STA 380 Homework 1

Charles Z. Aldrich

Friday, August 7th 2015

Problem 1: Exploratory Analysis

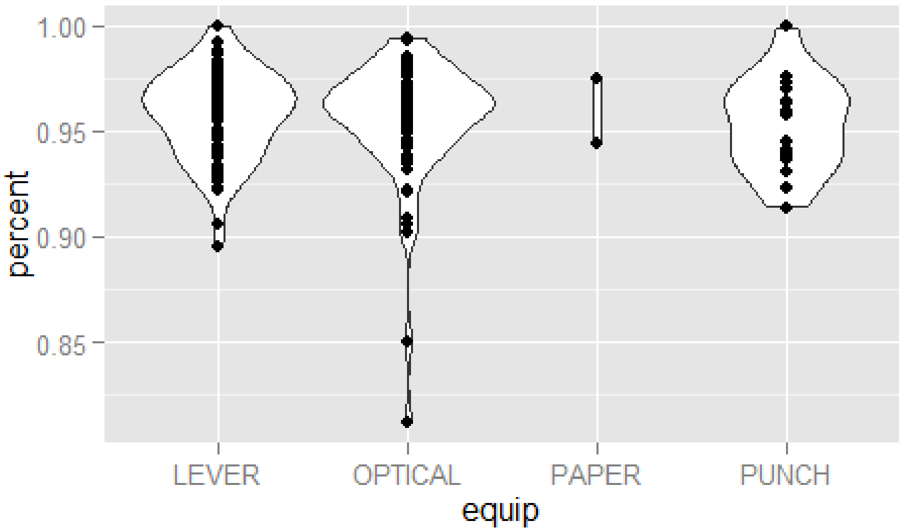
Issue 1: The impact of equipment type on the undercount rate of votes cast

In order to properly analyze each equipment type and its associated level of undercounting, I elected to use violin plots generated by the package ggplot2. Violin plots provide several advantages over other graphical representations. First, they provide good visual representation of the spread for each category of data. For instance, from the plot below it can be easily found that optical equipment has a much larger spread in terms of undercount rates relative to all other equipment types. Although the spread provides some insights into the overal performance of each equipment type, the ability to analyze the density of each undercount rate for all equipment types helps in determining if any particular equipment type underpeforms relative to the others.

The width of each violin plot at any point on the graph below shows the relative density of that percentage level in terms of all recorded percents for that equipment type. Additionally, the widths can be compared across equipment types as they are normalized by relative density. When looking at the plot below in these terms, it can be found that the highest density of optical equipment is below that of lever equipment on top of the fact that optical equipment has the largest range as noted above. This suggests that the use of optical equipment has a higher likelyhood of leading to higher undercount rates.

The other interesting piece of about densities to point out is that the punch equipment type displays a relatively constant density across all observations. This suggests that the performance of punch equipment is hard to predict given there is no outcome more common than the others. Add this to the fact that the undercount spread is anywhere from 0% to 8%. For these two reasons, it would appear that punch equipment is also susceptible to a greater potential for undercounting.

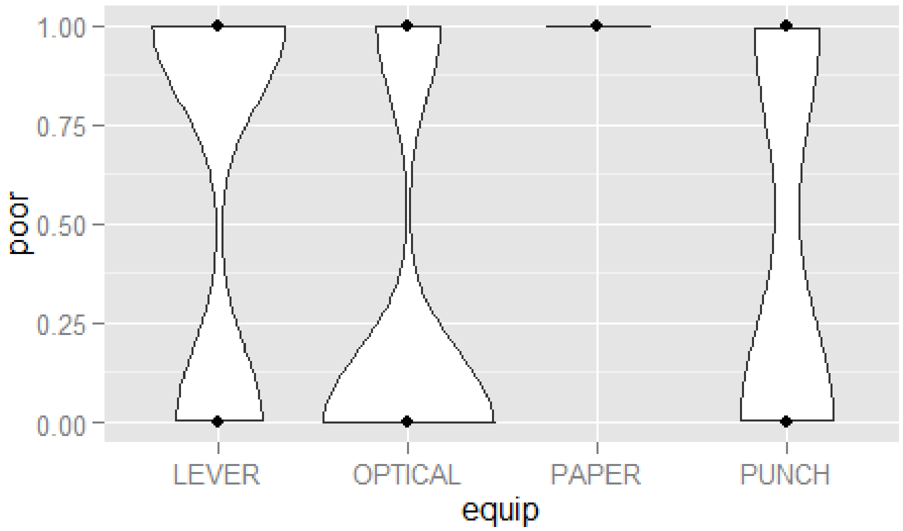
Therefore, after analyzing the violin plot created from the data set, I would suggest that the use of optical equipment and punch equipment lead to higher rates of undercounting thereby putting all counties that use they equipment types at a disadvantage.



Issue 2: The impact of equipment types on poor and minority communities

Based on the findings from Issue #1 I wanted to analyze the breakdown of poor and minority communities across all equipment types. If it can be shown that there is a higher percentage of poor and/or minority communities using the two "defective" equipment types found in Issue #1 then the data would suggest that these communities are at a disadvantage. In order to analyze the breakdown of poor and minority communities using each equipment type I again used the violin plots.

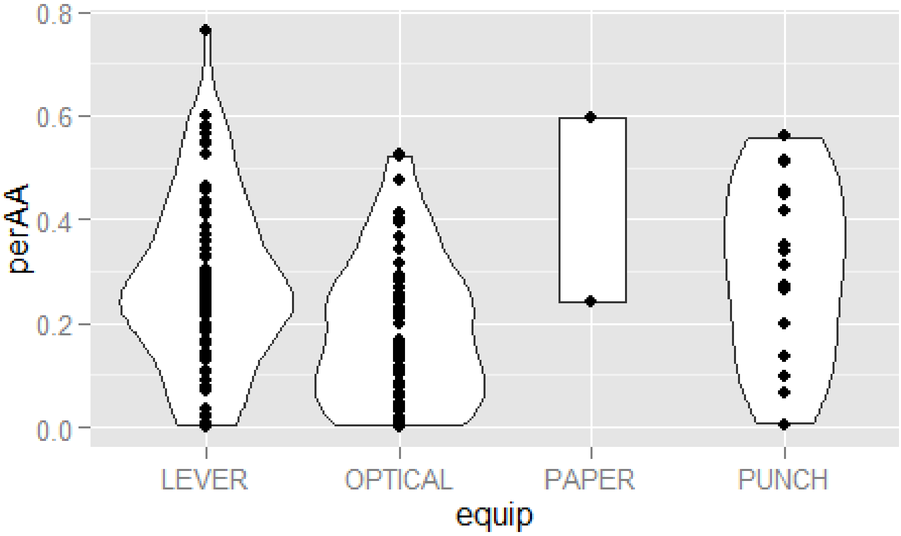
The plot below shows the density of usage for each type of equipment across poor and non-poor communities. From the violin plot it is relatively easy to determine that a larger proportion of poor communities use lever equipment. A smaller proportion use optical or punch equipment. It is important to see that the highest density of optical and punch equipment usage comes from the non poor communities. Therefore, given the two "defective" equipment types found in Issue 1, it does not appear that poor communities are at a larger disadvantage of vote undercounting. Conversly, it is the non-poor communities that are at a larger disadvantage.



Now, in terms of minority communities the charateristics of each violin plot are slightly different than those for poor communities due to the charateristics of the variable used. Unlike the variable poor which was binary, perAA measures the percentage of African-Americans within a given county. The lower the perAA number, the lower the perception of a a county being categorized as a minority county.

First, when looking at the density distribution for optical equipment in the plot below it is important to see that the greatest density occurs near a perAA level of 0.05. A second, smaller, density occurs at a level of roughly 0.25. Each of these is below the largest density of lever equipment and paper equipment. Therefore, in terms of the optical equipment, it appears that minority communities were not placed at a disadvantge because most of the optical equipment was used in non-minority communities. Thus, similar to the poor community analysis above, non minority communities were placed at a greate disadvange than minority communities.

Second, for the punch equipment, the density of usage is relatively even across all levels of perAA represented. Given an even distribution, there is no discernable disadvantage given to minority or non-minority communities.



Problem 2: Bootstrapping

Risk/Return charateristics of each major asset class

In order to show the level of volatility in each of the asset classes used in this portfolio, I chose to use the quantmod library and graph the ticker price for each asset over the past 5 years.

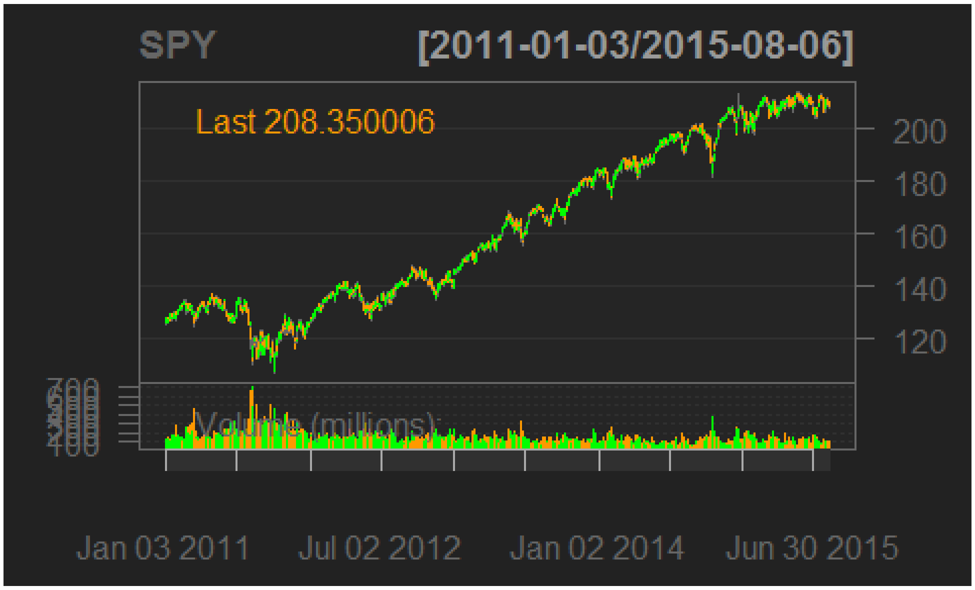
```
## Loading required package: timeDate
## Loading required package: timeSeries

## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following object is masked from 'package:timeSeries':
##
##     time<-
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
##
## Attaching package: 'xts'
##
## The following objects are masked from 'package:dplyr':
##
##     first, last
##
```

```
## Loading required package: TTR
## Version 0.4-0 included new data defaults. See ?getSymbols.
```

US Domestic Equity (SPY):

Over the last 5 years it is pretty easy to see that the SPY has relatively low risk and high return. Since 2011 the price of the SPY has risen nearly 70% with no long term negative trend. Therefore, it is a strong asset to hold in a high return, long-term risk adverse portfolio.

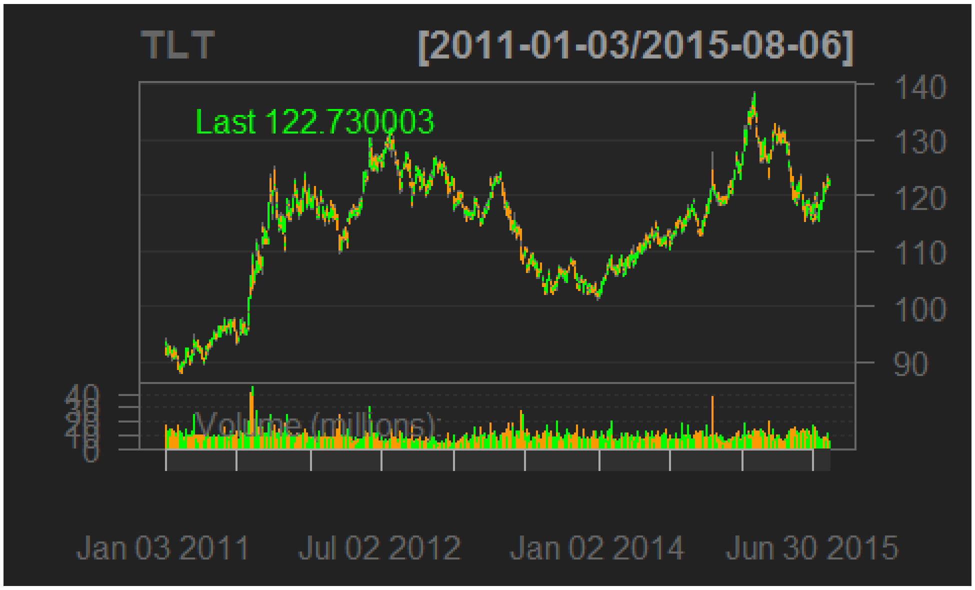
```
##     As of 0.4-0, 'getSymbols' uses env=parent.frame() and
##  auto.assign=TRUE by default.
##
##  This  behavior  will be  phased out in 0.5-0  when the call  will
##  default to use auto.assign=FALSE. getOption("getSymbols.env") and
##  getOptions("getSymbols.auto.assign") are now checked for alternate
defaults
##
##  This message is shown once per session and may be disabled by setting
##  options("getSymbols.warning4.0"=FALSE). See ?getSymbols for more details.

## [1] "SPY"
```



US Treasury Bonds (TLT):

Over the lat 5 years the volatility on TLT has been much larger than that of the SPY. Rather than a steady upward trend, TLT has seen rolling highs and lows. For exmaple, leading up to 2013 the stock was making steady positive progress. Then from 2013 to 2014 the stock retreated downward before finally rising again from 2014 to 2015. Finally, in 2015 the stock has moved back down. These fluctations provide great opportunities for investors to have large returns (if they time the market right) but also expose the portfolio to a sizable level of risk.

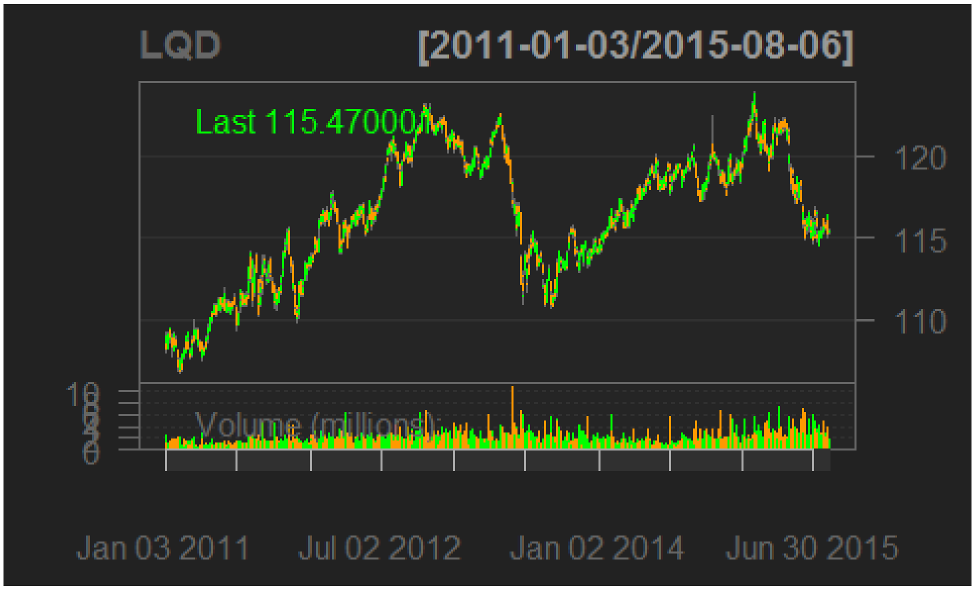
```
## [1] "TLT"
```



Investment-Grade Corporate Bonds (LQD):

Over the past 5 years LQD has moved relatively parallel to TLT. From 2011 to mid 2013 the prices of TLT rose slighly less than 10%. Then prices dropped ack to levels consistant with 2012 before making a strong upward movement starting in 2014. Now, although the volatitlity of the chart would suggest that LQD is highly volitile, it is important to understand the magnitude of the fluctuations. Over the last 5 years, the price of LQD has bounced around between 110 and 125. This is a small range in comparison to the SPY which has a range of 120. Therefore, LQD provides investors with an ability to capitalize on price fluctiations thereby increasing returns, however the size of these fluctuations are relatively small and therefore reduce a portfolios overall risk.

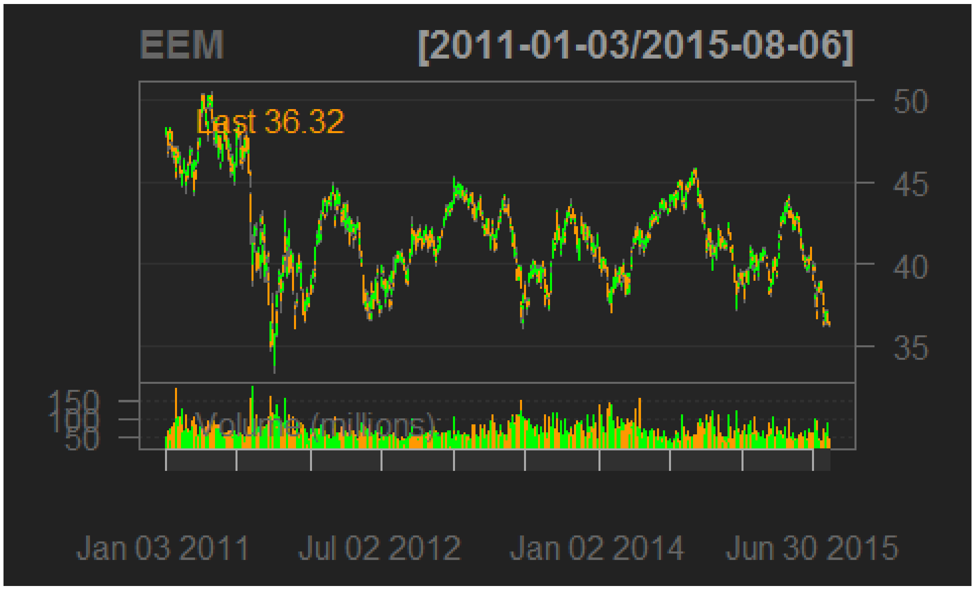
```
## [1] "LQD"
```



Emerging-Market Equities (EEM):

EEM is a highly volitile investment. This can be seen in the dramatic ups and downs in the ticker price over the last 5 years. There is no steady trend in either direction over the given time period. Rather the price has bounced around centered at $40 with 12.5% fluctuations occuraing regularly. EEM has the potential for large returns if the market is timed right, however this potential gain is offset by the high level of volatility and associated risk that it emposes on a portfolio.

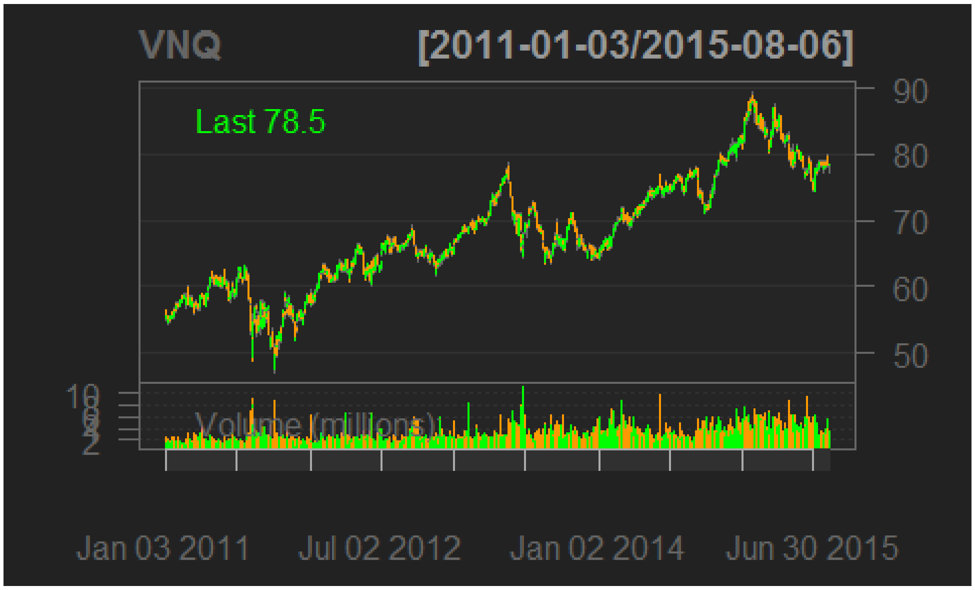
```
## [1] "EEM"
```



Real Estate (VNQ):

Similar to the SPY, VNQ has shown a positive trend over the last 5 years. from 2011 to 2015 the price has risen from a low of 50 to a high of nearly 90. For a long term investor, the inclusion of VNQ in a portfolio has both high reward and relatively low risk. However, for more short term investors, VNQ does has a relatively large degree of volatility within a given year. Unlike the SPY, VNQ experiences a lot of movement as it tracks upwards. Therefore, the level of return/risk for VNQ is dependent upon the investment style of the portfolio owner.

```
## [1] "VNQ"
```



Outline of the weighting for "Safe" and "Aggressive" porfolios:

Safe:

In terms of a safe porfolio, the goal is to have holdings that provide steady returns but do not subject the entire portfolio to large fluctuations in returns with the potential for large losses. Therefore, based on the 5 major asset classes that are available I would recommend the following weightings: 35% LQD, 35% TLT, 30% SPY. By assigning these weightings, the portfolio gets the volatility protection associated with owning bonds but the upward potential in larger returns oassociated with owning the entire stock market. At first glance, the prices of bonds has been all over the place. Even though this is true, as noted in the description of return/risk above, given the fluctuations are small in percentage terms to the total price of the asset, owning bonds is safe. The potential of large losses on a portfolio is dramatically reduced by having 70% of the portfolio value in bonds.

Aggressive:

In an aggressive portfolio we look to include assets that have a larger degree of volatility. The reson is, with greater volatility comes the chance to be on the upswing and see large percentage gains in a short period of time. Therefore, based on the 5 major asset classes that are available I would recommend the following weightings: 35% VNQ, 40% SPY, 25% EEM. Having a large position in Real Estate and Emerging Markets creates a lot of opportunity for large returns. When you look at the stock price of EEM over the past 5 years, if you were able to time the market right by buying low and selling high you have the potential of making a 30% return. Returns of this size are massive in comparison to most mutual funds or individual stocks. Therefore, by owning risker assets the chances of large gains increases. However, this position does take on a considerable amout of risk and is therefore very aggressive. The overall hope is that the size of each large gain will be enough to offset any major loss in value.

4 Week Value At Risk estimation (5% level):

Equal Split portfolio (20% SPY, 20% TLT, 20% LQD, 20% EEM, 20% VNQ)

```
##       5%
## -3340.14
```

Safe Portfolio: (35% LQD, 35% TLT, 30% SPY)

```
##        5%
## -1789.488
```

Aggressive Portfolio: (35% VNQ, 40% SPY, 25% EEM)

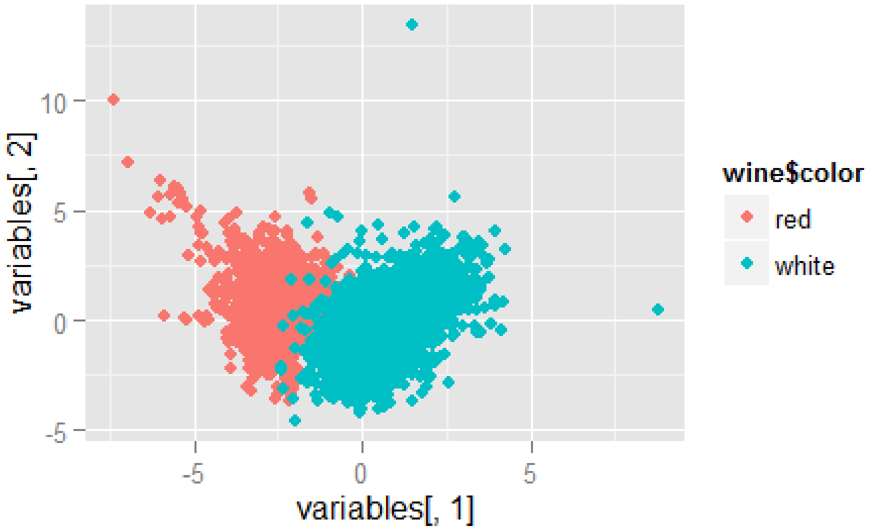
```
##        5%
## -5396.544
```

In order to obtain reproducability of portfolio outcomes I have set.seed(100). With set.seed(100) the following 5% value at risk were found: Even split = -$3,340.14, Safe = -$1,789.488, and Aggressive = -$5,396.544. The 5% value at risk provides a monetary amount of an investors portfolio that is being risked by choosing a particular portfolio asset weighting. As should be expected, the quantity of the portfolio at risk is smallest for the Safe Portfolio and largest for the Aggressive Portfolio. This illustrates the potential loss associated with the amount of risk in each portfolio. Additionally, the 5% value at risk of the evenly weighted portfolio falls within the Safe and Aggressive options. This shows the "average" amount at risk that an investor would face if they did not adjust the weighting to suit their strategy. From a decision standpoint, an investor should look at the 5% value at risk number for each portfolio option and choose the one that they are most comfortable with losing. If an investor is comfortable with losing $5,397 for the chance at seeing larger gains then the aggressive strategy suits them. However, if an investor is only comfortable risking $2,000 of their portfolio then they should look to use the Safe strategy as it is the only portfolio that does not risk more than the investor is comfortable with.

Problem 3: Clustering and PCA:

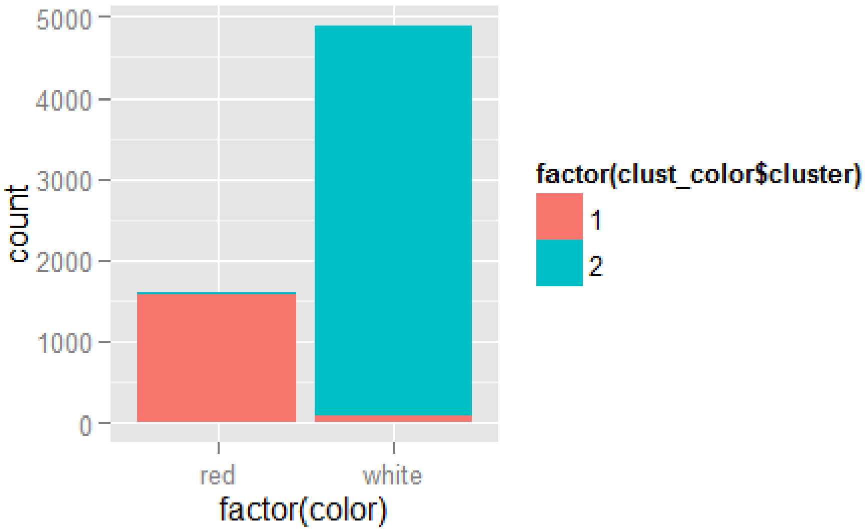
Determining Wine Color:

Principle Component Analysis:



The chart output from PCA is a reprenstation of the first priciple component relative to the second principle component with red dots being red wines and blue dots being white wines. From the scatter plot, it is fairly easy to discern two main groups. The biggest problem however is the small section of overlap between the two clusters. Within this overlap it is hard to tell just how many data points are within each color cluster. Despite the overlapped area, PCA does a relatively good job of clustering red and white wine based on the 11 chemical variables provided in the dataset.

K Means:



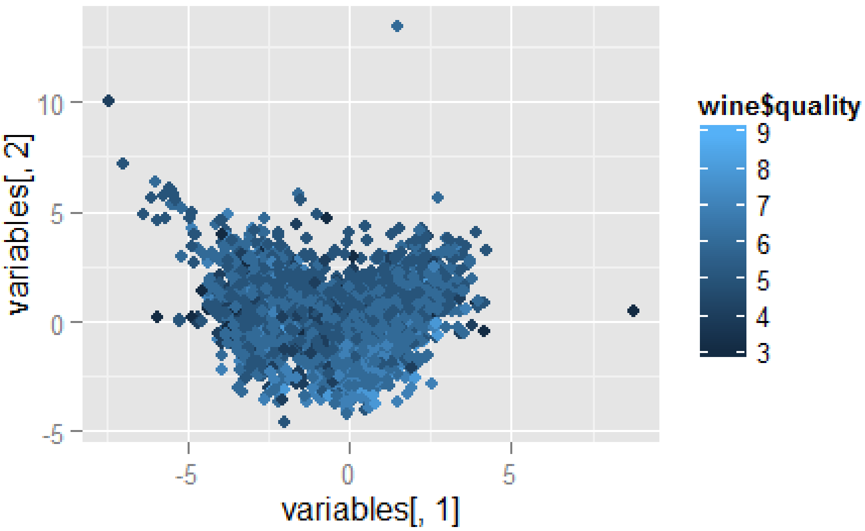
The bar chart generated from K Means displays the type of wine, red and white, and a count of their occurance based on clustering kmeans generated when using 2 clusters. The coloring shows the occurances of color prediction relative to the actual color of each wine. As can be seen in the bar charts, a small sliver of red wines were clustered as white and vise versa. The use of an area bar chart allows for an easy comparison of the number of misclassifications of both types. In general, using 2 centers, kmeans did a good job classifying wine colors using the 11 chemical variables provided in the dataset.

Wine Color Analysis:

Based on the graphical outputs from the two dimension reduction techniques analyzed above, it is pretty clear that kmeans did a better job of clustering red and white wine according to 11 chemical variables. The reason for this is that the amount of misclassification is less across both types of wine. Additionally, kmeans has an output that is far easier to interpret. It is easy for anyone to look at the bar graph and understand that it is showing a count of red and white wine categorized by the count of accurate classifications. The output from PCA does not have an easy interpretation due in large part to the area of overlap. Given the overlap it is hard to see both the red and blue points in that small area. Additionally, any reds that were misclassified are hidden under the blue points representing white wine. Plus it is hard to understand what the first and second principle components mean in the context of the chart. Therefore, based on the two techniques used, I would recommend kmeans for two reasons. First, it does a very good job of classifying red and white wines in the dataset. Second, and in some instances more importantly, it provides easy interpretability.

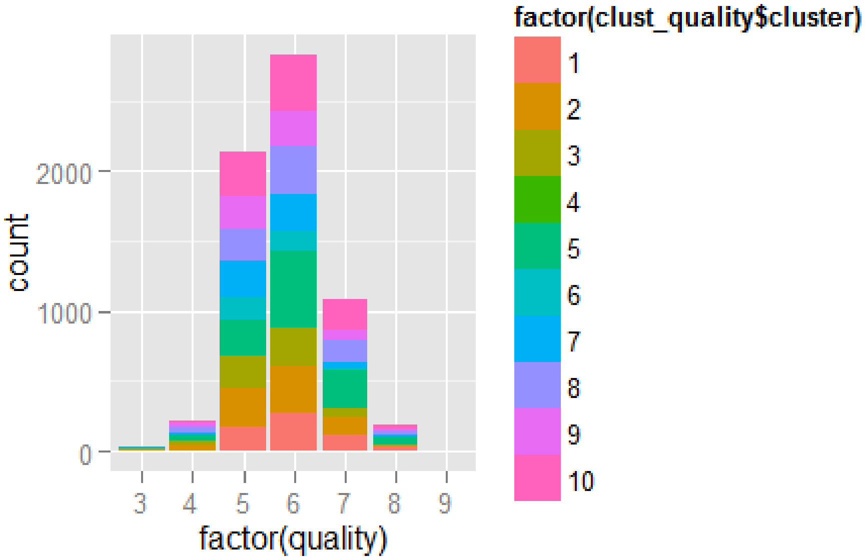
Determining Wine Quality:

Principle Component Analysis:



The chart output from PCA for quality of wines is trying to cluster the quality of wine based on the 11 chemicle variables from the dataset. The problem is, with upwards of 10 bins for quality rating, the color scheming of the plot becomes hard to analyze. It is hard to see where colors begin to cluster together, indicating a particular quality of wine. In general, the plot shows a wide variety of wine qualities spread all over the place in terms of the first and second principle components.

K Means:



The output chart from kmeans is trying to show the clustering of wine qualities using 10 different clusters. 10 clusters were used given the scale of quality rating was from 1-10. From the bar chart it is relatively easy to see that overall most of the wines were rated a 5 or 6. The biggest problem is that across the board, for all rating centers, there is a fairly even representation of each cluster indicating that kmeans did not do a very good job of clustering quality ratings in terms of the 11 chemical variables.

Wine Quality Analysis:

Based on the outputs from PCA and K-Means in terms of clustering for quality it is easy to see that neither tool was able to do a good job of clustering quality relative to the 11 chemical variables provided in the data set. This outcome is to be expected though. Quality is an ambiguous varible that depends on the end user. No two individuals are going to have an identical outlook on quality. Therefore, using the chemical components of a wine to predict quality is very difficult and potentially impossible. This assertion is supported by the outputs of PCA and K-Means. Neither tool was able to accurately cluster quality based on 11 chemical variables.

Overall Analysis:

After running PCA and K-means on the wine dataset it is clear that both dimensionality reduction tools performed well in trying to classify wine color. This would suggest that the color of a wine is dependent in large part on the chemical makeup (really the type of grapes used) of the wine. However, when trying to cluster quality based on the same 11 chemical variables, neither tool was able to do so. This poor performance is not unexpected though given the notion of quality is dependent upon the individual and not necessarily the chemical makeup. Therefore, dimension reduction tools work in regards to wine color but not wine quality.

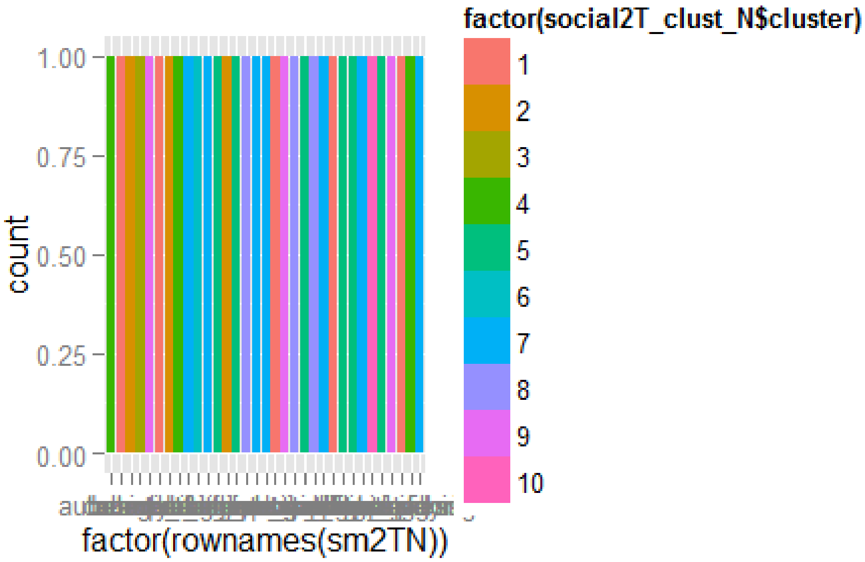
Problem 4: Market Segmentation

Market Segmentation in general can be a challenging task. In terms of the Twitter dataset provided looking at the theme of tweets for a selection of NutrientH2O followers, the task becomes no easier. One of the biggest challenges is the subjectiveness in how tweets are categorized manually. Additionally, some tweets do not really fall within any defined category, such as the tweets grouped in the chatter catergory. In order to generate a better result from the dataset, the following catergories were removed as it was determined that they had little relevance on the overall segmentation goal; chatter, adult, and spam. The remaining 33 categories provided a wide variety of tweet topics that followers referenced regularly.

The next big challenge for interpretation came from finding a way to properly interpret an individuals tweets for a certain category relative to their total number of tweets. For instance, an individual who tweets a lot and has one tweet about food should potentially be distanced farther from food than an individual who has only tweeted 5 times, one of which being about food. In order to account for the frequency of tweets in a given category by all users, the dataset was normalized. Normalization allows the analysis to look at the interaction of users in a particular catergory in constant terms across all users and all categories.

After the various cleaning steps taken to ensure the data provided an accurate representation of all followers within the dataset, a variety of clustering tools were used in order to try and find the best segmentation based on category charateristics. Tools used included, K-Means, K-Means ++, Hierarchical Clustering, and Principle Component Analysis. In order to accurately evaluate the results from each tool, two criteria were looked at. First, doe the resulting clusters make sense in terms of what followers would likely tweet about. For example, the potential for a common grouping of business, home_garden, and beauty may not make sense as a generalized cluster. Second, the output from the tool needed to be interpretable. After evaluating on these two criteria it was determined that kmeans provided the best output of possible market segmentation based on the data collected by NutrientH2O.

Below is a bar chart that provides a colored representaiton of all 33 categories used, subdivided across 10 different clusters. 10 clusters was chosen in the end because it provided the greatest number of clusters that resulted in catergory variables which made intuitive sense together. More clusters led to a smaller number of catergories in each cluster which resulted in less segmentation explanatory power. Fewer cluster led to market segments that were for to general and would be hard to accurately target.
The table at the end out the output shows the number of categories included in each of the 10 clusters.



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

From the 10 clusters created there are a few that standout a key market segments that should be targeted by the company. The choice of these clusters is due to the fact that the catergories inlcuded within the cluster make intuitive sense in terms of subjects in which one individual would likely talk about.

The first cluster that stands out is Cluster #5

```
## sports_fandom          food        family      religion     parenting 
##             6             8             9            26            28 
##        school 
##            30
```

Based on the categories included in this cluster it is likely that these types of tweets can be generalized to individuals such as stay at home moms. The reason for this generalization is that family, food, parenting, and school all appear to be associated with stay at home moms. In terms of the business, the idea of segmenting out moms and looking at targeting them for products of NutrientH2O makes sense given this segment is likely going to be the one going to the store and making food/beverage purchases the majority of the time in a given household. Therefore, targeting this market segment is key.

The second cluster that stands out is Cluster #1

```
##     travel   politics       news  computers automotive 
##          2          7         12         20         24
```

From the categories included in this cluster it can be generalized that followers in this category are likely young professionals. Young professionals are likely to be traveling frequently for work, actively engaged in the political landscape/current news, and in constant contact/up to date with the latest computers and automotibles. Given the nature of this cluster, it is likely these individuals are going to be constantly on the move but yet connected the entire time. Therefore, product placement is going to be far different than from Cluster #5. Additionally, marketing streams must be placed where the consumer is. For Cluster #1 this would mean through banner ads on the internet, such places as commercials on news and polticial channels, and potentially as inserts in newspapers. Knowing the target market segment has a huge impact on the marketing campaigns of the company.

The third cluster that stands out is Cluster #8

```
## health_nutrition         outdoors personal_fitness 
##               15               22               31
```

Cluster 8, based on the categories included, is made up of health concious, nature oriented individuals. These are likely to be the individuals that spend a good deal of their free time "out doing" rather than at home watching television or reading a book. Given this nature, the ways of targeting such a segment are going to be different than that of Cluster #5 and Cluster #1. For instance, product placement for this segment would likely include fitnes centers and sporting goods stores. Additionally, running television advertisements on channels such as ESPN or the Outdoor Channel will likely grab the attention of this segment more so than on news/politics channels for Cluster #1. Again, marketing and product placement for this cluster is going to far different from that of the two other clusters outlined above.

The fourth cluster that stands out is Cluster #9

```
##  online_gaming    college_uni sports_playing
##             13             16             17
```

The final cluster that made stroing intuitive sense was cluster #9. The categories included in this cluster would suggest that the NutrientH2O followers are college students. This is because the cluster explicitly includes colleges along with video games and playing sports. It is likely that this individual is a male college student. Knowing this, similar to the other cluster outlined above, allows NutrientH2O to do a better job with product placement in an attemtp to target this particular segment. It is fairly easy to predict where this type of segment individual is going to be.

Therefore, from the K-Means analysis run on the dataset provided there are four clear marketing segements that NutrientH2O should look at targeting. Each of these segments is unique from the rest and shows fairly strong descriptive charateristics. If the company can work to actively target these segments then it is likely that they will see positive returns both in top line revenues as well as a growing customer base due to free marketing generated by simple customer product usage. Now, it is likely that there are other potential market segements that could be found from the data provided in connection with additional follower information. For example, if the gender and/or age of the followers were known, then there is likley going to be more explanatory power for some of the categories. Even so, as noted above, the data provided gives the company relatively clear insights into four diverse market segments that the company can now attempt to accurately target.