Sprint Review 6

EDA Gone Rogue

Scraping and Parsing Income Data from the Web

By: Victoria Larson

devleague

From these state links!

http://livingwage.mit.edu/

States

Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

Typical Annual Salaries

These are the typical annual salaries for various professions in this location.

Occupational Area	Typical Annual Salary
Management	$98,267
Business & Financial Operations	$64,780
Computer & Mathematical	$77,521
Architecture & Engineering	$83,773
Life, Physical, & Social Science	$58,241
Community & Social Service	$41,923
Legal	$64,985
Education, Training, & Library	$46,853
Arts, Design, Entertainment, Sports, & Media	$33,938
Healthcare Practitioners & Technical	$52,798
Healthcare Support	$24,241
Protective Service	$34,153
Food Preparation & Serving Related	$19,229

This table!

```
# giving URL a name
URL_Weber <- http://livingwage.mit.edu/counties/49057

#reading in the URL
read_html(URL_Weber)

# Giving the read URL a name
LW_Weber <- read_html(URL_Weber)
# Getting all of the table outall_tbls <- html_nodes(LW_Weber, "table")

# just getting the Occupations Table
occ_tbls <- html_nodes(LW_Weber, "table.occupations_table")

# Turning the List into a data.frame
html_table(occ_tbls)

# renaming the table Weber_Occ1
Weber_Occ1 <- html_table(occ_tbls[[1]])
```

```
state_url_from_num = function(state_num)

{state_url = paste("http://livingwage.mit.edu/states/",state_num, sep = "")

return(state_url)
}
```

EXAMPLE: State_url_from_num("04") = http://livingwage.mit.edu/states/04

Input — Function — Output

```
get_occ_table_from_url = function(state_url) {
state_html <- read_html(state_url)
occ_node <- html_nodes(state_html, "table.occupations_table")
occ_table <- html_table(occ_node[[1]])
return(occ_table)
```

Grabs the table that we want from the html page using a CSS selector

Turns the html table into a data.frame

EXAMPLE:

```
get_occ_table_from_url("http://livingwage.mit.edu/states/05")
```

That's seems too hard…..

	Occupational Area	Typical Annual Salary
1	Management	$75,061
2	Business & Financial Operations	$57,011
3	Computer & Mathematical	$66,451
4	Architecture & Engineering	$64,268
5	Life, Physical, & Social Science	$51,578
6	Community & Social Service	$36,613
7	Legal	$55,237
8	Education, Training, & Library	$43,716
9	Arts, Design, Entertainment, Sports, & Media	$36,572
10	Healthcare Practitioners & Technical	$51,506
11	Healthcare Support	$23,391
12	Protective Service	$33,559
13	Food Preparation & Serving Related	$19,291
14	Building & Grounds Cleaning & Maintenance	$20,982
15	Personal Care & Service	$19,742
16	Sales & Related	$23,011
17	Office & Administrative Support	$29,500
18	Farming, Fishing, & Forestry	$25,666
19	Construction & Extraction	$34,471
20	Installation, Maintenance, & Repair	$38,007
21	Production	$30,299

```
get_occ_from_state_num2 = function(state_num){
state_url = state_url_from_num(state_num)
occ_table = get_occ_table_from_url(state_url)
return(occ_table)
}
```

Example:

```
get_occ_from_state_num2("05")
```

Input ___ Function ___ Output

devleague

So much easier!!!

	Occupational Area	Typical Annual Salary
1	Management	$75,061
2	Business & Financial Operations	$57,011
3	Computer & Mathematical	$66,451
4	Architecture & Engineering	$64,268
5	Life, Physical, & Social Science	$51,578
6	Community & Social Service	$36,613
7	Legal	$55,237
8	Education, Training, & Library	$43,716
9	Arts, Design, Entertainment, Sports, & Media	$36,572
10	Healthcare Practitioners & Technical	$51,506
11	Healthcare Support	$23,391
12	Protective Service	$33,559
13	Food Preparation & Serving Related	$19,291
14	Building & Grounds Cleaning & Maintenance	$20,982
15	Personal Care & Service	$19,742
16	Sales & Related	$23,011
17	Office & Administrative Support	$29,500
18	Farming, Fishing, & Forestry	$25,666
19	Construction & Extraction	$34,471
20	Installation, Maintenance, & Repair	$38,007
21	Production	$30,299

```
function.states = c("Alabama","Alaska","Arizona")
url_num = c("01", "02", "04”)
state_info = data.frame(
states,
url_num
)

gst = function(state_name,state_num){
Test_table <- get_occ_from_state_num2(state_num)
Test_table$State <- state_name
return(Test_table)}

State_Occupations = bind_rows(apply(state_info,1,function(row) gst(row[1],row[2])))
```

	Occupational Area	Typical Annual Salary	State
1	Management	$98,267	Alabama
2	Business & Financial Operations	$64,780	Alabama
3	Computer & Mathematical	$77,521	Alabama
4	Architecture & Engineering	$83,773	Alabama
5	Life, Physical, & Social Science	$58,241	Alabama
6	Community & Social Service	$41,923	Alabama
7	Legal	$64,985	Alabama
8	Education, Training, & Library	$46,853	Alabama
9	Arts, Design, Entertainment, Sports, & Media	$33,938	Alabama
10	Healthcare Practitioners & Technical	$52,798	Alabama
11	Healthcare Support	$24,241	Alabama
12	Protective Service	$34,153	Alabama
13	Food Preparation & Serving Related	$19,229	Alabama
14	Building & Grounds Cleaning & Maintenance	$22,191	Alabama
15	Personal Care & Service	$19,506	Alabama
16	Sales & Related	$23,975	Alabama
17	Office & Administrative Support	$31,252	Alabama
18	Farming, Fishing, & Forestry	$30,668	Alabama
19	Construction & Extraction	$37,433	Alabama
20	Installation, Maintenance, & Repair	$42,978	Alabama
21	Production	$31,406	Alabama

Input — Function — Output

Cleaned up the data…

Moved state to the left

- State_Occupations <- State_Occupations[,c(3,1:2)]

Changed the column names

names(State_Occupations)[names(State_Occupations) == 'Typical Annual Salary'] <-
'Annual_Salary'names(State_Occupations)[names(State_Occupations) == 'Occupational Area'] <- 'Occupational_Area'

Change to numeric…. Round one…

Took out the dollar sign

parse_number(State_Occupations$Annual_Salary)State_Occupations1 <-
parse_number(State_Occupations$Annual_Salary)State_Occupations$Annual_Salary <- State_Occupations1

Change to Numeric… round two…

Change to numeric

State_Occupations$Annual_Salary <- as.numeric(State_Occupations$Annual_Salary)

Behold!

	State	Occupational_Area	Annual_Salary
1	Alabama	Management	98267
2	Alabama	Business & Financial Operations	64780
3	Alabama	Computer & Mathematical	77521
4	Alabama	Architecture & Engineering	83773
5	Alabama	Life, Physical, & Social Science	58241
6	Alabama	Community & Social Service	41923
7	Alabama	Legal	64985
8	Alabama	Education, Training, & Library	46853
9	Alabama	Arts, Design, Entertainment, Sports, & Media	33938
10	Alabama	Healthcare Practitioners & Technical	52798
11	Alabama	Healthcare Support	24241
12	Alabama	Protective Service	34153
13	Alabama	Food Preparation & Serving Related	19229
14	Alabama	Building & Grounds Cleaning & Maintenance	22191

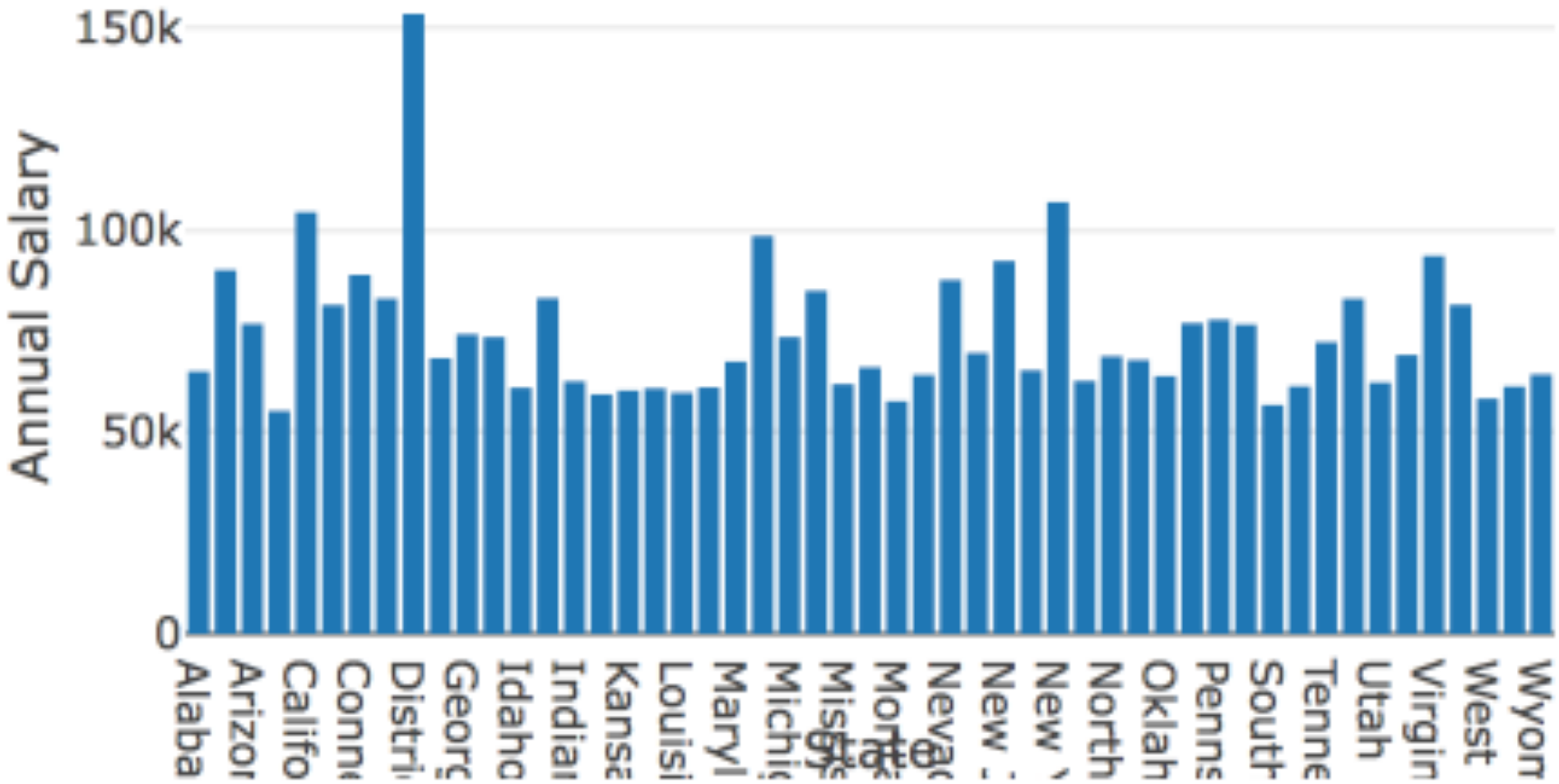
What was the real assignment?

EDA!!!!!

In which field and which state is the highest paying job?

State	Occupational_Area	Annual_ Salary
District of Columbia	Legal	153535

Compare Legal Salary in all States



This leads to more questions…

Fin