STA 380: Intro to Machine Learning, Part 2: Exercises

Aritra Chowdhury, Soumi Basu, Vishu Agarwal, Yashpreet Kaur

08/16/2021

Github link : https://github.com/aric95/STA-380-Intro-to-ML-Exercises-2

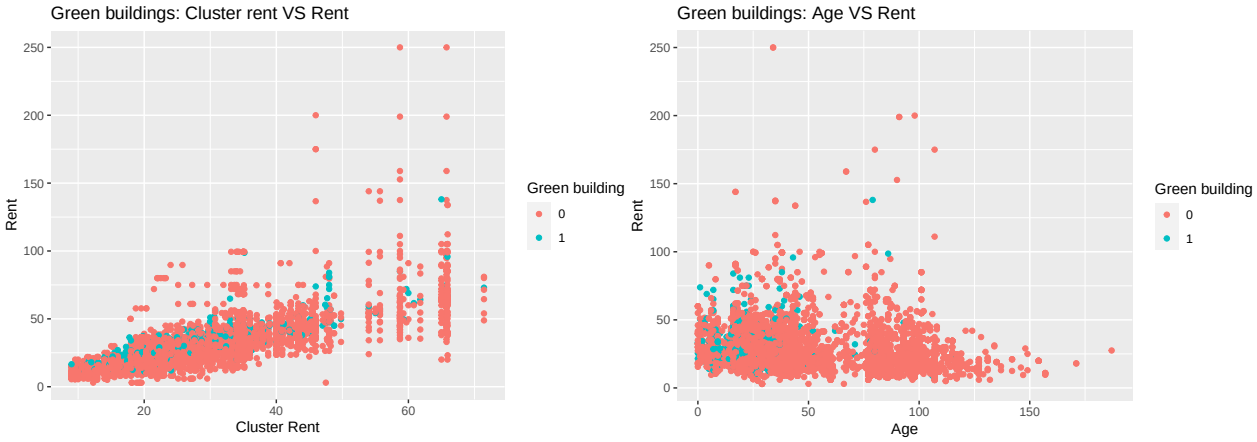
Visual story telling part 1: green buildings

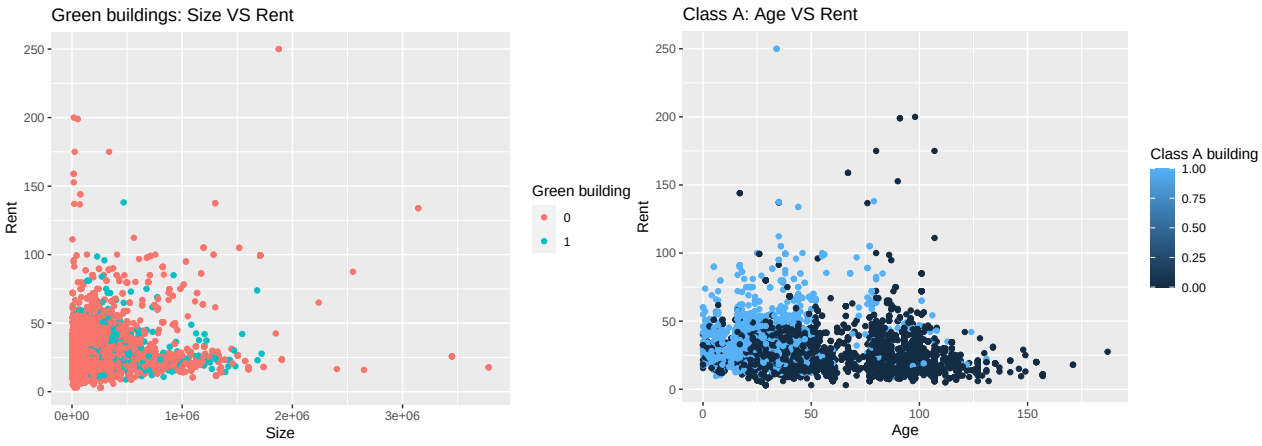
Objective : Evaluating the economic impact of going green for a new real estate project

[1] "We seggregated rent for green vs. regular building to review preliminary rent patterns" Min. 1st Qu. Median Mean 3rd Qu. Max. 8.87 21.50 27.60 30.02 35.50 138.07 Min. 1st Qu. Median Mean 3rd Qu. Max. 2.98 19.18 25.00 28.27 34.00 250.00

Step 1 :

We will Visualize the data to identify confounding variables affecting rent and then arrive at the numbers that can be used for calculations





Few preliminary findings from the correlation plots -

- Rent is correlated with the cluster rent
- Rent is correlated with the size, as expected
- Most of the class A buildings are also younger and have higher rent as they are premium buildings
- Age does not have a high correlation with rent

To further explore the rent dynamics, lets run a linear regression

Call: lm(formula = Rent ~ . - (cluster_rent), data = greenbuildings)

Residuals: Min 1Q Median 3Q Max -39.495 -6.822 -1.390 4.758 186.444

Coefficients: (1 not defined because of singularities) Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.306e+01 1.303e+00 -10.020 < 2e-16 ***CS_PropertyID -1.261e-06 2.000e-07 -6.305 3.03e-10*** cluster 2.761e-03 3.624e-04 7.617 2.90e-14 ***size 5.631e-06 8.413e-07 6.693 2.34e-11***
empl_gr 4.158e-01 2.086e-02 19.934 < 2e-16 ***leasing_rate 6.158e-02 6.774e-03 9.091 < 2e-16***
stories -7.686e-02 2.073e-02 -3.708 0.00021 ***age -1.162e-02 6.050e-03 -1.920 0.05490 .***
renovated -2.202e+00 3.296e-01 -6.682 2.52e-11 class_a 5.487e+00 5.594e-01 9.809 < 2e-16
class_b 2.114e+00 4.393e-01 4.811 1.53e-06 LEED 5.083e+00 4.594e+00 1.106 0.26859
Energystar 3.215e+00 4.896e+00 0.657 0.51142
green_rating1 -4.310e+00 4.923e+00 -0.876 0.38130
net -5.590e+00 7.586e-01 -7.370 1.88e-13 ***amenities 4.819e-01 3.231e-01 1.492 0.13586***
cd_total_07 -1.896e-03 1.851e-04 -10.242 < 2e-16 hd_total07 9.895e-04 1.148e-04 8.619 < 2e-16
total_dd_07 NA NA NA NA
Precipitation 5.437e-01 1.863e-02 29.183 < 2e-16 Gas_Costs -1.776e+03 9.726e+01 -18.263 < 2e-16 ***Electricity_Costs 1.110e+03 2.730e+01 40.669 < 2e-16*** — Signif. codes: 0 '*** 0.001 **' *0.01* ' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.07 on 7799 degrees of freedom (74 observations deleted due to missingness) Multiple R-squared: 0.3625, Adjusted R-squared: 0.3609 F-statistic: 221.8 on 20 and 7799 DF, p-value: < 2.2e-16

Insights :

From the linear regression, it can be concluded that rent is dependent on a lot of variables and we need to consider them before calculating the median rent for green buildings -
1. Rent is dependent on the size 2. Rent depends on the class of the building (a vs b) 3. Rent depends on a lot of other factors which we do not know about this new project, such as gas cost, electricity cost

Most importantly, rent is highly dependent on the location. Post performing some desk research, we find that East and Downtown seem to have the highest rent per square feet in Austin.

Links for prices based on location - https://www.commercialcafe.com/office-market-trends/us/tx/austin/ ; https://www.propertyshark.com/cre/office/us/tx/austin/east-cesar-chavez/ ; https://aquilacommercial.com/learning-center/cost-to-lease-office-space-austin-tx/ ;

Additionally, since our building is supposed to be 250,000 Square feet, we should also filter for buildings with similar size - in the range of 200,000 - 300,000 Square Feet

(we will also exclude buildings with occupancy less than 10% to avoid distortion in our analysis)

Insights :

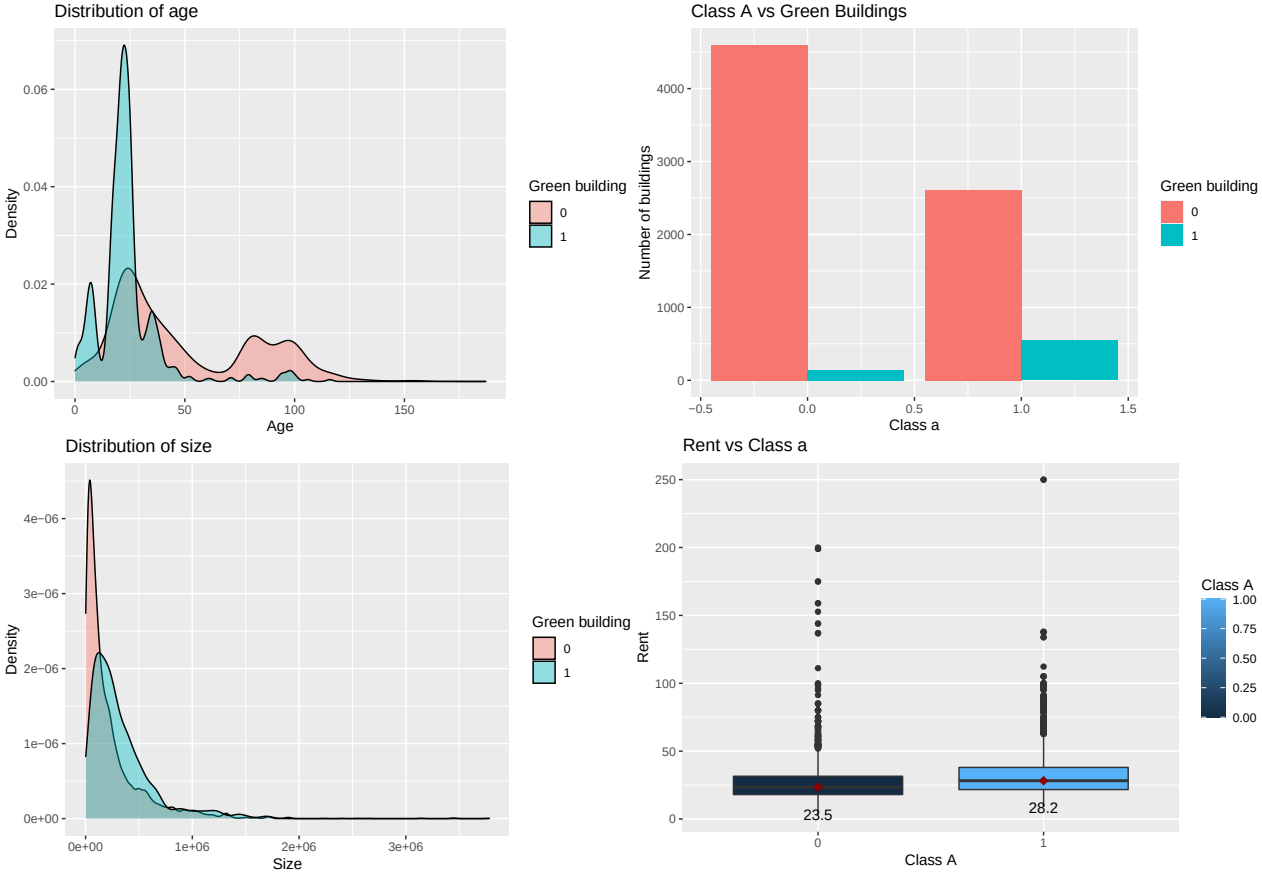
Rent per Square feet for regular building seems to be more than Green building, contrary to what excel guru calculated. Hence, excel guru had the median rates wrong since he directly calculated the median from the entire data.

A tibble: 2 x 2

green_bld_identifier n 1 Green building 108 2 Regular Building 713

Caveat : After filtering the data based on the above specificaitons, we are left with very few data points (10% of the total points) which might not yield correct results. Ideally, we would require a lot more data points to go ahead with this approach and correctly estimate median rents for green vs. regular buildings.

Hence, we will proceed ahead with the entire data set for further analysis.

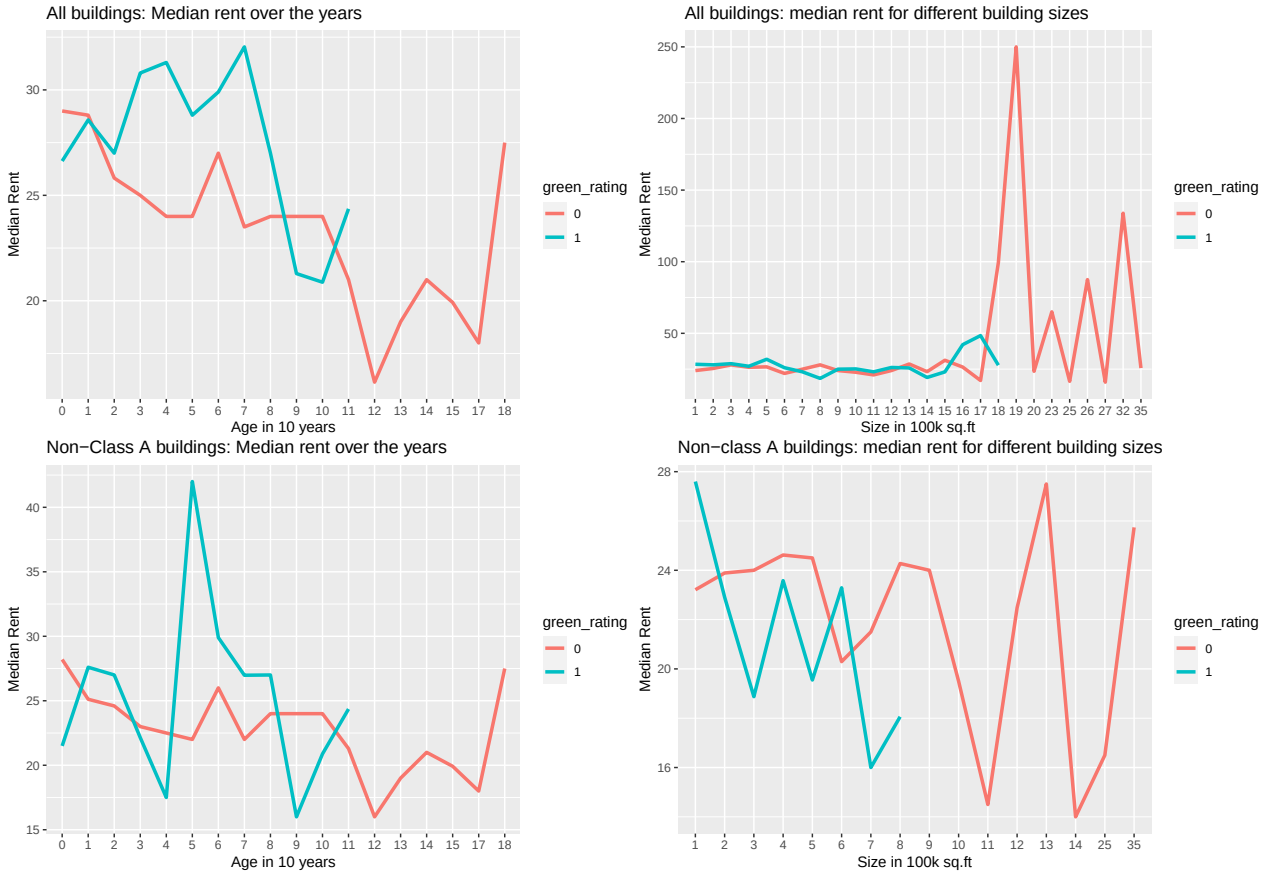


Insights :

- Most of the green buildings are younger than non-green buildings

- The proportion of class a buildings is higher in green buildings
- The proportion of green and non-green building increases as the size of buildings increases
- The is a significant difference in the of rent of class a and non-class a buildings

Next, we identified patterns between median rent for all buildings with age and size, and class a and non class a median rent with age and size.



Insights :

- For all buildings, the median rent over the years is higher than non green buildings for age between 2 to 9 years
- The rent of green buildings is lower than non-green ones when they are not class a buildings
- The rent difference is not uniform across different sizes and ages

Next, we calculated median leasing rate for class a buildings of sizes ranging from 200k to 300k sq.ft, difference in rent for the first 5 years for class a and non class a buildings.

[1] "Median leasing rate for class a buildings of sizes ranging from 200k to 300k sq.ft 91.605" [1] "Difference in rent for the first 5 years class a buildings: 3.097" [1] "Difference in rent for the first 5 years for non-class a buildings: -1.167"

Insights :

We have seen that the analysis by stats guru is flawed since he fails to account for all the factors that affect the rent. He used the median rent of all buildings to calculate the returns where he fails to account for other factors such as size and class of the buildings into his analysis. For instance, we have seen that class a green buildings have a higher rent than non green class a buildings.

The rent difference is not uniform across different sizes and age, so we cannot use a fixed difference in rent to calculate the returns. We should also use the median leasing rate of such buildings instead of 90% rate.

- The builder should invest in a Class-A green building to yield positive returns
- We can expect a occupancy rate of 91.6% on such buildings
- The average difference in rent for green and non-green buildings that are class a and whose sizes ranging from 200k to 300k is 3.097

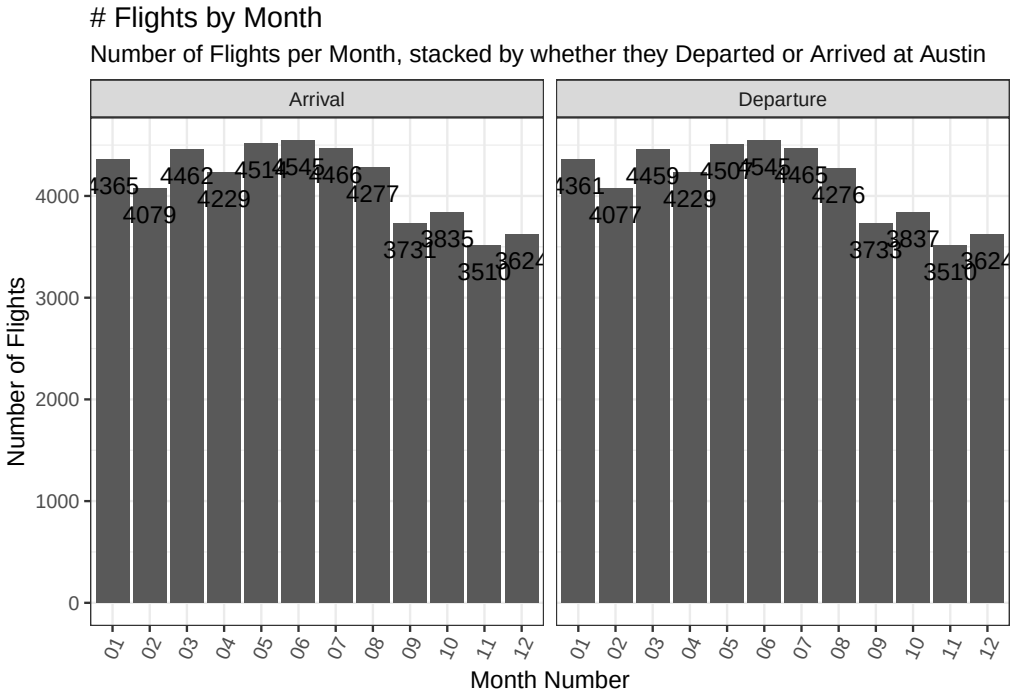
From the above analysis we conclude that the benefits of going for green buildings do not outweigh non-green buildings. Moreover, the constricted dataset on the basis of regression results also imply we should not go ahead with green-buildings. Additionally, due to absence of location data, we are not sure of the range of prices we will be able to charge. Overall, we would recommend to not go ahead with developing the green building due to uncertain factors involved.

End of Problem

Visual story telling part 2: flights at ABIA

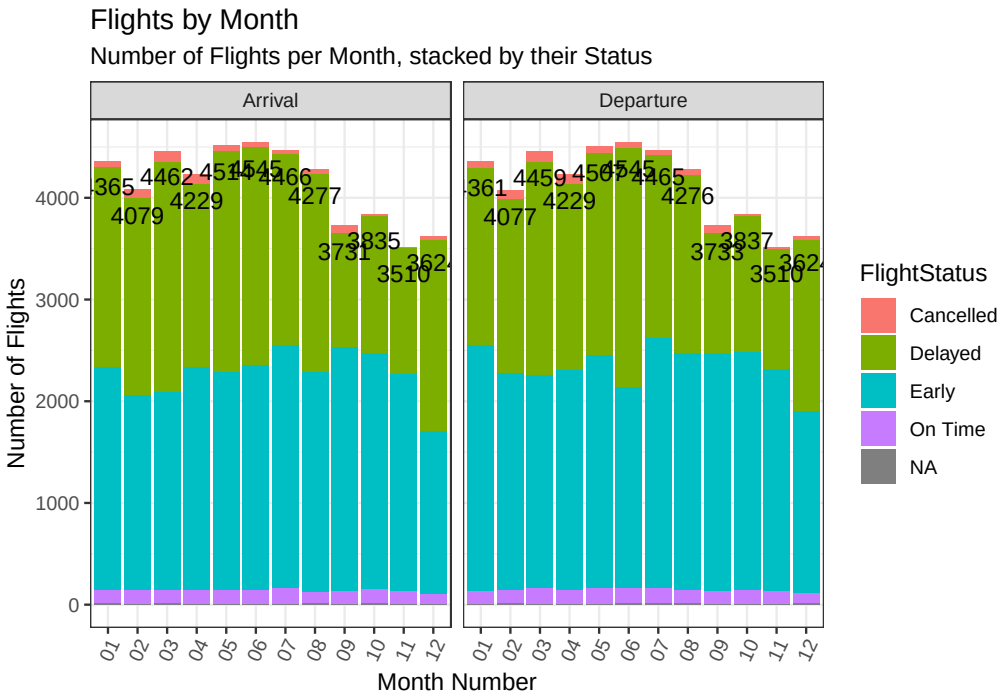
Objective : Using the data, tell an interesting story about flights into and out of Austin

To start analyzing the data, lets get an idea about the monthly number of flight incoming/outbound from Austin

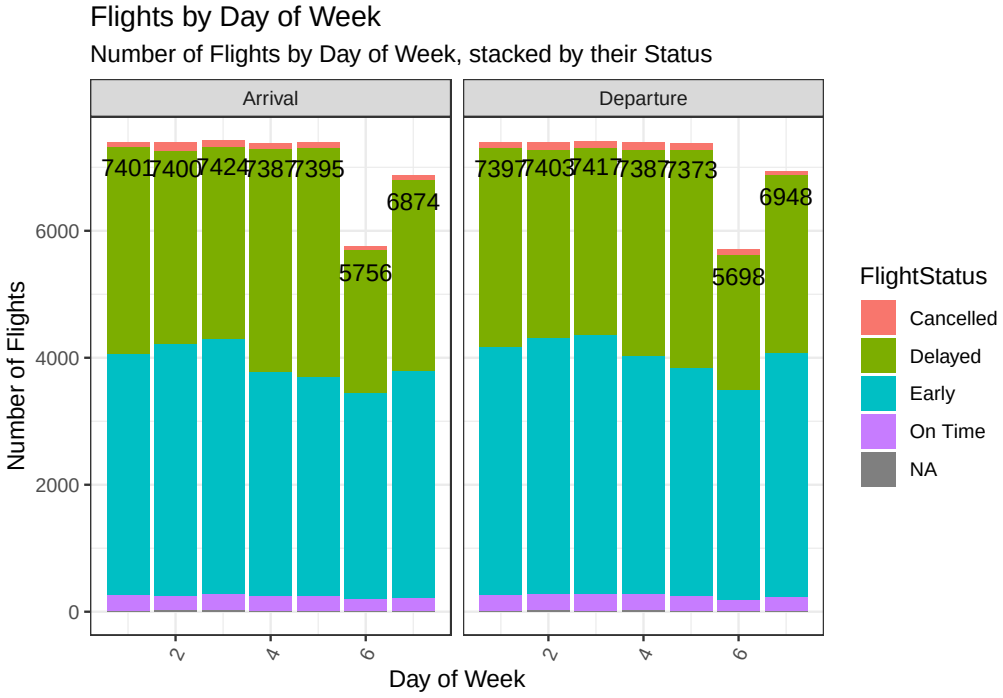


We see that the first 8 months of the year had the most flights operating to and from Austin. The last 4 months had lesser flights departing and arriving in Austin. **Interestingly, number of flights arriving and departing from Austin are pretty similar every month.**

Lets further break this chart down by the flight status (delayed/early/cancelled...)

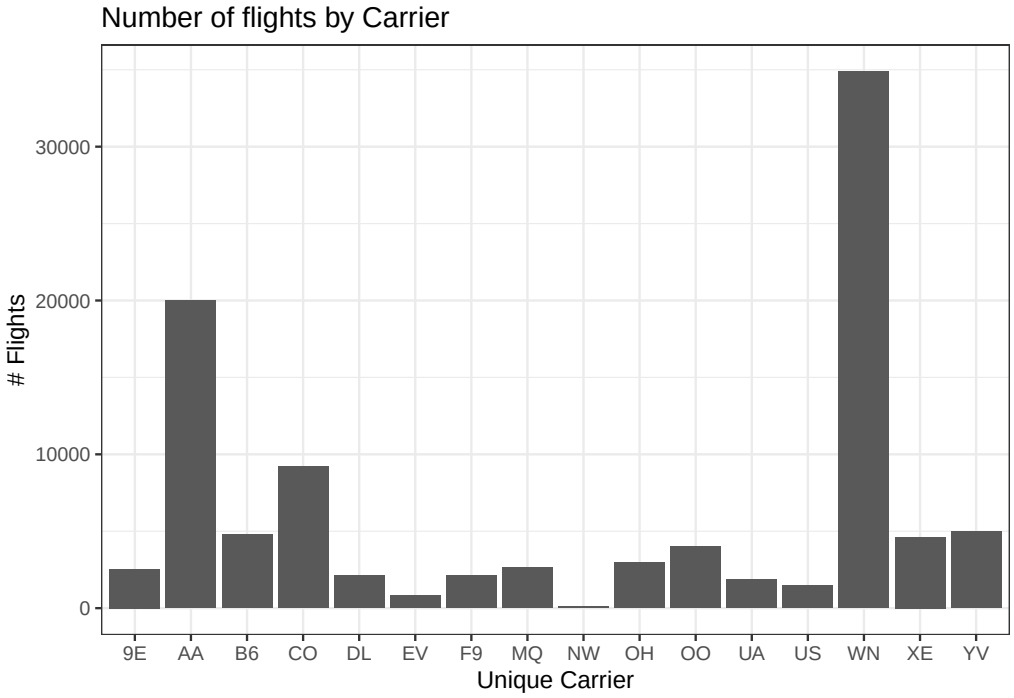


We see a very similar trend of flight status among flights arriving and departing from Austin. Hence, nothing much can be inferred.

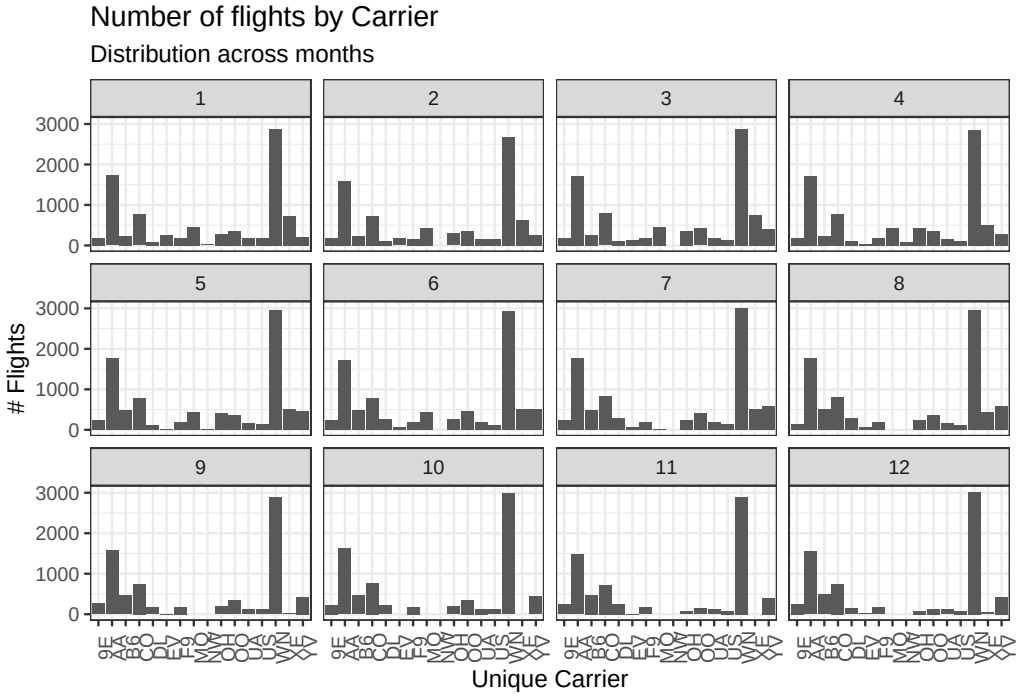


Ratio of cancelled/delayed/early... flights seem to be similar from what we observed in monthly trends. But we do see significantly low volume of flights on Saturday.

Next, lets analyze number of operations for each flight carrier in Austin.



Southwest (WN) seems to have the most number of flights operating at Austin airport. Lets see if this trend changes month over month.

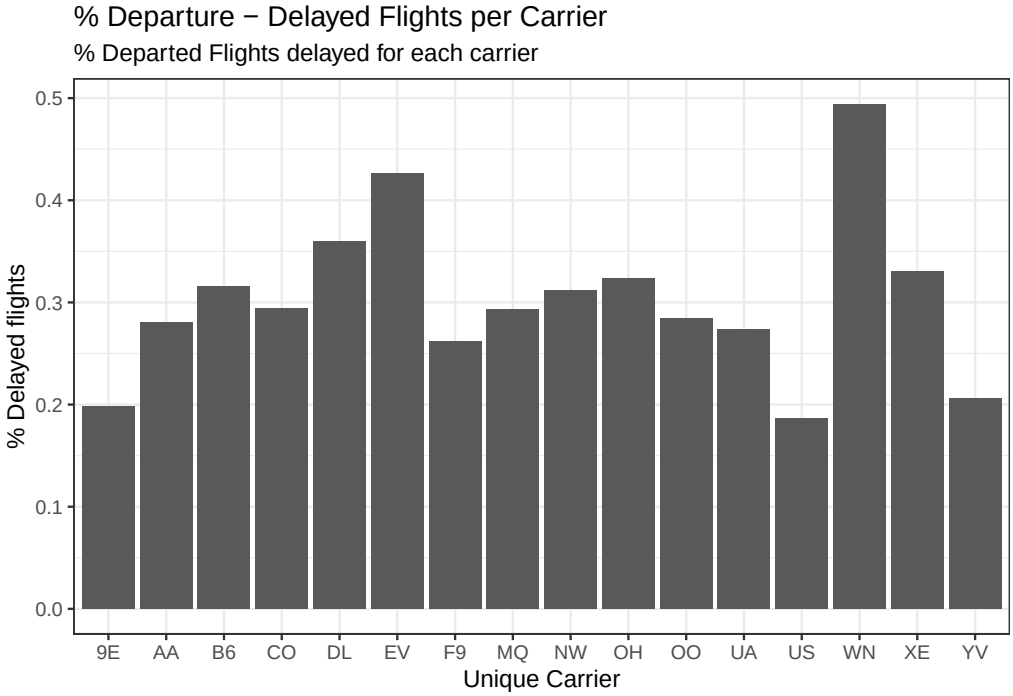


We can see that the overall trend does not change month over month. However, we have one peculiar observation. **MQ flight carrier has reduced operations in July and no operations after that. Also, NW carrier (which has least operations overall) operated in just January, April and May.**

Now, lets look at the flight delay dynamics to identify reliable vs. non reliable flights to travel!

We will filter for flights originating from Austin, and calculate % flights getting delayed for each carrier.

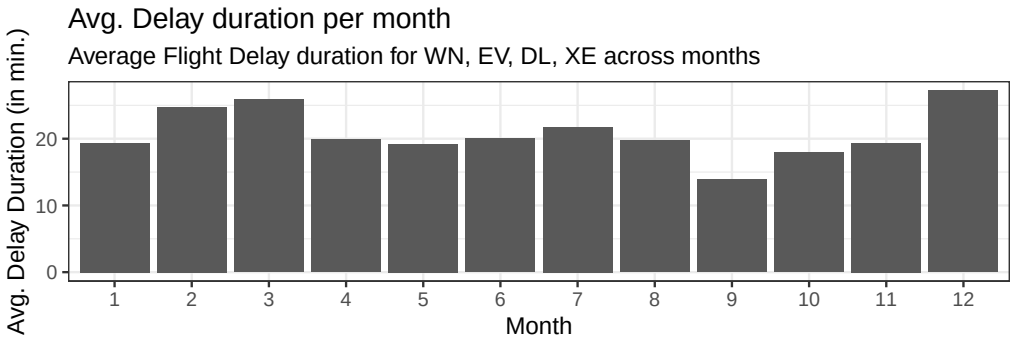
Also, we will exclude the negative delay values which probably indicates flight departed before time

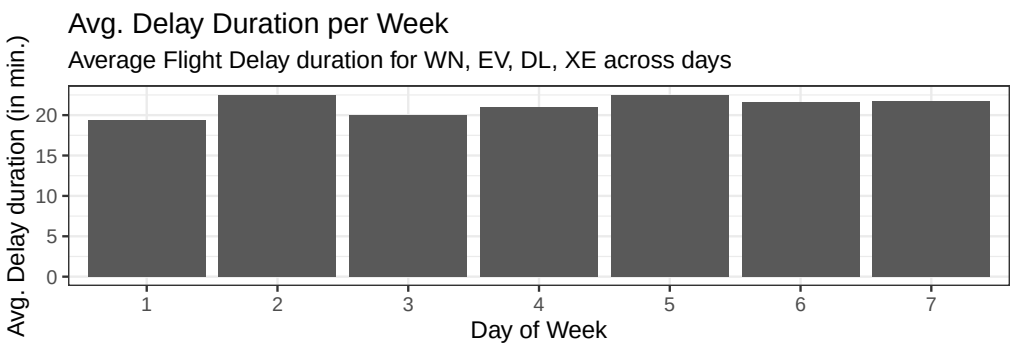


WN, which had maximum number of flights operating at Austin airport, also had the maximum % of delayed departed flights, followed by EV, DL and XE. US and 9E had the least % of delayed departed flights making them pretty reliable.

We are not looking at the arrival delays, since at the end of the day, a passenger only cares if his/her flight departed on time or how late did the flight depart

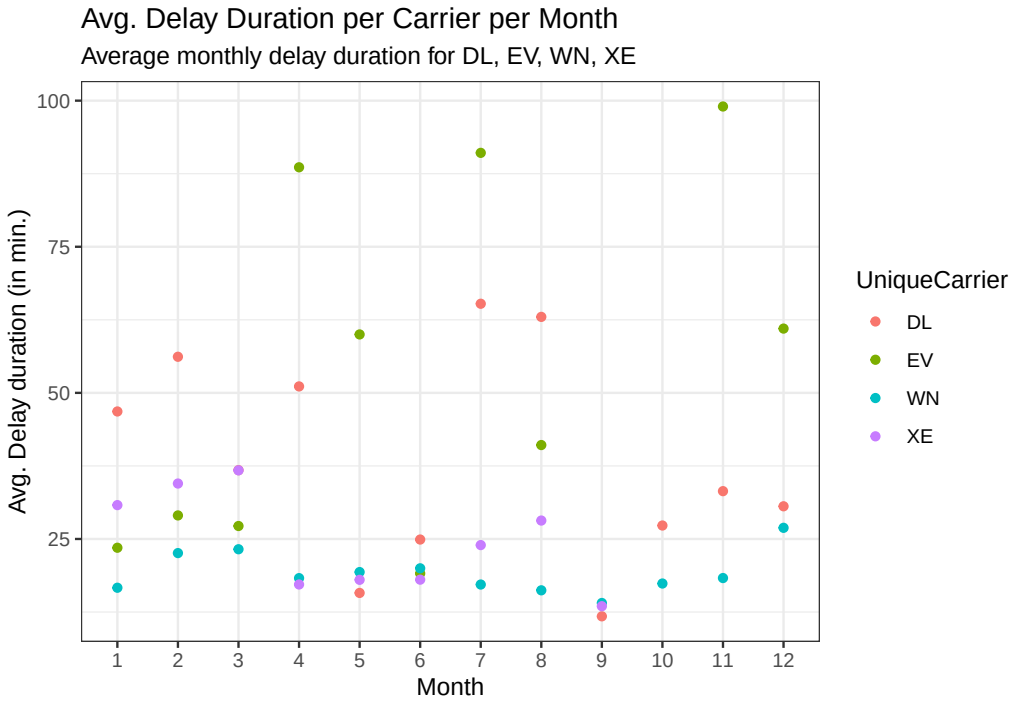
Lets see if we can find any pattern in the flight delays (for carriers with highest proportion of delayed flights - WN, EV, DL, XE) based on month, day, hour, etc. across different carriers.





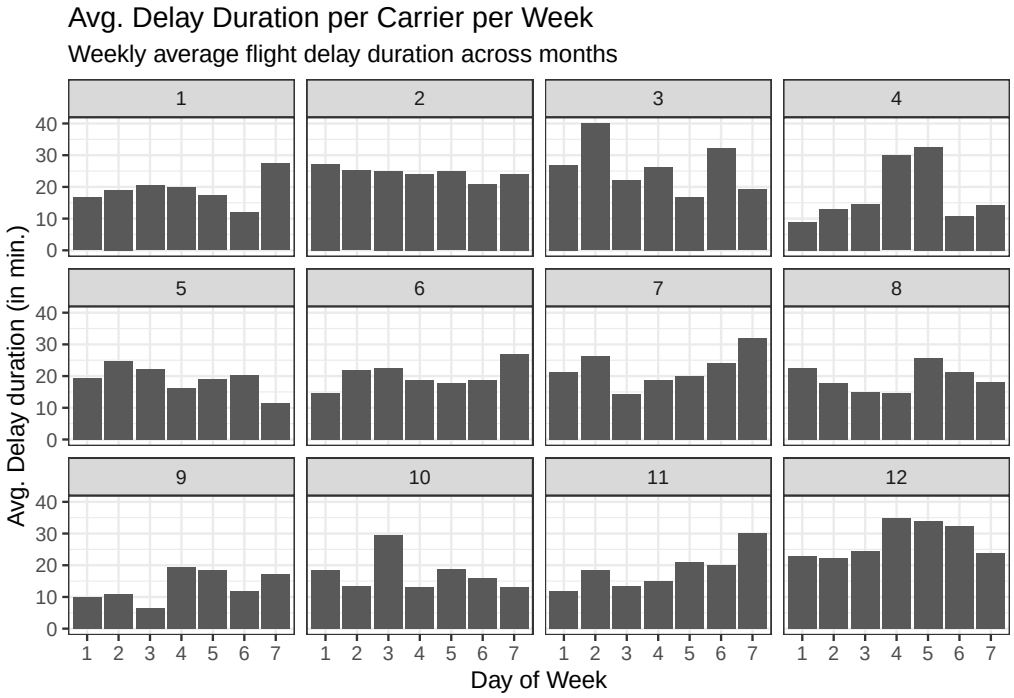
Average delay duration seem to be 25 min or more in February, March and December 2008. This is potentially due to lot of people travelling around Spring break and Christmas holidays which causes air traffic and delays. However, average delay duration across the week seems to be pretty consistent without much variation, hence it does not provide us much insight.

Lets see if the delay duration trend vary across each carrier.



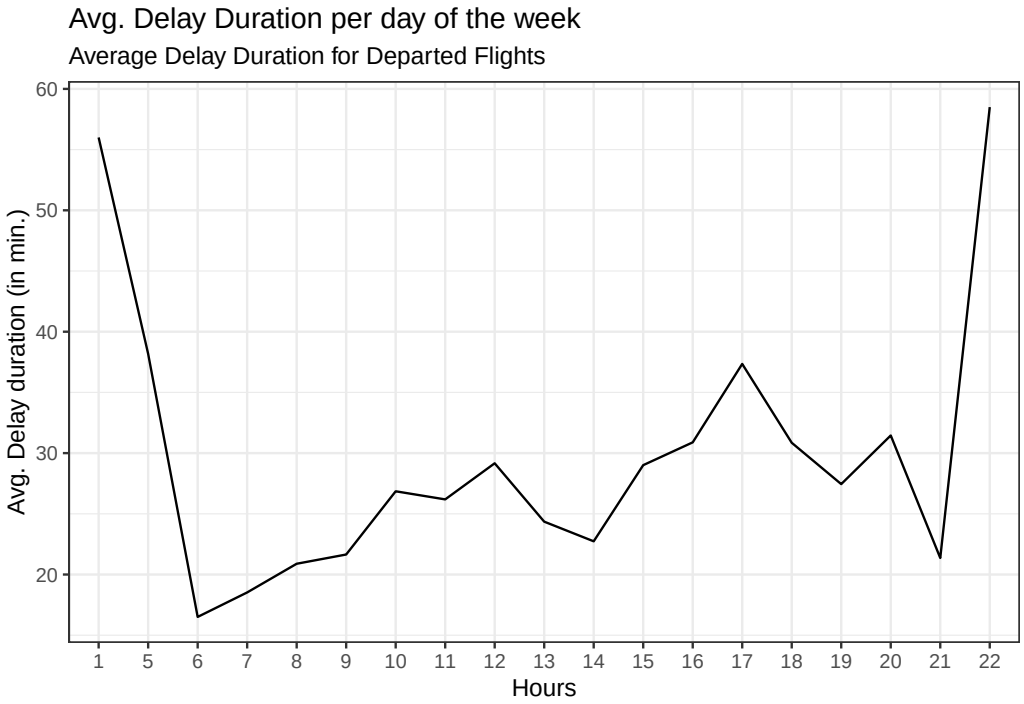
This gives us pretty interesting insights. EV seems to be pretty unreliable carrier due to their frequent large duration delays ($>$ 1 hour). WN which had maximum operations, had pretty consistent delay of $>$30 mins, throughout the year.

Lets us delve a little further into weekly delays, since we could not extract much insight earlier (due to consistent delay duration throughout the week). We will now look at the weekly delay duration for carriers with maximum proportion of their flights delayed (DL, EV, WN, XE) across months.



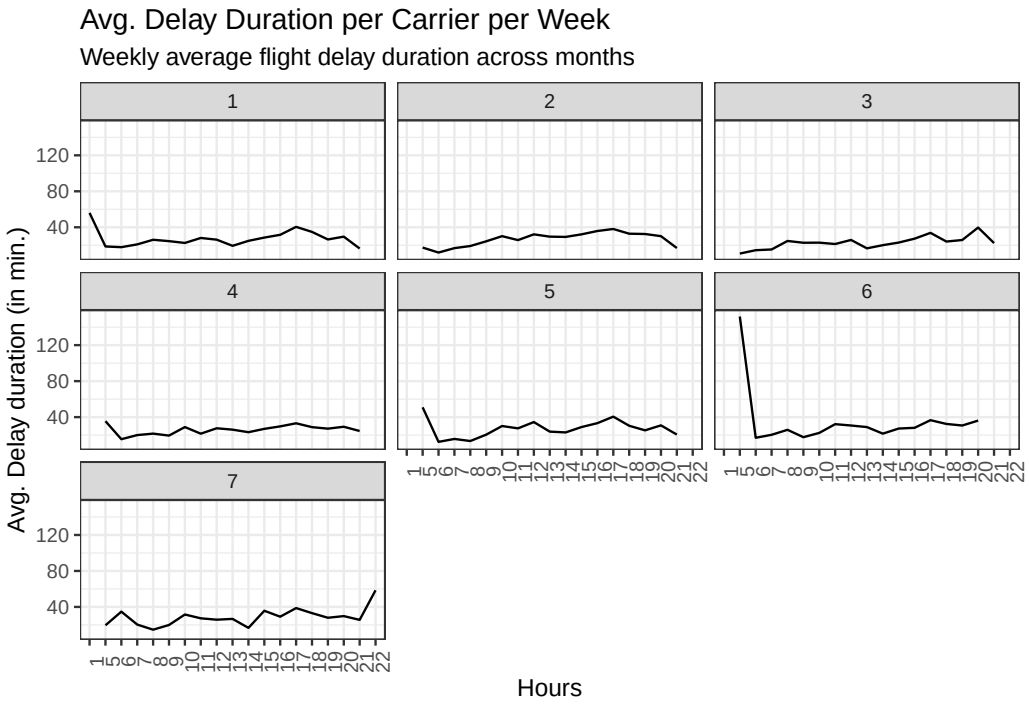
On a quick glance, March, April & December seem to have higher delays on Friday when compared to other months. March has higher delay for Tuesday as well.

Lastly, lets look at the hourly delay duration at the airport. This will help us determine the best time to travel via Austin airport.



Delay seems to be the maximum at 1AM and 10 PM, and the least at 6 AM. Hence, early morning, mid afternoon (around 2) or late nights are good time to travel.

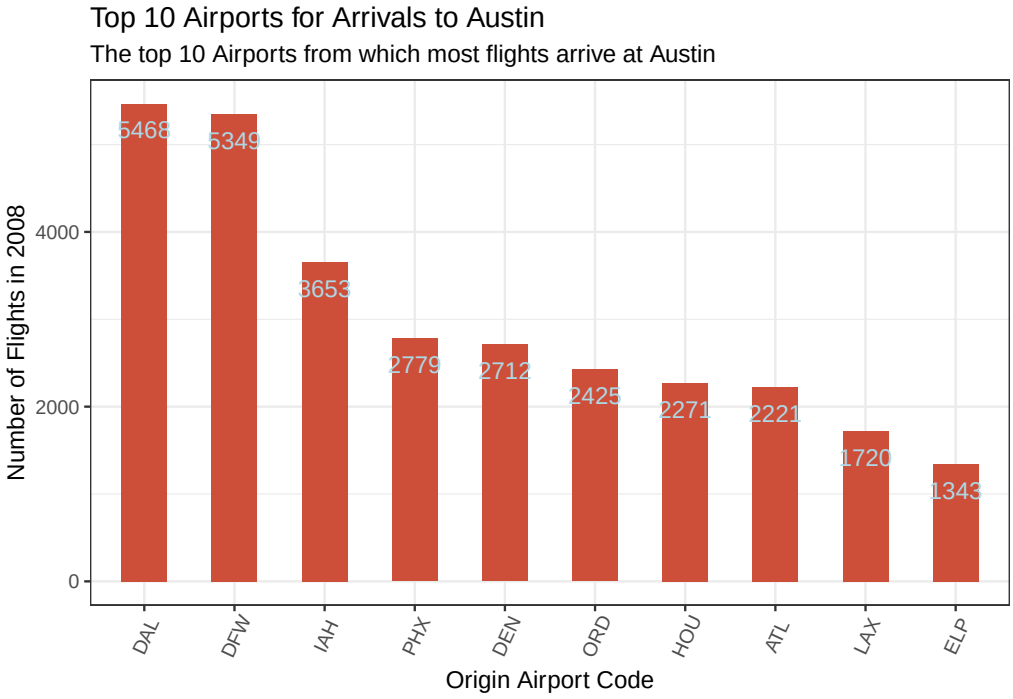
Lets look if this trend holds true for all days of the week or not.

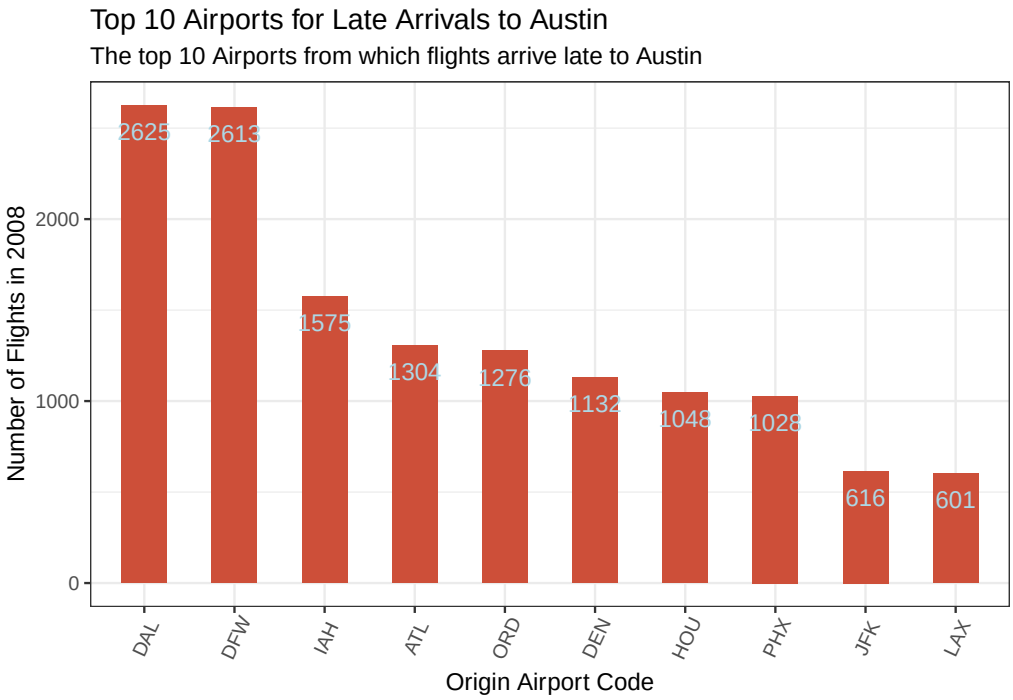


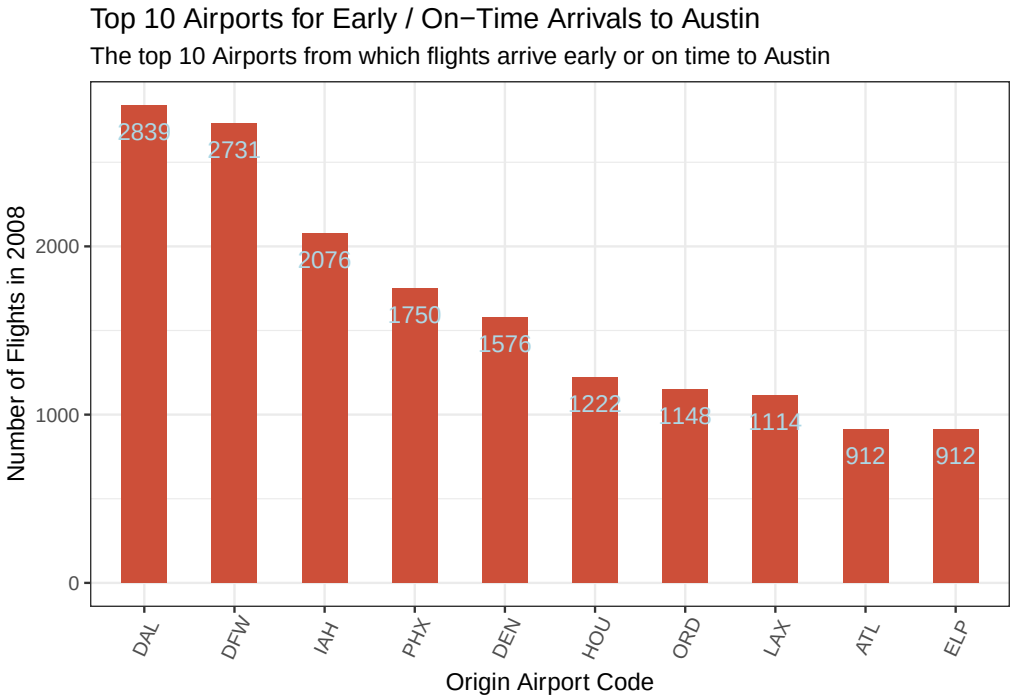
We can see that across all days of the week, early morning, mid afternoons and late nights are good time to travel.

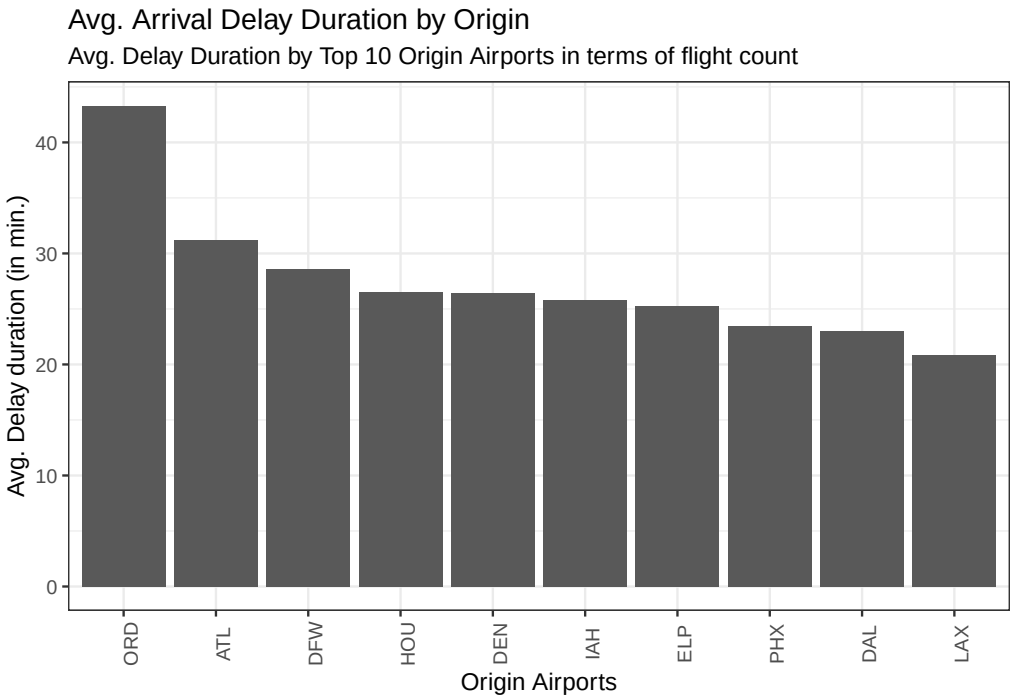
Now that we have analyzed Delays, lets see some summaries around arrivals and departures from Austin.

Flights arriving in Austin



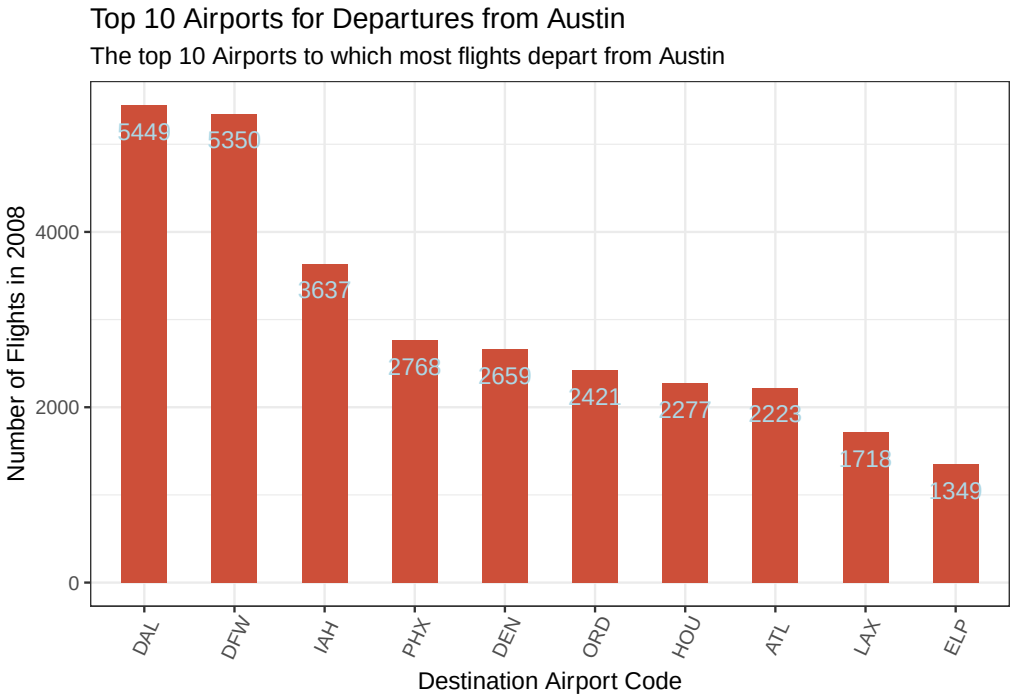

Top 10 Airports for Late Arrivals to Austin
The top 10 Airports from which flights arrive late to Austin
Number of Flights in 2008
0
1000
2000
2625
2613
1575
1304
1276
1132
1048
1028
616
601
DAL
DFW
IAH
ATL
ORD
DEN
HOU
PHX
JFK
LAX
Origin Airport Code

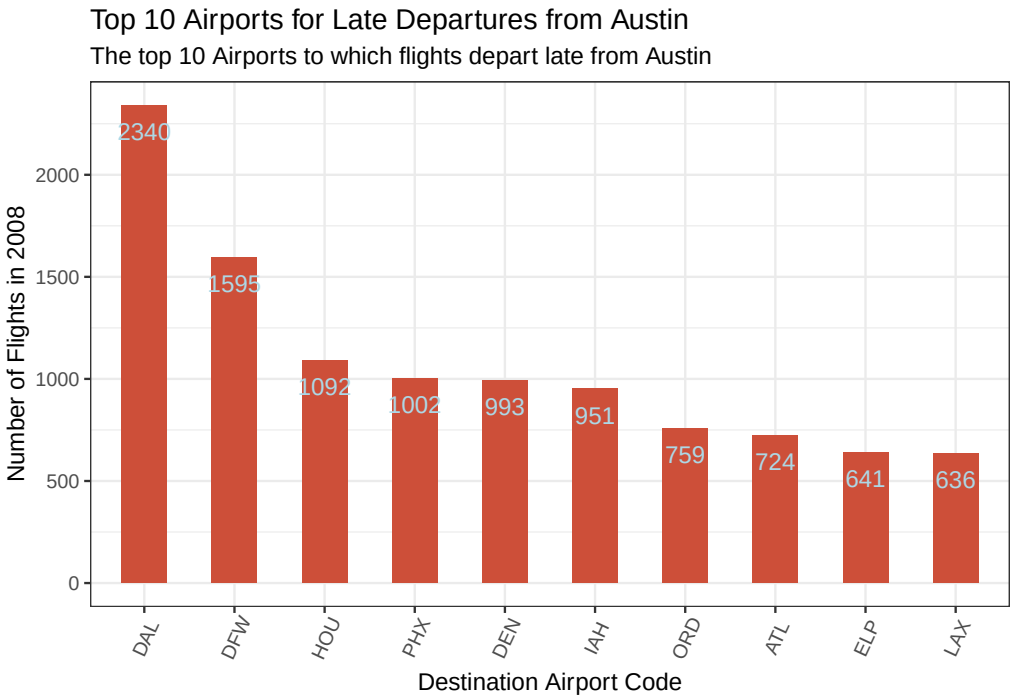

Top 10 Airports for Early / On−Time Arrivals to Austin
The top 10 Airports from which flights arrive early or on time to Austin
Number of Flights in 2008
0
1000
2000
2839
2731
2076
1750
1576
1222
1148
1114
912
912
DAL
DFW
IAH
PHX
DEN
HOU
ORD
LAX
ATL
ELP
Origin Airport Code

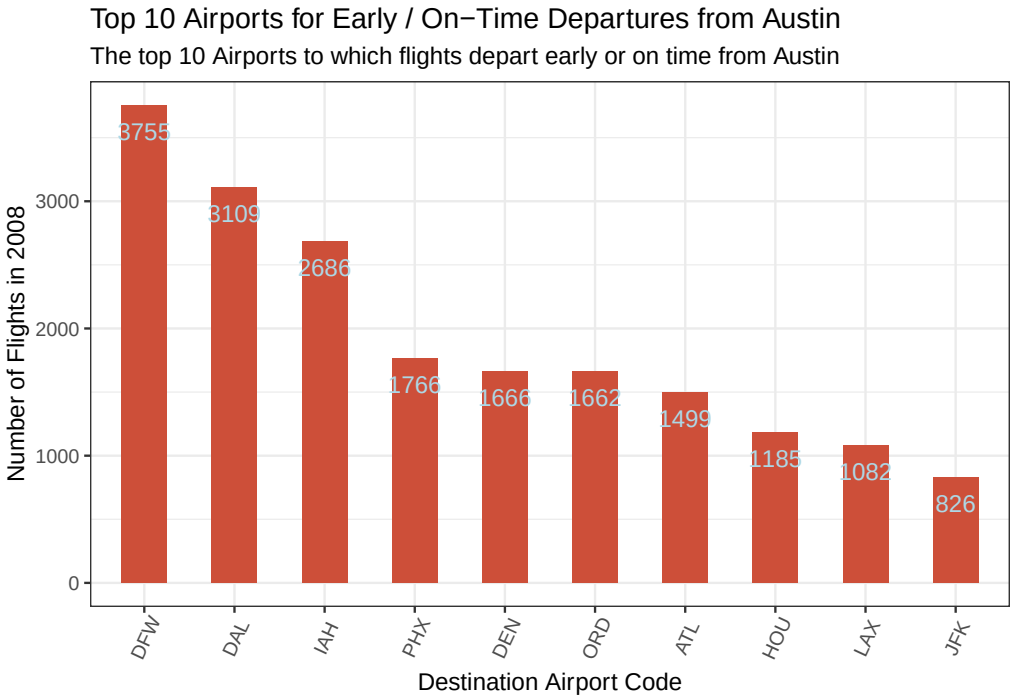


Maximum number of flights arrive from Dallas, DFW and Houston, i.e. flight seem to arrive from mostly Texas itself. Additionally for each airport, number of early/on-time arrivals are more than delayed arrivals.

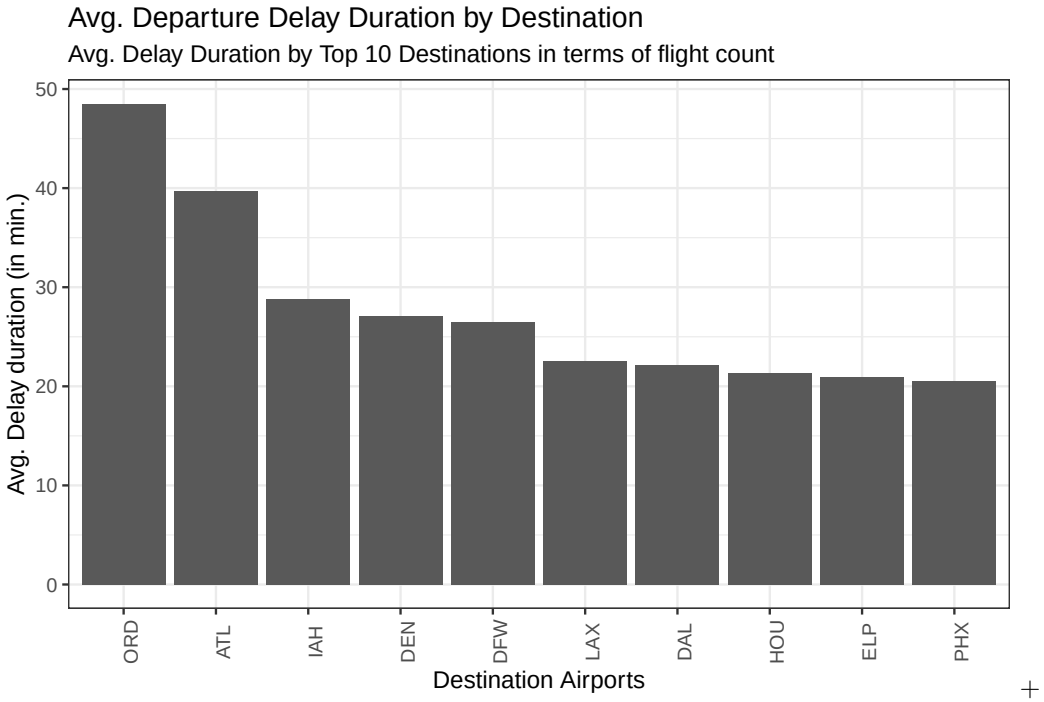
Flights departing from Austin







Maximum number of flights departing from Austin, again arrive within Texas (among DFW, Dallas and Houston). Again, number of early/on-time departures from Austin are greater than delayed departures for each destination. Lets see which average delay flight faces to depart for a particular destination.



The maximum delay in departed flights was observed for from Austin to ORD, JFK and ATL with avg. delay duration of ~40-50 minutes. There seems to be lesser delay in flights flying within Texas.

Final Summary :

- Almost equal number of flights arrived and departed from Austin every month in 2008.
- A little less than half of the arrived/departed flights got delayed
- WN (35K), followed by AA(20K) and CO (10K) carriers had maximum number of flights operating at Austin airport (MQ and NW carriers only operated in few months of the year)
- WN had the maximum proportion of delayed flights followed by EV, DL and XE - hence making them unreliable. On the other hand, US and 9E had least proportion of delayed flights, hence are reliable.
- During spring/Christmas breaks, average flight delay duration seem to be high which is understandable given the vacation rush.
- Among our unreliable carriers (WN, EV, DL, XE), delay duration during the spring/Christmas breaks was particularly high for EV (> 1 hour)
- Based on delay duration on a hourly basis, early morning, mid afternoons and late nights are good time to travel.
- Maximum number of flights departing from Austin, again arrive within Texas (among DFW, Dallas and Houston).
- The maximum delay in departed flights was observed for from Austin to ORD, JFK and ATL with avg. delay duration of ~40-50 minutes. There seems to be lesser delay in flights flying to destinations within Texas.

End of Problem

Portfolio modeling

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following object is masked from 'package:cowplot':
##
##     theme_map

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:purrr':
##
##     cross

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum

## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

## Loading required package: TTR

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

##
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

We choose 3 different ETF-based portfolios and estimate the 5% VaR (Value-at-Risk) for all 3 portfolios. We will create 3 different portfolios with different risk levels - diversified, high-risk high-reward, and low-risk low-reward. This will provide us with proper insights into the differences among various asset allocations in a portfolio.

To proceed, we choose the same ETFs for all 3 portfolios. However, we differ on the weights assigned to each ETF based on the risk tolerance for each portfolio. The ETFs we choose are as follows.

- Risky ETFs
  - Large Cap Growth Equities
    - SPDR S&P 500 ETF Trust (SPY)
    - Vanguard S&P 500 ETF (VOO)
  - Financials Equities
    - Invesco KBW High Dividend Yield Financial ETF (KBWD)
- Safe ETFs
  - Government Bonds
    - iShares U.S. Treasury Bond ETF (GOVT)
  - Oil and Gas ETFs
    - United States Brent Oil Fund LP (BNO)
  - Real Estate ETFs
    - SPDR Dow Jones REIT ETF (RWR)

We fetched data for the above selected ETFs for previous 5-years. Next, we created a return matrix containing returns corresponding to the five ETFs. We also generated a summary table containing min, max and mean returns.

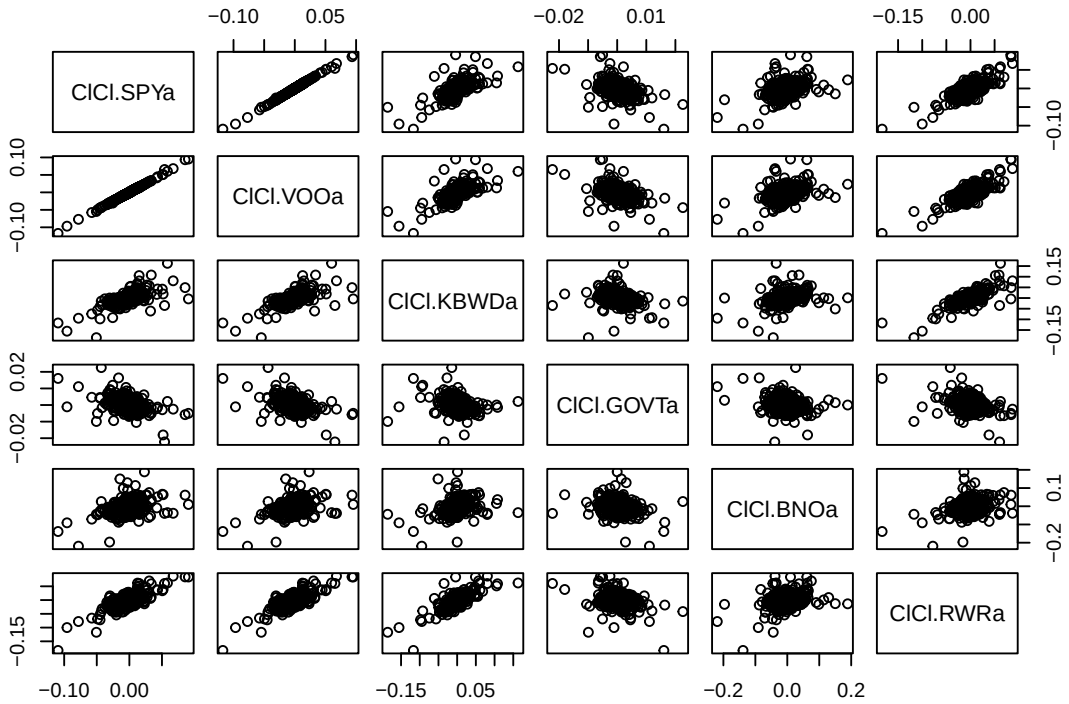
```
## 'getSymbols' currently uses auto.assign=TRUE by default, but will
## use auto.assign=FALSE in 0.5-0. You will still be able to use
## 'loadSymbols' to automatically load data. getOption("getSymbols.env")
## and getOption("getSymbols.auto.assign") will still be checked for
## alternate defaults.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

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

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/SPY?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/SPY?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/VOO?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/VOO?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/KBWD?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/KBWD?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/GOVT?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/GOVT?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'
```

```
## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/BNO?
## period1=-2208988800&period2=1629072000&interval=1d&events=div&crumb=SiIP9kwlGv2'


## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/BNO?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/BNO?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query1.finance.yahoo.com/v7/finance/download/RWR?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'

## Warning in read.table(file = file, header = header, sep = sep,
## quote = quote, : incomplete final line found by readTableHeader
## on 'https://query2.finance.yahoo.com/v7/finance/download/RWR?
## period1=-2208988800&period2=1629072000&interval=1d&events=split&crumb=SiIP9kwlGv2'
```

Adjusted Returns - Top Rows: ClCl.SPYa ClCl.VOOa ClCl.KBWDa ClCl.GOVTa 2016-08-18 2.243925e-03 2.395982e-03 0.0068618021 0.0011468272 2016-08-19 -1.462158e-03 -1.394353e-03 0.0004543389 -0.0026727758 2016-08-22 -4.573076e-05 -9.967588e-05 0.0009083106 0.0022970520 2016-08-23 2.013463e-03 1.845178e-03 0.0063520415 0.0000000000 2016-08-24 -5.114833e-03 -4.828518e-03 -0.0022543282 -0.0007639419 2016-08-25 -6.885885e-04 -8.003351e-04 -0.0013555807 -0.0011468272 ClCl.BNOa ClCl.RWRa 2016-08-18 0.0197010870 -0.0007958118 2016-08-19 -0.0006662225 -0.0081640679 2016-08-22 -0.0293333333 0.0046175365 2016-08-23 0.0116758242 0.0006994404 2016-08-24 -0.0183299389 -0.0047928506 2016-08-25 0.0138312586 0.0056185915

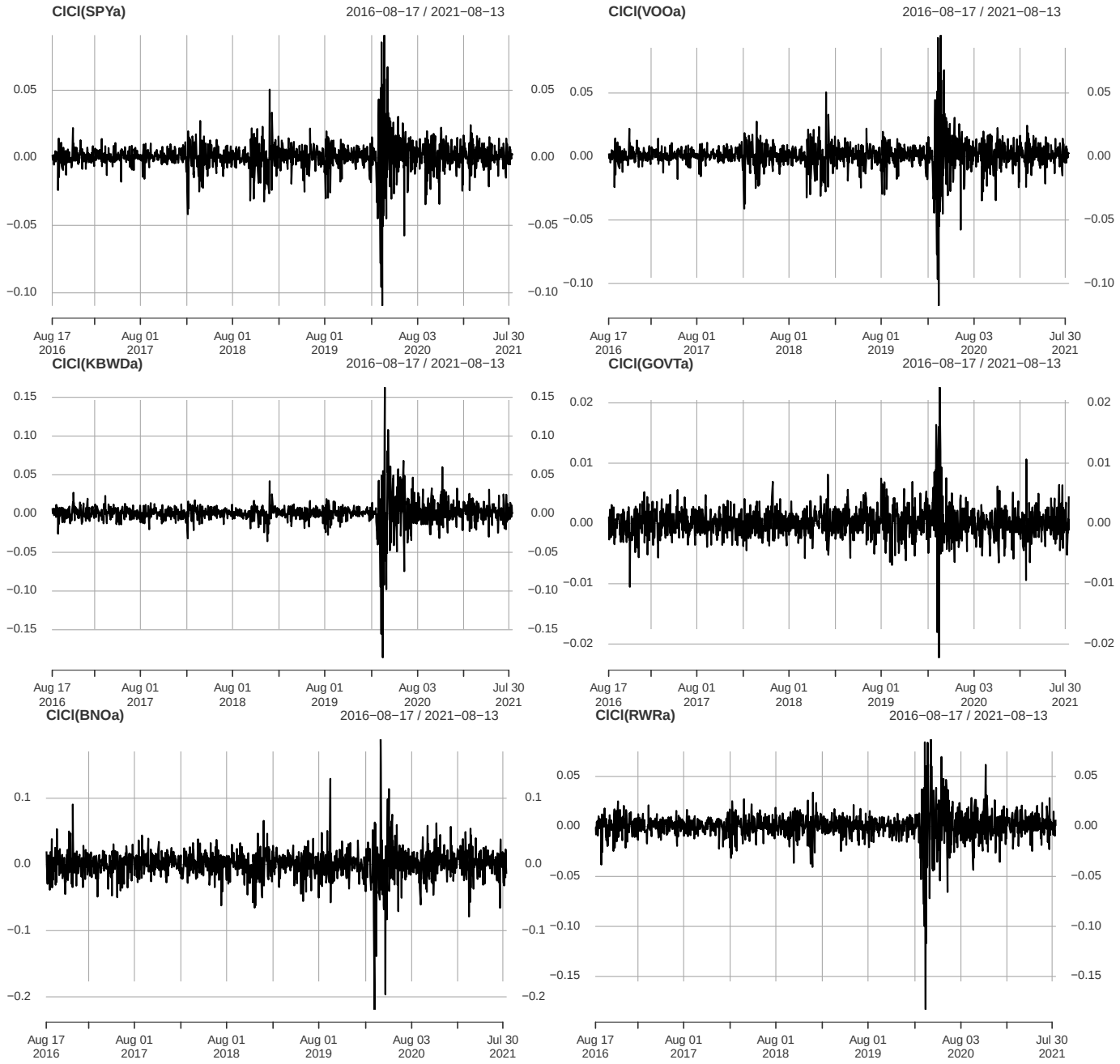
Summary Table ClCl.SPYa ClCl.VOOa ClCl.KBWDa
Min. :-0.1094237 Min. :-0.1173881 Min. :-0.1858134
1st Qu.:-0.0028198 1st Qu.:-0.0028318 1st Qu.:-0.0043777
Median : 0.0007673 Median : 0.0008654 Median : 0.0008286
Mean : 0.0007124 Mean : 0.0007179 Mean : 0.0004489
3rd Qu.: 0.0055504 3rd Qu.: 0.0056239 3rd Qu.: 0.0053437
Max. : 0.0906033 Max. : 0.0953640 Max. : 0.1628122
ClCl.GOVTa ClCl.BNOa ClCl.RWRa
Min. :-2.225e-02 Min. :-0.2191316 Min. :-0.1831070
1st Qu.:-1.475e-03 1st Qu.:-0.0098640 1st Qu.:-0.0049734
Median : 0.000e+00 Median : 0.0014444 Median : 0.0009768
Mean : 9.266e-05 Mean : 0.0004526 Mean : 0.0003324
3rd Qu.: 1.566e-03 3rd Qu.: 0.0118229 3rd Qu.: 0.0066903
Max. : 2.258e-02 Max. : 0.1889764 Max. : 0.0870632

In line with our portfolio selection criteria, we observe that SPY, VOO and KBWD have slightly higher mean returns as they are risky ETFs. GOVT, BNO and RWR have lower mean returns as they are less risky. As returns are proportional to risks, the mean returns for chosen ETFs follow the risk-return pattern.



We see from the above pairwise scatter plots that the Riskier ETFs are correlated with each other. This can be intuitively thought of as them increasing as the market index increases (SPY is a good indicator of the market in general) and vice versa. For the safer ETFs, we see that they don't change much and stay clustered towards the middle of the plots, showcasing lack of variability in the prices.

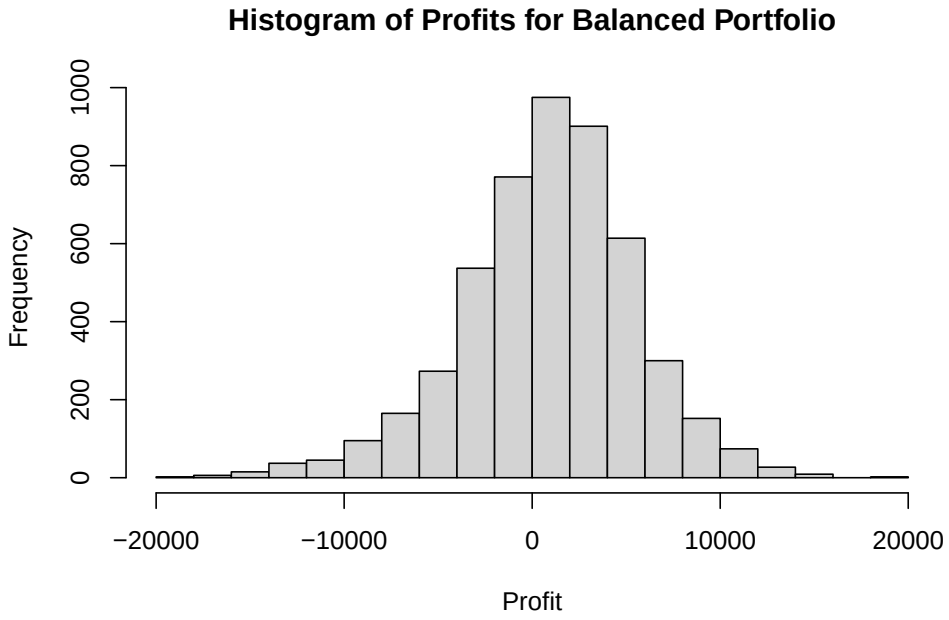
Volatility Plots



We checked the volatility of the ETFs across the 5 years and observed that year 2020 was the most volatile year for both safe and risky ETFs, which is in line with what we expected. For years prior to 2020, we see slightly more volatility in for risky ETFs compared to the safe ETFs.

Balanced Portfolio

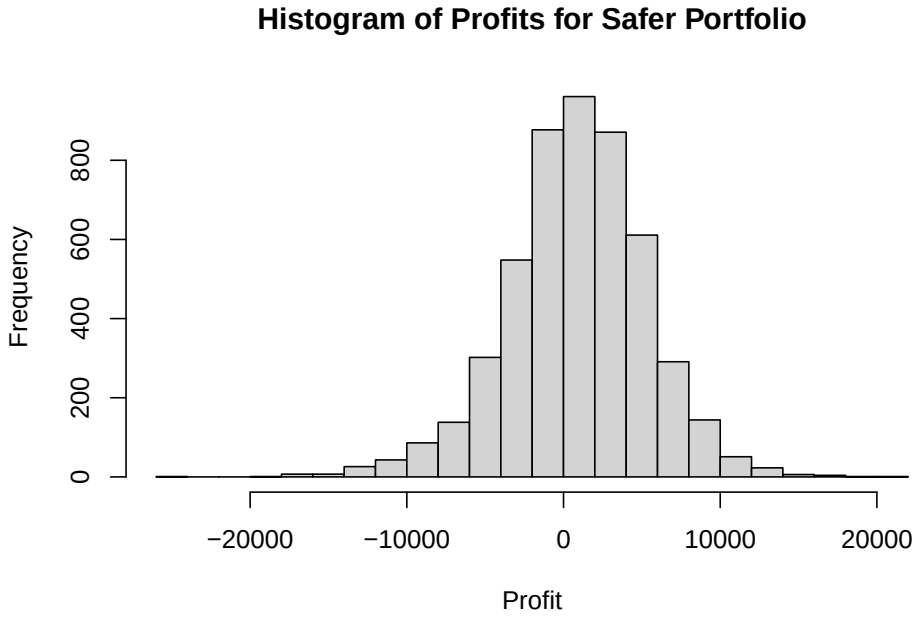
We begin with a Diversified Portfolio, providing near-equal weightage to all ETFs.



Profit Mean: 892.2765%
Value at Risk (VaR): -7318.086 We see from the above that the highest frequency (around 1000) is for marginal profits. The spread of the balanced portfolio is less, which indicates lesser variance and hence lesser risk.

Safe Portfolio

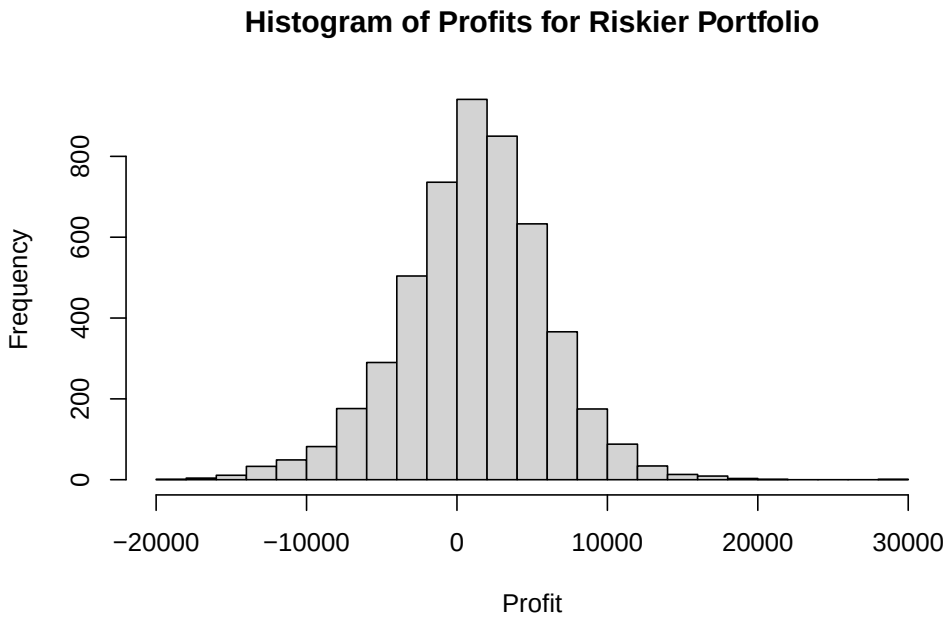
We now analyze a Safer Portfolio, providing higher weightage to the safer ETFs.



Profit Mean: 834.46195%
Value at Risk: -6780.393 The plot and results for the safer portfolio are similar to that of balanced (with a higher frequency at marginal profits and a lower spread). This is also intuitive since a diversified portfolio is also safe, so they are expected to give similar results.

Aggressive Portfolio

We finally analyze a Riskier Portfolio, providing higher weightage to the riskier ETFs.



Profit Mean: 1138.1725%

Value at Risk (VaR): -7045.593

The plot and the results for the riskier portfolio show a wider spread. This indicates a higher variance and as a result a higher risk compared to the other portfolios.

Summary of Portfolio Results

```
 Portfolio  Profit Mean        5% Value at Risk
```

[1,] “Balanced” “892.275999004083” “-7318.08601703537” [2,] “Safe” “834.461920991688” “-6780.39316040742” [3,] “Risky” “1138.17197265241” “-7045.5934723103”

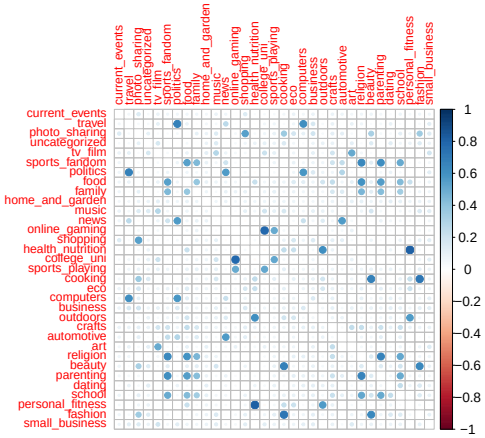
Thus, we see that the Riskier Portfolio gives us a high risk - high reward return, the safer gives us a low risk - low reward return, and the balanced portfolio gives us an investment somewhere in the middle. The portfolio we choose to go forward with depends on the investor and their preferences.

End of Problem

Market Segmentation

Objective : Analyze social_marketing.csv data, and prepare a concise report for NutrientH20 identifying any interesting market segments that appear to stand out in their social-media audience.

Step 1: Loading the libraries and Pre-Processing



The file social_marketing.csv had data corresponding to 7,882 users and 37 interest categories. Of these 37 categories (columns), two categories; "spam" and "adult" are mainly caused by spam and pornography "bots" on Twitter. Hence, we have removed any remaining user who fell in category spam or adult. This results in 7,259 remaining users after 7.3% of "adult" and "spam" users were removed. We also removed these two categories post removing the corresponding users.

We figured out the following from the correlation matrix that health_nutrition and personal_fitness are the most correlated topics with a correlation of 0.81.

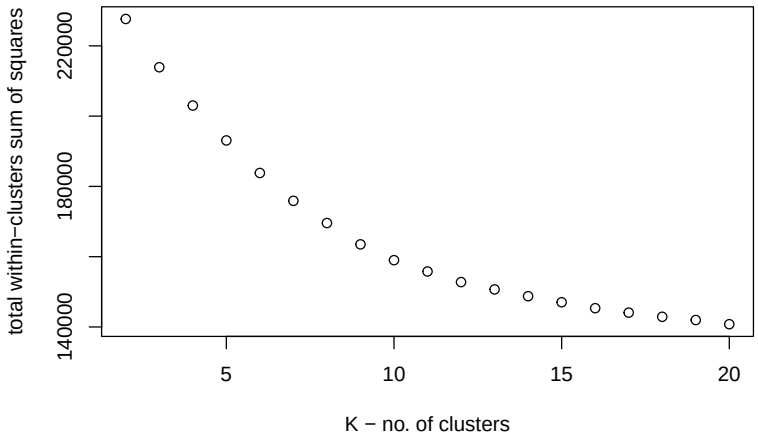
Step 2: Preliminary Data Patterns

We looked for the most frequently occuring categories among users by counting the total number of posts for a given category. We observed that the top 10 categories were: chatter, photo_sharing, health_nutrition, cooking, politics, sports_fandom, travel, college_uni, current_events and personal_fitness.

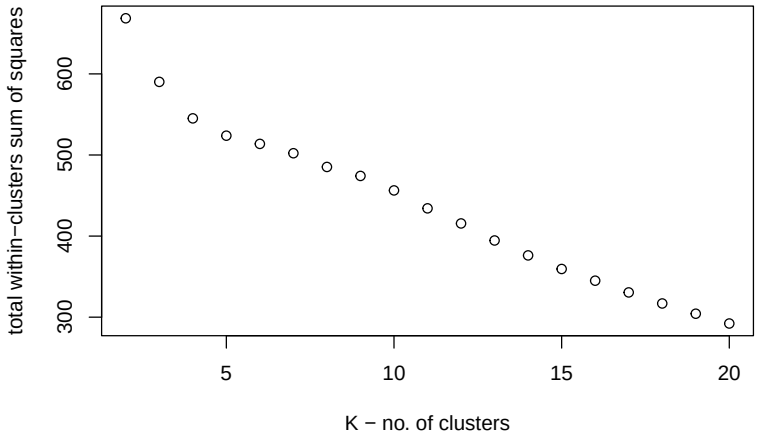
We also plotted correlation matrix to identify the correlations among all the variables. We observed following six catgories of correlations:

-> Travel, politics and news -> Online playing, college_univ, sports_playing -> Cooking, fashion, beauty -> Health_nutrition, personal_fitness, outdoors -> sports_fandom, food, religion

For defining the number of clusters, we use the elbow method (scree plot)



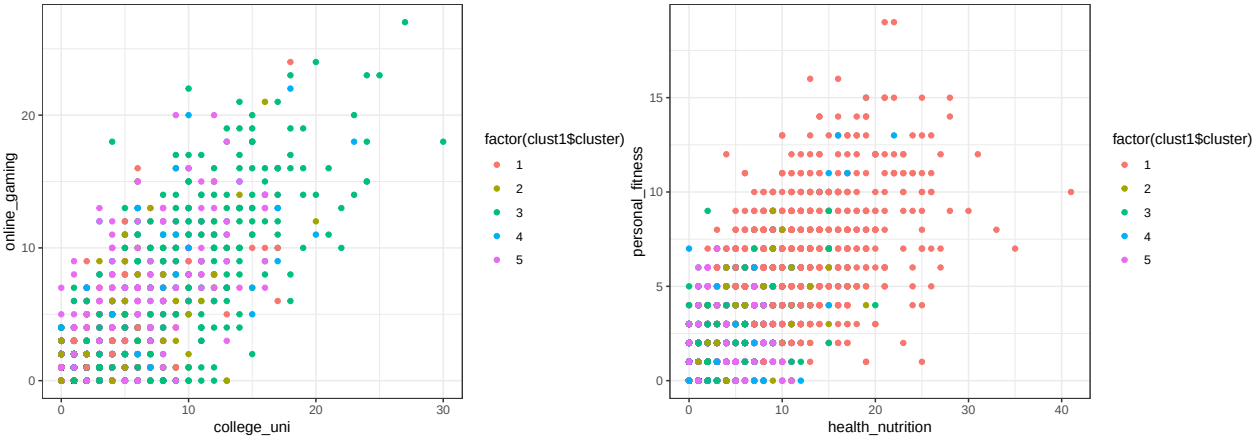
Clearly from the Scree plot, we aren't able to derive any clear elbow, hence we decided to explore model with CH index.

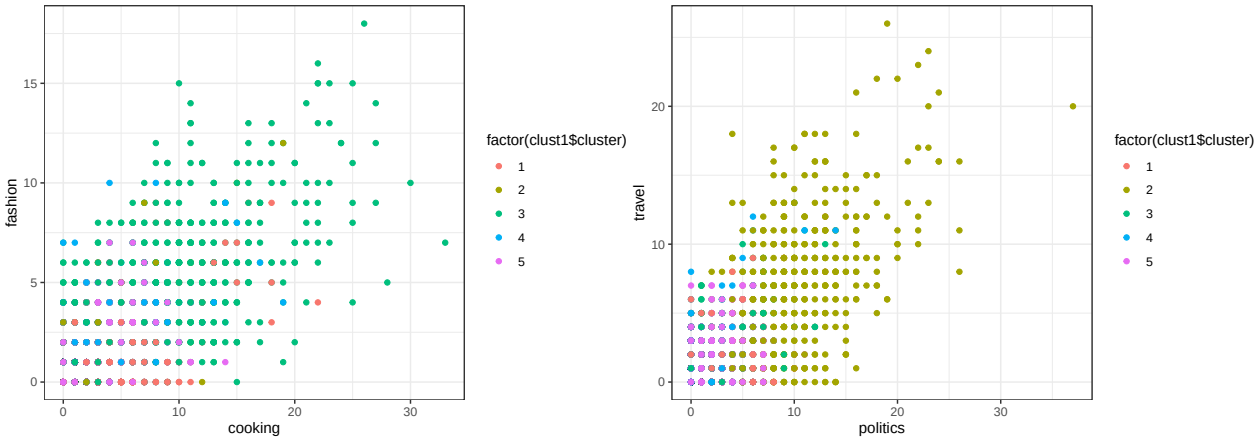


Step 3: Determining number of clusters:

We plotted Scree graph to determine elbow point which would give us the optimal K for the clustering model. However, we were not able to determine elbow point from Scree plot. Hence, we plotted CH index graph to determine K. K came out to be 5 for the plot.

Hence, we proceeded with K = 5 in our analysis





Step 4: Clustering under two parts:

Part A: K means and K means ++ clustering model

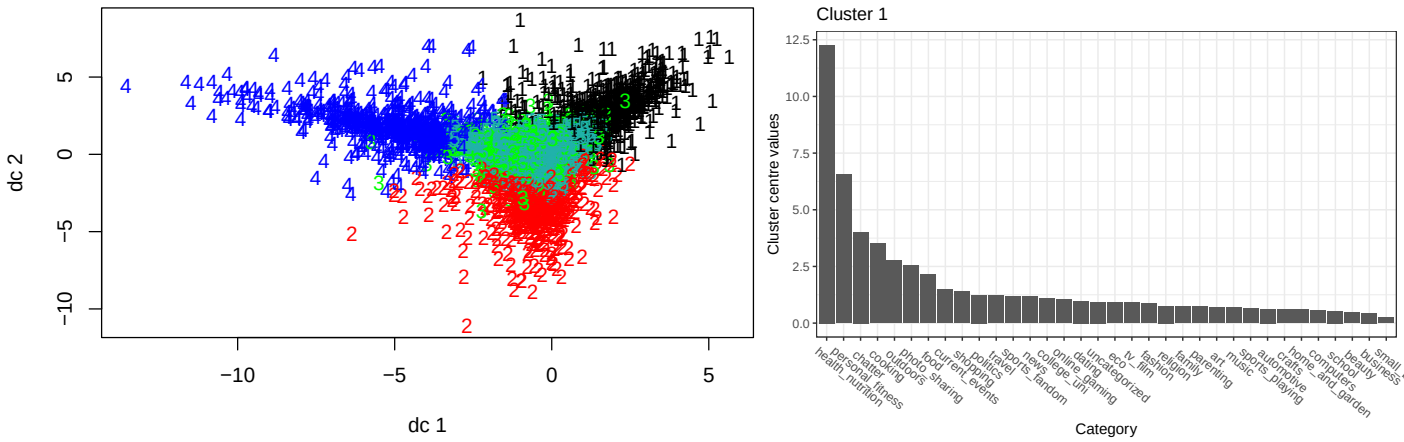
Part B: K means with PCA analysis

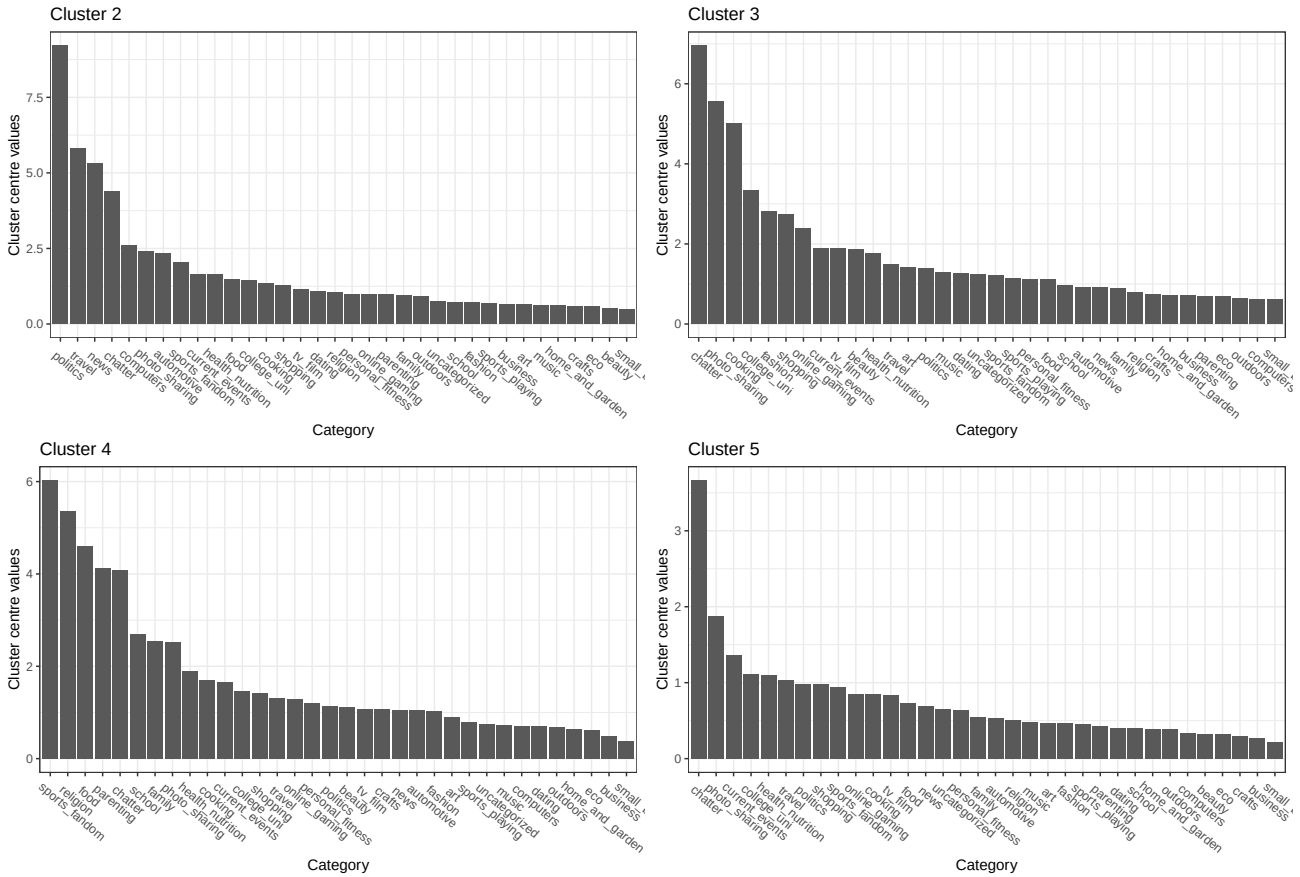
Part A - Running K means and K means ++ clustering model:

We chose K = 5 on the basis of CH plot and took nstart = 50 to run K-means and K-means++. The respective distance between 5 clusters for K-means and K-means++ were as below:

Type: Cluster1 Cluster2 Cluster3 Cluster4 Cluster5 K-means: 28401.89 21501.29 82946.21 25211.44 27857.10 K-means++: 82946.21 25211.44 21501.29 28401.89 27857.10

Overall within and between distance in clusters for K-means and K-means++ is similar. Hence, we proceeded with K-means cluster output.

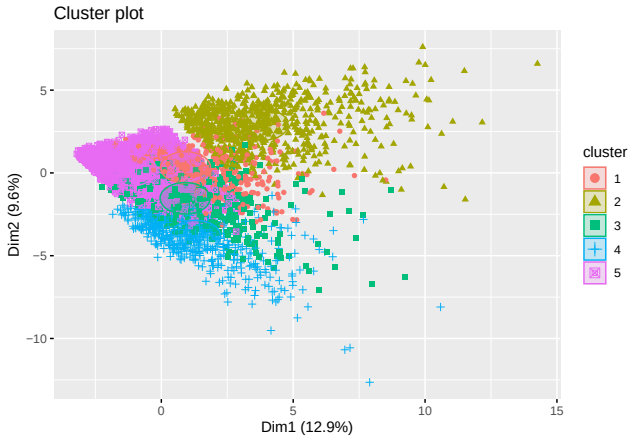




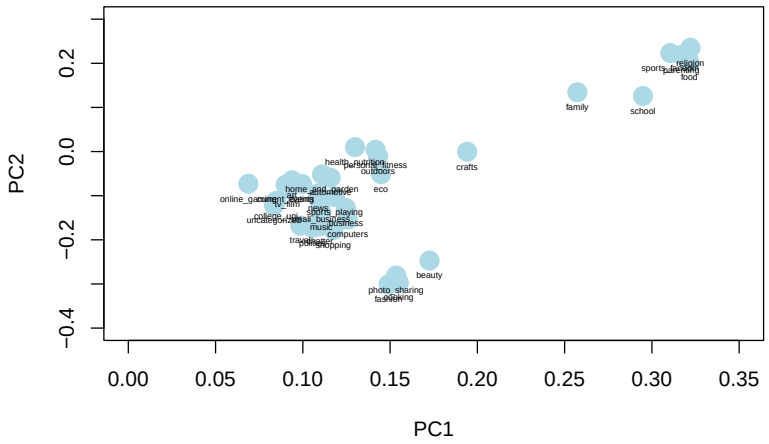
We de-scaled the variables and plotted the five clusters identified by K-means clustering.

Cluster 1: politics, travel and news Cluster 2: cooking, fashion, beauty Cluster 3: photo_sharing, college_uni, current_events, shopping, online_gaming Cluster 4: health_nutrition, personal_fitness Cluster 5: sports_fandom, religion, food, parenting

Part B - Next we also ran K-means clustering with PCA to identify the clusters.



Further we use PCA (Principle Component Analysis) to understand the composition of each cluster



Identifying the Clusters

Comparing both the clusters, we can identify the following 5 segments:

1 - The Homies

This cluster represents the people with most number of tweets in the **food**, **family**, **school**, **parenting**, **religion** and **sports_fandom** categories. These mostly represent people who majorly focused on topics relevant to parenting and hence mostly represent middle-aged to old-aged people groups.

2 - The Aware ones

This cluster represents people with most number of tweets in **crafts**, **computers**, **automotive**, **travel**, **news** and **politics** categories. This mostly represents people who are mostly youngsters, possibly new in their jobs, traveling and aware of politics.

3 - The Youngsters

This cluster represents people with most number of tweets in **eco**, **sports**, **music**, **shopping**, **online-gaming**, **college**, **dating**, **current events**, etc. categories. This majorly represents users in their schools/colleges, majorly the younger crowd.

4 - The Influencers

This cluster represents people with most number of tweets in **cooking**, **fashion**, **photo sharing** and **beauty** categories. This majorly represents influencers who are aware of their beauty and social media presence.

5 - The Health conscious

This cluster represents people with most number of tweets in **personal_fitness**, **health_nutrition** and **outdoors** categories. This majorly represents people who are more into fitness and healthy lifestyle.

Comparing Analysis from PART A and PART B

We notice that from both the analysis, using **K = 5** we get similar clusters using both K means with PCA and K means ++. These clusters are quite interesting and differ based on demographics. All the clusters are well defined based on the frequency of their tweets in respective categories.

Cluster 1 represents people who possibly are parents and belong to Gen-X. These people majorly tweet about parenting, religion, family and food. **Cluster 2** represents the people who are aware and travel, are aware of the news and politics. **Cluster 3** represents young consumers who are people more interested in photo sharing, college, current events, shopping, online_gaming, etc. This cluster majorly would represent school/college going individuals. **Cluster 4** represents influencers who are concerned about their social media image and their major interest lies in cooking, fashion and beauty. **Cluster 5** represents people who are fitness and health conscious and their interests lie in personal fitness, health nutrition and outdoors.

NutrientH20 can use these market segments to derive their marketing and positioning strategies and also can use these insights to understand consumer preferences to design products to cater to different market segments. However, as consumer preferences keep changing over-time, these segments and consumer categories need to be monitored regularly and the company needs to be flexible with their product and marketing strategy to accommodate changes.

End of Problem

Author Attribution

Objective: Predict the author of an article on the basis of that article's textual content

Step 1 : Create a 'readerPlain' function to help read the data

Step 2 : Create a combined corpus of train and test data. We are creating a combined corpus since there might be few words present in only one of the data set.

Step 3 : Tokenize the corpus by converting it to lowercase, removing numbers, punctuation, excess whitespace, stop words

Step 4 : Create a document term matrix and construct TF-IDF weights for corpus

Step 5 : Splitting the corpus to training and test data

Step 6 : Modeling 1st Iteration : We will not apply any dimensional reduction

Random Forest Model

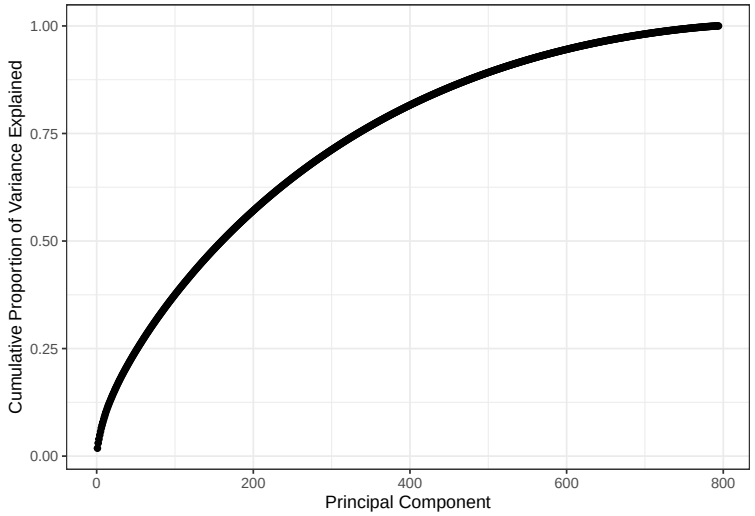
Training Random Forest without dimensional reduction and predicting on test data yields an accuracy of : 61.2 %

Lets try running KNN model and compare the accuracy with Random Forest.

Training KNN model without dimensional reduction and predicting on test data yields an accuracy of : 40.56 %

KNN generally performs poorly on data with large dimensions. Lets try reducing the data dimension using PCA and again calculate accuracy using Random Forest and KNN and check for any improvement in the accuracy.

Lets try improving the accuracy by reducing the dimensionality using PCA



We see that 500 components explain more than 50% variance in the data. Lets try training the Random Forest and KNN model with just 500 PCA components.

Training Random Forest post dimensional reduction and predicting on test data yields an accuracy of : 49 %

We see reduced accuracy post PCA dimension reduction. Hence, random forest seems to perform better on the whole data.

Lets see if there in an improvement in the accuracy of KNN model post dimension reduction.

Training KNN model without dimensional reduction and predicting on test data yields an accuracy of : 33.88 %

We don't see any improvement in the KNN accuracy after reducing the dimension as well. Hence, we get the **best accuracy of 65% using Random Forest model without reducing the dimension of the data**.

————————————————————**End of Problem**————————————————————

Association rule mining

Objective : Using the data, find interesting associations in grocery items in a market grocery list.

We notice that the file *groceries.txt* is not present in a format expected by the **arules** package. However, the **arules** package contains a function called **read.transactions** that processes files with transactions data and creates a transactions object, which is in the format of a Sparse Matrix.

On summarizing the resulting dataset, we found that a total of 169 items are present in the grocery list, spread across 9835 baskets. The most common items are as follows.

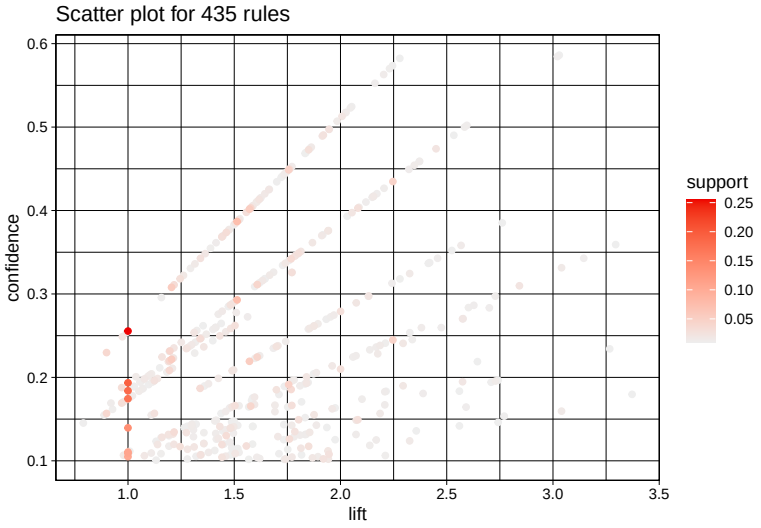
- Whole milk (2513 baskets)
- Other vegetables (1903 baskets)
- Rolls/buns (1809 baskets)
- Soda (1715) baskets
- Yogurt (1372 baskets)

On an average, there were around 4-5 items present in each basket, with a minimum of 1 item and a maximum of 32 items in a basket.

Association Set up and Analysis

We now use the **apriori** function to analyze associations. We use initial values of **support = 1%** and **confidence = 10%**. There are 435 association rules that created as a result, which we feel is a good number to move forward with.

Note that we pick the above support and confidence values keeping in mind that there are many grocery items and we would want to pick up a high number of items, which the support of 1% allows us to do. Moreover, we want to be sufficiently confident about one's purchases, thus choosing a confidence of 10%.



From the above plot, we see that higher the lift, higher the confidence. This is also shown from the color scale where we see that high support items have low lift and confidence, as compared to low support items.

Top 20 Association Rules by lift

```
                                                    rules     lift
```

342 {whole milk,yogurt} => {curd} 3.372304 371 {citrus fruit,other vegetables} => {root vegetables} 3.295045 357 {other vegetables,yogurt} => {whipped/sour cream} 3.267062 386 {other vegetables,tropical fruit} => {root vegetables} 3.144780 60 {root vegetables} => {beef} 3.040367 59 {beef} => {root vegetables} 3.040367 370 {citrus fruit,root vegetables} => {other vegetables} 3.029608 385 {root vegetables,tropical fruit} => {other vegetables} 3.020999 417 {other vegetables,whole milk} => {root vegetables} 2.842082 348 {other vegetables,whole milk} => {butter} 2.770630 341 {curd,whole milk} => {yogurt} 2.761356 67 {whipped/sour cream} => {curd} 2.742150 66 {curd} => {whipped/sour cream} 2.742150 363 {other vegetables,whole milk} => {whipped/sour cream} 2.729417 405 {other vegetables,yogurt} => {root vegetables} 2.728698 360 {whole milk,yogurt} => {whipped/sour cream} 2.709053 393 {other vegetables,yogurt} => {tropical fruit} 2.700550 372 {other vegetables,root vegetables} => {citrus fruit} 2.644626 411 {other vegetables,rolls/buns} => {root vegetables} 2.627525 389 {tropical fruit,whole milk} => {root vegetables} 2.602365 The associations seen above make a lot of sense. For example, the most likely item to buy is curd provided one buys whole milk and yogurt. We see that vegetables, root and otherwise make up the most likely items in a basket provided one buys fruits, milk or other categories of vegetables. From the top 20, we can hence see that milk products, fruits and vegetables make up the most likely items in a basket provided other items from the aforementioned list are also picked up.

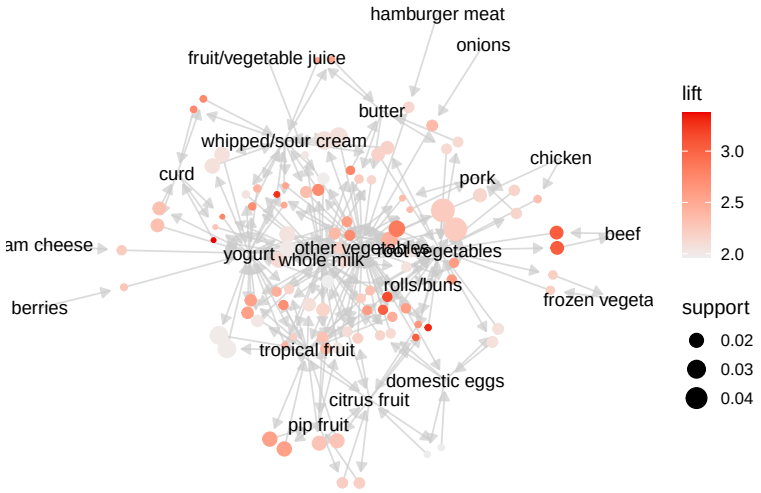
Top 20 Association Rules by confidence

```
                                                    rules confidence
```

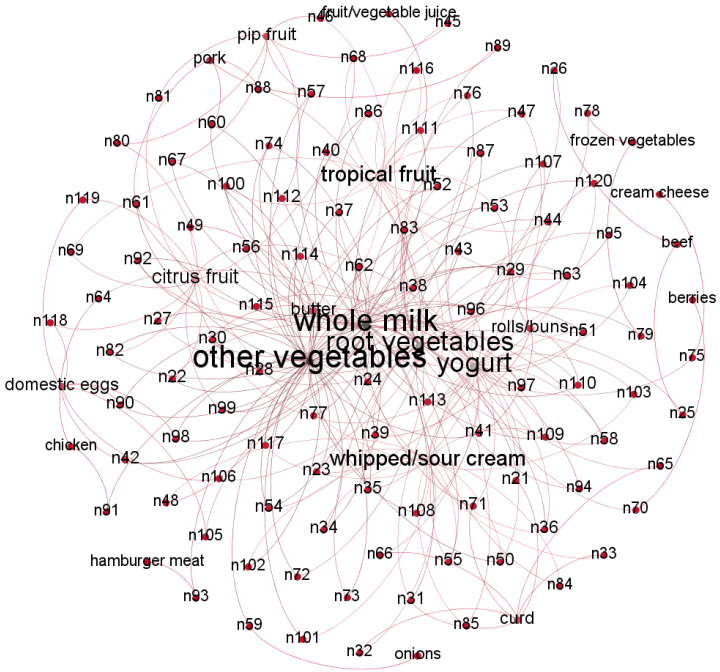
370 {citrus fruit,root vegetables} => {other vegetables} 0.5862069 385 {root vegetables,tropical fruit} => {other vegetables} 0.5845411 340 {curd,yogurt} => {whole milk} 0.5823529 346 {butter,other vegetables} => {whole milk} 0.5736041 388 {root vegetables,tropical fruit} => {whole milk} 0.5700483 406 {root vegetables,yogurt} => {whole milk} 0.5629921 349 {domestic eggs,other vegetables} => {whole milk} 0.5525114 358 {whipped/sour cream,yogurt} => {whole milk} 0.5245098 412 {rolls/buns,root vegetables} => {whole milk} 0.5230126 364 {other vegetables,pip fruit} => {whole milk} 0.5175097 394 {tropical fruit,yogurt} => {whole milk} 0.5173611 430 {other vegetables,yogurt} => {whole milk} 0.5128806 361 {other vegetables,whipped/sour cream} => {whole milk} 0.5070423 409 {rolls/buns,root vegetables} => {other vegetables} 0.5020921 403 {root vegetables,yogurt} => {other vegetables} 0.5000000 352 {fruit/vegetable juice,other vegetables} => {whole milk} 0.4975845 134 {butter} => {whole milk} 0.4972477 74 {curd} => {whole milk} 0.4904580 355 {whipped/sour cream,yogurt} => {other vegetables} 0.4901961 415 {other vegetables,root vegetables} => {whole milk} 0.4892704 This again reinforces the observation that vegetables and whole milk make up for the items we are most confident about after analyzing the data.

Network Plots

We now plot the graphical network of the top 100 associations by lift.



The high number of rules makes it difficult to visualize the network. However, the labels indicate the most prominent items in the baskets, in terms of degree and betweenness. We show a more sophisticated image below that takes the top 100 associations by lift.



End of Problem