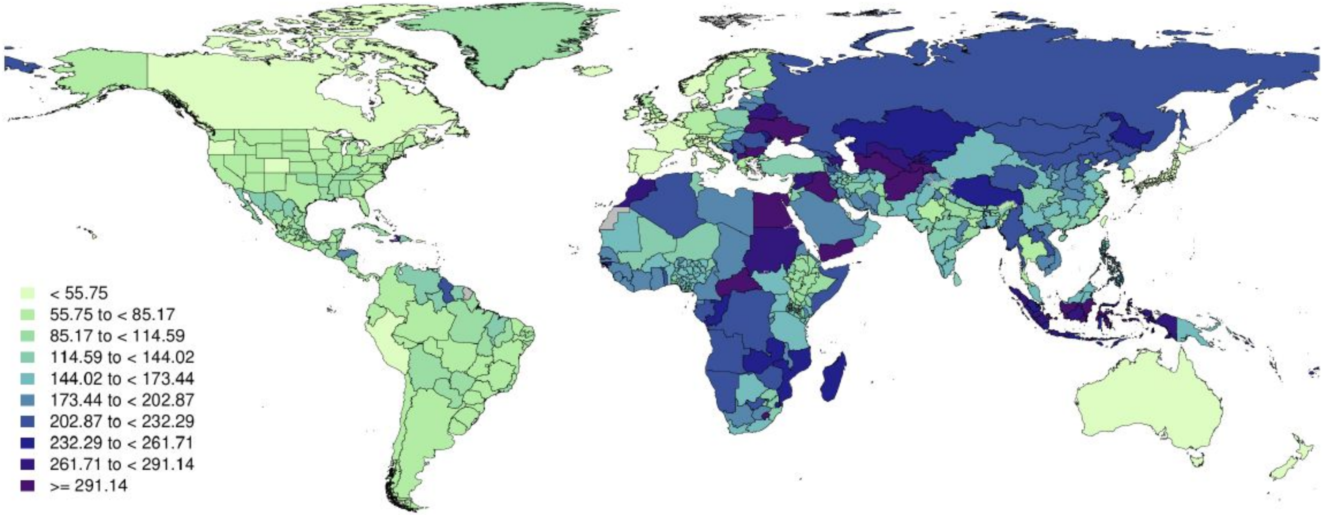
# HMS 520 Final Project: Predicting High Blood Pressure in USA using NHANES data

Nikki DeCleene & Aisha Twalibu

# CVD mortality attributable to high systolic blood pressure per 100,000 in 2022



https://www.jacc.org/doi/10.1016/j.jacc.2023.11.007

# Despite large burden of high blood pressure, many locations do not have population-representative data on measured blood pressure

# Project goal: predict measured high blood pressure from interview responses alone

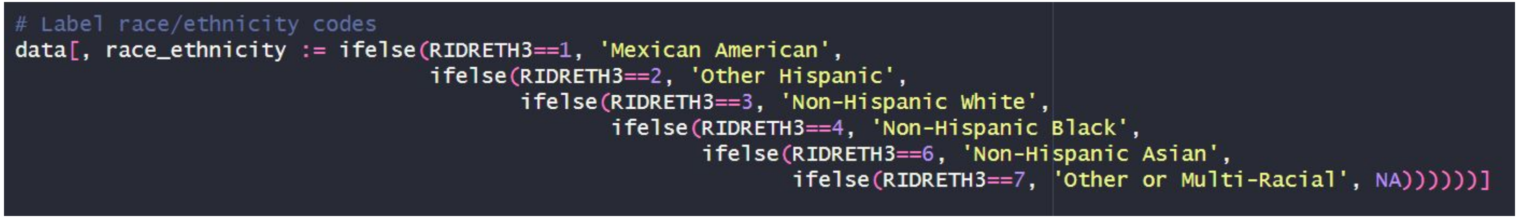
# National Health and Nutrition Examination Survey (NHANES)

- Nationally representative study in the United States designed to study health and nutrition
- Began in the early 1960s, became a continuous program
- Combines interviews and physical examinations
    - Including systolic and diastolic blood pressure measurements
- Limitation: does not report data at the state-level

https://www.cdc.gov/nchs/images/nhanes/nhanes_apple_color_tagline.jpg

# Part 1: Data processing

# Demographics data

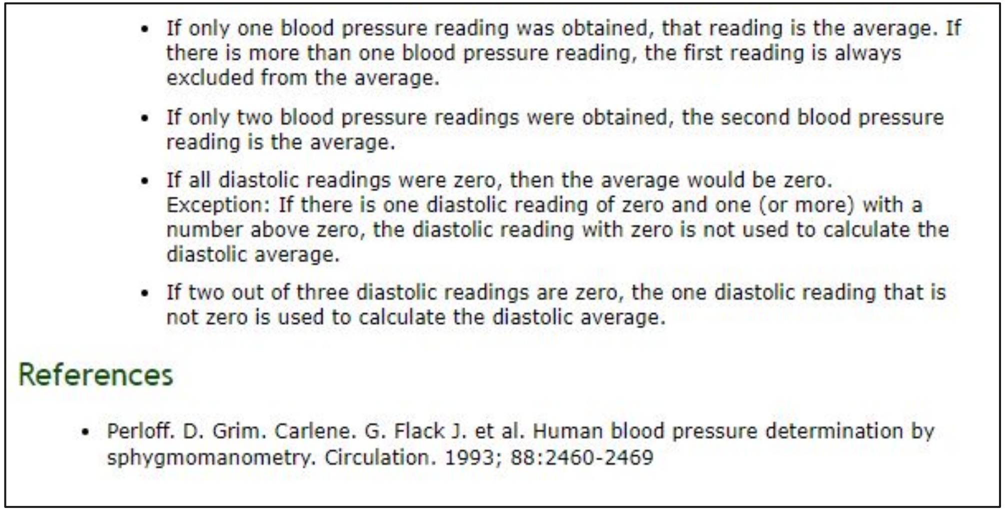


```
# Label race/ethnicity codes
data[, race_ethnicity := ifelse(RIDRETH3==1, 'Mexican American',
                                ifelse(RIDRETH3==2, 'Other Hispanic',
                                       ifelse(RIDRETH3==3, 'Non-Hispanic White',
                                              ifelse(RIDRETH3==4, 'Non-Hispanic Black',
                                                     ifelse(RIDRETH3==6, 'Non-Hispanic Asian',
                                                            ifelse(RIDRETH3==7, 'Other or Multi-Racial', NA))))))]
```

- Label codes for:
  - Sex
  - Race/ethnicity
  - Education
  - Marital status
- Create:
  - Proportion of time lived in US
- Subset:
  - Age

# Examination data

- Create:
  - Average of 3 systolic blood pressure readings
  - Average of 3 diastolic blood pressure readings

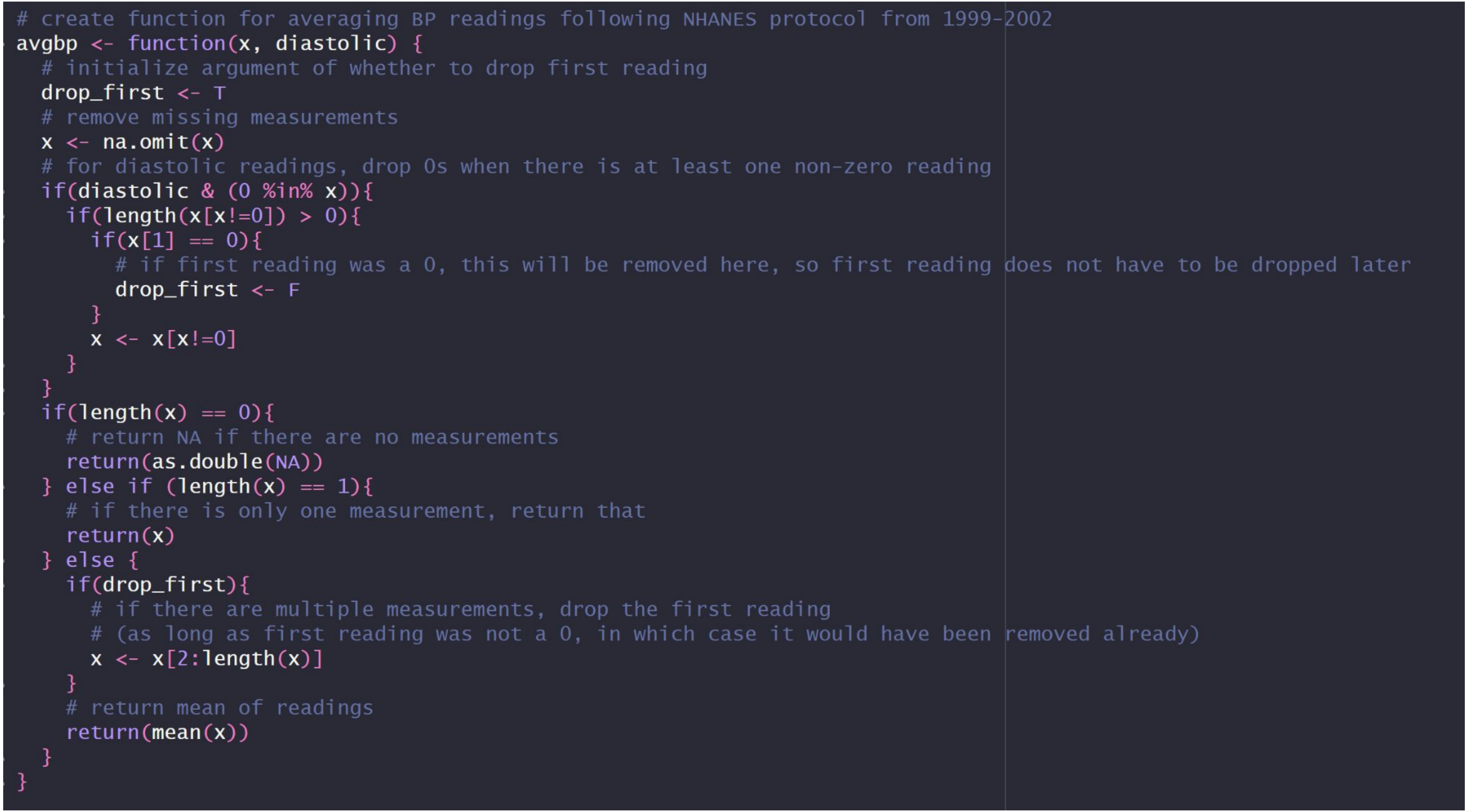


- If only one blood pressure reading was obtained, that reading is the average. If there is more than one blood pressure reading, the first reading is always excluded from the average.
- If only two blood pressure readings were obtained, the second blood pressure reading is the average.
- If all diastolic readings were zero, then the average would be zero. Exception: If there is one diastolic reading of zero and one (or more) with a number above zero, the diastolic reading with zero is not used to calculate the diastolic average.
- If two out of three diastolic readings are zero, the one diastolic reading that is not zero is used to calculate the diastolic average.

## References

- Perloff. D. Grim. Carlene. G. Flack J. et al. Human blood pressure determination by sphygmomanometry. Circulation. 1993; 88:2460-2469

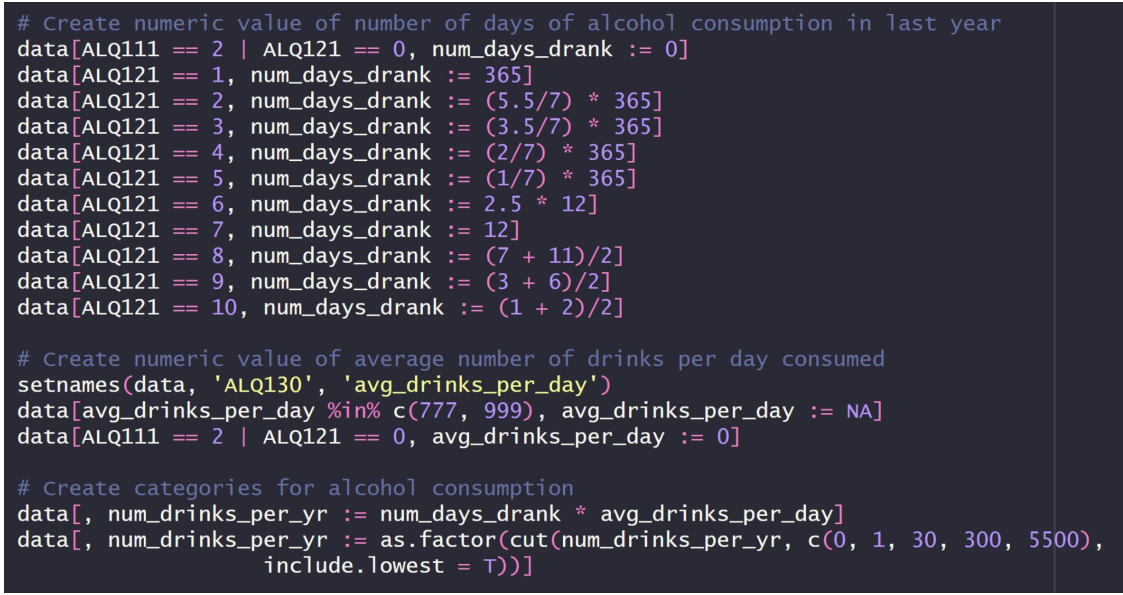
https://wwwn.cdc.gov/Nchs/Nhanes/1999-2000/BPX.htm



```
# create function for averaging BP readings following NHANES protocol from 1999-2002
avgbp <- function(x, diastolic) {
  # initialize argument of whether to drop first reading
  drop_first <- T
  # remove missing measurements
  x <- na.omit(x)
  # for diastolic readings, drop 0s when there is at least one non-zero reading
  if(diastolic & (0 %in% x)){
    if(length(x[x!=0]) > 0){
      if(x[1] == 0){
        # if first reading was a 0, this will be removed here, so first reading does not have to be dropped later
        drop_first <- F
      }
      x <- x[x!=0]
    }
  }
  if(length(x) == 0){
    # return NA if there are no measurements
    return(as.double(NA))
  } else if (length(x) == 1){
    # if there is only one measurement, return that
    return(x)
  } else {
    if(drop_first){
      # if there are multiple measurements, drop the first reading
      # (as long as first reading was not a 0, in which case it would have been removed already)
      x <- x[2:length(x)]
    }
    # return mean of readings
    return(mean(x))
  }
}
```

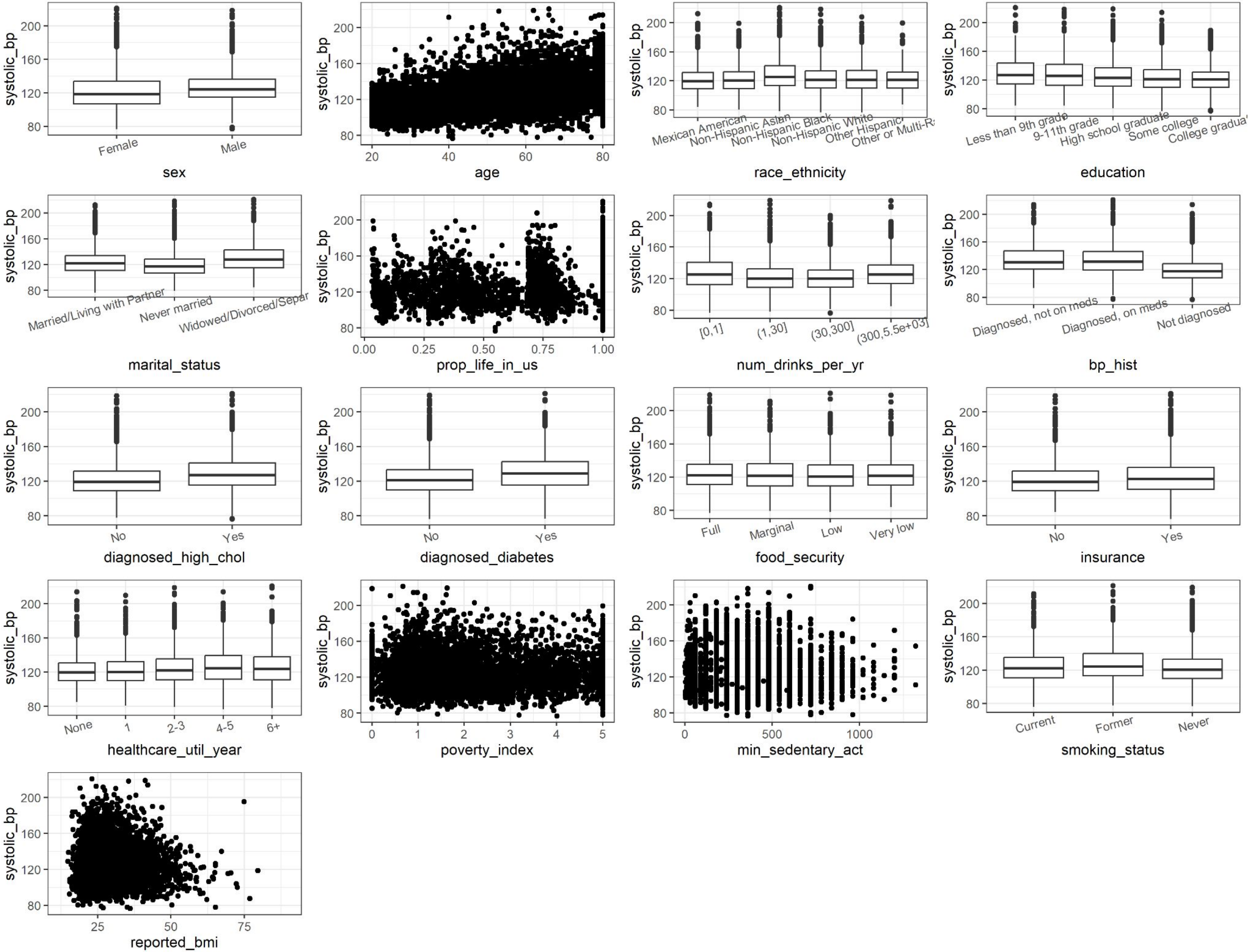
# Questionnaire data

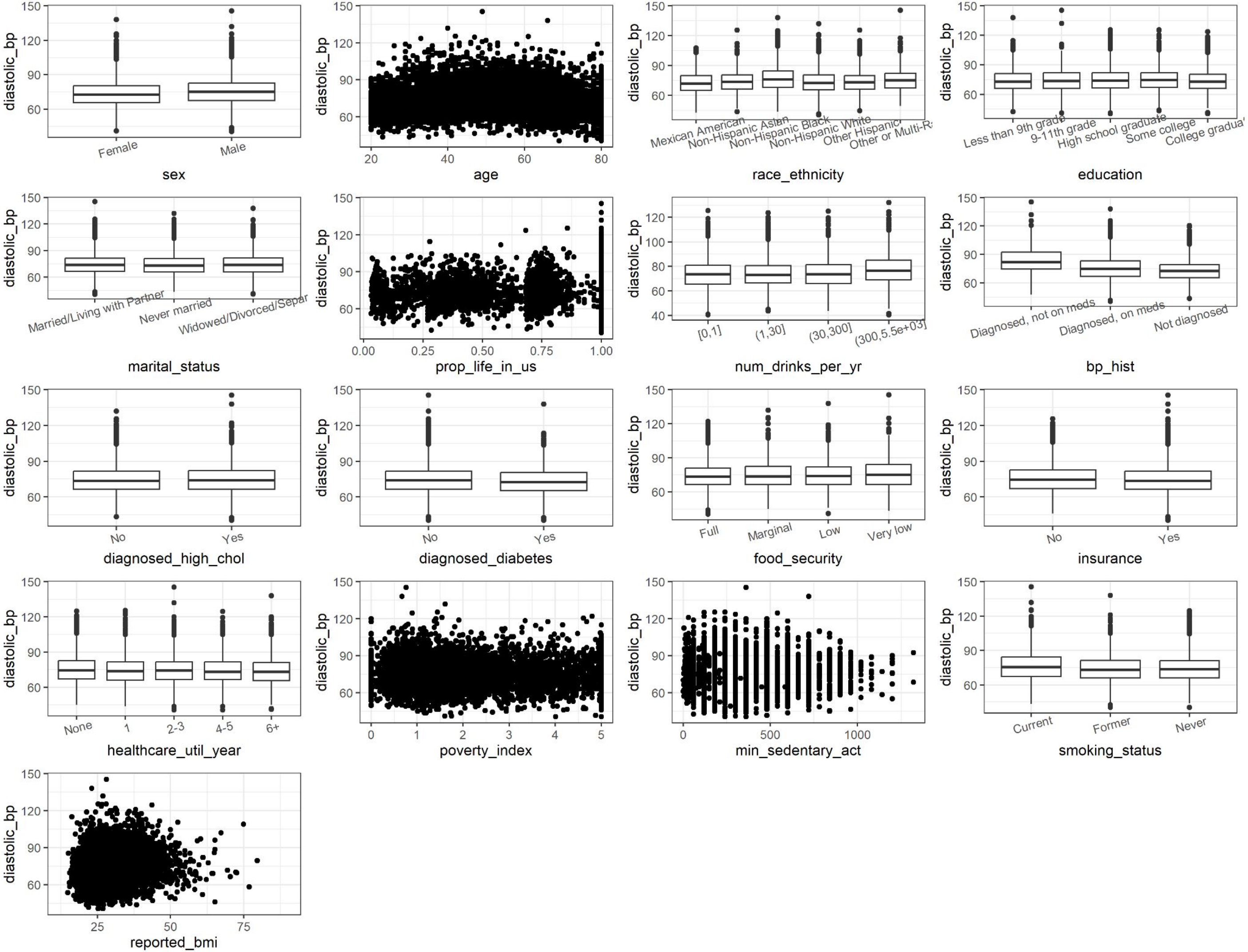
- Label codes for:
  - Diagnosed high cholesterol
  - Diagnosed diabetes
  - Food security
  - Insurance status
  - Healthcare utilization
- Create:
  - Number of alcoholic drinks consumed in a year
  - Blood pressure history
  - Smoking status
  - Self-reported BMI
- Deal with missing codes:
  - Minutes of sedentary activity
- Rename:
  - Poverty index



```
# Create numeric value of number of days of alcohol consumption in last year
data[ALQ111 == 2 | ALQ121 == 0, num_days_drank := 0]
data[ALQ121 == 1, num_days_drank := 365]
data[ALQ121 == 2, num_days_drank := (5.5/7) * 365]
data[ALQ121 == 3, num_days_drank := (3.5/7) * 365]
data[ALQ121 == 4, num_days_drank := (2/7) * 365]
data[ALQ121 == 5, num_days_drank := (1/7) * 365]
data[ALQ121 == 6, num_days_drank := 2.5 * 12]
data[ALQ121 == 7, num_days_drank := 12]
data[ALQ121 == 8, num_days_drank := (7 + 11)/2]
data[ALQ121 == 9, num_days_drank := (3 + 6)/2]
data[ALQ121 == 10, num_days_drank := (1 + 2)/2]

# Create numeric value of average number of drinks per day consumed
setnames(data, 'ALQ130', 'avg_drinks_per_day')
data[avg_drinks_per_day %in% c(777, 999), avg_drinks_per_day := NA]
data[ALQ111 == 2 | ALQ121 == 0, avg_drinks_per_day := 0]

# Create categories for alcohol consumption
data[, num_drinks_per_yr := num_days_drank * avg_drinks_per_day]
data[, num_drinks_per_yr := as.factor(cut(num_drinks_per_yr, c(0, 1, 30, 300, 5500),
                   include.lowest = T))]
```

# Part 2: Exploratory analysis


systolic_bp
200
160
120
80
Female
Male
sex
20
40
60
80
age
Mexican American
Non-Hispanic Asian
Non-Hispanic Black
Non-Hispanic White
Other Hispanic
Other or Multi-Ra
race_ethnicity
Less than 9th grade
9-11th grade
High school graduate
Some college
College gradua
education
Married/Living with Partner
Never married
Widowed/Divorced/Separ
marital_status
0.00
0.25
0.50
0.75
1.00
prop_life_in_us
[0,1]
(1,30]
(30,300]
(300,5.5e+03]
num_drinks_per_yr
Diagnosed, not on meds
Diagnosed, on meds
Not diagnosed
bp_hist
No
Yes
diagnosed_high_chol
No
Yes
diagnosed_diabetes
Full
Marginal
Low
Very low
food_security
No
Yes
insurance
None
1
2-3
4-5
6+
healthcare_util_year
0
1
2
3
4
5
poverty_index
0
500
1000
min_sedentary_act
Current
Former
Never
smoking_status
25
50
75
reported_bmi

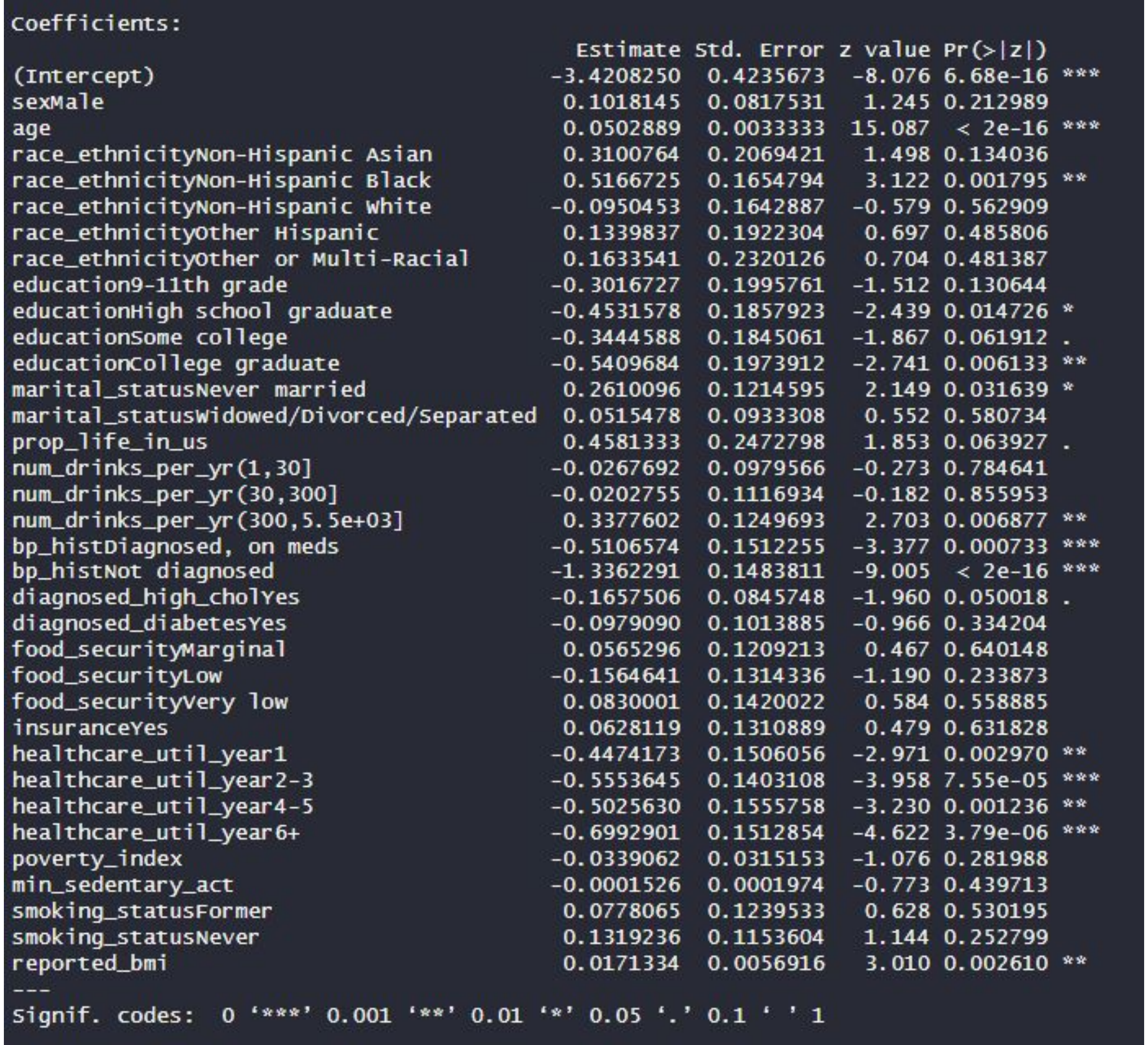

diastolic_bp
sex
Female
Male
age
race_ethnicity
Mexican American
Non-Hispanic Asian
Non-Hispanic Black
Non-Hispanic White
Other Hispanic
Other or Multi-Ra
education
Less than 9th grade
9-11th grade
High school graduate
Some college
College graduat
marital_status
Married/Living with Partner
Never married
Widowed/Divorced/Separ
prop_life_in_us
num_drinks_per_yr
[0,1]
(1,30]
(30,300]
(300,5.5e+03]
bp_hist
Diagnosed, not on meds
Diagnosed, on meds
Not diagnosed
diagnosed_high_chol
diagnosed_diabetes
No
Yes
food_security
Full
Marginal
Low
Very low
insurance
healthcare_util_year
None
1
2-3
4-5
6+
poverty_index
min_sedentary_act
smoking_status
Current
Former
Never
reported_bmi

# Part 3: Prediction

# Fitting logistic regression model on training data

```
high_measured_bp =

ifelse(systolic_bp >= 140 | diastolic_bp >= 90, 1,

            ifelse(systolic_bp < 140 | diastolic_bp < 90, 0, NA))
```

```
glm(high_measured_bp ~ sex + age + race_ethnicity + education + marital_status + prop_life_in_us +

     num_drinks_per_yr + bp_hist + diagnosed_high_chol + diagnosed_diabetes + food_security + insurance +

     healthcare_util_year + poverty_index + min_sedentary_act + smoking_status + reported_bmi,

    family = binomial(link='logit'), data = data[test == 0])
```

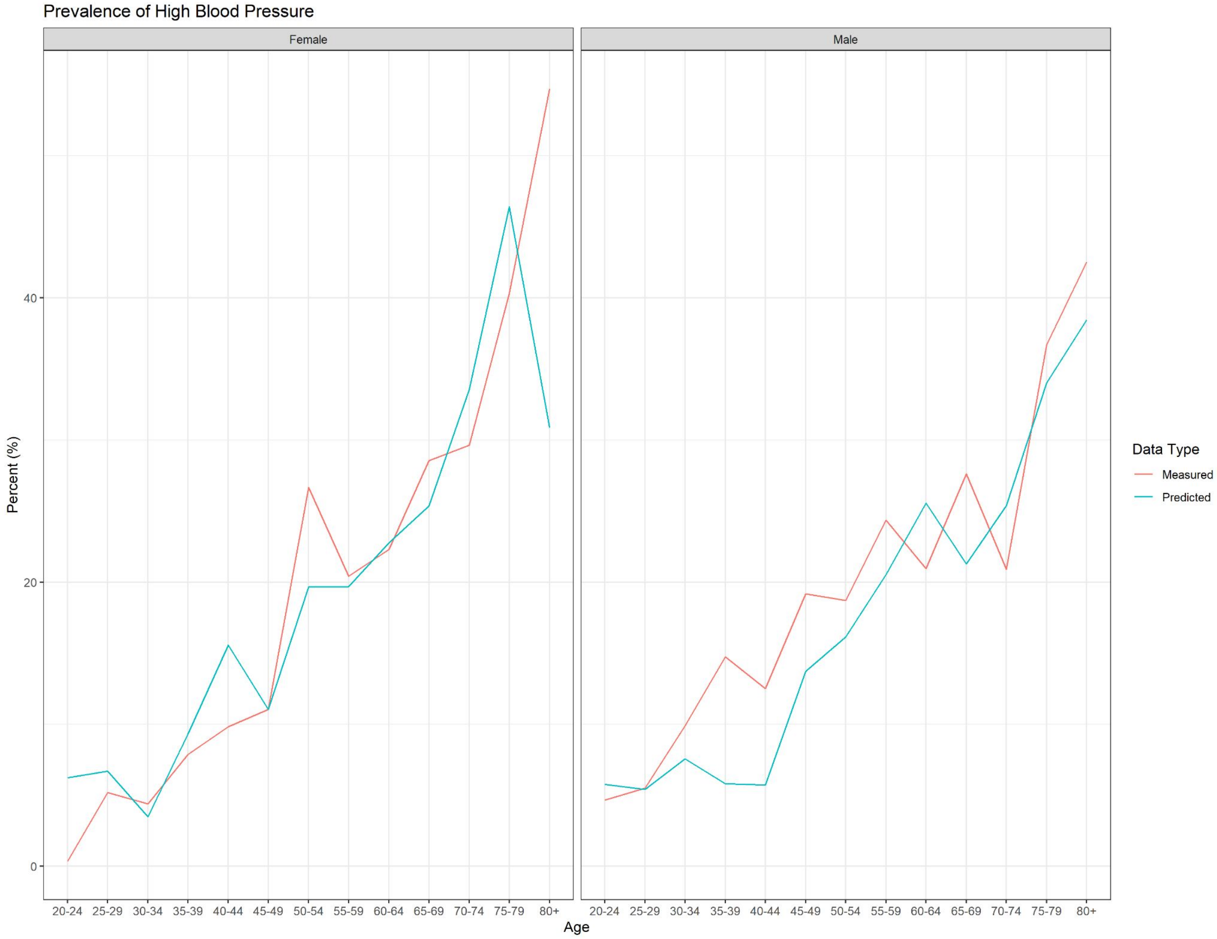


```
Coefficients:
                                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)                              -3.4208250  0.4235673  -8.076 6.68e-16 ***
sexMale                                   0.1018145  0.0817531   1.245 0.212989    
age                                       0.0502889  0.0033333  15.087  < 2e-16 ***
race_ethnicityNon-Hispanic Asian          0.3100764  0.2069421   1.498 0.134036    
race_ethnicityNon-Hispanic Black          0.5166725  0.1654794   3.122 0.001795 ** 
race_ethnicityNon-Hispanic White         -0.0950453  0.1642887  -0.579 0.562909    
race_ethnicityOther Hispanic              0.1339837  0.1922304   0.697 0.485806    
race_ethnicityOther or Multi-Racial       0.1633541  0.2320126   0.704 0.481387    
education9-11th grade                    -0.3016727  0.1995761  -1.512 0.130644    
educationHigh school graduate            -0.4531578  0.1857923  -2.439 0.014726 *  
educationSome college                    -0.3444588  0.1845061  -1.867 0.061912 .  
educationCollege graduate                -0.5409684  0.1973912  -2.741 0.006133 ** 
marital_statusNever married               0.2610096  0.1214595   2.149 0.031639 *  
marital_statusWidowed/Divorced/Separated  0.0515478  0.0933308   0.552 0.580734    
prop_life_in_us                           0.4581333  0.2472798   1.853 0.063927 .  
num_drinks_per_yr(1,30]                  -0.0267692  0.0979566  -0.273 0.784641    
num_drinks_per_yr(30,300]                -0.0202755  0.1116934  -0.182 0.855953    
num_drinks_per_yr(300,5.5e+03]            0.3377602  0.1249693   2.703 0.006877 ** 
bp_histDiagnosed, on meds                -0.5106574  0.1512255  -3.377 0.000733 ***
bp_histNot diagnosed                     -1.3362291  0.1483811  -9.005  < 2e-16 ***
diagnosed_high_cholYes                   -0.1657506  0.0845748  -1.960 0.050018 .  
diagnosed_diabetesYes                    -0.0979090  0.1013885  -0.966 0.334204    
food_securityMarginal                     0.0565296  0.1209213   0.467 0.640148    
food_securityLow                         -0.1564641  0.1314336  -1.190 0.233873    
food_securityVery low                     0.0830001  0.1420022   0.584 0.558885    
insuranceYes                              0.0628119  0.1310889   0.479 0.631828    
healthcare_util_year1                    -0.4474173  0.1506056  -2.971 0.002970 ** 
healthcare_util_year2-3                  -0.5553645  0.1403108  -3.958 7.55e-05 ***
healthcare_util_year4-5                  -0.5025630  0.1555758  -3.230 0.001236 ** 
healthcare_util_year6+                   -0.6992901  0.1512854  -4.622 3.79e-06 ***
poverty_index                            -0.0339062  0.0315153  -1.076 0.281988    
min_sedentary_act                        -0.0001526  0.0001974  -0.773 0.439713    
smoking_statusFormer                      0.0778065  0.1239533   0.628 0.530195    
smoking_statusNever                       0.1319236  0.1153604   1.144 0.252799    
reported_bmi                              0.0171334  0.0056916   3.010 0.002610 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Assessing model performance

- In-sample RMSE: 0.537
- Out-of-sample RMSE: 0.5441

- Measured prevalence of high blood pressure: 22.79%
- Predicted prevalence of high blood pressure: 20.69%


Prevalence of High Blood Pressure
Female
Male
Percent (%)
Age
Data Type
Measured
Predicted