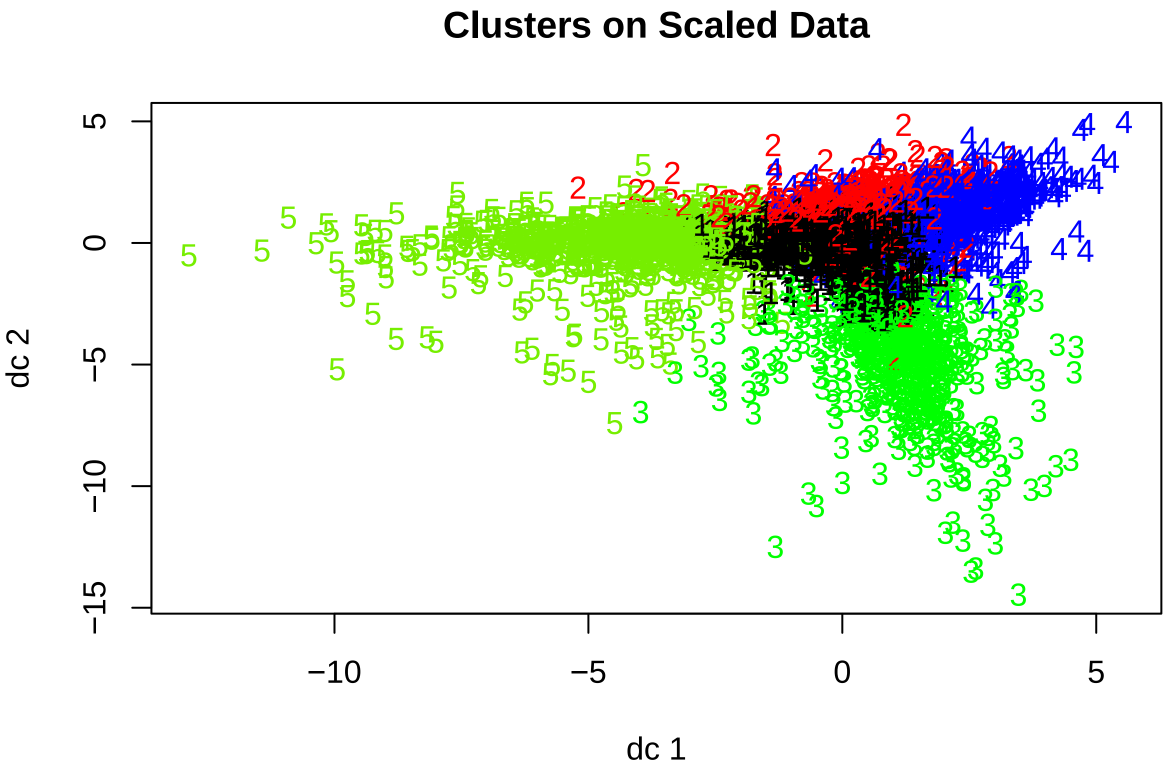
After the elbow plot, we ran a CH index, which has the same purpose as the elbow plot, but provides more context reagrding optimal K.

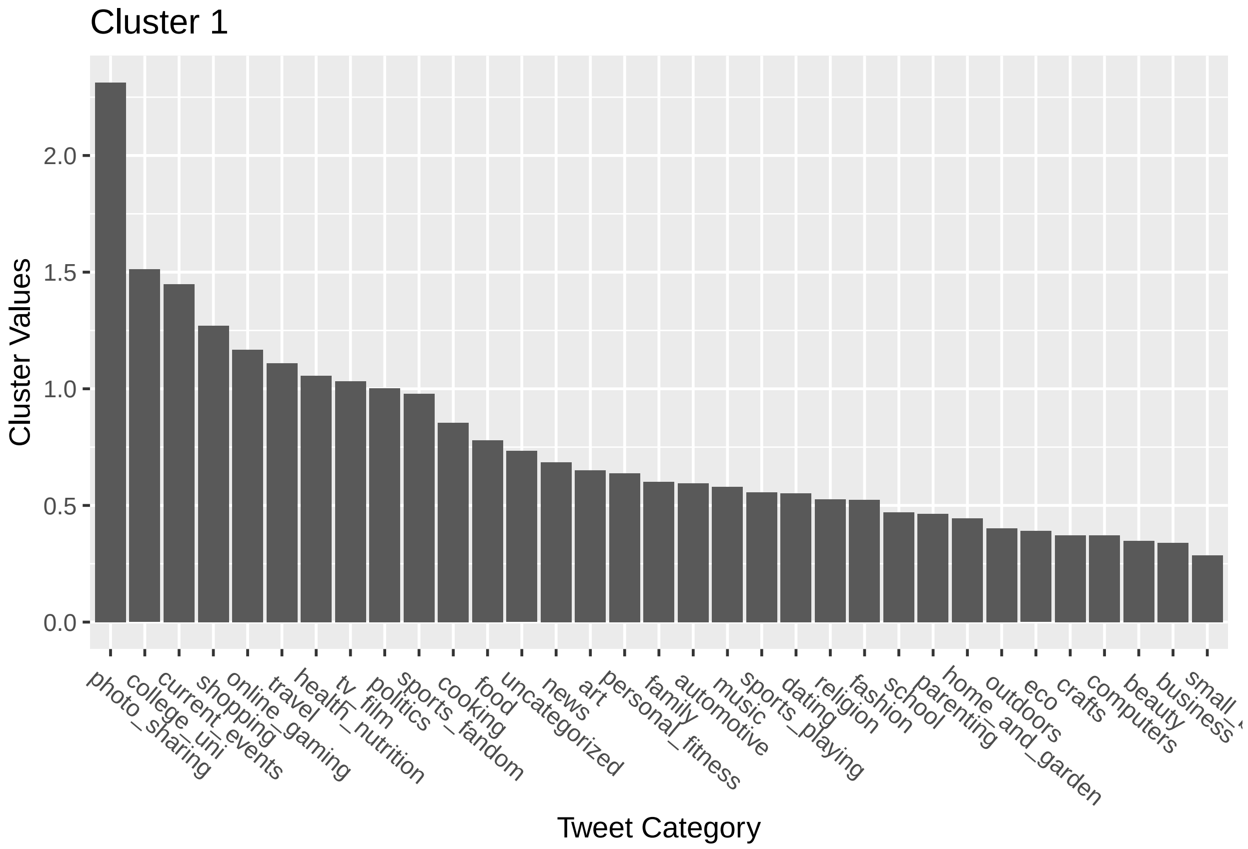


The CH index suggests 5 clusters.

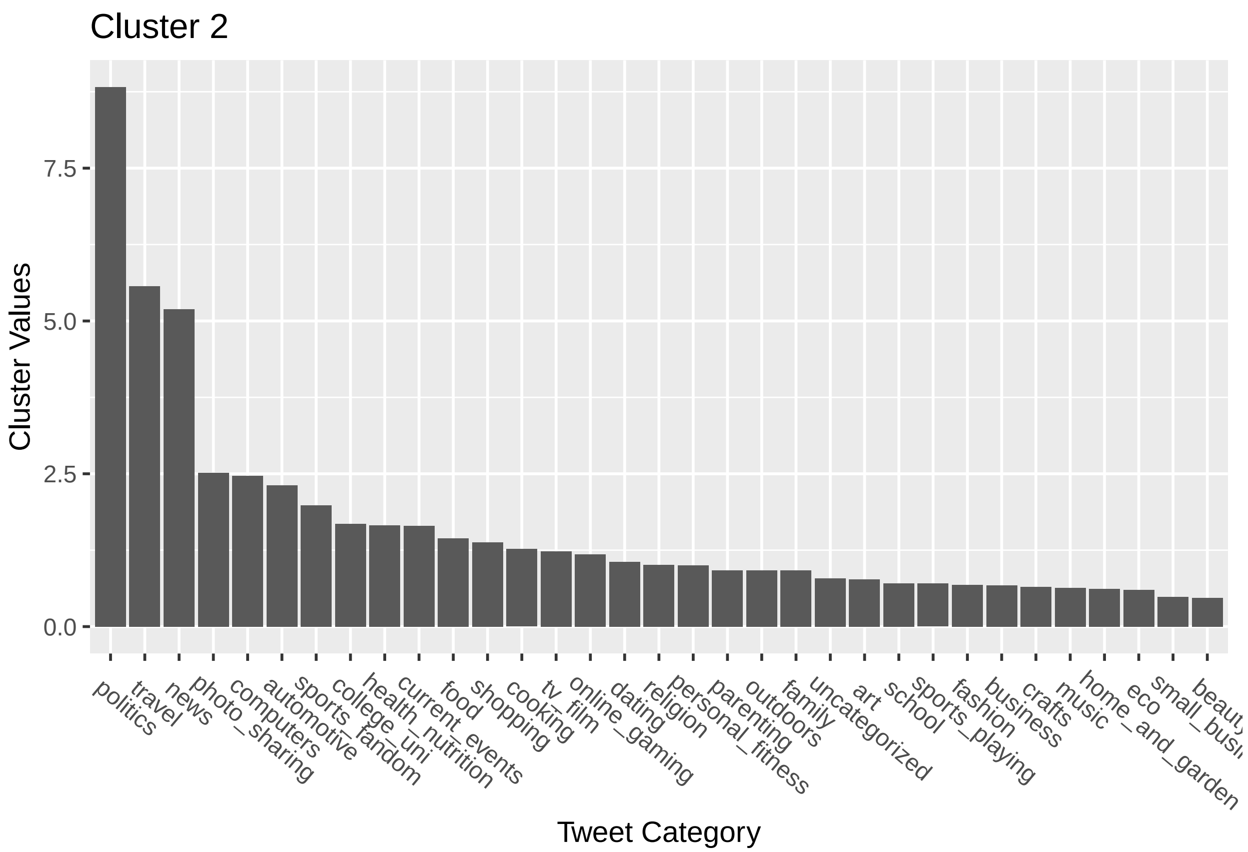
With our elbow plot and CH suggestions in mind, we ran a K-means++ with 5 clusters.


Clusters on raw Data
dc 2
dc 1
15
10
5
0
−5
−10
−5
0
5
1
2
3
4
5

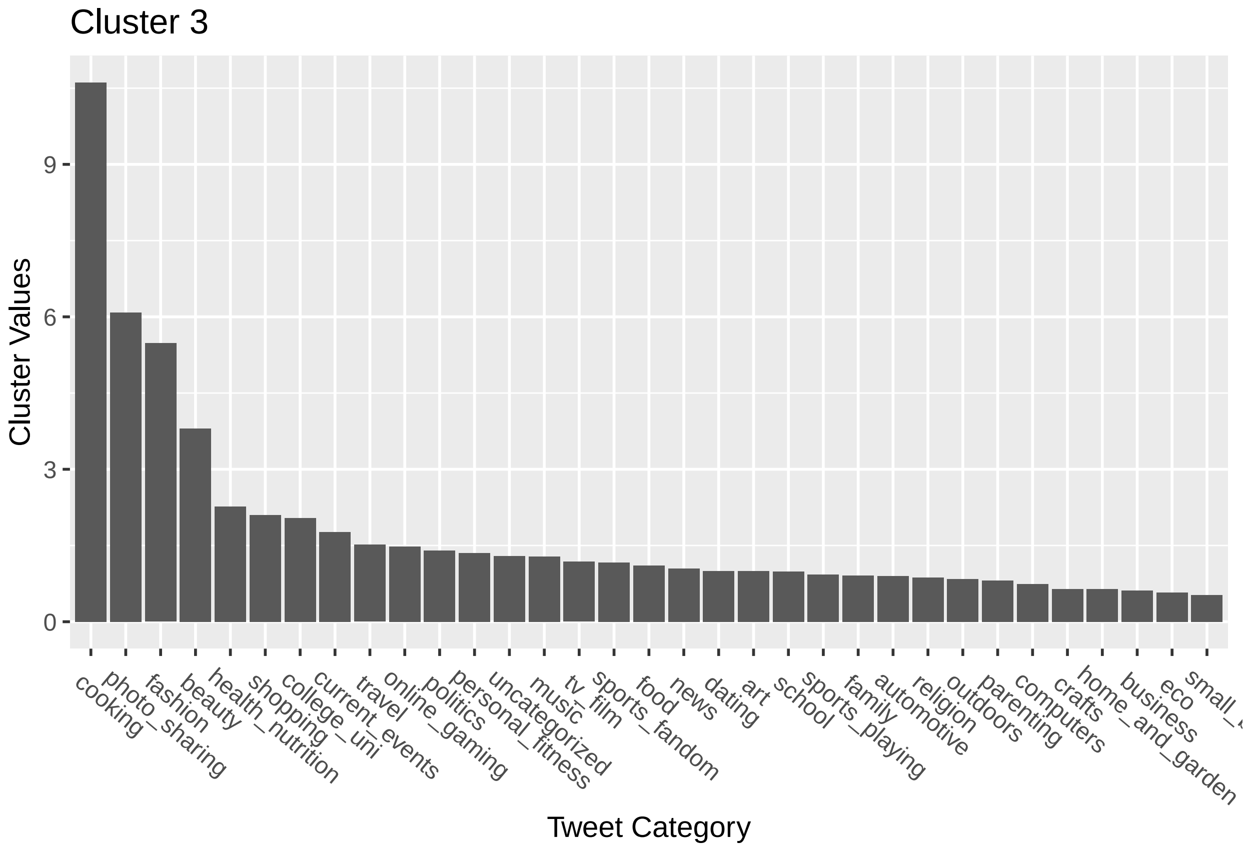

Clusters on Scaled Data
dc 2
5
0
−5
−10
−15
−10
−5
0
5
dc 1
1
2
3
4
5



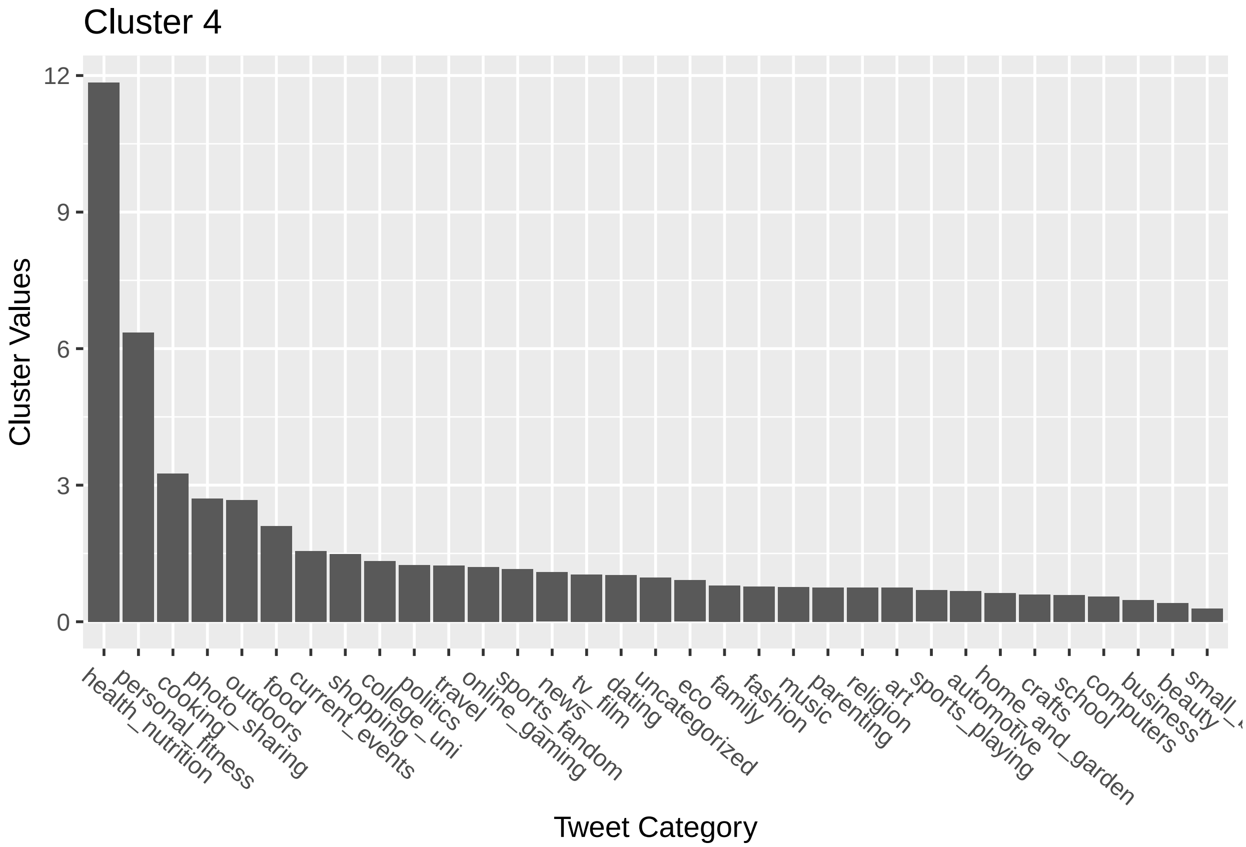
Cluster 1 —> Top 3 Categories are “photo_sharing”, “college_uni”, and “current_events”



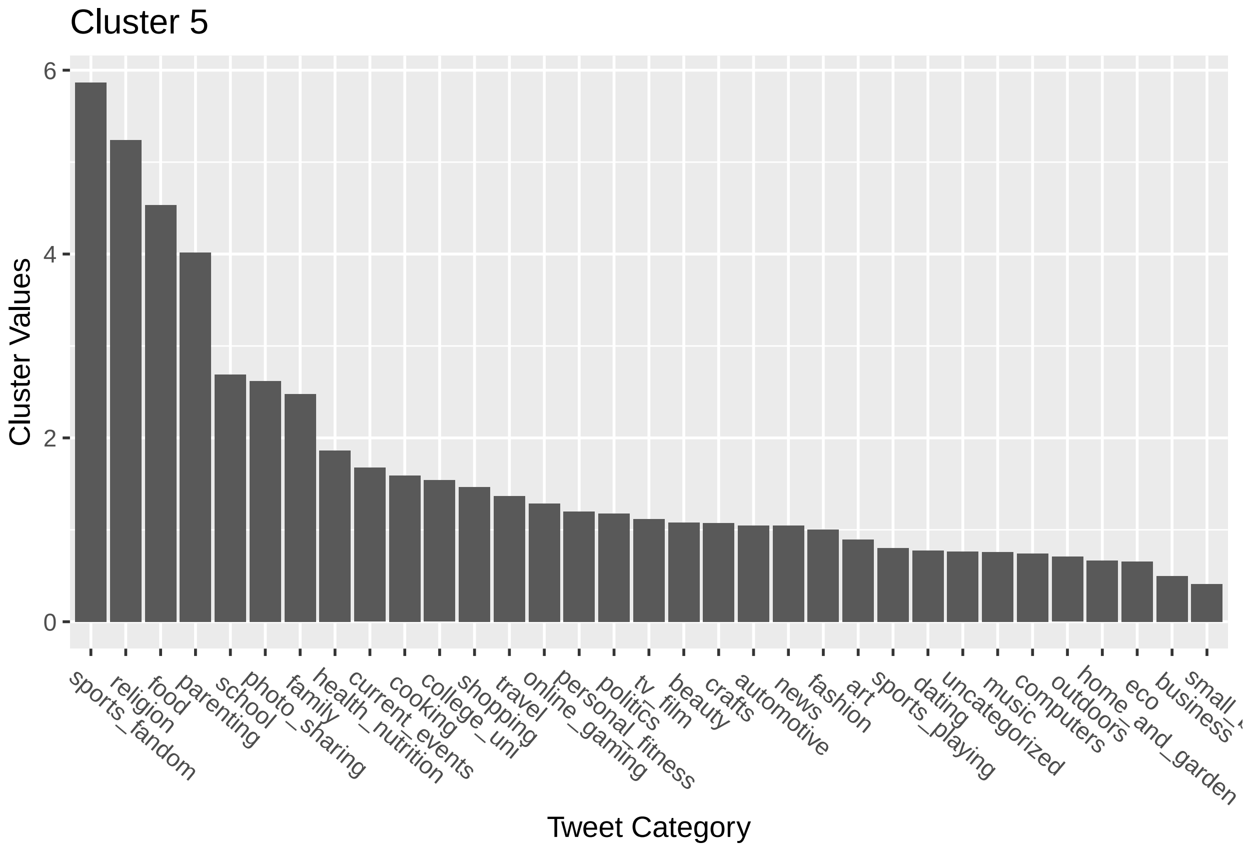
Cluster 2 —> Top 3 Categories are “politics”, “travel”, and “news”



Cluster 3 —> Top 3 Categories are “cooking”, “photo_sharing”, “fashion”

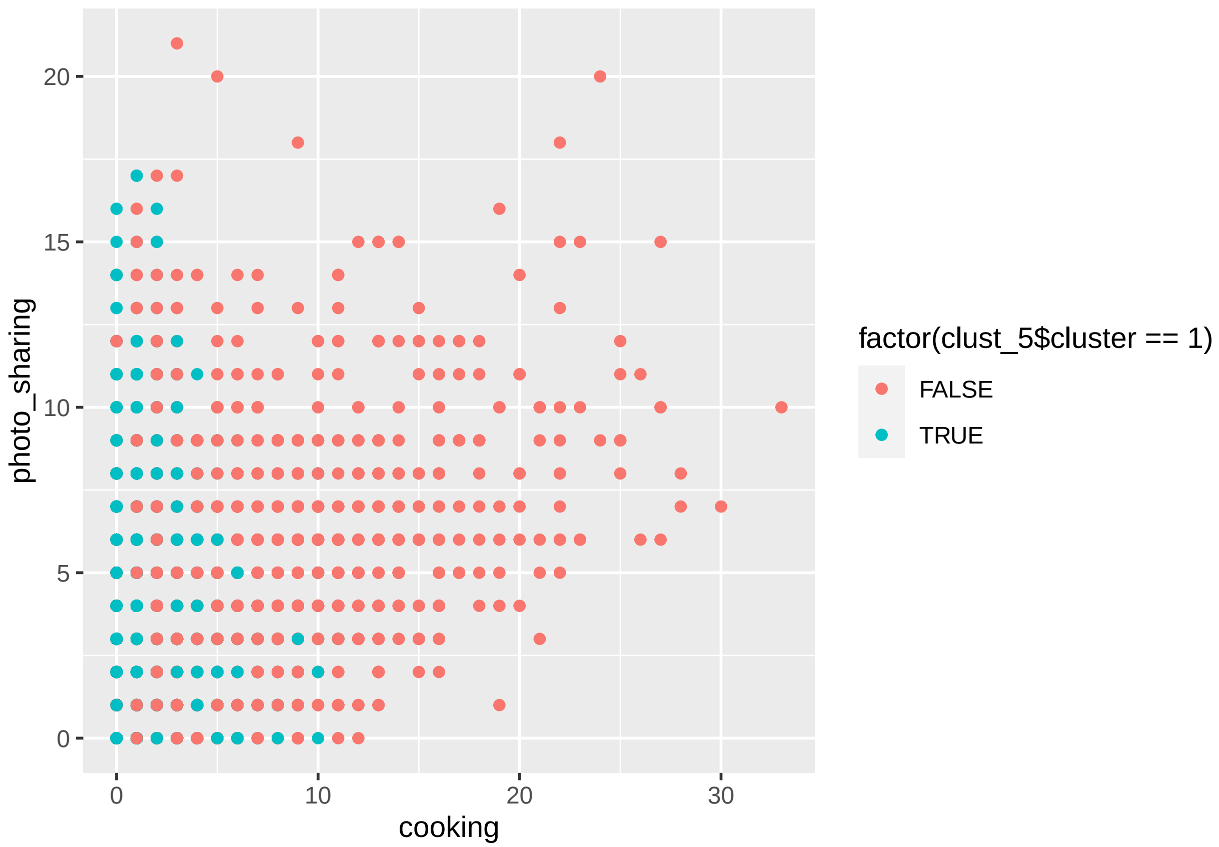


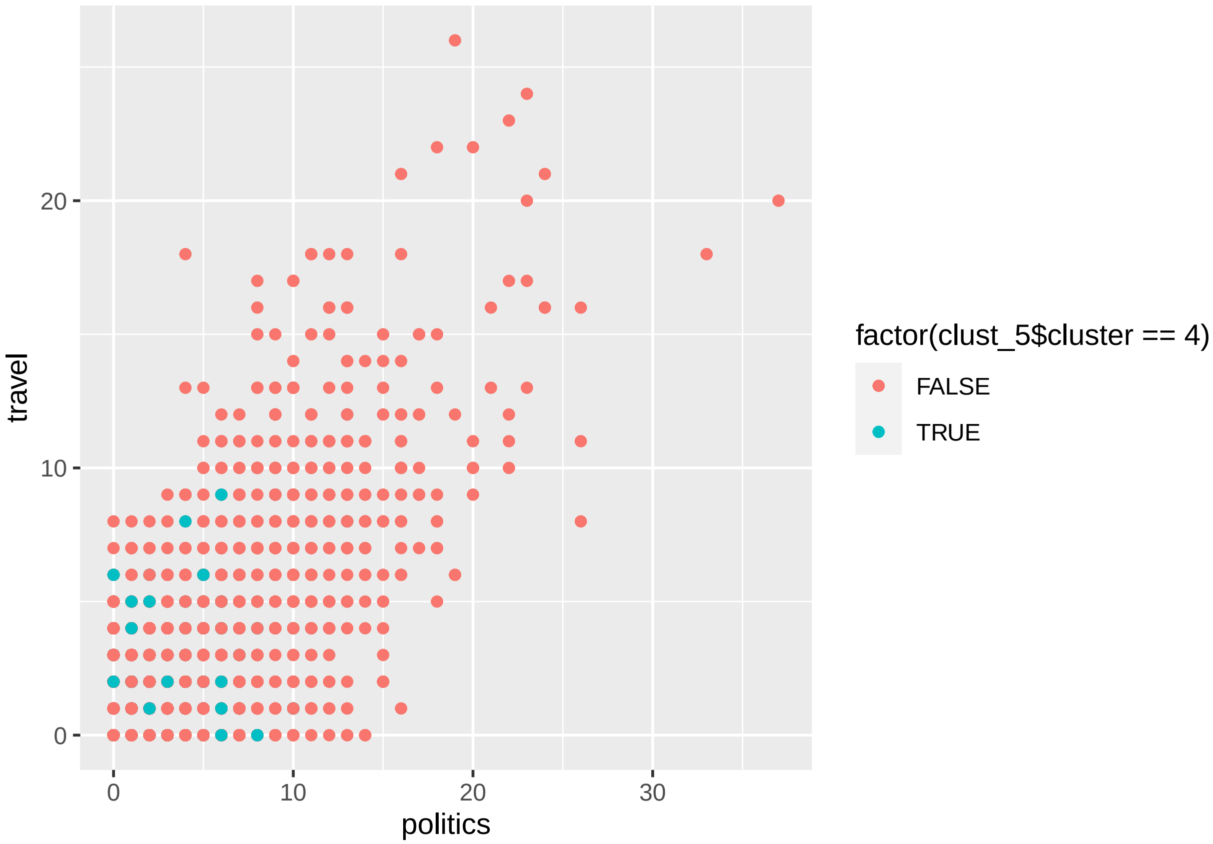
Cluster 4 —> Top 3 Categories are “health_nutrition”, “personal_fitness”, “cooking”



Cluster 5 —> Top 3 Categories are “sports_fandom”, “religion”, “food”

Just for illustrative purposes, we show a few plots that visualize cluster membership below.


photo_sharing
20
15
10
5
0
0
10
20
30
cooking
factor(clust_5$cluster == 1)
● FALSE
● TRUE



Final Results Based on the results of the K-Means with 5 clusters, we can identify some interesting market segments that appear to stand out in their social media audience. The large consumer brand "NutrientH20" can focus their marketing and advertising on specific cluster groups in order to maximize their outreach. For example, with one of the Clusters, they can target consumers that are interested in "health_nutrition", "personal_fitness" and "cooking" all at the same time. On the other hand, using the results from another Cluster, they can target consumers that are interested in "politics", "travel" and "news" all at the same time.

#Problem 5 Author Attribution

```
## Loading required package: NLP
```

```
## 
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
## 
##     annotate
```

```
## 
## Attaching package: 'tm'
```

```
## The following object is masked from 'package:mosaic':
## 
##     inspect
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(tolower)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(my_corpus, removeNumbers): transformation drops
## documents

## Warning in tm_map.SimpleCorpus(my_corpus,
## content_transformer(removePunctuation)): transformation drops documents

## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(stripWhitespace)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(removeWords), :
## transformation drops documents

## [1] "DocumentTermMatrix"    "simple_triplet_matrix"

## Warning in tm_map.SimpleCorpus(test_corpus, content_transformer(tolower)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(test_corpus, content_transformer(removeNumbers)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(test_corpus,
## content_transformer(removePunctuation)): transformation drops documents

## Warning in tm_map.SimpleCorpus(test_corpus,
## content_transformer(stripWhitespace)): transformation drops documents

## Warning in tm_map.SimpleCorpus(test_corpus, content_transformer(removeWords), :
## transformation drops documents

## [1] "DocumentTermMatrix"    "simple_triplet_matrix"

##
## ### stylo version: 0.7.4 ###
##
## If you plan to cite this software (please do!), use the following reference:
##     Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R:
##     a package for computational text analysis. R Journal 8(1): 107-121.
##     <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>
##
## To get full BibTeX entry, type: citation("stylo")
```

For Author Attributions, the modeling package Stylo was used. The task was to predict which author from a set of 50 authors wrote a Reuters article from a set of 2500 articles.

1) The training set was read into a vector with each row being one Article and the value being the text.
2) This vector was then converted to a 'Corpus' Object in R.
3) Tokenization was performed on the Corpus by:

- Converting all characters to lower case
- Removing numbers
- Removing punctuation
- Stripping the white space
- Removing the ‘SMART’ set of stop words.

4) The tokenized corpus was converted to a Document Term Matrix, with common tokens as columns and each article as a row.
5) Steps 1-4 are repeated for the test set
6) The training and test matrices are the given the original authors as row names (As required by Stylo)
7) Initially, we ran a K-Nearest Neighbor with K = 10. Followed by K = 1 and a Naive Bayes Model.

```
## [1] "K Nearest Neighbors (K=10) Accuracy:"
```

```
## [1] 0.3156
```

```
## [1] "K Nearest Neighbors (K=1) Accuracy:"
```

```
## [1] 0.37
```

```
## [1] "Naive Bayes Accuracy:"
```

```
## [1] 0.0988
```

Out of these three classification models, K-Nearest neighbors with K = 1had the highest accuracy at 36.96%. - Accuracy for K = 10: 32.28% - Accuracy for Naive Bayes: 9.88%

This performance of the K=1 model is likely due to authors maintaining a consistent writing style so the Nearest Neighbor to each test document is an article by the same author.

#Problem 6 Association Rule Mining

We set our max length for the left hand side to three because there are only 126 unique items in our data set and we did not want to pick up much noise.

In performing our inspection of association rules, we first looked at support values.

```
##      lhs                   rhs                     support    confidence
## [1]  {}                 => {canned beer}           0.07768175 0.07768175
## [2]  {}                 => {coffee}                0.05805796 0.05805796
## [3]  {}                 => {beef}                  0.05246568 0.05246568
## [4]  {}                 => {curd}                  0.05327911 0.05327911
## [5]  {}                 => {napkins}               0.05236401 0.05236401
## [6]  {}                 => {pork}                  0.05765125 0.05765125
## [7]  {}                 => {frankfurter}           0.05897306 0.05897306
## [8]  {}                 => {bottled beer}          0.08052872 0.08052872
## [9]  {}                 => {brown bread}           0.06487036 0.06487036
## [10] {}                 => {margarine}             0.05856634 0.05856634
## [11] {}                 => {butter}                0.05541434 0.05541434
## [12] {}                 => {newspapers}            0.07981698 0.07981698
## [13] {}                 => {domestic eggs}         0.06344687 0.06344687
## [14] {}                 => {fruit/vegetable juice} 0.07229283 0.07229283
## [15] {}                 => {whipped/sour cream}    0.07168277 0.07168277
## [16] {}                 => {pip fruit}             0.07564820 0.07564820
## [17] {}                  => {pastry}                0.08896797 0.08896797
## [18] {}                  => {citrus fruit}          0.08276563 0.08276563
## [19] {}                  => {shopping bags}         0.09852567 0.09852567
## [20] {}                  => {sausage}               0.09395018 0.09395018
## [21] {}                  => {bottled water}         0.11052364 0.11052364
## [22] {}                  => {tropical fruit}        0.10493137 0.10493137
## [23] {}                  => {root vegetables}       0.10899847 0.10899847
## [24] {}                  => {soda}                  0.17437722 0.17437722
## [25] {}                  => {yogurt}                0.13950178 0.13950178
## [26] {}                  => {rolls/buns}            0.18393493 0.18393493
## [27] {}                  => {other vegetables}      0.19349263 0.19349263
## [28] {}                  => {whole milk}            0.25551601 0.25551601
## [29] {yogurt}            => {whole milk}            0.05602440 0.40160350
## [30] {whole milk}        => {yogurt}                0.05602440 0.21925985
## [31] {rolls/buns}        => {whole milk}            0.05663447 0.30790492
## [32] {whole milk}        => {rolls/buns}            0.05663447 0.22164743
## [33] {other vegetables} => {whole milk}            0.07483477 0.38675775
## [34] {whole milk}        => {other vegetables}      0.07483477 0.29287704
##      coverage  lift     count
## [1]  1.0000000 1.000000  764
## [2]  1.0000000 1.000000  571
## [3]  1.0000000 1.000000  516
## [4]  1.0000000 1.000000  524
## [5]  1.0000000 1.000000  515
## [6]  1.0000000 1.000000  567
## [7]  1.0000000 1.000000  580
## [8]  1.0000000 1.000000  792
## [9]  1.0000000 1.000000  638
## [10] 1.0000000 1.000000  576
## [11] 1.0000000 1.000000  545
## [12] 1.0000000 1.000000  785
## [13] 1.0000000 1.000000  624
## [14] 1.0000000 1.000000  711
## [15] 1.0000000 1.000000  705
## [16] 1.0000000 1.000000  744
## [17] 1.0000000 1.000000  875
## [18] 1.0000000 1.000000  814
## [19] 1.0000000 1.000000  969
## [20] 1.0000000 1.000000  924
## [21] 1.0000000 1.000000 1087
## [22] 1.0000000 1.000000 1032
## [23] 1.0000000 1.000000 1072
## [24] 1.0000000 1.000000 1715
## [25] 1.0000000 1.000000 1372
## [26] 1.0000000 1.000000 1809
## [27] 1.0000000 1.000000 1903
## [28] 1.0000000 1.000000 2513
## [29] 0.1395018 1.571735  551
## [30] 0.2555160 1.571735  551
## [31] 0.1839349 1.205032  557
## [32] 0.2555160 1.205032  557
## [33] 0.1934926 1.513634  736
## [34] 0.2555160 1.513634  736
```

The subsets with the highest support are all single items, which makes sense, as a series of items (how R interprets support) is most often less likely than a single item.

These discoveries all make logical sense.

These high-support single items hit a maximum around .25, and those individual items with the most support are whole milk, other vegetables, rolls/buns, yogurt and soda.

Of the combinations, the support maxes out around .075. This is what we care about. The top combinations are rolls/buns and whole milk, other vegetables and whole milk, and whole milk and yogurt.

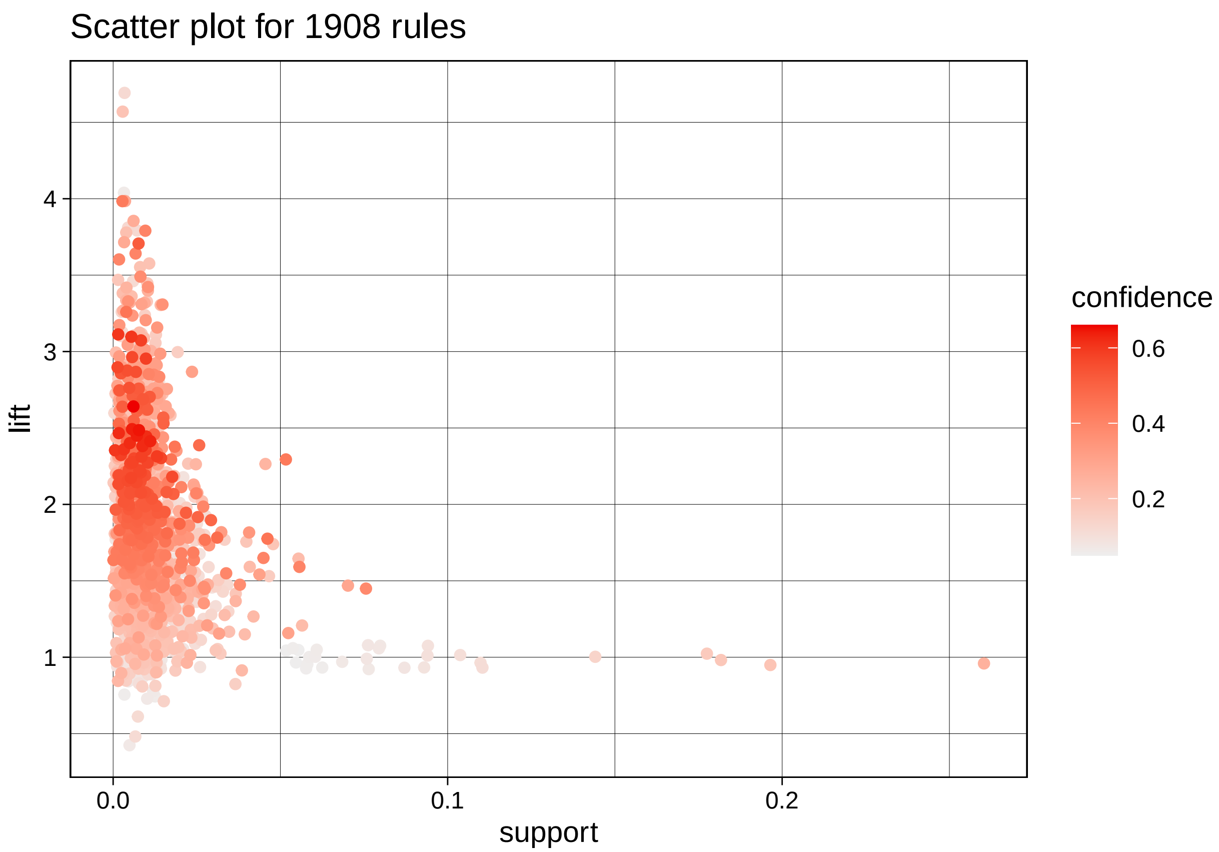
Next, we looked at confidence and located those groups with high confidence ratings.

```
##      lhs                               rhs                support
## [1]  {onions,root vegetables}       => {other vegetables} 0.005693950
## [2]  {curd,tropical fruit}          => {whole milk}       0.006507372
## [3]  {domestic eggs,margarine}      => {whole milk}       0.005185562
## [4]  {butter,domestic eggs}         => {whole milk}       0.005998983
## [5]  {butter,whipped/sour cream}    => {whole milk}       0.006710727
## [6]  {bottled water,butter}         => {whole milk}       0.005388917
## [7]  {butter,tropical fruit}        => {whole milk}       0.006202339
## [8]  {butter,root vegetables}       => {whole milk}       0.008235892
## [9]  {butter,yogurt}                => {whole milk}       0.009354347
## [10] {domestic eggs,pip fruit}      => {whole milk}       0.005388917
## [11] {domestic eggs,tropical fruit} => {whole milk}       0.006914082
## [12] {pip fruit,whipped/sour cream} => {other vegetables} 0.005592272
## [13] {pip fruit,whipped/sour cream} => {whole milk}       0.005998983
##      confidence coverage    lift     count
## [1]  0.6021505  0.009456024 3.112008 56
## [2]  0.6336634  0.010269446 2.479936 64
## [3]  0.6219512  0.008337570 2.434099 51
## [4]  0.6210526  0.009659380 2.430582 59
## [5]  0.6600000  0.010167768 2.583008 66
## [6]  0.6022727  0.008947636 2.357084 53
## [7]  0.6224490  0.009964413 2.436047 61
## [8]  0.6377953  0.012913066 2.496107 81
## [9]  0.6388889  0.014641586 2.500387 92
## [10] 0.6235294  0.008642603 2.440275 53
## [11] 0.6071429  0.011387900 2.376144 68
## [12] 0.6043956  0.009252669 3.123610 55
## [13] 0.6483516  0.009252669 2.537421 59
```

The subsets with the highest confidence are all combinations, and very many of them involve whole milk as the right hand consequent. In fact, all of the consequents are either whole milk or other vegetables. Many of the antecedents include butter or domestic eggs. Confidence maxes out around .65.

Here is a Plot of support and lift, with shading indicating confidence. This helps illustrate the relationships between the three.

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```

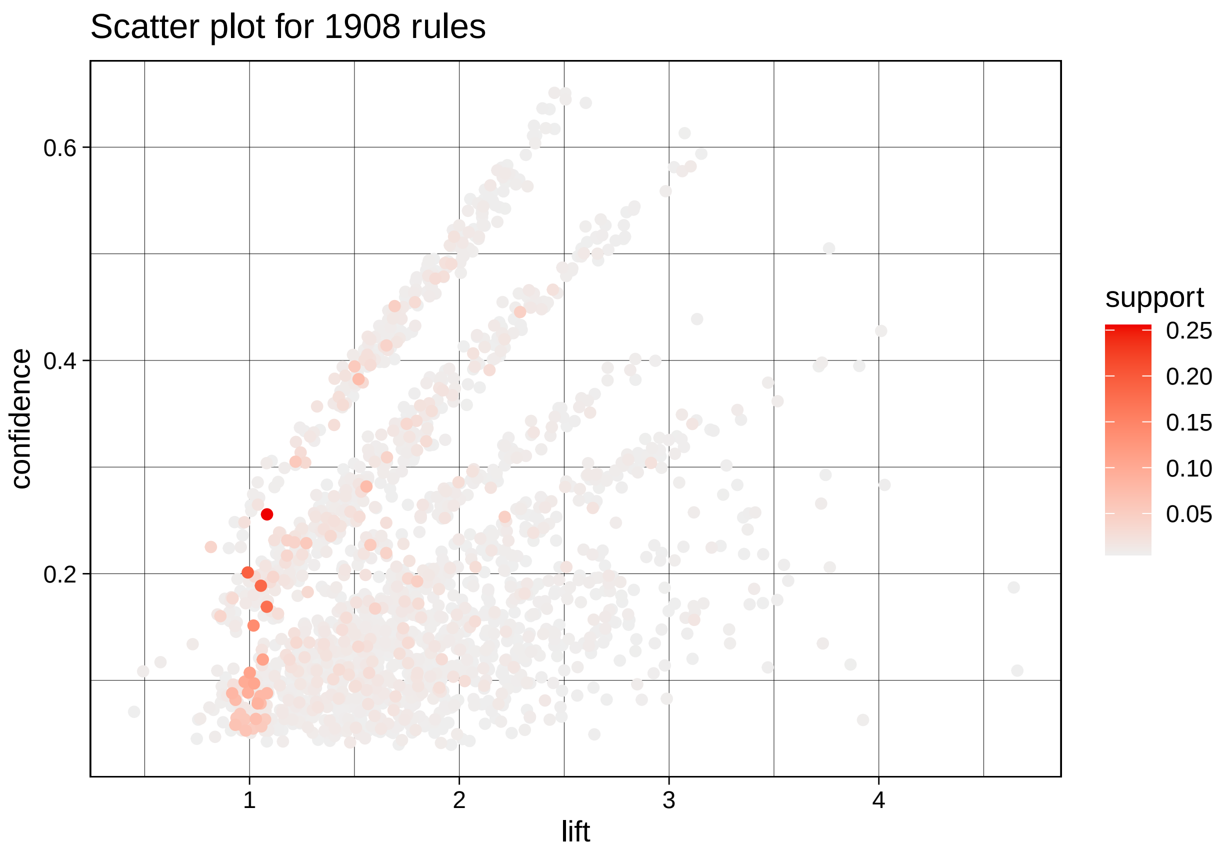


We then created a group of antecedents and consequents that had a lift of one or greater, and a confidence of 0.25 or greater. This means that they are not substitutes, and lhs occurs at least .25 percent of the time given rhs.

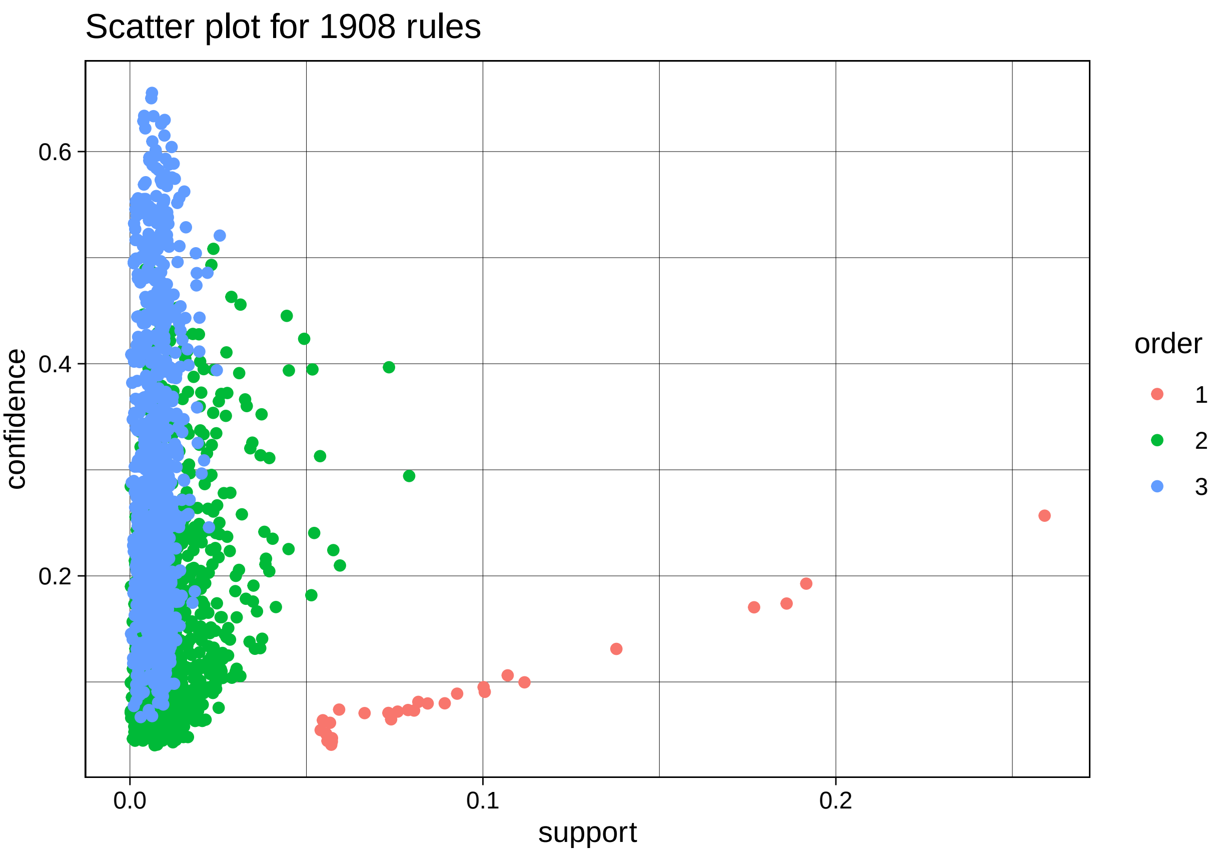
This group is called "related." There are 615 pairs in our "related" subset.

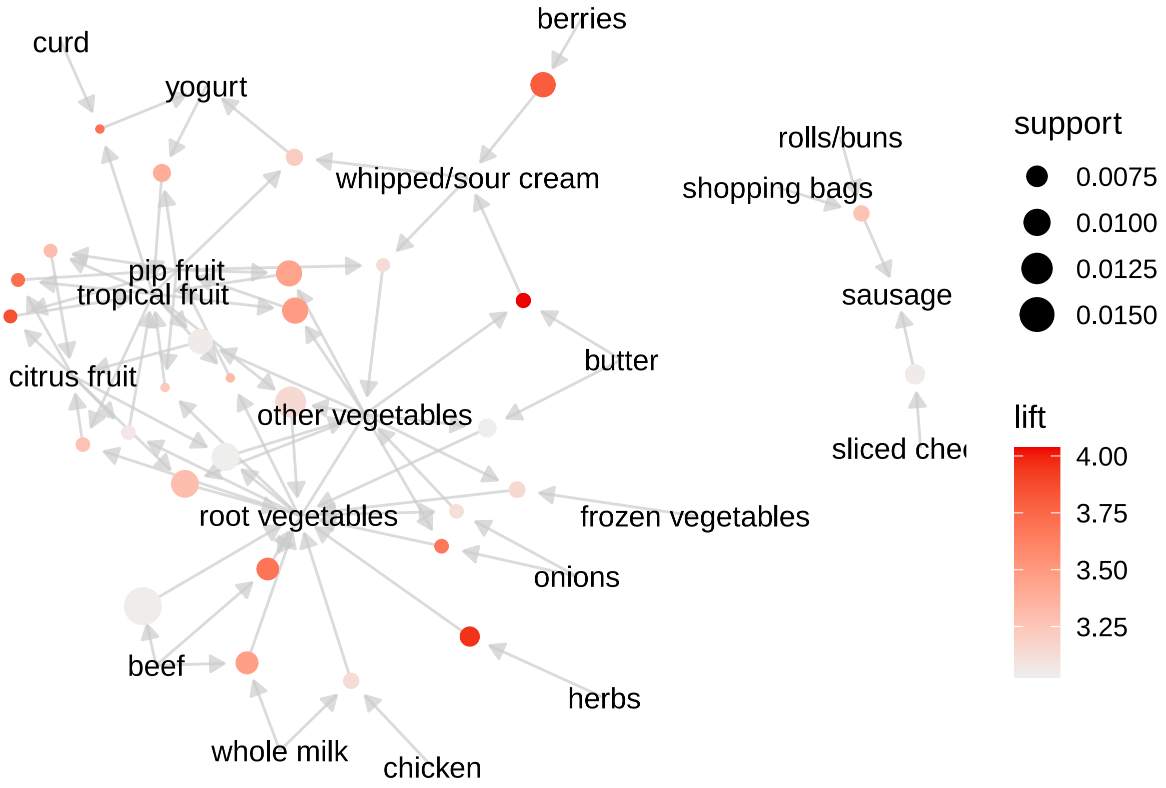
Next, we plot our "related" subset out in a way that visualizes associations.

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```


Scatter plot for 1908 rules
confidence
0.6
0.4
0.2
support
0.0
0.1
0.2
order
1
2
3



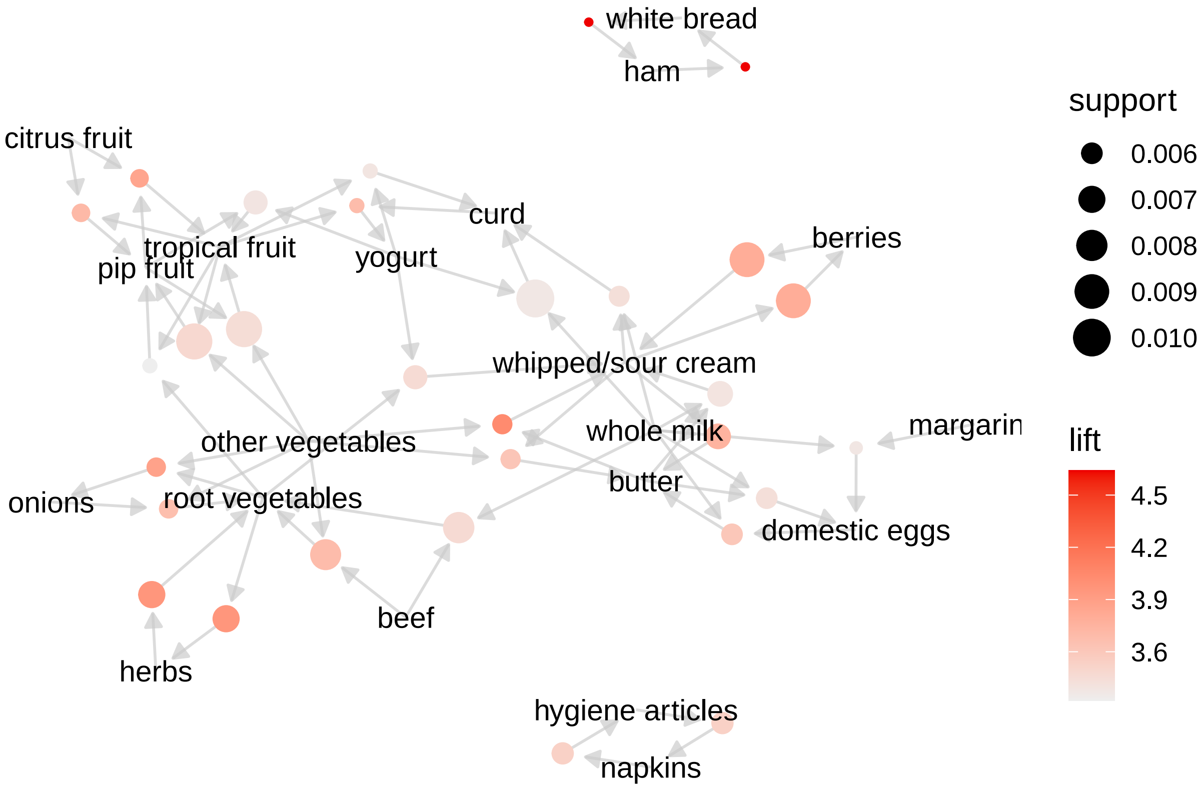
To evaluate substitute groups, we created a subset where lift is less than .75

```
substitutes = subset(groceryrules, subset=lift < .75)
```

This subset shows that beer and yogurt, beer and vegetables, beer and whole milk, and bottled beer and shopping bags are all substitutes. People who buy yogurt and vegetables are less likely to buy beer than those who do not. This may be because "beer runs" and actual grocery shopping fall into two different behavioral categories.

&nsbp;

Finally, we created another grouping to signify compliments, and identified these as groupings with high lift. Our lift cutoff is 3.05, which gives us a nice even group of 50 pairings.



The highest lifts are ham and white bread, butter/other vegetables and whipped/sour cream, and citrus fruit/pip fruit and tropical fruit. In looking through each of these high lift combinations and searching for unexpected associations, we find that they all make sense.