131 Project

Nicole Gibson

12/6/2020

R Markdown

```r
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
```

```r
library(zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
d_2003 <- read.delim("Deaths Dataset 2003.txt", header = TRUE)
d_2004 <- read.delim("Deaths Dataset 2004.txt", header = TRUE)
d_2005 <- read.delim("Deaths Dataset 2005.txt", header = TRUE)
d_2006 <- read.delim("Deaths Dataset 2006.txt", header = TRUE)
d_2007 <- read.delim("Deaths Dataset 2007.txt", header = TRUE)
d_2008 <- read.delim("Deaths Dataset 2008.txt", header = TRUE)
death_data <- rbind(d_2003, d_2004, d_2005, d_2006, d_2007, d_2008)

death_data$Month <- gsub("., ", "_", death_data$Month)
death_data$Month <- (sapply (death_data$Month, function (x) gsub ("\"", "", x)))

death_data$date <- as.yearmon(death_data$Month, "%b_%Y")
death_data$date <- as.Date(death_data$date)
class(death_data$date)
```

```
## [1] "Date"
```

```r
summary(death_data)
```

```
##     Notes              Race             Race.Code            Gender
##  Length:61888       Length:61888       Length:61888       Length:61888
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Gender.Code             Year          Year.Code         Month
##  Length:61888       Min.   :2003    Min.   :2003     Length:61888
##  Class :character   1st Qu.:2004    1st Qu.:2004     Class :character
##  Mode  :character   Median :2006    Median :2006     Mode  :character
##                     Mean   :2006    Mean   :2006
##                     3rd Qu.:2007    3rd Qu.:2007
##                     Max.   :2008    Max.   :2008
##                     NA's   :291     NA's   :291
##   Month.Code        UCD...ICD.10.113.Cause.List
##  Length:61888       Length:61888
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
##  UCD...ICD.10.113.Cause.List.Code     Deaths          Population
##  Length:61888                      Min.   :    1.0   Length:61888
##  Class :character                  1st Qu.:    6.0   Class :character
##  Mode  :character                  Median :   30.0   Mode  :character
##                                    Mean   :  585.5
##                                    3rd Qu.:  208.0
##                                    Max.   :42076.0
##                                    NA's   :291
##   Crude.Rate        Crude.Rate.Lower.95..Confidence.Interval
##  Length:61888       Length:61888
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
##  Crude.Rate.Upper.95..Confidence.Interval      date
##  Length:61888                              Min.   :2003-01-01
##  Class :character                          1st Qu.:2004-07-01
##  Mode  :character                          Median :2006-01-01
##                                            Mean   :2005-12-21
##                                            3rd Qu.:2007-07-01
##                                            Max.   :2008-12-01
##                                            NA's   :5425
```

```r
unique(death_data$Race)
```

```
## [1] "American Indian or Alaska Native" "Asian or Pacific Islander"
## [3] "Black or African American"        "White"
## [5] ""
```

```r
unique(death_data$date)
```

```
##  [1] "2003-01-01" "2003-02-01" "2003-03-01" "2003-04-01" NA
##  [6] "2003-06-01" "2003-07-01" "2003-08-01" "2003-09-01" "2003-10-01"
## [11] "2003-11-01" "2003-12-01" "2004-01-01" "2004-02-01" "2004-03-01"
## [16] "2004-04-01" "2004-06-01" "2004-07-01" "2004-08-01" "2004-09-01"
## [21] "2004-10-01" "2004-11-01" "2004-12-01" "2005-01-01" "2005-02-01"
## [26] "2005-03-01" "2005-04-01" "2005-06-01" "2005-07-01" "2005-08-01"
## [31] "2005-09-01" "2005-10-01" "2005-11-01" "2005-12-01" "2006-01-01"
## [36] "2006-02-01" "2006-03-01" "2006-04-01" "2006-06-01" "2006-07-01"
## [41] "2006-08-01" "2006-09-01" "2006-10-01" "2006-11-01" "2006-12-01"
## [46] "2007-01-01" "2007-02-01" "2007-03-01" "2007-04-01" "2007-06-01"
## [51] "2007-07-01" "2007-08-01" "2007-09-01" "2007-10-01" "2007-11-01"
## [56] "2007-12-01" "2008-01-01" "2008-02-01" "2008-03-01" "2008-04-01"
## [61] "2008-06-01" "2008-07-01" "2008-08-01" "2008-09-01" "2008-10-01"
## [66] "2008-11-01" "2008-12-01"
```

```r
length(death_data[death_data$Race == "Asian or Pacific Islander",]$Deaths)
```

```
## [1] 14515
```

```r
unique(death_data$Race)
```

```
## [1] "American Indian or Alaska Native" "Asian or Pacific Islander"
## [3] "Black or African American"        "White"
## [5] ""
```

identify the 5 largest causes of death

```r
library(ggplot2)
library(survminer)
```

```
## Loading required package: ggpubr
```

```r
library(devtools)
```

```
## Loading required package: usethis
```

```r
library(easyGgplot2)

total_deaths <- aggregate(death_data$Deaths, list(death_data$UCD...ICD.10.113.Cause.List), sum)
total_deaths <- total_deaths[order(-total_deaths$x),]

list(head(total_deaths, n=5)$Group.1)[[1]][1]
```

```
## [1] "Major cardiovascular diseases (I00-I78)"
```

```
important_diseases <- death_data[death_data$UCD...ICD.10.113.Cause.List ==list(head(total_deaths, n=5)$
                                 death_data$UCD...ICD.10.113.Cause.List ==list(head(total_deaths, n=5)$
                                 death_data$UCD...ICD.10.113.Cause.List ==list(head(total_deaths, n=5)$
                                 death_data$UCD...ICD.10.113.Cause.List ==list(head(total_deaths, n=5)$
                                 death_data$UCD...ICD.10.113.Cause.List ==list(head(total_deaths, n=5)$
summary(important_diseases)
```

```
##     Notes               Race             Race.Code           Gender
##  Length:2880        Length:2880        Length:2880        Length:2880
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Gender.Code            Year          Year.Code         Month
##  Length:2880        Min.   :2003    Min.   :2003    Length:2880
##  Class :character   1st Qu.:2004    1st Qu.:2004    Class :character
##  Mode  :character   Median :2006    Median :2006    Mode  :character
##                     Mean   :2006    Mean   :2006
##                     3rd Qu.:2007    3rd Qu.:2007
##                     Max.   :2008    Max.   :2008
##
##   Month.Code        UCD...ICD.10.113.Cause.List
##  Length:2880        Length:2880
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
##  UCD...ICD.10.113.Cause.List.Code     Deaths          Population
##  Length:2880                      Min.   :   16.0   Length:2880
##  Class :character                 1st Qu.:  166.2   Class :character
##  Mode  :character                 Median :  883.0   Mode  :character
##                                   Mean   : 5759.2
##                                   3rd Qu.: 5951.2
##                                   Max.   :42076.0
##
##   Crude.Rate        Crude.Rate.Lower.95..Confidence.Interval
##  Length:2880        Length:2880
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
##  Crude.Rate.Upper.95..Confidence.Interval      date
##  Length:2880                              Min.   :2003-01-01
##  Class :character                         1st Qu.:2004-07-01
##  Mode  :character                         Median :2005-12-16
##                                           Mean   :2005-12-20
```
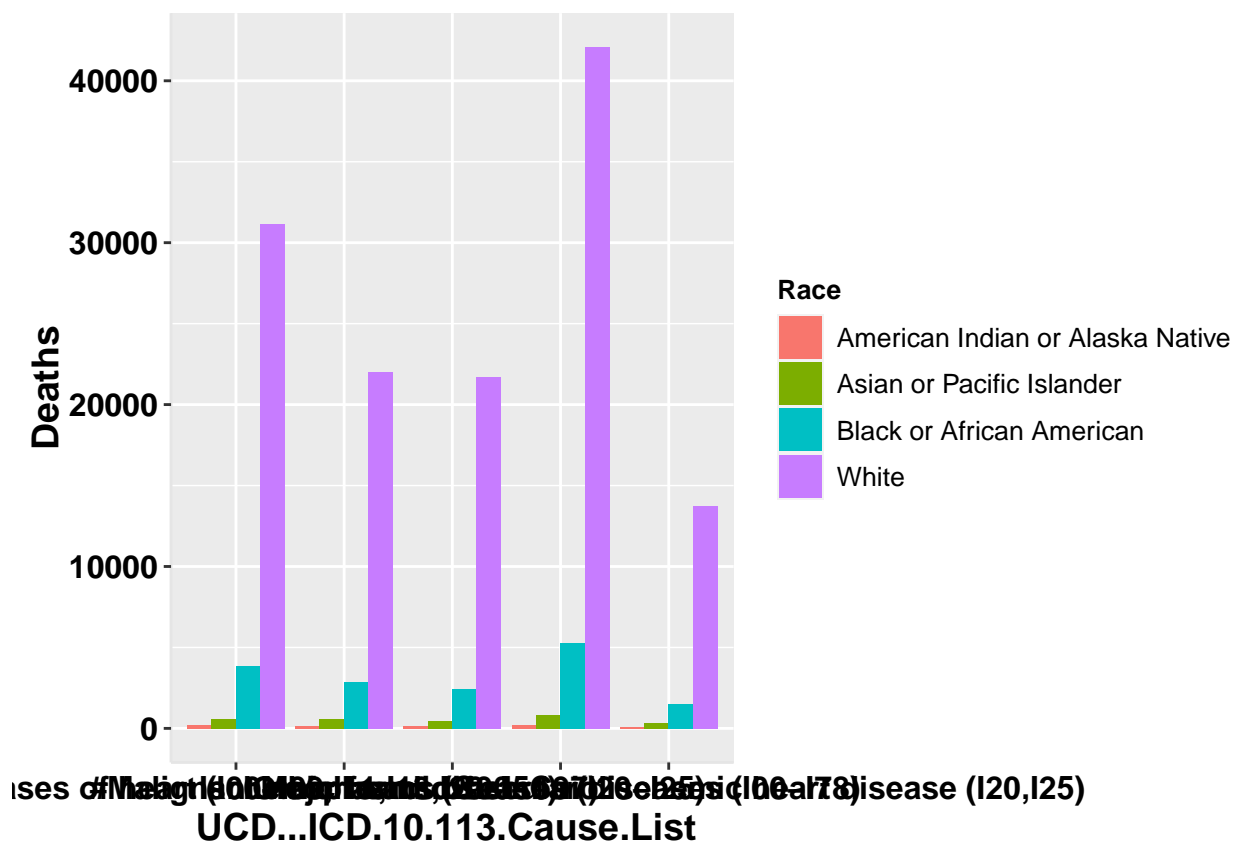
4

```
##                                            3rd Qu.:2007-07-01
##                                            Max.   :2008-12-01
##                                            NA's   :240
```

```r
counts <- important_diseases[, c('Deaths', 'Race', 'UCD...ICD.10.113.Cause.List')]

ggplot2.barplot(data=counts,
                xName='UCD...ICD.10.113.Cause.List',
                yName="Deaths",
                groupName='Race',
                position=position_dodge())
```

```
## Warning: Ignoring unknown parameters: binwidth
```



```r
api_data <- death_data[death_data$Race == "Asian or Pacific Islander",]

api_total_deaths <- aggregate(api_data$Deaths, list(api_data$UCD...ICD.10.113.Cause.List), sum)
api_total_deaths <- api_total_deaths[order(-api_total_deaths$x),]

list(head(api_total_deaths, n=5)$Group.1)[[1]][1]
```
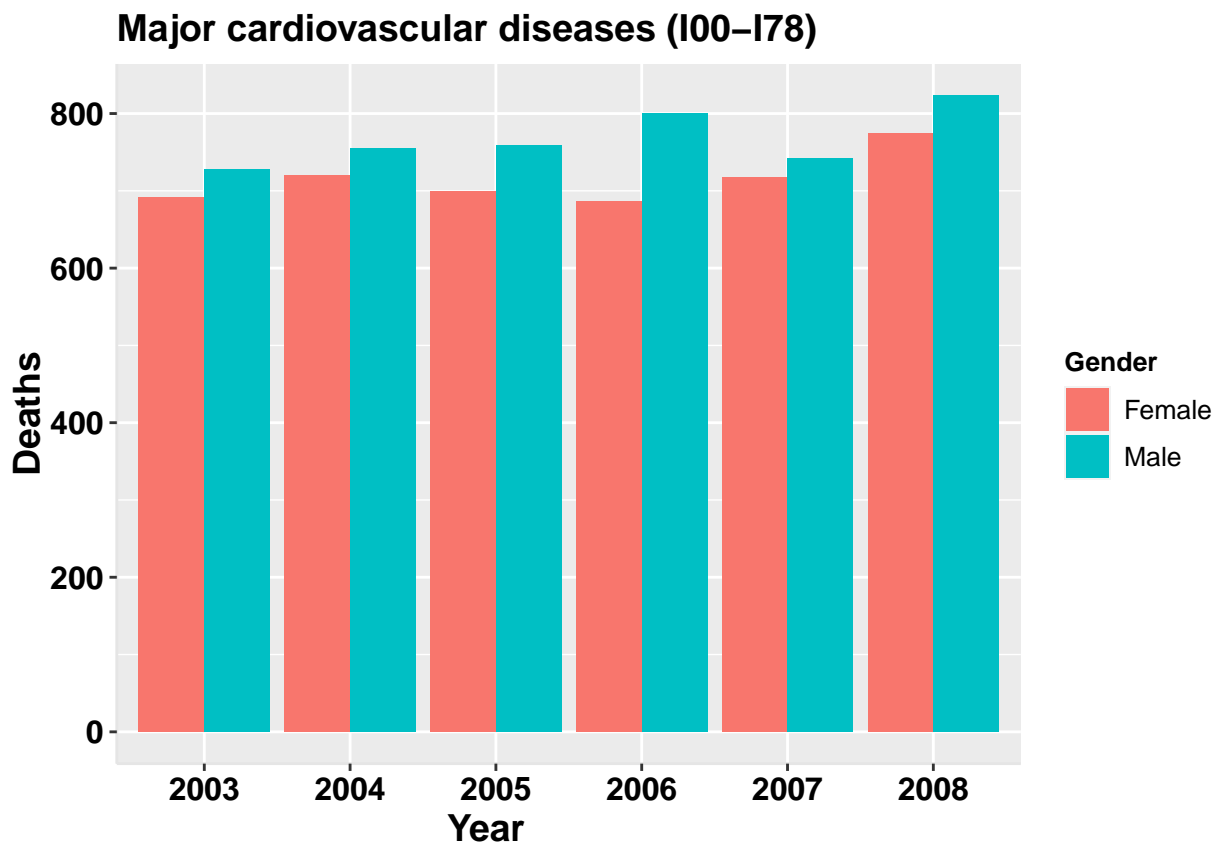
```
## [1] "Major cardiovascular diseases (I00-I78)"
```

```
api_important_diseases1 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_total_deaths, n
api_important_diseases2 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_total_deaths, n
api_important_diseases3 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_total_deaths, n
api_important_diseases4 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_total_deaths, n
api_important_diseases5 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_total_deaths, n

count1 <- api_important_diseases1[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count1,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][1])
```

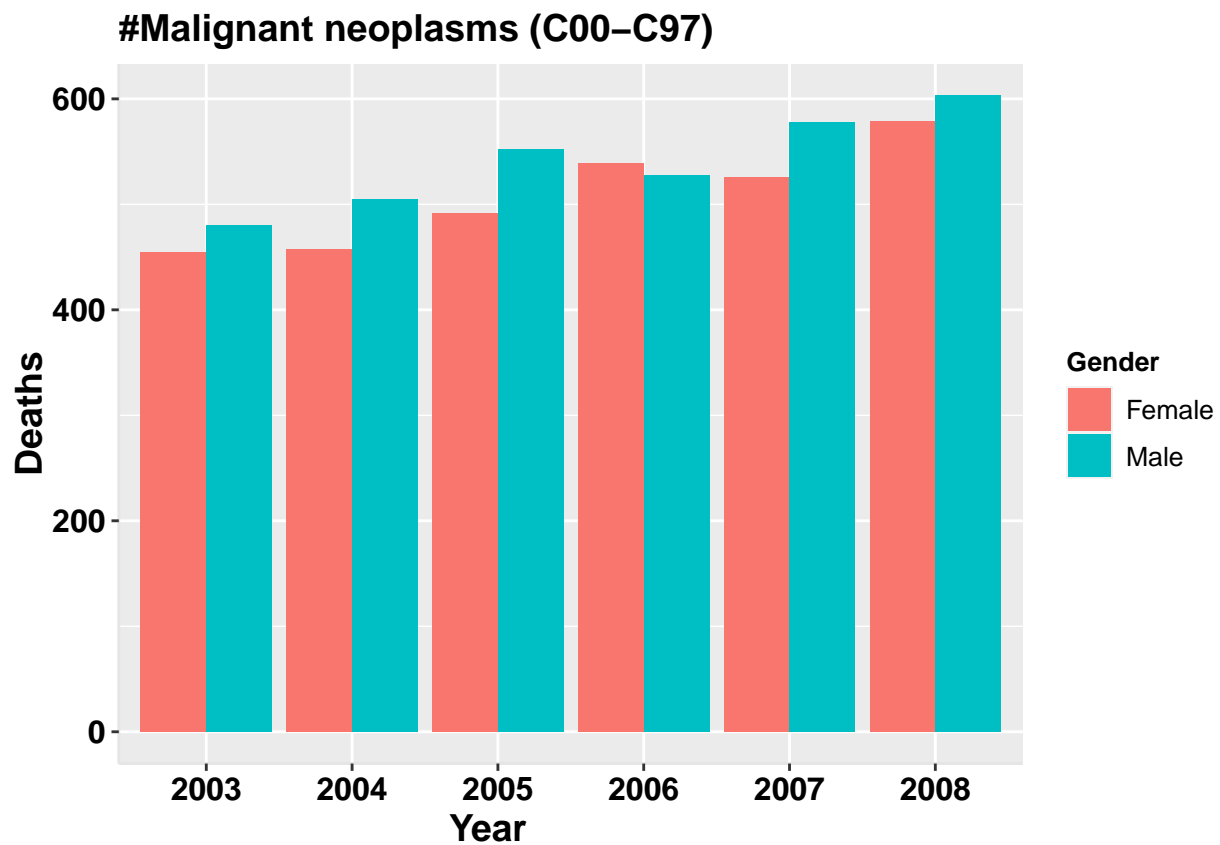## Warning: Ignoring unknown parameters: binwidth

Major cardiovascular diseases (I00–I78)

```
count2 <- api_important_diseases2[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count2,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][2])
```
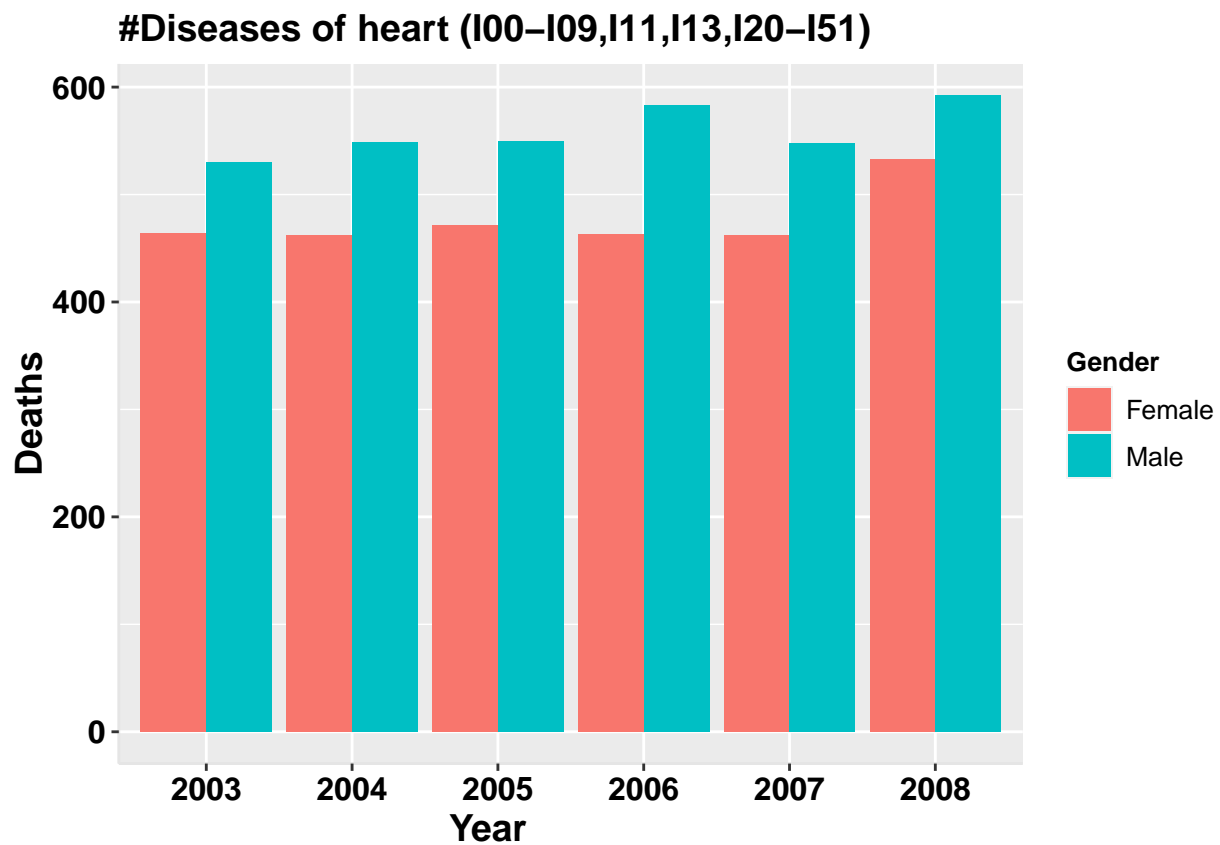
Warning: Ignoring unknown parameters: binwidth

#Malignant neoplasms (C00–C97)


```
count3 <- api_important_diseases3[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count3,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][3])
```
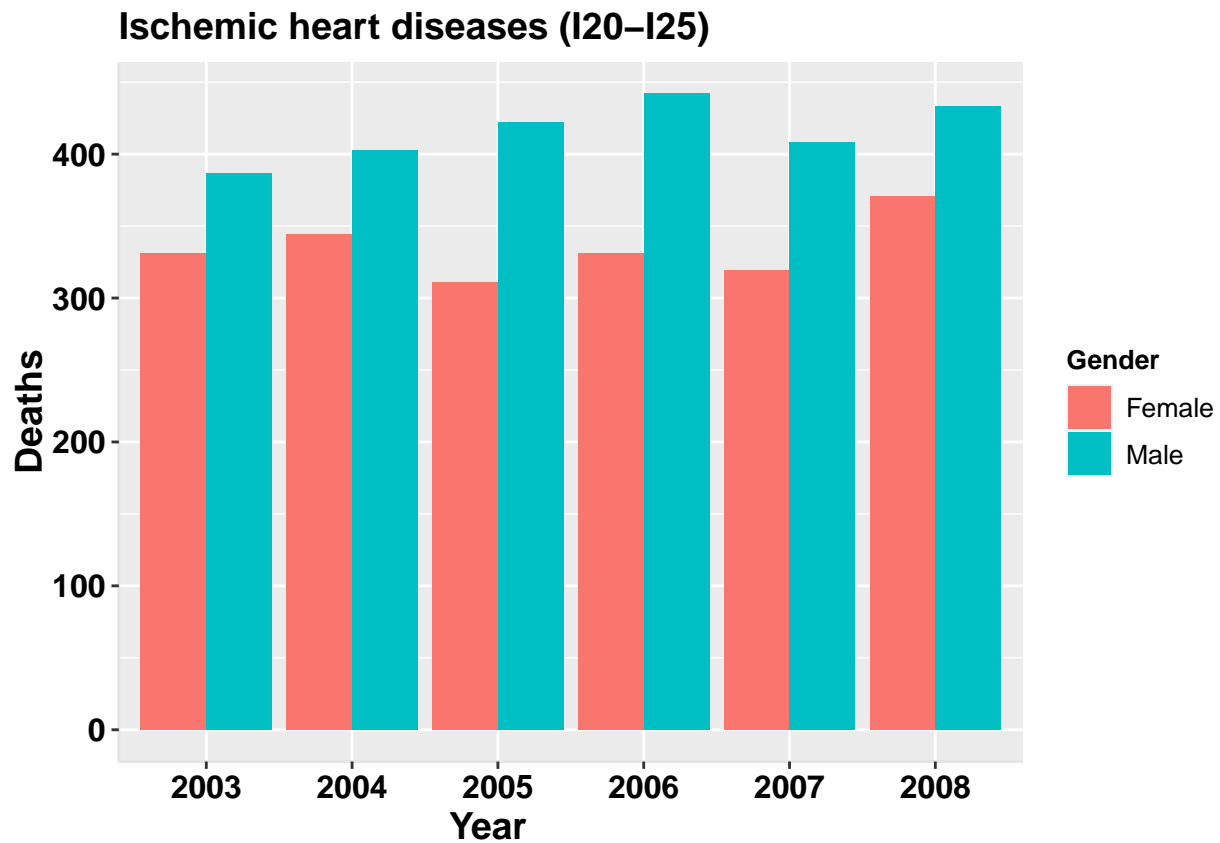
```
## Warning: Ignoring unknown parameters: binwidth
```

#Diseases of heart (I00–I09,I11,I13,I20–I51)


```
count4 <- api_important_diseases4[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count4,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][4])
```

```
## Warning: Ignoring unknown parameters: binwidth
```
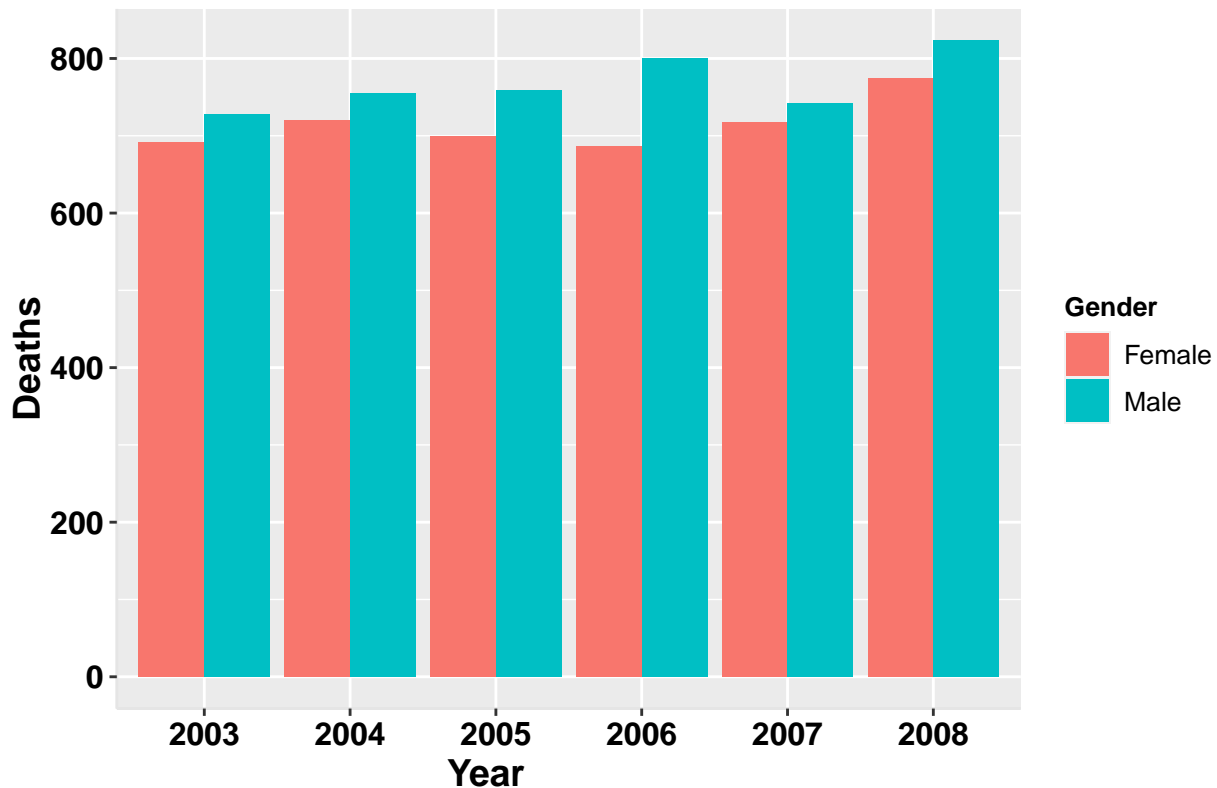
Ischemic heart diseases (I20–I25)


```
count5 <- api_important_diseases1[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count5,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][5])
```

```
## Warning: Ignoring unknown parameters: binwidth
```

Other forms of chronic ischemic heart disease (I20,I25)


```
female_data <- death_data[death_data$Gender == 'Female',]

female_total_deaths <- aggregate(female_data$Deaths, list(female_data$UCD...ICD.10.113.Cause.List), sum)
female_total_deaths <- female_total_deaths[order(-female_total_deaths$x),]

female_important_diseases1 <- female_data[female_data$UCD...ICD.10.113.Cause.List ==list(head(female_tot
female_important_diseases2 <- female_data[female_data$UCD...ICD.10.113.Cause.List ==list(head(female_tot
female_important_diseases3 <- female_data[female_data$UCD...ICD.10.113.Cause.List ==list(head(female_tot
female_important_diseases4 <- female_data[female_data$UCD...ICD.10.113.Cause.List ==list(head(female_tot
female_important_diseases5 <- female_data[female_data$UCD...ICD.10.113.Cause.List ==list(head(female_tot

count1 <- female_important_diseases1[, c('Deaths', 'Year', 'Race')]
ggplot2.barplot(data=count1,
                xName='Year',
                yName="Deaths",
                groupName='Race',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][1],
                ylim = c(0,50000))+
  geom_bar(stat = "identity", position = "dodge")+
  geom_text(
    aes(label = stat(y), group = Race),
    stat = 'summary', fun = sum, vjust = -1
  )
```
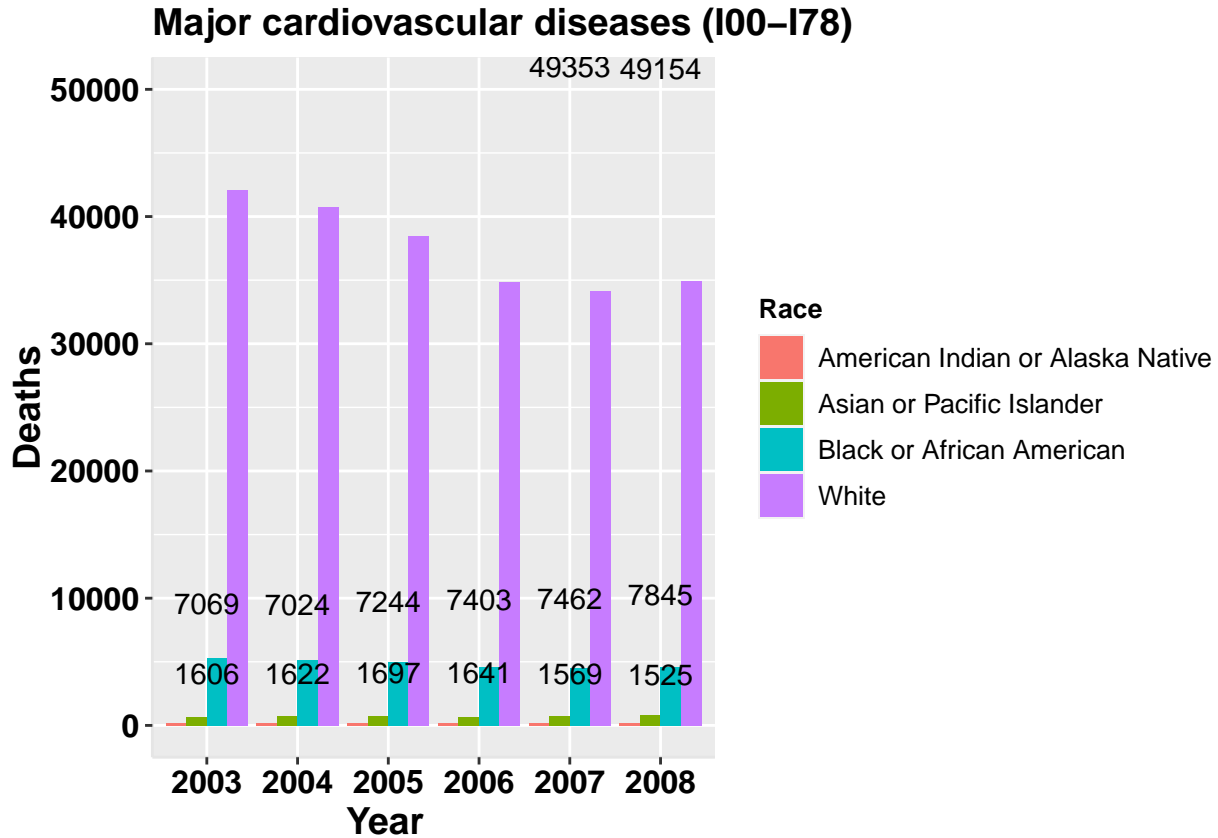
```
## Warning: Ignoring unknown parameters: binwidth
```

Warning: Removed 10 rows containing missing values (geom_text).

Major cardiovascular diseases (I00–I78)
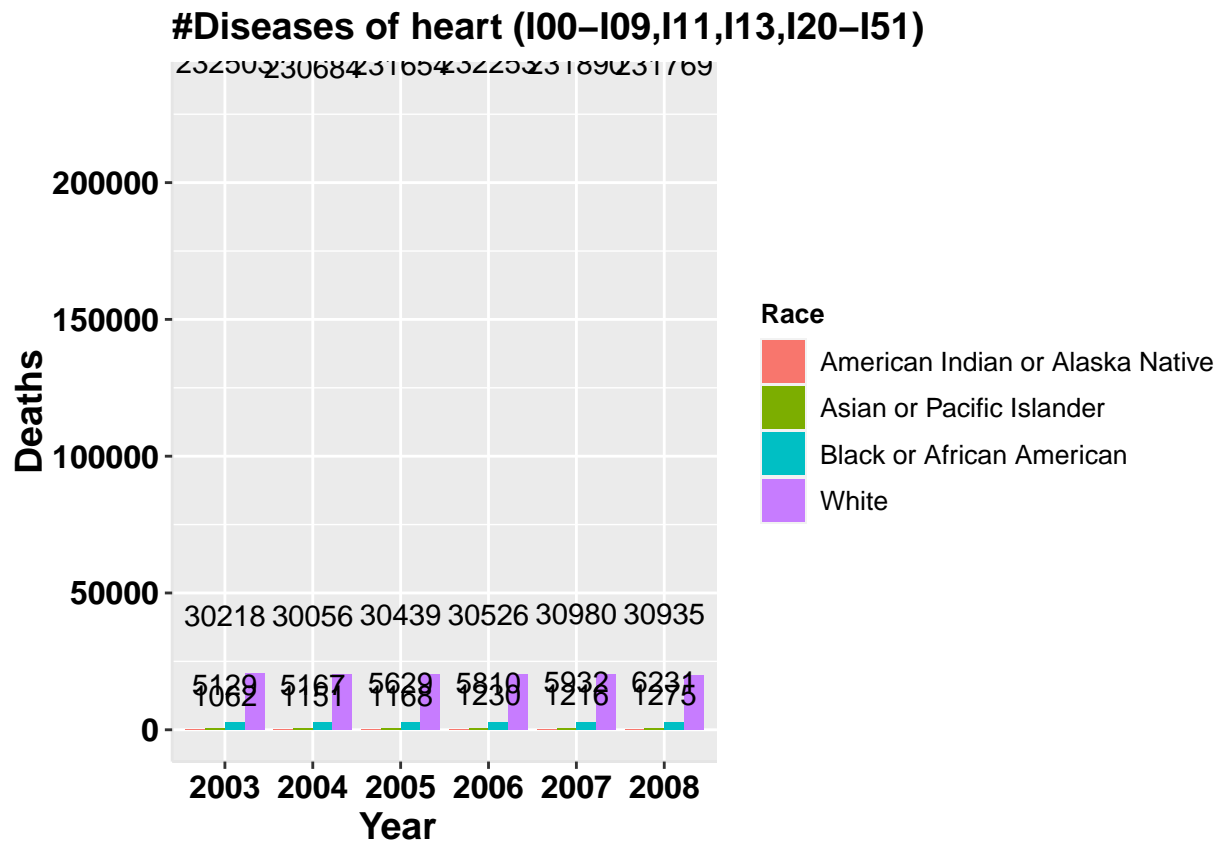


```
count2 <- female_important_diseases2[, c('Deaths', 'Year', 'Race')]
ggplot2.barplot(data=count2,
                xName='Year',
                yName="Deaths",
                groupName='Race',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][2])+
  geom_bar(stat = "identity", position = "dodge")+
  geom_text(
    aes(label = stat(y), group = Race),
    stat = 'summary', fun = sum, vjust = -1
  )
```

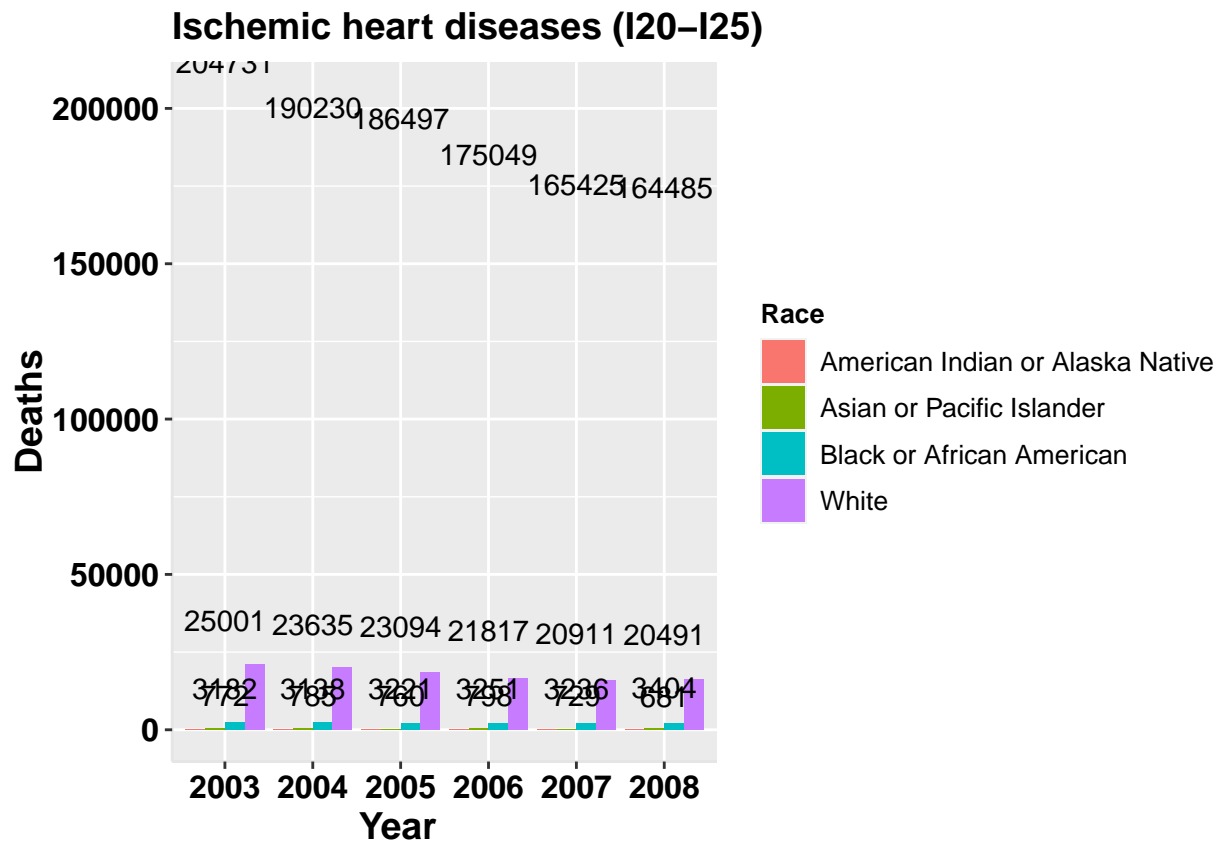## Warning: Ignoring unknown parameters: binwidth

#Malignant neoplasms (C00–C97)
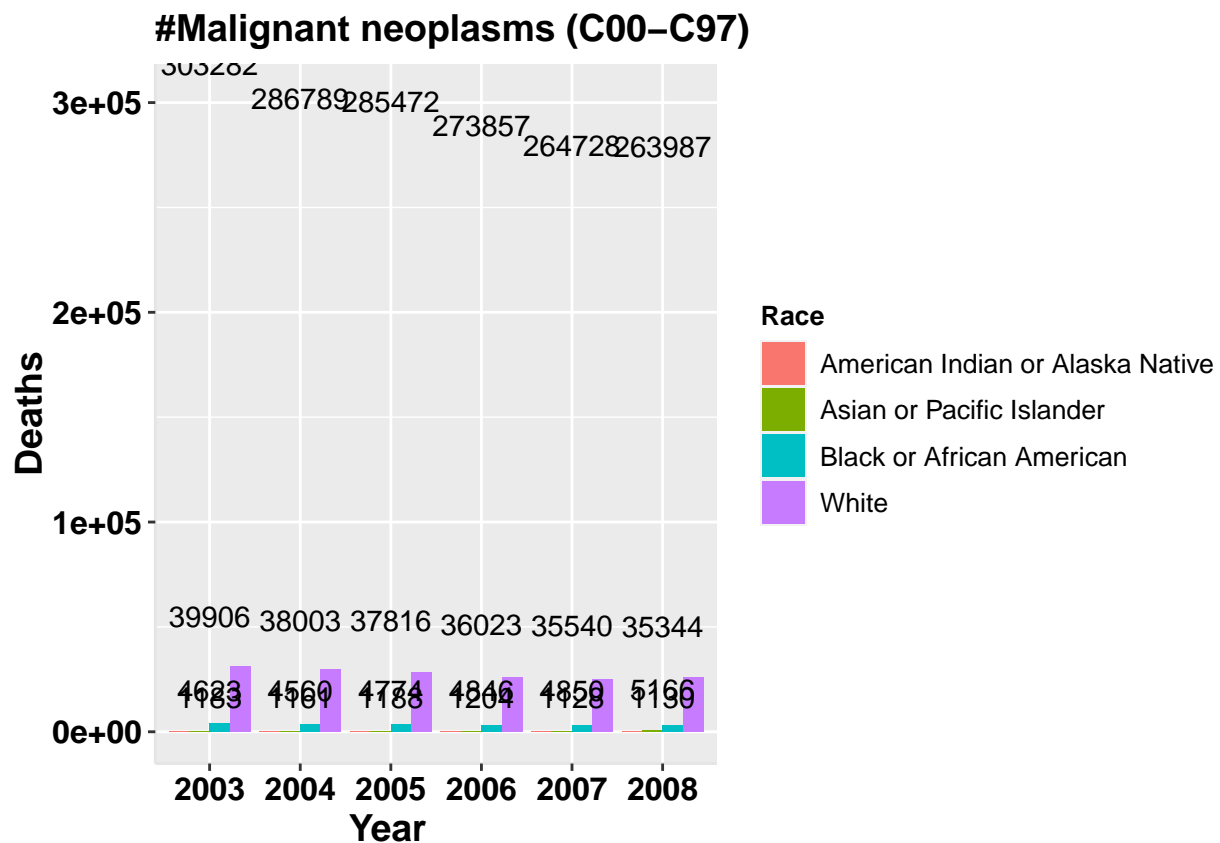

```
count3 <- female_important_diseases3[, c('Deaths', 'Year', 'Race')]
ggplot2.barplot(data=count3,
                xName='Year',
                yName="Deaths",
                groupName='Race',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][3])+
  geom_bar(stat = "identity", position = "dodge")+
  geom_text(
    aes(label = stat(y), group = Race),
    stat = 'summary', fun = sum, vjust = -1
  )
```

```
## Warning: Ignoring unknown parameters: binwidth
```

#Diseases of heart (I00–I09,I11,I13,I20–I51)

232503230684316543225923189231769

30218 30056 30439 30526 30980 30935

5129 5167 5629 5810 5932 6231
1062 1151 1168 1230 1216 1275

Race
- American Indian or Alaska Native
- Asian or Pacific Islander
- Black or African American
- White

```
count4 <- female_important_diseases4[, c('Deaths', 'Year', 'Race')]
ggplot2.barplot(data=count4,
                xName='Year',
                yName="Deaths",
                groupName='Race',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][4])+
  geom_bar(stat = "identity", position = "dodge")+
  geom_text(
    aes(label = stat(y), group = Race),
    stat = 'summary', fun = sum, vjust = -1
  )
```

## Warning: Ignoring unknown parameters: binwidth

Ischemic heart diseases (I20–I25)


```
count5 <- female_important_diseases5[, c('Deaths', 'Year', 'Race')]
ggplot2.barplot(data=count5,
                xName='Year',
                yName="Deaths",
                groupName='Race',
                position = "dodge",
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][5],
                ylim = c(0,20000))+
  geom_bar(stat = "identity", position = "dodge")+
  geom_text(
    aes(label = stat(y), group = Race),
    stat = 'summary', fun = sum, vjust = -1
  )
```

```
## Warning: Ignoring unknown parameters: binwidth
```

```
## Warning: Removed 6 rows containing missing values (geom_text).
```

Other forms of chronic ischemic heart disease (I20,I25)
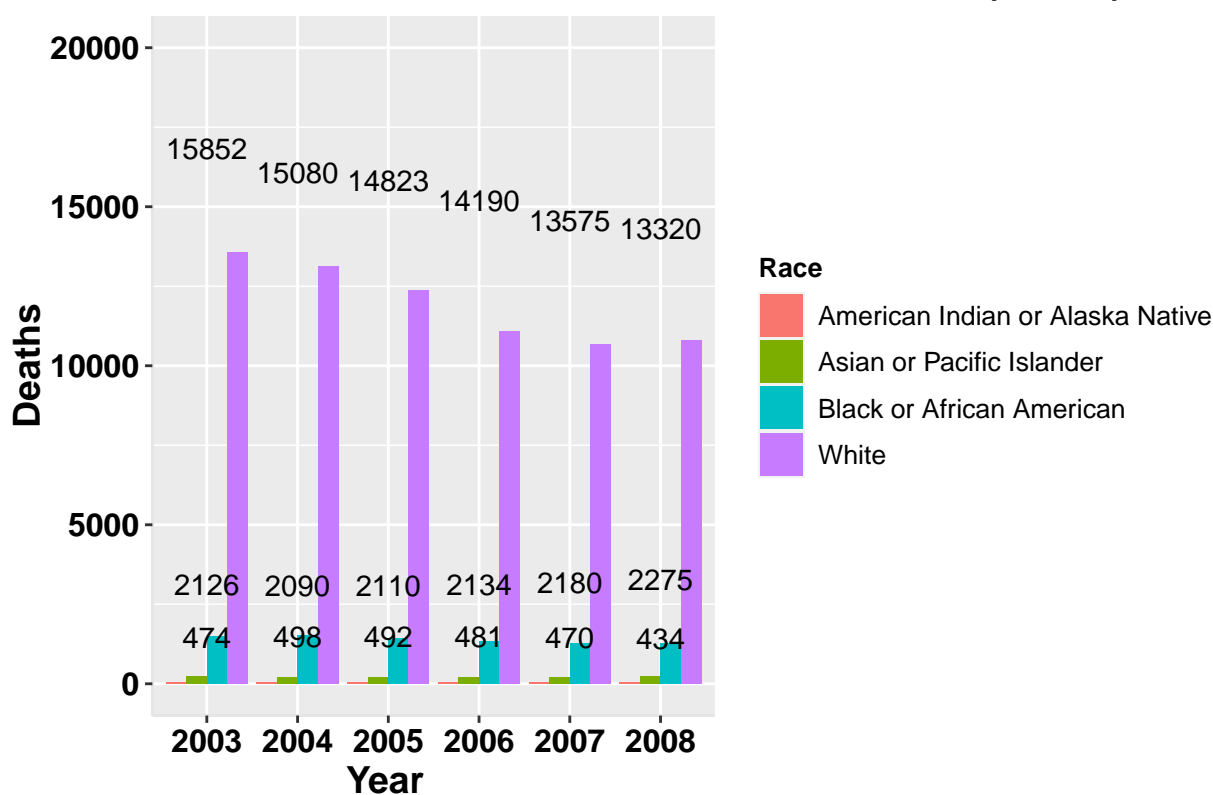

```
ggplot2.barplot(data=female_data,
                xName='Year',
                yName="Deaths",
                groupName='Race',
                position = "dodge",
                mainTitle= "Female Death Breakdown",
                ylim = c(0,45000))+
  geom_bar(stat = "identity", position = "dodge")+
  geom_text(
    aes(label = stat(y), group = Race),
    stat = 'summary', fun = sum, vjust = -1
  )
```

Warning: Ignoring unknown parameters: binwidth

Warning: Removed 18 rows containing missing values (geom_text).

Female Death Breakdown
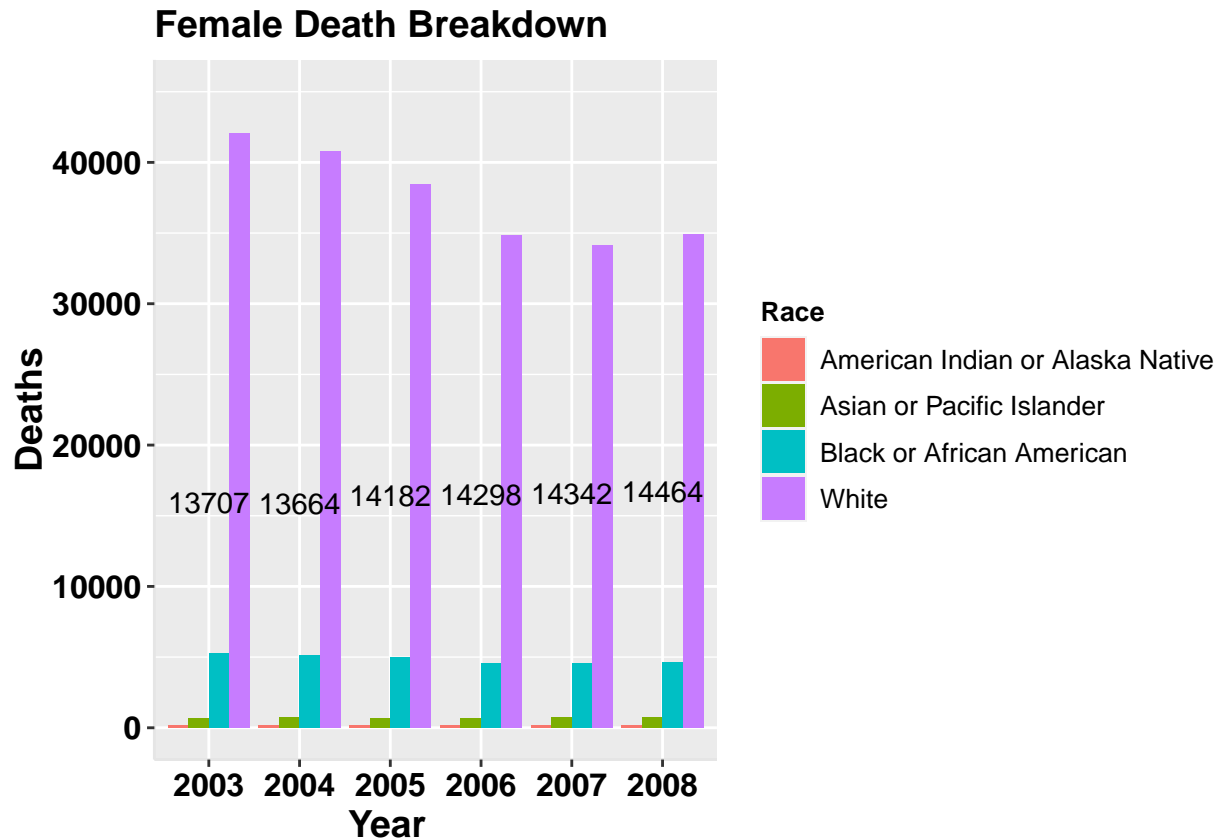

```
api_data_f <- death_data[death_data$Race == "Asian or Pacific Islander" & death_data$Gender == 'Female'

api_f_total_deaths <- aggregate(api_data_f$Deaths, list(api_data_f$UCD...ICD.10.113.Cause.List), sum)
api_f_total_deaths <- api_f_total_deaths[order(-api_f_total_deaths$x),]

list(head(api_total_deaths, n=5)$Group.1)[[1]][1]
```
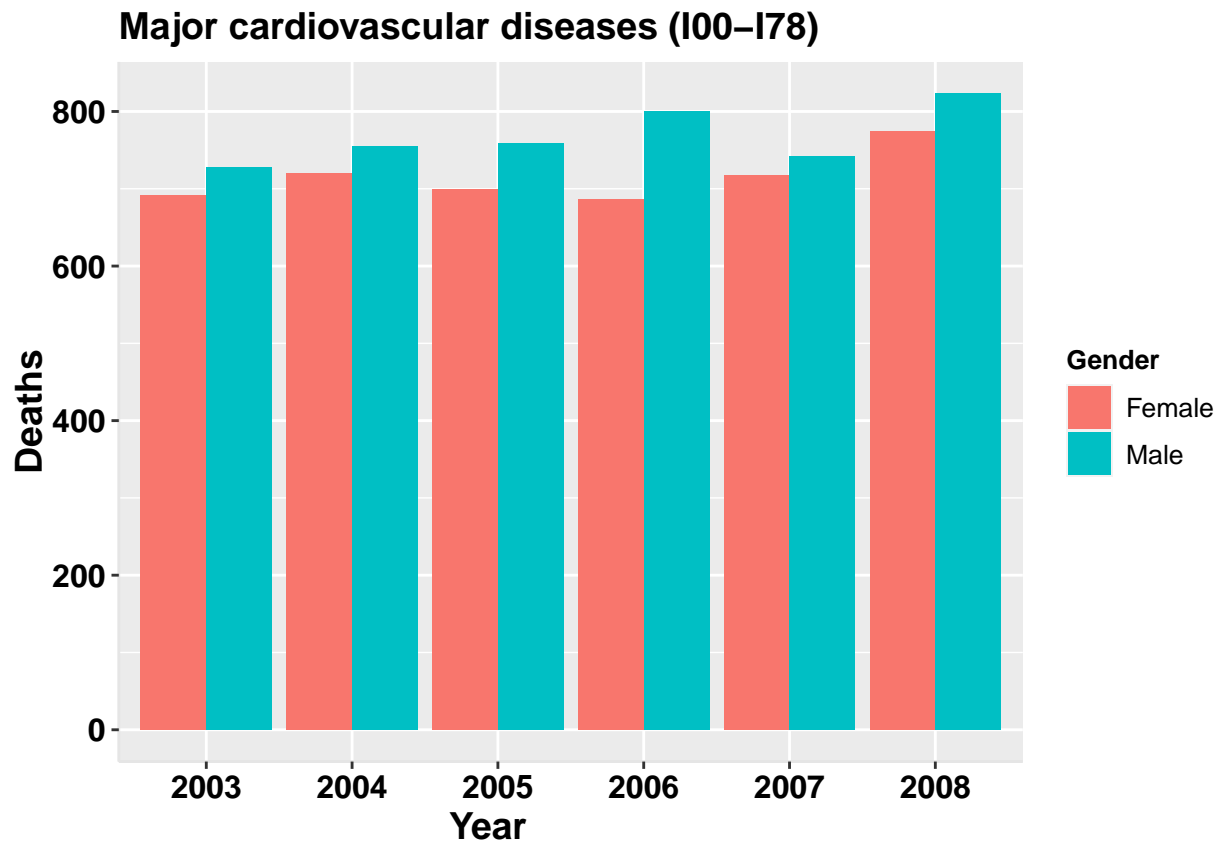
```
## [1] "Major cardiovascular diseases (I00-I78)"
```

```
api_important_diseases1 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_f_total_deaths
api_important_diseases2 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_f_total_deaths
api_important_diseases3 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_f_total_deaths
api_important_diseases4 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_f_total_deaths
api_important_diseases5 <- api_data[api_data$UCD...ICD.10.113.Cause.List ==list(head(api_f_total_deaths

count1 <- api_important_diseases1[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count1,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][1])
```
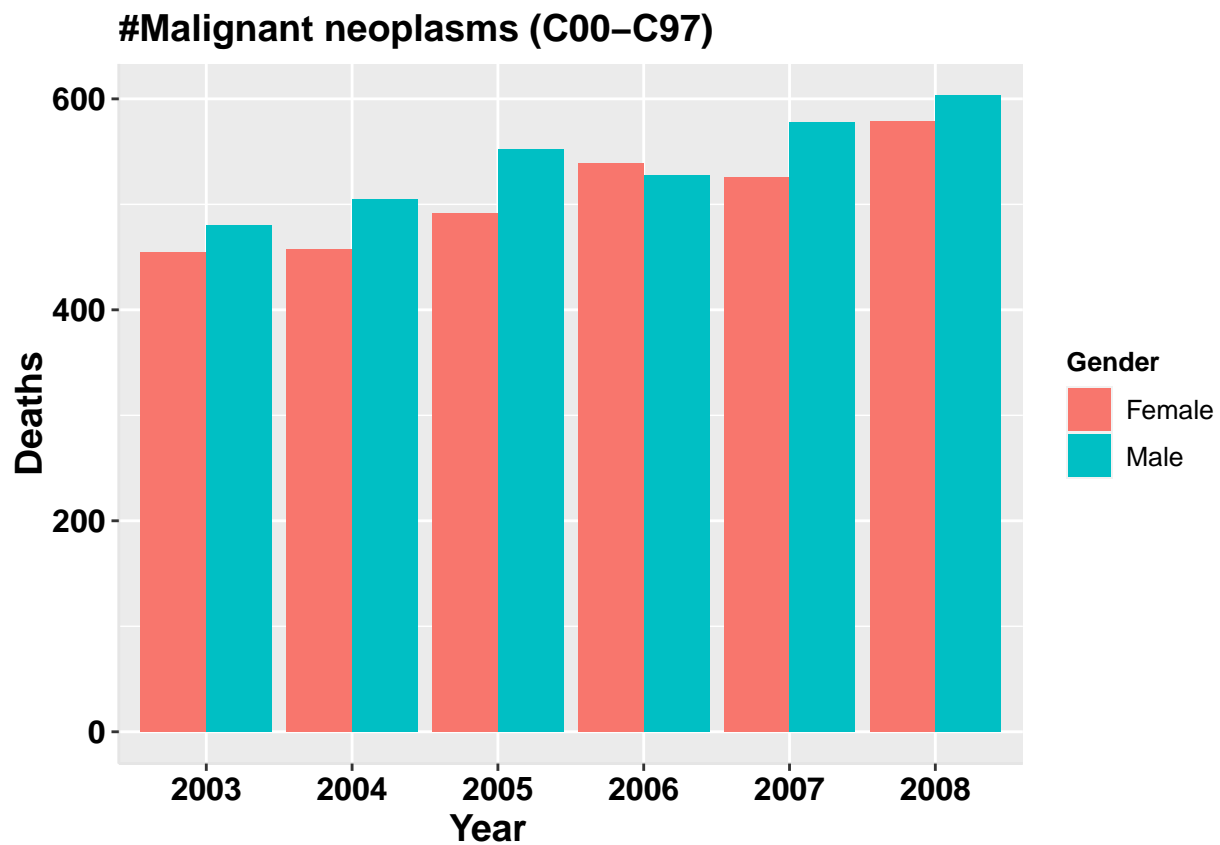
```
## Warning: Ignoring unknown parameters: binwidth
```
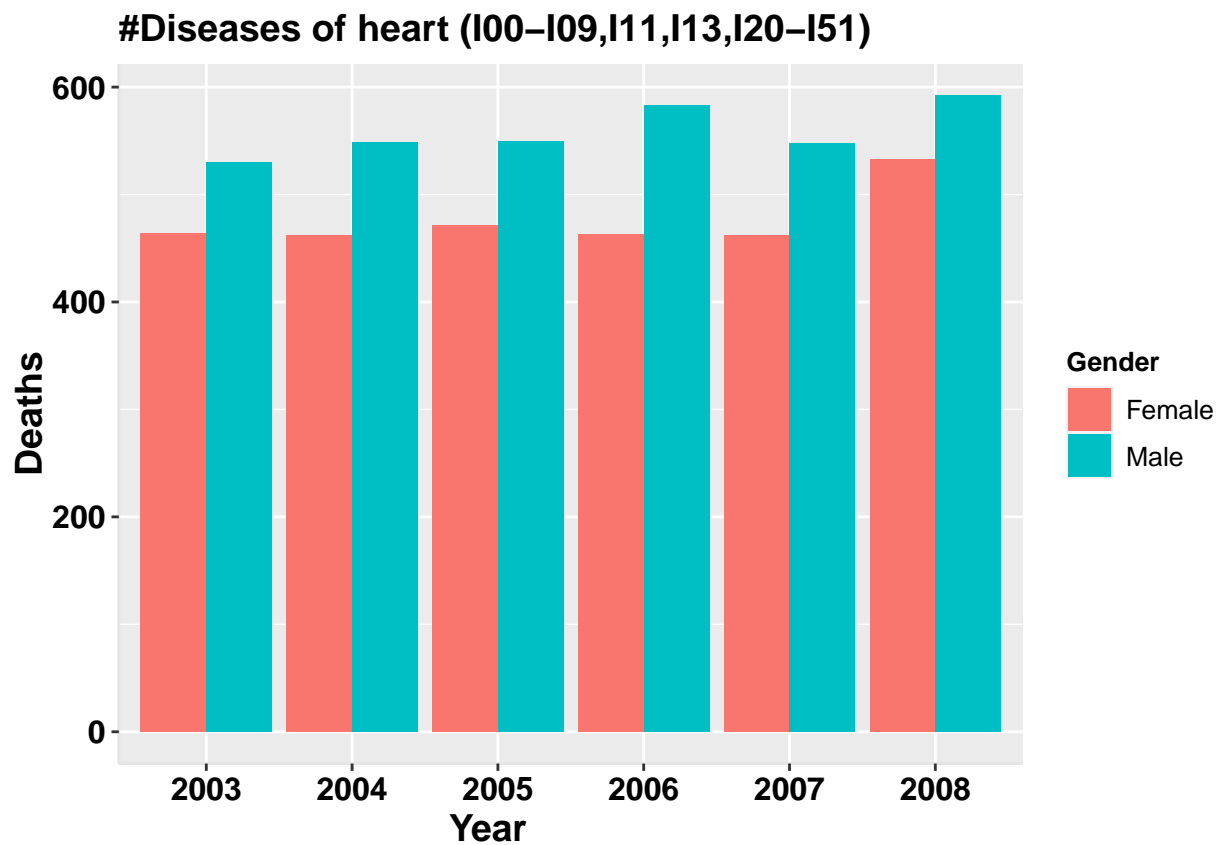
Major cardiovascular diseases (I00–I78)


```
count2 <- api_important_diseases2[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count2,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][2])
```

```
## Warning: Ignoring unknown parameters: binwidth
```
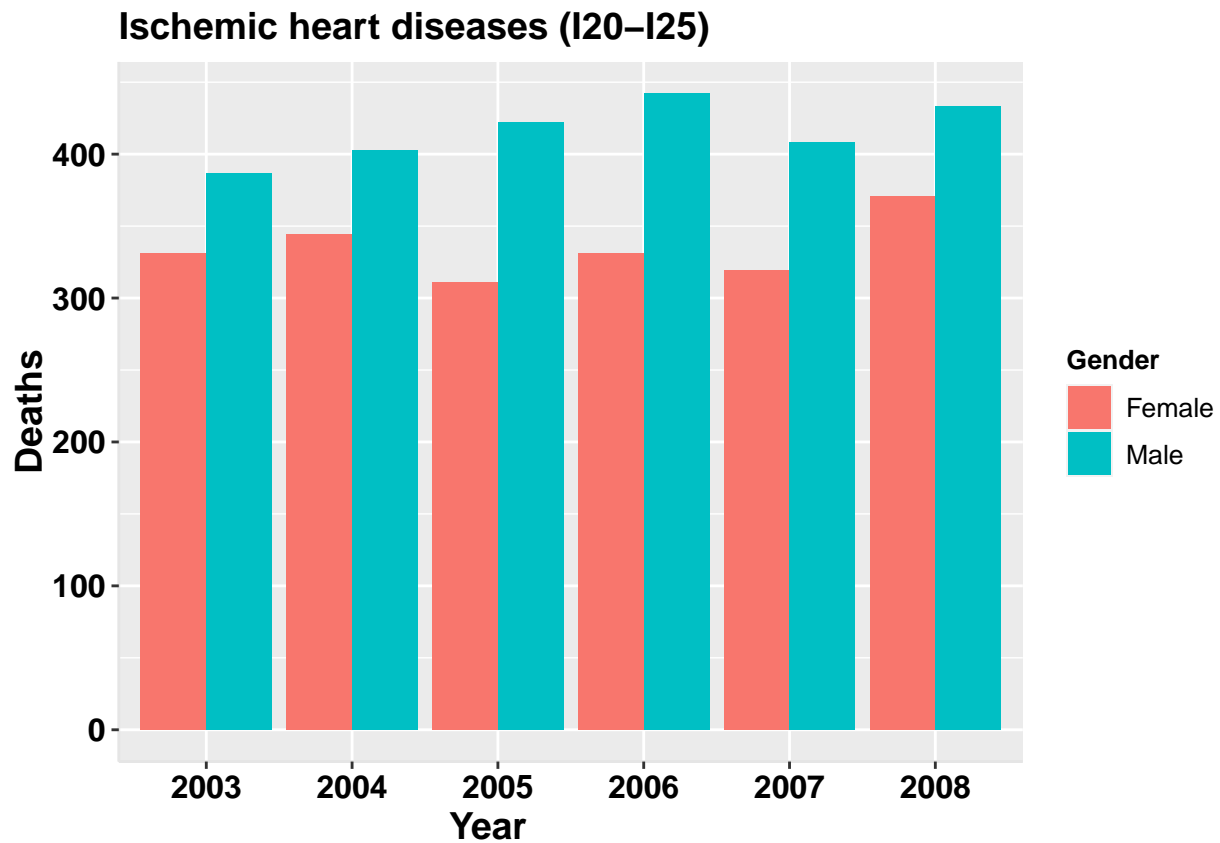
#Malignant neoplasms (C00–C97)


```
count3 <- api_important_diseases3[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count3,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][3])
```

```
## Warning: Ignoring unknown parameters: binwidth
```

#Diseases of heart (I00–I09,I11,I13,I20–I51)


```
count4 <- api_important_diseases4[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count4,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][4])
```

```
## Warning: Ignoring unknown parameters: binwidth
```
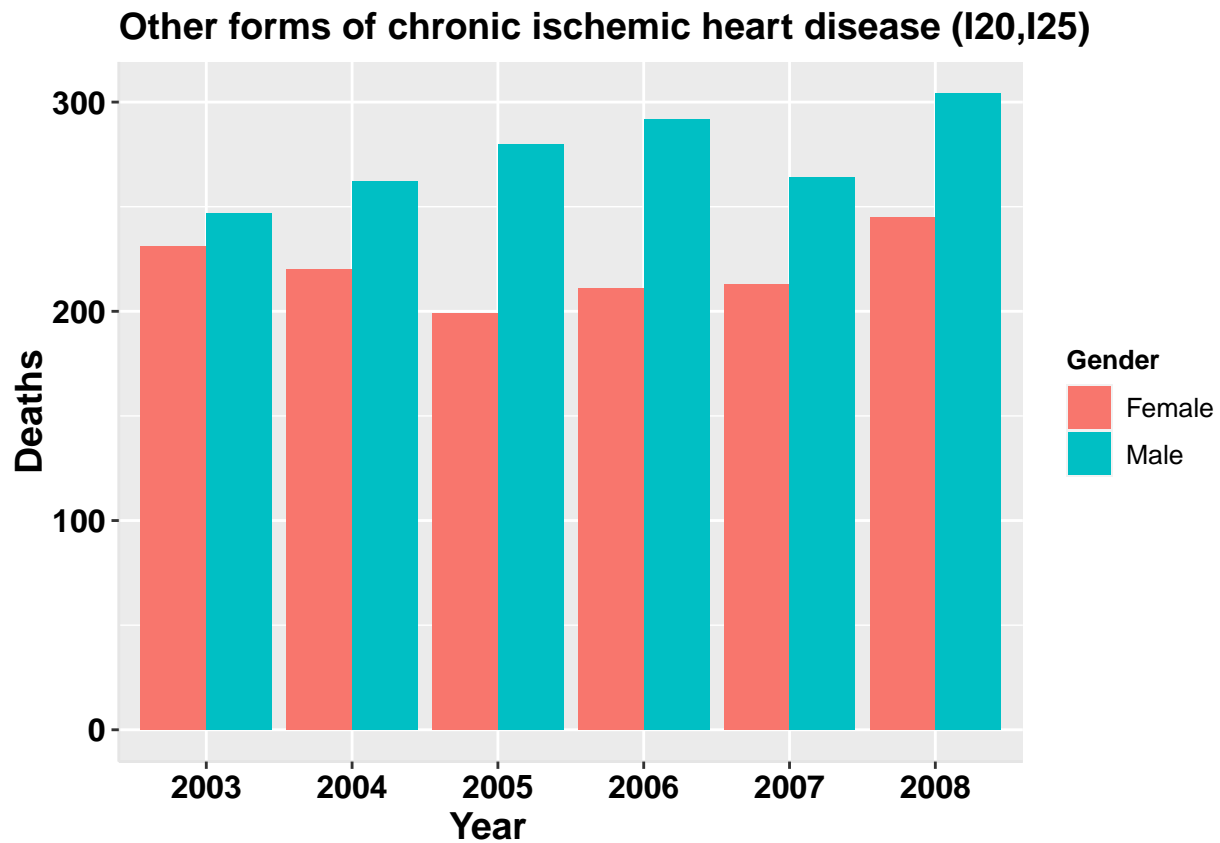
Ischemic heart diseases (I20–I25)


```
count5 <- api_important_diseases5[, c('Deaths', 'Year', 'Gender')]
ggplot2.barplot(data=count5,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle=list(head(api_total_deaths, n=5)$Group.1)[[1]][5])
```

```
## Warning: Ignoring unknown parameters: binwidth
```
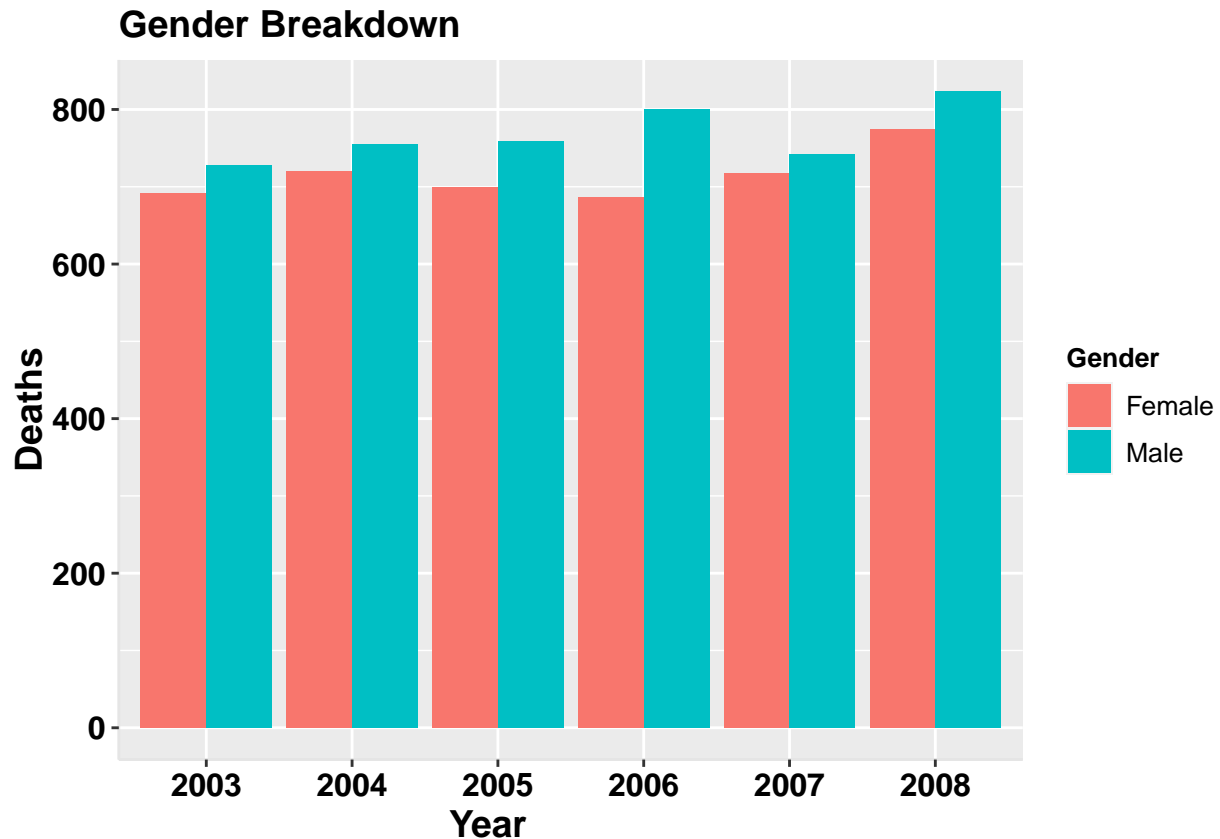
Other forms of chronic ischemic heart disease (I20,I25)


```
ggplot2.barplot(data=api_data,
                xName='Year',
                yName="Deaths",
                groupName='Gender',
                position=position_dodge(),
                mainTitle='Gender Breakdown')
```

```
## Warning: Ignoring unknown parameters: binwidth
```

Gender Breakdown


```
api_data_m <- death_data[death_data$Race == "Asian or Pacific Islander" & death_data$Gender == 'Male', ]

api_m_total_deaths <- aggregate(api_data_m$Deaths, list(api_data_m$UCD...ICD.10.113.Cause.List), sum)
api_m_total_deaths <- api_m_total_deaths[order(-api_m_total_deaths$x),]


list(head(api_f_total_deaths, n=5)$Group.1)[[1]]
```

```
## [1] "Major cardiovascular diseases (I00-I78)"
## [2] "#Malignant neoplasms (C00-C97)"
## [3] "#Diseases of heart (I00-I09,I11,I13,I20-I51)"
## [4] "Ischemic heart diseases (I20-I25)"
## [5] "Other forms of chronic ischemic heart disease (I20,I25)"
```

```
list(head(api_m_total_deaths, n=5)$Group.1)[[1]]
```

```
## [1] "Major cardiovascular diseases (I00-I78)"
## [2] "#Malignant neoplasms (C00-C97)"
## [3] "#Diseases of heart (I00-I09,I11,I13,I20-I51)"
## [4] "Ischemic heart diseases (I20-I25)"
## [5] "Other forms of chronic ischemic heart disease (I20,I25)"
```

```
unique(death_data$date_diff)
```

```
## NULL
```

```r
death_data$date_diff <- as.Date(as.character(death_data$date), format="%Y-%m-%d")-as.Date(('2003/01/01')

api_data <- death_data[death_data$Race == "Asian or Pacific Islander",]

api_model <- lm(Deaths ~ date_diff + Gender, data = api_data)
summary(api_model)
```

```
##
## Call:
## lm(formula = Deaths ~ date_diff + Gender, data = api_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.38  -39.00  -31.15   -5.99  773.53
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.846759   1.876223  20.172  < 2e-16 ***
## date_diff    0.002727   0.001333   2.046   0.0408 *
## GenderMale   6.638679   1.686649   3.936 8.33e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.28 on 13304 degrees of freedom
##   (1208 observations deleted due to missingness)
## Multiple R-squared:  0.001477,   Adjusted R-squared:  0.001327
## F-statistic: 9.841 on 2 and 13304 DF,  p-value: 5.36e-05
```

```r
plot(Deaths ~ date_diff + Gender, data = api_data,
     col = c("lightblue", "lightpink"),
     pch = 20,
     cex = 1.5,
     main = "API Gender Comparison",
     xlim = c(0, 2161),
     ylim = c(0, 1000))
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by coercion
```

API Gender Comparison
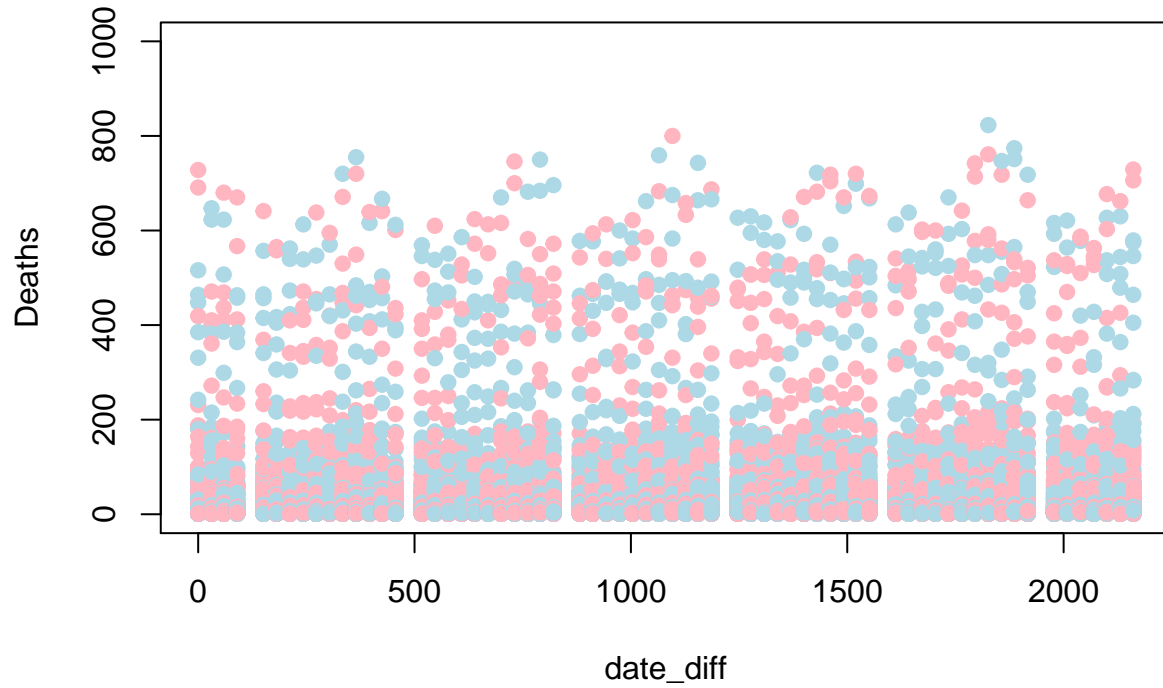

```
abline(api_model, lwd = 3, lty = 1, col = "Blue")
```

```
## Warning in abline(api_model, lwd = 3, lty = 1, col = "Blue"): only using the
## first two of 3 regression coefficients
```

```
legend("topright",
       c("Male","Female"),
       lty = c(1, 2),
       lwd = 2,
       col = c("darkorange", "red", "blue", "darkgreen"))
```

API Gender Comparison

Male
Female

Deaths

Gender