Lecture 15: In Class Exercise

Reevaluating Lecture 3 Plot

In today's in class activity, we will be going back to an old graphic we created and seeing if we can expand upon the concepts with anything new we have learned about data visualization. Hopefully, the story that we can tell with these graphics will become more robust and easier to understand for a potential viewer without prior knowledge of the data set.

So first things first, we are going to review the COVID data set from Lecture 03:

Example: Colorado Covid Data The following data is downloaded from the Colorado Department of Public Health and Environment[1]. It contains four main variables:

- The date
- The particular utility
- SARS CoV2 copies of RNA (measured as RNA/liter of water) in wastewater
- The number of new Covid-19 cases

SARS-CoV-2 is the virus that causes COVID-19, and RNA is the genetic material in each copy of the virus. SARS-CoV-2 copies per liter is one measure of how much of the virus is in the wastewater, expressed as a concentration. Studies have shown that individuals who develop COVID-19 often shed detectable SARS-CoV-2 RNA from their systems before, during, and after their infection, so higher levels of SARS-CoV-2 RNA can indicate a rise in cases in a community. Many universities used this method to monitor the wastewater from residence halls to obtain an early warning of a disease outbreak.

```
covid <- read.csv(file="dat/CDPHE_COVID19_Wastewater_Dashboard_Data.csv", header=T)
```

From here we do need to repeat some of the data wrangling to get it to a state we can work with a little easier. This is an equivalent tidyverse way to clean the data compared to what we did with base R in unit 1.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

[1] https://cdphe.maps.arcgis.com/apps/opsdashboard/index.html#/d79cf93c3938470ca4bcc4823328946b

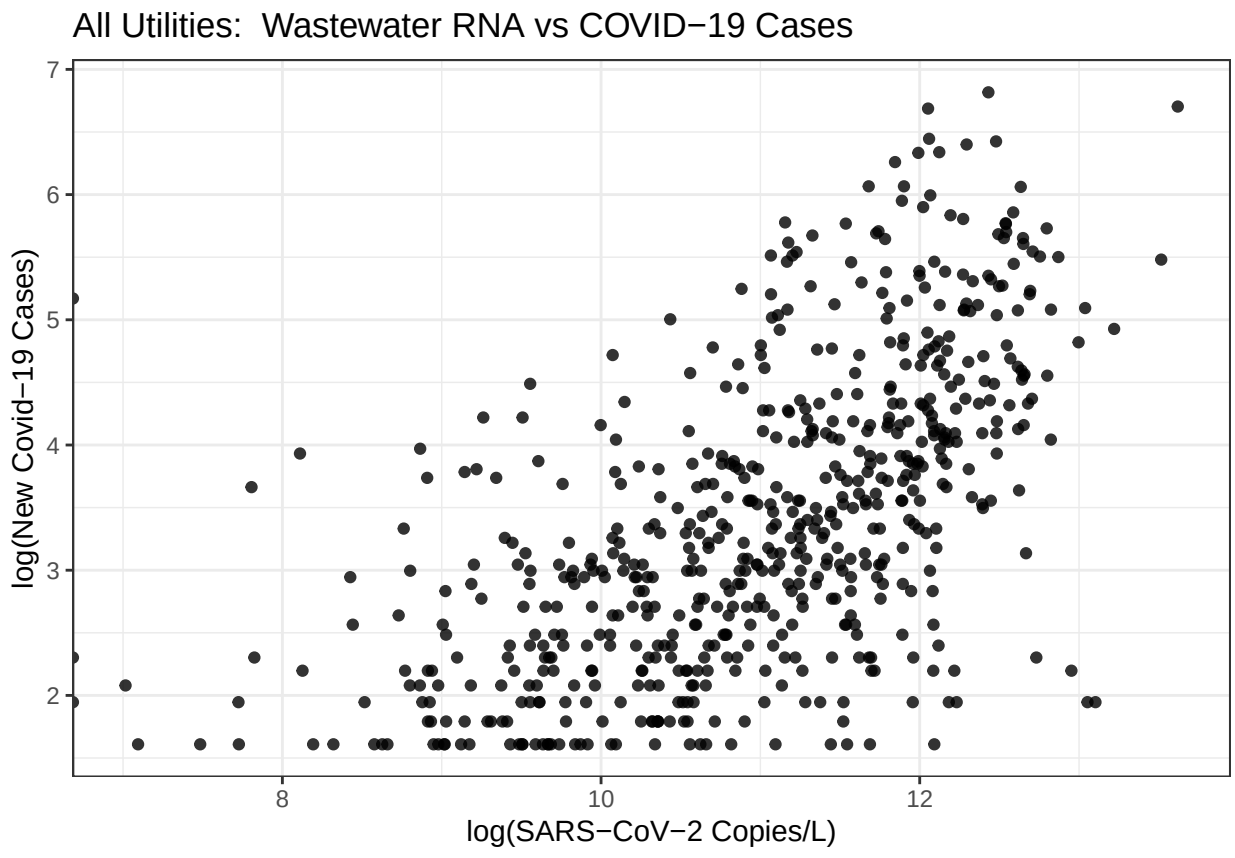
```
colnames(covid)
```

```
## [1] "Date"                                "Utility"
## [3] "SARS_CoV_2_copies_L"                 "Number_of_New_COVID19_Cases_by_"
## [5] "ObjectId"
```

```
covid %>%
  select(-ObjectId) %>%
  rename(sars_rna_copies = SARS_CoV_2_copies_L,
         new_covid_cases = Number_of_New_COVID19_Cases_by_) %>%
  filter(is.na(sars_rna_copies)==FALSE & new_covid_cases>=5)  -> covid_df
```

Finally, recreate the plot with the log of SARS-CoV-2 RNA the x-axis and the log of the number of new cases on the y-axis.

```
covid_df %>%
  mutate(sars_rna_copies = log(sars_rna_copies),
         new_covid_cases = log(new_covid_cases)) %>%
  ggplot(aes(x = sars_rna_copies, y = new_covid_cases)) +
  geom_point(alpha = 0.8) +
  xlab("log(SARS-CoV-2 Copies/L)") +
  ylab("log(New Covid-19 Cases)") +
  ggtitle("All Utilities:  Wastewater RNA vs COVID-19 Cases") +
  theme_bw()
```



```r
#m <- lm(log(covid_df$new_covid_cases)~log(covid_df$sars_rna_copies + 1))

#p + geom_abline(intercept = coef(m)[1], slope = coef(m)[2])
```

Expanding the Plot

After all of that, we can now move into our actual task for the day. You are tasked with taking this old plot of the data and altering it with the expressed purpose of redesigning the graphic to either support or end the funding for wastewater COVID testing.

Challenge: If you would like to explore geospatial coding, the coordinates in the `dat` folder can be used to make a map of all of Colorado. Set `eval = TRUE` to follow along with the in class demo.

```r
library(ggmap)

#ggmap::register_google("")

tot_covid <- as.data.frame(aggregate(new_covid_cases ~ Utility, covid_df, sum))

#create df for geocoding
addr_df <- data.frame(
  Utility = paste(unique(covid$Utility), rep(", CO", 22))) #add any helpful geographic info

#covid_lat_longs <- ggmap::geocode(addr_df$Utility, lat = latitude, lon = longitude)
load("dat/spatialCovid.RData")
covid_lat_longs$Utility <- unique(covid$Utility)

merged_df <- merge(covid_lat_longs, tot_covid, by = "Utility", all = TRUE)
#merged_df <- merge(covid_lat_longs, tot_covid, by.x = "UtilityinLatLon", by.y = "Utlity", all = TRUE)

#co <- qmap("Colorado, USA", zoom = 8, color = "bw")#how to pull map tile


co +
geom_point(aes(x = lon, y = lat, color = 'red', size = new_covid_cases), data = merged_df)
```

Telling Your Story

In our next session, we will be coming back together and sharing our final products to the rest of the class. You should be prepared to speak about the specifics of your graphic, the story you are trying to tell, and the choices you made and why. You should have a fully completed figure at this point.

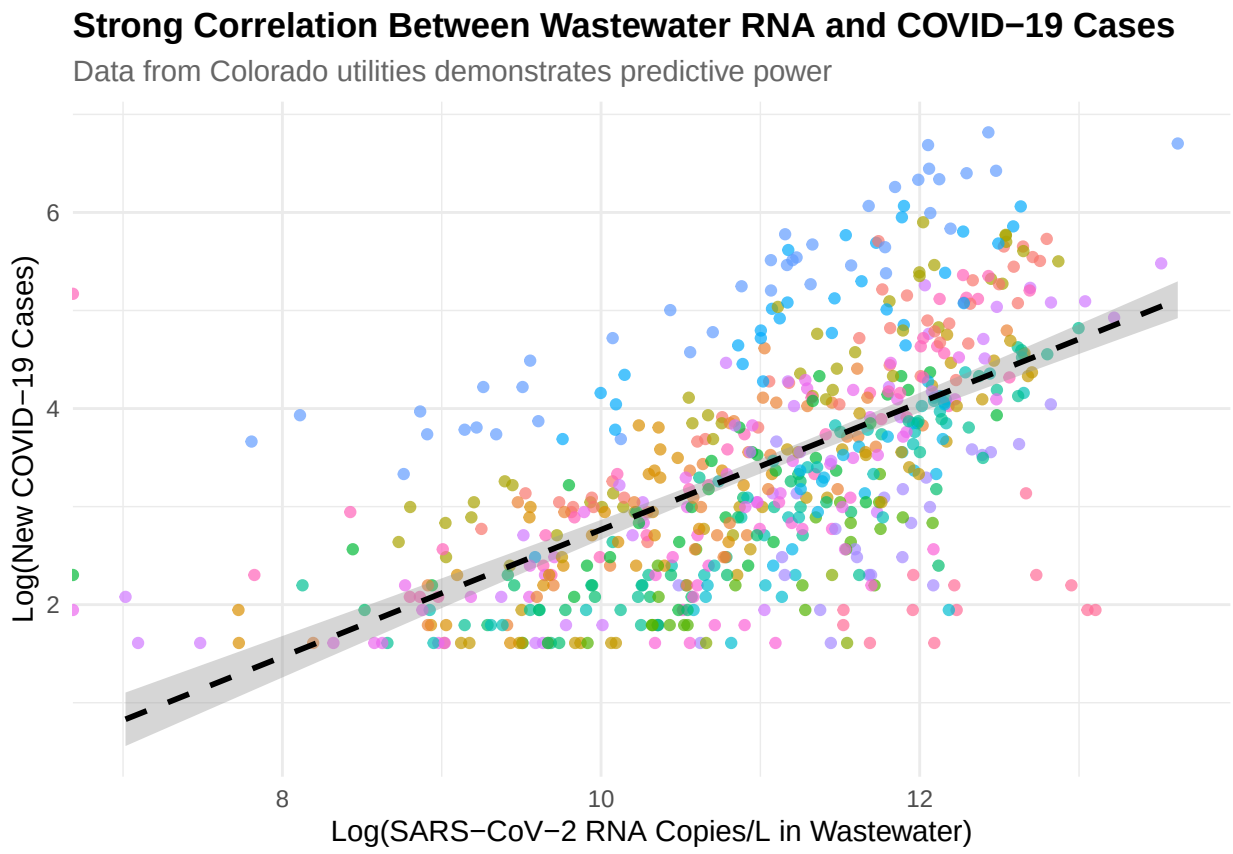
```r
library(tidyverse)

# Create enhanced visualization
covid_df %>%
  mutate(
    log_sars = log(sars_rna_copies),
    log_cases = log(new_covid_cases)
  ) %>%
```

```r
  ggplot(aes(x = log_sars, y = log_cases)) +
  geom_point(aes(color = Utility), alpha = 0.7) +
  geom_smooth(method = "lm", color = "black", linetype = "dashed") +
  labs(
    title = "Strong Correlation Between Wastewater RNA and COVID-19 Cases",
    subtitle = "Data from Colorado utilities demonstrates predictive power",
    x = "Log(SARS-CoV-2 RNA Copies/L in Wastewater)",
    y = "Log(New COVID-19 Cases)",
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    plot.title = element_text(face = "bold"),
    plot.subtitle = element_text(color = "gray40")
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



Submission

For this in class exercise, you will be submitting your final graphic and a short write up about what your graphic is trying to accomplish. Any additional information about your design process, abandoned visual design ideas, or alternative plots are welcomed to be submitted at the end of the document in their own section.

Purpose and Design:

1. Both axes are log-transformed to better display the relationship across different scales and make the pattern more visible.
2. Different utilities are shown in different colors to demonstrate that this relationship holds across various locations.
3. A fitted line shows the overall positive correlation between RNA levels and cases.

Key Findings:

1. There's considerable scatter around the trend line, with many points deviating significantly from it. This suggests the relationship is not as "strong" as I initially claimed.
2. For any given RNA level, there's a wide range of possible case numbers (y-axis values), sometimes varying by several orders of magnitude (since this is on a log scale).
3. The data points show some clustering patterns that might indicate other factors affecting the relationship that weren't accounted for.

Implications for Funding:

1. It demonstrates a clear predictive relationship
2. The method works across different utilities
3. It could provide early warning of COVID-19 surges before clinical cases are reported

Alternative Approaches Considered:

1. Time series plots showing lag between RNA detection and cases
2. Geographic visualization of testing sites
3. Plots by utility size

The current approach was chosen for its clarity in demonstrating the core relationship between wastewater testing and case numbers, which is the most crucial factor in funding decisions.