SCC Client Report

Reproduce Holder et al. (2021) Results

2024-02-08

1 Background

Holder et al. published a paper titled "The Early Impact of Covid-19 on Job Losses among Black Women in the United States" [1]. Focusing on the early phase of the pandemic, they investigated the disproportionate job losses experienced by Black women in the United States, particularly in essential sectors of the economy. They used the United States Census Bureau's Basic Monthly CPS data [2] from February, March, and April 2020. Their quantitative analysis, grounded in feminist economic and stratification economic theories, identified cashier positions in the hotel and restaurant industry and childcare worker roles in the healthcare and social services sector as the occupations where Black women encountered the most significant employment declines. These occupations, predominantly low-wage and considered essential, were argued to be particularly vulnerable due to Black women's strong attachment to the workforce, their overrepresentation in these industries, and the broader trend of women being concentrated in low-wage occupations. The SCC has been asked to reproduce and critically assess Figures 1 through 4 in Holder et al. (2021).

2 Methods

2.1 Libraries

We used the `tidyverse` and `survey` packages in R to conduct our analysis and create visualizations. `dplyr` functions were used to wrangle data and `ggplot` was used to build plots. `svydesign()` and `svytable()` functions from the `survey` library were used to create objects for handling weighted survey data.

2.2 Data

Data was obtained from the United States Census Bureau's Basic Monthly CPS from the months of February, March, and April in 2020. The following variables were used:

Table 1: Variables used in the data set

Variable	Renamed	Type	Description	Role
`pwsswgt`	`personWeight`	integer	final weight	weighting survey data
`primind1`	`industry1`	nominal factor	industry	figure 1 & 2
`prdtocc1`	`occupation1`	nominal factor	occupation	figure 3 & 4

Variable	Renamed	Type	Description	Role
pesex	sex	nominal factor	male (1) or female (2)	filtered for female
ptdtrace	race	nominal factor	race	filtered for 'black only' or 'white and black'
prtage	age	integer	age	filtered for above 15-years-old
prempnot	employStatus	nominal factor	employment status	filtered for in labor force (employed and unemployed)
prempnot	hispanic	nominal factor	hispanic (1) or non-hispanic (2)	filtered for non-hispanic

They were renamed for convenience. `primind1` and `prdtocc1` refer to the respective industry and occupation for job 1. Since Holder et al. did not specify whether they only used job 1, job 2, or some combination of job 1 and job 2, we only used job 1 to avoid the problem of over-counting.

Respondents that reported their race as "Black only" or "White and Black," were treated as Black. We also filtered for respondents that identified as female, above the age of 15, in the labor force, and non-hispanic. All NAs were removed. We divided `pwsswgt` by 10,000 because the documentation said there were 4 implicit decimal places. This was done for all four figures.

3 Computations

3.1 Load Data & Packages

Here we load the data, packages, and labels we'll be using throughout the project.

```
## Load packages
library(tidyverse)
library(survey)

## Load data
# feb <- read.csv("feb20pub.csv")
# mar <- read.csv("mar20pub.csv")
# apr <- read.csv("apr20pub.csv")
```

```
## Database variable names:
variables <- c("pwsswgt","primind1","primind2", "prdtocc1", "prdtocc2",
               "pesex", "prtage", "ptdtrace","prempnot","pehspnon")

## Rename these for our convenience:
c.names <- c("personWeight","industry1", "industry2", "occupation1",
             "occupation2", "sex", "age", "race","employStatus","hispanic")

## 22 Industry code labels:
ind.labels <- c("Agriculture, forestry, fishing, and hunting",
                "Mining", "Construction", "Manufacturing - durable goods",
                "Manufacturing - non-durable goods", "Wholesale trade",
                "Retail trade", "Transportation and warehousing",
```

```
                "Utilities", "Information", "Finance and insurance",
                "Real estate and rental and leasing",
                "Professional and technical services",
                "Management, administrative and waste management services",
                "Educational services", "Health care and social services",
                "Arts, entertainment, and recreation",
                "Accommodation and food services",
                "Private households",
                "Other services, except private households",
                "Public administration", "Armed Forces")

## 23 Occupation code labels:
occ.labels <- c("Management", "Business and financial operations",
                "Computer and mathematical science",
                "Architecture and engineering",
                "Life, physical, and social science",
                "Community and social service", "Legal",
                "Education, training, and library",
                "Arts, design, entertainment, sports, and media",
                "Healthcare practitioner and technical",
                "Healthcare support", "Protective service",
                "Food preparation and serving related",
                "Building and grounds cleaning and maintenance",
                "Personal care and service", "Sales and related",
                "Office and administrative support",
                "Farming, fishing, and forestry",
                "Construction and extraction",
                "Installation, maintenance, and repair", "Production",
                "Transportation and material moving", "Armed Forces")
```

The `clean_data` function selects and renames the relavent variables. NAs are removed and `pwsswgt` is divided by 10,000.

```
clean_data <- function(df){
  df <- df[, variables] # get variables
  df <- df[!is.na(df$pwsswgt),] # remove NAs
  df$pwsswgt <- (df$pwsswgt/10000) # `4 implied decimal places'
  colnames(df) <- c.names # rename variables
  return(df)
}

feb <- clean_data(feb)
mar <- clean_data(mar)
apr <- clean_data(apr)
```

3.2 Figure 1

Figure 1 shows the share of Black women's employment by industry in February 2020. `feb20_svy` is a survey object that contains metadata to handle the weighted survey data. It was then subsetted to the

target population. This yielded 10,812,814 Black women.

```
## Create survey object
feb20_svy <- svydesign(ids=~1, data = feb, weights = ~personWeight)

feb20.sub <- subset(feb20_svy, ((race == 2)|(race == 6))
                    & (hispanic == 2) & (sex == 2) & (age >= 15))

## Preview industry1
# svytable(~industry1, feb20.sub)

## In labor force
feb.labor.sub <- subset(feb20.sub, (employStatus==1)|(employStatus==2))

## Size of target population represented by sample subset
sum(weights(feb.labor.sub,"sampling"))
```

```
[1] 10812814
```

`getIndPct` is a function that returns a dataframe with the industry name, count, and percentage of black women employment. Observations with `industry1 == -1` were removed. The formula for calculating percentage is $Pct = 100 \times \frac{IndustryCount}{sum(Count)}$.

```
## Function to calculate percentage
getIndPct <- function(target, svy_object){
  formula_target <- as.formula(paste("~", target)) # create formula
  indTable <- svytable(formula_target, svy_object) # create survey table
  indTable <- indTable[-1] # exclude -1
  names(indTable) <- ind.labels[as.numeric(names(indTable))] # industry labels
  indTablePct <- round(100*indTable/sum(indTable), 2) # calculate Pct
  tbl <- cbind(indTable, indTablePct) # combine columns
  colnames(tbl) <- c("Count", "Pct") # rename columns
  tbl <- tbl[order(tbl[,1]),] # rows in ascending order
  df <- as.data.frame(tbl) # convert to dataframe
  df$Industry <- rownames(df) # create column for industry labels
  rownames(df) <- NULL # remove row index names
  df <- df[, c(3, 1, 2)] # reorder columns
  return(df)
}

ind1_df <- getIndPct("industry1", feb.labor.sub)
```

`ind1_df` is the dataframe with the final calculations that were used to create Figure 1.

```
## Create Figure 1 plot
fig1 <- ind1_df |>
  ggplot(aes(x = reorder(Industry, Pct),
             y = Pct)) +
  geom_bar(stat = "identity", fill = "#7FA5F9") +
  geom_text(aes(label = sprintf("%.1f%%", Pct)),
```

```
            hjust = -0.25, color = "black", size = 2) +
  theme_minimal() +
  scale_y_continuous(breaks = seq(0, 30, by = 5),
                     limits = c(0, 30),
                     labels = sprintf("%.1f%%", seq(0, 30, 5))) +
  coord_flip() +
  labs(title = "Share of Black Women's Employment by Industry\n(Feb 2020)",
       x = "Industry",
       y = "Percentage") +
  theme(panel.grid.major.y = element_blank(),
        panel.grid.minor.x = element_blank(),
        axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y = element_text(size = 5),
        axis.text.x = element_text(size = 5),
        plot.title = element_text(size = 10))
```

3.3 Figure 2

Figure 2 shows the change in Black women's employment by industry from February to April 2020. We use `feb.labor.sub`, a survey object after filtering for the same parameters as explained in our Methods section. See Figure 1 code above. The same process was done to obtain `mar.labor.sub` and `apr.labor.sub`. We then create tables from the survey objects for each month and turn them into dataframes.

`merged_df` is the dataframe with Black women's employment by industry in February, March, and April 2020. We calculate the percent change using the standard percent change formula: $\frac{aprilpercentemployed - februarypercentemployed}{februarypercentemployed} \times 100$.

```
merged_df$percentage_diff <- round(((merged_df$febpct - merged_df$aprpct)
                                    / merged_df$aprpct) * 100, 1)

merged_df$count_diff <- ((merged_df$febcount - merged_df$aprcount)
                         / merged_df$aprcount) * 100
```

We plot our results in a bar plot.

```
fig2 <- merged_df |>
  ggplot(aes(x = percentage_diff, y = reorder(Industry,
                                              desc(percentage_diff)))) +
  geom_bar(stat = "identity", fill = "#7FA5F9") +
  geom_text(aes(label = sprintf("%.1f%%", percentage_diff)),
            vjust = ifelse(merged_df$percentage_diff >= 0, -1, 1.75),
            size = 1, color = "#212121") +  # Smaller text
  labs(title="Change in Black Women's Employment by Industry (Feb-Apr 2020)",
       x="Percentage",
       y="Industry") +
  theme_minimal() +
  coord_flip() +
  scale_x_continuous(breaks = seq(-125, 150, by = 25),
```

```
                    limits = c(-125, 150),
                    labels = sprintf("%.1f%%", seq(-125, 150, 25))) +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.y = element_blank(),
        axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y = element_text(size = 5),  # Smaller text on the y-axis
        axis.text.x = element_text(size = 5, angle = 45, hjust = 1),
        plot.margin = margin(0.5, 1, 0.5, 2, "cm"),
        plot.title = element_text(size = 10)
        )
```

3.4 Figure 3

Figure 3 shows the share of Black women's employment by occupation in February 2020. The same analysis as Figure 1 was done. `occupation1` was used in place of `industry1`. Observations with `occupation1 == -1` were removed. The formula for calculating percentage is $Pct = 100 \times \frac{OccupationCount}{sum(Count)}$.

`occ1_df` is the dataframe with the final calculations that were used to create Figure 3.

```
fig3 <- occ1_df |>
  ggplot(aes(x = reorder(Occupation, Pct),
             y = Pct)) +
  geom_bar(stat = "identity", fill = "#7FA5F9") +
  geom_text(aes(label = sprintf("%.1f%%", Pct)),
            hjust = -0.25, color = "black", size = 2) +
  theme_minimal() +
  scale_y_continuous(breaks = seq(0, 18, by = 2),
                     limits = c(0, 18),
                     labels = sprintf("%.1f%%", seq(0, 18, 2))) +
  coord_flip() +
  labs(title = "Share of Black Women's Employment by Occupation\n(Feb 2020)",
       x = "Industry",
       y = "Percentage") +
  theme(panel.grid.major.y = element_blank(),
        panel.grid.minor.x = element_blank(),
        axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y = element_text(size = 5),
        axis.text.x = element_text(size = 5),
        plot.title = element_text(size = 10))
```

3.5 Figure 4

Figure 4 displays the change in employment among Black women across various occupations from February to April 2020. Same as Figure 2, survey objects for February, March, and April were created. We then make a table for job loss count by occupation for February, March, and April. We then calculate change in employment: $\frac{apr\_pct - avg\_feb\_and\_mar\_pct}{avg\_feb\_and\_mar\_pct} \times 100$.

```r
avg_feb_mar <- (tbl2[, 2]+tbl3[, 2])/2
result <- (tbl4[, 2] - avg_feb_mar) / avg_feb_mar * 100

## Create a data frame with the result
result_df <- data.frame(percentage_diff = result)
result_df$Occupation <- rownames(result_df)
rownames(result_df) <- NULL
result_df <- result_df[, c(2, 1)]
```

`result_df` is the dataframe with the change in Black women's employment by occupation from February to April 2020. We graph our results below.

```r
fig4 <- ggplot(result_df, aes(x = reorder(Occupation, percentage_diff),
                              y = percentage_diff)) +
  geom_bar(stat = "identity", fill = "#7FA5F9") +
  geom_text(aes(label = sprintf("%.1f%%", percentage_diff)),
            vjust = ifelse(result_df$percentage_diff >= 0, -1, 1.75),
            size = 1, color = "#212121") +  # Smaller text +
  labs(title = "Change in Black Women's Employment by Occupation\n(February-April 2020)",
       x = "Occupation",
       y = "Percentage") +
  theme_minimal() +
  scale_y_continuous(breaks = seq(-60, 40, by = 20),
                     limits = c(-60, 40),
                     labels = sprintf("%.1f%%", seq(-60, 40, 20))) +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.y = element_blank(),
        axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y = element_text(size = 5),  # Smaller text on the y-axis
        axis.text.x = element_text(size = 5, angle = 45, hjust = 1),
        plot.margin = margin(0.5, 1, 0.5, 2, "cm"),
        plot.title = element_text(size = 10)
        )
```

4 Results

We compare our results with Holder et al.

4.1 Figure 1

Our reproduced graph somewhat mirrors Figure 1 in the publication, with an absolute margin of error not surpassing 0.6%. For instance, the industry with the 2nd greatest share of black women's employment is different. We calculated it was Accommodation and food services at 10.1% but Holder et al. calculated it was Education services at 10.3%. This could be due to the utilization of a different package.

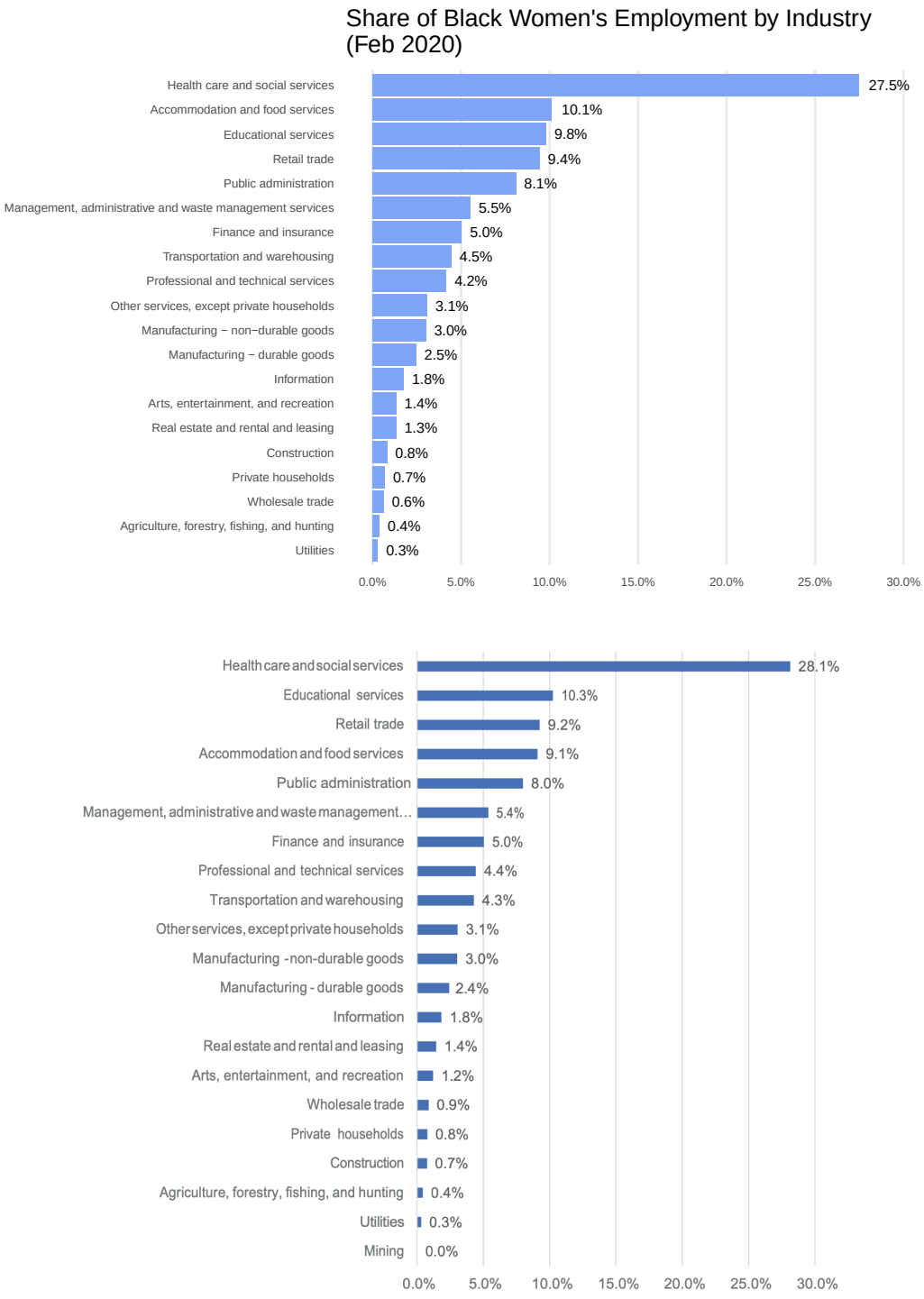

Figure 1: Holder et al. Figure 1

4.2 Figure 2

Our graphs had noticeable differences, with the percent change being positive for roughly half of the industries. The author's graph on the other hand was negative for all but Utilities and Mining.

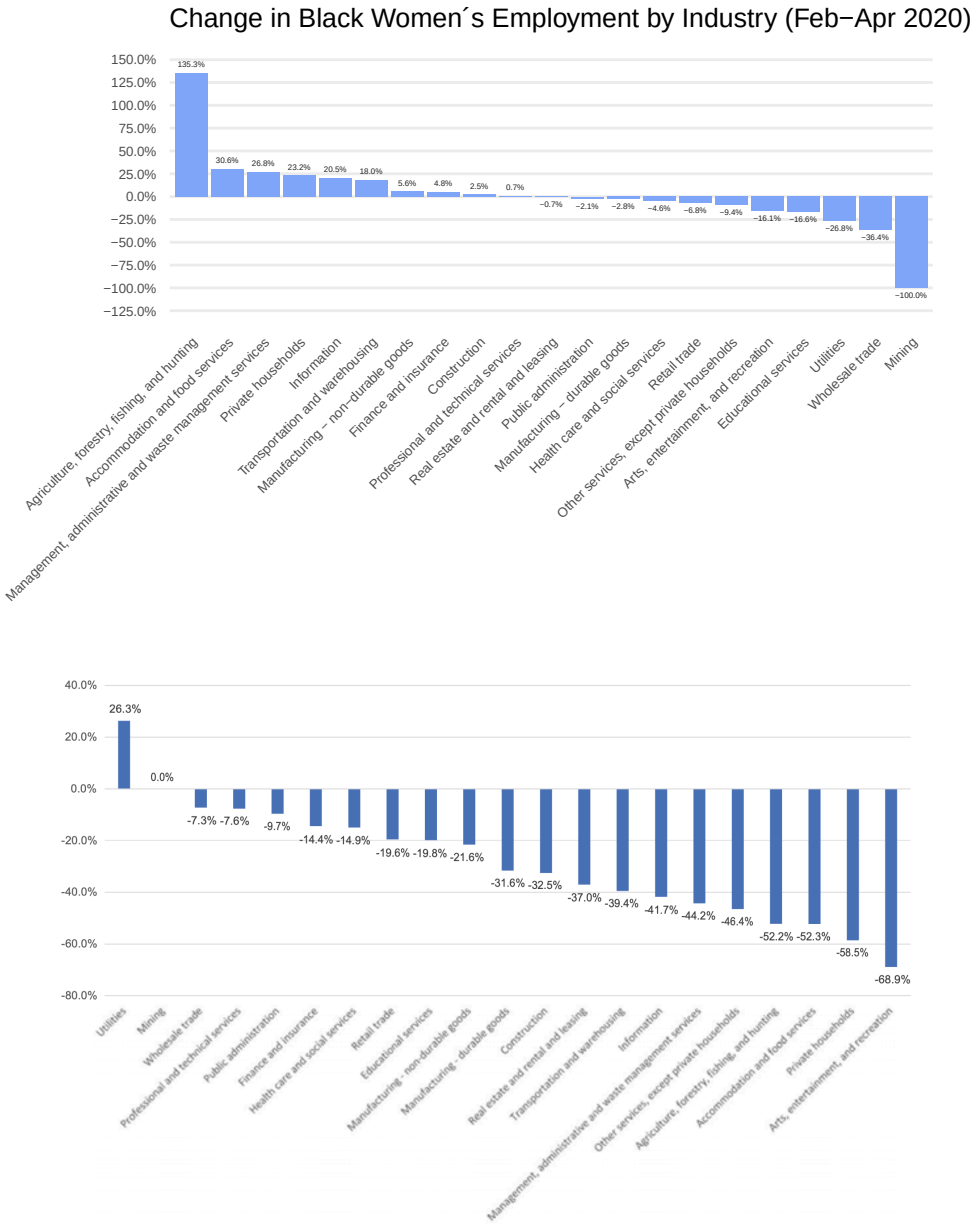

Figure 2: Holder et al. Figure 2

4.3 Figure 3

Only the percentages for Farming, fishing, and forestry, Construction and extraction, Healthcare practitioner and technical, Transportation material and moving, and Personal care and service matched that of Holder et al. The order of the occupations with the greatest to smallest share of Black women's employment is also slightly off.

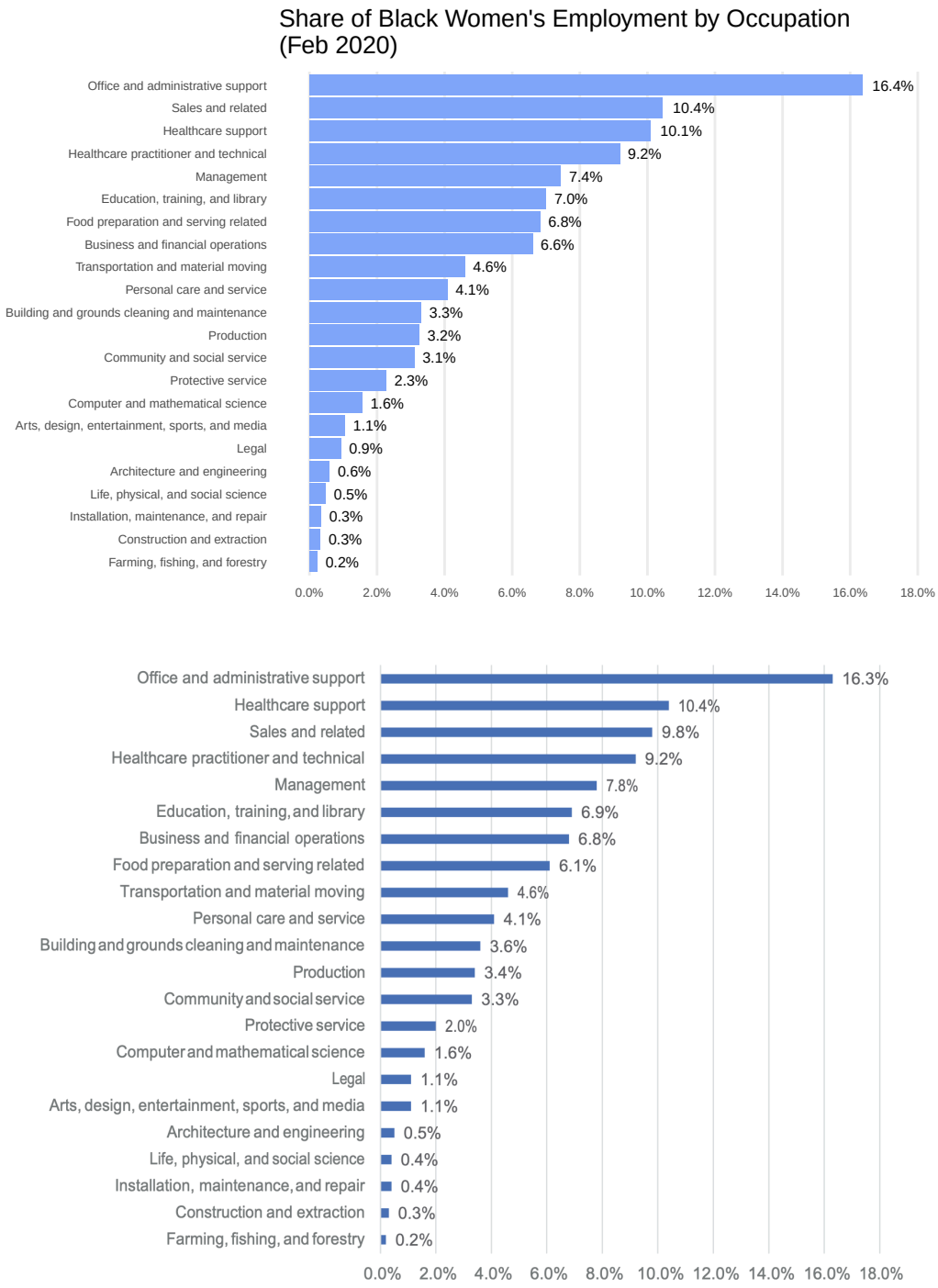

Figure 3: Holder et al. Figure 3

4.4 Figure 4

The resultant graph exhibits disparities from the original graph featured in the article, particularly in occupation categories such as Construction and Extraction, Life, Physical, and Social Science, and Legal. It is hypothesized that these discrepancies may arise from variations in the percentage formulas employed. Unfortunately, the original publication lacks supplementary information elucidating the specific percentage formula employed.

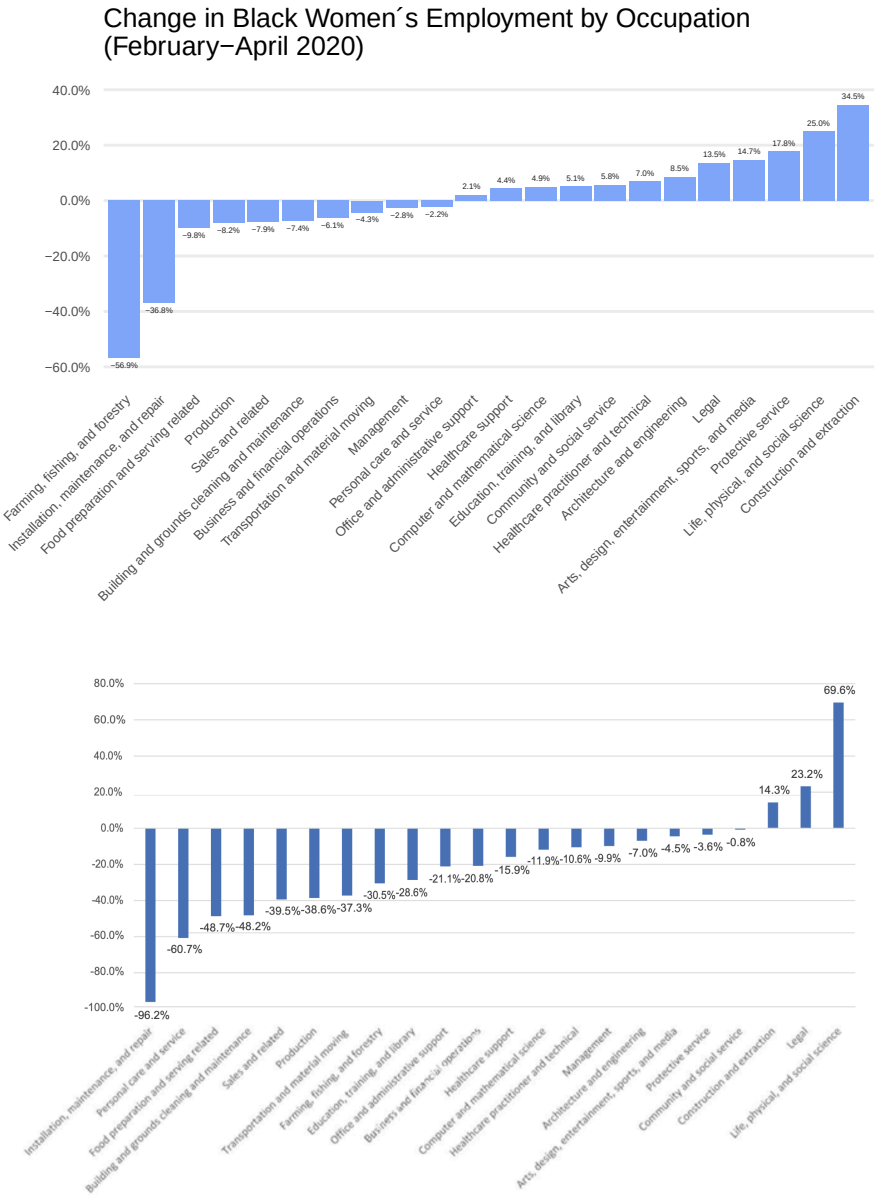

Figure 4: Holder et al. Figure 4

5 Conclusion

5.1 Assumptions

We had to make assumptions about the formulas for calculating the percentages and the usage of Job 1. We also attempted just on Job 2 and adding the counts for Job 1 and Job 2 but the results were not any closer to Holder et al.

5.2 Issues with Reproducibility

To enhance the reproducibility of the publication, it would be advantageous if the authors could disclose the specific package utilized and the weighted variable employed for survey data weighting. This trans-

parency would contribute to a more comprehensive understanding and facilitate the accurate replication of the study's findings.

We were unsure how the author defined "percent change," as they only stated that their graph's values were from " Author's calculations based on CPS Basic Monthly Data for February, March, and April 2020". Looking more closely at the percent employment rate for Agriculture, forestry, fishing, and hunting, for example, we found that our percent employment for February 2020 matched (both being 0.4%). However, the author didn't specify the percent employment for April 2020 in the paper or the article. Using the CPS data, we found that the percent employment was 0.89%, which doesn't match the author's - 52.2% decrease in employment logically or with the standard percent change formula. We were also unsure how the author incorporated data from March 2020 to calculate the percent change from February to April.

Another potential factor contributing to the variance in data could be disparities in our calculated March and April figures compared to the original data. However, this seems less plausible, as our calculated February data aligns with the original February data as presented in Appendix Table C, detailing the share of employment and the change in employment between February and April 2020.

5.3 Recommendations

1. Specify formulas used for each figure - this will ensure that other researchers are able to reproduce results and perform the correct calculations.
2. Specify how data is filtered (and by what parameters) - this will ensure that researchers are working with the same dataset.
3. Publish the codebase used to find results - this will allow other researchers to easily compare methodologies and reproduce a similar result.
4. Specify packages used for the analysis - this will avoid creating differences that are due to changing software.

6 References

[1] Holder et al. (2021), The Early Impact of Covid-19 on Job Losses among Black Women in the United States. https://doi.org/10.1080/13545701.2020.1849766

[2] Basic Monthly CPS. https://www.census.gov/data/datasets/time-series/demo/cps/cps-basic.2020.html#list-tab-1979780401