exercise-2

Matthew McAvoy

August 25, 2016

1. Load XML file

Requires XML to read in xml files.

```
require("XML")
```

```
## Loading required package: XML
```

Sets the working directory and loads the xml file into local memory.

```
setwd("C:/Users/homur/OneDrive/New College/EDA/Week 1")
courseData <- xmlParse("reed-courses.xml")
```

xmlToList to convert the xml file into a more usable form, topxml access's the top node, then xmlSApply to extract xml values.

```
courseXml <- xmlToList(courseData)
topxml <- xmlRoot(courseData)
topxml <- xmlSApply(topxml, function(x) xmlSApply(x,xmlValue))
```

If we were to inspect it now, it would look like an xml file; meaning not very pretty. The next step converts the xml file into an R data frame for further operations.

```
xml_df <- data.frame(t(topxml))
```

```
## Warning in data.row.names(row.names, rowsi, i): some row.names duplicated:
## 2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,3
## --> row.names NOT used
```

```
head(xml_df)
```

```
##   reg_num subj crse sect                              title units
## 1   10577 ANTH  211  F01      Introduction to Anthropology   1.0
## 2   20573 ANTH  344  S01                    Sex and Gender   1.0
## 3   10624 BIOL  431  F01       Field Biology of Amphibians   0.5
## 4   10626 BIOL  431  F03            Bacterial Pathogenesis   0.5
## 5   20626 BIOL  431  S04                Seminar in Biology   0.5
## 6   10543 CHEM  101    F MolecularStructure and Properties   1.0
##    instructor  days         time
## 1   Brightman   M-W 03:10PM04:30
## 2      Makley  T-Th 10:30AM11:50
## 3      Kaplan     T 06:10PM08:00
## 4
## 5   Yezerinac    Th 06:10PM08:00
```

```
## 6 Geselbracht M-W-F 11:00AM11:50
##                                          place
## 1                                      ELIOT414
## 2                                      VOLLUM120
## 3                                     PHYSIC240A
## 4 Mellies         RESCHEDULED TO OTHER SEMESTER
## 5                                       BIOL200A
## 6                                     VOLLUMVLH
```

Now upon inspection, it looks much better as a data frame. We then use the str() function to look at metadata of the xml file.

```
str(xml_df)
```

```
## 'data.frame':    703 obs. of  10 variables:
##  $ reg_num   : Factor w/ 699 levels "10072","10073",..: 260 606 302 303 651 228 229 230 231 430 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ subj      : Factor w/ 31 levels "ANTH","ART","BIOL",..: 1 1 3 3 3 4 4 4 4 4 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ crse      : Factor w/ 151 levels "100","101","102",..: 27 88 139 139 139 2 2 2 2 2 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ sect      : Factor w/ 129 levels "AE6","AE7","AFD",..: 18 62 18 20 65 17 18 19 20 25 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ title     : Factor w/ 394 levels "17th Cent French Drama",..: 194 330 127 54 327 258 258 258 258 1
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ units     : Factor w/ 3 levels "0.0","0.5","1.0": 3 3 2 2 2 3 1 1 1 1 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ instructor: Factor w/ 136 levels "","Ahmadi","Alonso",..: 16 74 61 1 136 43 43 43 43 43 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ days      : Factor w/ 14 levels "","F","M","M-T-W-F",..: 7 10 9 1 12 8 3 3 3 9 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ time      : Factor w/ 57 levels "","01:10PM02:00",..: 14 55 31 1 31 56 2 9 13 56 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
##  $ place     : Factor w/ 84 levels "ARTDRAW","ARTPAINT",..: 30 75 56 47 8 82 16 16 16 52 ...
##   ..- attr(*, "names")= chr  "course" "course" "course" "course" ...
```

From this, we can answer some of the questions.

2. There are 31 distinct subjects listed in the document.
3. There are 136 distinct insturctors listed.

Using the following summary, we see an empty field in instructor, which allows us to answer number 3.

3. empty row corresponds to professor names of NULL = 15.

```
summary(xml_df)
```

```
##     reg_num         subj          crse          sect
##  10436  :  2   PE     : 71   101    : 77   S      :160
##  10437  :  2   CHEM   : 64   102    : 52   F      :149
##  10747  :  2   BIOL   : 56   110    : 40   F01    : 27
```

```
##  10799  :  2   PHYS   : 49   201    : 28   F02    : 27
##  10072  :  1   MATH   : 45   211    : 22   S01    : 20
##  10073  :  1   HUM    : 42   100    : 16   S02    : 20
##  (Other):693   (Other):376   (Other):468   (Other):300
##                                  title     units          instructor
##  West Humanities: Greece  and Rome: 26   0.0:269   Casey      : 71
##  Intro Biology Lect and Lab       : 18   0.5: 36   Geselbracht: 21
##  MolecularStructure and Properties: 17   1.0:398   Bonfim     : 16
##  General Physics I                : 16             :  15
##  Introduction to Physics          : 12             Glasfeld   : 14
##  Chemical Reactivity              : 11             Hancock    : 13
##  (Other)                          :603             (Other)    :553
##       days               time           place
##  T-Th   :204   02:40PM04:00: 52   SPORTS   : 28
##  M-W-F  :152   11:00AM11:50: 50   PSYCH108 : 25
##  M-W    : 74   10:30AM11:50: 48   CHEM301  : 23
##  T      : 57   03:10PM04:30: 47   PHYSIC123: 22
##  W      : 55   01:10PM02:00: 43   VOLLUM120: 21
##  Th     : 51   01:10PM02:30: 41   VOLLUM134: 20
##  (Other):110   (Other)     :422   (Other)  :564
```