BSTT413_Final-Project_NHANES_DM.R

shujunxu

2020-05-11

```
##########################
# Download NHANES Dataset #
##########################

library(tidyverse) # data management (dplyr) and plots (ggplot2)

## ── Attaching packages ──────────────────────────────────────────────────────────────────
tidyverse 1.3.0 ──

## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.4
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0

## ── Conflicts ───────────────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(foreign) # reading xport (generated by SAS) files
require(dplyr)

# 1.1 Diagnosed diabetes
diabfile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DIQ_J.XPT",
              destfile = diabfile,mode = "wb") # use 'wb' to read as binary file
diabdf = read.xport(diabfile)
# using foreign package (file must be binary format) to read the export into R dataframe

# 1.2 Undiagnosed diabetes (A1c blood sugar lab data)
a1cfile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/GHB_J.XPT",
              destfile = a1cfile,mode = "wb")
a1cdf = read.xport(a1cfile)

# 2. Demographics - demo and weighting variables
demofile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.XPT",
              destfile = demofile,mode = "wb")
demodf = read.xport(demofile)

# 3. Health Insurance
insurfile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/HIQ_J.XPT",
              destfile = insurfile,mode = "wb")
insurdf = read.xport(insurfile)

# 4. Physical Activity (Adults Section)
pafile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/PAQ_J.XPT",
              destfile = pafile,mode = "wb")
padf = read.xport(pafile)
```

```r
# 5. Diet
dietfile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DBQ_J.XPT",
              destfile = dietfile,mode = "wb")
dietdf = read.xport(dietfile)

# 6. BMI
bmifile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BMX_J.XPT",
              destfile = bmifile,mode = "wb")
bmidf = read.xport(bmifile)

# 7. Blood Pressure
bpfile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BPX_J.XPT",
              destfile = bpfile,mode = "wb")
bpdf = read.xport(bpfile)

# 8. Mental health
mhfile = tempfile()
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DPQ_J.XPT",
              destfile = mhfile,mode = 'wb')
mhdf = read.xport(mhfile)



##############################
# Merging and Transformation #
##############################


nhanesdat = full_join(diabdf, a1cdf, by="SEQN") %>%
  select(SEQN,DIQ010,LBXGH) %>%
  filter(DIQ010 != 9) %>%
  mutate(diabetes = case_when(
  DIQ010 %in% c(1, 3) ~ "Yes",
  DIQ010 =2 & LBXGH >=6.5 ~ "Yes",
  DIQ010 =2 & LBXGH <6.5 ~ "No",
  DIQ010 =2 & is.na(LBXGH) ~ "No")) %>%
  inner_join(demodf,by="SEQN") %>%
  left_join(insurdf, by="SEQN") %>%
  left_join(padf, by="SEQN") %>%
  left_join(dietdf, by="SEQN") %>%
  left_join(bmidf, by="SEQN") %>%
  left_join(bpdf, by="SEQN") %>%
  left_join(mhdf, by="SEQN") %>%
  select(SEQN,
         RIDAGEYR, # age
         RIAGENDR, # gender
         RIDRETH3, # Race
         INDHHIN2, # household income
         HIQ011,   # Health Insurance
         PAQ650,   # Physical Activity (Vigorous recreational activities)
         PAQ665,   # Physical Activity (Moderate recreational activities)
         DBQ700,   # How healthy the diet is
         BMXBMI,   # BMI
         BPXDI3,   # Diastolic: Blood pres (3rd rdg) mm Hg (range 0-134)
         BPXSY3,   # Systolic: Blood pres (3rd rdg) mm Hg (72-238)
         starts_with("DPQ"),  # MH screener questions
         diabetes,  # Diabates (Diagnosed and undiagnosed)
         SDMVPSU,  # Masked variance pseudo-PSU, weighting variable
```

```r
         SDMVSTRA, # Masked variance pseudo-stratum
         WTINT2YR, # Full sample 2 year interview weight
         WTMEC2YR) %>%   # Full sample 2 year MEC exam weight

# use "na_if" to convert an annoying value to NA
  mutate_at(c("HIQ011", "DBQ700", "PAQ650", "PAQ665"),
            na_if, 7) %>%  # use "na_if" to convert an annoying value to NA
  mutate_at(c("HIQ011", "DBQ700", "PAQ665", "PAQ665"),
            na_if, 9) %>%
  mutate_at(vars(starts_with("DPQ")), # identify the variables starting with "DPQ"
            na_if, 7) %>%
  mutate_at(vars(starts_with("DPQ")),
            na_if, 9) %>%
  mutate_at("INDHHIN2",na_if, 77) %>%
  mutate_at("INDHHIN2",na_if, 99) %>%

# Gender/Insurance/Diabates: Change to factor
  mutate(RIAGENDR = factor(RIAGENDR, labels = c('Male', 'Female'))) %>%
  mutate(HIQ011 = factor(HIQ011,labels = c('Insured', 'Uninsured'))) %>%
  mutate(HIQ011 = relevel(HIQ011,ref='Uninsured')) %>%
  mutate(diabetes=factor(diabetes)) %>%

# Race:
  mutate(RaceEth = case_when(
    RIDRETH3 %in% c(1:2) ~ "Hispanic Origin",
    RIDRETH3 %in% c(3) ~ "NH White",
    RIDRETH3 %in% c(4) ~ "NH Black",
    RIDRETH3 %in% c(6:7) ~ "NH Asian/other")) %>%
  mutate(RaceEth=factor(RaceEth)) %>%
  mutate(RaceEth=relevel(RaceEth,ref="NH White")) %>%

# Household income :
  mutate(income = case_when(
    INDHHIN2 %in% c(1:4, 13) ~ "0-20K",
    INDHHIN2 %in% c(5:8, 12) ~ "20-55K",
    INDHHIN2 %in% c(9:10, 14,15) ~ "55-100K+")) %>%

# Diet: How healthy
  mutate(diet = case_when(
    DBQ700 %in% c(1:2) ~ "Very good/Excellent",
    DBQ700 %in% c(3) ~ "Good",
    DBQ700 %in% c(4:5) ~ "Poor/Fair")) %>%

# Physical Activity: Moderate or vigorous
  mutate(pa=ifelse(PAQ650 ==1 | PAQ665==1 ,"Yes (active)","No (inactive)")) %>%

# BMI (kg/m**2): cut into four levels, make factor and add labels
  mutate(bmi_cat = cut(BMXBMI,
                       c(0,18.5,25,30,Inf),right=F,
                       labels = c("underweight","normal","overweight","Obese"))) %>%
  mutate(bmi_cat=factor(bmi_cat)) %>%
  mutate(bmi_cat=relevel(bmi_cat,ref="normal")) %>%

# Depression scale total -> binary variable
  mutate(deptot = rowSums(select(., starts_with("DPQ")), na.rm = T)) %>%
  # create new variable "deptot", and remove "NA" to simplify the question
  mutate(depressed = factor(deptot > 9, labels = c("Min/mild", "Mod/Severe"))) %>%

# HBP: DBP>=80 or SBP>=130 mmHg
  mutate(hbp=ifelse(BPXDI3 >=80 | BPXSY3 >=130 ,"Yes (hypertension)","No (normal)")) %>%
```

```r
# Age: select adults, cut into tertiles, make factor and add labels
  filter(RIDAGEYR>=18) %>%
  mutate(age_cat = cut(RIDAGEYR,
                       breaks = quantile(RIDAGEYR, probs = c(0, 1/3, 2/3, 1)),
                       include.lowest = T,
                       labels = c("18-39", "40-61", "62+"))) %>% #includes 80+

# filter based on inclusion criteria (The primary exposure)
  filter(!is.na(pa))

# Save to local drive
save(nhanesdat, file="nhanesdat.RData")

################################################
# Analysis of NHANES 2017-2018 data on Diabates#
################################################

# Use load() to retrieve saved data
load("nhanesdat.RData")
library(tidyr)
library(gridExtra)
require(dplyr)
require(ggplot2)

##### Univariate analysis #####
library(arsenal)
Table_Univariate <-tableby(~ diabetes + pa + diet +
                       BMXBMI + BPXDI3 + BPXSY3 + deptot +
                       RIDAGEYR + RaceEth + RIAGENDR + income + HIQ011, data = nhanesdat)
summary(Table_Univariate)
```

	Overall (N=5852)
diabetes	
No	4649 (79.4%)
Yes	1203 (20.6%)
pa	
No (inactive)	3117 (53.3%)
Yes (active)	2735 (46.7%)
diet	
N-Miss	2
Good	2290 (39.1%)
Poor/Fair	1932 (33.0%)
Very good/Excellent	1628 (27.8%)
BMXBMI	
N-Miss	422
Mean (SD)	29.692 (7.443)
Range	14.200 - 86.200
BPXDI3	
N-Miss	737
Mean (SD)	72.088 (13.082)
Range	0.000 - 134.000
BPXSY3	
N-Miss	737
Mean (SD)	125.794 (20.087)
Range	72.000 - 238.000
deptot	
Mean (SD)	3.005 (4.431)
Range	0.000 - 28.000

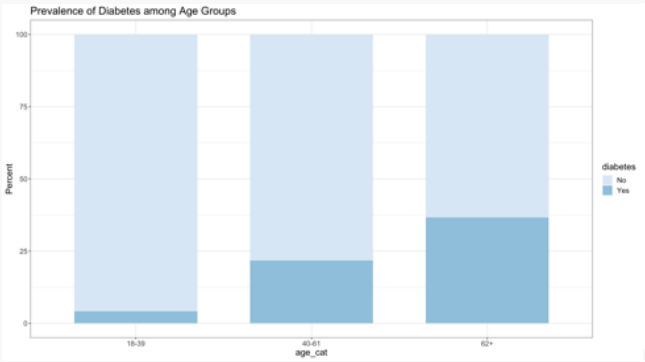
```
## |**RIDAGEYR**                          |                  |
## |   Mean (SD)           | 49.888 (18.772)  |
## |   Range               | 18.000 - 80.000  |
## |**RaceEth**                           |                  |
## |   NH White            |   2031 (34.7%)   |
## |   Hispanic Origin     |   1332 (22.8%)   |
## |   NH Asian/other      |   1146 (19.6%)   |
## |   NH Black            |   1343 (22.9%)   |
## |**RIAGENDR**                          |                  |
## |   Male                |   2839 (48.5%)   |
## |   Female              |   3013 (51.5%)   |
## |**income**                            |                  |
## |   N-Miss              |       613        |
## |   0-20K               |   1010 (19.3%)   |
## |   20-55K              |   2117 (40.4%)   |
## |   55-100K+            |   2112 (40.3%)   |
## |**HIQ011**                            |                  |
## |   N-Miss              |        19        |
## |   Uninsured           |   892 (15.3%)    |
## |   Insured             |   4941 (84.7%)   |
```

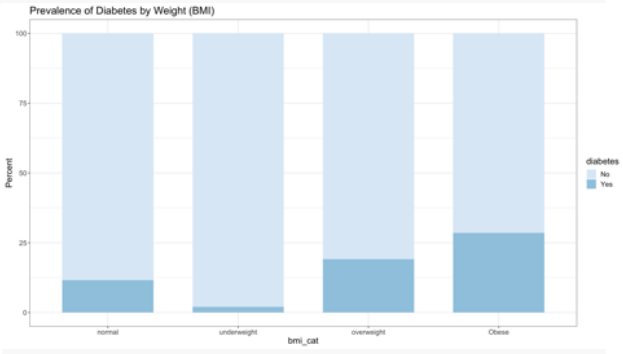
```
##### Bivariate Analysis, use CreateTableOne() #####
library(tableone)
Table_Bivariate <-CreateTableOne(vars=c("pa","diet","bmi_cat","hbp","depressed",
                                        "age_cat","RaceEth","RIAGENDR","income","HIQ011"),
                                 factorVars =c("pa","diet","bmi_cat","hbp","depressed",

"age_cat","RaceEth","RIAGENDR","income","HIQ011"),
                                 strata ="diabetes",
                                 data=nhanesdat)
summary(Table_Bivariate)

##
##      ### Summary of categorical variables ###
##
## diabetes: No
##        var    n miss p.miss              level freq percent cum.percent
##         pa 4649    0    0.0       No (inactive) 2346    50.5        50.5
##                                   Yes (active) 2303    49.5       100.0
##
##       diet 4649    2    0.0               Good 1810    38.9        38.9
##                                      Poor/Fair 1511    32.5        71.5
##                                 Very good/Excellent 1326    28.5       100.0
##
##    bmi_cat 4649  346    7.4             normal 1213    28.2        28.2
##                                     underweight   97     2.3        30.4
##                                      overweight 1395    32.4        62.9
##                                           Obese 1598    37.1       100.0
##
##        hbp 4649  606   13.0         No (normal) 2321    57.4        57.4
##                                 Yes (hypertension) 1722    42.6       100.0
##
##  depressed 4649    0    0.0            Min/mild 4260    91.6        91.6
##                                     Mod/Severe  389     8.4       100.0
##
##    age_cat 4649    0    0.0               18-39 1890    40.7        40.7
##                                           40-61 1579    34.0        74.6
##                                             62+ 1180    25.4       100.0
##
##    RaceEth 4649    0    0.0             NH White 1656    35.6        35.6
##                                   Hispanic Origin 1057    22.7        58.4
```

```
##                                 NH Asian/other  896    19.3        77.6
##                                       NH Black 1040    22.4       100.0
##
##   RIAGENDR 4649    0    0.0               Male 2216    47.7        47.7
##                                        Female 2433    52.3       100.0
##
##     income 4649  484   10.4               0-20K  797    19.1        19.1
##                                         20-55K 1658    39.8        58.9
##                                        55-100K+ 1710    41.1       100.0
##
##     HIQ011 4649   16    0.3           Uninsured  767    16.6        16.6
##                                           Insured 3866    83.4       100.0
##
## ------------------------------------------------------------
## diabetes: Yes
##        var    n miss p.miss              level freq percent cum.percent
##         pa 1203    0    0.0       No (inactive)  771    64.1        64.1
##                                    Yes (active)  432    35.9       100.0
##
##       diet 1203    0    0.0               Good  480    39.9        39.9
##                                      Poor/Fair  421    35.0        74.9
##                                 Very good/Excellent  302    25.1       100.0
##
##    bmi_cat 1203   76    6.3             normal  159    14.1        14.1
##                                     underweight    2     0.2        14.3
##                                      overweight  329    29.2        43.5
##                                           Obese  637    56.5       100.0
##
##        hbp 1203  131   10.9         No (normal)  450    42.0        42.0
##                                Yes (hypertension)  622    58.0       100.0
##
##  depressed 1203    0    0.0           Min/mild 1061    88.2        88.2
##                                      Mod/Severe  142    11.8       100.0
##
##    age_cat 1203    0    0.0              18-39   82     6.8         6.8
##                                           40-61  438    36.4        43.2
##                                             62+  683    56.8       100.0
##
##    RaceEth 1203    0    0.0           NH White  375    31.2        31.2
##                                   Hispanic Origin  275    22.9        54.0
##                                    NH Asian/other  250    20.8        74.8
##                                          NH Black  303    25.2       100.0
##
##   RIAGENDR 1203    0    0.0               Male  623    51.8        51.8
##                                          Female  580    48.2       100.0
##
##     income 1203  129   10.7               0-20K  213    19.8        19.8
##                                         20-55K  459    42.7        62.6
##                                        55-100K+  402    37.4       100.0
##
##     HIQ011 1203    3    0.2           Uninsured  125    10.4        10.4
##                                           Insured 1075    89.6       100.0
##
## p-values
##                 pApprox       pExact
## pa         4.058822e-17 2.136146e-17
## diet       4.875493e-02           NA
## bmi_cat    7.568968e-38           NA
## hbp        2.703745e-19 2.334345e-19
## depressed  2.702985e-04 3.079569e-04
## age_cat   7.639089e-136           NA
```

```
## RaceEth    1.970522e-02 1.886902e-02
## RIAGENDR   1.184432e-02 1.157851e-02
## income     9.010821e-02 8.906236e-02
## HIQ011     1.784482e-07 6.032088e-08
##
## Standardize mean differences
##               1 vs 2
## pa         0.27812968
## diet       0.08031334
## bmi_cat    0.47414737
## hbp        0.31235188
## depressed 0.11430108
## age_cat    0.94022441
## RaceEth    0.10213627
## RIAGENDR   0.08249199
## income     0.07541483
## HIQ011     0.18044207
```

```r
# Plots - may or may not be included in report
nhanesdat %>%
  filter(!is.na(age_cat)) %>%
  group_by(age_cat, diabetes) %>% # order matters
  summarise(n = n()) %>%
  mutate(P = n/sum(n),
         Percent = P * 100) %>%
  ggplot(., aes(y=Percent, x = age_cat, fill=diabetes)) + geom_col(width=0.7) +
  scale_fill_brewer(palette ='Blues')+
  theme_bw()+theme(text = element_text(size = 15))+
  ggtitle("Prevalence of Diabetes among Age Groups")

nhanesdat  %>%
  filter(!is.na(bmi_cat)) %>%
  group_by(bmi_cat, diabetes) %>% # order matters
  summarise(n = n()) %>%
  mutate(P = n/sum(n),
         Percent = P * 100) %>%
  ggplot(., aes(y=Percent, x = bmi_cat, fill=diabetes)) + geom_col(width=0.7) +
  scale_fill_brewer(palette ='Blues')+
  theme_bw()+theme(text = element_text(size = 15))+
  ggtitle("Prevalence of Diabetes by Weight (BMI)")
```

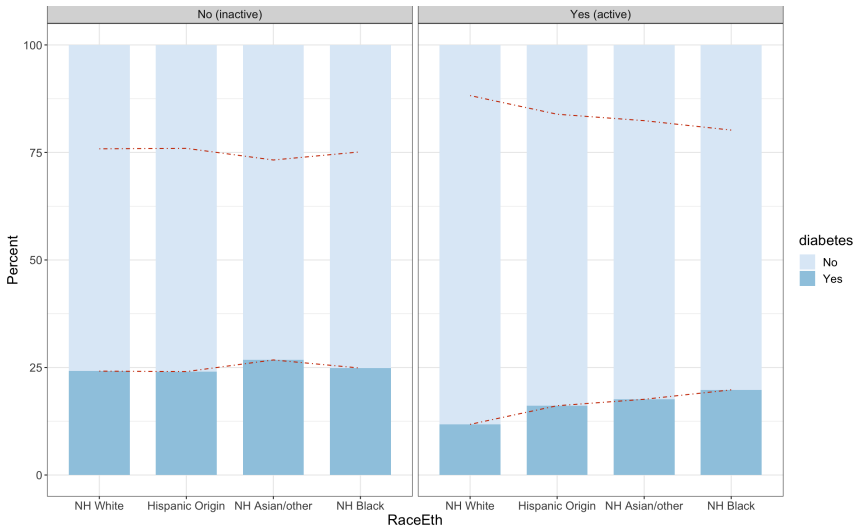





```r
nhanesdat %>%
  filter(!is.na(RaceEth)) %>%
  filter(!is.na(pa)) %>%
  group_by(RaceEth,pa,diabetes) %>% # order matters
  summarise(n = n()) %>%
  mutate(P = n/sum(n),
         Percent = P * 100) %>%
```

```
  ggplot(., aes(y=Percent, x = RaceEth,fill=diabetes)) + geom_col(width=0.7) +
  scale_fill_brewer(palette ='Blues')+
  geom_line(aes(y=Percent,x=RaceEth,group=diabetes),color="orangered3",linetype="dotdash")+
  theme_bw()+theme(text = element_text(size = 15))+
  facet_wrap(~pa)+
  ggtitle("Prevalence of Diabetes Among Race/Ethnicity Groups by Physical Activity")
```

FIGURE 2. Prevalence of Diabetes Among Race/Ethnicity Groups by Physical Recreation Activity




```
##### Multivariate Analysis #####

# MODEL 1: Unweighted basic model
g.base <- glm(diabetes ~ pa + diet + bmi_cat + hbp + depressed + age_cat + RaceEth + RIAGENDR
+ income + HIQ011+
                        pa:RaceEth,
                        data = nhanesdat, family = binomial())
summary(g.base)

##
## Call:
## glm(formula = diabetes ~ pa + diet + bmi_cat + hbp + depressed +
##     age_cat + RaceEth + RIAGENDR + income + HIQ011 + pa:RaceEth,
##     family = binomial(), data = nhanesdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5342  -0.7376  -0.3860  -0.1717   3.0337
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -3.98495    0.23507 -16.952  < 2e-16 ***
## paYes (active)                     -0.66267    0.15015  -4.413 1.02e-05 ***
## dietPoor/Fair                       0.10627    0.09614   1.105 0.269026
## dietVery good/Excellent            -0.13609    0.10287  -1.323 0.185854
## bmi_catunderweight                 -1.42217    0.73747  -1.928 0.053799 .
## bmi_catoverweight                   0.42142    0.12220   3.448 0.000564 ***
## bmi_catObese                        1.14150    0.11907   9.587  < 2e-16 ***
## hbpYes (hypertension)               0.04201    0.08254   0.509 0.610761
## depressedMod/Severe                 0.26032    0.13285   1.959 0.050058 .
## age_cat40-61                        1.75827    0.14316  12.282  < 2e-16 ***
## age_cat62+                          2.66271    0.14755  18.046  < 2e-16 ***
## RaceEthHispanic Origin              0.28463    0.14033   2.028 0.042536 *
```

```
## RaceEthNH Asian/other                   0.67428    0.15278   4.413 1.02e-05 ***
## RaceEthNH Black                         0.32616    0.13852   2.355 0.018543 *
## RIAGENDRFemale                         -0.28573    0.08085  -3.534 0.000410 ***
## income20-55K                            0.13903    0.11138   1.248 0.211938
## income55-100K+                          0.15352    0.11696   1.313 0.189336
## HIQ011Insured                           0.03849    0.13391   0.287 0.773768
## paYes (active):RaceEthHispanic Origin   0.52633    0.23099   2.279 0.022689 *
## paYes (active):RaceEthNH Asian/other    0.57624    0.22904   2.516 0.011874 *
## paYes (active):RaceEthNH Black          0.53100    0.22072   2.406 0.016139 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4663.4  on 4580  degrees of freedom
## Residual deviance: 3881.9  on 4560  degrees of freedom
##   (1271 observations deleted due to missingness)
## AIC: 3923.9
##
## Number of Fisher Scoring iterations: 6

# Variable selection, start with complete set (2858 observations used)
completeFun <- function(data, desiredCols) {
  completeVec <- complete.cases(data[, desiredCols])
  return(data[completeVec, ])
}
nhanes.cplt<-completeFun(nhanesdat,c("diabetes","pa","diet","bmi_cat","hbp","depressed",
                             "age_cat","RaceEth","RIAGENDR","income","HIQ011"))

g1 <- glm(diabetes ~ pa + diet + bmi_cat + hbp + depressed + age_cat + RaceEth + RIAGENDR
+income + HIQ011 +
                     pa:RaceEth,
                     data = nhanes.cplt, family = binomial())
g2 <-step(g1,direction = "both")

## Start:  AIC=3923.9
## diabetes ~ pa + diet + bmi_cat + hbp + depressed + age_cat +
##     RaceEth + RIAGENDR + income + HIQ011 + pa:RaceEth
##
##              Df Deviance    AIC
## - income      2   3883.9 3921.9
## - HIQ011      1   3882.0 3922.0
## - hbp         1   3882.2 3922.2
## <none>            3881.9 3923.9
## - diet        2   3886.6 3924.6
## - depressed   1   3885.7 3925.7
## - pa:RaceEth  3   3891.4 3927.4
## - RIAGENDR    1   3894.4 3934.4
## - bmi_cat     3   4015.8 4051.8
## - age_cat     2   4344.4 4382.4
##
## Step:  AIC=3921.89
## diabetes ~ pa + diet + bmi_cat + hbp + depressed + age_cat +
##     RaceEth + RIAGENDR + HIQ011 + pa:RaceEth
##
##              Df Deviance    AIC
## - HIQ011      1   3884.1 3920.1
## - hbp         1   3884.1 3920.1
## <none>            3883.9 3921.9
## - diet        2   3888.1 3922.1
## - depressed   1   3887.2 3923.2
```

```
## + income      2   3881.9 3923.9
## - pa:RaceEth  3   3893.1 3925.1
## - RIAGENDR    1   3896.7 3932.7
## - bmi_cat     3   4020.9 4052.9
## - age_cat     2   4345.7 4379.7
##
## Step:  AIC=3920.07
## diabetes ~ pa + diet + bmi_cat + hbp + depressed + age_cat +
##     RaceEth + RIAGENDR + pa:RaceEth
##
##              Df Deviance    AIC
## - hbp         1   3884.3 3918.3
## <none>            3884.1 3920.1
## - diet        2   3888.2 3920.2
## - depressed   1   3887.4 3921.4
## + HIQ011      1   3883.9 3921.9
## + income      2   3882.0 3922.0
## - pa:RaceEth  3   3893.3 3923.3
## - RIAGENDR    1   3896.8 3930.8
## - bmi_cat     3   4021.6 4051.6
## - age_cat     2   4363.4 4395.4
##
## Step:  AIC=3918.3
## diabetes ~ pa + diet + bmi_cat + depressed + age_cat + RaceEth +
##     RIAGENDR + pa:RaceEth
##
##              Df Deviance    AIC
## <none>            3884.3 3918.3
## - diet        2   3888.4 3918.4
## - depressed   1   3887.6 3919.6
## + hbp         1   3884.1 3920.1
## + HIQ011      1   3884.1 3920.1
## + income      2   3882.2 3920.2
## - pa:RaceEth  3   3893.7 3921.7
## - RIAGENDR    1   3897.2 3929.2
## - bmi_cat     3   4022.7 4050.7
## - age_cat     2   4401.9 4431.9
```

```
summary(g1)
```

```
##
## Call:
## glm(formula = diabetes ~ pa + diet + bmi_cat + hbp + depressed +
##     age_cat + RaceEth + RIAGENDR + income + HIQ011 + pa:RaceEth,
##     family = binomial(), data = nhanes.cplt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5342  -0.7376  -0.3860  -0.1717   3.0337
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -3.98495    0.23507 -16.952  < 2e-16 ***
## paYes (active)                      -0.66267    0.15015  -4.413 1.02e-05 ***
## dietPoor/Fair                        0.10627    0.09614   1.105 0.269026
## dietVery good/Excellent             -0.13609    0.10287  -1.323 0.185854
## bmi_catunderweight                  -1.42217    0.73747  -1.928 0.053799 .
## bmi_catoverweight                    0.42142    0.12220   3.448 0.000564 ***
## bmi_catObese                         1.14150    0.11907   9.587  < 2e-16 ***
## hbpYes (hypertension)                0.04201    0.08254   0.509 0.610761
## depressedMod/Severe                  0.26032    0.13285   1.959 0.050058 .
```

```
## age_cat40-61                              1.75827    0.14316  12.282  < 2e-16 ***
## age_cat62+                                2.66271    0.14755  18.046  < 2e-16 ***
## RaceEthHispanic Origin                    0.28463    0.14033   2.028 0.042536 *
## RaceEthNH Asian/other                     0.67428    0.15278   4.413 1.02e-05 ***
## RaceEthNH Black                           0.32616    0.13852   2.355 0.018543 *
## RIAGENDRFemale                           -0.28573    0.08085  -3.534 0.000410 ***
## income20-55K                              0.13903    0.11138   1.248 0.211938
## income55-100K+                            0.15352    0.11696   1.313 0.189336
## HIQ011Insured                             0.03849    0.13391   0.287 0.773768
## paYes (active):RaceEthHispanic Origin     0.52633    0.23099   2.279 0.022689 *
## paYes (active):RaceEthNH Asian/other      0.57624    0.22904   2.516 0.011874 *
## paYes (active):RaceEthNH Black            0.53100    0.22072   2.406 0.016139 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4663.4  on 4580  degrees of freedom
## Residual deviance: 3881.9  on 4560  degrees of freedom
## AIC: 3923.9
##
## Number of Fisher Scoring iterations: 6

summary(g2)

##
## Call:
## glm(formula = diabetes ~ pa + diet + bmi_cat + depressed + age_cat +
##     RaceEth + RIAGENDR + pa:RaceEth, family = binomial(), data = nhanes.cplt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5002  -0.7412  -0.3831  -0.1746   3.0380
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -3.82771    0.19441 -19.689  < 2e-16 ***
## paYes (active)                        -0.64765    0.14912  -4.343 1.40e-05 ***
## dietPoor/Fair                          0.09648    0.09569   1.008 0.313335
## dietVery good/Excellent               -0.12950    0.10260  -1.262 0.206913
## bmi_catunderweight                    -1.42857    0.73704  -1.938 0.052593 .
## bmi_catoverweight                      0.42943    0.12207   3.518 0.000435 ***
## bmi_catObese                           1.15609    0.11867   9.742  < 2e-16 ***
## depressedMod/Severe                    0.24140    0.13176   1.832 0.066941 .
## age_cat40-61                           1.76787    0.14148  12.496  < 2e-16 ***
## age_cat62+                             2.67207    0.14295  18.692  < 2e-16 ***
## RaceEthHispanic Origin                 0.27819    0.13925   1.998 0.045736 *
## RaceEthNH Asian/other                  0.68385    0.15246   4.485 7.28e-06 ***
## RaceEthNH Black                        0.32484    0.13808   2.353 0.018643 *
## RIAGENDRFemale                        -0.28917    0.08065  -3.585 0.000337 ***
## paYes (active):RaceEthHispanic Origin  0.51864    0.23036   2.251 0.024359 *
## paYes (active):RaceEthNH Asian/other   0.57219    0.22898   2.499 0.012460 *
## paYes (active):RaceEthNH Black         0.52796    0.22047   2.395 0.016634 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4663.4  on 4580  degrees of freedom
## Residual deviance: 3884.3  on 4564  degrees of freedom
## AIC: 3918.3
```

```
##
## Number of Fisher Scoring iterations: 6

# MODEL 2 : Unweighted final model
g.final <-glm(diabetes ~ pa + bmi_cat + depressed + age_cat + RaceEth + RIAGENDR +pa:RaceEth,
              data = nhanesdat, family = binomial())
summary(g.final)

##
## Call:
## glm(formula = diabetes ~ pa + bmi_cat + depressed + age_cat +
##     RaceEth + RIAGENDR + pa:RaceEth, family = binomial(), data = nhanesdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5223  -0.7389  -0.3790  -0.1789   2.9900
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -3.88495    0.17355 -22.386  < 2e-16 ***
## paYes (active)                       -0.57365    0.13811  -4.153 3.27e-05 ***
## bmi_catunderweight                   -1.60857    0.73049  -2.202  0.02766 *
## bmi_catoverweight                     0.43684    0.11183   3.906 9.37e-05 ***
## bmi_catObese                          1.18961    0.10751  11.065  < 2e-16 ***
## depressedMod/Severe                   0.32647    0.11844   2.756  0.00584 **
## age_cat40-61                          1.77887    0.13108  13.571  < 2e-16 ***
## age_cat62+                            2.69472    0.13158  20.480  < 2e-16 ***
## RaceEthHispanic Origin                0.28484    0.12629   2.255  0.02411 *
## RaceEthNH Asian/other                 0.69835    0.14017   4.982 6.29e-07 ***
## RaceEthNH Black                       0.22383    0.12568   1.781  0.07492 .
## RIAGENDRFemale                       -0.24228    0.07385  -3.281  0.00104 **
## paYes (active):RaceEthHispanic Origin  0.38679    0.21070   1.836  0.06640 .
## paYes (active):RaceEthNH Asian/other   0.43767    0.21193   2.065  0.03891 *
## paYes (active):RaceEthNH Black         0.52963    0.20205   2.621  0.00876 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5546.1  on 5429  degrees of freedom
## Residual deviance: 4616.4  on 5415  degrees of freedom
##   (422 observations deleted due to missingness)
## AIC: 4646.4
##
## Number of Fisher Scoring iterations: 6

# MODEL 3: Weighted final model
library(survey)
library(srvyr)
library(sjPlot)
require(sjPlot)

nhanessvy <- nhanesdat %>%
  as_survey_design(ids = SDMVPSU, strata = SDMVSTRA, nest = T, weights = WTMEC2YR)

g.svyfull <- svyglm(diabetes ~ pa + bmi_cat + depressed + age_cat + RaceEth + RIAGENDR
+pa:RaceEth,
               design=nhanessvy, family = quasibinomial(link = "logit"))
g.svyredu <- svyglm(diabetes ~ pa + bmi_cat + depressed + age_cat + RaceEth + RIAGENDR,
                    design=nhanessvy, family = quasibinomial(link = "logit"))
summary(g.svyfull)
```

```
##
## Call:
## svyglm(formula = diabetes ~ pa + bmi_cat + depressed + age_cat +
##     RaceEth + RIAGENDR + pa:RaceEth, design = nhanessvy, family = quasibinomial(link =
"logit"))
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -3.8009     0.2598 -14.631   0.0434 *
## paYes (active)                        -0.6670     0.1815  -3.675   0.1691
## bmi_catunderweight                     0.6142     0.9946   0.617   0.6478
## bmi_catoverweight                      0.4649     0.1694   2.745   0.2224
## bmi_catObese                           1.3397     0.1879   7.132   0.0887 .
## depressedMod/Severe                    0.3342     0.1479   2.259   0.2653
## age_cat40-61                           1.6023     0.1079  14.847   0.0428 *
## age_cat62+                             2.4618     0.2163  11.380   0.0558 .
## RaceEthHispanic Origin                 0.0559     0.1452   0.385   0.7661
## RaceEthNH Asian/other                  0.4183     0.2005   2.086   0.2846
## RaceEthNH Black                        0.1431     0.1317   1.087   0.4734
## RIAGENDRFemale                        -0.2714     0.1164  -2.332   0.2579
## paYes (active):RaceEthHispanic Origin  0.4441     0.1810   2.454   0.2464
## paYes (active):RaceEthNH Asian/other   0.5772     0.4353   1.326   0.4114
## paYes (active):RaceEthNH Black         0.6361     0.2721   2.338   0.2573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.064706)
##
## Number of Fisher Scoring iterations: 5

summary(g.svyredu)

##
## Call:
## svyglm(formula = diabetes ~ pa + bmi_cat + depressed + age_cat +
##     RaceEth + RIAGENDR, design = nhanessvy, family = quasibinomial(link = "logit"))
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.8936     0.2729 -14.265 0.000140 ***
## paYes (active)          -0.4526     0.1164  -3.888 0.017722 *
## bmi_catunderweight       0.6170     1.0117   0.610 0.574888
## bmi_catoverweight        0.4826     0.1755   2.750 0.051397 .
## bmi_catObese             1.3592     0.1900   7.155 0.002019 **
## depressedMod/Severe      0.3298     0.1497   2.203 0.092365 .
## age_cat40-61             1.5872     0.1077  14.732 0.000124 ***
## age_cat62+               2.4359     0.2122  11.480 0.000329 ***
## RaceEthHispanic Origin   0.2315     0.1299   1.782 0.149252
## RaceEthNH Asian/other    0.6760     0.1587   4.259 0.013069 *
## RaceEthNH Black          0.4099     0.1194   3.434 0.026430 *
## RIAGENDRFemale          -0.2659     0.1203  -2.211 0.091552 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.060655)
```

```
##
## Number of Fisher Scoring iterations: 5

anova(g.svyfull,g.svyredu)

## Working (Rao-Scott+F) LRT for pa:RaceEth
##  in svyglm(formula = diabetes ~ pa + bmi_cat + depressed + age_cat +
##     RaceEth + RIAGENDR + pa:RaceEth, design = nhanessvy, family = quasibinomial(link =
"logit"))
## Working 2logLR =  7.843778 p= 0.39029
## (scale factors:  2.2 0.74 0.11 );  denominator df= 1

# Compare three models, and the prevalence of diabetes between unweighted and weighted data
library(epiR)

tab_model(g.base,g.final,g.svyfull,g.svyredu,CSS = css_theme("cells"))
```

	diabetes			diabetes			diabetes			diabetes		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.02	0.01 – 0.03	**<0.001**	0.02	0.01 – 0.03	**<0.001**	0.02	0.01 – 0.04	**0.043**	0.02	0.01 – 0.03	**<0.001**
pa [Yes (active)]	0.52	0.38 – 0.69	**<0.001**	0.56	0.43 – 0.74	**<0.001**	0.51	0.36 – 0.73	0.169	0.64	0.51 – 0.80	**0.018**
diet [Poor/Fair]	1.11	0.92 – 1.34	0.269									
diet [Very good/Excellent]	0.87	0.71 – 1.07	0.186									
bmi_cat [underweight]	0.24	0.04 – 0.81	0.054	0.20	0.03 – 0.66	**0.028**	1.85	0.26 – 12.98	0.648	1.85	0.26 – 13.46	0.575
bmi_cat [overweight]	1.52	1.20 – 1.94	**0.001**	1.55	1.25 – 1.93	**<0.001**	1.59	1.14 – 2.22	0.222	1.62	1.15 – 2.29	0.051
bmi_cat [Obese]	3.13	2.49 – 3.97	**<0.001**	3.29	2.67 – 4.07	**<0.001**	3.82	2.64 – 5.52	0.089	3.89	2.68 – 5.65	**0.002**
hbp [Yes (hypertension)]	1.04	0.89 – 1.23	0.611									
depressed [Mod/Severe]	1.30	1.00 – 1.68	0.050	1.39	1.10 – 1.75	**0.006**	1.40	1.05 – 1.87	0.265	1.39	1.04 – 1.87	0.092
age_cat [40-61]	5.80	4.41 – 7.74	**<0.001**	5.92	4.61 – 7.71	**<0.001**	4.96	4.02 – 6.13	**0.043**	4.89	3.96 – 6.04	**<0.001**
age_cat [62+]	14.34	10.81 – 19.28	**<0.001**	14.80	11.51 – 19.28	**<0.001**	11.73	7.67 – 17.92	0.056	11.43	7.54 – 17.32	**<0.001**
RaceEth [Hispanic Origin]	1.33	1.01 – 1.75	**0.043**	1.33	1.04 – 1.70	**0.024**	1.06	0.80 – 1.41	0.766	1.26	0.98 – 1.63	0.149
RaceEth [NH Asian/other]	1.96	1.45 – 2.65	**<0.001**	2.01	1.53 – 2.65	**<0.001**	1.52	1.03 – 2.25	0.285	1.97	1.44 – 2.68	**0.013**
RaceEth [NH Black]	1.39	1.06 – 1.82	**0.019**	1.25	0.98 – 1.60	0.075	1.15	0.89 – 1.49	0.473	1.51	1.19 – 1.90	**0.026**
RIAGENDR [Female]	0.75	0.64 – 0.88	**<0.001**	0.78	0.68 – 0.91	**0.001**	0.76	0.61 – 0.96	0.258	0.77	0.61 – 0.97	0.092
income [20-55K]	1.15	0.92 – 1.43	0.212									
income [55-100K+]	1.17	0.93 – 1.47	0.189									
HIQ011 [Insured]	1.04	0.80 – 1.36	0.774									
pa [Yes (active)] * RaceEth [Hispanic Origin]	1.69	1.08 – 2.66	**0.023**	1.47	0.97 – 2.22	0.066	1.56	1.09 – 2.22	0.246			
pa [Yes (active)] * RaceEth [NH Asian/other]	1.78	1.14 – 2.79	**0.012**	1.55	1.02 – 2.35	**0.039**	1.78	0.76 – 4.18	0.411			
pa [Yes (active)] * RaceEth [NH Black]	1.70	1.10 – 2.62	**0.016**	1.70	1.14 – 2.53	**0.009**	1.89	1.11 – 3.22	0.257			
Observations	4581			5430			5430			5430		
R^2 Tjur	0.161			0.161			0.162 / -4549.256			0.160 / -1139.754		

```
plogis(-3.98495) # prob. of having diabetes for the reference group (base model)

## [1] 0.01825397

plogis(-3.88495) # prob. of having diabetes for the reference group (reduced model -
unweighted)

## [1] 0.0201351

plogis(-3.8009) # prob. of having diabetes for the reference group (reduced moderl -weighted)
```

```
## [1] 0.02186202
```

```
plogis(-3.8936) # prob. of having diabetes for the reference group (final model - weighted
without interaction)
```

```
## [1] 0.01996515
```

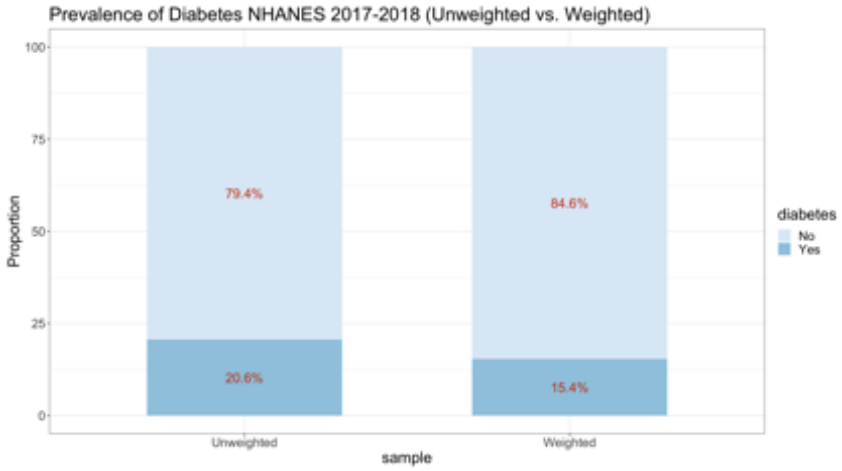
```
nhanesdat %>%
  group_by(diabetes) %>%
  summarise(n = n()) %>%
  mutate(P=n/sum(n),
         Proportion = P*100)
```

```
## # A tibble: 2 x 4
##   diabetes     n     P Proportion
##   <fct>    <int> <dbl>      <dbl>
## 1 No        4649 0.794       79.4
## 2 Yes       1203 0.206       20.6
```

```
nhanessvy %>%
  group_by(diabetes) %>%
  summarise(Proportion = survey_mean(),
            total = survey_total())
```

```
## # A tibble: 2 x 5
##   diabetes Proportion Proportion_se      total total_se
##   <fct>         <dbl>         <dbl>      <dbl>    <dbl>
## 1 No            0.846       0.00613 208898550. 7270330.
## 2 Yes           0.154       0.00613  38058436. 1850875.
```

```
diabetes <-c("Yes","Yes","No","No")
sample <-c("Weighted","Unweighted","Weighted","Unweighted")
Proportion <-c(15.4,20.6,84.6,79.4)
diab.prev<-data.frame(diabates=diabetes,Sample=sample,Proportion=Proportion)
ggplot(diab.prev, aes(y=Proportion, x = sample,fill=diabetes)) +
  geom_col(width=0.6) +
  geom_text(aes(label = paste0(round(Proportion,1),"%")),
            position = position_stack(vjust = 0.5),size=6,color="orangered3")+
  scale_fill_brewer(palette ='Blues')+
  theme_bw()+theme(text = element_text(size = 20))+
  ggtitle("Prevalence of Diabetes NHANES 2017-2018 (Unweighted vs. Weighted)")
```



```
p1<-plot_model(g.final, title = "ORs for Model 2",
               sortOdds = F, showModelSummary = T)
p2<-plot_model(g.svyfull, title = "ORs for Model 3",
        sortOdds = F, showModelSummary = T)
p3<-plot_model(g.svyredu, title = "ORs for Model 4 ",
               sortOdds = F, showModelSummary = T)

grid.arrange(p1, p2,p3,ncol=3)
```