Team 1:

Kaiyu Wang,Chinar,Urvashi,Chun,,

I.Setup

```
#install.packages('readr', dependencies = TRUE, repos='http://cran.rstudio.com/')
library(readr)
library(data.table)
library(ggplot2)
library(dplyr)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-2
```

```
library(ROCR)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(PRROC)
library(lattice)
library(caret)
library(e1071)
CHD <-fread("framingham.csv")
```

II.Clean Data

1. Summary

```
summary(CHD)
```

```
##       male             age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.003   Mean   :0.02963   Mean   :0.005899   Mean   :0.3105
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.00
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.00
##  Mean   :0.02572   Mean   :236.7   Mean   :132.4   Mean   : 82.89
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 89.88
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.50
##                    NA's   :50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.000
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.000
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.000
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.97   Mean   :0.152
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.000
##  NA's   :19      NA's   :1        NA's   :388
```

2. Replace NA

```
education_median<-median(CHD$education,na.rm=TRUE)
CHD[is.na(education),education:=education_median]

cigsPerDay_median<-median(CHD$cigsPerDay,na.rm=TRUE)
CHD[is.na(cigsPerDay),cigsPerDay:=cigsPerDay_median]

BPMeds_median<-median(CHD$BPMeds,na.rm=TRUE)
CHD[is.na(BPMeds),BPMeds:=BPMeds_median]

totChol_median<-median(CHD$totChol,na.rm=TRUE)
CHD[is.na(totChol),totChol:=totChol_median]

glucose_median<-median(CHD$glucose,na.rm=TRUE)
CHD[is.na(glucose),glucose:=glucose_median]

heartRate_median<-median(CHD$heartRate,na.rm=TRUE)
CHD[is.na(heartRate),heartRate:=heartRate_median]

BMI_median<-median(CHD$BMI,na.rm=TRUE)
CHD[is.na(BMI),BMI:=BMI_median]
```

```
colnames(CHD)[1] <- 'is_male'
```

III.EDA

1. Distribution of Ten Year Risk of CHD

```
count1 <- length(which(CHD$TenYearCHD == 1))
count1
```

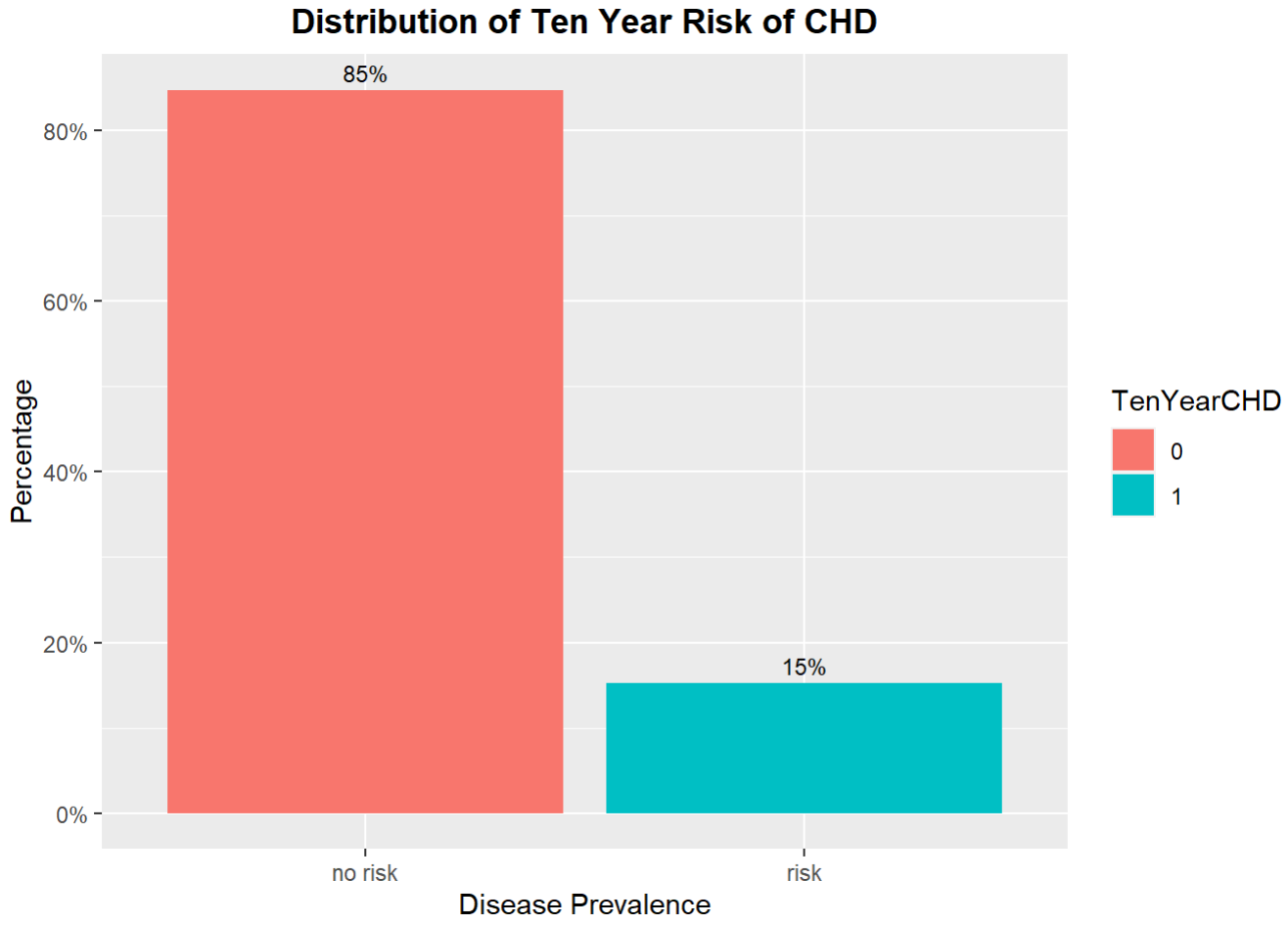
```
## [1] 644
```

```
count2 <- length(which(CHD$TenYearCHD == 0))
count2
```

```
## [1] 3594
```

```
common_theme <- theme(plot.title = element_text(hjust = 0.5, face = "bold"))

ggplot(data = CHD, aes(x = factor(TenYearCHD),
                       y = prop.table(stat(count)),
                       fill = factor(TenYearCHD),
                       label = scales::percent(prop.table(stat(count))))) +
    geom_bar(position = "dodge") +
    geom_text(stat = 'count',
              position = position_dodge(.9),
              vjust = -0.5,
              size = 3) +
    scale_x_discrete(labels = c("no risk", "risk"))+
    scale_y_continuous(labels = scales::percent)+
    labs(x = 'Disease Prevalence', y = 'Percentage',fill='TenYearCHD') +
    ggtitle("Distribution of Ten Year Risk of CHD") +
    common_theme
```



2. Distribution of Percentage of CHD with Age

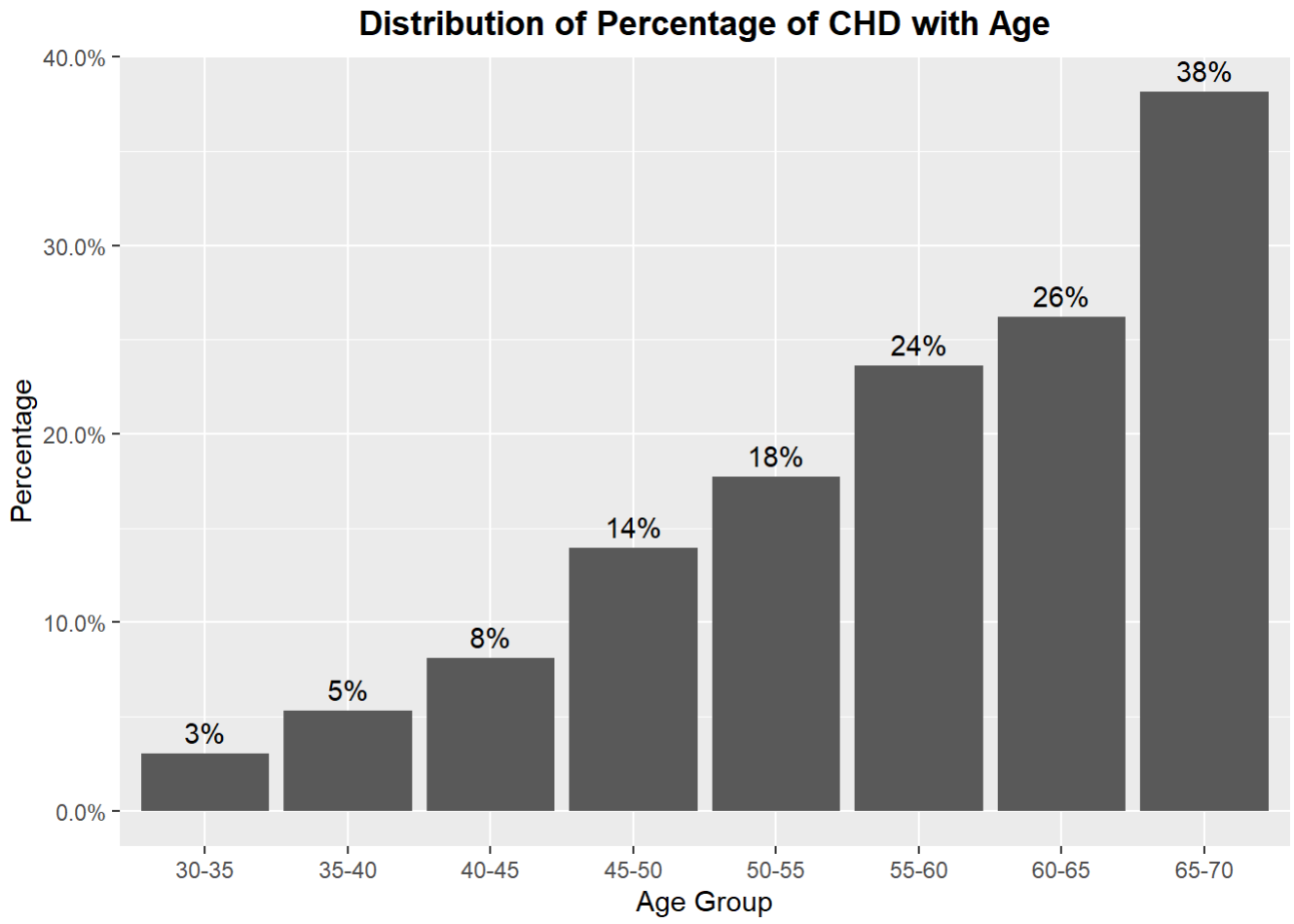
```
CHD$agec <-
  cut(CHD$age, breaks = c(30,35,40,45,50,55,60,65,70),
      labels = c("30-35","35-40","40-45","45-50","50-55","55-60","60-65","65-70"))

d <- CHD %>% group_by(agec) %>% summarise(perc = mean(TenYearCHD=='1'))
d$perc_r <- round(d$perc,2)*100
d$perc_r <- interaction(d$perc_r, "%", sep = "")
d
```

agec <fct>	perc <dbl>	perc_r <fct>
30-35	0.03030303	3%
35-40	0.05294118	5%
40-45	0.08085612	8%
45-50	0.13932292	14%
50-55	0.17721519	18%
55-60	0.23608769	24%
60-65	0.26226013	26%
65-70	0.38181818	38%

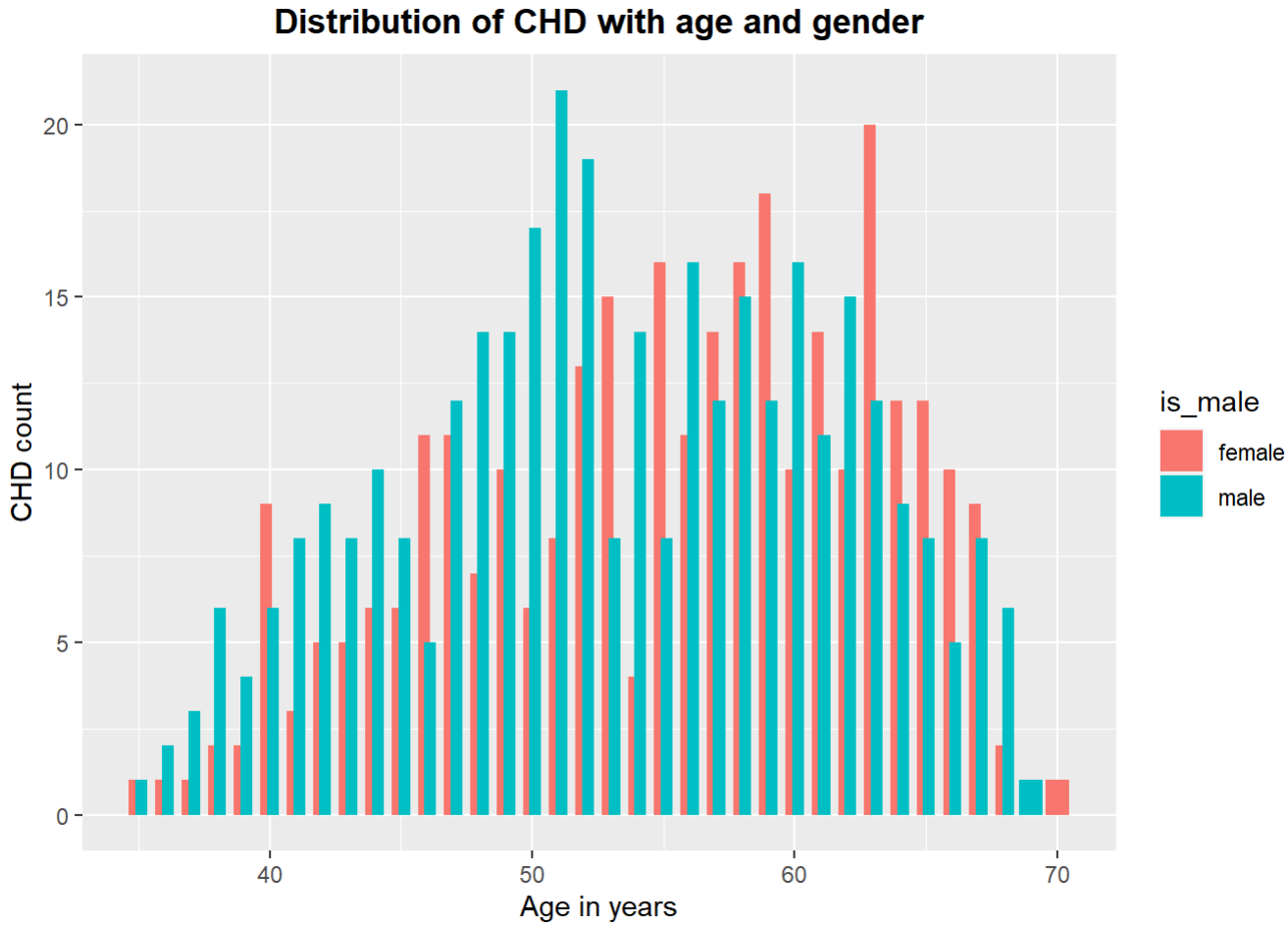
8 rows

```
ggplot(d,aes(x=agec,y=perc)) +
geom_col()+
scale_y_continuous(labels=scales::percent)+
geom_text(aes(label = perc_r), vjust = -0.5)+
labs(x='Age Group',y='Percentage')+
ggtitle("Distribution of Percentage of CHD with Age")+
common_theme
```



3. Histogram of CHD with age and gender

```
#cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73", #"#F0E442", "#0072B2", "#D55E00", "#
CC79A7")
CHD_1 <- CHD[ CHD$TenYearCHD=='1',]
CHD_1$is_male[CHD_1$is_male == 0] <- "female"
CHD_1$is_male[CHD_1$is_male == 1] <- "male"
ggplot(data=CHD_1,aes(age,fill=is_male))+
  geom_bar(position = position_dodge(width = 0.5))+
#  scale_fill_brewer(palette=cbPalette)+
  labs(x = "Age in years",y = "CHD count")+
  ggtitle("Distribution of CHD with age and gender")+
  common_theme
```



4. Probability of disease in smokers

```
d2 <- CHD %>% group_by(currentSmoker) %>% summarise(perc = mean(TenYearCHD=='1'))
d2
```

currentSmoker <int>	perc <dbl>
0	0.1450560
1	0.1590258

2 rows

5. Line Chart of Percentage of CHD with Age and Gender

```
d3 <- CHD %>% group_by(agec,factor(is_male)) %>% summarise(perc = mean(TenYearCHD=='1'))
```

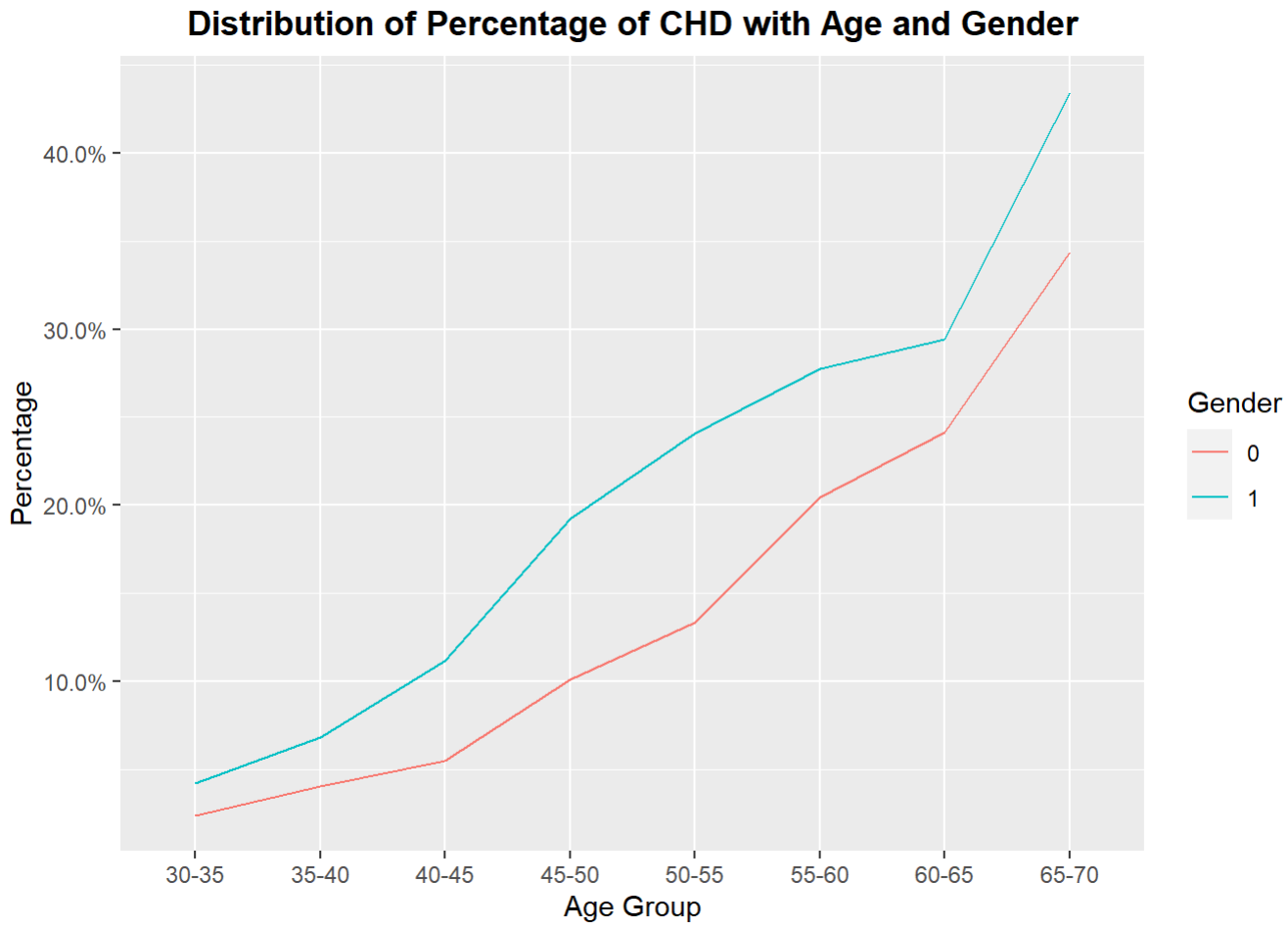
```
## `summarise()` has grouped output by 'agec'. You can override using the `.groups` argument.
```

```
d3
```

agec <fct>	factor(is_male) <fct>	perc <dbl>
30-35	0	0.02380952
30-35	1	0.04166667
35-40	0	0.04032258
35-40	1	0.06818182
40-45	0	0.05482456
40-45	1	0.11168831
45-50	0	0.10089686
45-50	1	0.19254658
50-55	0	0.13333333
50-55	1	0.24054983

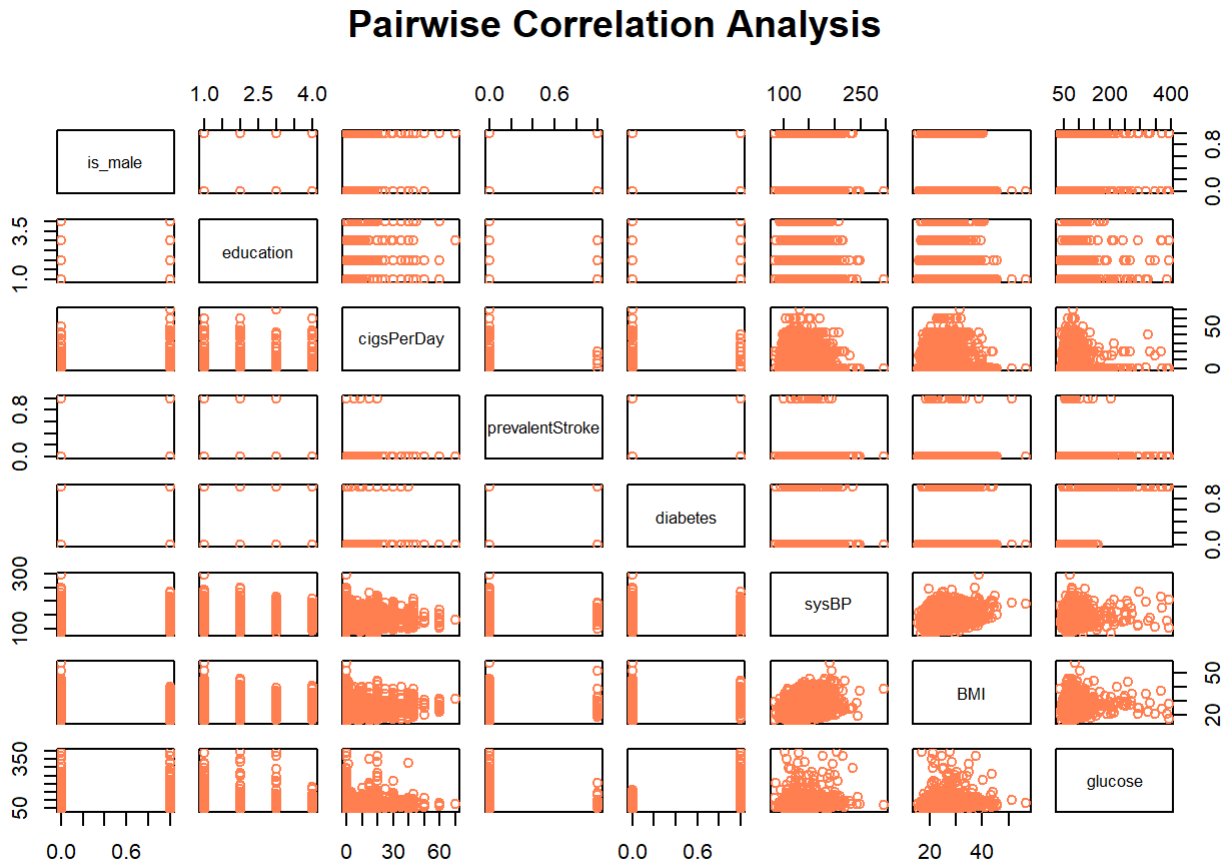
1-10 of 16 rows Previous **1** 2 Next

```
ggplot() +
geom_line(data=d3,aes(agec, perc,group =`factor(is_male)`,color =`factor(is_male)` ))+
scale_y_continuous(labels=scales::percent)+
labs(x='Age Group',y='Percentage',color='Gender' )+
ggtitle("Distribution of Percentage of CHD with Age and Gender")+
common_theme
```



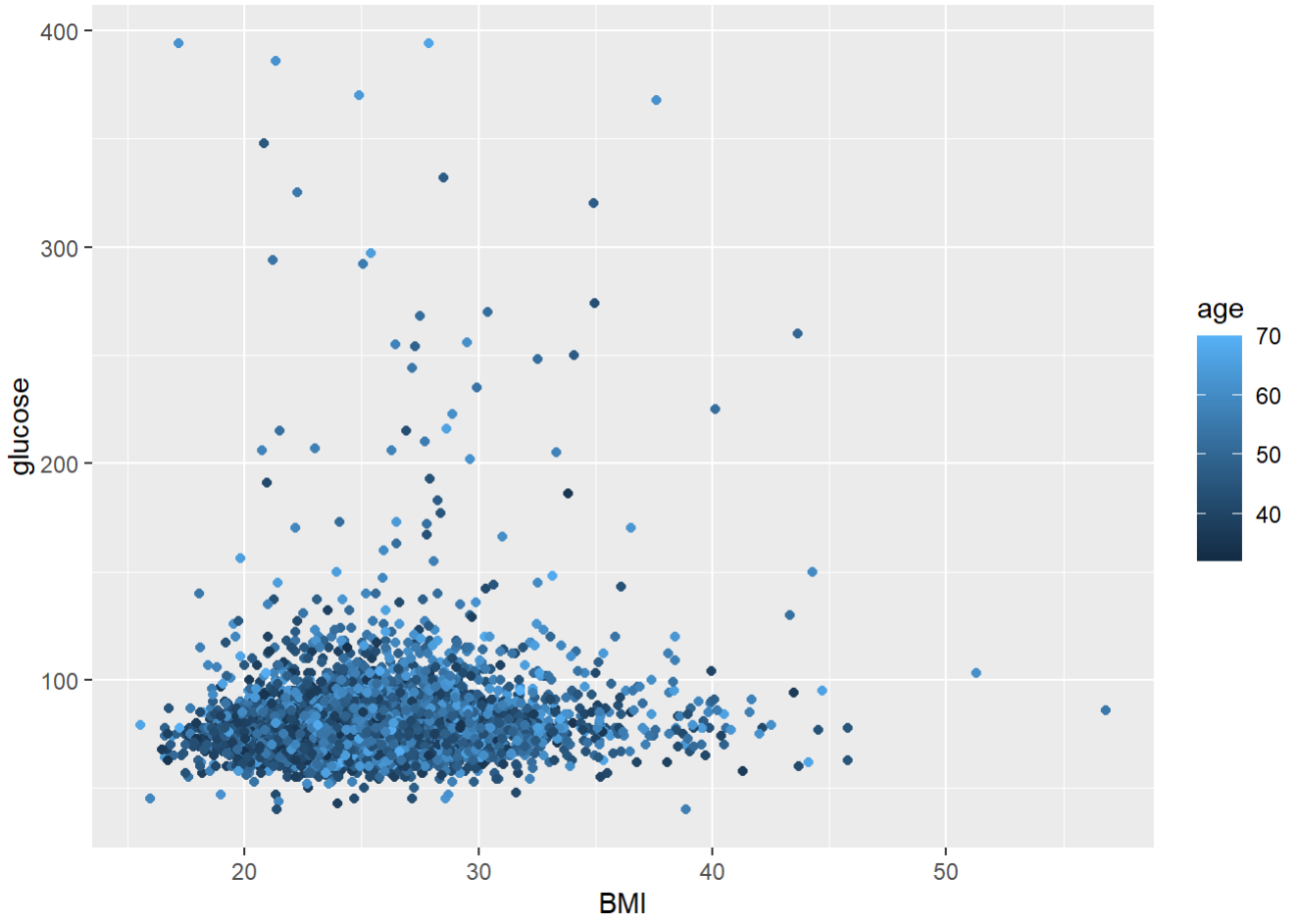
6. Pairwise Correlation Analysis

```
a <- CHD[,c(1,3,5,7,9,11,13,15)]
pairs(a, col = "coral",main = "Pairwise Correlation Analysis")
```



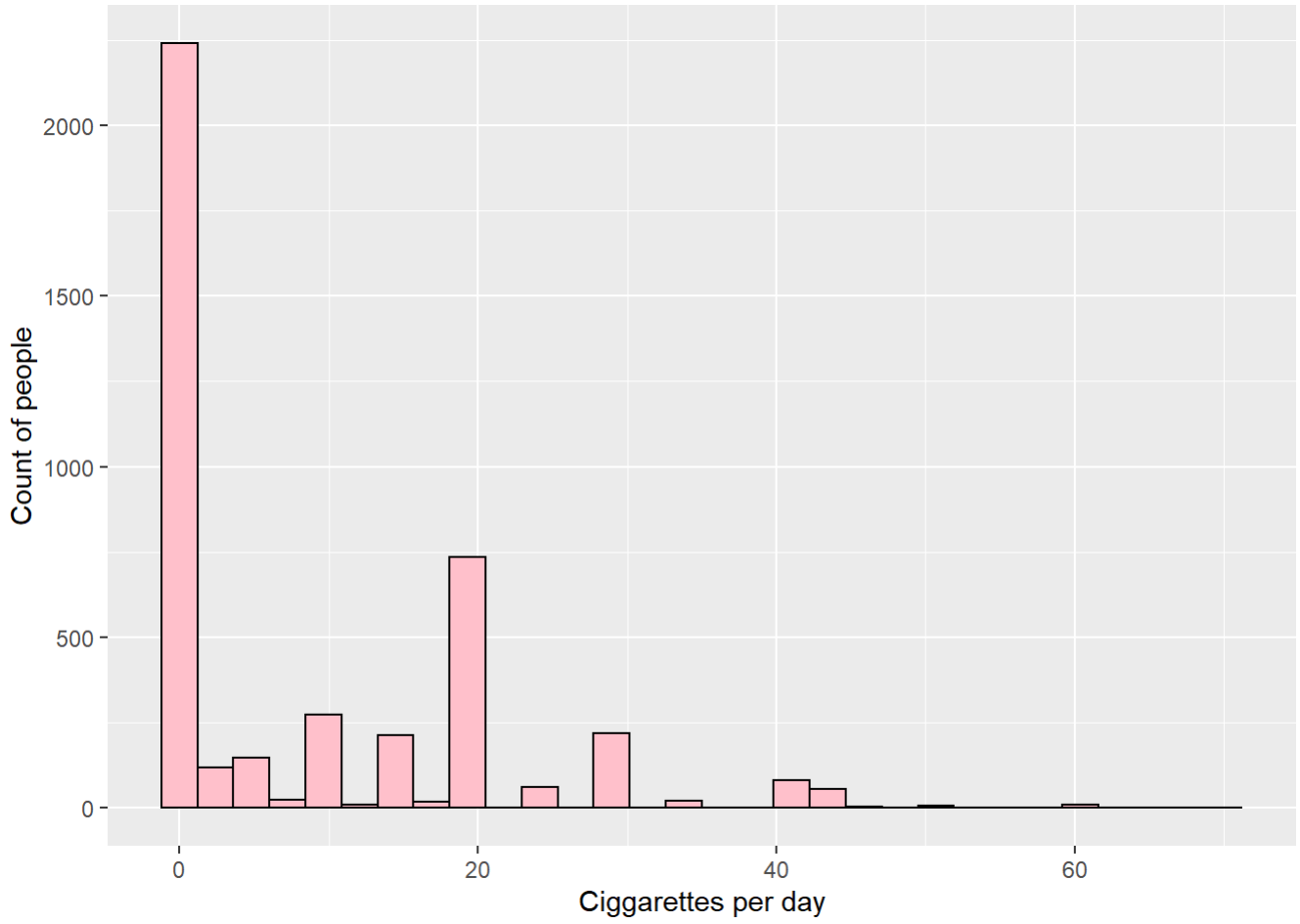
7. ???

```
ggplot(data = CHD, aes(BMI,glucose,color = age)) + geom_point(fill = "blue")
```




```
ggplot(data = CHD, aes(x = cigsPerDay, color = education)) + geom_histogram(color="black", fill=
"pink")+labs(x='Ciggarettes per day', y = 'Count of people')
```

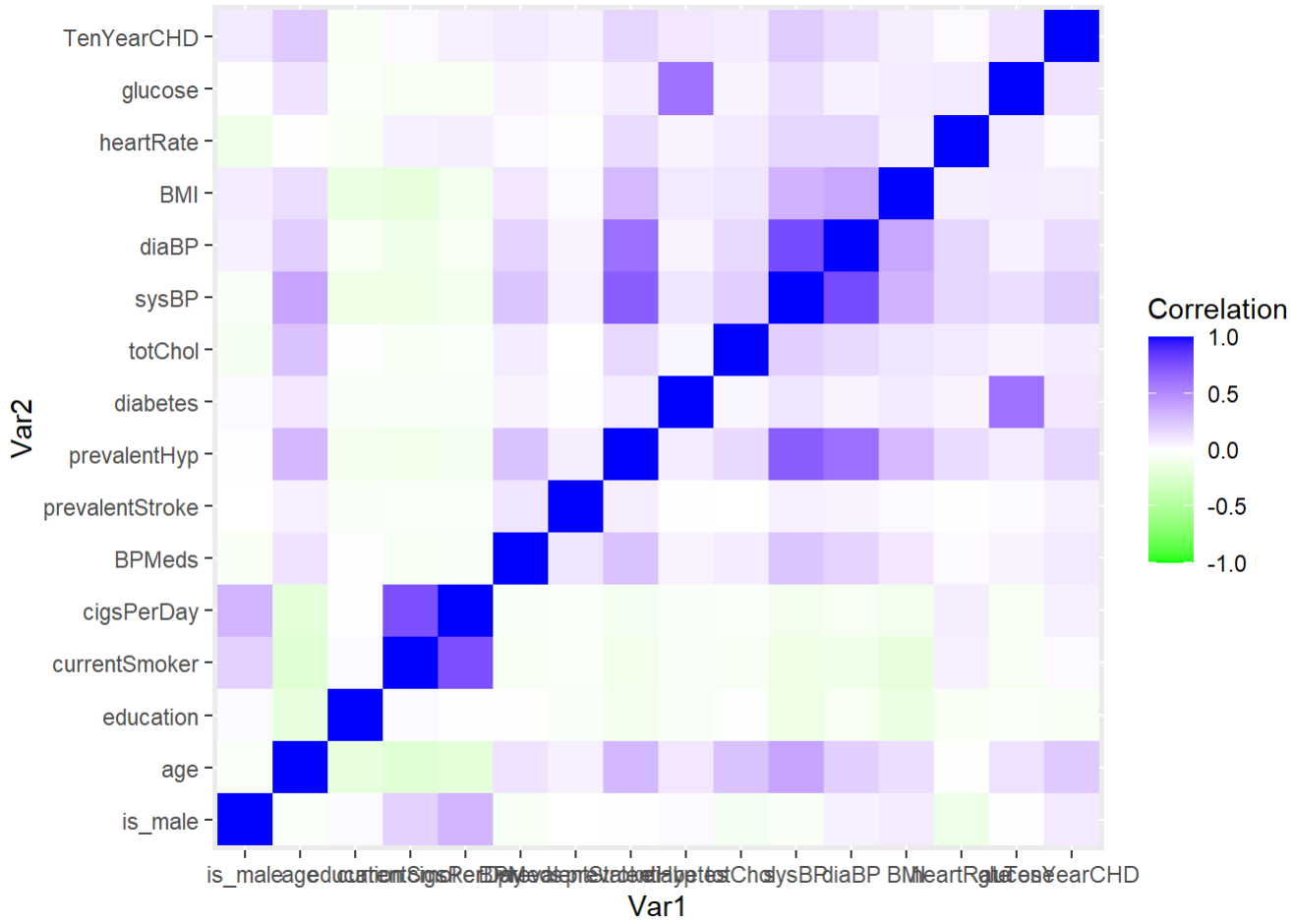
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#+  + geom_vline(aes(intercept=mean(cigsPerDay)), color="blue", linetype="dashed", size=12)
```

8. Corelation Heatmap

```
CHD<-subset(CHD,select=-c(17))
cormat <- round(cor(CHD),2)
melted_cormat <- melt(cormat)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
common_theme+
geom_tile()+
scale_fill_gradient2(low = "green", high = "blue",
   midpoint = 0, limit = c(-1,1),
   name="Correlation")
```



IV.Mechine Learning

1. Split Data

```
set.seed(1)
CHD_0<-CHD[CHD$TenYearCHD==0]
dt = sort(sample(nrow(CHD_0), nrow(CHD_0)*.8))
train_0<-CHD_0[dt,]
test_0<-CHD_0[-dt,]
CHD_1<-CHD[CHD$TenYearCHD==1]
dt = sort(sample(nrow(CHD_1), nrow(CHD_1)*.8))
train_1<-CHD_1[dt,]
test_1<-CHD_1[-dt,]
train<-rbind(train_0, train_1)
test<-rbind(test_0, test_1)
```

2. Lasso Regression

```
# Create formula
formula <- as.formula(TenYearCHD ~ .)

# Training set modeling
train.matrix <- model.matrix(formula, train)[, -1]
train_y <- train$TenYearCHD
fit <- cv.glmnet(train.matrix, train_y, family = "binomial", alpha = 1, nfolds = 10)

# Create testing matrices
test.matrix <- model.matrix(formula, test) [, -1]
```

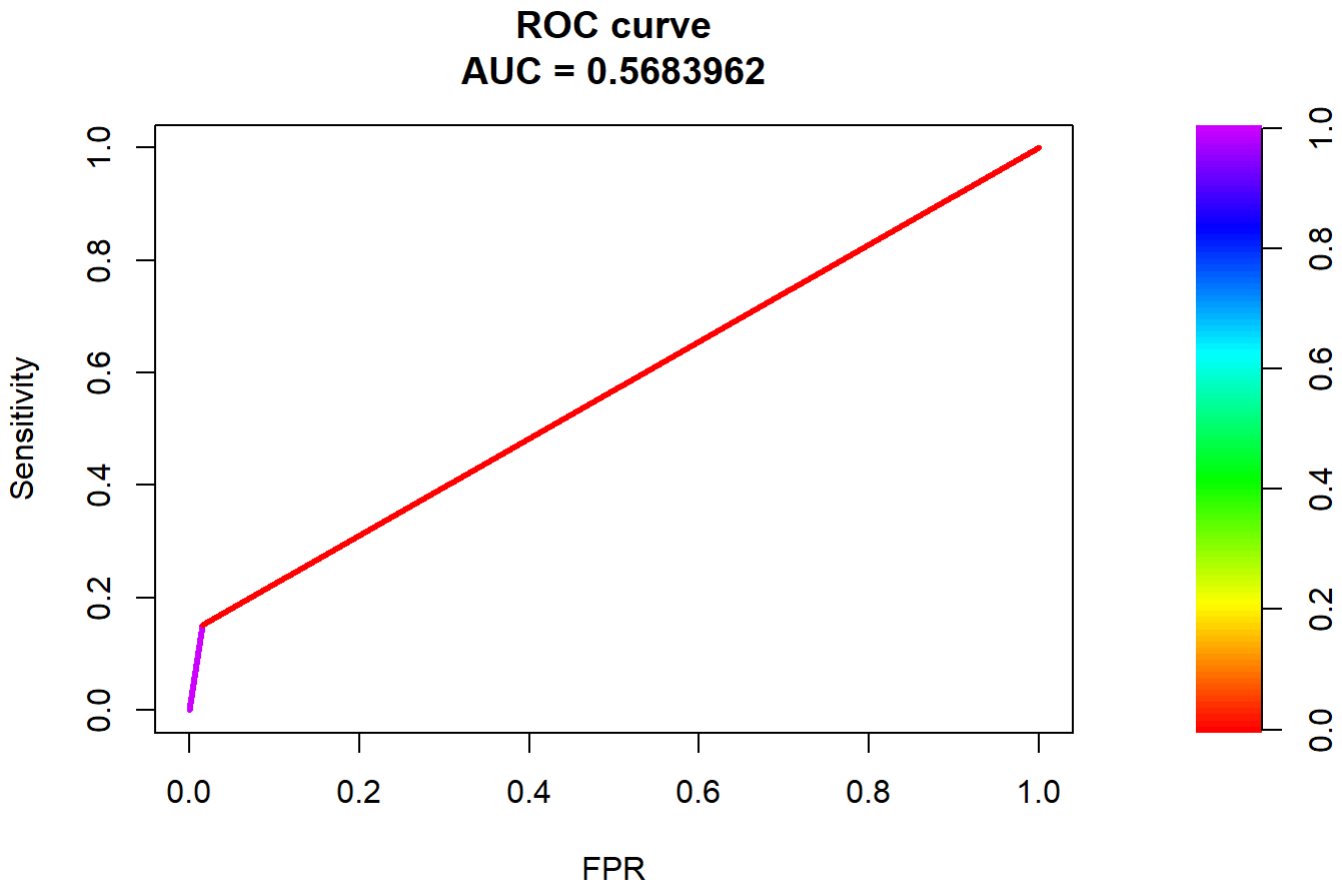
```
coef(fit,s=fit$lambda.min)
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                            s1
## (Intercept)     -7.962164835
## is_male          0.435117371
## age              0.059885184
## education        .
## currentSmoker    .
## cigsPerDay       0.018569190
## BPMeds           0.175797916
## prevalentStroke  0.414269086
## prevalentHyp     0.167806650
## diabetes         0.286122249
## totChol          0.001212990
## sysBP            0.014261096
## diaBP            .
## BMI              .
## heartRate        .
## glucose          0.005611137
```

```
# Predicting test data
test.predictions <- predict(fit, test.matrix, s = fit$lambda.min)
predicted.CHD <- ifelse(test.predictions > 0, 1, 0)
confusionMatrix(as.factor(predicted.CHD), as.factor(test$TenYearCHD), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 715 120
##          1   4   9
##
##                Accuracy : 0.8538
##                  95% CI : (0.8282, 0.8769)
##     No Information Rate : 0.8479
##     P-Value [Acc > NIR] : 0.3368
##
##                   Kappa : 0.1017
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.06977
##             Specificity : 0.99444
##          Pos Pred Value : 0.69231
##          Neg Pred Value : 0.85629
##              Prevalence : 0.15212
##          Detection Rate : 0.01061
##    Detection Prevalence : 0.01533
##       Balanced Accuracy : 0.53210
##
##        'Positive' Class : 1
##
```

```
c<-roc.curve(as.numeric(test$TenYearCHD), as.numeric(predicted.CHD), curve = TRUE)
plot(c)
```


ROC curve
AUC = 0.5683962
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
FPR
0.0
0.2
0.4
0.6
0.8
1.0