TakeHomeExamPart-2

Nittala Venkata Sai Aditya, Akhila Guttikonda, Vishwak Venkatesh, Srividya Rayaprolu

2022-08-08

Github Link : https://github.com/vishwak-venkatesh/STA380-Exam

Q1:Probability Practice

Part A

P(yes)=0.65 P(No)=0.35

P(Rc)=0.3 P(TC)=0.7

P(Y|RC)=p(N|RC)=0.5

P(Yes) = P(RC)*P(Y|RC) + P(TC)*P(Y|TC)

0.65 = 0.3*0.5+0.7*P(Y|TC) 0.65 = 0.15+0.7*P(Y|TC) 0.5/0.7 = P(Y|TC)

P(Y|TC) =0.714

- Probability of truthful clickers answering Yes is **0.714**

Part B

P(Positive |Disease) = 0.993 P(Negative | No Disease) = 0.9999

p(Disease) = 0.000025

- From Total Probability Rule

P(Positive) = P(Disease)* P(Positive|Disease)+P(Not Positive) * P(Positive | No Disease)

P(positive) = 0.000025 * 0.993 + 0.999975 * 0.0001 P(Positive) = 0.00002482 + 0.00009999 P(Positive) = 0.0001248

We have to find P(Disease | Positive)

- Going by Naive Bayes Theorem

P(disease | Positive) = (P(Positive|Disease)* P(Disease))/P(Positive)

P(disease | Positive) = (0.993*0.000025)/0.0001248

P(disease | Positive) ~ 0.2

- Final Probability = **0.2**

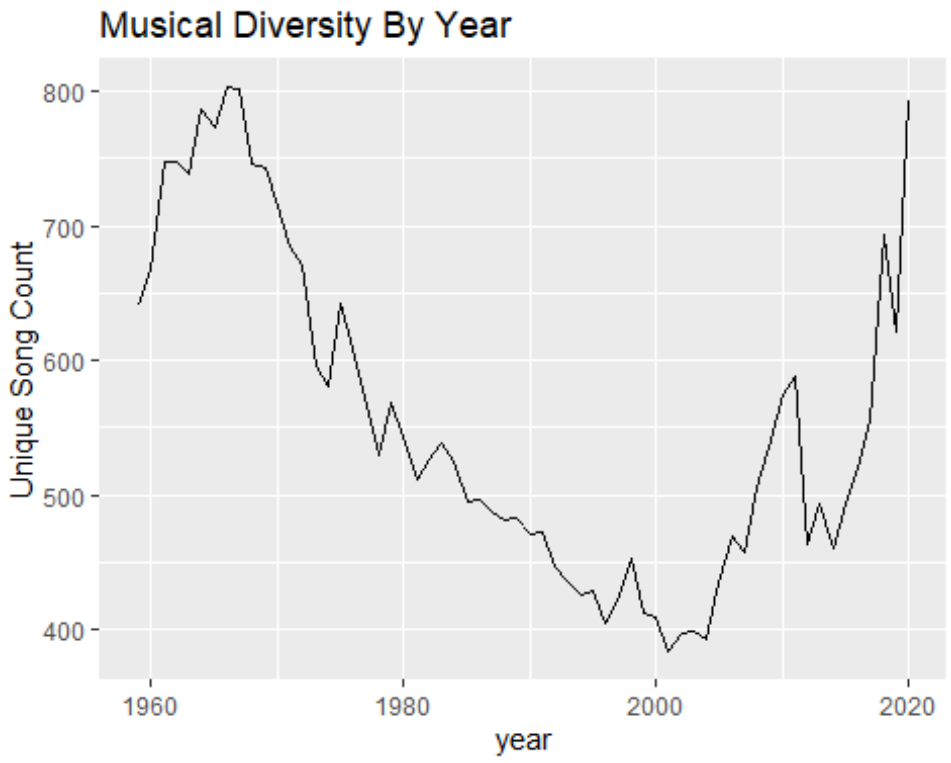
Q2: Wrangling the Billboard Top 100

Part A

Top 10 Popular Billboard songs

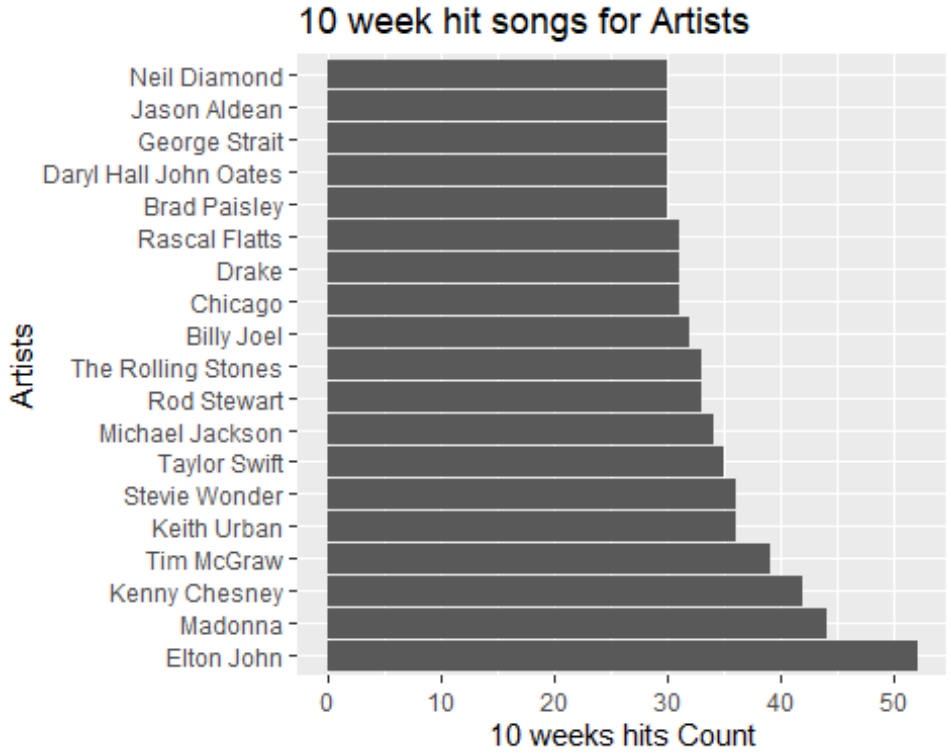
song	performer	count
Radioactive	Imagine Dragons	87
Sail	AWOLNATION	79
Blinding Lights	The Weeknd	76
I'm Yours	Jason Mraz	76
How Do I Live	LeAnn Rimes	69
Counting Stars	OneRepublic	68
Party Rock Anthem	LMFAO Featuring Lauren Bennett & GoonRock	68
Foolish Games/You Were Meant For Me	Jewel	65
Rolling In The Deep	Adele	65
Before He Cheats	Carrie Underwood	64

Part B



Musical Diversity Graph Per year.1960's was a great year for musical diversity

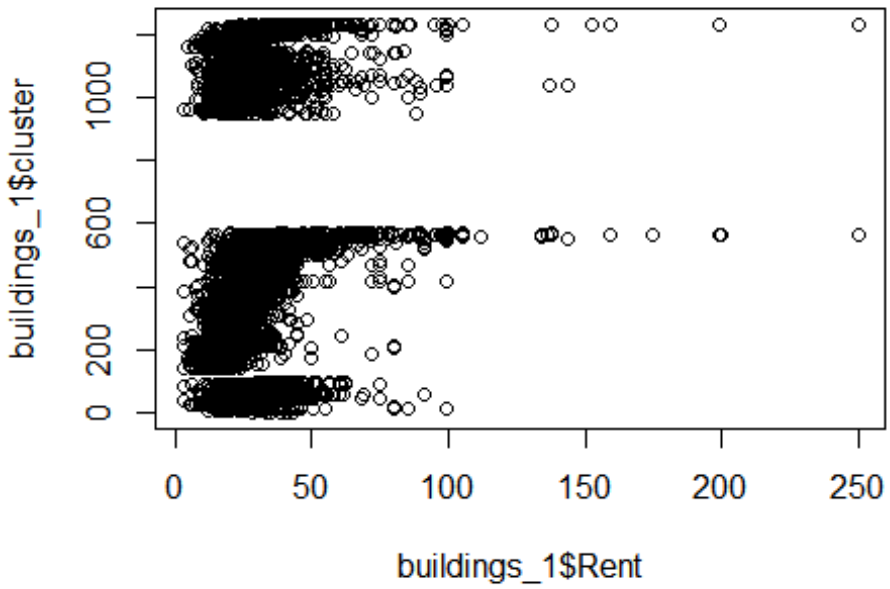
Part C

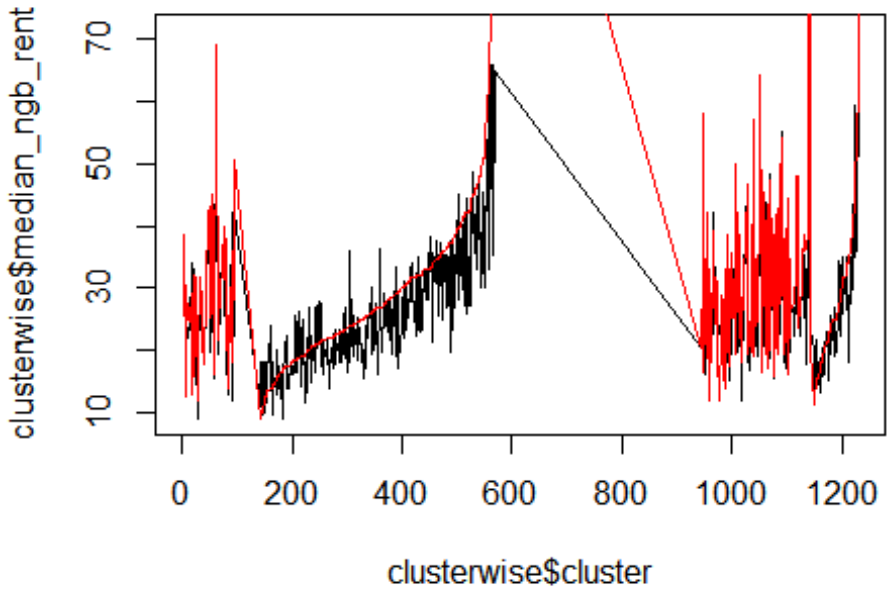


10 week hit songs count for Artists. Billy Joel has the most number of 10 week hit songs

Q3. Visual story telling part 1: green buildings

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## [1] 7894

## [1] 7679
```





Firstly we have compared the green buildings in each cluster to non-green buildings in the same cluster to have an advantage of a green building. It makes sense that buildings with occupancy<10% should be removed. There are 215 such buildings which are removed

As per the observations, the average rents of green buildings is higher but the average electricity cost of the buildings is almost equal. Average gas cost of both gb and ngb is almost equal.

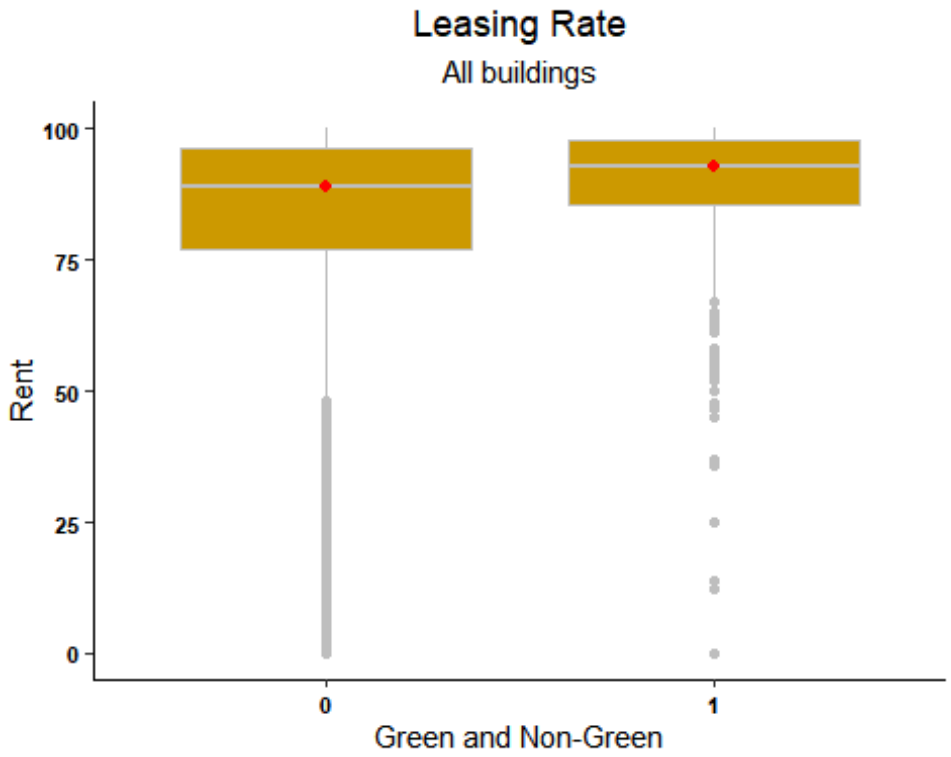
Instead of calculating the median values over the entire data of green and non green buildings, excel guru should have chosen only the buildings with similar number of floors and sq ft.

Exploratory analysis: green buildings

Step 1: Analysis on all buildings

1.Obtain a brief idea about the columns in the dataset

```
## 'data.frame':    7894 obs. of  23 variables:
##  $ CS_PropertyID    : int  379105 122151 379839 94614 379285 94765 236739
234578 42087 233989 ...
##  $ cluster          : int  1 1 1 1 1 1 6 6 6 6 ...
##  $ size             : int  260300 67861 164848 93372 174307 231633 210038
225895 912011 518578 ...
##  $ empl_gr          : num  2.22 2.22 2.22 2.22 2.22 2.22 4.01 4.01 4.01 4.
01 ...
##  $ Rent             : num  38.6 28.6 33.3 35 40.7 ...
##  $ leasing_rate     : num  91.4 87.1 88.9 97 96.6 ...
##  $ stories          : int  14 5 13 13 16 14 11 15 31 21 ...
##  $ age              : int  16 27 36 46 5 20 38 24 34 36 ...
##  $ renovated        : int  0 0 1 1 0 0 0 0 0 1 ...
##  $ class_a          : int  1 0 0 0 1 1 0 1 1 1 ...
##  $ class_b          : int  0 1 1 1 0 0 1 0 0 0 ...
##  $ LEED             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Energystar       : int  1 0 0 0 0 0 1 0 0 0 ...
##  $ green_rating     : int  1 0 0 0 0 0 1 0 0 0 ...
##  $ net              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ amenities        : int  1 1 1 0 1 1 1 1 1 1 ...
##  $ cd_total_07      : int  4988 4988 4988 4988 4988 4988 2746 2746 2746 27
46 ...
##  $ hd_total07       : int  58 58 58 58 58 58 1670 1670 1670 1670 ...
##  $ total_dd_07      : int  5046 5046 5046 5046 5046 5046 4416 4416 4416 44
16 ...
##  $ Precipitation    : num  42.6 42.6 42.6 42.6 42.6 ...
##  $ Gas_Costs        : num  0.0137 0.0137 0.0137 0.0137 0.0137 ...
##  $ Electricity_Costs: num  0.029 0.029 0.029 0.029 0.029 ...
##  $ cluster_rent     : num  36.8 36.8 36.8 36.8 36.8 ...

## [1] "Median rent of green buildings :  27.6"

## [1] "Median rent of green buildings :  25"
```



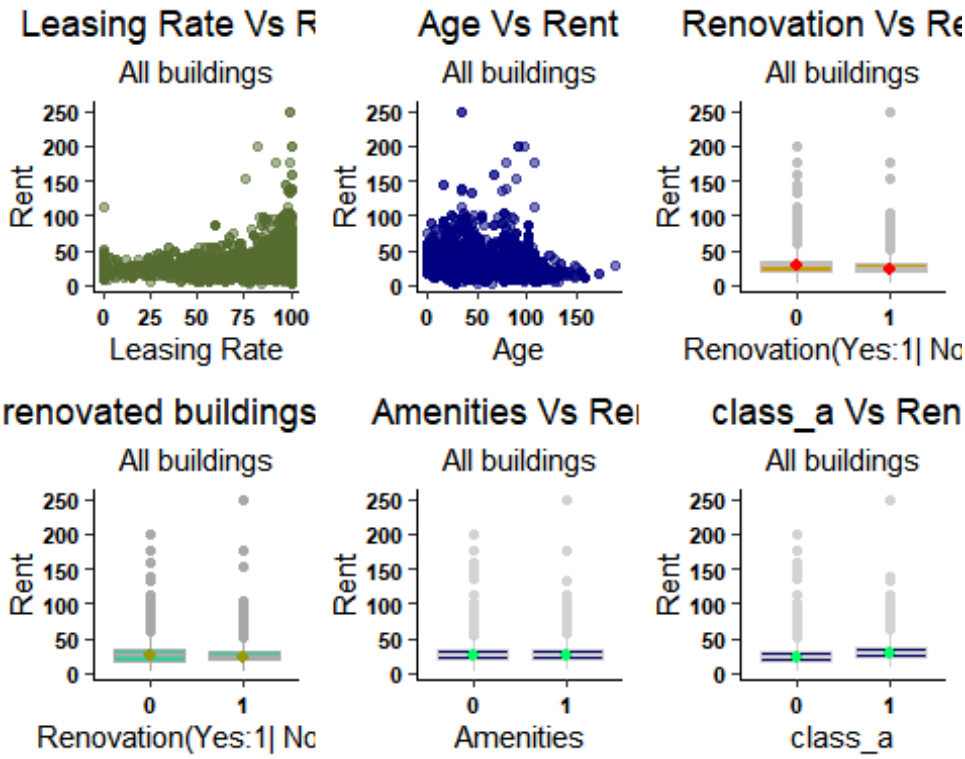

```
## [1] "Median leasing rate of green buildings :  92.92"

## [1] "Median leasing rate of non green buildings :  89.17"
```

- Green buildings have a higher occupancy rate when compared to non-green buildings

- As the stats guru, pointed out the median of green buildings($27.6) is higher than the median of non-green buildings($25). But he did not consider the effect of confouding variables while performing the analysis. In teh next section, we will check or the influence of confounding vairables on the Rent of green and non-green buildings

2.We will create some hypotheses using which we will steer through the data to understand if the data agrees with the respective hypotheses
a.Less leasing_rate might be a proxy for less demand for commercial real-estate
b.Rent decreases with age for buildings
c.Renovated buildings with age >30 years get higher rent than buidings with age < 30 without renovation
d.Buidings with amenities have higher rents than the other buildings e.class_a buildings have higher rent than the other buildings

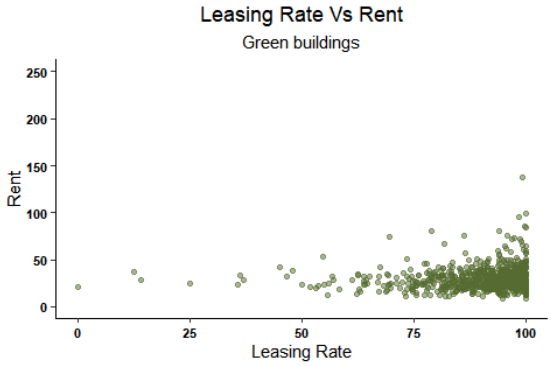
Let's plot the respective distribution to find if the hypotheses can be supported using the relationships

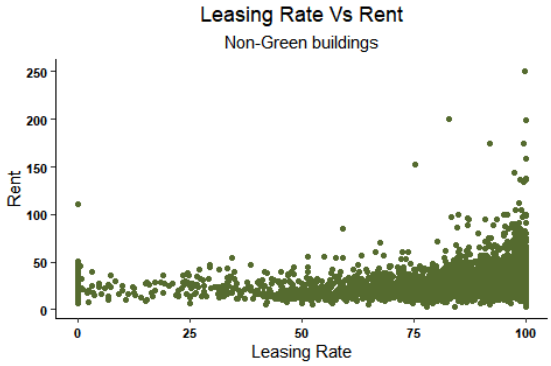


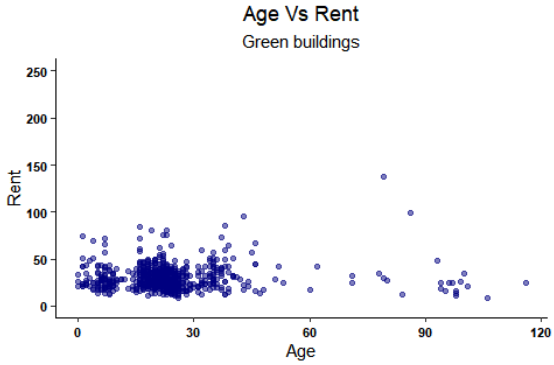
Findings:
* Age has no visible relation with Rent when all buildings are considered
* Buildings with Amenities and class_a quality material have slightly higher rent than the other buildings

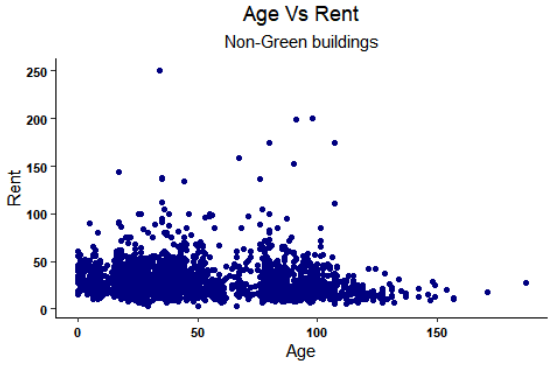
Step 2: Comparison of different variables for Green and Non-Green buildings

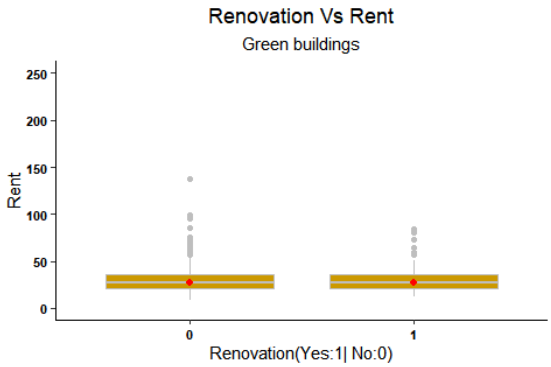
Lets check the above hypotheses for Green and Non-Green buildings separtely to see if there is any influence

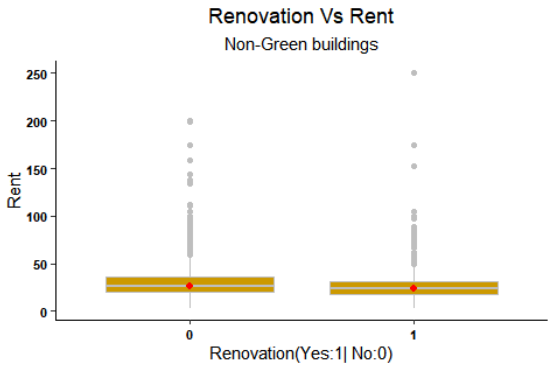

Leasing Rate Vs Rent
Green buildings
Rent
Leasing Rate

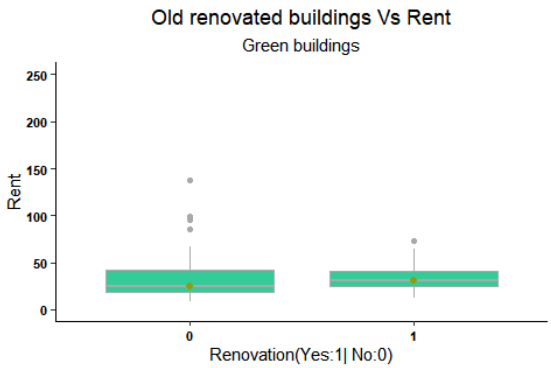

Leasing Rate Vs Rent
Non-Green buildings
Rent
Leasing Rate


Age Vs Rent
Green buildings
Rent
Age

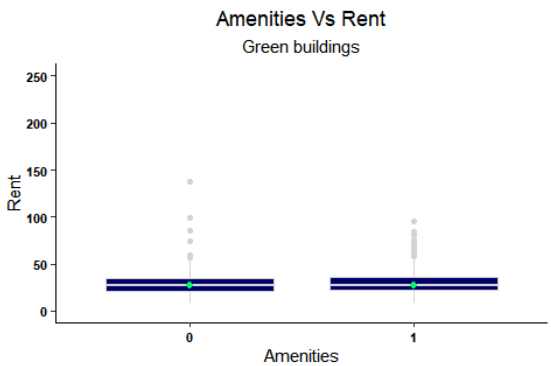

Age Vs Rent
Non-Green buildings
Rent
Age

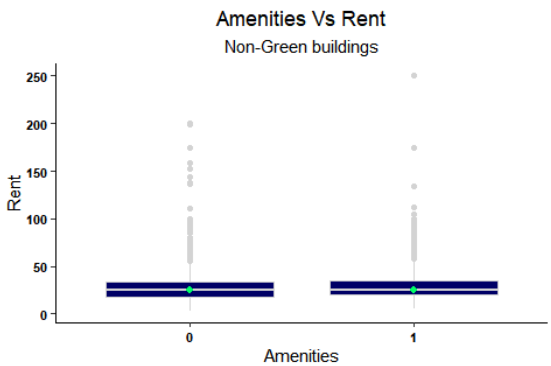

Renovation Vs Rent
Green buildings
Rent
Renovation(Yes:1| No:0)

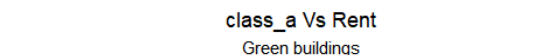

Renovation Vs Rent
Non-Green buildings
Rent
Renovation(Yes:1| No:0)

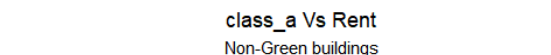

Old renovated buildings Vs Rent
Green buildings
Rent
Renovation(Yes:1| No:0)


Old renovated buildings Vs Rent
Non-Green buildings
Rent
Renovation(Yes:1| No:0)


Amenities Vs Rent
Green buildings
Rent
Amenities

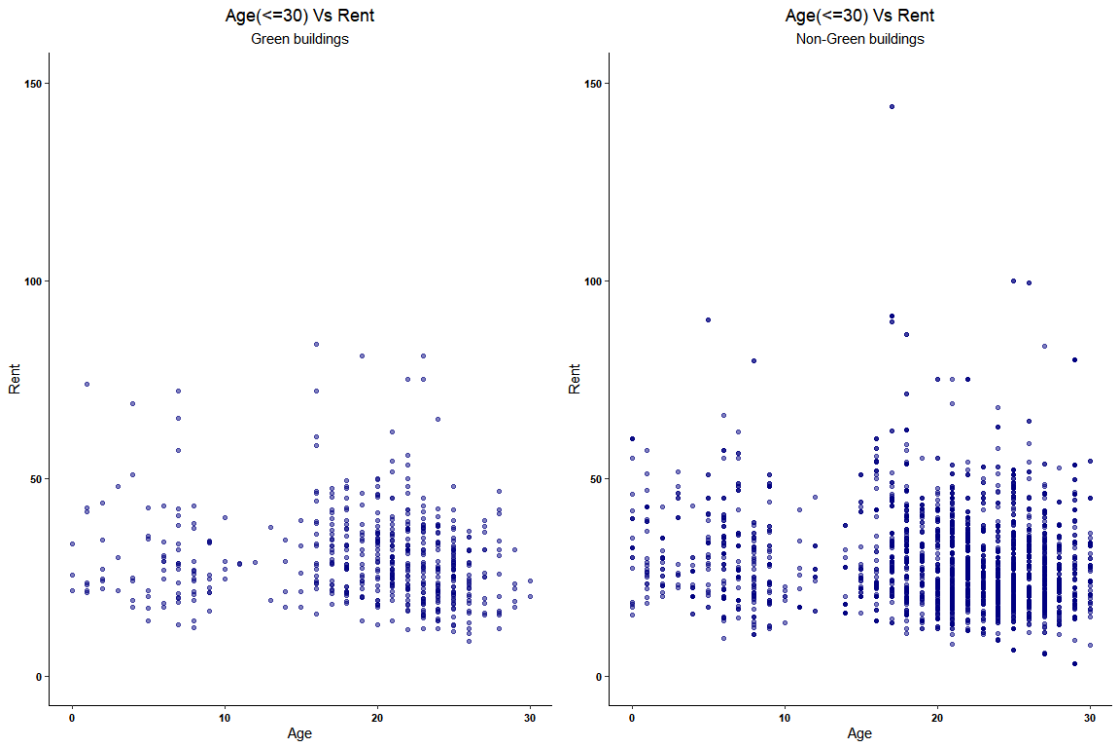

Amenities Vs Rent
Non-Green buildings
Rent
Amenities


class_a Vs Rent
Green buildings


class_a Vs Rent
Non-Green buildings

Findings:
* Older Green Buildings have the possibility of charging higher rents when they are renovated
* There are no variables that affect the distribution of rent even after the buildings are split into green and non-green buildings

Step 3: Deep Diving into some of the potential variables to see the difference between rents between green and non-green buildings




```
## [1] "Median rent of green buildings less than 30 years of age: 28"

## [1] "Median rent of non - green buildings less than 30 years of age: 27"
```

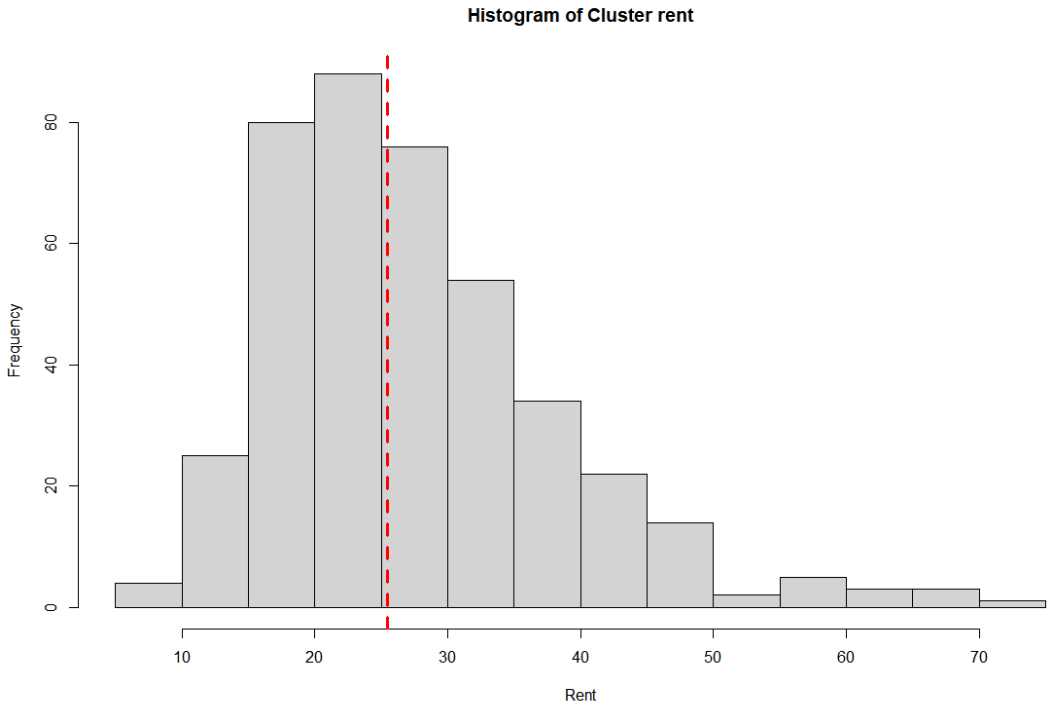
Findings:
* Age of the building does not affect the rent of the buildings as the green buildings have consistently higher rents across ages

After exploring multiple variables, it is clear that there is no one variable that affects the rent and clearly people are willing to pay more rent based on the green perception of the building although there is no way to quantify that experience

Step 4: As it is evident that people are willing to pay more for the green buildings,lets come up with an estimate for the returns on building a green building

1.Lets consider a local market(cluster) to check the probability of receiving a particular amount of rent
* Let us check the distribution of cluster rents to understand the local markets

* You can observe that more than 50% of the markets have rent less than $25 rent



We can further calculate the number of local markets in which the rent for green building is higher than the median cluster rent as median is more robust to outliers

- Green buildings have higher rents than the median rents in more than 75% of the local markets and on an averge it is $4.89

- In about 25% of the local markets, green buildings have lesser rent than the median rents and the value is $3 on an average

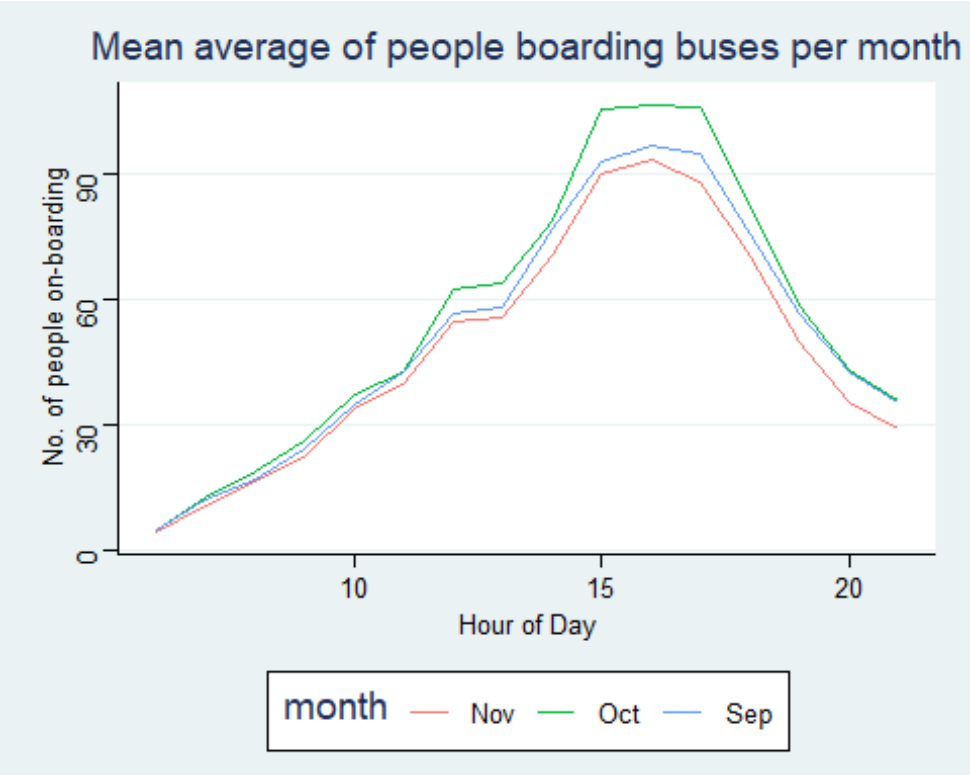
- With these observations,we can conclude that there is more than 75% chance that you will earn higher rents than the average in the local markets with a value more than $4.89

2.Estimate for calculating the returns on building a green building
* If we consider the mean of the differences between green buildings and the median local market rents, we see that green buildings get ~$3 more than the non-green buildings

When we do the calculation based on floors and sq ft, rent of green buildings is around $18.75 more than non-green buildings resulting in $18.75*250000 = $4,687,500 extra revenue per year which means that we have covered the extra 5% premium within an year. This is a great financial move to build a green building.

Q4. Visual story telling part 2: Capital Metro data

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



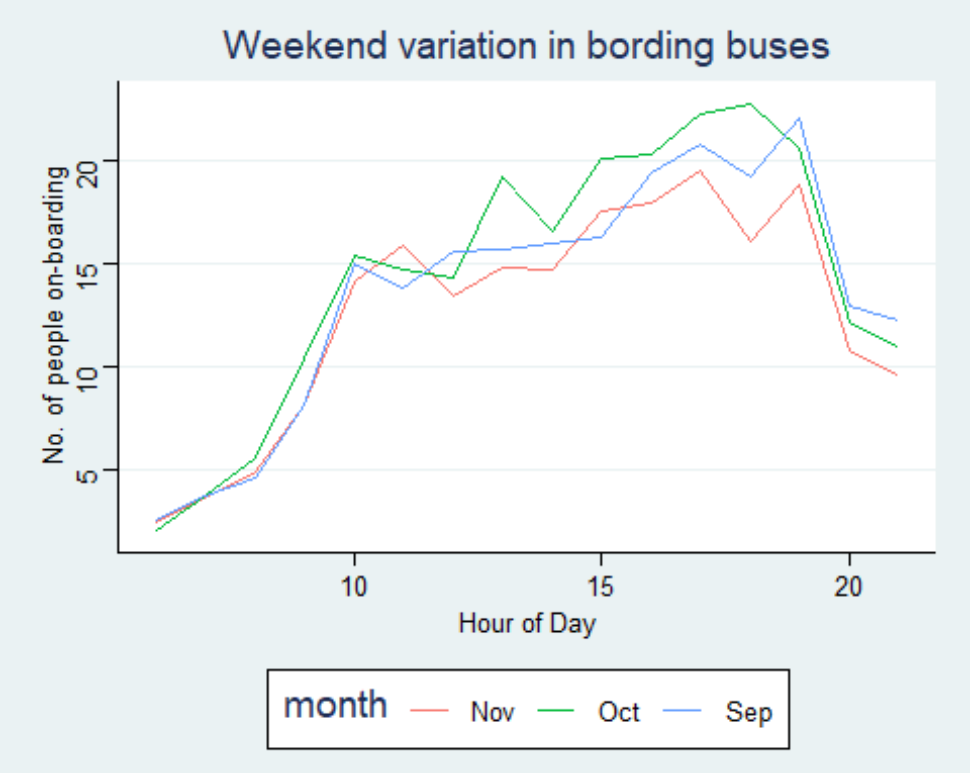
A look at the monthly average of people boarding the buses during different times of the day

A look at the monthly average of people off boarding the buses during different times of the day

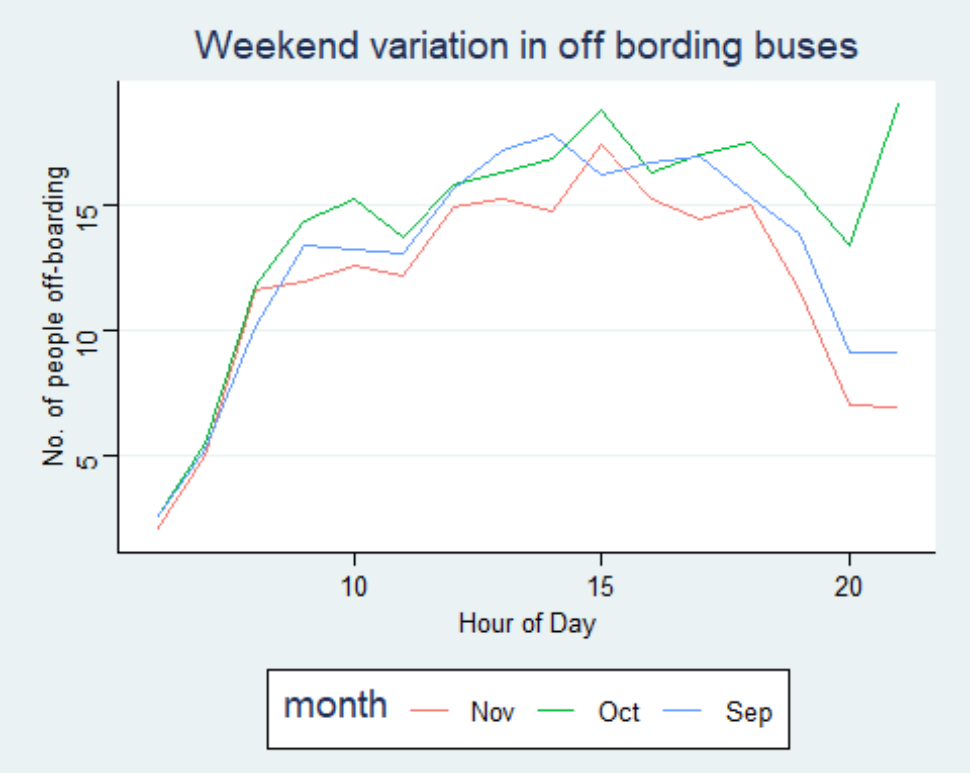
We can make the following observations -

- The no. of people on-boarding the bus peaks around 4-6pm in the evening which is when most classes get over and students are heading home.
- No. of people off-boarding the bus is highest in the morning hours, possibly when most students get off at campus for their morning lectures.
- The distribution for no. of people on-boarding and off-boarding doesn't change much in any month
- The graphs are not in sync (spikes in on-boarding don't coincide with spikes in off-boarding). This may be the case because students all get off at the same location in campus together but they board the bus over a span of couple of hours (so average boarding per hour is low but off-boarding per hour is high) with the same logic being applied to spikes in on-boarding count.
- Average ridership is the least in the month of November (maybe because it's too cold and students don't want to take public transport) and most in the month of October

This distribution seems to be heavily influenced by students going to and from campus for college. Let's try to look at the distribution on the weekends -



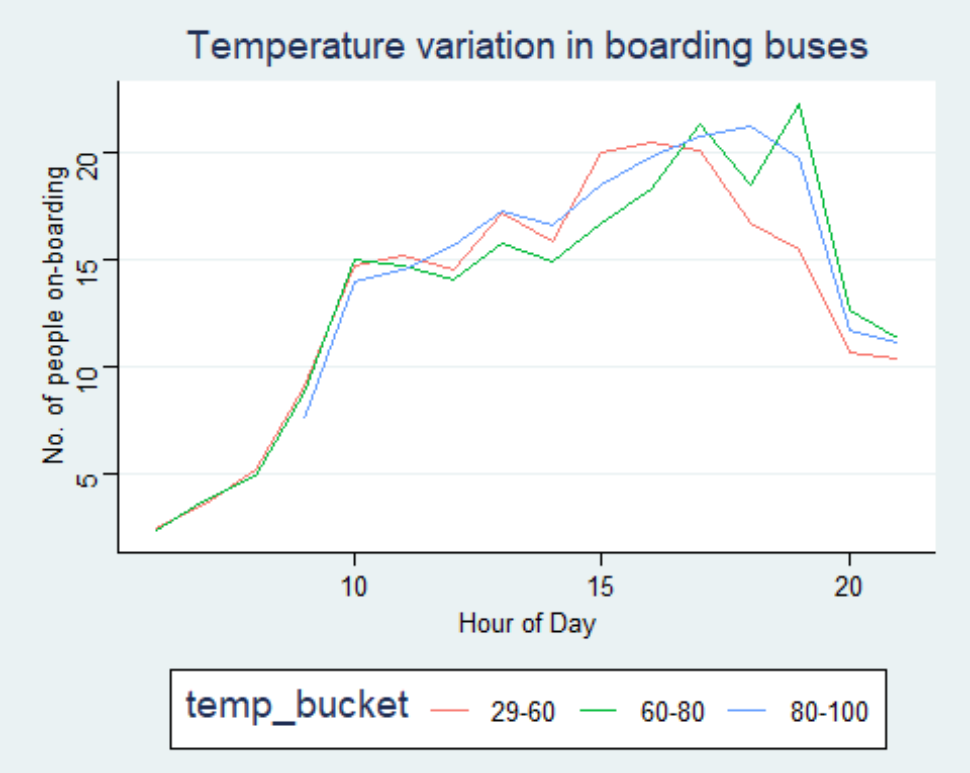
A look at the weekly average of people boarding the buses during different times of the day on weekends



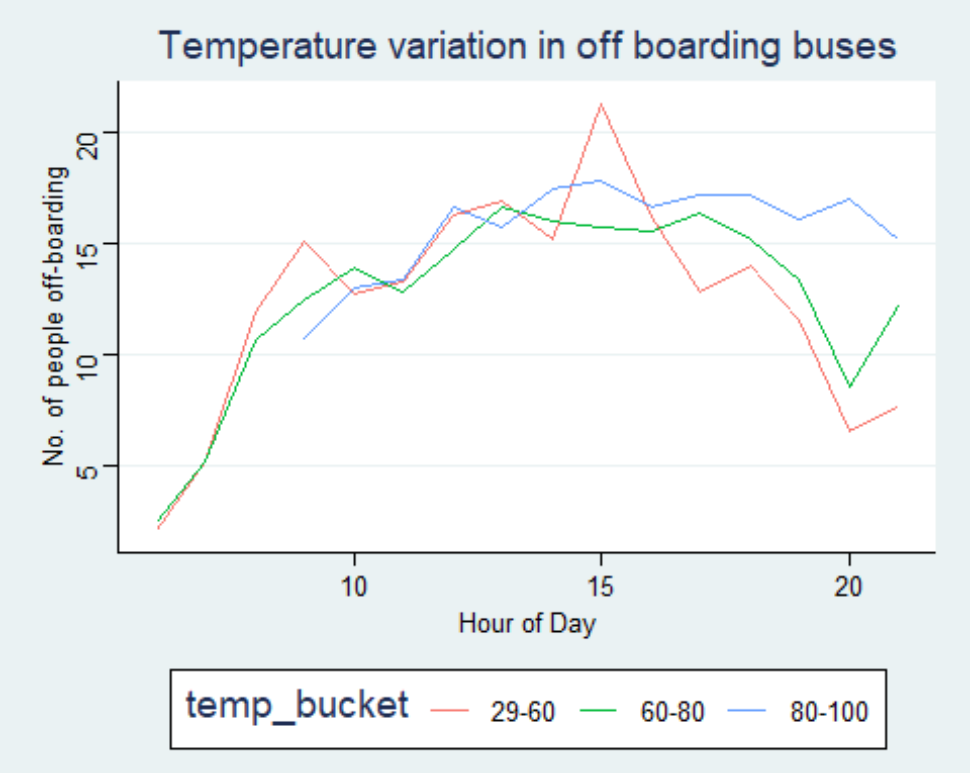
A look at the weekly average of people off boarding the buses during different times of the day on weekends

- The counts are much more varying throughout the day now
- There is an interesting spike in no. of off-boarding people in October towards the end of the day. Maybe this is because October is when students have their mid-terms for the semester so they tend to stay late on campus and go home during late hours of the day.

Let's also look at how weekend ridership changes based on temperature (since the weekday ridership is expected to not be affected by temperature since students have to go to college regardless) -



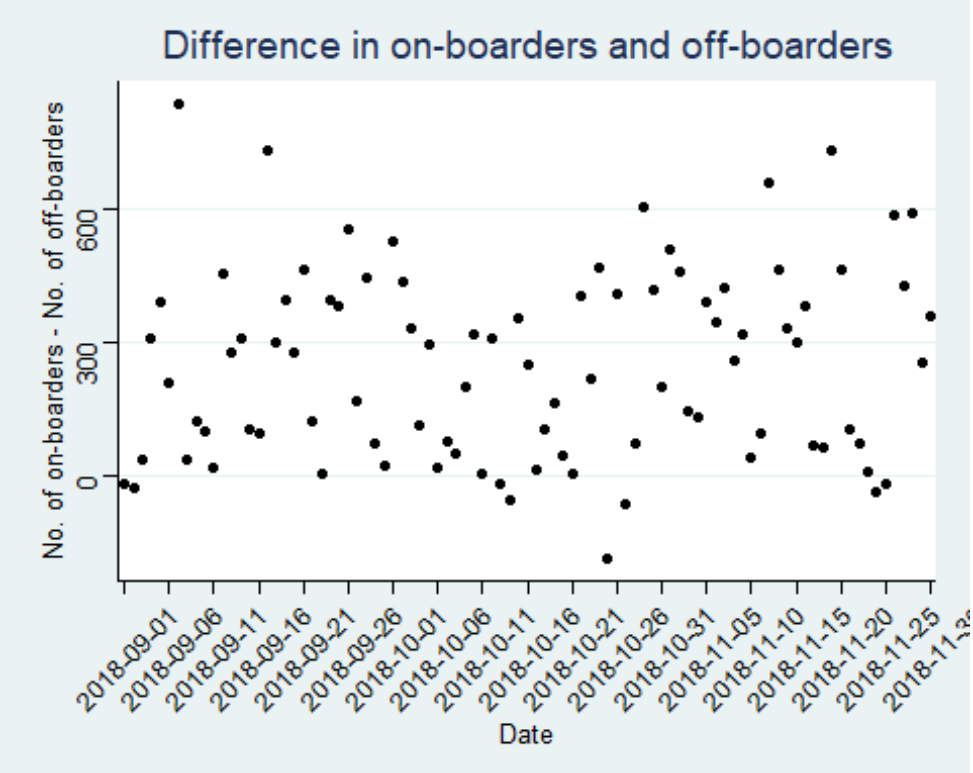
A look at the average of people boarding the buses during different times of the day based on the temperature



A look at the average of people off boarding the buses during different times of the day based on the temperature

- The temperature doesn't seem to affect ridership during the weekend so much since the patterns and numbers match those when we don't account for temperature separately. We've got some pretty interesting insights from these graphs!

Finally, let's look at the difference in total on-boardings and off-boardings per day -



A look at the differnet in the off boarders and on boarder during this period

- We can see that there is a huge discrepancy between no. of on-boarders and no. of off-boarders every day (ideally the difference should be 0 unless ~300 people are hiding in the bus at the end of each day). Capital Metro needs to work on the optical metro system a bit more to get an accurate count!

Q5: Portfolio Modelling

We want to pick a diversified portfolio which can give all the different combinations of ETFs and you try to get the maximum return possible.

Portfolio -1

Scenario-1: Assign equal weights

- EQUITY ETFS

We want to experiment with picking Equity ETF first and seeing if they have any impact on the VAR. The idea is to analyse if just picking different kinds of large cap/mid cap or small cap funds in the Equity asset class has any bearing on the final Var. The first portfolio consists of 3 large cap, 2 medium cap and 2 small cap and 1 multi cap equity ETF's. The aim is to see how the size of the assets have an impact on the overall VAR.

The following ETF's are part of portfolio - 1

- Vanguard (VOO) : This ETF tracks the S&P 500 Index, one of the most famous benchmarks in the world and one that tracks some of America's largest companies. VOO is more diversified than most, containing just over 500 securities in total. As a result, this fund could serve as a building block for many portfolios making it an excellent choice for many buy and holders, especially for those looking to keep costs at a minimum. It also has a low expense ratio of 0.03%. It has a mix of technology, financial and electronic sector as well. Asset Cap - 273,793 millions

- Invesco QQQ Trust (QQQ) : This ETF offers exposure to one of the world's most widely-followed equity benchmarks, the NASDAQ, and has become one of the most popular exchange-traded products. The significant average daily trading volumes reflect that QQQ is widely used as a trading vehicle, and less as a components of a balanced long-term strategy. This also is generally in the technological sector. Asset Cap - 177,912 millions

- Shares Core S&P 500 ETF (IVV) : IVV has become one of the largest ETFs in the world, offering exposure to one of the world's best-known and most widely followed stock indexes. This ETF tracks the S&P 500 Index, which includes many large and well known U.S. firms. It has a high P/E ratio and dividend along with a low expense ratio. Asset Cap - 308,926 millions

- iShares Core S&P Mid-Cap ETF (IJH) : This ETF is one of several ETFs available that offers exposure to mid cap U.S. stocks, an asset class that can make up a significant portion of long-term, buy-and-hold portfolios.The expense ratio is competitive with the other options out there. Finance and Producer Management are the more prevalent sectors

- Vanguard Mid-Cap ETF (VO) : VO offers exposure to a balanced portfolio of stocks, including close to 460 individual names and spreading exposure relatively evenly. The expense ratio is among the cheapest in the category making it an excellent choice for those looking to keep costs to an absolute minimum.

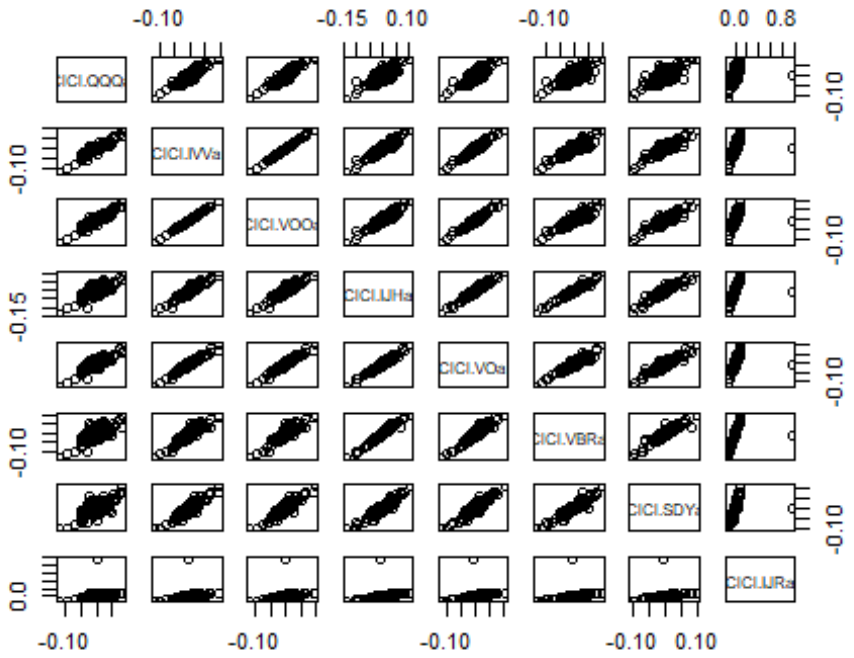
- Vanguard Small Cap Value ETF (VBR) : VBR seeks to replicate a benchmark which offers exposure small cap firms that exhibit value characteristics in the U.S. equity market. The investment thesis behind small caps is that these firms are likely to provide strong growth prospects to a portfolio and should have a much easier time growing then their large cap counterparts.

- iShares Core S&P Small-Cap ETF (IJR) : This ETF is linked to an index which tracks the performance of small cap U.S. stocks. This fund will make for a good investment for traders looking for growth and are aware of the risks that come along with investing in a small cap ETF.

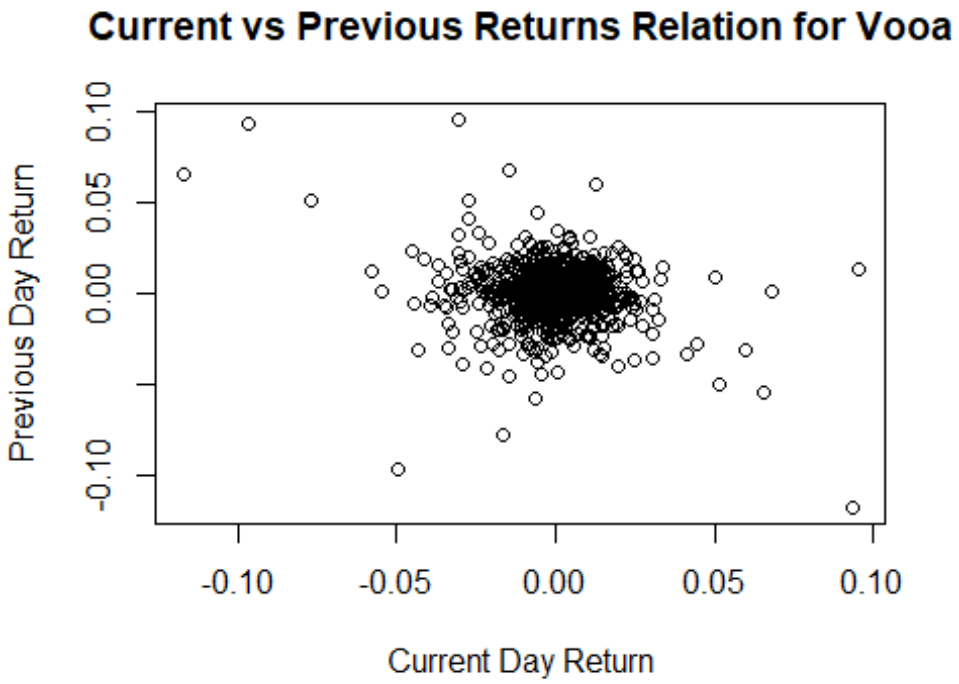
- SPDR S&P Dividend ETF (SDY) : This ETF is linked to the S&P High Yield Dividend Aristocrats Index, which offers exposure to dividend paying large-cap companies that exhibit value characteristics within the U.S. equity market.Only the highest yielding companies are chosen and these firms must have increased dividends every year for at least 25 consecutive years. Thanks to this focus, SDY only invests in

companies that are most likely to continue to pay out dividends in the future making it a solid pick for dividend focused investors even if the diversification is a little lacking.

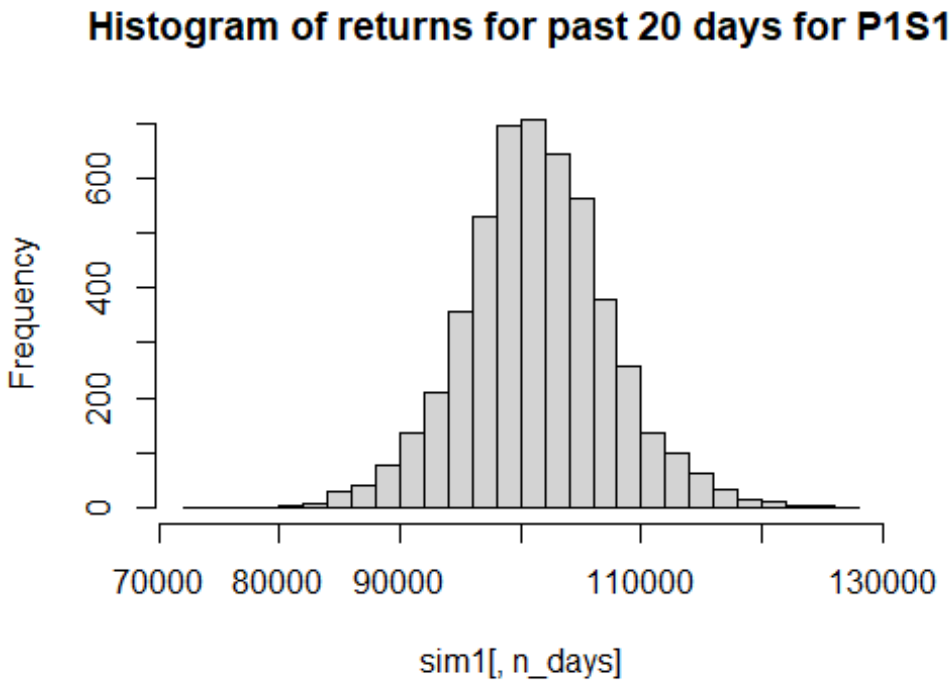
```
## [1] "QQQ" "IVV" "VOO" "IJH" "VO"  "VBR" "SDY" "IJR"
```



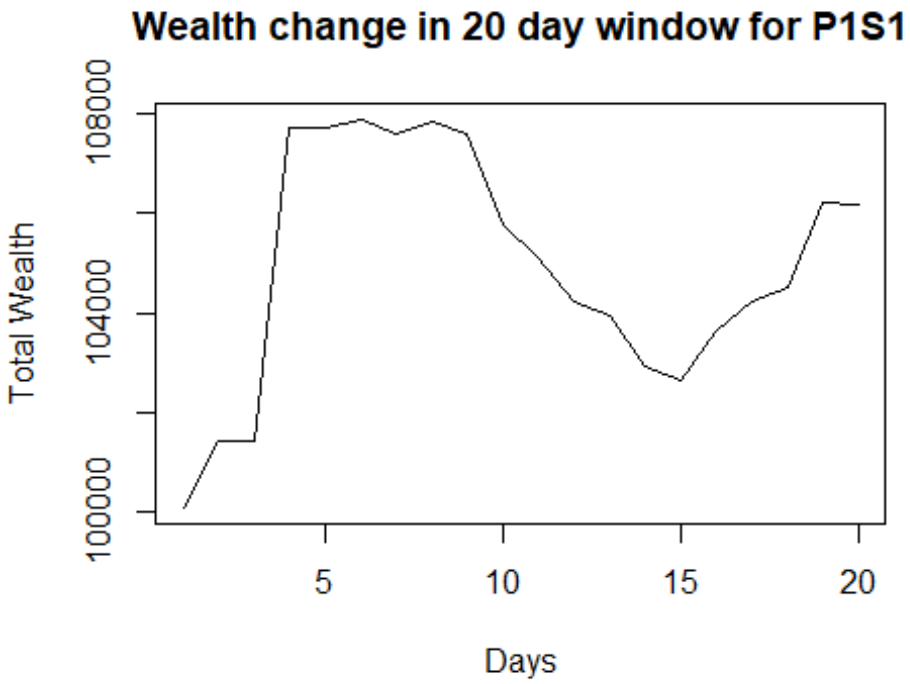
- Most of the ETFs seem to be correlated since most of them belong to the same sectors where Technology is the main contributor.
- Most ETFs are not closely correlated with IJR since IJR has more of a financial sector background and also is a small cap.



* The above plot shows that there is no correlation between current and yesterday's returns for a sample ETF which is on expected lines otherwise it would have been easy to predict and exploit the markets



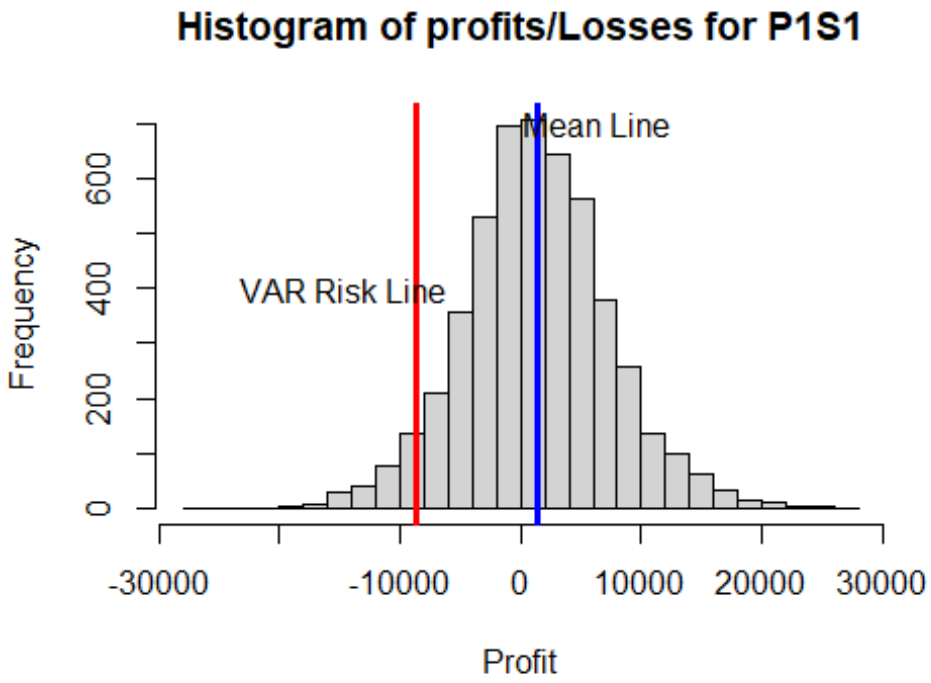
* The mean of the returns is plotted above and it can be seen that the mean is slightly beyond 100000 which was our initial amount



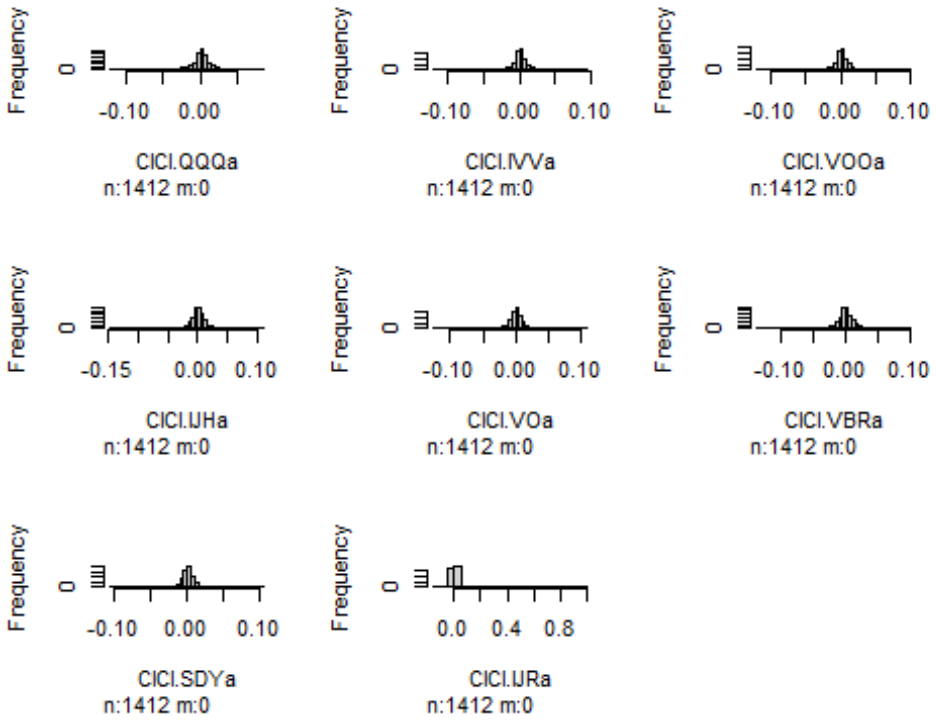
* The above plot shows the fluctuation in the wealth in the 20 day trading window. We see a small profit in all of the days and not a loss so that's a good measure of our portfolio.

```
## Mean of profits for P1S1is:

## [1] 1315.278

## Var for 5% return for Portfolio - 1 and scenario 1 is :

##        5%
## -8606.829

## $breaks
##  [1] -28000 -26000 -24000 -22000 -20000 -18000 -16000 -14000 -12000 -10000
## [11]  -8000  -6000  -4000  -2000      0   2000   4000   6000   8000  10000
## [21]  12000  14000  16000  18000  20000  22000  24000  26000  28000
##
## $counts
##  [1]   1   1   1   1   4   8  29  39  77 137 210 356 529 695 706 643 561 3
77 257
## [20] 136 100  63  34  15  11   5   3   1
##
## $density
##  [1] 1.00e-07 1.00e-07 1.00e-07 1.00e-07 4.00e-07 8.00e-07 2.90e-06 3.90e-
06
##  [9] 7.70e-06 1.37e-05 2.10e-05 3.56e-05 5.29e-05 6.95e-05 7.06e-05 6.43e-
05
## [17] 5.61e-05 3.77e-05 2.57e-05 1.36e-05 1.00e-05 6.30e-06 3.40e-06 1.50e-
```

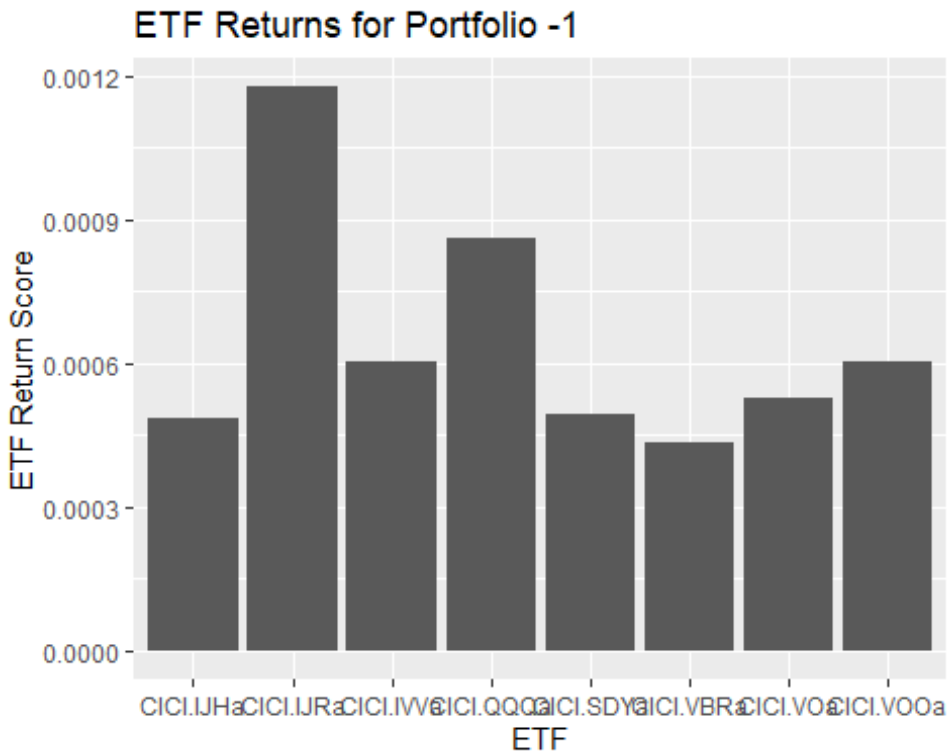
```
06
## [25] 1.10e-06 5.00e-07 3.00e-07 1.00e-07
##
## $mids
##  [1] -27000 -25000 -23000 -21000 -19000 -17000 -15000 -13000 -11000  -9000
## [11]  -7000  -5000  -3000  -1000   1000   3000   5000   7000   9000  11000
## [21]  13000  15000  17000  19000  21000  23000  25000  27000
##
## $xname
## [1] "profit_p1"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```



- There's a 0.05 probability that we will lose the mention VAR score over the course of this 20 day window for our portfolio.
- Let's calculate the mean returns for all the columns and try to assign more weight to the higher returnees and lower weights to the lower returnee ETF's
- Let's look at the histogram distribution of the ETF's returns.

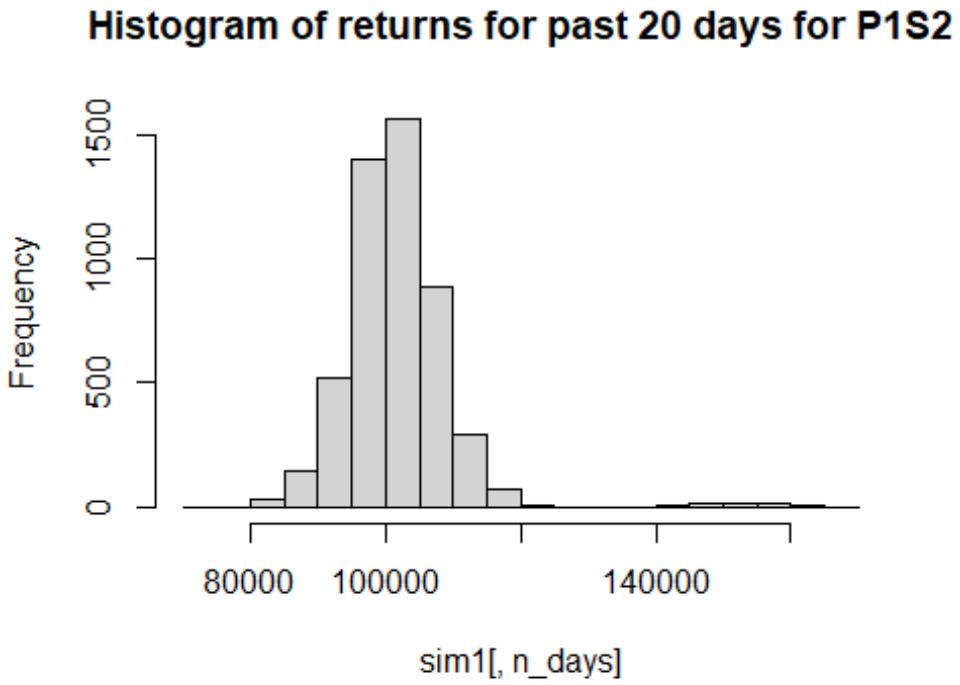


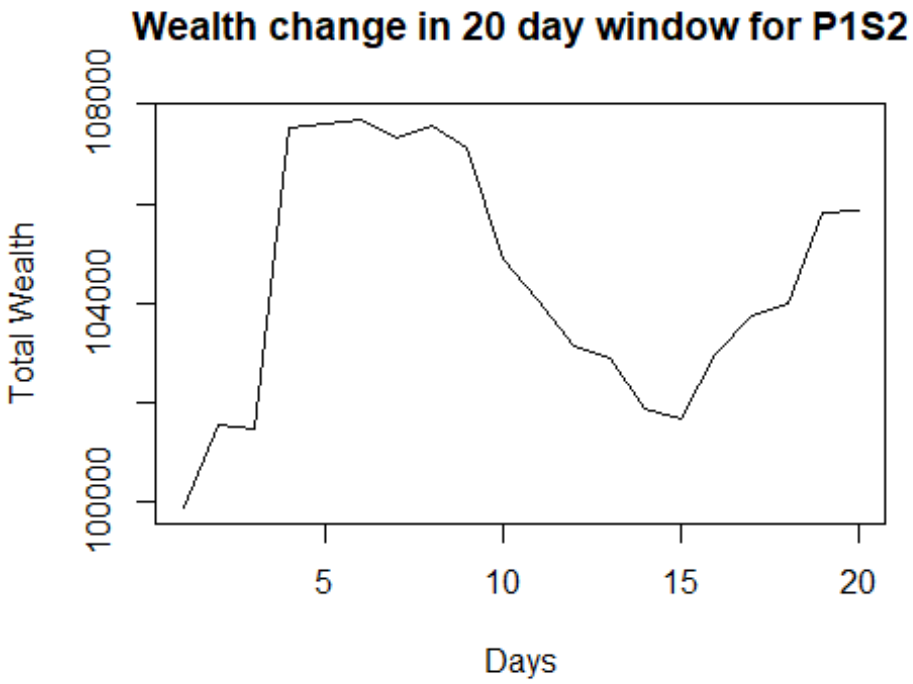
- IJR has more returns followed by QQQ so we will assign more weights to these 2 and assign the least to VBR and IJH



Scenario-2: Assigning more weight to highe return ETF

- Assign half weight to IJR and 0.2 to QQQ and least weight of 0.03 to VBR and IJH

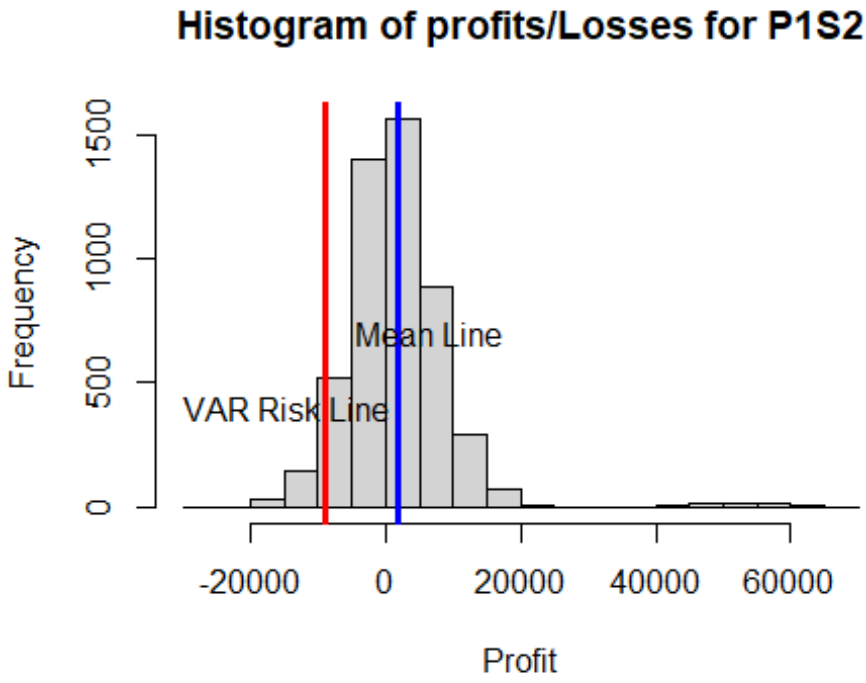




* Mean of profits is more when we assign more weight to IJH. This could be due to the presence of more financial sector stocks and we recently had a boom coming out of covid for these stocks. As a result the returns are more for those ETF

```
## Mean of profits for P1S2 is:

## [1] 1863.604

## Var for 5% return for Portfolio - 1 and scenario 2 is :

##        5%
## -8992.609

## $breaks
##  [1] -30000 -25000 -20000 -15000 -10000  -5000      0   5000  10000  15000
## [11]  20000  25000  30000  35000  40000  45000  50000  55000  60000  65000
## [21]  70000
##
## $counts
##  [1]    2    2   32  149  518 1403 1563  889  293   70    9    2    1    1
5
## [16]   18   17   18    5    3
##
## $density
##  [1] 8.000e-08 8.000e-08 1.280e-06 5.960e-06 2.072e-05 5.612e-05 6.252e-05
##  [8] 3.556e-05 1.172e-05 2.800e-06 3.600e-07 8.000e-08 4.000e-08 4.000e-08
## [15] 2.000e-07 7.200e-07 6.800e-07 7.200e-07 2.000e-07 1.200e-07
##
```

```
## $mids
##  [1] -27500 -22500 -17500 -12500  -7500  -2500   2500   7500  12500  17500
## [11]  22500  27500  32500  37500  42500  47500  52500  57500  62500  67500
##
## $xname
## [1] "profit_p1s2"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```



* Value of risk is more than the previous case since we bet more aggressively on one ETF.

Portfolio-2

Scenario-1 : Distribute all funds equally

- We will now create a portfolio with a mixture of Equity and Bond ETFs so that there is a bit more diversity in the asset classes.We will take 4 bond and 4 equity ETFs from the previous portfolio list
- Vanguard Total Bond Market ETF (BND) : This popular ETF offers exposure to entire investment grade bond market in a single ticker, with holdings in T-Bills, corporates, MBS, and agency bonds. This has the largest market capital and has a small expense ratio as well.

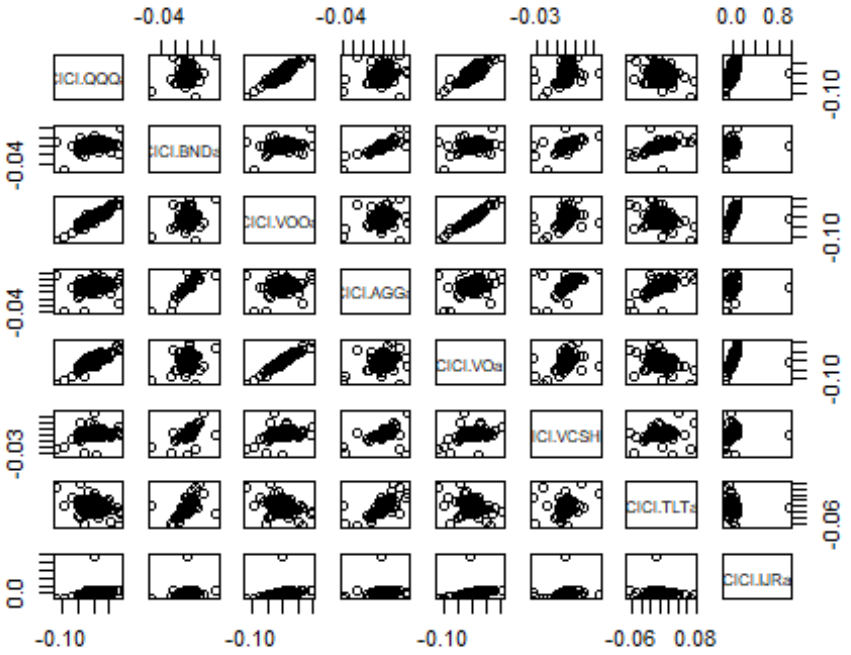
- iShares Core U.S. Aggregate Bond ETF (AGG) : The largest exchange-traded bond fund out there and one of the top 10 ETFs in the U.S. by assets, AGG boasts roughly $81 billion under management and is the simplest way to gain exposure to fixed-income markets. It is made up of more than 10,000 individual bond holdings to represent broad exposure to U.S. investment-grade bonds, including about 40% of its portfolio in U.S. Treasury bonds. The rest includes top-tier corporate bonds from firms like JPMorgan Chase & Co. (JPM) as well as mortgage-related debt. With loe expense ratio, AGG is a very affordable one-stop shop for bond exposure.

- Vanguard Short-Term Corporate Bond ETF (VCSH) : It is a $41 billion fund that focuses on high-quality corporate debt but with the typical bond in its portfolio maturing in just 2.8 years. That means investors can have a lot more certainty that those debts will be repaid in full, since it's a smaller window of time for unexpected disruptions to upend operations at these firms. The yield is a bit less than the longer-dated VCIT but is still very attractive when compared with the typical S&P 500 dividend stock – and offering a lot less risk, which is perhaps a big selling point all by itself in the current environment.

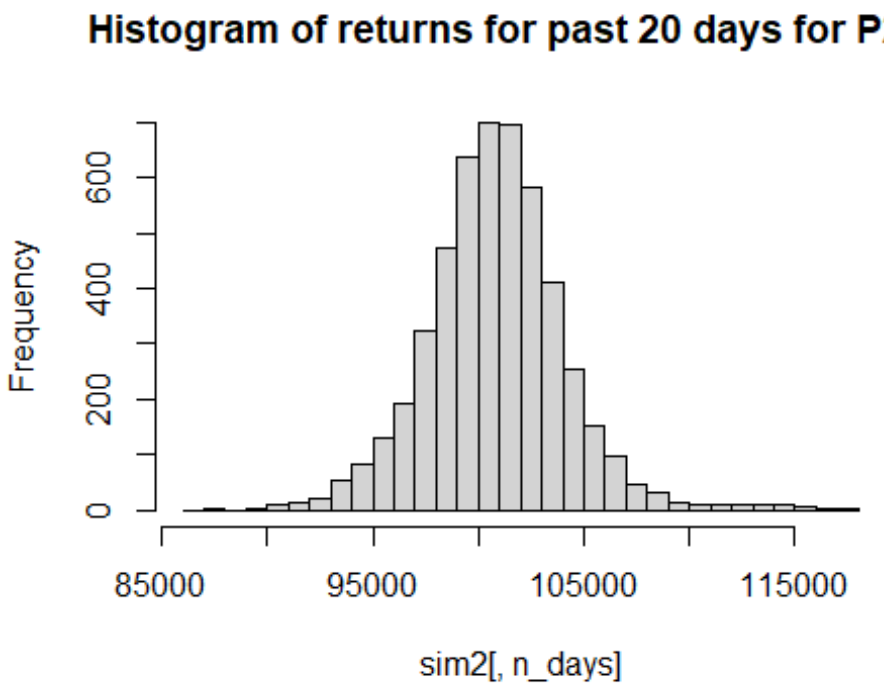
- iShares 20+ Year Treasury Bond ETF (TLT) : If you don't want to diversify into corporate debt and instead want the rock-solid comfort of the U.S. Treasury alone, perhaps the most popular low-risk way to play the bond market is TLT.The duration of this bond ETF's holdings are all 20 years or longer, which does provide some long-term interest rate risk; the fund is actually down 20% in the last 12 months as rates have moved higher and devalued its older positions. However, with a yield that is now roughly 60% higher than what it was just a year ago, it might be time to consider carving out a position once more in this $19 billion low-risk bond fund.

We will choose IJR, QQQ, VOO and VO from the above equity ETF's

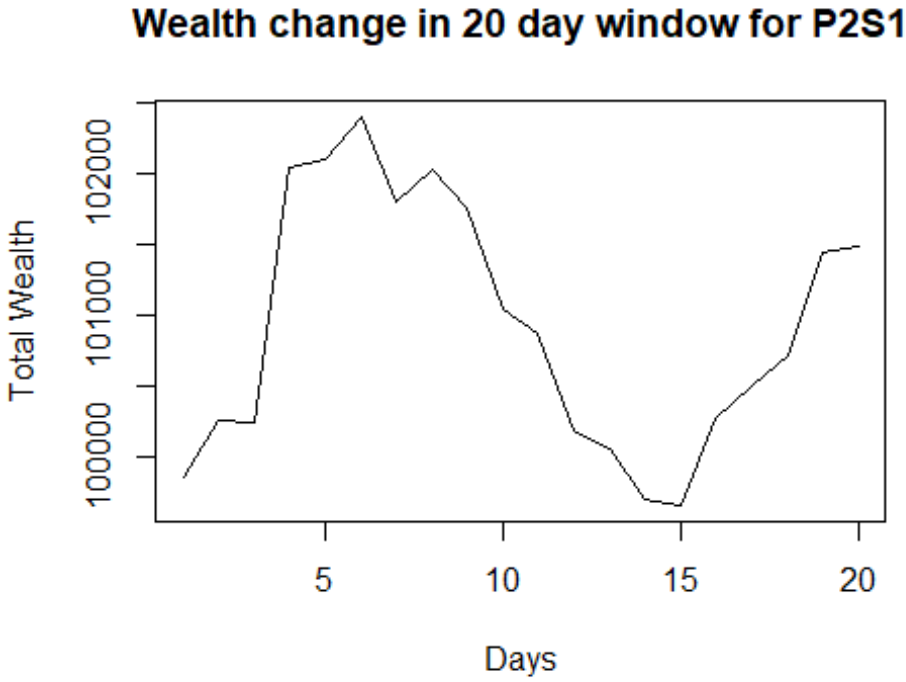
```
## [1] "QQQ"  "BND"  "VOO"  "AGG"  "VO"   "VCSH" "TLT"  "IJR"
```



- The Equity ETFs are not correlated with the bond ETFs as seen in the graph since its a different asset class itself.
- The Bond ETFs are not that correlated with themselves as well since bonds are individual stocks with no sector as such and as a result the correlation is not prevalent



* The returns are in the slightly positive side with normal distribution more towards the right side



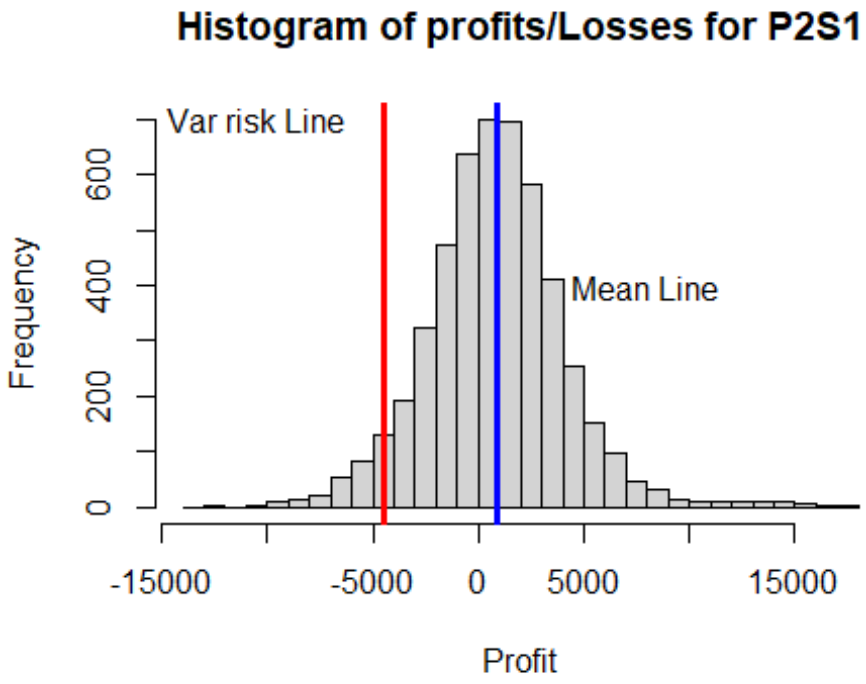
* Max returns were reached around day 18

```
## Mean of profits for P2S1 is:
```

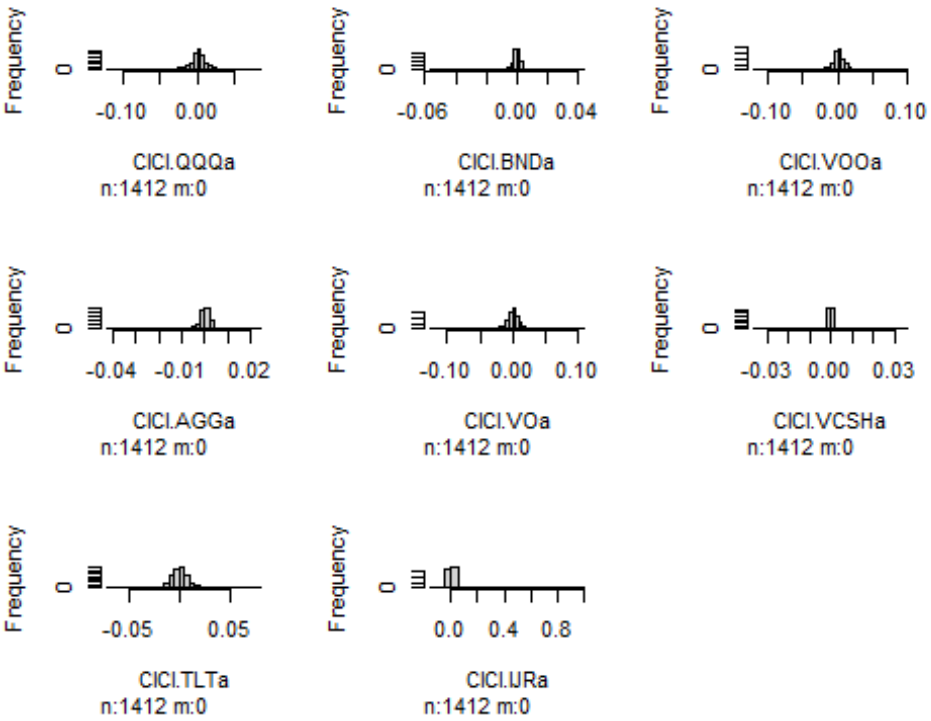
```
## [1] 851.6779

## Var for 5% return for Portfolio - 2 and scenario 1 is :

##        5%
## -4438.785
```

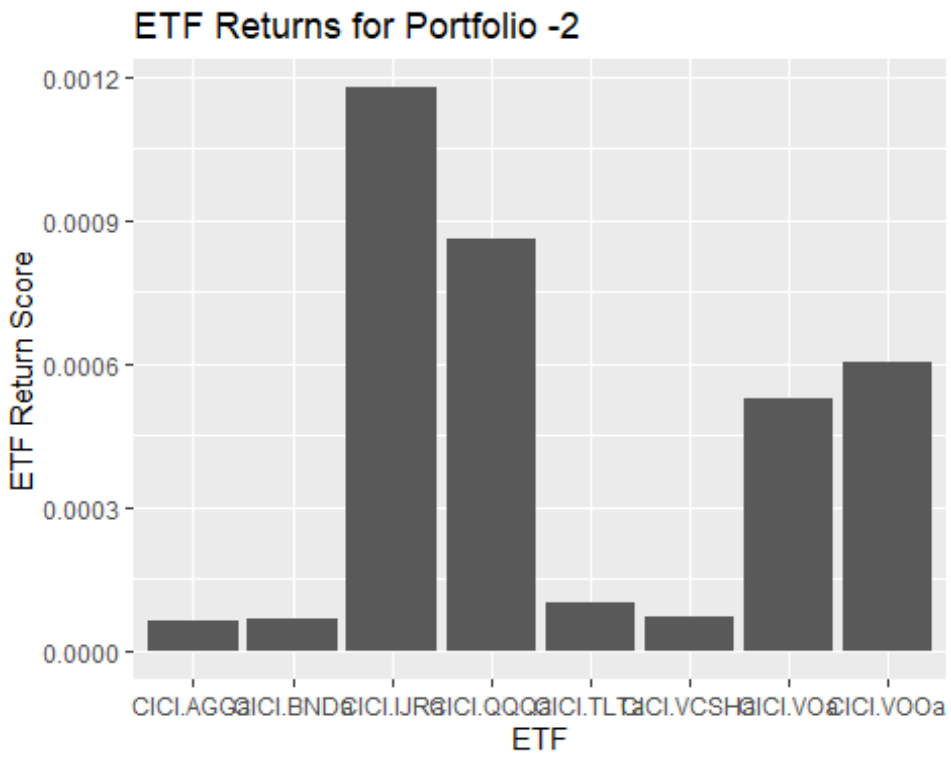
- There is lesser value at risk and lesser mean returns than portfolio 1. This is because of the presence of both equity and bond ETFs The higher volatility of stocks relative to bonds is due to the nature of the two types of investments. When you buy stocks, you're buying ownership in companies (albeit a small share). When you buy bonds, you're lending money, either to companies or to governments. Because creditors are paid before owners, it's riskier to own a company than it is to lend money, so the prices of stocks are more sensitive to changes in the economy.Thus bonds have lesser var but lesser returns

```
## $breaks
##  [1] -14000 -13000 -12000 -11000 -10000  -9000  -8000  -7000  -6000  -5000
## [11]  -4000  -3000  -2000  -1000      0   1000   2000   3000   4000   5000
## [21]   6000   7000   8000   9000  10000  11000  12000  13000  14000  15000
## [31]  16000  17000  18000
##
## $counts
##  [1]   1   2   1   2   9  13  21  54  82 132 194 324 474 635 699 694 583 4
10 254
## [20] 154  97  48  32  16  10   9  11  12  10   8   5   4
##
## $density
##  [1] 0.0000002 0.0000004 0.0000002 0.0000004 0.0000018 0.0000026 0.0000042
##  [8] 0.0000108 0.0000164 0.0000264 0.0000388 0.0000648 0.0000948 0.0001270
## [15] 0.0001398 0.0001388 0.0001166 0.0000820 0.0000508 0.0000308 0.0000194
## [22] 0.0000096 0.0000064 0.0000032 0.0000020 0.0000018 0.0000022 0.0000024
## [29] 0.0000020 0.0000016 0.0000010 0.0000008
##
## $mids
##  [1] -13500 -12500 -11500 -10500  -9500  -8500  -7500  -6500  -5500  -4500
## [11]  -3500  -2500  -1500   -500    500   1500   2500   3500   4500   5500
## [21]   6500   7500   8500   9500  10500  11500  12500  13500  14500  15500
## [31]  16500  17500
##
## $xname
## [1] "profit_p2"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

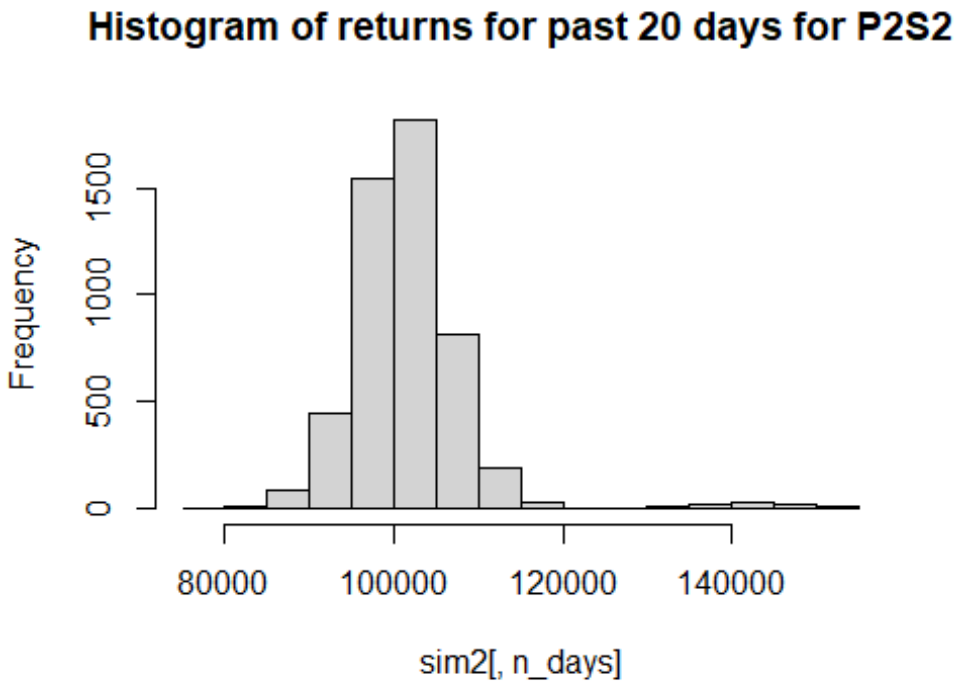


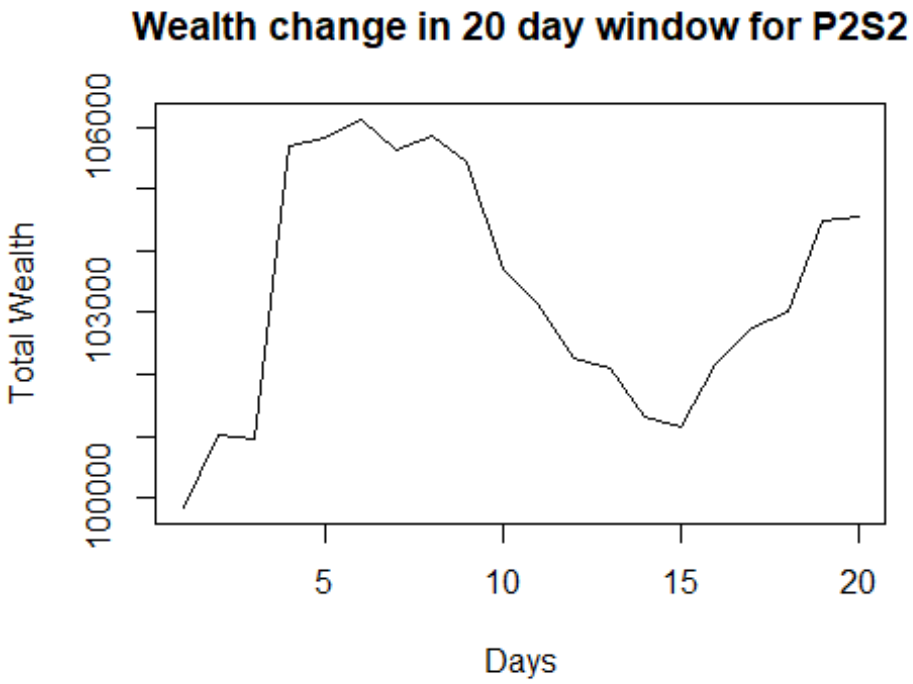
Scenario-2: Assign more weights to high return ETFs





* IJR, QQQ, VOO are the better ETFs so we will assign more weights to them and lesser weight to BND and Agg



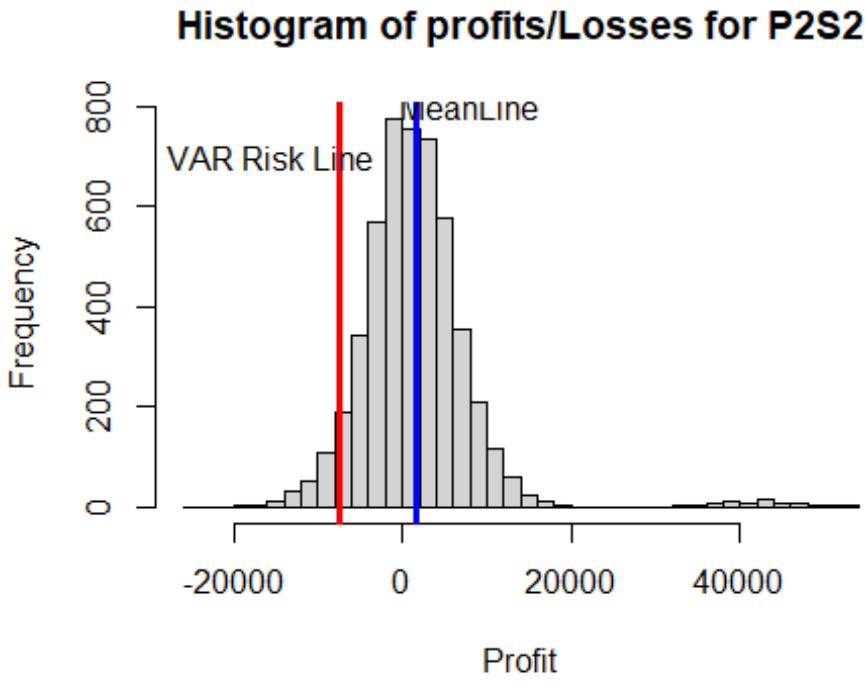



```
## Mean of profits for P2S2 is:

## [1] 1608.403

## Var for 5% return for Portfolio - 2 and scenario 2 is :

##        5%
## -7525.206
```

- Theres not much of a difference in mean return and var but scenario 2 tends to perform better than scenario - 1.

```
## $breaks
##  [1] -26000 -24000 -22000 -20000 -18000 -16000 -14000 -12000 -10000  -8000
## [11]  -6000  -4000  -2000      0   2000   4000   6000   8000  10000  12000
## [21]  14000  16000  18000  20000  22000  24000  26000  28000  30000  32000
## [31]  34000  36000  38000  40000  42000  44000  46000  48000  50000  52000
## [41]  54000
##
## $counts
##  [1]   1   0   1   2   3  11  30  51 108 188 344 570 776 754 734 577 356 2
10 115
## [20]  62  23  11   3   1   1   1   0   0   0   4   2   7  10   7  14   8
6   5
## [39]   2   2
##
## $density
##  [1] 1.00e-07 0.00e+00 1.00e-07 2.00e-07 3.00e-07 1.10e-06 3.00e-06 5.10e-
```

```
06
##  [9] 1.08e-05 1.88e-05 3.44e-05 5.70e-05 7.76e-05 7.54e-05 7.34e-05 5.77e-
05
## [17] 3.56e-05 2.10e-05 1.15e-05 6.20e-06 2.30e-06 1.10e-06 3.00e-07 1.00e-
07
## [25] 1.00e-07 1.00e-07 0.00e+00 0.00e+00 0.00e+00 4.00e-07 2.00e-07 7.00e-
07
## [33] 1.00e-06 7.00e-07 1.40e-06 8.00e-07 6.00e-07 5.00e-07 2.00e-07 2.00e-
07
##
## $mids
##  [1] -25000 -23000 -21000 -19000 -17000 -15000 -13000 -11000  -9000  -7000
## [11]  -5000  -3000  -1000   1000   3000   5000   7000   9000  11000  13000
## [21]  15000  17000  19000  21000  23000  25000  27000  29000  31000  33000
## [31]  35000  37000  39000  41000  43000  45000  47000  49000  51000  53000
##
## $xname
## [1] "profit_p2s2"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```



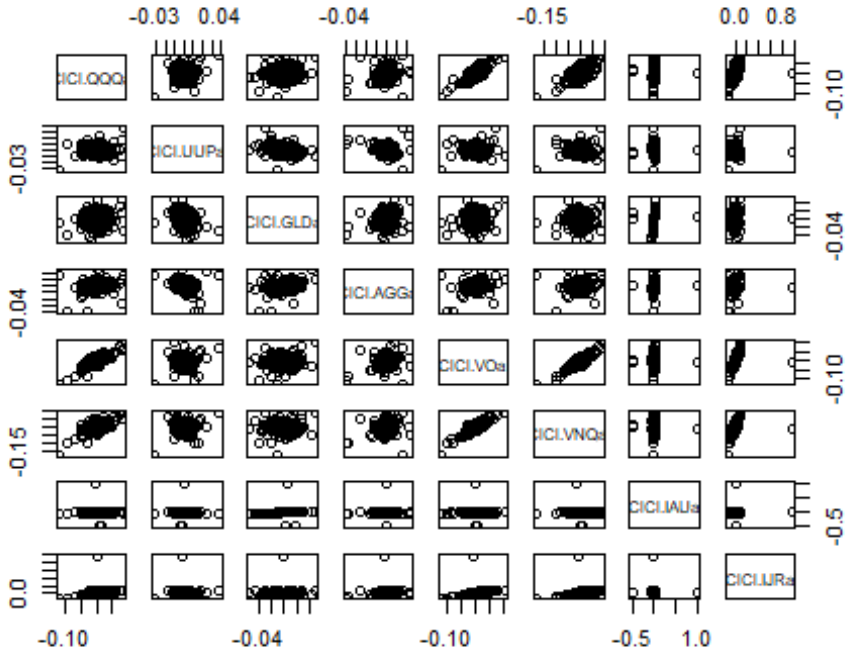
* The returns have increased but the var has gone down and the returns are still less than portfolio - 1

Portfolio-3

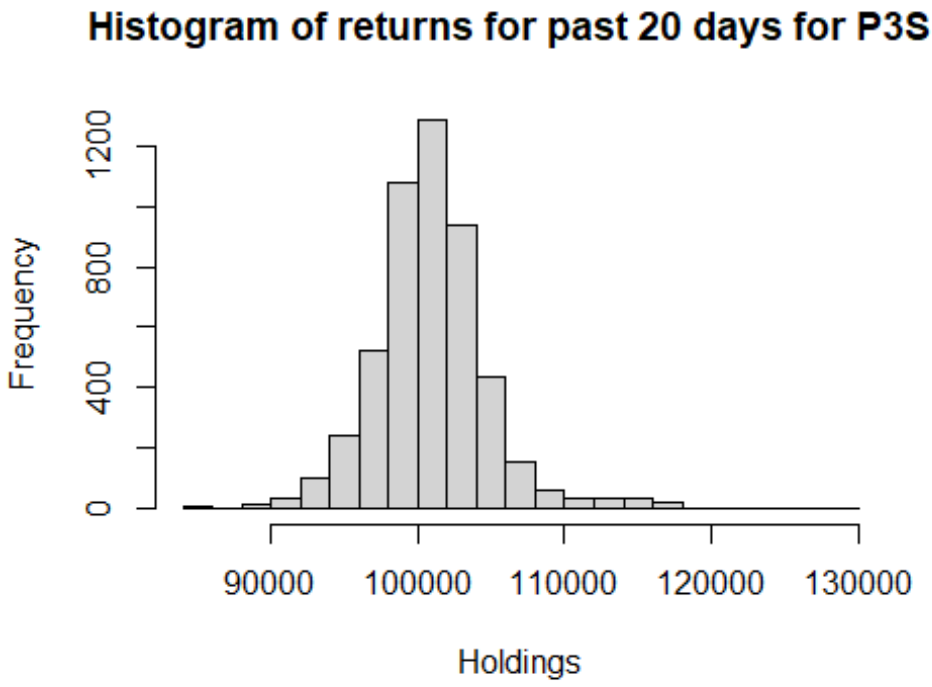
- We will now go for mix of equity, bonds, commodity,real estate and some currency ETF's
- Invesco DB US Dollar Index Bullish Fund (UUP) : This ETF offers exposure to a basket of currencies relative to the U.S. dollar, decreasing in value when the trade-weighted basket strengthens and increasing when the dollar appreciates. It is appropriate for investors seeking to a fund that is inversely correlated to the broad stock market or for those making a bet on a flight to quality so the aim is that it will help offset some losses
- SPDR Gold Shares (GLD) : GLD is one of the most popular ETFs in the world, offering exposure to an asset class that has become increasingly important to the asset allocation process in recent years. GLD can be used in a number of different ways; some may establish short term positions as a way of hedging against equity market volatility, dollar weakness, or inflation. Others may wish to include gold exposure as part of a long-term investment strategy. GLD is a relatively straightforward product; the underlying assets consist of gold bullion stored in secure vaults. As such, the price of this ETF can be expected to move in lock step with spot gold prices.
- Vanguard Real Estate ETF (VNQ) : The Vanguard Real Estate Trust (VNQ) offers broad exposure to U.S. equity REITs, alongside a small allocation to specialized REITs and real estate firms. Real estate has historically been embraced because of its ability to deliver excess returns during bull markets and for its low correlation with traditional stock and bond investments. REITs might appeal to investors seeking current income, as these trusts must distribute at least 90% of their income to investors. The fund offers an efficient way for investors to gain indirect exposure to real estate prices
- iShares Gold Trust (IAU) : This fund offers exposure to one of the world's most famous metals, gold. IAU is designed to track the spot price of gold bullion by holding gold bars in a secure vault, allowing investors to free themselves from finding a place to store the metal. While IAU isn't the most liquid way to gain exposure to gold, it does have among the lowest expense ratios, making it a solid choice for cost-conscious investors.

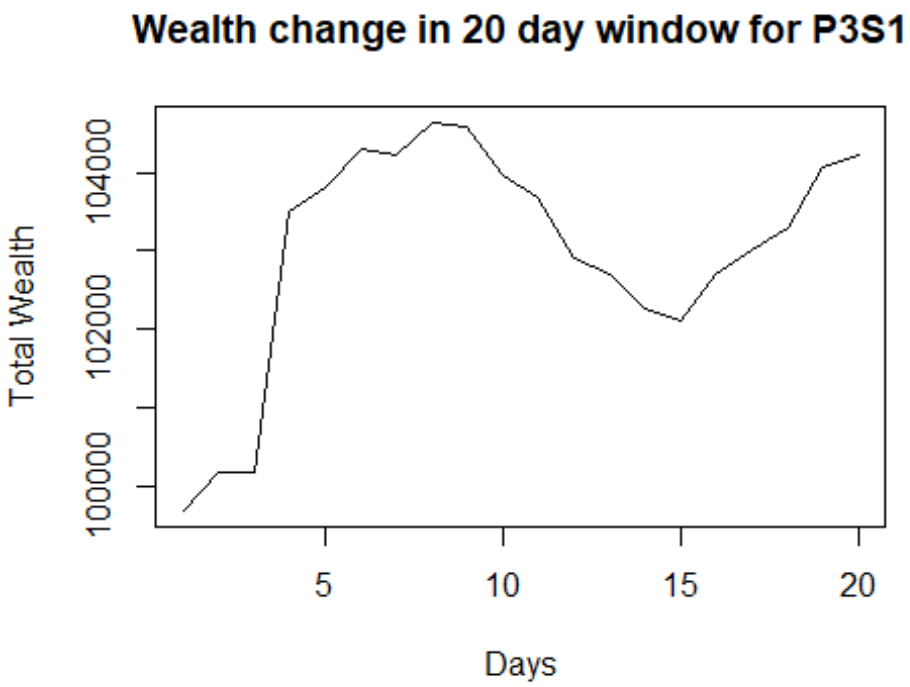
We will take 2 equity (IJR,QQQ) and 2 bond ETFS (TLT,Agg) from the above portfolios

```
## [1] "QQQ" "UUP" "GLD" "AGG" "VO"  "VNQ" "IAU" "IJR"
```



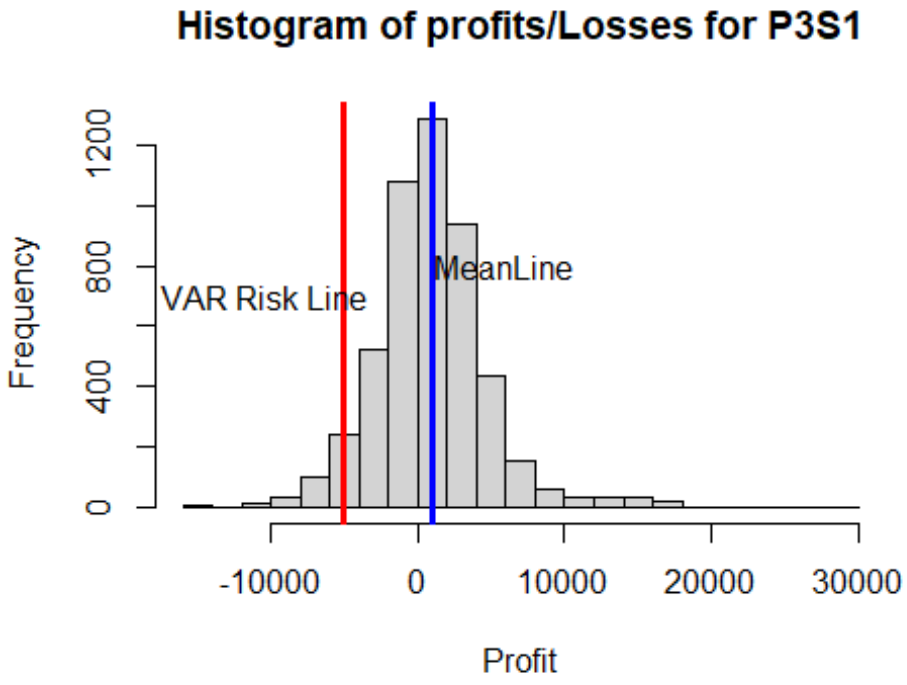
- Due to the different asset classes, there is not much correlation between the ETF's





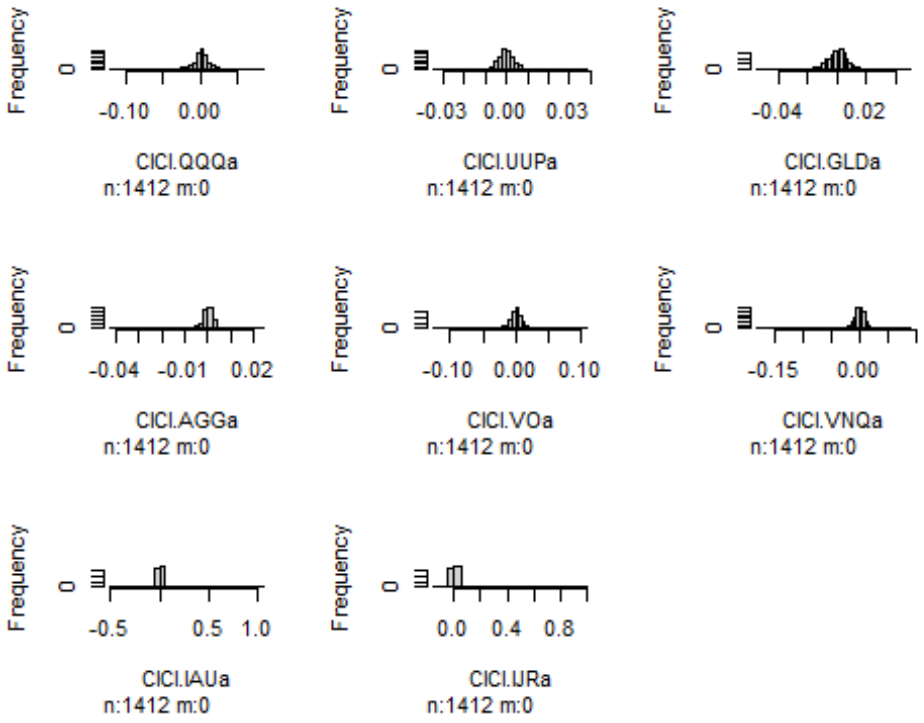

```
## Mean of profits for P3S1 is:

## [1] 926.3124

## Var for 5% return for Portfolio - 3 and scenario 1 is :

##        5%
## -5123.334

## $breaks
##  [1] -16000 -14000 -12000 -10000  -8000  -6000  -4000  -2000      0   2000
## [11]   4000   6000   8000  10000  12000  14000  16000  18000  20000  22000
## [21]  24000  26000  28000  30000
##
## $counts
##  [1]    3    2   16   35  103  240  519 1077 1287  935  435  155   63   35
36
## [16]   34   19    2    0    0    1    1    2
##
## $density
##  [1] 0.0000003 0.0000002 0.0000016 0.0000035 0.0000103 0.0000240 0.0000519
##  [8] 0.0001077 0.0001287 0.0000935 0.0000435 0.0000155 0.0000063 0.0000035
## [15] 0.0000036 0.0000034 0.0000019 0.0000002 0.0000000 0.0000000 0.0000001
## [22] 0.0000001 0.0000002
##
## $mids
##  [1] -15000 -13000 -11000  -9000  -7000  -5000  -3000  -1000   1000   3000
```

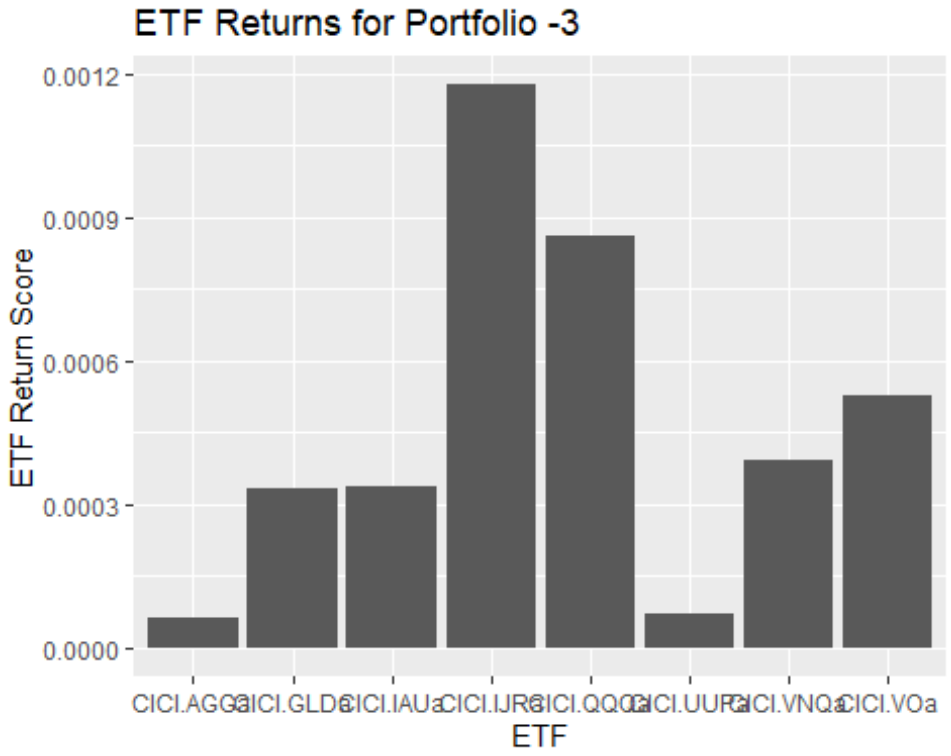
```
## [11]   5000   7000   9000  11000  13000  15000  17000  19000  21000  23000
## [21]  25000  27000  29000
##
## $xname
## [1] "profit_p3s1"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```



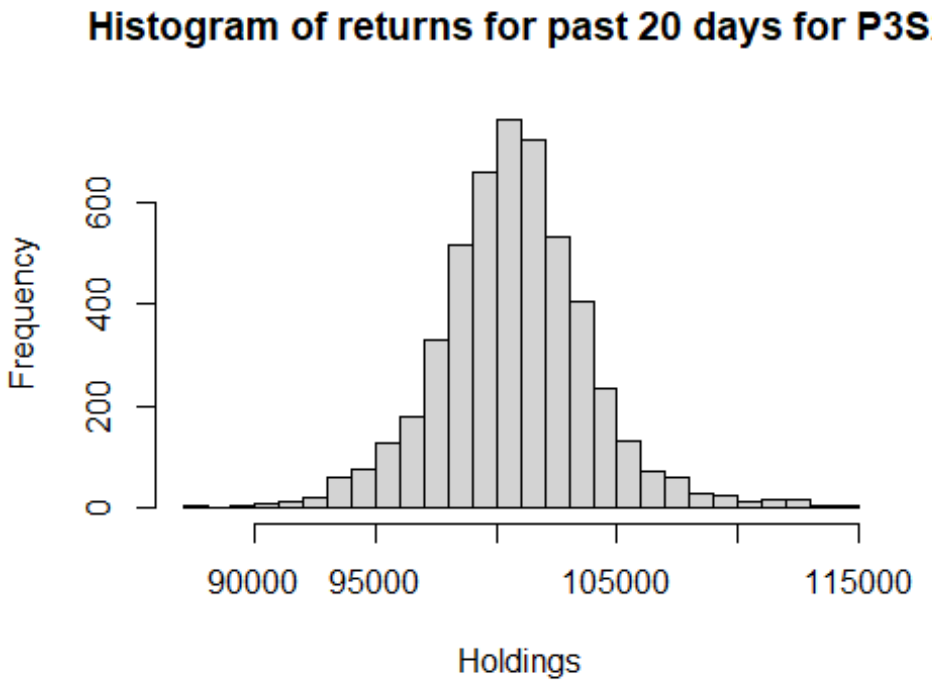
- The returns are slightly better than portfolio-2 but still less than portfolio - 1 since more returns are obtained for equity ETF's

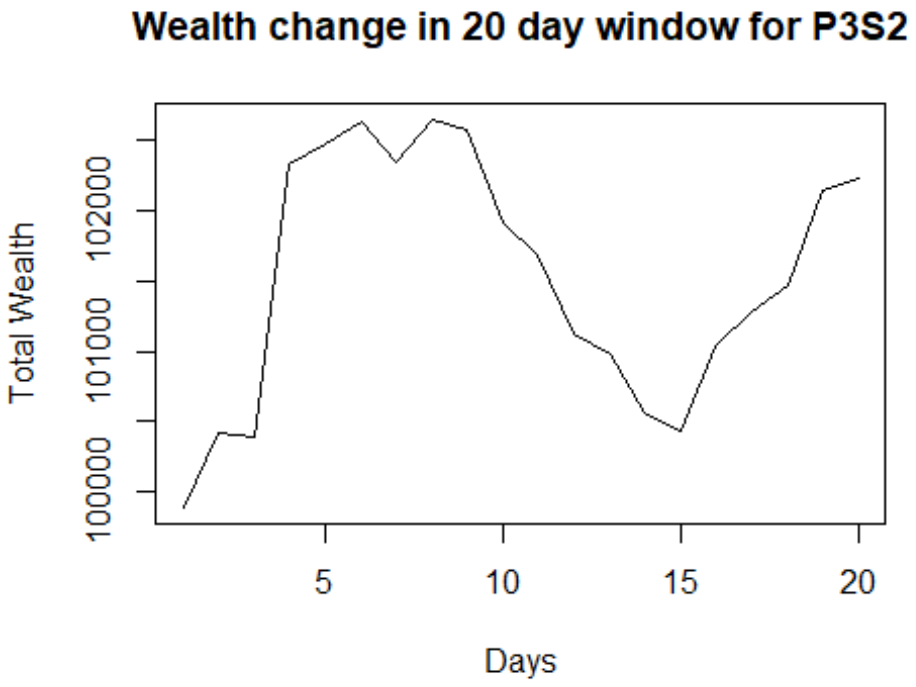
Scenario-2 : Assign weights to high performing ETF





* We will assign more weight to IJR, QQQ and Voa and less weight to Agg and UUP



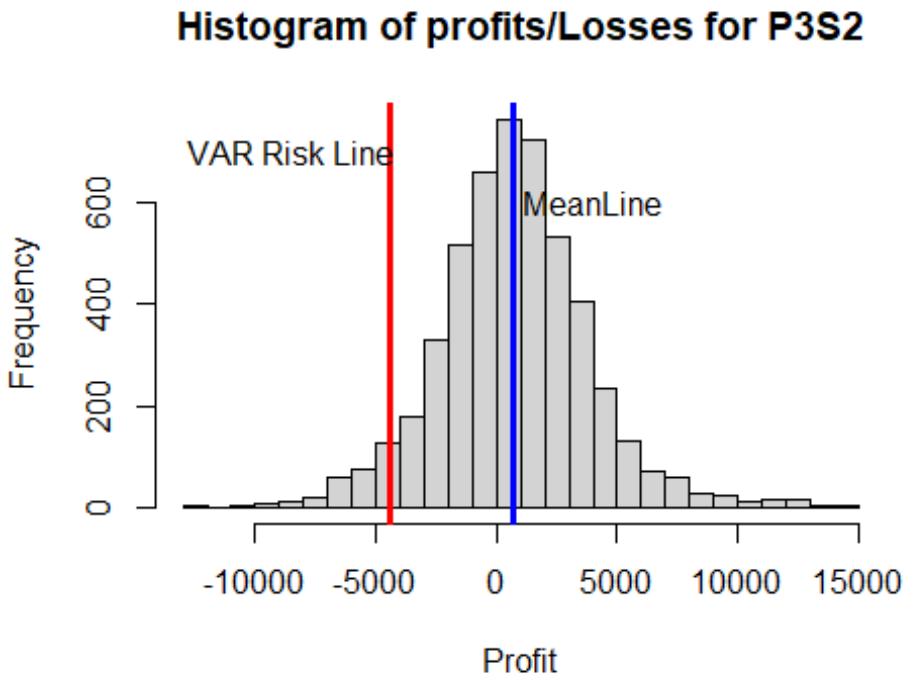


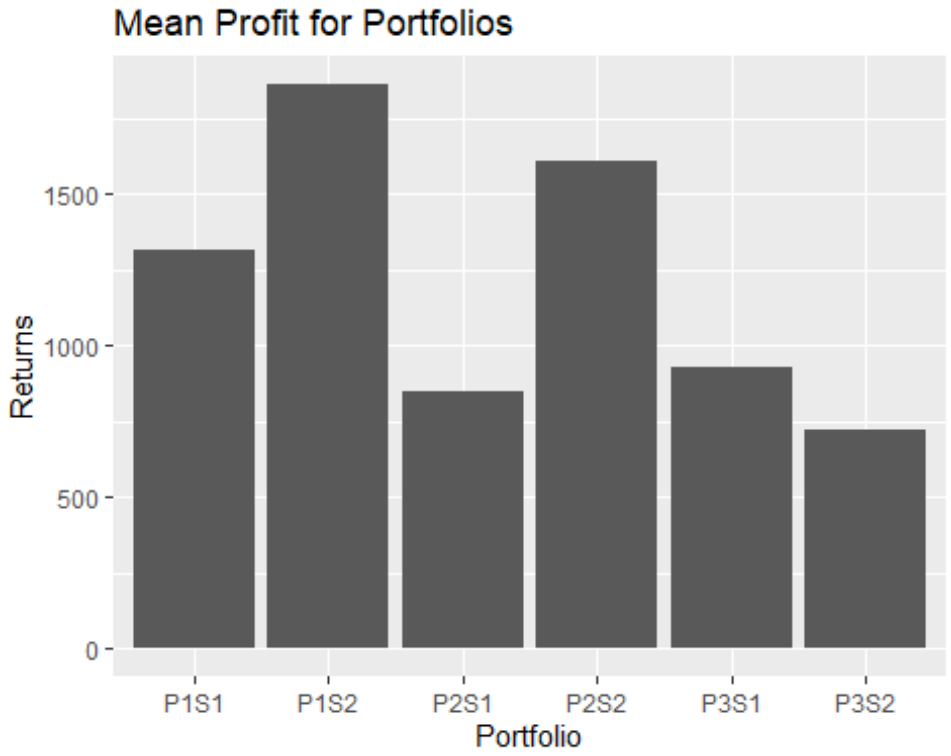

```
## Mean of profits for P3S2 is:

## [1] 724.816
```

```
## Var for 5% return for Portfolio - 3 and scenario 2 is :

##        5%
## -4412.074
```

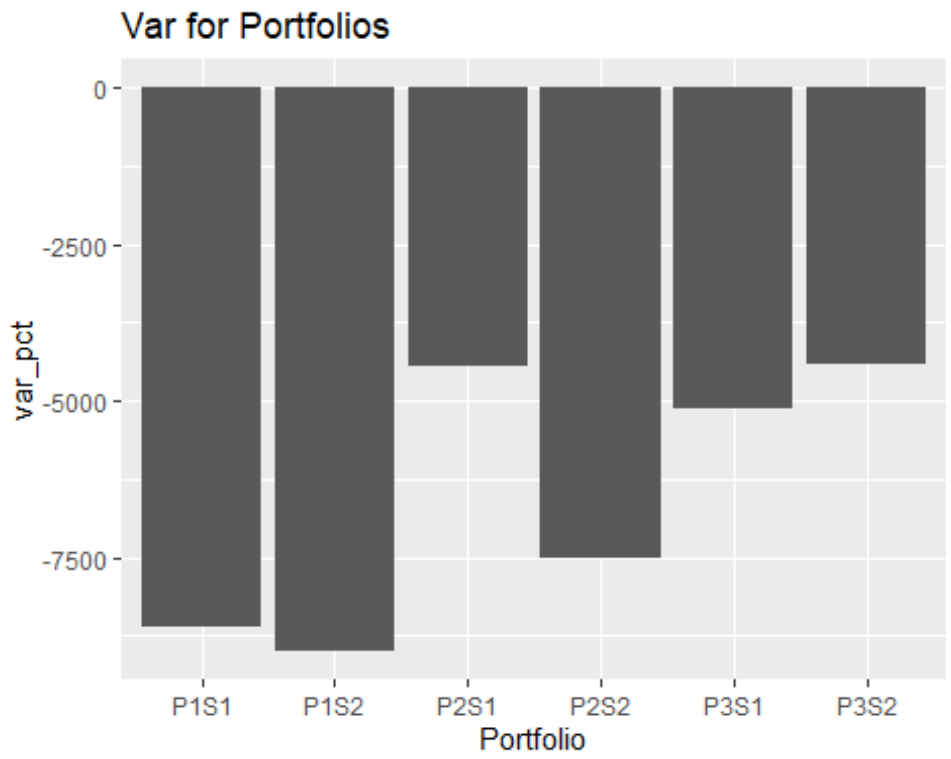
- The mean return is still less even though Var is lesser compared to Portfolio-1. This is due to the fact that equity ETFs can be more profitable due to their volatile nature and the increased market capitalization

```
## $breaks
##  [1] -13000 -12000 -11000 -10000  -9000  -8000  -7000  -6000  -5000  -4000
## [11]  -3000  -2000  -1000      0   1000   2000   3000   4000   5000   6000
## [21]   7000   8000   9000  10000  11000  12000  13000  14000  15000
##
## $counts
##  [1]   2   1   4   6  12  20  59  74 126 177 329 517 658 763 723 533 403 2
33 129
## [20]  70  59  27  25  12  16  14   4   4
##
## $density
##  [1] 0.0000004 0.0000002 0.0000008 0.0000012 0.0000024 0.0000040 0.0000118
##  [8] 0.0000148 0.0000252 0.0000354 0.0000658 0.0001034 0.0001316 0.0001526
## [15] 0.0001446 0.0001066 0.0000806 0.0000466 0.0000258 0.0000140 0.0000118
## [22] 0.0000054 0.0000050 0.0000024 0.0000032 0.0000028 0.0000008 0.0000008
##
## $mids
##  [1] -12500 -11500 -10500  -9500  -8500  -7500  -6500  -5500  -4500  -3500
## [11]  -2500  -1500   -500    500   1500   2500   3500   4500   5500   6500
## [21]   7500   8500   9500  10500  11500  12500  13500  14500
##
## $xname
## [1] "profit_p3s2"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```





* To summarize the mean returns for Portfolio-1, scenario - 2 are desirable when we assign more weight to IJR and QQQ ETFs. Portfolio - 1 gives more returns since equity ETFs give higher returns compared to other asset class as there are higher risk involved as seen in the below graph.



* Because bond ETFs never mature, they never offer the same protection for your initial investment the way that individual bonds can. In other words, you aren't guaranteed to get your money back at some point in the future. You can lose money if interest rates rise. Interest rates change over time. With so many incidents happening in the past 5 years like Covid, Russia-Ukraine War, inflation, it makes sense that with rising interest rates bonds give lesser returns as compared to ETFs.

As they say, High Risk means High Rewards

Q6: Clustering & PCA

Let's load in the data and take a look at it.

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

```
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality color
## 1       5   red
## 2       5   red
## 3       5   red
## 4       6   red
## 5       5   red
## 6       5   red
```

This data set has 13 columns about different bottles of wine. The first 11 columns describe the chemical properties of the wine and the last two columns contain information about the color of the wine & its quality (rated by experts).

Our goal is to run unsupervised learning algorithms on the 11 chemical properties of the wines. Through the results of the unsupervised techniques, we will check if differences between the labels (red/white & quality) emerge naturally.

First, let's run PCA (Principal Component Analysis)

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC
7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.7233
0
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.0475
6
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.9000
9
##                            PC8     PC9   PC10    PC11
## Standard deviation     0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.94568 0.97632 0.9970 1.00000
```

The third row of the summary table shows what is the cumulative proportion of variance that is captured through the corresponding principal components. In our case, we see that two principal components capture ~50% of the variance in the data set. So, let's take the number of principal components as two as this also helps us visualize the data easily.

Next, we'll interpret the loadings for each principal component, i.e., see how much weightage each component has given to our original 11 features.

```
##              Properties   PC1
## 1  total.sulfur.dioxide  0.49
## 2   free.sulfur.dioxide  0.43
## 3        residual.sugar  0.35
## 4           citric.acid  0.15
## 5               density -0.04
## 6               alcohol -0.11
## 7                    pH -0.22
## 8         fixed.acidity -0.24
## 9             chlorides -0.29
```

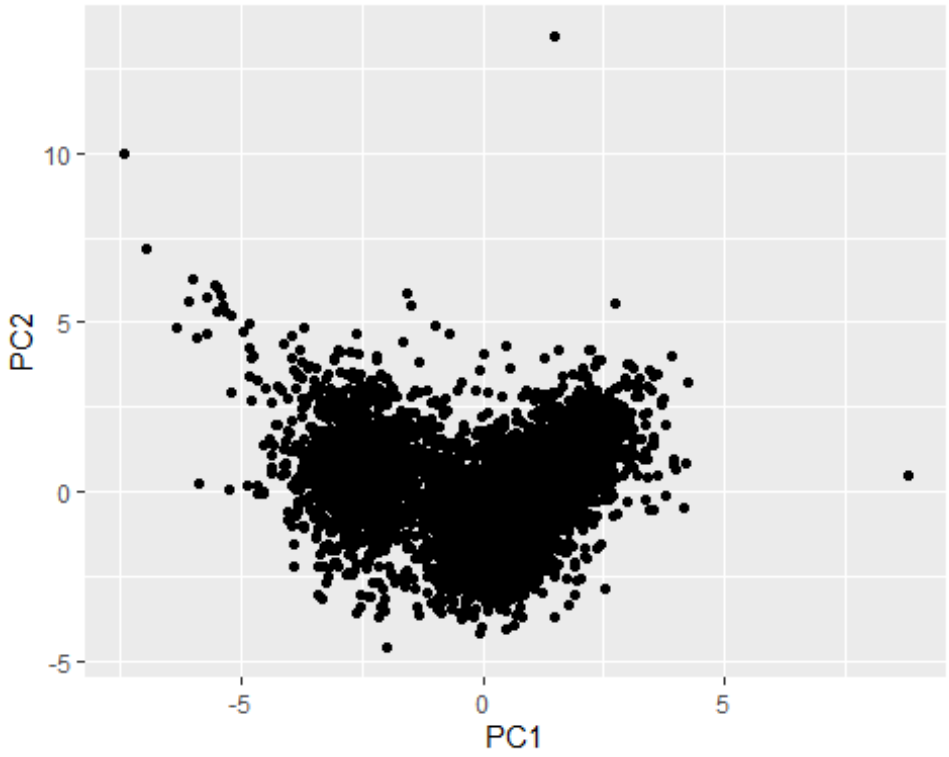
```
## 10            sulphates -0.29
## 11     volatile.acidity -0.38
```

We can see that PC1 seems to give more positive loadings to sulfur dioxide and residual sugar while chlorides, sulphates & acidity have negative loadings.

```
##              Properties   PC2
## 1               density  0.58
## 2         fixed.acidity  0.34
## 3        residual.sugar  0.33
## 4             chlorides  0.32
## 5             sulphates  0.19
## 6           citric.acid  0.18
## 7      volatile.acidity  0.12
## 8  total.sulfur.dioxide  0.09
## 9   free.sulfur.dioxide  0.07
## 10                  pH -0.16
## 11             alcohol -0.47
```

PC2 gives more positive loading to density and high negative to alcohol content.

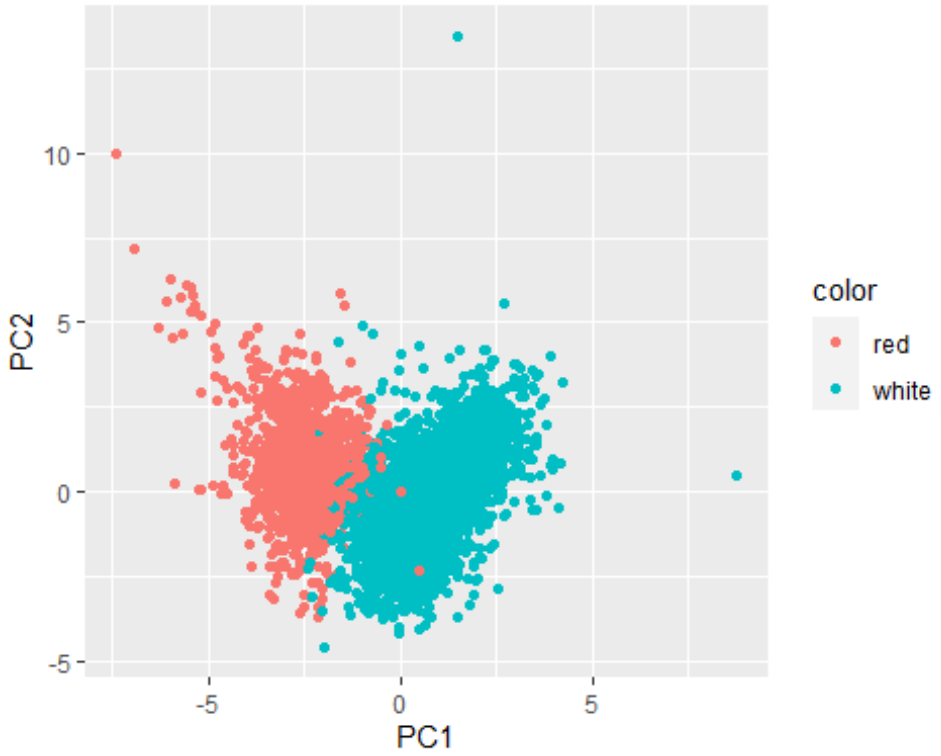
Now, let's plot our original data using these two principal components as the two axes and see if any pattern emerges.



We can see that there are two clusters emerging. One to the left having more negative PC1 & another to the right having more positive PC1. There doesn't appear to be much difference along PC2. Using the interpretation of the feature loadings of PC1, we can see

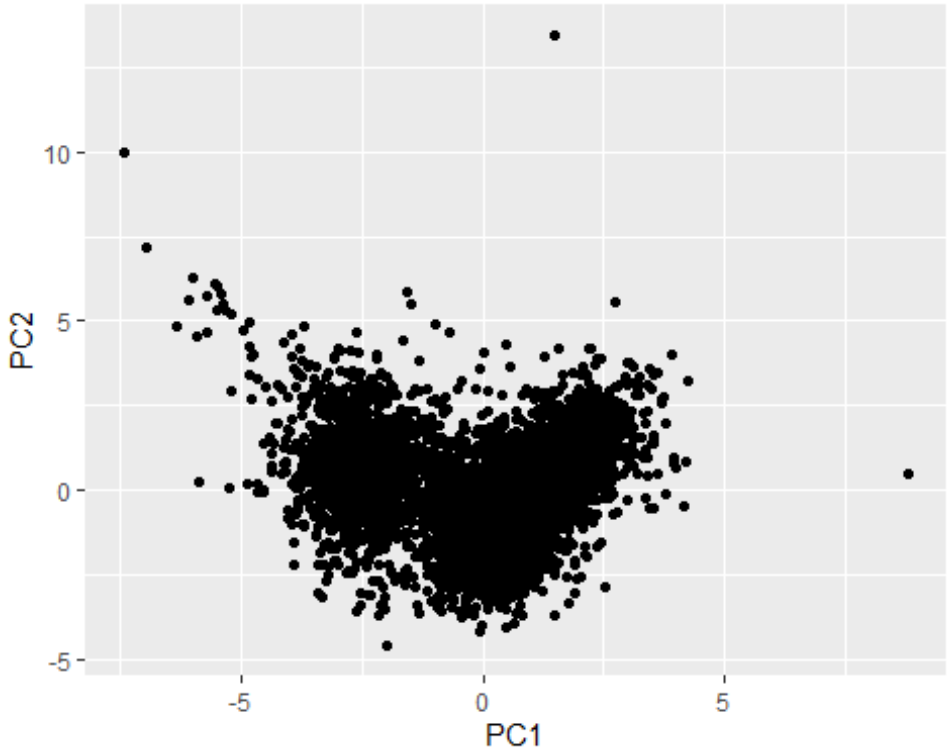
that wines having a high content of sulfur dioxides and residual sugars would have positive values while wines with high content of chlorides, sulphates and acidity would take negative values. Hence, we can say that these are two distinct clusters. However, we cannot say for certain which of these two clusters correspond to red & White wine without using the label information.

In the real world, we usually do not have access to this information. Nevertheless, for the purposes of validation, we can use the wine color to confirm if these two clusters do correspond to different types of wines.



We can clearly see that the principal components have clearly separated red & white wines. Hence, we can say that PCA is capable of distinguishing between red & white wines.

Next, we need to check if this model can distinguish between higher & lower quality wines. We continue to look at the original scatterplot using the principal components.



While we see two distinct clusters, it becomes harder to argue that these may correspond to higher & lower quality wines. This is because quality is a subjective metric. The data set describes quality as a rating on a 1-10 scale by a panel of wine experts. As a result, it may be possible that an expert's personal preference may have influenced their rating of the wine. Consequently, the actual chemical properties of the wine may not play an important role in the quality rating.

For checking if our argument is right, let's first bucket the quality score into 'High' & 'Low' and use this to color in the scatterplot. We will consider wines as high quality if they have a rating above 5.

As we thought, the principal components do not do a very good job of separating high vs low quality wines. Thus, PCA cannot distinguish between wine quality.

Next, let's run k-Means clustering on the data. First, let's center and scale the data.

Our aim is to predict either Red/White or High/Low quality wine. Hence, we need k-Means to return two clusters.

Let's take a look at the centers of the two clusters after bringing the units back to the original scale.

```
## [1] "Cluster 1:"

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

## [1] "Cluster 2:"

##        fixed.acidity     volatile.acidity          citric.acid
##            8.2895922            0.5319416            0.2695435
##       residual.sugar            chlorides  free.sulfur.dioxide
##            2.6342666            0.0883238           15.7647596
```

```
## total.sulfur.dioxide              density                   pH
##           48.6396835            0.9967404            3.3097200
##            sulphates              alcohol
##            0.6567194           10.4015216
```

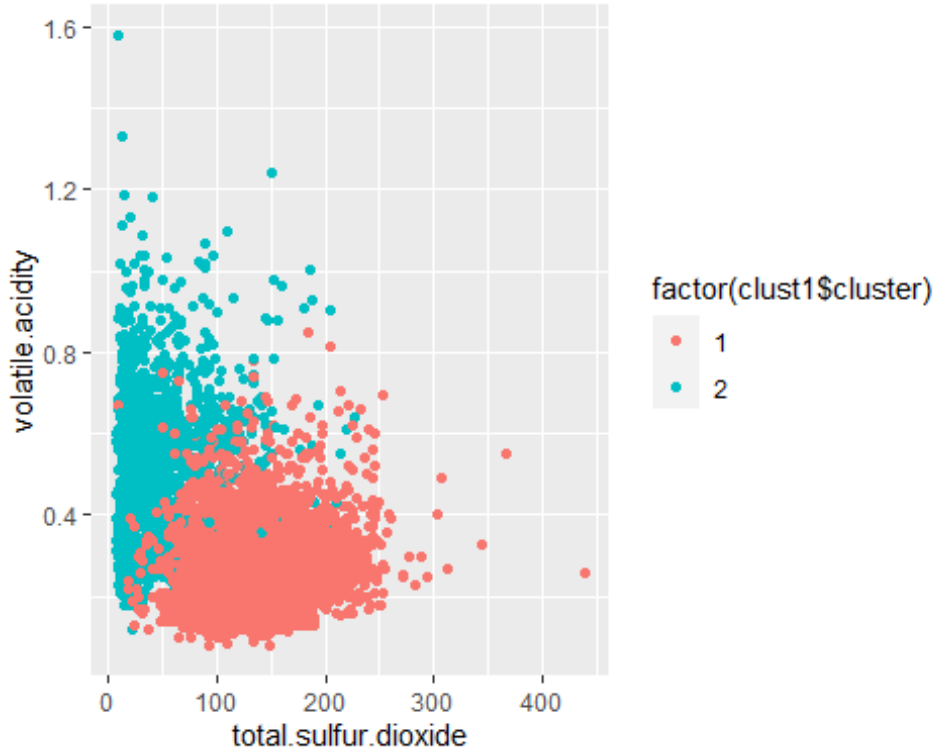
As it is a bit difficult to understand the differences between the two clusters due to the varying scales of the features, let's subtract the centers of the two clusters to get the difference.

```
##     volatile.acidity            chlorides            sulphates
##           -1.5631878           -1.2336600           -1.1284118
##        fixed.acidity              density                   pH
##           -1.1091297           -0.9122427           -0.7593601
##              alcohol          citric.acid       residual.sugar
##            0.1013131            0.4521520            0.7902299
##  free.sulfur.dioxide total.sulfur.dioxide
##            1.1130951            1.5890987
```

We see that the cluster 1 center has more sulfur dioxide & residual sugar compared to cluster 2. Similarly, cluster 2 center has more acidity, chlorides & sulphates compared to cluster 1.

It's becoming clear that these two clusters are distinct as the main differentiating features are the same as the ones we saw from PCA.

In order to visualize, let's create a scatterplot using the two features with the highest magnitude of difference, i.e., Total Sulfur Dioxide & Volatile Acidity. We'll use the two clusters to color in the data points.



We can see that these two clusters are distinctly separate from each other in terms of chemical characteristics. Again, we don't know which cluster corresponds to red vs white wine but we can be fairly confident that these two clusters denotes the wine colors.

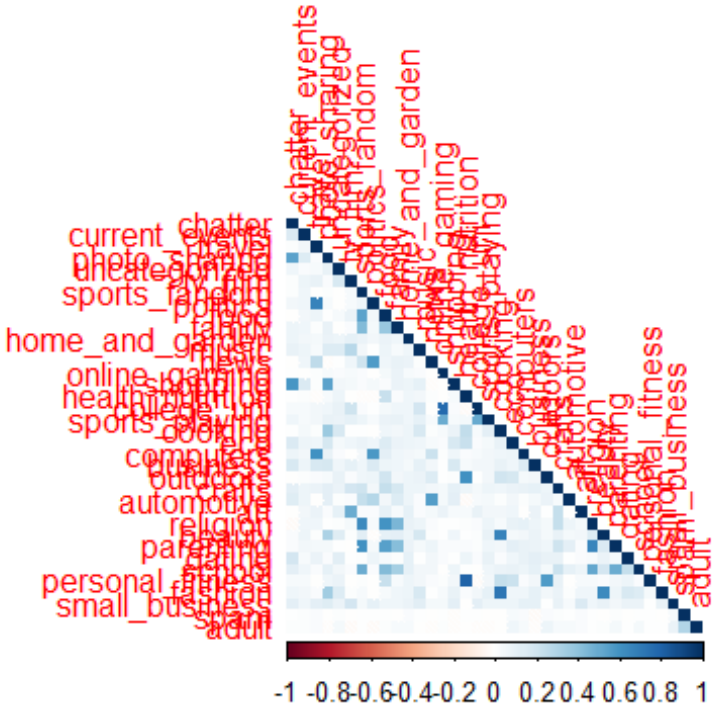
Similar to PCA, k-Means clustering cannot do a good job of separating high quality & low quality wines for the reasons mentioned earlier.

Thus, we can conclude that k-Means clustering can distinguish between red and white wines.

Q7: Market segmentation

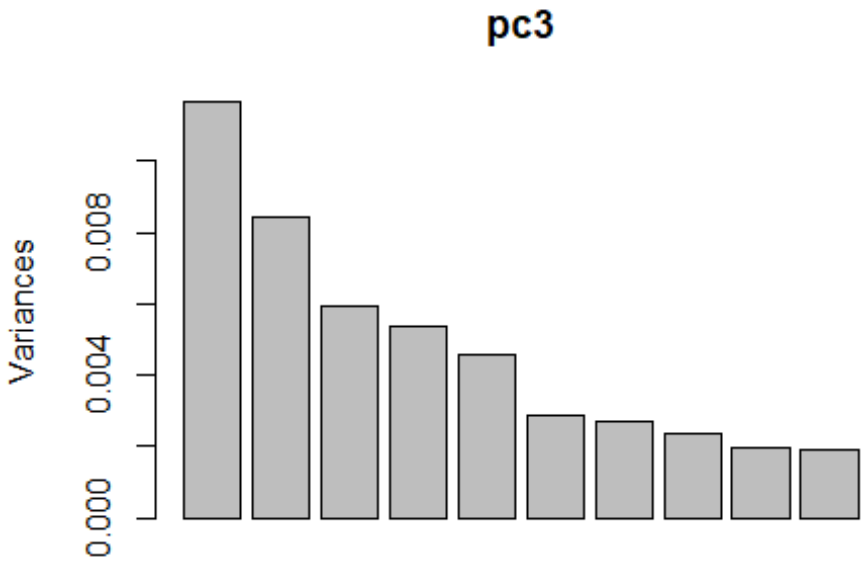
```
## 'data.frame':    7882 obs. of  37 variables:
##  $ X              : chr  "hmjoe4g3k" "clk1m5w8s" "jcsovtak3" "3oeb4hiln"
...
##  $ chatter        : int  2 3 6 1 5 6 1 5 6 5 ...
##  $ current_events : int  0 3 3 5 2 4 2 3 2 2 ...
##  $ travel         : int  2 2 4 2 0 2 7 3 0 4 ...
##  $ photo_sharing  : int  2 1 3 2 6 7 1 6 1 4 ...
##  $ uncategorized  : int  2 1 1 0 1 0 0 1 0 0 ...
##  $ tv_film        : int  1 1 5 1 0 1 1 1 0 5 ...
##  $ sports_fandom  : int  1 4 0 0 0 1 1 1 0 9 ...
##  $ politics       : int  0 1 2 1 2 0 11 0 0 1 ...
##  $ food           : int  4 2 1 0 0 2 1 0 2 5 ...
##  $ family         : int  1 2 1 1 1 1 0 0 2 4 ...
##  $ home_and_garden : int  2 1 1 0 0 1 0 0 1 0 ...
```

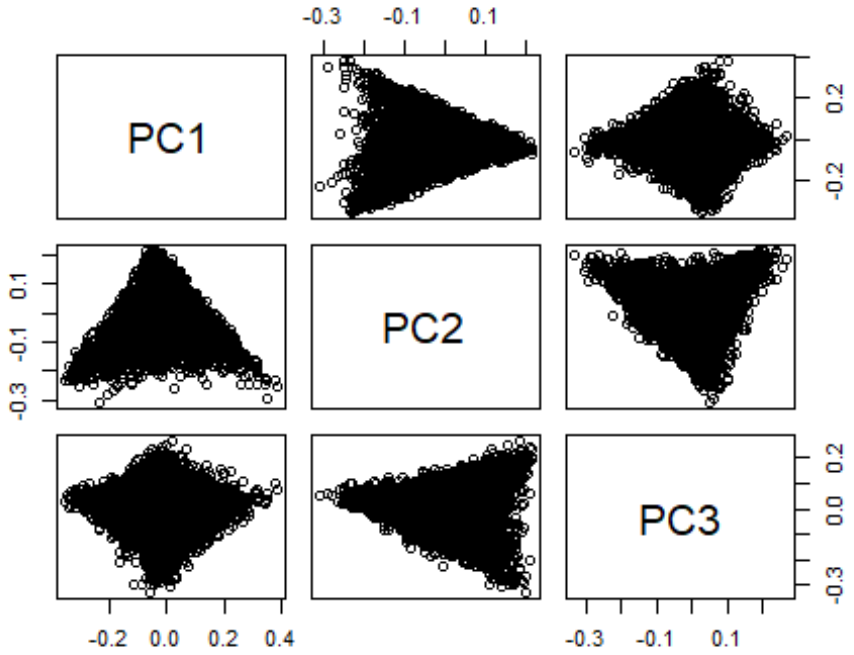
```
##  $ music           : int  0 0 1 0 0 1 0 2 1 1 ...
##  $ news            : int  0 0 1 0 0 0 1 0 0 0 ...
##  $ online_gaming   : int  0 0 0 0 3 0 0 1 2 1 ...
##  $ shopping        : int  1 0 2 0 2 5 1 3 0 0 ...
##  $ health_nutrition: int  17 0 0 0 0 0 1 1 22 7 ...
##  $ college_uni     : int  0 0 0 1 4 0 1 0 1 4 ...
##  $ sports_playing  : int  2 1 0 0 0 0 1 0 0 1 ...
##  $ cooking         : int  5 0 2 0 1 0 1 10 5 4 ...
##  $ eco             : int  1 0 1 0 0 0 0 0 2 1 ...
##  $ computers       : int  1 0 0 0 1 1 1 1 1 2 ...
##  $ business        : int  0 1 0 1 0 1 3 0 1 0 ...
##  $ outdoors        : int  2 0 0 0 1 0 1 0 3 0 ...
##  $ crafts          : int  1 2 2 3 0 0 0 1 0 0 ...
##  $ automotive      : int  0 0 0 0 0 1 0 1 0 4 ...
##  $ art             : int  0 0 8 2 0 0 1 0 1 0 ...
##  $ religion        : int  1 0 0 0 0 0 1 0 0 13 ...
##  $ beauty          : int  0 0 1 1 0 0 0 5 5 1 ...
##  $ parenting       : int  1 0 0 0 0 0 0 1 0 3 ...
##  $ dating          : int  1 1 1 0 0 0 0 0 0 0 ...
##  $ school          : int  0 4 0 0 0 0 0 0 1 3 ...
##  $ personal_fitness: int  11 0 0 0 0 0 0 0 12 2 ...
##  $ fashion         : int  0 0 1 0 0 0 0 4 3 1 ...
##  $ small_business  : int  0 0 0 0 1 0 0 0 1 0 ...
##  $ spam            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ adult           : int  0 0 0 0 0 0 0 0 0 0 ...
```



Findings:
* Shopping and photo-sharing are positively correlated
* College_uni and online_gaming stands out with a strong positive correlation
* Health_nutrition,peronal_fitness and outdoors have a high positive correlation showing these people are health conscious
* Fashion and beauty have a strong postive correlation

We can include all the variables in the cluster analysis to understand if the same points appear after profiling the clusters



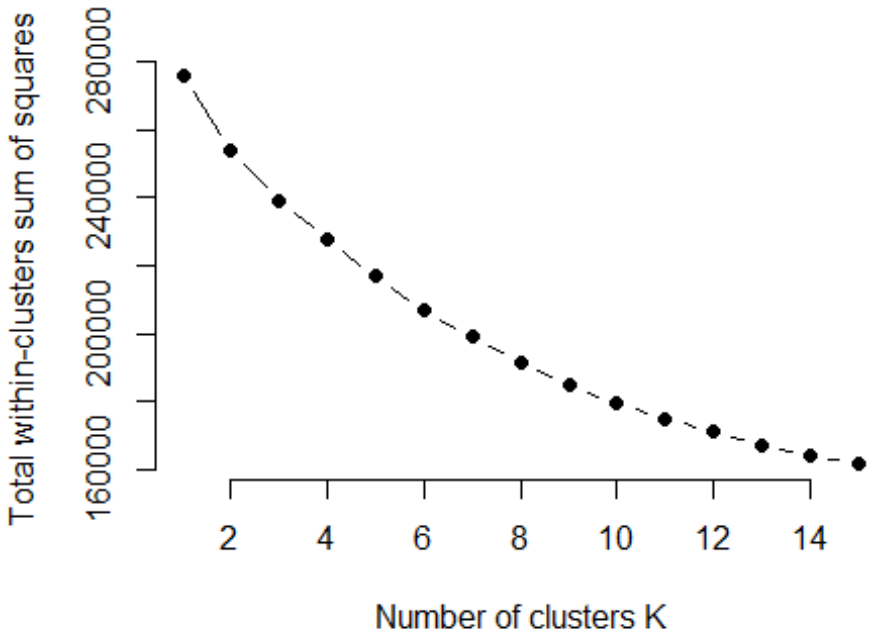


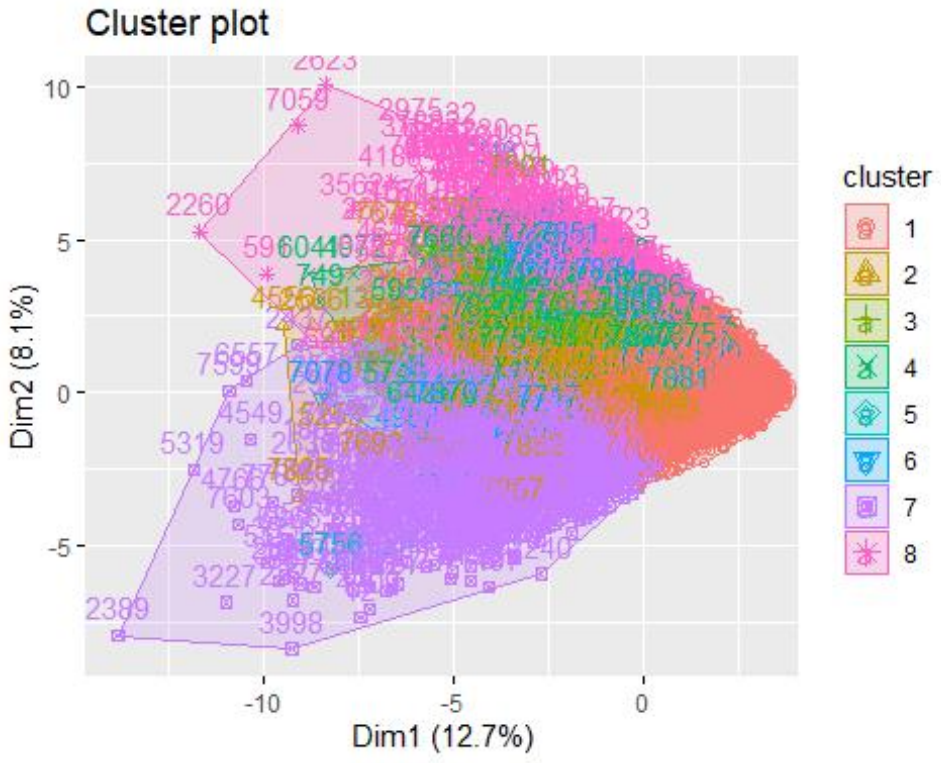
We observe the three categories

- Fitness Enthusiast - These are part of the active community who look for outdoor activities, cooking, and health and nutrition. This community is important as they could be the majority consumers of NutrientH20 but they could also be very particular about everything and detail oriented. So, Nutrient H20 should make sure that this segment is always satisfied and doesn't have any consumer dissonance.
- The Instagrammer/Vlogger - These are very social people and let everyone know about everything they do. They are buying the brand because of their friend, influencer or because of the brand value. consistent advertising to keep the brand image should work on this group.
- The Young College Kids - They probably buy it from their vending machines or close convenience stores. The key to this segment is purely through distribution and college events.

Step 2: Normalize the data and perform k - means clustering

There is no considerable decrease in the error after 8 clusters.
Hence 8 clusters were considered to be optimal for the analysis





The cluster separation is not very clear from the plot. So lets analyze the cluster centers to come up with the profiles based on the clusters

Findings:
* There are multiple interesting profiles that came out of the clusters

Cluster 2: People who are grouped under cluster 2 tweet a lot about photo sharing,cooking,beauty and fashion

Cluster 3: This segment is profoundly university students as they tweet about sports,universities and online games

Cluster 4: This segment of people are potentially health conscious people as they tweet mostly about health,nutrition,outdoors and personal fitness

Cluster 5: Most of the tweets that these people tweet are adult related tweets

Cluster 6: This segment of people are interested in tweeting about films, art, music, tv and dating

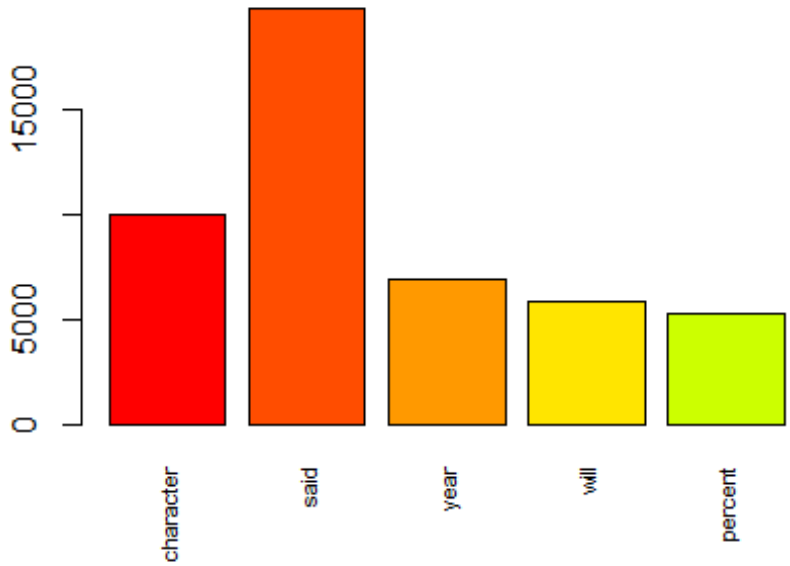
Cluster 7: This segment of people might be profoundly adults as they tweet a lot about religion, parenting, family, sports, school,food and crafts

Cluster 8: People from this segment are interested in tweeting about travel, politics, news and automotive

Q8: The Reuters corpus

- Now, we will perform tokenization by removing whitespaces, sparse terms, stopwords, punctuations and make the letters lowercase. We perform the same steps on both the test and train files.
- We create a Document Term matrix

```
## [1] "character" "said"      "year"      "will"      "percent"
```

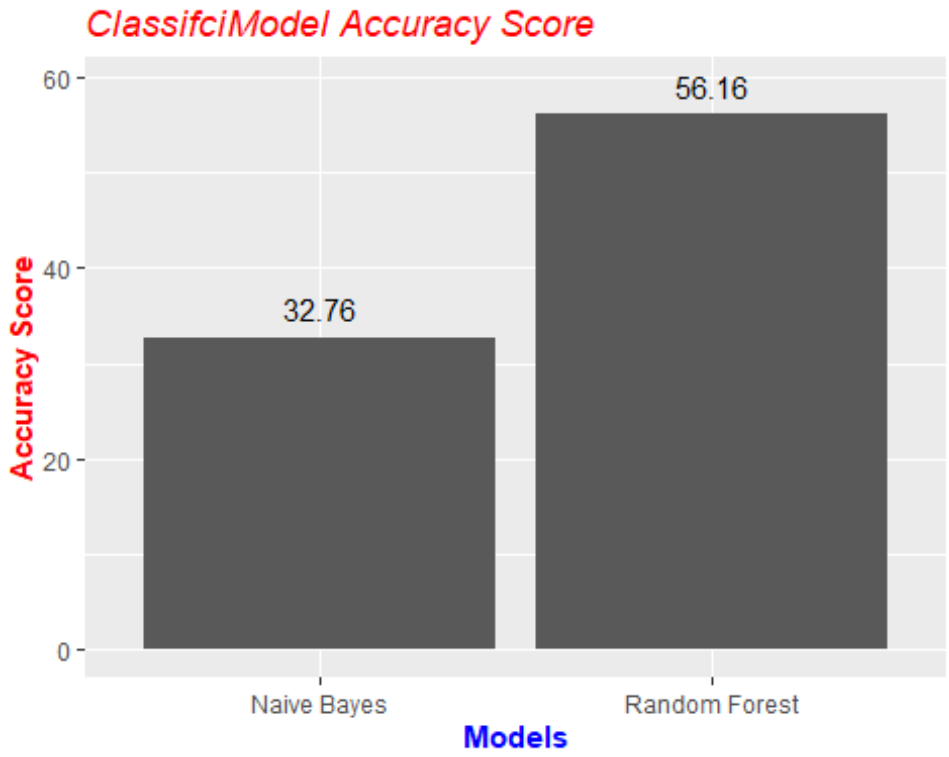


- The above are the most frequent words observed. Though we will not remove these words for now. In Inverse Term Frequency weighting, we could remove the word 'Said' as that may not be very relevant in our search.

Problem Statement: We try to predict the author by using the words in documents in the train folder.

We ran Naive Bayes to predict the author based on the word pattern count.

```
## The accuracy 0.3276

## The accuracy for Random Forest 0.5656
```



Results: We observed that Naive Bayes accuracy is quite low and hence, we check with Random Forest. We achieved a higher accuracy but there are some concerns. Technically, with just a word count, predicting an author would be tough. What we could think about is predicting a genre the author is known for. If we do have a table to train the model with authors and their respective genres, I believe we could achieve more accuracy with such problem statements.

Conclusions: This dataset is a goldmine to understand many other factors.For eg, if we have the book revenue/popularity of such authors, we would be able to predict what keywords produce more revenue or are more popular. The applications of such a dataset could also be extended to songs, podcasts, etc. This could then be utilized to think of strategies to increase interest among public as well.

Q9: Association rule mining

```
##
## Attaching package: 'igraph'

## The following object is masked from 'package:mosaic':
##
##     compare

## The following objects are masked from 'package:purrr':
##
##     compose, simplify
```

```
## The following object is masked from 'package:tidyr':
##
##     crossing

## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union

##
## Attaching package: 'arules'

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

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

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

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

Let's load in our data (groceries.txt). We see that the file consists of many rows with each row being a shopping basket. Unfortunately for us, this means that each row has different number of items and hence different number of columns. If we tried reading this file using the default function, R wouldn't read it correctly.

Hence, we first need to figure out what is the maximum number of columns present in the data set, i.e., maximum number of items in a shopping basket. R provides a function *count.fields* which can help us with this.

```
## Maximum number of columns in the file =  32
```

We see that the maximum number of columns is 32. Now that we know how many columns are there, we can read the file. We also need to tell R to not consider the first line as the

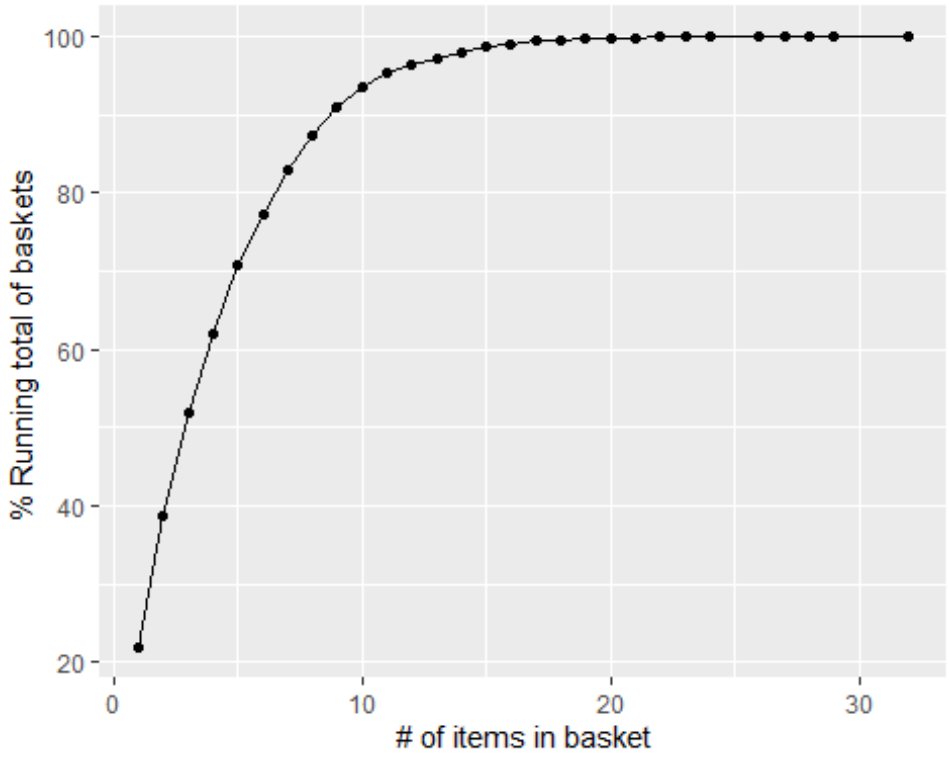
header. Further, R needs to be told that the rows have unequal lengths so that it can fill the missing columns with blanks.

Let's look at the size of the data set.

```
## [1] 9835   32
```

The data set has ~10k shopping baskets. Now, let's do some preprocessing to convert this data set into a format that is expected by the 'arules' package.

Before running the apriori algorithm, let's quickly check what is the proportion of the # of items present in shopping baskets. The below graph shows a running total of the % proportion for each # of items present in the basket.



We see ~75% of the shopping baskets have # of items <= 6. Therefore, we'll use 6 items as the maximum item length in the apriori algorithm. We'll also only look at item sets with support more than 0.005 (present in >50 baskets in our data set) and confidence more than 0.05. This will ensure we get discover rules backed by a significant number of shopping baskets.

Let's look at 10 rules for the baskets where lift > 3. This means that we want to look at rules where the presence of the LHS items in a basket will increase the probability of buying the RHS item by a factor of 3.

```
##      lhs                     rhs                      support confidence
coverage     lift count
## [1]  {ham}                => {white bread}        0.005083884 0.19531250 0
```

```
.02602949 4.639851    50
## [2]  {white bread}        => {ham}                0.005083884 0.12077295 0
.04209456 4.639851    50
## [3]  {citrus fruit,
##       other vegetables,
##       whole milk}         => {root vegetables}    0.005795628 0.44531250 0
.01301474 4.085493    57
## [4]  {butter,
##       other vegetables}   => {whipped/sour cream} 0.005795628 0.28934010 0
.02003050 4.036397    57
## [5]  {root vegetables}    => {herbs}              0.007015760 0.06436567 0
.10899847 3.956477    69
## [6]  {herbs}              => {root vegetables}    0.007015760 0.43125000 0
.01626843 3.956477    69
## [7]  {other vegetables,
##       root vegetables}    => {onions}             0.005693950 0.12017167 0
.04738180 3.875044    56
## [8]  {citrus fruit,
##       pip fruit}          => {tropical fruit}     0.005592272 0.40441176 0
.01382816 3.854060    55
## [9]  {berries}            => {whipped/sour cream} 0.009049314 0.27217125 0
.03324860 3.796886    89
## [10] {whipped/sour cream} => {berries}            0.009049314 0.12624113 0
.07168277 3.796886    89
```

The rules with the highest lift relate to ham & white bread. This makes sense. Most people would buy these two items to make a ham sandwich but usually not each item individually. The next few rules relate to fruits and vegetables. The last two rules are about cream and berries. These two items would usually not be purchased a lot individually. However, most people would buy both of these while making desserts, cakes, etc. which explains why it has a high lift.

Next, we'll look at 10 rules where the confidence is more than 0.5, i.e., the probability of the RHS item being bought given that the LHS items are present in the basket.

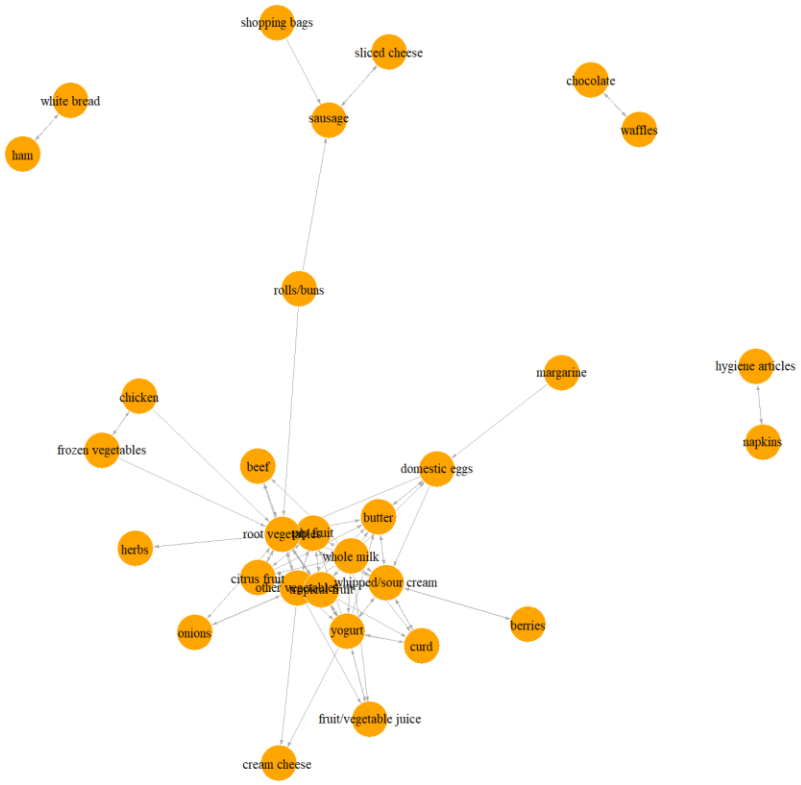
```
##      lhs                     rhs                    support confidence
coverage     lift count
## [1]  {root vegetables,
##       tropical fruit,
##       yogurt}             => {whole milk}       0.005693950  0.7000000 0.0
08134215 2.739554    56
## [2]  {other vegetables,
##       pip fruit,
##       root vegetables}    => {whole milk}       0.005490595  0.6750000 0.0
08134215 2.641713    54
## [3]  {butter,
##       whipped/sour cream} => {whole milk}       0.006710727  0.6600000 0.0
10167768 2.583008    66
## [4]  {pip fruit,
##       whipped/sour cream} => {whole milk}       0.005998983  0.6483516 0.0
```

```
09252669 2.537421    59
## [5]  {butter,
##       yogurt}             => {whole milk}       0.009354347  0.6388889 0.0
14641586 2.500387    92
## [6]  {butter,
##       root vegetables}    => {whole milk}       0.008235892  0.6377953 0.0
12913066 2.496107    81
## [7]  {curd,
##       tropical fruit}     => {whole milk}       0.006507372  0.6336634 0.0
10269446 2.479936    64
## [8]  {citrus fruit,
##       root vegetables,
##       whole milk}         => {other vegetables} 0.005795628  0.6333333 0.0
09150991 3.273165    57
## [9]  {other vegetables,
##       pip fruit,
##       yogurt}             => {whole milk}       0.005083884  0.6250000 0.0
08134215 2.446031    50
## [10] {domestic eggs,
##       pip fruit}          => {whole milk}       0.005388917  0.6235294 0.0
08642603 2.440275    53
```

We see that most of these rules involve Whole milk as the RHS item. Therefore, the confidence is interpreted as what is the probability of buying whole milk given one has bought the LHS items. For example, if one has already bought root vegetables, tropical fruit and yogurt, the probability of them buying whole milk is 0.7.

However, these rules have lower lifts compared to the rules seen earlier. Why is that? This is because milk is something that is bought very frequently ,i.e., it has a high probability of being bought. Hence, the increase in probability that milk will be bought given the other items were bought would be lesser.

Finally, let's visualize the network with all rules with lift > 3.



There are a lot of associations between commonly bought items like milk, vegetables, butter and fruits. We also see some other isolated associations which make sense. For example, napkins & hygiene articles, waffles & chocolate, ham & bread. Thus, if a grocery store should choose to act on this data, they should keep these items where the lift is > 3 close to each other.