

R Exam

Rahul Singla | Vishal Gupta | Prakhar Bansal | Ramya Madhuri Desineedi

Git hub link - [https://github.com/RahulSingla5209/data\\_viz\\_unsupervised\\_algo](https://github.com/RahulSingla5209/data_viz_unsupervised_algo)

## Question 1

```

## -- Attaching packages ----- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr    1.0.7
## v tidyr   1.1.3     v stringr  1.4.0
## v readr   1.4.0     vforcats  0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()

##   CS_PropertyID      cluster       size      empl_gr
## Min. : 1             Min. : 1.0     Min. : 1624   Min. :-24.950
## 1st Qu.: 157452      1st Qu.: 272.0   1st Qu.: 50891   1st Qu.: 1.740
## Median : 313253      Median : 476.0   Median : 128838   Median : 1.970
## Mean   : 453003      Mean   : 588.6   Mean   : 234638   Mean   : 3.207
## 3rd Qu.: 441188      3rd Qu.:1044.0   3rd Qu.: 294212   3rd Qu.: 2.380
## Max.   :6208103      Max.   :1230.0   Max.   :3781045   Max.   : 67.780
##
##          Rent        leasing_rate      stories      age
## Min. : 2.98      Min. : 0.00      Min. : 1.00      Min. : 0.00
## 1st Qu.: 19.50    1st Qu.: 77.85    1st Qu.: 4.00    1st Qu.: 23.00
## Median : 25.16    Median : 89.53    Median : 10.00   Median : 34.00
## Mean   : 28.42    Mean   : 82.61    Mean   : 13.58   Mean   : 47.24
## 3rd Qu.: 34.18    3rd Qu.: 96.44    3rd Qu.: 19.00   3rd Qu.: 79.00
## Max.   :250.00    Max.   :100.00    Max.   :110.00   Max.   :187.00
##
##      renovated      class_a      class_b      LEED
## Min. :0.00000      Min. :0.00000      Min. :0.00000      Min. :0.0000000
## 1st Qu.:0.00000      1st Qu.:0.00000      1st Qu.:0.00000      1st Qu.:0.0000000
## Median :0.00000      Median :0.00000      Median :0.00000      Median :0.0000000
## Mean   :0.3795      Mean   :0.3999      Mean   :0.4595      Mean   :0.006841
## 3rd Qu.:1.0000      3rd Qu.:1.0000      3rd Qu.:1.0000      3rd Qu.:0.0000000
## Max.   :1.0000      Max.   :1.0000      Max.   :1.0000      Max.   :1.0000000
##
##      Energystar      green_rating      net      amenities
## Min. :0.00000      Min. :0.00000      Min. :0.00000      Min. :0.00000
## 1st Qu.:0.00000      1st Qu.:0.00000      1st Qu.:0.00000      1st Qu.:0.00000

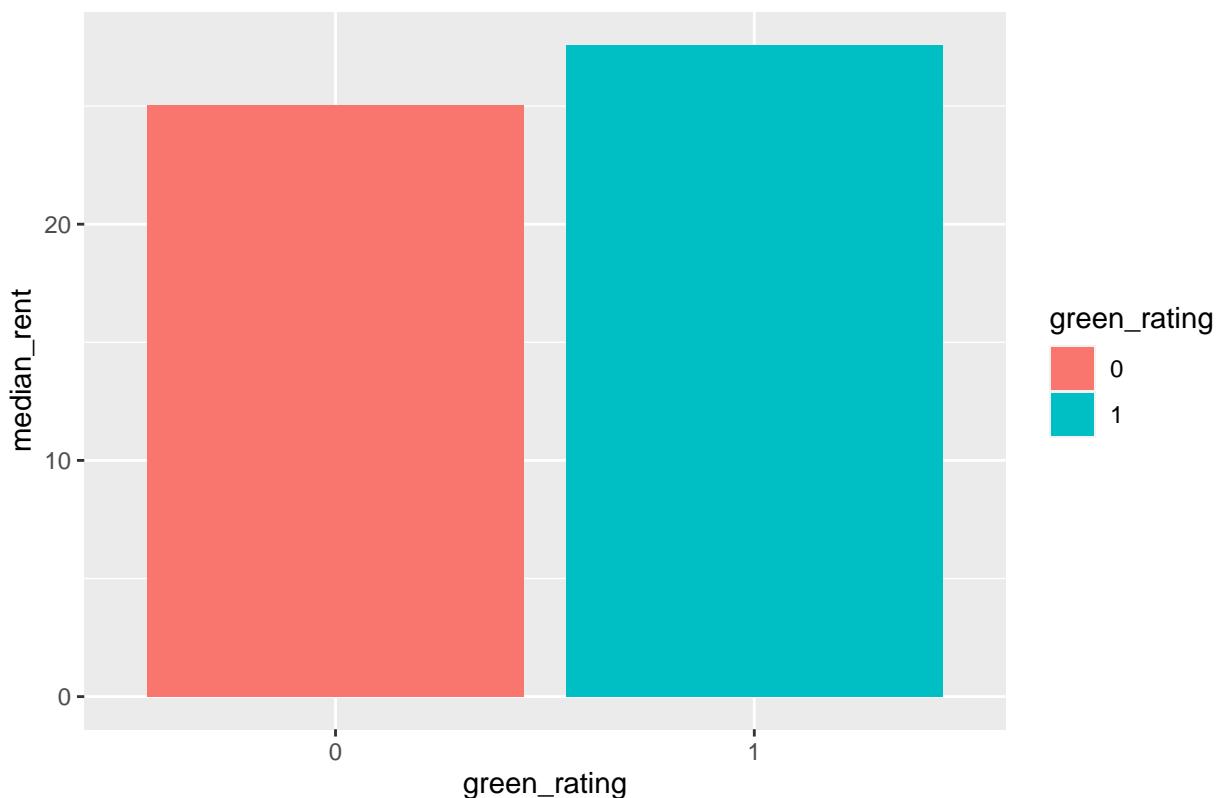
```

```

## Median :0.00000  Median :0.00000  Median :0.00000  Median :1.0000
## Mean   :0.08082  Mean   :0.08677  Mean   :0.03471  Mean   :0.5266
## 3rd Qu.:0.00000  3rd Qu.:0.00000  3rd Qu.:0.00000  3rd Qu.:1.0000
## Max.   :1.00000  Max.   :1.00000  Max.   :1.00000  Max.   :1.0000
##
##   cd_total_07      hd_total07      total_dd_07      Precipitation
## Min.   : 39       Min.   : 0       Min.   :2103       Min.   :10.46
## 1st Qu.: 684     1st Qu.:1419    1st Qu.:2869     1st Qu.:22.71
## Median : 966     Median :2739    Median :4979     Median :23.16
## Mean   :1229     Mean   :3432    Mean   :4661     Mean   :31.08
## 3rd Qu.:1620     3rd Qu.:4796    3rd Qu.:6413     3rd Qu.:43.89
## Max.   :5240     Max.   :7200    Max.   :8244     Max.   :58.02
##
##   Gas_Costs      Electricity_Costs  cluster_rent
## Min.   :0.009487  Min.   :0.01780  Min.   : 9.00
## 1st Qu.:0.010296  1st Qu.:0.02330  1st Qu.:20.00
## Median :0.010296  Median :0.03274  Median :25.14
## Mean   :0.011336  Mean   :0.03096  Mean   :27.50
## 3rd Qu.:0.011816  3rd Qu.:0.03781  3rd Qu.:34.00
## Max.   :0.028914  Max.   :0.06280  Max.   :71.44
##
## [1] 7894 23
## [1] 7679 23

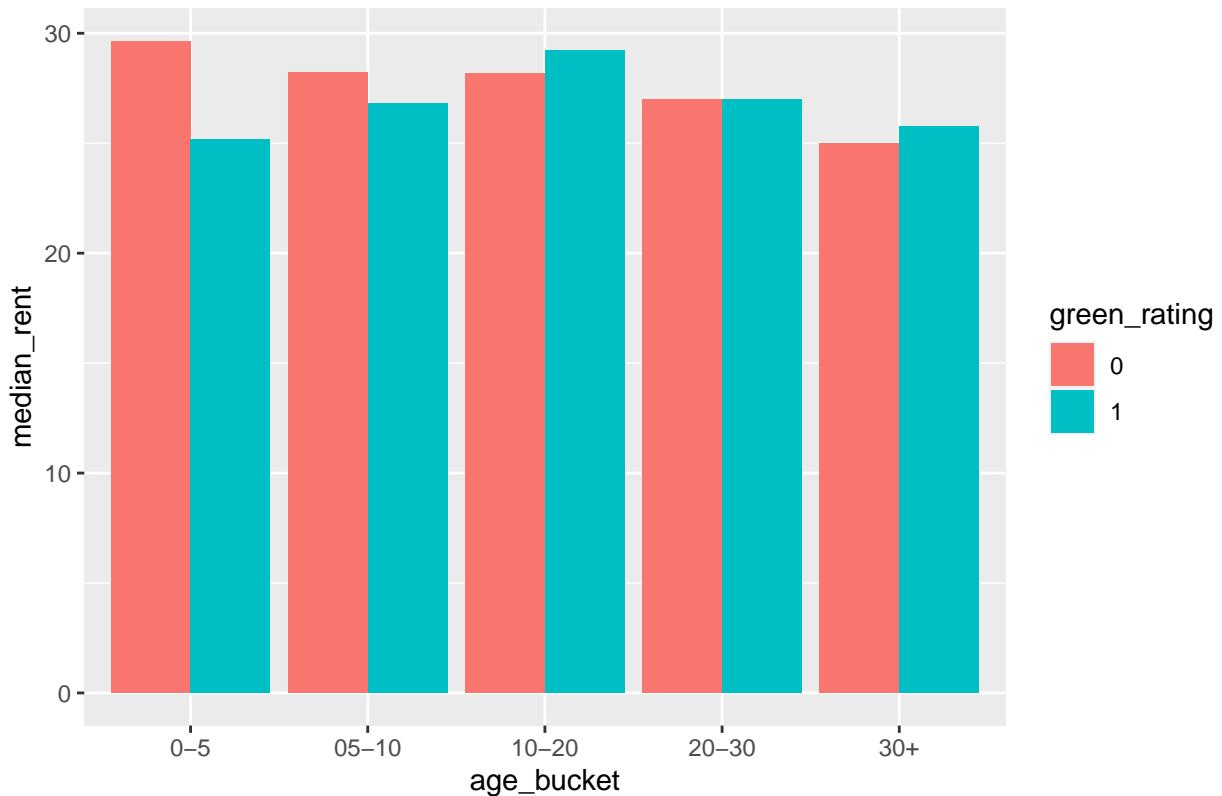
```

### Median rent of green and non-green buildings



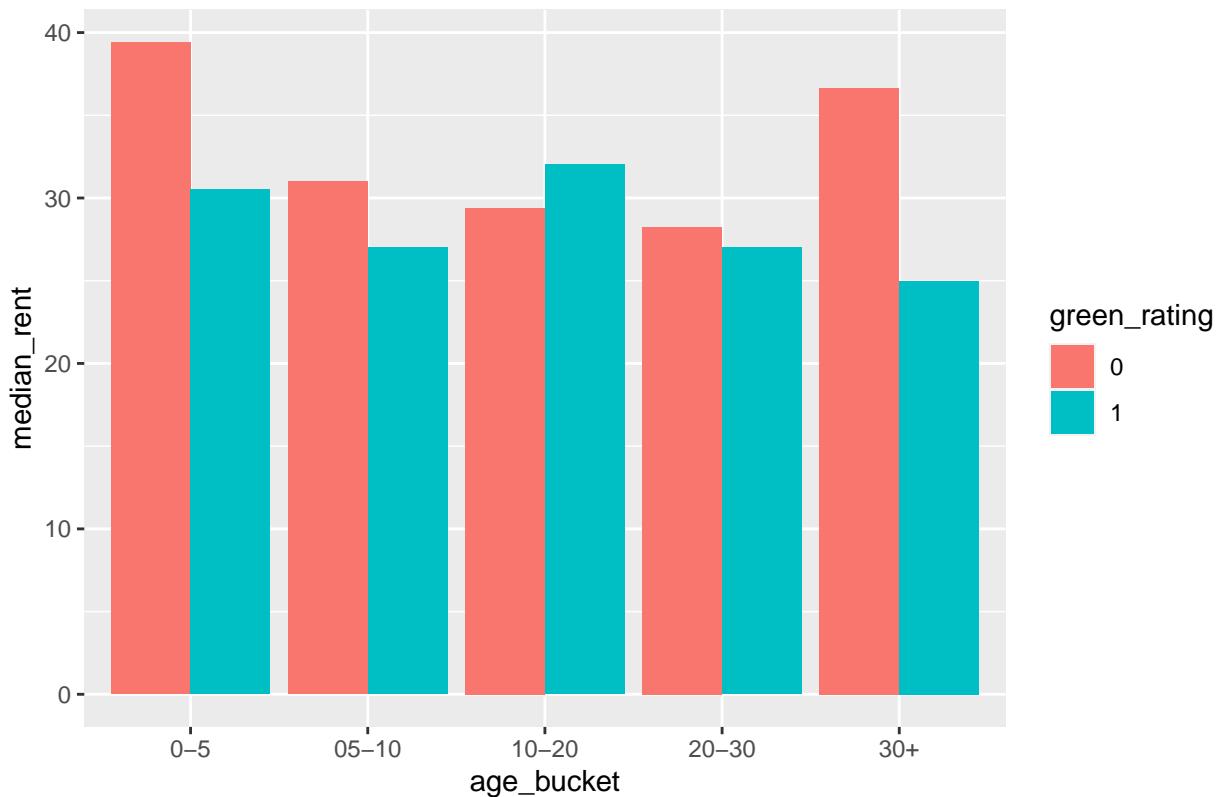
```
## `summarise()` has grouped output by 'age_buckets'. You can override using the '.groups' argument.
```

Median rent of non-renovated green and non-green buildings by age

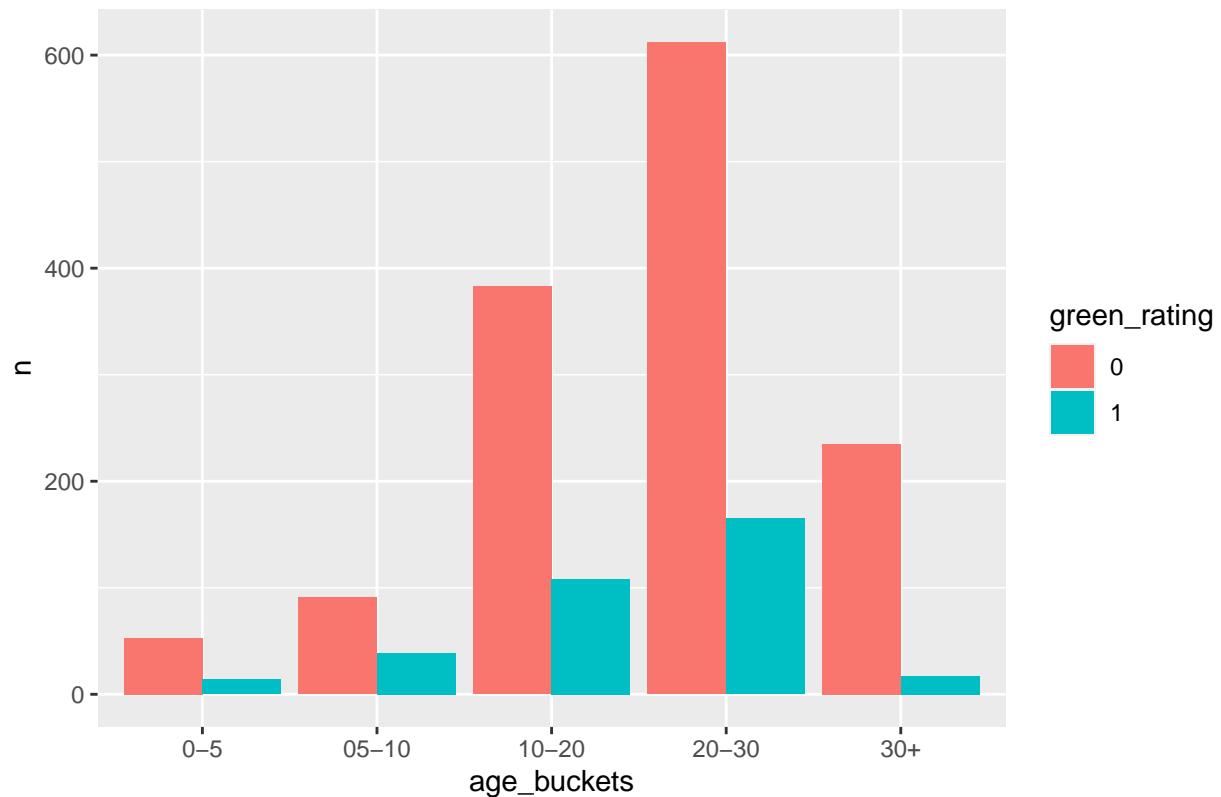


```
## `summarise()` has grouped output by 'age_buckets'. You can override using the '.groups' argument.
```

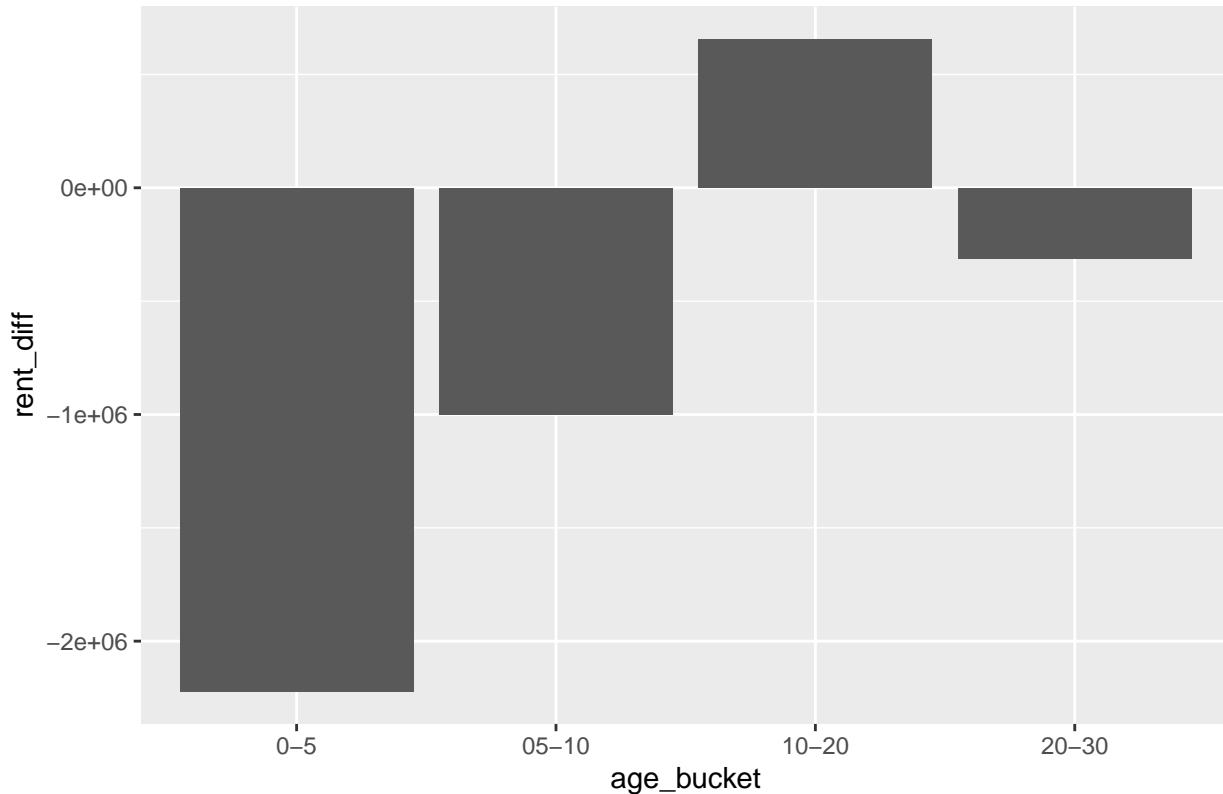
Median rent of non-renovated, class A and with amenities green and non-green



## Count rent of non-renovated, class A and with amenities green and non-green



## Different in rent of green vs non green buildings by different



over the lifetime of the building, assumed to be 30 years, green building won't be able to break even before the non-green building, as overall expected rent diff (based on the median rent of the subset) is -2.88 million dollars.

There are some variables other than green rating that are affecting the rent of the building.

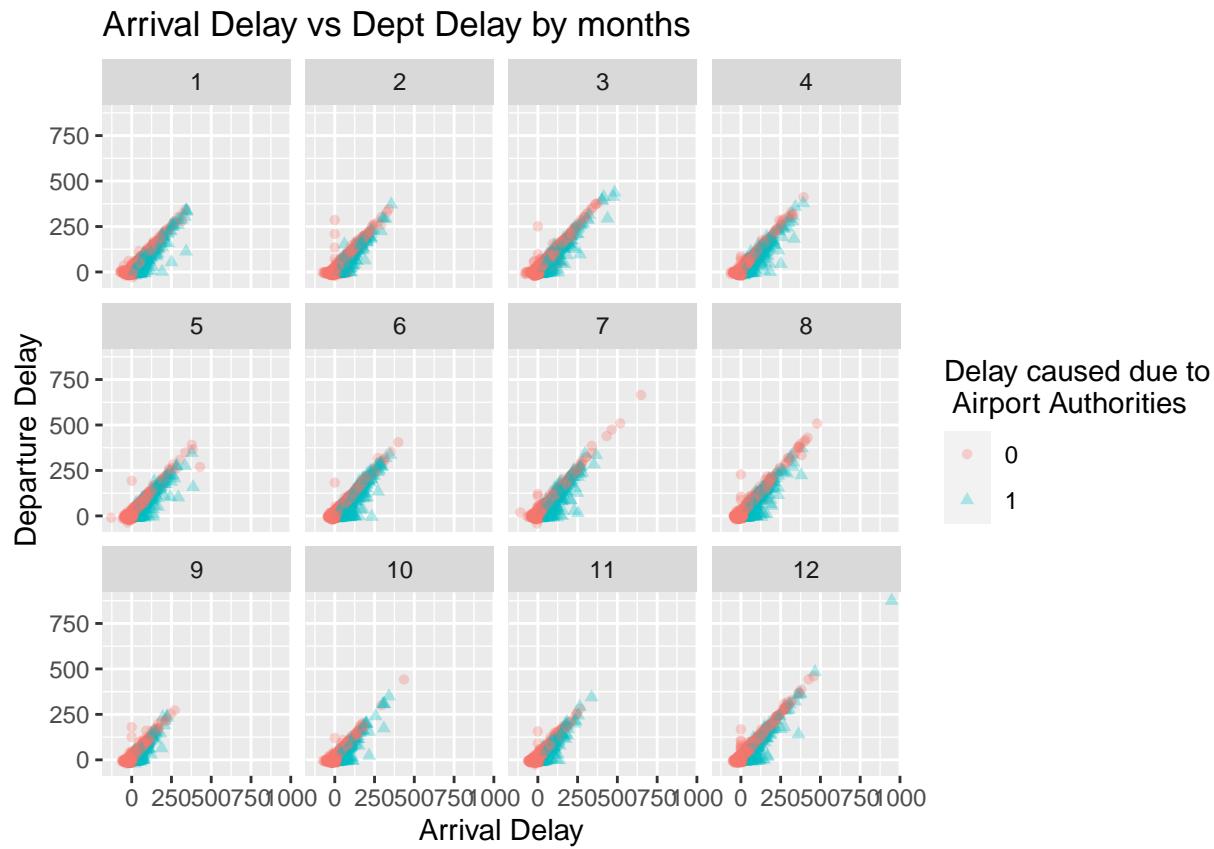
Namely, Class\_a, Age, Amenties, and Renovation variables are found to be other factors affecting the rent of the building.

Having the building in class\_a area with amenities, affect the rent of green building.

And since this is a new building, we need to filter for non-renovated buildings.

Only in 10-20 year range, The rent of new green buldings is higher than new non-green building in class\_a area with amenities.

## Question 2



we have plotted scatter plot for arrival delay vs departure delay. Facetted by months. Colored by Delay not caused due to airlines or airport authorities.

### insights

The delay count and delay times seems to be increased during vacation times - namely December, June, and July.

Delay caused due to authorities is increased in June and July.

## Question 3 -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 selected 4 ETFs to ensure diversity and different levels of risk for the portfolio. Below are the details:  
*GOVT - Bond ETF to minimize our risks*

*ISCV - Small Cap Value stock to expose us to the equity but with minimal risks*

*VT - Large Cap Growth stock with the maximum holding to increase our returns*

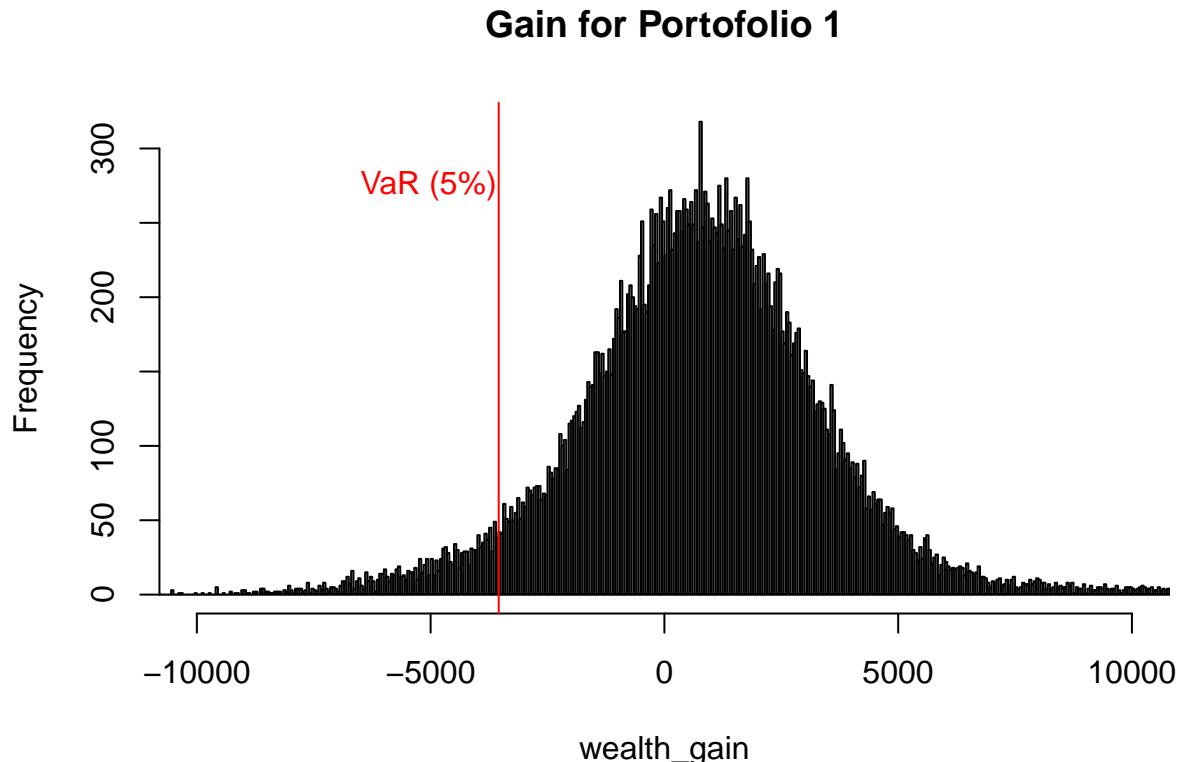
*IWS - Mid Cap Value stock with the maximum holding find a balance between growth and stability*

We have come up with 3 portfolios which suits 3 different age groups based on their risk tolerance. (Risk - low, medium, high)

## Portofolio 1 - Low Risk

Customers in their retirement period would want to have a stable income hence would not be interested to invest in high risk investments.

In this low risk portfolio, we gave maximum weightage(50%) in Govt. bonds, followed by Large Cap(25%) & Mid(25%) and least weightage to Small cap (5%).

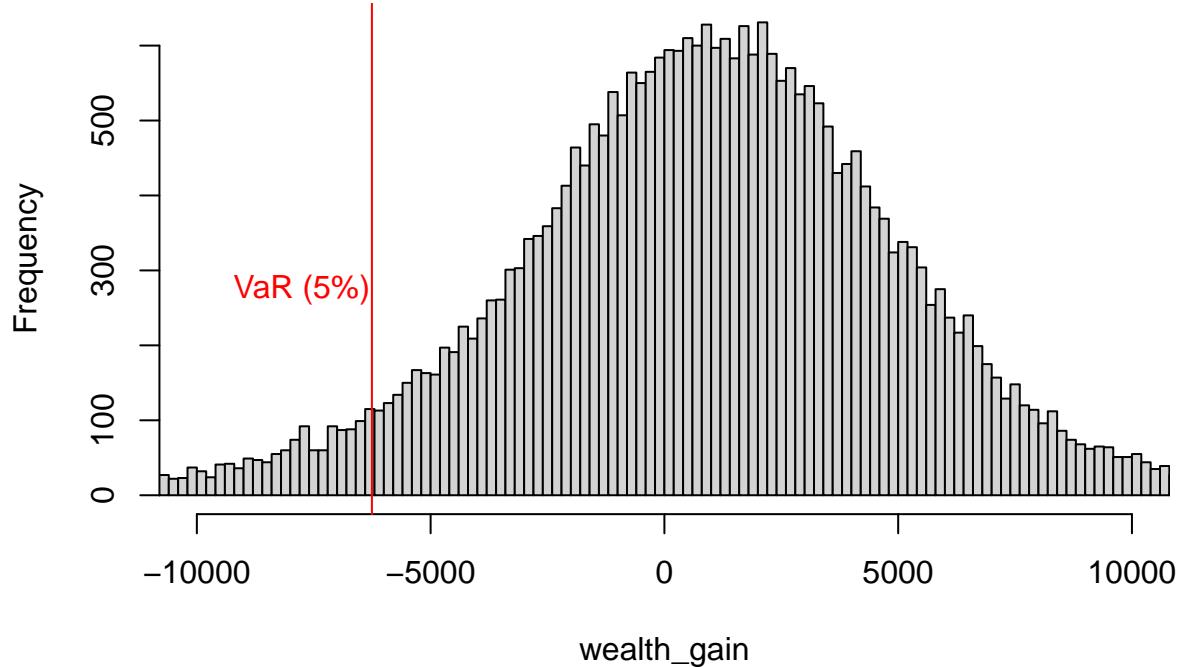


```
##  
## Average return of investement after 20 days for Low risk portfolio 100844  
  
##  
## 5% Value at Risk for Low risk portfolio- -3543.938  
  
-3543.9379427
```

**Portofolio 2 - Medium Risk** Customers in their mid age group (40-60 years) have moderate risk appetite and hence would be interested to invest equally in low & high risk investments.

In this medium risk portfolio, we gave equal weightage to all ETFs: Govt. bonds(25%) Large Cap(25%) & Mid(25%) and Small cap (25%).

## Gain for Portofolio 2



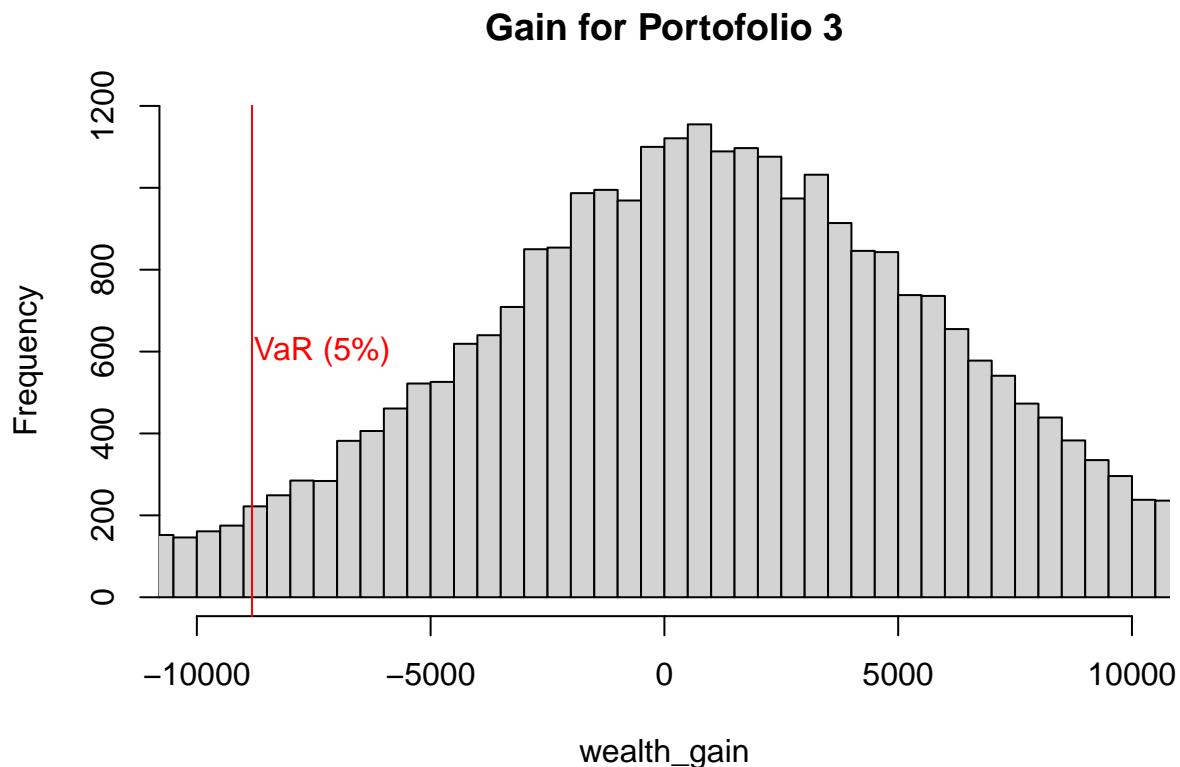
```
##  
## Average return of investement after 20 days for Medium risk portfolio 101685.3  
  
##  
## 5% Value at Risk for Medium risk portfolio- -6255.535  
  
-6255.5354385
```

## Portofolio 3 - High Risk

young customers would invest heavily in small cap and medium cap ETFs. And will have a minimal exposure in Bonds.

Customers in their youth have high risk appetite and hence would be interested to invest more in high risk investments.

In this high risk portfolio, we gave minimum weightage(5%) in Govt. bonds, Large Cap(50%) & Mid(15%) and Small cap (30%).



```

## 
## Average return of investement after 20 days for High risk portfolio 102412.3

## 
## 5% Value at Risk for High risk portfolio- -8820.138

-8820.1381177

```

**Summary** For the High risk portfolio, we are observing the maximum return of investment (\$102390.2) and the lowest 5% VaR(-\$8749.14). As the portfolio risk increases, we are able to witness the decrease in returns and increase in VaR value as expected.

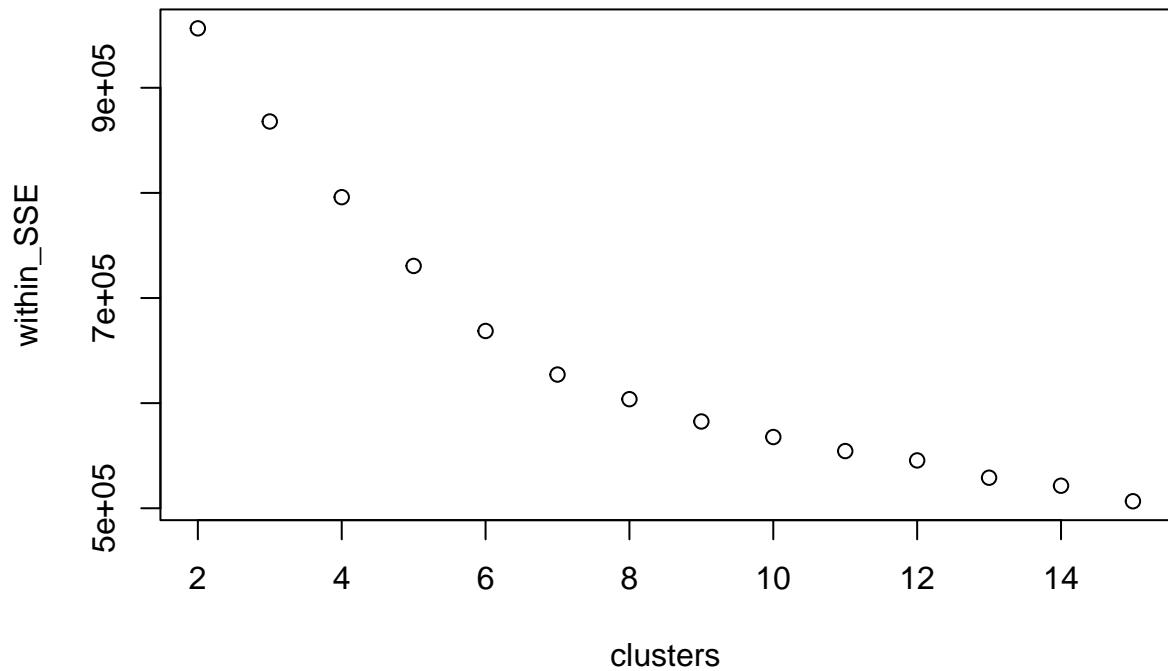
## Question 4 - Market segmentation

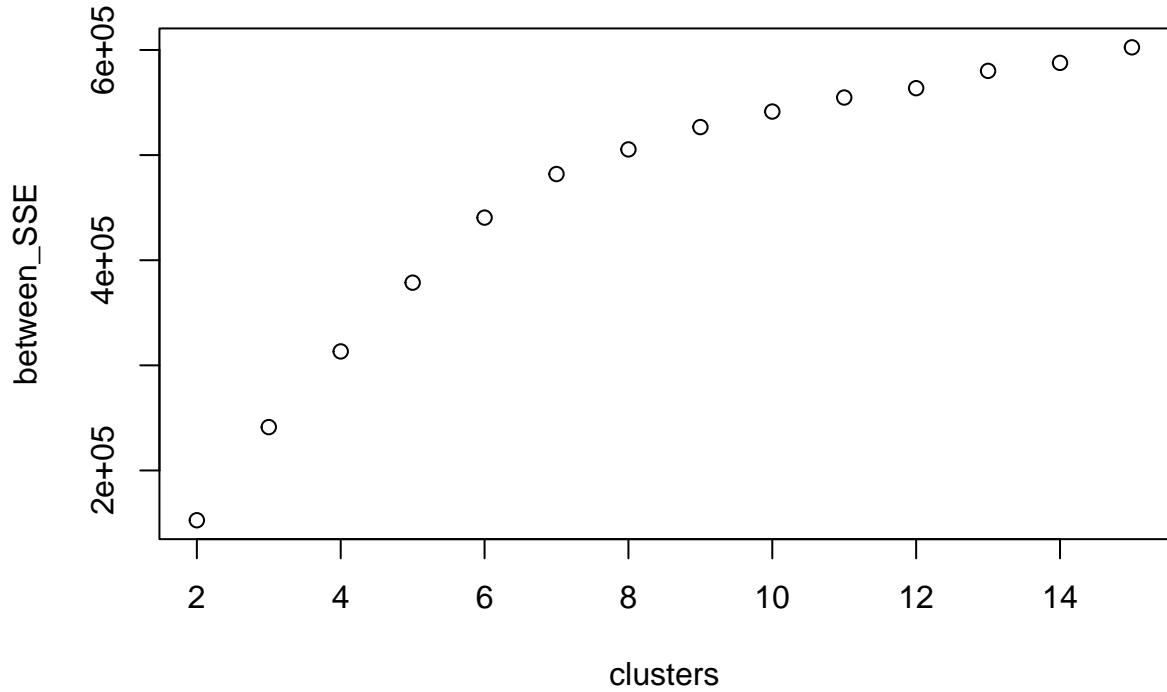
**K-Mean clustering** \*K-Means works better when the number of samples is high

\*scaling is not required because the unit of all the columns is same.

\*We will explore different K and different distance metrics, starting with K = 2 and Euclidean distance metric.

number of followers - 7882





```

##      chatter current_events   travel photo_sharing uncategorized   tv_film
## 1 -0.08868426    0.71841989  0.1232091    0.71557797 -0.2309572  0.6764819
## 2  5.50651991    0.47553995  0.7917713    1.78522778  0.3704899  0.9890781
## 3  0.64285961   -0.08586167  1.3854028    0.08999201 -0.3776038  0.1971119
## 4  0.47298582   -0.08222480  0.7858856   -0.00702703  0.6006889  0.6799747
##      sports_fandom   politics     food    family home_and_garden       music
## 1    0.5904936  0.41091153  0.267310 -0.9221228   -0.2584652 -0.32684509
## 2    0.6260885  0.82745476  1.641534 -0.3789840   -0.2777545  0.05799701
## 3    0.2393073  0.80204191 14.435906 -0.3600817   -0.5249874  0.04242885
## 4    0.6112907  0.04307279  1.026870 -0.6538126   -0.3824259  0.10737656
##      news online_gaming     shopping health_nutrition college_uni
## 1   -0.05592945  2.9791671  0.479351851   0.54918523  6.3962789
## 2    0.02092842  0.1581264  3.945788574   0.38428947  0.3156359
## 3   -0.11994078 -0.4151491  0.005811054   -0.08284061  0.1141709
## 4   -0.28034858  0.5421898  0.996939360   8.56158891 -0.1655863
##      sports_playing   cooking      eco computers business outdoors
## 1    0.7932445 -0.6713876 -0.3501236 -0.41202485 -0.3782863 -0.31043114
## 2    0.5540270  2.0111639 -0.2291664 -0.01949209 -0.2116605 -0.01935319
## 3    5.2891988 -0.3208730 -0.5648810  0.08143519 -0.5805220 -0.44139980
## 4    0.1484944  0.3929663 -0.3075304 -0.03743274 -0.4691700  2.46290734
##      crafts automotive         art religion   beauty parenting
## 1   -0.02591343  0.08653198  0.19640360  0.03767655 -1.06476038  0.182942139
## 2    0.36824648  0.24946818  0.02857516  1.35565269 -0.02893089  0.006215571
## 3   -0.30076300 -0.18822232 -0.10634196 12.68527902 -0.50210843 -0.348864281
## 4    0.15859074  0.01505361 -0.26309062  0.39886759 -0.73864016 -0.259250626
##      dating school personal_fitness fashion small_business      spam

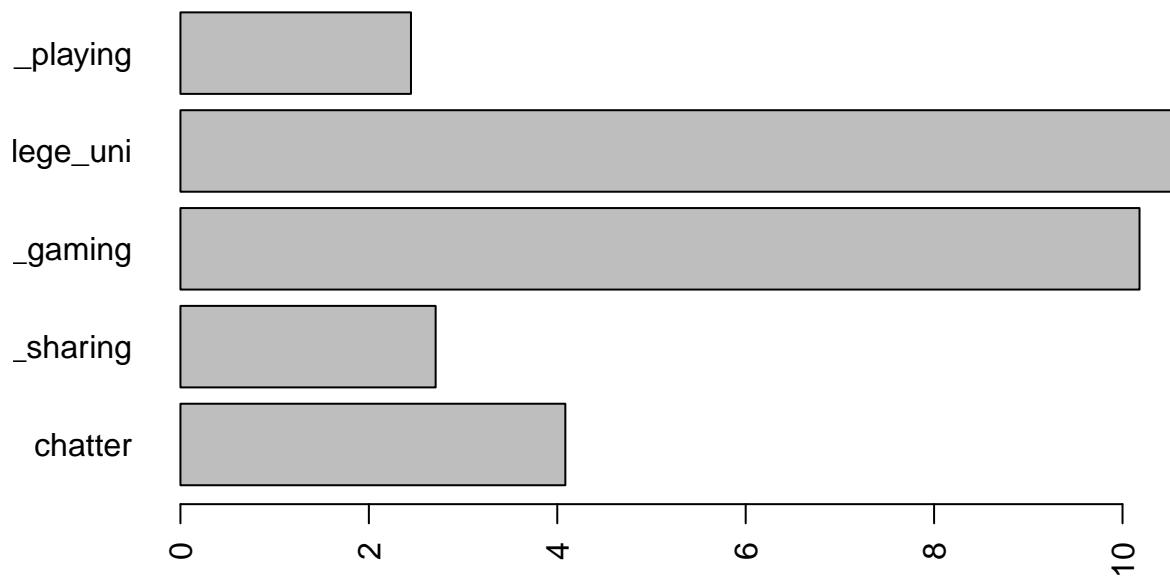
```

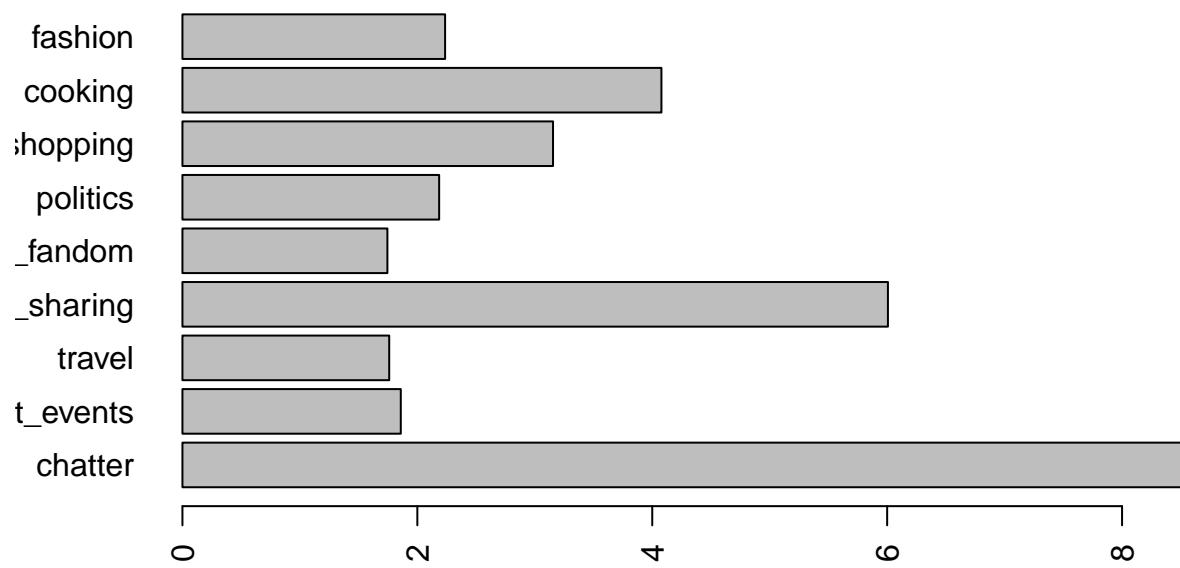
```

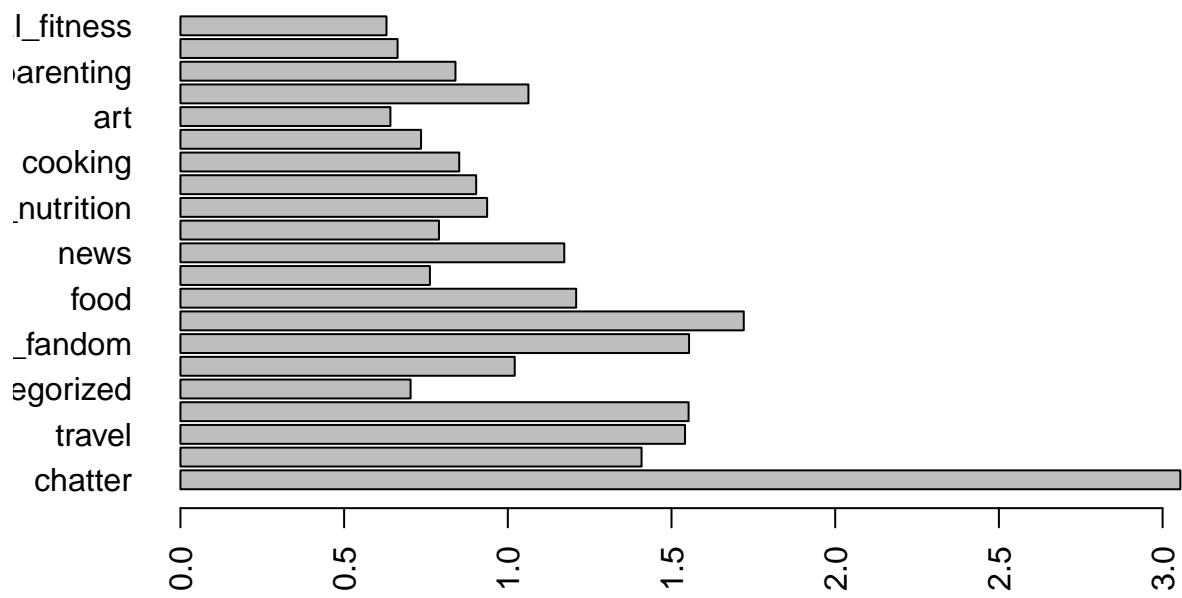
## 1 -0.35862858 -0.24826904      -0.1725205  0.4161937      -0.2935046 -0.6020426
## 2  0.22457222 -0.04149517      0.4029258  2.6187610      -0.1516985 -0.3974940
## 3 -0.03616171 -0.40142016     -0.3990760 -0.2025607      -0.4223298 -0.6398003
## 4  0.31684914 -0.40532539      7.1013913  0.6064756      -0.3092033 -0.6047456
##          adult
## 1 -0.31665025
## 2  0.02765828
## 3  5.13074792
## 4 -0.03411253

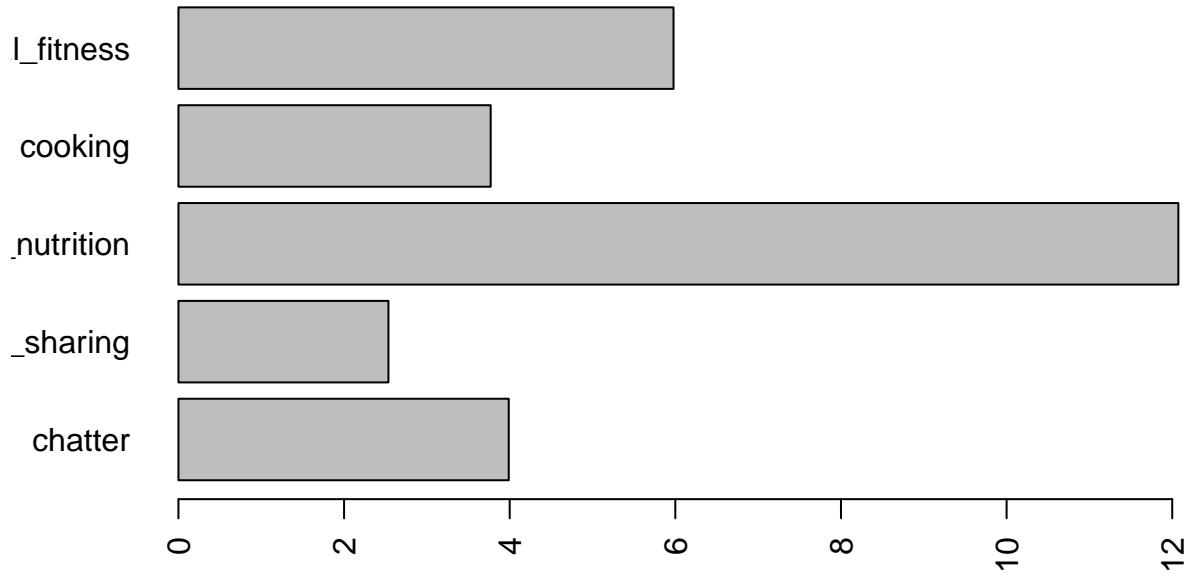
##          1         2         3         4
## chatter    4.085778781 8.51343284 3.054278416 3.98874296
## current_events 1.485327314 1.85910448 1.408471690 1.53939962
## travel     1.616252822 1.76059701 1.541294168 1.48874296
## photo_sharing 2.708803612 6.00656716 1.552149851 2.53564728
## uncategorized 0.880361174 1.00059701 0.703065134 0.97467167
## tv_film     1.446952596 1.10805970 1.021072797 1.07129456
## sports_fandom 1.636568849 1.74507463 1.553639847 1.51688555
## politics    1.544018059 2.18567164 1.720732226 1.56566604
## food        1.485327314 1.35104478 1.208812261 2.26547842
## family      1.144469526 1.04537313 0.762026394 0.91088180
## home_and_garden 0.571106095 0.60955224 0.459557258 0.62945591
## music       0.817155756 0.92955224 0.551936994 0.78986867
## news        1.088036117 1.26507463 1.172413793 1.30675422
## online_gaming 10.180586907 0.79343284 0.574286931 0.92964353
## shopping    1.214446953 3.15522388 0.789697744 1.33020638
## health_nutrition 1.580135440 1.35104478 0.936781609 12.07410882
## college_uni   10.613995485 1.27343284 0.903363133 1.06378987
## sports_playing 2.446952596 0.67880597 0.446998723 0.67260788
## cooking      2.029345372 4.07820896 0.851638995 3.77016886
## eco          0.485327314 0.69014925 0.373350362 0.85647280
## computers    0.665914221 0.84179104 0.580672627 0.64071295
## business     0.410835214 0.64000000 0.339293316 0.45778612
## outdoors     0.650112867 0.62029851 0.484248616 2.40806754
## crafts        0.618510158 0.67820896 0.418901660 0.64540338
## automotive   0.948081264 1.13134328 0.734993614 0.72514071
## art           1.218961625 0.76179104 0.641336739 0.82926829
## religion     1.065462754 1.17432836 1.063005534 1.12664165
## beauty        0.641083521 1.48955224 0.437420179 0.67917448
## parenting    0.984198646 1.08059701 0.840144742 1.00281426
## dating        0.760722348 1.11820896 0.494040017 1.00562852
## school        0.683972912 1.09731343 0.663260962 0.74484053
## personal_fitness 1.049661400 1.03223881 0.629416773 5.97842402
## fashion       1.139954853 2.23641791 0.537675607 1.01125704
## small_business 0.428893905 0.47761194 0.286717752 0.29455910
## spam          0.009029345 0.00238806 0.007449979 0.00750469
## adult         0.417607223 0.35343284 0.433801618 0.34146341

```









```
## [1] 443
```

```
## [1] 1675
```

```
## [1] 4698
```

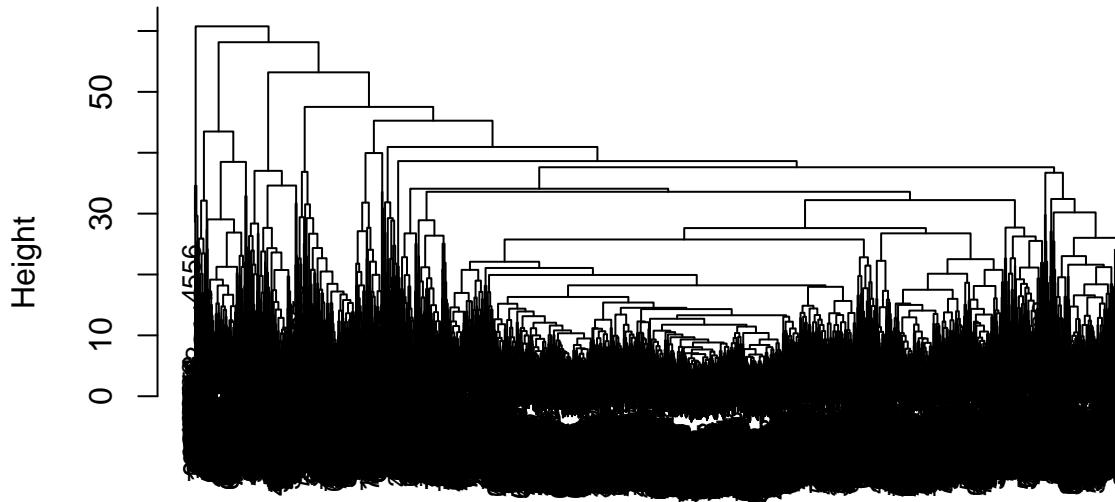
```
## [1] 1066
```

Profiling Methodology: \* We have ran iterations with k=2,15 and based on the results we got, we finalized on clustering into 4 groups \* For these clusters, we have found the cluster means through k-means algorithm and assigned tweets based on  $0.2 * \text{max}(\text{feature value of these clusters})$  to a cluster

Insights from clustering: KMeans clustering has provided good insights from data.Below are groups identified- Cluster 1. Teen girls? - High shopping, fashion, school, chatter, photosharing Cluster 2. Fitness enthusiasts - high - Outdoors, personal\_fitness , health\_nutrition Cluster 3. Middle aged group - religion, adult, food Cluster 4. College students - high - college\_uni, online\_gaming, sport\_playing

How these insights can help NutrientH20? \* If we define Cluster 2 as fitness enthusiasts and Cluster 4 as College students, we can do focussed marketing on these group and strategise marketing campaigns to improve revenue from these nische groups and work on their retention

## Cluster Dendrogram



`dist(df, method = "euclidean")`  
`hclust (*, "complete")`

**Heirarchical Clustering**

```
##      1     2     3     4
##  417 7014  420   31
```

H-clustering is biased for one cluster and is not giving any useful insights.

## Question 5 - Author attribution

```
## Importance of first k=200 (out of 793) components:
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation 3.95900 3.31200 2.94949 2.89421 2.74722 2.66560 2.54703
## Proportion of Variance 0.01977 0.01383 0.01097 0.01056 0.00952 0.00896 0.00818
## Cumulative Proportion 0.01977 0.03360 0.04457 0.05513 0.06465 0.07361 0.08179
##          PC8      PC9      PC10     PC11     PC12     PC13     PC14
## Standard deviation 2.38025 2.32864 2.24640 2.21198 2.18396 2.13510 2.08190
## Proportion of Variance 0.00714 0.00684 0.00636 0.00617 0.00601 0.00575 0.00547
## Cumulative Proportion 0.08893 0.09577 0.10214 0.10831 0.11432 0.12007 0.12553
##          PC15     PC16     PC17     PC18     PC19     PC20     PC21
## Standard deviation 2.0507 1.99909 1.99246 1.96315 1.91945 1.91681 1.87148
## Proportion of Variance 0.0053 0.00504 0.00501 0.00486 0.00465 0.00463 0.00442
## Cumulative Proportion 0.1308 0.13588 0.14088 0.14574 0.15039 0.15502 0.15944
##          PC22     PC23     PC24     PC25     PC26     PC27     PC28
## Standard deviation 1.86112 1.85750 1.83292 1.81635 1.78586 1.77879 1.76050
## Proportion of Variance 0.00437 0.00435 0.00424 0.00416 0.00402 0.00399 0.00391
## Cumulative Proportion 0.16381 0.16816 0.17239 0.17656 0.18058 0.18457 0.18848
```

```

##          PC29   PC30   PC31   PC32   PC33   PC34   PC35
## Standard deviation 1.75281 1.7367 1.72458 1.71488 1.69654 1.68775 1.68195
## Proportion of Variance 0.00387 0.0038 0.00375 0.00371 0.00363 0.00359 0.00357
## Cumulative Proportion 0.19235 0.1961 0.19990 0.20361 0.20724 0.21083 0.21440
##          PC36   PC37   PC38   PC39   PC40   PC41   PC42
## Standard deviation 1.66663 1.65774 1.65524 1.63988 1.63183 1.62287 1.60960
## Proportion of Variance 0.0035 0.00347 0.00346 0.00339 0.00336 0.00332 0.00327
## Cumulative Proportion 0.2179 0.22137 0.22482 0.22821 0.23157 0.23489 0.23816
##          PC43   PC44   PC45   PC46   PC47   PC48   PC49
## Standard deviation 1.60724 1.59563 1.5934 1.58935 1.58211 1.57644 1.5673
## Proportion of Variance 0.00326 0.00321 0.0032 0.00319 0.00316 0.00313 0.0031
## Cumulative Proportion 0.24142 0.24463 0.2478 0.25102 0.25417 0.25731 0.2604
##          PC50   PC51   PC52   PC53   PC54   PC55   PC56
## Standard deviation 1.56191 1.56014 1.55515 1.54520 1.54382 1.54015 1.53026
## Proportion of Variance 0.00308 0.00307 0.00305 0.00301 0.00301 0.00299 0.00295
## Cumulative Proportion 0.26348 0.26655 0.26960 0.27261 0.27562 0.27861 0.28156
##          PC57   PC58   PC59   PC60   PC61   PC62   PC63
## Standard deviation 1.52585 1.51380 1.51198 1.50956 1.50201 1.49820 1.49512
## Proportion of Variance 0.00294 0.00289 0.00288 0.00287 0.00284 0.00283 0.00282
## Cumulative Proportion 0.28450 0.28739 0.29027 0.29314 0.29599 0.29882 0.30164
##          PC64   PC65   PC66   PC67   PC68   PC69   PC70
## Standard deviation 1.48848 1.48584 1.48383 1.47971 1.47214 1.46834 1.4637
## Proportion of Variance 0.00279 0.00278 0.00278 0.00276 0.00273 0.00272 0.0027
## Cumulative Proportion 0.30443 0.30721 0.30999 0.31275 0.31548 0.31820 0.3209
##          PC71   PC72   PC73   PC74   PC75   PC76   PC77
## Standard deviation 1.45970 1.45767 1.45079 1.44926 1.44440 1.44020 1.43364
## Proportion of Variance 0.00269 0.00268 0.00265 0.00265 0.00263 0.00262 0.00259
## Cumulative Proportion 0.32359 0.32627 0.32893 0.33157 0.33421 0.33682 0.33941
##          PC78   PC79   PC80   PC81   PC82   PC83   PC84
## Standard deviation 1.42993 1.42680 1.42402 1.42251 1.41848 1.41217 1.41160
## Proportion of Variance 0.00258 0.00257 0.00256 0.00255 0.00254 0.00251 0.00251
## Cumulative Proportion 0.34199 0.34456 0.34712 0.34967 0.35220 0.35472 0.35723
##          PC85   PC86   PC87   PC88   PC89   PC90   PC91
## Standard deviation 1.41017 1.40472 1.40211 1.39999 1.39591 1.39485 1.38971
## Proportion of Variance 0.00251 0.00249 0.00248 0.00247 0.00246 0.00245 0.00244
## Cumulative Proportion 0.35974 0.36223 0.36471 0.36718 0.36964 0.37209 0.37452
##          PC92   PC93   PC94   PC95   PC96   PC97   PC98
## Standard deviation 1.38603 1.38221 1.3789 1.37539 1.37377 1.36977 1.36823
## Proportion of Variance 0.00242 0.00241 0.0024 0.00239 0.00238 0.00237 0.00236
## Cumulative Proportion 0.37695 0.37936 0.3817 0.38414 0.38652 0.38889 0.39125
##          PC99   PC100  PC101  PC102  PC103  PC104  PC105
## Standard deviation 1.36445 1.36045 1.35777 1.35492 1.35280 1.3515 1.34606
## Proportion of Variance 0.00235 0.00233 0.00232 0.00232 0.00231 0.0023 0.00228
## Cumulative Proportion 0.39359 0.39593 0.39825 0.40057 0.40288 0.4052 0.40746
##          PC106  PC107  PC108  PC109  PC110  PC111  PC112
## Standard deviation 1.34517 1.33862 1.33736 1.33507 1.32983 1.32892 1.32268
## Proportion of Variance 0.00228 0.00226 0.00226 0.00225 0.00223 0.00223 0.00221
## Cumulative Proportion 0.40975 0.41201 0.41426 0.41651 0.41874 0.42097 0.42317
##          PC113  PC114  PC115  PC116  PC117  PC118  PC119
## Standard deviation 1.3210 1.31683 1.31508 1.31384 1.31014 1.30947 1.30713
## Proportion of Variance 0.0022 0.00219 0.00218 0.00218 0.00216 0.00216 0.00215
## Cumulative Proportion 0.4254 0.42756 0.42974 0.43192 0.43408 0.43624 0.43840
##          PC120  PC121  PC122  PC123  PC124  PC125  PC126
## Standard deviation 1.30512 1.30229 1.30162 1.29920 1.29405 1.29296 1.28865

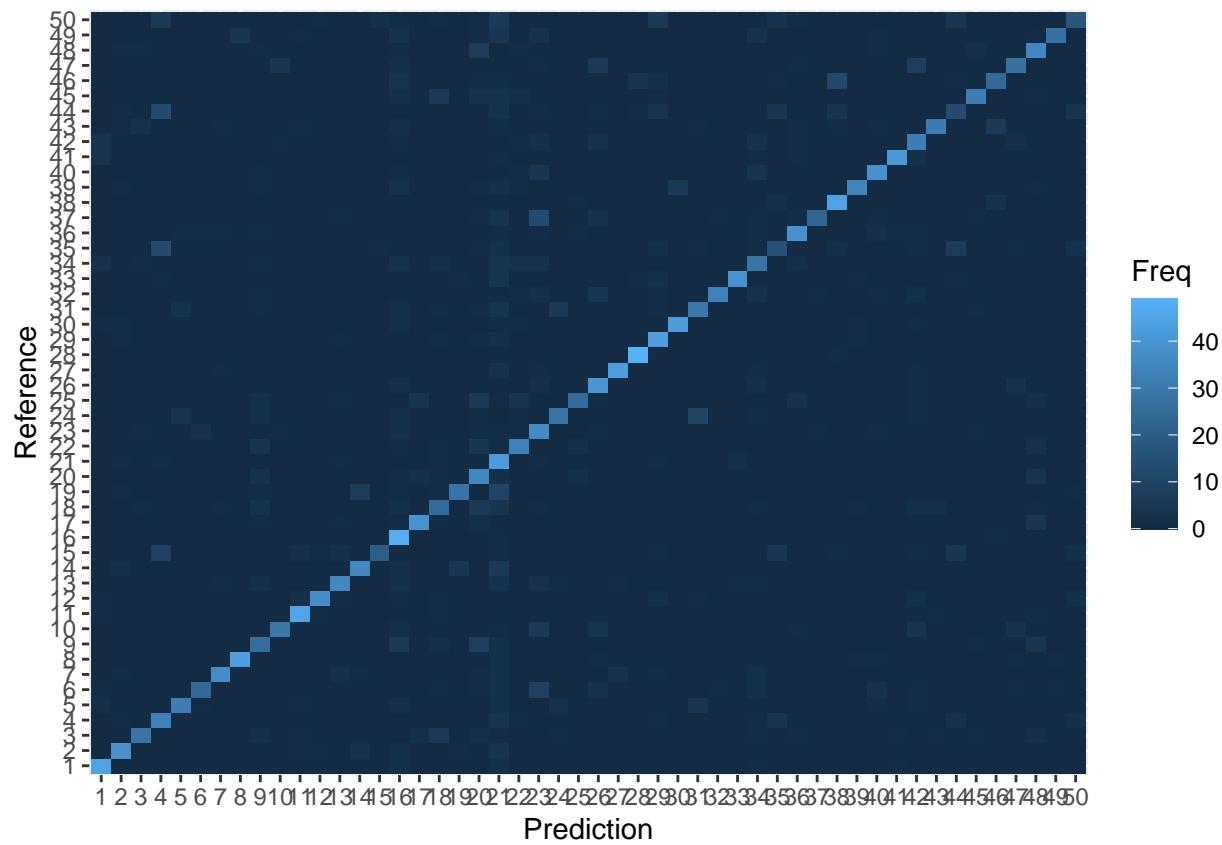
```

```

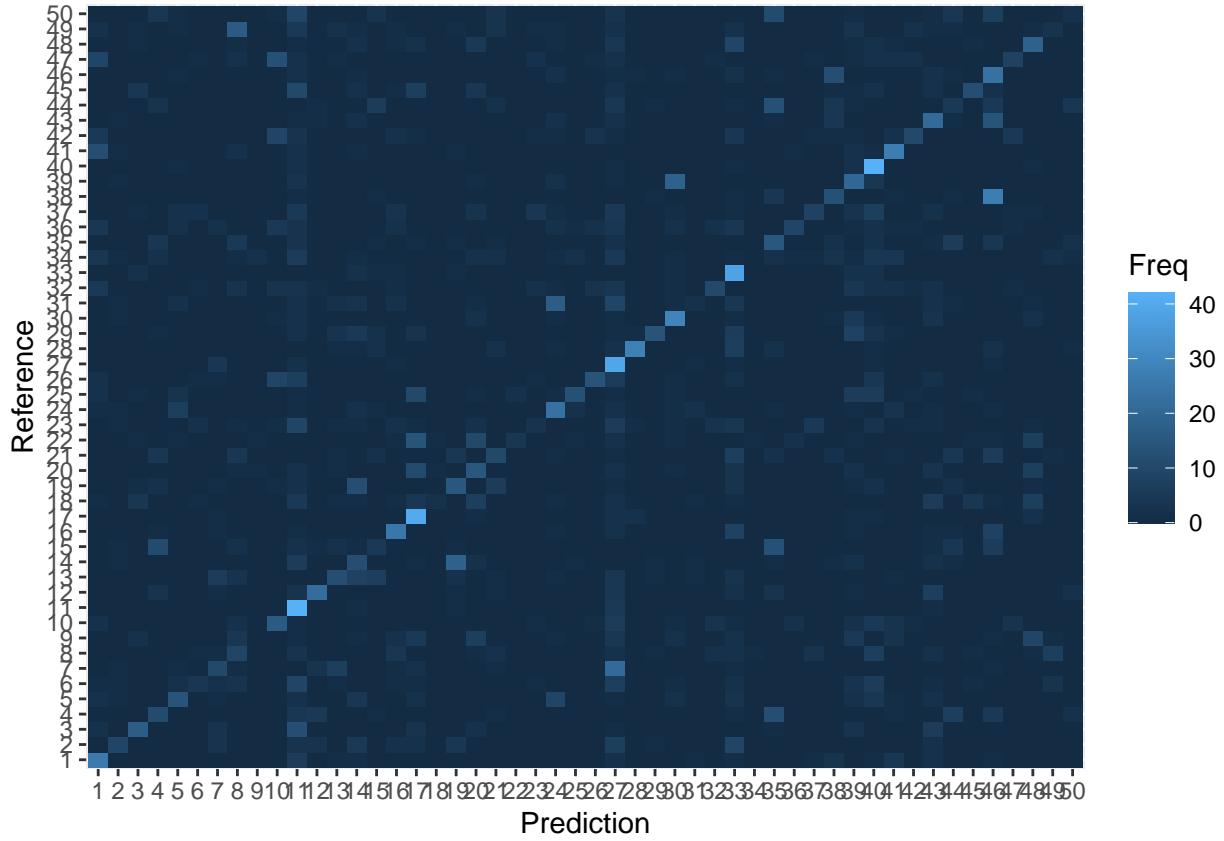
## Proportion of Variance 0.00215 0.00214 0.00214 0.00213 0.00211 0.00211 0.00209
## Cumulative Proportion 0.44055 0.44268 0.44482 0.44695 0.44906 0.45117 0.45326
## PC127 PC128 PC129 PC130 PC131 PC132 PC133
## Standard deviation 1.28749 1.28422 1.28185 1.28007 1.27810 1.27592 1.27398
## Proportion of Variance 0.00209 0.00208 0.00207 0.00207 0.00206 0.00205 0.00205
## Cumulative Proportion 0.45535 0.45743 0.45951 0.46157 0.46363 0.46568 0.46773
## PC134 PC135 PC136 PC137 PC138 PC139 PC140
## Standard deviation 1.27226 1.27026 1.26766 1.26559 1.26130 1.2602 1.25628
## Proportion of Variance 0.00204 0.00203 0.00203 0.00202 0.00201 0.0020 0.00199
## Cumulative Proportion 0.46977 0.47181 0.47383 0.47585 0.47786 0.4799 0.48185
## PC141 PC142 PC143 PC144 PC145 PC146 PC147
## Standard deviation 1.25416 1.25122 1.24959 1.24534 1.24366 1.24214 1.23926
## Proportion of Variance 0.00198 0.00197 0.00197 0.00196 0.00195 0.00195 0.00194
## Cumulative Proportion 0.48384 0.48581 0.48778 0.48974 0.49169 0.49363 0.49557
## PC148 PC149 PC150 PC151 PC152 PC153 PC154
## Standard deviation 1.23745 1.23609 1.23268 1.22934 1.2266 1.22539 1.22404
## Proportion of Variance 0.00193 0.00193 0.00192 0.00191 0.0019 0.00189 0.00189
## Cumulative Proportion 0.49750 0.49943 0.50134 0.50325 0.5051 0.50704 0.50893
## PC155 PC156 PC157 PC158 PC159 PC160 PC161
## Standard deviation 1.22174 1.22084 1.21707 1.21623 1.21385 1.21082 1.20843
## Proportion of Variance 0.00188 0.00188 0.00187 0.00187 0.00186 0.00185 0.00184
## Cumulative Proportion 0.51081 0.51269 0.51456 0.51642 0.51828 0.52013 0.52197
## PC162 PC163 PC164 PC165 PC166 PC167 PC168
## Standard deviation 1.20688 1.20458 1.20126 1.20024 1.19827 1.1935 1.19297
## Proportion of Variance 0.00184 0.00183 0.00182 0.00182 0.00181 0.0018 0.00179
## Cumulative Proportion 0.52381 0.52564 0.52746 0.52927 0.53108 0.5329 0.53468
## PC169 PC170 PC171 PC172 PC173 PC174 PC175
## Standard deviation 1.19157 1.18930 1.18690 1.18636 1.18534 1.18097 1.17856
## Proportion of Variance 0.00179 0.00178 0.00178 0.00177 0.00177 0.00176 0.00175
## Cumulative Proportion 0.53647 0.53825 0.54003 0.54180 0.54357 0.54533 0.54708
## PC176 PC177 PC178 PC179 PC180 PC181 PC182
## Standard deviation 1.17852 1.17541 1.17463 1.17210 1.17102 1.17005 1.16642
## Proportion of Variance 0.00175 0.00174 0.00174 0.00173 0.00173 0.00173 0.00172
## Cumulative Proportion 0.54883 0.55058 0.55232 0.55405 0.55578 0.55750 0.55922
## PC183 PC184 PC185 PC186 PC187 PC188 PC189
## Standard deviation 1.16505 1.1614 1.15932 1.15879 1.15664 1.15386 1.15359
## Proportion of Variance 0.00171 0.0017 0.00169 0.00169 0.00169 0.00168 0.00168
## Cumulative Proportion 0.56093 0.5626 0.56433 0.56602 0.56771 0.56939 0.57107
## PC190 PC191 PC192 PC193 PC194 PC195 PC196
## Standard deviation 1.15268 1.14994 1.14620 1.14475 1.14355 1.14100 1.13997
## Proportion of Variance 0.00168 0.00167 0.00166 0.00165 0.00165 0.00164 0.00164
## Cumulative Proportion 0.57274 0.57441 0.57607 0.57772 0.57937 0.58101 0.58265
## PC197 PC198 PC199 PC200
## Standard deviation 1.13723 1.13671 1.13342 1.13321
## Proportion of Variance 0.00163 0.00163 0.00162 0.00162
## Cumulative Proportion 0.58428 0.58591 0.58753 0.58915

## Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull
## 0.6652000 0.6583673 0.6463158 0.6836965 0.0200000
## AccuracyPValue McNemarPValue
## 0.0000000 NaN

```



```
##      Accuracy      Kappa AccuracyLower AccuracyUpper AccuracyNull
## 0.3076000 0.2934694 0.2895436 0.3261082 0.0200000
## AccuracyPValue McnemarPValue
## 0.0000000      NaN
```



Below is the summary of Steps followed:

1. Reading C50 train files 2. Pre-Processing (train & test):

- \* Created bag-of-words
- \* Change all words to lowercase
- \* Remove numbers and punctuations
- \* Removed white spaces
- \* Removed Stop words
- \* Created term frequency matrix for the independent features

3. Modelling:

- \* Created a y variable to find the author name for each document
- \* Created Principal components for train features and selected the 1st 200 features as they were able to explain 60% of variance of y.
- \* Multi-label Random Forest classification on the training set. Based on the confusion matrix table, we got train accuracy of 66%, while our baseline accuracy = 2% (50 classes).

Note:

- \* Adding IDF factors, along with RF model was giving poor prediction accuracy of just 6%.
- \* Going from 793 features to just 200 features with PCA. Explained proportion with 200 components is 60% which good enough dimensional reduction and without much loss in prediction accuracy.

Results:

- \* Test accuracy achieved through Random Forest model with PCA features: and PCA: 33.2%

## Question 6 - Association rule mining

##

```

## Attaching package: 'arules'

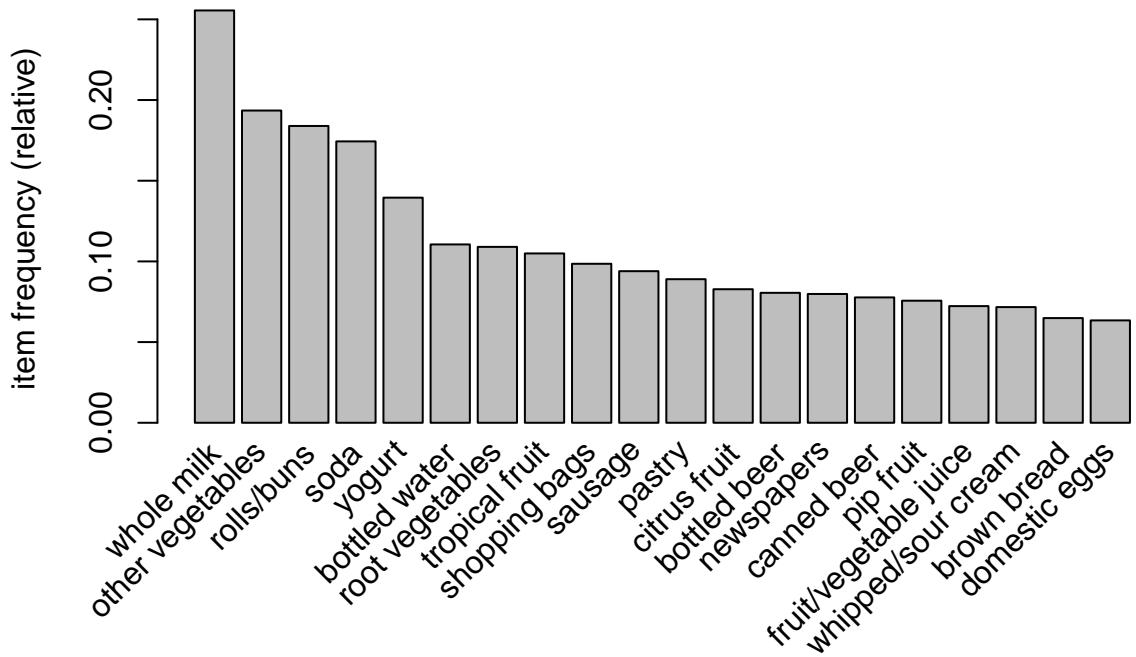
## The following object is masked from 'package:tm':
##   inspect

## The following objects are masked from 'package:mosaic':
##   inspect, lhs, rhs

## The following object is masked from 'package:dplyr':
##   recode

## The following objects are masked from 'package:base':
##   abbreviate, write

```



Several purchase patterns can be observed. For example:

- The most popular transactions were of milk, other vegetables, rolls/buns
- Interestingly soda is 4th most frequent with a count of atleast 1600.

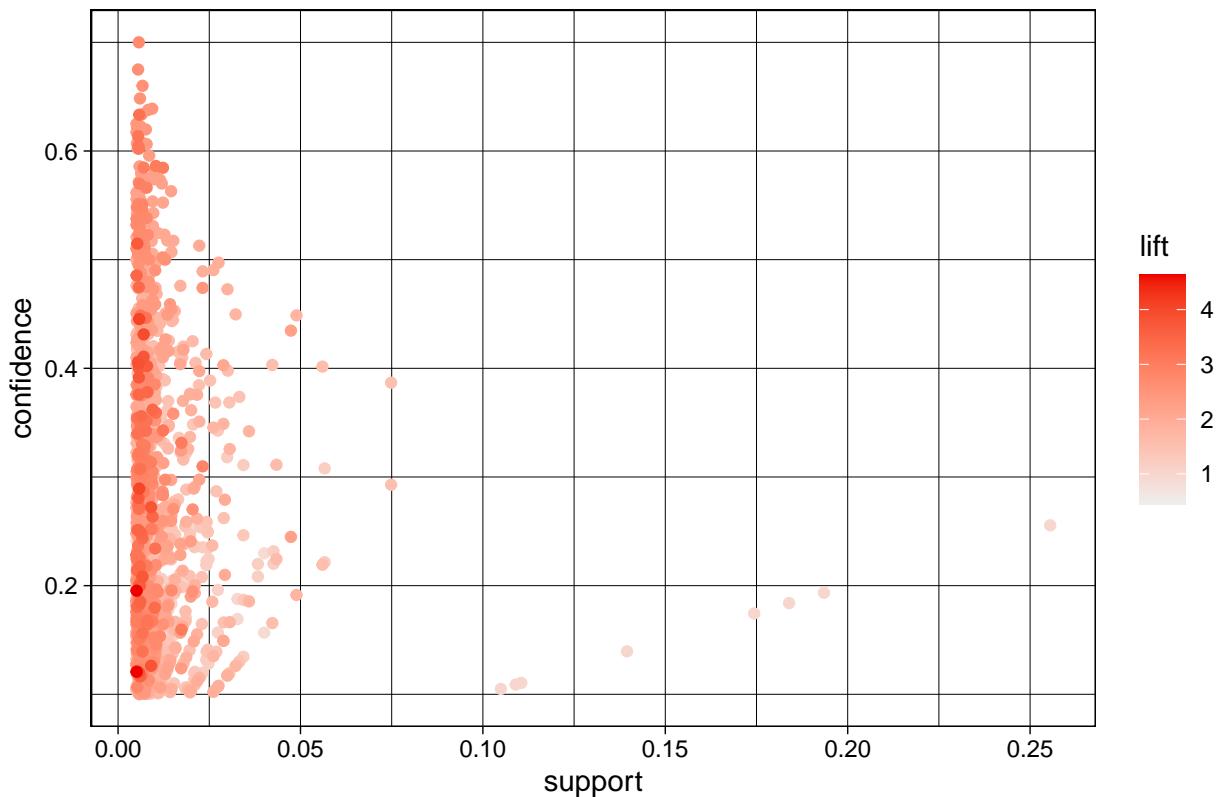
- If someone buys citrus fruit/tropical fruit, root vegetables, they are likely to have bought other vegetables/whole milk as well
- Relatively many people buy curd, yogurt along with whole milk

```

## Apriori
##
## Parameter specification:
##   confidence minval smax arem  aval originalSupport maxtime support minlen
##           0.1     0.1     1 none FALSE             TRUE      5  0.005     1
##   maxlen target  ext
##       5    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 ... [1582 rule(s)] done [0.00s].
## creating S4 object ... done [0.00s].

```

Scatter plot for 1582 rules

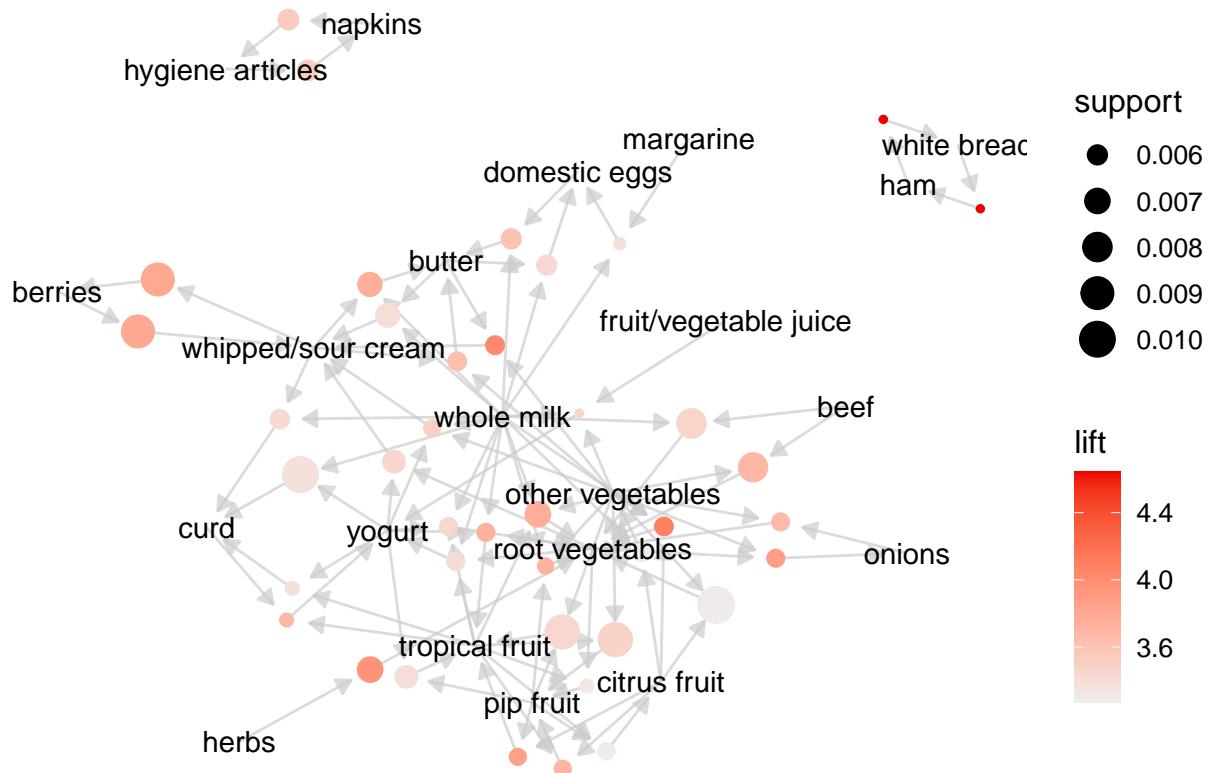


\*Anything with support >0.1 is clearly an outlier. Because these are individual items which have higher occurrency in the item list. So, we will remove those rules.

\*Looking at the graph above, Lift threshold should be 2 to make sure we evaluate only strong associations

##	lhs	rhs	support	confidence	coverage	lift	count
## [1]	{ham}	=> {white bread}	0.005083884	0.1953125	0.02602949	4.639851	50
## [2]	{white bread}	=> {ham}	0.005083884	0.1207729	0.04209456	4.639851	50
## [3]	{citrus fruit, other vegetables, whole milk}	=> {root vegetables}	0.005795628	0.4453125	0.01301474	4.085493	57
## [4]	{butter, other vegetables}	=> {whipped/sour cream}	0.005795628	0.2893401	0.02003050	4.036397	57
## [5]	{herbs}	=> {root vegetables}	0.007015760	0.4312500	0.01626843	3.956477	69
## [6]	{other vegetables, root vegetables}	=> {onions}	0.005693950	0.1201717	0.04738180	3.875044	56
## [7]	{citrus fruit, pip fruit}	=> {tropical fruit}	0.005592272	0.4044118	0.01382816	3.854060	55
## [8]	{berries}	=> {whipped/sour cream}	0.009049314	0.2721713	0.03324860	3.796886	89
## [9]	{whipped/sour cream}	=> {berries}	0.009049314	0.1262411	0.07168277	3.796886	89
## [10]	{other vegetables, tropical fruit, whole milk}	=> {root vegetables}	0.007015760	0.4107143	0.01708185	3.768074	69

## Association Graphs



Bidirectional rules with high lift and low support (which means highly complimentary items)

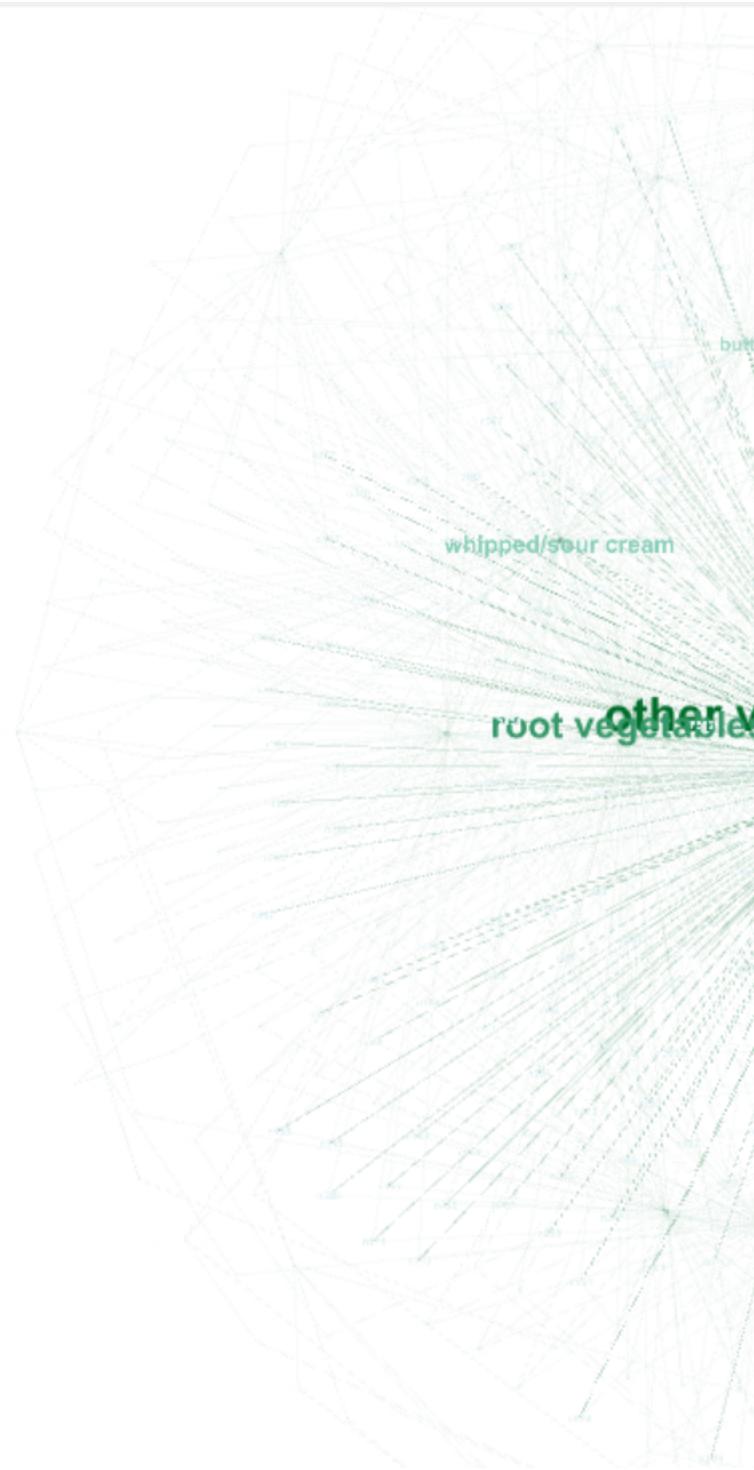
1. Ham -> whitebread & Whitebread -> Ham
2. Hygiene Articles -> Napkins & Napkins -> Hygiene Articles
3. whipped/sour cream -> berries & berries -> whipped/sour cream
4. Vegetables are brought together in general

In the grocery store, there are different centers. For example, below are often brought together:

- \* Fruits/vegetables/dairy products
- \* Hygiene products

Next steps:

- \* We can deep dive and check why white bread is bought only with Ham and not milk. It could be because of thresholds we have set, but it needs further speculation.



Association rule graph visualization in Gephi