



Aichemist Train Session

CHAP 05 회귀(1)

회귀 실습 목차

01. 시작하기 전에

01-1. Guide-Line

01-2. Dataset 자료 다운 & 소개

02. Medical Cost Personal 워크북



01.

시작하기 전에



Guide-Line

- Github에서 Notebook file 다운 & Kaggle 사이트 참고하기
- 조원과 워크북 질문 답하기 & Jupyter Notebook 코드 빈칸 채우기
- 팀별 발표🌟

Medical Cost Personal Dataset 자료 다운

Medical Cost Personal Datasets

Insurance Forecast by using Linear Regression






Data Card Code (1432) Discussion (15) Suggestions (0)

insurance.csv (55.63 kB)



Detail Compact Column

7 of 7 columns

# age Edad del asegurado	sex Género	# bmi Índice de masa corporal	# children Número de hijos	smoker Indicador si fuma	region Región do asegurado
 18 64	male 51% female 49%	 16 53.1	 0 5	true 0 0% false 0 0%	southeast southwest Other (64)

Data Explorer

Version 1 (55.63 kB)

insurance.csv

insurance.csv 다운로드

Medical Cost Personal Dataset 소개

개인 의료 보험료

- 의료 보험 데이터를 활용해 한 사람이 보험료를 얼마나 낼지 예측하는 회귀 문제
- 여러 feature를 가진 사람의 보험료를 예측

Feature

- Age : 피보험자의 나이
- Sex : 피보험자의 성별
- BMI : 피보험자의 체질량 지수
- Children : 피보험자의 자녀의 수
- Smoker : 흡연 여부 (yes / no)
- Region : 피보험자가 거주하는 지역 (Southeast / Southwest / Northeast / Northwest)
- Charges : 보험료 -> 타겟값

02.

Medical Cost Personal 워크북

Medical Cost Personal 워크북

1

```
<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
```

2

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Q1. 데이터 분석 결과 1번, 2번이 각각 어떤 정보를 담고 있는지 정리하고 어떤 메서드(함수)를 사용하여 구한 것인지 말해보세요.

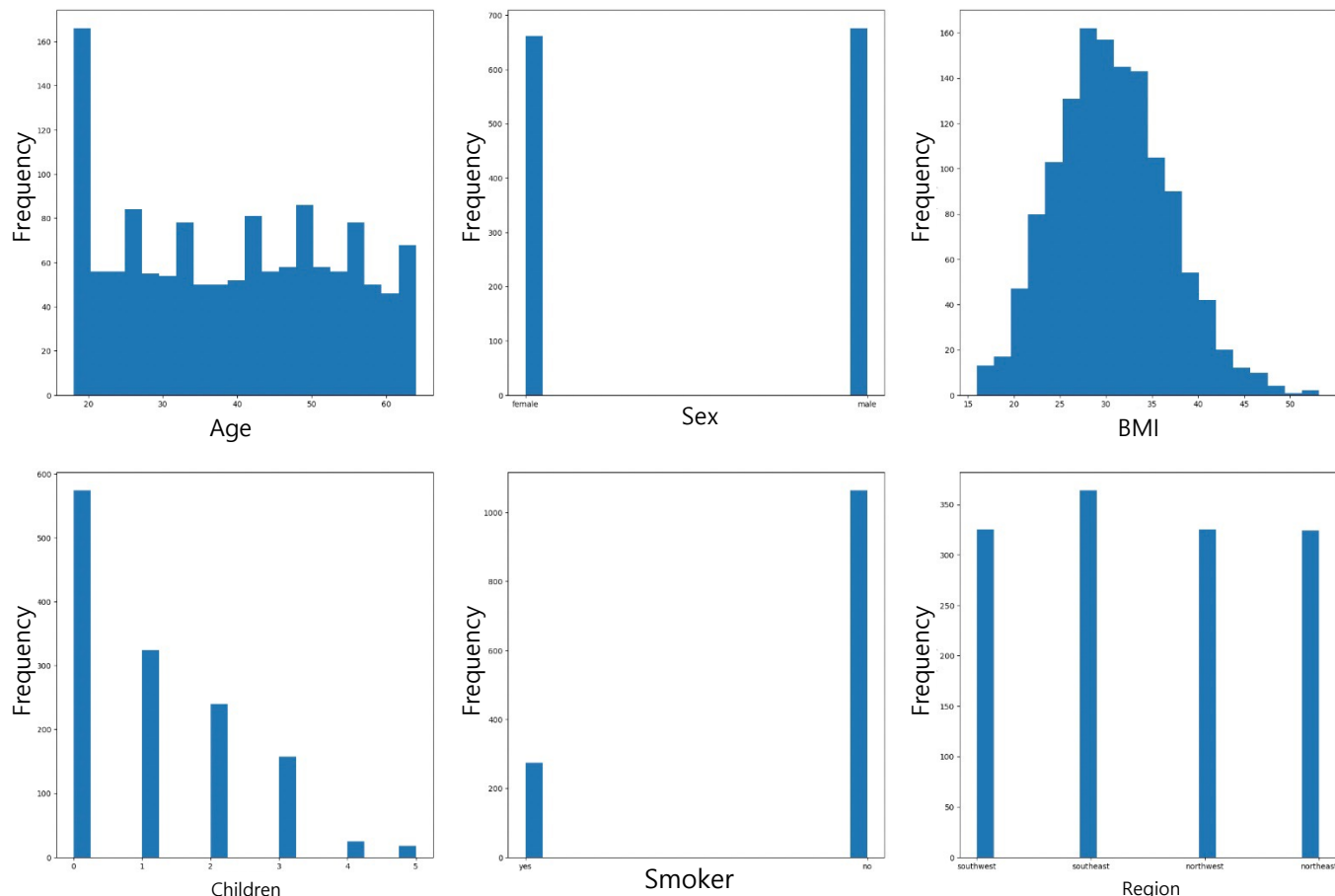
Q1-1. Dtype이 object인 것은 무엇을 의미하는가?

Q1-2. Dtype과 Null값 수를 보고 어떤 데이터 전처리를 진행해야 할지 말해보세요.

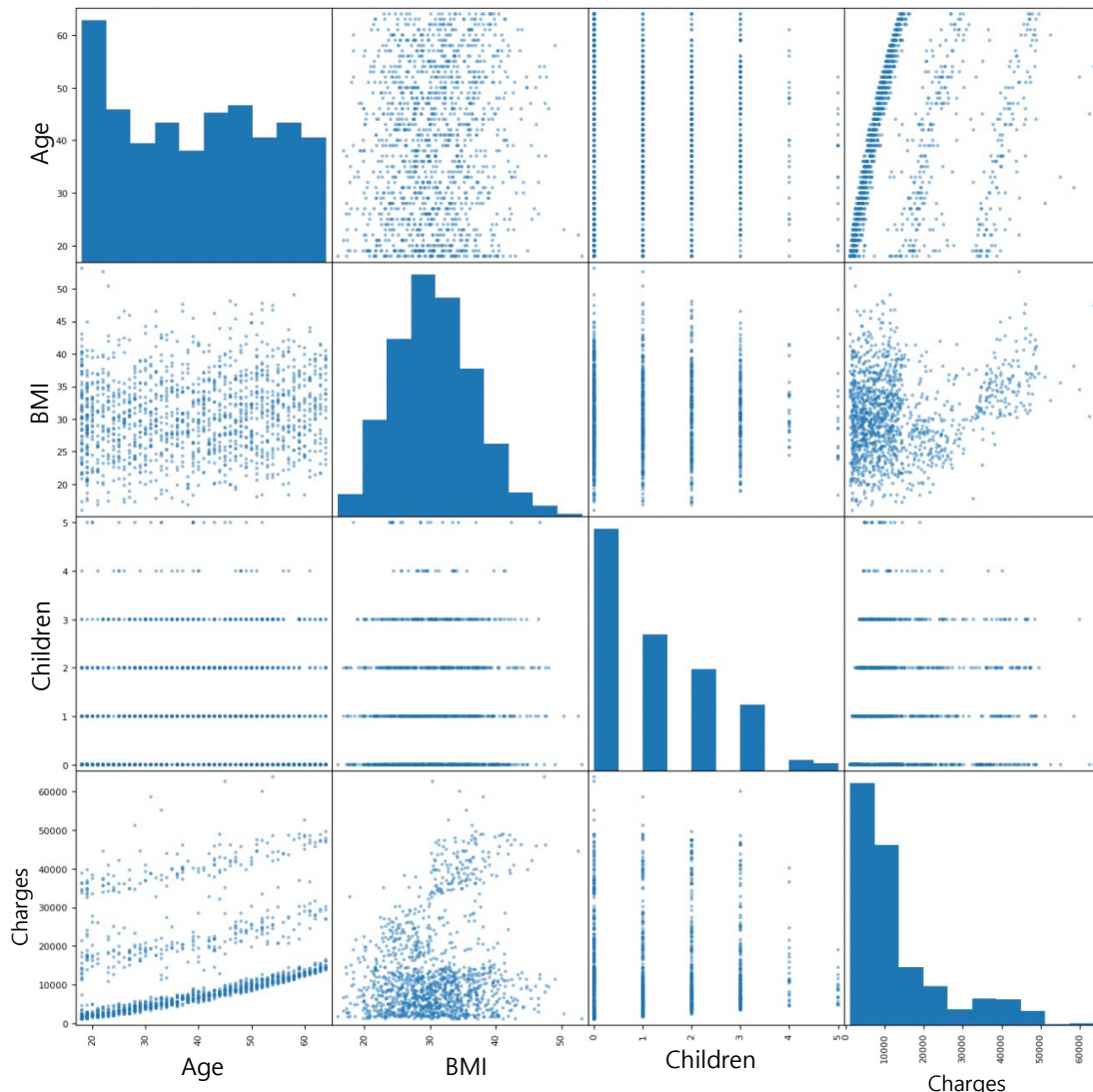
Medical Cost Personal 워크북

Q2. 변수들의 분포도를 시각화한 것을 보고 파악할 수 있는 정보를 나열하세요.

숫자형 / 카테고리형 칼럼 분류, 이상치, 피처별 불균형 정도와 피처별 분포 등을 기반으로



Medical Cost Personal 워크북



	age	bmi	children	charges
age	1.000000	0.109272	0.042469	0.299008
bmi	0.109272	1.000000	0.012759	0.198341
children	0.042469	0.012759	1.000000	0.067998
charges	0.299008	0.198341	0.067998	1.000000

Q3. 상관계수를 시각화한 것과 표로 정리한 것을 보면서

상관계수의 정의가 무엇인지 생각해보자.

Q3-1. 타겟값을 제외한 피쳐들간의 상관계수의 절댓값은 (클수록 / 작을수록)

이상적이다. 이유는?

Q3-2. 특정 피쳐와 타겟값의 상관계수의 절댓값이 (클수록 / 작을수록)

특정 피쳐가 타겟값에 큰 영향을 미친다.

Medical Cost Personal 워크북

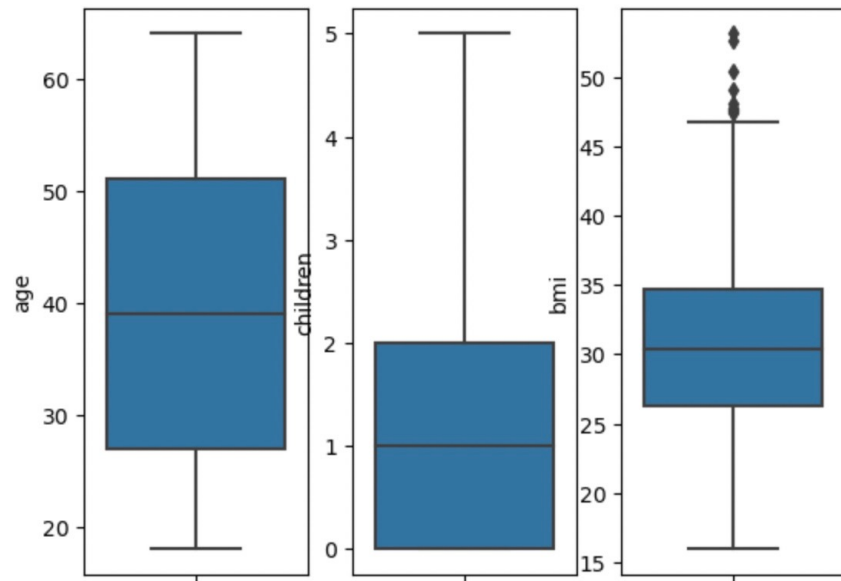
```
bmi_q1 = df['bmi'].quantile(q=0.25)
bmi_q3 = df['bmi'].quantile(q=0.75)
iqr = bmi_q3 - bmi_q1

condi1 = (df['bmi'] < (bmi_q1 - (1.5 * iqr)))
condi2 = (df['bmi'] > (bmi_q3 + (1.5 * iqr)))
outliers = df[condi1 | condi2]
outliers['bmi']
```

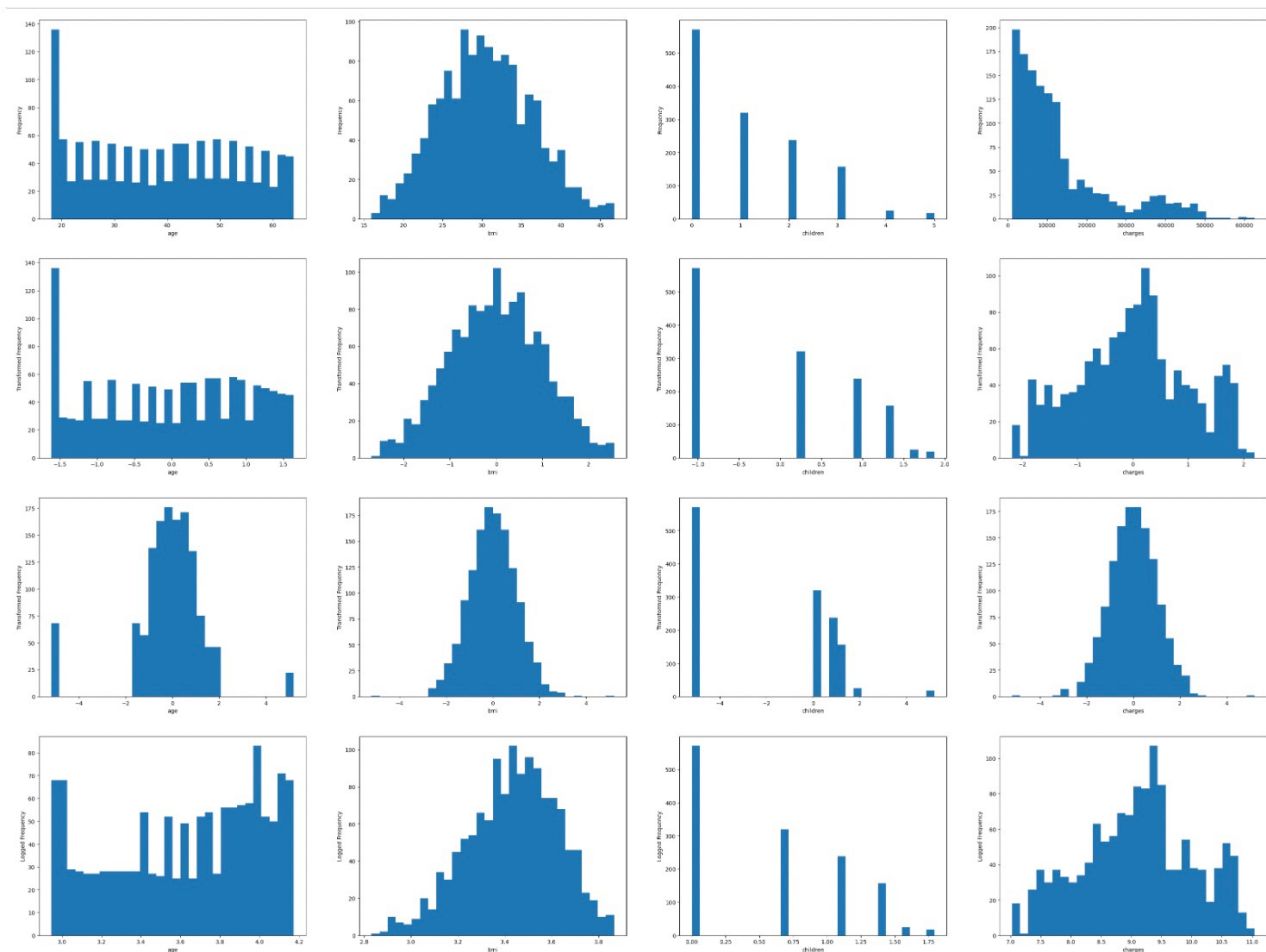
Q4. 이상치 탐지 코드와 시각화한 결과를 참고하여 IQR를 이용한 이상치 탐지가

어떤 로직으로 이루어지는지 조사해보세요.

Q4-1. 이상치 데이터 처리 방식에 대해 고민해보세요.



Medical Cost Personal 워크북



	age	sex	bmi	children	smoker	region_1	region_2	region_3
1180	0.159305	0	1.605816	-0.081877	0	0	0	0
489	0.588409	1	0.108793	-0.081877	0	1	0	0
619	0.674981	0	1.135173	-0.914017	0	0	0	1
301	0.588409	0	-1.378329	1.582402	1	0	0	0
1252	-0.688214	1	-0.563687	-0.914017	1	0	0	1

	age	bmi	children
count	9.960000e+02	9.960000e+02	9.960000e+02
mean	7.133963e-18	8.917454e-18	9.274152e-17
std	1.000502e+00	1.000502e+00	1.000502e+00
min	-3.167001e+00	-5.615665e+00	-9.140169e-01
25%	-3.299744e-01	-7.172856e-01	-9.140169e-01
50%	7.226309e-02	-1.632032e-02	-8.187720e-02
75%	5.034374e-01	6.890154e-01	7.502625e-01
max	3.311527e+00	5.622782e+00	3.246682e+00

질문은 다음 슬라이드에~~

Medical Cost Personal 워크북

Q5. 전 페이지에 있던 자료들은 아래의 과정을 모두 거친 데이터셋이다. <보기>를 참고하여 빈칸을 채워보자.

Q5-1. (A , B) 칼럼은 (C)형 칼럼이기 때문에 QuantileTransformer, PowerTransformer, log1p를 이용해 (D) 진행
이때 'children' 피쳐는 레이블이 적기 때문에 제외

Q5-2. 'age', 'bmi', 'children' 칼럼들의 단위와 분포를 맞추기 위해 (E)를 이용해 (F) 진행

Q5-3. 'sex', 'smoker' 칼럼은 (G)형 칼럼이기 때문에 (H) 진행

이때 'region' 칼럼은 yes / no 두개로 나뉘는 것이 아니기 때문에 (I)를 먼저 진행한 뒤 (H) 진행

<보기>

age, bmi, children, sex, smoker, charges, region

숫자, 카테고리, 표준화, 정규화, 시각화, 이상치 제거, 인코딩

StandardScaler, MinMaxScaler, log1p, QuantileTransformer, PowerTransformer, LabelEncoding, OneHotEncoding

Medical Cost Personal 워크북

회귀 모델 객체 생성

```
lr = LinearRegression()
enet = ElasticNet(random_state=42)
dt = DecisionTreeRegressor(random_state=42)
rf = RandomForestRegressor(random_state=42)
ada = AdaBoostRegressor(random_state=42)
gbr = GradientBoostingRegressor(random_state=42)
xgb = XGBRegressor(random_state=42)
lgbm = LGBMRegressor(random_state=42)

models = [lr, enet, dt, rf, ada, gbr, xgb, lgbm]
```

Q6. LinearRegression과 ElasticNet은 함수 기반 회귀이며

나머지는 트리 기반 회귀이다

Q6-1. LinearRegression과 ElasticNet은 회귀 함수식을 추론해내는 데에 목적을 두고 있다. 따라서 함수식에서 각 피처(= 독립변수)의 ()를 적절하게 설정해야 한다.

Q6-2. ElasticNet은 규제 선형 회귀이다. 이때 규제를 조절하는 파라미터가 무엇이며 ElasticNet의 규제 방식은 무엇인가?

Q6-3. 나머지 트리 기반 회귀는 ()을 통해 성능을 높일 수 있다.

-> 노트북을 참고하여 트리 기반 회귀에 무엇을 했는지 살펴보세요!

Medical Cost Personal 워크북

회귀 평가

```
for model in models:
    name = model.__class__.__name__
    scores = cross_val_score(model, X=X_train, y=y_train, cv=5, scoring='neg_mean_squared_error', n_jobs=-1)
    mse = (-1)*np.mean(scores)
    print('Model %s - RMSE: %.4f' % (name, np.sqrt(mse)))
```

```
Model LinearRegression - RMSE: 6428.2026
Model ElasticNet - RMSE: 9690.3391
Model DecisionTreeRegressor - RMSE: 6786.6027
Model RandomForestRegressor - RMSE: 5052.4484
Model AdaBoostRegressor - RMSE: 5236.0463
Model GradientBoostingRegressor - RMSE: 4849.8268
Model XGBRegressor - RMSE: 5441.5261
Model LGBMRegressor - RMSE: 5046.7237
```

다음은 각각의 회귀 모델에 RMSE 평가를 진행하는 코드와 결과이다

Q7. RMSE 평가의 식을 참고하여 회귀 모델에 대한 평가를 어떤 로직으로 진행하는지 생각해보자.

Q7-1. RMSE의 값이 (클수록 / 작을수록) 성능이 높다.

Q7-2. mse에 -1을 곱하는 이유가 무엇인가?



수고하셨습니다