Workfile

Group

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggpubr)
```

```
data <- read.csv("Data/initial_table.csv")
#mean(data$AGE, na.rm = T)
data.work <- dplyr::select(data, ID, AGE, SEX, IBS_POST, DLIT_AG, SIM_GIPERT, endocr_01, endocr_02, ZSN
data.work <- na.omit(data.work)
data.work <- filter(data.work, DLIT_AG != 10)
dim(data.work) # obs = 1380
```

```
## [1] 1380   10
```

```
# exploratory analysis
mean(data.work$AGE) #61.397
```

```
## [1] 61.3971
```

```
median(data.work$AGE) # 62
```

```
## [1] 62
```

```
table(data.work$SEX) # female(0): 502, male(1): 878
```

```
##
##   0   1
## 502 878
```

```r
table(data.work$IBS_POST) # no CHD(0): 353, exertional angina pectoris(1):443, unstable angina pectoris
```

```
##
##   0   1   2
## 353 443 584
```

```r
mean(data.work$DLIT_AG) # 3.34
```

```
## [1] 3.336232
```

```r
median(data.work$DLIT_AG) # 3
```

```
## [1] 3
```

```r
median(data.work$DLIT_AG) # 3
```

```
## [1] 3
```

```r
table(data.work$SIM_GIPERT) # no(0): 1336, yes(1): 44
```

```
##
##    0    1
## 1336   44
```

```r
table(data.work$endocr_01) # no(0): 1193, yes(1):187
```

```
##
##    0    1
## 1193  187
```

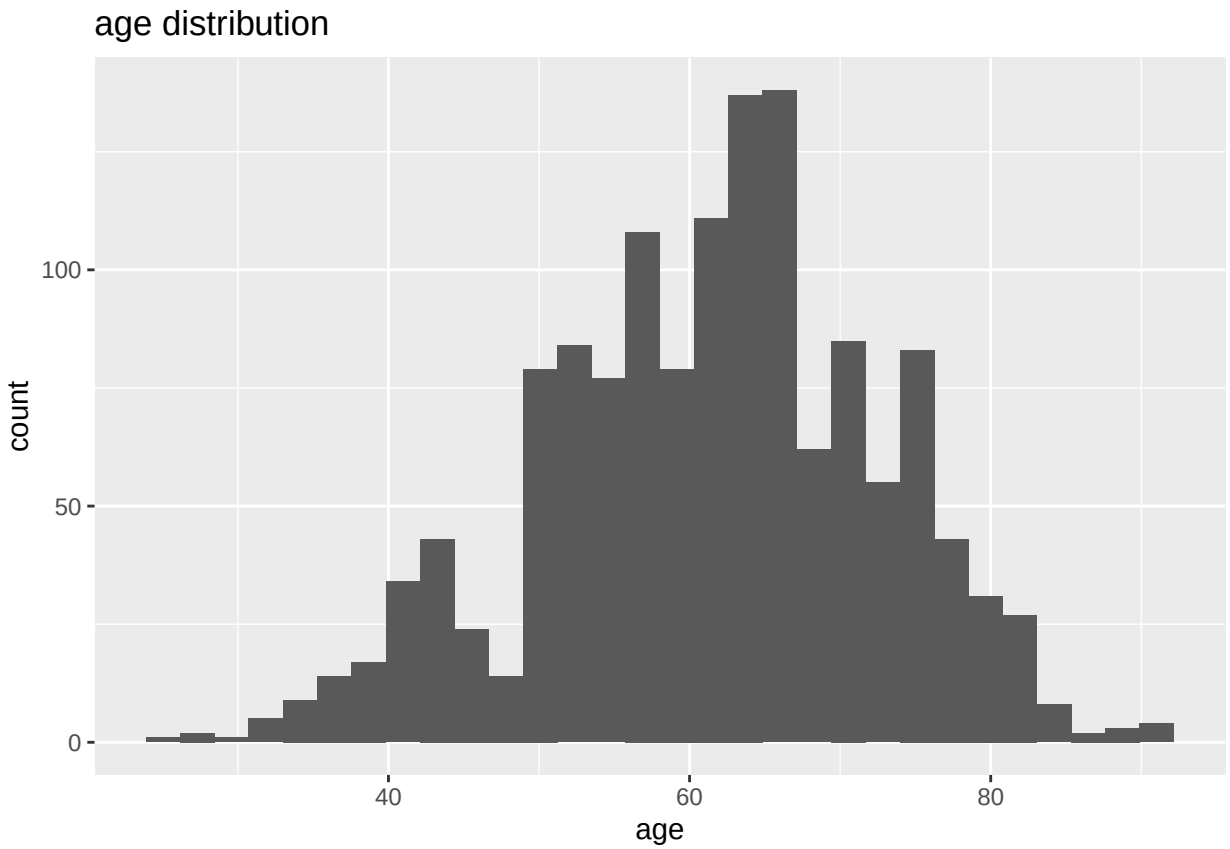
```r
table(data.work$endocr_02) # no(0): 1348, yes(1):32
```

```
##
##    0    1
## 1348   32
```

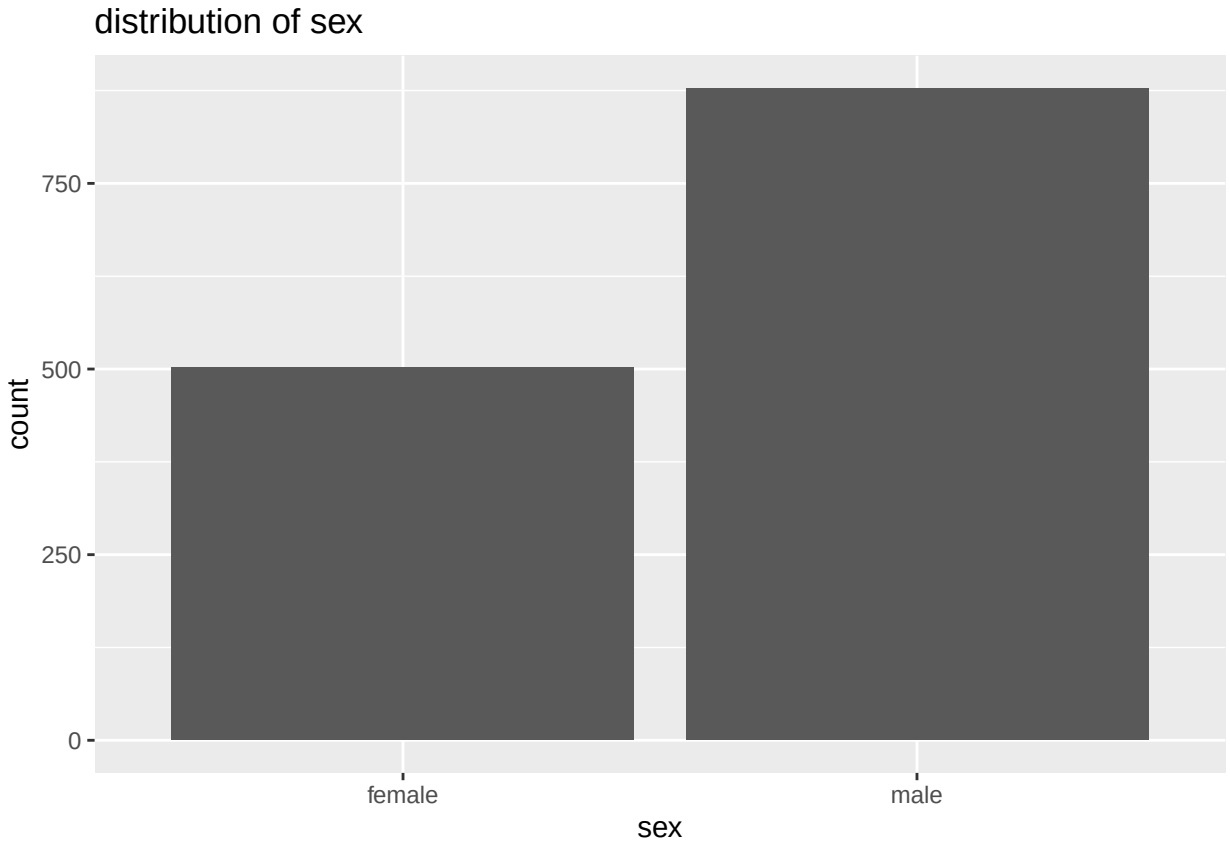
```r
table(data.work$ZSN) # no(0): 1052, yes(1): 328
```

```
##
##    0    1
## 1052  328
```

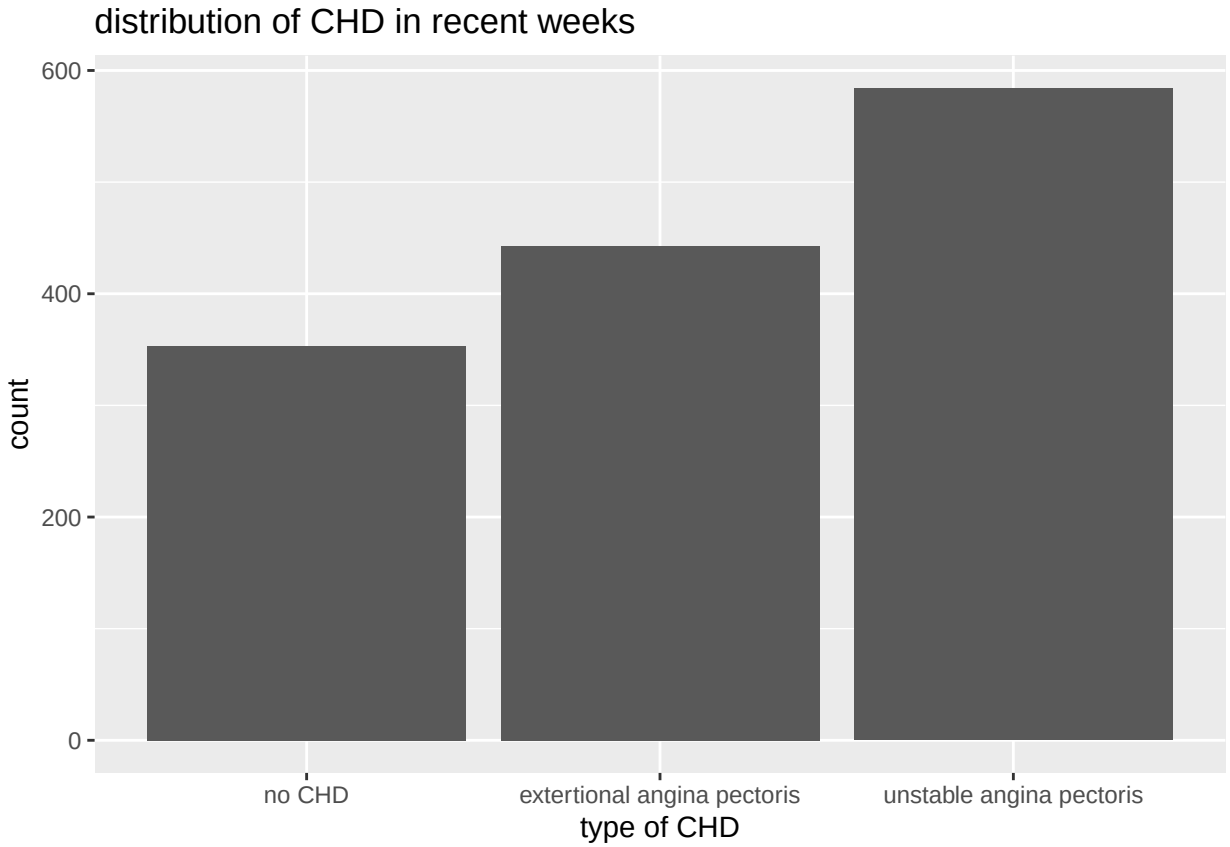
```r
# distribution plots for single variable
age.hist <- ggplot(data.work, aes(data.work$AGE)) + geom_histogram() + labs(title = "age distribution",
age.hist
```



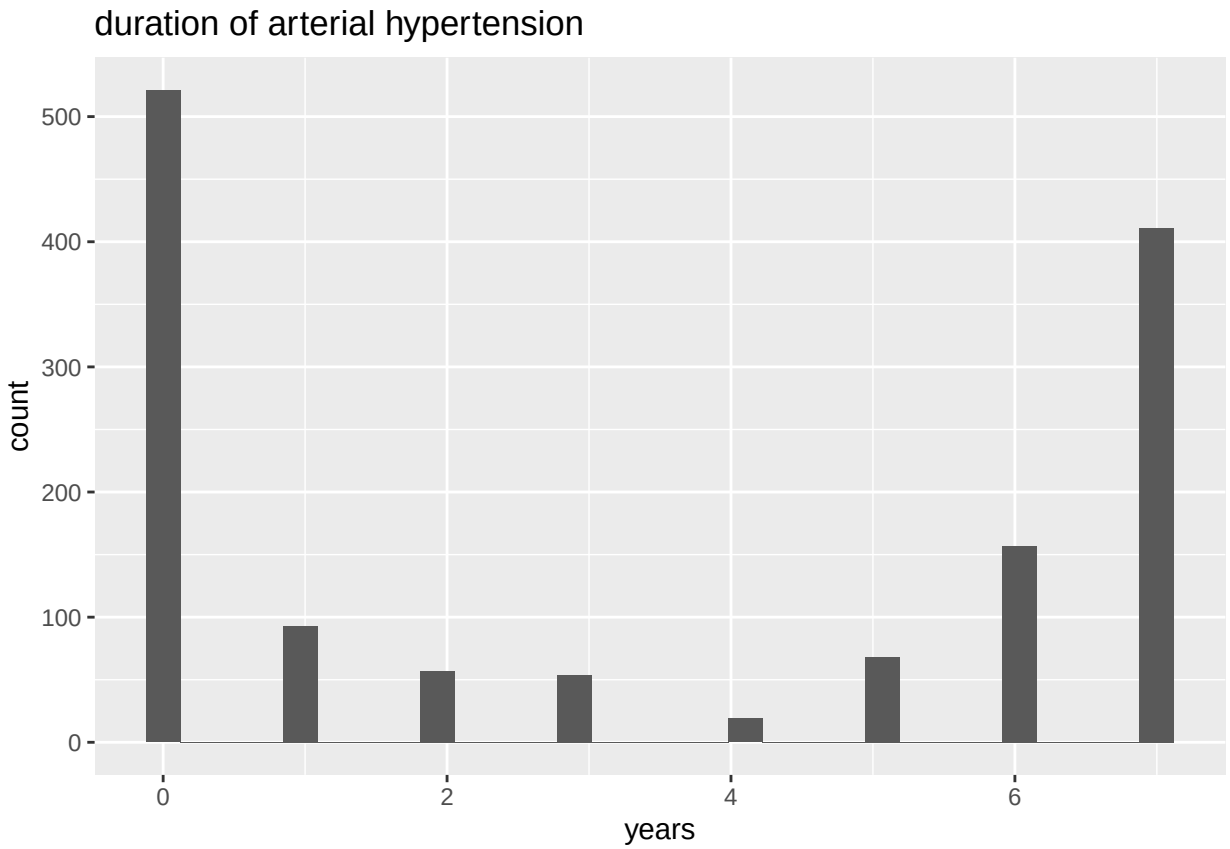

```
sex.plot <- ggplot(data.work, aes(as.factor(data.work$SEX))) + geom_bar() + labs(title = "distribution
sex.plot
```



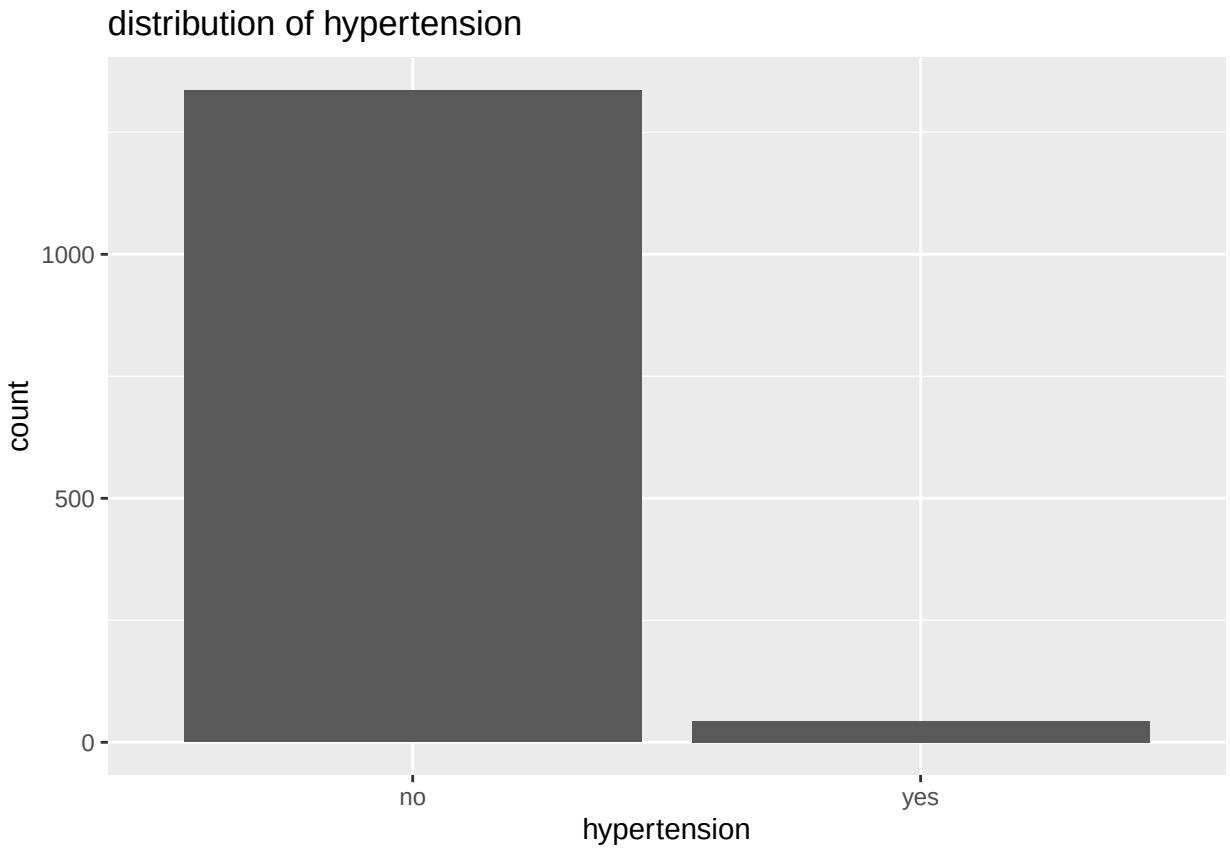

```
ibs.plot <- ggplot(data.work, aes(as.factor(data.work$IBS_POST))) + geom_bar() + labs(title = "distribu
ibs.plot
```



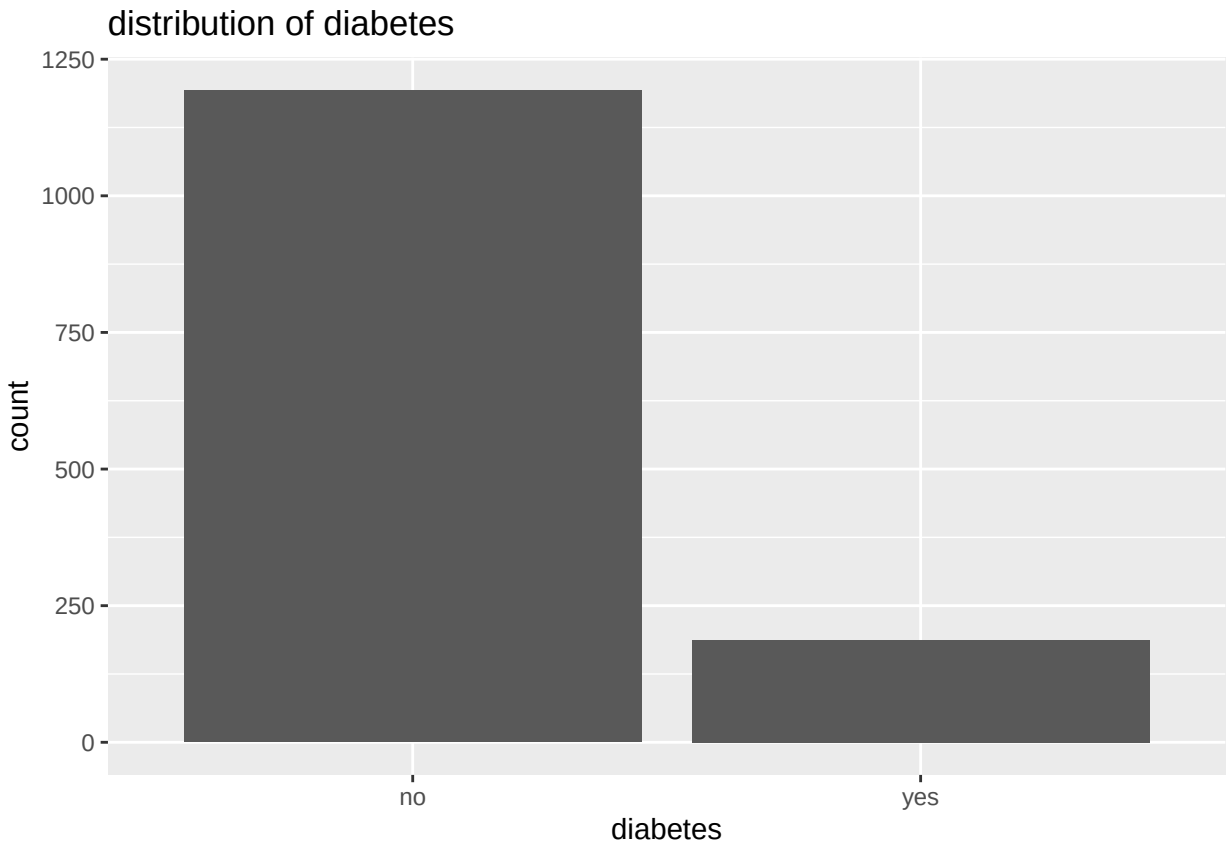

```
duration.hist <- ggplot(data.work, aes(data.work$DLIT_AG)) + geom_histogram() + labs(title = "duration
duration.hist
```



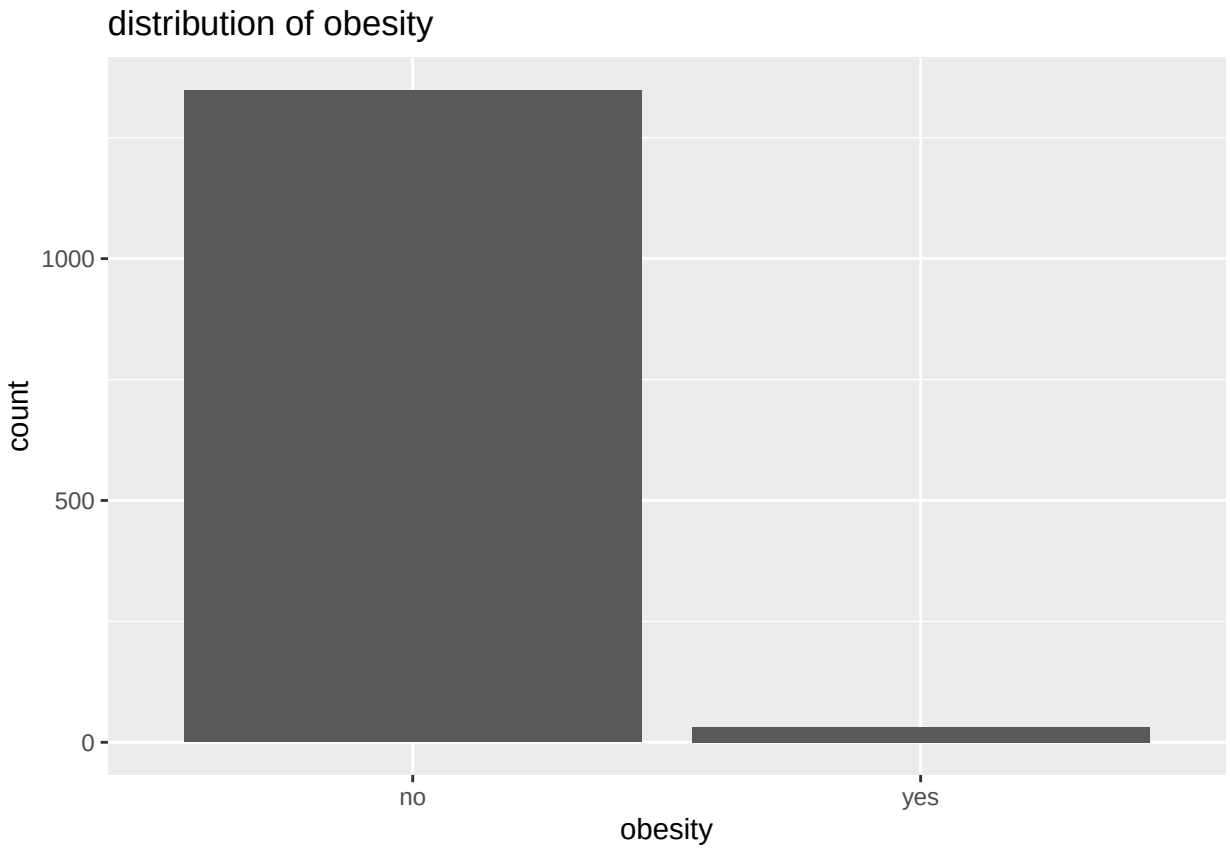

```
hypertension.plot <- ggplot(data.work, aes(as.factor(data.work$SIM_GIPERT))) + geom_bar() + labs(title =
hypertension.plot
```



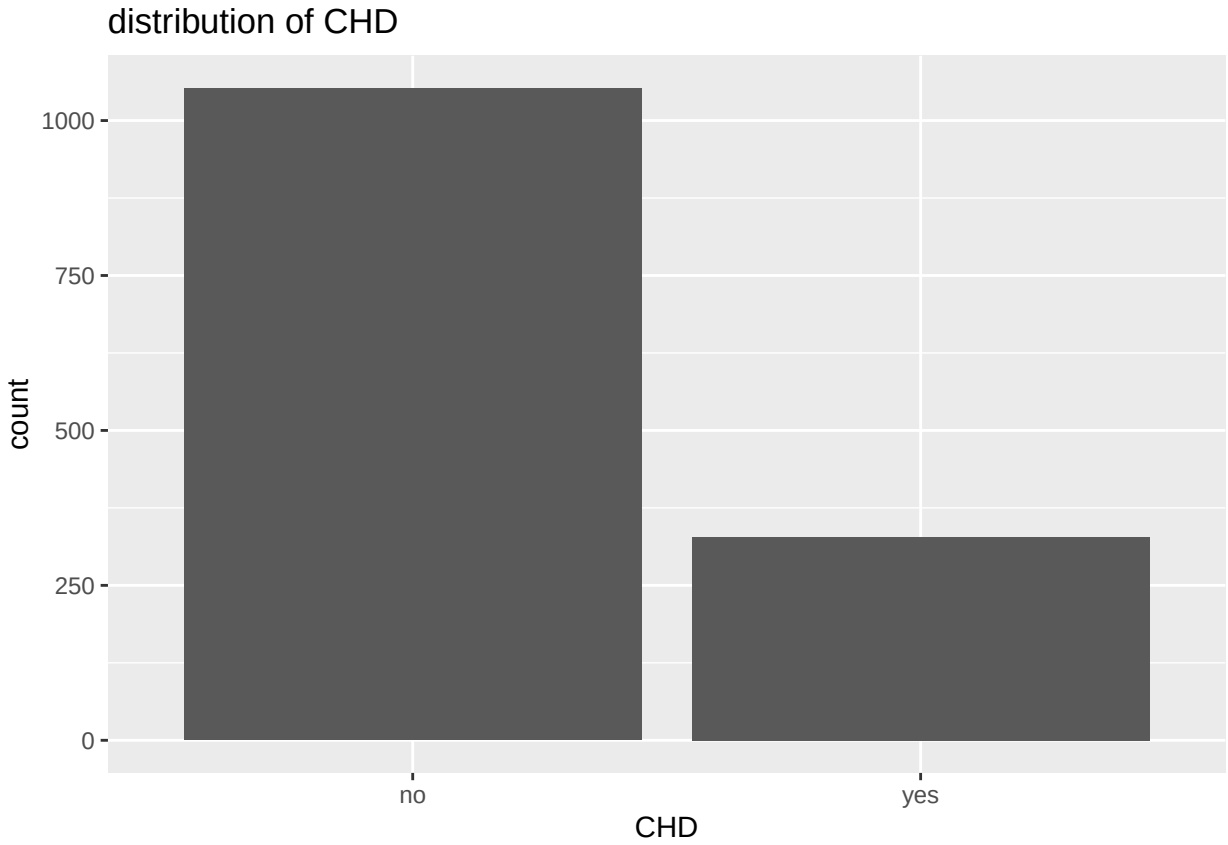

```
diabetes.plot <- ggplot(data.work, aes(as.factor(data.work$endocr_01))) + geom_bar() + labs(title = "dis
diabetes.plot
```




```
obesity.plot <- ggplot(data.work, aes(as.factor(data.work$endocr_02))) + geom_bar() + labs(title = "dis
obesity.plot
```




```
chd.plot <- ggplot(data.work, aes(as.factor(data.work$ZSN))) + geom_bar() + labs(title = "distribution
chd.plot
```




```
names(data)
```

```
##   [1] "ID"            "AGE"           "SEX"           "INF_ANAM"
##   [5] "STENOK_AN"     "FK_STENOK"     "IBS_POST"      "IBS_NASL"
##   [9] "GB"            "SIM_GIPERT"    "DLIT_AG"       "ZSN_A"
##  [13] "nr_11"         "nr_01"         "nr_02"         "nr_03"
##  [17] "nr_04"         "nr_07"         "nr_08"         "np_01"
##  [21] "np_04"         "np_05"         "np_07"         "np_08"
##  [25] "np_09"         "np_10"         "endocr_01"     "endocr_02"
##  [29] "endocr_03"     "zab_leg_01"    "zab_leg_02"    "zab_leg_03"
##  [33] "zab_leg_04"    "zab_leg_06"    "S_AD_KBRIG"    "D_AD_KBRIG"
##  [37] "S_AD_ORIT"     "D_AD_ORIT"     "O_L_POST"      "K_SH_POST"
##  [41] "MP_TP_POST"    "SVT_POST"      "GT_POST"       "FIB_G_POST"
##  [45] "ant_im"        "lat_im"        "inf_im"        "post_im"
##  [49] "IM_PG_P"       "ritm_ecg_p_01" "ritm_ecg_p_02" "ritm_ecg_p_04"
##  [53] "ritm_ecg_p_06" "ritm_ecg_p_07" "ritm_ecg_p_08" "n_r_ecg_p_01"
##  [57] "n_r_ecg_p_02"  "n_r_ecg_p_03"  "n_r_ecg_p_04"  "n_r_ecg_p_05"
##  [61] "n_r_ecg_p_06"  "n_r_ecg_p_08"  "n_r_ecg_p_09"  "n_r_ecg_p_10"
##  [65] "n_p_ecg_p_01"  "n_p_ecg_p_03"  "n_p_ecg_p_04"  "n_p_ecg_p_05"
##  [69] "n_p_ecg_p_06"  "n_p_ecg_p_07"  "n_p_ecg_p_08"  "n_p_ecg_p_09"
##  [73] "n_p_ecg_p_10"  "n_p_ecg_p_11"  "n_p_ecg_p_12"  "fibr_ter_01"
##  [77] "fibr_ter_02"   "fibr_ter_03"   "fibr_ter_05"   "fibr_ter_06"
##  [81] "fibr_ter_07"   "fibr_ter_08"   "GIPO_K"        "K_BLOOD"
##  [85] "GIPER_NA"      "NA_BLOOD"      "ALT_BLOOD"     "AST_BLOOD"
##  [89] "KFK_BLOOD"     "L_BLOOD"       "ROE"           "TIME_B_S"
##  [93] "R_AB_1_n"      "R_AB_2_n"      "R_AB_3_n"      "NA_KB"
```

```
##  [97] "NOT_NA_KB"     "LID_KB"        "NITR_S"        "NA_R_1_n"
## [101] "NA_R_2_n"      "NA_R_3_n"      "NOT_NA_1_n"    "NOT_NA_2_n"
## [105] "NOT_NA_3_n"    "LID_S_n"       "B_BLOK_S_n"    "ANT_CA_S_n"
## [109] "GEPAR_S_n"     "ASP_S_n"       "TIKL_S_n"      "TRENT_S_n"
## [113] "FIBR_PREDS"    "PREDS_TAH"     "JELUD_TAH"     "FIBR_JELUD"
## [117] "A_V_BLOK"      "OTEK_LANC"     "RAZRIV"        "DRESSLER"
## [121] "ZSN"           "REC_IM"        "P_IM_STEN"     "LET_IS"
```

Ariane

Exploring relationship between age and CHD

```
library("DescTools")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.1.0     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4

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

```
#sex and chronic heart failure
data_sex_chf <- table(data.work$SEX,data.work$ZSN)
dimnames(data_sex_chf) <- list(Sex=c("Female","Male"),
                     "Chronic Heart Failure"=c("No","Yes"))
data_sex_chf
```

```
##         Chronic Heart Failure
## Sex       No Yes
##   Female 353 149
##   Male   699 179
```

```
chi_sq_data_sex_chf <-chisq.test(data_sex_chf)
chi_sq_data_sex_chf
```

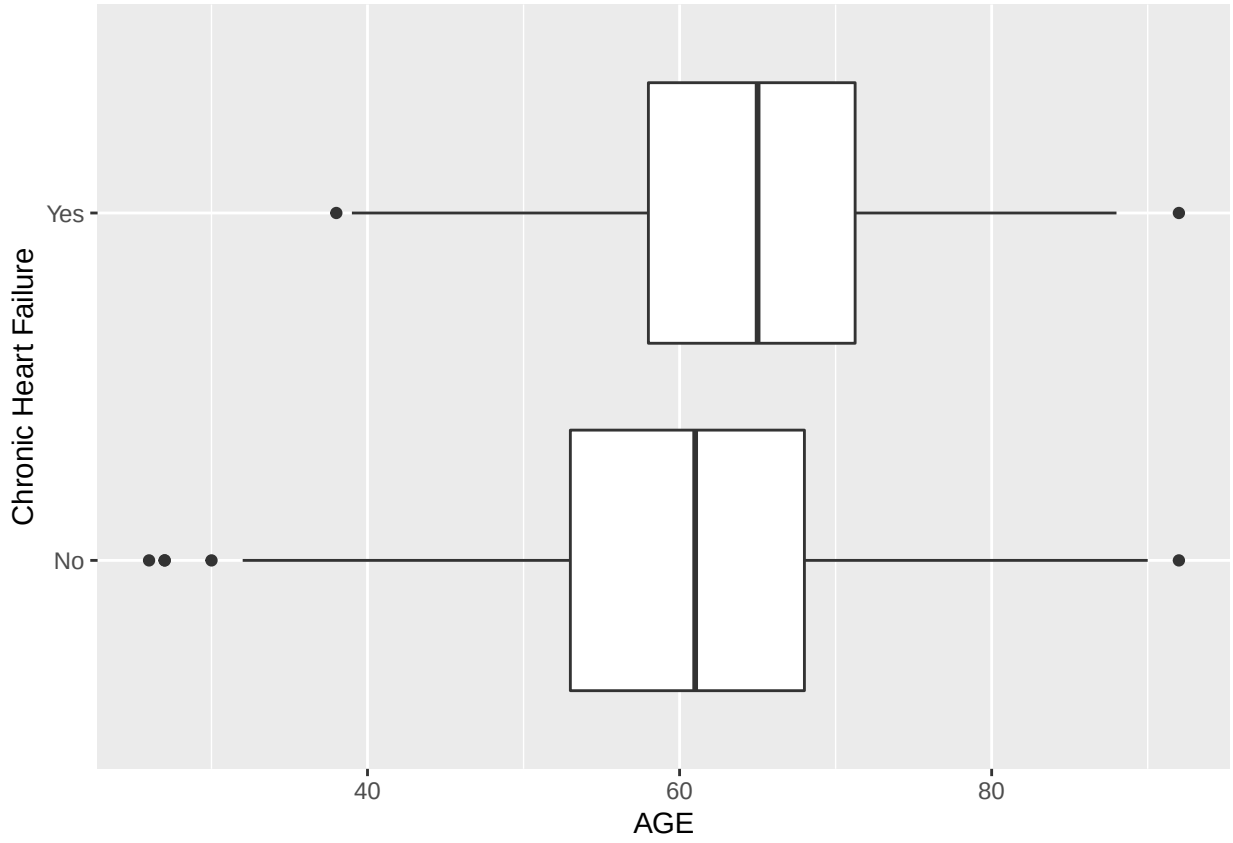
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data_sex_chf
## X-squared = 14.718, df = 1, p-value = 0.0001249
```

```
LR_data_sex_chf <- GTest(data_sex_chf)
LR_data_sex_chf
```

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  data_sex_chf
## G = 14.939, X-squared df = 1, p-value = 0.000111
```

With the p-value<0.01 we reject the null and conclude there is an association between Sex and chronic heart failure

```
# age and chronic heart failure
data_age_chf <- table(data.work$AGE,data.work$ZSN)
dimnames(data_age_chf) <- list(Age = names(data_age_chf[,1]),
                               "Chronic Heart Failure"=c("No","Yes"))
#data_age_chf
boxplot_age_chf <- data.work %>%
  ggplot() +
  geom_boxplot(mapping = aes(x=AGE, y=as.factor(ZSN),
                             group = as.factor(ZSN))) +
  ylab("Chronic Heart Failure") +
  scale_y_discrete(labels=c("No","Yes"))
boxplot_age_chf
```




```
#CHF NO
summary(data.work %>%
          filter(ZSN==0) %>%
          select(AGE))
```

```
##       AGE
##  Min.   :26.00
##  1st Qu.:53.00
##  Median :61.00
```

```
##  Mean   :60.43
##  3rd Qu.:68.00
##  Max.   :92.00
```

```
#CHF YES
summary(data.work %>%
          filter(ZSN==1) %>%
          select(AGE))
```

```
##       AGE
##  Min.   :38.00
##  1st Qu.:58.00
##  Median :65.00
##  Mean   :64.51
##  3rd Qu.:71.25
##  Max.   :92.00
```

```
wilcox.test(data.work$AGE[which(data.work$ZSN == 0)],
            data.work$AGE[which(data.work$ZSN == 1)])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data.work$AGE[which(data.work$ZSN == 0)] and data.work$AGE[which(data.work$ZSN == 1)]
## W = 136546, p-value = 1.113e-08
## alternative hypothesis: true location shift is not equal to 0
```

Results from Wilcoxon Rank Sum test rejects the null with the p-value <0.01 and concludes there is a difference and age between outcomes

```
#look at age categorically by decade
age_decade <- data.work %>%
  mutate(decade = floor(AGE/10)*10) %>%
  select(decade)
data_age_decade_chf <- table(age_decade$decade,data.work$ZSN)
dimnames(data_age_decade_chf) <-
  list(Age = paste0(names(data_age_decade_chf[,1]),"s"),
                                "Chronic Heart Failure"=c("No","Yes"))
data_age_decade_chf
```

```
##      Chronic Heart Failure
## Age    No Yes
##   20s   3   0
##   30s  44   2
##   40s 114  24
##   50s 294  67
##   60s 365 126
##   70s 197  86
##   80s  32  22
##   90s   3   1
```

```
chi_sq_data_age_decade_chf <-chisq.test(data_age_decade_chf)
```

```
## Warning in chisq.test(data_age_decade_chf): Chi-squared approximation may be
## incorrect
```

```
chi_sq_data_age_decade_chf
```

```
##
##  Pearson's Chi-squared test
##
## data:  data_age_decade_chf
## X-squared = 35.419, df = 7, p-value = 9.327e-06
```

```
LR_data_age_decade_chf <- GTest(data_age_decade_chf)
LR_data_age_decade_chf
```

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  data_age_decade_chf
## G = 38.862, X-squared df = 7, p-value = 2.077e-06
```

Using the age by decade we have a p-value<0.01 which like the wilcoxon test suggest an association between age an chronic heart failure due to the rejection of the null

Alona

Exploring the relationship between CHF and Duration of arterial hypertension.

```
library(knitr)
library(tidyverse)
library(vcdExtra, quietly = TRUE)
```

```
##
## Attaching package: 'vcdExtra'

## The following object is masked from 'package:dplyr':
##
##     summarise
```

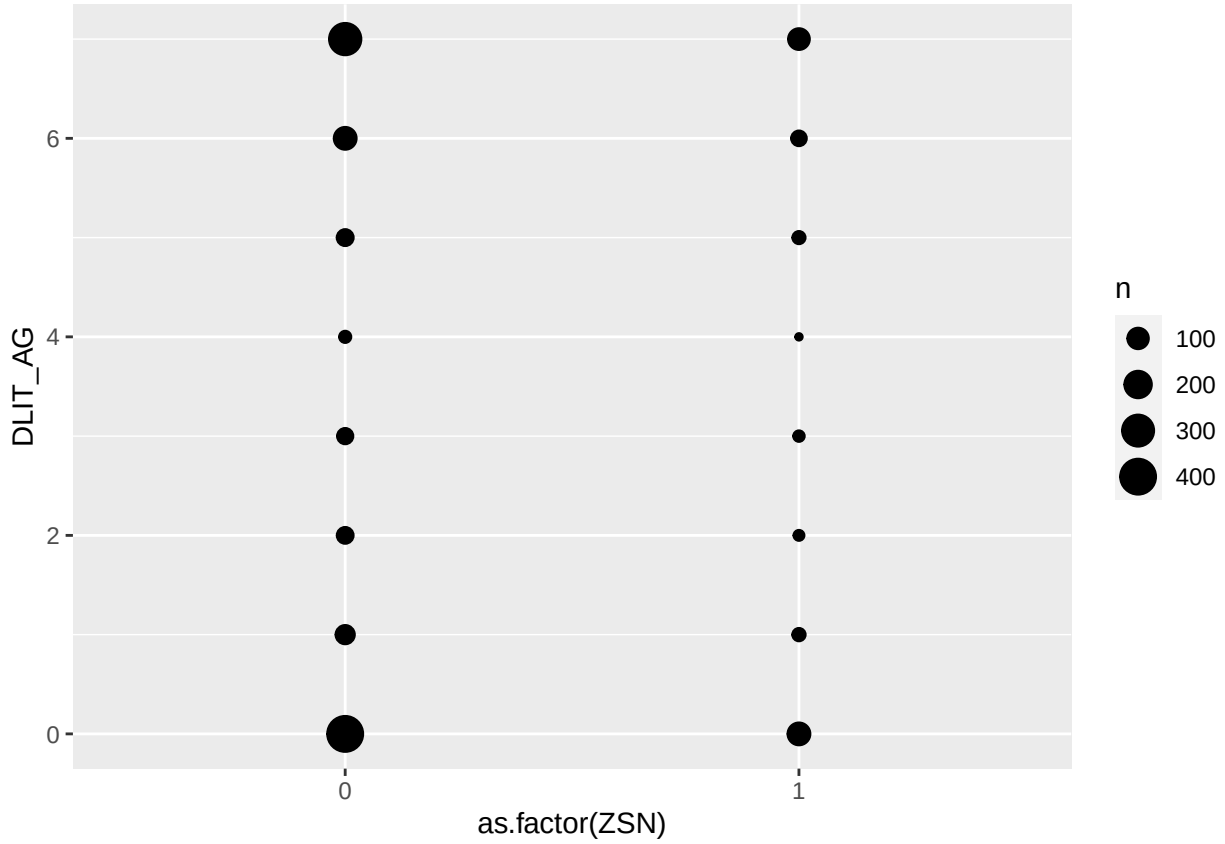
```
library("DescTools")
library("ResourceSelection")
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
# Duration of arterial hypertension (DLIT_AG): Ordinal
freq.dlitag <- data.work %>%
  group_by(DLIT_AG) %>%
  dplyr::summarize(n = n()) %>%
  mutate(freq = n/sum(n))
freq.dlitag
```

```
## # A tibble: 8 x 3
##   DLIT_AG     n   freq
##     <int> <int>  <dbl>
## 1       0   521 0.378
## 2       1    93 0.0674
## 3       2    57 0.0413
## 4       3    54 0.0391
## 5       4    19 0.0138
## 6       5    68 0.0493
## 7       6   157 0.114
## 8       7   411 0.298
```

```
ggplot(data.work, aes(x = as.factor(ZSN), y = DLIT_AG)) +
  geom_count()
```



The two classes of CHF have similar distribution of proportions across the level of duration of arterial hypertension. We will further test the hypothesis that theres is an association between the two variables.

```
# removing category 10 which is likely a mistake.
data.work.2 <- data.work %>%
  filter(DLIT_AG != 10)
data.work.3 <- data.work %>%
  mutate(DLIT_AG_N = case_when(DLIT_AG==6 ~ 8,
                               DLIT_AG==7 ~ 10,
                               DLIT_AG==0 ~ 0,
                               DLIT_AG==1 ~ 1,
```

```
                                DLIT_AG==2 ~ 2,
                                DLIT_AG==3 ~ 3,
                                DLIT_AG==4 ~ 4,
                                DLIT_AG==5 ~ 5
                                ))
mean(data.work.2$DLIT_AG) # 3.36
```

```
## [1] 3.336232
```

```
median(data.work.2$DLIT_AG) #3
```

```
## [1] 3
```

```
tab <- table(data.work.2$DLIT_AG,data.work.2$ZSN)
dimnames(tab) <- list("Duration of AH"=c("None","1-year","2-years","3-years","4-years",
                                         "5-years","6-10 years",">=10 years"),
                      "Chronic Heart Failure"=c("No","Yes"))
tab2 <- table(data.work.3$DLIT_AG_N,data.work.3$ZSN)
# contingency table
dlitag <- as.table(tab)
kable(dlitag,
      caption = "Duration of Arterial Hypertension by Chronic Heart Failure")
```

Table 1: Duration of Arterial Hypertension by Chronic Heart Failure

	No	Yes
None	401	120
1-year	72	21
2-years	47	10
3-years	42	12
4-years	15	4
5-years	48	20
6-10 years	120	37
>=10 years	307	104

Duration of Arterial Hypertension is an ordinal type variable. we therefore use ordinal trend tests

```
#Ordinal trend test
gamma.test <- GKgamma(dlitag)
pvalg=2*pnorm(q=gamma.test$gamma/gamma.test$sigma, lower.tail=FALSE)
pvalg
```

```
## [1] 0.3853727
```

```
# Cochran Armitage Test for Ix2 tables - section 5.3.5 in the book
coarm <- CochranArmitageTest(dlitag)
coarm
```

```
##
##  Cochran-Armitage test for trend
##
## data:  dlitag
## Z = -0.99455, dim = 8, p-value = 0.32
## alternative hypothesis: two.sided
```

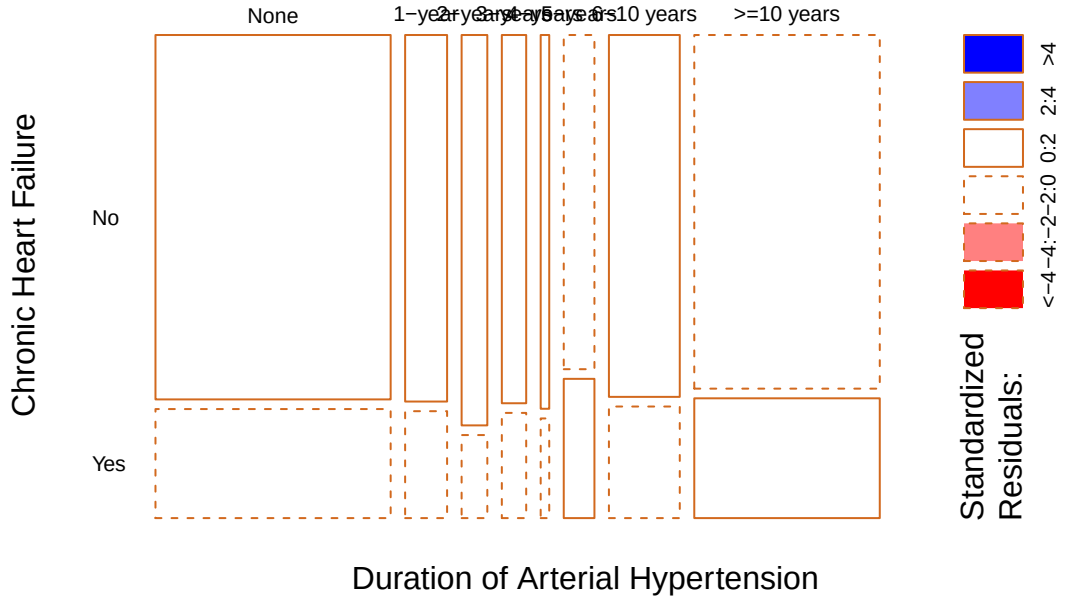
```
# chisq test can be used but is less powerful than the two above.
chisq <- round(chisq.test(dlitag)$statistic,3)
```

```
## Warning in chisq.test(dlitag): Chi-squared approximation may be incorrect
```

```
#pval <- round(chisq.test(dlitag)$p.value,3)
#lrt <- GTest(dlitag)
std.res <- chisq.test(dlitag)$stdres
```

```
## Warning in chisq.test(dlitag): Chi-squared approximation may be incorrect
```

```
# all p-values from all test are confirming the finding that there is no relationship between
# duration of arterial hypertension and chronic heart failure
# residual analysis
# this is just a cool plot - unfortunately nothing is significant so there is no color.
mosaicplot(dlitag,
           main = "",
           xlab = "Duration of Arterial Hypertension",
           ylab = "Chronic Heart Failure",
           las = 1,
           border = "chocolate",
           shade = TRUE)
```



All tests have non-significant p-value (>0.2) which suggest that we do not reject the null of no association.

```
# Logistic regression models for Chronic heart failure - ZSN as a function of DLIT_AG
# canonical link
fit.dlit.l <- glm(ZSN ~ DLIT_AG, data=data.work.2, family=binomial)
summary(fit.dlit.l)
```

```
##
## Call:
## glm(formula = ZSN ~ DLIT_AG, family = binomial, data = data.work.2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7607  -0.7540  -0.7148  -0.7148   1.7261
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.23418    0.09449 -13.062   <2e-16 ***
## DLIT_AG      0.02029    0.02041   0.994     0.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1513.6  on 1379  degrees of freedom
## Residual deviance: 1512.6  on 1378  degrees of freedom
## AIC: 1516.6
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
# fit.dlitn.l <- glm(ZSN ~ DLIT_AG_N, data=data.work.3, family=binomial)
# summary(fit.dlitn.l)
# cloglog link
fit.dlit.cll <- glm(ZSN ~ DLIT_AG, data=data.work.2, family=binomial(link="cloglog"))
summary(fit.dlit.cll)
```

```
##
## Call:
## glm(formula = ZSN ~ DLIT_AG, family = binomial(link = "cloglog"),
##     data = data.work.2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7607  -0.7540  -0.7148  -0.7148   1.7261
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.36466    0.08331 -16.380   <2e-16 ***
## DLIT_AG      0.01779    0.01787   0.996    0.319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1513.6  on 1379  degrees of freedom
## Residual deviance: 1512.6  on 1378  degrees of freedom
## AIC: 1516.6
##
## Number of Fisher Scoring iterations: 5
```

```
# identity link
fit.dlit.i <- glm(ZSN ~ DLIT_AG, data=data.work.2, family=binomial(link="identity"))
summary(fit.dlit.i)
```

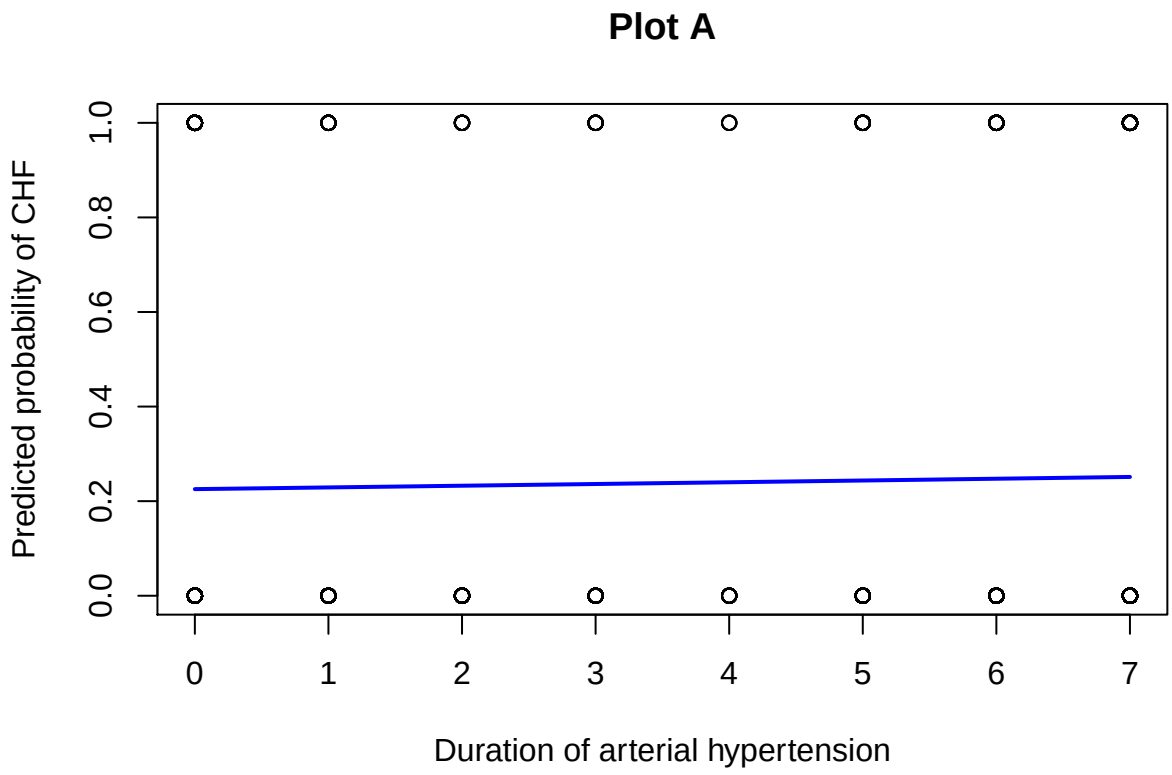
```
##
## Call:
## glm(formula = ZSN ~ DLIT_AG, family = binomial(link = "identity"),
##     data = data.work.2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7604  -0.7540  -0.7149  -0.7149   1.7260
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.225485   0.016566  13.611   <2e-16 ***
## DLIT_AG     0.003656   0.003701   0.988    0.323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1513.6  on 1379  degrees of freedom
## Residual deviance: 1512.6  on 1378  degrees of freedom
## AIC: 1516.6
##
## Number of Fisher Scoring iterations: 3
```

```
#goodness of fit
G.sq=deviance(fit.dlit.l)
df.fit <- fit.dlit.l$df.residual
p.val=1-pchisq(G.sq,df.fit)
p.val
```

```
## [1] 0.006261823
```

```
newdata <- data.frame(DLIT_AG=seq(min(data.work.2$DLIT_AG), max(data.work.2$DLIT_AG),len=23))
newdata$ZSN <- predict(fit.dlit.l, newdata=newdata, type="response")
plot(ZSN~DLIT_AG, data=data.work.2, col="black",
     main = "Plot A",
     ylab = "Predicted probability of CHF",
     xlab = "Duration of arterial hypertension")
lines(ZSN~DLIT_AG, newdata, col="Blue", lwd=2)
```



The logistic regression model for CHF as explained by duration of arterial hypertension is not predictive.

The predicted probabilities are effectively constant and the goodness of fit value is 0.0062618 suggesting we reject the null of the model fitting the data.

Considering the U shaped distribution of the variable, We also conducted the analysis for the dichotomized (at the median) variable. The results were no different than in the original form. We also evaluated if a binary cut of the duration of arterial hypertension to no arterial hypertension (category of 0) vs. duration of arterial hypertension > 0 has more meaningful association with CHF and here too, the results were not different.

In conclusion, the variable of duration of arterial hypertension by itself is not associated with the outcome of chronic heart failure. This ordinal variable was tested in the original form - with equally spaced categories - and was also evaluated with an adjustment of score assignment for the last two categories (that are not one-to-one mapping of name to value)

Minsu

Build a multivariable logistic regression model, identifying the best model, and calculating predictive power of the model.

```
data.work <- data.work %>%
  mutate(DLIT_AG_N = case_when(DLIT_AG==6 ~ 8,
                               DLIT_AG==7 ~ 10,
                               DLIT_AG==0 ~ 0,
                               DLIT_AG==1 ~ 1,
                               DLIT_AG==2 ~ 2,
                               DLIT_AG==3 ~ 3,
                               DLIT_AG==4 ~ 4,
                               DLIT_AG==5 ~ 5
                               ))
#fit a model with all 7 predictors
data.work$SIM.f <- factor(data.work$SIM_GIPERT, levels=c(0,1), labels = c("no","yes"))
data.work$endocr_01.f <- factor(data.work$endocr_01, levels=c(0,1), labels = c("no","yes"))
data.work$endocr_02.f <- factor(data.work$endocr_02, levels=c(0,1), labels = c("no","yes"))
chf.dat <- select(data.work, AGE, SEX, IBS_POST, DLIT_AG, SIM.f, endocr_01.f, endocr_02.f, ZSN)
fit<- glm(ZSN ~ . , data=chf.dat, family=binomial)
summary(fit)
```

```
##
## Call:
## glm(formula = ZSN ~ ., family = binomial, data = chf.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2617  -0.7582  -0.6408  -0.4645   2.1518
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.005772   0.467243  -6.433 1.25e-10 ***
## AGE             0.031965   0.006696   4.774 1.81e-06 ***
## SEX            -0.171381   0.150348  -1.140    0.254
## IBS_POST       -0.032669   0.082903  -0.394    0.694
## DLIT_AG        -0.037428   0.023127  -1.618    0.106
## SIM.fyes       -0.400837   0.408905  -0.980    0.327
## endocr_01.fyes  0.747247   0.177495   4.210 2.55e-05 ***
```

```
## endocr_02.fyes  0.146158   0.410614   0.356    0.722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1513.6  on 1379  degrees of freedom
## Residual deviance: 1456.6  on 1372  degrees of freedom
## AIC: 1472.6
##
## Number of Fisher Scoring iterations: 4
```

```
#overall test for model with 7 predictors
fit.0<- glm(ZSN ~ 1. , data=chf.dat, family=binomial)
summary(fit.0)
```

```
##
## Call:
## glm(formula = ZSN ~ 1, family = binomial, data = chf.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7367  -0.7367  -0.7367  -0.7367   1.6952
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.16543    0.06324  -18.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1513.6  on 1379  degrees of freedom
## Residual deviance: 1513.6  on 1379  degrees of freedom
## AIC: 1515.6
##
## Number of Fisher Scoring iterations: 4
```

```
lr <- deviance(fit.0) - deviance(fit)
df <- summary(fit.0)$df[2]-summary(fit)$df[2]
p.val <- 1 - pchisq(lr, df=df)
p.val
```

```
## [1] 6.188061e-10
```

There is strong evidence that at least one predictor has an effect. Although the overall test is highly significant, summary(fit) results show that only AGE and endocr_01 seems significant in the Wald test.

```
#add AGE and endocr_01 to the logistic model in subtopic 2.
fit.ini <- glm(ZSN~ DLIT_AG, data=chf.dat, family=binomial)
fit.add <- glm(ZSN~ DLIT_AG + AGE + endocr_01.f, data=chf.dat, family=binomial)
#goodnes of fit
```

```
G.sq=deviance(fit.add)
df.fit <- fit.add$df.residual
p.val=1-pchisq(G.sq,df.fit)
#compare this additive model with the initial model with only DLIT_AG
anova(fit.ini, fit.add)
```

```
## Analysis of Deviance Table
##
## Model 1: ZSN ~ DLIT_AG
## Model 2: ZSN ~ DLIT_AG + AGE + endocr_01.f
##   Resid. Df Resid. Dev Df Deviance
## 1      1378     1512.6
## 2      1376     1459.2  2   53.348
```

```
lr <- fit.ini$deviance - fit.add$deviance
df <- anova(fit.ini, fit.add, test="LRT")$Df[2]
p.val <- 1 - pchisq(lr, df=df)
p.val
```

```
## [1] 2.604583e-12
```

The model with AGE and endocr_01 in addition to DLIT_AG improves the goodness-of-fit. Next, we perform stepwise model selection through the forward and backward elimination methods to see if there is effect of interaction between predictors.

```
#Backward selection
fit.3 <- glm(ZSN~ DLIT_AG* AGE * endocr_01.f, data=chf.dat, family=binomial)
mod.back <- step(fit.3, scope=list(lower = ~ 1, upper = formula(fit.3)), scale = 1, trace = T, directio
```

```
## Start:  AIC=1467.2
## ZSN ~ DLIT_AG * AGE * endocr_01.f
##
##                            Df Deviance    AIC
## - DLIT_AG:AGE:endocr_01.f  1   1451.4 1465.4
## <none>                         1451.2 1467.2
##
## Step:  AIC=1465.44
## ZSN ~ DLIT_AG + AGE + endocr_01.f + DLIT_AG:AGE + DLIT_AG:endocr_01.f +
##     AGE:endocr_01.f
##
##                        Df Deviance    AIC
## - DLIT_AG:AGE           1   1452.0 1464.0
## - AGE:endocr_01.f       1   1452.0 1464.0
## <none>                      1451.4 1465.4
## - DLIT_AG:endocr_01.f   1   1458.7 1470.7
##
## Step:  AIC=1463.99
## ZSN ~ DLIT_AG + AGE + endocr_01.f + DLIT_AG:endocr_01.f + AGE:endocr_01.f
##
##                        Df Deviance    AIC
## - AGE:endocr_01.f       1   1452.4 1462.4
## <none>                      1452.0 1464.0
```

```
## - DLIT_AG:endocr_01.f  1   1459.2 1469.2
##
## Step:  AIC=1462.4
## ZSN ~ DLIT_AG + AGE + endocr_01.f + DLIT_AG:endocr_01.f
##
##                        Df Deviance    AIC
## <none>                      1452.4 1462.4
## - DLIT_AG:endocr_01.f  1   1459.2 1467.2
## - AGE                  1   1482.2 1490.2
```

```
res.back <- mod.back$anova
res.back
```

```
##                          Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                             NA        NA      1372   1451.198 1467.198
## 2 - DLIT_AG:AGE:endocr_01.f  1 0.2453927      1373   1451.444 1465.444
## 3             - DLIT_AG:AGE  1 0.5448708      1374   1451.989 1463.989
## 4         - AGE:endocr_01.f  1 0.4089297      1375   1452.398 1462.398
```

```
#Forward selection
fit.0 <- glm(ZSN ~ 1 , data=chf.dat, family=binomial)
mod.for <- step(fit.0, scope=list(lower = ~ 1, upper = formula(fit.3)), scale = 1, trace = T, direction
```

```
## Start:  AIC=1515.56
## ZSN ~ 1
##
##               Df Deviance    AIC
## + AGE          1   1479.0 1483.0
## + endocr_01.f  1   1489.9 1493.9
## <none>             1513.6 1515.6
## + DLIT_AG      1   1512.6 1516.6
##
## Step:  AIC=1483.05
## ZSN ~ AGE
##
##               Df Deviance    AIC
## + endocr_01.f  1   1461.7 1467.7
## <none>             1479.0 1483.0
## + DLIT_AG      1   1478.5 1484.5
##
## Step:  AIC=1467.65
## ZSN ~ AGE + endocr_01.f
##
##                   Df Deviance    AIC
## + DLIT_AG          1   1459.2 1467.2
## <none>                 1461.7 1467.7
## + AGE:endocr_01.f  1   1461.7 1469.7
##
## Step:  AIC=1467.23
## ZSN ~ AGE + endocr_01.f + DLIT_AG
##
##                        Df Deviance    AIC
## + DLIT_AG:endocr_01.f  1   1452.4 1462.4
```

```
## <none>                      1459.2 1467.2
## + DLIT_AG:AGE          1    1458.7 1468.7
## + AGE:endocr_01.f      1    1459.2 1469.2
##
## Step:  AIC=1462.4
## ZSN ~ AGE + endocr_01.f + DLIT_AG + endocr_01.f:DLIT_AG
##
##                   Df Deviance    AIC
## <none>                 1452.4 1462.4
## + AGE:endocr_01.f  1   1452.0 1464.0
## + DLIT_AG:AGE      1   1452.0 1464.0
```

```
res.for <- mod.for$anova
res.for
```

```
##                     Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                        NA        NA      1379   1513.563 1515.563
## 2                 + AGE -1 34.508682      1378   1479.054 1483.054
## 3         + endocr_01.f -1 17.399145      1377   1461.655 1467.655
## 4             + DLIT_AG -1  2.428388      1376   1459.226 1467.226
## 5 + DLIT_AG:endocr_01.f -1  6.828922      1375   1452.398 1462.398
```

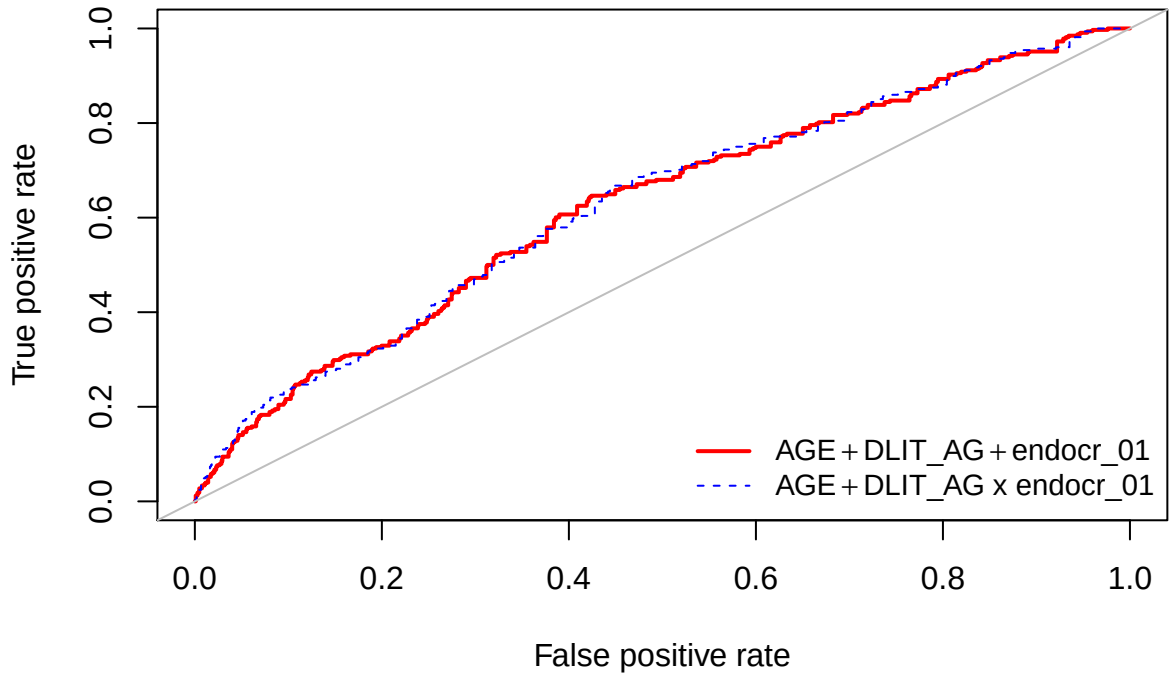
```
#fit the best model
fit.best <- glm(ZSN ~ AGE + DLIT_AG * endocr_01.f , data=chf.dat, family=binomial)
summary(fit.best)
```

```
##
## Call:
## glm(formula = ZSN ~ AGE + DLIT_AG * endocr_01.f, family = binomial,
##     data = chf.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4664  -0.7485  -0.6431  -0.4633   2.0988
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -3.383155   0.397092  -8.520  < 2e-16 ***
## AGE                     0.034151   0.006377   5.355 8.54e-08 ***
## DLIT_AG                -0.010223   0.023864  -0.428  0.66837
## endocr_01.fyes          1.472262   0.318678   4.620 3.84e-06 ***
## DLIT_AG:endocr_01.fyes -0.153230   0.058883  -2.602  0.00926 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1513.6  on 1379  degrees of freedom
## Residual deviance: 1452.4  on 1375  degrees of freedom
## AIC: 1462.4
##
## Number of Fisher Scoring iterations: 4
```

```r
#goodness of fit
G.sq=deviance(fit.best)
df.fit.best <- fit.best$df.residual
p.val=1-pchisq(G.sq,df.fit.best)
#compare this best model with the additive model
lr <- fit.add$deviance - fit.best$deviance
df <- anova(fit.ini, fit.best, test="LRT")$Df[2]
p.val <- 1 - pchisq(lr, df=df)
p.val
```

```
## [1] 0.07755518
```

```r
#predictive power using ROC curve
library(ROCR)
pred1 <- prediction(fitted(fit.add), chf.dat$ZSN)
val1 <- performance(pred1, 'tpr', 'fpr')
pred2 <- prediction(fitted(fit.best), chf.dat$ZSN)
val2 <- performance(pred2, 'tpr', 'fpr')
lab1 <- expression('AGE'+'DLIT_AG'+'endocr_01')
lab2 <- expression('AGE'+'DLIT_AG x endocr_01')
plot(val1@x.values[[1]], val1@y.values[[1]], type='s', ylab=val1@y.name, xlab=val1@x.name, col='red', lt
lines(val2@x.values[[1]], val2@y.values[[1]], type='s', col='blue', lty=2)
abline(0,1, col='gray')
legend('bottomright', c(lab1, lab2), col=c('red','blue'), lwd=c(2,1), lty=1:2, cex=.9, bty='n')
```



The model with the interaction between AGE and DLIT_AG and endocr_01.f doesn't improve the

goodness-of-fit.
As seen in the ROC Figure, their ROC curves are very close, thus we expect their performance for prediction will be almost same.

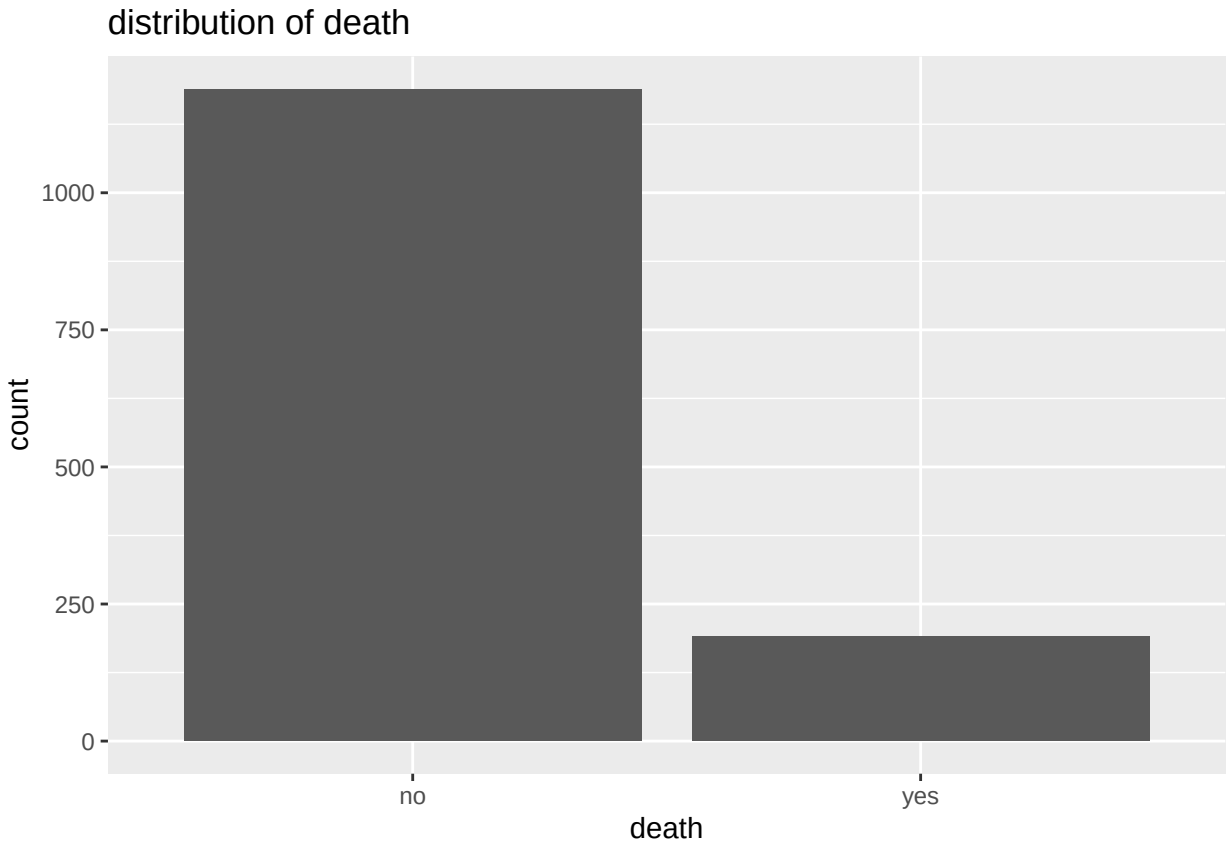
Jadey

Build a mutivariable logistic regression model to predict the death of the cohort and check model prediction accuracy.

```
data.work2 <- data.work
data.work2$death <- ifelse(data.work$LET_IS == 0, 0, 1)
table(data.work2$death) # survive: 1212, dead: 191
```

```
##
##    0    1
## 1189  191
```

```
ggplot(data.work, aes(as.factor(data.work2$death))) + geom_bar() + labs(title = "distribution of death"
```




```
# use stepwise selection to select variable
death.fit0 <- glm(death ~ 1, data = data.work2, family = binomial)
death.fit1 <- glm(death ~ AGE + as.factor(SEX) + as.factor(IBS_POST) + DLIT_AG + as.factor(SIM_GIPERT) +
step(death.fit1, death.fit0, direction = "both") # selected variable: AGE, IBS_POST, SIM_GIPERT, endocr
```

```
## Start:  AIC=1041.44
## death ~ AGE + as.factor(SEX) + as.factor(IBS_POST) + DLIT_AG +
##     as.factor(SIM_GIPERT) + as.factor(endocr_01) + as.factor(endocr_02) +
##     AGE * IBS_POST + AGE * DLIT_AG + AGE * SIM_GIPERT + AGE *
##     endocr_01 + AGE * endocr_02
##
##
## Step:  AIC=1041.44
## death ~ AGE + as.factor(SEX) + as.factor(IBS_POST) + DLIT_AG +
##     as.factor(SIM_GIPERT) + as.factor(endocr_01) + IBS_POST +
##     SIM_GIPERT + endocr_01 + endocr_02 + AGE:IBS_POST + AGE:DLIT_AG +
##     AGE:SIM_GIPERT + AGE:endocr_01 + AGE:endocr_02
##
##
## Step:  AIC=1041.44
## death ~ AGE + as.factor(SEX) + as.factor(IBS_POST) + DLIT_AG +
##     as.factor(SIM_GIPERT) + IBS_POST + SIM_GIPERT + endocr_01 +
##     endocr_02 + AGE:IBS_POST + AGE:DLIT_AG + AGE:SIM_GIPERT +
##     AGE:endocr_01 + AGE:endocr_02
##
##
## Step:  AIC=1041.44
## death ~ AGE + as.factor(SEX) + as.factor(IBS_POST) + DLIT_AG +
##     IBS_POST + SIM_GIPERT + endocr_01 + endocr_02 + AGE:IBS_POST +
##     AGE:DLIT_AG + AGE:SIM_GIPERT + AGE:endocr_01 + AGE:endocr_02
##
##                       Df Deviance    AIC
## - AGE:endocr_02        1   1013.4 1039.4
## - as.factor(SEX)       1   1013.5 1039.5
## - AGE:IBS_POST         1   1013.7 1039.7
## - AGE:DLIT_AG          1   1014.0 1040.0
## - AGE:SIM_GIPERT       1   1014.6 1040.6
## <none>                     1013.4 1041.4
## - AGE:endocr_01        1   1015.7 1041.7
## - as.factor(IBS_POST)  1   1016.2 1042.2
##
## Step:  AIC=1039.44
## death ~ AGE + as.factor(SEX) + as.factor(IBS_POST) + DLIT_AG +
##     IBS_POST + SIM_GIPERT + endocr_01 + endocr_02 + AGE:IBS_POST +
##     AGE:DLIT_AG + AGE:SIM_GIPERT + AGE:endocr_01
##
##                       Df Deviance    AIC
## - as.factor(SEX)       1   1013.5 1037.5
## - AGE:IBS_POST         1   1013.7 1037.7
## - AGE:DLIT_AG          1   1014.0 1038.0
## - AGE:SIM_GIPERT       1   1014.6 1038.6
## <none>                     1013.4 1039.4
## - AGE:endocr_01        1   1015.7 1039.7
## - as.factor(IBS_POST)  1   1016.2 1040.2
## - endocr_02            1   1018.6 1042.6
##
## Step:  AIC=1037.5
## death ~ AGE + as.factor(IBS_POST) + DLIT_AG + IBS_POST + SIM_GIPERT +
##     endocr_01 + endocr_02 + AGE:IBS_POST + AGE:DLIT_AG + AGE:SIM_GIPERT +
```

```
##     AGE:endocr_01
##
##                       Df Deviance    AIC
## - AGE:IBS_POST         1   1013.8 1035.8
## - AGE:DLIT_AG          1   1014.0 1036.0
## - AGE:SIM_GIPERT       1   1014.7 1036.7
## <none>                     1013.5 1037.5
## - AGE:endocr_01        1   1015.9 1037.8
## - as.factor(IBS_POST)  1   1016.2 1038.2
## - endocr_02            1   1018.8 1040.8
##
## Step:  AIC=1035.78
## death ~ AGE + as.factor(IBS_POST) + DLIT_AG + IBS_POST + SIM_GIPERT +
##     endocr_01 + endocr_02 + AGE:DLIT_AG + AGE:SIM_GIPERT + AGE:endocr_01
##
##
## Step:  AIC=1035.78
## death ~ AGE + as.factor(IBS_POST) + DLIT_AG + SIM_GIPERT + endocr_01 +
##     endocr_02 + AGE:DLIT_AG + AGE:SIM_GIPERT + AGE:endocr_01
##
##                       Df Deviance    AIC
## - AGE:DLIT_AG          1   1014.2 1034.2
## - AGE:SIM_GIPERT       1   1015.1 1035.1
## <none>                     1013.8 1035.8
## - AGE:endocr_01        1   1016.2 1036.2
## - endocr_02            1   1019.1 1039.1
## - as.factor(IBS_POST)  2   1029.2 1047.2
##
## Step:  AIC=1034.25
## death ~ AGE + as.factor(IBS_POST) + DLIT_AG + SIM_GIPERT + endocr_01 +
##     endocr_02 + AGE:SIM_GIPERT + AGE:endocr_01
##
##                       Df Deviance    AIC
## - DLIT_AG              1   1015.8 1033.8
## - AGE:SIM_GIPERT       1   1015.8 1033.8
## <none>                     1014.2 1034.2
## - AGE:endocr_01        1   1016.4 1034.4
## - endocr_02            1   1019.4 1037.4
## - as.factor(IBS_POST)  2   1029.5 1045.5
##
## Step:  AIC=1033.77
## death ~ AGE + as.factor(IBS_POST) + SIM_GIPERT + endocr_01 +
##     endocr_02 + AGE:SIM_GIPERT + AGE:endocr_01
##
##                       Df Deviance    AIC
## - AGE:SIM_GIPERT       1   1017.2 1033.2
## <none>                     1015.8 1033.8
## - AGE:endocr_01        1   1017.8 1033.8
## - endocr_02            1   1021.5 1037.5
## - as.factor(IBS_POST)  2   1031.6 1045.6
##
## Step:  AIC=1033.24
## death ~ AGE + as.factor(IBS_POST) + SIM_GIPERT + endocr_01 +
##     endocr_02 + AGE:endocr_01
```

```
##
##                       Df Deviance    AIC
## - AGE:endocr_01        1   1018.9 1032.9
## <none>                     1017.2 1033.2
## - SIM_GIPERT           1   1019.7 1033.7
## - endocr_02            1   1023.3 1037.3
## - as.factor(IBS_POST)  2   1033.0 1045.0
##
## Step:  AIC=1032.91
## death ~ AGE + as.factor(IBS_POST) + SIM_GIPERT + endocr_01 +
##     endocr_02
##
##                       Df Deviance    AIC
## <none>                     1018.9 1032.9
## - SIM_GIPERT           1   1022.0 1034.0
## - endocr_01            1   1024.1 1036.1
## - endocr_02            1   1025.2 1037.2
## - as.factor(IBS_POST)  2   1035.0 1045.0
## - AGE                  1   1071.6 1083.6
```

```
##
## Call:  glm(formula = death ~ AGE + as.factor(IBS_POST) + SIM_GIPERT +
##     endocr_01 + endocr_02, family = binomial, data = data.work2)
##
## Coefficients:
##          (Intercept)                   AGE  as.factor(IBS_POST)1
##              -6.01784               0.05764                0.07336
## as.factor(IBS_POST)2            SIM_GIPERT             endocr_01
##               0.69646               0.72551                0.47597
##            endocr_02
##               1.08069
##
## Degrees of Freedom: 1379 Total (i.e. Null);  1373 Residual
## Null Deviance:       1110
## Residual Deviance: 1019  AIC: 1033
```

```
# fit the best model
death.fit.logit <- glm(death ~ AGE + as.factor(IBS_POST) + as.factor(SIM_GIPERT) + as.factor(endocr_01)
deviance(death.fit.logit) # 1018.906
```

```
## [1] 1018.906
```

```
# Hosmer-Lemeshow test to check goodness of fit
library("ResourceSelection")
death.pred <- predict(death.fit.logit, data.work2, type = "response")
hoslem.test(data.work2$death, death.pred, g = 20) # p = 0.4291, fail to reject H0
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data.work2$death, death.pred
## X-squared = 18.253, df = 18, p-value = 0.4391
```

```
## Get indices of vector fit, from smallest to greatest
fit <- death.fit.logit$fitted.values
index <- sort.list(fit)
## check 10 smallest indices
index[1:10]
```

```
##  [1]  871  751 1038  460  522  448  454  485 1166 1169
```

```
## create a matrix of death and fit, using this index
hosmer <- matrix(c(data.work2$death[index], fit[index]), byrow = F, nrow = nrow(data.work2))
head(hosmer)
```

```
##      [,1]       [,2]
## [1,]    0 0.01078158
## [2,]    0 0.01353970
## [3,]    0 0.01605479
## [4,]    0 0.01630492
## [5,]    0 0.01630492
## [6,]    0 0.01699128
```

```
## group into 20 groups with 69 observations per group
observed <- rep(NA, 20)
for (i in 1:20){ observed[i] <- sum(hosmer[(69*(i-1) +1) : (69 *i), 1])/ 69 }
observed
```

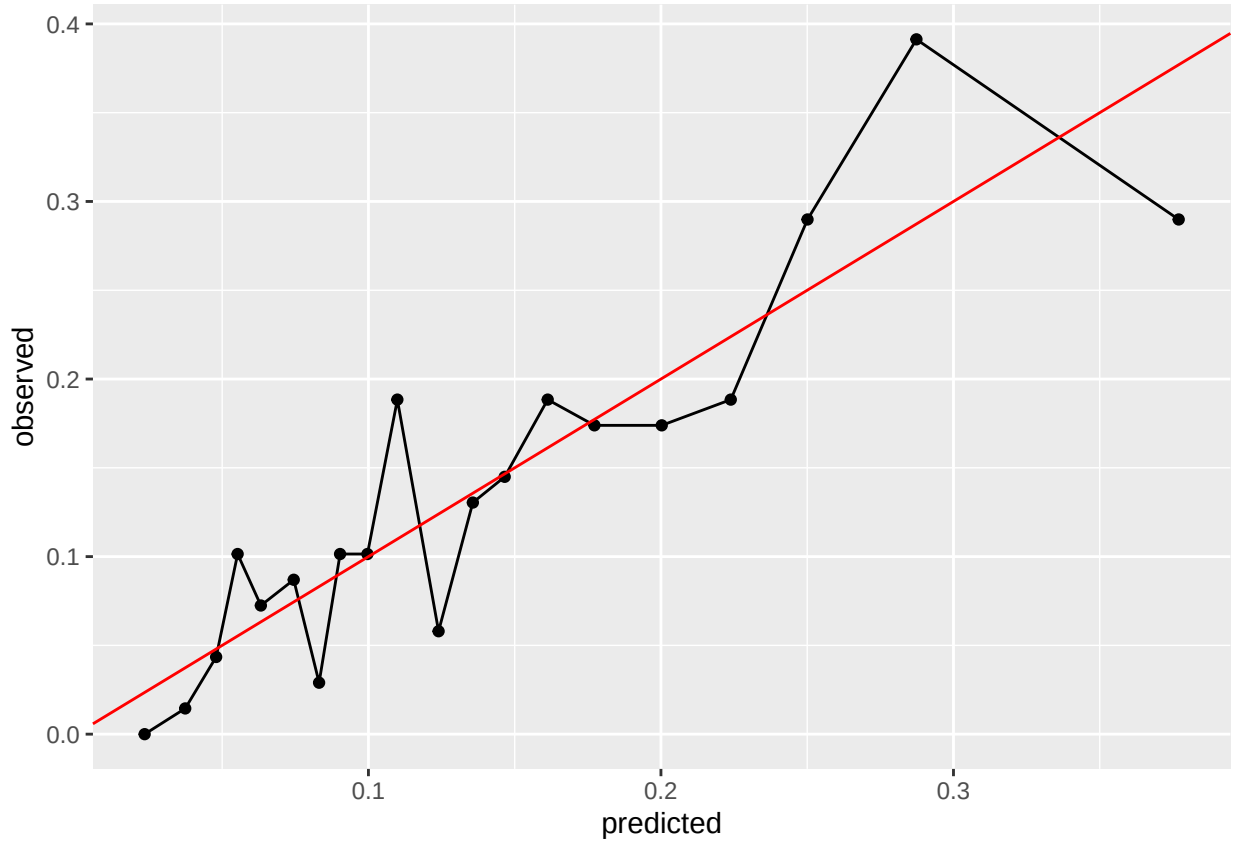
```
##  [1] 0.00000000 0.01449275 0.04347826 0.10144928 0.07246377 0.08695652
##  [7] 0.02898551 0.10144928 0.10144928 0.18840580 0.05797101 0.13043478
## [13] 0.14492754 0.18840580 0.17391304 0.17391304 0.18840580 0.28985507
## [19] 0.39130435 0.28985507
```

```
# repeat the previous step for the predicted probability
predicted <- rep(NA, 20)
for (i in 1:20){ predicted[i] <- sum(hosmer[(69*(i-1) +1) : (69 *i), 2])/ 69 }
predicted
```

```
##  [1] 0.02349284 0.03736897 0.04794239 0.05528165 0.06322126 0.07446455
##  [7] 0.08313026 0.09030576 0.09965501 0.10988757 0.12397682 0.13569291
## [13] 0.14658439 0.16128232 0.17725360 0.20026811 0.22387296 0.25010301
## [19] 0.28733826 0.37699331
```

```
# plot observed versus predicted
ggplot() + aes(x = predicted, y = observed) + geom_point() + geom_line() + geom_abline( a = 0, b = 1, c
```

```
## Warning: Ignoring unknown parameters: a, b
```




```
# model summary
summary(death.fit.logit)
```

```
##
## Call:
## glm(formula = death ~ AGE + as.factor(IBS_POST) + as.factor(SIM_GIPERT) +
##     as.factor(endocr_01) + as.factor(endocr_02), family = binomial,
##     data = data.work2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1636  -0.5915  -0.4345  -0.3086   2.5212
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.017841   0.571648 -10.527  < 2e-16 ***
## AGE                     0.057645   0.008303   6.943 3.84e-12 ***
## as.factor(IBS_POST)1    0.073359   0.249299   0.294  0.76856
## as.factor(IBS_POST)2    0.696463   0.227290   3.064  0.00218 **
## as.factor(SIM_GIPERT)1  0.725514   0.393102   1.846  0.06495 .
## as.factor(endocr_01)1   0.475974   0.202963   2.345  0.01902 *
## as.factor(endocr_02)1   1.080686   0.403462   2.679  0.00739 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1109.7  on 1379  degrees of freedom
## Residual deviance: 1018.9  on 1373  degrees of freedom
## AIC: 1032.9
##
## Number of Fisher Scoring iterations: 5
```

```
# calculate
glm.predict <- ifelse(predict(death.fit.logit, data.work2, type = "response") > 0.5, 1, 0)
sum(diag(table(glm.predict, data.work2$ZSN))) / nrow(data.work2) # 0.7616
```

```
## [1] 0.7615942
```

The final model fitted: $log\frac{\pi_i}{1-\pi_i} = -6.018 + 0.058 \times age + 0.073 \times I(IBS = 1) + 0.696 \times I(IBS = 2) + 0.726 \times I(SIM = 1) + 0.476 \times I(endocr01 = 1) + 1.081 \times I(endocr02 = 1)$. Hosmer Lemeshow tests shows adequate goodness of fit (p = 0.4291).

Fit logistic regression with multinomial response

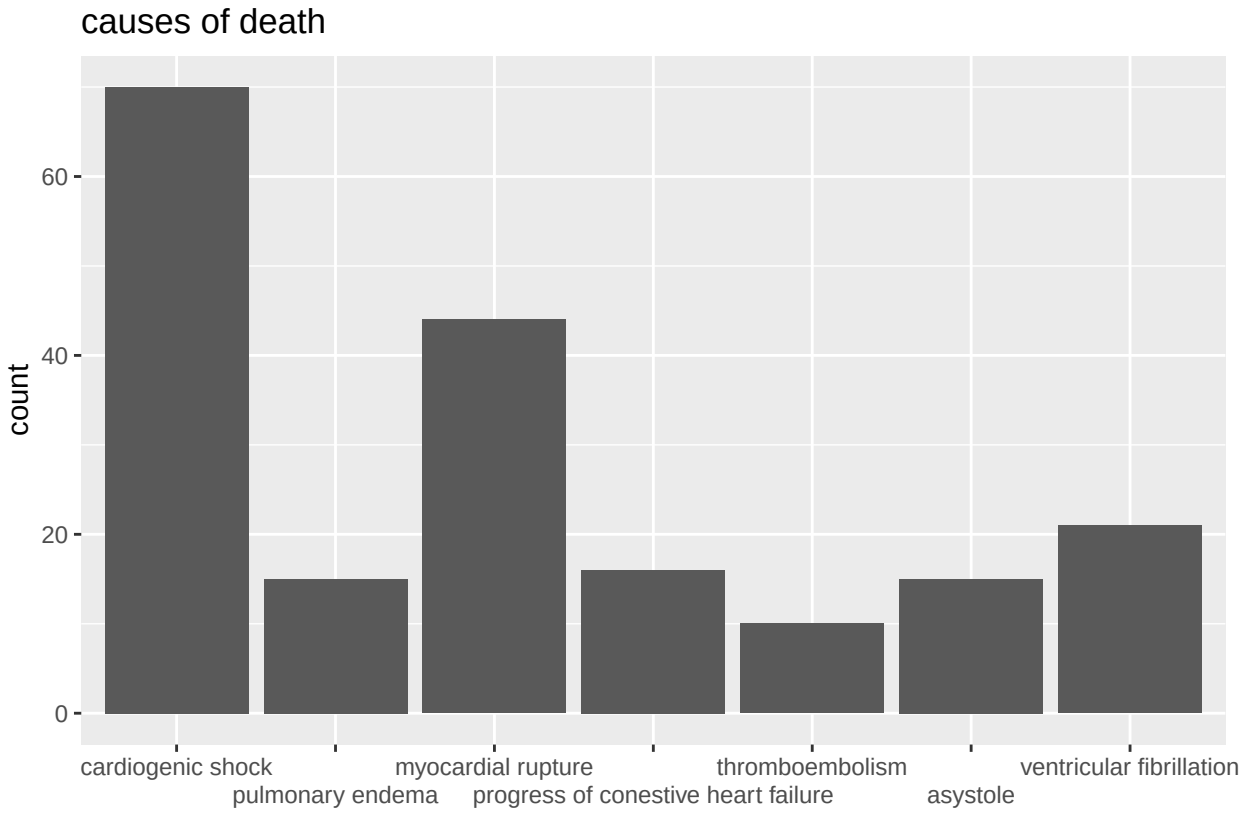
```
library(nnet)
data.work3 <- filter(data.work2, LET_IS != 0)
dim(data.work3) # n = 191
```

```
## [1] 191  15
```

```
table(data.work3$LET_IS)
```

```
##
##  1  2  3  4  5  6  7
## 70 15 44 16 10 15 21
```

```
ggplot(data.work3, aes(as.factor(data.work3$LET_IS))) + geom_bar() + labs(title = "causes of death") + 
```




```
multinom(LET_IS ~ AGE + as.factor(IBS_POST) + as.factor(SIM_GIPERT) + as.factor(endocr_01) + as.factor(
```

```
## # weights:  56 (42 variable)
## initial  value 371.668838
## iter  10 value 312.126386
## iter  20 value 300.807784
## iter  30 value 300.010607
## iter  40 value 299.933289
## iter  50 value 299.931699
## final  value 299.931683
## converged
```

```
## Call:
## multinom(formula = LET_IS ~ AGE + as.factor(IBS_POST) + as.factor(SIM_GIPERT) +
##     as.factor(endocr_01) + as.factor(endocr_02), data = data.work3)
##
## Coefficients:
##   (Intercept)         AGE as.factor(IBS_POST)1 as.factor(IBS_POST)2
## 2   -5.208477  0.05003237            0.4172287           -0.2605801
## 3   -2.662446  0.04515371           -0.8725386           -1.3250667
## 4   -3.189649  0.02970654           -0.3969242           -0.7216065
## 5    1.046965 -0.03766681           -0.2262002           -1.5074724
## 6   -2.551705  0.03088585           -2.0391936           -1.3081214
## 7    2.872844 -0.05676433           -0.5333366           -0.2875964
##   as.factor(SIM_GIPERT)1 as.factor(endocr_01)1 as.factor(endocr_02)1
```

```
## 2          -14.880228580            1.3482127          -14.0286129
## 3           -0.012590414            0.2372437            0.6681173
## 4            0.004074119            1.1727607          -15.2277093
## 5          -16.165192169            1.5726899          -16.3110403
## 6          -16.883580258            0.8243099            0.7648381
## 7            0.257029070           -0.7257783          -15.7083009
##
## Residual Deviance: 599.8634
## AIC: 683.8634
```