

STATISTICS 260/261: Introductory R for Data Science

FINAL PROJECT

Due Dates: March 24. Submit your report and R code to Crowdmark. We will send out the Crowdmark link a week before the due.

POLICY

1. This project is to be completed *independently*. You may use whatever materials you wish in completing this assignment. **BUT DO NOT DISCUSS RESULTS WITH ANYONE ELSE, WITHIN OR OUTSIDE OF THE CLASS.** Failure to follow this directive will result in a failing grade.
2. Late projects will be accepted at a penalty of 2 points/hour (it's a 100 point project).
3. You are allowed to clarify the project requirements, but you are not advised to show or discuss your answer with the TA and instructor and seek feedback from them.

ASSIGNMENT

One of the key skills of transferable programming skills is to be able to dive deeply on your own, read the help documents and online tutorials to learn about how to solve a problem in R on your own.

The project consists of two parts. In part 1, you will write an explanatory tutorial article on how to do web scraping in R. You can choose your own example webpage to scrape, and R packages for doing the web scraping. In part 2, you will be given a website, and you will write R code to scrape a specific data table into R.

Part 1:

The maximum page limit is 7 pages. There are no guidelines on font size and page margin so you can choose any size that you think is appropriate. You will write tutorial articles on how to do web scraping in R. The intended audience are people who have very basic R knowledge and don't know how to do web scraping. You will explain in your own words how to do. You can pick a website, and show R code example of how to web scrape the info from that website.

Your score in the report depends on the degree of comprehensiveness and complexity you demonstrated in the R code. There are various packages for R scraping (e.g., Rvest, RCrawler, RSelenium) and some packages are more versatile and powerful than others.

Tips on getting started:

You can start by going through online tutorial and videos on web scraping in R. There are abundant resources online for doing this, by simply google "web scraping tutorial R".

If you struggle with basic R, find two free mornings or evenings to go through either of the following online interactive course modules.

<https://www.dataquest.io/course/intro-to-r-rewrite/>

<https://app.datacamp.com/learn/courses/free-introduction-to-r>

From there, you should be able to navigate various tutorial R code on web scraping with ease.

The basic principle of learning by yourself is “*If you ever find it too difficult to comprehend, walk away and find one that is easier*”. Do not dwell on overly challenging things. There are always resources on the next item in the google page that suits your capabilities and needs. Learn it step by step and start with the ones that you are able to handle. Gradually you will be able to more difficult stuff.

There are other youtube videos or online course module (e.g., udemy.com) that you can rely on if you feel that watching videos is easier for you to learn.

If you ever encounter a concept hard to understand (e.g., HTML, XPath), take a step back and search for tutorials and videos on that to gain basic understanding of it first.

Citation:

To avoid plagiarism, at the end of your report, you should list out the online resources that you used for creating the tutorial. You should use your own word to explain the R code.

Part 2:

The task in Part 2 is to write R code for webscraping a specific website (detailed as follows). You should submit your R code which is free of error and installs all the necessary packages. The R code should achieve the following task.

- Go to <https://forecast.weather.gov/MapClick.php?lat=37.7772&lon=-122.4168>
- Create a web scraper in R that downloads the information from the website above. The program will download the data and summarize weather data from the following section of the website (screenshot attached below) into a table.



The downloaded table should look like the following.

	period	short_desc	temp
0	Tonight	Mostly Clear	Low: 49 °F
1	Thursday	Sunny	High: 63 °F
2	ThursdayNight	Mostly Clear	Low: 50 °F

	period	short_desc	temp
3	Friday	Sunny	High: 67 °F
4	FridayNight	Slight ChanceRain	Low: 57 °F
5	Saturday	Rain Likely	High: 64 °F
6	SaturdayNight	Rain Likely	Low: 57 °F
7	Sunday	Rain Likely	High: 64 °F
8	SundayNight	Chance Rain	Low: 55 °F

- If you study the example table above, you will see that each row comes from the forecast from the following section of the website (from left to right).
- In the last line of your R code, you should use the following code to write the table to your csv file. The title of the csv file should be your student ID. For example, suppose the table is downloaded into a variable called “tbl”, and suppose your Student ID is 301323011. The last line of R code should be

```
write.table(tbl, "301323011.csv", sep = ",", row.names = F,
col.names =F)
```

Some other notes for marking:

- No points will be deducted if there are some missing spaces in your extracted data, e.g., “FridayNight” rather than “Friday Night”.
- No points will be deducted if you missed the first indexing column, i.e., the column that contains 1,2,3,4...

GRADES

Your grade will be based partly on how well your report is written and partly based on whether the R code you submit achieves the desired target. Part 1 will count for 75% of your grade and part 2 will be 25%.

FINAL COMMENTS

I hope this is a useful experience for you. I hope that many of you can learn from this journey and prove to yourself that you are capable to handle challenges on your own. Remember, in real life you will face a situation where your job requires you to acquire a technically challenging skill on your own. This is practice...