개인 의료비 예측 프로젝트 (feat. 내 보험료는 얼마?)

고기4조
남태우,이수정,이정훈,정은정,정재석

팀원소개

남태우

데이터분석

이수정

데이터분석

모델링 구현

이정훈

데이터분석

모델링 구현

정은정

발표자료제작

정재석

발표자료제작

목차

01. 프로젝트 개요 - 목표

2023년 3월	23일 13:00	15:00	18:00	21:00	09:00	12:00	15:00
데이터 탐색							
데이터 전처리 및 분석							
모델링							
PPT 제작							

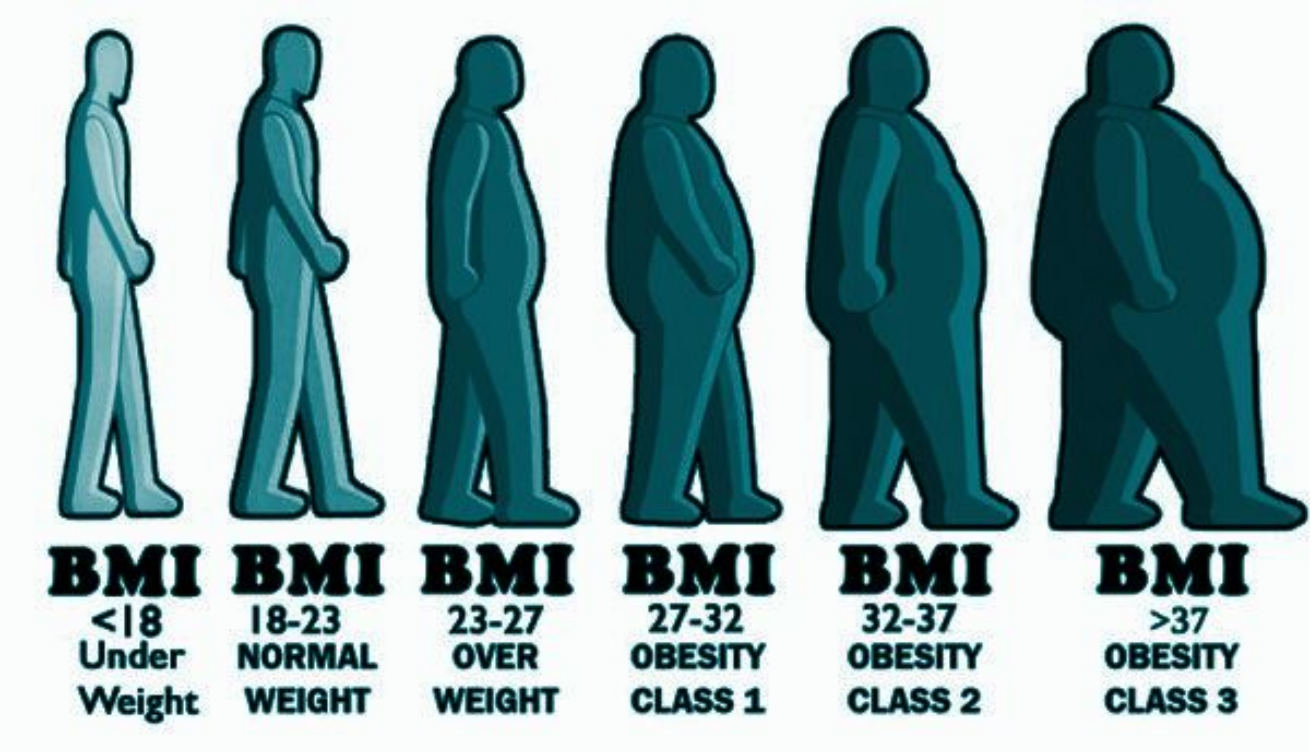
01. 프로젝트 개요 - 용어정리

1. Kaagle: 데이터 과학 및 머신러닝을 학습 하고 데이터 분석을 위한 데이터 셋을 제공하는 플랫폼
2. EDA(Exploratory Data Analysis): 시각화 및 통계 도구를 사용하여 데이터를 이해하기 위한 탐색적 데이터 분석
3. R-squared(결정계수): 목표 분산과 목표분산에 대한 예측 오차 분산간의 차의 비율

01. 프로젝트 개요 - 용어정리

BMI

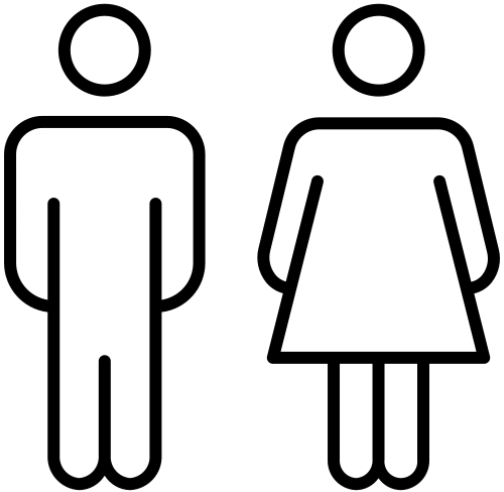
$$BMI = \frac{\text{몸무게(KG)}}{\text{키(M) X 키(M)}}$$



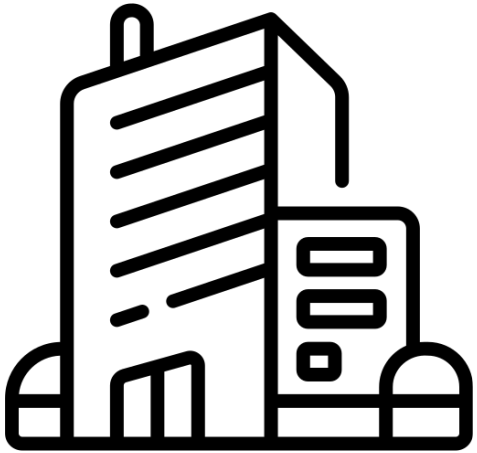
출처 : 미국 질병 통제 예방센터(CDC)
(https://www.cdc.gov)

01. 프로젝트 개요 - 목표

고객정보를 바탕의 보험료 예측



개인입장에서는 본인의 의료비를 직접 예측하여 과납을 막고, 그에 상응하는 보험금을 납부하도록 함



보험회사입장에서는 보험료 변화 추세를 파악하여, 그에 따른 맞춤형 상품을 기획할 수 있도록 함

01. 프로젝트 개요 - 데이터 소개

데이터셋 : Medical Cost Personal Datasets (Kaggle)

kaggle

Medical Cost Personal Datasets

Insurance Forecast by using Linear Regression

Data Card | Code (1063) | Discussion (10)

About Dataset

Context

Machine Learning with R by Brett Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book.

Usability ⓘ
8.82

License
Database: Open Database, Cont...

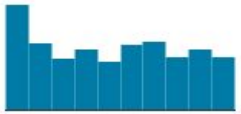
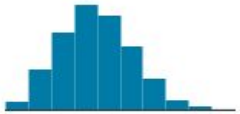
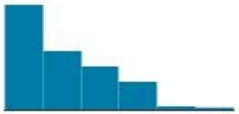
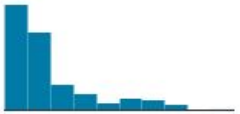
Expected update frequency
Not specified

01. 프로젝트 개요 - 데이터 소개

데이터셋 컬럼 체크

About this file

This dataset consists of 1338 rows.

# age	A sex	# bmi	# children	✓ smoker	A region	# charges
Edad del asegurado	Género	Indice de masa corporal	Número de hijos	Indicador si fuma	Región donde vive el asegurado	Prima del seguro
 18 64	male 51% female 49%	 16 53.1	 0 5	true 0 0% false 0 0%	southeast 27% southwest 24% Other (649) 49%	 1.12k 63.8k

Age - 보험 계약자(User)의 나이

Sex - 보험 계약자(User) 성별(여성, 남성)

Bmi - 체질량지수

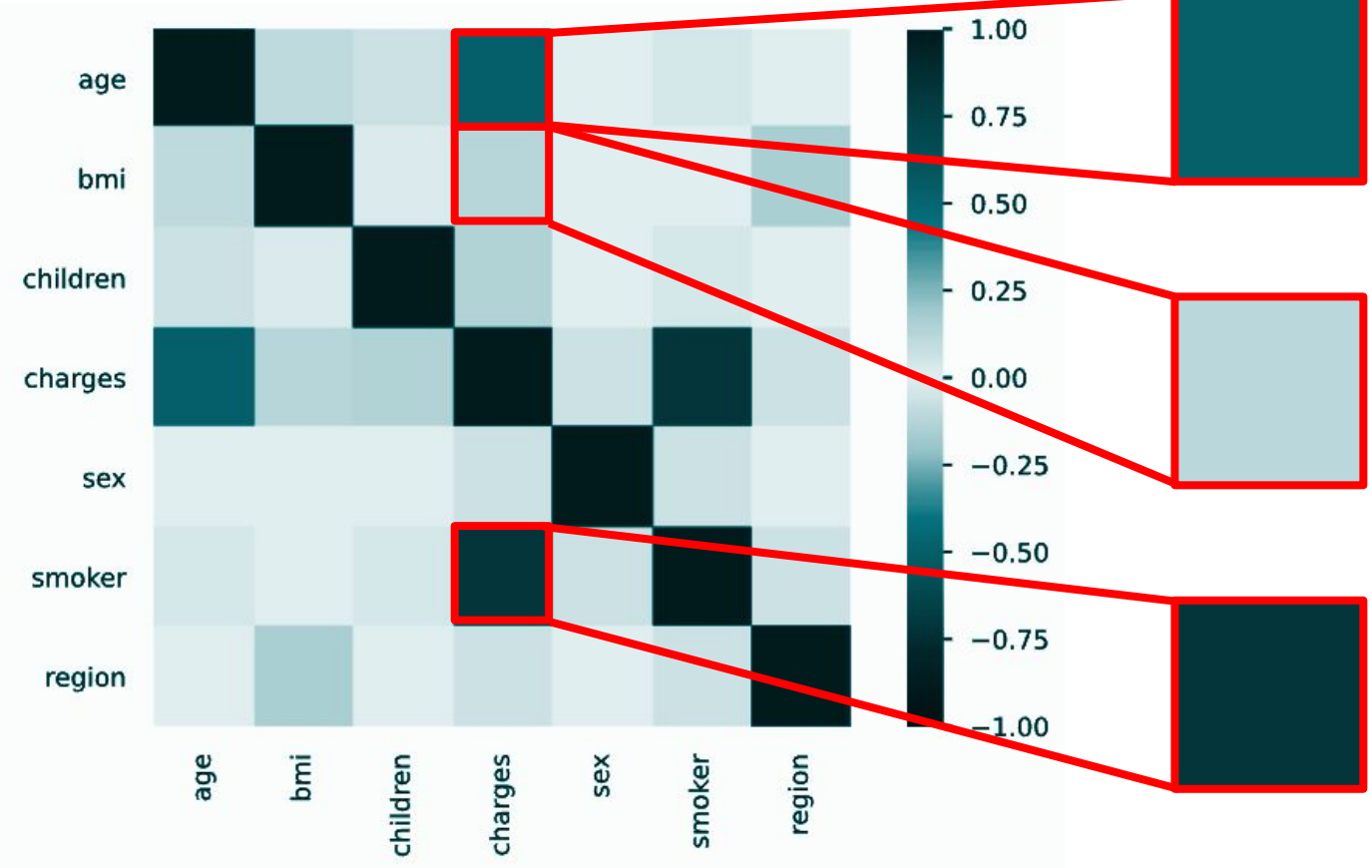
Children - 건강보험 적용 자녀 수/ 피부양자 수

Smoker - 흡연 여부

Region - 미국 수혜자(User) 주거 지역
(북동쪽, 남동쪽, 남서쪽, 북서쪽)

Charges - 건강보험에서 청구하는 개별 의료비

02. 데이터 탐색 및 전처리 - 히트맵

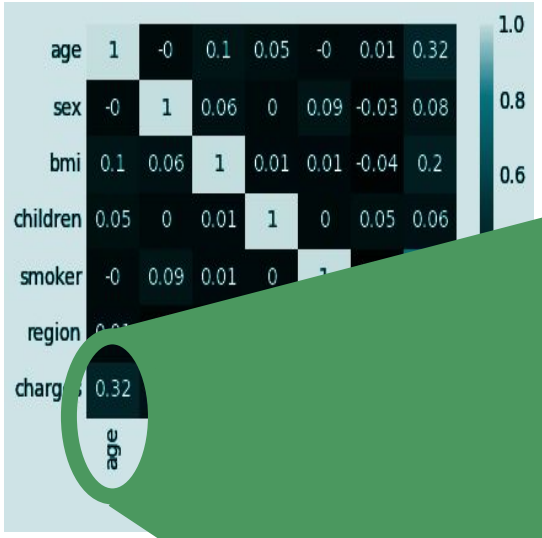


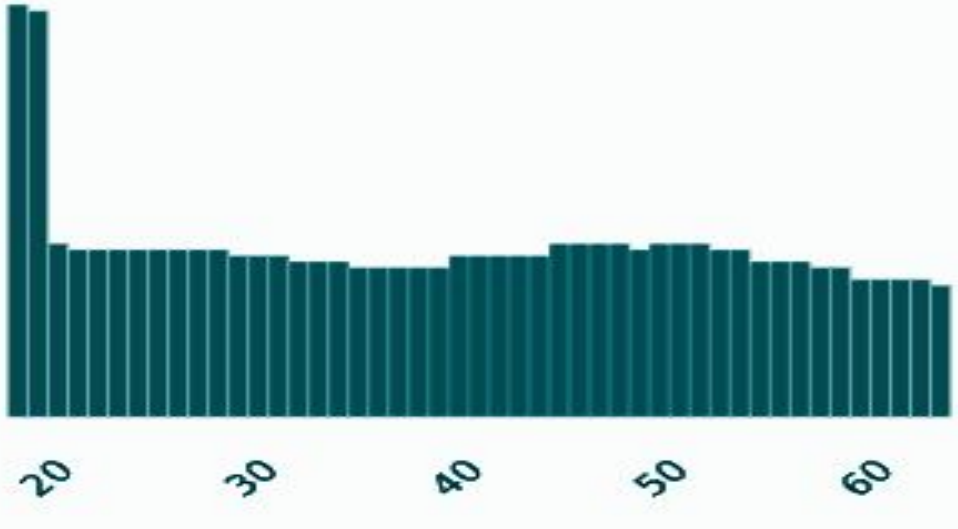
age
나이 : 0.32

BMI
체지방률 : 0.2

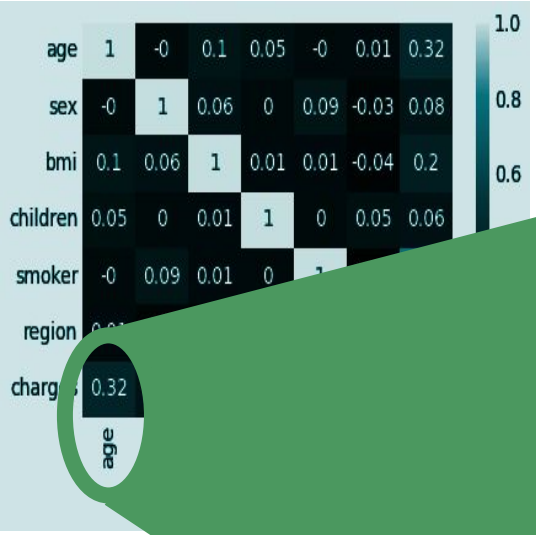
smoker
흡연 : 0.79

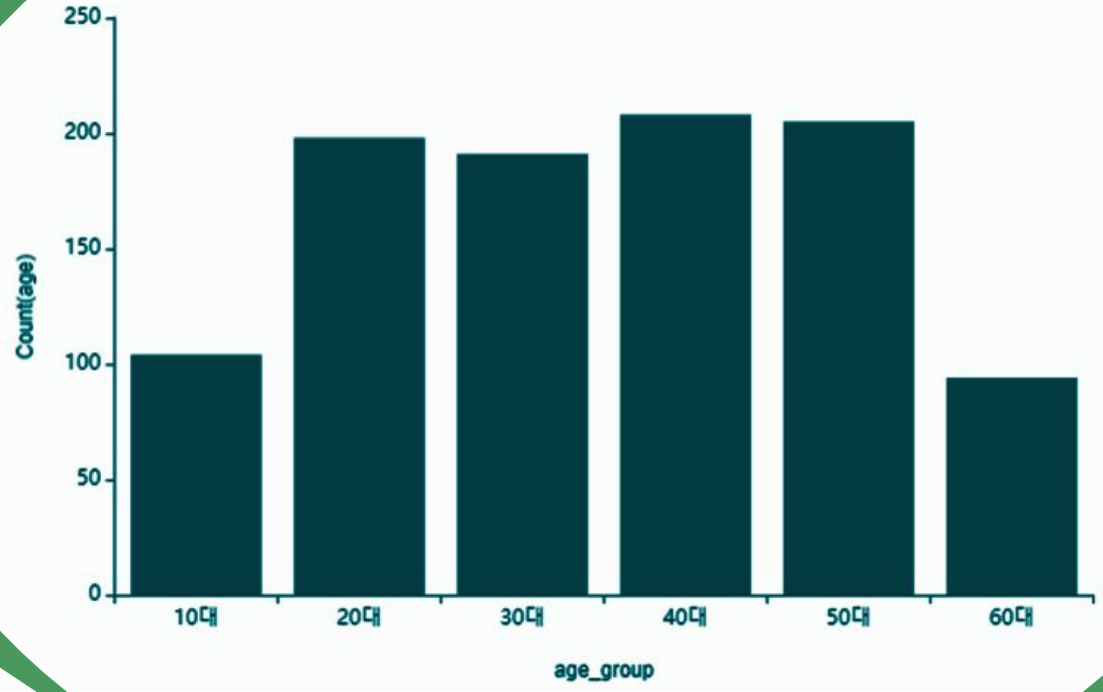
02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼



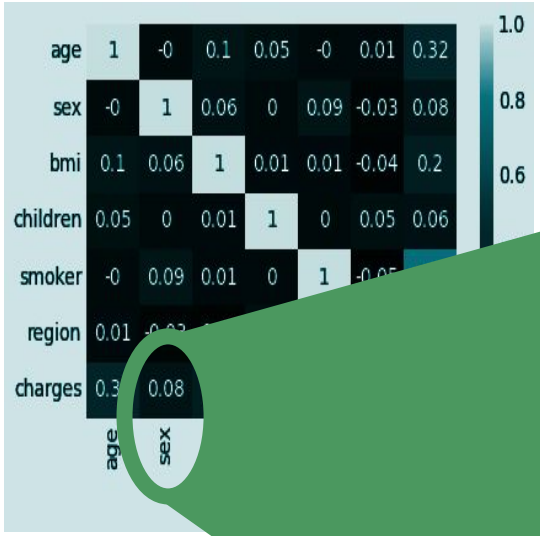


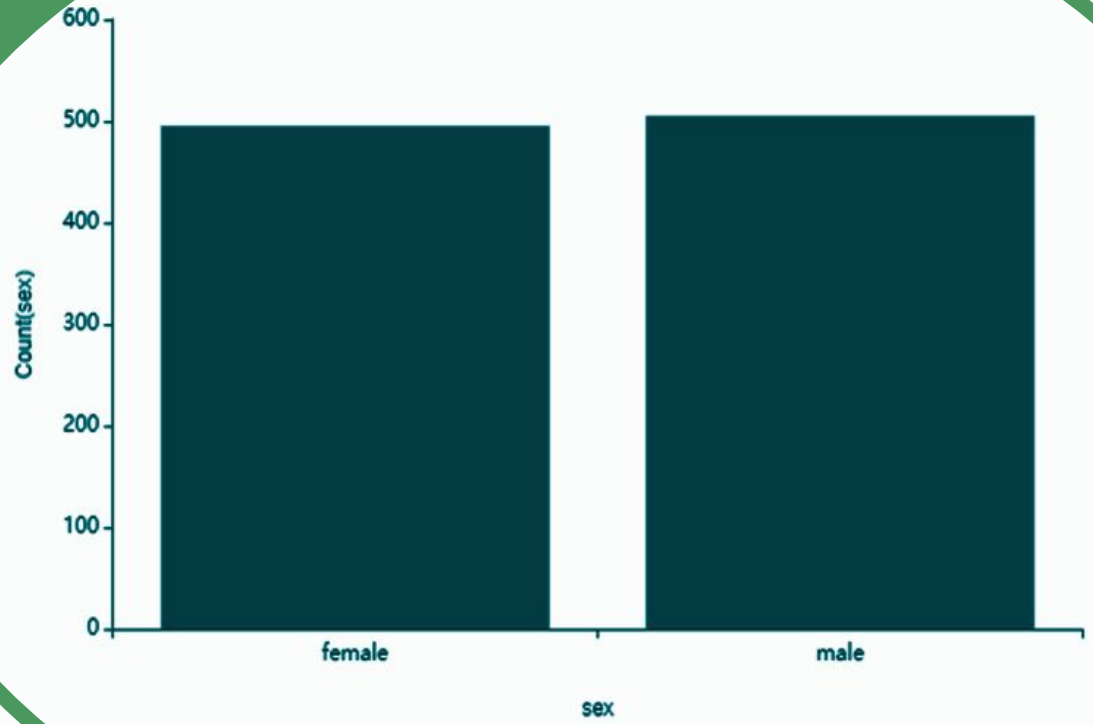
02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼



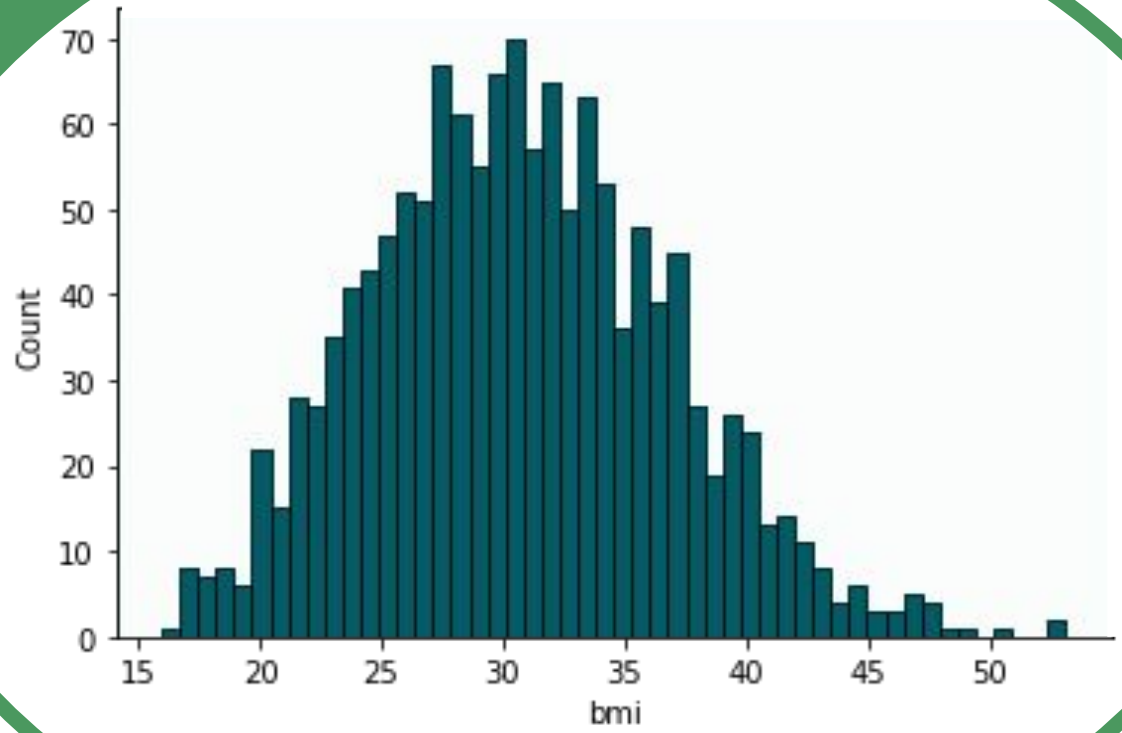


02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼

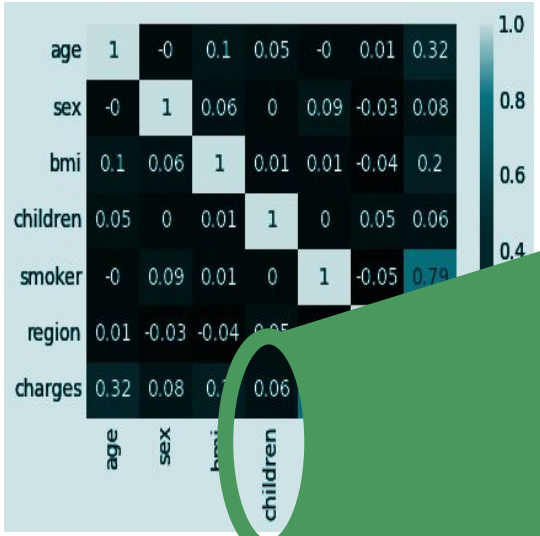


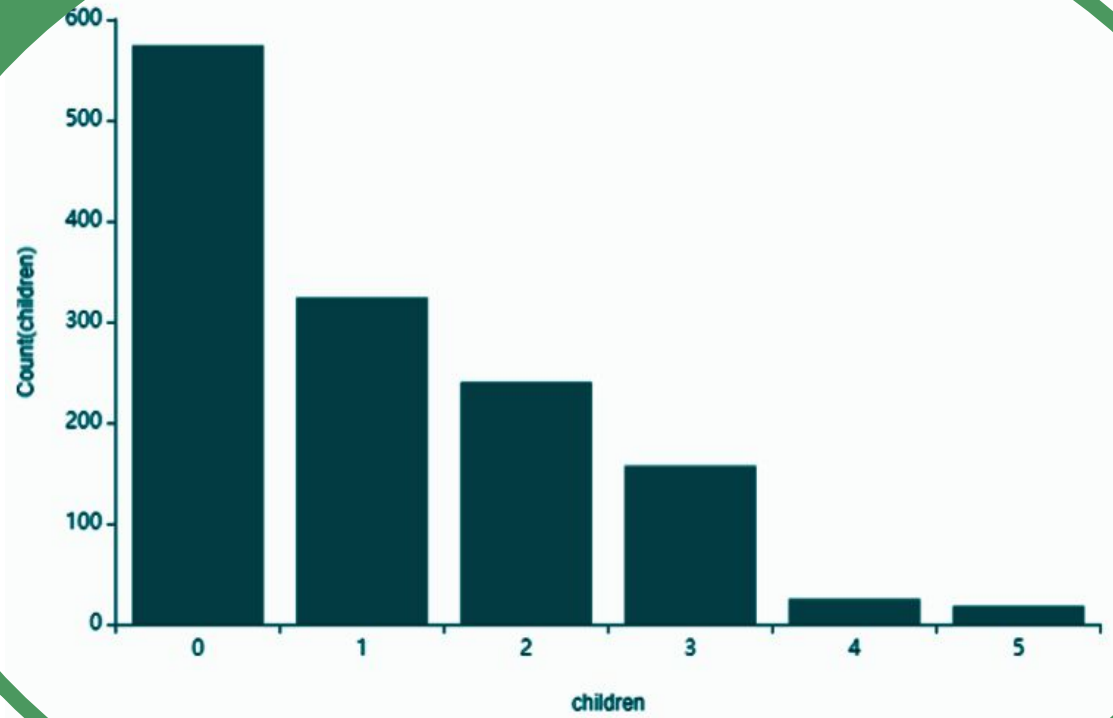


02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼

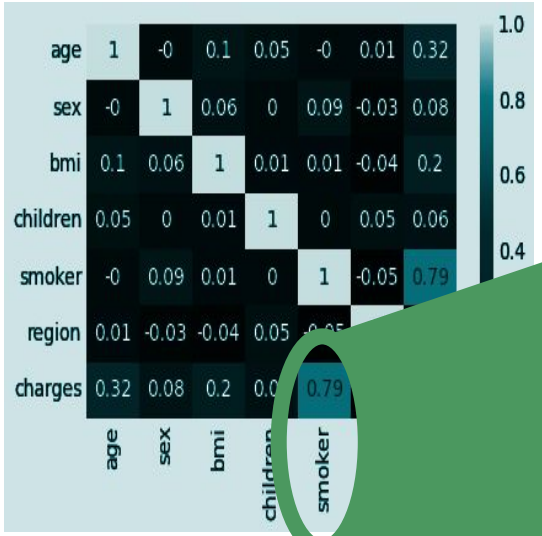


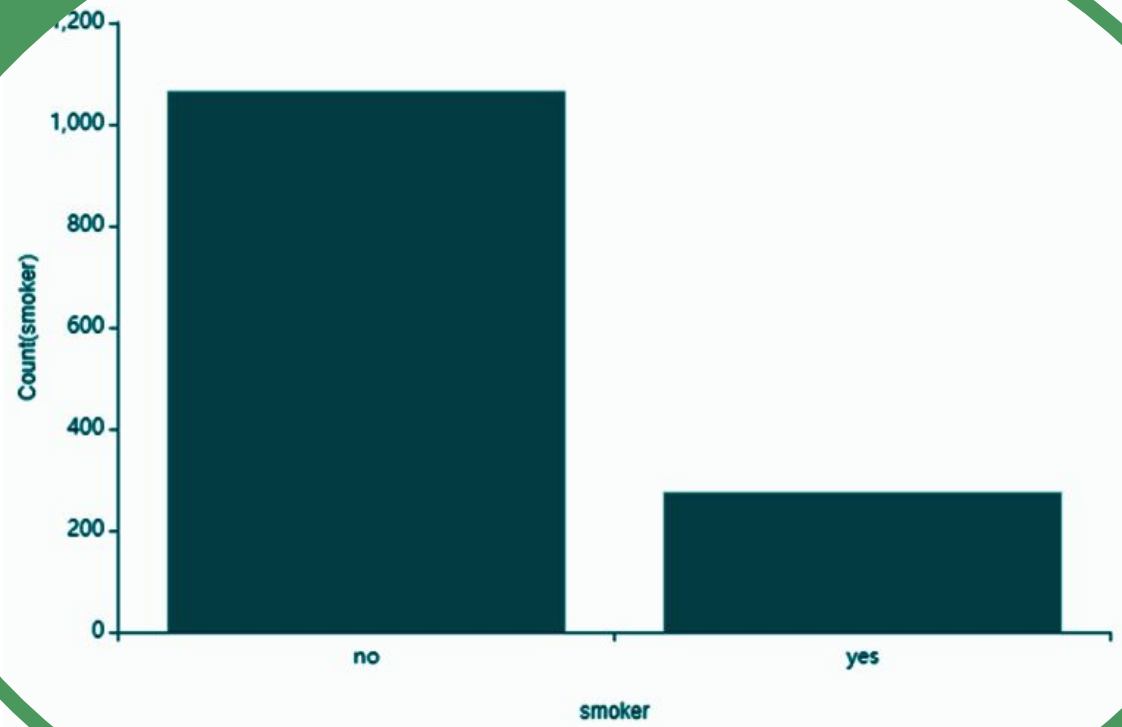
02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼



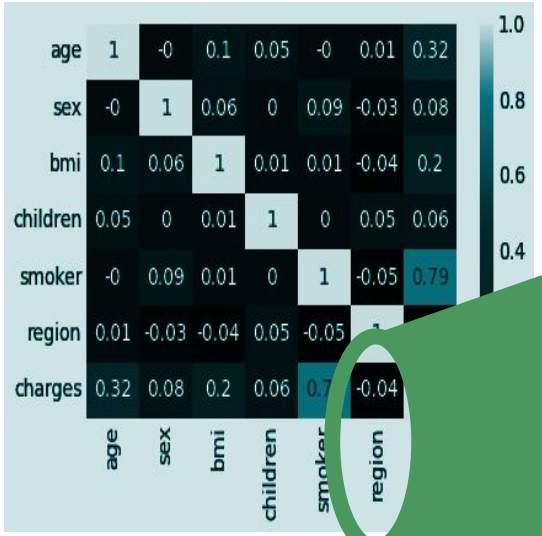


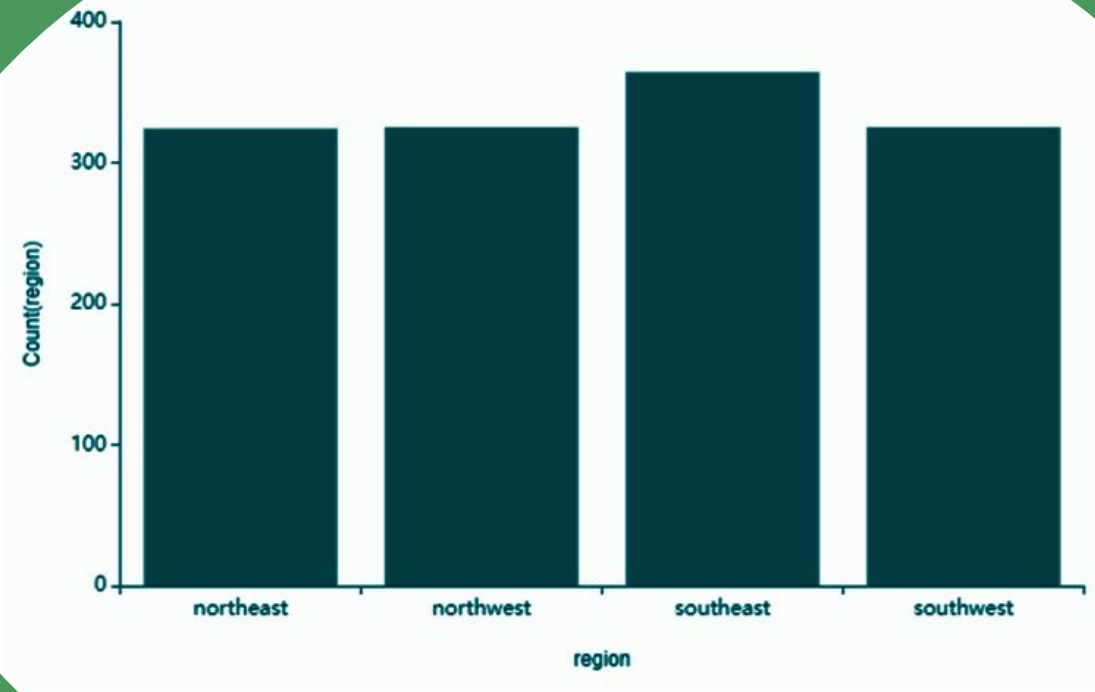
02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼



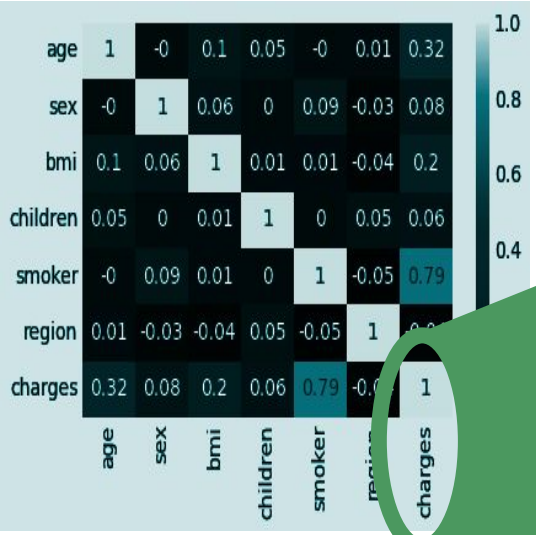


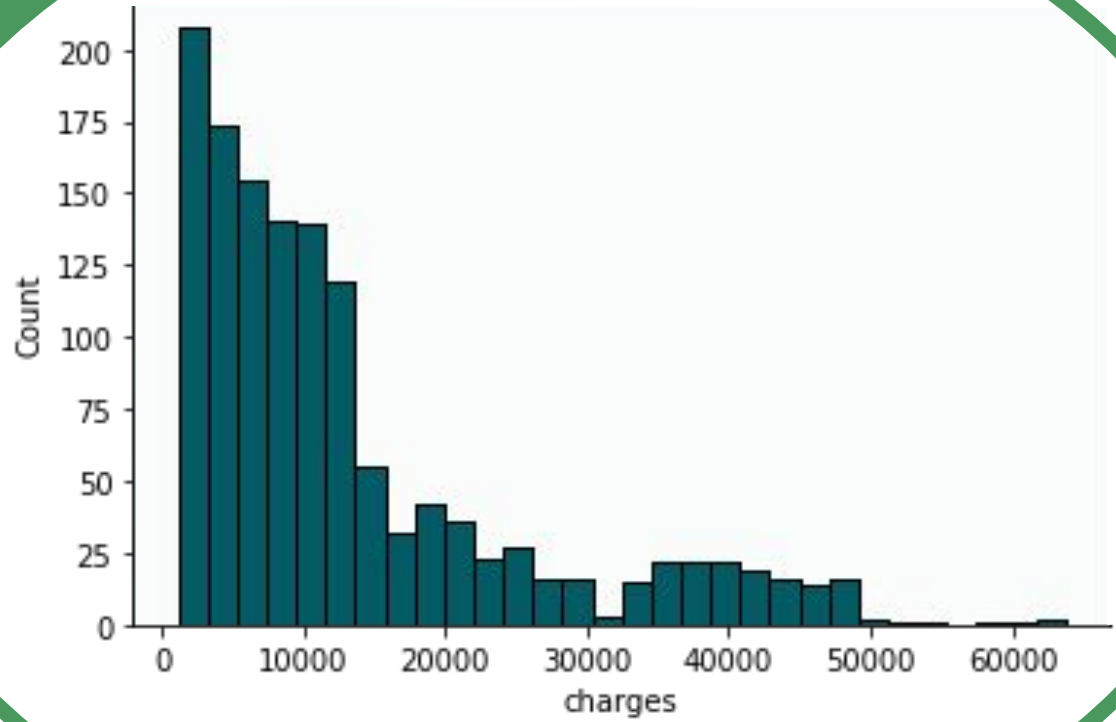
02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼



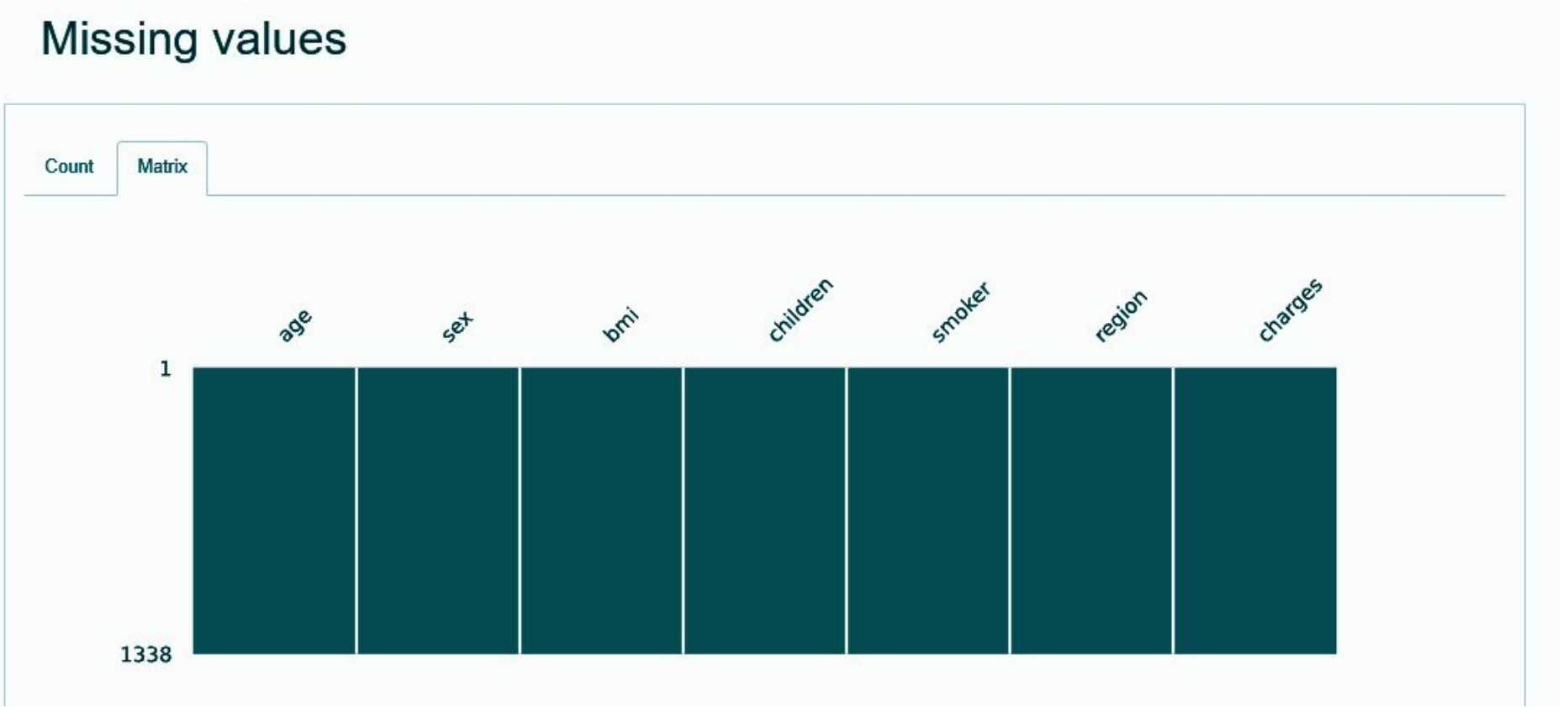


02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼

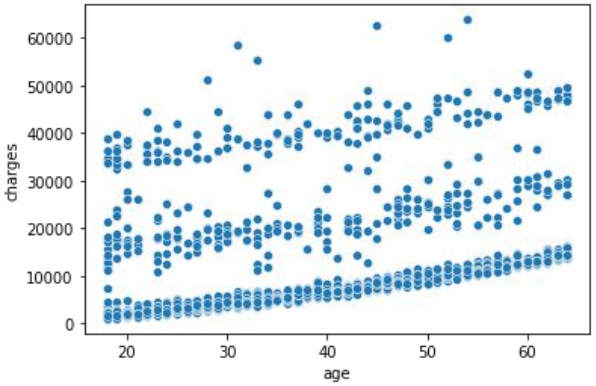


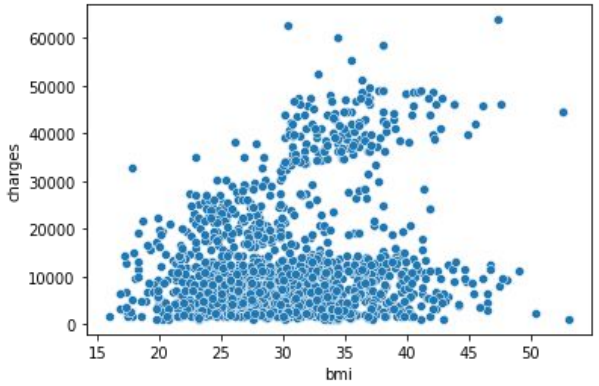


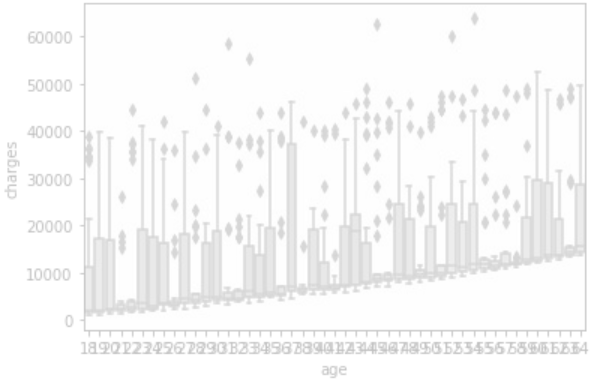
02. 데이터 탐색 및 전처리 - 결측치 분석

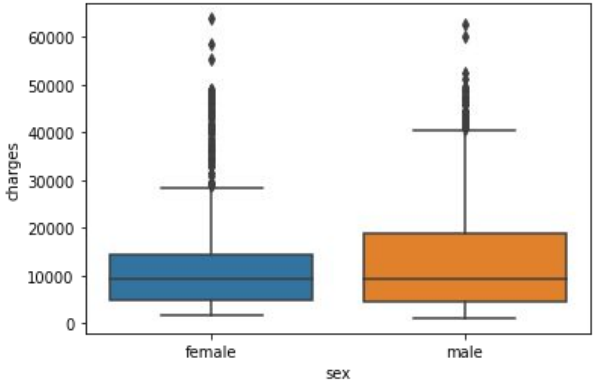


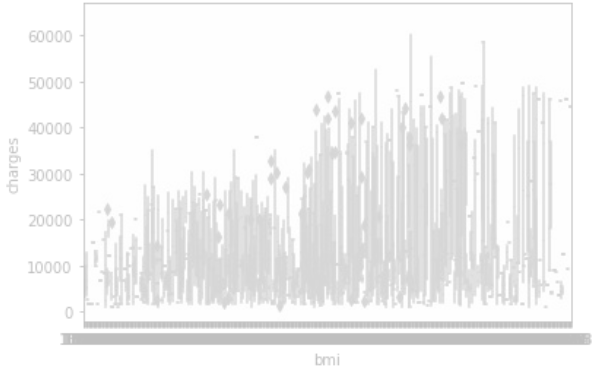
02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼2



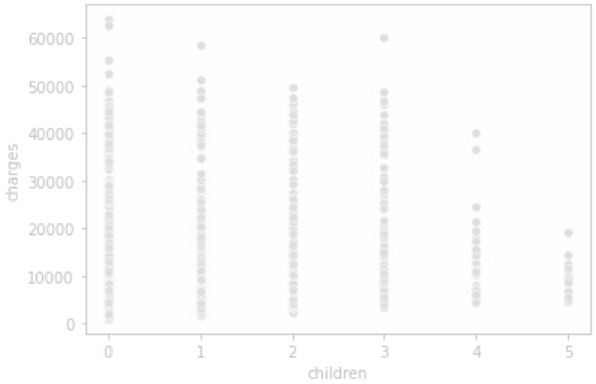


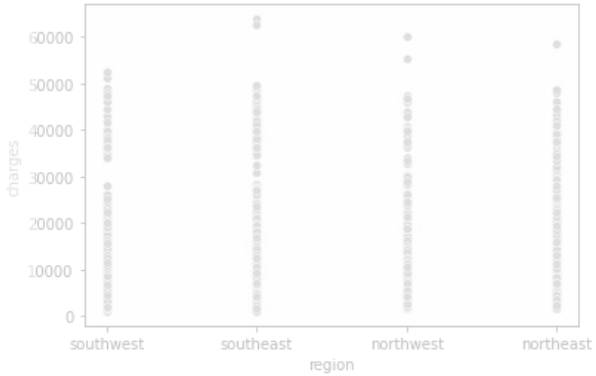


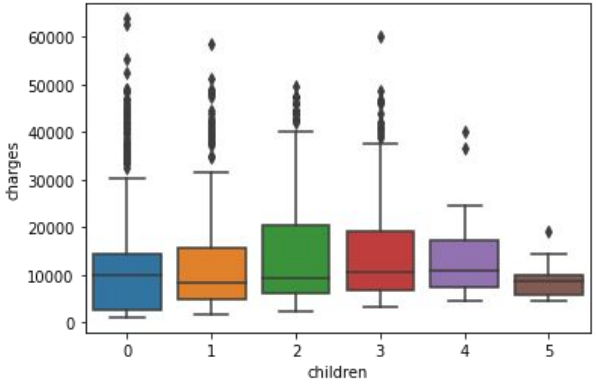


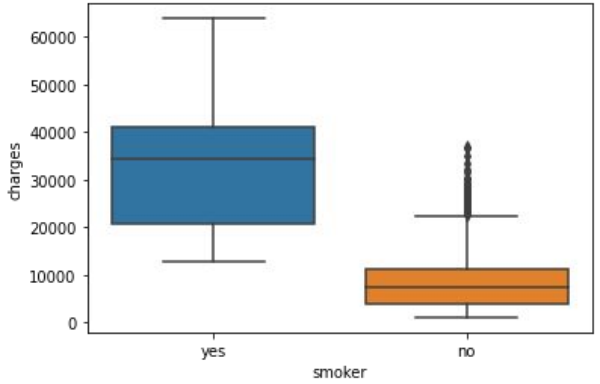


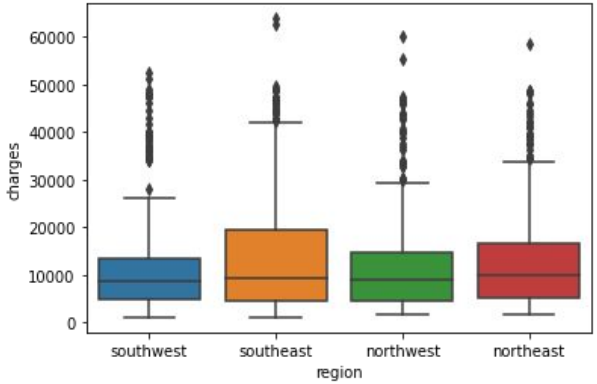
02. 데이터 탐색 및 전처리 - 데이터셋의 칼럼2



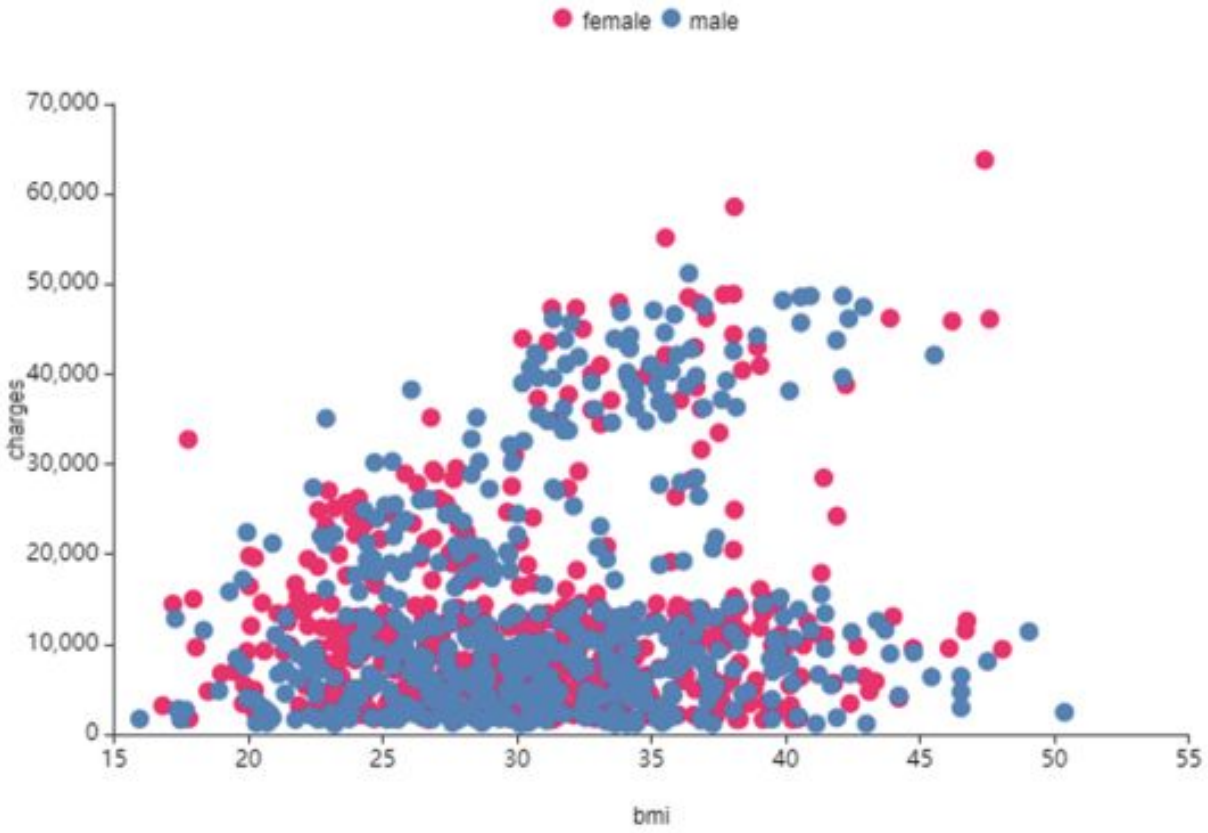




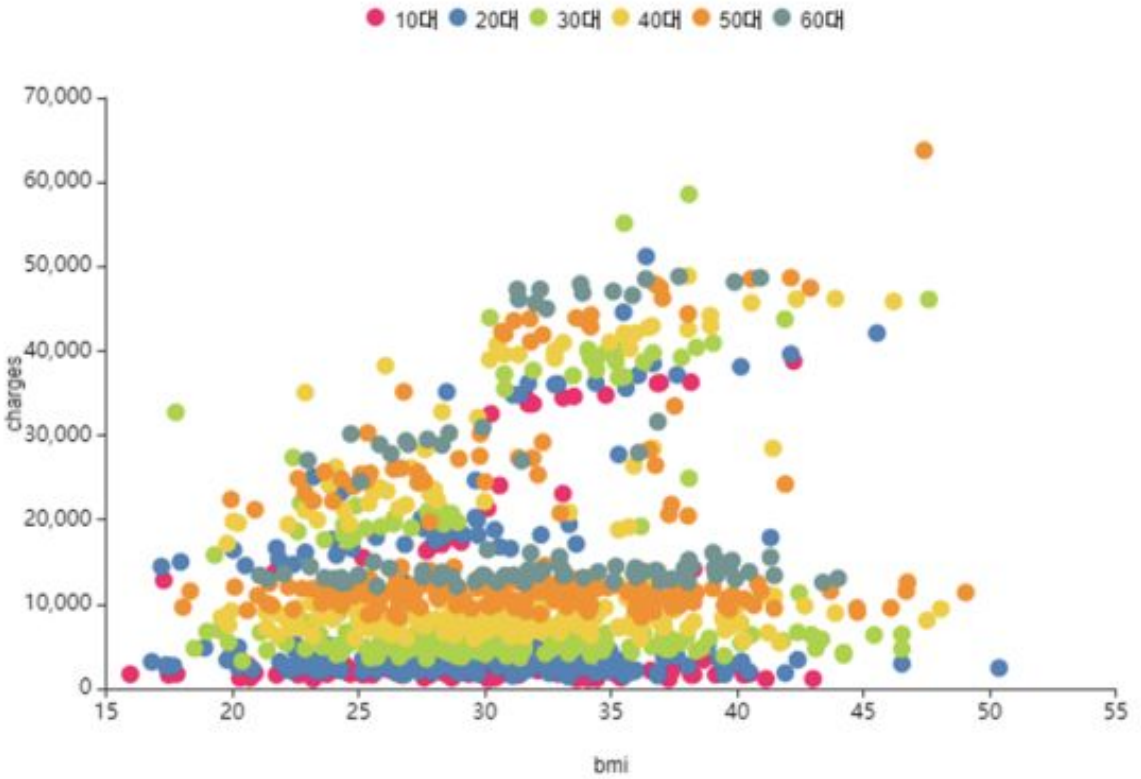




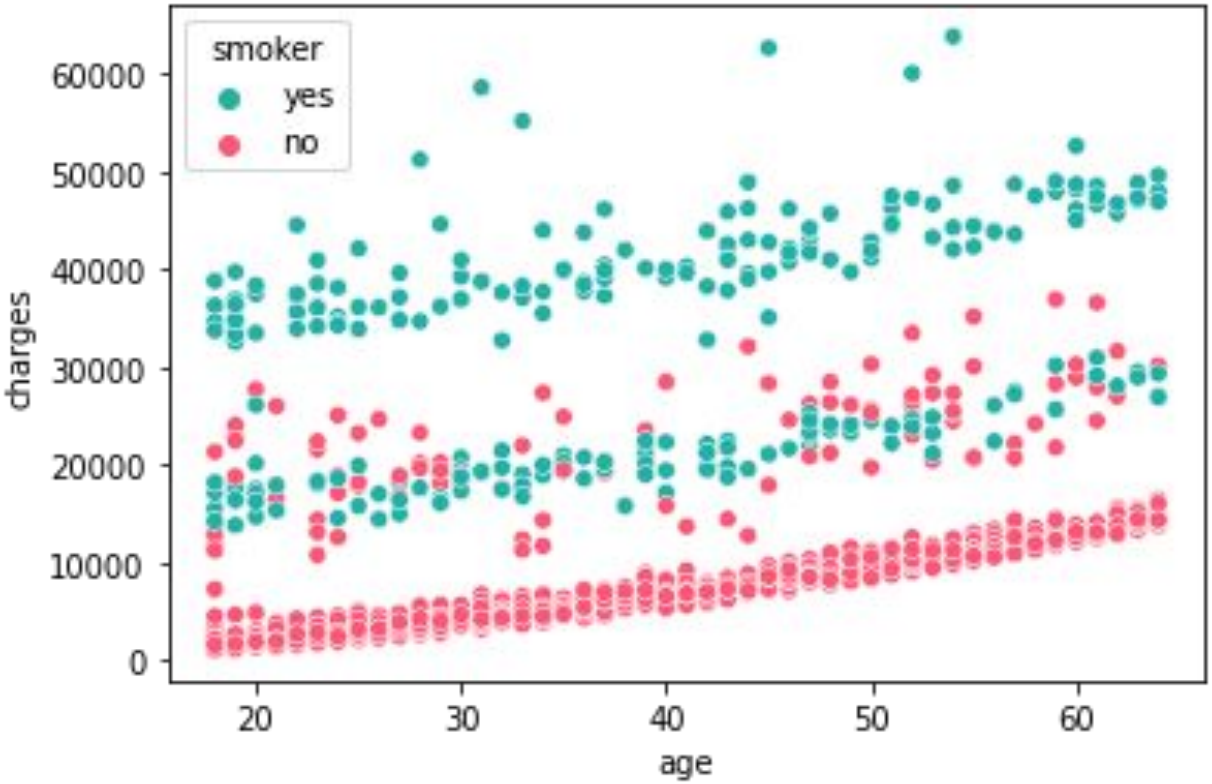
02. 데이터 탐색 및 전처리 - EDA 성별 기준



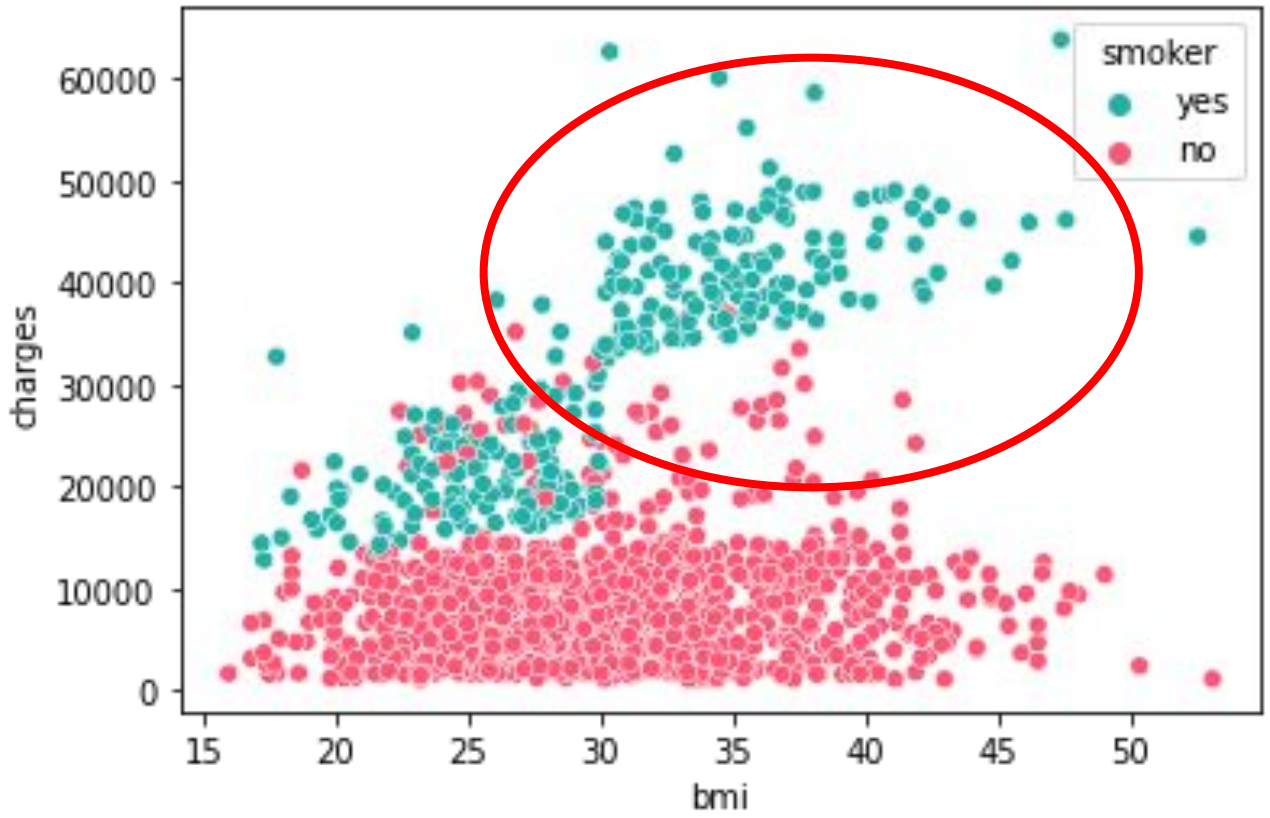
02. 데이터 탐색 및 전처리 - 복합적 EDA (bmi, age 기준)



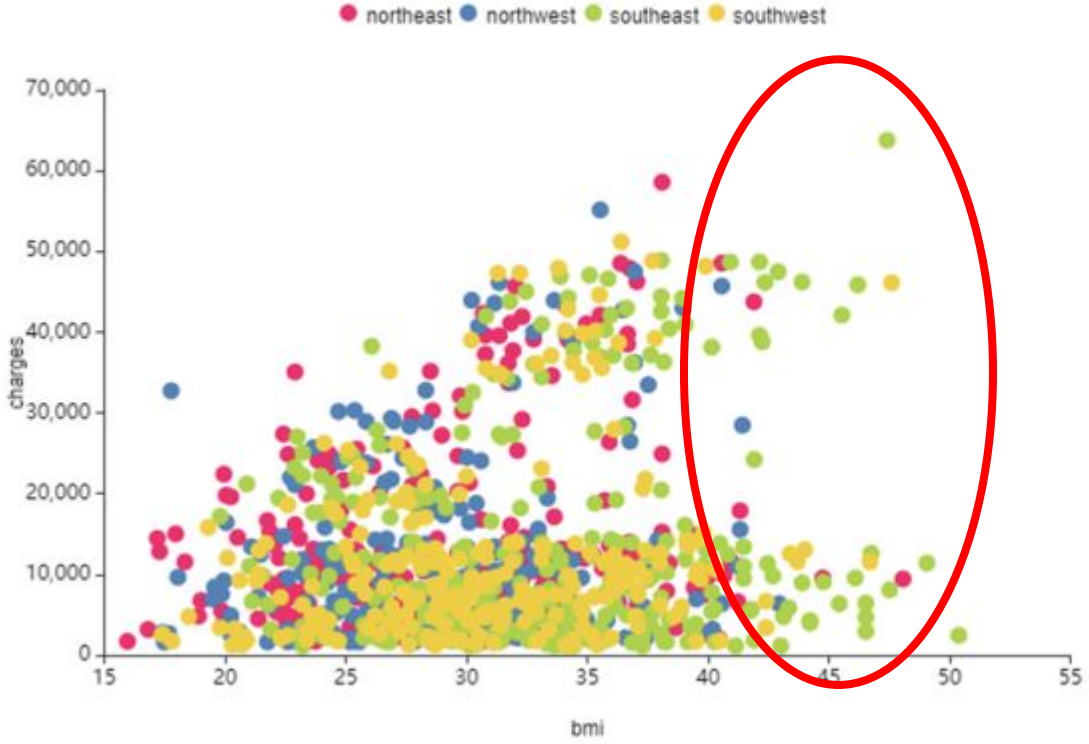
02. 데이터 탐색 및 전처리 - 기초통계량



02. 데이터 탐색 및 전처리 - EDA 흡연 기준

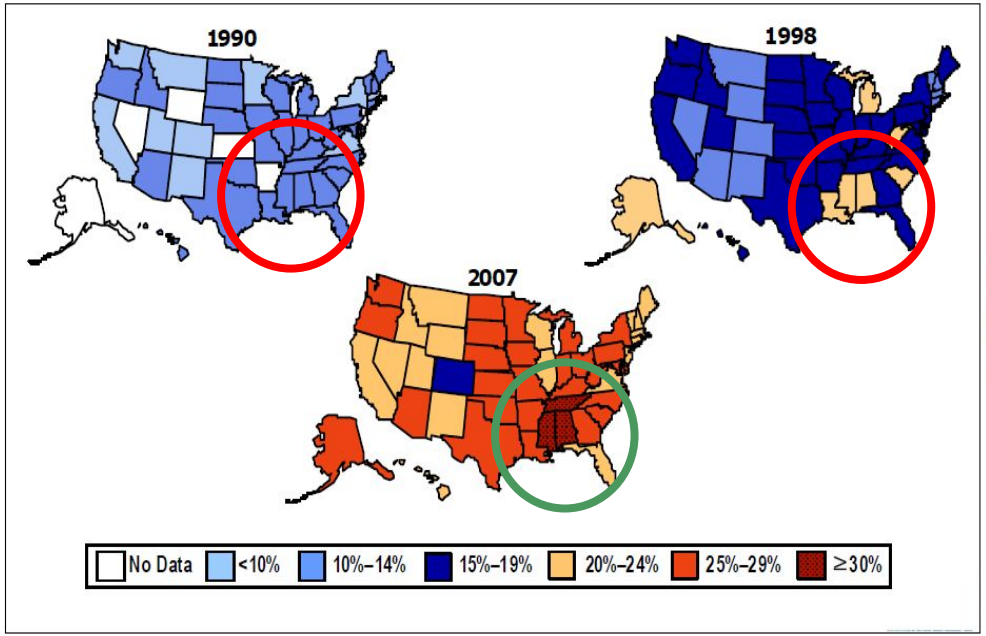


02. 데이터 탐색 및 전처리 - EDA - 지역별



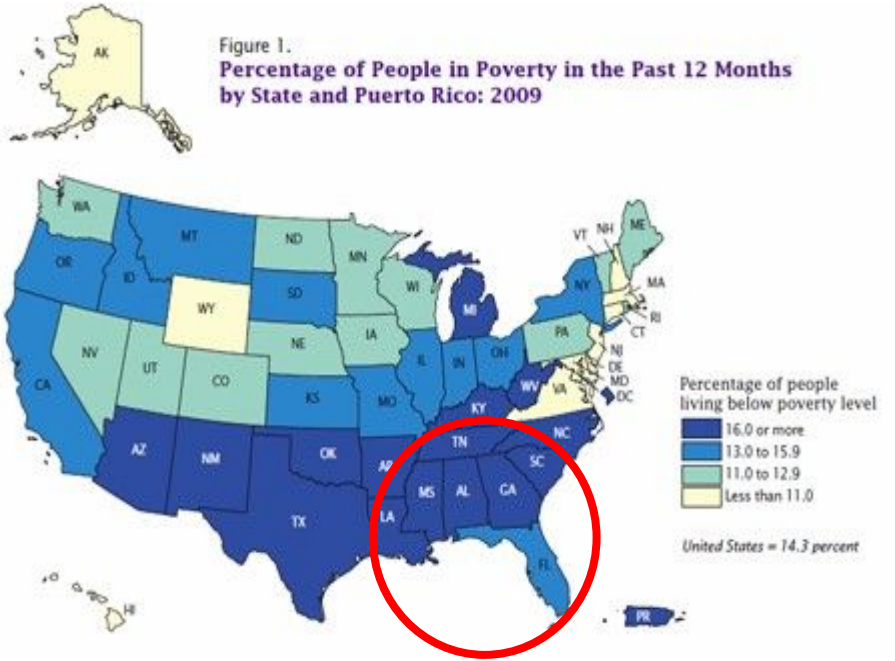
02. 데이터 탐색 및 전처리 - EDA - 지역별

미국 각 주별 비만율 (1990 - 2007)



주: CDC Behavioral Risk Factor Surveillance System.

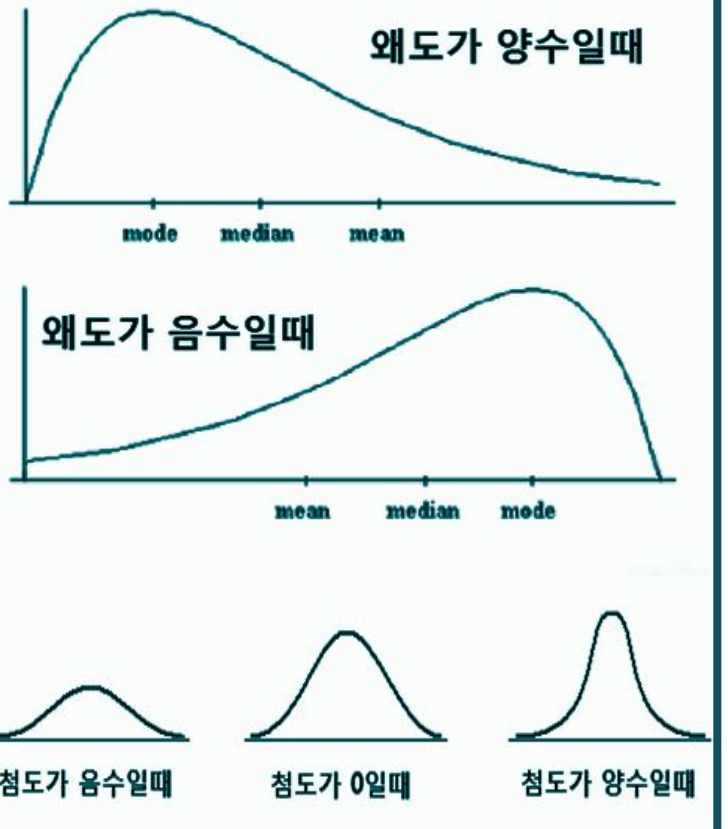
미국 각 주별 빈곤율 (2009년 기준)



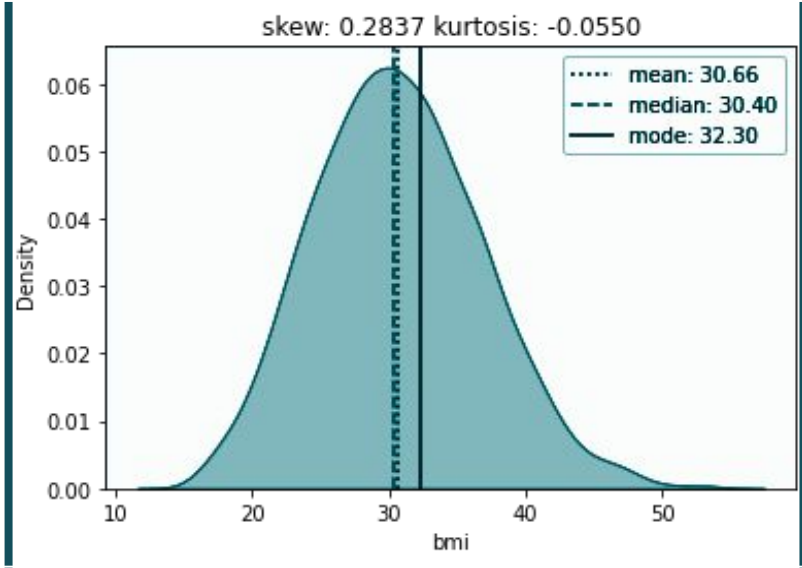
출처: 김은정, "액티브 리빙 리서치(Active Living Research) - 미국의 비만억제 노력과 우리나라에 대한 시사점",국토정책 Brief, 2008-09-22

02. 데이터 탐색 및 전처리 - 왜도, 첨도

이론



실습



skew : 0.2837

kurtosis : -0.055

```
data.isnull().sum()
```

✓ 0.0s

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

결측치 확인

```
data.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

데이터 타입 확인

데이터 전처리 과정 추가

{'sex':{'female':0, 'male':1},

'smoker': {'no': 0, 'yes':1},

'region':{'northeast':1, 'southeast':2, 'southwest':3, 'northwest':4}}

'object' -> 'int64'

X_train, X_test, y_train, y_test

= train_test_split(X, y, test_size=0.2, random_state=2023)

02. 데이터 탐색 및 전처리 - 데이터 전처리

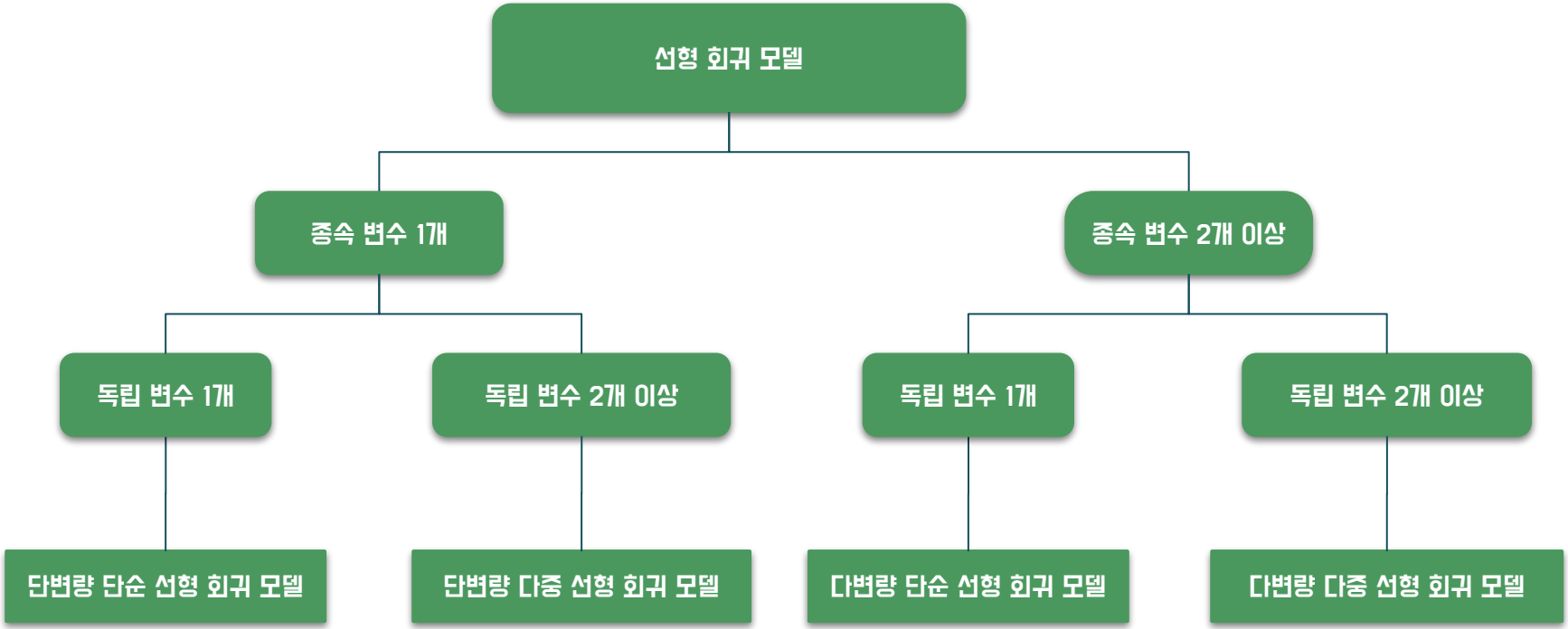
```
# object 형태의 타입을 수치형 데이터로 치환
data = data.replace({'sex':{'female':0, 'male':1},
                     'smoker': {'no': 0, 'yes':1},
                     'region':{'northeast':1, 'southeast':2, 'southwest':3, 'northwest':4}})
data.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   int64
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   int64
 5   region    1338 non-null   int64
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(5)
memory usage: 73.3 KB
```

object 형태의 데이터를 int형태로 치환

03. 모델링 - 선형회귀분석



03. 모델링

선택된 변수들을 바탕으로 다중 선형 회귀 모델을 구축

이 모델은 나이, 성별, 흡연 여부, BMI 등의 변수를 입력받아 의료비용을 예측

결과적으로 모델의 R-squared 값은 0.75로 나타남

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	998.1
Date:	Fri, 24 Mar 2023	Prob (F-statistic):	0.00
Time:	13:02:27	Log-Likelihood:	-13551.
No. Observations:	1338	AIC:	2.711e+04
Df Residuals:	1333	BIC:	2.714e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.21e+04	941.984	-12.848	0.000	-1.4e+04	-1.03e+04
age	257.8495	11.896	21.675	0.000	234.512	281.187
bmi	321.8514	27.378	11.756	0.000	268.143	375.559
smoker	2.381e+04	411.220	57.904	0.000	2.3e+04	2.46e+04
children	473.5023	137.792	3.436	0.001	203.190	743.814

Omnibus:	301.480	Durbin-Watson:	2.087
Prob(Omnibus):	0.000	Jarque-Bera (JB):	722.157
Skew:	1.215	Prob(JB):	1.53e-157
Kurtosis:	5.654	Cond. No.	292.

04. 결과 및 해석

우리가 구축한 모델은 Medical Cost Personal Datasets을 기반으로 보험 비용을 예측하는 데 있어서 높은 정확도를 보임

특히나 나이, BMI, 흡연 여부 등의 변수는 의료비용 예측에 큰 영향을 미치는 것으로 나타남

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	998.1
Date:	Fri, 24 Mar 2023	Prob (F-statistic):	0.00
Time:	13:02:27	Log-Likelihood:	-13551.
No. Observations:	1338	AIC:	2.711e+04
Df Residuals:	1333	BIC:	2.714e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.21e+04	941.984	-12.848	0.000	-1.4e+04	-1.03e+04
age	257.8495	11.896	21.675	0.000	234.512	281.187
bmi	321.8514	27.378	11.756	0.000	268.143	375.559
smoker	2.381e+04	411.220	57.904	0.000	2.3e+04	2.46e+04
children	473.5023	137.792	3.436	0.001	203.190	743.814

Omnibus:	301.480	Durbin-Watson:	2.087
Prob(Omnibus):	0.000	Jarque-Bera (JB):	722.157
Skew:	1.215	Prob(JB):	1.53e-157
Kurtosis:	5.654	Cond. No.	292.

05. 의의 및 시사점

이번 연구는 Medical Cost Personal Datasets을 활용하여 보험 비용을 예측하는 데 있어서 유용한 정보를 제공함

보험사는 이러한 예측 모델을 활용하여 보다 정확한 보험료를 책정할 수 있으며, 보험 계약자는 자신의 위험도를 파악하여 적절한 보험 상품을 선택할 수 있음

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	664.8
Date:	Fri, 24 Mar 2023	Prob (F-statistic):	0.00
Time:	11:47:28	Log-Likelihood:	-13551.
No. Observations:	1338	AIC:	2.712e+04
Df Residuals:	1331	BIC:	2.715e+04
Df Model:	6		
Covariance Type:	nonrobust		

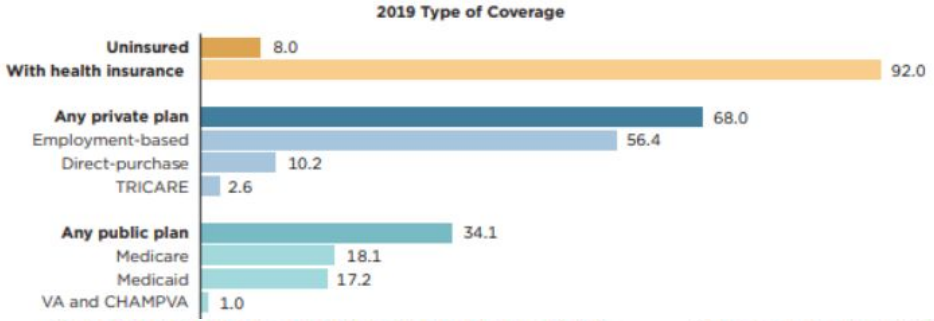
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.178e+04	1038.851	-11.340	0.000	-1.38e+04	-9742.634
age	257.7804	11.907	21.650	0.000	234.423	281.138
sex	-129.5026	333.435	-0.38	0.698	-783.618	524.613
bmi	321.3813	27.466	11.701	0.000	267.500	375.263
children	477.7866	137.983	3.463	0.001	207.099	748.474
smoker	2.381e+04	413.007	57.654	0.000	2.3e+04	2.46e+04
region	-98.2718	150.749	-0.652	0.515	-394.003	197.459

Omnibus:	301.228	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	722.385
Skew:	1.214	Prob(JB):	1.37e-157
Kurtosis:	5.658	Cond. No.	322.

06. 한계점 및 향후 발전 방향 - 한계점

자녀수에 따른 보험료 경향성

Figure 1.
Percentage of People by Type of Health Insurance Coverage: 2019
(Population as of March 2020)



보험 플랜	Bronze	Silver	Gold	Platinum
보험사가 내는 의료비 비율	60%	70%	80%	90%
보험 가입자가 내는 의료비 비율	40%	30%	20%	10%

〈그림 1〉 미국 건강보험 제도개혁 주요 내용과 의료비 영향

〈그림 2〉 국가별 GDP 대비 의료비 비중

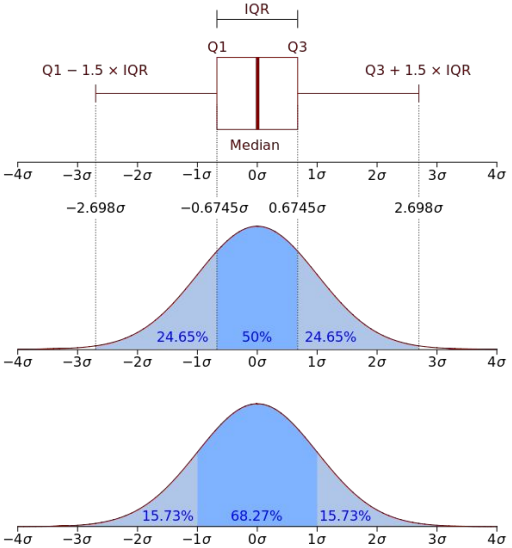
자료: OECD Health Expenditure.

출처:https://medicalhani.com/2021/04

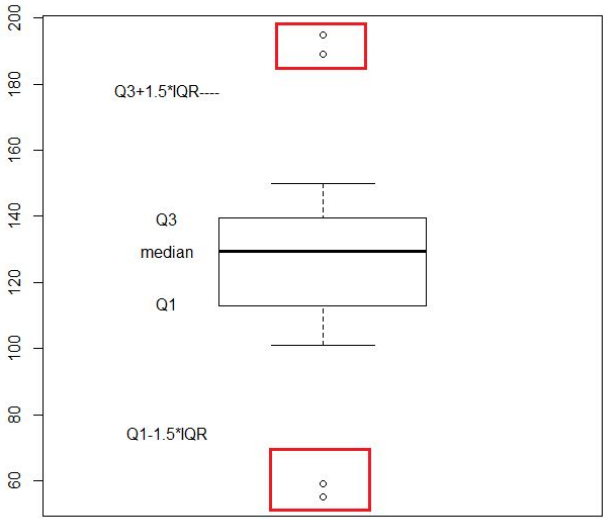
출처 :
http://www.dailypharm.com/Users/News/NewsView

06. 한계점 및 향후 발전 방향 - 향후 발전 방향 및 보완점

이상치 제거



출처 : https://medium.com/



출처 : https://sosal.kr/840

이상치(이상점, outlier)란, 관측된 데이터의 범위에서 많이 벗어난 아주 작은 값이나 아주 큰 값을 말하는데, 어떤 의사결정을 하는데 필요한 데이터를 분석할 경우 이렇게 이상한 값들에 의해서 의사결정에 영향을 미칠 수 있으므로 제거하는 것이 좋음

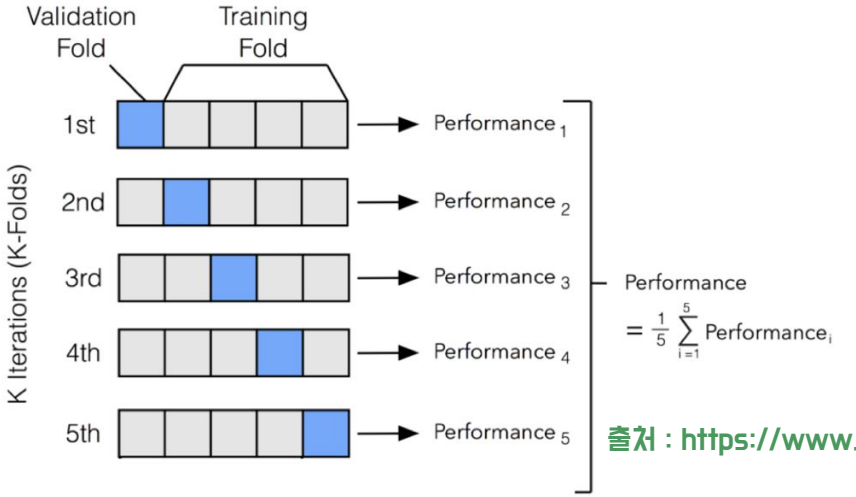
06. 한계점 및 향후 발전 방향 - 향후 발전 방향 및 보완점

K-fold 교차검증

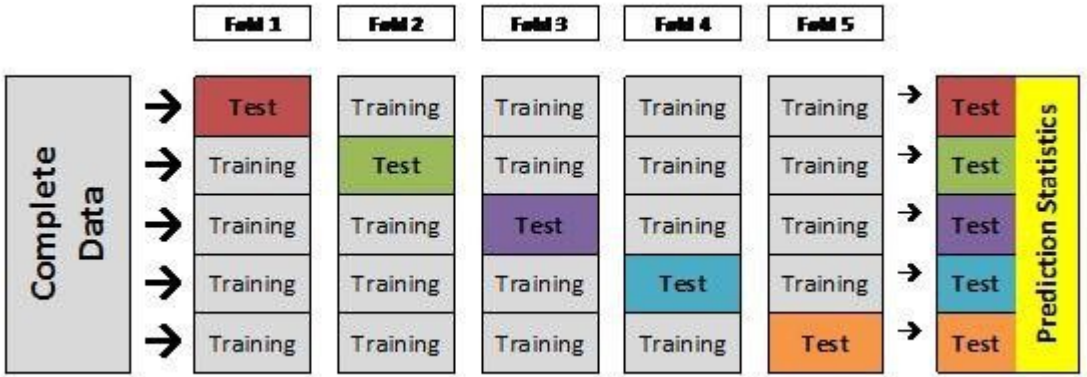
k-fold cross-validation
(k-겹 교차 검증)은 가장 널리
사용되는 교차 검증 방법의 하나로.

데이터를 k개로 분할한 뒤.
k-1개를 학습용(train) 데이터
세트로. 1개를 평가용(test) 데이터
세트로 사용하는데.

이 방법을 k번 반복하여 k개의 성능
지표를 얻어내는 방법



출처 : https://www.incodom.kr/



출처 : https://nonmeyet.tistory.com/entry/KFold-Cross-Validation

Question & Answer

(질의 응답)

고기4조
남태우,이수정,이정훈,정은정,정재석

Thank you!
(감사합니다!)

고기4조
남태우,이수정,이정훈,정은정,정재석