STA 380, Part 2: Exercises

Tanushree Devi Balaji, Spoorthi Anupuru

2022-08-15

1. Probability practice

Part A.   Visitors to your website are asked to answer a single survey question before they get access to the content on the page. Among all of the users, there are two categories: Random Clicker (RC), and Truthful Clicker (TC). There are two possible answers to the survey: yes and no. Random clickers would click either one with equal probability. You are also giving the information that the expected fraction of random clickers is 0.3. After a trial period, you get the following survey results: 65% said Yes and 35% said No. What fraction of people who are truthful clickers answered yes? Hint: use the rule of total probability.

This is a conditional probability question. The following data is available:

DATA: RCs click yes/no with equal probability. Therefore,

P(yes/RC) = 0.5

P(no/RC) = 0.5

Also,

P(RC) = 0.3

This implies, P(TC) = 0.7

Further,

P(yes) = 0.65

P(no) = 0.35

P(yes/TC) = ?

SOLUTION: By rules of conditional probability,

P(A/B) = P(A&B) / P(B)

This implies

P(yes & RC) = P(yes/RC) * P(RC) = 0.5*0.3 = 0.15

P(no & RC) = P(no/RC) * P(RC) = 0.5*0.3 = 0.15

By rules of total probability,

P(yes & TC) = P(yes) - P(yes & RC) = 0.65 - 0.15 = 0.5

P(no & TC) = P(no) - P(no & RC) = 0.35 - 0.15 = 0.2

Hence,

P(yes/TC) = P(yes & TC)/P(TC) = 0.5/0.7 = 5/7 = 0.71428

In other words, 71.43% of truthful clickers said yes

Part B. Imagine a medical test for a disease with the following two attributes:

The sensitivity is about 0.993. That is, if someone has the disease, there is a probability of 0.993 that they will test positive.

The specificity is about 0.9999. This means that if someone doesn't have the disease, there is probability of 0.9999 that they will test negative.

In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it (or 0.000025 as a decimal probability).

Suppose someone tests positive. What is the probability that they have the disease?

DATA: Sensitivity is 0.993 and hence, P(positive/disease) = 0.993

Specificity is 0.9999 and hence, P(negative/no disease) = 0.9999

Also,

P(disease) = 0.000025

This implies, P(no disease) = 1 - 0.000025 = 0.999975

P(disease/positive) = ?

SOLUTION: By rules of conditional probability,

P(A/B) = P(A&B) / P(B)

This implies

P(positive & disease) = P(positive/disease) * P(disease) = 0.993 * 0.000025 = 0.000024825

P(negative & no disease) = P(negative/no disease) * P(no disease) = 0.9999 * 0.999975 = 0.9998750025

By rules of total probability,

P(negative & disease) = P(disease) - P(positive & disease) = 0.000025 - 0.000024825 = 0.000000175

P(negative) = P(negative & disease) + P(negative & no disease) = 0.000000175 + 0.9998750025 = 0.9998751775

P(positive) = 1 - P(negative) = 1 - 0.9998751775 = 0.0001248225

Hence,

P(disease/positive) = P(positive & disease)/P(positive) = 0.000024825/0.0001248225 = 0.198882413

In other words, there is a probability of 19.89% that someone who tests positive actually has the disease

---

2. Wrangling the Billboard Top 100

Part A.

**SOLUTION:** 21st Century Music dominates All time Billboard Top 100!

The table shows top 10 songs that spent the most weeks on the Billboard Top 100 chart. It would seem that most of these songs were released in either the 2000s or 10s and considered some of the best of the generation. While it's hard to say whether it's Billboard influencing the music scene or if it's the music scene dictating the charts, it's clear that they go hand in hand, more so now than before.

```
## # A tibble: 10 x 3
## # Groups:   performer [10]
##    performer                                      song                          count
##    <chr>                                          <chr>                         <int>
##  1 Imagine Dragons                                Radioactive                      87
##  2 AWOLNATION                                     Sail                             79
##  3 Jason Mraz                                     I'm Yours                        76
##  4 The Weeknd                                     Blinding Lights                  76
##  5 LeAnn Rimes                                    How Do I Live                    69
##  6 LMFAO Featuring Lauren Bennett & GoonRock      Party Rock Anthem                68
##  7 OneRepublic                                    Counting Stars                   68
##  8 Adele                                          Rolling In The Deep              65
##  9 Jewel                                          Foolish Games/You Were Meant~    65
## 10 Carrie Underwood                               Before He Cheats                 64
```

Part B.

SOLUTION: The volume of songs making Top 100 were high in the 60s and took a dip around the later part of the 20th century. It stayed low until the mid 2010s and took an upward turn after that. This makes sense considering the fact that artists back then produced and loaded their albums with tons of music. The music scene was in its nascent stages and there was a ton of experimentation that was happening be it with motown or soul or rock. However, following that, every decade started having its own unique sound - heavy metal in the 70s, indie rock in the 80s, pop and hip hop in the 90s and so on. What came with it were artists who dominated the respective decades and produced limited but evergreen music that stayed on the charts for longer, staying true to quality over quantity .However, the trend took a turn in the mid 2015s with the advent of apps such as Tik Tok being a major influence on the music scene. That would also explain why the charts changed so much - specially in 2020 or the year of the pandemic, when people were looking for new sources of entertainment with every waking moment.

Variation in Musical Diversity over Time
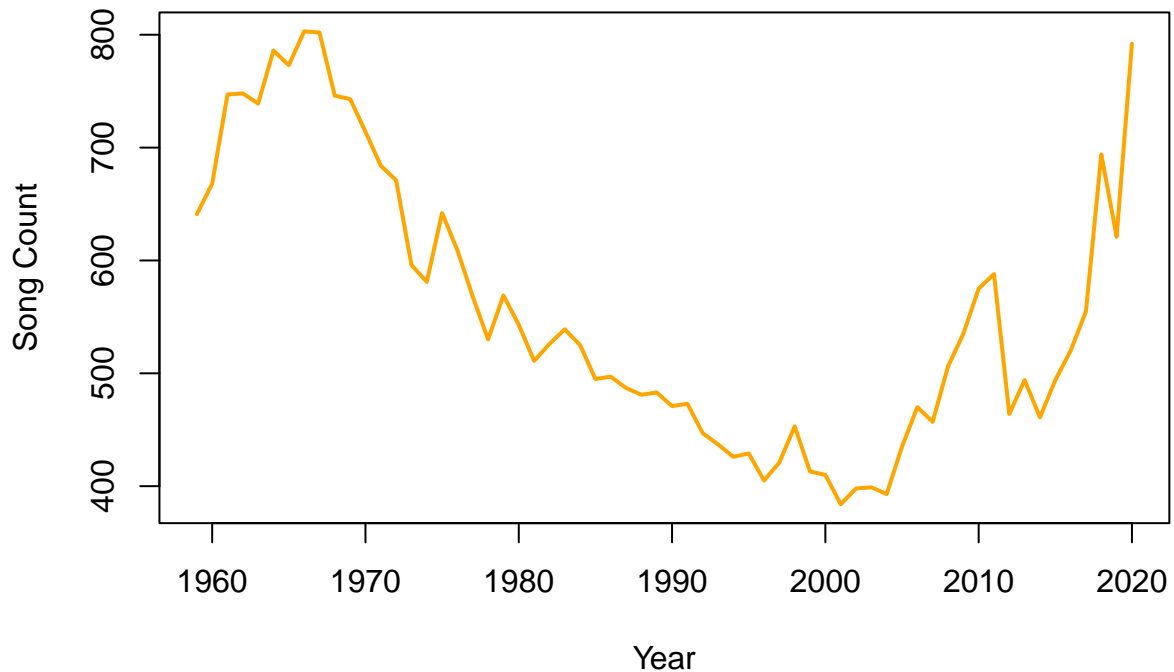
Part C.

SOLUTION: These artists who have had more than 30 10-week wonders on Billboard Top 100 are mostly global artists who are beloved and have made a name for themselves in their respective genres. People like Elton John, Stevie Wonder, Neil Diamond etc. have had very long careers and it's fair to say that they remained consistent throughout. There are also a few different country artists on there like Brad Paisley, Keith Urban etc.and also pop soloists like Taylor Swift, Madonna & Michael Jackson. It's an eclectic group and serves testament to popularity not being driven by genres.

Most Successful Artists on Billboard Top 100
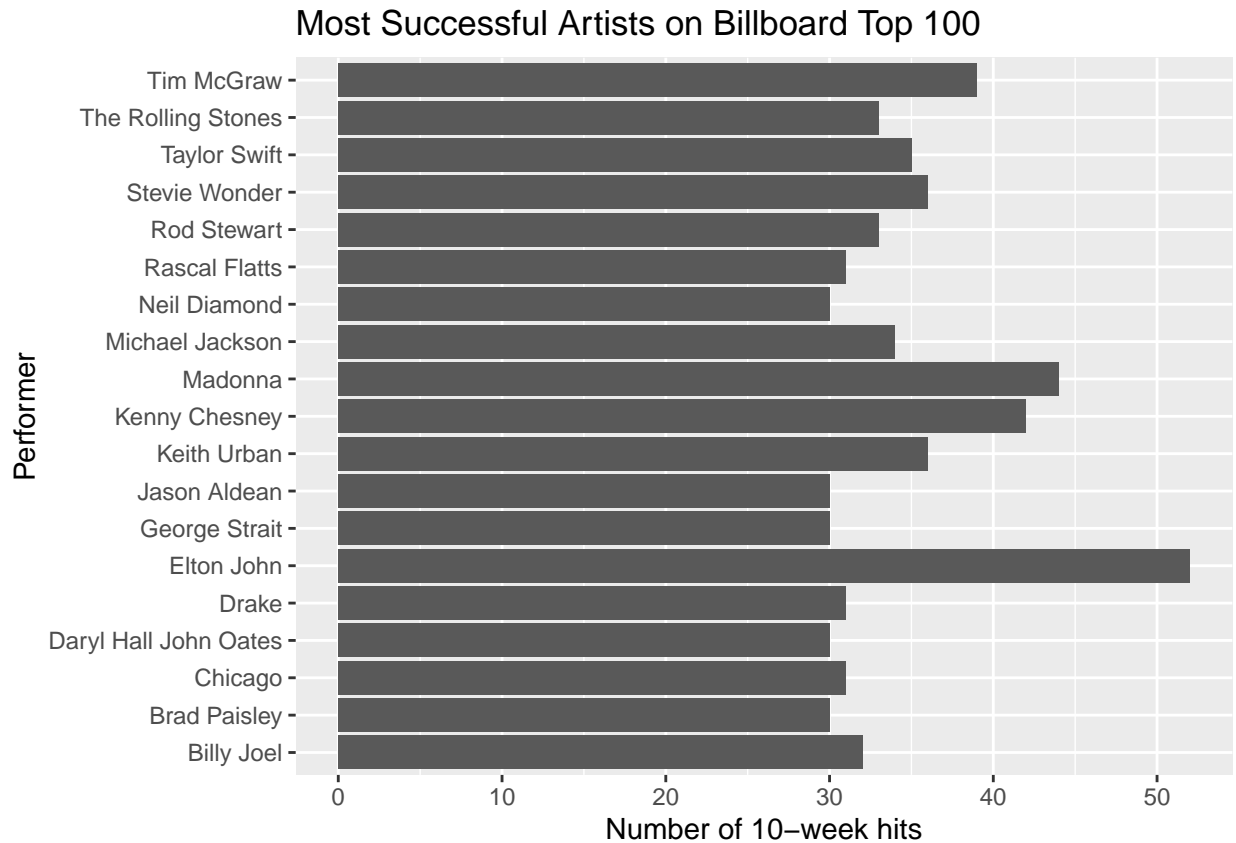
3. Visual story telling part 1: green buildings

SOLUTION: Although removing occupancy doesn't make a huge difference, let's stick with the analyst's assumption. In the "cleaned" data set, if you look at the average occupancy and size of green houses, they're very much in keeping with our desired design - around 250k in size and with about 90% expected occupancy.

```
## # A tibble: 2 x 4
##   green_rating median_size median_occupancy median_rent
##          <int>       <dbl>            <dbl>       <dbl>
## 1            0      123250             89.6        25.0
## 2            1      241199             92.9        27.6
```

However, what sticks out from analysing these factors is that the sizes of the non - green houses are rather different from the sizes of the green houses we've been looking at.While the green houses have a median sizing of around 220k, for non-green it is around 130k

Size for Green Houses

Size for Non–Green Houses
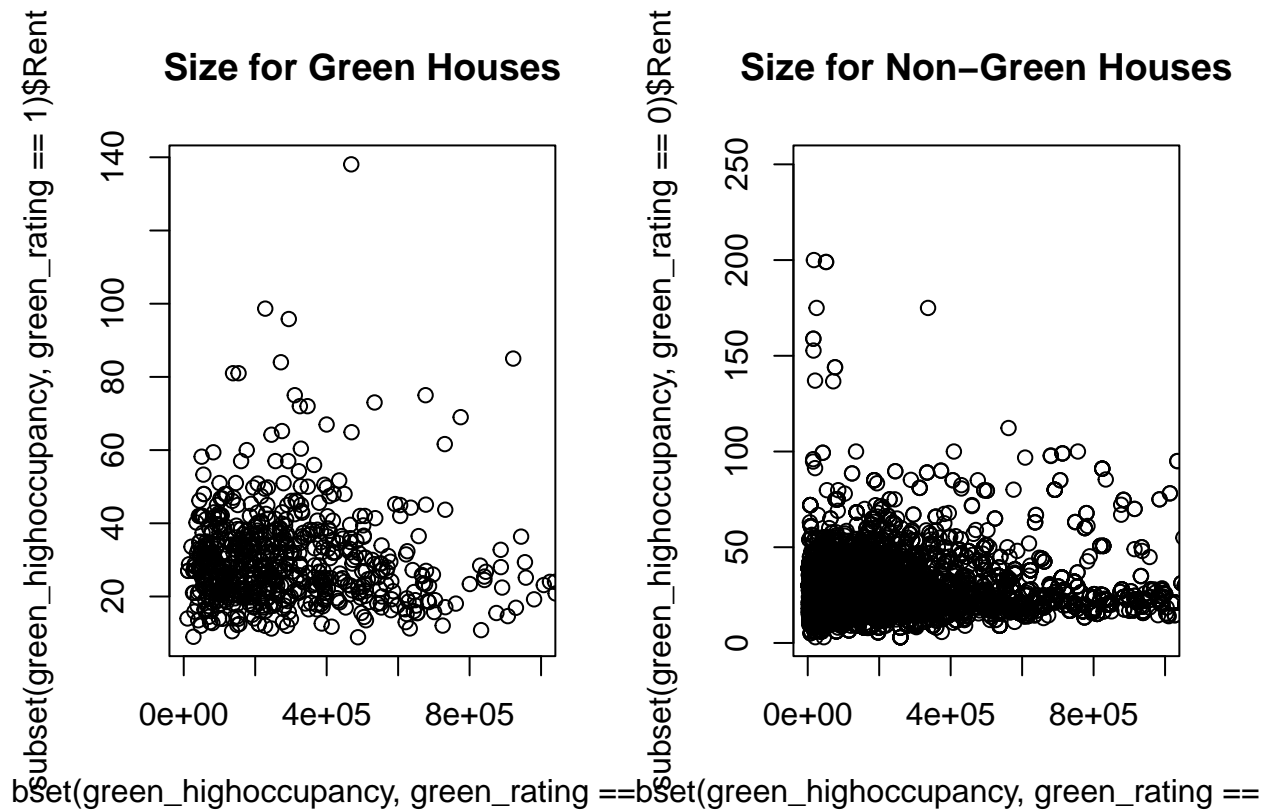
To account for size, let's define a variable is_big - If size < median value of size, 0, else 1. To "adjust" for size, let's restrict out analysis to is_big = 1 & compare only those rents. Let's look at mean rent first. Immediately what sticks out is that the rent for the "big" houses are not much different based on whether they're green or not.Given that, this is our bucket of interest, we wouldn't be making too much of extra revenue.

```
## # A tibble: 4 x 3
## # Groups:   green_rating [2]
##   green_rating is_big mean_rent
##          <int>  <dbl>     <dbl>
## 1            0      0      26.7
## 2            0      1      30.3
## 3            1      0      29.0
## 4            1      1      30.4
```

Maybe mean is not robust enough to outliers - let's consider median rent next. With this approach there is some change in the metrics again. While there is still a difference in rents, it's not as huge as was stated by the analyst. It's considerably smaller which means it would take longer to recuperate the money

```
## # A tibble: 4 x 3
## # Groups:   green_rating [2]
##   green_rating is_big median_rent
##          <int>  <dbl>       <dbl>
## 1            0      0        24.4
```

```
## 2                0       1        25.6
## 3                1       0        28.2
## 4                1       1        27.2
```

The other thing to note is that the analyst says they'll make additional "profit" based on greater rent prices. He doesn't take into account the premium in maintenance costs that comes with maintaining green properties along with other costs. Hence, this is not exactly "profit" but is additional revenue. Given that additional revenue does that necessarily represent additional cash flow, offsetting investment costs is a little trickier. To summarise, it will definitely take longer than 8 years to make that 5 million back.

---

4. Visual story telling part 2: Capital Metro data

The data includes UT bus trip information for the months of Sep, Oct and Nov 2018. Let's begin by understanding what the traffic volumes look like in each of these months. The maximum UT traffic on buses is in the Month of October (similar trends for boarding and alighting) - This is expected probably because midterms usually happen around October and students are utilizing buses to get to school & study groups

The lowest traffic however is in Novemebr, mostly because it's holiday time and people head back home for vacation.

```
## 'summarise()' has grouped output by 'month_id'. You can override using the
## '.groups' argument.
```
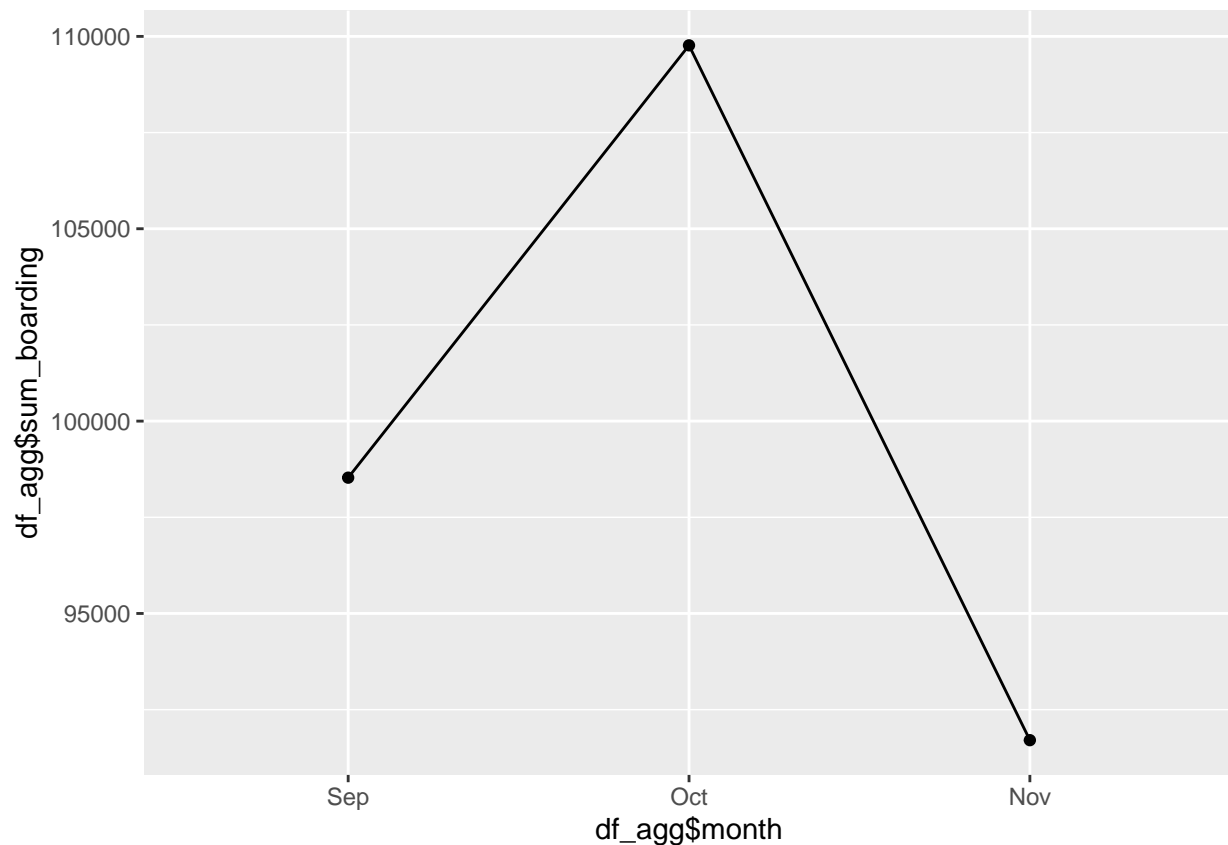
Looking at it at a slightly more granular level might help us understand when the buses are the busiest:
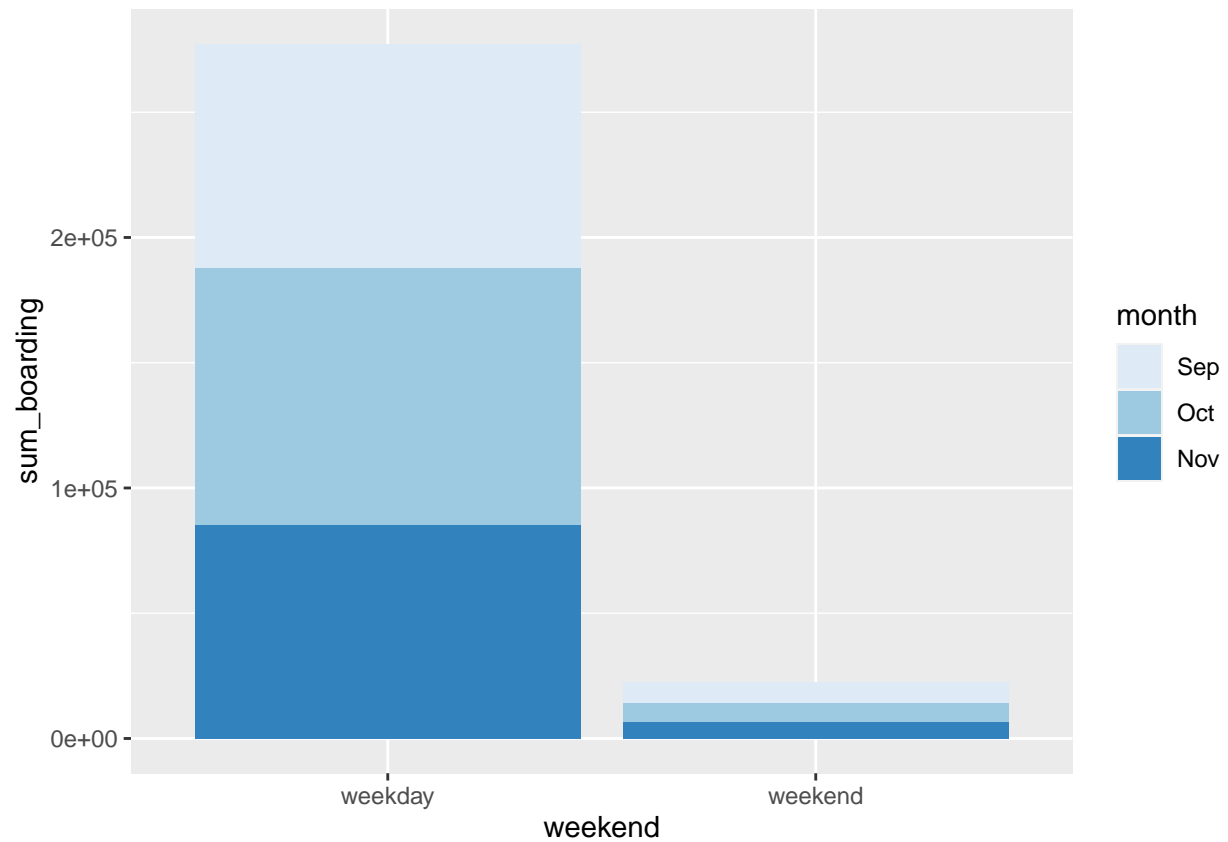
As expected, university traffic is lower during the weekends. But which days of the week are the busiest?

Interestingly, in Oct, Mon-Wed are pretty busy, also causing the spike noticed earlier. This strengthens our theory around exams and study groups.

Also, interestingly, Nov sees the lowest traffic on Wednesdays - wonder why that is. Maybe looking at this data at an hourly level might help.
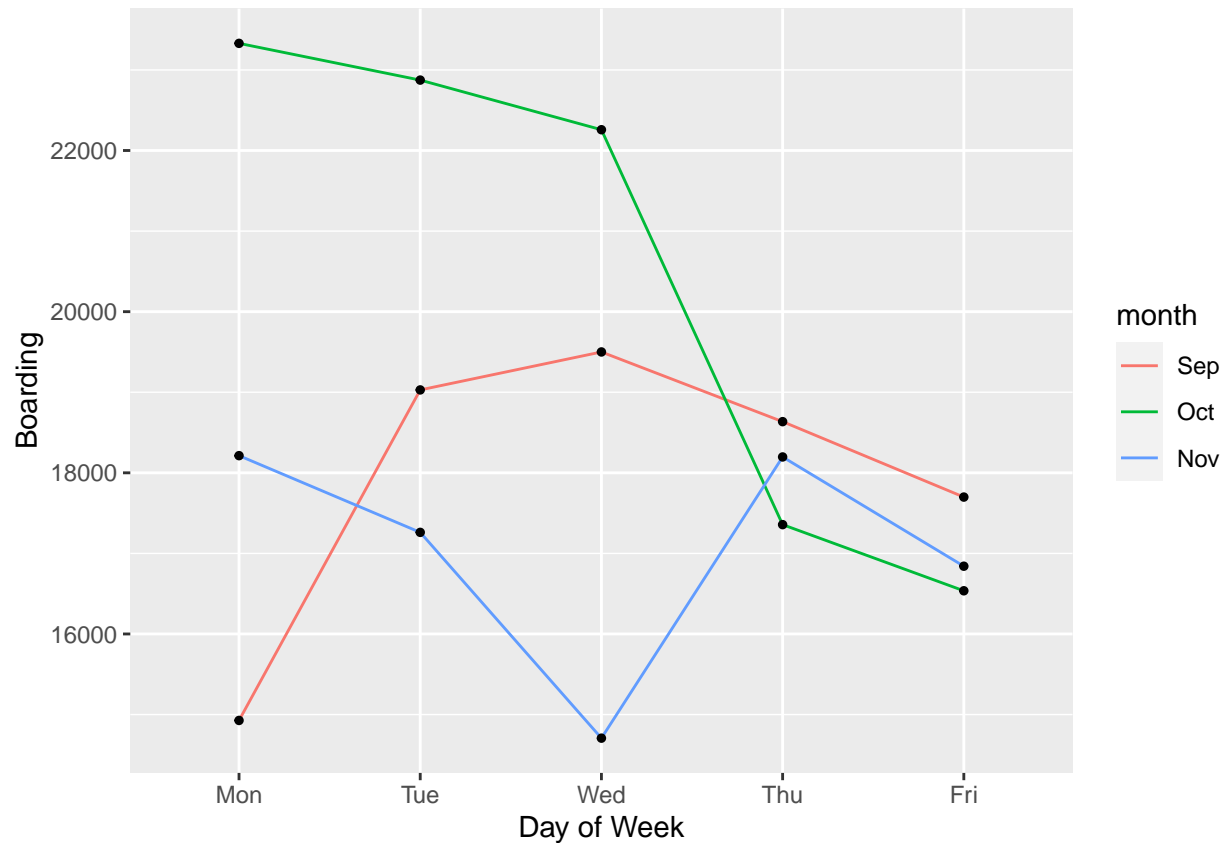
Interestingly, in Oct, MOn-Wed are pretty busy, also causing the spike noticed earlier. This strengthens our theory around exams and study groups. Also, interestingly, Nov sees the lowest traffic on Wednesdays - wonder why that is. Maybe looking at this data at an hourly level might help.

Looking at traffic on Wednesdays, you can see that the busiest hours are usually between 2 & 5 pm (ie) university hours Besides an overall dip in traffic, extreme drop during peak hours might indicate that a lot of the students had already headed back home.

With more granular data, it might be interesting to see if there were specific weeks in November when this happened. It might be an indication of the events surrounding it - ie, election week or
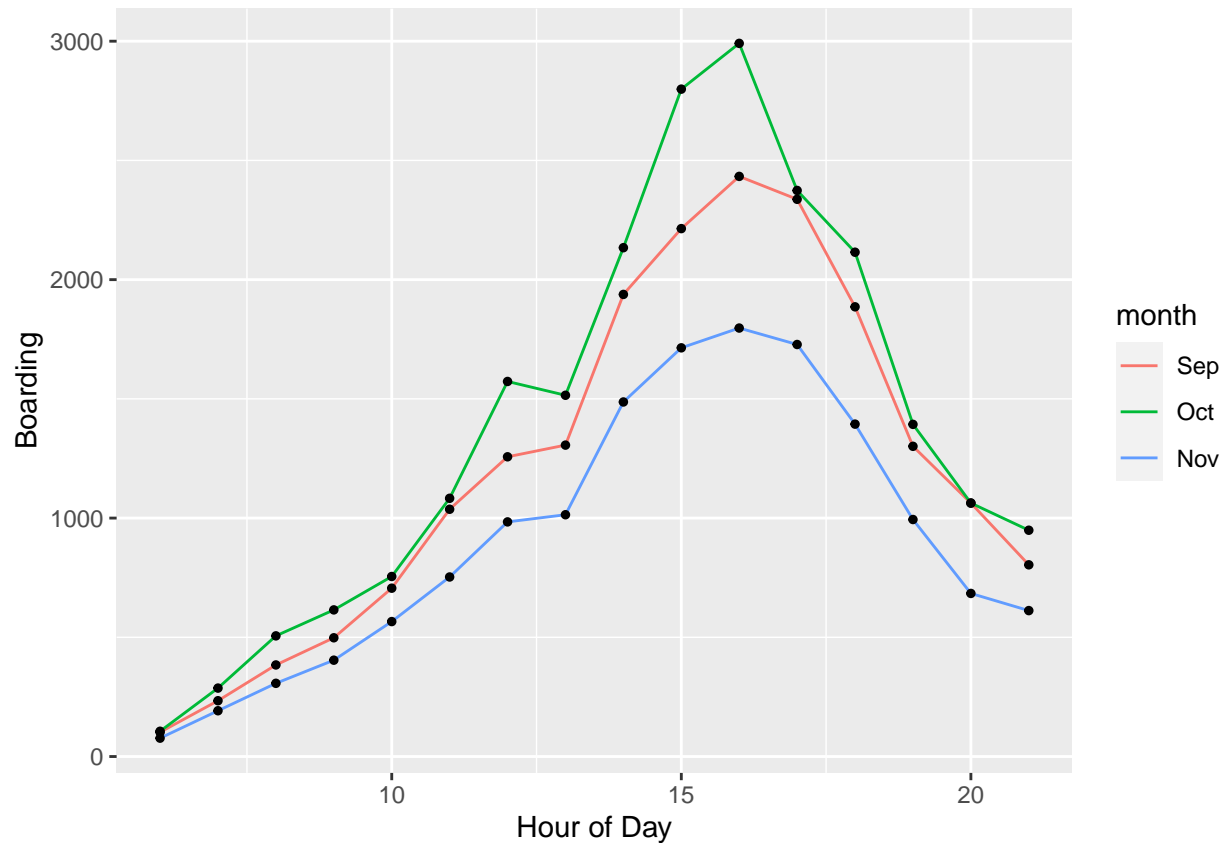
holidays etc.

---

5. Portfolio modeling

SOLUTION:

Part A.

HIGH RISK PORTFOLIO: Description: It's a tech portfolio consisting of Meta, PayPal, Visa & Intuit stocks

With this portfolio and equal allocation of costs, based on the simulations, an average cost profit of $1914.36 can be made. However, note that the standard deviation is $8396 (ie) it is highly volatile in nature.

The 5% value at risk for this stock is about -$11431 which means there is a possibility that 5% of the times you might lose over 10% of your original investment through this portfolio. However, the median profit/loss is around $1500 which means that at least 50% of the times, you would end up with a profit.

META   [2017−01−03/2022−08−12]

Last 180.5

Volume (millions):
21,239,600

PYPL   [2017−01−03/2022−08−12]

Last 101.099998

300

250

200

150

100

50

Volume (millions):
12,313,700

120
80
40
0

Jan 03   Jan 02   Jan 02   Jan 02   Jan 04   Jan 03
2017   2018   2019   2020   2021   2022

V   **[2017−01−03/2022−08−12]**

Last 211.330002

Volume (millions):
6,442,100

INTU   **[2017−01−03/2022−08−12]**

Last 486.660004

Volume (100,000s):
1,177,700

Histogram of sim1[, n_days]

[1] "Average Wealth 102052.435077415"

[1] "Standard Deviation of Wealth 8478.5020451045"

Histogram of sim1[, n_days] – initial_wealth
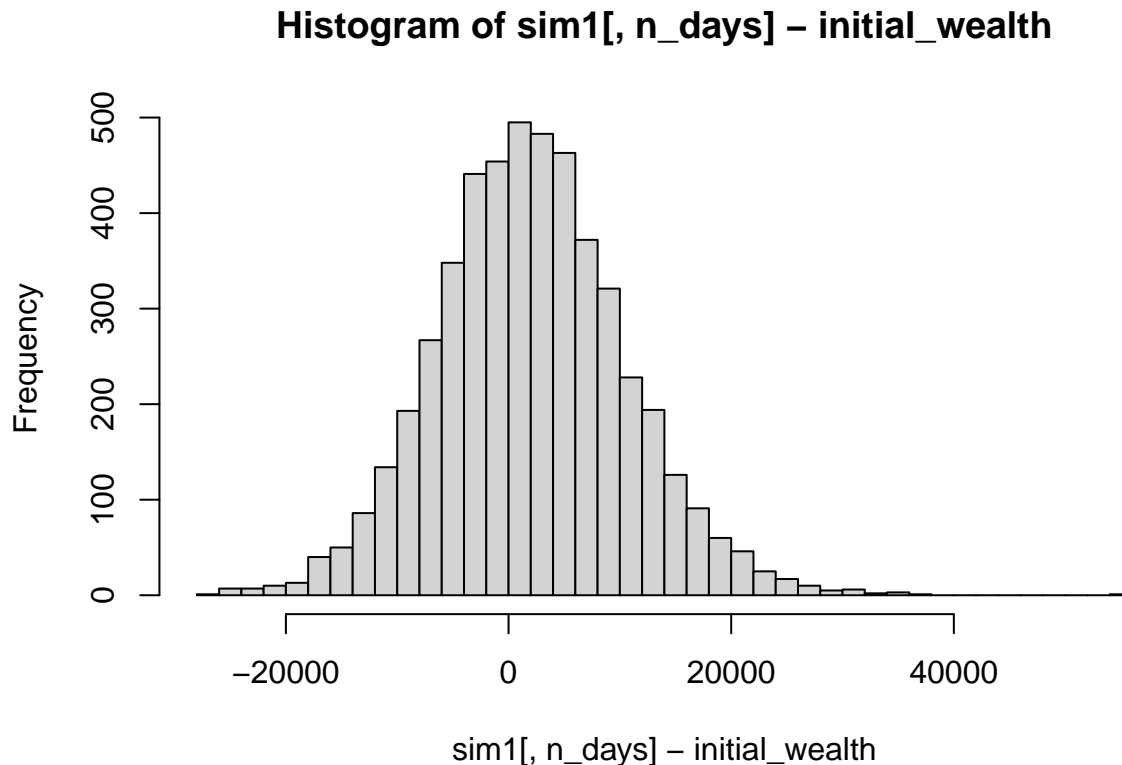
[1] "Average Profit/Loss 2052.43507741468"

[1] "Standard Deviation ofProfit/Loss 8478.5020451045"

[1] "5% VaR -11297.6032350762"

[1] "50% VaR 1788.70080209214"

Part B.

LOW RISK PORTFOLIO: Description: It's a telecom portfolio consisting of Verizon, AT&T & T-Mobile which have been relatively more stable

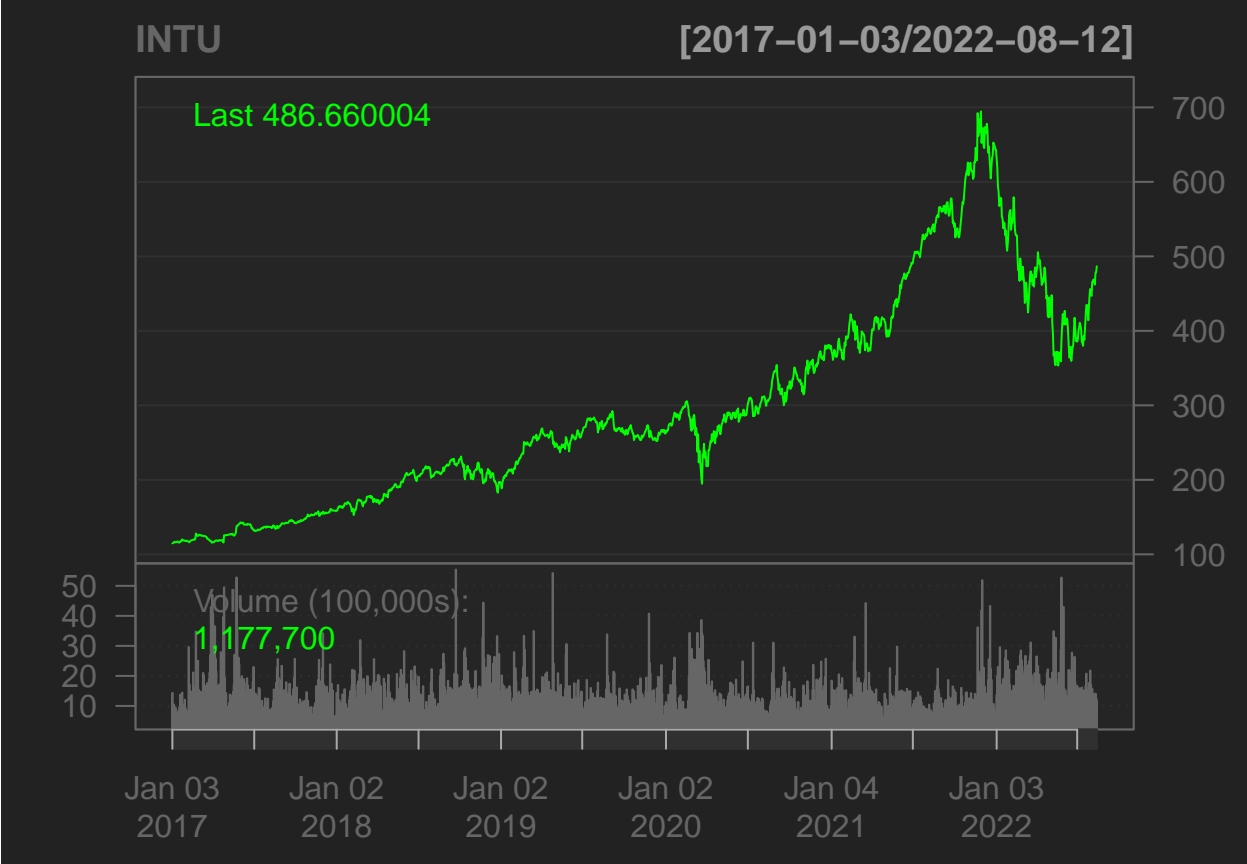With this portfolio and equal allocation of costs, based on the simulations, an average cost profit of $649.5 can be made with a standard deviation of $5652. The std dev is lower than the previous portfolio indicating that it's of lower risk but the profits are also lower in keeping with the phrase "High Risk High Return"
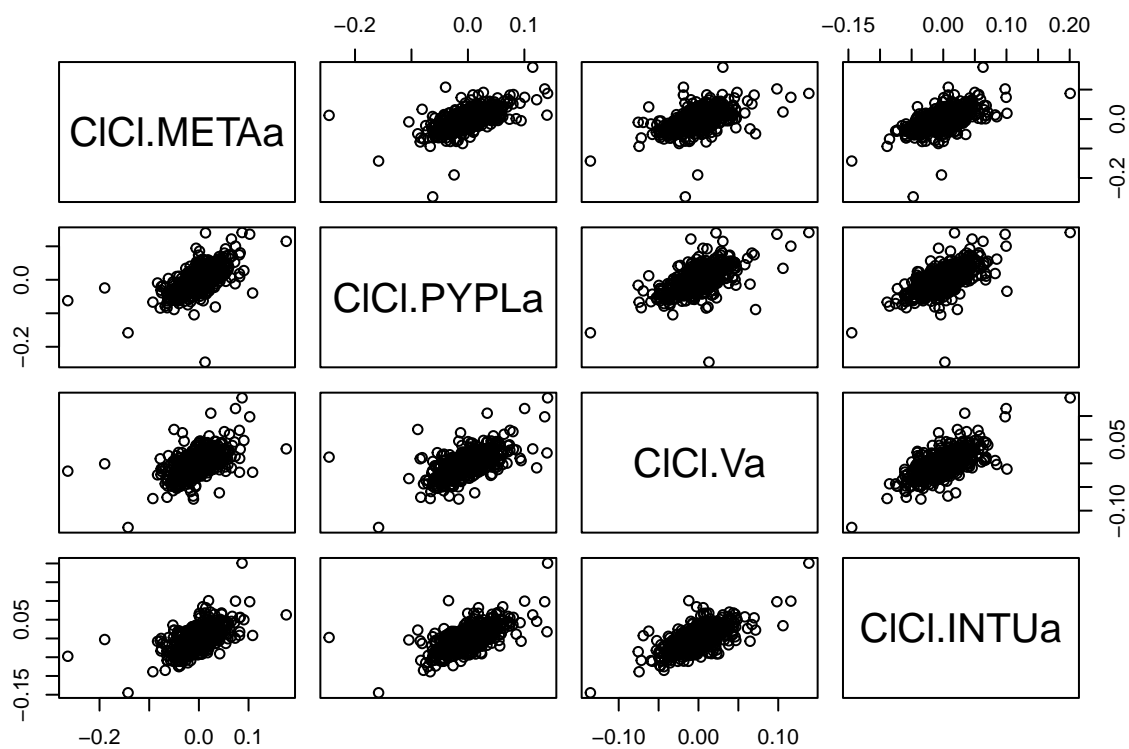
The 5% value at risk for this stock is about -$8211 which means there is a possibility that 5% of the times you might lose over 8% of your original investment through this portfolio. However, the median profit/loss is around $507 which means that at least 50% of the times, you would end up with a profit.
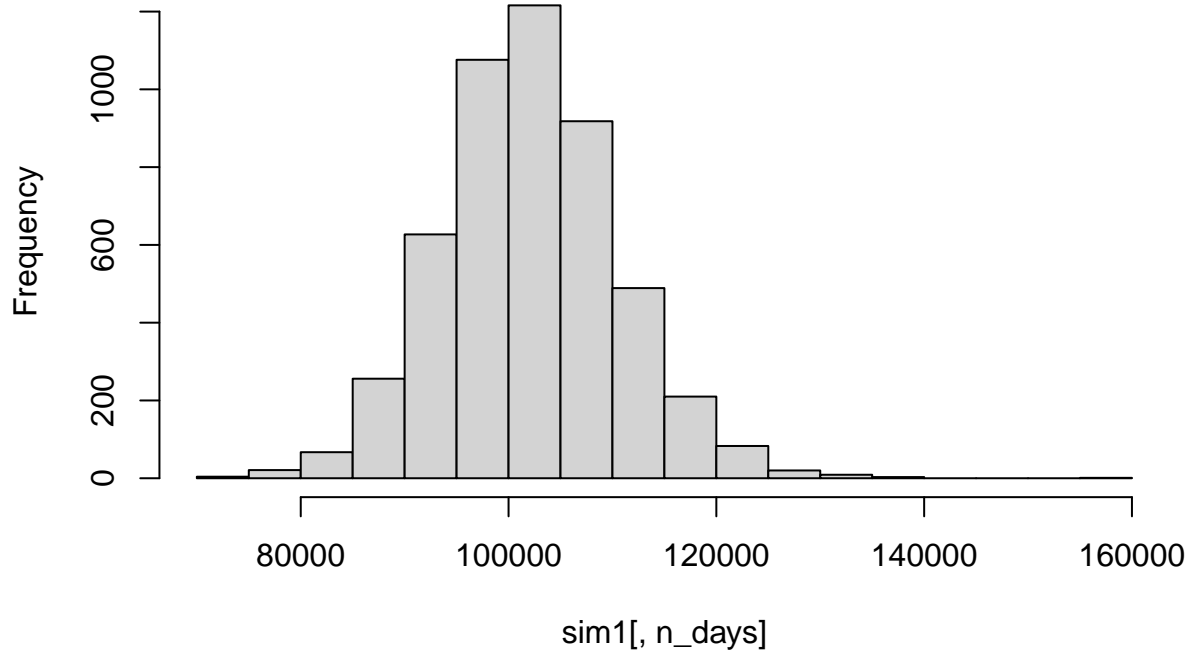
T   [2017−01−03/2022−08−12]

Last 18.27

30

25

20

Volume (millions):
25,800,800

300
200
100
0

Jan 03   Jan 02   Jan 02   Jan 02   Jan 04   Jan 03
2017   2018   2019   2020   2021   2022

TMUS **[2017−01−03/2022−08−12]**

Last 146.460007

140

120

100

80

60

Volume (millions):
3,290,300

60
40
20
0

Jan 03
2017

Jan 02
2018

Jan 02
2019

Jan 02
2020

Jan 04
2021

Jan 03
2022

Histogram of sim1[, n_days]


```
## [1] 100684.7

## [1] 5707.527
```

Histogram of sim1[, n_days] – initial_wealth
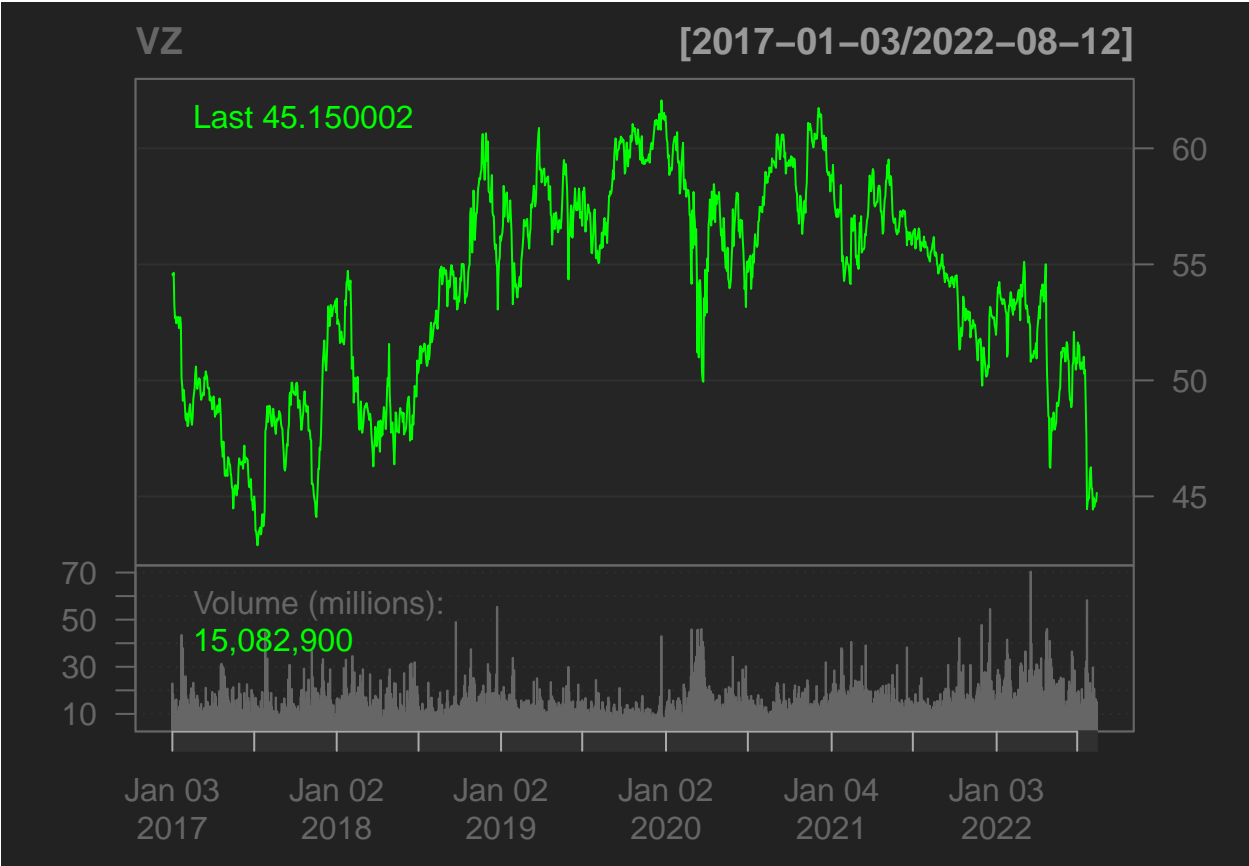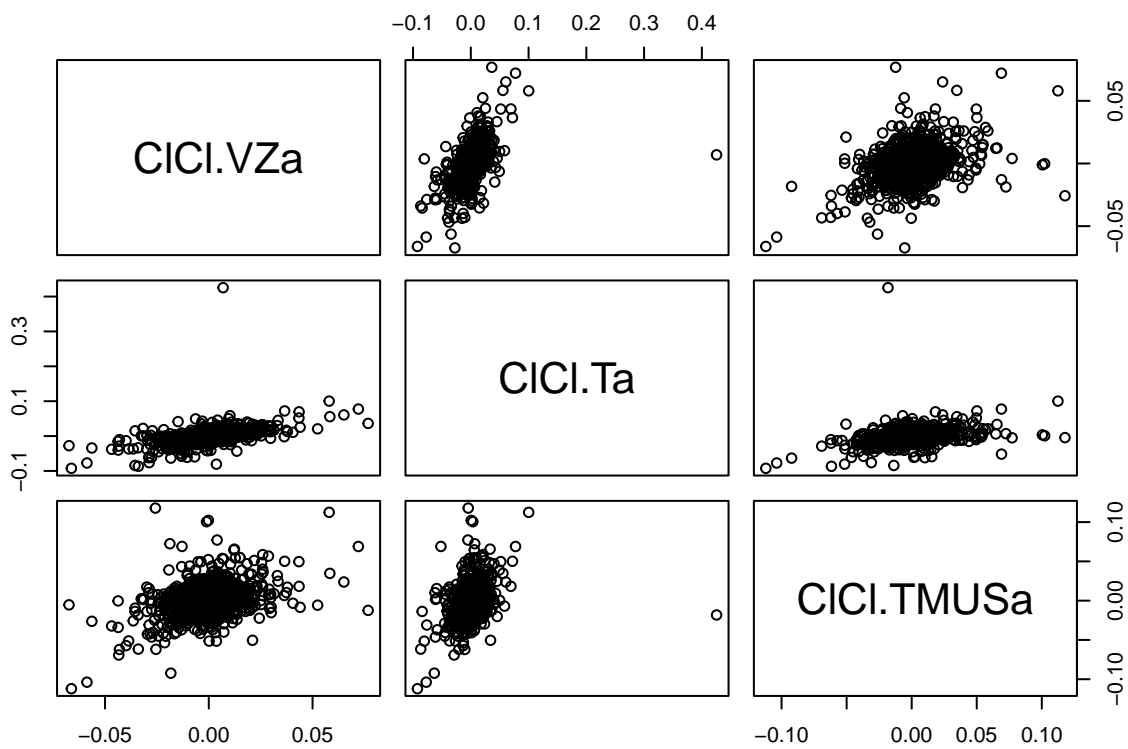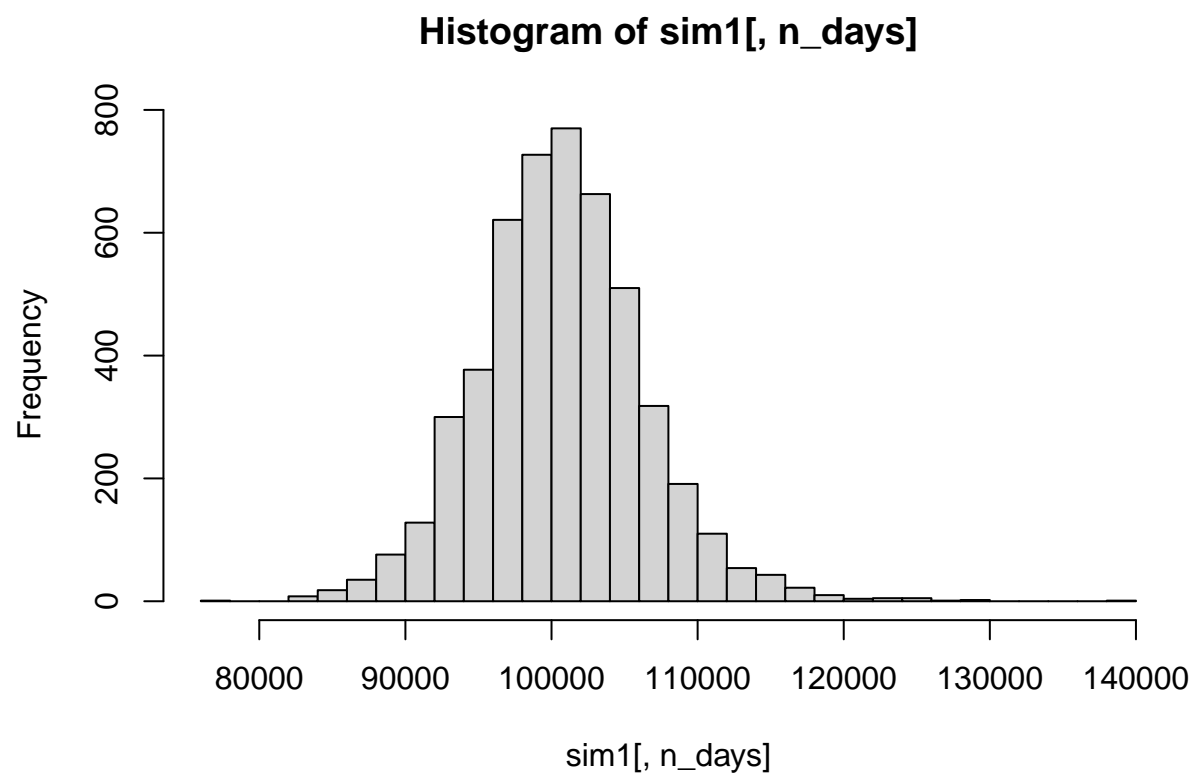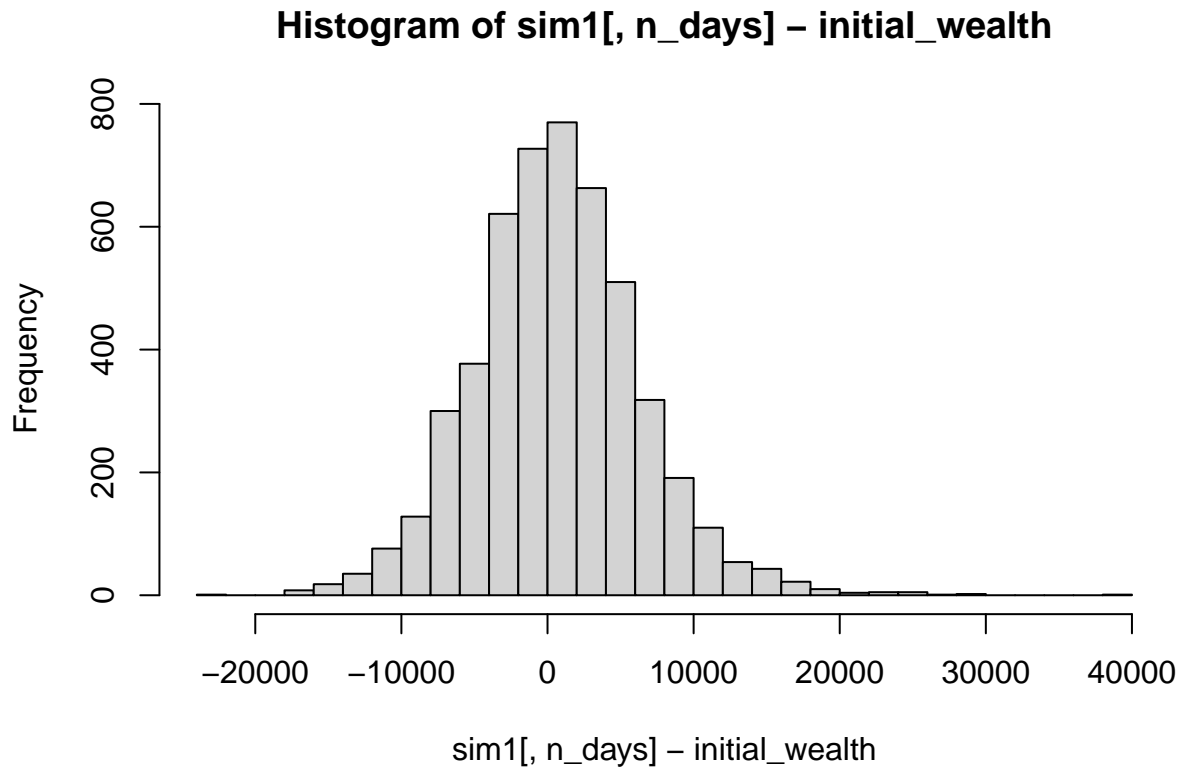

```
## [1] "Average Profit/Loss 684.713779519148"

## [1] "Standard Deviation ofProfit/Loss 5707.52680635004"

## [1] "5% VaR -8155.71493663456"

## [1] "50% VaR 561.189056931871"
```

---

6. Clustering and PCA:

SOLUTION: Approach: Use PCA to reduce dimensionality of the data set and come up with vectors that could used to predict the color of wine/quality.

Let's make a numerical variable to code for wine color (1 - red, 0 - white). Next, let's look at the correlation plot.

Color of wine is the highly correlated with sulphur content and volatile acidity. Quality of wine is the correlated with alcohol content

```
##                        color
## fixed.acidity       0.48673983
```

```
## volatile.acidity        0.65303559
## citric.acid           -0.18739650
## residual.sugar        -0.34882101
## chlorides              0.51267825
## free.sulfur.dioxide   -0.47164366
## total.sulfur.dioxide  -0.70035716
## density                0.39064532
## pH                     0.32912865
## sulphates              0.48721797
## alcohol               -0.03296955
## quality               -0.11932328
## color                  1.00000000


##                            quality
## fixed.acidity          -0.07674321
## volatile.acidity       -0.26569948
## citric.acid             0.08553172
## residual.sugar         -0.03698048
## chlorides              -0.20066550
## free.sulfur.dioxide     0.05546306
## total.sulfur.dioxide   -0.04138545
## density                -0.30585791
## pH                      0.01950570
## sulphates               0.03848545
## alcohol                 0.44431852
## quality                 1.00000000
## color                  -0.11932328
```
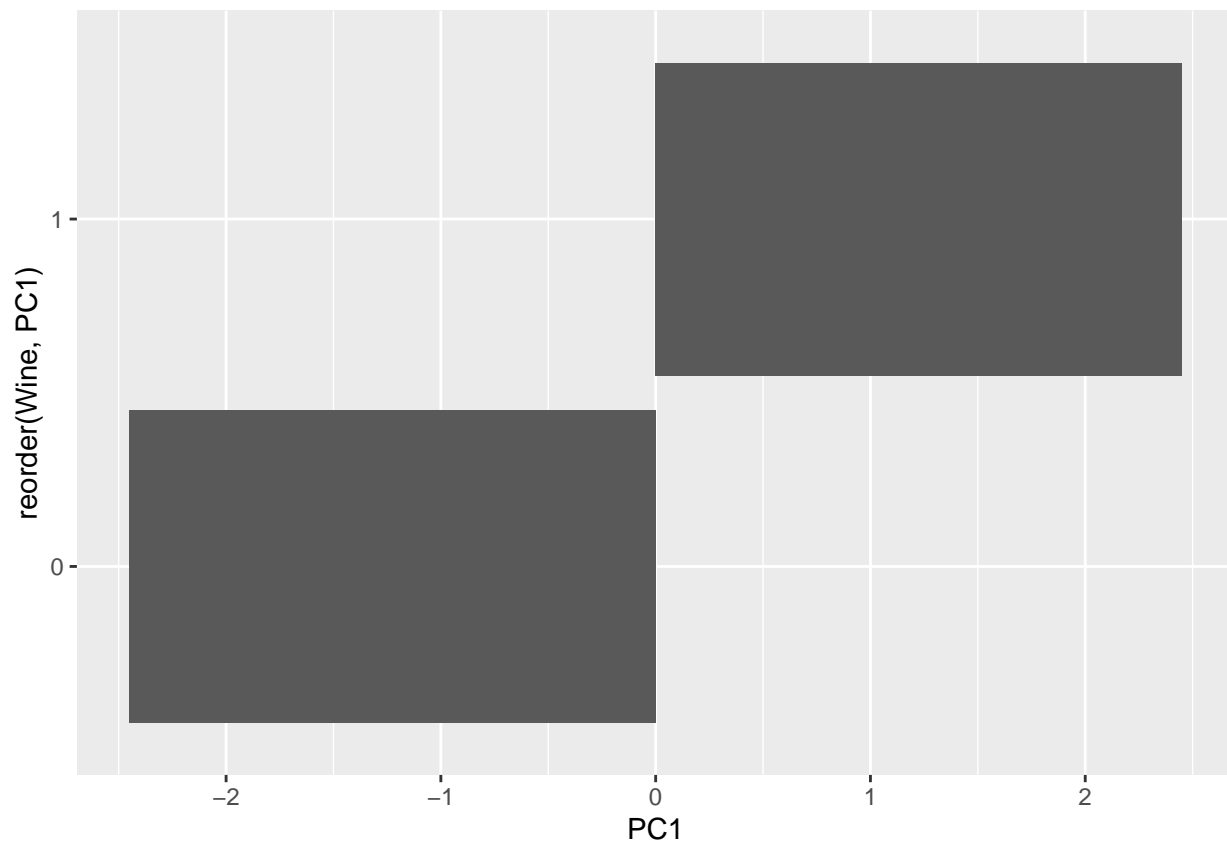
Predicting Color Let's work with PCA to reduce dimensionality of the data. From first glance, volatile.acidity and sulphur have opposing signs in PC1. Fair to say that a positive acidity & negative sulphur corresponds to red wine and vice-versa for white wine.

Also, PC1 seems adequately adept at predicting wine color on its own.

```
##                         PC1   PC2
## fixed.acidity          0.29 -0.10
## volatile.acidity       0.29 -0.10
## citric.acid           -0.29 -0.15
## residual.sugar        -0.29  0.09
## chlorides              0.29 -0.09
## free.sulfur.dioxide   -0.29  0.12
## total.sulfur.dioxide  -0.29  0.07
## density                0.29 -0.09
## pH                     0.29 -0.09
## sulphates              0.29 -0.22
## alcohol               -0.29  0.10
## quality               -0.29 -0.92
```

Predicting Quality Let's work with PCA to reduce dimensionality of the data. From first glance, it's hard to say how meaningful each vector is.

```
##                         PC1   PC2   PC3   PC4   PC5   PC6   PC7
## fixed.acidity         -0.17  0.51 -0.08  0.33 -0.38  0.18  0.06
## volatile.acidity      -0.31  0.31 -0.22 -0.07  0.34 -0.38 -0.02
## citric.acid            0.36  0.03  0.20  0.04 -0.56 -0.19  0.17
## residual.sugar        -0.19 -0.17  0.53 -0.27 -0.07  0.39 -0.28
## chlorides             -0.39  0.11  0.00  0.09  0.14  0.15 -0.11
## free.sulfur.dioxide    0.04  0.31  0.52  0.40  0.39  0.24  0.11
## total.sulfur.dioxide  -0.06  0.29  0.55 -0.31 -0.05 -0.53  0.01
## density               -0.39  0.03  0.00 -0.10 -0.24  0.24  0.70
## pH                     0.23  0.47 -0.12  0.11 -0.18  0.12 -0.43
## sulphates             -0.19 -0.44  0.16  0.66 -0.13 -0.22 -0.16
## alcohol                0.39  0.02  0.08  0.25  0.31 -0.13  0.41
## color                 -0.40  0.00  0.01  0.17 -0.17 -0.38 -0.07
```

Let's look at variance to understand better. Seems like PC1, PC2, PC3 are able to collectively explain 94% of the variance.

PCAwine_quality
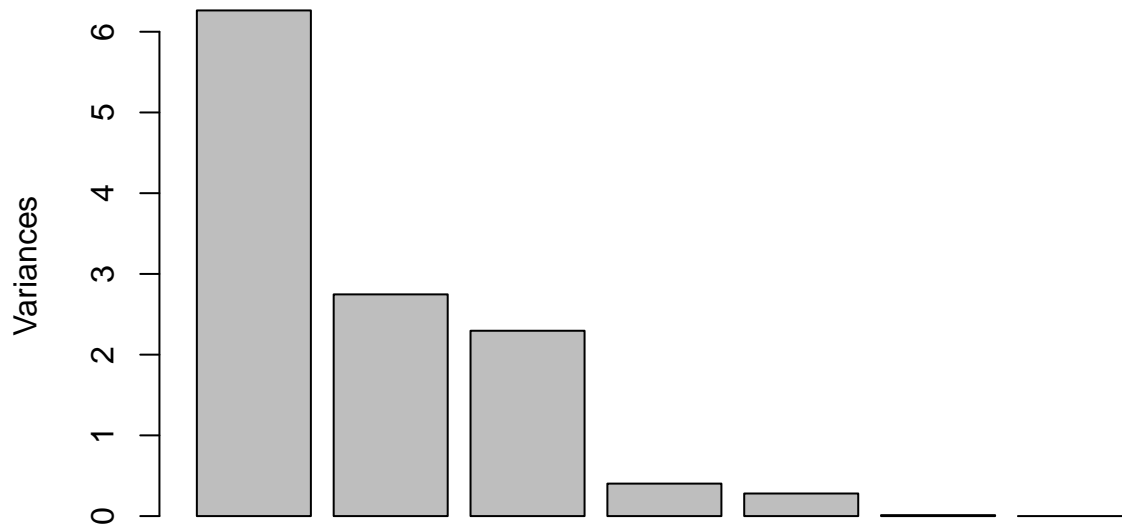

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6      PC7
## Standard deviation      2.503 1.6572 1.5152 0.63440 0.52900 0.10769 1.083e-14
## Proportion of Variance  0.522 0.2289 0.1913 0.03354 0.02332 0.00097 0.000e+00
## Cumulative Proportion   0.522 0.7509 0.9422 0.97571 0.99903 1.00000 1.000e+00
```

Let's run a regression model to see how influential these elements are. These are pretty good numbers

```
##
## Call:
## lm(formula = Quality ~ PC1 + PC2 + PC3, data = wines_quality)
##
## Residuals:
##         1         2         3         4         5         6         7
## -0.238037 -0.001246  0.460515 -0.125887 -0.114880 -0.147833  0.167368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.00000    0.12866  46.633 2.17e-05 ***
## PC1          0.81280    0.05553  14.638 0.000692 ***
## PC2         -0.26681    0.08386  -3.182 0.050035 .
## PC3          0.34612    0.09172   3.774 0.032581 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3404 on 3 degrees of freedom
## Multiple R-squared:  0.9876, Adjusted R-squared:  0.9752
## F-statistic: 79.54 on 3 and 3 DF,  p-value: 0.00234
```

Let's build a tree to cluster these PCA dimensions now

Cluster Dendrogram

dist_mat
hclust (*, "average")
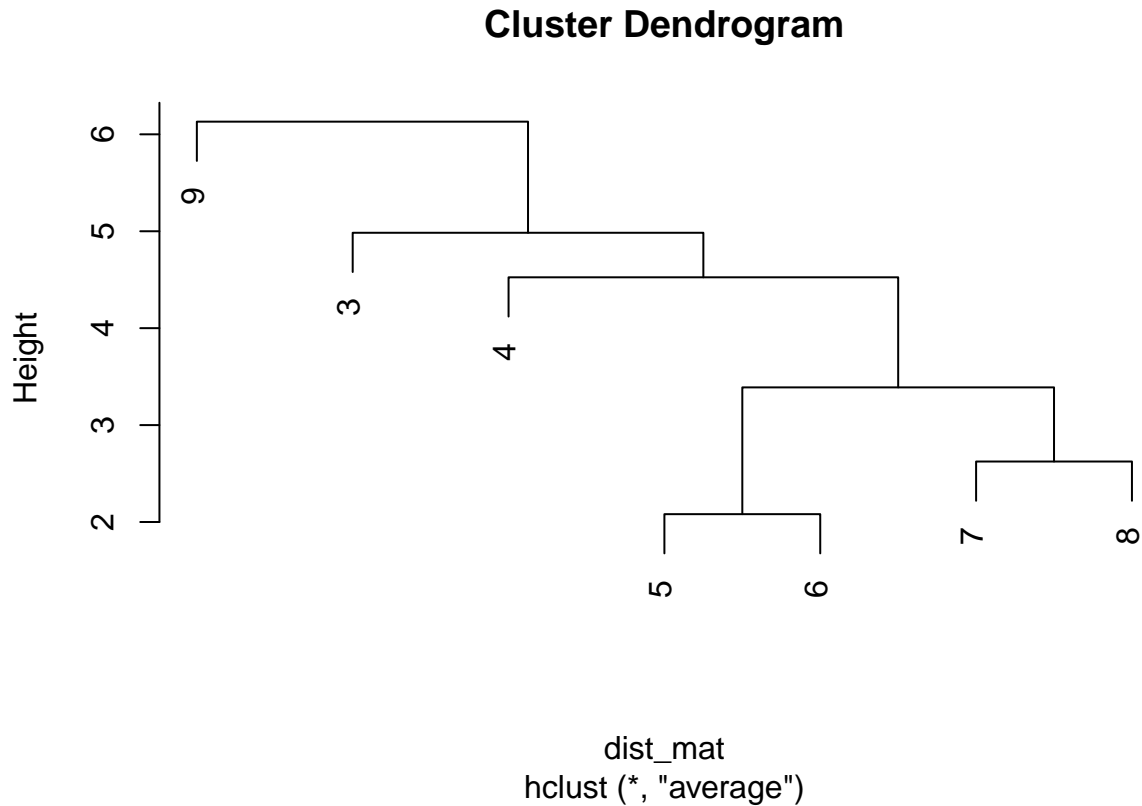
7. Market segmentation:

SOLUTION: To understand the tweets better, let's define a ratio which gives us the average tweets per user for each category listed in the dataset. Based on the results below, it would seem like 'chatter' is the most common - given that it's a proxy for 'unknown', let's ignore that one.

The next biggest categories are photo sharing/ health nutrition and cooking

This could mean that:

1. Some of these followers are food bloggers

2. Some of them chase a healthy lifestyle

3. Some of them are professional chefs or like posting cooking content

4. They could also be a sum of all 3

It's also interesting that some of these members are interested in politics & current events -

5. maybe they tend to support socially conscious brands

Also, there's a lot of sports fandom & college uni tweets too -

6. Might be an indicator that university sports fans are also your followers

```
##                    Tweet_Ratio
## chatter             4.3987567
## photo_sharing       2.6967775
## health_nutrition    2.5672418
## cooking             1.9982238
## politics            1.7886323
## sports_fandom       1.5940117
## travel              1.5850038
## college_uni         1.5494798
## current_events      1.5262624
## personal_fitness    1.4620655
## food                1.3974879
## shopping            1.3893682
## online_gaming       1.2088302
## news                1.2055316
## religion            1.0954073
## tv_film             1.0702867
## fashion             0.9965745
## parenting           0.9213398
## family              0.8638670
## automotive          0.8298655
```

However, these are just directional hypotheses. Let's look at the correlation matrix to confirm some of our theories:

1. Health conscious users who are interested in a highly physical, outdoorsy lifestlyle

2. People who blog about fashion & beauty also blog about cooking. This is interesting considering that these are some of the topics for which a lot of short, viral tik toks are available online. Consumers of one kind of topic are also the consumers of other 2 & hence tweet about it

3. Some of these members are interested in politics & current events - maybe they tend to support socially conscious brands

4. A new pattern that has emerged is a subset that tweets about parenting/religion/school/family and also food. These are typical topics associated with people in their 30s or 40s. Maybe they're looking for nutritious brands that might work well for their families

In summary, these four demographics derived based on the correlation matrix and intuition are good starting points to test out some of the marketing campaigns. But it is key to remember that correlation is not always causation. Hence, running A/B tests and then ramping up the experiments based on results would be a good idea.
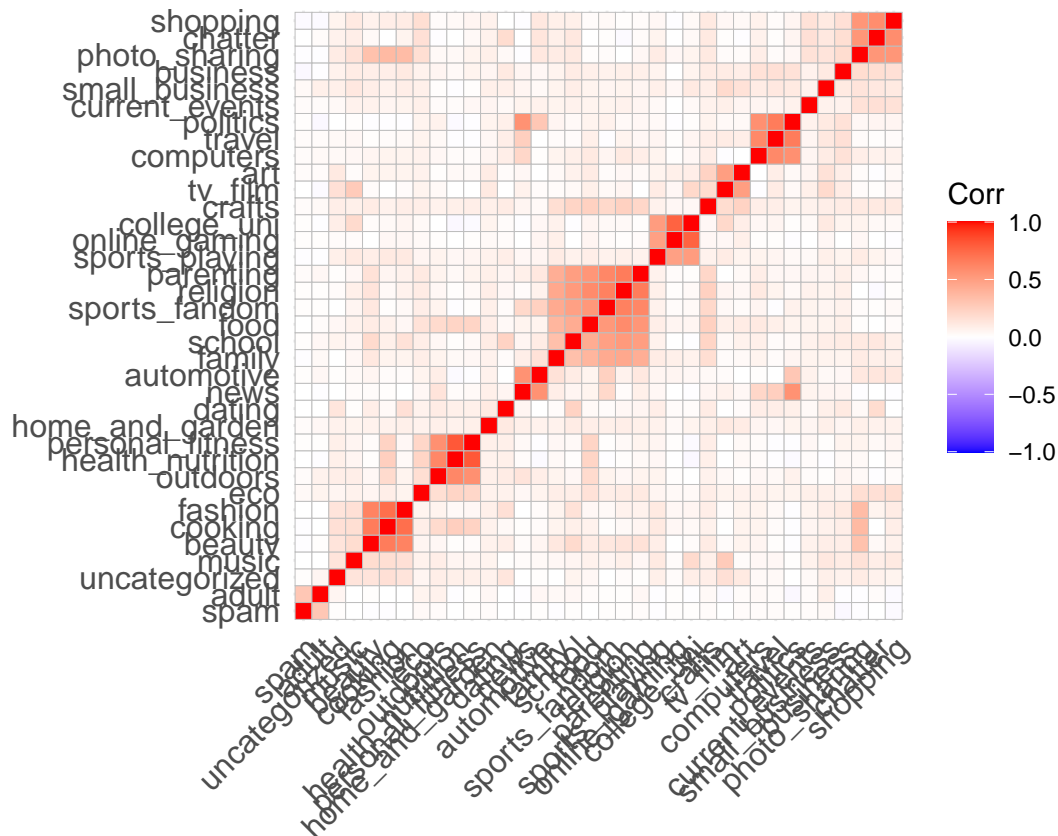
8. The Reuters corpus

SOLUTION:

QUESTION: One of Jane Macartney's articles begins by describing the beauty of Tibet and another one talks about political tensions in China. There are 50 different articles in her repository and while most of them feel seemingly unrelated at first glance,it's hard to gauge what she writes about due to the vivid swing in themes and topics.

The thriving questions are:

- What are her most frequently covered topics?
- Given the diversity of the topics she covers, is there a way to broadly categorize her articles based on the themes in each of them?

APPROACH: 1. A good way to understand the content of her articles would be through text mining. The process has the ability to broad stroke embedded sentiments

2. Given the quantity and verbosity of these articles, dimensionality reduction would also be necessary to condense all of that data into explainable, meaningful by concise vectors.

3. Once condensed, it helps to cluster based on these key vectors to get a sense of related articles

RESULTS:

Text Mining: Digging through 50 of her articles, the most used words seems to be 'China', 'Beijing', 'Chinese', 'government', 'state' and so on. While directionally it might seem like she's talking about the political situation in China, to understand the situation deeper, let's look at some of the associations we see.

```
## <<DocumentTermMatrix (documents: 50, terms: 4251)>>
## Non-/sparse entries: 12872/199678
## Sparsity           : 94%
## Maximal term length: 28
## Weighting          : term frequency (tf)


##  [1] "beijing"                    "character"
##  [3] "china"                      "chinas"
##  [5] "chinese"                    "communist"
##  [7] "ctrainjanemacartneynewsmltxt" "dalai"
##  [9] "datetimestamp"              "description"
## [11] "foreign"                    "heading"
## [13] "hour"                       "isdst"
## [15] "just"                       "lama"
## [17] "language"                   "last"
## [19] "law"                        "lhasa"
## [21] "listauthor"                 "listcontent"
## [23] "listsec"                    "many"
## [25] "may"                        "mday"
## [27] "meta"                       "million"
## [29] "min"                        "mon"
## [31] "next"                       "officials"
## [33] "one"                        "origin"
## [35] "party"                      "people"
## [37] "percent"                    "political"
## [39] "region"                     "said"
## [41] "say"                        "saying"
## [43] "several"                    "since"
## [45] "tibet"                      "trade"
## [47] "two"                        "wday"
## [49] "world"                      "yday"
## [51] "year"                       "years"
## [53] "also"                       "democracy"
## [55] "dissidents"                 "expected"
## [57] "government"                 "human"
## [59] "labour"                     "leader"
## [61] "liu"                        "police"
## [63] "rights"                     "state"
## [65] "wang"                       "week"
## [67] "market"                     "new"
## [69] "official"                   "economic"
## [71] "will"                       "states"
## [73] "united"                     "wto"
```

China: Looking at China's closest associations, words such as 'membership', 'demands', 'disadvantages', 'neighbours', 'liberalisations' etc. seem to emerge. Maybe she is talking about the political situation in China and it's relationship with its neighbours.

```
## $china
##      membership        country        sciences             pei            wait
##            0.80           0.76            0.75            0.71            0.69
##         actively      admittance      advantages    asianpacific       attending
##            0.69           0.69            0.69            0.69            0.69
##            block      catastrophe         demands    disadvantages        explain
##            0.69           0.69            0.69            0.69            0.69
##             head   liberalisations     neighbours       nontariff       presented
##            0.69           0.69            0.69            0.69            0.69
##        principle         proceed        proceeded         pulling           shelf
##            0.69           0.69            0.69            0.69            0.69
##         shopping           slash        solution        stumbling         weekend
##            0.69           0.69            0.69            0.69            0.69
##             wont            wtos          member      concessions        concrete
##            0.69           0.69            0.68            0.67            0.66
##           matter           stand          status           enter            gain
##            0.65           0.65            0.65            0.64            0.63
##        changhong         concede        exclusion         founding          hunker
##            0.61           0.61            0.61            0.61            0.61
##             pace        patience          seemed          snails        unlikely
##            0.61           0.61            0.61            0.61            0.59
##            eager          qichen         academy     application           sides
##            0.58           0.58            0.58            0.57            0.57
##         outlining          parts           seize          access        barriers
##            0.57           0.56            0.56            0.56            0.56
##            forum      philippines        populous         protect     recognition
##            0.56           0.56            0.56            0.56            0.56
##           insists         existed         economy       developed            cost
##            0.55           0.55            0.54            0.54            0.53
##          requires             see           every         service      remarkably
##            0.53           0.53            0.53            0.53            0.53
##           citing        observer         opening           world          nation
##            0.53           0.53            0.53            0.52            0.52
##           import       optimistic          states        distance         shelves
##            0.52           0.51            0.51            0.51            0.51
##          prepared    disappointed        meetings           ranks       sidelines
##            0.51           0.50            0.50            0.50            0.50
```

Tibet: When you look at Tibet's closest associations, it's with unrest, autonomy, uprising, antichinese etc. Feels like she's talking about Tibetian relations with China

```
## $tibet
##           region            lama         restive         tibetan           india
##            0.87           0.76            0.75            0.75            0.74
##          regions          tibets         network          unrest       himalayan
##            0.73           0.72            0.72            0.71            0.69
##         autonomy           dalai           lhasa      londonbased          exiled
##            0.68           0.68            0.68            0.67            0.66
##         abortive          robbie        uprising        buddhist      antichinese
##            0.66           0.66            0.66            0.65            0.65
##      traditional         godking        homeland        tibetans     inaccessible
##            0.61           0.60            0.60            0.60            0.59
##          peaceful         barnett            fled        reported       nonviolent
##            0.59           0.58            0.58            0.58            0.58
```

```
##      monasteries            temples           borders           rocked          campaign
##             0.57               0.57              0.57             0.57              0.56
##            devout            largest             blast           offers              easy
##             0.56               0.56              0.56             0.55              0.54
##         community             deeply             monks        situation            stable
##             0.54               0.53              0.53             0.53              0.53
##          hampered             hoping    infrastructure             dawn              know
##             0.52               0.52              0.52             0.52              0.52
##           peopled    selfgovernment              rest            radio            radius
##             0.52               0.52              0.51             0.50              0.50
```

Dimensionality Reduction/PCA: Initially the data was 94% sparse & removal of sparse terms, there's still a 83% sparsity in the data. We clearly don't need all these dimensions. PCA might be able to help with this further. Using even about 20 compents is able to capture 65% of the variance in the data. Even just looking at 2 dimensions, you can see patterns emerge

```
## <<DocumentTermMatrix (documents: 50, terms: 1302)>>
## Non-/sparse entries: 9133/55967
## Sparsity           : 86%
## Maximal term length: 28
## Weighting          : term frequency (tf)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.04972 0.06959 0.08364 0.10092 0.57966
```

pca_jane

32
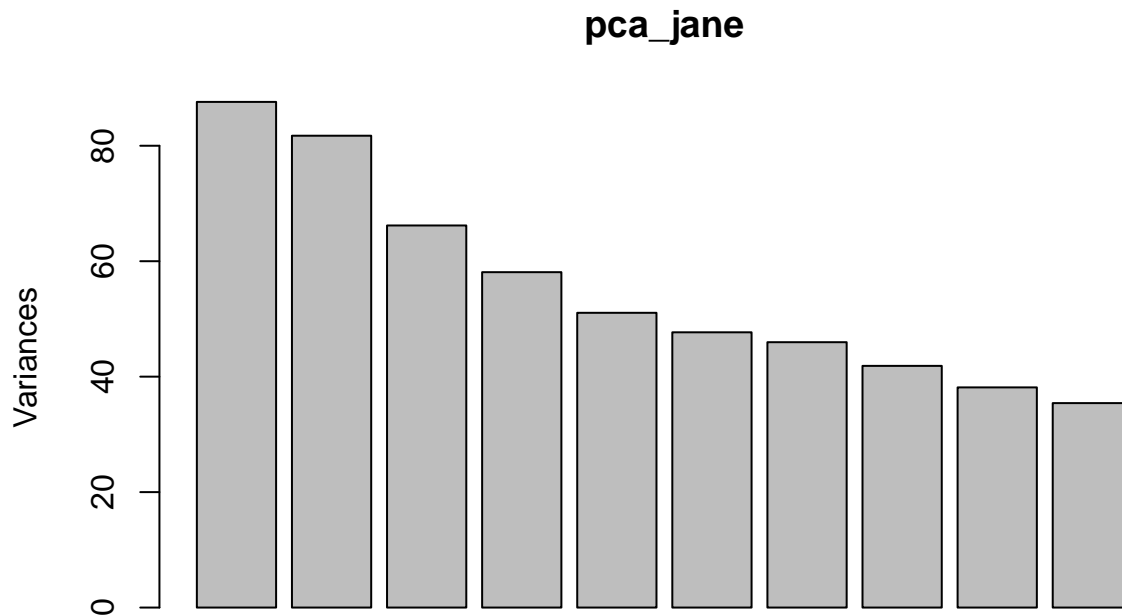
```
## Importance of first k=20 (out of 50) components:
##                              PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation       9.35918 9.04071 8.13545 7.62206 7.14558 6.90513 6.7799
## Proportion of Variance 0.06822 0.06366 0.05155 0.04525 0.03977 0.03713 0.0358
## Cumulative Proportion  0.06822 0.13188 0.18342 0.22867 0.26843 0.30557 0.3414
##                              PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation       6.4703  6.1754 5.95122 5.76050 5.64704 5.50245 5.45072
## Proportion of Variance  0.0326  0.0297 0.02758 0.02584 0.02484 0.02358 0.02314
## Cumulative Proportion   0.3740  0.4037 0.43126 0.45710 0.48194 0.50552 0.52866
##                             PC15    PC16    PC17    PC18    PC19    PC20
## Standard deviation       5.40034 5.29540 5.25485 5.18872 5.15952 5.05334
## Proportion of Variance 0.02271 0.02184 0.02151 0.02097 0.02073 0.01989
## Cumulative Proportion  0.55137 0.57321 0.59471 0.61568 0.63641 0.65630


##
## Docs          PC1          PC2
##   1    -1.23864710    6.08578850
##   2    21.21895942  -13.73300879
##   3     1.91414062    7.23160775
##   4    22.59900893  -13.93525939
##   5    -2.08194770    3.85756430
##   6    22.13423853  -13.99258797
##   7    -1.95568189    3.67516936
##   8    10.05345353   -6.39229616
##   9    -1.75421036   -0.19257810
##  10    -1.84988850    3.70199762
##  11     0.65804194   -1.71722196
##  12    12.55853743   -8.27080993
##  13    -1.19952388   -1.33871306
##  14    11.99981267   -7.28443507
##  15     0.19690784    0.05061722
##  16    -2.01789181   -3.39915414
##  17    -2.75651600   -4.21553275
##  18     0.09168762   -0.04490987
##  19     8.56369154   -7.01566872
##  20     4.65878068   -4.56052201
##  21    -4.06253629   -0.77035762
##  22     7.04111070   -5.60071365
##  23    -4.91839478   -1.07992701
##  24    -4.91839478   -1.07992701
##  25   -14.80913836   -5.32138883
##  26   -14.67698487   -5.26480442
##  27   -15.66684295   -6.83866672
##  28   -21.61654240   -6.52944805
##  29   -20.87254340   -6.27850937
##  30   -15.47505499   -6.40791687
##  31   -13.57822695   -5.55578442
##  32    -2.67616100   -1.52766236
##  33     3.26451523    6.48714502
##  34    -2.75815400    0.15574523
##  35     0.28917622   -2.11206984
##  36     3.76816467   -2.92702448
##  37     2.73358881    5.94463919
##  38     0.52308561   -1.30357800
```
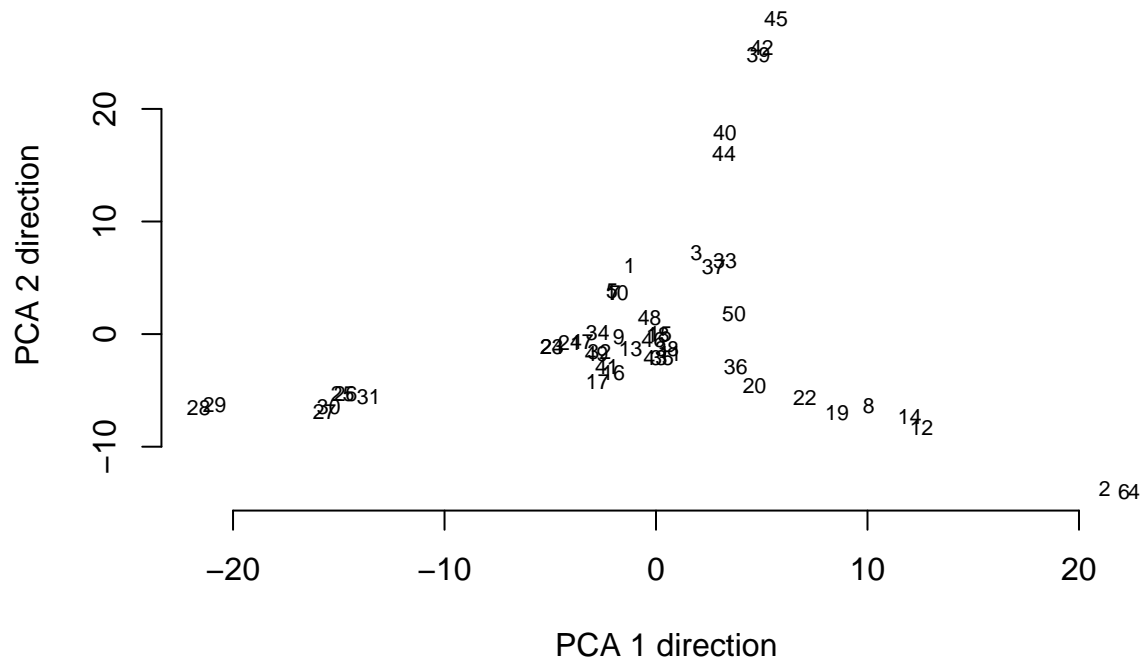
```
## 39    4.84442371   24.76974554
## 40    3.26051410   17.91443886
## 41   -2.32309633   -2.91432530
## 42    5.00587492   25.41499130
## 43   -0.02932604   -2.13117341
## 44    3.22120031   16.00945836
## 45    5.67687785   28.00555619
## 46   -0.12249002   -0.45941072
## 47   -3.50153588   -0.65707919
## 48   -0.30099293    1.48277430
## 49   -2.81113356   -1.70812963
## 50    3.69606390    1.77335608
```


Clustering:

Once we have condensed data, we could cluster the articles - maybe build a dendrogram to show which of there are related

Cluster Dendrogram

dist_mat
hclust (*, "average")
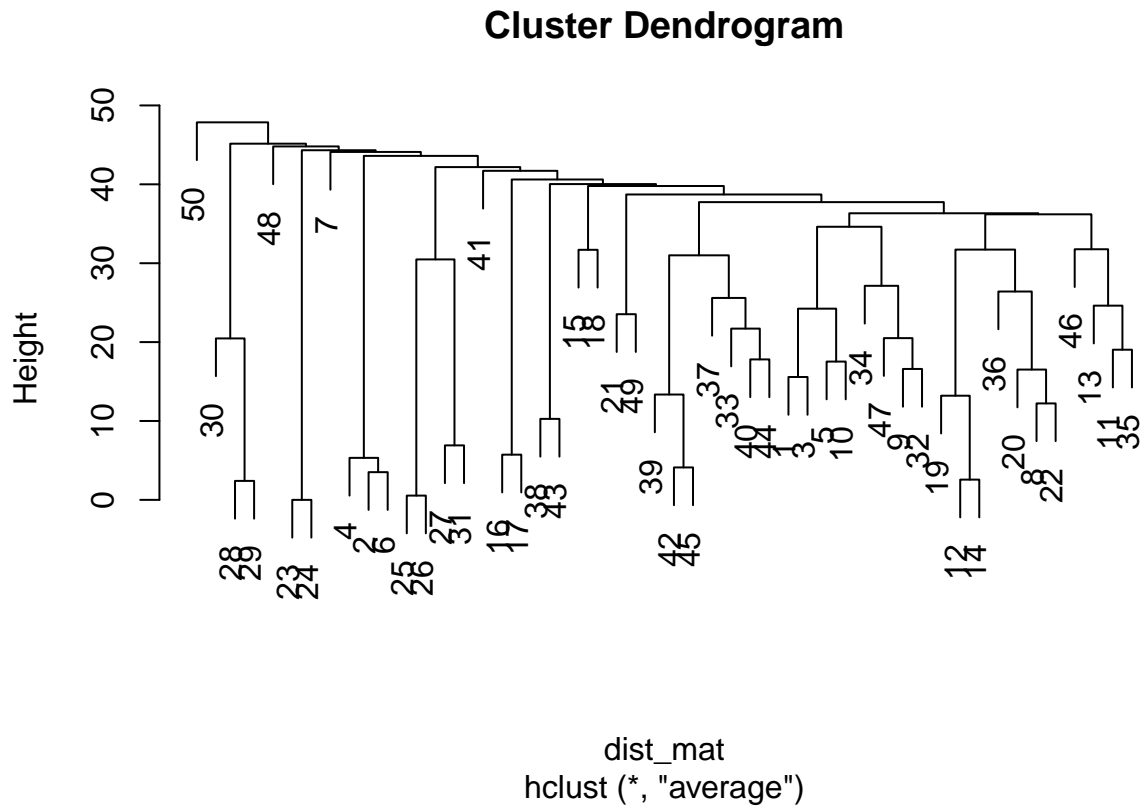
Conclusions: 1. To answer question 1, she generally covers topics relating to the political unrest in China and Tibet and how they're related to each other. She also talks about the various people involved in these situations (eg: Wang Dan)

2. To answer question 2, building a dendrogram helps us directionally understand which of these articles are related

But how do we consume this information? Jane Macartney discusses Chinese politics. Analysing the text of other authors, you would recognise that maybe a few others also had takes on these topics. Building a collective pool of these authors and their takes might help understand the overall sentiment. Also, combining this information with association rule mining might help build good recommendation systems on book websites eg: Goodreads or even Google Scholar

---

9. Association rule mining

SOLUTION:

PROCESSING THE DATA: The data was present as a single column with each row corresponding to a basket. Processing steps include: ###### 1. Storing text as excel ###### 2. Splitting column by comma in R ###### 3. Indexing the data to individual rows

```
##
## 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

## [1] "Before:"

##                 citrus fruit,semi-finished bread,margarine,ready soups
## 1                                       tropical fruit,yogurt,coffee
## 2                                                        whole milk
## 3                         pip fruit,yogurt,cream cheese ,meat spreads
## 4 other vegetables,whole milk,condensed milk,long life bakery product
## 5                         whole milk,butter,yogurt,rice,abrasive cleaner
## 6                                                        rolls/buns

## [1] "After:"

## # A tibble: 8 x 2
##   index basket
##   <chr> <chr>
## 1 1     "tropical fruit"
## 2 1     "yogurt"
## 3 1     "coffee"
## 4 2     "whole milk"
## 5 3     "pip fruit"
## 6 3     "yogurt"
## 7 3     "cream cheese "
## 8 3     "meat spreads"
```

ASSOCIATION RULE MINING: Once the data was processed, the next step was to build association rules. Given that the max lift was around 4.6, let's look at what it shows.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5   0.005      1
```

```
##  maxlen target  ext
##      4  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 49
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[171 item(s), 9834 transaction(s)] done [0.00s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [1581 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].


## set of 1581 rules
##
## rule length distribution (lhs + rhs):sizes
##   1   2   3   4
##   8 754 771  48
##
##    Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
##   1.000   2.000   3.000  2.543   3.000   4.000
##
## summary of quality measures:
##      support          confidence        coverage            lift
##  Min.   :0.005084   Min.   :0.1000   Min.   :0.008135   Min.   :0.4456
##  1st Qu.:0.005898   1st Qu.:0.1466   1st Qu.:0.022981   1st Qu.:1.4741
##  Median :0.007322   Median :0.2190   Median :0.037116   Median :1.8187
##  Mean   :0.010539   Mean   :0.2559   Mean   :0.053292   Mean   :1.9036
##  3rd Qu.:0.010372   3rd Qu.:0.3315   3rd Qu.:0.058064   3rd Qu.:2.2461
##  Max.   :0.255542   Max.   :0.7000   Max.   :1.000000   Max.   :4.6394
##      count
##  Min.   :  50.0
##  1st Qu.:  58.0
##  Median :  72.0
##  Mean   : 103.6
##  3rd Qu.: 102.0
##  Max.   :2513.0
##
## mining info:
##       data ntransactions support confidence
##   groctrans          9834   0.005        0.1
##                                                                         call
##  apriori(data = groctrans, parameter = list(support = 0.005, confidence = 0.1, maxlen = 4))


##     lhs            rhs             support    confidence coverage   lift
## [1] {ham}        => {white bread} 0.005084401 0.1953125  0.02603213 4.63938
## [2] {white bread} => {ham}        0.005084401 0.1207729  0.04209884 4.63938
##     count
## [1] 50
## [2] 50
```

Seems like buying ham is a strong indication that you're also going to buy white bread, probably for a ham sandwich. Let's look at a few more with smaller lifts.

All of these make total sense: 1. Buying veggies with herbs

2. Buying berries and cream - maybe for some fruit and cream dessert

3. Waffles and chocolate - Cheat day perhaps.

```
##      lhs                      rhs                   support    confidence
## [1]  {herbs}               => {root vegetables}     0.007016473 0.4312500
## [2]  {ham}                 => {white bread}         0.005084401 0.1953125
## [3]  {white bread}         => {ham}                 0.005084401 0.1207729
## [4]  {sliced cheese}       => {sausage}             0.007016473 0.2863071
## [5]  {berries}             => {whipped/sour cream}  0.009050234 0.2721713
## [6]  {whipped/sour cream}  => {berries}             0.009050234 0.1262411
## [7]  {hygiene articles}    => {napkins}             0.006101281 0.1851852
## [8]  {napkins}             => {hygiene articles}    0.006101281 0.1165049
## [9]  {waffles}             => {chocolate}           0.005796217 0.1507937
## [10] {chocolate}           => {waffles}             0.005796217 0.1168033
##      coverage   lift     count
## [1]  0.01627008 3.956075 69
## [2]  0.02603213 4.639380 50
## [3]  0.04209884 4.639380 50
## [4]  0.02450681 3.047125 69
## [5]  0.03325198 3.796499 89
## [6]  0.07169005 3.796499 89
## [7]  0.03294692 3.536138 60
## [8]  0.05236933 3.536138 60
## [9]  0.03843807 3.038739 57
## [10] 0.04962375 3.038739 57
```

The highest confidence is 0.7. Let's see what that corresponds to:
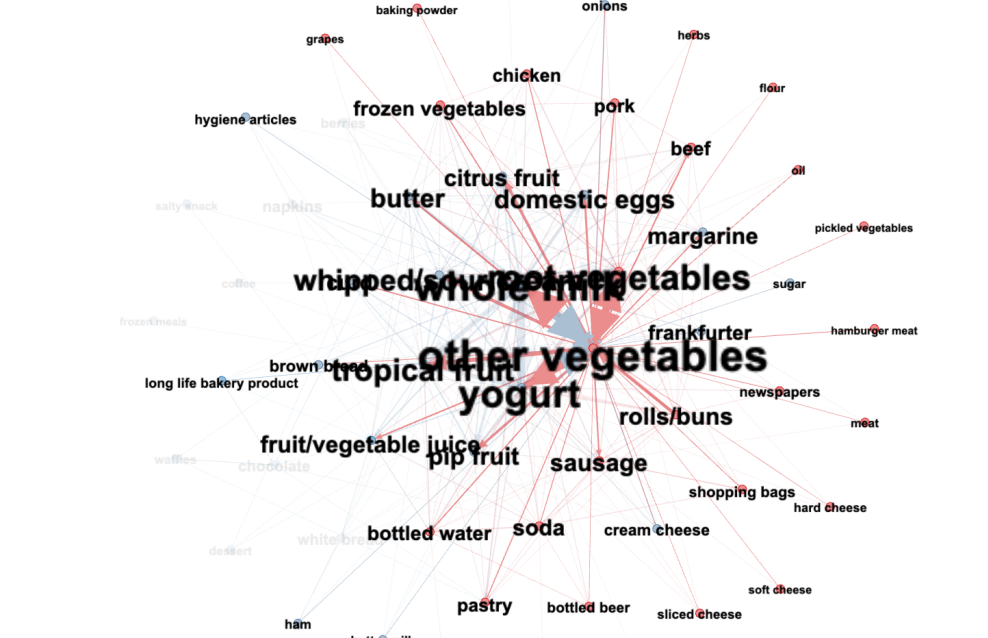
```
##      lhs                  rhs              support confidence   coverage      lift count
## [1] {root vegetables,
##      tropical fruit,
##      yogurt}           => {whole milk} 0.005694529        0.7 0.008135042 2.739276    56
```

If you're buying TROPICAL fruits,ROOT veg & yogurt, you might just buy milk too. There's an above average lift associated with it too. These are staples in a lot of households (eg:Indian households)

Let's make a Gephi image out of this first:

Here are some interesting visuals from Gephi

Whole Graph: Graph of all those with a lift greater than median lift(1.8)

General Grocery Bag: These seem like the staples on every monthly grocery list for a household

UHT Milk & Bottled Water - Generally, people get water off the tap in the US. However, some

people who buy Ultra pasteurized UHT milk also seem to buy purified mineral water. These sound like consumers of clean food.