P-SAT

1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 피피티(템플릿은 P-SAT기본 템플릿), 마크다운, HTML, PDF 모두 됩니다. .R이나 .ipynb 등의 소스코드 파일은 안됩니다. 완료 후 psat2009@naver.com로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는시간 이후에 하게 되며, 역시 랜덤으로 5시00분에 발표됩니다.
- 제출기한은 목요일 자정까지이며 지각시 벌금 5000원이 있습니다. 미제출 시 만원입니다. 패키지 2회 무단 미제출 시 퇴출이니 유의해 주세요.

Chapter 1. 전처리

좋은 데이터 분석을 위해서는 자유자재로 데이터를 가공하는 능력이 필요합니다. R에서는 전처리 시 tidyverse 패키지를 가장 많이 사용합니다. dplyr, magrittr, ggplot2 등이 포함 되어있기에 이것 만으로도 대부분의 전처리가 가능합니다. 이번 챕터에서는 코로나 확진자 정보 데이터 'data.csv' 전처리를 통해 대표적인 전처리 기법들에 익숙해져 봅시다.

tidyverse(정확히는 magrittr)에서 가장 기본적인 %>% 연산자를 알려드리겠습니다. pipe 연산자라고 하며, Ctrl + Shift + M 단축키로 쉽게 쓸 수 있습니다. 예를 들어 'A %>% B'는 'A를 전달받아서 B를 처리해라' 라는 뜻으로, 왼쪽에서 오른쪽으로의 직관적인 코드 이해를 가능하게 합니다. 패키지 문제 해결 시 %>%를 최대한 활용하여 직관적인 코드 작성에 익숙해 지시길 바랍니다.

[조건 : tidyverse, plyr, data.table 패키지 모두 사용(이외 패키지 금지), %>% 최대 활용]

문제 0 기본 세팅. 패키지를 plyr, tidyverse, data.table 순으로 부른 후, setwd로 'data.csv'가 있는 폴더로 경로를 설정하고, fread로 'data.csv'를 불러오세요.

문제 1 데이터 확인하기. str으로 데이터의 구성을 살펴보고, 각 열마다 NA 개수와 unique 값 및 개수를 확인해 보세요. (colSums, unique, length 이용 시 편리)

문제 2-1. NA가 있는 행을 삭제하세요.

문제 2-2 빈 문자열("")이 있는 행을 삭제하고, 각 열마다 NA 개수와 unique 값 및 개수를 다시 확인해 보세요. (which, 논리연산자 이용 시 편리)

문제 3. country가 'Korea'인 행만 남긴 다음, country 열을 제거하세요.

문제 4. province 변수 내 '서울, 부산, 대구, 인천, 대전, 세종, 울산, 제주도' 값을 다음과 같이 바꾸세요.

서울 -> 서울특별시, 부산 -> 부산광역시, 대구 -> 대구광역시, 인천 -> 인천광역시, 대전 -> 대전광역시,
세종 -> 세종특별자치시, 울산 -> 울산광역시, 제주도 -> 제주특별자치도

문제 5. confirmed_date를 날짜 자료형(Date)으로 바꾸세요.

문제 6. 확진날짜(confirmed_date) 별 확진자 수에 대한 파생변수를 만드세요. (파생변수 이름 : confirmed_number)

문제 7. 확진날짜(confirmed_date)의 주말 여부에 대한 파생변수를 만드세요. (파생변수 이름 : wday)

예 : 2021-03-05 -> 주중, 2021-03-06 -> 주말

문제 8. 나이대 별 일별 확진자 수에 대한 summary를 확인해 보세요. (예 : 10대의 날짜별 확진자 수의 분포)

(tapply 이용 시 편리)

출력 예시 :

```
$`0s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.488   2.000   3.000

$`100s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       1       1       1       1       1

$`10s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.276   3.000   8.000

$`20s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   3.500   7.526  11.000  43.000

$`30s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   5.051   8.000  17.000

$`40s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   4.931   6.000  33.000

$`50s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   3.000   6.112   9.000  31.000
```

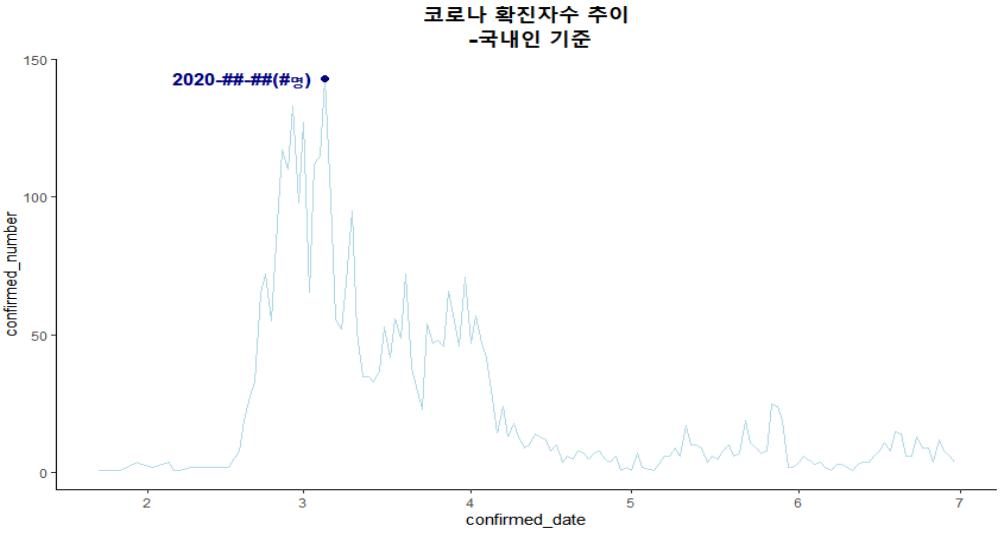
```
$`60s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   3.000   4.722   7.000  22.000

$`70s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.908   3.250  13.000

$`80s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   3.019   3.750  24.000

$`90s`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   2.286   2.000  16.000
```

Chapter 2. 시각화

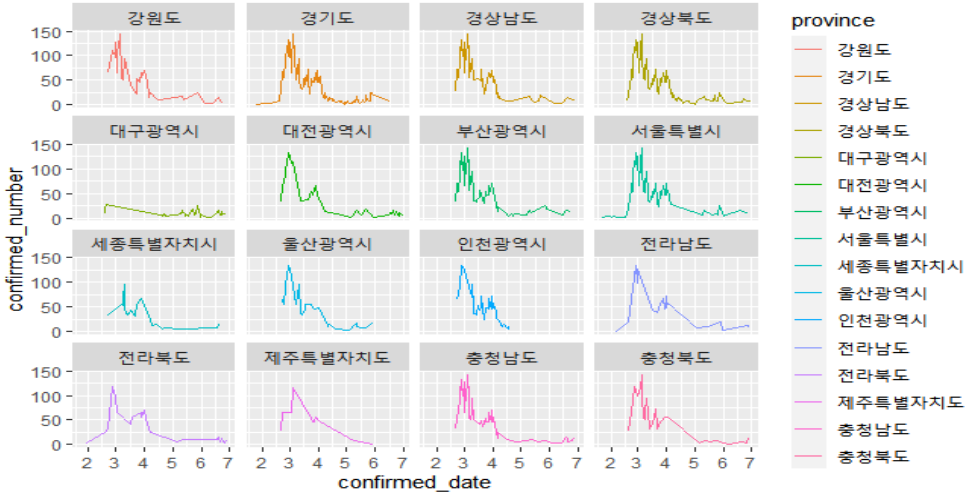
시각화를 통해 데이터에 쉬운 이해가 가능하기에, 데이터의 인사이트를 직관적으로 알 수 있도록 그래프를 그리는 것이 중요합니다. 이번 챕터에서는 ggplot2을 이용하여 대표적인 시각화 기법을 활용해 봅시다.

[조건 :주어진 그래프와 최대한 비슷하게 만들 것(데이터 값, 색, 라벨, 테마 등), %>% 최대 활용]

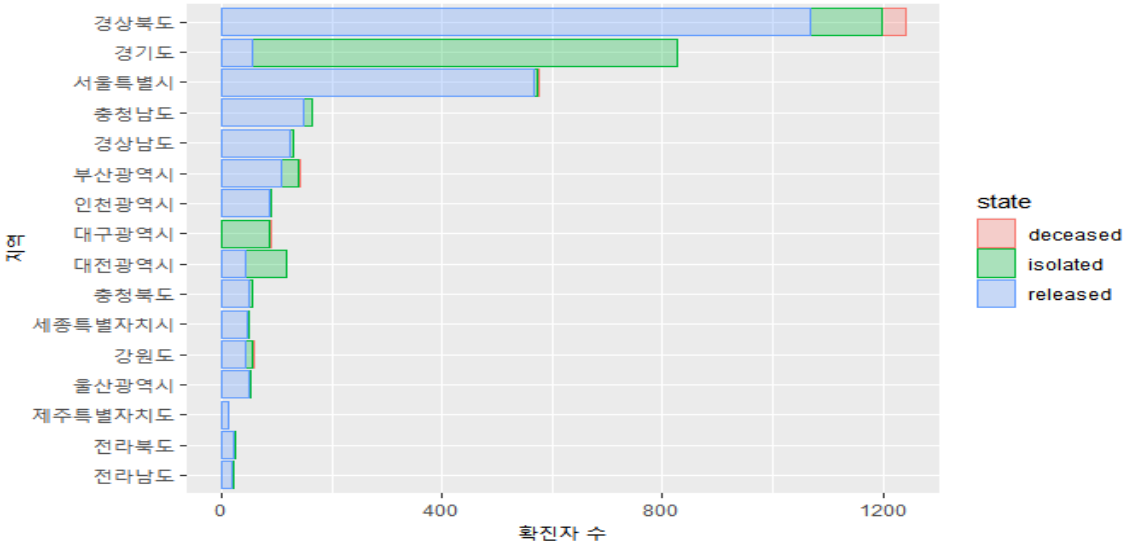
문제 1 Line Plot. confirmed_date와 confirmed_number 변수를 이용하여 확진자수 추이 그래프를 그리고, 최대 확진자에 대한 정보도 표시하세요. (적절한 함수를 이용하여 #대신 정확한 날짜 및 수를 쓸 것 , 색:navy, lightblue, 제목:bold체)



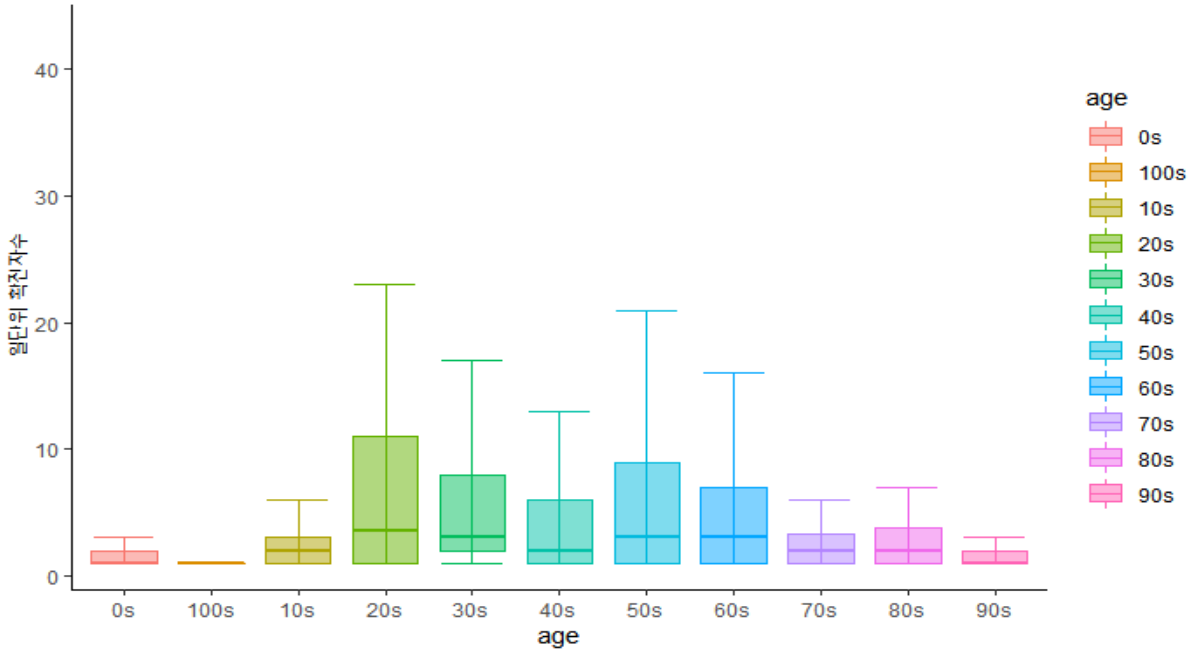
문제 1-2 Line Plot. province 별 확진자 수 추이 그래프를 그리세요.



문제 2 Bar Plot. 지역별 확진자 수를 state(확진자 상황) 그룹 별로 나누어 그래프를 그리세요.



문제 3 Box Plot. 나이대별 일별 확진자 수 box plot을 그리세요



문제 3-2 . 나이대별 일별 확진자 수에 대한 일원분산분석(one-way ANOVA)을 실시하여 해석해 보세요.

문제 4 Map (심화_기존기수는 필수, 신입기수는 선택) .

'CTPRVN_202101' 폴더 내 지도 데이터(TL_SCCO_CTPRVN.shp)를 이용하여 province(지역) 별 확진자 수에 대한 지도 시각화를 하세요.

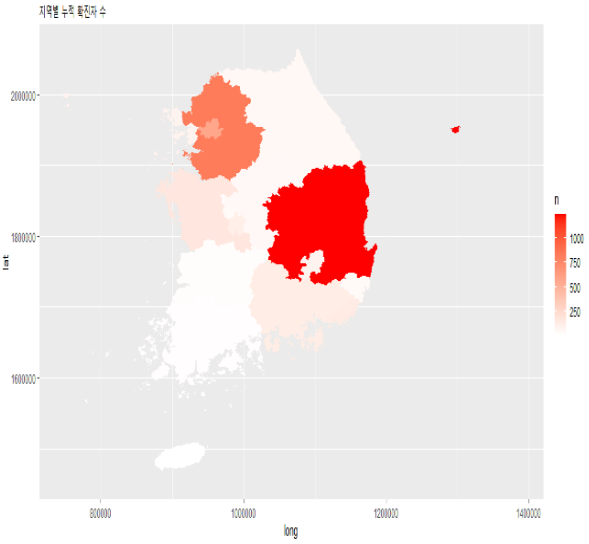
<힌트_아래 순서를 참고할 것>

1. raster, rgeos, maptools, rgdal 패키지 이용

2. 지도 데이터(TL_SCCO_CTPRVN.shp)를 readOGR로 부른 후, fortify로 데이터 프레임 화

3. left_join을 이용하여 데이터와 지도데이터 결합

4. geom_polygon 이용하여 시각화



Chapter 3. 모델링_회귀분석

회귀분석은 예측을 위한 지도학습 모형 중 기본 모형으로, 해석이 쉽기에 실무에 자주 쓰입니다. 회귀분석은 간단해 보이지만, 회귀 가정 만족 및 변수 선택 등 고려할 것이 많고 다양한 응용 모델들이 있습니다. 3주간의 회귀분석 클린업을 통해 자세한 내용을 알기 전에, 이번 챕터를 통해 간단히 회귀분석 수업에서 배운 내용을 복습해봅시다.

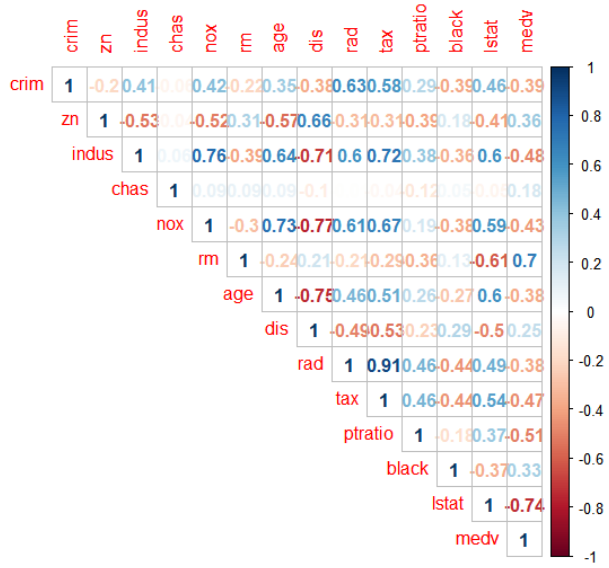
또한 **모델의 예측 및 분류 값을 평가**하기 위해서는 데이터를 train/test로 나누어 test set으로 모델의 성능을 평가해야 합니다. 과적합을 방지하기 위해 train/validation/test로 나누거나 Cross Validation(CV)를 통해 모델의 성능을 평가하는 방식이 더 선호되지만, 이는 데이터마이닝 클린업 및 다음 주 패키지를 통해 더 자세히 알아보도록 합시다.

[조건 : MASS 패키지에 있는 'Boston' 데이터를 사용할 것,

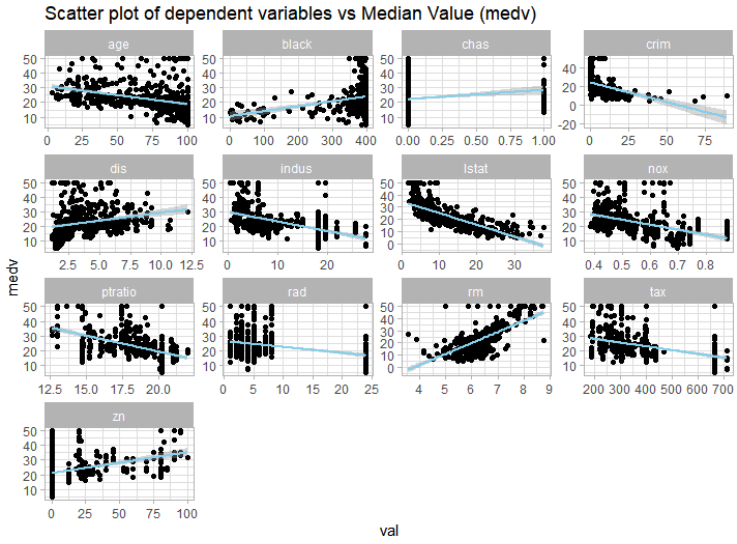
MASS, tidyverse(다시 부를 필요 없음),corrplot, caret, MLmetrics 패키지 모두 사용(이외 금지),

그래프를 최대한 비슷하게 나타낼 것]

문제 1. 아래처럼 상관계수 플랏을 만들고 간단히 해석해 보세요.



문제 2. 아래와 같이 종속변수로 사용할 medv와 이외 변수 간의 관계 파악을 위한 scatterplot 및 회귀 추세선을 그려보세요. (gather 이용 시 편리, 색 : lightblue)



문제 3. 데이터를 train/test 데이터를 7:3으로 나누세요 (1234로 시드 고정 필수).

문제 3-2. train 데이터로 medv를 종속변수로 하고 나머지를 모두 독립 변수로 하는 회귀 모형을 만든 후 간단히 결과를 해석한 후, test에 대한 RMSE를 구하세요.

문제 3-3. 모델의 RMSE를 낮출 수 있는 방법에 대해 간략히 설명해 주세요.

문제 4. 적합한 회귀모형의 계수에 대해 아래와 같이 시각화 해주세요 (색 : red, yellow, blue)