클린업 1주차 패키지

- 분석 툴은 R/Python 둘 다 사용할 예정입니다. 클린업 1주차 패키지 문제의 조건 및 힌트는 R, Python 모두 제공합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실시도 퇴출이니 유의해주세요.

Chapter 1 : Data Preprocessing & EDA

데이터 분석을 위해서 저희가 주로 다루게 되는 언어는 R과 Python입니다. R과 Python은 세부적인 문법이나, 자료의 구조 등이 세세하게 다르기는 하지만, 각 언어 마다 강점을 보이는 부분들이 있기 때문에, 단순히 익숙한 한 가지 언어만을 사용하는 것보단, 필요에 따라 R과 Python을 선택하여 사용하는 것이 중요합니다.

R에서는 데이터 분석을 위해 **Tidyverse**라는 패키지를 많이 사용합니다. 해당 패키지에는 tidyr, dplyr, ggplot 등 데이터를 전처리하고 시각화하는 과정에서 많이 사용되는 패키지들로 구성되어 있습니다.

또한, 데이터를 정제하는 과정에서 **pipe 연산자(%>%)**를 활용하면 코드를 간결하고 깔끔하게 짤 수 있다는 장점이 있습니다. 해당 연산자는 ctrl+shift+M을 통해 불러올 수 있으며, 데이터 흐름을 왼쪽에서 오른쪽으로 흐르도록 하여 직관적으로 파악할 수 있게 하는 장점이 있습니다.

Python은 R과 같이 데이터 분석/통계 만을 위한 프로그래밍 언어는 아니지만, **Pandas**라는 모듈을 사용하는 것으로 R에서 데이터프레임을 다루는 것과 비슷하게 데이터를 다룰 수 있고, matplotlib이나 seaborn등의 모듈을 통하여 시각화를 진행할 수도 있습니다.

이번주에는 R과 Python에서 데이터 전처리 과정에서 많이 사용되는 패키지를 이용하여 데이터를 효율적으로 파악해보고, 이를 통해 각 프로그래밍 언어 별로 가지는 차이점이 무엇인지 알아보는 시간을 가져보도록 하겠습니다.

문제 조건 :

① **문제에서 요구하는 패키지 및 모듈 외에는 최대한 사용 자제**

② **코드는 최대한 간결하게 작성**

③ **시각화는 최대한 문제 예시와 비슷하게 작성**

문제 1. R과 Python에서 각각 데이터를 불러온 뒤, 데이터의 구조를 파악하세요.

(R HINT) Tidyverse 패키지를 설치하여 사용하면 되고, R에서는 데이터 구조 파악을 위해 head, tail, summary, glimpse, str 등 다양한 함수가 쓰입니다.

(Python HINT)Pandas 모듈을 불러 온 후 데이터를 불러올 수 있고, Python에서는 데이터 구조 파악을 위해 마찬가지로 head, tail, info, describe 등 다양한 함수가 쓰입니다.

문제2. 데이터에서 각 변수별로 unique한 값이 몇 개씩 존재하는지 파악해주세요.

(R HINT) lapply 함수, n_distinct 함수를 사용하면 편합니다.

(Python HINT) Pandas에서는 함수를 통해 변수 별 unique 개수를 한 번에 확인할 수 있습니다.

문제3. 문제 1과 문제 2에서 얻어낸 정보들을 바탕으로, 데이터에서 각 변수들이 범주형 변수인지 수치형 변수인지 판단해보고, 그 이유에 대해서 간략하게 서술해주세요.

(+R) 수치형에 해당하는 변수들의 경우에는 numeric형으로, 범주형에 해당하는 변수들의 경우에는 factor형으로 자료를 변환해주세요. 이때, mutate_if() 함수를 사용하면 좀 더 편하게 자료형을 변환할 수 있습니다.

문제4. 데이터 내에 결측치가 있는지 간단히 파악해보세요.

(HINT) R과 Python에선 모두 해당 값이 결측치(NA)인지 아닌지를 반환해주는 함수를 가지고 있습니다. 이를 활용하여, 변수 별 나타나는 결측치의 개수를 확인하면 됩니다.

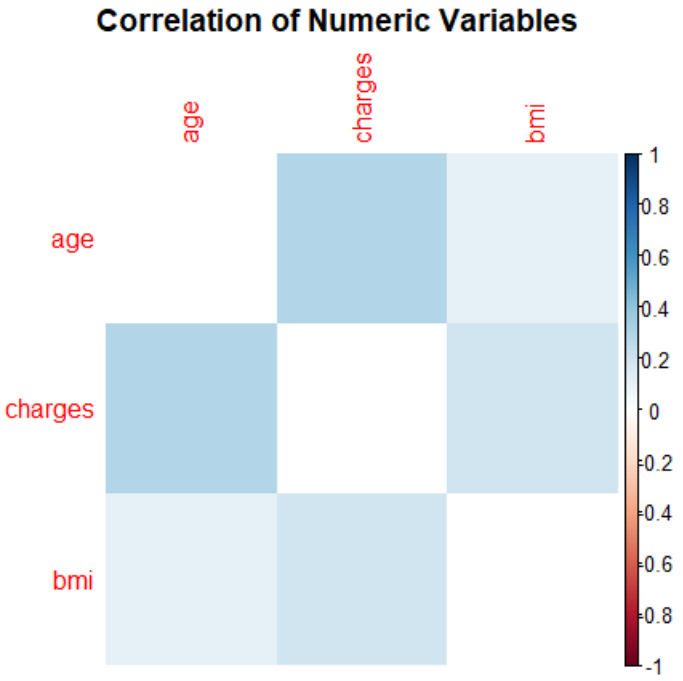
문제5. 데이터에 있는 'age' 변수를 바탕으로 'age_group'이라는 범주형 파생변수를 생성해주세요.

- 파생변수 'age_group'은 'young', 'senior', 'elder' 이렇게 3개의 값 만을 가지는 변수입니다.
- 만약 age 변수의 값이 18-35일 때는 'young',
  만약 age 변수의 값이 36-55일 때는 'senior',
  만약 age 변수의 값이 56 이상일 때는 'elder'의 값을 가지도록 파생변수를 생성해주세요.

(R HINT) mutate() 함수와 조건문을 사용하면 편하게 파생변수를 생성할 수 있습니다.

(Python HINT) 파생변수를 만들 때 lambda식을 활용하면, 좀 더 깔끔하게 파생변수를 만들 수 있습니다.

문제6. 데이터에서 수치형 변수들만을 사용하여, 수치형 변수들 간의 상관관계를 다음과 같은 상관관계 Plot을 통해 확인해주세요. 그리고 그 결과에 대해서 간단히 해석해주세요.



(HINT) R과 Python에서 상관관계 Plot을 그리기 위해선 먼저 상관계수 행렬이 필요합니다. 수치형 변수들간의 상관관계를 파악할 때는 피어슨(Pearson) 상관계수를 확인합니다.

(R HINT) corrplot 패키지를 활용하면, 상관계수 행렬로 쉽게 상관관계 Plot을 그릴 수 있습니다.

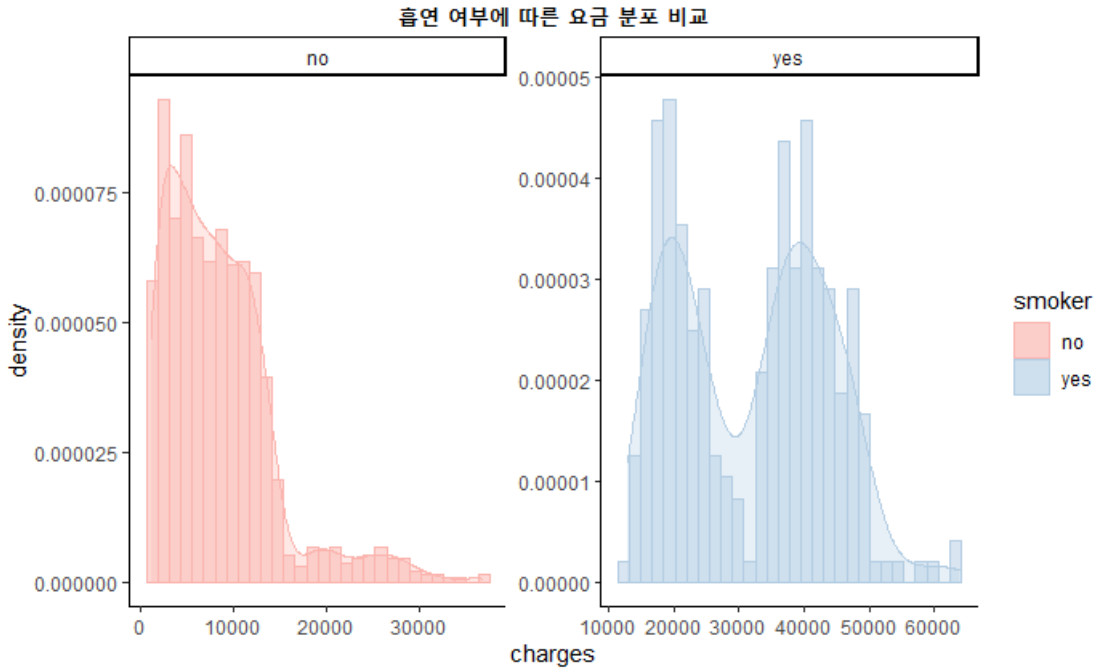
이때, method = shade, order = AOE로 설정 후 상관관계 Plot을 그리시면 됩니다.

(Python HINT) 상관관계 Plot을 그릴 때, 상관계수 행렬로 heatmap을 그리면, 상관관계 Plot을 그릴 수 있습니다.

(+보너스 문제1). 문제5에서는 수치형 변수들에 대해서만 상관관계 Plot을 그려보았지만, 범주형 자료들에 대해서도 상관계수를 계산할 수 있는 방법이 있습니다. 해당 방법에 대하여 조사해 본 후, 그 내용에 대해서 간단히 요약해서 서술하고, 수치형 변수때와 마찬가지로 범주형 자료들 간의 상관관계를 상관관계 Plot을 통해 확인해주세요.

문제7. 흡연 여부(smoker)에 따른 요금(charges)의 분포를 각각 비흡연자(smoker==no), 흡연자(smoker==yes)로 나누어서 다음과 시각화해주세요. 그 후, 그려진 plot을 통해 알 수 있는 점들을 간략하게 적어주세요.

- Plot을 그릴 때 palette는 'Pastel1'으로 지정해주세요.
- 제목의 위치는 Plot의 정중앙, 굵기는 'bold'로 설정해주세요.
- 그림과 같이 범주형 변수에 따라 Plot을 따로 그릴 때, FacetGrid 또는 facet_wrap등을 활용해주세요.



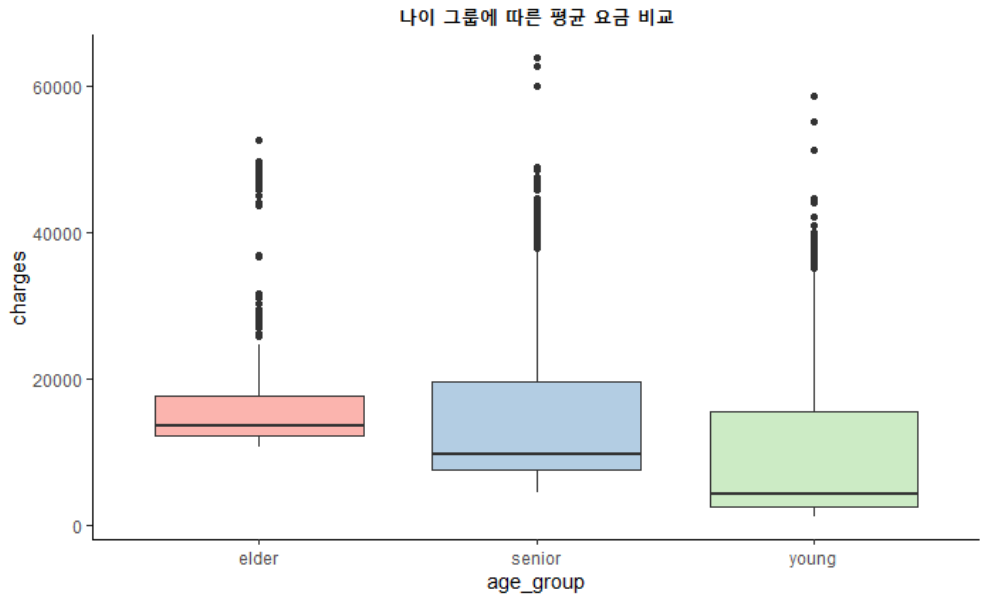
(R HINT)

1) 다음과 같이 히스토그램과 함께 분포 또한 같이 그리기 위해서는 geom_histogram과 geom_density를 함께 사용해야 합니다.
2) ggplot에서 선의 색깔을 조정할 때는 'color', 면적의 색을 조정할 때는 'fill'을 사용합니다.

(Python HINT)

1) R과 달리 Python에서는 seaborn 모듈의 distplot()함수를 사용하면 히스토그램과 분포를 동시에 그릴 수 있습니다.
2) Seaborn을 사용하여 시각화를 진행할 때, matplotlib에서 제공하는 함수를 사용하여 직접 옵션(제목 폰트 위치, 굵기 등)을 조절할 수 있습니다.

문제8. 나이 그룹(age_group)에 따른 평균 요금(charge)의 차이를 확인하기 위해 박스 플롯을 그려 다음과 같이 시각화해주세요. 그 후, 그림을 통해 확인할 수 있는 점을 간략하게 적어주세요.



Chapter 2 : ANOVA(ANalysis Of Variance, 분산분석)

본격적으로 모델링으로 넘어가기 전에, 굉장히 고전적인 모델 중 하나이지만, EDA 단계에서 유용하게 사용할 수 있는 분석 방법 중 하나인 분산분석(Analysis of Variance, ANOVA)를 다뤄보고자 합니다.

앞서 Chapter 1의 8번 문제처럼, 시각화를 통해서도 몇몇 그룹 간의 평균 값의 차이를 확인하기도 하는데, 이런 시각화를 통해서도 충분히 인사이트를 얻어 추후 파생변수 생성 등에 활용할 수도 있겠지만, 이를 통계적인 검정을 통해 정말 그룹 간의 평균 값의 차이가 유의한지 알아낼 수 있다면, 시각화를 통해 얻어낼 수 있는 결론에 좀 더 강한 설득력을 부여할 수 있을 것입니다.

ANOVA는 이렇게 전처리 과정 혹은 탐색적 데이터 분석(Exploratory Data Analysis, EDA) 과정에서 활용하기 좋은 모델입니다. 이번 챕터에서는 ANOVA가 어떤 모델이고, ANOVA를 통해서 어떻게 데이터에서 정보를 얻어내면 좋을지 알아보는 시간을 가지도록 하겠습니다.

문제1. ANOVA가 어떤 모델인지 조사해본 후, ANOVA가 어떤 모델인지에 대해서 간단히 서술해주세요. 이때, 다음과 같은 내용을 넣어서 서술해주세요.

- ANOVA를 통해 알아내고자 하는 것
- ANOVA의 귀무가설과 대립가설
- 3그룹 이상의 비교를 진행할 때 Z-test나 T-test가 아닌 ANOVA를 사용해야 되는 이유

문제2. ANOVA를 사용하여 나이 그룹(age_group) 간의 평균 요금(charges) 차이가 있는지를 검정해보고, 이를 ANOVA표를 통해 검정 결과를 해석하세요.

(R HINT) R의 경우에는 aov()함수를 사용하여 ANOVA를 진행할 수 있고, 회귀 분석을 진행할 때와 마찬가지로 summary()함수를 통해 결과를 확인할 수 있습니다.

(Python HINT) Python의 경우에는 statsmodels 모듈을 import한 후 ols().fit()을 통해 모델을 적합할 수 있고, 결과는 anova_lm()를 통해 확인할 수 있습니다.

```
                    Df       Sum Sq    Mean Sq F value             Pr(>F)
factor(age_group)
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

문제3. ANOVA를 사용하여 나이 그룹(age_group) 간의 평균 BMI 값에 차이가 있는지를 검정해보고, 이를 ANOVA표를 통해 검정 결과를 해석하세요.

```
                    Df Sum Sq Mean Sq F value  Pr(>F)
factor(age_group)
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

문제4. ANOVA를 사용하여 지역(region) 간의 평균 나이(age)의 차이가 있는지를 검정해보고, 이를 ANOVA 표를 통해 검정 결과를 해석하세요.

```
                   Df Sum Sq Mean Sq F value Pr(>F)
factor(region)
Residuals
```

Chapter 2에서 다뤄본 ANOVA는 결과적으로 input으로 범주형의 변수를 가지고, output으로 실수형의 변수를 가지는 선형 회귀 모형으로도 생각해볼 수 있습니다. 이런 점에서 ANOVA의 경우에는 일반화 선형 모델(Generalized Linear Model, GLM)에 속하기도 합니다.

(+보너스 문제2). ANOVA 또한 선형 회귀 모델과 같이 기본적인 모델의 가정이 존재할 것입니다. ANOVA 모델의 기본적인 가정에 대해서 살펴본 후, 이를 Plot이나 검정 등을 통해 확인 할 수 있는 방법을 찾아 설명해보세요.

(+보너스 문제3). ANOVA는 단순히 하나의 요인에 대한 평균 값들의 차이만을 검정할 수 있는 것이 아니라(One Way ANOVA), 2개 이상의 요인에 대한 평균 값들의 차이에 대해서도 검정할 수 있습니다. Two Way ANOVA에 대해서 조사해본 후, 나이 그룹(age_group)과 흡연여부(smoker)에 따라 평균 요금(charge)에 차이가 있는지 확인해보세요.