Problem 1

Daniel Peng, Valerie Roth, Rachel Wang

August 8, 2016

Probability Practice

Part A

First, I compute the probability that a user said "yes" given that they were a Truthful Clicker.

P(Y|RC) + P(Y|TC) = .65

.15 + P(Y|TC) = .65

P(Y|TC) = .5

Then, I compute the probability that a user said "no" given that they were a Truthful Clicker.

P(N|RC) + P(N|TC) = .35

.15 + P(N|TC) = .35

P(N|TC) = .2

I can check that I did not make a mistake by applying the law of total probability. The sum of all possible outcomes should be one.

Total Probability:

P(Y|TC) = .5

P(N|TC) = .2

P(Y|RC) = .15

P(N|RC) = .15

Total: 1

To calculate the fraction of Truthful Clickers who answered "yes," I divide the probability that a Truthful Clicker responds "yes" by the sum of the probabilities of every response a Truthful Clicker could give (this is "yes" and "no").

P(Y|TC)/(P(Y|TC) + P(N|TC)) = .5/(.5+.2) = 5/7

I find that the fraction of people who are Truthful Clickers who answered "yes" is 5/7.

Part B

From the problem description, we know that:

P(tests positive|has disease) = 0.993

P(tests positive|doesn't have disease) = 0.0001

P(has disease) = 0.000025

P(doesn't have disease) = 0.999975

P(tests positive) = P(tests positive|has disease)*P(has disease) + P(tests positive|doesn't have disease)*P(doesn't have disease) = 0.993*0.000025 + 0.0001*0.999975 = 0.0001248225

We want to find P(has disease|tests positive). To do this we can use Bayes' Rule.

With Bayes' Rule, this is equivalent to (P(tests positive|has disease)*P(has disease))/P(tests positive).

(0.993 * 0.000025)/0.0001248225 = 0.1988824130265

If someone tests positive, there is about a 19.9% chance that they have the disease. In light of this calculation, having a universal testing policy for this disease does not make sense. If someone tests positive, it is still unlikely that they actually have the disease.

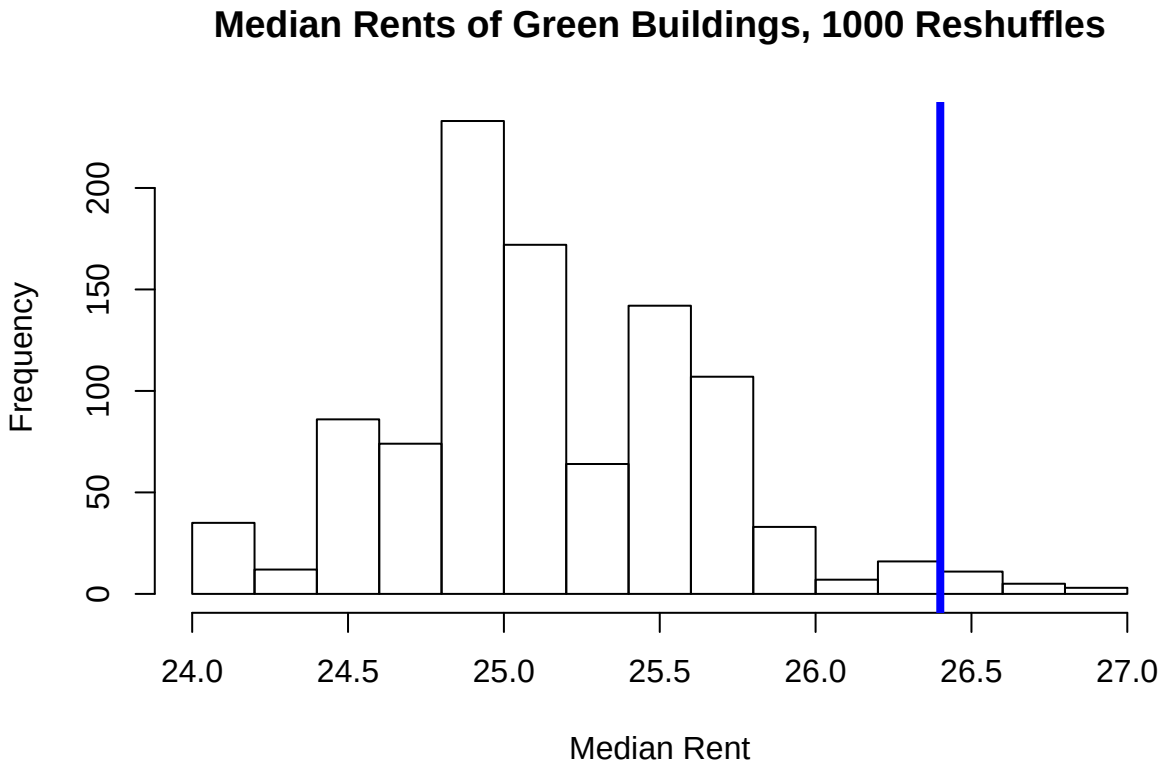
Problem 2

The guru suggested that the developer should seek green certification because green buildings supposedly result in higher rent. This would mean that a green certification would pay for itself within nine years assuming the lease rate was 90$+.

There are three flaws with this analysis:

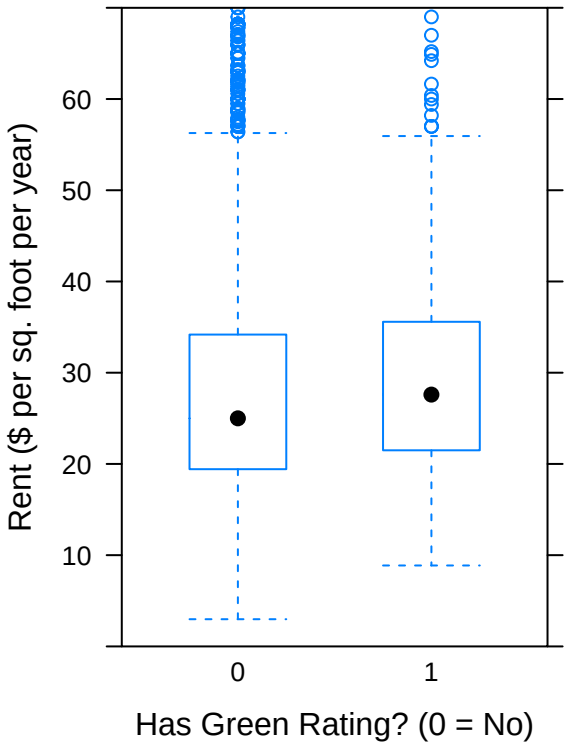
1. Using median price as the sole basis on deciding whether to "go green"
2. Failing to analyze confounding variables that may explain why green buildings are pricier
3. Using the difference in green and non-green median rents to determine profit without subsetting the data

Flaw 1: The guru points out that the difference in median rents of green vs. non-green buildings is $2.60 per square foot, which he uses to justify an increase in value by going green. We will run a permutation tests to determine whether this increase in median value is significant. This involves shuffling the green status and then creating a histogram of the median rents of only green buildings. The histogram below shows the result from 1000 resamples. This shows us that $26.40, the guru's calculation of median rent of green buildings, is statistically significant at the 95% level.



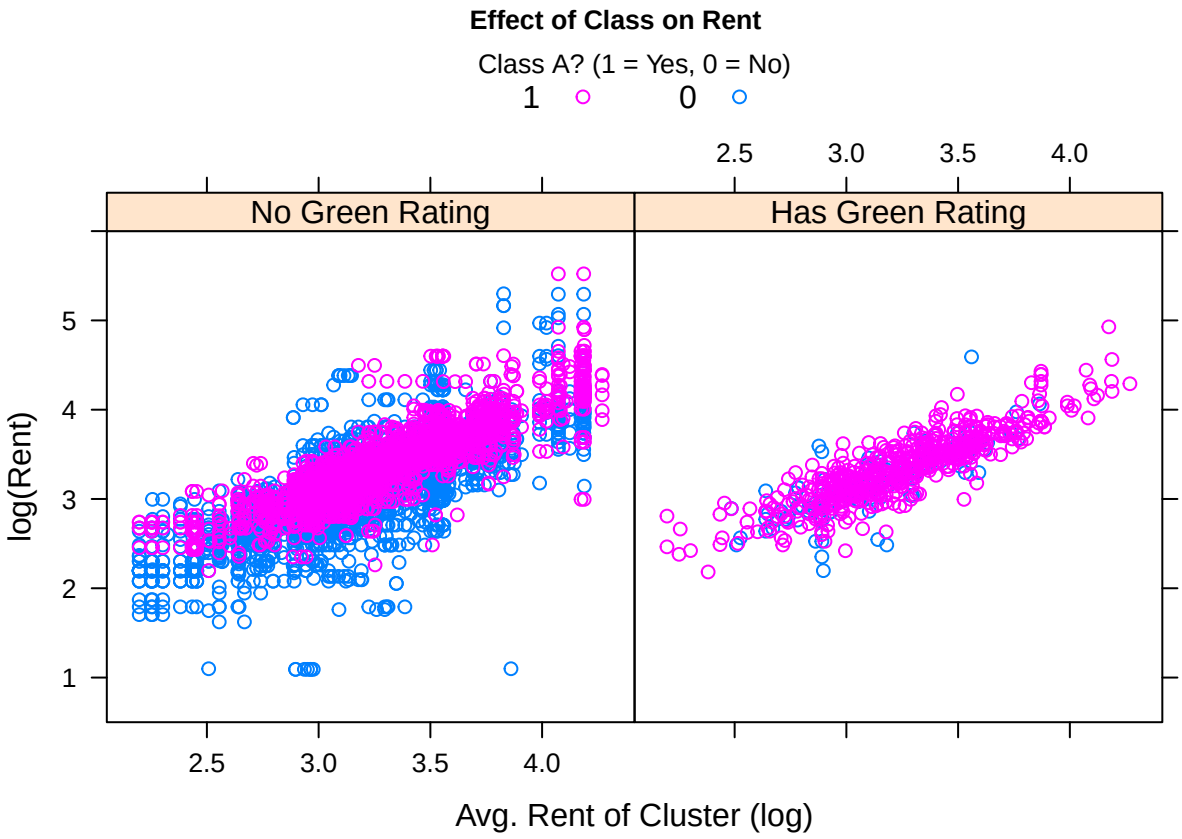
That said, a median, even if it is statistically significant, cannot prove the value of getting a green certification. There is still a 50% chance that the "actual" median rent that the real estate developer would have to set for her building would be lower depending on other variables. Maybe no company in Austin would pay $27.60 per square foot per year, even the building were "green." Or perhaps companies in Austin care more about other building features.

The box plot below illustrates that there is a significant amount of overlap between green and non-green buildings. This shows knowing whether a building is green or not does help you to predict the rent. Rather, other variables must be used to explain the variation in rent prices.



Flaw 2: There are confounding variables

Class, which measures the quality of a building compared to others within a specific market, can be used to explain why a building has higher rent than others. Below, we examine the effect of class on rent.



This plot shows the log of the average rent in a building's cluster on the x-axis and the log of the rent on the y-axis. This shows us the relative value of a building compared to other buildings in its price group. If a building has higher rent than the average of its cluster, this must be due to factors besides cluster location.

The magenta points are buildings in class A, and the blue points are buildings in class B or C. The left plot contains buildings with no green rating. Most of the magenta points appear to be higher than the blue points, meaning that being in class A tends to have a positive effect on rent.

The right plot contains only "green" buildings. They are almost all in class A. Additionally, class A buildings with green ratings have almost identical rents to class A buildings without green ratings. This tells us that being in class A likely matters much more than having a green status.

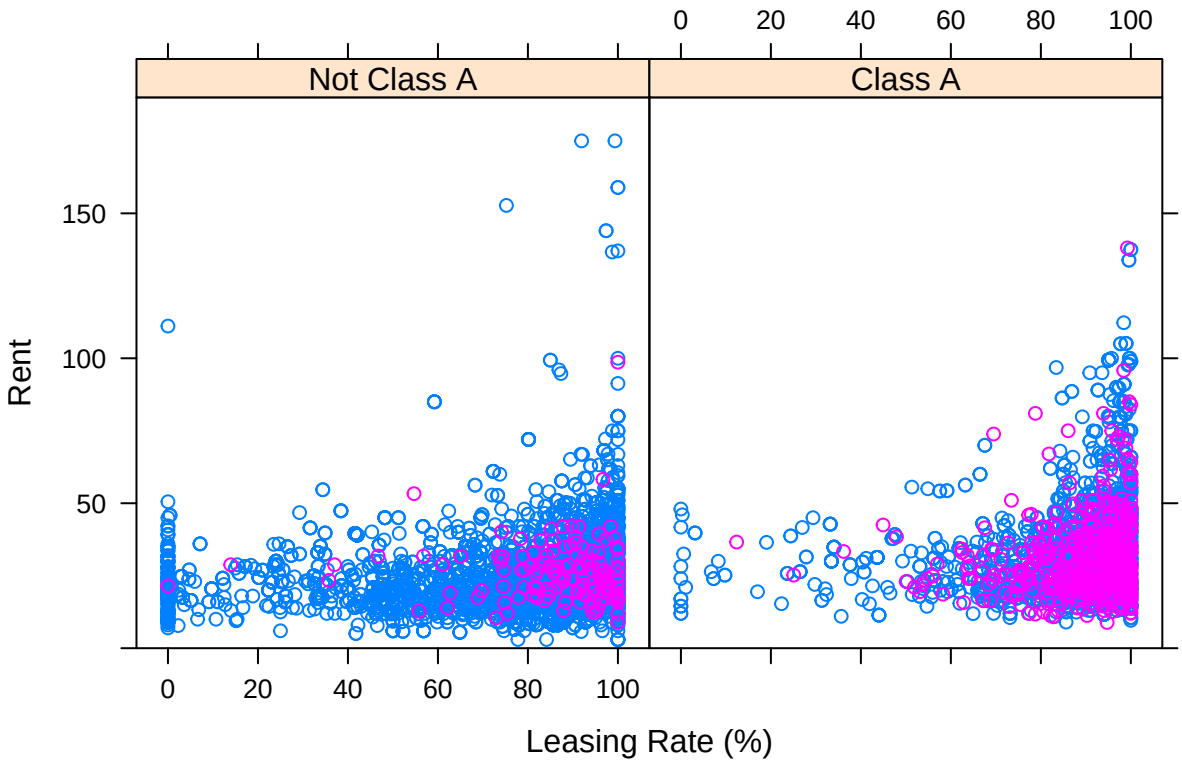
The median prices after controlling for the class of a building are as follows:

Table 1: Median Rent Prices

	Class A	~Class A	
Green	28.44	25.55	27.60
~Green	28.20	23.43	25.00
	28.20	23.50	25.16

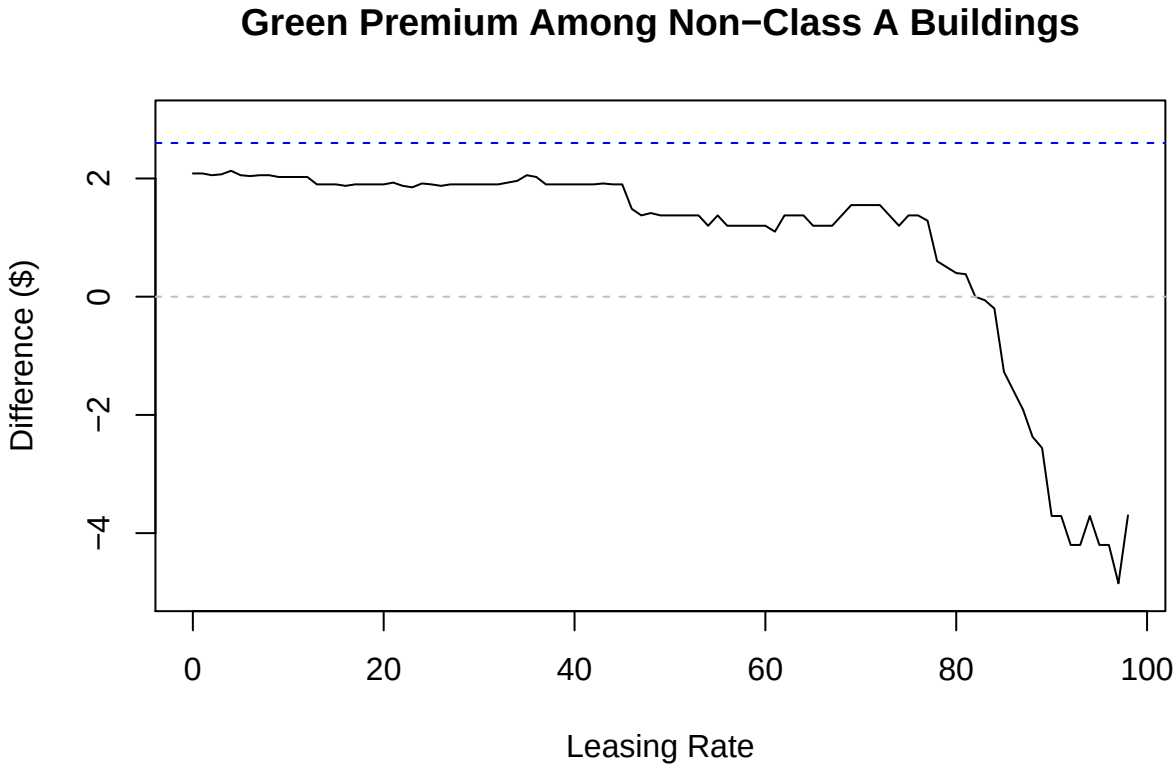
The table shows that having a green status vs. not having a green status in a class A building does not make a significant difference with regards to rent. However, it does seem to make a difference for buildings that are not in class A.

That said, we cannot use the median of non-class A buildings to make predicitons. Different variables have an effect on the median price, which can be shown by subsetting the data by 3 variables - class, green rating, and leasing rate. The distribution of leasing rates based on these is shown below.



There were many non-class A buildings with leasing rates below 10%, which the guru was smart to remove because the developer likely does not plan to rent out only a small fraction of her building.

Among class A buildings, we will now look at how much having a green building increases the difference between median prices of green and non-green buildings (the "green premium"). Recall that we saw previously that the "green premium" has little impact on the median price with class A buildings.



The guru's prediction for the "green premium" is shown in blue. Note that this value is lower at each leasing rate than the guru predicted. Surprisingly, the green status would actually cost the developer money at very high leasing rates!

One improvement that could be made to this graph would be to bucket leasing rates and report the average difference for each bucket. This is because there are many leasing rates with few values for each so this could cause the line to look quite erratic even though it follows a fairly smooth pattern. This did not occur too much, however, which shows that the "green premium" followed leasing rates very tightly.

Now consider only buildings with leasing rates of 90%+, which we can assume that the developer would be aiming for.

Table 2: Median Rent Prices, Leasing Rate > 90%

	Class A	~Class A	Green Rating Median
Green	29.00	24.36	28.485
~Green	29.67	26.92	27.275
Class Median	29.55	26.64	27.500

Here, even if the guru had ignored the class of the building and used the difference in median price between green and non-green buildings as his estimate for marginal revenue, the difference in medians among buildings with >90% leasing rates is $1.21, which is more than half of his original estimate of $2.60.

We see that using the median to compute the green premium was misleading. Rather, we needed to understand the effects of otehr variables, which may make the "green premium" seem artificially high.

If having a green status significantly helps the developer to attain class A status, it may be worth it. To determine whether she should go green or not, careful consideration of each feature she choses to add to her building must be made.

While this is a dataset that can be analyzed extensively, another point brought up by a classmate was that

there are two ways that people “go green.” The first focuses on energy efficiency and the second focuses on using green materials.

With this in mind, it is important to consider whether the net variable is true to better understand the effect of green buildings that were built for the purpose of energy efficiency. In cases where net is true, tenants pay their own for their own utility costs, and thus these buildings could be more valuable.

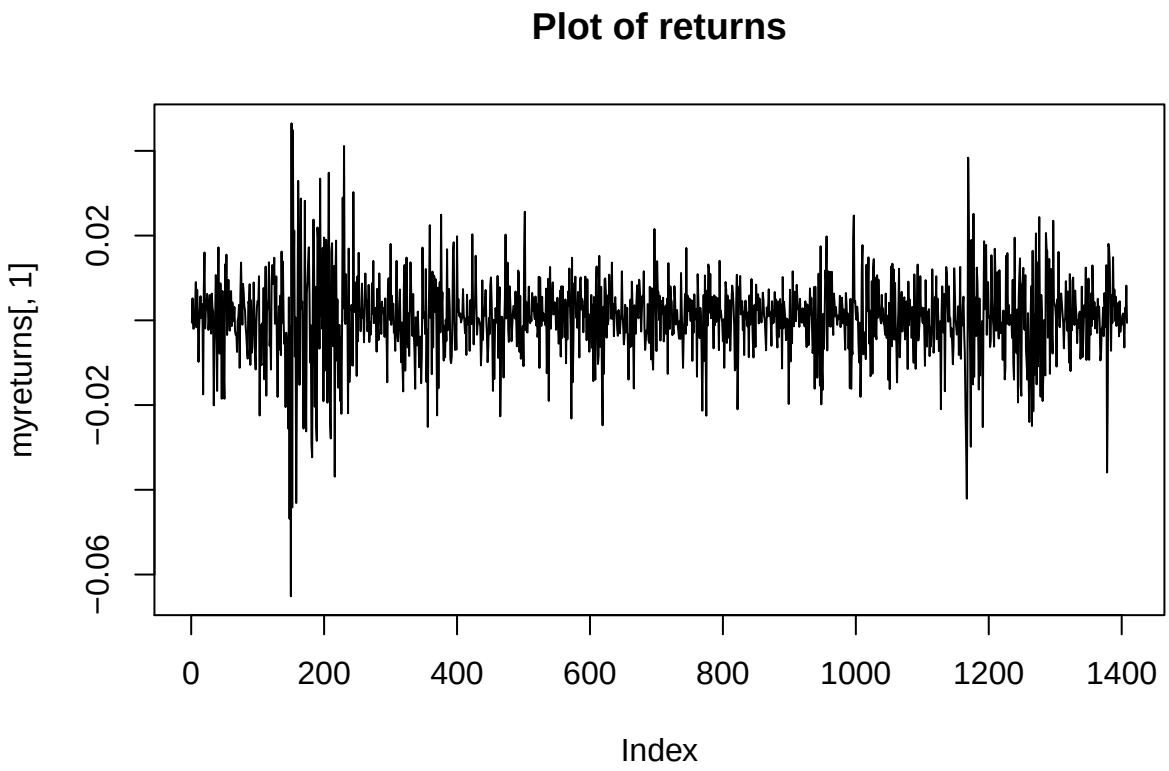
Again, there are many different ways to analyze this dataset so students’ reports may vary based on their different approaches.

Problem 3

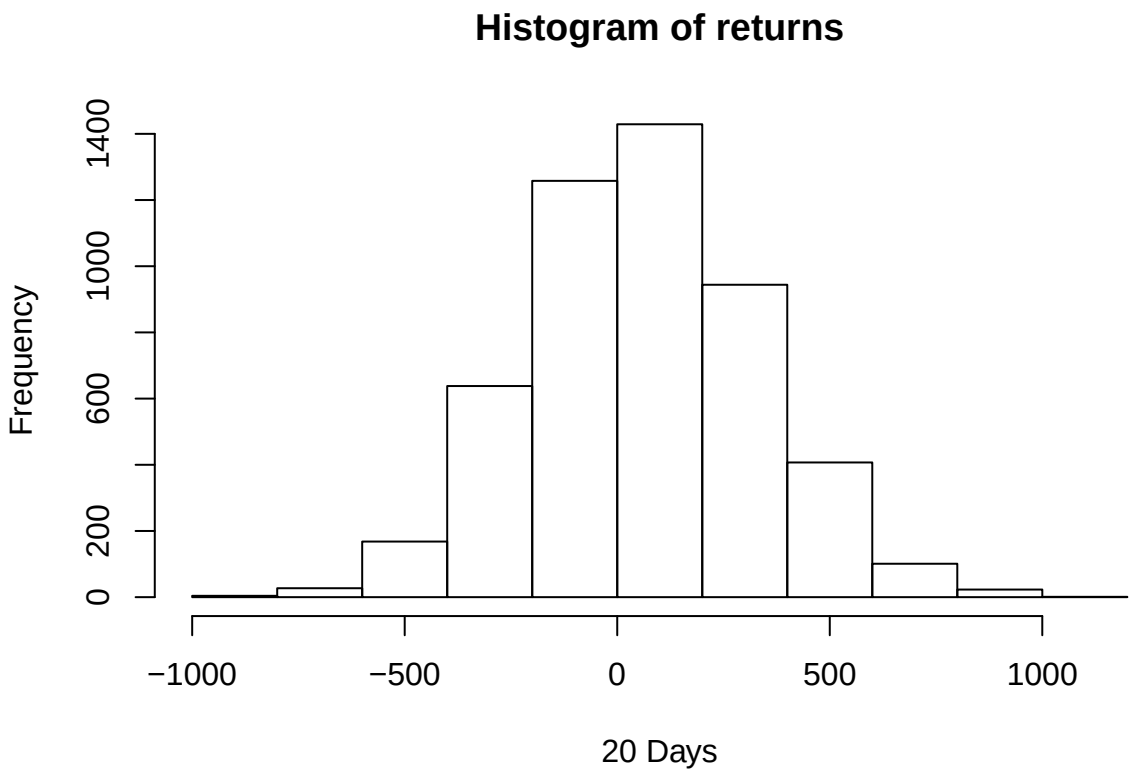
Bootstrapping

For an even split among 5 stocks: SPY, TLT, LQD, EEM, VNQ.

```
mystocks = c("SPY", "TLT", "LQD", "EEM", "VNQ")

myprices = yahooSeries(mystocks, from = "2011-01-01", to = Sys.timeDate())

YahooPricesToReturns = function(series) {
  mycols = grep('Adj.Close', colnames(series))
  closingprice = series[,mycols]
  N = nrow(closingprice)
  percentreturn = as.data.frame(closingprice[2:N,]) / as.data.frame(closingprice[1:(N-1),]) - 1
  mynames = strsplit(colnames(percentreturn), '.', fixed=TRUE)
  mynames = lapply(mynames, function(x) return(paste0(x[1], ".PctReturn")))
  colnames(percentreturn) = mynames
  as.matrix(na.omit(percentreturn))
}

myreturns = YahooPricesToReturns(myprices)

#plot daily returns % for SPY
plot(myreturns[,1], type='l', main = "Plot of returns")
```



```
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  totalwealth = 10000
  n_days = 20
  weights = c(0.2, 0.2, 0.2, 0.2, 0.2)
  holdings = weights * totalwealth
  wealthtracker = rep(0, n_days) # Set up a placeholder to track total wealth
  for(today in 1:n_days) {
    return.today = resample(myreturns, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    totalwealth = sum(holdings)
    wealthtracker[today] = totalwealth
  }
  wealthtracker
}

hist(sim1[,n_days]- 10000, main = "Histogram of returns", xlab = "20 Days")
```



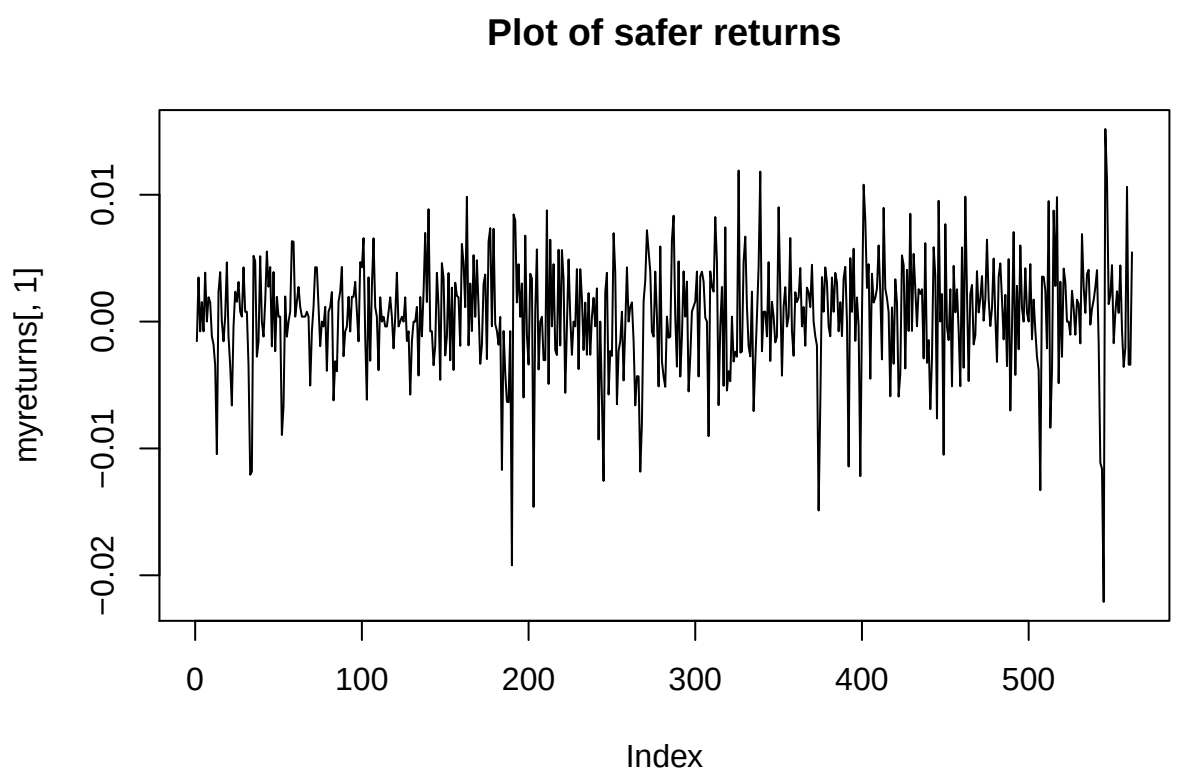

```
quantile(sim1[,n_days], 0.05) - 10000
```

```
##        5%
## -374.6939
```

For a safer choice than the even split above. I chose to only invest in top rated bonds and fixed-income ETFs. Historically the bond market has been less vulnerable to price swings or volatility than the stock market.

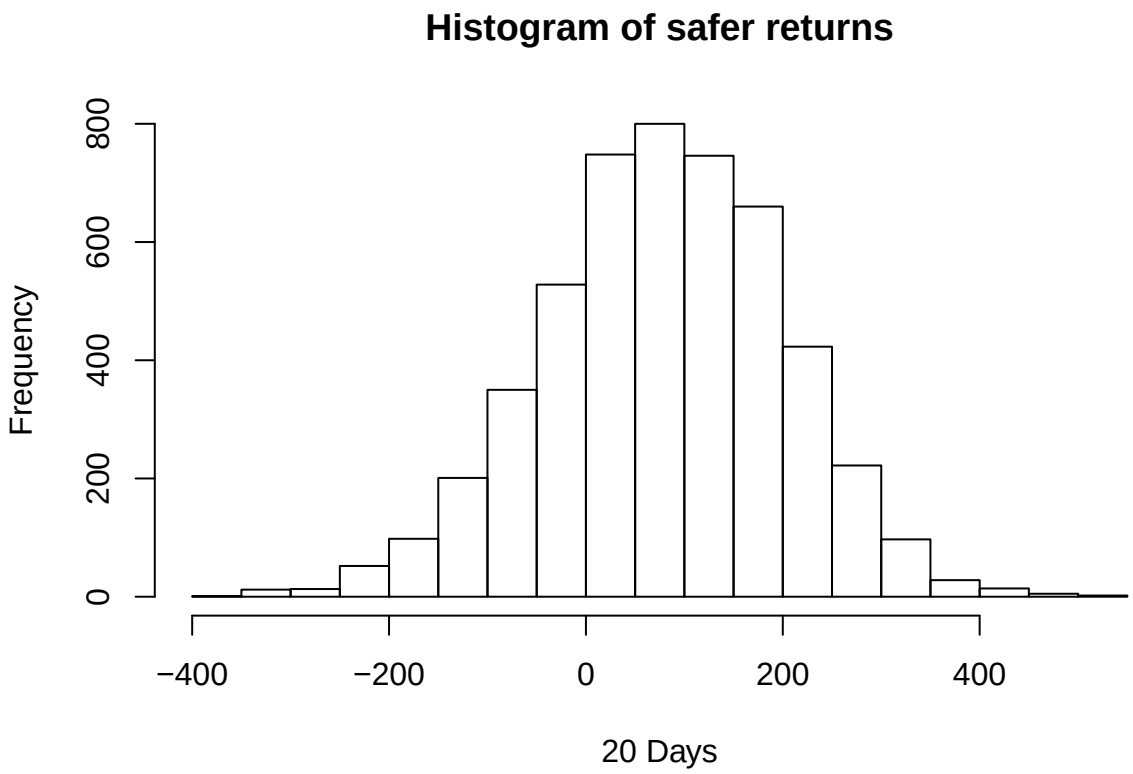
```
mystocks = c("XMPT", "BABS", "SPHD", "FMB", "PWZ")
myprices = yahooSeries(mystocks, from = "2011-01-01", to = Sys.timeDate())
```

```r
YahooPricesToReturns = function(series) {
  mycols = grep('Adj.Close', colnames(series))
  closingprice = series[,mycols]
  N = nrow(closingprice)
  percentreturn = as.data.frame(closingprice[2:N,]) / as.data.frame(closingprice[1:(N-1),]) - 1
  mynames = strsplit(colnames(percentreturn), '.', fixed=TRUE)
  mynames = lapply(mynames, function(x) return(paste0(x[1], ".PctReturn")))
  colnames(percentreturn) = mynames
  as.matrix(na.omit(percentreturn))
}

myreturns = YahooPricesToReturns(myprices)

plot(myreturns[,1], type='l', main = "Plot of safer returns")
```




```r
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  totalwealth = 10000
  n_days = 20
  weights = c(0.2, 0.2, 0.2, 0.2, 0.2)
  holdings = weights * totalwealth
  wealthtracker = rep(0, n_days) # Set up a placeholder to track total wealth
  for(today in 1:n_days) {
    return.today = resample(myreturns, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    totalwealth = sum(holdings)
    wealthtracker[today] = totalwealth
  }
  wealthtracker
```

```
}

hist(sim1[,n_days]- 10000, main = "Histogram of safer returns", xlab = "20 Days")
```



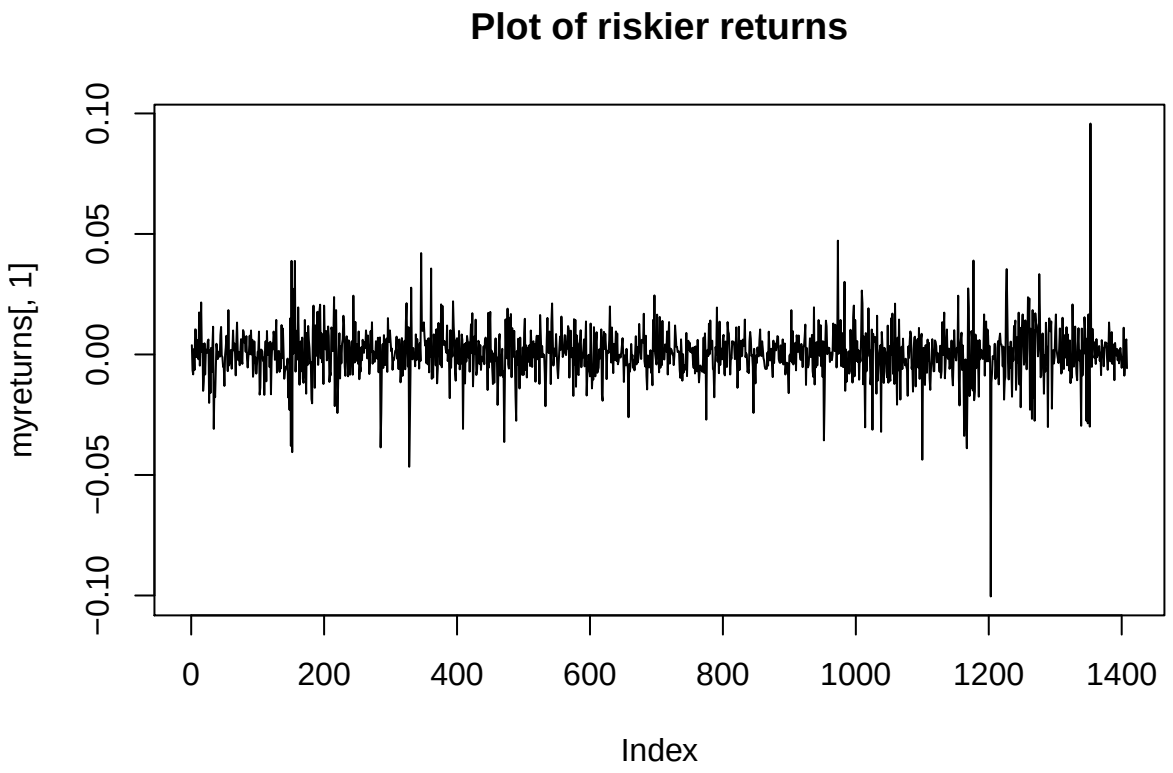

```
quantile(sim1[,n_days], 0.05) - 10000
```

```
##        5%
## -128.6049
```

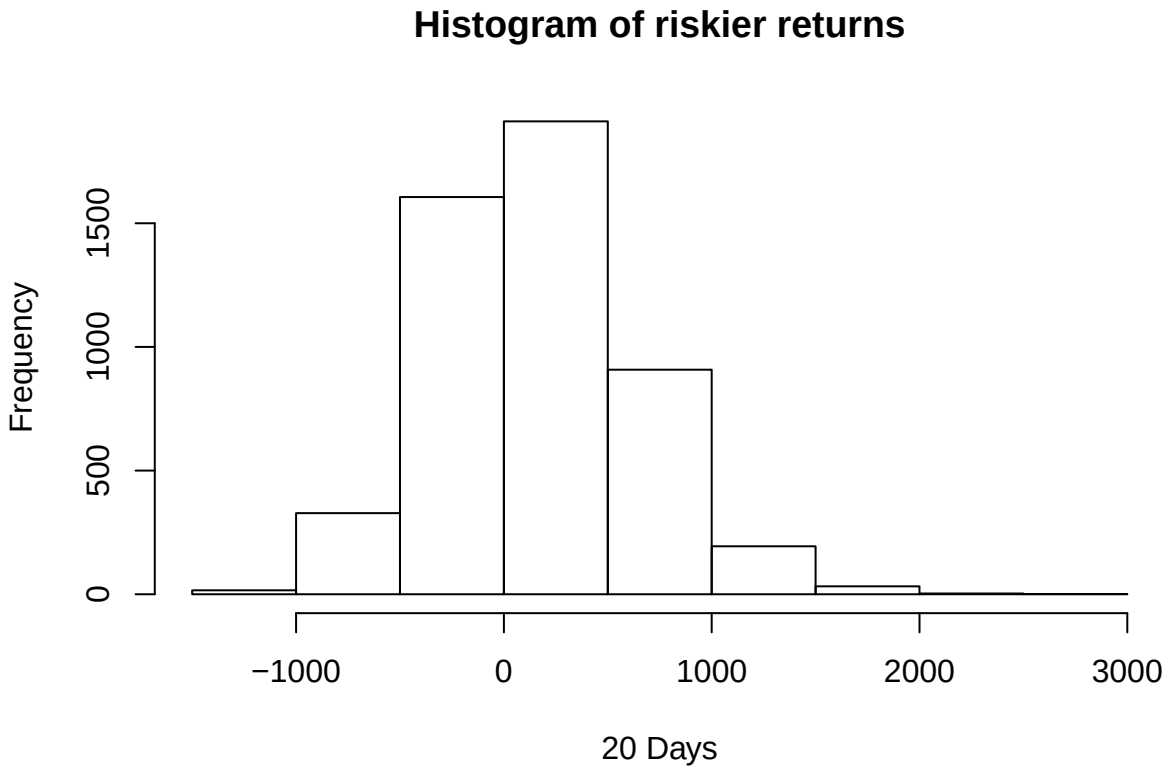
For a risker position than the even split above. These are high-yeild stocks, which provide the opportunity have much higher returns, but they are also have the potential to lose money as well

```
mystocks = c("WMT", "WM", "WFC", "KO", "LOAN")
myprices = yahooSeries(mystocks, from = "2011-01-01", to = Sys.timeDate())

YahooPricesToReturns = function(series) {
  mycols = grep('Adj.Close', colnames(series))
  closingprice = series[,mycols]
  N = nrow(closingprice)
  percentreturn = as.data.frame(closingprice[2:N,]) / as.data.frame(closingprice[1:(N-1),]) - 1
  mynames = strsplit(colnames(percentreturn), '.', fixed=TRUE)
  mynames = lapply(mynames, function(x) return(paste0(x[1], ".PctReturn")))
  colnames(percentreturn) = mynames
  as.matrix(na.omit(percentreturn))
}

myreturns = YahooPricesToReturns(myprices)
```

```r
plot(myreturns[,1], type='l', main = "Plot of riskier returns")
```




```r
sim1 = foreach(i=1:5000, .combine='rbind') %do% {
  totalwealth = 10000
  n_days = 20
  weights = c(0.2, 0.2, 0.2, 0.2, 0.2)
  holdings = weights * totalwealth
  wealthtracker = rep(0, n_days) # Set up a placeholder to track total wealth
  for(today in 1:n_days) {
    return.today = resample(myreturns, 1, orig.ids=FALSE)
    holdings = holdings + holdings*return.today
    totalwealth = sum(holdings)
    wealthtracker[today] = totalwealth
  }
  wealthtracker
}

hist(sim1[,n_days]- 10000, main = "Histogram of riskier returns", xlab = "20 Days")
```




```
quantile(sim1[,n_days], 0.05) - 10000
```

```
##        5%
## -569.2957
```

The histogram of the evenly spilt portfolio and the riskier porfolio both have a mean of gains/losses that are typically centered around zero. Whereas The safer portfolio has a mean that is usually a bit higher than zero.

The relative risk of these three portfolios can also be seen in the differences in the value at risk(VaR) of each portfolio at the 5% level. The riskier portfolio has the highest VaR and the safer portfolio has the lowest. The original porfolio's VaR sits between them.

$$y_{it} = \mu_{it} + \sigma_{it}\varepsilon_{it}$$

where i is asset, t is time, y is yield, μ is conditional mean, σ is conditional standard deviation, and ε is residual.

Bootstrapping can break the temporal correlations, but this could be avoided by bootstrappring only ε. Most stocks don't have temporal correlations. Volatility index, such as VIX, can have some temporal correlations.

Problem 4

NutrientH2O Twitter Market Segmentation

Section 1 - Overview

Introduction

The goal of this analysis was to identify any segments of NutrientH2O's twitter audience that may be useful for marketing purposes using statistical analysis. Taking a subset of approximately 8,000 NutrientH2O and 325,000 of their tweets, unsupervised analysis was conducted to group followers into meaningful groups. Multiple methods were attempted and the results for each are discussed and/or displayed in the sections below.

Analysis Details

Initially, simple summary tables of tweet types and correlations are presented below. From there, k-means clustering on the data was completed and the resulting groups are presented and discussed along with marketing implications. Next, more advanced clustering methods were attempted including hierarchical clustering with a variety of linkage functions and k-means clustering on a principal component analysis (PCA) of the twitter data.

Results

At a high level, the simplest clustering approach - k-means on the original dataset - yielded the most interpretable and practical results. Hierarchical clustering did not result in groupings that appear to have any relevance to marketing efforts. K-means clustering on PCAs, while lacking in interpretability, may have useful marketing implications when combined with other datasets.

Section 2 - Summary Analysis

Summary

Presented below are tables that display the top 10 tweet categories of tweets made by NutrientH2O followers, as well as any correlations between tweet type counts that exceeded .5 in absolute value. The idea of these high-level statistics are to get an overview of the data as well as confirm with data insights generated by human heuristics. For example, that followers of a nutritional beverage company will tend to tweet about health-related topics.

Top Tweet Categories

```
##                cat count  mean    sd
## 1          chatter 34671 4.399 3.529
## 2    photo_sharing 21256 2.697 2.732
## 3 health_nutrition 20235 2.567 4.496
## 4          cooking 15750 1.998 3.430
## 5         politics 14098 1.789 3.031
## 6    sports_fandom 12564 1.594 2.161
```

```
## 7            travel 12493 1.585 2.286
## 8       college_uni 12213 1.549 2.897
## 9    current_events 12030 1.526 1.269
## 10 personal_fitness 11524 1.462 2.405
```

The table above shows the top 10 tweet categories along with the associated mean per user and standard deviation across all users. The top two categories, chatter and photo sharing, can be expected to be the top categories across all tweets given the normal usage of Twitter. The next two categories - Health/Nutrition and Cooking - give insight into NutrientH2O's follower base, inidicating that they are health-conscious and pay attention to what they eat, which confirm normal intuition. This analysis could be vastly improved with the incorporation of a similar distribution across all of Twitter, which would allow for normalization of the data to understand which categories are significantly above or below their overall average.

Correlations

```
##                  Var1             Var2      Freq
## 1132 health_nutrition personal_fitness 0.8099024
## 590     online_gaming      college_uni 0.7728393
## 1171          cooking          fashion 0.7214027
## 991           cooking           beauty 0.6642389
## 255            travel         politics 0.6602100
## 1035         religion        parenting 0.6555973
## 943     sports_fandom         religion 0.6379748
## 1180           beauty          fashion 0.6349739
## 808  health_nutrition         outdoors 0.6082254
## 1015    sports_fandom        parenting 0.6077181
## 723             travel        computers 0.6029349
## 945               food         religion 0.5913181
## 505            chatter         shopping 0.5833732
## 728           politics        computers 0.5721506
## 1139         outdoors personal_fitness 0.5677903
## 440           politics             news 0.5618422
## 877               news       automotive 0.5554175
## 1017              food        parenting 0.5449481
## 109            chatter    photo_sharing 0.5362666
## 508     photo_sharing          shopping 0.5356210
## 295     sports_fandom              food 0.5326384
## 1107         religion            school 0.5162180
## 629       college_uni   sports_playing 0.5063748
```

This table shows any correlations between tweet types that exceed .5 in absolute value. Correlation is a measure of how two things vary together and ranges from -1 to 1. Values closer to 1 indicate a strong postive relationship, so that when one of the pair increases, so does the other. Values closer to -1 indicate a strong negative relationship, and values near zero indicate that the two items do not vary together. Looking at the table, results again seem to confirm human intuition. The highest correlation is between Health/Nutrition and Personal Fitness, indicating that among the users in the sample, these two tweet types often occur together. We also see some less-obvious relationships that yield insight into the overall audience. For example, Online Gaming and College/University often occur together, which indicates that there may be a group of followers who are college-age and into gaming, so that a promotion inside of an online game or located on a gaming website may be of value. Or, the high correlation between Cooking and Fashion, and Cooking and Beauty, means that advertisements in magazines that cover both of these topics could have high returns.

Section 3 - Initial Clustering

Summary

In an effort to identify meaningful groups, k-means clustering was carried out on the dataset to create clusters of similar users, which could ideally then be used to inform targeted marketing actions. The k-means algorithm defines a "distance function" to approximate how close or far away users are from each other, and iteratively tries different combinations to form the groups that best cluster the data together. A number of clustering attempts were carried out in an effort to identify the best parameters, and after they were determined, the created clusters were analyzed based on their Tweet characteristics. To determine the optimal number of clusters, a plot of within-cluster sum of squares was used along with the "elbow method" heuristic to identify the point where marginal returns have diminshed enough to justify not including an additional cluster. Once the cluster specifics were determined, the same clustering approach was carried out ten times and the averaged results are presented below.

Results - Cluster Details

```
##       [,1] [,2] [,3] [,4] [,5]
##  [1,]    0    0    0    0    0
##  [2,]    0    0    0    0    0
##  [3,]    0    0    0    0    0
##  [4,]    0    0    0    0    0
##  [5,]    0    0    0    0    0
##  [6,]    0    0    0    0    0
##  [7,]    0    0    0    0    0
##  [8,]    0    0    0    0    0
##  [9,]    0    0    0    0    0
## [10,]    0    0    0    0    0
```

```
##       Cluster Name                 Avg Size
##  [1,] "1"     "The College Gamers" "357 (5%)"
##  [2,] "2"     "The Worldly Nerds"  "334 (4%)"
##  [3,] "3"     "The Parents"        "601 (8%)"
##  [4,] "4"     "The Granolas"       "769 (10%)"
##  [5,] "5"     "The Romantics"      "120 (2%)"
##  [6,] "6"     "The Autocrats"      "344 (4%)"
##  [7,] "7"     "The Holy Rollers"   "72 (1%)"
##  [8,] "8"     "The Binge Watchers" "377 (5%)"
##  [9,] "9"     "The Stepford Wives" "496 (6%)"
## [10,] "10"    "The Rest"           "4325 (55%)"
##       Outstanding Tweet Types
##  [1,] "College/University, Online Gaming"
##  [2,] "Politics, Travel, News, Computers"
##  [3,] "School, Parenting, Religion, Sports, Food, Family"
##  [4,] "Personal Fitness, Health/Nutrition, Outdoors"
##  [5,] "Dating"
##  [6,] "Politics, News, Automotive"
##  [7,] "Religion"
##  [8,] "TV/Film"
##  [9,] "Fashion, Beauty, Cooking"
## [10,] "None"
##       Occurrences (out of 10)
##  [1,] "10"
```

```
##  [2,] "9"
##  [3,] "10"
##  [4,] "10"
##  [5,] "6"
##  [6,] "8"
##  [7,] "1"
##  [8,] "9"
##  [9,] "10"
## [10,] "10"
```

Results Interpretation

Displayed above are the average results for the 10 clusters created in the data, along with the average count/percentage for each (across all trials), a name to categorize them, and the types of tweets that categorize each cluster. The model contains some randomness, so it was repeated 10 times and results were averaged to reduce variability. Not every cluster appeared in every run, but the most common clusters are reported along with their appearance count across the 10 trials. Overall, on average, 45% of the data was grouped into a meaningful cluster, and the remaining 55% were grouped into what is effective an "other" category. These users did not fall neatly into one of the other groups, so this group exists in order to ensure the other groups are useful. Based on the results of these clusters, it is possible to create targeted marketing and promotional campaigns to different user groups that will likely be more successful than a blanket approach aimed at all users. For example, the "Parents" group may be more inclined to purchase with promotions relevant to their family/children, whereas the "Granolas" are likely inclined to listen to messaging related to health, nutrition, and exercise.

Section 4 - Alternate Clustering Approaches

In addition to the basic k-means cluster approach, the more complex approach of hierarchical clustering was attempted. This method attempts to create a hierarchy of groups of users where users in any subsequent level of grouping are more similar than in the level above. Multiple different linkage parameters were attemped, but none of the methods resulted in a hierarchy that appeared to be useful for segmentation purposes. That is to say, the hiearchy was imbalanced and grouped nearly all users into the same category with a few outliers elsewhere. Considering the structure of the data, and that there are not any readily apparent tiers or levels among what defines a twitter user's tweet distribution, it can be concluded that this method is not as useful as the k-means approach that uses only one layer/tier to form groups.

Section 5 - Principal Component Analysis and Clustering

Background

Because the user data contains so many dimensions - over 30 different types of tweet categories - a dimension reduction method may provide value in terms of simplifying computational complexity and summarizing types of users. Principal Component Analysis (PCA) attempts to represent a large number of dimensions with a relatively few number of components - usually less than five - with minimal loss in information. In the context of Twitter user data for NutrientH2O, the various tweet categories were condensed into only two components for each user and then the same clustering algorithm was applied to this smaller set of data. Details are provided in the appendix.

Results Interpretation and Usage

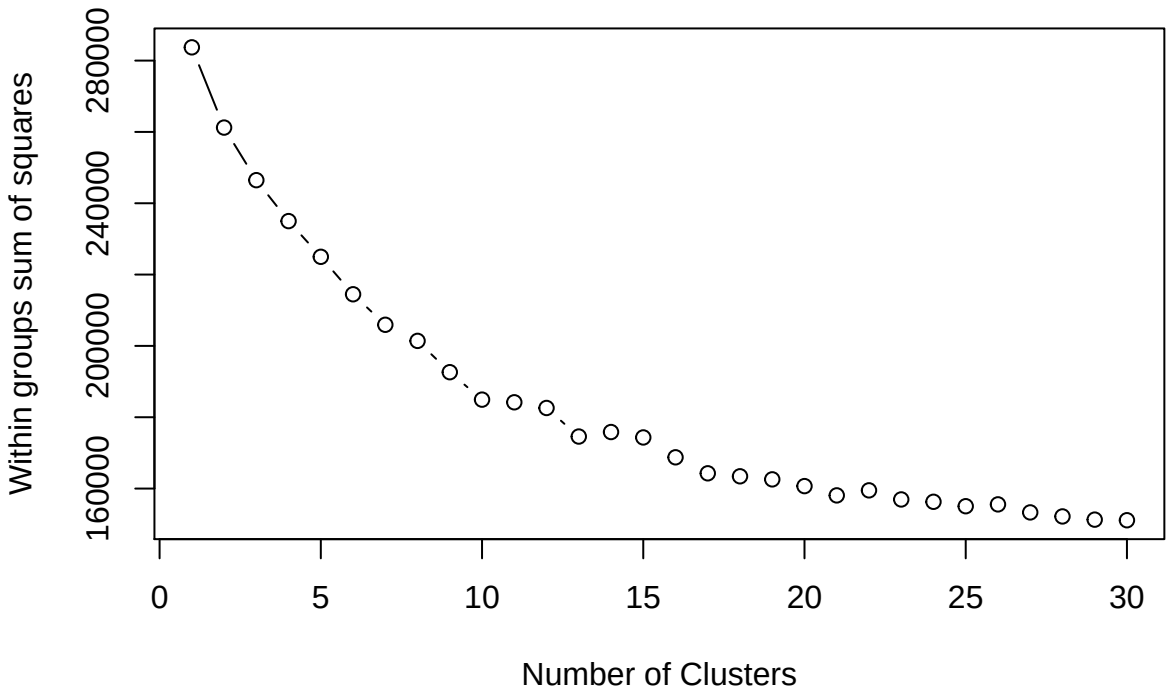
A key drawback of PCA is that because each component is composed of bits and pieces of each of the "real" dimensions, the results are not easily interpretable. As such, they are not presented here in detail. A review of

these components would do nothing to characterize individual users based on their tweeting habits, which is the only goal of this exercise as well as the only activity possible with the available data. However, PCA may have meaningful implications for marketing applications when combined with other types of data not provided here. For example, the Principal Components of the twitter users in this dataset may be correlated with other indicators of potential opportunities, such as user engagement (likes, re-tweets, etc.) or demographic information.

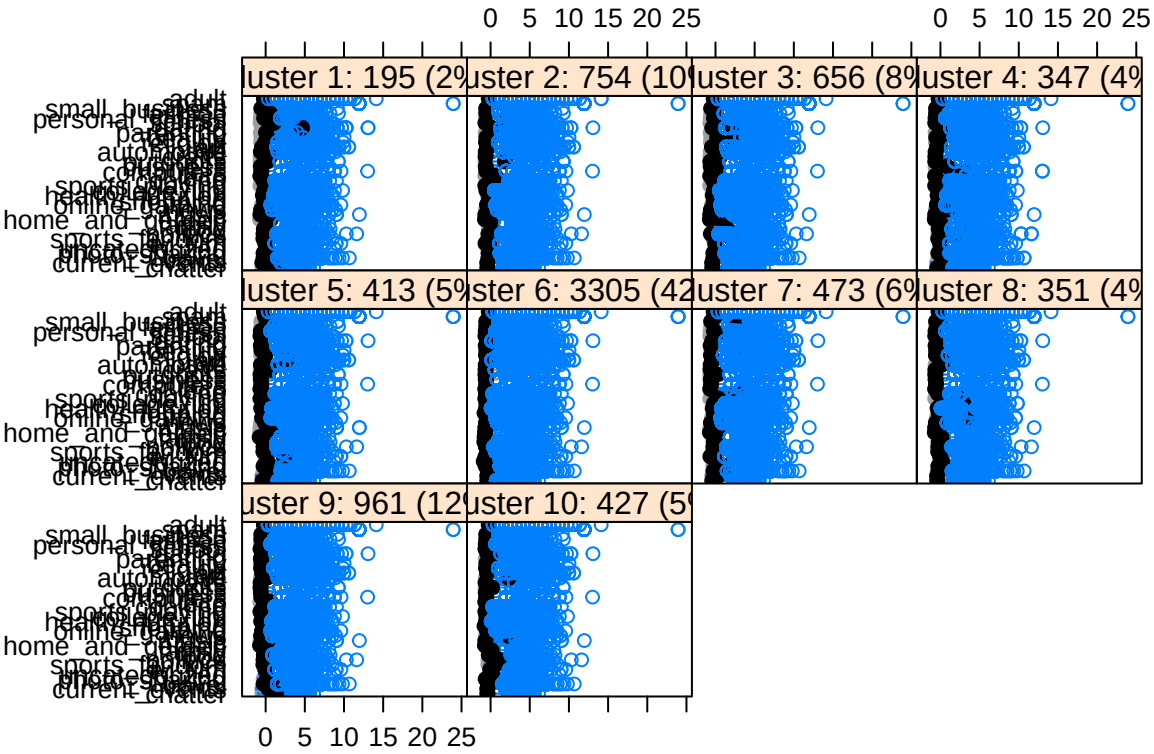
Appendix

Section 3 - Initial Clustering

Choosing Number of Clusters

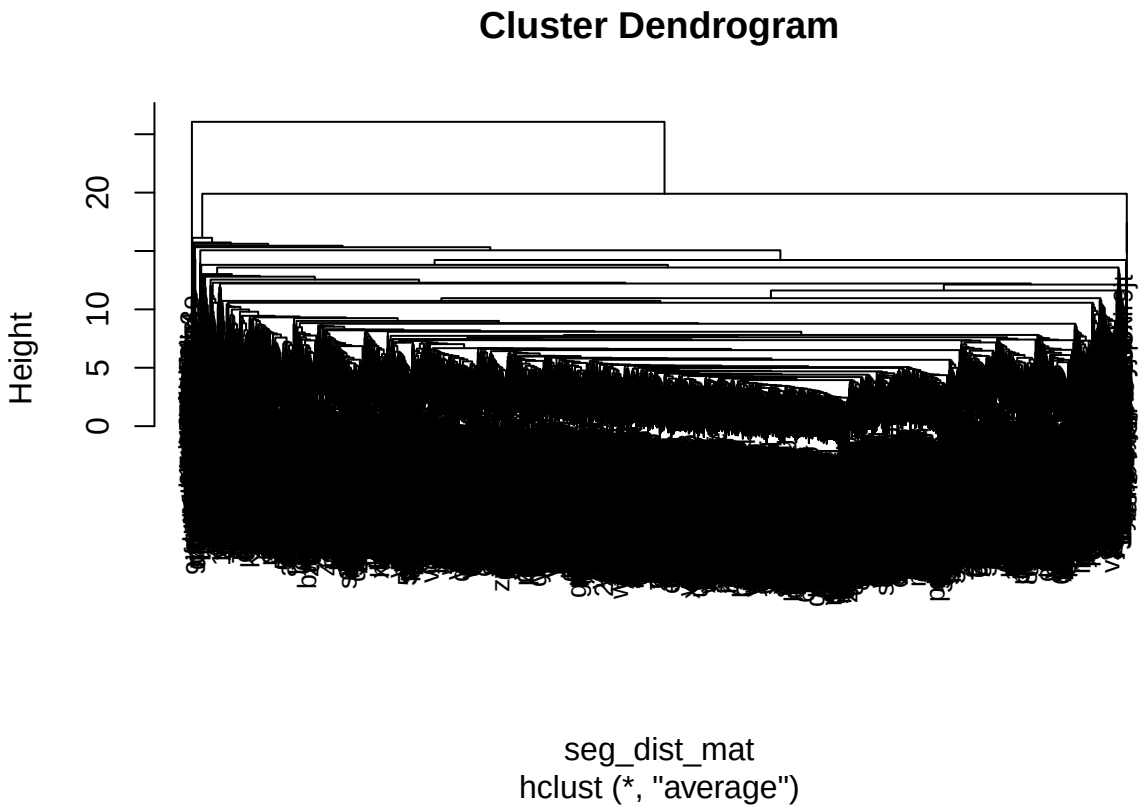


Box and Whisker Plot of Tweet Categories by Cluster



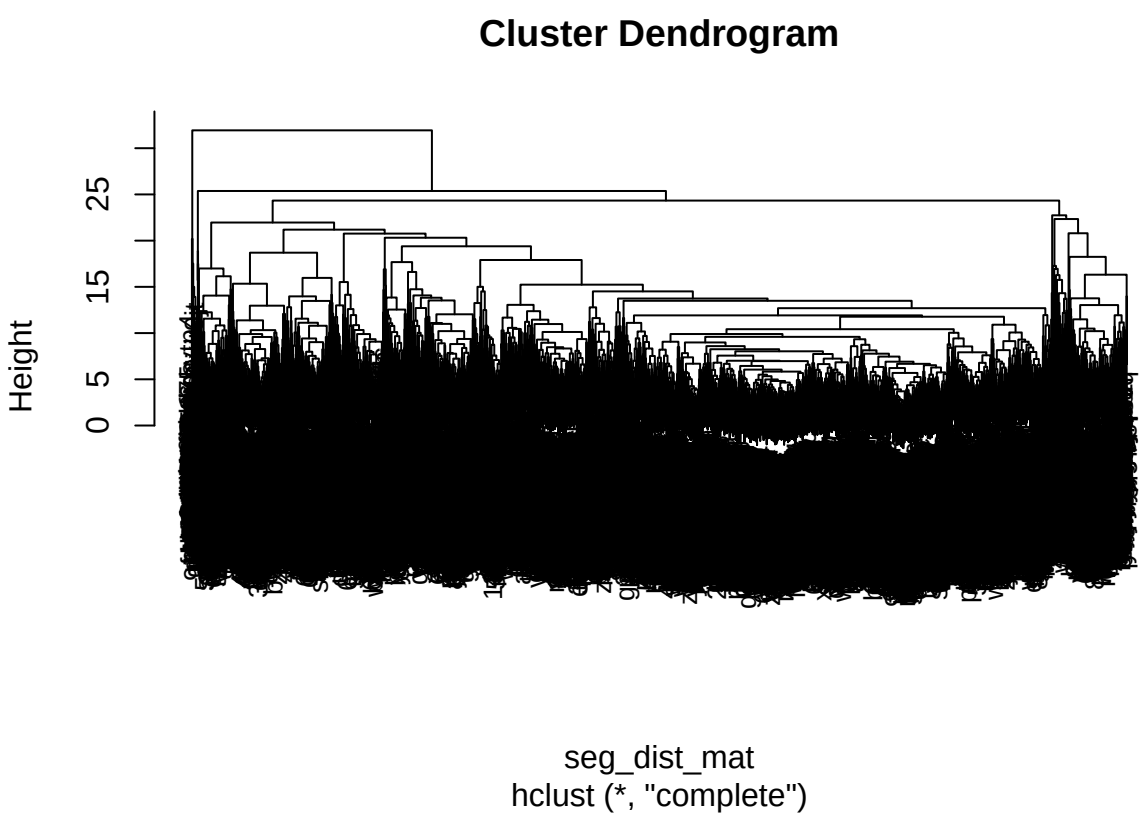
This plot shows a box-and-whisker plot for each cluster displaying the distribution of tweet categories for each cluster. Shaded clusters in each plot represent a significant increase over the entire population, so these are used to categorize each group. Please keep in mind that the results presented in Section 3 represent an average of 10 trials, whereas this plot is a sample from one result, so the numbers and clusters will not align exactly.

Section 4 - Alternate Clustering Approaches



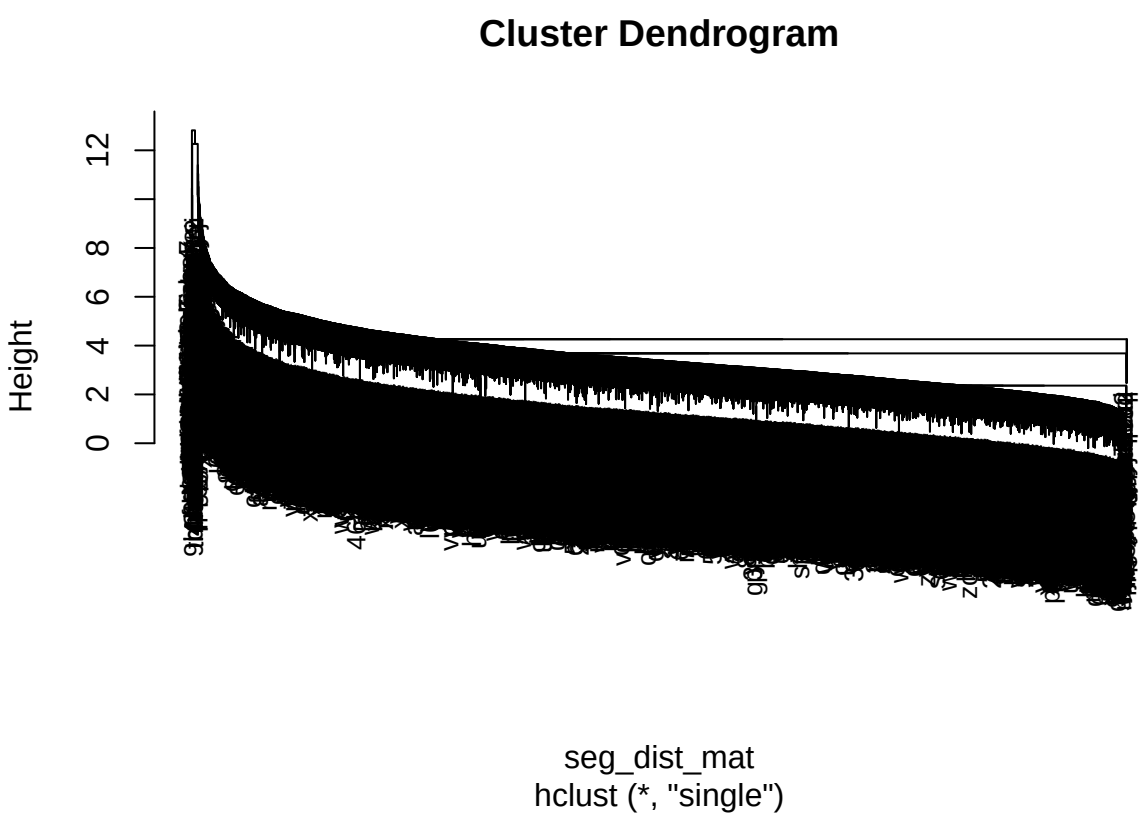

```
##    1    2    3    4    5    6    7    8    9   10
## 7802   10   47    7    8    2    2    1    2    1
```

Using an average linkage method, the dendrogram and table above show that this instance of hierarchical clustering does not yield meanigful groups, as nearly all users are located in the first group, with a few ouliers elsewhere.




```
##    1    2    3    4    5    6    7    8    9   10
##  487 5628  284  859  130   16  410   49    9   10
```

The complete linkage method provides a greater spread across 10 segments of the hierarchical cluster, but does not perform as well as the k-means method. Combined with the significant incrase in computational complexity over k-means, this method is a worse performer.

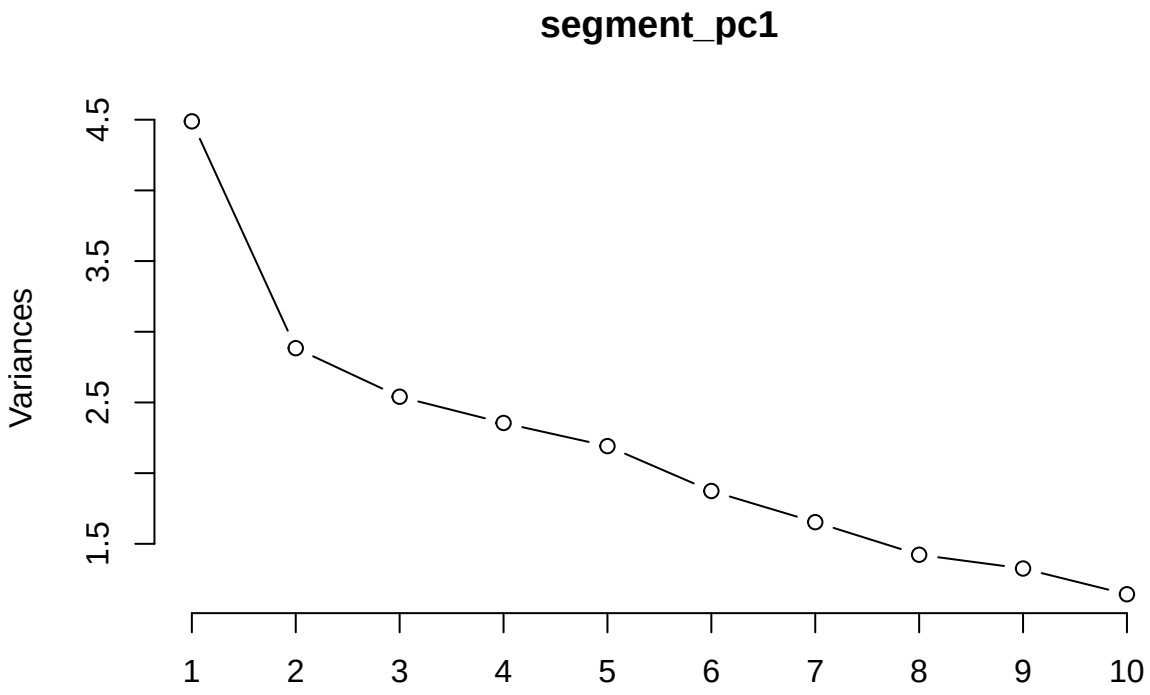



```
##    1    2    3    4    5    6    7    8    9   10
## 7827   47    1    1    1    1    1    1    1    1
```

The single linkage method performs approximately as poorly as the average method, and is not useful nor interpretable.

Section 5 - PCA Clustering

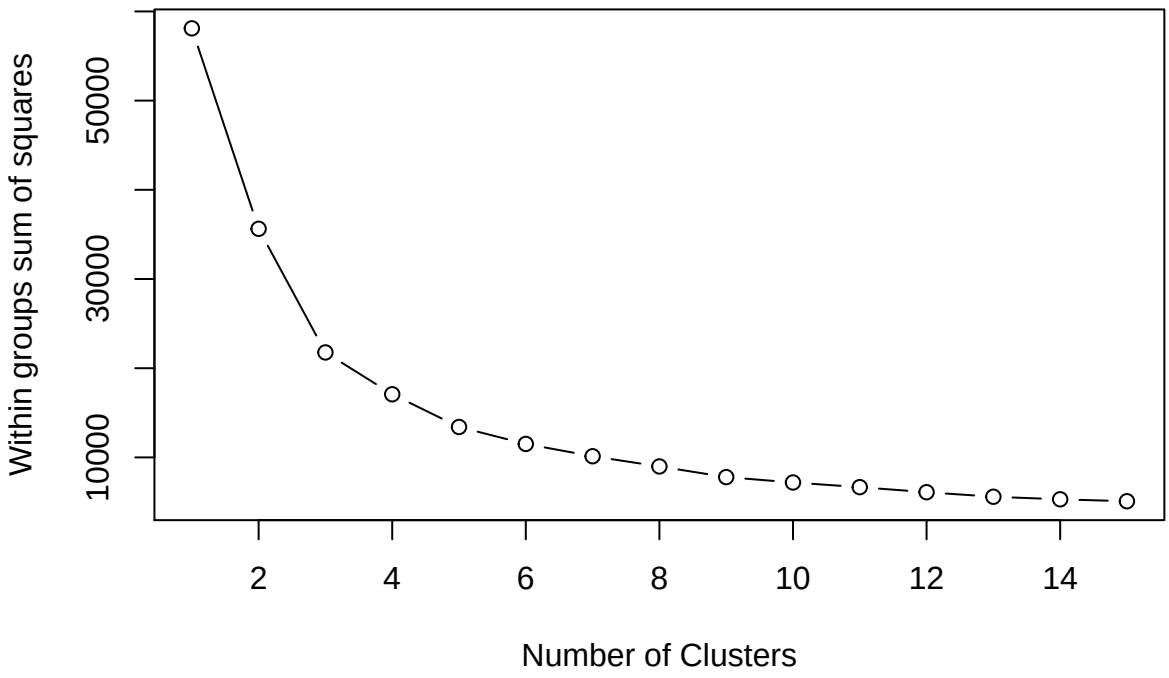
Choose Optimal # PCAs



The plot above reveals that 1-2 principal components represent most of the variance across the entire dataset. For analsis purposes, 2 PCs were chosen.

Determine Optimal # Clusters Using PCA

```
## Warning: did not converge in 10 iterations
```




```
## [1] 5916.338
```

Clustering on the two PCs above is most effective with 3 clusters based on the elbow method. The residuals and total SSE is calculated, but the result is not comparable to the results from clustering the original dataset.

Repeat with 10 Clusters for Sake of Comparison

```
## [1] 1973.356
```

For comparison, the clustering using two PCs was completed again using the same number of clusters as with the original dataset (10). The total SSE is much lower than for 3 clusters, but again is not comparable to the original attempt due to the conversion in units. Incorporating additional data sources, as well as a predictive/supervised element may yield insight as to whether PCA is useful from a marketing and segmentation perspective.

Class comments:

1. Sometimes one market segment is not only defined by one single category, if we try to use less clusters (less than 10), some market segments may get grouped together and they are likely to be more the more precise market segment. The following code shows another possible way to do the clustering and the possible intepretation in terms of the market segmentation could be:

second cluster: high positive weight on sports_fandom,food,family,religion,parenting and school,so this group may include married people who pay more attention to their family and parenting related topic.

third cluster: high positive weight on cooking,beauty and fashion,so this group should consist of younger woman and younger housewives who cook a lot and pay a lot of attention to beauty and fashion.

fourth cluster: high positive weight on politics,news,travel,computer and automotive,so this group might be younger man who are interested in politics, read lots of news online, loves computer and automotive and travels a lot.

fifth cluster: high positive weight on health_nutrition, outdoors and personal_fitness, so this group of people really care about nutrition and fitness, and they have lots of outdoor activity to keep fit.

sixth cluster: high positive weight on online_gaming,college_uni and sports_playing,this group is very likely to be college student whose main entertainment is online_gaming and sports.

first cluster's highest positive weight is on chatter, and nothing else really stands out. So we think this cluster may be just everything that is left out and is hard to get into any of the other market segment.

```
library(fpc)

cluster_6 <- kmeans(segment, centers=6, nstart=50)
names(cluster_6)
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
cluster_6$centers
```

```
##    chatter current_events   travel photo_sharing uncategorized  tv_film
## 1 4.093137       1.436275 1.529412      2.647059     0.8455882 1.409314
## 2 2.919512       1.379024 1.114390      1.507317     0.7107317 1.037561
## 3 9.883775       1.862715 1.194228      5.622465     0.9001560 1.053822
## 4 4.002026       1.548126 1.320162      2.428571     0.9574468 1.063830
## 5 4.027304       1.680887 6.337884      2.259386     0.7218430 1.165529
## 6 3.949904       1.712909 1.448940      5.909441     1.2080925 1.007707
##   sports_fandom  politics     food    family home_and_garden     music
## 1      1.627451 1.2818627 1.465686 1.1421569       0.5563725 0.7745098
## 2      1.512927 0.9587805 1.188537 0.7541463       0.4492683 0.5602439
## 3      1.628705 1.4656786 1.242590 0.9937598       0.5951638 0.7893916
## 4      1.531915 1.3019250 2.294833 0.9108409       0.6220871 0.7507599
## 5      2.211604 9.8771331 1.767918 0.9846416       0.5853242 0.6279863
## 6      1.543353 1.3333333 1.252408 0.9653179       0.6069364 1.1946050
##        news online_gaming shopping health_nutrition college_uni
## 1 0.8651961    10.4240196 1.227941        1.6053922  10.7916667
## 2 0.8051220     0.5607317 0.735122        0.9485366   0.8985366
## 3 0.8338534     0.7917317 3.546022        1.3104524   1.2316693
## 4 1.2289767     0.9493414 1.328267       12.2826748   1.0425532
## 5 5.1501706     0.8293515 1.186007        1.4232082   1.3242321
## 6 1.0558767     1.0366089 1.703276        2.0308285   1.4296724
##   sports_playing    cooking       eco computers  business  outdoors
## 1      2.4901961  1.5563725 0.4705882 0.5784314 0.3848039 0.6372549
## 2      0.4385366  0.8217073 0.3587805 0.3895122 0.3190244 0.4585366
## 3      0.5982839  1.1716069 0.6942278 0.6146646 0.5858034 0.5257410
```

```
## 4      0.6433637  3.3617021 0.8662614 0.5623100 0.4468085 2.4397163
## 5      0.6604096  1.2508532 0.5972696 2.6484642 0.6587031 0.8924915
## 6      0.8381503 11.9325626 0.5394990 0.7475915 0.5645472 0.8169557
##      crafts automotive       art  religion    beauty parenting    dating
## 1 0.6250000  0.9485294 1.2230392 1.0514706 0.5220588 0.9901961 0.7279412
## 2 0.4051220  0.6060976 0.6519512 1.0780488 0.4397561 0.8387805 0.4612195
## 3 0.6411856  1.0499220 0.6645866 0.9461778 0.5483619 0.9188768 1.1817473
## 4 0.6474164  0.6889564 0.8226950 1.1124620 0.5460993 1.0111449 0.9949341
## 5 0.6365188  2.0443686 0.6860068 1.4027304 0.5153584 1.1808874 1.0443686
## 6 0.6088632  0.8574181 0.9152216 1.2562620 3.8497110 1.0616570 0.5895954
##      school personal_fitness   fashion small_business        spam
## 1 0.6764706        1.0245098 0.9411765      0.4142157 0.009803922
## 2 0.6665854        0.6395122 0.5285366      0.2778049 0.007560976
## 3 0.9906396        0.9555382 0.8876755      0.4313573 0.003120125
## 4 0.7304965        6.1175279 0.8176292      0.2695035 0.007092199
## 5 0.8583618        0.9300341 0.6928328      0.4863481 0.005119454
## 6 1.0558767        1.3025048 5.6897881      0.4605010 0.003853565
##       adult
## 1 0.4534314
## 2 0.4534146
## 3 0.3510140
## 4 0.3242148
## 5 0.2474403
## 6 0.4238921
```

```
head(sort(cluster_6$centers[1,], decreasing=TRUE), 10)
```

```
##      college_uni    online_gaming          chatter    photo_sharing
##        10.791667        10.424020         4.093137         2.647059
##   sports_playing    sports_fandom health_nutrition          cooking
##         2.490196         1.627451         1.605392         1.556373
##           travel             food
##         1.529412         1.465686
```

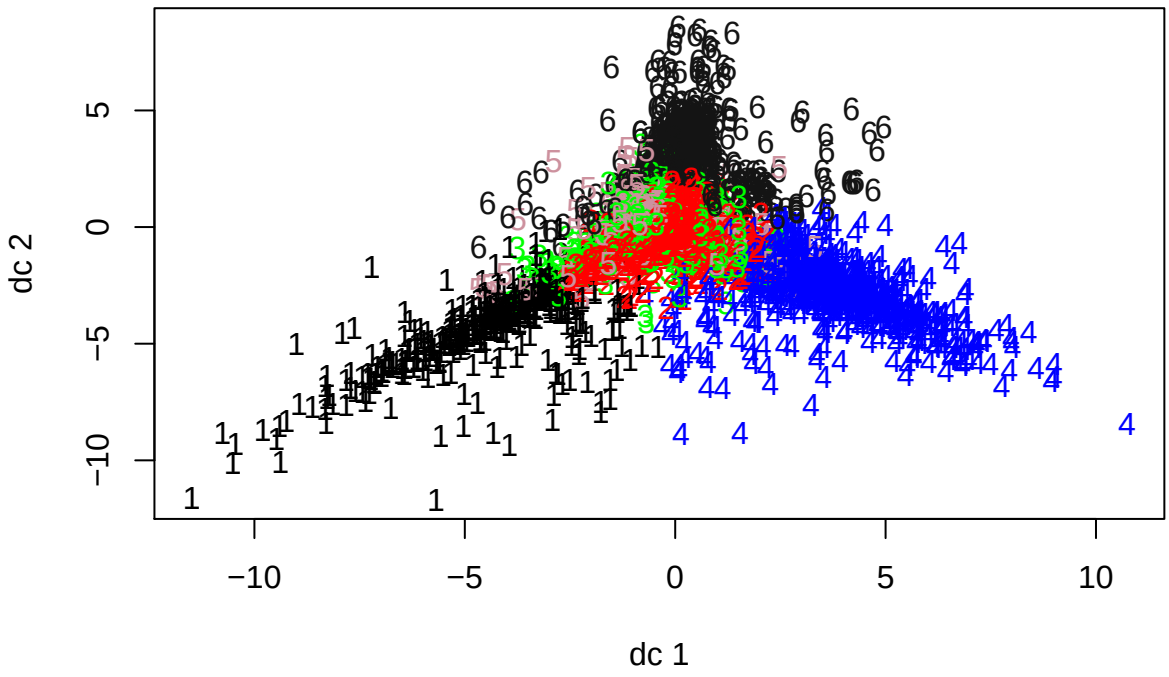
```
head(sort(cluster_6$centers[2,], decreasing=TRUE), 10)
```

```
##          chatter    sports_fandom    photo_sharing   current_events
##        2.9195122        1.5129268        1.5073171        1.3790244
##             food           travel         religion          tv_film
##        1.1885366        1.1143902        1.0780488        1.0375610
##         politics health_nutrition
##        0.9587805        0.9485366
```

```
head(sort(cluster_6$centers[3,], decreasing=TRUE), 10)
```

```
##          chatter    photo_sharing         shopping   current_events
##         9.883775         5.622465         3.546022         1.862715
##    sports_fandom         politics health_nutrition             food
##         1.628705         1.465679         1.310452         1.242590
##      college_uni           travel
##         1.231669         1.194228
```

```
plotcluster(segment, cluster_6$cluster)
```



2. We can also use wordcluster to plot out the top words in each category to show the characteristics of each market segment more intuitively.

3. There are more and more companies using clusters or PCA to get the clout score, then using that score to formulate product/marketing strategy.