Data translation challange

Data Cleaning

1. Loading Libraries

```
library(fixest)
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.1     ✓ tibble    3.2.1
✓ lubridate 1.9.2     ✓ tidyr     1.3.0
✓ purrr     1.0.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(rio)
library(lubridate)
library(rdrobust)
library(vtable)
```

```
Loading required package: kableExtra

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
library(ggplot2)
library(stringr)
library(dplyr)
library(ipumsr)
```

2. Loading Datasets

```
ddi <- read_ipums_ddi("DDI codebook.xml")
data <- read_ipums_micro(ddi)
```

```
Use of data from IPUMS CPS is subject to conditions including that users should cite the data
appropriately. Use command `ipums_conditions()` for more details.
```

```
#CPS data (by_month)
cps_data <- data %>%
  filter(is.na(ASECFLAG))
```

Creating Variables

```
cps_data <- cps_data %>%
  mutate(RetailEmployment = 5790 < IND & IND > 4670,       # RetailEmployment: TRUE if employed
in Retail, FALSE ifelse
         date = ymd(paste(YEAR, MONTH, "-01", sep = '-')))  # date: YYYY-MM-DD (useable format for
date)
```

Data Cleaning

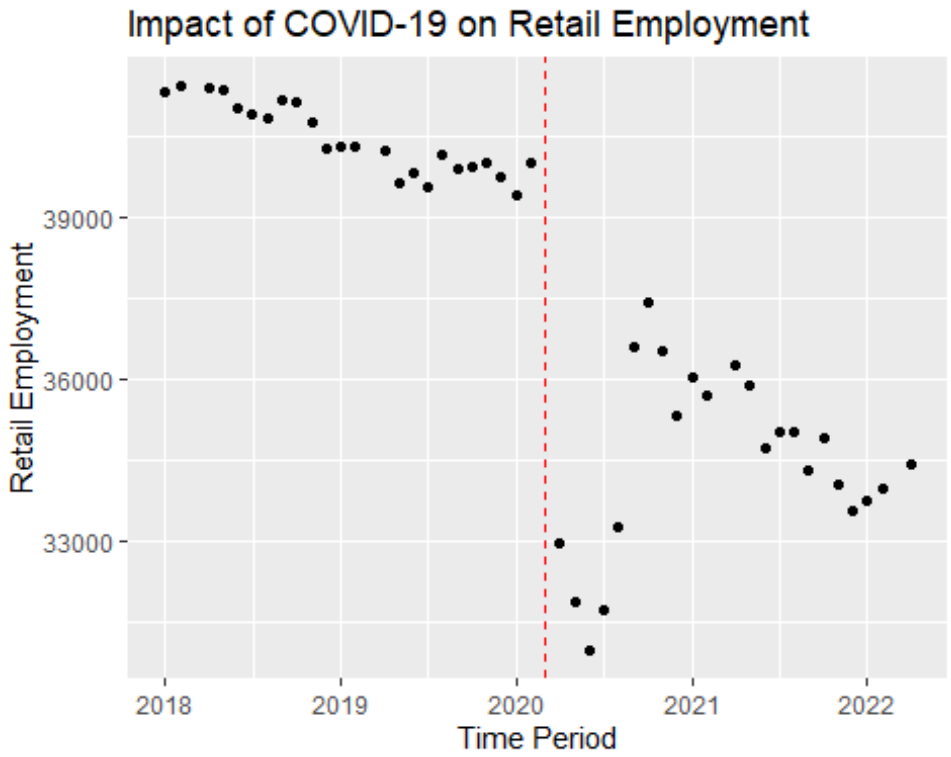
```
#get rid of everything before 2018 and after apr 2022
cps_data <- cps_data %>%
  filter(date >= ymd("2017-01-01") & date <= ymd("2022-04-01"))
```

Analysis Questions

1. **How has COVID affected the health of the retail industry, as measured by employment?**

The goal is to find some effect that COVID had on the retail industry, and perform an analysis on the impacts of COVID. Before building a regression model, it might be helpful to visualize the effect COVID had on the retail industry by looking at the total employment in the retail industry by month. We can see in this graph, that before COVID, there was high employment in the retail industry. Right around the time COVID happened, in March 2020, Retail employment dropped significantly. A couple months later employment in the retail industry recovered, but employment was not nearly as high as before COVID occurred.

```
#Retail Employment by Month
data_agg <- cps_data %>%
  group_by(date) %>%
  summarize(total_retail = sum(RetailEmployment))
# Graph
ggplot(data_agg, aes(x = date, y = total_retail)) +
  geom_point() +
  geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), linetype = "dashed", color = "red") +
  labs(title = "Impact of COVID-19 on Retail Employment",
       x = "Time Period",
       y = "Retail Employment")
```



Based on this graph we can find a 'COVID effect' by conducting an Interuppted Time Series model, in doing so assessing a continuous variable (time) with a cut off period during COVID (March 2020). First we must first center the date around March 2020, in other words center our data around the cut off period, so we can determine whether or not COVID independently had an impact on the retail industry. We will then run a interrupted time series model to show what COVID's effect was on the retail industry, measuring the impact on the retail industry by amout of employment.

Interrupted Time Series Model (Regression Discontinuity)

```
# dateCentered: center the date around March 2020
cps_data <- cps_data %>%
  mutate(dateCentered = as.numeric(interval(ymd("2020-03-01"), date) / months(1)))  # Assigned: any
date after March 2020 (COVID had happened)
cps_data <- cps_data %>%
  mutate(Assigned = dateCentered > 0)
# Regression Discontinuity Model
m1 <- feols(RetailEmployment ~ Assigned*dateCentered, data = cps_data, vcov = 'hetero')
etable(m1)

                                  m1
Dependent Var.:           RetailEmployment

Constant                  0.3386*** (0.0006)
AssignedTRUE               -0.0112*** (0.0008)
dateCentered              0.0003*** (3.7e-5)
AssignedTRUE x dateCentered -0.0002** (5.67e-5)
```

```
_____ _____
S.E. type                Heteroskedast.-rob.
Observations                       5,314,036
R2                                  5.32e-5
Adj. R2                             5.27e-5
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

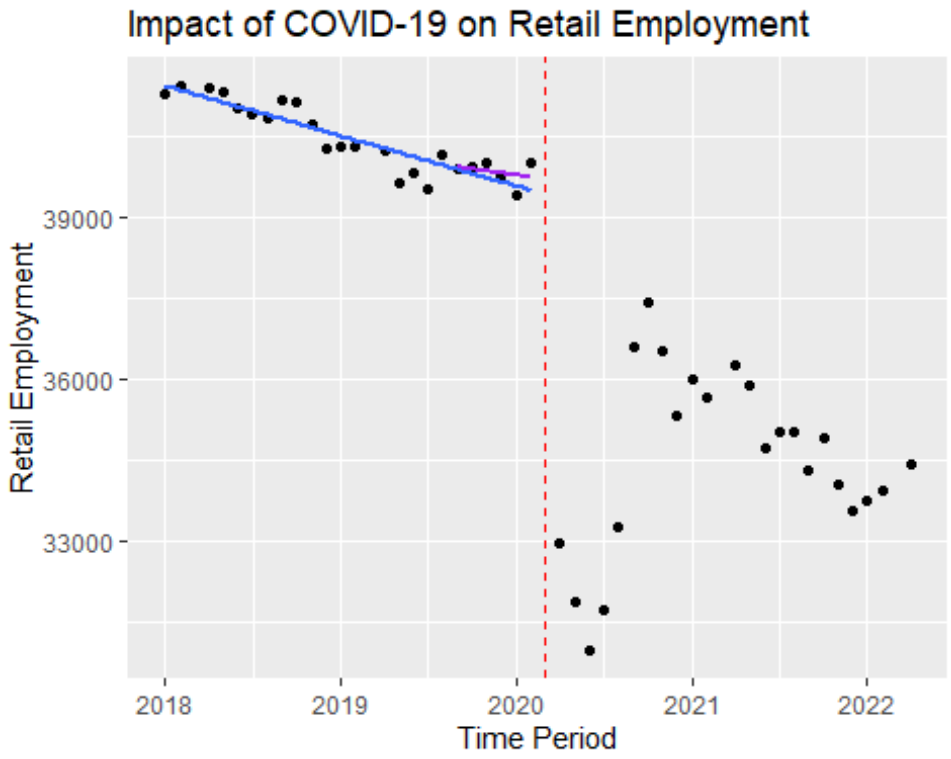
The result of our interrupted time series analysis shows that RetailEmployment decreased by 0.0112 units after COVID happened in March 2020, compared to before COVID. By using the amount of individuals employed in the retail industry as our measurement, we can say with 99.9% significance that COVID negatively impacted the retail industry. In looking at the big picture, some reasons COVID could have negatively impacted the retail industry might include reducing employment, potentially through layoffs, job loss, and so forth.

Interrupted Time Series Model: Narrowing the Window

Referring back to our graph, we can see that during the pre COVID period there was a slight downward trend in Retail Employment, even before COVID had occurred. Since we are only interested in the effect of the cut-off period (COVID), we can improve our model by narrowing the window around COVID so that we don't capture that downwards trend happening prior to the event.

```
# New Data Set
data_agg1 <- data_agg %>%
  filter(date < as.numeric(as.Date("2020-03-01")))
data_agg2 <- data_agg %>%
  filter(date > as.numeric(as.Date("2019-08-01")),
         date < as.numeric(as.Date("2020-03-01")))
# Graph
ggplot(data_agg, aes(x = date, y = total_retail)) +
  geom_point() +
  geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), linetype = "dashed", color = "red") +
  labs(title = "Impact of COVID-19 on Retail Employment",
       x = "Time Period",
       y = "Retail Employment") +
  geom_smooth(data = data_agg1, method = "lm", se = FALSE, span = 0.8) +   geom_smooth(data =
data_agg2, method = "lm", se = FALSE, span = 0.8, color = 'purple')

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



As seen in the graph above, the blue line represents the best fit line from January 2018 to March 2020, which shows an obvious downward trend. The purple line represent the best fit line from August 2019 to the cut-off point at March 2020, which we can see is more flat compared to the blue line. Because the purple line is more flat than the blue line, it shows that narrowing the window around these dates will reduce the effect of prior trends being captured in our model and we will capture a more true effect of COVID. In our next model, we will narrow the window, capturing data only from August 2019 to October 2020 to more accurately reflect the COVID effect on Retail Employment. Also note that our data still has a significant number of observations (1,307,769), despite narrowing down our window.

```
# Narrowing Window
window <- cps_data %>%
  filter(date > as.numeric(as.Date("2019-08-01")),
         date < as.numeric(as.Date("2020-10-01"))) %>%
  mutate(dateCentered = as.numeric(interval(ymd("2020-03-01"), date) / months(1))) %>%
  mutate(Assigned = dateCentered > 0)
# Running Narrow Window Model
m2 <- feols(RetailEmployment ~ Assigned*dateCentered, data = window, vcov = 'hetero')
etable(m2)

                              m2
Dependent Var.:       RetailEmployment

Constant              0.3386*** (0.0013)
AssignedTRUE            -0.0139*** (0.0019)
```

```
dateCentered                  0.0003 (0.0003)
AssignedTRUE x dateCentered   0.0010* (0.0005)
_____________________________ ____________________
S.E. type                     Heteroskedast.-rob.
Observations                        1,307,769
R2                                   8.36e-5
Adj. R2                              8.13e-5
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In narrowing the window of our initial interrupted time series analysis, we can see that RetailEmployment now decreases by 0.0139 units after COVID happened in March 2020, compared to before COVID instead of 0.0112 units. When we narrow the window around the COVID period, we can see that COVID had greater negative impact on retail employment than we previously expected. This could be because the economy (including the Retail Industry) was able to recover in the months following COVID, so including that data in our model would taint our regression to truly see the pure effects of just COVID. Based on these two regression models we can confidently say that by measuring the success of the retail industry by retail employment, COVID had a negative impact on the retail industry.

2. **How has retail fared relative to other industries?**

The goal of this question is to compare the success of the retail industry to other industries after COVID. We can do this by running a regression model and interacting the time period after COVID with industries. An interaction term between these two variables will show us how the retail industry changes as the other industries change in the time period after COVID. Since industry is a categorical variable, we used the retail industry as a reference group, so that all other industries can be compared to the retail industry.

```r
#join indnames with data
indnames <- read.csv("indnames.csv")
cps_data <- cps_data %>%
  filter(IND != 0)
cps_data <- full_join(cps_data, indnames, by = c('IND' = 'ind'))

#before covid
cps_data <- cps_data %>%
  mutate(before_covid = date <= ymd("2020-03-01"))

#number of employed and unemployed
  # 10 - At work
  # 20-22 - Unemployed
cps_data <- cps_data %>%
  mutate(employment = case_when(
    EMPSTAT == 10 ~ 1,
    EMPSTAT >= 20 & EMPSTAT <= 22 ~ 0)) %>%
  filter(!is.na(employment))
```

```
#regression
retail_relative <- feols(employment ~ before_covid * i(indname, ref = 'Retail Trade'), data = cps_data)
```

The variables 'before_covidTRUE:indname::Agriculture, Forestry, Fishing, and Hunting, and Mining', 'before_covidTRUE:indname::Arts, Entertainment, and Recreation, and Accommodation and Food Services' and eleven others have been removed because of collinearity (see $collin.var).

```
etable(retail_relative)
```

	retail_relative
Dependent Var.:	employment
Constant	0.9301*** (0.0006)
before_covidTRUE	0.0259*** (0.0008)
indname = Agriculture,Forestry,Fishing,andHunting,andMining	0.0016 (0.0012)
indname = Arts,Entertainment,andRecreation,andAccommodationandFoodServices	-0.0114*** (0.0008)
indname = Construction	-0.0040*** (0.0009)
indname = EducationalServices,andHealthCareandSocialAssistance	0.0185*** (0.0007)
indname = FinanceandInsurance,andRealEstateandRentalandLeasing	0.0230*** (0.0009)
indname = Information	0.0092*** (0.0014)
indname = Manufacturing	0.0127*** (0.0008)
indname = Military	-0.9560*** (0.0165)
indname = OtherServices,ExceptPublicAdministration	0.0132*** (0.0010)
indname = Professional,Scientific,andManagement,andAdministrativeandWasteManagementServices	0.0072*** (0.0007)
indname = PublicAdministration	0.0257*** (0.0010)
indname = TransportationandWarehousing,andUtilities	0.0125*** (0.0009)
indname = WholesaleTrade	0.0162*** (0.0013)
before_covidFALSE x indname = Agriculture,Forestry,Fishing,andHunting,andMining	0.0107*** (0.0018)
before_covidFALSE x indname = Arts,Entertainment,andRecreation,andAccommodationandFoodServices	-0.0598*** (0.0012)
before_covidFALSE x indname = Construction	0.0018 (0.0013)
before_covidFALSE x indname = EducationalServices,andHealthCareandSocialAssistance	0.0072*** (0.0010)
before_covidFALSE x indname = FinanceandInsurance,andRealEstateandRentalandLeasing	0.0145*** (0.0013)
before_covidFALSE x indname = Information	-0.0025 (0.0022)
before_covidFALSE x indname = Manufacturing	0.0054*** (0.0012)
before_covidFALSE x indname = Military	0.0259 (0.0249)
before_covidFALSE x indname = OtherServices,ExceptPublicAdministration	-

```
0.0131*** (0.0015)
before_covidFALSE x indname =
Professional,Scientific,andManagement,andAdministrativeandWasteManagementServices  0.0064***
(0.0011)
before_covidFALSE x indname = PublicAdministration                                  0.0212***
(0.0014)
before_covidFALSE x indname = TransportationandWarehousing,andUtilities                    -
0.0109*** (0.0014)
before_covidFALSE x indname = WholesaleTrade                                        0.0091***
(0.0019)
_____________________________________________
_____________________
S.E. type                                                IID
Observations                                         2,515,365
R2                                                     0.01558
Adj. R2                                                0.01556
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Employment of most industries increased compared to the "Retail Trade" Industry after COVID. "Public Administration" after COVID increased by 2.21% compared to "Retail Trade" Industry, which was the highest increase of all other industries. However, the industry of "Arts, Entertainment, and Recreation, and Accommodation and Food Services" decreased the most after COVID compared to Retail by 5.98%. Other industries like "Information", "Other Services, Except Public Administration", and"Transportation and Warehousing, and Utilities" also decreased compared to "Retail Trade" after COVID by 0.25%, 1.31%, and 1.09% respectively. Even though most of the other industries are experiences less of an impact from COVID compared to the retail industry, our company should try to retain the employees we already have and build a good company reputation, knowing that other retail companies might be laying off more employees.

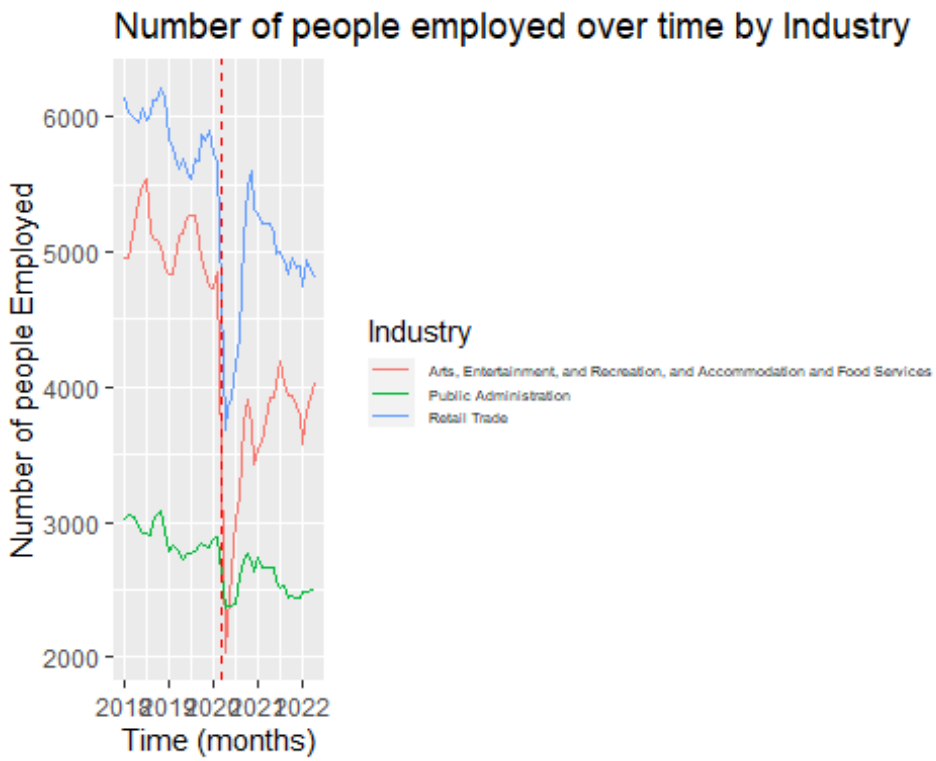
As we can see from the regression, the employment of the "Arts, Entertainment, and Recreation, and Accommodation and Food Services" industry decreased the most after COVID compared to "Retail Trade", and "Public Administration" increased the most compared to "Retail Trade" after COVID. So, we can visually see these differences in the following graph. The vertical line represents the specific time of March 2020 when COVID affected all the industries.

```
#dataset for graph
graph2 <- cps_data %>%
  group_by(indname, date) %>%
  summarize(number_employed = sum(employment == 1))

`summarise()` has grouped output by 'indname'. You can override using the
`.groups` argument.

#picking highest and lowest employed to compare to retail
graph2 <- graph2 %>%
  filter(indname %in% c("Public Administration","Retail Trade","Arts, Entertainment, and Recreation,
and Accommodation and Food Services"))
```

```
ggplot(graph2, aes(x=date, y=number_employed, color = indname)) +
  geom_line() +
  geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), linetype = "dashed", color = "red") +
  labs(title = "Number of people employed over time by Industry", y = "Number of people Employed", x
= "Time (months)", color = "Industry") +
  theme(legend.text = element_text(size=5),
        legend.key.height = unit(0.3, 'cm'))
```



The goal of the graph was to capture the true effect of COVID on employment in various industries, We chose to portray the retail industry (blue) and compare it to Arts industry (red) which decreased the most post COVID. We also compared Public Admin (green) to retail, which showed the greatest positive change compared to retail.

3. **Retail needs to worry about who has money to spend - what has changed about who is working and earning money?**

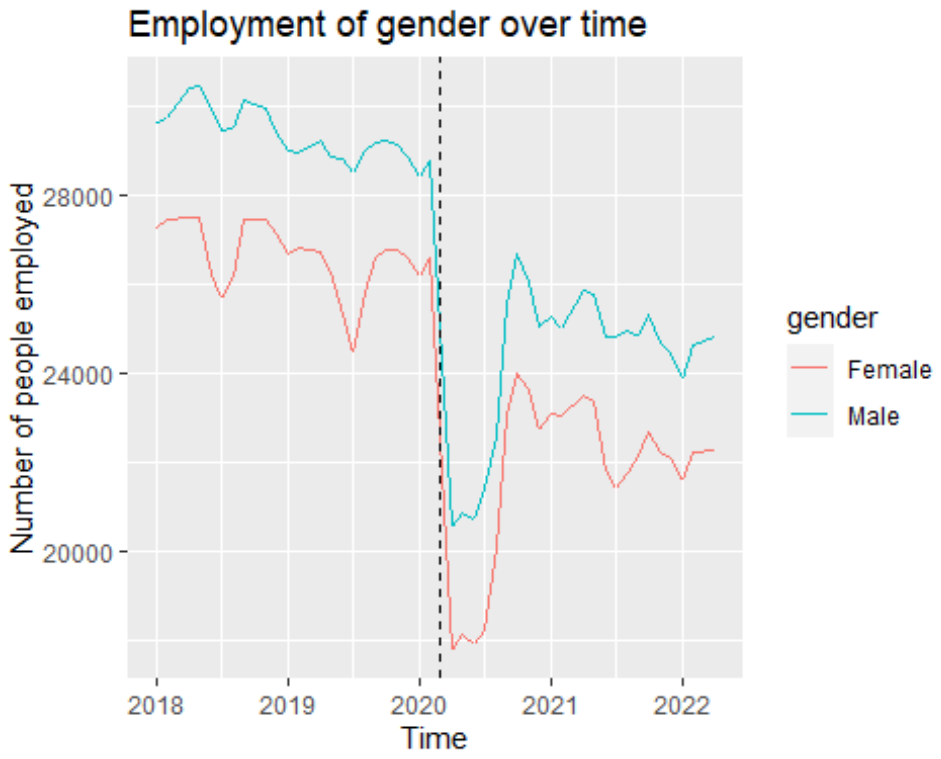
In order to analyze who is working and earning money, we can look at the breakdown between genders, generations, and different race groups.

Gender

```
#sex to categorical
cps_data <- cps_data %>%
  mutate(gender = factor((ifelse(SEX == 1, "Male", "Female"))))

#gender df
```

```r
gender_emp <- cps_data %>%
  group_by(date, gender) %>%
  mutate(numberofEmployed = sum(employment == 1),
         numberofUnemployed = sum(employment == 0))
#plot
ggplot(gender_emp, aes(x = date, y = numberofEmployed, color = gender)) +
  geom_line() +
  geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), linetype = "dashed") +
  labs(title = "Employment of gender over time",
       x = "Time",
       y = "Number of people employed")
```



The graph above shows that Males and Females both saw a significant decrease in employment after COVID (March 2020). While the number of people employed recovered a few months later, it hasn't reached pre-COVID levels. We can also see that throughout this time period (2018-2022) there are way more men employed than women. We can run a regression to see the effects of gender on employment and if it has changed since COVID.

```r
#regression
cps_data <- cps_data %>%
  mutate(after_covid = ifelse(date >= ymd("2020-03-01"), 1, 0))

moneyToSpend_gender <- feols(employment ~ gender*after_covid, data = cps_data)
etable(moneyToSpend_gender)

                    moneyToSpend_gender
Dependent Var.:              employment
```

```
Constant                0.9660*** (0.0003)
genderMale               -0.0010** (0.0004)
after_covid             -0.0282*** (0.0004)
genderMale x after_covid  0.0033*** (0.0005)
_____________________________ ___________________
S.E. type                            IID
Observations                   2,515,365
R2                                0.00392
Adj. R2                           0.00392
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the results from the regression, in the period before COVID, the coefficient on genderMale shows us that male employment decreased by 0.001.The coefficient on after_covid, tells us that during the period after COVID, female employment decreases by 0.0282. The positive coefficient on the interaction term of genderMale and after_covid tells us that the effect of being male on employment is 0.0033 units higher among males after COVID compared to males before COVID. Since male employment remained higher than female employment after COVID, our company should consider targeting males, as the market is more stable for males.

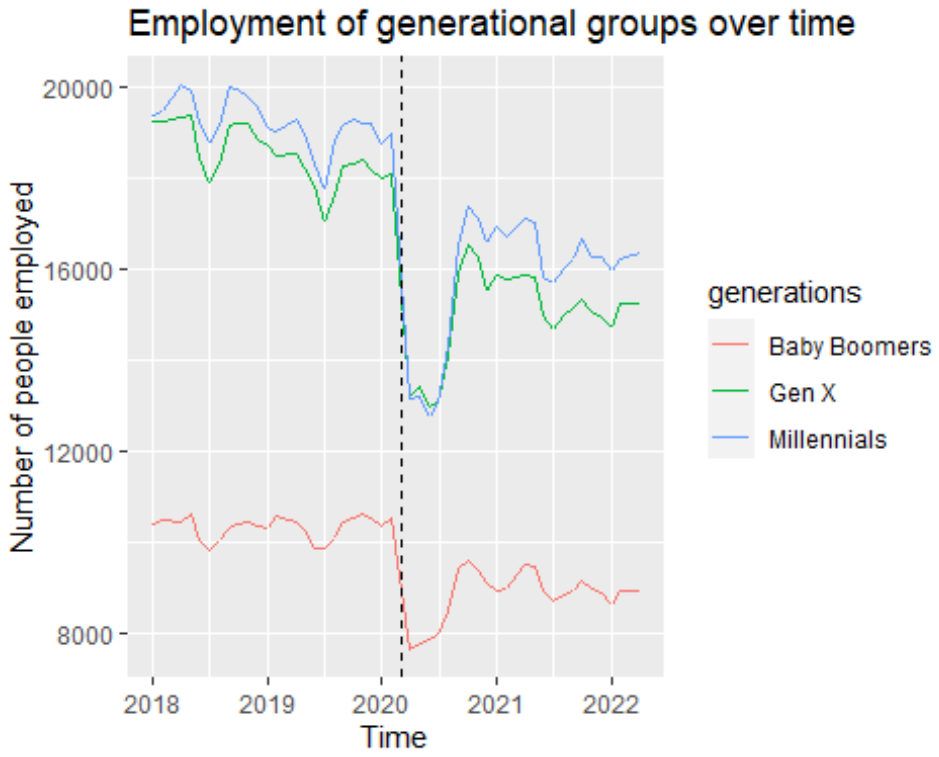
Age

Since Gen Alpha and Gen Z and includes individuals between 1 and 25 years old, we would want to drop them because most people that have a job are 25+ years old. The legal age to work starts at 16, but from the ages of 16-25, they are less likely to be working. In the Silent generation, these individuals are likely retired, so we are also dropping this generation.

```r
#age to categorical
##Gen alpha 9, generation Z 10-25, millenials 26-41, generation X 42-57, baby boomers 58-76, silent generation 77-97
cps_data <- cps_data %>%
  mutate(generations = case_when(
    26 <= AGE & AGE <= 41 ~ "Millennials",
    42 <= AGE & AGE <= 57 ~ "Gen X",
    58 <= AGE & AGE <= 76 ~ "Baby Boomers",
  )) %>%
  filter(!is.na(generations))

#age df
age_emp <- cps_data %>%
  group_by(date, generations) %>%
  mutate(numberofEmployed = sum(employment == 1),
         numberofUnemployed = sum(employment == 0))
#plot
ggplot(age_emp, aes(x = date, y = numberofEmployed, color = generations)) +
  geom_line() +
  geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), linetype = "dashed") +
  labs(title = "Employment of generational groups over time",
```

```
    x = "Time",
    y = "Number of people employed")
```



Before COVID, Millennials had the highest the number of people employed to begin with but was comparable to Gen X, which had slightly less. Baby Boomers had the least. However when COVID happened, all three generations had a decrease in employment, with Millennials slipping under Gen X during that drop. All generations increased again after COVID, but did not recover back to the pre-COVID employment numbers. After COVID, Millennials still have the highest number of people employed, followed by Gen X, then Baby Boomers, however, in general there are less people employed across the board, meaning less people have money to spend post-COVID.

```
#reference group millennials
cps_data$generations <- factor(cps_data$generations)
cps_data$generations <- relevel(cps_data$generations, ref = "Millennials")

#regression
cps_data <- cps_data %>%
  mutate(after_covid = date >= ymd("2020-03-01"))


moneyToSpend_age <- feols(employment ~ generations*after_covid, data = cps_data)
etable(moneyToSpend_age)

                                 moneyToSpend_age
Dependent Var.:                        employment
```

```
Constant                                  0.9671*** (0.0003)
generationsBabyBoomers                     0.0042*** (0.0005)
generationsGenX                           0.0050*** (0.0004)
after_covidTRUE                           -0.0263*** (0.0004)
generationsBabyBoomers x after_covidTRUE     0.0002 (0.0007)
generationsGenX x after_covidTRUE         0.0021*** (0.0006)
_____________________________________________ ____________________
S.E. type                                      IID
Observations                                 2,151,178
R2                                          0.00421
Adj. R2                                      0.00421
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the coefficients on the interaction terms between generations and after_covid shows how the effect of COVID on employment levels differ for each generational group (Baby boomers and Gen X) compared to the reference group (Millennials). The positive coefficients on the results indicate that COVID had less of a negative impact on employment of the Baby Boomer and Gen X generations compared to the reference group (Millennials). However although these two generations had less of a severe drop in employment compared to Millennials post-COVID, Millennials still had more people employed overall. This suggests that retail can take notice that Millennials have higher employment levels compared to the other generations, and therefore are the ones with money to spend.

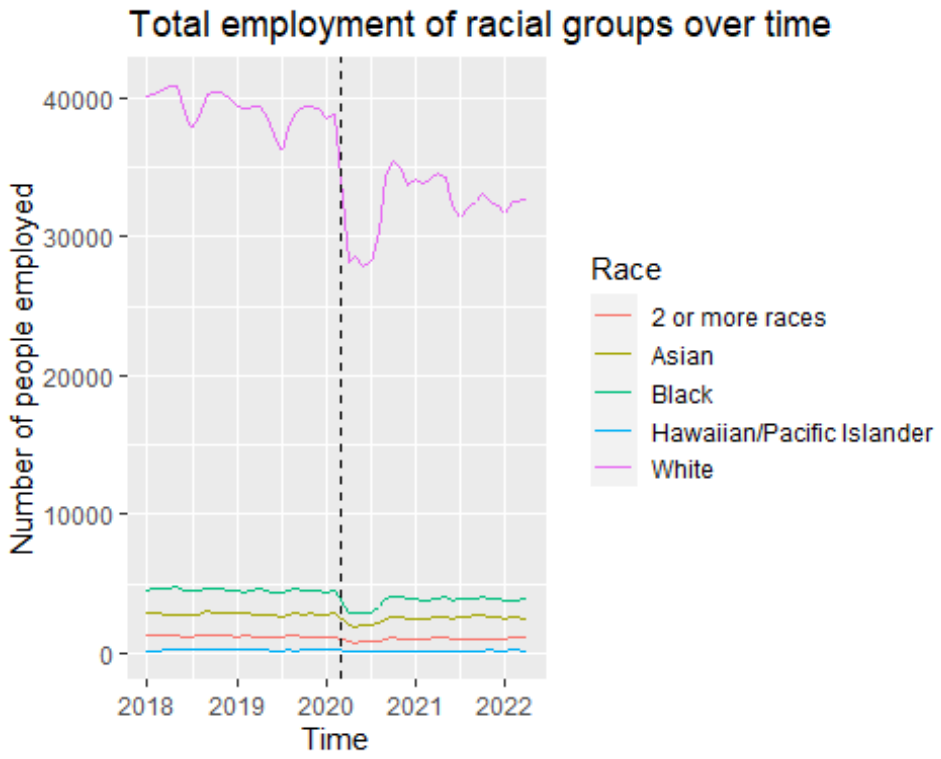
Race

```
#race to categorical
cps_data <- cps_data %>%
  mutate(races = case_when(
    RACE == 100 ~ "White",
    RACE == 200 ~ "Black",
    RACE == 652 ~ "Hawaiian/Pacific Islander",
    RACE == 651 ~ "Asian",
    TRUE ~ "2 or more races"
  ))

#race df
race_emp <- cps_data %>%
  group_by(date, races) %>%
  mutate(numberofEmployed = sum(employment == 1),
         numberofUnemployed = sum(employment == 0))
```

We looked at the total number of people employed in each race, which can provide insight into the Market size of each race group. Knowing the number of people employed in each race can help our company see how each race group has been affected by the pandemic in terms of employment, which can help inform decisions about where our company should focus their efforts to address any potential changes in consumer spending.

```r
#plot
ggplot(race_emp, aes(x = date, y = numberofEmployed, color = races)) +
  geom_line() +
  geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), linetype = "dashed") +
  labs(title = "Total employment of racial groups over time",
       x = "Time",
       y = "Number of people employed",
       color = "Race") +
  theme(legend.text = element_text(size=9),
        legend.key.height = unit(0.5, 'cm'))
```



As we can see from the graph, before COVID, White individuals have the highest number of employed individuals, and all other races (Asian, Black, Hawaiian/Pacific Islander, 2 or more races) have lower levels of employment in comparison. Employment levels for every race dropped after COVID. After COVID, all racial groups' number of employed individuals were on the rise, but regardless of the time period, White individuals made up most of the employed. Because of this, we picked White to be the reference group when running our regression.

```r
#after covid
cps_data <- cps_data %>%
  mutate(after_covid = date >= ymd("2020-03-01"))

#reference group white
cps_data$races <- factor(cps_data$races)
cps_data$races <- relevel(cps_data$races, ref = "White")
```

```
#regression for number of unemployed
moneyToSpend_race <- feols(employment ~ races*after_covid, data = cps_data)
etable(moneyToSpend_race)
```

	moneyToSpend_race
Dependent Var.:	employment
Constant	0.9727*** (0.0002)
races2ormoreraces	-0.0250*** (0.0011)
racesAsian	0.0054*** (0.0008)
racesBlack	-0.0246*** (0.0006)
racesHawaiian/PacificIslander	-0.0061* (0.0028)
after_covidTRUE	-0.0230*** (0.0003)
races2ormoreraces x after_covidTRUE	-0.0083*** (0.0017)
racesAsian x after_covidTRUE	-0.0164*** (0.0011)
racesBlack x after_covidTRUE	-0.0109*** (0.0009)
racesHawaiian/PacificIslander x after_covidTRUE	-0.0216*** (0.0041)
S.E. type	IID
Observations	2,151,178
R2	0.00653
Adj. R2	0.00652

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The coefficients on the interaction terms show us how the effect of COVID on employment levels differ for each race group compared White. The negative coefficient results indicate that COVID had a greater negative impact on employment of other races (Black, Hawaiian/Pacific Islander, 2 or more races) compared to the reference group (White). Since these races had lower employment compared to White individuals post-COVID, this suggests that our company can take notice that White individuals have higher employment levels compared to other races, and therefore are the ones with money to spend.