DATA SCIENCE WITH R

DATA SCIENCE WITH R

Class 4 – Data Manipulation in R

Topic 3

★ Working with Date Objects ★

INDEX

Manipulating Date objects

Manipulating date objects

- Dates are treated as a special data type in most programming languages
- R also treats dates as a separate data type
- Doing so, one can easily work with dates
- Usual date operations include:
  - Finding time interval between two points in data: age
  - Extracting months and week days

Manipulating date objects

- Character to date conversion-Date Class
- Extracting months and weekdays
- Using difftime()
- Manipulating data involving dates
- POSIXct and POSIXlt Classes
- Working with lubridate()

Manipulating date objects

- Using Date class to convert a character into a Date

```
> fd<-read.csv("Fd.csv")
> str(fd)
'data.frame': 30443 obs. of  25 variables:
 $ FlightDate        : Factor w/ 74 levels "01-Feb-14","01-Jan-14",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ UniqueCarrier     : Factor w/ 13 levels "AA","AS","B6",..: 4 4 4 4 11 11 11 11 11 4 ...
 $ AirlineID         : int  19790 19790 19790 19790 20355 20355 20355 20355 20355 19790 ...
 $ Carrier           : Factor w/ 13 levels "AA","AS","B6",..: 4 4 4 4 11 11 11 11 11 4 ...
 $ TailNum           : Factor w/ 2816 levels "","D942DN","N0EGMQ",..: 2641 2512 2490 2581 1657 34 60 17
98 1443 473 ...
 $ FlightNum         : int  335 1095 2422 1607 657 894 1843 2041 413 2030 ...
 $ OriginAirportID   : int  11057 11057 11057 11057 11057 11057 11057 11057 11057 13232 ...
 $ OriginAirportSeqID: int  1105703 1105703 1105703 1105703 1105703 1105703 1105703 1105703 1105703 132
```

Manipulating date objects

- Using Date class to convert a character into a Date

```
> fd$FlightDate<-as.Date(fd$FlightDate,"%d-%b-%y")
> str(fd)
'data.frame': 30443 obs. of  25 variables:
 $ FlightDate        : Date, format: "2014-01-01" "2014-01-01" "2014-01-01" ...
 $ UniqueCarrier     : Factor w/ 13 levels "AA","AS","B6",..: 4 4 4 4 11 11 11 11 11 4 ...
 $ AirlineID         : int  19790 19790 19790 19790 20355 20355 20355 20355 20355 19790 ...
 $ Carrier           : Factor w/ 13 levels "AA","AS","B6",..: 4 4 4 4 11 11 11 11 11 4 ...
 $ TailNum           : Factor w/ 2816 levels "","D942DN","N0EGMQ",..: 2641 2512 2490 2581 1657 34 60 17
98 1443 473 ...
 $ FlightNum         : int  335 1095 2422 1607 657 894 1843 2041 413 2030 ...
 $ OriginAirportID   : int  11057 11057 11057 11057 11057 11057 11057 11057 11057 13232 ...
 $ OriginAirportSeqID: int  1105703 1105703 1105703 1105703 1105703 1105703 1105703 1105703 1105703 132
```

Manipulating date objects

- Using Date class to convert a character into a Date

Code	Value
%d	Day of month (decimal number)
%m	Month (decimal number)
%b	Month (abbreviated)
%B	Month (full name)
%y	Year (2 digits)
%Y	Year (4 digits)

- 25/Aug/04: “%d/%b/%y
- 25-August-2004:%d-%B-%Y

Manipulating date objects

- Extracting months and weekdays from data:

```
> head(months(fd$FlightDate))
[1] "January" "January" "January" "January" "January" "January"
> unique(months(fd$FlightDate))
[1] "January"  "February" "March"
> head(weekdays(fd$FlightDate))
[1] "Wednesday" "Wednesday" "Wednesday" "Wednesday" "Wednesday" "Wednesday"
> unique(weekdays(fd$FlightDate))
[1] "Wednesday" "Thursday"  "Friday"    "Saturday"  "Sunday"    "Monday"    "Tuesday"
```

Manipulating date objects

- Computing time intervals and using difftime()

```
> fd$FlightDate[60]-fd$FlightDate[900]
Time difference of -3 days
> difftime(fd$FlightDate[3000],fd$FlightDate[90],units = "weeks")
Time difference of 1.571429 weeks
> difftime(fd$FlightDate[3000],fd$FlightDate[90],units = "days")
Time difference of 11 days
> difftime(fd$FlightDate[3000],fd$FlightDate[90],units = "hours")
Time difference of 264 hours
```

Manipulating date objects

- Manipulating data involving dates
- Sub-setting data: All rows when the day is Sunday

```
> library(dplyr)
> #Subset the data for day=Sunday
> dim(fd)
[1] 30443    25
> fd_s<-fd%>%filter(weekdays(FlightDate)=="Sunday")
> dim(fd_s)
[1] 4015   25
```

Manipulating date objects

- Manipulating data involving dates
- Find the number of flights on Sundays for destination Atlanta

```
> #Find the number of flights on Sundays for destination Atlanta
> fd%>%filter(weekdays(FlightDate)=="Sunday",DestCityName=="Atlanta, GA")%>%nrow()
[1] 683
```

Manipulating date objects

- Manipulating data involving dates
- Find the number of flights on Sundays for all cities

```
> #Find the number of flights on Sundays for all cities
> fd%>%filter(weekdays(FlightDate)=="Sunday")%>%group_by(DestCityName)%>%summarize(n())
Source: local data frame [10 x 2]

      DestCityName n()
1      Atlanta, GA 683
2    Charlotte, NC 342
3      Chicago, IL 193
4       Denver, CO 448
5      Houston, TX 155
6    Las Vegas, NV 507
7  Los Angeles, CA 603
8     New York, NY 349
9      Phoenix, AZ 466
10  Washington, DC 269
```

Manipulating date objects

- Whenever data has time information along with date, R uses POSIXct and POSIXlt classes to deal with dates

```
> date1<-Sys.time()
> date1
[1] "2015-03-02 17:35:47 IST"
> class(date1)
[1] "POSIXct" "POSIXt"
> weekdays(date1)
[1] "Monday"
> months(date1)
[1] "March"
```

Manipulating date objects

- Whenever data has time information along with date, we use POSIXct and POSIXlt classes to deal with dates

```
> date2<-as.POSIXlt(date1)
> date2
[1] "2015-03-02 17:35:47 IST"
> str(date2)
 POSIXlt[1:1], format: "2015-03-02 17:35:47"
```

```
> date2$wday
[1] 1
> date2$zone
[1] "IST"
> date2$hour
[1] 17
> date2$wday
[1] 1
> date2$zone
[1] "IST"
> date2$hour
[1] 17
```

Manipulating date objects

- lubridate() is a package that is a wrapper for POSIXct class
- It has a very simple syntax

```
> library(lubridate)
> fd$FlightDate<-dmy(fd$FlightDate)
> str(fd)
'data.frame': 30443 obs. of  25 variables:
 $ FlightDate        : POSIXct, format: "2014-01-01" "2014-01-01" "2014-01-01" ...
 $ UniqueCarrier     : Factor w/ 13 levels "AA","AS","B6",..: 4 4 4 4 11 11 11 11 11 4 ...
 $ AirlineID         : int  19790 19790 19790 19790 20355 20355 20355 20355 20355 19790 ...
 $ Carrier           : Factor w/ 13 levels "AA","AS","B6",..: 4 4 4 4 11 11 11 11 11 4 ...
 $ TailNum           : Factor w/ 2816 levels "","D942DN","N0EGMQ",..: 2641 2512 2490 2581 1657 34 60 17
98 1443 473 ...
 $ FlightNum         : int  335 1095 2422 1607 657 894 1843 2041 413 2030 ...
 $ OriginAirportID   : int  11057 11057 11057 11057 11057 11057 11057 11057 11057 13232 ...
 $ OriginAirportSeqID: int  1105703 1105703 1105703 1105703 1105703 1105703 1105703 1105703 1105703 132
3202 ...
 $ OriginCityMarketID: int  31057 31057 31057 31057 31057 31057 31057 31057 31057 30977 ...
 $ Origin            : Factor w/ 10 levels "ATL","CLT","DCA",..: 2 2 2 2 2 2 2 2 2 9 ...
 $ OriginCityName    : Factor w/ 10 levels "Atlanta, GA",..: 2 2 2 2 2 2 2 2 2 3 ...
 $ OriginState       : Factor w/ 10 levels "AZ","CA","CO",..: 6 6 6 6 6 6 6 6 6 5 ...
```

Manipulating date objects

- lubridate() is a package that is a wrapper for POSIXct class
- It has a very simple syntax.

Function	Date
dmy()	26/11/2008
ymd()	2008/11/26
mdy()	11/26/2008
dmy_hm()	26/11/2008 20:15
dmy_hms()	26/11/2008 20:15:30

DATA SCIENCE WITH R

RECAP

- Character to date conversion-Date Class
- Extracting months and weekdays
- Using difftime()
- Manipulating data involving dates
- POSIXct and POSIXlt Classes
- Working with lubridate()