ExcercisesSTA380

Zachary Hall

8/17/2020

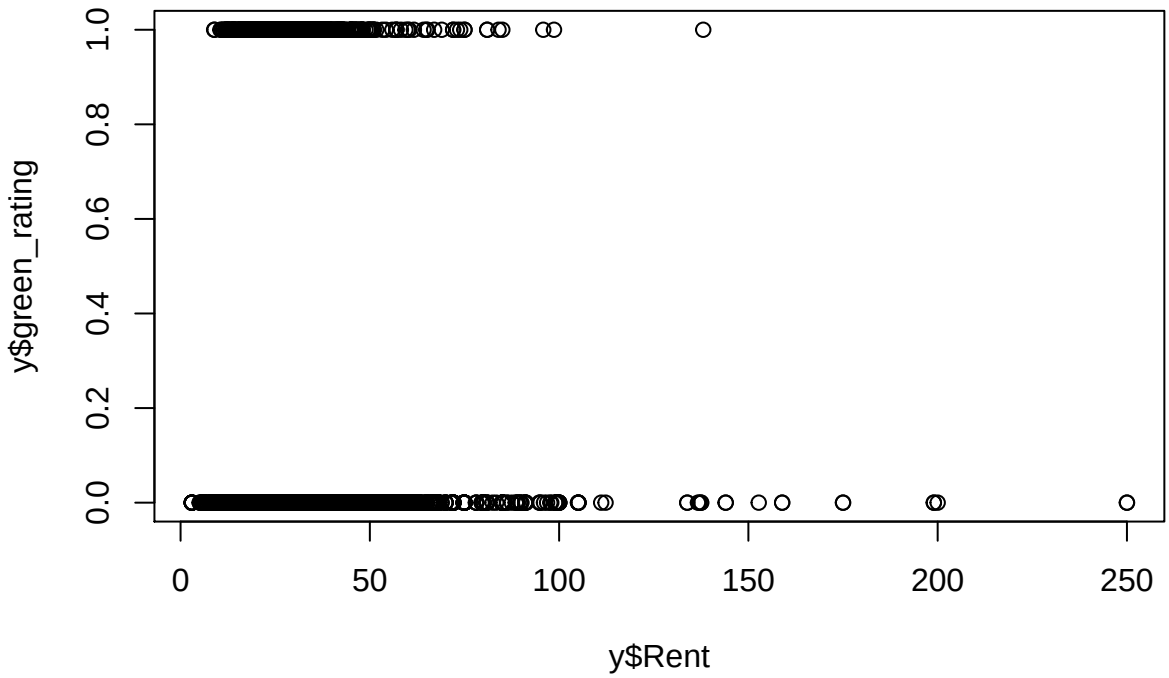
#Problem 1

```
library(RCurl)
x <- getURL('https://raw.githubusercontent.com/jgscott/STA380/master/data/greenbuildings.csv')
y <- read.csv(text = x)
```

```
library(devtools)
```

```
## Loading required package: usethis
```

```
plot(y$Rent,y$green_rating)
```



```r
median(y$Rent)
```

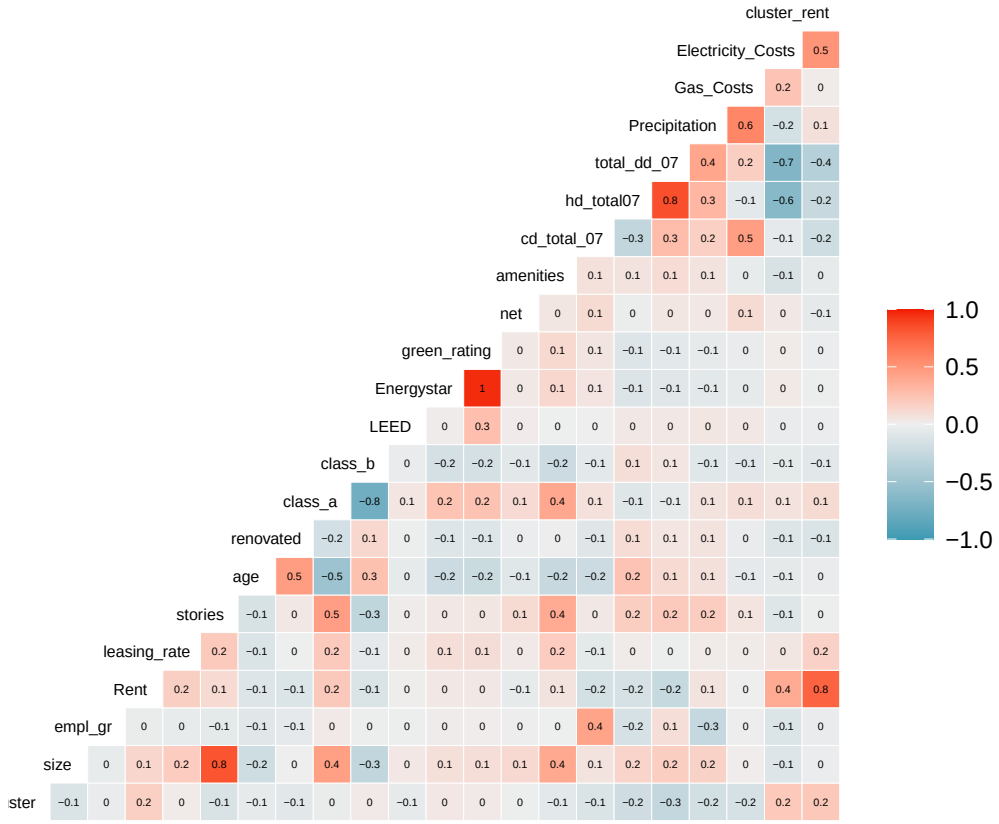
```
## [1] 25.16
```

```r
library(dplyr)
greendata <- as.data.frame(y)
glimpse(greendata)
```

```
## Rows: 7,894
## Columns: 23
## $ CS_PropertyID     <int> 379105, 122151, 379839, 94614, 379285, 94765, 236...
## $ cluster           <int> 1, 1, 1, 1, 1, 1, 6, 6, 6, 6, 6, 6, 6, 6, 6, 8...
## $ size              <int> 260300, 67861, 164848, 93372, 174307, 231633, 210...
## $ empl_gr           <dbl> 2.22, 2.22, 2.22, 2.22, 2.22, 2.22, 4.01, 4.01, 4...
## $ Rent              <dbl> 38.56, 28.57, 33.31, 35.00, 40.69, 43.16, 12.50, ...
## $ leasing_rate      <dbl> 91.39, 87.14, 88.94, 97.04, 96.58, 92.74, 94.33, ...
## $ stories           <int> 14, 5, 13, 13, 16, 14, 11, 15, 31, 21, 11, 15, 15...
## $ age               <int> 16, 27, 36, 46, 5, 20, 38, 24, 34, 36, 32, 25, 26...
## $ renovated         <int> 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0...
## $ class_a           <int> 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1...
## $ class_b           <int> 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0...
## $ LEED              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Energystar        <int> 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ green_rating      <int> 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ net               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ amenities         <int> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0...
## $ cd_total_07       <int> 4988, 4988, 4988, 4988, 4988, 4988, 2746, 2746, 2...
## $ hd_total07        <int> 58, 58, 58, 58, 58, 58, 1670, 1670, 1670, 1670, 1...
## $ total_dd_07       <int> 5046, 5046, 5046, 5046, 5046, 5046, 4416, 4416, 4...
## $ Precipitation     <dbl> 42.57, 42.57, 42.57, 42.57, 42.57, 42.57, 25.55, ...
## $ Gas_Costs         <dbl> 0.01370000, 0.01373149, 0.01373149, 0.01373149, 0...
## $ Electricity_Costs <dbl> 0.02900000, 0.02904455, 0.02904455, 0.02904455, 0...
## $ cluster_rent      <dbl> 36.78, 36.78, 36.78, 36.78, 36.78, 36.78, 17.50, ...
```

```r
variables <- greendata %>% select(2:23)
```

```r
GGally::ggcorr(variables, hjust = .9,size=2,label=TRUE, label_size = 1.5)
```



+As we can see in the correlation matrix, Rent and green rating have close to no correlation, however things like the local market;s average rent, cost of electricity, heating/cooling days, the class of the building, size, location and the leasing rate tend to affect rent costs much more. Number of stories, age, renovations, and net utilities also affect rent more than green rating.

```
fit <- lm(Rent ~ green_rating, data=variables)
summary(fit)
```

```
##
## Call:
## lm(formula = Rent ~ green_rating, data = variables)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.287  -9.044  -3.267   5.733 221.733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28.2668     0.1775 159.275   <2e-16 ***
## green_rating   1.7493     0.6025   2.903   0.0037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.07 on 7892 degrees of freedom
## Multiple R-squared:  0.001067,   Adjusted R-squared:  0.0009405
## F-statistic:  8.43 on 1 and 7892 DF,  p-value: 0.003701
```

+This shows that the relationship between green rating and rent is statistically significant ($p<0.05$). Running a multilinear regression to see how coefficients change to find confounding variables.

```
multivar_reg <- t(cov(variables$Rent, variables) %*% solve(cov(variables)))
model2 <- lm(Rent ~ size + renovated + leasing_rate+ empl_gr+ stories+ age+ class_a+ class_b+ green_rati
summary(model2)
```

```
##
## Call:
## lm(formula = Rent ~ size + renovated + leasing_rate + empl_gr +
##     stories + age + class_a + class_b + green_rating + net +
##     amenities + cd_total_07 + hd_total07 + total_dd_07 + Precipitation +
##     Gas_Costs + Electricity_Costs + cluster_rent, data = variables)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.877  -3.593  -0.523   2.491 174.500
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.754e+00  9.970e-01  -7.778 8.34e-15 ***
## size               6.720e-06  6.554e-07  10.253  < 2e-16 ***
## renovated         -2.156e-01  2.565e-01  -0.840 0.400711
## leasing_rate       8.846e-03  5.320e-03   1.663 0.096360 .
## empl_gr            6.099e-02  1.693e-02   3.602 0.000317 ***
## stories           -3.680e-02  1.617e-02  -2.276 0.022866 *
## age               -1.277e-02  4.713e-03  -2.709 0.006763 **
## class_a            2.854e+00  4.379e-01   6.517 7.62e-11 ***
## class_b            1.182e+00  3.428e-01   3.449 0.000565 ***
## green_rating       6.709e-01  3.980e-01   1.686 0.091906 .
## net               -2.532e+00  5.930e-01  -4.270 1.97e-05 ***
## amenities          5.890e-01  2.501e-01   2.355 0.018541 *
## cd_total_07       -1.339e-04  1.464e-04  -0.915 0.360107
## hd_total07         5.396e-04  8.944e-05   6.033 1.68e-09 ***
## total_dd_07               NA         NA      NA       NA
## Precipitation      4.430e-02  1.597e-02   2.774 0.005558 **
## Gas_Costs         -3.400e+02  7.606e+01  -4.469 7.96e-06 ***
## Electricity_Costs  1.933e+02  2.489e+01   7.767 9.09e-15 ***
## cluster_rent       1.008e+00  1.402e-02  71.942  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.418 on 7802 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.612,  Adjusted R-squared:  0.6111
## F-statistic: 723.8 on 17 and 7802 DF,  p-value: < 2.2e-16
```

+When all other variables are taken into account, not only are confounding variables found (10% change in the green rating coefficient), but also the p-value shows that green rating does not have a statistically significant correlation with rent ($p>0.05$). This shows that the employee that simply used excel assumed a correlation without taking in the affect that all of the other variables would have on rent. +The best way to adjust for confounding variables would be to either select more random locations or to only do an analysis on the specific area in which we are looking to rent a green office.

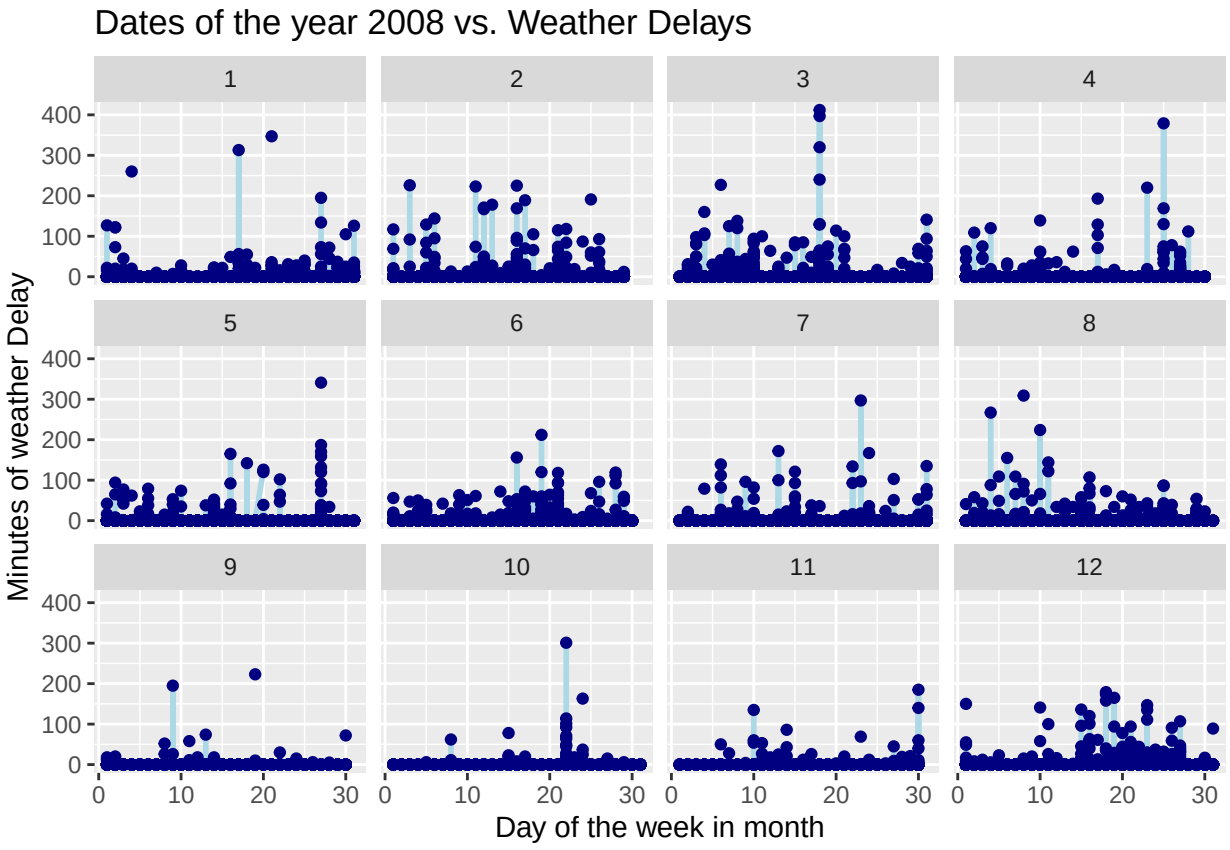
#Problem 2

```
library(tidyverse)
library(knitr)

a <- getURL('https://raw.githubusercontent.com/jgscott/STA380/master/data/ABIA.csv')
b <- read.csv(text = a)
summary(b)
```

```
##       Year          Month         DayofMonth      DayOfWeek        DepTime
##  Min.   :2008   Min.   : 1.00   Min.   : 1.00   Min.   :1.000   Min.   :   1
##  1st Qu.:2008   1st Qu.: 3.00   1st Qu.: 8.00   1st Qu.:2.000   1st Qu.: 917
##  Median :2008   Median : 6.00   Median :16.00   Median :4.000   Median :1329
##  Mean   :2008   Mean   : 6.29   Mean   :15.73   Mean   :3.902   Mean   :1329
##  3rd Qu.:2008   3rd Qu.: 9.00   3rd Qu.:23.00   3rd Qu.:6.000   3rd Qu.:1728
##  Max.   :2008   Max.   :12.00   Max.   :31.00   Max.   :7.000   Max.   :2400
##                                                                 NA's   :1413
##    CRSDepTime      ArrTime       CRSArrTime   UniqueCarrier        FlightNum
##  Min.   :  55   Min.   :   1   Min.   :   5   Length:99260       Min.   :   1
##  1st Qu.: 915   1st Qu.:1107   1st Qu.:1115   Class :character   1st Qu.: 640
##  Median :1320   Median :1531   Median :1535   Mode  :character   Median :1465
##  Mean   :1320   Mean   :1487   Mean   :1505                      Mean   :1917
##  3rd Qu.:1720   3rd Qu.:1903   3rd Qu.:1902                      3rd Qu.:2653
##  Max.   :2346   Max.   :2400   Max.   :2400                      Max.   :9741
##                 NA's   :1567
##    TailNum          ActualElapsedTime CRSElapsedTime     AirTime
##  Length:99260       Min.   : 22.0     Min.   : 17.0   Min.   :  3.00
##  Class :character   1st Qu.: 57.0     1st Qu.: 58.0   1st Qu.: 38.00
##  Mode  :character   Median :125.0     Median :130.0   Median :105.00
##                     Mean   :120.2     Mean   :122.1   Mean   : 99.81
##                     3rd Qu.:164.0     3rd Qu.:165.0   3rd Qu.:142.00
##                     Max.   :506.0     Max.   :320.0   Max.   :402.00
##                     NA's   :1601      NA's   :11      NA's   :1601
##     ArrDelay          DepDelay          Origin              Dest
##  Min.   :-129.000   Min.   :-42.000   Length:99260       Length:99260
##  1st Qu.:  -9.000   1st Qu.: -4.000   Class :character   Class :character
##  Median :  -2.000   Median :  0.000   Mode  :character   Mode  :character
##  Mean   :   7.065   Mean   :  9.171
##  3rd Qu.:  10.000   3rd Qu.:  8.000
##  Max.   : 948.000   Max.   :875.000
##  NA's   :1601       NA's   :1413
##     Distance        TaxiIn           TaxiOut         Cancelled
##  Min.   :  66   Min.   :  0.000   Min.   :  1.00   Min.   :0.00000
##  1st Qu.: 190   1st Qu.:  4.000   1st Qu.:  9.00   1st Qu.:0.00000
##  Median : 775   Median :  5.000   Median : 12.00   Median :0.00000
##  Mean   : 705   Mean   :  6.413   Mean   : 13.96   Mean   :0.01431
##  3rd Qu.:1085   3rd Qu.:  7.000   3rd Qu.: 16.00   3rd Qu.:0.00000
##  Max.   :1770   Max.   :143.000   Max.   :305.00   Max.   :1.00000
##                 NA's   :1567      NA's   :1419
##  CancellationCode      Diverted         CarrierDelay     WeatherDelay
##  Length:99260       Min.   :0.000000   Min.   :  0.00   Min.   :  0.00
##  Class :character   1st Qu.:0.000000   1st Qu.:  0.00   1st Qu.:  0.00
##  Mode  :character   Median :0.000000   Median :  0.00   Median :  0.00
##                     Mean   :0.001824   Mean   : 15.39   Mean   :  2.24
```

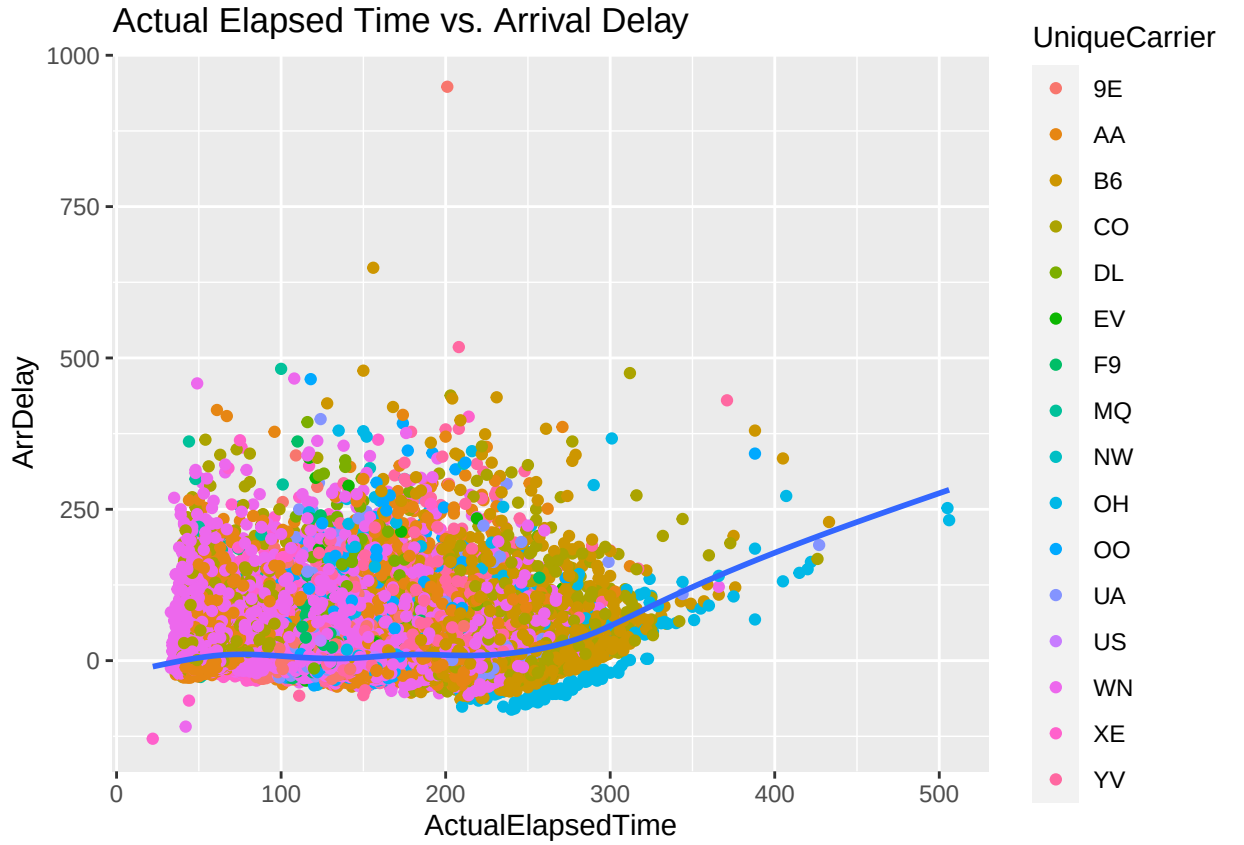
```
##                    3rd Qu.:0.000000   3rd Qu.: 16.00   3rd Qu.:  0.00
##                    Max.   :1.000000   Max.   :875.00   Max.   :412.00
##                                       NA's   :79513    NA's   :79513
##     NASDelay      SecurityDelay    LateAircraftDelay
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Median :  2.00   Median :  0.00   Median :  6.00
##  Mean   : 12.47   Mean   :  0.07   Mean   : 22.97
##  3rd Qu.: 16.00   3rd Qu.:  0.00   3rd Qu.: 30.00
##  Max.   :367.00   Max.   :199.00   Max.   :458.00
##  NA's   :79513    NA's   :79513    NA's   :79513
```

```
ggplot(data = b, aes(DayofMonth, WeatherDelay)) +
  geom_line(color = "lightblue", size = 1) +
  geom_point(color = "navyblue") +
  labs(title = "Dates of the year 2008 vs. Weather Delays",
       y = "Minutes of weather Delay", x = "Day of the week in month") +
  facet_wrap(~ Month)
```



+As we can see in the plot, we can predict days of worse weather which cause delays, like in the winter months there are probably more delays from cities coming from the north. Mixed in are days that could be violent thunderstroms throughout the year in Austin or other cities.

```
ggplot(b, aes(ActualElapsedTime, ArrDelay)) +
  geom_point(aes(color = UniqueCarrier)) +
  geom_smooth(se = FALSE) +
  labs(title = "Actual Elapsed Time vs. Arrival Delay")
```



+This shows us that if a flight is 250 minutes or less, it is on average going to arrive on time. As the elapsed time increases after 250, the arrival delay does as well which makes sense as planes may have to take alternate routes in inclement weather which would take more time. This chart also shows us that OH flights tend to be the most common to spend a long time in the air coinciding with a delay.

#Problem 3

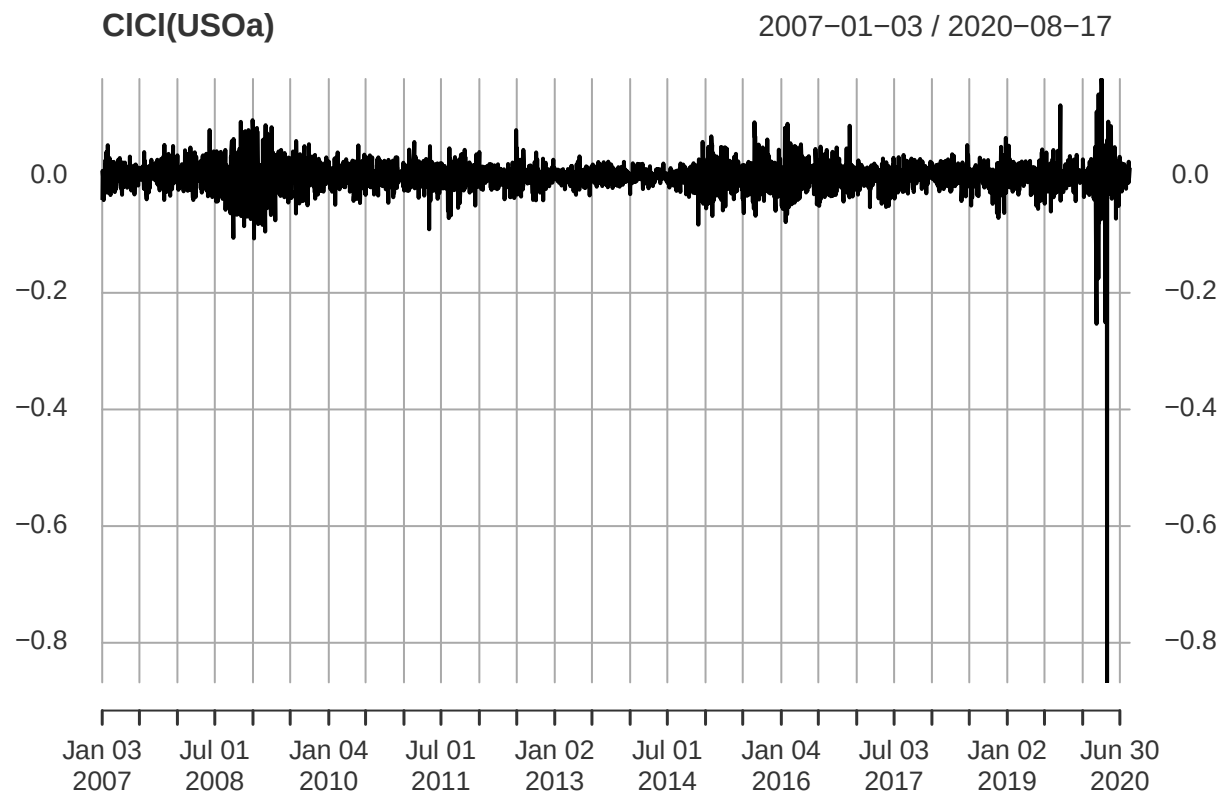
```
library(mosaic)
library(quantmod)
library(foreach)

# Import a few stocks
mystocks = c("USO", "GLD", "BND","HYG","XOP","XLE","VNQ","SHY","AOR","AOM","AOA","AOK","SWAN")
getSymbols(mystocks)
```

```
##  [1] "USO"  "GLD"  "BND"  "HYG"  "XOP"  "XLE"  "VNQ"  "SHY"  "AOR"  "AOM"
## [11] "AOA"  "AOK"  "SWAN"
```

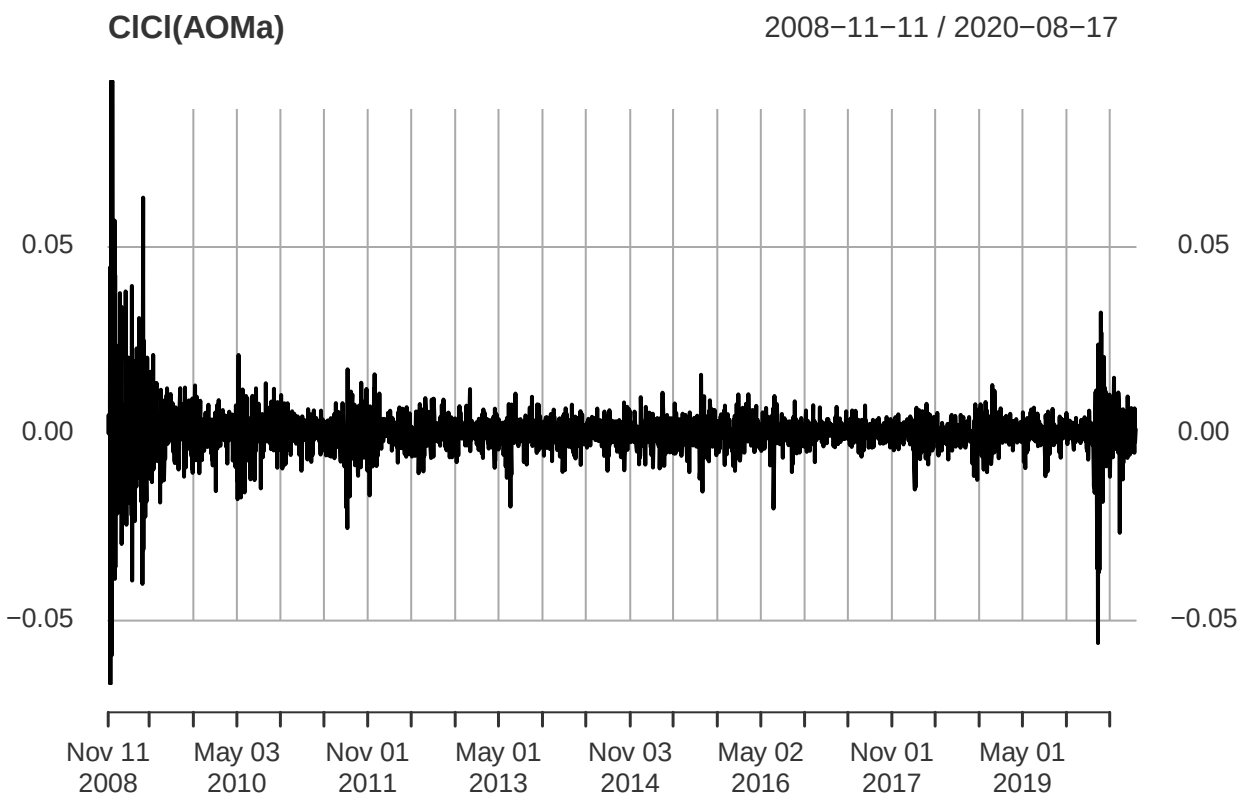
```
USOa = adjustOHLC(USO)
GLDa = adjustOHLC(GLD)
BNDa = adjustOHLC(BND)
HYGa = adjustOHLC(HYG)
XOPa = adjustOHLC(XOP)
XLEa = adjustOHLC(XLE)
```

```
VNQa = adjustOHLC(VNQ)
SHYa = adjustOHLC(SHY)
AORa = adjustOHLC(AOR)
AOMa = adjustOHLC(AOM)
AOAa = adjustOHLC(AOA)
AOKa = adjustOHLC(AOK)
SWANa = adjustOHLC(SWAN)
```

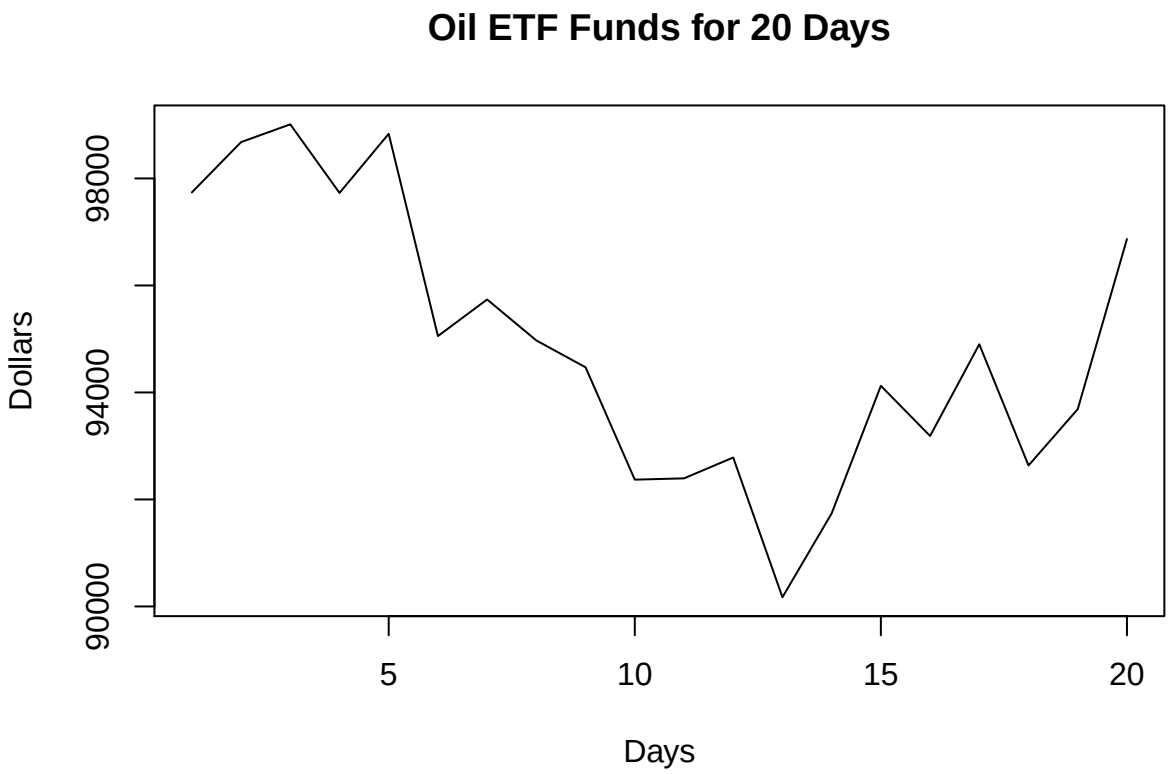
```
plot(ClCl(USOa))
```



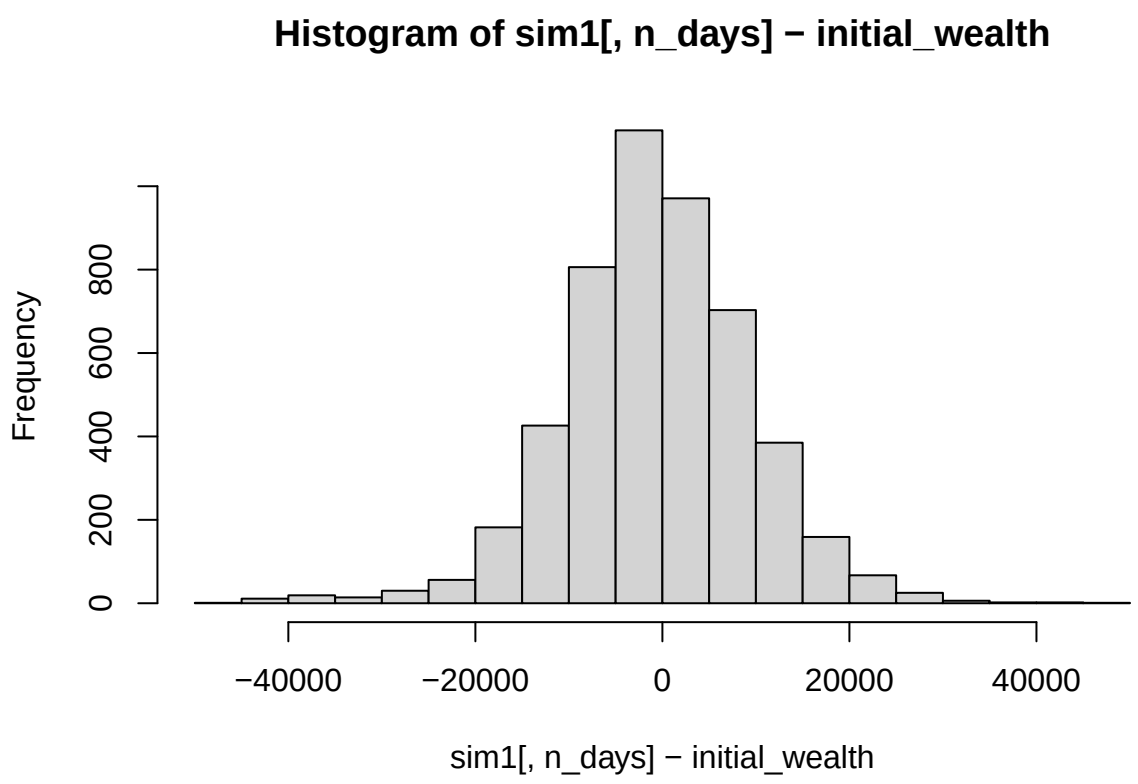

```
plot(ClCl(AOMa))
```




```
oil_returns = cbind(ClCl(USOa),ClCl(XOPa),ClCl(XLEa))
oil_returns = as.matrix(na.omit(oil_returns))
N = nrow(oil_returns)
initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  total_wealth = initial_wealth
  weights = c(0.5, 0.25, 0.25)
  holdings = weights * total_wealth
  n_days = 20
  wealthtracker = rep(0, n_days)
  for(today in 1:n_days) {
    return.today = resample(oil_returns, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
  }
  wealthtracker
}
plot(wealthtracker, type='l',main = 'Oil ETF Funds for 20 Days',xlab = 'Days',ylab = 'Dollars')
```



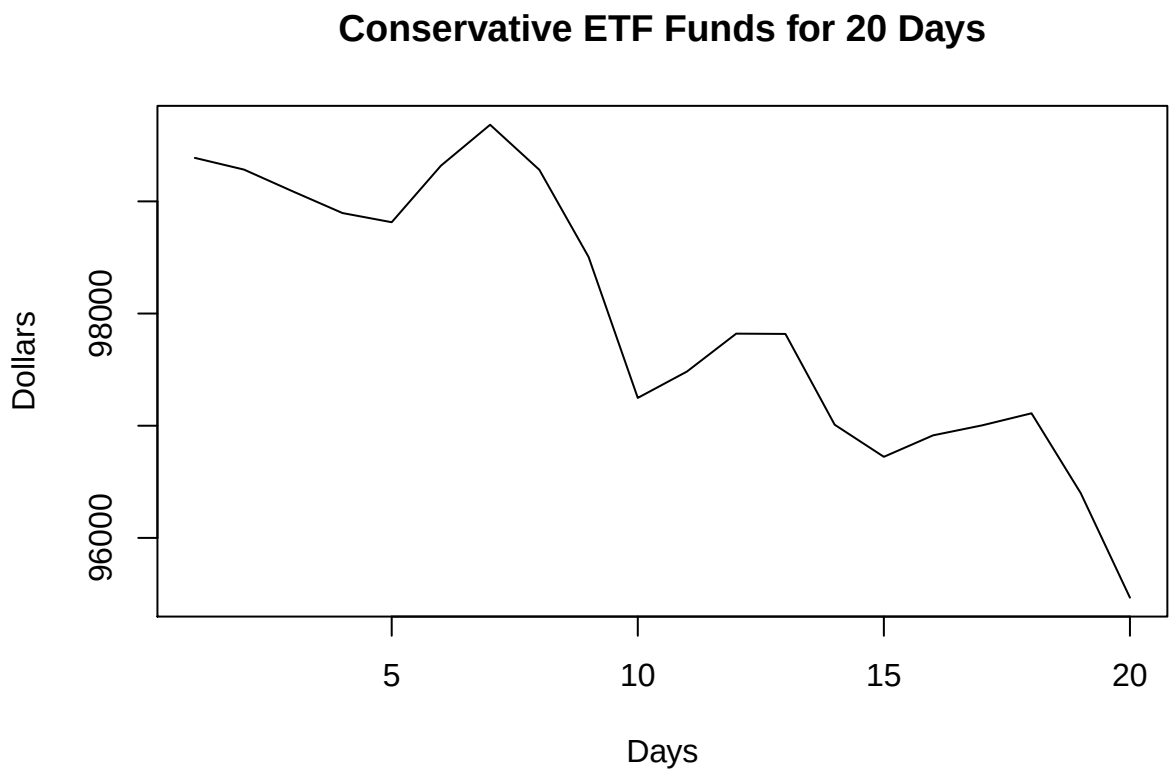

```
hist(sim1[,n_days]- initial_wealth, breaks=30)
```



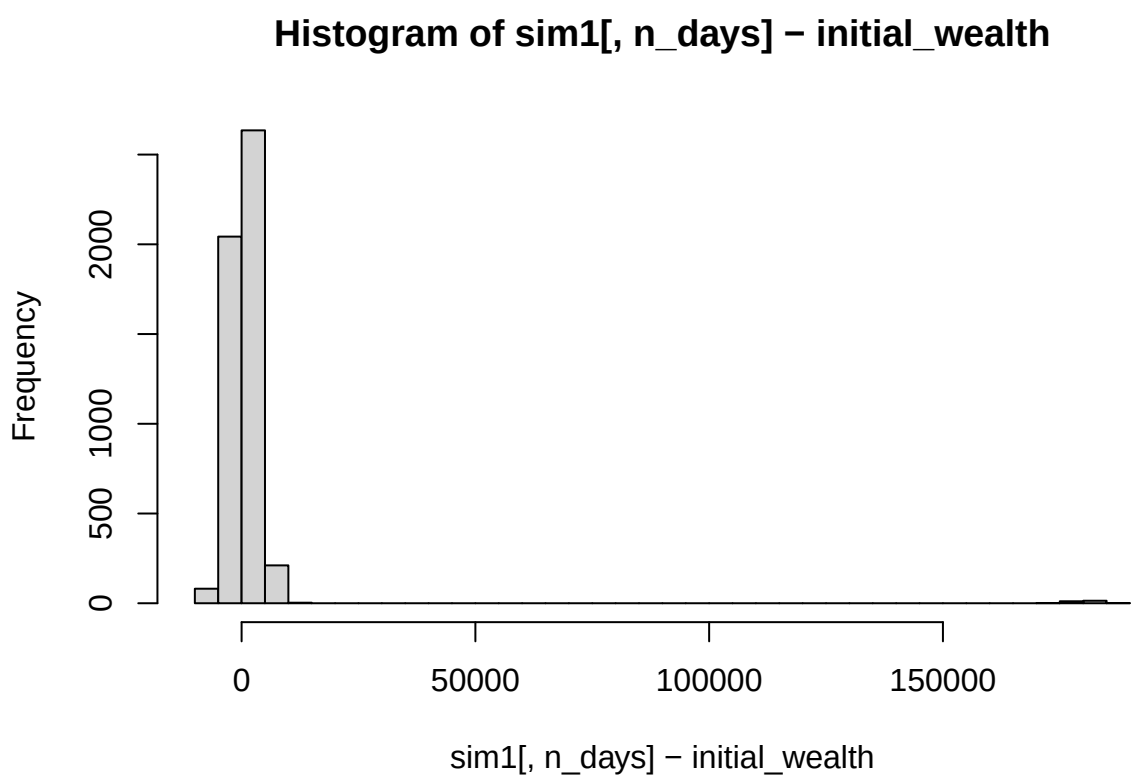

```
# 5% value at risk:
oilVaR = quantile(sim1[,n_days]- initial_wealth, prob=0.05)
oilVaR
```

```
##        5%
## -16044.44
```

```
conservative = cbind(ClCl(GLDa),ClCl(BNDa),ClCl(HYGa),ClCl(VNQa),ClCl(SHYa))
conservative = as.matrix(na.omit(conservative))
N = nrow(conservative)
initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  total_wealth = initial_wealth
  weights = c(0.2, 0.2, 0.2, 0.2, 0.2)
  holdings = weights * total_wealth
  n_days = 20
  wealthtracker = rep(0, n_days)
  for(today in 1:n_days) {
    return.today = resample(conservative, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
  }
  wealthtracker
}
plot(wealthtracker, type='l',main = 'Conservative ETF Funds for 20 Days',xlab = 'Days',ylab = 'Dollars')
```



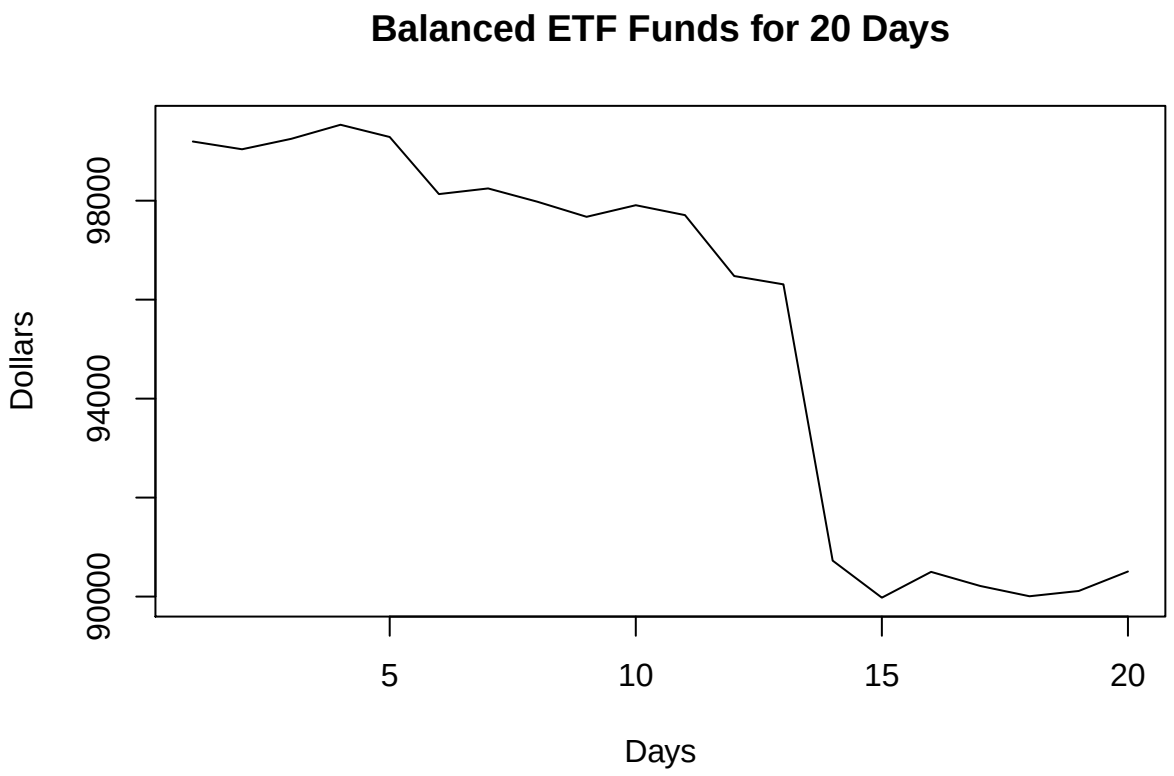

```
hist(sim1[,n_days]- initial_wealth, breaks=30)
```



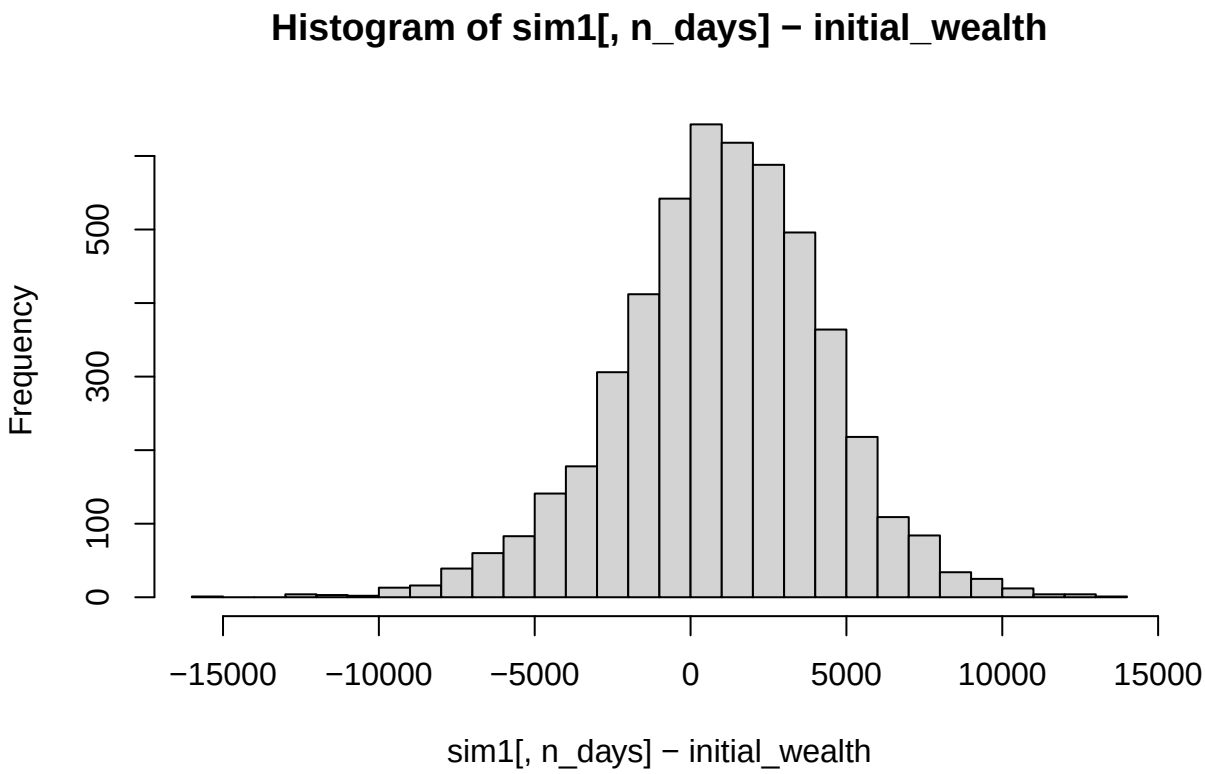

```
# 5% value at risk:
ConservativeVaR = quantile(sim1[,n_days]- initial_wealth, prob=0.05)
ConservativeVaR
```

```
##        5%
## -3689.614
```

```
balanced = cbind(ClCl(AOAa),ClCl(AOKa),ClCl(AOMa),ClCl(AORa),ClCl(SWANa))
balanced = as.matrix(na.omit(balanced))
N = nrow(balanced)
initial_wealth = 100000
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  total_wealth = initial_wealth
  weights = c(0.2, 0.2, 0.2, 0.2, 0.2)
  holdings = weights * total_wealth
  n_days = 20
  wealthtracker = rep(0, n_days)
  for(today in 1:n_days) {
    return.today = resample(balanced, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    total_wealth = sum(holdings)
    wealthtracker[today] = total_wealth
  }
  wealthtracker
}
plot(wealthtracker, type='l',main = 'Balanced ETF Funds for 20 Days',xlab = 'Days',ylab = 'Dollars')
```




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




```
# 5% value at risk:
BalancedVaR = quantile(sim1[,n_days]- initial_wealth, prob=0.05)
BalancedVaR
```

```
##        5%
## -4701.816
```

```
oilVaR
```

```
##        5%
## -16044.44
```

```
ConservativeVaR
```

```
##        5%
## -3689.614
```

```
BalancedVaR
```

```
##        5%
## -4701.816
```

+As we can see from the VaR's printed next to eachother, the best performing collection of ETFs is the conservative one filled with mostly bonds and relatively stable commodities like gold. The portfolio consisting

only of oil/energy based ETFs performed very poorly. The balanced ETF was a little more aggressive than the conservative and therefore lost more money. All three of these are at a loss which may just show how the economy is heading during the coronavirus pandemic.

#Problem 4

```
library(RCurl)
x <- getURL('https://raw.githubusercontent.com/jgscott/STA380/master/data/social_marketing.csv')
social <- read.csv(text = x)
rownames(social) <- social$X
social$X <- NULL
```

```
library(ggplot2)
library(LICORS)  # for kmeans++
library(foreach)
library(mosaic)

X = social[,-(1:9)]
X = scale(X, center=TRUE, scale=TRUE)


mu = attr(X,"scaled:center")
sigma = attr(X,"scaled:scale")

clust1 = kmeans(X, 5, nstart=50)
```

```
clust2 = kmeanspp(X, k=5, nstart=50)

clust2$center[1,]*sigma + mu
```

```
##           family  home_and_garden            music             news
##      0.803516029      0.652533609      0.752843847      1.324715615
##    online_gaming         shopping health_nutrition      college_uni
##      0.866597725      1.489141675     11.723888314      0.987590486
##   sports_playing          cooking              eco        computers
##      0.631851086      3.244053775      0.915201655      0.672182006
##         business         outdoors           crafts       automotive
##      0.472595657      2.715615305      0.576008273      0.748707342
##              art         religion           beauty        parenting
##      0.744570838      0.788004137      0.432264736      0.775594623
##           dating           school personal_fitness          fashion
##      1.089968976      0.600827301      6.260599793      0.788004137
##   small_business             spam            adult
##      0.307135471      0.006204757      0.430196484
```

```
clust2$center[2,]*sigma + mu
```

```
##           family  home_and_garden            music             news
##      0.584314508      0.444621832      0.566154460      1.152065456
##    online_gaming         shopping health_nutrition      college_uni
##      0.594891239      1.284174815      1.058670924      0.944522051
##   sports_playing          cooking              eco        computers
##      0.428257833      0.844941130      0.398722810      0.561564558
```

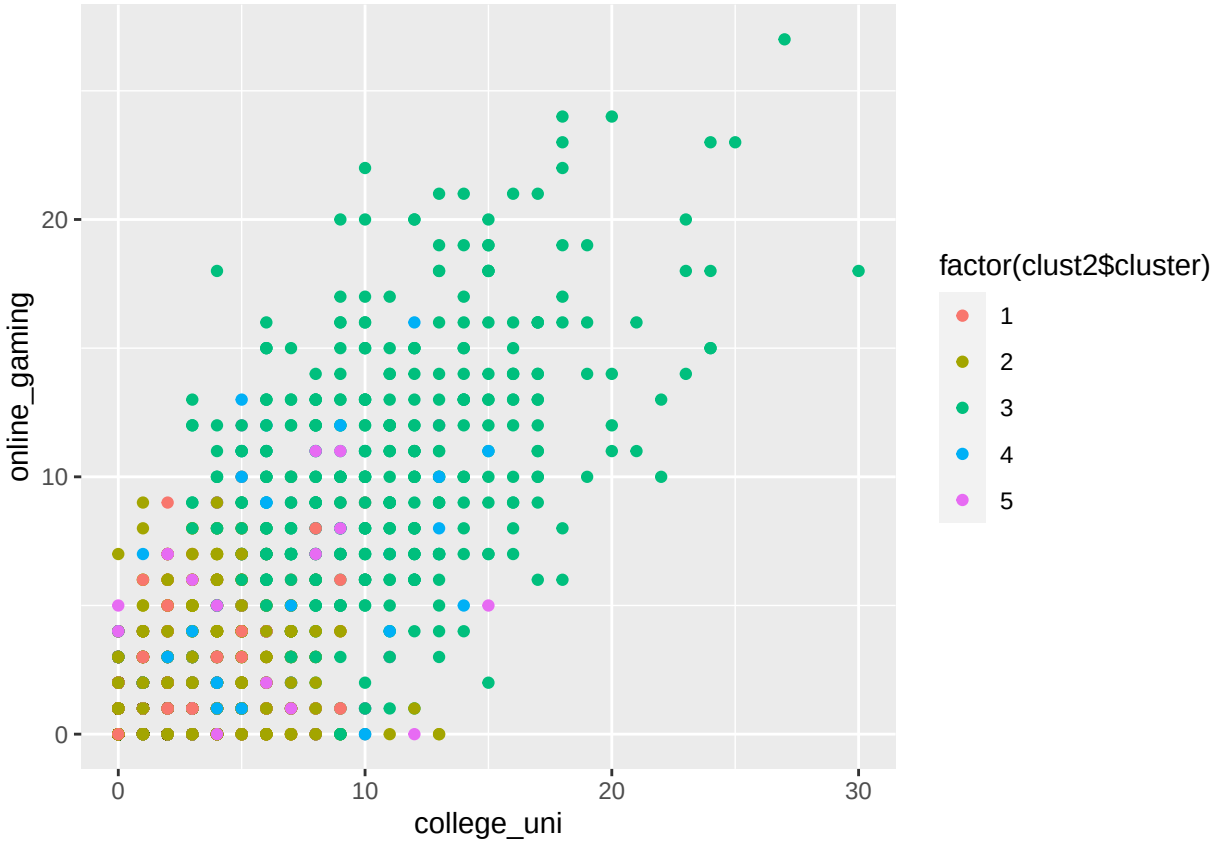
```
##         business         outdoors           crafts       automotive
##      0.365595689      0.435641589      0.377369786      0.758930353
##              art         religion           beauty        parenting
##      0.620235482      0.540211535      0.336858910      0.483336659
##           dating           school personal_fitness          fashion
##      0.529435242      0.456395929      0.644981042      0.503691878
##   small_business             spam            adult
##      0.296747156      0.006984634      0.392935542
```

```
clust2$center[5,]*sigma + mu
```

```
##           family  home_and_garden            music             news
##      2.378874856      0.702640643      0.741676234      1.490241102
##    online_gaming         shopping health_nutrition      college_uni
##      0.901262916      1.585533869      1.724454650      1.225028703
##   sports_playing          cooking              eco        computers
##      0.735935706      1.391504018      0.725602755      1.024110218
##         business         outdoors           crafts       automotive
##      0.588978186      0.670493685      1.138920781      1.194029851
##              art         religion           beauty        parenting
##      0.987370838      4.867967853      1.032146958      3.732491389
##           dating           school personal_fitness          fashion
##      1.278989667      2.704936854      1.115958668      1.000000000
##   small_business             spam            adult
##      0.442020666      0.004592423      0.429391504
```

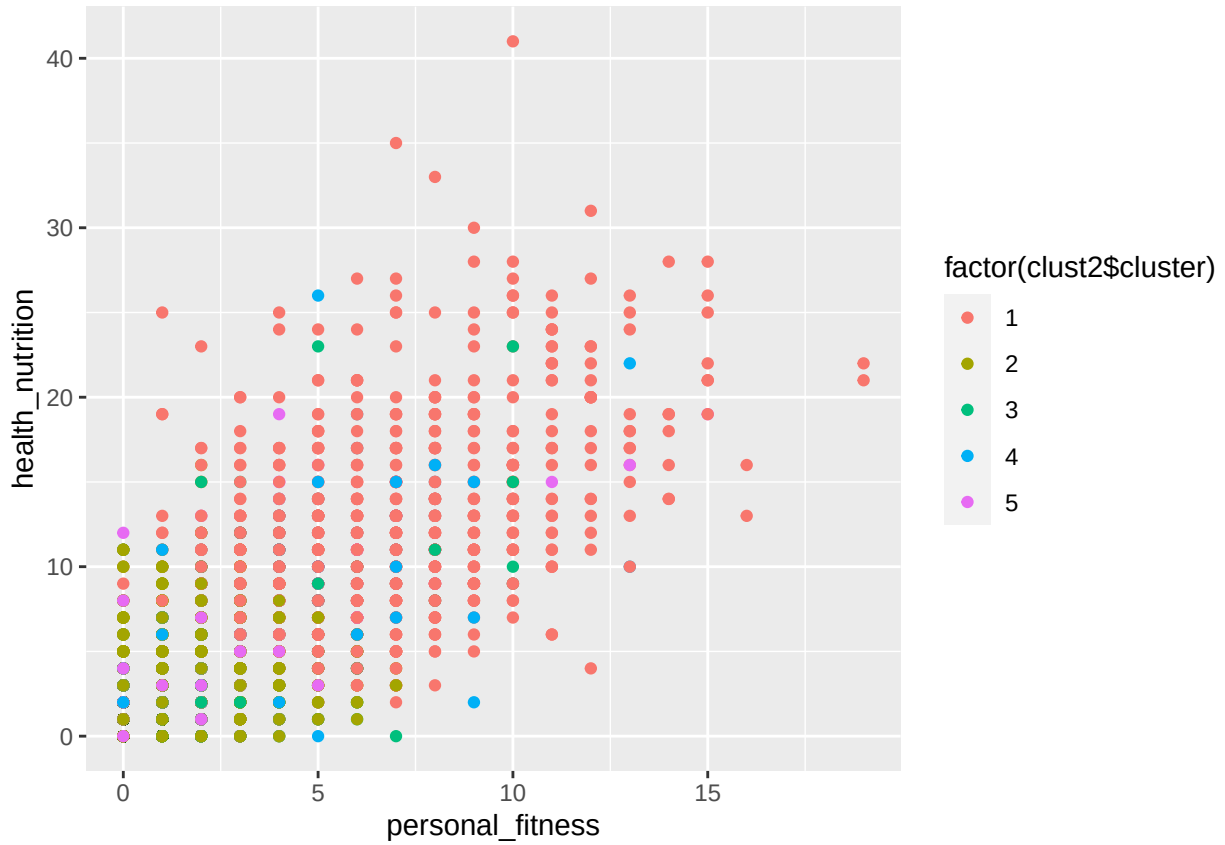
+On this question I ran a kmeans++ model in order to find five different clusters of people in the audience. I inspected their centers and plotted the top qualities of each cluster against eachother (cluster 2 had no significant attributes).

```
qplot(college_uni, online_gaming, data=social, color=factor(clust2$cluster))
```



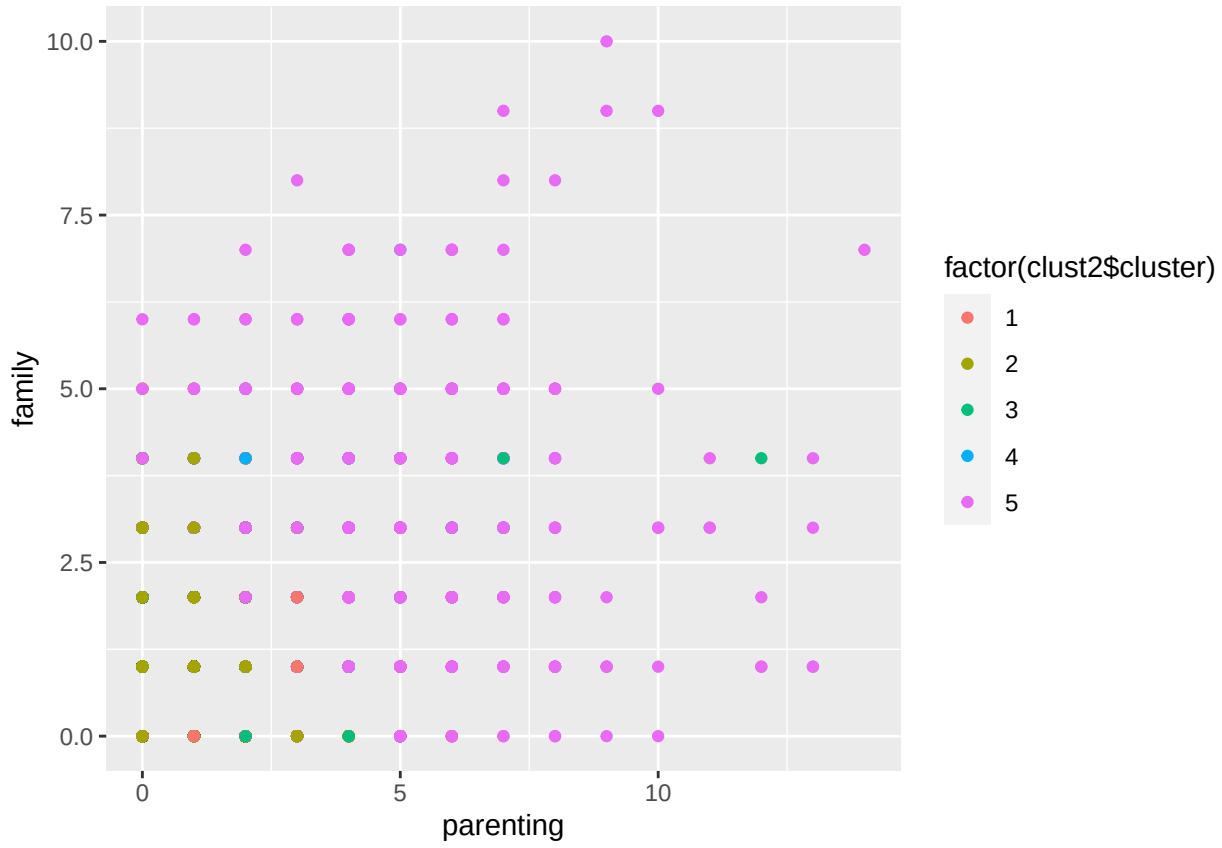
+There is a large section of the products followers that has high scores for college tweets as well as gaming, meaning that the audience of the twitter account has a lot of young people following it. This could be because if the products appeal to young people, the overall demographics of twitter or both.

```
qplot(personal_fitness, health_nutrition, data=social, color=factor(clust2$cluster))
```



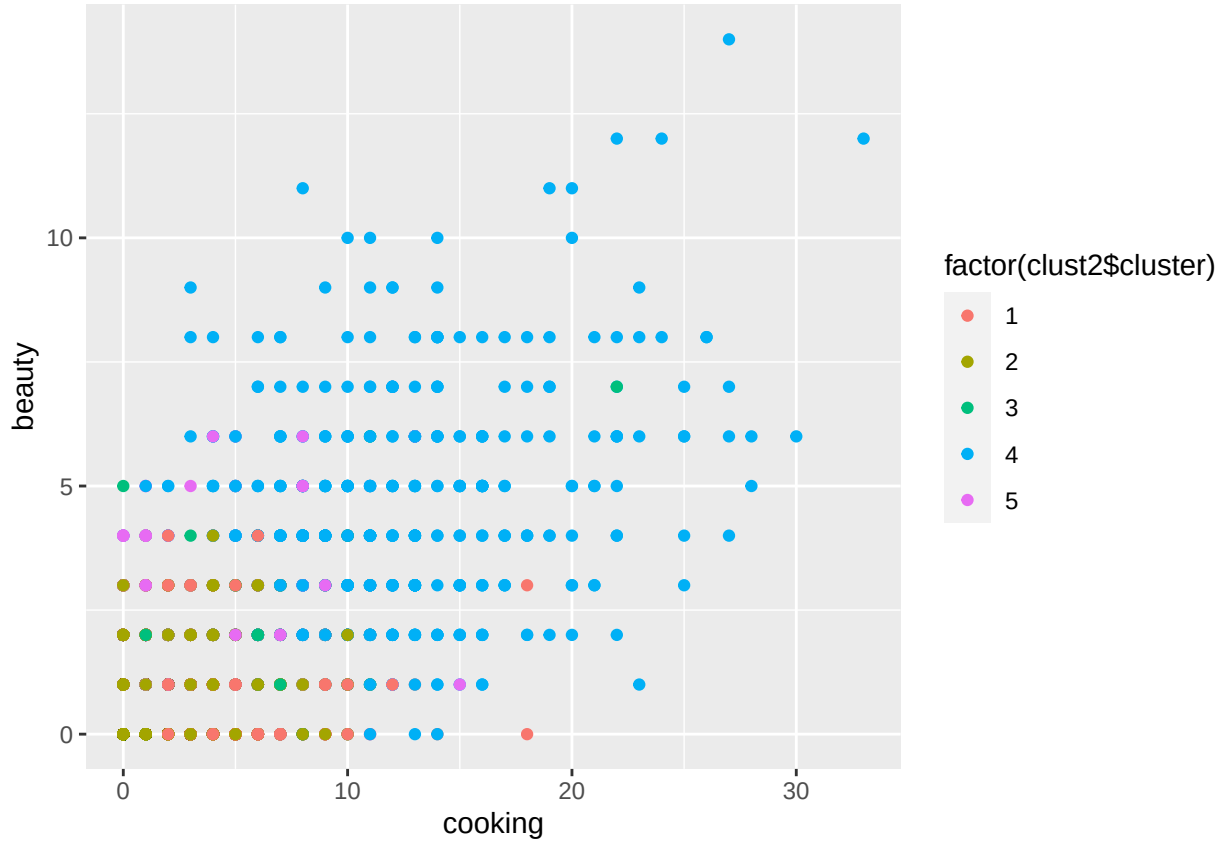
+There is also a cluster that places high values on health, fitness and nutrition meaning that this product is most likely healthy and reaching the people that care about health.

```
qplot(parenting, family, data=social, color=factor(clust2$cluster))
```



+Though not as strong of a correlation, there is still a good portion of followers who care about family and parenting.

```
qplot(cooking, beauty, data=social, color=factor(clust2$cluster))
```



+Another cluster seems to care most about cooking and beauty. +My analysis is that this company should be targeting college kids, parents, people who like to cook, and people who care a lot about health and fitness.

#Problem 5

```
library(tm)
library(tidyverse)
library(slam)
library(proxy)
readerPlain = function(fname){
  readPlain(elem=list(content=readLines(fname)),
            id=fname, language='en') }

file_list = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/SimonCowell/*.txt')
file_list1 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/AaronPressman/*.txt')
file_list2 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/AlanCrosby/*.txt')
file_list3 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/AlexanderSmith/*.txt')
file_list4 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/BenjaminKangLim/*.txt')
file_list5 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/BernardHickey/*.txt')
file_list6 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/BradDorfman/*.txt')
file_list7 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/DarrenSchuettler/*.txt')
file_list8 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/DavidLawder/*.txt')
file_list9 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/EdnaFernandes/*.txt')
file_list10 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/EricAuchard/*.txt')
file_list11 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/FumikoFujisaki/*.txt')
file_list12 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/GrahamEarnshaw/*.txt')
```

```
file_list13 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/HeatherScoffield/*.txt')
file_list14 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/JanLopatka/*.txt')
file_list15 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/JaneMacartney/*.txt')
file_list16 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/JimGilchrist/*.txt')
file_list17 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/JoWinterbottom/*.txt')
file_list18 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/JoeOrtiz/*.txt')
file_list19 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/JohnMastrini/*.txt')
file_list20 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/JonathanBirt/*.txt')
file_list21 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/KarlPenhaul/*.txt')
file_list22 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/KeithWeir/*.txt')
file_list23 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/KevinDrawbaugh/*.txt')
file_list24 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/KevinMorrison/*.txt')
file_list25 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/KirstinRidley/*.txt')
file_list26 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/KouroshKarimkhany/*.txt
file_list27 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/LydiaZajc/*.txt')
file_list28 = Sys.glob("/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/LynneO'Donnell/*.txt")
file_list29 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/LynnleyBrowning/*.txt')
file_list30 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/MarcelMichelson/*.txt')
file_list31 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/MarkBendeich/*.txt')
file_list32 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/MartinWolk/*.txt')
file_list33 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/MatthewBunce/*.txt')
file_list34 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/MichaelConnor/*.txt')
file_list35 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/MureDickie/*.txt')
file_list36 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/NickLouth/*.txt')
file_list37 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/PatriciaCommins/*.txt')
file_list38 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/PeterHumphrey/*.txt')
file_list39 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/PierreTran/*.txt')
file_list40 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/RobinSidel/*.txt')
file_list41 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/RogerFillion/*.txt')
file_list42 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/SamuelPerry/*.txt')
file_list43 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/SarahDavison/*.txt')
file_list44 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/ScottHillis/*.txt')
file_list45 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/TanEeLyn/*.txt')
file_list46 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/TheresePoletti/*.txt')
file_list47 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/TimFarrand/*.txt')
file_list48 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/ToddNissen/*.txt')
file_list49 = Sys.glob('/Users/zachhall/Rstudio/STA380/data/ReutersC50/C50train/WilliamKazer/*.txt')
files = c(file_list, file_list1,file_list10,file_list11,file_list12,file_list13,file_list14,file_list15
readfiles = lapply(files, readerPlain)
```

+Here I read in all of the files manually as I couldn't find a for loop that worked, merged them all together and read them in.

```
mynames = files %>%
    { strsplit(., '/', fixed=TRUE) } %>%
    { lapply(., tail, n=2) } %>%
    { lapply(., paste0, collapse = '') } %>%
    unlist
names(readfiles) = mynames
documents_raw = Corpus(VectorSource(readfiles))
my_documents = documents_raw %>%
  tm_map(content_transformer(tolower))  %>%             # make everything lowercase
  tm_map(content_transformer(removeNumbers)) %>%        # remove numbers
```

```
  tm_map(content_transformer(removePunctuation)) %>%    # remove punctuation
  tm_map(content_transformer(stripWhitespace))          # remove excess white-space
```

+Altered the documents here for easier processing

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

```
## <<DocumentTermMatrix (documents: 2500, terms: 32570)>>
## Non-/sparse entries: 537861/80887139
## Sparsity           : 99%
## Maximal term length: 67
## Weighting          : term frequency (tf)
```

```
inspect(DTM_quotes[1:10,1:20])
```

```
## <<DocumentTermMatrix (documents: 10, terms: 20)>>
## Non-/sparse entries: 37/163
## Sparsity           : 82%
## Maximal term length: 11
## Weighting          : term frequency (tf)
## Sample             :
##     Terms
## Docs act activity ago agreed also alternative amid andersen authorised average
##   1    1        1   1      1    3           2    1        2          1       1
##   10   0        0   0      0    2           0    0        0          0       0
##   2    0        1   0      1    0           0    0        0          0       0
##   3    0        0   0      0    2           0    0        0          0       0
##   4    0        0   0      0    2           0    0        0          0       0
##   5    1        0   0      0    1           0    0        0          1       0
##   6    0        0   0      0    4           0    1        0          0       0
##   7    0        0   0      0    2           0    0        0          0       0
##   8    0        1   0      0    1           0    0        0          0       0
##   9    0        1   1      0    2           0    0        0          0       1
```

```
findFreqTerms(DTM_quotes, 2000)
```

```
##  [1] "also"          "character"     "datetimestamp" "description"
##  [5] "heading"       "hour"          "isdst"         "language"
##  [9] "last"          "listauthor"    "listcontent"   "listsec"
## [13] "market"        "mday"          "meta"          "million"
## [17] "min"           "mon"           "new"           "origin"
## [21] "said"          "wday"          "will"          "yday"
## [25] "year"          "analysts"      "company"       "one"
## [29] "percent"       "billion"       "china"
```

+This is my matrix of numbers as well as terms that are used 2000 times.

```
DTM_quotes = removeSparseTerms(DTM_quotes, 0.95)
DTM_quotes
```

```
## <<DocumentTermMatrix (documents: 2500, terms: 801)>>
## Non-/sparse entries: 280686/1721814
## Sparsity           : 86%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

```
tfidf_quotes = weightTfIdf(DTM_quotes)
cosine_dist_mat = proxy::dist(as.matrix(tfidf_quotes), method='cosine')
tree_quotes = hclust(cosine_dist_mat)
clust5 = cutree(tree_quotes, k=5)
#Doing PCA
X = as.matrix(tfidf_quotes)
summary(colSums(X))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.543   6.075   6.799   8.168  29.795
```

```
scrub_cols = which(colSums(X) == 0)
X = X[,-scrub_cols]
pca_quotes = prcomp(X, scale=TRUE)
summary(pca_quotes)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.8006 3.09144 2.72834 2.70201 2.56291 2.51168 2.34230
## Proportion of Variance 0.0184 0.01217 0.00948 0.00930 0.00837 0.00804 0.00699
## Cumulative Proportion  0.0184 0.03057 0.04006 0.04936 0.05773 0.06576 0.07275
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     2.31372 2.24416 2.20230 2.14601 2.12115 2.04812 2.00348
## Proportion of Variance 0.00682 0.00642 0.00618 0.00587 0.00573 0.00534 0.00511
## Cumulative Proportion  0.07957 0.08599 0.09216 0.09803 0.10376 0.10911 0.11422
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     1.96376 1.88514 1.8798 1.86045 1.84216 1.83588 1.81066
## Proportion of Variance 0.00491 0.00453 0.0045 0.00441 0.00432 0.00429 0.00418
## Cumulative Proportion  0.11913 0.12366 0.1282 0.13257 0.13689 0.14119 0.14536
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     1.79061 1.78776 1.78210 1.76488 1.74204 1.72199 1.71932
## Proportion of Variance 0.00408 0.00407 0.00405 0.00397 0.00387 0.00378 0.00377
## Cumulative Proportion  0.14945 0.15352 0.15756 0.16153 0.16540 0.16918 0.17294
##                           PC29    PC30   PC31    PC32    PC33    PC34    PC35
## Standard deviation     1.69650 1.69041 1.6814 1.66377 1.65499 1.64142 1.63756
## Proportion of Variance 0.00367 0.00364 0.0036 0.00353 0.00349 0.00343 0.00342
## Cumulative Proportion  0.17661 0.18025 0.1839 0.18738 0.19086 0.19430 0.19771
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     1.62439 1.61925 1.61406 1.60688 1.59872 1.59233 1.58862
## Proportion of Variance 0.00336 0.00334 0.00332 0.00329 0.00326 0.00323 0.00321
## Cumulative Proportion  0.20107 0.20441 0.20773 0.21102 0.21428 0.21751 0.22072
##                           PC43    PC44    PC45    PC46    PC47    PC48    PC49
```

```
## Standard deviation     1.58748 1.57338 1.57250 1.56323 1.55360 1.54844 1.54022
## Proportion of Variance 0.00321 0.00315 0.00315 0.00311 0.00307 0.00305 0.00302
## Cumulative Proportion  0.22393 0.22709 0.23024 0.23335 0.23642 0.23948 0.24250
##                           PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     1.5336 1.53121 1.52557 1.52431 1.52274 1.51668 1.51138
## Proportion of Variance 0.0030 0.00299 0.00296 0.00296 0.00295 0.00293 0.00291
## Cumulative Proportion  0.2455 0.24848 0.25145 0.25441 0.25736 0.26029 0.26320
##                           PC57    PC58    PC59    PC60    PC61   PC62    PC63
## Standard deviation     1.50471 1.50037 1.49557 1.49174 1.48583 1.4828 1.47977
## Proportion of Variance 0.00288 0.00287 0.00285 0.00283 0.00281 0.0028 0.00279
## Cumulative Proportion  0.26609 0.26895 0.27180 0.27464 0.27745 0.2802 0.28304
##                           PC64    PC65    PC66    PC67    PC68   PC69    PC70
## Standard deviation     1.47486 1.47445 1.47039 1.46677 1.46365 1.4554 1.45058
## Proportion of Variance 0.00277 0.00277 0.00275 0.00274 0.00273 0.0027 0.00268
## Cumulative Proportion  0.28581 0.28858 0.29134 0.29408 0.29681 0.2995 0.30218
##                           PC71    PC72    PC73    PC74    PC75    PC76   PC77
## Standard deviation     1.44821 1.44393 1.44175 1.43977 1.43425 1.43252 1.4282
## Proportion of Variance 0.00267 0.00266 0.00265 0.00264 0.00262 0.00261 0.0026
## Cumulative Proportion  0.30486 0.30751 0.31016 0.31280 0.31542 0.31804 0.3206
##                           PC78    PC79    PC80    PC81    PC82    PC83    PC84
## Standard deviation     1.42324 1.41908 1.41274 1.41191 1.41011 1.40778 1.40391
## Proportion of Variance 0.00258 0.00257 0.00254 0.00254 0.00253 0.00252 0.00251
## Cumulative Proportion  0.32321 0.32578 0.32832 0.33086 0.33339 0.33592 0.33843
##                           PC85    PC86    PC87    PC88    PC89    PC90    PC91
## Standard deviation     1.4016 1.39786 1.39434 1.39328 1.39062 1.38775 1.38207
## Proportion of Variance 0.0025 0.00249 0.00248 0.00247 0.00246 0.00245 0.00243
## Cumulative Proportion  0.3409 0.34342 0.34590 0.34837 0.35083 0.35329 0.35572
##                           PC92   PC93    PC94    PC95    PC96    PC97    PC98
## Standard deviation     1.37851 1.3718 1.37101 1.36992 1.36519 1.36057 1.35848
## Proportion of Variance 0.00242 0.0024 0.00239 0.00239 0.00237 0.00236 0.00235
## Cumulative Proportion  0.35814 0.3605 0.36293 0.36532 0.36770 0.37006 0.37241
##                           PC99   PC100   PC101   PC102   PC103   PC104  PC105
## Standard deviation     1.35688 1.35439 1.35217 1.35001 1.34626 1.34612 1.3425
## Proportion of Variance 0.00235 0.00234 0.00233 0.00232 0.00231 0.00231 0.0023
## Cumulative Proportion  0.37475 0.37709 0.37942 0.38174 0.38405 0.38636 0.3886
##                          PC106   PC107   PC108   PC109   PC110   PC111   PC112
## Standard deviation     1.33969 1.33720 1.33048 1.32902 1.32598 1.32506 1.32393
## Proportion of Variance 0.00229 0.00228 0.00225 0.00225 0.00224 0.00224 0.00223
## Cumulative Proportion  0.39094 0.39322 0.39547 0.39772 0.39996 0.40220 0.40443
##                          PC113   PC114   PC115  PC116  PC117   PC118   PC119
## Standard deviation     1.32157 1.32044 1.31731 1.3153 1.3136 1.30806 1.30549
## Proportion of Variance 0.00222 0.00222 0.00221 0.0022 0.0022 0.00218 0.00217
## Cumulative Proportion  0.40666 0.40888 0.41109 0.4133 0.4155 0.41767 0.41984
##                          PC120   PC121   PC122   PC123   PC124   PC125   PC126
## Standard deviation     1.30494 1.30136 1.29930 1.29737 1.29385 1.29266 1.29043
## Proportion of Variance 0.00217 0.00216 0.00215 0.00214 0.00213 0.00213 0.00212
## Cumulative Proportion  0.42201 0.42417 0.42632 0.42846 0.43060 0.43272 0.43485
##                          PC127   PC128  PC129   PC130   PC131   PC132   PC133
## Standard deviation     1.28815 1.28728 1.2836 1.27976 1.27924 1.27397 1.27293
## Proportion of Variance 0.00211 0.00211 0.0021 0.00209 0.00208 0.00207 0.00206
## Cumulative Proportion  0.43696 0.43907 0.4412 0.44326 0.44534 0.44741 0.44947
##                          PC134   PC135   PC136   PC137   PC138   PC139   PC140
## Standard deviation     1.27031 1.26859 1.26801 1.26561 1.26095 1.26030 1.25865
## Proportion of Variance 0.00206 0.00205 0.00205 0.00204 0.00203 0.00202 0.00202
```

```
## Cumulative Proportion  0.45153 0.45358 0.45563 0.45767 0.45969 0.46172 0.46373
##                          PC141   PC142  PC143   PC144   PC145   PC146   PC147
## Standard deviation     1.25629 1.25532 1.2541 1.25045 1.24653 1.24451 1.24274
## Proportion of Variance 0.00201 0.00201 0.0020 0.00199 0.00198 0.00197 0.00197
## Cumulative Proportion  0.46574 0.46775 0.4698 0.47175 0.47373 0.47570 0.47767
##                          PC148   PC149   PC150   PC151   PC152   PC153   PC154
## Standard deviation     1.23842 1.23783 1.23576 1.23535 1.23358 1.22776 1.22627
## Proportion of Variance 0.00195 0.00195 0.00195 0.00194 0.00194 0.00192 0.00192
## Cumulative Proportion  0.47962 0.48157 0.48352 0.48546 0.48740 0.48932 0.49124
##                          PC155   PC156  PC157   PC158   PC159   PC160   PC161
## Standard deviation     1.22547 1.22411 1.2224 1.21965 1.21607 1.21301 1.21248
## Proportion of Variance 0.00191 0.00191 0.0019 0.00189 0.00188 0.00187 0.00187
## Cumulative Proportion  0.49315 0.49506 0.4970 0.49886 0.50074 0.50261 0.50449
##                          PC162   PC163   PC164   PC165   PC166   PC167   PC168
## Standard deviation     1.21105 1.20942 1.20746 1.20482 1.20342 1.20192 1.20051
## Proportion of Variance 0.00187 0.00186 0.00186 0.00185 0.00184 0.00184 0.00184
## Cumulative Proportion  0.50636 0.50822 0.51008 0.51193 0.51377 0.51561 0.51745
##                          PC169   PC170   PC171   PC172  PC173   PC174   PC175
## Standard deviation     1.19810 1.19654 1.19408 1.19127 1.1890 1.18624 1.18392
## Proportion of Variance 0.00183 0.00182 0.00182 0.00181 0.0018 0.00179 0.00179
## Cumulative Proportion  0.51927 0.52110 0.52292 0.52472 0.5265 0.52832 0.53010
##                          PC176   PC177   PC178   PC179   PC180   PC181   PC182
## Standard deviation     1.18194 1.18127 1.17977 1.17763 1.17561 1.17476 1.17187
## Proportion of Variance 0.00178 0.00178 0.00177 0.00177 0.00176 0.00176 0.00175
## Cumulative Proportion  0.53188 0.53366 0.53543 0.53720 0.53896 0.54072 0.54247
##                          PC183   PC184   PC185   PC186   PC187   PC188   PC189
## Standard deviation     1.16795 1.16754 1.16521 1.16276 1.16164 1.16011 1.15816
## Proportion of Variance 0.00174 0.00174 0.00173 0.00172 0.00172 0.00171 0.00171
## Cumulative Proportion  0.54420 0.54594 0.54767 0.54939 0.55111 0.55283 0.55454
##                          PC190  PC191   PC192   PC193   PC194   PC195   PC196
## Standard deviation     1.15755 1.1542 1.15259 1.15010 1.14882 1.14782 1.14729
## Proportion of Variance 0.00171 0.0017 0.00169 0.00169 0.00168 0.00168 0.00168
## Cumulative Proportion  0.55624 0.5579 0.55963 0.56132 0.56300 0.56468 0.56635
##                          PC197   PC198   PC199   PC200   PC201   PC202   PC203
## Standard deviation     1.14498 1.14138 1.13973 1.13840 1.13564 1.13382 1.13350
## Proportion of Variance 0.00167 0.00166 0.00165 0.00165 0.00164 0.00164 0.00164
## Cumulative Proportion  0.56802 0.56968 0.57134 0.57299 0.57463 0.57627 0.57791
##                          PC204   PC205   PC206   PC207   PC208   PC209  PC210
## Standard deviation     1.13253 1.12959 1.12775 1.12551 1.12399 1.12338 1.1220
## Proportion of Variance 0.00163 0.00163 0.00162 0.00161 0.00161 0.00161 0.0016
## Cumulative Proportion  0.57954 0.58116 0.58278 0.58440 0.58601 0.58762 0.5892
##                         PC211   PC212   PC213   PC214   PC215   PC216   PC217
## Standard deviation     1.1198 1.11881 1.11531 1.11391 1.11183 1.11073 1.10879
## Proportion of Variance 0.0016 0.00159 0.00158 0.00158 0.00157 0.00157 0.00157
## Cumulative Proportion  0.5908 0.59241 0.59400 0.59558 0.59715 0.59872 0.60029
##                          PC218   PC219   PC220   PC221   PC222   PC223   PC224
## Standard deviation     1.10625 1.10536 1.10347 1.10164 1.10101 1.09753 1.09701
## Proportion of Variance 0.00156 0.00156 0.00155 0.00155 0.00154 0.00153 0.00153
## Cumulative Proportion  0.60185 0.60340 0.60496 0.60650 0.60805 0.60958 0.61111
##                          PC225   PC226   PC227   PC228   PC229  PC230  PC231
## Standard deviation     1.09475 1.09421 1.09192 1.08897 1.08764 1.0864 1.0843
## Proportion of Variance 0.00153 0.00153 0.00152 0.00151 0.00151 0.0015 0.0015
## Cumulative Proportion  0.61264 0.61417 0.61568 0.61719 0.61870 0.6202 0.6217
##                          PC232   PC233   PC234   PC235   PC236   PC237   PC238
```

```
## Standard deviation     1.08147 1.07839 1.07746 1.07601 1.07468 1.07320 1.07104
## Proportion of Variance 0.00149 0.00148 0.00148 0.00147 0.00147 0.00147 0.00146
## Cumulative Proportion  0.62319 0.62467 0.62615 0.62763 0.62910 0.63057 0.63203
##                           PC239   PC240   PC241   PC242   PC243   PC244   PC245
## Standard deviation     1.06995 1.06777 1.06577 1.06357 1.06181 1.06084 1.05873
## Proportion of Variance 0.00146 0.00145 0.00145 0.00144 0.00144 0.00143 0.00143
## Cumulative Proportion  0.63349 0.63494 0.63639 0.63783 0.63926 0.64070 0.64212
##                           PC246   PC247   PC248   PC249   PC250   PC251  PC252
## Standard deviation     1.05788 1.05665 1.05484 1.05392 1.05349 1.05031 1.0494
## Proportion of Variance 0.00143 0.00142 0.00142 0.00141 0.00141 0.00141 0.0014
## Cumulative Proportion  0.64355 0.64497 0.64639 0.64780 0.64922 0.65062 0.6520
##                           PC253   PC254   PC255   PC256   PC257   PC258   PC259
## Standard deviation     1.04631 1.04576 1.04307 1.04228 1.04149 1.03947 1.03836
## Proportion of Variance 0.00139 0.00139 0.00139 0.00138 0.00138 0.00138 0.00137
## Cumulative Proportion  0.65342 0.65481 0.65620 0.65758 0.65897 0.66034 0.66172
##                           PC260   PC261   PC262   PC263   PC264   PC265   PC266
## Standard deviation     1.03735 1.03622 1.03497 1.03281 1.03049 1.02911 1.02768
## Proportion of Variance 0.00137 0.00137 0.00136 0.00136 0.00135 0.00135 0.00135
## Cumulative Proportion  0.66309 0.66445 0.66582 0.66718 0.66853 0.66988 0.67123
##                           PC267   PC268   PC269   PC270   PC271   PC272   PC273
## Standard deviation     1.02453 1.02336 1.02318 1.02233 1.01971 1.01772 1.01678
## Proportion of Variance 0.00134 0.00133 0.00133 0.00133 0.00132 0.00132 0.00132
## Cumulative Proportion  0.67256 0.67390 0.67523 0.67656 0.67789 0.67921 0.68052
##                           PC274   PC275  PC276  PC277  PC278  PC279   PC280
## Standard deviation     1.01597 1.01382 1.0114 1.0105 1.0097 1.0085 1.00655
## Proportion of Variance 0.00131 0.00131 0.0013 0.0013 0.0013 0.0013 0.00129
## Cumulative Proportion  0.68184 0.68315 0.6845 0.6857 0.6871 0.6883 0.68964
##                           PC281   PC282   PC283   PC284   PC285   PC286   PC287
## Standard deviation     1.00446 1.00280 1.00202 1.00019 0.99883 0.99697 0.99596
## Proportion of Variance 0.00129 0.00128 0.00128 0.00127 0.00127 0.00127 0.00126
## Cumulative Proportion  0.69092 0.69220 0.69348 0.69476 0.69603 0.69729 0.69856
##                           PC288   PC289   PC290   PC291   PC292   PC293   PC294
## Standard deviation     0.99380 0.99194 0.98975 0.98904 0.98792 0.98497 0.98420
## Proportion of Variance 0.00126 0.00125 0.00125 0.00125 0.00124 0.00124 0.00123
## Cumulative Proportion  0.69981 0.70107 0.70232 0.70356 0.70480 0.70604 0.70727
##                           PC295   PC296   PC297   PC298   PC299   PC300   PC301
## Standard deviation     0.97998 0.97883 0.97838 0.97754 0.97661 0.97573 0.97378
## Proportion of Variance 0.00122 0.00122 0.00122 0.00122 0.00121 0.00121 0.00121
## Cumulative Proportion  0.70850 0.70972 0.71094 0.71216 0.71337 0.71458 0.71579
##                          PC302  PC303   PC304   PC305   PC306   PC307   PC308
## Standard deviation     0.9716 0.9704 0.96803 0.96679 0.96551 0.96342 0.96283
## Proportion of Variance 0.0012 0.0012 0.00119 0.00119 0.00119 0.00118 0.00118
## Cumulative Proportion  0.7170 0.7182 0.71939 0.72058 0.72177 0.72295 0.72413
##                           PC309   PC310   PC311   PC312   PC313   PC314   PC315
## Standard deviation     0.96127 0.96015 0.95883 0.95714 0.95617 0.95501 0.95204
## Proportion of Variance 0.00118 0.00117 0.00117 0.00117 0.00116 0.00116 0.00115
## Cumulative Proportion  0.72531 0.72648 0.72765 0.72882 0.72998 0.73114 0.73230
##                           PC316   PC317   PC318   PC319   PC320   PC321   PC322
## Standard deviation     0.95057 0.94977 0.94890 0.94730 0.94512 0.94397 0.94200
## Proportion of Variance 0.00115 0.00115 0.00115 0.00114 0.00114 0.00114 0.00113
## Cumulative Proportion  0.73345 0.73460 0.73575 0.73689 0.73803 0.73916 0.74029
##                           PC323   PC324   PC325   PC326   PC327   PC328   PC329
## Standard deviation     0.94013 0.93892 0.93816 0.93717 0.93548 0.93357 0.93231
## Proportion of Variance 0.00113 0.00112 0.00112 0.00112 0.00111 0.00111 0.00111
```

```
## Cumulative Proportion  0.74142 0.74254 0.74366 0.74478 0.74590 0.74701 0.74811
##                          PC330  PC331  PC332   PC333   PC334   PC335   PC336
## Standard deviation     0.9313 0.9295 0.9281 0.92652 0.92484 0.92416 0.92315
## Proportion of Variance 0.0011 0.0011 0.0011 0.00109 0.00109 0.00109 0.00109
## Cumulative Proportion  0.7492 0.7503 0.7514 0.75251 0.75360 0.75469 0.75577
##                          PC337   PC338   PC339   PC340   PC341   PC342   PC343
## Standard deviation     0.92168 0.91974 0.91937 0.91666 0.91475 0.91422 0.91245
## Proportion of Variance 0.00108 0.00108 0.00108 0.00107 0.00107 0.00106 0.00106
## Cumulative Proportion  0.75686 0.75793 0.75901 0.76008 0.76115 0.76221 0.76327
##                          PC344   PC345   PC346   PC347   PC348   PC349   PC350
## Standard deviation     0.91079 0.91031 0.90775 0.90639 0.90593 0.90466 0.90302
## Proportion of Variance 0.00106 0.00106 0.00105 0.00105 0.00105 0.00104 0.00104
## Cumulative Proportion  0.76433 0.76538 0.76643 0.76748 0.76853 0.76957 0.77061
##                          PC351   PC352   PC353   PC354   PC355   PC356   PC357
## Standard deviation     0.90183 0.90150 0.89997 0.89952 0.89879 0.89743 0.89618
## Proportion of Variance 0.00104 0.00104 0.00103 0.00103 0.00103 0.00103 0.00102
## Cumulative Proportion  0.77164 0.77268 0.77371 0.77474 0.77577 0.77680 0.77782
##                          PC358   PC359   PC360   PC361  PC362  PC363  PC364
## Standard deviation     0.89507 0.89250 0.89064 0.88910 0.8869 0.8853 0.8843
## Proportion of Variance 0.00102 0.00101 0.00101 0.00101 0.0010 0.0010 0.0010
## Cumulative Proportion  0.77884 0.77985 0.78087 0.78187 0.7829 0.7839 0.7849
##                          PC365   PC366   PC367   PC368   PC369   PC370   PC371
## Standard deviation     0.88162 0.88025 0.87884 0.87852 0.87595 0.87568 0.87383
## Proportion of Variance 0.00099 0.00099 0.00098 0.00098 0.00098 0.00098 0.00097
## Cumulative Proportion  0.78586 0.78685 0.78783 0.78881 0.78979 0.79077 0.79174
##                          PC372   PC373   PC374   PC375   PC376   PC377   PC378
## Standard deviation     0.87321 0.87039 0.86916 0.86853 0.86694 0.86558 0.86474
## Proportion of Variance 0.00097 0.00097 0.00096 0.00096 0.00096 0.00095 0.00095
## Cumulative Proportion  0.79271 0.79368 0.79464 0.79560 0.79656 0.79751 0.79846
##                          PC379   PC380   PC381   PC382   PC383   PC384   PC385
## Standard deviation     0.86249 0.86156 0.85902 0.85844 0.85702 0.85586 0.85433
## Proportion of Variance 0.00095 0.00095 0.00094 0.00094 0.00094 0.00093 0.00093
## Cumulative Proportion  0.79941 0.80036 0.80130 0.80224 0.80317 0.80410 0.80503
##                          PC386   PC387   PC388   PC389   PC390   PC391   PC392
## Standard deviation     0.85258 0.85174 0.85064 0.84967 0.84843 0.84802 0.84417
## Proportion of Variance 0.00093 0.00092 0.00092 0.00092 0.00092 0.00092 0.00091
## Cumulative Proportion  0.80596 0.80688 0.80781 0.80873 0.80964 0.81056 0.81147
##                          PC393   PC394  PC395  PC396   PC397   PC398   PC399
## Standard deviation     0.84368 0.84288 0.8408 0.8386 0.83752 0.83701 0.83595
## Proportion of Variance 0.00091 0.00091 0.0009 0.0009 0.00089 0.00089 0.00089
## Cumulative Proportion  0.81237 0.81328 0.8142 0.8151 0.81597 0.81686 0.81775
##                          PC400   PC401   PC402   PC403   PC404   PC405   PC406
## Standard deviation     0.83397 0.83330 0.83218 0.83063 0.82977 0.82885 0.82739
## Proportion of Variance 0.00089 0.00088 0.00088 0.00088 0.00088 0.00088 0.00087
## Cumulative Proportion  0.81864 0.81952 0.82040 0.82128 0.82216 0.82304 0.82391
##                          PC407   PC408   PC409   PC410   PC411   PC412   PC413
## Standard deviation     0.82621 0.82493 0.82429 0.82343 0.82139 0.82036 0.81927
## Proportion of Variance 0.00087 0.00087 0.00087 0.00086 0.00086 0.00086 0.00086
## Cumulative Proportion  0.82478 0.82564 0.82651 0.82737 0.82823 0.82909 0.82995
##                          PC414   PC415   PC416   PC417   PC418   PC419   PC420
## Standard deviation     0.81741 0.81474 0.81428 0.81263 0.81255 0.81141 0.80871
## Proportion of Variance 0.00085 0.00085 0.00084 0.00084 0.00084 0.00084 0.00083
## Cumulative Proportion  0.83080 0.83164 0.83249 0.83333 0.83417 0.83501 0.83584
##                          PC421   PC422   PC423   PC424   PC425   PC426   PC427
```

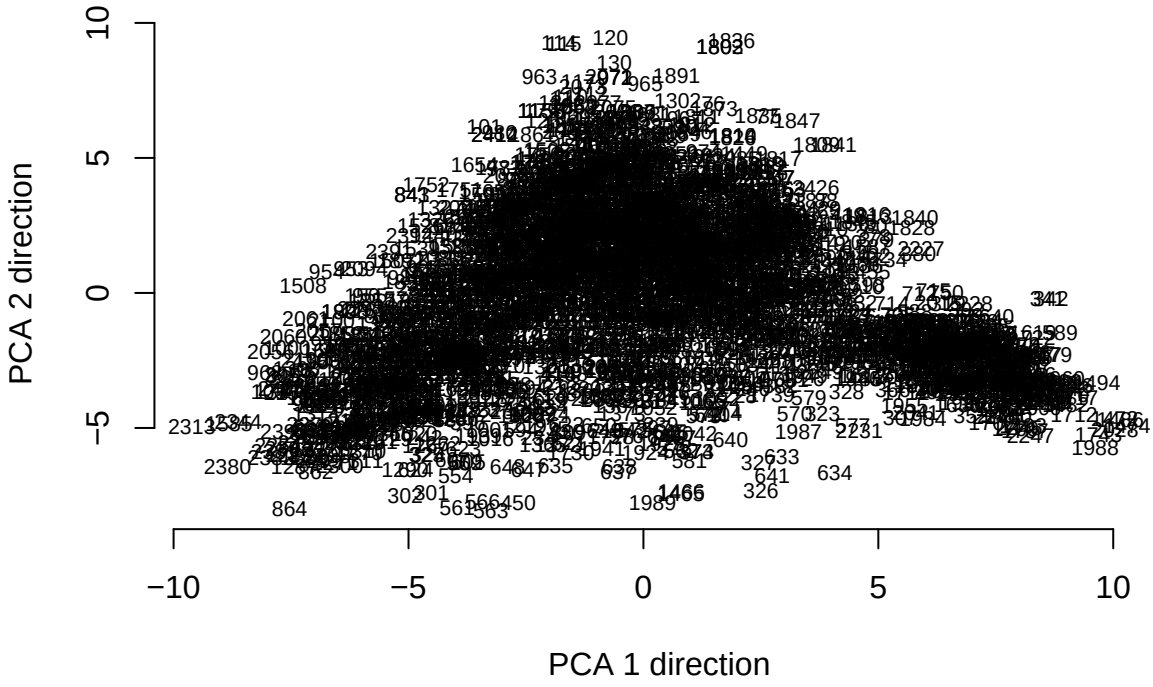
```
## Standard deviation     0.80819 0.80694 0.80555 0.80431 0.80255 0.80027 0.79853
## Proportion of Variance 0.00083 0.00083 0.00083 0.00082 0.00082 0.00082 0.00081
## Cumulative Proportion  0.83667 0.83750 0.83833 0.83915 0.83997 0.84079 0.84160
##                           PC428   PC429   PC430   PC431  PC432  PC433  PC434
## Standard deviation     0.79752 0.79691 0.79642 0.79532 0.7937 0.7926 0.7920
## Proportion of Variance 0.00081 0.00081 0.00081 0.00081 0.0008 0.0008 0.0008
## Cumulative Proportion  0.84241 0.84322 0.84403 0.84483 0.8456 0.8464 0.8472
##                          PC435   PC436   PC437   PC438   PC439   PC440   PC441
## Standard deviation     0.7909 0.78918 0.78837 0.78591 0.78508 0.78439 0.78178
## Proportion of Variance 0.0008 0.00079 0.00079 0.00079 0.00079 0.00078 0.00078
## Cumulative Proportion  0.8480 0.84883 0.84962 0.85041 0.85119 0.85197 0.85275
##                           PC442   PC443   PC444   PC445   PC446   PC447   PC448
## Standard deviation     0.78044 0.77935 0.77720 0.77517 0.77468 0.77391 0.77290
## Proportion of Variance 0.00078 0.00077 0.00077 0.00077 0.00076 0.00076 0.00076
## Cumulative Proportion  0.85353 0.85430 0.85507 0.85584 0.85660 0.85737 0.85813
##                           PC449   PC450   PC451   PC452   PC453   PC454   PC455
## Standard deviation     0.77148 0.77093 0.76817 0.76663 0.76571 0.76542 0.76411
## Proportion of Variance 0.00076 0.00076 0.00075 0.00075 0.00075 0.00075 0.00074
## Cumulative Proportion  0.85888 0.85964 0.86039 0.86114 0.86189 0.86263 0.86338
##                           PC456   PC457   PC458   PC459   PC460   PC461   PC462
## Standard deviation     0.76296 0.76165 0.75968 0.75802 0.75694 0.75625 0.75449
## Proportion of Variance 0.00074 0.00074 0.00074 0.00073 0.00073 0.00073 0.00073
## Cumulative Proportion  0.86412 0.86486 0.86559 0.86633 0.86706 0.86778 0.86851
##                           PC463   PC464   PC465   PC466   PC467   PC468   PC469
## Standard deviation     0.75364 0.75158 0.75117 0.75064 0.75058 0.74806 0.74660
## Proportion of Variance 0.00072 0.00072 0.00072 0.00072 0.00072 0.00071 0.00071
## Cumulative Proportion  0.86923 0.86995 0.87067 0.87139 0.87211 0.87282 0.87353
##                           PC470   PC471  PC472  PC473  PC474  PC475   PC476
## Standard deviation     0.74570 0.74398 0.7430 0.7428 0.7409 0.7396 0.73729
## Proportion of Variance 0.00071 0.00071 0.0007 0.0007 0.0007 0.0007 0.00069
## Cumulative Proportion  0.87424 0.87494 0.8757 0.8763 0.8770 0.8778 0.87844
##                           PC477   PC478   PC479   PC480   PC481   PC482   PC483
## Standard deviation     0.73698 0.73611 0.73569 0.73394 0.73312 0.73209 0.73099
## Proportion of Variance 0.00069 0.00069 0.00069 0.00069 0.00068 0.00068 0.00068
## Cumulative Proportion  0.87913 0.87982 0.88051 0.88120 0.88188 0.88256 0.88324
##                           PC484   PC485   PC486   PC487   PC488   PC489   PC490
## Standard deviation     0.73082 0.72896 0.72589 0.72483 0.72423 0.72276 0.72148
## Proportion of Variance 0.00068 0.00068 0.00067 0.00067 0.00067 0.00067 0.00066
## Cumulative Proportion  0.88392 0.88460 0.88527 0.88594 0.88661 0.88728 0.88794
##                           PC491   PC492   PC493   PC494   PC495   PC496   PC497
## Standard deviation     0.72015 0.71937 0.71806 0.71783 0.71515 0.71480 0.71256
## Proportion of Variance 0.00066 0.00066 0.00066 0.00066 0.00065 0.00065 0.00065
## Cumulative Proportion  0.88860 0.88926 0.88992 0.89057 0.89122 0.89187 0.89252
##                           PC498   PC499   PC500   PC501   PC502   PC503   PC504
## Standard deviation     0.71138 0.71056 0.70917 0.70833 0.70762 0.70635 0.70431
## Proportion of Variance 0.00064 0.00064 0.00064 0.00064 0.00064 0.00064 0.00063
## Cumulative Proportion  0.89317 0.89381 0.89445 0.89509 0.89573 0.89636 0.89699
##                           PC505   PC506   PC507   PC508   PC509   PC510   PC511
## Standard deviation     0.70309 0.70135 0.70025 0.69972 0.69843 0.69744 0.69654
## Proportion of Variance 0.00063 0.00063 0.00062 0.00062 0.00062 0.00062 0.00062
## Cumulative Proportion  0.89762 0.89825 0.89888 0.89950 0.90012 0.90074 0.90136
##                           PC512   PC513   PC514   PC515  PC516  PC517  PC518
## Standard deviation     0.69423 0.69217 0.69116 0.69026 0.6891 0.6885 0.6874
## Proportion of Variance 0.00061 0.00061 0.00061 0.00061 0.0006 0.0006 0.0006
```

```
## Cumulative Proportion  0.90197 0.90258 0.90319 0.90380 0.9044 0.9050 0.9056
##                          PC519  PC520  PC521   PC522   PC523   PC524   PC525
## Standard deviation     0.6866 0.6848 0.6844 0.68307 0.68262 0.68084 0.67936
## Proportion of Variance 0.0006 0.0006 0.0006 0.00059 0.00059 0.00059 0.00059
## Cumulative Proportion  0.9062 0.9068 0.9074 0.90800 0.90859 0.90918 0.90977
##                           PC526   PC527   PC528   PC529   PC530   PC531   PC532
## Standard deviation     0.67832 0.67659 0.67495 0.67370 0.67208 0.67131 0.67077
## Proportion of Variance 0.00059 0.00058 0.00058 0.00058 0.00058 0.00057 0.00057
## Cumulative Proportion  0.91036 0.91094 0.91152 0.91210 0.91267 0.91325 0.91382
##                           PC533   PC534   PC535   PC536   PC537   PC538   PC539
## Standard deviation     0.66913 0.66761 0.66628 0.66506 0.66348 0.66211 0.66161
## Proportion of Variance 0.00057 0.00057 0.00057 0.00056 0.00056 0.00056 0.00056
## Cumulative Proportion  0.91439 0.91496 0.91552 0.91609 0.91665 0.91721 0.91776
##                           PC540   PC541   PC542   PC543   PC544   PC545   PC546
## Standard deviation     0.66040 0.65824 0.65786 0.65611 0.65366 0.65341 0.65230
## Proportion of Variance 0.00056 0.00055 0.00055 0.00055 0.00054 0.00054 0.00054
## Cumulative Proportion  0.91832 0.91887 0.91942 0.91997 0.92052 0.92106 0.92160
##                           PC547   PC548   PC549   PC550   PC551   PC552   PC553
## Standard deviation     0.65050 0.64974 0.64931 0.64850 0.64766 0.64451 0.64399
## Proportion of Variance 0.00054 0.00054 0.00054 0.00054 0.00053 0.00053 0.00053
## Cumulative Proportion  0.92214 0.92268 0.92322 0.92375 0.92429 0.92481 0.92534
##                           PC554   PC555   PC556   PC557   PC558   PC559   PC560
## Standard deviation     0.64329 0.64199 0.64090 0.63927 0.63732 0.63577 0.63488
## Proportion of Variance 0.00053 0.00053 0.00052 0.00052 0.00052 0.00051 0.00051
## Cumulative Proportion  0.92587 0.92640 0.92692 0.92744 0.92796 0.92847 0.92898
##                           PC561   PC562   PC563   PC564   PC565  PC566  PC567
## Standard deviation     0.63411 0.63283 0.63223 0.63138 0.62966 0.6280 0.6264
## Proportion of Variance 0.00051 0.00051 0.00051 0.00051 0.00051 0.0005 0.0005
## Cumulative Proportion  0.92950 0.93001 0.93052 0.93102 0.93153 0.9320 0.9325
##                          PC568  PC569   PC570   PC571   PC572   PC573   PC574
## Standard deviation     0.6259 0.6241 0.62261 0.62163 0.62105 0.61858 0.61789
## Proportion of Variance 0.0005 0.0005 0.00049 0.00049 0.00049 0.00049 0.00049
## Cumulative Proportion  0.9330 0.9335 0.93402 0.93451 0.93500 0.93549 0.93598
##                           PC575   PC576   PC577   PC578   PC579   PC580   PC581
## Standard deviation     0.61689 0.61602 0.61520 0.61234 0.61137 0.61073 0.61000
## Proportion of Variance 0.00048 0.00048 0.00048 0.00048 0.00048 0.00048 0.00047
## Cumulative Proportion  0.93646 0.93695 0.93743 0.93791 0.93838 0.93886 0.93933
##                           PC582   PC583   PC584   PC585   PC586   PC587   PC588
## Standard deviation     0.60706 0.60642 0.60629 0.60541 0.60427 0.60231 0.60102
## Proportion of Variance 0.00047 0.00047 0.00047 0.00047 0.00047 0.00046 0.00046
## Cumulative Proportion  0.93980 0.94027 0.94074 0.94120 0.94167 0.94213 0.94259
##                           PC589   PC590   PC591   PC592   PC593   PC594   PC595
## Standard deviation     0.60007 0.59977 0.59914 0.59591 0.59549 0.59515 0.59351
## Proportion of Variance 0.00046 0.00046 0.00046 0.00045 0.00045 0.00045 0.00045
## Cumulative Proportion  0.94305 0.94351 0.94397 0.94442 0.94487 0.94532 0.94577
##                           PC596   PC597   PC598   PC599   PC600   PC601   PC602
## Standard deviation     0.59194 0.58981 0.58867 0.58715 0.58654 0.58424 0.58379
## Proportion of Variance 0.00045 0.00044 0.00044 0.00044 0.00044 0.00043 0.00043
## Cumulative Proportion  0.94622 0.94666 0.94710 0.94754 0.94798 0.94841 0.94885
##                           PC603   PC604   PC605   PC606   PC607   PC608   PC609
## Standard deviation     0.58312 0.58145 0.58031 0.57771 0.57646 0.57550 0.57453
## Proportion of Variance 0.00043 0.00043 0.00043 0.00043 0.00042 0.00042 0.00042
## Cumulative Proportion  0.94928 0.94971 0.95014 0.95057 0.95099 0.95141 0.95183
##                           PC610   PC611   PC612   PC613   PC614   PC615   PC616
```

```
## Standard deviation     0.57295 0.57268 0.57118 0.57038 0.56873 0.56748 0.56710
## Proportion of Variance 0.00042 0.00042 0.00042 0.00041 0.00041 0.00041 0.00041
## Cumulative Proportion  0.95225 0.95267 0.95308 0.95350 0.95391 0.95432 0.95473
##                           PC617   PC618  PC619  PC620  PC621  PC622  PC623
## Standard deviation     0.56659 0.56439 0.5630 0.5621 0.5616 0.5611 0.5603
## Proportion of Variance 0.00041 0.00041 0.0004 0.0004 0.0004 0.0004 0.0004
## Cumulative Proportion  0.95514 0.95554 0.9559 0.9564 0.9567 0.9571 0.9576
##                          PC624   PC625   PC626   PC627   PC628   PC629   PC630
## Standard deviation     0.5591 0.55658 0.55545 0.55393 0.55197 0.55069 0.55051
## Proportion of Variance 0.0004 0.00039 0.00039 0.00039 0.00039 0.00039 0.00039
## Cumulative Proportion  0.9579 0.95835 0.95874 0.95913 0.95952 0.95990 0.96029
##                           PC631   PC632   PC633   PC634   PC635   PC636   PC637
## Standard deviation     0.54864 0.54761 0.54712 0.54511 0.54477 0.54363 0.54195
## Proportion of Variance 0.00038 0.00038 0.00038 0.00038 0.00038 0.00038 0.00037
## Cumulative Proportion  0.96067 0.96106 0.96144 0.96182 0.96219 0.96257 0.96294
##                           PC638   PC639   PC640   PC641   PC642   PC643   PC644
## Standard deviation     0.54033 0.54007 0.53896 0.53703 0.53671 0.53498 0.53416
## Proportion of Variance 0.00037 0.00037 0.00037 0.00037 0.00037 0.00036 0.00036
## Cumulative Proportion  0.96332 0.96369 0.96406 0.96442 0.96479 0.96516 0.96552
##                           PC645   PC646   PC647   PC648   PC649   PC650   PC651
## Standard deviation     0.53308 0.53232 0.52985 0.52888 0.52590 0.52459 0.52252
## Proportion of Variance 0.00036 0.00036 0.00036 0.00036 0.00035 0.00035 0.00035
## Cumulative Proportion  0.96588 0.96624 0.96660 0.96696 0.96731 0.96766 0.96801
##                           PC652   PC653   PC654   PC655   PC656   PC657   PC658
## Standard deviation     0.52173 0.52092 0.52089 0.51991 0.51877 0.51634 0.51549
## Proportion of Variance 0.00035 0.00035 0.00035 0.00034 0.00034 0.00034 0.00034
## Cumulative Proportion  0.96835 0.96870 0.96905 0.96939 0.96973 0.97007 0.97041
##                           PC659   PC660   PC661   PC662   PC663   PC664   PC665
## Standard deviation     0.51434 0.51253 0.51148 0.50956 0.50807 0.50754 0.50710
## Proportion of Variance 0.00034 0.00033 0.00033 0.00033 0.00033 0.00033 0.00033
## Cumulative Proportion  0.97075 0.97108 0.97142 0.97175 0.97208 0.97240 0.97273
##                           PC666   PC667   PC668   PC669   PC670   PC671   PC672
## Standard deviation     0.50577 0.50456 0.50294 0.50141 0.50071 0.49924 0.49846
## Proportion of Variance 0.00033 0.00032 0.00032 0.00032 0.00032 0.00032 0.00032
## Cumulative Proportion  0.97306 0.97338 0.97370 0.97402 0.97434 0.97466 0.97498
##                           PC673   PC674   PC675   PC676   PC677   PC678  PC679
## Standard deviation     0.49770 0.49654 0.49414 0.49353 0.49261 0.48956 0.4888
## Proportion of Variance 0.00032 0.00031 0.00031 0.00031 0.00031 0.00031 0.0003
## Cumulative Proportion  0.97529 0.97561 0.97592 0.97623 0.97654 0.97684 0.9771
##                          PC680  PC681  PC682  PC683  PC684   PC685   PC686
## Standard deviation     0.4872 0.4863 0.4831 0.4823 0.4816 0.48069 0.47910
## Proportion of Variance 0.0003 0.0003 0.0003 0.0003 0.0003 0.00029 0.00029
## Cumulative Proportion  0.9775 0.9778 0.9780 0.9783 0.9786 0.97893 0.97923
##                           PC687   PC688   PC689   PC690   PC691   PC692   PC693
## Standard deviation     0.47789 0.47763 0.47570 0.47354 0.47253 0.47157 0.47034
## Proportion of Variance 0.00029 0.00029 0.00029 0.00029 0.00028 0.00028 0.00028
## Cumulative Proportion  0.97952 0.97981 0.98010 0.98038 0.98067 0.98095 0.98123
##                           PC694   PC695   PC696   PC697   PC698   PC699   PC700
## Standard deviation     0.46880 0.46786 0.46733 0.46572 0.46438 0.46307 0.46195
## Proportion of Variance 0.00028 0.00028 0.00028 0.00028 0.00027 0.00027 0.00027
## Cumulative Proportion  0.98151 0.98179 0.98207 0.98234 0.98262 0.98289 0.98316
##                           PC701   PC702   PC703   PC704   PC705   PC706   PC707
## Standard deviation     0.45936 0.45829 0.45770 0.45562 0.45418 0.45333 0.45306
## Proportion of Variance 0.00027 0.00027 0.00027 0.00026 0.00026 0.00026 0.00026
```

```
## Cumulative Proportion  0.98343 0.98370 0.98397 0.98423 0.98449 0.98476 0.98502
##                           PC708   PC709   PC710   PC711   PC712   PC713   PC714
## Standard deviation     0.45089 0.44941 0.44860 0.44655 0.44577 0.44526 0.44278
## Proportion of Variance 0.00026 0.00026 0.00026 0.00025 0.00025 0.00025 0.00025
## Cumulative Proportion  0.98528 0.98553 0.98579 0.98604 0.98630 0.98655 0.98680
##                           PC715   PC716   PC717   PC718   PC719   PC720   PC721
## Standard deviation     0.44161 0.44019 0.43913 0.43524 0.43487 0.43339 0.43318
## Proportion of Variance 0.00025 0.00025 0.00025 0.00024 0.00024 0.00024 0.00024
## Cumulative Proportion  0.98705 0.98730 0.98754 0.98778 0.98802 0.98826 0.98850
##                           PC722   PC723   PC724   PC725   PC726   PC727   PC728
## Standard deviation     0.43259 0.43190 0.42852 0.42727 0.42659 0.42547 0.42374
## Proportion of Variance 0.00024 0.00024 0.00023 0.00023 0.00023 0.00023 0.00023
## Cumulative Proportion  0.98874 0.98898 0.98921 0.98944 0.98968 0.98991 0.99014
##                           PC729   PC730   PC731   PC732   PC733   PC734   PC735
## Standard deviation     0.42319 0.42081 0.41952 0.41717 0.41623 0.41468 0.41283
## Proportion of Variance 0.00023 0.00023 0.00022 0.00022 0.00022 0.00022 0.00022
## Cumulative Proportion  0.99036 0.99059 0.99081 0.99103 0.99126 0.99147 0.99169
##                           PC736   PC737   PC738   PC739   PC740   PC741   PC742
## Standard deviation     0.41037 0.40959 0.40828 0.40686 0.40521 0.40262 0.40164
## Proportion of Variance 0.00021 0.00021 0.00021 0.00021 0.00021 0.00021 0.00021
## Cumulative Proportion  0.99191 0.99212 0.99233 0.99254 0.99275 0.99296 0.99316
##                          PC743  PC744  PC745  PC746  PC747  PC748   PC749
## Standard deviation     0.4009 0.3983 0.3981 0.3969 0.3950 0.3940 0.39060
## Proportion of Variance 0.0002 0.0002 0.0002 0.0002 0.0002 0.0002 0.00019
## Cumulative Proportion  0.9934 0.9936 0.9938 0.9940 0.9942 0.9944 0.99456
##                           PC750   PC751   PC752   PC753   PC754   PC755   PC756
## Standard deviation     0.39033 0.38785 0.38495 0.38325 0.38275 0.38059 0.37952
## Proportion of Variance 0.00019 0.00019 0.00019 0.00019 0.00019 0.00018 0.00018
## Cumulative Proportion  0.99476 0.99495 0.99514 0.99533 0.99551 0.99570 0.99588
##                           PC757   PC758   PC759   PC760   PC761   PC762   PC763
## Standard deviation     0.37535 0.37474 0.37284 0.37180 0.36843 0.36740 0.36628
## Proportion of Variance 0.00018 0.00018 0.00018 0.00018 0.00017 0.00017 0.00017
## Cumulative Proportion  0.99606 0.99624 0.99642 0.99659 0.99677 0.99694 0.99711
##                           PC764   PC765   PC766   PC767   PC768   PC769   PC770
## Standard deviation     0.36470 0.36239 0.36023 0.35821 0.35765 0.35399 0.35163
## Proportion of Variance 0.00017 0.00017 0.00017 0.00016 0.00016 0.00016 0.00016
## Cumulative Proportion  0.99728 0.99744 0.99761 0.99777 0.99794 0.99810 0.99825
##                           PC771   PC772   PC773   PC774   PC775   PC776   PC777
## Standard deviation     0.34639 0.34528 0.34145 0.33952 0.33758 0.33456 0.32953
## Proportion of Variance 0.00015 0.00015 0.00015 0.00015 0.00015 0.00014 0.00014
## Cumulative Proportion  0.99841 0.99856 0.99871 0.99885 0.99900 0.99914 0.99928
##                           PC778   PC779   PC780   PC781  PC782   PC783   PC784
## Standard deviation     0.32567 0.32365 0.31814 0.30003 0.2826 0.26760 0.09241
## Proportion of Variance 0.00014 0.00013 0.00013 0.00011 0.0001 0.00009 0.00001
## Cumulative Proportion  0.99941 0.99955 0.99968 0.99979 0.9999 0.99998 1.00000
##                           PC785
## Standard deviation     0.05791
## Proportion of Variance 0.00000
## Cumulative Proportion  1.00000
```

```
plot(pca_quotes$x[,1:2], xlab="PCA 1 direction", ylab="PCA 2 direction", bty="n",
     type='n')
text(pca_quotes$x[,1:2], labels = 1:length(readfiles), cex=0.7)
```



+In this graph a lot of these are grouped together in the middle due to the size of our dataset, however we can inspect some pairs of outliers.

```
content(readfiles[[40]])
```

```
##  [1] "British insurer General Accident PLC reported a fall in annual profits on Tuesday, despite a re
##  [2] "Operating profits slipped 15 million pounds to 421 million ($680 million) although a rise in we
##  [3] "The recent strength of the pound was offset by a change to an average translation of overseas e
##  [4] "Despite the fall, profits were above market forecasts and a 10.5 percent increase in dividend i
##  [5] "Net assets per share rose to 675 pence from 653 pence."
##  [6] "By late morning General Accident shares had added 8-1/2 pence to 845-1/2 pence."
##  [7] "General Accident chief executive Bob Scott described the market as increasingly competitive but
##  [8] "Scott said the group wanted to focus life operations in Britain and the goal was to increase th
##  [9] "The proportion of dividend covered by net life earnings in 1996 was 44 percent but Scott ruled
## [10] "\"At the present time there's nothing wrong with organic growth in life. We've proved we're cap
## [11] "Provident Mutual, acquired in January 1996, contributed 18 million pounds to the results, net e
## [12] "The key for the life company was the profitability of new business and the group had done well
## [13] "Scott dismissed speculation of a possible merger or takeover of General Accident."
## [14] "\"We've got the benefits of scale in our major businesses. We don't have to do anything,\" he s
## [15] "The rating environment continues to be mixed and General Accident echoed comments from Commerci
## [16] "\"There are certainly signs that in personal motor rating increases are coming through. We're p
## [17] "There were, however no signs of rate increases in the homeowner or commercial sectors."
## [18] "Like other insurers, General Accident results saw an increase in weather losses over the year a
```

```
content(readfiles[[28]])
```

```
##  [1] "Scottish Amicable on Thursday asked for firm offers for the company to be tabled by the end of
```

```
##  [2] "Three companies were already known to have signed confidentiality agreements which give access
##  [3] "But a source close to the deal said that others had also signed the agreements and that potenti
##  [4] "ScotAm said that all interested parties would be requested to table firm offers by February 28
##  [5] "A spokesman for ScotAm declined to comment on either the number or the identity of companies wh
##  [6] "One possibility is Dutch insurance group Aegon whose acquistion of Scottish Equitable has alrea
##  [7] "Another firm rumoured to have contemplated a bid, Belgian-Dutch insurer Fortis is reported to h
##  [8] "Lloyds chief executive Peter Ellwood said that the company had run a slide-rule over ScotAm but
##  [9] "Meanwhile, The Scotsman newspaper on Thursday reported that AMP is considering moving its Brist
## [10] "The paper said that AMP would use Scotland to expand into Europe, and that it placed considerab
## [11] "The Prudential and Abbey National are believed to have plans to downscale or close SAIM, the pa
## [12] "Following the receipt of bids, ScotAm will hold further discussions after which bidders will be
## [13] "At the end of the process, the board will recommend a single proposal and an announcement of th
## [14] "Under the terms of the confidentiality agreements the decision of the board is final."
## [15] "\"We believe the process we are publishing today will maximise value for policyholders and will
## [16] "In a move which may help deflect criticism of the sealed bid process, ScotAm said it would make
## [17] "In addition, the Department of Trade &amp;amp; Industry and an independent actuary will be consulte
## [18] "ScotAm will also outline the board's reasons for its recommendation, the criteria used and summ
## [19] "The circular to policyholders outlining the final bid is likely some time in May."
## [20] "Abbey National sparked the auction at the end of January when it announced it was prepared to p
## [21] "The Abbey offer was topped a week later by the Pru's 1.9 billion bid."
## [22] "Both companies have said that their offers may be raised when they have more detailed financial
```

+These two are similar which makes sense because they seem to be focused on financial issues. You can do a tree or knn in order to find the author attribution of these quotes and then find the accuracy based on that.

#Problem 6

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


library(RCurl)
x <- getURL('https://raw.githubusercontent.com/jgscott/STA380/master/data/groceries.txt')
y <- read.table(text=x, header = FALSE, sep = ",", fill = TRUE)
library(data.table)
setDT(y, keep.rownames = TRUE)[]
```

```
##           rn                 V1                 V2              V3
##     1:     1       citrus fruit semi-finished bread        margarine
##     2:     2    tropical fruit              yogurt           coffee
##     3:     3         whole milk
##     4:     4          pip fruit              yogurt    cream cheese
##     5:     5  other vegetables          whole milk   condensed milk
##    ---
## 15292: 15292              yogurt      frozen dessert    domestic eggs
## 15293: 15293                rum     cling film/bags
## 15294: 15294 semi-finished bread       bottled water             soda
## 15295: 15295             chicken      tropical fruit other vegetables
## 15296: 15296       shopping bags
##                            V4
##     1:             ready soups
##     2:
```

```
##     3:
##     4:             meat spreads
##     5: long life bakery product
##    ---
## 15292:               rolls/buns
## 15293:
## 15294:             bottled beer
## 15295:                  vinegar
## 15296:
```

```
y.list <- split(y, seq(nrow(y)))
y$rn = factor(y$rn)
playtrans = as(y, "transactions")
summary(playtrans)
```

```
## transactions as itemMatrix in sparse format with
##  15296 rows (elements/itemsets/transactions) and
##  15951 columns (items) and a density of 0.00031346
##
## most frequent items:
##           V4=           V3=           V2= V1=whole milk    V1=sausage
##          8217          6115          3485           872           825
##       (Other)
##         56966
##
## element (itemset/transaction) length distribution:
## sizes
##     5
## 15296
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5       5       5       5       5       5
##
## includes extended item information - examples:
##   labels variables levels
## 1   rn=1        rn      1
## 2  rn=10        rn     10
## 3 rn=100        rn    100
##
## includes extended transaction information - examples:
##   transactionID
## 1             1
## 2             2
## 3             3
```

+Here is a summary of the different values that are counted the most among the grocery lists, as well as some empty values of shorter lists. You can get rid of these empties by using a nonzero lift.

```
groceryrules = apriori(playtrans,
    parameter=list(support=.005, confidence=.1, maxlen=5))
```

```
## Apriori
```

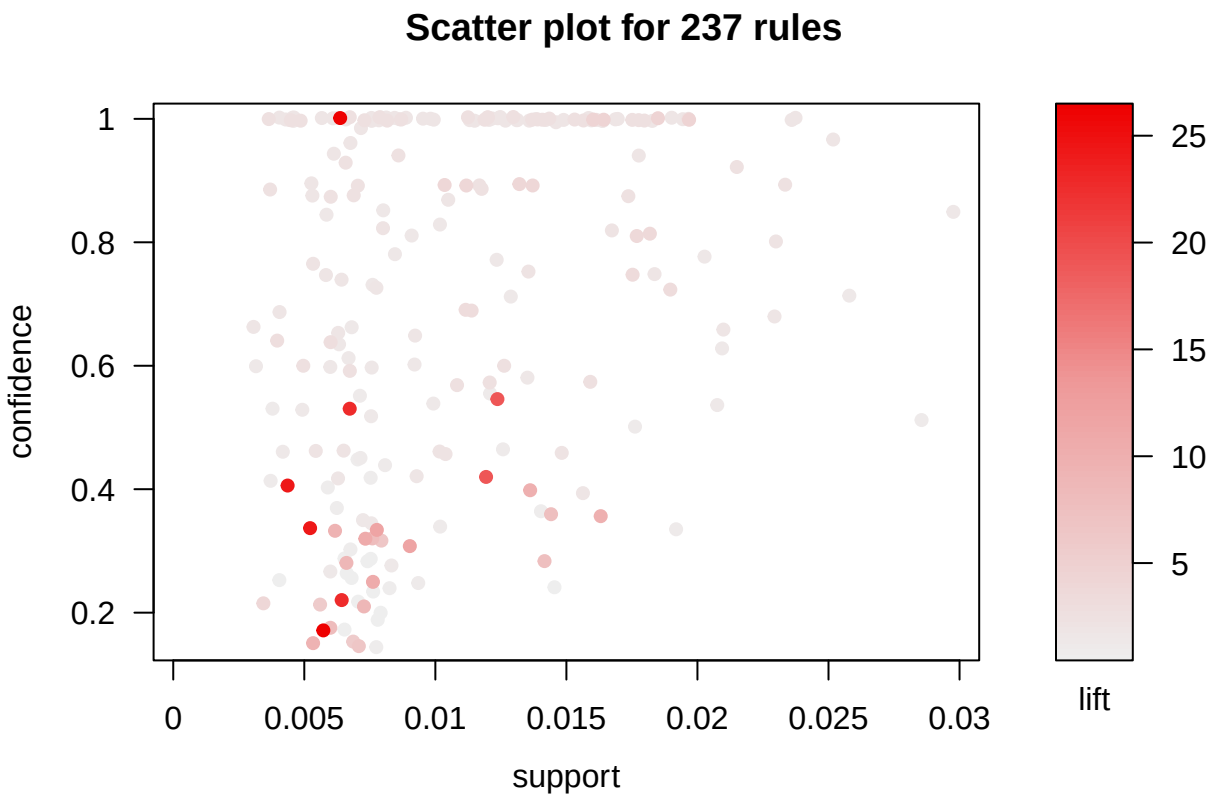
```
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5   0.005      1
##  maxlen target  ext
##       5  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 76
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[15951 item(s), 15296 transaction(s)] done [0.05s].
## sorting and recoding items ... [168 item(s)] done [0.00s].
## creating transaction tree ... done [0.01s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [237 rule(s)] done [0.00s].
## creating S4 object  ... done [0.01s].
```

```
inspect(subset(groceryrules, subset=lift > 5))
```

```
##      lhs                     rhs                    support     confidence
## [1]  {V2=sausage}          => {V1=frankfurter}       0.006472280 1.0000000
## [2]  {V1=frankfurter}      => {V2=sausage}           0.006472280 0.1706897
## [3]  {V3=pip fruit}        => {V2=tropical fruit}    0.005164749 0.5337838
## [4]  {V2=tropical fruit}   => {V3=pip fruit}         0.005164749 0.2200557
## [5]  {V3=root vegetables}  => {V4=other vegetables} 0.005753138 0.4093023
## [6]  {V4=other vegetables} => {V3=root vegetables}  0.005753138 0.3397683
## [7]  {V4=yogurt}           => {V3=whole milk}       0.005360879 0.3346939
## [8]  {V3=whole milk}       => {V4=yogurt}           0.005360879 0.1472172
## [9]  {V4=whole milk}       => {V3=other vegetables} 0.012160042 0.5454545
## [10] {V3=other vegetables} => {V4=whole milk}       0.012160042 0.4217687
## [11] {V1=root vegetables}  => {V2=other vegetables} 0.007191423 0.3225806
## [12] {V2=other vegetables} => {V1=root vegetables}  0.007191423 0.1774194
## [13] {V3=yogurt}           => {V2=whole milk}       0.007191423 0.3170029
## [14] {V2=whole milk}       => {V3=yogurt}           0.007191423 0.1480485
## [15] {V2=tropical fruit}   => {V1=citrus fruit}     0.007518305 0.3203343
## [16] {V1=citrus fruit}     => {V2=tropical fruit}   0.007518305 0.2505447
## [17] {V1=bottled water}    => {V2=soda}             0.006603033 0.2782369
## [18] {V2=soda}             => {V1=bottled water}    0.006603033 0.2069672
## [19] {V2=root vegetables}  => {V3=other vegetables} 0.008891213 0.3333333
## [20] {V3=other vegetables} => {V2=root vegetables}  0.008891213 0.3083900
## [21] {V2=root vegetables}  => {V3=whole milk}       0.005622385 0.2107843
## [22] {V3=whole milk}       => {V2=root vegetables}  0.005622385 0.1543986
## [23] {V3=whole milk}       => {V2=other vegetables} 0.014513598 0.3985637
## [24] {V2=other vegetables} => {V3=whole milk}       0.014513598 0.3580645
## [25] {V1=other vegetables} => {V2=whole milk}       0.013663703 0.3584906
## [26] {V2=whole milk}       => {V1=other vegetables} 0.013663703 0.2812921
##      coverage    lift      count
## [1]  0.006472280 26.372414  99
## [2]  0.037918410 26.372414  99
## [3]  0.009675732 22.743055  79
```

```
## [4]  0.023470188 22.743055  79
## [5]  0.014055962 24.172542  88
## [6]  0.016932531 24.172542  88
## [7]  0.016017259  9.191163  82
## [8]  0.036414749  9.191163  82
## [9]  0.022293410 18.918986 186
## [10] 0.028831067 18.918986 186
## [11] 0.022293410  7.958377 110
## [12] 0.040533473  7.958377 110
## [13] 0.022685669  6.526078 110
## [14] 0.048574791  6.526078 110
## [15] 0.023470188 10.675017 115
## [16] 0.030007845 10.675017 115
## [17] 0.023731695  8.721131 101
## [18] 0.031903766  8.721131 101
## [19] 0.026673640 11.561602 136
## [20] 0.028831067 11.561602 136
## [21] 0.026673640  5.788432  86
## [22] 0.036414749  5.788432  86
## [23] 0.036414749  9.832953 222
## [24] 0.040533473  9.832953 222
## [25] 0.038114540  7.380177 209
## [26] 0.048574791  7.380177 209
```

```
sub0 = subset(groceryrules, subset=lift > 0.01)
plot(sub0, xlim = c(0,.03))
```

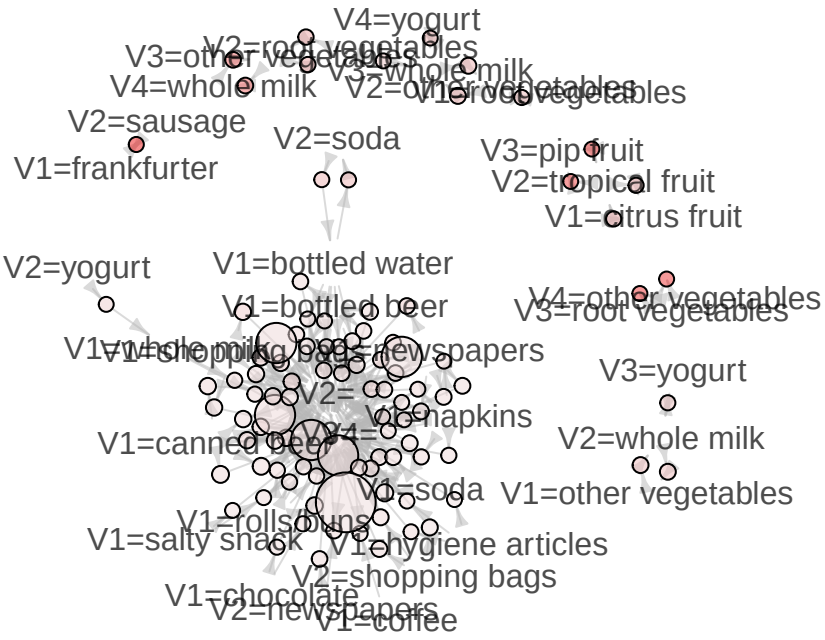


+Here we get a plot of all of the rules in support confidence space, as well as the subset of values with a lift value greater than 5. We see that things like yogurt and milk, frankfurters and sauasages, and fruits and vegetables have high lift values due to their likeness.

```
sub1 = subset(groceryrules, subset=confidence > 0.2 & lift > 0.01)
summary(sub1)
```

```
## set of 228 rules
##
## rule length distribution (lhs + rhs):sizes
##   1   2   3   4
##   3 109  89  27
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.614   3.000   4.000
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.005034   Min.   :0.2070   Min.   :0.005099   Min.   : 0.4299
##  1st Qu.:0.006472   1st Qu.:0.4666   1st Qu.:0.009349   1st Qu.: 1.5798
##  Median :0.009055   Median :0.8121   Median :0.015690   Median : 1.8615
##  Mean   :0.025973   Mean   :0.7291   Mean   :0.043394   Mean   : 2.9725
##  3rd Qu.:0.014971   3rd Qu.:1.0000   3rd Qu.:0.024516   3rd Qu.: 2.5014
##  Max.   :0.537199   Max.   :1.0000   Max.   :1.000000   Max.   :26.3724
##      count
##  Min.   :  77.0
##  1st Qu.:  99.0
##  Median : 138.5
##  Mean   : 397.3
##  3rd Qu.: 229.0
##  Max.   :8217.0
##
## mining info:
##       data ntransactions support confidence
##  playtrans         15296   0.005        0.1
```

```
plot(head(sub1, 100, by='lift'), method='graph')
```




```
options(warn = defaultW)
```

+This connection graph shows an exterior of produce and foods and then the middle is things such as drinks and hygiene materials. This may show that when shoppers get groceries that they are going for one or the other more often than both at the same time.