compiled Project

Seth Billiau

4/30/2020

EDA

Read in the data and make note of missing values:

```
data_raw = read.csv("data/chd_risk.csv")
summary(data_raw)
```

```
##       age                    education      cigsPerDay        totChol
##  Min.   :32.00   College or Higher : 473   Min.   : 0.000   Min.   :107.0
##  1st Qu.:42.00   High School or GED:1253   1st Qu.: 0.000   1st Qu.:206.0
##  Median :49.00   Some College      : 687   Median : 0.000   Median :234.0
##  Mean   :49.58   Some High School  :1720   Mean   : 9.003   Mean   :236.7
##  3rd Qu.:56.00   NA's              : 105   3rd Qu.:20.000   3rd Qu.:263.0
##  Max.   :70.00                             Max.   :70.000   Max.   :696.0
##                                            NA's   :29       NA's   :50
##      sysBP           diaBP             BMI           heartRate
##  Min.   : 83.5   Min.   : 48.00   Min.   :15.54   Min.   : 44.00
##  1st Qu.:117.0   1st Qu.: 75.00   1st Qu.:23.07   1st Qu.: 68.00
##  Median :128.0   Median : 82.00   Median :25.40   Median : 75.00
##  Mean   :132.4   Mean   : 82.89   Mean   :25.80   Mean   : 75.88
##  3rd Qu.:144.0   3rd Qu.: 89.88   3rd Qu.:28.04   3rd Qu.: 83.00
##  Max.   :295.0   Max.   :142.50   Max.   :56.80   Max.   :143.00
##                                   NA's   :19       NA's   :1
##     glucose           sex            smoker      OnBPMeds    PrevStroke
##  Min.   : 40.00   female:2419   Nonsmoker:2144   No  :4061   No :4213
##  1st Qu.: 71.00   male  :1819   Smoker   :2094   Yes : 124   Yes:  25
##  Median : 78.00                                  NA's:  53
##  Mean   : 81.97
##  3rd Qu.: 87.00
##  Max.   :394.00
##  NA's   :388
##  Hyp         Diab       CHD_Risk
##  No :2922   No :4129   No :3594
##  Yes:1316   Yes: 109   Yes: 644
##
##
##
##
##
```

Check for missingness

Count number of missing predictors in each variable:

```r
# Generate the number of missing values for each predictor
apply(is.na(data_raw), 2, sum)
```

```
##         age   education cigsPerDay     totChol       sysBP       diaBP         BMI
##           0         105          29          50           0           0          19
##   heartRate     glucose         sex      smoker    OnBPMeds PrevStroke         Hyp
##           1         388           0           0          53           0           0
##        Diab    CHD_Risk
##           0           0
```

```r
missing_preds = c("education", "cigsPerDay", "totChol", "BMI",
                  "heartRate", "glucose", "OnBPMeds")
```

Visualizing the Response

```r
# Labels for ticks
N_label = paste("No (", round(mean(data_raw$CHD_Risk == "No"), 2)*100, "%)", sep="")
Y_label = paste("Yes (", round(mean(data_raw$CHD_Risk == "Yes"), 2)*100, "%)", sep="")

# Plot bar chart
p <- ggplot(data=data_raw, aes(x=as.factor(CHD_Risk))) +
  geom_histogram(stat="count",fill="red",
                 aes(y=..count../sum(..count..)),
                 alpha = 0.5) +
    labs(title="Frequency of CHD_Risk") +
    xlab("CHD_Risk") +
    ylab("Frequency") +
    scale_x_discrete(breaks=c("No","Yes"), labels=c(N_label,Y_label)) +
    theme_bw()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```r
p
```

Frequency of CHD_Risk
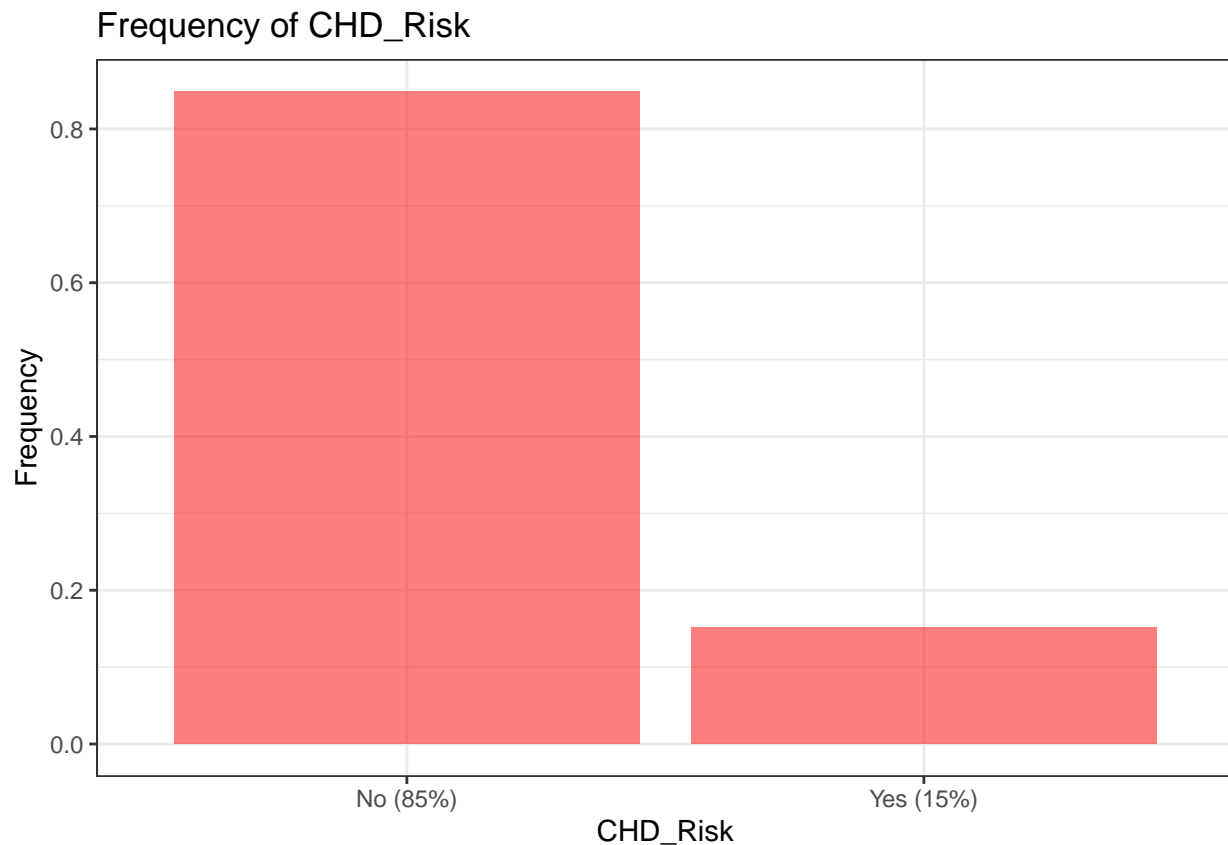
Visualizing the Predictors

Visualize distribution of quantitative predictors conditional on the CHD outcome:

```r
quant_preds = c("age", "cigsPerDay", "totChol", "sysBP",
                "diaBP", "BMI", "heartRate", "glucose")

make_cond_hist = function(varname) {
  p1 = ggplot(data_raw, aes_string(x=varname)) +
    geom_histogram(aes(y = ..density..),
                   fill = "red", alpha = 0.5) +
    labs(title=paste(varname, "given CHD_Risk")) +
    xlab(varname) +
    ylab("Density") +
    facet_grid(. ~ CHD_Risk) +
    theme_bw()
  return(p1)
}

graphs = lapply(quant_preds, make_cond_hist)
figure1 = ggarrange(graphs[[1]], graphs[[2]], graphs[[3]], graphs[[4]],
        graphs[[5]], graphs[[6]], graphs[[7]], graphs[[8]],
        ncol = 2, nrow = 4)
annotate_figure(figure1,
            top = text_grob("Visualizing Quantitative Predictors given CHD_Risk (prevalence = 0.152)
)
```

Visualizing Quantitative Predictors given CHD_Risk (prevalence = 0.152)
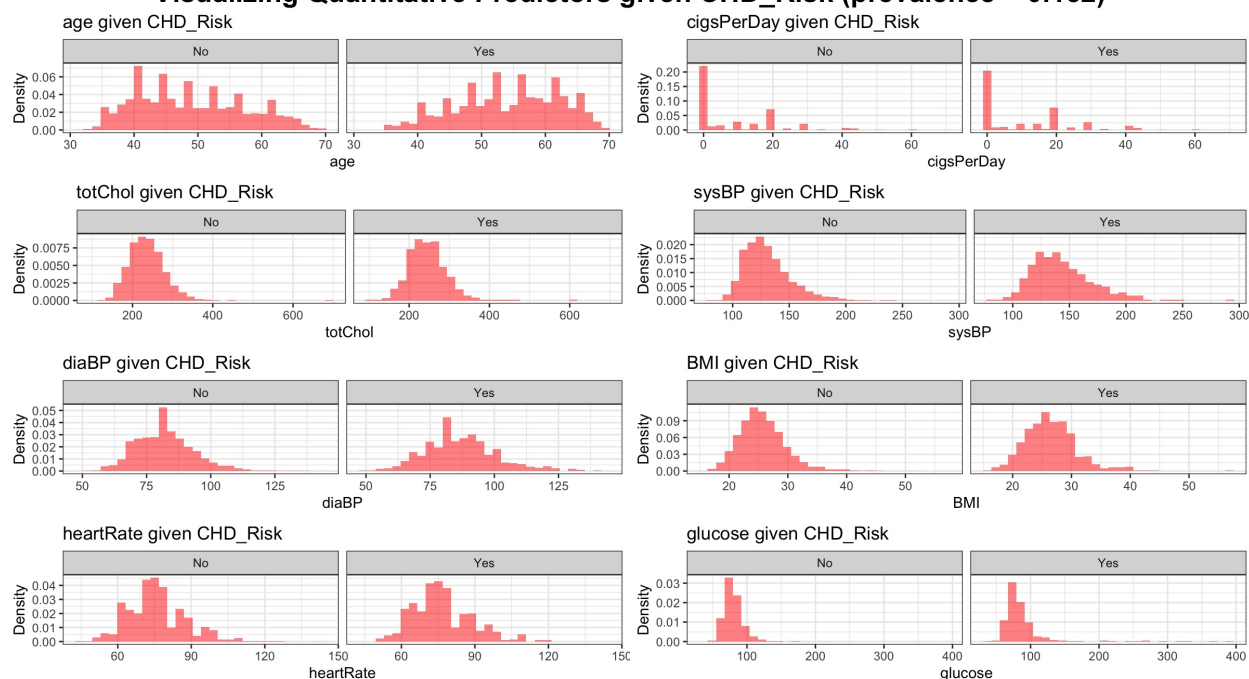

Figure 1: Quantitative EDA

Visualizing the Qualitative predictors by showing their distributions conditional on the outcome:

```r
# Address Categorical predictors
cat_preds = c("education", "sex", "smoker", "OnBPMeds",
              "PrevStroke", "Hyp", "Diab")
get_cond_prob_table = function(TABLE) {
  col1 = TABLE[,1] / sum(TABLE[,1])
  col2 = TABLE[,2] / sum(TABLE[,2])
  return(cbind(No=col1, Yes=col2))
}
tab_education = get_cond_prob_table(table(data_raw$education, data_raw$CHD_Risk))
tab_sex = get_cond_prob_table(table(data_raw$sex, data_raw$CHD_Risk))
tab_smoker = get_cond_prob_table(table(data_raw$smoker, data_raw$CHD_Risk))
tab_OnBPMeds = get_cond_prob_table(table(data_raw$OnBPMeds, data_raw$CHD_Risk))
tab_PrevStroke = get_cond_prob_table(table(data_raw$PrevStroke, data_raw$CHD_Risk))
tab_Hyp = get_cond_prob_table(table(data_raw$Hyp, data_raw$CHD_Risk))
tab_Diab = get_cond_prob_table(table(data_raw$Diab, data_raw$CHD_Risk))

tab_prob_Yes = rbind(tab_education, tab_sex, tab_smoker,
                     tab_OnBPMeds, tab_PrevStroke, tab_Hyp,
                     tab_Diab)
round(tab_prob_Yes,3)
```

```
##                      No    Yes
## College or Higher  0.115 0.111
## High School or GED 0.316 0.234
## Some College       0.171 0.140
## Some High School   0.399 0.514
## female             0.589 0.467
```

```
## male             0.411 0.533
## Nonsmoker         0.510 0.483
## Smoker            0.490 0.517
## No                0.977 0.935
## Yes               0.023 0.065
## No                0.996 0.983
## Yes               0.004 0.017
## No                0.724 0.495
## Yes               0.276 0.505
## No                0.981 0.938
## Yes               0.019 0.062
```

Check for collinearity with GVIF.

```
# Check for multicollinearity
mod.vif.lm <- lm(as.numeric(CHD_Risk) ~ ., data=data_raw)
vif(mod.vif.lm)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## age          1.397737  1        1.182259
## education    1.124453  3        1.019742
## cigsPerDay   2.732416  1        1.653002
## totChol      1.116842  1        1.056808
## sysBP        3.767158  1        1.940917
## diaBP        3.000260  1        1.732126
## BMI          1.246685  1        1.116550
## heartRate    1.095015  1        1.046429
## glucose      1.638312  1        1.279966
## sex          1.223718  1        1.106218
## smoker       2.585357  1        1.607904
## OnBPMeds     1.111774  1        1.054407
## PrevStroke   1.017647  1        1.008785
## Hyp          2.051447  1        1.432287
## Diab         1.616622  1        1.271465
```

Because all values in the last column are less than $3.1623 = \sqrt{10}$, there is not significant/strong evidence of multicollinearity.