STA 380: Intro to Machine Learning Pt.2

Anthony Huang

8/13/2021

Visual story telling part 1: green buildings

The "Excel Guru"'s analysis has several oversights that ultimately makes his forecast unconvincing. Since the analysis is based on the median rent of all buildings, the rent variable is confounded with multiple variables that are not accounted for. Size of the building, occupancy rate, location, and class are all features that may affect the rent level, which are not considered in the "Excel Guru"'s analysis.

We first check whether the cleaning the outliers from the data set is warranted by checking whether the "Excel Guru"'s theory is valid. The claim is that the buildings with leasing rate of lower than 10% are ones that have something weird going on and could potentially distort the analysis. We find that the staff cleaned out 215 buildings out of 7894, which does not affect the data set too much.

Summary of Rent (Green Buildings) after Data Cleaning:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.50   27.60   30.03   35.54  138.07
```

Summary of Rent (Non-Green Buildings) after Data Cleaning:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.43   25.03   28.44   34.18  250.00
```

Summary of Rent (Green Buildings) of entire dataset:

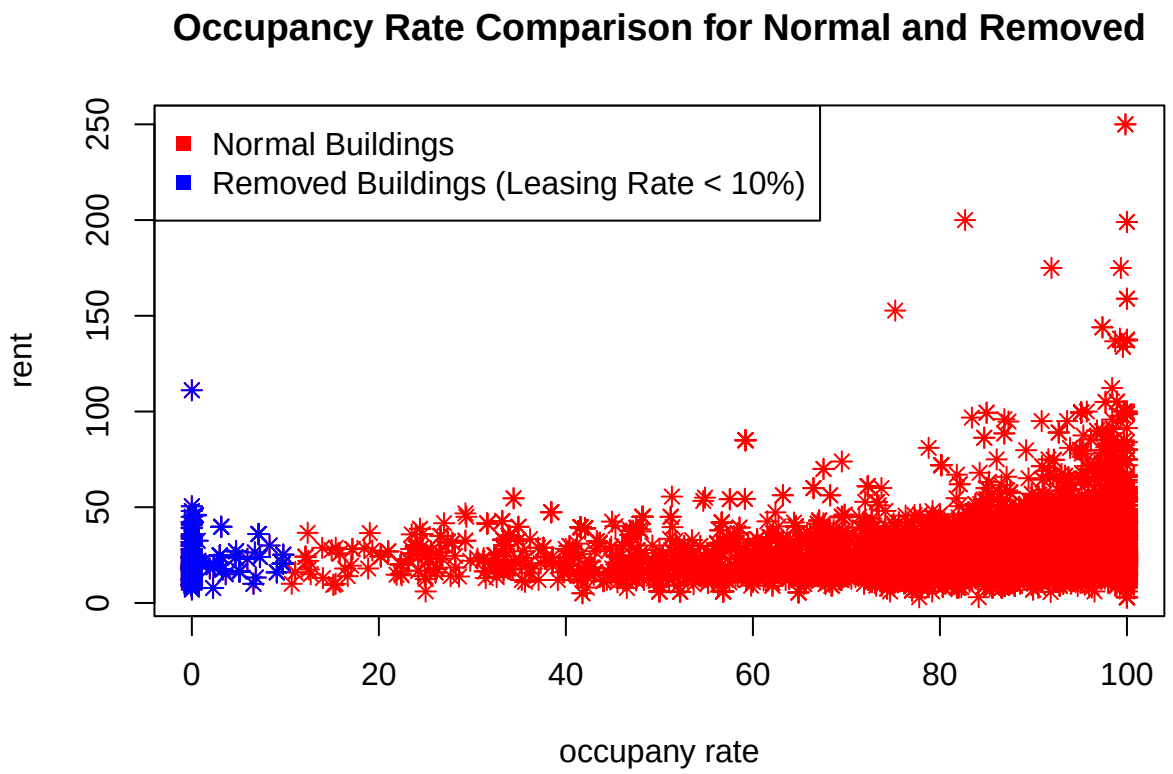
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.50   27.60   30.02   35.50  138.07
```

Summary of Rent (Non-Green Buildings) of entire dataset:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.18   25.00   28.27   34.00  250.00
```

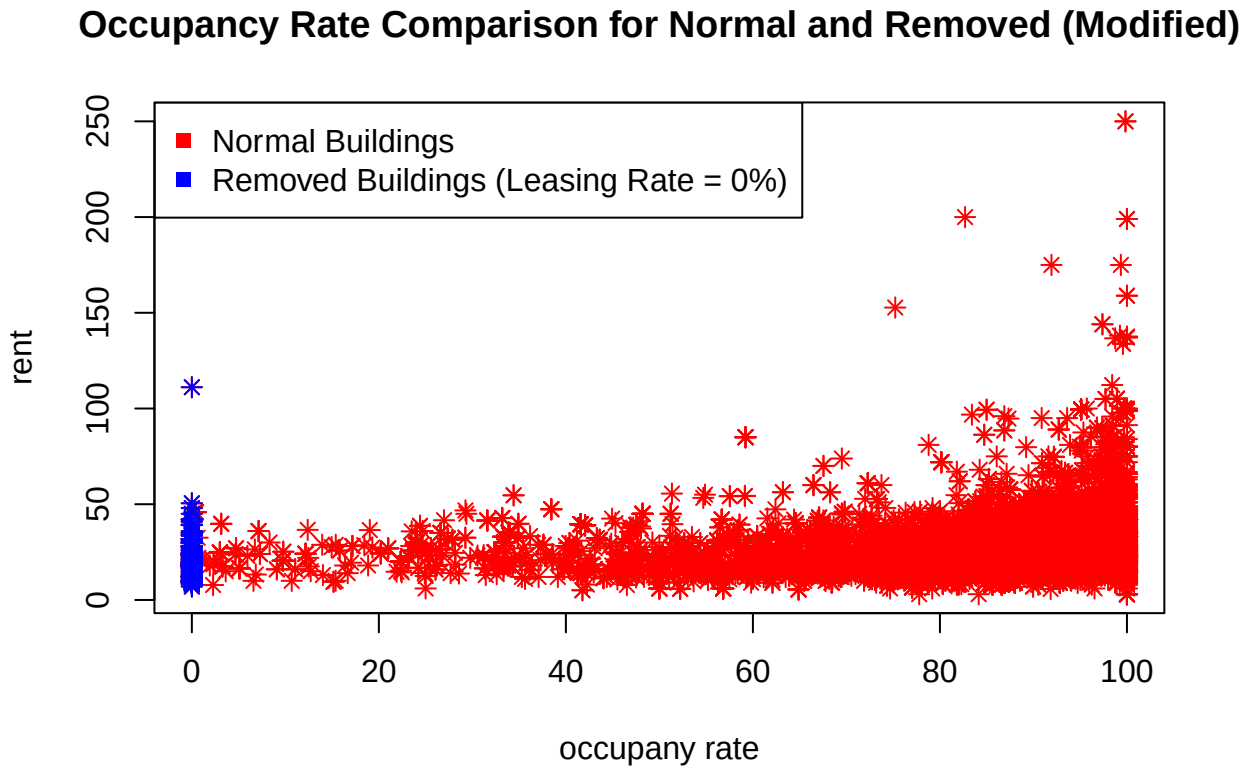
After checking the summary of rent for pre-data cleaning and after data cleaning, we find that removing the outliers does not change the mean by much. This may suggest that the data cleaning step was useless, and other actions should be taken. Let's try to plot occupancy rate versus rent to see if there are obvious outliers on the graph.

```
## integer(0)
```

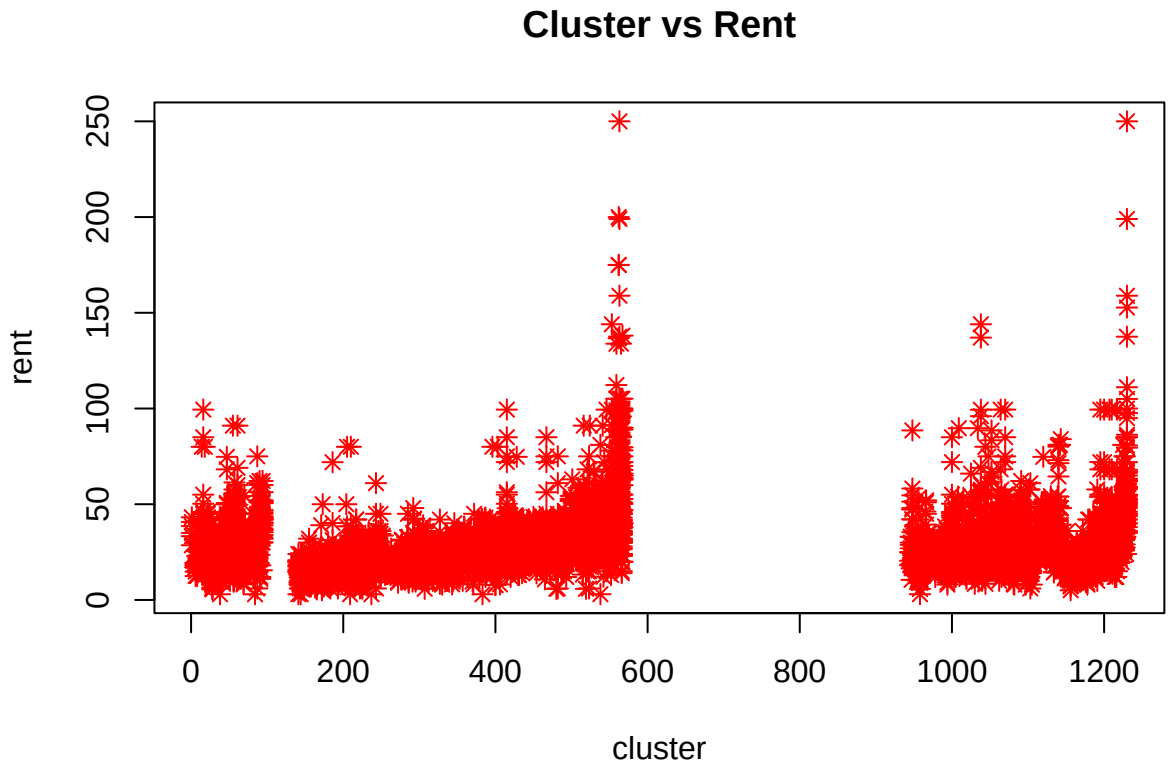


There are still several outliers not removed from the data set, which are buildings with rent higher than 100. On the other hand, some data points removed might not be outliers, which may cause the data to be skewed. Based on the pattern in the graph, instead of occupancy lower than 10%, removing data points with occupancy rate of 0% might result in a more accurate dataset.

```
## integer(0)
```

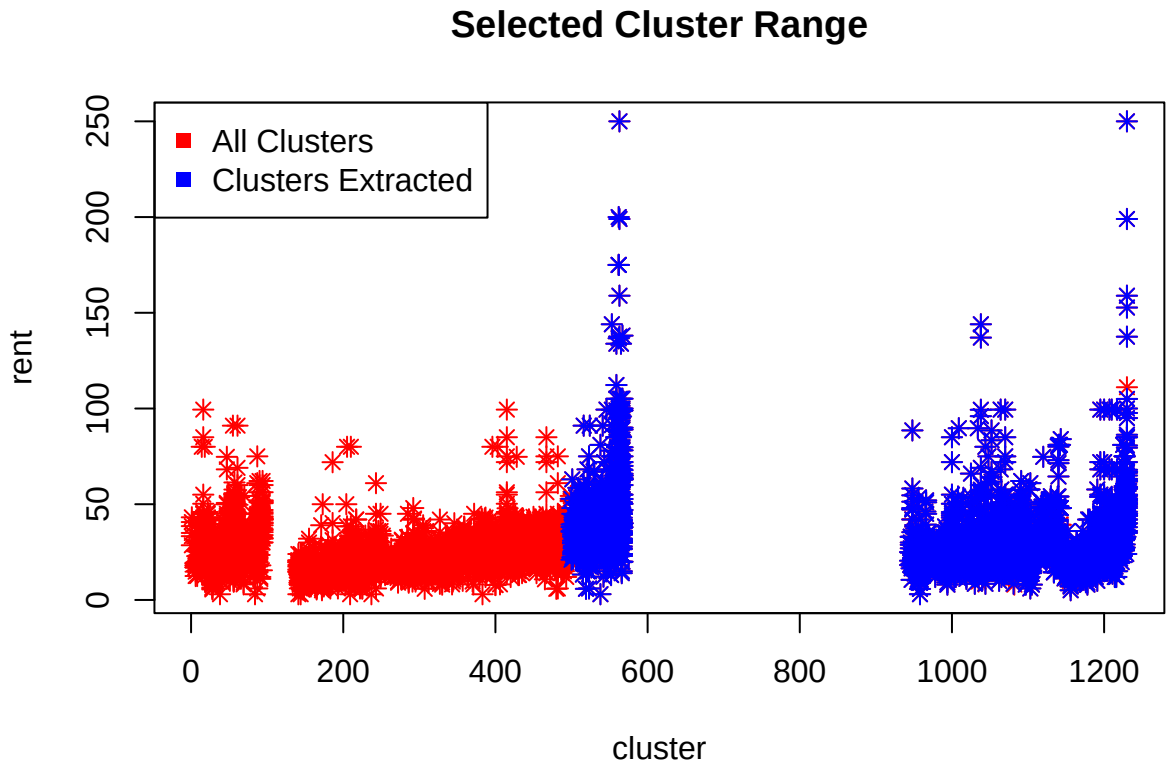


Additionally, other confounding variables may also affect the per-square-foot rent, instead of only buildings with zero occupancy rates. Cluster number determine the location the building is located in. By plotting cluster versus rent, we find that some clusters are in more luxurious area, and these clusters have a higher ceiling compared to areas that cost much less.



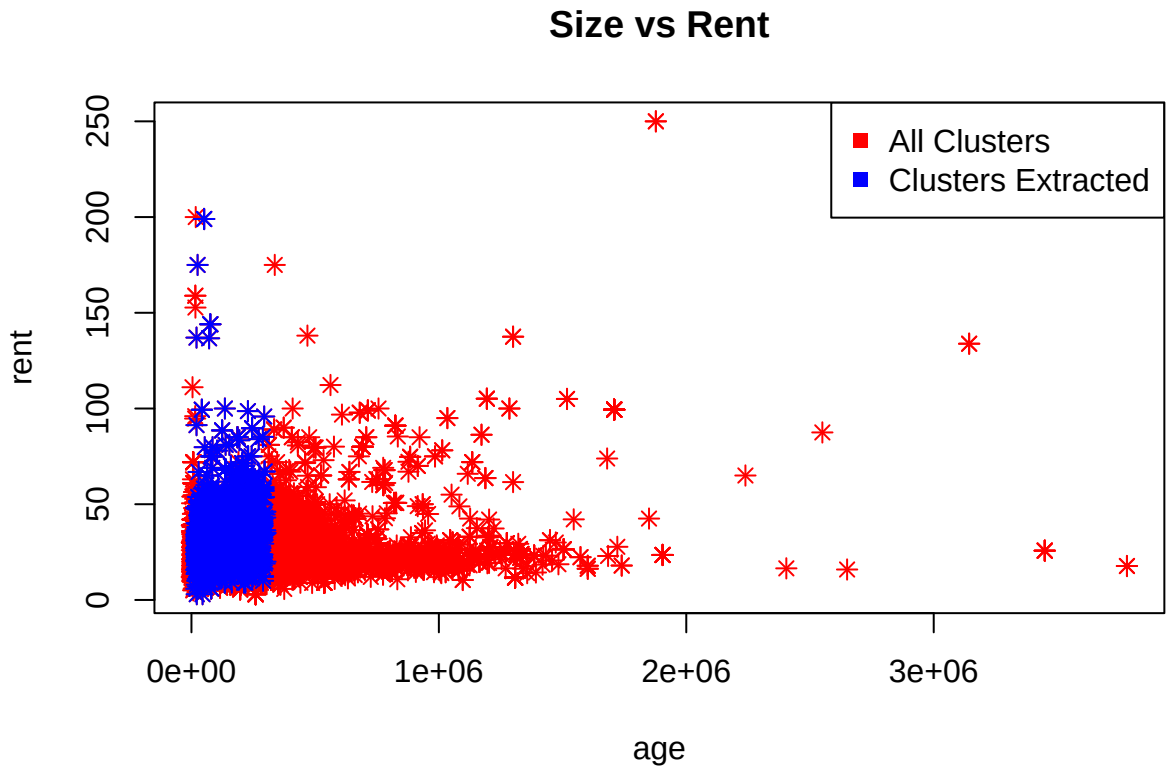
Since the building is projected to be built in East Cesar Chavez, just across I-35 from downtown, the average rent should be fairly high. The cluster value identifier should be in the region between 500-600 or 1000-1200. Now let's visualize the data points we will extract from the data set to get one that has similar features to our project.

```
## integer(0)
```



There also seems to be a correlation between rent and size of the building. The projected building size will be 250,000 square feet, so we will extract data points that are close to the level.

```
## integer(0)
```



Now we can extract the data points based on the features selected to minimize confounding variables that may affect the average rent.

Summary of Rent (Green Buildings) in new dataset:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.25   25.27   35.62   37.15   42.83   98.65
```

Summary of Rent (Non-Green Buildings) in new dataset:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   24.50   35.40   36.84   45.00   89.70
```

Previously, the first report concluded that there is a 2.7 dollars per square foot difference on average, while in the new dataset with lowered bias, the difference is very minimal. If we look at median there is a 0.22 dollars per square foot difference, whereas if we look at mean there is a 0.31 dollars per square foot difference. If we calculate the extra revenue earned based on the "Excel Guru"'s calculations, the green building provides 250000 x 0.22 = 55,000 dollars extra revenue. 100 million * 5% / 55,000 dollars will result in roughly 90 years to repay the 5% premium. However, the extra revenue earned from rent is not the only positives of going green - we can also explore the difference in energy, water, and waste disposal costs between green and non-green buildings.

Summary for Gas Costs(Green Building) in new dataset:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00950 0.01030 0.01030 0.01084 0.01052 0.01450
```

Summary for Gas Costs(Non-Green Building) in new dataset:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.009487 0.010296 0.010296 0.011600 0.012774 0.028914
```

Summary for Electricity Costs(Green Building) in new dataset:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01820 0.02890 0.03780 0.03339 0.03780 0.04550
```

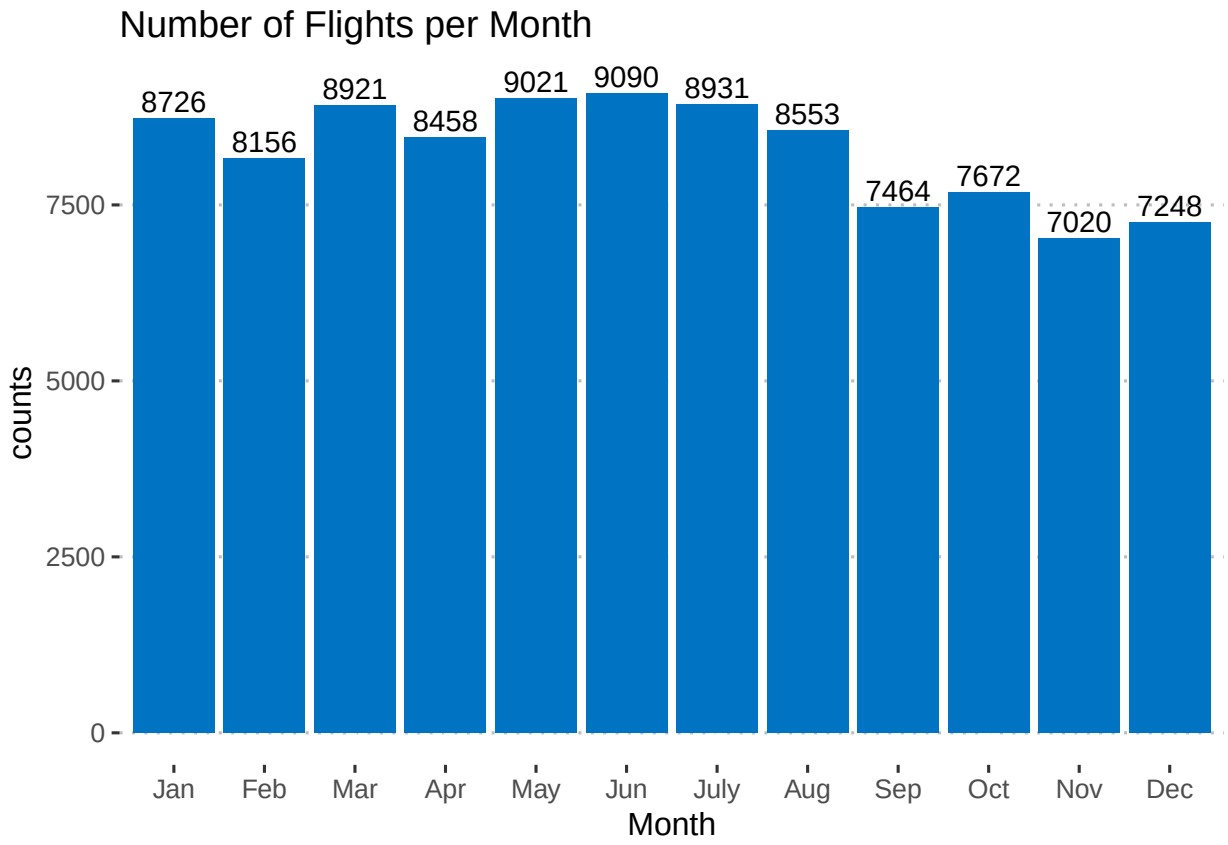
Summary for Electricity Costs(Non-Green Building) in new dataset:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01820 0.02351 0.03274 0.03213 0.03781 0.06278
```

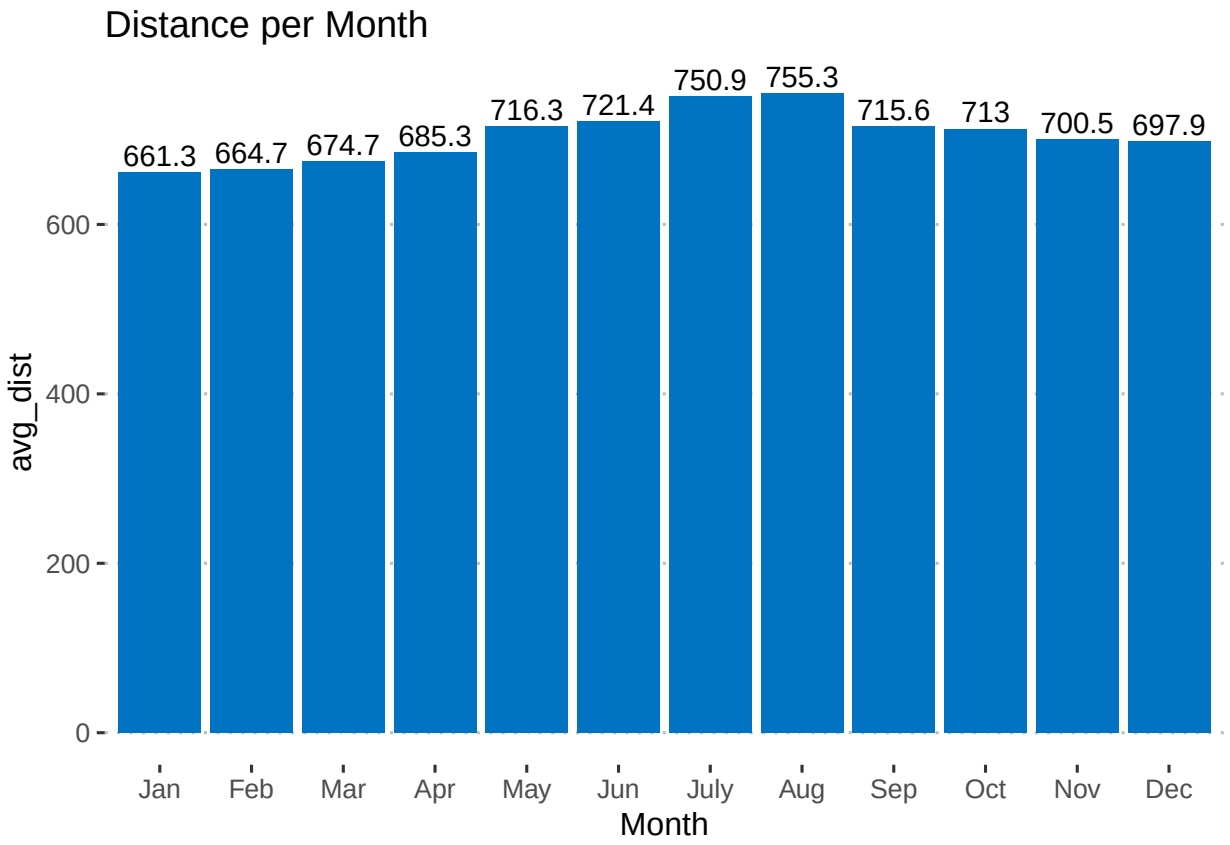
Surprisingly, the average costs for electricity and gas for green and non-green buildings are approximately the same, which may be due to the geographic region being not entirely green or non-green. These summary outputs suggest that the data set might not be fit for predicting whether green buildings save more money or earn more revenue than non green buildings. Features that are specific to green and non-green buildings should be included for a better comparison, such as gas costs for the specific building instead of a cluster.

Visual story telling part 2: flights at ABIA

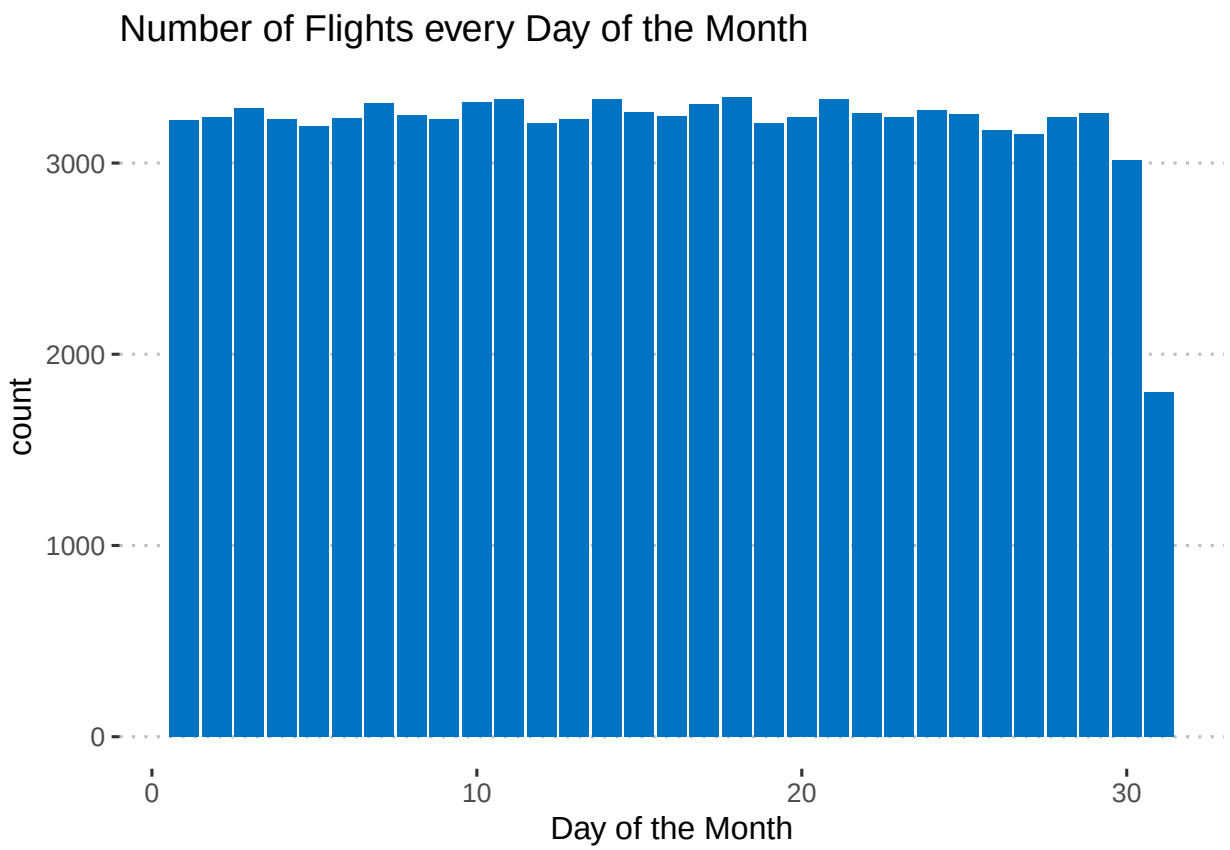
First, we take a look at simple plots that show the frequency of flights of each month, day of month, and day of week to find most common flight periods.



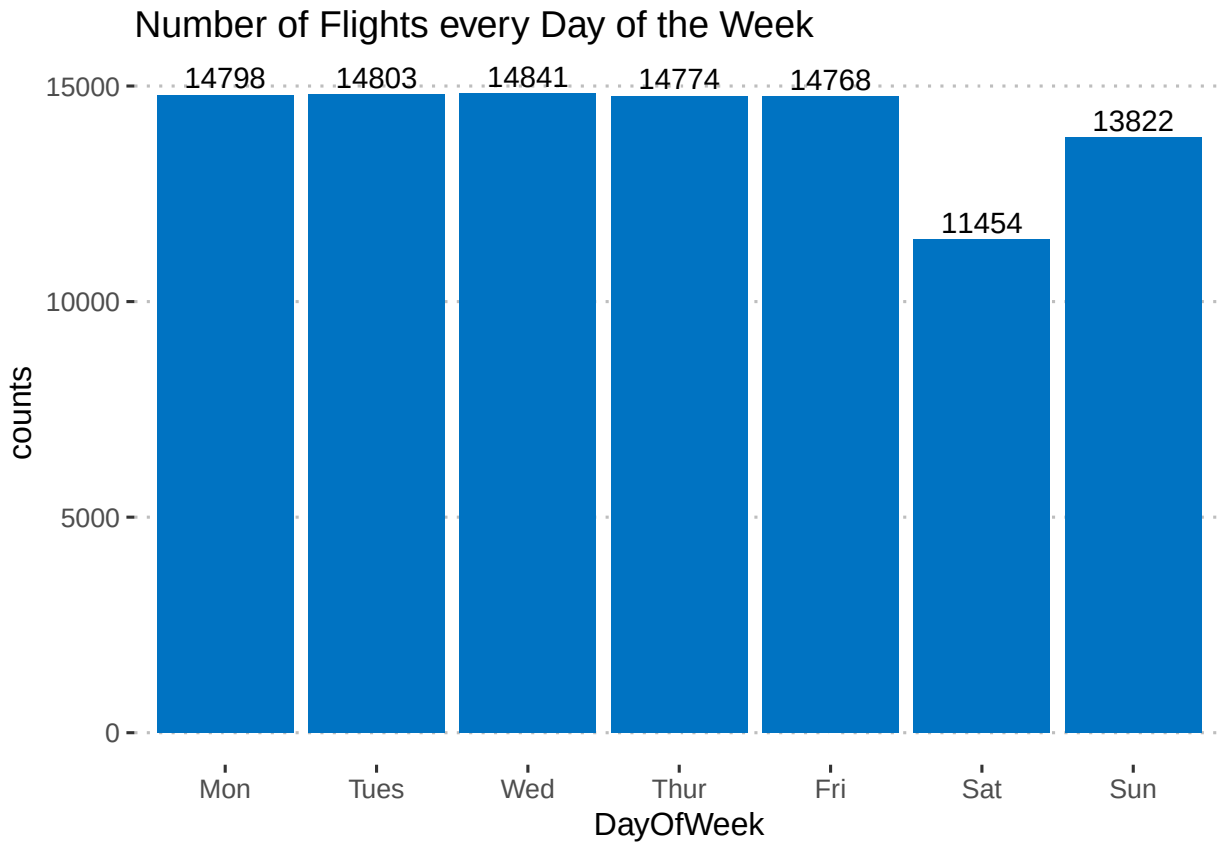
The graph shows an influx of flights mid-year, while there is a decrease in flights at the end of the year.



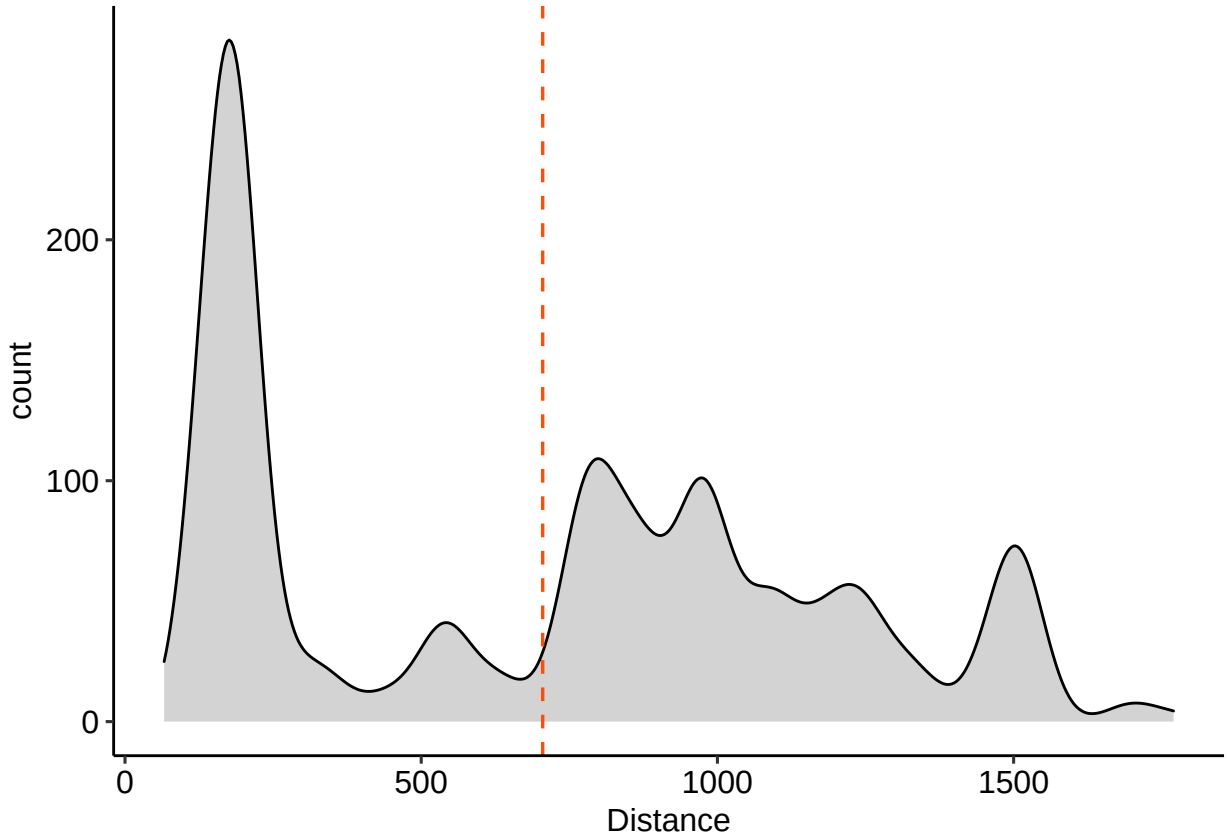
The chart shows that May to August, on average, have more long distance flights than other months.



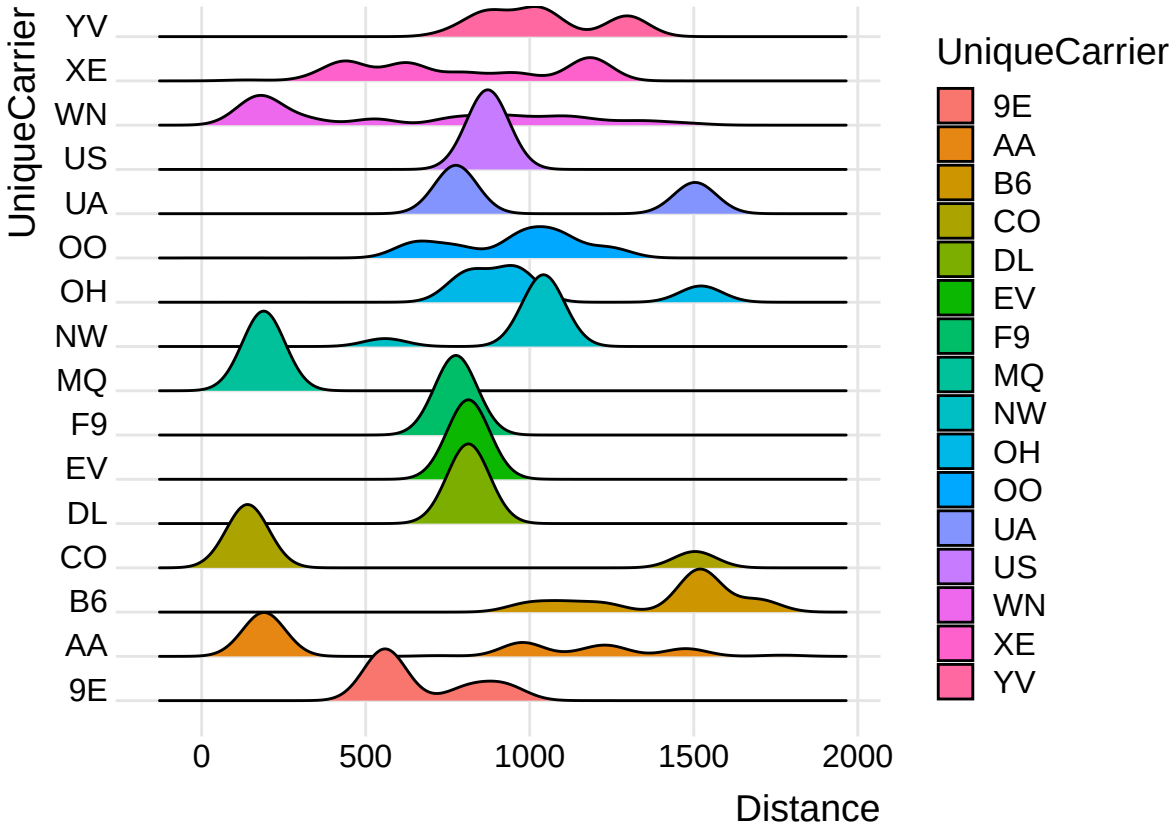
The graph shows a sharp drop in the frequency of flights at the end of each month.



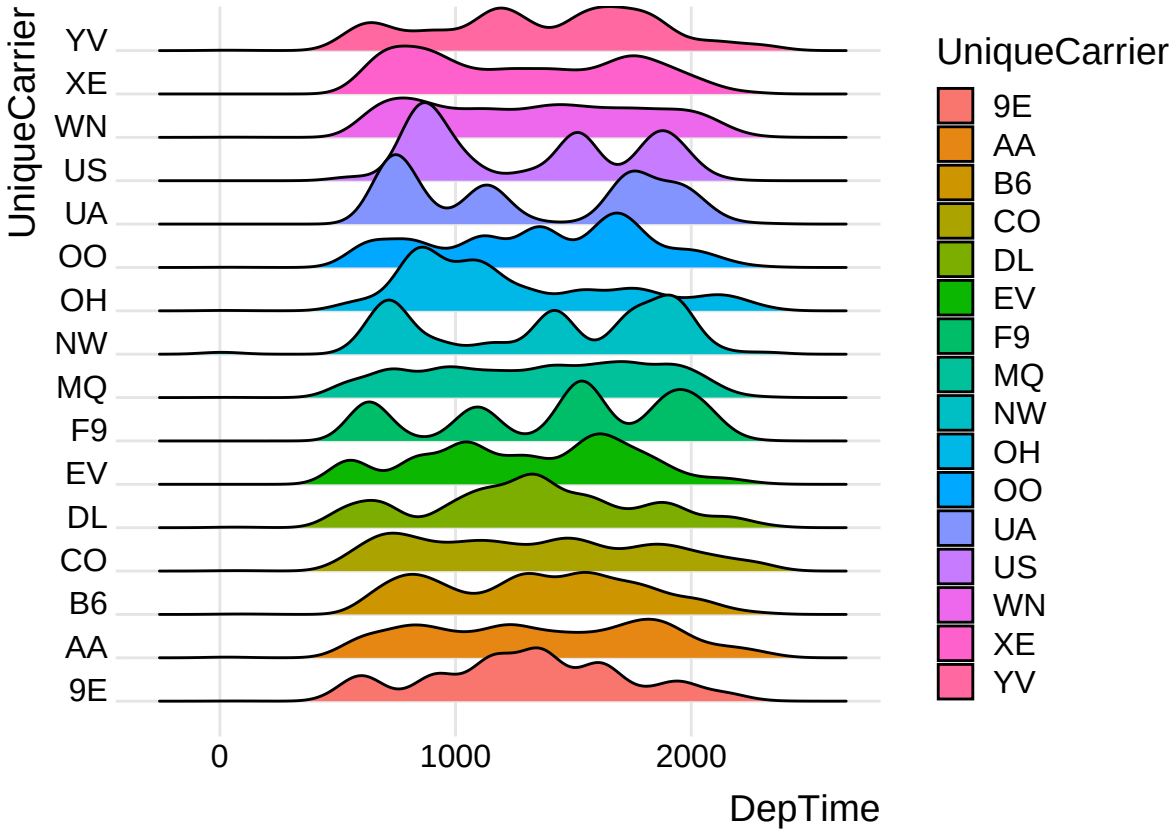
The graph shows a steady rate of flights during the week days, and a lower number on weekends. This is perhaps due to the amount of business travelers.



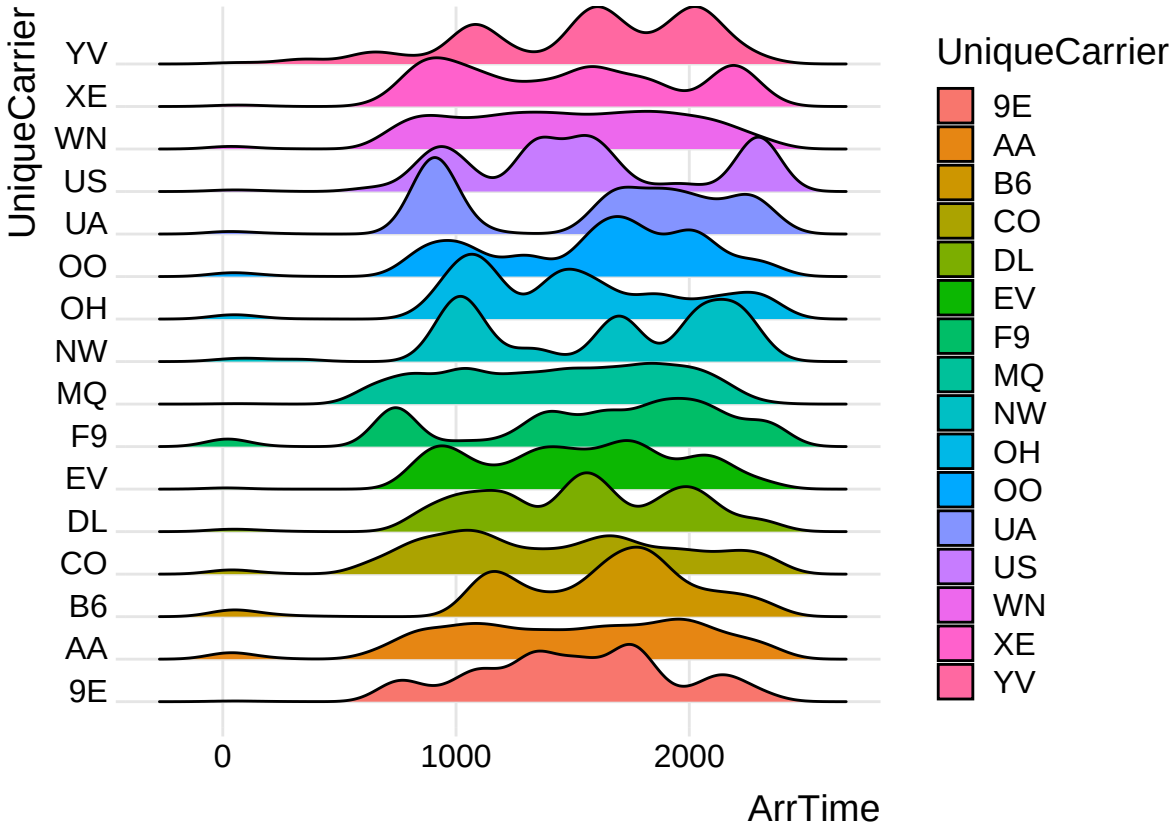
The density chart shows a high number of short distance flights around 150 miles to 250 miles, and drops sharply to around 400-750 miles. The latter half of the chart shows that there is also a high number of long distance flights that spread from 750 to 1750 miles.



Lastly, we find that the each Unique carrier are specific to a range of distance. For example, MQ specializes in short-distance flights, while B6 only has long distance flights. CO, on the other hand, has both short-distance flights and long-distance flights.

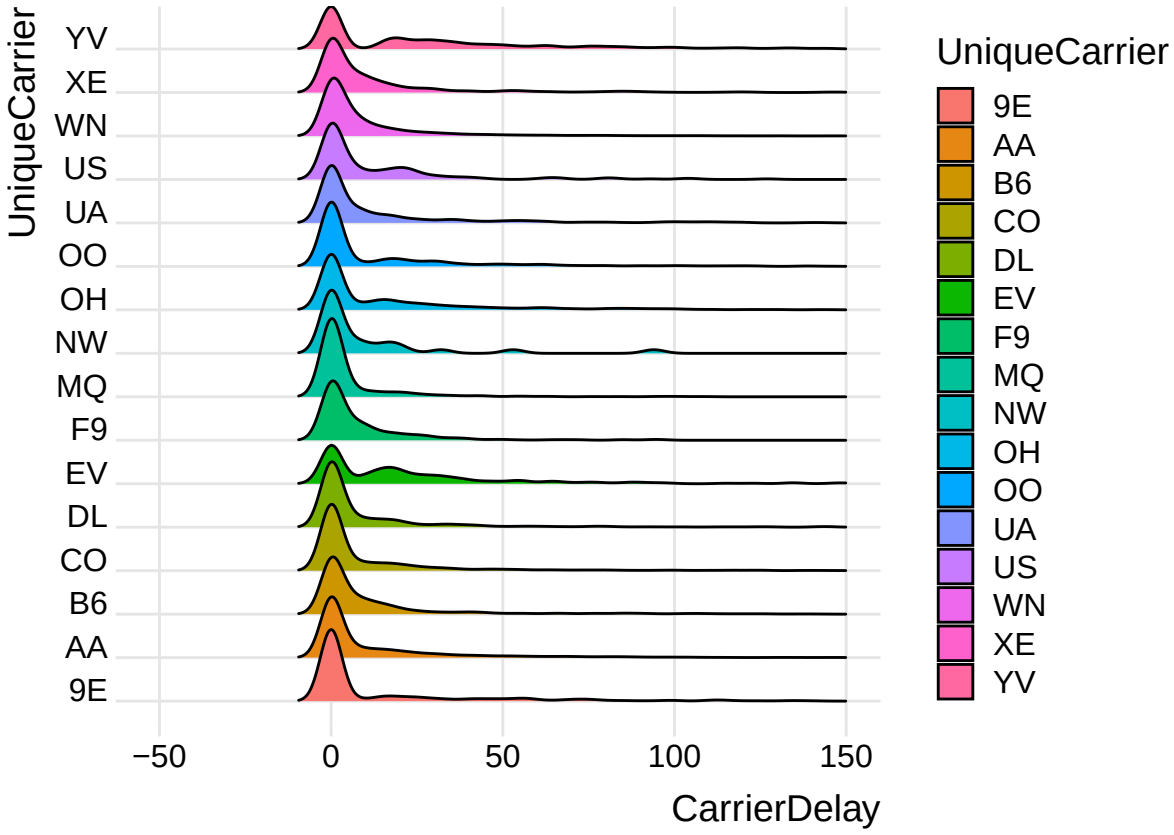


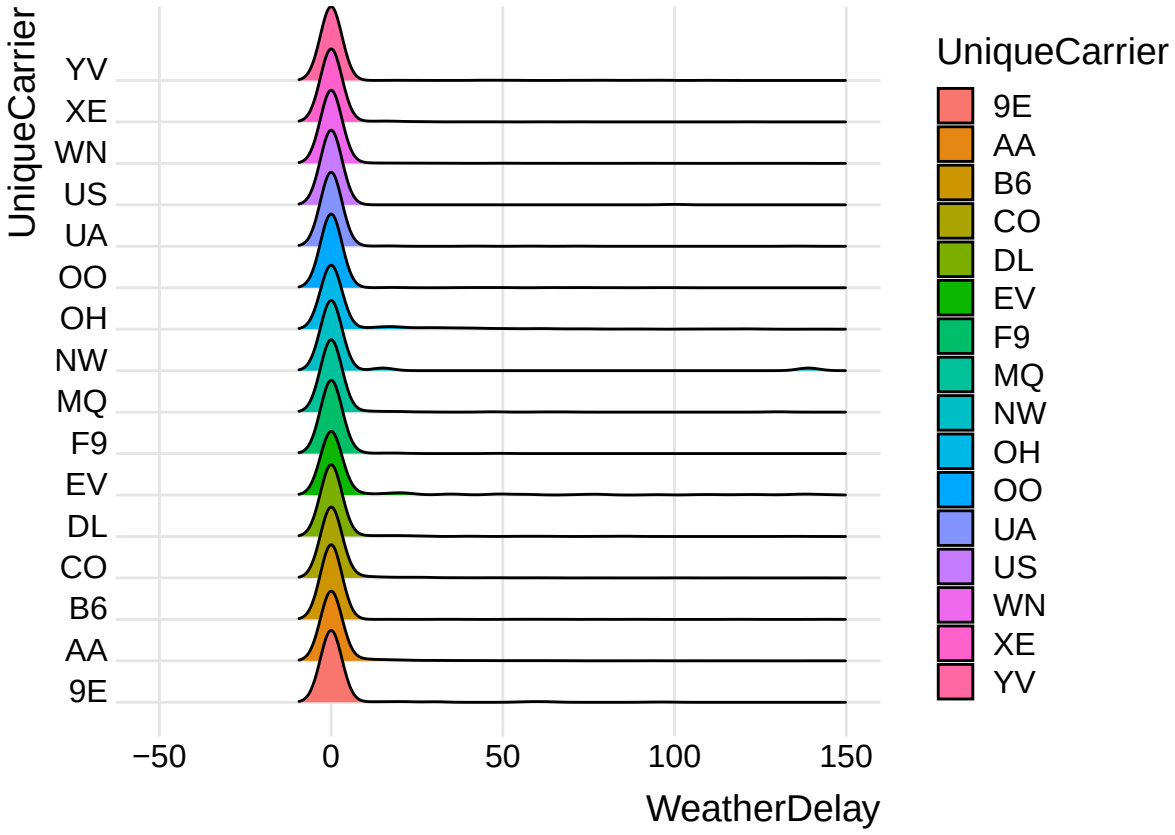
The departure time spread is also different for each unique carrier. Some have departure time centered around 13:00 (9E), while others may have 4 prime departure times at 8:00, 11:00, 15:00, and 19:00 (F9).

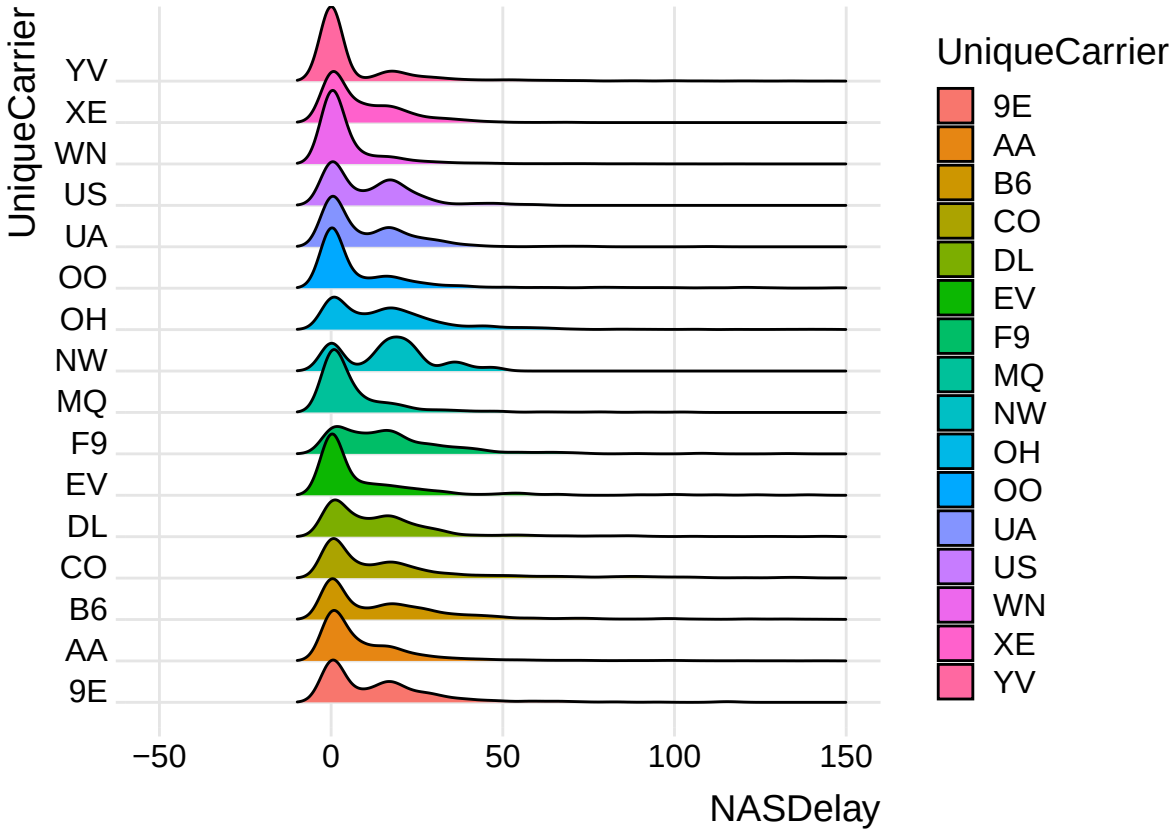


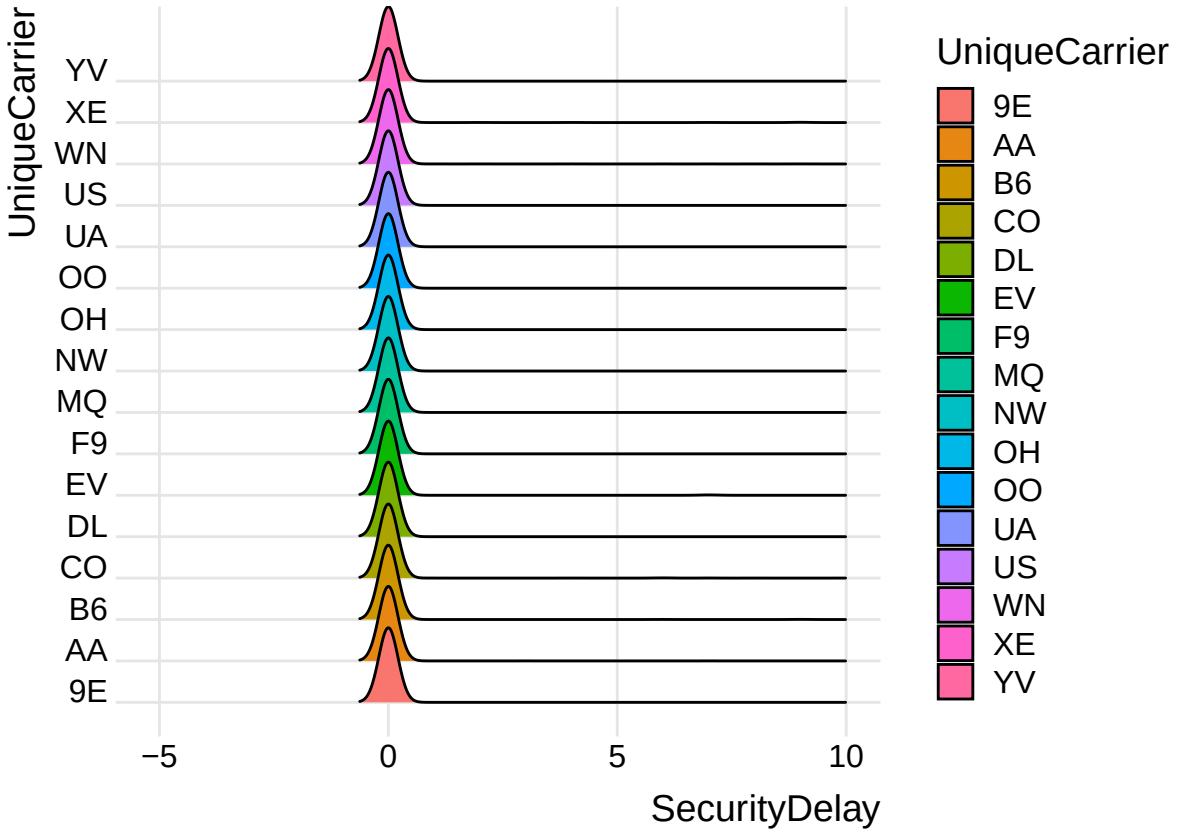
The arrival time spread for each unique carrier is similar to the departure time spread, with a major difference in the midnight time slot. Some unique carriers do not have flights that arrive at midnight, while others, such as F9 and B6, have slightly more flights that do.

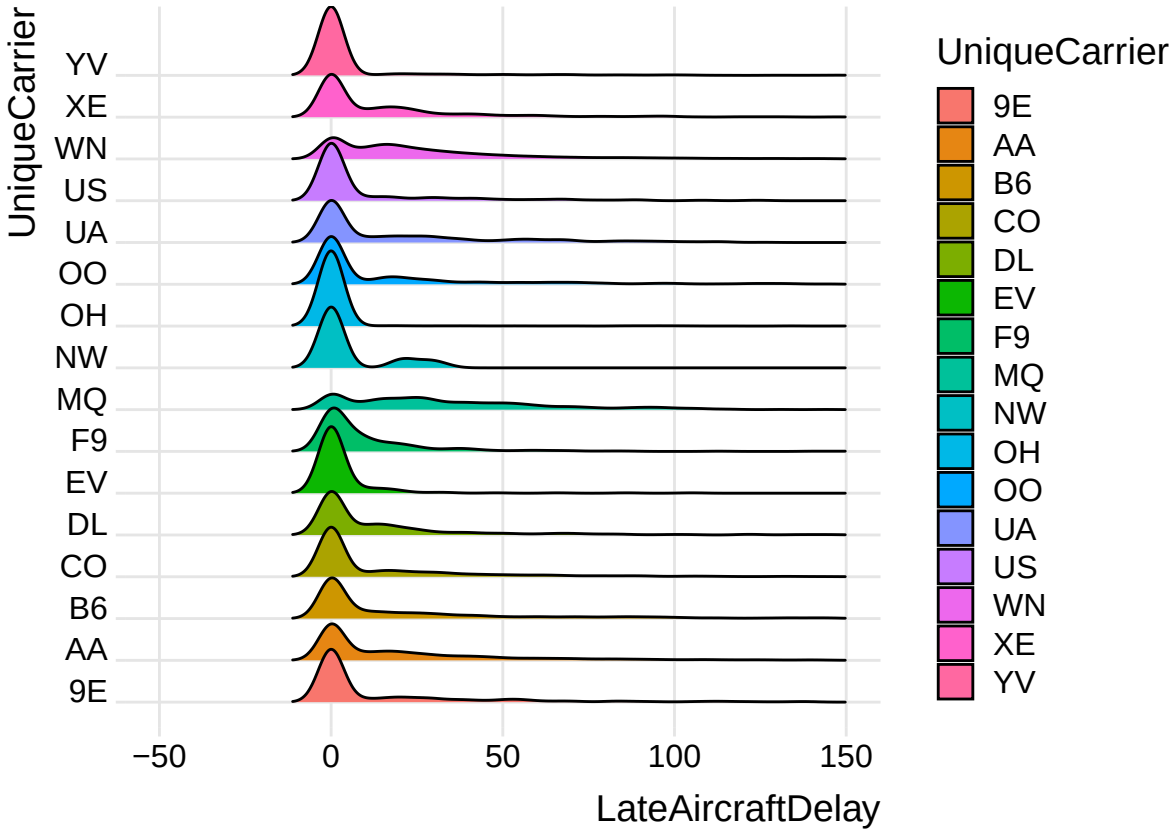
In the following graphs, we can observe the difference in delay time that is caused by Carrier, Weather, NAS, Security, and Late Aircraft


UniqueCarrier
YV
XE
WN
US
UA
OO
OH
NW
MQ
F9
EV
DL
CO
B6
AA
9E
−50
0
50
100
150
CarrierDelay
UniqueCarrier
9E
AA
B6
CO
DL
EV
F9
MQ
NW
OH
OO
UA
US
WN
XE
YV


UniqueCarrier
YV
XE
WN
US
UA
OO
OH
NW
MQ
F9
EV
DL
CO
B6
AA
9E
−50
0
50
100
150
WeatherDelay
UniqueCarrier
9E
AA
B6
CO
DL
EV
F9
MQ
NW
OH
OO
UA
US
WN
XE
YV


UniqueCarrier
YV
XE
WN
US
UA
OO
OH
NW
MQ
F9
EV
DL
CO
B6
AA
9E
−50
0
50
100
150
NASDelay
UniqueCarrier
9E
AA
B6
CO
DL
EV
F9
MQ
NW
OH
OO
UA
US
WN
XE
YV


UniqueCarrier
YV
XE
WN
US
UA
OO
OH
NW
MQ
F9
EV
DL
CO
B6
AA
9E
−5
0
5
10
SecurityDelay
UniqueCarrier
9E
AA
B6
CO
DL
EV
F9
MQ
NW
OH
OO
UA
US
WN
XE
YV



There does not seem to be much of a difference for weather and security delay, as these are external factors not caused by the unique carriers. As for the other reasons, we find that ,YV and EV have the largest spread for Carrier delay, NW and F9 have the largest spread for NAS delay, MQ has the largest spread for Late Aircraft delay. It would probably be the best to avoid these carriers!

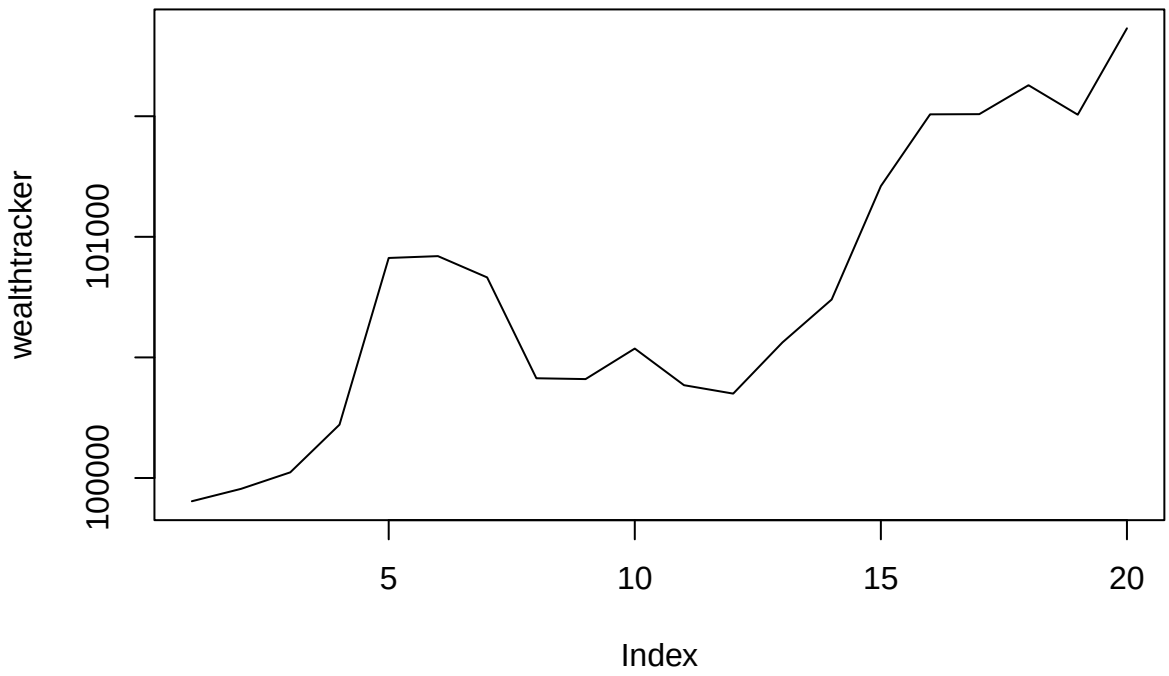
Portfolio Modeling

Portfolio 1: Risk Averse Portfolio

In the first portfolio, only low risk low yield funds will be included. The funds will be chosen from corporate and government bonds funds and mortgage backed securities funds, which are generally considered safe assets. The weights of each fund will also be equal, at 20% for each fund.

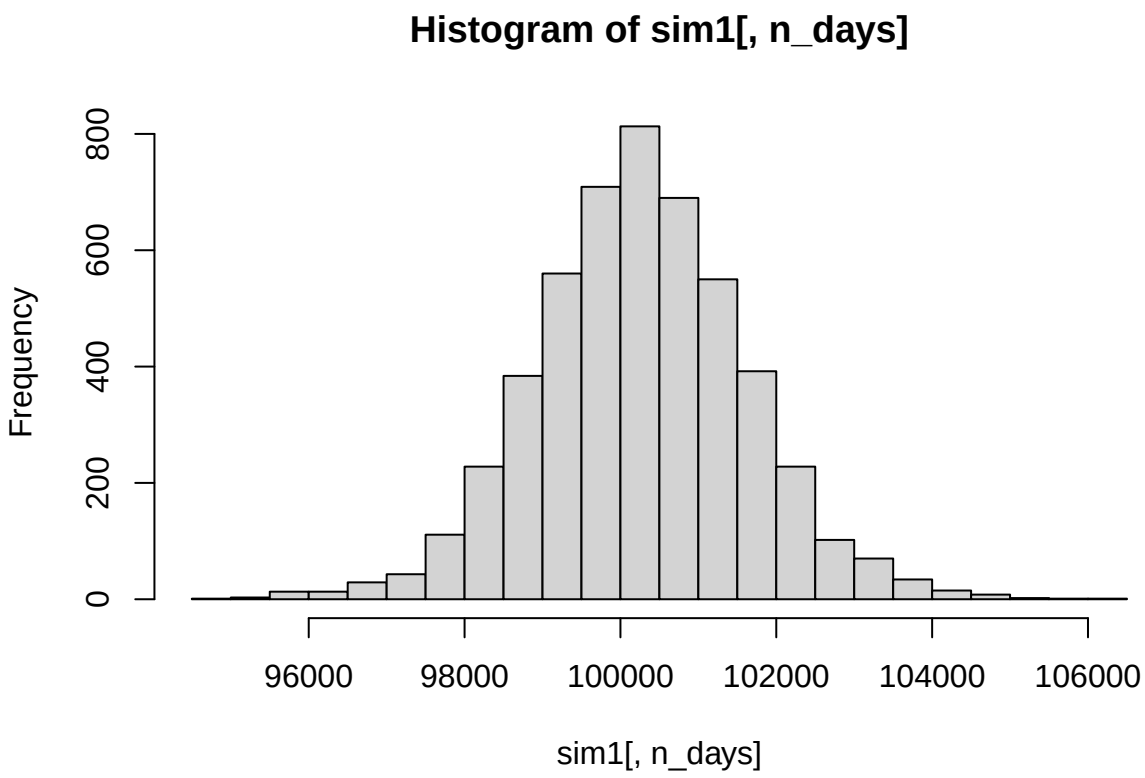
The following graph shows a sample run over 4 weeks, which shows a steady return.

```
## [1] 101864.4
```



The following histogram shows a simulation of over 5000 runs, with a mean slightly above the initial wealth at 100,111 dollars. For the risk levels, the standard deviation is 1375.43 dollars and the 5% VaR is 1997.9 dollars, which shows that less than 5% of the time the wealth will drop to that level.

```
##              [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## result.1 100410.30 100857.23 100731.22 100787.24 100246.57 100135.97 100434.22
## result.2  99575.75  99766.51  99729.68  99512.34  99521.23  99878.79  99947.56
## result.3  99751.83  99655.14  99844.59  99952.37  99171.59  98803.23  98920.02
## result.4  99886.13  98478.89  98536.00  98492.69  98185.98  97955.95  97753.33
## result.5  99788.37  99583.82  99437.49  99321.46  99679.77  99882.71  99786.41
## result.6 100207.92  99721.87  99717.02  99563.02  99814.31 100133.34 100277.19
##              [,8]      [,9]     [,10]     [,11]     [,12]     [,13]     [,14]
## result.1 100408.74 100771.47 100595.65 101211.30 101425.69 101443.91 101436.56
## result.2 100080.08 100187.02 100012.43 100079.92 100332.50 100418.14 100543.19
## result.3  99237.89  99021.13  99142.27  99026.59  98917.34  99189.05  98962.09
## result.4  97497.94  97614.16  97417.74  97319.10  97380.21  97258.12  97088.43
## result.5  99282.57  99038.18  98863.04  98920.99  98909.37  98492.95  98488.48
## result.6 100284.16 100476.84 100365.08 100488.32 100720.27 100889.55 101050.61
##             [,15]     [,16]     [,17]     [,18]     [,19]     [,20]
## result.1 101470.85 101897.03 101756.93 101616.20 101456.99 101487.59
## result.2 100221.65 100322.23 100212.16 100456.30 100567.95 100811.50
## result.3  98593.78  98980.96  99086.37  98953.85  98645.70  98834.19
## result.4  96967.68  96931.89  96837.94  97280.42  97436.99  97104.45
## result.5  98469.24  98630.08  98580.45  98625.14  98506.80  98646.22
## result.6 100902.13 100835.33 100782.83 100915.57 100947.95 101083.53
```




```
## [1] 1352.037
```

```
##        5%
## -1899.631
```

Portfolio 2: Semi-Risk Averse Portfolio

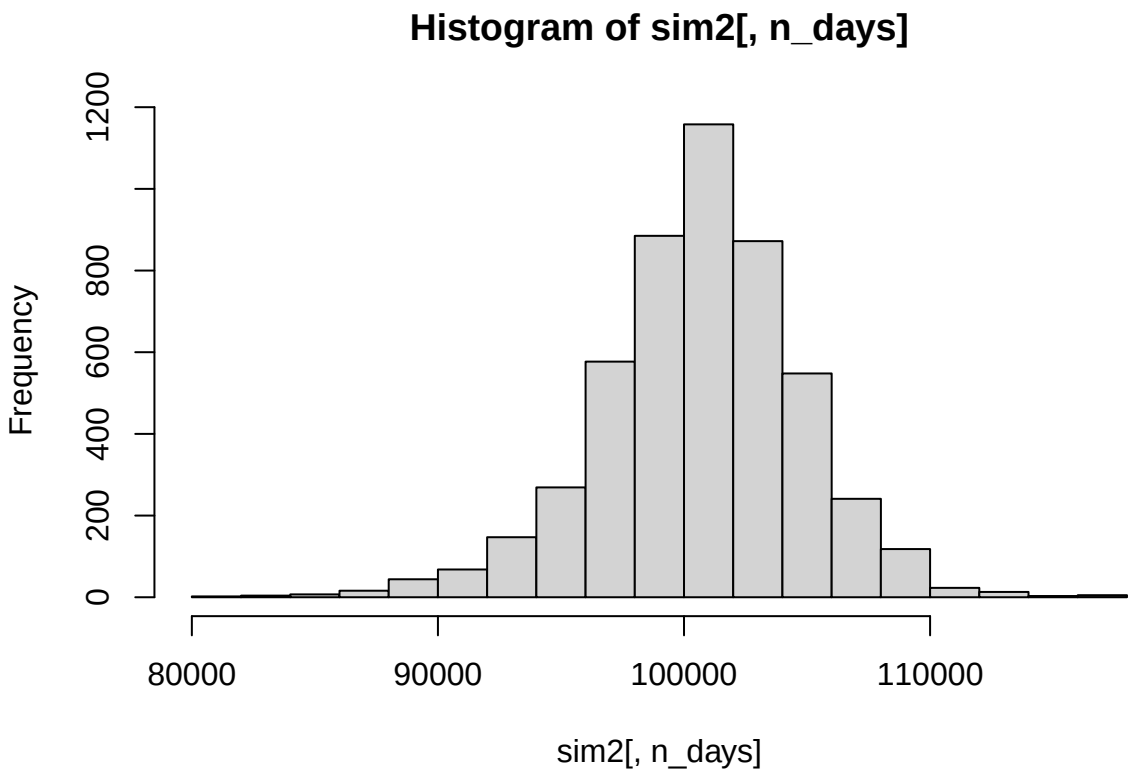
In the second portfolio, both low risk and high risk funds will be included. The portfolio will strive to be more diversified to reduce volatility and ensure there is good return. 7 funds will be chosen from fixed income, equities, commodities, and real estate funds, which the weights of each will be split based on sector (25% fixed income, 25% equities, 25% commodities, 25% real estate).

The sample run in the following graph shows a high return, possibly due to the higher risk involved in this portfolio.

```
## [1] 100594.5
```

The average return over 5000 simulations for medium risk portfolio, which is 100,531 dollars, is higher than that of low risk portfolio. The standard deviation, however, is much higher at 4152.93, and similarly the 5% VaR is at 6678.28, which is over 6 percent of initial wealth.

```
##              [,1]      [,2]     [,3]      [,4]      [,5]      [,6]      [,7]
## result.1 100163.73  99452.18 98712.01  98632.69  98172.53  98065.44  97695.60
## result.2 100025.70 100776.13 99854.99  99970.54 106589.47 107071.53 107492.58
## result.3  99838.82  99719.87 99512.79 100125.44 100343.80 100841.13 100339.71
## result.4  99461.48  99237.34 99397.06  99712.50  99531.06  99218.55  99861.00
## result.5  99726.48  99298.01 99681.09 100156.41  99626.39  99336.29  99923.52
## result.6  99696.63  99642.78 99602.64  99936.15 100051.80 100234.12  98506.48
##              [,8]      [,9]     [,10]     [,11]     [,12]     [,13]     [,14]
## result.1  97675.34  97954.40  98827.82  98447.56  98342.74  98471.72  98493.67
## result.2 107299.50 107607.45 106843.03 106205.24 106478.24 106606.61 107476.27
## result.3 101901.51 100107.73  99158.46  99419.55  98947.76  99425.49  99429.73
## result.4 100313.64 100791.97 100834.39 100497.06 100848.53 100991.20 101653.19
## result.5 101089.97 100060.51 100286.76 101063.06 101016.23 100888.25 101047.16
## result.6  99551.31  99768.41  99731.19 100100.75 100578.06 100426.52 100783.48
##             [,15]     [,16]     [,17]     [,18]     [,19]     [,20]
## result.1  98456.93  97192.73  96733.73  97362.24  97242.81  98817.23
## result.2 107282.61 108176.53 107981.67 107899.25 108977.91 108565.96
## result.3  97811.20  97880.35  98181.81  98319.06  97854.47  98217.54
## result.4 101467.57 101914.13 101902.01 102406.11 102251.82 102638.50
## result.5 100112.22 100653.16 100787.81 101244.64 101430.32 102188.60
## result.6 100612.58 100865.20 101886.21 102422.00 105829.93 106497.82
```




```
## [1] 4115.593
```

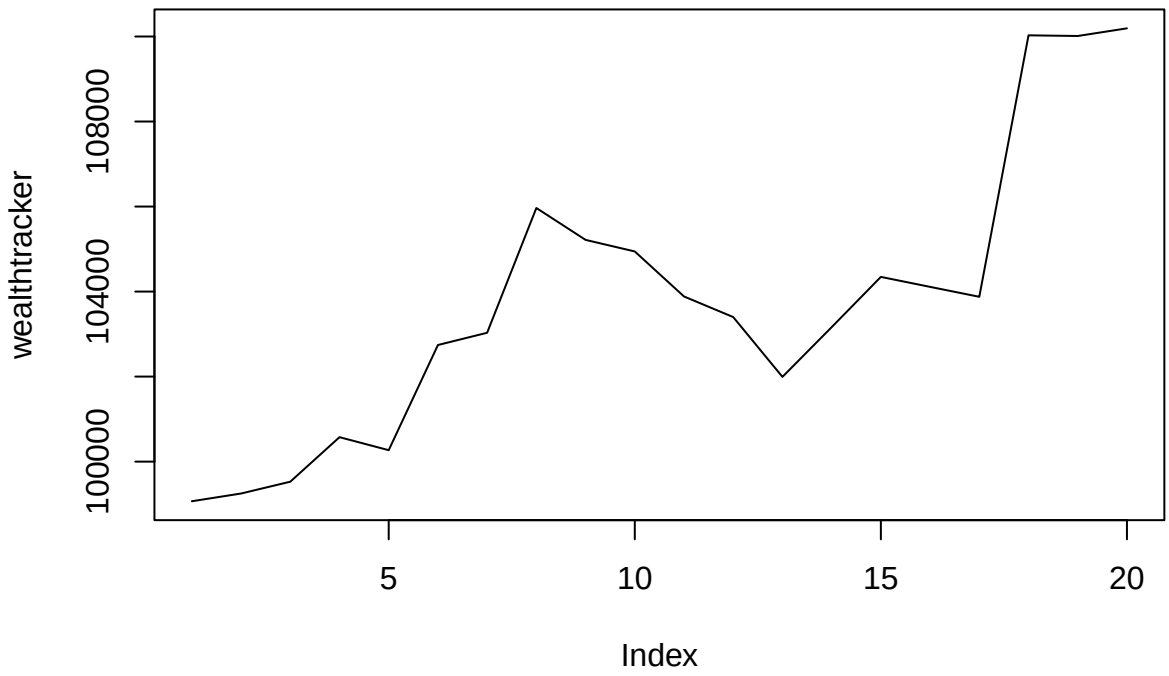
```
##        5%
## -6353.827
```

Portfolio 3 Risk Tolerant Portfolio

Finally, the third portfolio will be constructed based on risky assets, such as leveraged securities, commodity, and emerging markets funds. The portfolio will try to diversify to obtain a lower volatility, but the ultimate goal is to achieve maximum return. The portfolio will include 5 carefully selected funds with weights for higher risk funds increased.

The graph for risky portfolio returns show a steady return into a sharp rise in return. The ceiling for risky portfolio is higher due to the innate risks involved.

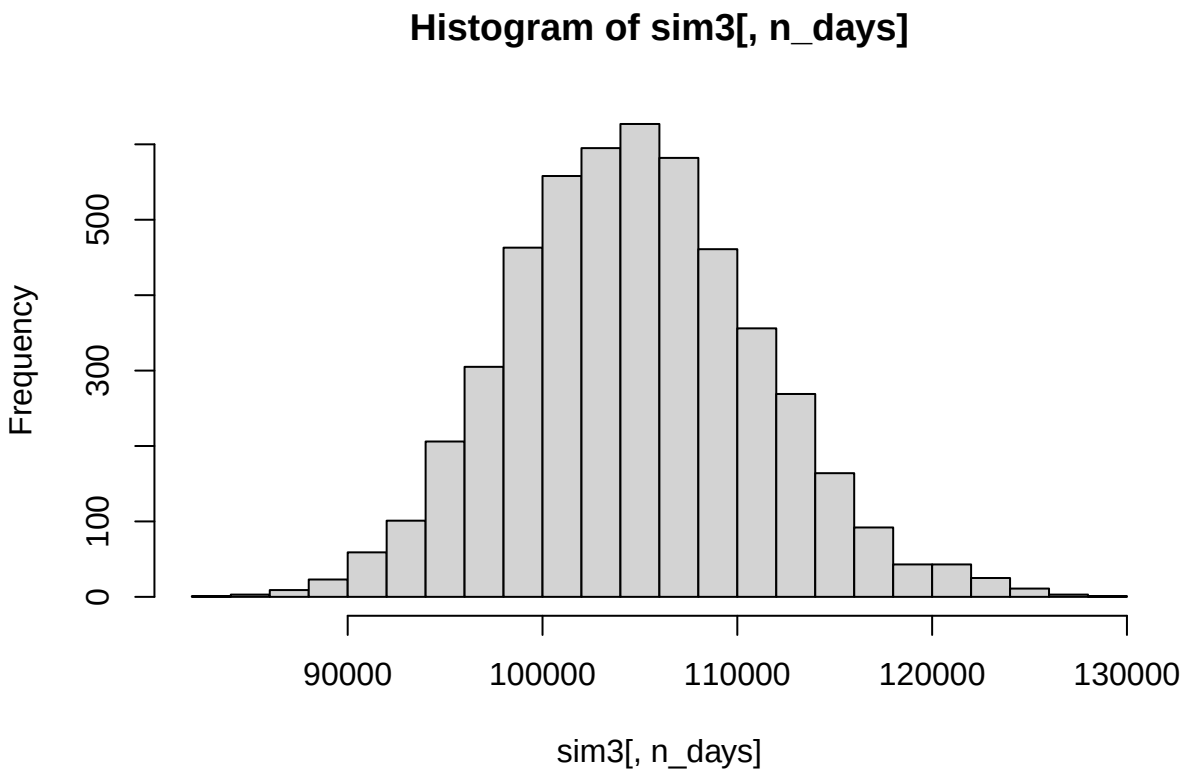
```
## [1] 110191.8
```



The average return for the 5000 simulations is 104870.5 dollars, which is a 4.87% increase. In contrast with other portfolios, the average return for risky portfolio is approximately 9 times that of the medium risk portfolio and 43 times that of the low risk portfolio. The standard deviation, however, is 6609.168, which is 1.6 times that of the medium risk portfolio and 4.8 times that of the low risk portfolio. The 5% VaR, surprisingly, is lower than that of the medium risk portfolio at 5563 dollars. This may mean that the medium risk portfolio is not well diversified, so the portfolio may be more volatile than intended.

```
##              [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## result.1 101427.50 100310.54 101057.88 102401.57 103578.94 104345.29 104942.41
## result.2 100304.35 100057.18 100490.53 100331.36  97107.67  96439.75  97065.86
## result.3  98872.28  98348.72  98959.05  99400.21  97645.24  96570.59  97896.11
## result.4  96702.80  94898.53  94055.08  94802.28  94568.67  92154.63  92058.83
## result.5  98899.43 100156.26 103064.04 101154.40 102026.26 101065.44 101373.03
## result.6  96552.35  98248.06  98322.57  98976.73 100744.88 101611.10 102342.10
##              [,8]      [,9]     [,10]     [,11]     [,12]     [,13]     [,14]
## result.1 105848.38 107627.31 109517.53 109332.67 108168.67 109478.42 108807.61
## result.2  96721.92  96732.00  96360.14  96149.92  96185.38  94867.26  94253.14
## result.3  97286.17  96608.54  96244.00  96758.26  98420.28  97734.75  96749.11
## result.4  93552.51  94895.56  95824.76  95615.71  95302.46  95288.66  95298.59
## result.5 101914.70 102203.60 104282.09 105178.72 105740.73 106982.09 107269.20
## result.6 101882.69 101389.60 100914.87 101057.05 100012.23 100552.82 100066.17
##             [,15]     [,16]     [,17]     [,18]     [,19]     [,20]
## result.1 108985.75 111866.62 110344.55 109363.82 111284.54 110919.96
## result.2  93948.00  93545.46  92928.88  93371.45  93203.56  94549.88
## result.3  97584.35  96223.23  96773.87  98082.69  97862.65  99212.82
## result.4  95008.31  93856.34  94268.37  93864.46  94012.27  94904.55
```

```
## result.5 106889.10 105398.21 105052.91 107177.29 106096.43 107008.66
## result.6 100370.72 102094.78 101560.14 103027.17 102348.10 100574.12
```




```
## [1] 6419.652
```

```
##        5%
## -5349.617
```

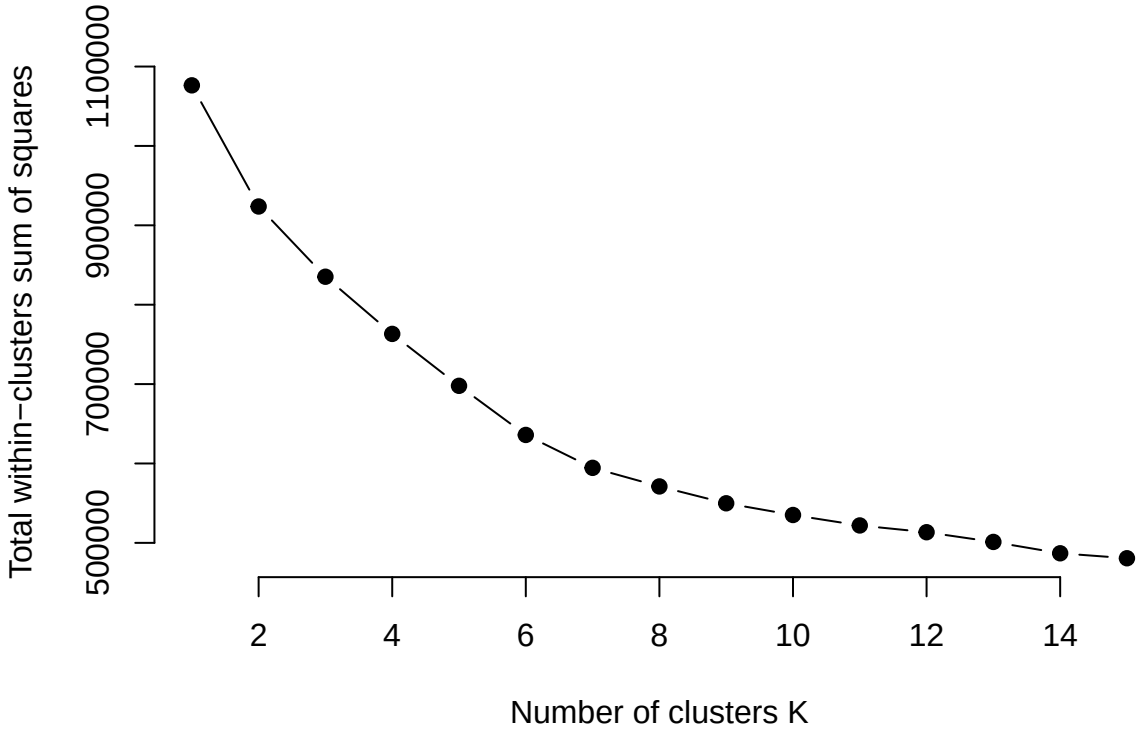
In conclusion, high risk equals to high return. The more risk averse investors will be more secure with slower, but steadier return. On the other hand, risk taking investors will have much higher return, but in turn, have a much lower ceiling for the lowest return, as suggested by the Value at Risk metric. Diversifying the portfolio is also crucial in achieving a more stable return, whether high risk or not, so do not put all the eggs in one basket!

Market Segmentation

K means clustering

First, we do some data cleaning. Since we do not want spam bots to affect our data, and uncategorized tweets may not provide anything useful, we will remove these variables from the data set. In addition, we do not need to scale our dataset, as all the variable are in the units for number of tweets.

First we check the optimal number of clusters before we perform K means clustering on the data set using the elbow method.



We find that the elbow seems to be around 6 clusters, so we will set k equal to 6.

```
## [1] 4100  588  403  519 1284  988
```

Cluster one has the largest size, with 4100 data points inside, followed by 1284 data points in cluster 5 and 988 data points in cluster 6. The remaining clusters have similar sizes. Now let's explore and try to define each cluster!

```
##    chatter current_events   travel photo_sharing   tv_film sports_fandom
## 1 2.918780       1.379756 1.114146      1.507805 1.0378049      1.514146
## 2 4.032313       1.680272 6.335034      2.256803 1.1632653      2.204082
## 3 4.039702       1.416873 1.526055      2.625310 1.3771712      1.620347
## 4 3.940270       1.710983 1.452794      5.899807 0.9922929      1.545279
## 5 9.883178       1.866044 1.192368      5.619938 1.0599688      1.637072
## 6 4.018219       1.548583 1.315789      2.440283 1.0789474      1.521255
##    politics     food    family home_and_garden     music      news
## 1 0.9578049 1.189024 0.7543902       0.4500000 0.5604878 0.8051220
## 2 9.8724490 1.767007 0.9863946       0.5850340 0.6275510 5.1343537
## 3 1.2555831 1.464020 1.1389578       0.5508685 0.7518610 0.8635236
## 4 1.3352601 1.252408 0.9653179       0.6069364 1.1907514 1.0578035
## 5 1.4657321 1.244548 0.9953271       0.5950156 0.7959502 0.8380062
## 6 1.3006073 2.290486 0.9089069       0.6214575 0.7530364 1.2236842
##   online_gaming  shopping health_nutrition college_uni sports_playing   cooking
## 1     0.5600000 0.7346341        0.9507317   0.8980488      0.4392683  0.820000
## 2     0.8384354 1.1870748        1.4183673   1.3418367      0.6598639  1.248299
## 3    10.4789082 1.2133995        1.5707196  10.8163772      2.4937965  1.553350
```

```
## 4     1.0462428 1.7013487        2.0134875   1.4373796      0.8381503 11.932563
## 5     0.8014019 3.5475078        1.3123053   1.2422118      0.6012461  1.170561
## 6     0.9554656 1.3299595       12.2874494   1.0546559      0.6447368  3.372470
##         eco computers  business  outdoors    crafts automotive       art
## 1 0.3597561 0.3880488 0.3195122 0.4585366 0.4058537  0.6053659 0.6514634
## 2 0.5952381 2.6462585 0.6581633 0.8928571 0.6343537  2.0374150 0.6853741
## 3 0.4689826 0.5781638 0.3796526 0.6327543 0.6253102  0.9429280 1.2332506
## 4 0.5394990 0.7475915 0.5645472 0.8150289 0.6069364  0.8593449 0.9094412
## 5 0.6923676 0.6160436 0.5864486 0.5264798 0.6417445  1.0568536 0.6643302
## 6 0.8653846 0.5637652 0.4453441 2.4392713 0.6457490  0.6862348 0.8269231
##    religion    beauty parenting    dating    school personal_fitness   fashion
## 1 1.0770732 0.4395122 0.8370732 0.4629268 0.6660976        0.6397561 0.5287805
## 2 1.3979592 0.5153061 1.1802721 1.0442177 0.8554422        0.9285714 0.6904762
## 3 1.0471464 0.5235732 0.9950372 0.7270471 0.6774194        1.0173697 0.9429280
## 4 1.2562620 3.8458574 1.0616570 0.5857418 1.0539499        1.2986513 5.6897881
## 5 0.9540498 0.5482866 0.9244548 1.1744548 0.9929907        0.9595016 0.8870717
## 6 1.1103239 0.5485830 1.0091093 0.9979757 0.7307692        6.1123482 0.8188259
##   small_business
## 1      0.2775610
## 2      0.4846939
## 3      0.4168734
## 4      0.4605010
## 5      0.4322430
## 6      0.2692308
```

To begin, we will look primarily on the mean values for cluster 1, which is the largest market segment for NutrientH20. Surprisingly, cluster 1 has the lowest average number of tweets for all of the fields, except for religion and small business. Even for religion and small business, they have one of the lowest mean. This means that a majority of the brand's twitter followers do not even post much.

Next, we examine the mean values for cluster 5, the second largest market segment. Cluster 5 has the highest average number of tweets for chatter, current events, shopping, dating, and the second highest average number of tweets for photo sharing, family, music, eco, business, crafts, and school. On the other end, they have the lowest average number of tweets for religion. This cluster's demographic appears to be a female audience.

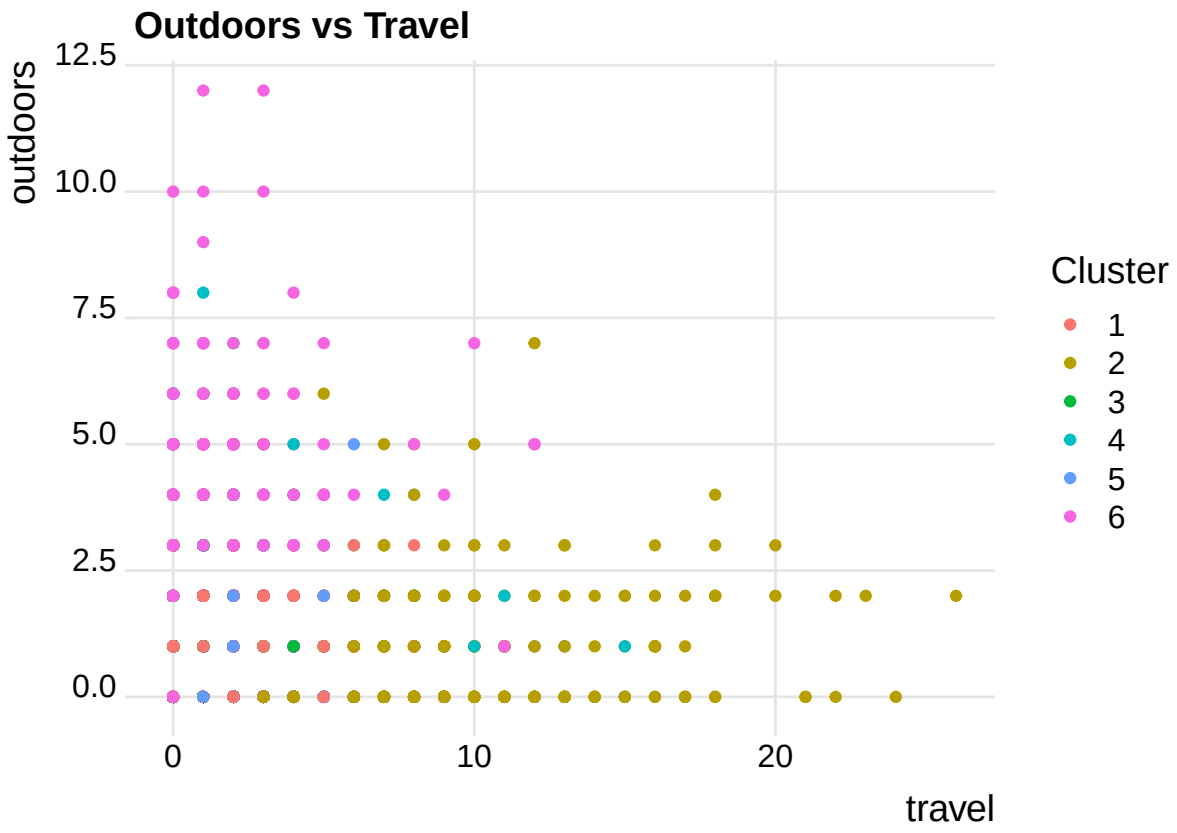
Now we infer upon the mean values for cluster 6. This cluster has the highest average number of tweets for food, home/garden, health nutrition, eco, outdoors, crafts, personal fitness. It seems obvious that this cluster is health and environmentally conscious.

Lastly, we will check the highest mean values for the remaining fields, and see what clusters they represent:

2 = Travel, Sports Fandom, Politics, News, Computers, Business, Automotive, Religion, Parenting, Small Business 3 = Tv Film, Family, , Online Gaming, College, Sports Playing, Art 4 = Photo Sharing, Music, Cooking, Beauty, Dating, School, Fashion

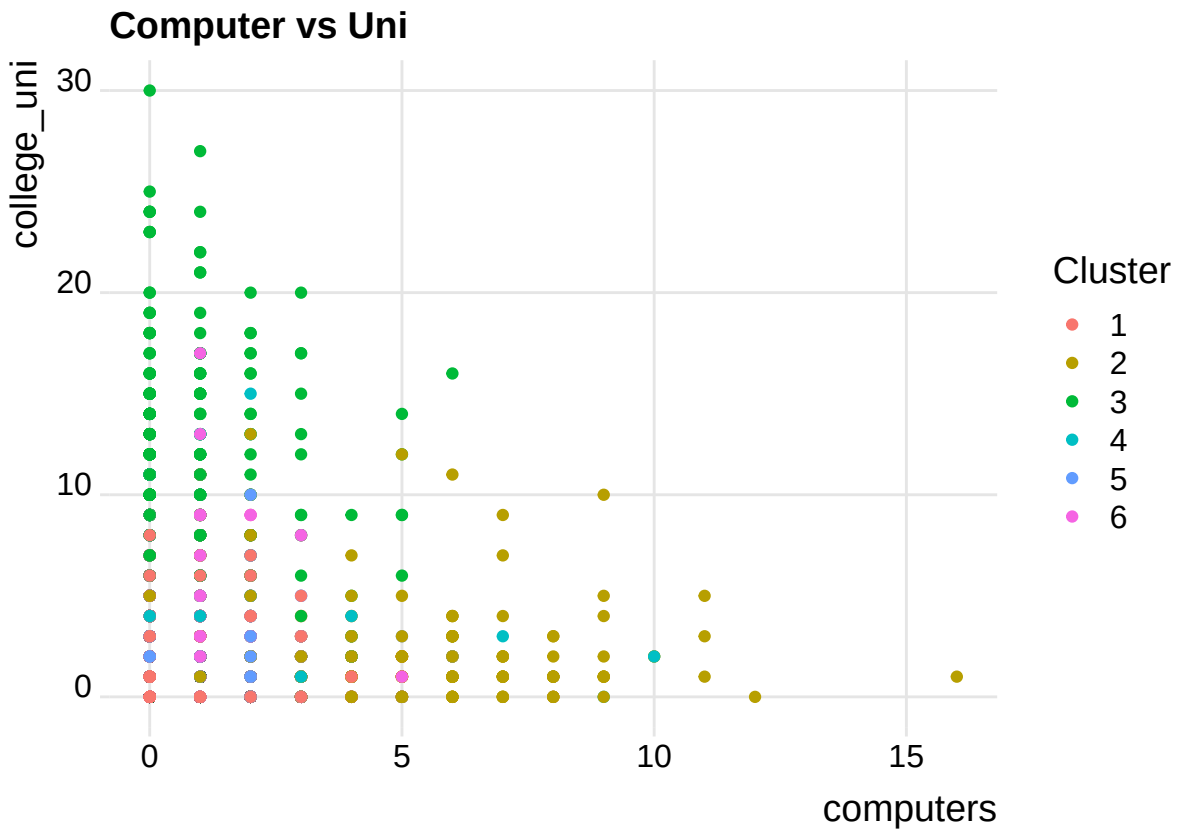
Based on these features, cluster 2 seems to be older adults with an active lifestyle. They tend to have a much higher number of tweets in contrast with other clusters, which shows that they are more vocal. Cluster 3 appears to be college students, who post more about their hobbies, while cluster 4 seems to be very similar to cluster 5.

Now let's visualize some data!

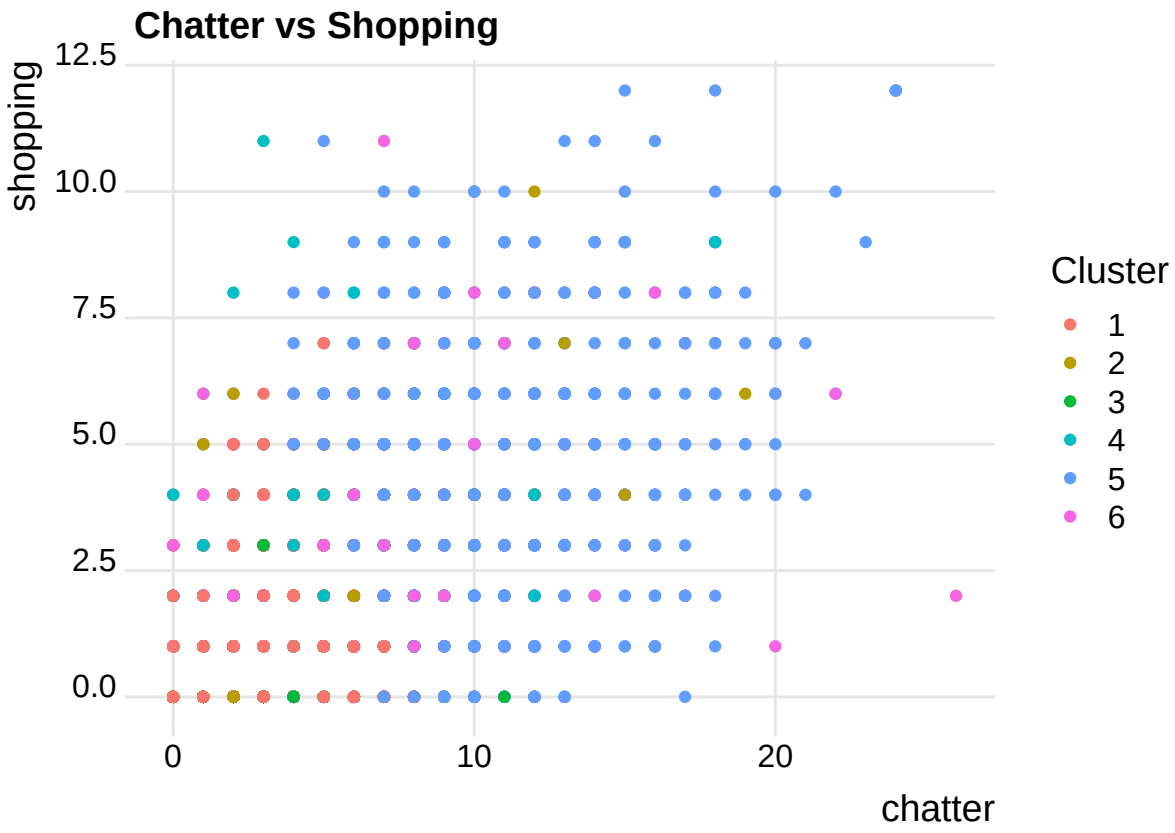


For Travel and Outdoors, the dominant clusters are 2 and 6, with cluster 2 leaning more towards travel and cluster 6 having more tweets on outdoors.

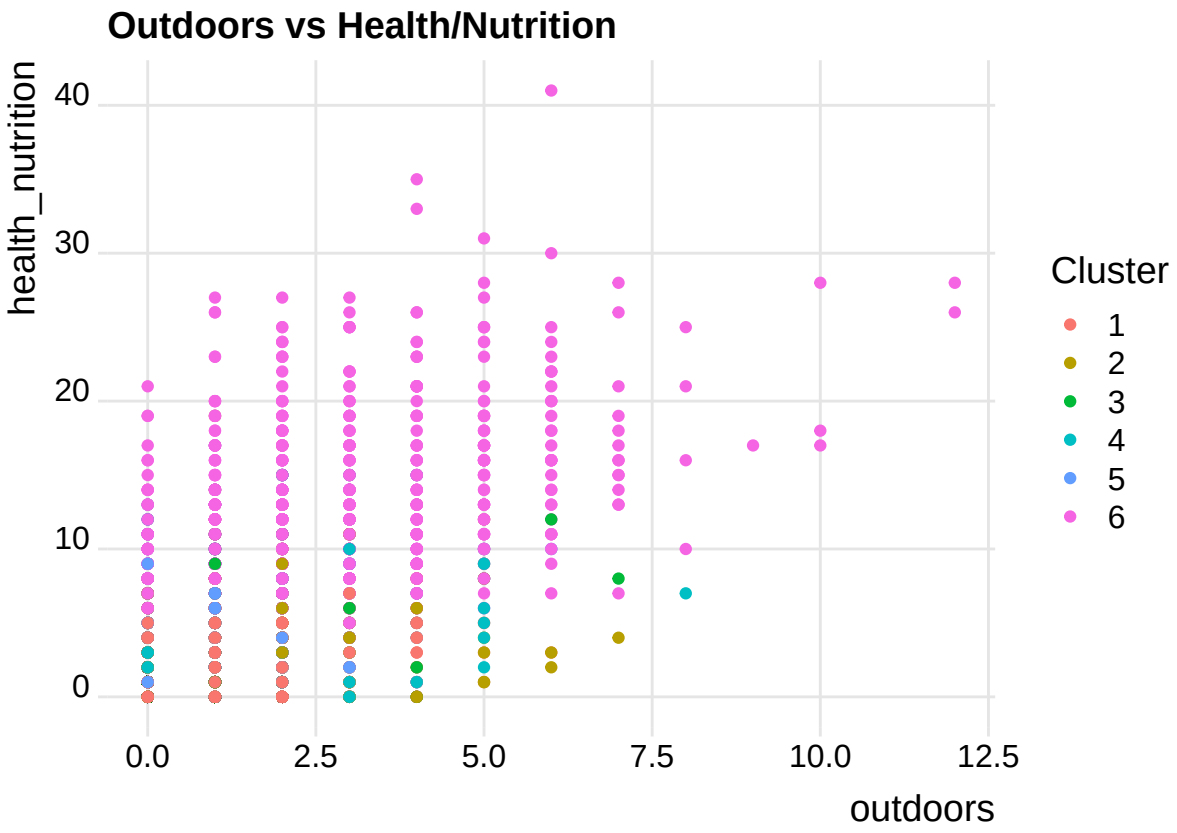
As suspected earlier, clusters 4 and 5 both have a high number of tweets in the beauty and shopping features. Cluster 5 lean more towards shopping, while cluster 4 have more tweets about beauty.



Cluster 2 and 3 are both the major clusters that tweet about computers and uni.



The majority of followers who tweet about chatter and shopping is cluster 5.



Cluster 6 has a high amount of tweets related to outdoors and health/nutrition.

In summary, by clustering the twitter followers, we find that the market is segmented into 4 main segments. The largest follower base is not as vocal and does not post on Twitter much. The second largest follower base is a female customer base that is mainly concerned with chatter, shopping, and current events. The third largest follower base also has a similar size as the female customer base, and this demographic is primarily health and environmentally conscious. By tailoring the Twitter content towards these two main demographics, customer impressions and engagement may see an increase.

Author Attribution

For author Attribution, we decided to build models to recognize text from three different authors: Benjamin Kang Lim, Darren Schuettler, and Fumiko Fujisaki. To briefly introduce these authors, Benjamin Kang Lim is a Philipino who worked in China and Taiwan. He regularly writes about news related to the Communist Party of China. Darren Schuettler, on the other hand, is a Canadian working in Asia. He writes topics on both Canada and parts of Asia. Lastly, Fumiko Fujisaki usually writes about Japan.

The models that will be considered are Principle Components Analysis, Naive Bayes, and and Hierarchical Clustering. The files will be preprocessed by removing numbers, punctuation, white spaces, english and SMART stop words, and will all be lowercased. The weighting used will be term frequency to remove the super rare words or phrases that appear in the text.

```
## <<DocumentTermMatrix (documents: 150, terms: 6489)>>
## Non-/sparse entries: 29865/943485
## Sparsity           : 97%
## Maximal term length: 39
## Weighting          : term frequency (tf)
```

We find that the sparsity within the matrix is 97%, which means that 97% of the matrix consists of zeroes. This most likely means that the three authors write about very distinct topics from each other, resulting in little overlaps in the terms used.

After removing the sparese terms, the number of entries decreased from 940k to 262k, as the sparsity percentage dropped to 92%.

```
## <<DocumentTermMatrix (documents: 150, terms: 1905)>>
## Non-/sparse entries: 22980/262770
## Sparsity           : 92%
## Maximal term length: 39
## Weighting          : term frequency (tf)
```

Naive Bayes

Now we try building a Naive Bayes classifier model. A smoothing factor is applied to ensure that the output probability is not extremely small. Applying the smoothing factor ensures interpretability of the results.

Now let's try testing on the documents unused to see which author is predicted.

Test data 1 first 25 entries:

```
##   newspaper  investment  government enterprises  industries   character
##           8           7           6           6           5           4
##        year        high     central    products      chinas         mao
##           4           4           4           4           3           3
##       state     economy  technology    economic     problem        plan
##           3           3           3           2           2           2
##        curb         due       local      listed highprofile   economist
##           2           2           2           2           2           2
##     regions
##           2
```

It seems like key terms such as chinas and mao suggested that perhaps Benjamin Kang Lim wrote this. Let's see if the model predicts accurately.

The log probability was -1384.005 for Benjamin Kang Lim, -1490.33 for Darren Schuettler, and -1475.14 for Fumiko Fujisaki. The model predicted correctly! Now let's try another document.

Test data 2 first 25 entries:

```
##               zhang               court               lawyer
##                   9                   7                   5
##         authorities           character              chiang
##                   4                   4                   4
##               china             chinese           democracy
##                   4                   4                   4
##               human              rights                hong
##                   4                   4                   3
##            province              taiwan               years
##                   3                   3                   3
##              leader counterrevolutionary           dissident
##                   3                   3                   3
##               found             accused                june
##                   3                   2                   2
```

```
##                kong           political            standing
##                   2                   2                   2
##                year
##                   2
```

Terms such as Zhang, chinese, taiwan may seem obvious to a human reader who the author is (Benjamin Kang Lim). Let's check the model predictions.

The log probability was -1712.924 for Benjamin Kang Lim, -2298.87 for Darren Schuettler, and -2336.6 for Fumiko Fujisaki. The probabilities have a much bigger contrast this time. Let's try another document.

Test data 3 first 25 entries:

```
##      banks        tax       bank       year      major    capital     budget
##         18         13          8          6          6          5          5
##      taxes  character      years   canadian       week government    percent
##          5          4          4          4          3          3          3
##    profits      small   business     canada    ontario    bankers    billion
##          3          3          3          3          3          3          2
##    reforms   analysts   expected     record
##          2          2          2          2
```

Terms such as canadian, canada, ontario suggest that this is probably an article written by Darren Schuettler.

The log probability was -1996.181 for Benjamin Kang Lim, -1713.539 for Darren Schuettler, and -1783.817 for Fumiko Fujisaki. The probabilities for Darren Schuettler and Fumiko Fujisaki are surprisingly close. This might be due to similar issues discussed in their articles (banks, government, business). Let's try another document.

Test data 4 first 25 entries:

```
##           land           real         estate     government         market
##             13              8              8              7              6
##          firms      financial      character        problem          owned
##              6              5              4              4              4
##            buy          panel         public          funds        finance
##              4              4              3              3              3
## collateralised      liquidity       troubled         japans       minister
##              3              3              3              2              2
##       official         tokyos         impact            led          state
##              2              2              2              2              2
```

Terms like japan and toyko are indicators that Fumiko Fujisaki was the author. However, the terms are indeed strikingly similar to that of the previous test.Let us see if the model will predict accurately.

The log probability was -2025.16 for Benjamin Kang Lim, -2008.82 for Darren Schuettler, and -1669.8 for Fumiko Fujisaki. This time, model has a high change of predicting fumiko fujisaki.

Test data 5 first 25 entries:

```
##       banks     foreign      canada        bank    canadian        year
##          17          13          12           9           7           5
##     banking   character    domestic       rules competition   financial
##           5           4           4           4           4           3
##     canadas    services   operating     bankers  government         big
```

```
##           3           3           3           3           2           2
##      number    recently    chairman     capital       plans         tax
##           2           2           2           2           2           2
##      access
##           2
```

The log probability was -2278.31 for Benjamin Kang Lim, -1926.6 for Darren Schuettler, and -2017.87 for Fumiko Fujisaki. The log probability difference was close, but the model still predicted correctly.

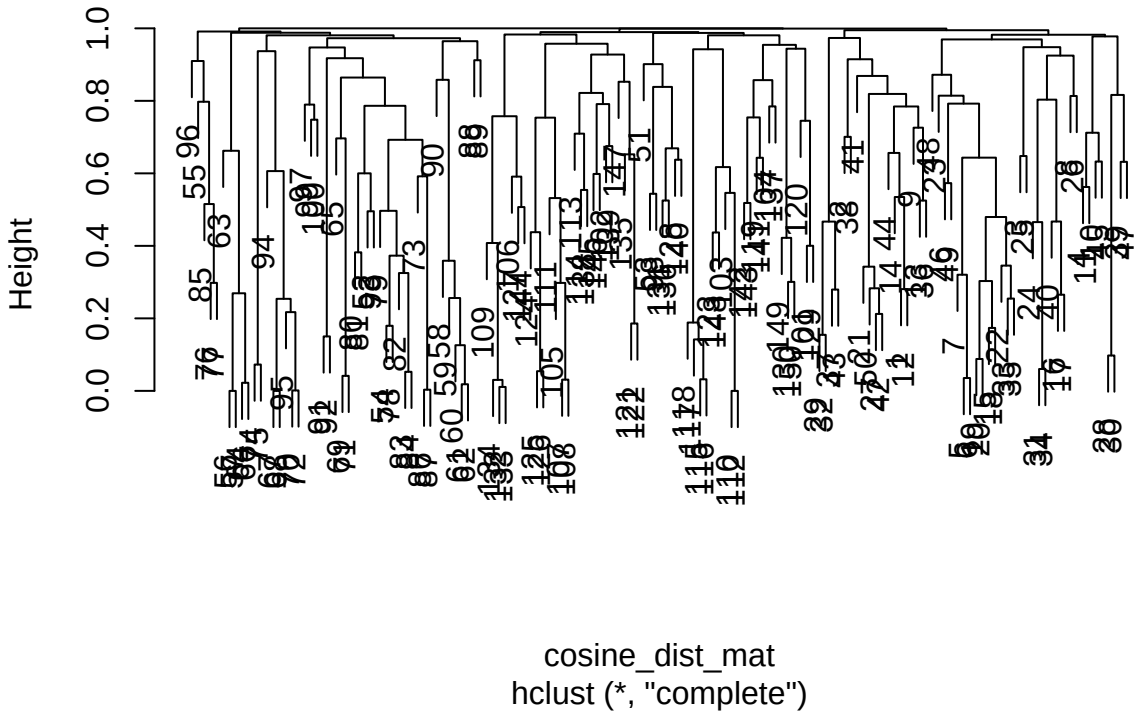
Test data 6 first 25 entries:

```
##         daiwa       billion          bank           yen        states
##            15             8             8             6             5
##        united      business        daiwas     character         march
##             5             5             5             4             4
##          year         banks restructuring      japanese        japans
##             4             4             4             3             3
##       million      analysts     september        damage    operations
##             3             3             3             3             3
##           end international         japan      ministry       company
##             2             2             2             2             2
```

The log probability was -2300.79 for Benjamin Kang Lim, -2199.6 for Darren Schuettler, and -1885.85 for Fumiko Fujisaki. The model predicted correctly again. Next, We will move on to the hierarchical clustering model to see if it can predict better than Naive Bayes.

Hierarchical Clustering

For Hierarchical Clustering, we will use Inverse Document Frequency to find terms that appear too often and remove them from the dataset. Then we will calculate cosine distance to form clusters.



Cluster 1:

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
```

After forming a hierarchical cluster, we cut the tree into 3 separate cluster. The first cluster seems to be Benjamin Kang Lim, with accuracy of 100%. All of BKL's documents are correctly in cluster 1.

Cluster 2:

```
##  51  53  98 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117
##  51  53  98 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117
## 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137
## 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137
## 138 139 140 141 142 143 144 145 146 147 148 149 150
## 138 139 140 141 142 143 144 145 146 147 148 149 150
```

The second cluster seems to be Fumiko Fujisaki. This time, however, the prediction accuracy is 90%. The cluster is missing documents 99 and 100, while containing documents 51,53, and 98. Cluster 3 should also have the same prediction accuracy of 90%. This error in predictor may be attributed to the observation made above: Fumiko Fujisaki and Darren Shuettler both write about similar topics, albeit on different countries and governments.

By observing the incorrectly predicted documents, we find that these documents are all on the same topic - banks, financial services.

Principle Component Analysis

Lastly, we perform Principle Component Analysis.

From the analysis, we find that 32 components explain approximately 50% of the variation over almost 2000 features. If we look at the loadings for component 1 and 2 below, there actually does not seem to much in common for the two components.

```
##    bombardier       busiest     commuters     factories      gingrich
##     0.1096599     0.1096599     0.1096599     0.1096599     0.1096599
##     havilland inconvenience          newt         plant      revamped
##     0.1096599     0.1096599     0.1096599     0.1096599     0.1096599
##      servants        shifts       speaker         stall        trains
##     0.1096599     0.1096599     0.1096599     0.1096599     0.1087800
##    protesters     picketers        harris      workfare       disrupt
##     0.1087601     0.1081809     0.1065524     0.1064558     0.1060113
##   communities       offices      baseball           bat       bedroom
##     0.1057567     0.1043836     0.1040779     0.1040779     0.1040779
```

```
##                                dissident                                    crushed
##                                0.09176463                                0.09161500
##                                    served                                subversion
##                                0.09053329                                0.09008988
##                               prodemocracy                           demonstrations
##                                0.08837294                                0.08835608
##                                    family                                   student
##                                0.08712022                                0.08510498
##                                 jingsheng                                     trial
##                                0.08463985                                0.08416560
##                                 overthrow                                background
##                                0.08398973                                0.08330421
##                                    daring                                   defying
##                                0.08330421                                0.08330421
##                                   maximum                                    museum
##                                0.08330421                                0.08330421
##                                  sentence                                      wang
##                                0.08317704                                0.08237946
##                                    mother                                     court
##                                0.08226509                                0.08143652
## reutersccctrainbenjaminkanglimnewsmltxt                                   lingyun
##                                0.08073200                                0.08018256
##                                       dan                           surveillance
##                                0.07988639                                0.07961998
##                    counterrevolutionary
##                                0.07934538
```

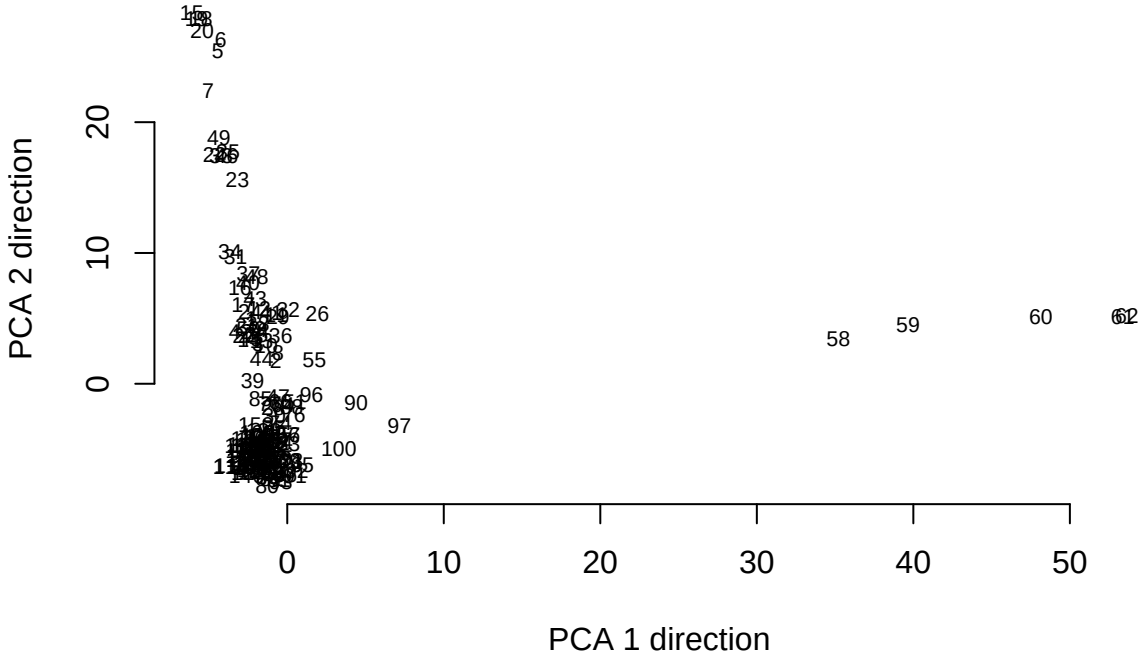
The graph of the two principal components seem to clutter in two big groups, one for components 80 and above, and another for 10-50. This shows that Fumiko Fujisaki and Darren Shuettler use words or phrases that are extremely similar.

```
##
## Docs          PC1        PC2
##   1   -2.30328992  4.9870108
##   2   -0.73578695  1.7856378
```

```
##   3   -1.85496906  3.0905906
##   4   -1.52589536  4.0639724
##   5   -4.44781189 25.4495956
##   6   -4.24508569 26.2786451
##   7   -5.04878526 22.4427948
##   8   -0.59004057  2.4068741
##   9   -1.70967096  4.2389120
##   10  -1.35785602  2.8883522
##   11  -0.69152789  5.2750856
##   12  -1.78812563  5.7754056
##   13  -1.63591762  3.2973122
##   14  -2.41576402  3.3708172
##   15  -6.09568073 28.3972417
##   16  -3.02022315  7.3554128
##   17  -2.80430827  6.0978084
##   18  -5.51537366 27.9289358
##   19  -5.78946742 27.8762202
##   20  -5.44396191 26.9893984
##   21  -2.57189840  4.4320319
##   22  -4.63849955 17.5258993
##   23  -3.18520489 15.6289423
##   24  -2.36921209  5.5081882
##   25  -1.95776953  3.7562004
##   26   1.92817123  5.3628165
##   27  -2.74686649  3.7211263
##   28  -0.89993796 -1.7927859
##   29  -0.59201276  5.1202142
##   30  -0.83387701 -2.4431662
##   31  -3.30201395  9.6958084
##   32   0.06324731  5.6339098
##   33  -4.22519680 17.4700101
##   34  -3.64757136 10.1359859
##   35  -3.78034637 17.7309243
##   36  -0.40948625  3.6471847
##   37  -2.48086143  8.3881554
##   38  -1.87963643  4.5876918
##   39  -2.21624768  0.2705045
##   40  -2.50627152  7.7024770
##   41  -1.02335504  5.4692352
##   42  -2.97657369  3.9938545
##   43  -2.04553136  6.5335339
##   44  -1.62346714  1.9621306
##   45  -2.35215750  3.4338528
##   46  -3.87990729 17.4516792
##   47  -0.57191089 -0.9967656
##   48  -1.94559318  8.1859976
##   49  -4.36113018 18.8160258
##   50  -2.47808768  4.1952885
##   51   0.48659435 -1.4043375
##   52  -0.74368490 -6.8982196
##   53   0.08629603 -4.5457135
##   54  -0.78869027 -6.8721331
##   55   1.73470934  1.7933210
##   56   0.08530406 -3.8815275
```

```
##   57   0.08530406 -3.8815275
##   58  35.21241889  3.4447043
##   59  39.66566808  4.5080718
##   60  48.15400687  5.1608389
##   61  53.37953800  5.0986132
##   62  53.64672103  5.2233067
##   63  -0.73384125 -5.6664158
##   64  -0.29280000 -1.5679818
##   65   0.96372368 -6.1524392
##   66  -0.44389094 -1.4060018
##   67  -1.25922062 -5.9322339
##   68  -1.25922062 -5.9322339
##   69   0.14771761 -5.9164563
##   70  -1.74567330 -4.9812065
##   71   0.28908569 -5.9193699
##   72  -1.89634039 -4.9867102
##   73   0.26519353 -5.8724426
##   74  -0.44797062 -2.9683975
##   75  -0.86011521 -3.3563834
##   76   0.42868522 -2.3972591
##   77  -1.10552237 -1.5275888
##   78  -0.98860496 -6.8292469
##   79  -0.42242557 -5.7551454
##   80  -1.28158270 -7.7582974
##   81  -0.52705594 -7.2992277
##   82  -0.99019860 -7.2874773
##   83  -1.24048955 -7.0710378
##   84  -0.94618576 -6.4437724
##   85  -1.70287809 -1.1629197
##   86  -0.12145382 -6.9382751
##   87  -0.16734457 -6.8885685
##   88   0.28188807 -1.7214129
##   89  -0.74115983 -4.3362932
##   90   4.39170951 -1.4648711
##   91   0.49575567 -7.0171636
##   92   0.61990575 -6.6621938
##   93  -0.40406618 -7.4550869
##   94  -0.39456617 -4.4078616
##   95  -1.27939305 -5.7854591
##   96   1.55729009 -0.8497476
##   97   7.16025716 -3.1948882
##   98  -1.59694534 -4.0011966
##   99  -0.79642675 -3.7111749
##  100   3.29542152 -4.9740508
##  101 -1.62709412 -5.4802730
##  102 -0.95985385 -4.5058314
##  103 -2.77969438 -5.4999144
##  104 -1.49073608 -4.6731193
##  105 -2.23822348 -5.1166966
##  106 -1.83364583 -3.9917184
##  107 -2.32285621 -6.3718448
##  108 -2.36596470 -6.4156838
##  109 -1.90592368 -3.7158445
##  110 -2.77559816 -5.7953842
```

```
##   111 -2.69706945 -4.6332280
##   112 -2.77559816 -5.7953842
##   113 -1.95611381 -5.6565824
##   114 -2.64941064 -6.2686508
##   115 -3.61546471 -6.3590838
##   116 -3.52461529 -6.2897407
##   117 -2.80108059 -5.0998424
##   118 -2.87506100 -4.7314356
##   119 -2.14279131 -4.2628237
##   120 -1.61799858 -4.3655459
##   121 -1.88839073 -4.1018749
##   122 -1.89464334 -4.9005085
##   123 -2.48788869 -4.7552479
##   124 -2.44755966 -6.7604782
##   125 -2.08323202 -6.5426092
##   126 -1.97448064 -6.7215824
##   127 -1.23923142 -5.6823653
##   128 -1.51190937 -3.5519575
##   129 -2.16890316 -4.6946258
##   130 -2.47445816 -4.1936852
##   131 -1.56971950 -6.3332418
##   132 -2.31798457 -6.1339385
##   133 -2.28505698 -6.1208488
##   134 -2.23048536 -6.0152896
##   135 -0.75176794 -5.6176009
##   136 -1.61224987 -5.6704568
##   137 -0.80123994 -4.2613609
##   138 -2.05475071 -6.1776261
##   139 -1.36031744 -6.6453371
##   140 -2.18310269 -5.0264291
##   141 -1.17175773 -5.8310807
##   142 -1.31660276 -4.6689615
##   143 -1.25145001 -5.0025600
##   144 -1.74921951 -5.7412513
##   145 -2.09664068 -4.8446814
##   146 -2.60643529 -7.0547506
##   147 -1.62485704 -4.0075421
##   148 -2.43356512 -5.0454228
##   149 -2.17796219 -5.0248097
##   150 -2.00604241 -3.1131301
```



In conclusion, all three models predict Benjamin Kang Lim very well, yet sometimes get mixed up between the words in Darren Shuettler and Fumiko Fujisaki. The best model was Naive Bayes Classifier, since all the tests conducted predicted correctly, even ones that are clustered incorrectly.

Association rule mining

We will inspect whether there are any interesting associations between the grocery list and provide insights on the data. First we read the text file as basket format in order to process the transaction data.

From the summary we find that milk is the most frequent item, followed by vegetables and rolls/buns.

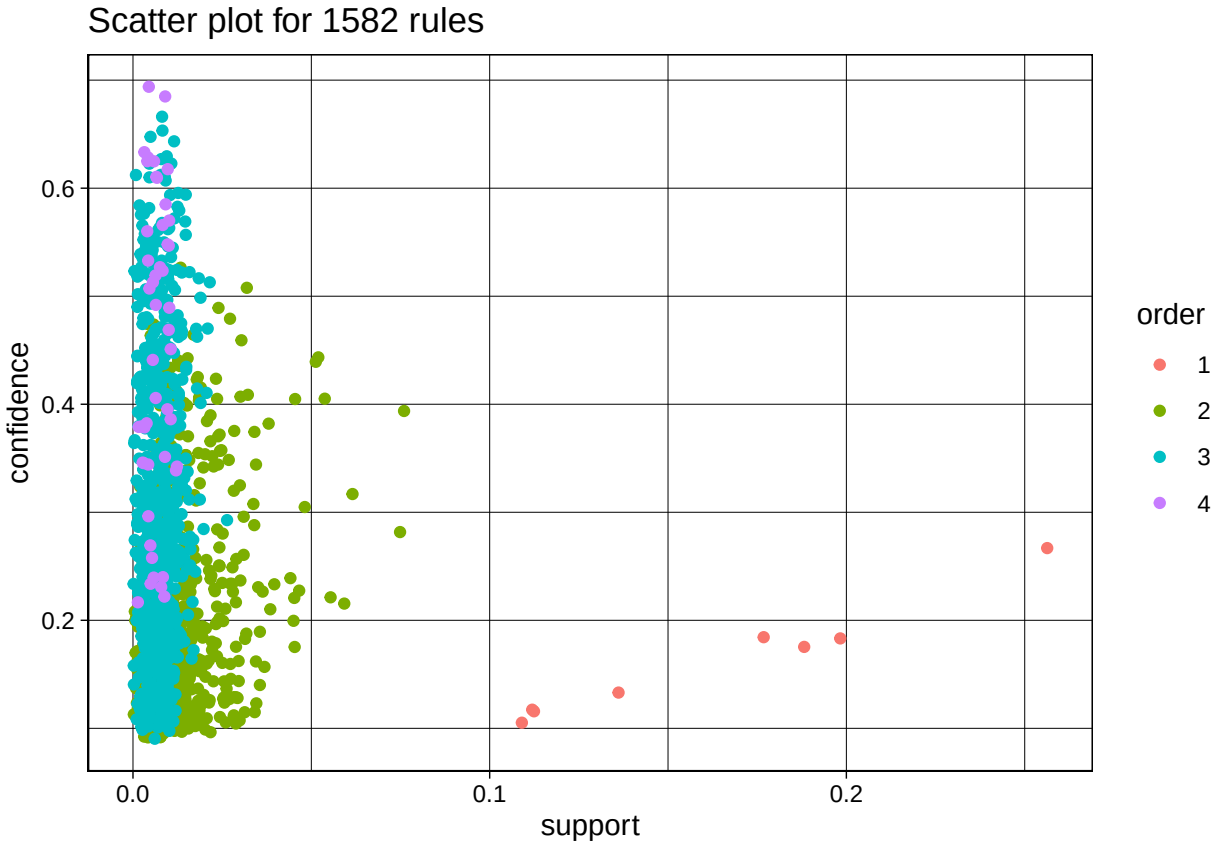
After building association rules, we find that the maximum lift is approximately 4.5 and the maximum confidence is around 0.6. Next we will filter the association rules by high confidence and high lift to find interesting associations.

We first try setting the condition to filter rules with lift higher than 4. Only 4 associations have lift higher than 4, with ham and white bread having the highest lift. This highly suggests that ham and white bread are complements of each other. Next we see shoppers who buy citrus fruits, other vegetables, whole milk also buy root vegetables, and shoppers who buy butter and other vegetables also buy sour cream. Interestingly enough the last association for butter, other vegetables and sour cream is not as obviously a complement as the others.

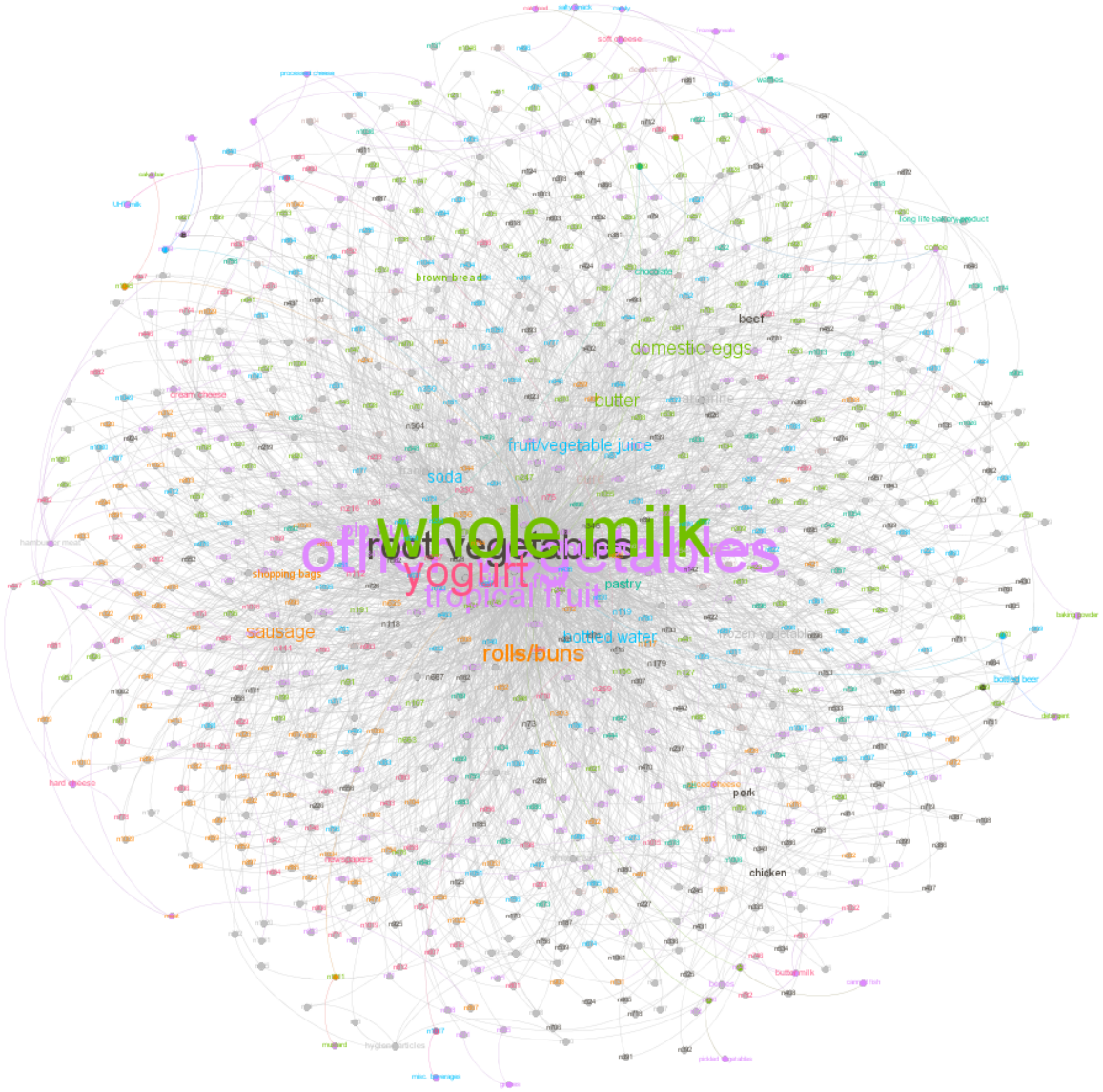
Then, we check rules with high confidence of higher than 0.6. We find that there is 22 association rules with such high confidence, with whole milk on the right hand side of the rule for most of the rules. The remaining item in the right hand side is other vegetables. This shows that whole milk and vegetables are most commonly bought with other items.

Lastly, we filter rules with both high confidence and high lift to see what associations are present. The condition we set is life above 3 and confidence above 0.5. We find that mainly 'other vegetables' is on the right hand side, implying how complementary and commonly bought this item is.

After plotting the association rules, we find that support of each association rule is generally quite low, while confidence varies from 0 to 0.6. The association rules with order 3 seem to be the most prominent with a smaller support value on average. Association rules with order 2 are more scattered, with a slightly higher support on average than order 3. Association rules with order 4 generally have high confidence and low support, while association rules with order 1 have much higher support with lower confidence.



Next, we build a network graph to show the associations and the major players of the grocery items.



As shown in the association network, we find that items from the same cluster have the same color. Fruits and vegetables form the purple cluster, dairy products form the green cluster, and beverages form the blue cluster.

In summary, the association rule results show some product associations that are expected. Milk and vegetables are generally bought with other items for every transaction, and bread and ham are complements of each other. The only surprising association rule with high degree of association is butter, vegetable, and sour cream. Sour cream does not seem to fit in with vegetables and butter, but these are most likely common grocery goods that Americans buy.