Processing Large Amounts of Data in R (AKA "The Tidyverse Tutorial")

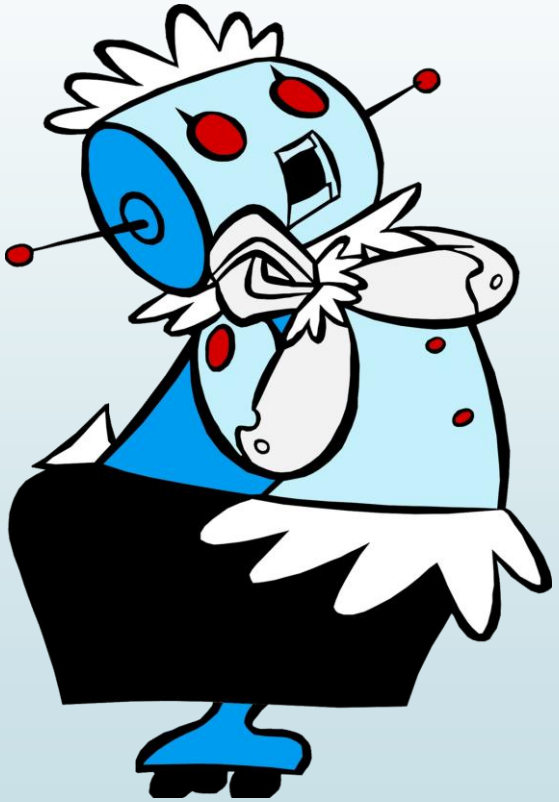
By Travis Oishi

About Me

- Central Shenandoah Health District Epidemiologist
- Jr Epi for Central Shenandoah from Oct 2019 - March 2022
- Limited experience in R when I first started in 2019
- Started coding around Spring of 2020 to improve epi processes
- travis.oishi@vdh.virginia.gov

Benefits of R

- Completing tasks that you need to do more than once
- Processing complex equations
- Navigating large datasets
- Large scale data manipulation
- Less prone to human error



https://hero.fandom.com/wiki/Rosie_The_Robot_Maid

What is Tidyverse?

- Collection of packages designed for data science
- Quality of life improvements
- Packages complement one another

https://bookdown.org/michela_cameletti/rcodingfordatascience/lecture-4-21122020.html

dplyr

- Data manipulation
- List of verbs that help you solve the most common data manipulation challenges
- Simplifies and streamlines codes
- Uses piping, %>% to create a sequence of actions that flow together

https://dplyr.tidyverse.org/

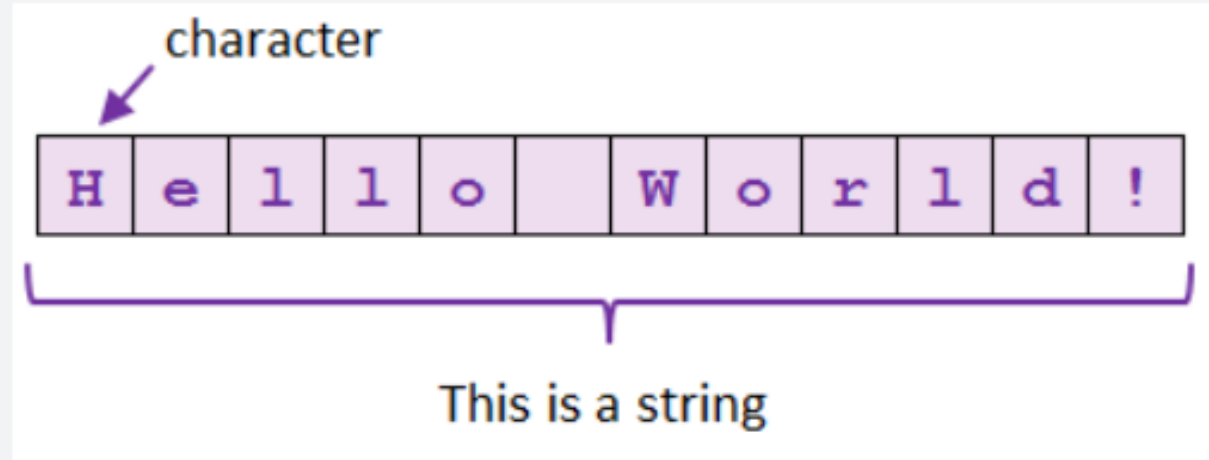
https://www.pixtastock.com/illustration/44869386

Operation	base R example	dplyr function	dplyr example
select some rows	my_data[c(2,3,10),]	slice()	slice(my_data, c(2,3,10))
select some columns	my_data[,1:2] OR my_data[,c("Var_1", "Var_2")]	select()	select(my_data, Var_1, Var_2)
subset	my_data[my_data$Var_2>80,] OR subset(my_data, Var_2>80)	filter()	filter(my_data, Var_2>80)
order the rows	my_data[order(my_data$Var_2),]	arrange()	arrange(my_data, Var_2)
add a column	my_data$logVar_2 <- log(my_data$Var_2) OR transform(my_data, logVar_2=log(Var_2)	mutate()	mutate(my_data, logVar_2 = log(Var_2)
define groups of data	Done within other functions.	group_by()	my_data %>% group_by(Var_3)
summarise the data	aggregate(Var_2 ~ Var_3, data = my_data, FUN = mean) OR tapply(my_data$Var_2, list(my_data$Var_3), mean)	summarise() AND group_by()	my_data %>% group_by(Var_3) %>% meanVar_2 = mean(Var_2)

https://insightsfromdata.io/more-R.html

stringr

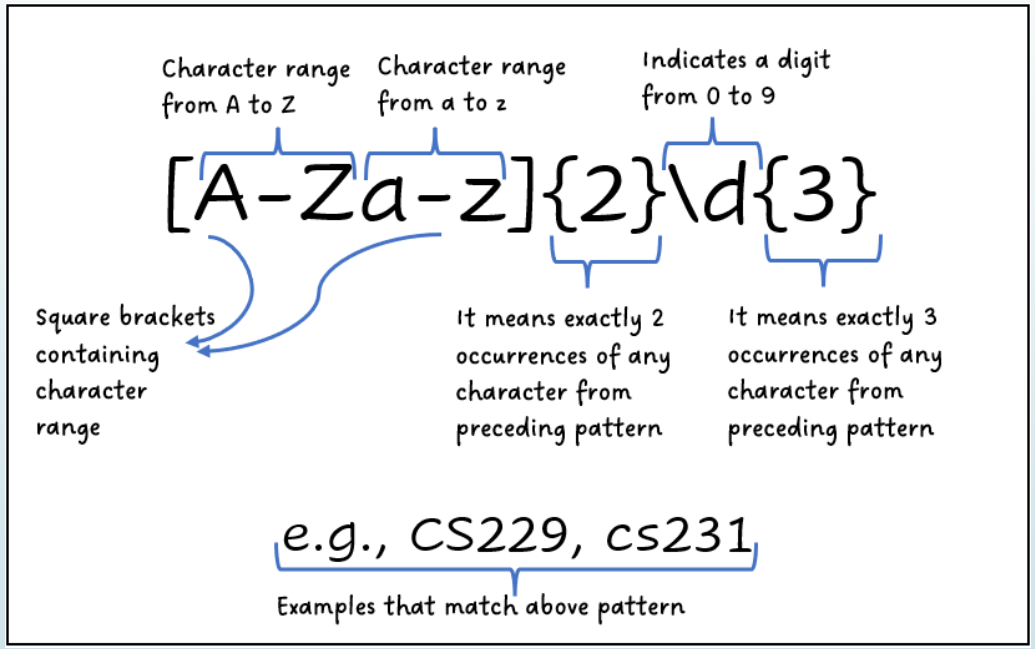
- String identification and manipulation
- Pattern recognition
- Data cleaning
- Use of regular expressions



https://www.geeksforgeeks.org/string-functions-and-operators-in-swift/

Identifying Patterns in Strings: Regular Expressions

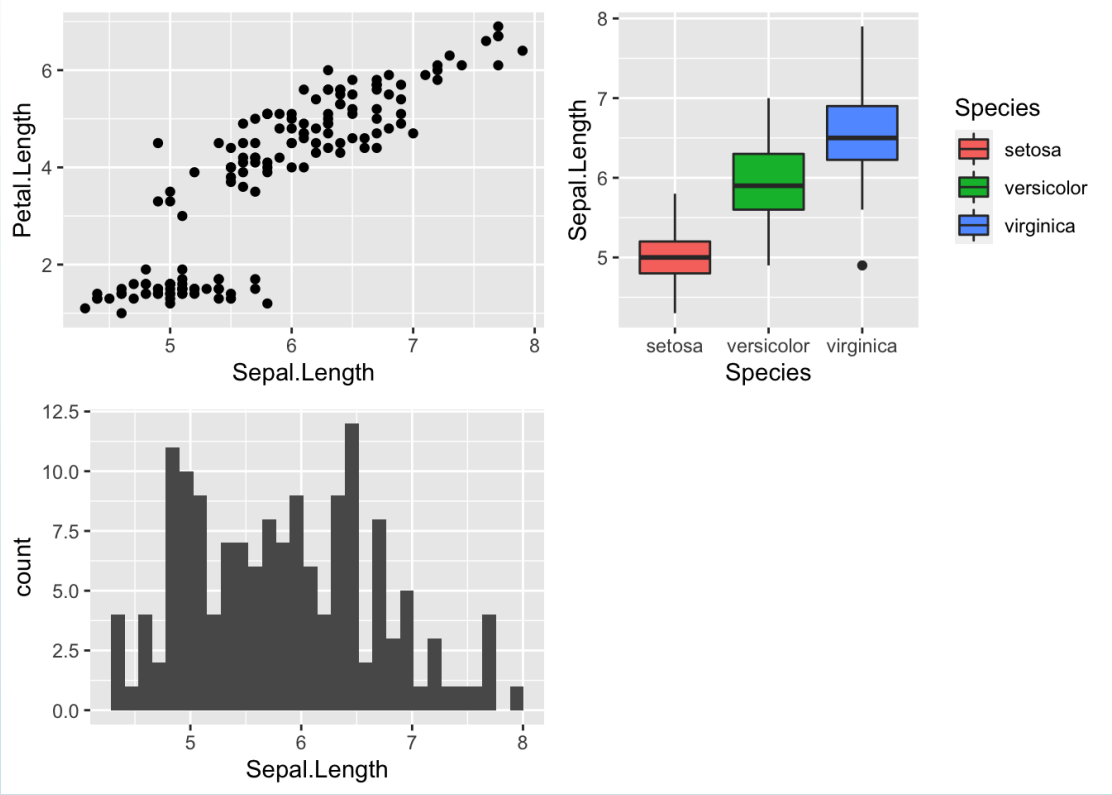
- Method for the computer to process text
- Identifies specific patterns
- Practical uses:
  - Address processing
  - Identifying emails



https://towardsai.net/p/l/regular-expression-regex-in-python-the-basics

ggplot2

- Data visualization
- Control the aesthetics of plots
- Create a variety of plots based on one or more data sources

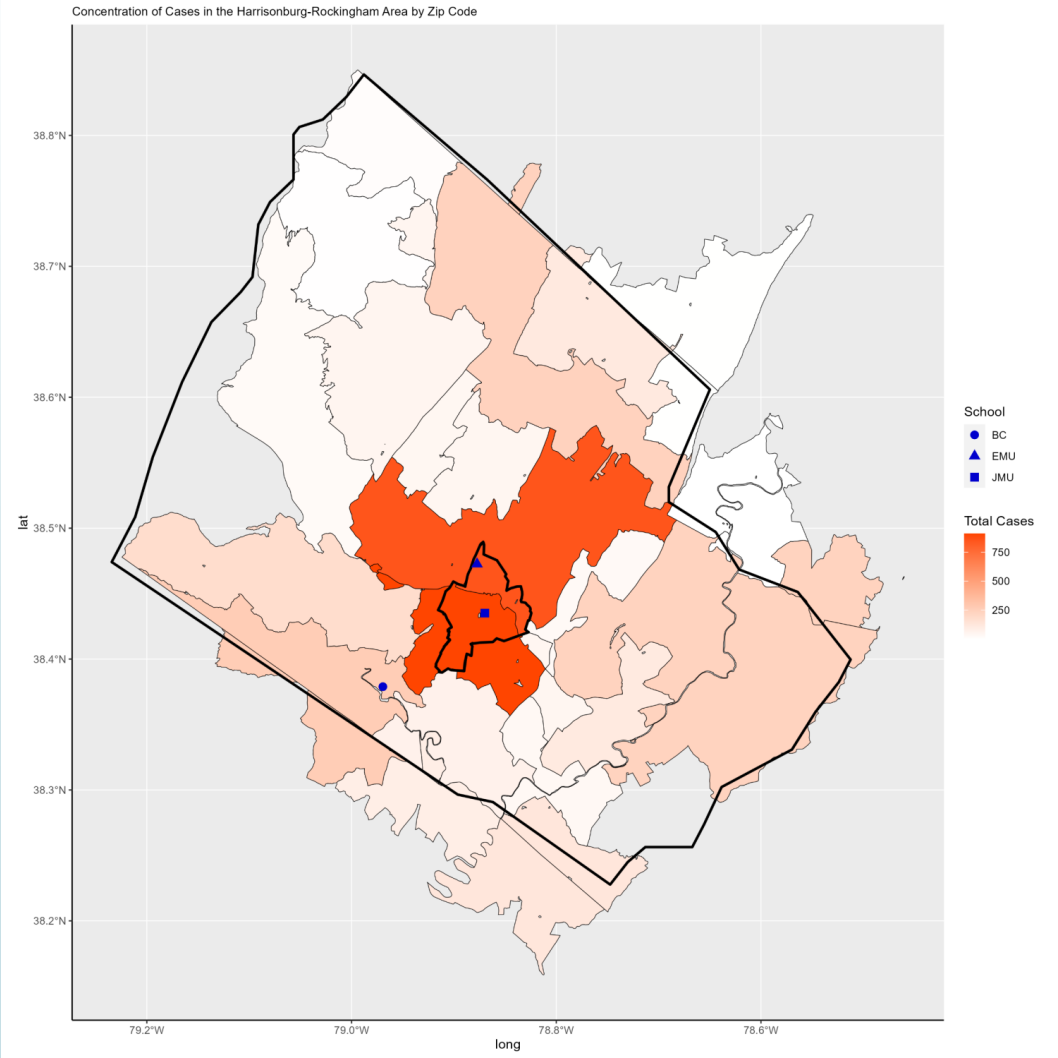


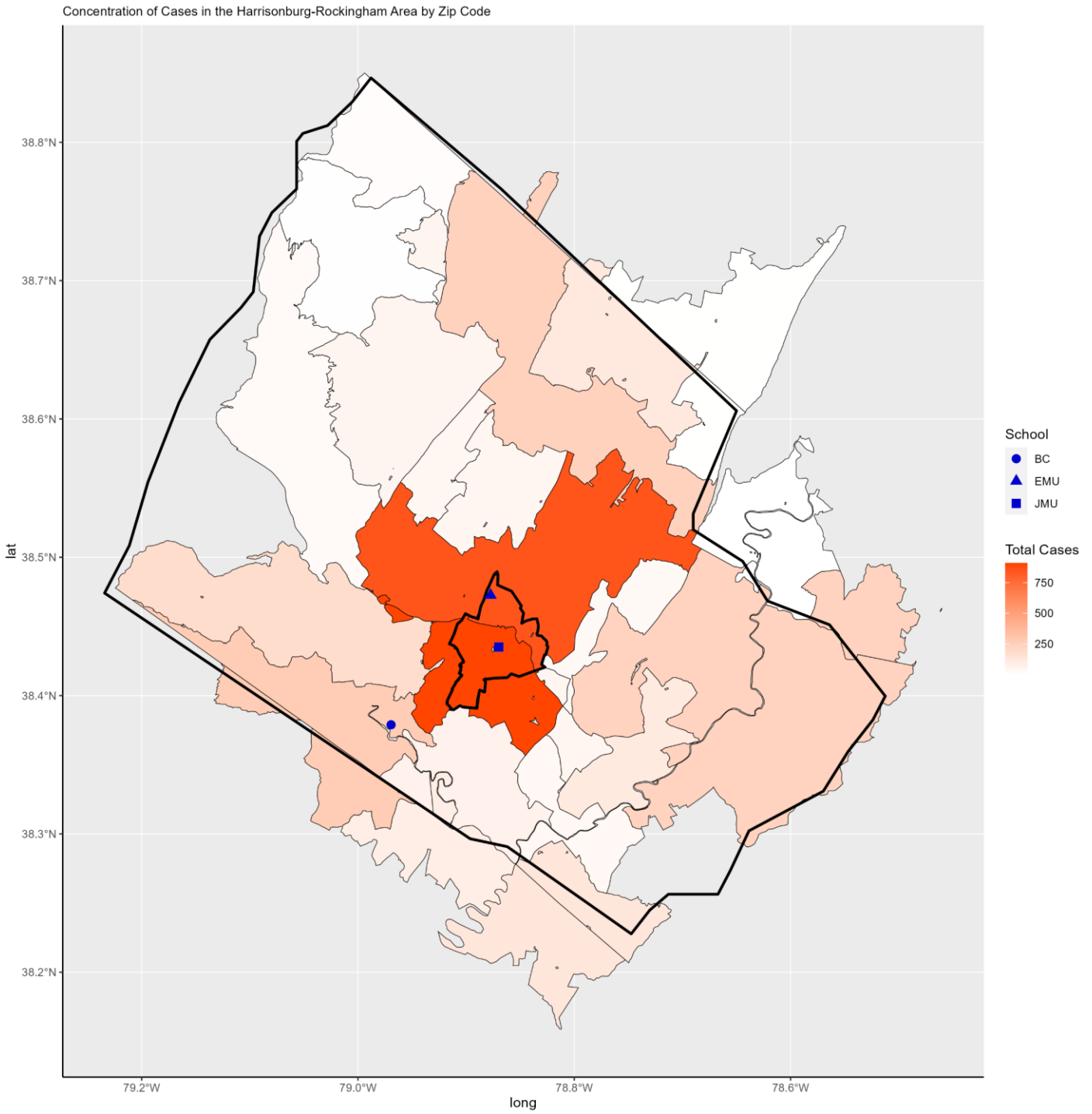
https://environmentalcomputing.net/graphics/ggplot/ggplot-colour-shapes/

https://ggplot2.tidyverse.org/

Tidyverse in Action: Mapping Concentration of Disease Cases

- Identify complete zip codes reported in Rockingham County case reports **(stringr)**
- Summarize the total cases in each zip code **(dplyr)**
- Match zip code data to a zip code map shape file **(dplyr)**
- Plot the map **(ggplot2)**




Concentration of Cases in the Harrisonburg-Rockingham Area by Zip Code
38.8°N
38.7°N
38.6°N
38.5°N
38.4°N
38.3°N
38.2°N
lat
79.2°W
79.0°W
78.8°W
78.6°W
long
School
BC
EMU
JMU
Total Cases
750
500
250

Type in the Chat

- 1: You are familiar with story of *The Odyssey*
- 0: You are not familiar with the story of *The Odyssey*

https://www.shutterstock.com/image-illustration/odysseus-his-companions-stormy-sea-computer-772680487

Learn More

- https://study.com/learn/lesson/tidyverse-packages-examples-r-programming.html#:~:text=Tidyverse%20is%20an%20R%20programming,constantly%20being%20modified%20and%20improved.
- https://www.tidyverse.org/