titanic EDA

Nam Kyeong Hun

2020 5 25

library(dplyr)

##   
## Attaching package: 'dplyr'

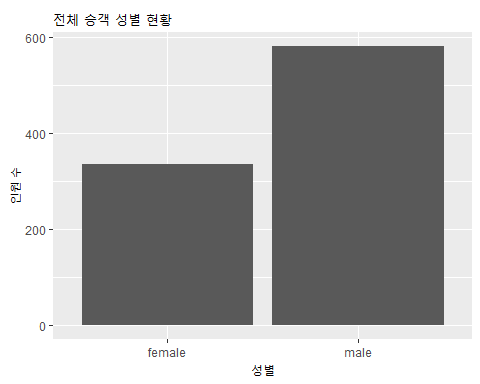
## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

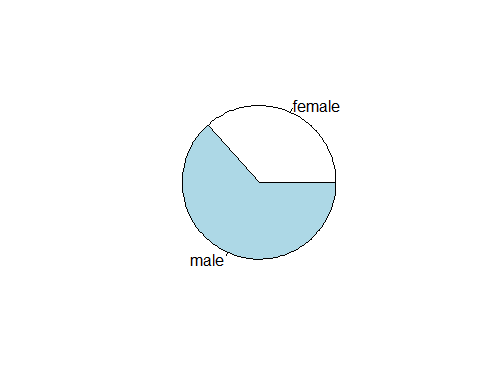
library(ggplot2)  
  
setwd ("C:\\Rworkspace\\new-R-working-space")  
  
ti <- read.csv(file="titanic\_train.csv",  
 header = T)

## 타이타닉 탑승자 현황

ggplot(ti, aes(x=sex))+  
 geom\_bar()+  
 labs(x="성별",y="인원 수")+  
 ggtitle("전체 승객 성별 현황")



pie(table(ti$sex))

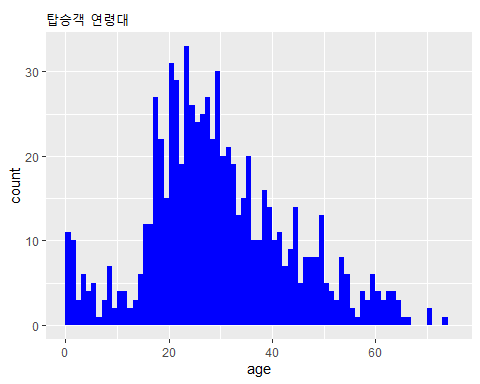


총 탑승자는 916명이며, 그중 남성 탑승객은 581명, 여성 탑승객은 335명이다.

## 탑승객 연령

ggplot(ti,aes(age))+  
 geom\_histogram(breaks=seq(0,75,by=1),  
 fill="blue")+  
 ggtitle("탑승객 연령대")

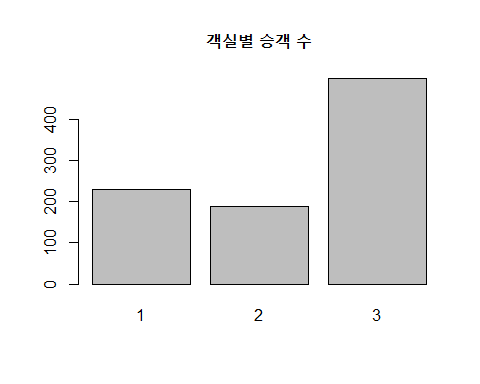
## Warning: Removed 175 rows containing non-finite values (stat\_bin).



주로 20~30세 사이의 연령을 기록하였다.

## 객실별 승객

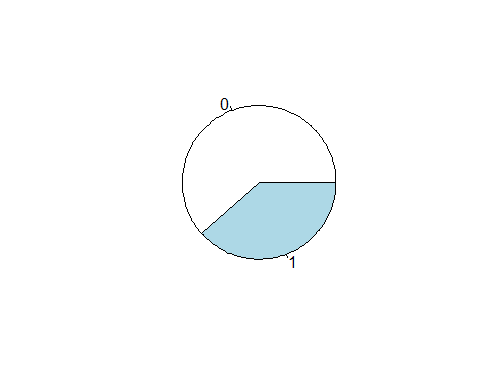
tc <- table(ti$pclass)  
barplot(tc,main = "객실별 승객 수")



1등급 객실 승객 수는 230명, 2등급 객실 승객 수는 188명, 3등급 객실 승객 수는 498명으로 나타났다.

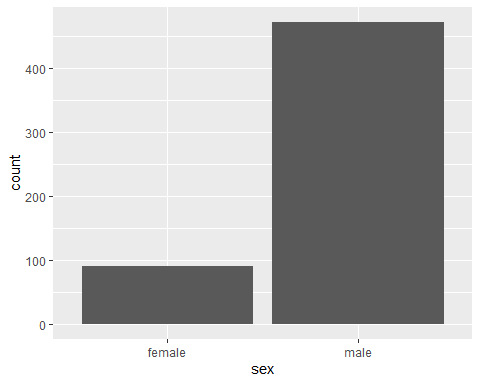
## 생존/사망 현황

td <-ti$survived  
td1<- table(td)  
pie(td1)

 총 사망자는 563명, 생존자는 353명으로 나타났다.

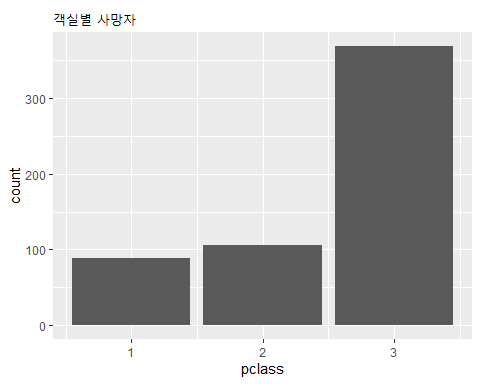
## 사망자 정보

ti.dead <- ti%>%  
 group\_by(survived)%>%  
 filter(survived==0)  
  
ggplot(ti.dead,aes(x=sex))+  
 geom\_bar()



남성이 여성보다 4배 넘게 사망했으며,

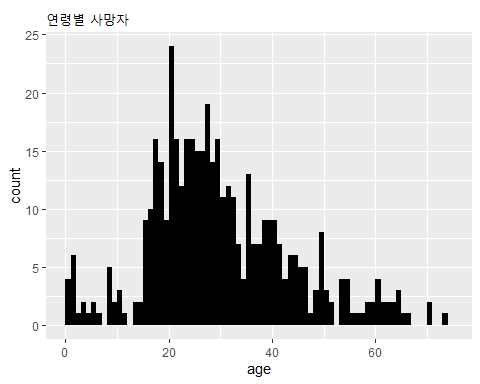
ti.d.c <- ti%>%  
 group\_by(pclass)%>%  
 filter(survived==0)  
  
ggplot(ti.d.c,aes(x=pclass))+  
 geom\_bar()+  
 ggtitle("객실별 사망자")



3등급 객실에서의 사망자가 다른 객실 사망자에 비해 3배가량 많은 것으로 나타났다.

ggplot(ti.dead,aes(age))+  
 geom\_histogram(breaks=seq(0,75,by=1),  
 fill="black")+  
 ggtitle("연령별 사망자")

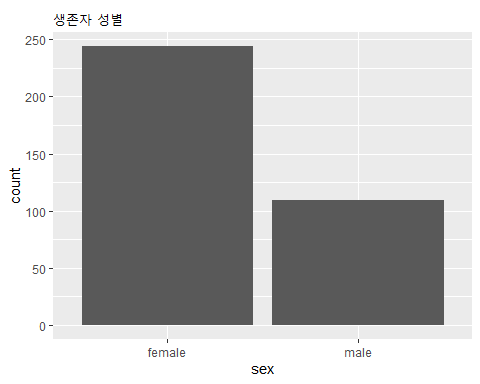
## Warning: Removed 128 rows containing non-finite values (stat\_bin).



연령대별로는 20대에서 30대 사이에서 가장 많은 사망자가 발생하였다.

## 생존자 정보

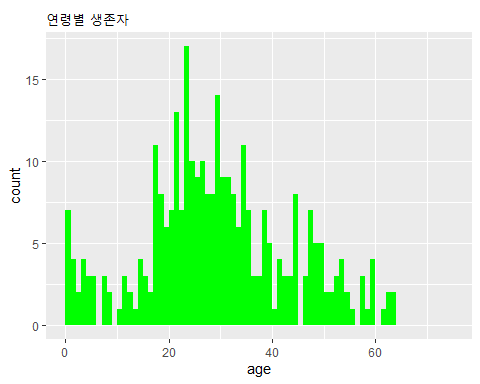
ti.alive <- ti%>%  
 group\_by(survived)%>%  
 filter(survived==1)  
  
ggplot(ti.alive,aes(x=sex))+  
 geom\_bar()+  
 ggtitle("생존자 성별")



여성이 남성보다 2배가량 더 생존하였으며,

ggplot(ti.alive,aes(age))+  
 geom\_histogram(breaks=seq(0,75,by=1),  
 fill="green")+  
 ggtitle("연령별 생존자")

## Warning: Removed 47 rows containing non-finite values (stat\_bin).



연령적으로는 20~30세 사이인 것으로 나타났다.