Stat 6021: Guided Question Set 11

Tom Lever

11/22/22

The Western Collaborative Study Group (WCGS) is one of the earliest studies regarding heart disease. Data were collected from 3,154 males aged 39 to 59 in the San-Francisco area in 1960. They all did not have coronary heart disease at the beginning of the study. The data set comes from the `faraway` package and is called `wcgs`. We will focus on predicting the likelihood of developing coronary heart disease based on the following predictors:

- `age`: age in years
- `sdp`: systolic blood pressure in $mm\ Hg$
- `dbp`: diastolic blood pressure in $mm\ Hg$
- `cigs`: number of cigarettes smoked per day
- `dibep`: behavior type, labeled A and B for aggressive and passive, respectively

The response variable is `chd`, whether the person developed coronary heart disease during annual follow ups in the study. Read the data in. We will also randomly split the data into two: half the data will be the training data set, and the remaining half will be the test data set. We will explore the training-test split in more detail in the next module. For this exercise, perform all analysis on the training data. The code below will randomly split the data into two halves.

```
library(faraway)
data_set <- wcgs
set.seed(6021)
number_of_observations <- nrow(data_set)
indices_of_observations <- sample.int(number_of_observations, floor(0.5 * number_of_observations), repla
training_data_set <- data_set[indices_of_observations, ]
testing_data_set <- data_set[-indices_of_observations, ]
head(training_data_set, n = 3)
```

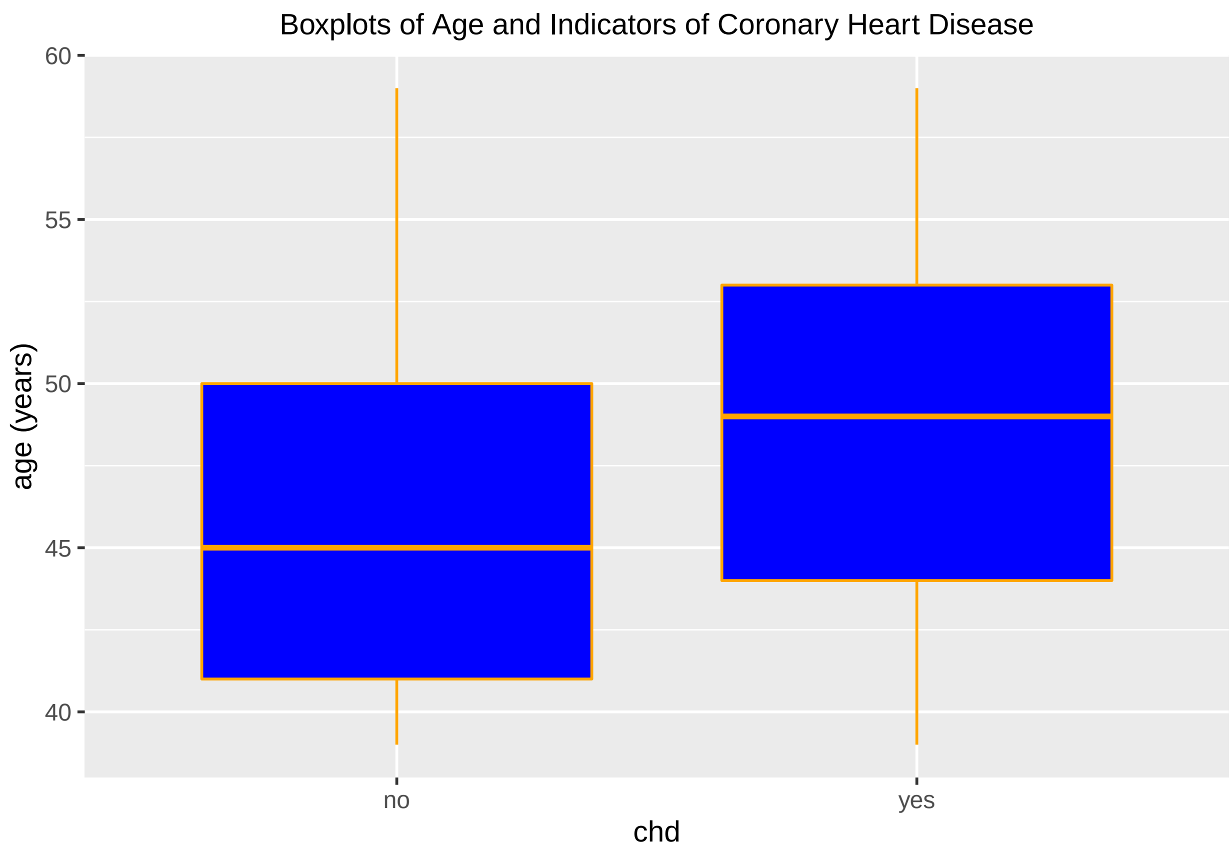
```
##      age height weight sdp dbp chol behave cigs dibep chd typechd timechd
## 225   53     71    142 150  78  218     A2   40     A  no    none    3127
## 2905  46     71    180 110  80  260     B3    0     B  no    none    2887
## 1644  39     71    180 114  78  234     B3    0     B  no    none    2985
##        arcus
## 225  present
## 2905  absent
## 1644 present
```

1. Before fitting a model, create some data visualizations to explore the relationship between these predictors and whether a middle-aged male develops coronary heart disease.

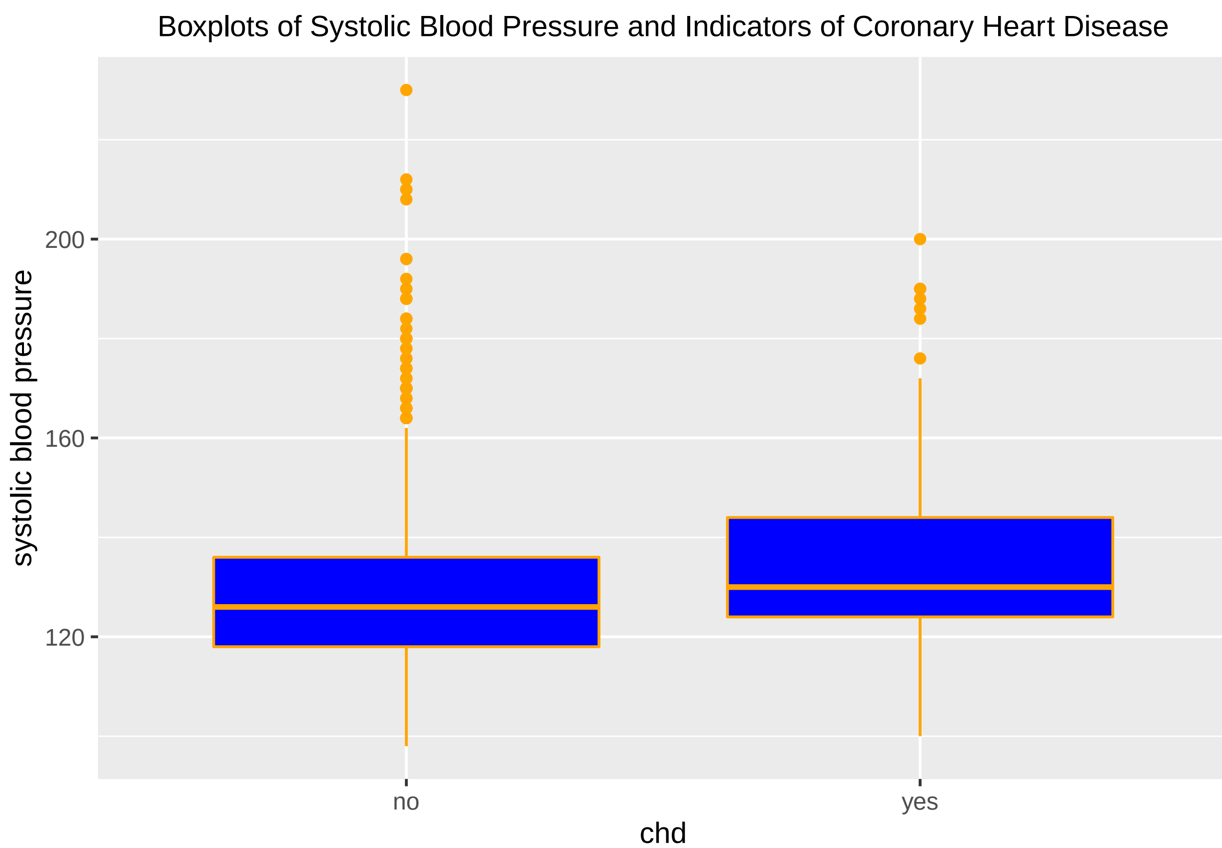
Boxplots of age (years), systolic blood pressure, diastolic blood pressure, and smoking rate (cigarettes per day) are presented below.

```
library(ggplot2)
ggplot(data_set, aes(x = chd, y = age)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
```

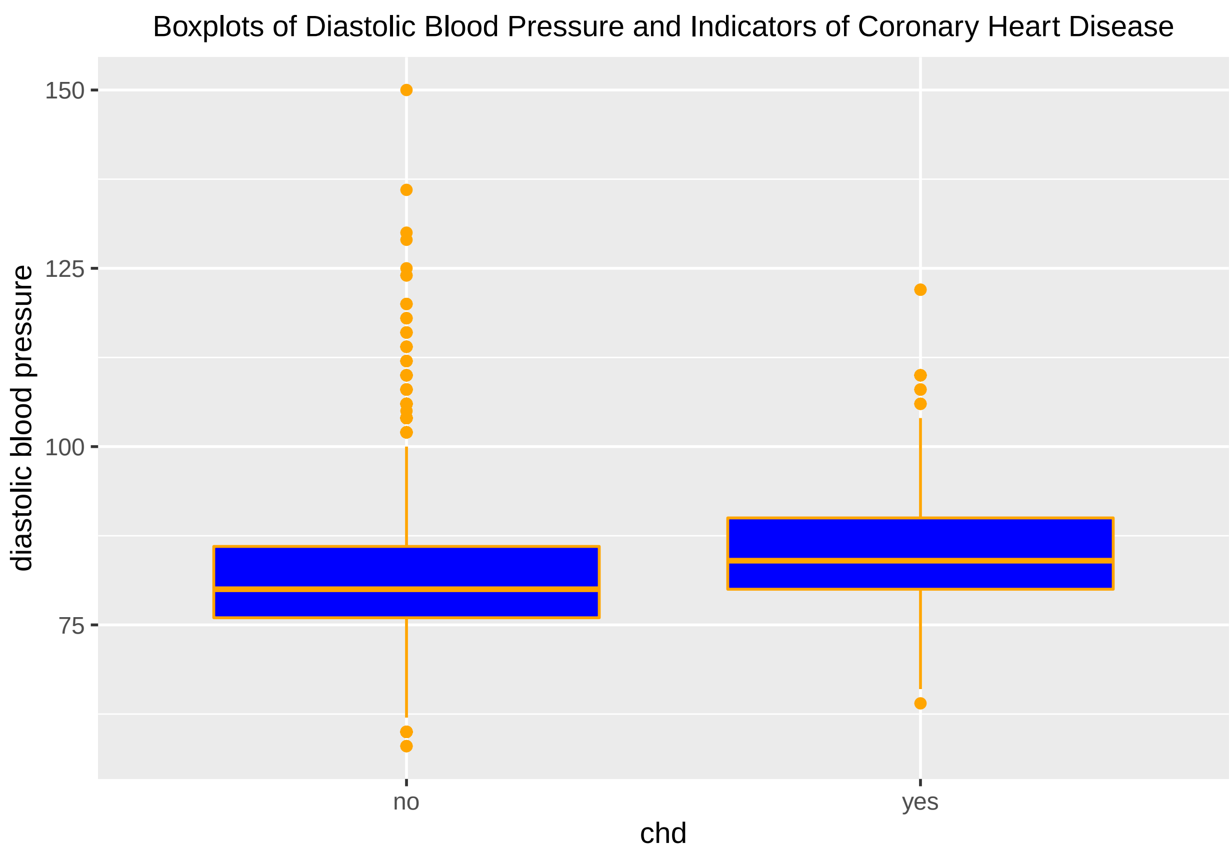
```r
    labs(
        y = "age (years)",
        title = "Boxplots of Age and Indicators of Coronary Heart Disease"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
    )
```



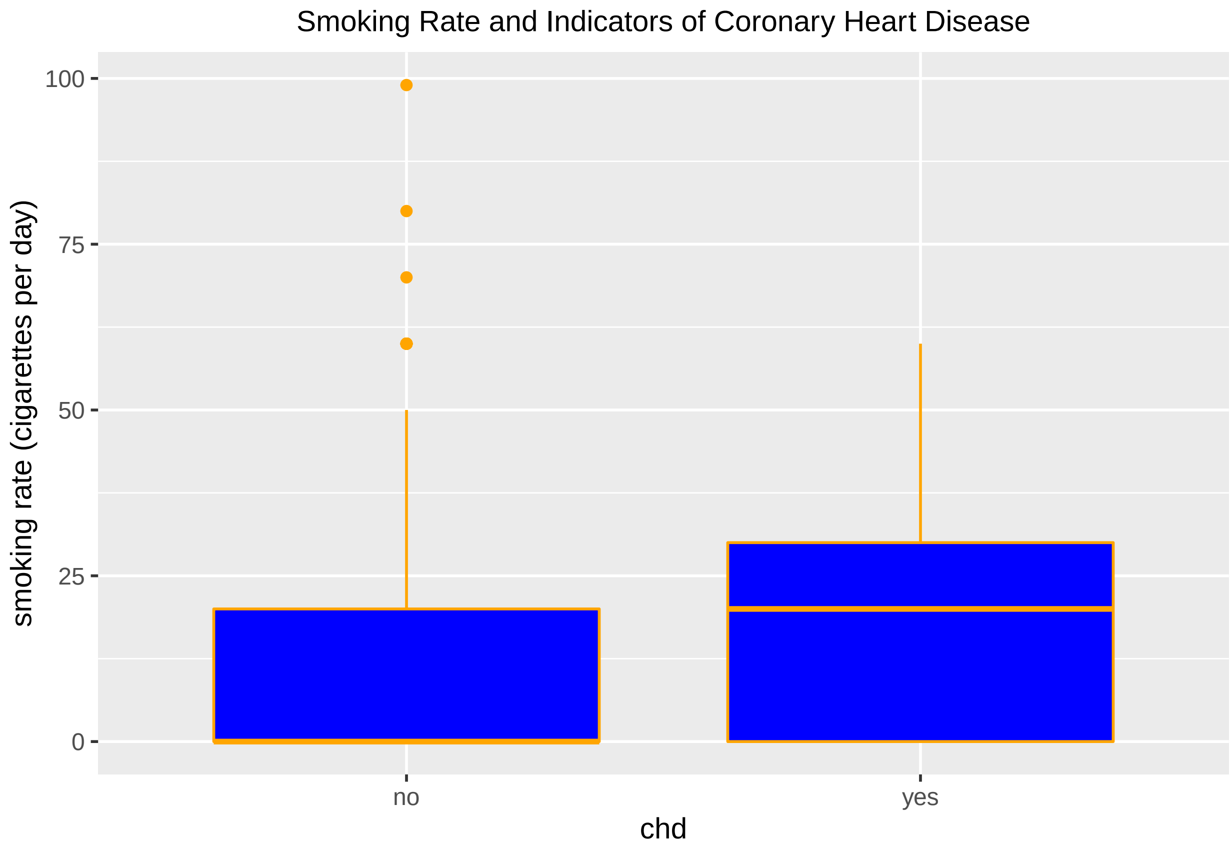

```r
ggplot(data_set, aes(x = chd, y = sdp)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        y = "systolic blood pressure",
        title = "Boxplots of Systolic Blood Pressure and Indicators of Coronary Heart Disease"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
    )
```




```
ggplot(data_set, aes(x = chd, y = dbp)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        y = "diastolic blood pressure",
        title = "Boxplots of Diastolic Blood Pressure and Indicators of Coronary Heart Disease"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
    )
```




```
ggplot(data_set, aes(x = chd, y = cigs)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        y = "smoking rate (cigarettes per day)",
        title = "Smoking Rate and Indicators of Coronary Heart Disease"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
    )
```



People who developed heart disease tend to be older, have higher blood pressures, and smoke more cigarettes per day. There is a high variability in a lot of these predictors for each indicator of coronary heart disease.

The number of cigarettes smoked per day appears to be the biggest factor in whether one develops coronary heart disease, as their distributions are most different. Among those with no heart disease, 50 percent of them did not smoke. Among those with heart disease, 25 percent of them did not smoke.

There is a lot of overlap in the boxplots for the blood-pressure variables, so blood pressure may not differentiate between those who develop heart disease from those who did not.

Density plots for age (years), systolic blood pressure, diastolic blood pressure, and smoking rate (cigarettes per day) are presented below. The density plot of age for those without heart disease is right skewed; a higher proportion of those without heart disease are younger than 45. The distributions of age for those with heart disease is more symmetric, with a peak around 50. Age could be a good predictor for whether someone develops heart disease.

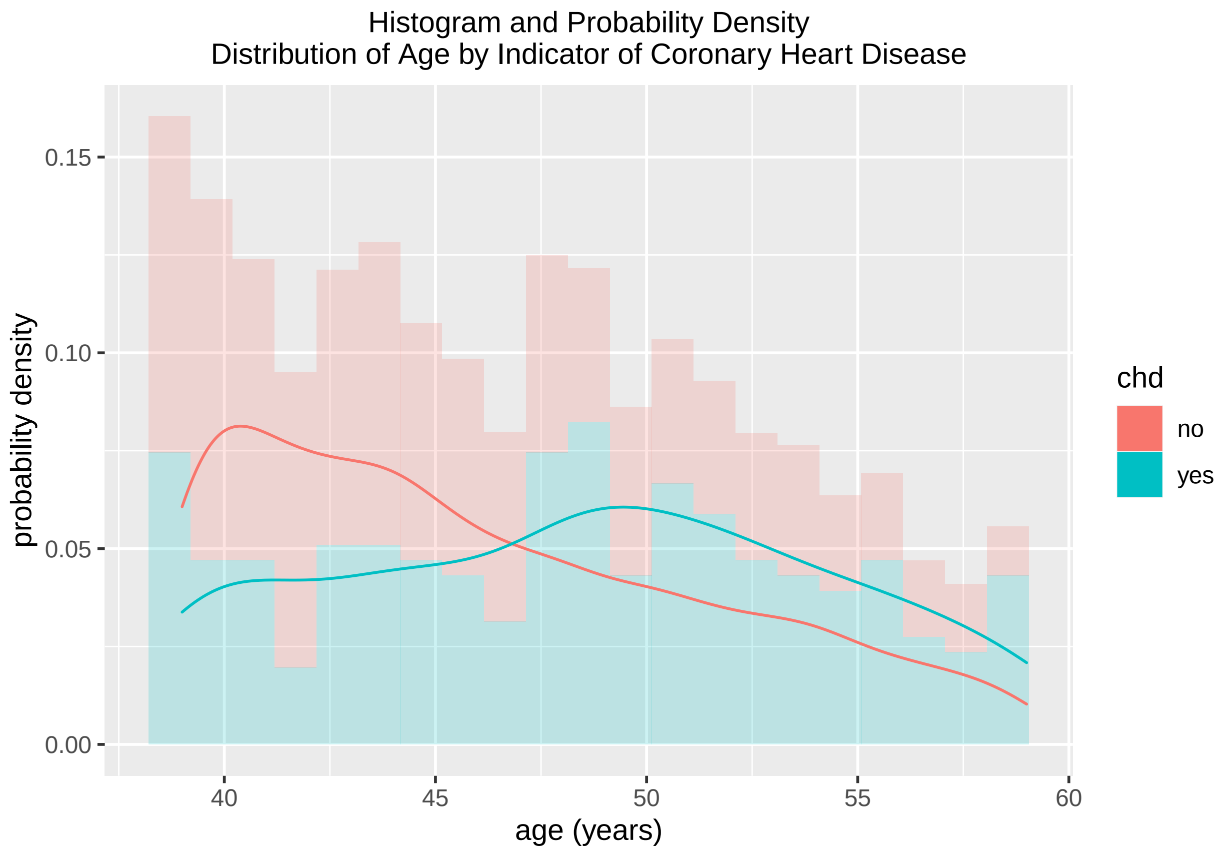
The density plots of blood pressure are similar for those with heart disease and those without heart disease. The blood pressure variables are less likely to be good predictors for whether someone develops heart disease.

A much larger proportion of those who did not develop heart disease do not smoke, compared to those who did develop heart disease.

```
ggplot() +
    geom_histogram(data = data_set, aes(x = data_set$age, y = ..density.., fill = chd), , alpha = 0.2, b
    geom_density(data = data_set, aes(x = data_set$age, color = chd)) +
    labs(
        x = "age (years)",
        y = "probability density",
        title = "Histogram and Probability Density\nDistribution of Age by Indicator of Coronary Heart D
```

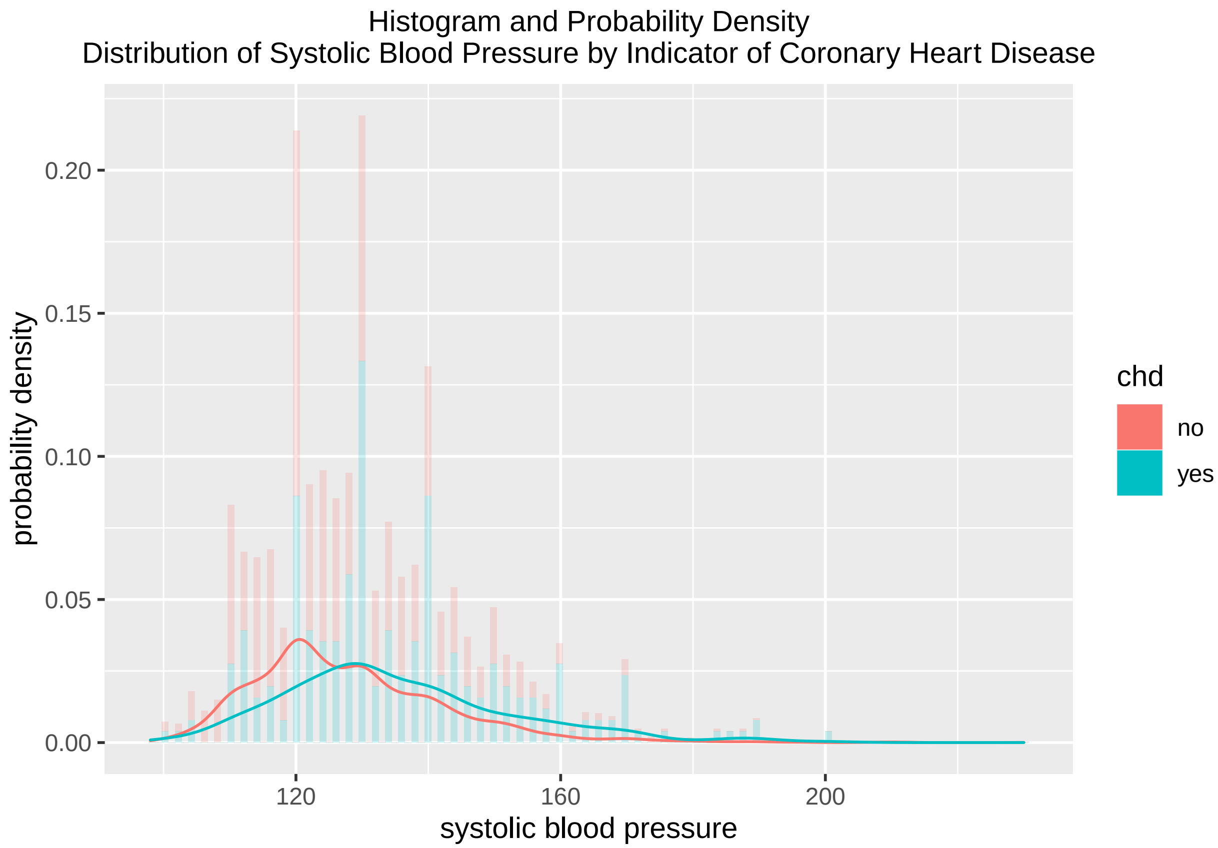
```r
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

```
## Warning: Use of `data_set$age` is discouraged. Use `age` instead.
## Use of `data_set$age` is discouraged. Use `age` instead.
```



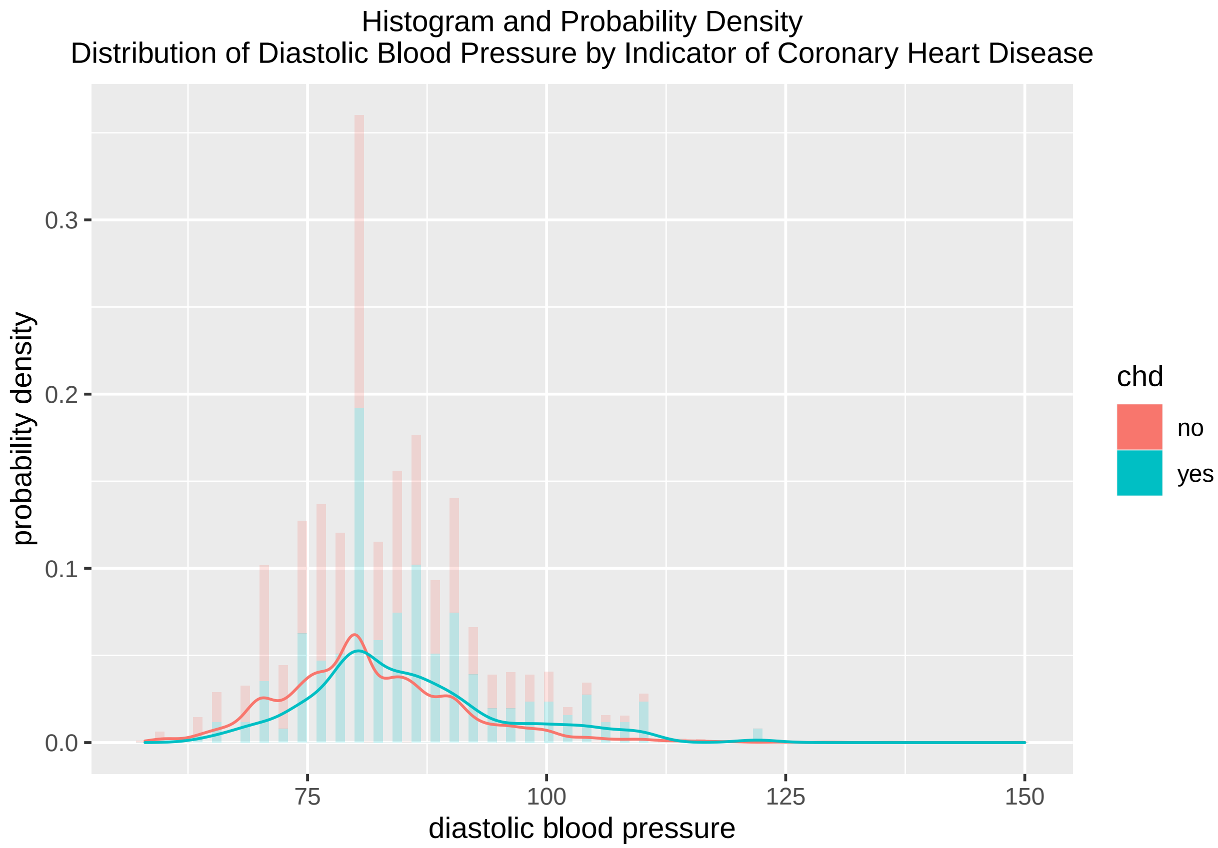

```r
ggplot() +
    geom_histogram(data = data_set, aes(x = data_set$sdp, y = ..density.., fill = chd), , alpha = 0.2, b
    geom_density(data = data_set, aes(x = data_set$sdp, color = chd)) +
    labs(
        x = "systolic blood pressure",
        y = "probability density",
        title = "Histogram and Probability Density\nDistribution of Systolic Blood Pressure by Indicato
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

```
## Warning: Use of `data_set$sdp` is discouraged. Use `sdp` instead.

## Warning: Use of `data_set$sdp` is discouraged. Use `sdp` instead.
```



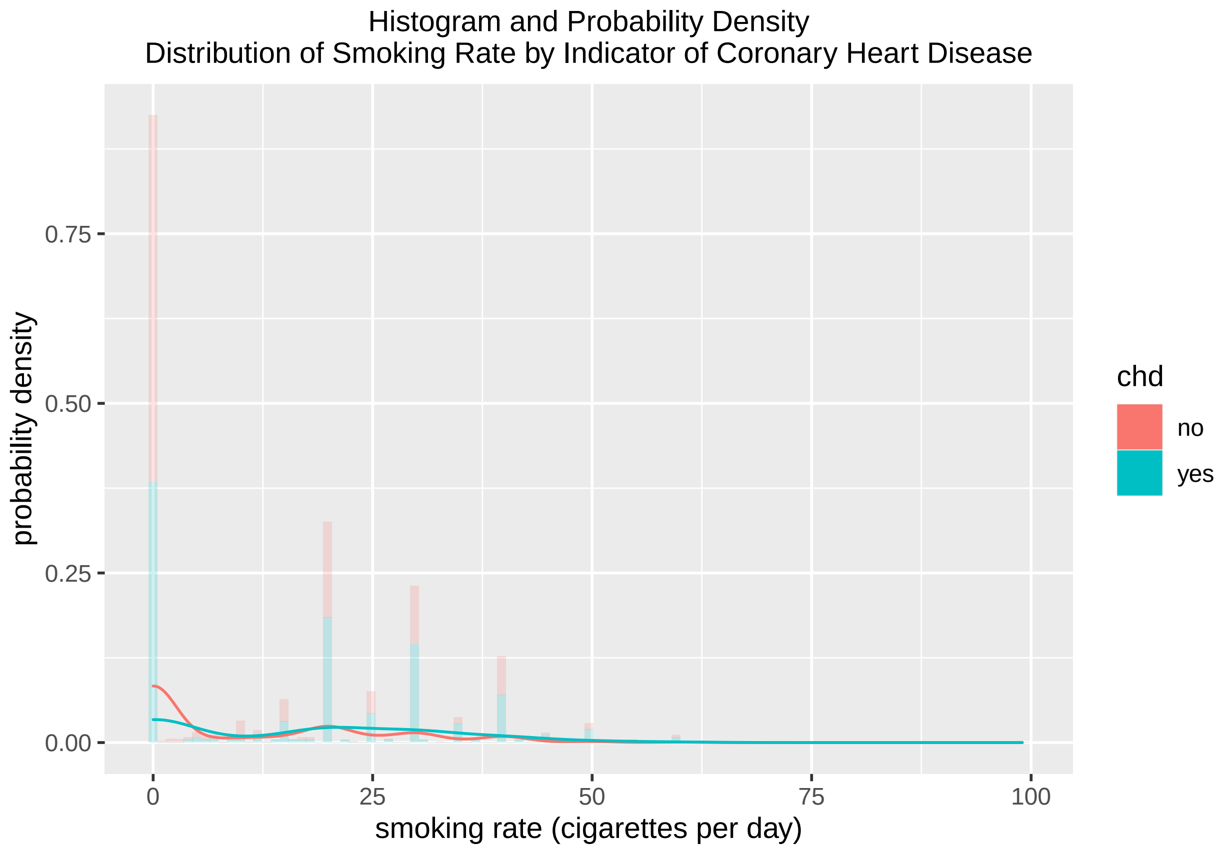

```
ggplot() +
    geom_histogram(data = data_set, aes(x = data_set$dbp, y = ..density.., fill = chd), , alpha = 0.2, b
    geom_density(data = data_set, aes(x = data_set$dbp, color = chd)) +
    labs(
        x = "diastolic blood pressure",
        y = "probability density",
        title = "Histogram and Probability Density\nDistribution of Diastolic Blood Pressure by Indicat
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

```
## Warning: Use of `data_set$dbp` is discouraged. Use `dbp` instead.

## Warning: Use of `data_set$dbp` is discouraged. Use `dbp` instead.
```



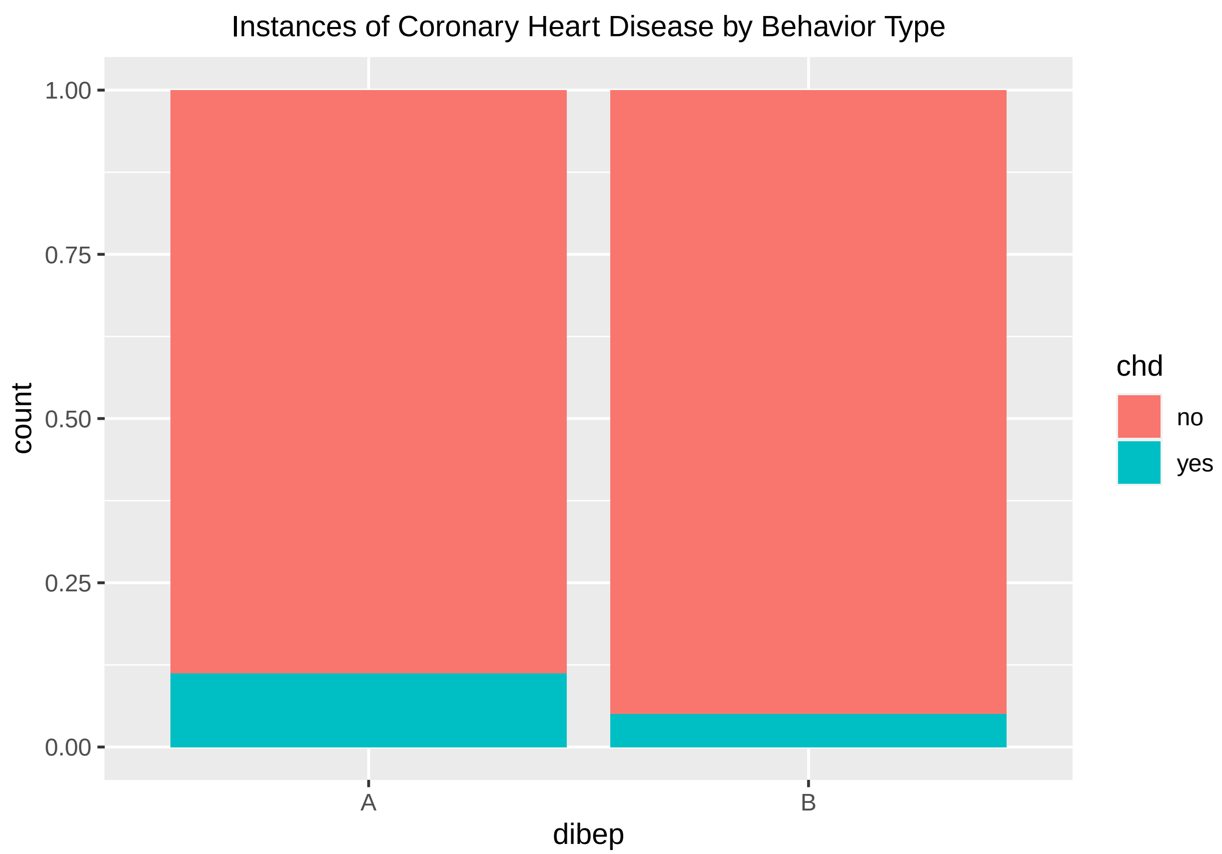

```
ggplot() +
    geom_histogram(data = data_set, aes(x = data_set$cigs, y = ..density.., fill = chd), , alpha = 0.2,
    geom_density(data = data_set, aes(x = data_set$cigs, color = chd)) +
    labs(
        x = "smoking rate (cigarettes per day)",
        y = "probability density",
        title = "Histogram and Probability Density\nDistribution of Smoking Rate by Indicator of Corona
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

```
## Warning: Use of `data_set$cigs` is discouraged. Use `cigs` instead.

## Warning: Use of `data_set$cigs` is discouraged. Use `cigs` instead.
```



A bar chart comparing the rate of developing heart disease by behavior type is shown below.

```
library(dplyr)
ggplot(data_set, aes(x = dibep, fill = chd)) +
    geom_bar(position = "fill") +
    labs(title = "Instances of Coronary Heart Disease by Behavior Type") +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```




```
two_way_table <- table(training_data_set$dibep, training_data_set$chd)
two_way_table
```

```
##
##      no yes
##   A 726  86
##   B 722  43
```

```
prop.table(two_way_table, 1)
```

```
##
##             no        yes
##   A 0.89408867 0.10591133
##   B 0.94379085 0.05620915
```

The rate of developing heart disease is low for all behavior types, but is higher for middle-aged males with passive behavior type than for males with aggressive behavior type. The two way table confirms this. The rate is about 5.6 percent for males with aggressive behavior type, and is about 10.6 percent for males with passive behavior type.

2. Use R to fit the logistic regression model using all the predictors listed above, and write the estimated logistic regression equation.

```
generalized_linear_model <- glm(chd ~ age + sdp + dbp + cigs + dibep, family = "binomial", data = traini
generalized_linear_model
```

```
##
## Call:  glm(formula = chd ~ age + sdp + dbp + cigs + dibep, family = "binomial",
##     data = training_data_set)
```

```
##
## Coefficients:
## (Intercept)          age          sdp          dbp         cigs       dibepB
##    -8.30877      0.06021      0.01512      0.01203      0.02137     -0.52691
##
## Degrees of Freedom: 1576 Total (i.e. Null);  1571 Residual
## Null Deviance:       893
## Residual Deviance: 837.5     AIC: 849.5
```

The logistic regression equation is

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -8.836 + 0.06\ age + 0.015\ sdp + 0.012\ dbp + 0.021\ cigs + 0.527\ I_1$$

where $I_1 = 1$ for behavior type B, and 0 for behavior type A.

3. Interpret the estimated coefficient for `cigs` in context.

The regression coefficient for `cigs` is 0.021.

For an additional cigarette smoked per day on average, the estimated log odds of developing coronary heart disease increases by 0.021, while controlling for the other predictors (age, systolic blood pressure, diastolic blood pressure, and behavior type).

For an additional cigarette smoked per day on average, the estimated odds of developing coronary heart disease gets multiplied by a factor of $exp(0.021) = 1.021$, while controlling for the other predictors.

4. Interpret the estimated coefficient for `dibep` in context.

The regression coefficient for `dibep` is 0.527.

The estimated log odds of developing heart disease for males with type B (passive) behavior is 0.527 higher than for males with type A (aggressive) behavior, while controlling for the other predictors.

The estimated odds of developing heart disease for males with type B behavior is $exp(0.69) = 1.694$ times the odds for males with type A behaviors, while controlling for the other predictors.

5. What are the estimated odds of developing heart disease for an adult male who is 45 years old, has a systolic blood pressure of 110 $mm\ Hg$, has diastolic blood pressure of 70 $mm\ Hg$, does not smoke, and has type B (passive) personality? What is this person's corresponding probability of developing heart disease?

```r
log_odds <- predict(generalized_linear_model, data.frame(age = 45, sdp = 110, dbp = 70, cigs = 0, dibep
odds <- exp(log_odds)
odds
```

```
##          1
## 0.02675027
```

```r
probability <- odds / (1 + odds)
probability
```

```
##          1
## 0.02605333
```

The estimated odds of developing heart disease is 0.045. The corresponding probability is 0.043.

6. Carry out the relevant hypothesis test to check if this logistic regression model with five predictors is useful in estimated the odds of heart disease. Clearly state the null and alternate hypotheses, test statistic, and conclusion in context.

We test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ that the logistic regression coefficients are all 0. The alternate hypothesis is $H_1 : \beta_k \neq 0$ for at least one index of a regression coefficient k.

The test statistic is $D_{dropped} = D_0 - D_{full}$. The deviance of the dropped predictors follows a χ^2 distribution with degrees of freedom equal to the number of predictors dropped.

```
D_0 <- generalized_linear_model$null.deviance
D_full <- generalized_linear_model$deviance
D_full
```

```
## [1] 837.5471
```

```
D_dropped <- D_0 - D_full
D_dropped
```

```
## [1] 55.49501
```

```
number_of_predictors_dropped <- 5
pchisq(D_dropped, number_of_predictors_dropped, lower.tail = FALSE)
```

```
## [1] 1.032455e-10
```

The associated p value is above. We reject the null hypothesis. The data support the claim that our model is useful compared to the intercept-only model.

7. Suppose a coworker of yours suggests fitting a logistic regression model without the two blood pressure variables. Carry out the relevant hypothesis test to check if this model without the blood pressure variables should be chosen over the previous model with all five predictors.

We test the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ that the logistic regression coefficients are all 0. The alternate hypothesis is $H_1 : \beta_k \neq 0$ for at least one index of a regression coefficient k.

The test statistic is $D_{dropped} = D_0 - D_{full}$. The deviance of the dropped predictors follows a χ^2 distribution with degrees of freedom equal to the number of predictors dropped.

```
reduced_generalized_linear_model <- glm(chd ~ age + cigs + dibep, family = "binomial", data = training_d
D_reduced <- reduced_generalized_linear_model$deviance
D_reduced
```

```
## [1] 851.253
```

```
D_dropped <- D_reduced - D_full
D_dropped
```

```
## [1] 13.70587
```

```
number_of_predictors_dropped <- 2
pchisq(D_dropped, number_of_predictors_dropped, lower.tail = FALSE)
```

```
## [1] 0.00105635
```

The associated p value is above. We reject the null hypothesis. The data support going with the full model, because we reject the null hypothesis that the regression coefficients for the blood pressure predictors are 0. We do not drop both blood pressure predictors.

8. Biased on the Wald test, is diastolic blood pressure a significant predictor of heart disease, when the other predictors are already in the model?

```
summary(generalized_linear_model)
```

```
##
## Call:
## glm(formula = chd ~ age + sdp + dbp + cigs + dibep, family = "binomial",
##     data = training_data_set)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1764  -0.4505  -0.3480  -0.2712   2.7006
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.308765   1.080141  -7.692 1.45e-14 ***
## age          0.060212   0.016604   3.626 0.000287 ***
## sdp          0.015119   0.008805   1.717 0.085950 .
## dbp          0.012026   0.014345   0.838 0.401818
## cigs         0.021366   0.006095   3.506 0.000456 ***
## dibepB      -0.526914   0.198429  -2.655 0.007921 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 837.55  on 1571  degrees of freedom
## AIC: 849.55
##
## Number of Fisher Scoring iterations: 5
```

The Wald statistic for testing the significance of β_3 is $Z = 0.838$ with a large p value. So we can drop diastolic blood pressure from the logistic regression model, while leaving the other predictors in the model.

9. Based on all the analysis performed, which of these predictors would you use in your logistic regression model?

We only remove diastolic blood pressure as a predictor from the logistic regression model, and the keep in the other predictors (age, systolic blood pressure, smoking rate, and behavior type).