STA 380 Homework 1

Pooshan Shah

August 8, 2018

```
##  [1]  0.27459028 -0.02857371  1.45856512  0.99038489  2.09651319
##  [6] -0.12090097 -0.08492216  0.21884096  0.34243960  0.76324466
```

```
##  [1]  1.0355153 -1.8963492 -0.6809532 -0.3328306 -0.9303412  0.7316681
##  [7] -1.2623699 -0.6749605  0.4671968 -1.0254782
```

```
##  [1]  0.156703769  1.373811191  0.730670244 -1.350800927 -0.008514961
##  [6]  0.320981863 -1.778148409  0.909503835 -0.919404336 -0.157714831
```

```
##  [1]  0.156703769  1.373811191  0.730670244 -1.350800927 -0.008514961
##  [6]  0.320981863 -1.778148409  0.909503835 -0.919404336 -0.157714831
```

Problem A

Problem B

Exploratory Analysis: green buildings

Summary of Green Data:

```
##  CS_PropertyID        cluster            size            empl_gr
##  Min.   :      1   Min.   :   1.0   Min.   :   1624   Min.   :-24.950
##  1st Qu.: 157452   1st Qu.: 272.0   1st Qu.:  50891   1st Qu.:  1.740
##  Median : 313253   Median : 476.0   Median : 128838   Median :  1.970
##  Mean   : 453003   Mean   : 588.6   Mean   : 234638   Mean   :  3.207
##  3rd Qu.: 441188   3rd Qu.:1044.0   3rd Qu.: 294212   3rd Qu.:  2.380
##  Max.   :6208103   Max.   :1230.0   Max.   :3781045   Max.   : 67.780
##                                                       NA's   :74
##       Rent         leasing_rate       stories            age
##  Min.   :  2.98   Min.   :  0.00   Min.   :  1.00   Min.   :  0.00
##  1st Qu.: 19.50   1st Qu.: 77.85   1st Qu.:  4.00   1st Qu.: 23.00
##  Median : 25.16   Median : 89.53   Median : 10.00   Median : 34.00
##  Mean   : 28.42   Mean   : 82.61   Mean   : 13.58   Mean   : 47.24
##  3rd Qu.: 34.18   3rd Qu.: 96.44   3rd Qu.: 19.00   3rd Qu.: 79.00
##  Max.   :250.00   Max.   :100.00   Max.   :110.00   Max.   :187.00
##
##    renovated         class_a          class_b            LEED
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000000
##  Mean   :0.3795   Mean   :0.3999   Mean   :0.4595   Mean   :0.006841
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000
##
##    Energystar       green_rating          net            amenities
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :0.08082   Mean   :0.08677   Mean   :0.03471   Mean   :0.5266
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##   cd_total_07     hd_total07    total_dd_07   Precipitation
##  Min.   :  39   Min.   :   0   Min.   :2103   Min.   :10.46
##  1st Qu.: 684   1st Qu.:1419   1st Qu.:2869   1st Qu.:22.71
##  Median : 966   Median :2739   Median :4979   Median :23.16
##  Mean   :1229   Mean   :3432   Mean   :4661   Mean   :31.08
##  3rd Qu.:1620   3rd Qu.:4796   3rd Qu.:6413   3rd Qu.:43.89
##  Max.   :5240   Max.   :7200   Max.   :8244   Max.   :58.02
##
##    Gas_Costs        Electricity_Costs  cluster_rent
##  Min.   :0.009487   Min.   :0.01780   Min.   : 9.00
##  1st Qu.:0.010296   1st Qu.:0.02330   1st Qu.:20.00
##  Median :0.010296   Median :0.03274   Median :25.14
##  Mean   :0.011336   Mean   :0.03096   Mean   :27.50
##  3rd Qu.:0.011816   3rd Qu.:0.03781   3rd Qu.:34.00
##  Max.   :0.028914   Max.   :0.06280   Max.   :71.44
##
```

Hypothesis: I do not agree with the approach the stats-guru took in finding the answer to the problem. There are other factors involved that could affect the rent and green status of a building. Some factors I believe that could be significant are: cluster_rent, gas/electricity costs, net, amenities, class, size, and leasing rate.

Extract the buildings with green ratings

```
green_only = subset(green, green_rating==1)
not_green = subset(green, green_rating ==0)
```

Basic EDA

Summary of Green Data Rent

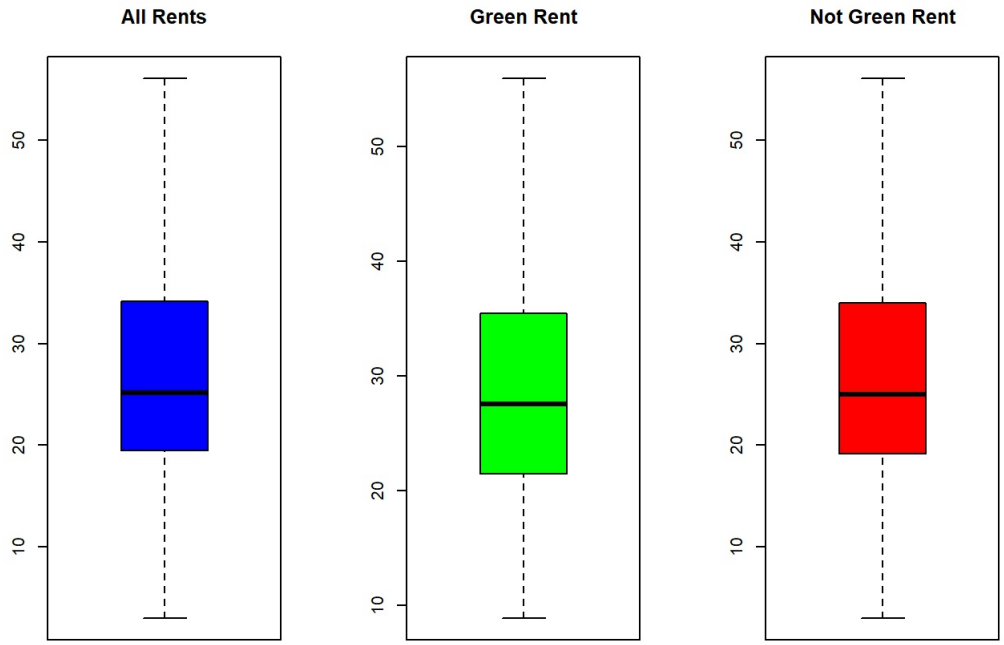
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.50   25.16   28.42   34.18  250.00
```

Summary of only Green Buildings' Rent

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.50   27.60   30.02   35.50  138.07
```

Summary of non-Green Buildings' Rent

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.18   25.00   28.27   34.00  250.00
```



Comments: Here we get a sense of the green data as a whole, and once we split the data into green buildings, and non-green buildings, we can see holistically how they differ at the rent level. A difference is see-able between green and non-green rent.

```
summary(green_only$size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10560  120000  241150  325781  417446 1721242
```

```
summary(not_green$size)
```

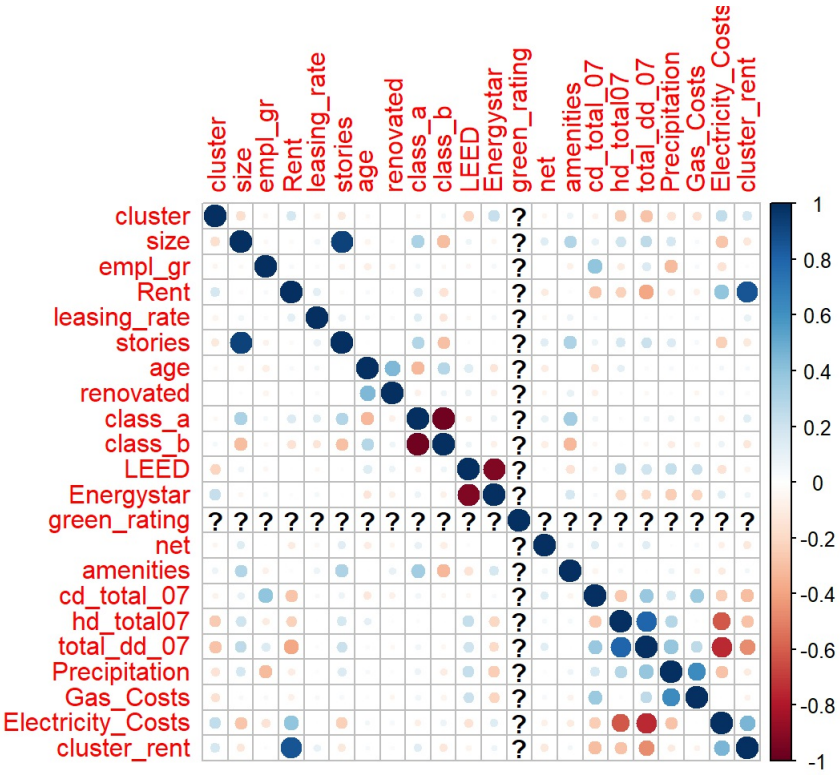
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1624   46043  118696  225977  279411 3781045
```

Comments: the size of this bulding will be close to the median of all green buildings and between the mean and 3rd quartile for non-green.

```
summary(green_only$age)
```

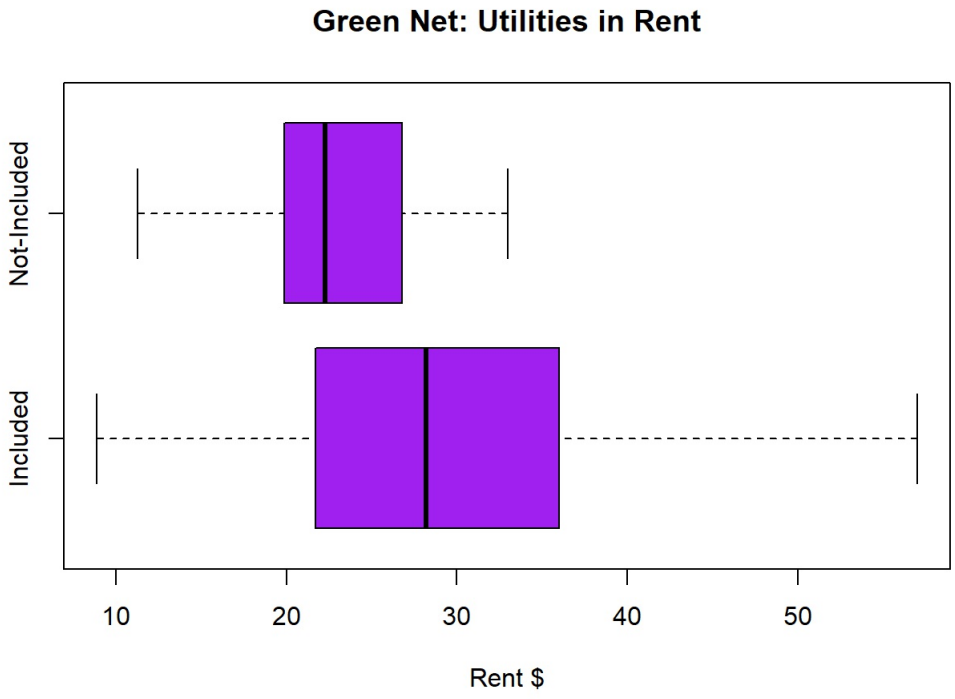
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   18.00   22.00   23.85   26.00  116.00
```

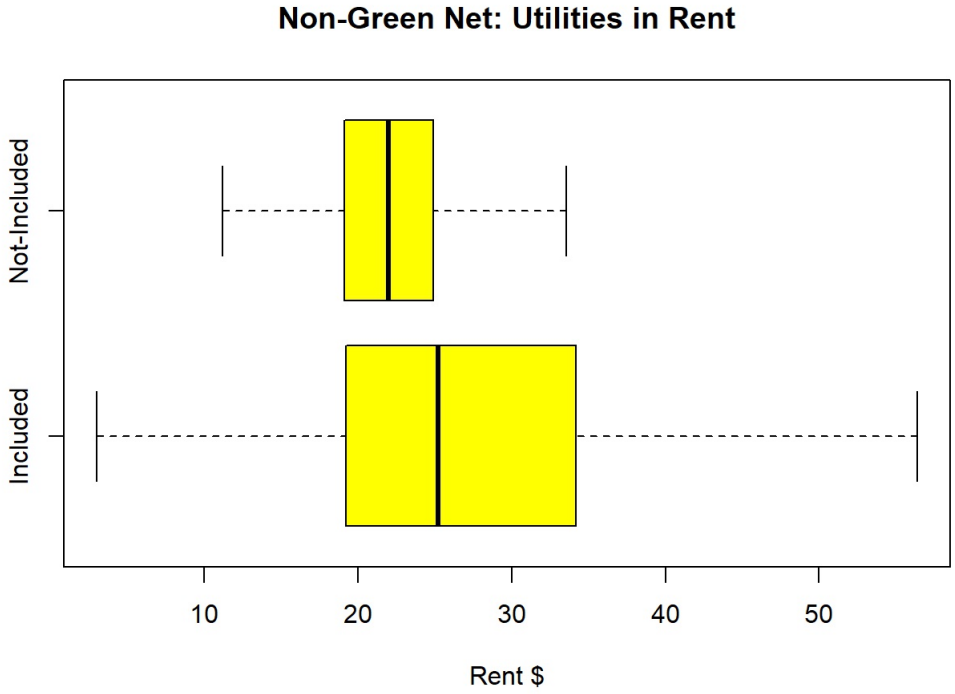
Comments: from this summary of green buildings, we see that the median age for a green building is around 22, so the stats-guru might be overestimating the age of the upcoming building at 30 years.



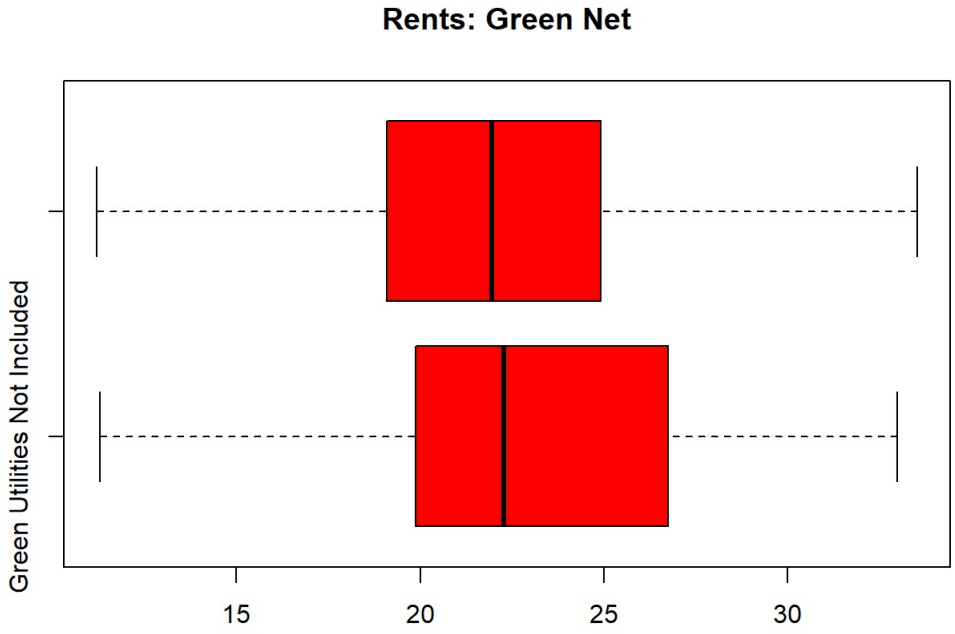
Comments: this helps to show which variables might be correlated to get a better idea of which variables to look at. We see: class, net, cold/hot and total days, and electricity.

Net

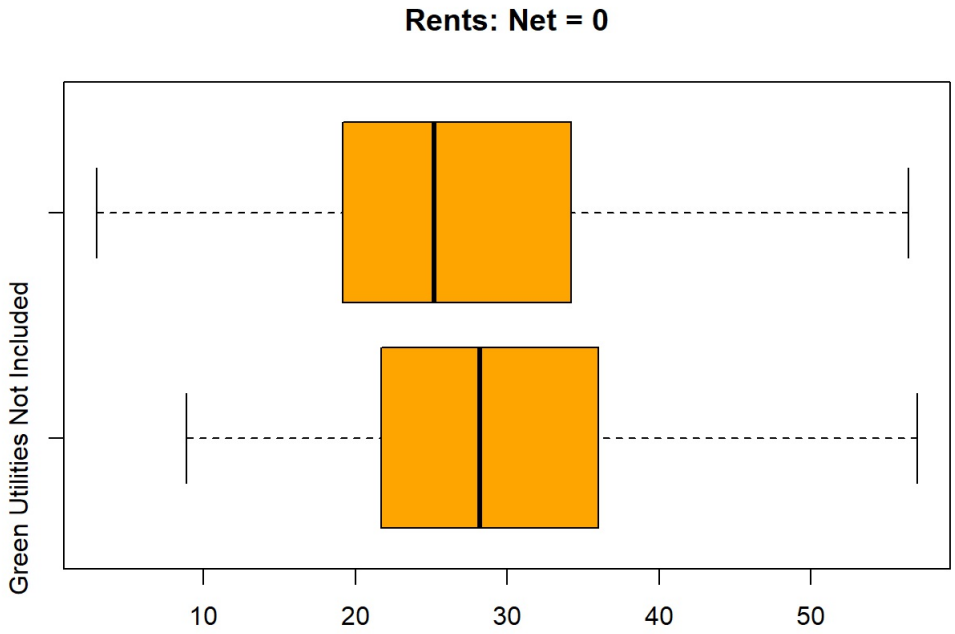




Comments: with Net, we see a big difference- both green and non green have big differences in rent based on net. Seems like rent wise, we should split the groups based on net as well.

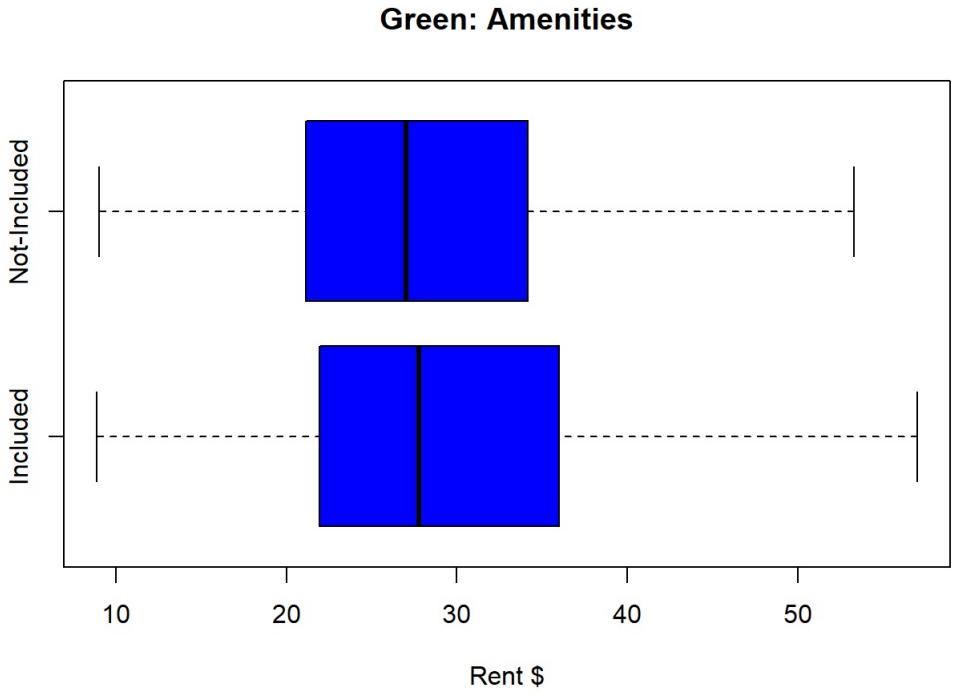


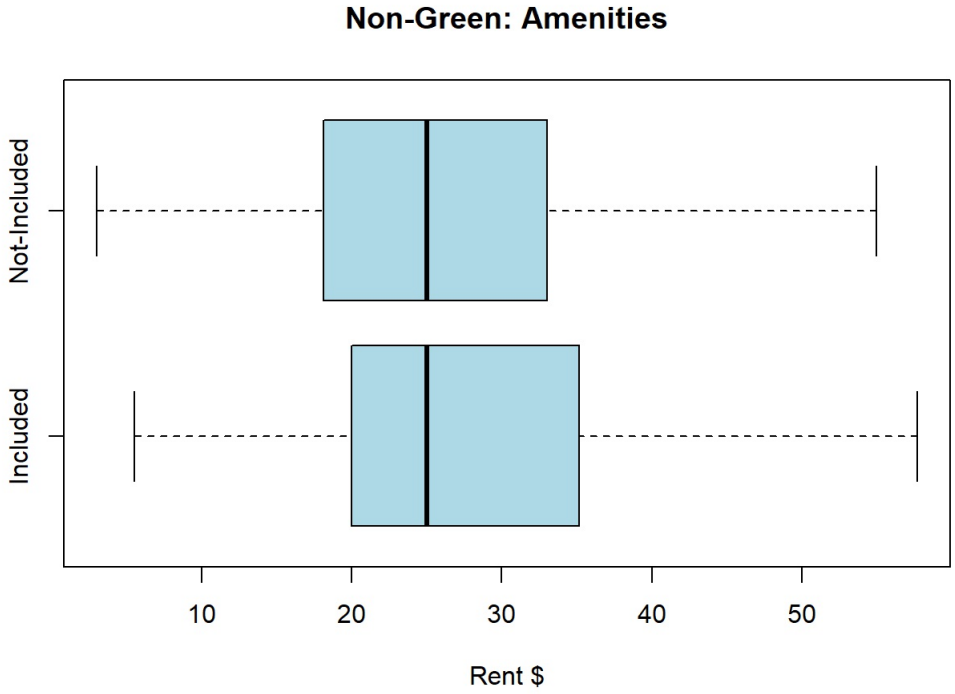
Comments: we see how rent is almost the same for both green and non-green when not including utilities- which is expected



Comments: this is could be a better indicator of the difference between green and non-green rents

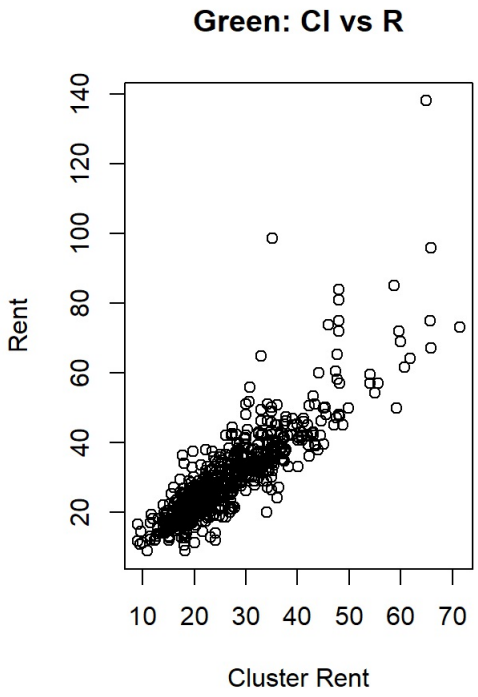
Amenities

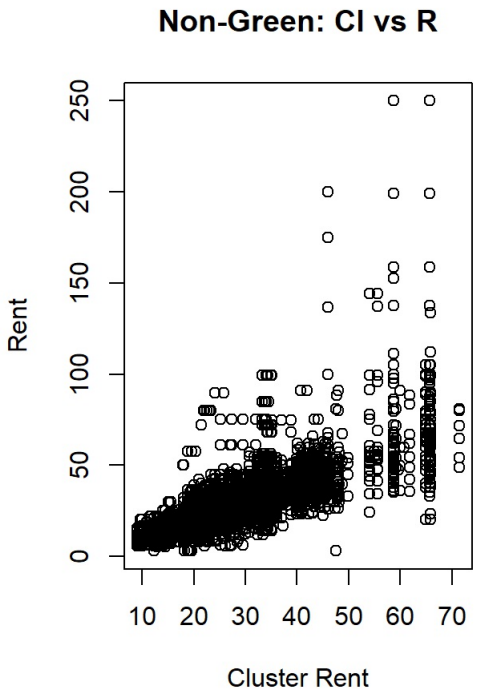




Comments: very similar in rent for both- no significant enough through graph

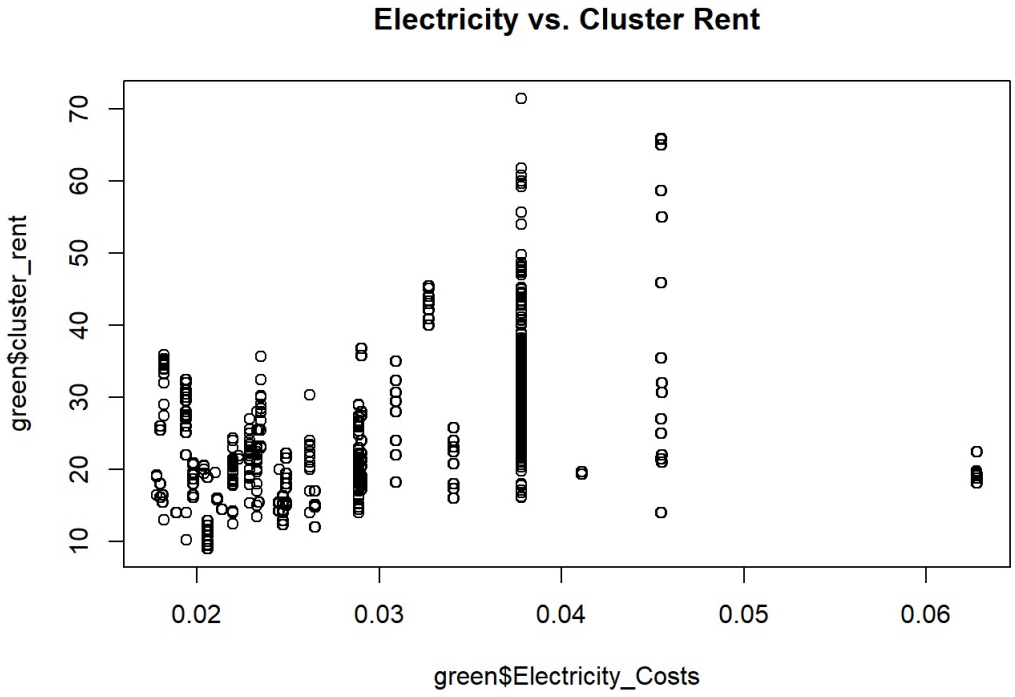
Cluster-Rate

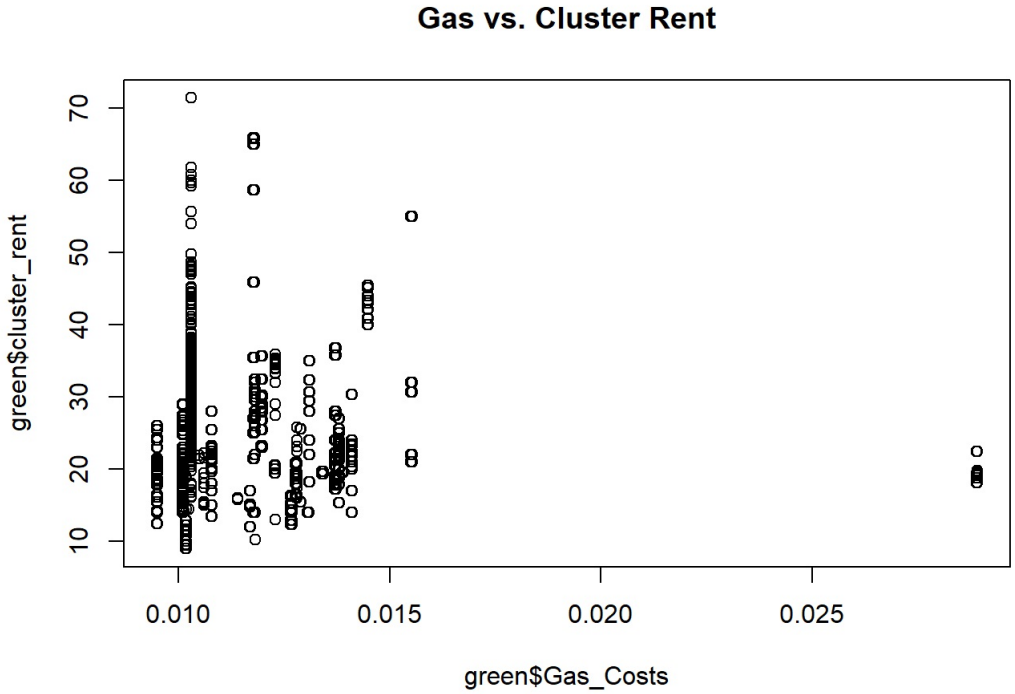




Comments: graphs show a positive correlation for both green and non-green. The slope for green buildings looks a lot steeper than for non-green. This deinitely could affect the rent price based on which cluster the green building is in. Could be that the higher the cluster price, the greater the difference between the green rent and cluster rent.

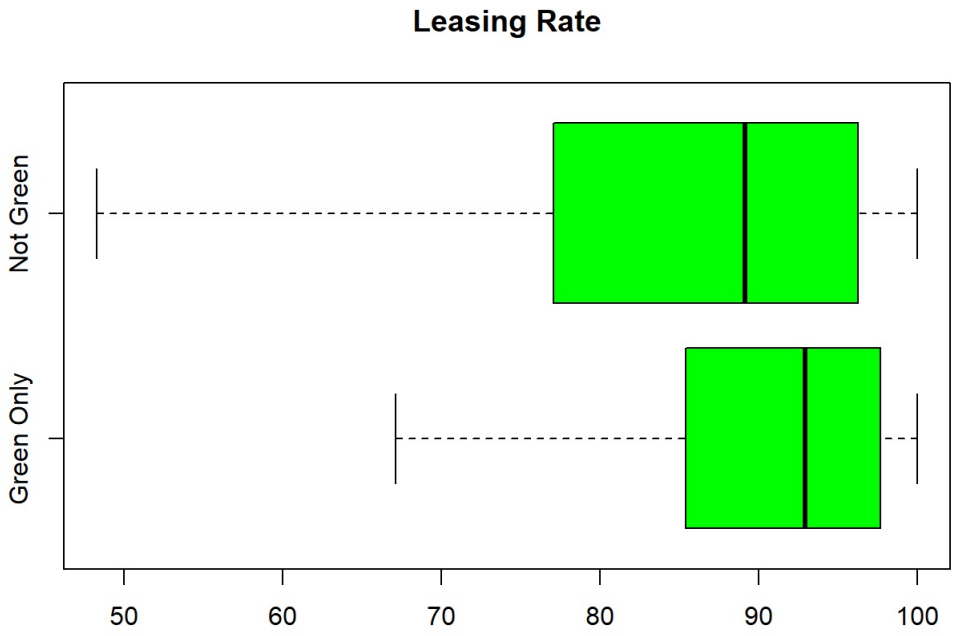
Cluster Rate and Utility Costs





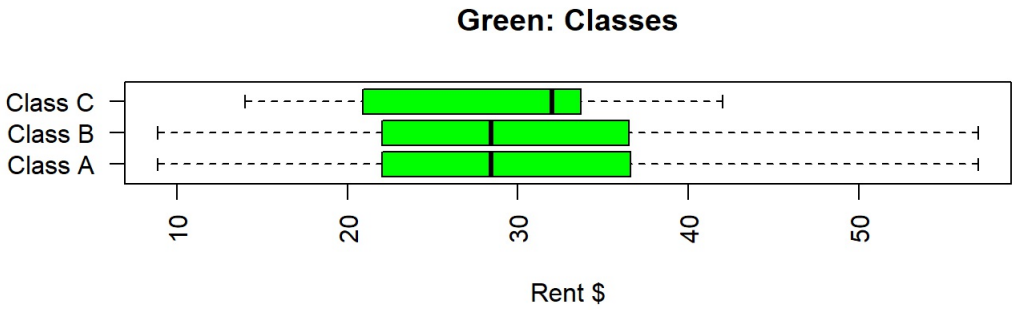
Comments: there seems to be some positive correlation between cluster rent and electricity cost but not really any for gas costs.

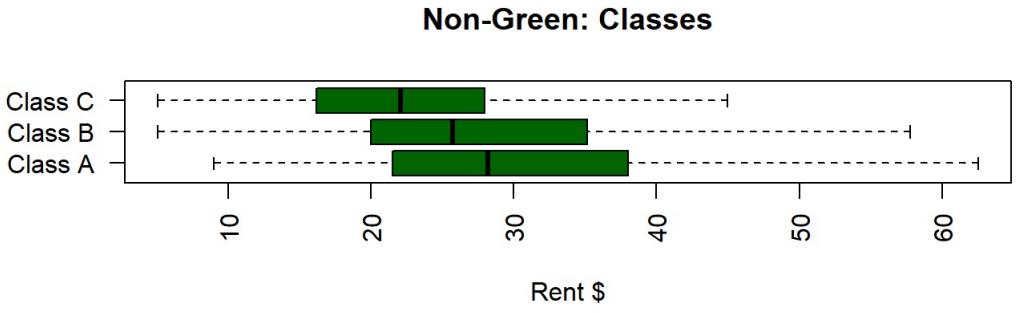
Leasing



Comments: leasing rate seems to be higher for green buildings. This may confirm that green buildings are more attractive to work in, and also the developer may not have to worry about empty space just sitting there as much as non-green buildings.

Class A,B,C





Comments: classes don't seem to be very different for green except for class c which is very surprising (could just be an anomaly), but since the pattern isn't apparent, we will not worry about it. But you can easily see positive correlation for non-green classes.

Size

Comments: not really any correlation for all buildings- mostly clustered below 250,000 sqrfeet

More analysis based on EDA

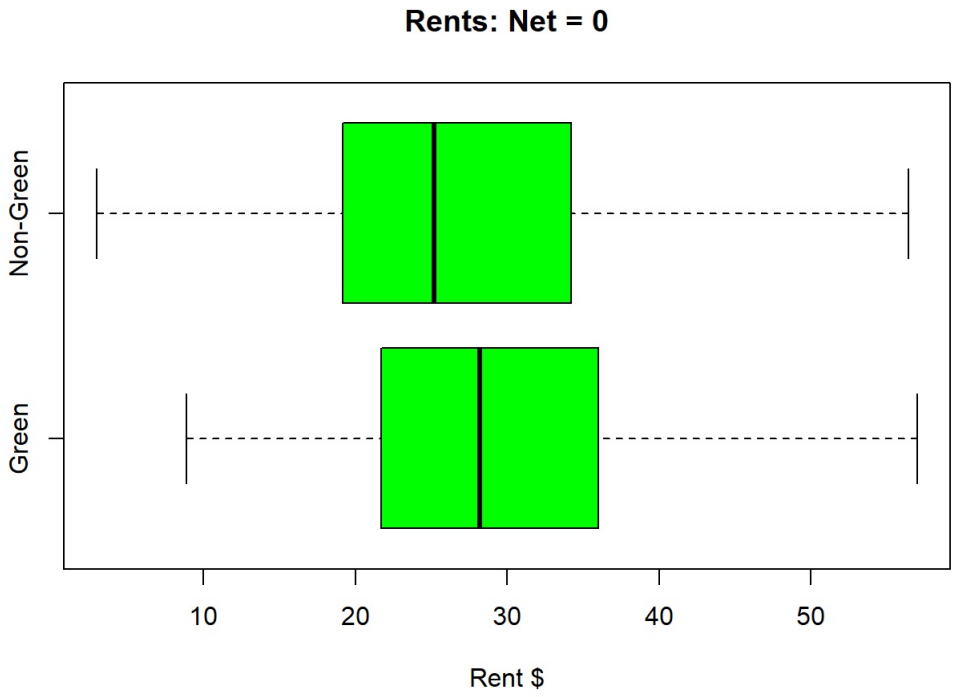
Using Net to further subset the data into: green_only_net and not_green_net. These include the buildings that do include utilities in the rent.

```
summary(green_only_net$Rent)
```

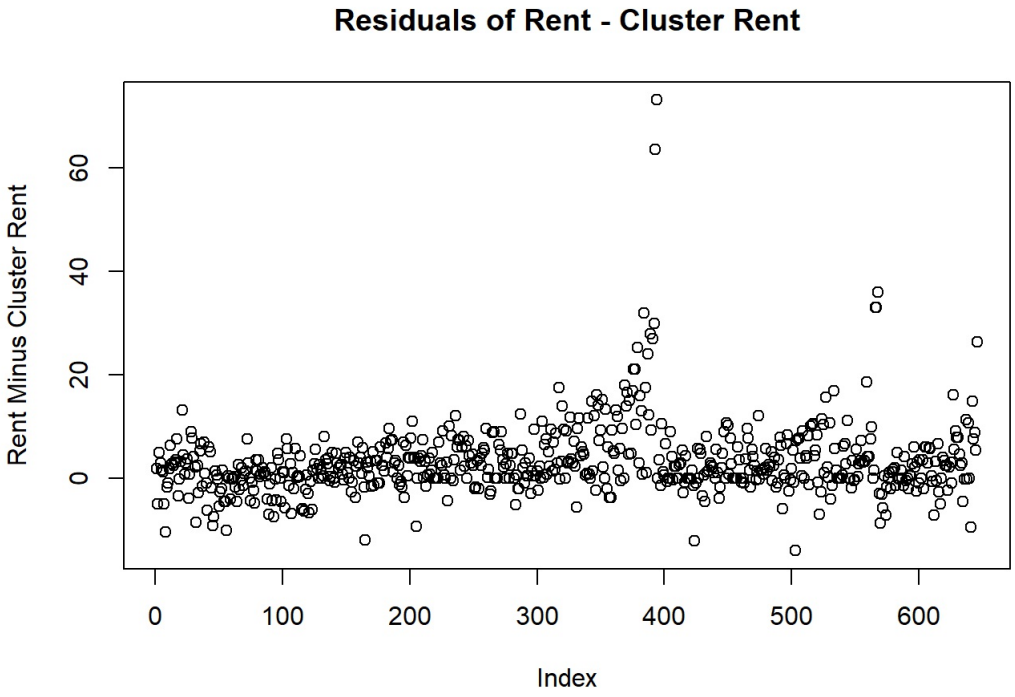
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.69   28.20   30.36   35.98  138.07
```

```
summary(not_green_net$Rent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.20   25.19   28.40   34.20  250.00
```



Comments: given the summaries and box plots, we can use these to furter deduce a more accurate difference in overall rent prices.



Summary of Rent - Cluster Rent from the green_only_net data

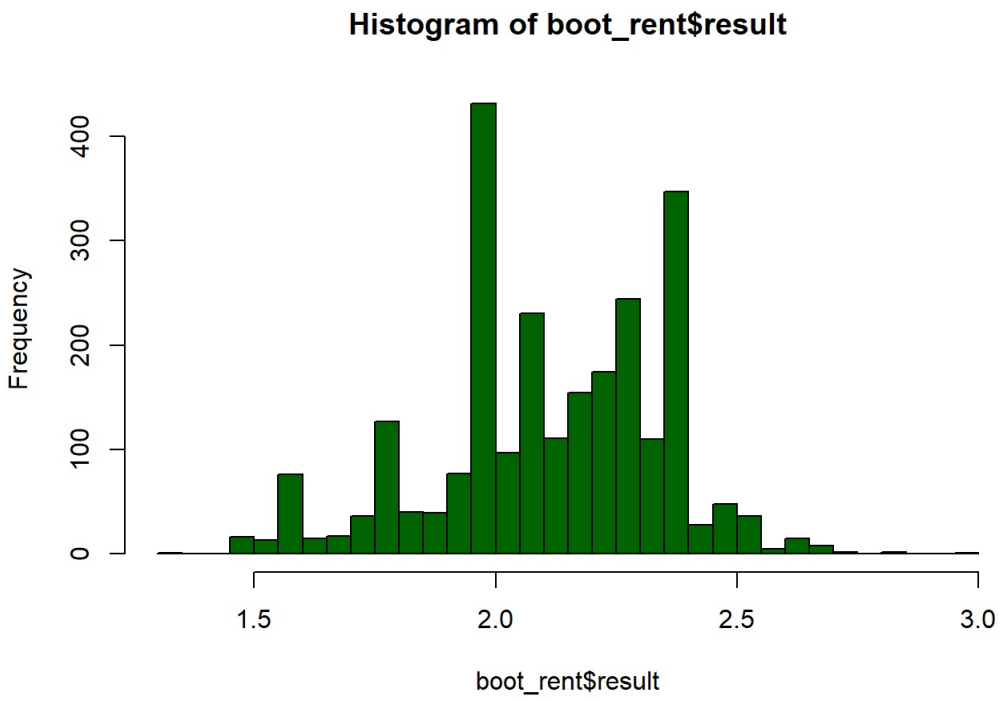
```
summary(rent_diff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -14.000   0.000   2.118   3.187   5.487  73.070
```

Comments: this is another way we could find the difference in the rent prices but this seems a little more bias as it weighs every data point the same- so outliers and other external effects have a stronger effect. This method is more appealing as it accounts for cluster similarities instead of just being general. Location plays an important role in rent and the clusters are the only source of location data.

Let’s try bootstrapping the rent difference for a more robust answer

Bootstrap the Rent-Cluster Rent median for the most subsetted table: green_only where net = 0




```
##      result
##  Min.   :1.340
##  1st Qu.:2.000
##  Median :2.118
##  Mean   :2.122
##  3rd Qu.:2.300
##  Max.   :2.960
```

We subtracted the subsetted green building cluster rent from the green building actual rent and bootstrapped the median. In this, we leveraged the cluster rent which I believe is more correlated to the green building rent than the overall mean of the non-green building rent.

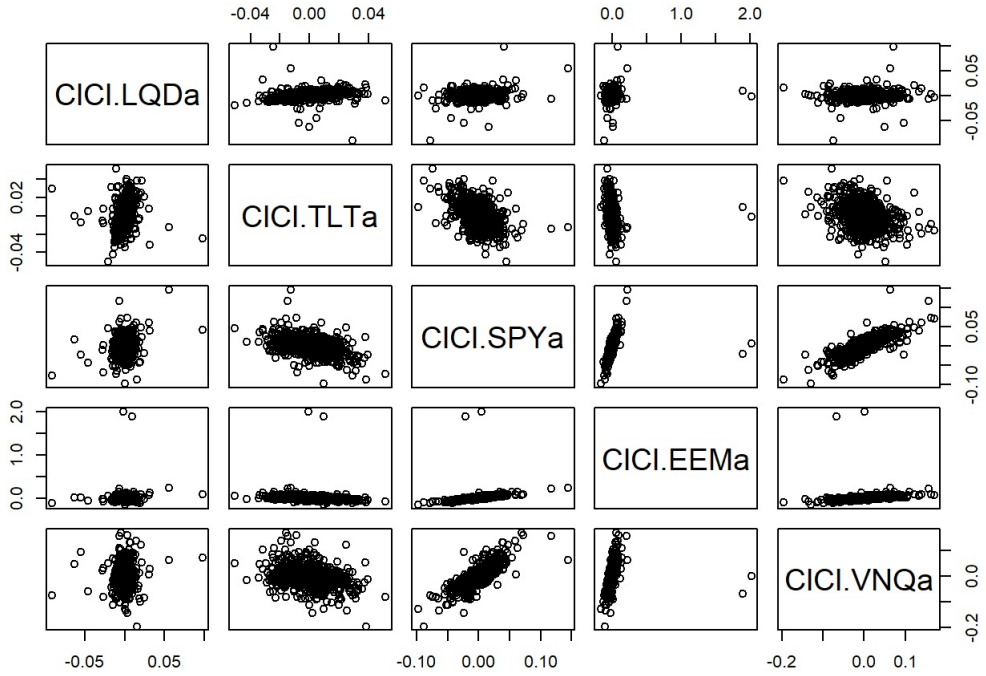
In conclusion, I do not agree with how the stats-guru got his answer. From our analysis, we see that there are multiple other variables involved in significantly changing the rent of a green building such as presence of utilitie and their prices, cluster rent, hot/cold days (from correlation plot), etc. In this analysis, we saw how net and cluster rent were correlated to the green building rent and we used this knowledge to help us find more evidence of a more accurate profit margin. The analysis can be improved even more by narrowing down our sample to more relevant and similar types of buildings of which we can compare. It would also be a lot of help to get more data from the developer on the type of building she wants to make regarding utilities, class, cluster, etc. It would be the most help to us to find the location of each cluster presented in the data in order to figure out which cluster is closest to the developer's building's location.

Bootstrapping

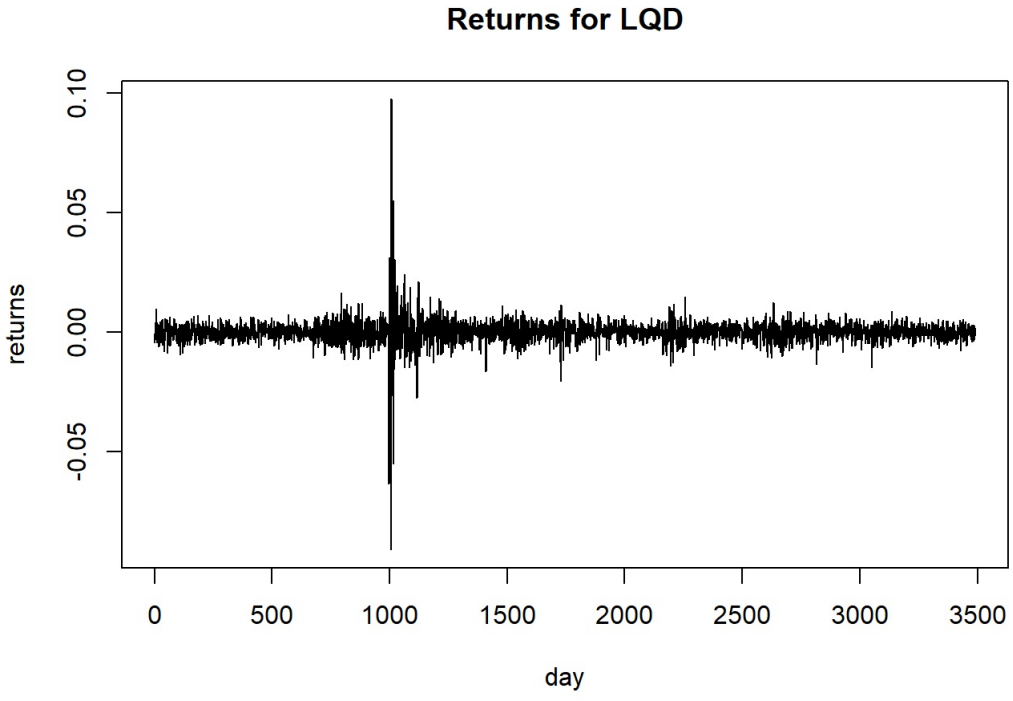
```
##               ClCl.LQDa    ClCl.TLTa     ClCl.SPYa    ClCl.EEMa
## 2004-09-30 -0.002944009 -0.003837860 -0.0007152540  0.005244650
## 2004-10-01 -0.004428211 -0.004994971  0.0169112381  0.019130453
## 2004-10-04 -0.000631532  0.001486621  0.0016717466  0.017064881
## 2004-10-05  0.004152803  0.000000000  0.0005271082 -0.007382617
## 2004-10-06 -0.002876913 -0.004567276  0.0068480947  0.011550711
## 2004-10-07 -0.002163971 -0.004358763 -0.0107255232 -0.008522252
##                ClCl.VNQa
## 2004-09-30  0.0080241127
## 2004-10-01  0.0185074627
## 2004-10-04  0.0042985932
## 2004-10-05 -0.0001946109
## 2004-10-06  0.0046702083
## 2004-10-07 -0.0102654075
```

```
##                ClCl.LQDa     ClCl.TLTa     ClCl.SPYa     ClCl.EEMa
## 2018-08-03  0.0031263569  0.0049734807  0.0042848222  0.007750171
## 2018-08-06  0.0009523071  0.0005871498  0.0036671685 -0.007690568
## 2018-08-07 -0.0031136567 -0.0051974432  0.0033023185  0.010257602
## 2018-08-08 -0.0015616866  0.0010112328 -0.0004201835 -0.001353836
## 2018-08-09  0.0007820299  0.0084182170 -0.0013661599 -0.002937122
## 2018-08-10  0.0016497613  0.0072627347 -0.0067001191 -0.021300747
##                ClCl.VNQa
## 2018-08-03  0.0108285677
## 2018-08-06 -0.0013240370
## 2018-08-07 -0.0018078944
## 2018-08-08 -0.0057957498
## 2018-08-09  0.0004858028
## 2018-08-10 -0.0076474511
```

Comment: We can see that after cleaning, the data starts at: 9/30/2004 and ends: 8/7/2018

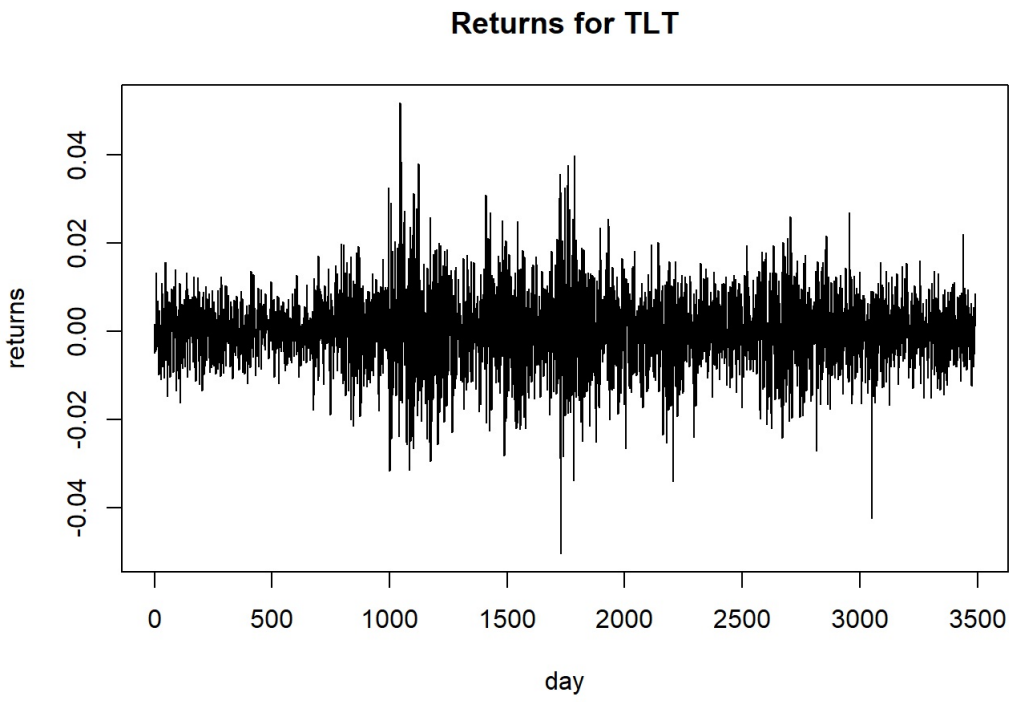


Comment: From this graph we can see the various correlation between the stocks. We can see that the stocks which produce a relatively flat or horizontal line means that one of the variables is more immune to change (ex. LQD)– helpful for portfolio 2 because that means the stock is low risk and low return. We see a positive correlation between VNQ and SPY which might be helpful in the portfolio 3 for high risk and high return.



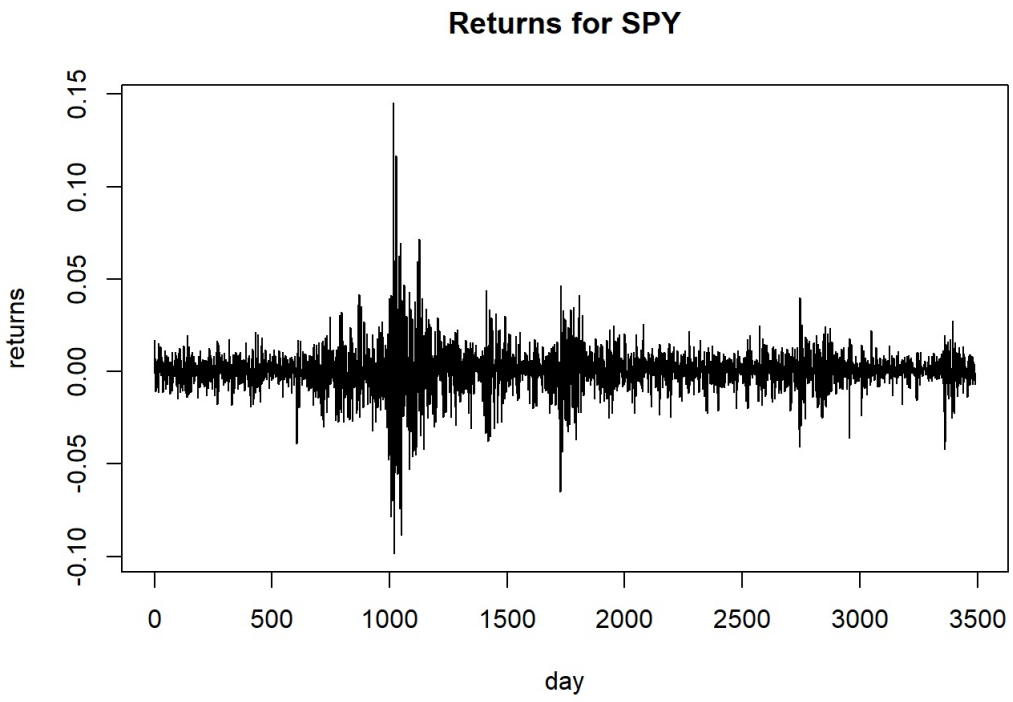
```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## -0.0911111 -0.0018985  0.0003581  0.0001916  0.0023453  0.0976772
```

Comment: **Graph** First thing that jumps out at me is the high and low around day: 1000, there was a high positive return as well as a relatively low negative return. However, the remaining points seem to be stable and within a small margin. Looks like a safe stock.
Summary To steer clear of outliers, I look mainly at the median and 1st/3rd Quartiles for the returns. Median is small and quartile range is minimum. Safe stock.



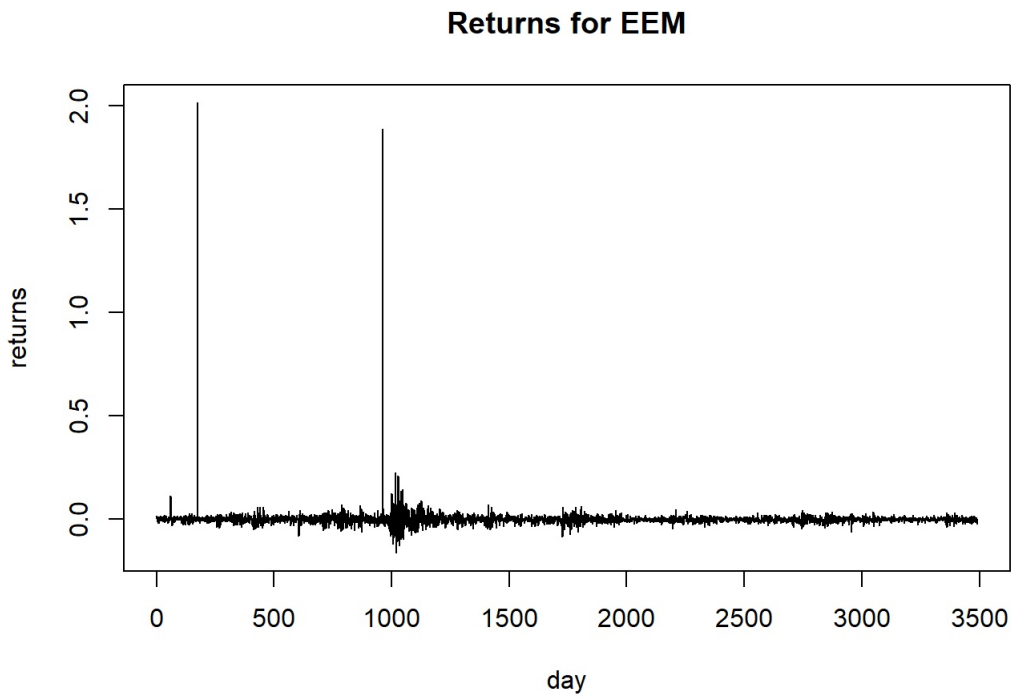

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## -0.0504495 -0.0047801  0.0005119  0.0002656  0.0053240  0.0516616
```

Comment: **Graph** The graph makes the stock look as if it is volatile, but in fact, it is relatively stable looking at the y-value. The highest it goes is 0.04 return and lowest is -0.04. Looks like a safe stock.
Summary Median is small and quartile range is minimum. Safe stock.



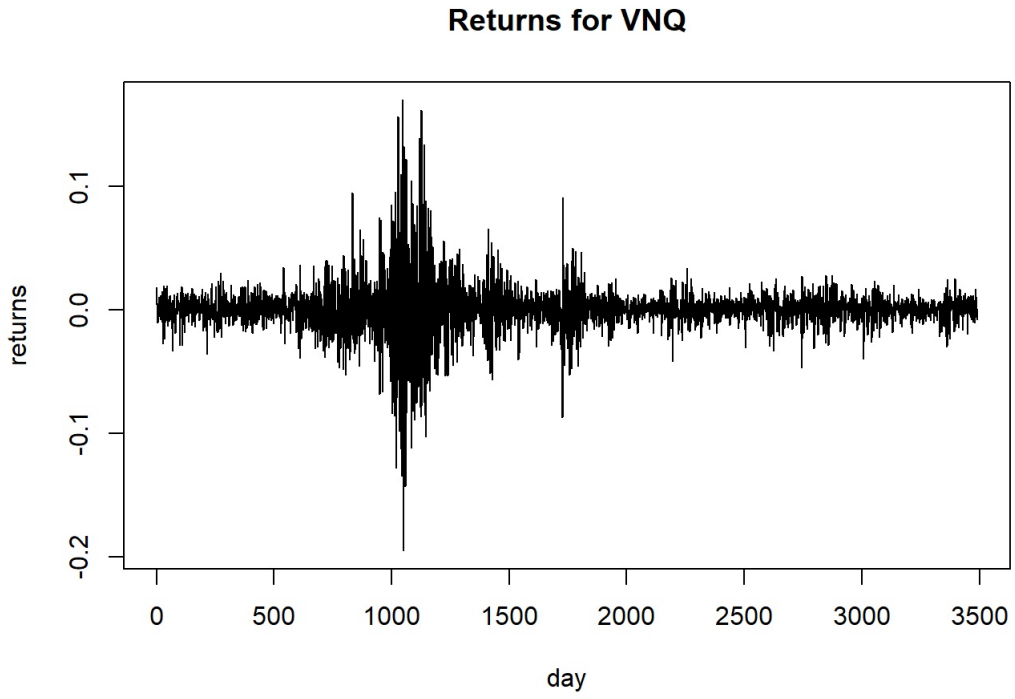

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## -0.0984477 -0.0037570  0.0007041  0.0004143  0.0053883  0.1451977
```

Comment: **Graph** Between day 100 and 1700ish there seems to be some more extreme high and low returns compared to other graphs. This stock seems change a lot but not too much as it is mostly margined between -0.05 and 0.05. Seems like could be either safe or risky. **Summary** Median is larger but quartile range is still relatively small. Safe stock.



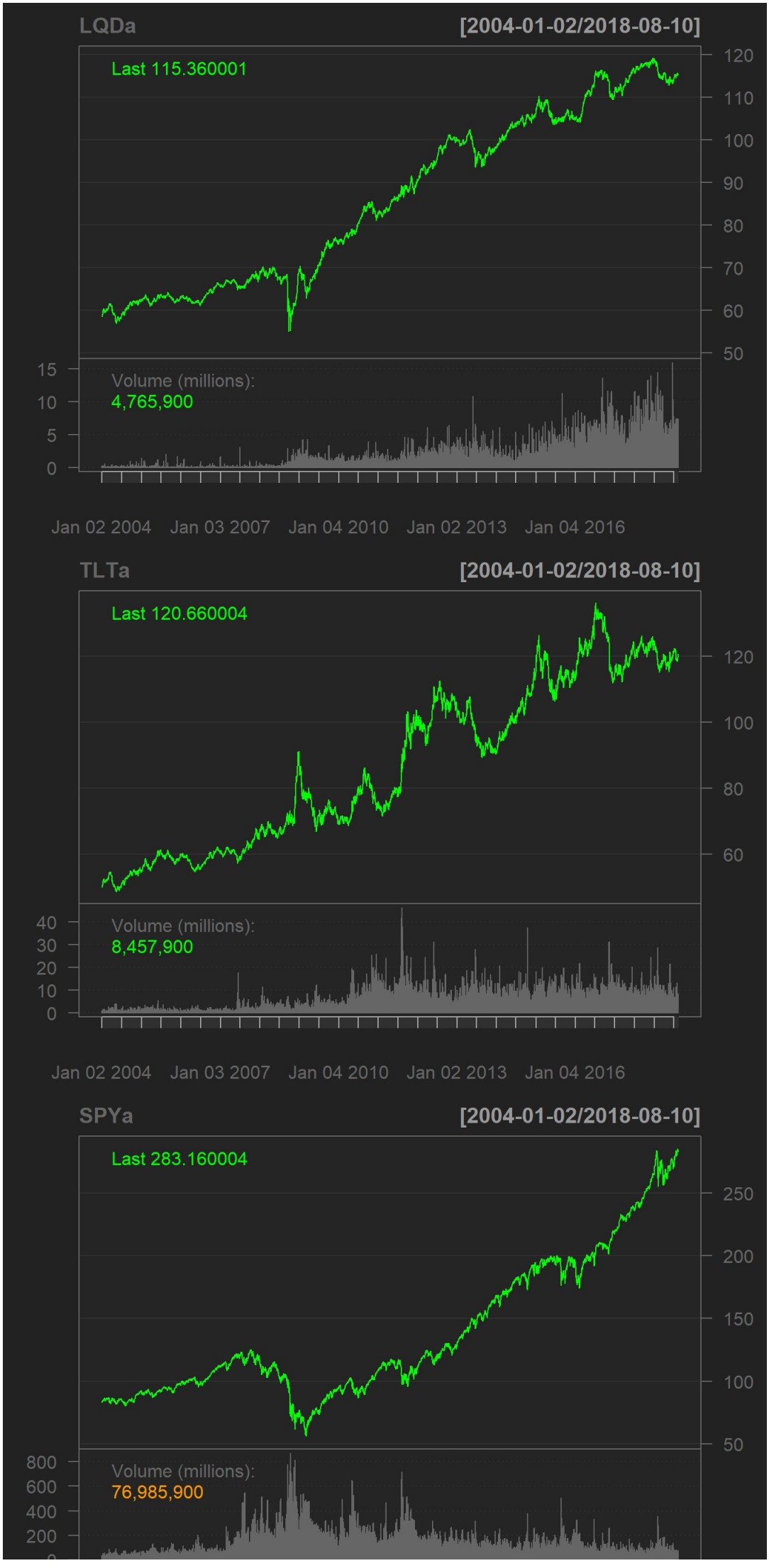

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.1616620 -0.0082770  0.0009536  0.0016786  0.0094616  2.0157291
```

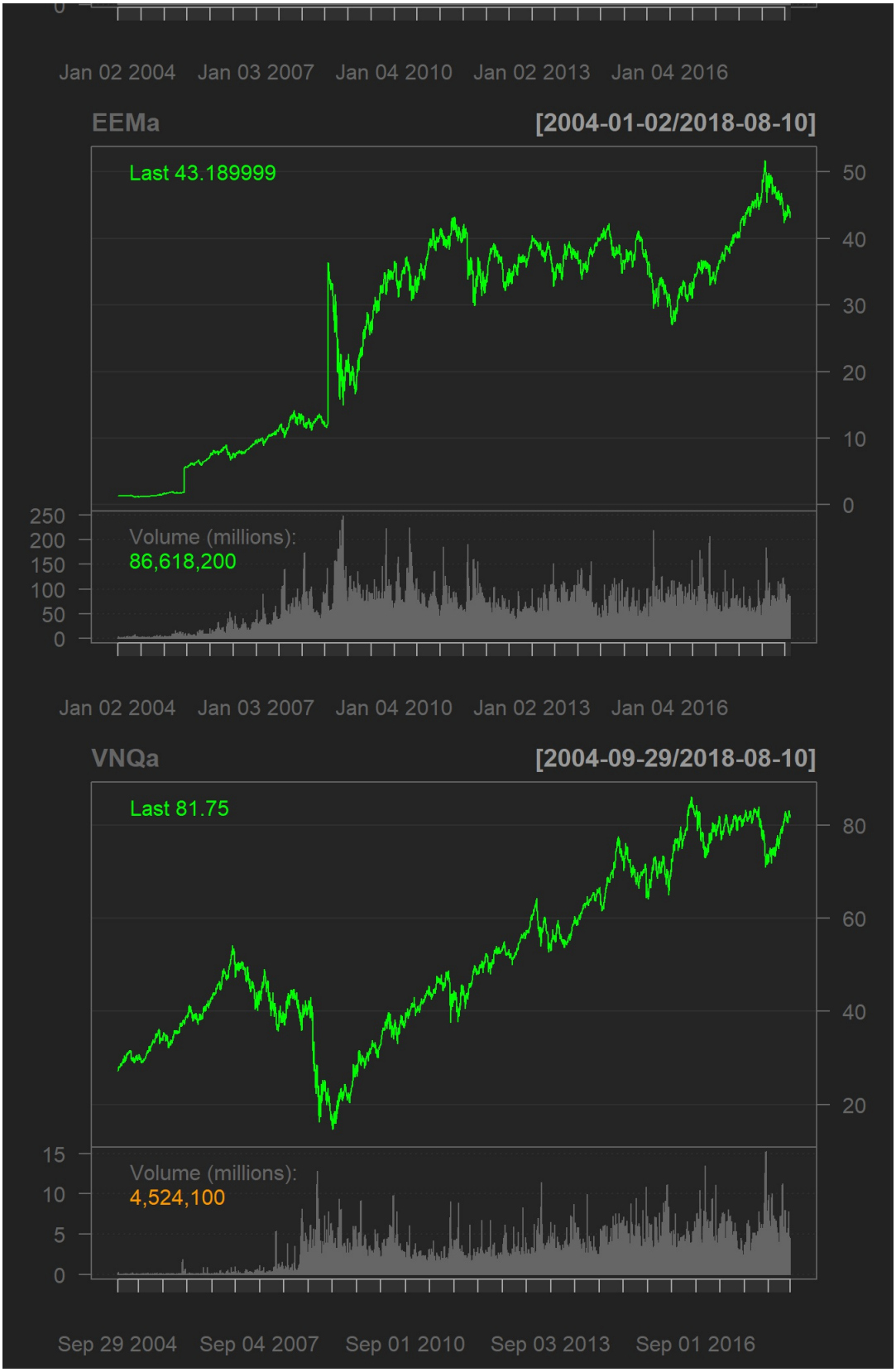
Comment: **Graph** It is hard to tell exactly how the non-outlier days are doing on this graph but when looking at its subsets (ex. days: 1500-3400), the returns are relatively high and low– going above 0.05 and below -0.05. Seems like more of a riskier stock. **Summary** Median is large. Quartile range is much larger comparitively on both negative and positive sides. Risky stock.




```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.1951372 -0.0065597  0.0007728  0.0005096  0.0076721  0.1700654
```

Comment: **Graph** Similar to EEM, this stock has more static and thus seems more volatile. Seems like a riskier stock. **Summary** Similar median to SPY but larger quartile range. Risky stock.


LQDa
[2004-01-02/2018-08-10]
Last 115.360001
120
110
100
90
80
70
60
50
15
10
5
0
Volume (millions):
4,765,900
Jan 02 2004
Jan 03 2007
Jan 04 2010
Jan 02 2013
Jan 04 2016
TLTa
[2004-01-02/2018-08-10]
Last 120.660004
120
100
80
60
40
30
20
10
0
Volume (millions):
8,457,900
Jan 02 2004
Jan 03 2007
Jan 04 2010
Jan 02 2013
Jan 04 2016
SPYa
[2004-01-02/2018-08-10]
Last 283.160004
250
200
150
100
50
800
600
400
200
Volume (millions):
76,985,900

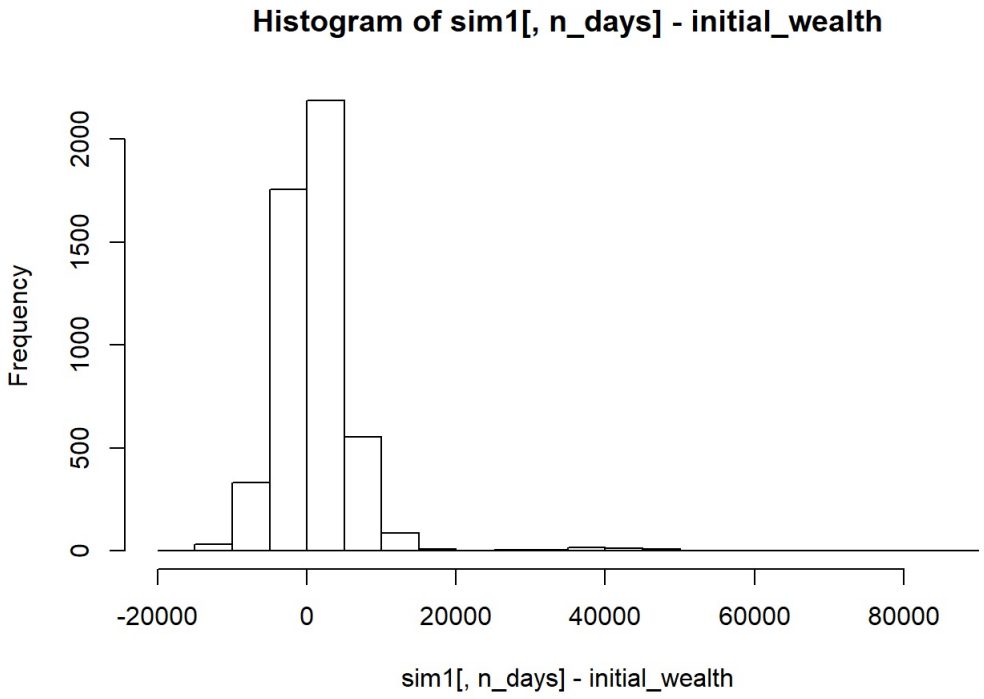


Comment: These chart series graphs show the prices over time. All the graphs seems to match up to my hypothesis described in the previous return graphs except for SPY. The riskier stocks show steep rises and falls of the price. What's interesting for SPY is that while it looks like a smoother curve, the range for the y-values are 250. There seems to be a solid increase in price but the volatility could be large. I will keep it as a safe stock.

Portfolio 1

This portfolio splits the wealth and weight evenly among the stocks at 0.2 each. Initial Wealth: $100,000 Runs 5000 times.

Mean wealth: 1.010887410^{5}



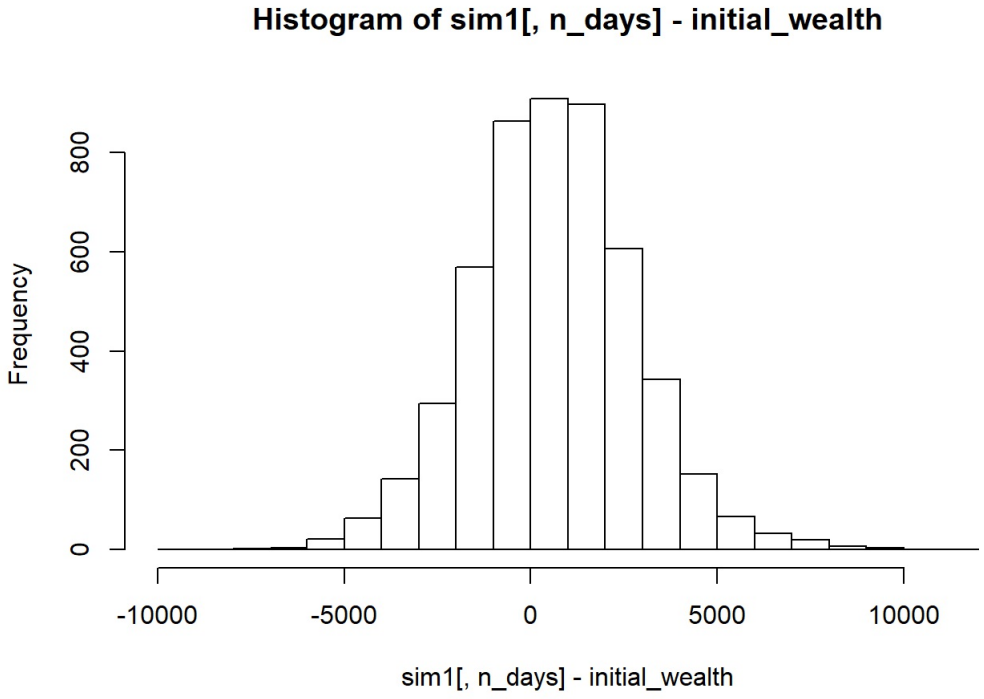
Comment: Graph shows the distribution of returns over the 5000 runs. Mean seems to be greater than 0 with many outlier in the 1000s positive.

Calculate 5% value at risk -5795.4496449 Calculate 95% value at risk 7909.774586

Portfolio 2

This portfolio leverages something that seems safer than the even split, comprising investments in at least three classes. Stocks used: LQD: 0.35, TLT: 0.3, SPY: 0.35 I selected these weights after many instances of trial and error to find the minimum 5% at risk number. Initial Wealth: $100,000 Runs 5000 times.

Mean wealth: 1.006126810^{5}



Comment: Graph shows the distribution of returns over the 5000 runs. Mean seems to be very close to 0 with a short distribution from about -$5000 to $7000.

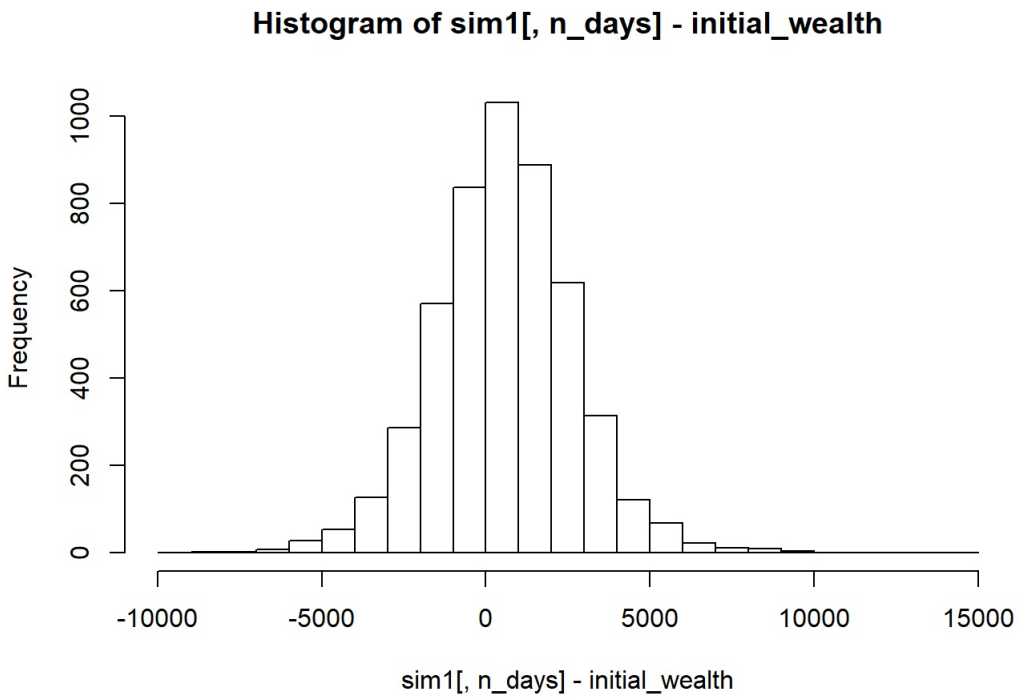
Calculate 5% value at risk -2950.9927336 Calculate 95% value at risk 4154.464743

Portfolio 3

This portfolio leverages something more aggressive, comprising investments in at least two classes/assets. By more aggressive, I mean a portfolio that looks like it has a chance at higher returns, but also involves more risk of loss.

Stocks used: EEM: 0.8, VNQ: 0.2 I selected these weights after many instances of trial and error to find the maximum lowest 5% and highest 95% at risk number. Initial Wealth: $100,000 Runs 5000 times.

Mean wealth: 1.005701710^{5}



Comment: Graph shows the distribution of returns over the 5000 runs. Mean seems to be above 0 with a large distribution from about -$50,000 to $200,000.

Calculate 5% value at risk -2819.7602608 Calculate 95% value at risk 3957.1704119

Comparison of Portfolios:

Looking at the portfolios, we see exactly what we expect from the returns: Portfolio 1 (even): +$1,200.2 Portfolio 2 (safe): +$659 Portfolio 3 (risky): +2,817.5 All of the portfolios gave us a positive return back with 2 being the lowest and 3 being the highest. Looking at the 5% and 95% risk, each portfolio essentially doubles its risk and returns as it gets riskier. Risk: -$3000(P2) to -$6000(P1) to -$11,000(P3). Returns: $4000(P2) to $8000(P1) to $16,000(P3). This doubling is represented in the mean returns as well: $600 to $1,200 to $3,000. From this model, it seems that each portfolio is scaled similarly– the decision would come down to how long does the invester want to play in this market with this money.

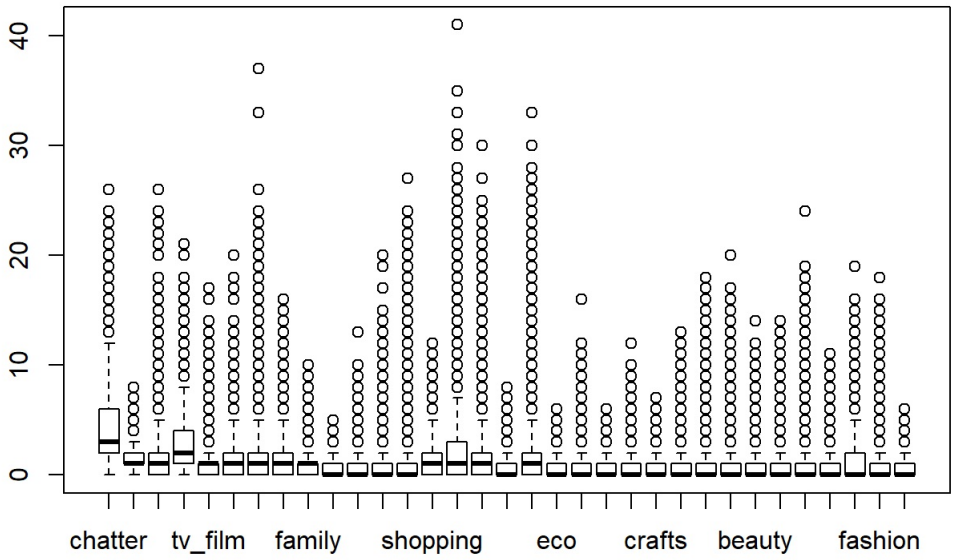
Market Segment Problem

Summary of Social Marketing Data

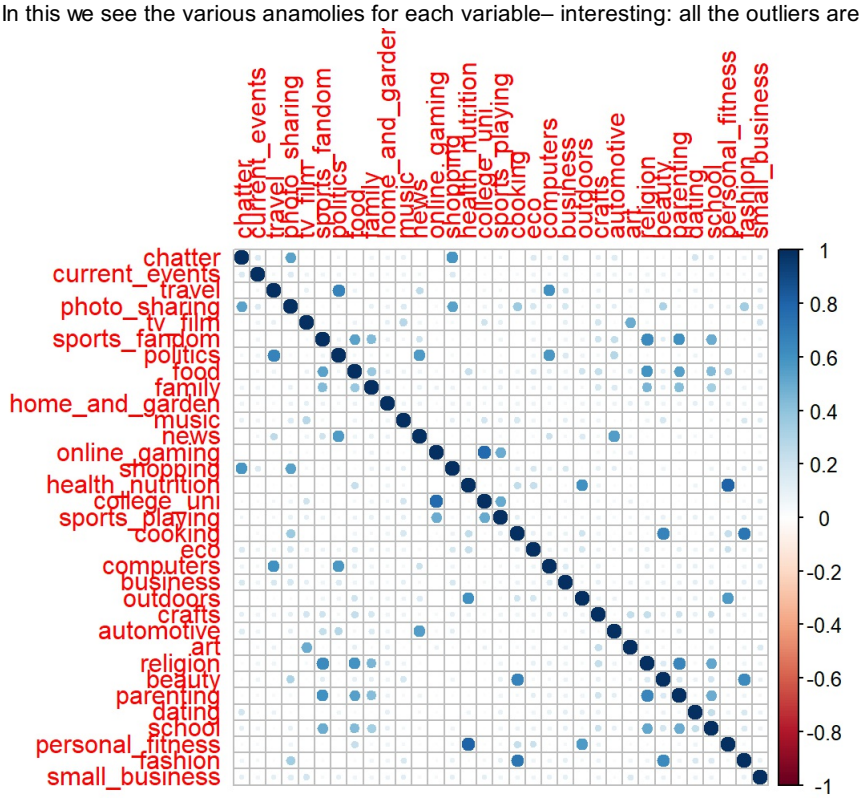
```
##          X           chatter       current_events      travel
##  123pxkyqj:   1   Min.   : 0.000   Min.   :0.000   Min.   : 0.000
##  12grikctu:   1   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 0.000
##  12klxic7j:   1   Median : 3.000   Median :1.000   Median : 1.000
##  12t4msroj:   1   Mean   : 4.399   Mean   :1.526   Mean   : 1.585
##  12yam59l3:   1   3rd Qu.: 6.000   3rd Qu.:2.000   3rd Qu.: 2.000
##  132y8f6aj:   1   Max.   :26.000   Max.   :8.000   Max.   :26.000
##  (Other)  :7876
##  photo_sharing    uncategorized      tv_film      sports_fandom
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 1.000   1st Qu.:0.000   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 2.000   Median :1.000   Median : 1.00   Median : 1.000
##  Mean   : 2.697   Mean   :0.813   Mean   : 1.07   Mean   : 1.594
##  3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 1.00   3rd Qu.: 2.000
##  Max.   :21.000   Max.   :9.000   Max.   :17.00   Max.   :20.000
##
##     politics           food            family        home_and_garden
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 1.000   Median : 1.000   Median : 1.0000   Median :0.0000
##  Mean   : 1.789   Mean   : 1.397   Mean   : 0.8639   Mean   :0.5207
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000
##  Max.   :37.000   Max.   :16.000   Max.   :10.0000   Max.   :5.0000
##
##      music              news        online_gaming       shopping
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
```

```
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.0000   Median : 0.000   Median : 0.000   Median : 1.000
##  Mean   : 0.6793   Mean   : 1.206   Mean   : 1.209   Mean   : 1.389
##  3rd Qu.: 1.0000   3rd Qu.: 1.000   3rd Qu.: 1.000   3rd Qu.: 2.000
##  Max.   :13.0000   Max.   :20.000   Max.   :27.000   Max.   :12.000
##
##  health_nutrition  college_uni     sports_playing      cooking
##  Min.   : 0.000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median : 1.000   Median : 1.000   Median :0.0000   Median : 1.000
##  Mean   : 2.567   Mean   : 1.549   Mean   :0.6392   Mean   : 1.998
##  3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.: 2.000
##  Max.   :41.000   Max.   :30.000   Max.   :8.0000   Max.   :33.000
##
##       eco           computers          business         outdoors
##  Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.0000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   :0.5123   Mean   : 0.6491   Mean   :0.4232   Mean   : 0.7827
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.: 1.0000
##  Max.   :6.0000   Max.   :16.0000   Max.   :6.0000   Max.   :12.0000
##
##      crafts         automotive           art              religion
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.000
##  Median :0.0000   Median : 0.0000   Median : 0.0000   Median : 0.000
##  Mean   :0.5159   Mean   : 0.8299   Mean   : 0.7248   Mean   : 1.095
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.000
##  Max.   :7.0000   Max.   :13.0000   Max.   :18.0000   Max.   :20.000
##
##      beauty          parenting           dating            school
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.7052   Mean   : 0.9213   Mean   : 0.7109   Mean   : 0.7677
##  3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000
##  Max.   :14.0000   Max.   :14.0000   Max.   :24.0000   Max.   :11.0000
##
##  personal_fitness    fashion        small_business        spam
##  Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.000   Median : 0.0000   Median :0.0000   Median :0.00000
##  Mean   : 1.462   Mean   : 0.9966   Mean   :0.3363   Mean   :0.00647
##  3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :19.000   Max.   :18.0000   Max.   :6.0000   Max.   :2.00000
##
##      adult
##  Min.   : 0.0000
##  1st Qu.: 0.0000
##  Median : 0.0000
##  Mean   : 0.4033
##  3rd Qu.: 0.0000
##  Max.   :26.0000
##
```

Comments: chatter and uncategorized are similar so get rid of one of them- I got rid of uncategorized.I also go rid of spam and adult and ID for easier analysis.



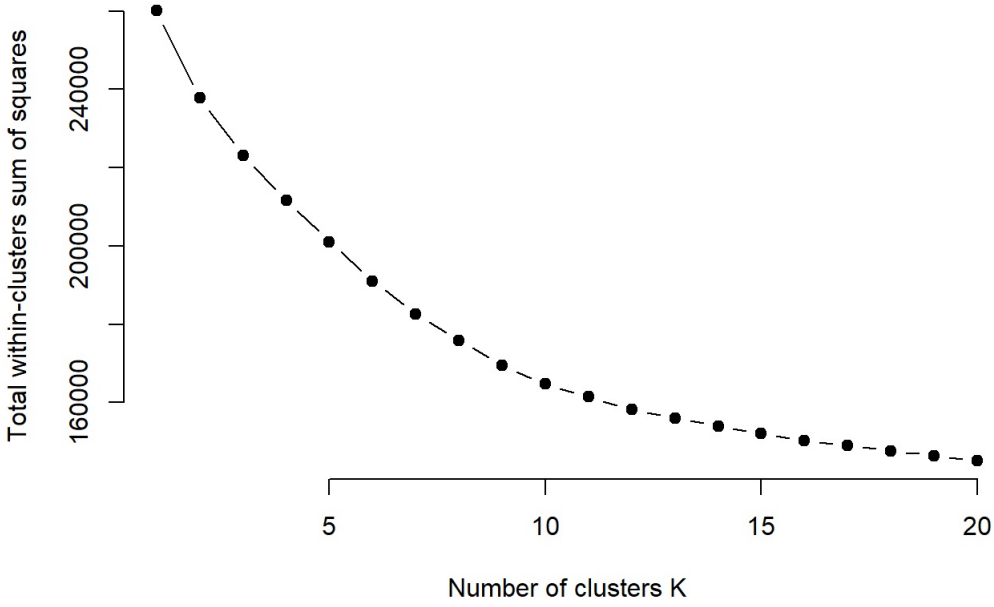
Comments: In this we see the various anamolies for each variable– interesting: all the outliers are only positive ones



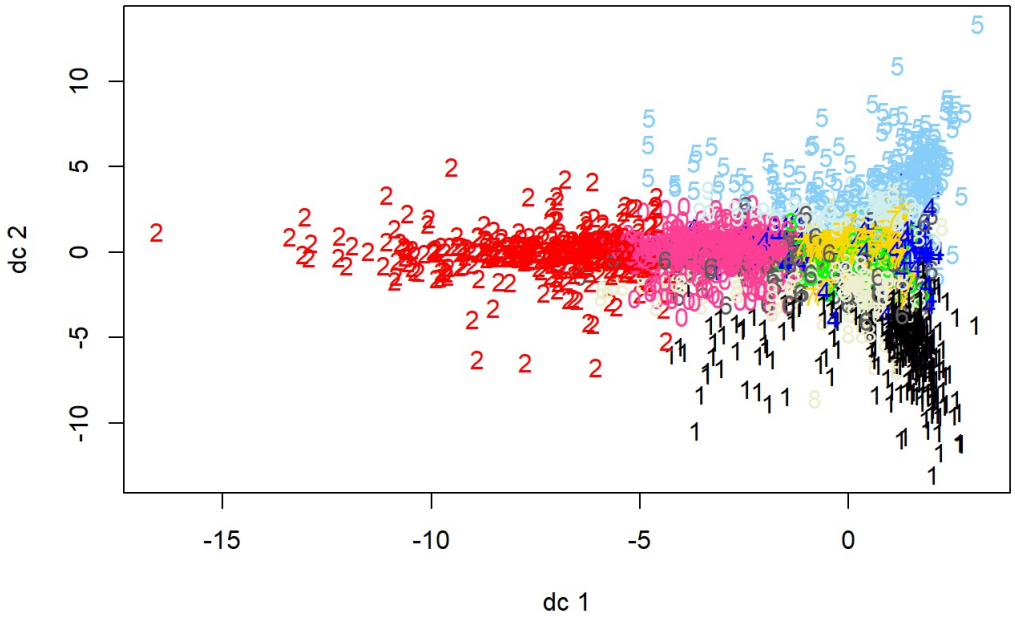
Comments: we see some high correlations for: - politics and travel - college/uni and online gaming - personal fitness and health_nutrition - fashion and cooking - beauty and fashion - parenting and religion - religion and politcs - news and politics - shopping chatter

K-Means Approach

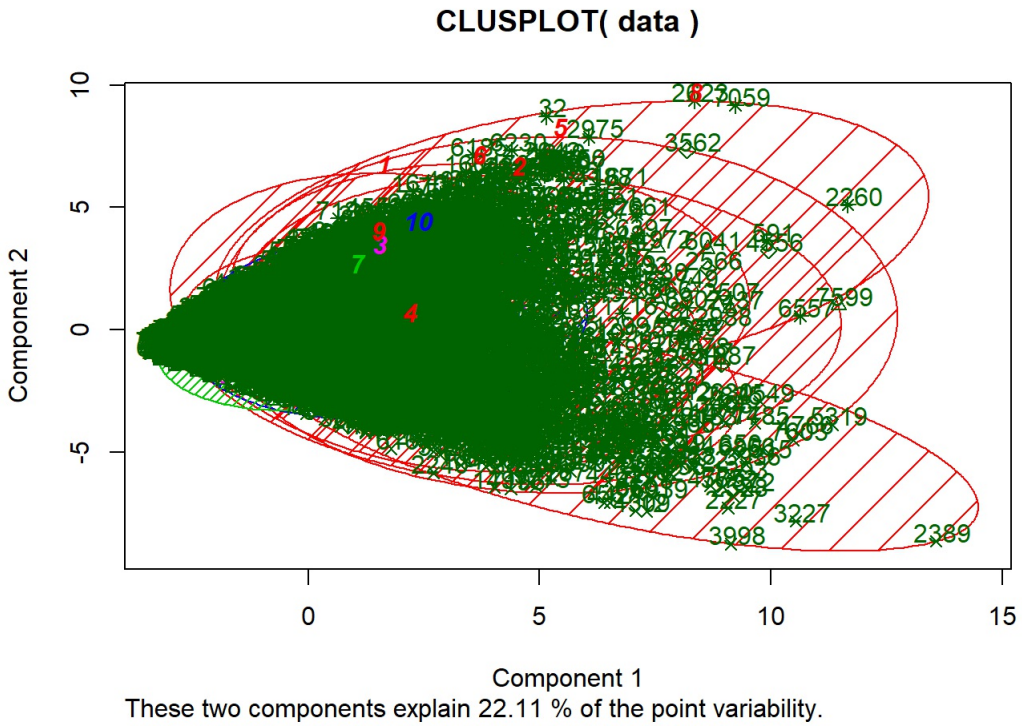
```
##  [1] 260073.0 237751.5 223180.2 211701.0 201103.6 191040.8 182562.4
##  [8] 175932.7 169484.3 164759.1 161424.9 158250.9 156005.2 153949.1
## [15] 152118.0 150273.2 149104.0 147698.5 146458.0 145104.2
```



Comments: Started the k-means algorithm– plot shows the minimization of the sum of squares for the various number of clusters up to the max of 20. Looks like 10 is the best cluster size.



Comments: This graph plots the various clusters created and is displayed by the cluster number they are in. It looks relatively split and grouped well.

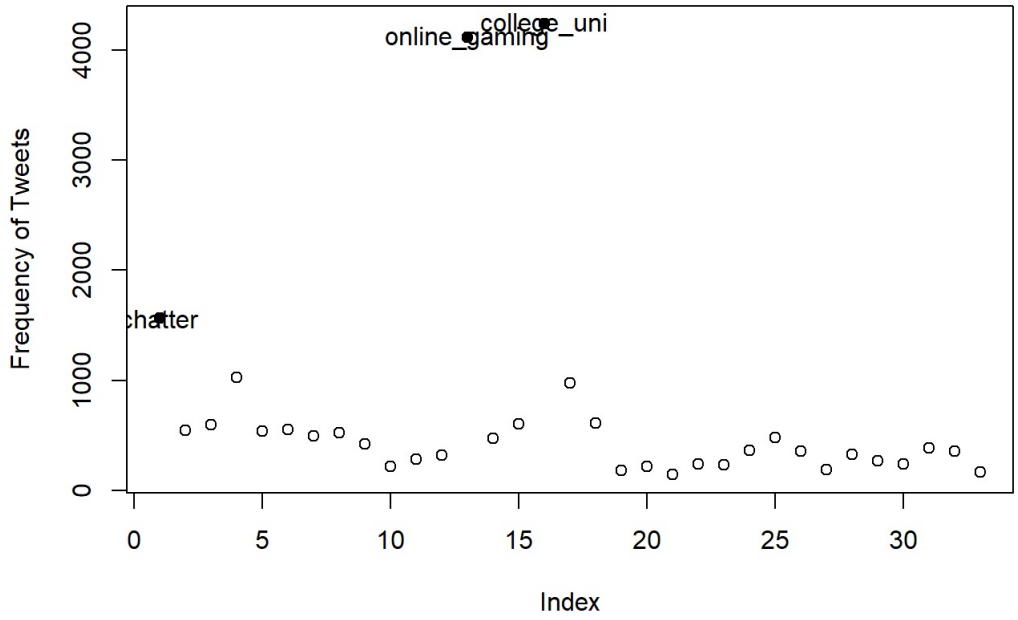


Comments: Gross and messy graph- probably not usable or interpretable

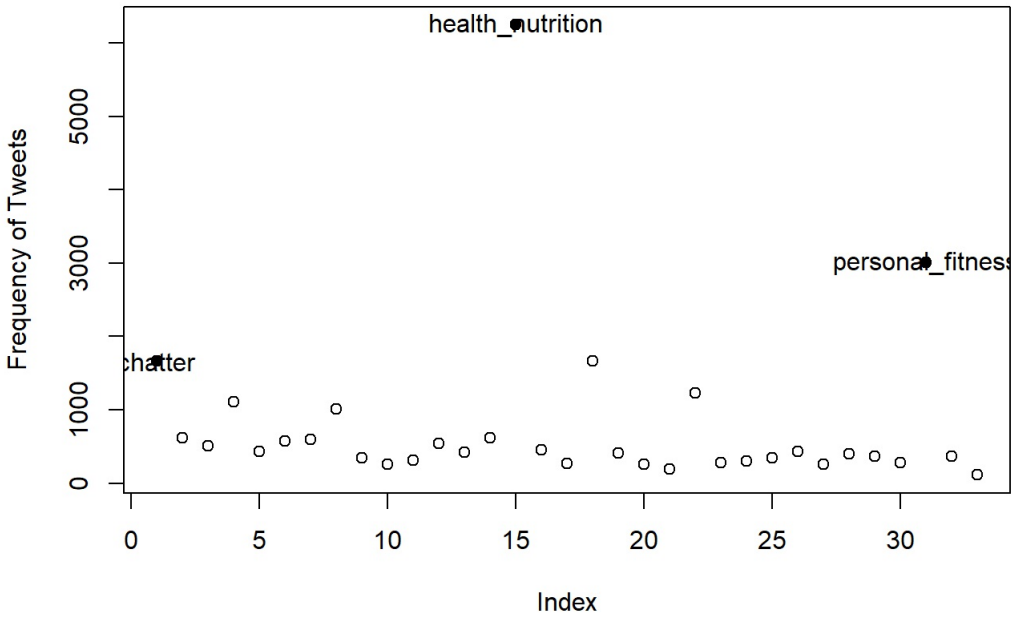
```
## List of 9
##  $ cluster     : int [1:7882] 2 7 3 7 3 3 5 8 2 4 ...
##  $ centers     : num [1:10, 1:33] 4.05 4.6 7.2 3.51 4.15 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:10] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:33] "chatter" "current_events" "travel" "photo_sharing" ...
##  $ totss       : num 1076310
##  $ withinss    : num [1:10] 44367 44314 64701 51718 42045 ...
##  $ tot.withinss: num 535087
##  $ betweenss   : num 541223
##  $ size        : int [1:10] 386 363 1269 574 344 489 2724 473 407 853
##  $ iter        : int 6
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

Comments: Here is the summary of the clusters created. We can see the sizes and other attributes.

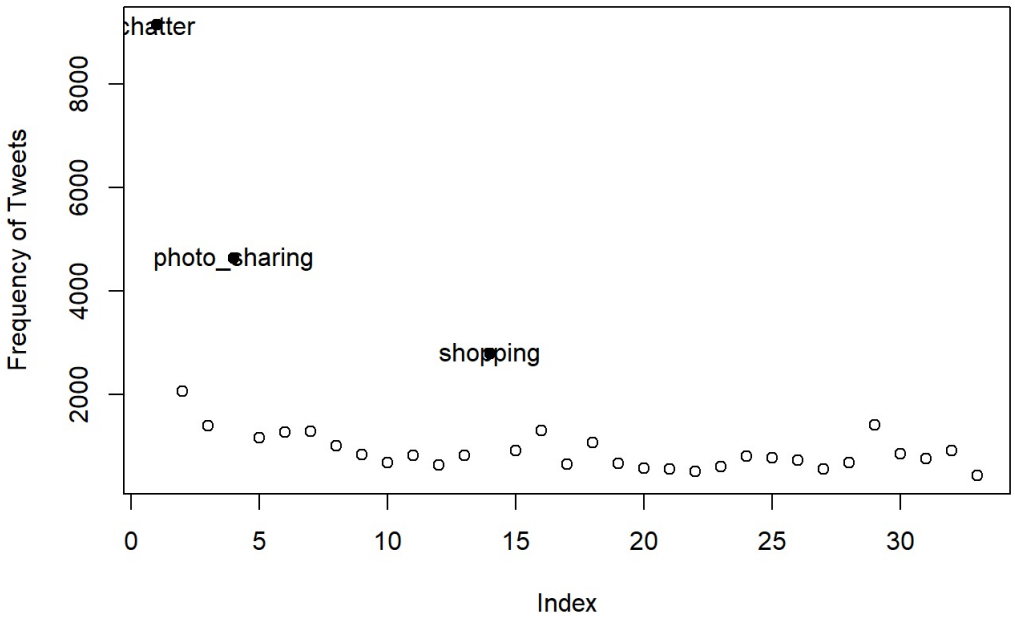
Top 3 Categories for Cluster 1



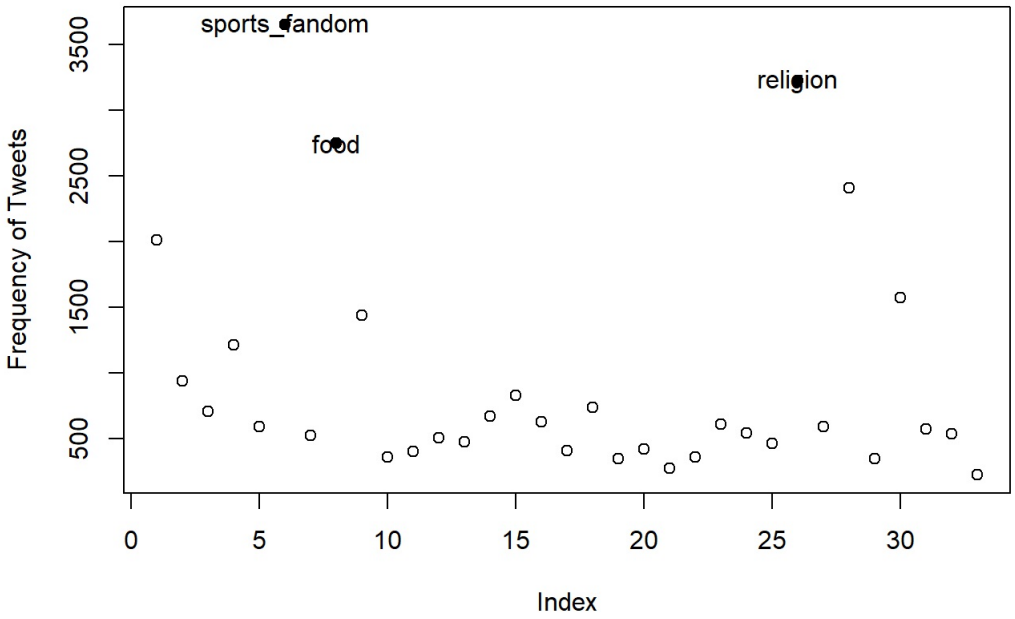
Top 3 Categories for Cluster 2



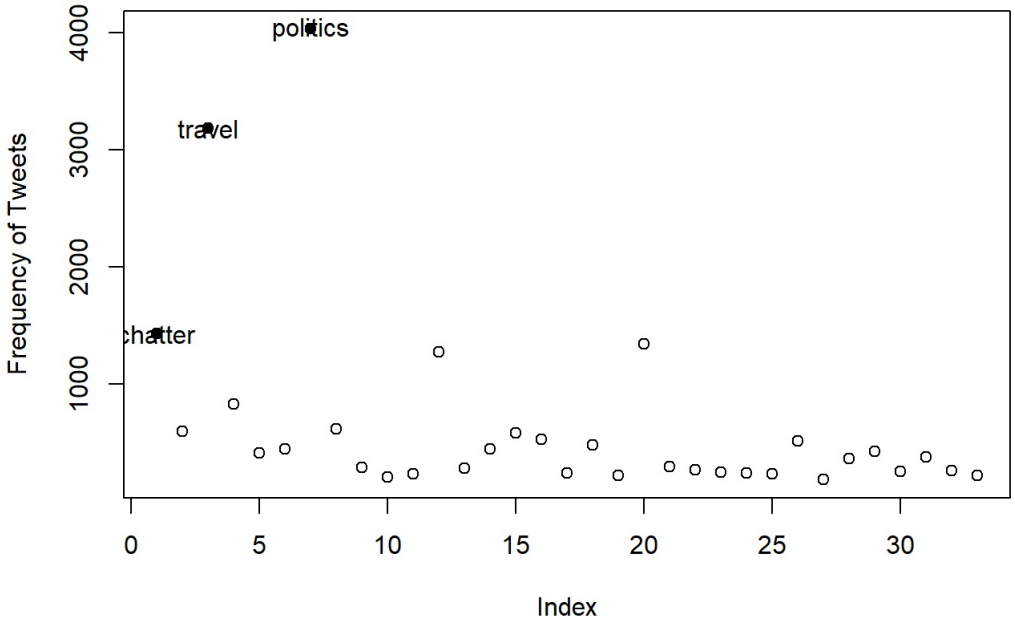
Top 3 Categories for Cluster 3



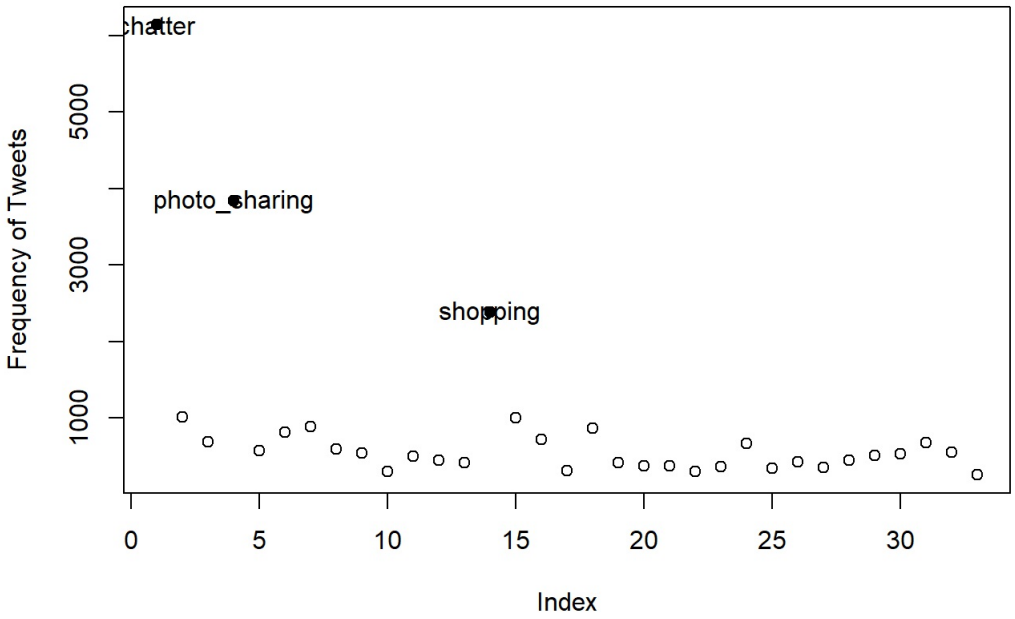
Top 3 Categories for Cluster 4



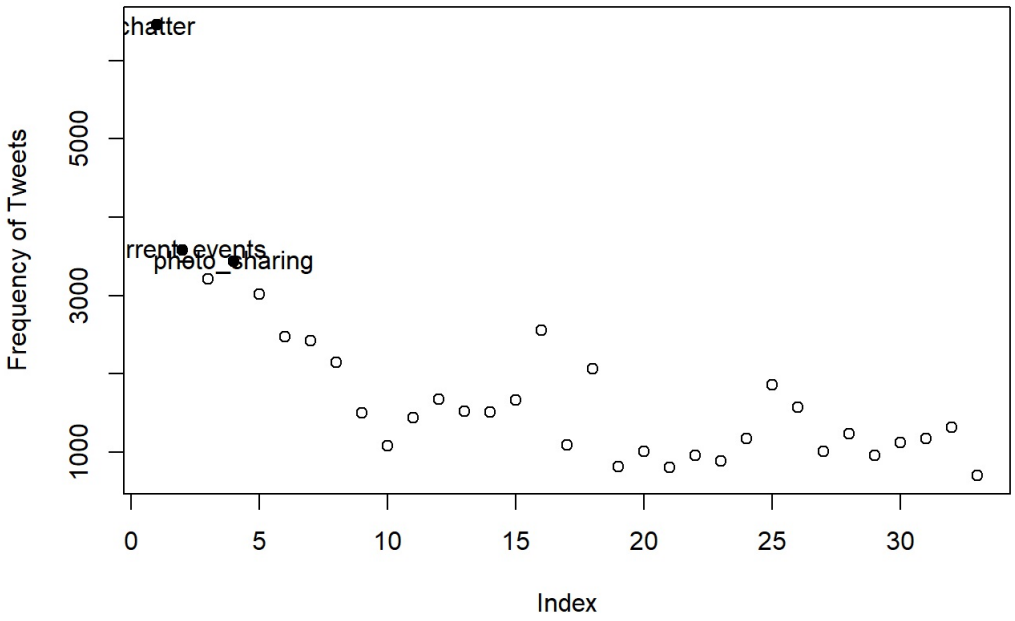
Top 3 Categories for Cluster 5



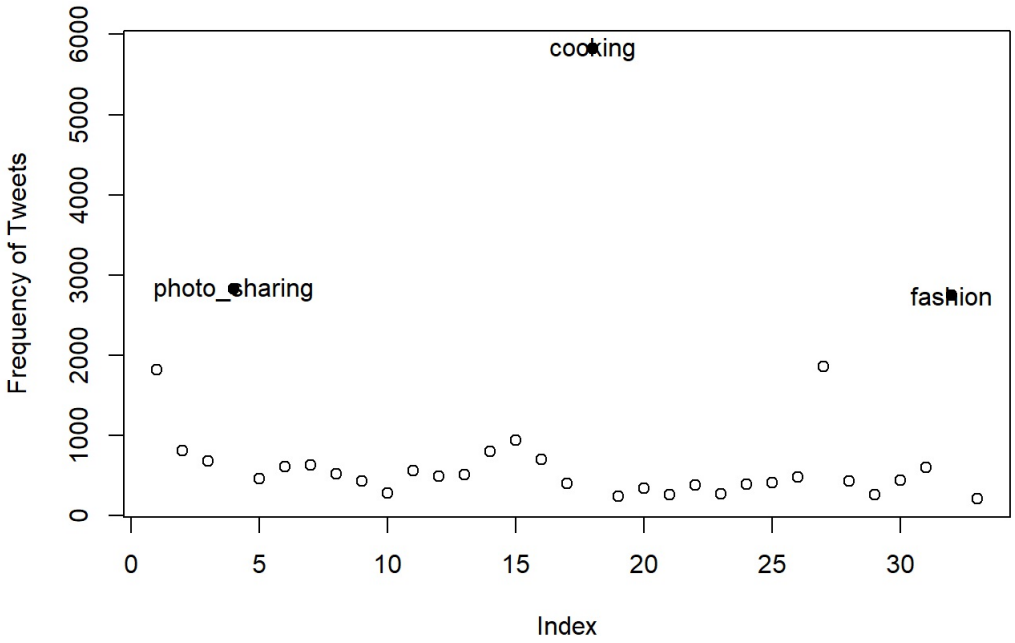
Top 3 Categories for Cluster 6



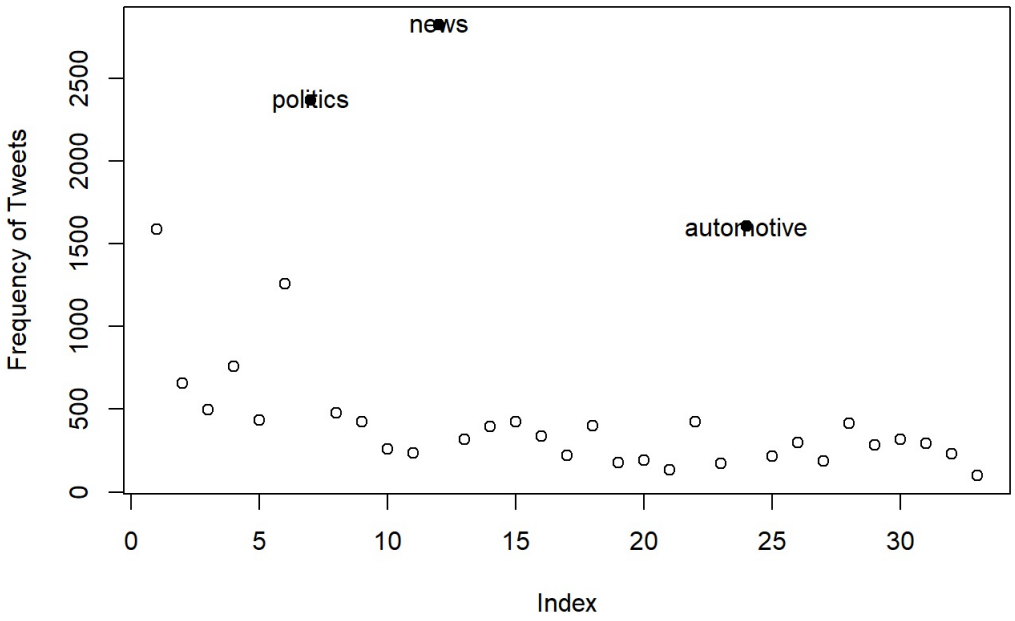
Top 3 Categories for Cluster 7



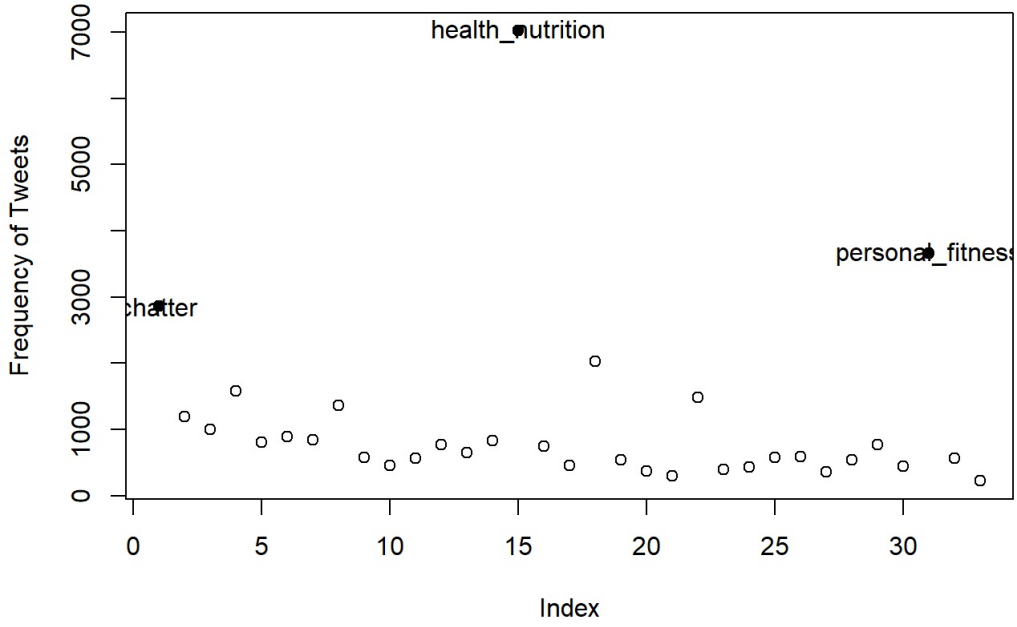
Top 3 Categories for Cluster 8



Top 3 Categories for Cluster 9



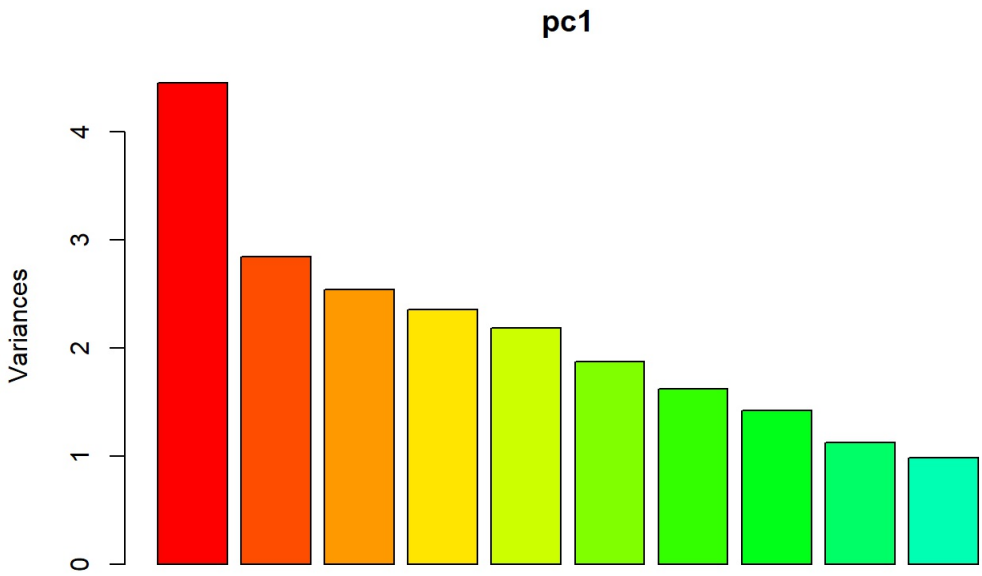
Top 3 Categories for Cluster 10



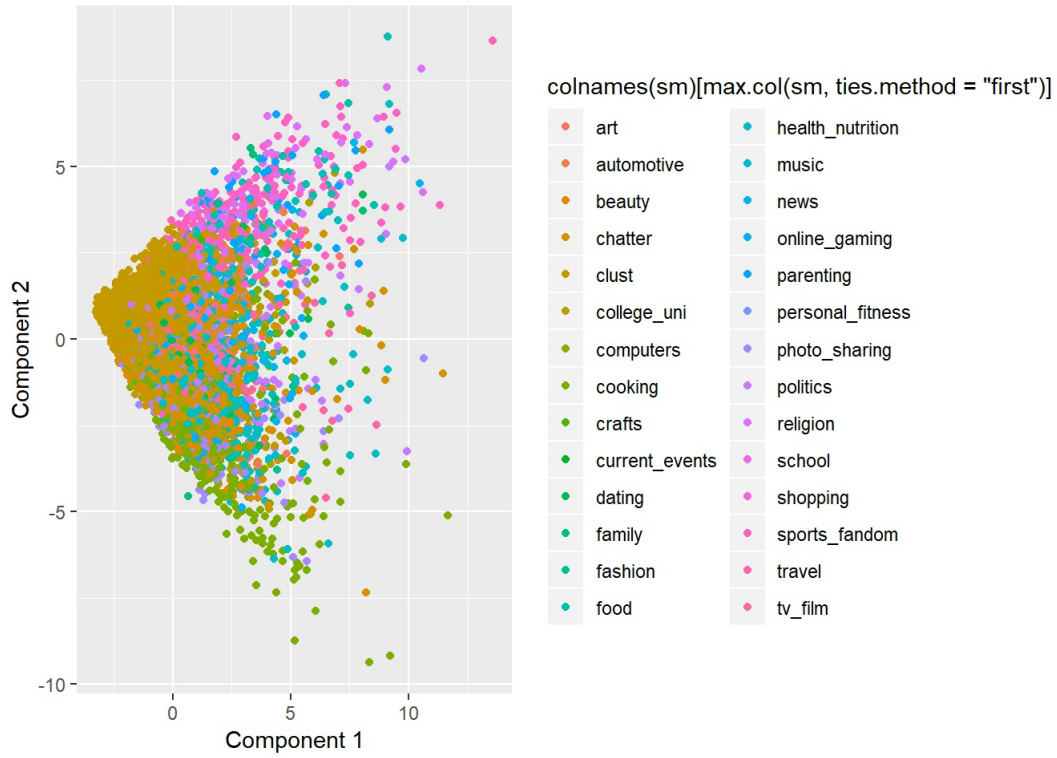
Other methods tried:

PCA

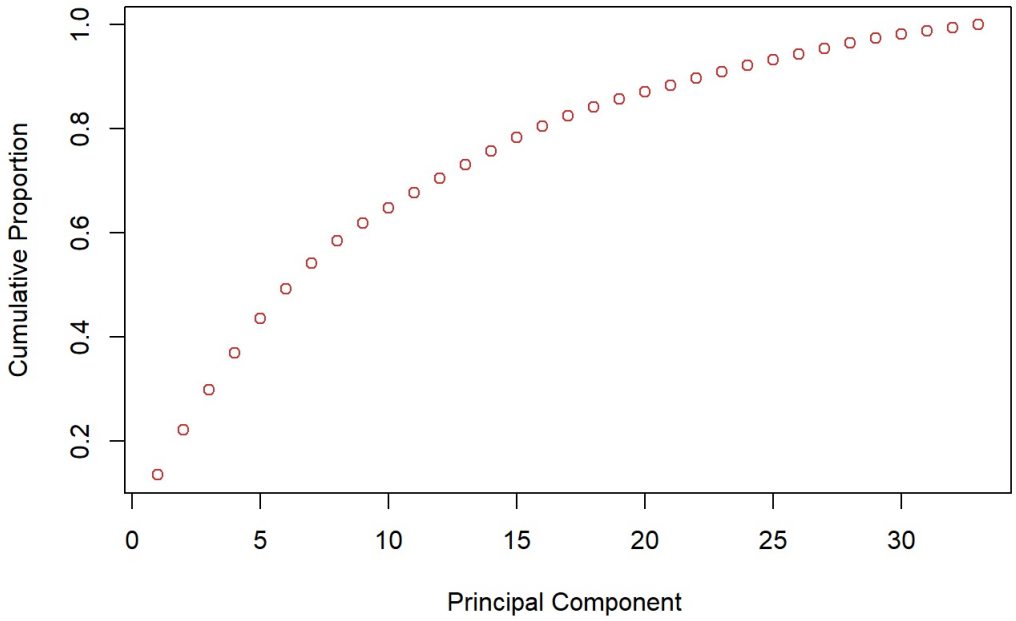
```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.111 1.68580 1.59326 1.53403 1.47848 1.36826
## Proportion of Variance 0.135 0.08612 0.07692 0.07131 0.06624 0.05673
## Cumulative Proportion  0.135 0.22112 0.29804 0.36935 0.43559 0.49233
##                           PC7     PC8     PC9   PC10   PC11    PC12
## Standard deviation     1.27409 1.19215 1.06092 0.9934 0.9681 0.96181
## Proportion of Variance 0.04919 0.04307 0.03411 0.0299 0.0284 0.02803
## Cumulative Proportion  0.54152 0.58458 0.61869 0.6486 0.6770 0.70503
##                           PC13    PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.93984 0.92317 0.91605 0.85455 0.80890 0.75383
## Proportion of Variance 0.02677 0.02583 0.02543 0.02213 0.01983 0.01722
## Cumulative Proportion  0.73180 0.75762 0.78305 0.80518 0.82501 0.84223
##                           PC19    PC20    PC21    PC22   PC23   PC24
## Standard deviation     0.69805 0.68770 0.65471 0.65041 0.6398 0.6372
## Proportion of Variance 0.01477 0.01433 0.01299 0.01282 0.0124 0.0123
## Cumulative Proportion  0.85699 0.87133 0.88432 0.89713 0.9095 0.9218
##                           PC25    PC26    PC27    PC28    PC29    PC30
## Standard deviation     0.61729 0.60206 0.59482 0.58767 0.55066 0.48580
## Proportion of Variance 0.01155 0.01098 0.01072 0.01047 0.00919 0.00715
## Cumulative Proportion  0.93339 0.94437 0.95509 0.96556 0.97475 0.98190
##                           PC31    PC32   PC33
## Standard deviation     0.47616 0.43858 0.4222
## Proportion of Variance 0.00687 0.00583 0.0054
## Cumulative Proportion  0.98877 0.99460 1.0000
```



Comments: Shows the variance of PC1



Comments: Here we see the PC1 vs PC2 plot. The points are colored by the corresponding category it is in. We can see that green and orange points are more towards the left negative and purple and blue more towards the right positive.

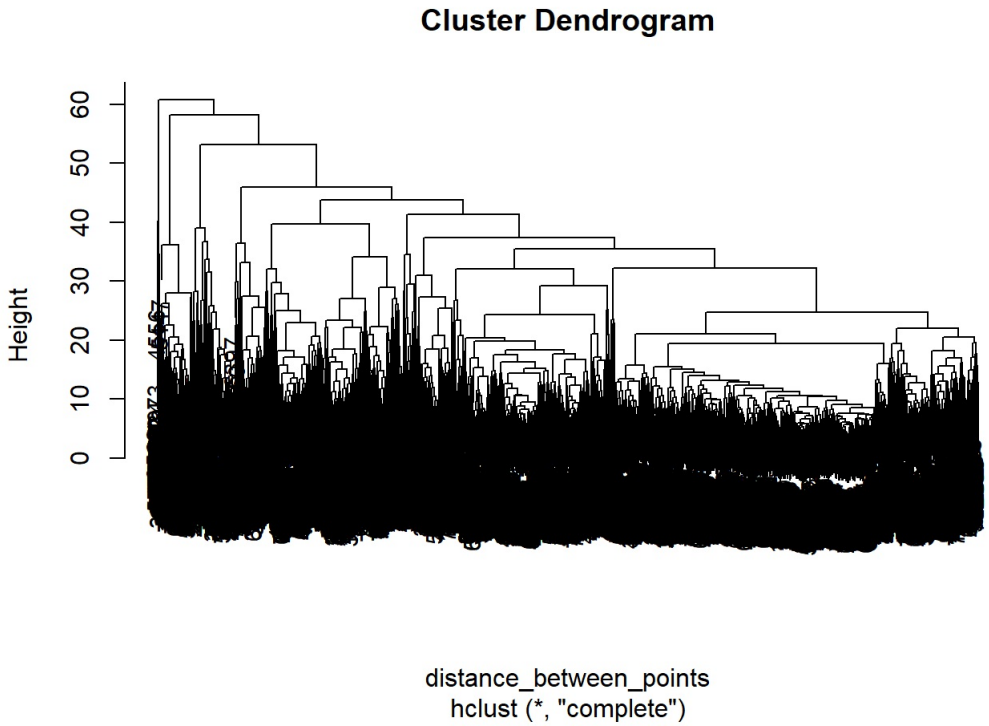


Comments: This graph shhows the cumulative proportion against the PC– it looks like between 15 and 20 would be the ideal number of components.

Hierarchical Clustering

```
##    1    2    3    4    5    6    7    8    9   10
##  299 5409  785  295  540  107  340   68   31    8
```

Comments: The summary shows the sizes of the different clusters– I don't like it that much due to how different the sizes are.



Comments: plot of the dendrogram

Report

In my analysis, I found k-means to be the most appropriate and usable given my knowledge. I split the data into 10 clusters– in the voice of the business, these would be the 10 market segments. I only plotted the top 3 max frequencies of tweets for each category, but ideally we would do a little more cleaning and retrieve the medians of each category for frequency. This would then show us how each category would be different. But using the analysis I found, we would market towards the higher frequency categories based on the clusters. I found many had chatter as a top contender so it seems that it would be helpful to create more categories to break down chatter even more.