Stat 240 Week 8
Dealing with time
more moving averages

Week 8

Dr. Dave Campbell

Table "National Parks and National Park Reserves" from https://en.wikipedia.org/wiki/List_of_National_Parks_of_Canada

NationalParks = c(" Aulavik\t\tNorthwest Territories\t12,200 km2 (4,710 sq mi)\t1992\n Auyuittuq\tPangnirtung Fiord S 2 2001-07-15.jpg\tNunavut\t19,089 km2 (7,370 sq mi) \t2001\n Banff\tMoraine Lake 17092005.jpg\tAlberta\t6,641 km2 (2,564 sq mi) \t1885\n Bruce Peninsula\tCyprusLake - Bruce Peninsula.jpg\tOntario\t154 km2 (59 sq mi) \t1987\n Cape Breton Highlands\tNS CapeBretonHighlands1 tango7174.jpg\tNova Scotia\t949 km2 (366 sq mi)\t1936\n Elk Island\tBison Elk Island.jpg\tAlberta\t194 km2 (75 sq mi)\t1913\n Forillon\tForillon National Park of Canada 1.jpg\tQuebec\t244 km2 (94 sq mi)\t1970\n Fundy\tFundy NP New Brunswick 1.jpg\tNew Brunswick\t206 km2 (80 sq mi)\t1948\n Georgian Bay Islands\tBeausoleilIslandCedarSprings2004.jpg\tOntario\t14 km2 (5 sq mi)\t1929\n Glacier\tGlacier np canada.JPG\tBritish Columbia\t1,349 km2 (521 sq mi)\t1886\n Grasslands\tSaskatchewan - Grasslands National Park 02.JPG\tSaskatchewan\t907 km2 (350 sq mi) \t1981\n Gros Morne\tNLW GrosMorne4 tango7174.jpg\tNewfoundland and Labrador\t1,805 km2 (697 sq mi)\t1973\n Gulf Islands\n (Reserve)\tGulfislfromair.jpg\tBritish Columbia\t36 km2 (14 sq mi)\t2003\n Gwaii Haanas[A]\n (Reserve)\tHaida Heritage Centre.jpg\tBritish Columbia\t1,495 km2 (577 sq mi)\t1988\n Ivvavik[B]\tCanada--yukon--ivvavik-np--spe 3021.jpg\tYukon\t10,168 km2 (3,926 sq mi)\t1984\n Jasper\tFryatt Valley top.jpg\tAlberta\t10,878 km2 (4,200 sq mi)\t1907\n Kejimkujik\tKejimkujik NP Nova Scotia 3.jpg\tNova Scotia\t404 km2 (156 sq mi)\t1968\n Kluane[C]\n (two units: a Park and a Reserve)\tDonjek Valley.jpg\tYukon\t22,013 km2 (8,499 sq mi)\t1976 (Reserve)\n 1993 (Park)\n Kootenay\tKootenay National Park.jpg\tBritish Columbia\t1,406 km2 (543 sq mi)\t1920\n Kouchibouguac\tKouchibouguac.JPG\tNew Brunswick\t239 km2 (92 sq mi)\t1969\n La Mauricie\tIle aux pins.jpg\tQuebec\t536 km2 (207 sq mi)\t1970\n Mingan Archipelago\n (Reserve)\tMonolithes de L'Archipel de Mingan.jpg\tQuebec\t151 km2 (58 sq mi)\t1984\n Mount Revelstoke\tRevelstoke from Mount Revelstoke.jpg\tBritish Columbia\t260 km2 (100 sq mi) \t1914\n Naats'ihch'oh[4]\n (Reserve)\tHoward's Pass Yukon Territory 1.jpg\tNorthwest Territories\t4,850 km2 (1,873 sq mi) \t2014\n Nahanni\n (Reserve) \tNahanni - VirginiaFalls.jpg\tNorthwest Territories\t30,000 km2 (11,583 sq mi)\t1976\n Pacific Rim\n (Reserve)\tLongbeach prnp.jpg\tBritish Columbia\t511 km2 (197 sq mi) \t1970\n Point Pelee\tPoint Pelee looking south.jpg\tOntario\t15 km2 (6 sq mi)\t1918\n Prince Albert\tPrince Albert National Park.jpg\tSaskatchewan\t3,874 km2 (1,496 sq mi) \t1927\n Prince Edward Island\tPeicoast.jpg\tPrince Edward Island\t22 km2 (8 sq mi)\t1937\n PukaskwaPukaskwaPark23.jpg\tOntario\t1,878 km2 (725 sq mi)\t1978\n Qausuittuq\tPeary caribou - looking west towards Evan's Bay.jpg\tNunavut\t11,000 km2 (4,247 sq mi) \t2015\n Quttinirpaaq[E]\tTanquary Fiord 16 1997-08-05.jpg\tNunavut\t37,775 km2 (14,585 sq mi) \t2001\n Riding Mountain[F]\tBison herd - Lake Audy - Riding Mountain National Park.JPG\tManitoba\t2,973 km2 (1,148 sq mi)\t1933\n Rouge\tLittle Rouge River Lookout.jpg\tOntario\t36 km2 (14 sq mi)\t2015\n Sable Island\n (Reserve)\tSableHorses.jpg\tNova Scotia\t34 km2 (13 sq mi)\t2013\n Sirmilik\tSirmilik Glacier 2 1997-08-06.jpg\tNunavut\t22,200 km2 (8,571 sq mi) \t2001\n Terra Nova\tNLC TerraNova3 tango7174.jpg\tNewfoundland and Labrador\t400 km2 (154 sq mi)\t1957\n Thousand Islands\tThousand Islands 2.JPG\tOntario\t24 km2 (9 sq mi)\t1904\n Torngat Mountains\tNachvak Fjord Labrador 2008.JPG\tNewfoundland and Labrador\t9,700 km2 (3,745 sq mi)\t2008\n Tuktut Nogait\tHornaday River.jpg\tNorthwest Territories\t16,340 km2 (6,309 sq mi)\t1996\n Ukkusiksalik\tEisbär 1996-07-23.jpg\tNunavut\t20,885 km2 (8,064 sq mi)\t2003\n Vuntut\tVontut National Park.jpg\tYukon\t4,345 km2 (1,678 sq mi) \t1995\n Wapusk\tBärenmutter & Junges 3 2004-11-17.jpg\tManitoba\t11,475 km2 (4,431 sq mi) \t1996\n Waterton Lakes[G]\tUpper Waterton Lake.JPG\tAlberta\t505 km2 (195 sq mi) \t1895\n Wood Buffalo\tWood-Buffalo-NP Gros Beak Lake 2 98-07-02.jpg\tAlberta\n Northwest Territories\t44,807 km2 (17,300 sq mi) \t1922\n Yoho\tYohoNP-Takakkaw IMG 1372-800x533byBMK.jpg\tBritish Columbia\t1,313 km2 (507 sq mi)\t1886")

Table “National Parks and National Park Reserves” from https://en.wikipedia.org/wiki/List_of_National_Parks_of_Canada

We split data into rows and were working on splitting it into columns.

We separated & cleaned size in km2 and sqmi

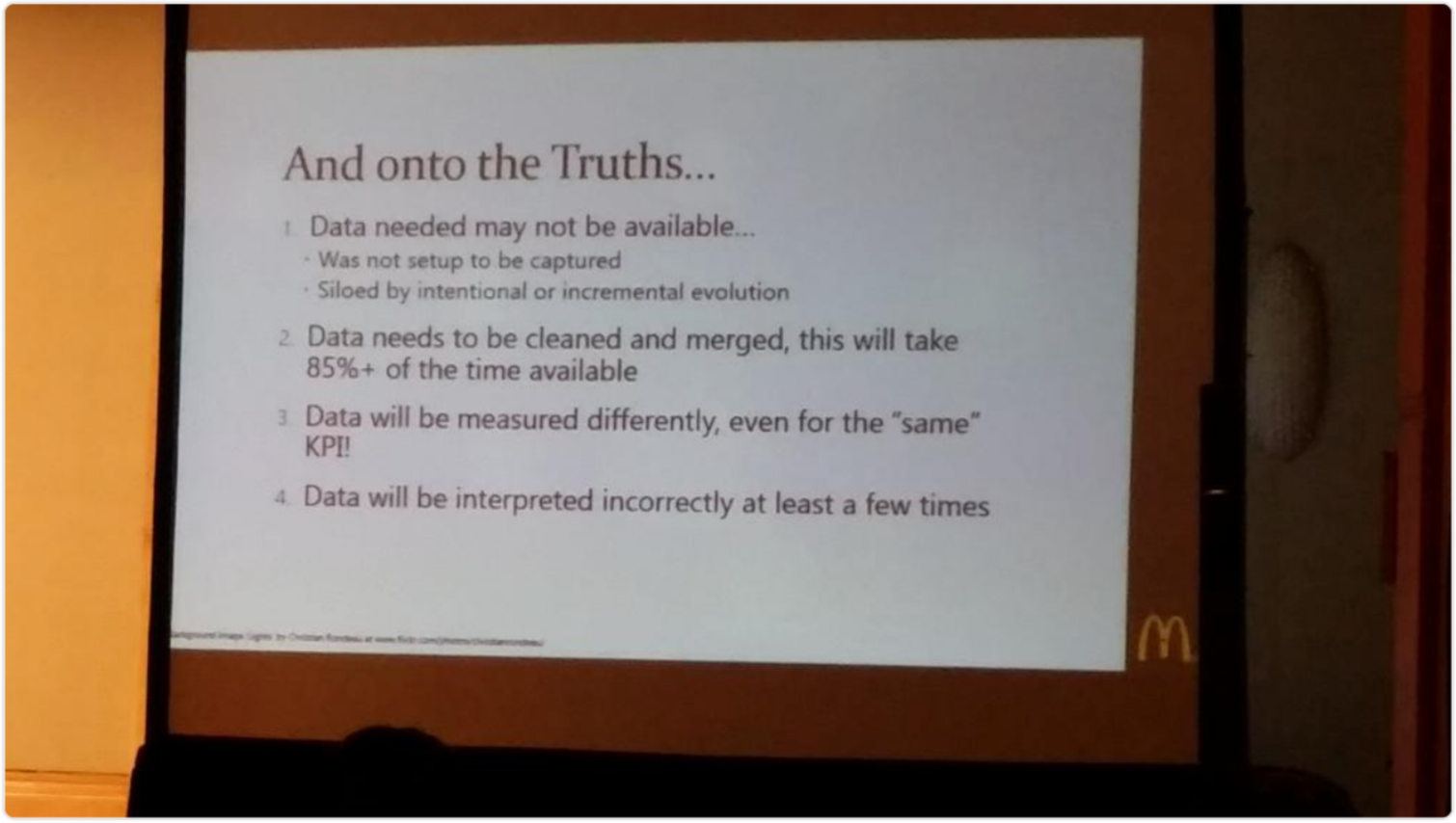
```
splitup = unlist(strsplit(NationalParks,"\n|\t")) #Make a
data.frame


N = length(splitup)

Name        = splitup[seq(1,N,by=5)]

Photo       = splitup[seq(2,N,by=5)]

Location    = splitup[seq(3,N,by=5)]

Area        = splitup[seq(4,N,by=5)]

Established  = splitup[seq(5,N,by=5)]

(Parks = cbind(Name,Location,Area,Established))
```

https://twitter.com/CIPSToronto/status/831610306111614977

CIPSToronto
@CIPSToronto

Follow

Diep of @McD_Canada tells #BigDataCA that 85% of time spent on data cleansing merging



RETWEET
1

1:05 PM - 14 Feb 2017 from Toronto, Ontario

Data Cleaning

Let’s look and fix the problem

```
grep("Reserve",splitup,ignore.case = TRUE)

grep("Reserve",splitup,ignore.case =
TRUE,value = TRUE)
```

Find where it breaks

```
strsplit(NationalParks,"(Reserve)")[[1:2]]
```

#OR

```
strsplit(NationalParks,"\\(Reserve\\)")[[1:2]]
```

Also could have found the problem by looking at the original data source:

Gulf Islands (Reserve)		British Columbia	36 km^2 (14 sq mi)	2003

Repair the break

#Find the error

substr(NationalParks,start = 950,stop = 1400)

#Try to fix the error

Repair the break

#Find the error

substr(NationalParks,start = 950,stop = 1400)

#Try to fix the error

substr(gsub("\n \\(Reserve\\)", "\\(Reserve\\)", NationalParks),start = 950,stop = 1400)

Repair the break

```
#Find the error

substr(NationalParks,start = 950,stop = 1400)

#Try to fix the error

substr(gsub("\n \\(Reserve\\)", "\\(Reserve\\)",
NationalParks),start = 950,stop = 1400)#nope

substr(gsub("\n[[:blank:]]\\(Reserve\\)", "****",
NationalParks,perl=TRUE),start = 950,stop = 1400)
```

Repair the break

```
#Find the error

substr(NationalParks,start = 950,stop = 1400)

#Try to fix the error

substr(gsub("\n \\(Reserve\\)", "\\(Reserve\\)",
NationalParks),start = 950,stop = 1400)#nope

substr(gsub("\n[[:blank:]]\\(Reserve\\)", "****",
NationalParks,perl=TRUE),start = 950,stop = 1400)#nope
```

Repair the break

```
#Find the error

substr(NationalParks,start = 950,stop = 1400)

#Try to fix the error

substr(gsub("\n \\(Reserve\\)", "\\(Reserve\\)",
NationalParks),start = 950,stop = 1400)#nope

substr(gsub("\n[[:blank:]]*\\(Reserve\\)", "****",
NationalParks,perl=TRUE),start = 950,stop = 1400)#YAY!
```

Repair the break

```
#Find the error

substr(NationalParks,start = 950,stop = 1400)

#Try to fix the error

substr(gsub("\n \\(Reserve\\)", "\\(Reserve\\)",
NationalParks),start = 950,stop = 1400)#nope

substr(gsub("\n[[:blank:]]*\\(Reserve\\)", "****",
NationalParks,perl=TRUE),start = 950,stop = 1400)#YAY!

substr(gsub("\n\\s*\\(Reserve\\)", "****",
NationalParks,perl=TRUE),start = 950,stop = 1400)#YAY!
```

```
Nat = gsub("\n[[:blank:]]*\\(Reserve\\)", "\\(Reserve\\)",
NationalParks,perl=TRUE)

splitup = unlist(strsplit(Nat,"\n|\t")) #Make a data.frame


N = length(splitup)

Name         = splitup[seq(1,N,by=5)]

Photo        = splitup[seq(2,N,by=5)]

Location     = splitup[seq(3,N,by=5)]

Area         = splitup[seq(4,N,by=5)]

Established  = splitup[seq(5,N,by=5)]

(Parks = cbind(Name,Location, Photo,Area,Established))
```

```
Warning message:
In cbind(Name, Location, Photo, Area, Established) :
  number of rows of result is not a multiple of vector length (arg 4)
```

^Not all pieces are the same length^

Look at the data:

Parks[17:20,]

Kluane[C] (two units: a Park and a Reserve)		Yukon	22,013 km^2 (8,499 sq mi)	1976 (Reserve) 1993 (Park)

How should we deal with:

2 dates and (two units: a Park and a Reserve) ?

There is a best answer! See

https://en.wikipedia.org/wiki/List_of_National_Parks_of_Canada

```
Nat = gsub("\n[[:blank:]]*\\(Reserve\\)", "\\(Reserve\\)",
NationalParks,perl=TRUE)

Nat2 = gsub("\\t1976\\s*\\(Reserve\\)", "\\(Reserve\\)", Nat,perl=TRUE)

Nat3 = gsub("\\n\\s*\\(two units: a Park and a Reserve\\)", "", Nat2,perl=TRUE)


splitup = unlist(strsplit(Nat3,"\n|\t")) #Make a data.frame


Name         = splitup[seq(1,N,by=5)]

Photo        = splitup[seq(2,N,by=5)]

Location     = splitup[seq(3,N,by=5)]

Area         = splitup[seq(4,N,by=5)]

Established  = splitup[seq(5,N,by=5)]

(Parks = cbind(Name,Location, Photo,Area,Established))
```

Table "National Parks and National Park Reserves" from https://en.wikipedia.org/wiki/List_of_National_Parks_of_Canada

```
NationalParks = c("  Aulavik\t\tNorthwest Territories\t12,200 km2 (4,710 sq mi)\t1992\n  Auyuittuq\tPangnirtung Fiord S 2 2001-07-15.jpg\tNunavut\t19,089 km2 (7,370 sq mi)\t2001\n Banff\tMoraine Lake 17092005.jpg\tAlberta\t6,641 km2 (2,564 sq mi)\t1885\n  Bruce Peninsula\tCyprusLake - Bruce Peninsula.jpg\tOntario\t154 km2 (59 sq mi)\t1987\n  Cape Breton Highlands\tNS CapeBretonHighlands1 tango7174.jpg\tNova Scotia\t949 km2 (366 sq mi)\t1936\n  Elk Island\tBison Elk Island.jpg\tAlberta\t194 km2 (75 sq mi)\t1913\n Forillon\tForillon National Park of Canada 1.jpg\tQuebec\t244 km2 (94 sq mi)\t1970\n  Fundy\tFundy NP New Brunswick 1.jpg\tNew Brunswick\t206 km2 (80 sq mi)\t1948\n Georgian Bay Islands\tBeausoleilIslandCedarSprings2004.jpg\tOntario\t14 km2 (5 sq mi)\t1929\n  Glacier\tGlacier np canada.JPG\tBritish Columbia\t1,349 km2 (521 sq mi)\t1886\n Grasslands\tSaskatchewan - Grasslands National Park 02.JPG\tSaskatchewan\t907 km2 (350 sq mi)\t1981\n Gros Morne\tNLW GrosMorne4 tango7174.jpg\tNewfoundland and Labrador\t1,805 km2 (697 sq mi)\t1973\n  Gulf Islands\n (Reserve)\tGulfIslfromair.jpg\tBritish Columbia\t36 km2 (14 sq mi) \t2003\n  Gwaii Haanas[A]\n  (Reserve)\tHaida Heritage Centre.jpg\tBritish Columbia\t1,495 km2 (577 sq mi) \t1988\n  Ivvavik[B]\tCanada--yukon--ivvavik-np--spe 3021.jpg\tYukon\t10,168 km2 (3,926 sq mi)\t1984\n Jasper\tFryatt Valley top.jpg\tAlberta\t10,878 km2 (4,200 sq mi) \t1907\n  Kejimkujik\tKejimkujik NP Nova Scotia 3.jpg\tNova Scotia\t404 km2 (156 sq mi)\t1968\n Kluane[C]\n  (two units: a Park and a Reserve)\tDonjek Valley.jpg\tYukon\t22,013 km2 (8,499 sq mi)\t1976 (Reserve)\n 1993 (Park)\n  Kootenay\tKootenay National Park.jpg\tBritish Columbia\t1,406 km2 (543 sq mi)\t1920\n Kouchibouguac\tKouchibouguac.JPG\tNew Brunswick\t239 km2 (92 sq mi)\t1969\n  La Mauricie\tIle aux pins.jpg\tQuebec\t536 km2 (207 sq mi)\t1970\n  Mingan Archipelago\n  (Reserve)\tMonolithes de L'Archipel de Mingan.jpg\tQuebec\t151 km2 (58 sq mi)\t1984\n  Mount Revelstoke\tRevelstoke from Mount Revelstoke.jpg\tBritish Columbia\t260 km2 (100 sq mi)\t1914\n Naats'ihch'oh[4]\n  (Reserve)\tHoward's Pass Yukon Territory 1.jpg\tNorthwest Territories\t4,850 km2 (1,873 sq mi) \t2014\n  Nahanni\n  (Reserve)\tNahanni - VirginiaFalls.jpg\tNorthwest Territories\t30,000 km2 (11,583 sq mi) \t1976\n  Pacific Rim\n  (Reserve)\tLongbeach prnp.jpg\tBritish Columbia\t511 km2 (197 sq mi)\t1970\n Point Pelee\tPoint Pelee looking south.jpg\tOntario\t15 km2 (6 sq mi)\t1918\n  Prince Albert\tPrince Albert National Park.jpg\tSaskatchewan\t3,874 km2 (1,496 sq mi)\t1927\n Prince Edward Island\tPeicoast.jpg\tPrince Edward Island\t22 km2 (8 sq mi)\t1937\n PukaskwaPukaskwaHorseshoeBay\tPukaskwaPark23.jpg\tOntario\t1,878 km2 (725 sq mi)\t1978\n  Qausuittuq\tPeary caribou - looking west towards Evan's Bay.jpg\tNunavut\t11,000 km2 (4,247 sq mi) \t2015\n  Quttinirpaaq[E]\tTanquary Fiord 16 1997-08-05.jpg\tNunavut\t37,775 km2 (14,585 sq mi)\t2001\n Riding Mountain[F]\tBison herd - Lake Audy - Riding Mountain National Park.JPG\tManitoba\t2,973 km2 (1,148 sq mi)\t1933\n Rouge\tLittle Rouge River Lookout.jpg\tOntario\t36 km2 (14 sq mi) \t2015\n  Sable Island\n  (Reserve) \tSableHorses.jpg\tNova Scotia\t34 km2 (13 sq mi)\t2013\n Sirmilik\tSirmilik Glacier 2 1997-08-06.jpg\tNunavut\t22,200 km2 (8,571 sq mi)\t2001\n  Terra Nova\tNLC TerraNova3 tango7174.jpg\tNewfoundland and Labrador\t400 km2 (154 sq mi) \t1957\n  Thousand Islands\tThousand Islands 2.JPG\tOntario\t24 km2 (9 sq mi)\t1904\n  Torngat Mountains\tNachvak Fjord Labrador 2008.JPG\tNewfoundland and Labrador\t9,700 km2 (3,745 sq mi)\t2008\n  Tuktut Nogait\tHornaday River.jpg\tNorthwest Territories\t16,340 km2 (6,309 sq mi)\t1996\n  Ukkusiksalik\tEisbär 1996-07-23.jpg\tNunavut\t20,885 km2 (8,064 sq mi)\t2003\n Vuntut\tVontut National Park.jpg\tYukon\t4,345 km2 (1,678 sq mi) \t1995\n  Wapusk\tBärenmutter & Junges 3 2004-11-17.jpg\tManitoba\t11,475 km2 (4,431 sq mi)\t1996\n Waterton Lakes[G]\tUpper Waterton Lake.JPG\tAlberta\t505 km2 (195 sq mi)\t1895\n  Wood Buffalo\tWood-Buffalo-NP Gros Beak Lake 2 98-07-02.jpg\tAlberta\n  Northwest Territories\t44,807 km2 (17,300 sq mi)\t1922\n Yoho\tYohoNP-Takakkaw IMG 1372-800x533byBMK.jpg\tBritish Columbia\t1,313 km2 (507 sq mi)\t1886")

Nat = gsub("\n[[:blank:]]*\\(Reserve\\)", "\\(Reserve\\)",
NationalParks,perl=TRUE)

substr(Nat,start=1500,stop=2000)

Nat2 = gsub("\n[[:blank:]]*\\(two units: a Park and a Reserve\\)",
    "",Nat,perl=TRUE)
substr(Nat2,start=1500,stop=2000)

Nat3 = gsub("1976 \\(Reserve\\)\n[[:blank:]]*1993 \\(Park\\)",
          "1976",Nat2,perl=TRUE)
substr(Nat3,start=1500,stop=2000)
substr(Nat3,start=4000,stop=5000)

#Change to Alberta since that's the majority location for Wood Buffalo Park

Nat4 = gsub("Alberta\n[[:blank:]]*Northwest Territories",
          "Alberta",Nat3,perl=TRUE)

splitup = unlist(strsplit(Nat4,"\n|\t")) #Make a data.frame


N = length(splitup)
Name          = splitup[seq(1,N,by=5)]
Photo         = splitup[seq(2,N,by=5)]
Location      = splitup[seq(3,N,by=5)]
Area          = splitup[seq(4,N,by=5)]
km2 = as.numeric(gsub("[[:punct:]]|km2.*\\)", "", Area) )
sqmi = as.numeric(gsub(".*\\(|[[:alpha:]]|[[:punct:]]", "", Area) )
Established  = splitup[seq(5,N,by=5)]

(Parks =
data.frame(cbind(Name,Location,Area,km2,sqmi,Established)))
Parks$km2 = as.numeric(as.character(Parks$km2))
Parks$sqmi = as.numeric(as.character(Parks$sqmi))
Parks$Established = as.numeric(as.character(Parks$Established))
```

Viewing the data
number of new parks per year

After data cleaning:

head(Parks)

table(Parks$Established)

plot(sort(unique(Parks$Established)), table(Parks$Established),xlab="year",ylab="numbe r",main= "Number of New National Parks and Reserves")

Viewing the data:
number of new parks per year
inserting zero

After data cleaning:

head(Parks)

range(Parks$E)

time = seq(from = min(Parks$E),to=max(Parks$E),by = 1)

ParksPerYear = table(time) #Make a vector in the right format

ParksPerYear = ParksPerYear*0 #Lazy set all values to 0

Viewing the data
number of new parks per year

names(table(Parks$E))

table(Parks$E)

ParksPerYear[names(table(Parks$E))]

ParksPerYear[names(table(Parks$E))] = table(Parks$E)

Viewing the data
number of new parks per year

```
ParksPerYear[names(table(Parks$E))] = table(Parks$E)

#Alternatively use a (slower) for loop

for(yr in names(table(Parks$E))){

        ParksPerYear[yr] = table(Parks$E)[yr]

 }
```

For loops are slower in R, nested for loops can be excruciatingly slow.

Slow code is often easier to understand and read.

ParksPerYear = as.array(ParksPerYear)# fix the table because we broke it by filling the table with new values

plot(time, ParksPerYear,xlab="year",ylab="number",main= "Number of New National Parks and Reserves")

Moving averages are back!
Recall density? Recall Moving Average in SQL?

in R: ksmooth(x,y,bandwidth = .5)

bandwidth is in units of x, choose for visual appeal

small bandwidth is too wiggly

big bandwidth is too smooth

plot(time,ParksPerYear,col=2,lwd=3)

lines(ksmooth(time,ParksPerYear,bandwidth = .5),col=2,lwd=3)

lines(ksmooth(time,ParksPerYear,bandwidth = 5),col=3,lwd=3)

lines(ksmooth(time,ParksPerYear,bandwidth = 50),col=4,lwd=3)

legend("topleft",c(".5","5","50"),col=2:4,lwd=3)

Moving averages are back!
Recall density? Recall Moving Average in SQL?

in R: ksmooth(x,y,bandwidth = .5)

bandwidth is in units of x, choose for visual appeal

small bandwidth is too wiggly

big bandwidth is too smooth

plot(Parks$E,Parks$km2,col=2,lwd=3)

lines(ksmooth(Parks$E,Parks$km2,bandwidth = .5),col=2,lwd=3)

lines(ksmooth(Parks$E,Parks$km2,bandwidth = 5),col=3,lwd=3)

lines(ksmooth(Parks$E,Parks$km2,bandwidth = 25),col=4,lwd=3)

legend("topleft",c(".5","5","25"),col=2:4,lwd=3)

Viewing the data number of parks per year

After data cleaning:

head(Parks)

table(Parks$Established)

cumsum(table(Parks$Established))

plot(sort(unique(Parks$Established)), cumsum(table(Parks$Established)),xlab="year",ylab ="cumulative number",main= "Number of National Parks and Reserves")

Dealing with Dates

```
load("Speed.Rdata")

head(speed)

speed[,"Date"]
```

Time difference

```
difftime(speed[1,"Date"],speed[2,"Date"])

difftime(speed[1,"Date"],speed[2,"Date"],units="hours")
```

Portable Operating System Interface

```
trunc(speed[,"Date"],"hours")#fails

trunc(as.POSIXct(speed[,"Date"]),"hours")

trunc(as.POSIXct(speed[,"Date"]),"days")

cut(as.POSIXct(speed[,"Date"]),"years")
```

Extracting time

strftime(speed[,"Date"], format="%H:%M:%S")

```
#Extract time of day:


difftime( as.POSIXct(speed[,"Date"]),

        trunc(as.POSIXct(speed[,"Date"]),"days"))


plot(difftime( as.POSIXct(speed[,"Date"]),
trunc(as.POSIXct(speed[,"Date"]),"days"))  ,speed$Downl
oad,xlab="time of day",ylab="Download Speed")
```