Data Translation Challenge

## Data Translation Challenge

Oliver, Ariana, Alice, Jiale, Annika

## Data Cleaning

#loading in lirbaries  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ readr 2.1.4  
✔ ggplot2 3.4.4 ✔ stringr 1.5.1  
✔ lubridate 1.9.3 ✔ tibble 3.2.1  
✔ purrr 1.0.2 ✔ tidyr 1.3.0

── 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(fixest)  
library(vtable)

Loading required package: kableExtra  
  
Attaching package: 'kableExtra'  
  
The following object is masked from 'package:dplyr':  
  
 group\_rows

library(rio)  
library(lubridate)  
library(ipumsr)  
library(rdrobust)  
library(multcomp)

Loading required package: mvtnorm  
Loading required package: survival  
Loading required package: TH.data  
Loading required package: MASS  
  
Attaching package: 'MASS'  
  
The following object is masked from 'package:dplyr':  
  
 select  
  
  
Attaching package: 'TH.data'  
  
The following object is masked from 'package:MASS':  
  
 geyser

#reading in data  
ddi <- read\_ipums\_ddi("cps\_00002.xml")  
cps\_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.

#importing indnames.csv file  
indnames <- import("indnames.csv")  
  
#Create post-lockdown variables (TRUE if after March 2020)  
cps\_data <- cps\_data%>%  
 mutate(Date = ym(paste0(YEAR, MONTH)),  
 post\_lockdown = Date > ymd("2020/03-01"))  
  
#Joining data ('joined\_data' will be our base data for each question)  
indnames <- indnames %>%  
 mutate(IND = ind)  
joined\_data <- cps\_data %>%  
 inner\_join(indnames, by = "IND")

## Building/Running Regression Question One

**Research Question:** How has COVID affected the health of the retail industry, as measured by employment?

outcome variable = employment

treatment variable = COVID (true if after March 2020)

cutoff = March 2020

#Filter results to only get retail industry data  
joined\_data\_1 <- joined\_data%>%  
 filter(indname == "Retail Trade")  
  
#Create plot to see trend of employment in the retail industry throughout time  
joined\_data\_1 <- joined\_data\_1%>%  
 group\_by(Date)%>%  
 summarize(Employed = sum(EMPSTAT %in% c(10,12)))  
  
ggplot(joined\_data\_1, aes(x = Date, y = Employed)) + geom\_line() + geom\_vline(xintercept = ymd("2020-03-01")) + labs(title = 'Retail Industry Employment from 2019-2022', y = 'Number of Employed Retail Workers')



#Create center variable and treated variable  
joined\_data\_1 <- joined\_data\_1%>%  
 mutate(COVID\_center = as.numeric(Date) - as.numeric(ymd("2020-03-01")),  
 post\_lockdown = Date > ymd("2020-03-01"))%>%  
 filter(COVID\_center <= 350)%>%  
 filter(COVID\_center >= -350)  
  
#regress amount of employed people  
model1 <- feols(Employed ~ post\_lockdown + COVID\_center + post\_lockdown\*COVID\_center, data = joined\_data\_1, vcov = 'hetero')  
etable(model1)

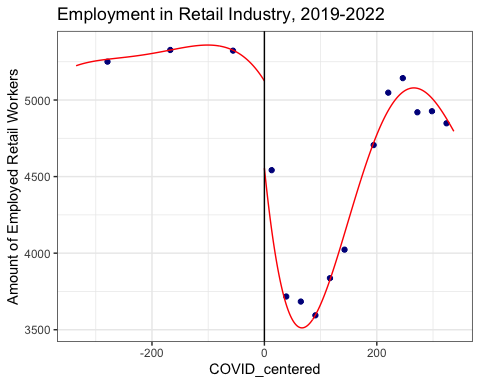
model1  
Dependent Var.: Employed  
   
Constant 5,112.8\*\*\* (198.5)  
post\_lockdownTRUE -1,703.8\*\*\* (249.4)  
COVID\_center -0.7300 (0.8529)  
post\_lockdownTRUE x COVID\_center 6.150\*\*\* (1.200)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 23  
R2 0.83762  
Adj. R2 0.81199  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Use rdrobust package to estimate effect of COVID on employment in the retail industry

model1b <- rdrobust(joined\_data\_1$Employed, joined\_data\_1$COVID\_center, c = 0, kernel = 'uniform')  
summary(model1b)

Sharp RD estimates using local polynomial regression.  
  
Number of Obs. 23  
BW type mserd  
Kernel Uniform  
VCE method NN  
  
Number of Obs. 11 12  
Eff. Number of Obs. 3 4  
Order est. (p) 1 1  
Order bias (q) 2 2  
BW est. (h) 110.292 110.292  
BW bias (b) 169.362 169.362  
rho (h/b) 0.651 0.651  
Unique Obs. 11 12  
  
=============================================================================  
 Method Coef. Std. Err. z P>|z| [ 95% C.I. ]   
=============================================================================  
 Conventional -876.068 546.796 -1.602 0.109 [-1947.769 , 195.633]   
 Robust - - -1.212 0.226 [-1969.225 , 464.654]   
=============================================================================

rdplot(joined\_data\_1$Employed, joined\_data\_1$COVID\_center, c = 0, title = "Employment in Retail Industry, 2019-2022", x.label = "COVID\_centered", y.label = "Amount of Employed Retail Workers")



## Write Up Question One

Running a regression discontinuity design would be best to measure the effect of COVID on the health of the retail industry. The running variable is Date, with the cutoff value being March 2020 because majority of COVID lockdown policies were implemented in the United States at that time.  The outcome variable is the number of employed retail workers, aggregated to the month level for simplicity. The regression model takes data just right below/above the cutoff, which closes back doors and leaves us with variation around the cutoff, allowing for variation in the treatment (COVID). By creating the post\_lockdown and COVID\_center variables, the regression model can capture the potential impact of the pandemic on retail employment levels, both in terms of an immediate shift (the post\_lockdown variable) and a gradual trend over time (the COVID\_center variable and its interaction with post\_lockdown).  The variable COVID\_center is created by taking the Date from each data point and measuring how far it is relative to March 2020.  The post\_lockdown variable is TRUE if the Date of the data point comes after March 2020, and FALSE if the Date is March 2020 or prior.  We then took those variables and regressed Employed (total amount of employed workers) on the interaction term (COVID\_center\*post\_lockdown).

The analysis aims to quantify the effect of the pandemic and associated lockdowns on retail employment, providing insights into the industry’s resilience or vulnerability during this period. Overall, this analysis contributes to understanding the dynamics of retail employment, especially in the context of the unprecedented disruptions caused by the COVID-19 pandemic. One assumption that needs to be made to run this RDD is that the window of data we chose (April 2019 to February 2022) is small enough that only allows for variation from the treatment (COVID) and all other sources of endogeneity are closed. Additionally, data being manipulated is implausible in this context as the data comes from a credible source and businesses were most likely unaware of the consequences of the COVID lockdown until March 2020 came. We also assume the only variable jumping at the cutoff is COVID, and no other variable is changing employment levels during March 2020. We also need the difference to be at the cutoff, which we can assume to be true based on the graph having a steep decline in employment levels right in March 2020 (indicated by vertical line). This is a sharp RDD rather than a fuzzy RDD because everyone in the Unites States after March 2020 was impacted by the COVID protocol, and no one prior to March 2020 was affected by COVID.

From our regression, we estimated that the retail industry had an average of about 5,112.8 employed workers prior to COVID and the lockdown (Constant = 5112.8).  Comparatively, the retail industry had 1,703.8 less employed workers after March 2020 or during COVID (post\_lockdown = -1,703.8), meaning that employment levels in the retail industry dropped 33.32% when COVID lockdown was implemented.  Both the average number of employed workers in the retail industry before and after March 2020 are statistically significant. The coefficient for COVID\_center (-.7300) is not statistically significant, indicating that there was no significant linear trend in retail employment before the lockdown period. The interaction term post\_lockdownTRUE x COVID\_center has a positive coefficient (6.150) that is statistically significant at the 5% level. This suggests that after the initial drop in employment following the lockdown, there was a gradual increase in the number of employed retail workers over time. Around 83.76% of the variation in total amount of employed workers is explained by the treatment of COVID.

We also used the rdrobust package to run the RDD as rdrobust does not rely on linearity. From the rdrobust function, we found that the retail industry during COVID had an average of 876.07 less employed workers than before COVID, but this result is not statistically significant at the 5% level. This could be a result of having a small dataset, making our estimate noisy.

In summary, the regression results suggest that the COVID-19 lockdown had a significant negative impact on the health of the retail industry, with a substantial immediate drop in the number of employed workers. However, there was a gradual recovery in retail employment over time, as indicated by the positive interaction term between the post-lockdown period and the number of days since the lockdown. This could be due to other policies implemented to fix the economic consequences of COVID, such as better unemployment benefits motivating workers to be unemployed.

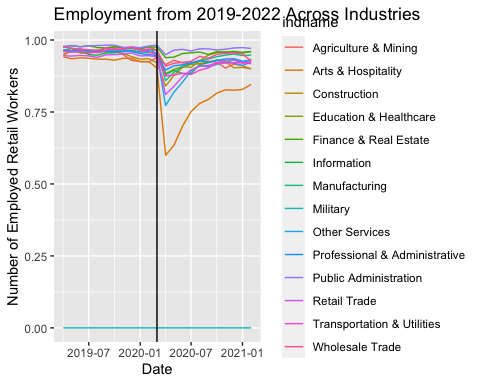
## Building/Running Regression Question Two

**Research Question:** How has retail fared relative to other industries?

#Create 'joined\_data\_2' using 'joined\_data' for this regression  
   
#Create plot to see how employment in different industries changed over time (visually bad)  
joined\_data\_2 <- joined\_data%>%  
 mutate(indname = case\_when(  
 indname == "Agriculture, Forestry, Fishing, and Hunting, and Mining" ~ "Agriculture & Mining",  
 indname == "Arts, Entertainment, and Recreation, and Accommodation and Food Services" ~ "Arts & Hospitality",  
 indname == "Construction" ~ "Construction",  
 indname == "Educational Services, and Health Care and Social Assistance" ~ "Education & Healthcare",  
 indname == "Finance and Insurance, and Real Estate and Rental and Leasing" ~ "Finance & Real Estate",  
 indname == "Information" ~ "Information",  
 indname == "Manufacturing" ~ "Manufacturing",  
 indname == "Military" ~ "Military",  
 indname == "Other Services, Except Public Administration" ~ "Other Services",  
 indname == "Professional, Scientific, and Management, and Administrative and Waste Management Services" ~ "Professional & Administrative",  
 indname == "Public Administration" ~ "Public Administration",  
 indname == "Retail Trade" ~ "Retail Trade",  
 indname == "Transportation and Warehousing, and Utilities" ~ "Transportation & Utilities",  
 indname == "Wholesale Trade" ~ "Wholesale Trade",  
 TRUE ~ "Other" # for any other cases not specified  
 )) %>%  
 group\_by(indname, Date)%>%  
 summarize(Employed = sum(EMPSTAT %in% c(10,12))/sum(!is.na(EMPSTAT)))%>%  
 filter(Date >= ymd("2019-04-01"))%>%  
 filter(Date <= ymd("2021-02-01"))

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

ggplot(joined\_data\_2, aes(x = Date, y = Employed, color = indname)) + geom\_line() + geom\_vline(xintercept = ymd("2020-03-01")) + labs(title = 'Employment from 2019-2022 Across Industries', y = 'Number of Employed Retail Workers')



#Create center variable and treated variable  
joined\_data\_2 <- joined\_data\_2%>%  
 mutate(COVID\_center = as.numeric(Date) - as.numeric(ymd("2020-03-01")),  
 post\_lockdown = Date > ymd("2020-03-01"),  
 indname = factor(indname))  
  
#Regression  
model2 <- feols(Employed ~ post\_lockdown + COVID\_center + post\_lockdown\*COVID\_center + i(indname, ref = "Retail Trade"), data = joined\_data\_2, vcov = 'hetero')  
etable(model2)

model2  
Dependent Var.: Employed  
   
Constant 0.9412\*\*\* (0.0043)  
post\_lockdownTRUE -0.0795\*\*\* (0.0090)  
COVID\_center -2.43e-5. (1.43e-5)  
indname = Agriculture&Mining 0.0114. (0.0059)  
indname = Arts&Hospitality -0.0716\*\*\* (0.0163)  
indname = Construction 0.0049 (0.0051)  
indname = Education&Healthcare 0.0286\*\*\* (0.0043)  
indname = Finance&RealEstate 0.0422\*\*\* (0.0050)  
indname = Information 0.0145\*\*\* (0.0041)  
indname = Manufacturing 0.0218\*\*\* (0.0042)  
indname = Military -0.9228\*\*\* (0.0068)  
indname = OtherServices 0.0033 (0.0068)  
indname = Professional&Administrative 0.0181\*\*\* (0.0042)  
indname = PublicAdministration 0.0501\*\*\* (0.0056)  
indname = Transportation&Utilities 0.0102\* (0.0041)  
indname = WholesaleTrade 0.0303\*\*\* (0.0044)  
post\_lockdownTRUE x COVID\_center 0.0002\*\*\* (3.75e-5)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 322  
R2 0.98884  
Adj. R2 0.98826  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Write Up Question Two

1. The analysis utilizes a fixed-effects regression model to isolate the impact of COVID-19 lockdowns on employment across various industries, using retail as a baseline for comparison. By incorporating industry-specific dummy variables (with Retail as the reference category), the model can compare the employment changes in each industry relative to Retail. Interaction terms between the post-lockdown period and industries enable an assessment of differential impacts over time. The outcome reveals how employment in Retail compares to other sectors before, during, and after lockdowns, providing a direct answer to the research question.

2.The graph analysis shows the trend in the number of employed retail workers over time across different industries. Retail is one of the lines on the graph, and its trajectory relative to others can visually indicate how it’s fared. If the retail line remains above other lines or recovers more quickly after the vertical line (indicating the start of lockdowns), it suggests retail fared better than those industries. The regression model directly answers the question by providing a numerical basis for comparison. Since “Retail Trade” is the reference category for the indname factor, the coefficients for the other industries show how they compare to retail. The model includes an interaction term for post-lockdown trends, giving insight into how retail employment trends compared to other sectors after the onset of the pandemic.

3. Regression results analysis

Graph Result: If Retail Trade is one of the lines in the graph, its position and trajectory relative to other lines indicate how retail fared. For instance, if Retail Trade’s line drops less than others around March 2020, it suggests retail fared relatively better during the lockdown.

Regression Result:

* indname Coefficients: These show the difference in employment between each industry and the Retail Trade industry, which is the reference category. Negative coefficients for other industries suggest they have fared worse than retail in terms of employment numbers, and positive coefficients suggest they have fared better.
* post\_lockdownTRUE: This shows the change in employment in the retail industry after the lockdown was imposed. Since the coefficient is negative, it indicates a decrease in retail employment after the lockdown compared to before.
* post\_lockdownTRUE x COVID\_center: The interaction term’s positive coefficient suggests that while there was a drop in employment after the lockdown, the relative decline slowed over time.

Overall Interpretation Based on the Retail Industry:

* The analyses show that retail employment experienced a significant decrease immediately following the lockdown (as indicated by the post\_lockdownTRUE
* coefficient). However, the positive and significant interaction term suggests that the rate of decline in retail employment may have slowed over time compared to the initial impact of the lockdown.
* Comparing retail to other industries (as shown by the indname coefficients), most industries have a negative coefficient, indicating they fared worse than retail since these coefficients represent the difference from retail employment. However, Education & Healthcare has a positive coefficient, indicating this sector fared better than retail in terms of employment numbers during the period studied.
* The high R-squared of the regression model implies that these variables explain a large proportion of the variance in employment, suggesting the model fits the data well.

4. Assumptions

* The industry categories accurately capture and group the relevant industries. This assumption seems reasonable as the industry categories are based on standard industry classifications.
* The data is representative of the overall employment trends in each industry.
* The model includes all relevant variables, and there are no omitted variable biases. It is possible that other relevant factors, such as industry-specific policies, regulations, or technological changes, could have influenced employment levels.

5. Retail employment saw a decrease following the COVID-19 lockdowns but fared relatively better than some industries like Agriculture & Mining, Arts & Hospitality, and Other Services. The rate of decline in retail employment seemed to slow over time, indicating some recovery or adaptation. In contrast, sectors like Education & Healthcare experienced growth in employment, likely due to increased demand in healthcare services during the pandemic. In conclusion, the retail sector was impacted by the lockdown but showed signs of resilience when compared to several other sectors. The analysis, while robust, is based on several assumptions, most of which are plausible given the context but may require further data or analysis to fully validate.

## Building/Running Regression Question Three

**Research Question:** Retail needs to worry about who has money to spend - what has changed about *who* is working and earning money?

#Create 'joined\_data\_3' using 'joined\_data' for this regression  
  
#Create categorical variables for sex, race, and age  
joined\_data\_3 <- joined\_data %>%  
 mutate(  
 Employed = EMPSTAT %in% c(10, 12), # Assuming codes 10 and 12 indicate employment  
 Male = SEX == 1,  
 RACE = case\_when(  
 RACE == 100 ~ "White",  
 RACE == 200 ~ "Black",  
 RACE == 651 ~ "Asian",  
 TRUE ~ "Other"  
 ),  
 AGE = case\_when(  
 AGE >= 18 & AGE <= 29 ~ "Young Adult",  
 AGE >= 30 & AGE <= 41 ~ "Adult",  
 AGE >= 42 & AGE <= 53 ~ "Middle Aged",  
 AGE >= 54 & AGE <= 65 ~ "Senior"  
 )  
 )  
  
employment\_summary <- joined\_data\_3 %>%  
 group\_by(post\_lockdown, RACE) %>%  
 summarize(Average\_Employed = mean(as.numeric(Employed), na.rm = TRUE), .groups = 'drop')  
  
  
#regression:  
model\_race\_age\_gender <- feols(Employed ~ post\_lockdown \* RACE + AGE + Male, data = joined\_data\_3, vcov = "hetero")  
  
# To print the regression results  
summary(model\_race\_age\_gender)

OLS estimation, Dep. Var.: Employed  
Observations: 1,993,703   
Standard-errors: Heteroskedasticity-robust   
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.975365 0.000851 1145.95764 < 2.2e-16 \*\*\*  
post\_lockdownTRUE -0.039317 0.001241 -31.69043 < 2.2e-16 \*\*\*  
RACEBlack -0.035205 0.001189 -29.60235 < 2.2e-16 \*\*\*  
RACEOther -0.031674 0.001691 -18.72923 < 2.2e-16 \*\*\*  
RACEWhite -0.006215 0.000833 -7.45753 8.8194e-14 \*\*\*  
AGEMiddle Aged 0.004093 0.000406 10.08902 < 2.2e-16 \*\*\*  
AGESenior -0.003824 0.000435 -8.79829 < 2.2e-16 \*\*\*  
AGEYoung Adult -0.035423 0.000511 -69.37450 < 2.2e-16 \*\*\*  
MaleTRUE 0.000723 0.000327 2.20934 2.7151e-02 \*   
post\_lockdownTRUE:RACEBlack 0.004281 0.001757 2.43714 1.4804e-02 \*   
post\_lockdownTRUE:RACEOther 0.004808 0.002439 1.97143 4.8675e-02 \*   
post\_lockdownTRUE:RACEWhite 0.015711 0.001286 12.21262 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
RMSE: 0.230415 Adj. R2: 0.010123

# Or, for a formatted output  
etable(model\_race\_age\_gender)

model\_race\_age\_ge..  
Dependent Var.: Employed  
   
Constant 0.9754\*\*\* (0.0009)  
post\_lockdownTRUE -0.0393\*\*\* (0.0012)  
RACEBlack -0.0352\*\*\* (0.0012)  
RACEOther -0.0317\*\*\* (0.0017)  
RACEWhite -0.0062\*\*\* (0.0008)  
AGEMiddleAged 0.0041\*\*\* (0.0004)  
AGESenior -0.0038\*\*\* (0.0004)  
AGEYoungAdult -0.0354\*\*\* (0.0005)  
MaleTRUE 0.0007\* (0.0003)  
post\_lockdownTRUE x RACEBlack 0.0043\* (0.0018)  
post\_lockdownTRUE x RACEOther 0.0048\* (0.0024)  
post\_lockdownTRUE x RACEWhite 0.0157\*\*\* (0.0013)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 1,993,703  
R2 0.01013  
Adj. R2 0.01012  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Write Up Question Three

In the wake of the COVID-19 pandemic, the retail sector faces unprecedented challenges, chief among them understanding the shifts in consumer spending power due to changes in employment status across various demographics. This complex scenario necessitates a detailed analysis to answer the critical question: “Retail needs to worry about who has money to spend - what has changed about who is working and earning money?” Such an inquiry is not only vital for gauging the pandemic’s impact on employment but also for identifying potential consumer bases for targeted recovery efforts.

Employing a regression analysis using the feols function from the fixest package, this study meticulously models the likelihood of employment against the backdrop of the COVID-19 lockdown, taking into account variables such as race, age, and sex, and including interaction terms between the lockdown periods and race. This approach enables a nuanced examination of the lockdown’s effects on employment across different racial groups, with adjustments for age and sex considerations. Through the interpretation of interaction term coefficients, we can ascertain the differential impacts of the lockdown on employment among these groups.

The findings unearth several insightful observations:

**Lockdown Impact**: A marked decrease in the probability of employment following the lockdown is signaled by a significant negative coefficient for post\_lockdownTRUE. This suggests a widespread downturn in employment opportunities as a direct consequence of pandemic-induced restrictions.

**Racial Disparities**: Initial analysis reveals inherent disparities in employment probabilities among racial groups, with Black, Other, and White groups all showing lower employment chances compared to Asians, the reference group. The most pronounced decline is observed among Black individuals.

**Age and Sex Differences**: The analysis also sheds light on age and sex as determinants of employment probability. Young adults and seniors exhibit notable variances in employment prospects relative to the baseline group of adults aged 30-41. Interestingly, middle-aged individuals appear slightly more likely to be employed, while males exhibit a marginally higher employment probability, though the effect is minimal.

**Lockdown by Race Interaction**: The interaction effects reveal a slight mitigation of the lockdown’s negative employment impacts for Black and Other races relative to Asians. Conversely, White individuals manifest a significant relative improvement in employment prospects during the lockdown period.

Addressing the underlying assumptions of this analysis is crucial for its validity. The assumption of linearity, the inclusion of all relevant variables, the independence of observations, and the correct model specification are foundational yet subject to scrutiny. For instance, the exclusion of variables representing economic conditions, industry-specific employment shifts, or levels of educational attainment could potentially skew the results. As such, these assumptions warrant careful consideration and validation against the data and broader economic indicators.

Synthesizing these insights, the analysis unequivocally demonstrates the profound and varied impacts of the COVID-19 lockdown on employment across different demographics, underscoring significant racial disparities. While all demographics have suffered, the differential effects highlight the necessity for the retail sector to adapt its strategies to address the evolving economic landscape and consumer capacities. Specifically, the relative improvement in employment probabilities for White individuals during the lockdown period offers a glimmer of hope and a potential focus area for retail targeting strategies. Ultimately, these findings emphasize the need for a nuanced understanding of demographic shifts and employment trends in crafting effective retail recovery plans post-pandemic.