Web scraping with R
rvest & RSelenium

Barret Schloerke
Statistics PhD Candidate
Purdue University

Web Scraping

- Main Goals
  - Parse information from a given website using R
  - Retrieve information from a non-“html & css only” website using R

Software - R

- magrittr - forward-piping
  - github.com/smbache/magrittr
- rvest - web scraper
  - github.com/hadley/rvest
- RSelenium - headless browser
  - github.com/ropensci/RSelenium

rvest

- “It is designed to work with magrittr to make it easy to express common web scraping tasks, inspired by libraries like beautiful soup.”
- Installation
  - `install.packages("rvest")`
- Requirements
  - url
  - css selectors - selectorgadget.com

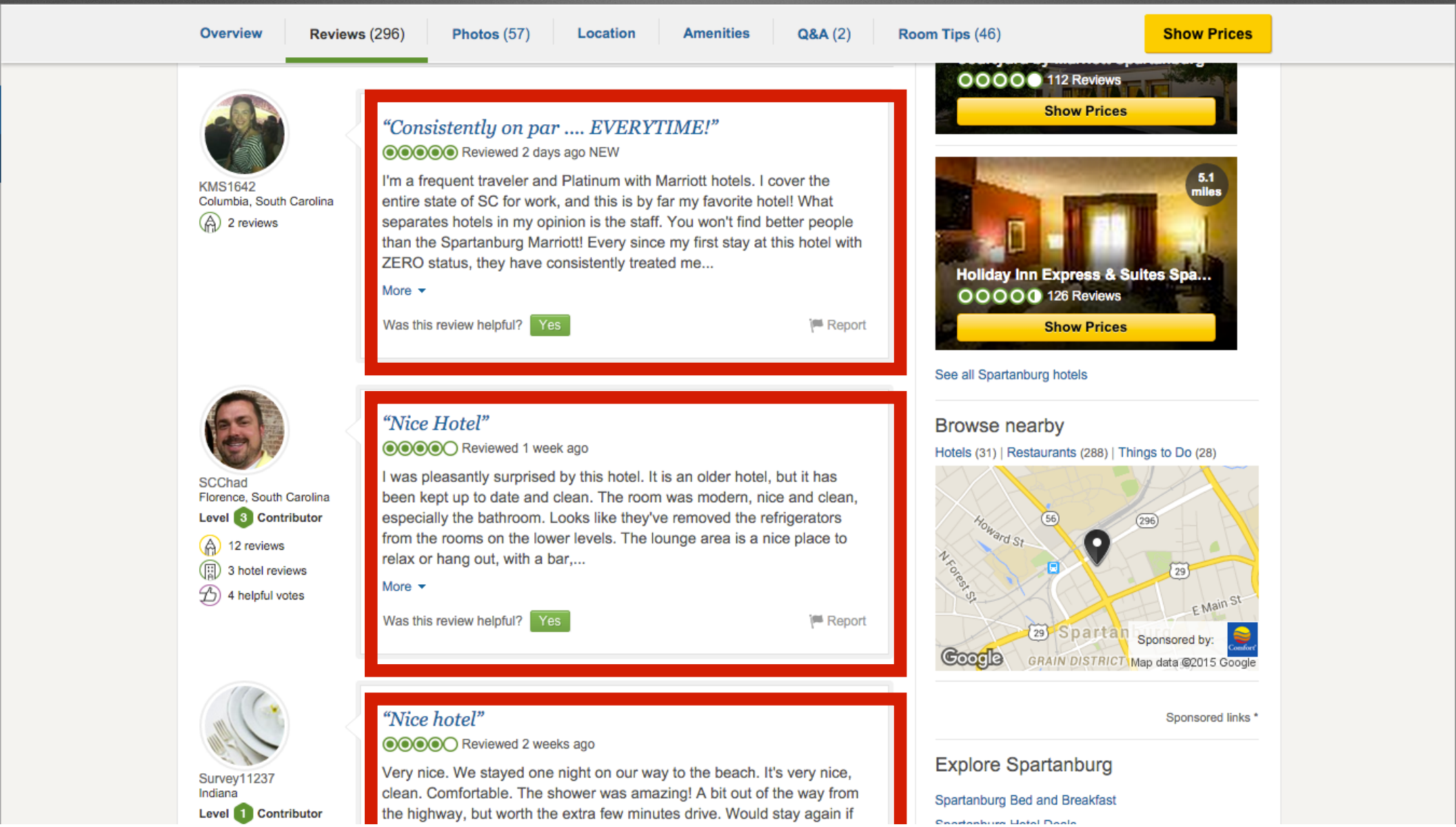
https://github.com/hadley/rvest

rvest - Common Functions

- `read_html()` - parse a url into html objects
- `html_nodes()` - select html objects with css selector
- `html_attr()` - retrieve html attributes from html objects
- `html_text()` - retrieve html object text

rvest - Example

- http://www.tripadvisor.com/Hotel_Review-g54448-d288616-Reviews-Spartanburg_Marriott-Spartanburg_South_Carolina.html



rvest - Example

- For each review
  - ID
  - Quote
  - Rating
  - Review

rvest - Example

```
library(rvest)

url <- "http://www.tripadvisor.com/Hotel_Review-
g54448-d288616-Reviews-Spartanburg_Marriott-
Spartanburg_South_Carolina.html"

# Grab each review
reviews <- url %>%
  read_html() %>%
  html_nodes("#REVIEWS .innerBubble")
```

rvest - Example

```
id <- reviews %>%
  html_node(".quote a") %>%
  html_attr("id")

quote <- reviews %>%
  html_node(".quote span") %>%
  html_text()

rating <- reviews %>%
  html_node(".ui_bubble_rating") %>%
  html_attr("class") %>%
  str_replace("ui_bubble_rating", "") %>%
  str_trim() %>%
  str_replace("bubble_", "") %>%
  as.numeric() %>%
  divide_by(10)

review <- reviews %>%
  html_node(".entry .partial_entry") %>%
  html_text()
```

rvest - Example

```
hotel <- tibble::tibble(id, quote, rating, review)

tibble::glimpse(hotel)
```

```
###
Observations: 10
Variables: 4
$ id     <chr> "rn468094751", "rn463240832", "rn462587870", "rn461414709", ...
$ quote  <chr> "Train sounds all night, employees entering room throughout ...
$ rating <dbl> 2, 3, 4, 5, 5, 4, 5, 5, 5, 3
$ review <chr> "The room was nice and clean. The staff was friendly. Someti…
###
```

rvest

- Pros
  - **very** small and efficient interface
  - one moving part
- Cons
  - only reads plain text
  - does not execute javascript or flash
    - ex: autonation.com car information will fail

PhantomJS

- http://phantomjs.org/
- “PhantomJS is a headless WebKit scriptable with a JavaScript API. It has fast and native support for various web standards: DOM handling, CSS selector, JSON, Canvas, and SVG.”
- Executes javascript and flash
  - renders autonation.com

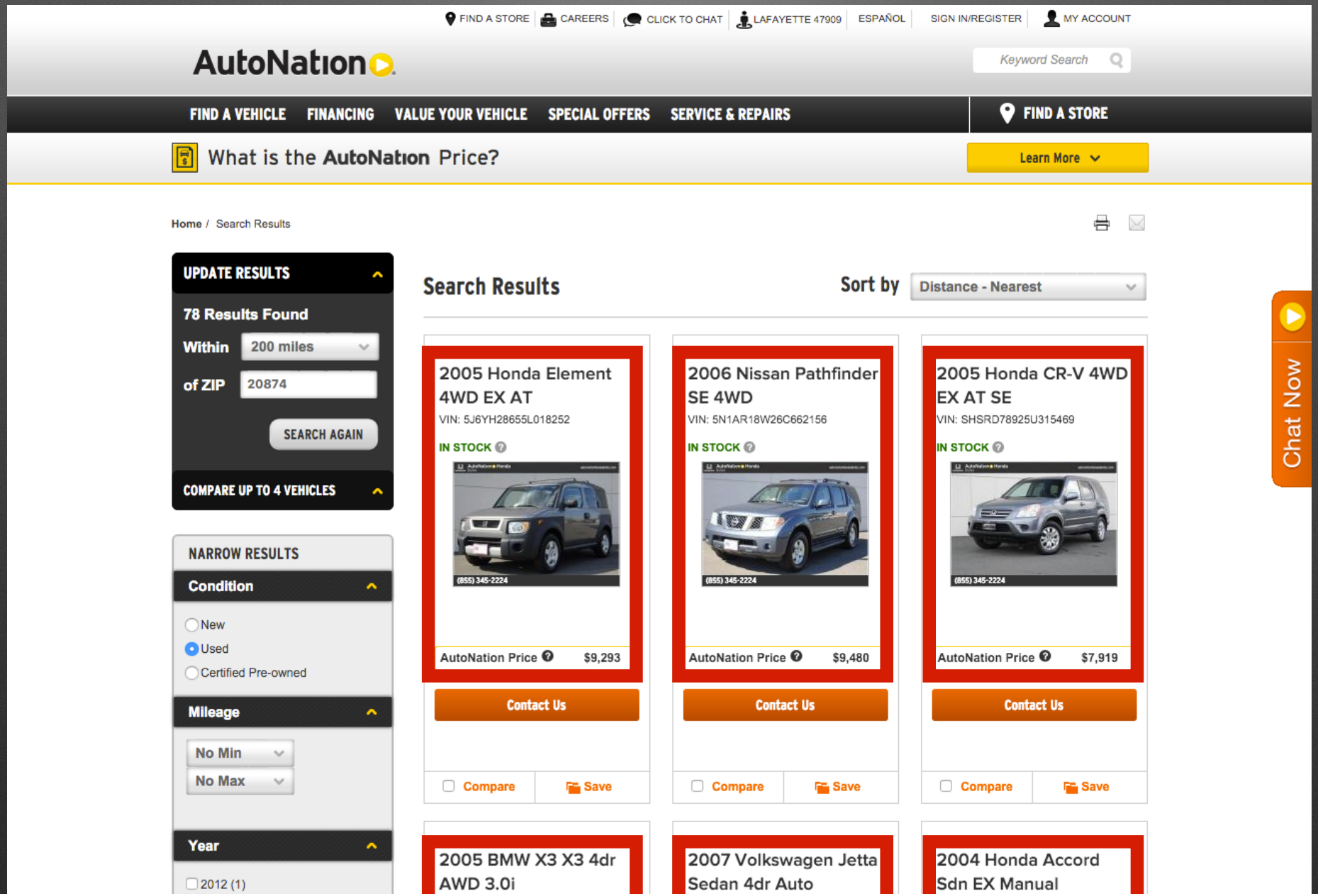
PhantomJS - Installation

```
pJS <- wdman::phantomjs(port = 4444L)
pJS$stop()
```

RSelenium

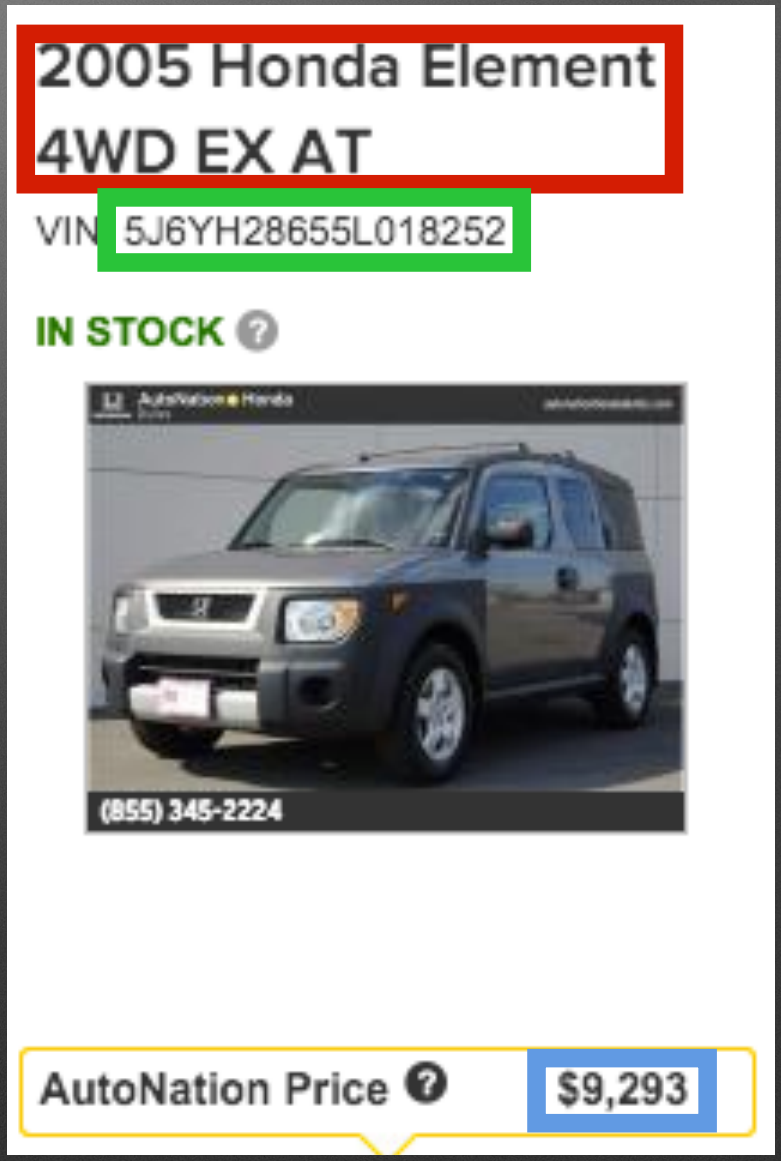
- “R Bindings for Selenium 2.0 Remote WebDriver”
- `browserName = “phantomjs”`
- `vignette(“RSelenium-headless”)`
- Retrieves web content and interacts with “browser”

AutoNation Example



AutoNation Example

- For each search result
  - **Name**
  - **VIN**
  - **Price**



AutoNation Example

```
library(RSelenium)
library(rvest)
library(magrittr)
library(stringr)

autoNationUrl <-
  "https://www.autonation.com/search/?cnd=USED&dst=200"

read_html(autoNationUrl) %>%
  html_nodes(".srp-filter-main-content .grid-tile")
# {xml_nodeset (1)}
[1] <div class="grid-tile col-lg-4 col-md-6 col-sm-12" ng-repeat="carDetails  …

# will not immediately work with rvest!
# must process website first!
```

AutoNation Example

```
# start phantomjs
pJS <- wdman::phantomjs(port = 4444L)
Sys.sleep(5) # give the binary a moment

remDr <- remoteDriver(browserName = 'phantomjs')
remDr$open()
remDr$setWindowSize(1800, 1800) # make not mobile

# navigate to page and let it load completely.
# will return when loaded
remDr$navigate(autoNationUrl)
Sys.sleep(5) # give site a moment
remDr$screenshot(file = "0-initial_load.png")

# retrieve processed page source
autoNationPageSource <- remDr$getPageSource()[[1]]

# close the PhantomJS process
remDr$close()
pJS$stop()
```

AutoNation Example

```
# post process information
searchResults <- read_html(autoNationPageSource) %>%
  html_nodes(".srp-filter-main-content .grid-tile")

searchResults
# {xml_nodeset (24)}
 [1] <div class="grid-tile col-lg-4 col-md-6 col-sm-12 ng-scope" ng-repeat="c ...
 [2] <div class="grid-tile col-lg-4 col-md-6 col-sm-12 ng-scope" ng-repeat="c ...
 [3] <div class="grid-tile col-lg-4 col-md-6 col-sm-12 ng-scope" ng-repeat="c ...
 [4] <div class="grid-tile col-lg-4 col-md-6 col-sm-12 ng-scope" ng-repeat="c ...
 [5] <div class="grid-tile col-lg-4 col-md-6 col-sm-12 ng-scope" ng-repeat="c ...
 [6] <div class="grid-tile col-lg-4 col-md-6 col-sm-12 ng-scope" ng-repeat="c ...
  …
```

AutoNation Example

```
vinNumbers <- searchResults %>%
  html_nodes(".vehicle-vin") %>%
  html_text() %>%
  str_trim() %>%
  str_replace("^VIN: ", "")

carName <- searchResults %>%
  html_nodes(".vehicle-name") %>%
  html_text() %>%
  str_trim()

carValue <- searchResults %>%
  html_nodes(".vehicle-price .price") %>%
  html_text() %>%
  str_trim() %>%
  str_replace("\\$", "") %>%
  str_replace(",", "") %>%
  as.numeric()
```

AutoNation Example

```
carDt <- tibble::tibble(
  name = carName, value = carValue, vin = vinNumbers
)

carDt
###
# A tibble: 24 × 3
                                      carName carValue               vin
                                        <chr>    <dbl>             <chr>
1   2014 BMW 4 Series 2dr Cpe 435i xDrive AWD    29600 WBA3R5C55EK186957
2                2014 BMW X3 AWD 4dr xDrive28i    33492 5UXWX9C50E0D31882
3   2011 BMW 3 Series 4dr Sdn 328i xDrive AWD    12595 WBAPK7C53BA463639
4                2013 BMW X1 AWD 4dr xDrive28i    19490 WBAVL1C54DVR84723
5            2008 Acura TL 4dr Sdn Auto Type-S    14885 19UUA76508A026616
6   2014 BMW 3 Series 4dr Sdn 328i xDrive AWD    30991 WBA3B3C52EJ981754
7                2014 BMW X5 AWD 4dr xDrive35i    38900 5UXKR0C51E0H25539
8                2014 BMW X5 AWD 4dr xDrive35i    47991 5UXKR0C50E0H19070
9  2014 BMW 6 Series 2dr Conv 640i xDrive AWD    55491 WBAYP5C50EDZ28097
10         2016 BMW 4 Series 2dr Conv 435i RWD    44900 WBA3T3C59G5A41494
# ... with 14 more rows
###
```

https and interactivity (OLD EXAMPLE, but works in theory!)

```
remDr$setWindowSize(1800, 1800)

# navigate to page and let it load completely.  will return when loaded
remDr$navigate("https://jobs.autonation.com")
remDr$screenshot(file = "0-initial_load.png")

# transitions to accounting page
# open list
remDr$findElement(using = "css selector", "div.job-category h2")$clickElement()
remDr$screenshot(file = "1-open_link.png")

# go to accounting
remDr$findElement(
  using = "css selector",
  "div.dropdowns.job-category div ul li a"
)$clickElement()
remDr$screenshot(file = "2-transition.png")

pageSource <- remDr$getPageSource()[[1]]
```

https - load

AutoNation. Careers

ABOUT AUTONATION | AUTONATION.COM | NEWS | CREATE A PROFILE-STORES | CREATE A PROFILE-CORPORATE
THIRD PARTY AGENCIES

Car Sales Jobs,
Auto Service Careers,
and Corporate
Opportunities at
AutoNation

Search Jobs | Keyword | City, State, or ZIP | Radius

Search Jobs by Category | Search Jobs by Location | Search Jobs by Group

You have the drive. We have the vehicle.

AutoNation is proud to be America's largest auto retailer representing domestic, import and luxury brands, both on and off the web. Headquartered in Fort Lauderdale, FL, and a member of the S&P 500 (NYSE AN), AutoNation employs approximately 25,000 people at over 290 store locations, representing over 290 new vehicle franchises across 15 states. We are driven to be the best, and we're always looking for passionate, motivated professionals who share the same drive to join us. Perhaps you're one of them.

Let us Search

With just 1 click and a connection with LinkedIn, you can receive job listings that best match your experience.

Jobs in AutoNation

Sign up for Job Alerts

First Name
First Name

Last Name
Last Name

Phone Number (Optional)
Phone Number

Career Resources

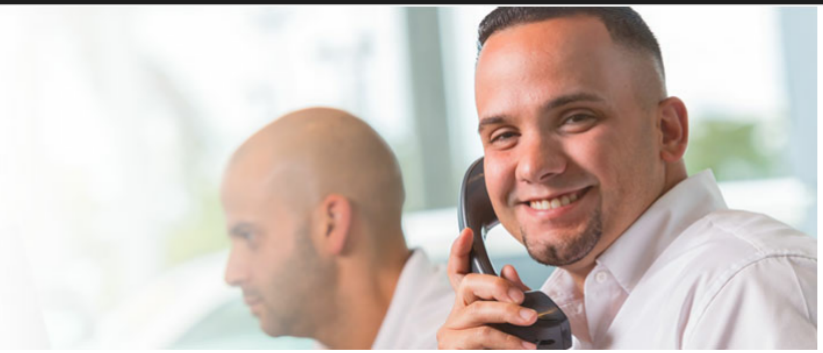
- What makes a great Salesperson at AutoNation
- Learn about being a Service Technician at AutoNation
- Opportunities for current students and recent Auto Technician Graduates
- Application Tips

https - click

AutoNation. Careers

ABOUT AUTONATION | AUTONATION.COM | NEWS | CREATE A PROFILE-STORES | CREATE A PROFILE-CORPORATE
THIRD PARTY AGENCIES

Car Sales Jobs,
Auto Service Careers,
and Corporate
Opportunities at
AutoNation



Search Jobs | Keyword | City, State, or ZIP | Radius

Search Jobs by Category

- Accounting Jobs
- Administrative Jobs
- Auto Body Jobs
- Auto Detail Jobs
- Auto Sales Jobs

view all

Search Jobs by Location

Search Jobs by Group

ve the drive. We have the vehicle.

s largest auto retailer representing domestic, import and luxury brands, both on and off the web.
nd a member of the S&P 500 (NYSE AN), AutoNation employs approximately 25,000 people at over
290 new vehicle franchises across 15 states. We are driven to be the best, and we're always looking
ted professionals who share the same drive to join us. Perhaps you're one of them.

Let us Search

With just 1 click and a connection with LinkedIn, you can receive job listings that best match your experience.

Jobs in AutoNation

Sign up for Job Alerts

First Name
First Name

Last Name
Last Name

Phone Number (Optional)
Phone Number

Career Resources

- What makes a great Salesperson at AutoNation
- Learn about being a Service Technician at AutoNation
- Opportunities for current students and recent Auto Technician Graduates
- Application Tips

https - follow link

AutoNation. Careers

ABOUT AUTONATION | AUTONATION.COM | NEWS | CREATE A PROFILE-STORES | CREATE A PROFILE-CORPORATE
THIRD PARTY AGENCIES

Accounting

AutoNation - America's largest automotive retailer is looking for qualified people to join our team. If you are looking for a career that will allow you the opportunity to create results and accomplish goals, join us with a job on our Accounting team. Our ability to perform at such a high level is directly related to the efforts of our talented team of professionals, and we're looking for more great people to join us in our Accounting team. If you have the Drive, We have the Vehicle. Learn more about a job with AutoNation today.

Search Jobs | Keyword | City, State, or ZIP | Radius

Filter

State [-]
California (1)
Florida (1)
Virginia (1)

City [+]

Store [+]

Position Type [+]

Reset

3 Results found for Accounting

Accounting Associate - Mercedes-Benz of Delray
Delray Beach, Florida 09/21/2015

Accounting Associate Mercedes-Benz Stevens Creek
San Jose, California 09/23/2015

Accounting Associate - Honda Dulles
Sterling, Virginia 10/09/2015

Share:

Sign up for Job Alerts

First Name
First Name

Last Name
Last Name

Phone Number (Optional)
Phone Number

Email Address*
Example@domain.com

Interested In*
Search for a category, location, or category/location pair, select a term from the suggestions, and click "Add".

Category
Type to search

Location

https - Results

```
jobs <- read_html(pageSource) %>%
  html_nodes("section#search-results-list ul li a")

jobTitle <- jobs %>%
  html_nodes("h2") %>%
  html_text()

jobLocation <- jobs %>%
  html_nodes("span.job-location") %>%
  html_text()

jobDatePosted <- jobs %>%
  html_nodes("span.job-date-posted") %>%
  html_text()
```

https - Results

```
jobDt <- data.frame(title = jobTitle, location =
jobLocation, datePosted = jobDatePosted)
jobDt

###
                                                title              location
1   Accounting Associate - Mercedes-Benz of Delray Delray Beach, Florida
2 Accounting Associate Mercedes-Benz Stevens Creek  San Jose, California
3              Accounting Associate - Honda Dulles    Sterling, Virginia
  datePosted
1 09/21/2015
2 09/23/2015
3 10/09/2015
###
```

Words of Caution

- IP
- It’s obvious

Recap

- rvest
  - use as much as possible!
  - small interface
- RSelenium
  - big and clunky, but much more powerful
  - parses full websites (javascript and flash)

Questions?

NCAA Example

- Every team info per year:
  - http://www.espn.com/mens-college-basketball/statistics/team/_/stat/scoring-per-game/sort/avgPoints/year/2016/seasontype/2
- Player statistics:
  - http://www.espn.com/mens-college-basketball/team/stats/_/id/2509/year/2016