그래프 그리기 (4) - 파이,산점도 그래프

선 그래프와 막대 그래프 이외에 자주 사용되는 그래프로는 **파이 그래프**와 **산점도 그래프**가 있습니다.

#01. 그래프 작업 준비하기

1) 패키지 로드

```
# 한국 서버를 통해 라이브러리 로드
REPO_URL = "http://healthstat.snu.ac.kr/CRAN/"

# 그래프 패키지
if (!require(ggplot2)) install.packages("ggplot2", repos=REPO_URL)
library(ggplot2)

# 폰트 설정 패키지
if (!require(extrafont)) install.packages("extrafont", repos=REPO_URL)
library(extrafont)

# `%>%`가 적용되는 기능을 사용하고자 할 경우
if (!require("dplyr"))     install.packages("dplyr", repos=REPO_URL)
library(dplyr)
```

2) 한글 사용을 위한 폰트 로드

나눔고딕 검색하기

아래 구문 실행시 y/n를 묻는 화면이 표시된다. y를 입력해야 실행된다.

시간이 다소 오래 소요되며 장치별로 1회만 수행하면 된다.

```
font_import(pattern = 'NanumGothic.ttf')
```

▶ 출력결과

```
Scanning ttf files in C:\WINDOWS\Fonts ...
Extracting .afm files from .ttf files...
C:\Windows\Fonts\NanumGothic.ttf : NanumGothic already registered in fonts database. Skipping.
Found FontName for 0 fonts.
Scanning afm files in D:/leekh/R-3.6.1/library/extrafontdb/metrics
```

설치된 폰트 목록 확인

fonttable() 함수를 통해 반환받는 DataFrame에서 FamilyName 컬럼을 확인하면 R 소스코드에 적용해야 할 폰트 이름을 확인할 수 있다.

출력결과는 시스템에 설치되어 있는 글꼴의 상태에 따라 다를 수 있다.

> `unique()` 함수는 파라미터로 전달된 항목들에 대해서 중복을 제거한 결과를 리턴한다.

```
ftable <- fonttable()
unique(ftable$FamilyName)
```

▶ 출력결과

```
1. 'NanumGothic'
```

설치된 폰트들 로드하기

```
# mac의 경우 `device="win"` 생략
loadfonts(device="win")
# loadfonts()
```

▶ 출력결과

```
NanumGothic already registered with windowsFonts().
NanumGothicExtraBold already registered with windowsFonts().
```

3) 샘플 데이터 가져오기

학생들의 나이와 좋아하는 계절 조사 자료 (임의의 예시 데이터)

```
설문 <- read.csv("http://itpaper.co.kr/demo/r/season.csv", stringsAsFactors=F, fileEncoding="euc-kr")
설문
```

▶ 출력결과

A data.frame: 500 × 3

이름	계절	나이
<chr>	<chr>	<int>
학생1	봄	15
학생2	봄	17
학생3	여름	18
학생4	겨울	19
학생5	가을	15
학생6	봄	16
학생7	여름	19
학생8	여름	18
학생9	여름	17
학생10	여름	17
학생11	가을	19
학생12	가을	19
학생13	여름	18
학생14	가을	19
학생15	겨울	15
학생16	겨울	16
학생17	봄	16
학생18	여름	18
학생19	여름	17
학생20	봄	18
학생21	겨울	19
학생22	가을	15

이름	계절	나이
<chr>	<chr>	<int>
학생23	가을	17
학생24	여름	18
학생25	봄	19
학생26	여름	15
학생27	겨울	16
학생28	겨울	19
학생29	가을	18
학생30	가을	17
…	…	…
학생471	가을	19
학생472	여름	15
학생473	봄	15
학생474	여름	17
학생475	겨울	18
학생476	겨울	19
학생477	가을	15
학생478	가을	16
학생479	가을	19
학생480	봄	18
학생481	봄	17
학생482	여름	17
학생483	겨울	19
학생484	가을	19
학생485	봄	18
학생486	여름	19
학생487	여름	15
학생488	여름	17
학생489	여름	18
학생490	가을	19
학생491	가을	15
학생492	여름	16
학생493	가을	19
학생494	겨울	18
학생495	겨울	17
학생496	봄	17
학생497	여름	19

이름	계절	나이
<chr>	<chr>	<int>
학생498	여름	19
학생499	봄	18
학생500	겨울	19

#02. 파이 그래프

데이터프레임의 특정 컬럼 안에서 전체를 100%로 봤을 때 얼만큼의 비중을 차지하는지를 시각화 한 자료.

1) 각 계절별로 빈도수 검사

파이 그래프를 표현하기 위해서는 각 데이터별로 빈도수를 확인하는 분석이 수행되어야 한다.

```
df <- data.frame(table(설문$계절))
df
```

▶ 출력결과

A data.frame: 4 × 2

Var1	Freq
<fct>	<int>
가을	134
겨울	101
봄	91
여름	174

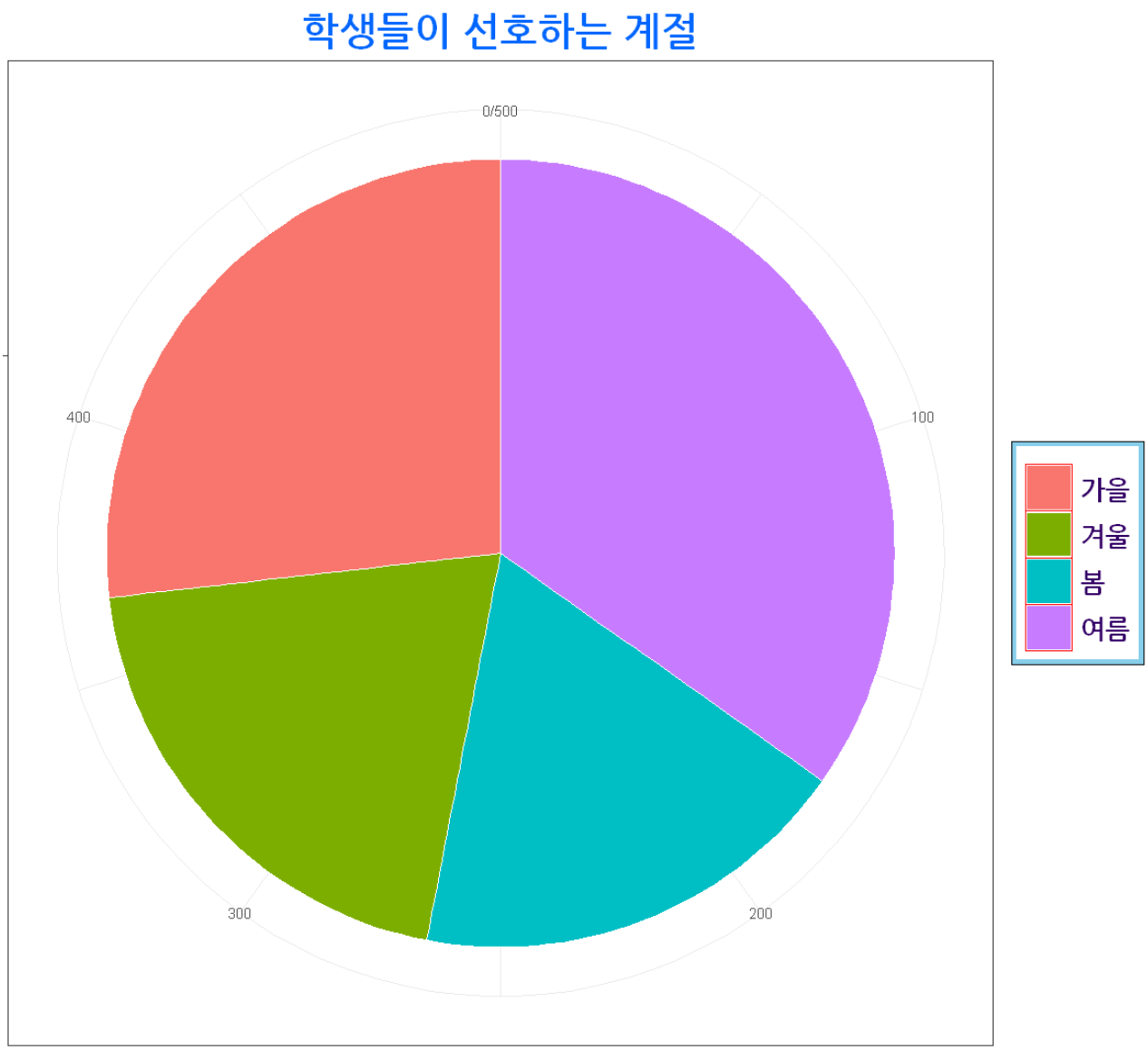
2) 파이 그래프 시각화

```
options(repr.plot.width=10, repr.plot.height=10, warn=-1)
```

```
ggplot(data=df) +
    # x축은 지정 안함, y축은 값의 종류, fill은 각 종류별 빈도
    geom_col(aes(x="", y=Freq, fill=Var1), width = 1, color = "white") +
    # y축을 기준으로 회전시킴 --> 파이그래프 표현
    coord_polar('y', start = 0) +
    # 배경을 흰색으로 설정
    theme_bw() +
    # 그래프 타이틀 설정
    ggtitle("학생들이 선호하는 계절") +
    # x축 제목 설정 --> 표시안함을 위해 빈 문자열 설정
    xlab("") +
    # y축 제목 설정 --> 표시안함을 위해 빈 문자열 설정
    ylab("") +
    theme(
        # 각 텍스트의 색상, 크기, 각도, 글꼴 설정
        plot.title=element_text(family="NanumGothic", color="#0066ff", size=24, face="bold", hjust=0.5),
        # 범주의 제목 표시 안함
        legend.title = element_blank(),
        # 범주의 각 항목별 텍스트
        legend.text=element_text(family="NanumGothic", face="bold", size=16, color="#330066"),
        # 범주를 구분하는 각 색상 박스에 대한 테두리와 배경
        legend.key=element_rect(color="red", fill="white"),
        # 범주의 배경상자 색상 설정
        legend.box.background = element_rect(fill="skyblue"),
```

```
        # 범주의 배경상자 여백
        legend.box.margin = margin(3, 3, 3, 3),
        # 범주를 구분하는 각 색상 박스에 대한 크기
        legend.key.size = unit(1,"cm"))
```

▶ 출력결과



#02. 산점도 그래프

두 변수 간의 영향력을 보여주기 위해 가로 축과 세로 축에 데이터 포인트를 그리는 그래프.

포인터들이 오밀조밀 뭉쳐 있으면 두 변수는 서로 관련성 정도가 높고 흩어져 있으면 관련성이 낮고 분석한다.

> 예) 여름철 온도와 아이스크림 판매량의 상관관계 분석

1) 산점도 그래프의 의미 -> 상관 관계

- 산점도에서 사용되는 두 변수 간의 관계
- 그래프에 표시되는 마커들의 배열이 직선에 가까운 경우 두 변수의 상관 관계가 높다.
- 마커가 산점도에 균등하게 분산되는 경우 상관 관계가 낮거나 0이다.
- **상관관계의 유형**

i. 정의관계 : x가 증가할 때 y도 증가
ii. 역의관계 : x가 증가할 때 y는 감소
iii. 선형관계 : 직선에 가까운 배치
iv. 비선형관계 : 곡선에 가까운 배치

2) 샘플 데이터 가져오기

```
교통사고 <- read.csv("http://itpaper.co.kr/demo/r/traffic.csv", stringsAsFactors=F, fileEncoding="euc-kr")
교통사고
```

▶ 출력결과

A data.frame: 168 × 5

년도	월	발생건수	사망자수	부상자수
<int>	<int>	<int>	<int>	<int>
2005	1	15494	504	25413
2005	2	13244	431	21635
2005	3	16580	477	25550
2005	4	17817	507	28131
2005	5	19085	571	29808
2005	6	18092	476	28594
2005	7	18675	528	29984
2005	8	19035	562	31603
2005	9	18759	577	29831
2005	10	19757	639	31597
2005	11	19129	574	30337
2005	12	18504	530	29750
2006	1	14971	420	24533
2006	2	14270	373	22903
2006	3	16767	465	26013
2006	4	17948	469	28725
2006	5	19140	531	30279
2006	6	17435	455	27032
2006	7	18634	516	29978
2006	8	18794	585	30882
2006	9	19293	580	30186
2006	10	19100	651	30715
2006	11	19877	701	31270
2006	12	17516	581	27713
2007	1	14914	468	23975
2007	2	14696	446	23717
2007	3	18166	476	28811

년도	월	발생건수	사망자수	부상자수
<int>	<int>	<int>	<int>	<int>
2007	4	18055	460	28555
2007	5	19264	516	30532
2007	6	18310	538	28662
...	...	...	...	...
2016	7	18955	358	28586
2016	8	18398	336	28017
2016	9	17883	375	26761
2016	10	19918	440	29635
2016	11	19234	416	28520
2016	12	18869	408	28192
2017	1	16970	353	26099
2017	2	14832	280	22323
2017	3	17047	295	25046
2017	4	17717	293	26530
2017	5	18502	366	27268
2017	6	18047	315	26454
2017	7	18158	357	27362
2017	8	18682	353	28162
2017	9	19891	419	29371
2017	10	18863	420	28698
2017	11	19377	379	28472
2017	12	18249	355	27044
2018	1	17026	304	25438
2018	2	16208	275	24630
2018	3	17022	310	25015
2018	4	17992	303	26643
2018	5	18636	309	27834
2018	6	18082	266	26574
2018	7	18699	315	28104
2018	8	18335	357	27749
2018	9	18371	348	27751
2018	10	19738	373	28836
2018	11	19029	298	28000
2018	12	18010	323	26463

3) 교통사고 발생건수와 부상자 수의 상관관계 분석

```
options(repr.plot.width=20, repr.plot.height=12, warn=-1)
```

```
ggplot(data=교통사고) +
    geom_point(aes(x=발생건수, y=부상자수), size=3, colour='blue') +
    # 배경을 흰색으로 설정
    theme_bw() +
    # 그래프 타이틀 설정
    ggtitle("교통사고 발생건수와 부상자 수의 상관관계") +
    # x축 제목 설정 --> 표시안함을 위해 빈 문자열 설정
    xlab("발생건수") +
    # y축 제목 설정 --> 표시안함을 위해 빈 문자열 설정
    ylab("부상자수") +
    # 각 텍스트의 색상, 크기, 각도, 글꼴 설정
    theme(plot.title=element_text(family="NanumGothic", color="#0066ff", size=25, face="bold"),
          axis.title.x=element_text(family="NanumGothic", color="#999999", size=18, face="bold"),
          axis.title.y=element_text(family="NanumGothic", color="#999999", size=18,face="bold", hjust=1),
          axis.text.x=element_text(family="NanumGothic", color="#000000", size=16, angle=45),
          axis.text.y  = element_text(family="NanumGothic", color="#000000", size=16, angle=45))
```

▶ 출력결과