GSS Analysis Report

Qiyu Huang& Yuhan Zhu

10/19/2020

Abstract

We are going to take a analysis on th Canadian General Social Survey. We have used logistic regression and linear regression to see relationships. Deetails are shown later in the discussion part.

Introduction

We want to find if people are older, will they have more kids, more relationships, and have early kids when they are young? We also want to find if the income level is related with gender, with total number of children, and with satisfaction of their lives. Will the family be richer if they have more kids or poorer is what we want to find out as well. ## Data

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-li
brary/4.0'
## (as 'lib' is unspecified)
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-li
brary/4.0'
## (as 'lib' is unspecified)
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-li
brary/4.0'
## (as 'lib' is unspecified)

## Warning in install.packages("readr", repos = "http://cran.us.r-proje
ct.org"):
## installation of package 'readr' had non-zero exit status

## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-li
brary/4.0'
## (as 'lib' is unspecified)

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

## ── Attaching packages ─────────────────────────────────────── tidyve
rse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ purrr   0.3.4
## ✓ tibble  3.0.4     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.0

## ── Conflicts ──────────────────────────────────────── tidyverse_co
nflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data cleaning Part

importing raw data

raw_data <- read_csv("/cloud/project/AAGe4G0U.csv") dict <- read_lines("gss_dict.txt", skip = 18) labels_raw <- read_file("gss_labels.txt")

set up dictionary

variable_descriptions <- as_tibble(dict) %>% filter(value!="}") %>% mutate(value = str_replace(value, ".+%[0-9].*f[ ]{2,}","")) %>% mutate(value = str_remove_all(value,"""")) %>% rename(variable_description = value) %>% bind_cols(tibble(variable_name = colnames(raw_data)[-1]))

variable names and values

labels_raw_tibble <- as_tibble(str_split(labels_raw, ";")[[1]]) %>% filter(row_number()!=1) %>% mutate(value = str_remove(value, "define")) %>% mutate(value = str_replace(value, "[ ]{2,}", "XXX")) %>% mutate(splits = str_split(value, "XXX")) %>% rowwise() %>% mutate(variable_name = splits[1], cases = splits[2]) %>% mutate(cases = str_replace_all(cases, "", "")) %>% select(variable_name, cases) %>% drop_na()

variable name

labels_raw_tibble <- labels_raw_tibble %>% mutate(splits = str_split(cases, "[ ]{0,}"[ ]{0,}"))

creating a function

add_cw_text <- function(x, y){ if(!is.na(as.numeric(x))){ x_new <- paste0(y, "==", x,"~") } else{ x_new <- paste0(""",x,"",") } return(x_new) } ### Another function cw_statements <- labels_raw_tibble %>% rowwise() %>% mutate(splits_with_cw_text = list(modify(splits, add_cw_text, y = variable_name))) %>% mutate(cw_statement = paste(splits_with_cw_text, collapse = "")) %>% mutate(cw_statement = paste0("case_when(",

cw_statement,”TRUE~"NA")“)) %>% mutate(cw_statement = str_replace(cw_statement,”,"",“,”,")) %>% select(variable_name, cw_statement)

Do some final cleans of this function

cw_statements <- cw_statements %>% mutate(variable_name = str_remove_all(variable_name, “\r”)) %>% mutate(cw_statement = str_remove_all(cw_statement, “\r”))

Apply that dictionary to the raw data

gss <- raw_data %>% select(CASEID, agedc, achd_1c, achdmpl, totchdc, acu0c, agema1c, achb1c, rsh_131a, arretwk, slm_01, sex, brthcan, brthfcan, brthmcan, brthmacr, brthprvc, yrarri, prv, region, luc_rst, marstat, amb_01, vismin, alndimmg, bpr_16, bpr_19, ehg3_01b, odr_10, livarr12, dwelc, hsdsizec, brthpcan, brtpprvc, visminpr, rsh_125a, eop_200, uhw_16gr, lmam_01, acmpryr, srh_110, srh_115, religflg, rlr_110, lanhome, lan_01, famincg2, ttlincg2, noc1610, cc_20_1, cc_30_1, ccmoc1c, cor_031, cor_041, cu0rnkc, pr_cl, chh0014c, nochricc, grndpa, gparliv, evermar, ma0_220, nmarevrc, ree_02, rsh_131b, rto_101, rto_110, rto_120, rtw_300, sts_410, csp_105, csp_110a, csp_110b, csp_110c, csp_110d, csp_160, fi_110) %>% mutate_at(vars(agedc:fi_110), .funs = funs(ifelse(.>=96, NA, .))) %>% mutate_at(.vars = vars(sex:fi_110), .funs = funs(eval(parse(text = cw_statements %>% filter(variable_name==deparse(substitute(.))) %>% select(cw_statement) %>% pull()))))

Change the attributes name

gss <- gss %>% clean_names() %>% rename(age = agedc, age_first_child = achd_1c, age_youngest_child_under_6 = achdmpl, total_children = totchdc, age_start_relationship = acu0c, age_at_first_marriage = agema1c, age_at_first_birth = achb1c, distance_between_houses = rsh_131a, age_youngest_child_returned_work = arretwk, feelings_life = slm_01, sex = sex, place_birth_canada = brthcan, place_birth_father = brthfcan, place_birth_mother = brthmcan, place_birth_macro_region = brthmacr, place_birth_province = brthprvc, year_arrived_canada = yrarri, province = prv, region = region, pop_center = luc_rst, marital_status = marstat, aboriginal = amb_01, vis_minority = vismin, age_immigration = alndimmg, landed_immigrant = bpr_16, citizenship_status = bpr_19, education = ehg3_01b, own_rent = odr_10, living_arrangement = livarr12, hh_type = dwelc, hh_size = hsdsizec, partner_birth_country = brthpcan, partner_birth_province = brtpprvc, partner_vis_minority = visminpr, partner_sex = rsh_125a, partner_education = eop_200, average_hours_worked = uhw_16gr, worked_last_week = lmam_01, partner_main_activity = acmpryr, self_rated_health = srh_110, self_rated_mental_health = srh_115, religion_has_affiliation = religflg, regilion_importance = rlr_110, language_home = lanhome, language_knowledge = lan_01, income_family = famincg2, income_respondent = ttlincg2, occupation = noc1610, childcare_regular = cc_20_1, childcare_type = cc_30_1, childcare_monthly_cost = ccmoc1c, ever_fathered_child = cor_031, ever_given_birth

= cor_041, number_of_current_union = cu0rnkc, lives_with_partner = pr_cl, children_in_household = chh0014c, number_total_children_intention = nochricc, has_grandchildren = grndpa, grandparents_still_living = gparliv, ever_married = evermar, current_marriage_is_first = ma0_220, number_marriages = nmarevrc, religion_participation = ree_02, partner_location_residence = rsh_131b, full_part_time_work = rto_101, time_off_work_birth = rto_110, reason_no_time_off_birth = rto_120, returned_same_job = rtw_300, satisfied_time_children = sts_410, provide_or_receive_fin_supp = csp_105, fin_supp_child_supp = csp_110a, fin_supp_child_exp = csp_110b, fin_supp_lump = csp_110c, fin_supp_other = csp_110d, fin_supp_agreement = csp_160, future_children_intention = fi_110)

Clean up

gss <- gss %>% mutate_at(vars(age:future_children_intention), .funs = funs(ifelse(.=="Valid skip"|.=="Refusal"|.=="Not stated", "NA", .))) gss <- gss %>% mutate(is_male = ifelse(sex=="Male", 1, 0)) gss <- gss %>% mutate_at(vars(fin_supp_child_supp:fin_supp_other), .funs = funs(case_when(.=="Yes"1, .=="No"0, .=="NA"~as.numeric(NA)))) main_act <- raw_data %>% mutate(main_activity = case_when(mpl_105a=="Yes"~ "Working at a paid job/business", mpl_105b=="Yes" ~ "Looking for paid work", mpl_105c=="Yes" ~ "Going to school", mpl_105d=="Yes" ~ "Caring for children", mpl_105e=="Yes" ~ "Household work", mpl_105i=="Yes" ~ "Other", TRUE~ "NA")) %>% select(main_activity) %>% pull() age_diff <- raw_data %>% select(marstat, aprcu0c, adfgrma0) %>% mutate_at(.vars = vars(aprcu0c:adfgrma0), .funs = funs(eval(parse(text = cw_statements %>% filter(variable_name==deparse(substitute(.))) %>% select(cw_statement) %>% pull())))) %>% mutate(age_diff = ifelse(marstat=="Living common-law", aprcu0c, adfgrma0)) %>% mutate_at(vars(age_diff), .funs = funs(ifelse(.=="Valid skip"|.=="Refusal"|.=="Not stated", "NA", .))) %>% select(age_diff) %>% pull() gss <- gss %>% mutate(main_activity = main_act, age_diff = age_diff) gss <- gss %>% rowwise() %>% mutate(hh_size = str_remove(string = hh_size, pattern = "\ ."*)) %>% mutate(hh_size = case_when(hh_size=="One" ~ 1, hh_size=="Two" ~ 2, hh_size=="Three" ~ 3, hh_size=="Four" ~ 4, hh_size=="Five" ~ 5, hh_size=="Six" ~ 6)) gss <- gss %>% rowwise() %>% mutate(number_marriages = str_remove(string = number_marriages, pattern = "\ ."*)) %>% mutate(number_marriages = case_when(number_marriages=="No" ~ 0, number_marriages=="One" ~ 1, number_marriages=="Two" ~ 2, number_marriages=="Three" ~ 3, number_marriages=="Four" ~ 4)) gss <- gss %>% rowwise() %>% mutate(number_total_children_known = ifelse(number_total_children_intention=="Don't know"|number_total_children_intention=="NA", 0, 1)) %>% mutate(number_total_children_intention = str_remove(string = number_total_children_intention, pattern = "\ .*")) %>% mutate(number_total_children_intention = case_when(number_total_children_intention=="None" ~ 0,

number_total_children_intention==“One” ~ 1,
number_total_children_intention==“Two” ~ 2,
number_total_children_intention==“Three” ~ 3,
number_total_children_intention==“Four” ~ 4,
number_total_children_intention==“Don’t” ~ as.numeric(NA)))

save to a new csv file and finish cleaning

write_csv(gss, “gss.csv”)

Model

I have used ggplot, histogram, logistic regression, linear regression to analaysis these datasets.

This is the mathmetical notation for linear regression :y = β0 + β1x + e, where β0 is the intercept, there could be β2,β3 and so on. Y must be numerical. Predictors can be both numerical and categorical. In this case, I have picked some numerical variables as predictors and age as Y.

This is the mathmetical notation for logistic regression : log(p/(1-p)) = β0 + β1x,where β0 is the intercept, there could be β2,β3 and so on. P is the probability of an event that is going to occur. β1 is a coefficient represents changes in log adds for every one unit increase in x. Y can either be 1 or 0.

Results

```
##
## ── Column specification ────────────────────────────────────────
──────────
## cols(
##   .default = col_character(),
##   caseid = col_double(),
##   age = col_double(),
##   age_first_child = col_double(),
##   age_youngest_child_under_6 = col_double(),
##   total_children = col_double(),
##   age_start_relationship = col_double(),
##   age_at_first_marriage = col_double(),
##   age_at_first_birth = col_double(),
##   distance_between_houses = col_double(),
##   age_youngest_child_returned_work = col_double(),
##   feelings_life = col_double(),
##   hh_size = col_double(),
##   number_total_children_intention = col_double(),
##   number_marriages = col_double(),
##   fin_supp_child_supp = col_double(),
##   fin_supp_child_exp = col_double(),
##   fin_supp_lump = col_double(),
##   fin_supp_other = col_double(),
```

```
##   is_male = col_double(),
##   main_activity = col_logical()
##   # ... with 1 more columns
## )
## ℹ Use `spec()` for the full column specifications.

##      caseid           age        age_first_child age_youngest_child_
under_6
##  Min.   :    1   Min.   :15.00   Min.   : 0.00   Min.   :0.000

##  1st Qu.: 5151   1st Qu.:37.30   1st Qu.:15.00   1st Qu.:1.000

##  Median :10302   Median :54.20   Median :32.00   Median :2.000

##  Mean   :10302   Mean   :52.19   Mean   :30.57   Mean   :2.412

##  3rd Qu.:15452   3rd Qu.:66.78   3rd Qu.:44.00   3rd Qu.:4.000

##  Max.   :20602   Max.   :80.00   Max.   :60.00   Max.   :5.000

##                                  NA's   :6835    NA's   :18488

##  total_children  age_start_relationship age_at_first_marriage
##  Min.   :0.000   Min.   :18.00          Min.   :15.0
##  1st Qu.:0.000   1st Qu.:25.00          1st Qu.:20.5
##  Median :2.000   Median :30.50          Median :22.8
##  Mean   :1.679   Mean   :33.63          Mean   :24.1
##  3rd Qu.:3.000   3rd Qu.:40.62          3rd Qu.:26.4
##  Max.   :7.000   Max.   :60.00          Max.   :50.0
##  NA's   :19      NA's   :18566          NA's   :15248
##  age_at_first_birth distance_between_houses age_youngest_child_retur
ned_work
##  Min.   :18.00      Min.   : 0.00           Min.   : 0.200

##  1st Qu.:22.80      1st Qu.: 4.00           1st Qu.: 0.500

##  Median :26.40      Median :10.00           Median : 6.000

##  Mean   :26.86      Mean   :17.13           Mean   : 6.589

##  3rd Qu.:30.30      3rd Qu.:24.75           3rd Qu.:12.000

##  Max.   :45.00      Max.   :90.00           Max.   :48.000

##  NA's   :7865       NA's   :19476           NA's   :19466

##  feelings_life        sex            place_birth_canada place_birth_
father
##  Min.   : 0.000   Length:20602       Length:20602       Length:20602
```

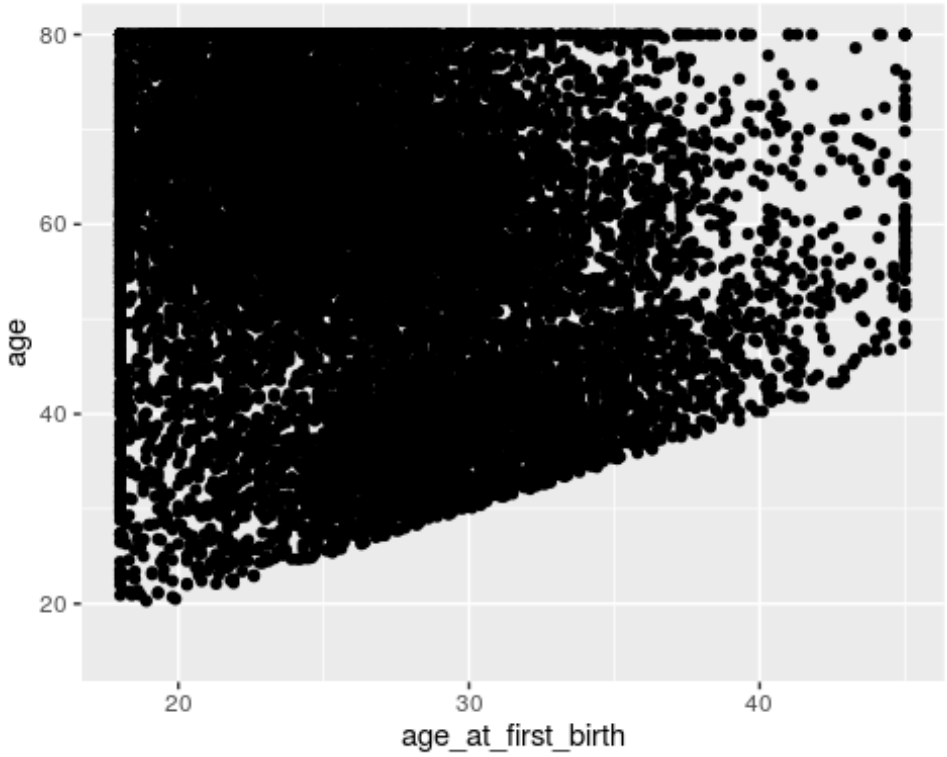
```
##  1st Qu.: 7.000   Class :character   Class :character   Class :chara
cter
##  Median : 8.000   Mode  :character   Mode  :character   Mode  :chara
cter
##  Mean   : 8.094

##  3rd Qu.: 9.000

##  Max.   :10.000

##  NA's   :271

##  place_birth_mother place_birth_macro_region place_birth_province
##  Length:20602       Length:20602             Length:20602
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  year_arrived_canada   province            region           pop_cent
er
##  Length:20602        Length:20602       Length:20602       Length:20
602
##  Class :character    Class :character   Class :character   Class :ch
aracter
##  Mode  :character    Mode  :character   Mode  :character   Mode  :ch
aracter
##

##

##

##

##  marital_status      aboriginal        vis_minority       age_immigr
ation
##  Length:20602       Length:20602       Length:20602       Length:206
02
##  Class :character   Class :character   Class :character   Class :cha
racter
##  Mode  :character   Mode  :character   Mode  :character   Mode  :cha
racter
##

##
```

```
##
##
##  landed_immigrant   citizenship_status  education           own_rent
##  Length:20602       Length:20602       Length:20602       Length:206
02
##  Class :character   Class :character   Class :character   Class :cha
racter
##  Mode  :character   Mode  :character   Mode  :character   Mode  :cha
racter
##
##
##
##
##  living_arrangement   hh_type             hh_size      partner_birth
_country
##  Length:20602       Length:20602       Min.   :1.000   Length:20602
##  Class :character   Class :character   1st Qu.:1.000   Class :charac
ter
##  Mode  :character   Mode  :character   Median :2.000   Mode  :charac
ter
##                                        Mean   :2.347
##                                        3rd Qu.:3.000
##                                        Max.   :6.000
##
##  partner_birth_province partner_vis_minority partner_sex
##  Length:20602           Length:20602         Length:20602
##  Class :character       Class :character     Class :character
##  Mode  :character       Mode  :character     Mode  :character
##
##
##
##
##  partner_education  average_hours_worked worked_last_week
##  Length:20602       Length:20602         Length:20602
##  Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character     Mode  :character
##
```

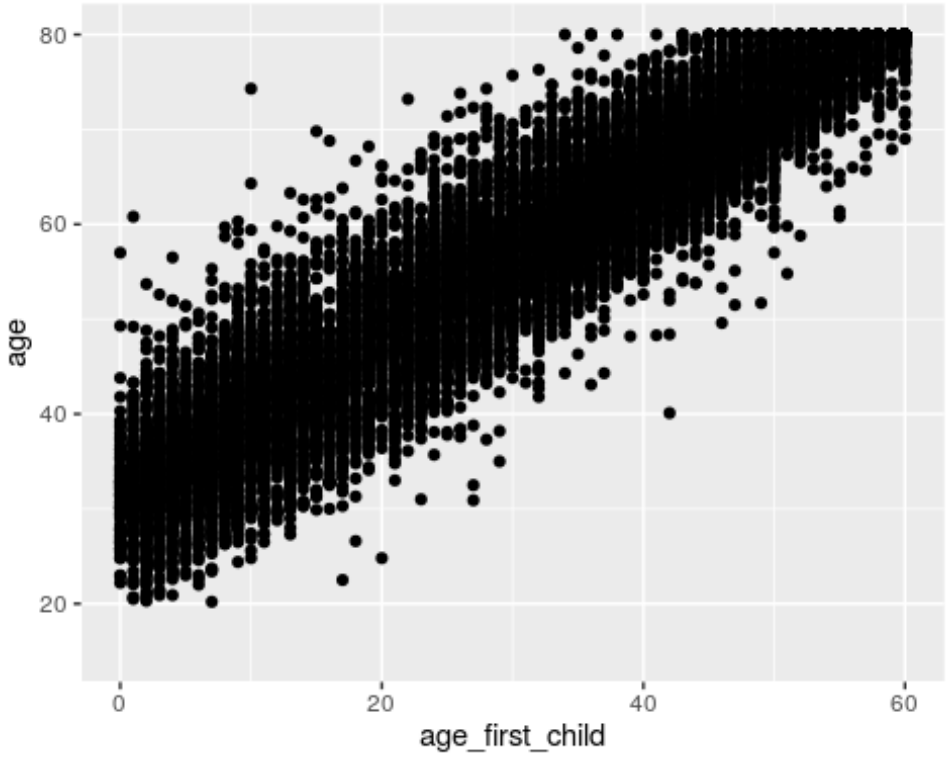
```
##
##
##
##  partner_main_activity self_rated_health  self_rated_mental_health
##  Length:20602          Length:20602       Length:20602
##  Class :character      Class :character   Class :character
##  Mode  :character      Mode  :character   Mode  :character
##
##
##
##
##  religion_has_affiliation regilion_importance language_home
##  Length:20602             Length:20602        Length:20602
##  Class :character         Class :character    Class :character
##  Mode  :character         Mode  :character    Mode  :character
##
##
##
##
##  language_knowledge income_family      income_respondent   occupatio
n
##  Length:20602       Length:20602       Length:20602       Length:206
02
##  Class :character   Class :character   Class :character   Class :cha
racter
##  Mode  :character   Mode  :character   Mode  :character   Mode  :cha
racter
##

##

##

##

##  childcare_regular  childcare_type     childcare_monthly_cost
##  Length:20602       Length:20602       Length:20602
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  ever_fathered_child ever_given_birth   number_of_current_union
##  Length:20602        Length:20602       Length:20602
##  Class :character    Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character
##
##
```

```
##
##
##  lives_with_partner children_in_household number_total_children_inte
ntion
##  Length:20602       Length:20602           Min.   :0.000
##  Class :character   Class :character       1st Qu.:0.000
##  Mode  :character   Mode  :character       Median :0.000
##                                            Mean   :0.903
##                                            3rd Qu.:2.000
##                                            Max.   :4.000
##                                            NA's   :12202
##  has_grandchildren  grandparents_still_living ever_married
##  Length:20602       Length:20602               Length:20602
##  Class :character   Class :character           Class :character
##  Mode  :character   Mode  :character           Mode  :character
##
##
##
##
##  current_marriage_is_first number_marriages religion_participation
##  Length:20602              Min.   :0.0000   Length:20602
##  Class :character          1st Qu.:0.0000   Class :character
##  Mode  :character          Median :1.0000   Mode  :character
##                            Mean   :0.7989
##                            3rd Qu.:1.0000
##                            Max.   :4.0000
##
##  partner_location_residence full_part_time_work time_off_work_birth
##  Length:20602               Length:20602        Length:20602
##  Class :character           Class :character    Class :character
##  Mode  :character           Mode  :character    Mode  :character
##
##
##
##
##  reason_no_time_off_birth returned_same_job  satisfied_time_children
##  Length:20602             Length:20602       Length:20602
##  Class :character         Class :character   Class :character
##  Mode  :character         Mode  :character   Mode  :character
##
##
##
```

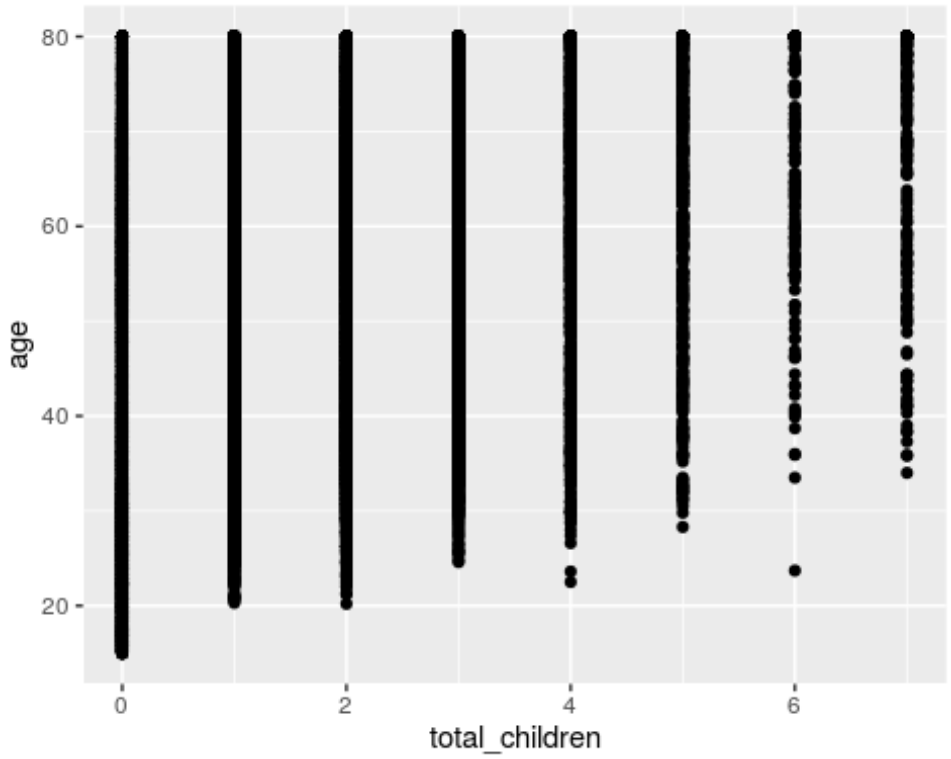
```
##
##  provide_or_receive_fin_supp fin_supp_child_supp fin_supp_child_exp
##  Length:20602                Min.   :0.000       Min.   :0.000
##  Class :character            1st Qu.:1.000       1st Qu.:0.000
##  Mode  :character            Median :1.000       Median :0.000
##                              Mean   :0.765       Mean   :0.339
##                              3rd Qu.:1.000       3rd Qu.:1.000
##                              Max.   :1.000       Max.   :1.000
##                              NA's   :20057       NA's   :20057
##  fin_supp_lump   fin_supp_other  fin_supp_agreement future_children_
intention
##  Min.   :0.000   Min.   :0.000   Length:20602       Length:20602

##  1st Qu.:0.000   1st Qu.:0.000   Class :character   Class :character

##  Median :0.000   Median :0.000   Mode  :character   Mode  :character

##  Mean   :0.055   Mean   :0.055

##  3rd Qu.:0.000   3rd Qu.:0.000

##  Max.   :1.000   Max.   :1.000

##  NA's   :20057   NA's   :20057

##     is_male       main_activity    age_diff         number_total_chi
ldren_known
##  Min.   :0.0000   Mode:logical   Length:20602       Min.   :0.0000

##  1st Qu.:0.0000   NA's:20602     Class :character   1st Qu.:0.0000

##  Median :0.0000                  Mode  :character   Median :0.0000

##  Mean   :0.4562                                     Mean   :0.4123

##  3rd Qu.:1.0000                                     3rd Qu.:1.0000

##  Max.   :1.0000                                     Max.   :1.0000

##

## Warning: Removed 7865 rows containing missing values (geom_point).
```



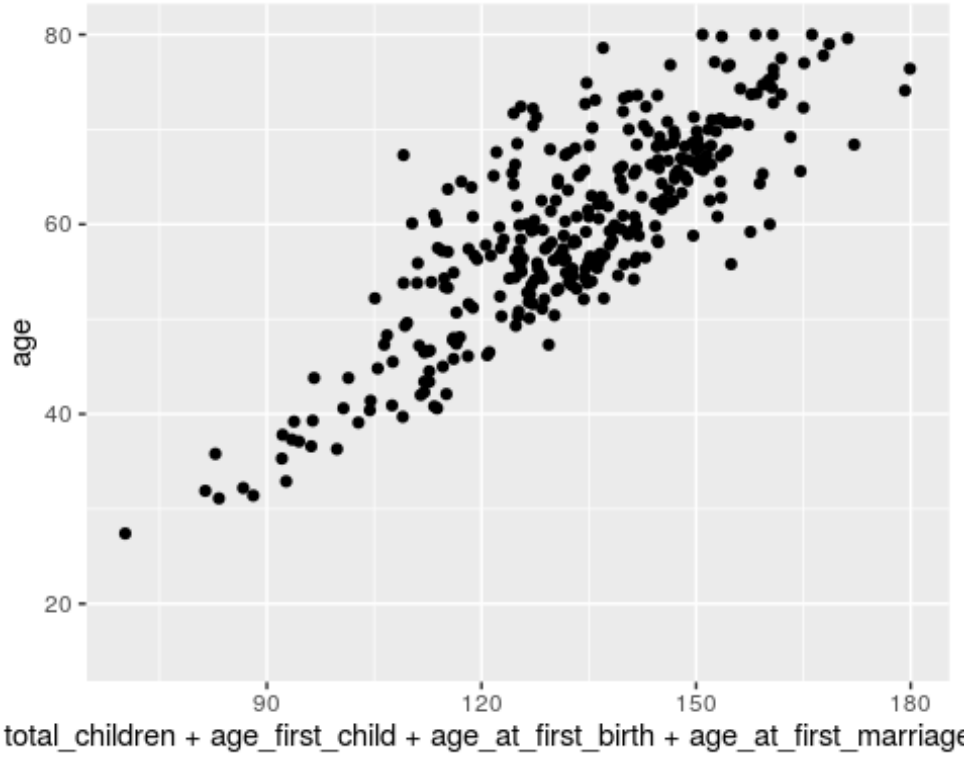

```
## Warning: Removed 6835 rows containing missing values (geom_point).
```




```
## Warning: Removed 19 rows containing missing values (geom_point).
```




```
## Warning: Removed 20272 rows containing missing values (geom_point).
```



linear regression

```
##
## Call:
## lm(formula = age ~ total_children + age_first_child + age_at_first_b
irth +
##     age_at_first_marriage + age_start_relationship, data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7577 -0.3441 -0.0071  0.2994 13.8028
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.541116   0.470578  -1.150   0.2510
## total_children         -0.010812   0.040020  -0.270   0.7872
## age_first_child         0.985121   0.005762 170.956   <2e-16 ***
## age_at_first_birth      1.018111   0.013588  74.930   <2e-16 ***
## age_at_first_marriage  -0.005123   0.012937  -0.396   0.6924
## age_start_relationship  0.015489   0.006343   2.442   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9223 on 324 degrees of freedom
##   (20272 observations deleted due to missingness)
## Multiple R-squared:  0.9925, Adjusted R-squared:  0.9924
## F-statistic:  8565 on 5 and 324 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = age ~ total_children + income_family, data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.970 -12.472   1.057  12.067  41.785
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)

## (Intercept)                      39.75129    0.35396 112.304  < 2e-16
 ***
## total_children                    5.30882    0.07226  73.473  < 2e-16
 ***
## income_family$125,000 and more   -1.53603    0.40167  -3.824 0.000132
 ***
## income_family$25,000 to $49,999   8.06564    0.40688  19.823  < 2e-16
 ***
## income_family$50,000 to $74,999   5.46642    0.41852  13.061  < 2e-16
 ***
## income_family$75,000 to $99,999   1.89703    0.43850   4.326 1.52e-05
```

```
 ***
## income_familyLess than $25,000   6.86210    0.44340  15.476  < 2e-16
 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.44 on 20576 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.2438, Adjusted R-squared:  0.2436
## F-statistic:  1106 on 6 and 20576 DF,  p-value: < 2.2e-16
```

Logistic Regression

```
## # A tibble: 6 x 81
##   caseid   age age_first_child age_youngest_ch… total_children age_s
tart_relat…
##    <dbl> <dbl>           <dbl>            <dbl>          <dbl>
      <dbl>
## 1      1  52.7              27               NA              1
       NA
## 2      2  51.1              33               NA              5
       NA
## 3      3  63.6              40               NA              5
       NA
## 4      4  80                56               NA              1
       NA
## 5      5  28                NA               NA              0
       25.3
## 6      6  63                37               NA              2
       NA
## # … with 75 more variables: age_at_first_marriage <dbl>,
## #   age_at_first_birth <dbl>, distance_between_houses <dbl>,
## #   age_youngest_child_returned_work <dbl>, feelings_life <dbl>, sex
 <chr>,
## #   place_birth_canada <chr>, place_birth_father <chr>,
## #   place_birth_mother <chr>, place_birth_macro_region <chr>,
## #   place_birth_province <chr>, year_arrived_canada <chr>, province
<chr>,
## #   region <chr>, pop_center <chr>, marital_status <chr>, aboriginal
 <chr>,
## #   vis_minority <chr>, age_immigration <chr>, landed_immigrant <ch
r>,
## #   citizenship_status <chr>, education <chr>, own_rent <chr>,
## #   living_arrangement <chr>, hh_type <chr>, hh_size <dbl>,
## #   partner_birth_country <chr>, partner_birth_province <chr>,
## #   partner_vis_minority <chr>, partner_sex <chr>, partner_education
 <chr>,
## #   average_hours_worked <chr>, worked_last_week <chr>,
## #   partner_main_activity <chr>, self_rated_health <chr>,
## #   self_rated_mental_health <chr>, religion_has_affiliation <chr>,
## #   regilion_importance <chr>, language_home <chr>, language_knowled
```

```
ge <chr>,
## #   income_family <chr>, income_respondent <chr>, occupation <chr>,
## #   childcare_regular <chr>, childcare_type <chr>,
## #   childcare_monthly_cost <chr>, ever_fathered_child <chr>,
## #   ever_given_birth <chr>, number_of_current_union <chr>,
## #   lives_with_partner <chr>, children_in_household <chr>,
## #   number_total_children_intention <dbl>, has_grandchildren <chr>,
## #   grandparents_still_living <chr>, ever_married <chr>,
## #   current_marriage_is_first <chr>, number_marriages <dbl>,
## #   religion_participation <chr>, partner_location_residence <chr>,
## #   full_part_time_work <chr>, time_off_work_birth <chr>,
## #   reason_no_time_off_birth <chr>, returned_same_job <chr>,
## #   satisfied_time_children <chr>, provide_or_receive_fin_supp <chr>,
## #   fin_supp_child_supp <dbl>, fin_supp_child_exp <dbl>, fin_supp_lu
mp <dbl>,
## #   fin_supp_other <dbl>, fin_supp_agreement <chr>,
## #   future_children_intention <chr>, is_male <dbl>, main_activity <l
gl>,
## #   age_diff <chr>, number_total_children_known <dbl>

## # A tibble: 6 x 82
##   caseid   age age_first_child age_youngest_ch… total_children age_s
tart_relat…
##    <dbl> <dbl>           <dbl>            <dbl>          <dbl>
      <dbl>
## 1      1  52.7              27               NA              1
       NA
## 2      2  51.1              33               NA              5
       NA
## 3      3  63.6              40               NA              5
       NA
## 4      4  80                56               NA              1
       NA
## 5      5  28                NA               NA              0
       25.3
## 6      6  63                37               NA              2
       NA
## # … with 76 more variables: age_at_first_marriage <dbl>,
## #   age_at_first_birth <dbl>, distance_between_houses <dbl>,
## #   age_youngest_child_returned_work <dbl>, feelings_life <dbl>, sex
 <chr>,
## #   place_birth_canada <chr>, place_birth_father <chr>,
## #   place_birth_mother <chr>, place_birth_macro_region <chr>,
## #   place_birth_province <chr>, year_arrived_canada <chr>, province
<chr>,
## #   region <chr>, pop_center <chr>, marital_status <chr>, aboriginal
 <chr>,
## #   vis_minority <chr>, age_immigration <chr>, landed_immigrant <ch
r>,
## #   citizenship_status <chr>, education <chr>, own_rent <chr>,
```

```
## #   living_arrangement <chr>, hh_type <chr>, hh_size <dbl>,
## #   partner_birth_country <chr>, partner_birth_province <chr>,
## #   partner_vis_minority <chr>, partner_sex <chr>, partner_education
 <chr>,
## #   average_hours_worked <chr>, worked_last_week <chr>,
## #   partner_main_activity <chr>, self_rated_health <chr>,
## #   self_rated_mental_health <chr>, religion_has_affiliation <chr>,
## #   regilion_importance <chr>, language_home <chr>, language_knowled
ge <chr>,
## #   income_family <chr>, income_respondent <chr>, occupation <chr>,
## #   childcare_regular <chr>, childcare_type <chr>,
## #   childcare_monthly_cost <chr>, ever_fathered_child <chr>,
## #   ever_given_birth <chr>, number_of_current_union <chr>,
## #   lives_with_partner <chr>, children_in_household <chr>,
## #   number_total_children_intention <dbl>, has_grandchildren <chr>,
## #   grandparents_still_living <chr>, ever_married <chr>,
## #   current_marriage_is_first <chr>, number_marriages <dbl>,
## #   religion_participation <chr>, partner_location_residence <chr>,
## #   full_part_time_work <chr>, time_off_work_birth <chr>,
## #   reason_no_time_off_birth <chr>, returned_same_job <chr>,
## #   satisfied_time_children <chr>, provide_or_receive_fin_supp <chr>,
## #   fin_supp_child_supp <dbl>, fin_supp_child_exp <dbl>, fin_supp_lu
mp <dbl>,
## #   fin_supp_other <dbl>, fin_supp_agreement <chr>,
## #   future_children_intention <chr>, is_male <dbl>, main_activity <l
gl>,
## #   age_diff <chr>, number_total_children_known <dbl>, income_status
 <chr>

##
##    1    2    3    4    5    6
## 2775 4345 3696 2921 2158 4707

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion
```

```
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion
```

```
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion
```

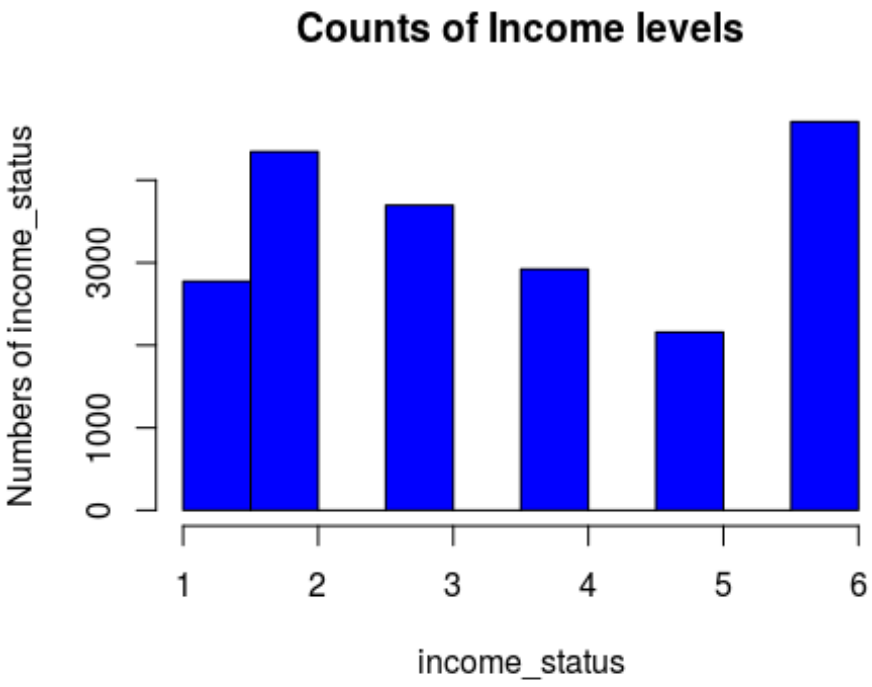
```
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

##   caseid  age age_first_child age_youngest_child_under_6 total_child
ren
## 1      1 52.7              27                          NA
  1
## 2      2 51.1              33                          NA
  5
## 3      3 63.6              40                          NA
  5
## 4      4 80.0              56                          NA
  1
## 5      5 28.0              NA                          NA
  0
## 6      6 63.0              37                          NA
  2
##   age_start_relationship age_at_first_marriage age_at_first_birth
## 1                     NA                    NA               25.9
## 2                     NA                    NA                 NA
## 3                     NA                    NA               23.2
## 4                     NA                    NA               27.3
## 5                   25.3                    NA                 NA
## 6                     NA                    NA               25.8
##   distance_between_houses age_youngest_child_returned_work feelings_
life sex
## 1                      30                                NA
   8   1
## 2                      NA                                NA
  10   0
## 3                      NA                                NA
   8   1
## 4                      NA                                NA
  10   1
## 5                      NA                                NA
   8   0
## 6                      NA                                NA
   9   1
##   place_birth_canada place_birth_father place_birth_mother
## 1                 NA                 NA                 NA
## 2                 NA                 NA                 NA
## 3                 NA                 NA                 NA
## 4                 NA                 NA                 NA
## 5                 NA                 NA                 NA
## 6                 NA                 NA                 NA
##   place_birth_macro_region place_birth_province year_arrived_canada
province
## 1                       NA                   NA                  NA
```

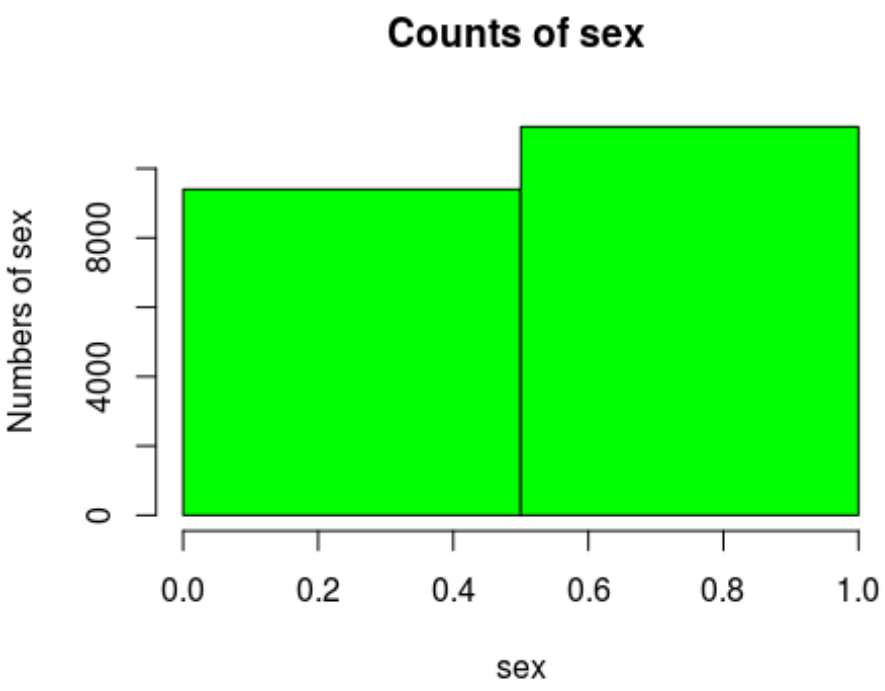
```
      NA
## 2                         NA                    NA                    NA
      NA
## 3                         NA                    NA                    NA
      NA
## 4                         NA                    NA                    NA
      NA
## 5                         NA                    NA                    NA
      NA
## 6                         NA                    NA                    NA
      NA
##   region pop_center marital_status aboriginal vis_minority age_immig
ration
## 1     NA         NA             NA         NA            NA
    NA
## 2     NA         NA             NA         NA            NA
    NA
## 3     NA         NA             NA         NA            NA
    NA
## 4     NA         NA             NA         NA            NA
    NA
## 5     NA         NA             NA         NA            NA
    NA
## 6     NA         NA             NA         NA            NA
    NA
##   landed_immigrant citizenship_status education own_rent living_arra
ngement
## 1               NA                  NA        NA       NA
     NA
## 2               NA                  NA        NA       NA
     NA
## 3               NA                  NA        NA       NA
     NA
## 4               NA                  NA        NA       NA
     NA
## 5               NA                  NA        NA       NA
     NA
## 6               NA                  NA        NA       NA
     NA
##   hh_type hh_size partner_birth_country partner_birth_province
## 1      NA       1                    NA                     NA
## 2      NA       2                    NA                     NA
## 3      NA       2                    NA                     NA
## 4      NA       2                    NA                     NA
## 5      NA       2                    NA                     NA
## 6      NA       2                    NA                     NA
##   partner_vis_minority partner_sex partner_education average_hours_w
orked
## 1                   NA          NA                NA
   NA
```

```
## 2                   NA          NA                NA
   NA
## 3                   NA          NA                NA
   NA
## 4                   NA          NA                NA
   NA
## 5                   NA          NA                NA
   NA
## 6                   NA          NA                NA
   NA
##   worked_last_week partner_main_activity self_rated_health
## 1               NA                    NA                NA
## 2               NA                    NA                NA
## 3               NA                    NA                NA
## 4               NA                    NA                NA
## 5               NA                    NA                NA
## 6               NA                    NA                NA
##   self_rated_mental_health religion_has_affiliation regilion_importa
nce
## 1                       NA                       NA
 NA
## 2                       NA                       NA
 NA
## 3                       NA                       NA
 NA
## 4                       NA                       NA
 NA
## 5                       NA                       NA
 NA
## 6                       NA                       NA
 NA
##   language_home language_knowledge income_family income_respondent o
ccupation
## 1            NA                 NA            NA                 NA
       NA
## 2            NA                 NA            NA                 NA
       NA
## 3            NA                 NA            NA                 NA
       NA
## 4            NA                 NA            NA                 NA
       NA
## 5            NA                 NA            NA                 NA
       NA
## 6            NA                 NA            NA                 NA
       NA
##   childcare_regular childcare_type childcare_monthly_cost ever_fathe
red_child
## 1                NA             NA                      NA
       NA
## 2                NA             NA                      NA
```

```
       NA
## 3                 NA              NA                      NA
       NA
## 4                 NA              NA                      NA
       NA
## 5                 NA              NA                      NA
       NA
## 6                 NA              NA                      NA
       NA
##   ever_given_birth number_of_current_union lives_with_partner
## 1               NA                      NA                 NA
## 2               NA                      NA                 NA
## 3               NA                      NA                 NA
## 4               NA                      NA                 NA
## 5               NA                      NA                 NA
## 6               NA                      NA                 NA
##   children_in_household number_total_children_intention has_grandchi
ldren
## 1                    NA                             NA
   NA
## 2                    NA                             NA
   NA
## 3                    NA                             NA
   NA
## 4                    NA                             NA
   NA
## 5                    NA                              2
   NA
## 6                    NA                             NA
   NA
##   grandparents_still_living ever_married current_marriage_is_first
## 1                        NA           NA                         NA
## 2                        NA           NA                         NA
## 3                        NA           NA                         NA
## 4                        NA           NA                         NA
## 5                        NA           NA                         NA
## 6                        NA           NA                         NA
##   number_marriages religion_participation partner_location_residence
## 1                0                      NA                         NA
## 2                1                      NA                         NA
## 3                1                      NA                         NA
## 4                1                      NA                         NA
## 5                0                      NA                         NA
## 6                1                      NA                         NA
##   full_part_time_work time_off_work_birth reason_no_time_off_birth
## 1                  NA                  NA                       NA
## 2                  NA                  NA                       NA
## 3                  NA                  NA                       NA
## 4                  NA                  NA                       NA
## 5                  NA                  NA                       NA
```

```
## 6                    NA                    NA                          NA
##   returned_same_job satisfied_time_children provide_or_receive_fin_s
upp
## 1                NA                       NA
 NA
## 2                NA                       NA
 NA
## 3                NA                       NA
 NA
## 4                NA                       NA
 NA
## 5                NA                       NA
 NA
## 6                NA                       NA
 NA
##   fin_supp_child_supp fin_supp_child_exp fin_supp_lump fin_supp_othe
r
## 1                  NA                 NA            NA              N
A
## 2                  NA                 NA            NA              N
A
## 3                  NA                 NA            NA              N
A
## 4                  NA                 NA            NA              N
A
## 5                  NA                 NA            NA              N
A
## 6                  NA                 NA            NA              N
A
##   fin_supp_agreement future_children_intention is_male main_activity
 age_diff
## 1                 NA                        NA       0             NA
       NA
## 2                 NA                        NA       1             NA
       NA
## 3                 NA                        NA       0             NA
       NA
## 4                 NA                        NA       0             NA
       NA
## 5                 NA                        NA       1             NA
       NA
## 6                 NA                        NA       0             NA
       NA
##   number_total_children_known income_status
## 1                           0             2
## 2                           0             4
## 3                           0             4
## 4                           0             5
## 5                           1             3
## 6                           0             3
```





Standard Logistic Regression

```
##
## Call:
## glm(formula = sex ~ age + age_first_child + total_children +
```

```
##     feelings_life + as.factor(income_status), family = "binomial",
##     data = gss_numeric)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4244  -1.1885   0.7674   1.0340   1.9379
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.786023   0.170562  22.197  < 2e-16 ***
## age                         -0.097040   0.003906 -24.841  < 2e-16 ***
## age_first_child              0.090983   0.003602  25.260  < 2e-16 ***
## total_children              -0.189921   0.017017 -11.161  < 2e-16 ***
## feelings_life                0.025987   0.011501   2.260   0.0238 *
## as.factor(income_status)2 -0.293959   0.070921  -4.145 3.40e-05 ***
## as.factor(income_status)3 -0.582232   0.072268  -8.057 7.85e-16 ***
## as.factor(income_status)4 -0.632205   0.075822  -8.338  < 2e-16 ***
## as.factor(income_status)5 -0.733849   0.080806  -9.082  < 2e-16 ***
## as.factor(income_status)6 -0.633354   0.071529  -8.855  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18560  on 13584  degrees of freedom
## Residual deviance: 17585  on 13575  degrees of freedom
##   (7017 observations deleted due to missingness)
## AIC: 17605
##
## Number of Fisher Scoring iterations: 4
```

Survey Estimation for Logistic Regression

install.packages("survey") library(survey)

Using the Survey Library

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-li
brary/4.0'
## (as 'lib' is unspecified)

## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

##
## Call:
## svyglm(formula = sex ~ age + age_first_child + total_children +
##     feelings_life + as.factor(income_status), design = a, family = "
binomial")
##
## Survey design:
## svydesign(id = ~1, data = gss_numeric, fpc = b)
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.786023   0.140641  26.920  < 2e-16 ***
## age                        -0.097040   0.003257 -29.794  < 2e-16 ***
## age_first_child             0.090983   0.002992  30.410  < 2e-16 ***
## total_children             -0.189921   0.013716 -13.847  < 2e-16 ***
## feelings_life               0.025987   0.009315   2.790  0.00528 **
## as.factor(income_status)2 -0.293959   0.056767  -5.178 2.27e-07 ***
## as.factor(income_status)3 -0.582232   0.058178 -10.008  < 2e-16 ***
## as.factor(income_status)4 -0.632205   0.061268 -10.319  < 2e-16 ***
## as.factor(income_status)5 -0.733849   0.065778 -11.157  < 2e-16 ***
## as.factor(income_status)6 -0.633354   0.058028 -10.915  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.002454)
##
## Number of Fisher Scoring iterations: 4
```

Discussion

By using summary, we can see the details of each variable.

Then we created several ggplots to determine relationships between factors.

There is a positive linear relationship between age_first_child and age.

There is a postive linear relationship between age and total_children, age_first_children, age_at_first_birth, age_at_first_marriage and age_start_relationship.

Then we run linear regression since we foud some relationship.

The function will be Y = (-0.54 + -0.01* total_children + 0.99* age_first_child +1.02* age_at_first_birth-0.01* age_at_first_marriage+0.02*age_start_relationship)

Since there are many missing values that has been removed, about 20272 rows. So we recreated a linear regression, this time, I added categorical variables income_family to test if age and income are related.

The function will be Y = 39.75 + 5.31* total_children - 1.54* '125,000 and more' + 8.07* '25,000 to $49,999' + 5.47*'50,000 to $74,999' + 1.90* '75,000 to $99,999' + 6.86*'Less than $25,000'

In order to generate logistic regression. We created a binary variable as our Y value which is sex. We set female = 1 and male = 0.

Then we sepertaed the income_family by 1,2,3,4,5,6 levels. '1' is the poorest family and '6' is richest family as the rank.

Since logistic regression Y is a binary.Current Y value is a character. So we changed GSS file to numerical and named it gss_numeric.

We created a histogram to show the income status from 1 through 6. According to the histogram, there are many income_family that have income over $125,000 and between $25,000 to $49,000.

We also created a histogram of female and male. According to the histogram, there are more females than males in the dataset.

The logistic regression function we got is Y = 3.79-0.097* age+0.091* age_first_child-0.19* total_children+0.026*feelings_life-0.294x2-0.58x3-0.63x4-0.73x5-0.63x6
There is about 1000 difference between null deviance and residual deviance, the larger the difference is, the better the model is.

Weaknesses

There are many missing values even after data cleaning. If by removing all the NAs,there will be 0 rows left. So in order to run some analysis, we have to obtain some missing values. But due to this is a questionaire, it's hard to obtain perfect answers. There are too many columns, factors in the dataset, which can create mislead when doing analysis.

Next Steps

Since we have produced predictive model, which is the logistic regression. We can try to predict some values. We also would like to create other algorithms such as random forest, or decision tree. Because there are too many factors in the dataset random forest and decision tree can deal with large dataset with multidimensions. we think GSS dataset is multidimensional.And of course, we should also set up a following survey to fill up these missing values by doing our best. Obtaining good dataset can lead to a better result.

References

Technology, A. (n.d.). Data Centre. Retrieved October 20, 2020, from http://dc.chass.utoronto.ca/myaccess.html Tyagi, P. (2018, December 25). Decision Tree. Retrieved from https://medium.com/@pytyagi/decision-tree-ac0c9e3b8258