DataViz

Darren Liang

October 15, 2021

# Data Visualization Assignment  
# For this assignment you will need to find/secure access to a dataset that you think can be used to tell an interesting story through visualization. Possible sources of data sets include: 1) a data set you've already worked with before (e.g., your undergraduate thesis project); 2) an openly available psychology dataset; 3) one of the previous #TidyTuesday datasets (note, you may not "double count" this as a bonus assignment submission); 4) one of the open data sets used by FiveThirtyEight; 5) some other interesting an open data set of your choice (e.g., from a research article with an open data badge). You will need to make 2-4 visualizations (as Wilke notes, it's difficult to tell a story with only one!), and two different "versions" of each visualization. Your submission should be submitted as a fully reproducible workflow that must mindlessly knit to word.   
  
# Note: Part of the correctness points portion of this assignment will draw from the coherence and convincingness of your story, plots, explanations, and justifications. Further, in addition to the normal reproducibility + correctness grading scheme, we will be grading this assignment using a "complexity multiplier" that's designed to incentivize folks who want to go the extra mile (tastefully) incorporating more advanced features. This multiplier will be applied following the tallying of reproducibility + correctness points (but the assignment will still cap out at its allotted 10% of final grade amount).:  
#Advanced: 1.10x  
#Expected: 1.0x  
#Easier: 0.90x

# Planning

# Write a # Planning section in your .Rmd to describe the ## Story you hope to tell about the data you are exploring, and the ## Plots that you think will help to convey that story.  
  
file\_list = list.files(pattern="\*.csv")  
last\_modified\_dt = max(file.info(file\_list)$mtime)  
  
# import spreadsheets of interest  
dose.df <- read\_csv("./vaccine\_doses.csv", show\_col\_types = FALSE)  
case.df <- read\_csv("./covidtesting.csv", show\_col\_types = FALSE)  
  
# reduce dataframe through column select  
dose.reduce <- dose.df %>%   
 select(report\_date, previous\_day\_fully\_vaccinated, total\_individuals\_fully\_vaccinated)  
case.reduce <- case.df %>%   
 select(`Reported Date`,`Confirmed Positive`, `Total Cases`)  
covid.df <- left\_join(dose.reduce, case.reduce, by = c("report\_date" = "Reported Date"))  
covid.daily <- covid.df %>%   
 select(report\_date, previous\_day\_fully\_vaccinated, `Confirmed Positive`)  
covid.total <- covid.df %>%   
 select(report\_date, total\_individuals\_fully\_vaccinated, `Total Cases`)  
  
# pivot dataframes for ggplot  
covid.daily.long <- covid.daily %>%   
 pivot\_longer(!report\_date, names\_to = "Type", values\_to = "Total") %>%   
 mutate(Type = case\_when(Type == "Confirmed Positive" ~ "Case", TRUE ~ "Vaccine"))  
covid.total.long <- covid.total %>%   
 pivot\_longer(!report\_date, names\_to = "Type", values\_to = "Total") %>%   
 mutate(Type = case\_when(Type == "Total Cases" ~ "Case", TRUE ~ "Vaccine"))

The current project takes two open datasets made available by the Government of Ontario published to open Canada regarding COVID-19 Vaccine Data in Ontario and Status of COVID-19 cases in Ontario licensed by the Open Government Licence - Ontario. Both datasets are avaliable for download at <https://open.canada.ca/data/en/dataset/752ce2b7-c15a-4965-a3dc-397bf405e7cc> and <https://open.canada.ca/data/en/dataset/f4f86e54-872d-43f8-8a86-3892fd3cb5e6> respectively. The combined dataset includes the following files: cases\_by\_age\_vac\_status.csv, cases\_by\_vac\_status.csv, covidtesting.csv, daily\_change\_in\_cases\_by\_phu.csv, vac\_status\_hosp\_icu.csv, vaccine\_doses.csv, vaccines\_by\_age.csv, vaccines\_by\_age\_phu.csv last downloaded 2021-10-22 16:30:30 which includes information up until 8pm the previous day.

The files provided by COVID-19 Vaccine Data in Ontario dataset is seperated into six spreadsheets with information regarding daily and total doses administered; individuals with at least one dose; individuals fully vaccinated; total doses given to fully vaccinated individuals; vaccinations by age; percentage of age group; individuals with at least one dose, by public health unit (PHU), by age group; individuals fully vaccinated, by PHU, by age group; COVID-19 cases by status: unvaccinated, partially vaccinated, fully vaccinated; individuals in hospital due to COVID-19 (excluding intensive care unit [ICU]) by status: unvaccinated, partially vaccinated, fully vaccinated; individuals in ICU due to COVID-19 by status: unvaccinated, partially vaccinated, fully vaccinated, unknown; rate of COVID-19 cases per 100,000 by status and age group: unvaccinated, partially vaccinated, fully vaccinated (calculated by dividing the number of cases for a vaccination status, by the total number of people with the same vaccination status and then multiplying by 100,000); rate per 100,000 (7-day average) by status and age group: unvaccinated, partially vaccinated, fully vaccinated (the average rate of COVID-19 cases per 100,000 for each vaccination status for the previous 7 days).

The files provided by the Status of COVID-19 cases in Ontario dataset is separated into two spreadsheets with information regarding daily tests completed; total tests completed; test outcomes; total case outcomes (resolutions and deaths); current tests under investigation; current hospitalizations; current patients in Intensive Care Units (ICUs) due to COVID-related critical Illness; current patients in Intensive Care Units (ICUs) testing positive for COVID-19; current patients in Intensive Care Units (ICUs) no longer testing positive for COVID-19; current patients in Intensive Care Units (ICUs) on ventilators due to COVID-related critical illness; current patients in Intensive Care Units (ICUs) on ventilators testing positive for COVID-19; current patients in Intensive Care Units (ICUs) on ventilators no longer testing positive for COVID-19; Long-Term Care (LTC) resident and worker COVID-19 case and death totals; Variants of Concern case totals; number of new deaths reported (occurred in the last month); number of historical deaths reported (occurred more than one month ago); change in number of cases from previous day by Public Health Unit (PHU).

There is an inconsistency in the first date reported for each of the spreadsheets which also makes data not immediately comparable. From the vaccine dataset, case rates by vaccination status and age group begins logging from 2021-09-13; hospitalizations by vaccination status begins logging from 2021-08-10; cases and rates by vaccination status begins logging from 2021-08-09; COVID-19 Vaccine data by Public Health Unit (PHU) and by age begins logging from 2021-07-26; COVID-19 Vaccine data begins logging from 2020-12-24; COVID-19 Vaccine data by age begins logging from 2020-12-16. From the cases dataset, daily change in cases by PHU begins logging from 2020-03-04; status of COVID-19 cases in Ontario begins logging from 2020-01-26. Data and plots which will be presented will only include data which data is avaliable for each report date.

## Story

The presentation of the COVID-19 vaccine and case rate record to the public has been notiously problematic with even statisitician and research supporting issues regarding improper or poor presentation of the information. A specific concern that made interpretation difficult was that each week that data was presented, the scale may have changed thus graphs could not be directly compared from week to week. With the current cumulation of data, a cumulative presentation of the data will be able to display the vaccination and case rate in a data-driven manner. This method will leverage the reusability and reproducibility aspect of R to focus on data as a whole rather than individual time points. As this dataset is updated (almost) daily, the use of RMarkdown will allow this data presentation to be continually updated (via. knit) with their respective information and plots to include the latest updated dataset that is downloaded.

By plotting the daily and cumulative rates of both vaccine and case count data, we should be able to visualize how the reported cases changes as vaccine rates changes in a short term (daily) manner as well as a long term (cumulative) manner. This report focuses on the vaccine rate of fully vaccinated individuals and the case rate of confirmed positive COVID tests in all PHU of Ontario. It is important to remember that an individual can only be accounted for once in the vaccine rate but could be accounted for one or more times in the case rate due to multiple contractions of the virus.

## Plots

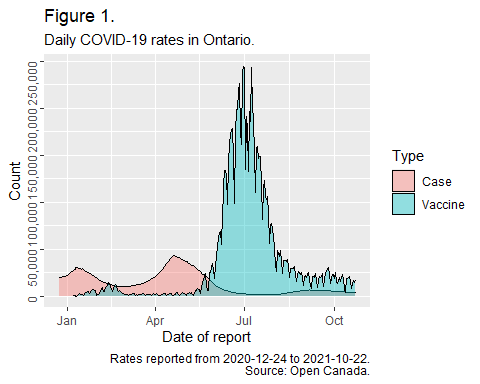
The plots presented displays both the case and vaccination data together to show how the vaccination rate and COVID case rate may be related. It is important to notice the slope or trend of the plots as it is informative about relationship between vaccine and case rate. Vaccine rate does not encompass all the factors related to case rate however a trend for each can show an interesting relationship and a potential ceiling that vaccines have on case rate.

Figures 1 and 3 will present the number of COVID case and number individuals fully vaccinated from the virus per day or the daily count from 2020-12-24 to 2021-10-22.

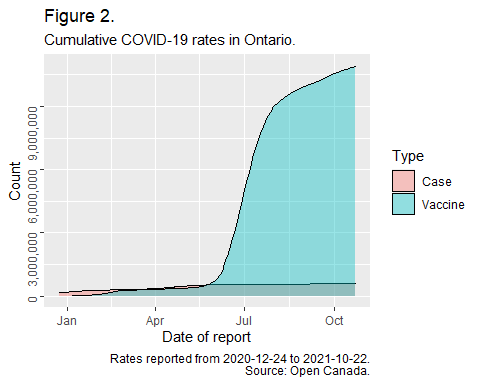
Figures 2 and 4 will present the cumulation of the number of total COVID case and total individuals fully vaccinated from 2020-12-24 to 2021-10-22.

# Publication

# In a # Publication section, create your ggplots styled in a way befitting of being published in a peer-reviewed journal article. Provide a brief ## Summary of what you think each plot tells us, and a ## Justification of the styling/formatting decisions that you made  
  
covid.daily.long %>%   
 ggplot(aes(x = report\_date, y = Total, fill = Type)) +  
 # geom\_bar(stat = "identity", alpha = 0.4) +  
 geom\_density(stat = "identity", alpha = 0.4) +  
 scale\_y\_continuous(name = "Count", labels = scales::comma) +  
 theme(axis.text.y = element\_text(angle = 90)) +  
 labs(x = "Date of report", title = "Figure 1.",  
 subtitle = "Daily COVID-19 rates in Ontario.",  
 caption = paste("Rates reported from ", covid.daily.long$report\_date[[1]], " to ", tail(covid.daily.long$report\_date, n=1), ".\n",  
 "Source: Open Canada.", sep=""))



covid.total.long %>%   
 ggplot(aes(x = report\_date, y = Total, fill = Type)) +  
 geom\_density(stat = "identity", alpha = 0.4) +  
 scale\_y\_continuous(name = "Count", labels = scales::comma) +  
 theme(axis.text.y = element\_text(angle = 90)) +  
 labs(x = "Date of report", title = "Figure 2.",  
 subtitle = "Cumulative COVID-19 rates in Ontario.",  
 caption = paste("Rates reported from ", covid.total.long$report\_date[[1]], " to ", tail(covid.total.long$report\_date, n=1), ".\n",  
 "Source: Open Canada.", sep=""))



## Summary

Figure 1 displays how as more individuals are vaccinated each day, the confirmed positive case rate does tend to decrease signifying a potential inverse relationship or negative correlation. In early 2021, as individuals began to be (fully) vaccinated there was a slight decrease in reported cases however as vaccine rates staggered in late February or early March 2021, the case rate increased. It was until approximately May 2021 which daily case rate reached an all time high but was able to taper as vaccine rate also began to sharply rise. As the daily vaccination rate continued to increase, the case rate continued to decreased and reached an all time low in late July or early August. An increase once again occurred in cases as daily vaccination rate staggered since September but rate for both continue to stay relatively stable.

Figure 2 displays how the total number of COVID cases are compared to total individuals vaccinated. Total vaccination rate was relative low in the first half of 2021 and case rate was relatively uncontrolled. As the vaccinations risen in about Feburary 2021, the trend of case rate slowed. Since approximately June 2021, as vaccinations rates sharply increased, the case rate has been remained stable.

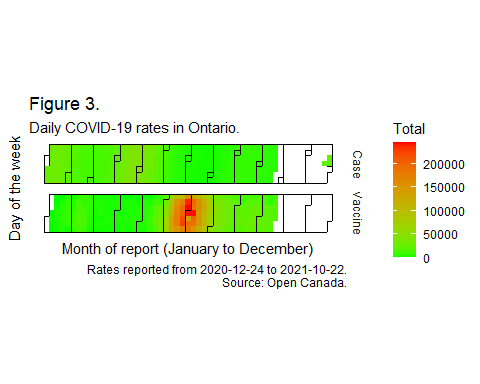
## Justification

For easier comparision between the graphs and since the x-axis remains consistent, both figures (Figure 1 and Figure 2) were made using a density plot with the same styling and formatting.

The density plot was chosen as this is smoother than creating the plots using a bar graph which due to a lack of reported data for either vaccine or case rate during weekends would have created skipping in the bars. The gaps of the dates in the x-axis is smoothed out to show a more normalized pattern. The y-axis required a change in the tick labels due to a large count number in the 100,000 range for Figure 1 and 10,000,000 range for Figure 2. The use of exponential tick labels are relatively difficult to interpret so rotated numerical values allowed the y-axis to take up less space and be more informative. Label choices were made to fill in potential gaps in information such as date and source. Thematic choices were made to be as compliant to APA 7th figure guidelines as possible.

# Dynamic

# In a # Dynamic section, create your ggplots styled in a way befitting of a more "fun" and "dynamic" outlet, such as a blog, tweet, etc., (this is where you should strut your stuff/push your boundaries). Again provide a brief ## Summary of what you think each plot tells us, and a ## Justification of the styling/formatting decisions that you made  
  
covid.daily.long %>%   
 ggplot(aes(date = as.Date(report\_date), fill = Total)) +  
 labs(x = "Month of report (January to December)", y = "Day of the week",  
 title = "Figure 3.", subtitle = "Daily COVID-19 rates in Ontario.",  
 caption = paste("Rates reported from ", covid.daily.long$report\_date[[1]], " to ", tail(covid.daily.long$report\_date, n=1), ".\n",  
 "Source: Open Canada.", sep="")) +  
 facet\_grid(rows = vars(Type))+  
 stat\_calendar\_heatmap()+  
 scale\_fill\_continuous(low = 'green', high = 'red') +  
 theme(axis.text = element\_blank(), axis.ticks = element\_blank(),  
 strip.background = element\_blank(), panel.background = element\_blank(), panel.border = element\_blank() )



covid.total.long %>%   
 ggplot(aes(x = report\_date, y = Total, color = Type)) +  
 scale\_y\_continuous(name = "Count", labels = scales::comma) +  
 theme(axis.text.y = element\_text(angle = 90)) +  
 labs(x = "Date of report", title = "Figure 4.",  
 subtitle = "Cumulative COVID-19 rates in Ontario.",  
 caption = paste("Rates reported from ", covid.total.long$report\_date[[1]], " to ", tail(covid.total.long$report\_date, n=1), ".\n",  
 "Source: Open Canada.", sep="")) +  
 geom\_path() +  
 geom\_point() +  
 transition\_reveal(along = report\_date) +  
 view\_follow()

## Warning in min(x): no non-missing arguments to min; returning Inf

## Warning in max(x): no non-missing arguments to max; returning -Inf

## Warning: Removed 2 row(s) containing missing values (geom\_path).

## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning in min(x): no non-missing arguments to min; returning Inf

## Warning in max(x): no non-missing arguments to max; returning -Inf

## Warning: Removed 2 row(s) containing missing values (geom\_path).

## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning in min(x): no non-missing arguments to min; returning Inf

## Warning in max(x): no non-missing arguments to max; returning -Inf

## Warning: Removed 2 row(s) containing missing values (geom\_path).

## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning in min(x): no non-missing arguments to min; returning Inf

## Warning in max(x): no non-missing arguments to max; returning -Inf

## Warning: Removed 4 row(s) containing missing values (geom\_path).

## Warning: Removed 1 rows containing missing values (geom\_point).

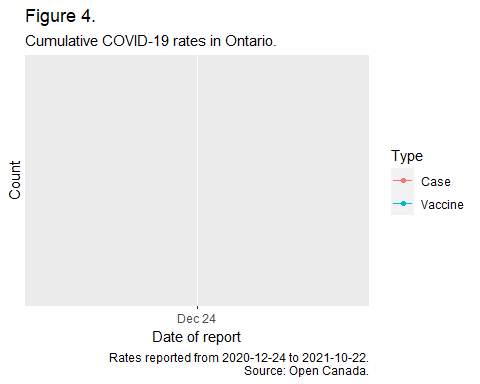
## Warning in min(x): no non-missing arguments to min; returning Inf

## Warning in max(x): no non-missing arguments to max; returning -Inf

## Warning: Removed 7 row(s) containing missing values (geom\_path).

## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).  
  
## Warning: Removed 8 row(s) containing missing values (geom\_path).



## Summary

Figure 3 displays the daily total faceted by COVID cases and number of people fully vaccinated in a calendar. This shows the exact day and month which higher counts occurs for each of the faceted rates. Over the past year, the highest number of reported COVID cases occured in the months of January and April. Over the past year, the greatest number of reported vaccinations occurred in June and July with less but still occurring in February. Note that the month of December currently represents data from December 2020.

Figure 4 displays how the total number of COVID cases are compared to total individuals vaccinated animated over time to follow each rate. COVID cases relatively exceeded maintained higher total numbers compared to vaccine rate by a considerable margin. Until May of 2020, the vaccine rate intercepted and continued to increase the gap between case count and vaccine count.

## Justification

Figure 3 reformatted the data present in Figure 1 in a format which is more specific to the data present. As we are presenting specific day by day information, the calendar heatmap format is more specific about when the case information occurs. This format visualizes specific dates when significant changes occurs in rate and vaccination as well as how gradual or steep the changes are. The rates were faceted to visualize the difference but maintained the same scale so that the data could be directly compared. Thematic choices were made to make it similar to a heatmap with signature heat colors.

Figure 4 reformatted the data present in Figure 2 as a animation to track the rates of COVID cases and vaccination. As we have a cumulative presentation, the tracking of case and vaccine rate over time show a large increase in both factors due to the y-axis movement but after the intercept exxagerates the gap which vaccine rate overtakes the case rate. The following transition draws focus to the scaling and change of rate over time to draw attention in a way that the static version cannot. Thematic choices remained consistent with Figure 2.