STA_380N_PART2_EXAM

Mihir Deshpande

2022-08-04

STA 280N Intro to Machine Learning Part 2 Take Home Exam

A link to a GitHub repo where the final report has been knitted and stored *in Markdown (.md) or PDF format.*

Type	Link
RMD File	

A link to a GitHub repo where the final report has been knitted and stored *in Markdown (.md) or PDF format.*

Problem 1) Probability practice

Part A. Visitors to your website are asked to answer a single survey question before they get access to the content on the page. Among all of the users, there are two categories: Random Clicker (RC), and Truthful Clicker (TC). There are two possible answers to the survey: yes and no. Random clickers would click either one with equal probability. You are also giving the information that the expected fraction of random clickers is 0.3. After a trial period, you get the following survey results: 65% said Yes and 35% said No. What fraction of people who are truthful clickers answered yes? Hint: use the rule of total probability.

Answer:

Let's assume the total no. of people who answered survey = 100

$$S = 100$$

Now, we know expected fraction of Random Clickers (RC) = 0.3

$$P(RC) = 0.3$$

$$=> N(RC) = S * 0.3 = 30$$

Now, Random Clickers are equally likely to answer Yes/No.

Therefore, of the 100 people who answered the survey, 30 were random clickers out of whom 15 answered **Yes** and 15 answered **No**

We know that at the end of the survey, we got 65% who voted Yes and 35% who voted No.

i.e

65 people voted **Yes** out of which 15 were Random Clickers => 50 who voted **Yes** were Truthful Clickers (TC)

35 people voted **No** out of which 15 were Random Clickers => 20 who voted **No** were Truthful Clickers (TC)

Therefore, Fraction of people who are Truthful Clickers (TC) who answered **Yes**

= 50/(50+20)

= **5/7**

Part B. Imagine a medical test for a disease with the following two attributes:

- The sensitivity is about 0.993. That is, if someone has the disease, there is a probability of 0.993 that they will test positive.
- The specificity is about 0.9999. This means that if someone doesn't have the disease, there is probability of 0.9999 that they will test negative.
- In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it (or 0.000025 as a decimal probability).

Suppose someone tests positive. What is the probability that they have the disease?

Answer:

Let's define the following events -

P = Event that you test Positive

N = Event that you test Negative

D = Event that you have the Disease

ND = Event that you don't have the Disease

What is the probability that youi have the disease given you tested positive?

i.e

$$P(D \mid P) = ?$$

Now, we know the following -

$P(P \mid D) = 0.9330$

$P(N \mid ND) = 0.9999$

$P(D) = 0.000025$

Now,

$P(ND) = 1 - 0.000025$

$\therefore P(ND) = 0.999975$

We know,

$$P(D \mid P) = \frac{P(D \cap P)}{P(P)}$$

Let's find the value of Numerator -

$P(P \mid D) = \frac{P(D \cap P)}{P(D)}$

$\therefore P(D \cap P) = P(P \mid D) \cdot P(D)$

$\therefore P(D \cap P) = 0.933 * 0.000025$

$\therefore P(D \cap P) = 0.000023325$.....(1)

Now,

$P(P) = \sum_{X=0}^{N} P(X \cap P)$

$P(P) = P(D \cap P) + P(ND \cap P)$

$P(P) = 0.000023325 + P(ND \cap P)$.....$(from 1)$...(2)

Now,

$P(P \mid ND) = 1 - P(N \mid ND)$

$\therefore P(P \mid ND) = 1 - 0.9999$

$\therefore P(P \mid ND) = 0.0001$

We know,

$P(P \mid ND) = \frac{P(ND \cap P)}{P(ND)}$

$\therefore P(ND \cap P) = P(P \mid ND) \cdot P(ND)$

$\therefore P(ND \cap P) = 0.0001 * 0.999975$

$\therefore P(ND \cap P) = 0.0000999975$.....(3)

Now, from 2 and 3,

$P(P) = 0.000023325 + 0.0000999975$

$P(P) = 0.0001233225$.....(4)

From 1 and 4,

$P(D \mid P) = \frac{0.000023325}{0.0001233225}$

$\therefore P(D \mid P) = 0.1891$

Hence, if you test positive, you have an 18.91% chance of having the disease.

Problem 2: Wrangling the Billboard Top 100

Consider the data in billboard.csv containing every song to appear on the weekly Billboard Top 100 chart since 1958, up through the middle of 2021. Each row of this data corresponds to a single song in a single week. For our purposes, the relevant columns here are:

- performer: who performed the song
- song: the title of the song
- year: year (1958 to 2021)
- week: chart week of that year (1, 2, etc)
- week_position: what position that song occupied that week on the Billboard top 100 chart.

Use your skills in data wrangling and plotting to answer the following three questions.

Part A: Make a table of the top 10 most popular songs since 1958, as measured by the *total number of weeks that a song spent on the Billboard Top 100.* Note that these data end in week 22 of 2021, so the most popular songs of 2021 will not have up-to-the-minute data; please send our apologies to The Weeknd.

Solution:

Let's group by song and performer, summarize and sort in descending order by week count and print the top 10 entries -

Table 2: Table 2.1 Top 10 Songs and their artists with maximum number of weeks in Billboard 100

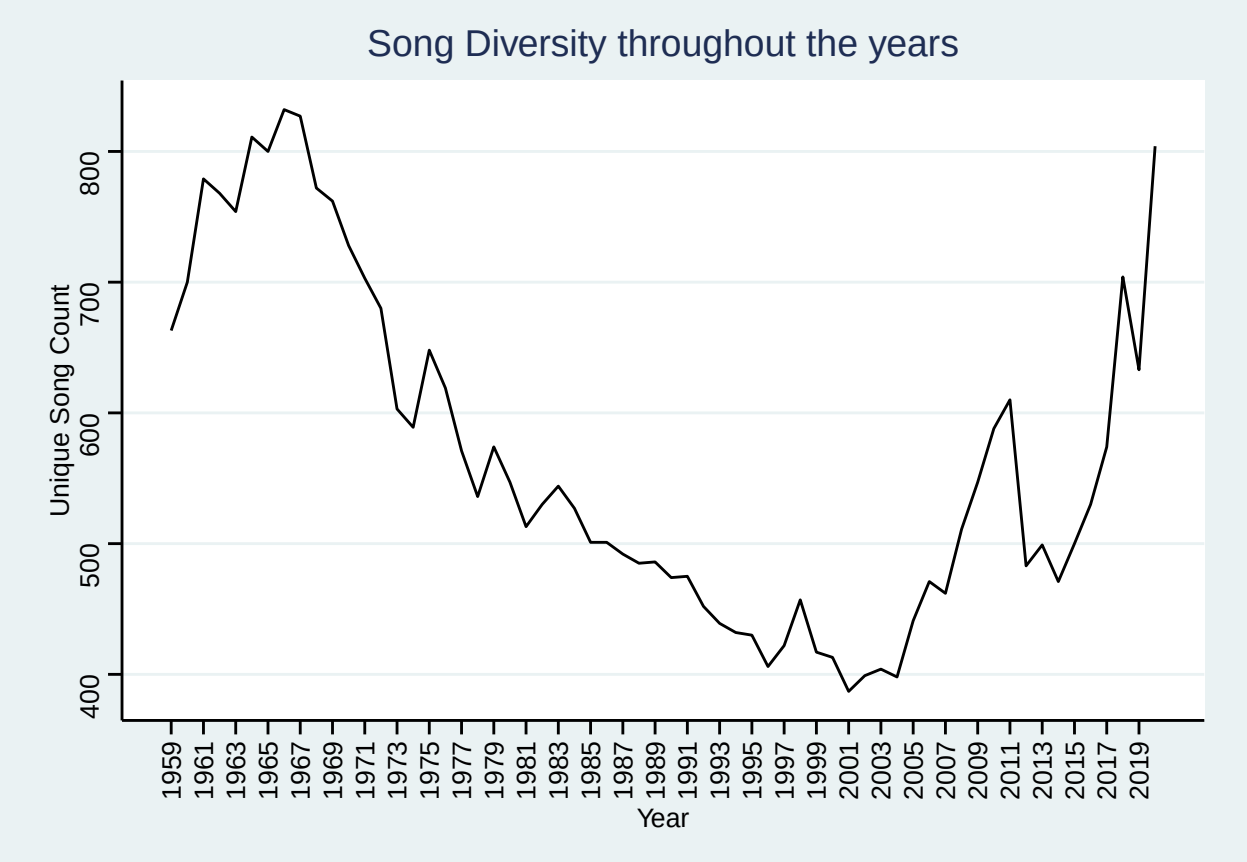
song	performer	week_count
Radioactive	Imagine Dragons	87
Sail	AWOLNATION	79
Blinding Lights	The Weeknd	76
I'm Yours	Jason Mraz	76
How Do I Live	LeAnn Rimes	69
Counting Stars	OneRepublic	68
Party Rock Anthem	LMFAO Featuring Lauren Bennett & GoonRock	68
Foolish Games/You Were Meant For Me	Jewel	65
Rolling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

Part B: Is the "musical diversity" of the Billboard Top 100 changing over time? Let's find out. We'll measure the musical diversity of given year as *the number of unique songs that appeared in the Billboard Top 100 that year.* Make a line graph that plots this measure of musical diversity over the years. The x axis should show the year, while the y axis should show the number of unique songs appearing at any position on the Billboard Top 100 chart in any week that year. For this part, please filter the data set so that it excludes the years 1958 and 2021, since we do not have complete data on either of those years. Give the figure an informative caption in which you explain what is shown in the figure and comment on any interesting trends you see.

There are number of ways to accomplish the data wrangling here. We offer you two hints on two possibilities:

1) You could use two distinct sets of data-wrangling steps. The first set of steps would get you a table that counts the number of times that a given song appears on the Top 100 in a given year. The second set of steps operate on the result of the first set of steps; it would count the number of unique songs that appeared on the Top 100 in each year, *irrespective of how many times* it had appeared.
2) You could use a single set of data-wrangling steps that combines the `length` and `unique` commands.

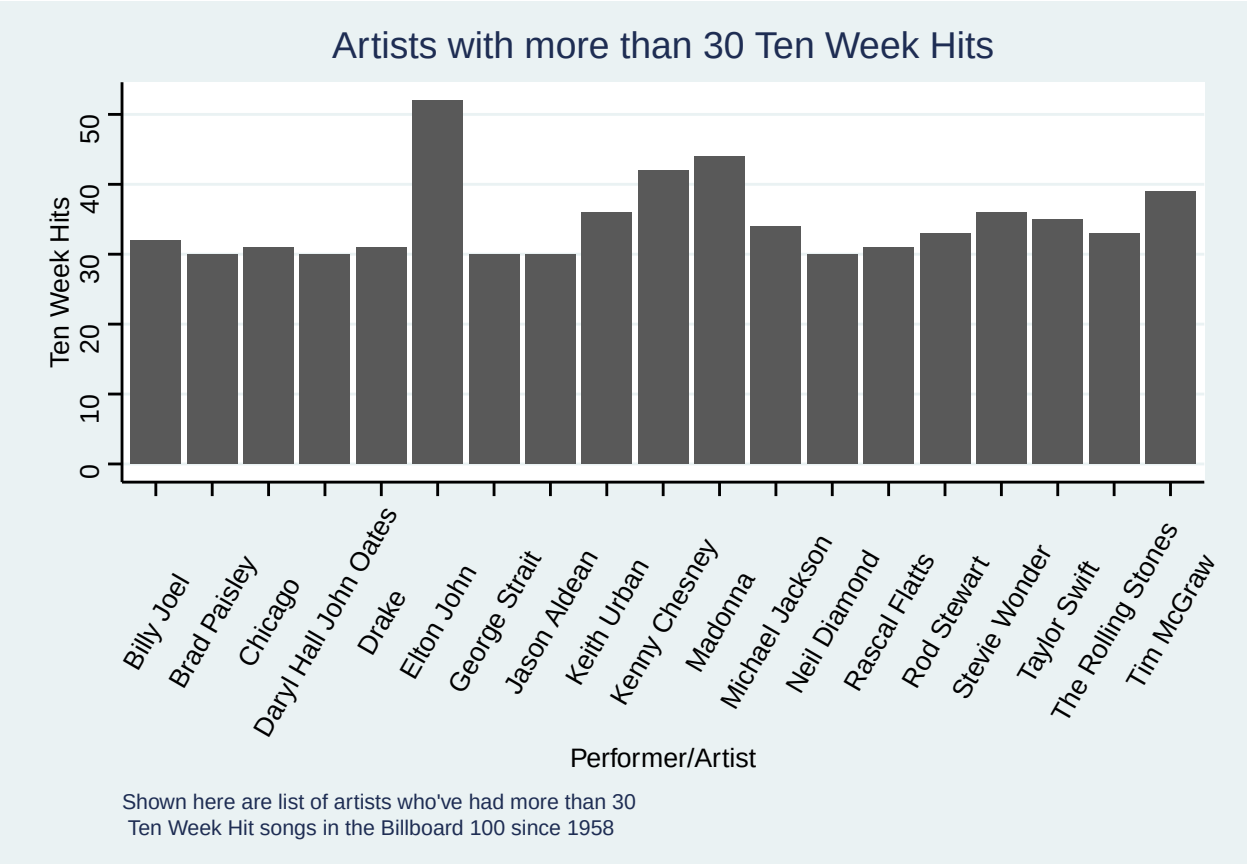
Solution:



We can see that from the mid 1980s to the mid 2000s, we don't really have a lot of song diversity! The unique no. of songs that top the **Billboard 100** significantly drop in this era but we are seeing a rise in diversity from 2007 onwards.

Part C: Let's define a "ten-week hit" as a single song that appeared on the Billboard Top 100 for at least ten weeks. There are 19 artists in U.S. musical history since 1958 who have had *at least 30 songs* that were "ten-week hits." Make a bar plot for these 19 artists, showing how many ten-week hits each one had in their musical career. Give the plot an informative caption in which you explain what is shown.

Solution:



Visual story telling part 1: green buildings

The case

Over the past decade, both investors and the general public have paid increasingly close attention to the benefits of environmentally conscious buildings. There are both ethical and economic forces at work here. In commercial real estate, issues of eco-friendliness are intimately tied up with ordinary decisions about how to allocate capital. In this context, the decision to invest in eco-friendly buildings could pay off in at least four ways.

1. Every building has the obvious list of recurring costs: water, climate control, lighting, waste disposal, and so forth. Almost by definition, these costs are lower in green buildings.

2. Green buildings are often associated with better indoor environments—the kind that are full of sunlight, natural materials, and various other humane touches. Such environments, in turn, might result in higher employee productivity and lower absenteeism, and might therefore be more coveted by potential tenants. The financial impact of this factor, however, is rather hard to quantify ex ante; you cannot simply ask an engineer in the same way that you could ask a question such as, "How much are these solar panels likely to save on the power bill?"

3. Green buildings make for good PR. They send a signal about social responsibility and ecological awareness, and might therefore command a premium from potential tenants who want their customers to associate them with these values. It is widely believed that a good corporate image may

enable a firm to charge premium prices, to hire better talent, and to attract socially conscious investors.

4. Finally, sustainable buildings might have longer economically valuable lives. For one thing, they are expected to last longer, in a direct physical sense. (One of the core concepts of the green-building movement is "life-cycle analysis," which accounts for the high front-end environmental impact of acquiring materials and constructing a new building in the first place.) Moreover, green buildings may also be less susceptible to market risk—in particular, the risk that energy prices will spike, driving away tenants into the arms of bolder, greener investors.

Of course, much of this is mere conjecture. At the end of the day, tenants may or may not be willing to pay a premium for rental space in green buildings. We can only find out by carefully examining data on the commercial real-estate market.

The file greenbuildings.csv contains data on 7,894 commercial rental properties from across the United States. Of these, 685 properties have been awarded either LEED or EnergyStar certification as a green building. You can easily find out more about these rating systems on the web, e.g. at www.usgbc.org. The basic idea is that a commercial property can receive a green certification if its energy efficiency, carbon footprint, site selection, and building materials meet certain environmental benchmarks, as certified by outside engineers.

A group of real estate economists constructed the data in the following way. Of the 1,360 green-certified buildings listed as of December 2007 on the LEED or EnergyStar websites, current information about building characteristics and monthly rents were available for 685 of them. In order to provide a control population, each of these 685 buildings was matched to a cluster of nearby commercial buildings in the CoStar database. Each small cluster contains one green-certified building, and all non-rated buildings within a quarter-mile radius of the certified building. On average, each of the 685 clusters contains roughly 12 buildings, for a total of 7,894 data points.

The columns of the data set are coded as follows:

- CS.PropertyID: the building's unique identifier in the CoStar database.
- cluster: an identifier for the building cluster, with each cluster containing one green-certified building and at least one other non-green-certified building within a quarter-mile radius of the cluster center.
- size: the total square footage of available rental space in the building.
- empl.gr: the year-on-year growth rate in employment in the building's geographic region.
- Rent: the rent charged to tenants in the building, in dollars per square foot per calendar year.
- leasing.rate: a measure of occupancy; the fraction of the building's available space currently under lease.
- stories: the height of the building in stories.
- age: the age of the building in years.
- renovated: whether the building has undergone substantial renovations during its lifetime.
- class.a, class.b: indicators for two classes of building quality (the third is Class C). These are relative classifications within a specific market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.
- green.rating: an indicator for whether the building is either LEED- or EnergyStar-certified.

- LEED, Energystar: indicators for the two specific kinds of green certifications.
- net: an indicator as to whether the rent is quoted on a "net contract" basis. Tenants with net-rental contracts pay their own utility costs, which are otherwise included in the quoted rental price.
- amenities: an indicator of whether at least one of the following amenities is available on-site: bank, convenience store, dry cleaner, restaurant, retail shops, fitness center.
- cd.total.07: number of cooling degree days in the building's region in 2007. A degree day is a measure of demand for energy; higher values mean greater demand. Cooling degree days are measured relative to a baseline outdoor temperature, below which a building needs no cooling.
- hd.total07: number of heating degree days in the building's region in 2007. Heating degree days are also measured relative to a baseline outdoor temperature, above which a building needs no heating.
- total.dd.07: the total number of degree days (either heating or cooling) in the building's region in 2007.
- Precipitation: annual precipitation in inches in the building's geographic region.
- Gas.Costs: a measure of how much natural gas costs in the building's geographic region.
- Electricity.Costs: a measure of how much electricity costs in the building's geographic region.
- cluster.rent: a measure of average rent per square-foot per calendar year in the building's local market.

The goal

An Austin real-estate developer is interested in the possible economic impact of "going green" in her latest project: a new 15-story mixed-use building on East Cesar Chavez, just across I-35 from downtown. Will investing in a green building be worth it, from an economic perspective? The baseline construction costs are $100 million, with a 5% expected premium for green certification.

The developer has had someone on her staff, who's been described to her as a "total Excel guru from his undergrad statistics course," run some numbers on this data set and make a preliminary recommendation. Here's how this person described his process.

> I began by cleaning the data a little bit. In particular, I noticed that a handful of the buildings in the data set had very low occupancy rates (less than 10% of available space occupied). I decided to remove these buildings from consideration, on the theory that these buildings might have something weird going on with them, and could potentially distort the analysis. Once I scrubbed these low-occupancy buildings from the data set, I looked at the green buildings and non-green buildings separately. The median market rent in the non-green buildings was $25 per square foot per year, while the median market rent in the green buildings was $27.60 per square foot per year: about $2.60 more per square foot. (I used the median rather than the mean, because there were still some outliers in the data, and the median is a lot more robust to outliers.) Because our building would be 250,000 square feet, this would translate into an additional $250000 x 2.6 = $650000 of extra revenue per year if we build the green building.
>
> Our expected baseline construction costs are $100 million, with a 5% expected premium for green certification. Thus we should expect to spend an extra $5 million on the green building. Based on the extra revenue we would make, we would recuperate these costs in $5000000/650000 = 7.7 years. Even if our occupancy rate were only 90%, we would still recuperate the costs in a little over 8 years. Thus from year 9 onwards, we would be making an extra $650,000 per year in profit. Since the building will be earning rents for 30 years or more, it seems like a good financial move to build the green building.

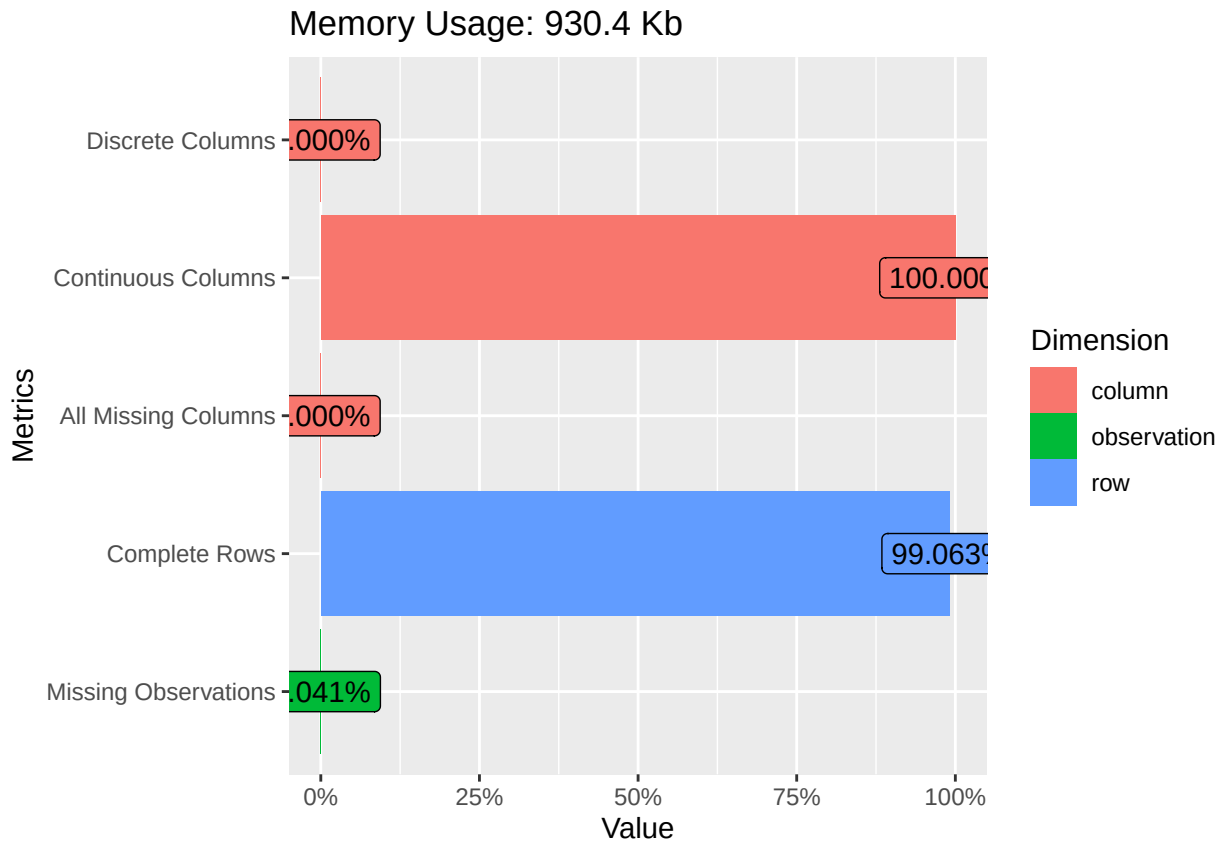
The developer listened to this recommendation, understood the analysis, and still felt unconvinced. She has therefore asked you to revisit the report, so that she can get a second opinion.

Do you agree with the conclusions of her on-staff stats guru? If so, point to evidence supporting his case. If not, explain specifically where and why the analysis goes wrong, and how it can be improved. Do you see the possibility of confounding variables for the relationship between rent and green status? If so, provide evidence for confounding, and see if you can also make a picture that visually shows how we might "adjust" for such a confounder. *Tell your story in pictures, with appropriate introductory and supporting text.*

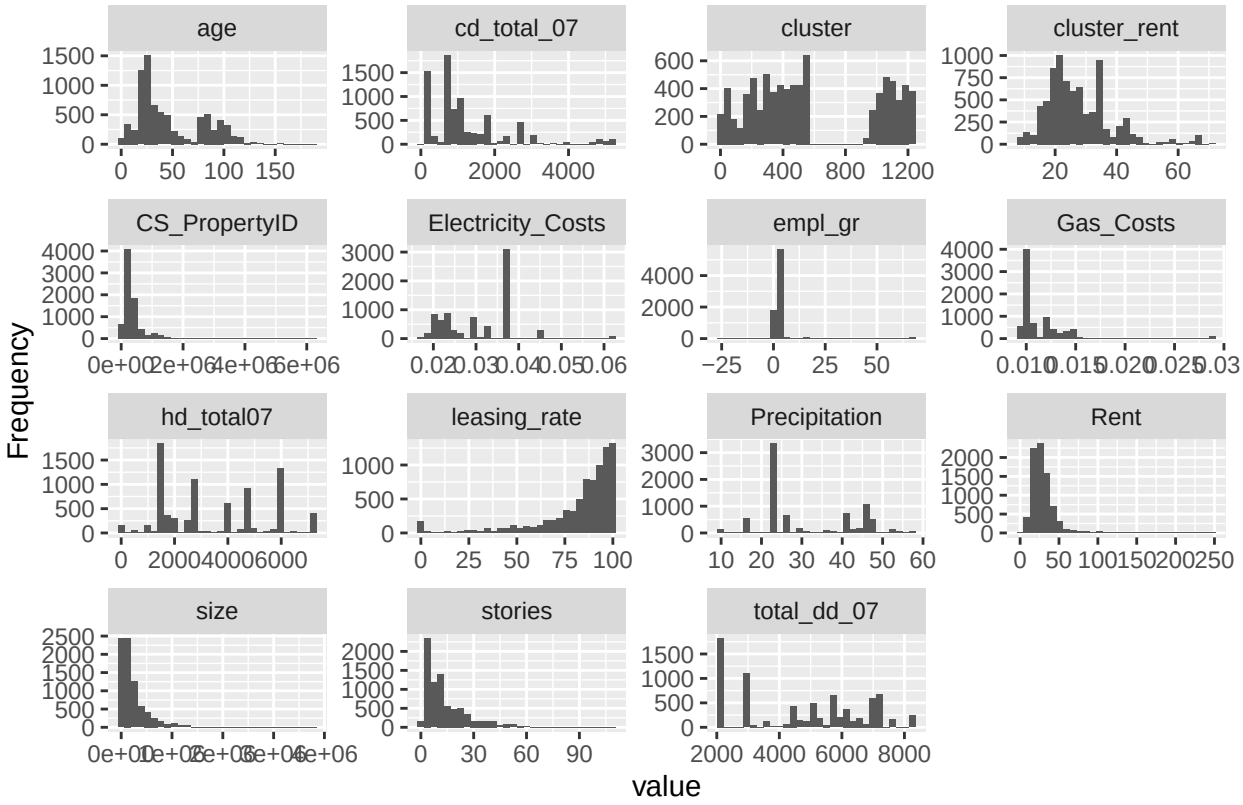
Note: this is intended as an exercise in visual and numerical story-telling. Your approach should rely on pictures and/or tables, not a regression model. Tell a story understandable to a non-technical audience. Keep it concise.

Solution:

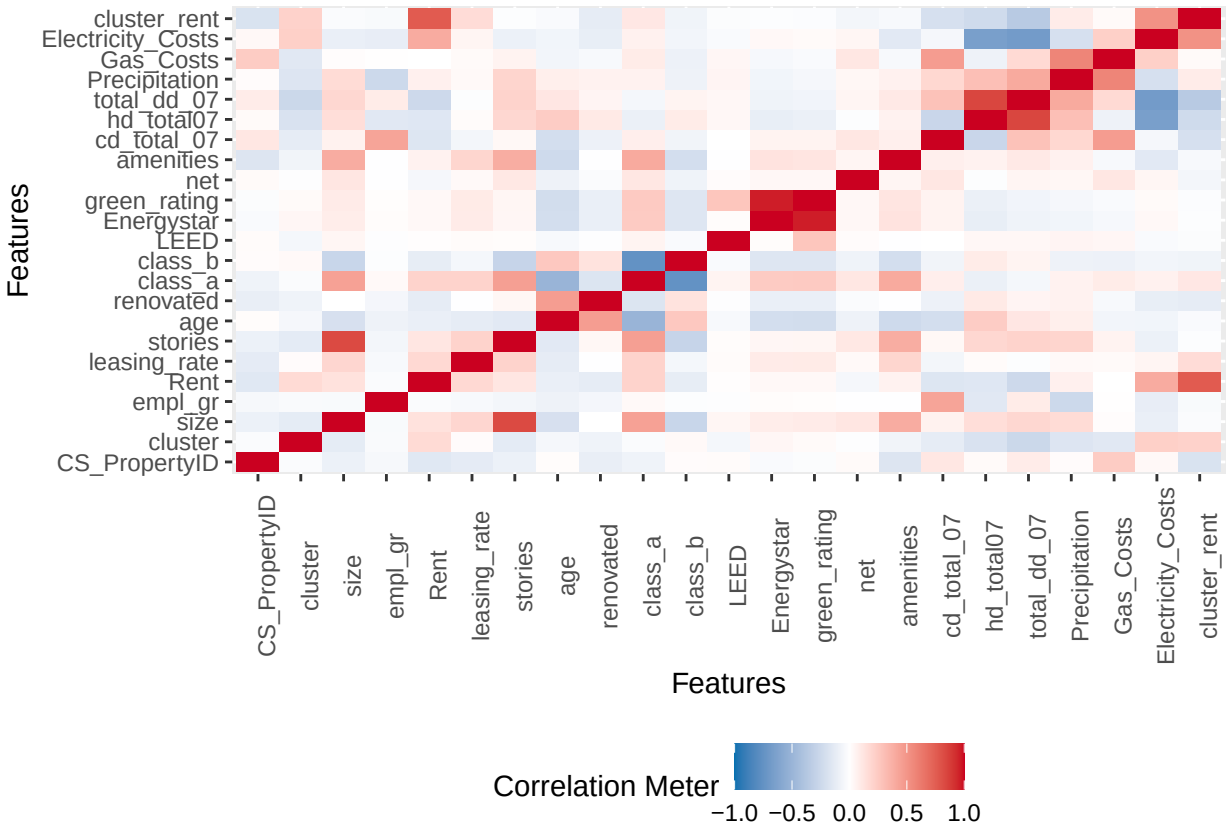
```
## [1] "Lets check if we have any missing values"
```



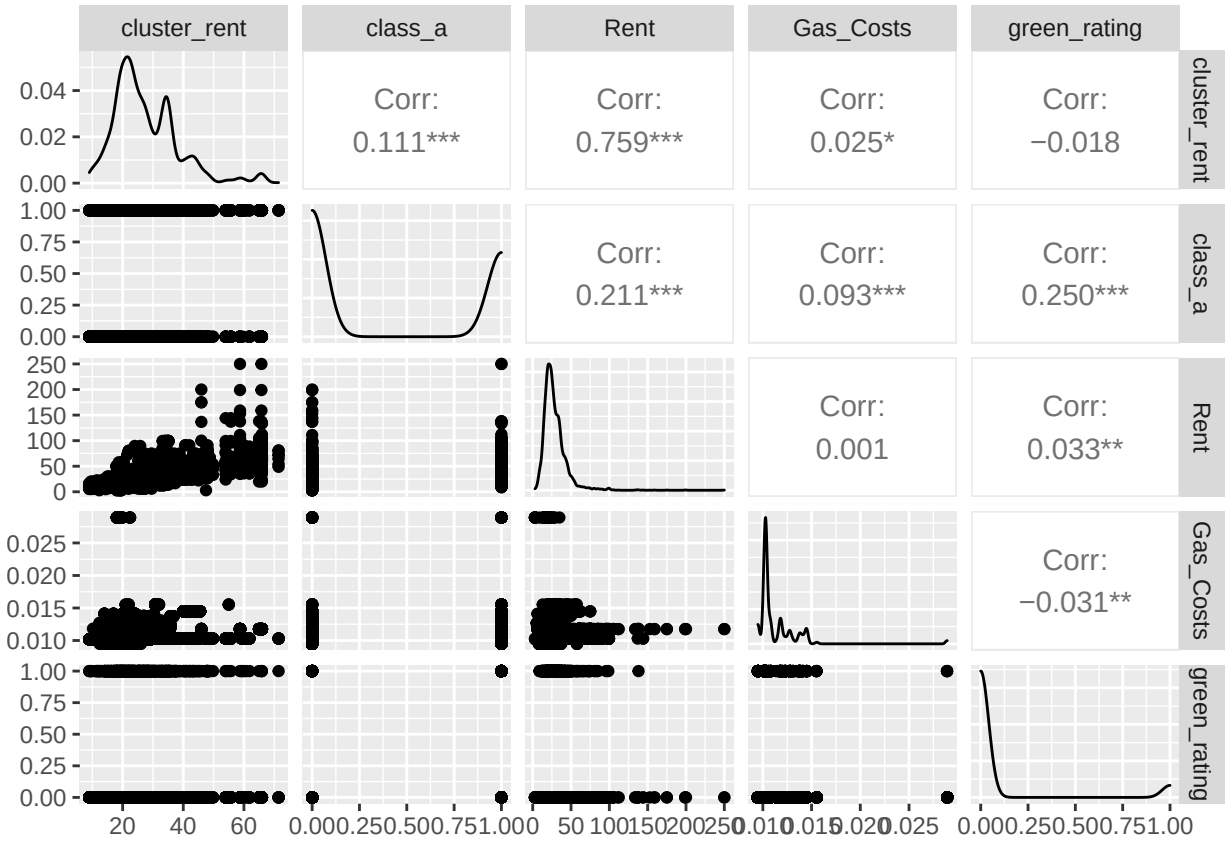

```
## [1] "Lets look at the distribution of the data"
```



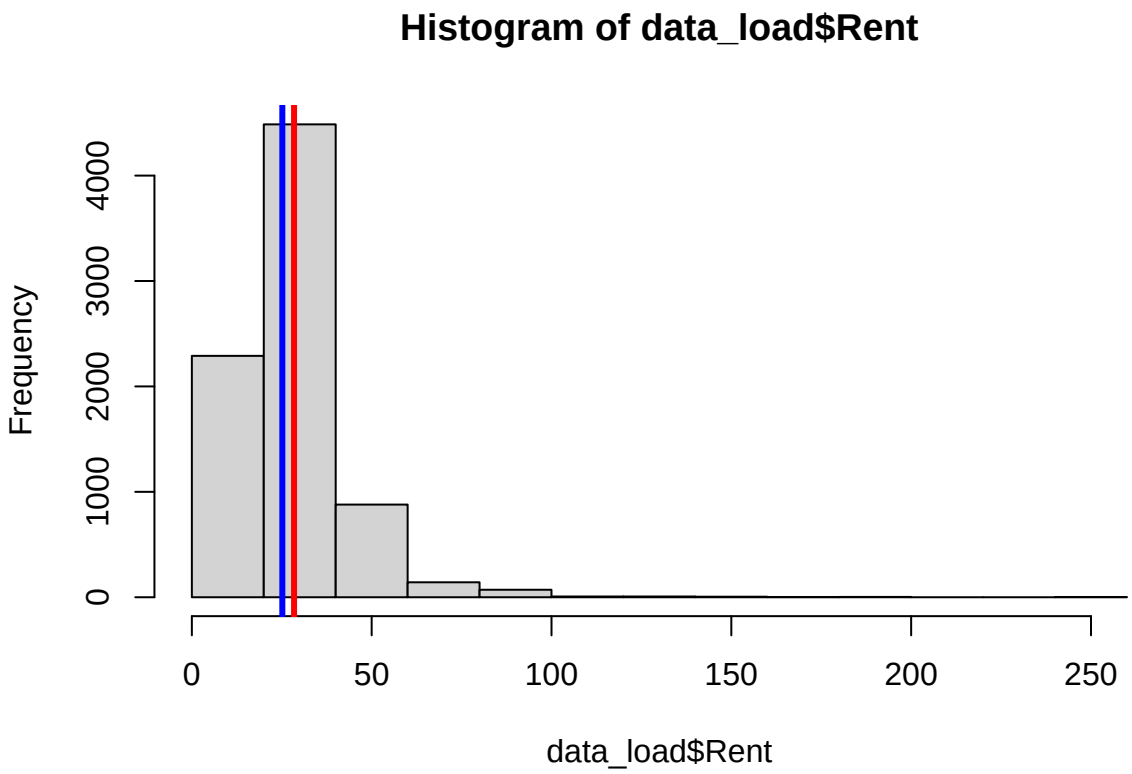

```
## [1] "Lets look at the correltion between these variables"
```



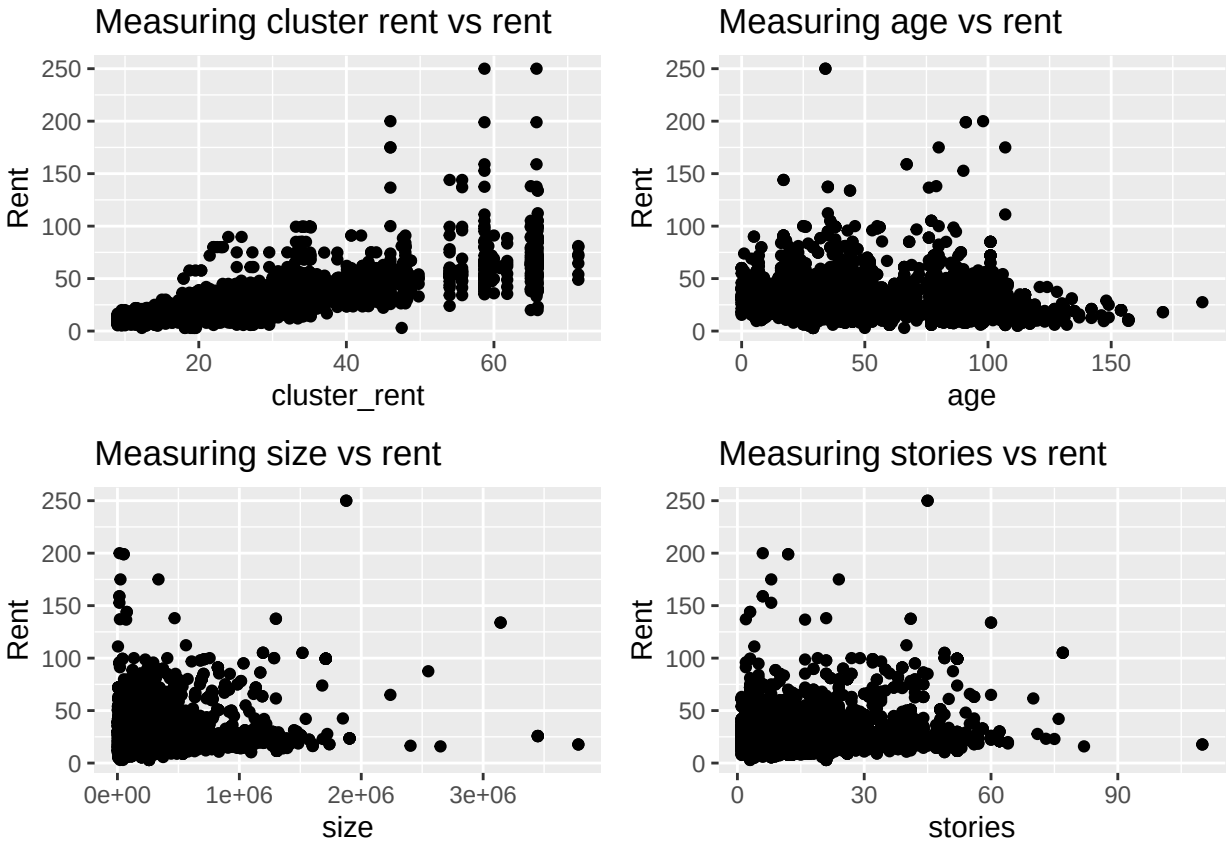

```
## [1] "Lets look at the pairwise correltion between these variables"
```




```
## [1] "Lets look at the mean and median rent in the data"
```

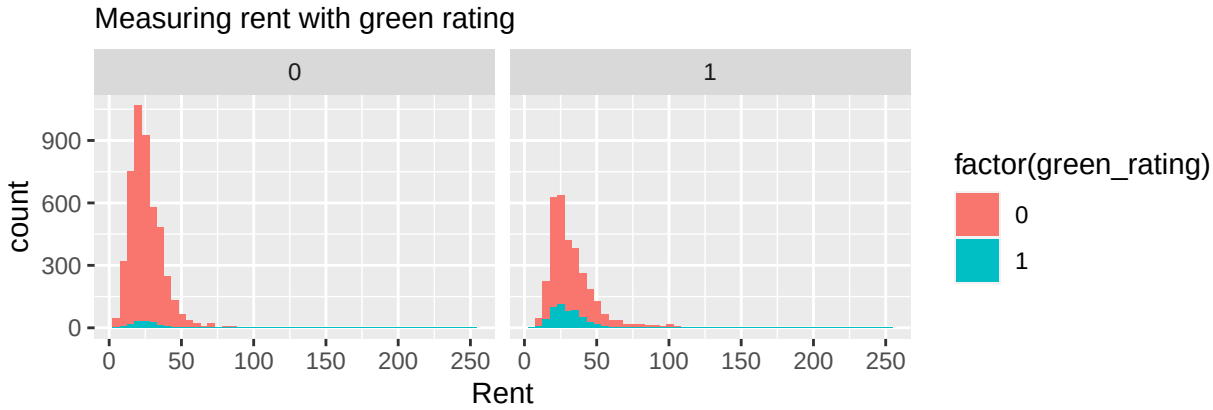



```
## [1] "Lets look at the distribution of these features across various metrics"
```



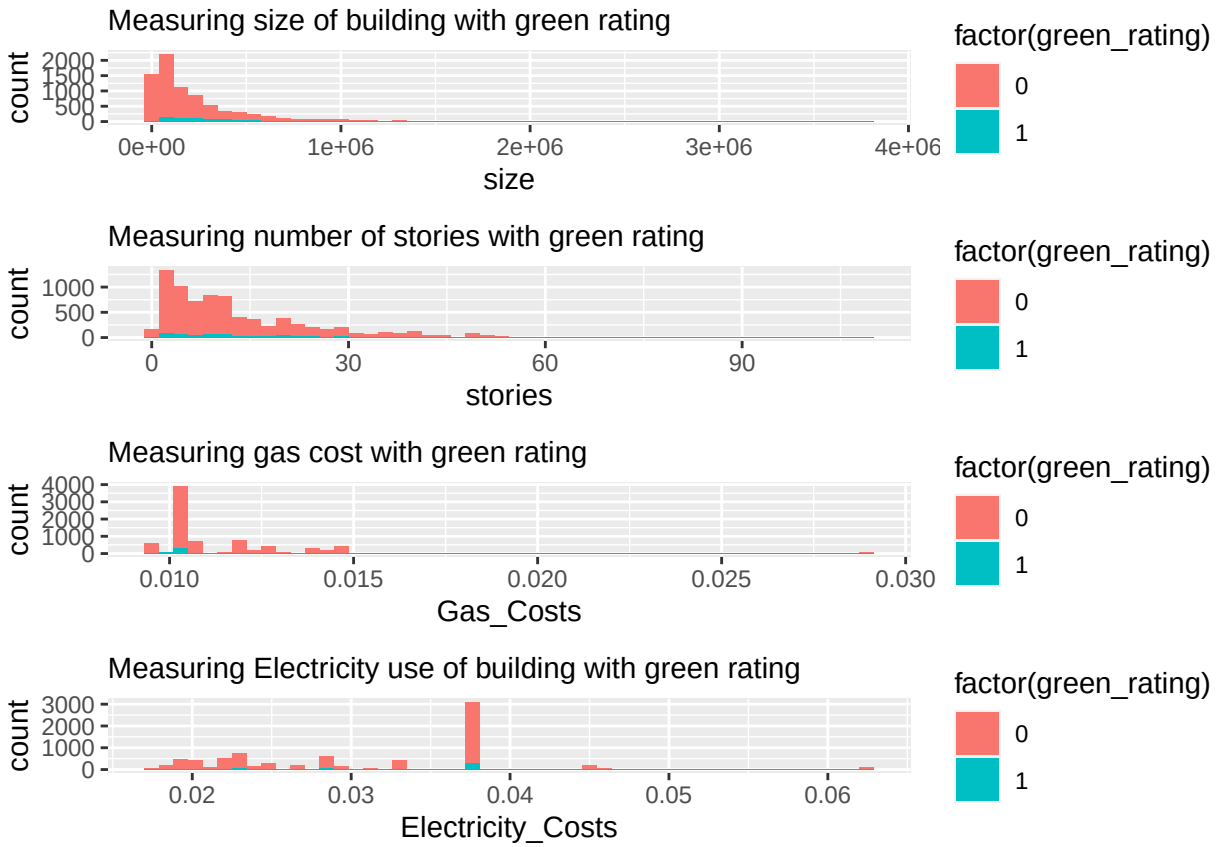

```
## [1] "we observed that Rent is correlated with the cluster rent, size, class A. Additionally, Class a
```

```
## [1] "Lets look at the impact of Class A buildings on rent of green rated houses"
```

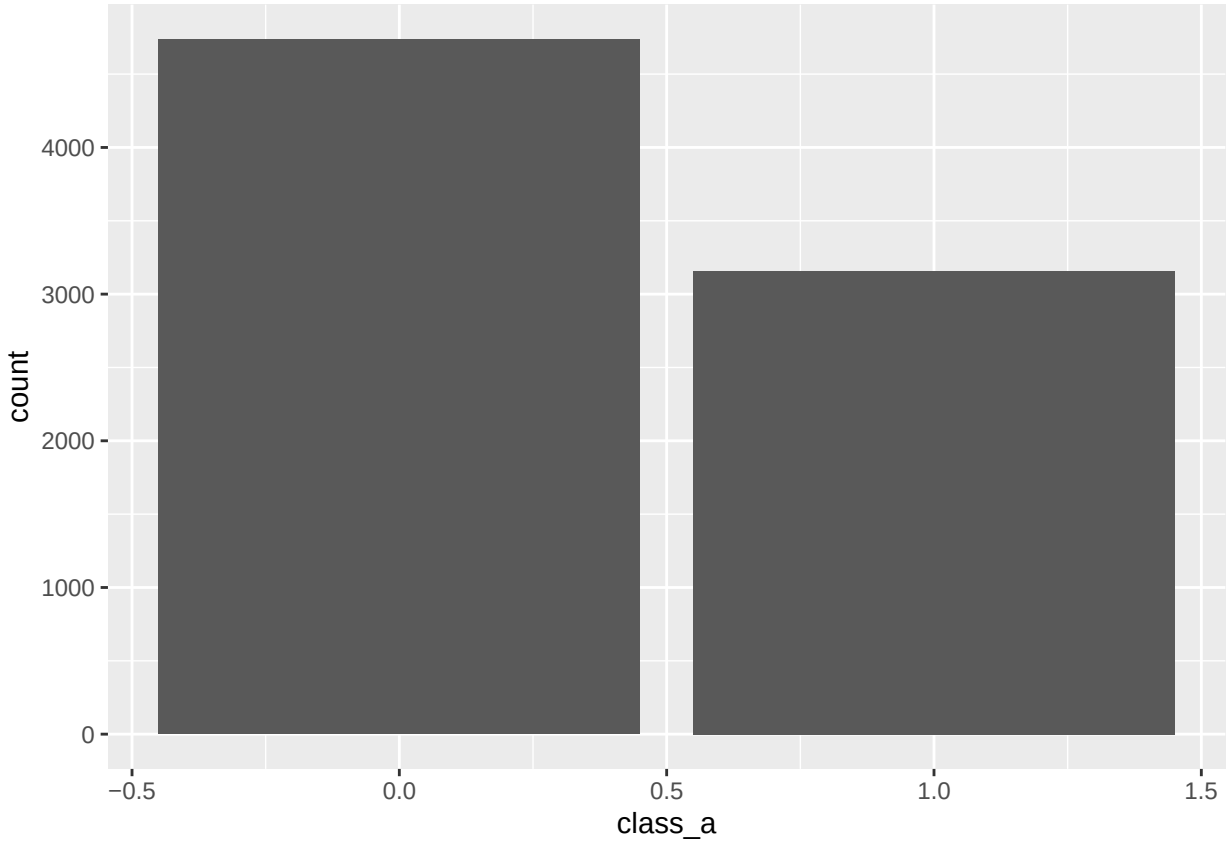


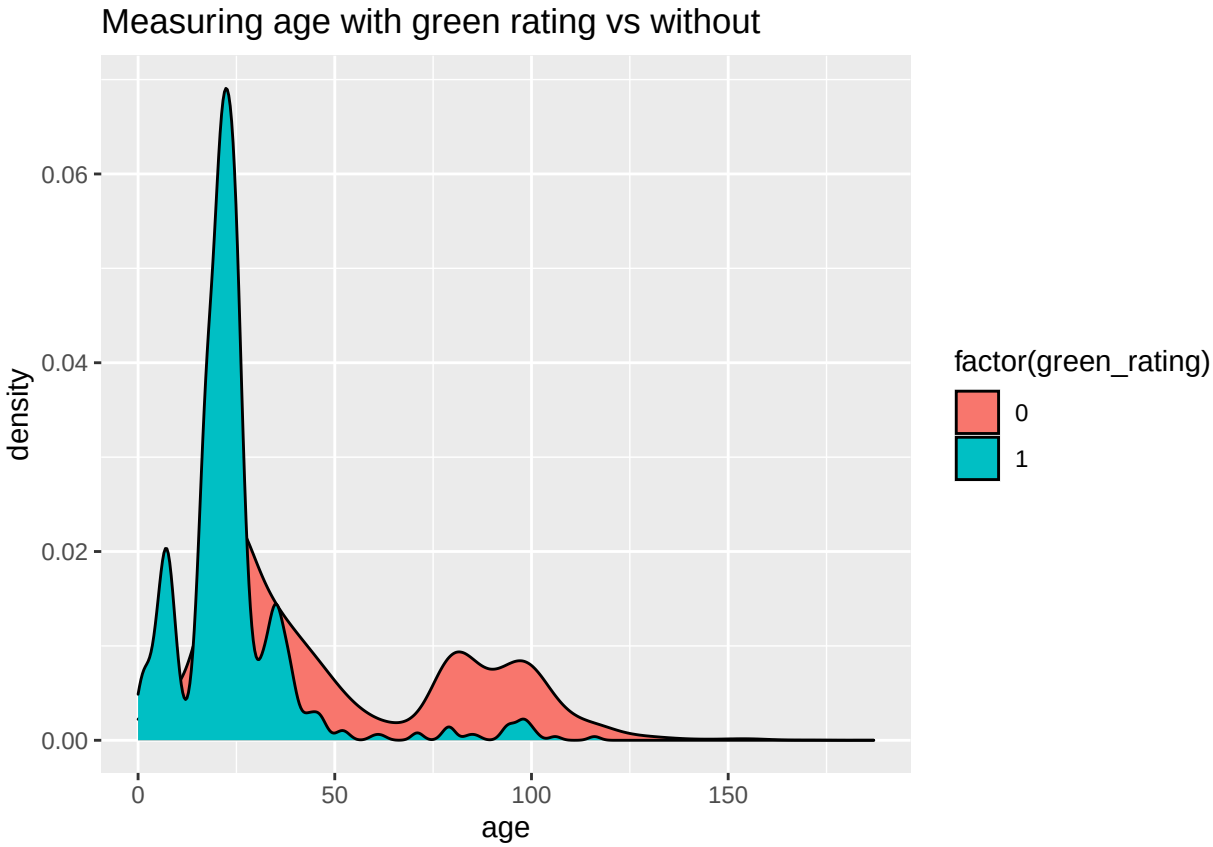

```
## [1] "we can observe that the class A buildings have better rent if they are green rated here"
```

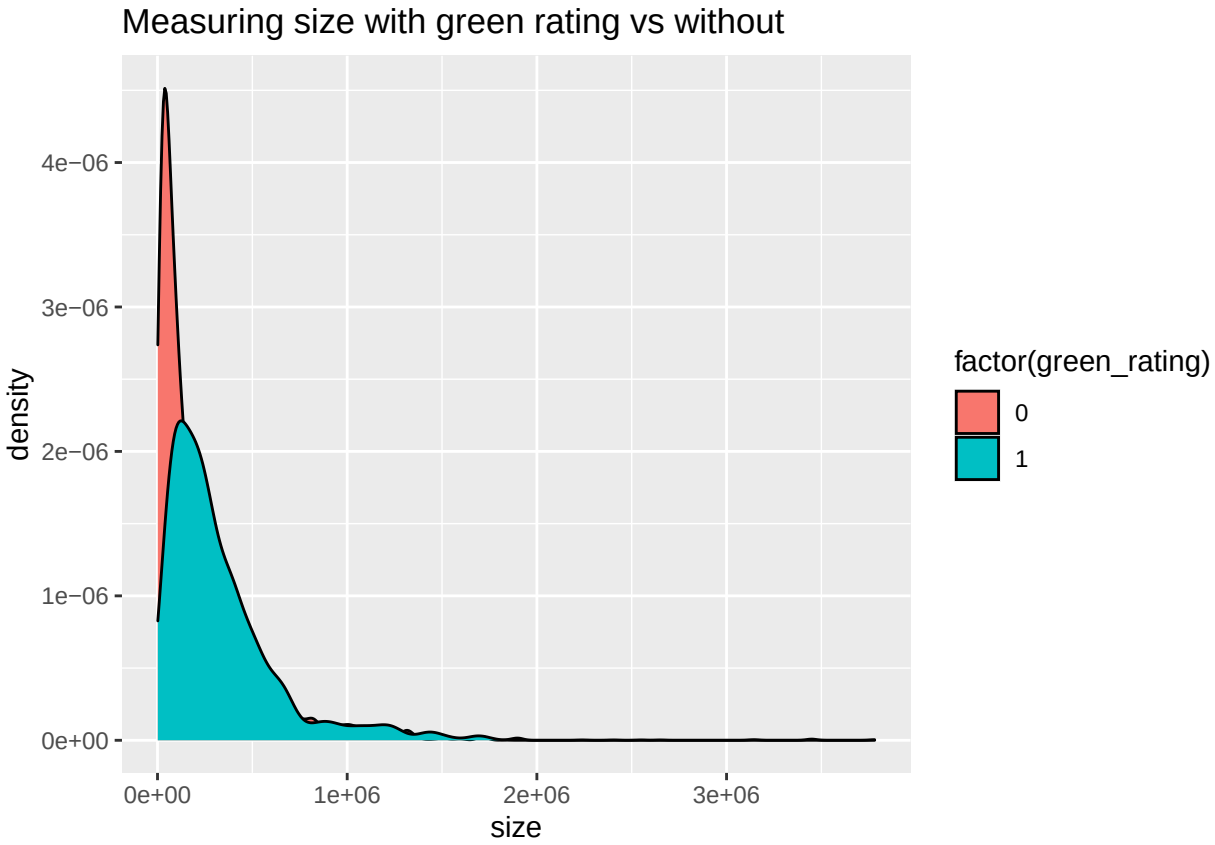
```
## [1] "Lets look at the consumption of resources for green buildings"
```



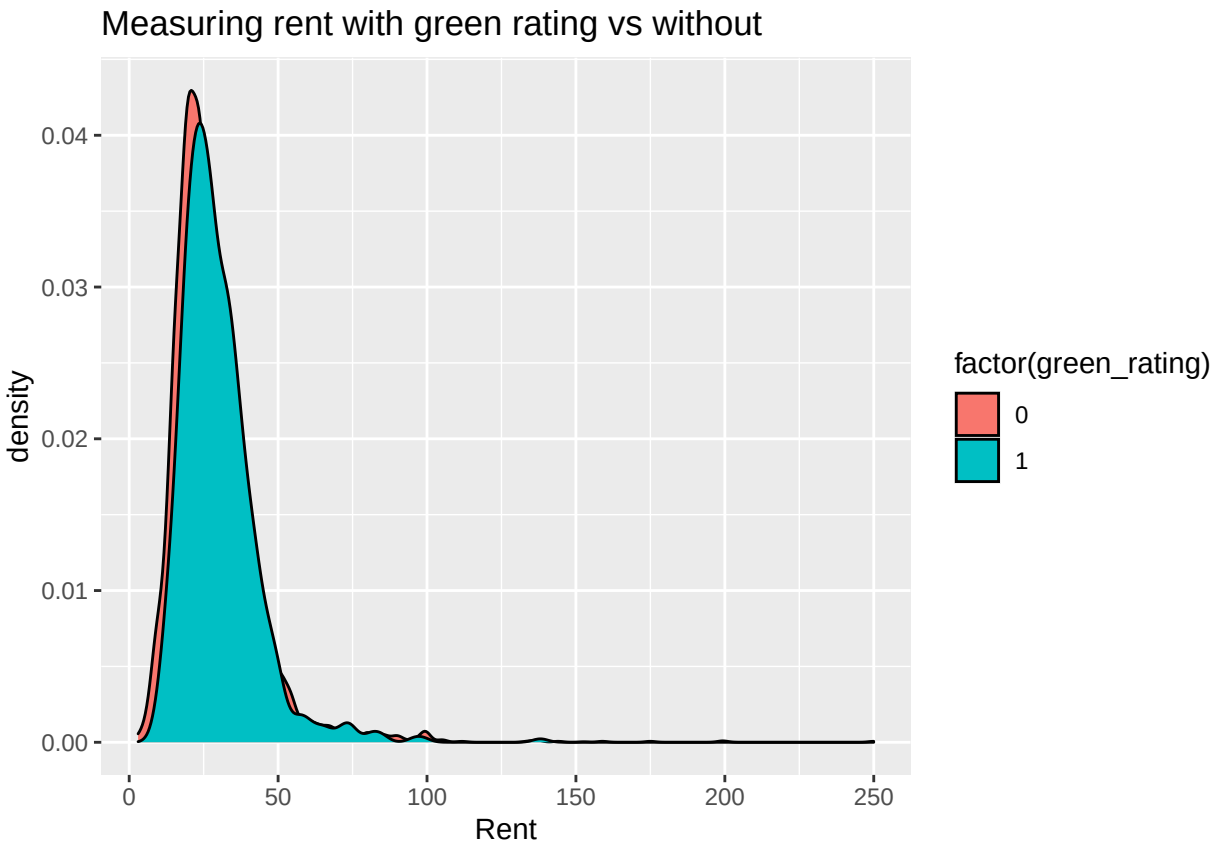

```
## [1] "Lets look at the density of green rated houses with various metrics"
```


4000
3000
2000
1000
0
count
−0.5
0.0
0.5
1.0
1.5
class_a


Measuring age with green rating vs without
density
0.06
0.04
0.02
0.00
0
50
100
150
age
factor(green_rating)
0
1


Measuring size with green rating vs without
density
4e−06
3e−06
2e−06
1e−06
0e+00
0e+00
1e+06
2e+06
3e+06
size
factor(green_rating)
0
1


Measuring cluster_rent with green rating vs without
density
0.04
0.02
0.00
20
40
60
cluster_rent
factor(green_rating)
0
1



We observed most of the green buildings are younger than non-green buildings and the proportion of class a buildings is higher in green buildings

Since Stats Guru fails to account for all factors that affect rent, his analysis is wrong In order to calculate the returns, he began by using the median rent for all buildings.Because of this, he fails to factor in other factors, such as the size and class of the buildings,into his analysis. For instance, we have a class A building will yield a higher rent than a non-green building.

Visual story telling part 2: Capital Metro data

The file capmetro_UT.csv contains data from Austin's own Capital Metro bus network, including shuttles to, from, and around the UT campus. These data track ridership on buses in the UT area. Ridership is measured by an optical scanner that counts how many people embark and alight the bus at each stop. Each row in the data set corresponds to a 15-minute period between the hours of 6 AM and 10 PM, each and every day, from September through November 2018. The variables are:

- *timestamp*: the beginning of the 15-minute window for that row of data
- *boarding*: how many people got on board any Capital Metro bus on the UT campus in the specific 15 minute window
- *alighting*: how many people got off ("alit") any Capital Metro bus on the UT campus in the specific 15 minute window
- *day_of_week* and *weekend*: Monday, Tuesday, etc, as well as an indicator for whether it's a weekend.
- *temperature*: temperature at that time in degrees F
- *hour_of_day*: on 24-hour time, so 6 for 6 AM, 13 for 1 PM, 14 for 2 PM, etc.
- *month*: July through December

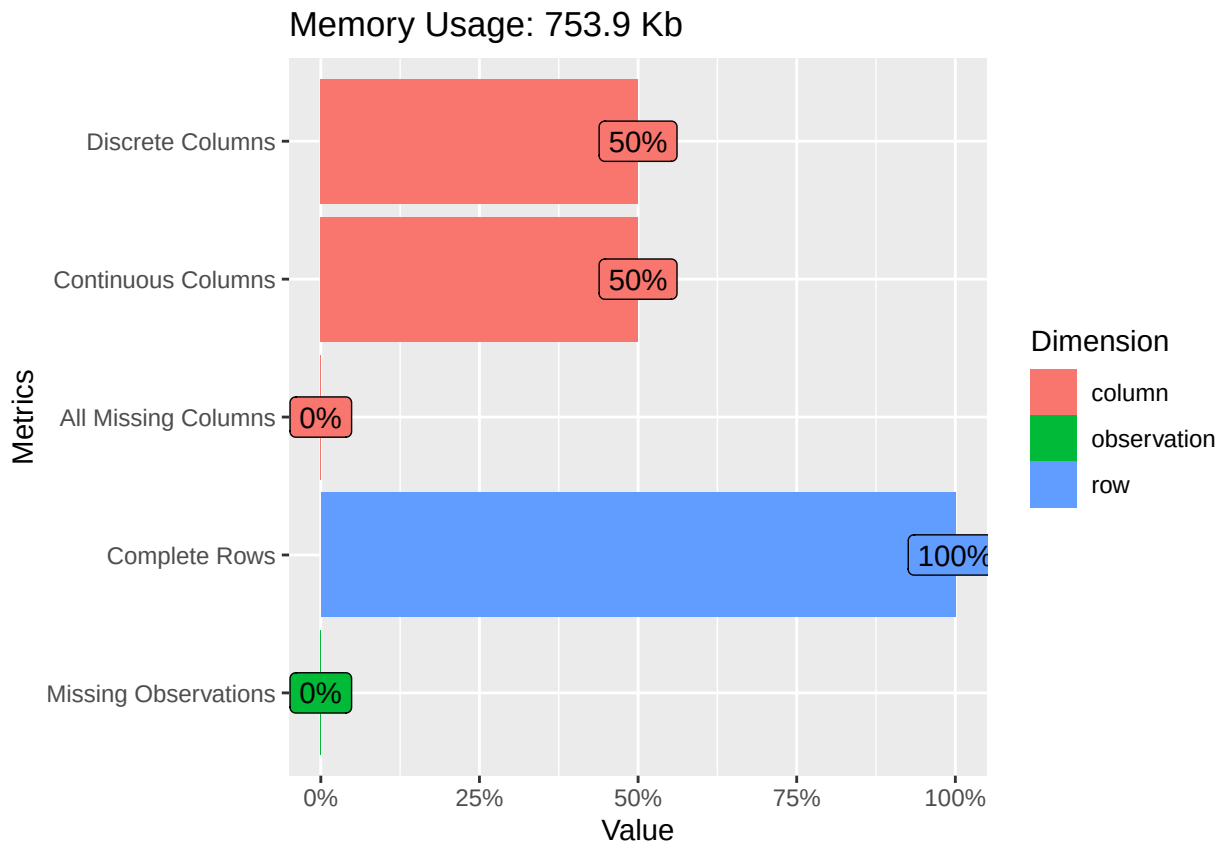
Your task is to create a figure, or set of related figures, that tell an interesting story about Capital Metro ridership patterns around the UT-Austin campus during the semester in question. Provide a clear annotation/caption for each figure, but the figure(s) should be more or less stand-alone, in that you shouldn't need many, many paragraphs to convey its meaning. Rather, the figure together with a concise caption should speak for itself as far as possible.

You have broad freedom to look at any variables you'd like here – try to find that sweet spot where you're showing genuinely interesting relationships among more than just two variables, but where the resulting figure or set of figures doesn't become overwhelming/confusing. (Faceting/panel plots might be especially useful here.)

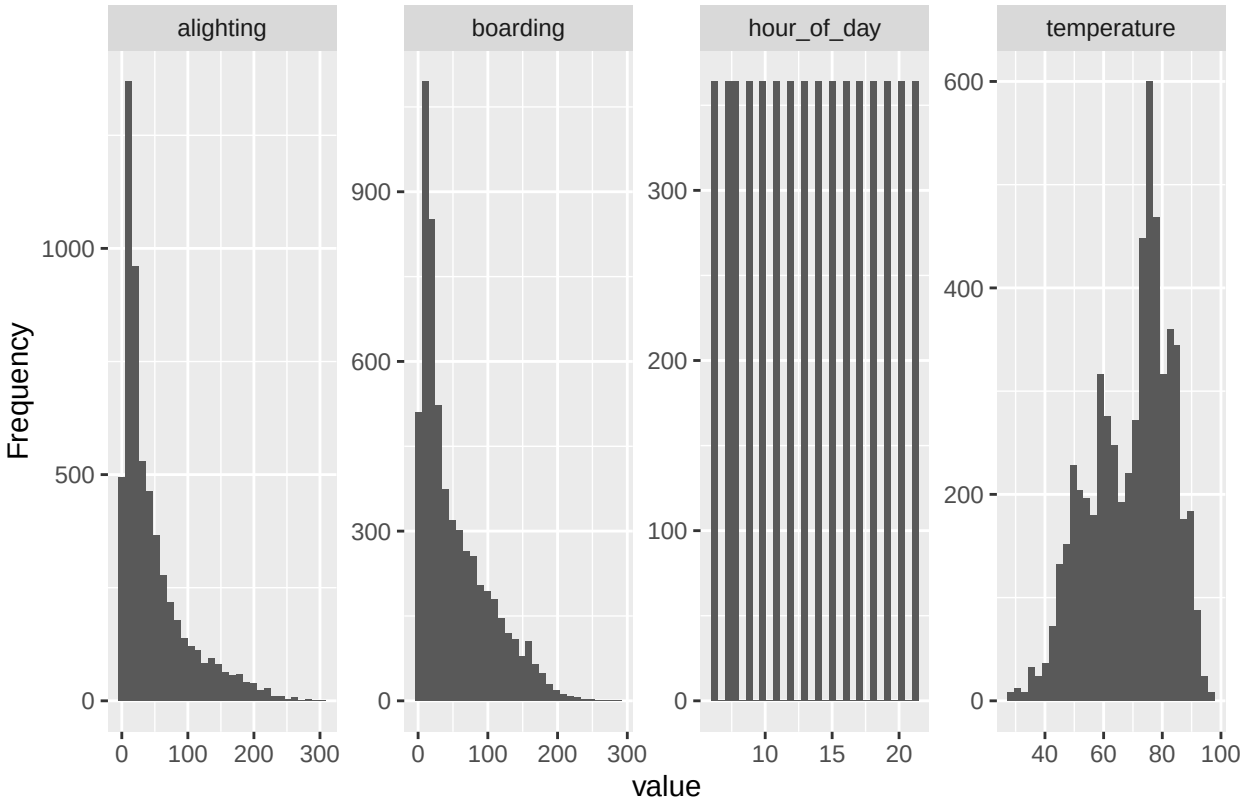
Solution:

To figure out interesting Capital Metro ridership patterns around the UT-Austin we started with loading the data and then looking at various aspects of the same. We focused on weekly, monthly ridership and also the riding trends during the day.

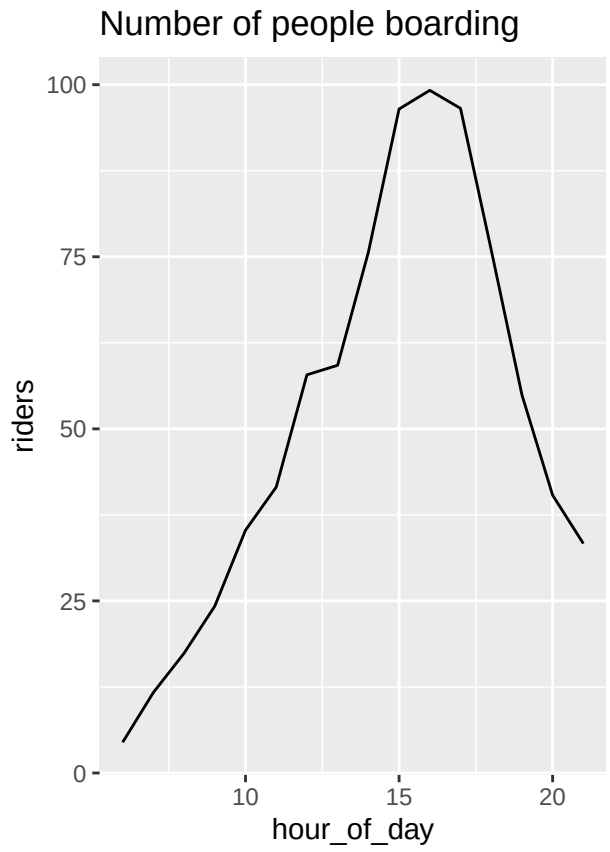
```
## [1] "Lets check if we have any missing values"
```

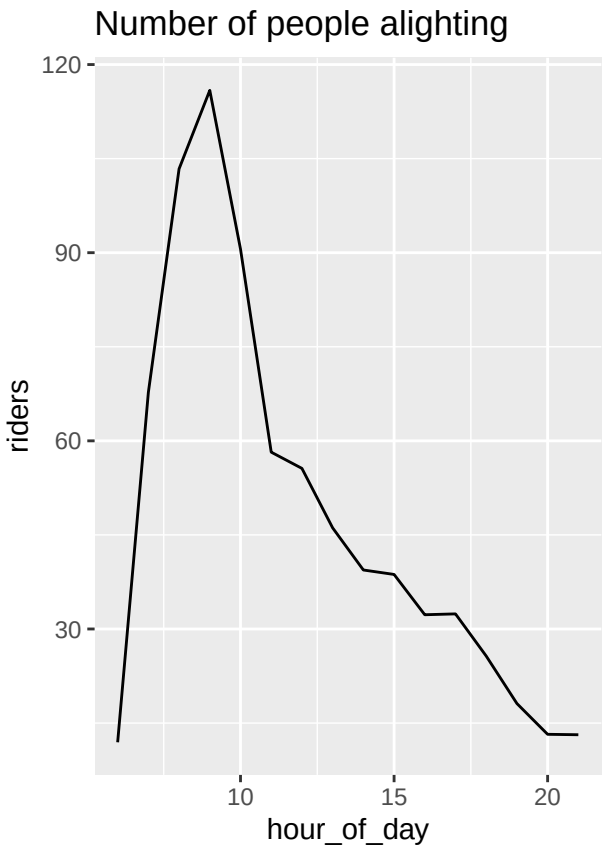



```
## [1] "We can now plot the distribution of this data"
```

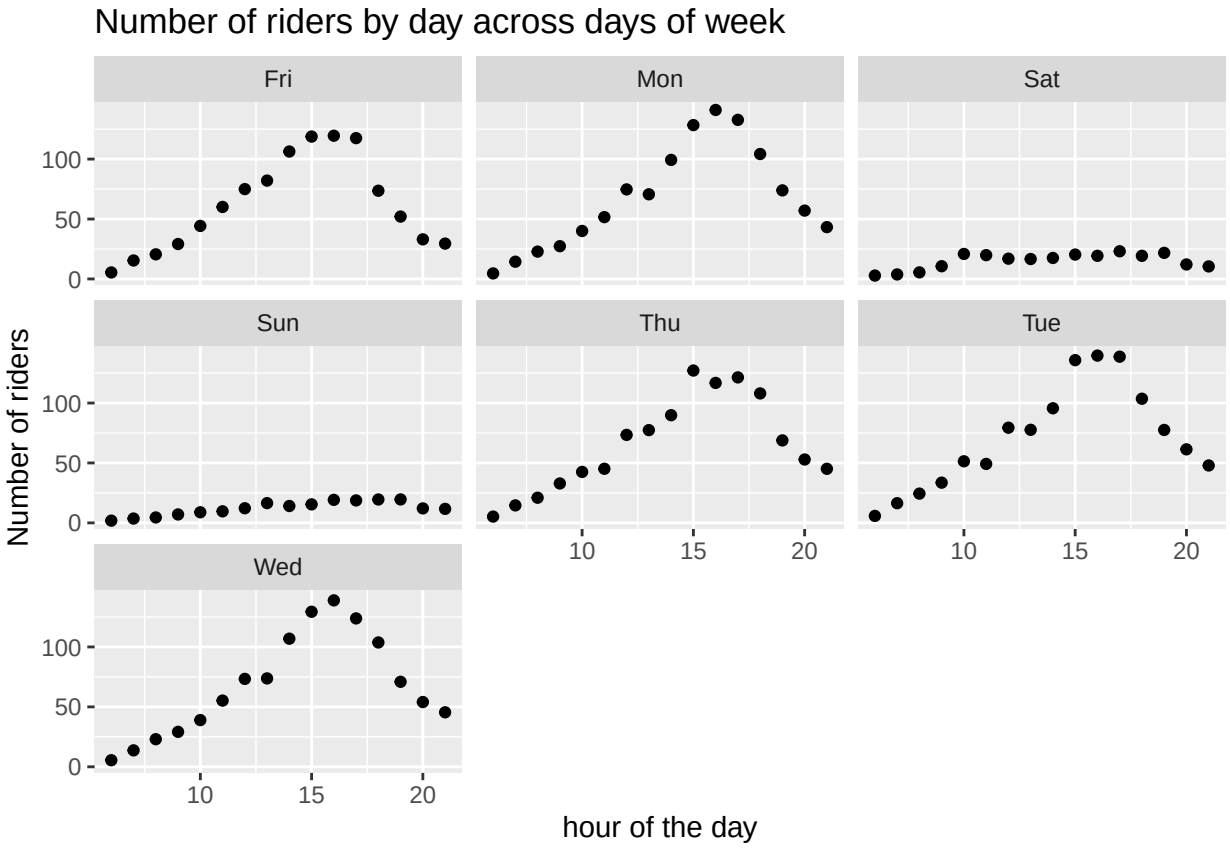



```
## [1] "lets look at the hourly trend of ridership"
```

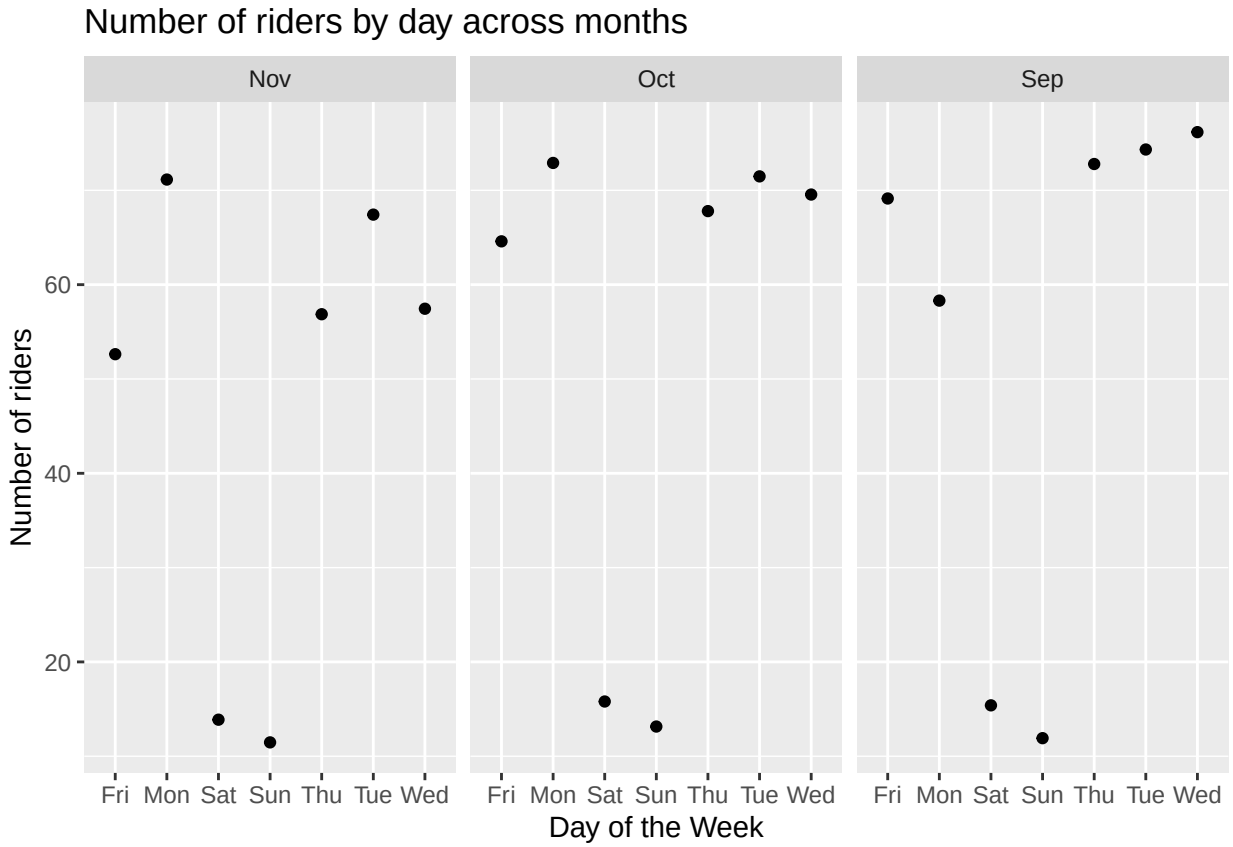


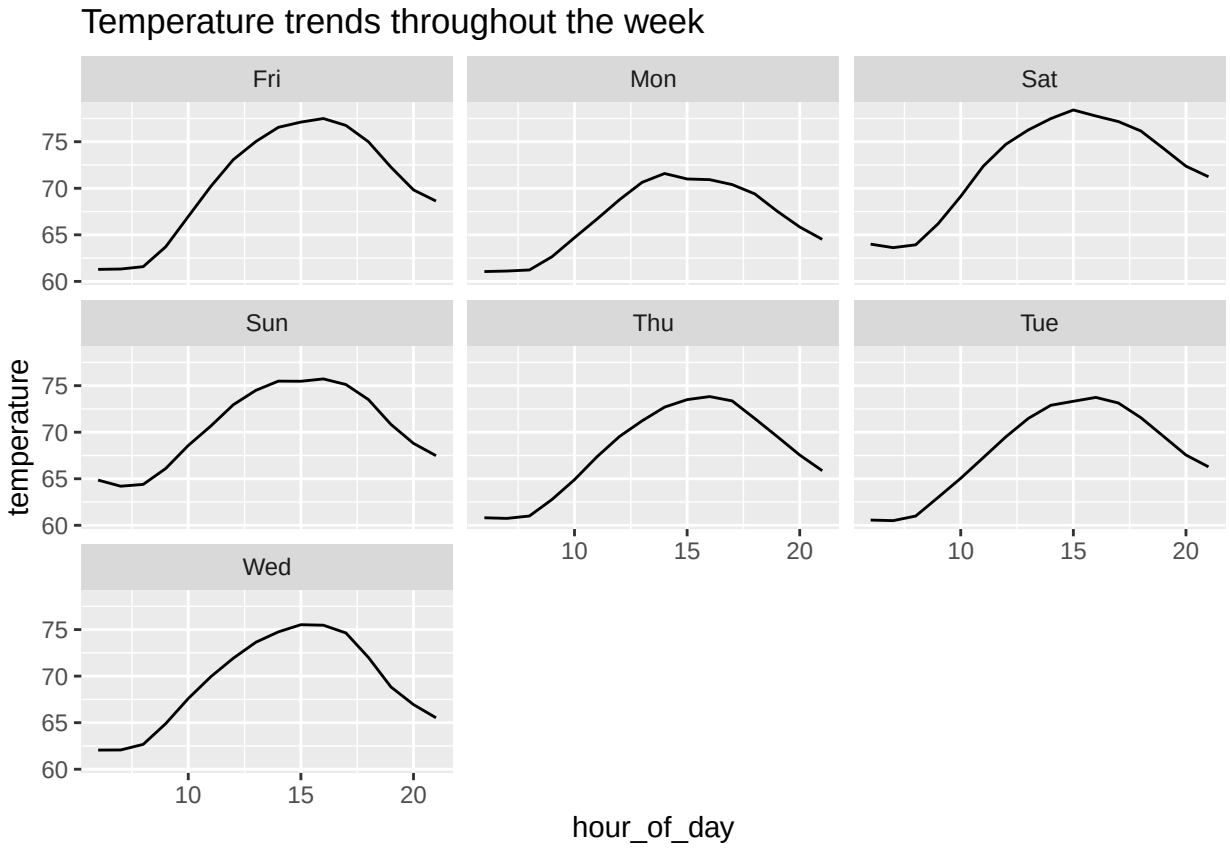


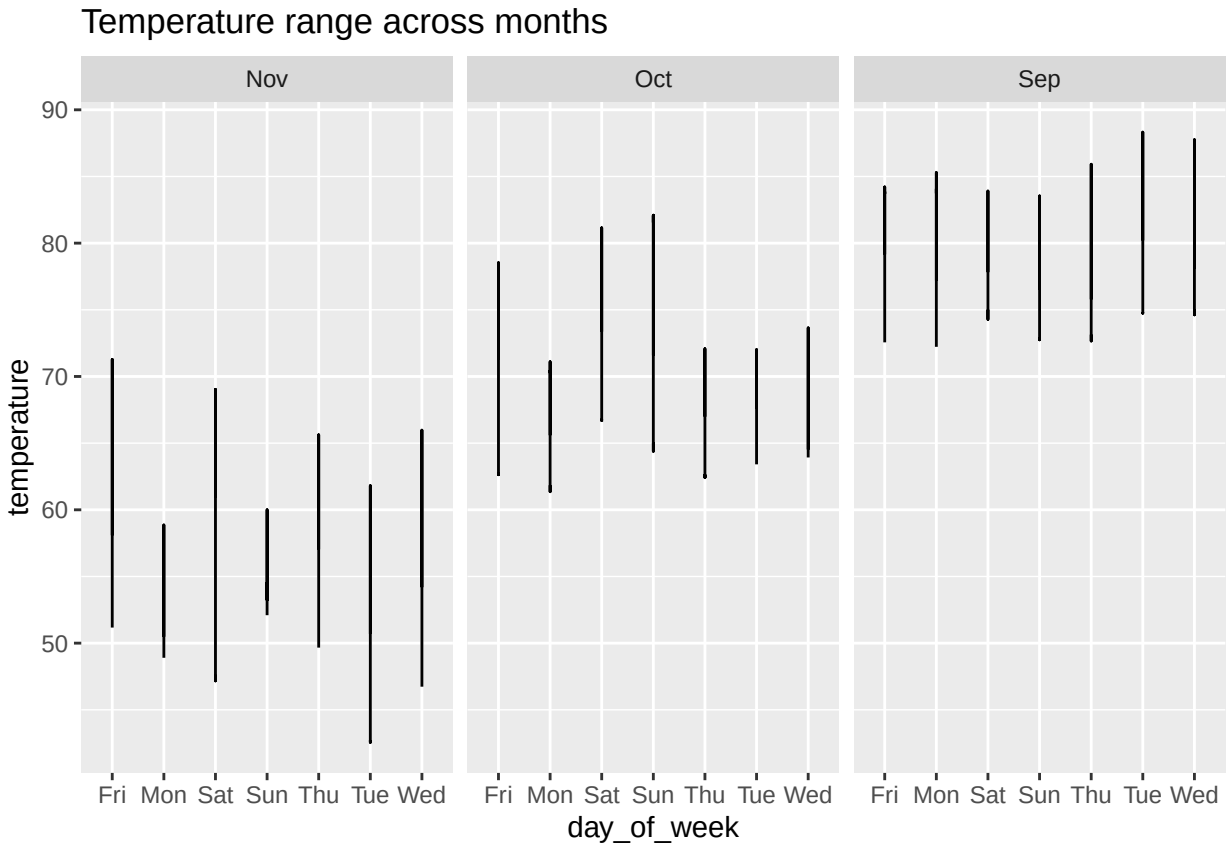

```
## [1] "Lets look at the weekly trend of ridership"
```




```
## [1] "lets look at the monthly trend of ridership"
```


Number of riders by day across months
Nov
Oct
Sep
60
40
20
Number of riders
Fri Mon Sat Sun Thu Tue Wed
Day of the Week

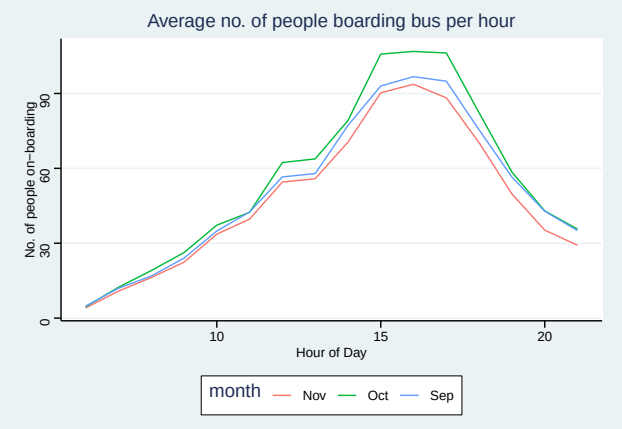

Temperature trends throughout the week
Fri
Mon
Sat
Sun
Thu
Tue
Wed
temperature
hour_of_day
75
70
65
60
10
15
20

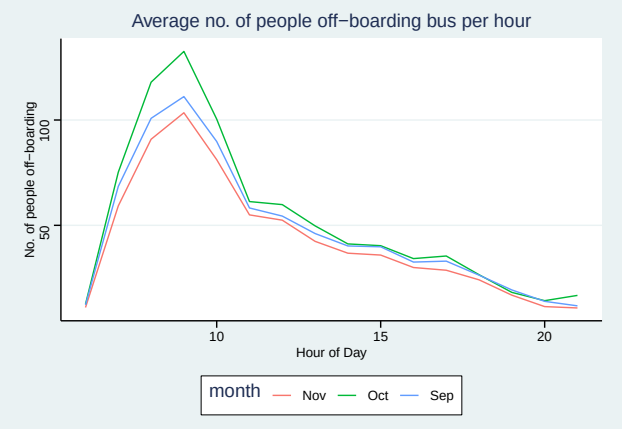


We looked at the daily, weekly and monthly distributions of the data and saw interesting trends like more people leave the bus during the early part of the day and interestingly the temperature of day can also be measure from this data. We also looked at monthly trends in ridership across the months of Oct, Sep and Nov. We also looked at temperature trends and ridership throughout the week.

Let's also look at hour-wise traffic distribution for months from July to December -

```
##   timestamp            boarding        alighting      day_of_week
##  Length:5824        Min.   :  0.00   Min.   :  0.00   Length:5824
##  Class :character   1st Qu.: 13.00   1st Qu.: 13.00   Class :character
##  Mode  :character   Median : 33.00   Median : 28.00   Mode  :character
##                     Mean   : 51.51   Mean   : 47.65
##                     3rd Qu.: 79.25   3rd Qu.: 64.00
##                     Max.   :288.00   Max.   :304.00
##   temperature     hour_of_day       month             weekend
##  Min.   :29.18   Min.   : 6.00   Length:5824        Length:5824
##  1st Qu.:59.20   1st Qu.: 9.75   Class :character   Class :character
##  Median :72.75   Median :13.50   Mode  :character   Mode  :character
##  Mean   :69.28   Mean   :13.50
##  3rd Qu.:79.29   3rd Qu.:17.25
##  Max.   :97.64   Max.   :21.00
```



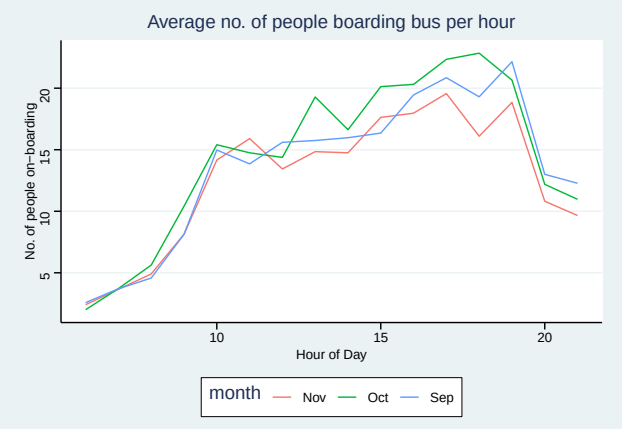


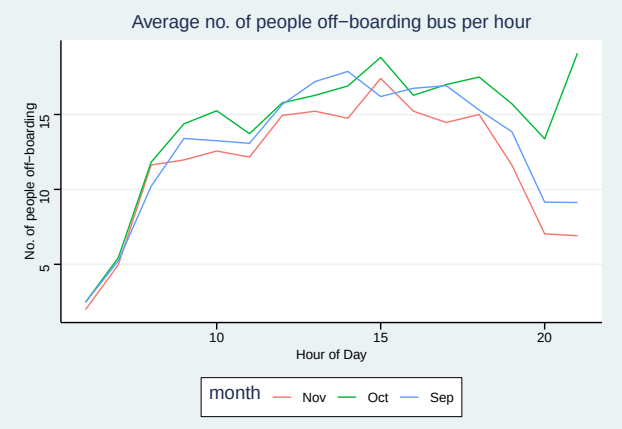
There are some weird discrepancies if you look at it without thinking much.

We can make the following observations -

- The distribution for no. of people on-boarding and off-boarding doesn't change much in any month
- 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
- 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 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.

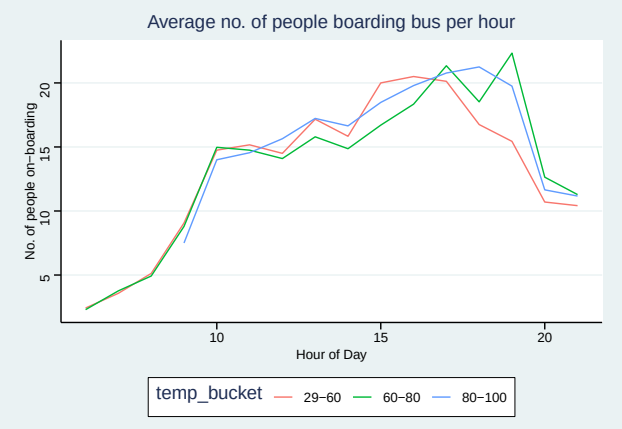
This distribution seems to be heavily influenced by students going to and fro from campus for college. Let's try to look at the distribution on the weekends -

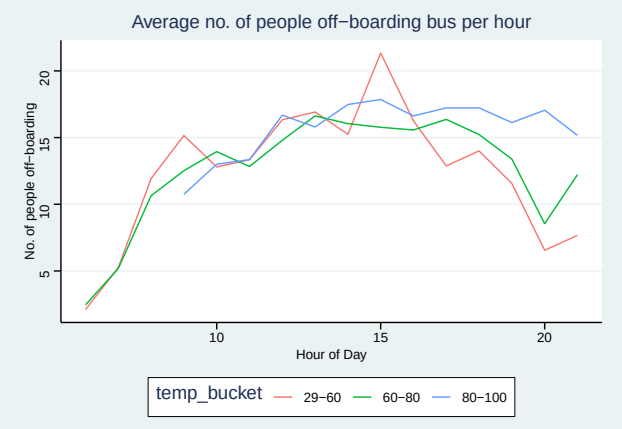




- The counts are much more varying throughout the day now
- There is an interesting spike in no. of off-boarding people in October towards the end of the day. Maybe this is because October is when students have their mid-terms for the semester so they tend to stay late on campus and go home during late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

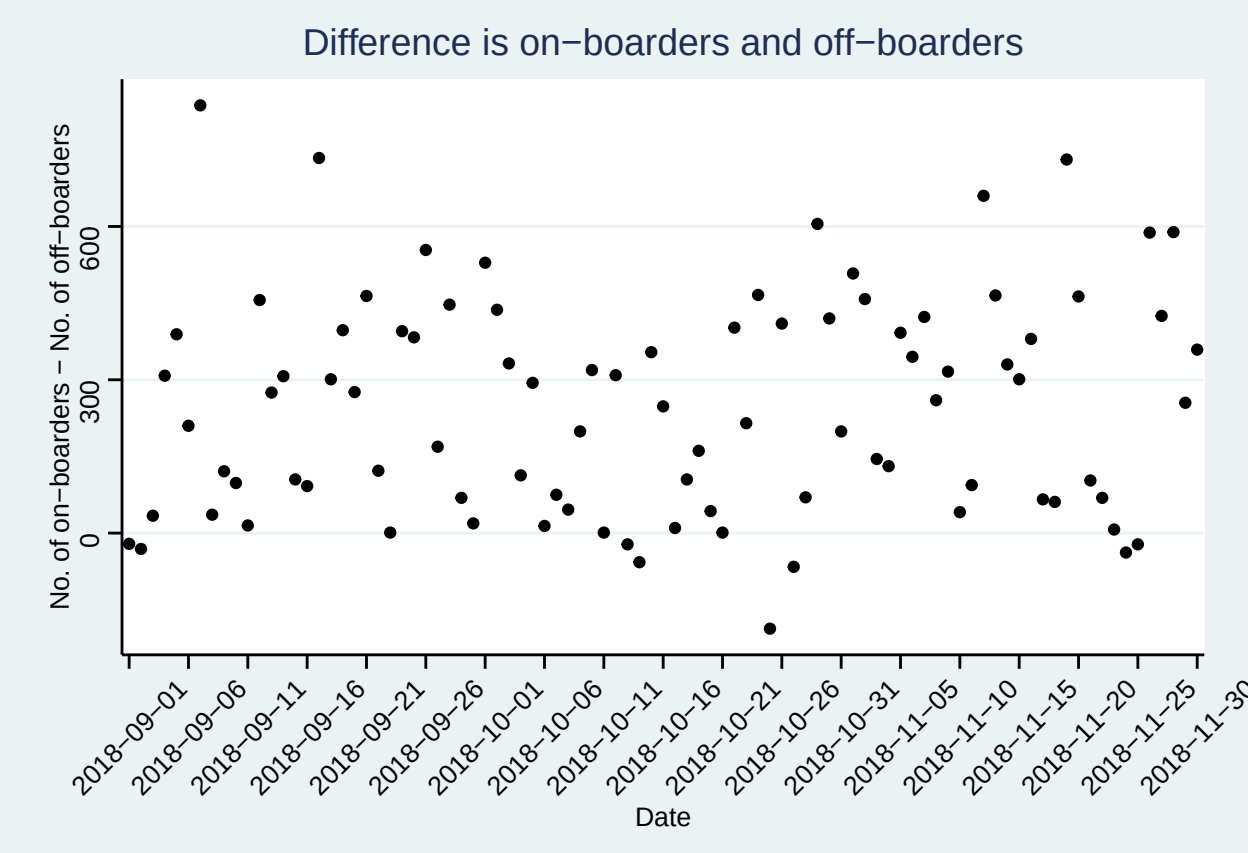
Let's also look at how weekend ridership changes based on temperature (since the weekday ridership is expected to not be affected by temperature since students have to go to college regardless) -





- 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 -



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!

Portfolio modeling

Background

In this problem, you will construct three different portfolios of exchange-traded funds, or ETFs, and use bootstrap resampling to analyze the short-term tail risk of your portfolios. If you're unfamiliar with exchange-traded funds, you can read a bit about them here.

The goal

Suppose you have $100,000 in capital. Your task is to:
- Construct two different possibilities for an ETF-based portfolio, each involving an allocation of your $100,000 in capital to somewhere between 3 and 10 different ETFs. You can find a big database of ETFs here.
- Download the last five years of daily data on your chosen ETFs, using the functions in the `quantmod` package, as we used in class. Note: make sure to choose ETFs for which at least five years of data are available. There are tons of ETFs and some are quite new!
- Use bootstrap resampling to estimate the 4-week (20 trading day) value at risk of each of your three portfolios at the 5% level.
- Write a report summarizing your portfolios and your VaR findings.

You should assume that your portfolios are rebalanced each day at zero transaction cost. For example, if you're allocating your wealth evenly among 5 ETFs, you always redistribute your wealth at the end of each day so that the equal five-way split is retained, regardless of that day's appreciation/depreciation.

Notes: - Make sure the portfolios are different from each other! (Maybe one seems safe, another aggressive, or something like that.) You're not being graded on what specific portfolios you choose... just provide some context for your choices.

Solution:

Let's create a 5-ETF portfolio with 3 ETFs from the Oil sector and hedge them with 2 Green Energy ETFs. How would such a portfolio perform throughout the past 5 years?

Let's construct it out of the following ETFs-

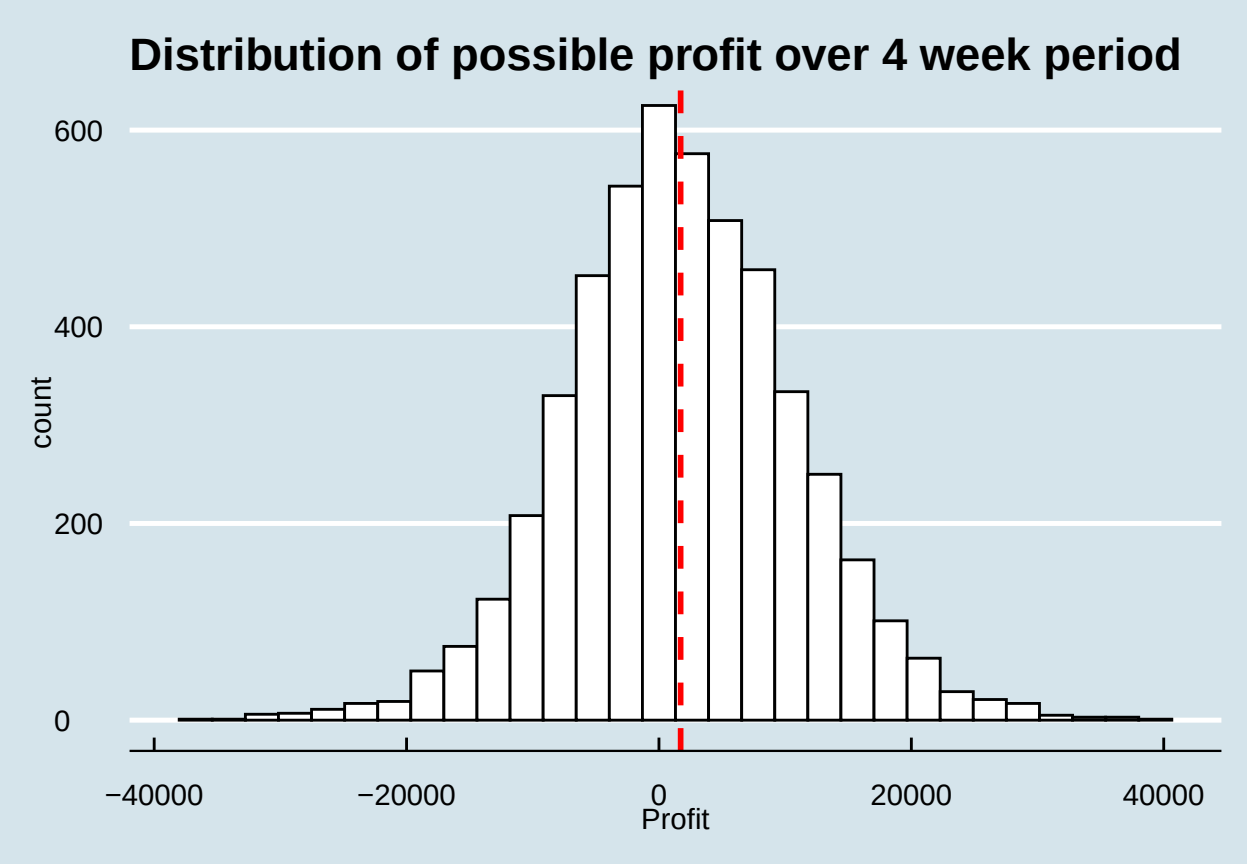
1) BNO: United States Brent Oil Fund LP
2) IEO: iShares US Oil & Gas Exploration & Production ETF
3) IEZ: iShares U.S. Oil Equipment & Services ETF
4) ICLN: iShares Global Clean Energy ETF
5) TAN: Invesco Solar ETF

Let's assign $20,000 to each ETF and look at what our 4 week return distribution looks like when bootstrapped for 5000 iterations -

```
## [1] "BNO"  "IEO"  "TAN"  "IEZ"  "ICLN"
```

```
##               ClCl.BNOa     ClCl.IEOa    ClCl.IEZa    ClCl.TANa   ClCl.ICLNa
## 2017-08-08           NA            NA           NA           NA           NA
## 2017-08-09  0.013818182 -0.0011148644 -0.010286845 -0.006369427 -0.008743169
```

```
## 2017-08-10 -0.018651363 -0.0167410162 -0.018267717 -0.027472527 -0.013230430
## 2017-08-11  0.005847953  0.0007567348 -0.003849888  0.010828578  0.001117318
## 2017-08-14 -0.026889535 -0.0085066348 -0.003542609  0.011644155  0.008928571
## 2017-08-15  0.004480956 -0.0041944899 -0.013897899  0.009668555  0.002212389
```



What's our Value at Risk for this portfolio for a 4 week period with a 5% confidence?

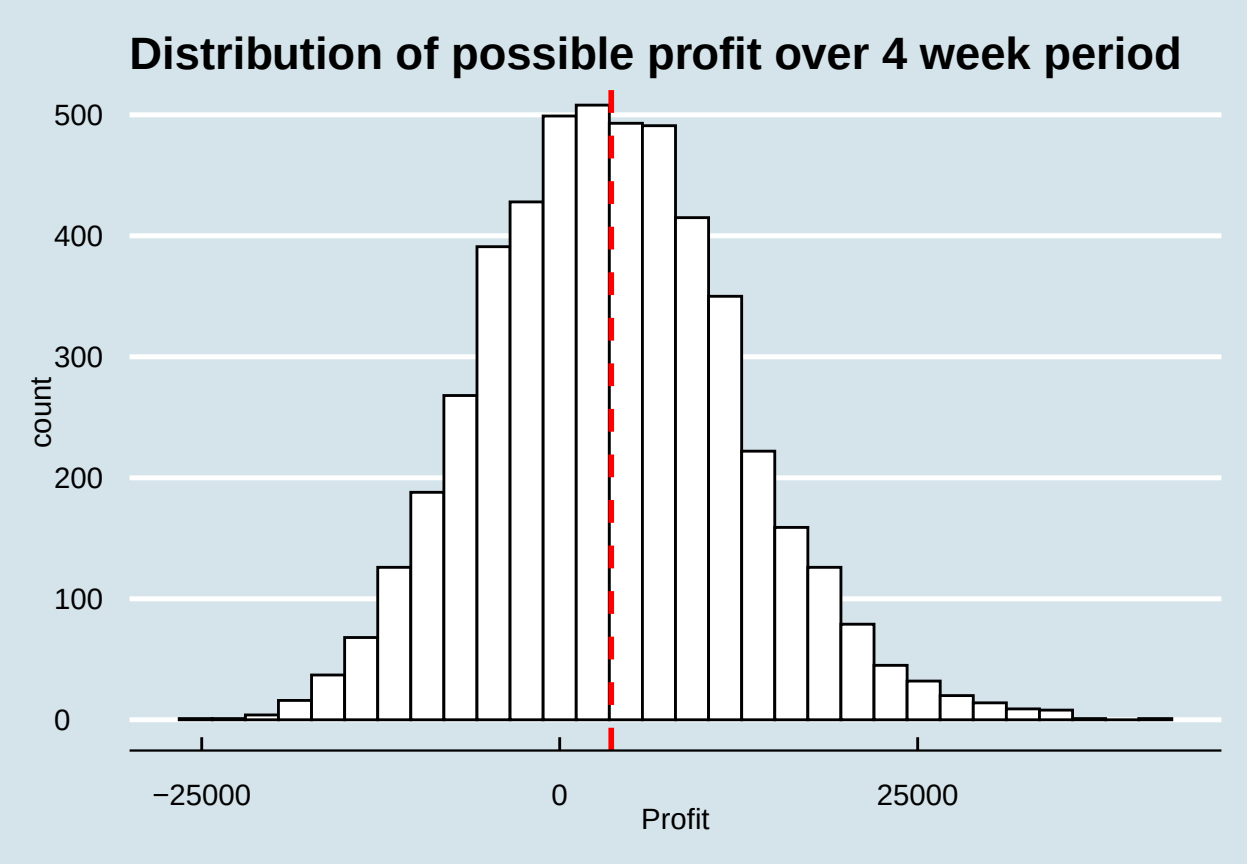
```
##       5%
## -12859.2
```

Quick side note - What if we take data from the past 2 years for this portfolio and check? (given that most of these clean energy ETFs started to pick up the pace from Mid 2020)

Let's look at the profit distributions -

```
## [1] "BNO"  "IEO"  "TAN"  "IEZ"  "ICLN"
```

```
##              ClCl.BNOa    ClCl.IEOa    ClCl.IEZa   ClCl.TANa   ClCl.ICLNa
## 2020-08-10           NA           NA           NA          NA           NA
## 2020-08-11 -0.011996572 -0.014133200  0.009980040 -0.03120123 -0.016311167
## 2020-08-12  0.019080659  0.016968929  0.008893281  0.02878424  0.017219388
## 2020-08-13 -0.004255319 -0.022727188 -0.020568071  0.02543532  0.016927837
## 2020-08-14 -0.005128205  0.017368266  0.008000000 -0.01621826 -0.007398213
## 2020-08-17  0.009450172 -0.006944473 -0.012896825  0.01532190  0.011801304
```



What's our Value at Risk for this portfolio for a 4 week period with a 5% confidence?

```
##        5%
## -10417.46
```

Looks like the Value at Risk (VaR) at 5% confidence decreased by almost 30%!

Okay, sidebar over.

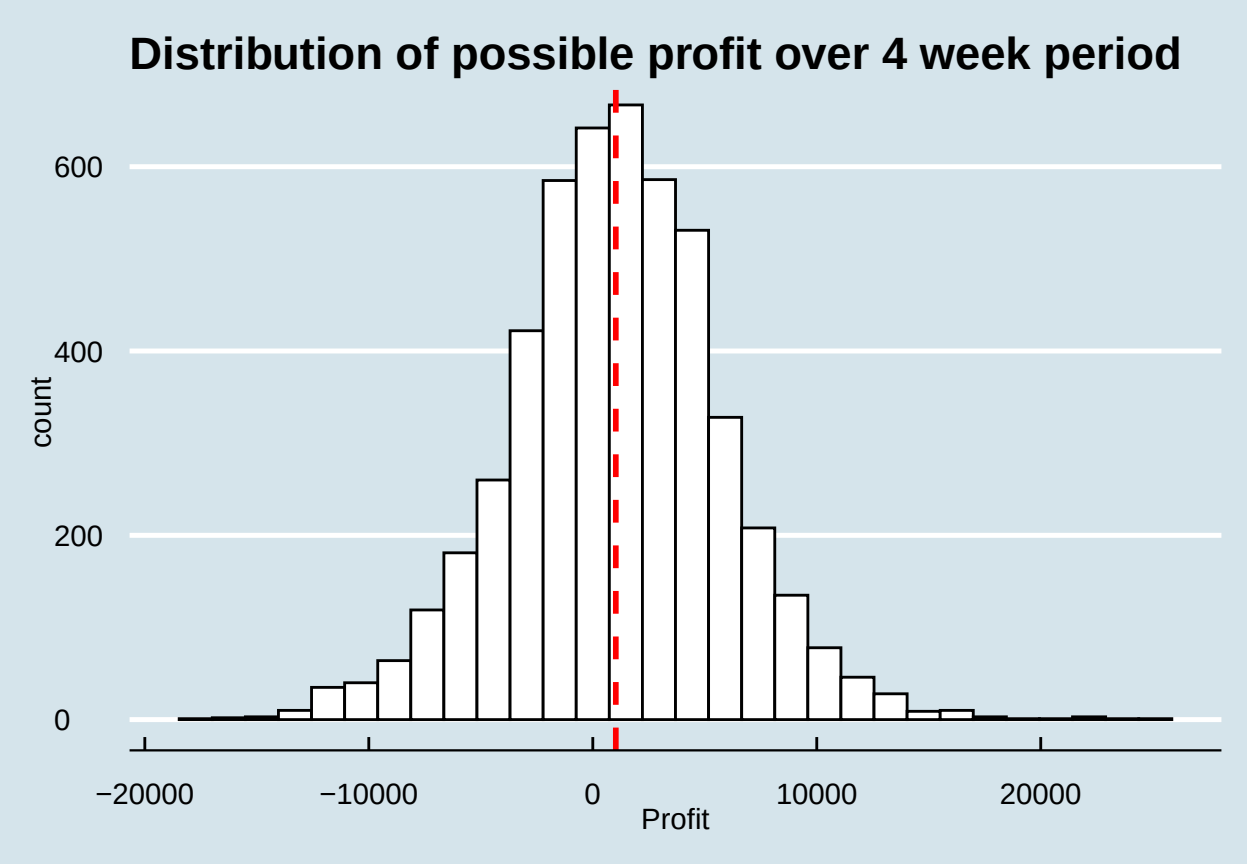
Let's construct another portfolio but a bit more diversified this time-

1) XLF: Financial Select Sector SPDR Fund
2) IYH: iShares U.S. Healthcare ETF
3) XLY: Consumer Discretionary Select Sector SPDR Fund
4) QQQ: Invesco QQQ Trust
5) XLK: Technology Select Sector SPDR Fund
6) XLE: Energy Select Sector SPDR Fund
7) FXZ: First Trust Materials AlphaDEX Fund
8) FLOT: iShares Floating Rate Bond ETF
9) BKLN: Invesco Senior Loan ETF
10) VMBS: Vanguard Mortgage-Backed Securities ETF

Let's assign $20,000 to each ETF in this distributed portfolio and look at what our 4 week return distribution looks like when bootstrap for 5000 iterations -

```
##  [1] "XLF"  "IYH"  "XLY"  "QQQ"  "XLK"  "XLE"  "FXZ"  "FLOT" "BKLN" "VMBS"
```

```
##                ClCl.XLFa     ClCl.IYHa    ClCl.XLYa     ClCl.QQQa     ClCl.XLKa
## 2017-08-08            NA            NA           NA            NA            NA
## 2017-08-09  0.0000000000  0.0017454590 -0.005805663 -0.0013166309 -0.0005183656
## 2017-08-10 -0.0178006718 -0.0138788571 -0.014984574 -0.0214404393 -0.0197061200
## 2017-08-11 -0.0040273862  0.0035337963  0.005145403  0.0075870879  0.0068770765
## 2017-08-14  0.0137484836  0.0063747675  0.007122179  0.0128781627  0.0159369533
## 2017-08-15  0.0007977663  0.0006636101 -0.009171293  0.0006948239  0.0015514567
##                ClCl.XLEa    ClCl.FXZa    ClCl.FLOTa    ClCl.BKLNa    ClCl.VMBSa
## 2017-08-08            NA           NA           NA            NA            NA
## 2017-08-09  0.0009232959 -0.008305217  0.0001965612 -0.0012903656  0.0003786445
## 2017-08-10 -0.0106072555 -0.013870689 -0.0001965226 -0.0030145996  0.0007570212
## 2017-08-11 -0.0065258234  0.002653901  0.0001965612  0.0004319654  0.0005673033
## 2017-08-14 -0.0029715202  0.008470064 -0.0005894499  0.0008635579 -0.0009449820
## 2017-08-15 -0.0037647373  0.000000000  0.0005897975 -0.0004759638 -0.0007567348
```



What's our Value at Risk for this portfolio for a 4 week period with a 5% confidence?

```
##        5%
## -6846.212
```

Would you look at that, with a more diversified portfolio, we got almost half the value at risk (VaR) at 5% confidence intervals compared to the more concentrated portfolio we analyzed before.

Let's take a more standard appraoch and analyze 3 more portfolios with varying levels of risk -

Portfolio 1 - 10 ETFs Balanced

We will consider the following portfolio for this case :

Table 3: PORTFOLIO 1

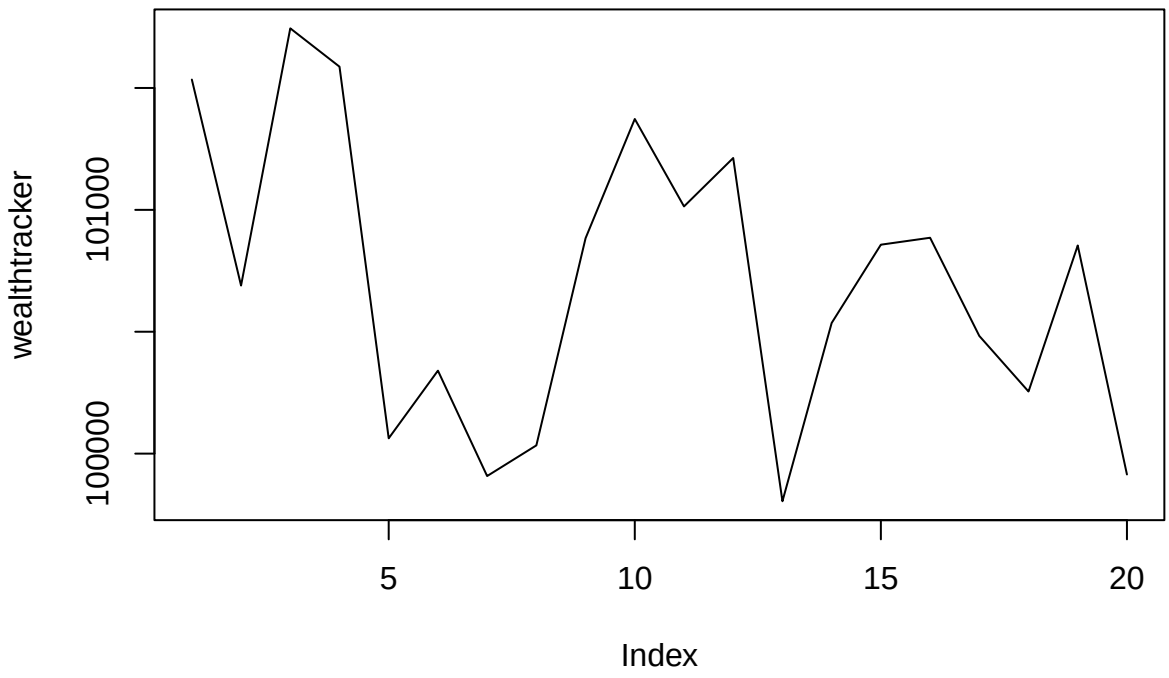
S.No	ETF	Definition
1	WMT	Walmart
2	TGT	Target
3	XOM	Exxon Mobil
4	MRK	Merck
5	IDV	iShares International Select Dividend ETF
6	SPY	SPDR S&P 500 ETF Trust
7	VTI	Vanguard
8	DIA	Dow Jones Industrial
9	JNJ	Johnson and Johnson
10	URTY	Ultra share pro

```
##  [1] "WMT"  "TGT"  "XOM"  "URTY" "JNJ"  "MRK"  "IDV"  "SPY"  "VTI"  "DIA"
```

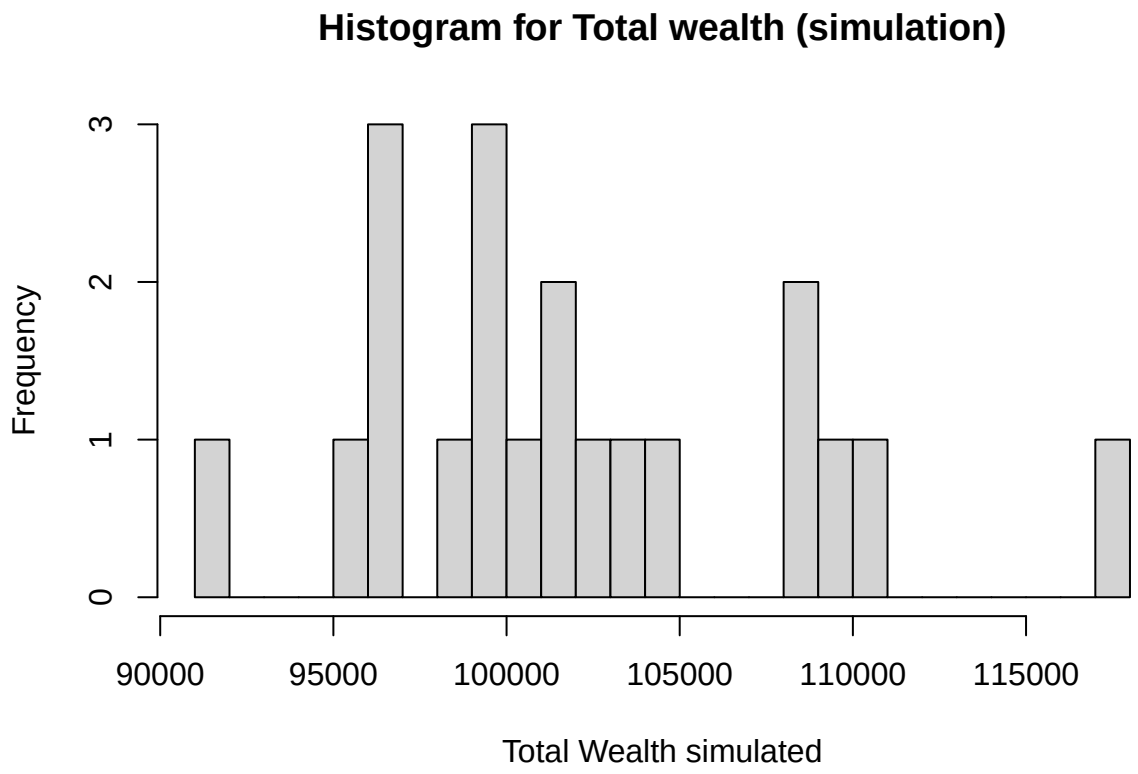
```
## [1] "Combine all the returns in a matrix"
```

```
## [1] 99914.4
```

```
## [1] "Total wealth through 20 days"
```



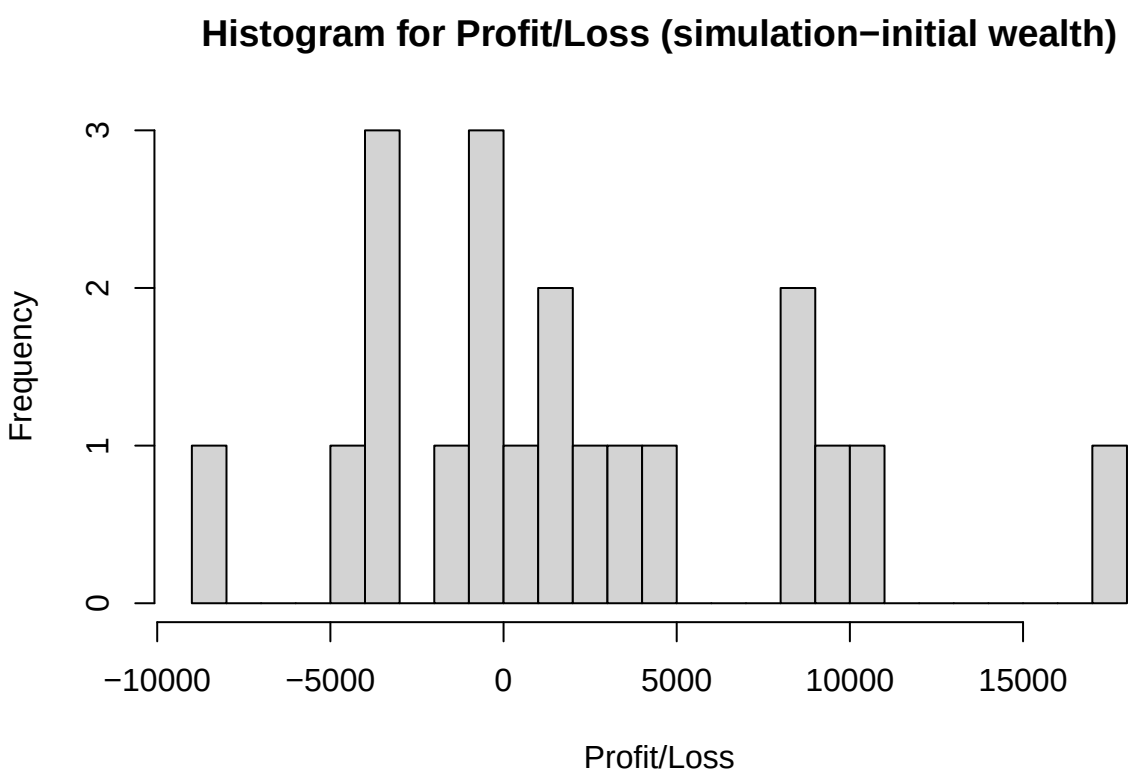

```
## [1] "Total wealth through simulate many different possible futures"
```




```
## [1] 102198.6
```

```
## [1] 2198.599
```

```
## [1] "Mean profit Loss"
```




```
## [1] "5% value at risk"
```

```
##        5%
## -4521.687
```

This first portfolio performs well and this time with a -ve VAR value but a lower profit margin overall compared to our next portfolio. This is potentially due to the high number of stocks which can cause our profits to average out over the stocks that we have.

Let's test out this theory by using a lower total number of stocks in our portfolio and try to measure the profit and VAR value.

Portfolio 2 - 5 ETFs Risky and balanced

We will consider the following portfolio for this case :

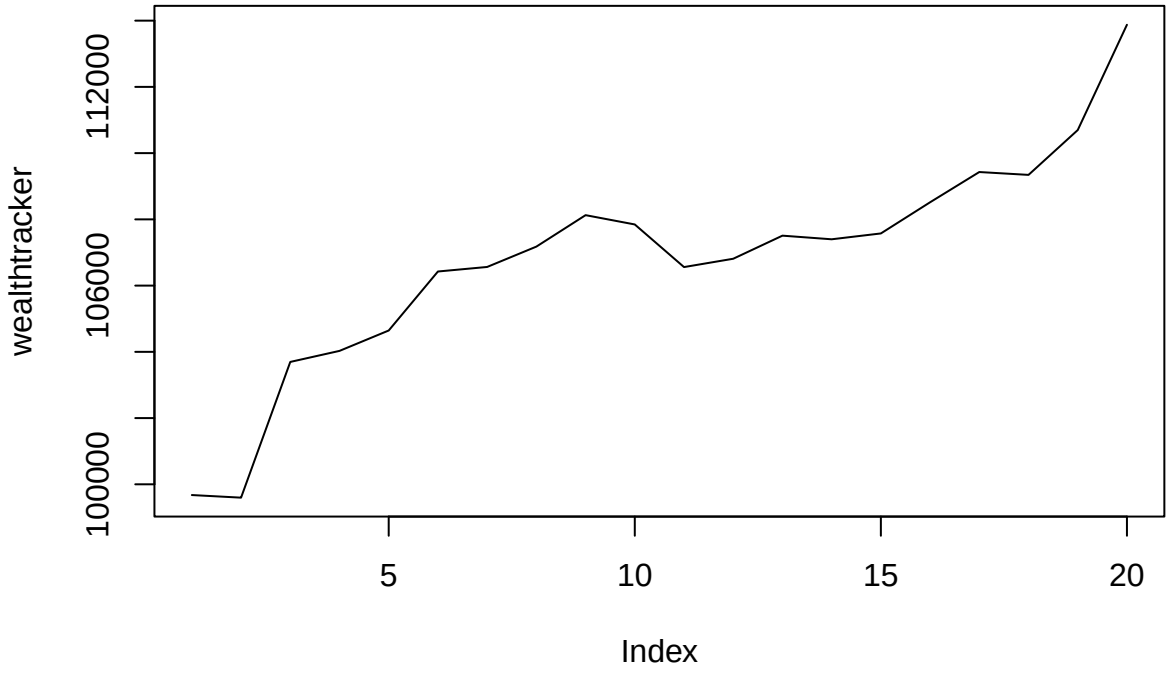
Table 4: PORTFOLIO 2

S.No	ETF	Definition
1	MRK	Merck
2	IDV	iShares International Select Dividend ETF
3	SPY	SPDR S&P 500 ETF Trust
4	VTI	Vanguard
5	DIA	Dow Jones Industrial

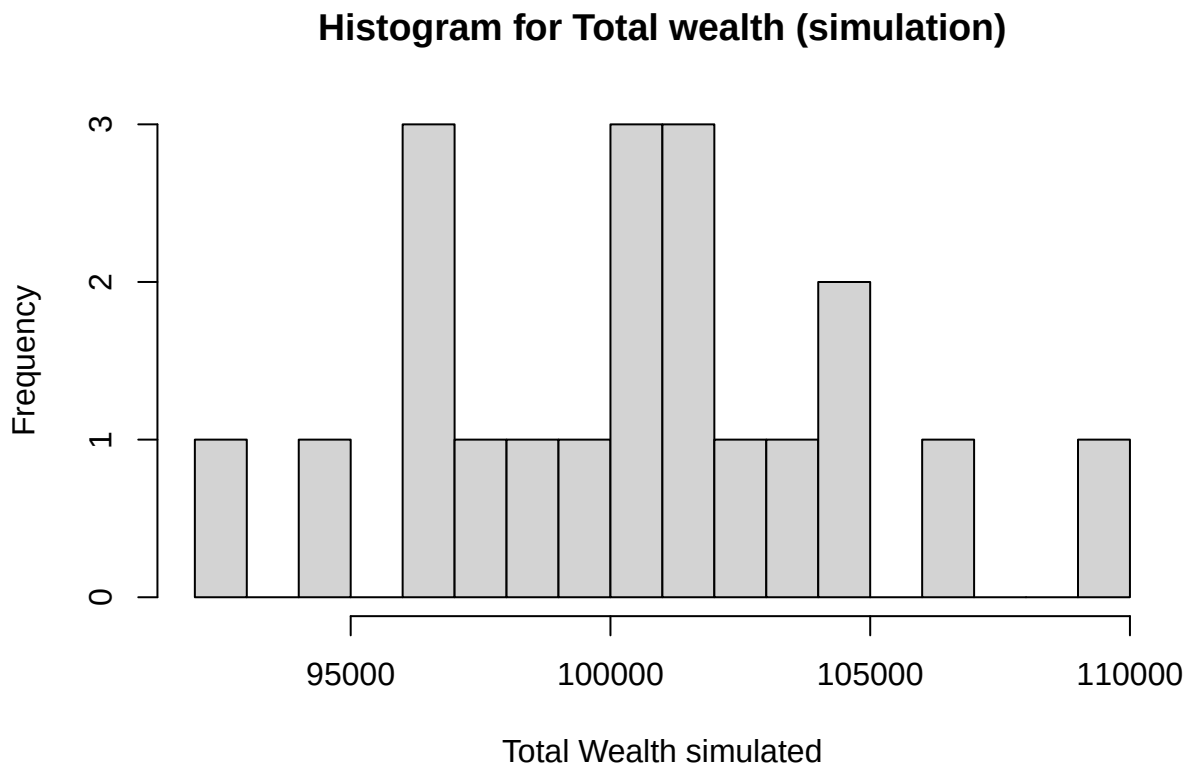
```
## [1] "MRK" "IDV" "SPY" "VTI" "DIA"
```

```
## [1] 113875.7
```

```
## [1] "Total wealth through 20 days"
```



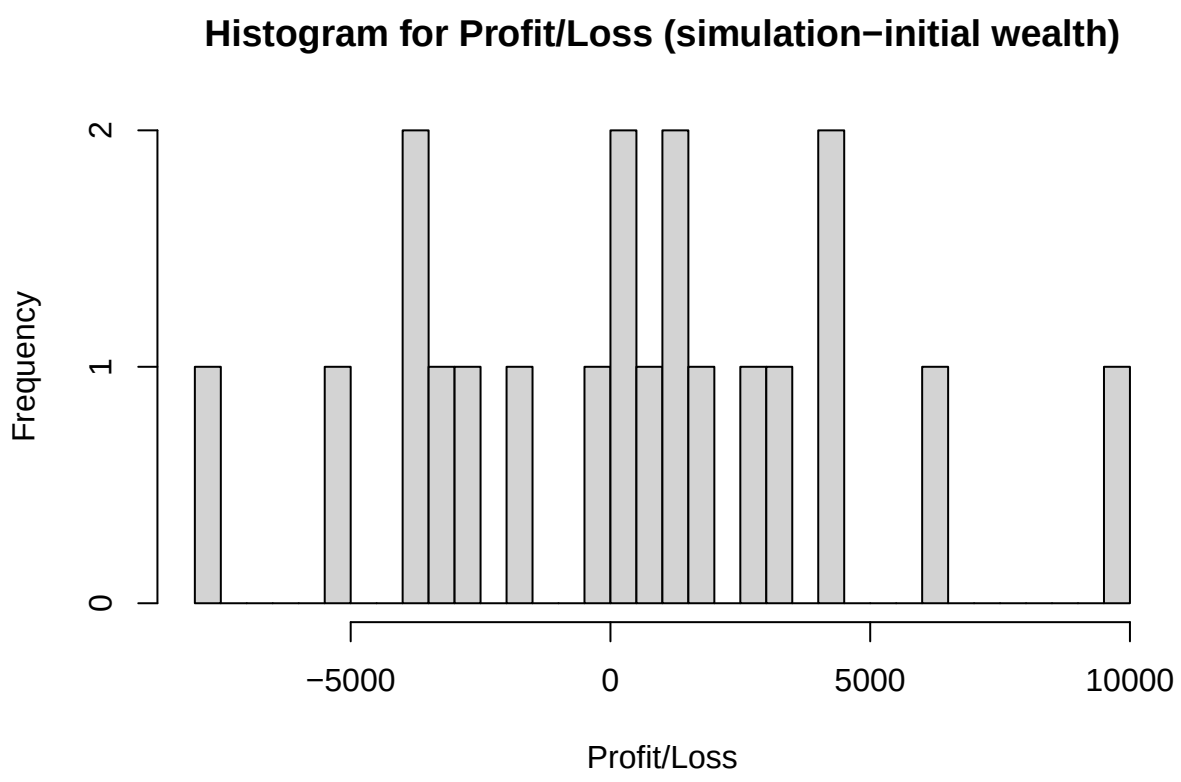

```
## [1] "Total wealth through simulate many different possible futures"
```




```
## [1] 100333.1
```

```
## [1] 333.1087
```

```
## [1] "Mean profit Loss"
```




```
## [1] "5% value at risk"
```

```
##       5%
## -5377.31
```

This second portfolio performs worse than portfolio 1 with a lower profit margin overall. This means that it was not helpful to have a lower number of total stocks in the portfolio and as a result assign higher weights to individual stocks in the portfolio.

Portfolio 3 - Smallcap- Very Risky

Leveraged U.S. Size Factor TR ,VTI:Vanguard We will consider the following portfolio for this case :

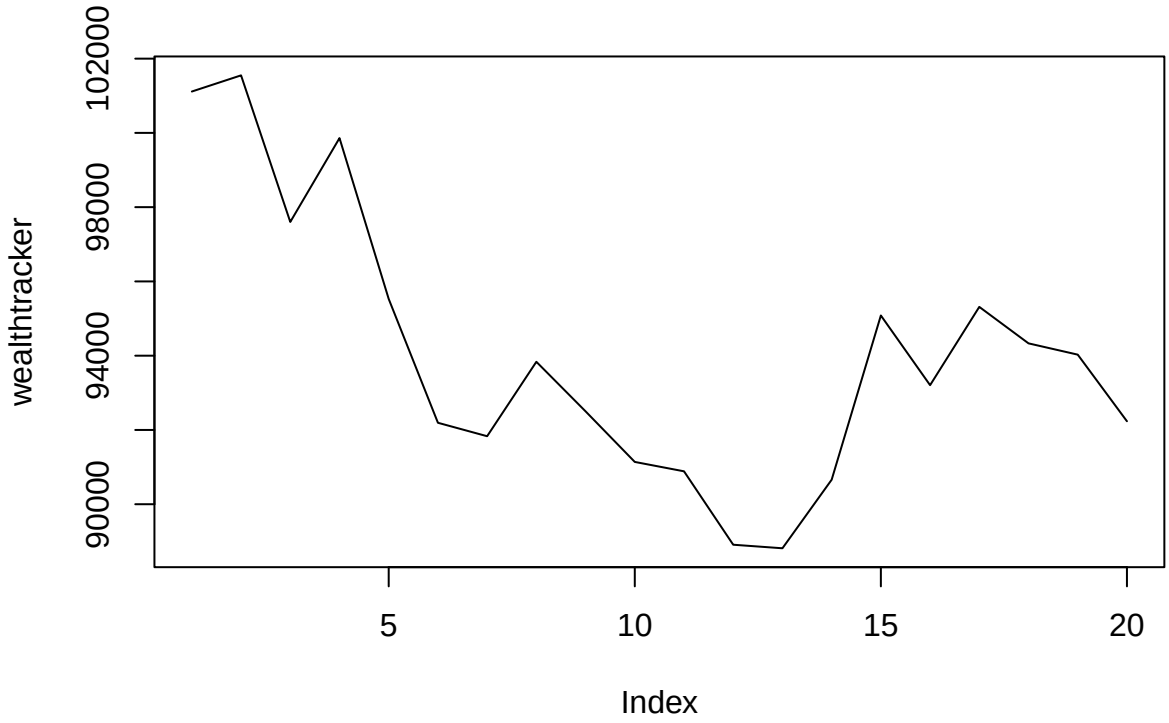
Table 5: PORTFOLIO 3

S.No	ETF	Definition
1	TNA	Direxion Daily Small Cap Bull 3x
2	URTY	ProShares UltraPro
3	UWM	ProShares Ultra
4	VTI	Vanguard
5	IWML	ETRACS 2x Leveraged U.S. Size Factor TR

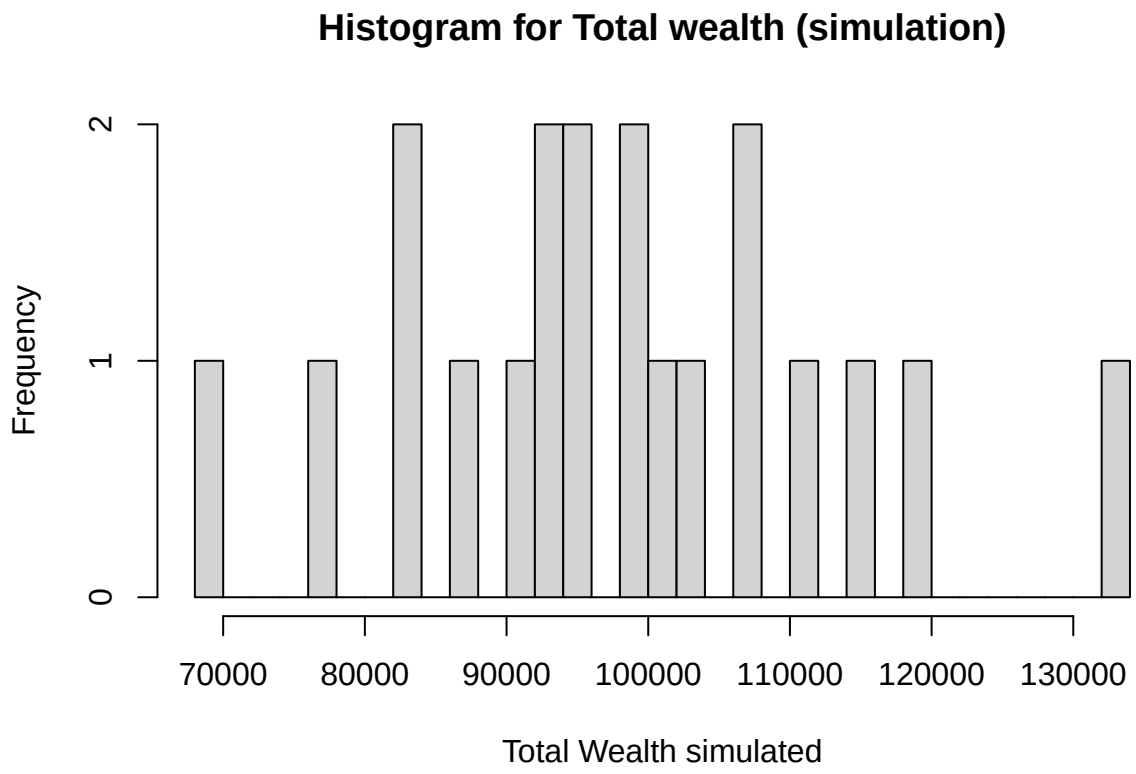
```
## [1] "TNA"  "URTY" "UWM"  "IWML" "VTI"
```

```
## [1] 92234.03
```

```
## [1] "Total wealth through 20 days"
```



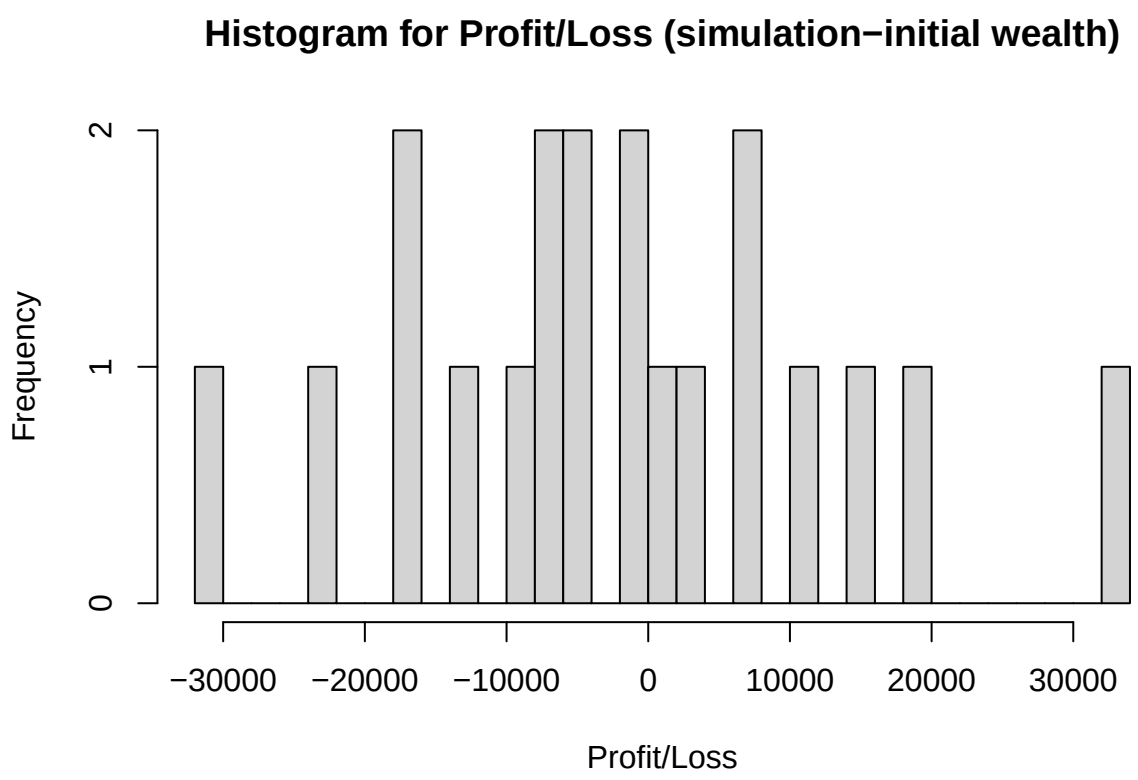

```
## [1] "Total wealth through simulate many different possible futures"
```




```
## [1] 98080.45
```

```
## [1] -1919.555
```

```
## [1] "Mean profit Loss"
```




```
## [1] "5% value at risk"

##        5%
## -22947.71
```

This third portfolio performs better than portfolio 2 with a higher profit margin overall. This means that it was helpful to have a risky set of stocks in the portfolio and as a result assign higher weights to individual stocks in the portfolio.Also this gives us a high VAR which makes sense as the portfolio was very risky.

Clustering and PCA

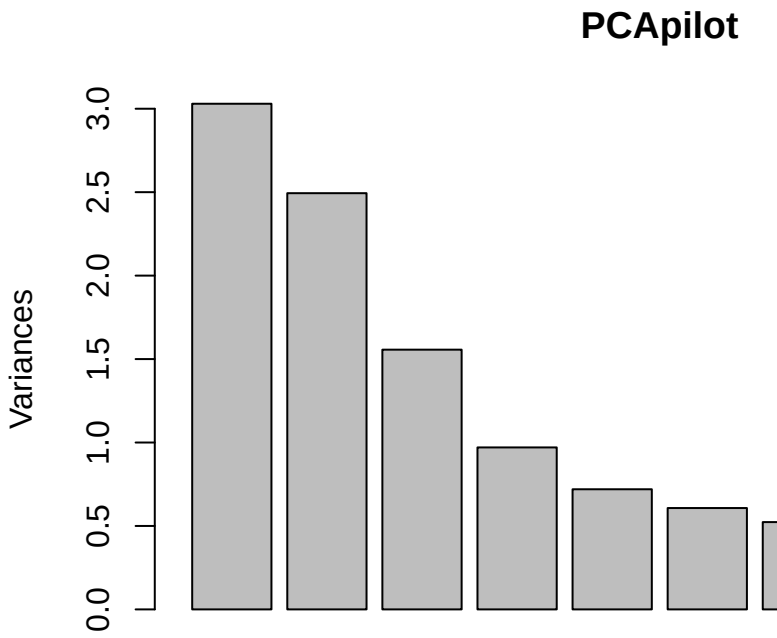
The data in wine.csv contains information on 11 chemical properties of 6500 different bottles of *vinho verde* wine from northern Portugal. In addition, two other variables about each wine are recorded: - whether the wine is red or white
- the quality of the wine, as judged on a 1-10 scale by a panel of certified wine snobs.

Run both PCA and a clustering algorithm of your choice on the 11 chemical properties (or suitable transformations thereof) and summarize your results. Which dimensionality reduction technique makes more sense to you for this data? Convince yourself (and me) that your chosen method is easily capable of distinguishing the reds from the whites, using only the "unsupervised" information contained in the data on chemical properties. Does your unsupervised technique also seem capable of distinguishing the higher from the lower quality wines?

To clarify: I'm not asking you to run an supervised learning algorithms. Rather, I'm asking you to see whether the differences in the labels (red/white and quality score) emerge naturally from applying an unsupervised technique to the chemical properties. This should be straightforward to assess using plots.

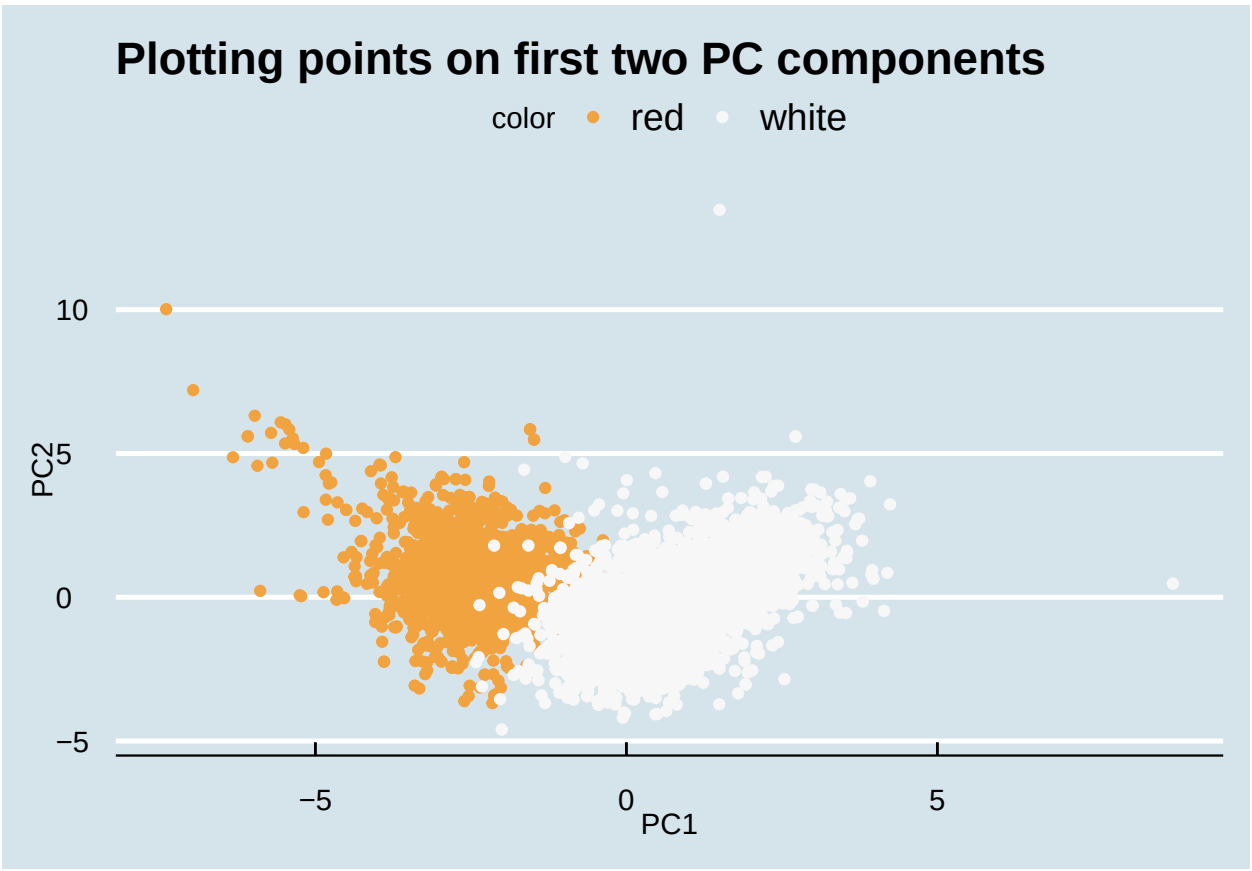
Solution:

Let's first look at Principal Component Analysis and if it can help us classify red and white wine.



Let's look at the variances explained by different PC axes-

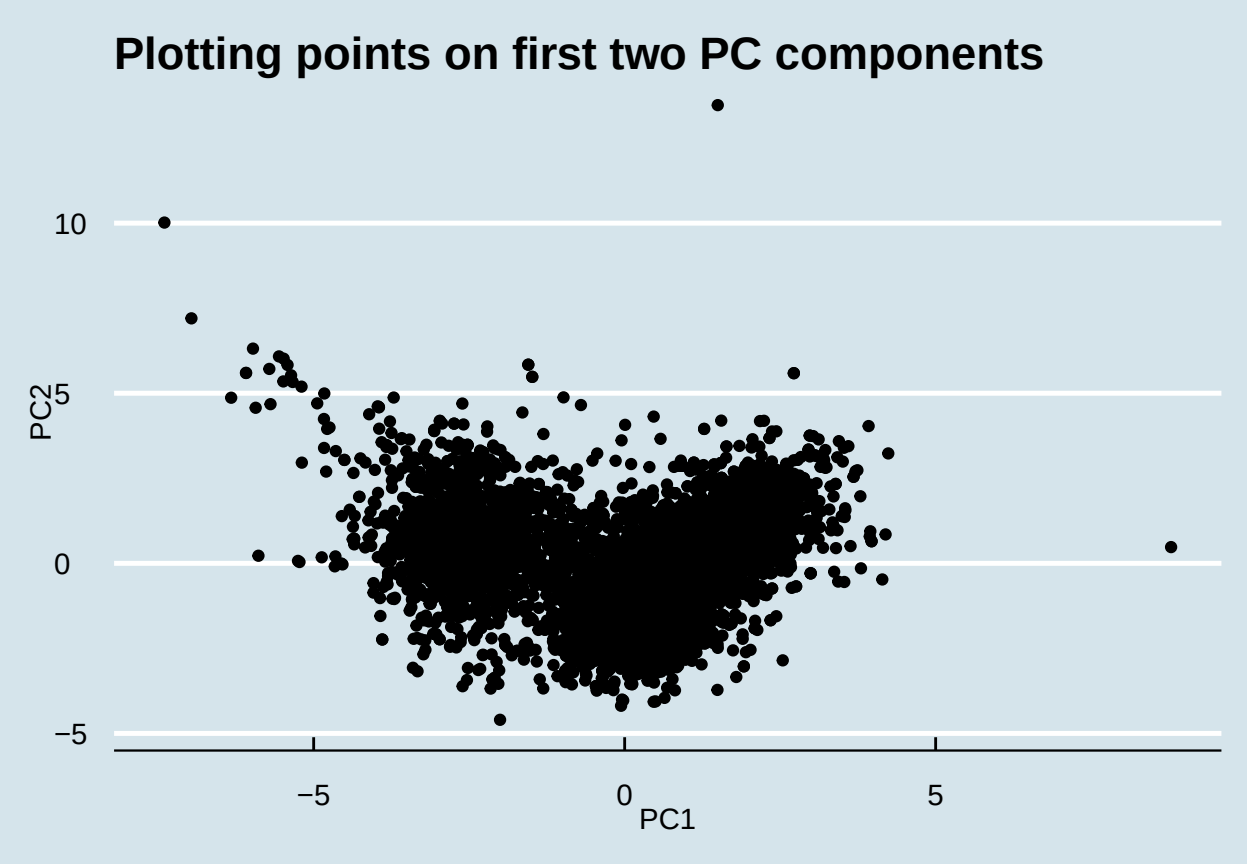
Most of the variance is explained by the first two PC axes. Let's plot the observations along these two PC axes



-

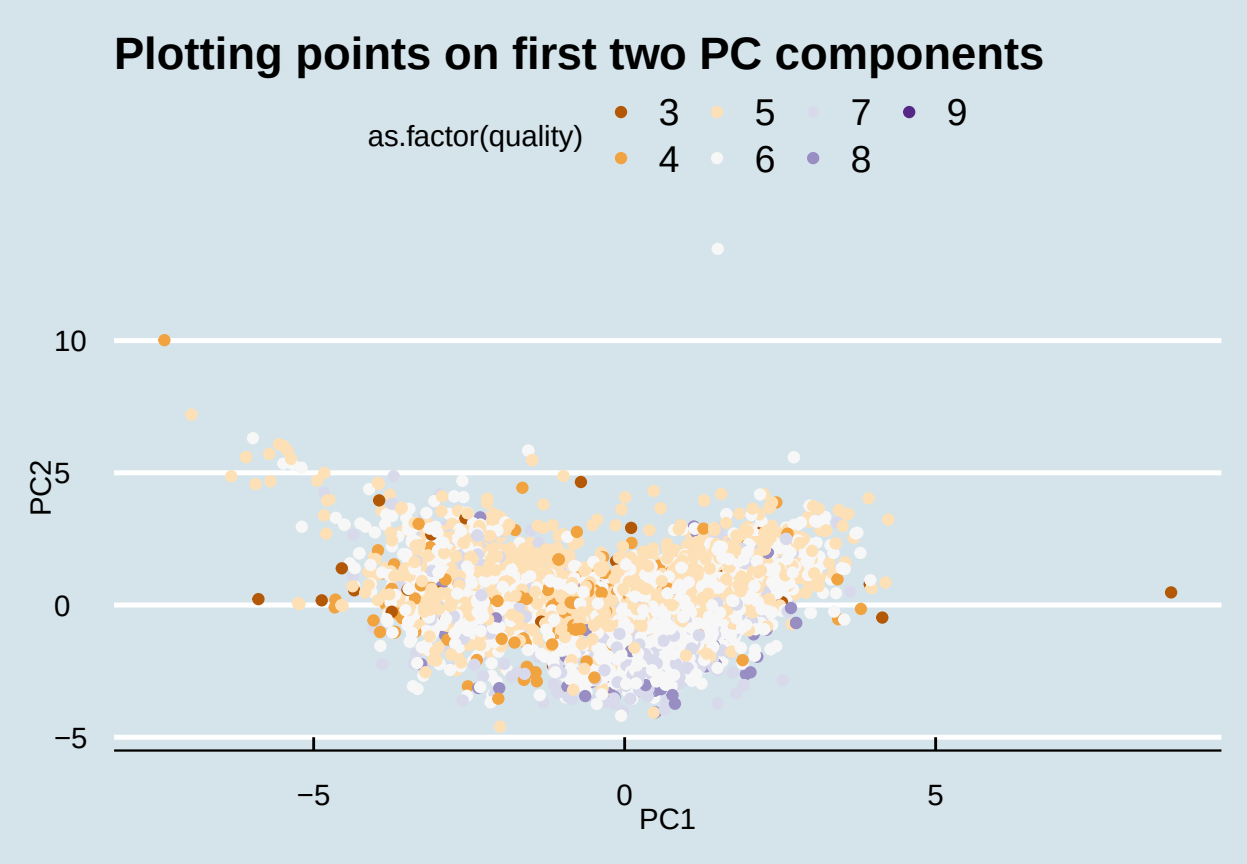
We can clearly see that the first two PC axes help us distinguish wine color!

However, in tyhe absence of the predicted variable such as color, let's see what outr PCA plot looks like -



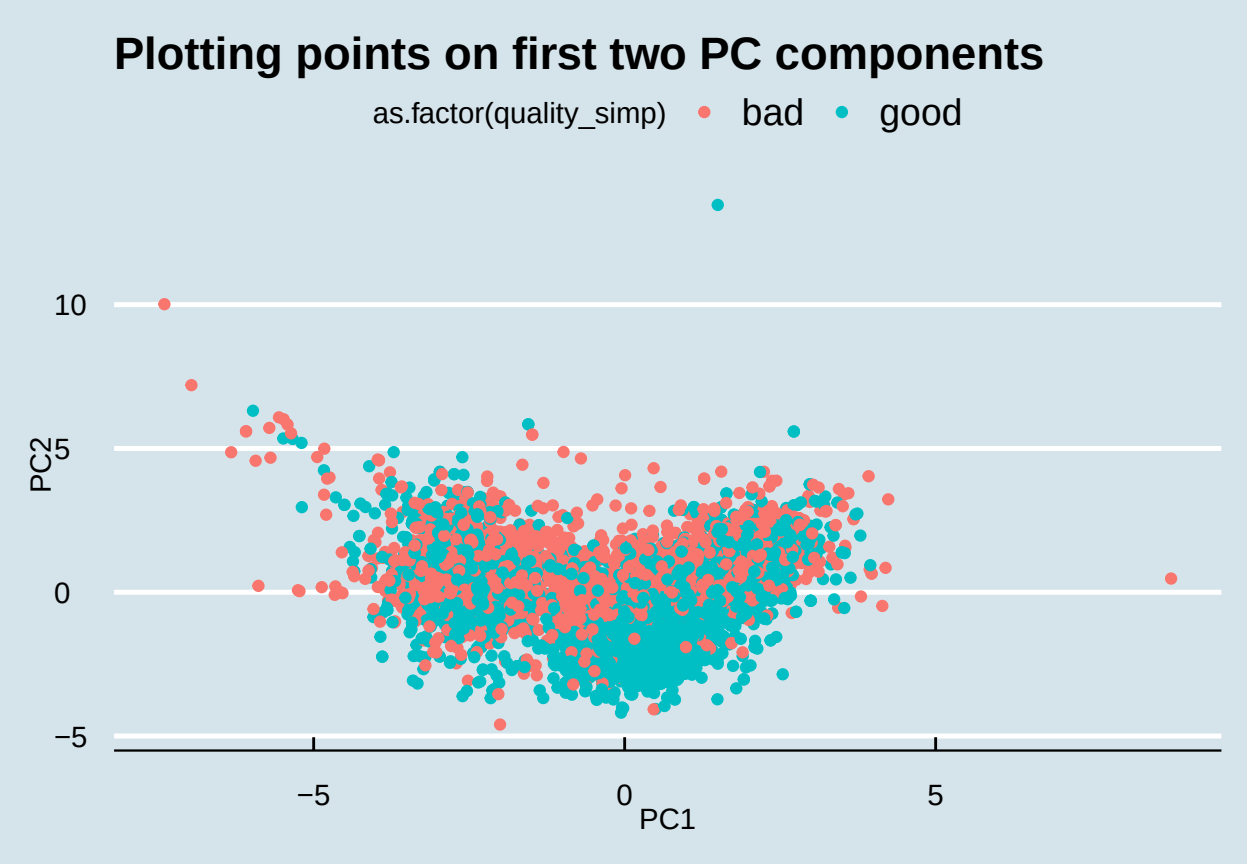
We can kind of see two different clusters without distinguiishing the points with the color related information we already have.

Let's see if they help us distinguish quality -



Nope! Does not help in distinguishing quality.

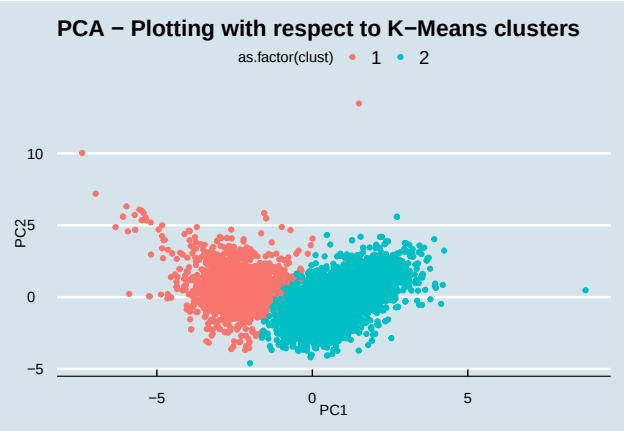
Let's classify any wines with quality lesser or equal to 5 as 'bad' and greater than 5 as 'good'. Maybe PCA could help classify wines in these simplified categories?



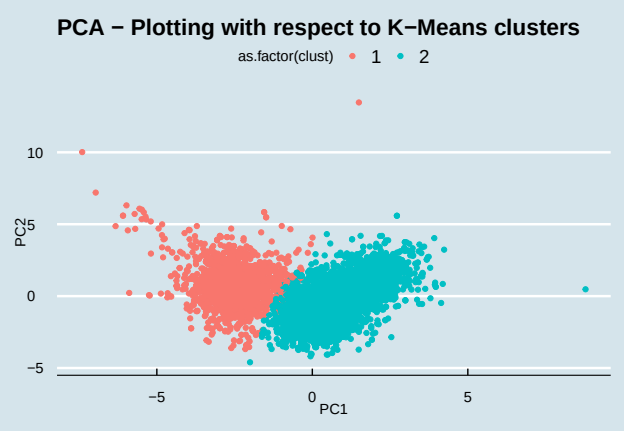
Nope, the data points with respect to quality cannot be distinguished even after simplifying the categories.

Let's see if K-means clustering can classify color of wine as well.

Using simple K-means, let's create 2 clusters and perform PCA. We can then plot the points wrt the first two PC axes, colour each point based on the cluster it's assigned to and compare that plot to the PCA plot we got earlier.



We can see that K-means performs pretty well in distinguishing the color of the wine except for a few outliers. However, if we plot the two clusters with beside the wine quality (good or bad) along the first two PC aces, we can see it doesn't perform too well -



Market segmentation

Consider the data in social_marketing.csv. This was data collected in the course of a market-research study using followers of the Twitter account of a large consumer brand that shall remain nameless—let's call it "NutrientH20" just to have a label. The goal here was for NutrientH20 to understand its social-media audience a little bit better, so that it could hone its messaging a little more sharply.

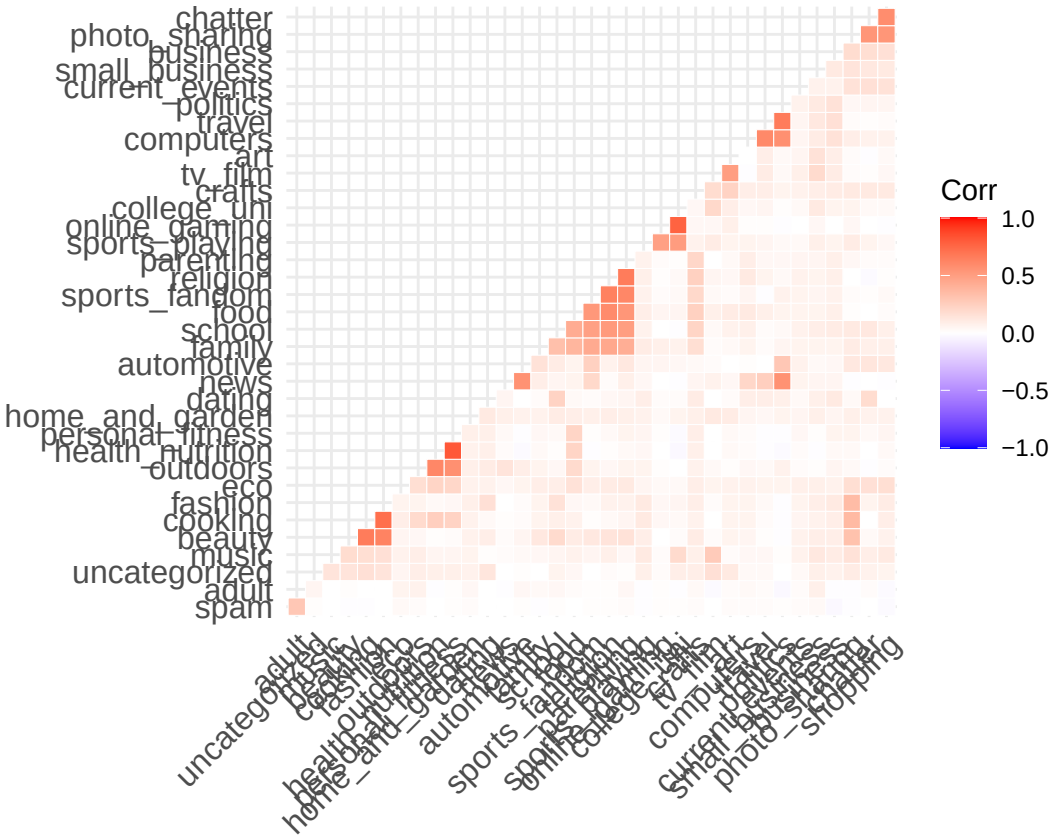
A bit of background on the data collection: the advertising firm who runs NutrientH20's online-advertising campaigns took a sample of the brand's Twitter followers. They collected every Twitter post ("tweet") by each of those followers over a seven-day period in June 2014. Every post was examined by a human annotator contracted through Amazon's Mechanical Turk service. Each tweet was categorized based on its content using a pre-specified scheme of 36 different categories, each representing a broad area of interest (e.g. politics, sports, family, etc.) Annotators were allowed to classify a post as belonging to more than one category. For example, a hypothetical post such as "I'm really excited to see grandpa go wreck shop in his geriatic soccer league this Sunday!" might be categorized as both "family" and "sports." You get the picture.

Each row of social_marketing.csv represents one user, labeled by a random (anonymous, unique) 9-digit alphanumeric code. Each column represents an interest, which are labeled along the top of the data file. The entries are the number of posts by a given user that fell into the given category. Two interests of note here are "spam" (i.e. unsolicited advertising) and "adult" (posts that are pornographic, salacious, or explicitly sexual). There are a lot of spam and pornography "bots" on Twitter; while these have been filtered out of the data set to some extent, there will certainly be some that slip through. There's also an "uncategorized" label. Annotators were told to use this sparingly, but it's there to capture posts that don't fit at all into any of the listed interest categories. (A lot of annotators may used the "chatter" category for this as well.) Keep in mind as you examine the data that you cannot expect perfect annotations of all posts. Some annotators might have simply been asleep at the wheel some, or even all, of the time! Thus there is some inevitable error and noisiness in the annotation process.

Your task to is analyze this data as you see fit, and to prepare a concise report for NutrientH20 that identifies any interesting market segments that appear to stand out in their social-media audience. You have complete freedom in deciding how to pre-process the data and how to define "market segment." (Is it a group of correlated interests? A cluster? A latent factor? Etc.) Just use the data to come up with some interesting, well-supported insights about the audience, and be clear about what you did.

Solution:

Let's first look at a correlation graph of each interest -

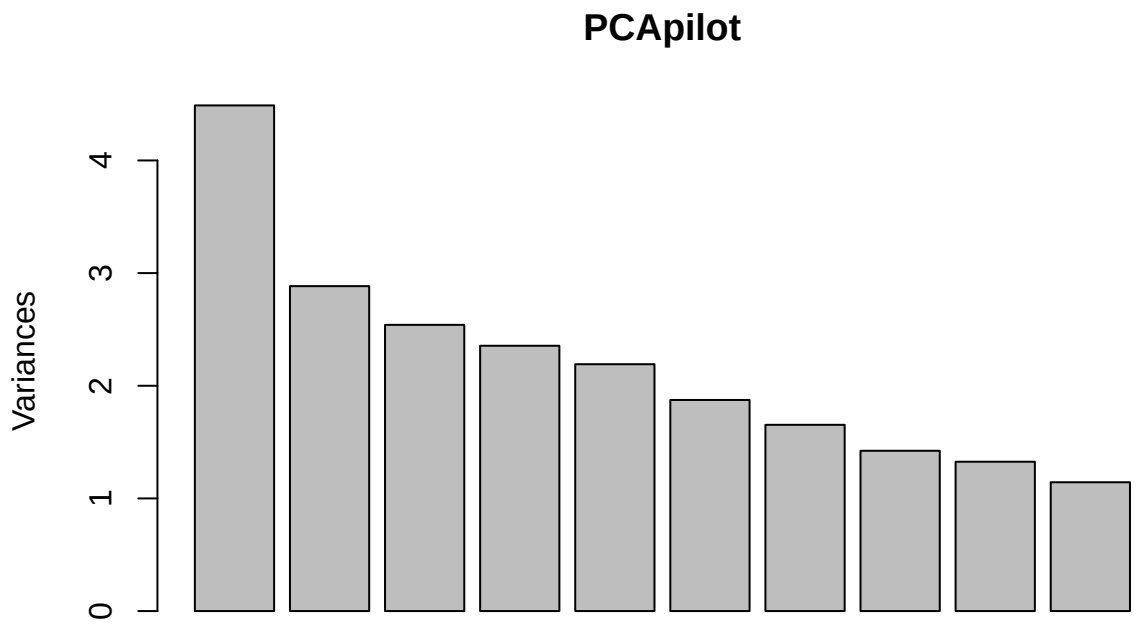


The interests of users seem to be quite correlated!

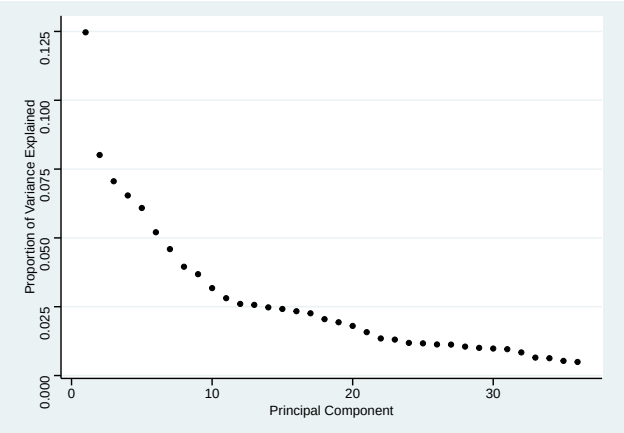
Let's perform PCA and look at the variance explained plot -

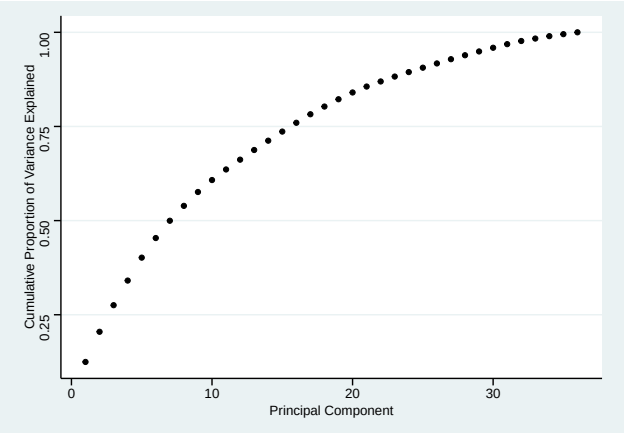
```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1186 1.69824 1.59388 1.53457 1.48027 1.36885 1.28577
## Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087 0.05205 0.04592
## Cumulative Proportion  0.1247 0.20479 0.27536 0.34077 0.40164 0.45369 0.49961
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.19277 1.15127 1.06930 1.00566 0.96785 0.96131 0.94405
## Proportion of Variance 0.03952 0.03682 0.03176 0.02809 0.02602 0.02567 0.02476
## Cumulative Proportion  0.53913 0.57595 0.60771 0.63580 0.66182 0.68749 0.71225
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.93297 0.91698 0.9020 0.85869 0.83466 0.80544 0.75311
## Proportion of Variance 0.02418 0.02336 0.0226 0.02048 0.01935 0.01802 0.01575
## Cumulative Proportion  0.73643 0.75979 0.7824 0.80287 0.82222 0.84024 0.85599
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.69632 0.68558 0.65317 0.64881 0.63756 0.63626 0.61513
## Proportion of Variance 0.01347 0.01306 0.01185 0.01169 0.01129 0.01125 0.01051
## Cumulative Proportion  0.86946 0.88252 0.89437 0.90606 0.91735 0.92860 0.93911
##                           PC29    PC30    PC31   PC32    PC33    PC34    PC35
## Standard deviation     0.60167 0.59424 0.58683 0.5498 0.48442 0.47576 0.43757
## Proportion of Variance 0.01006 0.00981 0.00957 0.0084 0.00652 0.00629 0.00532
## Cumulative Proportion  0.94917 0.95898 0.96854 0.9769 0.98346 0.98974 0.99506
##                           PC36
## Standard deviation     0.42165
```

```
## Proportion of Variance 0.00494
## Cumulative Proportion  1.00000
```



Let's look at the Scree plot and the cumulative scree plot -





We can maybe see an elbow point around the 8th PC axis.

Let's examine these axes and their weights -

```
##
## Loadings:
##                 PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
## chatter                                                  0.574
```

```
## current_events                                    0.178
## travel                       -0.580
## photo_sharing          0.188                       0.473
## uncategorized          0.113                             0.238
## tv_film                                                  0.602
## sports_fandom  -0.424                                           -0.147
## politics                     -0.476                             -0.301
## food           -0.381              -0.138
## family         -0.325
## home_and_garden                                          0.161
## music                  0.118                             0.234
## news                                                            -0.628
## online_gaming                                0.611
## shopping                                           0.548
## health_nutrition                    -0.581
## college_uni                                  0.603
## sports_playing                               0.498
## cooking                0.556        -0.102
## eco                                 -0.199         0.174
## computers                    -0.567
## business                     -0.178                0.159  0.107
## outdoors                            -0.496
## crafts         -0.138                                    0.272
## automotive                   0.154                 0.103        -0.653
## art                                                      0.550
## religion       -0.453
## beauty                 0.541
## parenting      -0.435
## dating                       -0.153                              0.106
## school         -0.366
## personal_fitness                    -0.565
## fashion                0.553
## small_business                                           0.260
## spam
## adult
##
##                  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
## SS loadings    1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
## Proportion Var 0.028 0.028 0.028 0.028 0.028 0.028 0.028 0.028
## Cumulative Var 0.028 0.056 0.083 0.111 0.139 0.167 0.194 0.222
```

Looking at PC1 - It has negative weights for topics such as family, food, parenting, crafts, sports, etc. This could possibly be a segment who don't care for family or recreational topics, somewhat of a no-nonsense business only segment (although the lack of positive weights is a little confusing).

For PC2, Positive weights are observed for photo_sharing, cooking, beauty, fashion. This could be our young adults segment who are more interested in sharing photos via social media, probably of the dishes they cook or interested in beauty and fashion tips.

For PC3, We can look at this as a casual user who's not interested in politics or business and maybe just looking at more light-hearted posts.

For PC4, This segment seems to be averse towards any physical activity related posts. We can't be too sure of their interests.

For PC5, This could be the sports user segment where the users are more interested in sports (maybe college

games, etc).'

For PC6, This segment seems to have interest in uncategorized topics or banter, shopping and photo sharing. We can call this segment the 'instagrammers'.

For PC7, Interests include tv and film, art, craft, music etc. We can call this segment 'art lovers'.

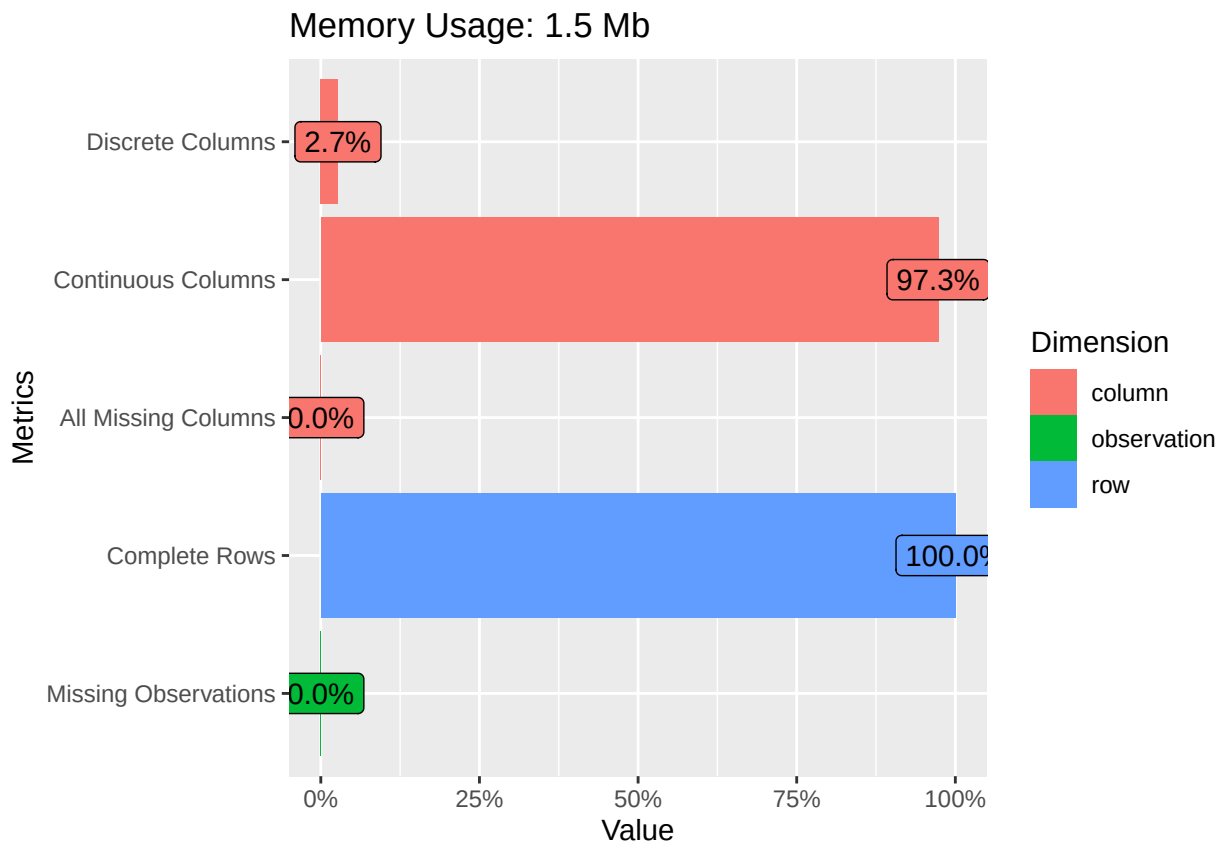
For PC8, Interests include dating but they're averse to news, business, etc. This might be the 'kids' segment.

We require at least 8 PC axes to define over 50% variance. This means that not a lot of clearly demarcated groups are visible and the interests are relatively independent.

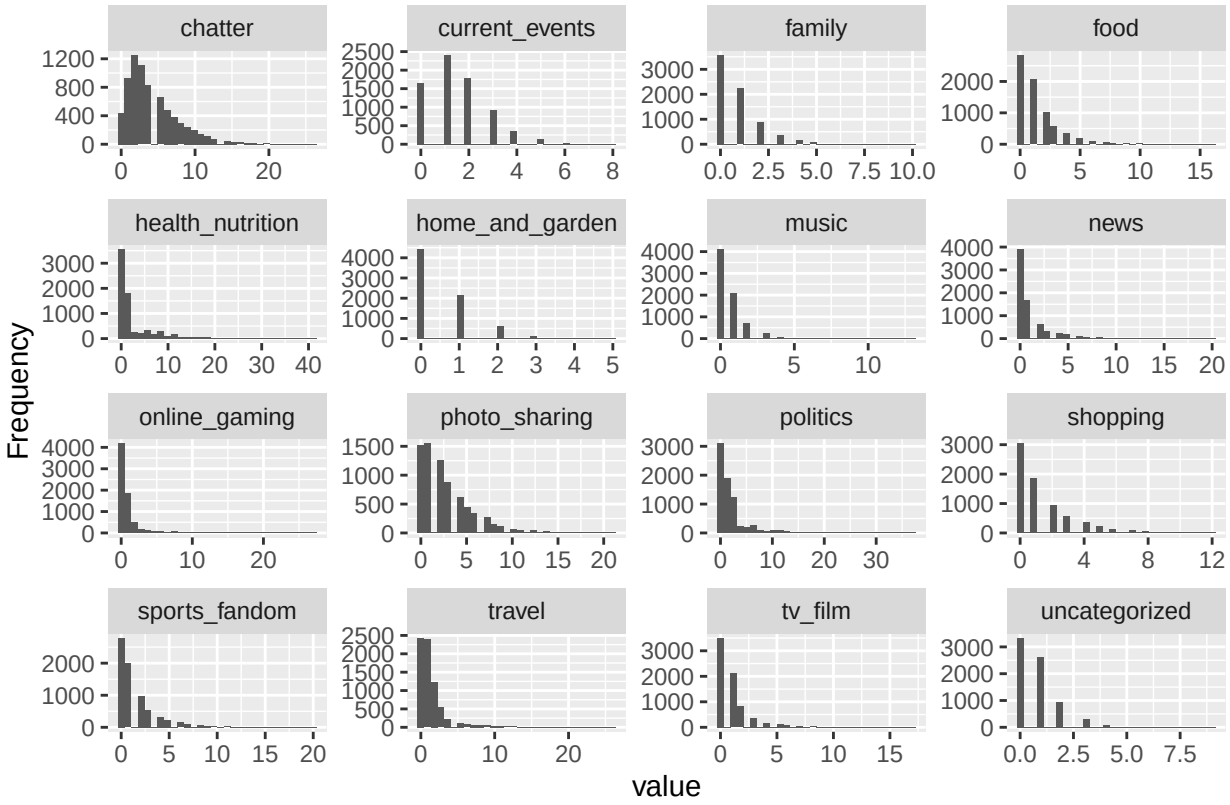
Now, let's try to combine PCA and clustering to obtain possibly better results that are more interpretable -

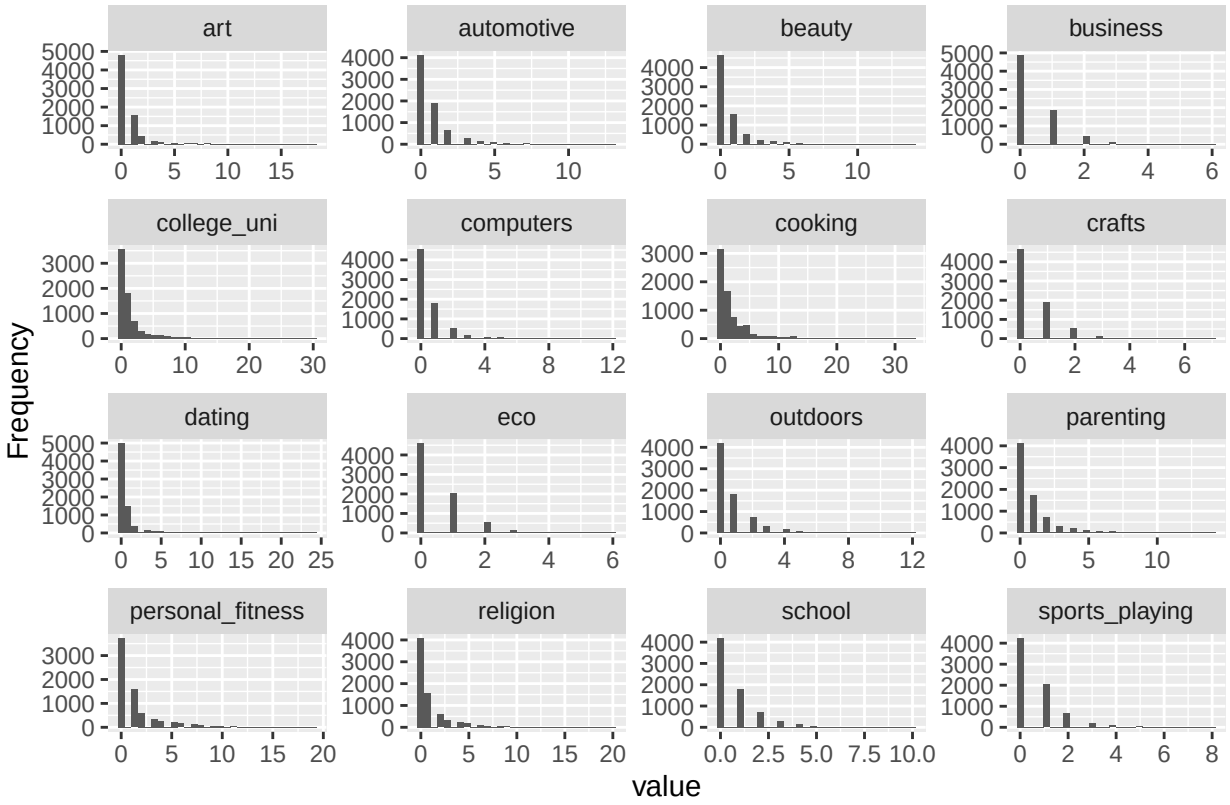
```
## [1] "Filter dataset on spam and Adult flags"
```

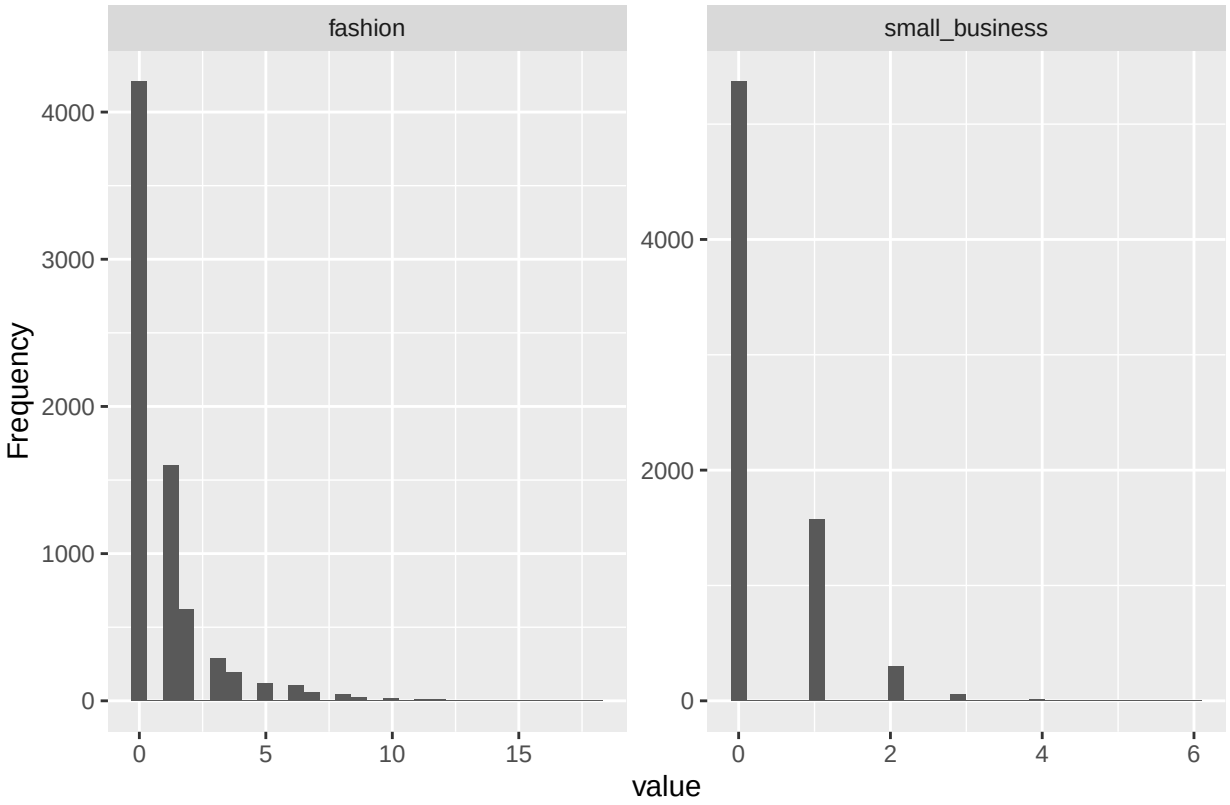
```
## [1] "Plotting the missing values and categorical columns"
```



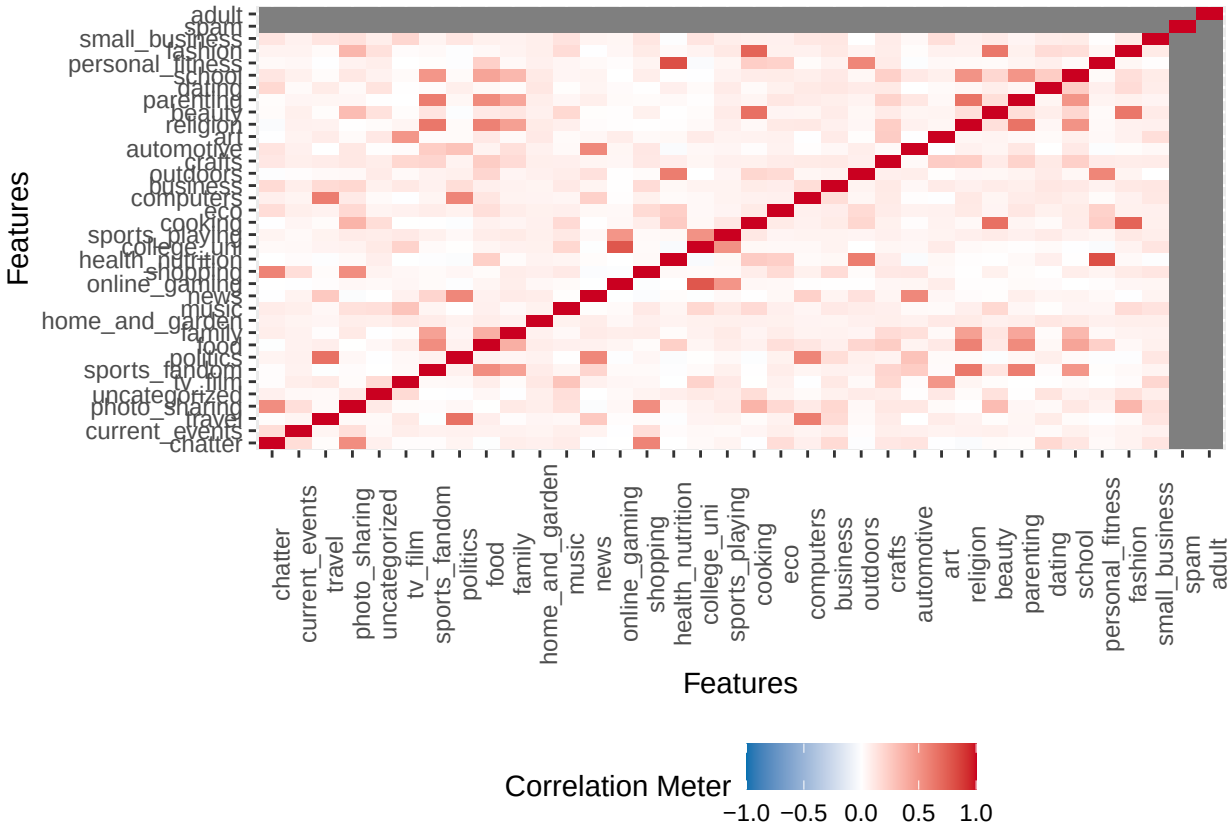

```
## [1] "Plotting the distribution of variables"
```

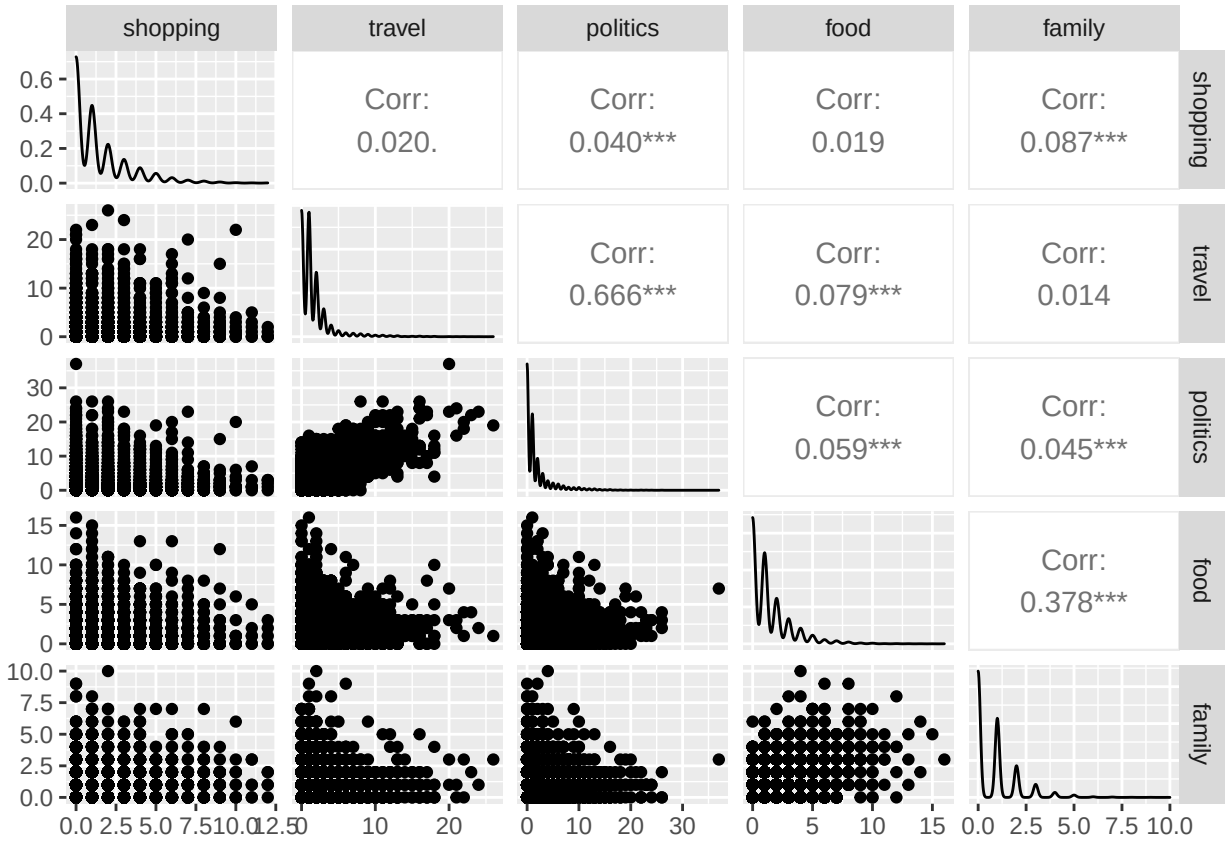

chatter
current_events
family
food
health_nutrition
home_and_garden
music
news
online_gaming
photo_sharing
politics
shopping
sports_fandom
travel
tv_film
uncategorized
Frequency
value


art
automotive
beauty
business
college_uni
computers
cooking
crafts
dating
eco
outdoors
parenting
personal_fitness
religion
school
sports_playing
Frequency
value



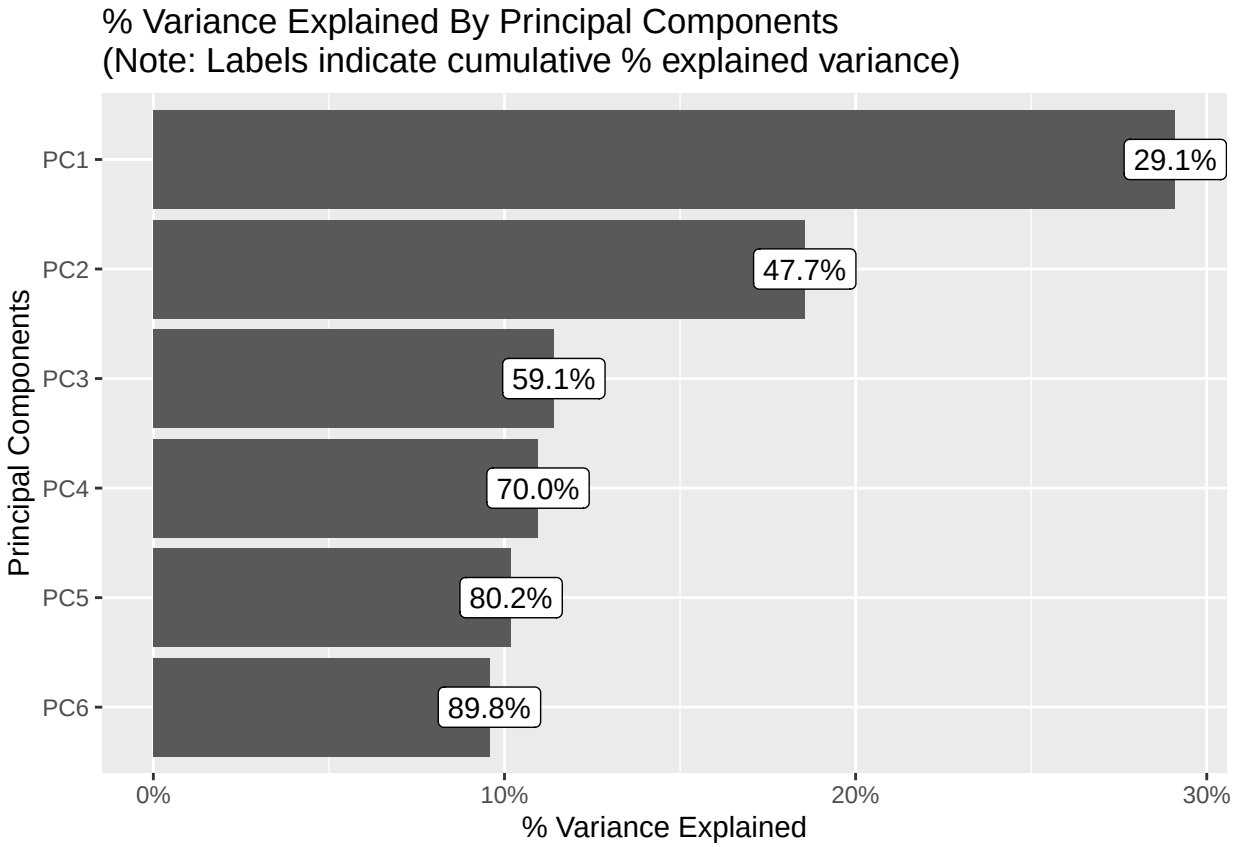

```
## [1] "Plotting the correlation of items"
```

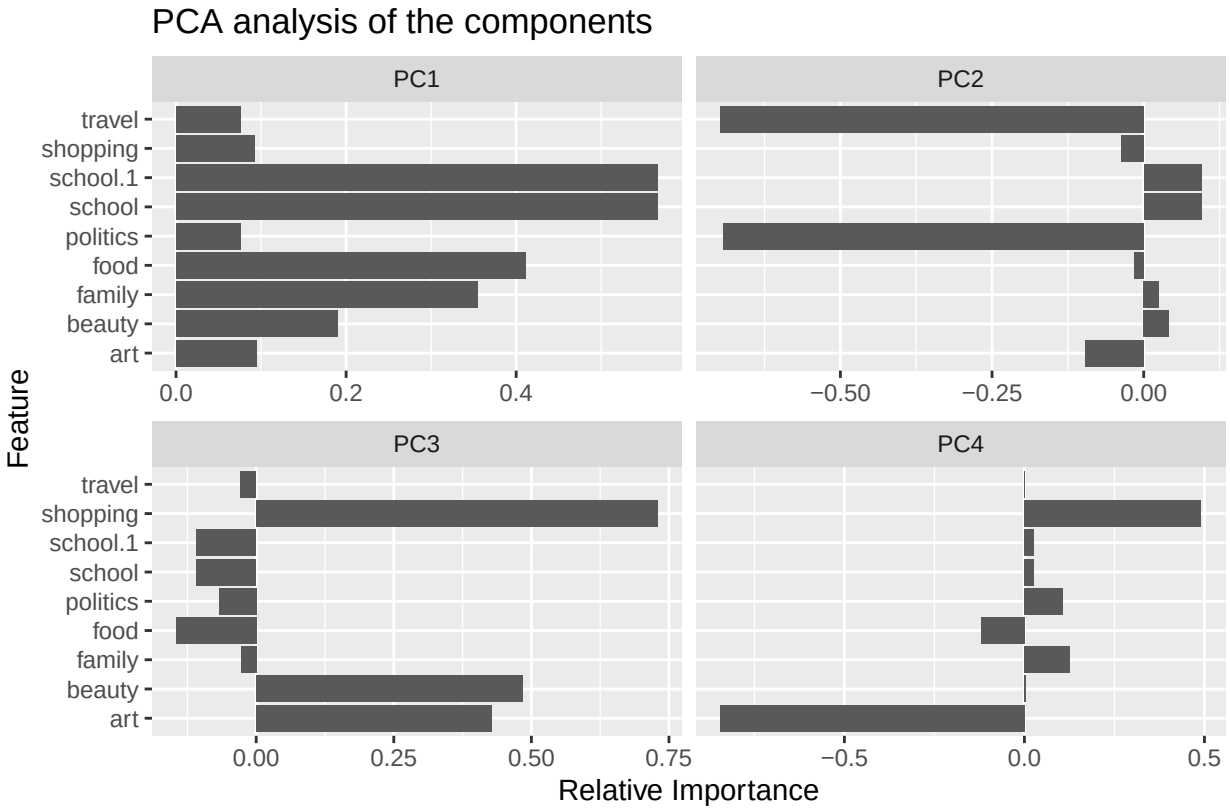

Features
adult
spam
small_business
fashion
personal_fitness
school
dating
parenting
beauty
religion
art
automotive
crafts
outdoors
business
computers
eco
cooking
sports_playing
college_uni
health_nutrition
shopping
online_gaming
news
music
home_and_garden
family
food
politics
sports_fandom
tv_film
uncategorized
photo_sharing
travel
current_events
chatter
chatter
current_events
travel
photo_sharing
uncategorized
tv_film
sports_fandom
politics
food
family
home_and_garden
music
news
online_gaming
shopping
health_nutrition
college_uni
sports_playing
cooking
eco
computers
business
outdoors
crafts
automotive
art
religion
beauty
parenting
dating
school
personal_fitness
fashion
small_business
spam
adult
Features
Correlation Meter
−1.0 −0.5 0.0 0.5 1.0



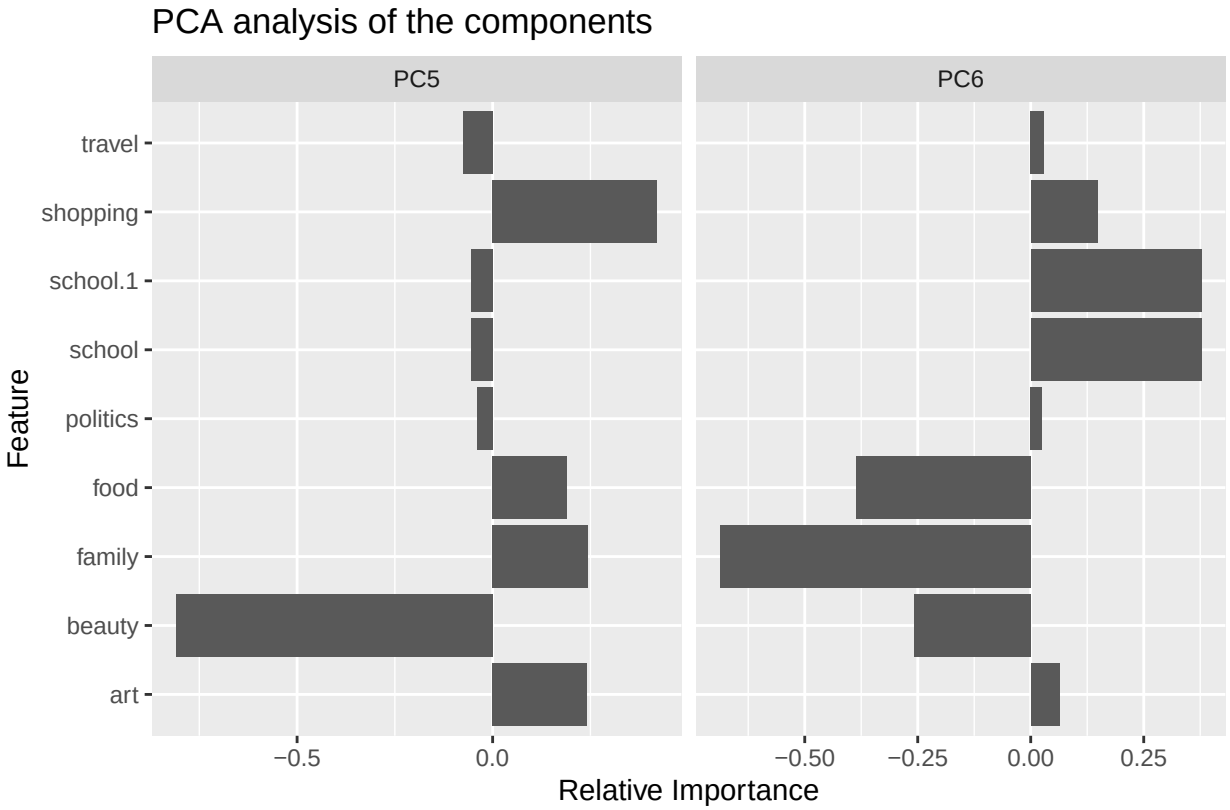

```
## [1] "It seems that tv_films is correlated with school, beauty and art, lets try to build a PCA model
```

```
## [1] "Performaing a PCA analysis to understand the composition"
```


% Variance Explained By Principal Components
(Note: Labels indicate cumulative % explained variance)
Principal Components
PC1
PC2
PC3
PC4
PC5
PC6
29.1%
47.7%
59.1%
70.0%
80.2%
89.8%
0%
10%
20%
30%
% Variance Explained


PCA analysis of the components
PC1
PC2
PC3
PC4
travel
shopping
school.1
school
politics
food
family
beauty
art
Feature
0.0
0.2
0.4
−0.50
−0.25
0.00
0.00
0.25
0.50
0.75
−0.5
0.0
0.5
Relative Importance

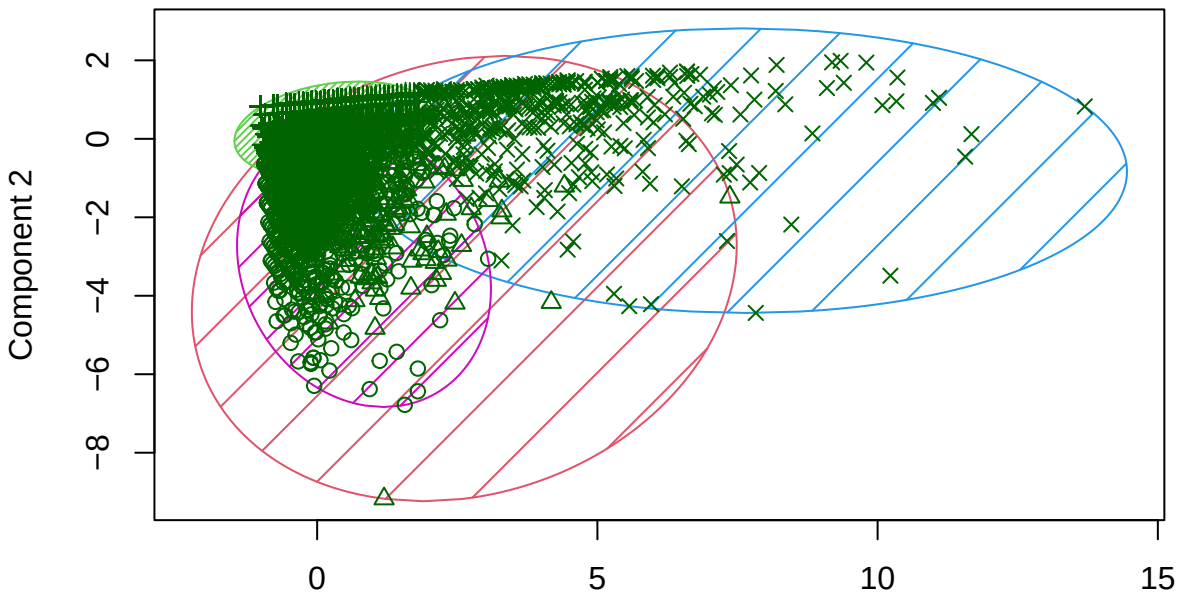
Page 1




```
## [1] "Lets try to cluster the most correlated variables of this data using Kmeans clustering"
```

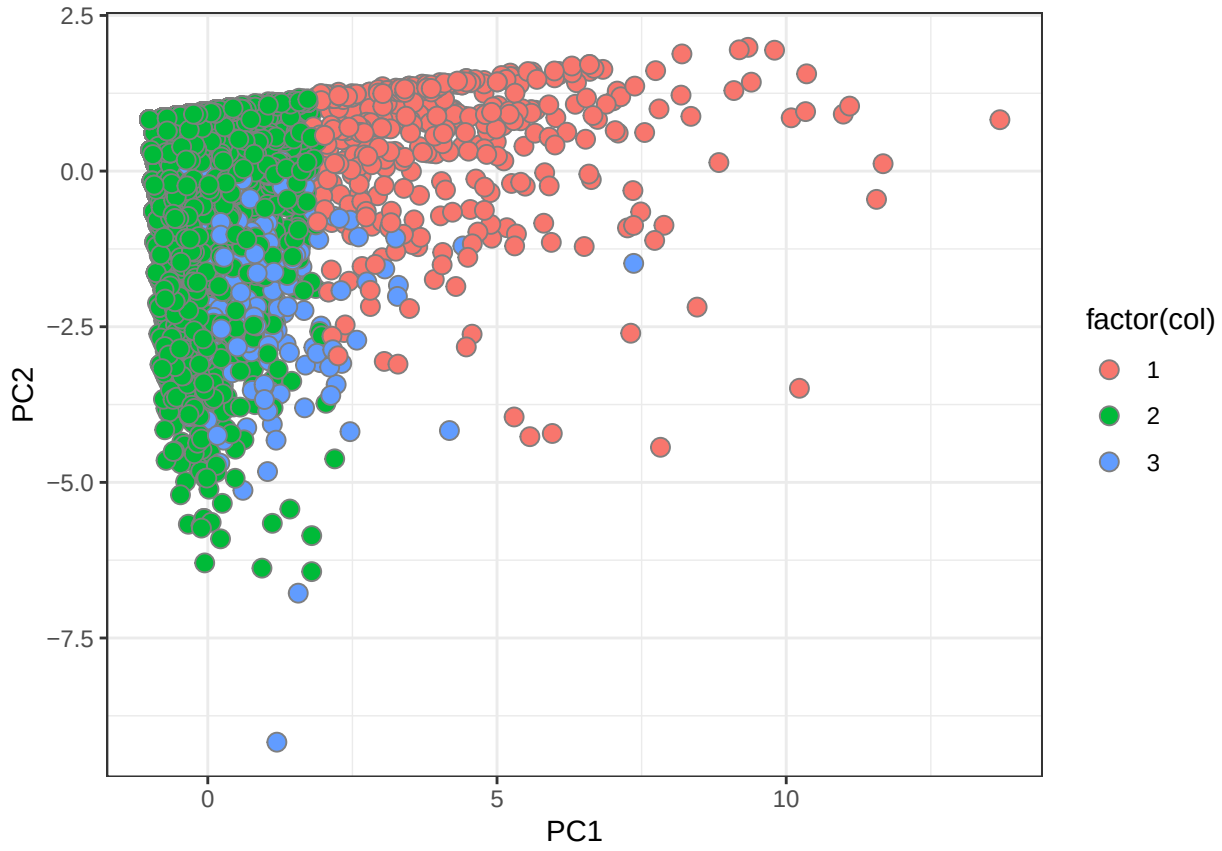
```
##      politics     travel      school      beauty         art
## 1 -0.19572837 -0.1808596  1.23985885  1.26962059 -0.08900009
## 2 -0.05124863  0.2661466  0.13111182 -0.01769018  3.74215419
## 3 -0.20037142 -0.2186171 -0.33186097 -0.32436755 -0.20093181
## 4  2.96827449  2.9051058 -0.01004564 -0.08008951 -0.10407236
```

```
## [1] "Lets look at the clusters for this set"
```

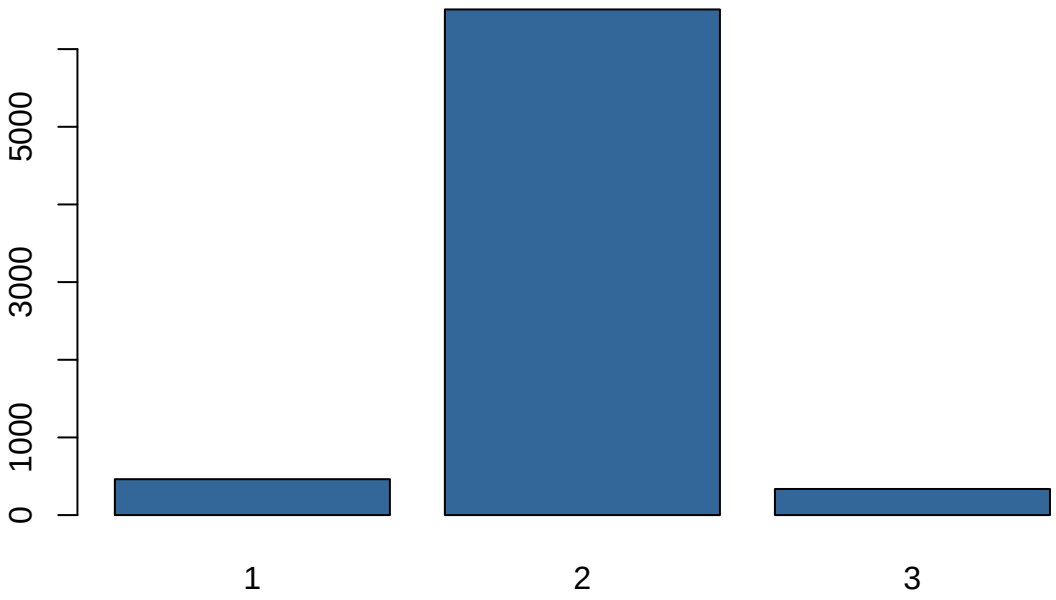



```
## [1] "Now lets try to combine PCA with Clustering for the most correlated variables"
```

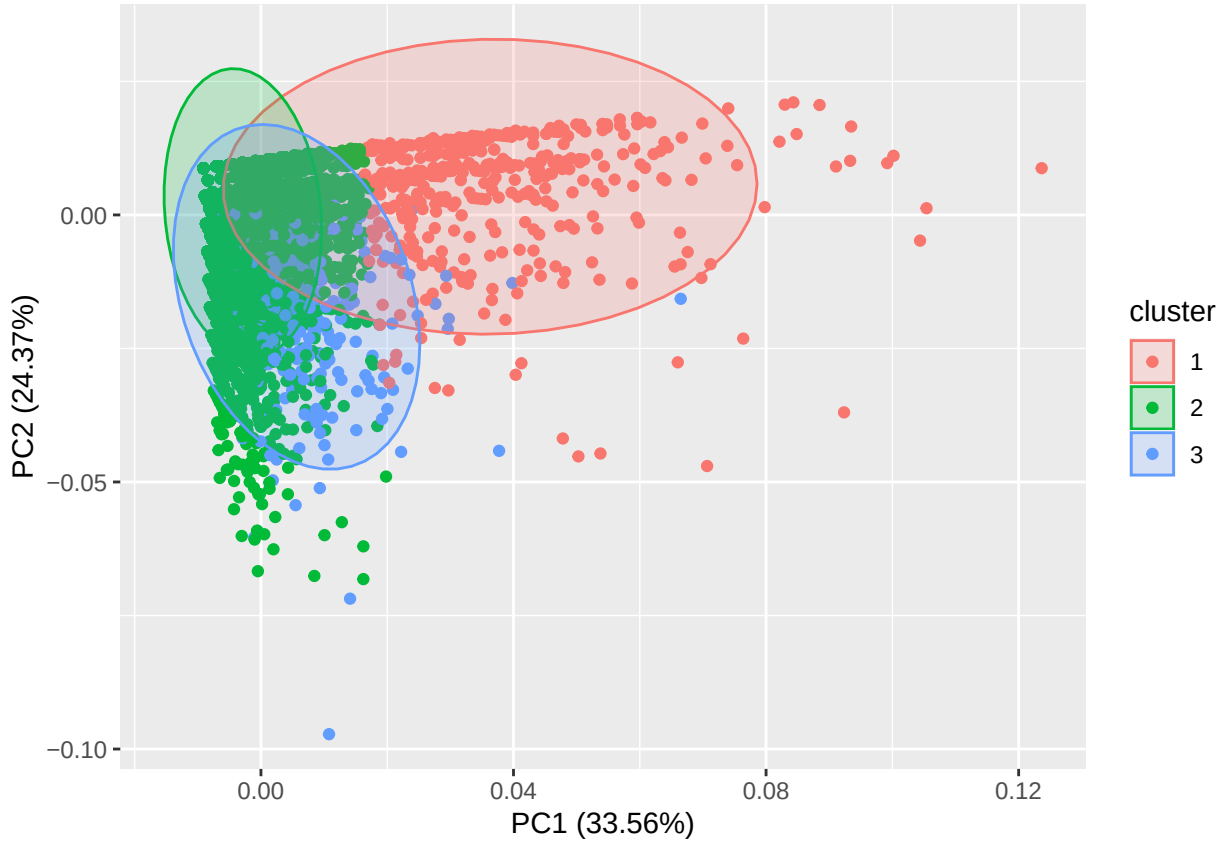
```
## [1] "We want to examine the cluster memberships for each #observation"
```



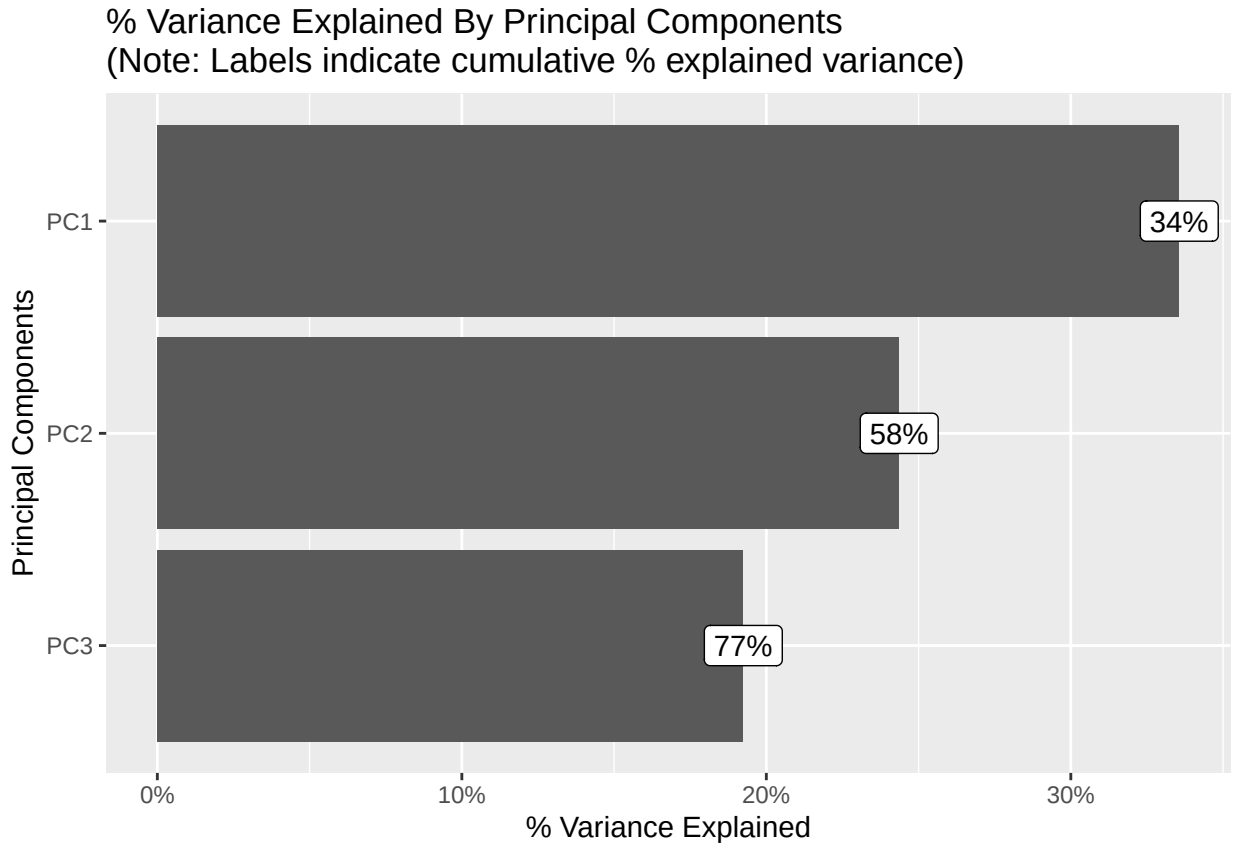

```
## [1] "Number of members in each cluster"
```

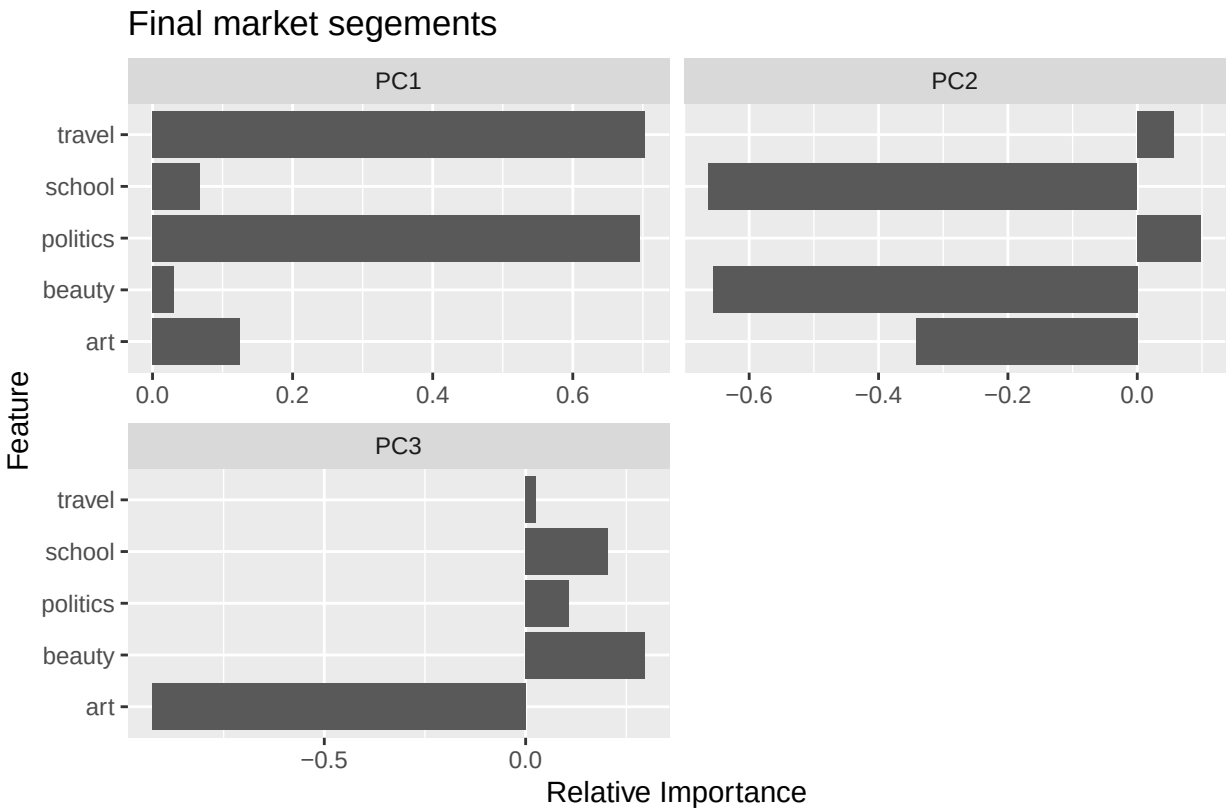



```
## [1] "Plotting the cluster"
```




```
## [1] "The Final Market segments"
```


% Variance Explained By Principal Components
(Note: Labels indicate cumulative % explained variance)
Principal Components
PC1
PC2
PC3
34%
58%
77%
0%
10%
20%
30%
% Variance Explained



We observed that using PCA in combination with Kmeans clustering gave us the best separation for components, we can use these market segments as our final set.

The Reuters corpus

Revisit the Reuters C50 text corpus that we briefly explored in class. Your task is simple: tell an interesting story, anchored in some analytical tools we have learned in this class, using this data. For example:
- you could cluster authors or documents and tell a story about what you find.
- you could look for common factors using PCA.
- you could train a predictive model and assess its accuracy. (Yes, this is a supervised learning task, but it potentially draws on a lot of what you know about unsupervised learning, since constructing features for a document might involve dimensionality reduction.)
- you could do anything else that strikes you as interesting with this data.

Describe clearly what question you are trying to answer, what models you are using, how you pre-processed the data, and so forth. Make sure you include at least *one* really interesting plot (although more than one might be necessary, depending on your question and approach.)

Format your write-up in the following sections, some of which might be quite short:
- Question: What question(s) are you trying to answer? - Approach: What approach/statistical tool did you use to answer the questions? - Results: What evidence/results did your approach provide to answer the questions? (E.g. any numbers, tables, figures as appropriate.) - Conclusion: What are your conclusions about your questions? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set.

Regarding the data itself: In the C50train directory, you have 50 articles from each of 50 different authors (one author per directory). Then in the C50test directory, you have another 50 articles from each of those

same 50 authors (again, one author per directory). This train/test split is obviously intended for building predictive models, but to repeat, you need not do that on this problem. You can tell any story you want using any methods you want. Just make it compelling!

Note: if you try to build a predictive model, you will need to figure out a way to deal with words in the test set that you never saw in the training set. This is a nontrivial aspect of the modeling exercise. (E.g. you might simply ignore those new words.)

This question will be graded according to three criteria:
1. the overall "interesting-ness" of your question and analysis.
2. the clarity of your description. We will be asking ourselves: could your analysis be reproduced by a competent data scientist based on what you've said? (That's good.) Or would that person have to wade into the code in order to understand what, precisely, you've done? (That's bad.)
3. technical correctness (i.e. did you make any mistakes in execution or interpretation?)

Solution:

- **Question:** What are the most frequently used words across the entire dataset and are they the same words which are used in the most number of documents?
- **Approach:** We used the DocumentTermMatrix to calculate the frequencies of word across the entire dataset across authors and we used the inverse document frequency to calculate the usage of these words across documents and plotted these on word cloud to understand the same.

```
## [1] "load data for all authors"
```

```
## [1] "list of directory names/Authors"
```

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

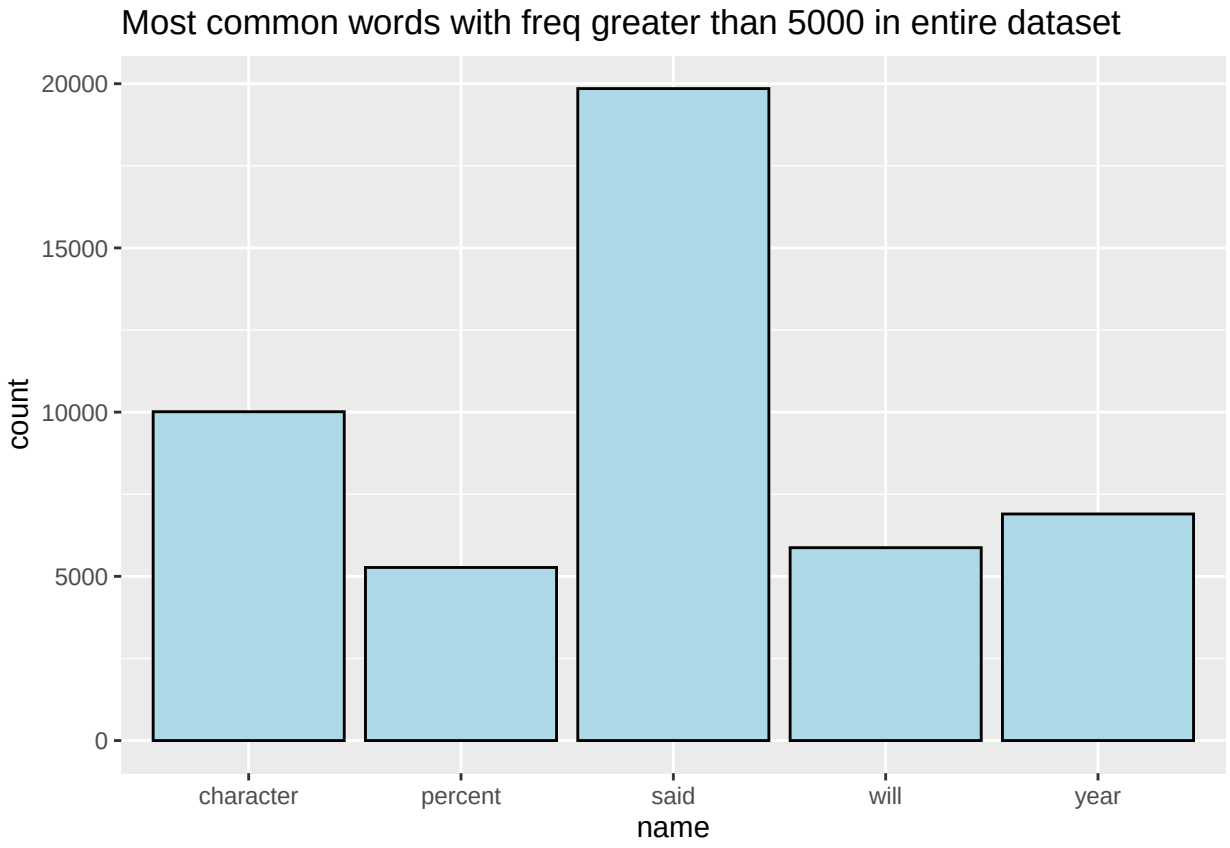
```
## [1] 801
```

```
## [1] "most frequent and least frequent words across the entire dataset"
```

```
##          name count
##  1:      said 19851
##  2: character 10013
##  3:      year  6899
##  4:      will  5873
##  5:   percent  5270
##  6:   million  4848
##  7:       new  3508
##  8:    market  3215
##  9:   company  3200
## 10:   billion  3012
```

```
##           name count
##  1:      thing   130
##  2:     cchina   132
##  3:     moving   133
##  4: considered   133
##  5:     either   134
##  6:    success   134
```

```
##  7:       ever   137
##  8:    initial   138
##  9:     figure   138
## 10:    quickly   139
```




```
## [1] "Word cloud of words with min frequency of 1000 in entire dataset"
```

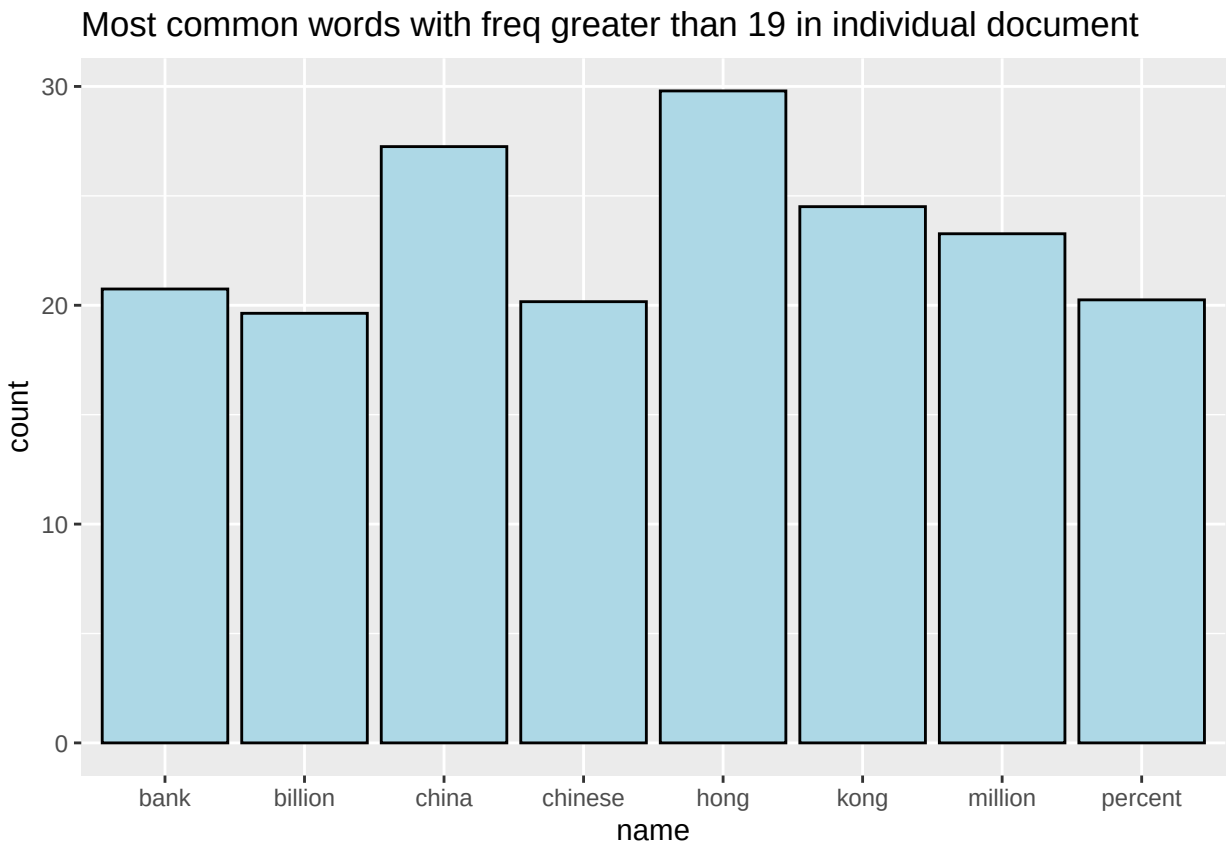

```
## [1] 801
```

```
## [1] "most frequent and least frequent words by number of documents"
```

```
##        name    count
##  1:    hong 29.79466
##  2:   china 27.24980
##  3:    kong 24.50645
##  4: million 23.26764
##  5:    bank 20.74283
##  6: percent 20.24720
##  7: chinese 20.16345
##  8: billion 19.63257
##  9:   sales 18.38453
## 10: company 18.26743
```

```
##             name count
##  1:     character     0
##  2: datetimestamp     0
##  3:   description     0
##  4:       heading     0
##  5:          hour     0
##  6:         isdst     0
##  7:      language     0
##  8:   listcontent     0
```

```
##  9:          mday     0
## 10:          meta     0
```




```
## [1] "Word cloud of words with min frequency of 15 in individual dataset"
```

- **Results:** The word clouds have been generated and we can see that these are not the same, which means that the words which have the highest frequency are not he ones which have been used in most number of documents.
- **Conclusion:** We can conclude that the data distribution is such that some words are used much more frequently within certain documents here.
- **Question:** Based on TF-IDF values, can we predict the author of a document?
- **Approach:** Let's first start by getting names of all the authors -

```
##  [1] "AaronPressman"     "AlanCrosby"        "AlexanderSmith"
##  [4] "BenjaminKangLim"   "BernardHickey"     "BradDorfman"
##  [7] "DarrenSchuettler"  "DavidLawder"       "EdnaFernandes"
## [10] "EricAuchard"       "FumikoFujisaki"    "GrahamEarnshaw"
## [13] "HeatherScoffield"  "JanLopatka"        "JaneMacartney"
## [16] "JimGilchrist"      "JoWinterbottom"    "JoeOrtiz"
## [19] "JohnMastrini"      "JonathanBirt"      "KarlPenhaul"
## [22] "KeithWeir"         "KevinDrawbaugh"    "KevinMorrison"
## [25] "KirstinRidley"     "KouroshKarimkhany" "LydiaZajc"
## [28] "LynneO'Donnell"    "LynnleyBrowning"   "MarcelMichelson"
## [31] "MarkBendeich"      "MartinWolk"        "MatthewBunce"
## [34] "MichaelConnor"     "MureDickie"        "NickLouth"
## [37] "PatriciaCommins"   "PeterHumphrey"     "PierreTran"
## [40] "RobinSidel"        "RogerFillion"      "SamuelPerry"
## [43] "SarahDavison"      "ScottHillis"       "SimonCowell"
```

```
## [46] "TanEeLyn"          "TheresePoletti"    "TimFarrand"
## [49] "ToddNissen"        "WilliamKazer"
```

Now, let's create a vectorized matrix of word counts from the entire corpus against each author (Bag of Words model where 'author' is the dependent variable).

What does this matrix consist of?

```
## <<DocumentTermMatrix (documents: 5000, terms: 45853)>>
## Non-/sparse entries: 1083147/228181853
## Sparsity           : 100%
## Maximal term length: 45
## Weighting          : term frequency (tf)
```

It's a very sparse matrix.

Let's remove any words which are in the bottom 1% by count (select the top 99%) and see how our matrix is affected -

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

```
## <<DocumentTermMatrix (documents: 5000, terms: 3378)>>
## Non-/sparse entries: 844216/16045784
## Sparsity           : 95%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

Wow! That's quite a reduction! Now that we've removed a lot of such randomly occurring words (variance), let's train a Multinomial Naive Bayes model and look at it's accuracy -

```
## [1] 0.793
```

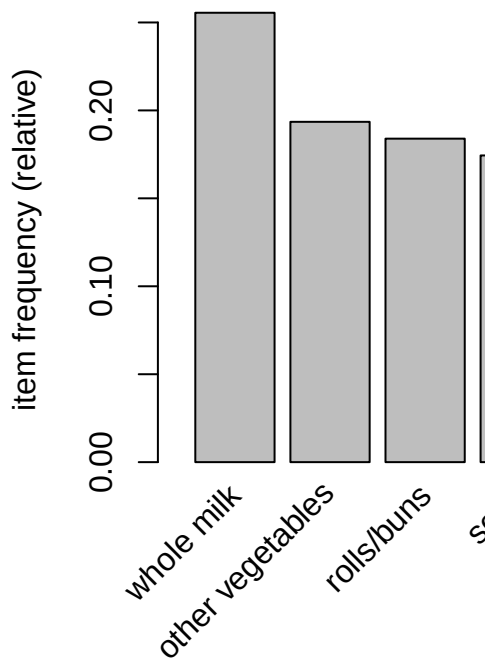
We get almost 80% accuracy (79.3%) on a simple Naive Bayes model.

Association rule mining

Revisit the notes on association rule mining and the R example on music playlists: playlists.R and playlists.csv. Then use the data on grocery purchases in groceries.txt and find some interesting association rules for these shopping baskets. The data file is a list of shopping baskets: one person's basket for each row, with multiple items per row separated by commas. Pick your own thresholds for lift and confidence; just be clear what these thresholds are and say why you picked them. Do your discovered item sets make sense? Present your discoveries in an interesting and visually appealing way.

Notes: - This is an exercise in visual and numerical story-telling. Do be clear in your description of what you've done, but keep the focus on the data, the figures, and the insights your analysis has drawn from the data, rather than technical details.
- The data file is a list of baskets: one row per basket, with multiple items per row separated by commas. You'll have to cobble together your own code for processing this into the format expected by the "arules" package. This is not intrinsically all that hard, but it is the kind of data-wrangling wrinkle you'll encounter frequently on real problems, where your software package expects data in one format and the data comes in a different format. Figuring out how to bridge that gap is part of the assignment, and so we won't be giving tips on this front.

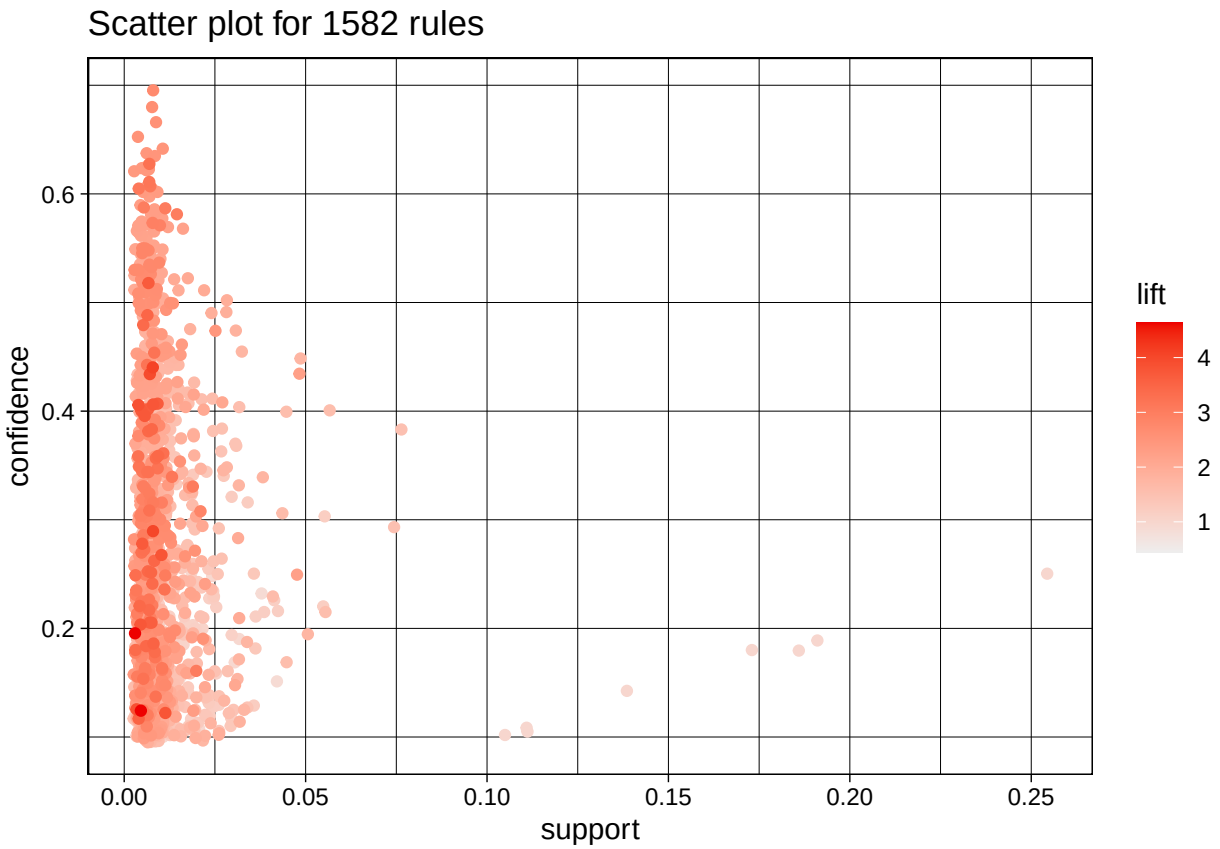
Solution:



From the given list, let's first look at which items are bought the most by frequency -

Now, let's look at a graph of confidence vs support for all the rules we obtain from apriori -

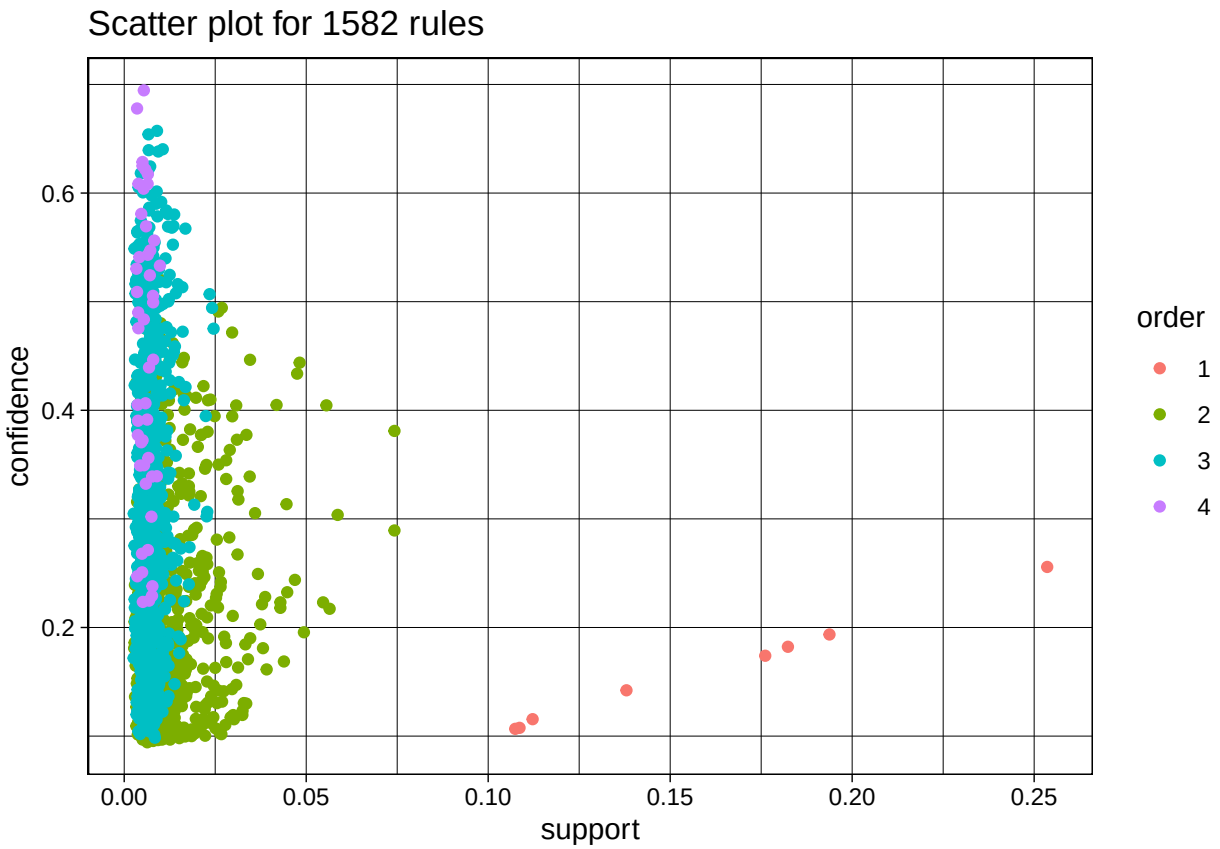
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5   0.005      1
##  maxlen target  ext
##       4  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 49
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [1582 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```



We can see most of the points that have low support have high confidence.

We can also see that a majority of rules that have high lift (3 or more) tend to have lower confidence.

Let's color the points based on order of the rules -



Most of the rules are of order 2 and 3. Most of order 1 rules have very low confidence and most of order 4 rules have relatively high confidence and low support.

Let's look at a graphical representation of all rules that have a confidence of more than 0.15 and a lift of more than 2 -

A majority of rules contain whole milk, vegetables, root vegetables, yogurt, eggs, etc.

Items like napkins, buttermilk, salty snacks, berries, etc are not bought too often as compared to other items and they have low confidence as well. Items like sliced cheese, oil etc aren't bought that often either but have comparatively high confidence.

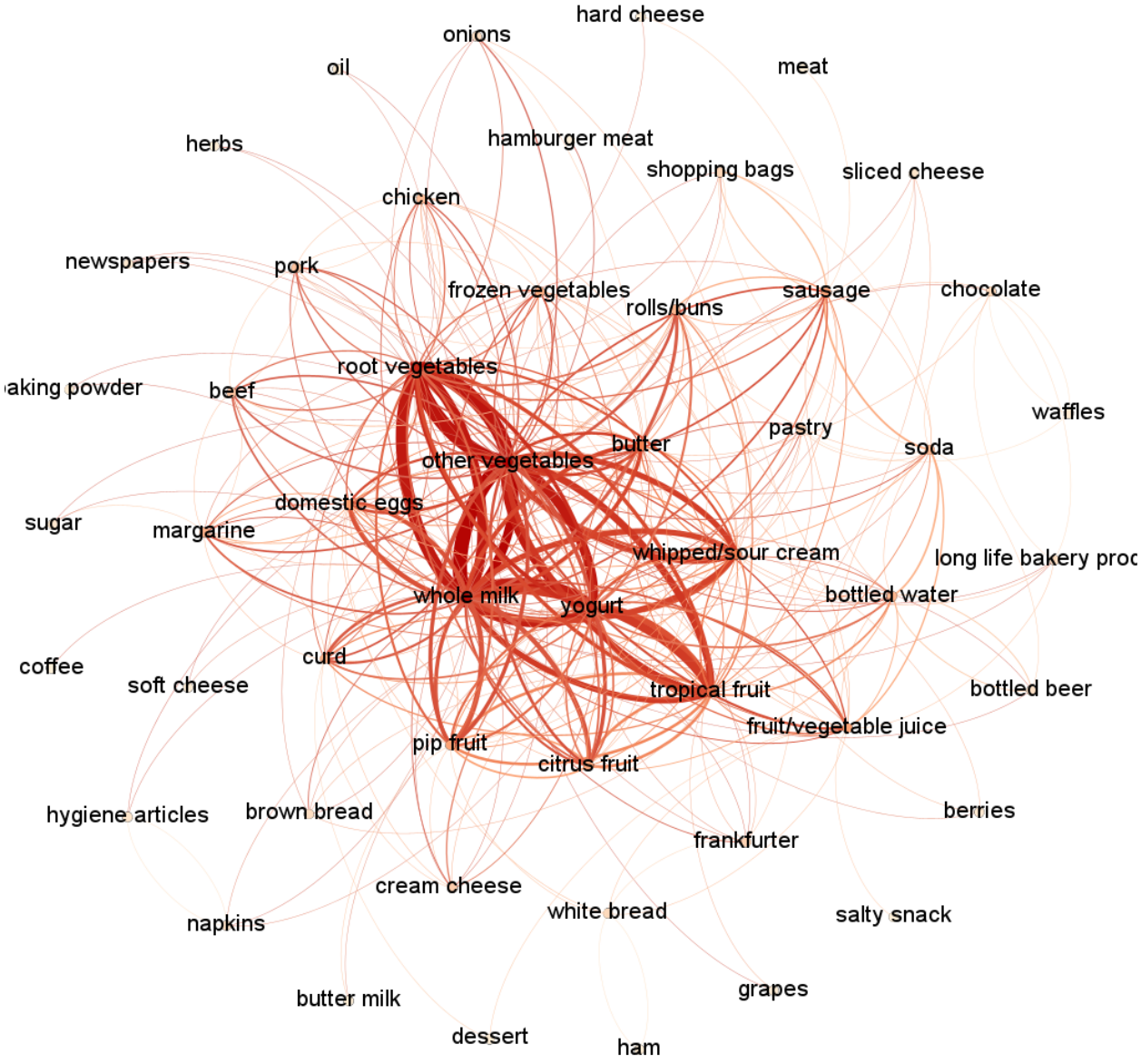


Figure 1: Graph weighted by confidence