

# ESM 206 – ASSIGNMENT 2

Get, wrangle, visualize and present data in reports with R Markdown

Due Wednesday 2019-10-16 at 8:00am

Read through the entire assignment before starting. **Follow instructions carefully.**  
There is a very detailed “What you will submit” section at the end of the assignment.

**Submit exactly and only what is requested.**

**For Assignment 2, you will:**

- Ensure that RStudio and GitHub are talking to each other
- Read Wilson et al. (2017) and Ch. 4 - 7 of Claus O. Wilke's *Fundamentals of Data Visualization*
- Create two R Markdown documents that contain text, code, and graphs for Tasks 1 & 2

**General notes for Assignment 2:**

- All files needed for this assignment are posted on Gauchospace under the assignment prompt
- You will not submit anything for the readings in this assignment, but you **are responsible for understanding** the reading materials
- You will submit both your .Rmd file and your knitted HTML for each task
- All figures should be customized & finalized
- You will still be graded on organization, reproducibility and good coding practices

## A. CONNECT RSTUDIO & YOUR GITHUB ACCOUNT

If you haven't already, follow along with [this document](#) to connect your local RStudio with your GitHub account so that you can push and pull between the two. You need to have this **working by Friday, October 18th**, so try early, be resourceful with troubleshooting and come to office hours during Weeks 2 & 3 with questions.

## B. READINGS

- [Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK \(2017\) Good enough practices in scientific computing. PLoS Comput Biol 13\(6\): e1005510.   
<https://doi.org/10.1371/journal.pcbi.1005510>](#)
- [Chapters 4 - 7 in Fundamentals of Data Visualization](#) by Claus O. Wilke

## C. R MARKDOWN TASKS

### Task 1. Pollution burden inequity in California

For Task 1, you will create a mini-report using R Markdown to explore and visualize indicators for pollution burden (e.g. air quality, asthma occurrence) and investigate their relationships with demographics in California. For Task 1, **all of your code, any code outputs (incl. warnings/messages) and all graphs should show up** in the knitted HTML that you submit, along with any text you are asked to include.

**Download & read** this [short flyer from the California Environmental Justice Alliance](#) to understand how CA OEHHA defines/calculates data *pollution burden* and *CES 3.0*.

**Data source:** California Office of Environmental Health Hazard Assessment (OEHHA)'s CalEnviroScreen database: <https://oehha.ca.gov/calenviroscreen/maps-data/download-data>

**Files you'll use for Task 1 (download from GauchoSpace):**

- **ca\_pollution\_burden.csv:** data for metrics of pollution burden in California by census tract (2018)
- **ca\_census\_demographics\_2010.csv:** selected demographics (race, poverty, age) for California census tracts, from 2010 census

**Variable information:** More detailed descriptions of the variables (columns) are included in the "Data dictionary" worksheet of the "ca\_pollution\_burden\_metadata.xlsx" file. Note: you do not need to import that file into R - it's just so that you have the variable descriptions.

**To complete Task 1:**

- A. Create a new R project for Task 1 within your ESM 206 Assignments > Assignment 2 (or whatever yours is named) folder on your computer
- B. Copy and paste the two CSV files (**ca\_pollution\_burden.csv** and **ca\_census\_demographics\_2010.csv**) into the working directory
- C. Create a new well-formatted, organized R Markdown document (.Rmd) saved as **"a2\_task1\_firstname\_lastname.Rmd"** in which you first:
  - a. In a very brief (3 - 4 sentence) "introduction," define/describe pollution burden and how it relates to environmental justice (using references as necessary), and introduce what you are specifically exploring in this report. It may be easiest to write the introduction *last*.

- b. Provide a brief (1 - 2 sentence) text overview of the dataset, with a citation (always give credit to collectors & sharers of data)
  - c. Attach any necessary packages
  - d. Read in both data files, and clean up the column names
  - e. Join the two data frames together by census tract number
- D. In thoughtfully labeled subsections and using good practices for writing code and naming things, complete the following:
  - a. Calculate the mean **CES 3.0 Score by County**, by averaging the census tract level scores for each county in CA, then determine the top 10 **worst (highest)** counties by average CES 3.0 Score. Create a graph to visualize the CES 3.0 Score for those 10 worst counties (you can choose to either show the county mean scores, or all census tract scores within each county, or something that combines the two). Customize and finalize your graph, adding appropriate axis labels and a helpful title.
  - b. Create a plot to explore the relationship between demographic estimate for “% white” and asthma rates across all census tracts ([hint: click here for some cool options](#)). Customize and finalize your graph, adding appropriate axis labels and a helpful title. In 1 - 2 sentences below the final graph, describe any trends and/or patterns that you observe.
  - c. **Coder’s choice.** Pick another relationship that you want to explore between 2 (or more) variables in the datasets for Task 1. You can pick how you want to limit your exploration spatially, demographically, and/or based on the pollution burden variables. In 1 - 2 sentences above your code to create the graph, describe the relationship you are going to explore. Visualize the relationship, customize & finalize your graph with updated axis labels and a useful title. Below the graph, add 1 - 2 sentences describing any patterns/trends (or lack of) that you notice.
- E. Save your .Rmd and knitted HTML (**a2\_task1\_firstname\_lastname.Rmd & a2\_task1\_firstname\_lastname.html**). You will submit both.

## **Task 2. Mismanaged plastics: from rivers to oceans**

For Task 2, you'll create a document with R Markdown to visualize plastic (micro- and macroplastic) loads in river catchments. In your Task 2 .Rmd, set code chunks to **hide all code and code warnings/messages in the knitted HTML** that you submit. The HTML you submit for Task 2 should contain only well-formatted headers and subsections, text, and your finalized figures.

- **Citation:** [Schmidt C, Krauth T, Wagner S \(2017\). \*Export of plastic debris by rivers into the sea. Environ. Sci. Technol.\* v. 51 \(21\), p. 12246-12253.](#)
- **Data:** posted to Gauchospace, **plastic\_rivers.csv**

### **To complete Task 2:**

- A. Create a new R project for Task 2 within your ESM 206 Assignments > Assignment 2 (or whatever yours is named) folder on your computer
- B. Copy and paste the plastic\_rivers.csv file into your Task 2 working directory
- C. Create a new well-formatted, organized R Markdown document (.Rmd) saved as **"a2\_task2\_firstname\_lastname.Rmd"** in which you first:
  - a. Add a brief (3 - 4 sentence) paragraph introducing what you are exploring in this report
  - b. Include the citation
  - c. Attach packages
  - d. Read in the plastic\_rivers.csv
- D. In thoughtfully labeled subsections and using good practices for writing code and naming things, complete the following:
  - a. Calculate the total estimated annual **microplastics** load (using Model 2, which "takes data sets into account where both microplastic and macroplastic fraction were measured") for each 'Receiving Sea' by totalling the microplastics load (Model 2) across river catchments flowing into each sea. Then create a graph of the top 15 most impacted seas by microplastic load. You should choose the graph type, and customize to make it clear, professional and nice to look at. Add concise axis labels, with units as relevant.
  - b. Which 5 river catchments are contributing the **greatest total plastic loads** (including microplastics from Model 2, *and* macroplastics) to the Gulf of Mexico?

Write code to answer the question, then create and finalize a graph which shows the total plastics loads associated with those 5 rivers.

- E. Save your .Rmd and knitted HTML (**a2\_task2\_firstname\_lastname.Rmd & a2\_task2\_firstname\_lastname.html**). You will submit both.

### **What you will submit for Assignment 2:**

#### **Your .Rmd and knitted .html files for Tasks 1 & 2**

That's it. Make sure you've read the instructions carefully - all code, messages & warnings should appear for your knitted html in Task 1, but should be hidden for your knitted Task 2 html.

### **What you will be graded on for Assignment 2:**

The two tasks are worth 15 points each for a total of 30 points. Scores for each task are based on code functionality (8 pts), organization & good coding habits (3 pts), and overall clarity & correctness of graphs (4 pts).

- Functionality:
  - Code does not work = 0 points
  - Code partially works = 4 points
  - Code all works correctly = 8 points
- Organization/good coding habits:
  - No organization/annotation = 0 points
  - Average organization/annotation = 1 point
  - Excellent organization/annotation = 3 points
- Graph clarity & correctness:
  - No graphs, or incorrect graphs = 0 points
  - Correct data, poorly presented = 2 point
  - Correct & well presented graphs = 4 points

### **How to submit your Assignment 2 files:**

Upload your four files to the 'Submit Assignment 2' link on Gauchospace. Once you've submitted, congratulations! You're done with your second ESM 206 assignment.