February Survey Report

Maksim Zubok

Date

Intro

To create weights, I am working with the 2020 census data, particularly the cross tabbed gender, age, and university education file here.

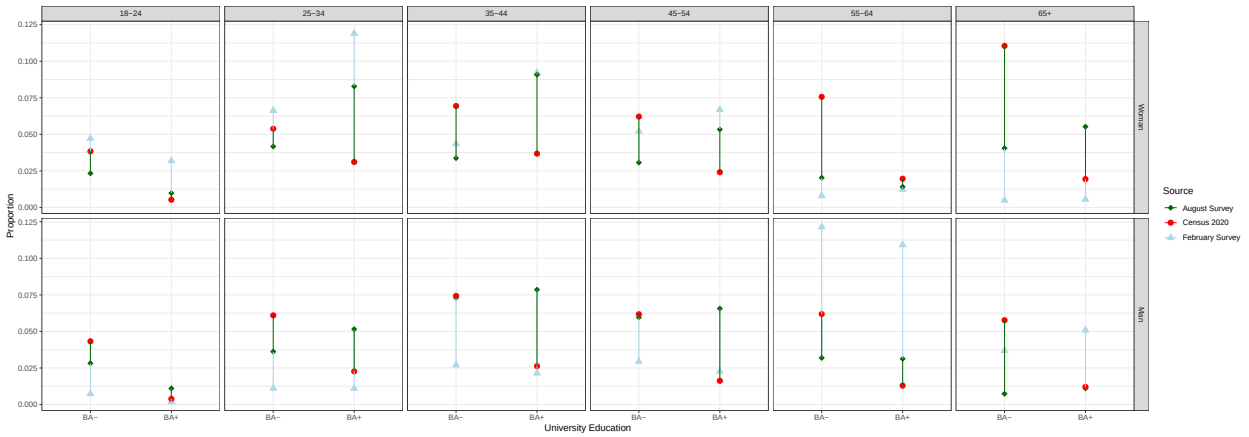
I did not include information about region of residence, even though we could do it after harmonising census data with the survey.

```
kable(head(ru_population_frame, 5),
      caption = "Population Frame: Census 2020",
      align = "c",
      format = "markdown")
```

Table 1: Population Frame: Census 2020

gender	age_group	university_education	Freq
Man	18-24	BA+	463546
Man	18-24	BA-	5164261
Man	25-34	BA+	2707302
Man	25-34	BA-	7287296
Man	35-44	BA+	3136411

Sample to Population Comparison



The **main disparities between the February and the 2020 Census* are:

- The biggest disparities are is women 55-64 and 65+ without education. The survey has 0‘ women in the latter category.
- We have oversampled young women, especially 25-34 years old with university education, and undersampled men without university education across all age categories except 55- and 65+.

Weights witth Survey package

To compute post-stratification weights we rely on the `postStratify` function from the `survey` package. The function adjusts the sampling and replicate weights so that the joint distribution of a set of post-stratifying variables matches the known population joint distribution. **However, the package documentation does not describe how exactly the adjustment is implemented.**

```
## survey library ##
unweighted_data <- svydesign(ids = ~1, data = survey_feb)


weighted <- postStratify(unweighted_data, ~age_group + gender + university_education,
             ru_population_frame, partial=TRUE)

# save weights
survey_feb$weight_poststratify <- weights(weighted)

sum_feb <- round(summary(weights(weighted)), 2)

sum_feb_mat <-  matrix(as.numeric(sum_feb), nrow = 1,
                           dimnames = list(c("Value"),
                           names(sum_feb)
                           )
                           )

kable(sum_feb_mat,
      caption = "February Survey PostStratify Weights Summary",
      align = 'c',
      format = "markdown")
```

Table 2: February Survey PostStratify Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	8670.35	19144.08	37348.11	73292.34	87350.22	1649629

Note: some strata had no observations in the survey (NA on education for some age gender groups). This means we had to ignore them in producing weights.

```
## survey library ##
unweighted_data <- svydesign(ids = ~1, data = survey_aug)


weighted <- postStratify(unweighted_data, ~age_group + gender + university_education,
             ru_population_frame, partial=TRUE)

# save weights
```

```r
survey_aug$weight_poststratify <- weights(weighted)

sum_aug <- round(summary(weights(weighted)), 2)

sum_aug_mat <-  matrix(as.numeric(sum_aug), nrow = 1,
                       dimnames = list(c("Value"),
                       names(sum_aug)
                       )
                       )

kable(sum_aug_mat,
      caption = "August Survey PostStratify Weights Summary",
      align = 'c',
      format = "markdown")
```

Table 3: August Survey PostStratify Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	18133.48	27510.75	33083.25	74666.57	112266.5	574250.5

For August, we also see that some weights are much larger than others. As you can see in the graphs below, the distribution of weights is similarly skewed and the disparities between the bulk of the distribution and its tales are in the same orders of magnitude. However, the largest weight in Feb survey is three times bigger than the largest weight in Aug survey.

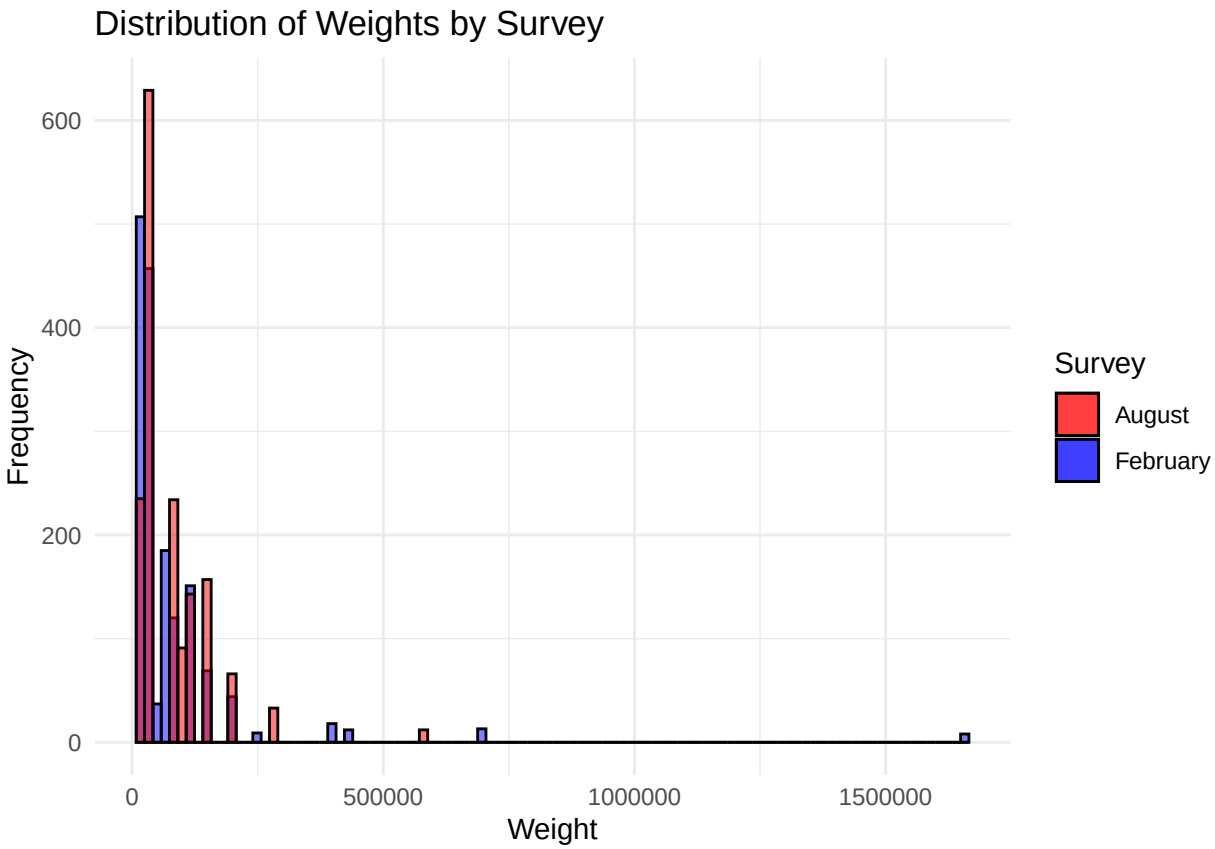
The largest weights in both surveys relate to different population groups.

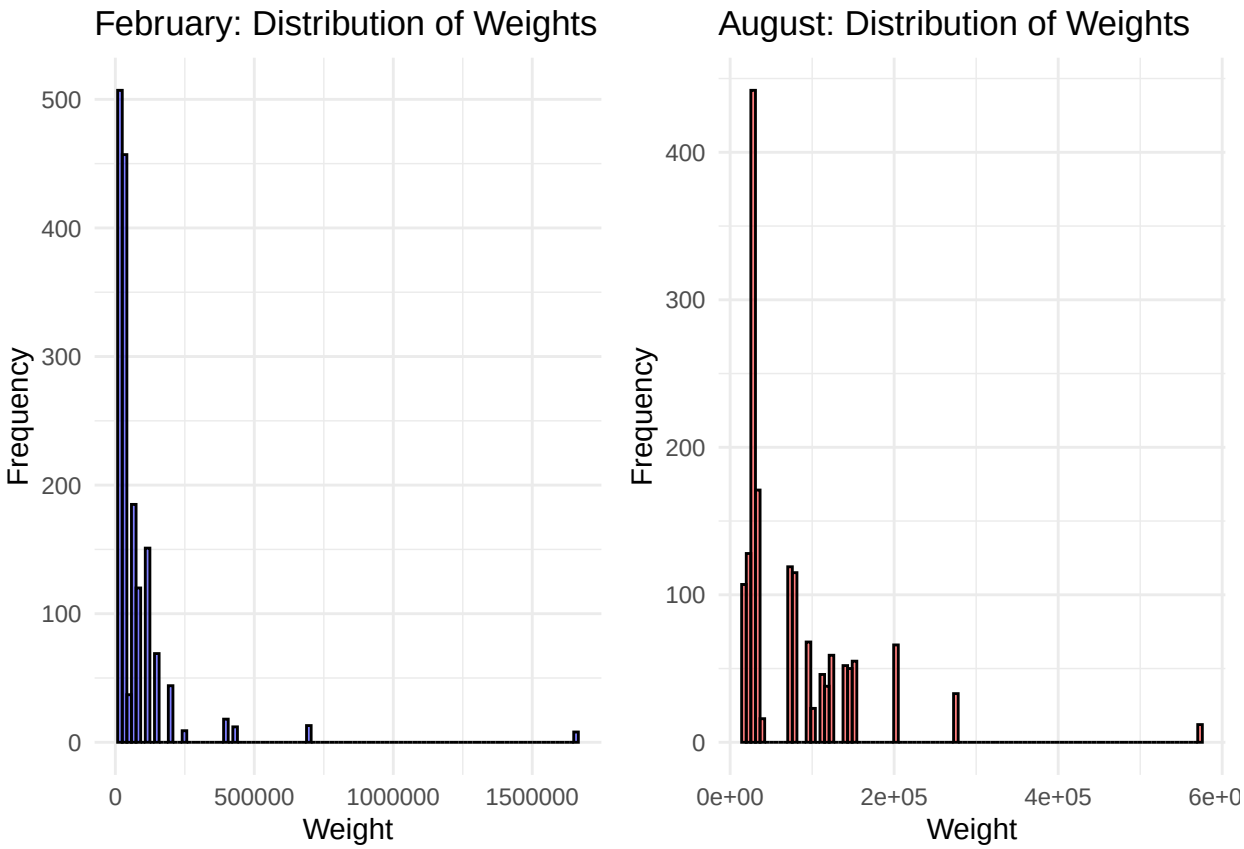
Table 4: February Survey, Top Five Rows by Unique Weight

age_group	gender	university_education	weight_poststratify
65+	Woman	BA-	1649629.1
55-64	Woman	BA-	695189.1
18-24	Man	BA-	430355.1
25-34	Man	BA-	404849.8
65+	Woman	BA+	256902.0

Table 5: August Survey, Top Five Rows by Weight

age_group	gender	university_education	weight_poststratify
65+	Man	BA-	574250.5
55-64	Woman	BA-	273862.4
65+	Woman	BA-	199955.0
35-44	Woman	BA-	150708.4
45-54	Woman	BA-	148495.4


Distribution of Weights by Survey
Frequency
600
400
200
0
0
500000
1000000
1500000
Weight
Survey
August
February



Weights created manually

To check the plausibility of resulting weights, we create alternative weights based on the population frequencies of the combination of the same strata (Yana's approach). The weights are calculated for each category:

$$\text{weight}_i = \frac{\text{population frequency}_i}{\text{sample frequency}_i}$$

```
# calculate weights and print the df with weights
weights_aug_strata_man <- left_join(survey_aug_strata,
                           pop_strata,
                           c("gender", "age_group", "university_education")) %>%
  rename(population_proportion = proportion.y,
         sample_proportion = proportion.x)  %>%
  # calculate weights as popul prop/sample prop
  mutate(weight = population_proportion / sample_proportion)

sum_man_aug <- round(summary(weights_aug_strata_man$weight), 2)

sum_man_aug_mat <-  matrix(as.numeric(sum_man_aug), nrow = 1,
                           dimnames = list(c("Value"),
                           names(sum_man_aug)
                           )
                           )
```

```
kable(sum_man_aug_mat,
      caption = "August Survey Weights Summary",
      align = 'c',
      format = "markdown")
```

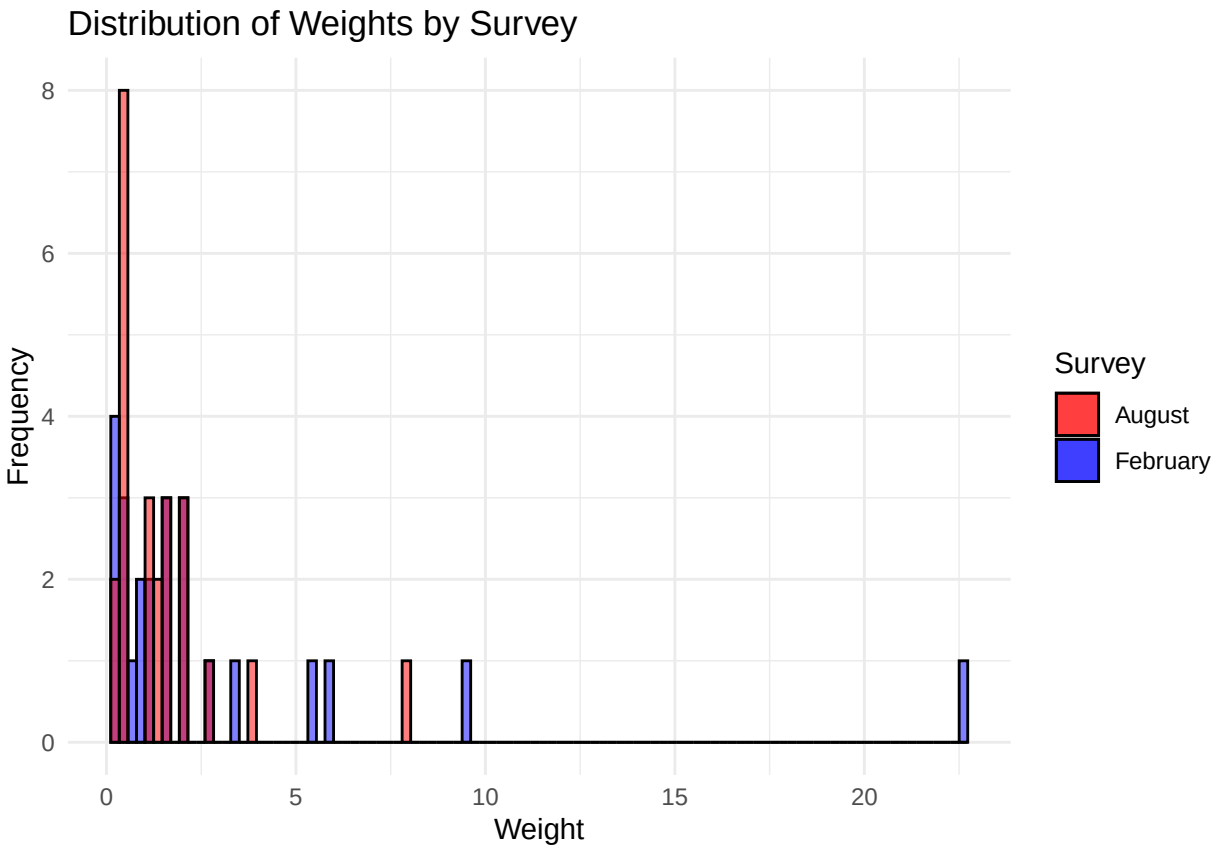
Table 6: August Survey Weights Summary

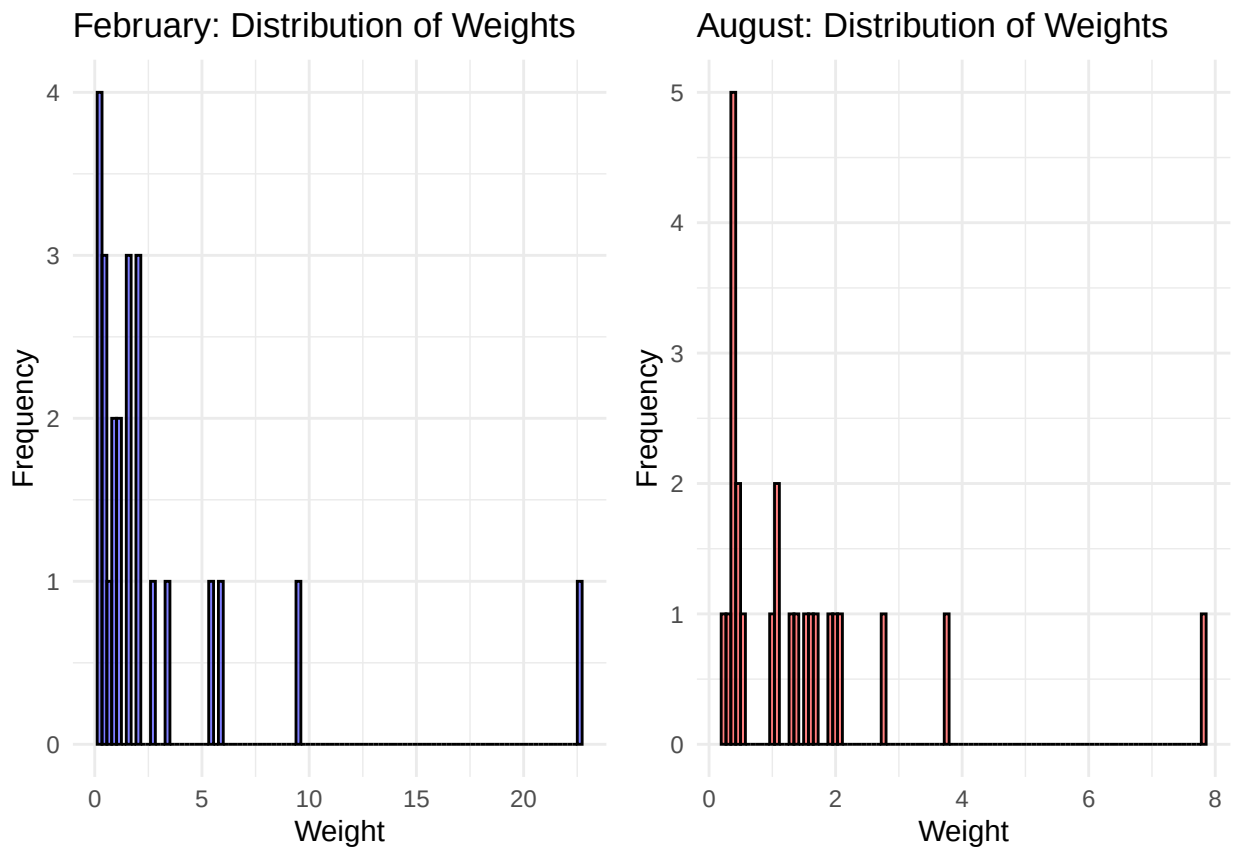
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	0.25	0.41	1.07	1.46	1.75	7.84

```
# calculate weights and print the df with weights
weights_feb_strata_man <- left_join(survey_feb_strata,
                            pop_strata,
                            c("gender", "age_group", "university_education")) %>%
  rename(population_proportion = proportion.y,
         sample_proportion = proportion.x)  %>%
  # calculate weights as popul prop/sample prop
  mutate(weight = population_proportion / sample_proportion)

sum_man_feb <- round(summary(weights_feb_strata_man$weight), 2)

sum_man_feb_mat <-  matrix(as.numeric(sum_man_feb), nrow = 1,
                           dimnames = list(c("Value"),
                           names(sum_man_feb)
                           )
                           )

kable(sum_man_feb_mat,
      caption = "February Survey Weights Summary",
      align = 'c',
      format = "markdown")
```

Table 7: February Survey Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	0.12	0.48	1.39	2.81	2.27	22.51


Distribution of Weights by Survey
Frequency
Weight
0
2
4
6
8
0
5
10
15
20
Survey
August
February



Calculating weights this way gives us the same overall picture. The main difference is in the orders of magnitude between the median weight and the tail (highest weight).

For **February** survey, the difference (max weight / median weight) with `postStratify` is **44.17**, but with simple manually created weights it is **16.19**. For **August** survey, the difference with `postStratify` is **17.36**, but with simple manually created weights it is **7.33**.

Note that some observations in February survey have a weight of zero because they do not exist in the population census. These are women 18-19 years old with university education. My suspicion on why we have these observations is that some respondents might have misreported their education attainment.

```
# survey_feb <- survey_feb %>%
#   left_join(weights_feb_strata_man, by = c("age_group", "gender", "university_education")) %>%
#   rename(weight_manually_calculated = weight) %>%
#   select(-Survey)

# write.csv(survey_feb, "data/surveys/survey_feb_weights.csv", row.names = FALSE)
```