SML 201 Project 2

[full names for each group member]

2024-10-21

Table of contents

Project 2 **2**
- Quality . . . 2

Setting the Scene **3**
- Task 1 . . . 3
- Task 2 . . . 3
- Task 3 . . . 3

Updating the Audience **3**
- Task 4 . . . 3
- Task 5 . . . 4

Building Regression Models **4**
- Task 6 . . . 4
- Task 7 . . . 4
- Task 8 . . . 4
- Task 9 . . . 4

Exploratory Bioinformatics **5**
- Task 10 . . . 5
- Task 11 . . . 5
- Task 12 . . . 5

Wrap Up **5**

Project 2

In this project, we will replicate studies made by biologists and epidemiologists about wastewater tracking for monitoring the Covid-19 pandemic in the United States.

- data files from NWSS (National Wastewater Surveillance System) with weekly observations
- shapefiles from the US Census Bureau

Quality

In addition to completing the tasks below with your group, we ask for the following level of quality

- all graphs are made with the `ggplot2` package
  - are labeled with at least a title and subtitle
  - caption includes the data source (namely: "Source: NWSS")
  - use colors beyond the default shades of gray
- all tables are made with the `gt` package
  - include the source (namely: "Source: NWSS")

At the end of each **task**, describe the results with about 3 to 5 sentences. Descriptions may include

- what you found interesting about the data
- discussion about pros and flaws in the analysis
- Some tasks specifically ask for certain elements in the description. You do not have to provide an additional paragraph of description.
- Contrary to Precept 1, do not place the description in a code block.

Load the R packages that we will need in this project.

```
library("gt")
library("sf")
library("tidyverse")
```

Load the data files that we will need in this project.

```
states_shp <- readr::read_rds("https://raw.githubusercontent.com/dsollberger/sml201slides/re
wastewater_regional_data <- readr::read_csv("https://raw.githubusercontent.com/dsollberger/s
wastewater_state_data <- readr::read_csv("https://raw.githubusercontent.com/dsollberger/sml2
wastewater_variant_data <- readr::read_csv("https://raw.githubusercontent.com/dsollberger/sml
```

Setting the Scene

Task 1

Write an introductory paragraph for this project. Preview some of the techniques that will be used. Mention the data source. (Tip: write this paragraph after finishing the other tasks.)

Task 2

Using the `states_shp` shapefile, draw a map that shows the 4 regions of the contiguous United States as defined in the NWSS national trends website. (If need be, you may make 4 maps.)

Task 3

Using this project's `wastewater_regional_data` data file, replicate the line chart seen in the NWSS national trends website, but in `ggplot`.

- You do not have to include the "activity levels" here
- For full credit on this task, your plot has to include an appropriate legend and 5 curves from `geom_line` usage.
- You do not have to use the same colors as the NWSS website

Updating the Audience

Task 4

Using the most recent day of observations in this project's `wastewater_state_data` data file, color-code a `states_shp` map according to the percent difference between `state_med_conc` and `national_value`

$$\text{percent difference} = \frac{100(\text{state med conc} - \text{national value})}{\text{national value}}$$

Task 5

Repeat the previous task, but with `region_value` instead of `national_value`.

Building Regression Models

Task 6

Use `mutate` in the `wastewater_state_data` data frame to create 4 new variables:

- 1-week and 2-week lag variables
- for `state_med_conc` and `region_value`
- grouped by `State`

These will be the explanatory variables for the next task.

Task 7

Build a linear regression model with `activity_level` as the response variable and the 4 new variables (from the previous task) along with `State` as the explanatory variables.

- Compute a coefficient of determination for this model
- Interpret the intercept for this model
- Interpret one additional coefficient for this model

Task 8

Using the two most recent weeks of observations, predict the upcoming `activity_level` for every state (of the United States). Display your results as a `gt` table.

Task 9

Expand your regression model with one additional term of your choice.

- Compute a coefficient of determination for this new model
- Be sure to explain how your intuition about the data led you to create this new term.

Exploratory Bioinformatics

Task 10

From the `wastewater_variant_data` data frame, build a bar plot in descending order for the number of weeks where each variant strain was featured in the data set.

- omit the `Other` column

Task 11

We will replicate the visualization found on the NWSS variants page. From the `wastewater_variant_data` data frame

- omit the `Other` column
- build a *stacked bar chart*
- the vertical axis should be in proportions (that is, from zero to one)
- place the legend of the graph below the bar chart
- You do not have to use the same colors

Task 12

Write a conclusion paragraph to summarize your work from the above tasks. Also address the notion of "If you had more time to work on this project, what additional directions would you take the work?"

Wrap Up

When you are done, please upload a copy of the output HTML file back here in Canvas

- also upload a copy of your `.qmd` file there as a attachment in a user comment in Canvas