Introduction to Machine Learning Exercises

Aishwarya Sarkar, Pratik Gawli, Aniket Patil, Rochan Nehete

2022-08-15

Link to .Rmd - https://github.com/Rochan79/STA380_part2_exercises

Question 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.

```
## Fraction of people who answered yes given that they are truthful speakers: 0.7142857
```

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?

```
## Probability of having the disease given they test positive: 0.1988824
```

Question 2: Wrangling the Billboard Top 100

Consider the data in billboard.csv containing every song to appear on the weekly Billboard Top 100 chart since 1958, up through the middle of 2021. Each row of this data corresponds to a single song in a single week.

Part A:

Make a table of the top 10 most popular songs since 1958, as measured by the total number of weeks that a song spent on the Billboard Top 100. Note that these data end in week 22 of 2021, so the most popular songs of 2021 will not have up-to-the-minute data; please send our apologies to The Weeknd.

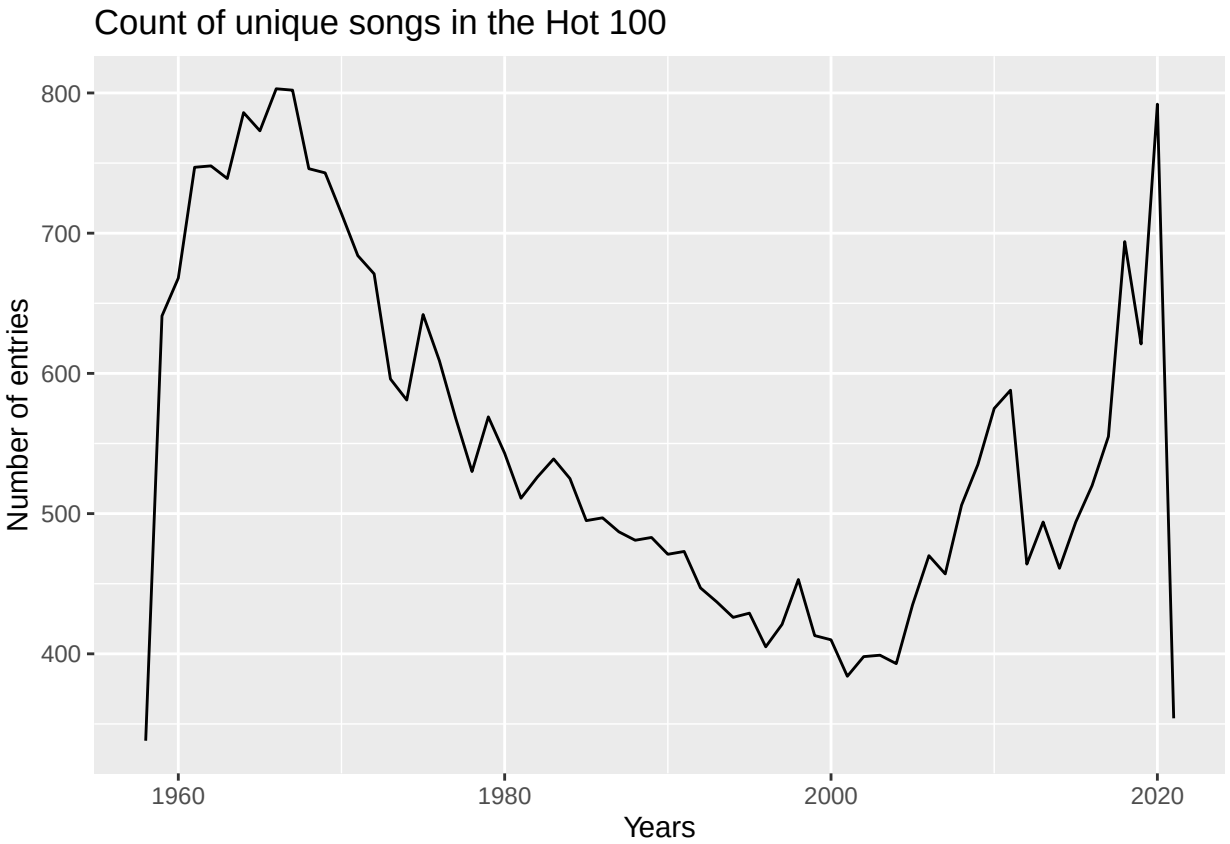
```
## # A tibble: 10 x 3
## # Groups:   performer [10]
##    performer                                song                           count
##    <chr>                                    <chr>                          <int>
##  1 Imagine Dragons                          Radioactive                       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
```

The above table shows the performers and their songs with a count of weeks for which that particular song was in the top 100 from 1958 to 2021.

Part B:

Is the "musical diversity" of the Billboard Top 100 changing over time? Let's find out. We'll measure the musical diversity of given year as the number of unique songs that appeared in the Billboard Top 100 that year. Make a line graph that plots this measure of musical diversity over the years. The x axis should show the year, while the y axis should show the number of unique songs appearing at any position on the Billboard Top 100 chart in any week that year. For this part, please filter the data set so that it excludes the years 1958 and 2021, since we do not have complete data on either of those years. Give the figure an informative caption in which you explain what is shown in the figure and comment on any interesting trends you see.

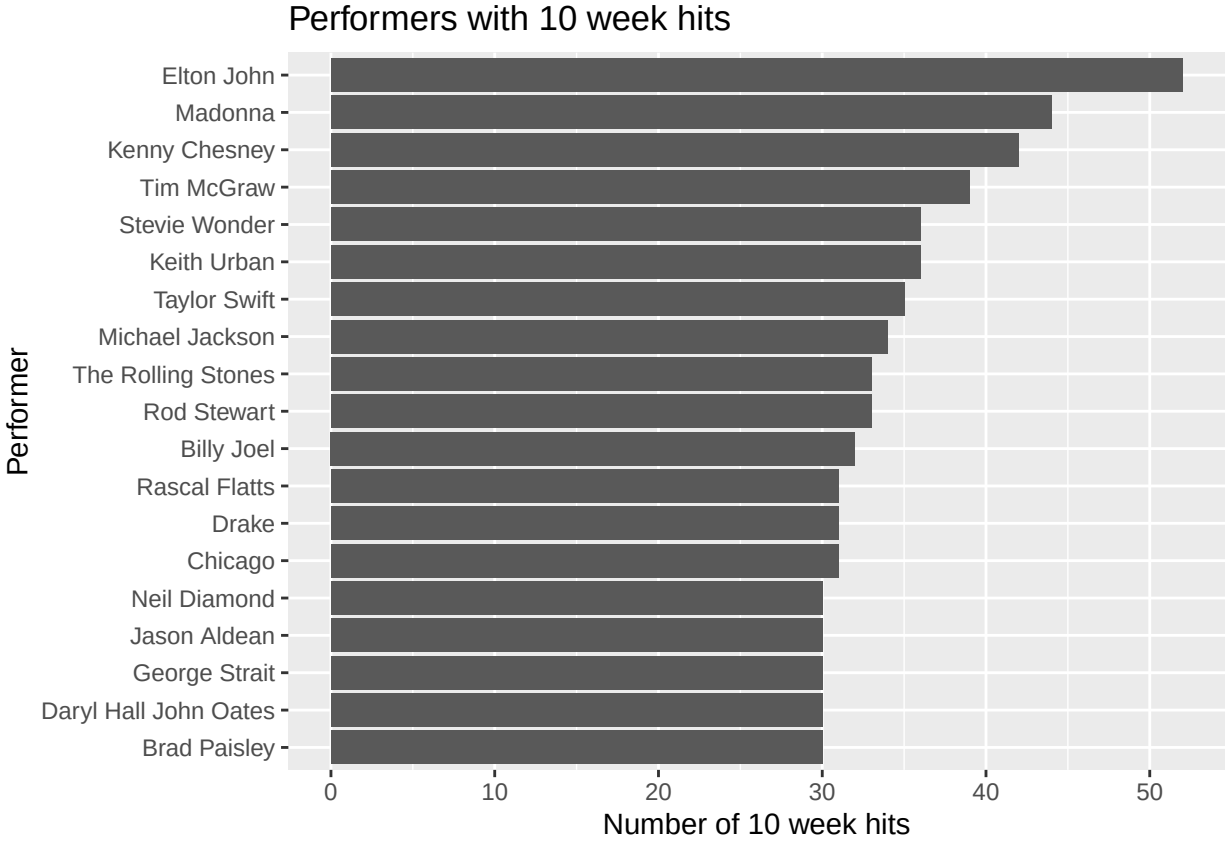


The above plot shows the level of diversity among the Billboard 100 songs from 1958 to 2021 with Number of entries as the metric to measure it. As evident from the plot, we can see diversity reahed its maximum around 1967 and from there it has been on a decline till the advent of the internet. With the arrival of internet which brought the world closer, the diversity in billboard 100 songs has been on the rise, touching the level at which it was in the late 1960s. The decline which we can see from 1960s to 2000 could be attributes to same genres and artists making songs and making it to the top 100 list possibly because of lack

of diversity among music discovered across the world by then. But since 2000 we can see the graph ends on a high and we can expect it to grow even further moving into 2022 and beyond.

Part C:

Let's define a "ten-week hit" as a single song that appeared on the Billboard Top 100 for at least ten weeks. There are 19 artists in U.S. musical history since 1958 who have had at least 30 songs that were "ten-week hits." Make a bar plot for these 19 artists, showing how many ten-week hits each one had in their musical career. Give the plot an informative caption in which you explain what is shown.



The above plot shows the number of 10 week hits given by 19 artists who have atleast 30 songs in the 10 week hits list. Elton John is the only one who has crossed 50 such hits with Madonna at 2nd place with close to 45 hits.

Question 3: Visual story telling part 1: green buildings

3-Visual story telling part 1: green buildings

The case Over the past decade, both investors and the general public have paid increasingly close attention to the benefits of environmentally conscious buildings. There are both ethical and economic forces at work here. In commercial real estate, issues of eco-friendliness are intimately tied up with ordinary decisions about how to allocate capital. In this context, the decision to invest in eco-friendly buildings could pay off in at least four ways.

Firstly, to fact check the Excel guru, I will calculate median rent for green buildings:

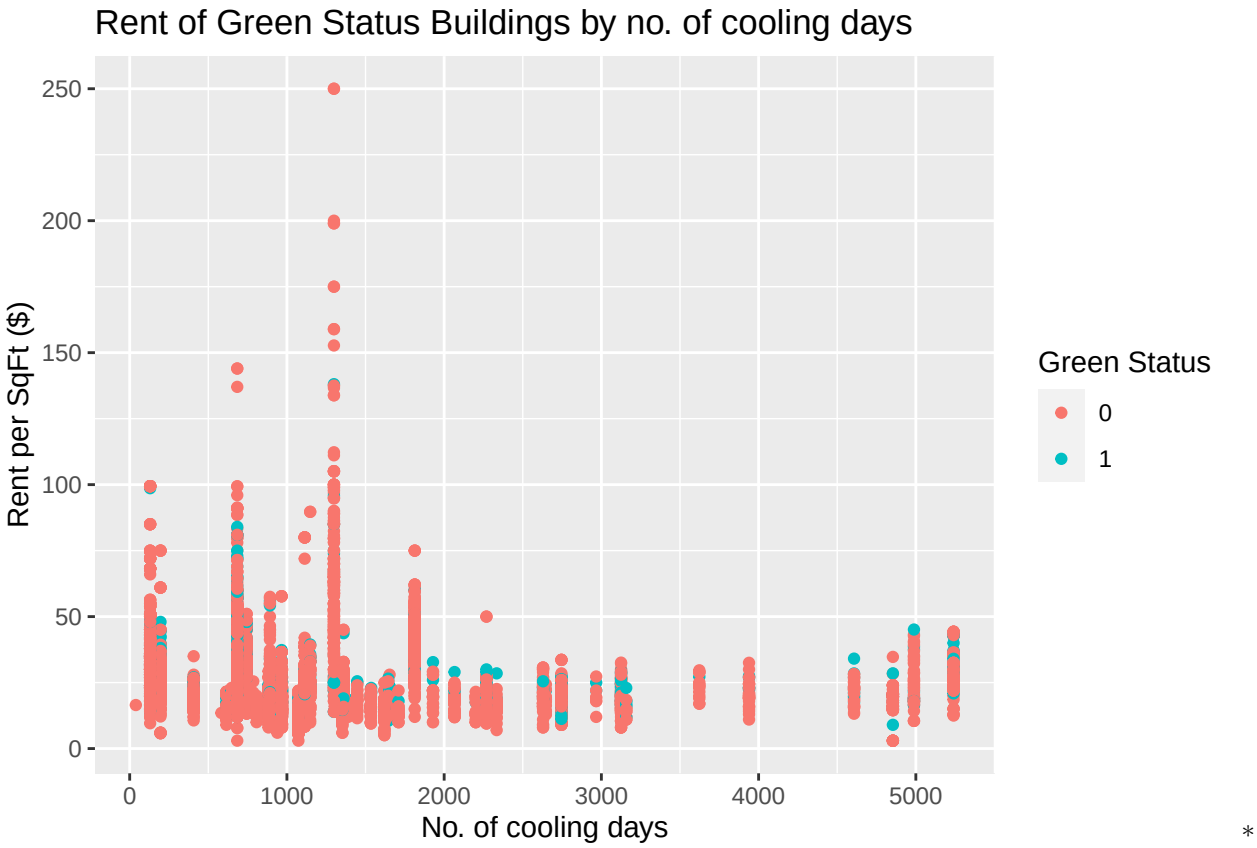
```
## [1] 27.6
```

```
## [1] 25
```

Turns out the guru calculated his median correctly

But in his analysis, I feel that he should've considered other variables before giving a conclusion. I shall firstly find out variables which are confounding with Rent by checking whether they are correlated to Rent or not using scatter plots

#First approach - check for correlation between cooling days, heating days, degree days



*

green buildings have a higher rent across all kinds of cooling days (barring 1000-2000 days),so we can conclude that no of cooling days does not contribute to the rent of the building

- No correlation observed between number of cooling days and rent

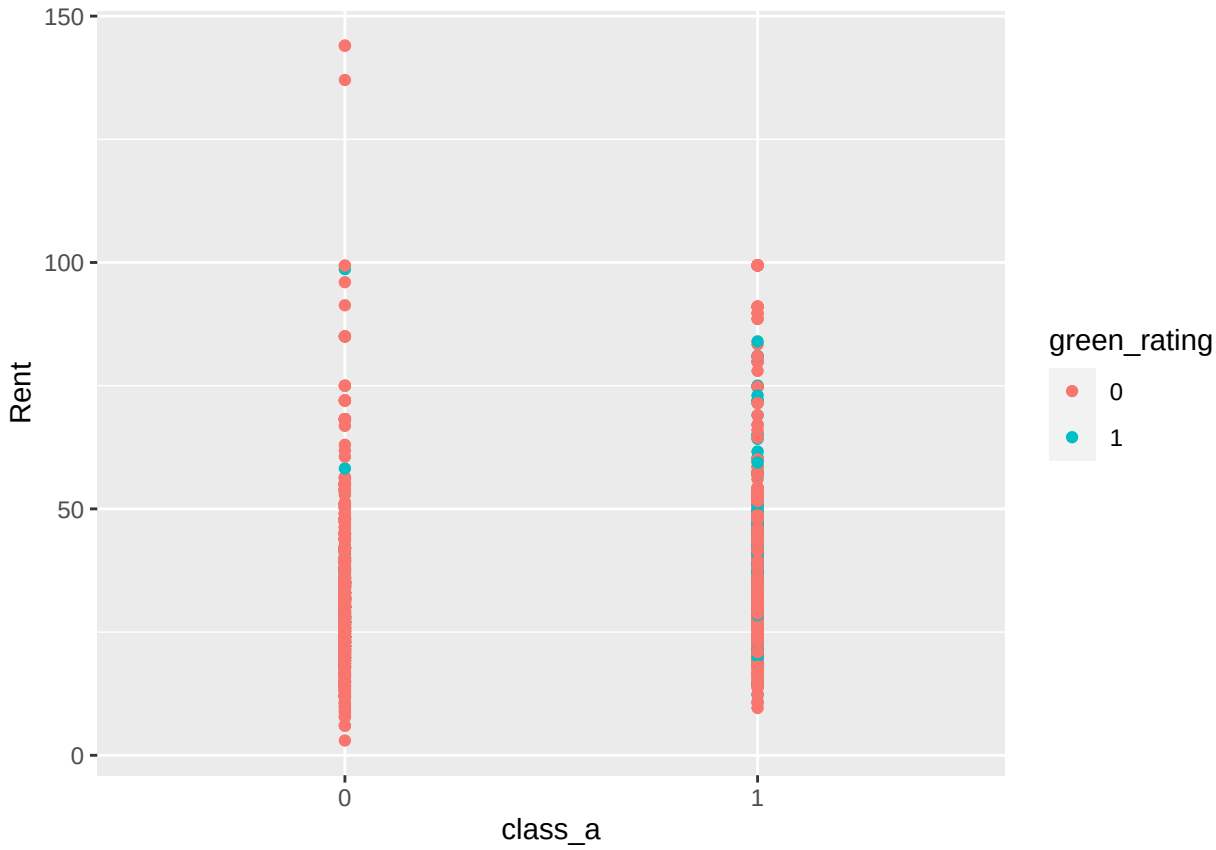
From the graph we see that green_rated buildings charge a higher rent when: * The no of heating degree days are high, implying that there is a need for heating on most days. This implies that the savings in energy costs are higher for a green building, thereby having a higher rent * No Correlation observed between number of heating days and Rent

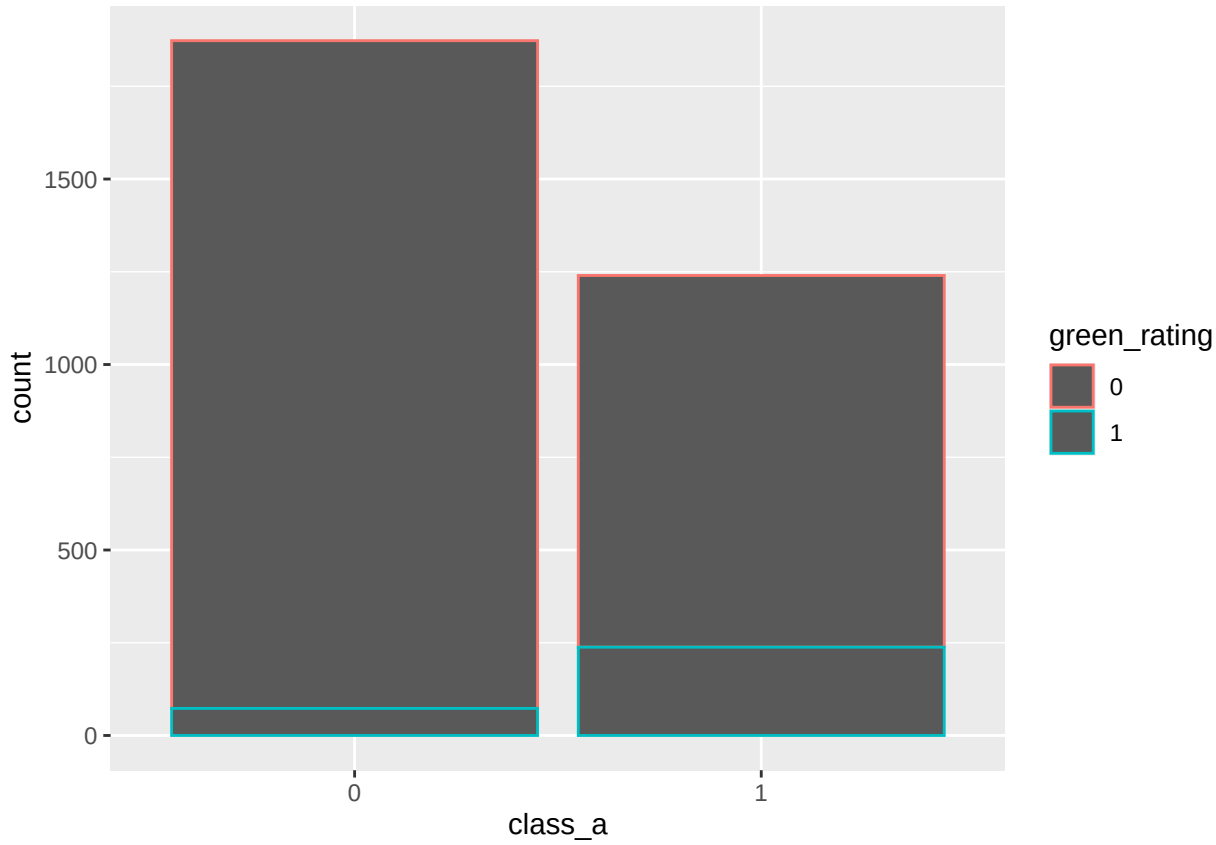
- We can now confidently say that places that have more than 4000q degree days (extreme temperatures), green buildings charge a higher rent. One possible reason for higher rent could be higher savings in energy costs.

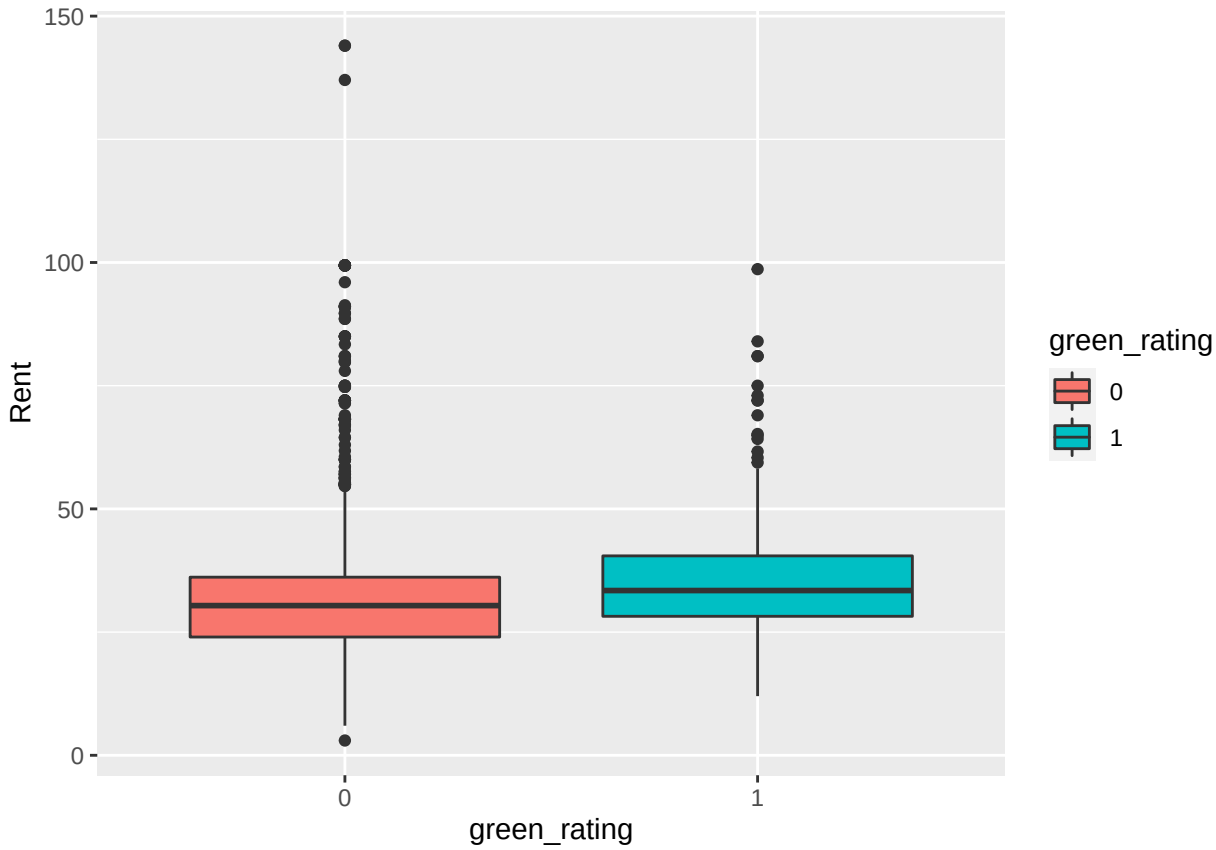
We will now hold the degree days constant and check if it is a confounding variable for Rent for different intervals of degree days.

It is possible that because buildings with degree days > 2000 are better built, and hence charge a premium rent

holding degree days < 4000 constant

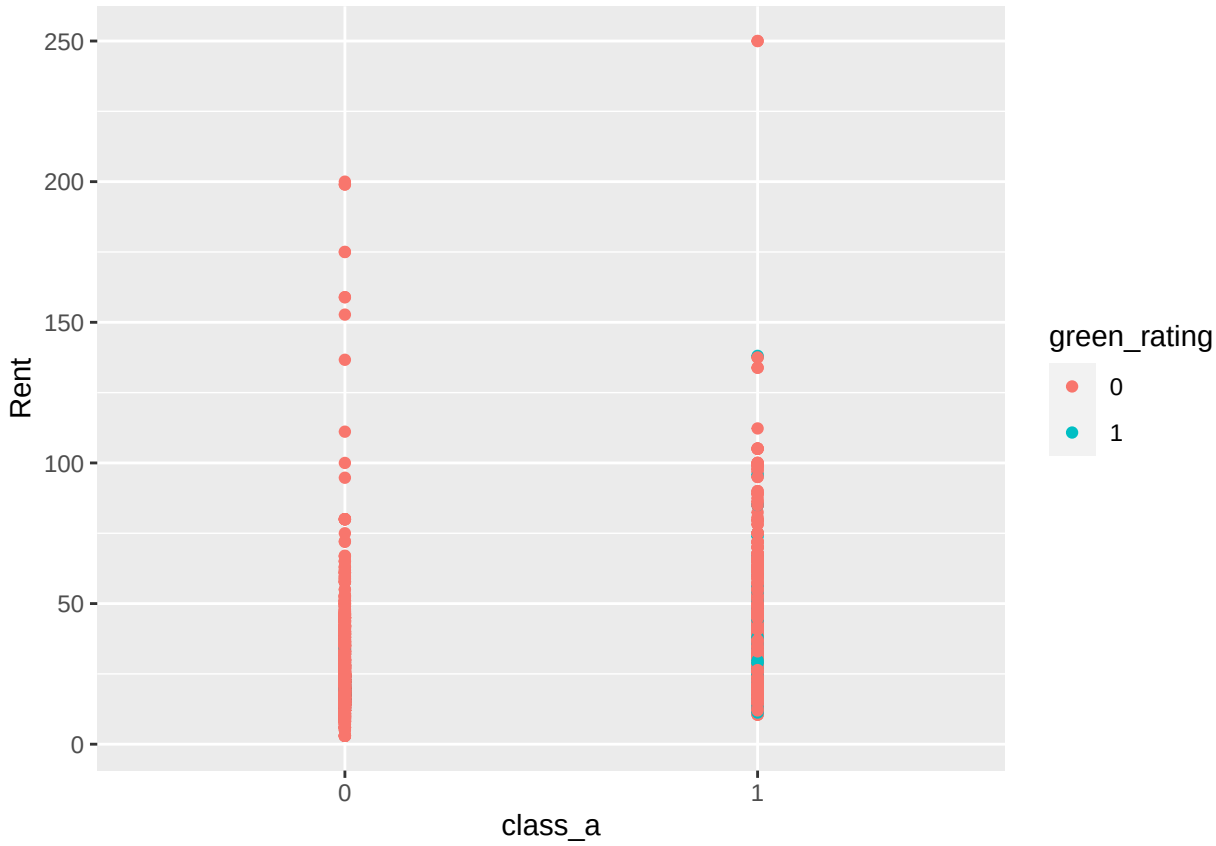

150
100
50
0
Rent
0
1
class_a
green_rating
0
1

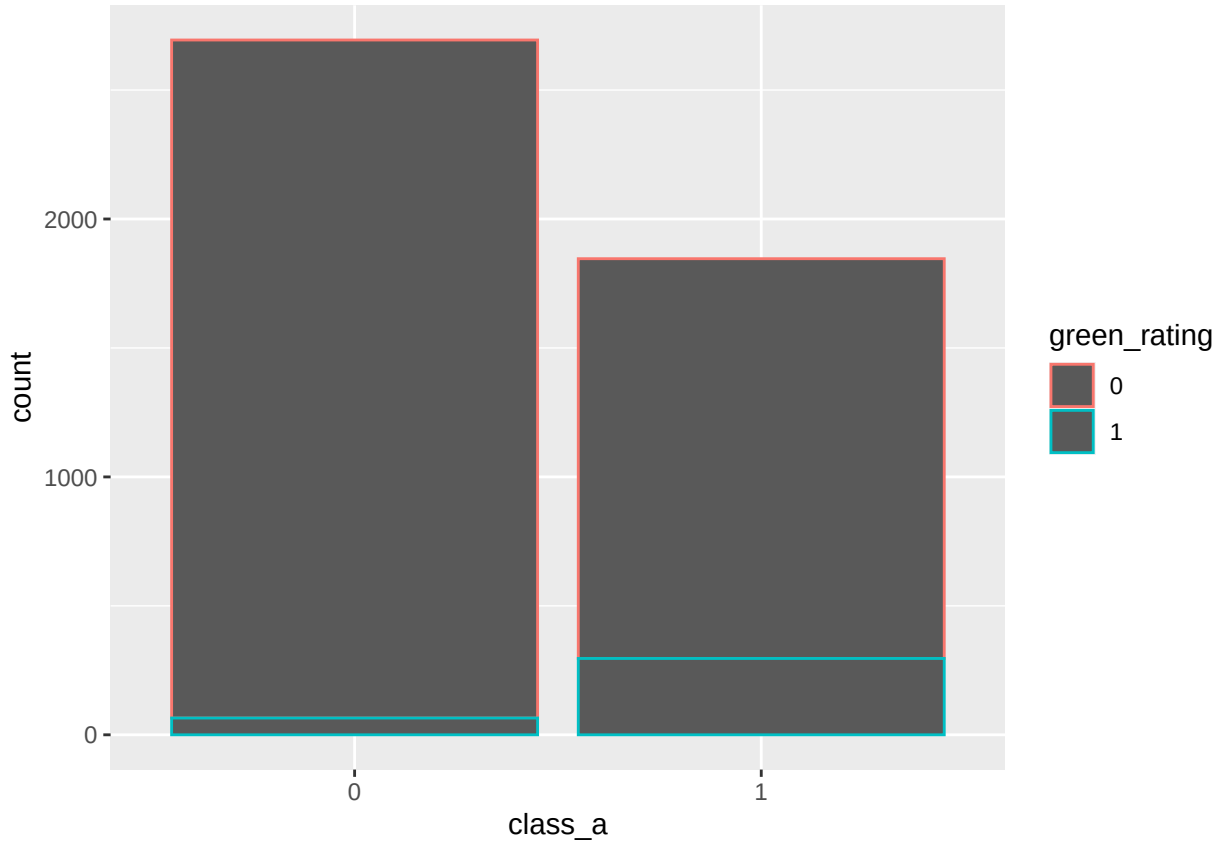

1500
1000
500
0
count
0
1
class_a
green_rating
0
1

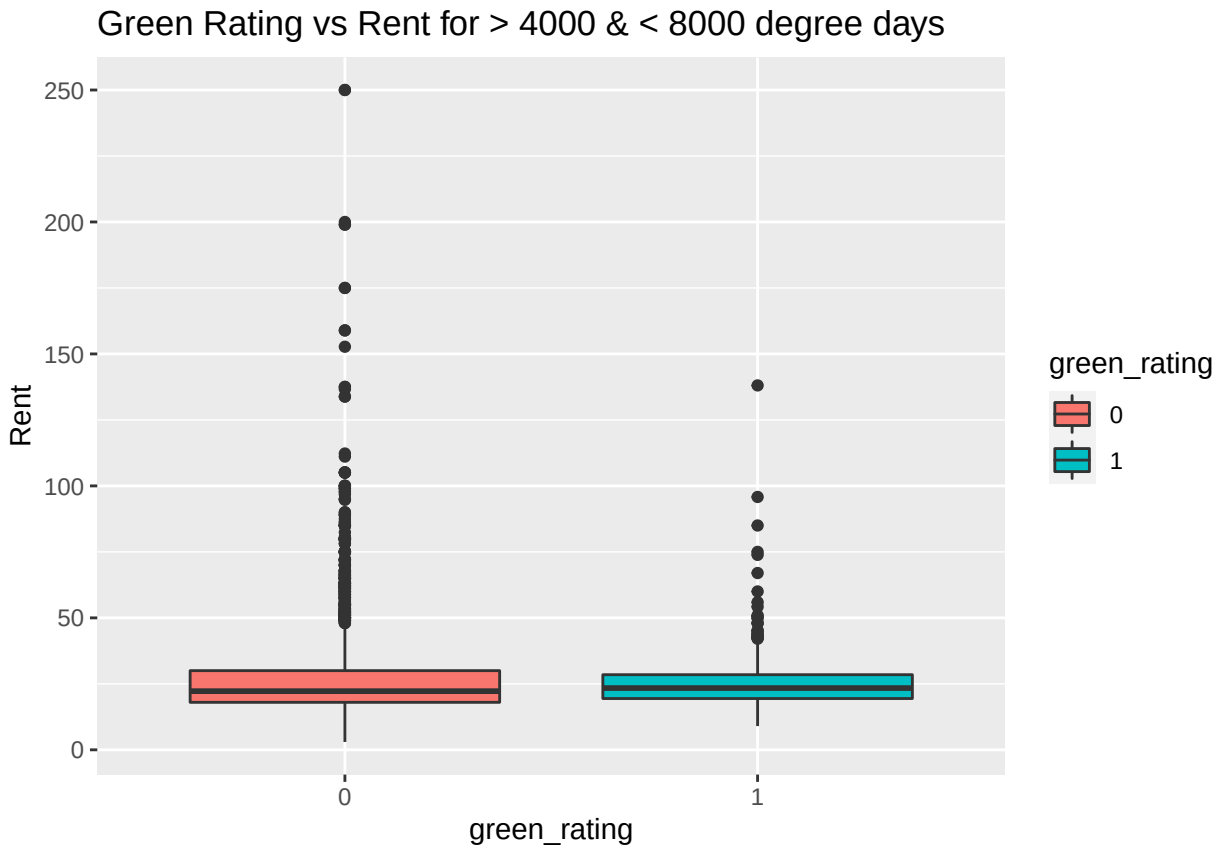


This is telling us that there is almost equal rent associated with a green building and non green building, if it is class_a and in an area with degree days < 4000

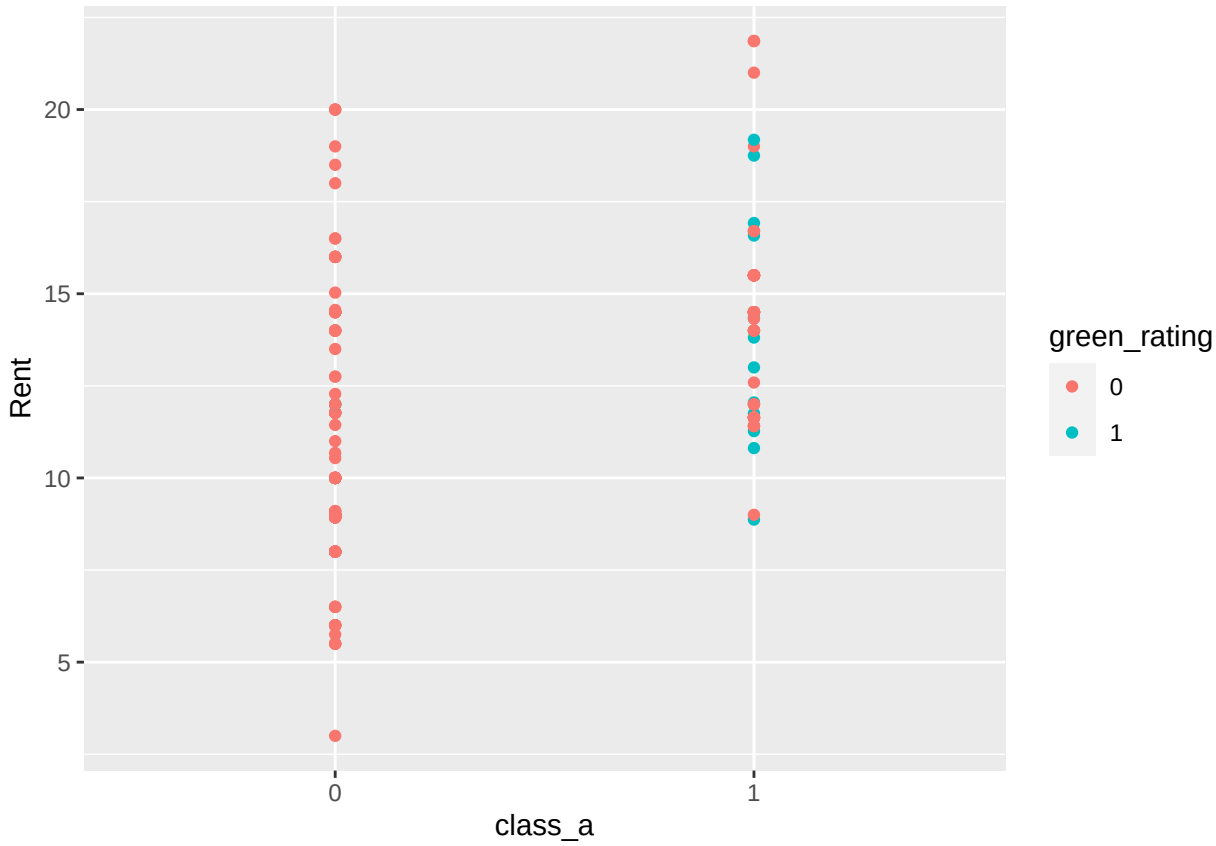
Now we will try for >4000 and <8000

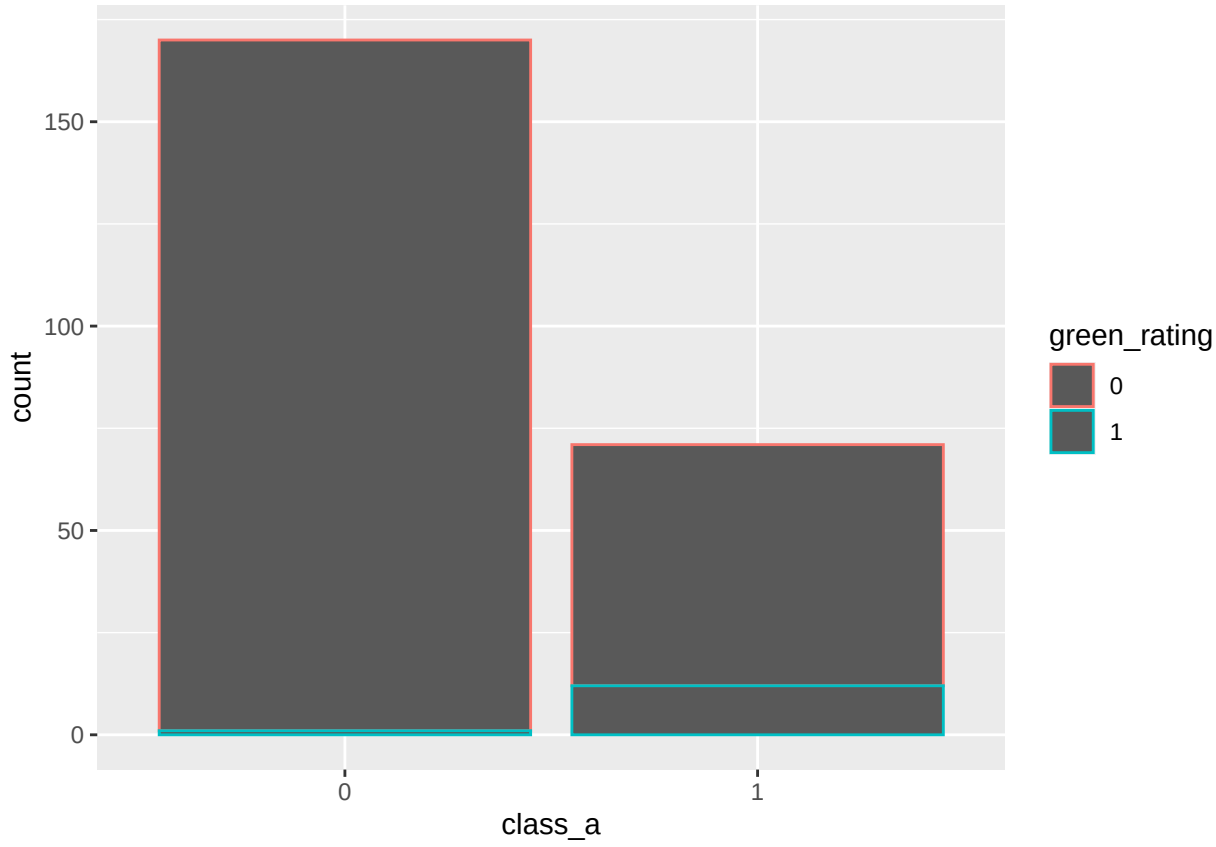

250
200
150
100
50
0
Rent
0
1
class_a
green_rating
0
1

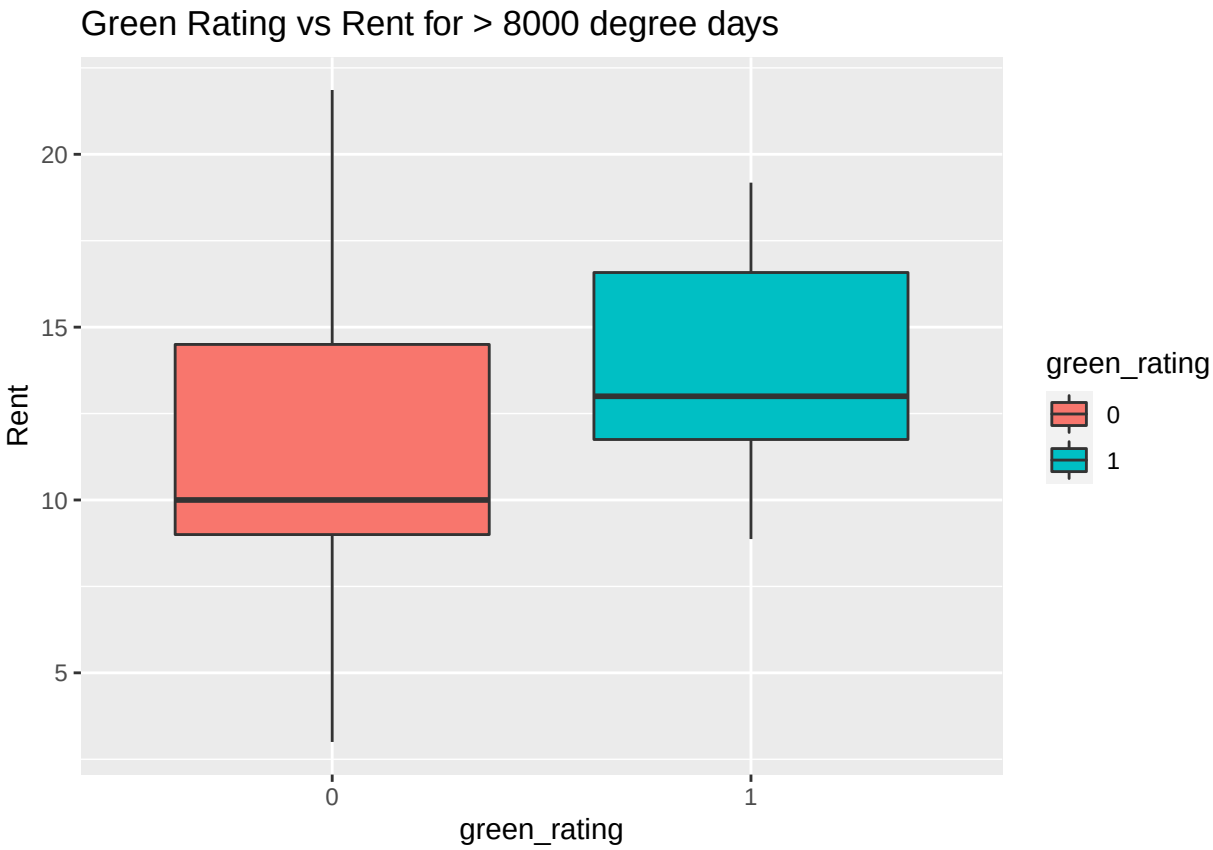

2000
1000
0
count
0
1
class_a
green_rating
0
1



the Rent is comparable for moderate degree days conditions, now we will check for extreme conditions where there are more than 8000 degree days.

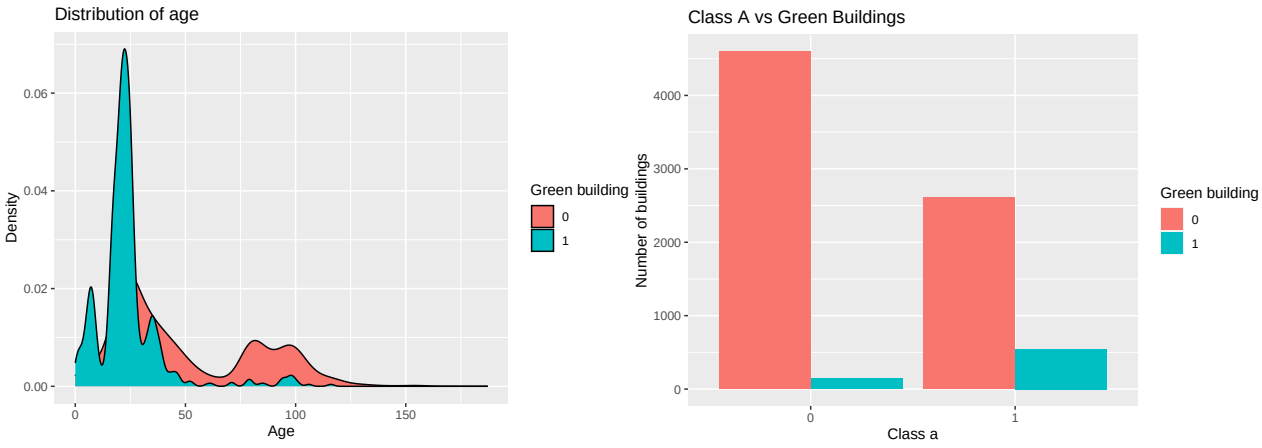

Rent
20
15
10
5
0
1
class_a
green_rating
0
1

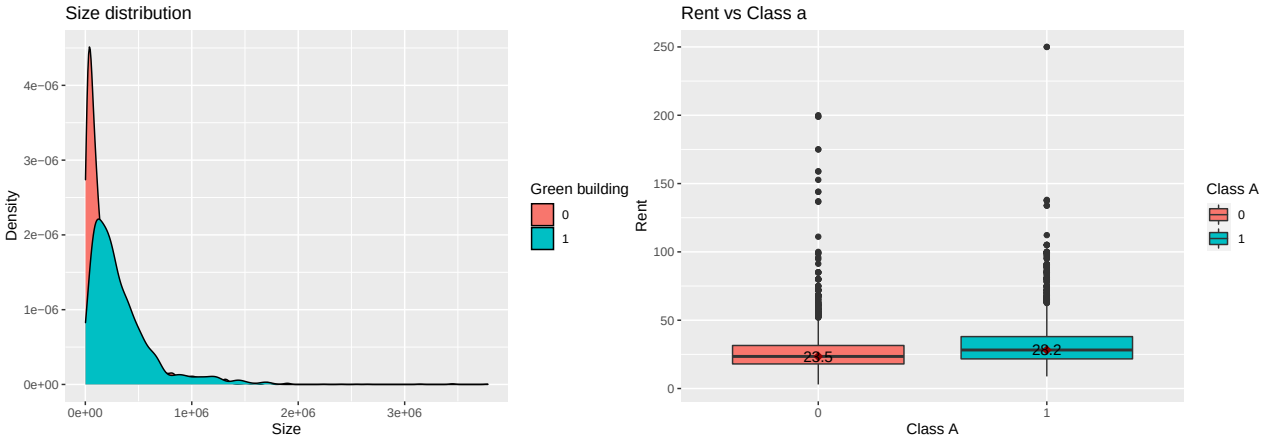

count
150
100
50
0
0
1
class_a
green_rating
0
1



Here we can again confirm that we see rent for green buildings higher for class a In all areas with greater than $>$ 8000 degree days, we see that across the class of the buildings, the rent is higher for green_rated buildings.

Comparing boxplots we can conclude that degree days is a confounding variable which should've been a part of the guru's analysis and we should invest in a green building if they are going to be built in areas with a high number of degree days (>8000), ie areas with extremes of temperature.





Observations: Most of the green buildings are newly built as compared to non-green buildings. There are more class a buildings in green buildings than non-class a buildings. More number of green buildings have higher size than non-green buildings. There is a difference in the rent of class a and non-class a buildings,class a buildings charging more rent.

Conclusion:

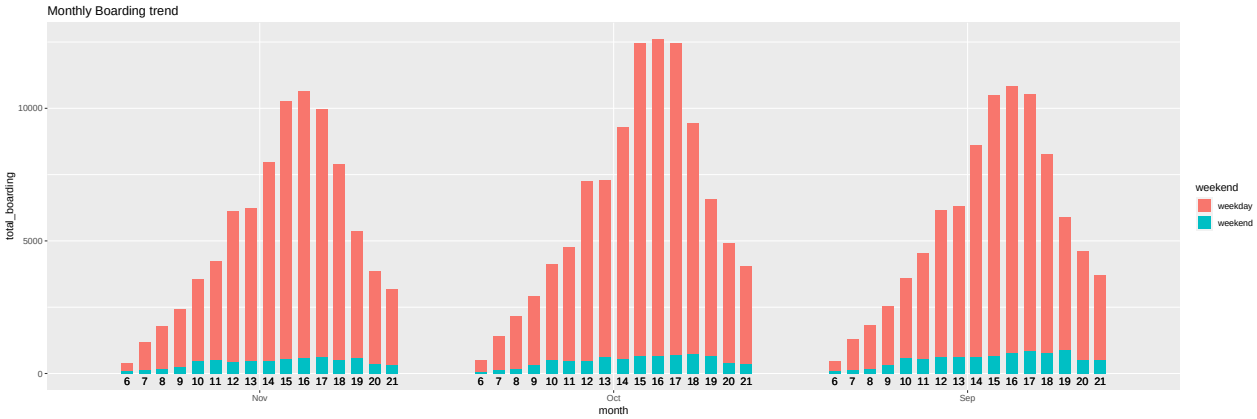
We can conclude that the guru should've taken other variables such as number of degree days, class a b into his consideration.

It would've been financially more feasible if the building is a class a building built in areas with higher number of degree days / more extreme weather conditions, thereby convincing renters to pay higher rent in expectations of having higher savings.

Question 4: Visual story telling part 2: Capital Metro data

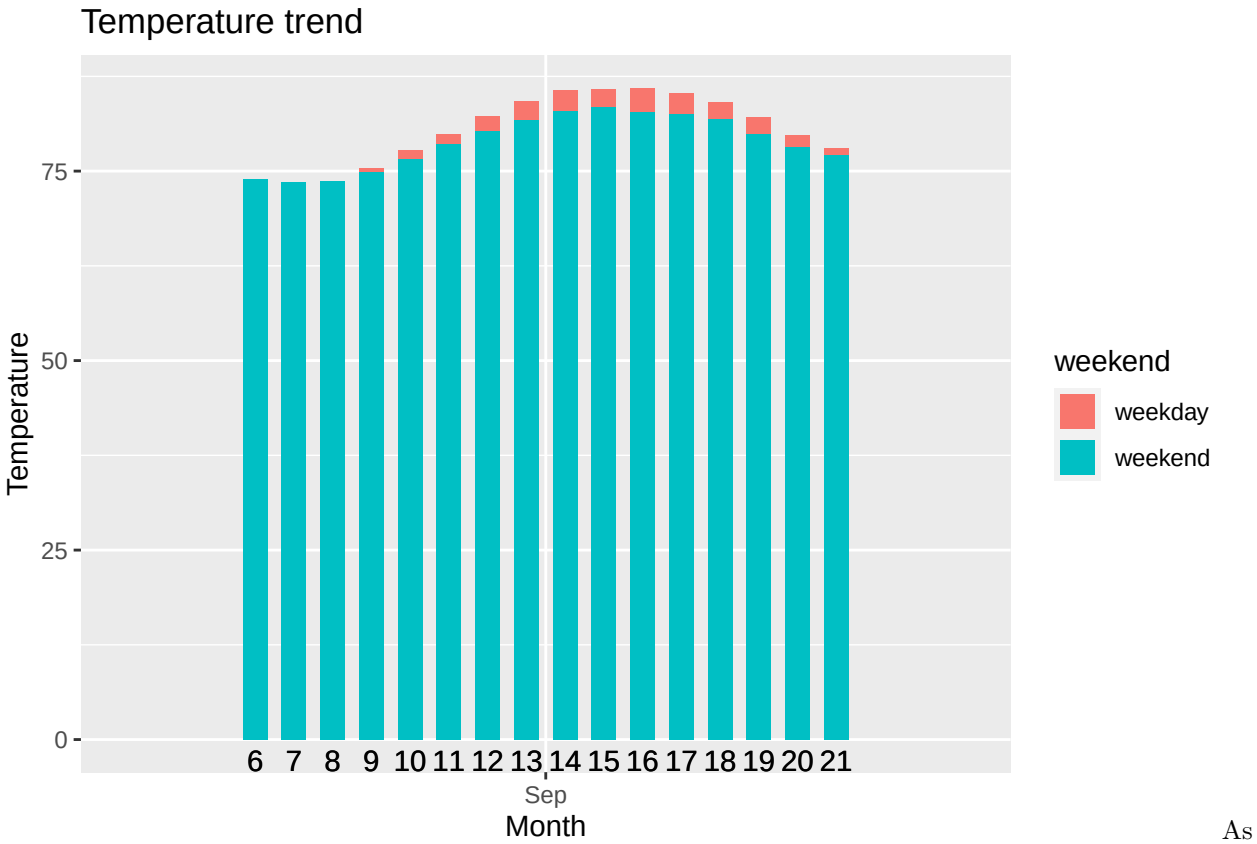
In this exercise, we dive into the Capital metro ridership data to uncover trends in ridership over a period of 6 months. For starters, we look at the basic statistical summary of data to catch some low-hanging fruits.

```
##    timestamp                     boarding        alighting
##  Min.   :2018-09-01 06:00:00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:2018-09-23 17:56:15   1st Qu.: 13.00   1st Qu.: 13.00
##  Median :2018-10-16 13:52:30   Median : 33.00   Median : 28.00
##  Mean   :2018-10-16 13:52:30   Mean   : 51.51   Mean   : 47.65
##  3rd Qu.:2018-11-08 09:48:45   3rd Qu.: 79.25   3rd Qu.: 64.00
##  Max.   :2018-11-30 21:45:00   Max.   :288.00   Max.   :304.00
##  day_of_week         temperature     hour_of_day       month
##  Length:5824        Min.   :29.18   Min.   : 6.00   Length:5824
##  Class :character   1st Qu.:59.20   1st Qu.: 9.75   Class :character
##  Mode  :character   Median :72.75   Median :13.50   Mode  :character
##                     Mean   :69.28   Mean   :13.50
##                     3rd Qu.:79.29   3rd Qu.:17.25
##                     Max.   :97.64   Max.   :21.00
##    weekend
##  Length:5824
##  Class :character
##  Mode  :character
##
##
##
```



For all 3 months, we can see that weekdays have more boarding than weekends. More boarding is observed in the middle of the day between 12pm to 6pm, this might be the time when the buses are in most use.

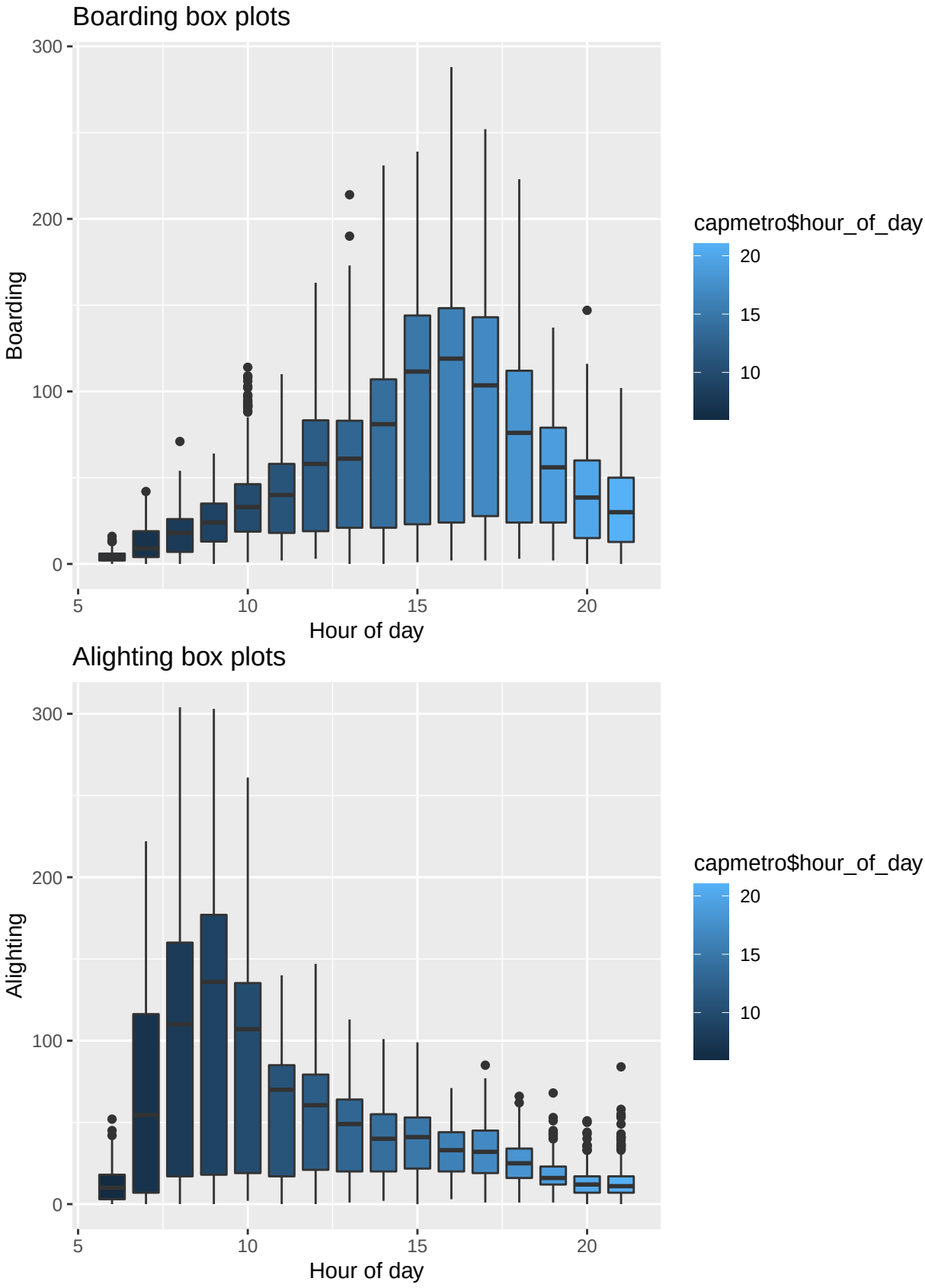
An observation I see while traveling on the buses is their advertisement : come in, cool off. Maybe because temperatures are more during the afternoon, people tend to use buses We will check to see if the temperature graph for a month aligns with this



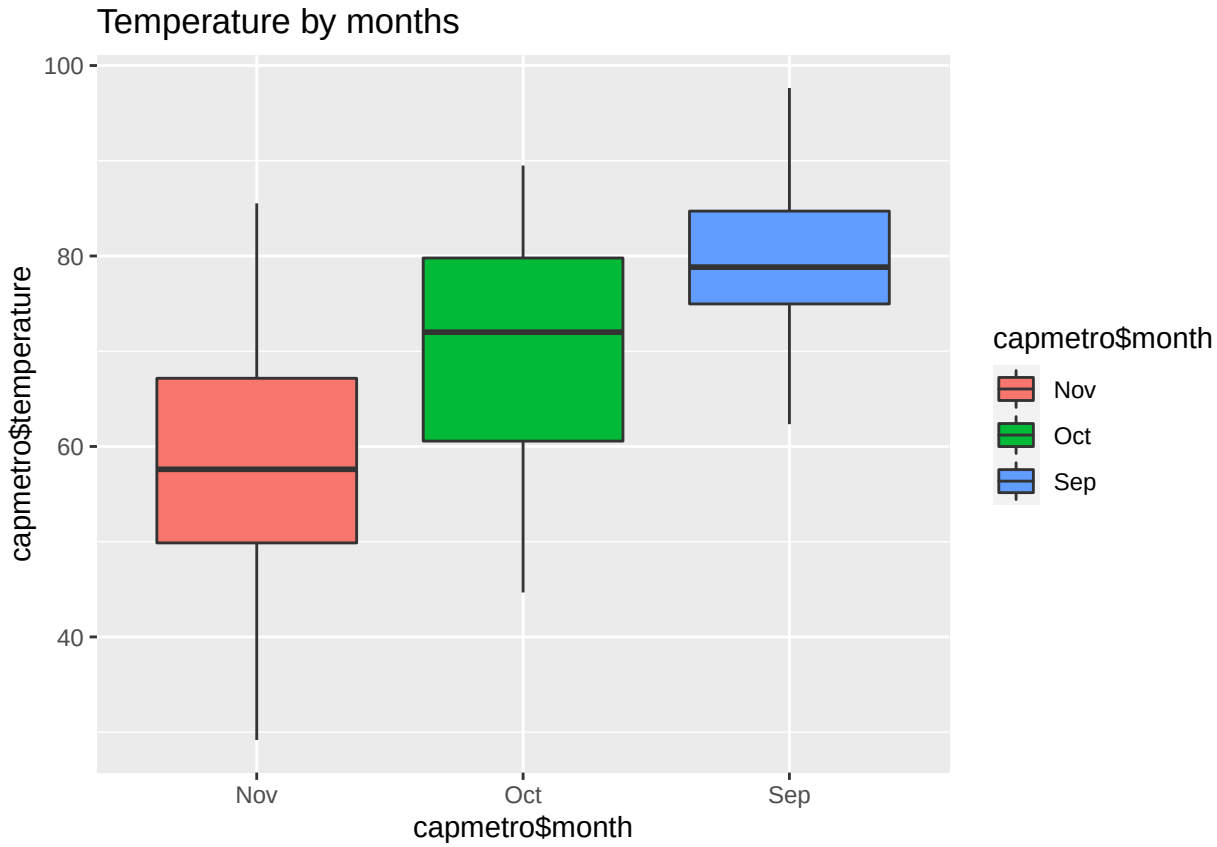
As we can see, the temperatures are more during the same time period boarding is at its peak. This concludes that temperatures along with the advertisement compels people to board more during the afternoon.

In morning, boarding-alighting is negative. Hence we can infer that the buses would be empty between time

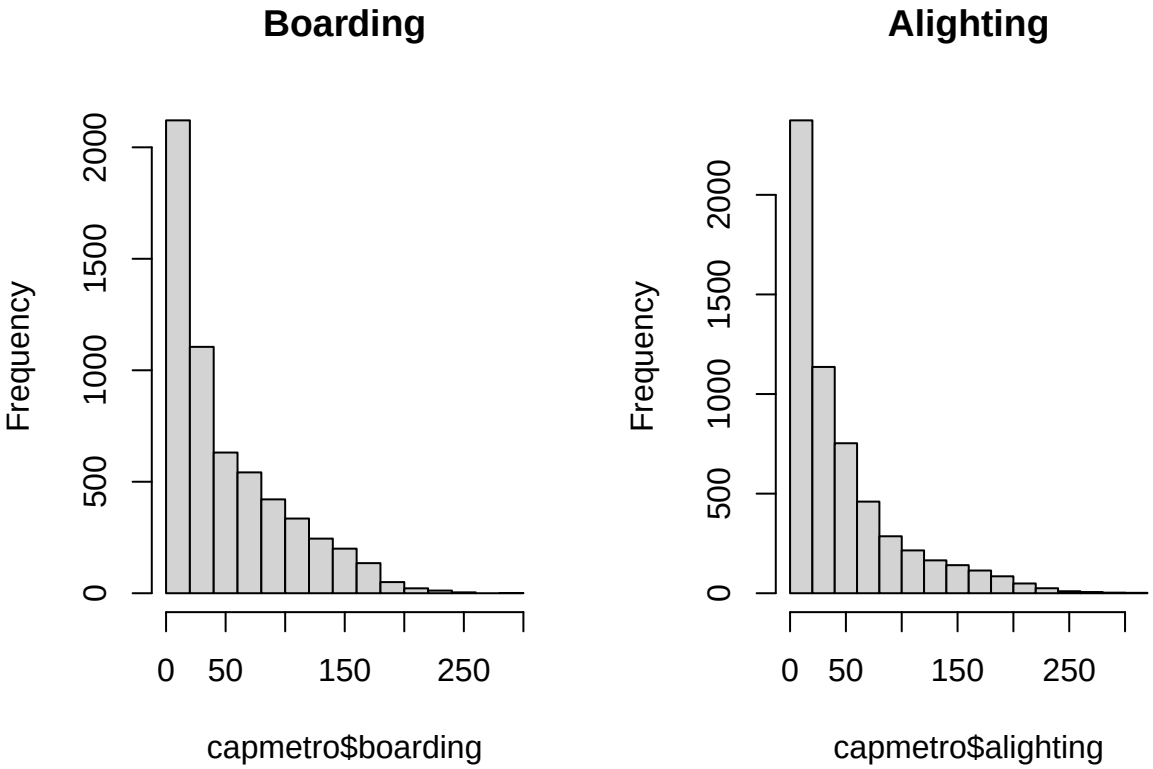
Here is the total line trend of boarding. We see a continuous pattern in the boarding trend, one exception being a prolonged drop in boarding in late November.



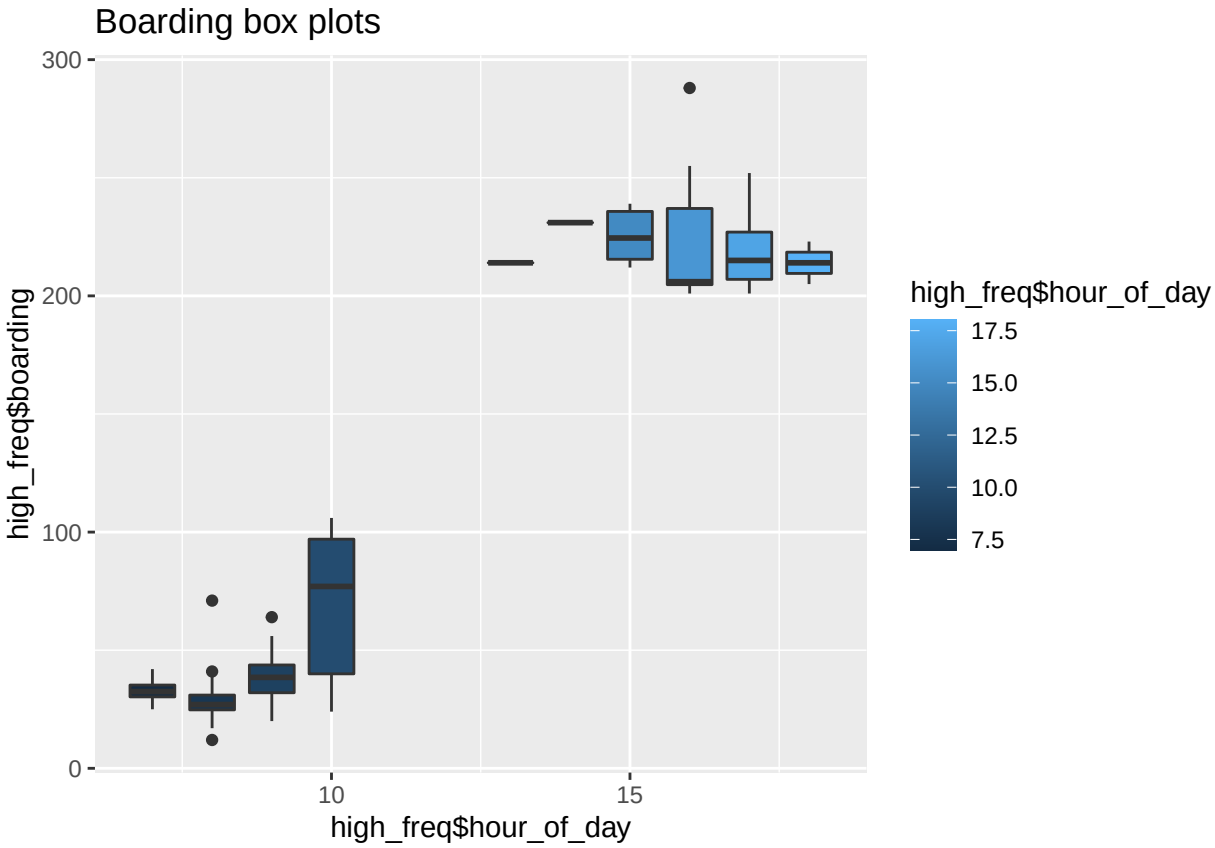
Boarding is at its peak in the middle of the day, alighting is at its peak at the start of the day.

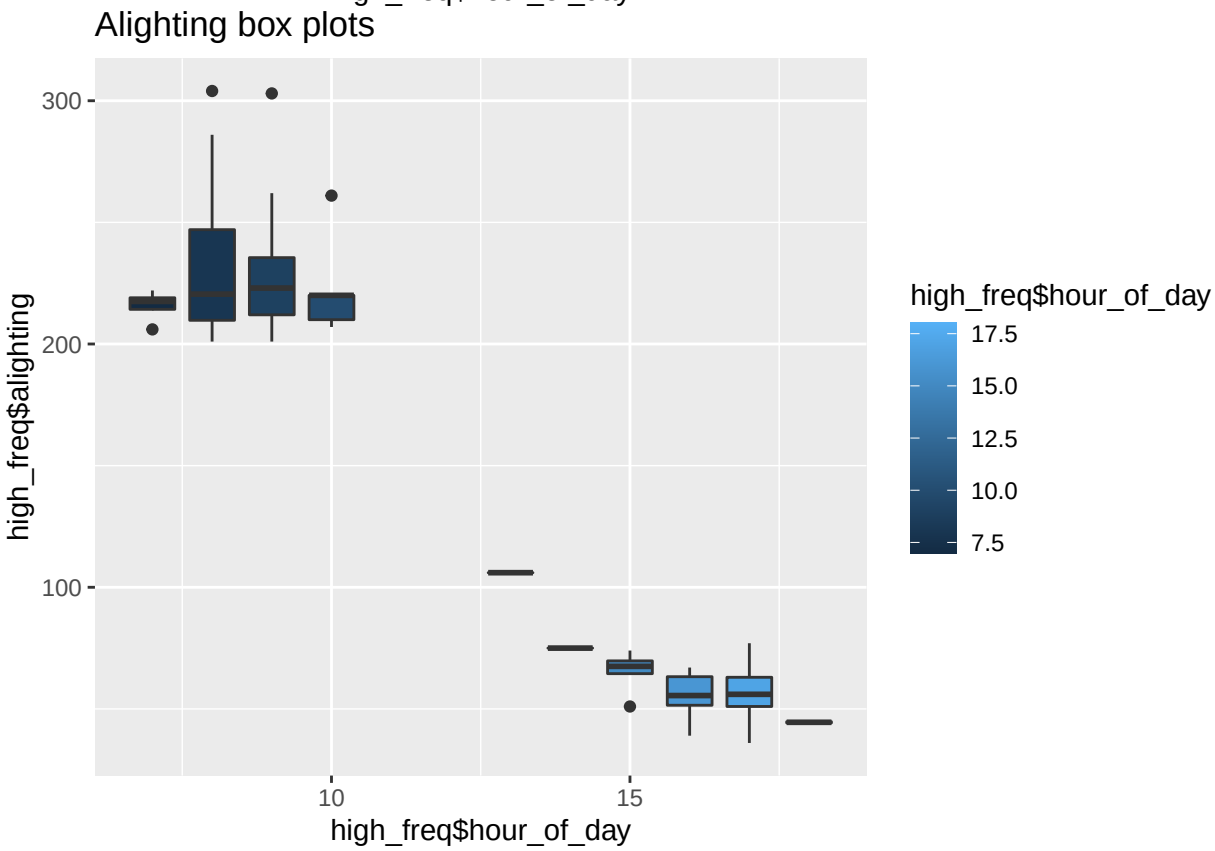


Average temperature is highest in Sep, followed by Oct and Nov



There are also some instances where more than 200 people are either boarding or alighting





High boarding(>200) happens usually around 3pm and high alighting happens usually around 8:30am

Question 5: Portfolio modeling

In this problem, you will construct three different portfolios of exchange-traded funds, or ETFs, and use bootstrap resampling to analyze the short-term tail risk of your portfolios.

We have selected 7 ETFs from the following categories in the ETF database: * 1. Mid Cap Growth Equities ETFs: IJH * 2.Diversified ETFs: DWAT * 3. Corporate Bonds ETFs: IBCE * 4. Vanguard S&P 500 ETF: VOO * 5. Oil and gas ETFs: USO,UNG * 6. All cap equities ETF: SDY

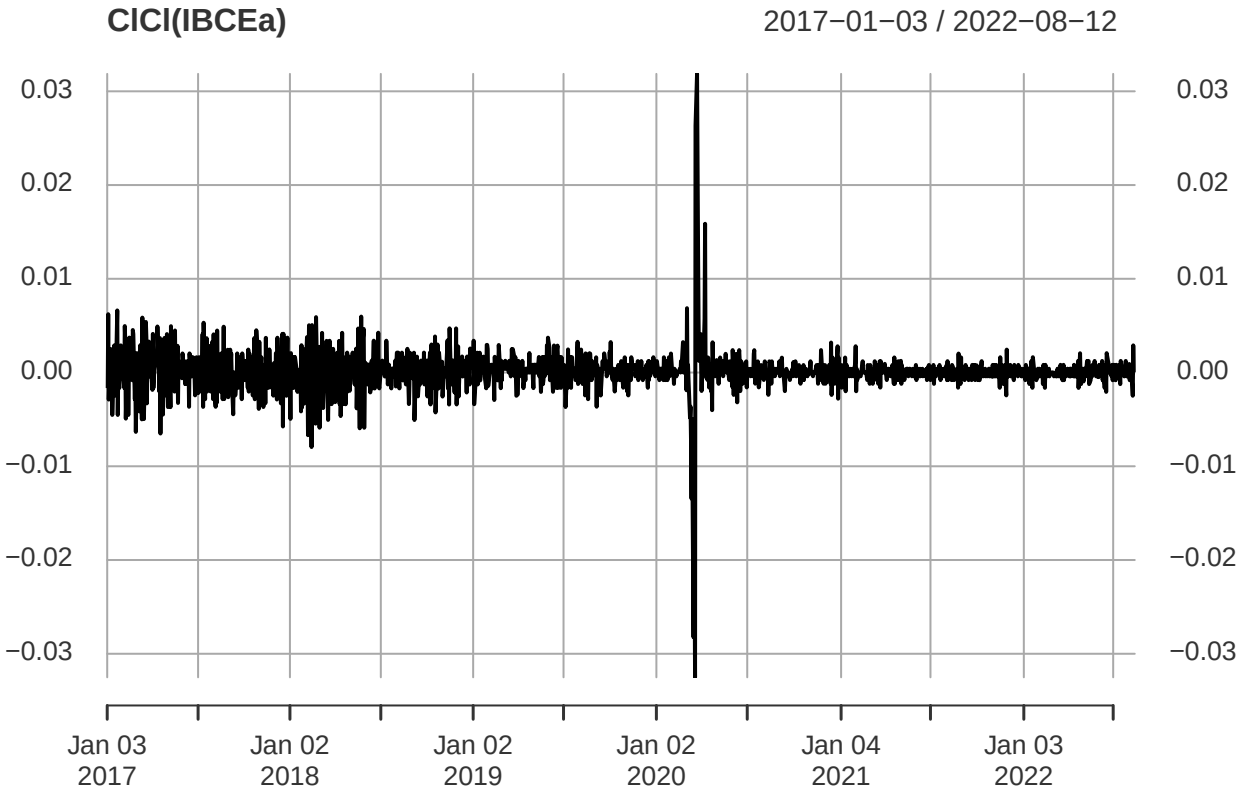
We will create 3 portfolios, one equally balanced across the above listed ETFs, one with more bias towards non-commodity market ETFs(IBCE, IJH, DWAT, VOO, SDY) and third with more bias towards commodity market ETFs (USO,UNG).

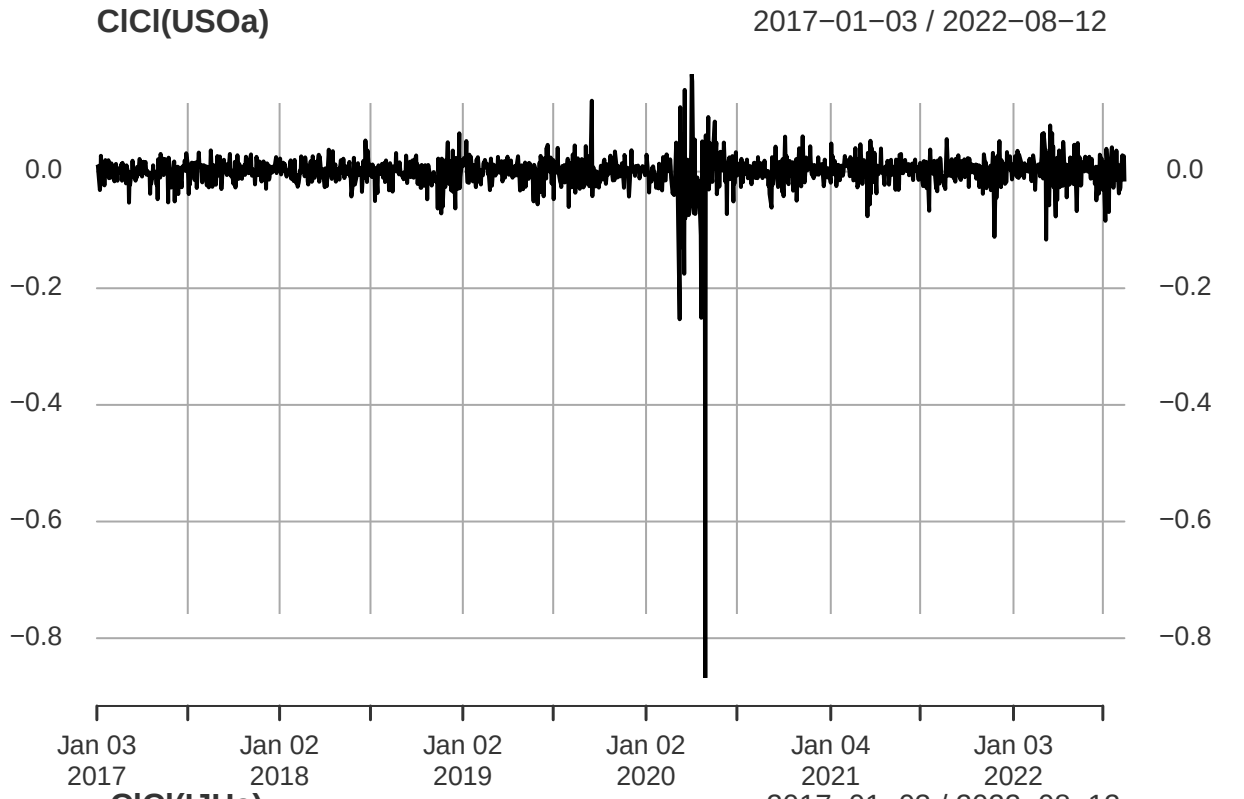
Preprocessing

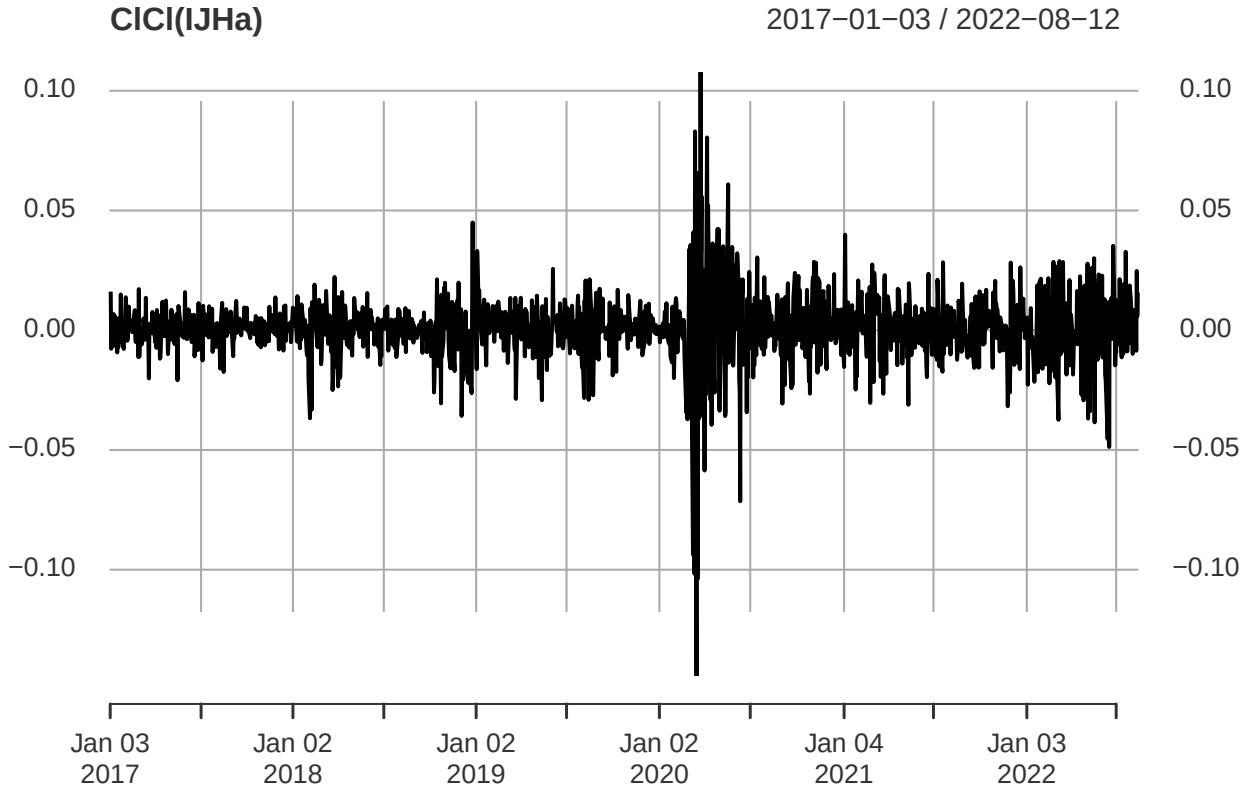
We've considered the past data of 5 years, i.e. starting January 2017 and all these ETFs have data for this range, providing necessary data for analysis.

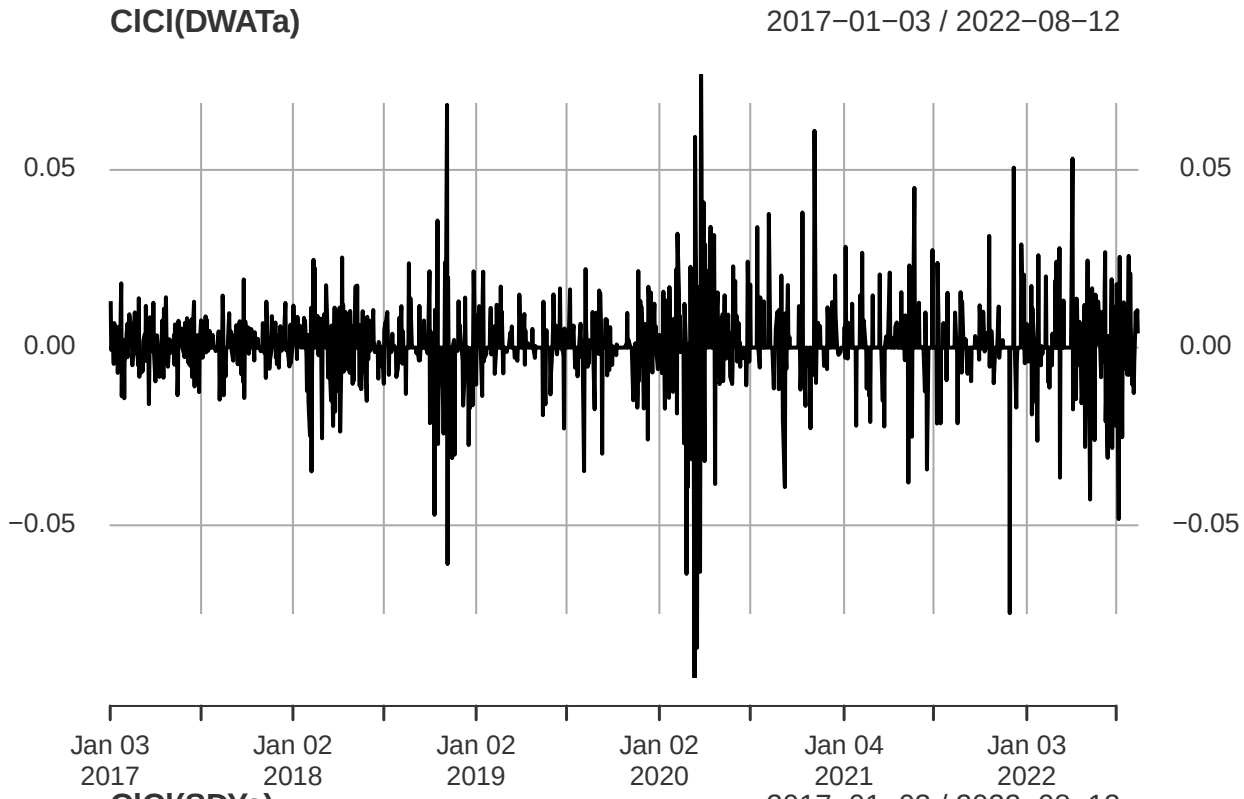
```
## [1] "IJH"  "DWAT" "SDY"  "UNG"  "IBCE" "VOO"  "USO"
```

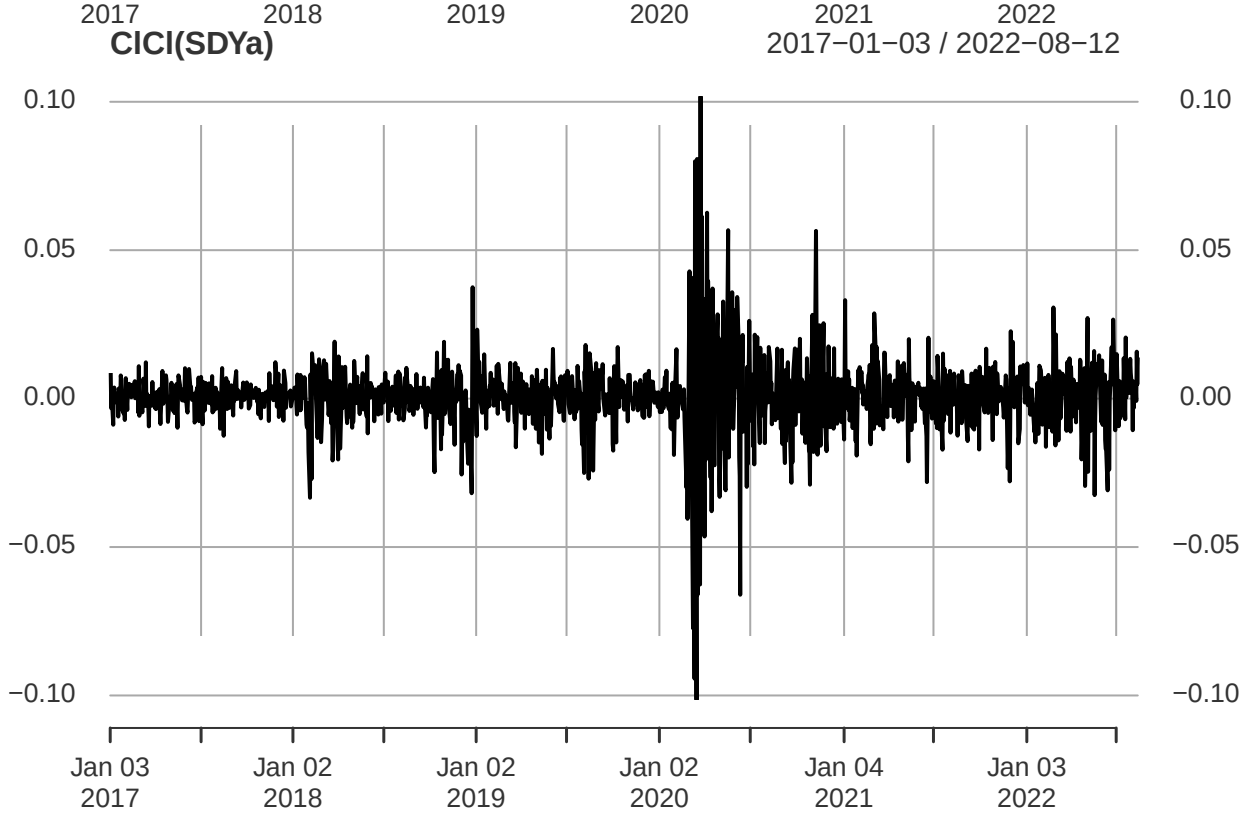
Volatility of the ETFs across the 5 year period.

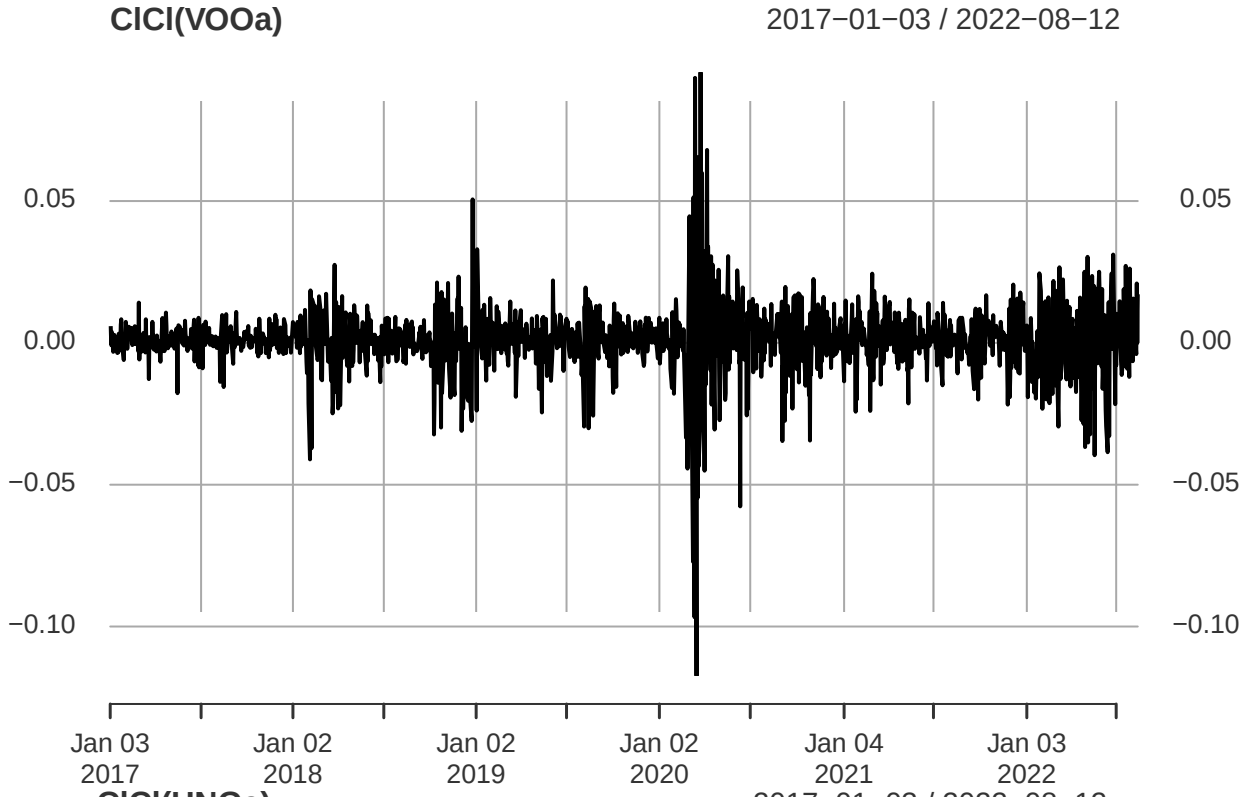


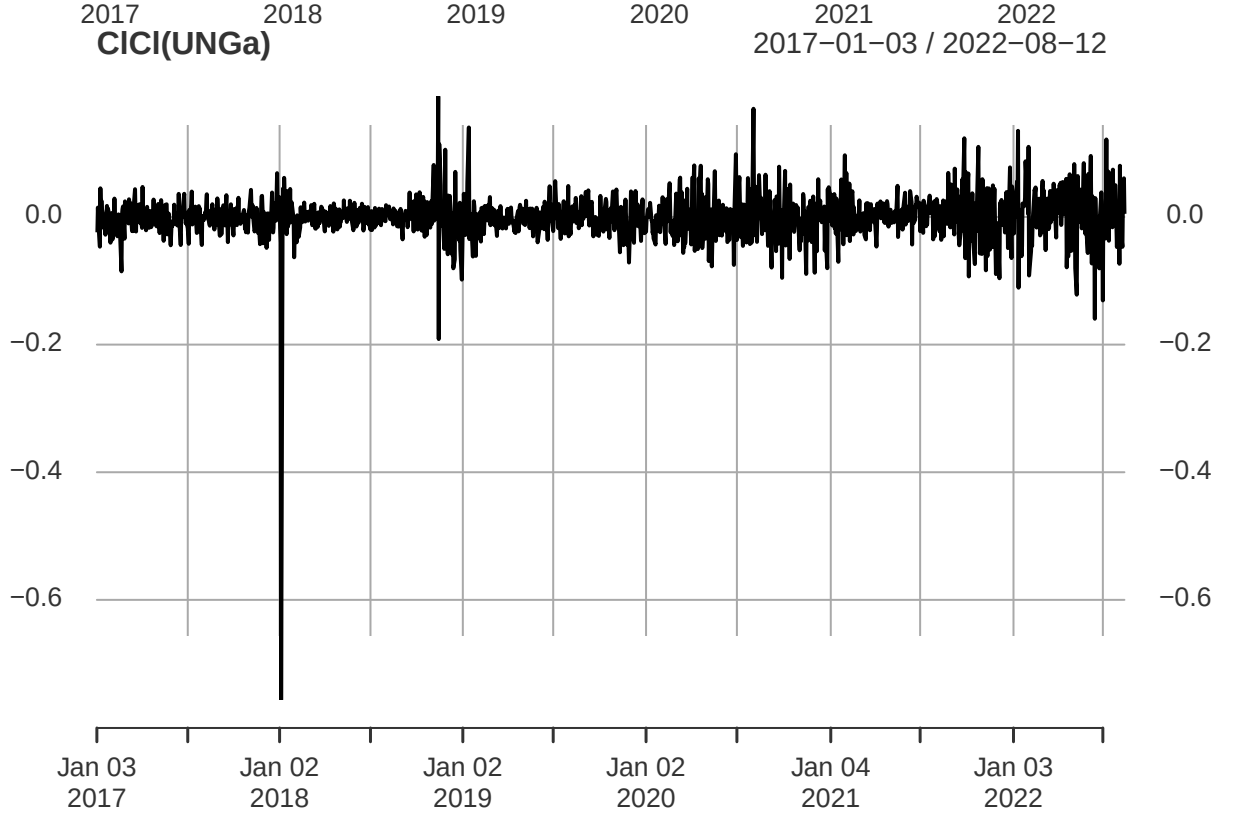

ClCl(USOa)
2017−01−03 / 2022−08−12
0.0
−0.2
−0.4
−0.6
−0.8
Jan 03
2017
Jan 02
2018
Jan 02
2019
Jan 02
2020
Jan 04
2021
Jan 03
2022


ClCl(IJHa)
2017−01−03 / 2022−08−12
0.10
0.05
0.00
−0.05
−0.10
Jan 03
2017
Jan 02
2018
Jan 02
2019
Jan 02
2020
Jan 04
2021
Jan 03
2022

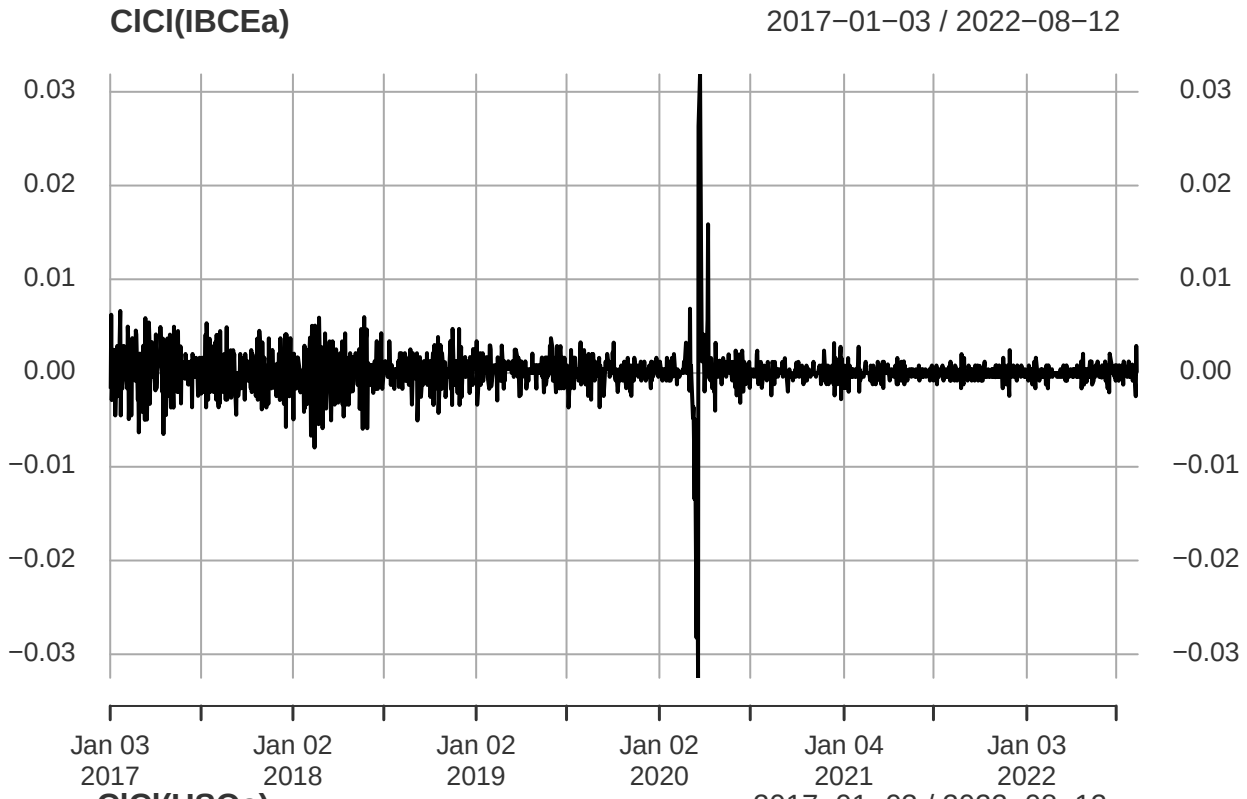

ClCl(DWATa)
2017−01−03 / 2022−08−12
0.05
0.00
−0.05
Jan 03
2017
Jan 02
2018
Jan 02
2019
Jan 02
2020
Jan 04
2021
Jan 03
2022

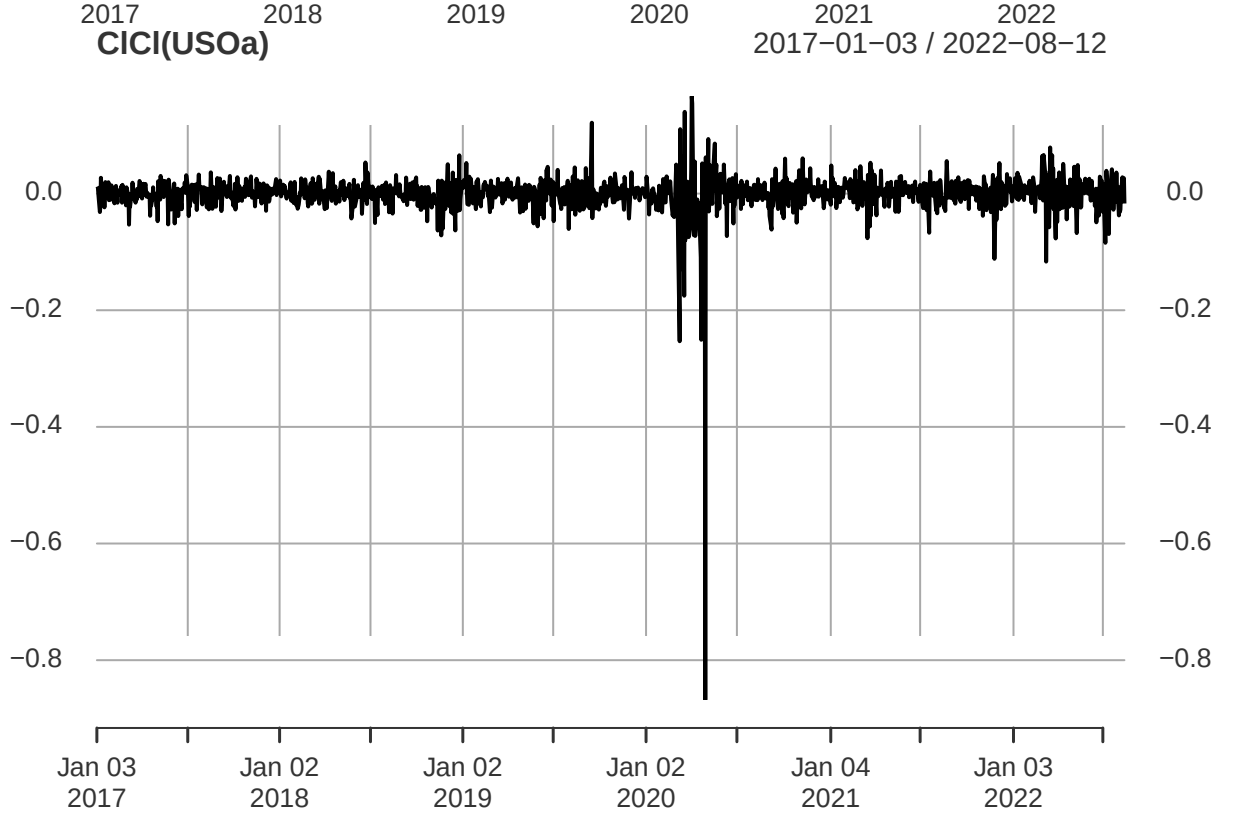

ClCl(SDYa)
2017−01−03 / 2022−08−12
0.10
0.05
0.00
−0.05
−0.10
Jan 03
2017
Jan 02
2018
Jan 02
2019
Jan 02
2020
Jan 04
2021
Jan 03
2022

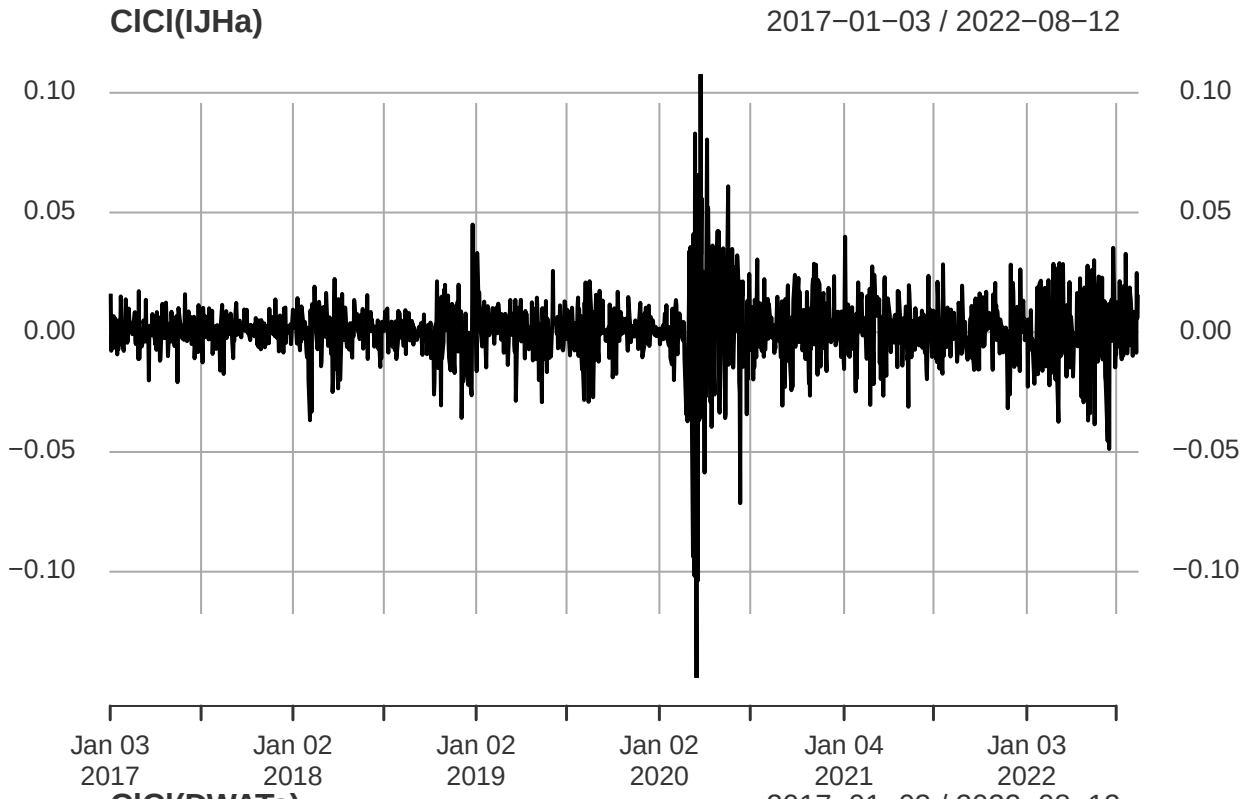


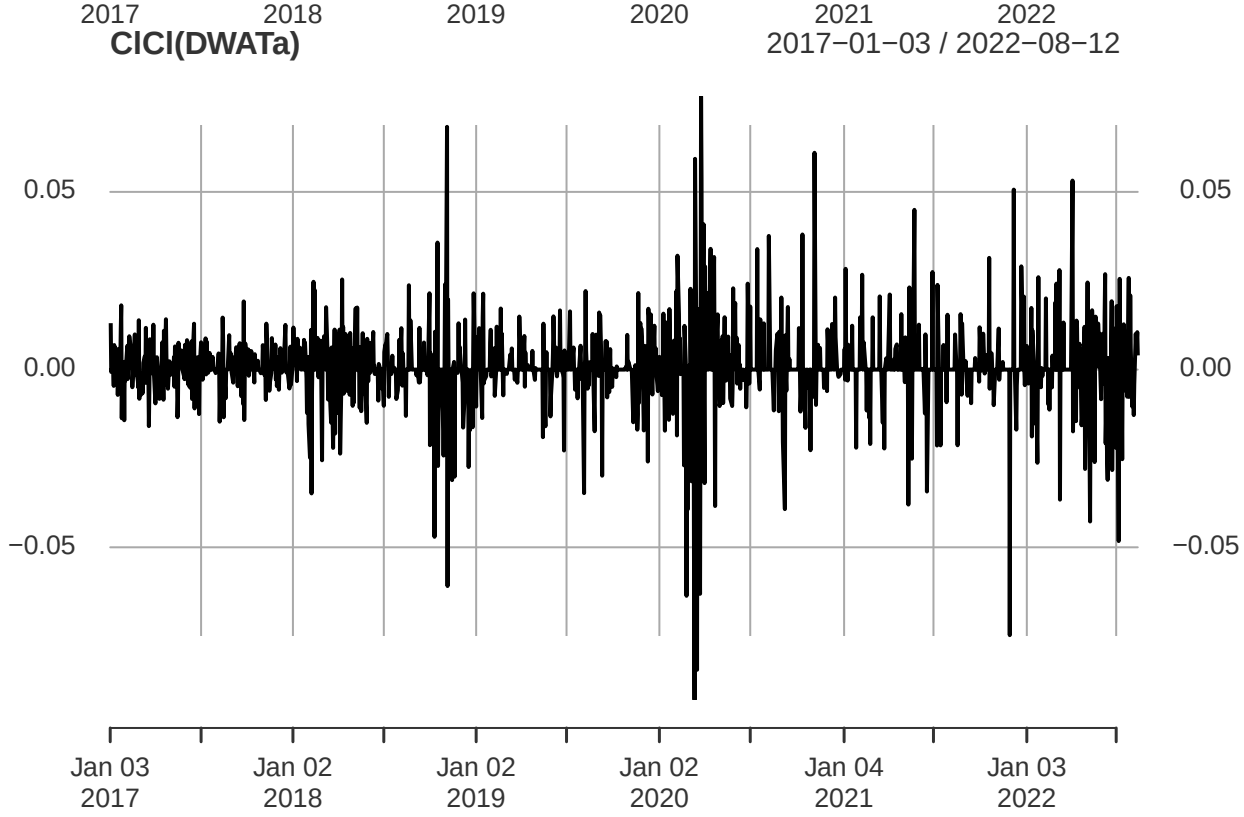


Print close to close changes for some of adjusted ETFs:


ClCl(IBCEa)
2017−01−03 / 2022−08−12
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
Jan 03 2017
Jan 02 2018
Jan 02 2019
Jan 02 2020
Jan 04 2021
Jan 03 2022

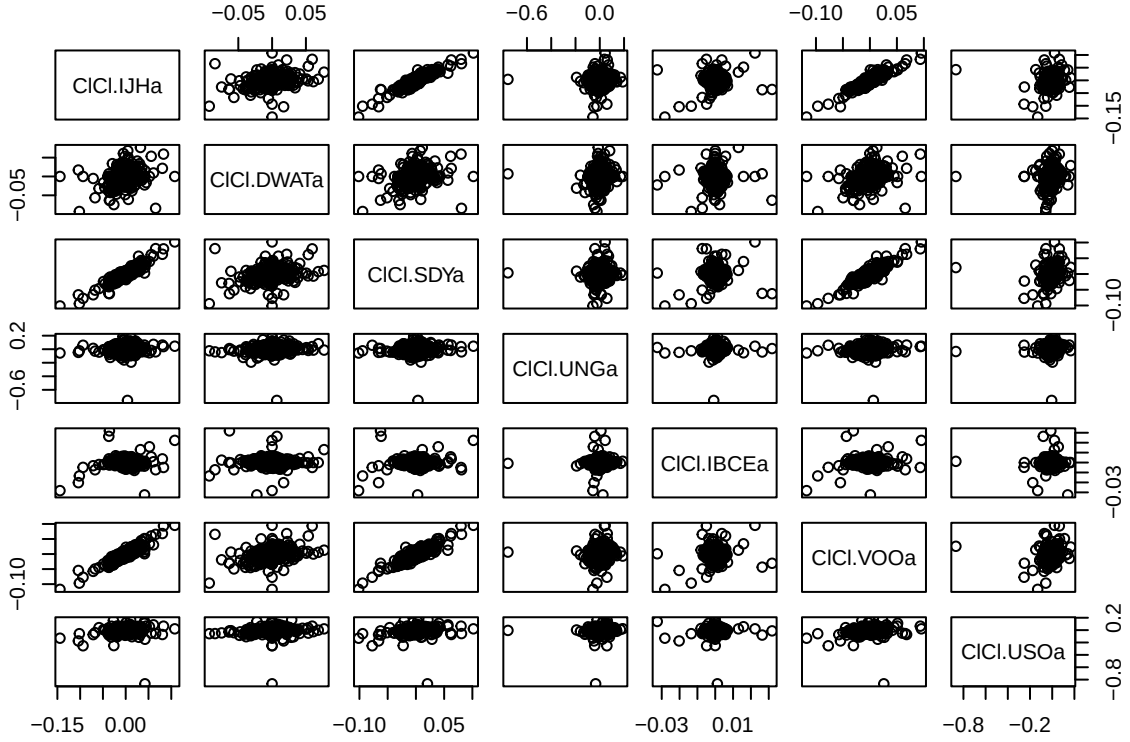

ClCl(USOa)
2017−01−03 / 2022−08−12
0.0
−0.2
−0.4
−0.6
−0.8
Jan 03 2017
Jan 02 2018
Jan 02 2019
Jan 02 2020
Jan 04 2021
Jan 03 2022

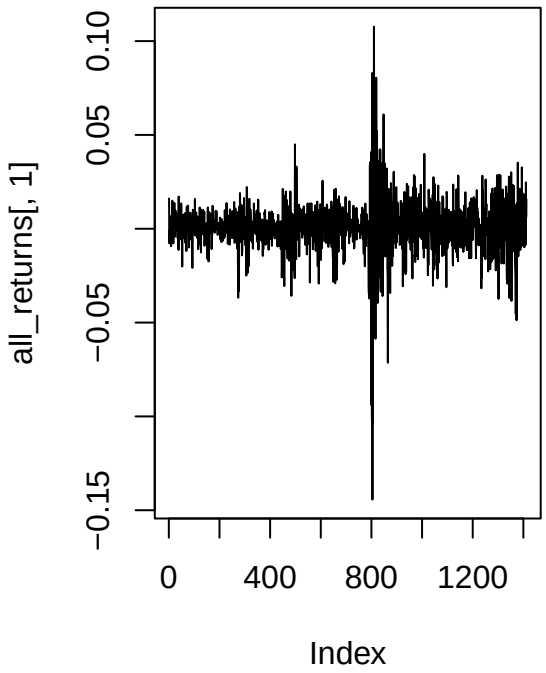





```
##              ClCl.IJHa    ClCl.DWATa    ClCl.SDYa    ClCl.UNGa    ClCl.IBCEa
## 2017-01-04  0.016122289  0.0130624093  0.008624685 -0.023980815 -0.0016562733
## 2017-01-05 -0.007755597 -0.0006685769 -0.003235486  0.014742014  0.0062214431
## 2017-01-06 -0.001014397  0.0002867246  0.001043311 -0.010895914 -0.0028853669
```

```
## 2017-01-09 -0.006629553  0.0000000000 -0.008801413 -0.046511628  0.0024803223
## 2017-01-10  0.006733976 -0.0045862794  0.000467356  0.044929429 -0.0008247423
## 2017-01-11  0.003762572  0.0069111154  0.003737008  0.006142414  0.0012381345
##                ClCl.VOOa    ClCl.USOa
## 2017-01-04  0.0059011414  0.012237784
## 2017-01-05 -0.0007693979  0.010362684
## 2017-01-06  0.0038979691 -0.001709359
## 2017-01-09 -0.0031638176 -0.031678071
## 2017-01-10 -0.0001442606 -0.021220214
## 2017-01-11  0.0028376155  0.027100283
```





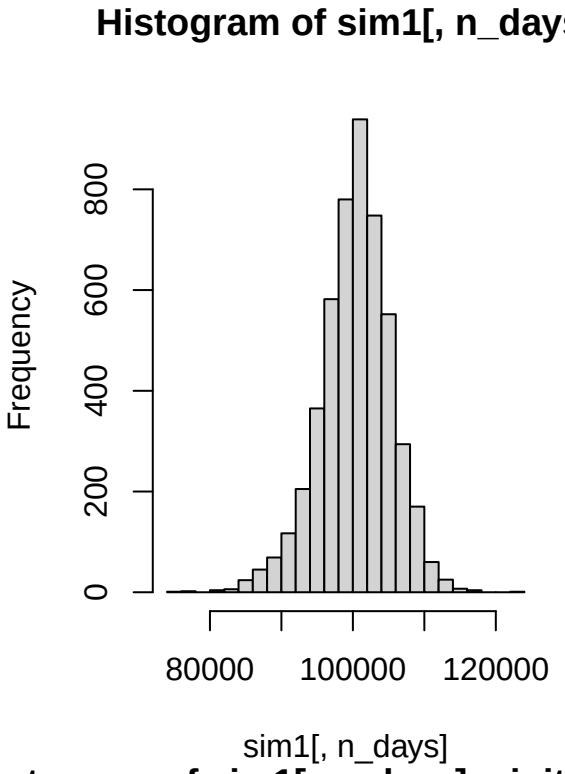
Plotting the pairwise plots of these ETFs, we can see a strong correlation among them. But it is linear for some and non-linear in some cases. A huge spike can be seen in the all_returns plot around 2020, (Beginning of the pandemic).

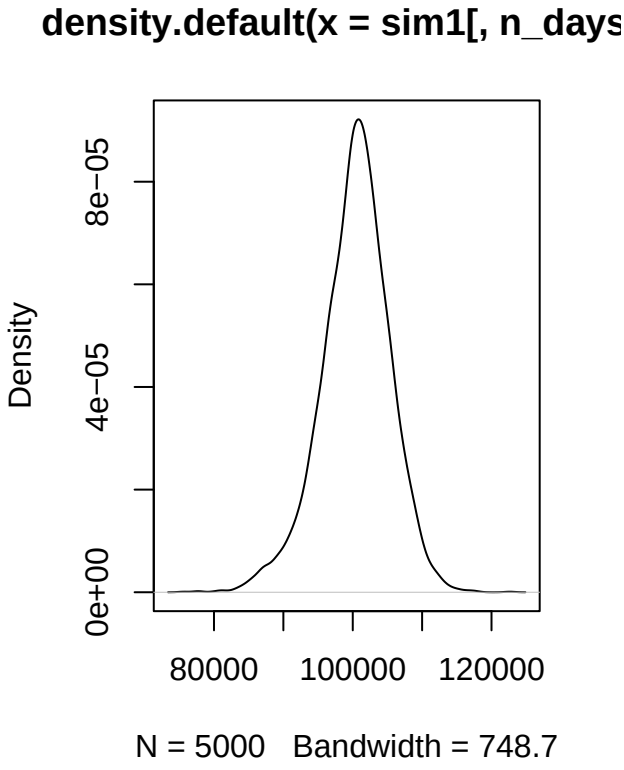
Simulations

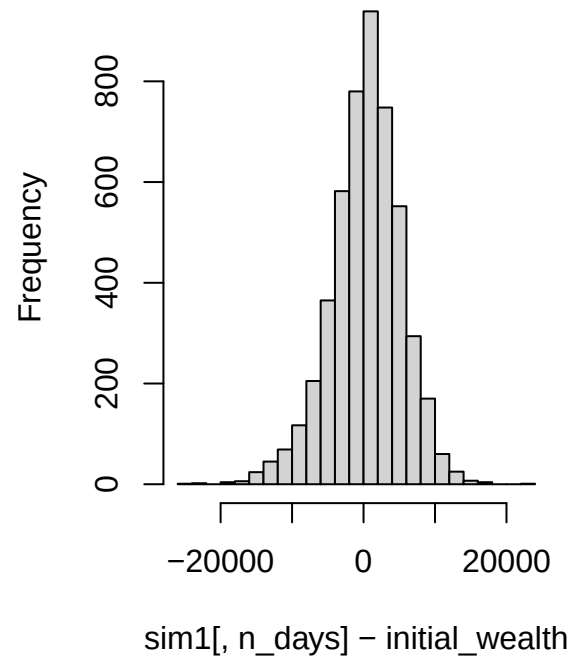
Task: With a capital of $100,000, Use bootstrap resampling to estimate the 4-week (20 trading day) value at risk of each of your three portfolios at the 5% level. Write a report summarizing your portfolios and your VaR findings.

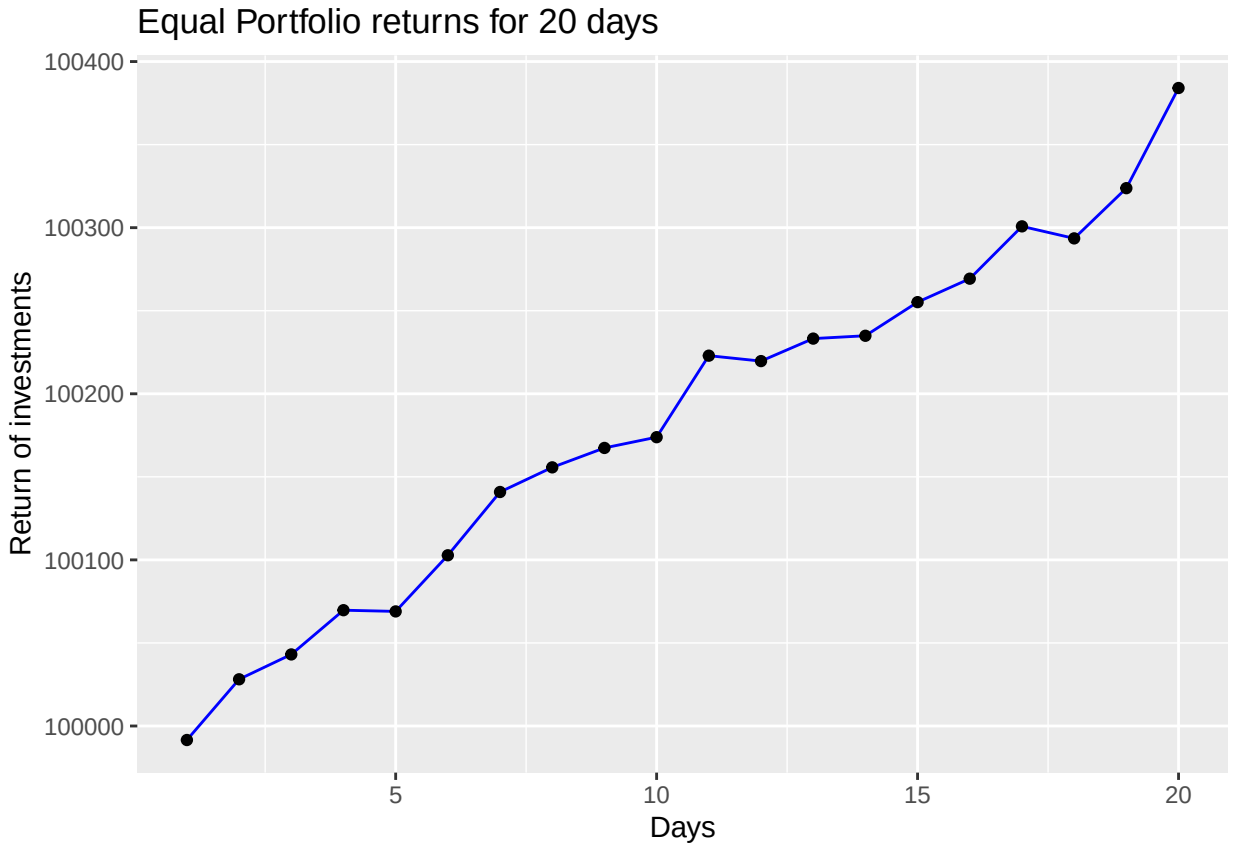
Portfolio 1 : Equal weight Results: Initial wealth: 100000 Average Final Wealth over 20 days :100427.7 Average Profit:427.7061 Value at risk of 5% level: 7912.454

```
##              [,1]      [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## result.1  99757.86  99868.29 102321.4 103813.1 103368.4 103222.0 103539.2
## result.2 100354.89 100274.16 100724.2 100881.0 101449.3 102179.5 101279.7
## result.3 100491.06 101386.56 102131.1 102761.2 102521.5 102783.2 102084.2
## result.4 101900.77 102250.70 103145.4 102944.0 102752.0 102190.4 102451.2
## result.5 100093.98 102161.43 102425.4 102930.2 103885.8 103623.4 103790.7
## result.6 101217.43 101346.51 101691.9 101650.1 104141.3 103399.5 104064.6
##              [,8]     [,9]     [,10]     [,11]     [,12]    [,13]     [,14]
## result.1 101678.0 101008.4  99476.59  99407.93  99519.75  99476.5  99454.22
## result.2 101297.3 100245.9 102210.75 101961.27 102573.73 102950.0 102289.51
## result.3 101681.6 102128.7 102868.93 102519.28 101704.31 102041.4  99089.92
## result.4 103067.1 102859.7 102807.71 102778.68 102401.42 102453.1 102792.62
## result.5 104237.4 103678.1 103470.40 104377.25 104492.53 103129.6 103738.15
## result.6 105092.5 104378.3 103983.89 104492.67 100876.80 101126.1 100216.65
##              [,15]     [,16]     [,17]     [,18]     [,19]     [,20]
## result.1  98405.31  98381.64  98593.59  95741.86  96842.50  96919.71
## result.2 101858.02 100771.64 100734.15 101712.60 103191.12 103108.10
## result.3  96352.71  96347.34  96926.53  96395.06  94913.25  94889.37
## result.4 102792.18 102181.00 100617.65  99942.44  99406.20  99118.20
## result.5 101480.85 101884.21 101946.82 101920.33 102415.96 102854.09
## result.6 100055.99 100753.11 101743.63 100516.42 100281.91 100365.25
```


Histogram of sim1[, n_days]
Frequency
0
200
400
600
800
80000
100000
120000
sim1[, n_days]

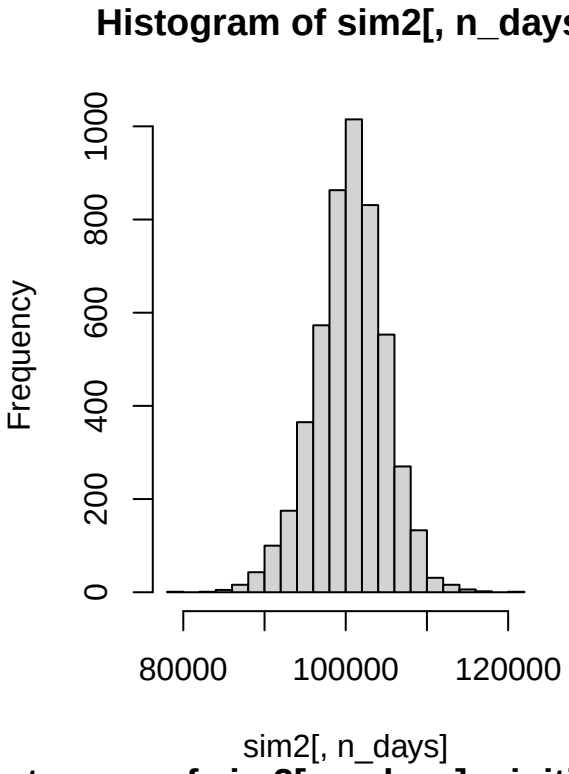

density.default(x = sim1[, n_days
Density
0e+00
4e−05
8e−05
80000
100000
120000
N = 5000 Bandwidth = 748.7

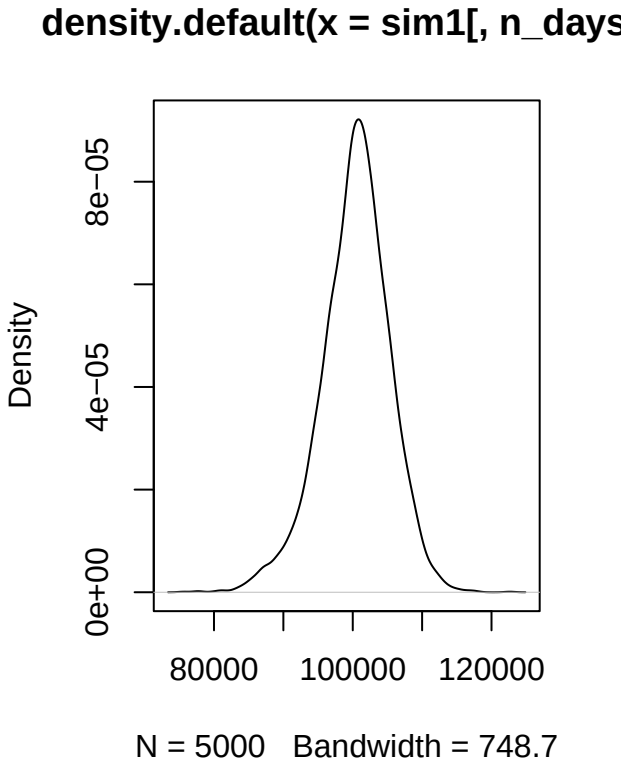

stogram of sim1[, n_days] − initial_
Frequency
0
200
400
600
800
−20000
0
20000
sim1[, n_days] − initial_wealth

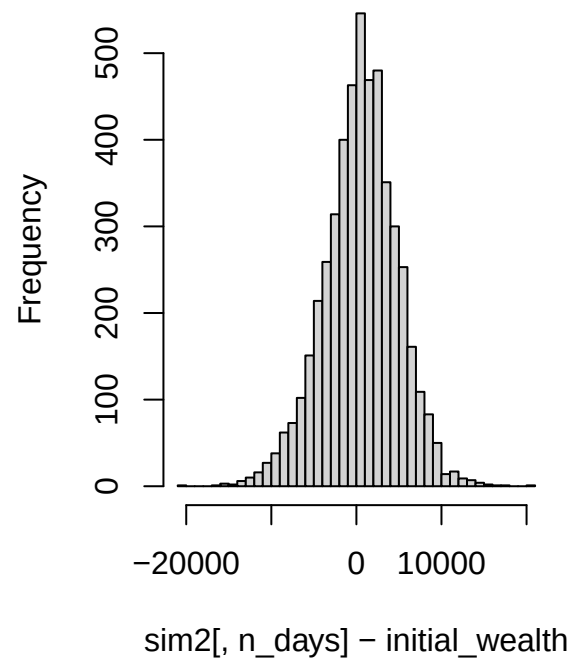


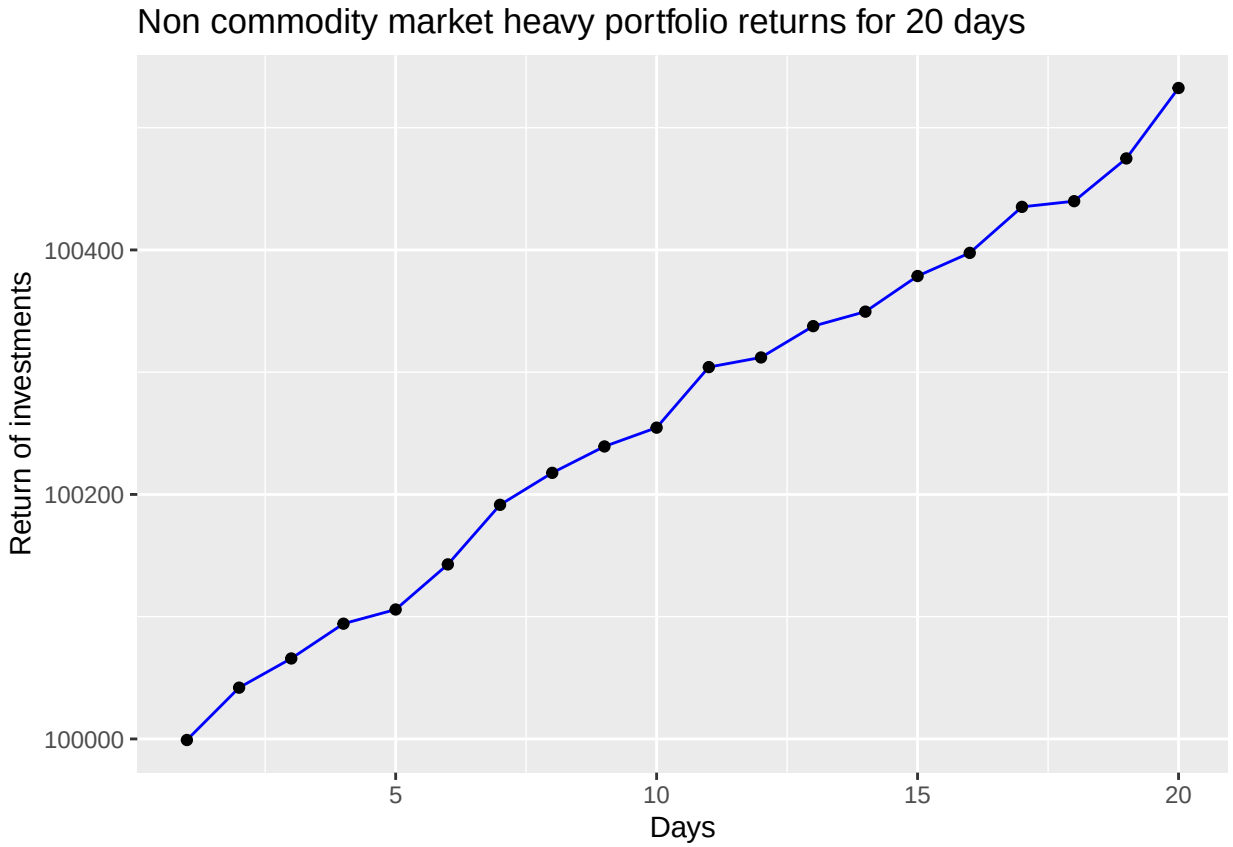
Portfolio 2 : More Bias in weights for non-commodity market ETFs 80% weightage among IBCE, IJH, DWAT, VOO, SDY 20% weightage among UNG, USO Results: Initial wealth: 100000 Average Final Wealth over 20 days :100348.7 Average Profit:348.6563 Value at risk of 5% level: 8389.12

```
##              [,1]      [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## result.1  99843.54  99922.08 101079.8 102045.3 101748.2 101639.3 101850.7
## result.2 100345.57 100514.02 101072.3 101363.0 101745.0 102294.8 101699.5
## result.3 100335.51 100981.58 101929.8 102459.0 102291.0 102693.9 102370.1
## result.4 101425.16 101664.28 102278.2 102102.9 101943.3 101374.1 101773.3
## result.5 100104.16 101958.40 102261.1 102819.8 103796.4 103402.6 103541.9
## result.6 101338.16 101349.51 101612.9 101480.2 103509.0 103728.1 104414.1
##              [,8]      [,9]     [,10]     [,11]     [,12]     [,13]     [,14]
## result.1 100054.2  99574.41  98041.37  97951.87  98060.86  98006.28  98087.47
## result.2 101388.9 100494.49 102144.42 101763.28 102325.49 102726.79 101865.95
## result.3 102157.7 102408.74 102965.71 102709.93 102187.59 102448.18  99987.56
## result.4 102192.8 102040.69 102156.44 102128.42 101769.48 101831.72 102178.75
## result.5 103982.3 103617.35 103466.50 104265.76 104095.56 102810.32 103316.18
## result.6 105181.2 104535.75 104152.38 104733.27 101298.82 101526.63 100826.40
##              [,15]     [,16]     [,17]     [,18]     [,19]     [,20]
## result.1  97188.24  97236.6  97366.38  95027.81  96064.34  95737.04
## result.2 101287.45 100246.2 100207.43 100770.89 102089.17 102260.54
## result.3  97962.81  97864.9  98268.23  97627.21  96156.86  96001.74
## result.4 102141.32 101615.9 100290.63  99712.84  99362.88  99290.82
## result.5 101066.66 101320.4 101319.55 101439.71 101929.82 102241.48
## result.6 100884.48 101473.9 102391.03 100783.32 100618.08 100754.10
```


Histogram of sim2[, n_days]
Frequency
0
200
400
600
800
1000
80000
100000
120000
sim2[, n_days]

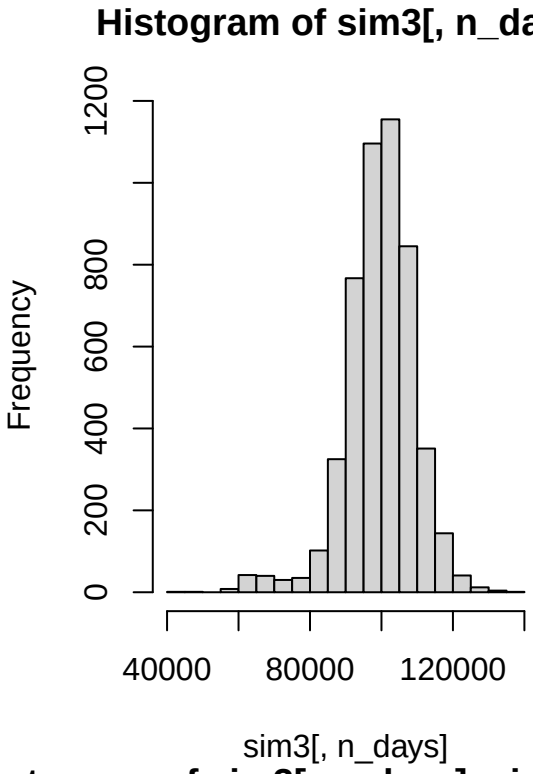

density.default(x = sim1[, n_days
Density
0e+00
4e−05
8e−05
80000
100000
120000
N = 5000 Bandwidth = 748.7

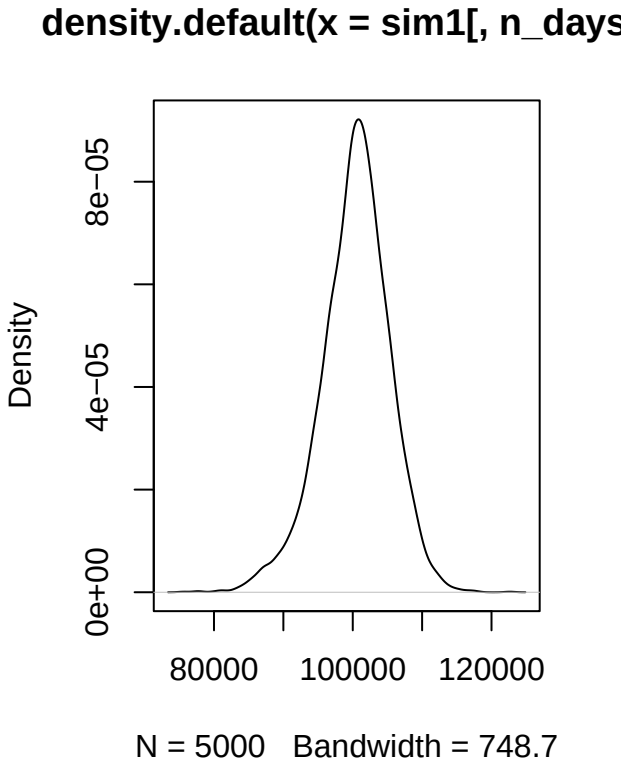

stogram of sim2[, n_days] − initial_
Frequency
0
100
200
300
400
500
−20000
0
10000
sim2[, n_days] − initial_wealth

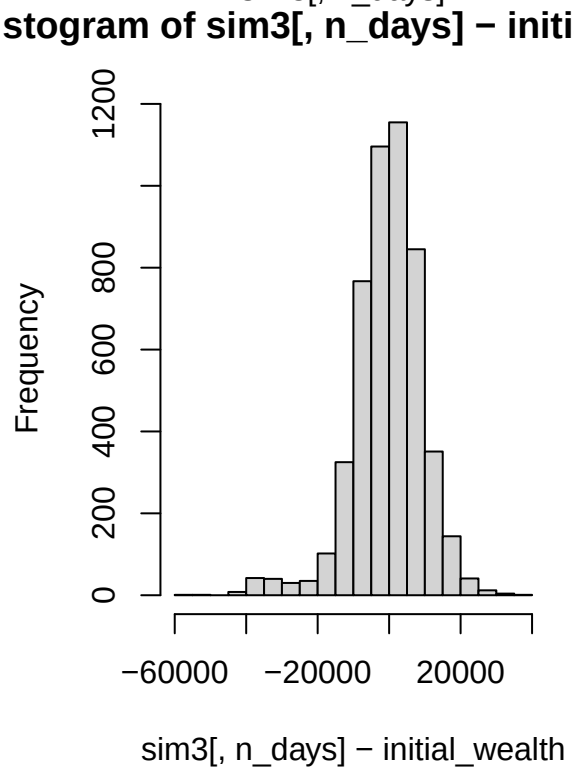


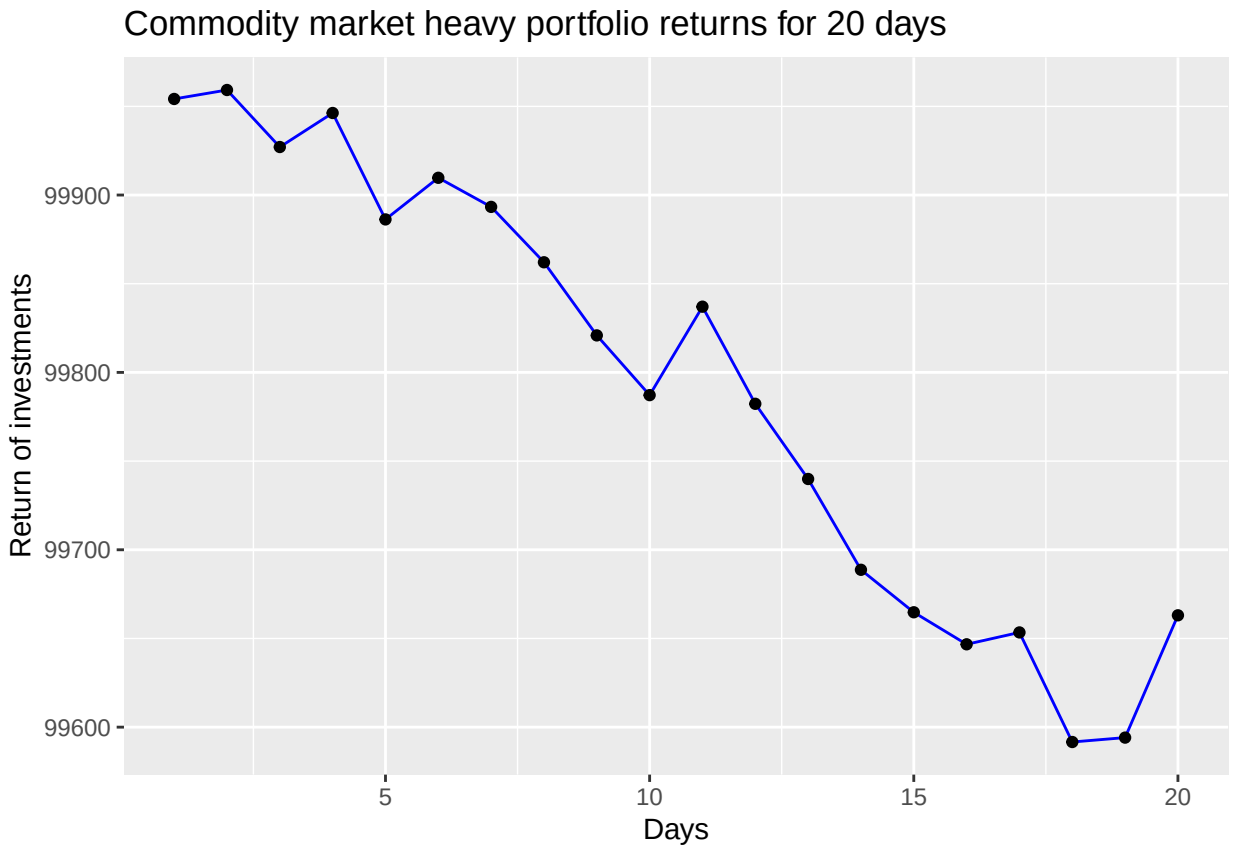
Portfolio 3 : More Bias in weights for Commodity Market ETFs: 80% weightage among UNG, USO 20% weightage among IBCE, IJH, DWAT, VOO, SDY Results: Initial wealth: 100000 Average Final Wealth over 20 days :100789.7 Average Profit:789.7232 Value at risk of 5% level: 6226.687

```
##              [,1]      [,2]      [,3]     [,4]      [,5]     [,6]      [,7]
## result.1  99102.94  99556.33 107904.11 112294.5 111068.05 110575.8 111577.38
## result.2 100201.54  99003.37  98704.87  98015.8  99460.02 101325.4  99175.01
## result.3 101044.72 103605.83 103476.83 104616.1 103861.51 103366.3 101080.48
## result.4 104918.67 106194.59 108751.86 108306.0 107766.63 107578.1 107065.19
## result.5 100046.30 103513.17 103393.28 104205.6 105005.17 105760.3 106078.83
## result.6 100742.27 101671.71 102564.16 102990.3 108271.87 103191.3 103976.04
##              [,8]      [,9]    [,10]    [,11]     [,12]    [,13]     [,14]
## result.1 109636.78 107726.32 106277.1 106335.3 106554.13 106318.4 105897.28
## result.2 100810.34  99166.09 102662.0 103220.9 104030.07 104218.9 104661.88
## result.3  99883.67 101369.25 102674.5 101904.4  99249.82 100124.1  95176.96
## result.4 108733.11 108218.04 107242.6 106944.8 106409.57 106317.8 106491.36
## result.5 106621.91 105048.49 104535.4 105903.5 107452.56 105861.0 107372.06
## result.6 106430.76 105305.31 104792.7 104961.1 100344.64 100877.0  98880.73
##             [,15]     [,16]     [,17]     [,18]     [,19]     [,20]
## result.1 103927.95 103399.58 104004.71  98865.87 100004.51 102689.38
## result.2 105162.40 104028.24 104305.32 107979.52 110081.44 108765.13
## result.3  88249.00  88732.76  90115.09  90054.68  88683.69  89404.79
## result.4 106812.34 105596.54 102828.38 101751.09 100249.54  98914.05
## result.5 105139.54 106473.14 107004.77 106383.23 106800.98 107688.00
## result.6  97049.04  98229.99 100244.57 101220.73 100490.67 100073.15
```


Histogram of sim3[, n_days]
Frequency
0
200
400
600
800
1200
40000
80000
120000
sim3[, n_days]

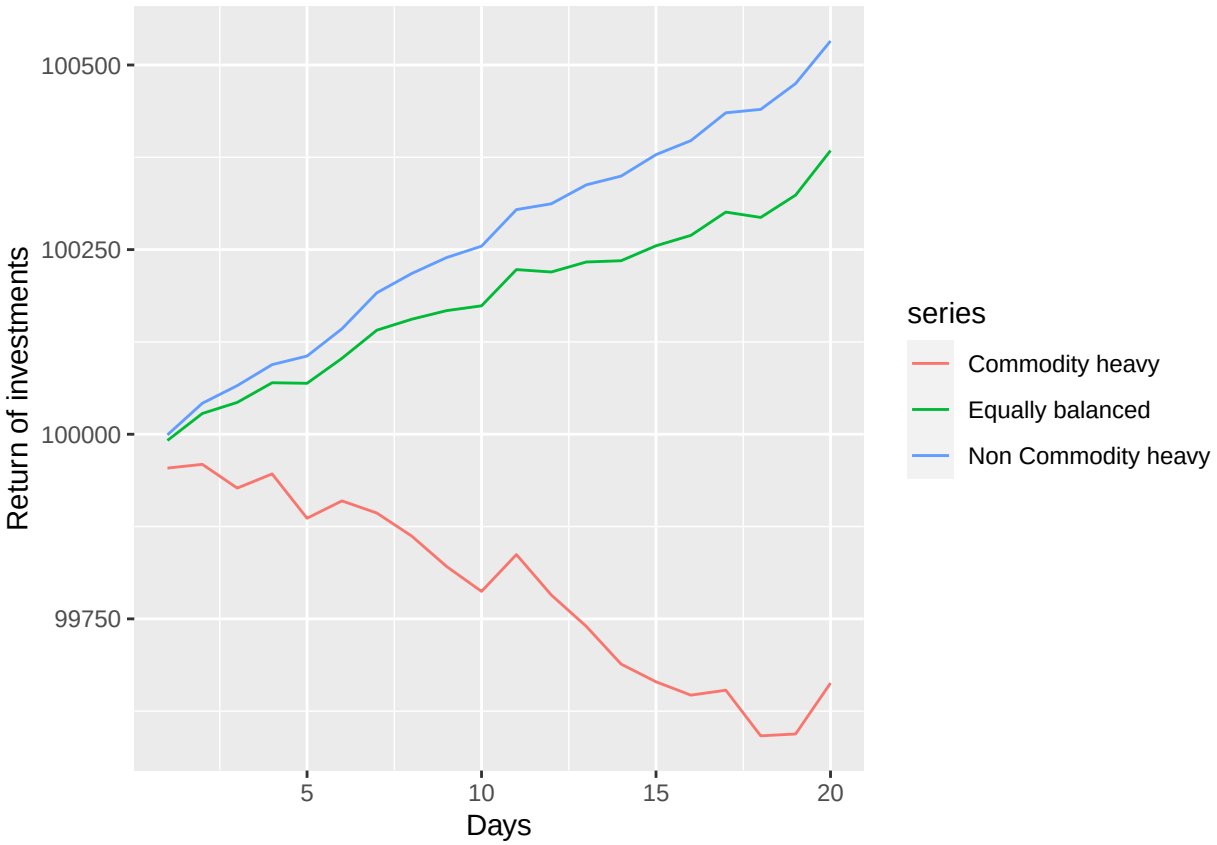

density.default(x = sim1[, n_days
Density
0e+00
4e−05
8e−05
80000
100000
120000
N = 5000 Bandwidth = 748.7


stogram of sim3[, n_days] − initial_
Frequency
0
200
400
600
800
1200
−60000
−20000
20000
sim3[, n_days] − initial_wealth



Summarizing all three portfolios:

```
##    days wealthbydaysim3 wealthbydaysim1 wealthbydaysim2
## 1     1        99954.17         99991.57         99999.11
## 2     2        99959.22        100028.13        100041.94
## 3     3        99927.07        100043.09        100065.82
## 4     4        99946.26        100069.72        100094.26
## 5     5        99886.25        100068.98        100105.87
## 6     6        99909.73        100102.78        100142.78
## 7     7        99893.32        100140.88        100191.49
## 8     8        99862.09        100155.71        100217.66
## 9     9        99820.87        100167.40        100239.24
## 10   10        99787.18        100173.87        100254.63
## 11   11        99837.06        100222.95        100304.15
## 12   12        99782.30        100219.68        100312.05
## 13   13        99739.91        100233.23        100337.65
## 14   14        99688.72        100234.93        100349.52
## 15   15        99664.77        100255.17        100378.63
## 16   16        99646.70        100269.29        100397.59
## 17   17        99653.41        100300.79        100435.23
## 18   18        99591.62        100293.55        100439.87
## 19   19        99594.11        100323.78        100474.85
## 20   20        99663.06        100384.09        100532.45
```




```
##        Portfolio_Type Mean_Wealth Profit_Loss Value_at_Risk
## 1    Equally Balanced   100384.09    384.0854     -8209.827
## 2 Non Commodity Heavy   100532.45    532.4550     -6869.723
## 3     Commodity Heavy    99663.06   -336.9430    -15421.300
```

Insights and Conclusions:

The above graph and the summary statistics table show the wealth value over 20 days and the variables calculated for these three portfolios. Considering the past performance of these stocks for 5 years,we have predicted the mean value, profit / loss and their value at risk at 5 %. The findings are as expected. The numbers are in line with what can be expected in a real-life scenario.

Firstly, the equally balanced portfolio gives a profit of 385 over 20 days which is not that high, but ensures a steady growth of wealth.The amount which we can expect to lose 5 percent of the time is 8209, which is less than 10 percent of the total wealth at the start. So if an investor is willing to take that level of risk, he/she can take up the equally balanced portfolio because based on the past 5 year performance, he is most likely to gain a reasonable amount within the term of analysis.

Furthermore, the portfolio with Non-commodity heavy ETFs gives a profit of 532 which is the highest of the three portfolios as expected because this portfolio lays more bias on the more diverse set of ETFs which also in turn reduces the VAR down to 6869, which is close to 7%. So based on our analysis, this portfolio would be the best one to take up, giving more returns accompanied with a lower risk level as well.

On the contrary, with the portfolio with more bias towards commodity based ETFs, we can see a decline in wealth in contrast to the growth observed in the first two portfolios. This is as expected for two reasons: Firstly, as the commodity market is more volatile, the investment can go down over a short term as there are a large number of variables affecting the ETF prices. Secondly, as this portfolio is more biased towards one type of ETFs. So, if the commodity market goes down, the other ETFs are not able to make up for the loss experienced in the commodity market ETFs because of having low weightage. Hence, the value at risk is also

high for this portfolio (close to 15.5%). So, clearly, it would not be advisable to take up a portfolio 3 or a similar portfolio like the third one which would be biased towards one particular market ETFs, especially in volatile markets like the commodities. As seen in the real life and our analysis as well, equally balanced and non commodity market bias portfolios perform better.

Question 6: Clustering and PCA

Clustering and PCA

Reading the wine.csv and retaining only first 11 variables for analysis, i.e. clustering and applying PCA.

Scale data and extract centre and scale from it to be effective to calculate distances whiile clustering

To run the clustering, we selected K means clustering algorithm starting with K = 2, one for type of wine to be clustered and nstarts = 10 to attempt with 30 different cluster centers in order to select the one with the least variance.

Assign the cluster determined by kmeans to individual rows and plotting with respect to x=fixed.acidity, y=citric.acid to see if clusters formed:

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality color cluster
## 1       5   red       1
## 2       5   red       1
## 3       5   red       1
## 4       6   red       1
## 5       5   red       1
## 6       5   red       1
```

Get average values of all 11 characteristics for the two types of wines. i.e. red and white:

```
## # A tibble: 2 x 14
##   color fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
##   <chr>         <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
## 1 red            8.32            0.528       0.271           2.54    0.0875
## 2 white          6.85            0.278       0.334           6.39    0.0458
## # ... with 8 more variables: free.sulfur.dioxide <dbl>,
## #   total.sulfur.dioxide <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,
## #   alcohol <dbl>, quality <dbl>, cluster <dbl>
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##            8.2895922            0.5319416            0.2695435
##       residual.sugar            chlorides  free.sulfur.dioxide
##            2.6342666            0.0883238           15.7647596
## total.sulfur.dioxide              density                   pH
##           48.6396835            0.9967404            3.3097200
##            sulphates              alcohol
##            0.6567194           10.4015216
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##           6.85167903           0.27458385           0.33524928
##       residual.sugar            chlorides  free.sulfur.dioxide
##           6.39402555           0.04510424           35.52152864
## total.sulfur.dioxide              density                   pH
##         138.45848785           0.99400486           3.18762464
##            sulphates              alcohol
##           0.48880511          10.52235888
```

```
##        cluster
## color    red white
##   red   1575    24
##   white   68  4830
```

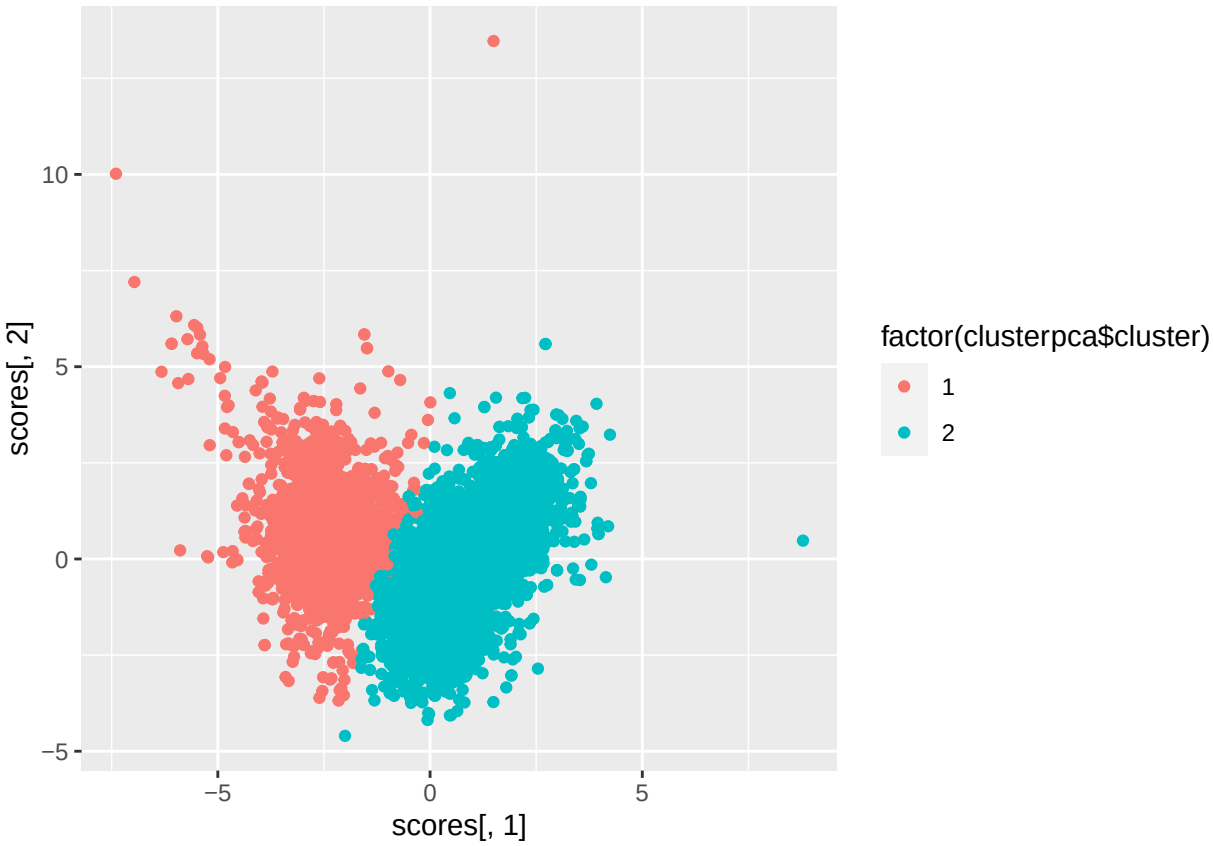
```
## [1] 0.9858396
```

As we can see in the classification matrix, we are able to cluster 98.5% of the data correctly using Kmeans clustering. Next we apply PCA as a dimensionality reduction method for this dataset:

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.90009
##                            PC8     PC9   PC10    PC11
## Standard deviation     0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.94568 0.97632 0.9970 1.00000
```

```
##                           PC1      PC2      PC3      PC4      PC5
## fixed.acidity        -0.23880  0.33635 -0.43430  0.16435 -0.14748
## volatile.acidity     -0.38076  0.11755  0.30726  0.21278  0.15146
## citric.acid           0.15239  0.18330 -0.59057 -0.26430 -0.15535
## residual.sugar        0.34592  0.32991  0.16469  0.16744 -0.35336
## chlorides            -0.29011  0.31526  0.01668 -0.24474  0.61439
## free.sulfur.dioxide   0.43091  0.07193  0.13422 -0.35728  0.22353
## total.sulfur.dioxide  0.48742  0.08727  0.10746 -0.20842  0.15813
## density              -0.04494  0.58404  0.17561  0.07272 -0.30656
## pH                   -0.21869 -0.15587  0.45532 -0.41455 -0.45338
## sulphates            -0.29414  0.19172 -0.07004 -0.64054 -0.13658
## alcohol              -0.10644 -0.46506 -0.26110 -0.10680 -0.18889
```

```
##        PC1        PC2        PC3        PC4        PC5
## -1.6602148  1.1646796  0.3630259 -1.0116941 -0.8602153
```




```
##        cluster
## color    red white
##   red   1574    25
##   white   78  4820
```

```
## [1] 0.9841465
```

As we can see above, the classification ratio in the class matrix is 98.4 i,e, close to 98.5 which was obtained during Kmeans algorithm. Hence we conclude that dimensionality reduction in this scenario is performing as well as kmeans clustering and is giving almost the same results with less number of predictors i.e. PCA components.

```
## [1] 5 6 7 4 8 3 9
```

```
##             kmeansclusterforquality$cluster
## wine$quality   1   2   3   4   5   6   7
##            3   7   7   1   2   4   5   4
##            4  61  24   2  29  15  64  21
##            5 470 656  20 298 200 414  80
##            6 347 645   9 503 266 538 528
##            7  42 122   1 179 140 144 451
##            8   2  21   0  30  14  30  96
##            9   0   1   0   0   0   0   4
```

```
## [1] 0.1120517
```

As we can see in the above table, the unique values for quality of wine in our data set is 3 to 9, which are 7 different values. So if we use a 7 class Kmeans clustering, the accuracy comes out to be only 11 percent. Hence, kmeans clustering does not seem to perform very well on determining quality of the wine. This possibly could be attributed to the fact that ratings given by the panel to that particular wine are purely

based on its taste and not its chemical compositions proportions. For eg. 2 wines with contrasting proportions might end up having a taste which the panel rated as 9. Hence, the clustering algorithm performs poorly on determining the quality of the wine.

Question 7: Market Segmentation

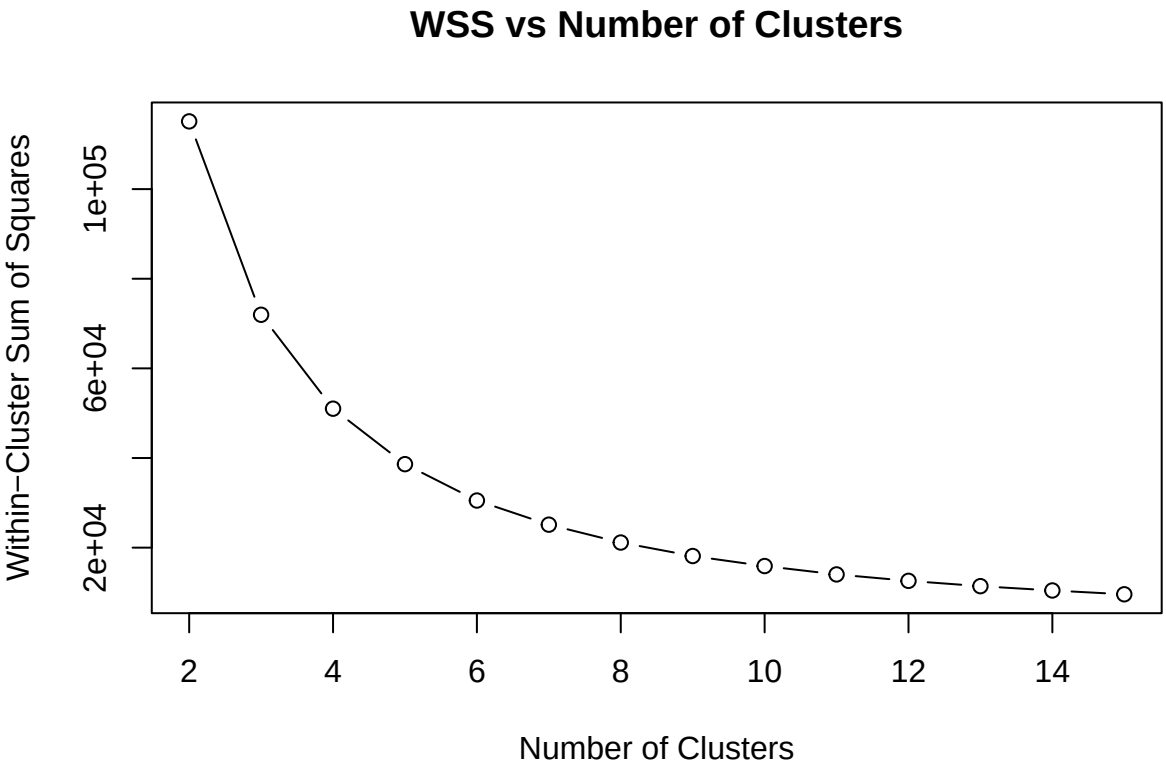
We want to understand the social media audience of a large consumer brand NutrientH20 so that they can tailor their messaging to their liking. The data on the audience includes annotated tweets of NutrientH20's Twitter followers bucketed into 36 unique categories over a seven-day period collected by a market-research study.

Clustering using K-Means++

Kmeans++ clustering achieves our objective by chunking similar audience in clusters which can be given personas to identify the broad meaning of each cluster and form market segments. We chose Kmeans++ over Kmeans since the latter chooses intial points for cluster centroids randomly.

Before clustering our data, we removed the following variables, as they would not provide any beneficial insight to our problem and: chatter, uncategorized, adult, and spam.

(a)



Choosing the Number of Clusters

Kmeans requires choosing the number of clusters beforehand, and it is difficult to speculate on the optimum number of clusters. We use the Elbow method that visualizes the within-cluster sum of squares (WSS) distance against the number of clusters. Looking at the plot, we choose the clusters so that adding more clusters does not add much improvement to the WSS.

Looking at the elbow plot above, we'll use 6 clusters for our analysis.

Cluster 1

	x
sports_fandom	5.88
religion	5.24
food	4.56
parenting	4.05
school	2.70
photo_sharing	2.63

```
## Percentage of total followers:  9.731033
```

Parents: With interests such as sports fandom, religion, food, parenting, and school, this segment represents the parents archetype. 9% of their

Cluster 2

	x
college_uni	10.70
online_gaming	9.95
photo_sharing	2.82
sports_playing	2.66
health_nutrition	1.78
travel	1.58

```
## Percentage of total followers:  5.239787
```

College-goers: With college & online gaming the predominant categories, this cluster represents the college-goers archetype. Tweets about college dominates this segment so the brand can possibly run campaigns in their universities to increase their outreach.

Cluster 3

	x
photo_sharing	2.29
shopping	1.28
health_nutrition	1.11
travel	1.10
politics	1.00
sports_fandom	0.97

```
## Percentage of total followers:  57.75184
```

Influencers: With health photo-sharing, shopping, health & fitness as the predominant categories, this cluster represents the Influencers archetype. They form the majority of the brand's follower signifying that the brand has many ambassadors.

Cluster 4

	x
health_nutrition	12.03
personal_fitness	6.45

	x
cooking	3.28
outdoors	2.76
photo_sharing	2.68
food	2.13

```
## Percentage of total followers:  11.17737
```

Health Conscious: With health & nutrition, fitness and cooking as the predominant categories, this cluster represents the Health Conscious archetype. They are careful about the products they use and possibly espouse them on social media. The brand can focus on increasing followers belonging to this cluster as they can provide good word-of-mouth publicity for their products. Currently, this cluster contains second-highest number of followers, after the Influencer cluster.

Cluster 5

	x
politics	8.90
travel	5.57
news	5.28
photo_sharing	2.54
computers	2.48
automotive	2.32

```
## Percentage of total followers:  8.779498
```

Working class: With politics, travel, and news as the predominant categories, this cluster represents the Working Class archetype. They travel frequently, either for work or for vacations. Either ways, this group tends to have high spending capacity and the brand can market high margin products to this group of followers.

Cluster 6

	x
cooking	10.84
photo_sharing	6.09
fashion	5.55
beauty	3.87
health_nutrition	2.28
shopping	2.04

```
## Percentage of total followers:  7.320477
```

Chefs: With food and photo-sharing as the predominant categories, this cluster represents the Chef archetype. They are fond of sharing recipes and cooking tips and possibly share a lot of food pictures. The NutrientH20 brand can market their consumer packaged goods product line to appeal to this group. They are also likely to retweet photos that the brand tweets about.

Market Segments

1) **Parents**
2) **College-goers**
3) **Influencers**

4) **Health Conscious**
5) **Working Class**
6) **Chefs**

NutrientH20 can run various tailored marketing campaigns on twitter according to these market segments to ensure their followers relate to the content and drive more followers in each segment!

Question 8: The Reuters corpus

Question: We will try to find the author of a given article based on the style of vocabulary used in the article. Further, we want to analyze which authors are most similar by looking at the probabilities of our classification model.

Approach: We break down our analysis into below sequential tasks -

1. Read in the raw data and pre-process it to make it easier for classification model to extract signal (consistent casing, stop word removal, etc.)
2. Create target column out of file names
3. Extract numerical features from text data by creating a TF-IDF matrix
4. Perform PCA dimensionality reduction to compress the features and wash out white noise
5. Develop classification models from the resulting features and analyze model parameters and output for interesting trends.

Details of the approach given below -

Pre-processing -

- Includes stop words removal, tokenization and consistency of casing, special characters removal, and space-like character consistency.

Text representation -

- We have created a document-term matrix (DTM) where the words in the documents are the columns and each document is a row. Standardization of the DTM is done using TF-IDF, meaning that each word is represented as a TF-IDF value.

Handling out-of-vocabulary (OOV) words -

- OOV words are the ones which are present in test set but not in the tran set, so our models cannot deal with them. Since there are many such words in the dataset, we decided to remove all OOV words from the test dataset and only keep the ones common in both the datasets. This means that our model needs to be trained periodically if it is deployed in production.

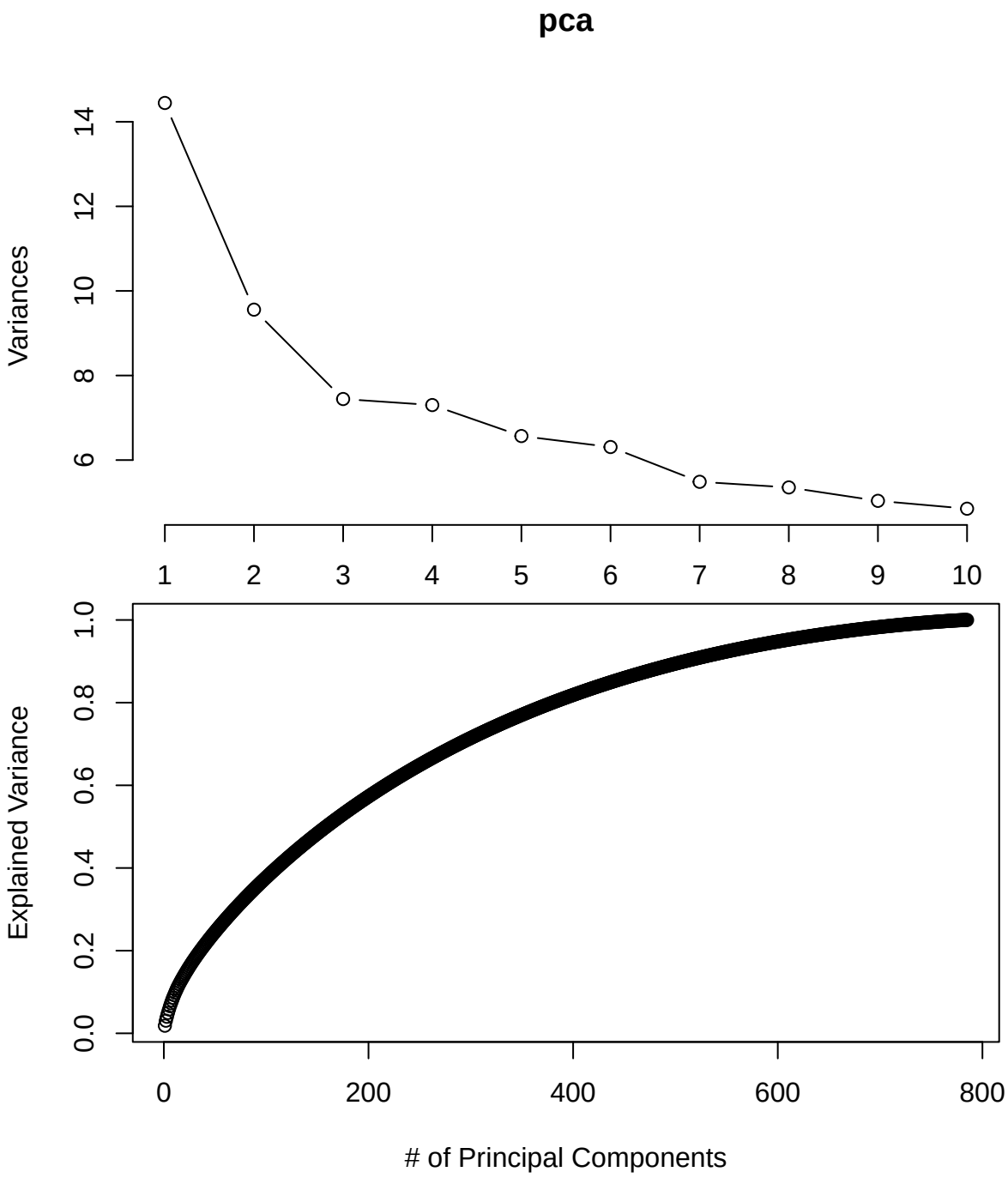
Dimensionality reduction -

- The resulting document-term-matrix (DTM) contains huge number of variables/words which can make the model complex and slow to train. We reduce the dimension of the DTM using Principal Component Analysis. Since there are hundreds of words (variables) in the training and test dataset, it is better to reduce the number of variables using principal component analysis.

Result:

Principal Component Analysis

The PCA shows that 400 principal components explain almost 80% of the variance. So we reduce our original matrix of 314k features to 400 features, giving a compression ratio of 99%.



Multi-label Classification models

KNN

As a baseline, we perform KNN classification with K=10 from cross-validation, achieving 35.72%. The score is on the lower end considering we have a multi-label classification problem, i.e., the number of authors are huge.

```
## [1] 886
```

```
## [1] 35.44
```

Naive Bayes

Naive Bayes improves the accuracy to 46.84%. The independent features assumption seems to fit well for this dataset.

```
##           Length Class  Mode
## apriori    50    table  numeric
## tables    400    -none- list
## levels     50    -none- character
## isnumeric 400    -none- logical
## call        4    -none- call
```

```
## [1] 46.92
```

Random Forest

We obtained best performance using the Random Forest classifier where the accuracy rose to 70.6%. The runtime of the model is slightly more due to 400 feature set. However, the accuracy more than compensates for the high model training time.

```
## [1] 69.96
```

Conclusion Author attribution on this dataset is a fairly complex task since its difficult for most ML models to generate signal. This is evident from the low accuracy scores of majority of the classifiers. Ultimately, the Random Forest Classifier reigns supreme with a ~70% accuracy over the test set.

For next steps, we recommend using a word embedding such as GloVe to further enrich text representation and boost accuracy.

We recommend using the 3 classification models are made and the most accurate model was found to be Random Forest with an accuracy of 70.6%.

Question 9: Association Rule Mining:

```
## Fraction of people who answered yes given that they are truthful speakers: 0.7142857
```

```
##
## Fraction of people who answered yes and are truthful speakers: 0.5
```

```
## Probability of having the disease given they test positive: 0.1988824
```

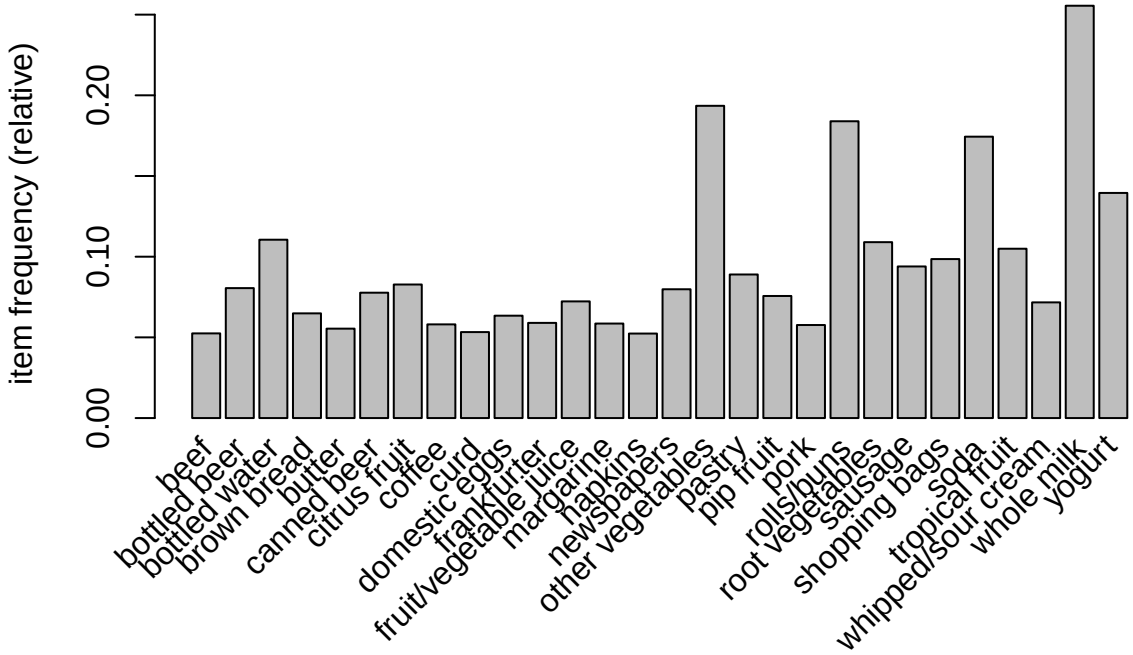
```
## [1] 9835  169
```

```
##     items
## [1] {citrus fruit,
##      margarine,
##      ready soups,
##      semi-finished bread}
## [2] {coffee,
##      tropical fruit,
##      yogurt}
## [3] {whole milk}
## [4] {cream cheese,
##      meat spreads,
##      pip fruit,
##      yogurt}
## [5] {condensed milk,
##      long life bakery product,
##      other vegetables,
```

```
##      whole milk}
```

Getting and plotting Item Frequency:

```
## abrasive cleaner artif. sweetener   baby cosmetics        baby food
##     0.0035587189     0.0032536858     0.0006100661     0.0001016777
##             bags    baking powder bathroom cleaner             beef
##     0.0004067107     0.0176919166     0.0027452974     0.0524656838
##          berries        beverages     bottled beer    bottled water
##     0.0332486019     0.0260294865     0.0805287239     0.1105236401
##           brandy      brown bread           butter
##     0.0041687850     0.0648703610     0.0554143366
```



Relative itemFrequency is highest for whole milk.

Apriori

Now we run the Apriori algorithm.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.2    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 ...[169 item(s), 9835 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 ... [873 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

Inspecting the first 10 rules:

```
##      lhs              rhs                support     confidence coverage
## [1]  {}            => {whole milk}       0.255516014 0.2555160  1.00000000
## [2]  {cake bar}    => {whole milk}       0.005592272 0.4230769  0.01321810
## [3]  {dishes}      => {other vegetables} 0.005998983 0.3410405  0.01759024
## [4]  {dishes}      => {whole milk}       0.005287239 0.3005780  0.01759024
## [5]  {mustard}     => {whole milk}       0.005185562 0.4322034  0.01199797
## [6]  {pot plants}  => {whole milk}       0.006914082 0.4000000  0.01728521
## [7]  {chewing gum} => {soda}             0.005388917 0.2560386  0.02104728
## [8]  {chewing gum} => {whole milk}       0.005083884 0.2415459  0.02104728
## [9]  {canned fish} => {other vegetables} 0.005083884 0.3378378  0.01504830
## [10] {pasta}       => {whole milk}       0.006100661 0.4054054  0.01504830
##      lift      count
## [1]  1.0000000 2513
## [2]  1.6557746   55
## [3]  1.7625502   59
## [4]  1.1763569   52
## [5]  1.6914924   51
## [6]  1.5654596   68
## [7]  1.4683033   53
## [8]  0.9453259   50
## [9]  1.7459985   50
## [10] 1.5866145   60
```

```
## Inspecting grocery rules with lift>4
```

```
##     lhs                    rhs                      support confidence   coverage     lift count
## [1] {butter,
##      other vegetables}  => {whipped/sour cream} 0.005795628  0.2893401 0.02003050 4.036397    57
## [2] {citrus fruit,
##      other vegetables,
##      whole milk}        => {root vegetables}    0.005795628  0.4453125 0.01301474 4.085493    57
```

```
## Inspecting top 10 grocery rules with confidence>0.3
```

```
##      lhs              rhs                support     confidence coverage
## [1]  {cake bar}    => {whole milk}       0.005592272 0.4230769  0.01321810
## [2]  {dishes}      => {other vegetables} 0.005998983 0.3410405  0.01759024
## [3]  {dishes}      => {whole milk}       0.005287239 0.3005780  0.01759024
## [4]  {mustard}     => {whole milk}       0.005185562 0.4322034  0.01199797
## [5]  {pot plants}  => {whole milk}       0.006914082 0.4000000  0.01728521
## [6]  {canned fish} => {other vegetables} 0.005083884 0.3378378  0.01504830
## [7]  {pasta}       => {whole milk}       0.006100661 0.4054054  0.01504830
## [8]  {herbs}       => {root vegetables}  0.007015760 0.4312500  0.01626843
## [9]  {herbs}       => {other vegetables} 0.007727504 0.4750000  0.01626843
## [10] {herbs}       => {whole milk}       0.007727504 0.4750000  0.01626843
##      lift     count
## [1]  1.655775 55
## [2]  1.762550 59
## [3]  1.176357 52
```

```
## [4]  1.691492 51
## [5]  1.565460 68
## [6]  1.745998 50
## [7]  1.586614 60
## [8]  3.956477 69
## [9]  2.454874 76
## [10] 1.858983 76
```

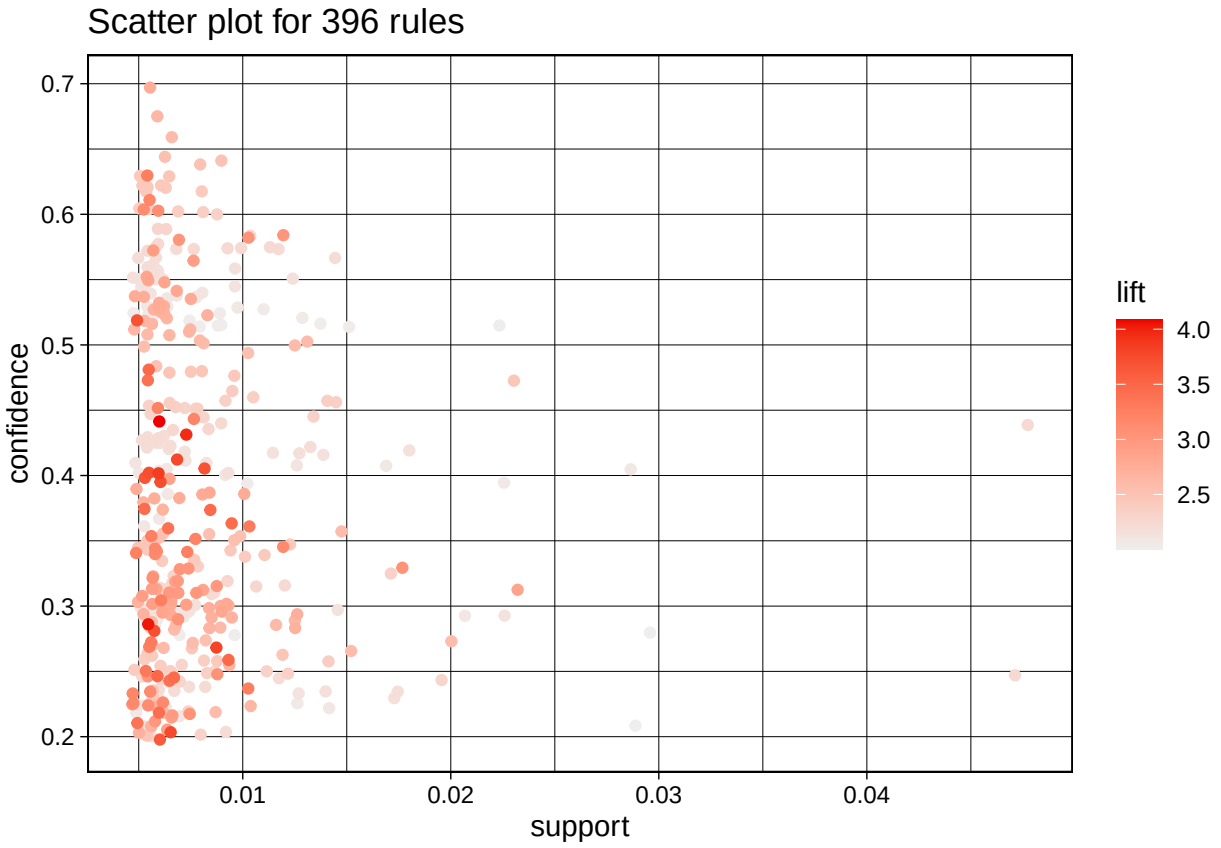
```
## Inspecting grocery rules with confidence>0.2 and lift>2
```

```
## set of 396 rules
```

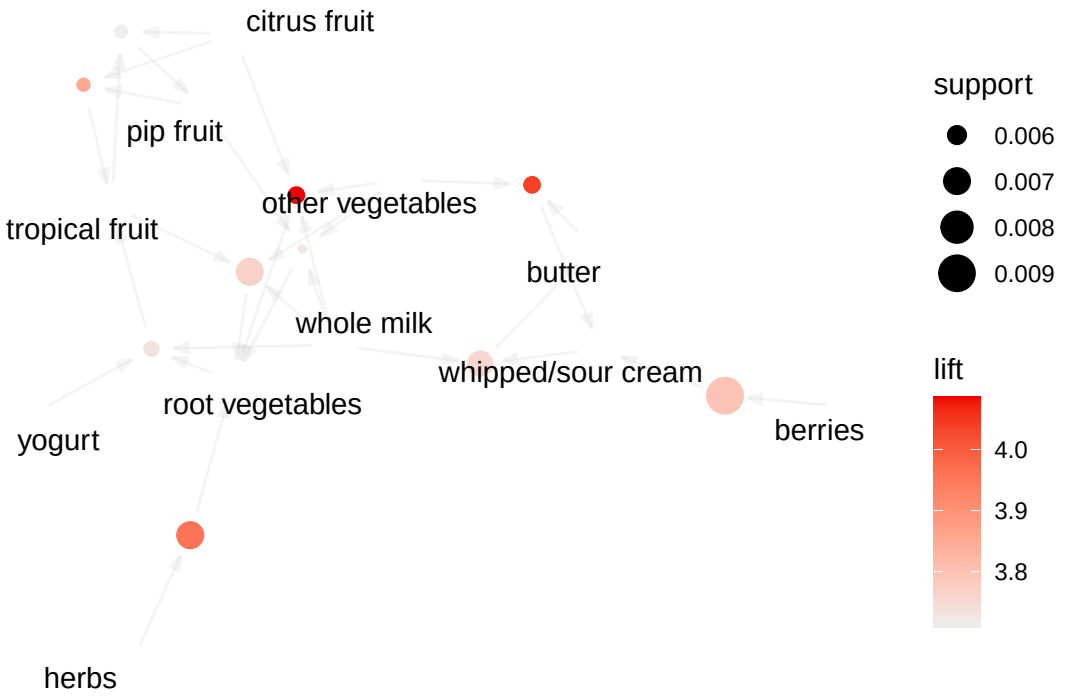
```
## Inspecting top 10 grocery rules sorted by lift
```

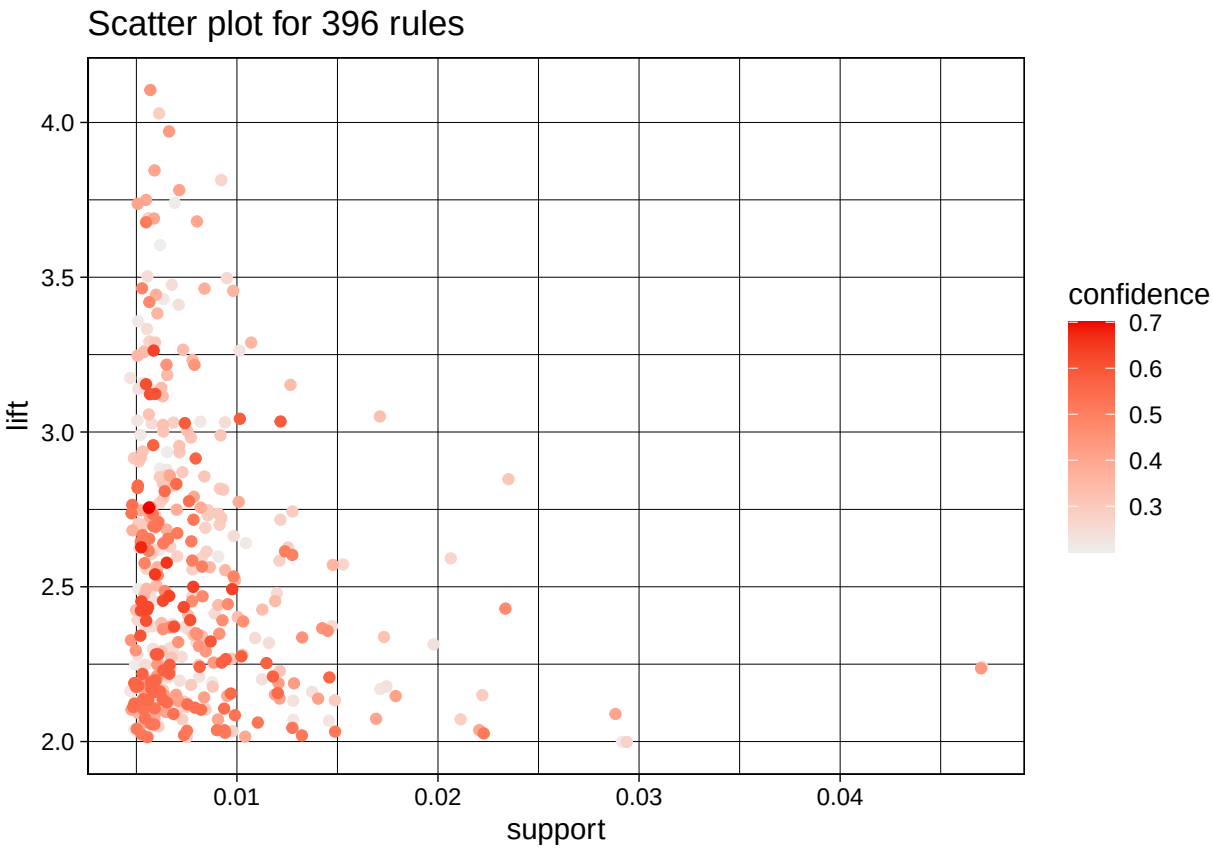
```
##      lhs                      rhs                      support confidence   coverage     lift count
## [1]  {citrus fruit,
##       other vegetables,
##       whole milk}          => {root vegetables}    0.005795628  0.4453125 0.01301474 4.085493    57
## [2]  {butter,
##       other vegetables}    => {whipped/sour cream} 0.005795628  0.2893401 0.02003050 4.036397    57
## [3]  {herbs}              => {root vegetables}    0.007015760  0.4312500 0.01626843 3.956477    69
## [4]  {citrus fruit,
##       pip fruit}           => {tropical fruit}     0.005592272  0.4044118 0.01382816 3.854060    55
## [5]  {berries}            => {whipped/sour cream} 0.009049314  0.2721713 0.03324860 3.796886    89
## [6]  {other vegetables,
##       tropical fruit,
##       whole milk}          => {root vegetables}    0.007015760  0.4107143 0.01708185 3.768074    69
## [7]  {whipped/sour cream,
##       whole milk}          => {butter}             0.006710727  0.2082019 0.03223183 3.757185    66
## [8]  {root vegetables,
##       whole milk,
##       yogurt}              => {tropical fruit}     0.005693950  0.3916084 0.01453991 3.732043    56
## [9]  {other vegetables,
##       pip fruit,
##       whole milk}          => {root vegetables}    0.005490595  0.4060150 0.01352313 3.724961    54
## [10] {citrus fruit,
##       tropical fruit}      => {pip fruit}          0.005592272  0.2806122 0.01992883 3.709437    55
```

Visualizing Apriori




```
## Graph Plot for grocery rules
```






```
## Inspecting top 10 grocery rules with support > 0.01

##      lhs                    rhs                support    confidence coverage
## [1]  {onions}            => {other vegetables} 0.01423488 0.4590164  0.03101169
## [2]  {berries}           => {yogurt}           0.01057448 0.3180428  0.03324860
## [3]  {hamburger meat}    => {other vegetables} 0.01382816 0.4159021  0.03324860
## [4]  {cream cheese}      => {yogurt}           0.01240468 0.3128205  0.03965430
## [5]  {chicken}           => {root vegetables}  0.01087951 0.2535545  0.04290798
## [6]  {chicken}           => {other vegetables} 0.01789527 0.4170616  0.04290798
## [7]  {frozen vegetables} => {root vegetables}  0.01159126 0.2410148  0.04809354
## [8]  {beef}              => {root vegetables}  0.01738688 0.3313953  0.05246568
## [9]  {curd}              => {yogurt}           0.01728521 0.3244275  0.05327911
## [10] {pork}              => {root vegetables}  0.01362481 0.2363316  0.05765125
##      lift     count
## [1]  2.372268 140
## [2]  2.279848 104
## [3]  2.149447 136
## [4]  2.242412 122
## [5]  2.326221 107
## [6]  2.155439 176
## [7]  2.211176 114
## [8]  3.040367 171
## [9]  2.325615 170
## [10] 2.168210 134

## Inspecting top 10 grocery rules  where rhs is yogurt sorted by lift

##      lhs                        rhs          support confidence   coverage     lift count
```

```
## [1]  {curd,
##       tropical fruit}         => {yogurt} 0.005287239  0.5148515 0.01026945 3.690645    52
## [2]  {fruit/vegetable juice,
##       other vegetables,
##       whole milk}             => {yogurt} 0.005083884  0.4854369 0.01047280 3.479790    50
## [3]  {root vegetables,
##       tropical fruit,
##       whole milk}             => {yogurt} 0.005693950  0.4745763 0.01199797 3.401937    56
## [4]  {tropical fruit,
##       whipped/sour cream}     => {yogurt} 0.006202339  0.4485294 0.01382816 3.215224    61
## [5]  {other vegetables,
##       tropical fruit,
##       whole milk}             => {yogurt} 0.007625826  0.4464286 0.01708185 3.200164    75
## [6]  {cream cheese,
##       whole milk}             => {yogurt} 0.006609049  0.4012346 0.01647178 2.876197    65
## [7]  {fruit/vegetable juice,
##       other vegetables}       => {yogurt} 0.008235892  0.3913043 0.02104728 2.805013    81
## [8]  {root vegetables,
##       tropical fruit}         => {yogurt} 0.008134215  0.3864734 0.02104728 2.770384    80
## [9]  {curd,
##       whole milk}             => {yogurt} 0.010066090  0.3852140 0.02613116 2.761356    99
## [10] {cream cheese,
##       other vegetables}       => {yogurt} 0.005287239  0.3851852 0.01372649 2.761149    52

## Inspecting top 10 grocery rules  where rhs is root vegetables sorted by lift

##      lhs                     rhs                   support confidence   coverage     lift count
## [1]  {citrus fruit,
##       other vegetables,
##       whole milk}          => {root vegetables} 0.005795628  0.4453125 0.01301474 4.085493    57
## [2]  {herbs}               => {root vegetables} 0.007015760  0.4312500 0.01626843 3.956477    69
## [3]  {other vegetables,
##       tropical fruit,
##       whole milk}          => {root vegetables} 0.007015760  0.4107143 0.01708185 3.768074    69
## [4]  {other vegetables,
##       pip fruit,
##       whole milk}          => {root vegetables} 0.005490595  0.4060150 0.01352313 3.724961    54
## [5]  {beef,
##       other vegetables}    => {root vegetables} 0.007930859  0.4020619 0.01972547 3.688692    78
## [6]  {onions,
##       other vegetables}    => {root vegetables} 0.005693950  0.4000000 0.01423488 3.669776    56
## [7]  {beef,
##       whole milk}          => {root vegetables} 0.008032537  0.3779904 0.02125064 3.467851    79
## [8]  {tropical fruit,
##       whole milk,
##       yogurt}              => {root vegetables} 0.005693950  0.3758389 0.01514997 3.448112    56
## [9]  {citrus fruit,
##       other vegetables}    => {root vegetables} 0.010371124  0.3591549 0.02887646 3.295045   102
## [10] {other vegetables,
##       whipped/sour cream,
##       whole milk}          => {root vegetables} 0.005185562  0.3541667 0.01464159 3.249281    51

## Inspecting top 10 grocery rules  where rhs is butter sorted by lift

##     lhs                                    rhs      support     confidence
## [1] {whipped/sour cream, whole milk}       => {butter} 0.006710727 0.2082019
```

```
## [2] {other vegetables, whipped/sour cream} => {butter} 0.005795628 0.2007042
##     coverage   lift     count
## [1] 0.03223183 3.757185 66
## [2] 0.02887646 3.621883 57
```

```
## Inspecting top 10 grocery rules  where rhs is sausage sorted by lift
```

```
##      lhs                                  rhs       support     confidence
## [1]  {rolls/buns, shopping bags}       => {sausage} 0.005998983 0.3072917
## [2]  {sliced cheese}                   => {sausage} 0.007015760 0.2863071
## [3]  {rolls/buns, soda}                => {sausage} 0.009659380 0.2519894
## [4]  {other vegetables, shopping bags} => {sausage} 0.005388917 0.2324561
## [5]  {shopping bags, soda}             => {sausage} 0.005693950 0.2314050
## [6]  {other vegetables, soda}          => {sausage} 0.007219115 0.2204969
## [7]  {hard cheese}                     => {sausage} 0.005185562 0.2116183
## [8]  {other vegetables, rolls/buns}    => {sausage} 0.008845958 0.2076372
## [9]  {bottled water, other vegetables} => {sausage} 0.005083884 0.2049180
## [10] {meat}                            => {sausage} 0.005287239 0.2047244
##      coverage   lift     count
## [1]  0.01952211 3.270794 59
## [2]  0.02450432 3.047435 69
## [3]  0.03833249 2.682160 95
## [4]  0.02318251 2.474249 53
## [5]  0.02460600 2.463060 56
## [6]  0.03274021 2.346956 71
## [7]  0.02450432 2.252452 51
## [8]  0.04260295 2.210078 87
## [9]  0.02480935 2.181135 50
## [10] 0.02582613 2.179074 52
```

We run the Apriori algorithm for multiple thresholds of lift and confidence. Selecting a value of lift above 4 gives us only two rules, while selecting only a confidence of >0.3 gives us 481 rules.

After multiple iterations, finally we select the values of lift and confidence as lift > 2 & confidence > 0.2 as this gave us sufficient number of rules with a reasonable value of confidence.

People who buy citrus fruits and root vegetables are more likely to buy tropical fruits, therefore it would be ideal to have them placed in the same corner.

Yogurt has a high lift value with fruit juices and vegetables. So the dairy section should ideally have the fruit section right next to it.

Even butter should be placed alongside the rest of the dairy products, as per the high lift value.

Root vegetables has a high lift value with other fruits and vegetables, so they can be placed together. This makes sense since most people shop for their vegetables and fruits together.

Sausage achieves a high lift with snacks like items like rolls/buns and sliced cheese and also with soda, so it makes sense to have a promotional discount clubbing the two items together. Somehow it has a high lift >2 with shopping bags.