Stat 240 Week 12
Last Steps

Week 12

Dr. Dave Campbell

Some links:

http://www.sfu.ca/bin/wcm/course-outlines?

http://www.sfu.ca/bin/wcm/course-outlines?2014

http://www.sfu.ca/bin/wcm/course-outlines?2014/fall/

http://www.sfu.ca/bin/wcm/course-outlines?2014/fall/stat

http://www.sfu.ca/bin/wcm/course-outlines?2014/fall/stat/285

http://www.sfu.ca/bin/wcm/course-outlines?2014/fall/stat/285/d900

JSON workflow

library(jsonlite)

To get a list of years: – {baseURL}

- To get a list of terms:

– {baseURL}?{year}

- To get a list of departments:

– {baseURL}?{year}/{term}

- To get a list of course numbers:

– {baseURL}?{year}/{term}/{department}

- To get a list of sections:

– {baseURL}?{year}/{term}/{department}/{courseNumber}

To get the actual course outline:
– {baseURL}?{year}/{term}/{department}/{courseNumber}/{courseSection}

.

```
baseURL = "http://www.sfu.ca/bin/wcm/course-outlines?"
for(year in fromJSON(baseURL)$value){
    URLthisYear = paste(baseURL,year,sep="")
    TermsthisYear = fromJSON(URLthisYear)
    for(term in TermsthisYear$value){
        DeptsThisTerm = fromJSON(paste(URLthisYear,term,sep="/"))
        for(Dept in DeptsThisTerm$value){
            CoursesThisTerm = fromJSON(paste(URLthisYear,term,Dept,sep="/"))
            for(Course in CoursesThisTerm $value){
                SectionThisTerm =
                fromJSON(paste(URLthisYear,term,Dept,Course,sep="/"))
                for(Section in SectionThisTerm $value){
                    CourseOutline=
                    (fromJSON(paste(URLthisYear,term,"econ",Course,"d100",sep="/")))
                    #Store the pieces you need…
```

Data frame to JSON

```
MCU_JSON = toJSON(MergedTable,    pretty =
TRUE)#Make json

#Look at json pieces:

length(course_info)
# names of each element

names(course_info)

# class of each element

lapply(course_info, class)  #apply to all list elements
```

JSON use cases

Sharing and importing information of different sizes across software

Frequently used in IOT and data from APIs

STAT 240: obtaining, cleaning, and not running away from modern data types

STAT 380: Stochastic Processes (time between tweets, queueing theory)

Stat 350: Fitting lines to data (regression models)

STAT 445: Multivariate Analysis

STAT 475: Discrete Data Models

STAT 440: Learning from Big Data (full of case studies and modern tools)

STAT 452: Statistical Learning and Prediction (unsupervised/ supervised learning)

Unsupervised Learning (Clustering)

Goal: find natural groupings in the data

Different types of shoppers / students

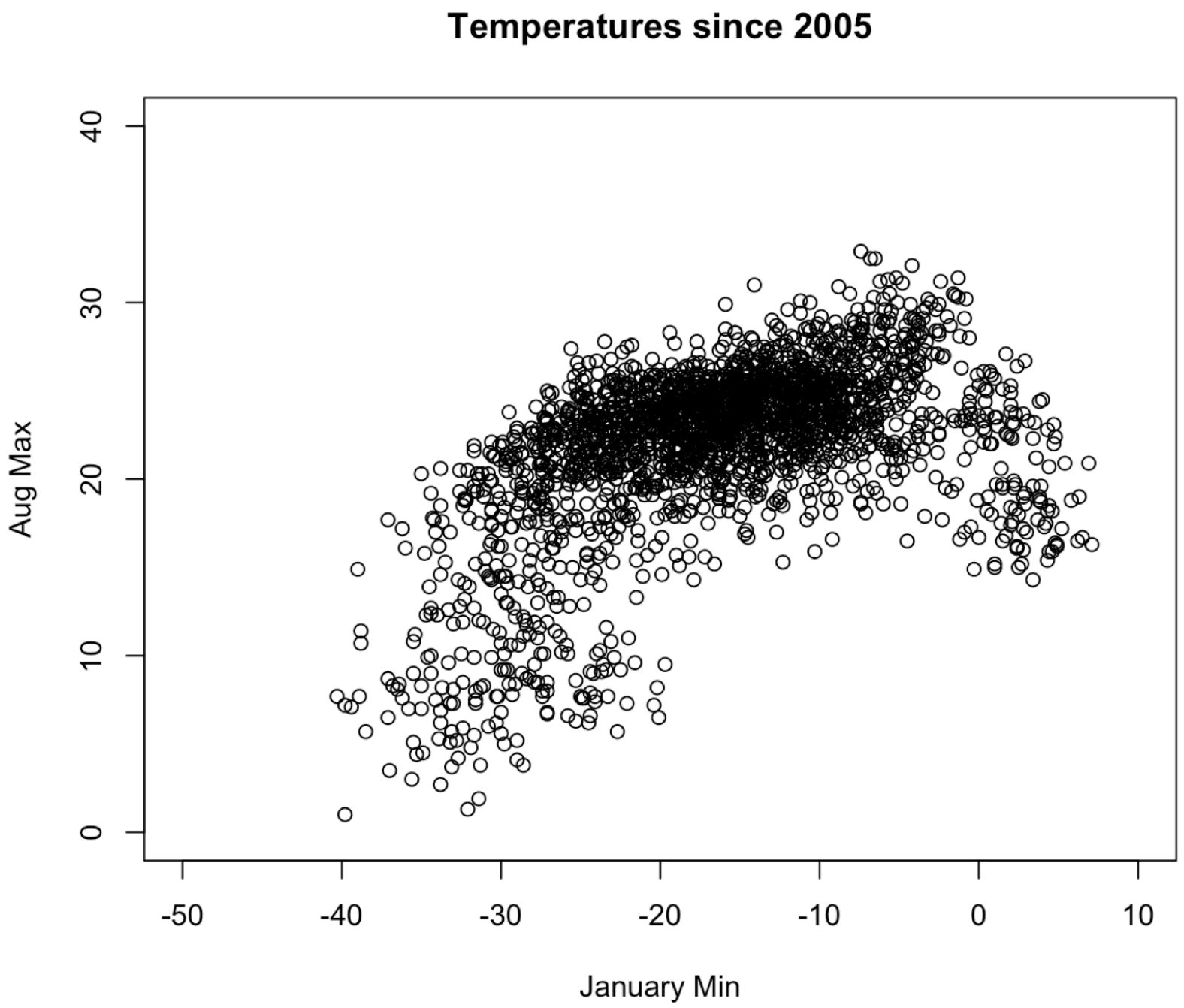
Different types of preferences

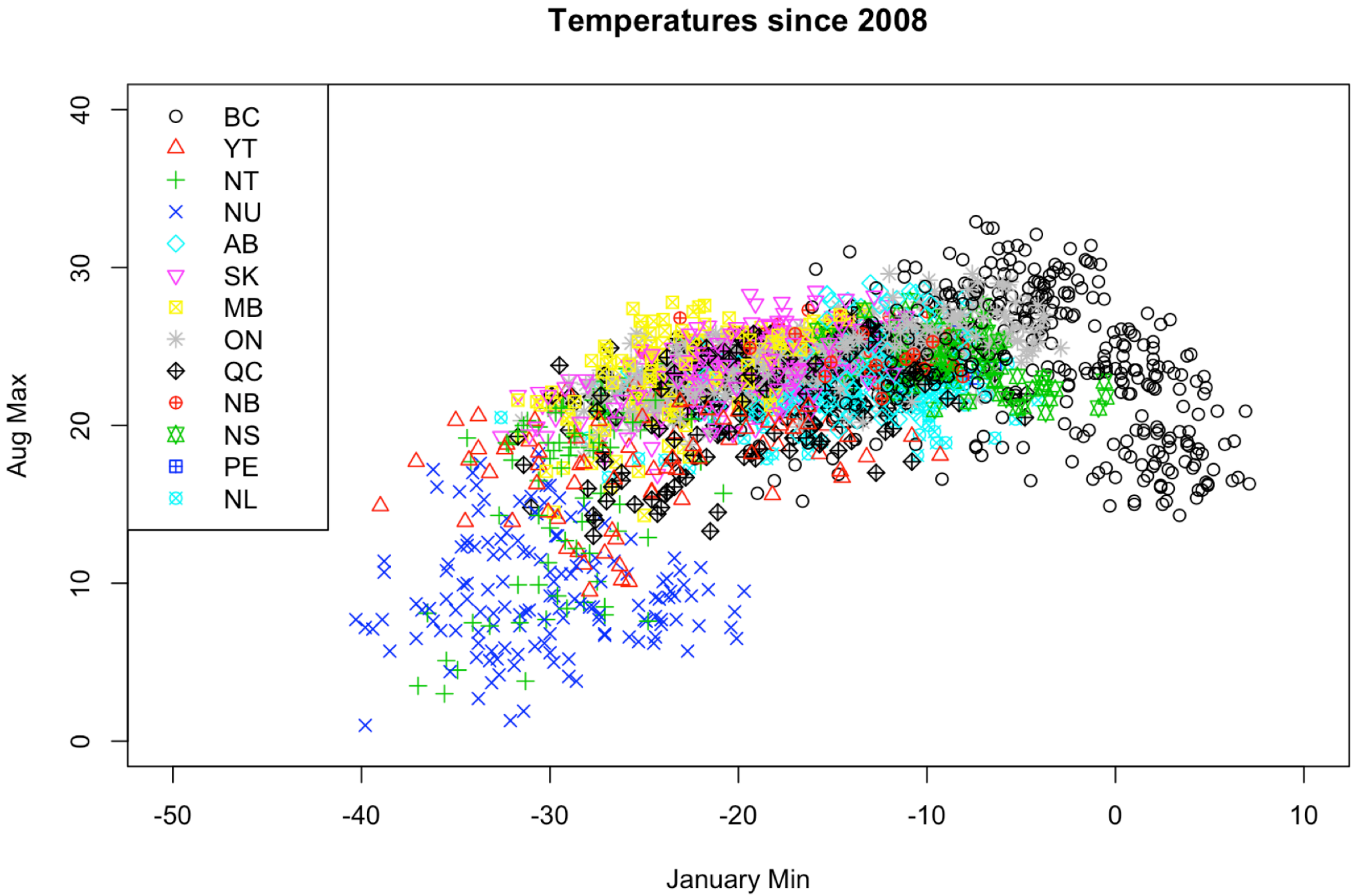
Can we find different climate zones?

Coastal temperatures are moderated by the water

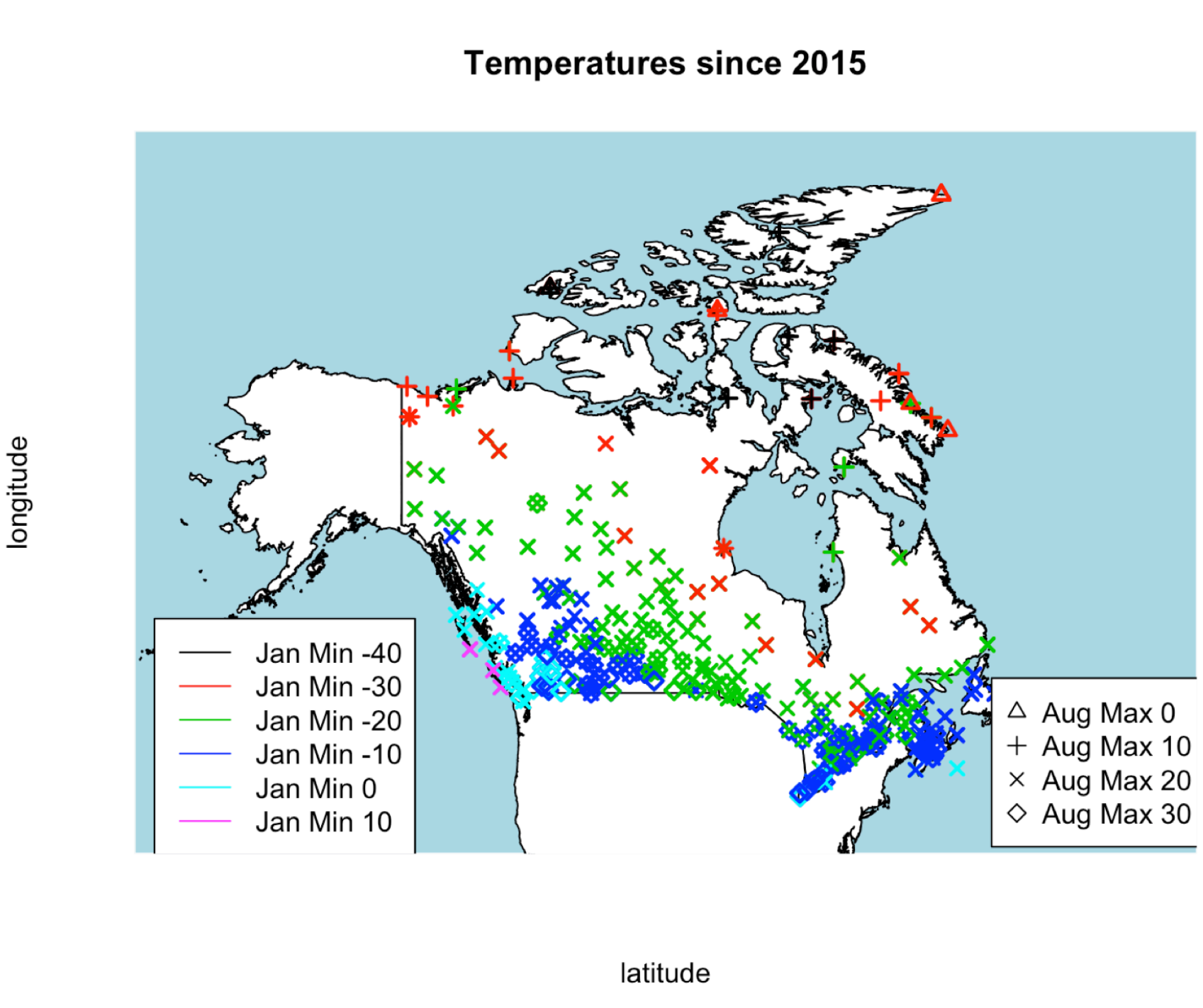
Continental temperatures are driven to arctic air masses and sun heated Earth

Clustering (Unsupervised Learning)




Temperatures since 2008
BC
YT
NT
NU
AB
SK
MB
ON
QC
NB
NS
PE
NL
Aug Max
0
10
20
30
40
-50
-40
-30
-20
-10
0
10
January Min

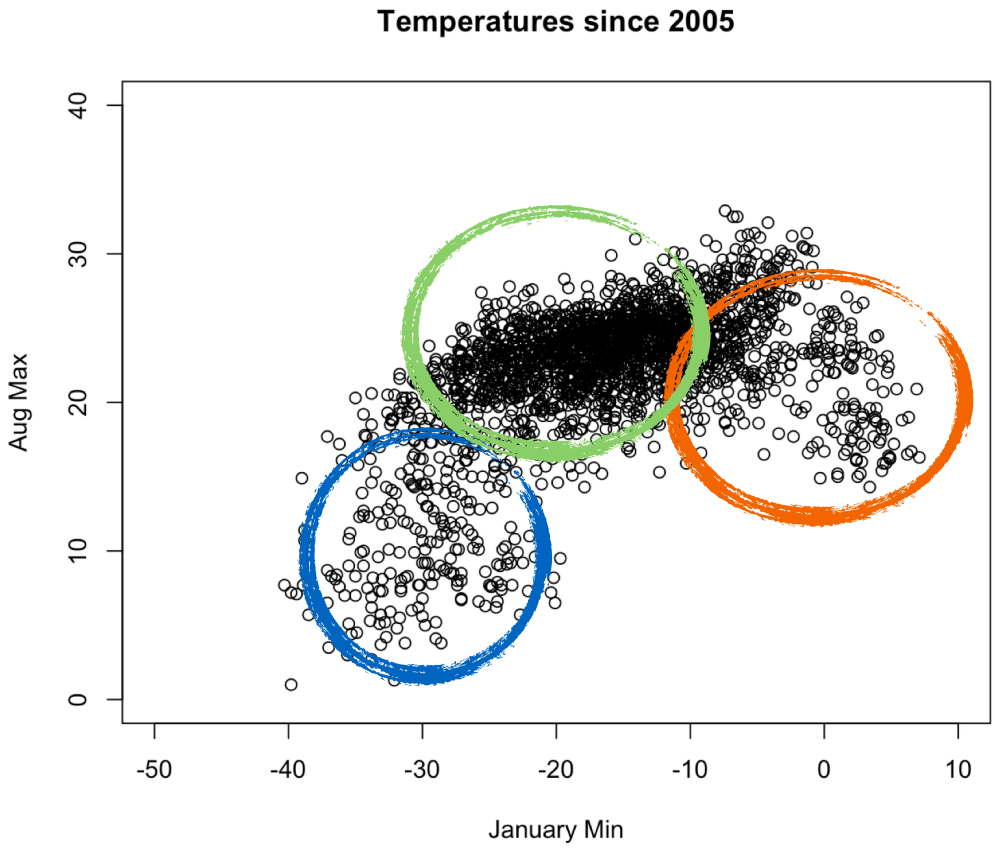
```
index2 = index&!is.na(temps[,"JanMin"])&!
is.na(temps[,"AugMax"])

temps2use =
as.matrix(temps[index2,c("JanMin","AugMax")])
```


Temperatures since 2015
longitude
latitude
Jan Min -40
Jan Min -30
Jan Min -20
Jan Min -10
Jan Min 0
Jan Min 10
Aug Max 0
Aug Max 10
Aug Max 20
Aug Max 30

kmeans

idea: data comes from a mixture of k Normal distributions; here k=3

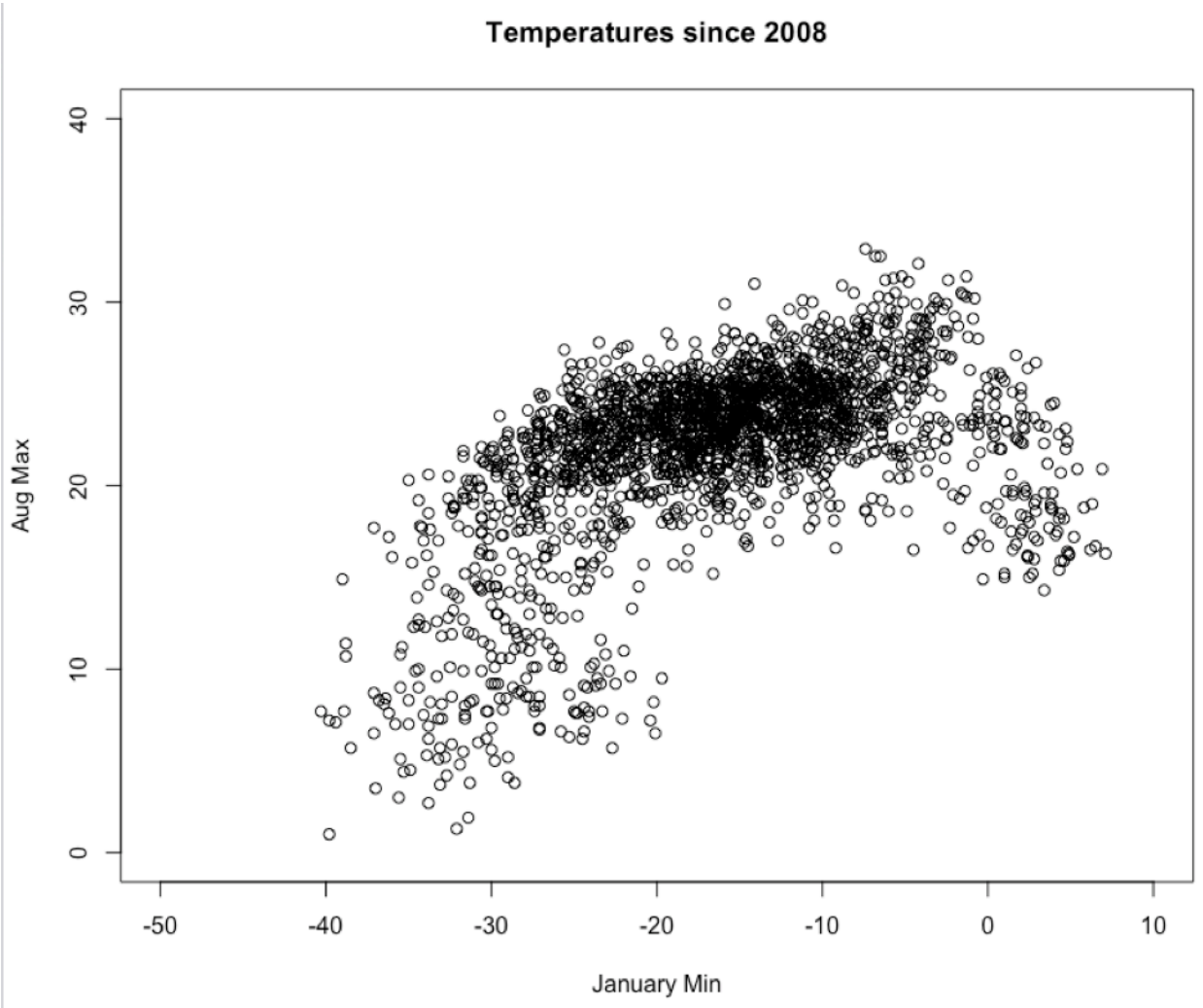
$$Y \sim p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2) + p_3 N(\mu_3, \sigma_3^2)$$



idea: data comes from a mixture of k Normal distributions; here k=3

$$Y \sim p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2) + p_3 N(\mu_3, \sigma_3^2)$$

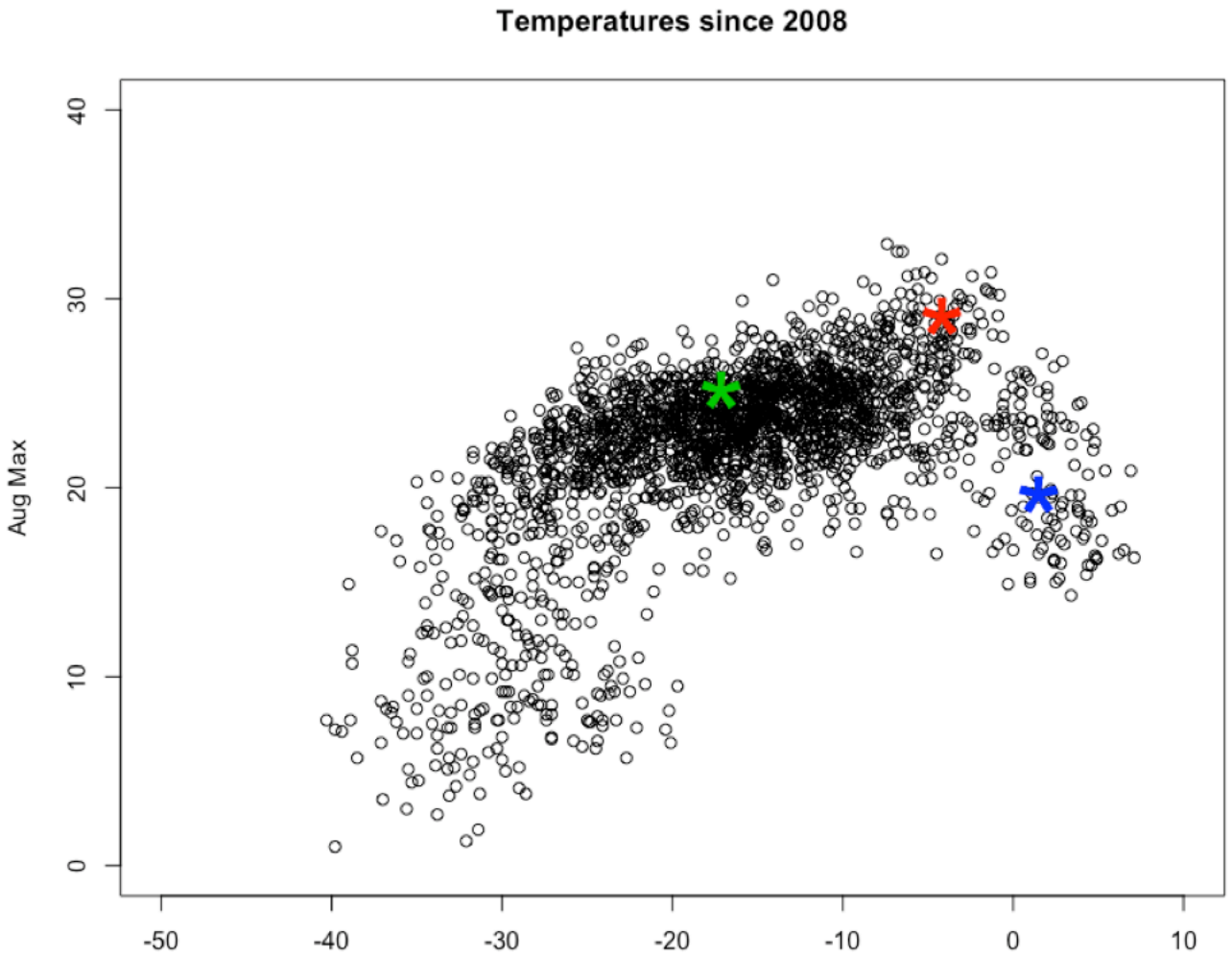
Raw data:



kmeans

kmeans algorithm:

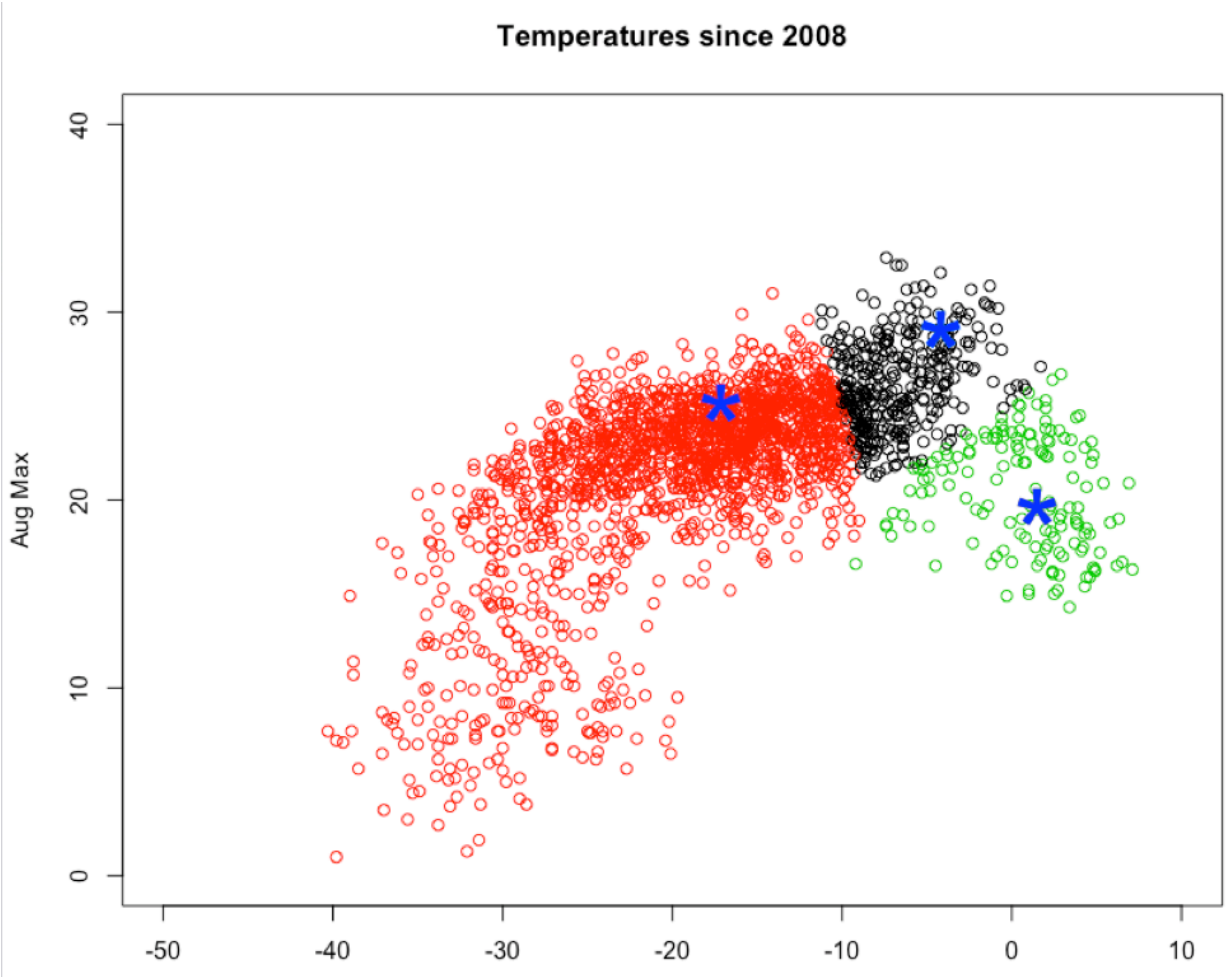
1: Randomly select k observations to be the centres



kmeans

kmeans algorithm:

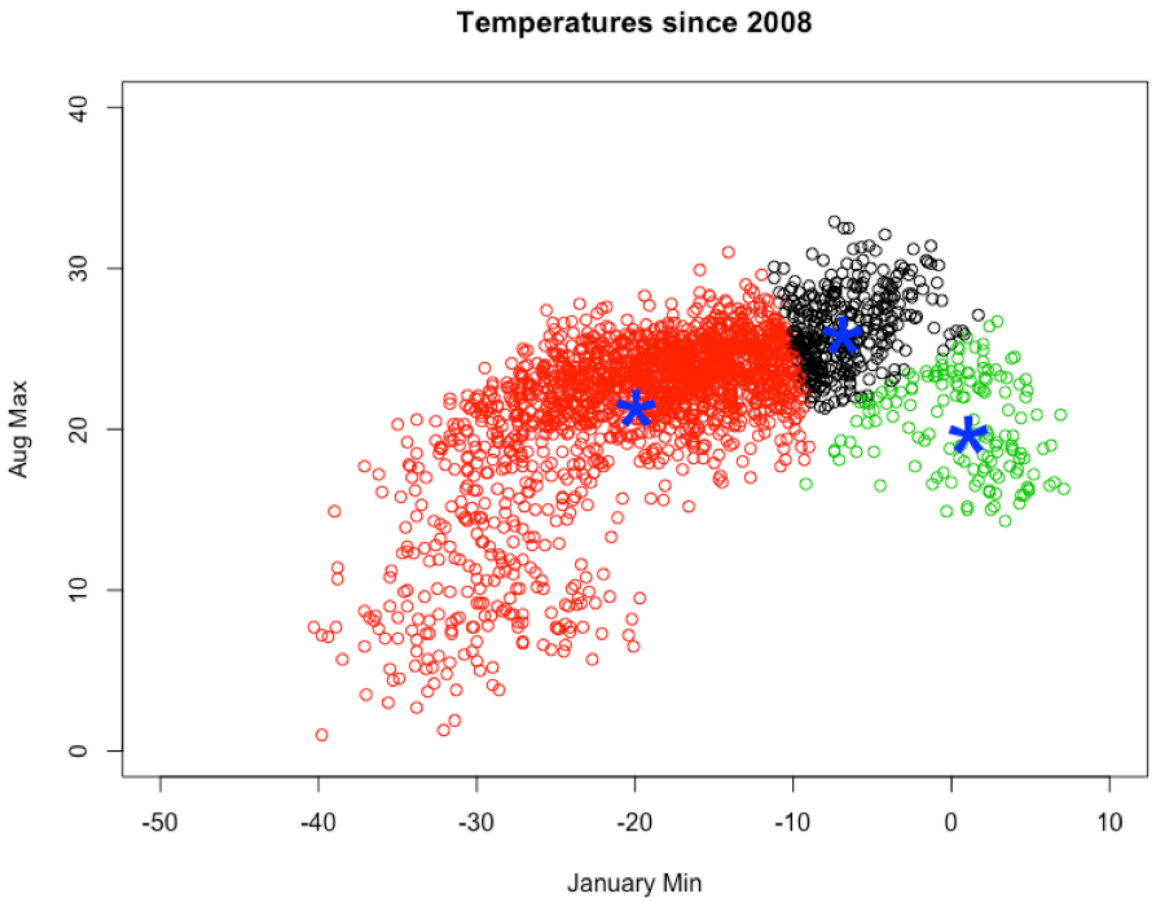
2: For all other observations, find the nearest centre



kmeans

kmeans algorithm:

3: Find the middle most point within each group



kmeans

kmeans algorithm:

iterate:

Find group Allocations

Find the mid point

kout = kmeans(x=temps2use, centers=3)

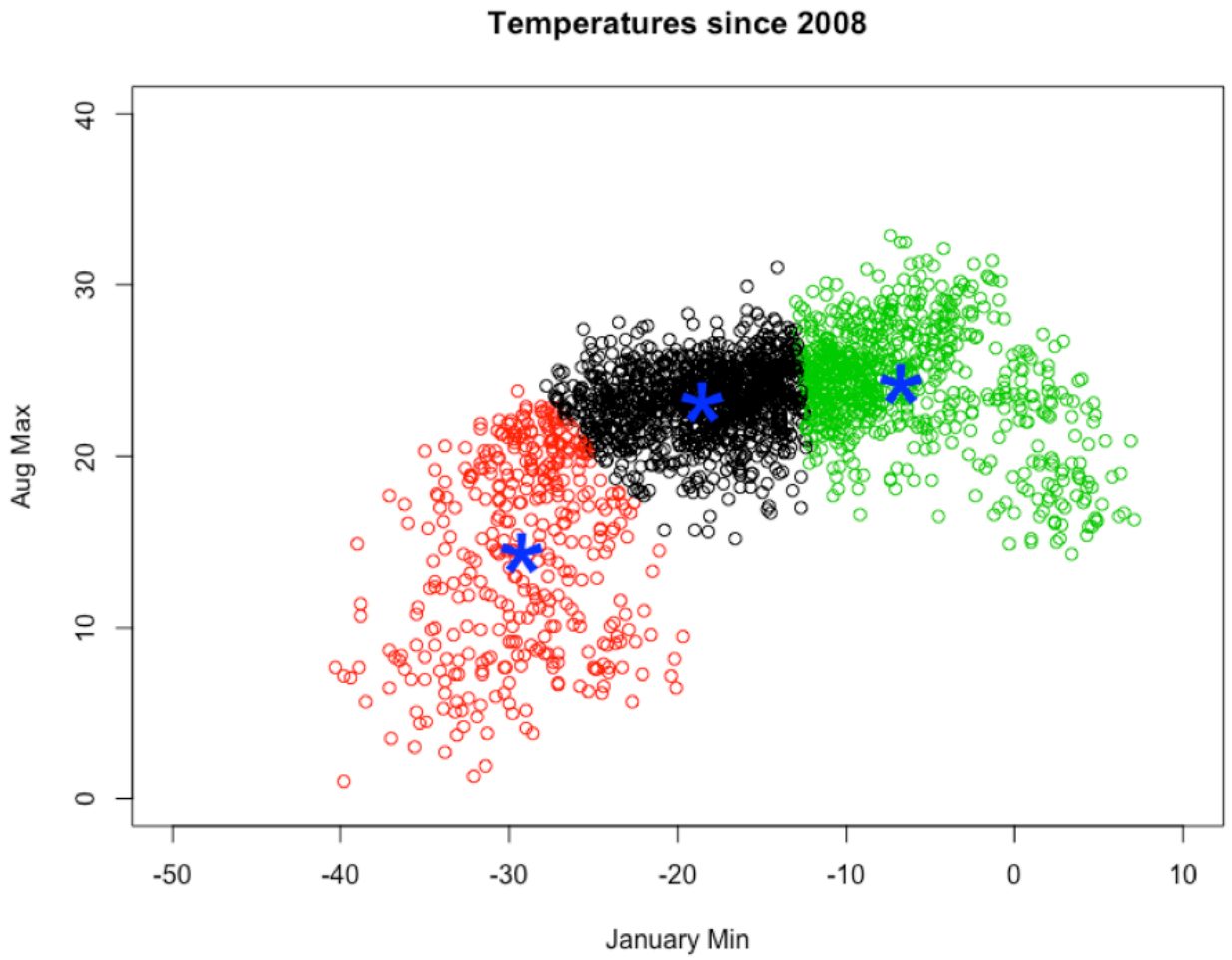
kout = kmeans(x, centers)

kout = kmeans(x=temps2use, centers=3)

```
> attributes(kout)
$names
[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"

$class
[1] "kmeans"
```

```
plot(temps2use[,"JanMin"],temps2use[,"AugMax"],

    main = "Temperatures since 2008",     ylab =
"Aug Max",     xlab = "January Min",     ylim =
c(0,40),     xlim = c(-50,10),col=kout$cluster)


points(kout$centers[,"JanMin"],kout$centers[,"Aug
Max"],       col=4,cex=5,       lwd=2,pch="*")
```

End result:


Temperatures since 2008
Aug Max
0
10
20
30
40
-50
-40
-30
-20
-10
0
10
January Min

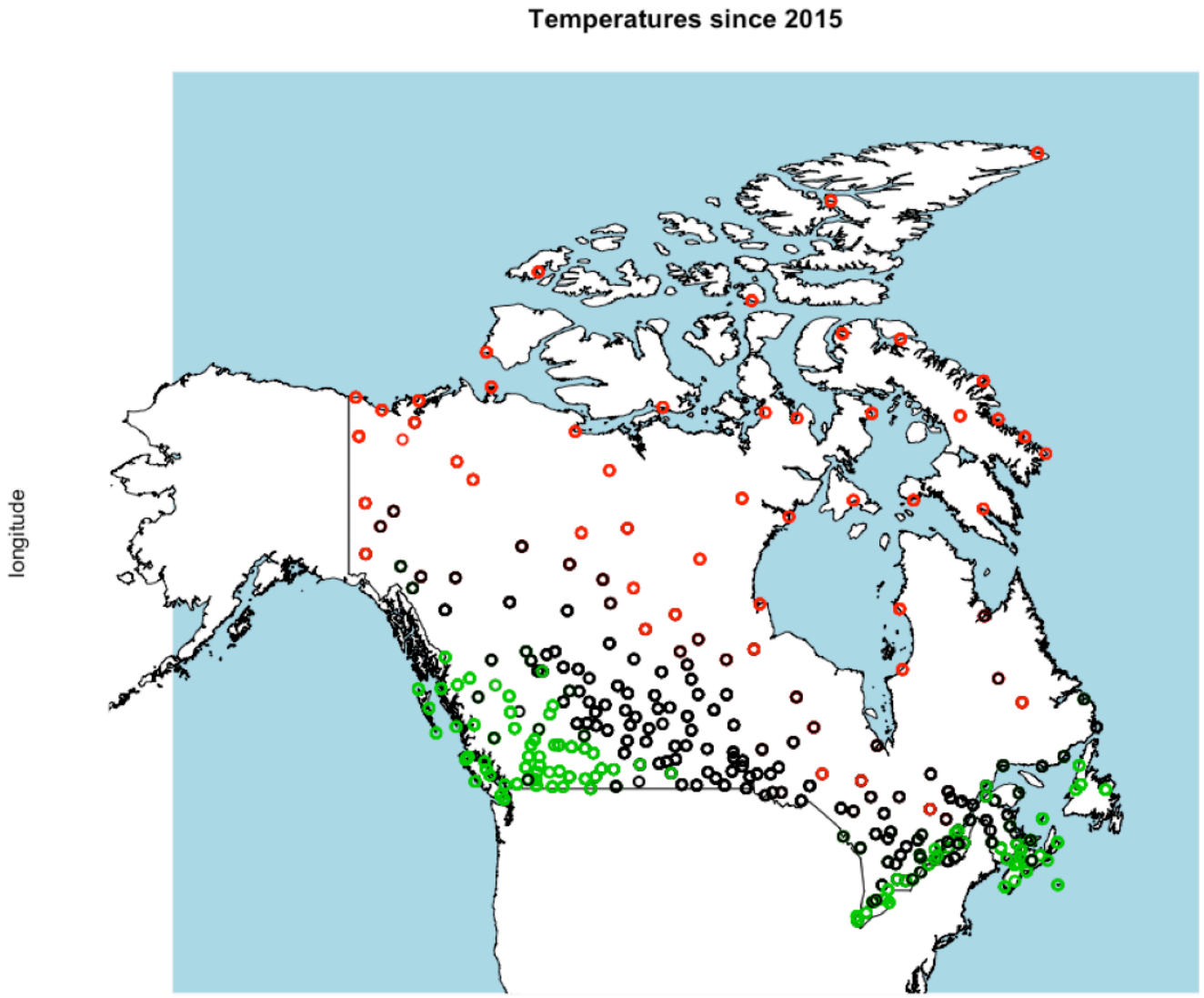
kout$size = cluster sizes

```
[1] 1095  383  733
```

kout$**centres**

```
      JanMin   AugMax
1 -18.497352 23.21352
2 -29.167885 14.41645
3  -6.691542 24.25007
```

Clusters



Risks

Choice of number of clusters is vital to analysis

Initialization is random ==> the end point might be random too

Even for a fixed k, you might not be at the best answer.

Software will always give an answer

Cluster names/interpretations will change with rerun

Benefits

Unsupervised means there is no known right answer to compare with, so result helps you explore

With more exploration and data insight explanations can be formed

New individuals can be assigned to a cluster without re-running the algorithm

Is this answer any good?

Best answer take STAT 440 or STAT 452

For now, treat selecting the number of clusters as an exploratory tool

Finding text documents that are similar

Within a topic certain words occur together more frequently:

{Data Science: [statistics, computing, machine learning, data],

Environment: [climate change, green energy, carbon tax, sea level rise],

US President: [collusion, definitely not lying, fake news]}

Topic Modelling

There are many excellent tutorials:

https://www.tidytextmining.com/topicmodeling.html

https://rpubs.com/wsevier/LDA

Unsupervised Learning

Idea:

Find groups of documents that have similar word counts relationships and consider them to be **clusters**

Here clusters are interpreted as **topics** and words are more common for certain topics

A document might have several different topics: for example:

40% data science, 55% Software Engineering,…

20% Environment 60% US President…

Latent Dirichlet Allocation

Consider the job ads from the first week of class

strip out formatting, set words to lowercase, remove stop words, stem words (remove ‘ing’ ‘ed’,…)

As with kmeans you must choose the number of topics

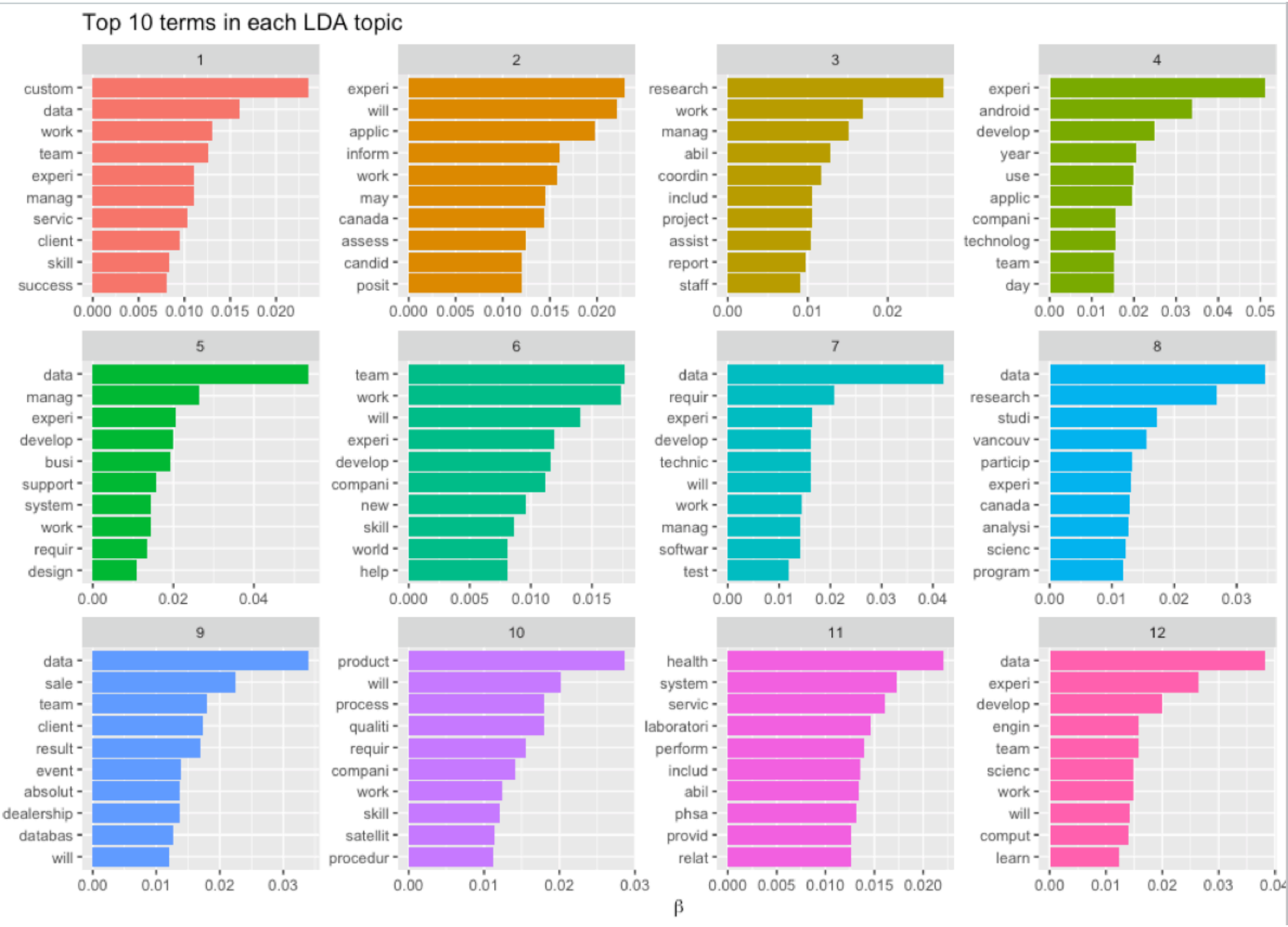
dtm = DocumentTermMatrix(myCorpus)

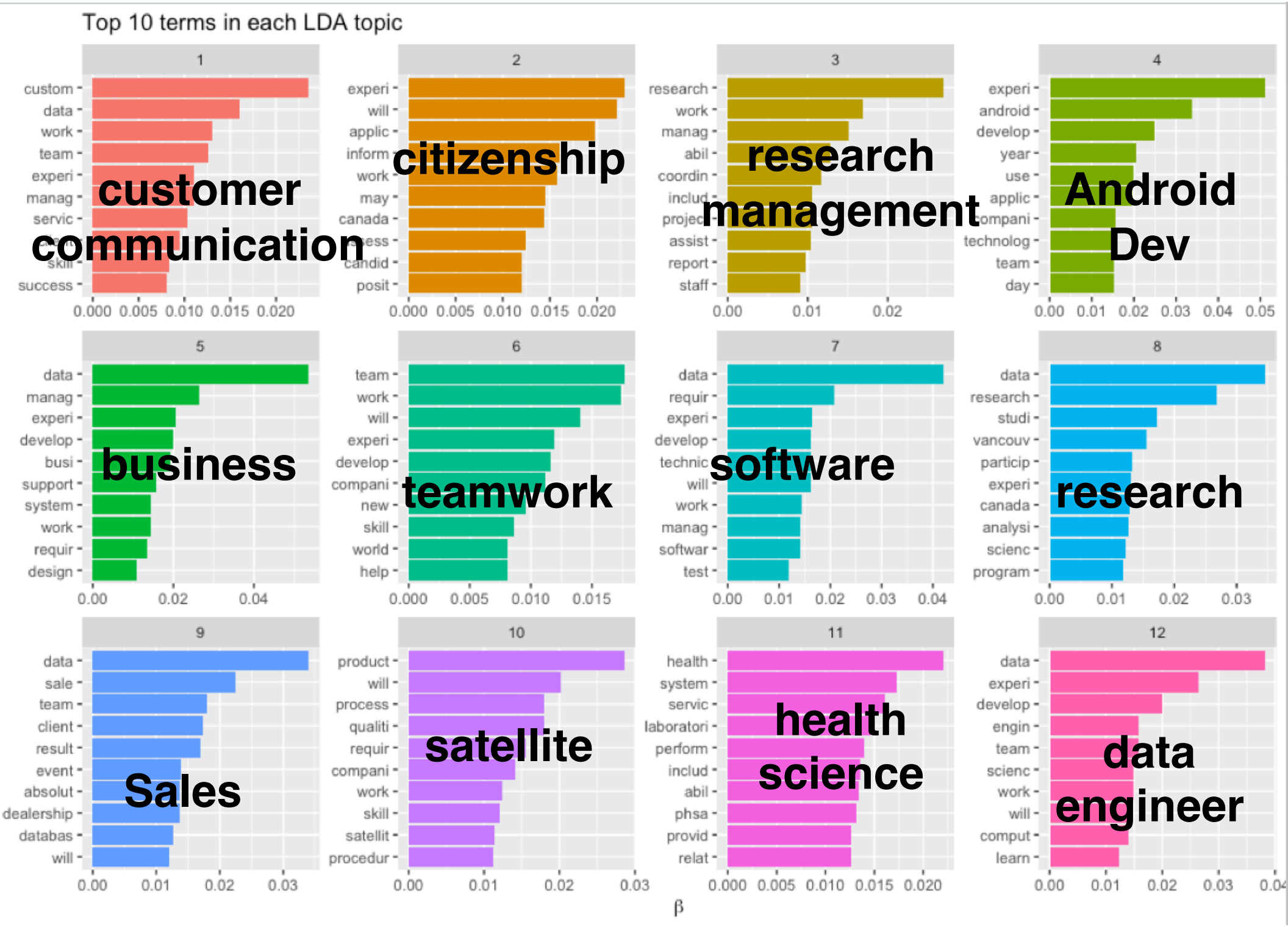
Make a matrix of documents (rows) vs words (columns)

Entries are counts of appearance

kmeans for text documents

```
lda_out = LDA(dtm,k=12)
```


Top 10 terms in each LDA topic
1
custom
data
work
team
experi
manag
servic
client
skill
success
0.000 0.005 0.010 0.015 0.020
2
experi
will
applic
inform
work
may
canada
assess
candid
posit
0.000 0.005 0.010 0.015 0.020
3
research
work
manag
abil
coordin
includ
project
assist
report
staff
0.00 0.01 0.02
4
experi
android
develop
year
use
applic
compani
technolog
team
day
0.00 0.01 0.02 0.03 0.04 0.05
5
data
manag
experi
develop
busi
support
system
work
requir
design
0.00 0.02 0.04
6
team
work
will
experi
develop
compani
new
skill
world
help
0.000 0.005 0.010 0.015
7
data
requir
experi
develop
technic
will
work
manag
softwar
test
0.00 0.01 0.02 0.03 0.04
8
data
research
studi
vancouv
particip
experi
canada
analysi
scienc
program
0.00 0.01 0.02 0.03
9
data
sale
team
client
result
event
absolut
dealership
databas
will
0.00 0.01 0.02 0.03
10
product
will
process
qualiti
requir
compani
work
skill
satellit
procedur
0.00 0.01 0.02 0.03
11
health
system
servic
laboratori
perform
includ
abil
phsa
provid
relat
0.000 0.005 0.010 0.015 0.020
12
data
experi
develop
engin
team
scienc
work
will
comput
learn
0.00 0.01 0.02 0.03 0.04
β


Top 10 terms in each LDA topic
1
custom
data
work
team
experi
manag
servic
skill
success
customer
communication
2
experi
will
applic
inform
work
may
canada
candid
posit
citizenship
3
research
work
manag
abil
coordin
includ
assist
report
staff
research
management
4
experi
android
develop
year
use
applic
compani
technolog
team
day
Android
Dev
5
data
manag
experi
develop
busi
support
system
work
requir
design
business
6
team
work
will
experi
develop
compani
new
skill
world
help
teamwork
7
data
requir
experi
develop
technic
will
work
manag
softwar
test
software
8
data
research
studi
vancouv
particip
experi
canada
analysi
scienc
program
research
9
data
sale
team
client
result
event
absolut
dealership
databas
will
Sales
10
product
will
process
qualiti
requir
compani
work
skill
satellit
procedur
satellite
11
health
system
servic
laboratori
perform
includ
abil
phsa
provid
relat
health
science
12
data
experi
develop
engin
team
scienc
work
will
comput
learn
data
engineer
β