STA 380, Part 2: Exercises 1

Ryan Hoff, Won Lee, Sam Malcolm, Cory Nguyen

8/6/2018

STA 380, Part 2: Exercises 1

Probability practice

Part A.

p = (.65 - (.5*.3))/.7

.71 of people who are truthful clickers answered Yes.

```
## [1] 0.7142857
```

Part B.

Suppose someone tests positive. What is the probability that they have the disease? In light of this calculation, do you envision any problems in implementing a universal testing policy for the disease?

PPV = ((0.993)*(0.000025)/((0.993)*(0.000025)+(1-0.9999)*(1-0.000025))*100)

The probability that they have the disease is 19.88824%. This means that approximately 80% of the positive results will be false positives. This is not a reliable test, and those who receive false positives would have to incur unnecessary medical costs and, potentially, treatment.

```
## [1] 19.88824
```

Part 2: Exploratory Analysis

Overall, we think it's a solid first step in an analysis. However, we question how he made certain decisions, believe he may have taken too vague of an approach, and ultimately think his conclusion is flawed.

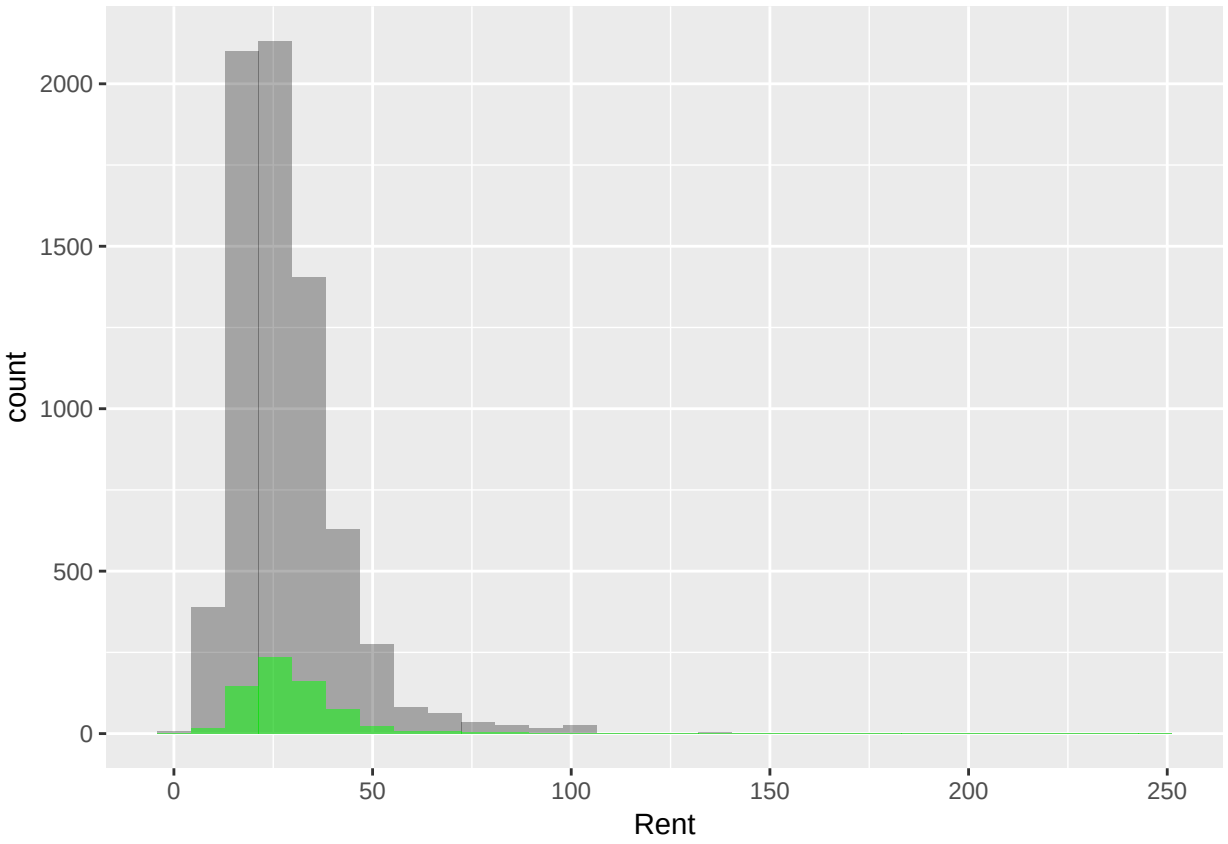
For example, he arbitrarily removes buildings with very low occupancy rates because they "might have something weird going on with them" but still chooses to evaluate his decision based on the median, effectively discounting outliers twice.

He is correct that the mean and median rents of Green buildings tend to be higher than non-Green buildings overall. This appears to maintain when comparing the rents of Green buildings only to those in their cluster. However, there are still a significant number of occurrences where a Green building charges less respectively for rent.

```
## [1] 30.01603
```

```
## [1] 28.26678
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



With this in mind, if you were to plan to build in bulk - say 100 buildings - you might have a better chance of seeing a return on investment with Green certification as rents regressed toward the mean. However, for just 1 building, we can't confidently say that you will be able to charge a premium.

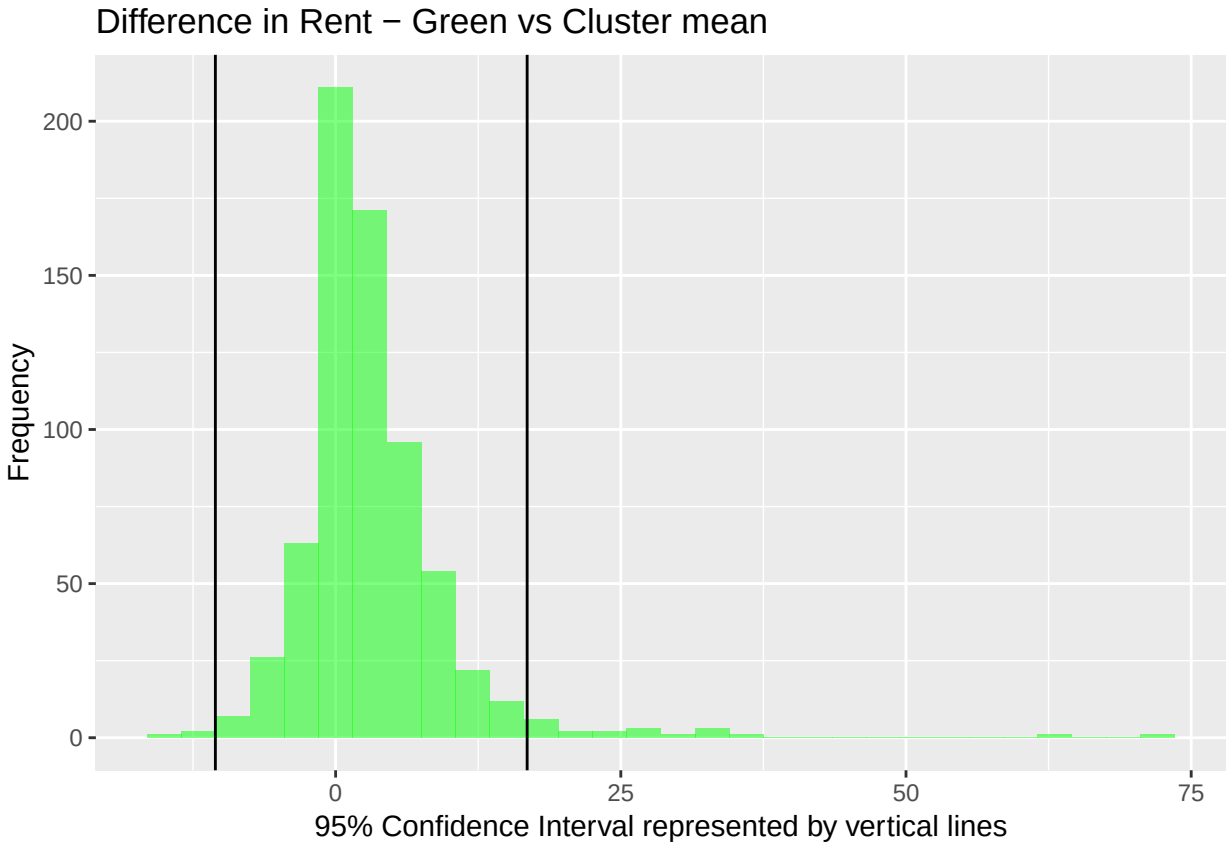
To show this, we can take a few different perspectives. First, we'll look at overall correlation between variables.

```
## [1] 0.03266587
```

'Green' has only a .03 correlation with 'Rent.' Immediately we can see that this relationship might not be as straightforward as he proposed.

Perhaps it would be more useful to compare the rent of a Green building with the mean rent in its cluster.

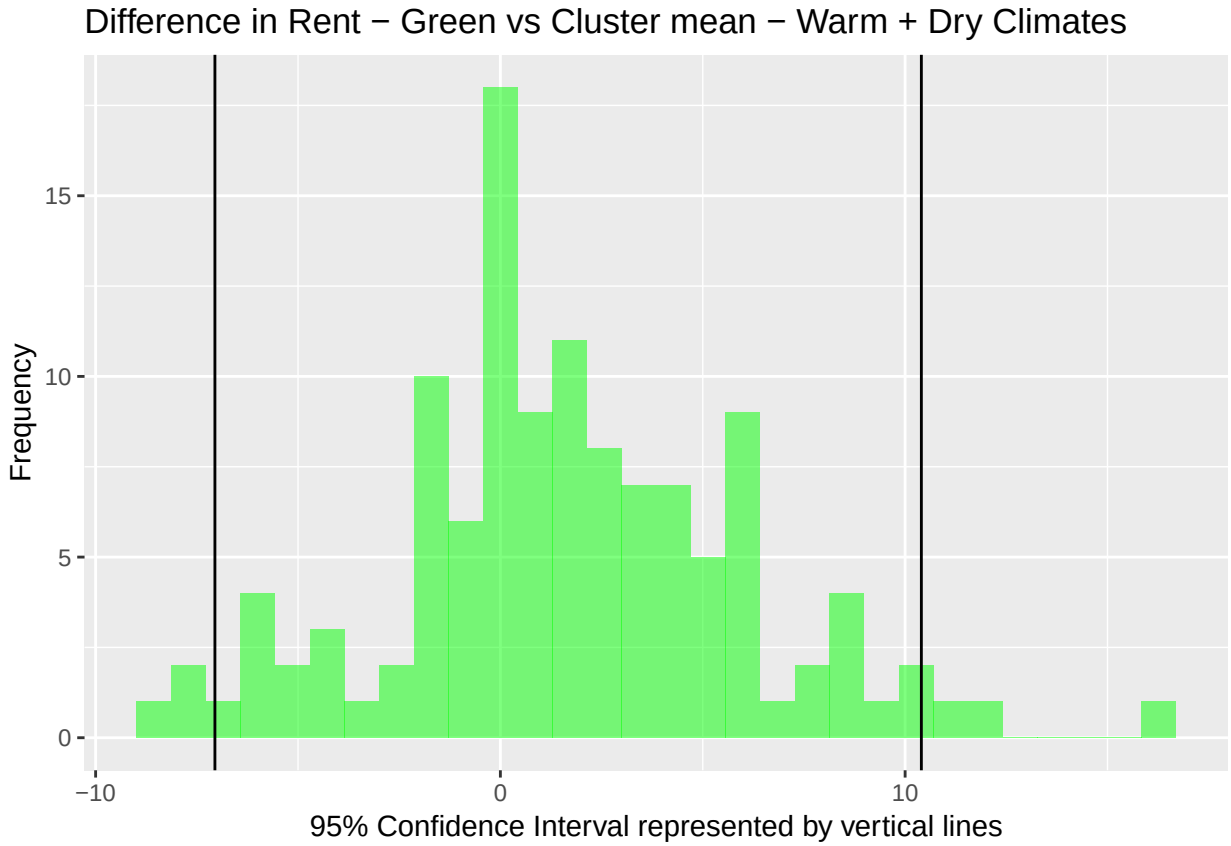
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We observe that the mean difference between a Green building and those in its cluster is about $3.12 while the median is $2.10. In either case, it is similar to what the staffmember had found. However, we do observe a couple outliers. Even with those, the 95% confidence interval shows that there is a significant amount of Green buildings whose rent was less than the mean of its cluster.

However, we have more information specific to our building. What if we drilled a little deeper and only compared the rent of Green buildings to those in its cluster in markets with similar climates to ours? Austin is relatively warm and dry, so we'll use those as a starting point.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

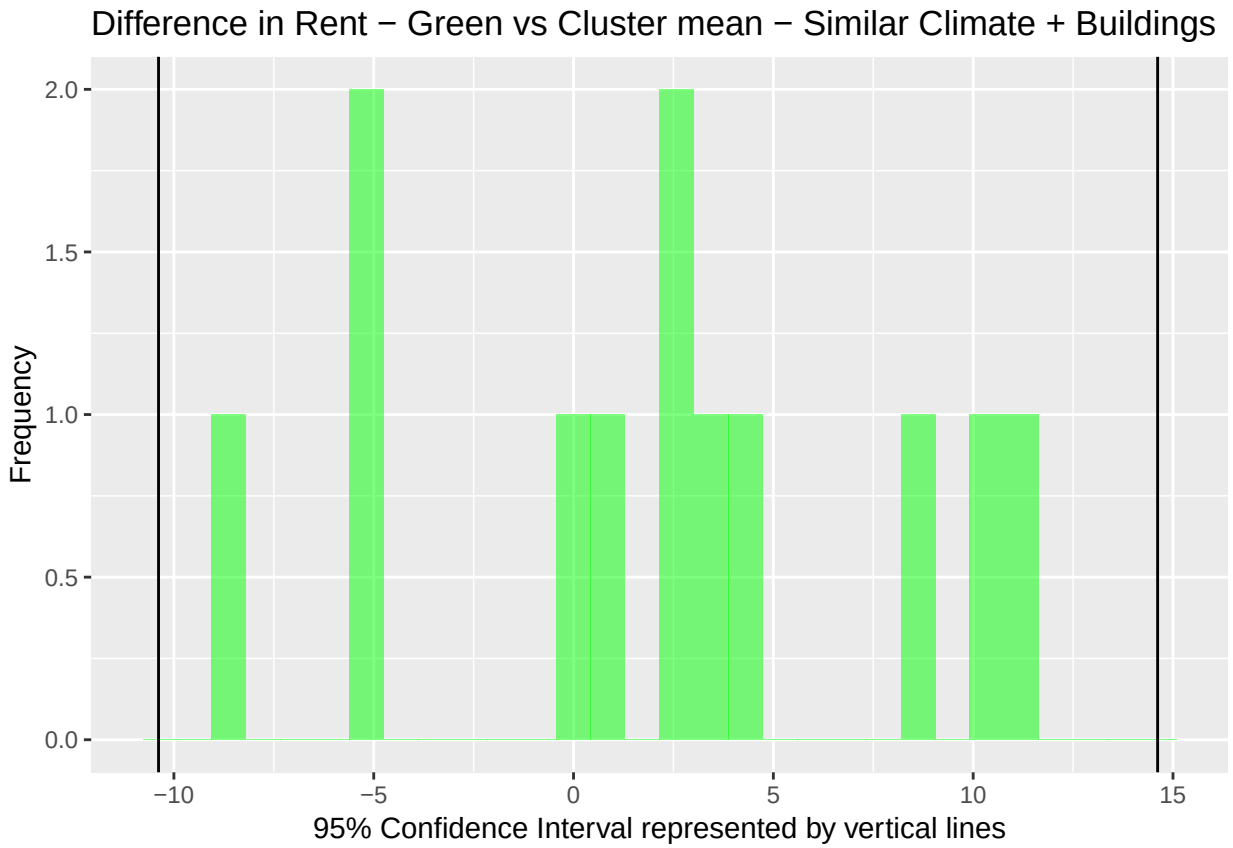



```
## [1] 10.40129
```

With these stipulations, we see a mean difference of just $1.67 and a median of $1.32. This likely is aided by losing the extreme outliers on the positive end from the previous metric. We also might posit that in areas with more extreme climates, people are more concerned with the climate-controlling functionality of a building rather than its green rating or features? This is supported by a correlation matrix where we can see that Electricity Costs and Total Days Heating / Cooling have the strongest relationship with Rent, but neither is particularly correlated with whether a building is or is not Green certified.

There is a risk of whittling away too many observations, but we can continue to get even more specific to the features of our building. In particular, we can focus on these aspects: * Size - 250,000sqft * Stories - 15 * Amenities - yes (mixed use)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Here we observe a mean difference within cluster of $2.12 and a median of $2.89. However, we likely do not have enough data points at this specificity.

To reiterate our conclusion, we cannot confidently predict that you will be able to charge a premium in Rent if you were to invest in meeting the requirements for Green certification.

Part 3: Bootstrapping

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: lattice
```

```
## Loading required package: ggformula
```

```
## Loading required package: ggstance
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh

##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Loading required package: Matrix

##
## 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.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##
## Attaching package: 'mosaic'

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

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

## 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

## Loading required package: xts

## Loading required package: zoo

##
## 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
```

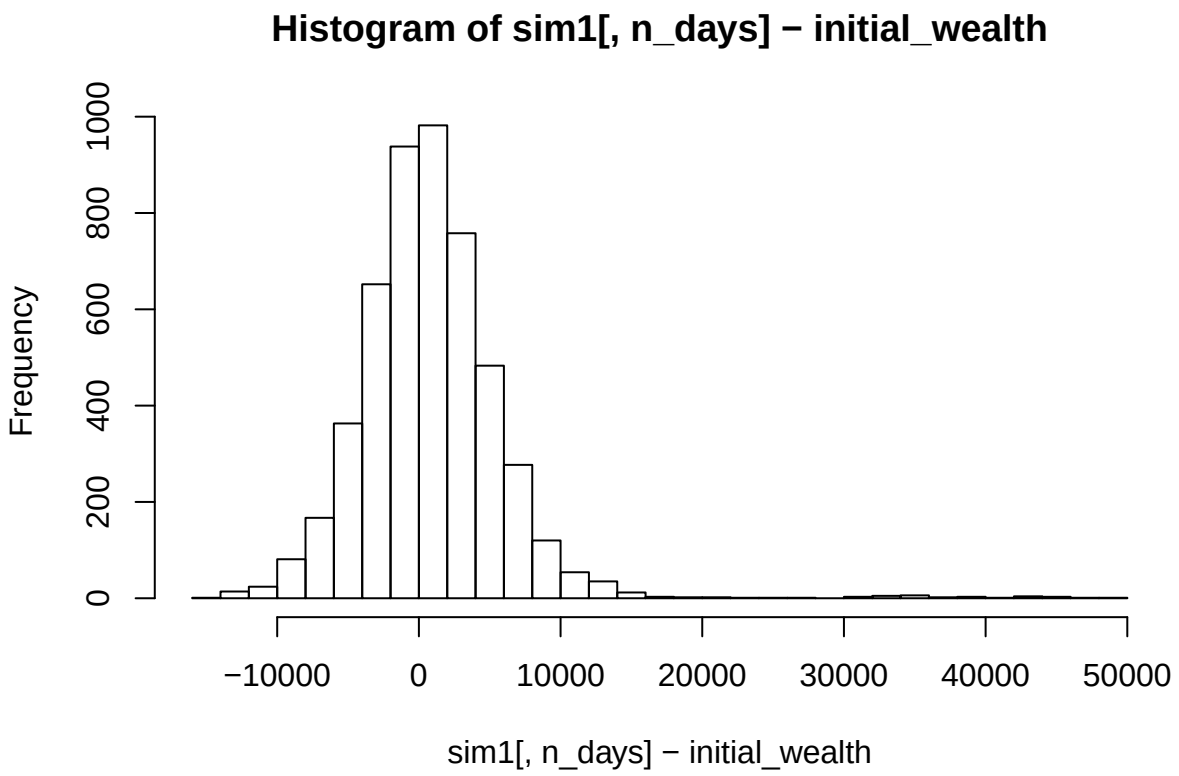
```
## Version 0.4-0 included new data defaults. See ?getSymbols.

## 'getSymbols' currently uses auto.assign=TRUE by default, but will
## use auto.assign=FALSE in 0.5-0. You will still be able to use
## 'loadSymbols' to automatically load data. getOption("getSymbols.env")
## and getOption("getSymbols.auto.assign") will still be checked for
## alternate defaults.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.warning4.0"=FALSE). See ?getSymbols for details.

##
## WARNING: There have been significant changes to Yahoo Finance data.
## Please see the Warning section of '?getSymbols.yahoo' for details.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.yahoo.warning"=FALSE).

##                ClCl.SPYa    ClCl.TLTa     ClCl.LQDa    ClCl.EEMa
## 2007-01-03            NA           NA             NA           NA
## 2007-01-04  0.0021221123  0.006063328  0.0075152938 -0.013809353
## 2007-01-05 -0.0079763183 -0.004352668 -0.0006526807 -0.029238205
## 2007-01-08  0.0046250821  0.001793566 -0.0002798843  0.007257535
## 2007-01-09 -0.0008498831  0.000000000  0.0001866169 -0.022336235
## 2007-01-10  0.0033315799 -0.004475797 -0.0013063264 -0.002303160
##               ClCl.VNQa
## 2007-01-03           NA
## 2007-01-04  0.001296655
## 2007-01-05 -0.018518518
## 2007-01-08  0.001451392
## 2007-01-09  0.012648208
## 2007-01-10  0.012880523
```

In our "aggressive" portfolio, we decided to allocate 75% in emerging markets, 15% in US domestic equities, and 10% in real estate, while our "conservative" portfolio had 80% allocated in US treasury bonds and then 20% in US domestic equities.

Traditionally, US treasury bonds (TTL) are the most secure with the lowest yield followed by investment-grade corporate bonds (LQD), US domestic equities (SPY), real-estate market (VNQ), and finally emerging-market equities (EEM).

After evaluating each asset and portfolio, we will then state our analysis on each asset's risk/return properties based on empirical data.

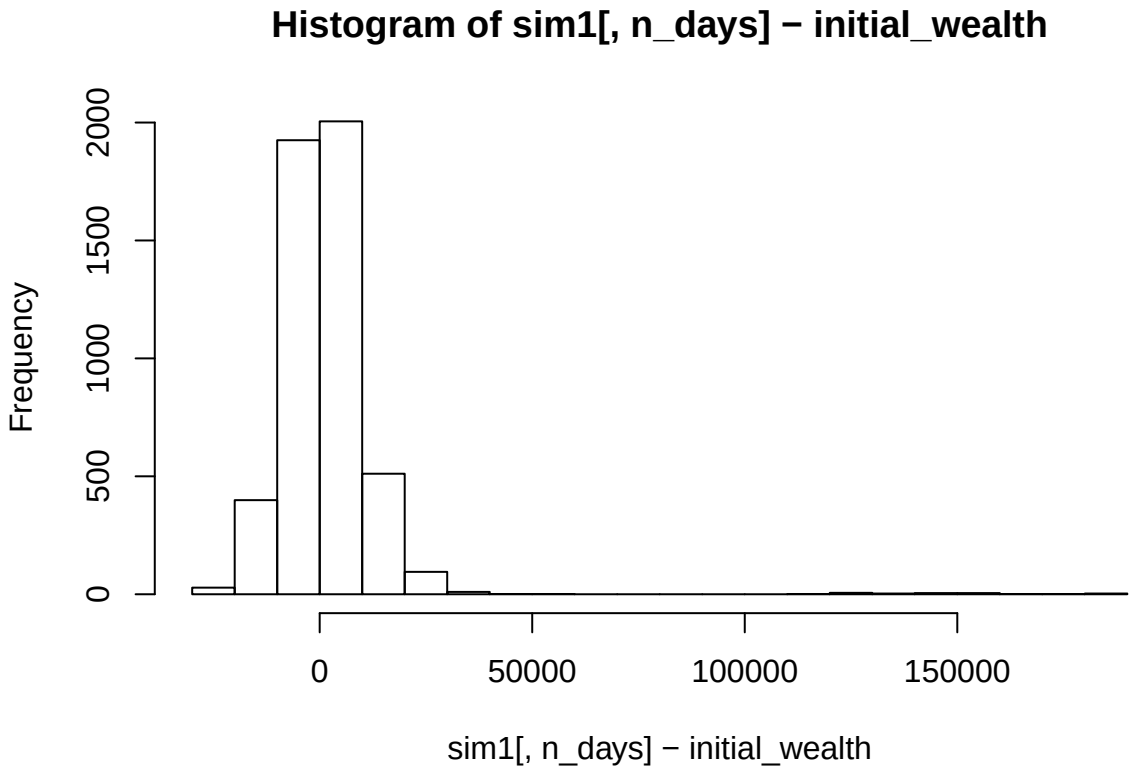
```
## [1] 100827.9
```




```
##        5%
## -6269.325
```

We simulated the even-weighted portfolio over a two-week period five thousand times to discover that the average return was approximately 966.50 USD, leaving us with 100,966.50 USD. According to the value-at-risk measure, we have a 5% probability of losing 5,906.92 USD or more in two weeks.

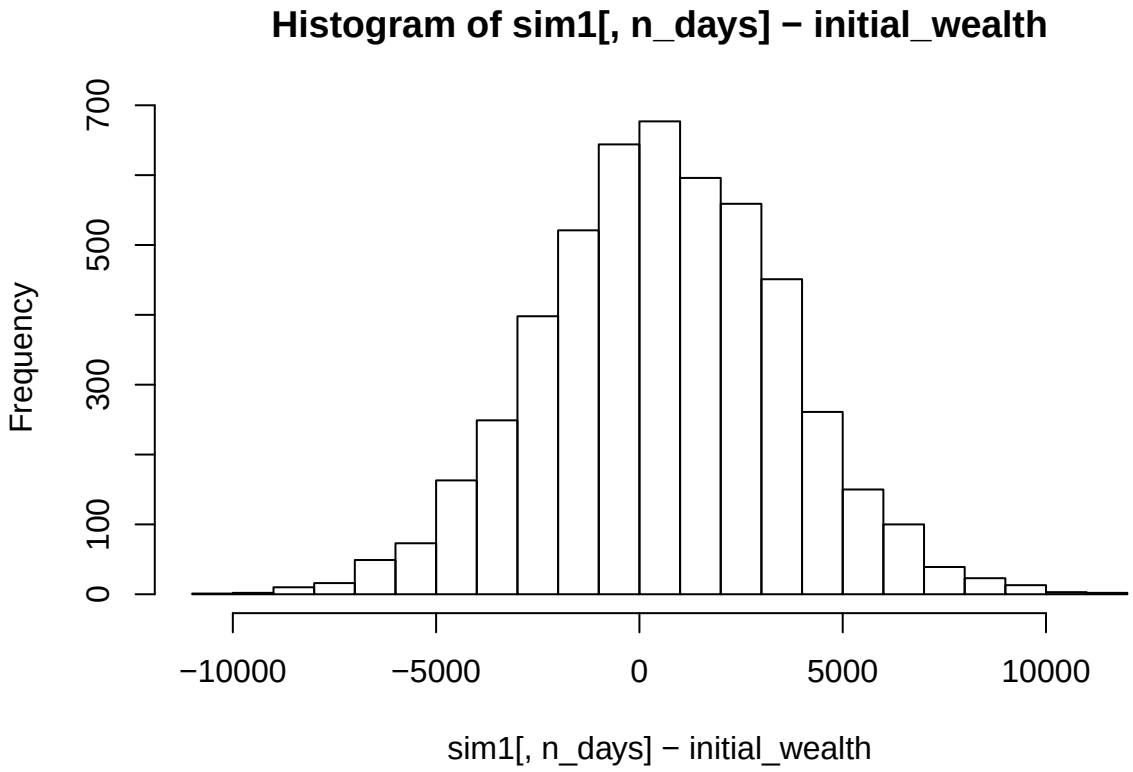
```
## [1] 101538.8
```




```
##        5%
## -12438.85
```

We then simulated our "aggressive" portfolio over a two-week period five thousand times to notice that we averaged a significantly higher return than the even-weighted portfolio, gaining 1,552.30 USD in the same span of time. The portfolio's heavy focus in emerging markets invites a far higher level of risk, since according to the value-at-risk measure, there is a 5% probability of losing 12,475.04 USD or more.

```
## [1] 100572.3
```




```
##        5%
## -4368.806
```

Finally, we simulated our more conservative portfolio that placed a heavy emphasis on US treasury bond over a two-week period, doing so five thousand times. This more secure portfolio only netted an average return of 612.00 USD, leaving us with 100,612 USD in the end. Unsurprisingly, the value-at-risk measure showed that we have a 5% probability of losing 4,166.68 USD or more, showing that it is far less risky than the two other portfolios in question.

```
##    ClCl.SPYa    ClCl.TLTa    ClCl.LQDa    ClCl.EEMa    ClCl.VNQa
## 0.0003981241 0.0002787785 0.0002095494 0.0009813682 0.0004157265
```

```
## [1] "ClCl.SPYa"
##          5%
## -0.01882661
## [1] "ClCl.TLTa"
##          5%
## -0.01473998
## [1] "ClCl.LQDa"
##           5%
## -0.006342492
## [1] "ClCl.EEMa"
##          5%
## -0.02864301
## [1] "ClCl.VNQa"
##          5%
## -0.02798942
```

Evaluation of Emerging Markets' Risk/Return Properties

Empirically, we can see that emerging-market equities (EEM) are the most volatile and riskiest of asset classes, yet they deliver a far higher return as evidenced by our simulation with the aggressive portfolio. It averages a daily return rate of approximately .0996%, the highest of all assets, while the 5% value-at-risk measure states that we will lose -2.8% in a day or more, matching up with our traditional understanding of emerging markets.

Evaluation of Real-Estate Market's and US Domestic Equities' Risk/Return Properties

We can see that the real-estate market (VNQ) and the US domestic equities (SPY) offers similar average daily return rates of .0412% and .0398% respectively. Both are in the middle of the pack when it comes to returns as anticipated. However, the stock market should offer higher rates of returns than real-estate markets due to being innately more volatile and therefore riskier than real estate. We have also noticed that real-estate market is slightly riskier than the stock market with potentially higher losses according to its 5% value-at-risk measure, comparable to emerging markets. It states that we have a 5% probability of losing -2.8% or more in a day. Yet, we know that the real-estate market should be less volatile or risky than the stock market due to being illiquid. Conversely, US domestic equities' 5% value-at-risk measure states that we have a 5% probability of losing 1.8% or more in a day, a potential loss rate that fits better with traditional rates. While their place in relative risk and returns compared to other markets remain true to traditional understanding, the same cannot be said when they are compared to one another. Something has caused this discrepancy to occur.

Evaluation of US Treasury Bonds' and Investment-Grade Corporate Bonds' Risk/Return Properties

As bonds, the US treasury and investment-grade corporate bonds averages a low daily return rate at .0273% and .0208% respectively which contradicts our traditional understanding. In fact, the 5% risk-measure reflects this anomalous discrepancy by showing that investment-grade corporate bonds has a 5% probability of losing .6% or more in a day compared to the treasury bonds' 1.4% or more. They certainly fit into the traditional picture when compared to other asset classes, however, the same cannot be said when the two bonds are compared to one another. Again, something must have occurred to make the most secure bond in the world to appear more volatile and riskier than investment-grade corporate bonds.

The slight deviance from traditional wisdom may be related to the housing crisis and the financial crisis that followed in 2006 and 2007 respectively.

In light of this revelation, we should stress to the investor that the value-at-risk model essentially ignores the tail-end of outcomes, leading to a skewed perception about an asset's risk. Certainly, the probability of the unlikely "bad" events are low in most day-to-day situations, however, we have seen in 2008, like when the housing bubble and then financial sector collapsed, that the unlikely can become quite likely under certain economic constraints. Furthermore, we have observed that the markets are correlated with one another to the point that any negative returns in one market may lead to simultaneous negative returns in other markets, similar to what was witnessed in the Financial Crisis of 2007-2008 that followed the collapse in the housing market in 2006. While the 2007-2008 years are included in this measure of risk, value-at-risk may be ignoring the risk present due to how it completely ignores the tail-end of its outcomes. This type of risk is simply not calculated here, and the investor should be aware that they may incur significantly higher losses than anticipated when the unlikely indeed does strike.

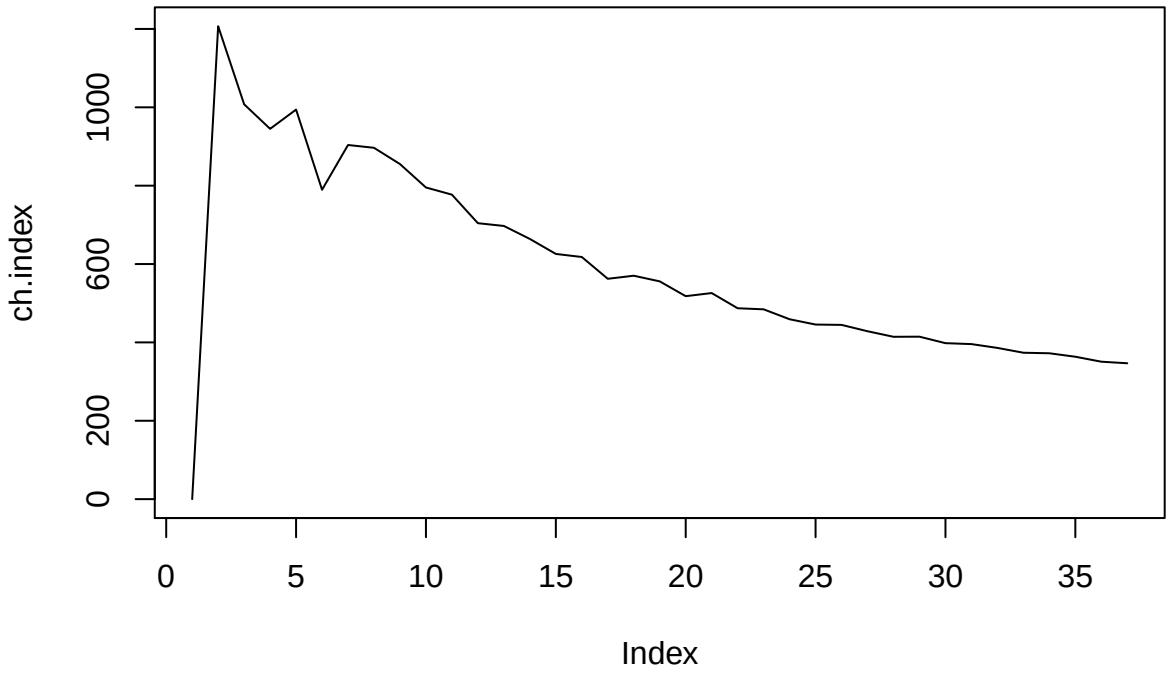
Part 4: Market Segmentation

Method

The data was pre-processed by removing all accounts that were flagged for spam or adult content. All missing values were omitted from the data. K-means clustering and principal component analysis was used on the twitter users data. K-means was performed on unscaled and scaled data and optimized for maximum score on the CH index. In both cases, the optimal number of clusters was two. When performing PCA, we found that the first component was only able to capture 12% of the variance in the data. In order to capture 80% of the variance in the data, over 20 components are required. Since there are so many components needed to describe a sufficient amount of variability in the data, only k-means clustering will be considered going forward.

Findings

When evaluating the means for each feature in the centers in K-means clustering, the two core market segments seem to be differentiable based strictly on the number of tweets in the "Food," "Health & Nutrition," "Cooking," "Outdoors," and "Personal Fitness" categories. That is to say, for all other categories, the two clusters have roughly the same number of tweets for each of the remaining categories. For the five categories previously mentioned, one cluster has exceedingly more tweets in each when compared to the other cluster. From this result, one of the major market segments for the company is an audience primarily focused on maintaining healthy diet and exercise. Unsatisfied with the simplicity of this finding, the k-means algorithm was implemented repeatedly to find the next best number of clusters. In these simulations, three was found to be the next best value of k based on CH index. When using three clusters, our first finding remains intact, as one cluster center around more tweets in the "Health & Nutrition," "Outdoors," and "Personal Fitness" categories. However, a new market segment becomes apparent as one cluster has a noticeably high amount of average tweets in the "Chatter," "Photo Sharing," "Beauty," and "Fashion" categories. We can refer to this cluster as the "Socialites." Their Twitter behavior reflects social media savvy, personal branding, and a familiarity with Instagram aesthetic.

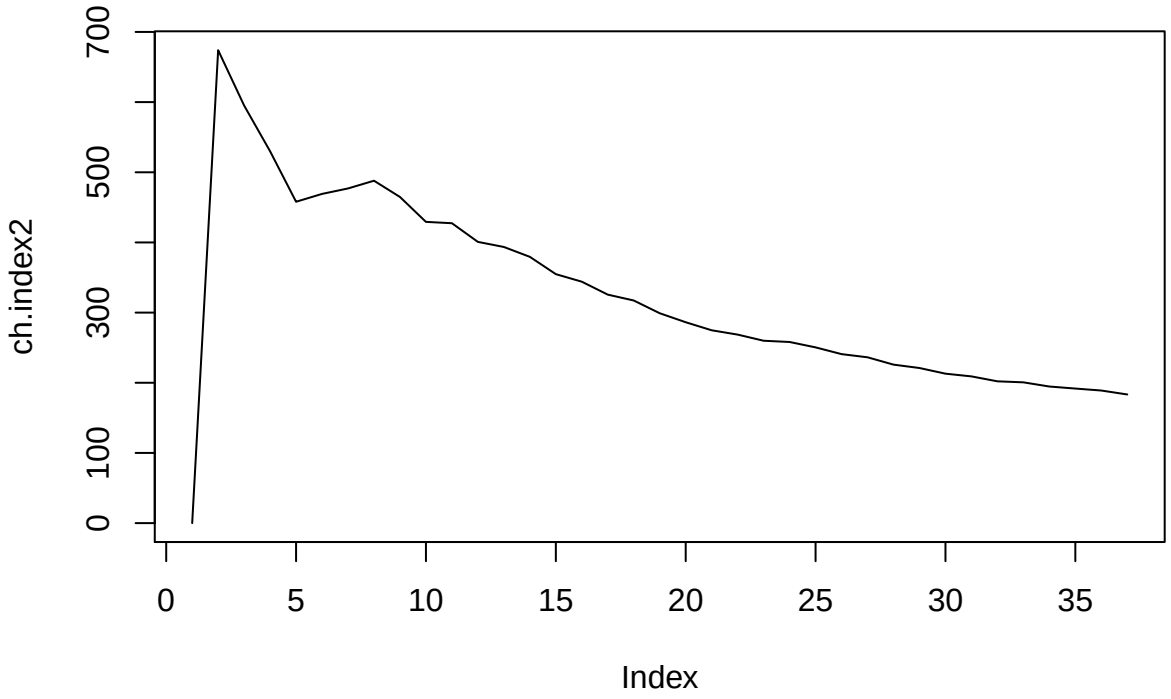



```
## [1] 1206.93
```

```
##  [1]    0.0000 1206.9303 1007.7146  945.0776  994.1894  789.5278  903.7859
##  [8]  896.6896  855.1470  795.3917  777.0567  704.2703  697.2080  664.0166
```

```
## [15]  625.7560  617.9704  562.3403  570.1584  555.7799  517.9916  525.9148
## [22]  487.1475  484.4568  459.1398  445.5986  444.7103  428.4630  414.3006
## [29]  414.5996  398.0583  395.6735  385.9139  373.6055  372.2590  363.4054
## [36]  350.7852  346.7670
```

```
## Warning: did not converge in 10 iterations
```

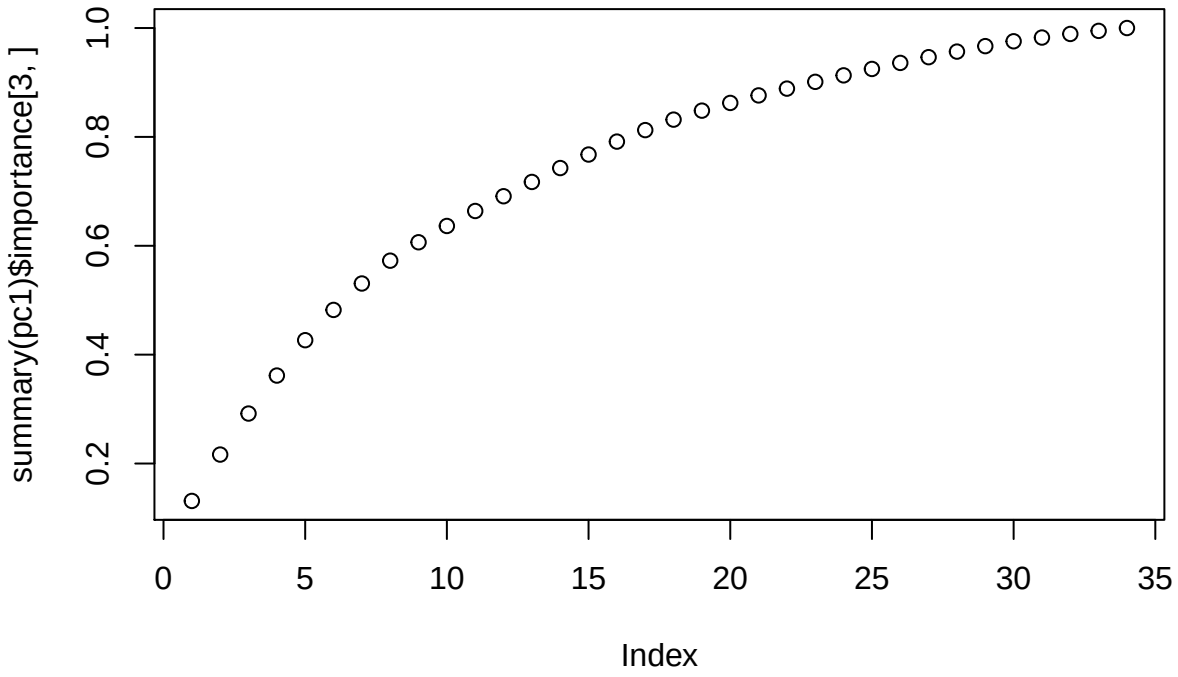



```
## [1] 673.9647
```

```
##  [1]   0.0000 673.9647 595.1317 529.9130 457.9952 469.1783 476.9707
##  [8] 487.9786 464.7958 429.2911 427.3880 400.7831 393.5170 379.4875
## [15] 354.8110 344.0347 325.6016 317.2395 299.1070 286.2907 274.8154
## [22] 268.6451 259.9088 258.1197 250.4571 240.8121 236.3048 225.8391
## [29] 221.0280 212.9553 209.0791 202.0564 200.6124 194.6338 191.7756
## [36] 188.9440 183.3073
```

```
## Importance of components:
##                            PC1     PC2     PC3    PC4     PC5     PC6
## Standard deviation     2.1128 1.70186 1.60046 1.5406 1.48585 1.37505
## Proportion of Variance 0.1313 0.08519 0.07534 0.0698 0.06493 0.05561
## Cumulative Proportion  0.1313 0.21648 0.29181 0.3616 0.42655 0.48216
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     1.28487 1.19405 1.07174 1.00921 0.96766 0.95978
## Proportion of Variance 0.04856 0.04193 0.03378 0.02996 0.02754 0.02709
## Cumulative Proportion  0.53072 0.57265 0.60643 0.63639 0.66393 0.69102
##                           PC13    PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.94424 0.93157 0.91924 0.89861 0.84765 0.80885
## Proportion of Variance 0.02622 0.02552 0.02485 0.02375 0.02113 0.01924
## Cumulative Proportion  0.71725 0.74277 0.76762 0.79137 0.81251 0.83175
##                           PC19    PC20    PC21    PC22   PC23    PC24
## Standard deviation     0.74863 0.69470 0.68513 0.65208 0.6493 0.63701
## Proportion of Variance 0.01648 0.01419 0.01381 0.01251 0.0124 0.01193
## Cumulative Proportion  0.84823 0.86243 0.87623 0.88874 0.9011 0.91307
##                           PC25    PC26    PC27    PC28    PC29    PC30
```

```
## Standard deviation     0.63245 0.61457 0.59777 0.59096 0.58382 0.55069
## Proportion of Variance 0.01176 0.01111 0.01051 0.01027 0.01002 0.00892
## Cumulative Proportion  0.92484 0.93595 0.94646 0.95673 0.96675 0.97567
##                           PC31    PC32    PC33    PC34
## Standard deviation     0.48261 0.47497 0.43660 0.42198
## Proportion of Variance 0.00685 0.00664 0.00561 0.00524
## Cumulative Proportion  0.98252 0.98916 0.99476 1.00000
```




```
## [1] 1201 6174
```

```
##    chatter current_events   travel photo_sharing uncategorized  tv_film
## 1 4.337219       1.578684 1.538718      3.041632     0.9708576 1.104913
## 2 4.408811       1.509232 1.580499      2.647554     0.7695173 1.073210
##   sports_fandom politics     food    family home_and_garden     music
## 1      1.516236 1.664446 2.183181 0.9450458       0.6136553 0.7935054
## 2      1.609006 1.834629 1.229673 0.8368967       0.4941691 0.6624555
##       news online_gaming shopping health_nutrition college_uni
## 1 1.282265      1.227311 1.509575       11.2206495    1.412157
## 2 1.199061      1.196955 1.380790        0.8995789    1.589083
##   sports_playing  cooking       eco computers  business  outdoors
## 1      0.7310575 4.115737 0.8376353 0.6910908 0.4854288 2.2539550
## 2      0.6256884 1.597344 0.4361840 0.6333009 0.4126984 0.4782961
##      crafts automotive       art religion    beauty parenting    dating
## 1 0.6627810  0.7352206 0.8659450 1.124896 0.8368027 1.0049958 1.0524563
## 2 0.4802397  0.8352770 0.6827017 1.085358 0.6763848 0.8896987 0.6448008
##      school personal_fitness  fashion small_business
## 1 0.7860117        5.6161532 1.264779      0.3097419
## 2 0.7529964        0.6528993 0.946874      0.3299320
```

```
## [1] 5494 1132  749
```

```
##    chatter current_events   travel photo_sharing uncategorized  tv_film
## 1 4.412632       1.486167 1.101201      2.660721     0.7721150 1.059701
## 2 4.318021       1.564488 1.294170      2.988516     0.9681979 1.090989
```

```
## 3 4.403204       1.706275 5.461949      2.667557     0.7730307 1.196262
##   sports_fandom  politics     food    family home_and_garden     music
## 1      1.507827 0.9446669 1.170914 0.8134328       0.4807062 0.6614489
## 2      1.521201 1.2985866 2.221731 0.9408127       0.6086572 0.7985866
## 3      2.335113 8.8998665 1.690254 1.0253672       0.6114820 0.6742323
##        news online_gaming shopping health_nutrition college_uni
## 1 0.6914816      1.205315 1.386422        0.9200946    1.595377
## 2 1.1590106      1.253534 1.500000       11.4558304    1.410777
## 3 5.1161549      1.098798 1.365821        1.3444593    1.528705
##   sports_playing  cooking       eco computers  business  outdoors
## 1      0.6179469 1.627776 0.4179104 0.4266473 0.3853295 0.4402985
## 2      0.7287986 4.051237 0.8507067 0.6051237 0.4717314 2.2888693
## 3      0.6955941 1.703605 0.5874499 2.2843792 0.6408545 0.8678238
##      crafts automotive       art religion    beauty parenting    dating
## 1 0.4615945  0.6543502 0.6842009 1.055333 0.6814707 0.8514743 0.6068438
## 2 0.6696113  0.7049470 0.8454064 1.127208 0.8091873 1.0035336 1.0220848
## 3 0.6234980  2.1989319 0.7196262 1.305741 0.6955941 1.1829105 1.0066756
##      school personal_fitness   fashion small_business
## 1 0.7358937        0.6570805 0.9639607      0.3165271
## 2 0.7729682        5.7402827 1.2120141      0.2871025
## 3 0.9012016        0.8918558 0.9305741      0.4606142
```