STA380 Part-2

Prathyusha Vedulla(pv6269), Shubhada Kapre(sk55489), Vivek Dhulipalla(vd6543)

2022-08-15

Prathyusha Vedulla(pv6269), Shubhada Kapre(sk55489), Vivek Dhulipalla(vd6543) Github link - https://github.com/fudgemallow/STA380-Take-Home-Exam-2

Probability Practice

Part A

- Total probability of an yes $P(Y) = 0.65$
- Probability of an yes given the click is by a random clicker $P(Y/RC) = 0.5$
- Probability of a clicker being a random clicker $P(RC) = 0.3$
- Probability of a clicker being a truthful clicker $P(TC) = 0.7$

Let's consider the total probability equation of yes to get started with: Total probability $P(Y) =$ Joint probability of Yes and Random Clickers $P(Y, RC)$ +Joint probability of Yes and Random Clickers $P(Y, TC)$
i.e $P(Y) = P(Y, RC) + P(Y, TC)$
=> $P(Y) = P(Y/RC) * P(RC) + P(Y/TC) * P(TC)$
=> $P(Y/TC) * P(TC) = P(Y) - P(Y/RC) * P(RC)$
=> $P(Y/TC) = (P(Y) - P(Y/RC) * P(RC))/P(TC)$
Substituting the given values
$P(Y/TC) = (0.65 - (0.5*0.3))/0.7

Fraction of people who are truthful clickers answered yes = 5/7

Part B

- Probability of positive test result given the presence of disease $P(P/D) = 0.993$
- Probability of negative result given there is no disease $P(N/NoDis) = 0.9999$
- Total probability of having a disease $P(D) = 0.000025$

Confusion matrix	Actual	
Test Result	Disease	No Disease
Positive	24,825	99,997
Negative	175	999,875,003

- Using the above calculations, we can see that there is a 0.1988 probability of having a disease if the test result is positive

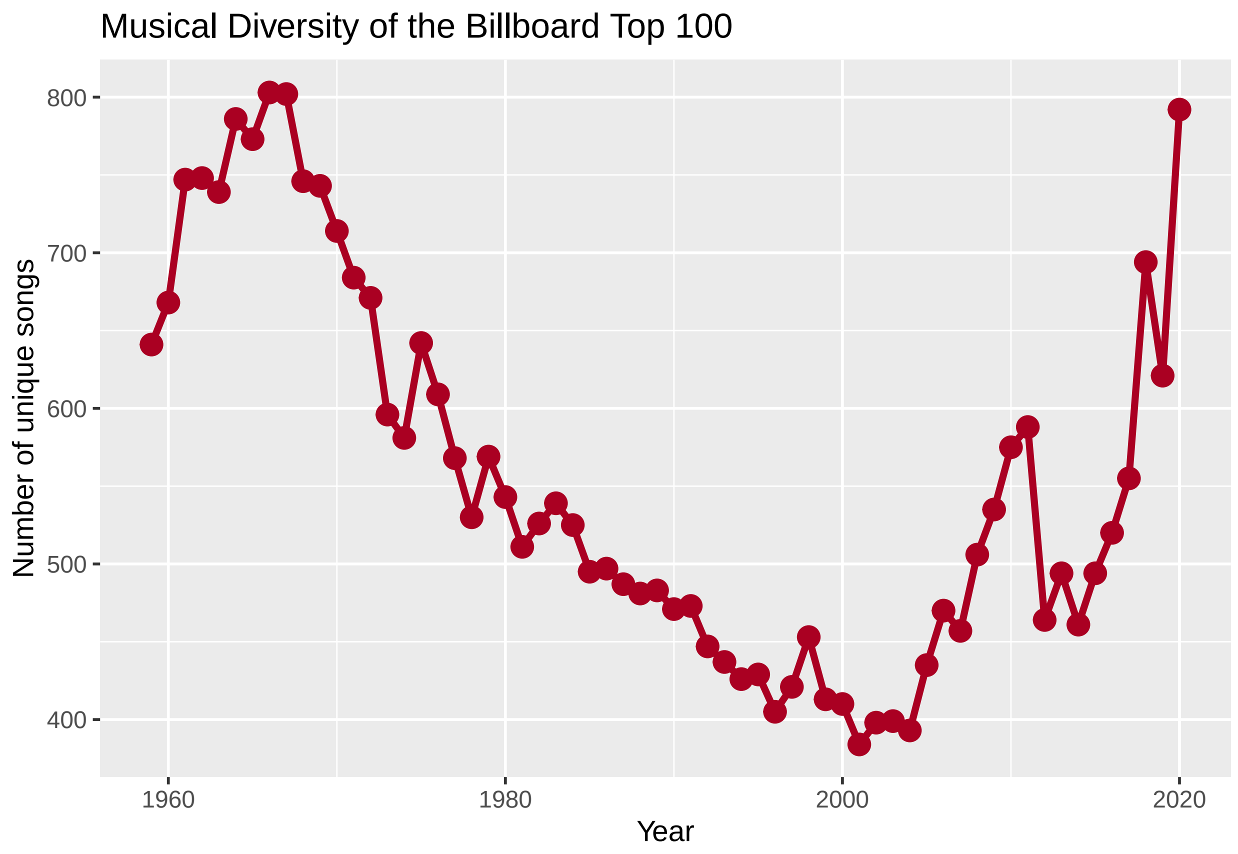
Wrangling the Billboard Top 100

Part A: Make a table of the top 10 most popular songs since 1958.

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

- The above table displays the top 10 most popular songs since 1958 based on the total number of weeks that a song spent on the Billboard Top 100.

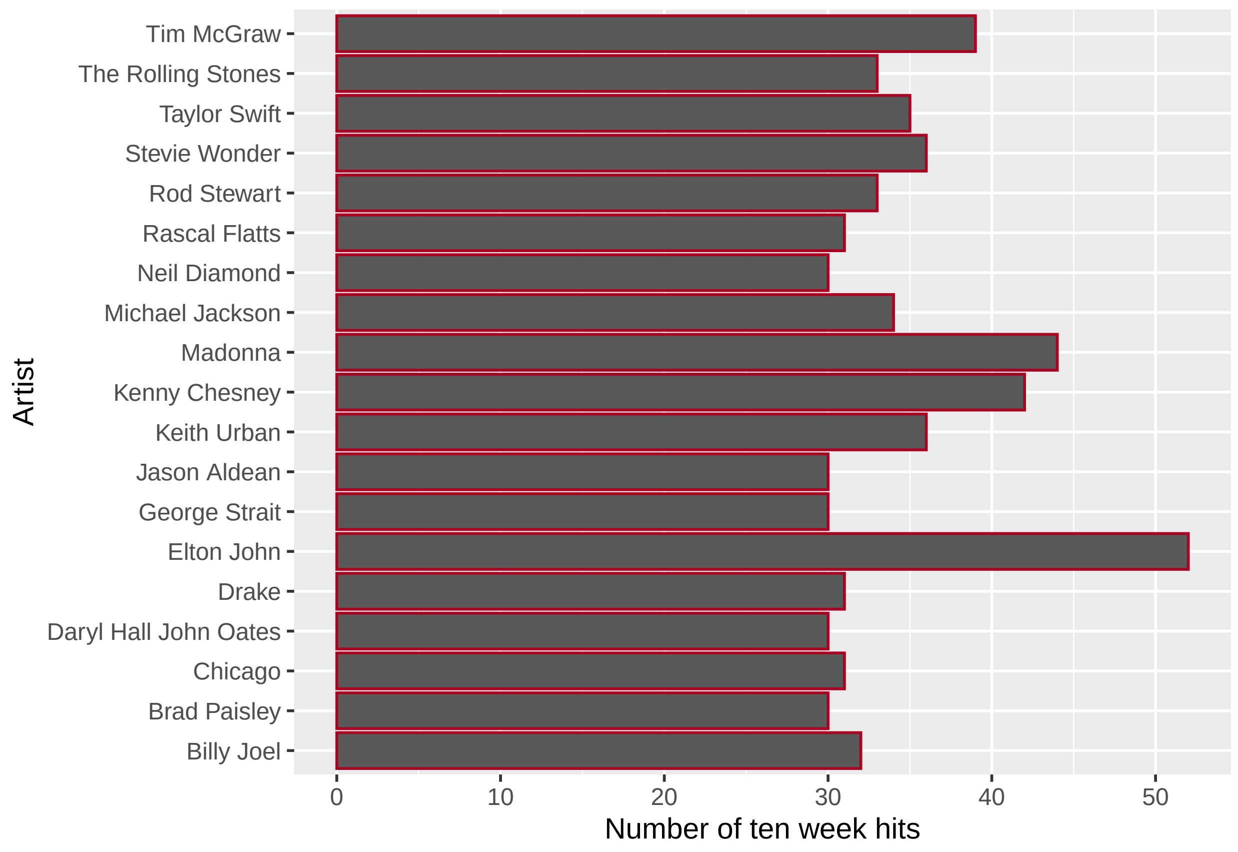
Part B: Is the "musical diversity" of the Billboard Top 100 changing over time?



- The above line graph shows the number of distinct songs in the Billboard top 100 in each year from 1959 to 2020. As we can see, the number of distinct songs in the years 1965-1970 were pretty high but consistently decreased after that until the year 2000. From the year 2000, the number of distinct songs consistently increased till the year 2020.

Part C: Ten-week hit

```
## 'summarise()' has grouped output by 'song'. You can override using the
## '.groups' argument.
```



- The bar plot shows the artists who have atleast 30 songs that appeared in the top 100 billboard list for more than 10 weeks. Elton John has is on the top of the list with 52 ten week hit songs followed by Madonna, Kenny Chesney, Tim McGraw and Keith Urban in the top 5

Visual story telling part 1: Green Buildings

```
## [1] 7894
```

```
## [1] 7679
```

- Green buildings in each cluster should be compared to non-green buildings in the same cluster. This will give us the advantage of a green building if any financially. I agree with the excel guru that buildings with less than 10% occupancy need to be removed. There are 215 such buildings

#similar green buildings

```
## [1] 39.542 39.542 39.542 39.542 39.542
```

```
##  [1] 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216
##  [9] 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216
## [17] 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216
## [25] 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216
## [33] 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216
## [41] 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216 26.05216
## [49] 26.05216 26.05216 26.05216
```

- Above table shows the average rent of green buildings and non-green which have 15 stories and are of size 250000. We see that the average rent of green buildings is higher.

```
## [1] 0.01026 0.01026 0.01026 0.01026 0.01026
```

```
##  [1] 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373
##  [8] 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373
## [15] 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373
## [22] 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373
## [29] 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373
## [36] 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373
## [43] 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373 0.0110373
## [50] 0.0110373 0.0110373
```

- Average Gas cost of green buildings is almost equal to that of non-green buildings

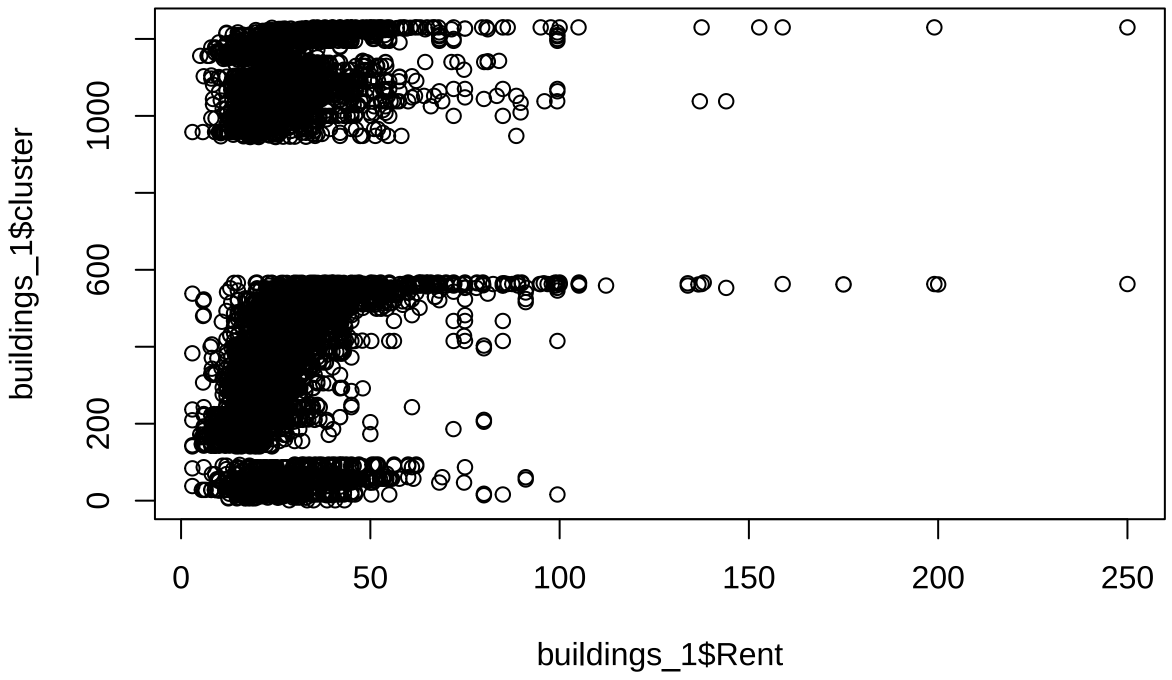
```
## [1] 0.03602 0.03602 0.03602 0.03602 0.03602
```

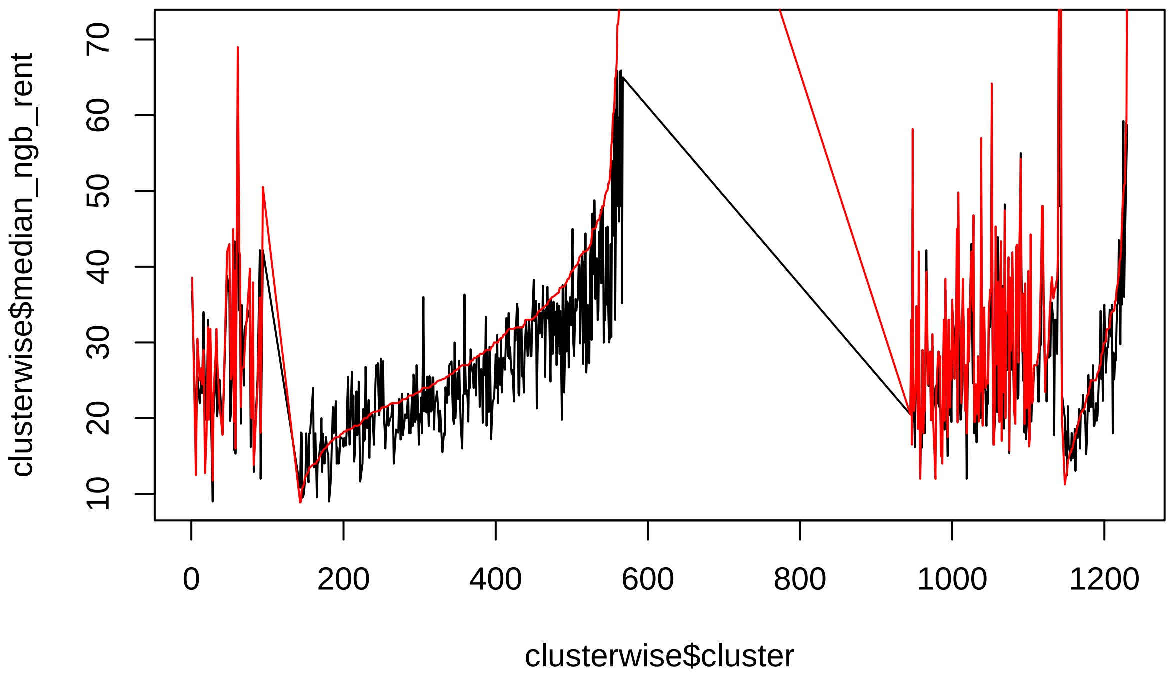
```
##  [1] 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945
##  [7] 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945
## [13] 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945
## [19] 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945
## [25] 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945
## [31] 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945
## [37] 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945
## [43] 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945 0.02803945
## [49] 0.02803945 0.02803945 0.02803945
```

- Average electricity cost of green buildings is higher
- The excel guru did not compare the electricity and gas costs which are a significant factor while constructing green buildings.
- From our analysis, we see that while the gas costs are comparable between green and non-green buildings, the electricity costs of green buildings are much higher. This is contrary to what we expected.
- I believe that the excel guru should not have calculated the median values over the entire dataset of green and non-green buildings.

*Instead she should have only chosen those buildings that had similar number of stories and square footage for comparison.

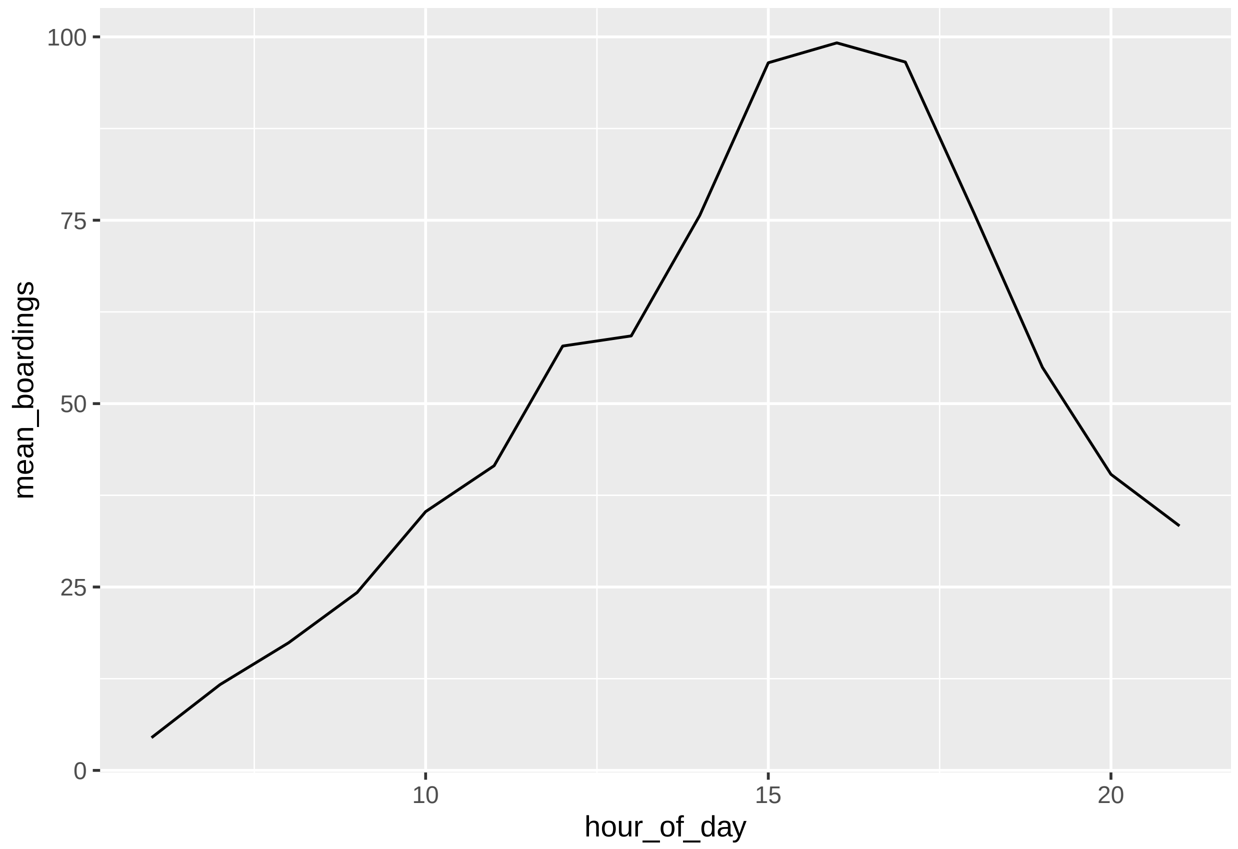
- After doing this, I found that the rent of green buildings seems to be around 18.75$more than non-green buildings. This translates to 18.75 * 250000 = 4,687,500 extra revenue per year.
- This means that we cover the extra 5% premium in a little over a year itself.
- Hence, it is a great financial move to build a green building.




clusterwise$median_ngb_rent
clusterwise$cluster
10
20
30
40
50
60
70
0
200
400
600
800
1000
1200

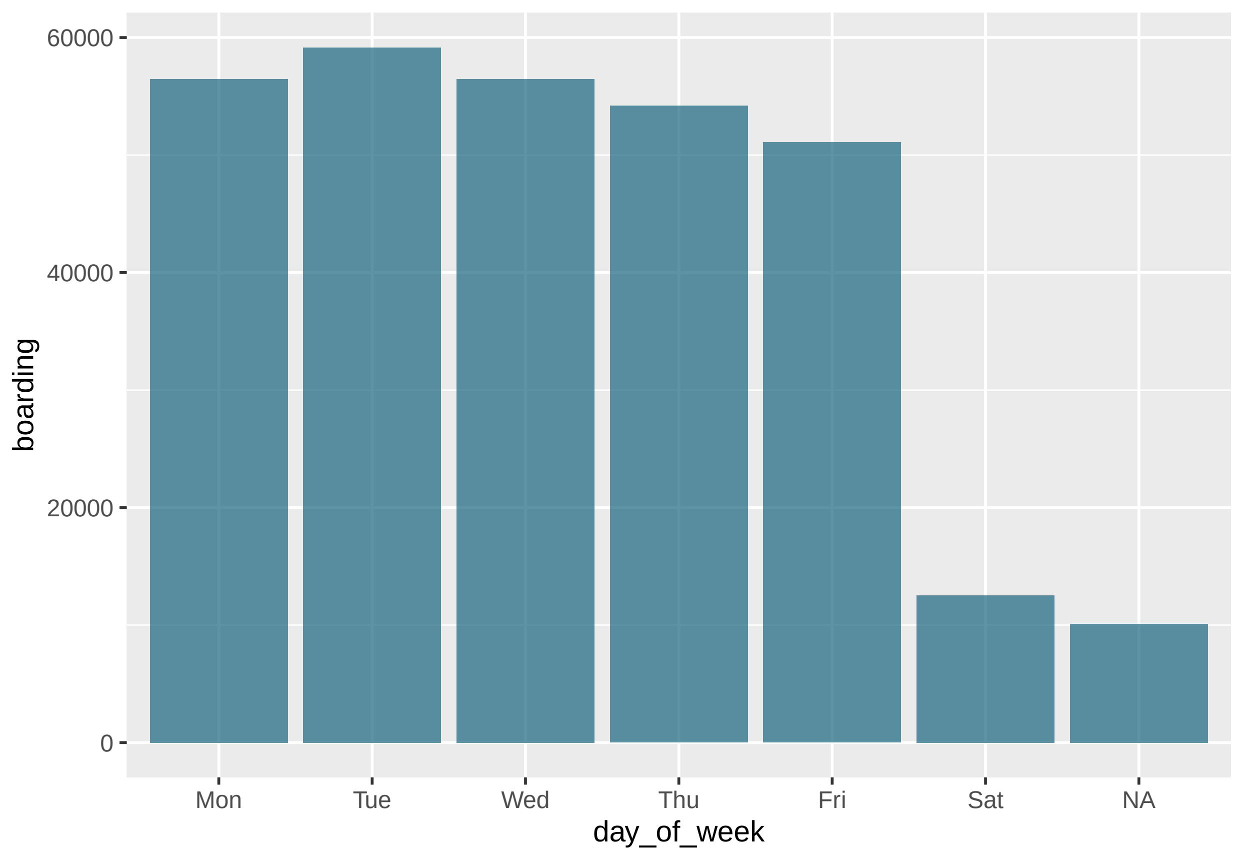
```
## integer(0)
```

Visual story telling part 2: Capital Metro data



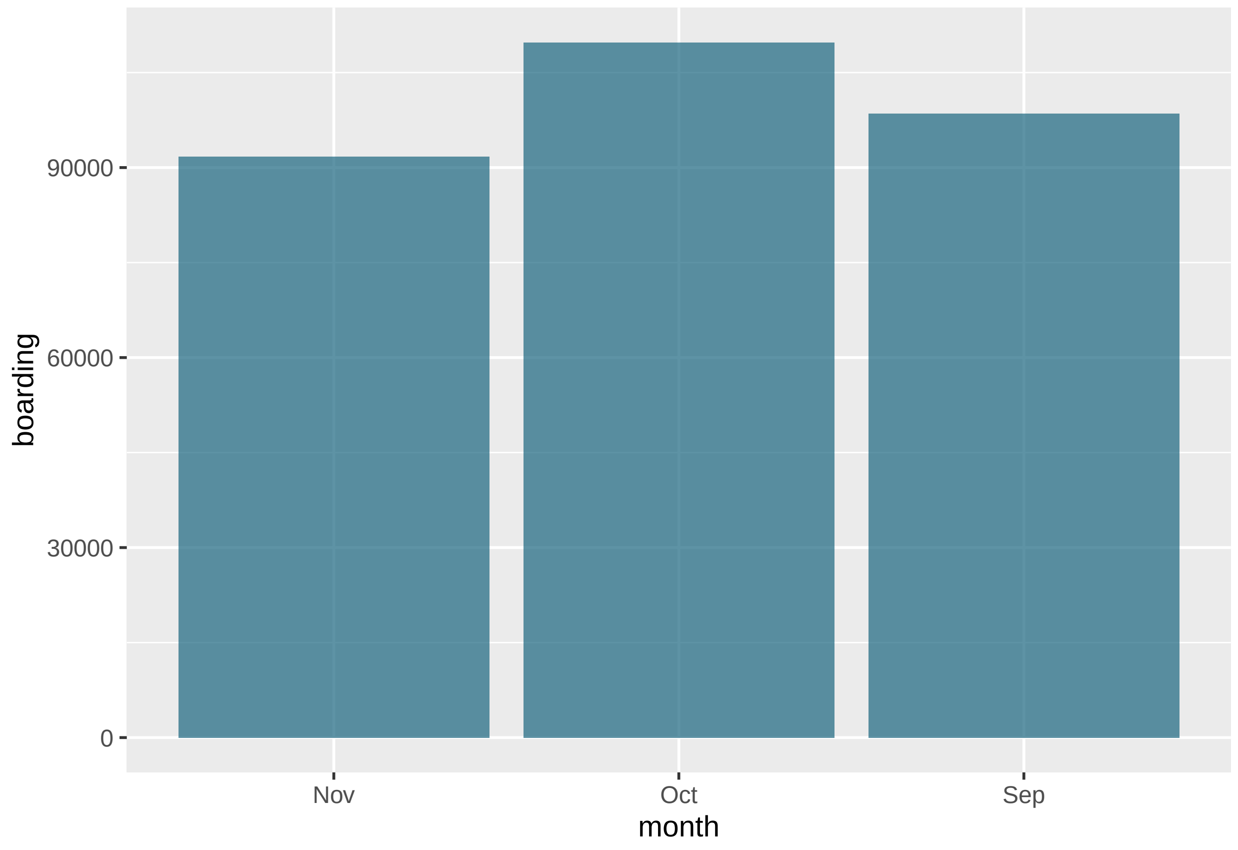
- We see that the boarding pattern is not uniform throughout the day. The peak is around 3 pm to 5 pm. This is the time when most people step out of their offices, graduate students attend classes etc.
- Number of people boarding is comparitively low in the morning and late evening

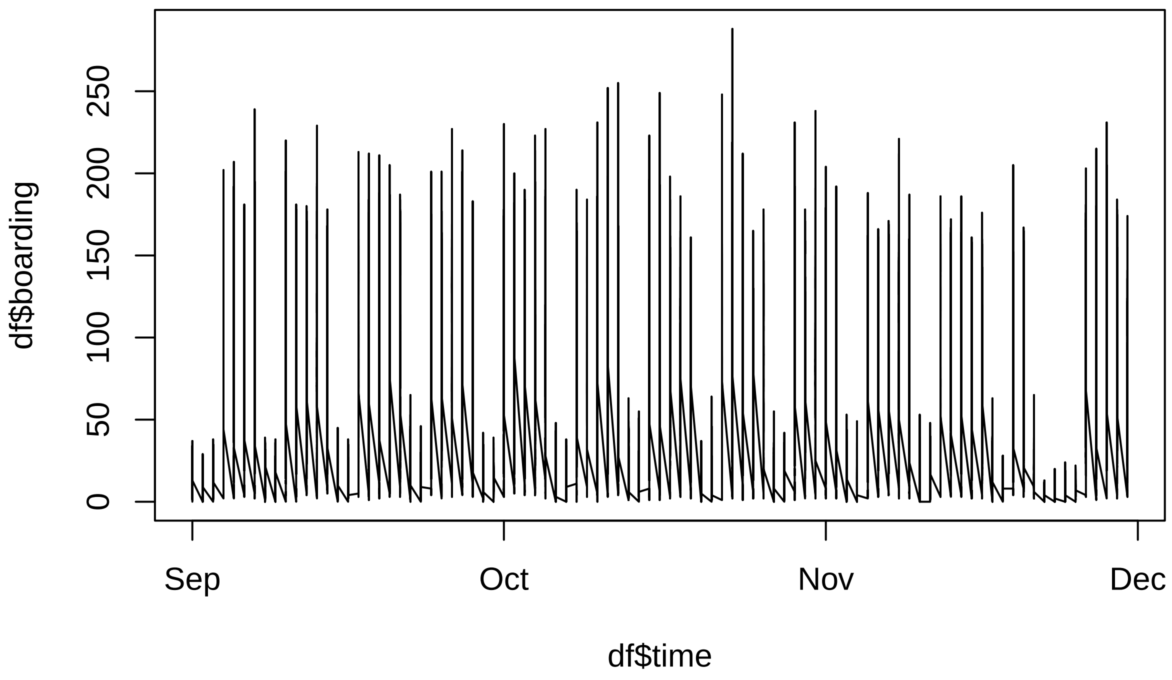
```
##   day_of_week boarding
## 1         Mon    56470
## 2         Tue    59164
## 3         Wed    56465
## 4         Thu    54189
## 5         Fri    51076
## 6         Sat    12525
```



- We see that number of people boarding is much higher during the weekdays as compared to the weekend. This is because students usually account for a high percentage of bus riders and since most classes are on the weekdays, the demand for buses is more then.

```
##   month boarding
## 1   Nov    91707
## 2   Oct   109767
## 3   Sep    98530
```


boarding
90000
60000
30000
0
Nov
Oct
Sep
month



- To get an idea of what the pattern looks over the 3-month period, when we plot the time-series boarding data we notice that there is a significant slump in boarding during the last week of November which is consistent with the Thanksgiving and Christmas holidays. We can attribute this decline in ridership to students visiting their families during the holidays.

Portfolio Modelling

- For the first portfolio, I picked 5 random stocks from the database of ETFs, thus giving me a diverse portfolio. I made sure for every stock there were at least 5 years of data.

```
##               ClCl.IWRa     ClCl.VNQa     ClCl.SPYa    ClCl.USOa   ClCl.GOVTa
## 2016-08-03           NA            NA            NA           NA           NA
## 2016-08-04  0.0004615438 -0.0040821051  0.0010639791  0.015479928  0.001909855
## 2016-08-05  0.0075538923  0.0006647281  0.0081788687  0.002032469 -0.004574876
## 2016-08-08 -0.0002861787  0.0017712388 -0.0005957925  0.024340848  0.000000000
## 2016-08-09 -0.0009158920  0.0054150184  0.0005961477 -0.002970359  0.002680927
## 2016-08-10 -0.0025212239 -0.0010992086 -0.0024749932 -0.029791386  0.001909855

##               ClCl.IWRa     ClCl.VNQa     ClCl.SPYa    ClCl.USOa   ClCl.GOVTa
## 2016-08-04  0.0004615438 -0.0040821051  0.0010639791  0.015479928  0.001909855
## 2016-08-05  0.0075538923  0.0006647281  0.0081788687  0.002032469 -0.004574876
## 2016-08-08 -0.0002861787  0.0017712388 -0.0005957925  0.024340848  0.000000000
## 2016-08-09 -0.0009158920  0.0054150184  0.0005961477 -0.002970359  0.002680927
## 2016-08-10 -0.0025212239 -0.0010992086 -0.0024749932 -0.029791386  0.001909855
## 2016-08-11  0.0053998392 -0.0113336161  0.0046406681  0.045035770 -0.003431224
```

- Understanding the return behavior of the stocks

```
##      Index              ClCl.IWRa
##  Min.   :2016-08-03   Min.   :-0.1379755
##  1st Qu.:2018-02-05   1st Qu.:-0.0041810
##  Median :2019-08-08   Median : 0.0009807
##  Mean   :2019-08-08   Mean   : 0.0024485
##  3rd Qu.:2021-02-09   3rd Qu.: 0.0064269
##  Max.   :2022-08-12   Max.   : 2.9363938
##                       NA's   :1
```

```
##      Index              ClCl.VNQa
##  Min.   :2016-08-03   Min.   :-0.1772773
##  1st Qu.:2018-02-05   1st Qu.:-0.0052662
##  Median :2019-08-08   Median : 0.0009221
##  Mean   :2019-08-08   Mean   : 0.0003271
##  3rd Qu.:2021-02-09   3rd Qu.: 0.0067669
##  Max.   :2022-08-12   Max.   : 0.0899666
##                       NA's   :1
```

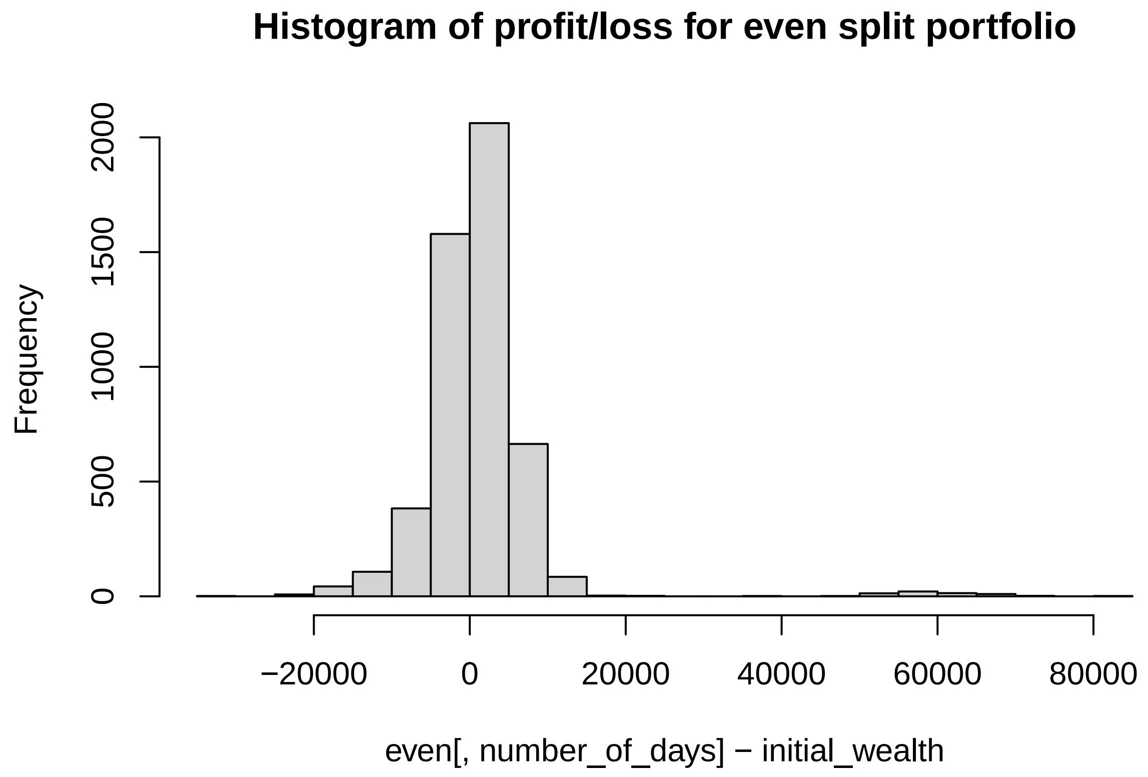
```
##      Index              ClCl.SPYa
##  Min.   :2016-08-03   Min.   :-0.1094237
##  1st Qu.:2018-02-05   1st Qu.:-0.0033009
##  Median :2019-08-08   Median : 0.0007051
##  Mean   :2019-08-08   Mean   : 0.0005914
##  3rd Qu.:2021-02-09   3rd Qu.: 0.0059158
##  Max.   :2022-08-12   Max.   : 0.0906033
##                       NA's   :1
```

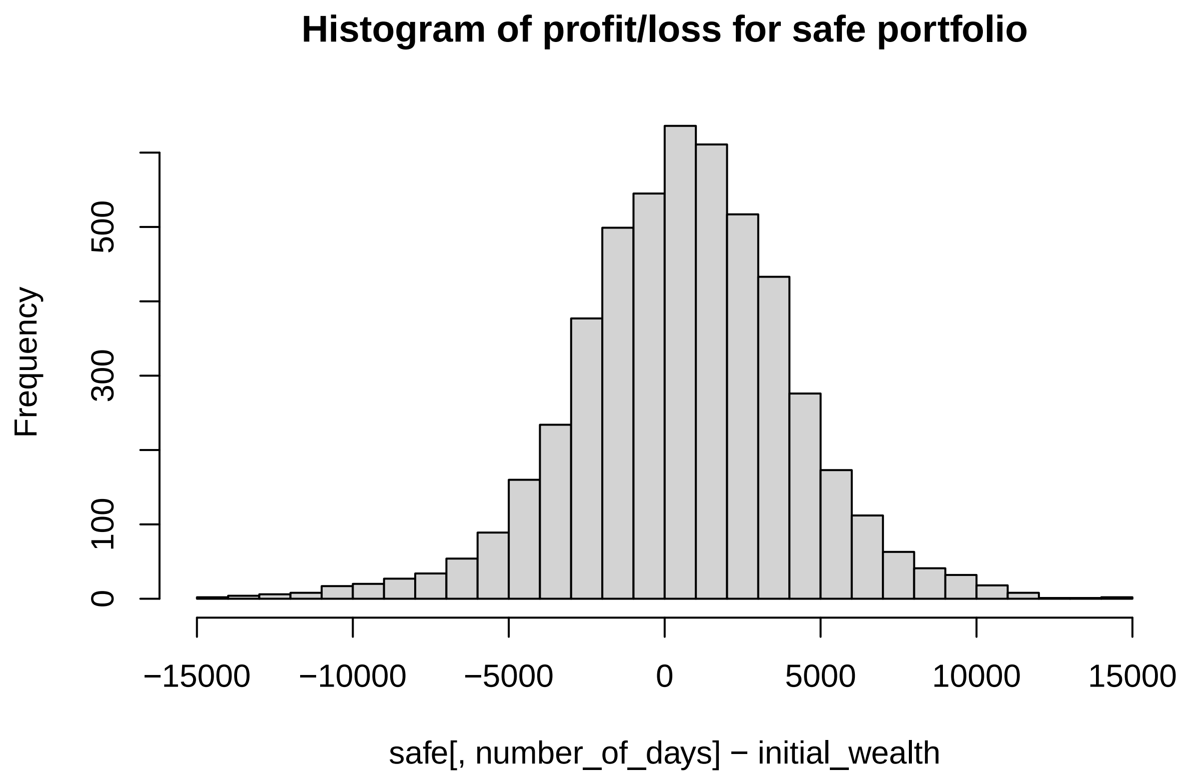
```
##      Index              ClCl.USOa
##  Min.   :2016-08-03   Min.   :-0.8679578
##  1st Qu.:2018-02-05   1st Qu.:-0.0106857
##  Median :2019-08-08   Median : 0.0017182
##  Mean   :2019-08-08   Mean   :-0.0002811
##  3rd Qu.:2021-02-09   3rd Qu.: 0.0128586
##  Max.   :2022-08-12   Max.   : 0.1666667
##                       NA's   :1
```

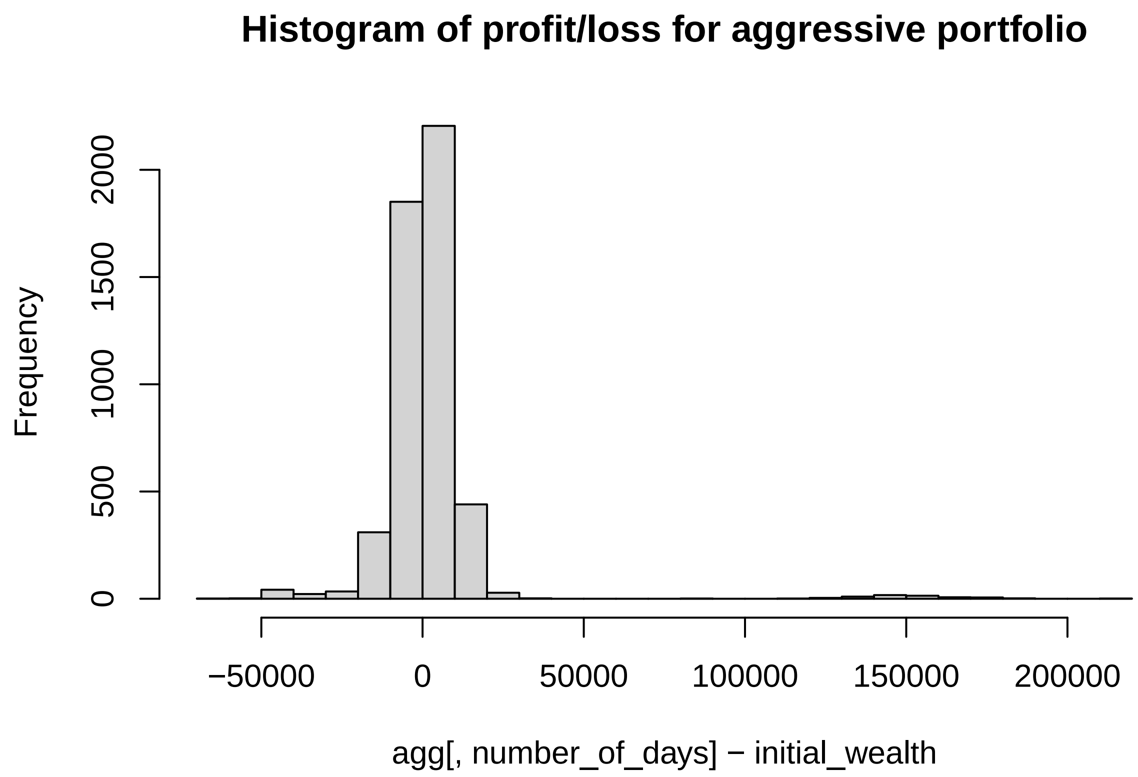
```
##      Index              ClCl.GOVTa
##  Min.   :2016-08-03   Min.   :-0.0222462
##  1st Qu.:2018-02-05   1st Qu.:-0.0015854
##  Median :2019-08-08   Median : 0.0000000
##  Mean   :2019-08-08   Mean   : 0.0000121
##  3rd Qu.:2021-02-09   3rd Qu.: 0.0016083
##  Max.   :2022-08-12   Max.   : 0.0225758
##                       NA's   :1
```

- The returns obtained for the ETFs are in the following ranges:- IWR = -13% to 293%, VNQ = -17% to 9%, SPY = -10% to 9%, USO = -86% to 16%, GOVT = -2% to 2%
- Based on the summary statistics of the ETFs, we can infer that VNQ, SPY and GOVT are in the safe category as their returns are in a reasonable range.

- Split considered for Safe portfolio is SPY:30%, VNQ:35% and GOVT:35%
- We can infer that USO and IWR are high risk/return ETFs since they have an extremely wide return range.
- Split considered for aggressive portfolio is IWR:50% and USO:50%
- Now simulate many different possible scenarios




Histogram of profit/loss for safe portfolio
Frequency
0
100
300
500
−15000
−10000
−5000
0
5000
10000
15000
safe[, number_of_days] − initial_wealth



```
## [1] "Value at risk for even split :  7970.7"

## [1] "Value at risk for safe portfolio :  5104.88"

## [1] "Value at risk for aggressive portfolio :  13152.63"
```

- On seeing the above histogram graphs we observe the following:
- For the evenly distributed portfolios the spread is quite uneven
- For the safe portfolio a relatively even curve is observed
- For the aggressive portfolio the there is a widespread.
- The value at risk for the varios portfolios are as follows:
- The safe portfolio is 5175.9 which is the least risky portfolio as the name suggests
- The aggressive portfolio has a VaR of 13549.54 which high
- The evenly distributed portfolio is more or less the average of the safe and aggressive portfolios

PCA and Clustering

Principal Component Analysis

```
## 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
```

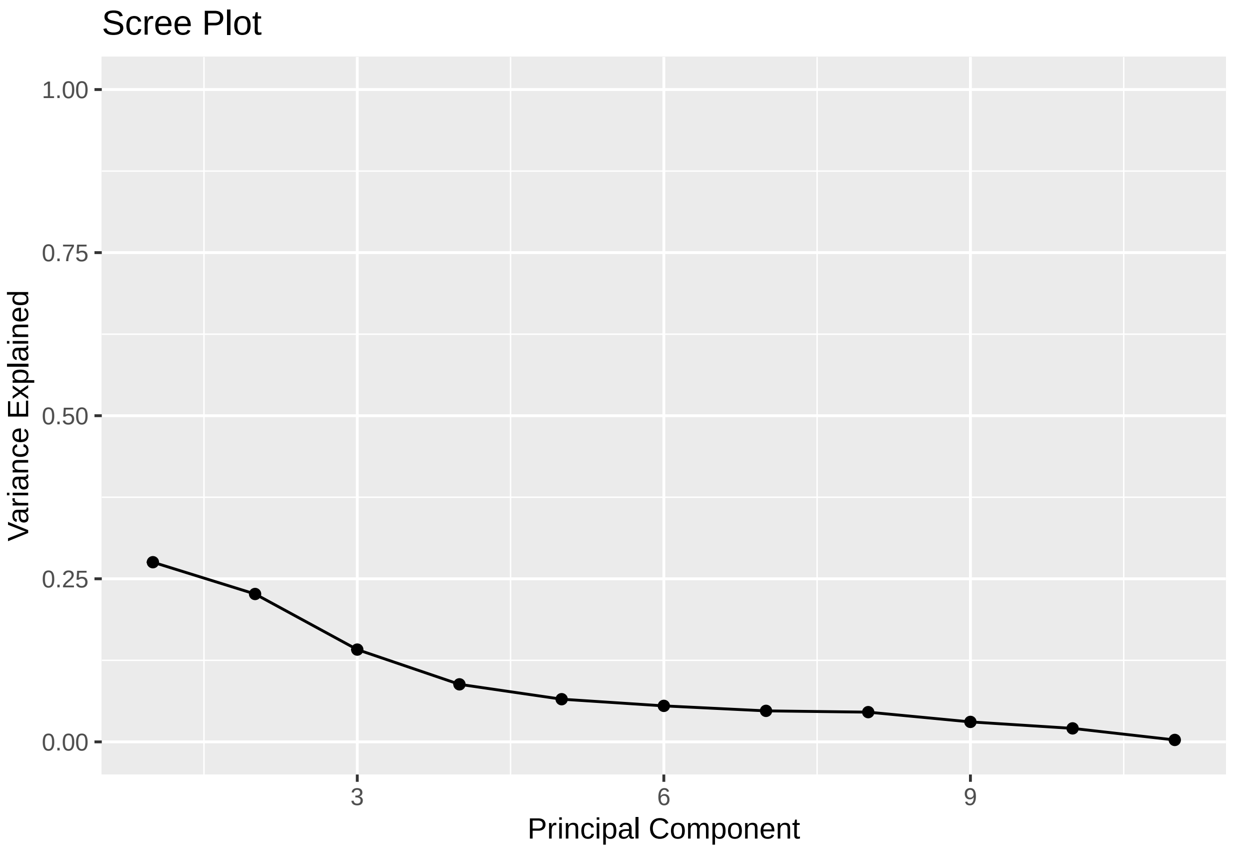


Figure 1: The above table shows the summary statistics for the PCA. It has total 11 principal components as there are 11 features in our dataset (not considering wine quality and wine color). It shows the Standard deviation, variance and cummulative proportion for each of the principal components. As we can see in the table, PC1 has 27% of total variance, PC2 has about 23% of total variance and so on. The same can also be seen in the Scree Plot. The sumary statistics show that we need five principal components to achieve a variance of about 85%

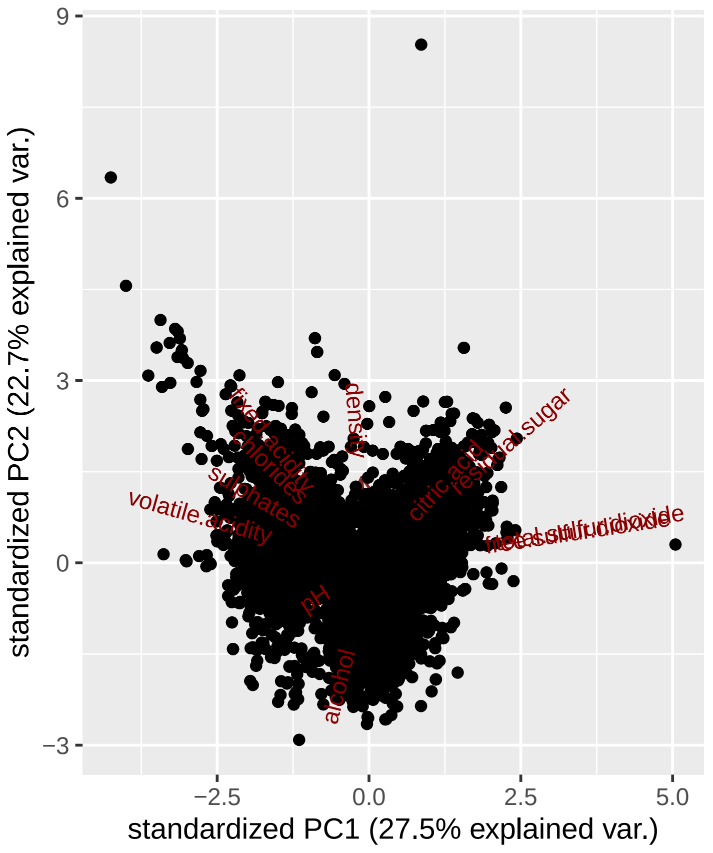
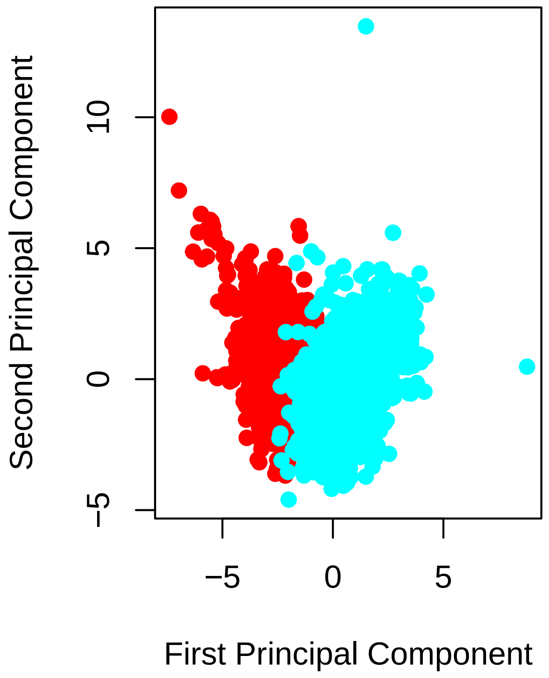
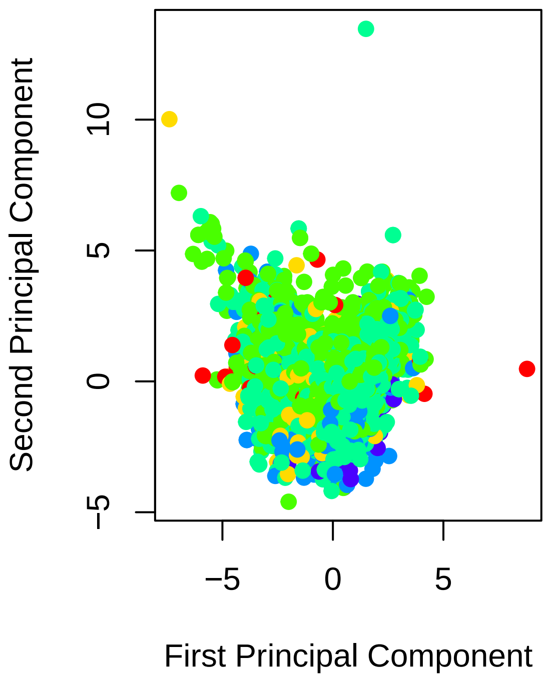


Figure 2: The above plot shows the first two principal components. The orange arrows indicate the first two principal component loading vectors



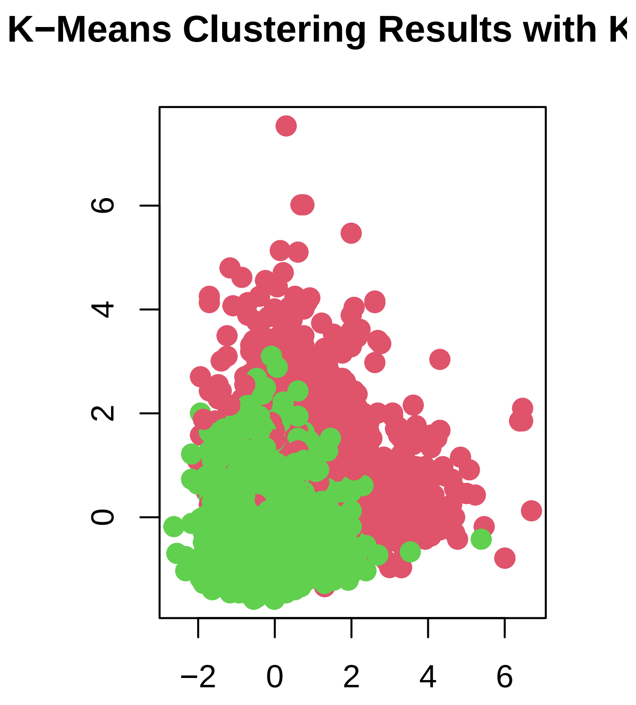
- The plot shows the scores for the first two principal components. The color of wine (Red wine or White wine) is used to color code the data points as red and blue respectively. We can see that observations belonging to red wine category tend to lie near each other and observations belonging to the white wine category cluster together.



- This plot shows the projections of the observations on the first two principal components. The quality of wine is used to color code the observations. As we can see, there is no evident clustering of the data points based on the quality of wine.
- In conclusion, PCA performs very well in clustering the observations based on the wine color. It would not have been possible to visualize the data without using a dimension reduction method such as PCA, since based on the full data set there are approx 6500 possible scatter plots, none of which would have been particularly informative.
- However, it performs poorly in clustering them based on the quality of wine.

Clustering

Using K means clustering instead of heirarchical clustering as we know that our datset has two types of wine.



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

```
## [1] 98.58396
```

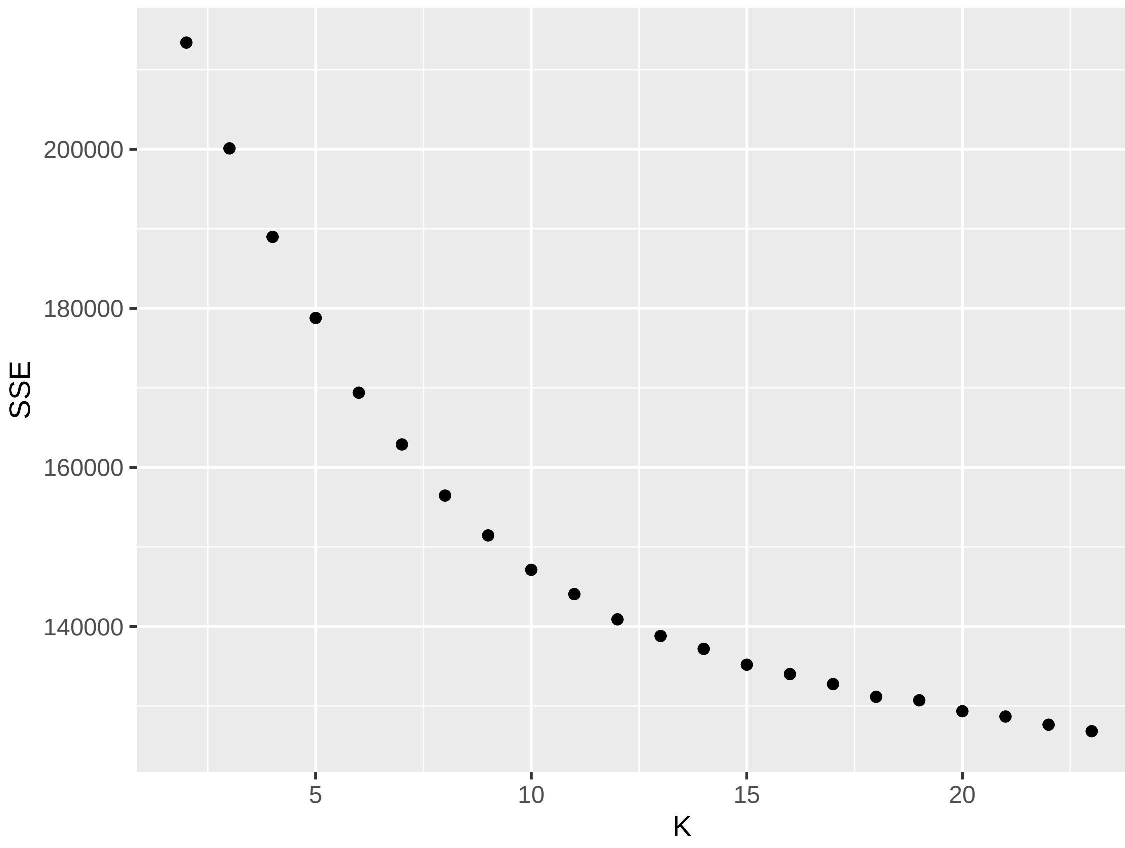
```
## Accuracy of K-means clustering is: 98.58396
```

- The confusion matrix shows how K-means clustering performed on our data set in distinguishing red wine from white wine. The accuracy turned out to be 98.58% which is very good. We can conclude that k-means clustering has done a very good job in clustering.

Market Segmentation

First step in analysing the trends in data set is by cleaning the data. The columns 'adult', 'spam' and 'chatter' have been removed from the data set. The data has been checked for NA or missing values but no missing values were found. The data is scaled and centred using the 'scale' function.

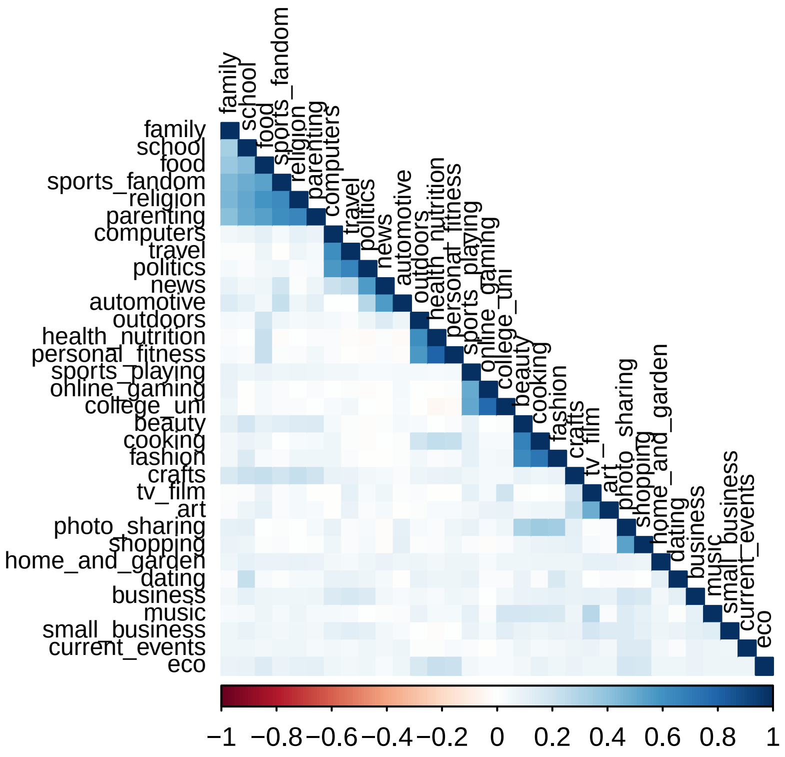
After data cleaning, K-means clustering algorithm was run on the data to find similar data points To find the optimal number of clusters, the within-cluster sum of squares was plotted for



K values from 1 to 15.

- From the above elbow plot, we can say that the optimal value of k is k=7. Hence, running K means++ clustering with K=7.

```
## corrplot 0.92 loaded
```

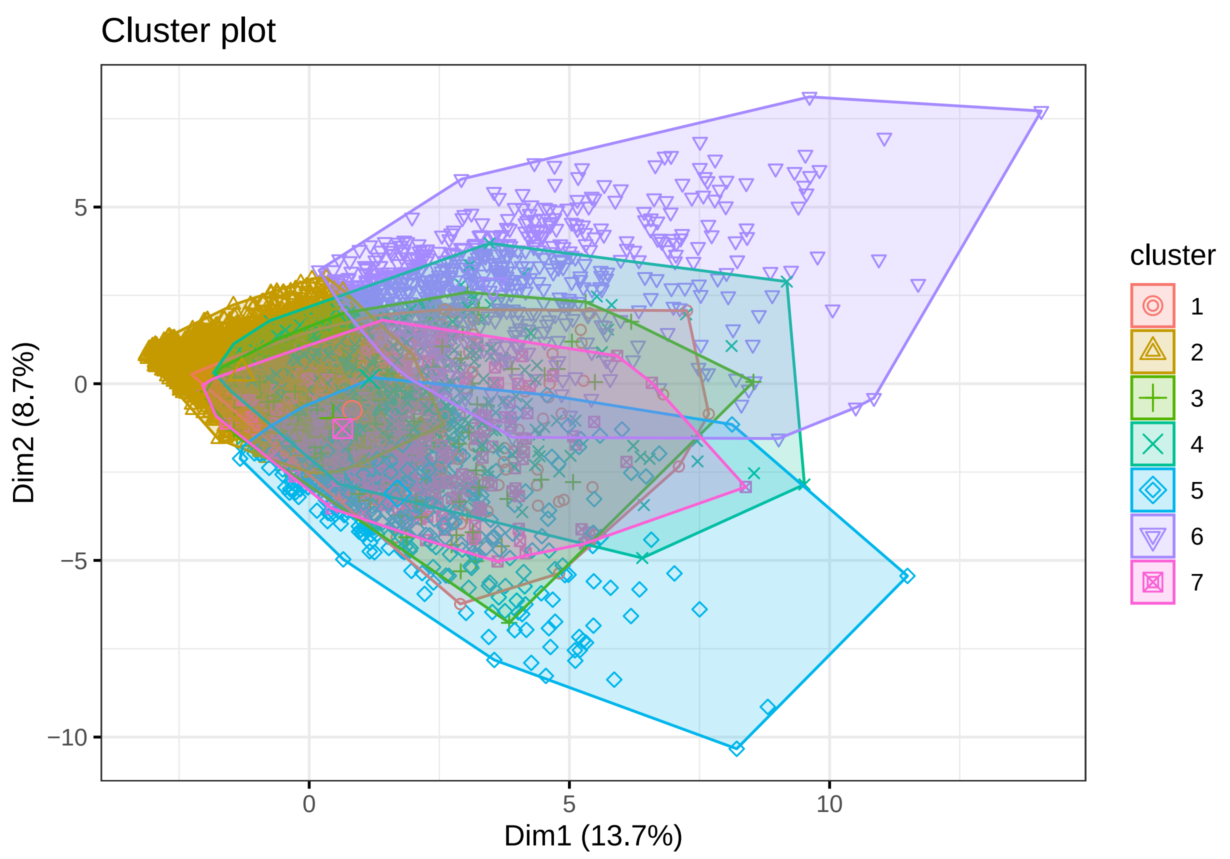


- From the above correlation plot we can infer that:
- Shopping and photo-sharing are positively correlated

- College_uni and online_gaming stands out with a strong positive correlation

- Health_nutrition ,personal_fitness and outdoors have a high positive correlation showing these people are health conscious

- Fashion and beauty have a strong positive correlation



- From the above plot, we cannot clearly differentiate between the clusters. Hence, extracting the cluster centers.

```
##                        c1        c2         c3        c4         c5        c6
## current_events  1.9706546 1.4083461  1.4485714 1.6380952  1.7685950 1.6412104
## travel          2.1196388 1.0184332  1.5028571 5.5730159  1.4752066 1.2651297
## photo_sharing   2.8487585 2.3102919  2.7628571 2.5317460  6.1797521 2.6628242
## tv_film         5.4198646 0.7081413  1.2742857 0.9857143  0.8409091 0.8919308
## sports_fandom   1.3724605 0.9534050  1.2885714 2.0587302  1.1280992 5.9279539
## politics        1.5778781 1.0038402  1.3142857 9.0317460  1.4297521 1.1152738
## food            1.5665914 0.7122376  1.2257143 1.4126984  1.0061983 4.5547550
## family          0.7065463 0.5673323  1.0714286 0.9285714  0.8863636 2.4899135
## home_and_garden 0.7494357 0.4178187  0.5828571 0.5952381  0.6115702 0.6383285
## music           1.8532731 0.5000000  0.6371429 0.6031746  1.2582645 0.6974063
## news            1.2618510 0.6559140  0.8200000 5.3920635  1.0330579 1.0000000
## online_gaming   0.7268623 0.5921659 10.5342857 0.8190476  1.1322314 1.0086455
## shopping        1.7471783 1.2782898  1.2457143 1.3571429  1.9710744 1.5230548
## health_nutrition 1.7674944 1.1080389  1.7514286 1.6206349  2.2231405 1.8904899
## college_uni     2.7246050 0.8402458 11.0514286 1.2666667  1.5475207 1.1657061
## sports_playing  0.7562077 0.4203789  2.7228571 0.6269841  0.8347107 0.7363112
## cooking         1.4401806 0.8853047  1.5514286 1.2873016 11.2458678 1.6484150
## eco             0.6049661 0.3620072  0.4857143 0.5888889  0.5227273 0.6484150
## computers       0.4808126 0.3653354  0.5371429 2.4920635  0.7500000 0.7204611
## business        0.7200903 0.3159242  0.3685714 0.6587302  0.5950413 0.4913545
## outdoors        0.6726862 0.3801843  0.5857143 0.8968254  0.7768595 0.6916427
## crafts           1.0744921 0.3202765  0.5342857 0.5634921  0.5805785 1.0561960
## automotive       0.5620767 0.5732207  0.8885714 2.4222222  0.8677686 1.0461095
## art              4.4943567 0.3259089  1.1428571 0.4412698  0.7148760 0.6657061
## religion         1.1196388 0.5061444  0.7485714 1.0095238  0.8140496 5.2391931
## beauty           0.6726862 0.3381976  0.3885714 0.4682540  4.1095041 1.1167147
## parenting        0.6275395 0.4470046  0.6828571 0.9507937  0.7623967 4.0749280
## dating           0.6297968 0.5576037  0.6800000 1.0809524  0.7851240 0.8847262
## school           0.7291196 0.4541731  0.5142857 0.7428571  0.9772727 2.7103746
## personal_fitness 1.0677201 0.6577061  1.0000000 1.0111111  1.3388430 1.1743516
## fashion          0.9525959 0.5176651  0.8657143 0.6777778  5.7541322 1.0504323
## small_business   0.8126411 0.2324629  0.4000000 0.4587302  0.4524793 0.3789625
##                         c7
## current_events    1.5037406
## travel            1.2032419
## photo_sharing     2.6857855
## tv_film           0.8004988
## sports_fandom     1.1408978
## politics          1.2643392
## food              2.1022444
## family            0.7493766
## home_and_garden   0.5947631
## music             0.6770574
## news              1.1034913
## online_gaming     0.8678304
## shopping          1.4800499
## health_nutrition 12.1184539
## college_uni       0.9064838
## sports_playing    0.5860349
## cooking           3.3067332
## eco               0.9114713
## computers         0.5648379
## business          0.4725686
## outdoors          2.7443890
## crafts            0.5598504
## automotive        0.6309227
## art               0.5835411
## religion          0.7356608
## beauty            0.4201995
## parenting         0.7406484
## dating            1.0349127
## school            0.5498753
## personal_fitness  6.4750623
## fashion           0.7805486
## small_business    0.2581047
```

```
## [1] 0.06061021 0.53440963 0.04788617 0.08619510 0.06621973 0.09495143 0.10972773
```

- As we can see from the above table, about 53 percent of the total number of data points are in a single cluster. This cluster includes the people who have tweeted less than 2 times on an average in all the categories. This could mean that most of the followers of "NutrientH20" are not active users of "twitter" or social media in general. Despite not being active users of social media, these people are following "Nutrient20" which means that the current social media marketing strategy is working quite well.

- The cluster with the lowest number of people, on an average tweeted more about ‘photo sharing’, ‘cooking’, and ‘fashion’. Therefore, in order to attract and appeal to more of the people who are more interested in photo sharing, cooking and fashion, the company should position their brand in a way that it seems like it is related to photo sharing or cooking or fashion.

Reuters Corpus

** Pre-processing data: **

- We take all the files to be used for training and store their names in ‘file_list’
- Our data does not have the column name for the Author name, so to extract it we will be using the file name

```
# Function to split the path name by '/'
extract_author <- function(x) {
  strsplit(x, "/")
}

#Make a dataframe with the author names of each document
df = as.data.frame(lapply(extract_author(file_list), function(x) x[length(x) - 1] ))
author <- t(df)
rownames(author)<-seq(1,2500)
#author
```

- We will then clean up the file name using piping operator from magrittr

```
mynames = file_list %>%
  { strsplit(., '/', fixed=TRUE) } %>%
  { lapply(., tail, n=2) } %>%
  { lapply(., paste0, collapse = '') } %>%
  unlist

# Rename the articles
#mynames
names(raw.data) = mynames
```

- Some pre-processing/tokenization steps uses tm_map which maps some function to every document in the corpus
- Using this we make everything lowercase, remove numbers, punctuations, excess white spaces and stopwords
- Creating a doc-term-matrix

```
DTM_raw = DocumentTermMatrix(my_documents)
DTM_raw # some basic summary statistics
```

```
## <<DocumentTermMatrix (documents: 2500, terms: 22479)>>
## Non-/sparse entries: 496196/55701304
## Sparsity           : 99%
## Maximal term length: 40
## Weighting          : term frequency (tf)
```

```r
class(DTM_raw)  # a special kind of sparse matrix format
```

```
## [1] "DocumentTermMatrix"    "simple_triplet_matrix"
```

- The final step in pre-processing is to drop those terms that only occur in one or two documents. This is done to remove rare terms from which there is nothing to learn from
- To do this we remove those terms that have count 0 in >97% of docs.

```
## <<DocumentTermMatrix (documents: 2500, terms: 1264)>>
## Non-/sparse entries: 360611/2799389
## Sparsity           : 89%
## Maximal term length: 13
## Weighting          : term frequency (tf)
```

- Constructing TF IDF weights

```r
# construct TF IDF weights
tfidf_raw = weightTfIdf(DTM_raw)
tfidf_raw
```

```
## <<DocumentTermMatrix (documents: 2500, terms: 1264)>>
## Non-/sparse entries: 320611/2839389
## Sparsity           : 90%
## Maximal term length: 13
## Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
```

```r
train.data <- data.frame(as.matrix(tfidf_raw), stringsAsFactors=FALSE)
#Merge with author names
train.data <- merge(train.data,author,by =0)
train.data$V1 <- as.factor(train.data$V1)
```

- We repeat the above steps to obtain the test data

```
## <<DocumentTermMatrix (documents: 2500, terms: 22987)>>
## Non-/sparse entries: 502733/56964767
## Sparsity           : 99%
## Maximal term length: 45
## Weighting          : term frequency (tf)
```

```
## [1] "DocumentTermMatrix"    "simple_triplet_matrix"
```

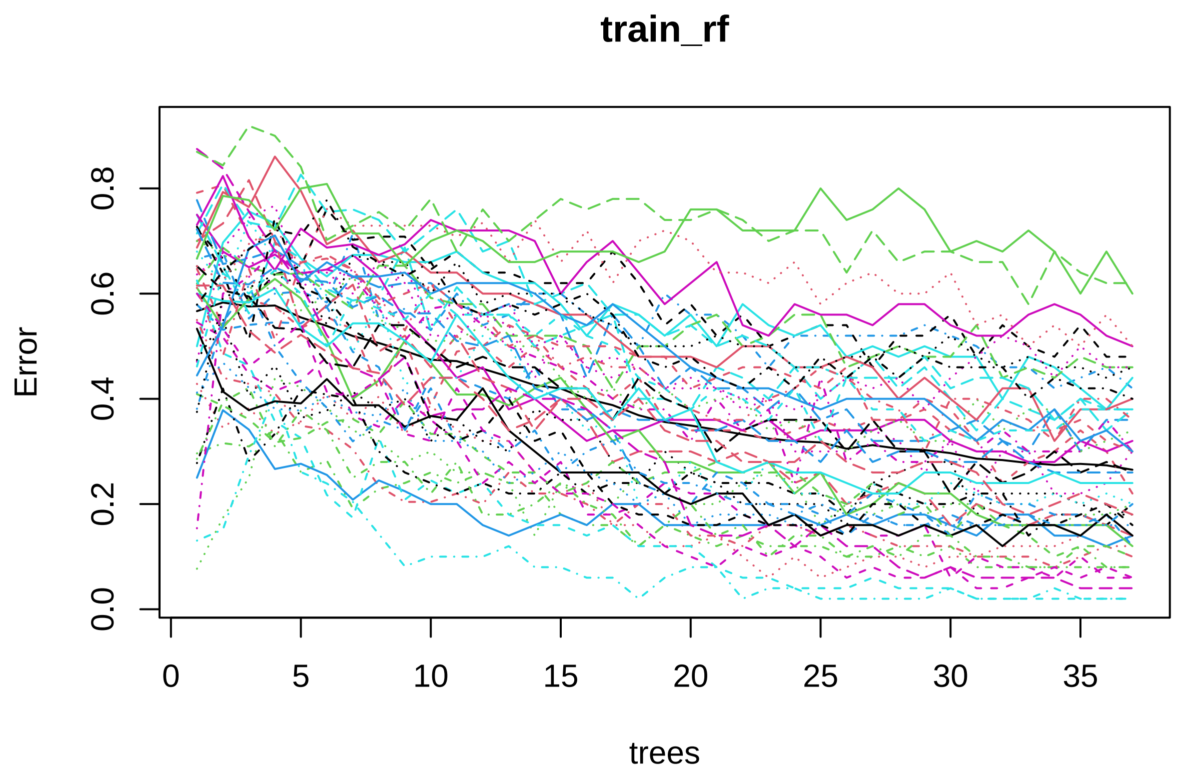
```
## <<DocumentTermMatrix (documents: 2500, terms: 1296)>>
## Non-/sparse entries: 366936/2873064
## Sparsity           : 89%
## Maximal term length: 13
## Weighting          : term frequency (tf)
```

```
## <<DocumentTermMatrix (documents: 2500, terms: 1296)>>
## Non-/sparse entries: 326936/2913064
## Sparsity           : 90%
## Maximal term length: 13
## Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
```

- We ignore words in the test set that are not present in the train set

** Running Random Forest on the Train Data **

```
##
## Call:
##  randomForest(formula = new_train$V1 ~ ., data = new_train, ntree = 37,      proximity = T)
##                Type of random forest: classification
##                      Number of trees: 37
## No. of variables tried at each split: 34
##
##         OOB estimate of  error rate: 26.52%
## Confusion matrix:
##                   AaronPressman AlanCrosby AlexanderSmith BenjaminKangLim
## AaronPressman                39          0              0               0
## AlanCrosby                    0         40              0               0
## AlexanderSmith                0          0             29               0
## BenjaminKangLim               1          0              0              31
## BernardHickey                 1          0              0               0
## BradDorfman                   1          1              0               0
## DarrenSchuettler              0          0              0               0
## DavidLawder                   0          0              0               0
## EdnaFernandes                 0          0              0               0
## EricAuchard                   1          1              1               0
## FumikoFujisaki                1          0              0               0
## GrahamEarnshaw                0          0              0               0
## HeatherScoffield              0          0              0               0
## JaneMacartney                 0          0              0               3
## JanLopatka                    0          4              0               0
## JimGilchrist                  0          0              0               0
## JoeOrtiz                      0          0              2               0
## JohnMastrini                  0          4              0               0
## JonathanBirt                  0          0              0               0
## JoWinterbottom                0          0              0               0
## KarlPenhaul                   0          0              1               0
## KeithWeir                     0          0              0               0
## KevinDrawbaugh                0          0              0               0
## KevinMorrison                 0          0              0               0
## KirstinRidley                 0          0              2               0
## KouroshKarimkhany             0          0              0               0
## LydiaZajc                     0          0              0               0
## LynneO'Donnell                0          0              0               0
## LynnleyBrowning               0          0              0               0
## MarcelMichelson               0          0              0               0
## MarkBendeich                  0          0              1               0
## MartinWolk                    0          0              0               0
## MatthewBunce                  0          0              0               0
## MichaelConnor                 0          0              0               0
## MureDickie                    0          0              0              10
## NickLouth                     0          0              0               0
## PatriciaCommins               0          0              0               0
## PeterHumphrey                 0          0              0               1
## PierreTran                    0          0              1               0
## RobinSidel                    1          0              0               0
## RogerFillion                1           0              0                0
## SamuelPerry                 5           0              0                0
## SarahDavison                1           0              1                1
## ScottHillis                 0           0              0               11
## SimonCowell                 2           0              1                0
## TanEeLyn                    0           0              0                0
## TheresePoletti              1           0              0                0
## TimFarrand                  0           0              2                0
## ToddNissen                  0           0              0                0
## WilliamKazer                0           0              0                2
##                    BernardHickey BradDorfman DarrenSchuettler DavidLawder
## AaronPressman               0           0                0           0
## AlanCrosby                  0           0                0           0
## AlexanderSmith              2           0                0           0
## BenjaminKangLim             0           0                0           0
## BernardHickey              34           0                0           1
## BradDorfman                 0          26                1           1
## DarrenSchuettler            0           0               44           0
## DavidLawder                 0           0                0          46
## EdnaFernandes               1           0                0           1
## EricAuchard                 0           0                0           0
## FumikoFujisaki              0           0                0           0
## GrahamEarnshaw              0           0                0           0
## HeatherScoffield            0           1                1           0
## JaneMacartney               0           0                0           0
## JanLopatka                  0           0                0           0
## JimGilchrist                0           0                0           0
## JoeOrtiz                    0           0                0           0
## JohnMastrini                0           0                0           0
## JonathanBirt                0           0                0           0
## JoWinterbottom              0           0                0           0
## KarlPenhaul                 0           0                0           0
## KeithWeir                   0           0                0           0
## KevinDrawbaugh              0           1                1           0
## KevinMorrison               6           0                0           0
## KirstinRidley               0           0                0           0
## KouroshKarimkhany           0           0                0           0
## LydiaZajc                   0           0                7           0
## LynneO'Donnell              0           0                0           0
## LynnleyBrowning             0           0                0           0
## MarcelMichelson             0           0                0           0
## MarkBendeich                6           0                0           0
## MartinWolk                  1           1                0           0
## MatthewBunce                0           0                0           0
## MichaelConnor               2           4                0           0
## MureDickie                  0           0                0           0
## NickLouth                   1           0                0           0
## PatriciaCommins             0           8                0           0
## PeterHumphrey               0           0                1           0
## PierreTran                  0           0                0           1
## RobinSidel                  0           0                0           0
## RogerFillion                0           0                1           0
## SamuelPerry                 0           1                0           0
## SarahDavison                0           0                1           0
## ScottHillis                0           0                 0            0
## SimonCowell                0           1                 0            0
## TanEeLyn                   0           0                 0            0
## TheresePoletti             0           1                 0            0
## TimFarrand                 1           0                 0            0
## ToddNissen                 0           1                 1            3
## WilliamKazer               0           0                 0            1
##                   EdnaFernandes EricAuchard FumikoFujisaki GrahamEarnshaw
## AaronPressman              0           0              1              0
## AlanCrosby                 0           0              0              0
## AlexanderSmith             0           0              1              0
## BenjaminKangLim            0           0              0              0
## BernardHickey              1           1              0              0
## BradDorfman                0           1              0              0
## DarrenSchuettler           0           0              0              0
## DavidLawder                0           0              0              0
## EdnaFernandes             32           0              0              0
## EricAuchard                0          28              0              0
## FumikoFujisaki             0           0             47              0
## GrahamEarnshaw             0           0              0             41
## HeatherScoffield           0           0              0              0
## JaneMacartney              0           0              0              3
## JanLopatka                 0           0              0              0
## JimGilchrist               0           0              0              0
## JoeOrtiz                   2           0              0              0
## JohnMastrini               0           0              0              0
## JonathanBirt               4           0              0              0
## JoWinterbottom             2           0              0              0
## KarlPenhaul                0           0              0              0
## KeithWeir                  2           0              0              0
## KevinDrawbaugh             0           1              0              0
## KevinMorrison              0           0              0              0
## KirstinRidley              0           0              1              0
## KouroshKarimkhany          0           2              0              0
## LydiaZajc                  0           0              0              0
## LynneO'Donnell             0           0              0              0
## LynnleyBrowning            0           0              0              0
## MarcelMichelson            0           0              1              0
## MarkBendeich               0           0              0              0
## MartinWolk                 0           0              1              0
## MatthewBunce               0           0              0              0
## MichaelConnor              1           0              0              0
## MureDickie                 0           0              0              1
## NickLouth                  0           1              0              0
## PatriciaCommins            1           2              1              0
## PeterHumphrey              0           0              0              0
## PierreTran                 0           0              1              0
## RobinSidel                 1           4              0              0
## RogerFillion               0           0              0              0
## SamuelPerry                0           3              0              0
## SarahDavison               0           0              2              1
## ScottHillis                1           0              0              0
## SimonCowell                1           0              0              0
## TanEeLyn                   1           0              0              1
## ThereseﾠPoletti               0            4              0              0
## TimFarrand                   3            0              1              0
## ToddNissen                   0            0              0              0
## WilliamKazer                 0            0              0              3
##                  HeatherScoffield JaneMacartney JanLopatka JimGilchrist
## AaronPressman                   0             0          0            0
## AlanCrosby                      0             0          5            0
## AlexanderSmith                  0             0          0            0
## BenjaminKangLim                 0             4          0            1
## BernardHickey                   1             0          0            1
## BradDorfman                     0             0          0            0
## DarrenSchuettler                0             0          0            0
## DavidLawder                     0             0          0            0
## EdnaFernandes                   0             0          0            1
## EricAuchard                     0             0          0            0
## FumikoFujisaki                  0             0          0            0
## GrahamEarnshaw                  0             2          0            0
## HeatherScoffield               45             0          0            0
## JaneMacartney                   0            27          0            0
## JanLopatka                      0             0         43            0
## JimGilchrist                    0             0          0           49
## JoeOrtiz                        0             0          0            0
## JohnMastrini                    0             0          6            0
## JonathanBirt                    0             0          0            0
## JoWinterbottom                  0             0          0            0
## KarlPenhaul                     0             0          0            0
## KeithWeir                       0             0          0            0
## KevinDrawbaugh                  0             0          0            0
## KevinMorrison                   1             0          0            2
## KirstinRidley                   0             0          0            0
## KouroshKarimkhany               0             0          0            0
## LydiaZajc                       1             0          0            0
## LynneO'Donnell                  0             0          0            0
## LynnleyBrowning                 1             0          0            0
## MarcelMichelson                 0             0          0            0
## MarkBendeich                    1             0          0            0
## MartinWolk                      1             0          0            0
## MatthewBunce                    0             0          1            0
## MichaelConnor                   0             0          0            0
## MureDickie                      0             2          0            0
## NickLouth                       0             0          0            0
## PatriciaCommins                 0             0          0            0
## PeterHumphrey                   0             0          0            0
## PierreTran                      0             0          0            0
## RobinSidel                      0             0          0            0
## RogerFillion                    0             0          0            0
## SamuelPerry                     0             0          0            0
## SarahDavison                    0             0          0            1
## ScottHillis                     0             5          0            0
## SimonCowell                     0             0          1            0
## TanEeLyn                        0             0          0            2
## TheresePoletti                  0             0          0            0
## TimFarrand                      0             0          0            0
## ToddNissen                      0             0          0            0
## WilliamKazer                     0            6          0            0
##                    JoeOrtiz JohnMastrini JonathanBirt JoWinterbottom KarlPenhaul
## AaronPressman             0            0            0              0           0
## AlanCrosby                0            4            1              0           0
## AlexanderSmith            2            0            0              3           0
## BenjaminKangLim           0            0            0              0           0
## BernardHickey             0            0            0              0           0
## BradDorfman               0            0            0              0           1
## DarrenSchuettler          0            0            0              0           0
## DavidLawder               0            0            0              0           0
## EdnaFernandes             0            0            0              2           0
## EricAuchard               0            0            0              0           0
## FumikoFujisaki            0            0            0              0           1
## GrahamEarnshaw            0            0            0              0           0
## HeatherScoffield          0            0            0              0           0
## JaneMacartney             0            0            0              0           0
## JanLopatka                0            3            0              0           0
## JimGilchrist              0            0            0              0           0
## JoeOrtiz                 35            0            0              4           0
## JohnMastrini              0           40            0              0           0
## JonathanBirt              1            0           43              1           0
## JoWinterbottom            3            0            0             44           0
## KarlPenhaul               0            0            0              0          44
## KeithWeir                 0            0            0              0           2
## KevinDrawbaugh            0            0            1              0           1
## KevinMorrison             0            0            0              0           0
## KirstinRidley             1            0            1              2           1
## KouroshKarimkhany         0            0            0              0           0
## LydiaZajc                 0            0            0              0           0
## LynneO'Donnell            0            0            0              0           0
## LynnleyBrowning           0            0            0              0           0
## MarcelMichelson           0            0            1              0           0
## MarkBendeich              0            0            0              0           1
## MartinWolk                0            0            1              0           1
## MatthewBunce              0            0            0              0           3
## MichaelConnor             0            0            0              0           0
## MureDickie                0            0            0              0           0
## NickLouth                 0            0            0              0           0
## PatriciaCommins           0            0            0              0           0
## PeterHumphrey             0            0            0              0           0
## PierreTran                1            0            0              0           0
## RobinSidel                0            0            0              0           0
## RogerFillion              0            0            0              0           0
## SamuelPerry               0            0            0              0           0
## SarahDavison              0            0            1              0           0
## ScottHillis               0            0            0              0           0
## SimonCowell               2            0            0              0           0
## TanEeLyn                  0            1            0              0           0
## TheresePoletti            0            0            0              0           0
## TimFarrand                2            0            1              2           0
## ToddNissen                0            0            0              0           0
## WilliamKazer              0            0            0              0           0
##                    KeithWeir KevinDrawbaugh KevinMorrison KirstinRidley
## AaronPressman              0              0             0             0
## AlanCrosby                 0              0             0             0
## AlexanderSmith             3              0             0             3
## BenjaminKangLim            0              0             0             0
## BernardHickey              0              0             3             0
## BradDorfman                0              5             0             0
## DarrenSchuettler           0              0             0             0
## DavidLawder                0              0             0             0
## EdnaFernandes              0              0             1             2
## EricAuchard                0              0             0             0
## FumikoFujisaki             0              0             0             0
## GrahamEarnshaw             0              0             0             0
## HeatherScoffield           0              0             0             0
## JaneMacartney              0              0             0             0
## JanLopatka                 0              0             0             0
## JimGilchrist               0              0             0             0
## JoeOrtiz                   1              0             0             0
## JohnMastrini               0              0             0             0
## JonathanBirt               0              0             0             0
## JoWinterbottom             0              0             0             0
## KarlPenhaul                0              0             0             0
## KeithWeir                 40              0             1             1
## KevinDrawbaugh             1             35             2             0
## KevinMorrison              0              1            37             0
## KirstinRidley              3              0             0            30
## KouroshKarimkhany          0              0             0             0
## LydiaZajc                  0              0             0             0
## LynneO'Donnell             0              0             0             0
## LynnleyBrowning            0              0             0             0
## MarcelMichelson            0              0             0             0
## MarkBendeich               0              0             3             0
## MartinWolk                 0              1             0             1
## MatthewBunce               1              0             0             0
## MichaelConnor              0              1             1             0
## MureDickie                 0              0             0             0
## NickLouth                  0              0             0             1
## PatriciaCommins            0              3             0             0
## PeterHumphrey              0              0             0             0
## PierreTran                 0              0             0             1
## RobinSidel                 0              3             0             0
## RogerFillion               0              0             0             0
## SamuelPerry                0              0             0             0
## SarahDavison               0              0             0             0
## ScottHillis                0              0             0             0
## SimonCowell                0              0             0             0
## TanEeLyn                   0              0             0             0
## TheresePoletti             0              0             0             0
## TimFarrand                 1              0             0             2
## ToddNissen                 0              0             0             0
## WilliamKazer               0              0             0             1
##                   KouroshKarimkhany LydiaZajc LynneO'Donnell LynnleyBrowning
## AaronPressman                     1         0              0               0
## AlanCrosby                        0         0              0               0
## AlexanderSmith                    0         0              0               0
## BenjaminKangLim                   0         0              1               0
## BernardHickey                 0         0              0               0
## BradDorfman                   0         0              0               0
## DarrenSchuettler              0         4              0               0
## DavidLawder                   0         0              0               0
## EdnaFernandes                 0         0              0               0
## EricAuchard                   0         1              0               0
## FumikoFujisaki                0         0              0               0
## GrahamEarnshaw                0         0              0               0
## HeatherScoffield              0         2              0               0
## JaneMacartney                 0         0              1               0
## JanLopatka                    0         0              0               0
## JimGilchrist                  0         0              1               0
## JoeOrtiz                      0         0              0               0
## JohnMastrini                  0         0              0               0
## JonathanBirt                  0         0              0               0
## JoWinterbottom                0         0              0               0
## KarlPenhaul                   0         0              0               2
## KeithWeir                     0         0              0               0
## KevinDrawbaugh                1         1              0               0
## KevinMorrison                 0         0              0               0
## KirstinRidley                 1         0              0               0
## KouroshKarimkhany            45         0              0               0
## LydiaZajc                     0        42              0               0
## LynneO'Donnell                0         0             49               0
## LynnleyBrowning               0         0              0              48
## MarcelMichelson               0         0              0               0
## MarkBendeich                  0         0              0               0
## MartinWolk                    3         0              0               0
## MatthewBunce                  0         0              0               2
## MichaelConnor                 0         0              0               2
## MureDickie                    0         0              1               0
## NickLouth                     0         0              0               0
## PatriciaCommins               0         0              0               0
## PeterHumphrey                 0         0              0               0
## PierreTran                    0         0              0               1
## RobinSidel                    0         0              0               0
## RogerFillion                  0         0              0               0
## SamuelPerry                   2         1              0               0
## SarahDavison                  0         0              0               0
## ScottHillis                   0         0              0               0
## SimonCowell                   0         0              0               0
## TanEeLyn                      0         0              3               0
## TheresePoletti                4         0              0               0
## TimFarrand                    0         0              0               0
## ToddNissen                    0         0              0               0
## WilliamKazer                  0         0              3               0
##                     MarcelMichelson MarkBendeich MartinWolk MatthewBunce
## AaronPressman                     0            0          1            0
## AlanCrosby                        0            0          0            0
## AlexanderSmith                    0            0          0            0
## BenjaminKangLim                   0            0          0            0
## BernardHickey                     0            4          1            0
## BradDorfman                       0            0          1            0
## DarrenSchuettler                  0            0          0            0
## DavidLawder                    0            0          0            0
## EdnaFernandes                  1            1          0            0
## EricAuchard                    0            0          1            0
## FumikoFujisaki                 0            0          0            0
## GrahamEarnshaw                 0            0          0            0
## HeatherScoffield               0            0          0            0
## JaneMacartney                  0            0          0            0
## JanLopatka                     0            0          0            0
## JimGilchrist                   0            0          0            0
## JoeOrtiz                       0            0          0            0
## JohnMastrini                   0            0          0            0
## JonathanBirt                   0            0          0            0
## JoWinterbottom                 0            0          0            0
## KarlPenhaul                    0            0          0            1
## KeithWeir                      1            0          0            1
## KevinDrawbaugh                 0            0          0            0
## KevinMorrison                  0            2          0            0
## KirstinRidley                  0            1          1            0
## KouroshKarimkhany              0            0          0            0
## LydiaZajc                      0            0          0            0
## LynneO'Donnell                 0            0          0            0
## LynnleyBrowning                0            0          0            1
## MarcelMichelson               43            0          1            0
## MarkBendeich                   0           35          0            0
## MartinWolk                     0            1         33            0
## MatthewBunce                   0            0          0           40
## MichaelConnor                  0            1          1            0
## MureDickie                     0            1          0            0
## NickLouth                      0            0          0            0
## PatriciaCommins                0            0          1            0
## PeterHumphrey                  0            0          0            0
## PierreTran                     6            0          0            1
## RobinSidel                     0            0          1            0
## RogerFillion                   0            0          0            0
## SamuelPerry                    0            0          1            0
## SarahDavison                   0            0          0            0
## ScottHillis                    0            0          0            1
## SimonCowell                    0            1          0            0
## TanEeLyn                       0            1          0            0
## TherasePoletti                 0            0          1            0
## TimFarrand                     0            3          0            0
## ToddNissen                     0            0          0            1
## WilliamKazer                   0            0          0            0
##                   MichaelConnor MureDickie NickLouth PatriciaCommins
## AaronPressman                  0          0         1               0
## AlanCrosby                     0          0         0               0
## AlexanderSmith                 0          0         0               1
## BenjaminKangLim                1          7         0               0
## BernardHickey                  1          0         0               1
## BradDorfman                    3          0         0               2
## DarrenSchuettler               0          0         0               0
## DavidLawder                    0          0         0               0
## EdnaFernandes                  0          0         0               0
## EricAuchard                    1          0         4               1
## FumikoFujisaki               0          0         0               0
## GrahamEarnshaw               0          1         0               0
## HeatherScoffield             0          0         0               0
## JaneMacartney                0          1         0               0
## JanLopatka                   0          0         0               0
## JimGilchrist                 0          0         0               0
## JoeOrtiz                     0          0         0               0
## JohnMastrini                 0          0         0               0
## JonathanBirt                 0          0         0               0
## JoWinterbottom               0          0         0               0
## KarlPenhaul                  0          1         0               0
## KeithWeir                    0          0         1               0
## KevinDrawbaugh               0          0         0               2
## KevinMorrison                0          0         0               0
## KirstinRidley                1          1         1               0
## KouroshKarimkhany            0          0         0               0
## LydiaZajc                    0          0         0               0
## LynneO’Donnell               0          0         0               0
## LynnleyBrowning              0          0         0               0
## MarcelMichelson              0          0         0               0
## MarkBendeich                 0          1         0               0
## MartinWolk                   1          0         0               0
## MatthewBunce                 0          0         0               0
## MichaelConnor               31          0         1               1
## MureDickie                   0         25         1               0
## NickLouth                    1          0        42               0
## PatriciaCommins              0          0         0              25
## PeterHumphrey                0          1         0               0
## PierreTran                   0          0         0               0
## RobinSidel                   0          0         0               1
## RogerFillion                 0          0         0               0
## SamuelPerry                  1          0         1               1
## SarahDavison                 0          1         0               0
## ScottHillis                  1          4         0               0
## SimonCowell                  1          0         0               0
## TanEeLyn                     0          0         0               0
## TheresePoletti               0          0         3               1
## TimFarrand                   0          0         0               0
## ToddNissen                   0          0         0               1
## WilliamKazer                 0          2         0               0
##                   PeterHumphrey PierreTran RobinSidel RogerFillion SamuelPerry
## AaronPressman                0          0          1            0           3
## AlanCrosby                   0          0          0            0           0
## AlexanderSmith               0          0          0            0           0
## BenjaminKangLim              1          0          0            0           0
## BernardHickey                0          0          0            0           0
## BradDorfman                  0          0          1            0           1
## DarrenSchuettler             0          0          1            0           0
## DavidLawder                  0          0          0            0           0
## EdnaFernandes                0          3          0            0           0
## EricAuchard                  0          0          0            0           3
## FumikoFujisaki               0          0          1            0           0
## GrahamEarnshaw               0          0          0            0           0
## HeatherScoffield             0          0          1            0           0
## JaneMacartney                0          0          0            0           0
## JanLopatka                   0          0          0            0           0
## JimGilchrist                 0          0          0            0           0
## JoeOrtiz                     0          1          0            0           0
## JohnMastrini                 0          0          0            0           0
## JonathanBirt                 0          0          0            0           0
## JoWinterbottom               0          0          0            0           0
## KarlPenhaul                  0          0          0            1           0
## KeithWeir                    0          0          0            0           0
## KevinDrawbaugh               0          0          1            0           0
## KevinMorrison                0          0          0            0           0
## KirstinRidley                0          0          0            0           0
## KouroshKarimkhany            0          0          0            0           2
## LydiaZajc                    0          0          0            0           0
## LynneO'Donnell               1          0          0            0           0
## LynnleyBrowning              0          0          0            0           0
## MarcelMichelson              0          4          0            0           0
## MarkBendeich                 0          0          0            0           0
## MartinWolk                   0          1          0            0           2
## MatthewBunce                 0          2          0            0           0
## MichaelConnor                0          0          1            1           0
## MureDickie                   3          0          0            0           1
## NickLouth                    0          0          0            2           2
## PatriciaCommins              0          1          4            1           1
## PeterHumphrey               46          0          0            0           0
## PierreTran                   0         37          0            0           0
## RobinSidel                   0          0         38            0           0
## RogerFillion                 0          0          1           47           0
## SamuelPerry                  0          0          0            0          27
## SarahDavison                 6          0          0            0           0
## ScottHillis                  0          0          0            0           0
## SimonCowell                  0          1          1            0           0
## TanEeLyn                     8          0          0            0           0
## TheresePoletti               0          0          0            0           8
## TimFarrand                   0          0          0            0           0
## ToddNissen                   0          0          1            0           0
## WilliamKazer                 0          0          0            0           0
##                   SarahDavison ScottHillis SimonCowell TanEeLyn TheresePoletti
## AaronPressman                0           0           0        0              2
## AlanCrosby                   0           0           0        0              0
## AlexanderSmith               1           0           3        0              0
## BenjaminKangLim              0           2           0        1              0
## BernardHickey                0           0           0        0              0
## BradDorfman                  0           0           1        0              0
## DarrenSchuettler             0           1           0        0              0
## DavidLawder                  0           0           0        0              0
## EdnaFernandes                0           0           2        0              0
## EricAuchard                  0           0           1        0              7
## FumikoFujisaki               0           0           0        0              0
## GrahamEarnshaw               2           0           0        0              0
## HeatherScoffield             0           0           0        0              0
## JaneMacartney                1           5           0        2              0
## JanLopatka                   0           0           0        0              0
## JimGilchrist                 0           0           0        0              0
## JoeOrtiz                 0           0           3        1              0
## JohnMastrini             0           0           0        0              0
## JonathanBirt             0           0           1        0              0
## JoWinterbottom           0           0           0        0              0
## KarlPenhaul              0           0           0        0              0
## KeithWeir                0           0           1        0              0
## KevinDrawbaugh           0           0           1        0              0
## KevinMorrison            1           0           0        0              0
## KirstinRidley            0           0           2        0              0
## KouroshKarimkhany        0           0           0        0              1
## LydiaZajc                0           0           0        0              0
## LynneO'Donnell           0           0           0        0              0
## LynnleyBrowning          0           0           0        0              0
## MarcelMichelson          0           0           0        0              0
## MarkBendeich             1           0           0        0              0
## MartinWolk               0           0           0        0              1
## MatthewBunce             0           0           0        1              0
## MichaelConnor            0           0           0        0              0
## MureDickie               0           2           0        0              0
## NickLouth                0           0           0        0              0
## PatriciaCommins          0           0           0        0              1
## PeterHumphrey            0           0           0        1              0
## PierreTran               0           0           0        0              0
## RobinSidel               0           0           0        0              0
## RogerFillion             0           0           0        0              0
## SamuelPerry              0           0           0        0              5
## SarahDavison            32           1           0        1              0
## ScottHillis              0          19           0        2              0
## SimonCowell              0           0          35        0              0
## TanEeLyn                 0           1           0       31              0
## TheresePoletti           0           0           0        0             27
## TimFarrand               0           0           2        0              0
## ToddNissen               0           0           0        0              1
## WilliamKazer             1           9           0        2              0
##                    TimFarrand ToddNissen WilliamKazer class.error
## AaronPressman               0          1            0        0.22
## AlanCrosby                  0          0            0        0.20
## AlexanderSmith              1          1            0        0.42
## BenjaminKangLim             0          0            0        0.38
## BernardHickey               0          0            0        0.32
## BradDorfman                 0          4            0        0.48
## DarrenSchuettler            0          0            0        0.12
## DavidLawder                 0          4            0        0.08
## EdnaFernandes               3          0            0        0.36
## EricAuchard                 0          0            0        0.44
## FumikoFujisaki              0          0            0        0.06
## GrahamEarnshaw              0          0            4        0.18
## HeatherScoffield            0          0            0        0.10
## JaneMacartney               0          1            6        0.46
## JanLopatka                  0          0            0        0.14
## JimGilchrist                0          0            0        0.02
## JoeOrtiz                    0          0            1        0.30
## JohnMastrini                0          0            0        0.20
## JonathanBirt                0          0            0        0.14
## JoWinterbottom          1       0       0       0.12
## KarlPenhaul             0       0       0       0.12
## KeithWeir               0       0       0       0.20
## KevinDrawbaugh          0       1       0       0.30
## KevinMorrison           0       0       0       0.26
## KirstinRidley           1       0       0       0.40
## KouroshKarimkhany       0       0       0       0.10
## LydiaZajc               0       0       0       0.16
## LynneO'Donnell          0       0       0       0.02
## LynnleyBrowning         0       0       0       0.04
## MarcelMichelson         0       0       0       0.14
## MarkBendeich            1       0       0       0.30
## MartinWolk              0       0       0       0.34
## MatthewBunce            0       0       0       0.20
## MichaelConnor           1       1       0       0.38
## MureDickie              0       0       3       0.50
## NickLouth               0       0       0       0.16
## PatriciaCommins         0       1       0       0.50
## PeterHumphrey           0       0       0       0.08
## PierreTran              0       0       0       0.26
## RobinSidel              0       1       0       0.24
## RogerFillion            0       0       0       0.06
## SamuelPerry             2       0       0       0.46
## SarahDavison            0       0       0       0.36
## ScottHillis             0       0       6       0.62
## SimonCowell             3       0       0       0.30
## TanEeLyn                0       0       1       0.38
## TheresePoletti          0       0       0       0.46
## TimFarrand             30       0       0       0.40
## ToddNissen              0      41       0       0.18
## WilliamKazer            0       0      20       0.60
```



** Running the model on the Test Data **

```
## [1] 0.672
```

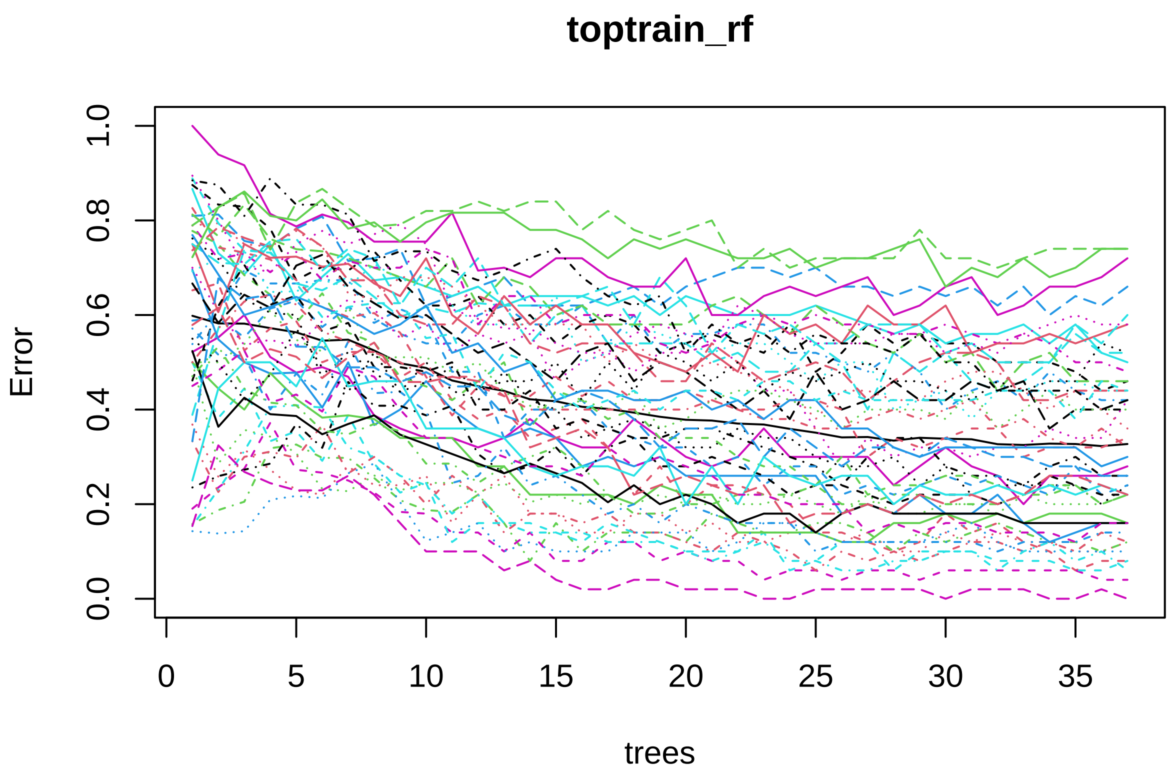
- Trying different values of top features to see if performance improves.

```
importanceOrder=order(-train_rf$importance)
topnames=rownames(train_rf$importance)[importanceOrder][2:151]
topnames = append(topnames,"V1")

top_test <- test.data %>% select(all_of(topnames))
top_train<-train.data %>% select(all_of(topnames))
```

** Running Random Forest on Train data

```
library("randomForest")
set.seed(1)
toptrain_rf = randomForest(top_train$V1~., data=top_train, ntree=37, proximity=T)
#table(predict(train_rf), train.data$V1)
#train_rf
plot(toptrain_rf)
```



** Checking model on test data **

```
toptestPred = predict(toptrain_rf, newdata=top_test, type = 'class')
#table(testPred, test.data$V1)

CM = table(toptestPred, top_test$V1)
accuracy = (sum(diag(CM)))/sum(CM)
accuracy
```

```
## [1] 0.528
```

- The performance does not improve that much for various values used to subset features.

** Summary of the Process: **

- We first extract the author name from the file paths.
- Next, we clean the file names and pre-process it in the following order:

a) Join all files and convert into one corpus. This corpus will have rows as each document.
b) Tokenize the documents(split each document into separate words) and convert to lower case c)Remove numbers, punctuation, extra white spaces and stop words (Words like 'as','the','so' do not add much meaning to the sentence without context. Hence we remove them to reduce any noise in the data) from the document

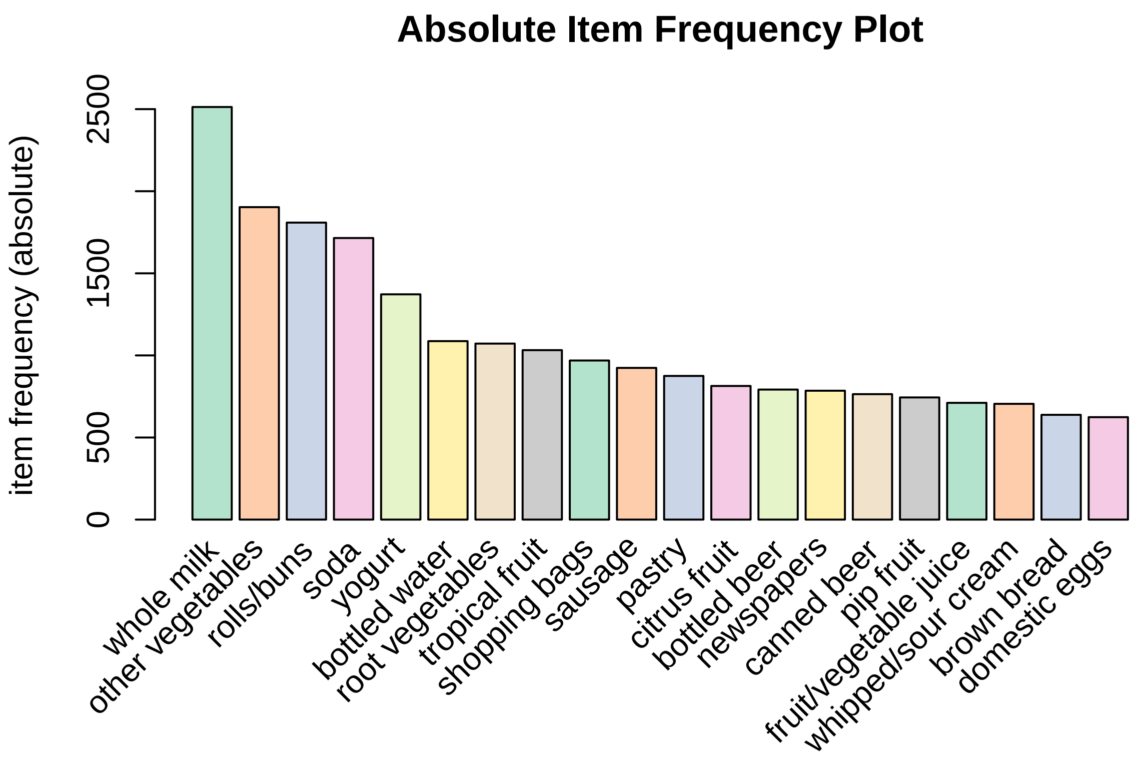
c) Stemming (Words such as ‘run’ and ‘running’ essentially mean the same. So in stemming, we take each word and use it’s root value. In this example, our root value will be ‘run’)

- We convert the output from step 3 to a sparse matrix. Rows in a sparse matrix represent data for each document. The columns in the sparse matrix represent each word identified after Step 2. The values for a particular row, column in the matrix is the number of times the word appears in the particular document.
- We drop terms that may occur only once or twice in the documents. This further removes some noise from the data and reduces number of features.
- Some texts can be small while some can be large. To compare several texts, the frequency of each word relative to the length of the text is more helpful than the count of each word in the text. We thus use the TF IDF values for this Purpose. So we replace values in the sparse matrix to TF IDF scores.
- We then merge the author names with output from 5 to get train data.
- Repeat steps 1 to step 6 using test data.
- We ignore words present in the test set but not in the train set.
- Using an intersection of words between train and test set, We now run Random Forest for classification to get accuracy of 70.8%.
- We try using a number of most important features to reduce dimensionality. However, this does not improve the performance.

Association Rule Mining

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

```
## Loading required package: RColorBrewer
```

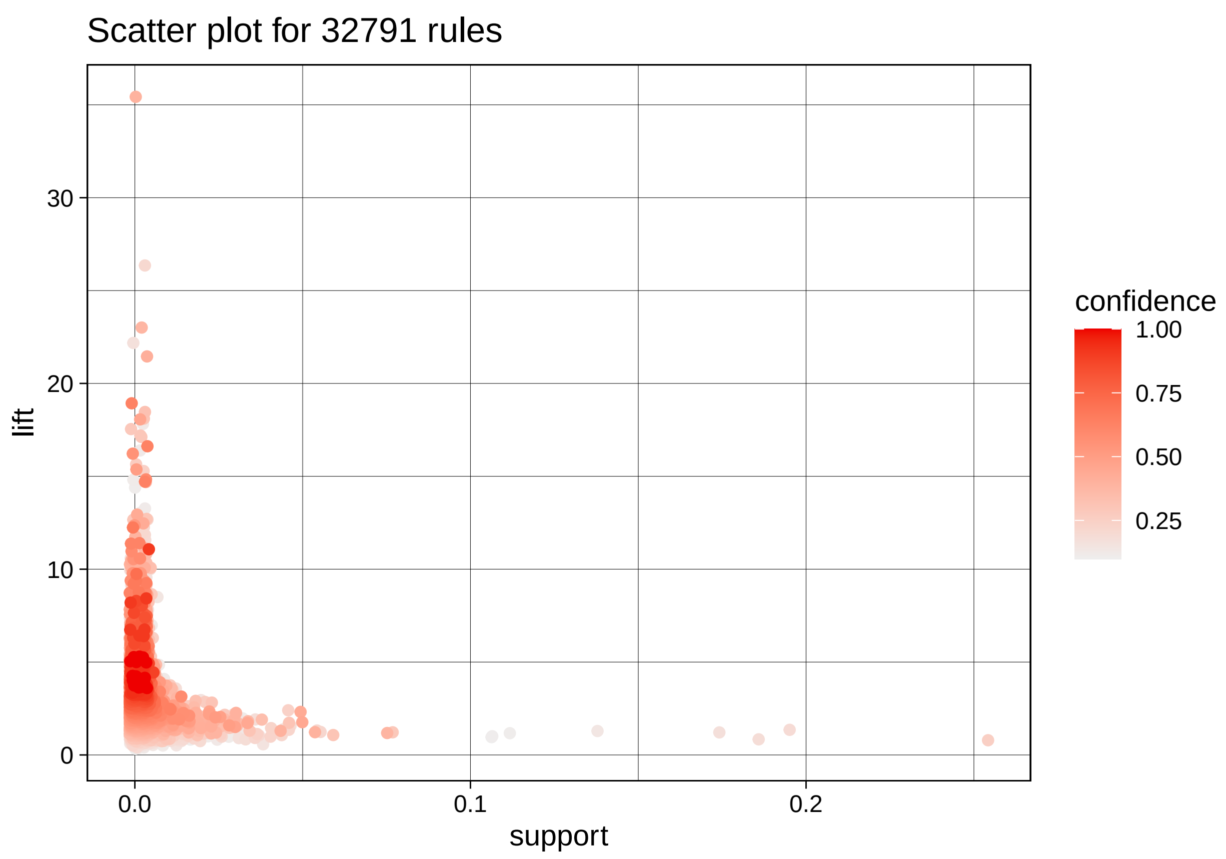


- As the first step in cleaning the data, the duplicates were removed and items in each row are separated by a comma.The summary statistics of the data set are shown above. We can see that the data set has the info of 9835 different shopping baskets. The plot shows the most frequently bought items which include whole milk and vegetables followed by rolls/buns, soda and yogurt.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5   0.001      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 9
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
## sorting and recoding items ... [157 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 done [0.01s].
## writing ... [32791 rule(s)] done [0.01s].
## creating S4 object  ... done [0.02s].
## set of 32791 rules
##
## rule length distribution (lhs + rhs):sizes
##     1     2     3     4     5     6
##     8  2121 16468 12254  1880    60
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.429   4.000   6.000
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.001017   Min.   :0.1000   Min.   :0.001017   Min.   : 0.4193
##  1st Qu.:0.001118   1st Qu.:0.1692   1st Qu.:0.003457   1st Qu.: 1.9568
##  Median :0.001423   Median :0.2609   Median :0.005897   Median : 2.5771
##  Mean   :0.002070   Mean   :0.3118   Mean   :0.009387   Mean   : 2.8378
##  3rd Qu.:0.002034   3rd Qu.:0.4167   3rd Qu.:0.009761   3rd Qu.: 3.3979
##  Max.   :0.255516   Max.   :1.0000   Max.   :1.000000   Max.   :35.7158
##      count
##  Min.   :  10.00
##  1st Qu.:  11.00
##  Median :  14.00
##  Mean   :  20.36
##  3rd Qu.:  20.00
##  Max.   :2513.00
##
## mining info:
##       data ntransactions support confidence
##  groceries          9835   0.001        0.1
##                                                                                call
##  apriori(data = groceries, parameter = list(supp = 0.001, conf = 0.1, maxlen = 10))
```

- The next step after the cleaning of data is to find the association rules in the data set. I used the apriori function with support = 0.001, confidence = 0.1 and max length = 10. As we can see in the output, a total of 32791 association rules were formed with these parameters.

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



- In order to find strong associations, we need the association rules which have high lift and high confidence. Based on the above plot, selecting the association rules which have lift greater than 5 and confidence greater than 0.5.This gives us 455 such associations. below are the first 10 of these 455 associations.

```
##      lhs                          rhs                   support confidence    coverage      lift co
## [1]  {liquor,
##       red/blush wine}          => {bottled beer}      0.001931876  0.9047619 0.002135231 11.235269
## [2]  {liquor,
##       soda}                    => {bottled beer}      0.001220132  0.5714286 0.002135231  7.095960
## [3]  {jam,
##       other vegetables}        => {root vegetables}   0.001016777  0.5555556 0.001830198  5.096911
## [4]  {instant coffee,
##       other vegetables}        => {whipped/sour cream} 0.001016777  0.5263158 0.001931876  7.342292
## [5]  {popcorn,
##       soda}                    => {salty snack}       0.001220132  0.6315789 0.001931876 16.697793
## [6]  {dog food,
##       yogurt}                  => {tropical fruit}    0.001016777  0.5263158 0.001931876  5.015810
## [7]  {Instant food products,
##       soda}                    => {hamburger meat}    0.001220132  0.6315789 0.001931876 18.995654
## [8]  {tropical fruit,
##       turkey}                  => {root vegetables}   0.001525165  0.5769231 0.002643620  5.292946
## [9]  {root vegetables,
##       turkey}                  => {tropical fruit}    0.001525165  0.6000000 0.002541942  5.718023
## [10] {butter,
##       rice}                    => {root vegetables}   0.001016777  0.5555556 0.001830198  5.096911
```

We can see that almost all of associations make sense. The first association in table shows that people who buy liquor are very likely to buy bottled beer. These two items definitely go together. Jam,other vegetables and root vegetables are also consumed together. Likewise, the association between popcorn, soda and salty snack makes sense.

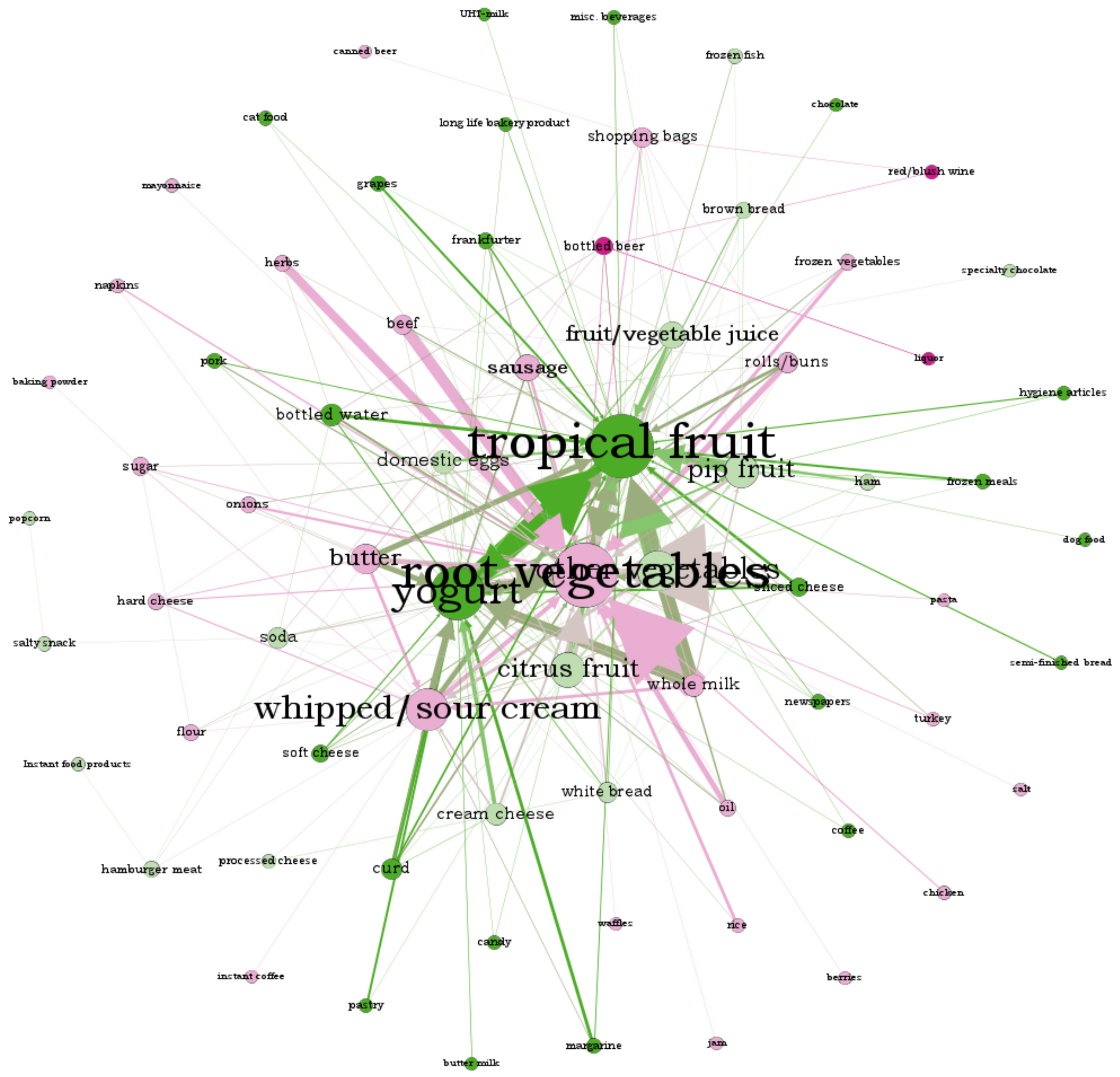


Figure 3: Association rule mining graph for groceries dataset using Gephy

The above can help us visualize the item sets. We can see other vegetables, butter milk, hamburger meat, onions form a set, which does make sense. Likewise, items like margarine, roll buns and frankfurter which are consumed together make a set. Similarly, we can see that items like citrus fruit, yogurt and cream cheese are grouped together. From an association point of view, it also makes perfect sense.