STA 380, Part 2: Exercises

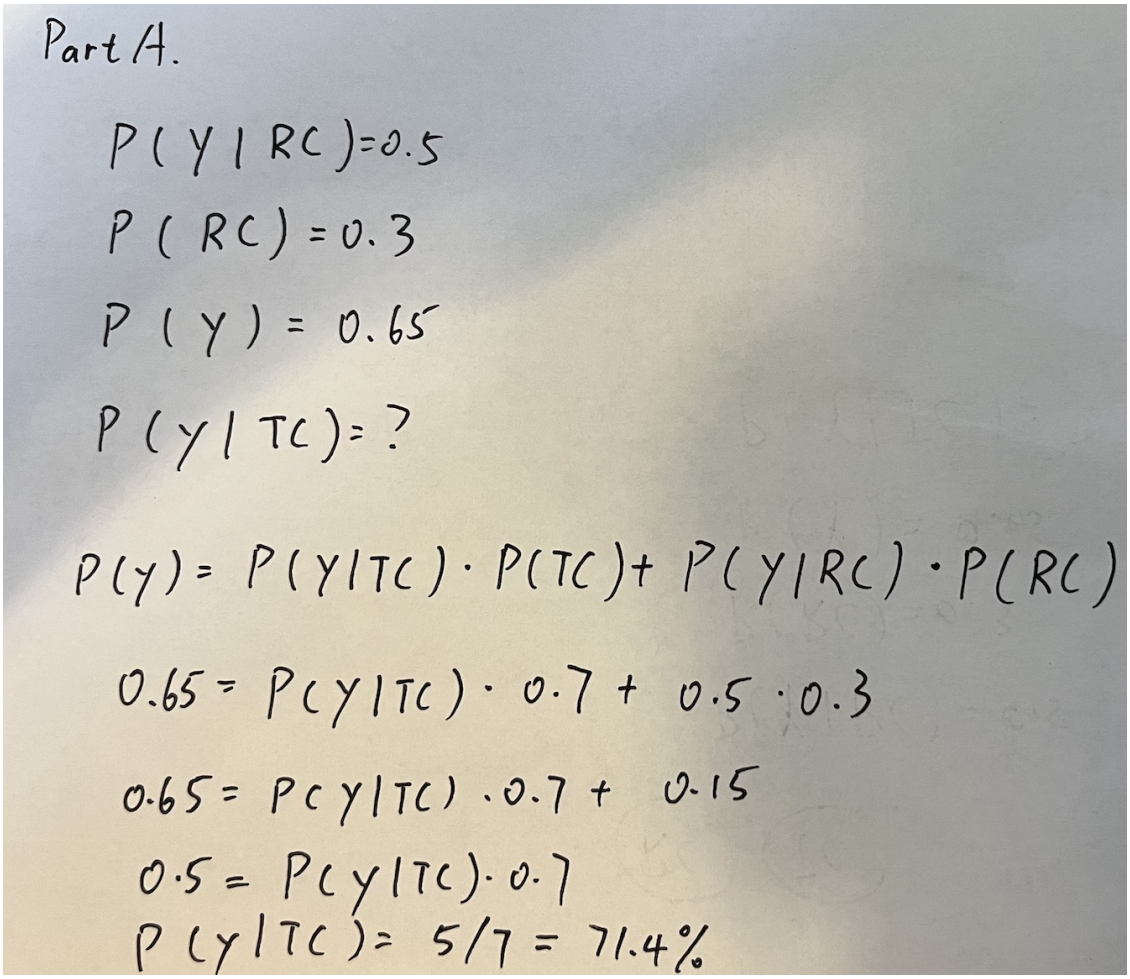
Anthony Moreno, Mark Moreno, Kai Zhang, Carlee Allen

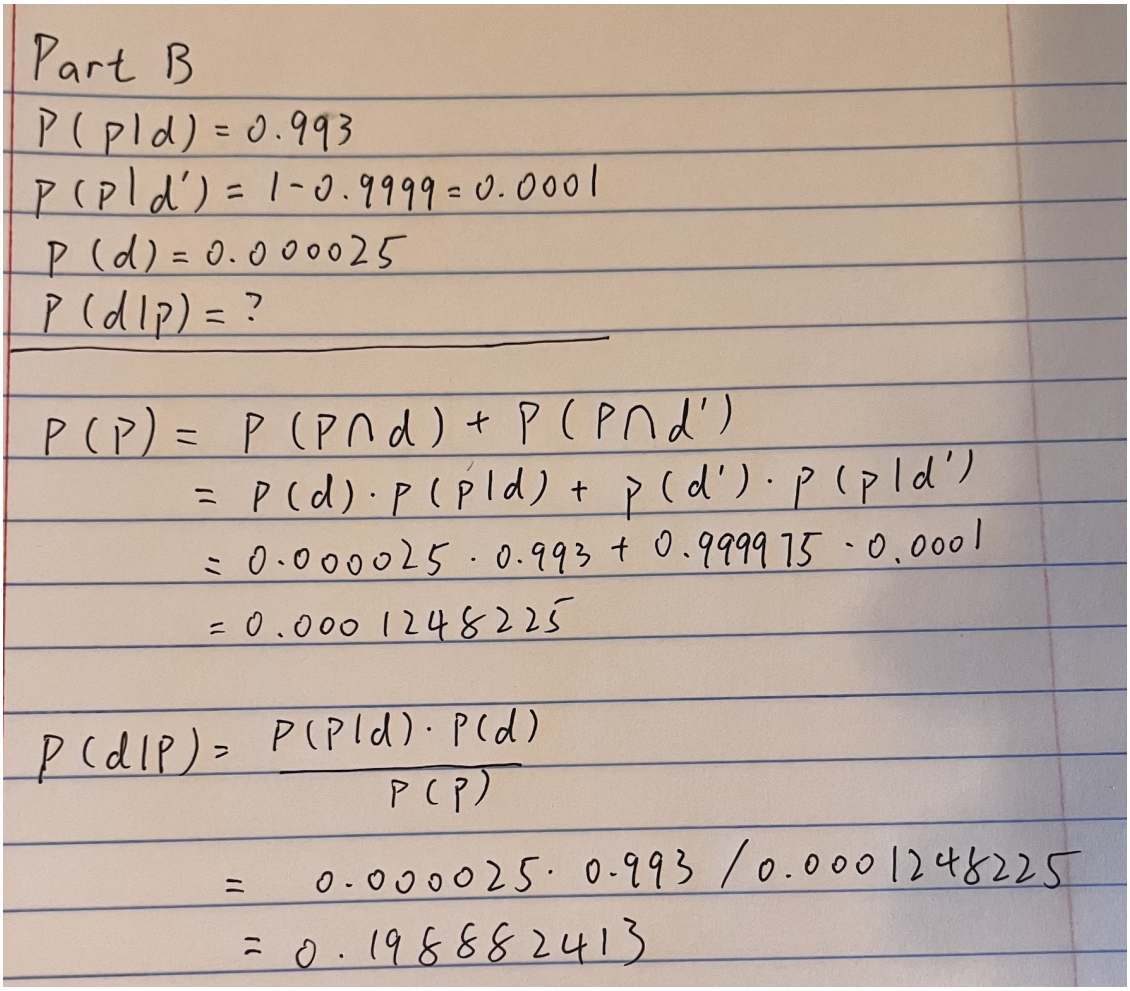
2022-08-15

STA 380, Part 2: Exercises

Contents

Probability Practice





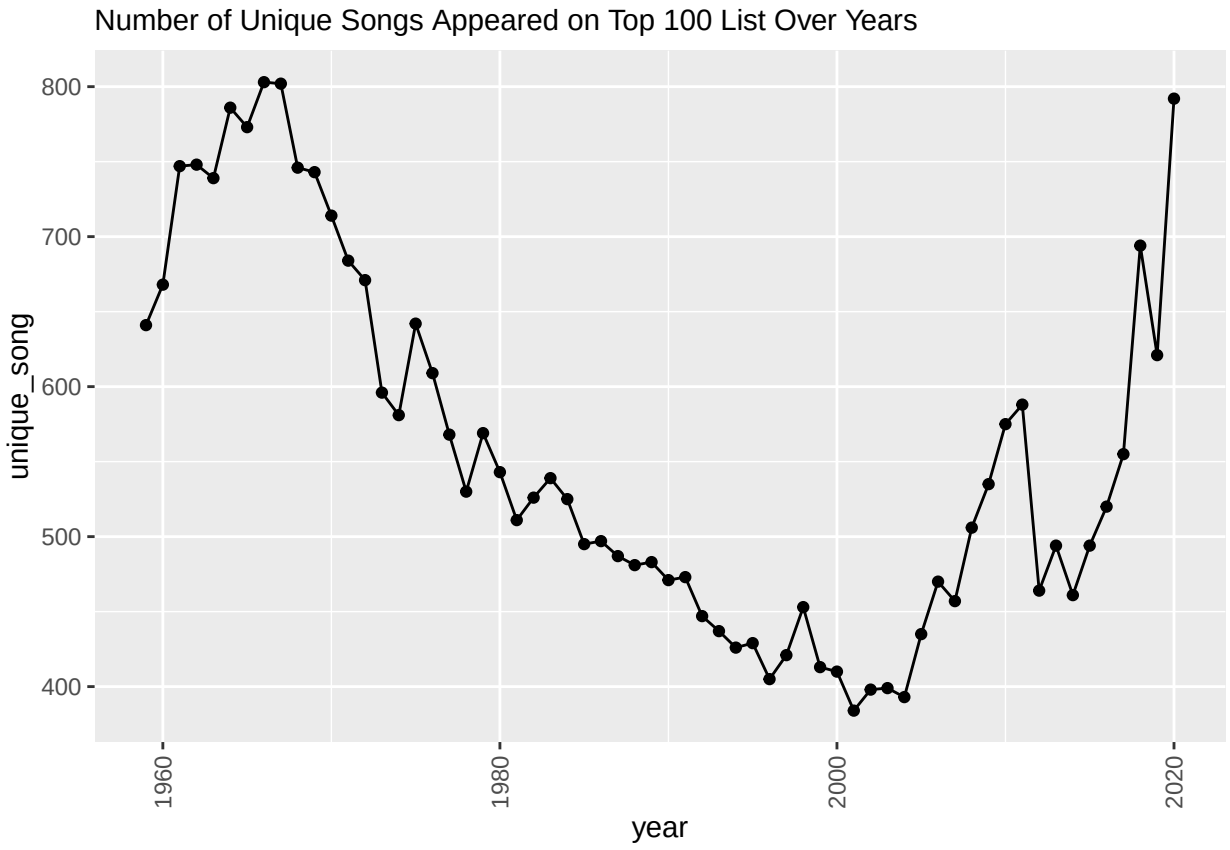
Wrangling the Billboard Top 100

Part A

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## `summarise()` has grouped output by 'song'. You can override using the
## `.groups` argument.

## # A tibble: 10 x 3
## # Groups:   song [10]
```

```
##    song                                     performer                          count
##    <chr>                                    <chr>                              <int>
##  1 Radioactive                              Imagine Dragons                       87
##  2 Sail                                     AWOLNATION                            79
##  3 Blinding Lights                          The Weeknd                            76
##  4 I'm Yours                                Jason Mraz                            76
##  5 How Do I Live                            LeAnn Rimes                           69
##  6 Counting Stars                           OneRepublic                           68
##  7 Party Rock Anthem                        LMFAO Featuring Lauren Bennett & G~    68
##  8 Foolish Games/You Were Meant For Me      Jewel                                 65
##  9 Rolling In The Deep                      Adele                                 65
## 10 Before He Cheats                         Carrie Underwood                      64
```

Part B

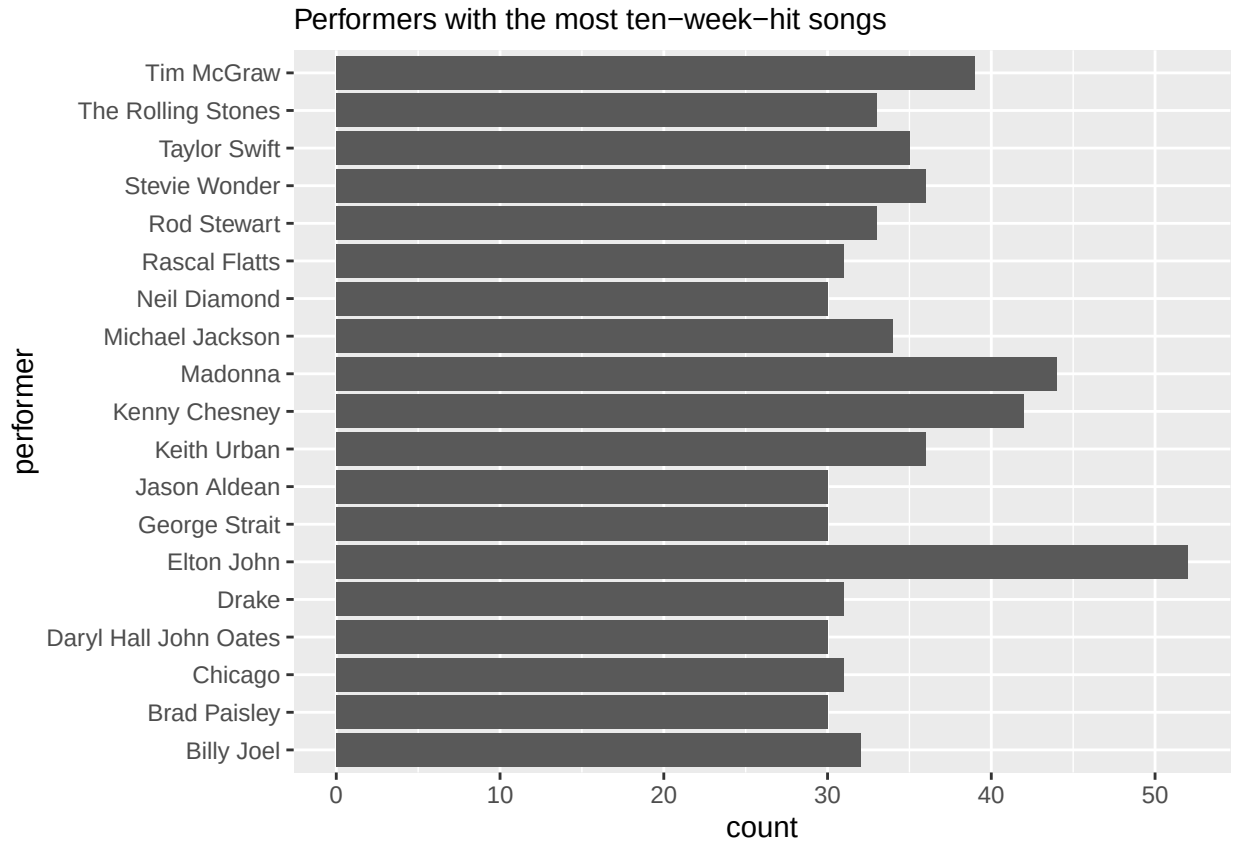


Music diversity declined starting from the 70. People seems to like the same music with less diversity from the 70s through mid 2000s. People's taste has become more diverse since then and has reached the diversity peak in mid 60s.

Part C

```
## # A tibble: 19 x 2
##    performer             count
##    <chr>                 <int>
```

```
##  1 Brad Paisley             30
##  2 Daryl Hall John Oates    30
##  3 George Strait            30
##  4 Jason Aldean             30
##  5 Neil Diamond             30
##  6 Chicago                  31
##  7 Drake                    31
##  8 Rascal Flatts            31
##  9 Billy Joel               32
## 10 Rod Stewart              33
## 11 The Rolling Stones       33
## 12 Michael Jackson          34
## 13 Taylor Swift             35
## 14 Keith Urban              36
## 15 Stevie Wonder            36
## 16 Tim McGraw               39
## 17 Kenny Chesney            42
## 18 Madonna                  44
## 19 Elton John               52
```



Visual story telling part 1: green buildings

Problem

An Austin real-estate developer is interested in the possible economic impact of "going green" in her latest project: a new 15-story mixed-use building on East Cesar Chavez, just across I-35 from downtown. Will investing in a green building be worth it, from an economic perspective? The baseline construction costs are $100 million, with a 5% expected premium for green certification.

The developer has had someone on her staff, who's been described to her as a "total Excel guru from his undergrad statistics course," run some numbers on this data set and make a preliminary recommendation. Here's how this person described his process:

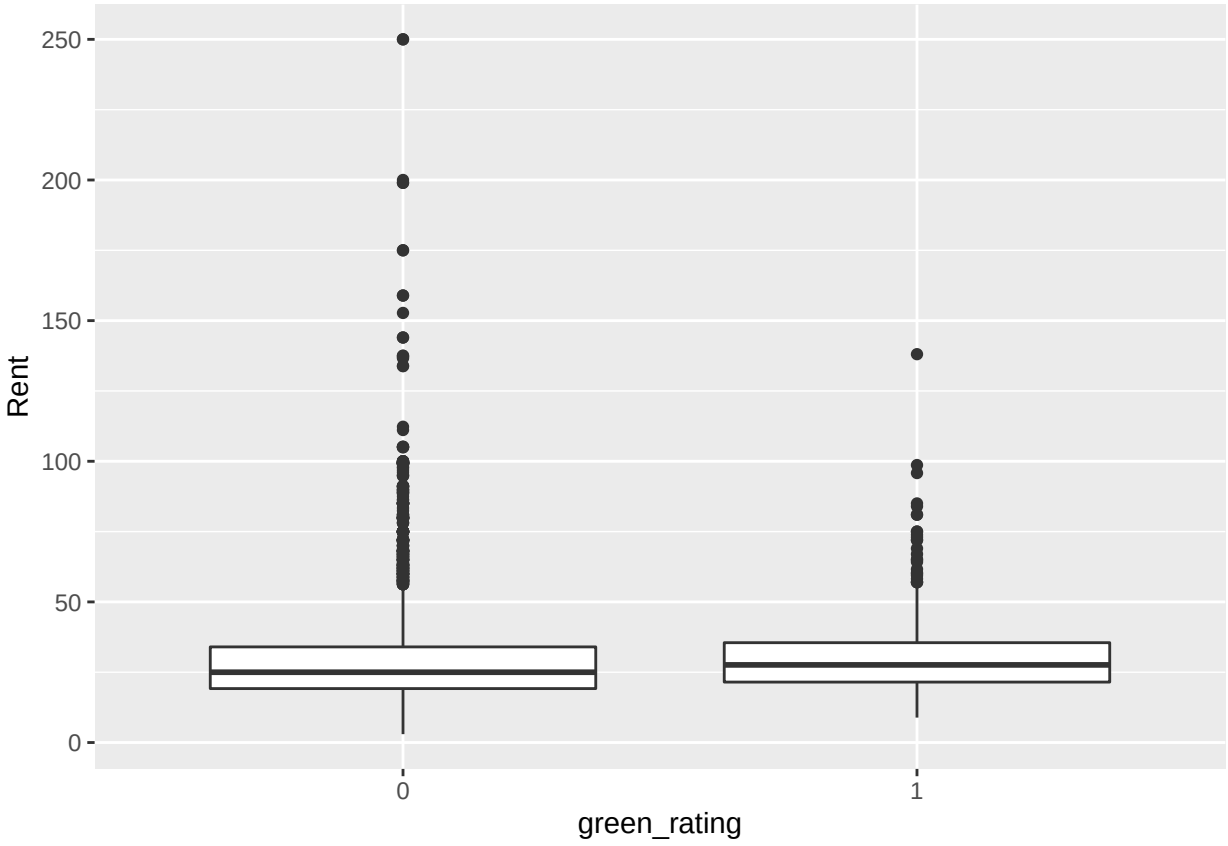
> I began by cleaning the data a little bit. In particular, I noticed that a handful of the buildings in the data set had very low occupancy rates (less than 10% of available space occupied). I decided to remove these buildings from consideration, on the theory that these buildings might have something weird going on with them, and could potentially distort the analysis. Once I scrubbed these low-occupancy buildings from the data set, I looked at the green buildings and non-green buildings separately. The median market rent in the non-green buildings was $25 per square foot per year, while the median market rent in the green buildings was $27.60 per square foot per year: about $2.60 more per square foot. (I used the median rather than the mean, because there were still some outliers in the data, and the median is a lot more robust to outliers.) Because our building would be 250,000 square feet, this would translate into an additional $250000 x 2.6 = $650000 of extra revenue per year if we build the green building.
>
> Our expected baseline construction costs are $100 million, with a 5% expected premium for green certification. Thus we should expect to spend an extra $5 million on the green building. Based on the extra revenue we would make, we would recuperate these costs in $5000000/650000 = 7.7 years. Even if our occupancy rate were only 90%, we would still recuperate the costs in a little over 8 years. Thus from year 9 onwards, we would be making an extra $650,000 per year in profit. Since the building will be earning rents for 30 years or more, it seems like a good financial move to build the green building. Goal: The developer listened to this recommendation, understood the analysis, and still felt unconvinced. She has therefore asked you to revisit the report, so that she can get a second opinion.

Exploratory Data Analysis

it may seem like the stats guru is on point with their analysis upon first glance when using the same assumption and removing the rows with occupancy rates lower than 10%, the green buildings still average $2.6 dollars more per square foot.

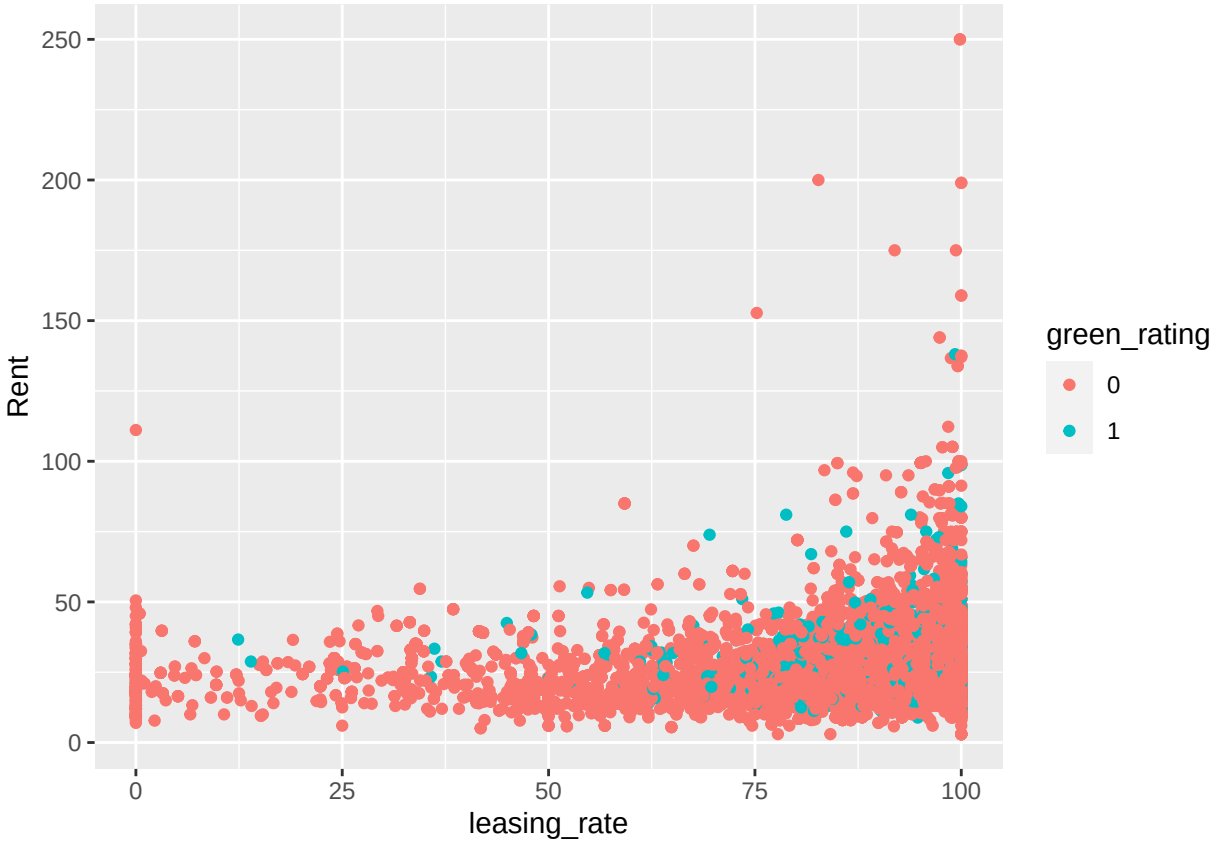
```
## # A tibble: 2 x 3
##   green_rating med_rent count
##   <fct>           <dbl> <int>
## 1 0                25.0  6995
## 2 1                27.6   684
```

While looking at the box plot we see that the non-green buildings have a fair amount more outliers, which explains the guru's reasoning for using the median as opposed to the mean.



Rent vs. Occupancy Rates

We can see that the minimum rent is fairly constant for all occupancy rates, but the maximum increases with higher occupancy rates:

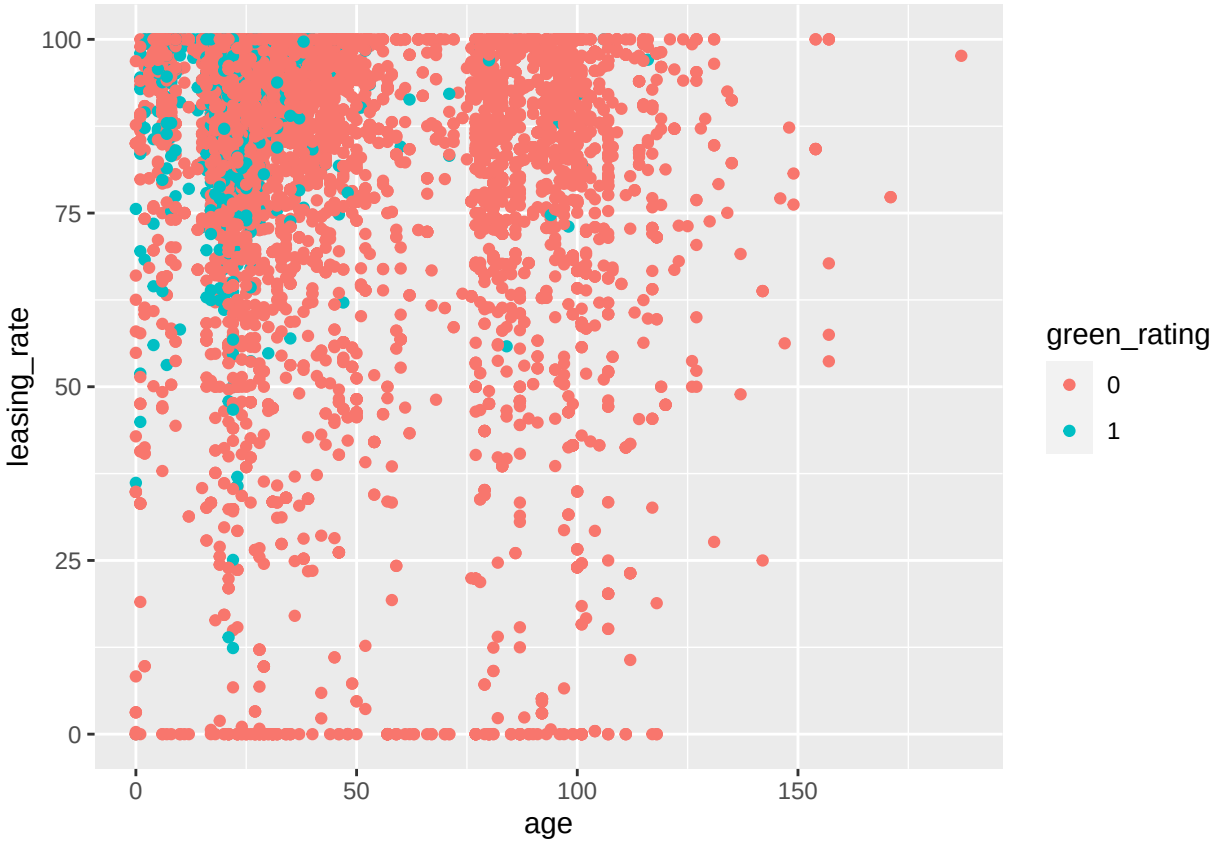


Looking further into this we can note that the green buildings have a higher median occupancy rate, which could be attributed to the awarness mentioned in the information provided with the problem, but the true cause can't be verified with the data provded.

```
## # A tibble: 2 x 3
##   green_rating med_occupancy count
##   <fct>                <dbl> <int>
## 1 0                     89.2  7209
## 2 1                     92.9   685
```

Age vs. Occupancy Rates

From the plot there is no clear relationship between age and occupancy rate.



POSSIBLE UNEXPECTED FACTORS AFFECTING RENT DIFFERENCE

Factors that may inderectly influence the rent for buildings

1. Number of stories

From the plot, we can see that the minimum rent charged increases with the number of stories, but the median height of a green building is only one story taller than a non-green building, which likely isn't a large enough difference. So while a taller building may be able to demand more rent, this likely isn't a strong enough factor to influence the decision to go green.


```
## # A tibble: 2 x 2
##   green_rating med_stories
##   <fct>              <int>
## 1 0                     10
## 2 1                     11
```

2. Amenities

When loooking at the possility of amenities being a factor in raising rent, we see that most green buildings have amenities (~73%), where about only half non-green buildings have them (~52%). When comparting rent for green and non-green buildings with and withput amenities, we see that green buildings still charge $2 to $2.8 more per square foot irreguarless of amenities, meaning this isn't a factor in the differnce in rent.

```
## `summarise()` has grouped output by 'green_rating'. You can override using the
## `.groups` argument.
```

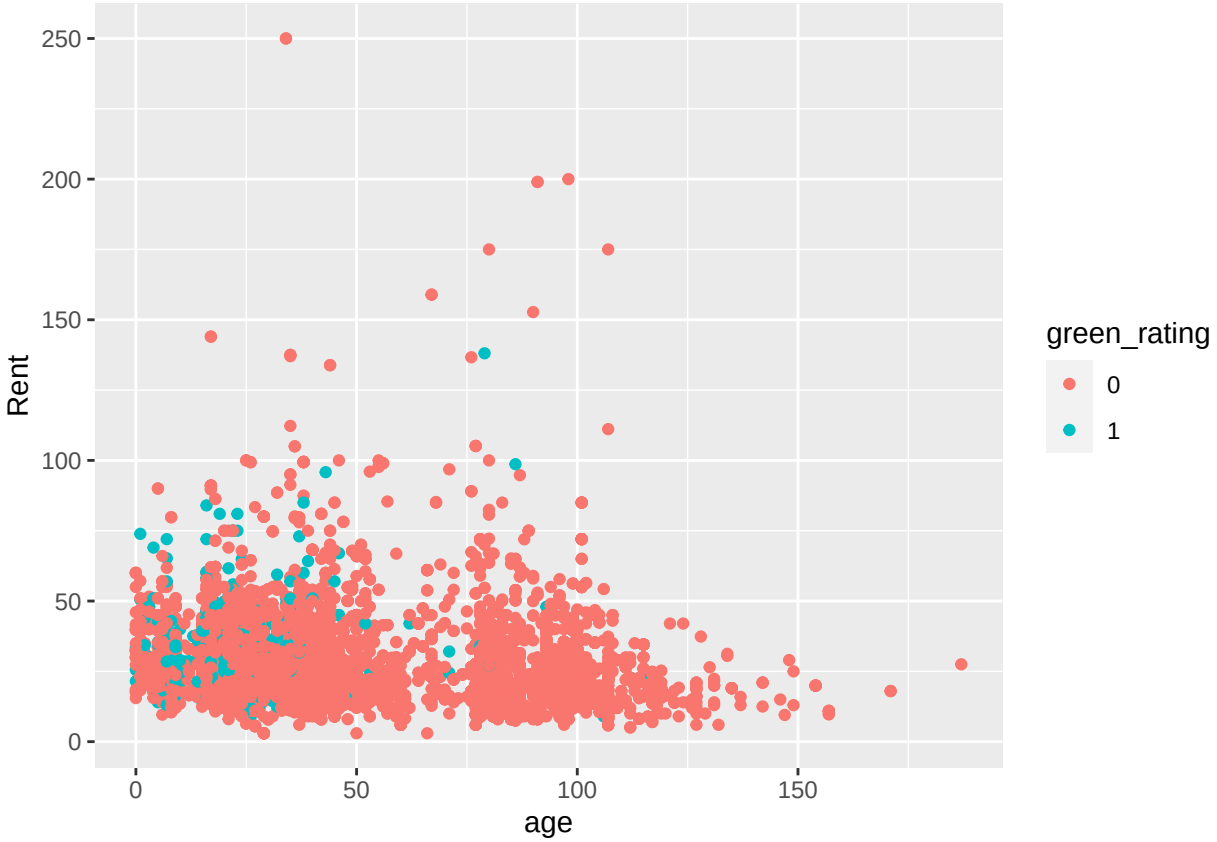
```
## # A tibble: 4 x 4
## # Groups:   green_rating [2]
##   green_rating amenities med_rent count
##   <fct>        <fct>        <dbl> <int>
## 1 0            0             25    3550
## 2 0            1             25    3659
## 3 1            0             27     187
## 4 1            1             27.8   498
```

3. Age

We can see that the median green building is about 15 years newer than the median non-green building, thus we should consider that maybe newer buildings demand a higher rent by default and that this upcharge isn't necessarily due to the fact the building is green.

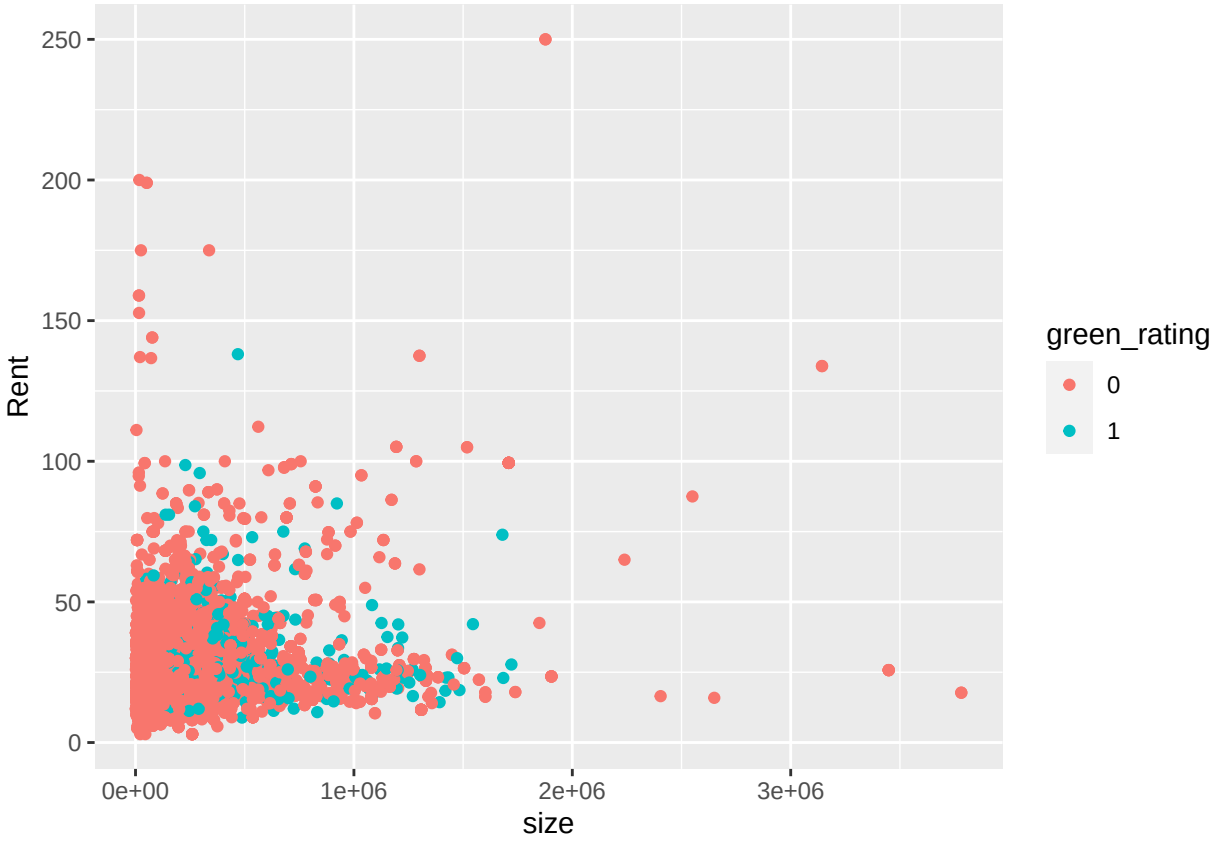
```
## # A tibble: 2 x 3
##   green_rating med_age count
##   <fct>          <int> <int>
## 1 0                 37  7209
## 2 1                 22   685
```

Looking at the plot of rent vs building age, there is hard to see a clear correlation in the data, with maybe a slight downward trend after the building has reached 100 years old, but there isn't enough data on green buildings at that age to manke any conclusions.



4. Space

When looking at the plot of rent vs square footage, we can see that the minimum rent charged increases with available leasing space.



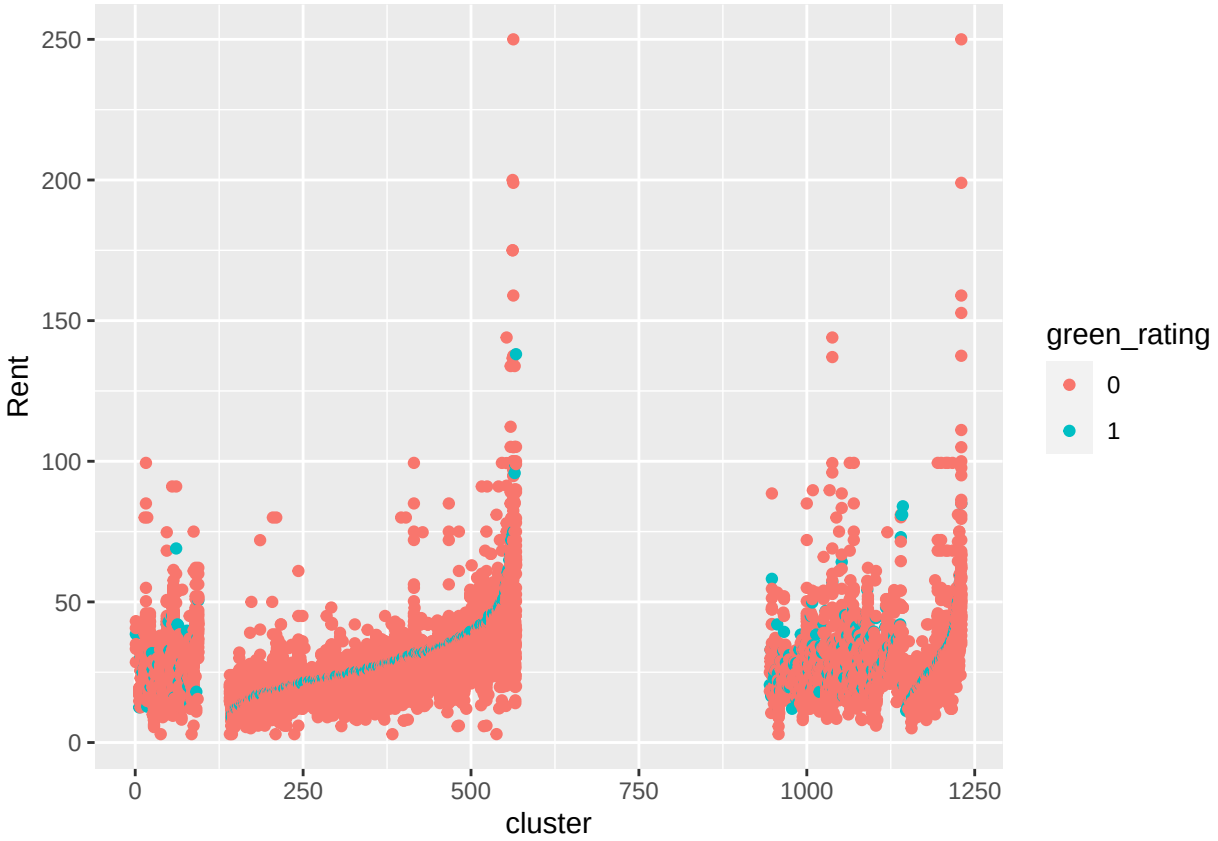
Then, when we look a the median and mean square footage for green and non-green buildings we see that green building have a greater value for each, so they have more space to lease in general, which could be a potential unexpected factor in the rent upcharge.

```
## # A tibble: 2 x 3
##   green_rating med_size mean_size
##   <fct>           <int>     <dbl>
## 1 0              118696   225977.
## 2 1              241150   325781.
```

5. Clustering (Building location)

Could location influence the rent price for buildings? Could we use location to maximize profits for our new building?

Looking at a plot of rent vs cluster, we can notice a distinct trend in clusters 300 to around 600. When looking at the median value for these clusters separating by if the building is green or not, we can see that the green buildings charge ~$4.8 more per square foot in the same cluster, meaning in these areas the perception of a green building is potentially more positive and people are willing to pay more to be viewed as environmentally conscious.



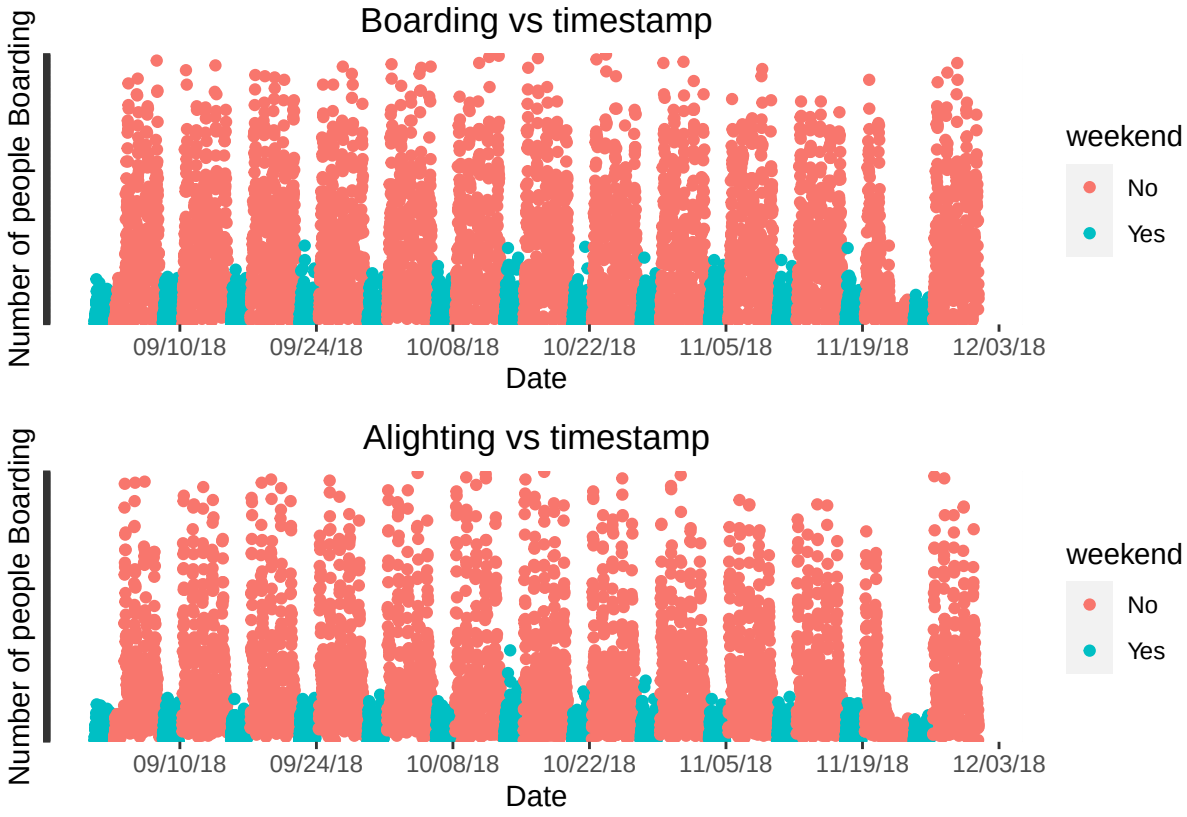

```
## # A tibble: 2 x 3
##   green_rating   med count
##   <fct>        <dbl> <int>
## 1 0             28.2  2640
## 2 1             33     243
```

Summary and Conclusions

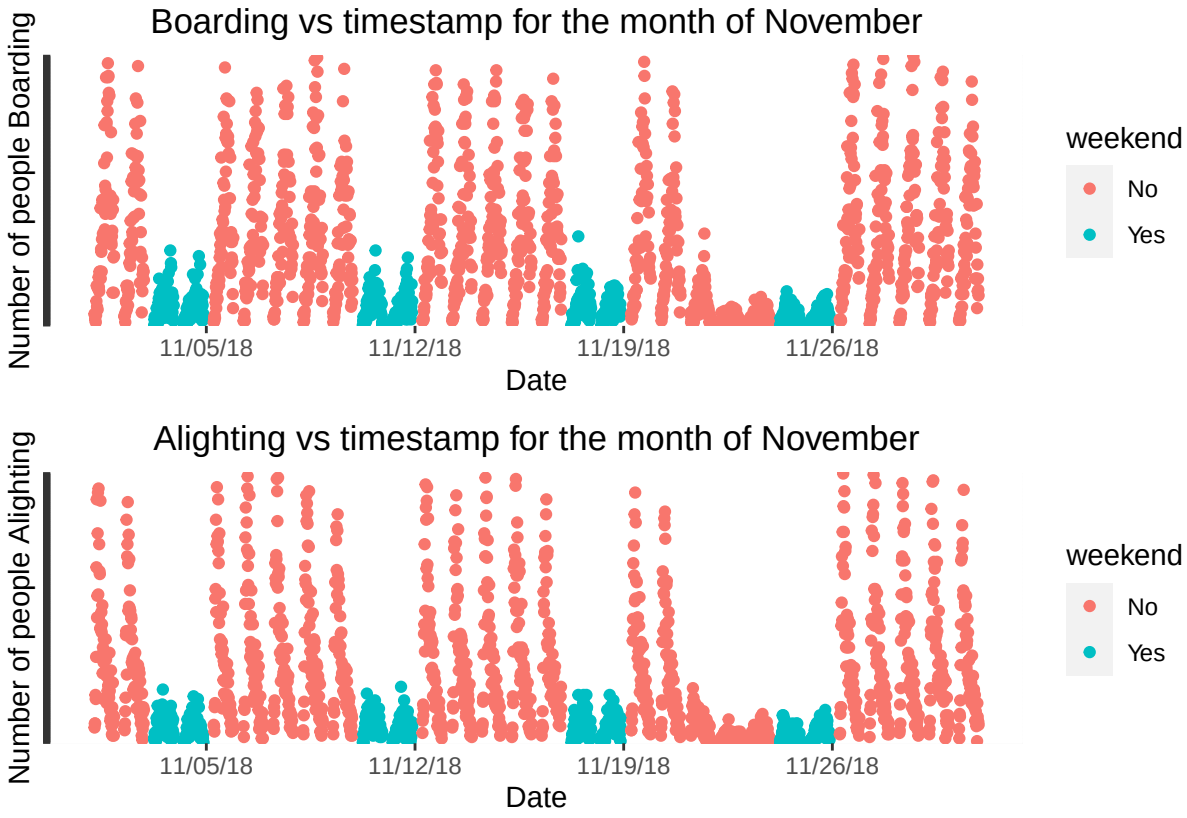
- There is a slight positive relationship between rent and occupancy rates.
- Green buildings have a slightly higher rate of occupancy.
- No clear corrolation between building age and occupancy rate
- There is an additional $2.6 per square foot in revenue for green buildings, and this difference increases to ~$4.8 for clusters 300 - 600.
- Rent and available square footage have a small positive corrolation.
- Green buildings are, on average, ~ 100,000 square feet larger than non-green buildings, but more data is needed to further explore this idea.

The guru seems to be correct, with the average age of green buildings being 22 years, we should expect to make our money back within 10 years then make additional revenue from then on.

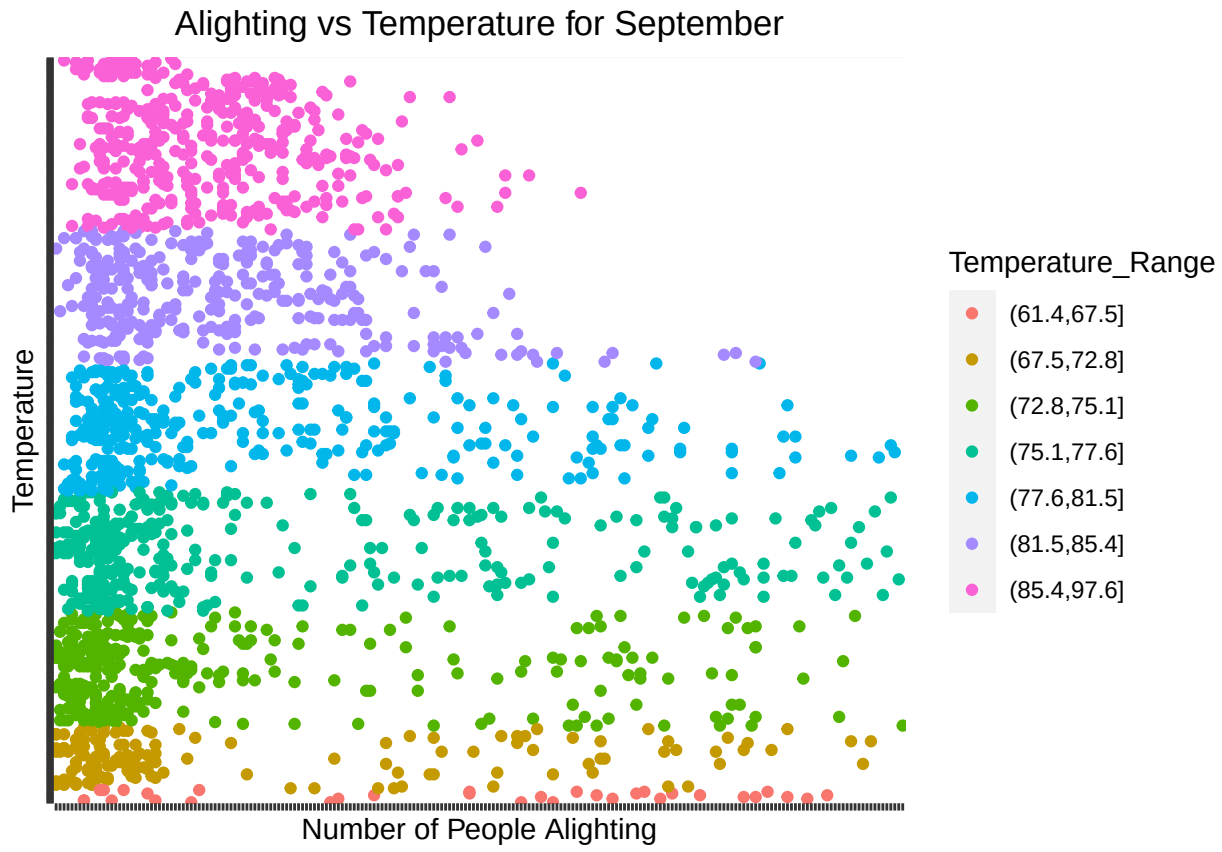
Visual story telling part 2: Capital Metro data



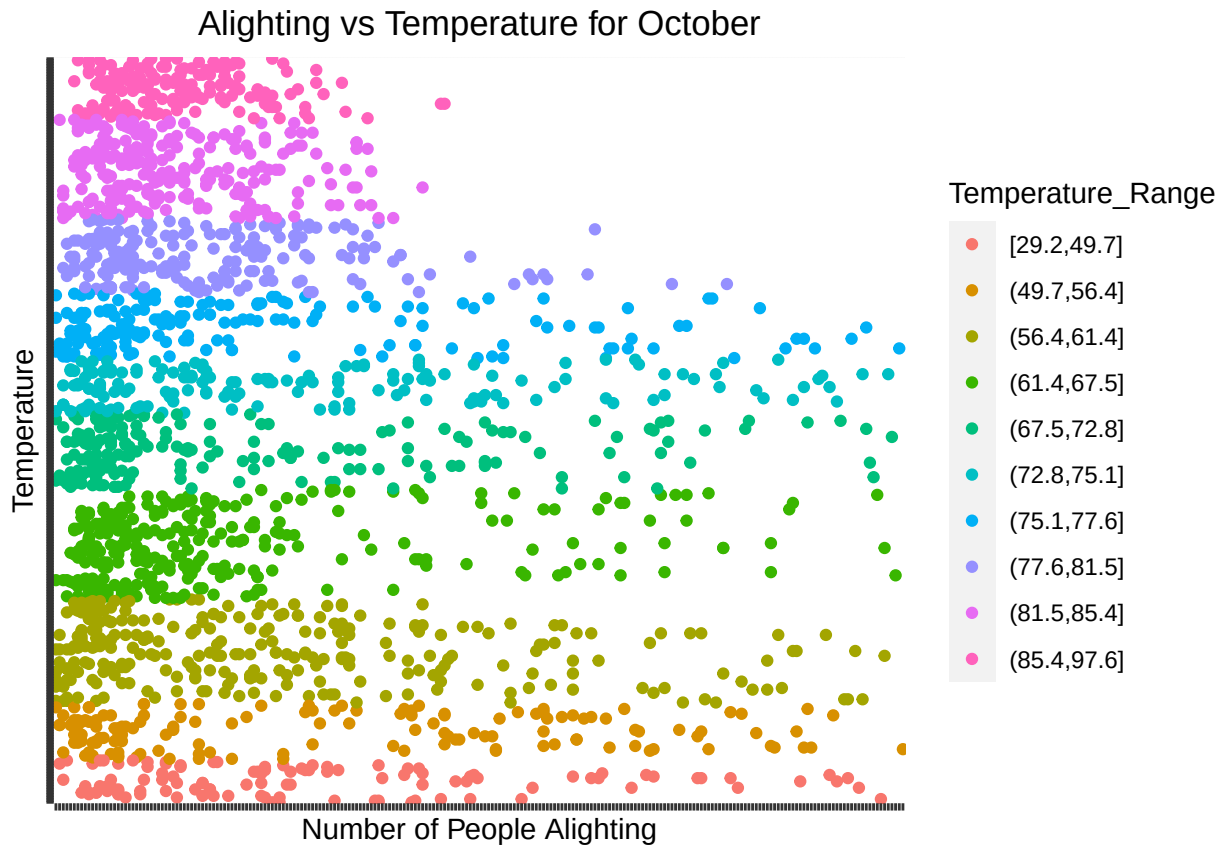
Boarding and alighting patterns over entire timeframe, with weekend indicator



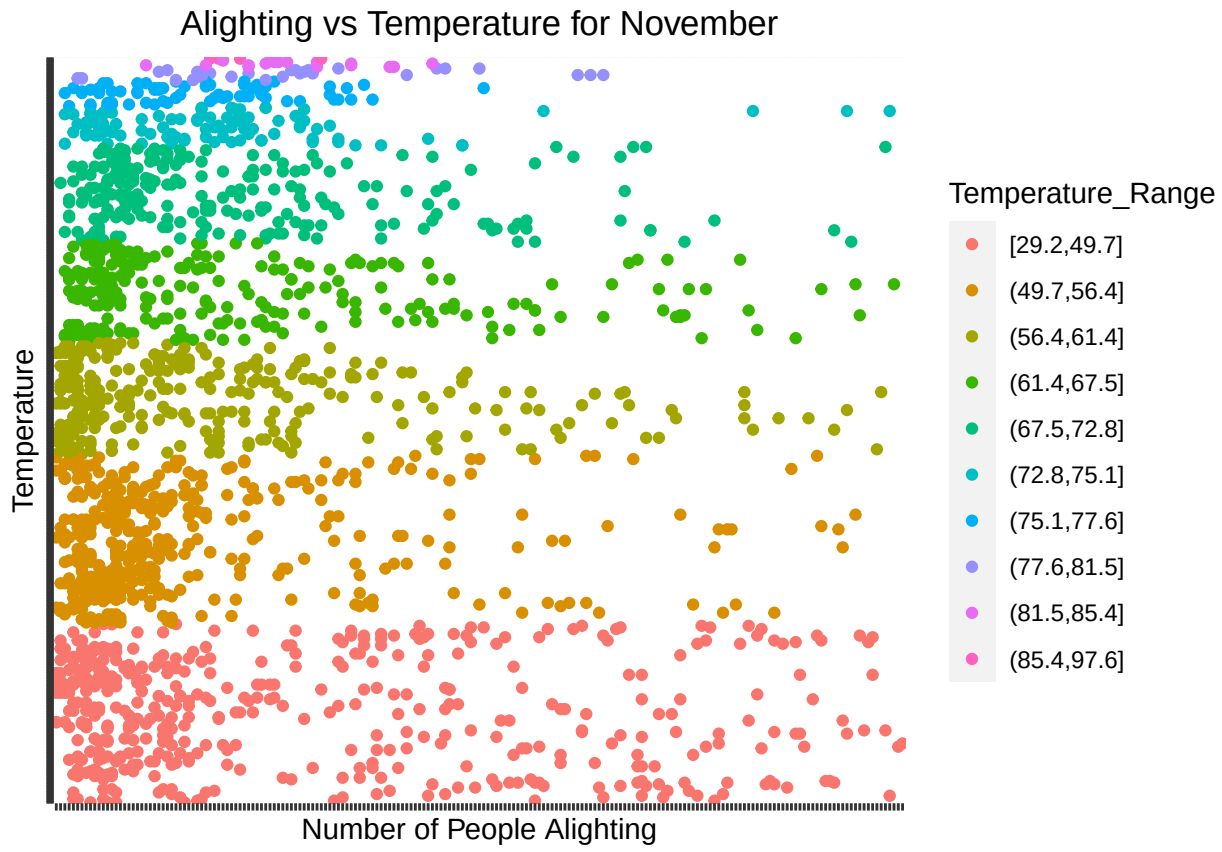
Boarding and alighting patterns for the month of Novenmber, with weekend indicator. Note the slow down in the second to last week of the month



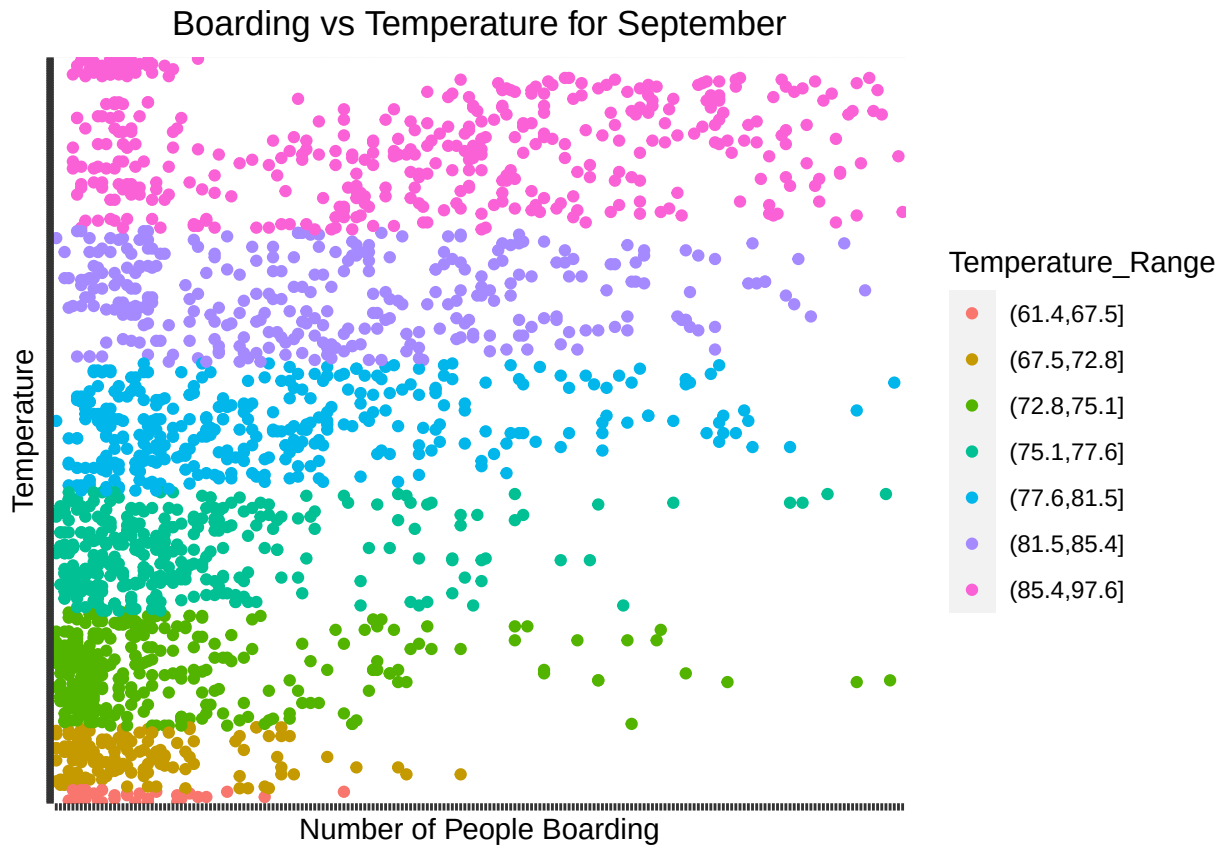
Number of people alighting per 15 minutue interval vs temperature range for the month of September, with color indicated by temperature range bucket the point falls into



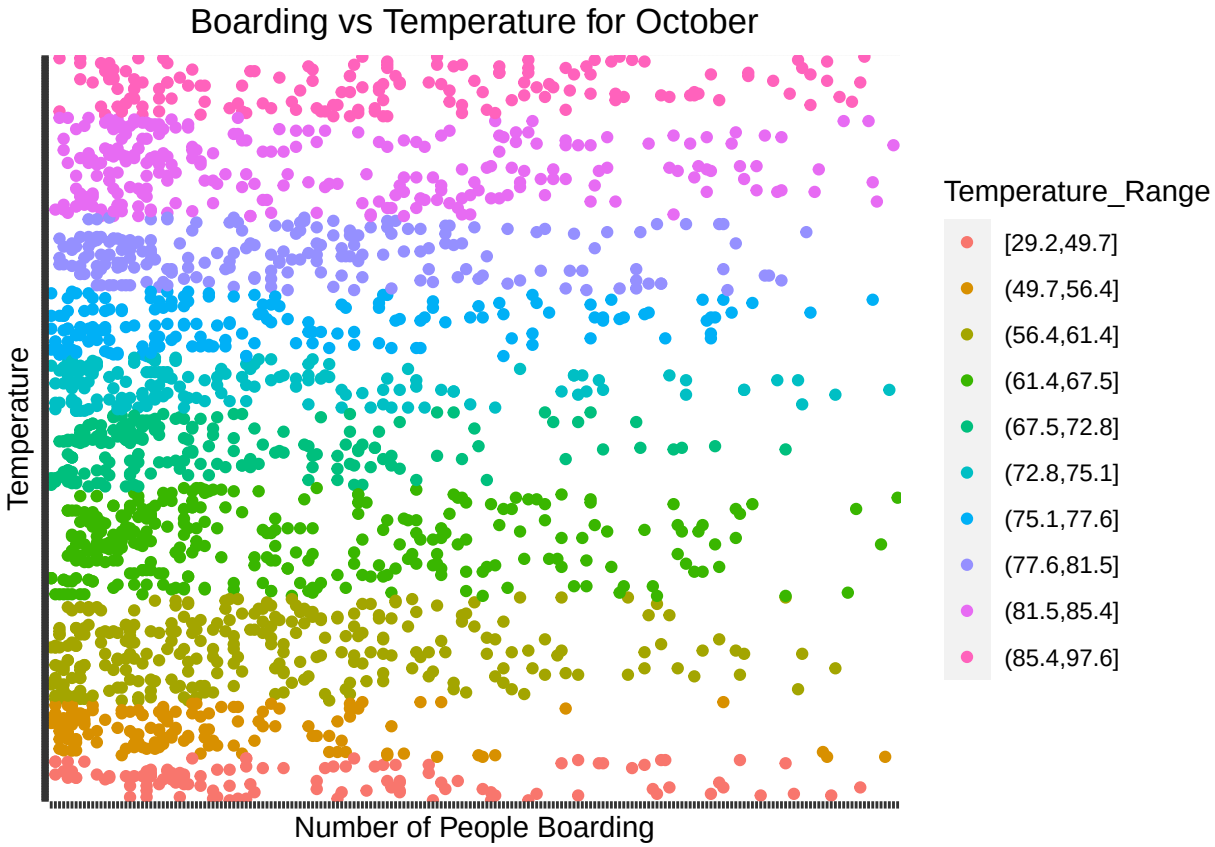
Number of people alighting per 15 minutue interval vs temperature range for the month of October, with color indicated by temperature range bucket the point falls into



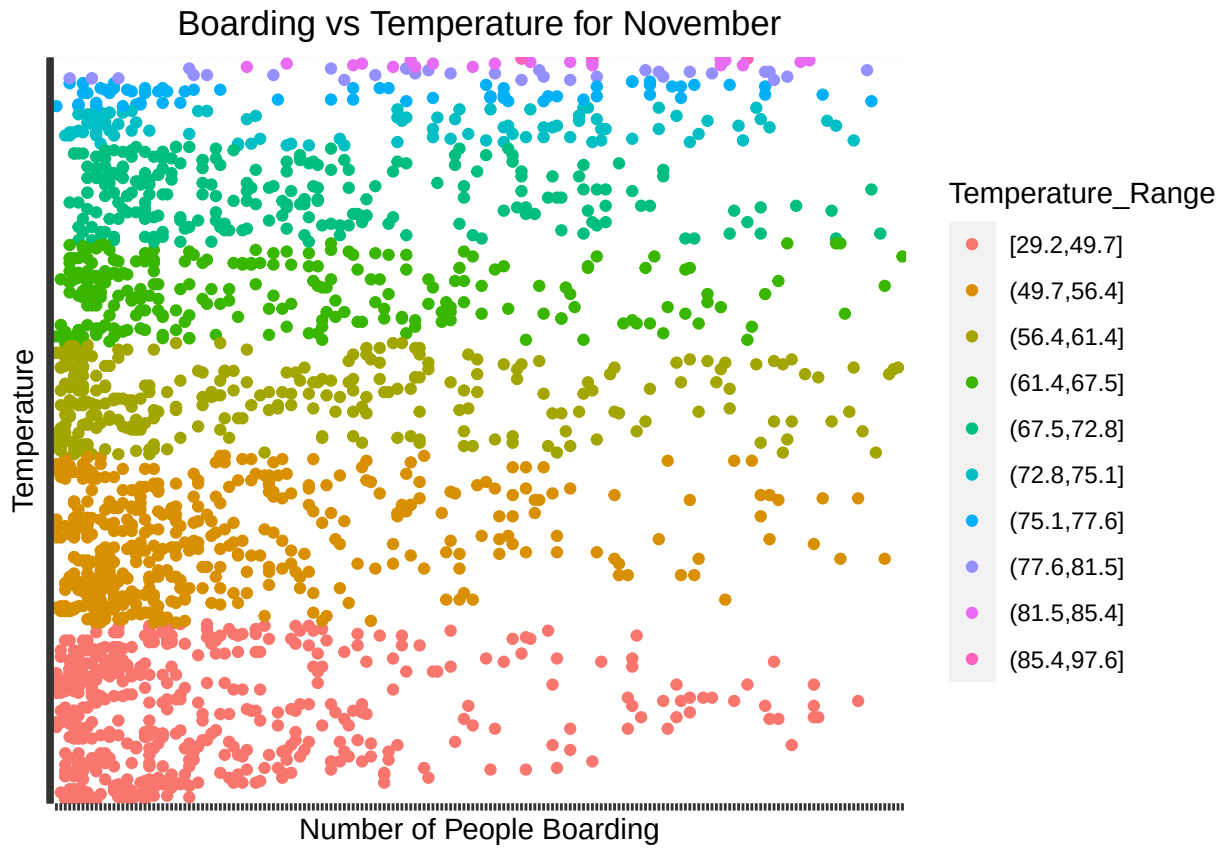
Number of people alighting per 15 minutue interval vs temperature range for the month of November, with color indicated by temperature range bucket the point falls into



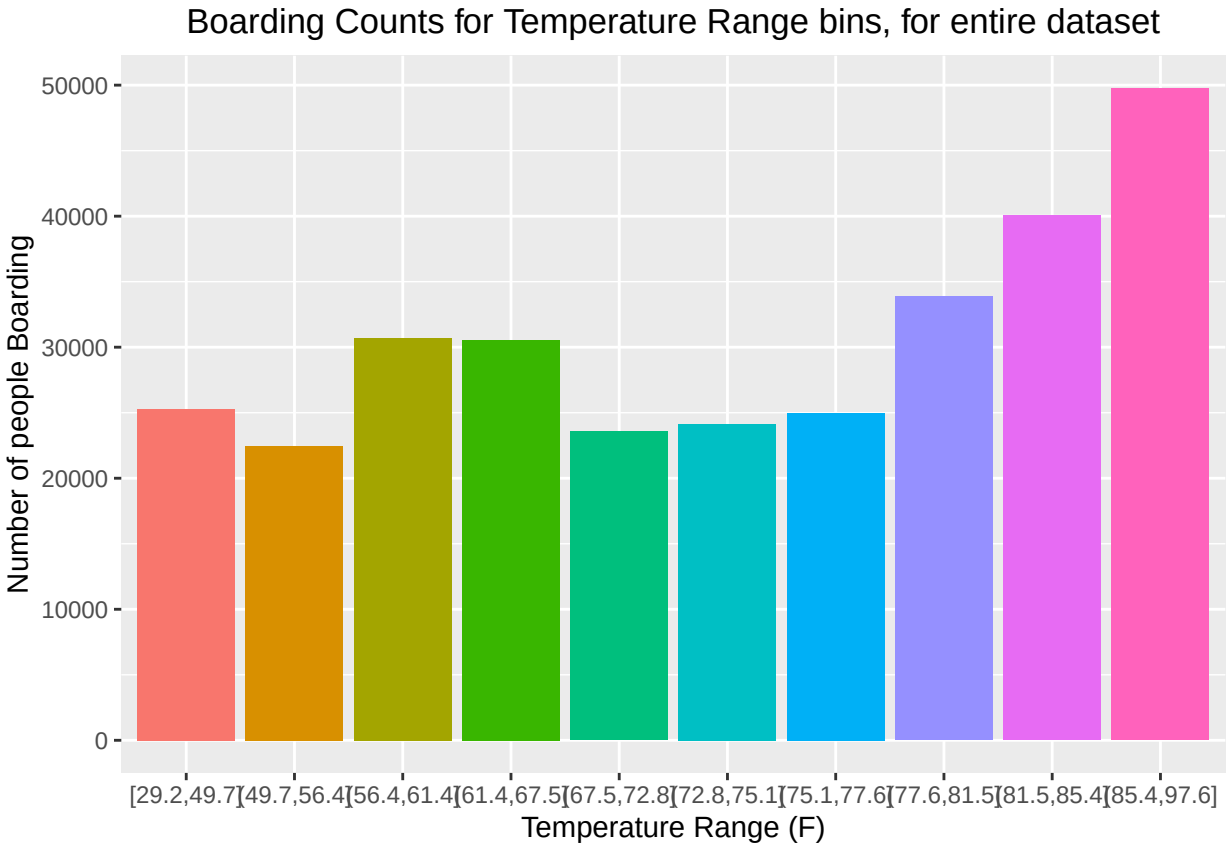
Number of people boarding per 15 minutue interval vs temperature range for the month of September, with color indicated by temperature range bucket the point falls into



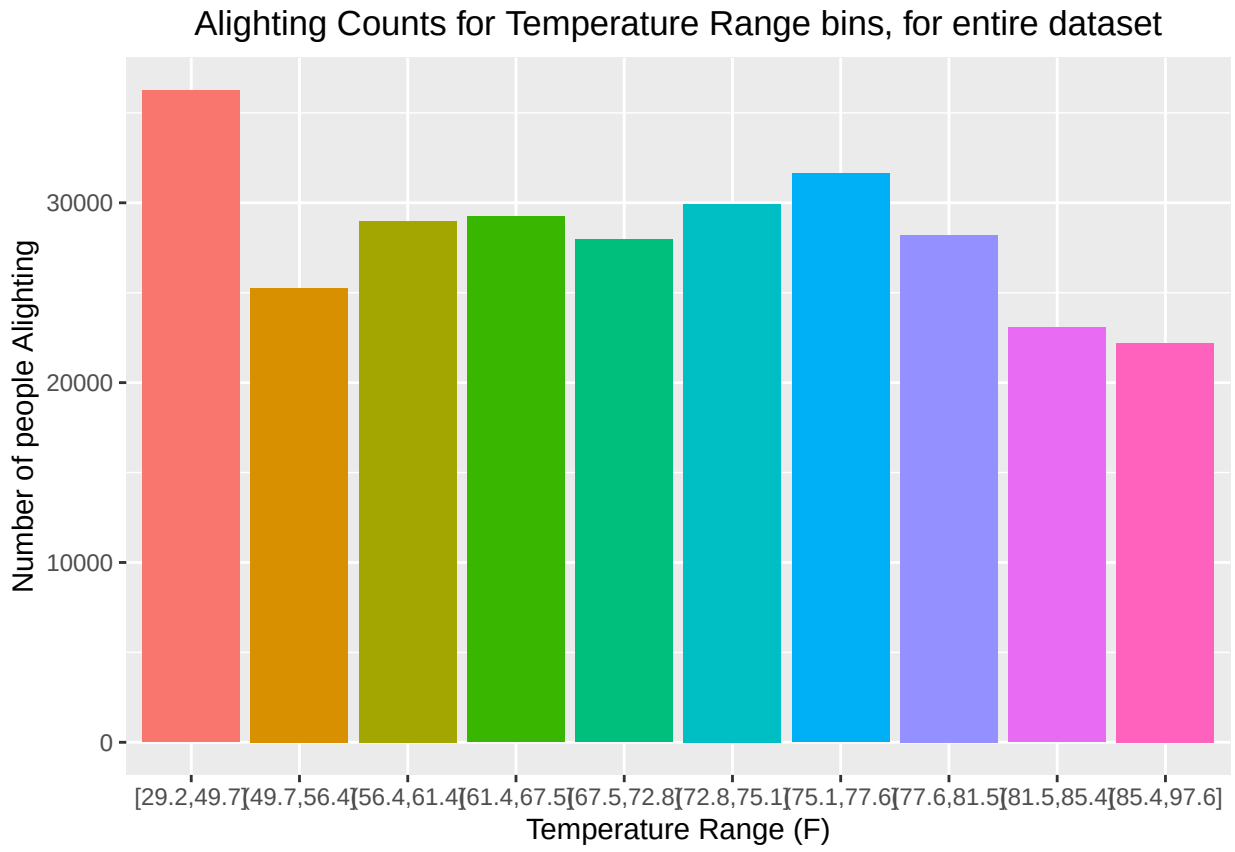
Number of people boarding per 15 minutue interval vs temperature range for the month of October, with color indicated by temperature range bucket the point falls into



Number of people boarding per 15 minutue interval vs temperature range for the month of November, with color indicated by temperature range bucket the point falls into



Bar plot showing aggregate boarding counts for entire timeframe based on temperature range bucket



Bar plot showing aggregate alighting counts for entire timeframe based on temperature range bucket

Portfolio Modeling

```
library(mosaic)
library(quantmod)
```

```
## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
## Loading required package: TTR
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(foreach)
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
mystocks = c( "SPLV", "VDE","USMV")
myprices = getSymbols(mystocks, from = "2017-08-14")
for(ticker in mystocks) {
    expr = paste0(ticker, "a = adjustOHLC(", ticker, ")")
    eval(parse(text=expr))
}
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/SPLV?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/SPLV?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/VDE?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/VDE?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

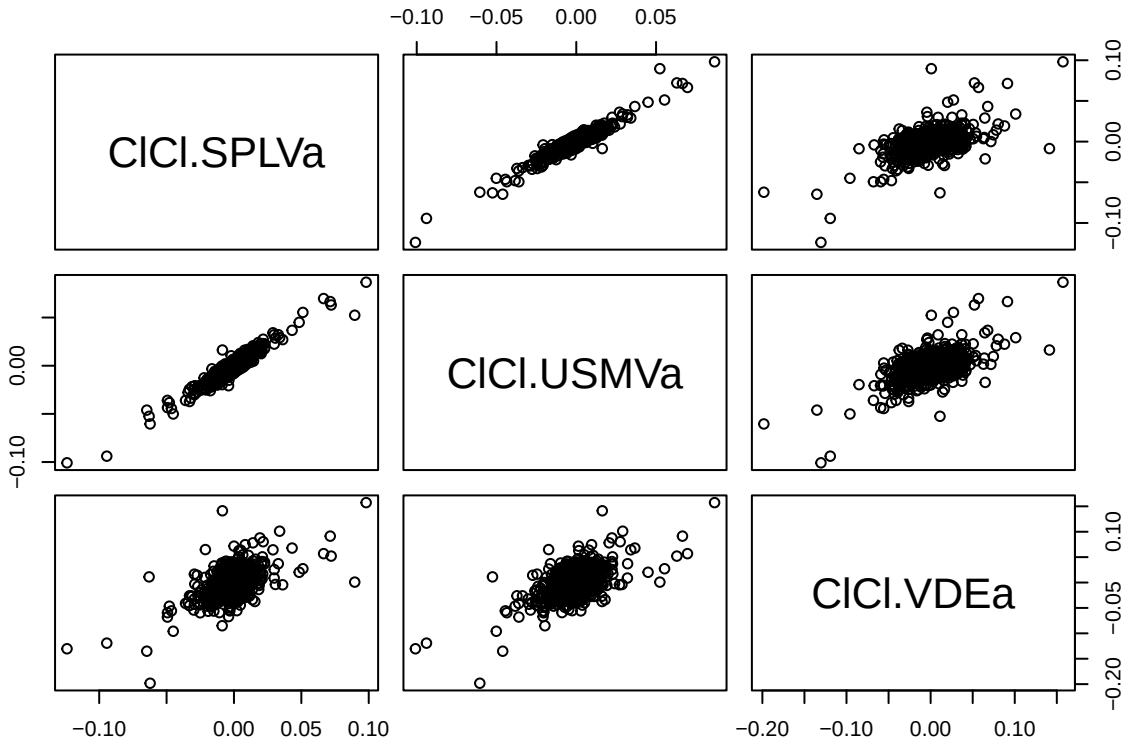
```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/USMV?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/USMV?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
# Combine all the returns in a matrix
all_returns = cbind( ClCl(SPLVa),
                                ClCl(USMVa),
                                ClCl(VDEa))
head(all_returns)
```

```
##              ClCl.SPLVa   ClCl.USMVa    ClCl.VDEa
## 2017-08-14           NA           NA           NA
## 2017-08-15  0.001188245  0.001199560 -0.004277893
## 2017-08-16  0.004175802  0.003394529 -0.009986077
## 2017-08-17 -0.012037645 -0.011343284 -0.014074642
## 2017-08-18 -0.001772220 -0.001811594  0.006542981
## 2017-08-21  0.003106968  0.003831397 -0.006855005
```

```
all_returns = as.matrix(na.omit(all_returns))
# Compute the returns from the closing prices
pairs(all_returns)
```




```
# Sample a random return from the empirical joint distribution
# This simulates a random day
return.today = resample(all_returns, 1, orig.ids=FALSE)
# Update the value of your holdings
# Assumes an equal allocation to each asset
total_wealth = 100000
```

```
my_weights = c(0.25,0.25,0.5)
holdings = total_wealth*my_weights
holdings = holdings*(1 + return.today)
# Compute your new total wealth
holdings
```

```
##            ClCl.SPLVa ClCl.USMVa ClCl.VDEa
## 2021-04-01   25085.98    25191.5  51345.59
```

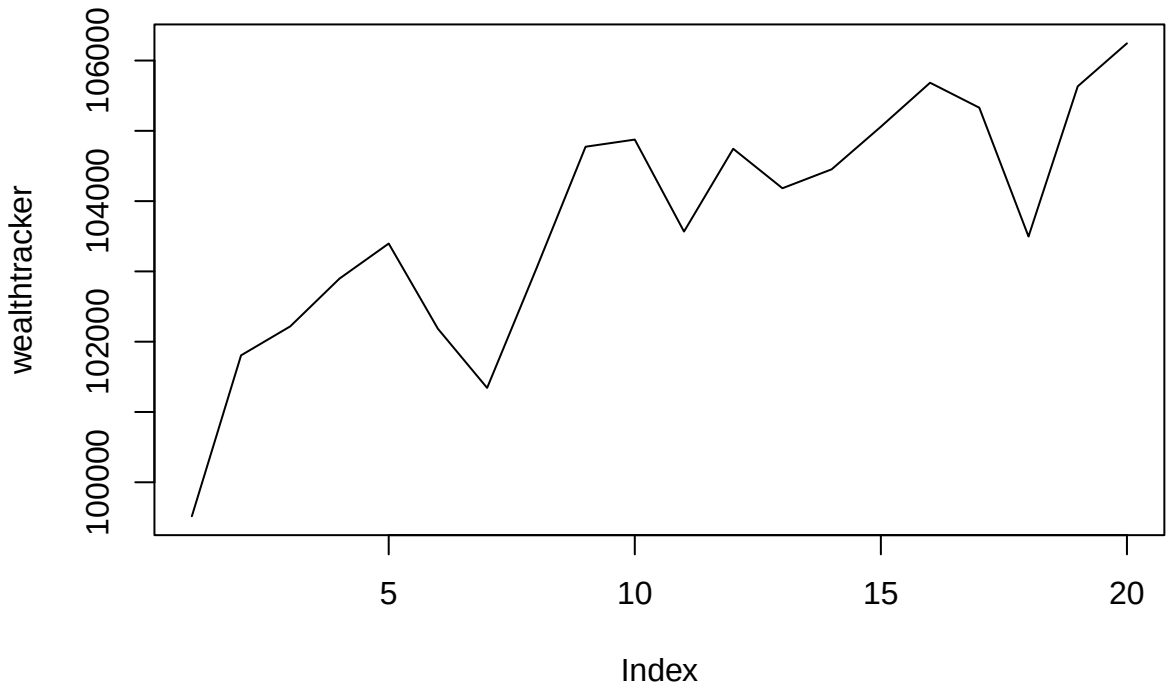
```
total_wealth = sum(holdings)
total_wealth
```

```
## [1] 101623.1
```

```
# Now loop over a 20 day period
## begin block
total_wealth = 100000
weights = c(0.25,0.25,0.5)
holdings = weights * total_wealth
n_days = 20  # capital T in the notes
wealthtracker = rep(0, n_days) # Set up a placeholder to track total wealth
for(today in 1:n_days) {
    return.today = resample(all_returns, 1, orig.ids=FALSE)  # sampling from R matrix in notes
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
}
total_wealth
```

```
## [1] 106244.6
```

```
plot(wealthtracker, type='l')
```

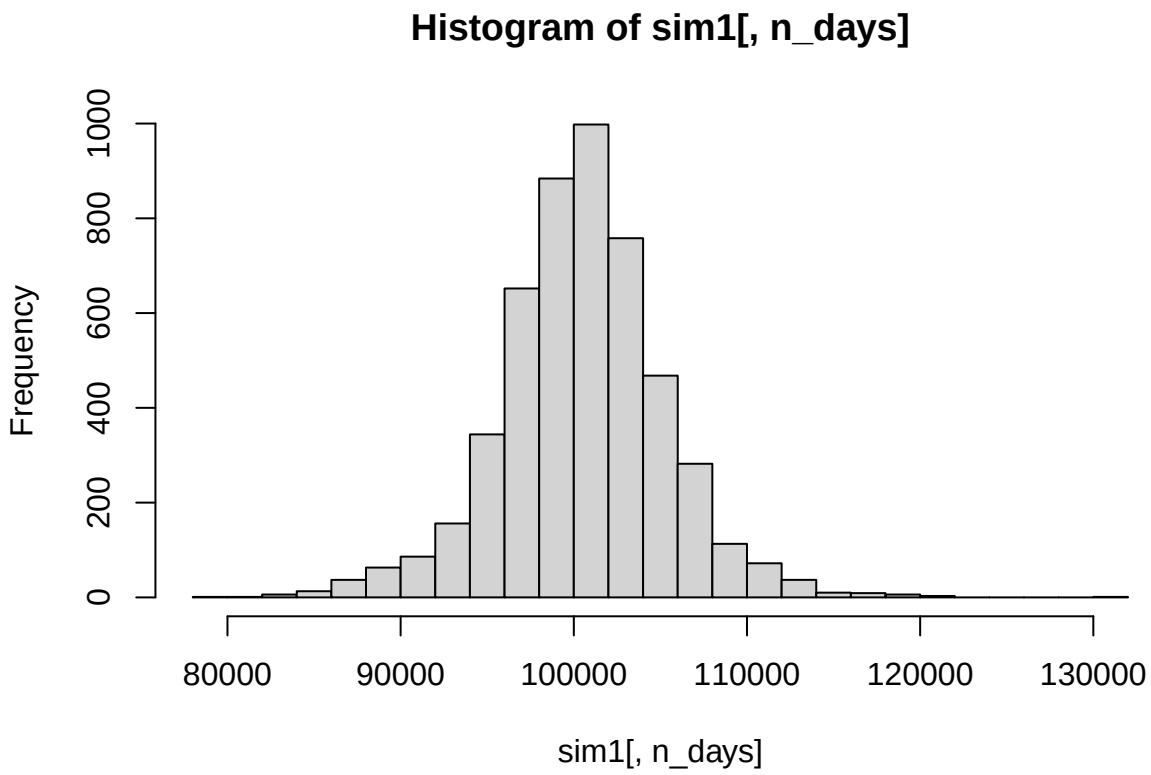



```
## end block
# Now simulate many different possible futures
# just repeating the above block thousands of times
initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
    total_wealth = initial_wealth
    weights = c(0.25,0.25,0.5)
    holdings = weights * total_wealth
    n_days = 10
    wealthtracker = rep(0, n_days)
    for(today in 1:n_days) {
        return.today = resample(all_returns, 1, orig.ids=FALSE)
        holdings = holdings + holdings*return.today
        total_wealth = sum(holdings)
        wealthtracker[today] = total_wealth
    }
    wealthtracker
}
# each row is a simulated trajectory
# each column is a data
head(sim1)
```

```
##              [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## result.1  99989.92 101101.21  99054.19  98571.10  99510.68  99240.56  99539.20
## result.2  99810.80 102072.83 101021.55 100892.79 100436.80 100142.25 101453.23
## result.3 100663.15 101109.36 100076.80 100475.48 103638.35 104773.96 104659.39
```

```
## result.4  97258.79  97697.93  98203.11  99106.05  98346.91  98388.54  98959.39
## result.5  98829.84  99650.17 100716.07 103316.16 104021.10 103761.62 102647.34
## result.6 100012.95  99978.16 101148.82  99323.85  99202.11  99545.06  99674.50
##               [,8]      [,9]     [,10]
## result.1  99422.17  97996.39  97903.66
## result.2 102038.47 101133.23 102206.53
## result.3 105185.02 104960.38 103530.89
## result.4  98841.90  96989.82  97360.37
## result.5 103587.56 105483.76 105188.70
## result.6 100336.34  99313.91  99330.33
```

```
hist(sim1[,n_days], 25)
```



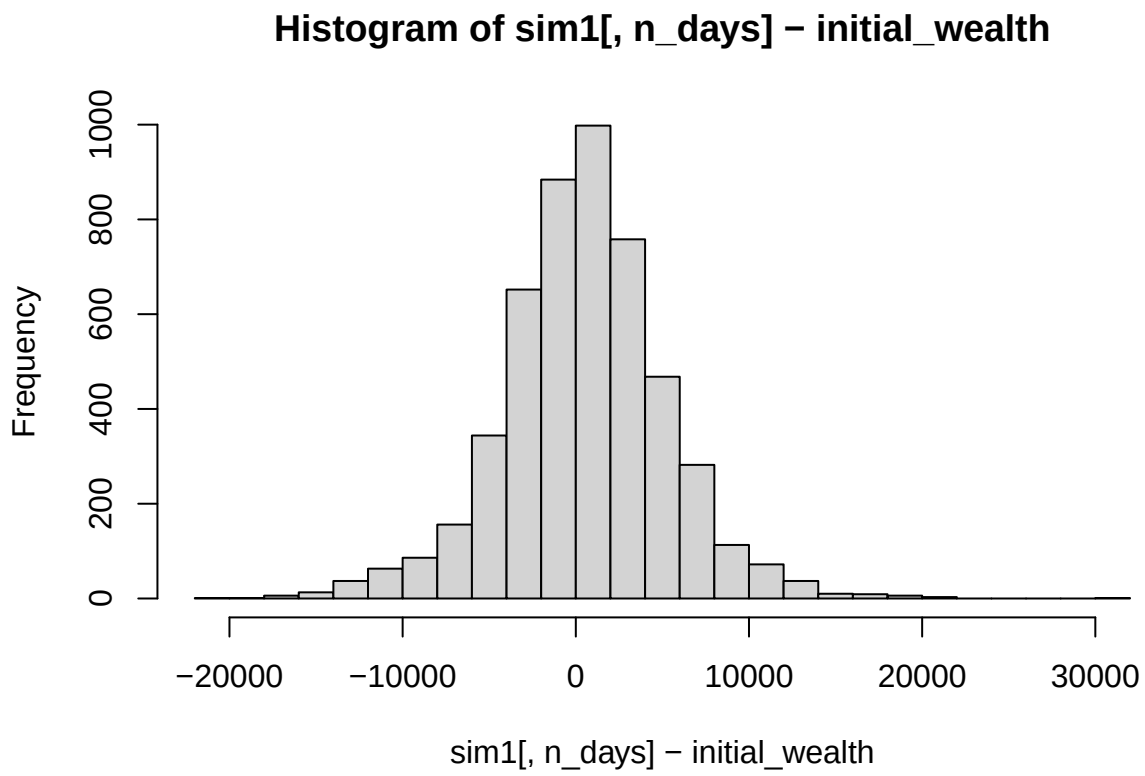

```
# Profit/loss
mean(sim1[,n_days])
```

```
## [1] 100494.4
```

```
mean(sim1[,n_days] - initial_wealth)
```

```
## [1] 494.4477
```

```
hist(sim1[,n_days]- initial_wealth, breaks=30)
```




```
# 5% value at risk:
quantile(sim1[,n_days]- initial_wealth, prob=.05)
```

```
##       5%
## -7238.86
```

In this portfolio, We have 3 ETFS. We half our portfolio USMV which tracks equities that in the past have had a lower risk in the past. We also have 1/4 of our portfolio in an energy ETF VDE. We have the final quarter of our portfolio in SPLV another low volatility index. This is our most conservative option with a very low risk and only a 5% chance of losing 6000 dollars in any 20 day period.

```
mystocks = c( "VOO", "QQQ","TQQQ")
myprices = getSymbols(mystocks, from = "2017-08-14")
for(ticker in mystocks) {
    expr = paste0(ticker, "a = adjustOHLC(", ticker, ")")
    eval(parse(text=expr))
}
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/VOO?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/VOO?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

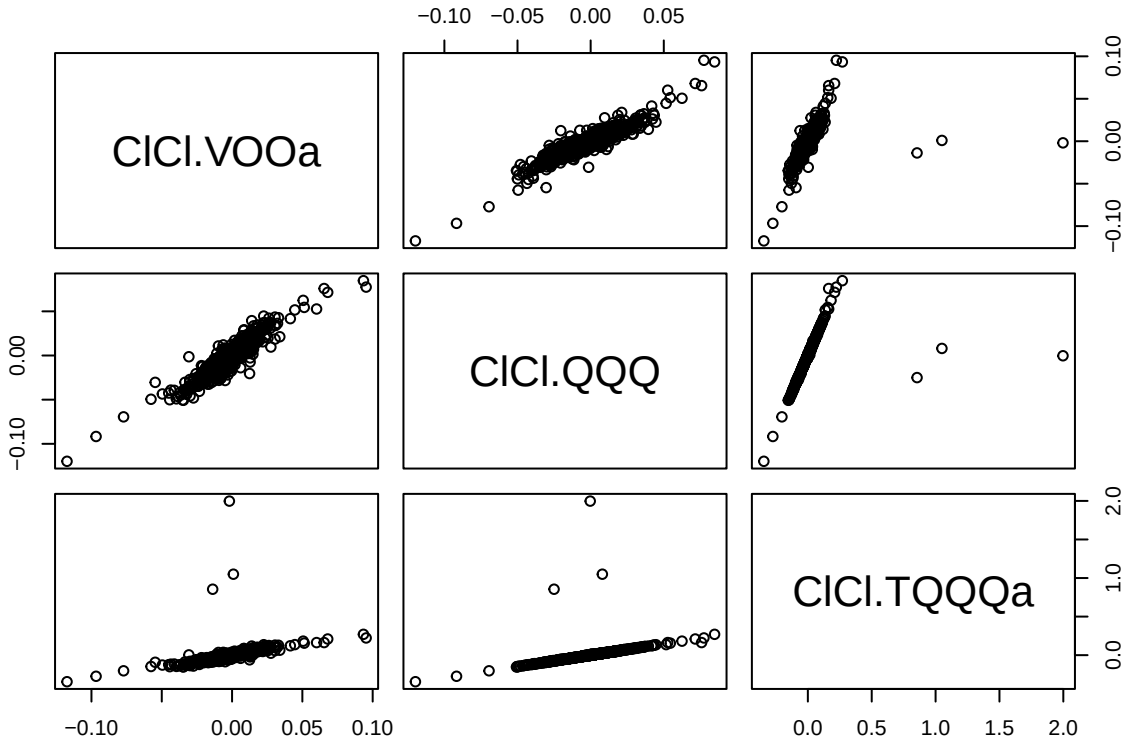
```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/QQQ?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/QQQ?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
# Combine all the returns in a matrix
all_returns = cbind( ClCl(VOOa),
                                ClCl(QQQ),
                                ClCl(TQQQa))
head(all_returns)
```

```
##                ClCl.VOOa      ClCl.QQQ   ClCl.TQQQa
## 2017-08-14            NA            NA           NA
## 2017-08-15  8.834519e-05  0.0006948239  0.001799892
## 2017-08-16  1.589759e-03  0.0017357495  0.005210169
## 2017-08-17 -1.556368e-02 -0.0204463337 -0.061483432
## 2017-08-18 -1.836277e-03 -0.0007076063 -0.003523214
## 2017-08-21  1.211487e-03 -0.0012744672 -0.001815538
```

```
all_returns = as.matrix(na.omit(all_returns))
# Compute the returns from the closing prices
pairs(all_returns)
```




```
# Sample a random return from the empirical joint distribution
# This simulates a random day
return.today = resample(all_returns, 1, orig.ids=FALSE)
# Update the value of your holdings
# Assumes an equal allocation to each asset
total_wealth = 100000
my_weights = c(0.3,0.3,.4)
holdings = total_wealth*my_weights
holdings = holdings*(1 + return.today)
# Compute your new total wealth
holdings
```

```
##             ClCl.VOOa ClCl.QQQ ClCl.TQQQa
## 2022-01-24  30125.16 30137.34   40578.79
```

```
total_wealth = sum(holdings)
total_wealth
```

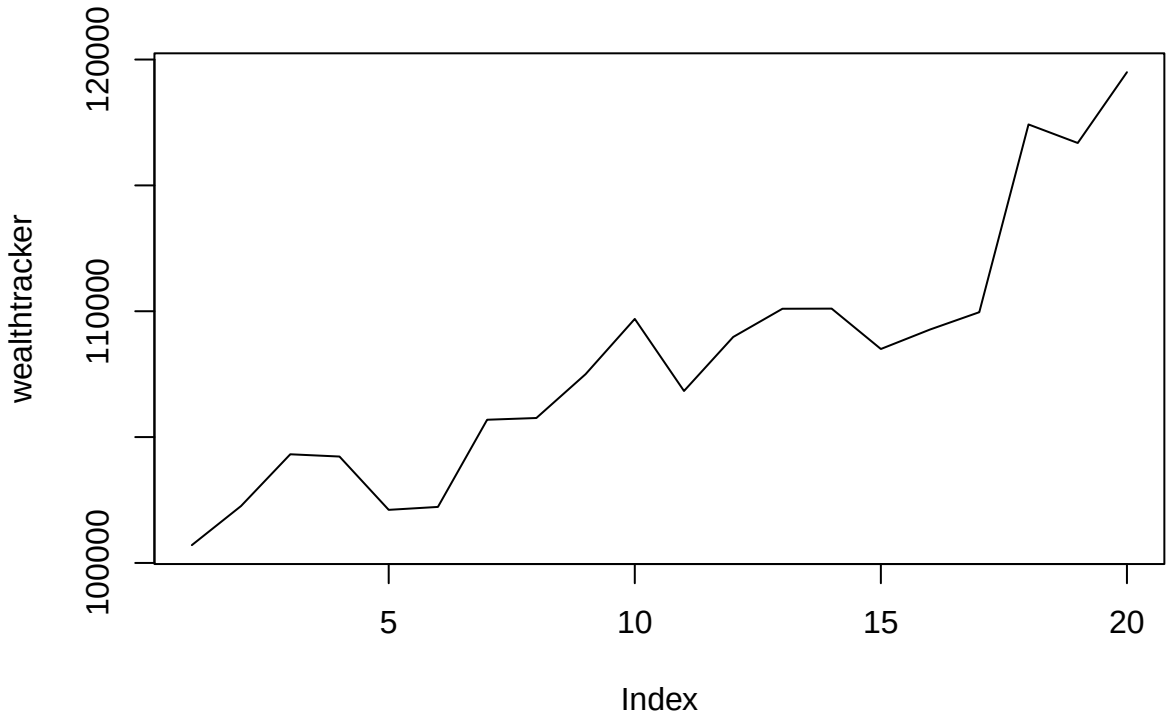
```
## [1] 100841.3
```

```
# Now loop over a 20 day period
## begin block
total_wealth = 100000
weights = c(0.3,0.4,.3)
```

```
holdings = weights * total_wealth
n_days = 20  # capital T in the notes
wealthtracker = rep(0, n_days) # Set up a placeholder to track total wealth
for(today in 1:n_days) {
    return.today = resample(all_returns, 1, orig.ids=FALSE)  # sampling from R matrix in notes
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
}
total_wealth
```

```
## [1] 119495.8
```

```
plot(wealthtracker, type='l')
```

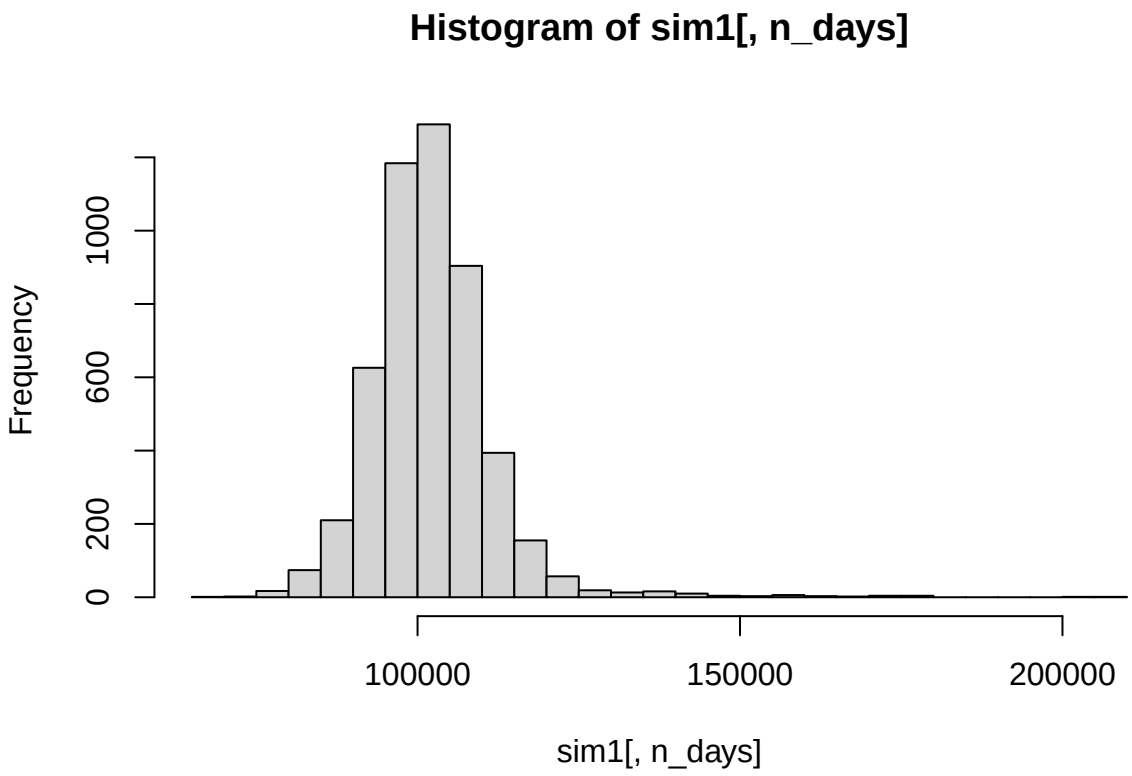



```
## end block
# Now simulate many different possible futures
# just repeating the above block thousands of times
initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
    total_wealth = initial_wealth
    weights =c(0.3,0.4,.3)
    holdings = weights * total_wealth
    n_days = 10
```

```
    wealthtracker = rep(0, n_days)
    for(today in 1:n_days) {
        return.today = resample(all_returns, 1, orig.ids=FALSE)
        holdings = holdings + holdings*return.today
        total_wealth = sum(holdings)
        wealthtracker[today] = total_wealth
    }
    wealthtracker
}
# each row is a simulated trajectory
# each column is a data
head(sim1)
```

```
##               [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## result.1 100658.98 101284.94 102934.46 104036.37 103087.14 108295.28 106872.65
## result.2 102229.94 105285.01 140671.18 140888.76 141468.53 140230.96 132677.73
## result.3 100669.70  97207.34  97192.86  98421.35  99552.89  99200.34 100328.96
## result.4 100270.35 100892.64 100398.61 102252.17 101115.86 100827.53 102217.13
## result.5 100051.32  94602.76  95889.70  96826.81  97984.83  98963.54  98184.82
## result.6  99996.64  99799.11  99856.94 103170.96 103518.45 104536.71 104828.66
##               [,8]     [,9]    [,10]
## result.1 107132.0 108786.6 109807.0
## result.2 132941.9 119990.5 120719.6
## result.3 103064.4 102901.9 103610.0
## result.4 103263.4 105324.6 106705.3
## result.5 100062.8 100338.4 100465.8
## result.6 106521.5 105979.9 108718.8
```

```
hist(sim1[,n_days], 25)
```



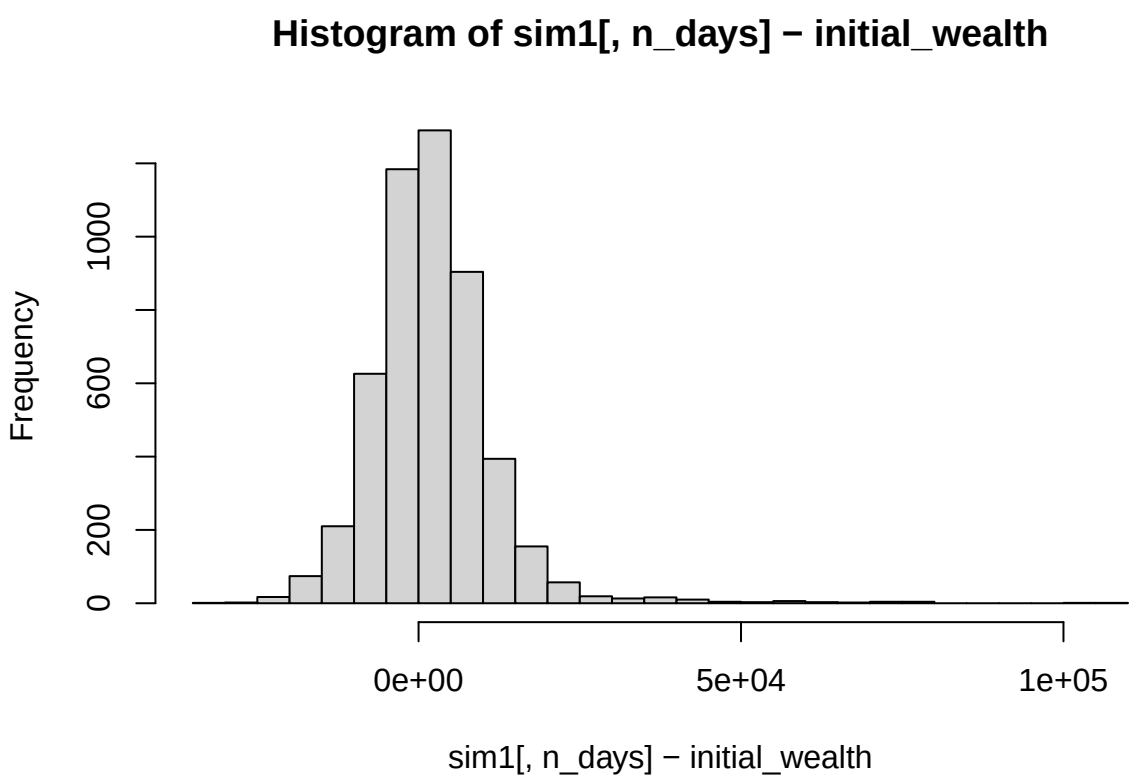

```
# Profit/loss
mean(sim1[,n_days])
```

```
## [1] 102137.1
```

```
mean(sim1[,n_days] - initial_wealth)
```

```
## [1] 2137.111
```

```
hist(sim1[,n_days]- initial_wealth, breaks=30)
```




```
# 5% value at risk:
quantile(sim1[,n_days]- initial_wealth, prob=.05)
```

```
##        5%
## -10829.43
```

In this portfolio, We have 3 ETFS. We 30% our portfolio VOO which tracks The S&P 500. We also have 40% of our portfolio in QQQ which also tracks the S&P 500. We have the 30% of our portfolio in TQQQ, this is a leveraged index of the S&P 500 meaning that all gains and losses are 3x. This is our most Aggressive option as we our betting heavily on companies in the S&P 500 and have a 5% chance of losing 11000 dollars in any 20 day period.

```
mystocks = c( "AOR", "VDE","XLE")
myprices = getSymbols(mystocks, from = "2017-08-14")
for(ticker in mystocks) {
    expr = paste0(ticker, "a = adjustOHLC(", ticker, ")")
    eval(parse(text=expr))
}
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/AOR?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'

## Warning in read.table(file = file, header = header, sep = sep,
```

```
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/AOR?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/VDE?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/VDE?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

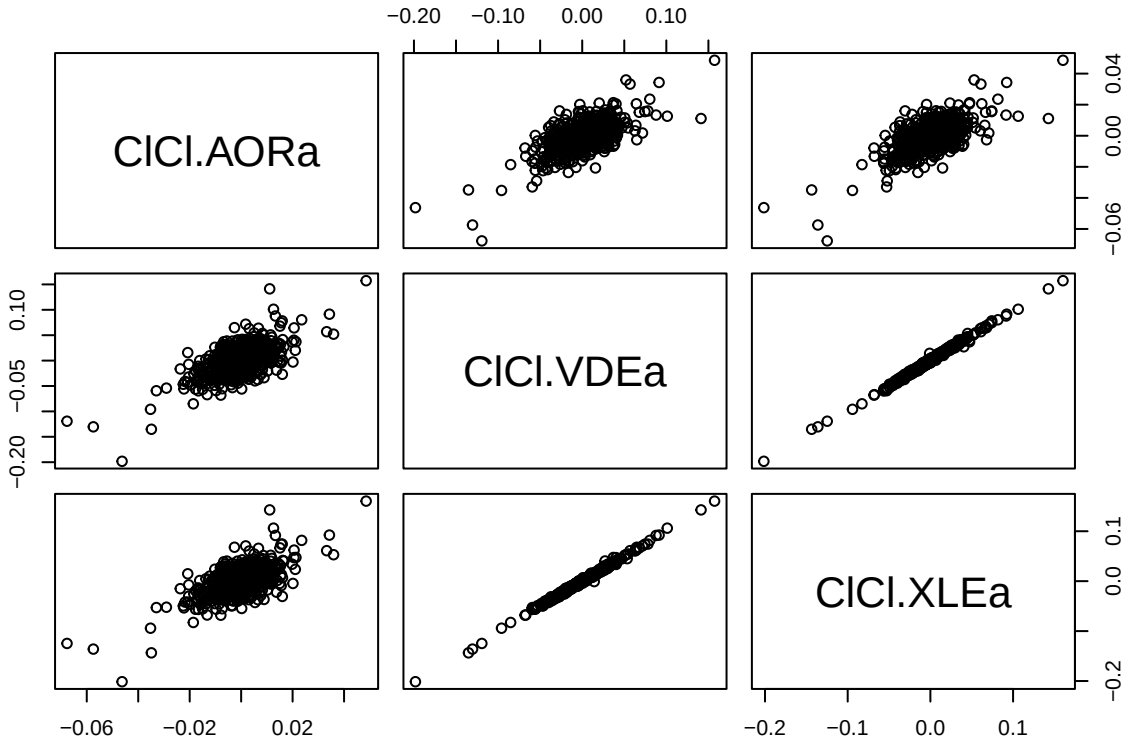
```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/XLE?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/XLE?
## period1=-2208988800&period2=1660521600&interval=1d&events=split'
```

```
# Combine all the returns in a matrix
all_returns = cbind( ClCl(AORa),
                             ClCl(VDEa),
                             ClCl(XLEa))
head(all_returns)
```

```
##                ClCl.AORa    ClCl.VDEa    ClCl.XLEa
## 2017-08-14            NA           NA           NA
## 2017-08-15 -0.0009093203 -0.004277893 -0.003764737
## 2017-08-16  0.0031854152 -0.009986077 -0.009604724
## 2017-08-17 -0.0074847358 -0.014074642 -0.014149507
## 2017-08-18  0.0002285649  0.006542981  0.004999226
## 2017-08-21  0.0006853781 -0.006855005 -0.005134788
```

```
all_returns = as.matrix(na.omit(all_returns))
# Compute the returns from the closing prices
pairs(all_returns)
```




```
# Sample a random return from the empirical joint distribution
# This simulates a random day
return.today = resample(all_returns, 1, orig.ids=FALSE)
# Update the value of your holdings
# Assumes an equal allocation to each asset
total_wealth = 100000
my_weights = c(0.3,0.3,.4)
holdings = total_wealth*my_weights
holdings = holdings*(1 + return.today)
# Compute your new total wealth
holdings
```

```
##            ClCl.AORa ClCl.VDEa ClCl.XLEa
## 2020-01-31  29725.46  29116.81  38752.72
```

```
total_wealth = sum(holdings)
total_wealth
```

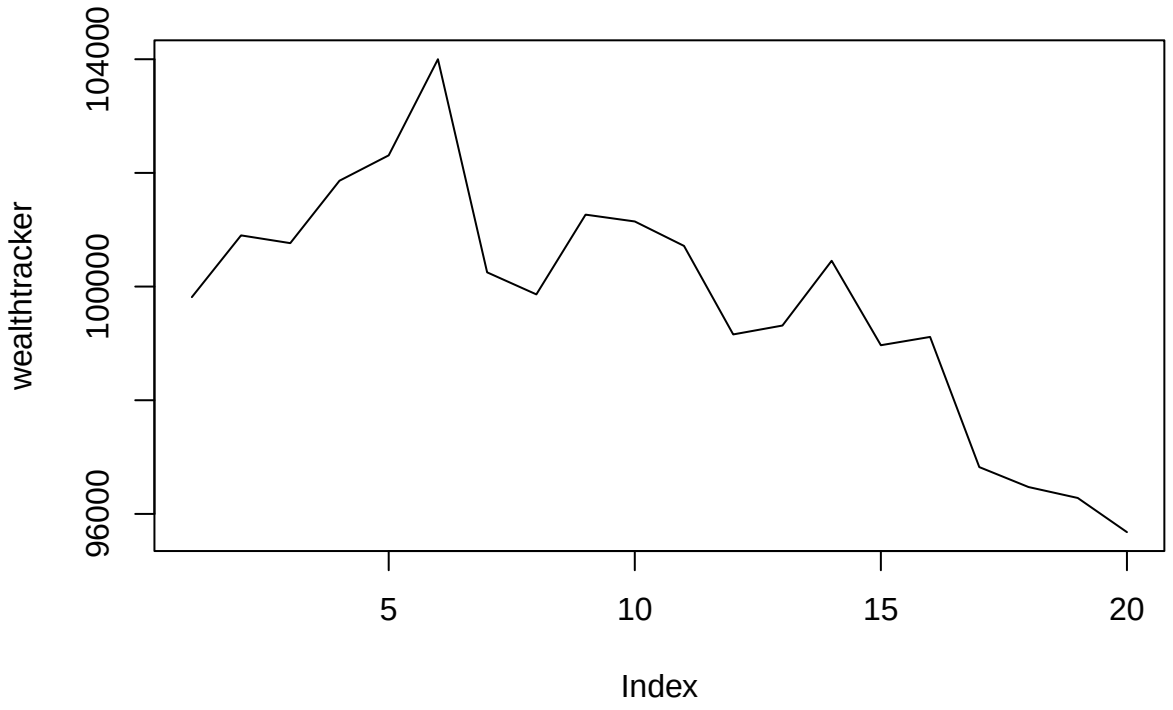
```
## [1] 97594.99
```

```
# Now loop over a 20 day period
## begin block
total_wealth = 100000
weights = c(0.3,0.4,.3)
```

```
holdings = weights * total_wealth
n_days = 20  # capital T in the notes
wealthtracker = rep(0, n_days) # Set up a placeholder to track total wealth
for(today in 1:n_days) {
    return.today = resample(all_returns, 1, orig.ids=FALSE)  # sampling from R matrix in notes
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
}
total_wealth
```

```
## [1] 95679.67
```

```
plot(wealthtracker, type='l')
```

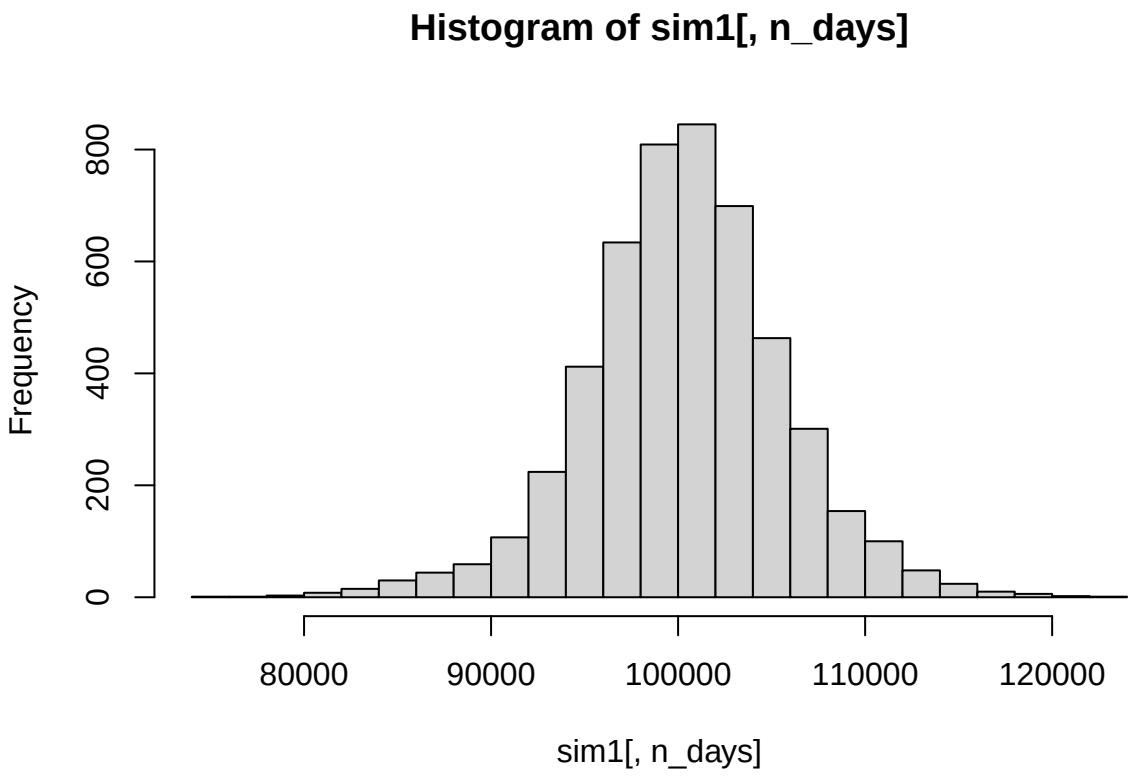



```
## end block
# Now simulate many different possible futures
# just repeating the above block thousands of times
initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
    total_wealth = initial_wealth
    weights =c(0.3,0.4,.3)
    holdings = weights * total_wealth
    n_days = 10
```

```
    wealthtracker = rep(0, n_days)
    for(today in 1:n_days) {
        return.today = resample(all_returns, 1, orig.ids=FALSE)
        holdings = holdings + holdings*return.today
        total_wealth = sum(holdings)
        wealthtracker[today] = total_wealth
    }
    wealthtracker
}
# each row is a simulated trajectory
# each column is a data
head(sim1)
```

```
##               [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## result.1 102503.99 102964.38 102873.05 101859.72 107079.32 107751.05 107347.86
## result.2  98541.91  97371.89  97817.20  97972.62  97990.10  98502.18  98377.78
## result.3  98241.57  98365.65  98652.17  99628.60  99931.80  99624.29 100312.52
## result.4 101829.25 103475.53 101998.04 102309.27 102858.52 102471.27 100710.11
## result.5  99487.32  98560.61  93668.54  93749.43  93956.82  91258.01  89907.54
## result.6  99774.75  99913.26 100761.78 100823.05 102195.93 100360.97  99273.54
##               [,8]      [,9]     [,10]
## result.1 107220.94 109567.71 109544.53
## result.2  99544.86  98984.35  98036.91
## result.3 100919.86 100996.28 101068.69
## result.4 101180.84  98739.18  97575.60
## result.5  90443.65  90300.70  88696.83
## result.6  99871.69  99637.44  99300.01
```

```
hist(sim1[,n_days], 25)
```



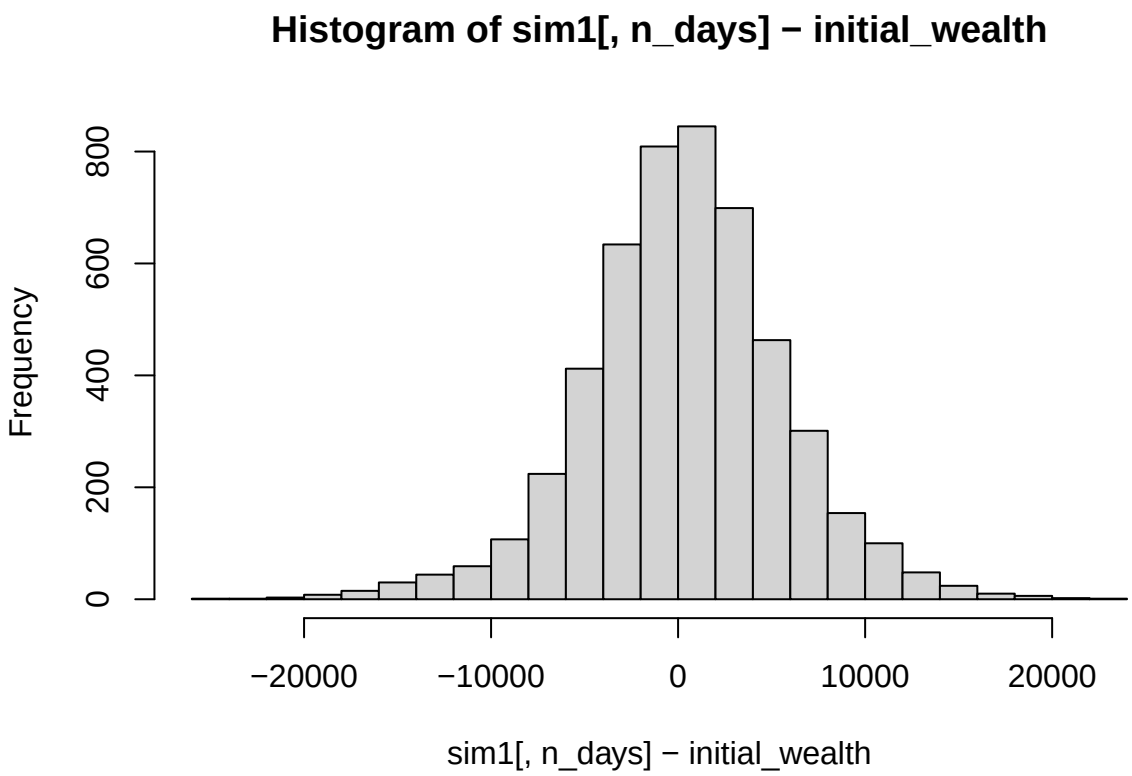

```
# Profit/loss
mean(sim1[,n_days])
```

```
## [1] 100346.1
```

```
mean(sim1[,n_days] - initial_wealth)
```

```
## [1] 346.0942
```

```
hist(sim1[,n_days]- initial_wealth, breaks=30)
```




```
# 5% value at risk:
quantile(sim1[,n_days]- initial_wealth, prob=.05)
```

```
##        5%
## -8215.941
```

In this portfolio, We have 3 ETFS. We 30% our portfolio AOR which tracks equities that are in the growth sector. We also have 40% of our portfolio in an energy ETF VDE. We have the final 40% our portfolio in XLE another energy index. This is our middle of the road option with higher risk as we bet heavily on energy a market that has been steady for 100+ years. We have a 5% chance of losing 8000 dollars in any 20 day period.

#PCA

```
library(readr)
wine <- read_csv("./wine.csv")
```

```
## Rows: 6497 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): color
## dbl (12): fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlo...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ind <- sample(2, nrow(wine),
              replace = TRUE,
              prob = c(0.8, 0.2))
training <- wine[ind==1,]
testing <- wine[ind==2,]
wine
```

```
## # A tibble: 6,497 x 13
##    fixed~1 volat~2 citri~3 resid~4 chlor~5 free.~6 total~7 density    pH sulph~8
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>
##  1     7.4    0.7     0        1.9   0.076      11      34   0.998  3.51    0.56
##  2     7.8    0.88    0        2.6   0.098      25      67   0.997  3.2     0.68
##  3     7.8    0.76    0.04     2.3   0.092      15      54   0.997  3.26    0.65
##  4    11.2    0.28    0.56     1.9   0.075      17      60   0.998  3.16    0.58
##  5     7.4    0.7     0        1.9   0.076      11      34   0.998  3.51    0.56
##  6     7.4    0.66    0        1.8   0.075      13      40   0.998  3.51    0.56
##  7     7.9    0.6     0.06     1.6   0.069      15      59   0.996  3.3     0.46
##  8     7.3    0.65    0        1.2   0.065      15      21   0.995  3.39    0.47
##  9     7.8    0.58    0.02     2     0.073       9      18   0.997  3.36    0.57
## 10     7.5    0.5     0.36     6.1   0.071      17     102   0.998  3.35    0.8
## # ... with 6,487 more rows, 3 more variables: alcohol <dbl>, quality <dbl>,
## #   color <chr>, and abbreviated variable names 1: fixed.acidity,
## #   2: volatile.acidity, 3: citric.acid, 4: residual.sugar, 5: chlorides,
## #   6: free.sulfur.dioxide, 7: total.sulfur.dioxide, 8: sulphates
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
wine.pca <- prcomp(training[,c(1:12)], center = TRUE,scale. = TRUE)
summary(wine.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7470 1.6195 1.2846 1.03565 0.91969 0.81689 0.74878
## Proportion of Variance 0.2543 0.2186 0.1375 0.08938 0.07049 0.05561 0.04672
## Cumulative Proportion  0.2543 0.4729 0.6104 0.69980 0.77029 0.82590 0.87262
##                           PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.71515 0.67728 0.54485 0.47701 0.18443
## Proportion of Variance 0.04262 0.03823 0.02474 0.01896 0.00283
## Cumulative Proportion  0.91524 0.95347 0.97820 0.99717 1.00000
```

```
wine
```

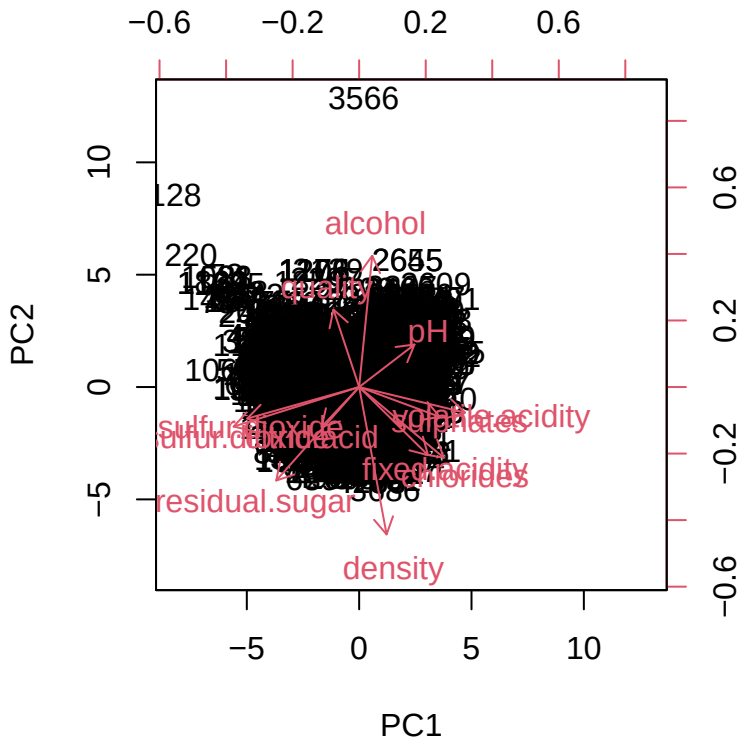
```
## # A tibble: 6,497 x 13
##    fixed~1 volat~2 citri~3 resid~4 chlor~5 free.~6 total~7 density    pH sulph~8
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>
##  1     7.4    0.7     0        1.9   0.076      11      34   0.998  3.51    0.56
##  2     7.8    0.88    0        2.6   0.098      25      67   0.997  3.2     0.68
##  3     7.8    0.76    0.04     2.3   0.092      15      54   0.997  3.26    0.65
##  4    11.2    0.28    0.56     1.9   0.075      17      60   0.998  3.16    0.58
##  5     7.4    0.7     0        1.9   0.076      11      34   0.998  3.51    0.56
##  6     7.4    0.66    0        1.8   0.075      13      40   0.998  3.51    0.56
##  7     7.9    0.6     0.06     1.6   0.069      15      59   0.996  3.3     0.46
##  8     7.3    0.65    0        1.2   0.065      15      21   0.995  3.39    0.47
```

```
##  9     7.8    0.58    0.02     2     0.073       9      18   0.997  3.36    0.57
## 10     7.5    0.5     0.36     6.1   0.071      17     102   0.998  3.35    0.8
## # ... with 6,487 more rows, 3 more variables: alcohol <dbl>, quality <dbl>,
## #   color <chr>, and abbreviated variable names 1: fixed.acidity,
## #   2: volatile.acidity, 3: citric.acid, 4: residual.sugar, 5: chlorides,
## #   6: free.sulfur.dioxide, 7: total.sulfur.dioxide, 8: sulphates
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
wine.pca$rotation <- -1*wine.pca$rotation
wine.pca$rotation
```

```
##                              PC1         PC2         PC3         PC4
## fixed.acidity         0.25854743 -0.25604160  0.46973005  0.14452485
## volatile.acidity      0.39784702 -0.09612355 -0.27889865  0.06692591
## citric.acid          -0.14127957 -0.14965420  0.58887446 -0.03779968
## residual.sugar       -0.31163862 -0.35131657 -0.07914015 -0.13407358
## chlorides             0.31822693 -0.26560178  0.04988896 -0.15703823
## free.sulfur.dioxide  -0.42068745 -0.11850717 -0.09437436 -0.30053088
## total.sulfur.dioxide -0.47163345 -0.14871437 -0.09954933 -0.12434572
## density               0.10306348 -0.55368269 -0.05303737 -0.16120822
## pH                    0.20800628  0.15980983 -0.39773029 -0.46571582
## sulphates             0.30245175 -0.11185605  0.17171278 -0.58045295
## alcohol               0.04894837  0.49325238  0.21785515 -0.08728097
## quality              -0.09667632  0.29114537  0.29879581 -0.48720054
##                              PC5          PC6         PC7         PC8
## fixed.acidity         0.17568172 -0.047289117 -0.37924225  0.11296039
## volatile.acidity      0.13133952  0.369249284 -0.51378709 -0.21395580
## citric.acid          -0.21070184 -0.373136233 -0.12024120 -0.44382892
## residual.sugar        0.49481001  0.106171041  0.06145865 -0.11988145
## chlorides            -0.40754066  0.423587184  0.36221651 -0.47569343
## free.sulfur.dioxide  -0.26384770  0.242971392 -0.38725298 -0.02388127
## total.sulfur.dioxide -0.24099953  0.074704615 -0.23871840  0.02534886
## density               0.33675253 -0.144584500 -0.01658111 -0.02290394
## pH                    0.01894413 -0.576259215 -0.12853167 -0.33286294
## sulphates            -0.23293048  0.008146656 -0.04769791  0.58991060
## alcohol               0.11282573  0.174691173 -0.39391418 -0.17498459
## quality               0.43372858  0.288636544  0.24813724 -0.11257719
##                              PC9         PC10         PC11         PC12
## fixed.acidity        -0.416842155  0.26153017 -0.288687825  0.332988055
## volatile.acidity      0.080145412 -0.49707307  0.162373609  0.083636113
## citric.acid           0.221586137 -0.31968355  0.251143372 -0.001676265
## residual.sugar        0.484541061  0.19617862 -0.003585357  0.455721674
## chlorides             0.003287997  0.23494847 -0.203250296  0.038917675
## free.sulfur.dioxide  -0.319730517  0.32261738  0.473547113 -0.003232659
## total.sulfur.dioxide -0.008602939 -0.32493851 -0.706379665 -0.064436963
## density              -0.062389224  0.07292282 -0.009459786 -0.713867097
## pH                   -0.144835563  0.11466136 -0.140211891  0.206344274
## sulphates             0.329188338 -0.08894165  0.046731381  0.080921414
## alcohol               0.404060876  0.39536656 -0.201580357 -0.331629492
## quality              -0.370306009 -0.31028795 -0.010544980 -0.007152425
```

```
biplot(wine.pca, scale = 0)
```




```
trg <- predict(wine.pca, training)
trg <- data.frame(trg, training[13])
tst <- predict(wine.pca, testing)
tst <- data.frame(tst, testing[13])
library(nnet)
mymodel <- multinom(color~PC1+PC2, data = trg)
```

```
## # weights:  4 (3 variable)
## initial  value 3607.831075
## iter  10 value 353.769489
## final  value 314.899930
## converged
```

```
summary(mymodel)
```

```
## Call:
## multinom(formula = color ~ PC1 + PC2, data = trg)
##
## Coefficients:
##                 Values  Std. Err.
## (Intercept)  4.5436196 0.22336181
## PC1         -3.9280859 0.17689396
## PC2          0.3167584 0.07172836
##
```

```
## Residual Deviance: 629.7999
## AIC: 635.7999
```

```
p <- predict(mymodel, trg)
tab <- table(p, trg$color)
tab
```

```
##
## p        red white
##   red   1261    41
##   white   42  3861
```

```
p1 <- predict(mymodel, tst)
tab1 <- table(p1, tst$color)
tab1
```

```
##
## p1      red white
##   red   283    13
##   white  13   983
```

```
# 1 % misclassification rate
```

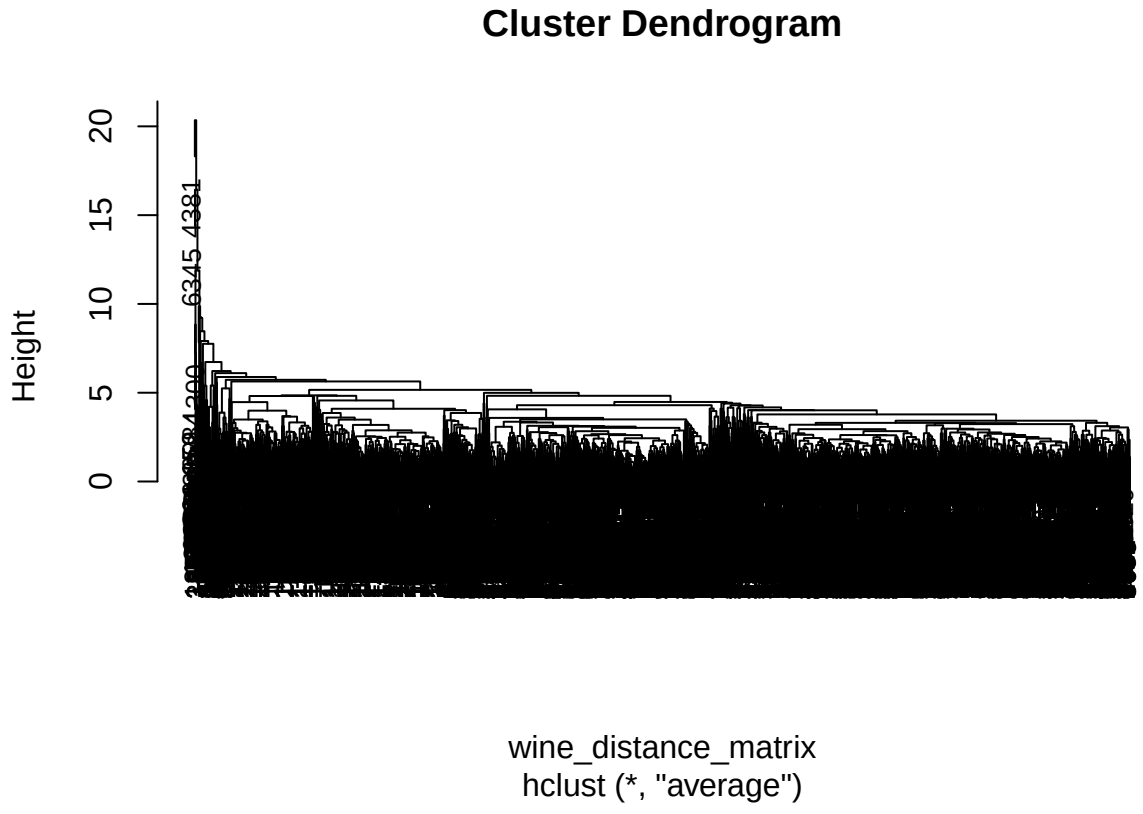
#clustering

```
wine_2 = wine[-13]
# Center/scale the data
wine_2 %>% na.omit(wine_2)
```

```
## # A tibble: 6,497 x 12
##    fixed~1 volat~2 citri~3 resid~4 chlor~5 free.~6 total~7 density    pH sulph~8
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>
##  1     7.4    0.7     0        1.9   0.076      11      34   0.998  3.51    0.56
##  2     7.8    0.88    0        2.6   0.098      25      67   0.997  3.2     0.68
##  3     7.8    0.76    0.04     2.3   0.092      15      54   0.997  3.26    0.65
##  4    11.2    0.28    0.56     1.9   0.075      17      60   0.998  3.16    0.58
##  5     7.4    0.7     0        1.9   0.076      11      34   0.998  3.51    0.56
##  6     7.4    0.66    0        1.8   0.075      13      40   0.998  3.51    0.56
##  7     7.9    0.6     0.06     1.6   0.069      15      59   0.996  3.3     0.46
##  8     7.3    0.65    0        1.2   0.065      15      21   0.995  3.39    0.47
##  9     7.8    0.58    0.02     2     0.073       9      18   0.997  3.36    0.57
## 10     7.5    0.5     0.36     6.1   0.071      17     102   0.998  3.35    0.8
## # ... with 6,487 more rows, 2 more variables: alcohol <dbl>, quality <dbl>, and
## #   abbreviated variable names 1: fixed.acidity, 2: volatile.acidity,
## #   3: citric.acid, 4: residual.sugar, 5: chlorides, 6: free.sulfur.dioxide,
## #   7: total.sulfur.dioxide, 8: sulphates
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
wine_scaled = scale(wine_2, center=TRUE, scale=TRUE)
# Form a pairwise distance matrix using the dist function
wine_distance_matrix = dist(wine_scaled, method='euclidean')
```

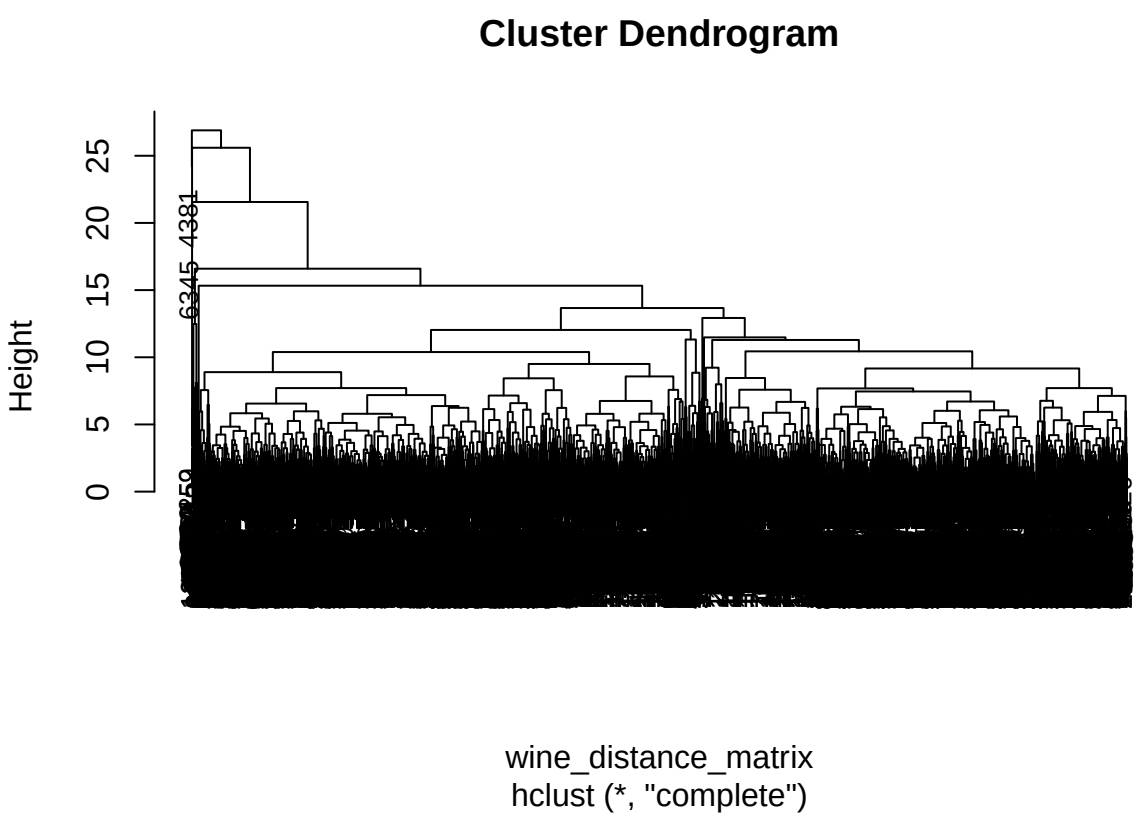
```r
# Now run hierarchical clustering
hier_wine = hclust(wine_distance_matrix, method='average')
# Plot the dendrogram
plot(hier_wine, cex=0.8)
```




```r
# Cut the tree into 5 clusters
cluster1 = cutree(hier_wine, k=2)
summary(factor(cluster1))
```

```
##    1    2
## 6496    1
```

```r
# Using max ("complete") linkage instead
hier_protein2 = hclust(wine_distance_matrix, method='complete')
# Plot the dendrogram
plot(hier_protein2, cex=0.8)
```




```
cluster2 = cutree(hier_protein2, k=5)
summary(factor(cluster2))
```

```
##    1    2    3    4    5
## 6450   43    2    1    1
```

After running both PCA and Clustering, it was aparent that PCA was the much more effective model. The heirarchal clustering values were not good This is because the values of the red and white wine were so close together that there simply was not a large enough difference in the data to effectively tell which wine was which However, in the PCA we ran could not have gone much better we were able to take the 13 variables down to 2 variables which we were able to plot against each other. This gave us a confusion matrix with 99% accuracy meaning that only a few wines were incorrectly placed in the wrong category. PCA was able to take these differences and reduce the number of variables allowing us to make us to make a very strong prediction about the type of wine.

#Market Segmentation

###Pre-Processing and First Correlation plot

```
library(readr)
soc <- read_csv("./social_marketing.csv")
```

```
## New names:
## Rows: 7882 Columns: 37
## -- Column specification
```

```
## ------------------------------------------------------- Delimiter: "," chr
## (1): ...1 dbl (36): chatter, current_events, travel, photo_sharing,
## uncategorized, tv_...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

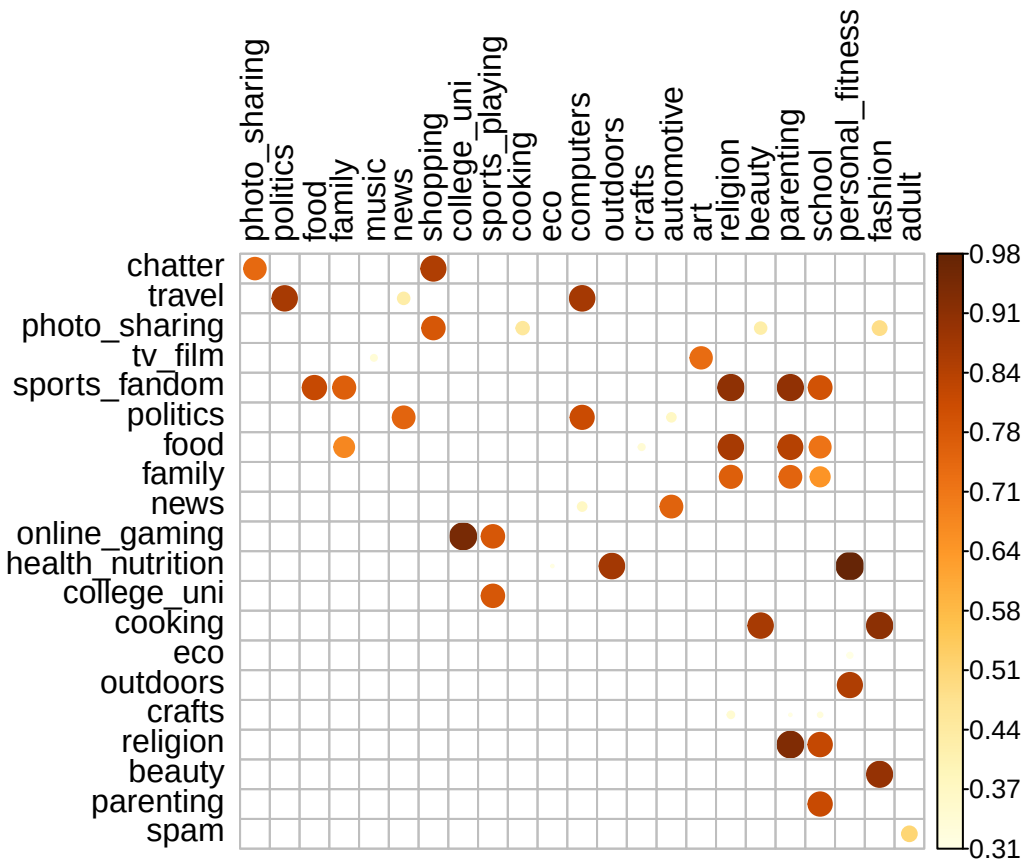
```
soc <-soc[-1]
d2 <- soc %>%
  as.matrix %>%
  cor %>%
  as.data.frame
mydata.cor = cor(soc, method = c("spearman"))
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corr_simple <- function(data=d2,sig=0.3){
  #convert data to numeric in order to run correlations
  #convert to factor first to keep the integrity of the data - each value will become a number rather t
  df_cor <- data %>% mutate_if(is.character, as.factor)
  df_cor <- df_cor %>% mutate_if(is.factor, as.numeric)
  #run a correlation and drop the insignificant ones
  corr <- cor(df_cor)
  #prepare to drop duplicates and correlations of 1
  corr[lower.tri(corr,diag=TRUE)] <- NA
  #drop perfect correlations
  corr[corr == 1] <- NA
  #turn into a 3-column table
  corr <- as.data.frame(as.table(corr))
  #remove the NA values from above
  corr <- na.omit(corr)
  #select significant values
  corr <- subset(corr, abs(Freq) > sig)
  #sort by highest correlation
  corr <- corr[order(-abs(corr$Freq)),]
  #print table
  print(corr)
  #turn corr back into matrix in order to plot with corrplot
  mtx_corr <- reshape2::acast(corr, Var1~Var2, value.var="Freq")

  #plot correlations visually
  corrplot(mtx_corr, is.corr=FALSE, tl.col="black", na.label=" ")
}
corr_simple()
```

```
##                  Var1             Var2      Freq
## 1132 health_nutrition personal_fitness 0.9784190
## 590     online_gaming      college_uni 0.9477743
## 1035         religion        parenting 0.9329145
## 1171          cooking          fashion 0.9126470
## 943     sports_fandom         religion 0.9110081
```

```
## 1015    sports_fandom        parenting 0.9052349
## 1180           beauty          fashion 0.9003331
## 808  health_nutrition         outdoors 0.8767125
## 723            travel        computers 0.8743096
## 991           cooking           beauty 0.8708852
## 945              food         religion 0.8702392
## 255            travel         politics 0.8686486
## 1139         outdoors personal_fitness 0.8569863
## 505           chatter         shopping 0.8558037
## 1017             food        parenting 0.8461772
## 1107         religion           school 0.8221967
## 295     sports_fandom             food 0.8200116
## 1109        parenting           school 0.8172017
## 728           politics        computers 0.8113277
## 1087    sports_fandom           school 0.7968535
## 629       college_uni   sports_playing 0.7823720
## 626     online_gaming   sports_playing 0.7821580
## 508     photo_sharing         shopping 0.7808589
## 946            family         religion 0.7667968
## 331     sports_fandom           family 0.7661486
## 877              news       automotive 0.7547770
## 1018           family        parenting 0.7530952
## 440           politics             news 0.7515664
## 109            chatter    photo_sharing 0.7416774
## 906            tv_film              art 0.7360172
## 1089             food           school 0.7185102
## 333              food           family 0.6775890
## 1090           family           school 0.6507692
## 1295             spam            adult 0.5097285
## 1156    photo_sharing          fashion 0.4862129
## 652     photo_sharing          cooking 0.4524290
## 435            travel             news 0.4320153
## 976     photo_sharing           beauty 0.4310986
## 733              news        computers 0.3772670
## 872           politics       automotive 0.3763098
## 960             crafts         religion 0.3492083
## 837              food           crafts 0.3422739
## 402           tv_film            music 0.3400537
## 1104           crafts           school 0.3270665
## 1032           crafts        parenting 0.3135940
## 700  health_nutrition              eco 0.3135571
## 1136              eco personal_fitness 0.3072261
```

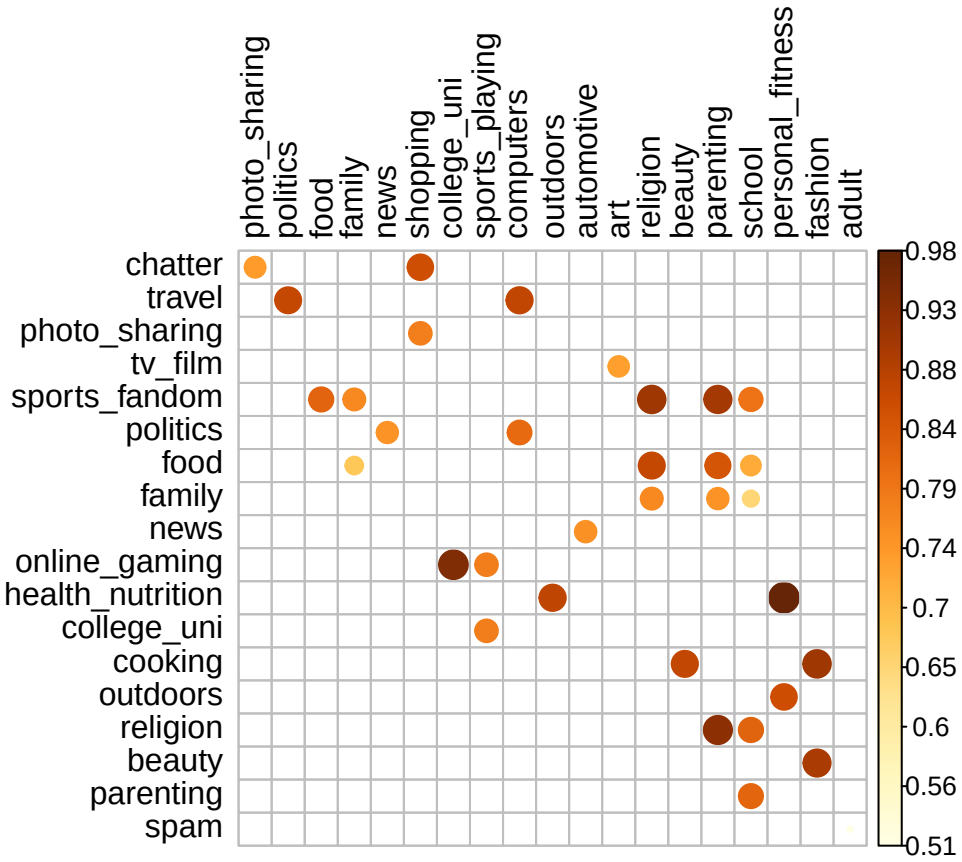


Second and third correlation plots along with data frames

```
corr_simple2 <- function(data=d2,sig=0.5){
  #convert data to numeric in order to run correlations
  #convert to factor first to keep the integrity of the data - each value will become a number rather t
  df_cor <- data %>% mutate_if(is.character, as.factor)
  df_cor <- df_cor %>% mutate_if(is.factor, as.numeric)
  #run a correlation and drop the insignificant ones
  corr <- cor(df_cor)
  #prepare to drop duplicates and correlations of 1
  corr[lower.tri(corr,diag=TRUE)] <- NA
  #drop perfect correlations
  corr[corr == 1] <- NA
  #turn into a 3-column table
  corr <- as.data.frame(as.table(corr))
  #remove the NA values from above
  corr <- na.omit(corr)
  #select significant values
  corr <- subset(corr, abs(Freq) > sig)
  #sort by highest correlation
  corr <- corr[order(-abs(corr$Freq)),]
  #print table
  print(corr)
  #turn corr back into matrix in order to plot with corrplot
```

```
  mtx_corr <- reshape2::acast(corr, Var1~Var2, value.var="Freq")

  #plot correlations visually
  corrplot(mtx_corr, is.corr=FALSE, tl.col="black", na.label=" ")
}
corr_simple2()
```

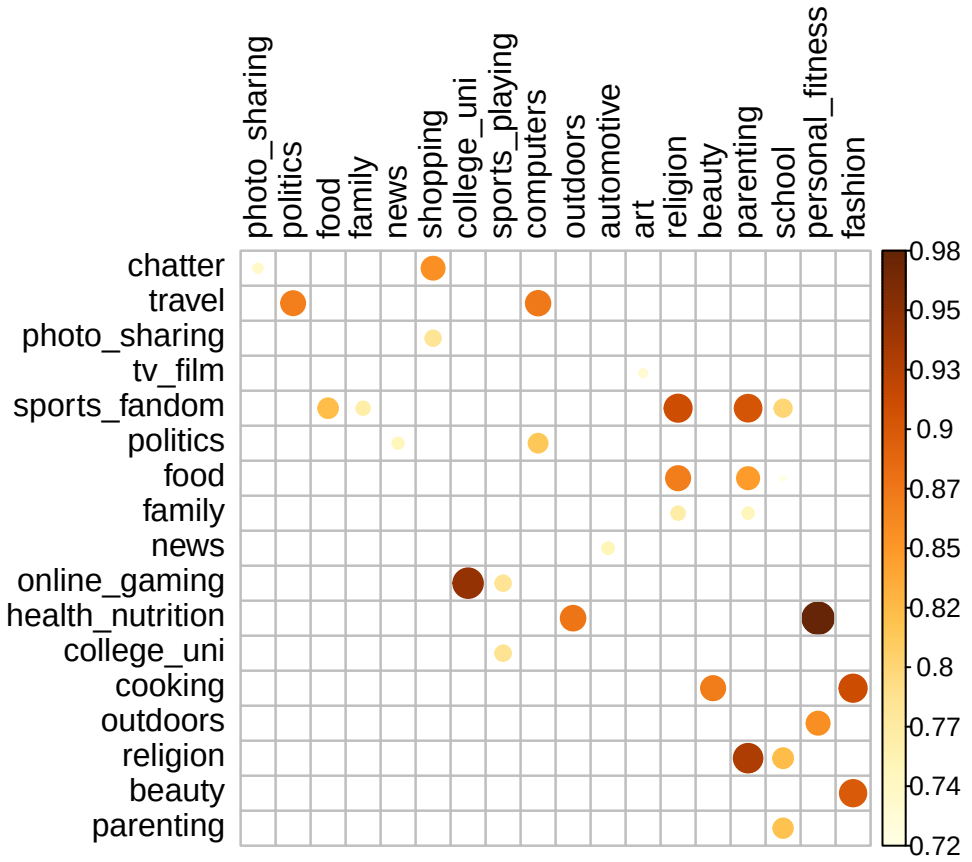
```
##                 Var1             Var2      Freq
## 1132 health_nutrition personal_fitness 0.9784190
## 590     online_gaming      college_uni 0.9477743
## 1035         religion        parenting 0.9329145
## 1171          cooking          fashion 0.9126470
## 943     sports_fandom         religion 0.9110081
## 1015    sports_fandom        parenting 0.9052349
## 1180           beauty          fashion 0.9003331
## 808  health_nutrition         outdoors 0.8767125
## 723            travel        computers 0.8743096
## 991           cooking           beauty 0.8708852
## 945              food         religion 0.8702392
## 255            travel         politics 0.8686486
## 1139         outdoors personal_fitness 0.8569863
## 505           chatter         shopping 0.8558037
## 1017             food        parenting 0.8461772
## 1107         religion           school 0.8221967
## 295     sports_fandom             food 0.8200116
## 1109        parenting           school 0.8172017
## 728           politics        computers 0.8113277
## 1087    sports_fandom           school 0.7968535
## 629       college_uni   sports_playing 0.7823720
## 626     online_gaming   sports_playing 0.7821580
## 508     photo_sharing         shopping 0.7808589
## 946            family         religion 0.7667968
## 331     sports_fandom           family 0.7661486
## 877               news       automotive 0.7547770
## 1018           family        parenting 0.7530952
## 440          politics             news 0.7515664
## 109           chatter    photo_sharing 0.7416774
## 906           tv_film              art 0.7360172
## 1089             food           school 0.7185102
## 333              food           family 0.6775890
## 1090           family           school 0.6507692
## 1295             spam            adult 0.5097285
```




```
corr_simple3 <- function(data=d2,sig=0.7){
  #convert data to numeric in order to run correlations
  #convert to factor first to keep the integrity of the data - each value will become a number rather t
  df_cor <- data %>% mutate_if(is.character, as.factor)
  df_cor <- df_cor %>% mutate_if(is.factor, as.numeric)
  #run a correlation and drop the insignificant ones
  corr <- cor(df_cor)
  #prepare to drop duplicates and correlations of 1
  corr[lower.tri(corr,diag=TRUE)] <- NA
  #drop perfect correlations
  corr[corr == 1] <- NA
  #turn into a 3-column table
  corr <- as.data.frame(as.table(corr))
  #remove the NA values from above
  corr <- na.omit(corr)
  #select significant values
  corr <- subset(corr, abs(Freq) > sig)
  #sort by highest correlation
  corr <- corr[order(-abs(corr$Freq)),]
  #print table
  print(corr)
  j <- corr %>% group_by(Var1,Var2) %>% count()
  print(j)
  #turn corr back into matrix in order to plot with corrplot
  mtx_corr <- reshape2::acast(corr, Var1~Var2, value.var="Freq")
```

```
  #plot correlations visually
  corrplot(mtx_corr, is.corr=FALSE, tl.col="black", na.label=" ")
}
corr_simple3()
```

```
##                  Var1             Var2      Freq
## 1132 health_nutrition personal_fitness 0.9784190
## 590     online_gaming      college_uni 0.9477743
## 1035         religion        parenting 0.9329145
## 1171          cooking          fashion 0.9126470
## 943     sports_fandom         religion 0.9110081
## 1015    sports_fandom        parenting 0.9052349
## 1180           beauty          fashion 0.9003331
## 808  health_nutrition         outdoors 0.8767125
## 723            travel        computers 0.8743096
## 991           cooking           beauty 0.8708852
## 945              food         religion 0.8702392
## 255            travel         politics 0.8686486
## 1139         outdoors personal_fitness 0.8569863
## 505           chatter         shopping 0.8558037
## 1017             food        parenting 0.8461772
## 1107         religion           school 0.8221967
## 295     sports_fandom             food 0.8200116
## 1109        parenting           school 0.8172017
## 728           politics        computers 0.8113277
## 1087    sports_fandom           school 0.7968535
## 629       college_uni   sports_playing 0.7823720
## 626     online_gaming   sports_playing 0.7821580
## 508     photo_sharing         shopping 0.7808589
## 946            family         religion 0.7667968
## 331     sports_fandom           family 0.7661486
## 877              news       automotive 0.7547770
## 1018           family        parenting 0.7530952
## 440          politics             news 0.7515664
## 109           chatter    photo_sharing 0.7416774
## 906           tv_film              art 0.7360172
## 1089             food           school 0.7185102
## # A tibble: 31 x 3
## # Groups:   Var1, Var2 [31]
##    Var1          Var2              n
##    <fct>         <fct>         <int>
##  1 chatter       photo_sharing     1
##  2 chatter       shopping          1
##  3 travel        politics          1
##  4 travel        computers         1
##  5 photo_sharing shopping          1
##  6 tv_film       art               1
##  7 sports_fandom food              1
##  8 sports_fandom family            1
##  9 sports_fandom religion          1
## 10 sports_fandom parenting         1
## # ... with 21 more rows
## # i Use ‘print(n = ...)‘ to see more rows
```



#PCA and clustering analysis along with network graph

```
# this gives us a good idea of what people looked at together
soc
```

```
## # A tibble: 7,882 x 36
##    chatter current~1 travel photo~2 uncat~3 tv_film sport~4 polit~5  food family
##      <dbl>     <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>  <dbl>
##  1       2         0      2       2       2       1       1       0     4      1
##  2       3         3      2       1       1       1       4       1     2      2
##  3       6         3      4       3       1       5       0       2     1      1
##  4       1         5      2       2       0       1       0       1     0      1
##  5       5         2      0       6       1       0       0       2     0      1
##  6       6         4      2       7       0       1       1       0     2      1
##  7       1         2      7       1       0       1       1      11     1      0
##  8       5         3      3       6       1       1       1       0     0      0
##  9       6         2      0       1       0       0       0       0     2      2
## 10       5         2      4       4       0       5       9       1     5      4
## # ... with 7,872 more rows, 26 more variables: home_and_garden <dbl>,
## #   music <dbl>, news <dbl>, online_gaming <dbl>, shopping <dbl>,
## #   health_nutrition <dbl>, college_uni <dbl>, sports_playing <dbl>,
## #   cooking <dbl>, eco <dbl>, computers <dbl>, business <dbl>, outdoors <dbl>,
## #   crafts <dbl>, automotive <dbl>, art <dbl>, religion <dbl>, beauty <dbl>,
## #   parenting <dbl>, dating <dbl>, school <dbl>, personal_fitness <dbl>,
## #   fashion <dbl>, small_business <dbl>, spam <dbl>, adult <dbl>, and ...
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

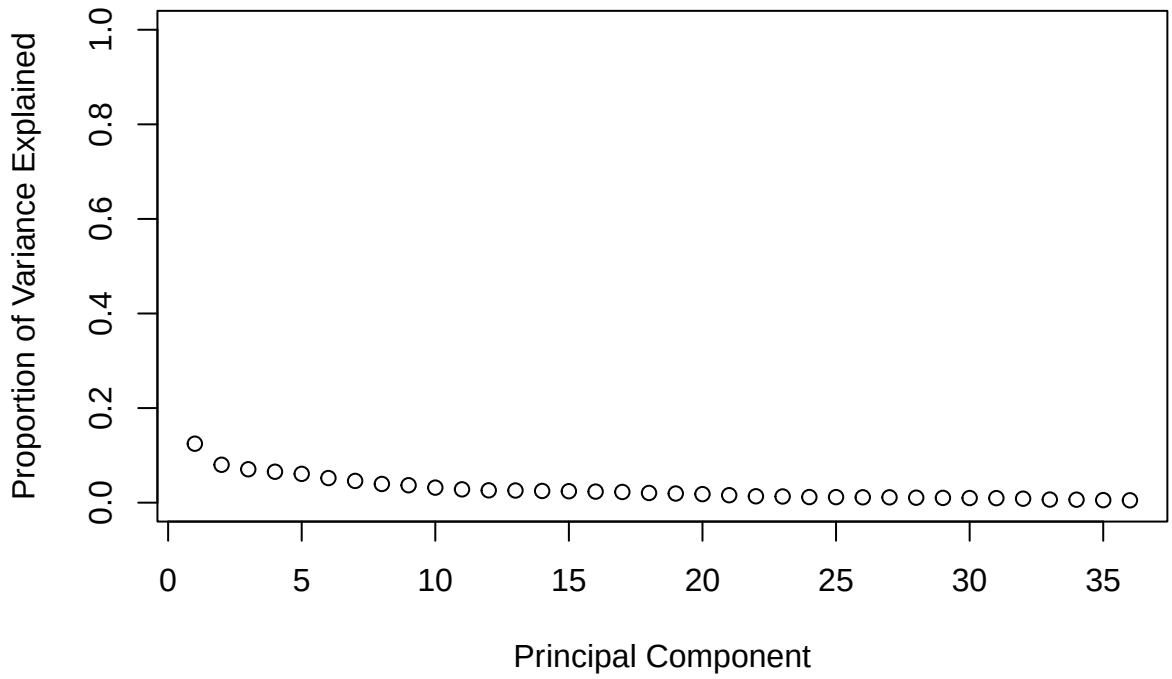
```
soc.pca <- prcomp(na.omit(soc),center = T,scale. = T)
summary(soc.pca)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1186 1.69824 1.59388 1.53457 1.48027 1.36885 1.28577
## Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087 0.05205 0.04592
## Cumulative Proportion  0.1247 0.20479 0.27536 0.34077 0.40164 0.45369 0.49961
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.19277 1.15127 1.06930 1.00566 0.96785 0.96131 0.94405
## Proportion of Variance 0.03952 0.03682 0.03176 0.02809 0.02602 0.02567 0.02476
## Cumulative Proportion  0.53913 0.57595 0.60771 0.63580 0.66182 0.68749 0.71225
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.93297 0.91698 0.9020 0.85869 0.83466 0.80544 0.75311
## Proportion of Variance 0.02418 0.02336 0.0226 0.02048 0.01935 0.01802 0.01575
## Cumulative Proportion  0.73643 0.75979 0.7824 0.80287 0.82222 0.84024 0.85599
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.69632 0.68558 0.65317 0.64881 0.63756 0.63626 0.61513
## Proportion of Variance 0.01347 0.01306 0.01185 0.01169 0.01129 0.01125 0.01051
## Cumulative Proportion  0.86946 0.88252 0.89437 0.90606 0.91735 0.92860 0.93911
##                           PC29    PC30    PC31   PC32    PC33    PC34    PC35
## Standard deviation     0.60167 0.59424 0.58683 0.5498 0.48442 0.47576 0.43757
## Proportion of Variance 0.01006 0.00981 0.00957 0.0084 0.00652 0.00629 0.00532
## Cumulative Proportion  0.94917 0.95898 0.96854 0.9769 0.98346 0.98974 0.99506
##                           PC36
## Standard deviation     0.42165
## Proportion of Variance 0.00494
## Cumulative Proportion  1.00000
```

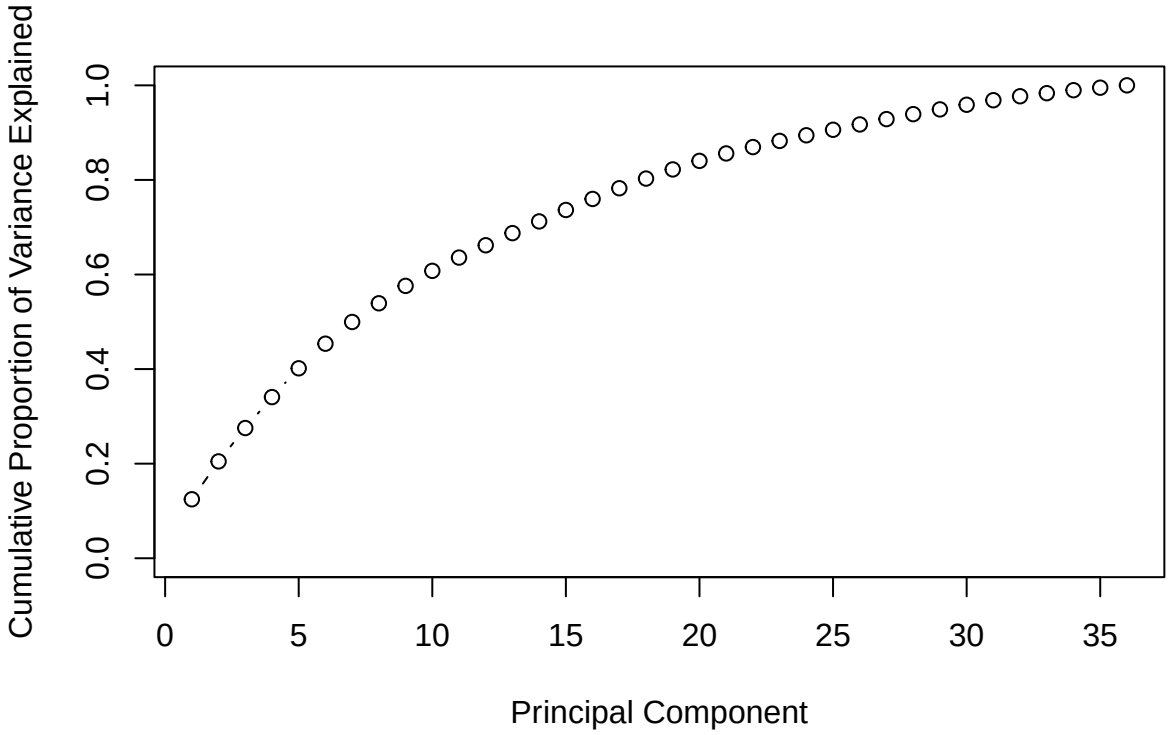
```
soc
```

```
## # A tibble: 7,882 x 36
##    chatter current~1 travel photo~2 uncat~3 tv_film sport~4 polit~5  food family
##      <dbl>     <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>  <dbl>
##  1       2         0      2       2       2       1       1       0     4      1
##  2       3         3      2       1       1       1       4       1     2      2
##  3       6         3      4       3       1       5       0       2     1      1
##  4       1         5      2       2       0       1       0       1     0      1
##  5       5         2      0       6       1       0       0       2     0      1
##  6       6         4      2       7       0       1       1       0     2      1
##  7       1         2      7       1       0       1       1      11     1      0
##  8       5         3      3       6       1       1       1       0     0      0
##  9       6         2      0       1       0       0       0       0     2      2
## 10       5         2      4       4       0       5       9       1     5      4
## # ... with 7,872 more rows, 26 more variables: home_and_garden <dbl>,
## #   music <dbl>, news <dbl>, online_gaming <dbl>, shopping <dbl>,
## #   health_nutrition <dbl>, college_uni <dbl>, sports_playing <dbl>,
## #   cooking <dbl>, eco <dbl>, computers <dbl>, business <dbl>, outdoors <dbl>,
## #   crafts <dbl>, automotive <dbl>, art <dbl>, religion <dbl>, beauty <dbl>,
## #   parenting <dbl>, dating <dbl>, school <dbl>, personal_fitness <dbl>,
## #   fashion <dbl>, small_business <dbl>, spam <dbl>, adult <dbl>, and ...
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

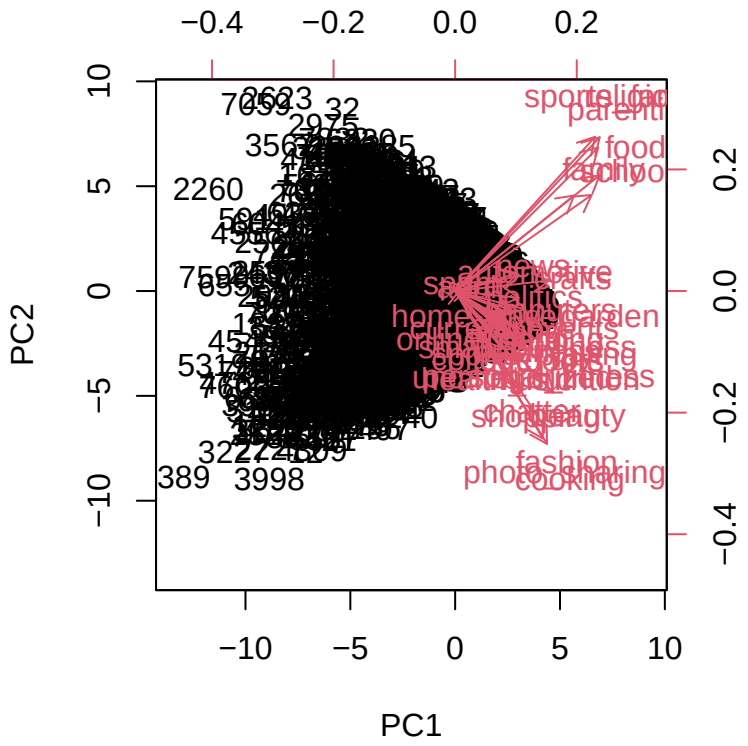
```r
soc.pca$rotation <- -1*soc.pca$rotation
pr_var <-  soc.pca$sdev ^ 2
pve <- pr_var / sum(pr_var)
plot(pve, xlab = "Principal Component", ylab = "Proportion of Variance Explained", ylim = c(0,1), type =
```




```r
plot(cumsum(pve), xlab = "Principal Component", ylab = "Cumulative Proportion of Variance Explained", yl
```




```
biplot(soc.pca, scale = 0)
```



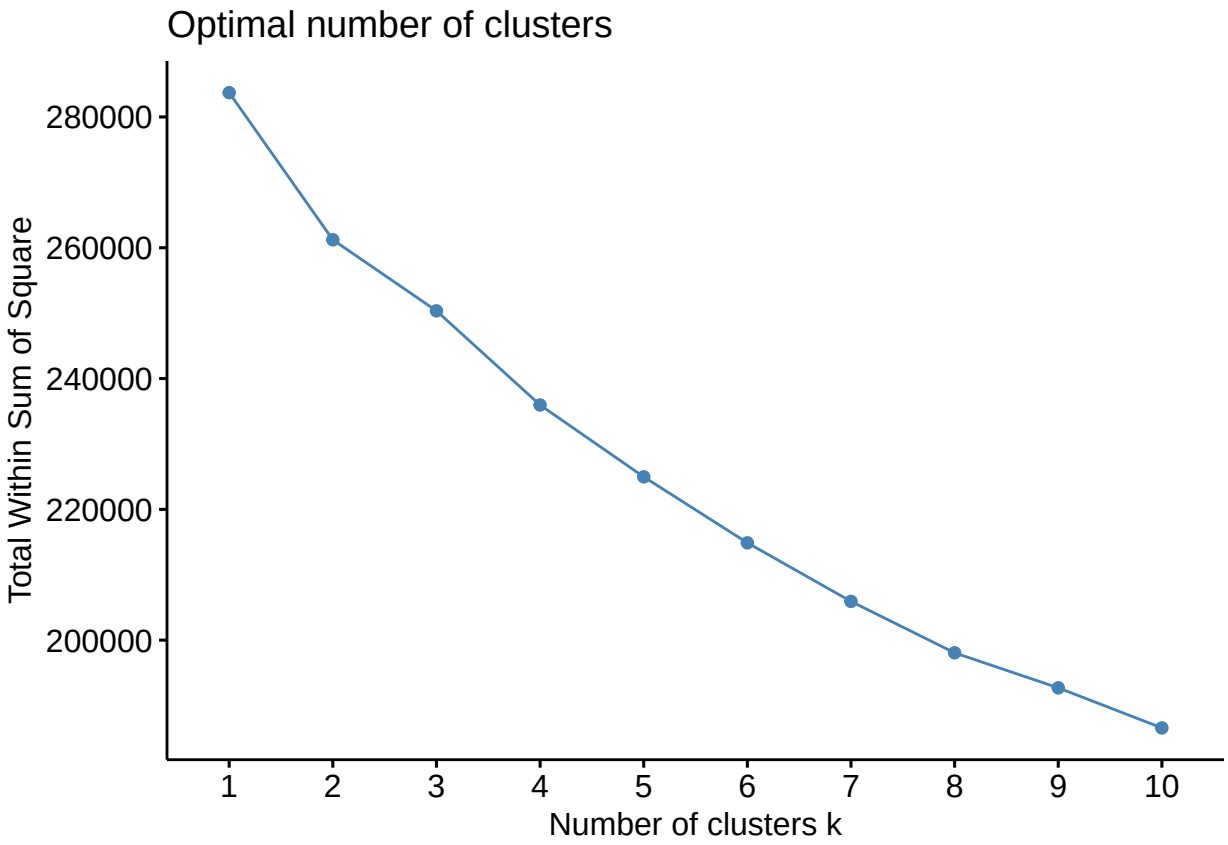

```
#install.packages('factoextra')
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

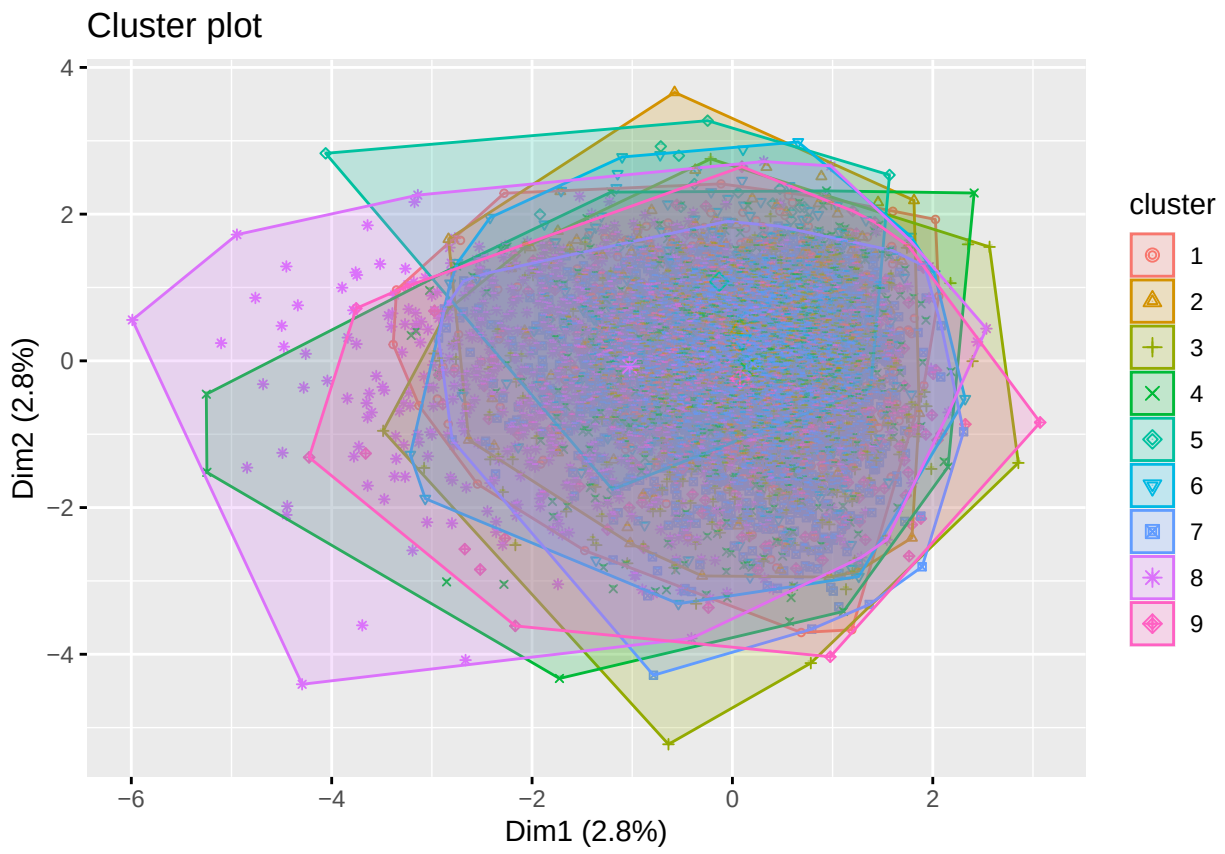
```
soc.pca$sdev
```

```
##  [1] 2.1185929 1.6982404 1.5938831 1.5345728 1.4802747 1.3688531 1.2857745
##  [8] 1.1927721 1.1512709 1.0693038 1.0056551 0.9678504 0.9613075 0.9440546
## [15] 0.9329689 0.9169821 0.9020217 0.8586866 0.8346572 0.8054378 0.7531070
## [22] 0.6963187 0.6855844 0.6531720 0.6488066 0.6375570 0.6362575 0.6151344
## [29] 0.6016711 0.5942420 0.5868324 0.5497655 0.4844242 0.4757637 0.4375693
## [36] 0.4216526
```

```
soc_transform = as.data.frame(-soc.pca$x[,1:36])
k = 9
fviz_nbclust(soc_transform, kmeans, method = 'wss')
```




```
kmeans_soc = kmeans(soc_transform, centers = k, nstart = 50)
fviz_cluster(kmeans_soc, data = soc_transform,pointsize = 1,labelsize = 1)
```

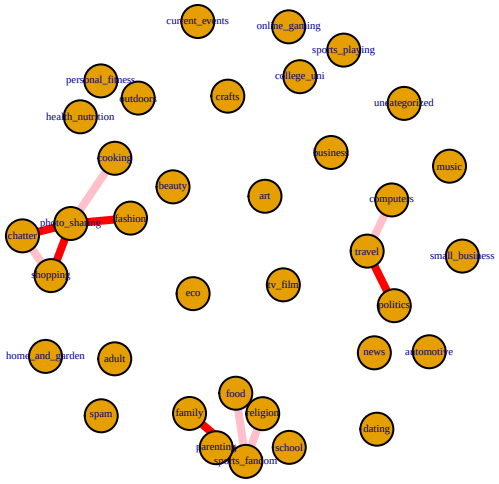



```
library(igraph)
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following object is masked from 'package:tidyr':
##
##     crossing

## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following object is masked from 'package:mosaic':
##
##     compare

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```
soc_mat <- as.matrix(soc)
soc_graph<- soc[1:100,]
soc_mat <- as.matrix(soc_graph)
library(Matrix)
mydata.cor = cor(soc, method = c("spearman"))
mydata.cor<- as.matrix(mydata.cor)
# takes out the weak correlations
mydata.cor[mydata.cor<abs(.25)] <- 0
#creates a network graph
network<-  graph_from_adjacency_matrix(mydata.cor, mode="undirected", diag=F,weighted = T,add.rownames =
plot(network,vertex.label.cex=.35,edge.color=rep(c("red","pink"),5),edge.width=4,layout = layout.frucht
```

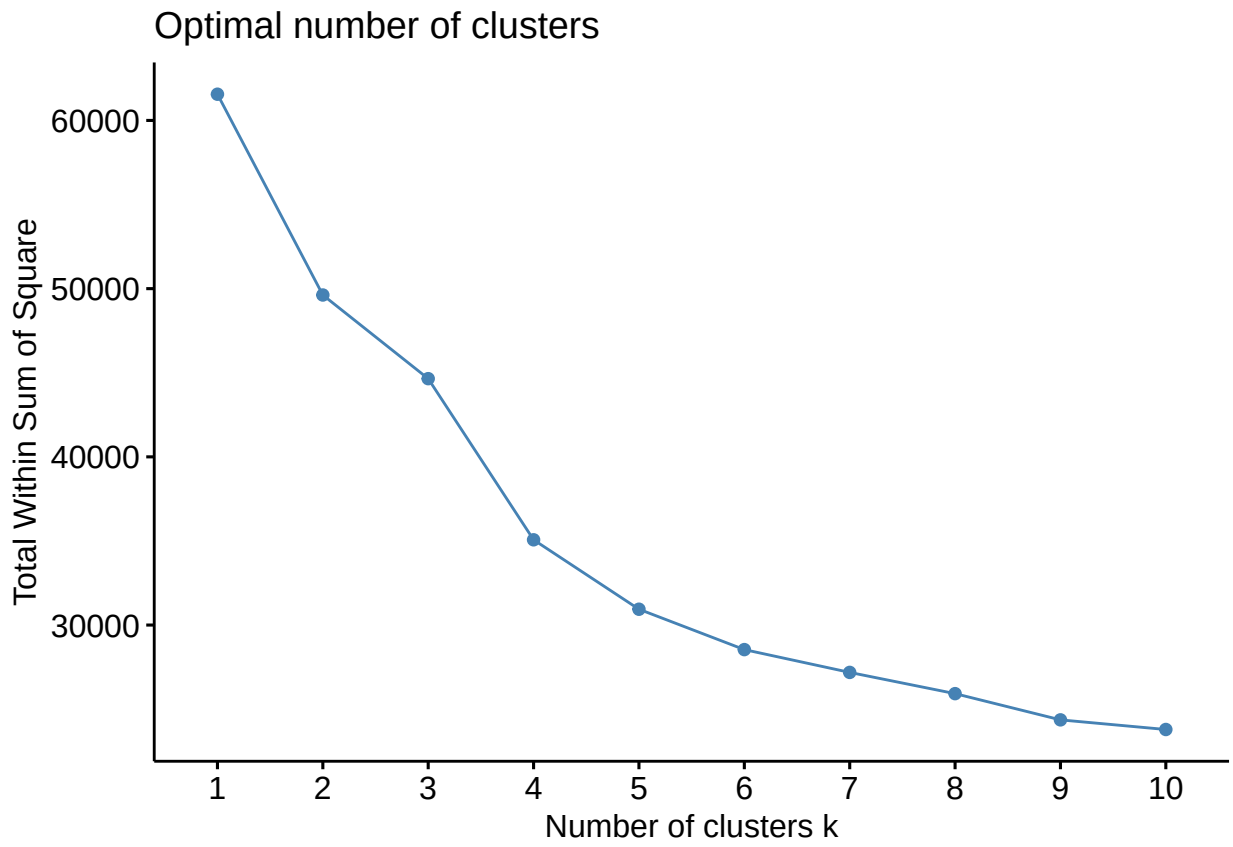



```
soc2 <- soc %>% select(sports_fandom,family,food,school,shopping,travel,health_nutrition,personal_fitnes
soc2.pca <- prcomp(soc2,center = T,scale. = T)
summary(soc2.pca)
```

```
## Importance of components:
```

```
##                            PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.5470 1.3372 1.0081 0.9952 0.82109 0.72888 0.64618
## Proportion of Variance 0.2992 0.2235 0.1270 0.1238 0.08427 0.06641 0.05219
## Cumulative Proportion  0.2992 0.5227 0.6497 0.7735 0.85778 0.92419 0.97638
##                            PC8
## Standard deviation     0.43467
## Proportion of Variance 0.02362
## Cumulative Proportion  1.00000
```

```
soc2.pca$rotation <- -1*soc.pca$rotation
soc_transform = as.data.frame(-soc2.pca$x[,1:7])
k = 4
fviz_nbclust(soc_transform, kmeans, method = 'wss')
```




```
kmeans_soc = kmeans(soc_transform, centers = k, nstart = 50)
fviz_cluster(kmeans_soc, data = soc_transform,pointsize = 1,labelsize = 1)
```

#conclusion Using the Market Segmentation data our task is to find trends in what people are looking at to try and figure how to more effectively market at all of our customers. The place I first wanted to start was looking at some correlation matrices. These matrices will help us see which categories are highly correlated with one another. This will help us get an idea of what categories people looked at together. The functions I used created a table with the relative correlation values. I had 3 different matrices each one getting more selective than the last. The first Matrix included all correlations with .3 or higher, this means even weak correlations were included. This matrix was hard to read and did not provide us with any valuable information. However, moving up the ladder and making removing the weaker correlation eventually up to .75 provides us with a much clearer picture. These correlations provide a good base on how the categories interact with one another. I wanted to use PCA to reduce the dimensions of my data set but after running it, it simply did not paint a clear picture of different groups. This can be seen in my cluster plot I created. However, using some of the highly correlated categories. Using these categories I ran another PCA this PCA heeded better results but still not great. Finally I created a network graph to show how some of the correlations related to one another. This helps us see which categories can be advertised to the most wide range of categories. I think this was the most helpful insight from my analysis. Sometimes unsupervised learning isn't the best choice and until I figure out what to do with the parameters to make them more understandable, I believe the best way to look at marketing would be to find the correlations and just make ad suggestions based on those with high correlations.

#Reuters Corpus -Aaron Pressman

```
library(tm)
```

```
## 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
```

```
library(tidyverse)
library(slam)
library(proxy)
```

```
##
## Attaching package: 'proxy'

## The following object is masked from 'package:Matrix':
##
##     as.matrix

## The following objects are masked from 'package:stats':
##
##     as.dist, dist

## The following object is masked from 'package:base':
##
##     as.matrix
```

```
readerPlain = function(fname){
    readPlain(elem=list(content=readLines(fname)),
              id=fname, language='en') }
file_list = Sys.glob('C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/R
aaron = lapply(file_list, readerPlain)
file_list
```

```
##  [1] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
##  [2] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
##  [3] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
##  [4] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
##  [5] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
##  [6] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
##  [7] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
##  [8] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
##  [9] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [10] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [11] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [12] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [13] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [14] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
```

```
## [15] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [16] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [17] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [18] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [19] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [20] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [21] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [22] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [23] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [24] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [25] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [26] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [27] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [28] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [29] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [30] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [31] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [32] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [33] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [34] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [35] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [36] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [37] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [38] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [39] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [40] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [41] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [42] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [43] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [44] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [45] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [46] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [47] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [48] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [49] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
## [50] "C:/Users/chama/OneDrive/Documents/SUMMER 2022/Machine Learning/2nd half/Project/ReutersC50/C50
```

```
mynames = file_list %>%
    { strsplit(., '/', fixed=TRUE) } %>%
    { lapply(., tail, n=2) } %>%
    { lapply(., paste0, collapse = '') } %>%
    unlist
# Rename the articles
mynames
```

```
##  [1] "AaronPressman106247newsML.txt" "AaronPressman120600newsML.txt"
##  [3] "AaronPressman120683newsML.txt" "AaronPressman136958newsML.txt"
##  [5] "AaronPressman137498newsML.txt" "AaronPressman14014newsML.txt"
##  [7] "AaronPressman156814newsML.txt" "AaronPressman182596newsML.txt"
##  [9] "AaronPressman186392newsML.txt" "AaronPressman193495newsML.txt"
## [11] "AaronPressman196805newsML.txt" "AaronPressman197734newsML.txt"
## [13] "AaronPressman206838newsML.txt" "AaronPressman231479newsML.txt"
## [15] "AaronPressman233150newsML.txt" "AaronPressman237175newsML.txt"
## [17] "AaronPressman249407newsML.txt" "AaronPressman2537newsML.txt"
```

```
## [19] "AaronPressman266038newsML.txt" "AaronPressman269995newsML.txt"
## [21] "AaronPressman269999newsML.txt" "AaronPressman270046newsML.txt"
## [23] "AaronPressman270084newsML.txt" "AaronPressman270134newsML.txt"
## [25] "AaronPressman270346newsML.txt" "AaronPressman275174newsML.txt"
## [27] "AaronPressman277117newsML.txt" "AaronPressman277513newsML.txt"
## [29] "AaronPressman290125newsML.txt" "AaronPressman299375newsML.txt"
## [31] "AaronPressman312178newsML.txt" "AaronPressman324896newsML.txt"
## [33] "AaronPressman325347newsML.txt" "AaronPressman330967newsML.txt"
## [35] "AaronPressman331411newsML.txt" "AaronPressman347226newsML.txt"
## [37] "AaronPressman354135newsML.txt" "AaronPressman354285newsML.txt"
## [39] "AaronPressman357147newsML.txt" "AaronPressman366020newsML.txt"
## [41] "AaronPressman369570newsML.txt" "AaronPressman371380newsML.txt"
## [43] "AaronPressman372744newsML.txt" "AaronPressman372989newsML.txt"
## [45] "AaronPressman372995newsML.txt" "AaronPressman378457newsML.txt"
## [47] "AaronPressman394237newsML.txt" "AaronPressman398094newsML.txt"
## [49] "AaronPressman401260newsML.txt" "AaronPressman407599newsML.txt"
```

```
names(aaron) = mynames
## once you have documents in a vector, you
## create a text mining 'corpus' with:
documents_raw = Corpus(VectorSource(aaron))
## Some pre-processing/tokenization steps.
## tm_map just maps some function to every document in the corpus
my_documents = documents_raw
my_documents = tm_map(my_documents, content_transformer(tolower)) # make everything lowercase
```

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

```
my_documents = tm_map(my_documents, content_transformer(removeNumbers)) # remove numbers
```

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

```
my_documents = tm_map(my_documents, content_transformer(removePunctuation)) # remove punctuation
```

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

```
my_documents = tm_map(my_documents, content_transformer(stripWhitespace)) ## remove excess white-space
```

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

```
## Remove stopwords.  Always be careful with this!
stopwords("en")
```

```
##   [1] "i"          "me"         "my"         "myself"     "we"
##   [6] "our"        "ours"       "ourselves"  "you"        "your"
##  [11] "yours"      "yourself"   "yourselves" "he"         "him"
```

```
##  [16] "his"        "himself"    "she"        "her"        "hers"
##  [21] "herself"    "it"         "its"        "itself"     "they"
##  [26] "them"       "their"      "theirs"     "themselves" "what"
##  [31] "which"      "who"        "whom"       "this"       "that"
##  [36] "these"      "those"      "am"         "is"         "are"
##  [41] "was"        "were"       "be"         "been"       "being"
##  [46] "have"       "has"        "had"        "having"     "do"
##  [51] "does"       "did"        "doing"      "would"      "should"
##  [56] "could"      "ought"      "i'm"        "you're"     "he's"
##  [61] "she's"      "it's"       "we're"      "they're"    "i've"
##  [66] "you've"     "we've"      "they've"    "i'd"        "you'd"
##  [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"
##  [76] "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
##  [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
##  [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
##  [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"
##  [96] "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
## [101] "who's"      "what's"     "here's"     "there's"    "when's"
## [106] "where's"    "why's"      "how's"      "a"          "an"
## [111] "the"        "and"        "but"        "if"         "or"
## [116] "because"    "as"         "until"      "while"      "of"
## [121] "at"         "by"         "for"        "with"       "about"
## [126] "against"    "between"    "into"       "through"    "during"
## [131] "before"     "after"      "above"      "below"      "to"
## [136] "from"       "up"         "down"       "in"         "out"
## [141] "on"         "off"        "over"       "under"      "again"
## [146] "further"    "then"       "once"       "here"       "there"
## [151] "when"       "where"      "why"        "how"        "all"
## [156] "any"        "both"       "each"       "few"        "more"
## [161] "most"       "other"      "some"       "such"       "no"
## [166] "nor"        "not"        "only"       "own"        "same"
## [171] "so"         "than"       "too"        "very"
```

```
stopwords("SMART")
```

```
##   [1] "a"             "a's"           "able"          "about"
##   [5] "above"         "according"     "accordingly"   "across"
##   [9] "actually"      "after"         "afterwards"    "again"
##  [13] "against"       "ain't"         "all"           "allow"
##  [17] "allows"        "almost"        "alone"         "along"
##  [21] "already"       "also"          "although"      "always"
##  [25] "am"            "among"         "amongst"       "an"
##  [29] "and"           "another"       "any"           "anybody"
##  [33] "anyhow"        "anyone"        "anything"      "anyway"
##  [37] "anyways"       "anywhere"      "apart"         "appear"
##  [41] "appreciate"    "appropriate"   "are"           "aren't"
##  [45] "around"        "as"            "aside"         "ask"
##  [49] "asking"        "associated"    "at"            "available"
##  [53] "away"          "awfully"       "b"             "be"
##  [57] "became"        "because"       "become"        "becomes"
##  [61] "becoming"      "been"          "before"        "beforehand"
##  [65] "behind"        "being"         "believe"       "below"
##  [69] "beside"        "besides"       "best"          "better"
##  [73] "between"       "beyond"        "both"          "brief"
```

```
##  [77] "but"           "by"            "c"             "c'mon"
##  [81] "c's"           "came"          "can"           "can't"
##  [85] "cannot"        "cant"          "cause"         "causes"
##  [89] "certain"       "certainly"     "changes"       "clearly"
##  [93] "co"            "com"           "come"          "comes"
##  [97] "concerning"    "consequently"  "consider"      "considering"
## [101] "contain"       "containing"    "contains"      "corresponding"
## [105] "could"         "couldn't"      "course"        "currently"
## [109] "d"             "definitely"    "described"     "despite"
## [113] "did"           "didn't"        "different"     "do"
## [117] "does"          "doesn't"       "doing"         "don't"
## [121] "done"          "down"          "downwards"     "during"
## [125] "e"             "each"          "edu"           "eg"
## [129] "eight"         "either"        "else"          "elsewhere"
## [133] "enough"        "entirely"      "especially"    "et"
## [137] "etc"           "even"          "ever"          "every"
## [141] "everybody"     "everyone"      "everything"    "everywhere"
## [145] "ex"            "exactly"       "example"       "except"
## [149] "f"             "far"           "few"           "fifth"
## [153] "first"         "five"          "followed"      "following"
## [157] "follows"       "for"           "former"        "formerly"
## [161] "forth"         "four"          "from"          "further"
## [165] "furthermore"   "g"             "get"           "gets"
## [169] "getting"       "given"         "gives"         "go"
## [173] "goes"          "going"         "gone"          "got"
## [177] "gotten"        "greetings"     "h"             "had"
## [181] "hadn't"        "happens"       "hardly"        "has"
## [185] "hasn't"        "have"          "haven't"       "having"
## [189] "he"            "he's"          "hello"         "help"
## [193] "hence"         "her"           "here"          "here's"
## [197] "hereafter"     "hereby"        "herein"        "hereupon"
## [201] "hers"          "herself"       "hi"            "him"
## [205] "himself"       "his"           "hither"        "hopefully"
## [209] "how"           "howbeit"       "however"       "i"
## [213] "i'd"           "i'll"          "i'm"           "i've"
## [217] "ie"            "if"            "ignored"       "immediate"
## [221] "in"            "inasmuch"      "inc"           "indeed"
## [225] "indicate"      "indicated"     "indicates"     "inner"
## [229] "insofar"       "instead"       "into"          "inward"
## [233] "is"            "isn't"         "it"            "it'd"
## [237] "it'll"         "it's"          "its"           "itself"
## [241] "j"             "just"          "k"             "keep"
## [245] "keeps"         "kept"          "know"          "knows"
## [249] "known"         "l"             "last"          "lately"
## [253] "later"         "latter"        "latterly"      "least"
## [257] "less"          "lest"          "let"           "let's"
## [261] "like"          "liked"         "likely"        "little"
## [265] "look"          "looking"       "looks"         "ltd"
## [269] "m"             "mainly"        "many"          "may"
## [273] "maybe"         "me"            "mean"          "meanwhile"
## [277] "merely"        "might"         "more"          "moreover"
## [281] "most"          "mostly"        "much"          "must"
## [285] "my"            "myself"        "n"             "name"
## [289] "namely"        "nd"            "near"          "nearly"
```

```
## [293] "necessary"     "need"          "needs"         "neither"
## [297] "never"         "nevertheless"  "new"           "next"
## [301] "nine"          "no"            "nobody"        "non"
## [305] "none"          "noone"         "nor"           "normally"
## [309] "not"           "nothing"       "novel"         "now"
## [313] "nowhere"       "o"             "obviously"     "of"
## [317] "off"           "often"         "oh"            "ok"
## [321] "okay"          "old"           "on"            "once"
## [325] "one"           "ones"          "only"          "onto"
## [329] "or"            "other"         "others"        "otherwise"
## [333] "ought"         "our"           "ours"          "ourselves"
## [337] "out"           "outside"       "over"          "overall"
## [341] "own"           "p"             "particular"    "particularly"
## [345] "per"           "perhaps"       "placed"        "please"
## [349] "plus"          "possible"      "presumably"    "probably"
## [353] "provides"      "q"             "que"           "quite"
## [357] "qv"            "r"             "rather"        "rd"
## [361] "re"            "really"        "reasonably"    "regarding"
## [365] "regardless"    "regards"       "relatively"    "respectively"
## [369] "right"         "s"             "said"          "same"
## [373] "saw"           "say"           "saying"        "says"
## [377] "second"        "secondly"      "see"           "seeing"
## [381] "seem"          "seemed"        "seeming"       "seems"
## [385] "seen"          "self"          "selves"        "sensible"
## [389] "sent"          "serious"       "seriously"     "seven"
## [393] "several"       "shall"         "she"           "should"
## [397] "shouldn't"     "since"         "six"           "so"
## [401] "some"          "somebody"      "somehow"       "someone"
## [405] "something"     "sometime"      "sometimes"     "somewhat"
## [409] "somewhere"     "soon"          "sorry"         "specified"
## [413] "specify"       "specifying"    "still"         "sub"
## [417] "such"          "sup"           "sure"          "t"
## [421] "t's"           "take"          "taken"         "tell"
## [425] "tends"         "th"            "than"          "thank"
## [429] "thanks"        "thanx"         "that"          "that's"
## [433] "thats"         "the"           "their"         "theirs"
## [437] "them"          "themselves"    "then"          "thence"
## [441] "there"         "there's"       "thereafter"    "thereby"
## [445] "therefore"     "therein"       "theres"        "thereupon"
## [449] "these"         "they"          "they'd"        "they'll"
## [453] "they're"       "they've"       "think"         "third"
## [457] "this"          "thorough"      "thoroughly"    "those"
## [461] "though"        "three"         "through"       "throughout"
## [465] "thru"          "thus"          "to"            "together"
## [469] "too"           "took"          "toward"        "towards"
## [473] "tried"         "tries"         "truly"         "try"
## [477] "trying"        "twice"         "two"           "u"
## [481] "un"            "under"         "unfortunately" "unless"
## [485] "unlikely"      "until"         "unto"          "up"
## [489] "upon"          "us"            "use"           "used"
## [493] "useful"        "uses"          "using"         "usually"
## [497] "uucp"          "v"             "value"         "various"
## [501] "very"          "via"           "viz"           "vs"
## [505] "w"             "want"          "wants"         "was"
```

```
## [509] "wasn't"        "way"          "we"           "we'd"
## [513] "we'll"         "we're"        "we've"        "welcome"
## [517] "well"          "went"         "were"         "weren't"
## [521] "what"          "what's"       "whatever"     "when"
## [525] "whence"        "whenever"     "where"        "where's"
## [529] "whereafter"    "whereas"      "whereby"      "wherein"
## [533] "whereupon"     "wherever"     "whether"      "which"
## [537] "while"         "whither"      "who"          "who's"
## [541] "whoever"       "whole"        "whom"         "whose"
## [545] "why"           "will"         "willing"      "wish"
## [549] "with"          "within"       "without"      "won't"
## [553] "wonder"        "would"        "would"        "wouldn't"
## [557] "x"             "y"            "yes"          "yet"
## [561] "you"           "you'd"        "you'll"       "you're"
## [565] "you've"        "your"         "yours"        "yourself"
## [569] "yourselves"    "z"            "zero"
```

```
my_documents = tm_map(my_documents, content_transformer(removeWords), stopwords("en"))
```

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

```
## create a doc-term-matrix
DTM_aaron = DocumentTermMatrix(my_documents)
DTM_aaron # some basic summary statistics
```

```
## <<DocumentTermMatrix (documents: 50, terms: 3158)>>
## Non-/sparse entries: 10434/147466
## Sparsity           : 93%
## Maximal term length: 47
## Weighting          : term frequency (tf)
```

```
class(DTM_aaron)  # a special kind of sparse matrix format
```

```
## [1] "DocumentTermMatrix"    "simple_triplet_matrix"
```

```
## You can inspect its entries...
## ...find words with greater than a min count...
findFreqTerms(DTM_aaron, 50)
```

```
##  [1] "also"
##  [2] "character"
##  [3] "cuserschamaonedrivedocumentssummer"
##  [4] "datetimestamp"
##  [5] "description"
##  [6] "federal"
##  [7] "halfprojectreuterscctrainaaronpressmannewsmltxt"
##  [8] "heading"
##  [9] "hour"
## [10] "internet"
## [11] "isdst"
```

```
## [12] "language"
## [13] "learningnd"
## [14] "listcontent"
## [15] "machine"
## [16] "mday"
## [17] "meta"
## [18] "min"
## [19] "mon"
## [20] "new"
## [21] "origin"
## [22] "said"
## [23] "wday"
## [24] "yday"
## [25] "year"
## [26] "online"
## [27] "will"
## [28] "companies"
## [29] "banks"
## [30] "congress"
## [31] "credit"
## [32] "financial"
## [33] "policy"
## [34] "court"
## [35] "encryption"
```

```
## ...or find words whose count correlates with a specified word.
findAssocs(DTM_aaron, "banks", .5)
```

```
## $banks
##   comptroller    securities         field       playing  underwriting
##          0.77          0.74          0.72          0.72          0.70
##       percent     brookings         litan        twoway        wanted
##          0.69          0.67          0.67          0.67          0.65
##    industries           act        action          move      pressure
##          0.65          0.64          0.64          0.64          0.64
##  subsidiaries     institute     governors       section glasssteagall
##          0.64          0.63          0.63          0.62          0.61
##           fed          bank      american          next         alice
##          0.60          0.59          0.59          0.58          0.58
##        clause        elated       engaged      engaging         gifts
##          0.58          0.58          0.58          0.58          0.58
##   principally       rivlin         santa         steve   unanimously
##          0.58          0.58          0.58          0.58          0.58
##     incentive     prohibits     generally     insurance           cap
##          0.57          0.57          0.56          0.56          0.56
##      activity        entire       larocca       nonbank    underwrite
##          0.55          0.55          0.55          0.55          0.55
##       bankers      director        street       revenue          feds
##          0.53          0.52          0.52          0.52          0.52
##         larry      managing          hope         firms         begin
##          0.52          0.52          0.51          0.51          0.51
##         limit     affiliate neighbourhood
##          0.51          0.50          0.50
```

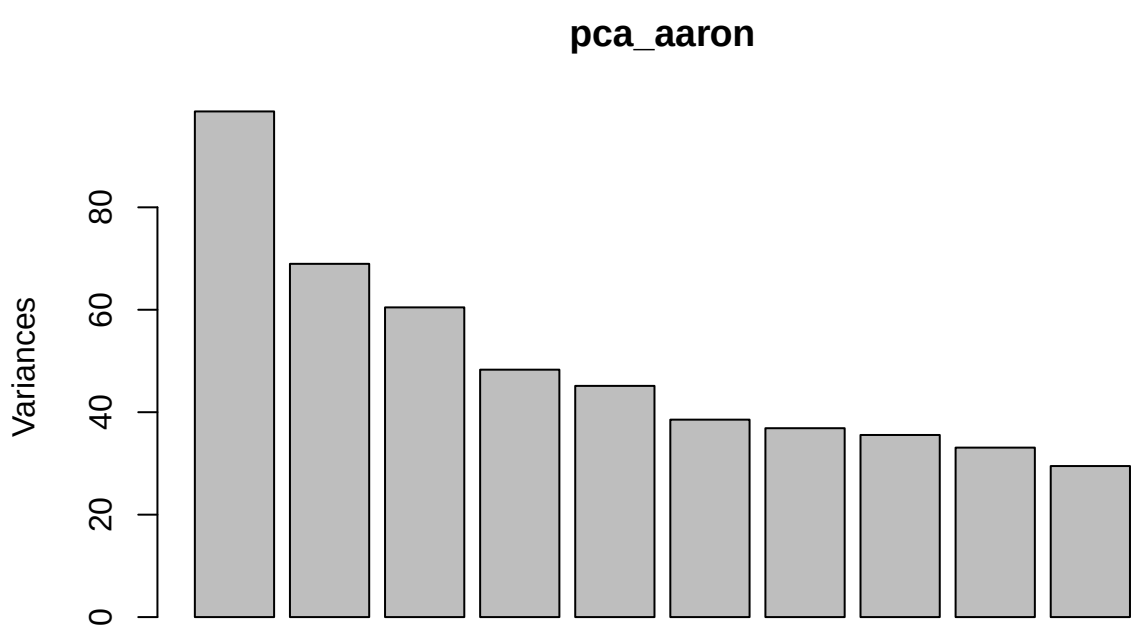
```
## Drop those terms that only occur in one or two documents
## This is a common step: the noise of the "long tail" (rare terms)
##  can be huge, and there is nothing to learn if a term occurred once.
## Below removes those terms that have count 0 in >95% of docs.
## Probably a bit extreme in most cases... but here only 50 docs!
DTM_aaron = removeSparseTerms(DTM_aaron, 0.95)
DTM_aaron # now ~ 1000 terms (versus ~3000 before)
```

```
## <<DocumentTermMatrix (documents: 50, terms: 1077)>>
## Non-/sparse entries: 7767/46083
## Sparsity           : 86%
## Maximal term length: 47
## Weighting          : term frequency (tf)
```

```
# construct TF IDF weights
tfidf_aaron = weightTfIdf(DTM_aaron)
####
# Compare documents
####
# could go back to the raw corpus
####
# Dimensionality reduction
####
# Now PCA on term frequencies
X = as.matrix(tfidf_aaron)
summary(colSums(X))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.05699 0.07653 0.09901 0.11768 0.68420
```

```
scrub_cols = which(colSums(X) == 0)
X = X[,-scrub_cols]
pca_aaron = prcomp(X, rank=2, scale=TRUE)
plot(pca_aaron)
```




```
# Look at the loadings
pca_aaron$rotation[order(abs(pca_aaron$rotation[,1]),decreasing=TRUE),1][1:25]
```

```
##          bag        blunt        books      careful    cinternet   cochairman
##   0.09960316   0.09960316   0.09960316   0.09960316   0.09960316   0.09960316
##     conceded    copyright        curve      decency     defeated      destroy
##   0.09960316   0.09960316   0.09960316   0.09960316   0.09960316   0.09960316
## developments       dodge         ever       evslin      express      fretted
##   0.09960316   0.09960316   0.09960316   0.09960316   0.09960316   0.09960316
##         gage        goals   greenhouse    guarantee    heartened    hollywood
##   0.09960316   0.09960316   0.09960316   0.09960316   0.09960316   0.09960316
##        hopes
##   0.09960316
```

```
pca_aaron$rotation[order(abs(pca_aaron$rotation[,2]),decreasing=TRUE),2][1:25]
```

```
##          com      control          edu     assigned      channel    employees
##  -0.08882045  -0.08882045  -0.08882045  -0.08687726  -0.08687726  -0.08687726
##       fought          mtv       mtvcom        music   outofcourt    regaining
##  -0.08687726  -0.08687726  -0.08687726  -0.08687726  -0.08687726  -0.08687726
##   registered   settlement       viacom      society   securities    standards
##  -0.08687726  -0.08687726  -0.08687726  -0.08560848   0.08552011  -0.08550342
##         name        helps        heath          gov designations underwriting
##  -0.08276945  -0.08270609  -0.08268504  -0.08230267  -0.08212610   0.08105300
##        names
```

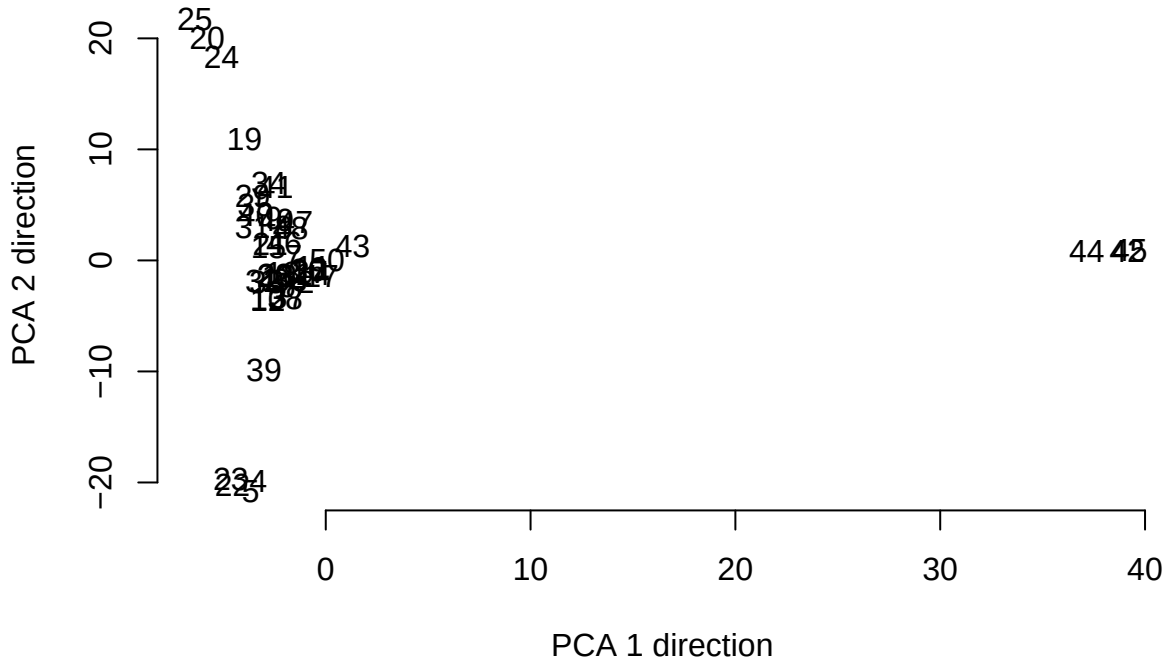
```
##  -0.08062340
```

```
## Look at the first two PCs..
# We've now turned each document into a single pair of numbers -- massive dimensionality reduction
pca_aaron$x[,1:2]
```

```
##
## Docs        PC1          PC2
##   1  -2.7631713  -1.76123094
##   2  -0.9754208  -1.89712178
##   3  -1.9750012  -1.37428154
##   4  -3.2883989 -19.85740444
##   5  -3.6681802 -20.81356590
##   6  -1.8953941  -2.27696590
##   7  -1.5822806   0.44013271
##   8  -2.0472546   2.96272925
##   9  -2.5431764   3.85480956
##   10 -2.8243986  -3.45226628
##   11 -0.6297834  -0.61279538
##   12 -2.8251968  -3.53673634
##   13 -2.2510759  -1.36804880
##   14 -0.6303470  -1.20942008
##   15 -2.7891993   1.18496188
##   16 -0.8722269  -0.83058266
##   17 -0.2243066  -1.41126851
##   18 -2.0941926  -1.14406161
##   19 -3.9619959  10.93688543
##   20 -5.7822558  20.02541005
##   21 -2.7069996   1.51874767
##   22 -4.5460985 -20.23663978
##   23 -4.6339794 -19.70064967
##   24 -5.0837478  18.35460655
##   25 -6.3912579  21.77794225
##   26 -2.4987925  -1.36294599
##   27 -0.7630410  -0.97955639
##   28 -2.4679625  -1.29620098
##   29 -3.5663162   5.78104350
##   30 -0.8545311  -0.82763338
##   31 -3.5666936   2.97504176
##   32 -3.0605973  -1.84213290
##   33 -2.8734144  -1.91907110
##   34 -2.7793157   6.94638095
##   35 -3.4356561   4.97390478
##   36 -1.7395259  -1.74223700
##   37 -1.8578260  -3.26979697
##   38 -1.9805015  -3.36199603
##   39 -3.0129042  -9.83007439
##   40 -3.4184641   4.04162054
##   41 -2.4400500   6.59375796
##   42 39.1278630   0.87905485
##   43  1.3126678   1.32830728
##   44 37.1600295   0.87821928
##   45 39.2638325   0.90544387
##   46 -2.0352303   1.53092895
```

```
##   47 -1.4840336   3.45728883
##   48 -1.7079088   2.87654917
##   49 -2.4095578   3.61994954
##   50  0.0732701   0.07096812
```

```
plot(pca_aaron$x[,1:2], xlab="PCA 1 direction", ylab="PCA 2 direction", bty="n",
     type='n', )
text(pca_aaron$x[,1:2], labels = 1:length(aaron), cex=1)
```




```
# all about the federal reserve
content(aaron[[25]])
```

```
##  [1] "The Federal Reserve took another step Friday toward eliminating the barriers between banking a
##  [2] "The Fed Board of Governors voted unanimously to increase the percentage of revenue bank subsidi
##  [3] "Bank subsidiaries have been bumping against the 10 percent limit and earlier this year the Fed
##  [4] "Banks were elated by the new rule."
##  [5] "\"This is a tremendous boost for banks that already underwrite securities and an incentive for
##  [6] "Securities firms, however, were not pleased."
##  [7] "\"It's as if Santa Clause delivered all the gifts to only one house in the neighbourhood,\" sa
##  [8] "For decades, securities underwriting was the exclusive purview of firms like Merrill Lynch &amp
##  [9] "While the 1933 Glass-Steagall Act generally prohibits banks from engaging in non-bank activity
## [10] "The Fed's action will add to pressure on Congress, which was already planning to review the en
## [11] "Securities firms, insurance companies and others who opposed lifting restriction on banks in t
## [12] "Raising the revenue cap, and an announcement last month by the Comptroller of the Currency all
## [13] "\"It's not a two-way street. Other industries will want to level the playing field,\" he said.
```

```
## [14] "At Friday's meeting, Fed governors called for Congress to follow their lead and and revise the
## [15] "\"We hope that the next move will be up to the Congress in this whole area,\" Fed Vice Chairwo
```

```
content(aaron[[20]])
```

```
##  [1] "The Federal Reserve took another step Friday toward eliminating the barriers between banking a
##  [2] "The Fed Board of Governors voted unanimously to increase the percentage of revenue bank subsidi
##  [3] "Bank subsidiaries have been bumping against the 10 percent limit and earlier this year the Fed
##  [4] "Banks were elated by the new rule."
##  [5] "\"This is a tremendous boost for banks that already underwrite securities and an incentive for
##  [6] "Securities firms, however, were not pleased."
##  [7] "\"It's as if Santa Clause delivered all the gifts to only one house in the neighbourhood,\" sa
##  [8] "Bruce Thompson, director of government relations at the nation's largest brokerage firm, Merril
##  [9] "\"We hope this does not reduce their incentive,\" he said. \"The time really is coming for havi
## [10] "Officials at leading banks said they still wanted Congress to take more dramatic action."
## [11] "\"We need complete financial reform,\" Rachel Robbins, general counsel of J.P. Morgan &amp;amp; Co
## [12] "For decades, securities underwriting was the exclusive purview of firms like Merrill Lynch and
## [13] "But since 1987, when the Fed first allowed bank subsidiaries to earn 5 percent of their revenue
## [14] "While the 1933 Glass-Steagall Act generally prohibits banks from engaging in non-bank activity
## [15] "The Fed's action will add to pressure on Congress, which was already planning to review the en
## [16] "Securities firms, insurance companies and others who opposed lifting restriction on banks in th
## [17] "Raising the revenue cap, and an announcement last month by the Comptroller of the Currency allo
## [18] "\"It's not a two-way street. Other industries will want to level the playing field,\" he said."
## [19] "At Friday's meeting, Fed governors called for Congress to follow their lead and and revise the
## [20] "\"We hope that the next move will be up to the Congress in this whole area,\" Fed Vice Chairwo
```

```
content(aaron[[24]])
```

```
##  [1] "U.S. banks will be able to significantly increase their securities underwriting activities unde
##  [2] "The Fed voted unanimously to allow so-called section 20 subsidiaries of banks to derive as much
##  [3] "The new rule takes effect in 60 days, the Fed said.  "
##  [4] "At the meeting, Fed governors called for Congressional action to reform the 1933 Glass-Steagall
##  [5] "\"We hope that the next move will be up to the Congress in this whole area,\" Fed vice chair Al
##  [6] "\"Legislation is needed but I think this is a constructive step and it is appropriate for us to
##  [7] "While the Glass-Steagall Act prohibits banks from engaging in non-bank activity, under section
##  [8] "The Fed said it currently authorizes 41 bank subsidiaries to deal and underwrite securities, ra
##  [9] "Banks were already allowed to handle other sorts of financial instruments, including Treasury s
## [10] "Both the Fed's action and an even broader move last month by the Comptroller of the Currency wi
## [11] "In the past, banks led the charge to change the law because they wanted to get into other busin
## [12] "Now, all industries will press for change, Litan said. \"The tables have been turned,\" he said
## [13] "Banking industry representatives said they were elated by the Fed vote.  "
## [14] "\"This is a tremendous boost for banks that already underwrite securities and an incentive for
## [15] "\"By lifting the 10 percent revenue cap to 25 percent for Section 20 ineligible securities, the
## [16] "Securities industry groups had written to the Fed strongly opposing the proposal and restated t
## [17] "\"It's as if Santa Clause delivered all the gifts to only one house in the neighborhood,\" Stev
## [18] "Judge also called for Congress to act. \"Nothing could benefit consumers and our economy more t
## [19] "((202-898-8312))"
```

```
# IPapers about banking and technology intersection.
content(aaron[[39]])
```

```
##  [1] "An international Internet group released its plan Wednesday to dramatically increase the numbe
```

```
##  [2] "The International Ad Hoc Committee, which includes members of Internet standards-setting bodies
##  [3] "If the plan is adopted, Net surfers will see addresses ending in \"web\", \"store\", \"info\",
##  [4] "Each new domain reflects a particular type or category of Internet site. For example, \"rec\" w
##  [5] "In drafting the plan, the group received more than 4,000 comments from around the world. The co
##  [6] "\"We are very pleased with the acceptance and broad consensus that we have achieved in this pro
##  [7] "Heath said the plan should be approved by Internet standard setters within a few weeks, allowin
##  [8] "The plan includes provisions to resolve disputes arising over the use of trademarked names as
##  [9] "Last year, for example, toymaker Hasbro Inc. won a lawsuit to regain control of the Internet ad
## [10] "But more complex disputes are arising where both parties may have a legitimate claim to an Inte
## [11] "Under the committee plan, anyone applying for an Internet address will have to agree to resolve
## [12] "All challenges and proposed decisions would be made public and time allowed for comment before
## [13] "The plan will not completely eliminate court battles, attorneys said. \"It's much better than t
## [14] "The plan also calls for establishing up to 28 competing registration firms to dole out the new
## [15] "Currently, one firm, Herndon, Va.-based Network Solutions Inc., hands out addresses in the most
## [16] "Network Solutions, which registers over 80,000 new Internet addresses a month, had no comment o
## [17] "The complete proposal is posted on the World Wide Web at http://www.iahc.org/draft-iahc-recomme
```

```
content(aaron[[43]])
```

```
##  [1] "Several U.S. legislators are planning to introduce measures soon to relax export restrictions o
##  [2] "Last year, the Clinton administration relaxed some of the Cold War era export limits on the tec
##  [3] "But computer companies said the administration did not go far enough, locking them out of the g
##  [4] "Rep. Bob Goodlatte, Republican of Virginia, plans to reintroduce a bill on Wednesday that would
##  [5] "Goodlatte's bill would allow computer companies like Netscape Communications Corp or Microsoft
##  [6] "In the Senate, Montana Republican Sen Conrad Burns is also planning to reintroduce his encrypti
##  [7] "At an informal gathering of the Congressional Internet Caucus Tuesday night, Burns said he expe
##  [8] "At one time, the Senator had said the 1997 bill would be identical to last year's bill which di
##  [9] "\"It's about ready to go,\" Burns said Tuesday night."
## [10] "Goodlatte's bill, under the jurisdiction of the House Judiciary Committee, had already garnered
## [11] "Staff said the bill almost had the backing of a majority of the members of the Judiciary Commit
## [12] "Others are also considering introducing bills, some seeking a middle approach between the admin
## [13] "Sen. Patrick Leahy, the Vermont Democrat and co-chair of the Internet Caucus, may have a bill a
## [14] "Some lobbyists involved in the debate said that the administration was considering introducing
## [15] "Administration officials were not immediately available for comment."
## [16] "((--202-898-8312))"
```

```
content(aaron[[34]])
```

```
##  [1] "Prospects for comprehensive reform of U.S. banking and financial services laws remain bright de
##  [2] "Most in government and the industry now agree that the 60-year-old Glass-Steagall Act separatin
##  [3] "But there is little consensus about removing the barriers between financial firms and other com
##  [4] "\"The feeling is still very, very good that everyone wants Glass-Steagall reformed,\" former Co
##  [5] "With almost two years left for the 105th Congress, legislators should have sufficient time to c
##  [6] "\"There's tremendous momentum building. Maybe it won't be in 1997 but certainly within the 105t
##  [7] "Federal courts and regulators spurred the momentum last year by lifting many restrictions on ba
##  [8] "As previously reported, Treasury Secretary Robert Rubin is weighing the recommendations of a ta
##  [9] "\"There has been an enormous amount of movement on the whole issue in the last two months,\" a
## [10] "Among various bills in Congress and the administration, the commerce issue \"is the last issue
## [11] "The remaining \"hang-up\" has raised strong opposition from House Banking Committee Chairman Ji
## [12] "\"There is no public support and no economic need for the conglomeration of financial institut
## [13] "If the administration pushes ahead with the Hawke plan, it would create a \"much longer debate
## [14] "Small bankers agree. \"It's a proposal that would essentially change and concentrate not only t
## [15] "Such concentration would be bad for the economy, Guenther added. \"It is based on very debatabl
```

```
## [16] "One bill in Congress includes a possible compromise approach, lobbyists noted. Representative
## [17] "The New Jersey Republican, chairwoman of the House Banking Committee's Financial Institutions
```

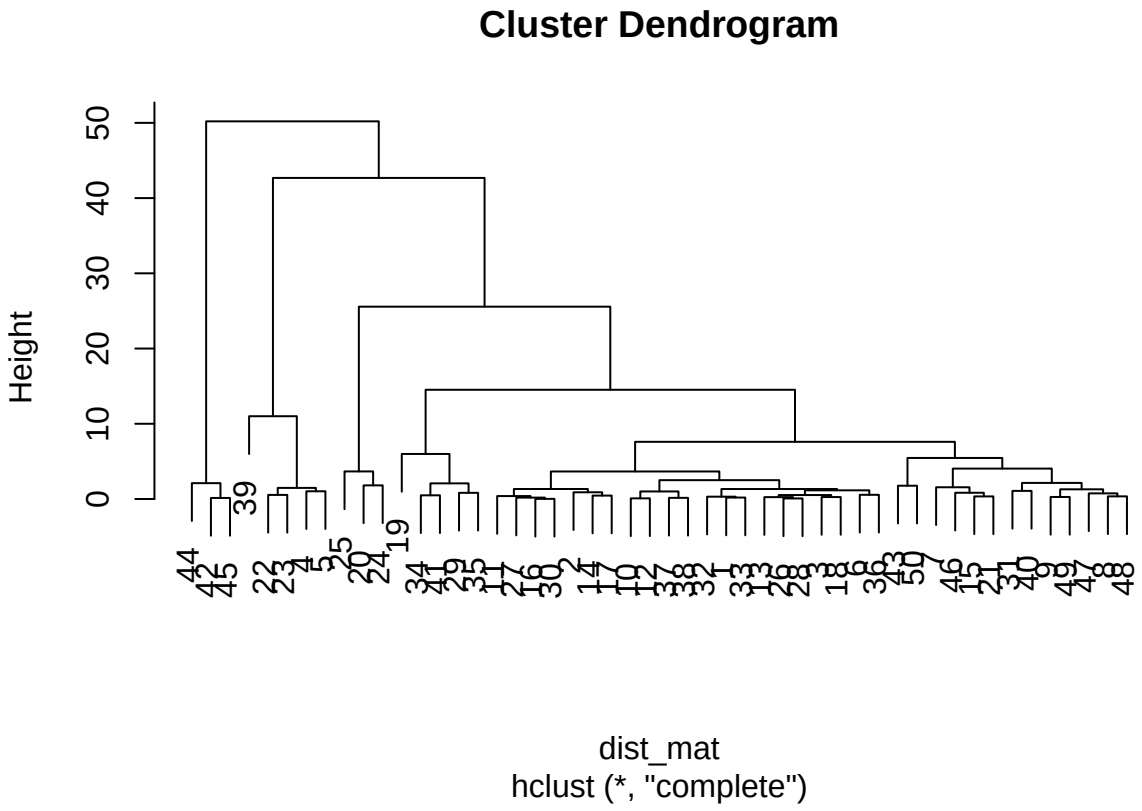
```
### Papers about strictly Technology
content(aaron[[44]])
```

```
##  [1] "Internet access providers and others who met with Net-savvy lawmakers said they fear Congress
##  [2] "Company representatives met Tuesday night at the Capitol with members of the Congressional Int
##  [3] "\"I think it's inevitable that there will be bills introduced that attempt to legislate the co
##  [4] "Evslin said it would be as difficult to keep Congress from making wrong moves as it would be t
##  [5] "Bill Schrader, president of Internet provider PSINet Inc., was more blunt. \"Be very careful o
##  [6] "The members of Congress who attended the meeting pledged to try to protect the Internet."
##  [7] "\"We have a lot of pretty tough work ahead of us,\" Representative Rick White, R-Pa., and a co
##  [8] "Encryption is the use of software to encode and decode information."
##  [9] "On Wednesday, White co-sponsored one of the first Internet-related bills in the 105th Congress
## [10] "Sen. Conrad Burns, also a co-chairman, warned that, despite the caucus, legislation will alway
## [11] "\"The government making policy is never ahead of the curve,\" the Montana Republican said. \"I
## [12] "Last year was a mixed bag for the Internet in Congress."
## [13] "Legislators passed the Communications Decency Act, opposed by online companies and civil liber
## [14] "Sun Microsystems Inc.'s chief scientist, John Gage, fretted that, with only 85 pro-Internet le
## [15] "Companies interested in the Internet offered some conflicting goals for Congress. Jack Valenti
## [16] "\"Congress cannot avoid that all intellectual property has to be protected -- this can't be Do
## [17] "PSINet's Schrader countered that existing laws already protected movies, books and other works
```

```
content(aaron[[45]])
```

```
##  [1] "Internet access providers and others who met with Net-savvy lawmakers said they fear Congress
##  [2] "Company representatives met Tuesday night at the Capitol with members of the Congressional Int
##  [3] "\"I think it's inevitable that there will be bills introduced that attempt to legislate the co
##  [4] "Evslin said it would be as difficult to keep Congress from making wrong moves as it would be t
##  [5] "Bill Schrader, president of Internet provider PSINet Inc., was more blunt. \"Be very careful o
##  [6] "The members of Congress who attended the meeting pledged to try to protect the Internet."
##  [7] "\"We have a lot of pretty tough work ahead of us,\" Representative Rick White, R-Pa., and a co
##  [8] "Encyrption is the use of software to encode and decode information."
##  [9] "Sen. Conrad Burns, also a co-chairman, warned that, despite the caucus, legislation will alway
## [10] "\"The government making policy is never ahead of the curve,\" the Montana Republican said. \"I
## [11] "Last year was a mixed bag for the Internet in Congress."
## [12] "Legislators passed the Communications Decency Act, opposed by online companies and civil liber
## [13] "Sun Microsystems Inc.'s chief scientist, John Gage, fretted that, with only 85 pro-Internet le
## [14] "Companies interested in the Internet offered some conflicting goals for Congress. Jack Valenti
## [15] "\"Congress cannot avoid that all intellectual property has to be protected -- this can't be Do
## [16] "PSINet's Schrader countered that existing laws already protected movies, books and other works
```

```
#####
# Cluster documents
#####
# define the distance matrix
# using the PCA scores
dist_mat = dist(pca_aaron$x)
tree_aaron = hclust(dist_mat)
plot(tree_aaron)
```




```
clust5 = cutree(tree_aaron, k=5)
# inspect the clusters
which(clust5 == 3)
```

```
## 19 29 34 35 41
## 19 29 34 35 41
```

For the reuters corpus text analysis problem, we decided to take a deeper dive into a particularly successful New York Times corpus fortune senior writer, Aaron Pressman. After his successful career reporting at Business Week, The Industry Standard, and Bloomberg, and his SABEW "Best in Business Award", our group was interested in further investigating where Pressman's passions lie within the reporting industry. We wanted to specifically analyze what Pressman's works look like on a deeper level, aling with these areas that have made him so successful in the writing industry. We decided to take a similar approach to that of class in order to further investigate Pressman's works and career. The file includes data relating to the New York Times annotated Corpus with text included of articles written by NYT authors.

In this analysis, we found out information on what Aaron Pressman was writing about, Aaron pressman wrote about a few topics mostly in the Banking and Technology we used principal components analysis to take down the number of variables to only two single variables. This was surprisingly good at taking the overall meaning of his papers. This was seen in the analysis of the content of each of the papers. A couple were strictly banking, most were a mix, and finally some were just tech. Using PCA is a great way to find out the overall meaning of a large set of documents.

Association rule mining

```
library(tidyverse)
library(igraph)
library(arules)  # has a big ecosystem of packages built around it
library(arulesViz)
```

```
groceries <- read.transactions(file="groceries.txt",sep = ',',format="basket",rm.duplicates=TRUE)
summary(groceries)
```

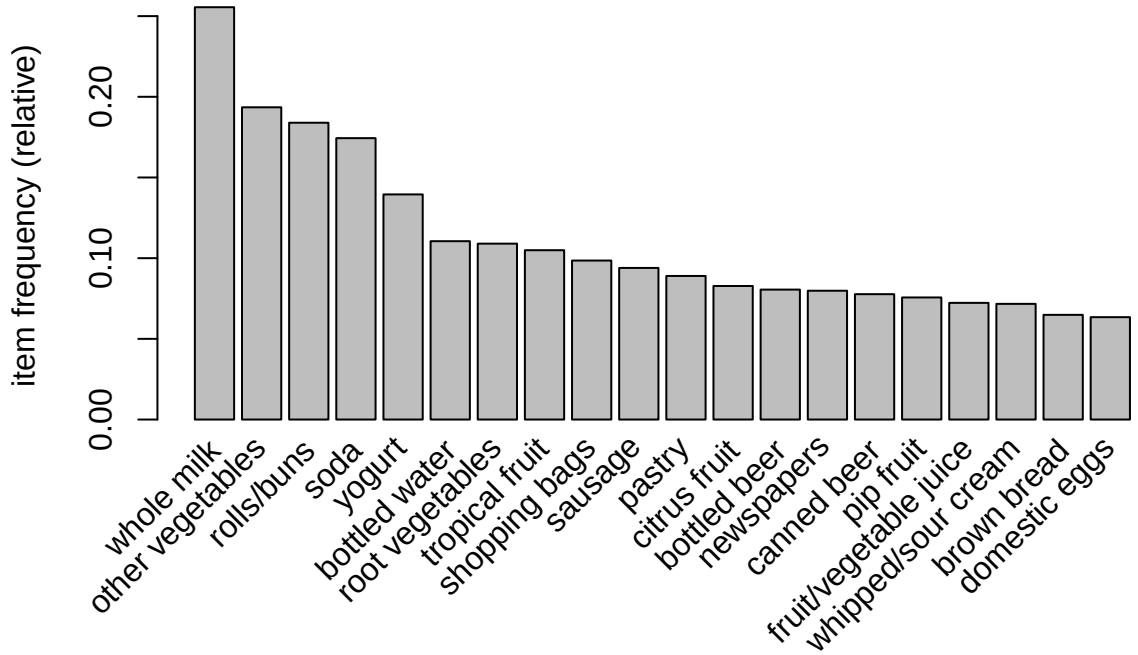
```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
```

```
groc_trans = as(groceries, "transactions")
summary(groc_trans)
```

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
```

```
itemFrequencyPlot(groc_trans, topN = 20)
```



Some Initial findings: There are total of 9835 transactions in our dataset. Whole milk is the most frequent item bought by shoppers, followed by other vegetables, then rolls & buns.

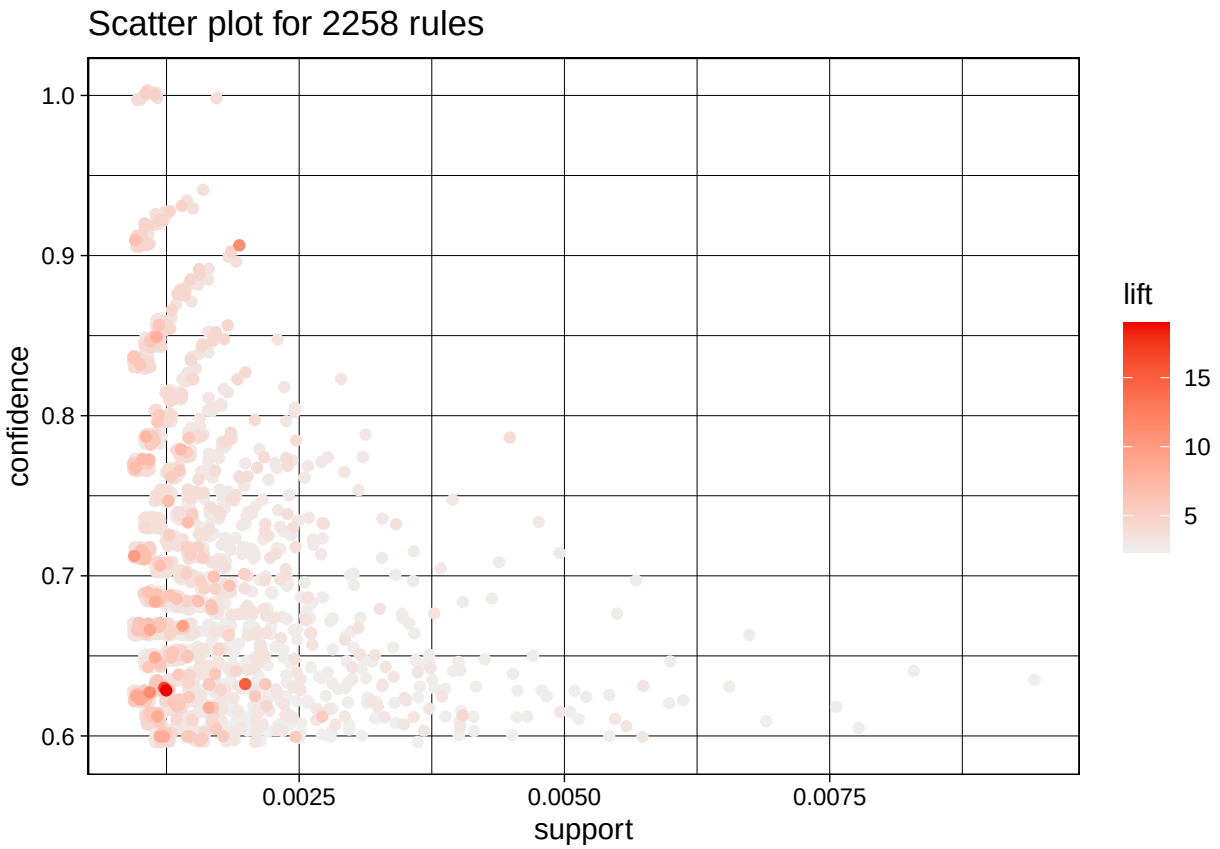
```
groc_rules <- apriori(groc_trans,
                      parameter=list(support=.001, confidence=0.6, maxlen=4)) # rules
```

```
## Apriori
##
## Parameter specification:
```

```
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.6    0.1    1 none FALSE            TRUE       5   0.001      1
##  maxlen target  ext
##       4  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 9
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [157 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4

## Warning in apriori(groc_trans, parameter = list(support = 0.001, confidence
## = 0.6, : Mining stopped (maxlen reached). Only patterns up to a length of 4
## returned!

##  done [0.01s].
## writing ... [2258 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
#arules::inspect(groc_rules)
#arules::inspect(subset(groc_rules, subset=lift > 4))
#arules::inspect(subset(groc_rules, subset=support > 0.002))
#arules::inspect(subset(groc_rules, subset=lift > 5 & confidence > 0.8))
plot(groc_rules)
```

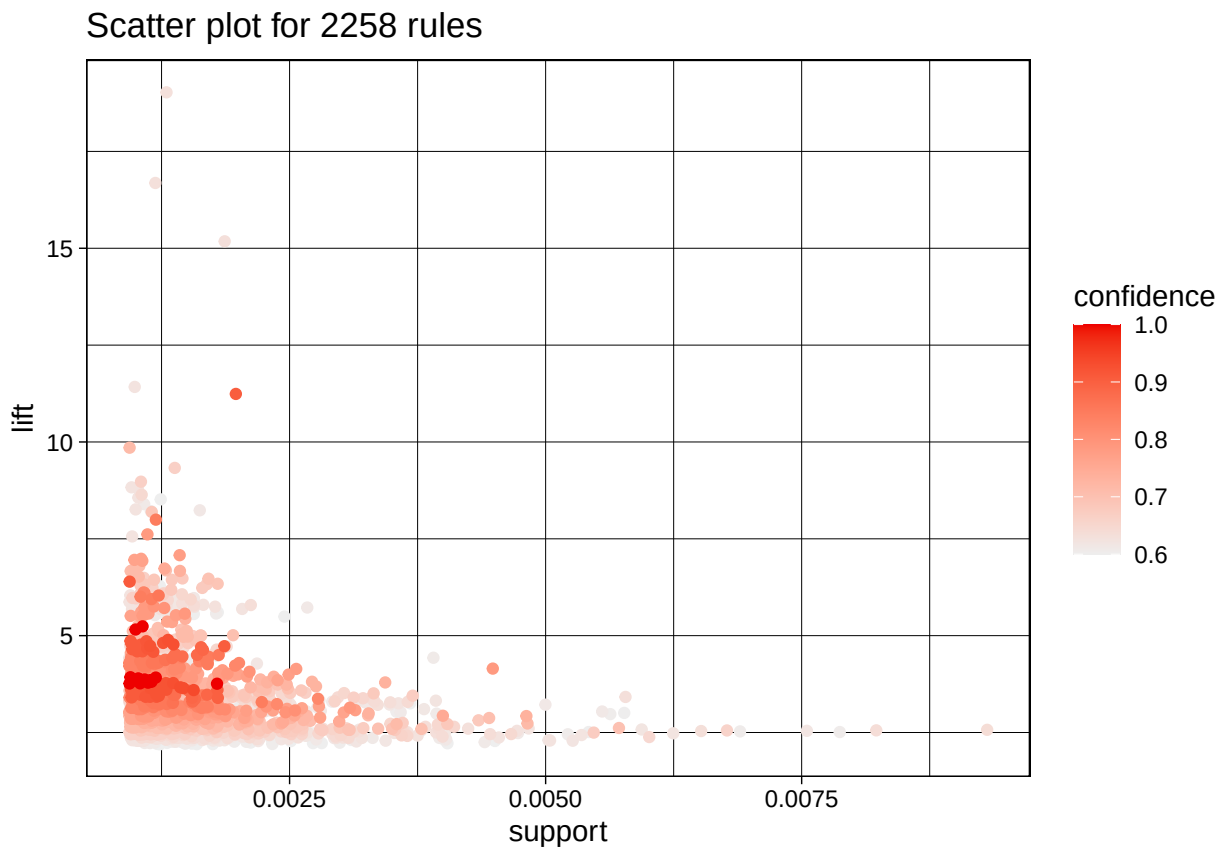
```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



Look at the output... so many rules! There are 2142 rules generated with support at .001, confidence at .6, and max length at 4. I set support=0.001 meaning the RHS appear to be in at least 0.1% of the baskets. I set confidence=0.6 meaning the RHS will be purchased given LHS were purchased 60% of the times. I set length=4 because our grocery baskets have maximum of 4 items. Then I took different subsets. Increasing lift, support, and confidence will all reduce the number of rules. This makes sense because not all rules have high accuracy, occurrence, or impact.

```
plot(groc_rules, measure = c("support", "lift"), shading = "confidence")
```

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




```
plot(groc_rules, method='two-key plot')
```

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

Plot all the rules in (support, confidence) space: we notice that high lift rules tend to have low support.This makes sense intuitively because lift is a fraction of confidence over support. From the two key graph: The more items a rule include, the lower support but higher confidence that rule will have. This makes sense intuitively.

```
groc_graph = associations2igraph(subset(groc_rules, lift>4), associationsAsNodes = FALSE)
igraph::write_graph(groc_graph, file='grocery.graphml', format = "graphml")
```

In the last step I opened the graph in Gephi: From the association graph from Gephi, we can see some of the clustering of associations. For example, Gephi1 shows the beverage purchase community; Gephi2 shows the veggie/fruit purchase community In Gephi3, we can see Root Veggie, Yogurt, Tropical Fruit are nodes with big centers.They are more prevalent in association rule mappings. It means that these products are put in the basket often in combination with of other different products.