**Data Visualization Project: COVID-19 Mobility Reports for Ontario**

Julia Sebastien (216971293)

Kamya Goenka (216611196)

PSYC 3031A

Dr. Monique Herbert

November 4th, 2020

**Description of data**

The original dataset, called mobilityreportCA.csv, was collected by Google via its Location History function. It tracks relative changes in sampled Canadians’ mobility patterns (visiting and lingering) at various kinds of public and private facilities relevant to social distancing each day between February 15, 2020 and September 19, 2020. Sampled users’ mobility pattern data are aggregated by region within Canada (i.e. by province, by district, county, and municipality), facility type (i.e. retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and residential) and calendar date. The positive and negative mobility values reflect percentage increases and decreases in visitations to each kind of facility each day relative to Google’s baseline: users’ relative mobility pattern changes, sampled between Jan 3-Feb 6 2020. All mobility data was sampled from users who opted-in to Location History for their Google Accounts, and only in cases where the Privacy threshold of sampling mobility data was met (meaning only when locations were sufficiently crowded such that recording the user’s location would not identify them). Raw mobility patterns were converted into relative “percent change[s]”, as recorded mobility patterns were compared to patterns pre-recorded during a baseline period before the COVID-19 pandemic: January 3rd to February 6th, 2020.

For our data visualization project, we chose to focus on mobility pattern changes in Ontario for two reasons: 1) it was personally relevant, and 2) it was the province with the most consistently recorded mobility information (despite a number of missing mobility percent change cells in the later dates). Since the specificity of locational information was sometimes inconsistent between rows, we chose not to distinguish between Ontario’s subregions. We also chose to group date-marked rows according to each stage of Ontario’s staged response to COVID-19 to see how mobility patterns might change over the course of changing (or, more specifically, decreasingly stringent) guidelines and restrictions. Lastly, for easy digestion of our graph, we also decided to aggregate our data by condition so that we could present a single mean percent mobility change value for each of the 6x4 combinations of facility type and stage of Ontario’s COVID-19 response. Please see **Appendix** for our script in its entirety. Our data visualization project will thus visualize the average percent change in Ontarians’ mobility patterns in each type of facility (as compared to the recorded baseline) across each stage of Ontario’s response to COVID, starting with the provincial Lockdown, which was from March 11-July 16, and ending with September 19th of Ontario’s Stage 3.

The dataset we used was extracted from the Datasearch platform run by Google (Google LLC. (2020). *Google COVID-19 Community Mobility Reports*. Retrieved from: https://www.google.com/covid19/mobility/), which includes a variety of datasets that are publicly available. The details about our chosen dataset - including the method of tracking, reasons for missing values, privacy of data, and a baseline mobility pattern to compare with are outlined in the provided description for the data.

The timeline for stages of reopening Ontario after the lockdown, which were implemented due to the COVID-19 pandemic, was taken from Ontario’s official website (Archived - Reopening Ontario in Stages. (2020). Retrieved from: https://www.ontario.ca/page/reopening-ontario-stages#section-0). As the dates for each stage outlined by Ontario’s government differed by the sub-regions of Ontario, a rough estimate of the date ranges were used to divide the stages for our dataset.

**Visualization Question**

Our visualization question was as follows: Since the baseline period (January 3-February 6), how have Ontarians’ mobility patterns changed across each type of facility over the course of each stage of the Ontario government’s response to COVID-19: Lockdown (March 11-July 16), Stage 1 (May 19-June11), Stage 2 (June 12-July 16), and Stage 3 (July 16-September 19)?

**Goals/Outcomes**

Our goal with our visualization of our question was to communicate, in a comprehensive and comparable manner, Ontarians’ relative mobility pattern changes in different private and public facilities throughout the staged progression of Ontario’s response to COVID-19. Ideally, we hope our re-framing of these aggregated statistics is more aesthetically pleasing, digestible, and relevant to our fellow Ontarians.

With our data wrangling, we hoped to reframe the dataset into a visualization that would display the results of a factorial-type investigation of the relationship between two independent variables (Stage of COVID-19 in Ontario, and Facility Type) and one dependent variable (Average percent change in Mobility). Our final, modified dataset consists of two categorical independent variables (Stage of Ontario Government’s Response to COVID-19 in Ontario and Facility type), and one continuous dependent variable (Average Percent Change in Mobility).

Our data aimed to make sense of our unwieldy and repetitive dataset about Ontarians’ movement trends during the pandemic’s initial stages. Thus, to visualize our particular question, we re-contextualized the data according to a unified timeline that allows for comparisons between each of the decreasingly stringent social distancing protocols the Ontario government placed upon citizens from Lockdown to Stage 3. Re-contextualizing the data according to Ontario’s response stages was of key interest to us, since we wondered which facilities in particular, if any, would show gradual changes in mobility as the stages of Ontario’s response to COVID-19 went on and the social distancing restrictions were gradually lifted.

**Describing the Layout of the Graph**

In Figure 1 below, the time variable, “Stage of Ontario Government’s Response to COVID-19”, is displayed in chronological order by stage along the x-axis, and the y-axis displays the “Average percent change in Mobility (%)”. The y-axis ranges from -150% to +150% with 0% change in the center, to encapsulate the full range of relative percent changes in mobility patterns with a reasonable amount of room (the highest absolute value of any one datapoint was ~135). Instead of “Percent Change in Mobility”, our y-axis tabulates mobility pattern changes by “Average Percent Change in Mobility (%)”, as all percent change data has been aggregated by stage and facility type. A grey, horizontal line was added at y=0 to show the baseline for easier comparison.

The broken lines and data-points denote the direction of change across stages, distinguishing between “Facility type” by colour, as indicated in the legend. Each point plotted on the graph indicates the average (aggregated) percent change in Ontarians’ mobility patterns for a given facility type during a particular stage of COVID-19. These points were selected to be wide rings to represent that they are averages and represent multiple clustered data points. The line was deliberately made to be dashed because although the masses’ mobility patterns may have changed from stage to stage, there is no confirming whether these aggregated points could be interpellated between each stage; it is also unknown whether the datapoints reflect the mobility patterns of the exact same sampled individuals each time, since it might be possible that different members of the population were sampled on various dates throughout this graph, depending on their Google Maps settings during this time.

A picture containing chart, timeline

Description automatically generated

Figure 1. Average percent changes in Ontarians’ mobility patterns across facility types over the course of Ontario’s multi-staged approach to the COVID-19 pandemic.

**Observable Trends in the graph**

One observable trend which we were expecting was a marked drop from the baseline in mobility for all facility types except residential; it makes sense that residential mobility patterns experienced a boost during Lockdown, as citizens were confined primarily in their homes. As various restrictions gradually lifted across each subsequent stage, it is no wonder that each facility type began to see more mobility, with time spent in residential facilities gradually decreasing accordingly. Interestingly, even by Stage 3, the most relaxed stage, people seem to have been in their homes more than they were on average during the baseline period; this might be due to their need to work and/or attend school remotely, as well as due to enduring restrictions placed on activities requiring large indoor gatherings.

Another foreseeable trend was the drastic rate at which mobility in parks shot up after Lockdown; we see a higher frequency and duration of park visits during the summer months as we proceed into Stage 1 and Stage 2 of the COVID-19 timeline in Ontario. This trend makes sense, of course, since being outdoors was, for many, one of the few acceptable ways to take a break from the confines of their homes, as well as the only way to see their friends while following social distancing guidelines. Additionally, on a non-COVID-19-related note, the weather was beginning to get warmer around this time, and so it makes sense Ontarians would flock to public parks and beaches to take advantage of this, even when congregating in large groups was not recommended.

Another observable trend is that mobility patterns in grocery stores and pharmacies underwent a relatively small change from baseline to the Lockdown period; this was foreseeable as these are facilities that, regardless of COVID-19 stages, are necessary to visit in order to get essential products such as food, medication, and household essentials; furthermore, having being deemed “essential services” from the start, mobility to these facilities was not heavily restricted even during Lockdown. What is slightly surprising is that mobility rates to grocery stores and pharmacies seem to have increased as Ontario’s response stages got more relaxed. At first glance, this trend is slightly curious, as it is not logical to assume that people would suddenly need more food and/or essentials as provincial regulations relaxed somewhat. Plausible interpretations of this increasing trend might be any of the following: 1) this change reflects an increase of the duration of each trip to grocery stores and pharmacies as the population became less nervous, 2) this change reflects an increase of non-essential shopping in pharmacies, 3) households began travelling to grocery stores in larger groups as the stages wore on, or 4) as people adjusted to life during the pandemic, households stopped stockpiling essentials as much, which would necessitate more frequent trips to buy essentials.

On a more minute scale, it seems that mobility patterns at Retail and Recreation, Transit Stations, and Workplace facilities were all similarly reduced during Lockdown, and mobility at all three increased gradually over the course of the subsequent stages. The increases in mobility patterns for all three facility types, of course, could be attributed to the opening up of non-essential services to the public in Stages 2 and Stage 3. However, it is interesting to note that mobility at Retail and Recreation facilities seems to have increased at a faster rate than did the other two, whose increase trends seem almost perfectly parallel to each other. The parallel increases for Workplaces and Transit stations seem to indicate that mobility in these two facility types are directly related (meaning people used transit primarily to go to work). Taken in the context of the greater trend for Retail and Recreation, it seems that perhaps all the time off work during this strange time of social distancing had a silver lining, after all, and people began prioritizing recouping their decreased mental wellbeing over going to work. Of course, this difference might simply suggest that workplaces were increasingly adjusting to new online formats, and using transit to travel to physical workplaces facilities became decreasingly necessary as more people began working from home.

**Limitations**

One of the main limitations of this dataset are the missing Stage 3 values for percentage change in these facility types: parks, grocery and pharmacy, retail and recreation, and transit stations. These missing values were read as N/A by R-studio when the dataset was imported and dropped from the final visualization entirely. These missing values were thus not available for aggregation into Average Percent Change scores by date and facility. Google’s description of its data collection methods suggests that it refrained from sampling data when there was insufficient mobility data such that recording it presented a risk to the anonymity and privacy of sampled individuals. If these data points reflect real people who were left out from the dataset in order to maintain the anonymity and privacy of individuals, this is not only a shame from a data standpoint but also a potential confound; any sudden decrease in mobility patterns would have been especially illuminating had we been able to include it in our visualization. Unfortunately, we could not have replaced low mobility values with ‘0’ to include them, since all mobility pattern scores must be plotted as relative percent changes, not absolutes, and the raw mobility rates during the baseline period were not provided.

Another limitation of this dataset is its lack of specifications for the exact sample size. The mobility pattern dataset we chose consists of individuals in Ontario who opted to share their location history for their Google accounts and allowed it to track their movement while their phone was switched on, during different stages of COVID-19 outlined by Ontario’s government. However, we cannot be sure of how many individuals the data was collected from, as the raw data for people using Google Maps in Ontario with their location history switched on is not publicly available. An important aspect of the sample that we are aware of is that the sample size for all the data collected had to meet the minimum threshold requirement for preserving the privacy of the individuals from whom the data was obtained.

Additionally, as the data was collected by tracking people through the location history function of the Google Maps smartphone application, the data also does not represent members of the population who do not have smartphones and/or who do not have Google Maps running, and/or who have opted out of the location history function. This could bring into question the generalizability and validity of the data, especially for lower income and/or older populations. Similarly, since no data was collected when the mobility was extremely low at any one time, it is difficult to use this data to make sweeping claims about the average citizen’s mobility patterns, as only the movements of the masses, *en masse,* are plotted.

Another potential limitation of the dataset is its accuracy level. While Google Maps is updated often and generally monitors changing landscapes and businesses, it is possible that some visited facilities could have been mislabeled or not recorded at all, especially new facilities and/or facilities that were recently converted from other facility types. For example, it is difficult to know whether a residential building, doubling as an out-of-home business, would or should be recorded under the facility type “residence” rather than “workplace”.

Another potential limitation with our data visualization is due to our own data-wrangling choices; by choosing not to display the mobility percent change data for the period between the baseline and Lockdown (February 7-March 10), our visualization does not provide as stark a mobility contrast as it could between pre-Lockdown and the stages of Ontario’s response to COVID-19, which also might also have been interesting and illuminating to view. Our reasoning for this was that we already have the recorded average mobility patterns for each facility type from the baseline period as our “pre-COVID” measure (against which every subsequent sampling was relatively scored), and the pre-Lockdown period was not part of our Data Visualization question. Even so, one remaining setback from this decision was that it may appear as though the mobility patterns for each facility were at “0” up until May 10th, the day before Lockdown, whereas a visualization of the data before Lockdown might have shown that this was not actually the case, whether due to early concerns about COVID-19 or not.

Another important potential limitation to our data is Google’s choice of using the brief period from January 3-February 6 as the baseline against which all subsequent measures should be relatively scored. Not only did COVID-19 already exist in Ontario during this period, but also it is difficult to ascertain whether this particular period is a reasonable reflection of the population’s typical mobility patterns as Google’s calculations would have us believe. A cursory glance at a Canadian Holidays calendar indicates that certain holidays regularly fall during this period–Orthodox Christmas Day, Orthodox New Year, Chinese New Year, and Groundhog Day–all of which might have driven variable holiday-specific mobility patterns among Ontario’s culturally diverse population that are not necessarily representative of “typical” monthly mobility in Ontario.

It should also be noted that our choice of using means to represent average mobility percent changes has left our data vulnerable to bias; while the trends visible in our final graph seem plausible, tests for normality would be required to determine whether or not the distribution of values per stage was normal or skewed, and these tests would have confirmed for us whether median would have been a less biased, and, arguably, superior, choice to represent the central tendencies for mobility percent change. We chose not to use medians or to test for normality in this project because a) it was largely outside the scope of what we wanted to achieve with this graph, and b) displaying medians can confuse the layperson, who is likely to interpret aggregated results as means anyway.

Finally, this dataset ended on September 19th, 2020, and so it is lacking the latter part of Stage 3; This is a potential issue for two reasons, 1) our dataset is technically incomplete given what it was trying to visualize (as Stage 3 technically lasted until October 14 in certain cities, and even longer in others). 2) From a data visualization standpoint, it also would have been interesting to see how mobility patterns have changed beyond Stage 3, as various cities in Ontario are already regressing to earlier, more stringent stages, since higher infection rates have made a comeback.

**Appendix**

Script

#load libraries into R for help with data wrangling and graphing  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## ── Conflicts ────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(here)

## here() starts at /Users/jsebastien/Desktop/IntStats/R Projects/dataviz

library(readr)  
  
#Importing data for mobility patterns in Canada by 4 stages of the Covid-19 pandemic, and by 6 different facility types.  
mobilitypattern\_CA <- read\_csv(file = here("data/mobilityreportCA.csv"))

## Parsed with column specification:  
## cols(  
## country\_region\_code = col\_character(),  
## country\_region = col\_character(),  
## sub\_region\_1 = col\_character(),  
## sub\_region\_2 = col\_character(),  
## metro\_area = col\_logical(),  
## iso\_3166\_2\_code = col\_character(),  
## census\_fips\_code = col\_logical(),  
## date = col\_date(format = ""),  
## retail\_and\_recreation\_percent\_change\_from\_baseline = col\_double(),  
## grocery\_and\_pharmacy\_percent\_change\_from\_baseline = col\_double(),  
## parks\_percent\_change\_from\_baseline = col\_double(),  
## transit\_stations\_percent\_change\_from\_baseline = col\_double(),  
## workplaces\_percent\_change\_from\_baseline = col\_double(),  
## residential\_percent\_change\_from\_baseline = col\_double()  
## )

View(mobilitypattern\_CA)  
  
#DATA CLEANUP AND TROUBLESHOOTING  
#To be sure our relative percent changes are being perceived of as numerical data, our regions are categorical,  
#and our dates are perceived of as dates (we will not be using the rest of the data), we must check the type of   
#variables R thinks we have by using the spec() function  
spec(mobilitypattern\_CA)

## cols(  
## country\_region\_code = col\_character(),  
## country\_region = col\_character(),  
## sub\_region\_1 = col\_character(),  
## sub\_region\_2 = col\_character(),  
## metro\_area = col\_logical(),  
## iso\_3166\_2\_code = col\_character(),  
## census\_fips\_code = col\_logical(),  
## date = col\_date(format = ""),  
## retail\_and\_recreation\_percent\_change\_from\_baseline = col\_double(),  
## grocery\_and\_pharmacy\_percent\_change\_from\_baseline = col\_double(),  
## parks\_percent\_change\_from\_baseline = col\_double(),  
## transit\_stations\_percent\_change\_from\_baseline = col\_double(),  
## workplaces\_percent\_change\_from\_baseline = col\_double(),  
## residential\_percent\_change\_from\_baseline = col\_double()  
## )

#use the problems function to check if there were any issues importing the data and defining the variable types  
problems(mobilitypattern\_CA)

## [1] row col expected actual   
## <0 rows> (or 0-length row.names)

#ISOLATING ONTARIO IN A NEW DATAFRAME FOR RELEVANT DATA ONLY  
mobilitypattern\_ON <- mobilitypattern\_CA %>%  
 #Removing unwanted rows, and keeping the ones with data on Ontario's mobility patterns.  
 filter(sub\_region\_1 == "Ontario") %>%  
 #Removing unwanted columns, since we are not focusing on the sub-regions of Ontario for our project.  
 select(one\_of("sub\_region\_1", "sub\_region\_2", "date"),  
 contains("change"))  
#removing rows, and keeping stats only for the entire province, after the beginning of lockdown.  
mobilitypattern\_ON <- slice(mobilitypattern\_ON, 26:218) %>%  
 #removing column for sub regions in ontario.  
 select(-"sub\_region\_2")  
View(mobilitypattern\_ON)  
  
#CREATING OUR DATAFRAME BY STAGE OF COVID FOR GRAPHING PURPOSES  
#ADDING A NUMERIC REPRESENTATION OF THE STAGES OF COVID  
#create a vector called covidstages\_ontario with the number coding for our stages of covid (order matters)  
#in our case, we are using "0", "1", "2", and "3" to represent the stages of Ontario's response to COVID   
#(lockdown, stage 1, stage 2 and stage 3, respectively), so it will be: 0, 1, 2, 3.  
#we will use the repeat function to have each value in that sequence repeat according to the number of rows   
#whose date column values fell into each stage of covid; in our case,   
#there were 69 rows in Lockdown, 24 rows in Stage 1, 35 rows in Stage 2 and 65 rows in Stage 3,   
#so this is what our multiplier sequence will contain (order here also matters): 69, 24, 35, 65.  
covidstages\_ontario <- rep(c(0, 1, 2, 3), times = c(69, 24, 35, 65))  
#we will now add covidstages\_ontario as a column to our mobilitypattern\_ON dataframe   
#(this is why order mattered; each value must have exactly the right number of repetitions in order to  
#align with the same row of data from before, when we had individual dates instead of number coding)  
view(covidstages\_ontario)  
#add covidstages\_ontario as a column to our mobilitypattern\_ON dataframe  
mobilitypattern\_ON$covidstages\_ontario = covidstages\_ontario  
view(mobilitypattern\_ON)  
  
#ORGANIZING PERCENT CHANGES AT FACILITY TYPES ACCORDING TO STAGES OF COVID, ONE AT A TIME  
#creating a selection of the dataframe with just Retail and Recreation column and the new covidstages\_ontario column.  
RetailandRecByStage <- mobilitypattern\_ON %>%  
 select(retail\_and\_recreation\_percent\_change\_from\_baseline, covidstages\_ontario)  
#assign the numeric value 1 to code for Retail and Recreation and make it repeat once for each of the 193 rows  
facility\_typeRetailandRec <- rep(1, times = 193)  
#add this coded column to our Retail and Recreation by stage dataframe  
RetailandRecByStage$facility\_typeRetailandRec = facility\_typeRetailandRec  
view(RetailandRecByStage)  
  
#creating a selection of the dataframe with just Grocery Stores and Pharmacies column and the new covidstages\_ontario column.  
GroceryandPharmaByStage <- mobilitypattern\_ON %>%  
 select(grocery\_and\_pharmacy\_percent\_change\_from\_baseline, covidstages\_ontario)  
#assign the numeric value 2 to code for Grocery and Pharma  
facility\_typeGroceryandPharma <- rep(2, times = 193)  
#add this coded column to our Grocery and Pharma by stage dataframe  
GroceryandPharmaByStage$facility\_typeGroceryandPharma = facility\_typeGroceryandPharma  
view(GroceryandPharmaByStage)  
  
#creating a selection of the dataframe with just Parks column and the new covidstages\_ontario column.  
ParksByStage <- mobilitypattern\_ON %>%  
 select(parks\_percent\_change\_from\_baseline, covidstages\_ontario)  
#assign the numeric value 3 to code for Parks  
facility\_typeParks <- rep(3, times = 193)  
#add this coded column to our Parks by stage dataframe  
ParksByStage$facility\_typeParks = facility\_typeParks  
view(ParksByStage)  
  
#creating a selection of the dataframe with just Transit and Stations column and the new covidstages\_ontario column.  
TransitByStage <- mobilitypattern\_ON %>%  
 select(transit\_stations\_percent\_change\_from\_baseline, covidstages\_ontario)  
#assign the numeric value 4 to code for Transit  
facility\_typeTransit <- rep(4, times = 193)  
#add this coded column to our Transit by stage dataframe  
TransitByStage$facility\_typeTransit = facility\_typeTransit  
view(TransitByStage)  
  
#creating a selection of the dataframe with just workplaces column and the new covidstages\_ontario column.  
WorkByStage <- mobilitypattern\_ON %>%  
 select(workplaces\_percent\_change\_from\_baseline, covidstages\_ontario)  
#assign the numeric value 5 to code for Workplaces  
facility\_typeWork <- rep(5, times = 193)  
#add this coded column to our Workplaces by stage dataframe  
WorkByStage$facility\_typeWork = facility\_typeWork  
view(WorkByStage)  
  
#creating a selection of the dataframe with just residential column and the new covidstages\_ontario column.  
RezByStage <- mobilitypattern\_ON %>%  
 select(residential\_percent\_change\_from\_baseline, covidstages\_ontario)  
#assign the numeric value 6 to code for Residential  
facility\_typeRez <- rep(6, times = 193)  
#add this coded column to our Residential by stage dataframe  
RezByStage$facility\_typeRez = facility\_typeRez  
view(RezByStage)  
  
#STACKING ALL INDIVIDUAL PERCENT CHANGES BY COVID STAGE AND FACILITY TYPE  
#in this stage, we will stack all our individual dataframe columns before combining them into one, long, comprehensive tibble.  
#note: it is very important we combine data from facilities in the same order each time (from 1 to 6) as we will be stacking them.  
#STACKING COVID STAGES COLUMNS  
covidstage\_ontario <- c(RetailandRecByStage$covidstages\_ontario, GroceryandPharmaByStage$covidstages\_ontario, ParksByStage$covidstages\_ontario, TransitByStage$covidstages\_ontario, WorkByStage$covidstages\_ontario, RezByStage$covidstages\_ontario)  
#converting this into a tibble since this list is over 1000 rows long   
as\_tibble(covidstage\_ontario)

## # A tibble: 1,158 x 1  
## value  
## <dbl>  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## 7 0  
## 8 0  
## 9 0  
## 10 0  
## # … with 1,148 more rows

view(covidstage\_ontario)  
#STACKING FACILITY TYPE COLUMNS  
#note: it is very important we combine data from facilities in the same order each time (from 1 to 6) as we will be stacking them.  
facility\_type <- c(RetailandRecByStage$facility\_typeRetailandRec, GroceryandPharmaByStage$facility\_typeGroceryandPharma, ParksByStage$facility\_typeParks, TransitByStage$facility\_typeTransit, WorkByStage$facility\_typeWork, RezByStage$facility\_typeRez)  
#converting this into a tibble since this list is over 1000 rows long   
as\_tibble(facility\_type)

## # A tibble: 1,158 x 1  
## value  
## <dbl>  
## 1 1  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 1  
## 7 1  
## 8 1  
## 9 1  
## 10 1  
## # … with 1,148 more rows

view(facility\_type)  
#STACKING PERCENT CHANGE COLUMNS  
#note: it is very important we combine data from facilities in the same order each time (from 1 to 6) as we will be stacking them.  
percent\_change <- c(RetailandRecByStage$retail\_and\_recreation\_percent\_change\_from\_baseline, GroceryandPharmaByStage$grocery\_and\_pharmacy\_percent\_change\_from\_baseline, ParksByStage$parks\_percent\_change\_from\_baseline, TransitByStage$transit\_stations\_percent\_change\_from\_baseline, WorkByStage$workplaces\_percent\_change\_from\_baseline, RezByStage$residential\_percent\_change\_from\_baseline)  
#converting this into a tibble since this list is over 1000 rows long   
as\_tibble(percent\_change)

## # A tibble: 1,158 x 1  
## value  
## <dbl>  
## 1 3  
## 2 6  
## 3 -3  
## 4 -10  
## 5 -10  
## 6 -12  
## 7 -29  
## 8 -36  
## 9 -38  
## 10 -45  
## # … with 1,148 more rows

view(percent\_change)  
  
#COMBINING OUR STACKED COLUMNS INTO A TIBBLE  
#the next step is to bind these three long columns in order, side by side, in a new tibble called rawstackeddata  
rawstackeddata <- data.frame(covidstage\_ontario, facility\_type, percent\_change)  
as\_tibble(rawstackeddata)

## # A tibble: 1,158 x 3  
## covidstage\_ontario facility\_type percent\_change  
## <dbl> <dbl> <dbl>  
## 1 0 1 3  
## 2 0 1 6  
## 3 0 1 -3  
## 4 0 1 -10  
## 5 0 1 -10  
## 6 0 1 -12  
## 7 0 1 -29  
## 8 0 1 -36  
## 9 0 1 -38  
## 10 0 1 -45  
## # … with 1,148 more rows

#SPECIFYING OUR VARIABLES ACCORDING TO VARIABLE TYPE  
#specify appropriate variable type for each variable with the typeof function and save the recoded  
#version of variables into new dataframe called rawdatarecoded  
rawdatarecoded <- rawstackeddata %>%  
 #use mutate to change our dataframe and save it into rawdatarecoded; in this  
#case we use mutate to specify that although levels of the gender variable were originally denoted  
#with a binary numerical system, levels are not continuous and numeric, but  
#rather divided into distinct categories (and is thus a categorical variable) that are best  
#represented with character strings (which we will now rename from numbers to "male" and "female").  
 mutate(covidstage\_ontario = as.character(covidstage\_ontario),  
 #tell R to recode gender as a factor with levels  
 covidstage\_ontario = fct\_recode(covidstage\_ontario,  
 #specify that the character string "Lockdown (March 11-May 18)" should overwrite   
 #any "0" in the "covidstage\_ontario" column as it is the true level name  
 "Lockdown (March 11-May 18)" = "0",  
 #specify that "Stage 1 (May 19-June 11)" should overwrite any "1" in this column  
 #as it is the true level name. Use the "then" function (%>%) to continue  
 #using mutate for another variable.  
 "Stage 1 (May 19-June 11)" = "1",  
 #specify that "Stage 2 (June 12-July 16)" should overwrite any "2" in this column  
 "Stage 2 (June 12-July 16)" = "2",  
 #specify that "Stage 3 (July 17-September 19)" should overwrite any "3" in this column  
 "Stage 3 (July 17-September 19)" = "3")) %>%  
 #specify that facility\_type is also a categorical variable, as each number value represents a type of facility.   
 #The levels of this categorical variable are thus also best represented with character strings.  
 mutate(facility\_type = as.character(facility\_type),  
 #tell R to recode target as a factor with levels  
 facility\_type = fct\_recode(facility\_type,  
 #specify that "Retail and Recreation" should overwrite any "1" in this  
 #column as the name of this level  
 "Retail and Recreation" = "1",  
 #specify that "Grocery and Pharmacy" should overwrite any "2" in this  
 #column as the name of this level, etc  
 "Grocery and Pharmacy" = "2",  
 "Parks" = "3",  
 "Transit Stations" = "4",  
 "Workplaces" = "5",  
 "Residential" = "6")) %>%  
 #specify that percent\_change is a numeric variable.  
 mutate(percent\_change = as.numeric(percent\_change))  
view(rawdatarecoded)  
  
#DATA CLEANUP: CONFIRMING OUR LEVELS  
#just to confirm that we added the levels of our variables properly, we can use the levels function to  
#check what the valid levels are in each variable  
levels(rawdatarecoded$covidstage\_ontario)

## [1] "Lockdown (March 11-May 18)" "Stage 1 (May 19-June 11)"   
## [3] "Stage 2 (June 12-July 16)" "Stage 3 (July 17-September 19)"

levels(rawdatarecoded$facility\_type)

## [1] "Retail and Recreation" "Grocery and Pharmacy" "Parks"   
## [4] "Transit Stations" "Workplaces" "Residential"

levels(rawdatarecoded$percent\_change)

## NULL

#we can then use the pipe function to add to each variable a length function to count the number  
#of valid levels in the factor.  
levels(rawdatarecoded$covidstage\_ontario) %>%  
 length()

## [1] 4

levels(rawdatarecoded$facility\_type) %>%  
 length()

## [1] 6

#this should be null as it is a continuous variable  
levels(rawdatarecoded$percent\_change) %>%  
 length()

## [1] 0

#DESCRIPTIVE DATA OF AVG PERCENT CHANGE PER FACILITY AND PER STAGE OF COVID  
#If we wanted to get a sense of how many conditions we have, and the sample size of each of these conditions for later if we want, use the group\_by  
#function to group the data into the cells and then count how many people are in each condition.  
#this is what we would do:  
#rawdatarecoded %>%  
 #group\_by(covidstage\_ontario,facility\_type) %>%  
 #count()   
  
#INTERACTION EFFECTS BETWEEN GENDER X TARGET  
#If we wanted to get a better understanding of our data as it pertains to our research question,  
#we could use the describeBy function to split the data into each condition (or cell)   
#and then calculate descriptive data for each condition.  
#To do so, you would set the descriptive statistics to focus on the numeric variable, percent\_change, using the $  
#symbol to denote the column name, and assign it to a new object, intfx, like this:  
#intfx <- describeBy(x = rawdatarecoded$percent\_change,  
 ##here we are going to wrap the two grouping variables (covidstage\_ontario,facility\_type) inside the list  
 ##function to create lists of descriptive statistics.  
 #group = list(rawdatarecoded$covidstage\_ontario,  
 #rawdatarecoded$facility\_type))  
 #you could see this data by entering: view(intfx)  
  
#AGGREGATING DATA BY COVID STAGE AND FACILITY TYPE  
#In our case, the only descriptives called for by our research question are the average percent change scores for each combination  
#of covid stage and facility type (4 x 6 = 24).  
#Use the aggregate function to create a new version of our rawdatarecoded dataframe that contains only the mean   
#percent change value, as determined by the distinct conditions that can be made with covidstages and facility types.  
#With this information, we can plot the average percent change value for each stage and facility type.   
#Assign the outcome to a new dataframe called aggregatecoviddata.  
aggregatecoviddata <- aggregate(x = rawdatarecoded$percent\_change,  
 by = list(rawdatarecoded$covidstage\_ontario, facility\_type),  
 FUN = mean)  
  
#RENAME AGGREGATED COLUMNS  
#figure out what our current column names are  
colnames(aggregatecoviddata)

## [1] "Group.1" "Group.2" "x"

# Rename column where names is "Group.1"  
names(aggregatecoviddata)[names(aggregatecoviddata) == "Group.1"] <- "covidstages\_ontario"  
# Rename column where names is "Group.2"  
names(aggregatecoviddata)[names(aggregatecoviddata) == "Group.2"] <- "facility\_type"  
# Rename column where names is "x"  
names(aggregatecoviddata)[names(aggregatecoviddata) == "x"] <- "avg\_percent\_change"  
  
#SPECIFY VARIABLES ONCE MORE IN AGGREGATE VERSION JUST IN CASE BECAUSE FACILITY\_TYPE WAS REVERTED  
#specify once more that facility\_type is a categorical variable, as each number value represents a type of facility.   
#The levels of this categorical variable are thus also best represented with character strings.  
aggregatecoviddatarecode <- aggregatecoviddata %>%  
 mutate(facility\_type = as.character(facility\_type),  
 #tell R to recode target as a factor with levels  
 facility\_type = fct\_recode(facility\_type,  
 #specify that "Retail and Recreation" should overwrite any "1" in this  
 #column as the name of this level  
 "Retail and Recreation" = "1",  
 #specify that "Grocery and Pharmacy" should overwrite any "2" in this  
 #column as the name of this level, etc  
 "Grocery and Pharmacy" = "2",  
 "Parks" = "3",  
 "Transit Stations" = "4",  
 "Workplaces" = "5",  
 "Residential" = "6"))  
#this is what we can graph to showcase our data visualization question  
view(aggregatecoviddatarecode)  
  
#CREATING THE GGPLOT#  
#change axis text style to have italicized, bolded, and grey labels  
grey.bold.italic.10.text <- element\_text(face = "bold.italic", color = "#939799", size = 10)  
##use ggplot package to plot the data in our tibble called aggregatecoviddatarecode, using a set of aesthetic rules whereby x-values are plotted   
#according to the 4 levels of our covidstages\_ontario categorical factor, and y-values are plotted according to the average percent   
#change rates associated with our numeric avg\_percent\_change variable. An aesthetic rule is applied to these lines to map facility\_type   
#to colour such that facility types are groups and a different coloured line is used for each facility type. This way, we can distinguish   
#between facility types in terms of how visitation rates to each type has changed over each stage of covid in Ontario.  
#The + sign is used to add on additional graphing-related design features.  
ggplot(data = aggregatecoviddatarecode,aes(x=covidstages\_ontario, y=avg\_percent\_change, group=facility\_type, colour=facility\_type)) +  
 #geom\_line is then used to create a broken line graph joining the average\_percent\_change scores across each stage of covid  
 #To further customize geom\_line(), add these specifications as desired for a dashed line type and light thickness:  
 geom\_line(linetype="dashed", size=1) +  
 #Add points to the line graph using geom\_point and change the appearance of the points by specifying the size, shape and fill colour of the points;   
 #if we want colours to vary by group colour, add this specification:  
 geom\_point(size=8, shape=21, stroke=3, fill="white") +  
 #Define range of Y from -150 to +150 to show entire range of possible relative percent change without bias  
 ylim(-150, 150) +  
 #set the labels of the title and later the x and y labels using labs.  
 labs(title = "Ontarians’ Mobility Pattern changes across Facility Types during Stages of Ontario Government’s Response to COVID-19",   
 #write out the x-axis label in full  
 x = "Stage of Ontario Government's Response to COVID-19",  
 #write out the y-axis label in full  
 y = "Average Percent Change Relative to Baseline (%)",  
 #title the legend  
 colour = "Facility Type") +   
 #add a layer to customize the theme to something more APA-appropriate such as a plain, white background   
 theme() +   
 #apply the classic ggplot graphing theme, which has x and y axes, a plain white background, and no gridlines.  
 theme\_classic() +  
 #add a horizontal line at y=0 to show the baseline  
 geom\_hline(yintercept=0, linetype="solid", color = "grey", size = 1) +  
 #apply x-axis theme to both x and y axes  
 theme(axis.text = grey.bold.italic.10.text)

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

## Warning: Removed 4 rows containing missing values (geom\_point).