Exercise 1 - Shreyas K.S.

Introduction

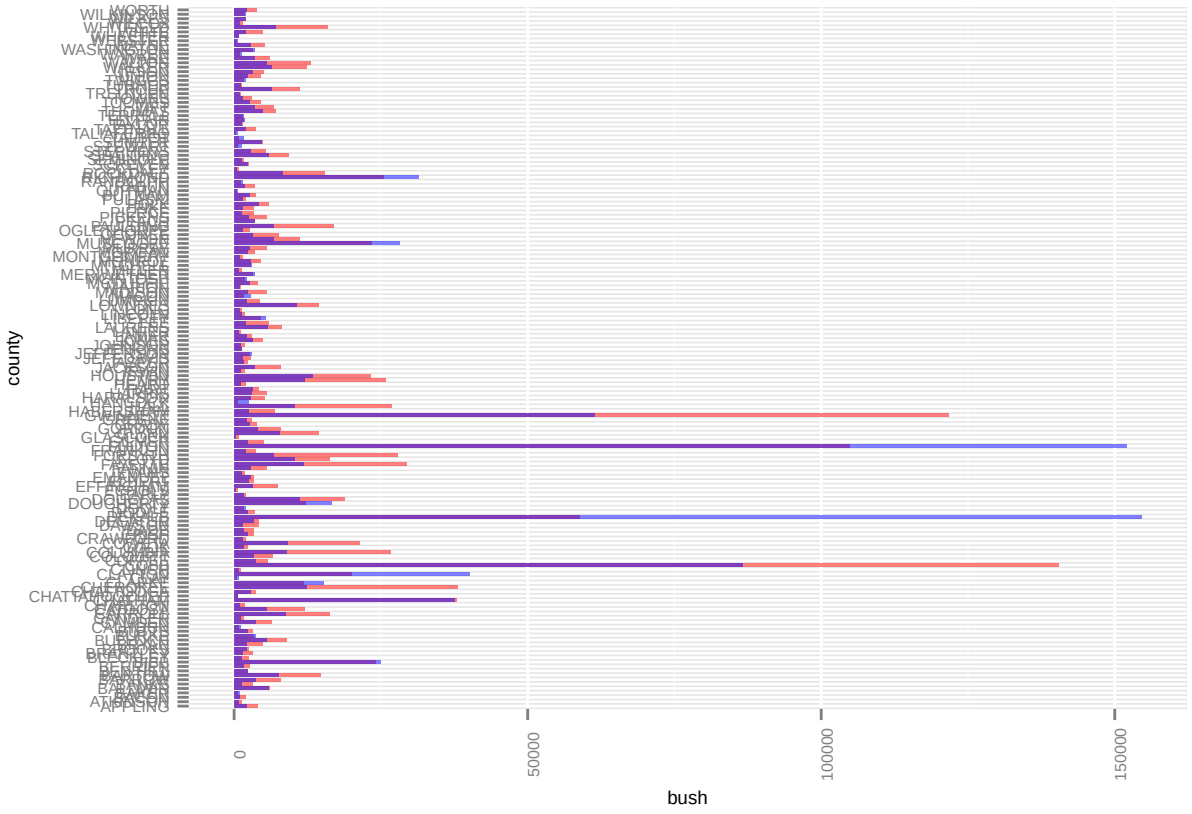
The following report contains 4 sections including an exploratory analysis of 2000 election results, bootstrapping in a financial context using ETFs, unsupervised learning on data on wines, and market segmentation.

Before starting on each of the individual sections, as a best practice, I loaded all required libraries and set a seed at the top of the R code.

```
my_seed = 59058
```

Exploratory Analysis

The following bars show the votes received by Bush (in red) and votes received by Gore (in blue) by county.

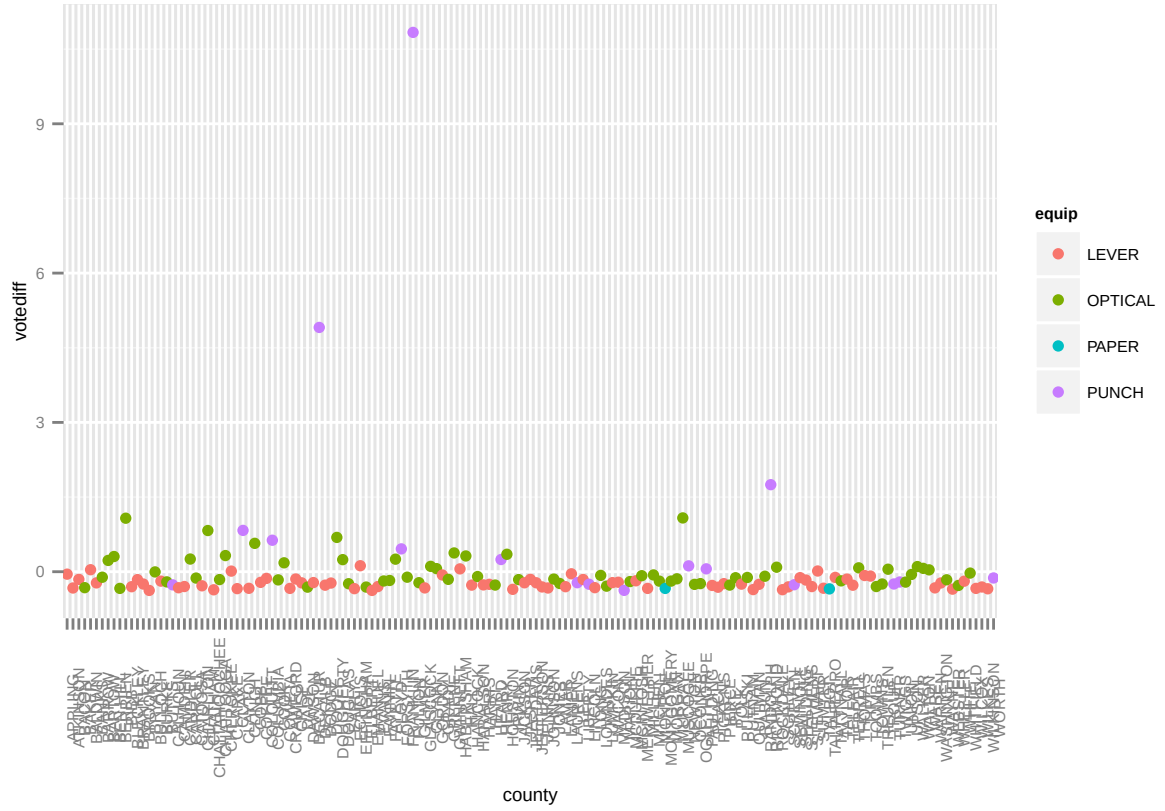


It is clear that Bush won in more counties than Gore, since the red bars overshadow blue bars in most cases. Other than the county of Dekalb, where there is a clear anomaly and Gore dominated Bush.

A new variable is created to measure the difference in votes and ballots:

```
georgia$votediff = georgia$ballots - georgia$votes
georgia$votediff = scale(georgia$votediff)
```

The variable needs to be scaled in order to adjust for differences in population across counties. For example, a `votediff` of 20 in a county with population of 1000 is more drastic than in a county with population of 200,000. Looking at the scaled difference in votes and ballots by state, we get the following:

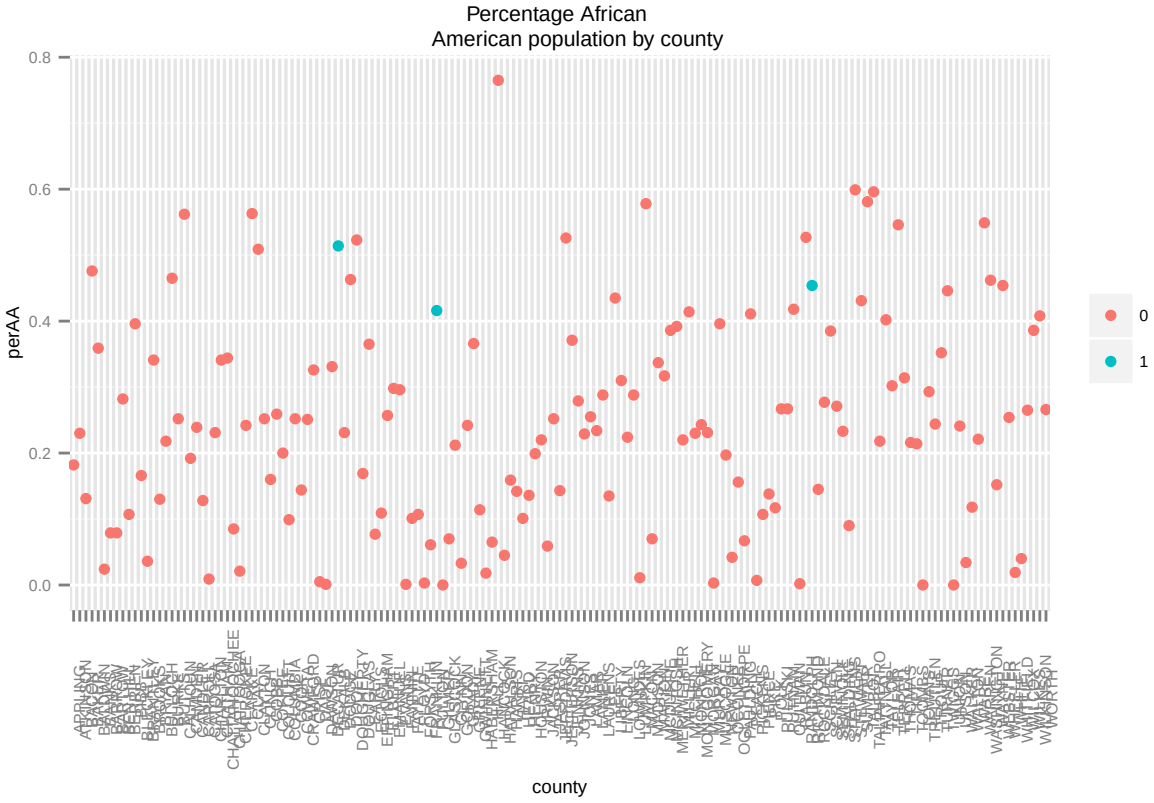


The following are factors which would affect voting in any location, regardless of the failure or success of equipment and voting mechanisms in any county. Hence, we can take them to be systematic risk in this context:

- voting for multiple candidates
- not voting for any candidate
- vote disqualified for other reasons

Lets analyze the outliers closely. In all the counties where there is a huge discrepancy (more ballots than votes recorded), punch ballots have been used. This is the clearest indicator of fraud in the given data. The normalized difference between ballots cast and votes recorded is more than 2 standard deviations from the mean.

Lets look at some metrics for these counties:



The 3 'suspect' counties have a relatively high proportion of African Americans compared to the rest of Georgia, as seen from the blue points.

```
##     County Poor
## 1   DEKALB    0
## 2   FULTON    0
## 3 RICHMOND    0
```

```
##     County Urban
## 1   DEKALB     1
## 2   FULTON     1
## 3 RICHMOND     1
```

```
##     County Bush Votes Gore Votes Difference
## 1   DEKALB       58807     154509      95702
## 2   FULTON      104870     152039      47169
## 3 RICHMOND       25485      31413       5928
```

Finally, we can also see that voters from these counties are not poor, are in urban areas, and predominantly voted for Gore over Bush (when the votes were recorded). We don't need to worry much about poor populations being affected, but we cannot conclude from this data whether areas which support Gore have been under represented due to the suspected fraud, or if Bush's dominance has been understated due to fraud.

Bootstrapping

The Even Split

For the case of the even split, the 5 specified ETFs are used:

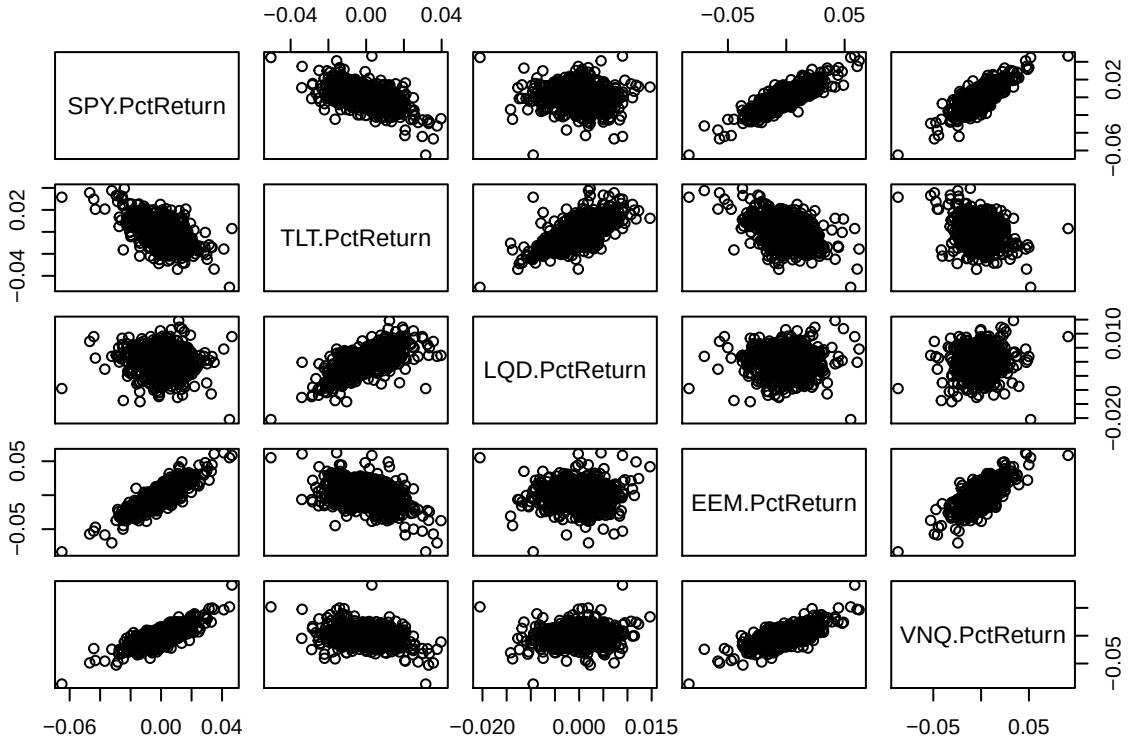
- US domestic equities (SPY)
- US Treasury bonds (TLT)
- Investment-grade corporate bonds (LQD)
- Emerging-market equities (EEM)
- Real estate (VNQ)

The mean percentage return of each ETF is as follows:

```
## SPY.PctReturn TLT.PctReturn LQD.PctReturn EEM.PctReturn VNQ.PctReturn
##  0.0006867895  0.0003101899  0.0002102527  0.0001528394  0.0006283819
```

We can see that on average, SPY returns about 0.068%, and other stocks return less.

The pairwise correlations for this portfolio is as follows:



It is clear from the pairwise correlations within the portfolio there is a strong positive correlation between EEM and SPY, LQD and TLT, VNQ and SPY, and EEM and VNQ. The SPY index is representative of the 500 largest stocks in the US. Stocks which have a positive correlation with SPY, VNQ and EEM, are relatively stable stocks to hold since they follow the market.

TLT and LQD, on the other hand, either have negative or no correlation with SPY, which is a good way to diversify the portfolio. TLT and LQD have lower returns. It is clear that these ETFs are in the portfolio to reduce the systematic risk involved in having ETFs positively correlated with the market. Idiosyncratic risk, however, cannot be reduced even through excessive diversification. 20% representation of the ETFs in this portfolio would be a good middle ground to compare a conservative and aggressive portfolio against.

Next, let's look at the covariance matrix:

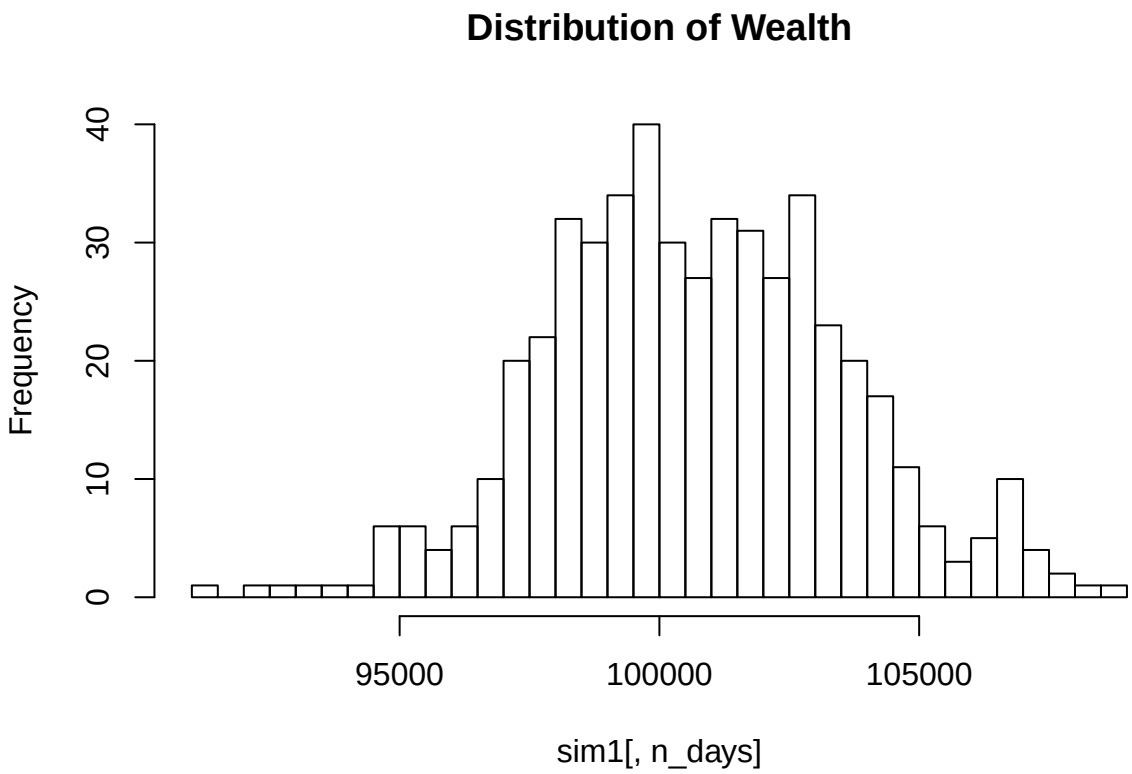
```
##                SPY.PctReturn TLT.PctReturn LQD.PctReturn EEM.PctReturn
## SPY.PctReturn   8.894429e-05 -4.986487e-05 -4.070904e-06  1.088871e-04
## TLT.PctReturn -4.986487e-05  9.527215e-05  2.546320e-05 -5.788960e-05
## LQD.PctReturn -4.070904e-06  2.546320e-05  1.271043e-05 -4.431753e-07
## EEM.PctReturn  1.088871e-04 -5.788960e-05 -4.431753e-07  1.890620e-04
## VNQ.PctReturn  8.630001e-05 -3.306871e-05  4.379381e-06  1.114575e-04
##                VNQ.PctReturn
## SPY.PctReturn   8.630001e-05
## TLT.PctReturn -3.306871e-05
## LQD.PctReturn  4.379381e-06
## EEM.PctReturn  1.114575e-04
## VNQ.PctReturn  1.380251e-04
```

The mean returns for the portfolio is as follows:

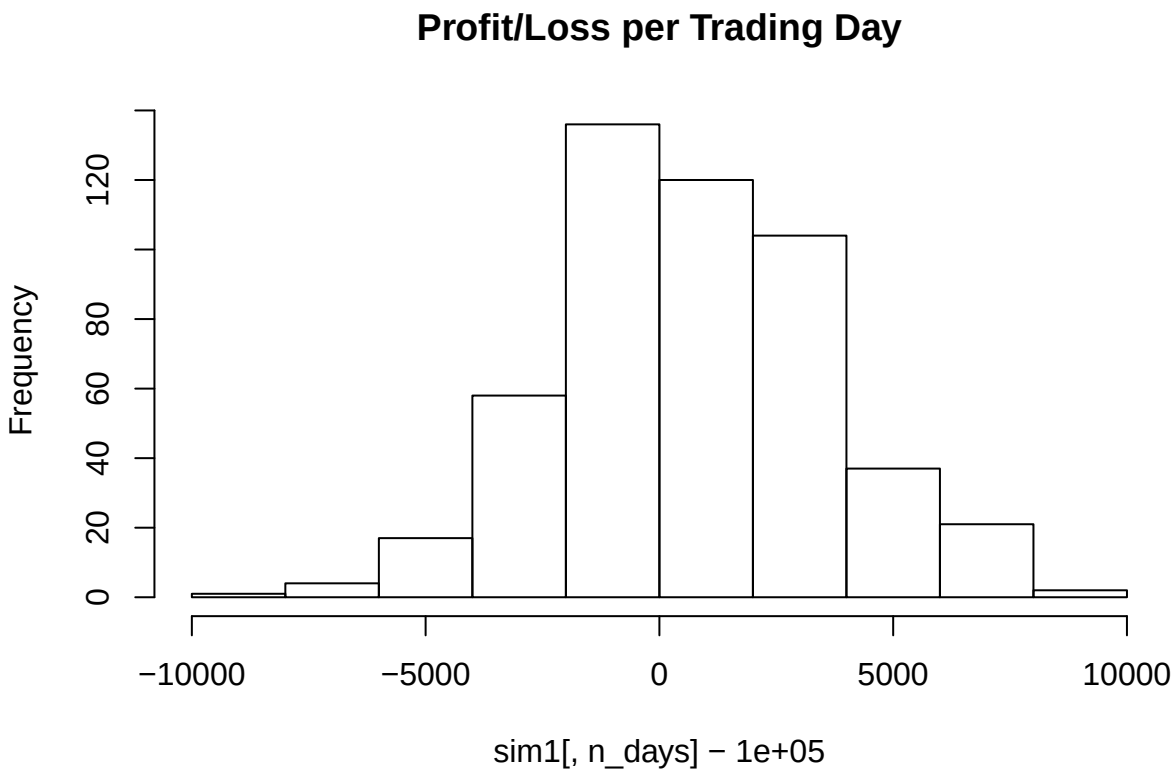
```
## [1] 0.004757295
```

On average, the portfolio returns 0.004%. Next, lets bootstrap returns on this portfolio for the 5 year trading period. We have returns of the portfolio over 1278 days, and the 4 week trading period consists of 20 days. Hence, we trade over 63.9 periods. Although this number seems odd, it gives us the precise returns (and total wealth) for a 20 day trading period in the given time frame. The selected seed is used to maintain reproducibility of results.

Each iteration of the bootstrap simulates one possible complete trading cycle. The following histogram shows the spread of total wealth each iteration of the bootstrap (for a total of 500 iterations).



The following histogram shows the profit and loss per trading day over the 500 iterations of the bootstrap:



The 5% VaR of this portfolio is -3701.3374863

The Conservative Portfolio

In the conservative portolio, I move away from emerging markets and real estate. I include the iShares Conservative allocation and Vanguard's Conservative ETF. I anticipate there will be a high correlation between these newly added ETFs, but the proportion of fixed income assets in these ETFs are favorable to the conservative investor.

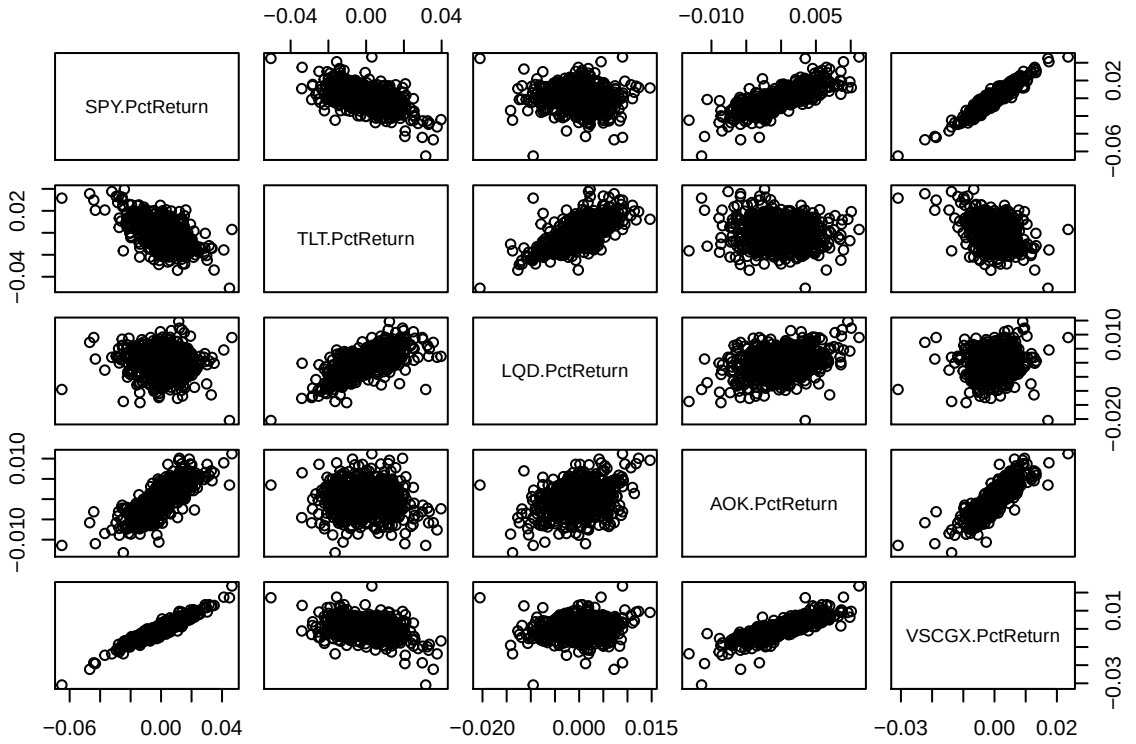
- US domestic equities (SPY)
- US Treasury bonds (TLT)
- Investment-grade corporate bonds (LQD)
- iShares Conservative allocation (AOK)
- Vanguard Conservative Growth Fund (VSCGX)

The mean percentage return of each ETF is as follows:

```
##   SPY.PctReturn   TLT.PctReturn   LQD.PctReturn   AOK.PctReturn
##    0.0006867895    0.0003101899    0.0002102527    0.0002205773
## VSCGX.PctReturn
##    0.0003059187
```

Returns on each ETF are significantly lower than the SPY return of 0.068%.

The pairwise correlations for this portfolio is as follows:



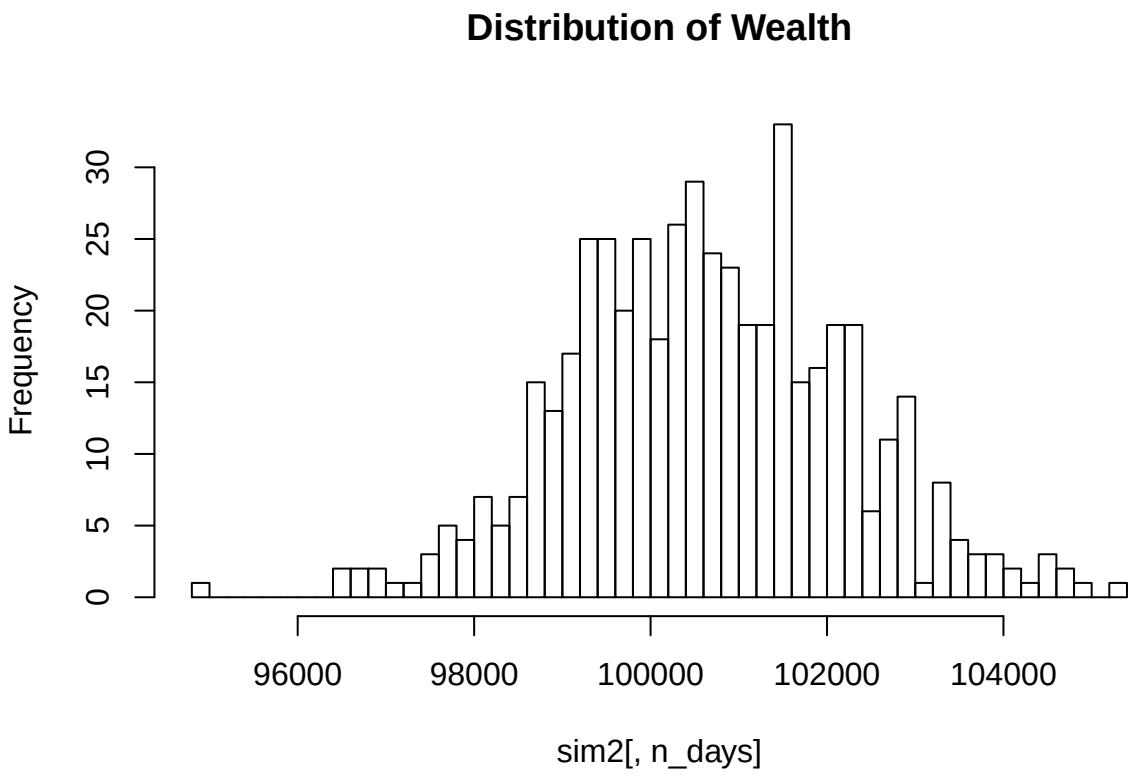
Next, let's look at the covariance matrix:

```
##                 SPY.PctReturn TLT.PctReturn LQD.PctReturn AOK.PctReturn
## SPY.PctReturn    8.894429e-05 -4.986487e-05 -4.070904e-06  2.088408e-05
## TLT.PctReturn   -4.986487e-05  9.527215e-05  2.546320e-05 -3.423033e-06
## LQD.PctReturn   -4.070904e-06  2.546320e-05  1.271043e-05  2.863709e-06
## AOK.PctReturn    2.088408e-05 -3.423033e-06  2.863709e-06  8.114087e-06
## VSCGX.PctReturn  3.415405e-05 -1.169145e-05  1.756773e-06  9.308725e-06
##                 VSCGX.PctReturn
## SPY.PctReturn       3.415405e-05
## TLT.PctReturn      -1.169145e-05
## LQD.PctReturn       1.756773e-06
## AOK.PctReturn       9.308725e-06
## VSCGX.PctReturn     1.479268e-05
```

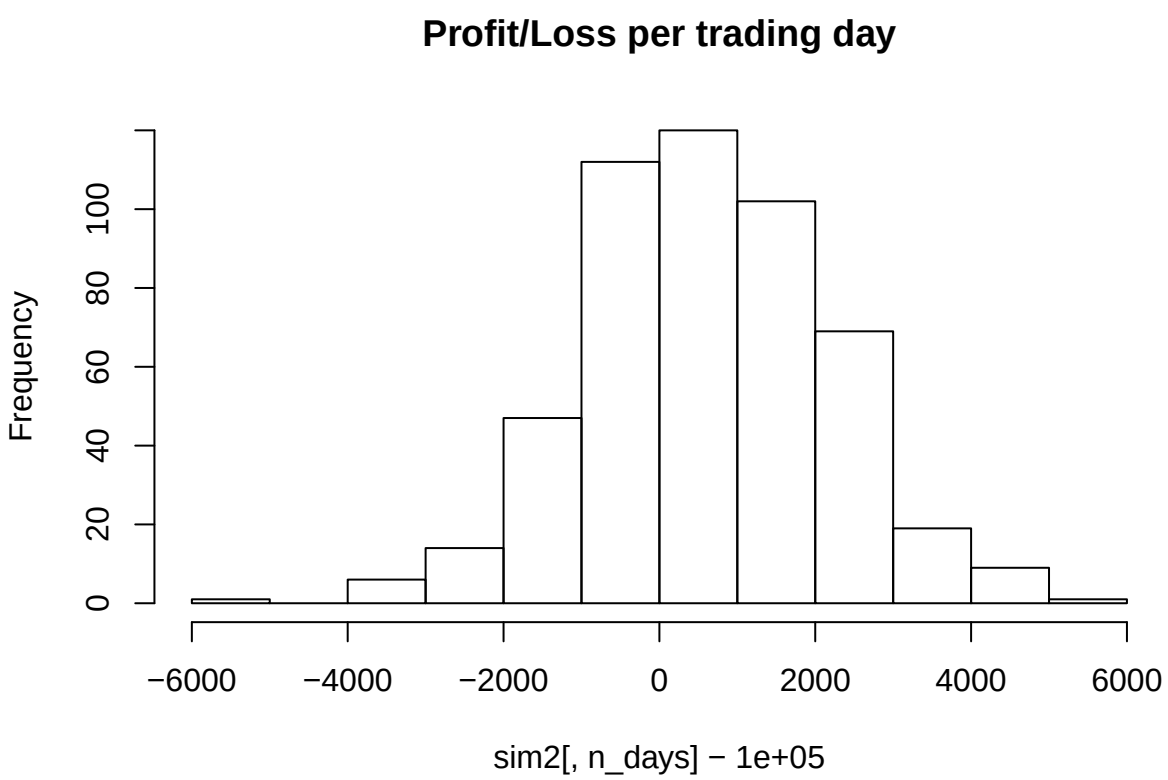
The variance of the new ETFs is smaller than the variance of ETFs included in the previous portfolio. The mean returns for the portfolio is as follows:

```
## [1] 0.00221019
```

On average, the portfolio returns 0.002%. Again, lets run a bootstrap with 500 iterations.



The following histogram shows the profit and loss per trading day over the 500 iterations of the bootstrap:



We can see clearly that compared to the previous portfolio, the variance in returns for the conservative portfolio is significantly lower. The 5% VaR of this portfolio is -1873.7711419, which is significantly lower than the 5% VaR of the balanced portfolio.

The Aggressive Portfolio

Next, lets look at a more aggressive portfolio. In the aggressive portolio, there is a focus on high return ETFs. Vanguard's high yield ETF and iShares Healthcare Providers ETF are included, along with EEM and VNQ. These were chosen due to their reputation for being aggressive, and availability of data at daily frequency as opposed to other ETFs which Yahoo collected data on at a lower frequency.

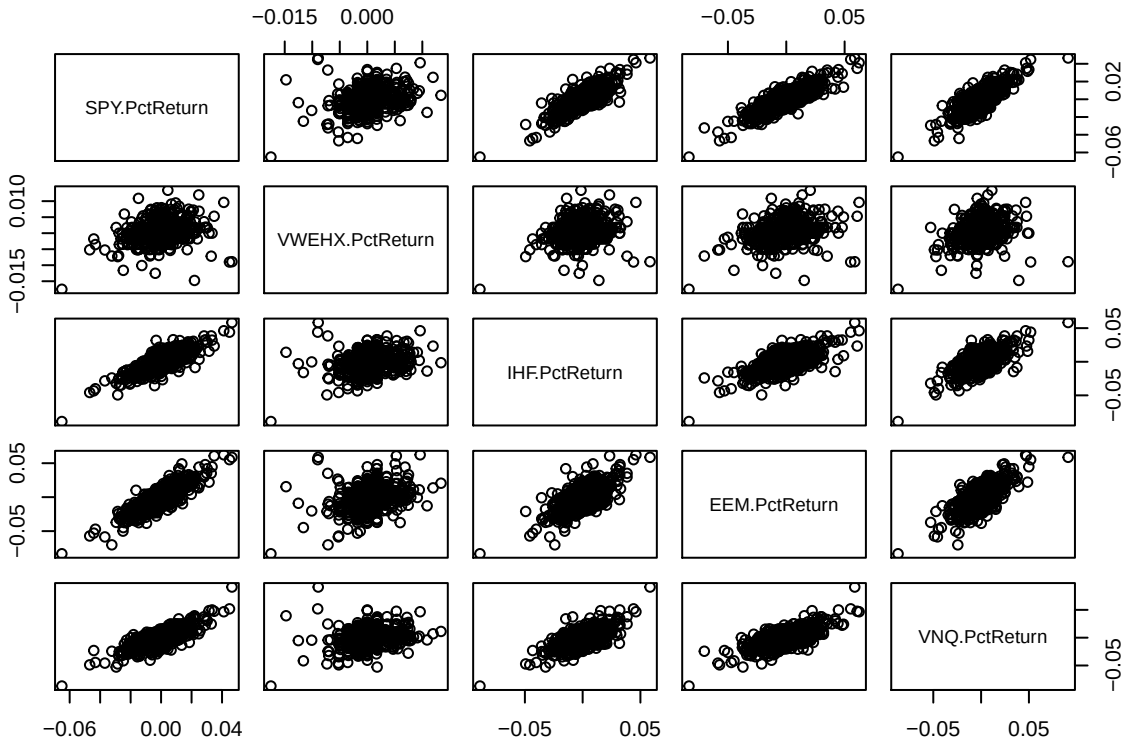
- US domestic equities (SPY)
- Vanguard High Yield Corporate Fund (VWEHX)
- iShares Healthcare Providers Fund (IHF)
- Emerging-market equities (EEM)
- Real estate (VNQ)

The mean percentage return of each ETF is as follows:

```
##   SPY.PctReturn VWEHX.PctReturn   IHF.PctReturn   EEM.PctReturn
##    0.0006867895    0.0003115168    0.0009597395    0.0001528394
##   VNQ.PctReturn
##    0.0006283819
```

Returns on VWEHX are surprisingly lower than the SPY return, which is counter intuitive for high risk portfolios. However, the return on IHF is relatively high.

The pairwise correlations for this portfolio is as follows:



IHF seems to be more strongly correlated with the market than VWEHX. This might explain the unexpected low return on VWEHX.

Next, let's look at the covariance matrix:

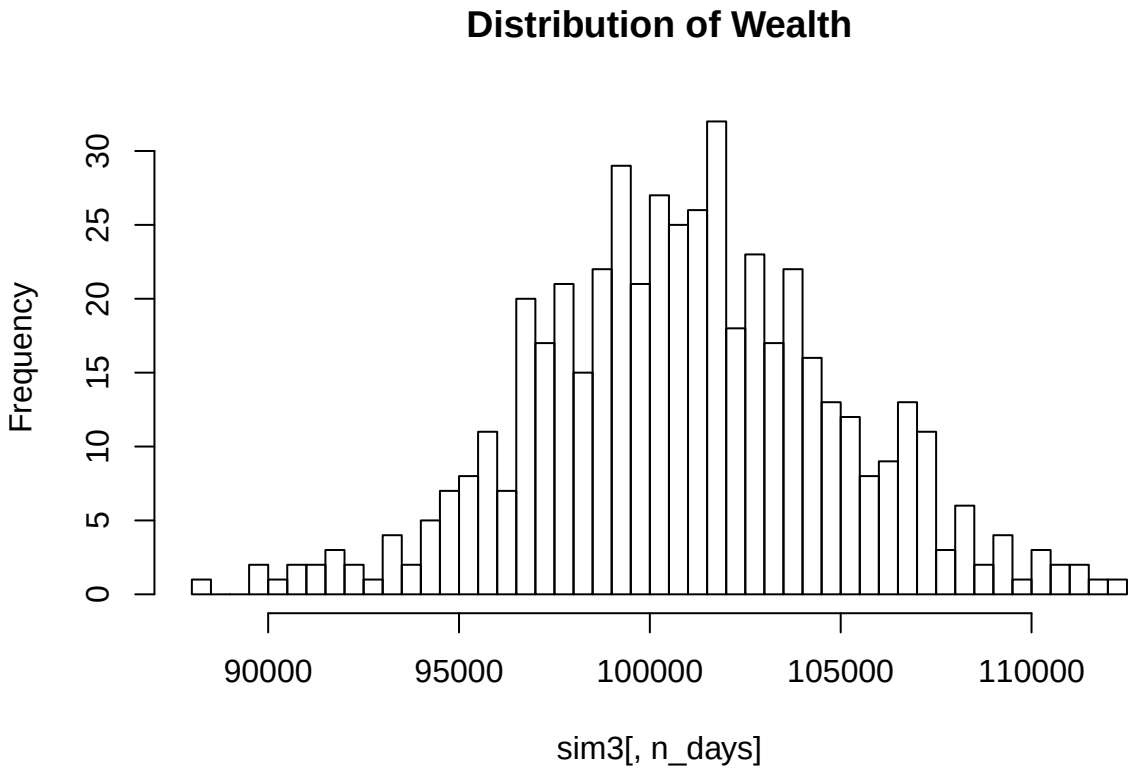
```
##                 SPY.PctReturn VWEHX.PctReturn IHF.PctReturn EEM.PctReturn
## SPY.PctReturn    8.894429e-05    5.850483e-06  8.554157e-05  1.088871e-04
## VWEHX.PctReturn  5.850483e-06    5.736560e-06  5.741738e-06  9.720435e-06
## IHF.PctReturn    8.554157e-05    5.741738e-06  1.251307e-04  1.021978e-04
## EEM.PctReturn    1.088871e-04    9.720435e-06  1.021978e-04  1.890620e-04
## VNQ.PctReturn    8.630001e-05    6.379331e-06  8.476497e-05  1.114575e-04
##                 VNQ.PctReturn
## SPY.PctReturn    8.630001e-05
## VWEHX.PctReturn  6.379331e-06
## IHF.PctReturn    8.476497e-05
## EEM.PctReturn    1.114575e-04
## VNQ.PctReturn    1.380251e-04
```

The variance of the new ETFs is generally larger than the variance of conservative ETFs included in the previous portfolio.

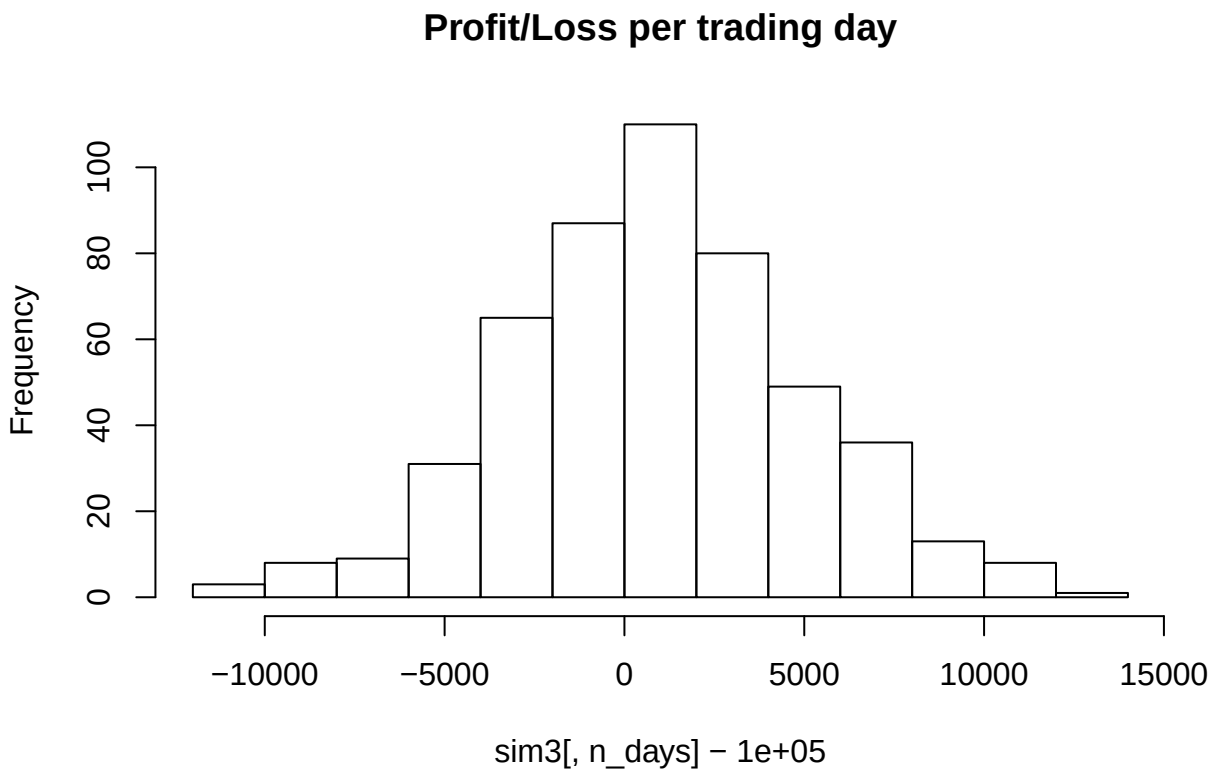
The mean returns for the portfolio is as follows:

```
## [1] 0.007219362
```

On average, the portfolio returns 0.007%. This is significantly larger than the returns from conservative (0.02%) and split (0.04%) portfolios. Again, lets run a bootstrap with 500 iterations.



The following histogram shows the profit and loss per trading day over the 500 iterations of the bootstrap:

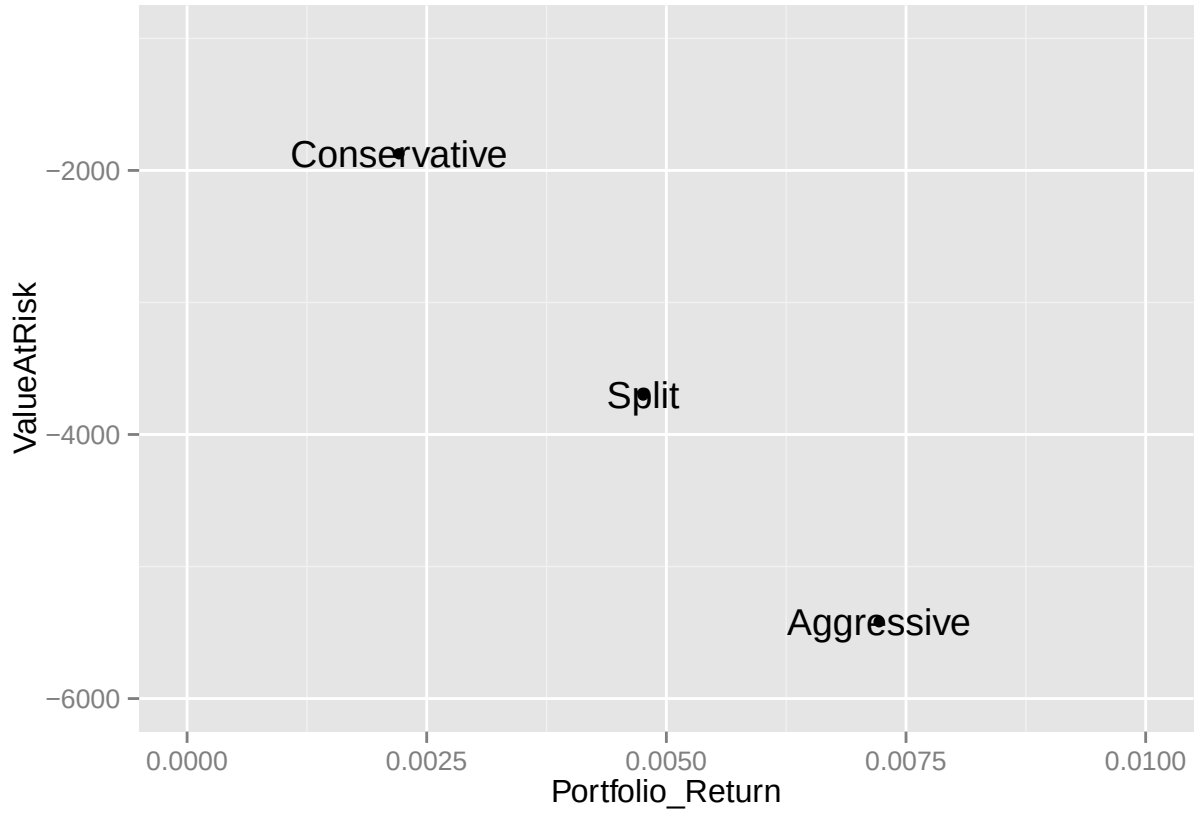


We can see clearly that compared to the other portfolios, the variance in returns for the aggressive portfolio is higher. The 5% VaR of this portfolio is -5418.3367582, which is significantly higher than the 5% VaR of the other portfolios.

The following table shows the mean returns and VaR for each portfolio:

Portfolio	Mean Return	VaR
Split	0.0047573	-3701.3374863
Conservative	0.0022102	-1873.7711419
Aggressive	0.0072194	-5418.3367582

The next graph shows each portfolio's bootstrapped mean returns (keep in mind a seed has been set and called to maintain consistency of 'events' that occur in each iteration of the bootstrap) and 5% VaR. As expected, the aggressive portfolio has the smallest VaR and conservative portfolio has the largest VaR. The returns are also increasing as the composition of risky assets in the portfolio increases.

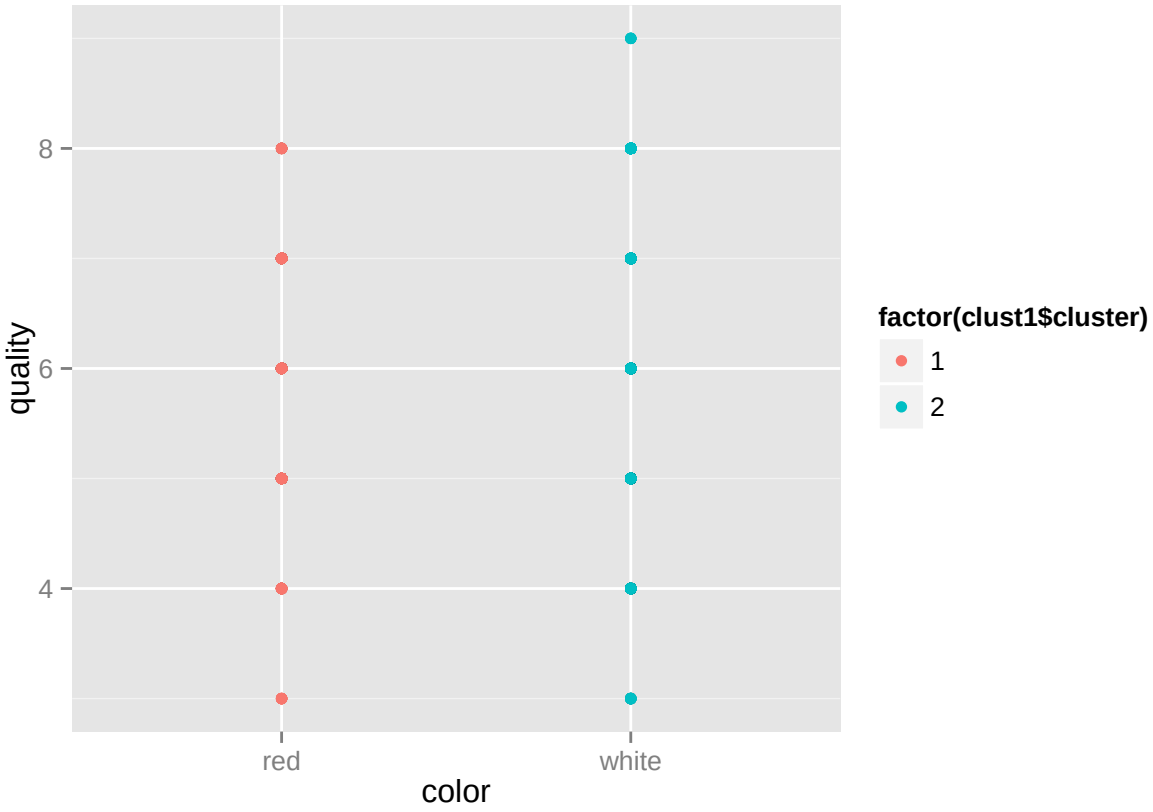


The table and graph of Mean Returns against VaR is a good measure of the risk profile of a portfolio. Based on the risk aversiveness of the investor, a portfolio can be chosen from the 3 choices. Additionally, this exercise can be repeated with different ETFs to compare how other portfolio combinations perform in a bootstrap test.

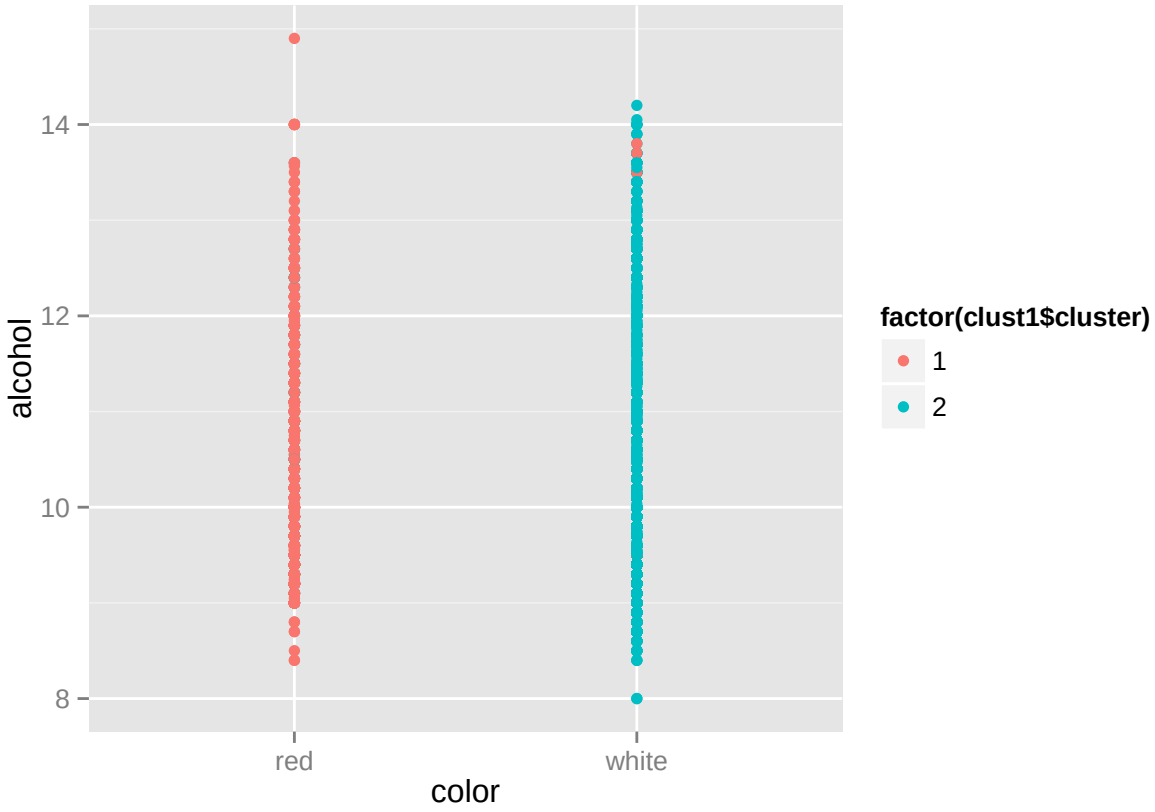
Clustering and PCA

Clustering and PCA will be used to classify the wines into their categories. I hypothesize that clustering will be inappropriate and difficult to understand or justify since the data contains the color of wine as categorical variables. On the other hand, PCA aims to reduce the dimensions in the attribute, and classification or dimension reduction will be less rigid than clustering.

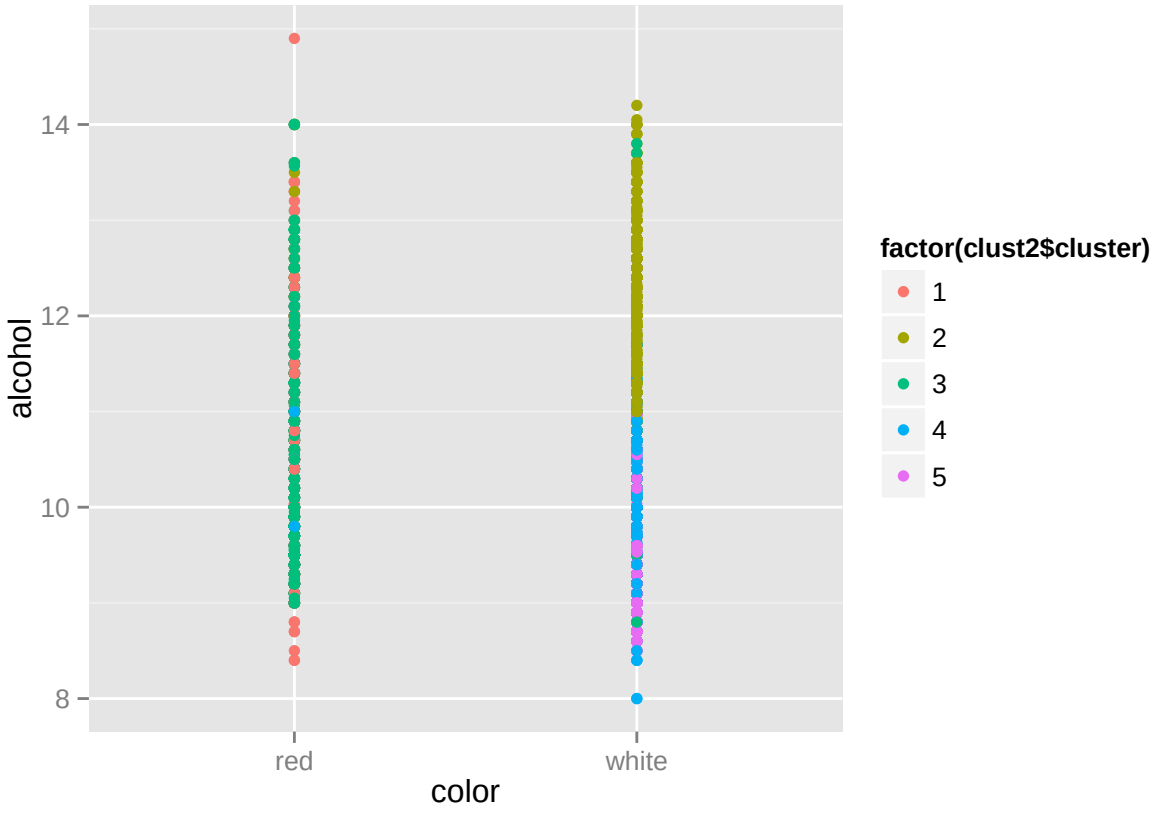
First, lets cluster the data and look at the clusters formed. First of all, let's look at how a run with 2 clusters looks based on quality:



We don't learn much from the above plot, so lets look at how well k means clustering identifies color in general. Lets look at how well clustering sorts the wines with respect to alcohol content:



Increasing the number of clusters to 5, however, yields some results which are difficult to interpret:

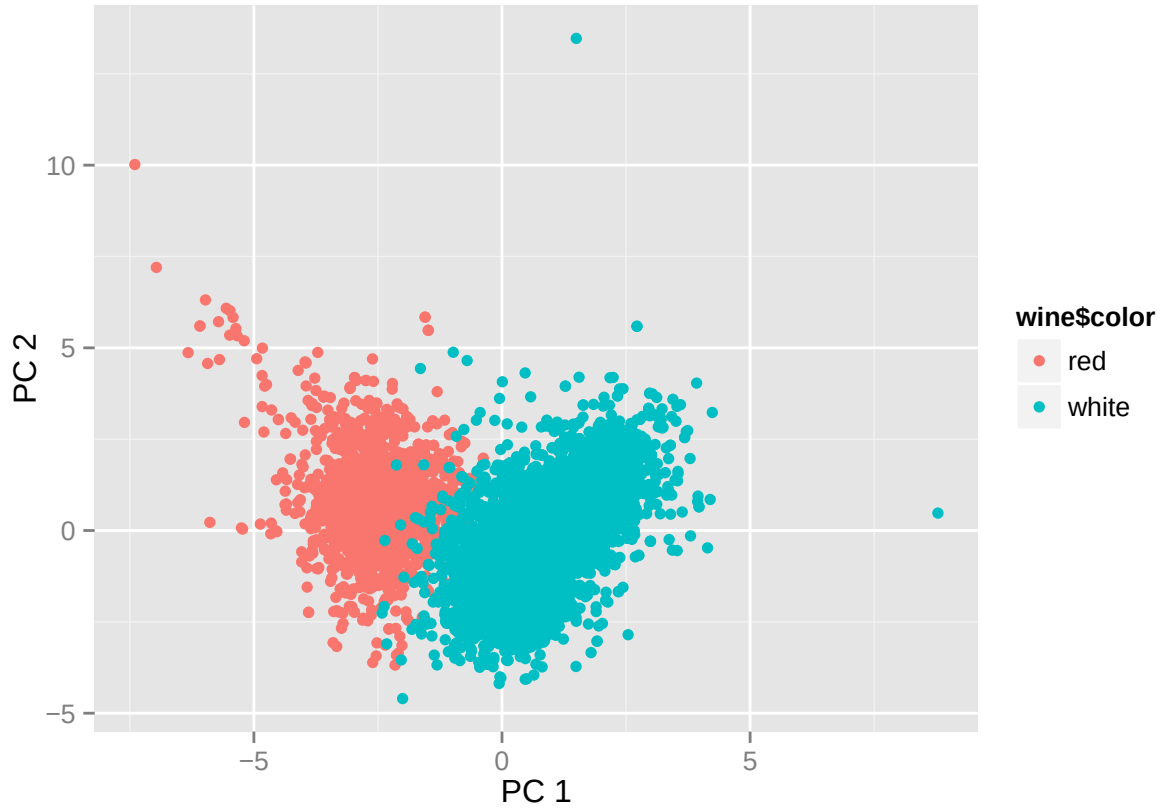


After some iterations with different means, I came to the conclusion that k-means clustering is only useful when k=2 for this data. Next, lets look at hierarchical clustering. Hierarchical clustering is done using average distance, and the following shows the number of points in each tree:

```
##    1    2    3    4    5    6    7    8    9   10
## 6454    6   22    2    6    1    2    2    1    1
```

A majority of the points fall under one of the 10 trees, which is not an even split. Similar results can be expected from a tree with a different number of levels. Hence, hierarchical clustering is not a good model for this data.

Next, I attempt PCA on this data. The following graph plots PC1 against PC2, sorted by color:



We can successfully distinguish between red and white wines from PC1. In general, a positive value in the vector of component 1 corresponds closely to white wine, and a negative value in the vector of component 1 corresponds to red wine. Let's look at these elements. The following is the loadings output of component 1, sorted by highest to lowest:

```
##    rownames(loadings)[pc1ord]
## 1        total.sulfur.dioxide
## 2         free.sulfur.dioxide
## 3              residual.sugar
## 4                 citric.acid
## 5                     density
## 6                     alcohol
## 7                          pH
## 8               fixed.acidity
## 9                   chlorides
## 10                  sulphates
## 11           volatile.acidity
```

Different forms of sulfur dioxide seem to be the detrimental chemicals in white wine. Residual sugar is the next largest component. 3 of those chemicals have high covariance with each other, and are a detrimental component of PC1. In the opposite direction, volatile acidity, sulphates, and chlorides have high covariance. From the first component, we are able to categorize the wines by 6 of the given 11 variables.

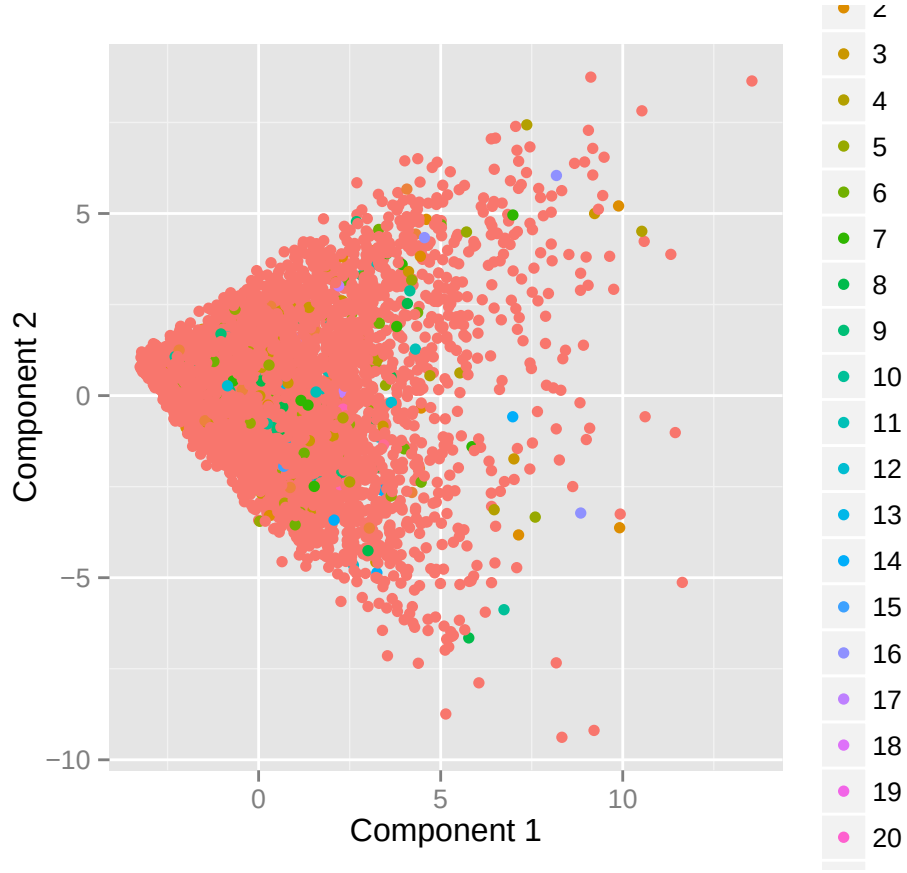
Both PCA and clustering are able to distinguish between red and white wines. Comparing PCA and clustering, I find that PCA has produced better results in two ways. From a methodological perspective, PCA is more efficient than clustering. PCA is naturally able to characterize the elements which give each color of wine its

properties. Additionally, the output from clustering might not yield desirable results with other variations of k.

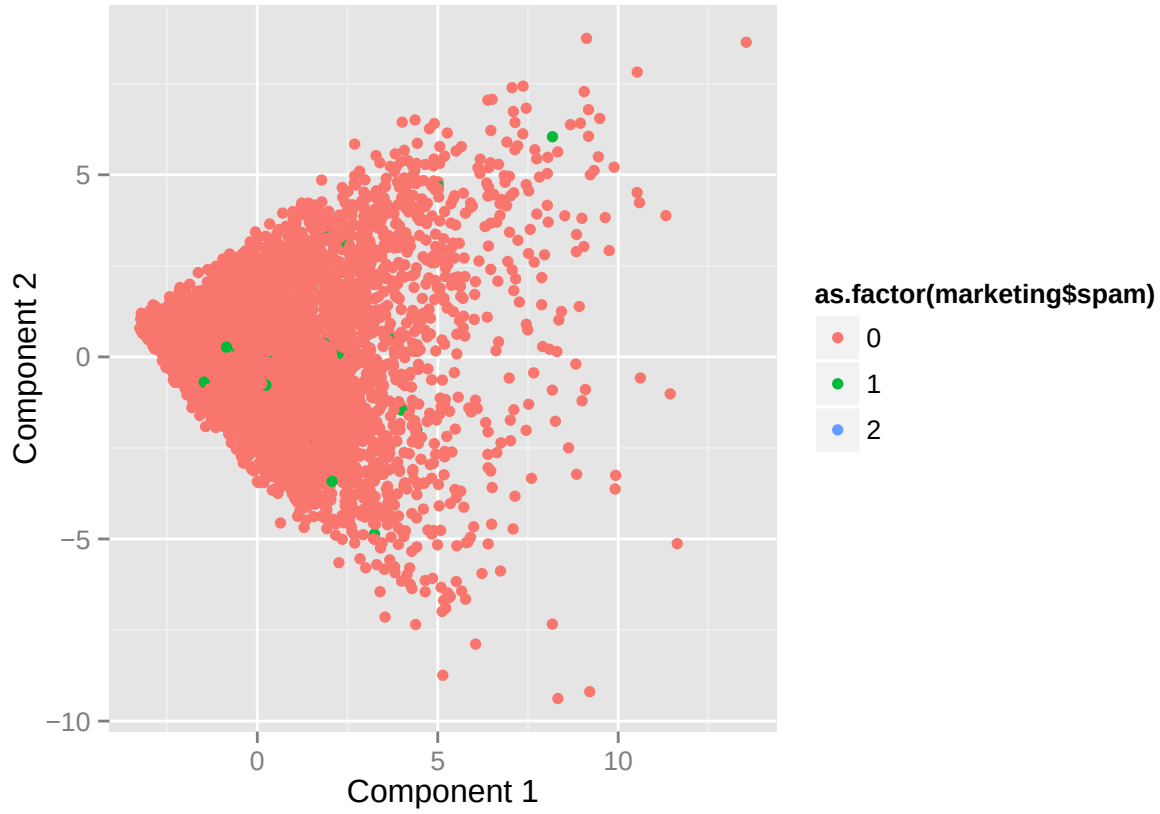
Market Segmentation

Given the dimensions of the data, the most useful tool in this situation would be a dimension reduction method. First of all, some clean up needs to be done. The column for uncategorized tweets is dropped from the data. Trying to categorize or explain an uncategorized variable using PCA defeats the purpose of the components. Measuring covariance between an uncategorized vector and any other vector is counter intuitive and not interpretable. Since the PCA algorithm takes into account all the variables and will consider the effect of being in no category, it is dropped.

The plot of components 1 against 2 is as follows:



In the above plot, the tweets which have been categorized as adults are spread out. Lets look at tweets which have been categorized as spam on a plot of principal components:



The tweets marked as spam are slightly better categorized, but we are unable to draw a decent conclusion due to the small number of data points and spread of points. Lets take a look at the first principal component:

```
##    rownames(loadings)[pc1ord]
## 1                    religion
## 2                        food
## 3                   parenting
## 4               sports_fandom
## 5                      school
## 6                      family
## 7                      beauty
## 8                      crafts
## 9                     cooking
## 10                    fashion
## 11              photo_sharing
## 12                       eco
## 13                 computers
## 14                   outdoors
## 15           personal_fitness
## 16                   business
## 17                 automotive
## 18                   politics
## 19                   shopping
## 20                       news
## 21             sports_playing
## 22                    chatter
```

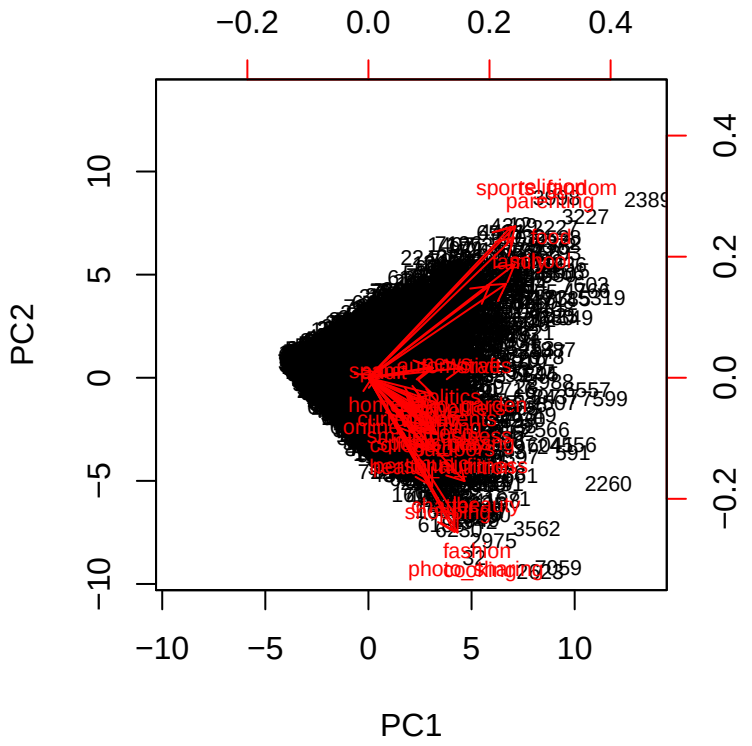
```
## 23           health_nutrition
## 24                      music
## 25                     travel
## 26             small_business
## 27            home_and_garden
## 28                     dating
## 29             current_events
## 30                        art
## 31                   tv_film
## 32               college_uni
## 33             online_gaming
## 34                     adult
## 35                      spam
```

We see that in the first principal component, adult and spam are actually the variables with the lowest values. Hence, the spam and adult labels do not vary much with the other variables. Religion, food, parenting, sports fandom, school, family, beauty, crafts, cooking, and fashion vary together in the first principal component. These generally relate to married adults.

Lets look closely at the categories of the second principal component:

```
##  [1] 26  6 28  8 30  9 12 23 24 35 34  7 10 20  3 25 29  5  2 13 19 33 21
## [24] 17 16 22 11 31 15  1 27 14 32  4 18
```

Again, we see some of the same components repeating. Religion, parenting, sports fandom, and food are among the top components. In the opposite direction, beauty, cooking, and fashion are also significant in this component. We can assume that the first customer segment is parents. Lets test this hypothesis using a biplot of both principal components.



Like seen earlier, sports fandom, food, religion, and parenting are strong components in the same direction. Hence, the hypothesis is right. There is a distinct 90 degree angle between 2 large sets of components, which suggests there could be 2 distinct customer segments. The second customer segment seems to be heavily based on fashion, photo sharing, beauty, outdoors, personal fitness, and online gaming. This customer segment seems to correspond more to the younger population, teenagers.

In conclusion, the two major customer segments we can distinguish from principal component analysis is parents and teenagers.