506 final report

Sunny Ma

2025-12-14

Setup

Data Preprocessing

```
data_raw <- read_csv("nonprofit-survey-spring-2021-puf.csv")
dim(data_raw)
```

```
[1] 2306  359
```

Variable Selection

Since there are so many variables in our dataset, after a thorough look through the dataset, The following varriables are considered to be the selected covariants. 1. Organizational variables: a. the location (CENSUSREGION4) b. the size, defined by the number of full-time staff (STAFF_1_1_1) c. the sector (NTMAJ5) 2. Leader's demographic variables: a. sexuality (CEOBCdem_1_1 CEOBCdem_1_2) b. race (CEORACE BCRACE) c. gender (CEOGENDER BCGENDER)

```
selected_df <- data_raw %>%
  select(
    org_id = ResponseId,
    weight = weight_complete_only,

    #Dependent Variable
    dv_lgbtq_focus = ProgDem_19,

    #Independent Variables
```

```
    #Organizational Controls
    cv_region = CensusRegion4,    # Census Region
    cv_staff_ft = Staff_1_1_1,    # Full-time Staff Size (as organization size)
    cv_sector = ntmaj5,      # Major Sector

    #Leader Demographics

    #Sexuality
    cv_ceo_lgbtq = CEOBCdem_1_1,
    cv_chair_lgbtq = CEOBCdem_1_2,

    #Age
    iv_ceo_age = CEOBCdem_3_1,
    iv_chair_age = CEOBCdem_3_2,

    #Race
    cv_ceo_race = CEOrace,
    cv_chair_race = BCrace,

    #Gender
    cv_ceo_gender = CEOgender,
    cv_chair_gender = BCgender
  )
```

Data Cleansing

```
young_codes = c(1, 2, 3) #<45
non_cismale_codes = c(2, 3, 4) #female, non-binary, trans

data_clean <- selected_df %>%
  mutate(
    #Dependent Variable (DV): LGBTQ+ Focus
    dv_lgbtq_focus = case_when(
      dv_lgbtq_focus %in% c(1, 2) ~ 1,
      is.na(dv_lgbtq_focus) ~ NA_real_,
      TRUE ~ 0
    ),
    dv_lgbtq_focus = factor(dv_lgbtq_focus, levels = c(0, 1), labels = c("No Focus", "LGBTQ

    #Independent Variable (IV): Younger Leadership
```

```
  younger_ceo = if_else(iv_ceo_age %in% young_codes,1,0),
  younger_chair = if_else(iv_chair_age %in% young_codes,1,0),

  iv_young_leadership = case_when(younger_ceo == 1 | younger_chair == 1 ~ 1,
                                  is.na(younger_ceo) & is.na(younger_chair) ~ NA_real_,
                                  TRUE ~ 0),
  iv_young_leadership = factor(iv_young_leadership, levels = c(0,1), labels  = c("Older (45
) %>%
#Control Variables (Covariates)
mutate(
  # Log Transformation
  cv_staff_ft = if_else(cv_staff_ft < 0, NA_real_, cv_staff_ft),
  log_staff = log10(cv_staff_ft + 1),

  cv_region = as.factor(cv_region),  #region
  cv_sector = as.factor(cv_sector), # sector

  # race
  poc_ceo = if_else(cv_ceo_race!= 5 & cv_ceo_race > 0, 1, 0),
  poc_chair = if_else(cv_chair_race != 5 & cv_chair_race > 0, 1, 0),
  cv_leadership_race = case_when(
    poc_ceo == 1 | poc_chair == 1 ~ 1,
    is.na(cv_ceo_race) & is.na(cv_chair_race) ~ NA_real_,
    TRUE ~ 0
  ),
  cv_leadership_race = factor(cv_leadership_race, levels = c(0, 1), labels = c("All White"
  # Sexuality

  lgbtq_ceo = if_else(cv_ceo_lgbtq == 1, 1, 0),
  lgbtq_chair = if_else(cv_chair_lgbtq == 1 , 1, 0),
  cv_leadership_lgbtq = case_when(
    lgbtq_ceo == 1 | lgbtq_chair == 1 ~ 1,
    is.na(cv_ceo_lgbtq) & is.na(cv_chair_lgbtq) ~ NA_real_,
    TRUE ~ 0
  ),
  cv_leadership_lgbtq = factor(cv_leadership_lgbtq, levels = c(0, 1),labels = c("No LGBTQ+

  # Gender
  non_cismale_ceo = if_else(cv_ceo_gender %in% non_cismale_codes,1,0),
  non_cismale_chair = if_else(cv_chair_gender %in% non_cismale_codes,1,0),
  cv_noncismale_leadership = case_when(non_cismale_ceo == 1| non_cismale_chair == 1 ~ 1, is
  cv_noncismale_leadership = factor(cv_noncismale_leadership,levels = c(0,1), labels = c("A
```

```
  ) %>%
  drop_na(weight, dv_lgbtq_focus,iv_young_leadership, cv_region, cv_sector, cv_leadership_ra
```

```
dim(data_clean)
```

```
[1] 1688   27
```

EDA

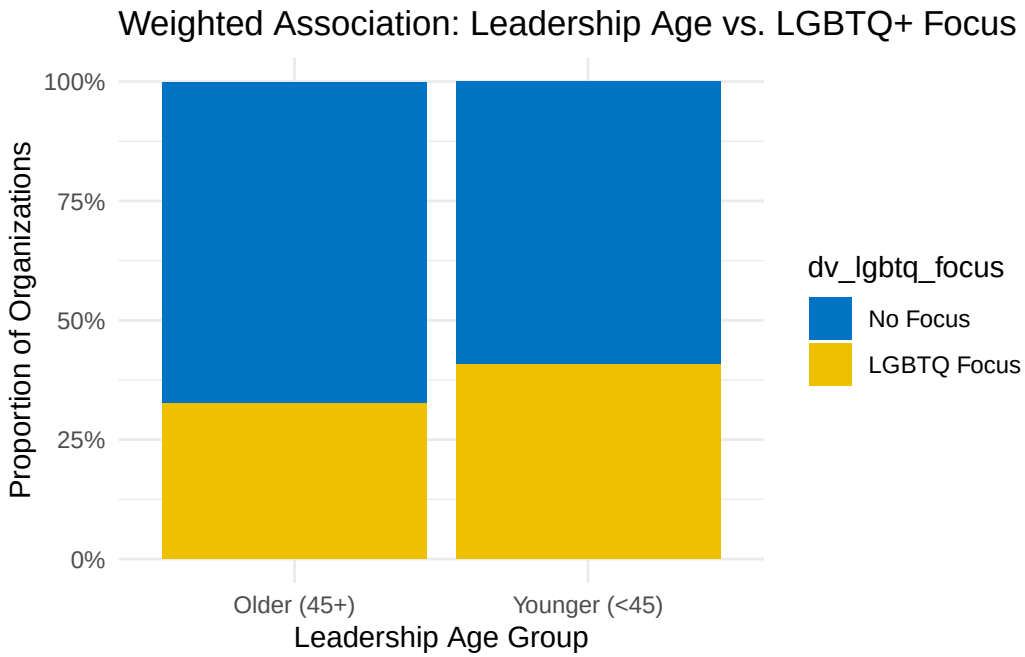
Since this survey has weights, Gemini AI taught me to deals with the weighted data using "survey" and "srvyr" packages.

```
svy_design <- data_clean %>% as_survey_design(ids = 1, weights = weight)
svy_design %>%
  summarise(
    unweighted_n = n(),
    weighted_n = survey_total(vartype = NULL)
  )
```

```
# A tibble: 1 x 2
  unweighted_n weighted_n
         <int>      <dbl>
1         1688      1548.
```

```
library(ggplot2)
library(ggsci)

svy_design %>%
  group_by(iv_young_leadership, dv_lgbtq_focus) %>%
  summarise(prop = survey_mean(), .groups = "drop") %>%
  ggplot(aes(x = iv_young_leadership, y = prop, fill = dv_lgbtq_focus)) +
  geom_col(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Weighted Association: Leadership Age vs. LGBTQ+ Focus",
    y = "Proportion of Organizations",
    x = "Leadership Age Group"
  ) +
  theme_minimal() +
  scale_fill_jco()
```



Modeling

```
final_model <- svyglm(
  dv_lgbtq_focus ~ iv_young_leadership +
                   log_staff +
                   cv_region +
                   cv_sector +
                   cv_leadership_race +
                   cv_leadership_lgbtq +
                   cv_noncismale_leadership,
  design = svy_design,
  family = binomial()
)

summary(final_model)
```

```
Call:
svyglm(formula = dv_lgbtq_focus ~ iv_young_leadership + log_staff +
    cv_region + cv_sector + cv_leadership_race + cv_leadership_lgbtq +
    cv_noncismale_leadership, design = svy_design, family = binomial())
```

```
Survey design:
Called via srvyr

Coefficients:
                                                    Estimate Std. Error t value
(Intercept)                                         -1.46861    0.23149  -6.344
iv_young_leadershipYounger (<45)                     0.25457    0.12699   2.005
log_staff                                            0.31086    0.11836   2.626
cv_region2: Midwest                                  0.09219    0.19233   0.479
cv_region3: South                                    0.04062    0.18349   0.221
cv_region4: West                                     0.25560    0.18376   1.391
cv_sectorED                                          0.06394    0.27229   0.235
cv_sectorHE                                          0.72080    0.26784   2.691
cv_sectorHU                                          0.11880    0.15462   0.768
cv_sectorOT                                         -0.56670    0.18407  -3.079
cv_leadership_raceBIPOC Representation               0.44965    0.12978   3.465
cv_leadership_lgbtqLGBTQ+ Representation             0.67210    0.18365   3.660
cv_noncismale_leadershipNon Cis-Male Representation  0.32023    0.14557   2.200
                                                    Pr(>|t|)
(Intercept)                                         3.05e-10 ***
iv_young_leadershipYounger (<45)                    0.045204 *
log_staff                                           0.008727 **
cv_region2: Midwest                                 0.631756
cv_region3: South                                   0.824846
cv_region4: West                                    0.164473
cv_sectorED                                         0.814380
cv_sectorHE                                         0.007210 **
cv_sectorHU                                         0.442430
cv_sectorOT                                         0.002122 **
cv_leadership_raceBIPOC Representation              0.000547 ***
cv_leadership_lgbtqLGBTQ+ Representation            0.000262 ***
cv_noncismale_leadershipNon Cis-Male Representation 0.027986 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.013564)

Number of Fisher Scoring iterations: 4
```