Intro to Machine Learning (STA380) - Part 2

Christian Alfonso, Musmin Zar, Satya Pal, Vinay Pahwa

16/08/2021

```
library(mosaic)
```

```
## Warning: package 'mosaic' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(quantmod)
```

```
## Warning: package 'quantmod' was built under R version 4.0.5
```

```
## Loading required package: xts
```

```
## Warning: package 'xts' was built under R version 4.0.5
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.0.5
```

```
##
## Attaching package: 'zoo'
```

```
## 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
```

```
## Warning: package 'TTR' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(foreach)
```

```
## Warning: package 'foreach' was built under R version 4.0.5
```

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 4.0.5
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate

##
## Attaching package: 'tm'

## The following object is masked from 'package:mosaic':
##
##     inspect
```

```
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 4.0.5
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.0.5
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.5

##
## Attaching package: 'caret'

## The following object is masked from 'package:mosaic':
##
##     dotPlot
```

```
library(dplyr)
library(doParallel)
```

```
## Warning: package 'doParallel' was built under R version 4.0.5

## Loading required package: iterators

## Warning: package 'iterators' was built under R version 4.0.5

## Loading required package: parallel
```

```
library(foreach)
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.0.5

## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 4.0.5

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following object is masked from 'package:mosaic':
##
##     count

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

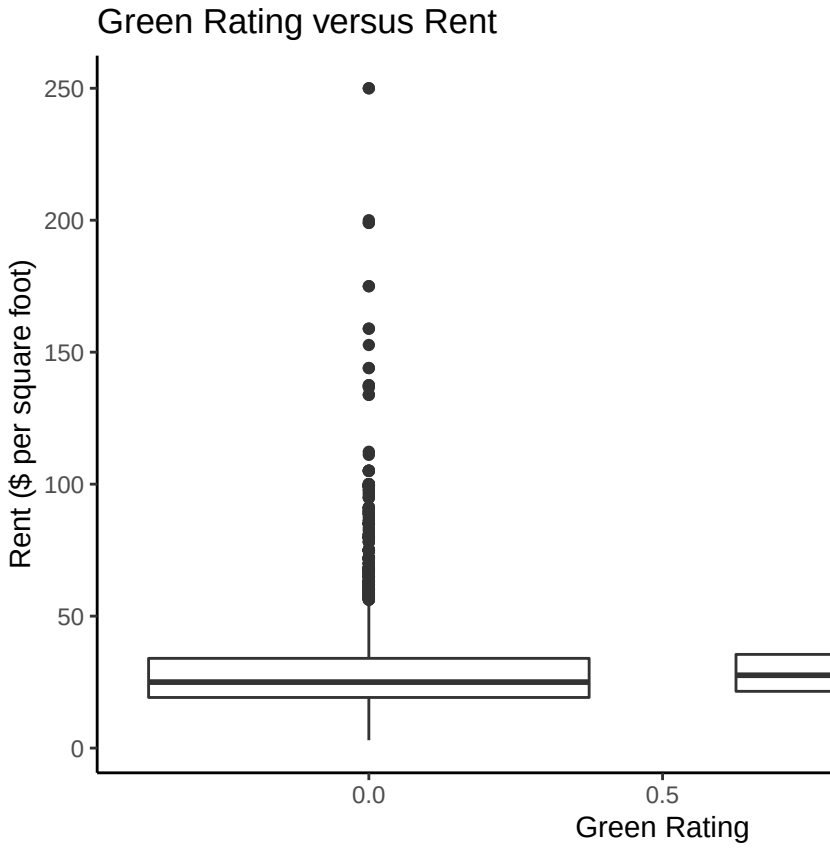
```
library(arules) ## install.packages('arules')
```

```
## Warning: package 'arules' was built under R version 4.0.5

##
## Attaching package: 'arules'

## The following object is masked from 'package:tm':
##
##     inspect

## The following objects are masked from 'package:mosaic':
##
##     inspect, lhs, rhs
```

```
## The following object is masked from 'package:dplyr':
##
##     recode

## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

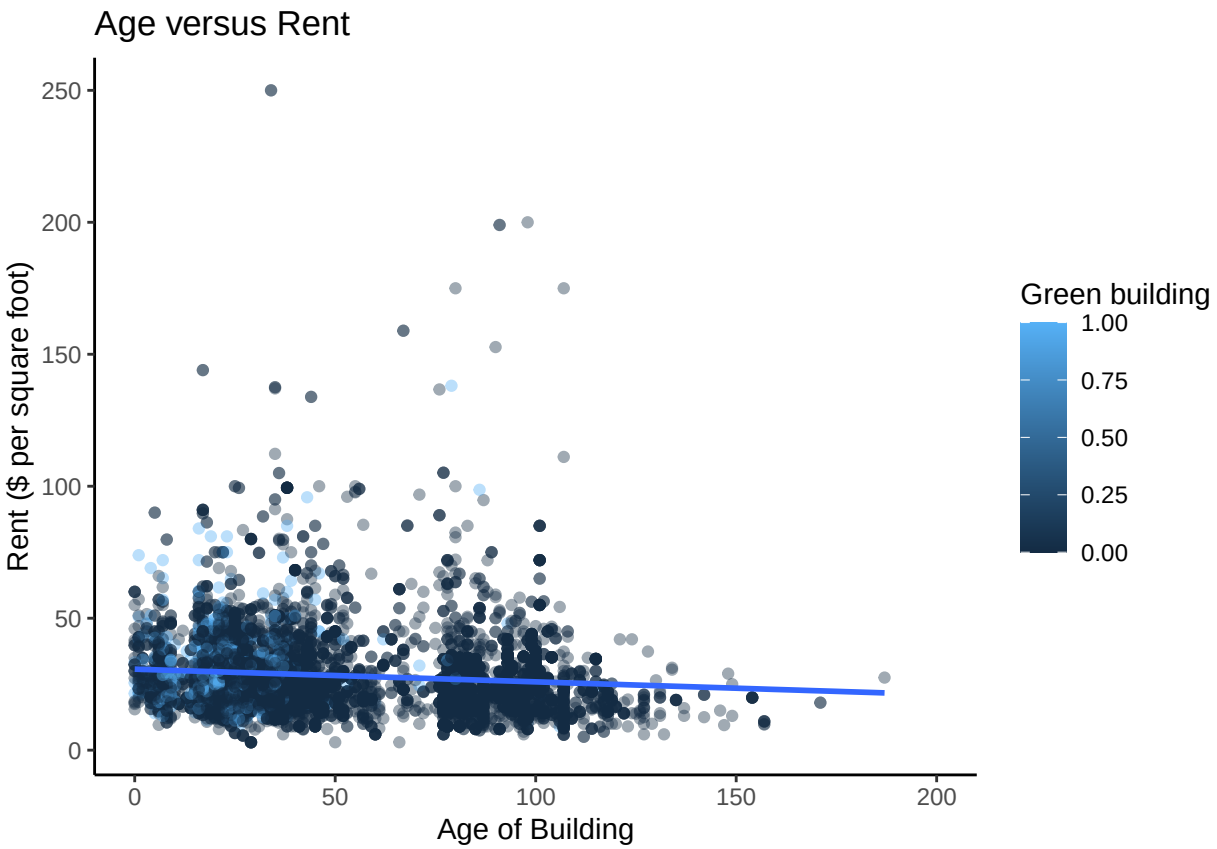
Question 1: Visual Story Telling (Part 1): Green Building

I do not agree with the stats guru. For starters, only 216 houses have an occupancy of less than 10%. This is a small amount of buildings, and there is no way to know what is happening in this buildins, so we shouldn't just write them off.



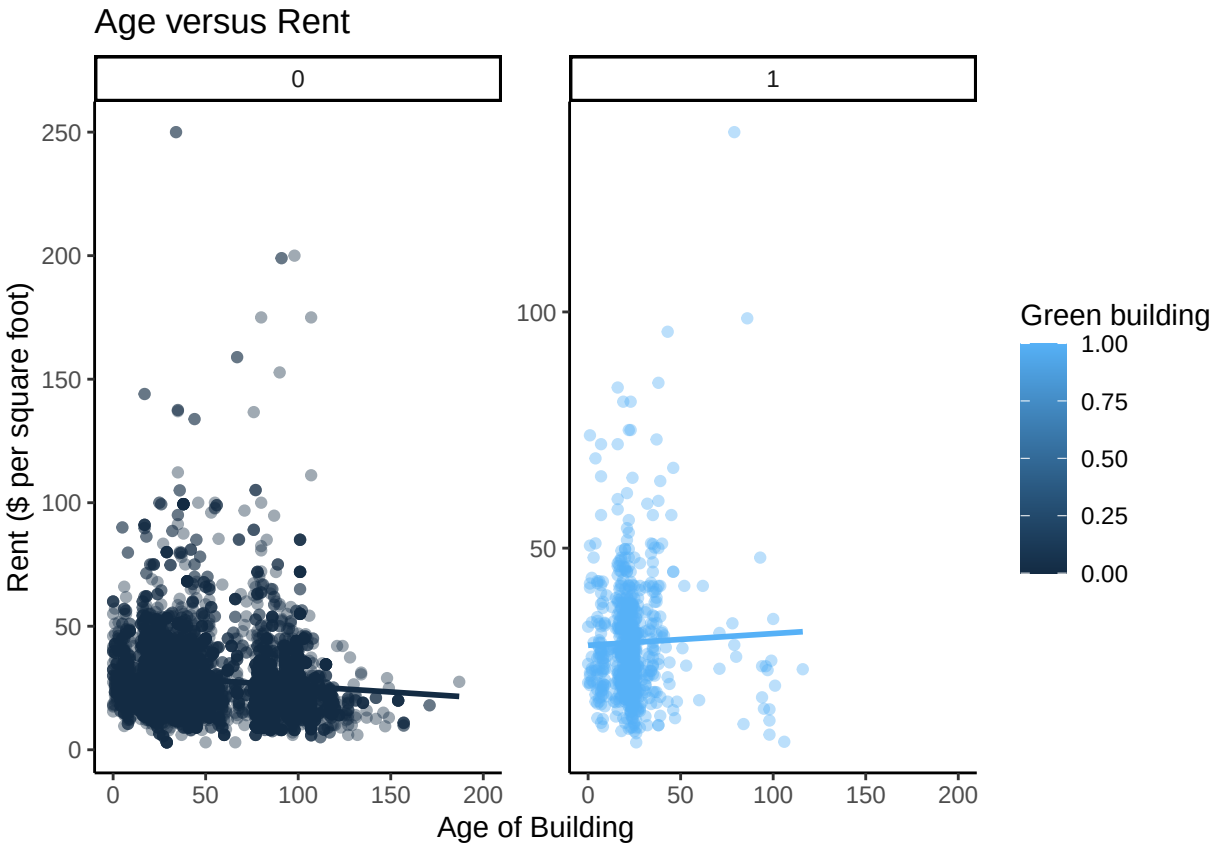
To begin, we have the plot of the data the Guru uses.
He is saying the rent is higher because the buildings are green. He seems right if you use only this data, but lets look at some more.

```
## `geom_smooth()` using formula 'y ~ x'
```

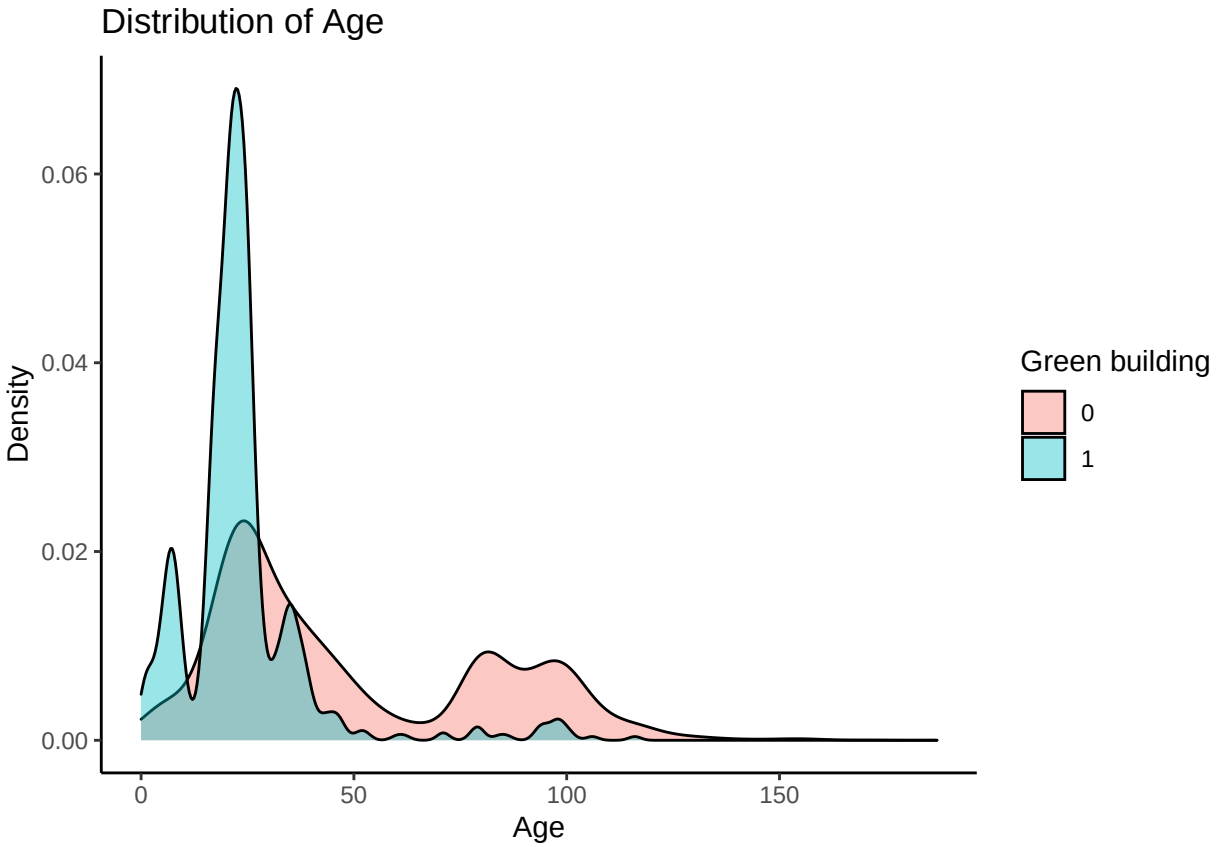


Here, I am looking at the relationship between Age and Rent. It is a bit hard to see so we will separate it.

```
## `geom_smooth()` using formula 'y ~ x'
```



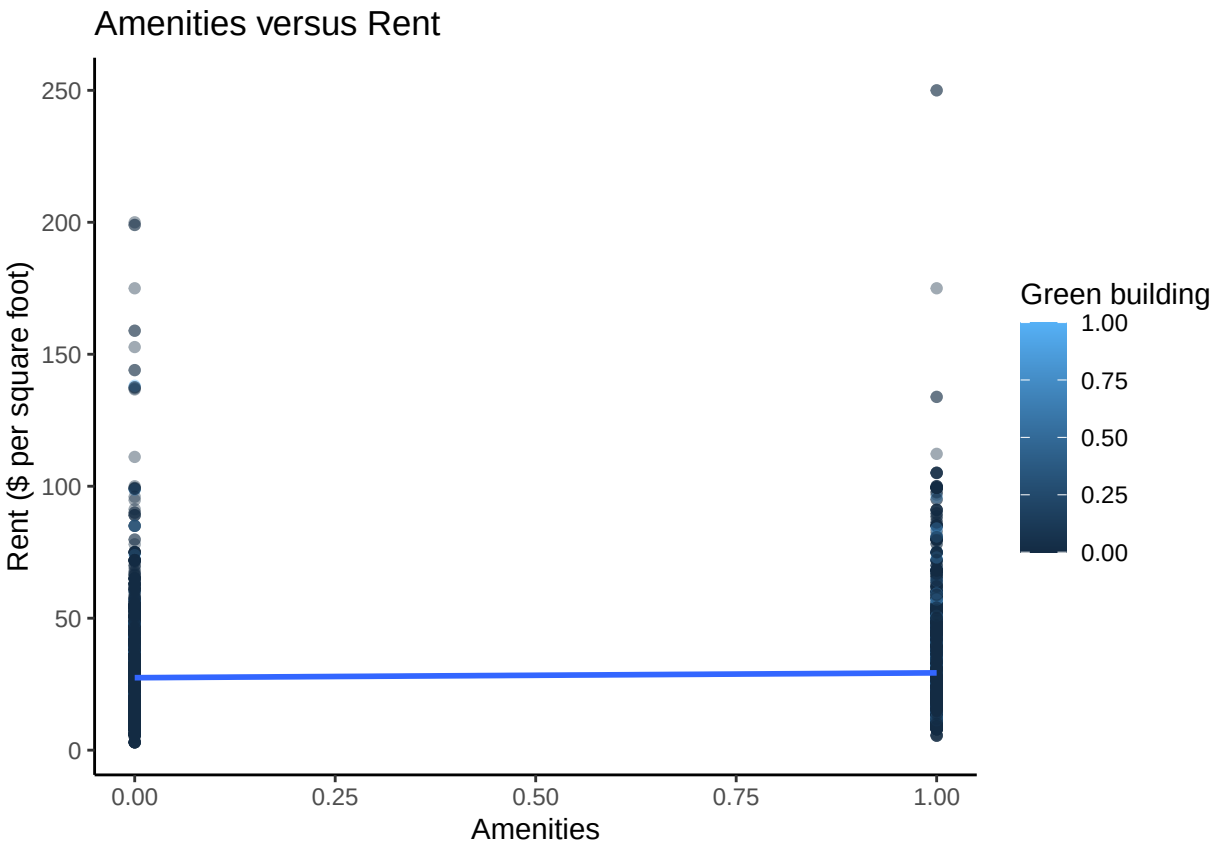
Now we can see how rent goes down as age goes up for non-green buildings, but the inverse is true for green buildings. We can already start to see a little of how the Guru is wrong.



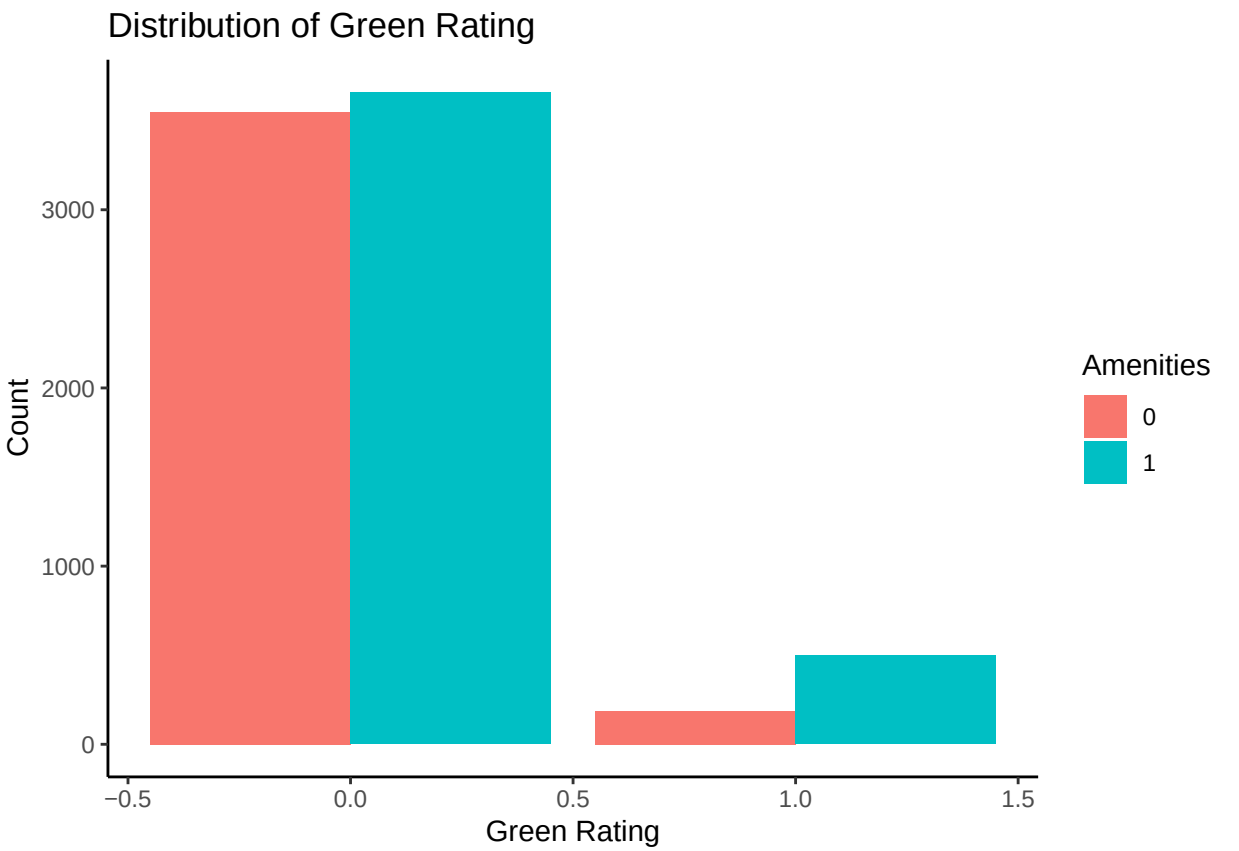
Here we can see a majority of the green buildings are newer. This shows us the Guru can also be wrong because he assumes the rent is higher because the buildings are green, when it could also be because the buildings are newer. Age is a confounding variable, and should have been taken into consideration.

The assumption made by the guru is wrong, the building would pay for itself faster as time when on and rent increased gradually.

```
## `geom_smooth()` using formula 'y ~ x'
```



To see if I could grab another confounding variable, I attempted using amenities. The plot doesn't go so well, but I am saying the cost of rent is higher where there are more amenities. Next I have a bar chart to show green rating.



Here, we can see the green buildings have more amenities, and rent is higher in both green buildings, and buildings with more amenities, making amenities another confounding variable.

All and all, the Guru was wrong for making quick assumptions of the data without trying to find if his correlation was true regarding other variables and if other variables play a factor.

Question 2: Visual Story Telling (Part 2): flights at ABIA

Question 3: Portfolio Modeling

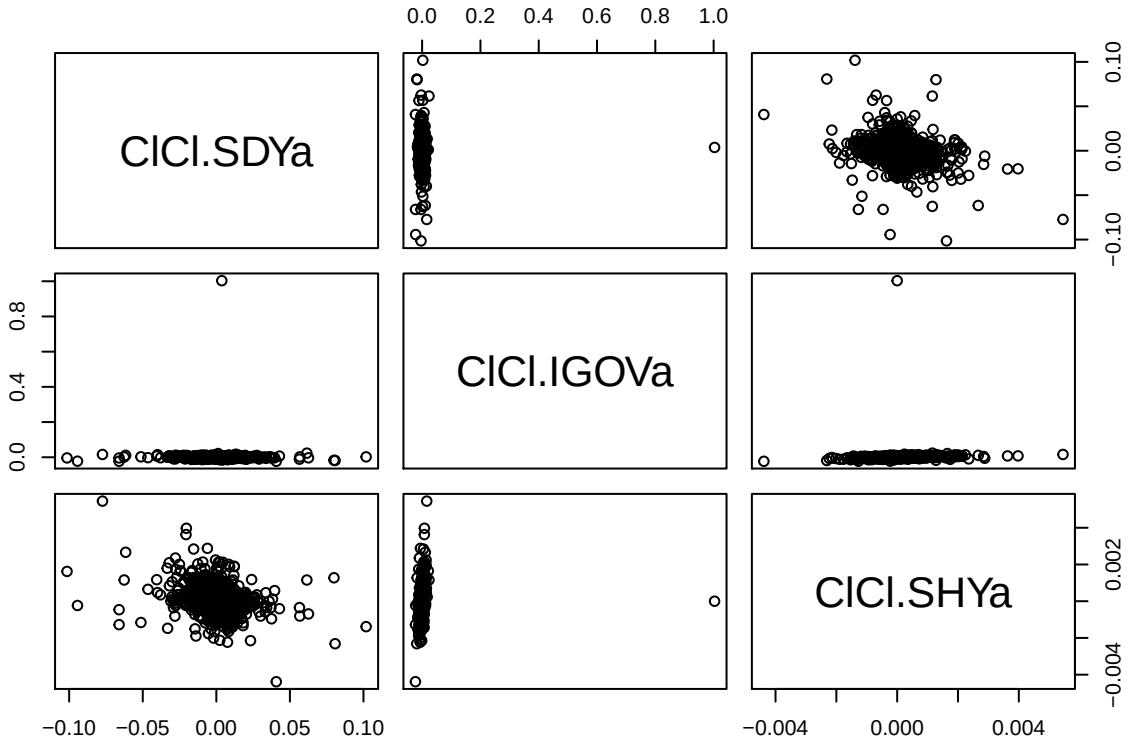
We have decided to choose three different portfolios with the following breakdowns:

Portfolio 1 (Low Risk Portfolio):

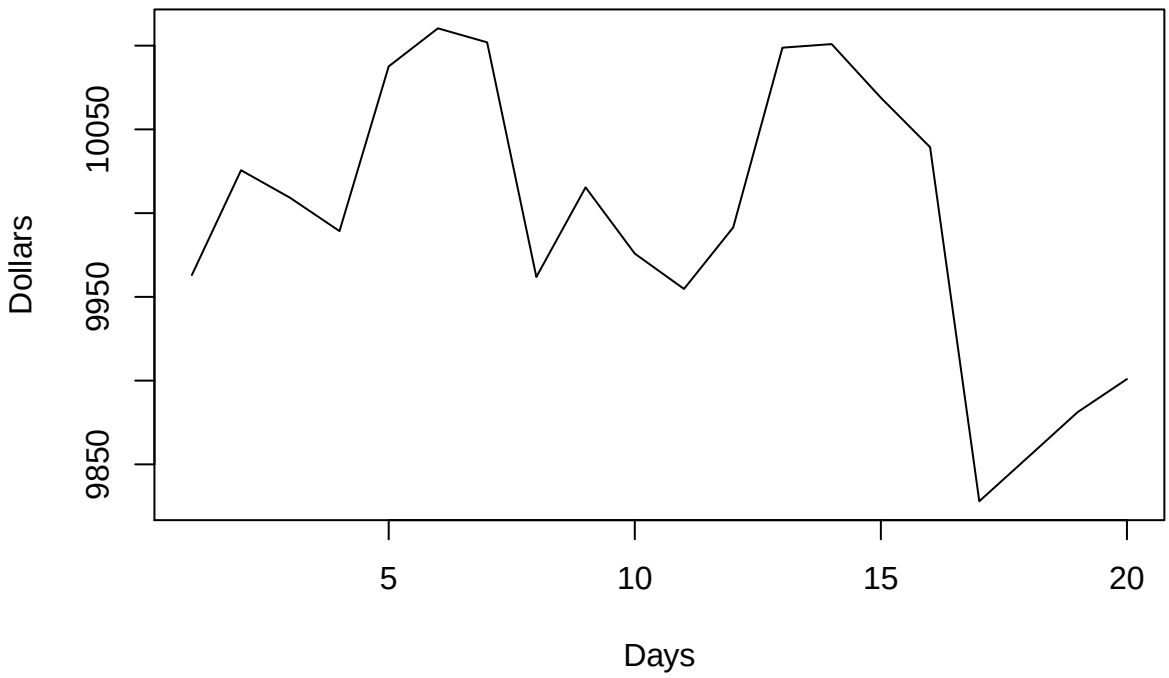
This portfolio is made up of index ETF's and bond ETF's. It is meant to represent a low risk investment strategy.

- SDY - SPDR S&P Dividend ETF (80%)
- IGOV - iShares International Treasury Bond ETF (10%)
- SHY - iShares 1-3 Year Treasury Bond ETF (10%)

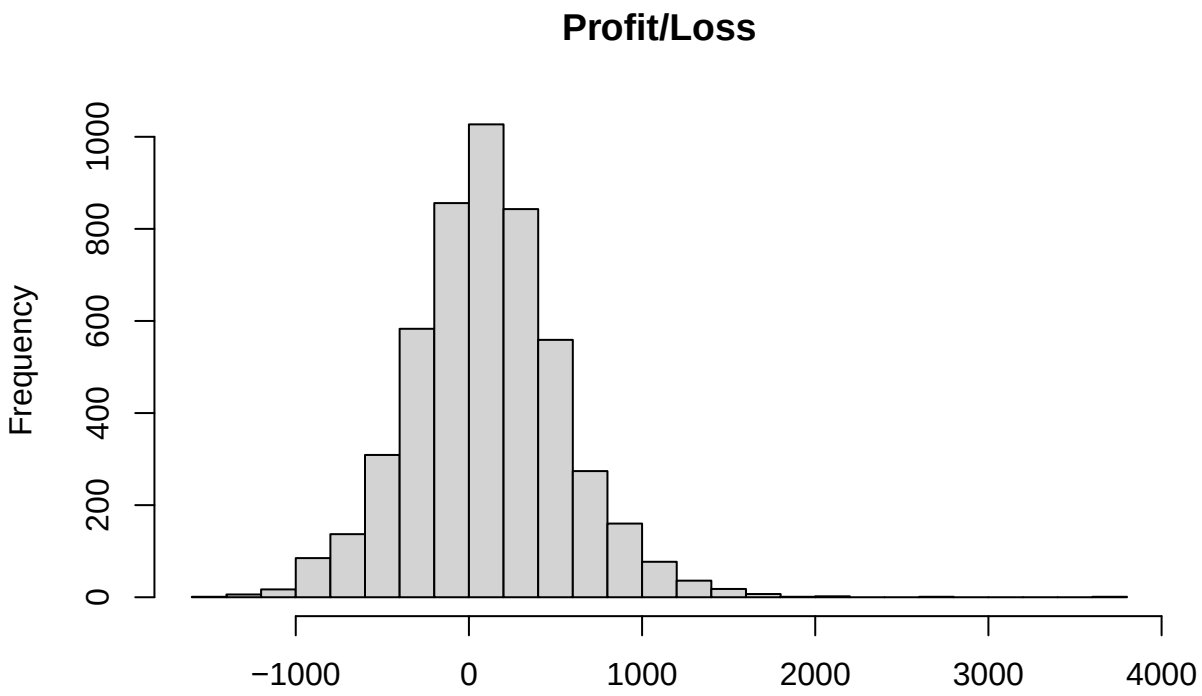
Below is a pairings plot of the selected ETFs:



We are running 5000, 20 trading day models. One of the models looks follows the graph seen below:



A summary of these 5000 models can be seen here:



The mean profit/loss was:

```
## [1] 108.1641
```

The 5% VaR was:

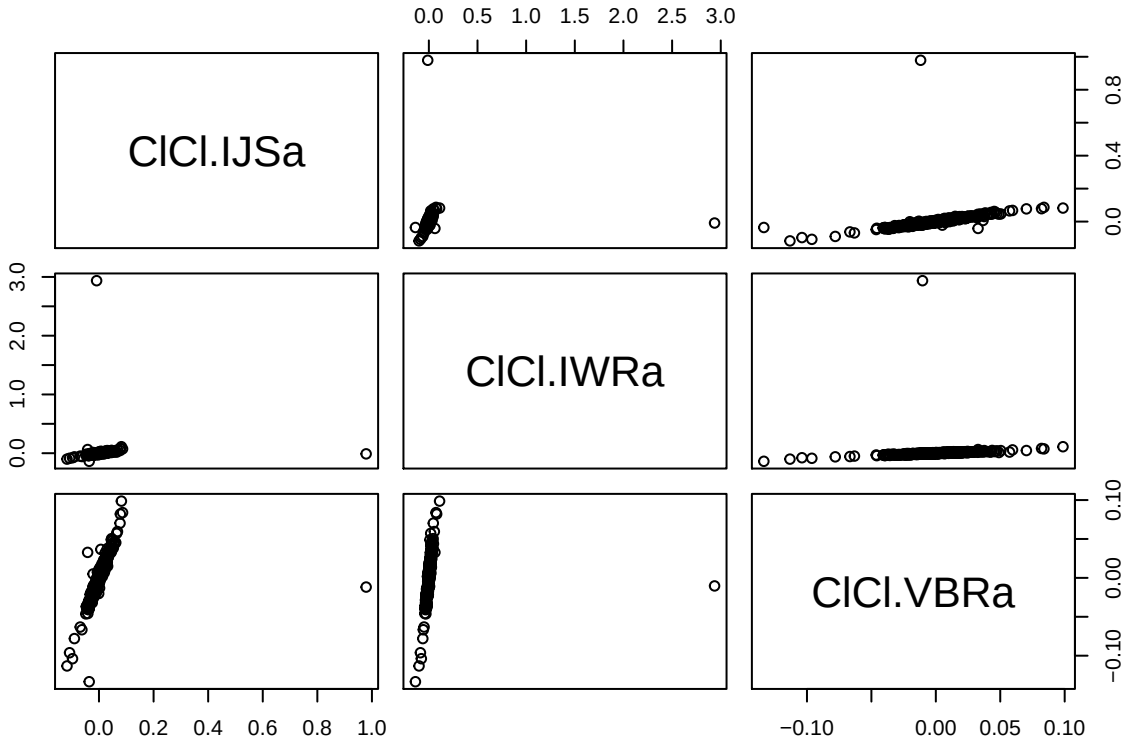
```
##        5%
## -595.6386
```

Portfolio 2 (High Risk):

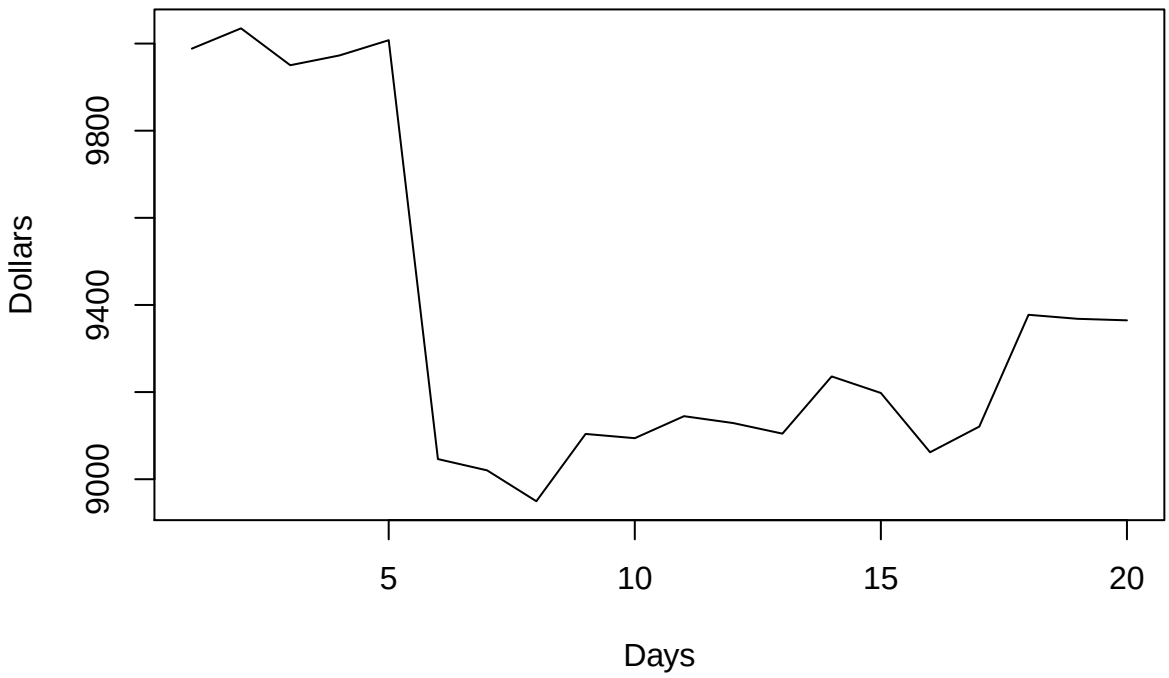
This portfolio is made up of small cap and mid cap index ETFs. It is meant to represent a high risk investment strategy.

- IJS - iShares S&P Small-Cap 600 Value ETF (40%)
- IWR - iShares Russell Mid-Cap ETF (30%)
- VBR - Vanguard Small-Cap Value Index Fund ETF (30%)

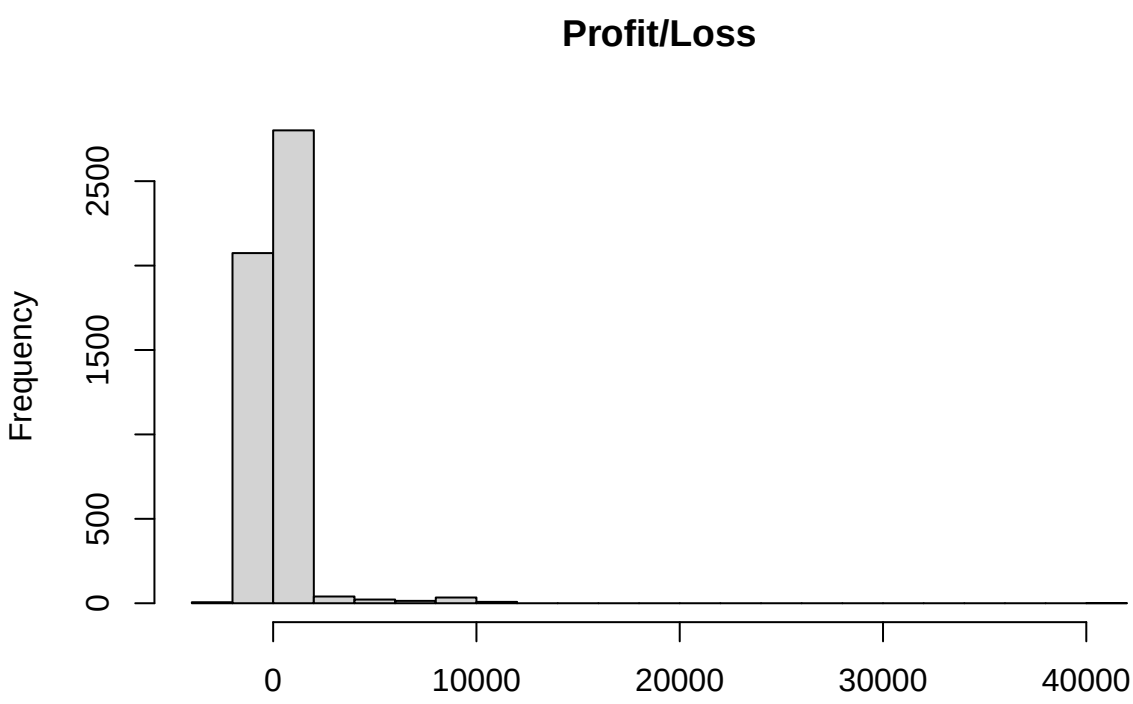
Below is a pairings plot of the selected ETFs:



We are running 5000, 20 trading day models. One of the models looks follows the graph seen below:



A summary of these 5000 models can be seen here:



The mean profit/loss was:

```
## [1] 253.7411
```

The 5% VaR was:

```
##        5%
## -898.6983
```

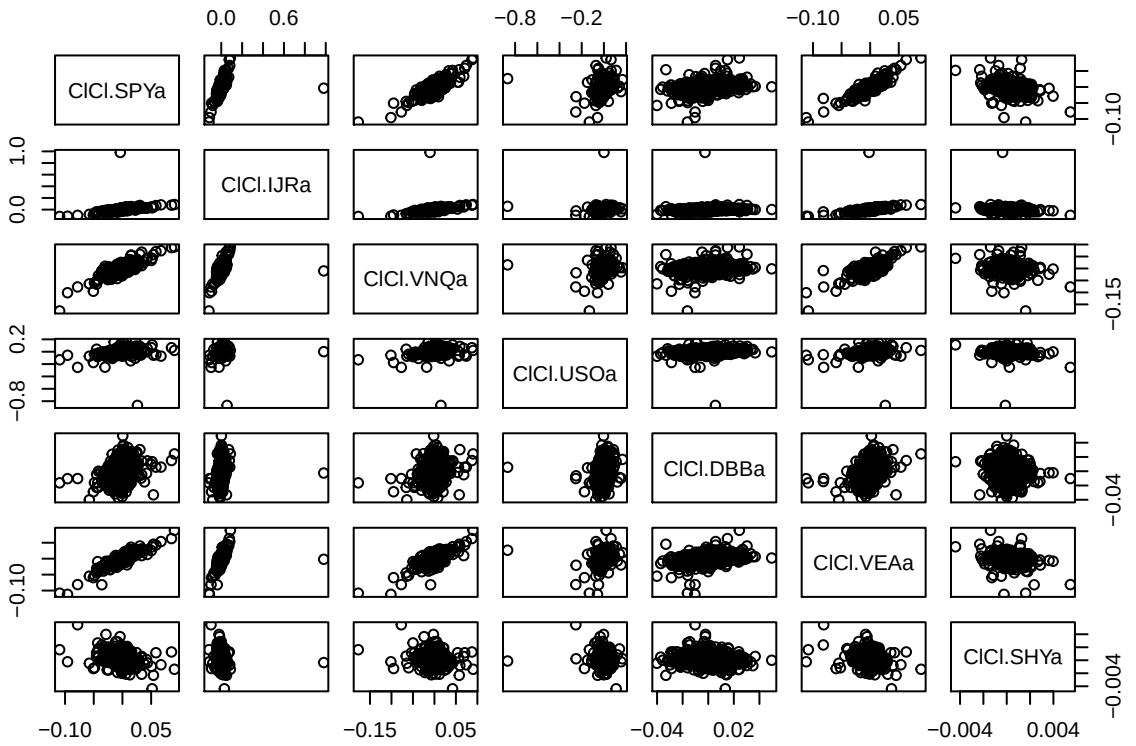
Portfolio 3 (Diverse):

This portfolio is made up of a diverse range of ETFs. It is meant to represent a highly diversified investment strategy.

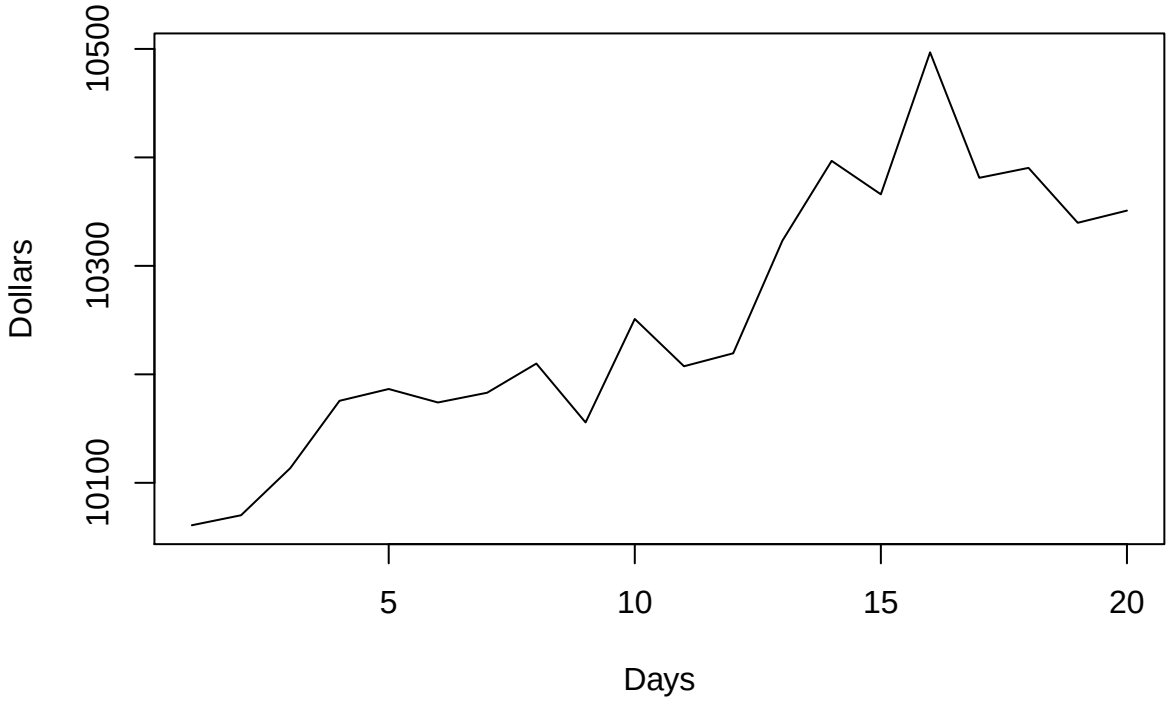
- SPY - SPDR S&P500 ETF (20%)
- IJR - iShares Core S&P Small-Cap ETF (20%)
- VNQ - Vanguard Real Estate Index Fund ETF (10%)
- USO - United State Oil Fund (10%)
- DBB - Investco DB Base Metals (10%)
- VEA - Vanguard Developed Markets Index Fund ETF (10%)
- SHY - iShares 1-3 Year Treasury Bond ETF (10%)

Below is a pairings plot of the selected ETFs:

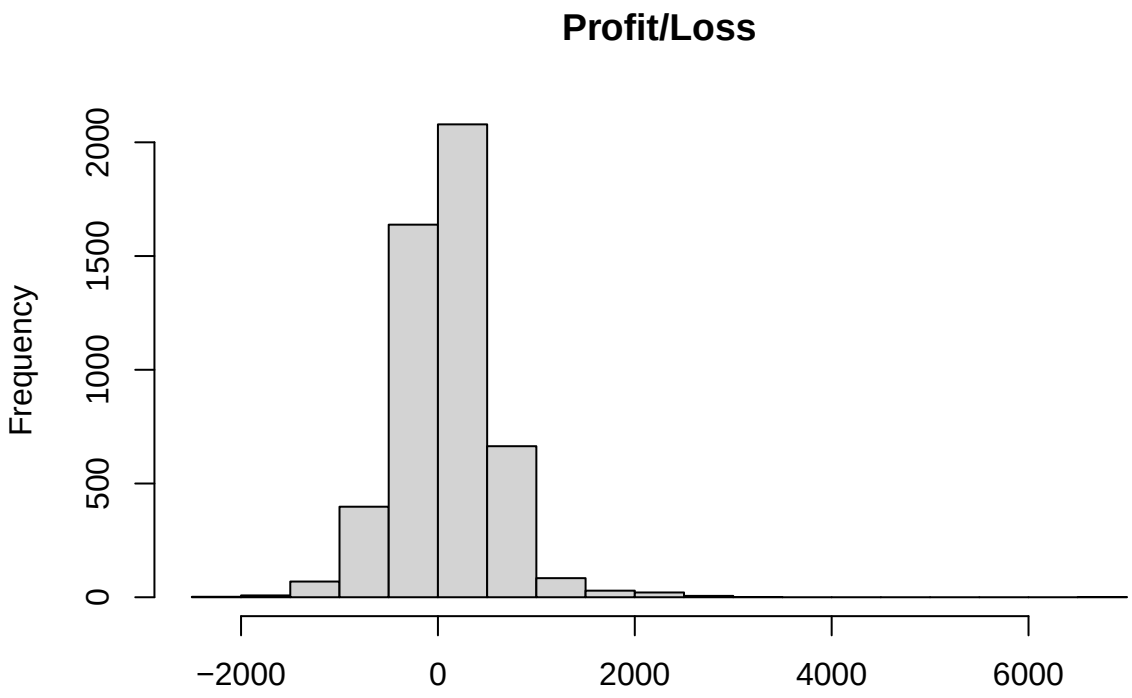
```
## pausing 1 second between requests for more than 5 symbols
## pausing 1 second between requests for more than 5 symbols
## pausing 1 second between requests for more than 5 symbols
```



We are running 5000, 20 trading day models. One of the models looks follows the graph seen below:



A summary of these 5000 models can be seen here:



The mean profit/loss was:

```
## [1] 86.0965
```

The 5% VaR was:

```
##        5%
## -689.4531
```

Question 4: Market Segmentation

Question 5: Author attribution

In this question, we need to build the best performing model to predict the author of an article on the basis of that article's textual content. We used Random Forests and Naive Base

We begin by loading in to the data using a for loop.

```
## Warning in tm_map.SimpleCorpus(train_Corpus, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(train_Corpus,
## content_transformer(removeNumbers)): transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(train_Corpus,
## content_transformer(removePunctuation)): transformation drops documents

## Warning in tm_map.SimpleCorpus(train_Corpus,
## content_transformer(stripWhitespace)): transformation drops documents

## Warning in tm_map.SimpleCorpus(train_Corpus, content_transformer(removeWords), :
## transformation drops documents
```

Second, we took some pre-processing/tokenization steps, including: 1) make everything lowercase 2) remove numbers 3) remove punctuation 4) remove excess white-space

```
## <<DocumentTermMatrix (documents: 2500, terms: 660)>>
## Non-/sparse entries: 224397/1425603
## Sparsity           : 86%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

Next we created DTM(doc-term-matrix) for the train data set. We have 2500 documents with 660 terms and a sparcity of 86%

```
## Warning in tm_map.SimpleCorpus(test_corpus, content_transformer(tolower)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(test_corpus, content_transformer(removeNumbers)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(test_corpus,
## content_transformer(removePunctuation)): transformation drops documents

## Warning in tm_map.SimpleCorpus(test_corpus,
## content_transformer(stripWhitespace)): transformation drops documents

## Warning in tm_map.SimpleCorpus(test_corpus, content_transformer(removeWords), :
## transformation drops documents

## <<DocumentTermMatrix (documents: 2500, terms: 660)>>
## Non-/sparse entries: 225031/1424969
## Sparsity           : 86%
## Maximal term length: 18
## Weighting          : term frequency (tf)

## [1] 0.2624

## [1] 0.5976
```

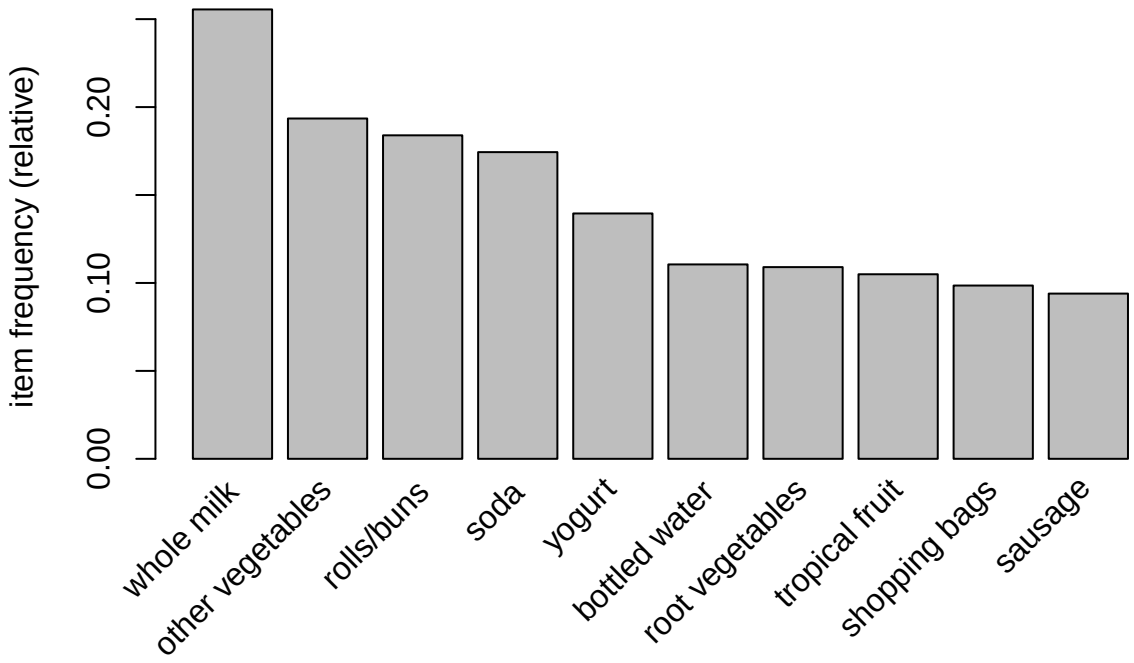
Finally, we pull the naive bayes analysis and get a 26% accuracy. This is not ideal compared to random forest. Random forest gets us an accuracy of 59% with 25 trees. This was our best model. Increasing or decreasing the number of trees only decreased accuracy. Based on this we think random forests is the best method for predicting the author of an article based on the articles textual content.

Question 6: Association Rule Mining

To start, we can take a small look into the dataset.

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
```

Here we can see the most frequent items bought, as well as the size of all the baskets in the dataset. Whole milk clearly dominates while Other Vegetables comes in a close 2nd.



Here we can see the frequency of the top 10 items found in carts. Whole milk comes up nearly 25% of the time in baskets with Other Vegetables coming in around 19%.

Moving forward to try and find association, we have a 125 rule set created by limiting support to 0.01 and confidence to 0.30.

```
## set of 125 rules
```

```
## set of 125 rules
##
## rule length distribution (lhs + rhs):sizes
##  2  3
## 69 56
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.000   2.000   2.448   3.000   3.000
##
## summary of quality measures:
##     support          confidence        coverage            lift
##  Min.   :0.01007   Min.   :0.3079   Min.   :0.01729   Min.   :1.205
##  1st Qu.:0.01149   1st Qu.:0.3454   1st Qu.:0.02888   1st Qu.:1.608
##  Median :0.01454   Median :0.3978   Median :0.03711   Median :1.789
##  Mean   :0.01859   Mean   :0.4058   Mean   :0.04783   Mean   :1.906
##  3rd Qu.:0.02217   3rd Qu.:0.4496   3rd Qu.:0.05663   3rd Qu.:2.155
##  Max.   :0.07483   Max.   :0.5862   Max.   :0.19349   Max.   :3.295
##      count
##  Min.   : 99.0
```

```
##  1st Qu.:113.0
##  Median :143.0
##  Mean   :182.8
##  3rd Qu.:218.0
##  Max.   :736.0
##
## mining info:
##  data ntransactions support confidence
##   df5          9835    0.01        0.3
```

Now we can see we have 69 times their is a correlation of one item to another, and 56 times we have 2 items that correlate to getting a 3rd item.

```
##      lhs                                rhs                support
## [1]  {citrus fruit,other vegetables}   => {root vegetables}  0.01037112
## [2]  {other vegetables,tropical fruit} => {root vegetables}  0.01230300
## [3]  {beef}                            => {root vegetables}  0.01738688
## [4]  {citrus fruit,root vegetables}    => {other vegetables} 0.01037112
## [5]  {root vegetables,tropical fruit}  => {other vegetables} 0.01230300
## [6]  {other vegetables,whole milk}     => {root vegetables}  0.02318251
## [7]  {curd,whole milk}                 => {yogurt}           0.01006609
## [8]  {rolls/buns,root vegetables}      => {other vegetables} 0.01220132
## [9]  {root vegetables,yogurt}          => {other vegetables} 0.01291307
## [10] {tropical fruit,whole milk}       => {yogurt}           0.01514997
##      confidence coverage   lift     count
## [1]  0.3591549  0.02887646 3.295045 102
## [2]  0.3427762  0.03589222 3.144780 121
## [3]  0.3313953  0.05246568 3.040367 171
## [4]  0.5862069  0.01769192 3.029608 102
## [5]  0.5845411  0.02104728 3.020999 121
## [6]  0.3097826  0.07483477 2.842082 228
## [7]  0.3852140  0.02613116 2.761356  99
## [8]  0.5020921  0.02430097 2.594890 120
## [9]  0.5000000  0.02582613 2.584078 127
## [10] 0.3581731  0.04229792 2.567516 149
```

So with this, I separated the ruleset by making the top rules those with higher lift. Lift being the likelyhood of someone buying the item because of the items they already have compared to the base rate. Those who buy citrus fruit and other vegetables are 3 times more liikely to buy root vegetables. What is interesting here is the 3rd rule. All of the rules of the top 10 by lift are rules that require at least 2 items before considering the 3rd, however the 3rd highest rule is only a one for one correlation. These shoppers are 3 times more likely to buy root vegetable when the person also buys beef. This makes sense if you think about how meat is commonly prepared, but the fact it beats out other teams of 2 is interesting. Another interesting element to this data is the infrequency of whole milk here. Whole milk was found to be the most common item by far, however in this comparison it only shows up in 3 combos. I had assumed milk would be all over the place, but in reality when it comes to association, people just seem to buy milk no matter what. Whereas the second highest item in terms of raw frequency is actually much more common to find in associations.