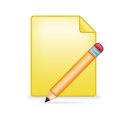
**세종대학교 데이터사이언스**

**확률통계 프로그래밍**

**중간고사 대체 온라인 과제**

|  |  |
| --- | --- |
| **담당교수** | **안수엽 교수님** |
| **성 명** | **정소윤** |
| **제 출 일** | **2021. 4 . 28 .** |





중간고사

R에서는 기본 내장 데이터 셋을 포함해서 각 패키지마다 내장된 데이터 셋을 제공한다.

현재 **410개의 패키지**에서 제공하는 **3천개가 넘는 데이터 셋**에 대한 패키지 명, 데이터 셋 명칭, 데이터 셋 주제(Topic), 데이터 행(Row)의 수, 데이터 열(Column)의 수에 대한 정보가 포함되어 있는 사이트이다.

[https://vincentarelbundock.github.io/Rdatasets/datasets.html](http://vincentarelbundock.github.io/Rdatasets/datasets.html)

# 데이터 셋 제공 사이트

[https://r-dir.com/reference/datasets.html](http://r-dir.com/reference/datasets.html)

datasets 패키지에 어떤 데이터 들이 있는지 확인하기 위하여 data() 함수를 이용한다.

> data( )

각 데이터 셋의 정보를 확인하려면 help(데이터 셋 명)을...

> help(데이터 셋 명)

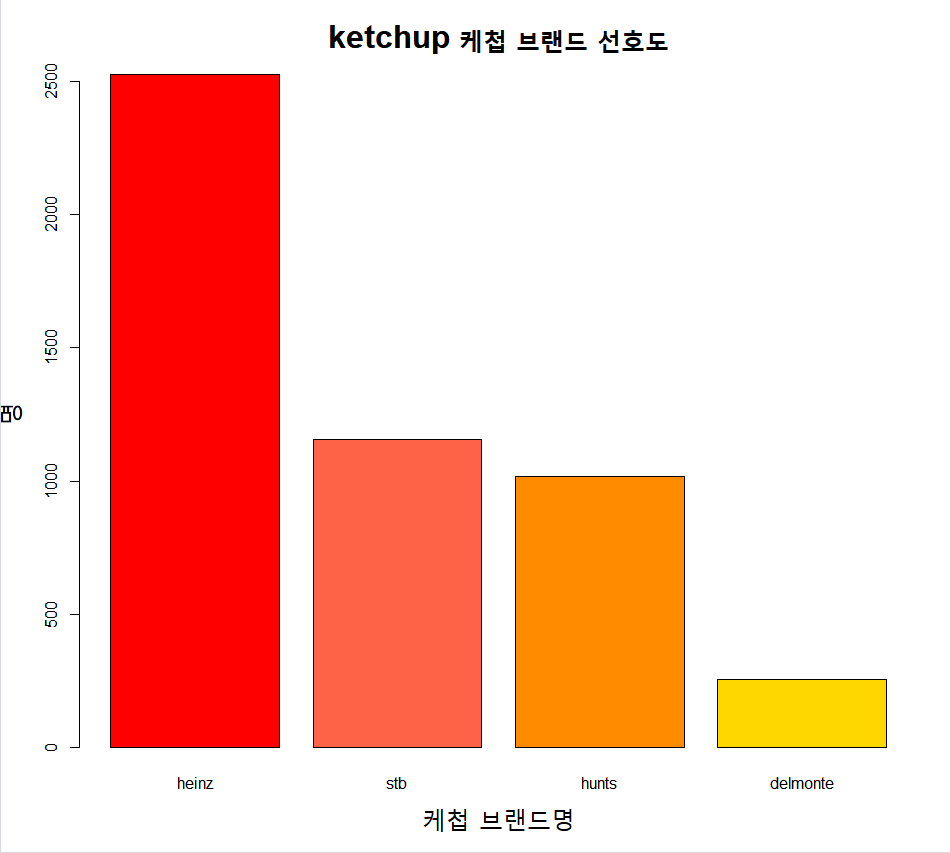
**[문제1] (15점)** 여러분이 임의로 선정한 패키지(package)에 내장되어 있는 데이터를 활용하여 다음의 내용을 **R**로 작성하여라. (기존에 많이 알려지지 않은 데이터 활용일수록 높은 점수)

1. 막대그래프(Bar Chart)를 예쁘게 작성하여라.

“Ecdat” 패키지의 Ketchup 데이터셋을 활용하였습니다.

이 데이터셋은 미국의 4956명에게 케첩 브랜드에 대한 선호조사 결과 데이터입니다.

저는 케첩 브랜드 선호도인 ketchup.choice 열을 선정하여 어떤 브랜드가 얼마만큼의 선호도를 가지고 있는지 막대그래프로 표현해보았습니다.

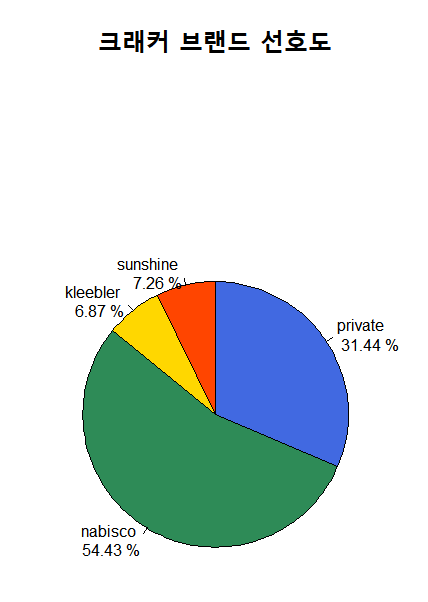
heinz, stb, hunts, delmonte순으로 선호도를 많이 얻고 있었습니다.

1. 원형그래프(Pie Chart)를 예쁘게 꾸며 보세요.

“Ecdat” 패키지의 Cracker 데이터셋을 활용하였습니다.

이 데이터셋은 미국의 3292명에게 크래커 브랜드에 대한 선호조사 결과 데이터입니다.

저는 크래커 브랜드 선호도인 choice 열을 선정하여 어떤 브랜드가 얼마만큼의 선호도를 가지고 있는지 원형그래프로 표현해보았습니다. 좀 더 정확하게 비교하기 위해 백분율도 더했습니다.

navisco이 반 이상을 차지하였고 다음으로 private, kleebler, sunshine순으로 선호도를 차지하였습니다.

**[문제2] (15점)** 여러분이 임의로 선정한 패키지(package)에 내장되어 있는 데이터를 활용하여 다음의 내용을 **R**로 작성하여라. (기존에 많이 알려지지 않은 데이터 활용일수록 높은 점수)

1. **줄기-잎 그림**(Stem-and-Leaf plot)을 나타내어라.

“DAAG” 패키지의 carprice 데이터셋을 활용하였습니다.

carprice는 미국의 차들의 가격 데이터로 이 중 베이직 모델의 가격인 Min.Price, 중간 모델 가격인 Price, 프리미엄 모델 가격인 Max.Price를 활용하였습니다. 가격의 단위는 $1000입니다. 이 가격들의 빈도와 분포를 알아보고자 줄기 잎 그림을 그려보았습니다.

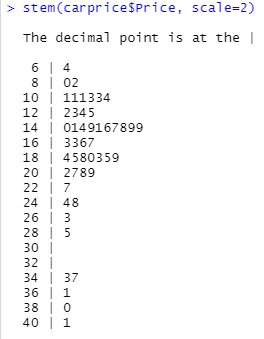
테이블이(가) 표시된 사진

자동 생성된 설명테이블이(가) 표시된 사진

자동 생성된 설명

먼저, 베이직 모델 가격인 Min.Price를 scale을 2로 하여 줄기 잎 그림으로 표현해 보았습니다. 20000달러를 넘는 차는 거의 드물었고 대부분 8000달러에서 15000달러 사이에 가장 많이 분포하였습니다. 더 자세한 분포를 알고 싶어 scale을 4로 하고 그려보았습니다. 그 결과 베이직 모델의 차는 14000달러 영역에 가장 많이 분포함을 알 수 있었습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

다음으로는 중간 모델 가격인 Price를 scale을 2로 하여 줄기 잎 그림으로 표현해 보았습니다. 중간 모델의 가격은 대부분 10000달러에서 21000달러 사이에 많이 분포하였습니다.

8000달러 이하 저렴한 가격의 차는 베이직 모델에 비해 드물었고 20000달러를 넘는 차도 있었음을 알 수 있었습니다.

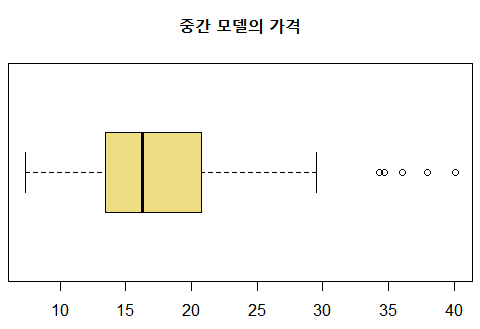
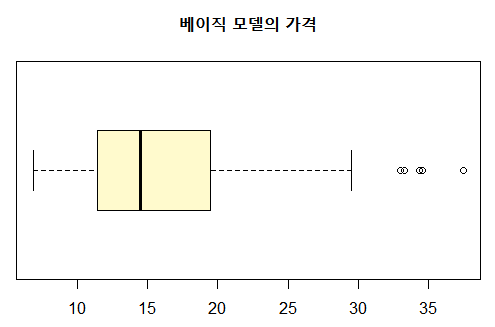
scale을 4로도 표현해보았는데 15000달러 후반에 가장 많이 분 포함을 알 수 있었습니다.

테이블이(가) 표시된 사진

자동 생성된 설명테이블이(가) 표시된 사진

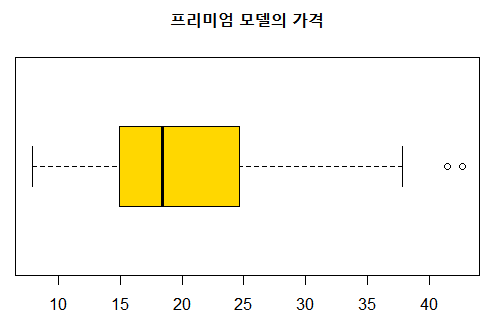
자동 생성된 설명

마지막으로 프리미엄 모델 가격인 Max.Price를 scale을 2로 하여 줄기 잎 그림으로 표현해 보았습니다. 다른 모델들 과는 다르게 40000달러를 넘는 초고가의 차들도 있었으며 대부분 14000달러에서 22000달러 사이에 많이 분포하였습니다. scale을 4로 해보니 21000달러 영역에 가장 많이 분포함을 알 수 있었습니다.

****(2) **상자그림**(Box Plot)을 나타내어라.

위에서 줄기 잎 그림으로 표현했던 carprice 데이터 셋을 활용하였습니다.

Min.Price, Price, Max.Price 순으로 제1사분위수, 제3사분위수와 중앙값이 점점 오른쪽으로 치우쳐져 있음을 알 수 있습니다. 그리고 세 모델 모두 안쪽울타리 경계를 벗어나는 이상점이 모두 오른쪽에 있습니다.

그리고 사분위수의 범위는 Max.Price가 가장 길었고, Price가 가장 짧은 것으로 보아 프리미엄 모델의 가격이 가장 넓게 퍼져있음을 알 수 있었습니다. 세 모델 모두 중앙값과 제 3분위수사이의 거리가 중앙갑과 제 1분위수 사이 거리보다 긴 것으로 보아 가격이 저렴한 쪽으로(자료가 왼쪽으로)치우친 것을 알 수 있었습니다.

(3) **히스토그램(histogram)**을나타내어라.

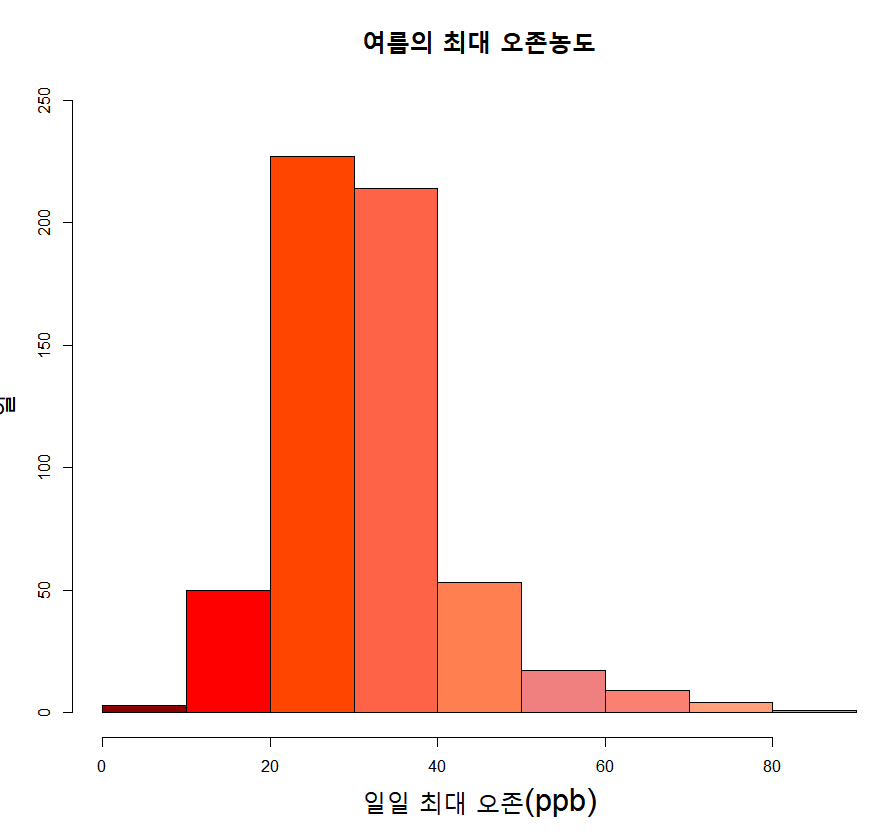
“texmex” 패키지의 summer와 winter 데이터셋을 활용하였습니다.

1994년부터 1998년까지 여름은 578일, 겨울은 532일동안 대기오염도 데이터를 모아놓은 것입니다. 여름의 대기오염도는 4월부터 7월동안, 겨울은 11월에서 1월동안 측정하였습니다.

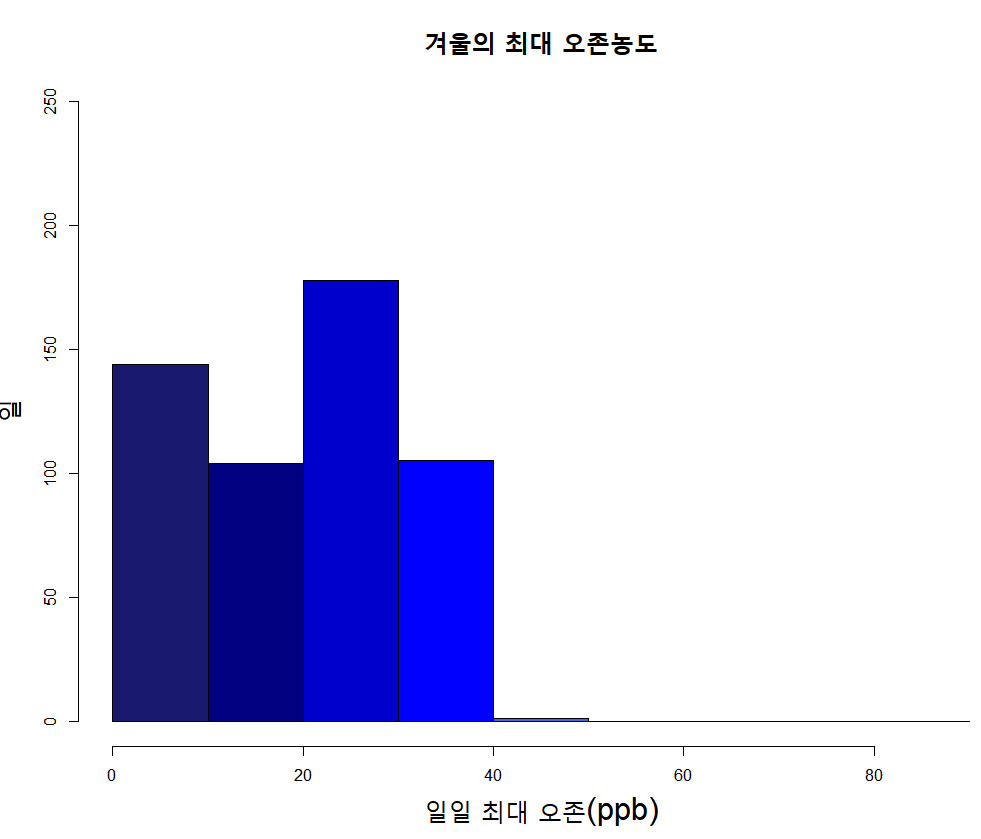
그 중 오존농도를 측정한 데이터를 선정하여 최대 오존 농도를 계급으로 나눴습니다.

그리고 계급에 해당되는 오존 농도를 가진 일이 며칠인지를 알아보려 이를 히스토그램 형태로 그려내었습니다.

그 결과 여름에는 일일 최대 오존 농도가 20에서 40ppb에 몰려 있었고 그보다 더 높은 오존 농도가 측정된 날도 드물게 있었음을 알 수 있었습니다.

****

그에 비해 겨울에는 일일 최대 오존 농도가 0에서 40ppb에 고루 분포해 있었고 그보다 더 높은 오존 농도는 거의 측정되지 않았습니다.

****

결과적으로 여름이 비교적 겨울보다 높은 일일 최대 오존 농도를 가진 날이 많았음을 알 수 있었습니다.

**[문제3] (70점) (자유주제) 지금까지 R을 학습한 내용 중에서 하나의 테마(주제, 통계자료)를 선정하여 심화 응용, 또는 확장된 내용을 작성하시오.**

1. 데이터 파일은 **정부 부처, 공공기관, 국제기구(적극 권장)** **등**에서 발표된 (되도록)최신 자료를 이용하며, 자료의 내용에 제한이 없다. 단, 데이터에 대한 자료출처를 반드시 밝힐 것(URL 표시)이며, **코로나 관련 주제는 일체 인정하지 않음**)

2. R을 구현할 때, 함수의 속성(옵션)에 대한 추가적인 사항에 따라 단계별로 작성하시오. 예를 들면, 표현하고자 하는 대상을 꾸며가는 단계가 있다면 그 과정을 설명까지 덧붙여 작성하라는 뜻

3. **작성방법과 분량도 제한이 없으며**, 불필요한 메일 질문은 감점!

과제(보고서)란 특별한 형식이 없으며, 단지 과학적 사실을 바탕으로 논리적 전개하는 과정을 작성하는 것입니다. 여기에 덧붙인다면 자신이 타인과 다르다는 것을 알리기는 것이 바로 창의성입니다.

4. 제출방법

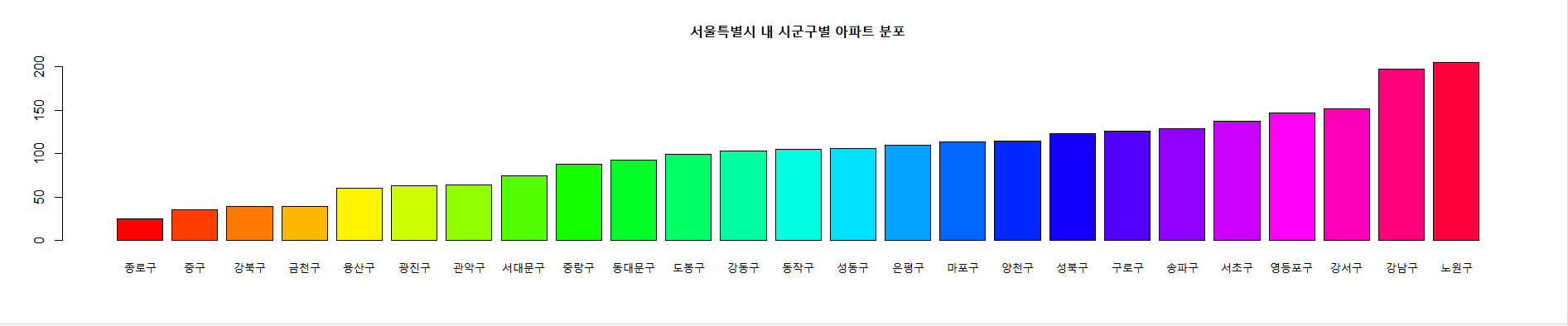
(1) 제출파일(제공된 표지에 작성한 한글 파일, R파일, 데이터 파일 등)을 압축파일로 블랙보드에 업로드 하시길 바랍니다.

(2) 파일명: 성명\_학번.zip

(3) 제출기한: 4월 28일(수요일) 오후11시59분59초까지(그 이후의 제출은 감점)

(4) 세상에 데이터는 무한(?)이다. 같은 내용, 같은 생각을 하는 학생이 있을까?

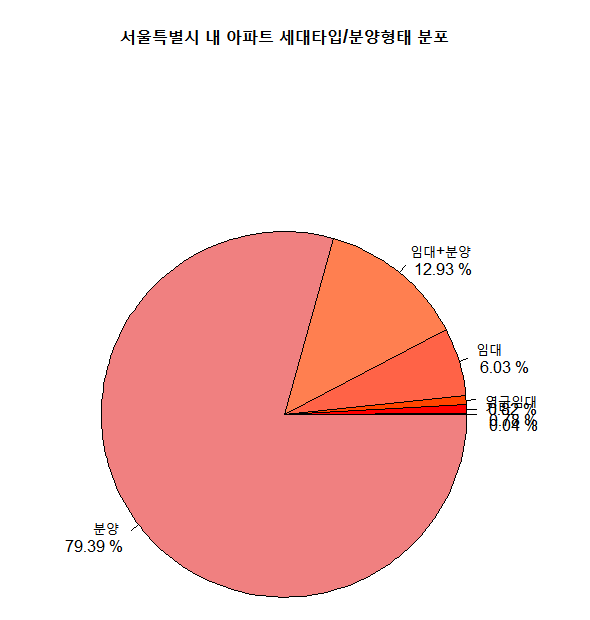
**<서울특별시 내 아파트 데이터 조사>**

1. **시군구별 아파트 분포**

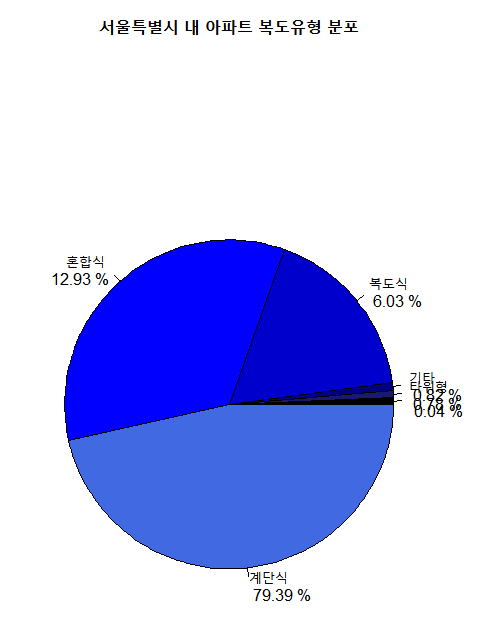
시군구별로 아파트 단지가 얼마나 분포되어 있는지를 파악하기 위해 아파트들의 위치 열을 골라 막대그래프를 그렸습니다.

그 결과 종로구, 중구, 강북구, 금천구에 50개 이하로 아파트단지가 적게 분포하고, 노원구, 강남구에 아파트 단지가 150개 이상으로 비교적 많이 분포함을 알 수 있었습니다.

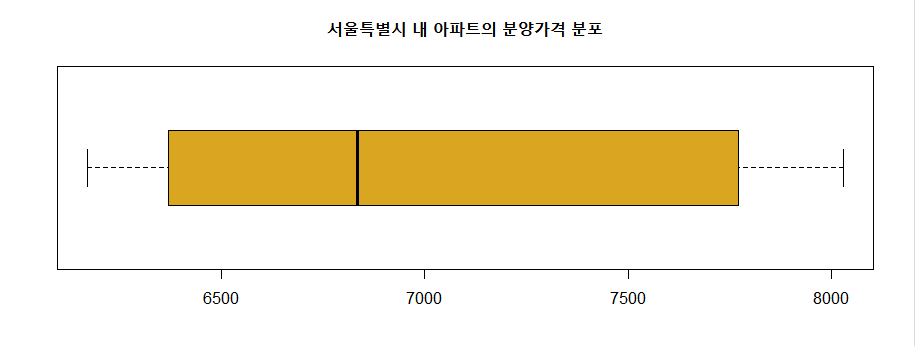
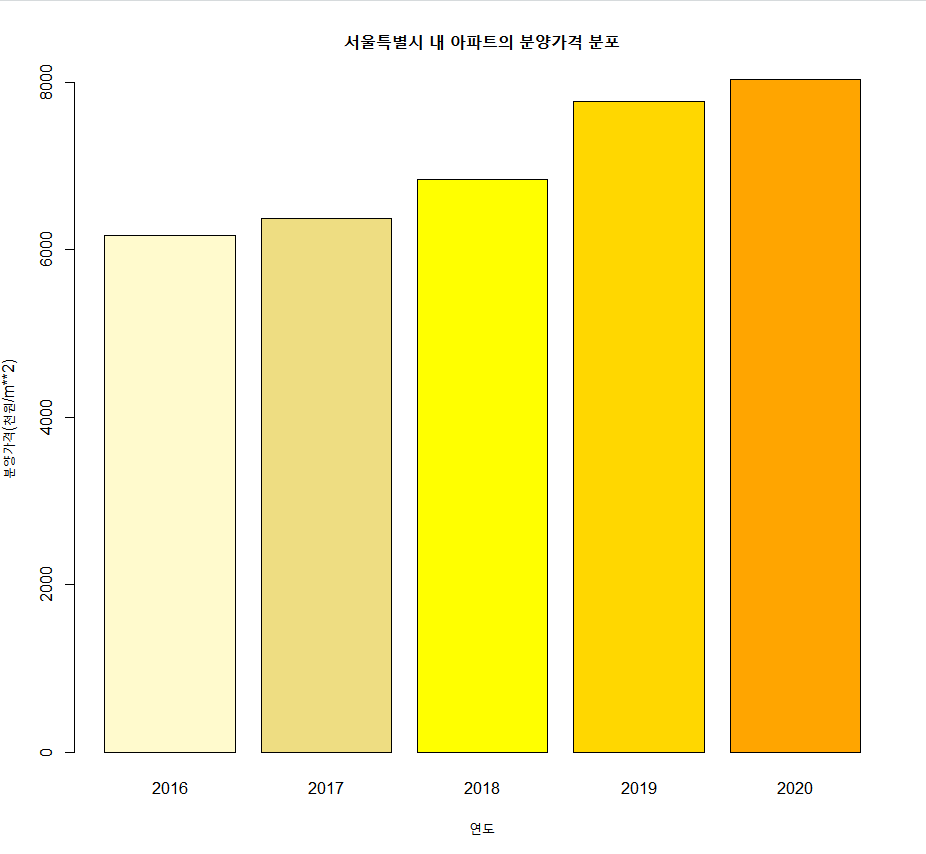
상위와 하위를 쉽게 알기 위해 sort()함수를 사용하였고, 색은 ‘rainbow(length(data))’ 이용하여 많은 막대들을 꾸몄습니다.

1. **아파트의 분양형태 분포**

서울특별시에 분양, 임대, 분양+임대 등의 다양한 분양형태의 아파드 단지들이 어떻게 분포하였는지 확인하기 위해 원형그래프를 그려보았습니다. 그 결과, 서울에는 대부분 분양 아파트가 제일 많이 있고 그 다음 임대아파트와 분양아파트를 같이 가지는 단지, 그리고 임대아파트 단지 순으로 분포함을 알 수 있었습니다. names()와 paste()함수를 이용하여 백분위와 해당 백분위를 가지는 분양형태를 표시하였고 col을 이용하여 붉은 계통으로 그라데이션을 주었습니다.

1. **아파트의 복도형태 분포**

서울특별시에 복도식, 계단식 등 복도형태가 어떻게 분포하였는지 확인하기 위해 원형그래프를 그려보았습니다. 그 결과, 서울에는 대부분 계단식 아파트가 제일 많이 있고 그 다음 계단과 복도의 혼합식, 복도식 아파트 순으로 분포함을 알 수 있었습니다. names()와 paste()함수를 이용하여 백분위와 해당 백분위를 가지는 복도형태를 표시하였고 col을 이용하여 푸른 계통으로 그라데이션을 주었습니다.

1. **신규 민간아파트의 분양가격 분포**

서울 특별시 내 신규 민간아파트의 분양가격을 연도별로 알기 위해 막대그래프와 상자그림을 그려보았습니다. 2016년부터 2020년까지 중 3월의 신규 민간 아파트 분양가격 데이터를 활용하였습니다. 막대그래프를 보아 2016년에는 m2 당 600만원을 맴돌다가 2020년에 **m2** 당 800만원을 찍을 만큼 2016년에서 2020년으로 갈수록 분양가격이 계속 증가함을 알 수 있었습니다. 또한, 그 증가폭이 2018년에서 2019년 사이 가장 크다는 것을 알 수 있었습니다. 또한, 상자그림에서 중앙값은 **m2** 당 650만원과 700만원 사이에 위치하여 있으며, 제1사분위수와 중앙값 사이의 거리가 제3사분위수와 중앙값 사이의 거리보다 짧은 것을 보아 분양가격이 적은 쪽으로 치우쳐져 있음을 알 수 있고, 분양가격이 급격하게 증가하였음을 알 수 있습니다.

막대그래프는 ‘xlab’과 ‘ylab’을 통해 x축, y축 이름을 정해주었고 col을 이용하여 노란 계통으로 그라데이션을 주었습니다. 상자그림은 horizontal= T를 통하여 수평 방향으로 바꾸었습니다.

<시군구별 아파트 분포>와 <분양형태 분포>, <복도형태 분포>는

<http://data.seoul.go.kr/dataList/OA-15818/S/1/datasetView.do>의 “서울특별시 공동주택 아파트 정보.csv” 데이터를 활용하였습니다.

<분양가격 분포>는 <https://www.data.go.kr/data/3035522/fileData.do>의 “주택도시보증공사\_전국 신규 민간 아파트 분양가격 동향.csv” 데이터를 활용하였습니다.