Predictive Modeling HWK1

Alec Grubbs

August 5, 2015

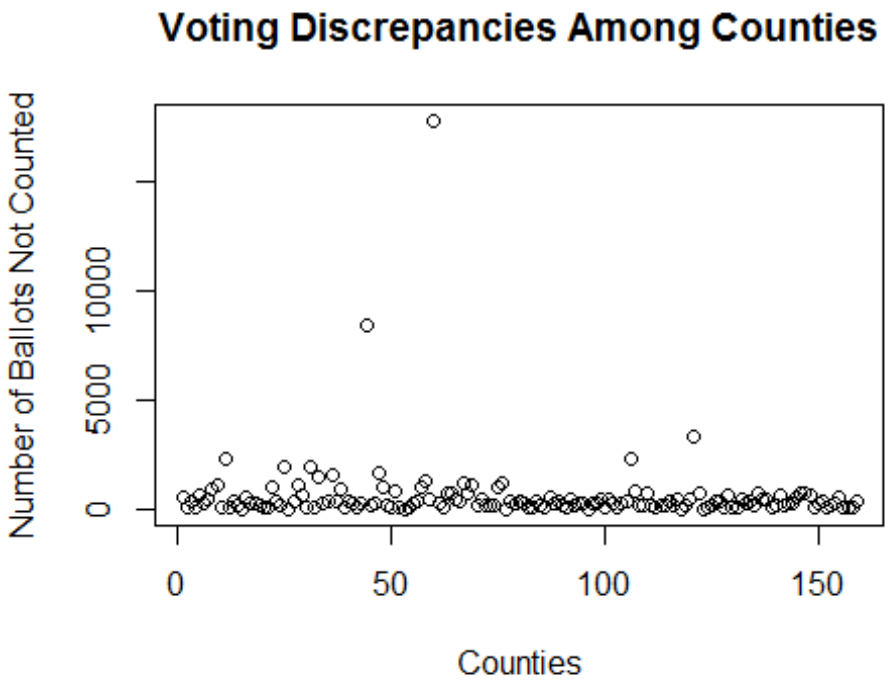
Exploratory Analysis

Here was are looking at a sample of the dataset. The data is composed of each Georgia county with the main variables we are interested in being: ballots, the number of ballots cast in that county; votes, the number of votes recorded in that county; equip, the type of voting equipment used with the categories of Lever, Optical, Paper, and Punch; poor, given a numerical value of 1 if more than 25% of the residents in a county live below 1.5 times the federal poverty line, or 0 otherwise; perAA, the percent of people in the county who are African-American.

```
##     county ballots votes   equip poor urban atlanta perAA gore bush
## 1  APPLING    6617  6099   LEVER    1     0       0 0.182 2093 3940
## 2 ATKINSON    2149  2071   LEVER    1     0       0 0.230  821 1228
## 3    BACON    3347  2995   LEVER    1     0       0 0.131  956 2010
## 4    BAKER    1607  1519 OPTICAL    1     0       0 0.476  893  615
## 5  BALDWIN   12785 12126   LEVER    0     0       0 0.359 5893 6041
## 6    BANKS    4773  4533   LEVER    0     0       0 0.024 1220 3202
```

Voting Equipment

We start out by comparing the number of ballots placed versus the votes counted for each county. We can clearly see several large outliers



We find the mean of the difference between ballots and votes among all counties to be 595.478

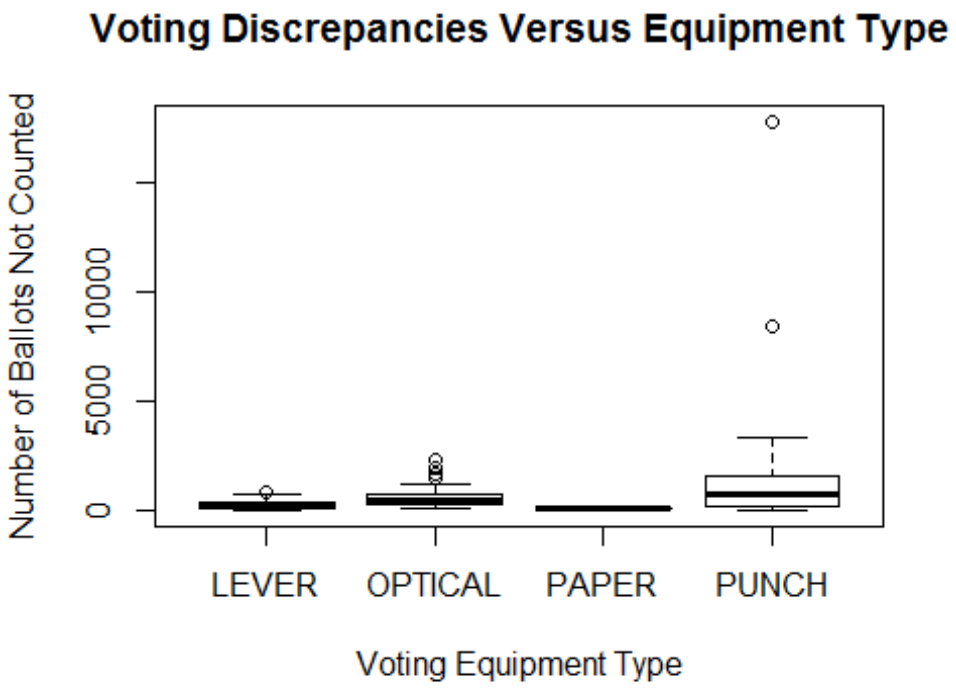
```
mean(ballots-votes)
```

```
## [1] 595.478
```

We can also quantify the number of counties with a difference greater than the mean and divide them up by the voting equipment that is used. Here we see that paper equipment does not contain any of these outliers, while 62% of the outliers are from optical voting machines.

```
## equip
##   LEVER OPTICAL   PAPER   PUNCH
##       5      23       0       9
```

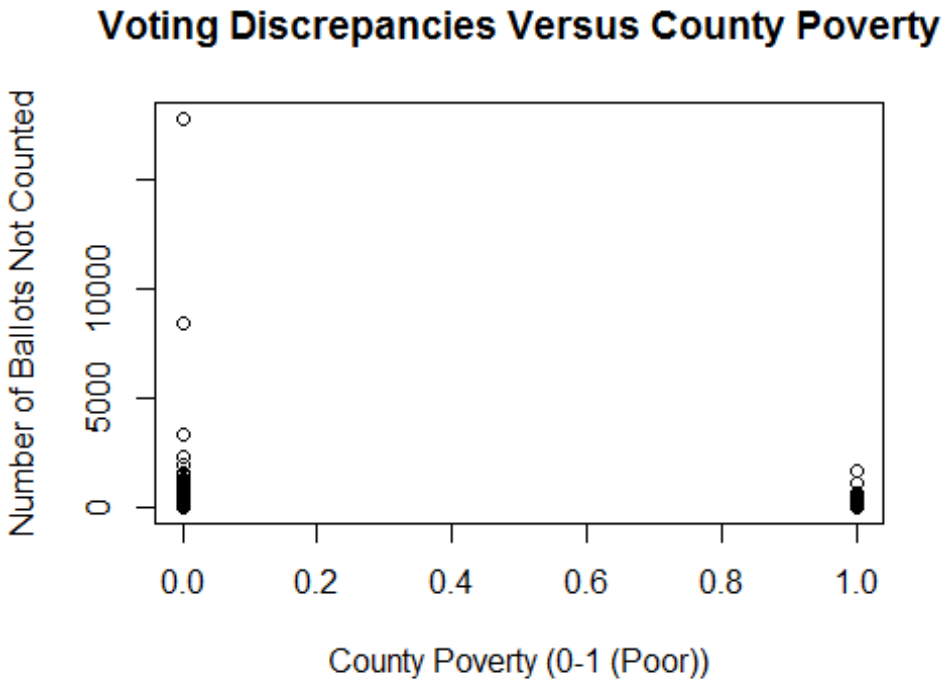
Performing a histogram plot of the difference and grouping by type of equipment also displays on which equipment the majority of outliers lie. We can see that both optical and punch methods have a large range of differences compared to the remaing two methods, but even averaging on those individual equipments shows outliers, especially the two outliers we see for the punch equip.



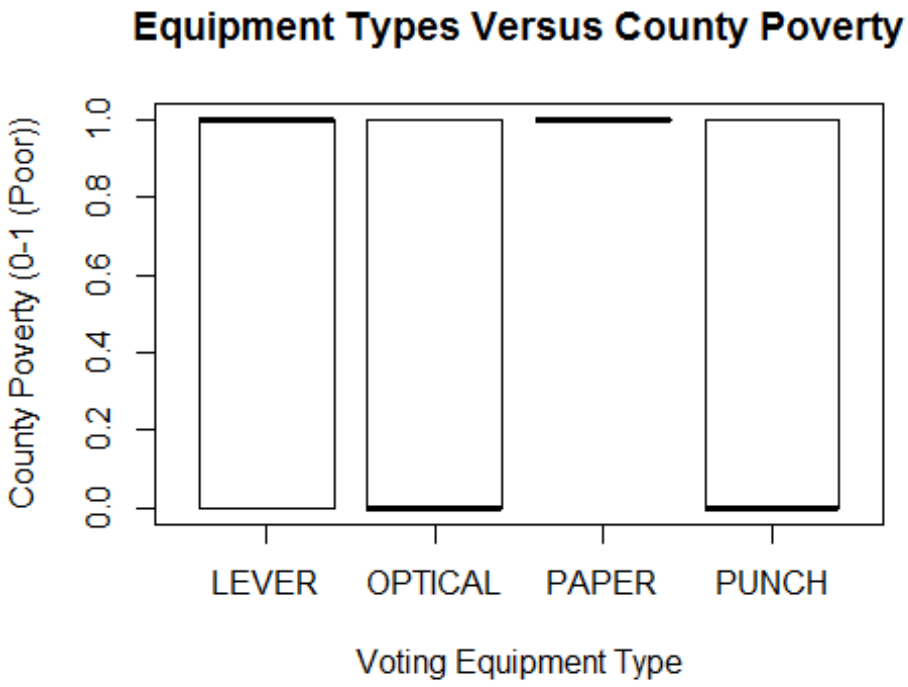
Poor and Minority Communities

Given that we have seen that the optical and punch card methods of voting are more likely to have higher rates of undercount, we now want to see if that bias impacts the counties that are considered poor or contain minority communities

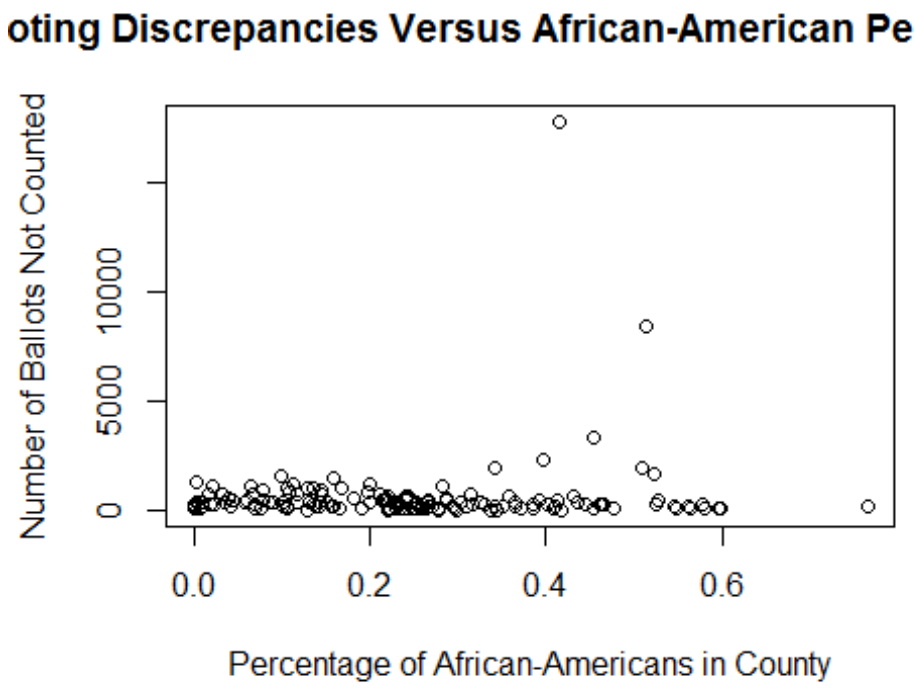
Graphing the difference versus whether the county is considered poor shows us visually the the majority of outliers exist in non-poor counties.



This is further seen by graphing which equipment was used by the counties, and we see all paper equipment, with the lowest number of difference outliers, was exclusively used by poor counties. The lever equipment was also primarily used by poor counties. The two equipments with the largest number of outliers, optical and punch, were more likely to be used by non-poor counties. From this we can conclude that the use of equipment that resulted in large differences between ballots and votes was in fact more impactful on non-poor counties, than poor ones.



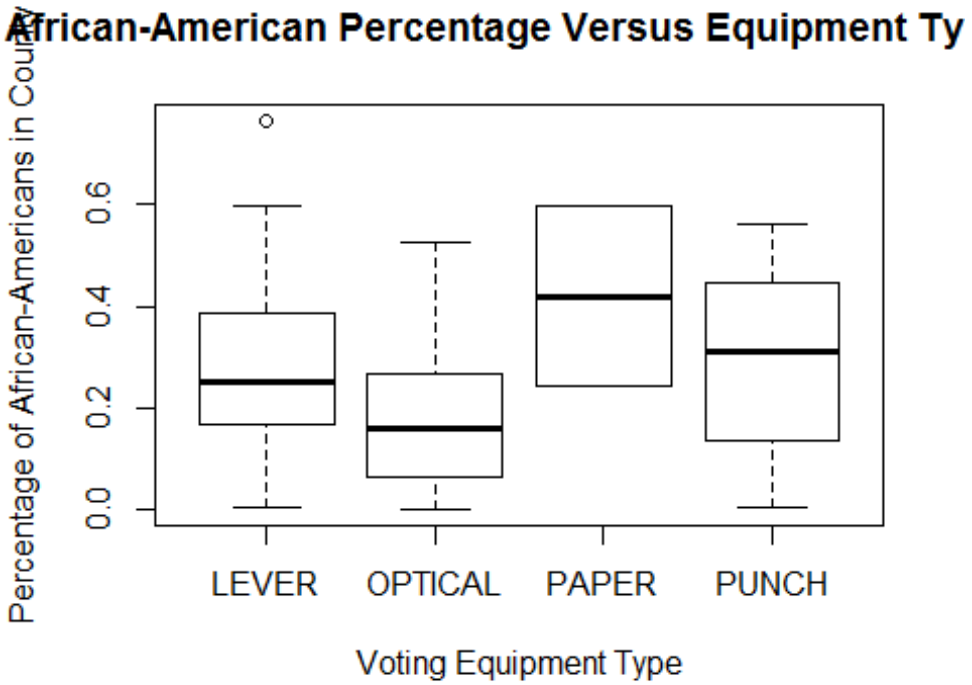
Moving to counties with a large minority percentage, we plot the difference versus the percentage of minorities and see a different result than with poor counties. The largest outliers are now in counties with a minority percentage greater than 0.4.



However, creating a table and seperating the counties at 30% minority population still shows that over 70% of outliers (difference greater than the mean difference of all counties) occur in counties with a low percentage of minorities.

```
## perAA < 0.3
## FALSE  TRUE
##    11    26
```

Creating a histogram plot of the minority percentage versus equipment used shows us that minority counties with a minority percentage greater than 30% were more likely to use a paper method, which had no difference outliers, and less likely to use the optical equipment which contained the majority of outliers.

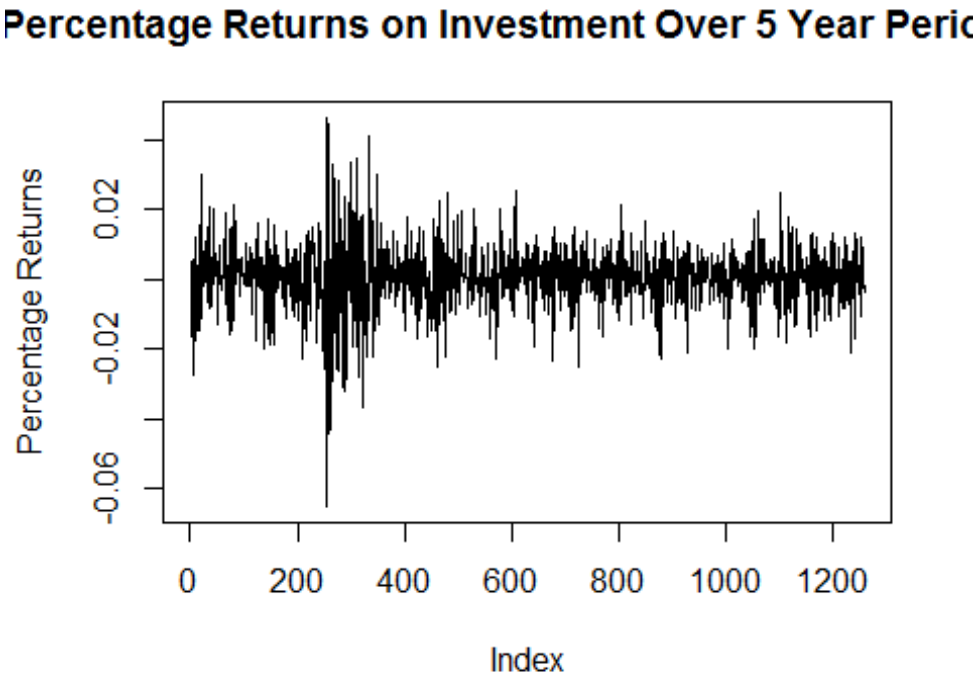


Based on this visual analysis it would be correct to say that we should not worry about difference bias from equipment affecting poor and minority counties, as these counties are more likely to use the types of equipment that do not contain the difference outliers.

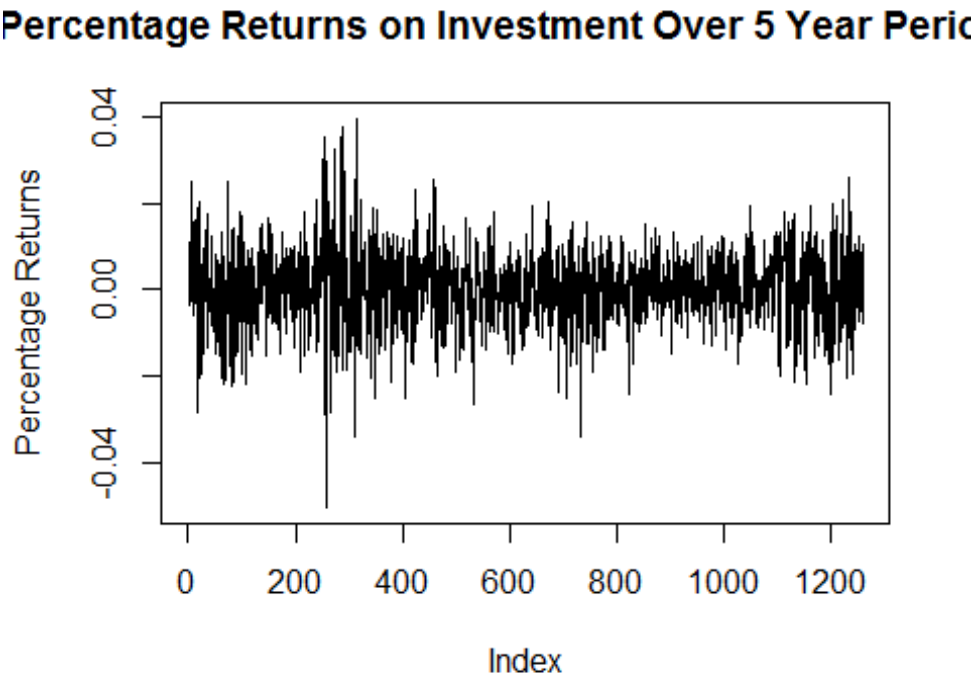
Bootstrapping

Characterizing Exchange Traded Funds (ETFs)

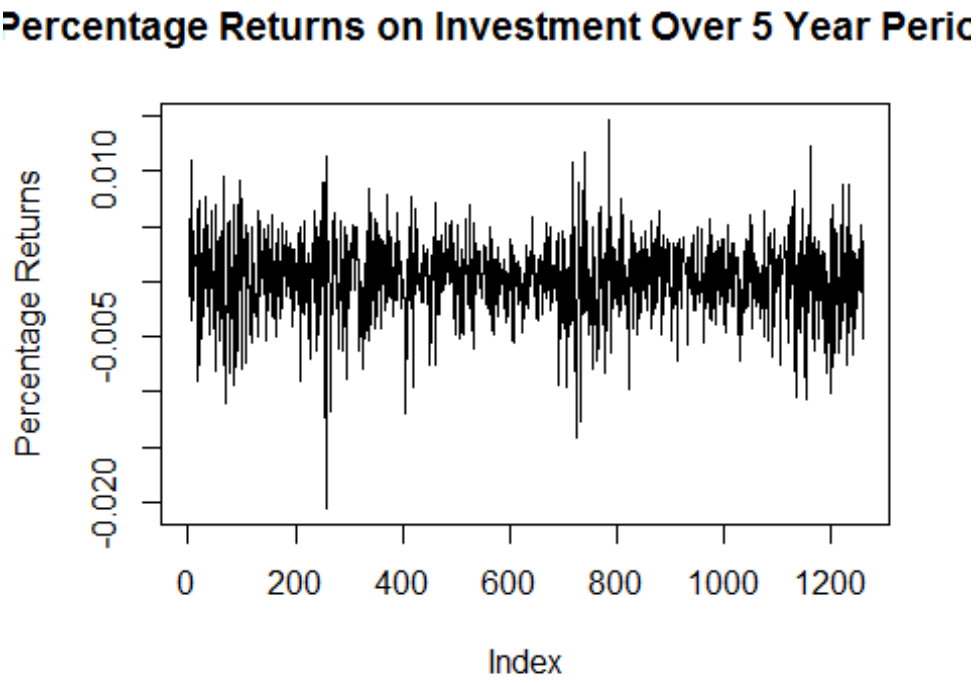
SPY has a mean of 0.000615 and a standard deviation of 0.00935



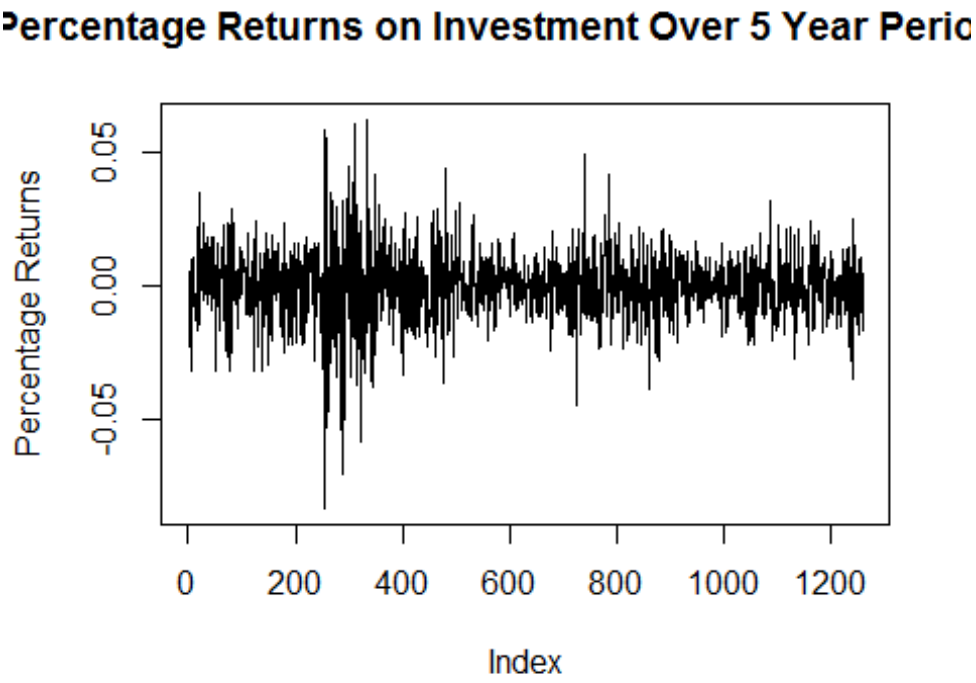
TLT has a mean of 0.000344 and a standard deviation of 0.00977



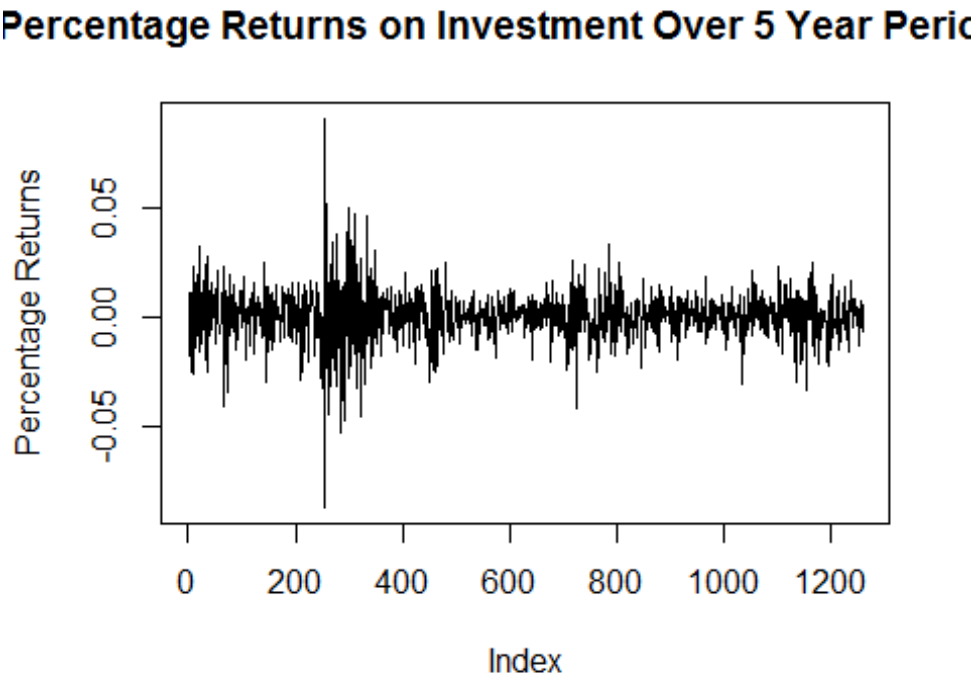
LQD has a mean of 0.000202 and a standard deviation of 0.00358. LQD produces the lowest range of deviations from the mean and therefore would be considered low risk



EEM has a mean of 5.76e-5 but a standard deviation of 0.01374. EEM produces the highest range of rate of returns with the largest standard deviation. The most risky ETF.



VNQ has a mean of 0.00054 and a standard deviation of 0.0115. Another risky ETF.



Initial Investment In All Three Portfolios will be $100,000

Three different Portfolios will be considered.

Equal weighting of investment among all 5 ETFs

```
holdings1 = total_wealth*c(0.2,0.2,0.2,0.2,0.2)
```

The 'safe-bet' weighting of the investment by investing in the three ETFs with the lowest standard deviation. Favors LQD which has the lowest standard deviation of all the ETFs in this study

```
holdings2 = total_wealth*c(0.2,0.2,0.6,0,0)
```

A riskier investment, focusing on the two ETFs with the greatest standard deviation

```
holdings3 = total_wealth*c(0,0,0,0.55,0.45)
```

The growth of the portfolios will be estimated over a 20 trading day period

The simulation will be run 500 times to produce an accurate estimate of the Value at Risk (VaR)

Calculating the 5% value at risk for each holding

Holding 1: Balanced Weight Distribution

```
##        5%
## -3897.946
```

Holding 2: 'Safe-bet' Weight Distribution

```
##        5%
## -1803.936
```

Holding 3: 'Risky' Weight Distribution

```
##        5%
## -8394.088
```

A Value at Risk (VaR) can best be explained with an example: if a portfolio of stocks has a one-day 5% VaR of $1000, then there is a 5% probability that the portfolio will fall in value by more than $1000 over that one-day period if no trading occurs. Since we are running our simulation over 500 times, we can expect that 5% of those interations will on average return the 5% VaR returned. We ran the simulation 500 times in order to get an average of the 5% VaR for each portfolio distribution.

The results we as expected given the weighting of the portfolios. The portfolio that was considered 'safer', holdings2, by weighting the ETFs to favor the three ETFs with the lowest standard deviations, had the lowest 5% VaR at -1896. When equal distribution of weight between the five ETFs was used in 'holdings1' produced a 5% VaR that was just about $1700 more than holdings2. The riskiest portfolio, holdings3, which was made up of the two ETFs with the greatest variability in rate of return, produced a 5% VaR that was almost 4 times the 5% VaR of the 'safe-bet'.

If this simulation is used to predict how real portfolios will respond in the market, then which one is the best depends on the buyer. If the buyer favors a conservative allocation of their wealth, than a distribution like that in holdings2 would be a recommendation. Although it is less likely to produce the same gains as a risker portfolio, is it also less likely to experience the market swings and losses associated with risker investments. A risker portfolio can produce great rewards for the investor if the market is on the rise, however it is also likely to see greater losses if the market experiences a recession. Finally the equal weight portfolio in this simulation was closer to the conservative portfolio that the risker one. If the buyer wanted a truely middle-of-the-road portfolio, they would be advised to invest more in the risker ETFs such as EEM and VNQ to increase their potential earnings at the potential loss of financial stability.

Clustering and PCA

We begin by examining the data in our dataset. We have been provided roughly 6500 bottles of wine as our observations, and these bottles are described using eleven chemical properties such as acidity and alcohol percentage.

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality color
## 1       5   red
## 2       5   red
## 3       5   red
## 4       6   red
## 5       5   red
## 6       5   red
```

In addition to the chemical properties, the color of each wine is provided: with red or white, as well as a subjective quality of the wine on a 1-10 scale, with our observations having qualities between 3 and 9.

If we want to visualize the data, we will need to reduce the dimensionality, as it would be very difficult to visualize in terms of all 11 chemical descriptors. In this case, it would be much easier to visual the data in just two dimensions, so principle component analysis (PCA) will be used to do just that. It should be noted that PCA is senstitive to the relative scaling of the original variables, so they must first be scaled in order to produce better results.

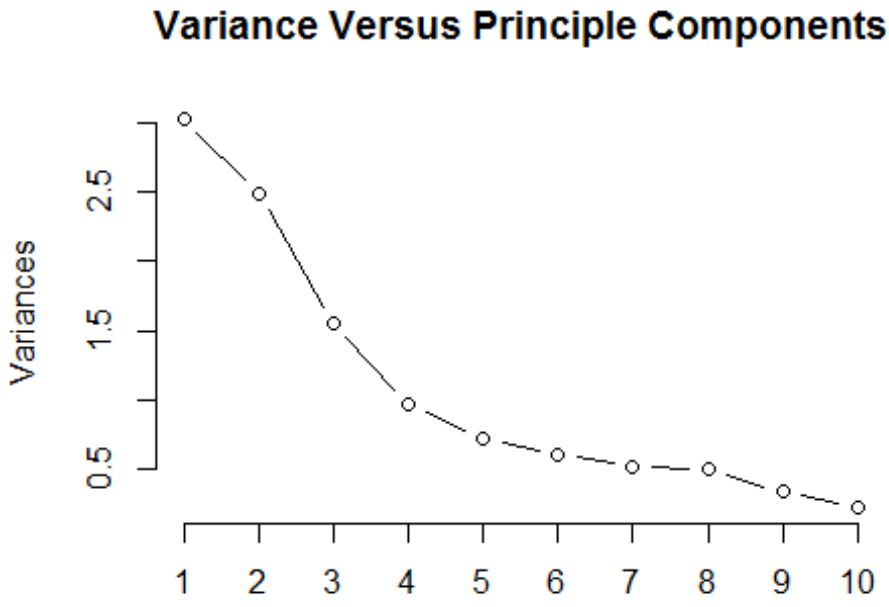
PCA will create a number of principle components (PC) equal to the number of variables in our data. Each PC contains a loading of the chemical variables to produce a new variable. If we examine the cumulative proprotion of variance amongst the PCs, we can understand how much of the information of the original data can be described by the combindation of the new variables by looking at the cumulative proportion row.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253
```

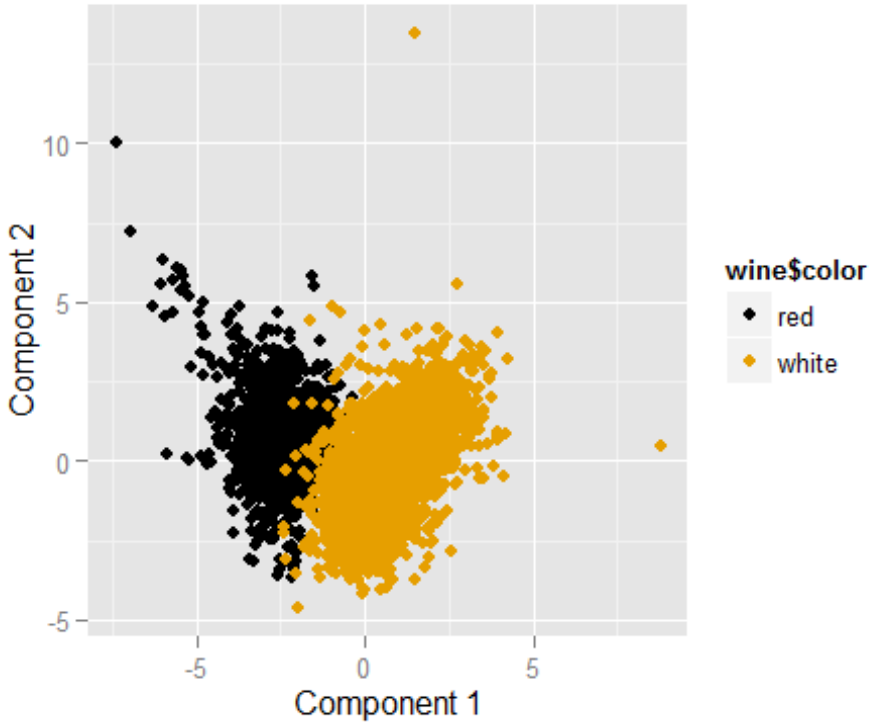
```
##                             PC7     PC8     PC9   PC10    PC11
## Standard deviation     0.72330 0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04756 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.90009 0.94568 0.97632 0.9970 1.00000
```

We can see that the first 3 PCs contribute to almost 50% of the variance just by themselves, so plotting on just those variables should provide a good dimensional visualize of the data.

Further, by plotting the variance over each PC, we can see that the amount of variance described by each PC quickly tapers off after the 8th PC, and from the above table, that corresponds to around 95% of the variance.

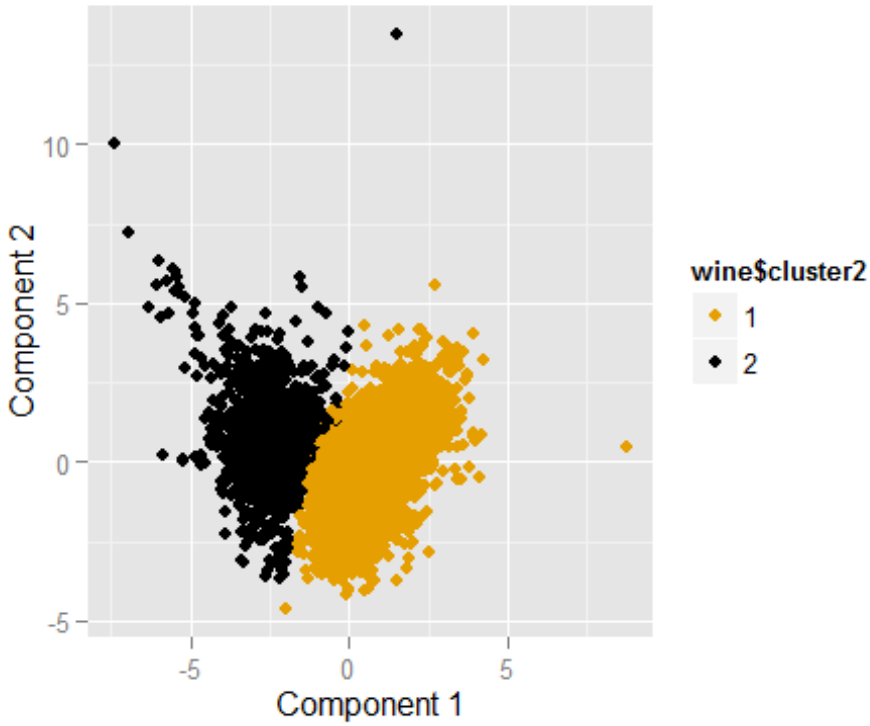


Given what we have oberserved so far, we can now plot of our wine data points expressed in terms of PC1 and PC2. For added visual effect, we can also color code each data point in terms of the wine's color, red or white. We see that the PCA analysis does a good job correctly grouping the two colors of wines with a lateral division seperating the majority of the red wines to the left, and the white wines to the right.



PCA is not the only dimensionality reduction technique we can use. Consider the use of clustering via kmeans to allocate a color of wine to each data point based on the datapoint's euchlidean distance from a cluster's center.

Here is the plot of the kmeans clustering using a number of clusters equal to 2 on the same principle components that we generated with PCA.

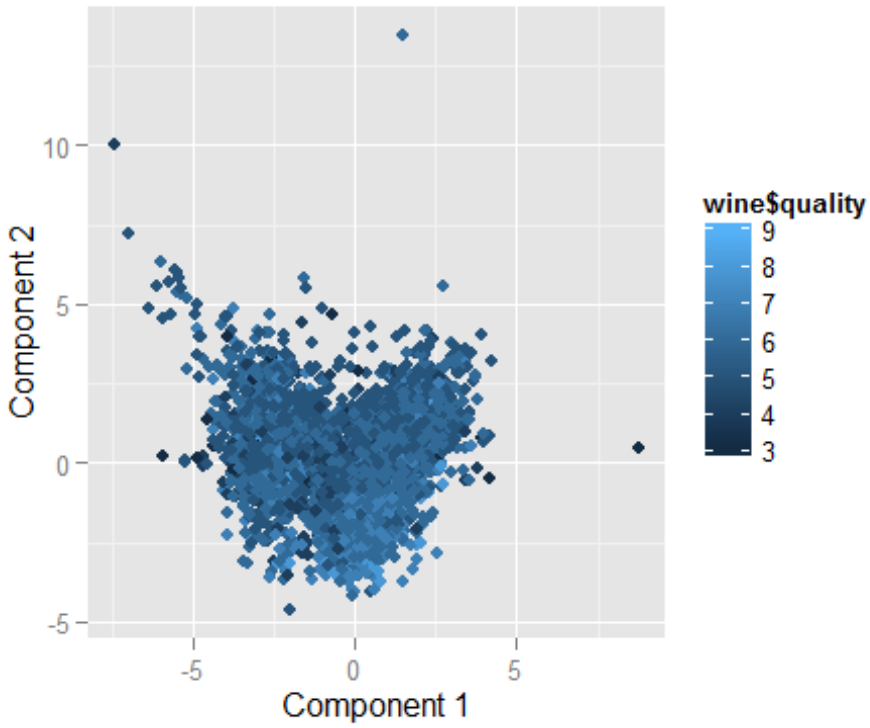


Just using 2 clusters appears to divide the two colors of wines just as well if not better than just with PCA. Cluster 1 correcly identifies 98.5% of the red wines, and Cluster 2 correctly identifies 98.6.

```
##          color
## cluster2  red white
##        1   24  4830
##        2 1575    68
```

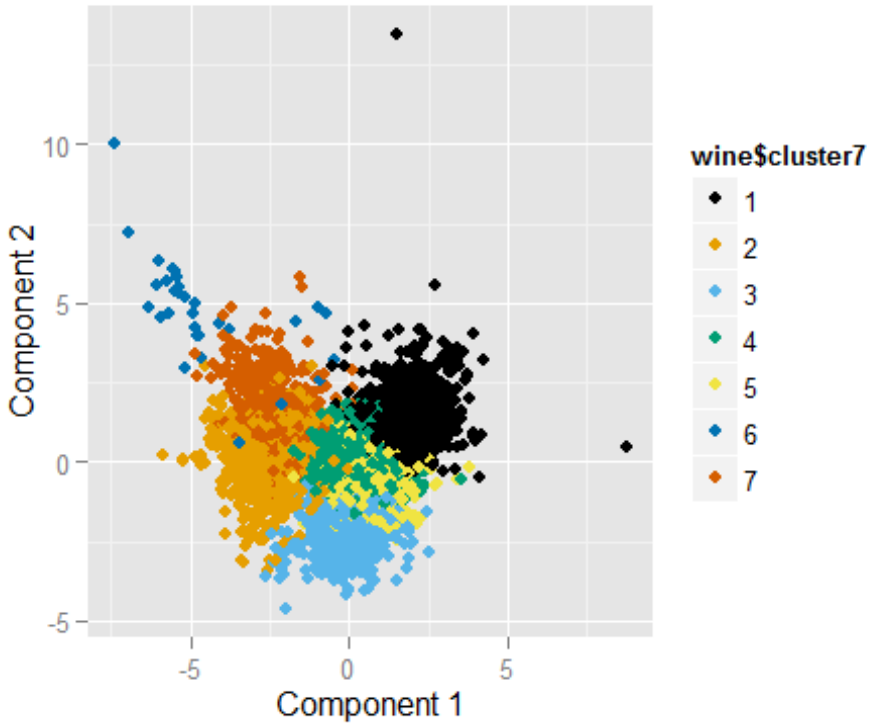
It is therefore a fair assessment to say that using kmeans is a stronger dimensionality reduction technique than pca as it makes it easier to visually distringuish between the colors of wines in our dataset. This is most easily seen by comparing the two groups and noting the crisp vertical division between the two clusters.

Shifting our attention over the quality descriptor, we can graph the qualities in our plot of reexpressed data in terms of the two principle components



Unlike with the color predictor, using PC1 and PC2 does not produce a visually seperate grouping of qualities and we can see each quality distributed across the plot of the data.

We can also perform kmeans to produce 7 clusters (one for each of the qualities awarded the wines in our dataset) and then compare these clusters to the qualities of the wines in our data.



Here we see a large amount of cross-over between the clusters. It is especially difficult to distinguish clusters 4 and 6 as they appear to be centered almost on top of one another.

Producing a table showing the distribution of qualities between our 7 clusters shows that a single quality are rarely associated with just one cluster and there are usually an assortment of qualities amongst each. For instance, the quality of 6 is the most numerous amongst the wines in our data set and is evenly distributed amongst many of the clusters.

```
##         cluster7
## quality   1   2   3   4   5   6   7
##       3   7   7   4   5   2   1   4
##       4  24  63  21  64  27   2  15
##       5 655 471  77 446 269  20 200
##       6 640 350 548 549 475   9 265
##       7 122  43 446 137 189   1 141
##       8  22   2  97  27  31   0  14
##       9   0   0   4   1   0   0   0
```

Market Segmentation

I first began by cleaning up the data. I did this by removing the columns that would not be relevant in terms of market segments, such as the name of the twitter user, and categories such as chatter, uncategorized or adult.

I then normalized the data by dividing each user's tweet by their total tweet count.

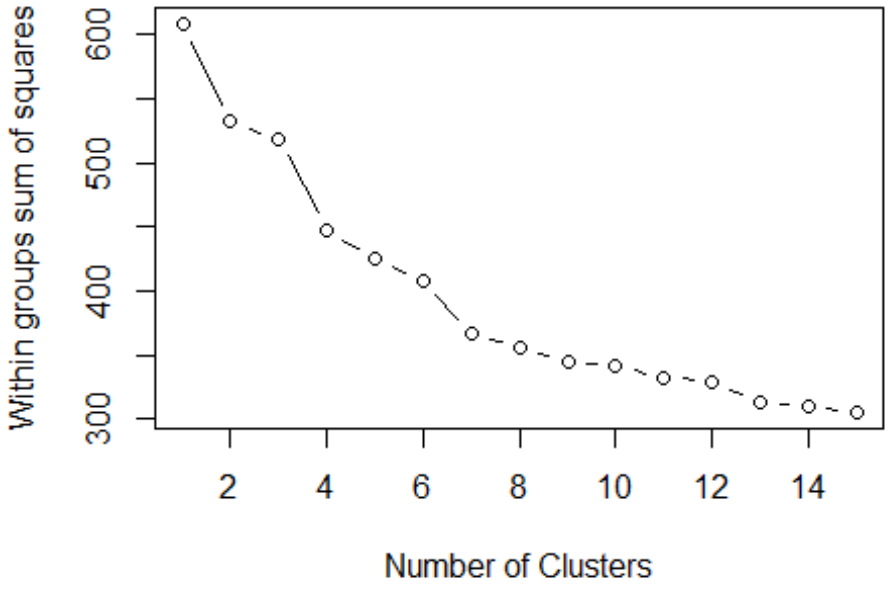
This provides us with each row representing one user and each column the percentage of their tweets that fall into a particular category.

```
##   current_events     travel photo_sharing    tv_film sports_fandom
## 1     0.00000000 0.03508772    0.03508772 0.01754386    0.01754386
## 2     0.11538462 0.07692308    0.03846154 0.03846154    0.15384615
## 3     0.07500000 0.10000000    0.07500000 0.12500000    0.00000000
## 4     0.25000000 0.10000000    0.10000000 0.05000000    0.00000000
## 5     0.08333333 0.00000000    0.25000000 0.00000000    0.00000000
## 6     0.14285714 0.07142857    0.25000000 0.03571429    0.03571429
##     politics       food     family home_and_garden      music  news
## 1 0.00000000 0.07017544 0.01754386      0.03508772 0.00000000 0.000
## 2 0.03846154 0.07692308 0.07692308      0.03846154 0.00000000 0.000
## 3 0.05000000 0.02500000 0.02500000      0.02500000 0.02500000 0.025
## 4 0.05000000 0.00000000 0.05000000      0.00000000 0.00000000 0.000
## 5 0.08333333 0.00000000 0.04166667      0.00000000 0.00000000 0.000
## 6 0.00000000 0.07142857 0.03571429      0.03571429 0.03571429 0.000
##   online_gaming   shopping health_nutrition college_uni sports_playing
## 1         0.000 0.01754386        0.2982456   0.0000000     0.03508772
## 2         0.000 0.00000000        0.0000000   0.0000000     0.03846154
## 3         0.000 0.05000000        0.0000000   0.0000000     0.00000000
## 4         0.000 0.00000000        0.0000000   0.0500000     0.00000000
## 5         0.125 0.08333333        0.0000000   0.1666667     0.00000000
## 6         0.000 0.17857143        0.0000000   0.0000000     0.00000000
##      cooking        eco  computers   business   outdoors     crafts
## 1 0.08771930 0.01754386 0.01754386 0.00000000 0.03508772 0.01754386
## 2 0.00000000 0.00000000 0.00000000 0.03846154 0.00000000 0.07692308
## 3 0.05000000 0.02500000 0.00000000 0.00000000 0.00000000 0.05000000
## 4 0.00000000 0.00000000 0.00000000 0.05000000 0.00000000 0.15000000
## 5 0.04166667 0.00000000 0.04166667 0.00000000 0.04166667 0.00000000
## 6 0.00000000 0.00000000 0.03571429 0.03571429 0.00000000 0.00000000
##   automotive art   religion beauty  parenting     dating    school
## 1 0.00000000 0.0 0.01754386  0.000 0.01754386 0.01754386 0.0000000
## 2 0.00000000 0.0 0.00000000  0.000 0.00000000 0.03846154 0.1538462
## 3 0.00000000 0.2 0.00000000  0.025 0.00000000 0.02500000 0.0000000
## 4 0.00000000 0.1 0.00000000  0.050 0.00000000 0.00000000 0.0000000
## 5 0.00000000 0.0 0.00000000  0.000 0.00000000 0.00000000 0.0000000
## 6 0.03571429 0.0 0.00000000  0.000 0.00000000 0.00000000 0.0000000
##   personal_fitness fashion small_business
## 1        0.1929825   0.000     0.00000000
## 2        0.0000000   0.000     0.00000000
```

```
## 3         0.0000000   0.025     0.00000000
## 4         0.0000000   0.000     0.00000000
## 5         0.0000000   0.000     0.04166667
## 6         0.0000000   0.000     0.00000000
```

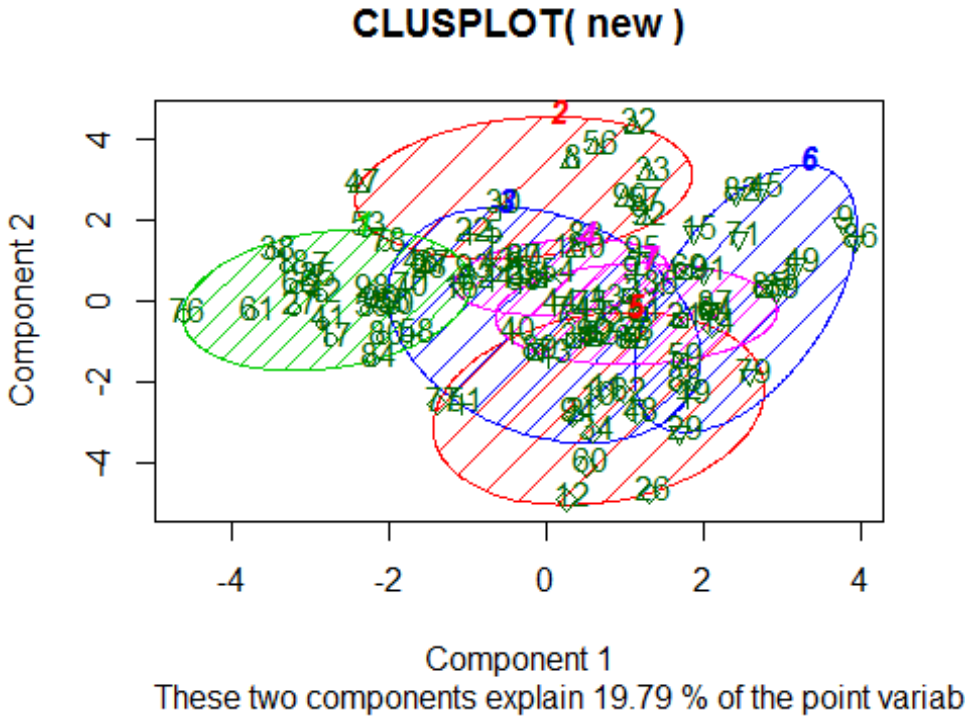
I used the process of k-means in order to cluster the data into market segments. K-means involves selecting a number of clusters (market segments) to group the data by. Then the distance between each data point (twitter user) and the center of each cluster is calculated. The data point is then assigned to the nearest cluster. This process is then repeated several hundred times in order to ensure that cluster centers are found that minimize the distance between the center and all data points associated with that cluster.

We can perform a quick analysis to determine the optimal number of clusters to use based on minimizing the distances between each data point and its cluster center.



The result is the above graph that shows we do not really improve how close the data points are to their cluster centers after the 7th cluster, so we will just divide the data into 7 markest segments.

It is difficult to visualize these clusters in just 2 dimensions, but if plot a hundred data points from our data, we can see how they roughly fall into each cluster. While it might appear as if some users are in overlapping clusters, this is again a product of the number of variables in the data set and the difficulty in visualizing the boundaries between each cluster.



We can now take a look at each of our clusters. I decided that the best way to help explain these clusters is to take the average value for each tweet category among all users in the cluster. Remember that we normalized the data already, so we are returning out of all the tweets of the users in the cluster, the percentage of those of a specific category. I then sorted these to show the top 4 categories.

```
## [1] "Business Man"

##   politics       news     travel automotive
## 0.16921263 0.11041501 0.10025193 0.05197428

## [1] "Mid-aged Woman"

##       cooking photo_sharing       fashion        beauty
##    0.20757293    0.10511058    0.10030379    0.06501491

## [1] "Young Professional"
```

```
##  photo_sharing       shopping current_events         travel
##     0.20622773     0.11059825     0.07399775     0.04274985

## [1] "College Male"

##    college_uni  online_gaming  photo_sharing sports_playing
##     0.20377040     0.19033973     0.05613751     0.04697395

## [1] "Father"

## sports_fandom      religion          food     parenting
##    0.12249805    0.09644103    0.08828672    0.07385373

## [1] "Health Enthusiast"

## health_nutrition personal_fitness          cooking    photo_sharing
##       0.23464577       0.11732524       0.05986224       0.04801825

## [1] "Post-College Arts & science Major"

##        tv_film current_events         travel            art
##     0.08894528     0.08557457     0.06205597     0.05434229
```

These market segments have been labled based on common interests among the users as seen by the categories of their tweets. Please note that photo sharing is common among several of the segments and has been used to guess the age group of the segment as younger users are more likely to share their photos than older twitter users. Based on your company's branding, I would recommend the 'Health Enthusiast' market segment be targeted.