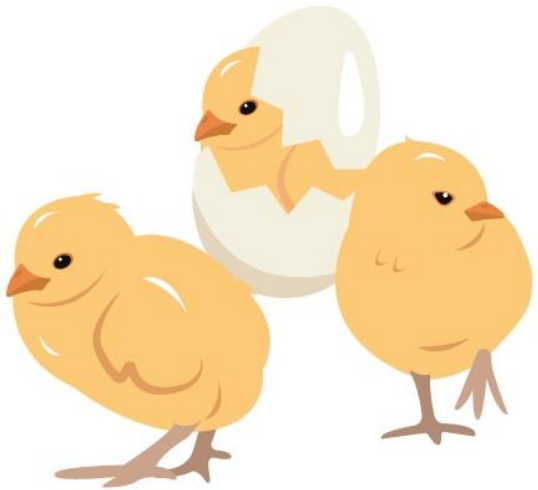
모두를 위한

R 데이터 분석 입문

2판



Chapter 14
데이터 분석 사례 II

목차

Section 01

데이터셋 설명

1. 데이터셋 설명

- 국민건강보험 가입자 중 요양기관(병, 의원 등)으로부터의 진료이력이 있는 각 연도별 수진자(진료를 받은 사람) 100만 명에 대한 기본정보(성, 연령대 등)와 진료내역(진료과목코드, 주상병 코드, 요양일수, 총처방일수 등)으로 구성된 개방 데이터.
- 2018년도와 2020년도의 데이터를 분석 대상으로 함.
- 데이터는 공공데이터 포털(https://www.data.go.kr/)에서 다운로드
- 일부 항목(열)은 삭제를 하였고, 수진자의 지역은 인천광역시로 제한.

1. 데이터셋 설명

표 14-1 진료 내역 데이터셋 변수 설명

No	변수명	설명
1	기준년도	진료 데이터 수집 연도
2	가입자일련번호	수진자 구분을 위한 ID
3	성별코드	성별코드(1-남자, 2-여자)
4	연령대코드	수진자 연령대코드([표 14-2) 참조)
5	요양개시일자	외래진료의 경우 병의원에 내원하여 진료받은 날짜이며, 입원진료의 경우는 입원 일자
6	서식코드	내원한 수진자의 진료형태([표 14-3] 참조)
7	진료과목코드	수진자의 진료과목코드([표 14-4] 참조)
8	주상병코드	병의원에서 진단받은 질병의 코드([표 14-5] 참조)
9	요양일수	수진자가 요양급여를 받은 실제일수(입원일수 ÷ 내원일수+투약일수)
10	입내원일수	입원일수 또는 내원일수
11	심결요양급여비용총액	심결본인부담금+심결보험자부담금
12	심결본인부담금	수진자가 부담해야 하는 진료비 금액
13	심결보험자부담금	보험에서 지급되는 진료비 금액
14	총처방일수	처방전을 발급한 경우에 해당 처방전에 따라 조제 투약하도록 처방한 일수의 합계

1. 데이터셋 설명

표 14-2 연령대코드의 예

코드	연령대
1	0~4세
2	5~9세
3	10~14세
4	15~19세
..	..
18	85세+

표 14-3 서식코드의 예

코드	진료형태
2	의과 입원
3	의과 외래
6	조산원 입원
7	보건기관 입원
..	..
11	정신과 외래

표 14-4 진료과목코드의 예

코드	진료형태
0	일반의
1	내과
2	신경과
3	정신과
..	..
88	한방응급

표 14-5 주상병코드의 예

코드	질병명
A00	콜레라
A000	비브리오 콜레라 – 콜레라형균에 의한 콜레라
A000	고전적 콜레라
A001	비브리오 콜레라 – 엘토르형균에 의한 콜레라
..	..
Z999	상세불명의 기능성 기계 및 장치에 의존

Section 02

데이터 탐색

코드 14-1 *(계속)*

```
library(ggplot2)
setwd('D:/source')

#################################################
## 데이터셋 준비
#################################################
ds.2019 <- read.csv('NHIS_INCHON_2019.csv')
treat.code <- read.csv('서식코드.csv')
age.code <- read.csv('연령대코드.csv')
disease.code <- read.csv('주상병코드.csv')
dept.code <- read.csv('진료과목코드.csv')

dim(ds.2019)
View(head(ds.2019))
```

코드 14-1 *(계속)*

```
#####################################################
## 데이터 탐색
#####################################################
# 수진자수
temp <- unique(ds.2019[,2:4])
nrow(temp)

# 수진자 1명당 평균 진료회수
nrow(ds.2019)/nrow(temp)

# 수진자 남녀 비율
table(temp$성별코드)
```

```
> # 수진자수
> temp <- unique(ds.2019[,2:4])
> nrow(temp)
[1] 58084
```

```
> # 수진자 1명당 평균 진료회수
> nrow(ds.2019)/nrow(temp)
[1] 12.38926
> # 수진자 남녀 비율
> table(temp$성별코드)
1     2
28056 30028
```

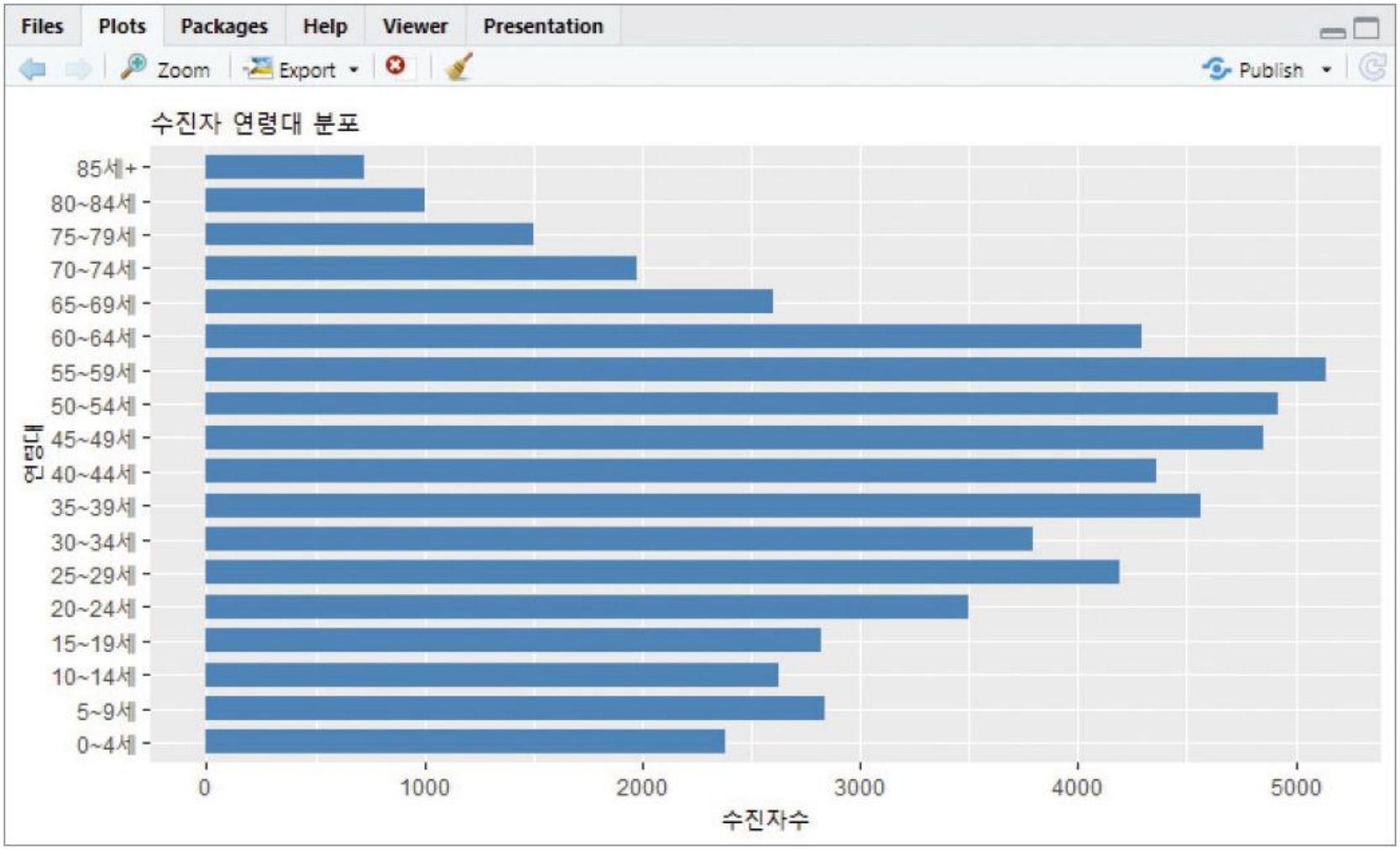
코드 14-1 *(계속)*

```
# 수진자 연령대
age <- aggregate(temp[,'연령대코드'], by=list(
                                연령대코드=temp$연령대코드), length)
age

age.new <- merge(age, age.code, by.x='연령대코드', by.y='코드')
names(age.new)[2] <- '수진자수'
head(age.new)

ggplot(age.new, aes(x=reorder(연령대, 연령대코드), y=수진자수)) +
geom_bar(stat='identity', width=0.7, fill='steelblue') +
   ggtitle('수진자 연령대 분포') +
   labs(x = '연령대') +
   coord_flip()
```

2. 데이터 탐색

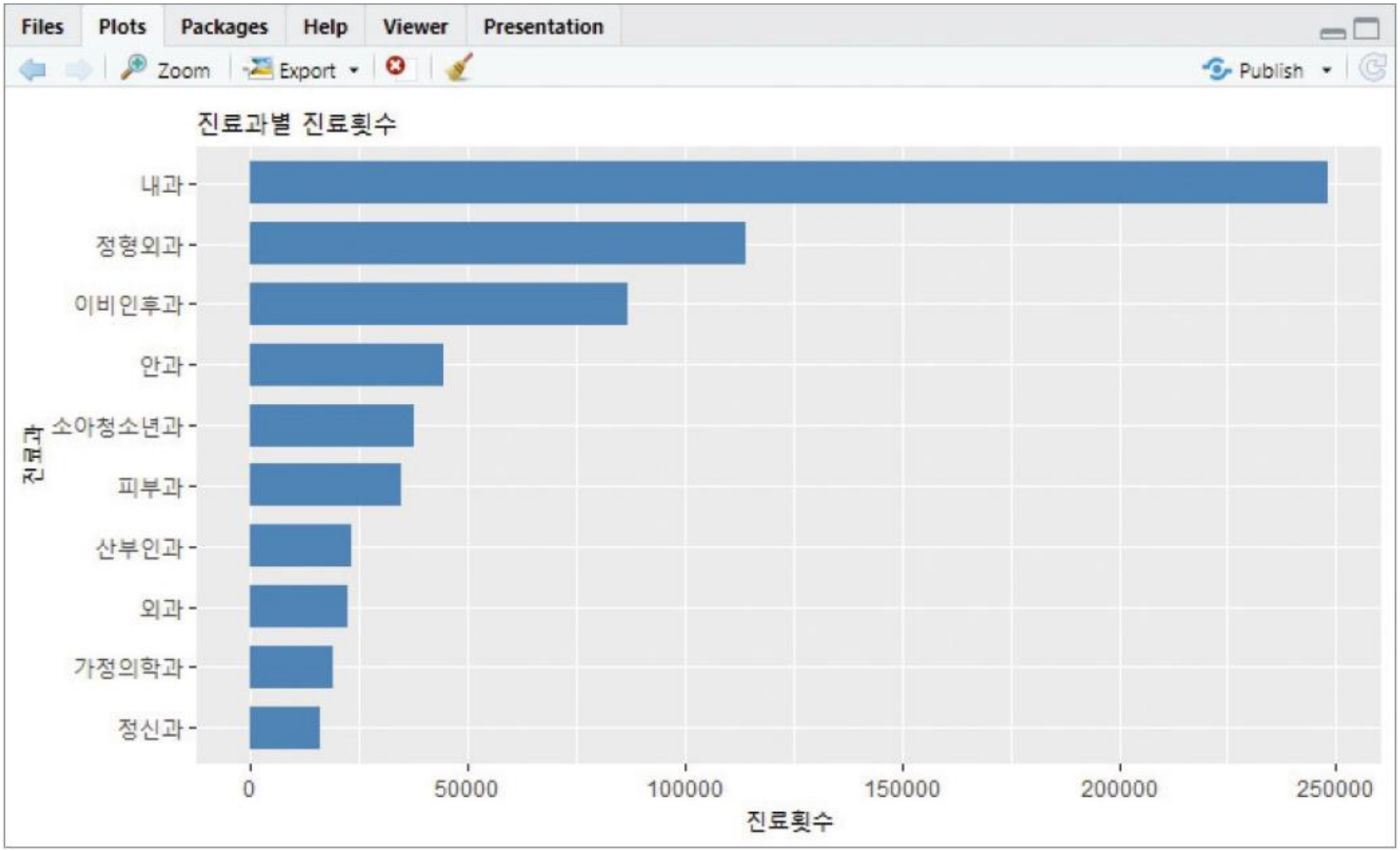

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
수진자 연령대 분포
85세+
80~84세
75~79세
70~74세
65~69세
60~64세
55~59세
50~54세
45~49세
40~44세
35~39세
30~34세
25~29세
20~24세
15~19세
10~14세
5~9세
0~4세
연령대
0
1000
2000
3000
4000
5000
수진자수

코드 14-1 *(계속)*

```
# 진료과별 진료회수 비교 (상위 10개)
dept <- aggregate(ds.2019[,'진료과목코드'], by=list(
                              진료과목코드=ds.2019$진료과목코드), length)
dept.new <- merge(dept, dept.code, by.x='진료과목코드', by.y='코드')
names(dept.new)[2] <- '진료횟수'
dept.new <- dept.new[order(-dept.new$진료횟수),]
head(dept.new)

ggplot(dept.new[1:10, ], aes(x=reorder(진료과, 진료횟수), y=진료횟수)) +
   geom_bar(stat='identity', width=0.7, fill='steelblue') +
   ggtitle('진료과별 진료횟수') +
   labs(x = '진료과') +
   coord_flip()
```

2. 데이터 탐색

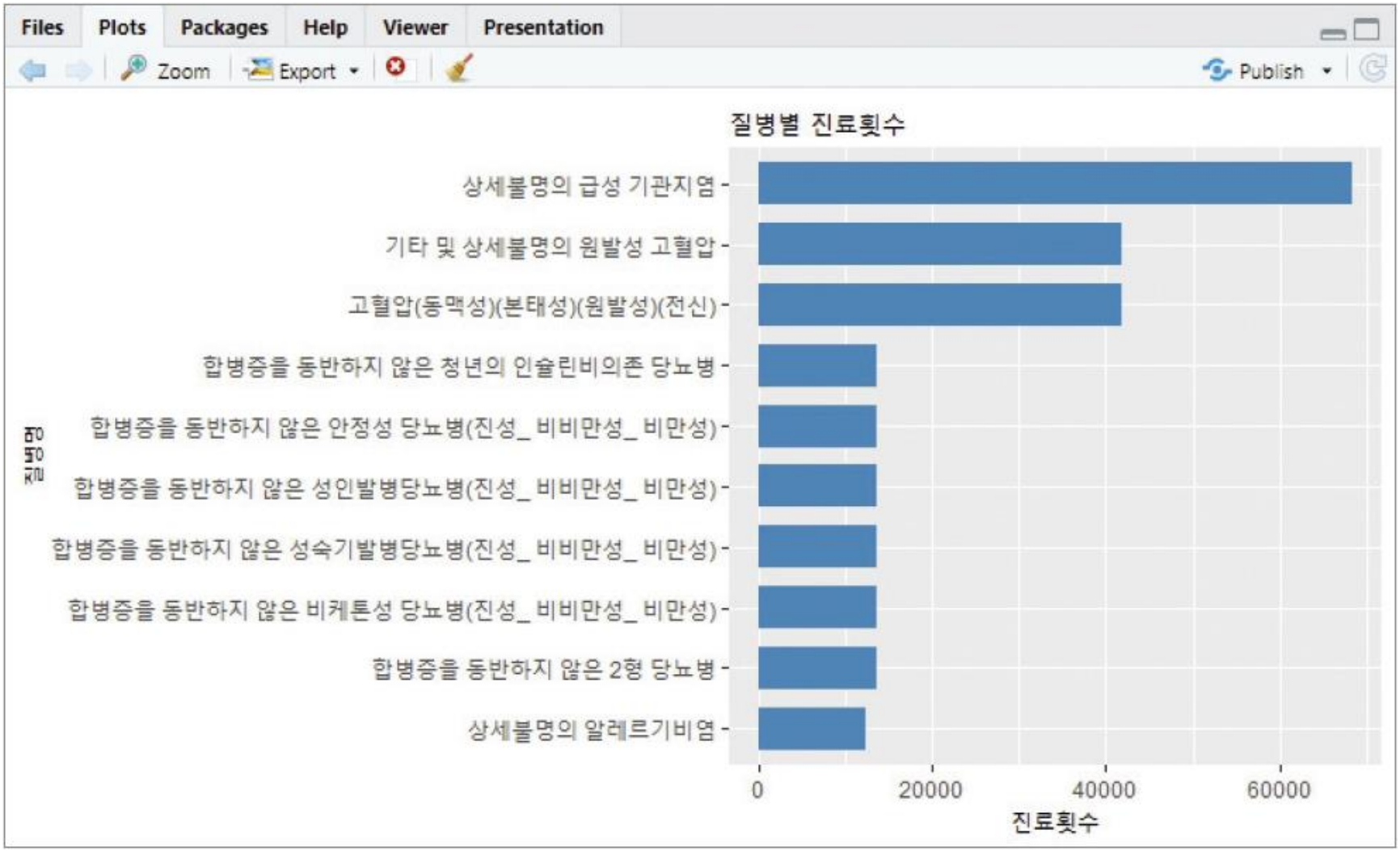

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
진료과별 진료횟수
내과
정형외과
이비인후과
안과
소아청소년과
피부과
산부인과
외과
가정의학과
정신과
진료과
0
50000
100000
150000
200000
250000
진료횟수

코드 14-1 *(계속)*

```
# 질병별 진료회수 비교 (상위 10개)
disease <- aggregate(ds.2019['주상병코드'], by=list(
                                주상병코드=ds.2019$주상병코드), length)
disease.new <- merge(disease, disease.code, by.x='주상병코드',
                             by.y='코드')
names(disease.new)[2] <- '진료횟수'
disease.new <- disease.new[order(-disease.new$진료횟수),]
head(disease.new)

ggplot(disease.new[1:10, ], aes(x=reorder(질병명, 진료횟수), y=진료횟수)) +
  geom_bar(stat='identity', width=0.7, fill='steelblue') +
  ggtitle('질병별 진료횟수') +
  labs(x = '질병명') +
  coord_flip()
```

2. 데이터 탐색

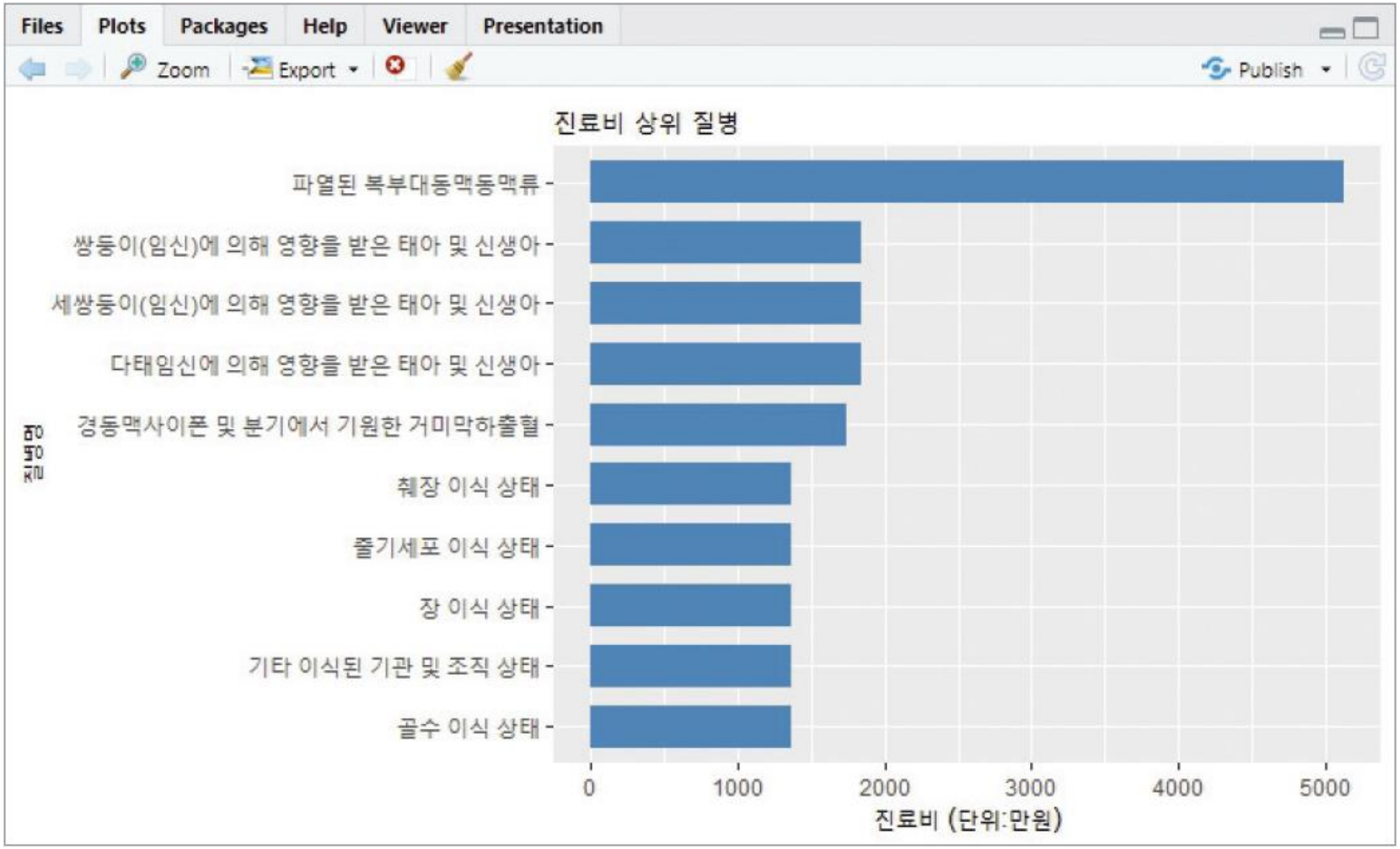

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
질병별 진료횟수
상세불명의 급성 기관지염
기타 및 상세불명의 원발성 고혈압
고혈압(동맥성)(본태성)(원발성)(전신)
합병증을 동반하지 않은 청년의 인슐린비의존 당뇨병
합병증을 동반하지 않은 안정성 당뇨병(진성_ 비비만성_ 비만성)
합병증을 동반하지 않은 성인발병당뇨병(진성_ 비비만성_ 비만성)
합병증을 동반하지 않은 성숙기발병당뇨병(진성_ 비비만성_ 비만성)
합병증을 동반하지 않은 비케톤성 당뇨병(진성_ 비비만성_ 비만성)
합병증을 동반하지 않은 2형 당뇨병
상세불명의 알레르기비염
질병명
0
20000
40000
60000
진료횟수

코드 14-1 *(계속)*

```
# 진료비가 높은 상위 10개 질병
cost <- aggregate(ds.2019[,'심결요양급여비용총액'], by=list(
                                        주상병코드=ds.2019$주상병코드), mean)
cost.new <- merge(cost, disease.code, by.x='주상병코드', by.y='코드')
names(cost.new)[2] <- '진료비'
cost.new <- cost.new[order(-cost.new$진료비),]
cost.new$진료비 <- cost.new$진료비/10000
head(cost.new)

ggplot(cost.new[1:10, ], aes(x=reorder(질병명, 진료비), y=진료비)) +
   geom_bar(stat='identity', width=0.7, fill='steelblue') +
   ggtitle('진료비 상위 질병') +
   labs(x = '질병명', y='진료비 (단위:만원)') +
   coord_flip()
```

2. 데이터 탐색

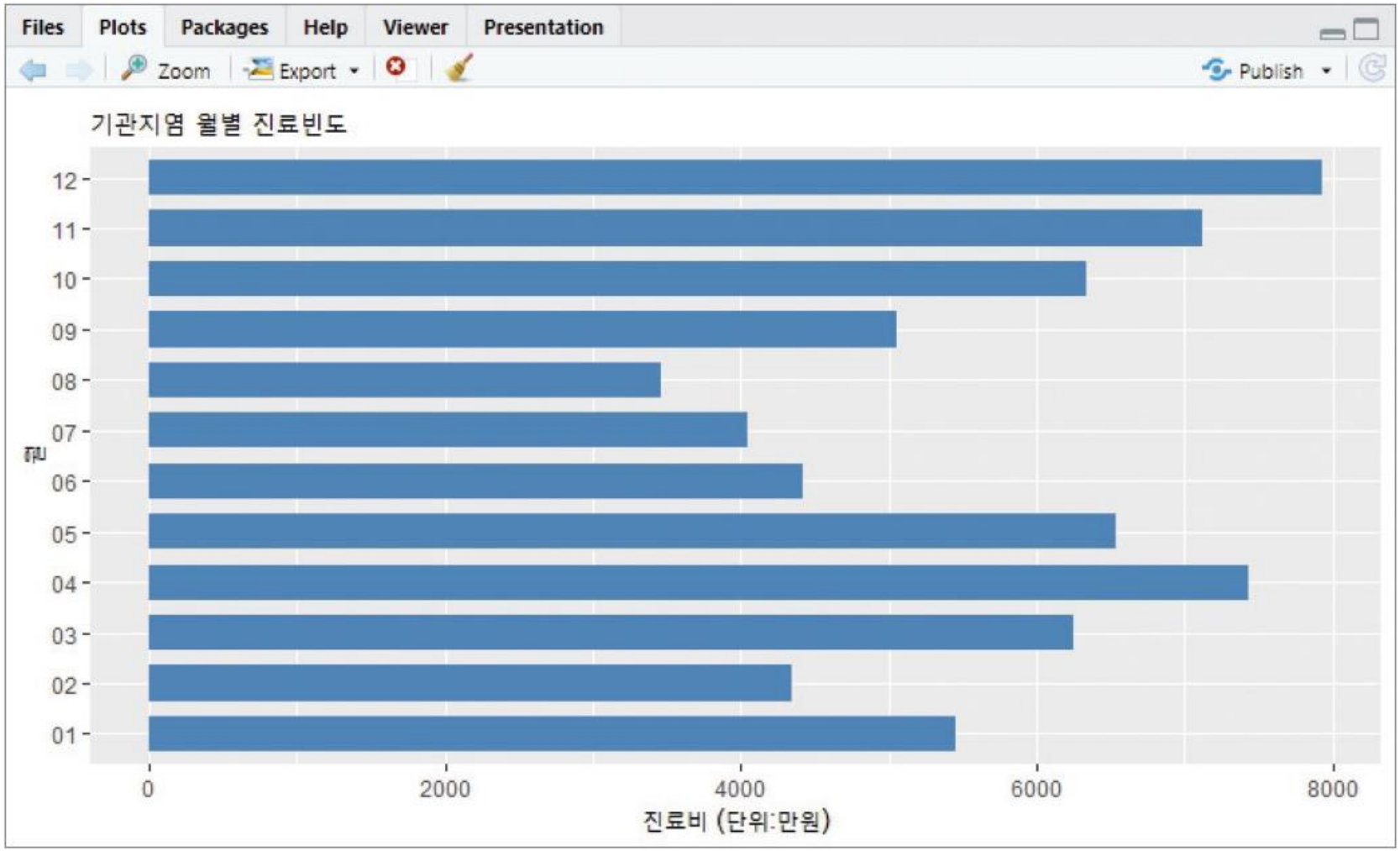

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
진료비 상위 질병
파열된 복부대동맥동맥류
쌍둥이(임신)에 의해 영향을 받은 태아 및 신생아
세쌍둥이(임신)에 의해 영향을 받은 태아 및 신생아
다태임신에 의해 영향을 받은 태아 및 신생아
경동맥사이폰 및 분기에서 기원한 거미막하출혈
췌장 이식 상태
줄기세포 이식 상태
장 이식 상태
기타 이식된 기관 및 조직 상태
골수 이식 상태
질병명
0
1000
2000
3000
4000
5000
진료비 (단위:만원)

코드 14-1 *(계속)*

```
# 기관지염(J209)의 월별 진료 빈도
temp <- ds.2019[ds.2019$주상병코드=='J209', '요양개시일자']
temp <- substr(temp,5,6)
temp.freq <- data.frame(table(temp))
names(temp.freq) <- c('월','진료빈도')
temp.freq

ggplot(temp.freq, aes(x=월,y=진료빈도)) +
geom_bar(stat='identity', width=0.7, fill='steelblue') +
    ggtitle('기관지염 월별 진료빈도') +
    labs(x = '월', y='진료비 (단위:만원)') +
    coord_flip()
```

2. 데이터 탐색


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
기관지염 월별 진료빈도
12
11
10
09
08
07
06
05
04
03
02
01
0
2000
4000
6000
8000
진료비 (단위:만원)

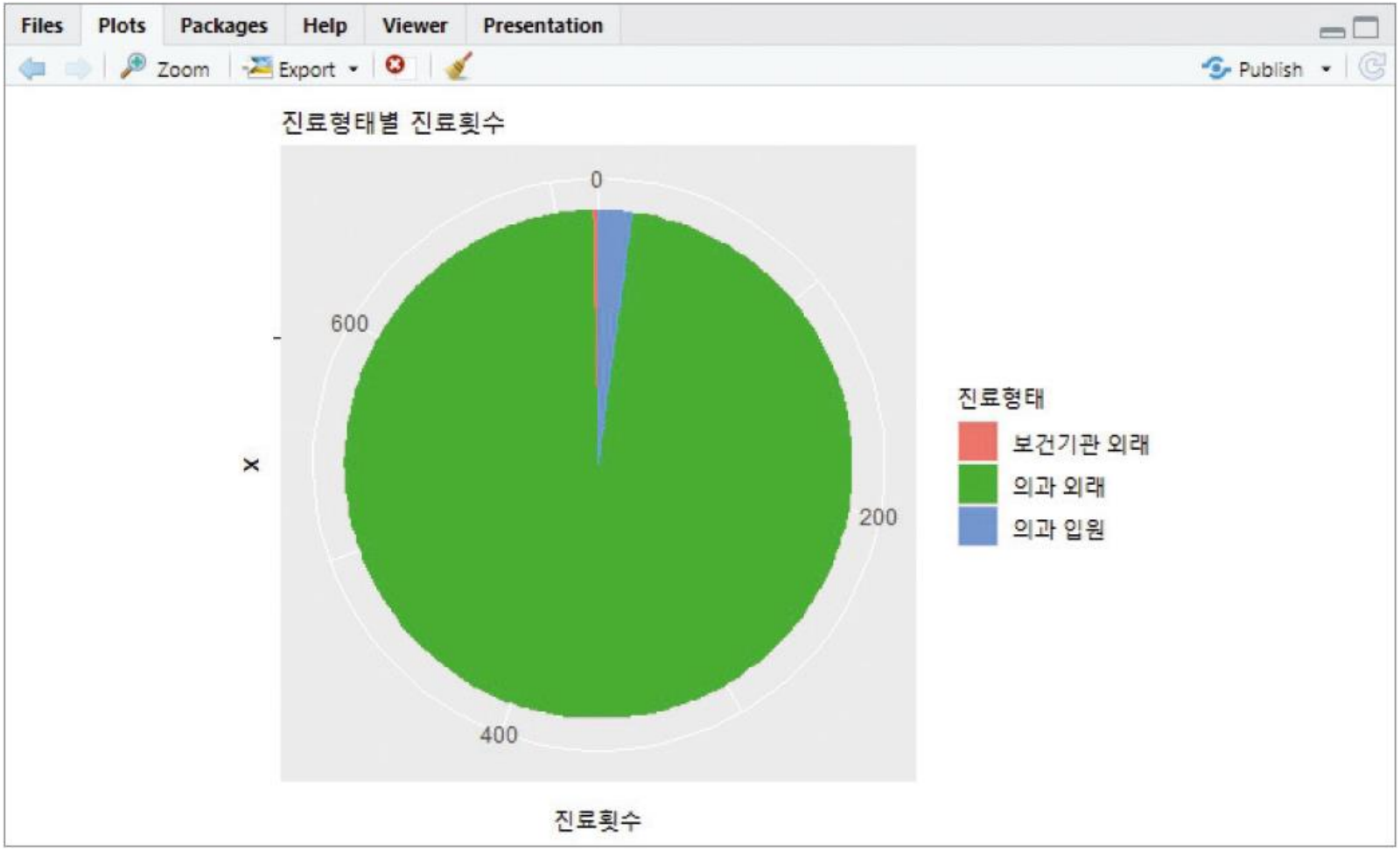
코드 14-1

```
# 진료형태별 진료빈도
treat <- aggregate(ds.2019[,'서식코드'], by=list(
                                        서식코드=ds.2019$서식코드), length)

treat.new <- merge(treat, treat.code, by.x='서식코드', by.y='코드')
names(treat.new)[2] <- '진료횟수'
treat.new$진료횟수 <- treat.new$진료횟수/1000
treat.new

ggplot(treat.new, aes(x="", y=진료횟수, fill=진료형태)) +
    geom_bar(stat="identity", width=1) +
    coord_polar("y", start=0)+
   ggtitle('진료형태별 진료횟수')
```

2. 데이터 탐색


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
진료형태별 진료횟수
0
200
400
600
x
진료횟수
진료형태
보건기관 외래
의과 외래
의과 입원

Section 03

코로나19 전후 비교 분석

코드 14-2 *(계속)*

```
library(ggplot2)
setwd('D:/source')

##################################################
## 데이터셋 준비
##################################################
ds.2019 <- read.csv('NHIS_INCHON_2019.csv')
ds.2020 <- read.csv('NHIS_INCHON_2020.csv')

treat.code <- read.csv('서식코드.csv')
age.code <- read.csv('연령대코드.csv')
disease.code <- read.csv('주상병코드.csv')
dept.code <- read.csv('진료과목코드.csv')

ds.2020$요양개시일자 <- gsub("-", "", ds.2020$요양개시일자)
ds.tot <- rbind(ds.2019, ds.2020)
ds.tot$기준년도 <- factor(ds.tot$기준년도)

dim(ds.tot)
```

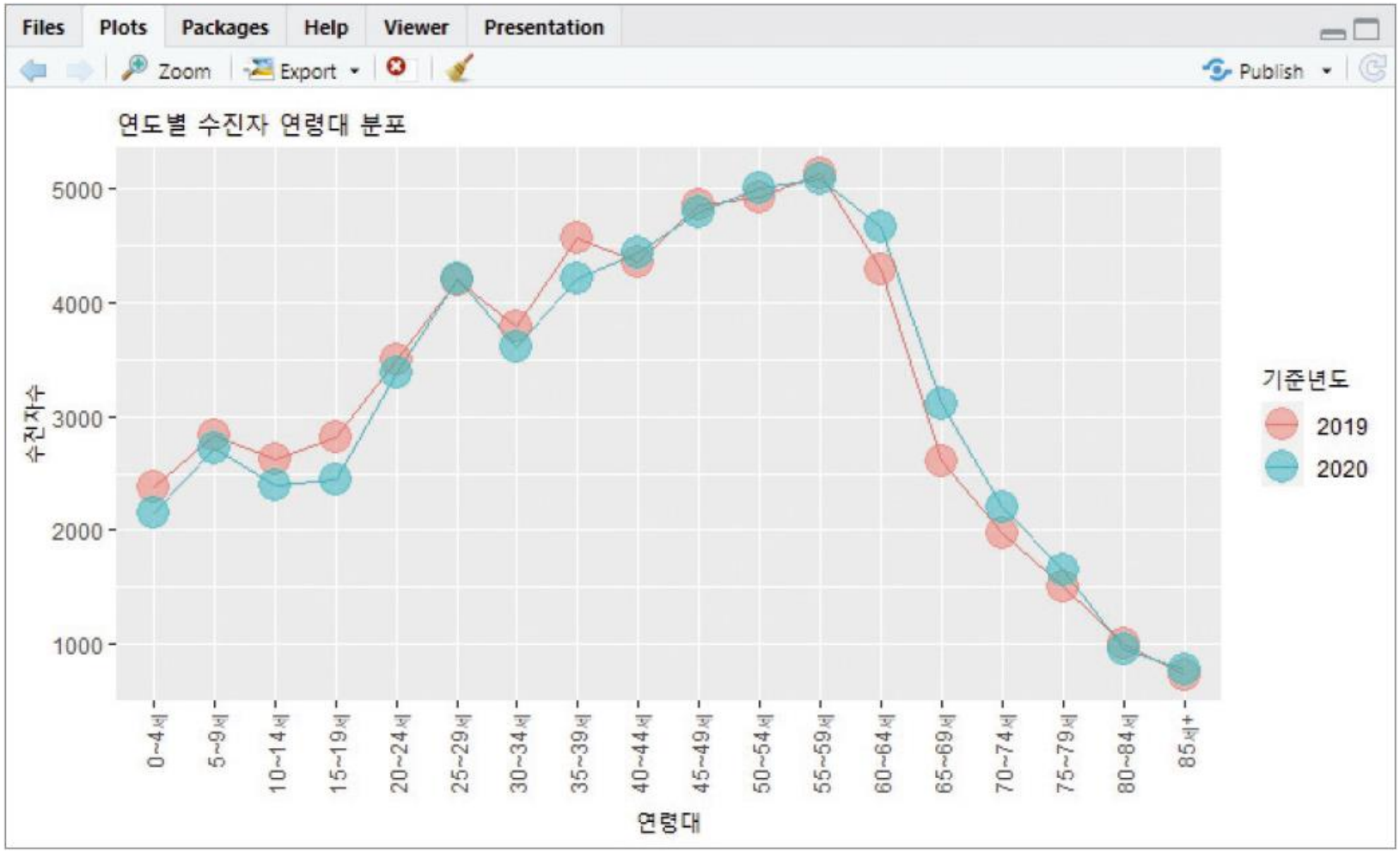
```
> dim(ds.tot)
[1] 1350311 14
```

코드 14-2 *(계속)*

```
# 연도별 연령대별 수진자수 비교
temp <- unique(ds.tot[,1:4])
age <- aggregate(temp[,c('기준년도','연령대코드')],
by=list(기준년도=temp$기준년도,연령대코드=temp$연령대코드),
          length)
age <- age[,1:3]
age.new <- merge(age, age.code, by.x='연령대코드', by.y='코드')
names(age.new)[3] <- '수진자수'
head(age.new)

ggplot(age.new, aes(x=reorder(연령대, 연령대코드), y=수진자수,
   colour=기준년도, group=기준년도)) +
   geom_line() +
   geom_point(size=6, shape=19, alpha=0.5) +
   ggtitle('연도별 수진자 연령대 분포') +
   labs(x = '연령대') +
   theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
            hjust=1))
```

3. 코로나19 전후 비교 분석



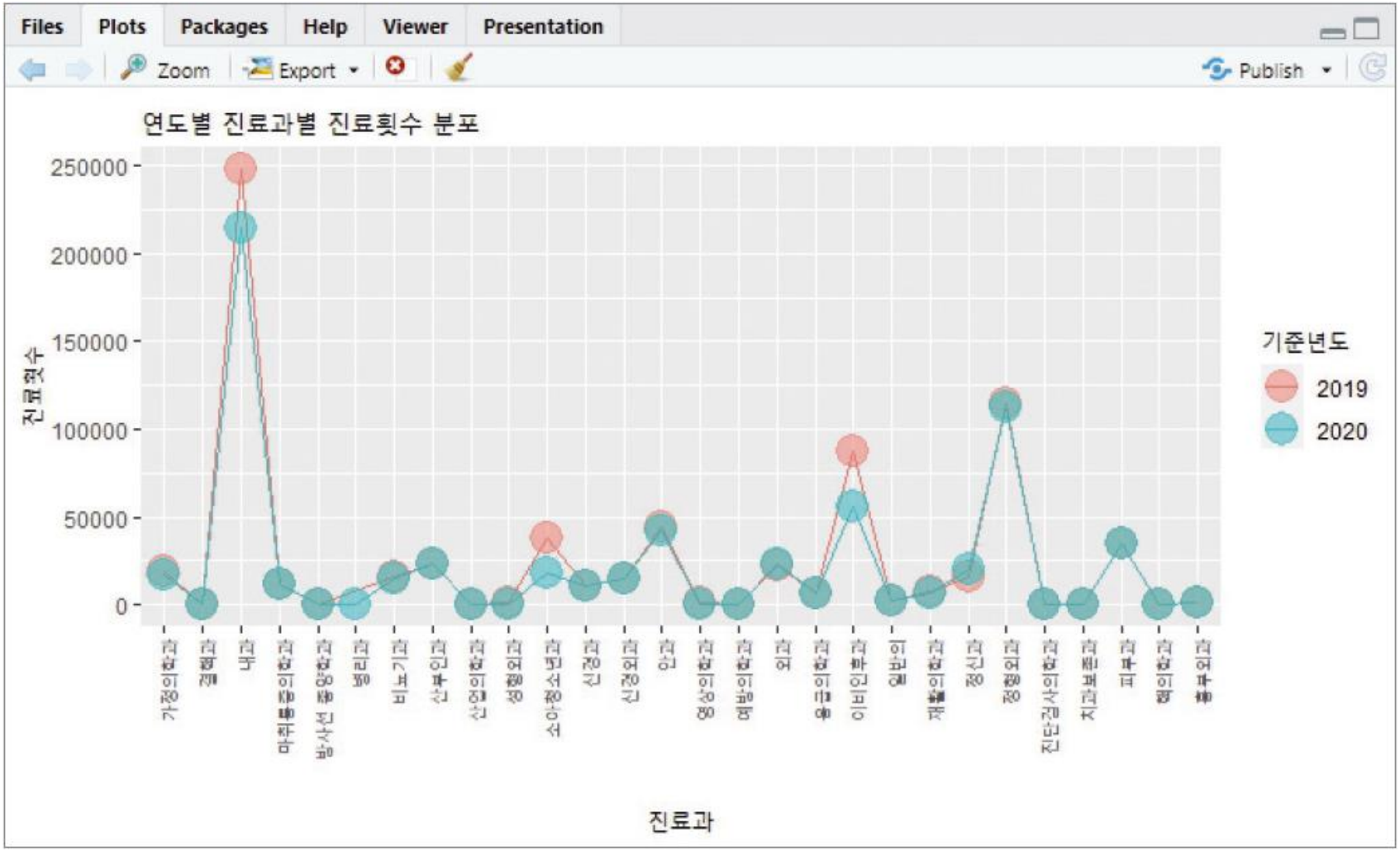
코드 14-2 *(계속)*

```
# 연도별 진료과별 진료횟수 비교
dept <- aggregate(ds.tot[,c('기준년도','진료과목코드')],
                  by=list(기준년도=ds.tot$기준년도,
                          진료과목코드=ds.tot$진료과목코드), length)

dept <- dept[,1:3]
dept.new <- merge(dept, dept.code, by.x='진료과목코드', by.y='코드')
names(dept.new)[3] <- '진료횟수'
head(dept.new)

ggplot(dept.new, aes(x=진료과, y=진료횟수,
   colour=기준년도, group=기준년도)) +
   geom_line() +
   geom_point(size=6, shape=19, alpha=0.5) +
   ggtitle('연도별 진료과별 진료횟수 분포') +
   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

3. 코로나19 전후 비교 분석



코드 14-2

```
# 내과 질환 월별 진료건수 추이 비교
temp <- ds.tot[ds.tot$진료과목코드==1, c('기준년도','요양개시일자')]
temp$month <- substr(temp$요양개시일자,5,6)
temp.agg <- aggregate(temp,
          by=list(기준년도=temp$기준년도,월=temp$month), length)
temp.agg <- temp.agg[,1:3]
names(temp.agg)[3] <- '진료횟수'
head(temp.agg)

ggplot(temp.agg, aes(x=월, y=진료횟수,
   colour=기준년도, group=기준년도)) +
   geom_line() +
   geom_point(size=6, shape=19, alpha=0.5) +
   ggtitle('연도별 월별 내과 진료횟수 분포') +
   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

3. 코로나19 전후 비교 분석

3. 코로나19 전후 비교 분석

그림 14-1 2020년도 월별 코로나19 확진자 통계 © news1

Thank you!

한빛아카데미 Hanbit Academy, Inc.