Team 1:

Kaiyu Wang,Chinar,,,,

Setup

```
install.packages('readr', dependencies = TRUE, repos='http://cran.rstudio.com/')
```

```
## package 'readr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\wky\AppData\Local\Temp\Rtmp4KvxjM\downloaded_packages
```

```
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
```

```
CHD_raw <-read_csv("framingham.csv")
```

```
## Rows: 4238 Columns: 16
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (16): male, age, education, currentSmoker, cigsPerDay, BPMeds, prevalent...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Clean Data

```
CHD<-na.omit(CHD_raw)
colnames(CHD)[1] <- 'is_male'
```

EDA

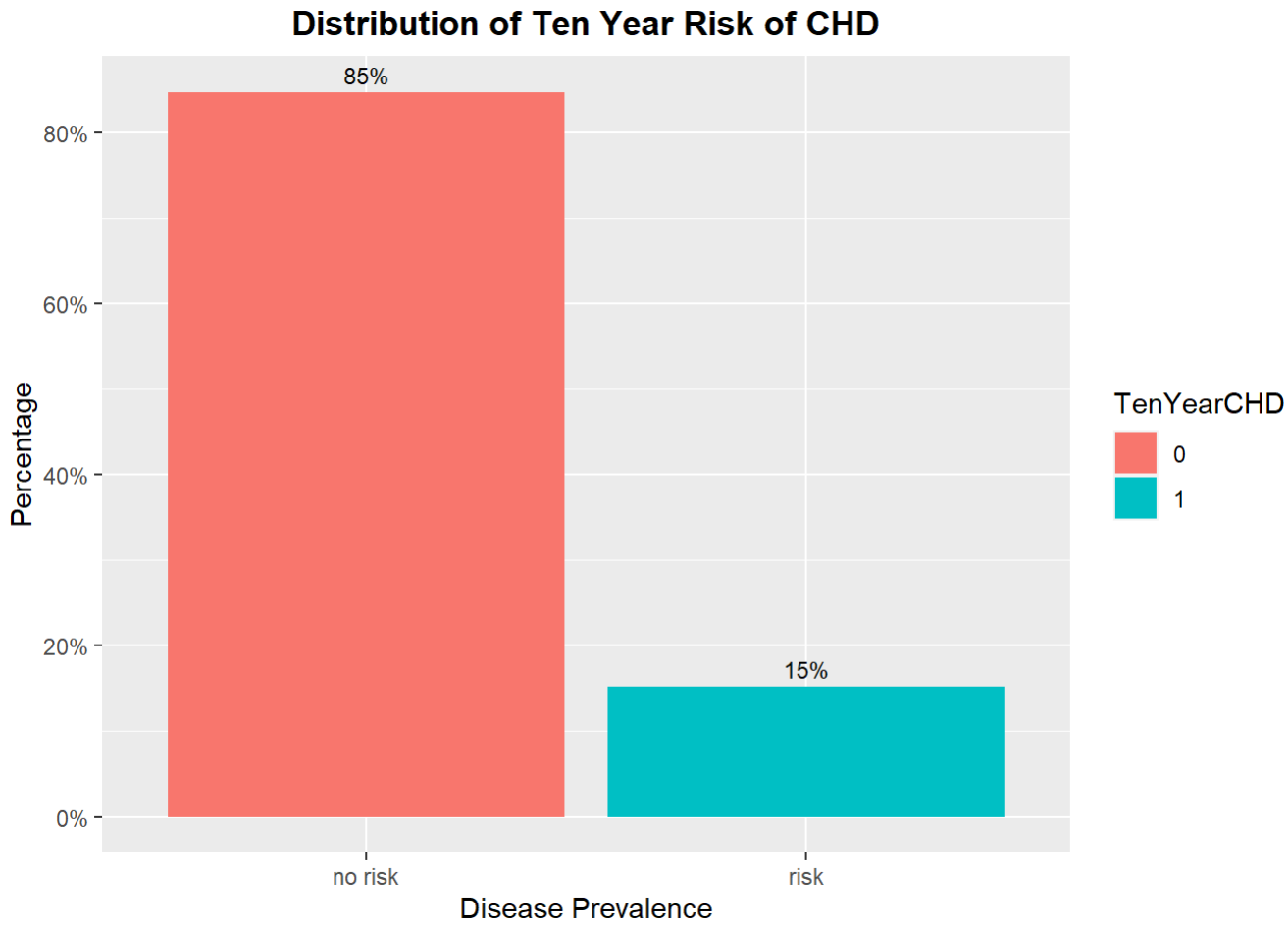
```
count1 <- length(which(CHD$TenYearCHD == 1))
count1
```

```
## [1] 557
```

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

```
## [1] 3099
```

```
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
```




```
#summary(CHD)
sapply(CHD, class)
```

```
##         is_male             age       education   currentSmoker      cigsPerDay
##       "numeric"       "numeric"       "numeric"       "numeric"       "numeric"
##          BPMeds prevalentStroke    prevalentHyp        diabetes         totChol
##       "numeric"       "numeric"       "numeric"       "numeric"       "numeric"
##           sysBP           diaBP             BMI       heartRate         glucose
##       "numeric"       "numeric"       "numeric"       "numeric"       "numeric"
##      TenYearCHD
##       "numeric"
```

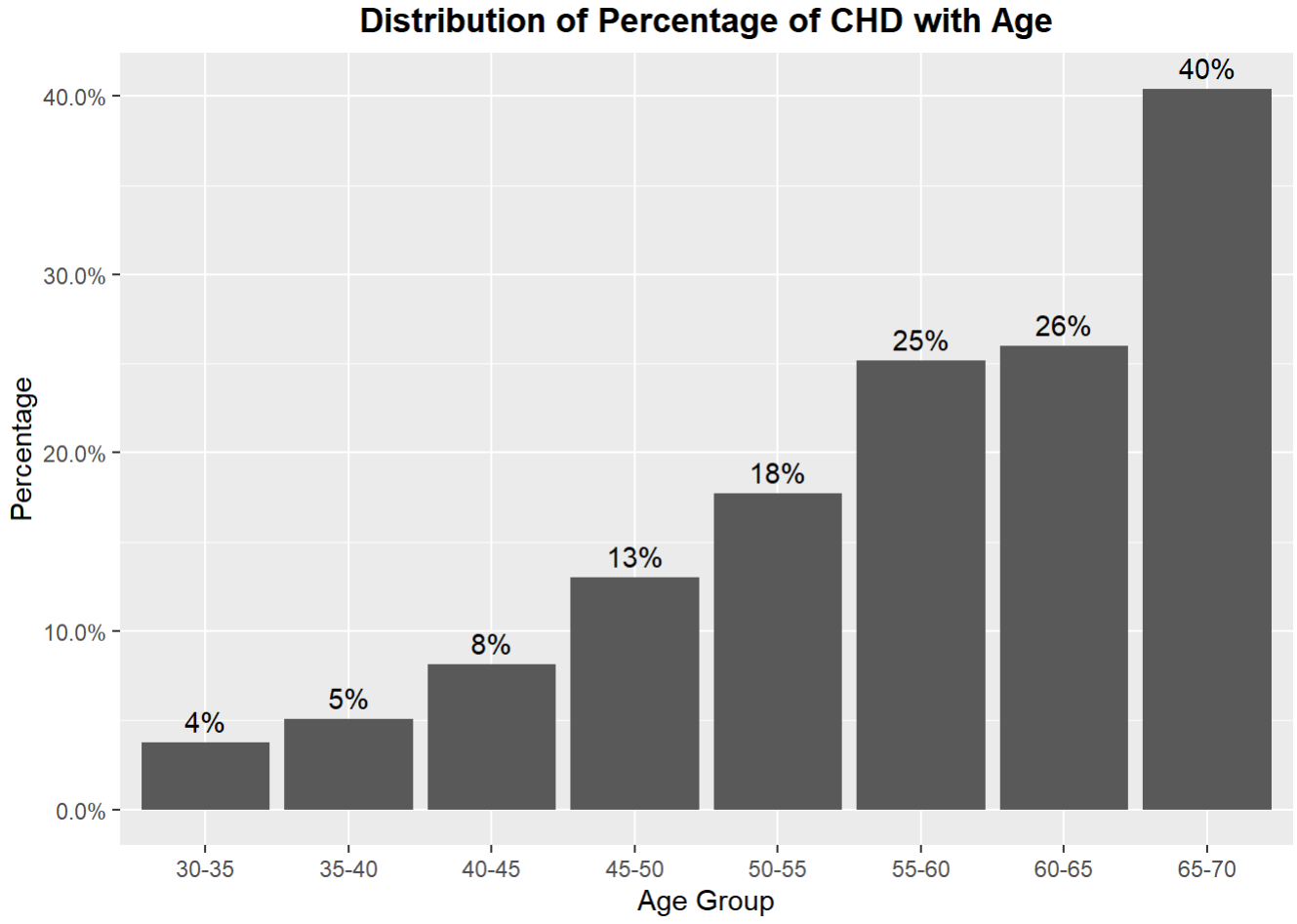
```
CHD$agec <-
  cut(CHD$age, breaks = c(30,35,40,45,50,55,60,65,70),
      labels = c("30-35","35-40","40-45","45-50","50-55","55-60","60-65","65-70"))

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	perc	perc_r
<fct>	<dbl>	<fct>

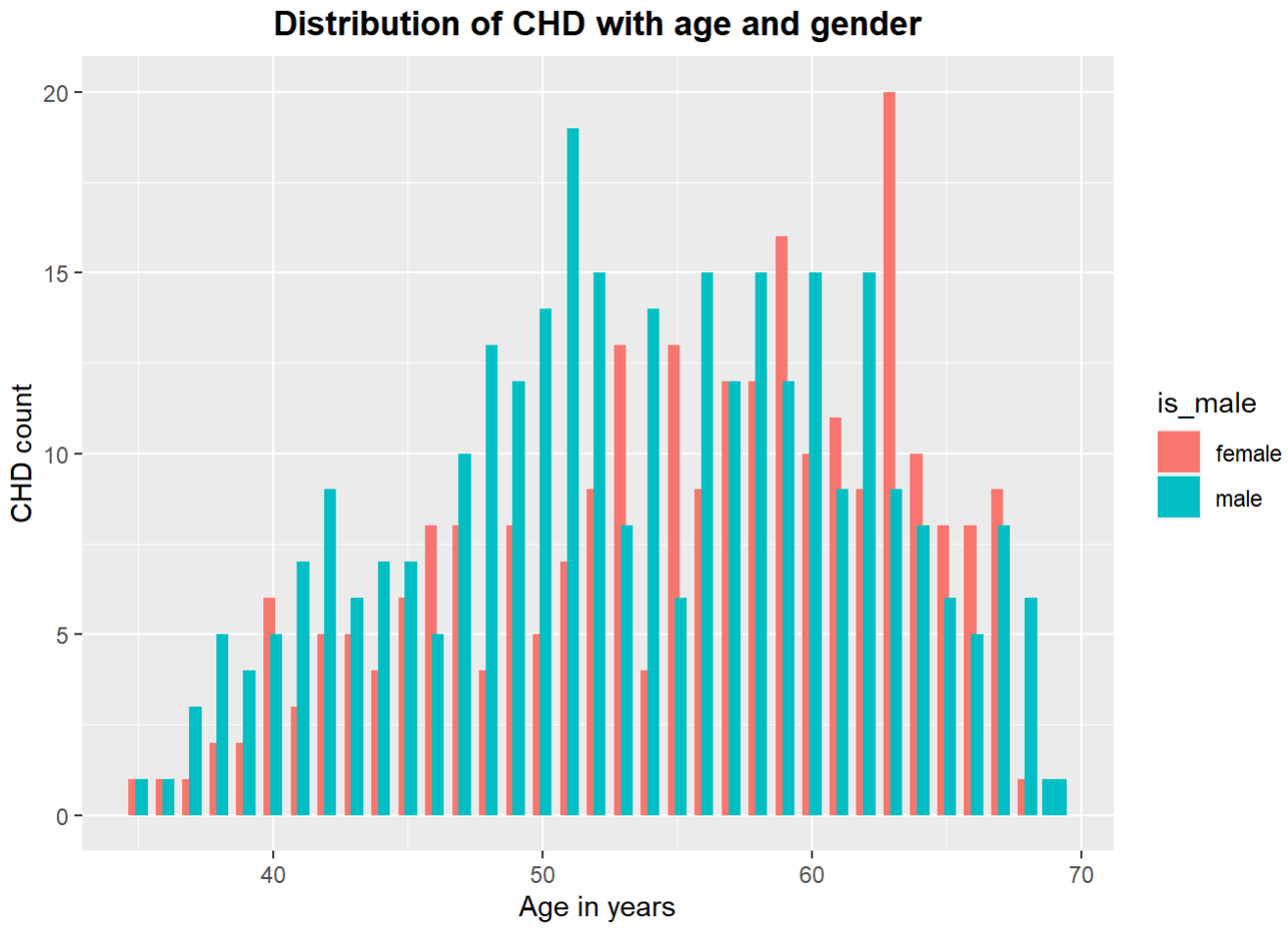
agec <fct>	perc <dbl>	perc_r <fct>
30-35	0.03773585	4%
35-40	0.05059022	5%
40-45	0.08126722	8%
45-50	0.13023952	13%
50-55	0.17704918	18%
55-60	0.25196850	25%
60-65	0.25990099	26%
65-70	0.40425532	40%

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
```



```
#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
```




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

currentSmoker <dbl>	perc <dbl>
0	0.1456103
1	0.1593960

2 rows

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

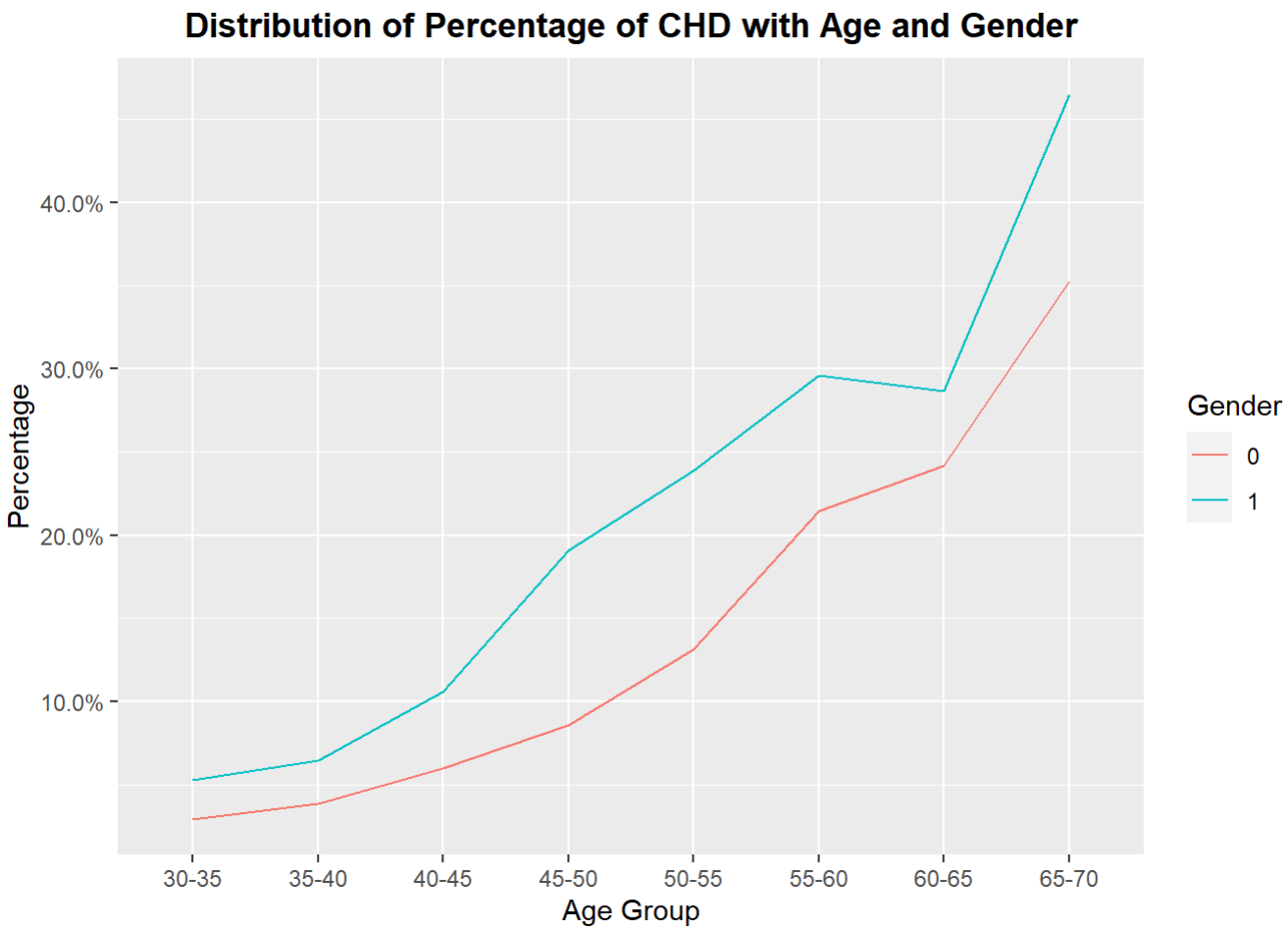
```
## `summarise()` has grouped output by 'agec'. You can override using the `.groups` argument.
```

```
d3
```

agec <fct>	factor(is_male) <fct>	perc <dbl>
30-35	0	0.02941176
30-35	1	0.05263158
35-40	0	0.03833866
35-40	1	0.06428571
40-45	0	0.05958549
40-45	1	0.10588235
45-50	0	0.08571429
45-50	1	0.19081272
50-55	0	0.13142857
50-55	1	0.23846154

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
```


Distribution of Percentage of CHD with Age and Gender
40.0%
30.0%
20.0%
10.0%
Percentage
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
Age Group
Gender
0
1