Web Scraping Final Project

Hassan Abbas

12/21/2021

Introduction

The assignment was done as a part of our final project of Web scraping. The aim of this project was to extract html data from **Payscale** and save it in json format and visualize it.Initially the attempt was to obtain a API for the site and scrape data using that method but the API was paid and could not be access with proper credentials. Therefore, we inspect the data and tried to locate the json file and extract the data through Xpath method.

Extracting data from Json

The first task was to extract links for each jobs within each industries. Since the all the information was not on one page, we had to create loop to to obtain links for specific industry and then jobs within each industry. This process was done with the help of Rvest function & XML. Once we had to link we extracted the json file through xpath and filtered to through the list to create data table for Jobs, Experience ,Employment & Location. The following is one of the functions we create to make our data table. The whole data scarping project file can be access **here**

```
#All_jobs <- function(x){
 # Links <- read_html(x)
#  Data <- fromJSON(Links %>% html_node(xpath = '//script[@type="application/json"]') %>%
 # job_header <- Data$props$pageProps$pageData$dimensions$job
 # Average_salary <- Data$props$pageProps$pageData$byDimension$`Average Salary Overall`$r
 # Skill_set <- Data$props$pageProps$pageData$occupationalDetails$skills
 # Average_Hourly_rate <- Data$props$pageProps$pageData$byDimension$`Average Hourly Rate
 # Description <- Data$props$pageProps$pageData$narratives$description
 # Sample_size <- Data$props$pageProps$pageData$occupationalDetails$sampleSize
 # Job_satifaction <- Data$props$pageProps$pageData$ratings$`Job Satisfaction Overall`$sc
 # Female_ratio <- round((Data$props$pageProps$pageData$byDimension$`Gender Breakdown`$ro
 # Male_ratio <- round((Data$props$pageProps$pageData$byDimension$`Gender Breakdown`$rows
 # Top_employer <- Data$props$pageProps$pageData$byDimension$`Job by Employer`$rows$name[

 # final_data <-  data.table(job_header,Average_Hourly_rate,Average_salary,Skill_set,Samp
#}
#Salary_data <- rbindlist(pblapply(All_links, All_jobs ), fill = T)
```

Process Overview

Since the data loading is lengthy process we saved all our data into RDs and Csv files. Then we called our data from **Github** and made visualization on them. The advantage of saving data in RDs is that you can directly load the data into your environment.Our final data has more than 2800 observation and 19 variable that provided us with the descriptive variable for each jobs including the changes in pay scale with the level of experience.Moreover, for each job there were further drill-downs that provided you with the salary based on employer and the location of your occupation.Therefore, we made further data tables which include the salary based on occupation location and the employer.

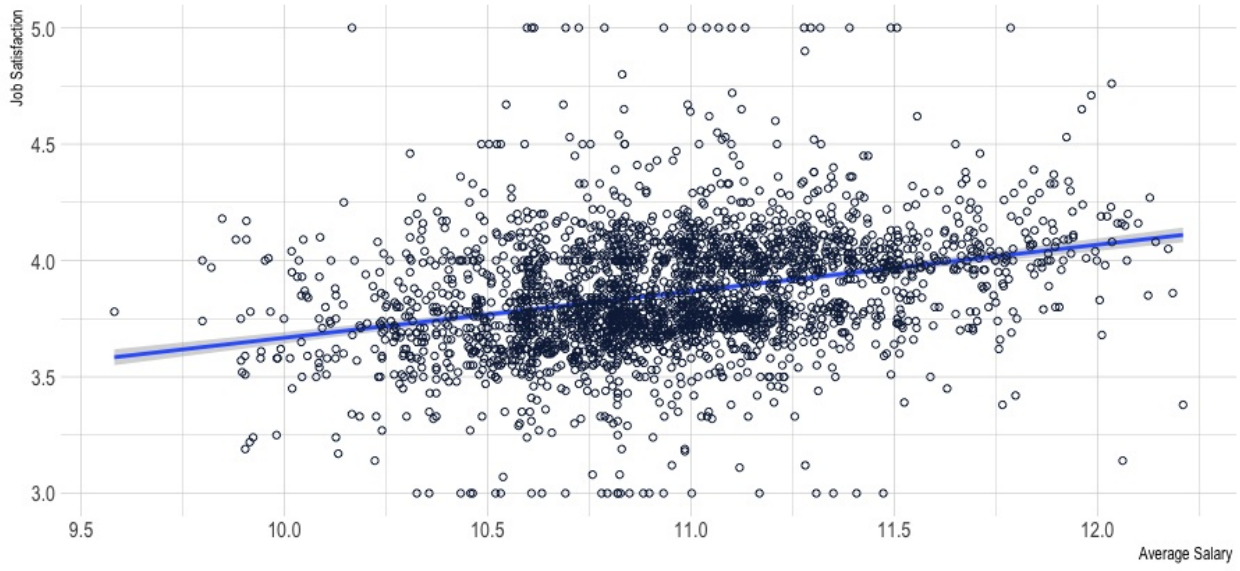
One of the problem with data was that some salary were in Hourly bases and others were in total average salary per year. Due to which we created an additional column stating the type of value mention in salary column. Once all the data table were created we combine the the experience and salary data into one single table based on job titles. However, we kept the location and employer data table separate to avoid overloading of variables in our data. Once we loaded our data we cleaned the data for NA values and removed the columns which were irrelevant.

Analysis

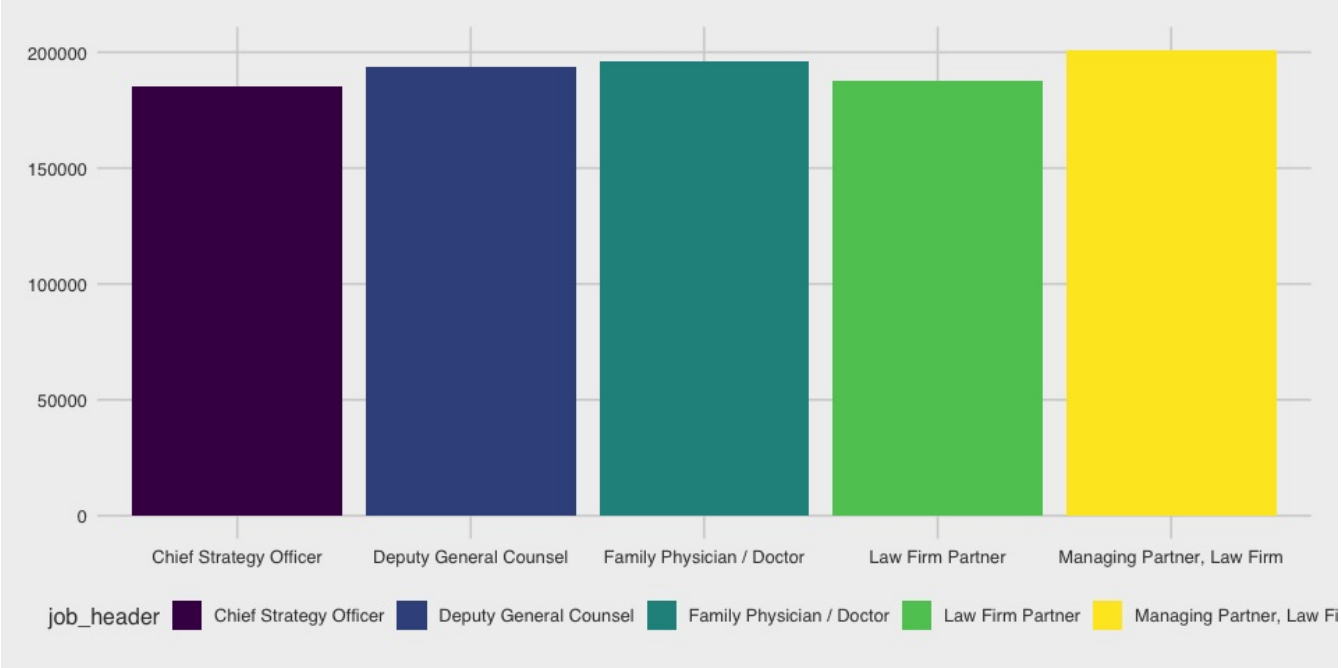
Now once we had the data in a clean format we wanted to create visualization and run regressions to see the association between the variables. Following our some meaningful interpretation abour our data.

Linear Regression of Job Satisfaction Vs Log Salary

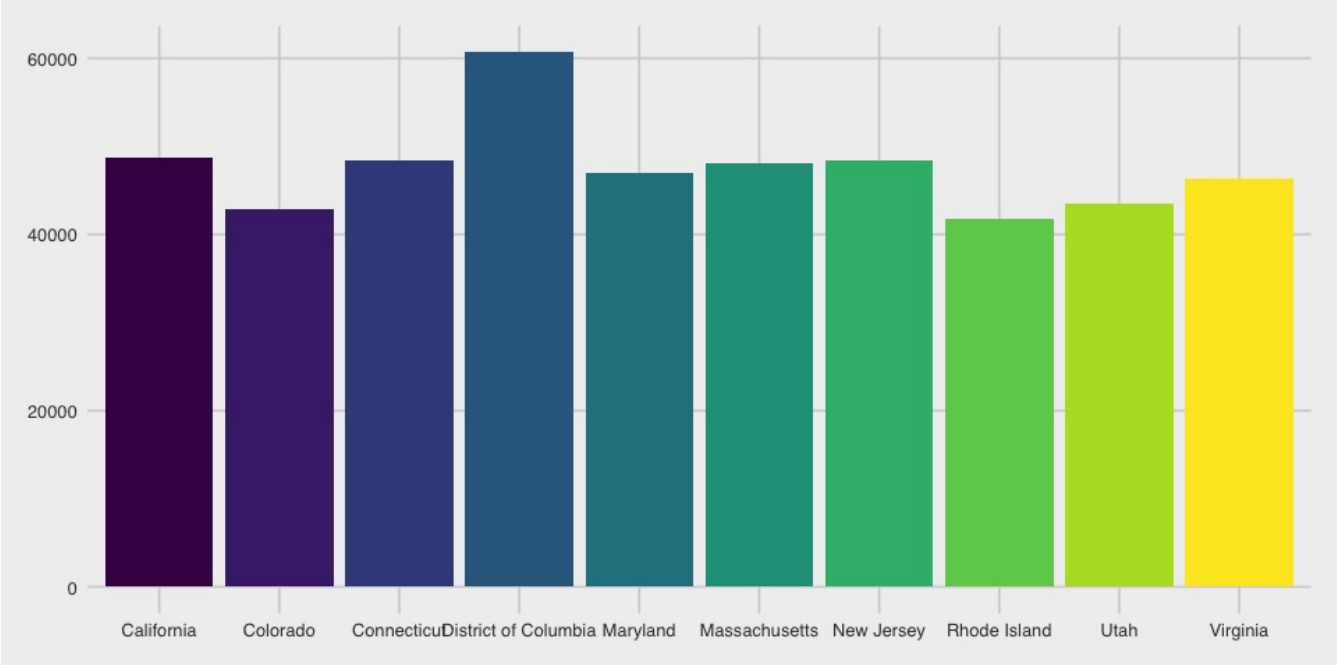
For the purpose of regression we took Job satisfaction and Average salary to see how are they related and what is the beta coefficient. Before we run the regression we had to check if our values in these variables were skewed or not. We witness skewness in the salary variable and therefore we took log. The regression result showed that with 1% increase in Average salary , the β coefficient of job satisfaction would increment by 0.03 units on an average.



The Top Five Highest Paying Occupation



Highest Number of Average salary by State



Company with Highest Average Salary