Guided Question Set 11 Solutions

```
library(faraway)
library(tidyverse)
library(gridExtra)
Data<-wcgs
set.seed(6021) ##for reproducibility
sample<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
train<-Data[sample, ] ##training data frame
test<-Data[-sample, ] ##test data frame
```

1)

The boxplots of the 4 quantitative variables across CHD status are shown below.

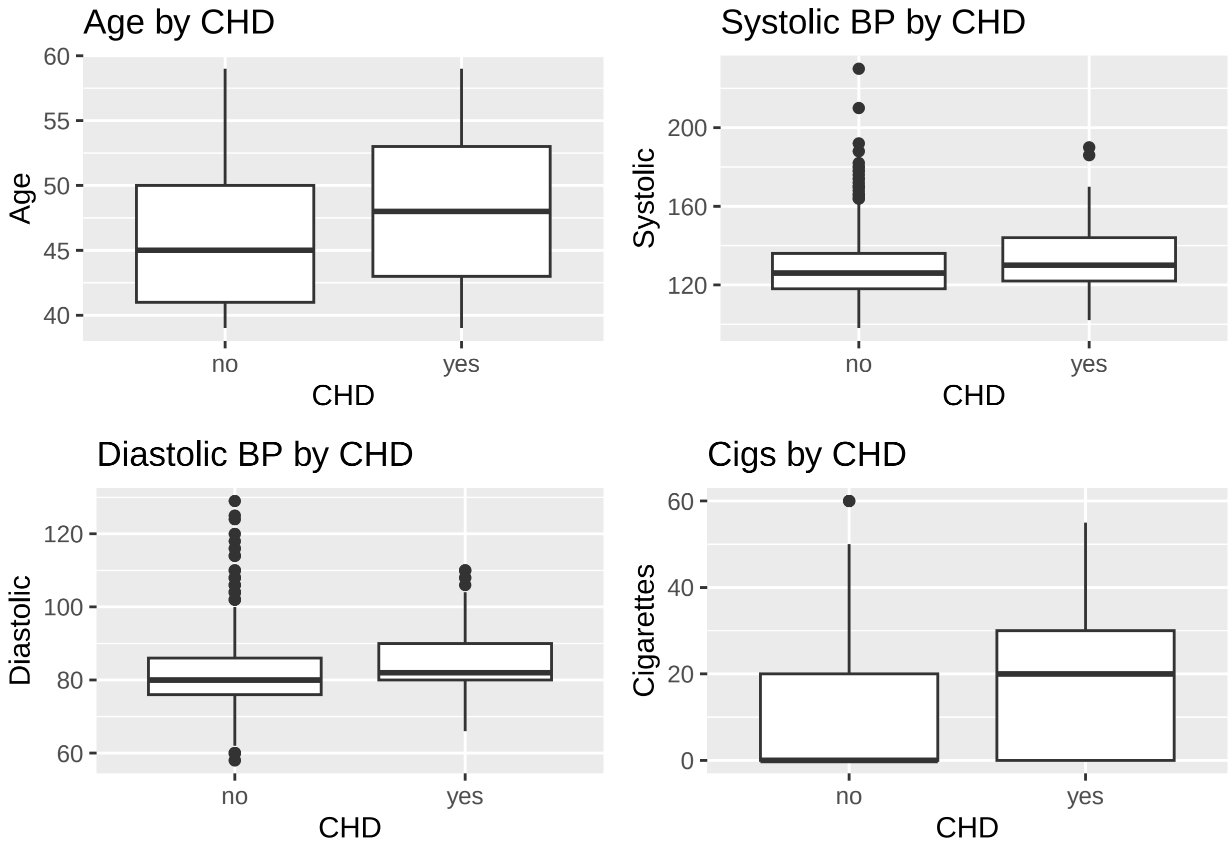
```
bp1<-ggplot(train, aes(x=chd, y=age))+
  geom_boxplot()+
  labs(x="CHD", y="Age", title="Age by CHD")

bp2<-ggplot(train, aes(x=chd, y=sdp))+
  geom_boxplot()+
  labs(x="CHD", y="Systolic", title="Systolic BP by CHD")

bp3<-ggplot(train, aes(x=chd, y=dbp))+
  geom_boxplot()+
  labs(x="CHD", y="Diastolic", title="Diastolic BP by CHD")

bp4<-ggplot(train, aes(x=chd, y=cigs))+
  geom_boxplot()+
  labs(x="CHD", y="Cigarettes", title="Cigs by CHD")

##produce the 4 boxplots in a 2 by 2 matrix
grid.arrange(bp1, bp2, bp3, bp4, ncol = 2, nrow = 2)
```



People who developed heart disease tend to be older, have higher blood pressures, as well as smoke more cigarettes. There is high variability in a lot of these predictors for each group (those without heart disease and those with heart disease).

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

The density plots of the 4 quantitative variables across CHD status are shown below.

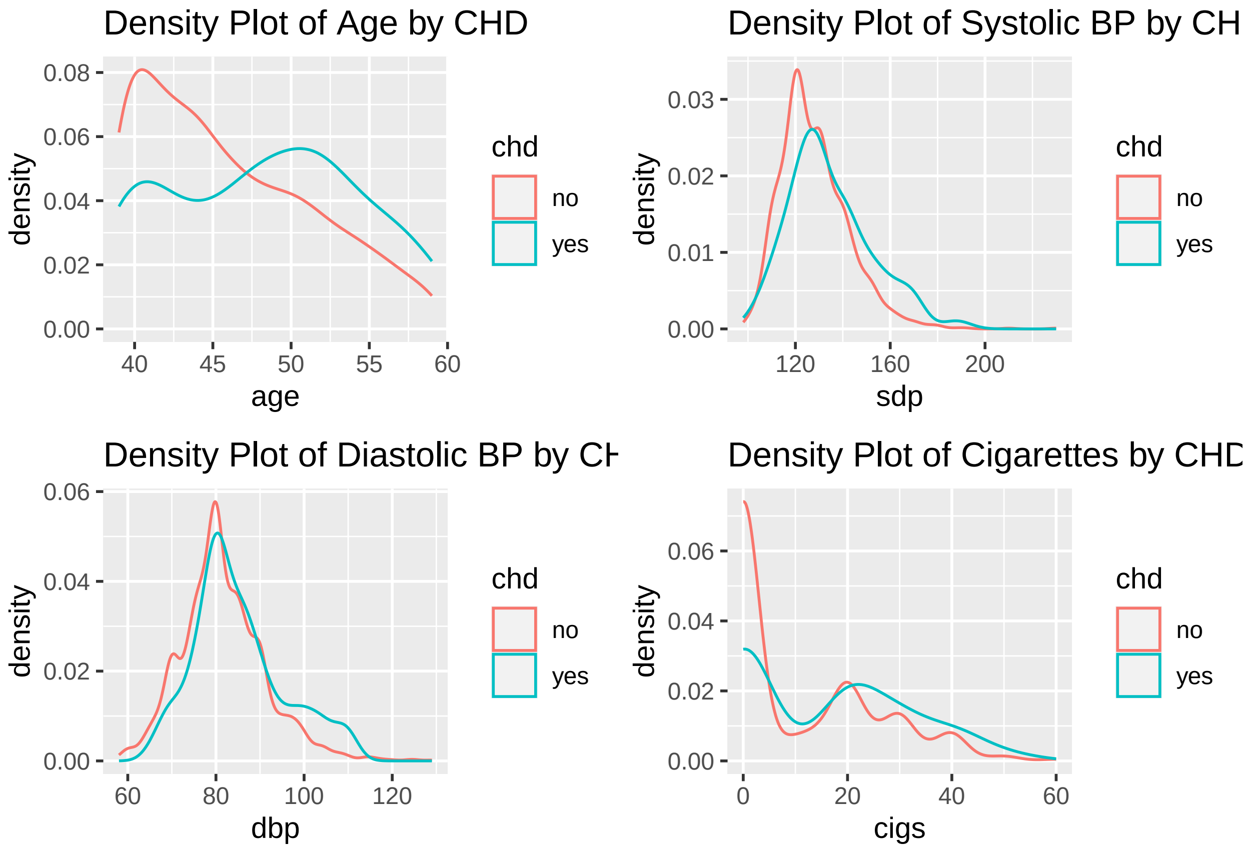
```r
dp1<-ggplot(train,aes(x=age, color=chd))+
  geom_density()+
  labs(title="Density Plot of Age by CHD")

dp2<-ggplot(train,aes(x=sdp, color=chd))+
  geom_density()+
  labs(title="Density Plot of Systolic BP by CHD")

dp3<-ggplot(train,aes(x=dbp, color=chd))+
  geom_density()+
  labs(title="Density Plot of Diastolic BP by CHD")

dp4<-ggplot(train,aes(x=cigs, color=chd))+
  geom_density()+
  labs(title="Density Plot of Cigarettes by CHD")

##produce the 4 density plots in a 2 by 2 matrix
grid.arrange(dp1, dp2, dp3, dp4, ncol = 2, nrow = 2)
```



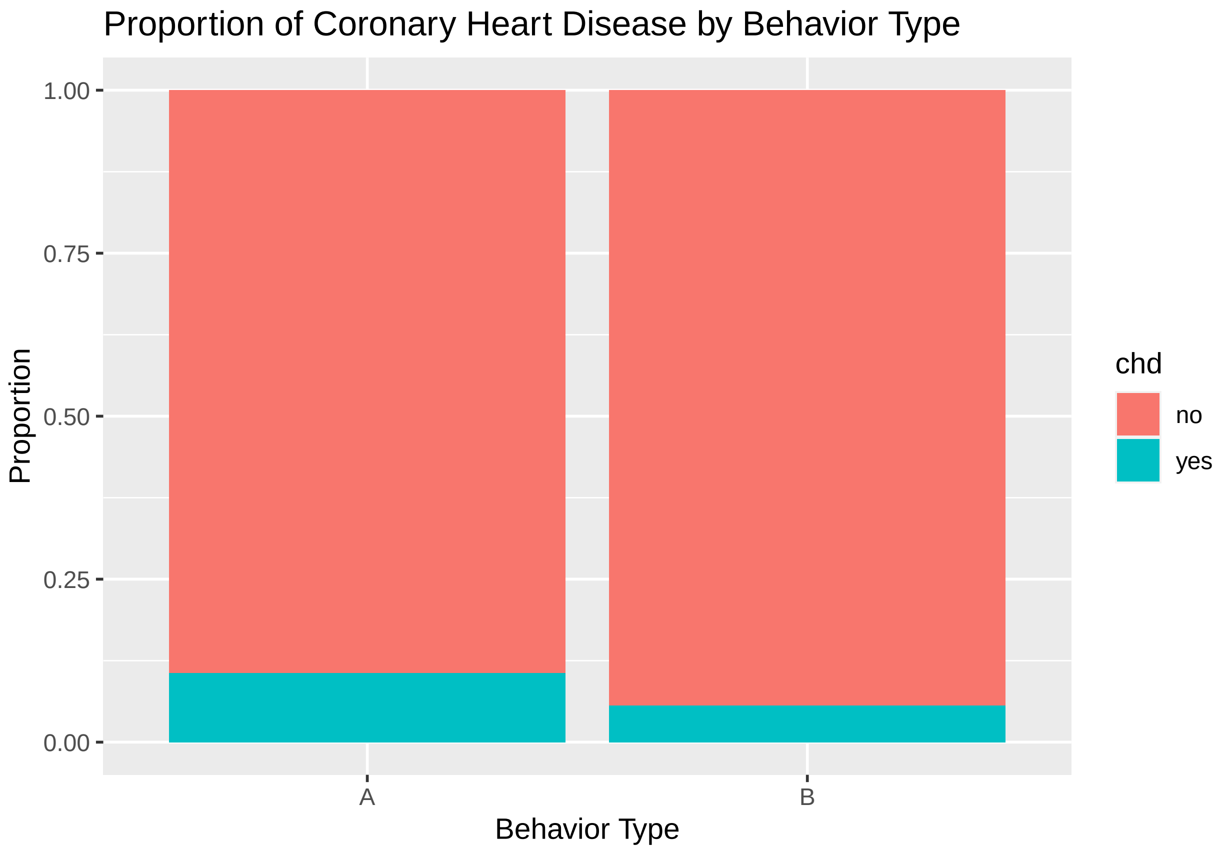
The density plot of age for those without heart disease is right skewed; a higher proportion of those without heart disease are younger (below 45). The distribution of age for those with heart disease tend is a more symmetric, with a peak around 50. Age could be a good predictor for whether someone develops heart disease.

The density plots of the blood pressure variables are similar for those with heart disease and for 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.

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

```r
ggplot(train, aes(x=dibep, fill=chd))+
  geom_bar(position = "fill")+
  labs(x="Behavior Type", y="Proportion",
       title="Proportion of Coronary Heart Disease by Behavior Type")
```



The rate of developing heart disease is low for both bahavior types, but is higher for males with aggressive behavior type than for males with passive behavior type. The two way table confirms this. The rate is about 10.6% for males with aggressive behavior type, and is about 5.6% for males with passive behavior type.

```r
##2 way table
mytab<-table(train$dibep, train$chd)
mytab
```

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

```r
prop.table(mytab, 1)
```

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

2)

```
result<-glm(chd ~ age + sdp + dbp + cigs + dibep, family="binomial", data=train)
result
```

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

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -8.309 + 0.060age + 0.015sdp + 0.012dbp + 0.021cigs - 0.527I_1$$

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

3)

Coefficient for `cigs` is 0.021. A couple of interpretations:

- For an additional cigarette smoked per day (on average), the estimated log odds of developing coronory heart disease increases by 0.021, while controlling for age, systolic BP, diastolic BP, and behavior type.
- For an additional cigarette smoked per day (on average), the estimated odds of developing coronory heart disease gets multiplied by a factor of $\exp(0.021) = 1.021$, while controlling for age, systolic BP, diastolic BP, and behavior type.

4)

Coefficient for `dibep` is -0.527. A couple of interpretations:

- The estimated log odds of developing heart disease for males with type B behaviors is 0.527 lower than for males with type A behaviors, while controlling for age, systolic BP, diastolic BP, and number of cigarettes smoked per day.

- The estimated odds of developing heart disease for males with type B behaviors is $\exp(-0.527) = 0.590$ times the odds for males with type A behaviors, while controlling for age, systolic BP, diastolic BP, and number of cigarettes smoked per day.

5)

```
##make prediction for log odds
newdata<-data.frame(age=45, sdp=110, dbp=70, cigs=0, dibep="B")
predict(result,newdata)
```

```
##         1
## -3.621211
```

```
##convert to odds
odds<-exp(predict(result,newdata))
odds
```

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

```
##convert odds to probability
prob<-odds/(1+odds)
prob
```

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

The estimated odds of developing heart disease for this male is 0.02675027. The corresponding probability is 0.02605333.

6)

```
deltaG2<-result$null.deviance-result$deviance
deltaG2
```

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

```
1-pchisq(deltaG2,5)
```

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

$H_0 : \beta_1 = \cdots = \beta_5 = 0$

H_a : at least one of the coefficients in H_0 is not zero.

The test statistic is $\Delta G^2 = 55.49501$ with a p-value virtually 0. So we reject the null hypothesis. The data support the claim that our model is useful, compared to the intercept-only model.

7)

```r
reduced<-glm(chd ~ age + cigs + dibep, family="binomial", data=train)
deltaG2_partial<-reduced$deviance-result$deviance
deltaG2_partial
```

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

```r
1-pchisq(deltaG2_partial,2)
```

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

$H_0 : \beta_2 = \beta_3 = 0$

H_a : at least one of the coefficients in H_0 is not zero.

The test statistic is $\Delta G^2 = 13.70587$ with a p-value virtually 0. So we reject the null hypothesis. The data support going with the full model, so we do not drop both blood pressure predictors.

8)

```r
summary(result)
```

```
##
## Call:
## glm(formula = chd ~ age + sdp + dbp + cigs + dibep, family = "binomial",
##     data = train)
##
## 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 β_3 is $Z = 0.838$ with a large p-value. So we can drop diastolic blood pressure from the model, while leaving the other predictors in the model.

9)

We only remove diastolic blood pressure as a predictor from the model, and keep the other predictors in: age, systolic blood pressure, number of cigarettes smoked per day, and behavior type.