1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 R markdown으로 HTML, PDF 모두 가능합니다. **.R이나 .ipynb 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 5시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요.

Chapter 1 전처리

데이터 분석을 위해선 먼저 데이터를 분석 가능한 형태로 가공하는 과정이 필요합니다. R에서는 데이터 전처리에 tidyverse 패키지를 가장 많이 사용합니다. dplyr, magrittr, ggplot2 등의 패키지가 포함 되어 있어 대부분의 전처리가 가능합니다. 이번 챕터에서는 금융 **고객정보(cus_info) 데이터**와 **계좌정보(act_info) 데이터**를 전처리해보면서 tidyverse 패키지 사용에 익숙해져 보도록 하겠습니다.

Tidyverse 패키지를 사용해 데이터를 정제할 때 pipe 연산자 **%>%**를 자주 사용하게 됩니다. %>% 연산자는 magrittr에 포함되어 있는 연산자로 단축키 **ctrl+shift+M**를 통해 부를 수 있습니다. 기존에 h(g(f(x)))처럼 안에서 밖으로 흐르던 데이터 흐름 연산을 x %>% f() %>% g() %>% h() 와 같은 형태로 왼쪽에서 오른쪽으로 바꿀 수 있습니다. 패키지 문제 해결 시 최대한 %>%를 사용하여 직관적인 코드를 작성해봅시다.

[조건: tidyverse, plyr, data.table, ggpubr 패키지 사용 (이외의 패키지 사용 금지), %>% 연산자를 최대한 사용하여 한 줄의 코드로 표현]

문제0. (기본 세팅) 0번 txt파일을 실행하세요. (패키지 불러오기, 디렉토리 설정 및 데이터 불러오기)

문제1. 데이터의 기본 구조를 파악하고 데이터 개수, 변수 개수, 데이터 형식을 확인해보세요. (head, tail, str, glimpse, summary 등 다양하게 사용해보세요)

문제2. 각 열별로 결측치(NA)의 개수를 확인한 후 결측치가 70% 이상인 열을 삭제하세요.. (colSums 사용 시 편리)

문제3. 각 열마다 unique한 값의 개수를 확인하세요. (apply, n_distinct 사용 시 편리)

문제4. act_info에서 계좌개설일(act_opn_ym)의 unique 값을 확인 후 이상치값을 갖는 행을 삭제하세요.

문제5. act_info에서 계좌개설일(act_opn_ym) 변수를 각각 년(act_opn_yy) 변수와 월(act_opn_mm) 변수로 나눈 뒤 수치형 변수로 변환하세요. (separate 사용시 편리)

문제6. cus_info에서 범주형 변수인데 수치형으로 읽힌 경우 mutate_if를 통해 범주형 변수로 변경하세요.

문제7. cus_info에서 연령대(cus_age) 변수를 10세 기준으로 재범주화 하세요. (데이터 명세 참고)

문제8. 데이터 레이블 변경하기

8-1. 성별(sex_dit_cd) 변수의 값이 1일 경우 "M", 0일 경우 "F" 로 변경하세요.

8-2. 주소(zip_ctp_cd) 변수의 값을 각각 다음과 같이 변경하세요.

(41:경기, 11:서울, 48:경남, 26:부산, 27:대구, 47:경북, 28:인천, 44:충남, 46:전남,

30:대전, 29:광주, 43:충북, 45:전북, 42:강원, 31:울산, 50:제주, 36:세종)

문제9. cus_id를 제외한 모든 변수들을 factor 형태로 변경한 후 자료형태를 다시 확인해보세요.

문제10. 문제 4번의 결과를 바탕으로 두 데이터셋을 병합한 뒤 data로 저장하고 이전 데이터셋은 삭제하세요.

문제11. 연령대별(cus_age)로 그룹화하여 고객 수(cus_cnt), 계좌 수(act_cnt), 그리고 1인당 평균 계좌 개수(mean_act_cnt) 파생변수를 만드세요. 이후의 시각화에 사용하기 위해 account__cnt로 저장하세요.

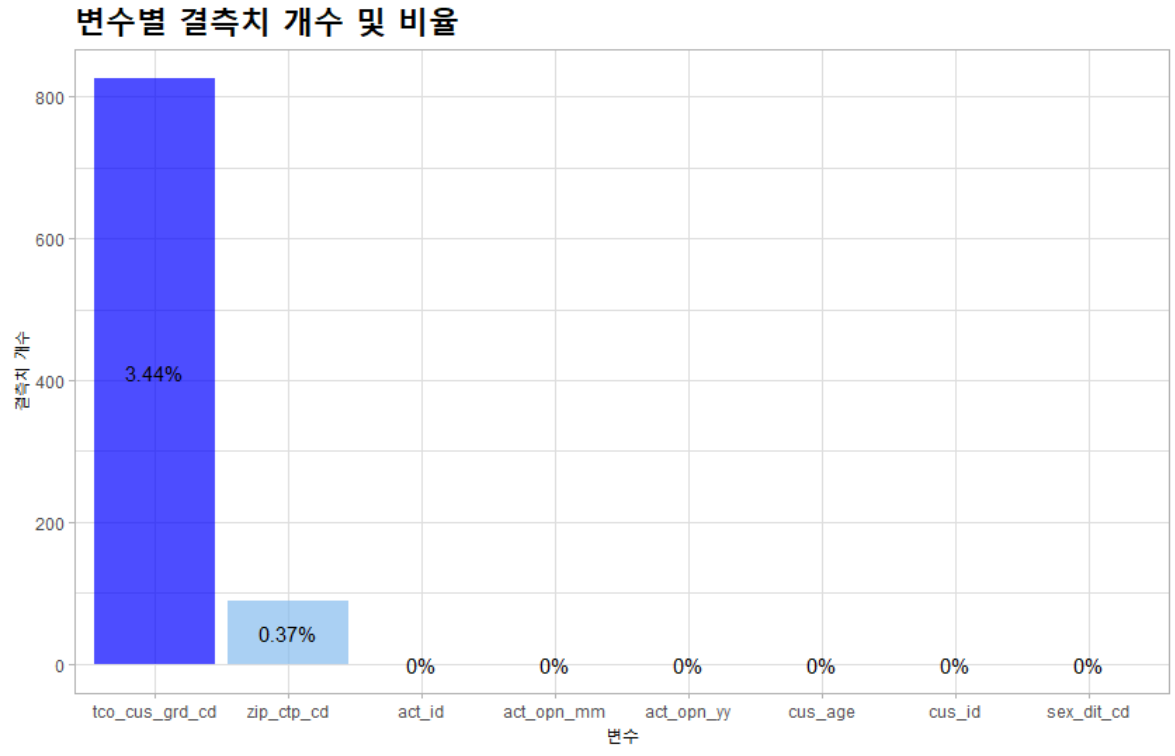
Chapter 2 시각화

데이터 시각화는 데이터 분석 결과를 시각적으로 표현해 **스토리텔링**을 하는 능력입니다. 데이터 시각화를 통해 우리는 많은 양의 데이터를 한눈에 볼 수 있고, 데이터에 담긴 **인사이트를 발굴**할 수도 있습니다. 같은 정보라도 시각화를 잘 하면 정보를 보다 더 쉽게 이해할 수 있습니다. R에서는 ggplot2 패키지를 사용하여 데이터를 다양하게 시각화할 수 있습니다. 이번 챕터에서는 **ggplot2**를 사용하여 그래프를 그리고 커스터마이징하는 방법을 익혀보도록 하겠습니다.

[조건: 주어진 그래프와 최대한 비슷하게 만들 것, %>% 연산자 최대한 사용하여 한 줄의 코드로 표현, Warnings가 뜨는 경우 R Markdown에서 warning=FALSE로 설정해서 뜨지 않게 해주세요.]

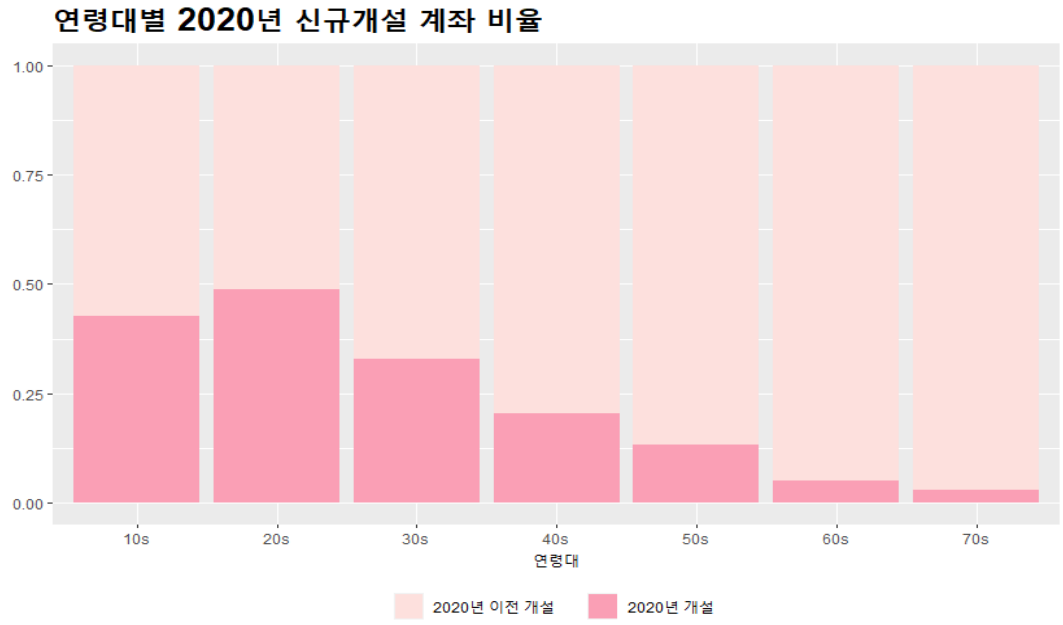
문제1. Bar Graph

1-1. data의 각 변수별 결측치 개수와 비율을 다음과 같이 시각화해서 보여주세요.



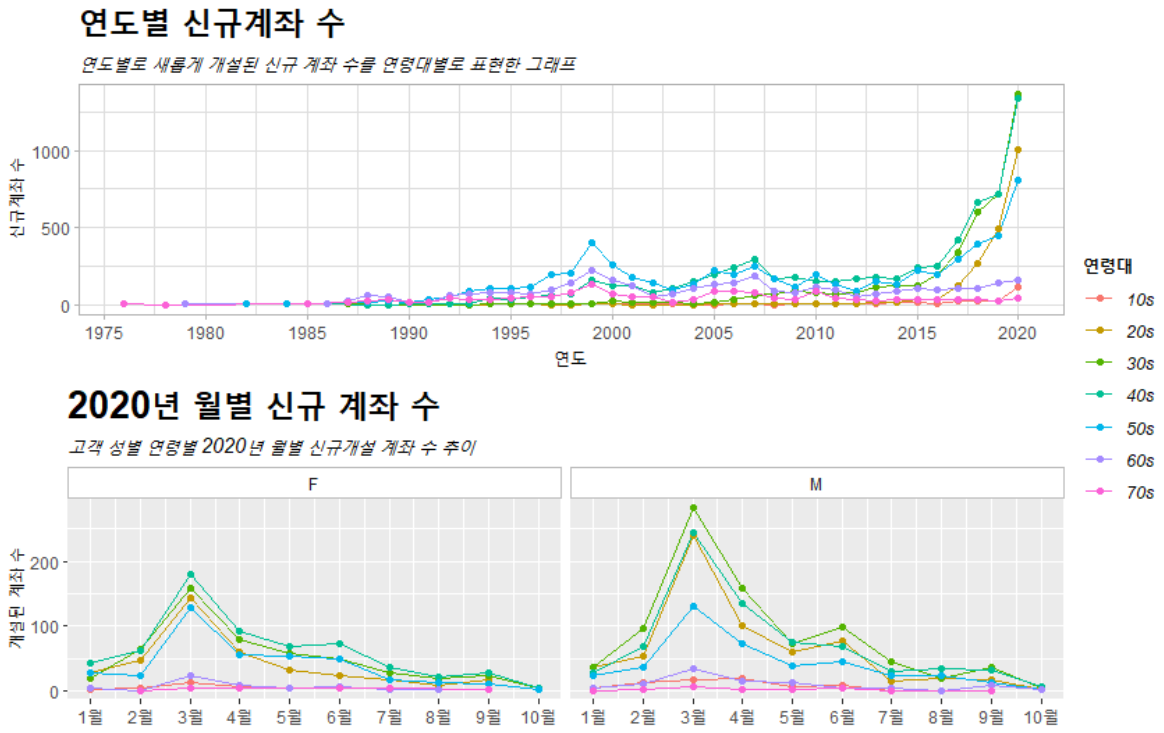
- 변수별 결측치 개수와 비율을 구한 데이터프레임을 만들면 편리합니다.
- 크기 순으로 정렬해주세요.
- Bar graph의 색상은 결측치 개수가 큰 쪽이 blue, 작은 쪽은 skyblue이고, 투명도는 0.7입니다.
- 플랏 제목은 size 20, 볼드 체입니다.

1-2. data에서 연령대별 2020년 신규 개설 계좌 비율을 다음과 같이 시각화해서 보여주세요.



- 계좌개설 년도가 2020인지 구분하는 파생변수를 생성합니다.
- 팔레트 색상은 palette = "RdPu", 플랏 제목은 size 20, 볼드 체입니다.

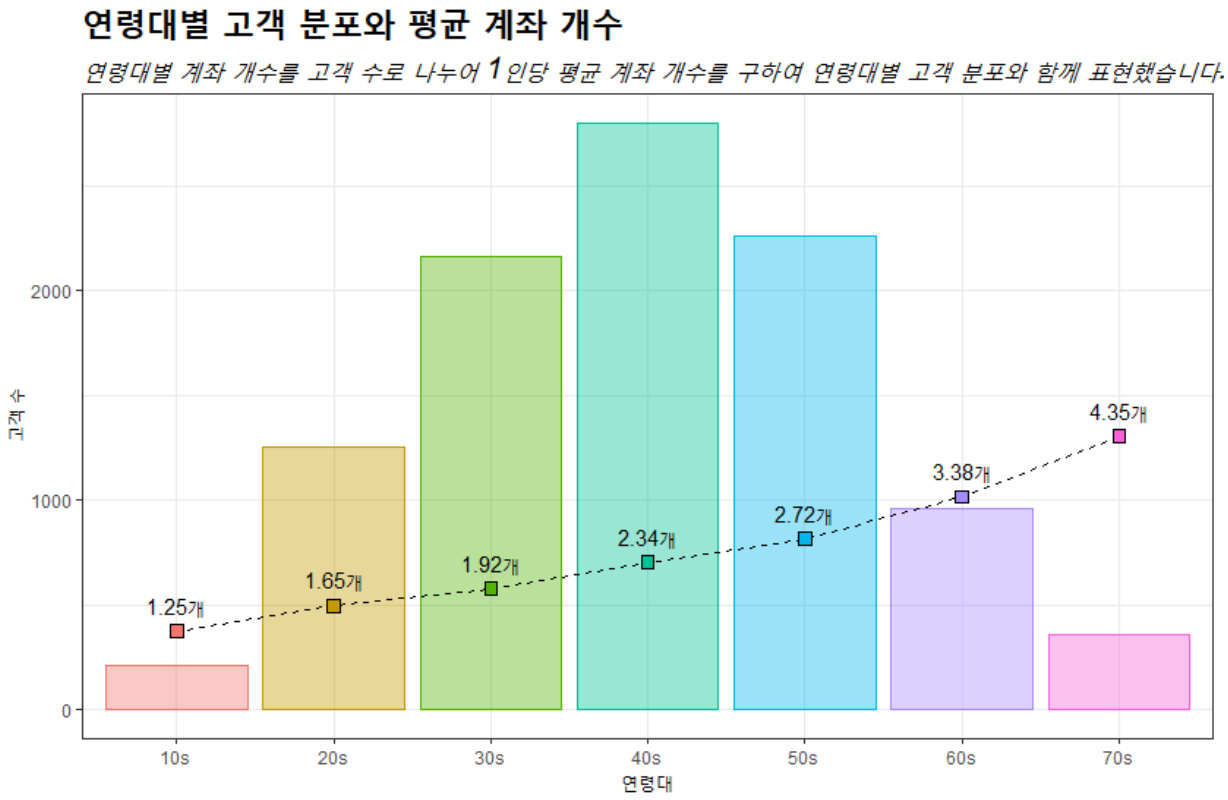
문제2. Line Graph (Time Series Graph)



- 두 그래프는 공통 범례를 사용하며, 범례 텍스트는 이탤릭 체입니다.
- 하단의 그래프는 2020년 내 개설된 계좌를 성별에 따라 구분한 것입니다.
- 두 플랏의 제목은 size20, 볼드 체이고, 부제목은 size 15, 이탤릭 체입니다.

문제3. Bar Graph + Line Graph

Chapter1 문제 11번에서 생성한 account_cnt를 사용하여 다음과 같이 연령대별 고객 분포를 bar graph로 표현하고, 인당 평균 계좌 개수를 line graph로 표현하여 시각화해주세요.



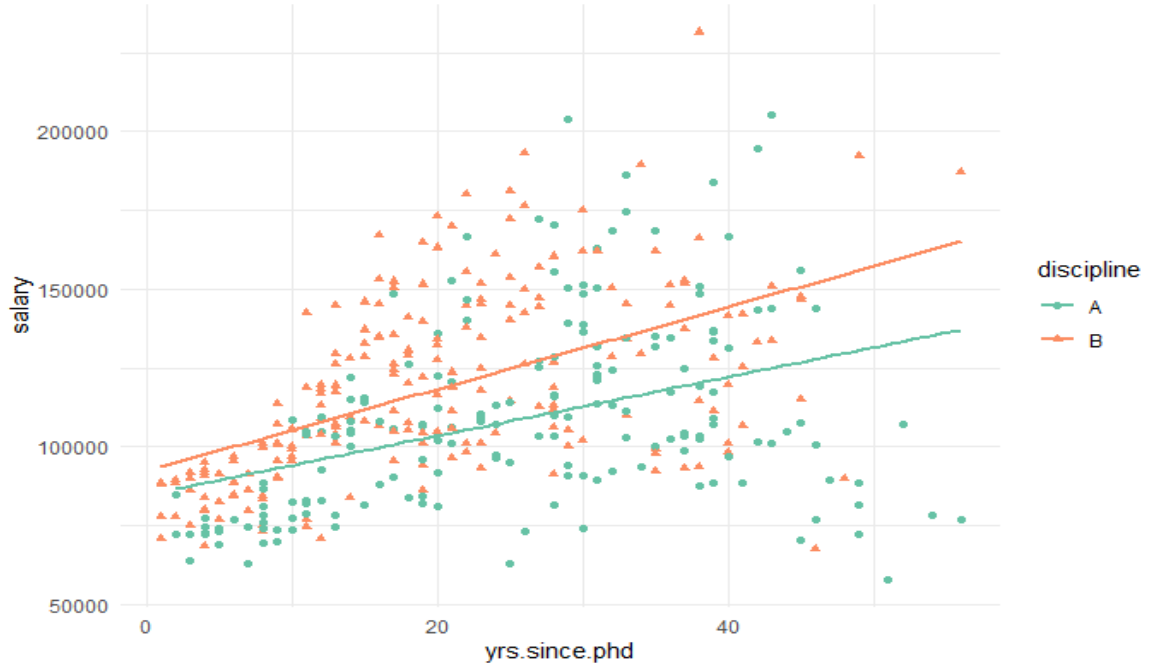
- Line graph는 인당 평균계좌 개수에 300을 곱해서 그려주세요. (왜 숫자를 곱해주는지 이유도 적어주세요.)
- Bar graph는 투명도 0.4이고, 선 그래프의 linetype은 "dashed", 점 shape은 22, size는 3입니다.
- 플랏 제목은 size 20, 볼드 체이고, 부제목은 size15, 이탤릭 체입니다.

문제4. Scatter Plot & Box Plot

환경에 저장된 데이터를 모두 삭제하고 carData 패키지의 Salaries 데이터셋을 불러주세요.

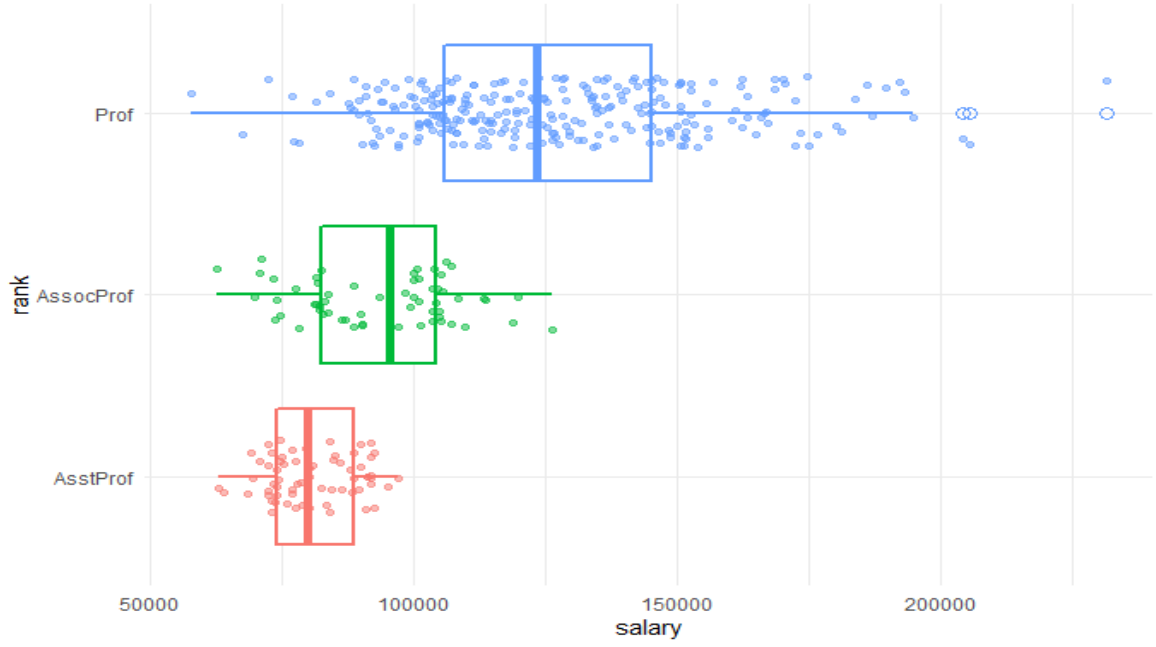
[rm(list=ls()) ; data(Salaries, package="carData") 코드 사용]

4-1. Salaries 데이터셋으로 다음과 같은 scatter plot을 시각화해주세요.



- X축, Y축, 범례에 사용된 변수명 그대로 사용하면 됩니다.
- 각 discipline 별 회귀선을 신뢰구간을 표시하지 않고 그려주세요.

4-2[심화_기존은 필수, 신입은 선택] Salaries 데이터셋으로 다음과 같은 box plot을 시각화해주세요.



- X축, Y축에 사용된 변수명 그대로 사용하면 됩니다.
- Scatter plot의 점은 투명도 0.5, 넓이 0.2입니다.

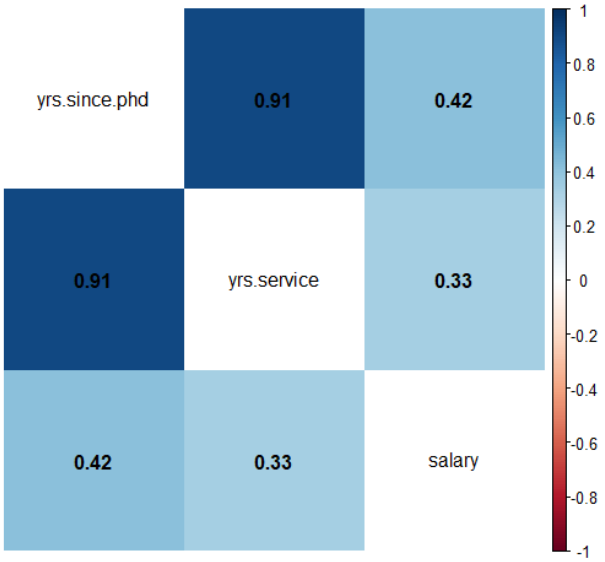
Chapter3 회귀분석

회귀분석은 예측을 위한 지도학습 모형 중 기본 모형으로 해석이 쉽기에 자주 쓰입니다. 회귀분석은 간단해 보이지만 회귀 가정 만족 및 변수 선택 등 고려할 것이 많고 다양한 응용 모델들이 있습니다. 3주 간의 회귀분석 클린업 동안 자세한 내용을 배우기 전에 이번 챕터를 통해 간단한 회귀분석 내용을 복습해봅시다.

또한 **모델의 예측 또는 분류 값을 평가**하기 위해서는 데이터를 train set과 test set으로 나누어 모델의 성능을 평가해야 합니다. 과적합을 방지하기 위해서 Hold-out Validation 또는 K-fold Cross Validation (CV) 등이 사용되는데, 자세한 내용은 데이터마이닝팀 클린업과 다음주 패키지를 통해 알아보도록 하겠습니다.

[조건: tidyverse(다시 부를 필요 X), corrplot, caret, Metrics 패키지 사용 (이외 금지)]

문제 1. Salaries 데이터셋의 수치형 변수만을 선택하여 상관계수 플랏을 그리고 간단히 해석해보세요.



문제 2. [심화_기존은 필수, 신입은 선택] Salaries 데이터셋에서 성별에 따른 salary의 평균이 유의미하게 다른지 통계적으로 검증하고 싶습니다. 어떤 검정 방법을 사용할지 선택하고, 검정을 진행한 뒤 그 결과를 해석해주세요.

문제3. 데이터를 7:3비율로 train/test를 분리하세요. (2728 시드 고정 필수, p=0.7으로 사용)

문제4. train 데이터를 이용하여 salary를 종속변수, 나머지 변수들을 독립변수로 하는 회귀 모형을 만든 뒤 결과를 간단히 해석해주세요. (범주형 독립변수들은 어떻게 해석할 수 있는지도 설명해주세요.)

문제5. 회귀모형의 성능을 평가할 수 있는 지표가 무엇인지 설명하고, 모델의 train error와 test error를 계산한 뒤 비교해주세요.