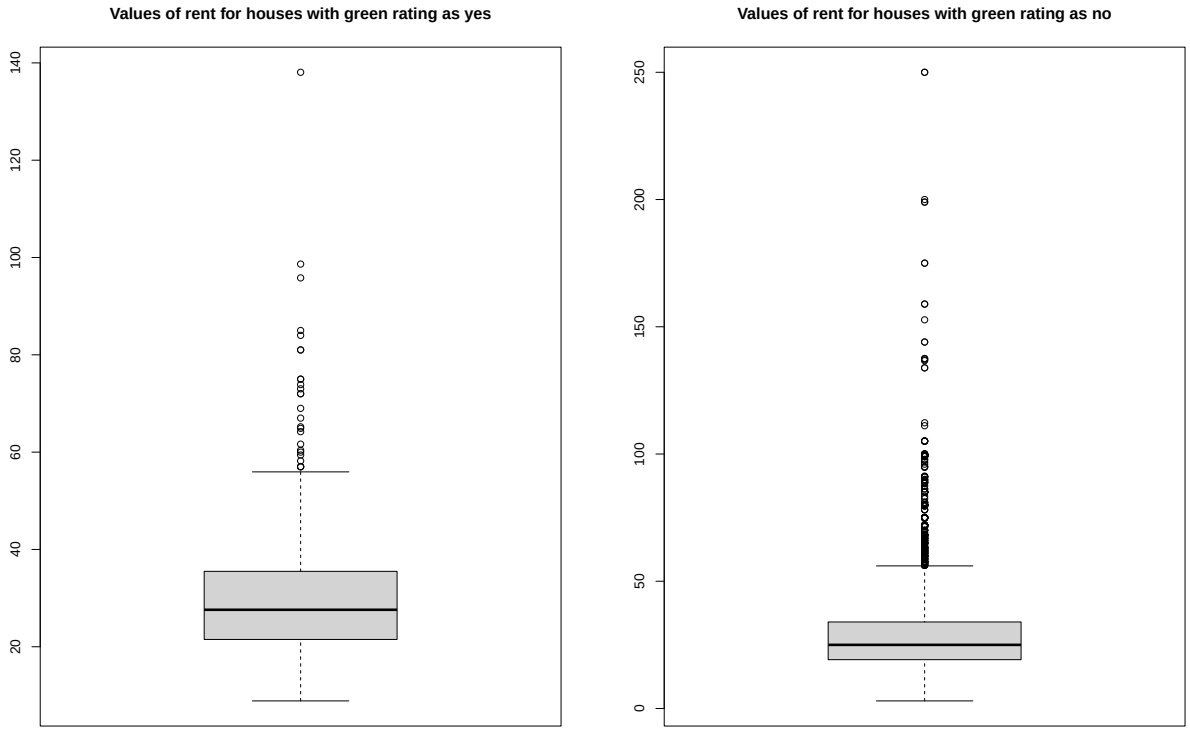
Exam 2

SHUBHAM SINGH, SOUMIK CHOUDHURI ,KARTHICK RAMASUBRAMANIAN, HARSH MEHTA

08/03/2020

Visual story telling part 1: green buildings

We started out with the analysis of the Stat Guru's Work.




```
## 27.6 25

##
## Call:
## lm(formula = Rent ~ green_rating, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.287  -9.044  -3.267   5.733 221.733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28.2668     0.1775 159.275   <2e-16 ***
## green_rating   1.7493     0.6025   2.903   0.0037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.07 on 7892 degrees of freedom
## Multiple R-squared:  0.001067,   Adjusted R-squared:  0.0009405
## F-statistic:  8.43 on 1 and 7892 DF,  p-value: 0.003701
```

1) The stat Guru made his conclusion on the basis of point estimate statistic, (median) of two subsets of data, and didn't even consider the variation in the rent price of two subsets for his conclusion remarks.

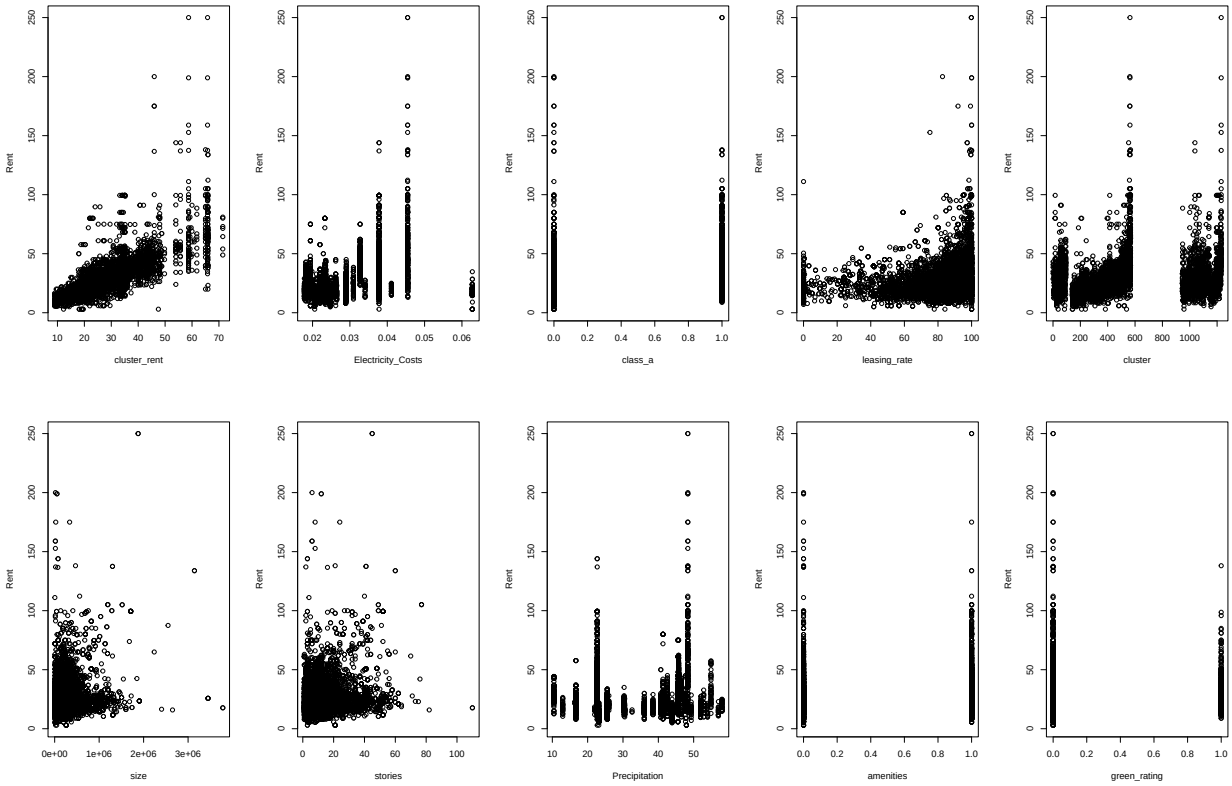
2) There is clear overlapping in rent price for two sets of data;

One with the green rating and the other without the one

3) Possibly this variation is due to other factors that need to be analysed for precisely deciding what other factors control the rent price

- The SLR model trained using only the indicator variable (whether the building is green or not) is not able to explain the variation in rent prices (adj r2 is quite low)
- We surely need to consider other variables

Analysis of individual variable




```
##           predictors         coeff        p_vals
## 21 Electricity_Costs  6.918652e+02 9.678206e-288
## 9             class_a  6.498923e+00  2.693258e-80
## 14                net -4.173730e+00  6.637145e-06
## 20         Gas_Costs  3.889947e+00  9.560120e-01
## 8          renovated -3.792060e+00  1.361204e-27
## 10           class_b -3.743735e+00  2.532107e-28
## 15         amenities  1.836687e+00  6.332899e-08
## 12        Energystar  1.802486e+00  3.780595e-03
## 13      green_rating  1.749252e+00  3.700575e-03
## 11              LEED  1.290631e+00  5.307220e-01
```

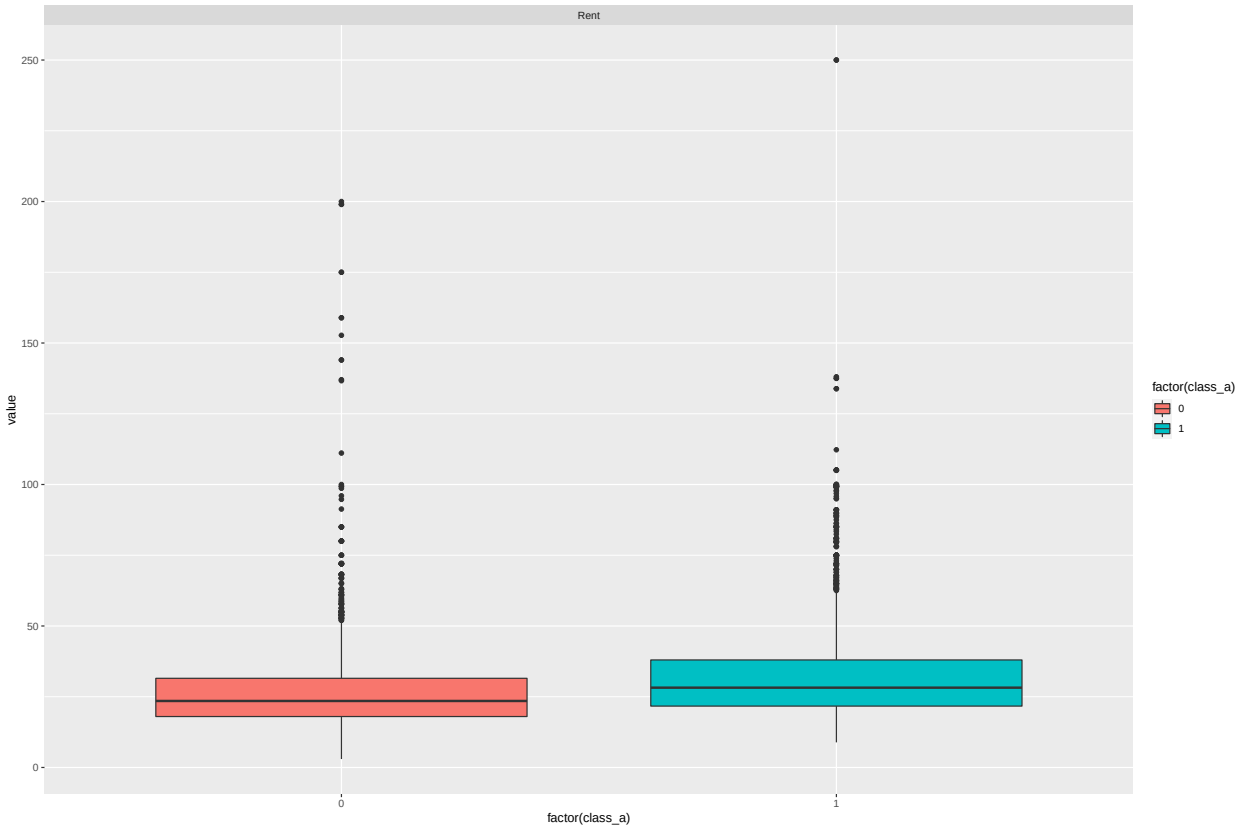
```
## 22      cluster_rent  1.080052e+00  0.000000e+00
## 6            stories  1.424611e-01  4.200110e-25
## 5       leasing_rate  1.262626e-01  7.034605e-58
## 19     Precipitation  8.652248e-02  3.461048e-09
## 7                 age -4.807381e-02  6.006126e-20
## 4            empl_gr -4.744971e-02  2.333542e-02
## 2            cluster  6.496323e-03  1.126627e-53
## 16       cd_total_07 -2.292038e-03  4.995907e-51
## 18       total_dd_07 -1.893035e-03 5.120155e-112
## 17        hd_total07 -1.191668e-03  2.442562e-44
## 3               size  6.968472e-06  1.220878e-34
## 1      CS_PropertyID -3.135488e-06  1.971085e-43
```

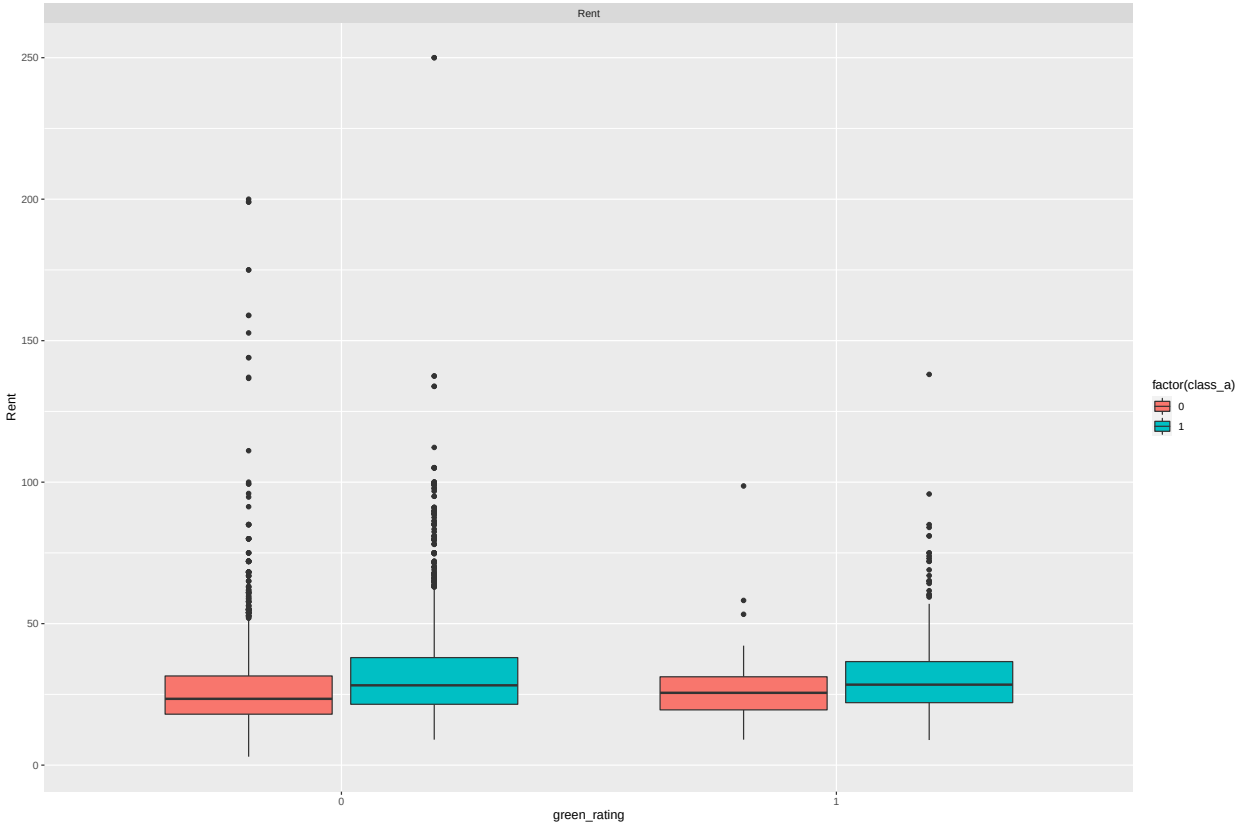
OBSERVATIONS

1) Individual predictor analysis suggest that the rent to an appreciable extent depends on the value of cost of electricity, class a and many other variables

2) These variables can have same value in two different subgroups (green not green) or can have remarkably different values inside same subset

Analysis of Class A variable, independent and combination with Green rating



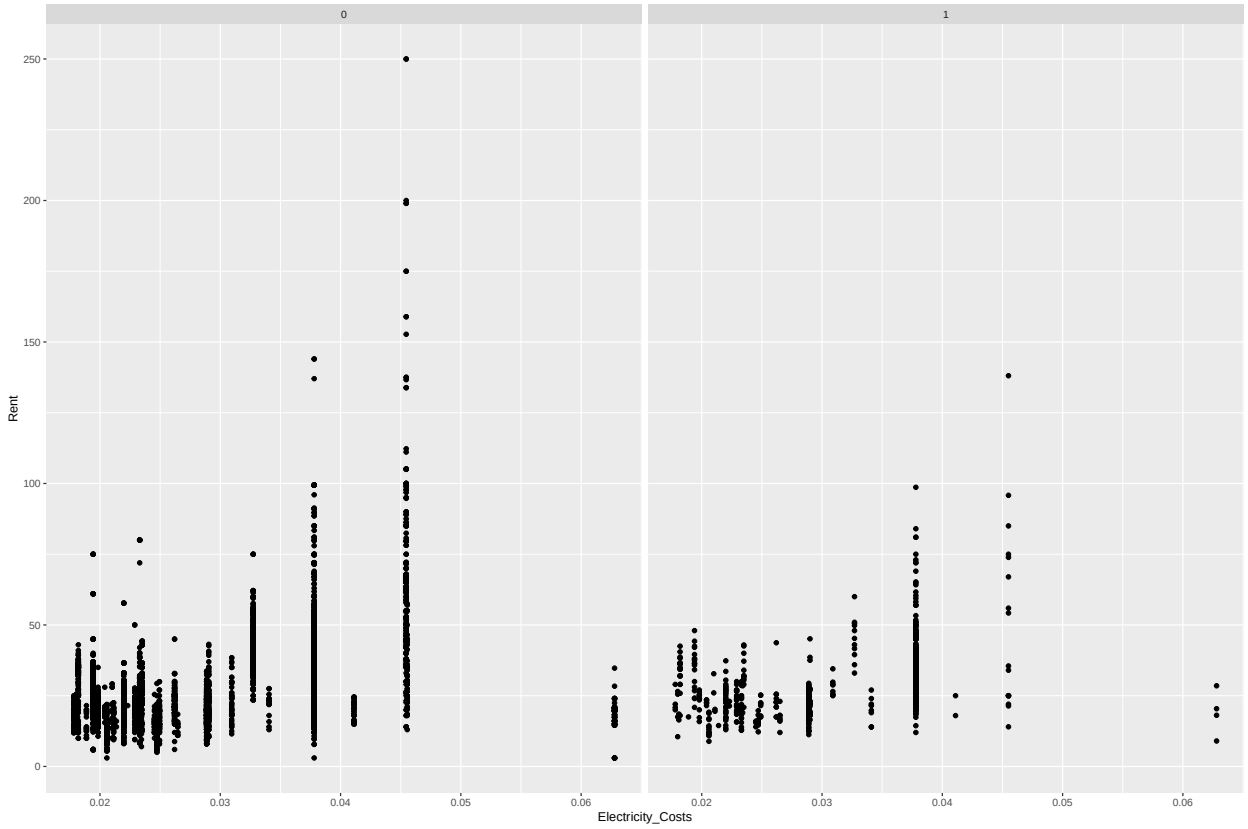


OBSERVATIONS

1) Class A seems to be more discriminatory when it comes to determine prices

2) Within each subset of green rating class, class_a variable is able to differentiate between rent prices It can be a strong confounding variable

Analysis of Electricity cost variable, in combination with Green rating



1) Within both kinds of houses rent increases with electricity cost

-There is a possibility that other predictors help to determine green_rating variable.

-Hence the variation in rent prices withing green_rating is because of those secondary (hidden) predictors

-These hidden predictors can be strong candidates for the confounding variables

Estimating the presence of confounding variable

```
##
## Call:
## glm(formula = factor(green_rating) ~ Gas_Costs, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -0.4474  -0.4373  -0.4373  -0.4076   2.6209
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.6891     0.2445  -6.909 4.87e-12 ***
## Gas_Costs   -59.2615    21.7104  -2.730  0.00634 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4657.6  on 7893  degrees of freedom
## Residual deviance: 4648.7  on 7892  degrees of freedom
## AIC: 4652.7
##
## Number of Fisher Scoring iterations: 5
```

```
##           predictors         coeff       p_vals
## 9            class_a  1.934033e+00 1.692003e-86
## 7                age -4.673134e-02 9.540872e-67
## 10           class_b -1.371780e+00 4.464668e-43
## 14         amenities  9.492492e-01 1.359988e-26
## 8          renovated -8.811853e-01 5.943881e-20
## 16        hd_total07 -1.931531e-04 2.838532e-18
## 5       leasing_rate  2.498687e-02 1.020214e-16
## 3               size  8.224380e-07 2.340708e-15
## 17       total_dd_07 -1.247612e-04 1.327380e-09
## 15       cd_total_07  1.556886e-04 1.319342e-06
## 18     Precipitation -1.607532e-02 8.664053e-06
## 6             stories  1.152974e-02 1.054071e-04
## 13                net  5.831209e-01 1.034586e-03
## 19         Gas_Costs -5.926150e+01 6.340368e-03
## 2            cluster  2.321198e-04 1.899723e-02
## 20 Electricity_Costs  9.233684e+00 4.558625e-02
## 1      CS_PropertyID -1.361314e-07 4.931130e-02
## 21      cluster_rent -6.101075e-03 1.176531e-01
## 4             empl_gr  4.425456e-03 3.182150e-01
## 11               LEED  1.800185e+01 9.275773e-01
## 12        Energystar  2.559901e+01 9.709089e-01
```

OBSERVATIONS

1) Predictors like class_a,age and class_b seems capable of determining green_rating for households

2) Additionally class_a, class_b based variables demonstrated appreciable amount of association with rent prices (from our previous analysis)

3) Class a from our combination of analysis, seems to be a foremost contender of being a confounding variables for relationship present between green_rating and rent price

Visual story telling part 2: flights at ABIA

-We started with creating an indicator variable stating whether the specific flight started from and ended at Austin

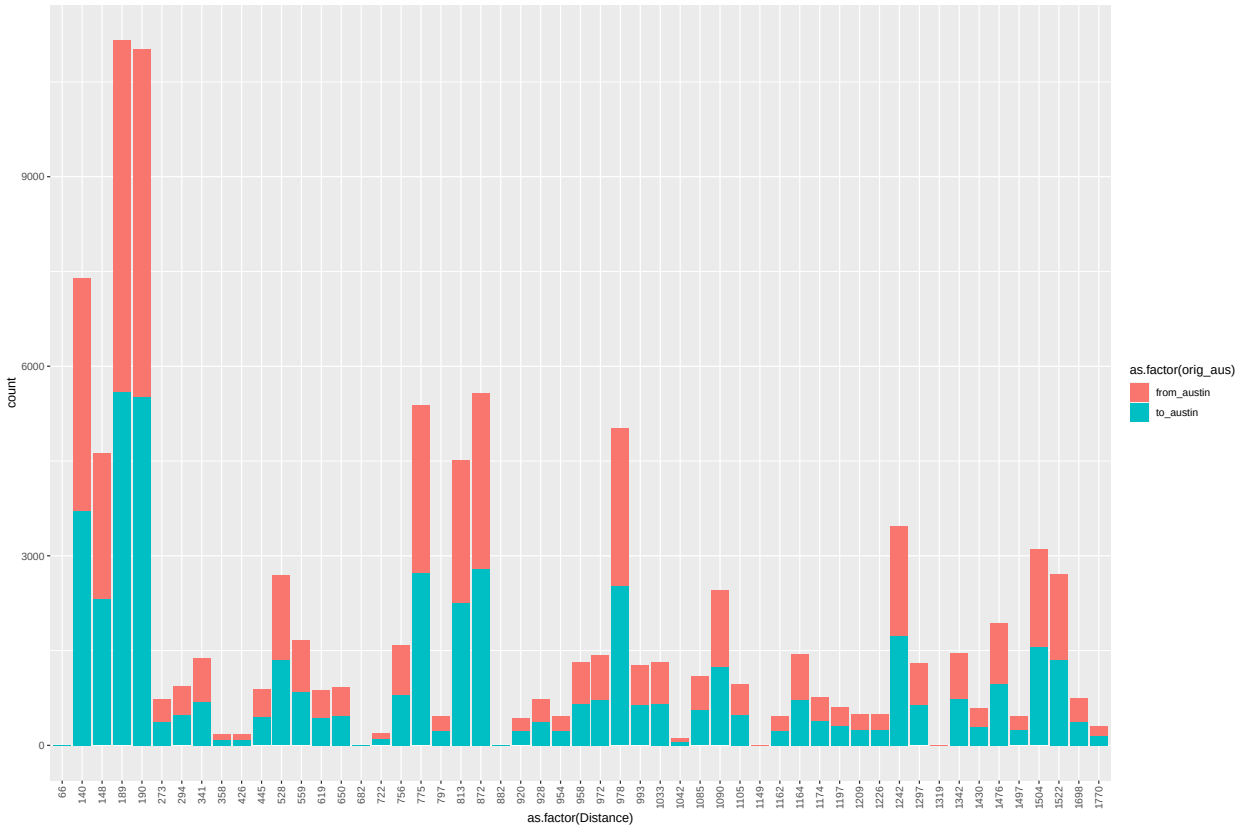
```
##
## from_austin   to_austin
##       49623       49637
```

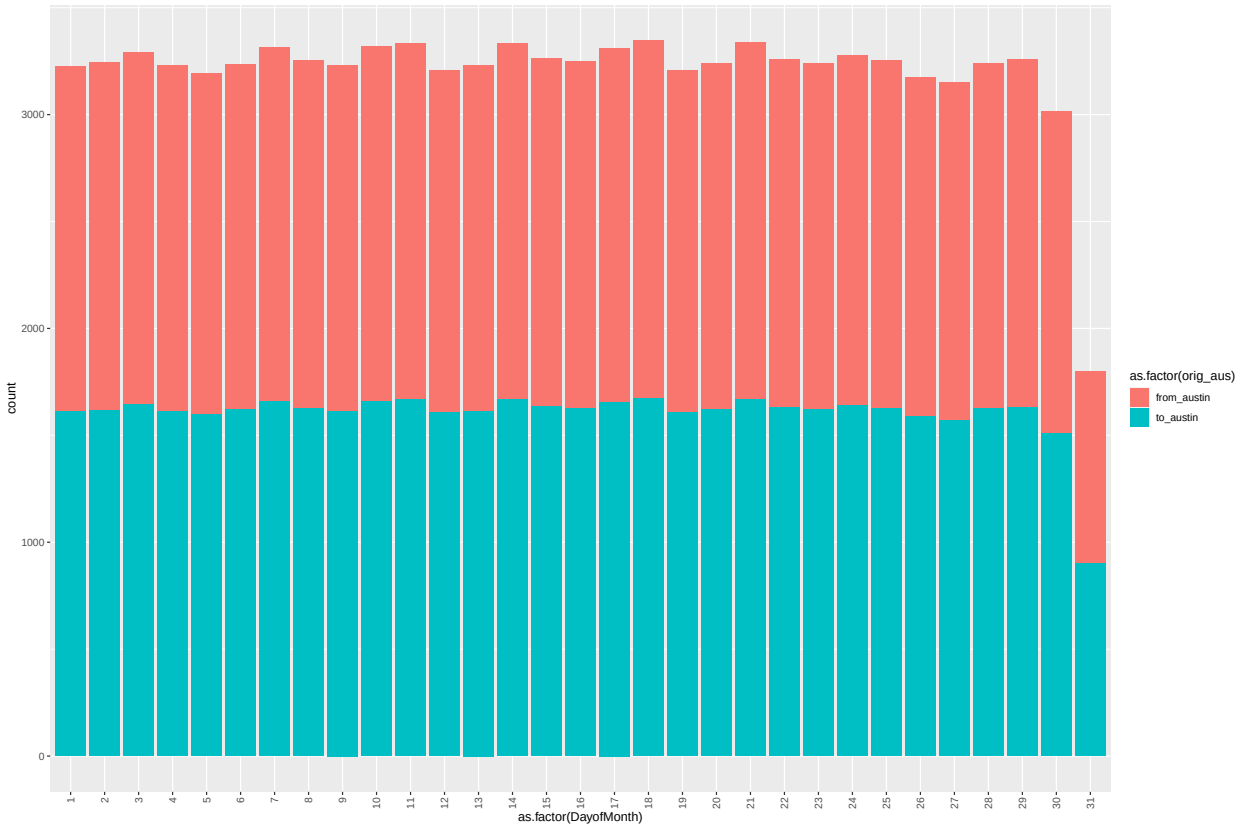
```
## Location of origin of flights arriving at Austin
```

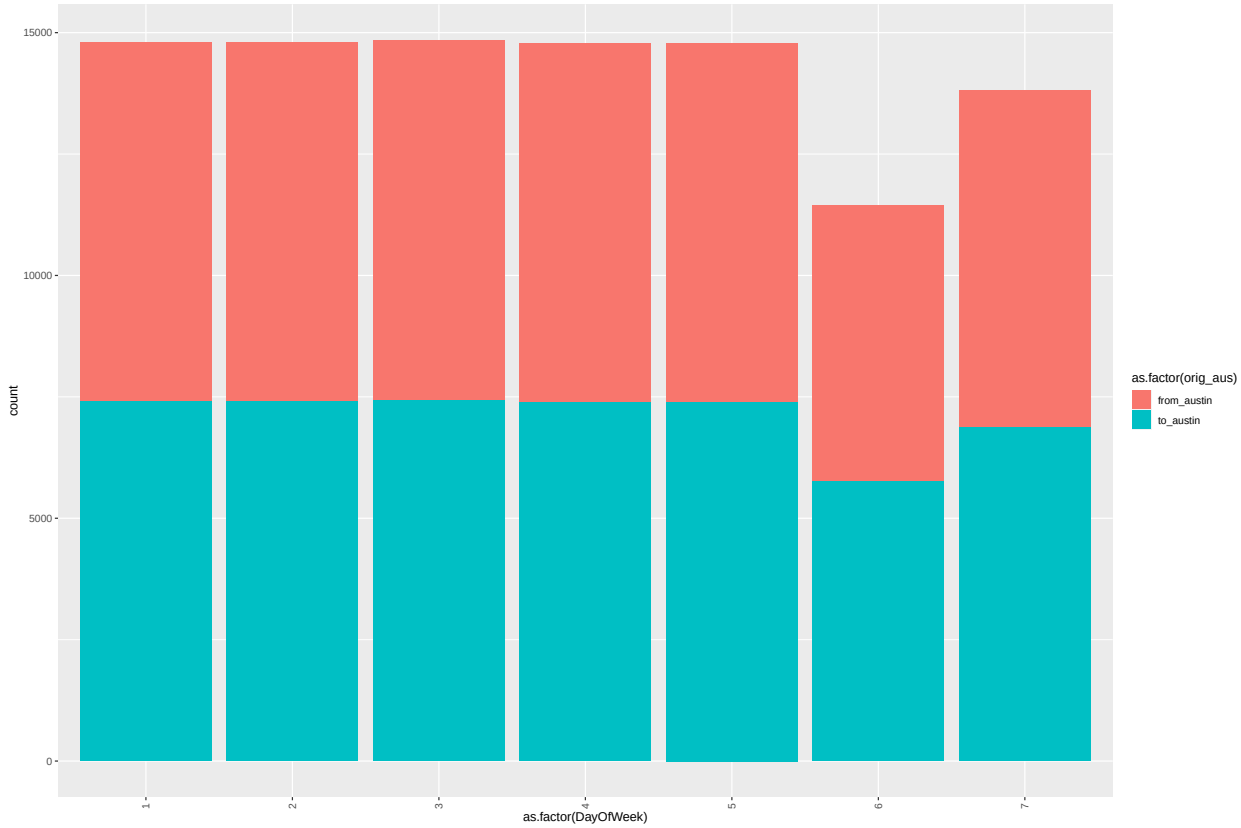

```
## Location of destination of flights departing from Austin
```


SEA
MSP
BOS
CLE
PHL
DSM
SLC
IND
CVG
DEN
MCI
STL
RDU
LAS
BNA
CLT
JUL
OKC
MEM
ABQ
ATL
PHX
LBB
SAN
TUS
JAX
FLP
MAF
MSY
MCO
TPA
FLL
HRL

Analysing the departing and landing flights with other catehgorical features







OBSERVATIONS

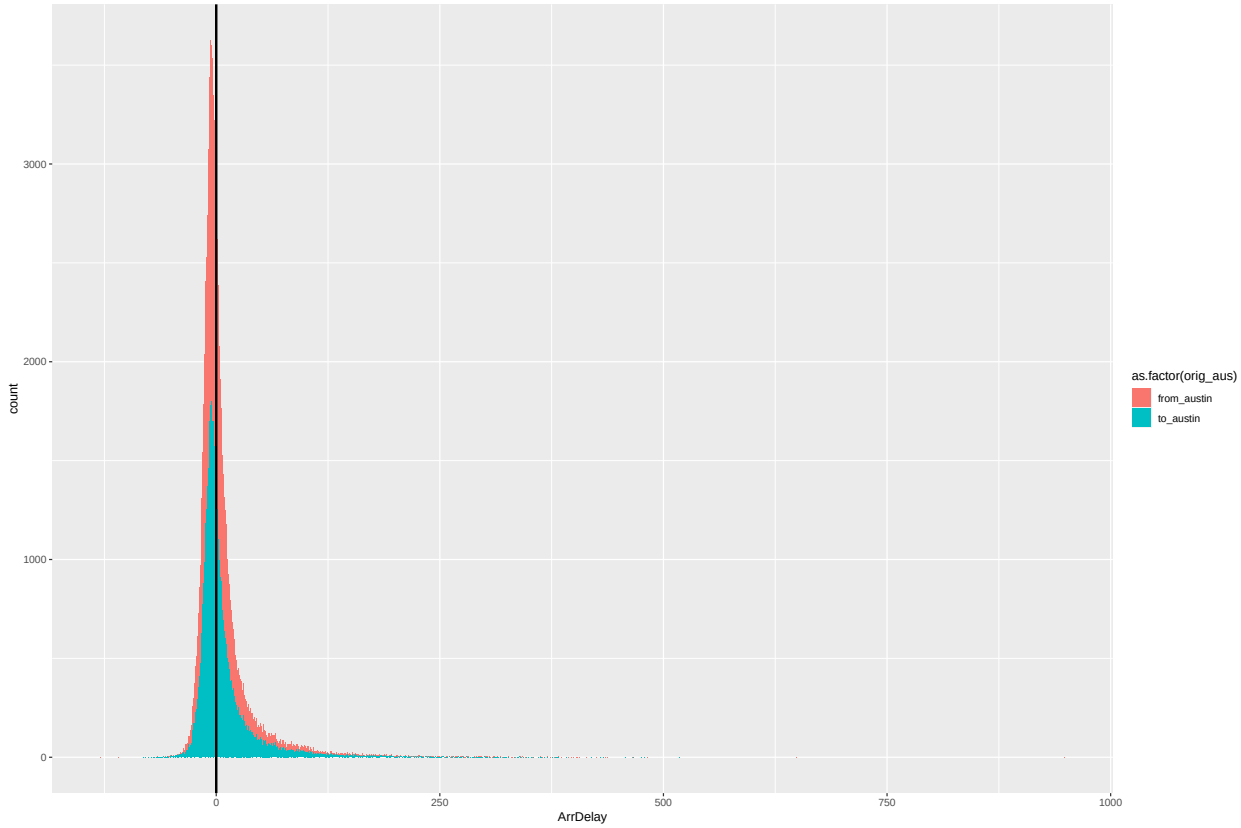
1)Short distance flights are more in number as compared to long distance flights However, the arriving and departing flights are evenly distributed for all range of distances

2) The number of incoming and outgoing flights are uniformly distributed across all days of the month

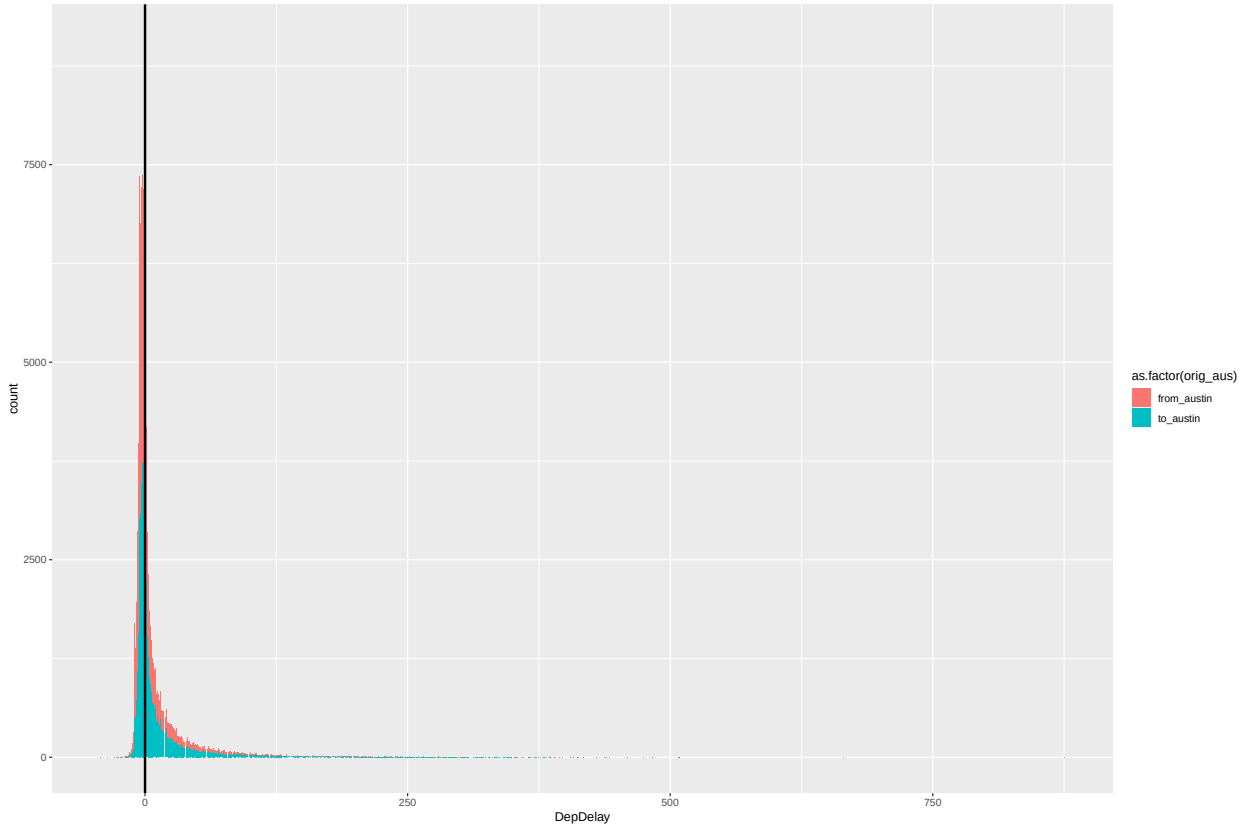
3) The number of flights are less on weekends

ANALYSING flights with respective delay times

```
## Warning: Removed 1601 rows containing non-finite values (stat_count).
```




```
## Warning: Removed 1413 rows containing non-finite values (stat_count).
```




```
## Flights departed to following cities have departure delay of more than 153

##
## ABQ ATL BNA BOS BWI CLE CLT CVG DAL DEN DFW ELP EWR FLL HOU IAD IAH JAX JFK LAS
##   5  41   4   5   5   4   3   4  32  32  23   8  23   3  19  23  30   7  27   4
## LAX LBB LGB MAF MCI MCO MDW MEM MSP MSY OAK ONT ORD PHL PHX RDU SAN SFO SJC SLC
##  12   2   4   3   2   7   7   5   2   3   1   8  46   2  10   2   1  10   8   3
## STL TPA
##   3   2

## Flights arriving to following cities have arrival delay of more than 153

##
## ABQ ATL BNA BOS BWI CLE CLT DAL DEN DFW ELP EWR FLL HOU IAD IAH JFK LAS LAX LBB
##   3  37  20   2   9   1  11  26  15  23   5  20  11  18  38  21  28   9   6   4
## LGB MAF MCI MCO MDW MEM MSP MSY OKC ONT ORD PHL PHX RDU SAN SEA SFO SJC SLC SNA
##   8   2   5   4  14   7   1   5   5   3  58  10  13  11   4   5   7  14   4   2
## STL TPA TUL TUS
##   1   2   1   3
```

OBSERVATIONS

1) Flights arriving to Austin seems to arrive early, arrival delay time is usually less than 0

-Flights having high order of delay seems to be arriving majorly from Chicago ORD, also Dulles IAD and Atlanta ATl airports

2) Flights departing from Austin, comparatively have high Departure delay times The delay time is occasionally quite high

-Flights majorly leaving for ATL (Atlanta) has high order of departure delay (0.99 quantile value of departure delay)

Portfolio MOdelling

We extracted all the ETF associated symbols and then documented their respective returns for the last 5 years

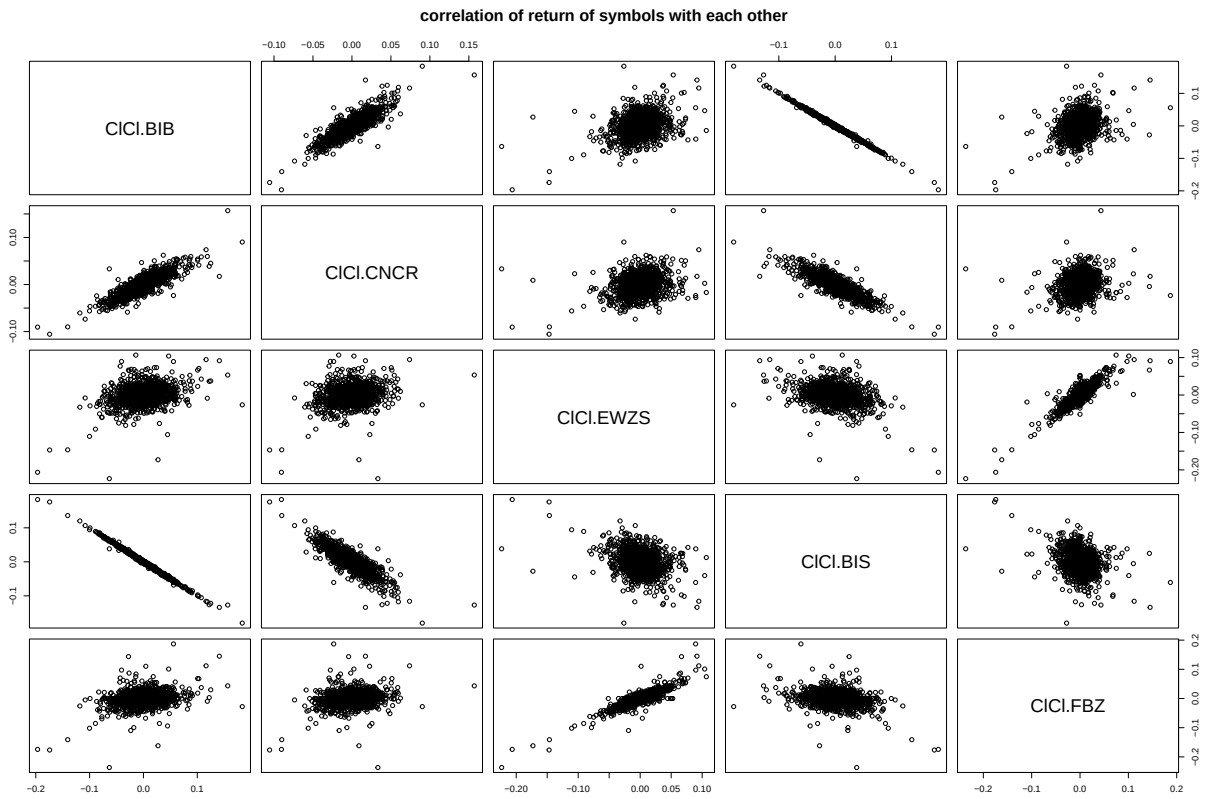
Following, we computed variance of their return for last five years. We sub selected those symbols that has been actively (providing return) for the last five years

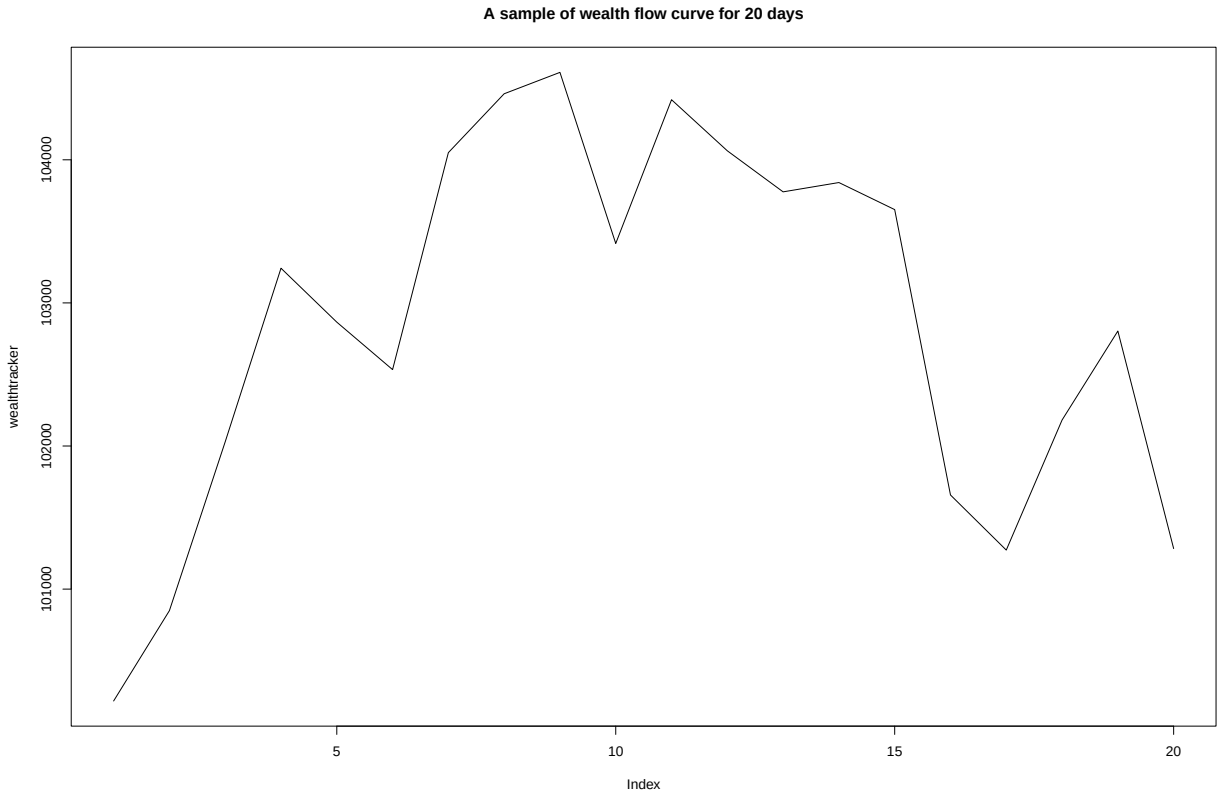
We created our aggressive, moderate and mixed profiles using the variation score of returns of symbols

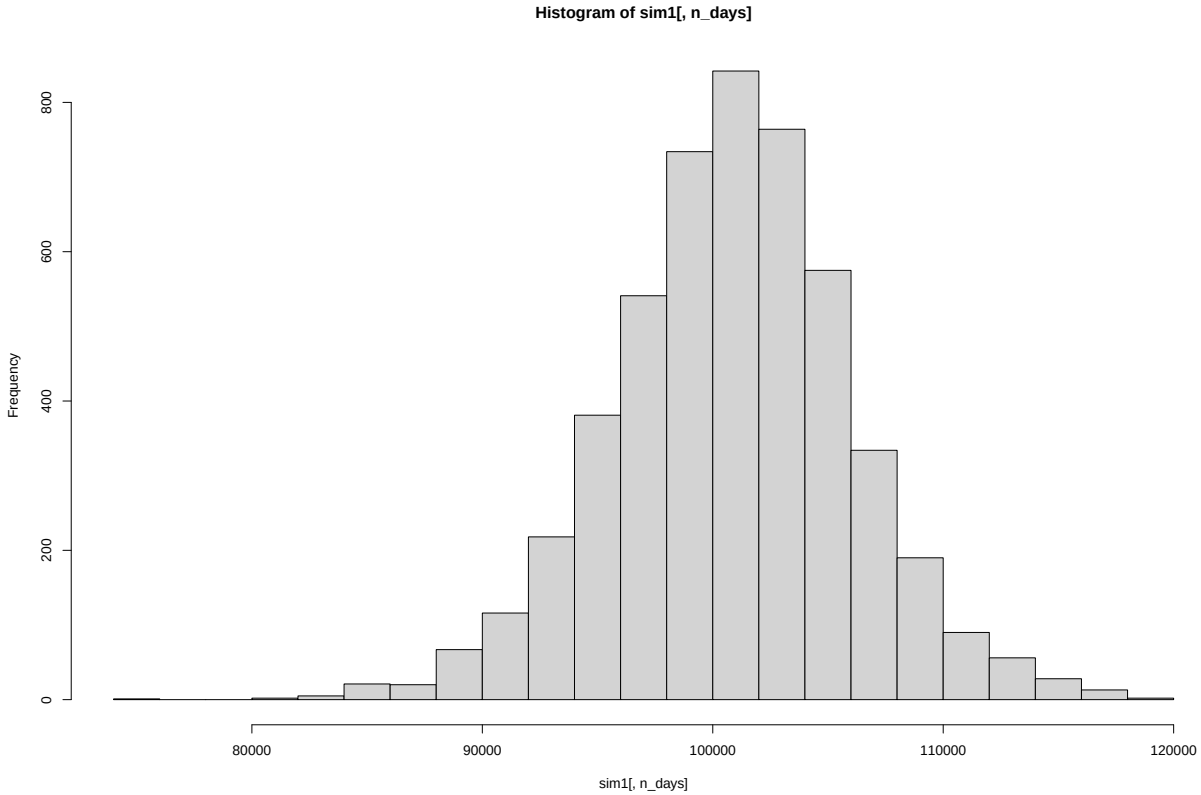
Aggressive Portfolio (With high variation)

Post sorting (decreasing order) the symbols on the basis of their 5 year return variation, we selected the top five symbols for our aggressive portfolio.

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

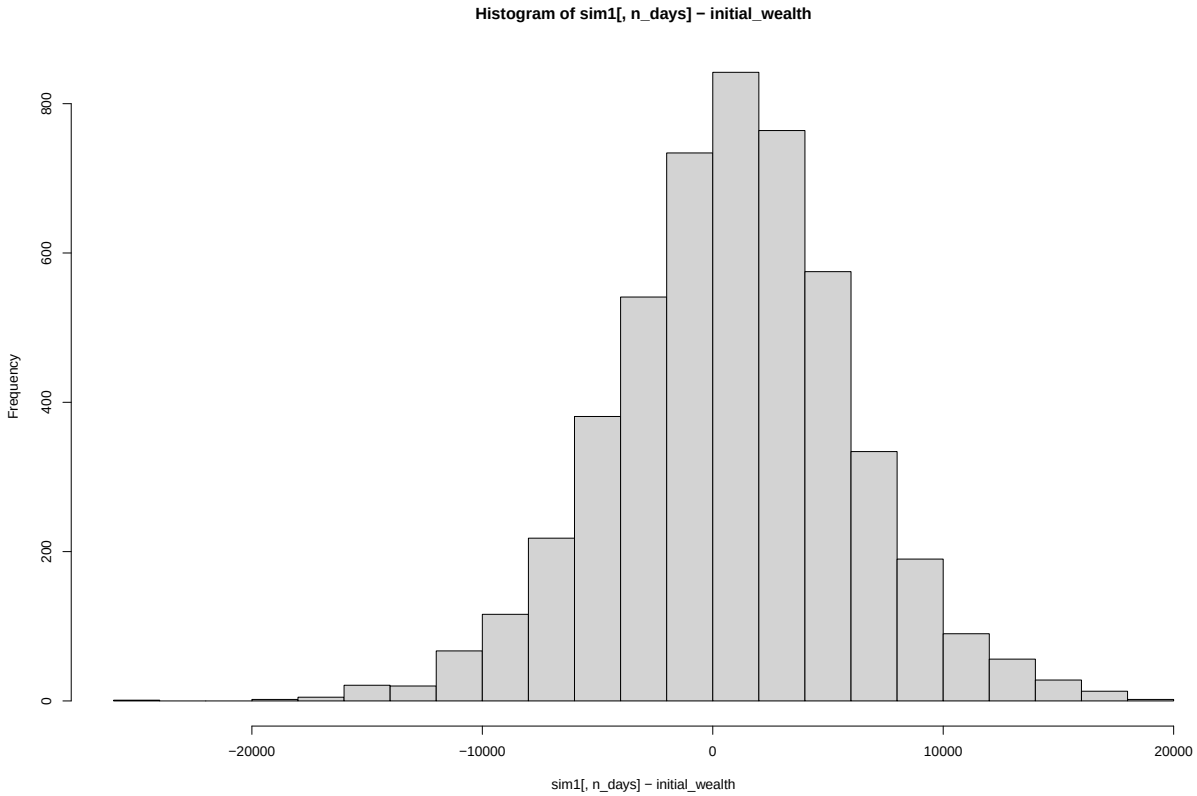



A sample of wealth flow curve for 20 days
wealthtracker
101000
102000
103000
104000
5
10
15
20
Index


Histogram of sim1[, n_days]
Frequency
0
200
400
600
800
80000
90000
100000
110000
120000
sim1[, n_days]

```
## [1] 100823.3
```

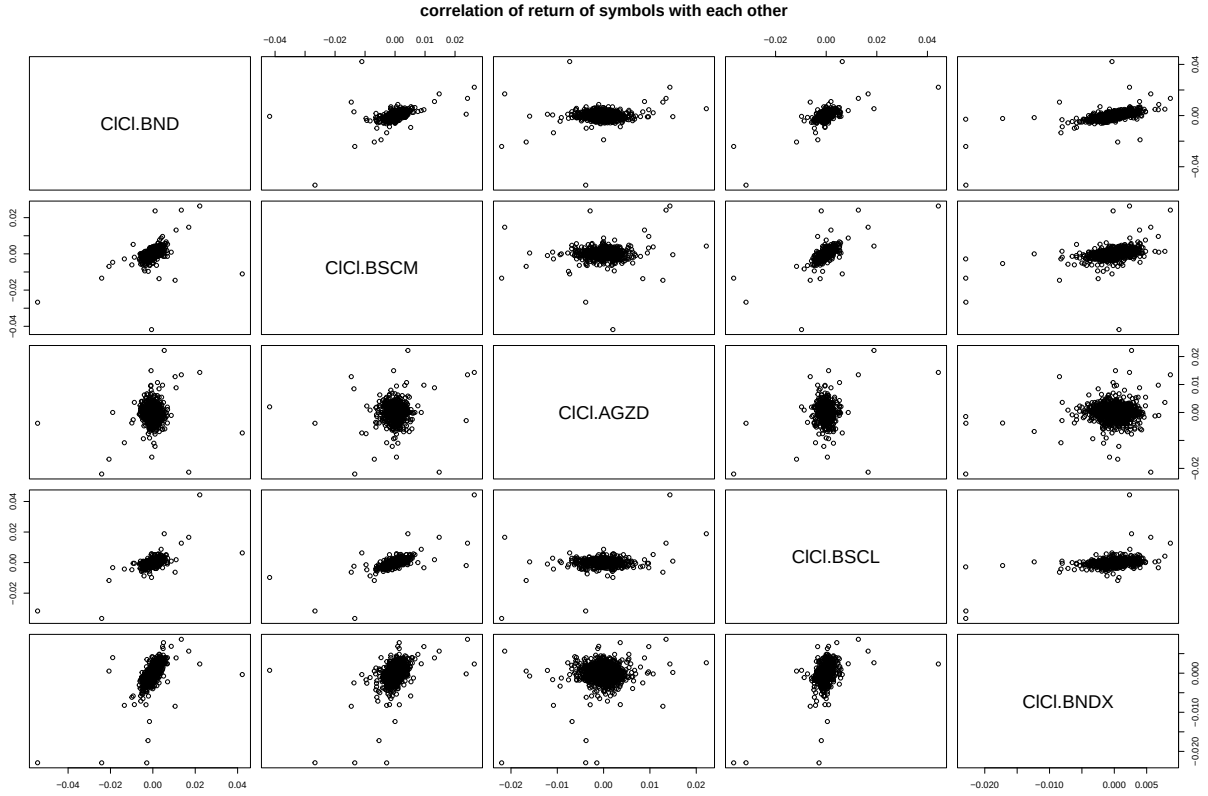
```
## [1] 823.2834
```

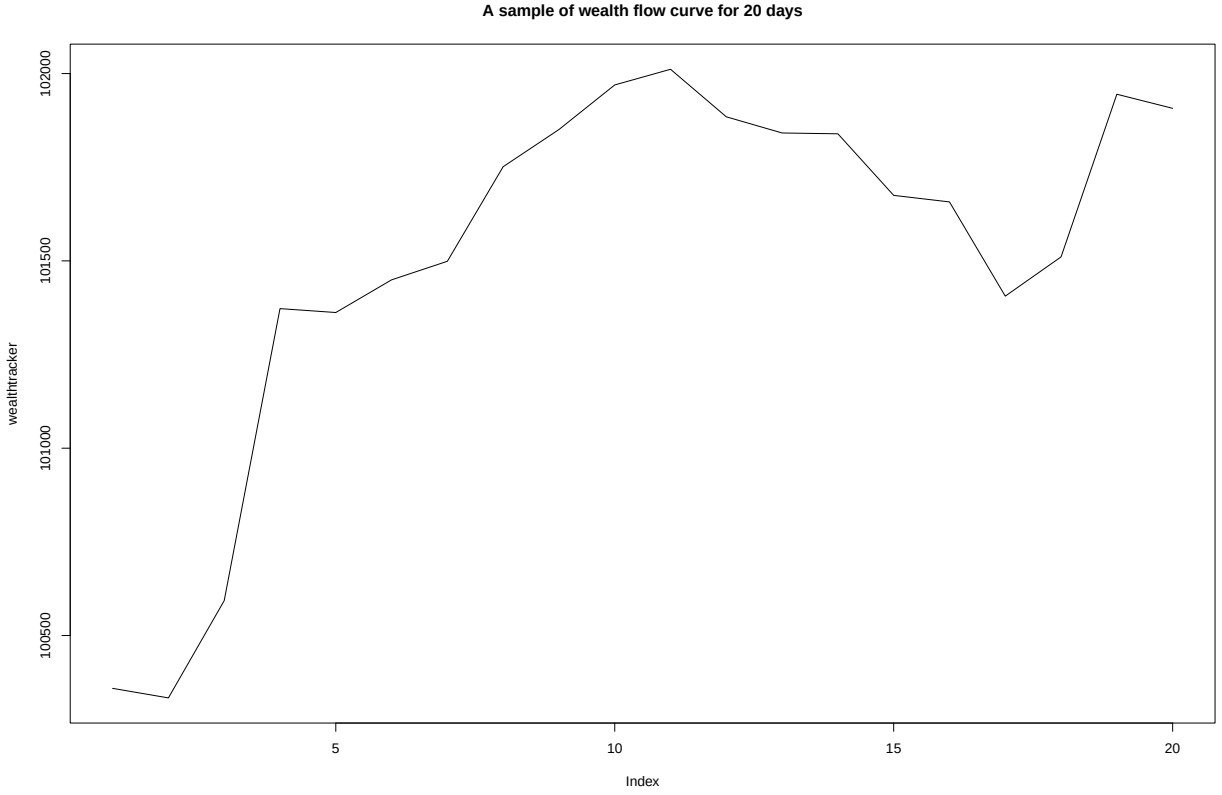


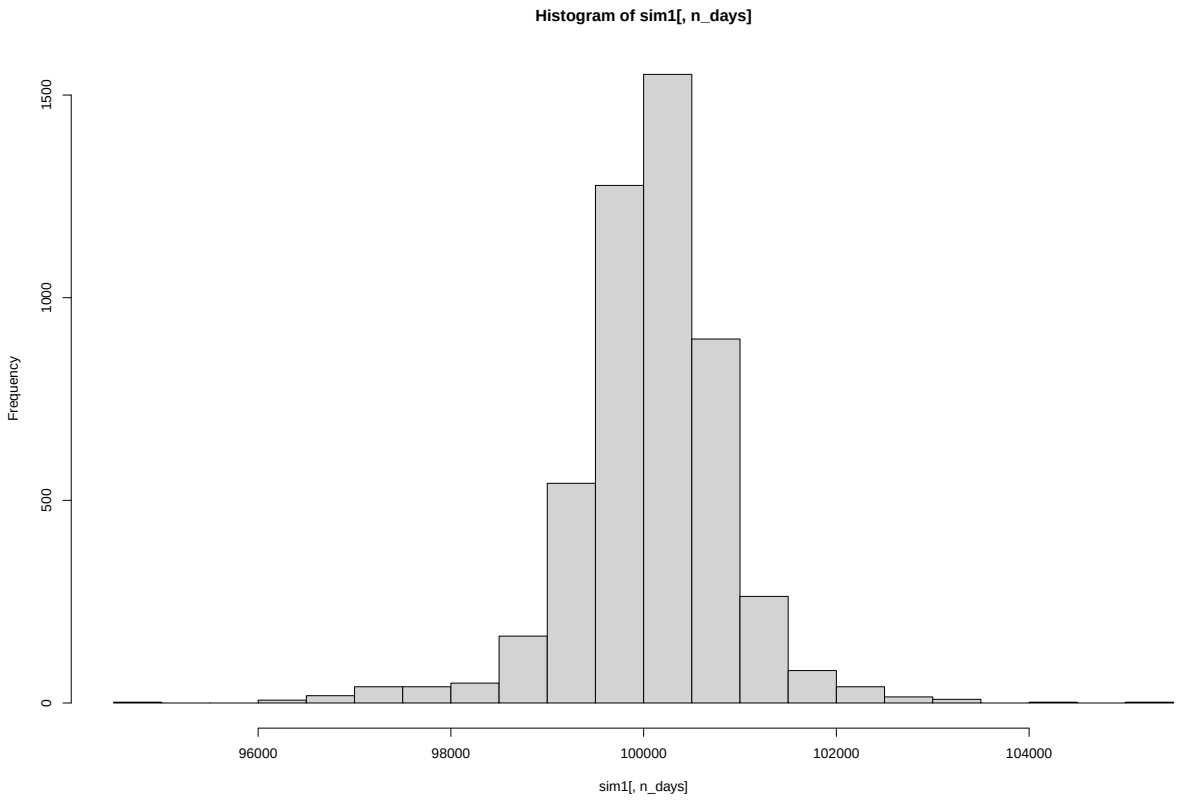

```
## stocks used BIB EWZS CNCR BIS FBZ
```

```
## 5% value at risk -7776.073
```

Non aggressive portfolio

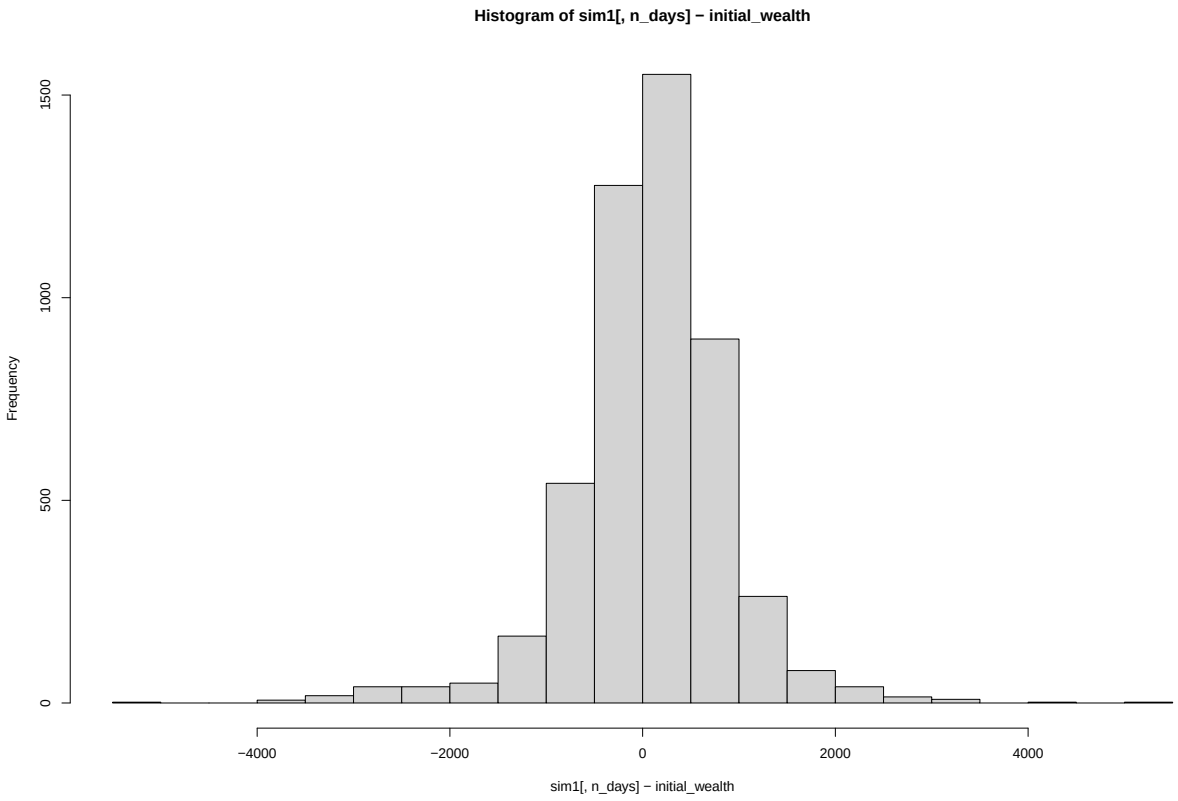







```
## [1] 100075.1
```

```
## [1] 75.131
```

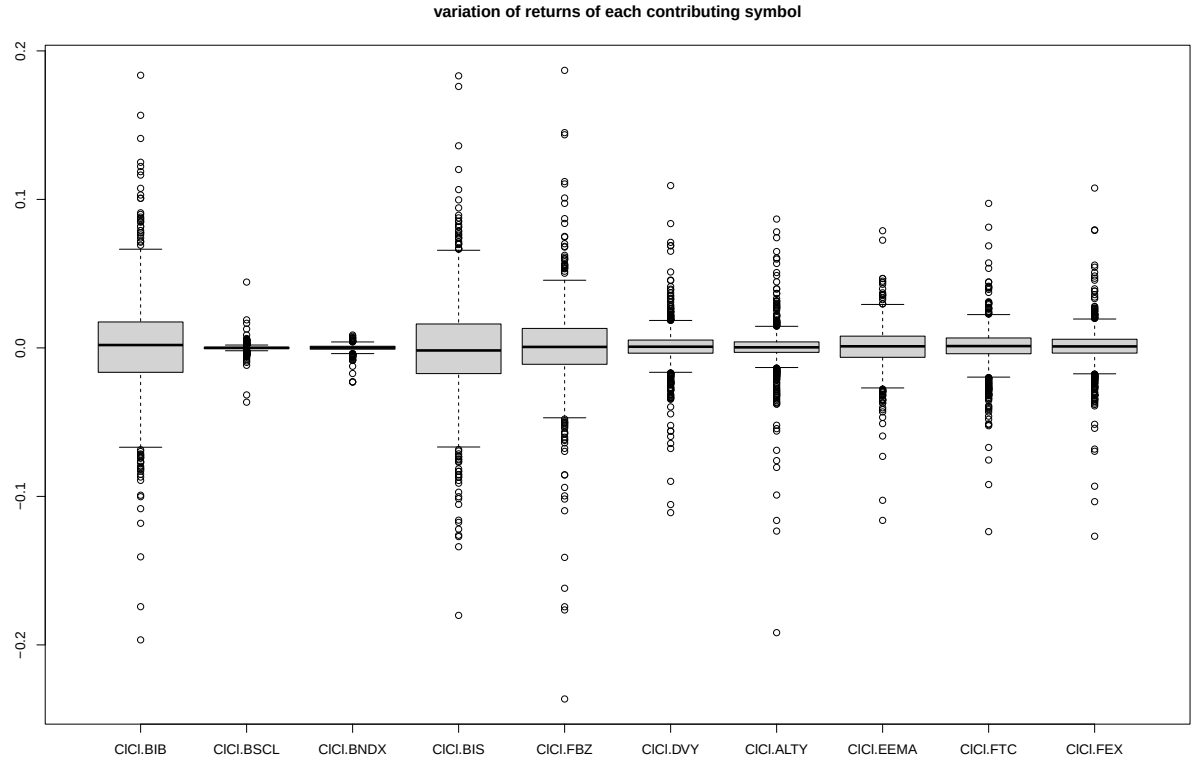


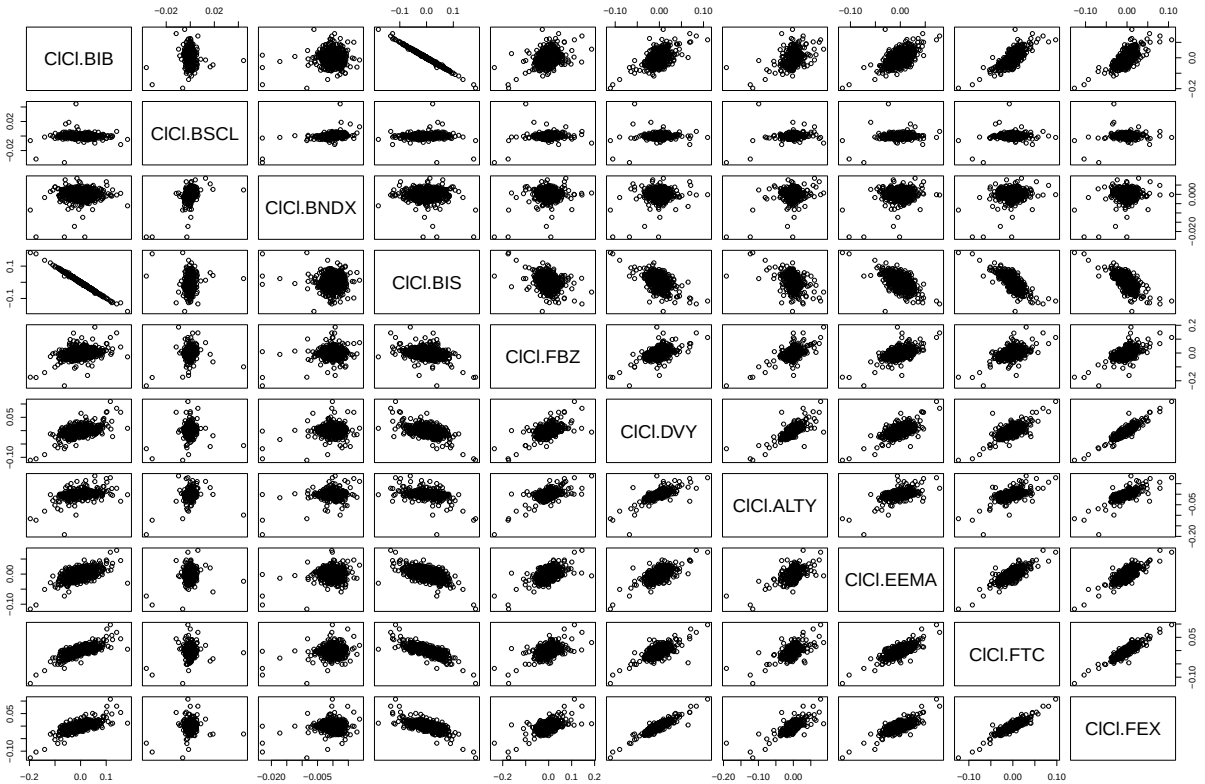

```
## stocks used BND BSCM AGZD BSCL BNDX

## 5% value at risk -1156.75

## pausing 1 second between requests for more than 5 symbols
## pausing 1 second between requests for more than 5 symbols
## pausing 1 second between requests for more than 5 symbols
## pausing 1 second between requests for more than 5 symbols
## pausing 1 second between requests for more than 5 symbols
## pausing 1 second between requests for more than 5 symbols

##               ClCl.BIB     ClCl.BSCL     ClCl.BNDX    ClCl.BIS     ClCl.FBZ
## 2016-01-04          NA            NA            NA          NA           NA
## 2016-01-05 -0.00119979  0.0019473709 -0.0005662011 0.003928012  0.004705882
## 2016-01-06 -0.03588587  0.0009718173  0.0037764540 0.036843797  0.000000000
## 2016-01-07 -0.08176297  0.0019416990 -0.0015048533 0.083018854 -0.043325527
## 2016-01-08 -0.03765260  0.0004845446 -0.0001884325 0.037456469  0.000000000
## 2016-01-11 -0.06979208 -0.0043584019 -0.0011306199 0.066610720  0.000000000
##                ClCl.DVY    ClCl.ALTY    ClCl.EEMA     ClCl.FTC     ClCl.FEX
## 2016-01-04           NA           NA           NA           NA           NA
## 2016-01-05  0.006035462  0.000000000  0.004707306  0.003799873  0.003043807
## 2016-01-06 -0.010132009  0.003646973 -0.018944795 -0.009463744 -0.015873016
## 2016-01-07 -0.017777778  0.000000000 -0.024501661 -0.018046667 -0.019449715
## 2016-01-08 -0.006855889 -0.017078488 -0.008514219 -0.012972930 -0.013788099
## 2016-01-11  0.002347066  0.000000000  0.000000000  0.001314283 -0.001962252
```



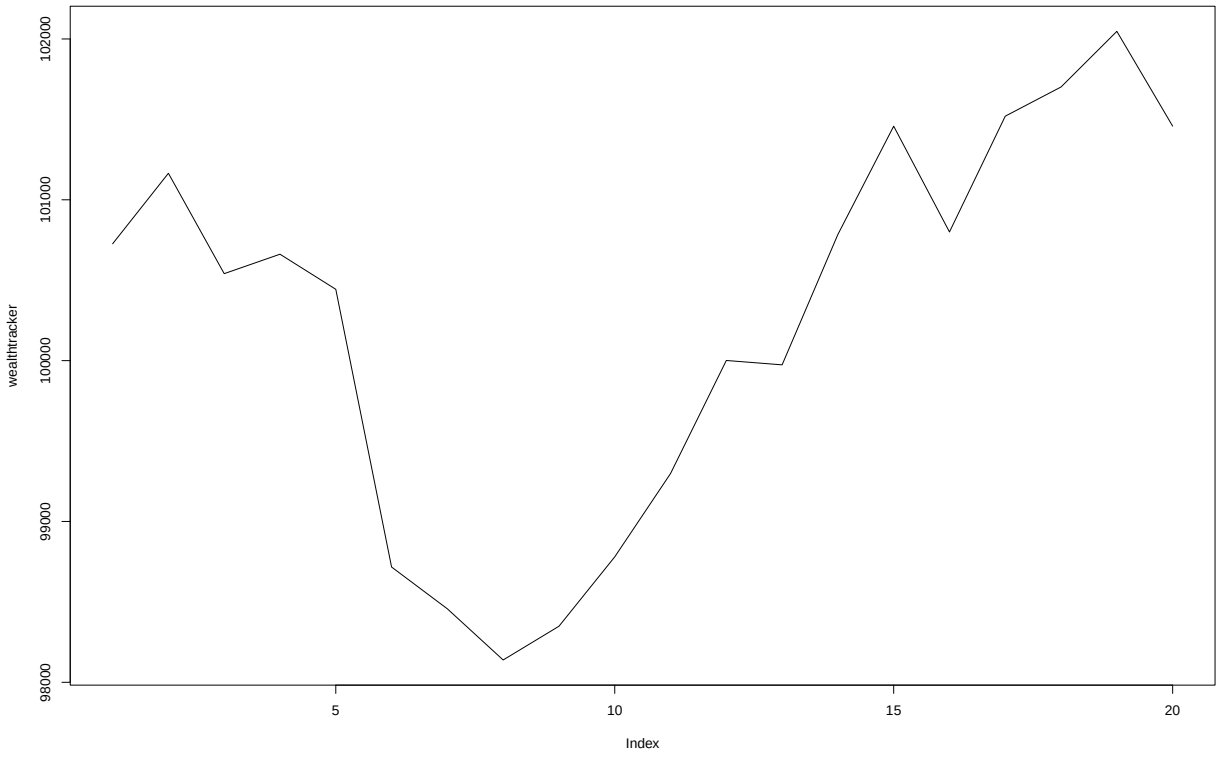


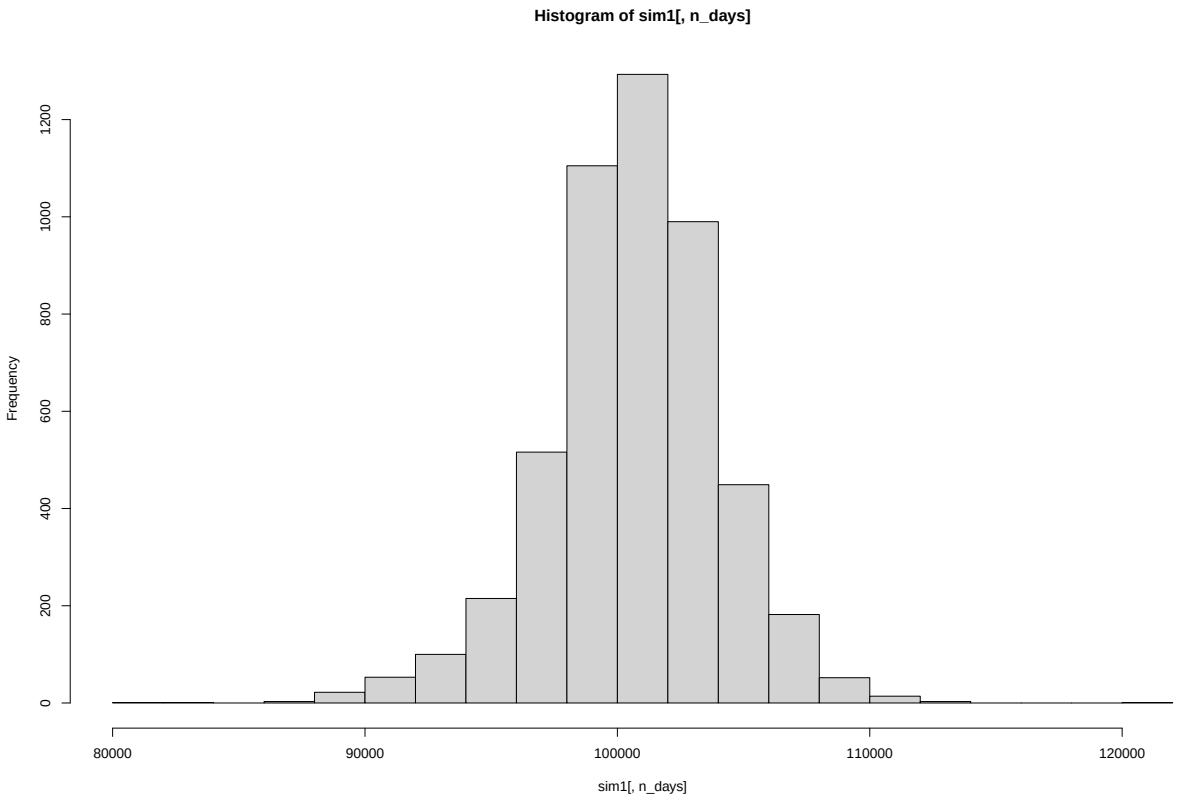

```
##             ClCl.BIB ClCl.BSCL ClCl.BNDX ClCl.BIS ClCl.FBZ ClCl.DVY ClCl.ALTY
```

```
## 2020-01-08 10175.96  9995.299  9980.637 9839.416 10087.31 10018.18   9986.55
##            ClCl.EEMA ClCl.FTC ClCl.FEX
## 2020-01-08  10080.38  10047.6 10029.14
```

```
## [1] 100240.5
```

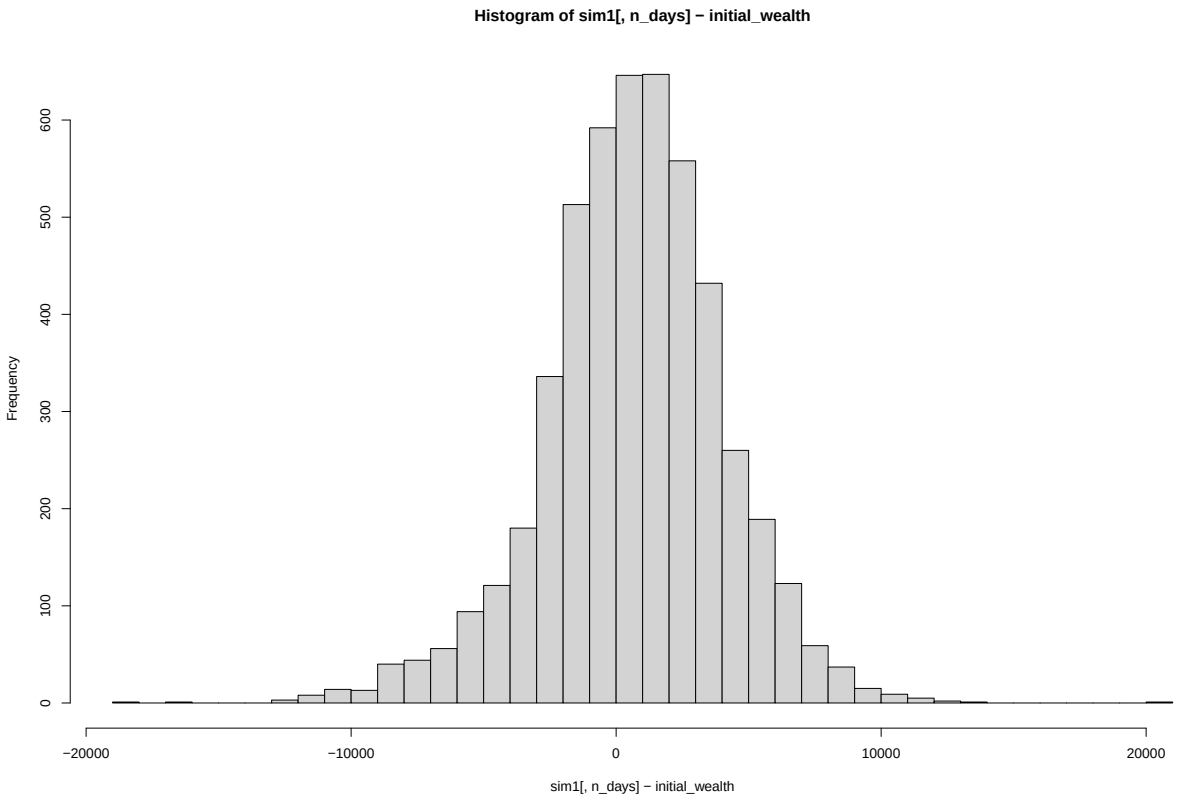
```
## 10072.63 10072.63 10072.63 10072.63 10072.63 10072.63 10072.63 10072.63 10072.63 10072.63
## 10116.5 10116.5 10116.5 10116.5 10116.5 10116.5 10116.5 10116.5 10116.5 10116.5
## 10054.07 10054.07 10054.07 10054.07 10054.07 10054.07 10054.07 10054.07 10054.07 10054.07
## 10066.16 10066.16 10066.16 10066.16 10066.16 10066.16 10066.16 10066.16 10066.16 10066.16
## 10044.35 10044.35 10044.35 10044.35 10044.35 10044.35 10044.35 10044.35 10044.35 10044.35
## 9871.673 9871.673 9871.673 9871.673 9871.673 9871.673 9871.673 9871.673 9871.673 9871.673
## 9845.678 9845.678 9845.678 9845.678 9845.678 9845.678 9845.678 9845.678 9845.678 9845.678
## 9813.87 9813.87 9813.87 9813.87 9813.87 9813.87 9813.87 9813.87 9813.87 9813.87
## 9834.849 9834.849 9834.849 9834.849 9834.849 9834.849 9834.849 9834.849 9834.849 9834.849
## 9877.98 9877.98 9877.98 9877.98 9877.98 9877.98 9877.98 9877.98 9877.98 9877.98
## 9929.78 9929.78 9929.78 9929.78 9929.78 9929.78 9929.78 9929.78 9929.78 9929.78
## 10000.09 10000.09 10000.09 10000.09 10000.09 10000.09 10000.09 10000.09 10000.09 10000.09
## 9997.371 9997.371 9997.371 9997.371 9997.371 9997.371 9997.371 9997.371 9997.371 9997.371
## 10078.47 10078.47 10078.47 10078.47 10078.47 10078.47 10078.47 10078.47 10078.47 10078.47
## 10145.76 10145.76 10145.76 10145.76 10145.76 10145.76 10145.76 10145.76 10145.76 10145.76
## 10080 10080 10080 10080 10080 10080 10080 10080 10080 10080
## 10152 10152 10152 10152 10152 10152 10152 10152 10152 10152
## 10170.15 10170.15 10170.15 10170.15 10170.15 10170.15 10170.15 10170.15 10170.15 10170.15
## 10204.72 10204.72 10204.72 10204.72 10204.72 10204.72 10204.72 10204.72 10204.72 10204.72
## 10145.81 10145.81 10145.81 10145.81 10145.81 10145.81 10145.81 10145.81 10145.81 10145.81
```






```
## [1] 100651.6
```

```
## [1] 651.5734
```




```
## stocks used BIB BSCL BNDX BIS FBZ DVY ALTY EEMA FTC FEX
```

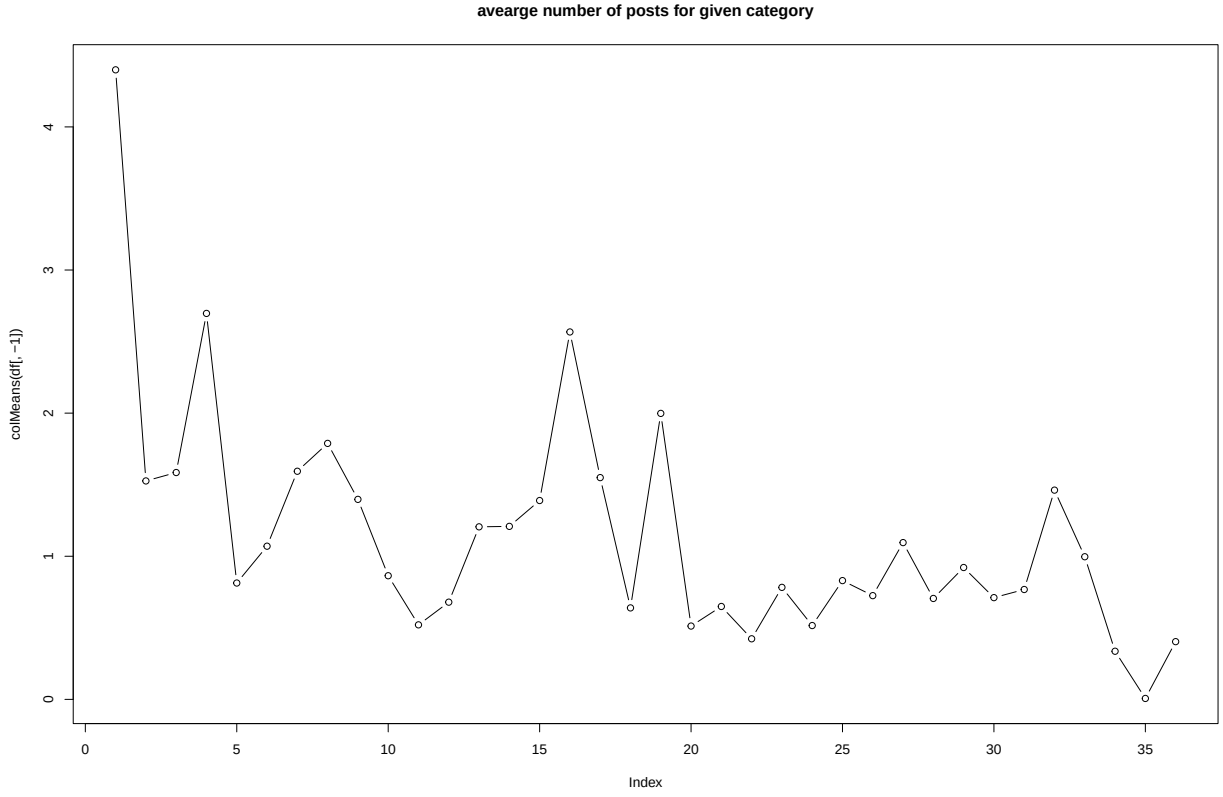
```
## 5% value at risk -5210.668
```

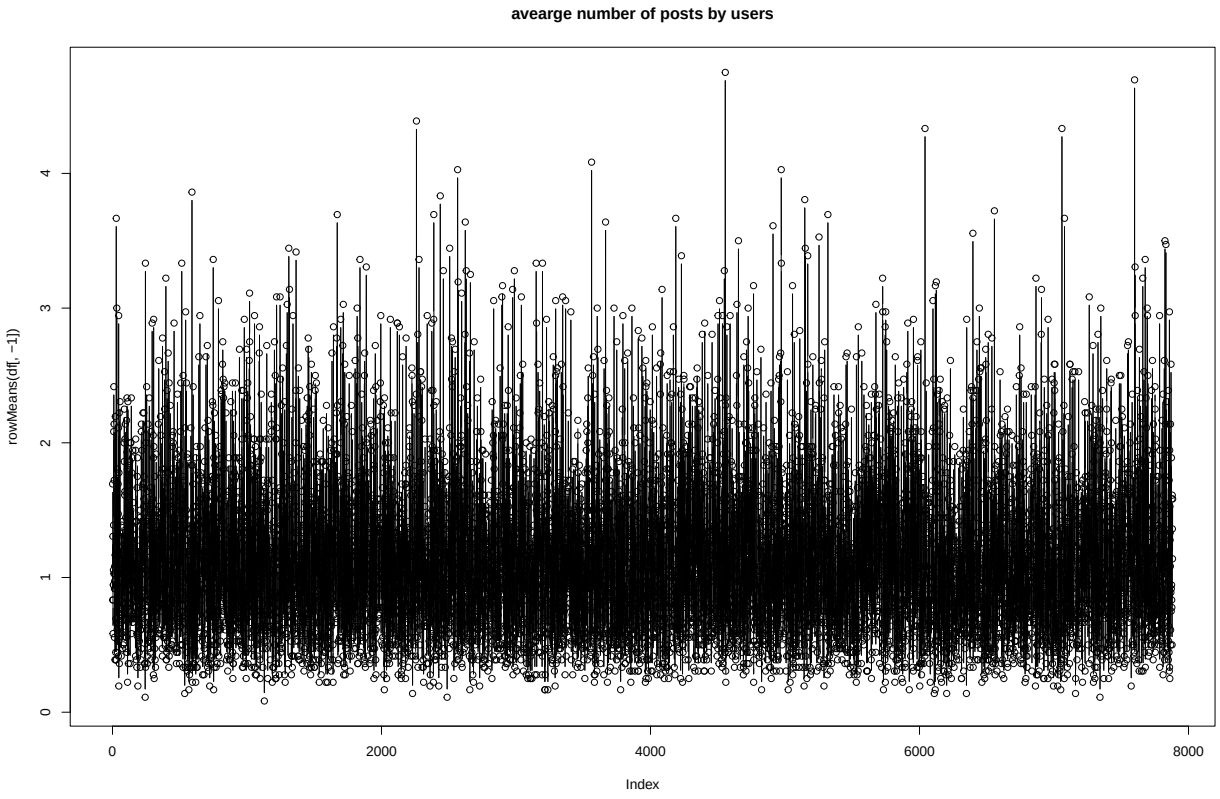
MARKET SEGMENTATION

Brief Analysis of the data

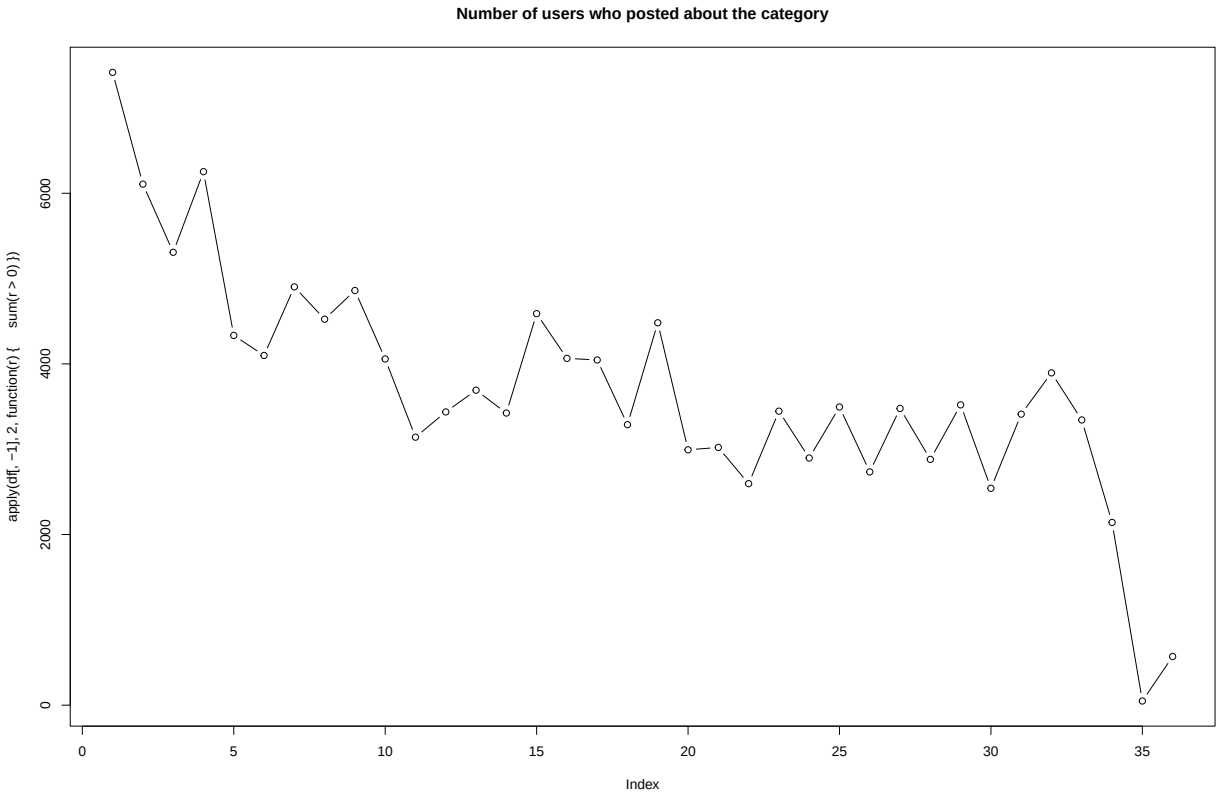
```
##       X                chatter        current_events      travel
##  Length:7882        Min.   : 0.000   Min.   :0.000   Min.   : 0.000
##  Class :character   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 0.000
##  Mode  :character   Median : 3.000   Median :1.000   Median : 1.000
##                     Mean   : 4.399   Mean   :1.526   Mean   : 1.585
##                     3rd Qu.: 6.000   3rd Qu.:2.000   3rd Qu.: 2.000
##                     Max.   :26.000   Max.   :8.000   Max.   :26.000
##  photo_sharing    uncategorized      tv_film      sports_fandom
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 1.000   1st Qu.:0.000   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 2.000   Median :1.000   Median : 1.00   Median : 1.000
##  Mean   : 2.697   Mean   :0.813   Mean   : 1.07   Mean   : 1.594
##  3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 1.00   3rd Qu.: 2.000
##  Max.   :21.000   Max.   :9.000   Max.   :17.00   Max.   :20.000
##     politics           food            family        home_and_garden
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 1.000   Median : 1.000   Median : 1.0000   Median :0.0000
##  Mean   : 1.789   Mean   : 1.397   Mean   : 0.8639   Mean   :0.5207
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000
##  Max.   :37.000   Max.   :16.000   Max.   :10.0000   Max.   :5.0000
##      music              news        online_gaming       shopping
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.0000   Median : 0.000   Median : 0.000   Median : 1.000
##  Mean   : 0.6793   Mean   : 1.206   Mean   : 1.209   Mean   : 1.389
##  3rd Qu.: 1.0000   3rd Qu.: 1.000   3rd Qu.: 1.000   3rd Qu.: 2.000
##  Max.   :13.0000   Max.   :20.000   Max.   :27.000   Max.   :12.000
##  health_nutrition  college_uni     sports_playing      cooking
##  Min.   : 0.000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median : 1.000   Median : 1.000   Median :0.0000   Median : 1.000
##  Mean   : 2.567   Mean   : 1.549   Mean   :0.6392   Mean   : 1.998
##  3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.: 2.000
##  Max.   :41.000   Max.   :30.000   Max.   :8.0000   Max.   :33.000
##       eco            computers          business         outdoors
##  Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.0000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   :0.5123   Mean   : 0.6491   Mean   :0.4232   Mean   : 0.7827
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.: 1.0000
##  Max.   :6.0000   Max.   :16.0000   Max.   :6.0000   Max.   :12.0000
##      crafts          automotive           art             religion
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.000
##  Median :0.0000   Median : 0.0000   Median : 0.0000   Median : 0.000
##  Mean   :0.5159   Mean   : 0.8299   Mean   : 0.7248   Mean   : 1.095
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.000
##  Max.   :7.0000   Max.   :13.0000   Max.   :18.0000   Max.   :20.000
##      beauty          parenting           dating            school
```

```
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.7052   Mean   : 0.9213   Mean   : 0.7109   Mean   : 0.7677
##  3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000
##  Max.   :14.0000   Max.   :14.0000   Max.   :24.0000   Max.   :11.0000
##  personal_fitness    fashion        small_business        spam
##  Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.000   Median : 0.0000   Median :0.0000   Median :0.00000
##  Mean   : 1.462   Mean   : 0.9966   Mean   :0.3363   Mean   :0.00647
##  3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :19.000   Max.   :18.0000   Max.   :6.0000   Max.   :2.00000
##      adult
##  Min.   : 0.0000
##  1st Qu.: 0.0000
##  Median : 0.0000
##  Mean   : 0.4033
##  3rd Qu.: 0.0000
##  Max.   :26.0000
```






```
##          chatter   current_events           travel    photo_sharing
##             7417             6107             5308             6253
##    uncategorized          tv_film    sports_fandom         politics
##             4334             4099             4903             4524
##             food           family  home_and_garden            music
##             4860             4058             3141             3437
##             news    online_gaming         shopping health_nutrition
##             3692             3424             4589             4065
##      college_uni   sports_playing          cooking              eco
##             4046             3288             4482             2993
##        computers         business         outdoors           crafts
##             3021             2596             3446             2897
##       automotive              art         religion           beauty
##             3495             2734             3478             2881
##        parenting           dating           school personal_fitness
##             3520             2542             3411             3895
##          fashion   small_business             spam            adult
##             3343             2142               49              570
```



1) There are 7882 turks and 36 categories of items

2) Almost all of the categories get posted on an average 1-2 number of times

3) There are some users who post more than 3-4 number of posts across all categories. However, mostly users post between 0-2 across all categories.

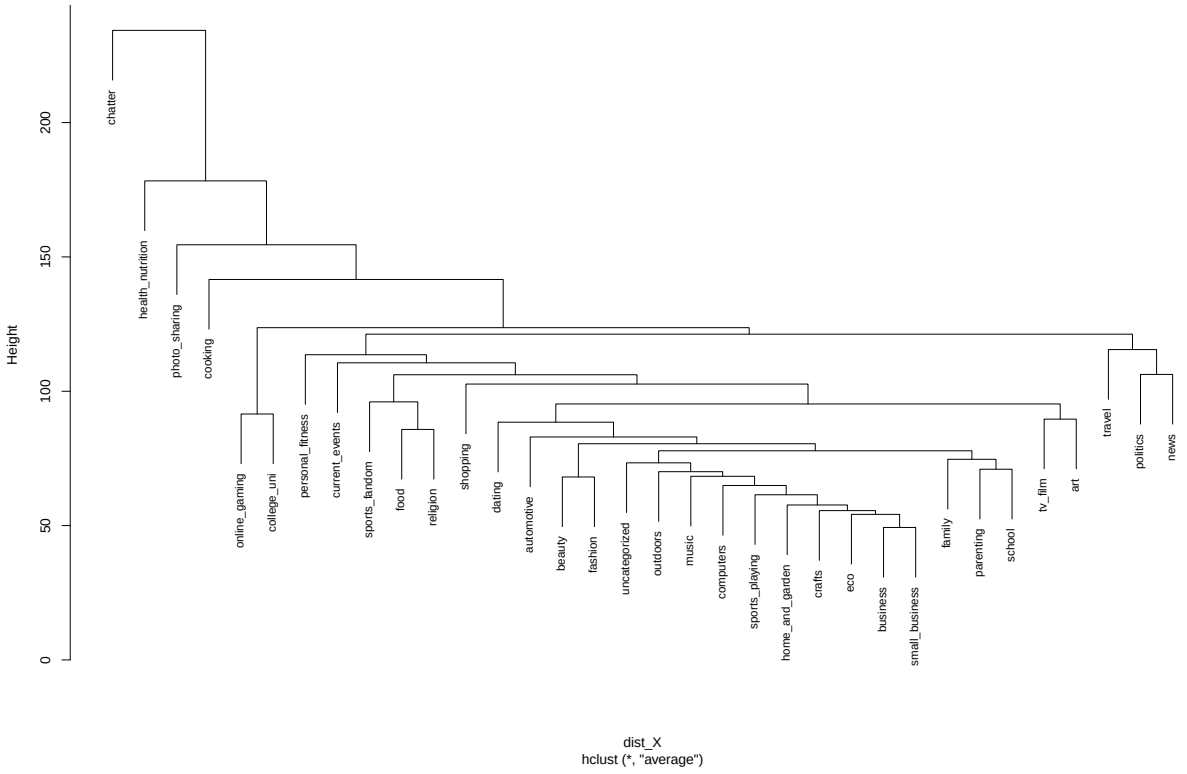
4) Topics related to categories like current events, health and sports got tweeted by more than 3000 users, whereas other topics like fashion and school got rarely covered.

5) We wanted to analyse which segments were closely related. We will be using the user_ids as different features/variables that will nhelp us determine the association between all the 36 segments

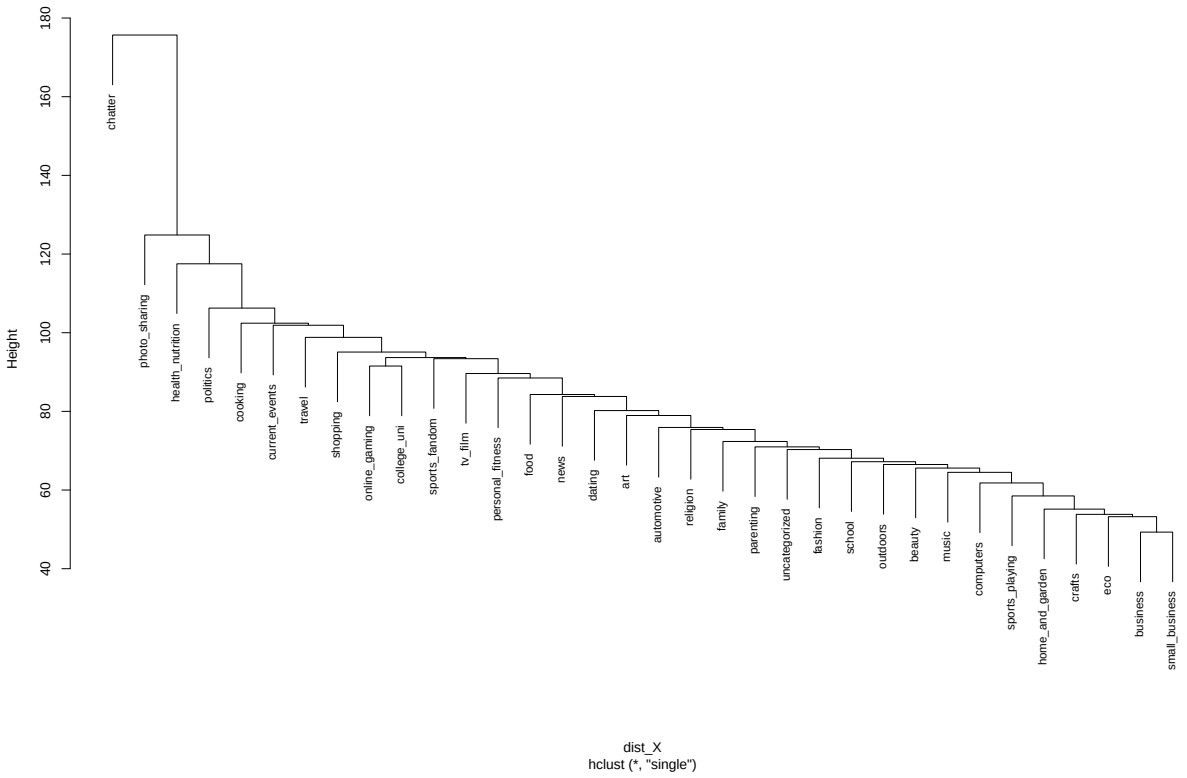
Hierarchial Clustering of the data

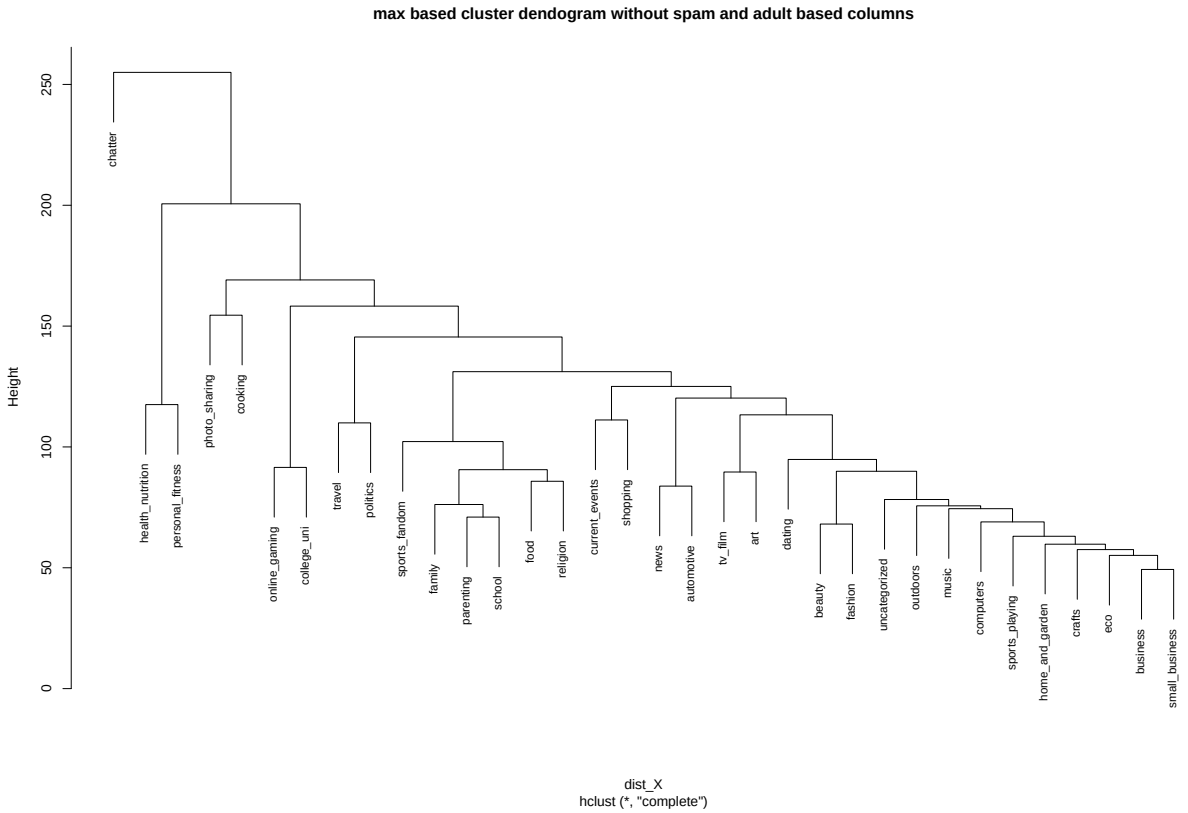
Steps - 1) Transposed the data, in order to analyse association between segments 2) Created a function providing with three different clustering plots, on the basis of association rule for point between two clusters. 3) Removed categories associated with span and adult based content

avg based cluster dendogram without spam and adult based columns



min based cluster dendogram without spam and adult based columns





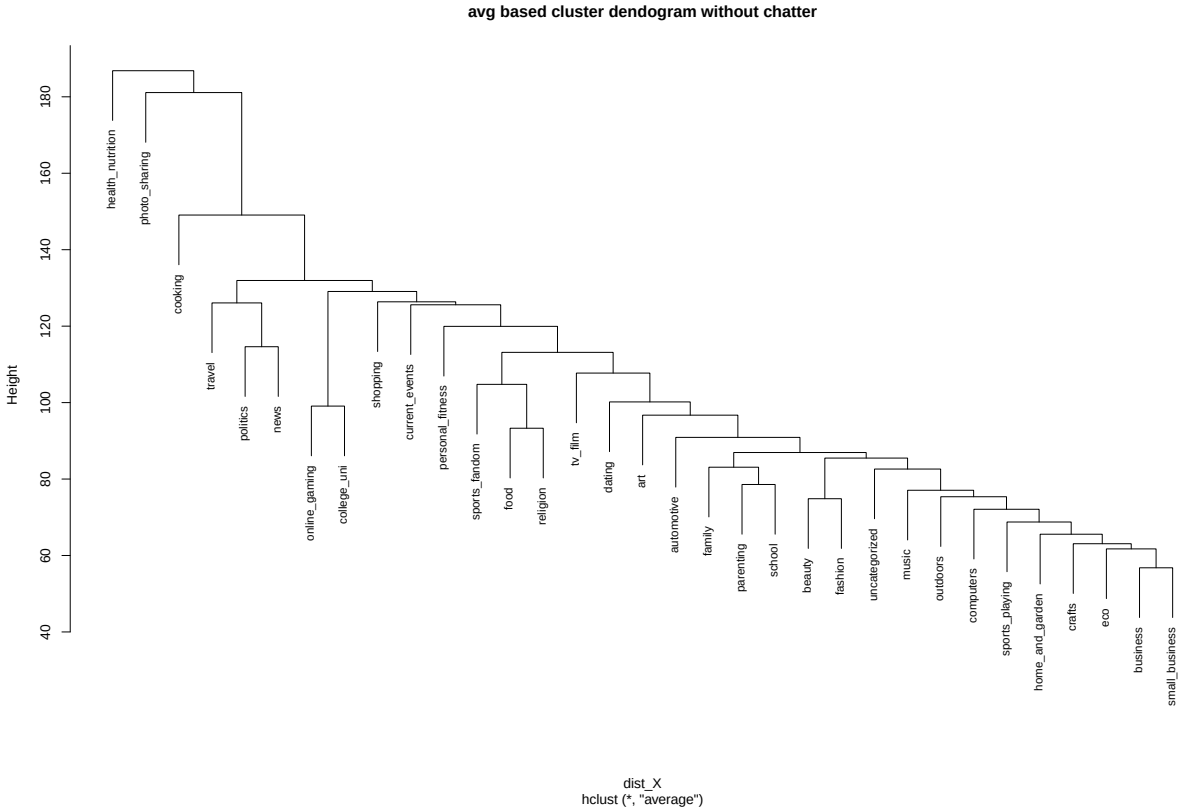
OBSERVATIONS

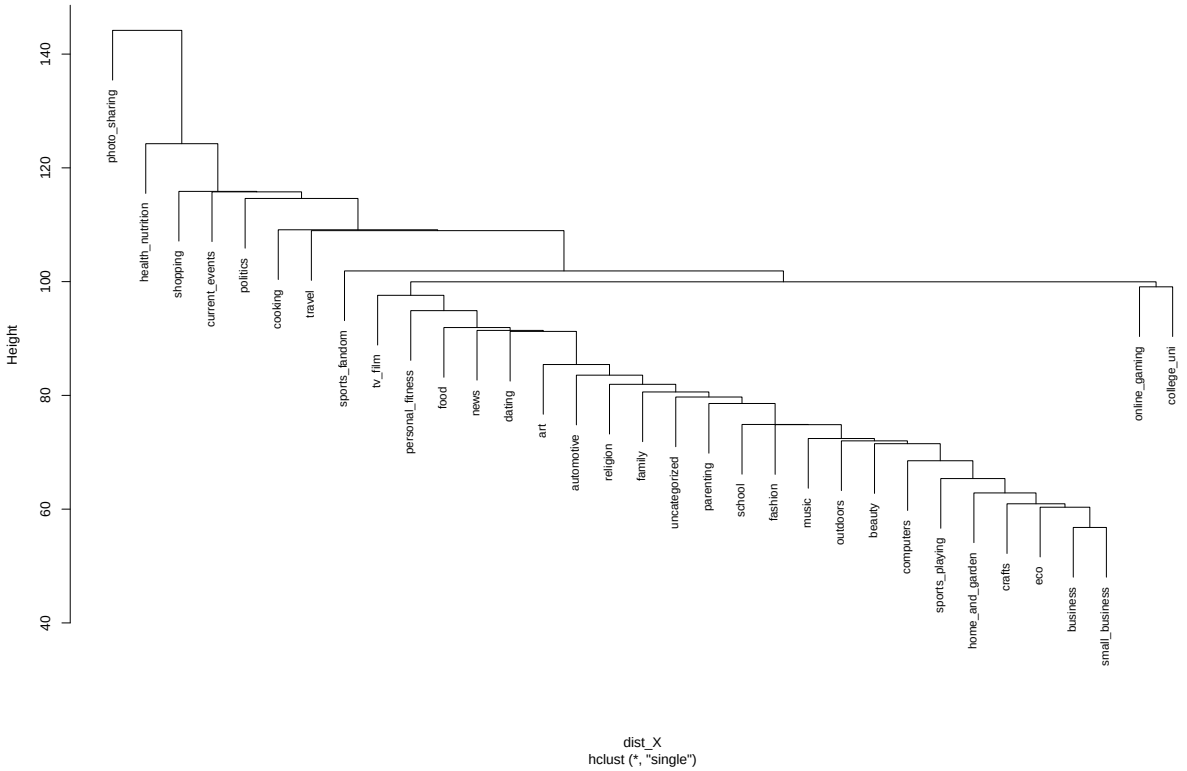
1) In all three kinds of associations used, chatter seems to be not unassociated and different(distance id large from other clusters) from other categories We decided to remove it from our further analysis

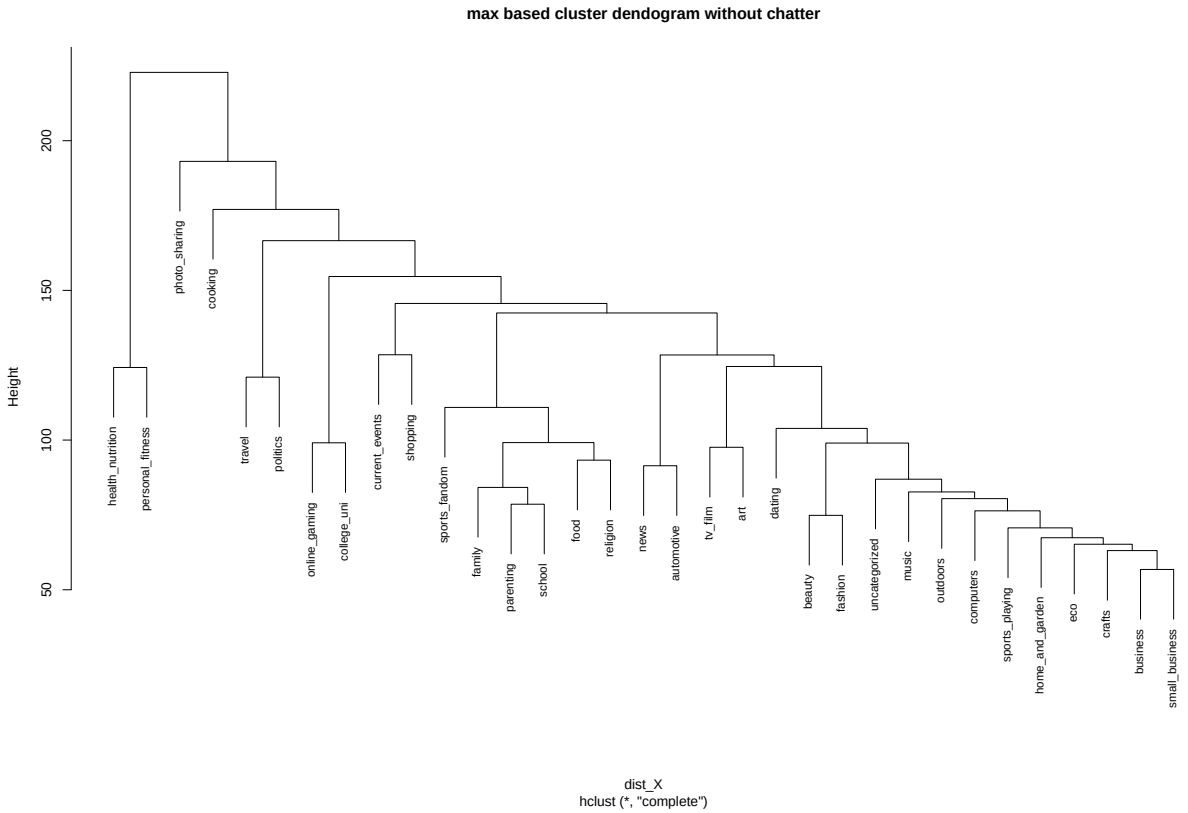
2) Minimum based association creates a large zone, comprising of all the categories. There can be at least one turk that has shared interest in two categories, therefore we can avoid using minimum association rule

3) Some categories like nutrition and fitness, family parenting school seems to be associated together in max based association rule, which make sense. Extreme turks of such cluster pars can be associated with each other.

WITHOUT CHATTER







OBSERVATIONS

1) Upon removing the "chatter" category and running hierarchical clustering on the others, we see that in the average based approach, there is a zone towards the right end with a significant number of clusters like businesses, sports, music, outdoors, home and gardening with crafts. Some clusters like online gaming with college or news with politics or beauty with fashion or parenting with school which were seen in the previous run are present here as well.

2) In the minimum based approach, we see that photo sharing or health nutrition does not seem to be associated with others which was also seen in the average approach. Online gaming and university are together and there is again a large zone with most of the clusters like businesses, computers, crafts and home or parenting and schools close to each other.

3) The maximum based approach gives a relatively more spread-out graph of the clustering maintaining the clusters like parenting with schools or gaming with college or tv with art and others. Moreover health nutrition and personal fitness are clubbed into one cluster and they are quite separated from the rest.

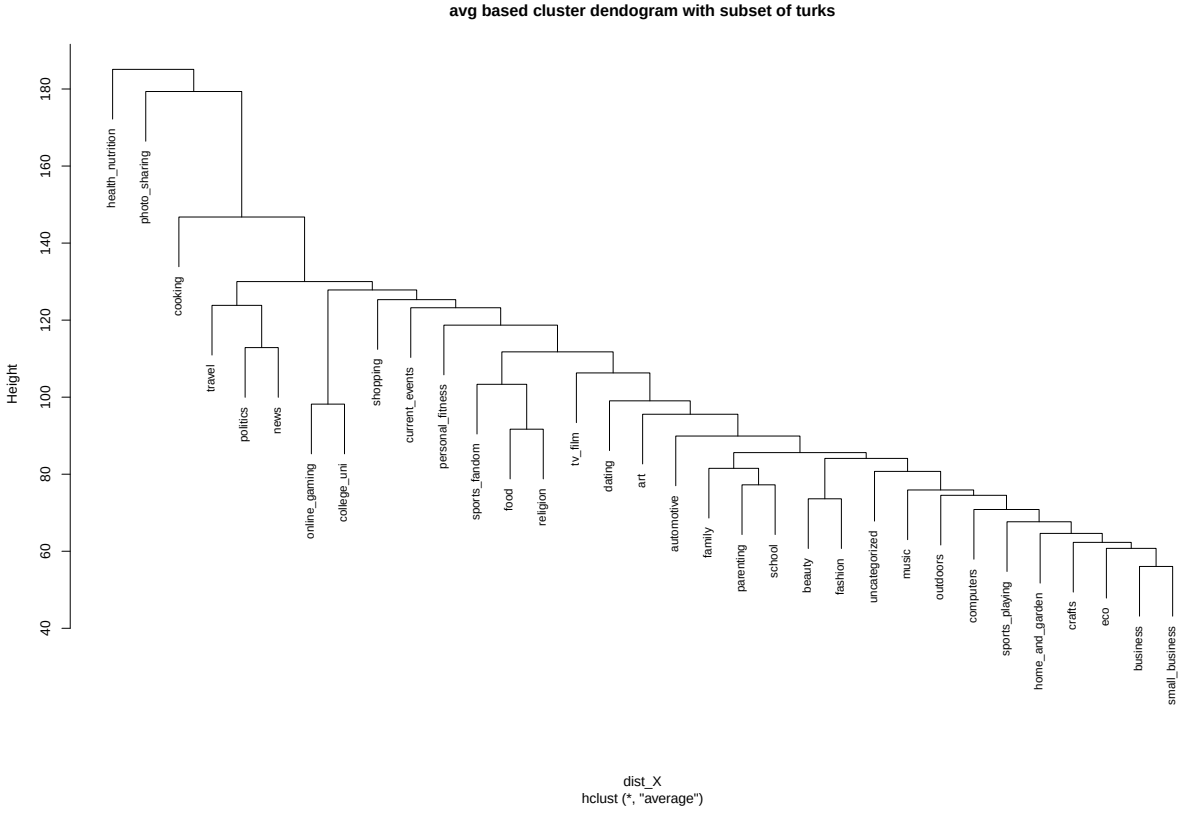
SUBSET of TURKS

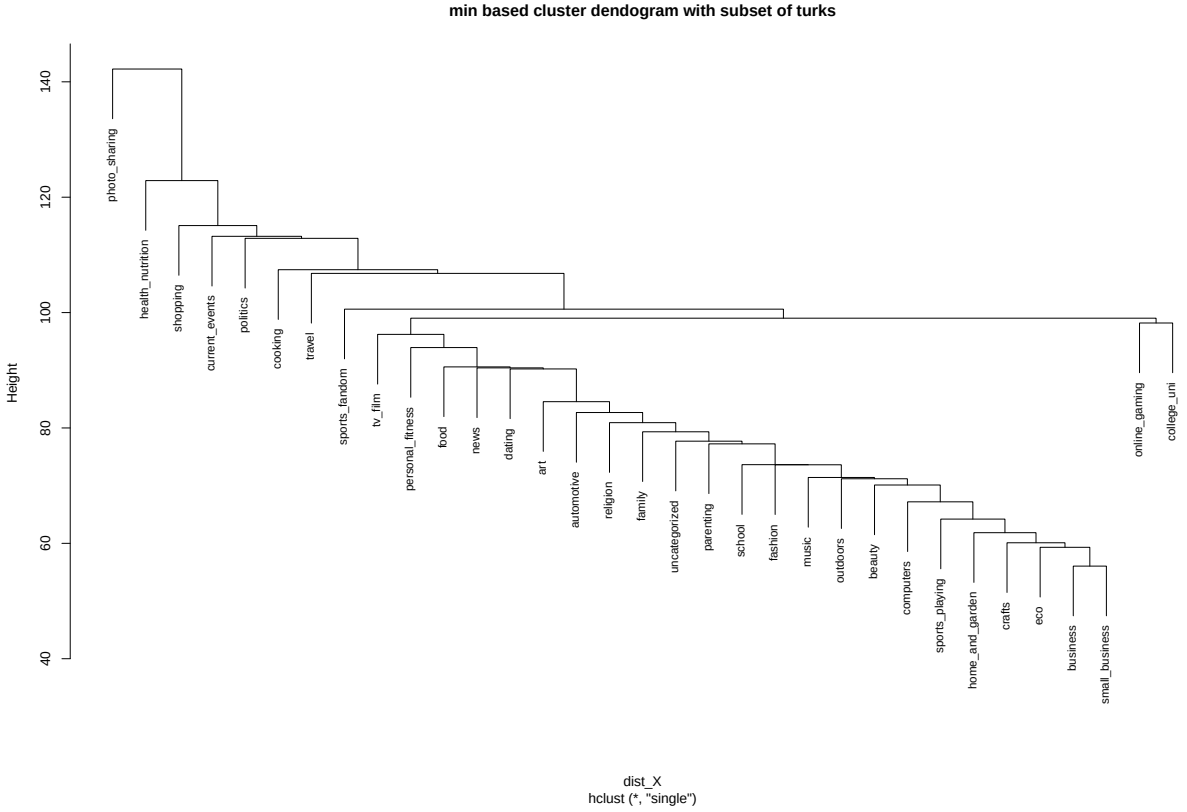
-We though of removing the turks who either post many tweets or who post less number of tweets

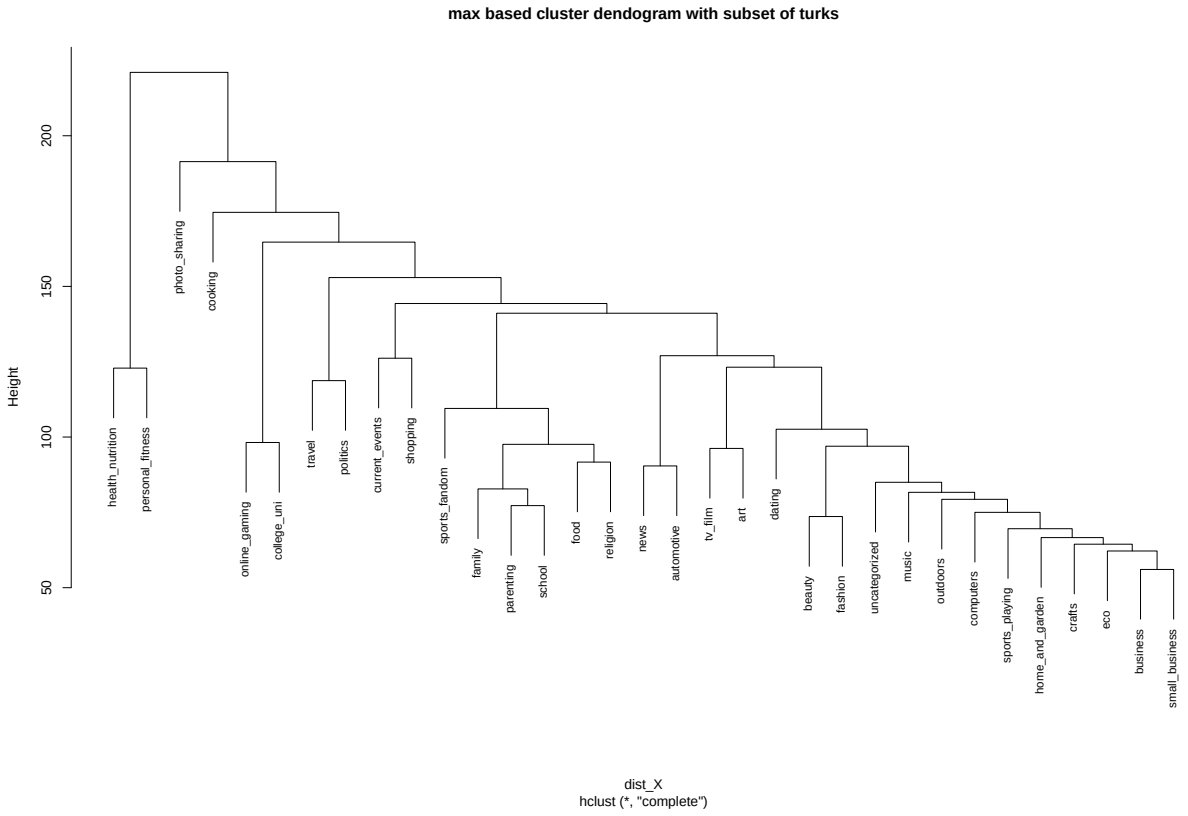
-Reliability on their tweets is low, especially while creating clusters of segments

-They act as outlier points lying on individual axes (people with less tweets) which will only distort the structure of association rules

-We used quantile based approach to perform this experiment


avg based cluster dendogram with subset of turks
Height
40
60
80
100
120
140
160
180
health_nutrition
photo_sharing
cooking
travel
politics
news
online_gaming
college_uni
shopping
current_events
personal_fitness
sports_fandom
food
religion
tv_film
dating
art
automotive
family
parenting
school
beauty
fashion
uncategorized
music
outdoors
computers
sports_playing
home_and_garden
crafts
eco
business
small_business
dist_X
hclust (*, "average")


min based cluster dendogram with subset of turks
Height
40
60
80
100
120
140
photo_sharing
health_nutrition
shopping
current_events
politics
cooking
travel
sports_fandom
tv_film
personal_fitness
food
news
dating
art
automotive
religion
family
uncategorized
parenting
school
fashion
music
outdoors
beauty
computers
sports_playing
home_and_garden
crafts
eco
business
small_business
online_gaming
college_uni
dist_X
hclust (*, "single")



OBSERVATIONS

We see that there is a reduction of 200 turks after removing the top and bottom 1 percent.

We receive an overall similar plot like the one received without chatters. The maximum based approach cluster is the most spread-out with the same clusters as seen previously. The minimum and the average approach clusters have a zone with significant number of clusters in it. Some clusters like news with politics or food with religion which occur together in the avg based approach get separated out in the min based approach.

AUTHOR ATTRIBUTION

##APPROACH 1 - PCA and RandomForest BAsed

1) TRAINING SET

- Created a corpus of all the fifty text files belonging to fifty different authors
- Applied transformations associated with lower casing, removal pf numbers, punctuations and whitespaces
- Removed stopwords, that are present in the smart vocabulary list, from the corpus
- Formulated a Bag of words based Sparse matrix for the training corpus

```
o   Number of terms 32241
o   Applied sparsity-based filter to remove columns with more than 98.5
sparsity, terms reduced to 2326
```

- Computed TFIDF scores for each document for the set of words
- Removed columns with sum across column values of TFIDF equal to zero
- Dimension changed to rows=2500 col/words=2310
- Stored the set of words used in training set in a placeholder named (words_from_trainset)
- Created a 'pseudo-word-column' for the words that may surface in test set but are not present in training set o Value for pseudo word was determined individually for each document, it was equal to the 0.2 quantile value of all the non-zero tfidf for words present in the specific row.
- Applied PCA transformation for the entire column sets

```
o   We plotted cumsum of variation_explaination with number of components
and found out that we will require at least 500 words to explain
more than 60 percent of the variation

o   We stored the feature contribution of each words (2311) in these
500 axes in another placeholder, we will use this to transform our test set
```

- Now we had our predictors, we formulated the target variable using rep function 50 authors repeated 50 times
- Then we trained a random forest-based classifier model between pca predictors and target variables

2) TEST SET

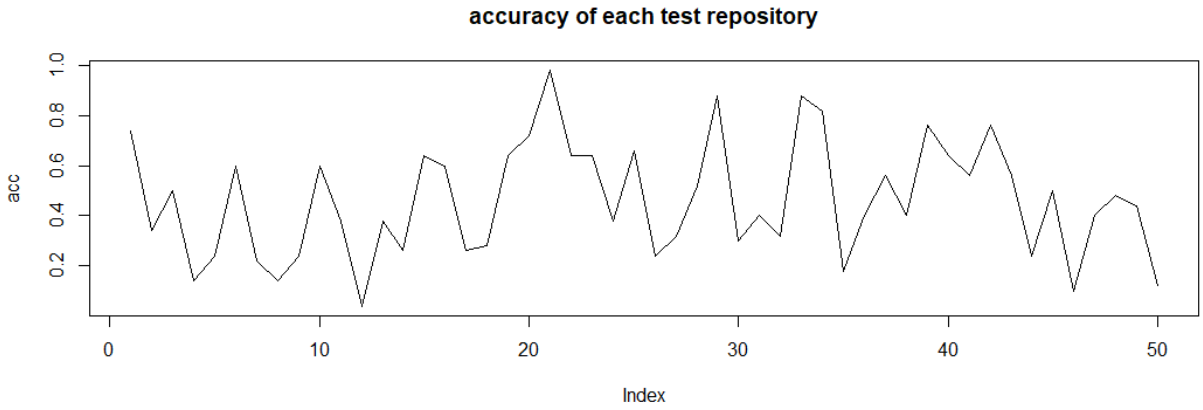
- We used similar transformation as applied on the train set to create TFIDF matrix of the test set
- Then we found out the words that are common in the training and test set using intersection function, it turned out to be 1999
- We realise that we can only use contribution of these 1999 and 1 pseudo word feature for creating our 50 test pca predictors
- We computed mean of the tfidf scores of the words that are not present in the train set and set it as pseudo variable
- We used words present in both sets and pseudo word based variable to select pca rotation contribution, followingly computed pca scores for each of the test document

Evaluated the test set accuracy

The approach seems to be appreciable in selecting right number of words for each repository. Additionally, The model was flexible enough to provide room for new words present in the test set

However, the accuracy on the test repository was not that good, so we decided to switch to Naive Bayes based simple composite model.

Approach 2- Naive Bayes based approach



Workflow Steps

• For each test set, we trained 50 different naïve bayes model based on the word bag count of each repository of documents belonging to 50 individual writers

• For all trained model we computed the probability of occurrence of each word specific to that repository/folder for individual writer

• Following, we took intersection of words that occur in both the train and the test set We had this (50 combinations of intersecting words for each test folder)

• Subsequently, we computed likelihood using the product of log of prior probability and counts of intersecting words.

o We obtained 50 likelihood scores from all the trained models for each of the 50 documents present in the repository

o We analysed which model gave the highest/max likelihood for each document and assigned the value of that model as predicted folder set for the document

o Then we had a series of 50 documents with their predicted set/author

- We computed accuracy of for every test folder and stored it in a place holder variable

Finally, we plotted the computed accuracy for each test folder.

We observed a large variance in test accuracy prediction, for example documents associated with author number 21, Karl Penhaul were predicted with an accuracy of more than 0.98.

Whereas, files attributed to author 12, Graham Earnshaw was predicted with just an accuracy of 0.04

We pondered over improving the time complexity of the model by creating a placeholder for all the weights/likelihood of words of training corpus.

ASSOCIATE RULE MINING

Treatment of the data

In this step, we first took the Groceries dataset where each row represented the items bought by a user We then created a dataframe to have only 2 columns - user_id and product where user_id is the row number for groceries datasets and grouped the products by user_id

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.2     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x purrr::accumulate()        masks foreach::accumulate()
## x dplyr::combine()           masks EBImage::combine()
## x mosaic::count()            masks dplyr::count()
## x purrr::cross()             masks mosaic::cross()
## x mosaic::do()               masks dplyr::do()
## x tidyr::expand()            masks Matrix::expand()
## x dplyr::filter()            masks stats::filter()
## x dplyr::first()             masks xts::first()
## x ggstance::geom_errorbarh() masks ggplot2::geom_errorbarh()
## x dplyr::lag()               masks stats::lag()
## x dplyr::last()              masks xts::last()
## x tidyr::pack()              masks Matrix::pack()
## x mosaic::stat()             masks ggplot2::stat()
## x mosaic::tally()            masks dplyr::tally()
## x purrr::transpose()         masks EBImage::transpose()
## x tidyr::unpack()            masks Matrix::unpack()
## x purrr::when()              masks foreach::when()

##
## Attaching package: 'arules'

## 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
```

Summary of the products distribution

```
## transactions as itemMatrix in sparse format with
##  15296 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.01677625
##
## 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
## 3485 2630 2102 7079
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.835   4.000   4.000
##
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
##
## includes extended transaction information - examples:
##   transactionID
## 1             1
## 2             2
## 3             3
```

OBSERVATIONS

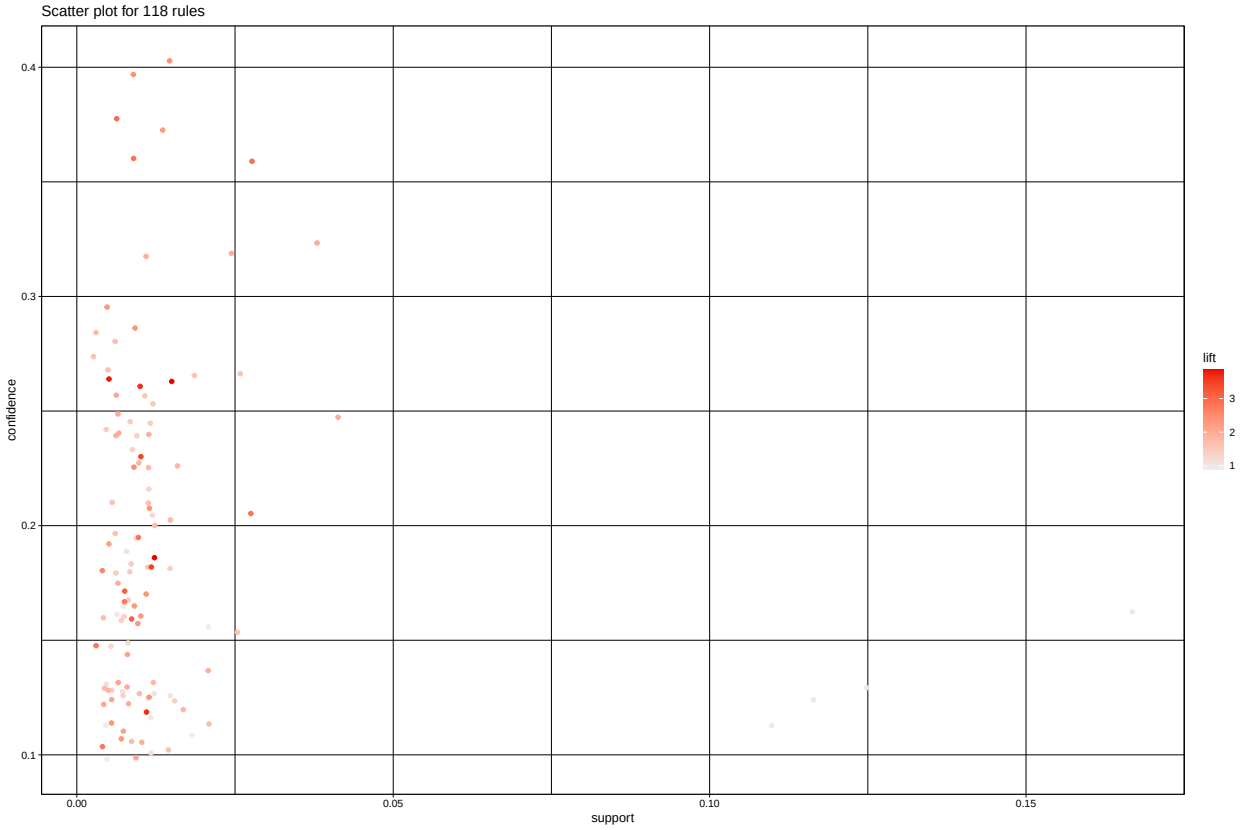
In the summary of the prod_dist, we find that "whole milk" was the highest bought item followed by "other vegetables" Also, we found that in majority (around 7079 times), 4 items were bought in a single transaction.

Running the apriori function to get the support, confidence and lift

```
## 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: 76
```

```
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 15296 transaction(s)] done [0.00s].
## sorting and recoding items ... [101 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [118 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].

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



OBSERVATIONS

From the scatter plot we found that there are a few items with support less than 0.05 but have a very high confidence above 0.3 which indicated that there is a good association between these items So we decided to further inspect these items.

Inspection steps and Observations

We started with the inspection of product combination with confidence above 0.3

```
##     lhs                    rhs                support
## [1] {onions}            => {other vegetables} 0.007452929
## [2] {curd}              => {whole milk}       0.012617678
```

```
## [3] {butter}                           => {whole milk}       0.014382845
## [4] {root vegetables}                  => {other vegetables} 0.025366109
## [5] {root vegetables}                  => {whole milk}       0.022620293
## [6] {other vegetables}                 => {whole milk}       0.040860356
## [7] {other vegetables,root vegetables} => {whole milk}       0.008172071
## [8] {root vegetables,whole milk}       => {other vegetables} 0.008172071
## [9] {other vegetables,yogurt}          => {whole milk}       0.006341527
##     confidence coverage   lift     count
## [1] 0.3737705  0.01993985 3.004306 114
## [2] 0.3683206  0.03425732 2.241875 193
## [3] 0.4036697  0.03563023 2.457036 220
## [4] 0.3619403  0.07008368 2.909216 388
## [5] 0.3227612  0.07008368 1.964566 346
## [6] 0.3284288  0.12441161 1.999064 625
## [7] 0.3221649  0.02536611 1.960937 125
## [8] 0.3612717  0.02262029 2.903842 125
## [9] 0.3991770  0.01588651 2.429690  97
```

Upon inspection, we found that "onions" and "other vegetables" have the highest lift indicating a strong tendency of other vegetables being bought with onions Same is the case with "root vegetables" and "other vegetables" or vegetables along with milk which have lift close to 3 Among these, we find the highest count of 625 for the combination of "other vegetables" and "whole milk".

Next, we inspect the product combinations with count > 150 and lift > 1.5

```
##      lhs                    rhs                support    confidence
## [1]  {curd}                 => {whole milk}       0.01261768 0.3683206
## [2]  {butter}               => {whole milk}       0.01438285 0.4036697
## [3]  {whipped/sour cream}   => {whole milk}       0.01144090 0.2482270
## [4]  {pip fruit}            => {tropical fruit}   0.01268305 0.2607527
## [5]  {tropical fruit}       => {pip fruit}        0.01268305 0.1879845
## [6]  {pip fruit}            => {other vegetables} 0.01091789 0.2244624
## [7]  {pip fruit}            => {whole milk}       0.01255230 0.2580645
## [8]  {pastry}               => {rolls/buns}       0.01019874 0.1782857
## [9]  {citrus fruit}         => {tropical fruit}   0.01248692 0.2346437
## [10] {tropical fruit}       => {citrus fruit}     0.01248692 0.1850775
## [11] {citrus fruit}         => {other vegetables} 0.01281381 0.2407862
## [12] {other vegetables}     => {citrus fruit}     0.01281381 0.1029953
## [13] {sausage}              => {rolls/buns}       0.01078713 0.1785714
## [14] {sausage}              => {other vegetables} 0.01261768 0.2088745
## [15] {other vegetables}     => {sausage}          0.01261768 0.1014188
## [16] {bottled water}        => {soda}             0.01464435 0.2060718
## [17] {soda}                 => {bottled water}    0.01464435 0.1306122
## [18] {tropical fruit}       => {root vegetables}  0.01098326 0.1627907
## [19] {root vegetables}      => {tropical fruit}   0.01098326 0.1567164
## [20] {tropical fruit}       => {other vegetables} 0.01549425 0.2296512
## [21] {other vegetables}     => {tropical fruit}   0.01549425 0.1245402
## [22] {tropical fruit}       => {whole milk}       0.01830544 0.2713178
## [23] {whole milk}           => {tropical fruit}   0.01830544 0.1114206
## [24] {root vegetables}      => {other vegetables} 0.02536611 0.3619403
## [25] {other vegetables}     => {root vegetables}  0.02536611 0.2038886
## [26] {root vegetables}      => {whole milk}       0.02262029 0.3227612
## [27] {whole milk}           => {root vegetables}  0.02262029 0.1376840
## [28] {yogurt}               => {whole milk}       0.02425471 0.2704082
```

```
## [29] {whole milk}         => {yogurt}           0.02425471 0.1476323
## [30] {other vegetables}   => {whole milk}       0.04086036 0.3284288
## [31] {whole milk}         => {other vegetables} 0.04086036 0.2487067
##      coverage   lift     count
## [1]  0.03425732 2.241875 193
## [2]  0.03563023 2.457036 220
## [3]  0.04609048 1.510895 175
## [4]  0.04864017 3.864800 194
## [5]  0.06746862 3.864800 194
## [6]  0.04864017 1.804191 167
## [7]  0.04864017 1.570774 192
## [8]  0.05720450 1.507495 156
## [9]  0.05321653 3.477820 191
## [10] 0.06746862 3.477820 191
## [11] 0.05321653 1.935400 196
## [12] 0.12441161 1.935400 196
## [13] 0.06040795 1.509911 165
## [14] 0.06040795 1.678898 193
## [15] 0.12441161 1.678898 193
## [16] 0.07106433 1.837944 224
## [17] 0.11212082 1.837944 224
## [18] 0.06746862 2.322805 168
## [19] 0.07008368 2.322805 168
## [20] 0.06746862 1.845898 237
## [21] 0.12441161 1.845898 237
## [22] 0.06746862 1.651444 280
## [23] 0.16429132 1.651444 280
## [24] 0.07008368 2.909216 388
## [25] 0.12441161 2.909216 388
## [26] 0.07008368 1.964566 346
## [27] 0.16429132 1.964566 346
## [28] 0.08969665 1.645907 371
## [29] 0.16429132 1.645907 371
## [30] 0.12441161 1.999064 625
## [31] 0.16429132 1.999064 625
```

Upon putting the above condition of count and lift, we find that “pip fruit” and “tropical fruit” have a lift close to 4 indicating high association between them

We find that people generally buy whole milk and other vegetables together upon inspecting for lift > 1 and support > 0.03

```
##     lhs                   rhs                support    confidence coverage
## [1] {other vegetables} => {whole milk}       0.04086036 0.3284288  0.1244116
## [2] {whole milk}       => {other vegetables} 0.04086036 0.2487067  0.1642913
##     lift     count
## [1] 1.999064 625
## [2] 1.999064 625
```

Upon putting the condition of support > 0.10, we find that these are values for individual items like “soda” or “rolls/buns”

```
##     lhs    rhs               support   confidence coverage lift count
## [1] {}  => {soda}             0.1121208 0.1121208  1        1    1715
```

```
## [2] {}  => {rolls/buns}       0.1182662 0.1182662  1        1    1809
## [3] {}  => {other vegetables} 0.1244116 0.1244116  1        1    1903
## [4] {}  => {whole milk}       0.1642913 0.1642913  1        1    2513
```

NETWORK GRAPHS

We then plot the network graph to see the association between the different items bought.

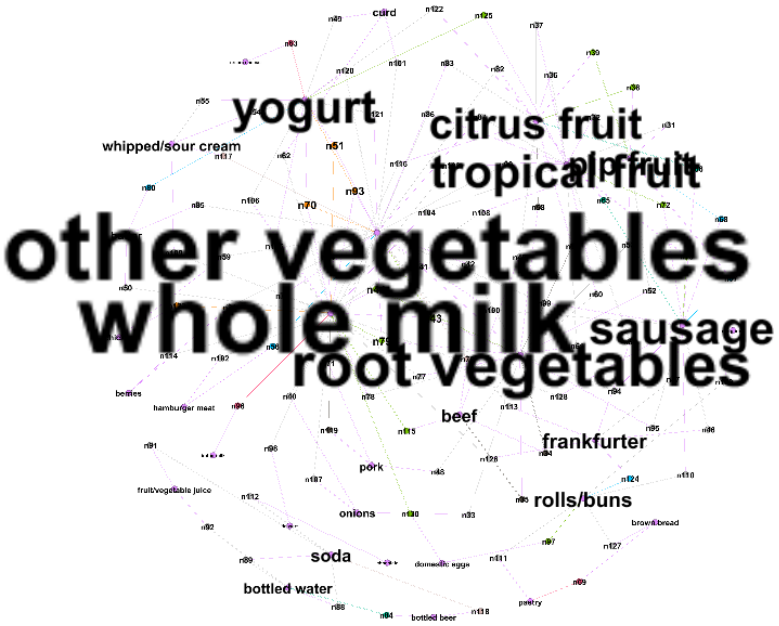
OBSERVATIONS

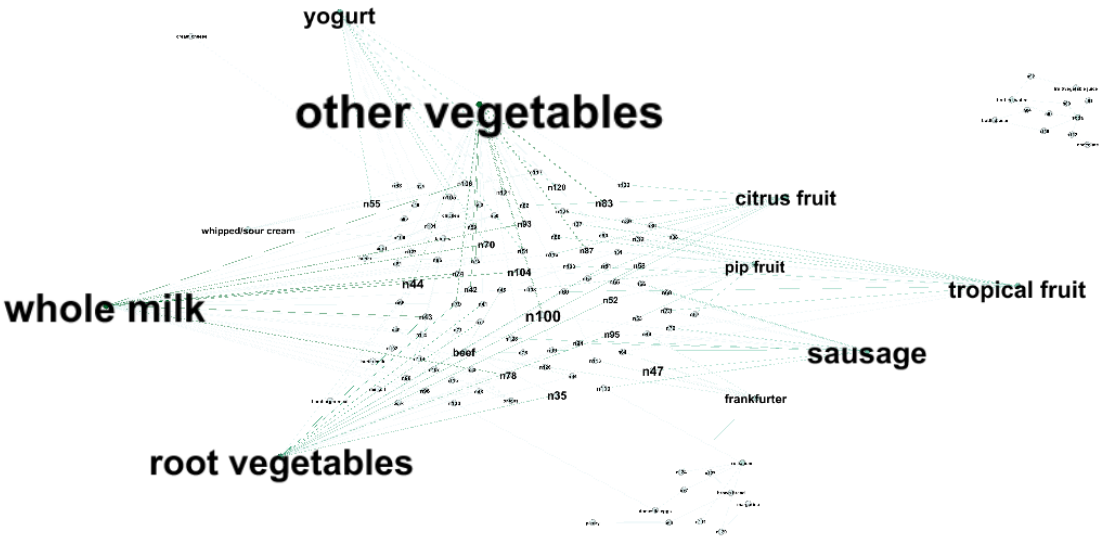
1)We get to see that other vegetables and whole milk are the items which are highest associated with other items i.e. they have the highest network.

2) Also, we find bottled water, beer, soda, chocolate and juice belong to a completely different cluster indicating they are not associated with vegetables or whole milk or other regular items

GEPHI PLOTS

Finally, we use the Gephi software to visualize the association between the items.


yogurt
citrus fruit
whipped/sour cream
tropical fruit
other vegetables
whole milk
sausage
root vegetables
beef
frankfurter
pork
rolls/buns
onions
soda
bottled water


yogurt
other vegetables
citrus fruit
whipped/sour cream
pip fruit
tropical fruit
whole milk
n100
sausage
beef
n47
frankfurter
root vegetables

OBSERVATIONS

1) Here we see that “other vegetables” and “whole milk” have the biggest font followed by “root vegetables”, “yogurt”, “sausage” and others.

2) The font of an item is proportional to the degree of the node which in turn is related to the association of that item. The Gephi plot also validates the earlier network plot indicating that vegetables and whole milk have highest association and are often bought with other items

3) In the second Gephi plot we find that there are 2 clusters validating our earlier inference that bottled water, beer, soda are not associated with the regular items and are mostly bought separately.