Personal Medical Cost Prediction.

개인 의료비 예측 프로젝트

HI:FIVE

서영석 오수민 정민경 최수빈 홍수정

Index.

Part 1.

Introduction

Part 1 개요

News

한국 의료비 증가 속도, OECD 중 가장 빨라

노인단독가구 부담 2배 증가 … "비급여의 급여화, 저소득층 본인부담 완화, 일차의료 강화" 제시

2022-07-26 10:55:44 게재

우리 국민이 부담하는 의료비 증가 속도가 OECD 중 가장 빠른 가운데 특히 노인단독가구의 부담은 2배로 증가해 대책마련이 시급하다는 지적이 나왔다. 비급여의 급여화, 저소득층의 본인부담 완화, 일차의료 강화 필요성이 제기됐다.

[이슈 In] 급증하는 의료비 지출...'행위별수가제'가 문제인가

송고시간 | 2022-08-05 06:03

특히 92.4%는 현재 우리나라가 채택한 행위별수가제는 불필요한 의료서비스를 유발하는 등 의료이용량 증가 유인의 단점이 있으니 진료비 지불제도를 개선할 필요가 있다고 답했다.

개인 의료비 예측을 통한
과잉진료,
무분별한 의료비 증가 방지

사회적·신체적 개인 정보를 기반으로 하는 **의료비 예측**을 통해,

개인은 본인의 의료비를 직접 예측하여 과납을 막고 그에 상응하는 **보험금을 납부**한다.

또한 **보험회사**는 보험료 변화 추세를 파악하여, 그에 맞는 **상품을 기획**한다.

Kaggle의
Open Dataset을 활용한다.

나이, 성별, 비만도, 출산 경험 수, 거주지, 흡연 여부를 고려한 **개인별 의료비 예측 모델**을 구축하여 평가한다.

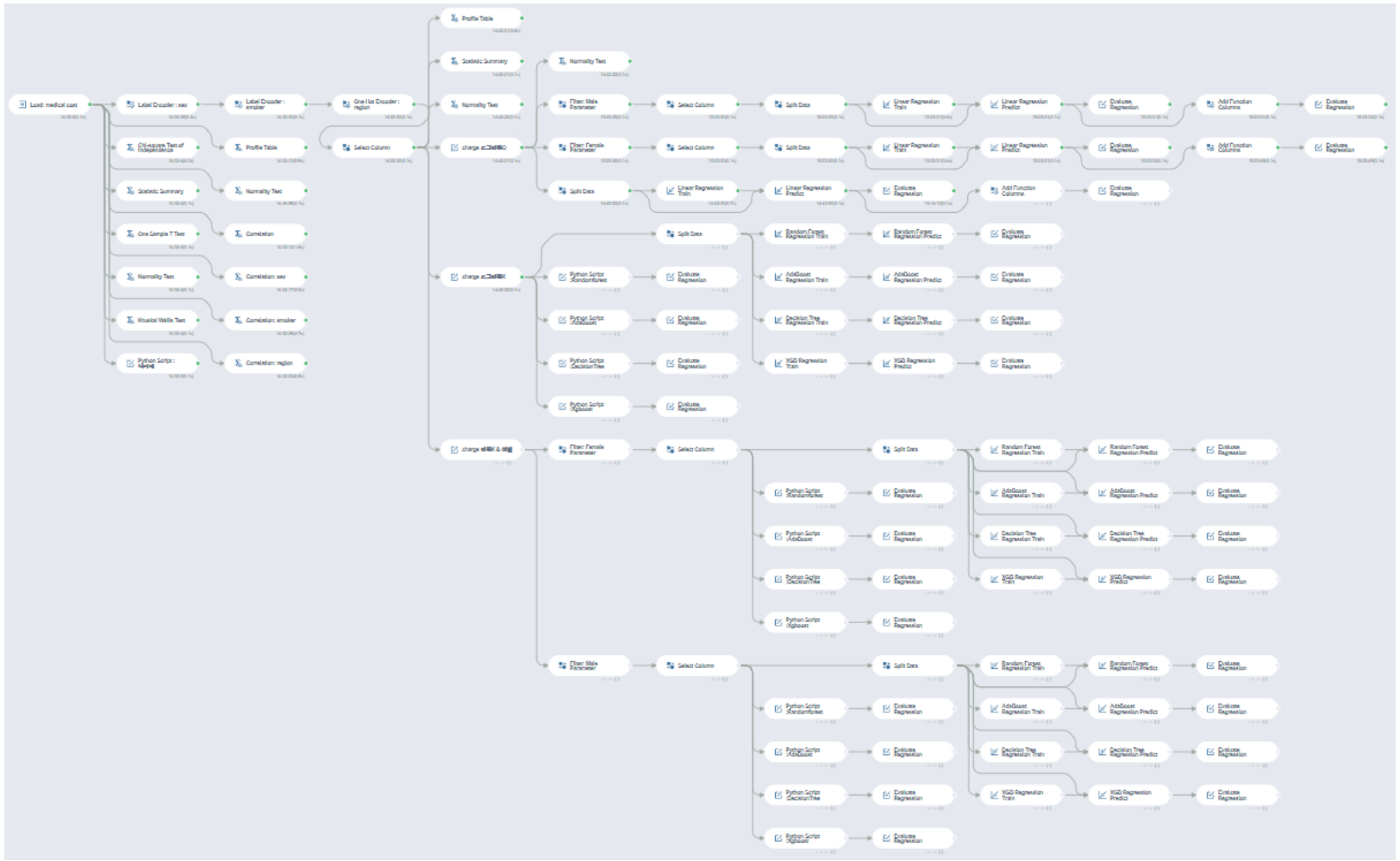
Brightics Studio의 다양한 회귀 모델을 활용하여 개인별 의료비를 예측한다.

전처리, EDA와 다양한 모델을 활용하는 모습을 **직관적으로 보여줄 수 있는 데이터**이다.

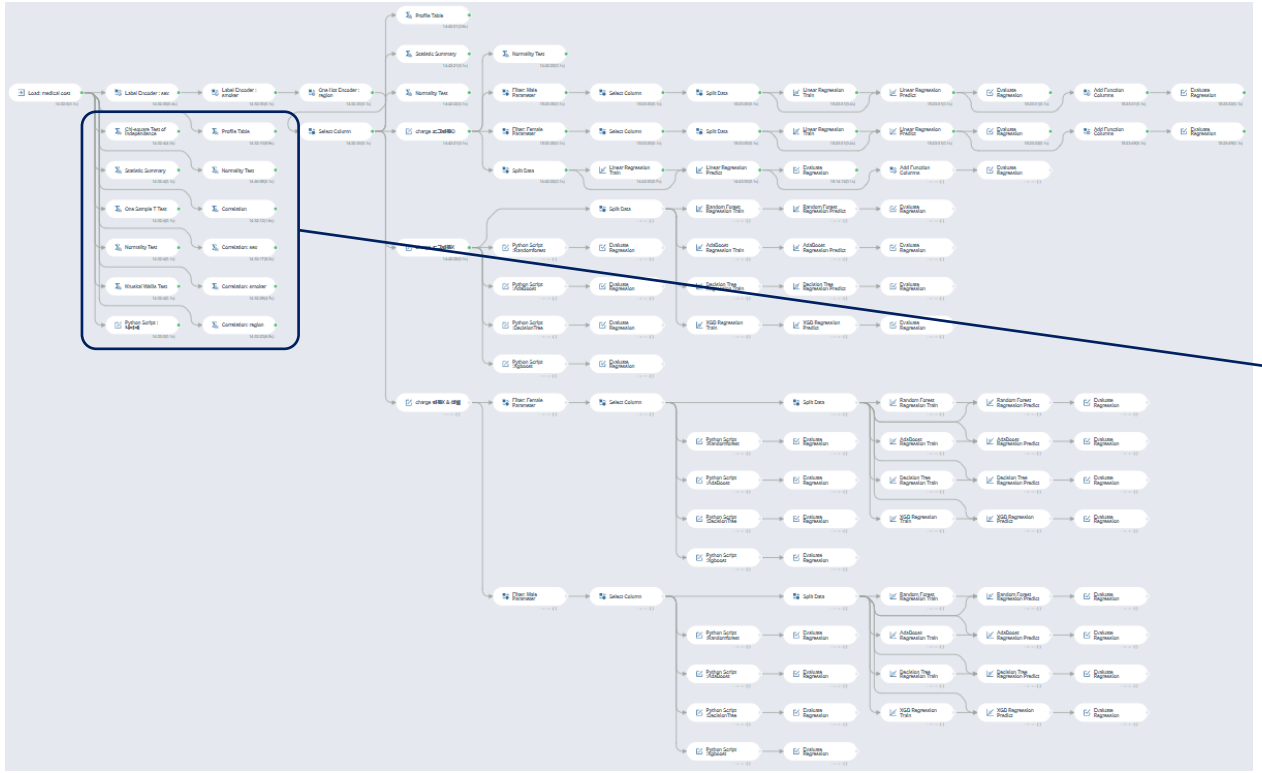
Column	About Column
age	보험 계약자의 **나이**
sex	보험 계약자의 **성별**
bmi	몸무게를 키의 제곱으로 나눈 **체질량 지수**
children	건강보험 적용 **자녀 수/부양 가족 수**
smoker	**흡연 여부**
region	미국에서 보험 계약자의 **거주지** (northeast, southeast, southwest, northwest)
charges	의료 보험에서 고지 받은 **의료 비용**

Part 1 개요

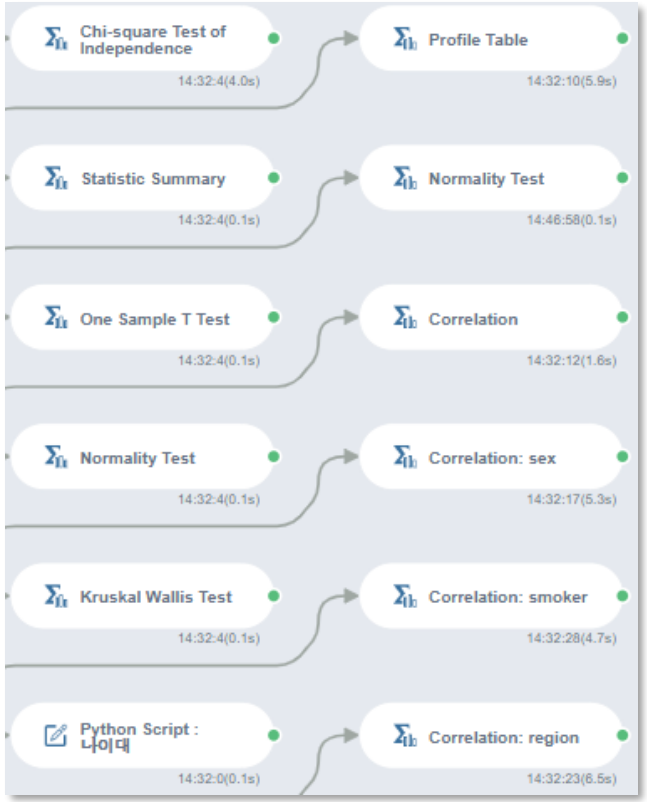
Brightics 흐름도



Brightics 흐름도

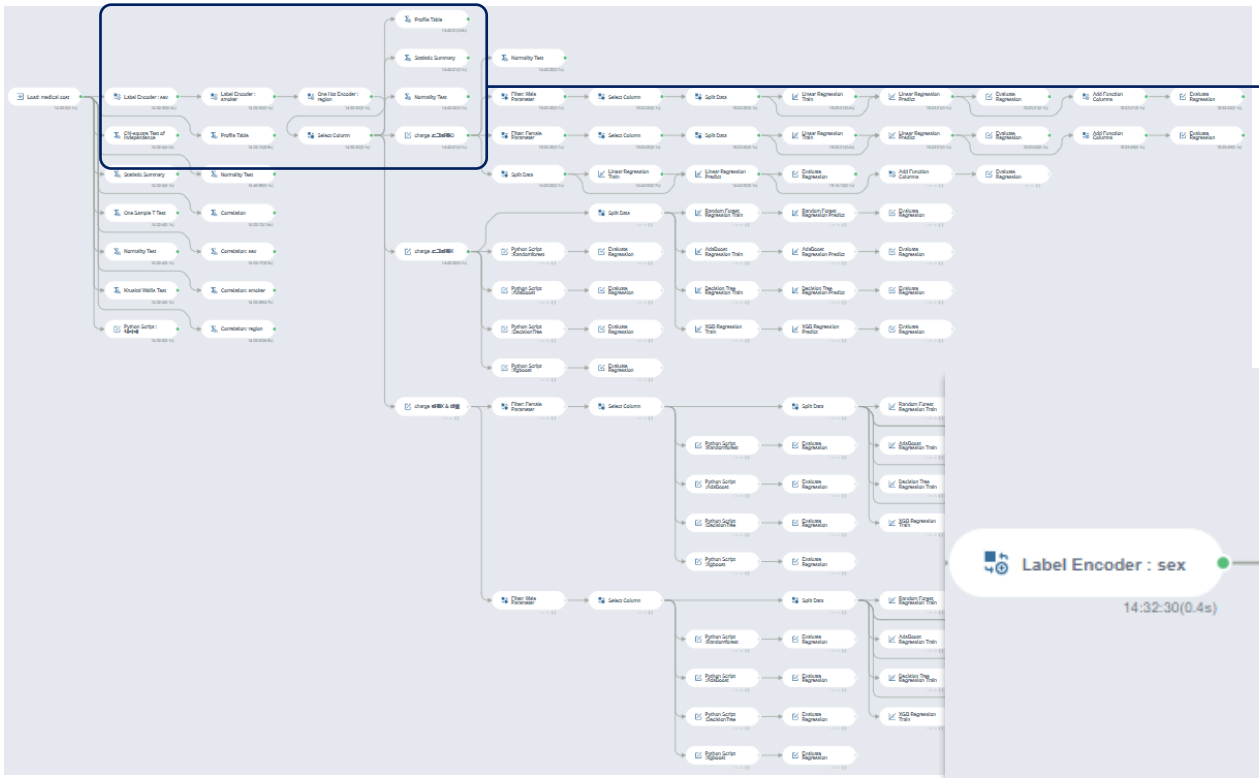


EDA

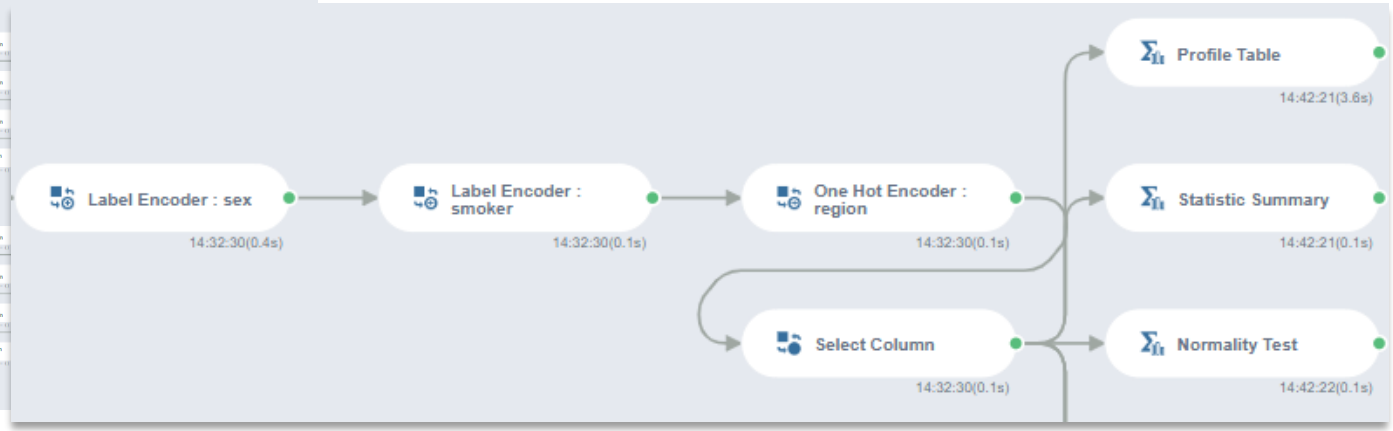


Part 1 개요

Brightics 흐름도

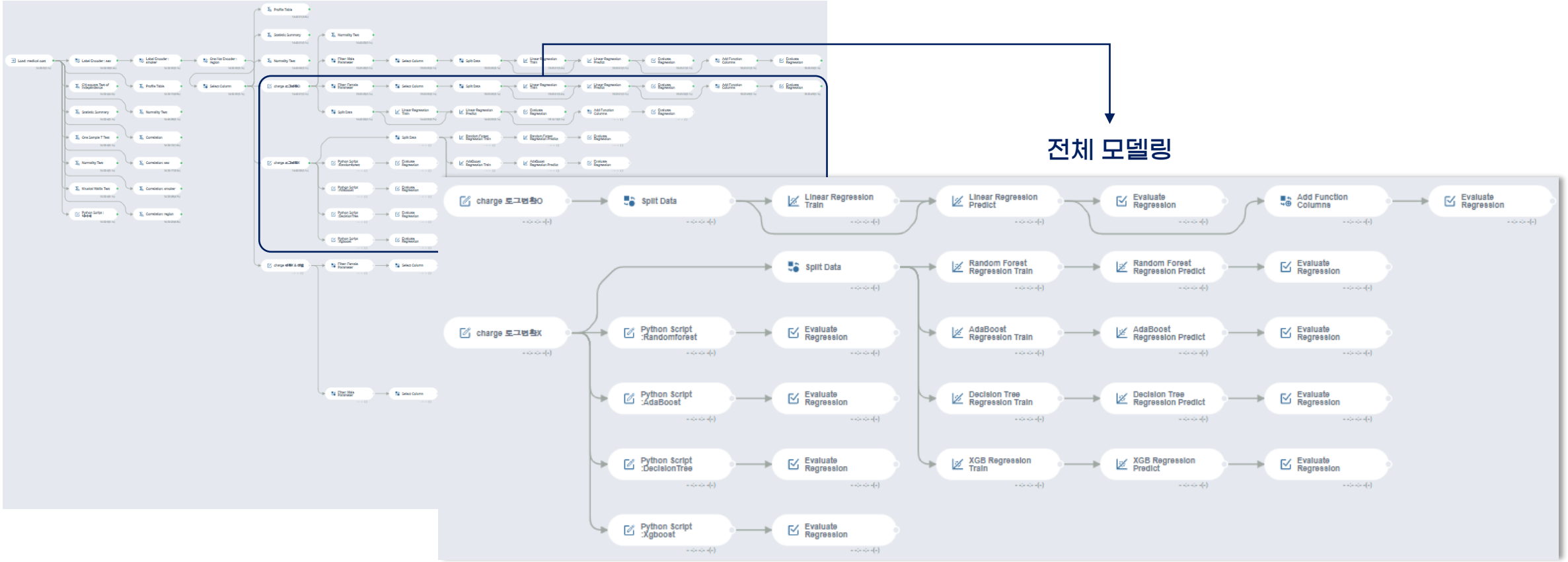


데이터 전처리



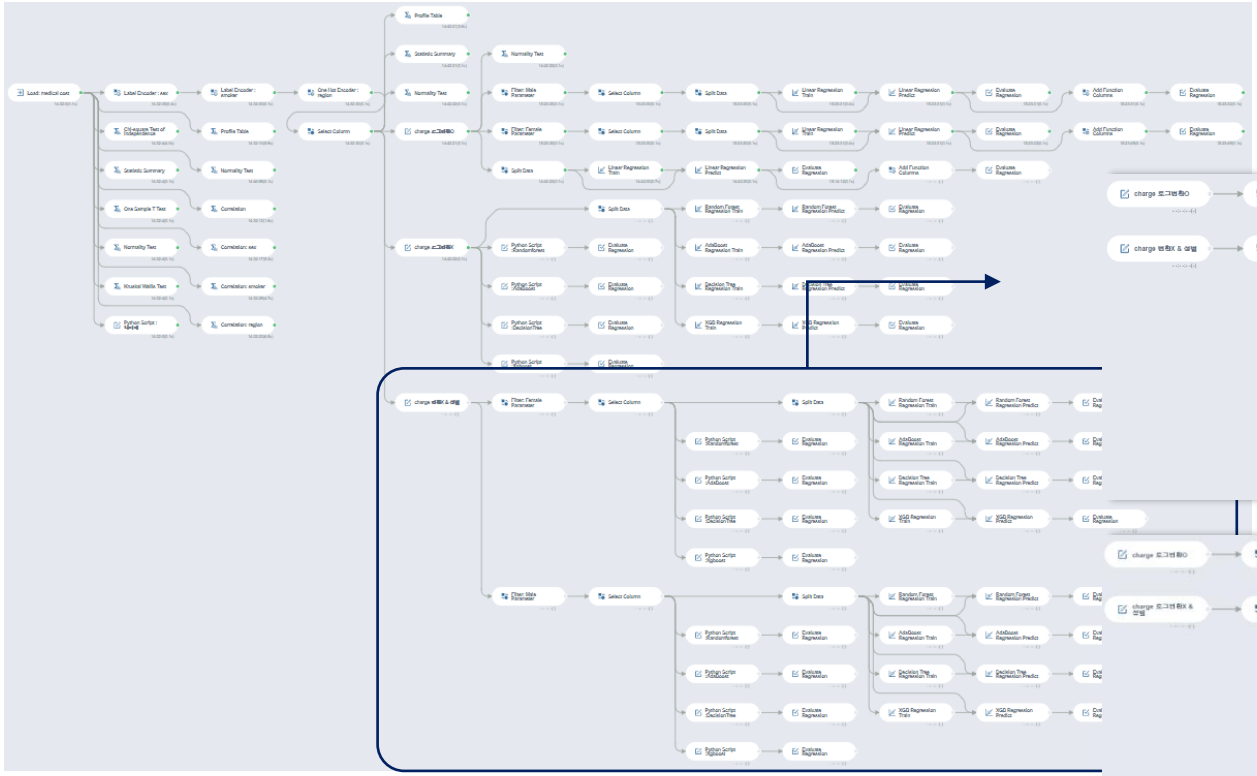
Part 1 개요

Brightics 흐름도

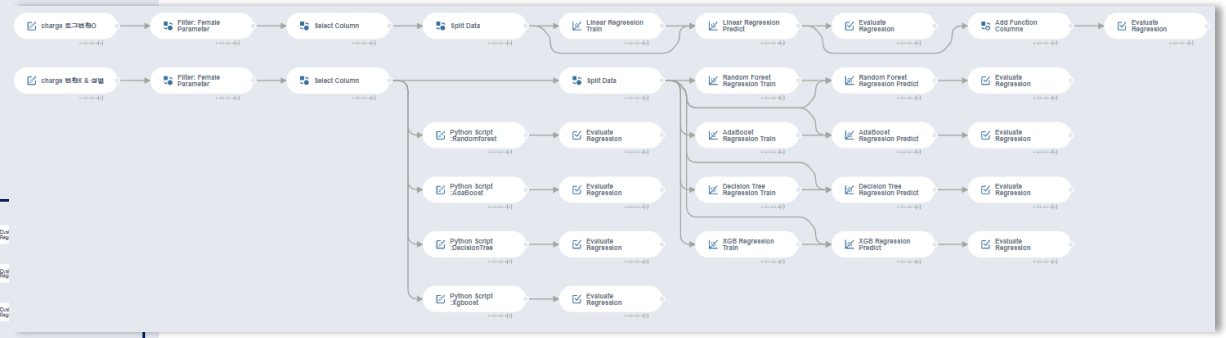


Part 1 개요

Brightics 흐름도



성별(남/여) 모델링



Part 2.

EDA

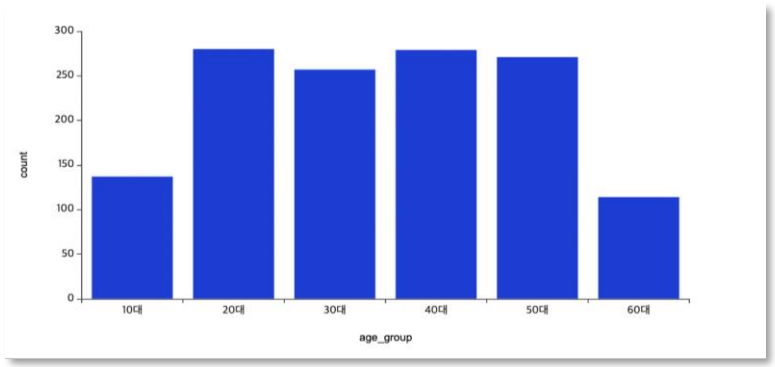
01 **Statistic Summary**

02 **Correlation**

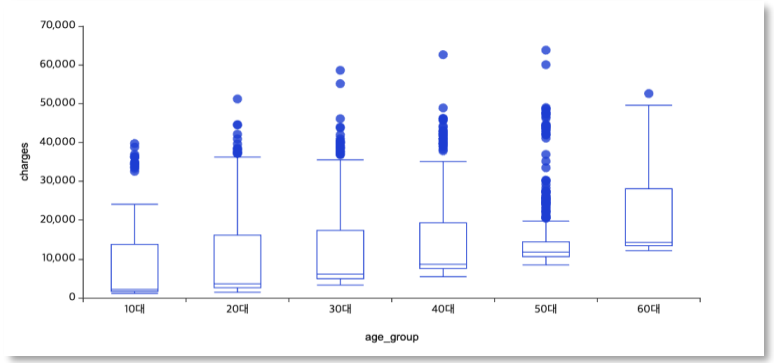
03 **Statistical Analysis**

Part 2 데이터 탐색

나이(age)

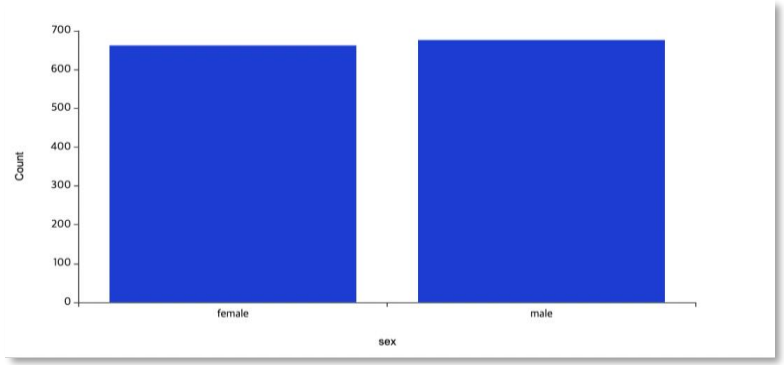


- 10대와 60대의 표본이 적다

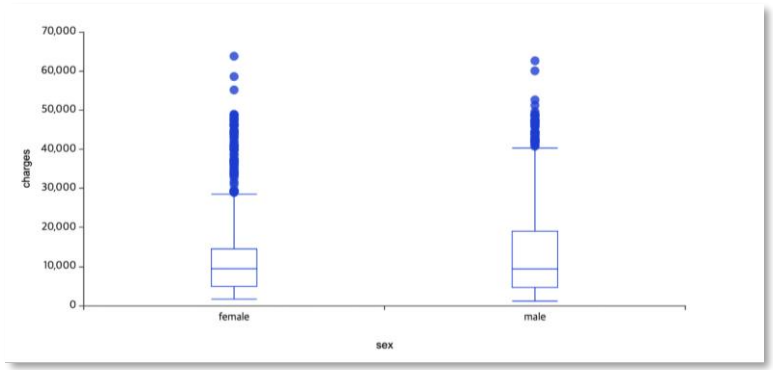


- 나이가 증가할 수록 의료비가 높아진다.
- 50대의 의료비 분포가 다른 나이대와 다르다.

성별(sex)



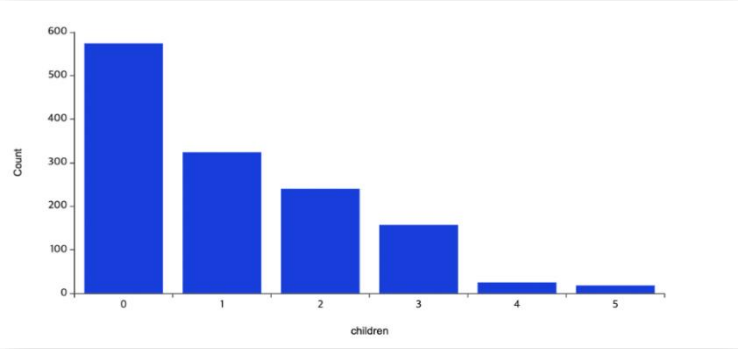
- 남/여의 표본 수 차이는 거의 없다.



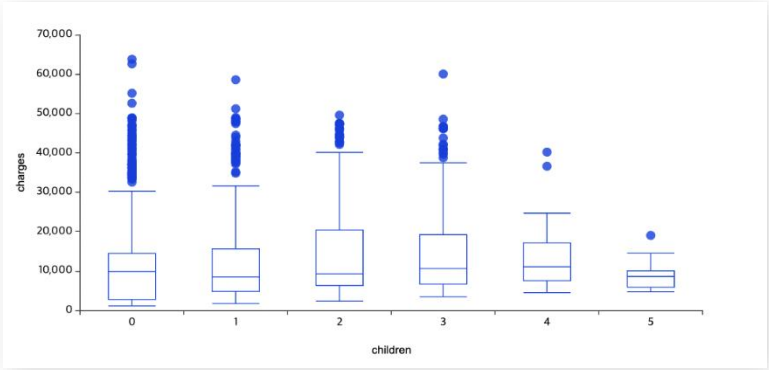
- 남성에 비해 여성에게 이상치 값이 더 많이 존재한다.
  (특이질환의 유무가 다른 것인지 추측)

Part 2 데이터 탐색

자녀수(children)

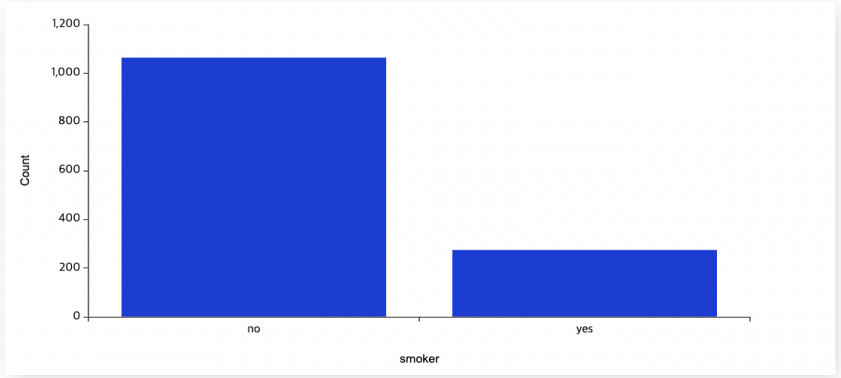


- 자녀수가 많을수록 표본수가 적어진다.

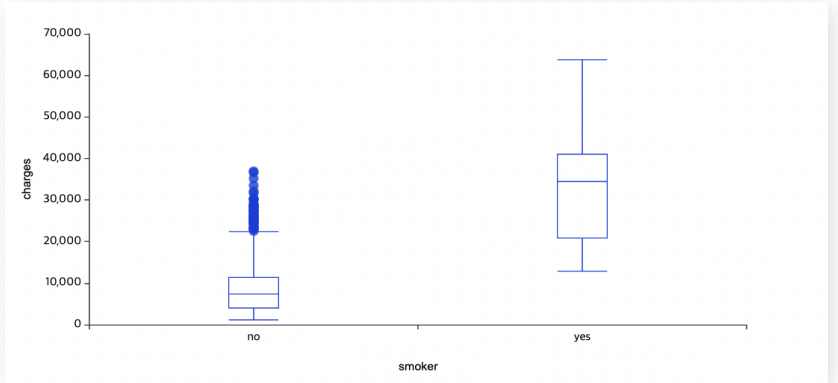


- **자녀수가 많다고 의료비가 높아지지는 않는다.**

흡연여부(smoker)



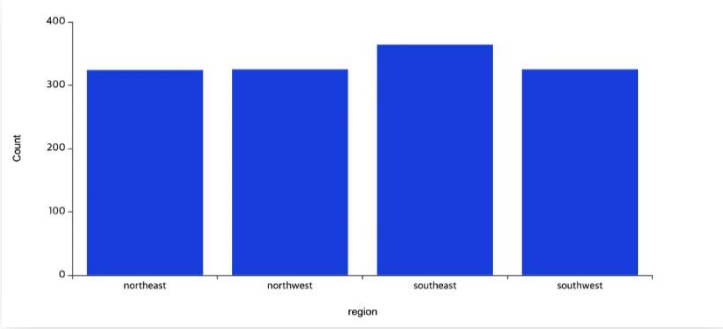
- 흡연자가 비흡연자에 비해 적다.



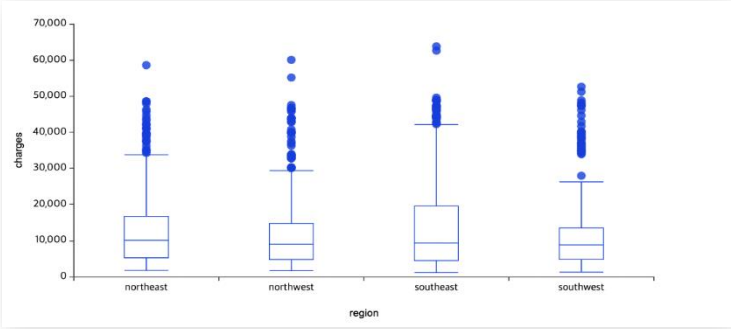
- **비흡연자보다 흡연자의 의료비가 높다.**

Part 2 데이터 탐색

지역 (region)

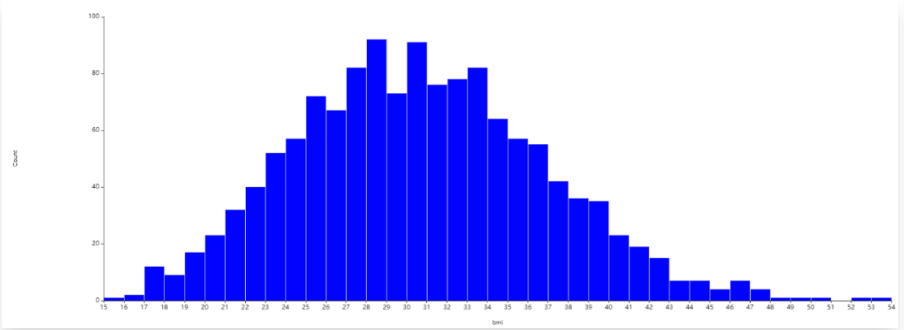


- 지역마다의 표본 수 차이 거의 없다.

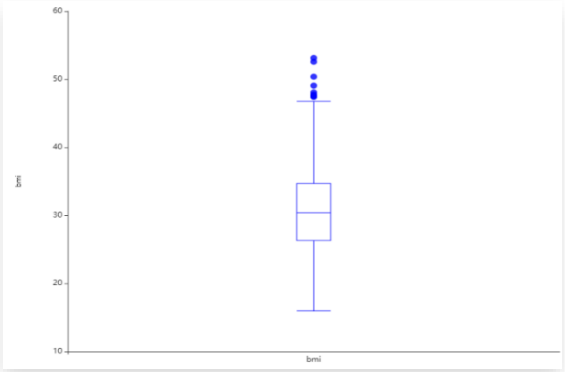


- 지역마다 의료비의 차이는 크지 않다.

비만도(bmi)



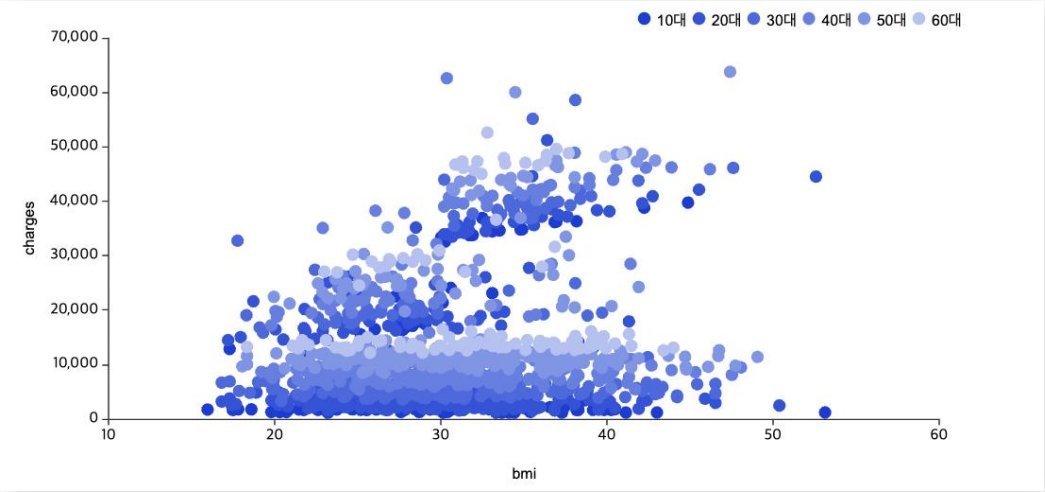
- 비만도가 높을수록 의료비가 높아진다.



- 15에서 54까지 분포하며, 평균은 30이다.

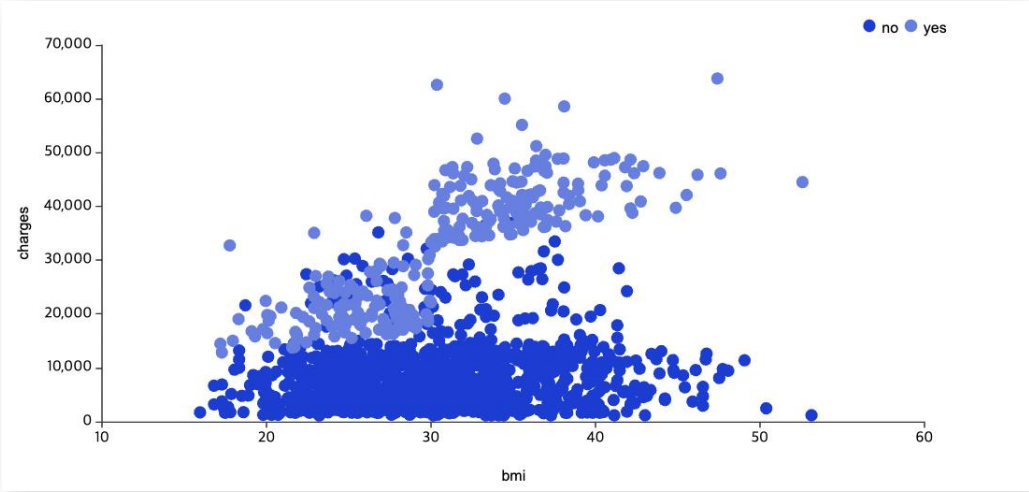
Part 2 데이터 탐색

나이(age) & 비만도(bmi)



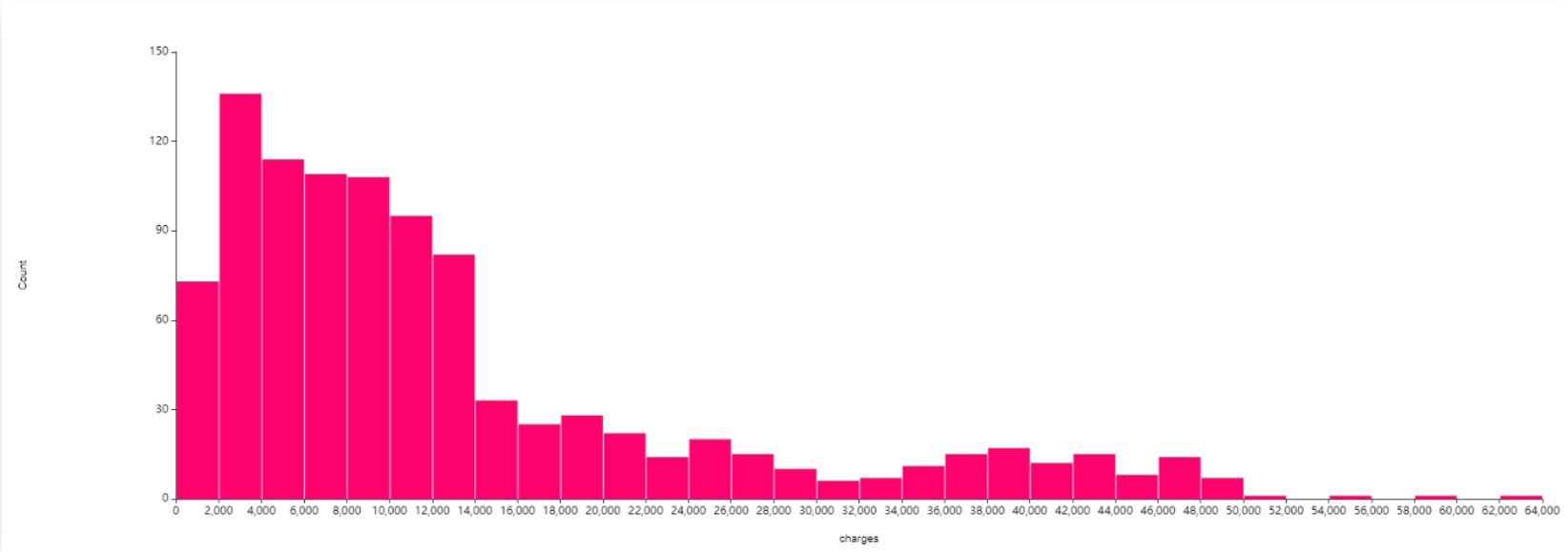
- 같은 비만도(bmi) 수치여도
  나이대에 따라 의료비가 달라진다.

흡연여부(smoker) & 비만도(bmi)

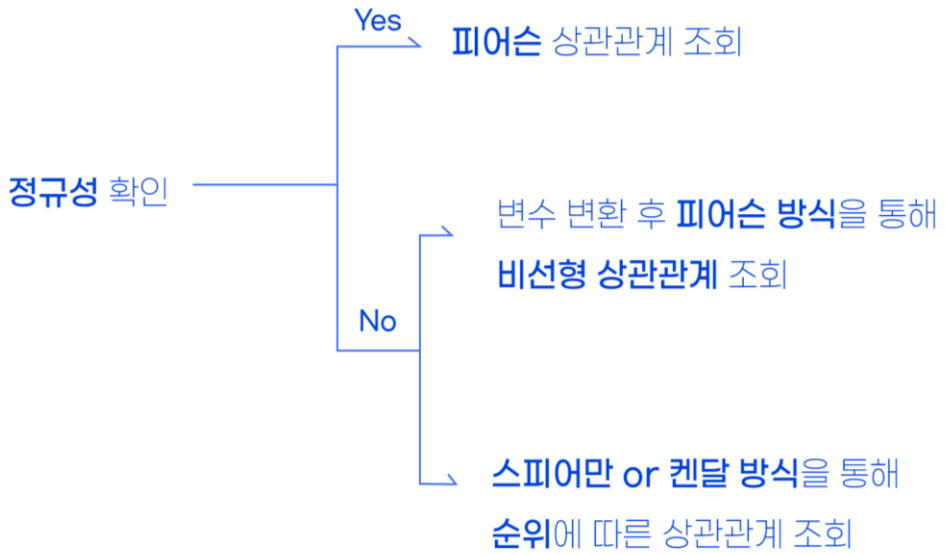


- 흡연자일 때, 비만도와 의료비는
  양의 상관관계를 가진다.

의료비(charges)



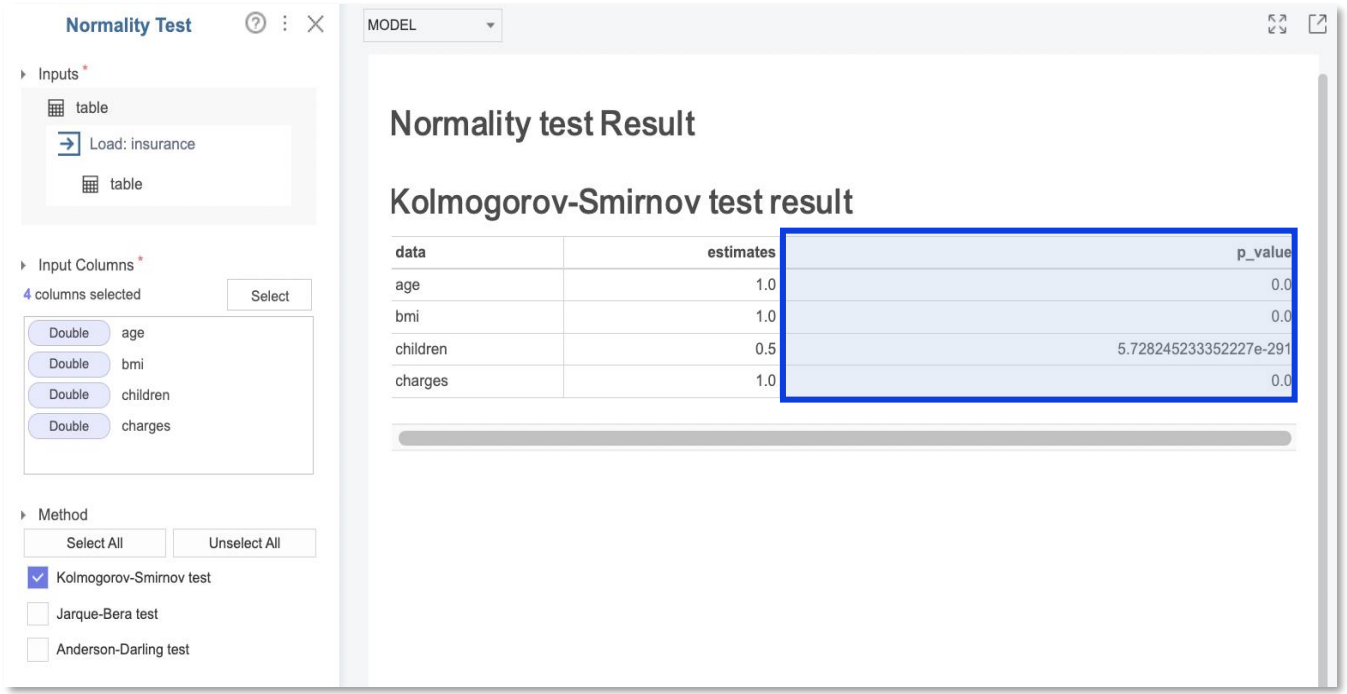
- $0에서 $64,000까지의 분포를 가지며, 왼쪽으로 치우친 경향이 있다.



Process 및 개념

정규성을 만족시키면 **피어슨 방식**,
비정규성을 띄면 **스피어만 방식**을 사용한다.

정규성 검정



Kolmogorov-Smirnov test 결과

p_value < 0.05

∴ 정규성을 만족하지 않는다.

스피어만(Spearman) 방식 채택

Part 2 데이터 탐색

숫자형 변수 전체(All)

Correlation 결과

- 의료비(charges) & 나이(age) 상관관계가 가장 크다.
- 자녀수(children) & 비만도(bmi)는 독립적이다.
- 나머지 변수들 간, 어느정도 상관관계가 존재하고 있다.

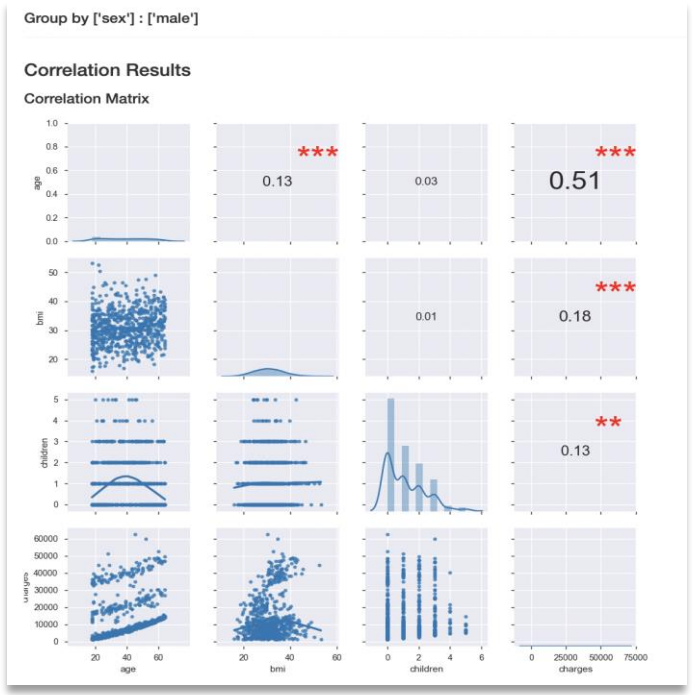
* 상관계수 0.05 이상 → 유의미

Part 2 데이터 탐색

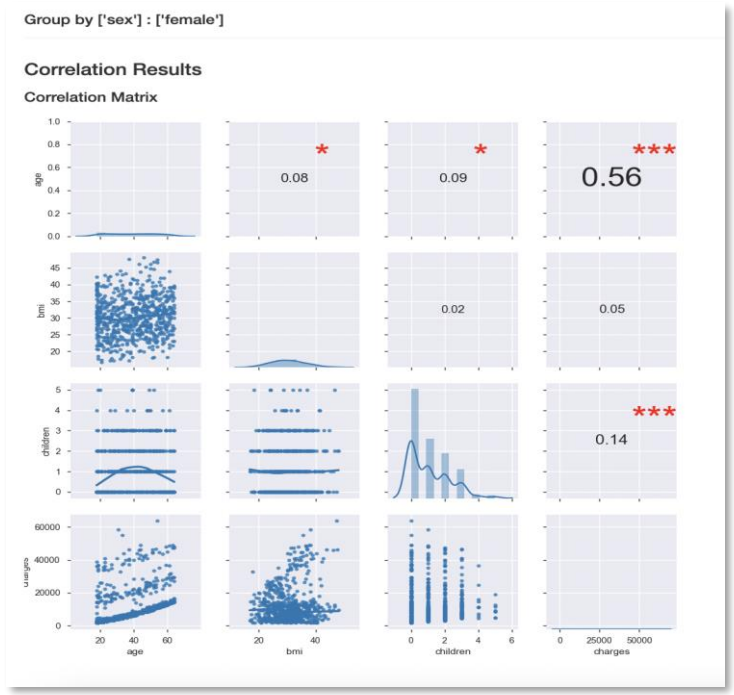
Statistic Summary | **Correlation** | Statistic Analysis

성별(sex)에 따른 상관관계

남성



여성



Correlation 결과

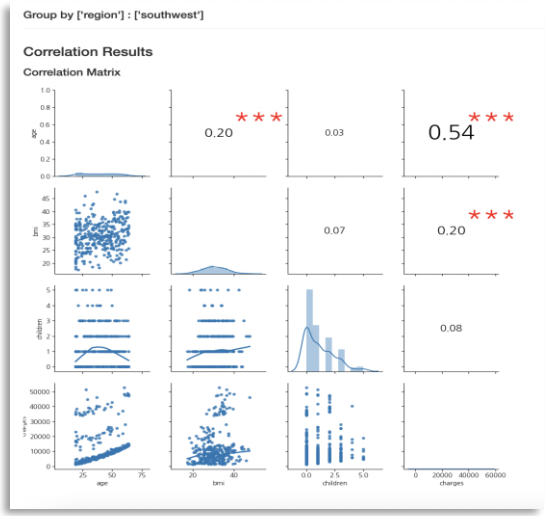
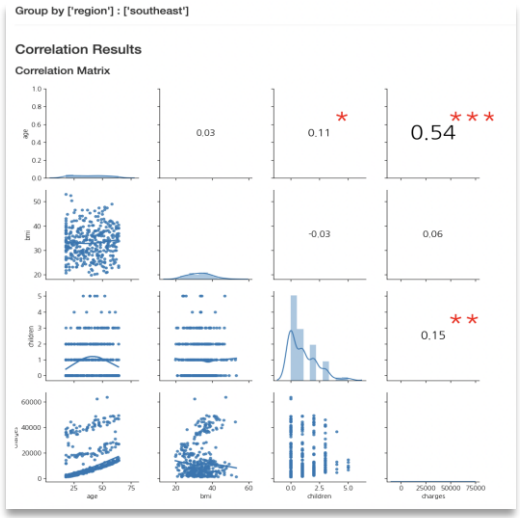
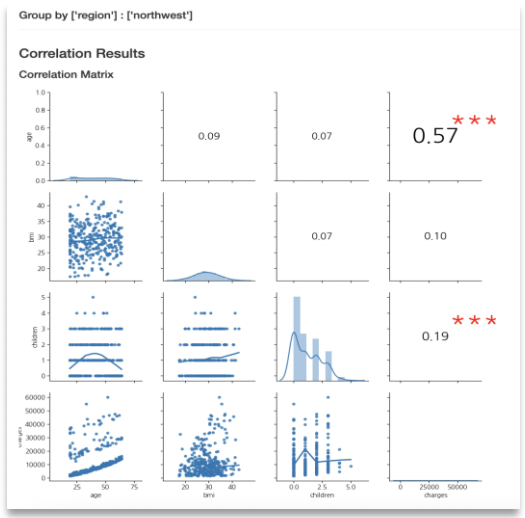
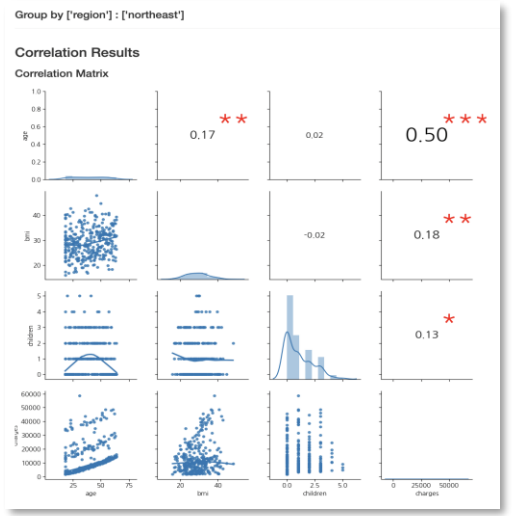
- 성별에 따라 나누었을 때,

  전체 데이터와 큰 차이가 없다.

Part 2 데이터 탐색

Statistic Summary | **Correlation** | Statistic Analysis

지역(region)에 따른 상관관계

southwest	southeast	northwest	northeast
			

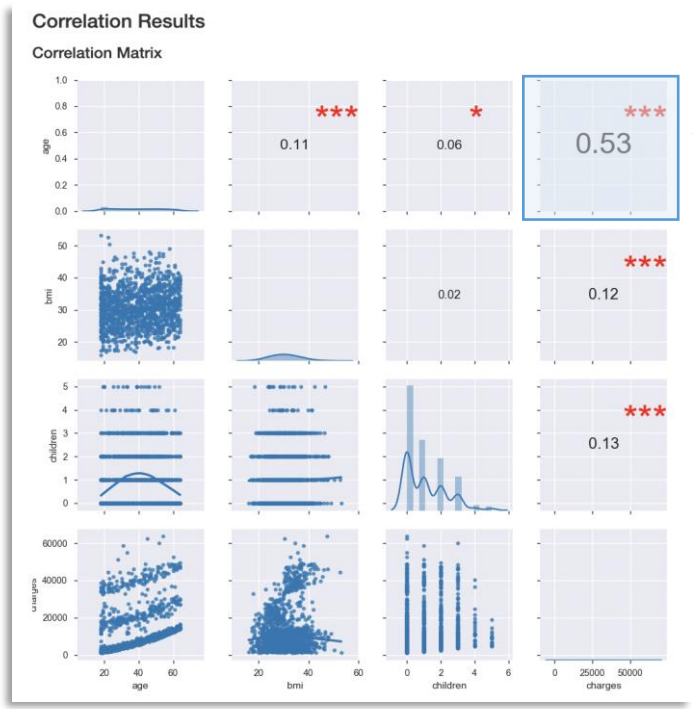
Correlation 결과

- 지역에 따라 나누었을 때, 전체 데이터와 큰 차이가 없다.

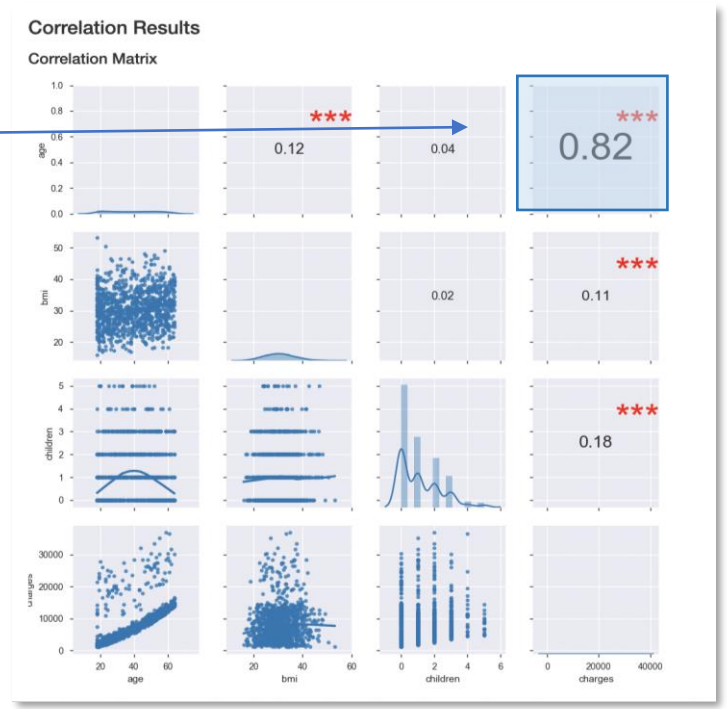
Part 2 데이터 탐색

흡연여부(smoke)에 따른 상관관계

전체



비흡연자



Correlation 결과

비흡연자의

나이(age) & 의료비(charge)

상관계수가 약 1.5배 증가

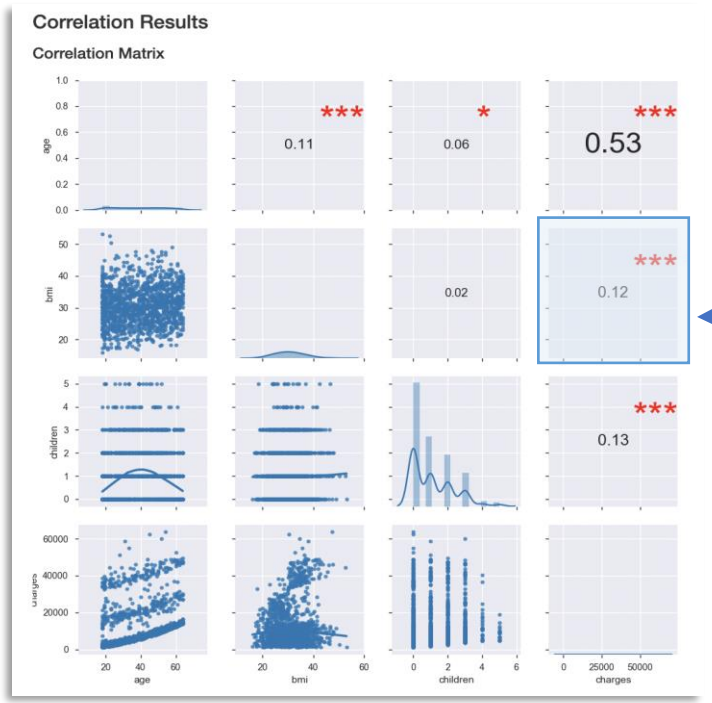
전체 : 0.53
비흡연자 : 0.82

Part 2 데이터 탐색

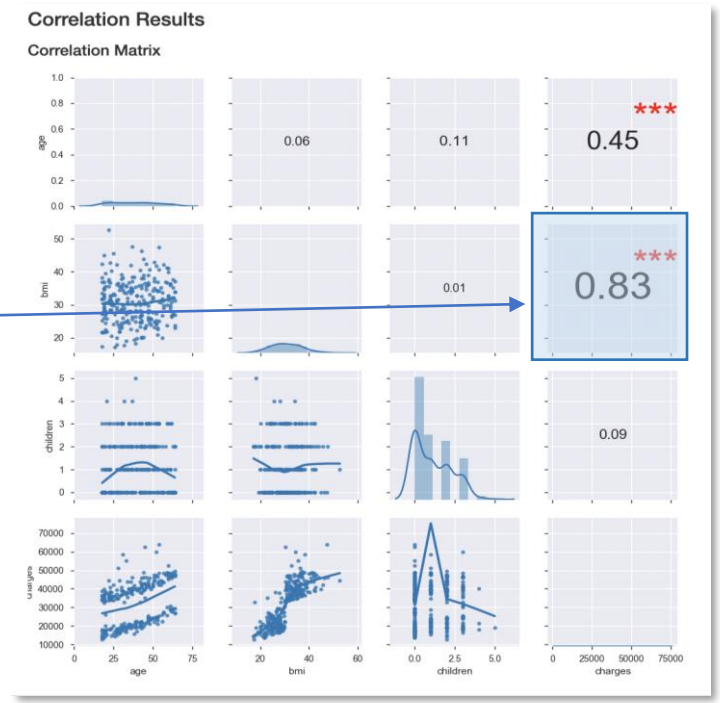
Statistic Summary | **Correlation** | Statistic Analysis

흡연여부(smoke)에 따른 상관관계

전체



흡연자



Correlation 결과

흡연자의

비만도(bmi) & 의료비(charge)

상관계수가 약 8배 증가

전체 : 0.12
흡연자 : 0.83

Process 및 개념

두 범주형 변수 관계 파악

상관성 검정

→ **Chi Square Test of Independence**

귀무가설(H0) : 범주형 변수가 서로 독립적이다.

대립가설(H1) : 범주형 변수가 서로 의존적이다.

정규성 검정

귀무가설(H0) : 표본 분포는 정규분포를 따른다.

대립가설(H1) : 표본 분포는 정규분포를 따르지 않는다.

Yes → **ANOVA**

No → **Kruskal-Wallis Test**

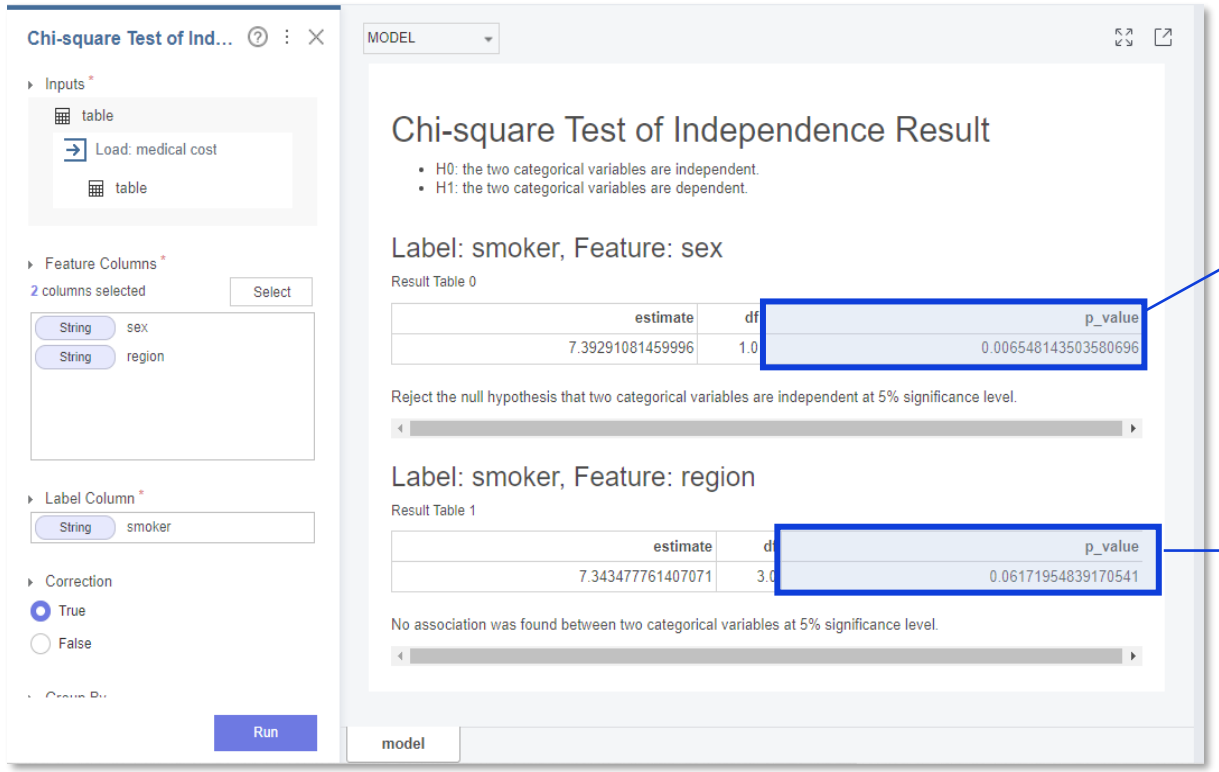
귀무가설(H0) : 분포가 모든 그룹이 같다.

대립가설(H1) : 분포가 다른 그룹이 적어도 하나 이상 존재한다.

Part 2 데이터 탐색

Chi Square Test of Independence

흡연여부(smoke) 기준



Test 결과

성별(sex) & 흡연여부(smoke)

∴ 두 변수 의존적 (서로 영향을 준다)

▎p_value <0.05 ▶ H1 채택

지역(region) & 흡연여부(smoke)

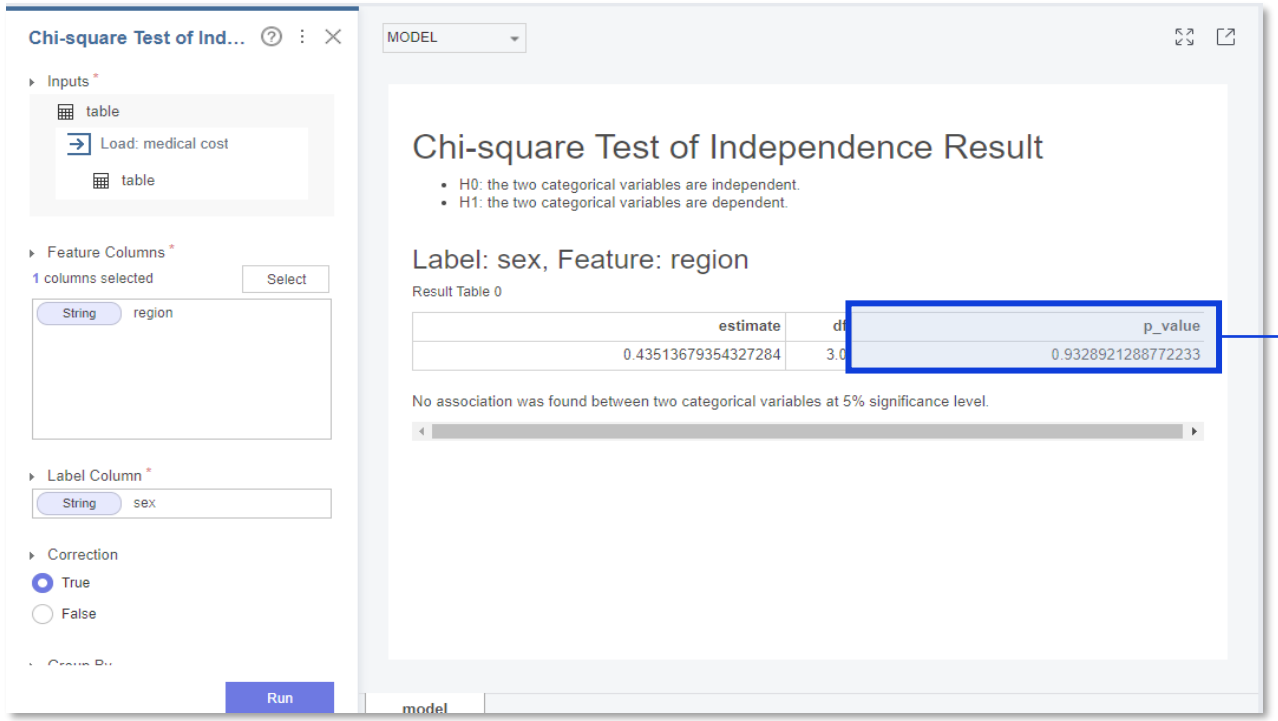
∴ 유의미한 연관이 있다고 보기에는 0.05를 충족하지 못한다.

▎p_value >0.05 ▶ H0 채택

Part 2 데이터 탐색

Chi Square Test of Independence

성별(sex) 기준



Test 결과

성별(sex) & 지역(region)

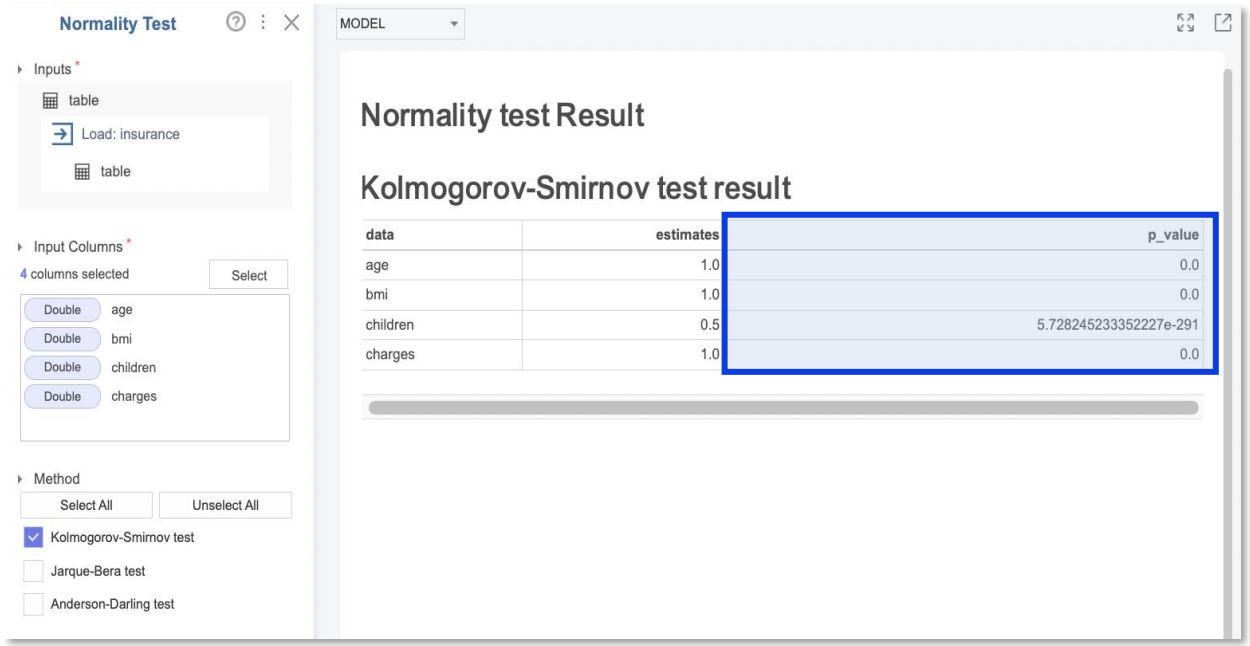
∴ 두 변수 독립적 (영향이 거의 없다)

p_value >0.05 ▶ H0 채택

Part 2 데이터 탐색

Kruskal-Wallis Test

정규성 검정



유의수준으로 H0, H1 가설 채택

Test 결과

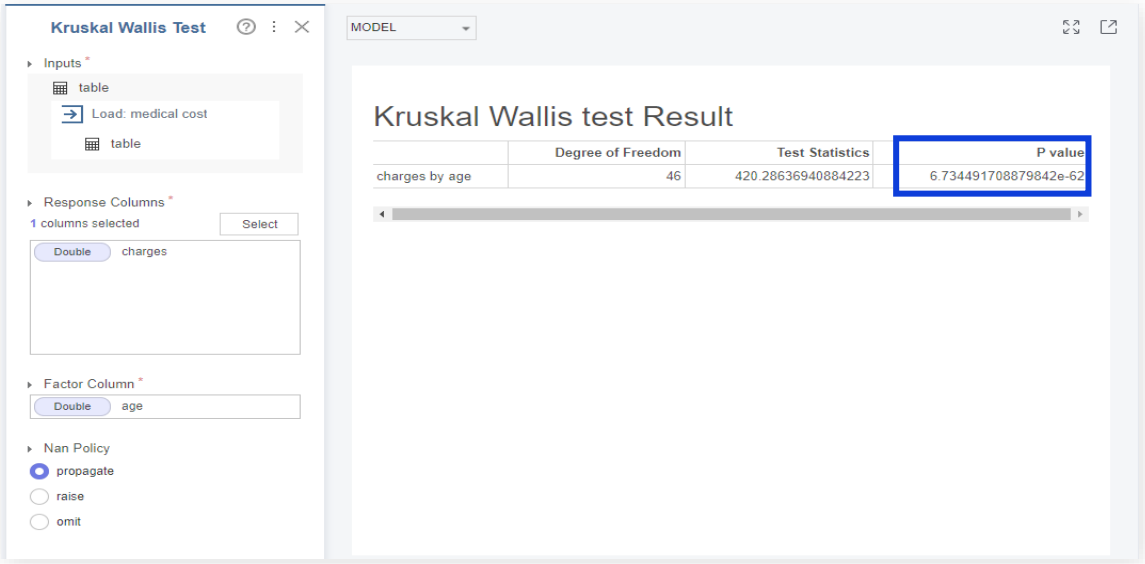
모두 표본분포가 정규분포를 따르지 않는다. (표본분포 ≠ 정규분포)

p_value <0.05 ▶ H1 채택

Part 2 데이터 탐색

Kruskal-Wallis Test

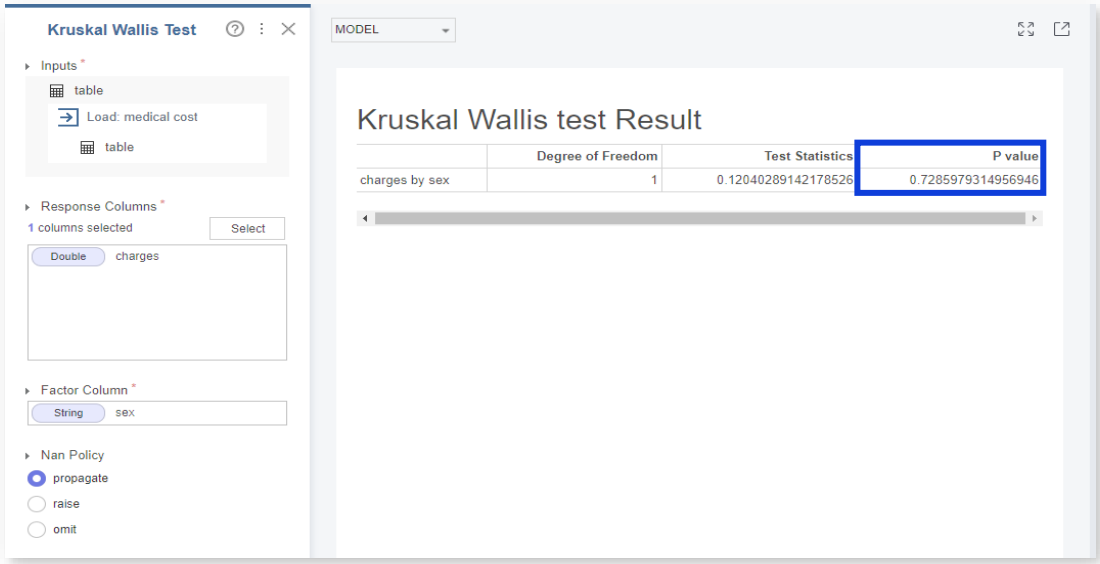
유의수준으로 H0, H1 가설 채택



의료비(Charge) & 나이(age)

분포가 다른 그룹이 적어도 하나 이상 존재한다.

p_value <0.05 ▶ H1 채택



의료비(Charge) & 성별(sex)

두 그룹의 분포는 같다.

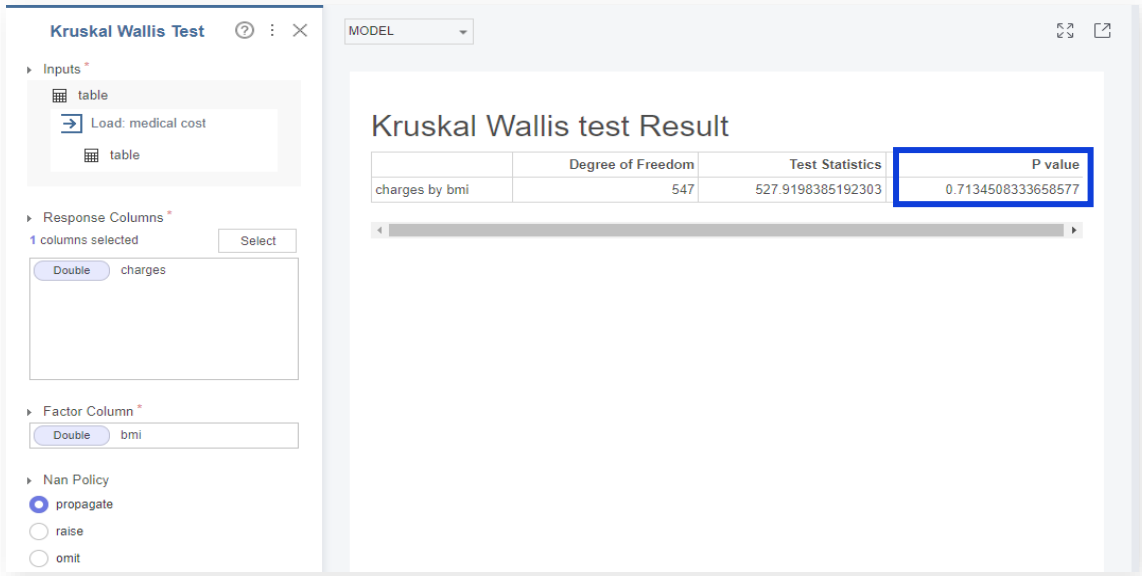
p_value >0.05 ▶ H0 채택

Part 2 데이터 탐색

Statistic Summary | Correlation | **Statistic Analysis**

Kruskal-Wallis Test

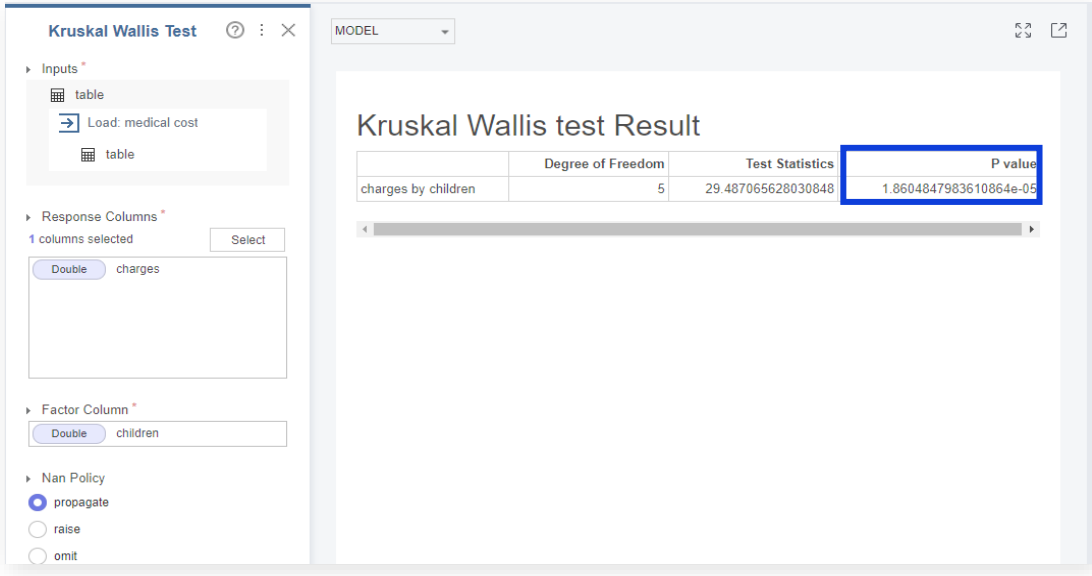
유의수준으로 H0, H1 가설 채택



의료비(Charge) & 비만도(bmi)

두 그룹의 분포는 같다

p_value >0.05 ▶ H0 채택



의료비(Charge) & 자녀수(children)

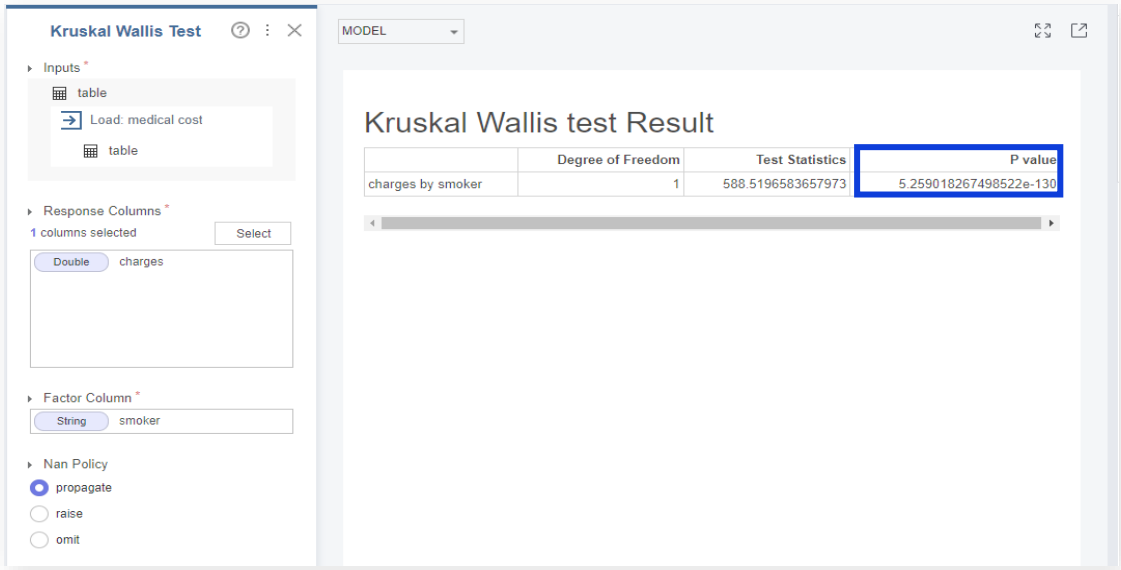
분포가 다른 그룹이 적어도 하나 이상 존재한다.

p_value <0.05 ▶ H1 채택

Part 2 데이터 탐색

Kruskal-Wallis Test

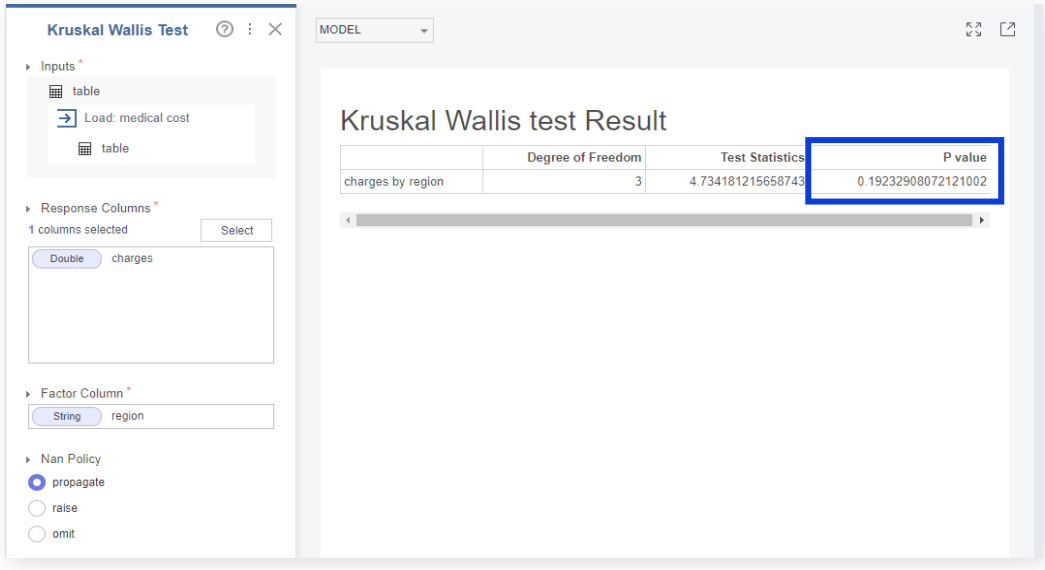
유의수준으로 H0, H1 가설 채택



의료비(Charge) & 흡연여부(smoke)

분포가 다른 그룹이 적어도 하나 이상 존재한다.

p_value <0.05 ▶ H1 채택



의료비(Charge) & 지역(region)

두 그룹의 분포는 같다.

p_value >0.05 ▶ H0 채택

Part 3.

Preprocessing

01 결측치 · 중복행 처리

02 Outlier Detection

03 Label Encoder

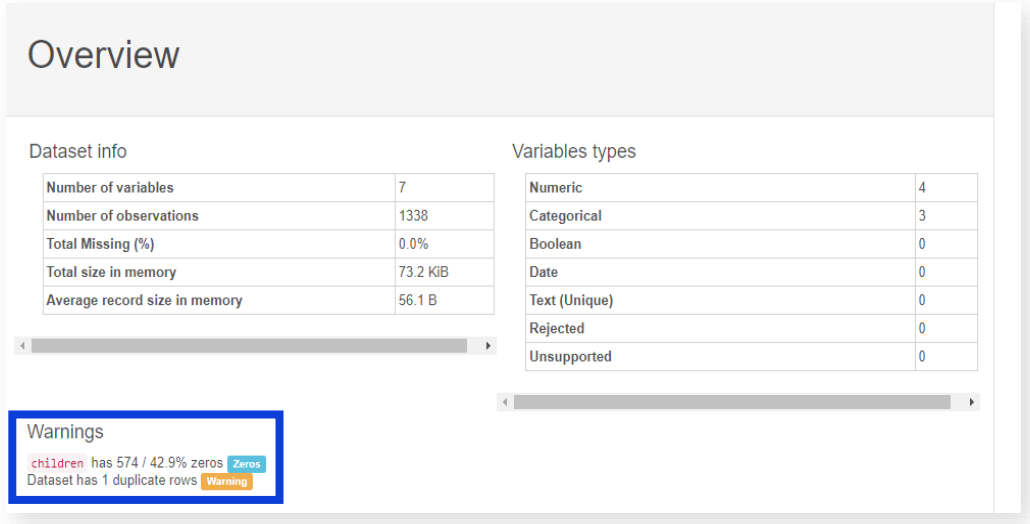
04 One Hot Encoder

05 로그변환 · 정규성 검정

Part 3 데이터 전처리

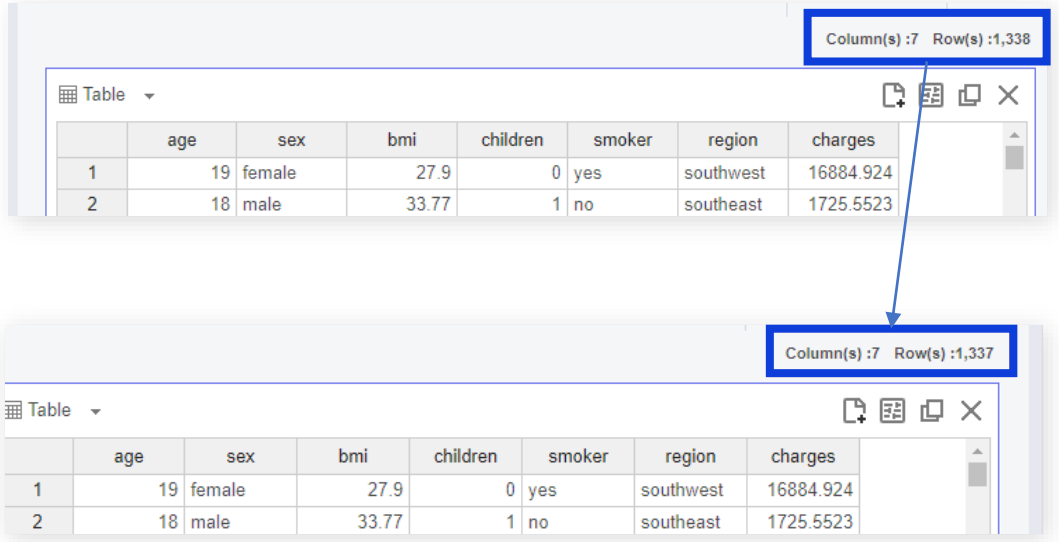
결측치 · 중복행 처리 | Outlier Detection | Label Encoder

Profile Table



- Profile Table ▶ **결측치 & 중복되는 행** 확인

Distinct

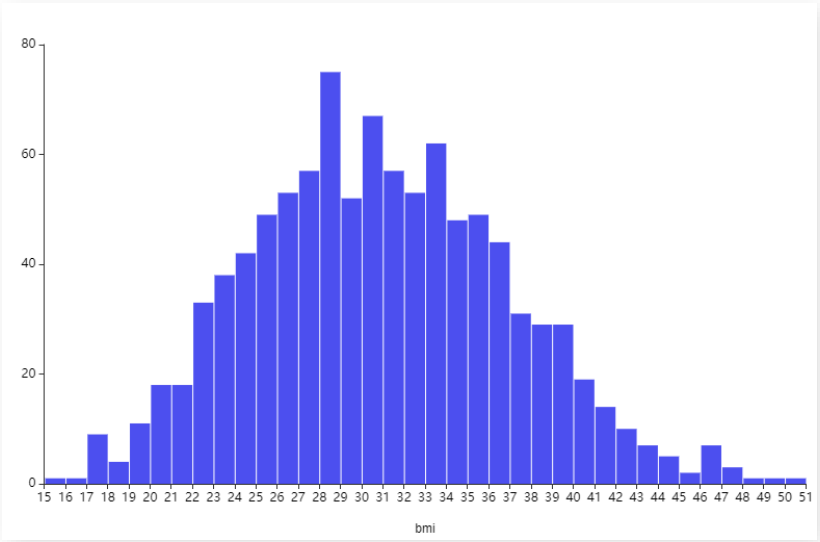


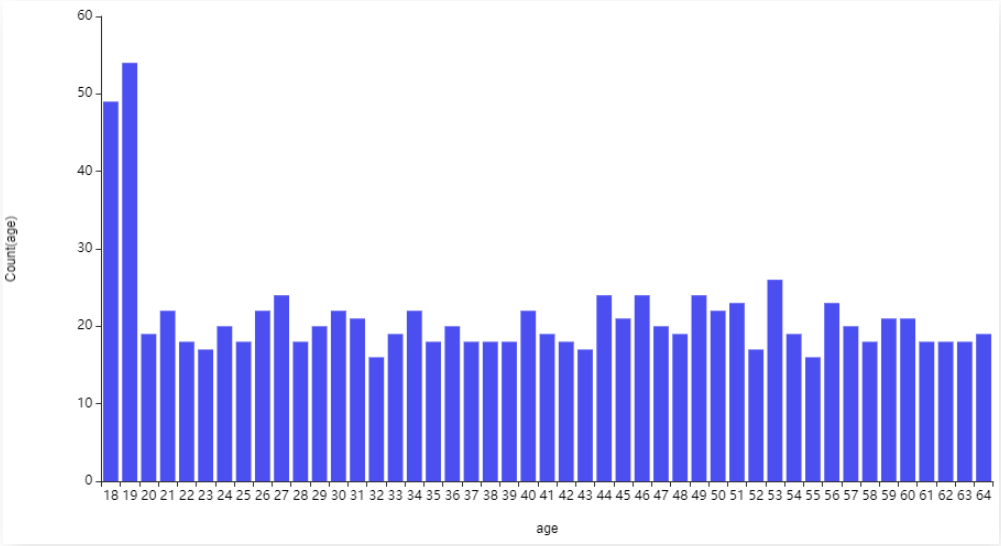
- 결측치는 0%로 존재하기 때문에 전처리 생략
- Distinct 블록에 전체 컬럼을 입력하여 중복행 제거

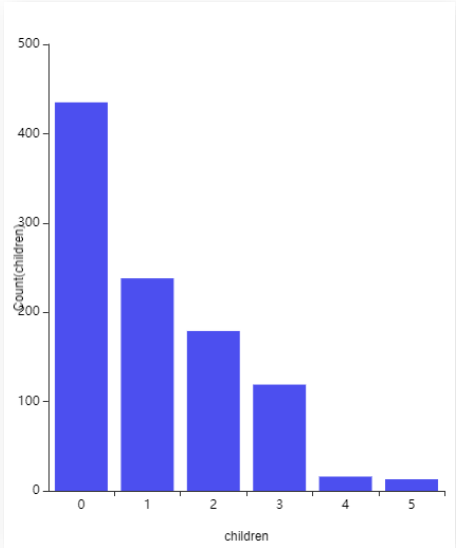
Part 3 데이터 전처리

결측치 · 중복행 처리 | **Outlier Detection** | Label Encoder

* 수치형 변수만 확인







Bmi

15~54까지 분포

bmi가 50 이상인 값도 계산상 가능한

정상치로 간주

Age

17-65세까지 분포

이상치 없음

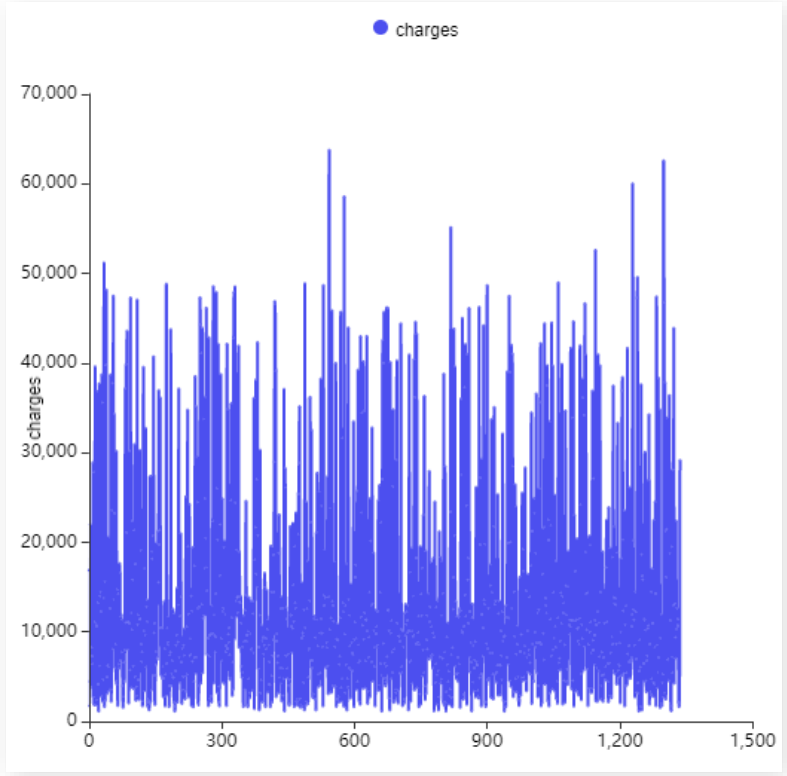
Children

0부터 5명까지 구성

이상치 없음

Part 3 데이터 전처리

결측치 · 중복행 처리 | Outlier Detection | Label Encoder



charges ↑
1121.8739
1131.5066
1135.9407
1136.3994
1137.011
1137.4697
1141.4451
1146.7966
1149.3959
1163.4627
1241.565
1242.26

charges ↓
63770.42...
62592.87...
60021.39...
58571.07...
55135.40...
52590.82...
51194.55...
49577.6624
48970.2476
48885.13...
48824.45

Charges

1121부터 63770으로 구성,

값들 간 다소 차이가 있지만

연소자, 비흡연자일수록 의료비가 낮다.

연장자, 흡연자일수록 의료비가 높다.

이상치 없음.

Part 3 데이터 전처리

결측치 · 중복행 처리 | Outlier Detection | Label Encoder

Sex & Smoker

	age	sex	bmi	children	smoker	region	charges	sex_f0m1	smoker_y1n0
1	19	female	27.9	0	yes	southwest	16884.924	0	1
2	18	male	33.77	1	no	southeast	1725.5523	1	0
3	28	male	33	3	no	southeast	4449.462	1	0
4	33	male	22.705	0	no	northwest	21984.47...	1	0
5	32	male	28.88	0	no	northwest	3866.8552	1	0
6	31	female	25.74	0	no	southeast	3756.6216	0	0
7	46	female	33.44	1	no	southeast	8240.5896	0	0
8	37	female	27.74	3	no	northwest	7281.5056	0	0
9	37	male	29.83	2	no	northeast	6406.4107	1	0
10	60	female	25.84	0	no	northwest	28923.13...	0	0
11	25	male	26.22	0	no	northeast	2721.3208	1	0

Sex

Female = 0, Male = 1 로 변경

Input Columns : sex, Suffix : _f0m1

입력하여 sex_f0m1 열을 추가

Smoker

흡연=1, 비흡연=0 로 변경

Input Columns : smoker, Suffix : _y1n0

입력하여 smoker_y1n0 열을 추가

Part 3 데이터 전처리

One-hot Encoder

로그변환 • 정규성 검정

Region

region	charges	sex_f0m1	smoker_y1n0	region_northeast	region_northwest	region_southeast	region_southwest
southwest	16884.924	0	1	0	0	0	1
southeast	1725.5523	1	0	0	0	1	0
southeast	4449.462	1	0	0	0	1	0
northwest	21984.47...	1	0	0	1	0	0
northwest	3866.8552	1	0	0	1	0	0
southeast	3756.6216	0	0	0	0	1	0
southeast	8240.5896	0	0	0	0	1	0
northwest	7281.5056	0	0	0	1	0	0
northeast	6406.4107	1	0	1	0	0	0
northwest	28923.13...	0	0	0	1	0	0
northeast	2721.3208	1	0	1	0	0	0
southeast	27808.7251	0	1	0	0	1	0
southwest	1826.843	1	0	0	0	0	1

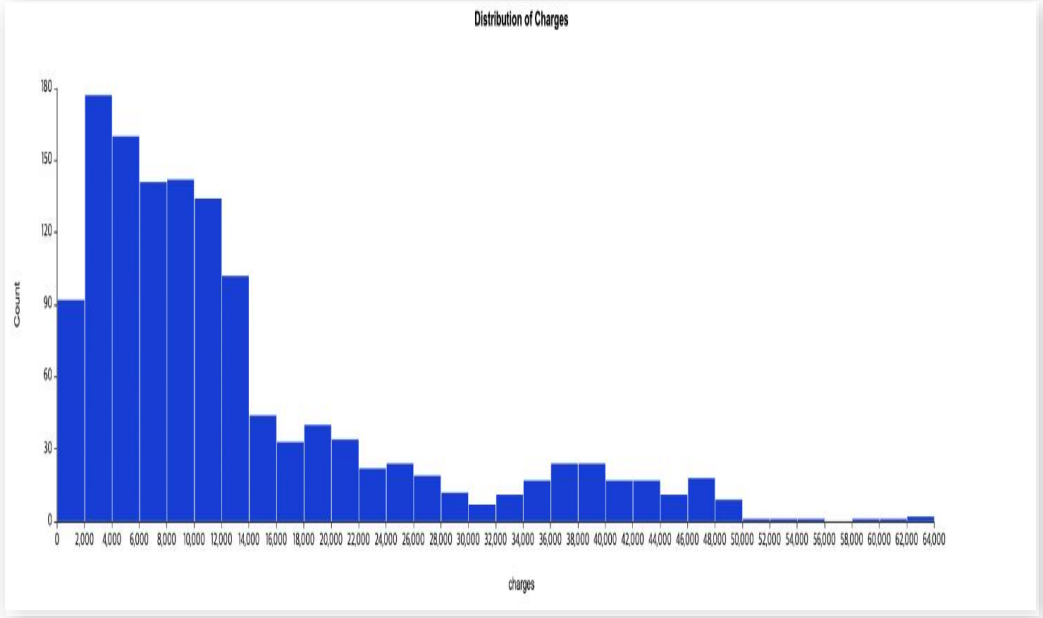
명목형 변수인 지역은 4개의 값으로 작성되어 있어서 **One-hot encoding으로 전처리**

Input Columns : region, Suffix_Type : Label

입력하여 region_지역명으로 추가된 4개의 열을 볼 수 있다.

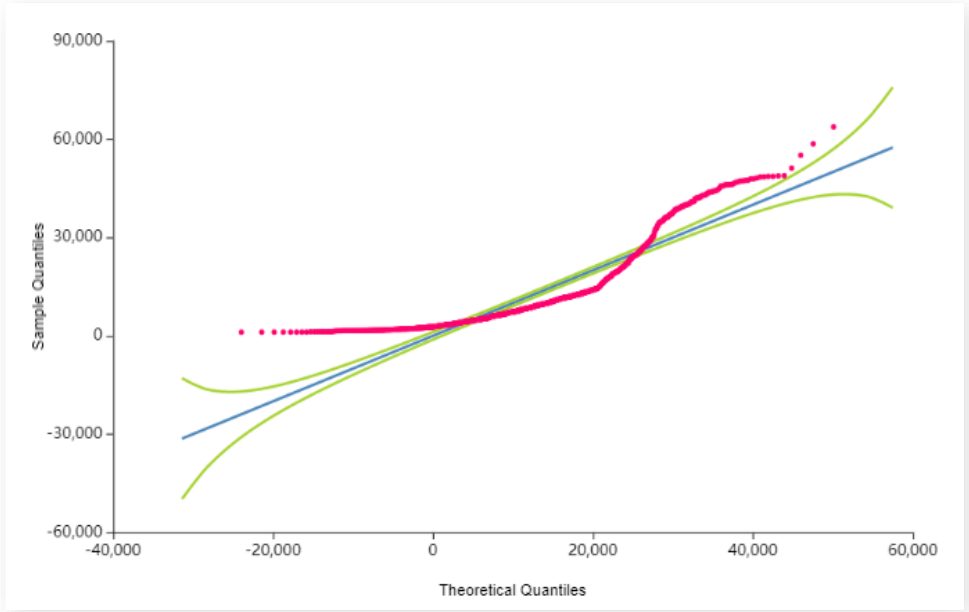
로그 변환 전

의료비 Histogram



- 그래프는 **왼쪽으로 치우쳐진 형태**

의료비 Normal 분포 (Q-Q plot)

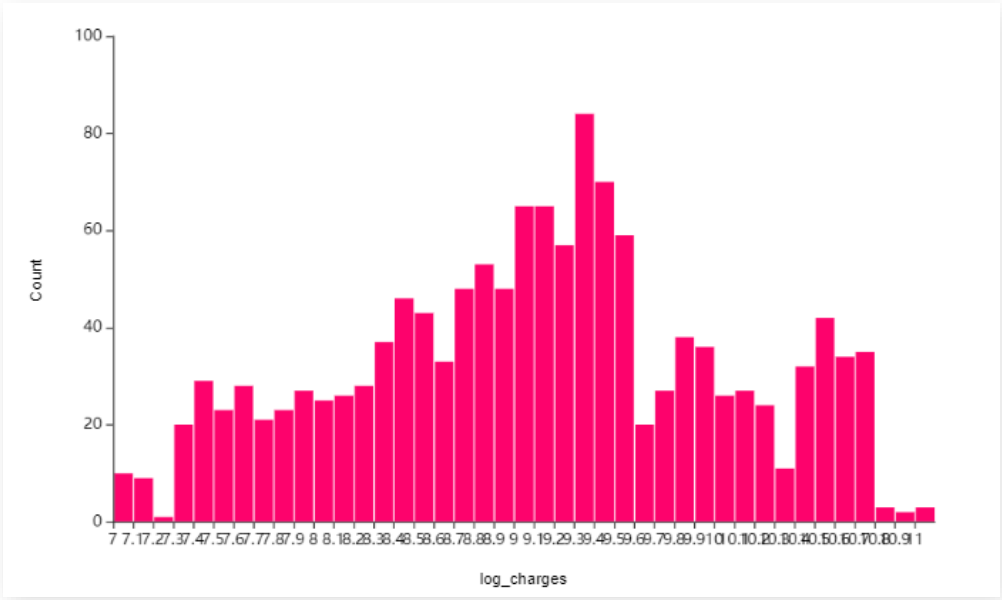


- Normal Distribution의 경우,
  연두선을 크게 벗어남. (정규성 만족 x)

Part 3 데이터 전처리

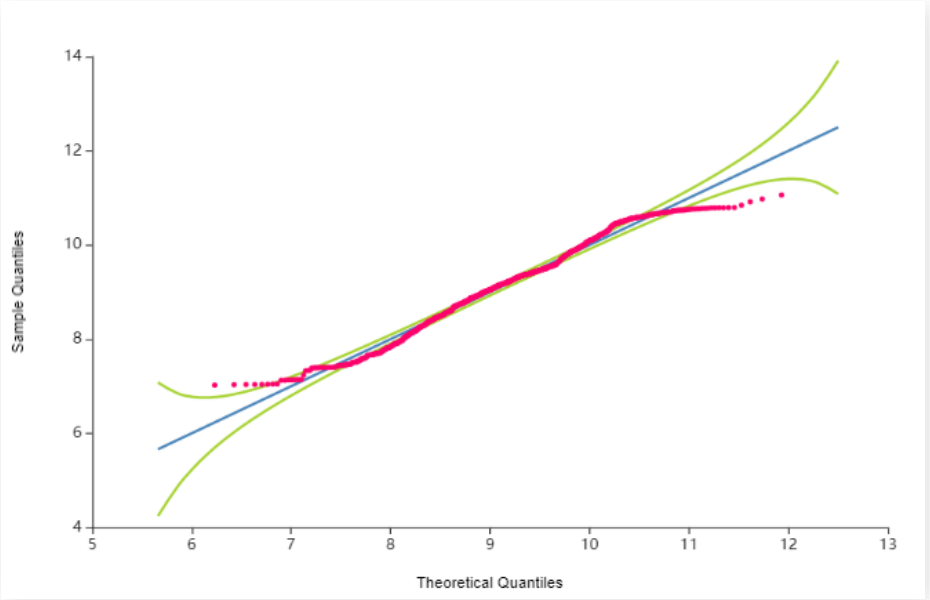
로그 변환 후 : Python Script에서 numpy 모듈의 log1p를 사용해 charges에 로그를 취한다.

의료비 Histogram



- 그래프가 **고르게 퍼져있다.**

의료비 Normal 분포 (Q-Q plot)



- 빨간색 선이 **연두색 선 내에서 많은 부분 겹쳐짐을 볼 수 있다.**

Normality Test

Anderson-Darling test result

data	estimates
charges	85.12851936032212

로그변환 **전**

Anderson-Darling test result

data	estimates
log_charges	3.929912675382411

로그변환 **후**

- Ad 통계량이 **작을수록 특정 분포를 잘 따르는 성질**을 띄는데,
  로그변환 전 측정한 **85**에서 **로그변환 후 3.9로 감소했음을 볼 수 있다.**

Part 4.

Modeling

01 Process

02 Linear Regression

03 Decision Tree

04 AdaBoost

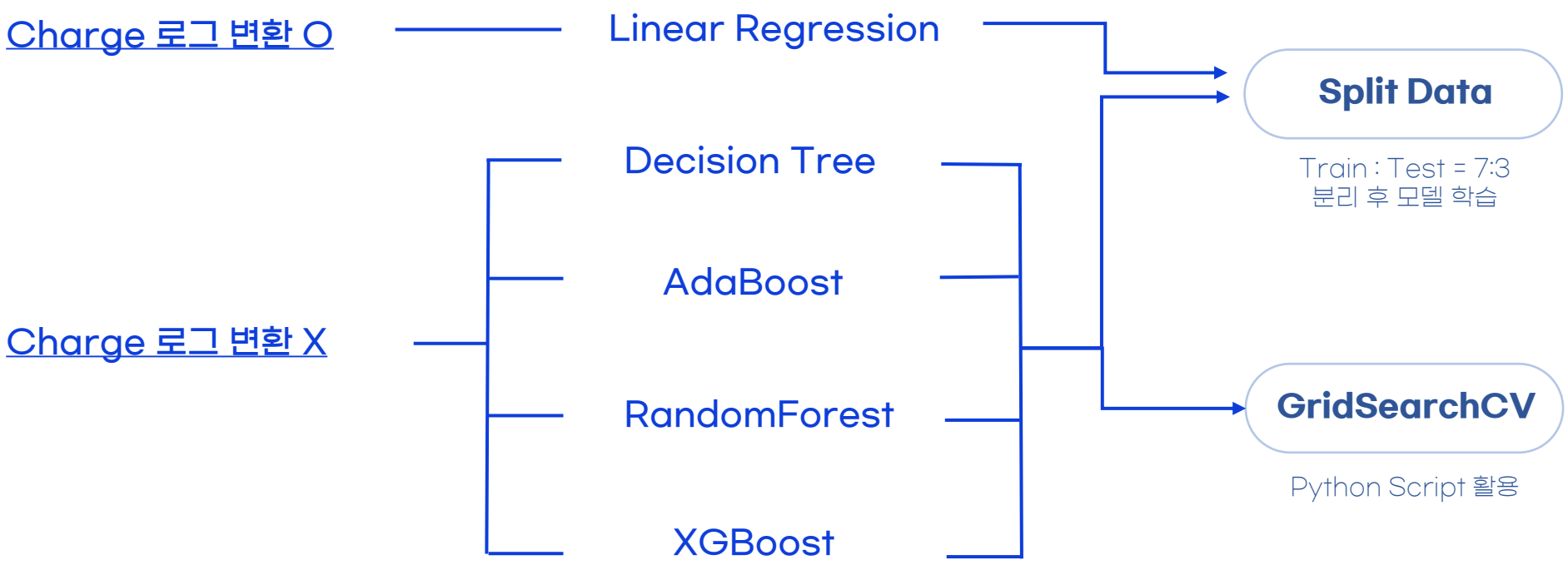
05 Random Forest

06 XGBoost

Part 4 전체 모델링

Process | 1) Linear Regression | 2) Decision Tree

전체(All Data)



* 로그변환은 선형회귀 알고리즘 적용시에는 유용하나, tree 계열 알고리즘에서는 로그변환이 예측결과에 전혀 영향을 주지 않음

→ Linear Regression에만 로그변환한 charge를 적용함.

Part 4 전체 모델링

python script

GridSearchCV (ex) RandomForest

```
data = inputs[0]
X_train, X_test, y_train, y_test = train_test_split(
  data.iloc[:,:-1], data.iloc[:,-1], test_size=0.3, random_state=0)

#X_train =
trainset.loc[:,['age','bmi','sex_f0m1','region_northeast','region_northwest','region_southeast','region_southwest']]
#y_train = trainset['log_charges'].values
#X_test =
testset.loc[:,['age','bmi','sex_f0m1','region_northeast','region_northwest','region_southeast','region_southwest']]
#y_test = testset['log_charges'].values

parameters = {  "n_estimators": [13,14,15, 16, 17, 18],
                "max_depth": [3,4, 5, 6, 7],
                "min_samples_leaf": [1, 2, 3, 4],
                "min_samples_split": [7,8,9,10,11],
                "max_features" : ['sqrt', 'log2', None],}

clf = RandomForestRegressor()

rf_model = GridSearchCV(clf, param_grid = parameters,cv =5)
rf_model.fit(X_train, y_train)

best_params = rf_model.best_params_
best_model = rf_model.best_estimator_

y_predict = rf_model.predict(X_test)

df = pd.DataFrame({'charges': y_test ,'prediction': y_predict})
```

Part 4 전체 모델링

사용한 평가 지표

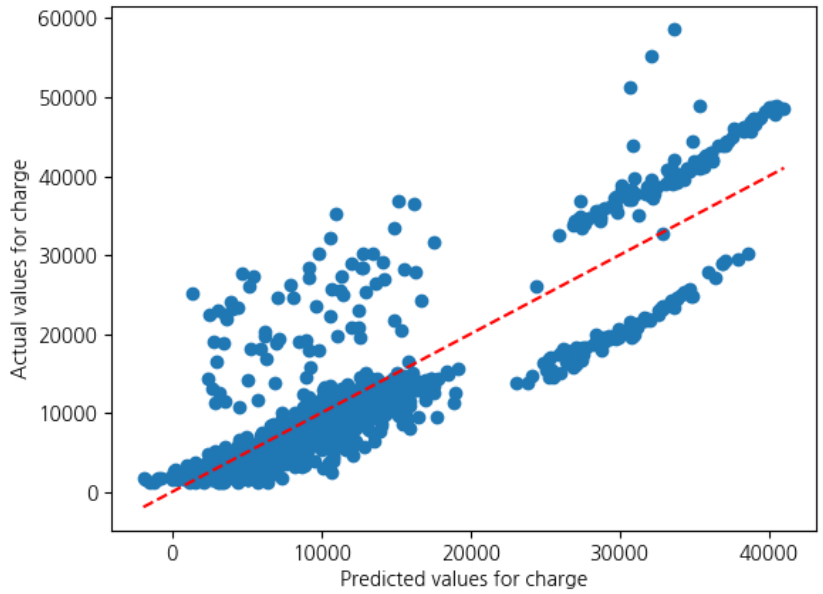
	거리		상관성
	MAE	MAPE	R-Square
	$MAE = \frac{\sum \lvert y - \hat{y} \rvert}{n}$	$MAPE = \frac{\sum \lvert \frac{y-\hat{y}}{y} \rvert}{n} * 100\%$	$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$
특징	- 이상치의 영향을 적게 받음	- MAE를 퍼센트로 변환한 개념	- 상대적으로 얼마나 성능이 나오는지 측정하는 지표
평가지표 선정 이유	- 정답 및 예측값과 같은 단위를 가진다.	- 다른 모델과 에러율 비교가 쉽다	- 모델이 데이터를 얼마나 잘 설명했는지 알 수 있다.

Part 4 전체 모델링

Process | 1) Linear Regression | 2) Decision Tree

1) Train - Split Data

Predicted vs Actual



Linear Regression Result

Summary

OLS Regression Results

Dep. Variable:	log_charges	R-squared:	0.755
Model:	OLS	Adj. R-squared:	0.753
Method:	Least Squares	F-statistic:	357.2
Date:	Sun, 04 Sep 2022	Prob (F-statistic):	6.68e-277
Time:	14:42:29	Log-Likelihood:	-589.74
No. Observations:	935	AIC:	1197.
Df Residuals:	926	BIC:	1241.
Df Model:	8		
Covariance Type:	nonrobust		

- R-square는 0.742, Adj R-squared 0.740으로
  약 74%정도의 설명력을 가지는 모델이 생성됨.

  * R-square값은 클수록 좋은 모델

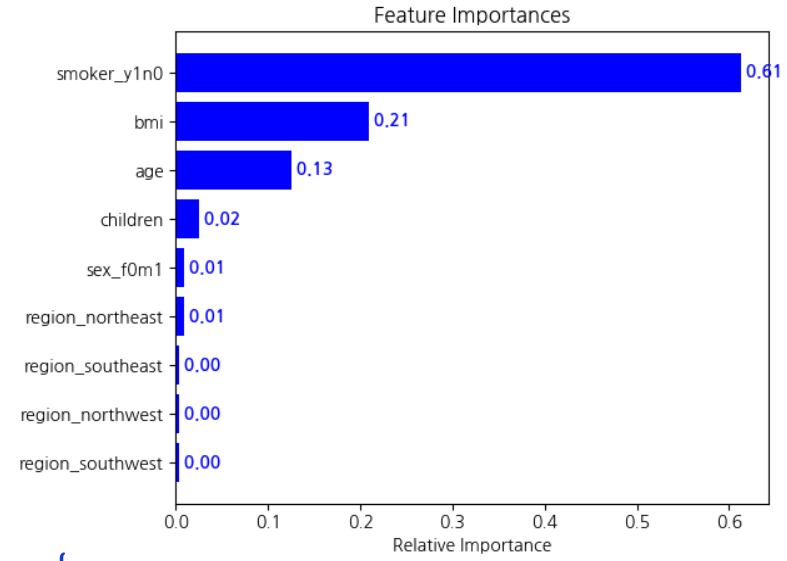
- AIC와 BIC의 값이 높아
  Linear Regression 모델의 적합성이 떨어짐을 알 수 있음

Part 4 전체 모델링

Process | 1) Linear Regression | 2) Decision Tree

1) Train - Split Data

Feature Importance



2) GridSearch - Cross Validation

[Model Best Parameters]

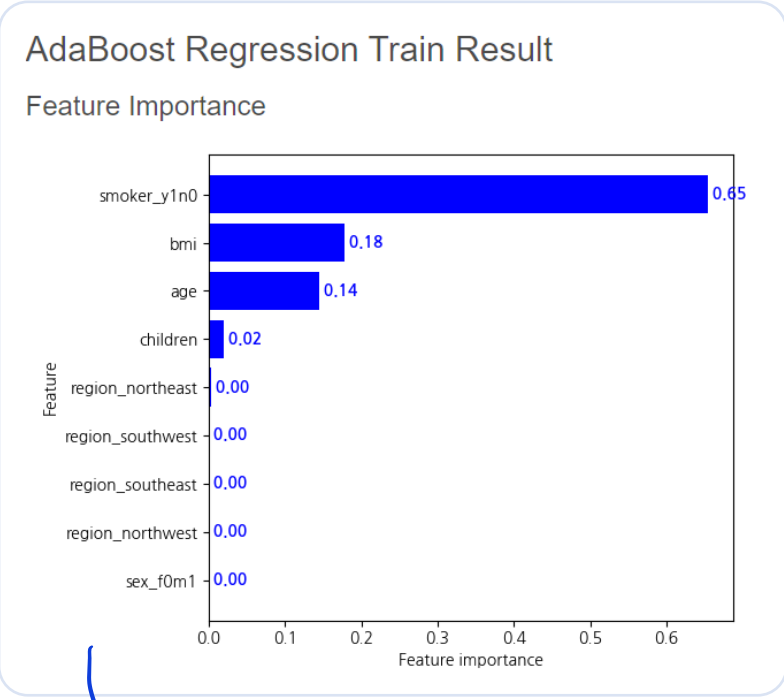
```
DecisionTreeRegressor(criterion='mse', max_depth=5,
max_features='auto',
max_leaf_nodes=None, min_impurity_decrease=0.0,
min_impurity_split=None, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.1,
presort=False, random_state=None, splitter='best')
```

- 흡연여부의 중요도가 0.61로 가장 높다. 그 다음, bmi, 나이순으로 중요도가 높다.
- 부양 자녀수/ 성별 / 거주 지역은 0.02에서 0.00 사이로 중요도가 낮은 편에 속한다.

Part 4 전체 모델링

3) Adaboost | 4) Random Forest | 5) XGBoost

1) Train - Split Data



2) GridSearch - Cross Validation

[Model Best Parameters]

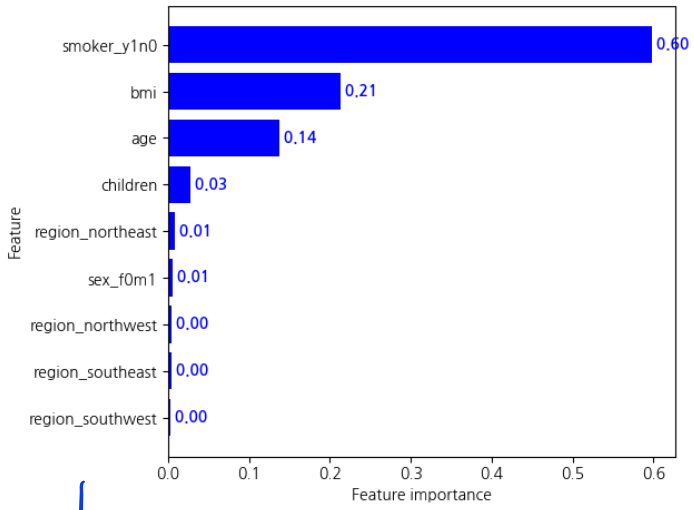
```
AdaBoostRegressor(base_estimator=None, learning_rate=0.0001,
loss='exponential',
n_estimators=80, random_state=None)
```

- 흡연여부의 중요도가 0.65로 가장 높다. 그 다음, bmi, 나이순으로 중요도가 높다.
- 부양 자녀수와 성별, 거주 지역은 0.02와 0.00으로 중요도가 낮은 편에 속한다.

Part 4 전체 모델링

3) Adaboost | 4) Random Forest | 5) XGBoost

1) Train - Split Data



2) GridSearch - Cross Validation

[Model Best Parameters]

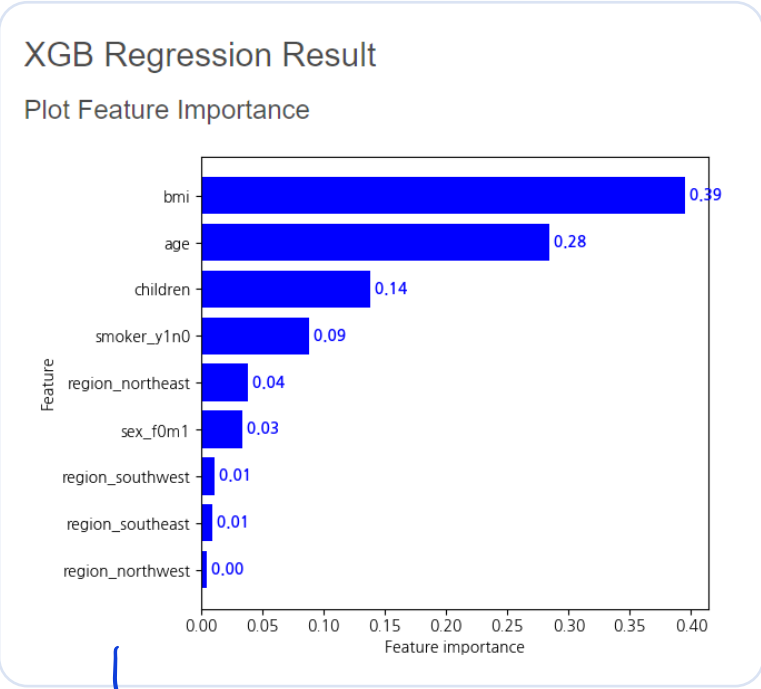
```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=4,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=4, min_samples_split=10,
min_weight_fraction_leaf=0.0, n_estimators=15,
n_jobs=None, oob_score=False, random_state=None,
verbose=0, warm_start=False)
```

- 흡연여부의 중요도가 0.60으로 가장 높다. 그 다음, bmi, 나이순으로 중요도가 높다.
- 부양 자녀수와 성별, 거주 지역은 0.03에서 0.00 사이로 중요도가 낮은 편에 속한다.

Part 4 전체 모델링

3) Adaboost | 4) Random Forest | 5) XGBoost

1) Train - Split Data



2) GridSearch - Cross Validation

[Model Best Parameters]

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bytree=0.7, gamma=0, learning_rate=0.1, max_delta_step=0,
max_depth=5, min_child_weight=1, missing=None, n_estimators=80,
n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
silent=True, subsample=0.5)
```

- BMI의 중요도가 0.39로 가장 높다. 그 다음 나이, 부양자녀수, 흡연여부 순으로 중요도가 높다.
- 성별, 거주 지역은 0.04에서 0.00 사이로 중요도가 낮은 편에 속한다.

What is the best model?

Best model

모델구성	구분	R-squared	MAE	MAPE
Linear Regression	-	0.5142	4326.1094	27.3802
DecisionTree Regression	Train	0.6649	3437.3757	56.4213
	GridSearchCV	0.8395	3243.4083	39.4671
AdaBoost Regression	Train	0.8395	3900.8289	65.1139
	GridSearchCV	0.8541	2987.9688	35.8850
RandomForest Regression	Train	0.8426	2788.7546	38.3939
	GridSearchCV	0.8602	2846.8738	33.7300
XGBoost Regression	Train	0.8805	2455.0297	31.8088
	GridSearchCV	0.8511	2893.5439	33.1910

XGBoost Regression의 경우 R-squared가 가장 높고, MAE,MAPE가 가장 낮다.

-> XGBoost Regression이 최적 모델

Part 5.

Modeling
: group by Sex

01 **Process**

02 **Linear Regression**

03 **Decision Tree**

04 **AdaBoost**

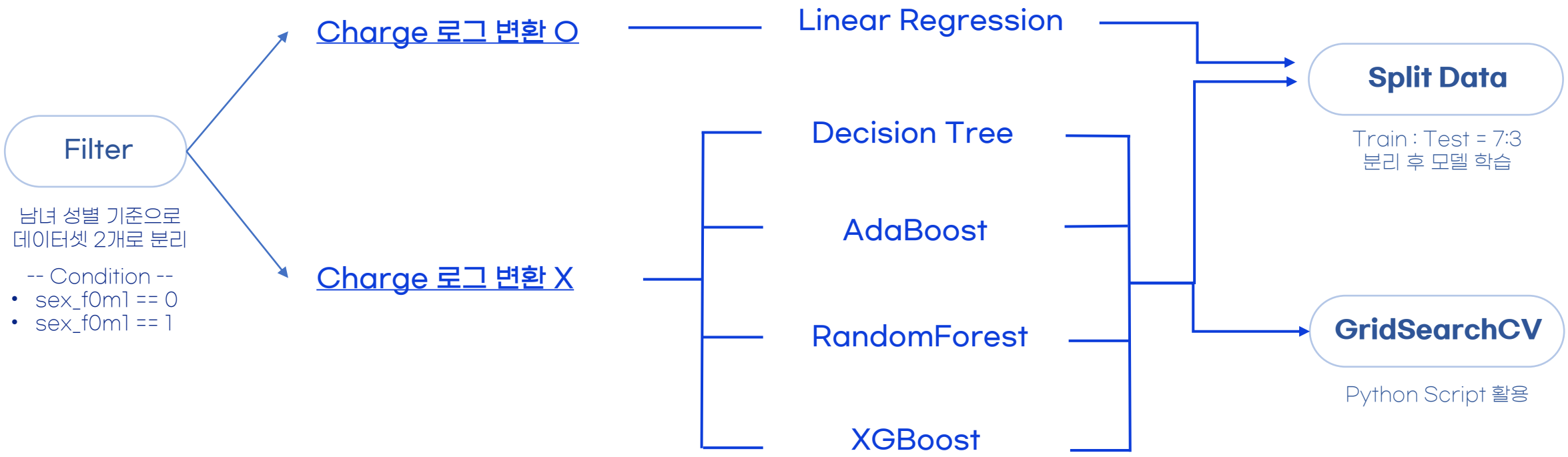
05 **Random Forest**

06 **XGBoost**

Part 5 성별 모델링

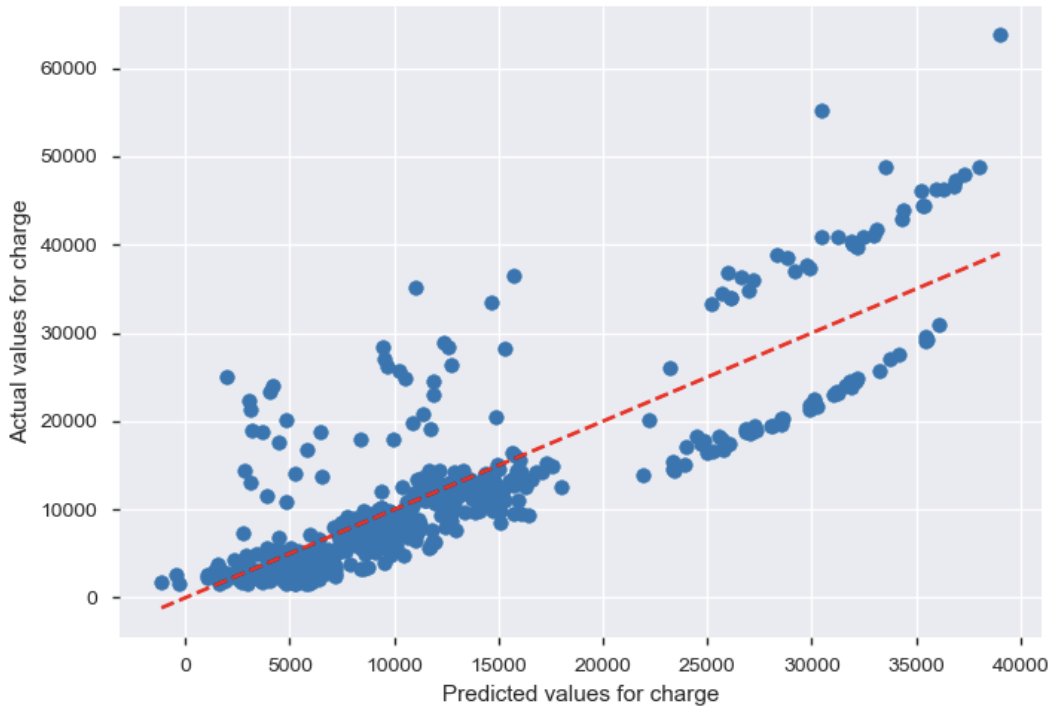
Process | 1) Linear Regression | 2) Decision Tree

성별(Group by Sex)



Part 5 성별 모델링

1) Train - Split Data



Predict vs Actual

R-squared:	0.694
Adj. R-squared:	0.689
F-statistic:	147.4
Prob (F-statistic):	9.65e-113
Log-Likelihood:	-4690.0
AIC:	9396.
BIC:	9429.

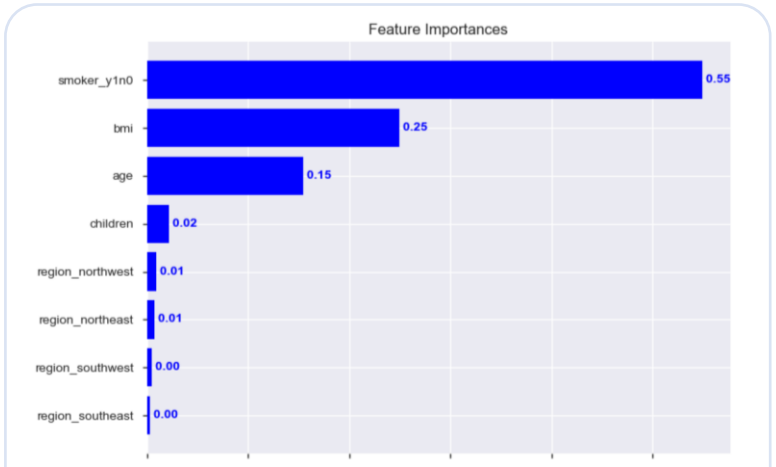
- R-squared = 0.694, Adj R-squared = 0.689
  약 68%정도의 설명력을 가지는 모델이 생성됨

- AIC = 9396으로 높은 값으로
  Linear Regression 모델의 적합성이 떨어짐을 알 수 있음

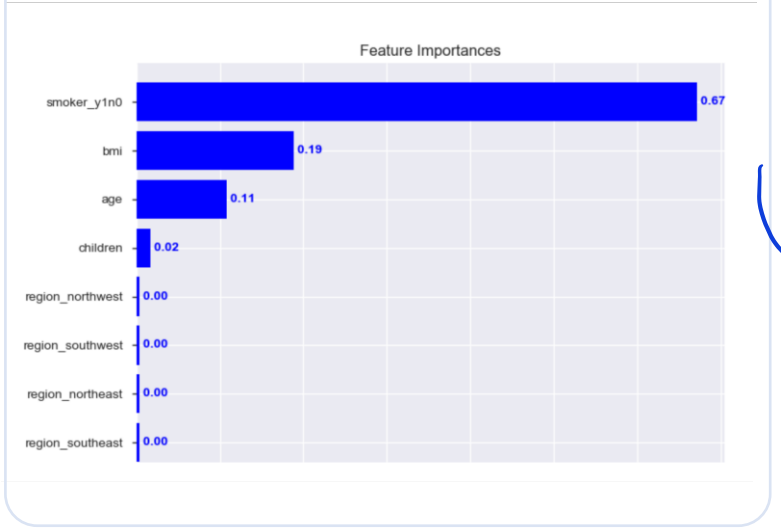
Part 5 성별 모델링

1) Train - Split Data

Female



Male



2) GridSearch - Cross Validation

Female

```
DecisionTreeRegressor(criterion='mse', max_depth=12,
max_features=None,
max_leaf_nodes=70, min_impurity_decrease=0.0,
min_impurity_split=None, min_samples_leaf=3,
min_samples_split=2, min_weight_fraction_leaf=0.1,
presort=False, random_state=None, splitter='random')
```

Male

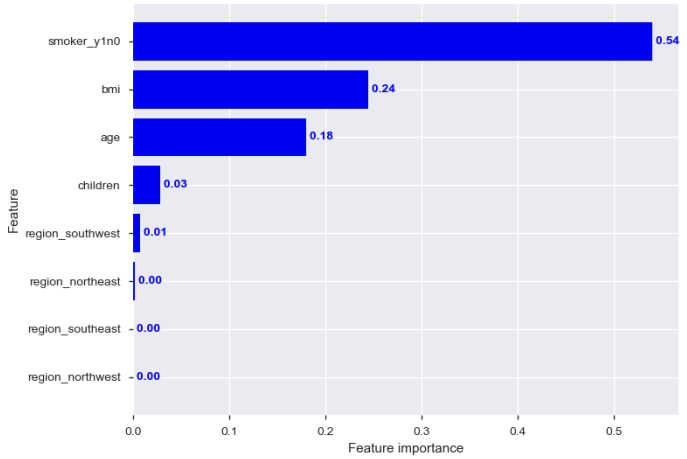
```
DecisionTreeRegressor(criterion='mse', max_depth=5,
max_features='auto',
max_leaf_nodes=None, min_impurity_decrease=0.0,
min_impurity_split=None, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.1,
presort=False, random_state=None, splitter='best')
```

- 전체적으로 보았을 때, 남녀 간의 큰 차이는 없음.
- 그러나, 여성이 남성보다 흡연 여부의 중요성이 0.12 정도 낮고, bmi의 중요성이 0.06 정도 높은 것으로 판단됨.

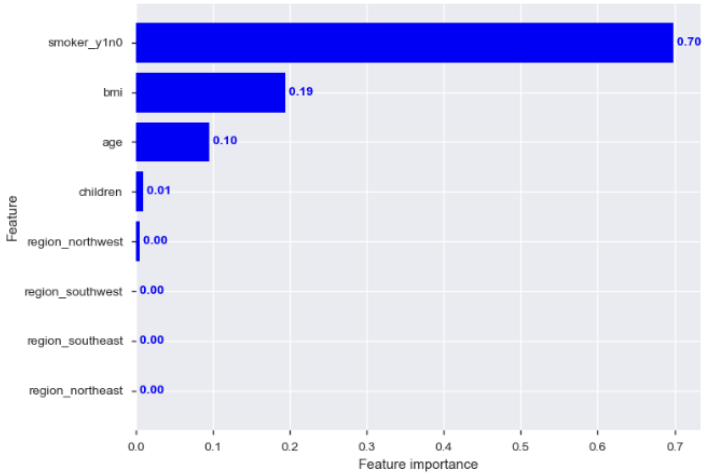
Part 5 성별 모델링

1) Train - Split Data

Female



Male



2) GridSearch - Cross Validation

Female

```
AdaBoostRegressor(base_estimator=None, learning_rate=0.0001, loss='square',
n_estimators=80, random_state=None)
```

Male

```
AdaBoostRegressor(base_estimator=None, learning_rate=0.001, loss='linear',
n_estimators=30, random_state=None)
```

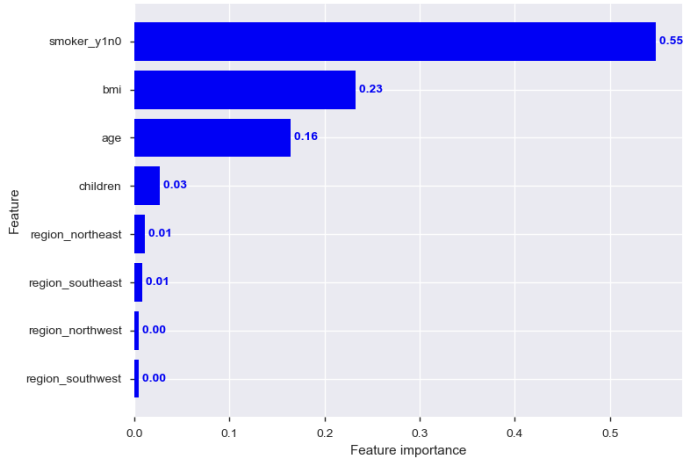
- 남,여 동일하게 Smoker_y1n0, bmi, age의 중요도 순위를 가진다.
- 여자는 남자보다 children에 미치는 영향이 0.02정도 더 큼

Part 5 성별 모델링

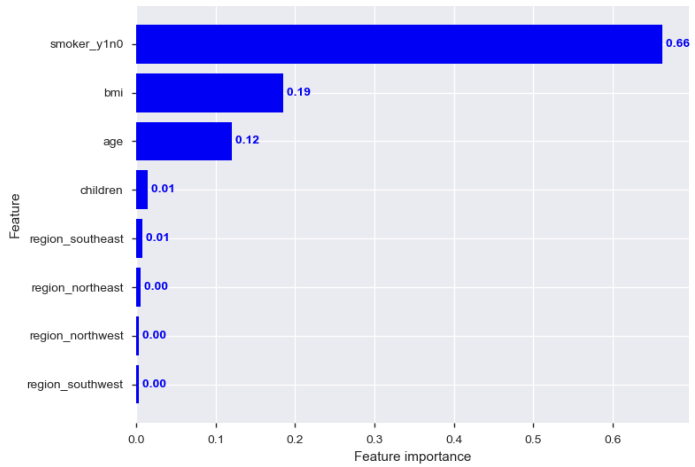
3) Adaboost | 4) Random Forest | 5) XGBoost

1) Train - Split Data

Female



Male



2) GridSearch - Cross Validation

Female

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=3,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=4, min_samples_split=9,
min_weight_fraction_leaf=0.0, n_estimators=13,
n_jobs=None, oob_score=False, random_state=None,
verbose=0, warm_start=False)
```

Male

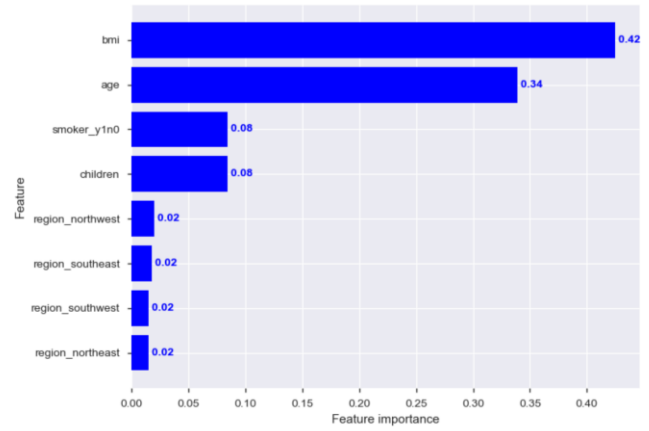
```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=3,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=4, min_samples_split=8,
min_weight_fraction_leaf=0.0, n_estimators=12,
n_jobs=None, oob_score=False, random_state=None,
verbose=0, warm_start=False)
```

- smoker_y1n0 / bmi / age 세 변수가 여성과 남성 둘 다 보다 더 유의미하다.
- smoker_y1n0은 남성에게, bmi와 age는 여성에게 더 영향을 미친다

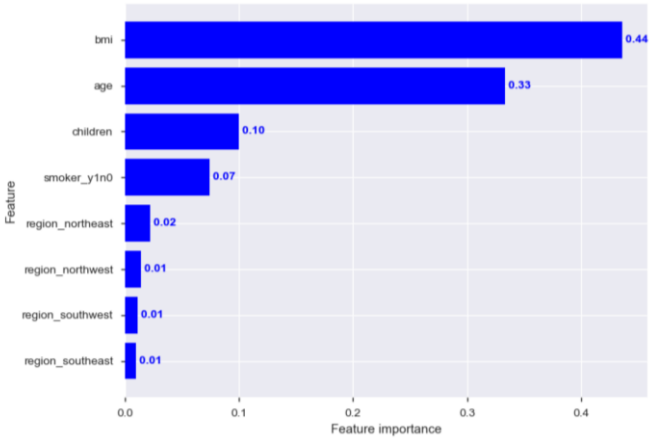
Part 5 성별 모델링

1) Train - Split Data

Female



Male



2) GridSearch - Cross Validation

Female

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bytree=1.0, gamma=0, learning_rate=0.05,
max_delta_step=0, max_depth=3, min_child_weight=1, missing=None,
n_estimators=90, n_jobs=1, nthread=None, objective ='reg:linear',
random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1,
seed=None, silent=True, subsample=0.7)
```

Male

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bytree=0.7, gamma=0, learning_rate=0.1, max_delta_step=0,
max_depth=5, min_child_weight=1, missing=None, n_estimators=50,
n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed= None,
silent=True, subsample=0.5)
```

- 전체적으로 남녀 간의 큰 차이는 없다.
- 여성의 경우 부양 자녀 수와 흡연 여부의 중요도가 0.08로 같지만, 남성의 경우 부양 자녀 수는 0.10, 흡연 여부는 0.07의 중요도를 가져 부양자녀 수의 중요도가 더 높다는 것을 알 수 있다.

What is the best model?

Best model (for. 여성)

모델구성	구분	R-squared	MAE	MAPE
Linear Regression	-	0.5999	3502.5495	21.3250
DecisionTree Regression	Train	0.6280	3393.6735	35.7606
	GridSearchCV	0.5878	4952.7778	46.1039
AdaBoost Regression	Train	0.7818	4518.3908	68.1321
	GridSearchCV	0.7990	2863.1551	29.3353
RandomForest Regression	Train	0.7847	2968.0723	32.4753
	GridSearchCV	0.7972	2902.0931	29.6373
XGBoost Regression	Train	0.8402	2535.7157	27.1813
	GridSearchCV	0.8063	2621.4972	23.6831

XGBoost Regression에서 R-squared가 가장 높고, MAE, MAPE가 가장 낮다.

-> XGBoost Regression이 최적 모델

What is the best model?

Best model (for. 남성)

모델구성	구분	R-squared	MAE	MAPE
Linear Regression	-	0.6374	4486.7413	29.6237
DecisionTree Regression	Train	0.7636	3378.2610	32.5729
	GridSearchCV	0.8678	3118.2285	34.2620
AdaBoost Regression	Train	0.8380	4491.3278	91.6623
	GridSearchCV	0.8886	2444.5435	29.8163
RandomForest Regression	Train	0.8596	2914.2939	34.4022
	GridSearchCV	0.8863	2490.8895	31.7018
XGBoost Regression	Train	0.8796	2562.6684	31.3735
	GridSearchCV	0.8559	3464.2745	50.8211

AdaBoost Regression에서 R-squared가 높고, MAE, MAPE가 가장 낮다.

-> AdaBoost Regression이 최적 모델

Part 6.

Result

Result.

Modeling(All)

“ 체질량지수(bmi)가 높을수록, 나이(age)가 많을수록 ,
부양자녀수(children)가 많을수록, 흡연(smoke)을 할수록
의료비가 많이 부과된다.”

Modeling : group by Sex

남성의 의료비는 부양자녀수(children)에 의해 더 많이 영향을 받는다.

여성의 의료비는 흡연여부(smoke)와 부양자녀수(children)에 의해 더 영향을 받는다.

Result.

“보험회사는 보험료 변화 추세를 파악하여, 그에 맞는 상품을 기획할 수 있다.”

“보험회사는 여성,남성 구분 없이 **체질량지수(bmi), 나이(age), 자녀 부양자수(chlidren),**
흡연여부(smoke)를 기준으로 맞춤형 상품 기획할 수 있다.”

소감.

서영석

브라이틱스로
팀 프로젝트를 같이
진행하여 뜻깊은
경험이었습니다.
재밌고 즐거웠던
프로젝트였습니다!

오수민

이번 팀 분석
활동으로 분석의
흐름을 설정하는
역량을 키우고
데이터를 다루는
시야가 넓어졌다고
생각합니다.

최수빈

팀프로젝트를
진행하면서
Brigthics를
더 알 수 있는
기회가 되었습니다.

정민경

브라이틱스로
협업하니 서로
무엇을 진행했는지
이해하기 한결
수월하더라고요.
그 덕에 즐겁고
다들 고마웠어요!

홍수정

팀원들과 함께
브라이틱스로
분석하며
많이 배웠습니다!