Project 20

Natalie McGuckin

Using RStudio, we start to learn how to extract data from the web.

Use the data from the Billboard Hot 100 for question 1. Please use the data from the week you were born. For instance, if I solve question 1, I would use the data located here: https://www.billboard.com/charts/hot-100/1976-10-13

On the Hot 100 chart, from the day of your birth:

1a. Extract the titles of the songs ranked #2 through #100.

```
library(XML)
library(RCurl)
```

```
## Loading required package: bitops
```

```
xpathSApply(htmlParse(getURL("https://www.billboard.com/charts/hot-100/2000-04-13")),
            "//*/div[@class='chart-list-item__title']", xmlValue)
```

```
##  [1] "\n\nSay My Name\n\n"
##  [2] "\n\nBreathe\n\n"
##  [3] "\n\nBye Bye Bye\n\n"
##  [4] "\n\nAmazed\n\n"
##  [5] "\n\nThong Song\n\n"
##  [6] "\n\nThere You Go\n\n"
##  [7] "\n\nI Try\n\n"
##  [8] "\n\nShow Me The Meaning Of Being Lonely\n\n"
##  [9] "\n\nGet It On Tonite\n\n"
## [10] "\n\nIt Feels So Good\n\n"
## [11] "\n\nSmooth\n\n"
## [12] "\n\nEverything You Want\n\n"
## [13] "\n\nThat's The Way It Is\n\n"
## [14] "\n\nI Wanna Know\n\n"
## [15] "\n\nNever Let You Go\n\n"
## [16] "\n\n I Knew I Loved You\n\n"
## [17] "\n\nYou Sang To Me\n\n"
## [18] "\n\nOnly God Knows Why\n\n"
## [19] "\n\nGoodbye Earl\n\n"
## [20] "\n\nTry Again\n\n"
## [21] "\n\nBring It All To Me\n\n"
## [22] "\n\nAll The Small Things\n\n"
## [23] "\n\nI Like It\n\n"
## [24] "\n \nI Need To Know\n\n"
## [25] "\n \nForgot About Dre\n\n"
## [26] "\n\nBack At One\n\n"
## [27] "\n\nHigher\n\n"
## [28] "\n\nOtherside\n \n"
## [29] "\n\nWhat A Girl Wants\n\n"
## [30] "\n\nHow Do You Like Me Now?!\n\n"
## [31] "\n \nParty Up (Up In Here)\n\n"
## [32] "\n\nThe Best Day\n\n"
## [33] "\n\nBe With You\n\n"
## [34] "\n\nHe Wasn't Man Enough\n\n"
```

```
## [35] "\n\nAnother Dumb Blonde\n\n"
## [36] "\n\nI Don't Wanna\n\n"
## [37] "\n\nHot Boyz\n\n"
## [38] "\n\nI Wish\n\n"
## [39] "\n\nLessons Learned\n\n"
## [40] "\n\nThen The Morning Comes\n\n"
## [41] "\n\nLove's The Only House\n\n"
## [42] "\n\nWhoa!\n\n"
## [43] "\n\nBeen There\n\n"
## [44] "\n\nFrom The Bottom Of My Broken Heart\n\n"
## [45] "\n\nBuy Me A Rose\n\n"
## [46] "\n\nCarlene\n \n"
## [47] "\n\nMy Best Friend\n\n"
## [48] "\n\nThe Way You Love Me\n\n"
## [49] "\n\nI Turn To You\n\n"
## [50] "\n\nUntitled (How Does It Feel)\n\n"
## [51] "\n\nThe Bad Touch\n\n"
## [52] "\n\nI Learned From The Best\n\n"
## [53] "\n\nShackles (Praise You)\n\n"
## [54] "\n\nShe's More\n\n"
## [55] "\n\nThat's What I'm Looking For\n\n"
## [56] "\n\nWobble Wobble\n\n"
## [57] "\n\nCrash And Burn\n\n"
## [58] "\n\nYou Owe Me\n\n"
## [59] "\n\nFalls Apart\n\n"
## [60] "\n\nFeelin' So Good\n\n"
## [61] "\n\nWhere You Are\n\n"
## [62] "\n\nI Don't Wanna Kiss You Goodnight\n\n"
## [63] "\n\nIt Was\n\n"
## [64] "\n\nBetter Off Alone\n\n"
## [65] "\n\nSwear It Again\n\n"
## [66] "\n\nCouldn't Last A Moment\n\n"
## [67] "\n\nWhat I Need To Do\n\n"
## [68] "\n\nHe Can't Love U\n\n"
## [69] "\n\nKryptonite\n\n"
## [70] "\n\nI Belong To You\n\n"
## [71] "\n\nBecause You Love Me\n\n"
## [72] "\n\nThe Chain Of Love\n\n"
## [73] "\n\nBack At One\n\n"
## [74] "\n\nU Don't Love Me\n\n"
## [75] "\n\nBlue (Da Ba Dee)\n\n"
## [76] "\n\nYes!\n\n"
## [77] "\n\nSmile\n\n"
## [78] "\n\nI Need A Hot Girl\n\n"
## [79] "\n\nPurest Of Pain (A Puro Dolor)\n\n"
## [80] "\n\nGraduation (Friends Forever)\n\n"
## [81] "\n\nThank God I Found You\n\n"
## [82] "\n\nReal Live Woman\n\n"
## [83] "\n\nWhistle While You Twurk\n\n"
## [84] "\n\nBest Of Me\n\n"
## [85] "\n\nMr. Too Damn Good\n\n"
## [86] "\n\nLove Is Blind\n\n"
## [87] "\n\nAmerican Pie\n\n"
## [88] "\n\nIf You Don't Wanna Love Me\n\n"
```

```
## [89] "\n\nUnconditional\n\n"
## [90] "\n\nRyde Or Die, Chick\n\n"
## [91] "\n\nAnything\n\n"
## [92] "\n\nStay Or Let It Go\n\n"
## [93] "\n\nMirror Mirror\n\n"
## [94] "\n\nNo Leaf Clover\n\n"
## [95] "\n\nLeft, Right, Left\n\n"
## [96] "\n\nGive Me You\n\n"
## [97] "\n\nCan't Stay\n\n"
## [98] "\n\nOne Night Stand\n\n"
## [99] "\n\nNo More Rain (In This Cloud)\n\n"
```

Explain solution:changed the code to pull the page for my birthday and scraped the top 99 songs.

1b. Extract the artists for those 99 songs.

```
xpathSApply(htmlParse(getURL("https://www.billboard.com/charts/hot-100/2000-04-13")),
            "//*/div[@class='chart-list-item__artist']", xmlValue)
```

```
##  [1] "\n\nDestiny's Child\n\n"
##  [2] "\n\nFaith Hill\n\n"
##  [3] "\n\n'N Sync\n\n"
##  [4] "\n\nLonestar\n\n"
##  [5] "\n\nSisqo\n\n"
##  [6] "\n\nP!nk\n\n"
##  [7] "\n\nMacy Gray\n\n"
##  [8] "\n\nBackstreet Boys\n\n"
##  [9] "\n\nMontell Jordan\n\n"
## [10] "\n\nSonique\n\n"
## [11] "\nSantana Featuring Rob Thomas\n"
## [12] "\n\nVertical Horizon\n\n"
## [13] "\n\nCeline Dion\n\n"
## [14] "\n\nJoe\n\n"
## [15] "\n\nThird Eye Blind\n\n"
## [16] "\n\nSavage Garden\n\n"
## [17] "\n\nMarc Anthony\n\n"
## [18] "\n\nKid Rock\n\n"
## [19] "\n\nDixie Chicks\n\n"
## [20] "\n\nAaliyah\n\n"
## [21] "\n\nBlaque\n\n"
## [22] "\n\nBlink-182\n\n"
## [23] "\n\nSammie\n\n"
## [24] "\n\nMarc Anthony\n\n"
## [25] "\nDr. Dre Featuring Eminem\n"
## [26] "\n\nBrian McKnight\n\n"
## [27] "\n\nCreed\n\n"
## [28] "\n\nRed Hot Chili Peppers\n\n"
## [29] "\n\nChristina Aguilera\n\n"
## [30] "\n\nToby Keith\n\n"
## [31] "\n\nDMX\n\n"
## [32] "\n\nGeorge Strait\n\n"
## [33] "\n\nEnrique Iglesias\n\n"
## [34] "\n\nToni Braxton\n\n"
## [35] "\n\nHoku\n\n"
## [36] "\n\nAaliyah\n\n"
## [37] "\nMissy \"Misdemeanor\" Elliott Featuring NAS, EVE & Q-Tip\n"
```

```
## [38] "\n\nCarl Thomas\n\n"
## [39] "\n\nTracy Lawrence\n\n"
## [40] "\n\nSmash Mouth\n\n"
## [41] "\n\nMartina McBride\n\n"
## [42] "\n\nBlack Rob\n\n"
## [43] "\nClint Black With Steve Wariner\n"
## [44] "\n\nBritney Spears\n\n"
## [45] "\nKenny Rogers With Alison Krauss & Billy Dean\n"
## [46] "\n\nPhil Vassar\n\n"
## [47] "\n\nTim McGraw\n\n"
## [48] "\n\nFaith Hill\n\n"
## [49] "\n\nChristina Aguilera\n\n"
## [50] "\n\nD'Angelo\n\n"
## [51] "\n\nBloodhound Gang\n\n"
## [52] "\n\nWhitney Houston\n\n"
## [53] "\n\nMary Mary\n\n"
## [54] "\n\nAndy Griggs\n\n"
## [55] "\n\nDa Brat\n\n"
## [56] "\n\n504 Boyz\n\n"
## [57] "\n\nSavage Garden\n\n"
## [58] "\n NAS Featuring Ginuwine\n"
## [59] "\n\nSugar Ray\n\n"
## [60] "\n\nJennifer Lopez Featuring Big Pun & Fat Joe\n\n"
## [61] "\nJessica Simpson Featuring Nick Lachey\n"
## [62] "\n\nLFO\n\n"
## [63] "\n\nChely Wright\n\n"
## [64] "\n\nAlice Deejay\n\n"
## [65] "\n\nWestlife\n\n"
## [66] "\n \nCollin Raye\n\n"
## [67] "\n\nKenny Chesney\n\n"
## [68] "\n\nJagged Edge\n\n"
## [69] "\n\n3 Doors Down\n\n"
## [70] "\n\nLenny Kravitz\n\n"
## [71] "\n\n Jo Dee Messina\n\n"
## [72] "\n\nClay Walker\n\n"
## [73] "\n\nMark Wills\n\n"
## [74] "\nKumbia Kings Featuring A.B. Quintanilla\n"
## [75] "\n\nEiffel 65\n\n"
## [76] "\n\nChad Brock\n\n"
## [77] "\n\nLonestar\n\n"
## [78] "\n\nHot Boys\n\n"
## [79] "\n\nSon By Four\n\n"
## [80] "\n \nVitamin C\n\n"
## [81] "\nMariah Carey Featuring Joe & 98 Degrees\n"
## [82] "\n\nTrisha Yearwood\n\n"
## [83] "\n\nYing Yang Twins\n \n"
## [84] "\nMya Featuring Jadakiss\n"
## [85] "\n\nGerald Levert\n\n"
## [86] "\nEve Featuring Faith Evans\n"
## [87] "\n\nMadonna\n\n "
## [88] "\n\nTamar Braxton\n\n"
## [89] "\n\nClay Davidson\n\n"
## [90] "\nThe Lox Featuring Timbaland And EVE\n"
## [91] "\n\nJAY-Z\n\n"
```

```
## [92] "\n\nBrian McKnight\n\n"
## [93] "\n\nM2M\n\n"
## [94] "\n \nMetallica\n\n"
## [95] "\n\nDrama\n\n"
## [96] "\n\nMary J. Blige\n\n"
## [97] "\n\nDave Hollister\n\n"
## [98] "\nJ-Shin Featuring LaTocha Scott\n"
## [99] "\n\nAngie Stone\n\n"
```

Explain solution: Changed class

1c. Extract the title of the number 1 song for that day.

```
xpathSApply(htmlParse(getURL("https://www.billboard.com/charts/hot-100/2000-04-13")),
            "//*/div[@class='chart-number-one__title']", xmlValue)
```

```
## [1] "Maria Maria"
```

Explain solution: changed class, used same code from 1a

1d. Extract the artist for the number 1 song for that day.

```
xpathSApply(htmlParse(getURL("https://www.billboard.com/charts/hot-100/2000-04-13")),
            "//*/div[@class='chart-number-one__artist']", xmlValue)
```

```
## [1] "\nSantana Featuring The Product G&B\n"
```

Explain solution: changed class

2a. Extract the city where the National Park property for Catoctin Mountain is located. This data is found at: https://www.nps.gov/cato/index.htm or in the file: /depot/statclass/data/parks/cato.htm

```
xpathSApply(htmlParse(getURL("https://www.nps.gov/cato/index.htm")), "//*/span[@itemprop='addressLocali
```

```
## [1] "Thurmont"
```

Explain solution:cato shortening the website uses for cactotin moutntain

2b. Extract the state where Catoctin Mountain is located.

```
xpathSApply(htmlParse(getURL("https://www.nps.gov/cato/index.htm")), "//*/span[@itemprop='addressRegion
```

```
## [1] "MD"
```

Explain solution: Changed class

2c. Extract the zip code where Catoctin Mountain is located.

```
xpathSApply(htmlParse(getURL("https://www.nps.gov/cato/index.htm")), "//*/span[@itemprop='postalCode']"
```

```
## [1] "21788     "
```

Explain solution:changed class

3a. Identify three potential websites that you are interested to try to scrape yourself, during the upcoming seminars. Look for websites with data that is (relatively) easy to scrape, for instance: Systematic URL's that are easy to understand; (relative) consistency in how the data is stored; and make sure that the data is embedded in the page, rather than in csv files that are already prepared for download. (We want to actually scrape some data.)

1. https://www.color-hex.com/popular-colors.php
2. http://norfolkdailynews.com/lite_rock/programs/dave_williams/most-popular-ice-cream-flavors/article_8e46a544-6a88-11e8-a65b-2bb11e94b092.html
3. https://buccetos.com/the-top-10-most-popular-pizza-toppings/

3b. For each of the three websites that you identified, give a very brief description of the kind of data that you want to scrape.

1. I want to scrape most popular colors

2. I want to scrape most popular ice cream flavors
3. I want to scrape most popular pizza toppings