# <오전>

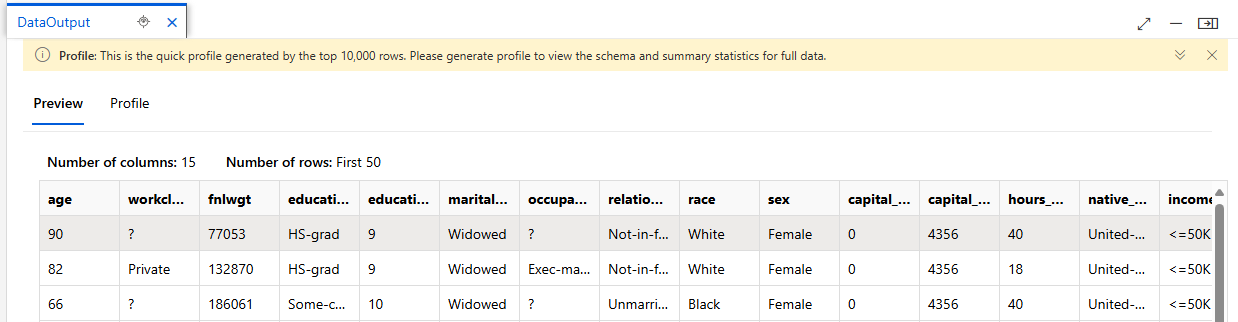
## 환경

* Azure
* .csv 파일

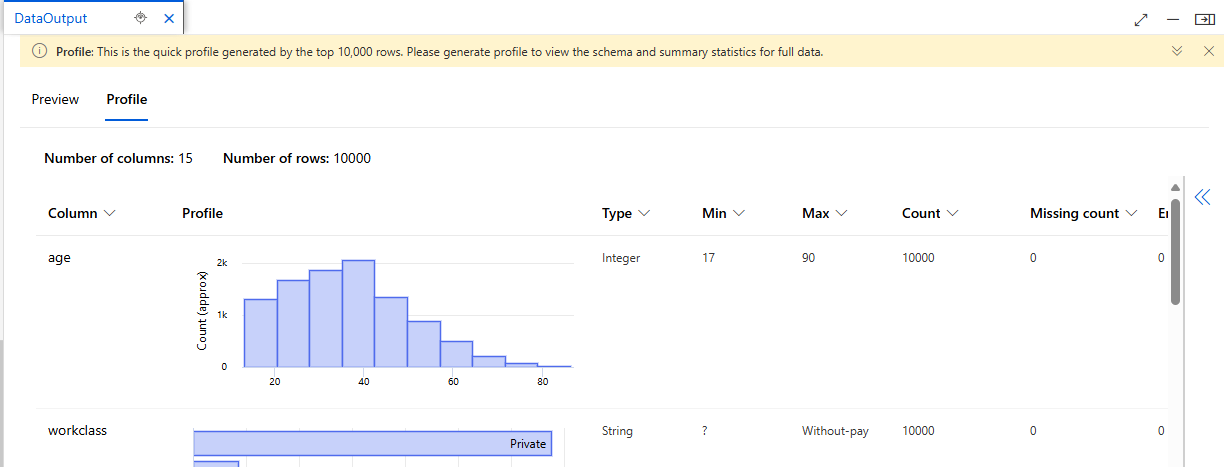
## [실습 - 분류MLD\_개인수입예측\_랜덤포레스트]

* 앞의 내용은 07\_03 참고

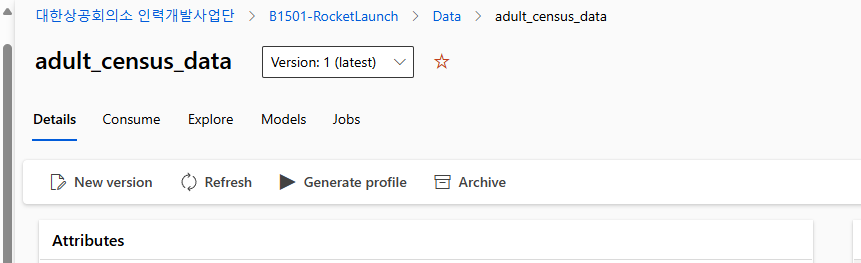
### 데이터 이해

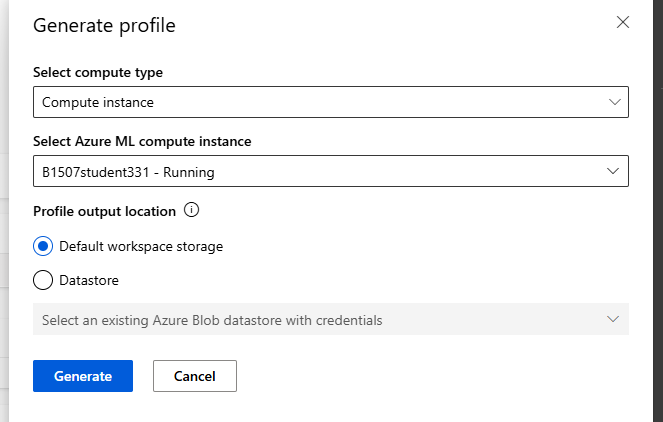


Profile은 10,000개 행에 대해서 보여지고 있으며, Profile 생성 기능을 통해 모든 데이터에 대한 Profile 확인이 가능하다.

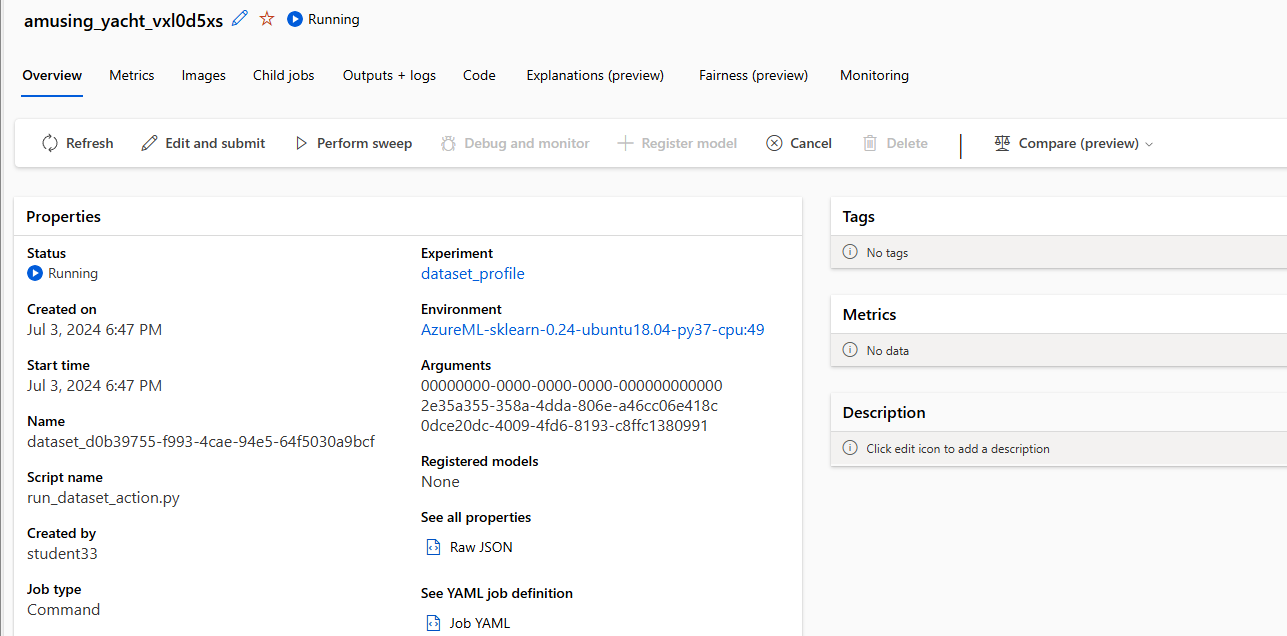


Data에서 선택한 후 Generate profile 클릭

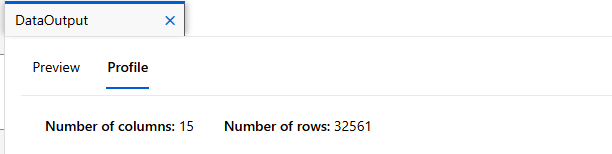




Jobs 메뉴에서 Profile 생성 작업의 진행상황 확인이 가능하다.

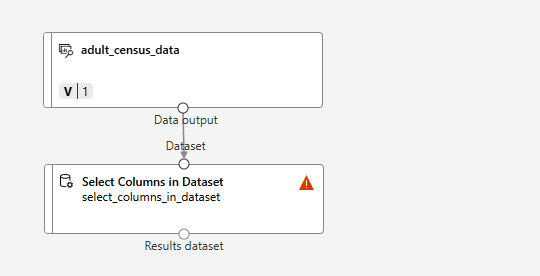


Generate profile 이후 행이 32561개로 변한 것을 확인할 수 있다.

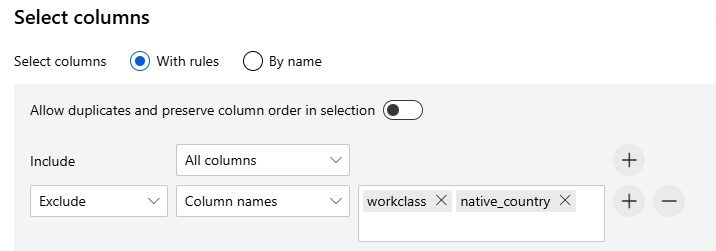


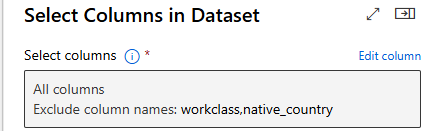
### 데이터 준비

#### 특성 선택 – 불필요한 컬럼 제외

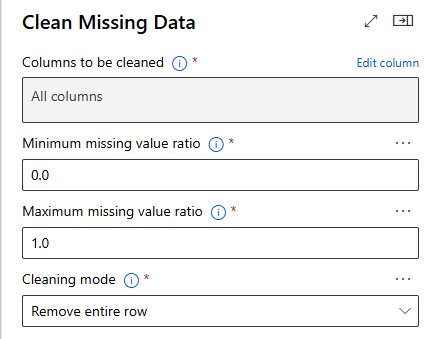


모든 정보를 선택하고 선택한 것을 제외한다.



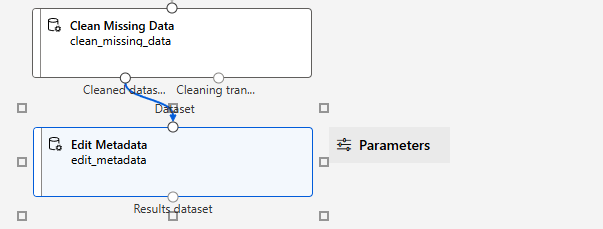


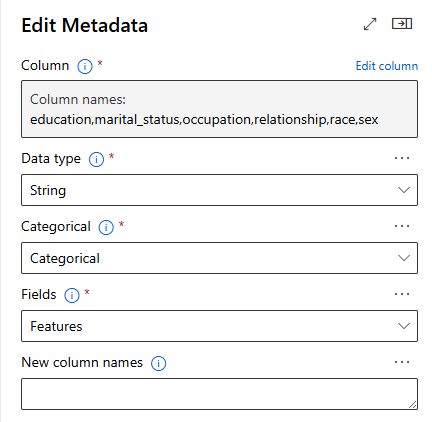
#### 누락값 처리 – 데이터 샘플 제거



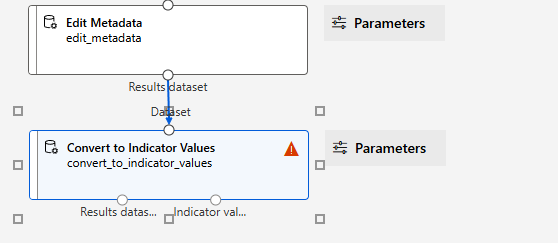
#### 데이터 변환 : String -> Category -> Indicator value

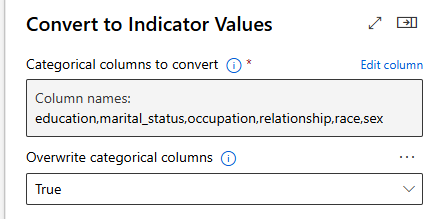
String -> Category





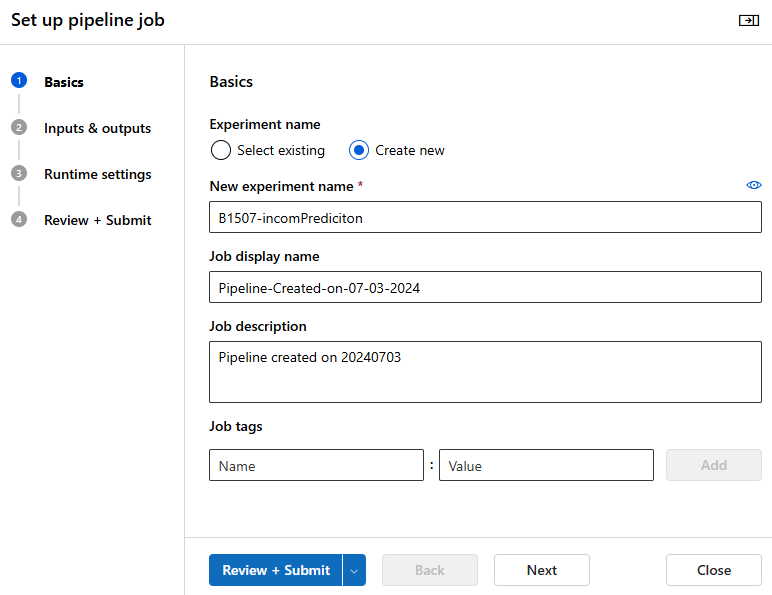
Category -> Indicator

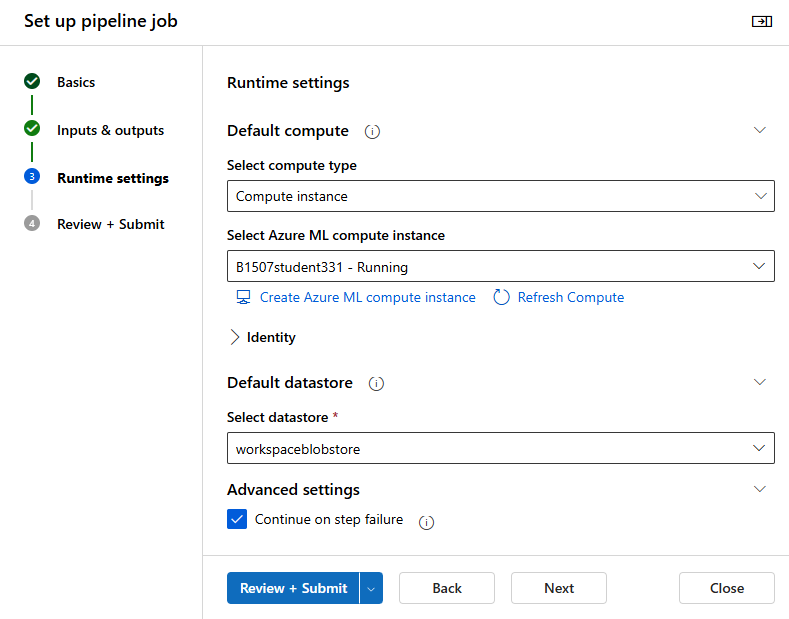




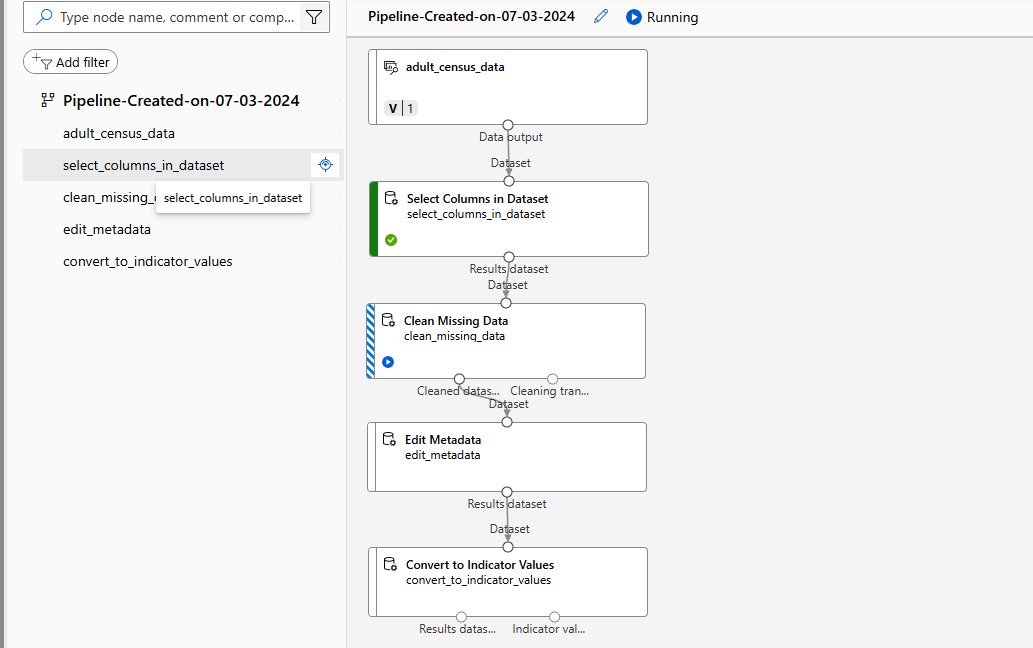
#### 중간 점검

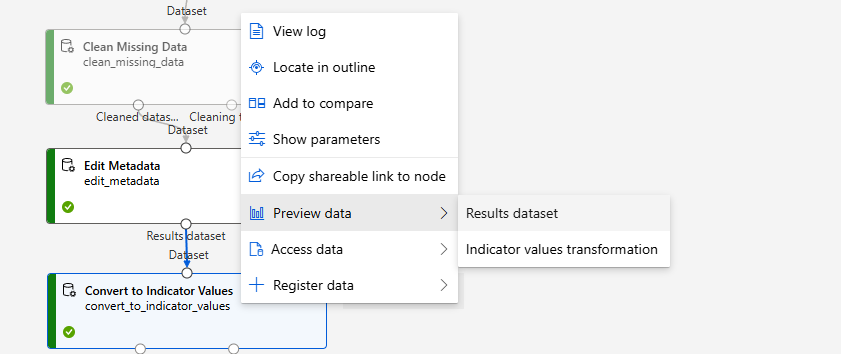
Save 후 Configure & Submit 클릭



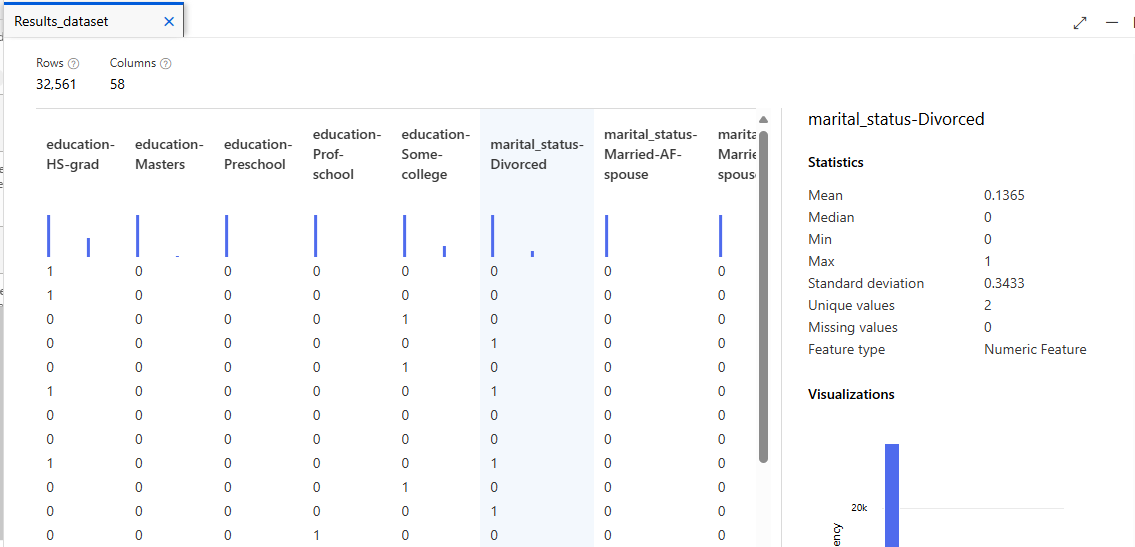


Jobs에서 캔버스에서 작성 한 파이프라인 실행 상황 확인

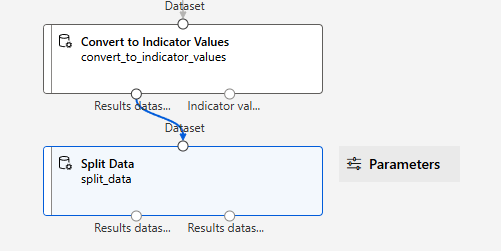




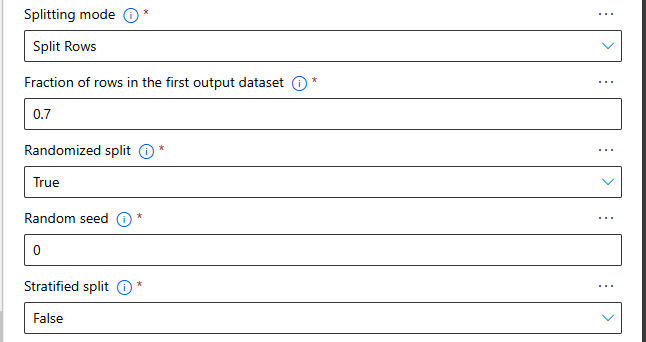
범주형 데이터가 많아 컬럼수가 많아진 것을 확인할 수 있다. (15 -> 58)



#### 데이터 분리 - 학습 데이터와 테스트 데이터로 분리

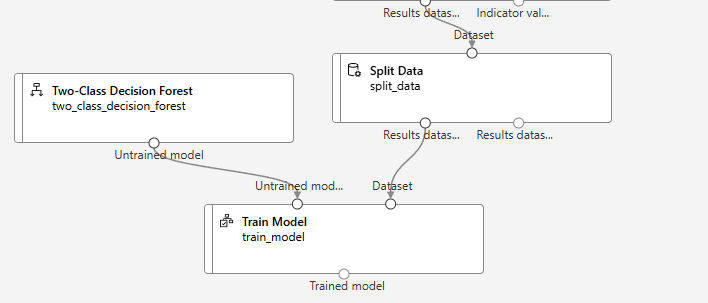


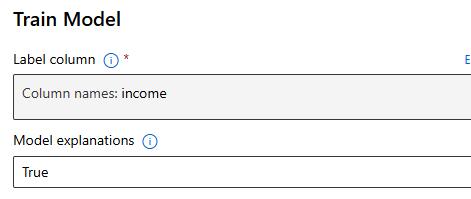
학습 데이터 70 / 테스트 데이터 30 으로 분리



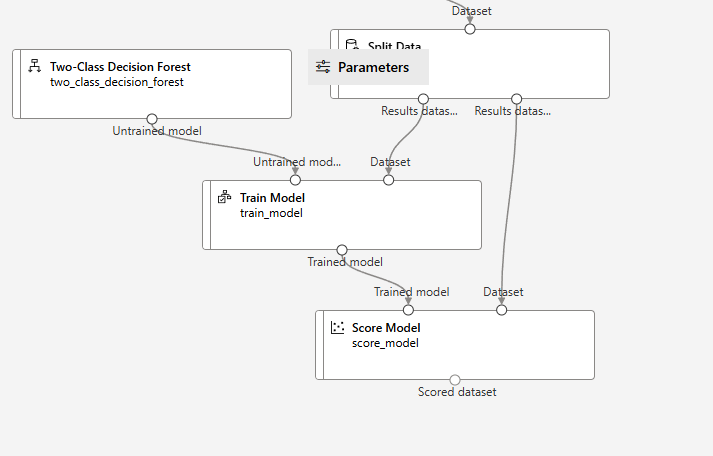
### 모델링/평가

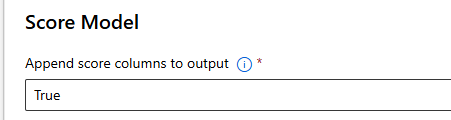
#### 모델 학습(훈련)



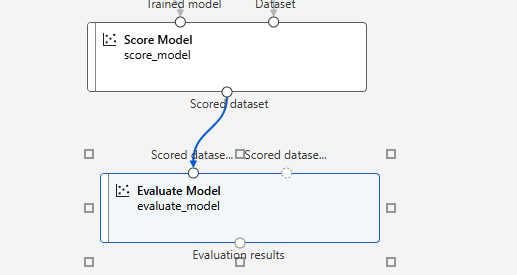


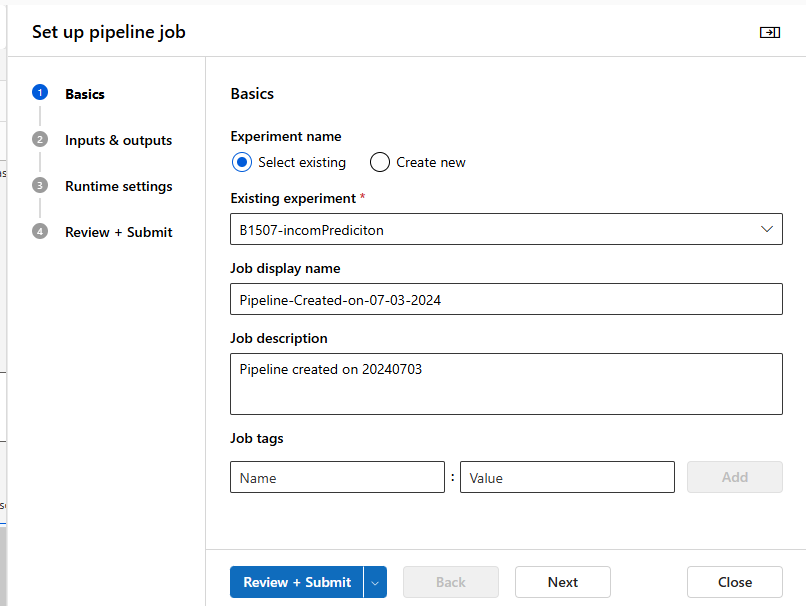
#### 모델 테스트



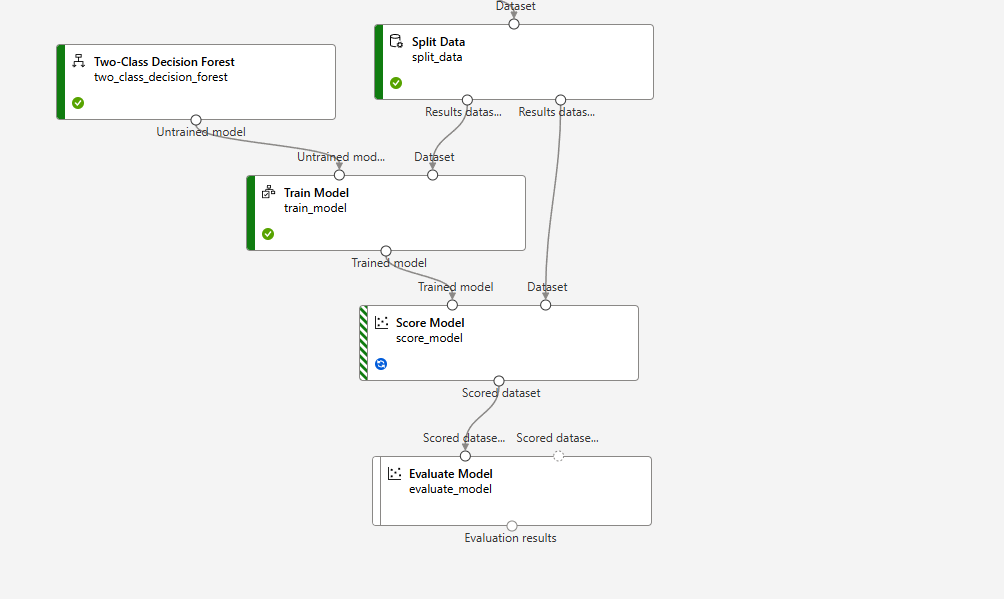


#### 모델 평가(테스트 결과 평가)



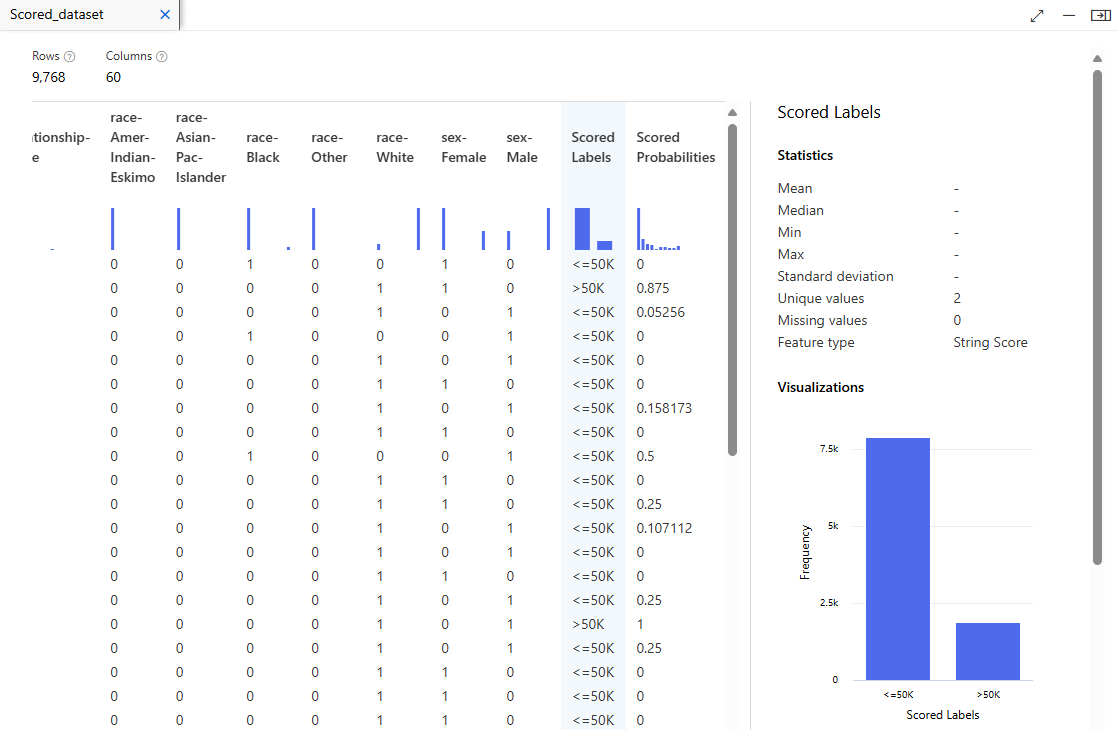


Jobs에서 확인하기

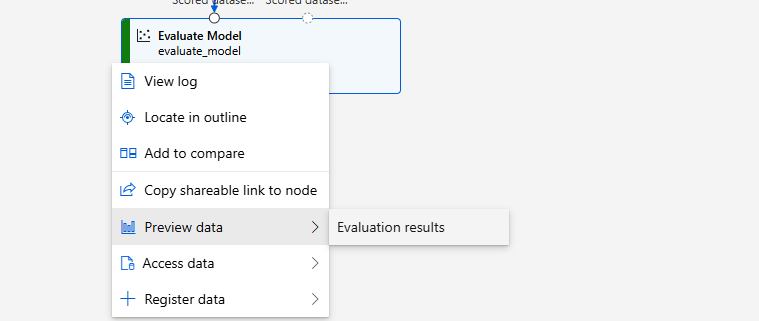


로켓 발사 가능 여부와 예측 확률이 표시됨

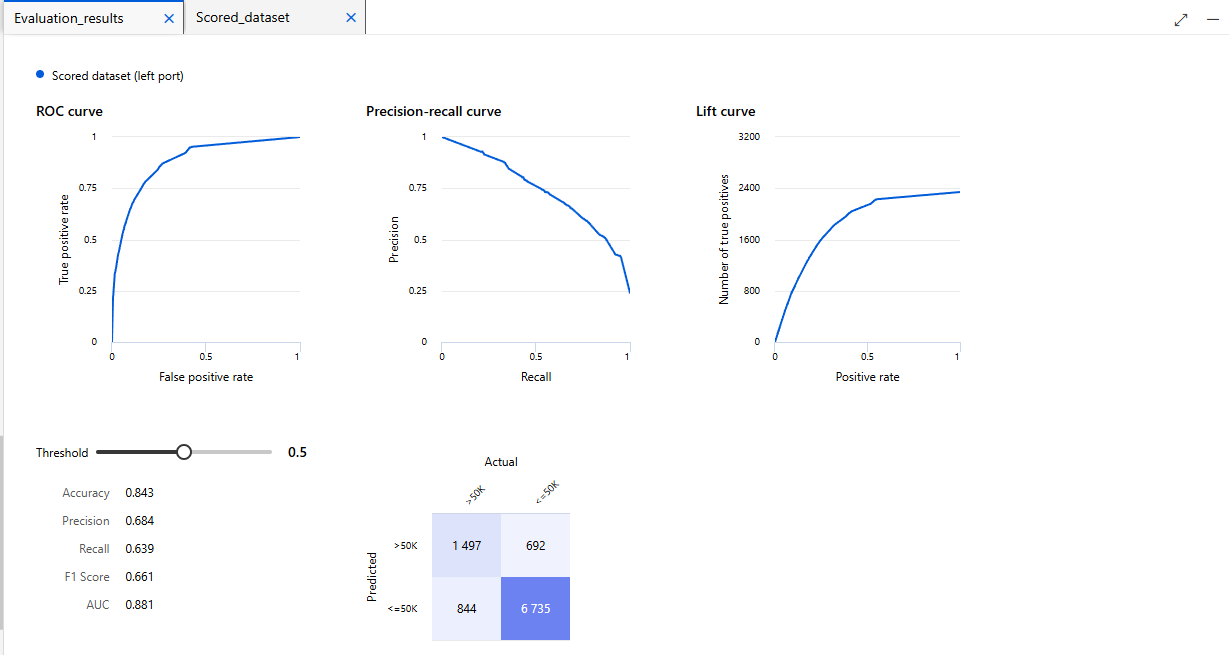


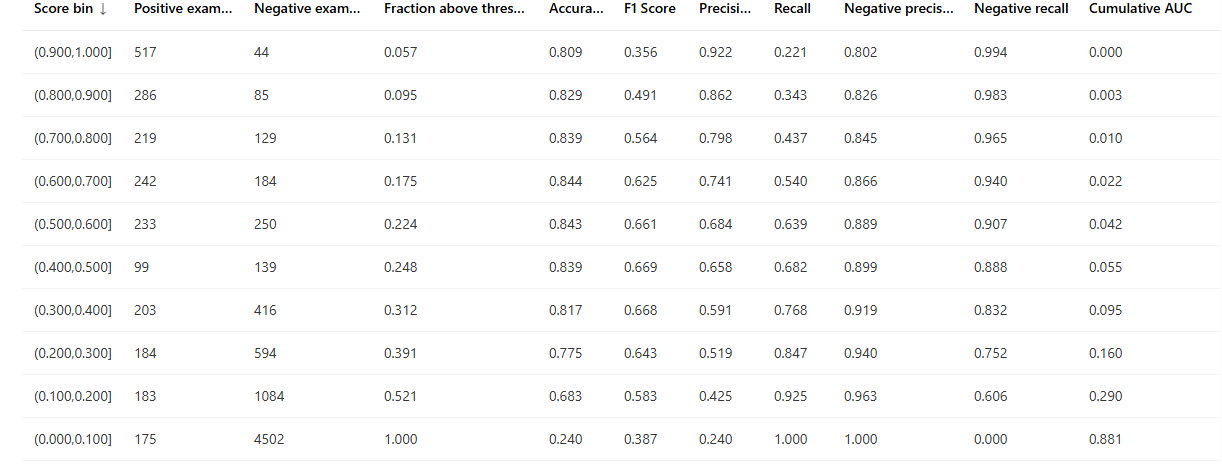


정확도 (Accuracy)를 포함한 다양한 모델 평가 결과가 표시됨



• 임계값(Threshold) : 특정 확률 이상의 예측값을 Positive로 예측





## [실습 - 회귀MLD\_자동차가격예측\_선형회귀]

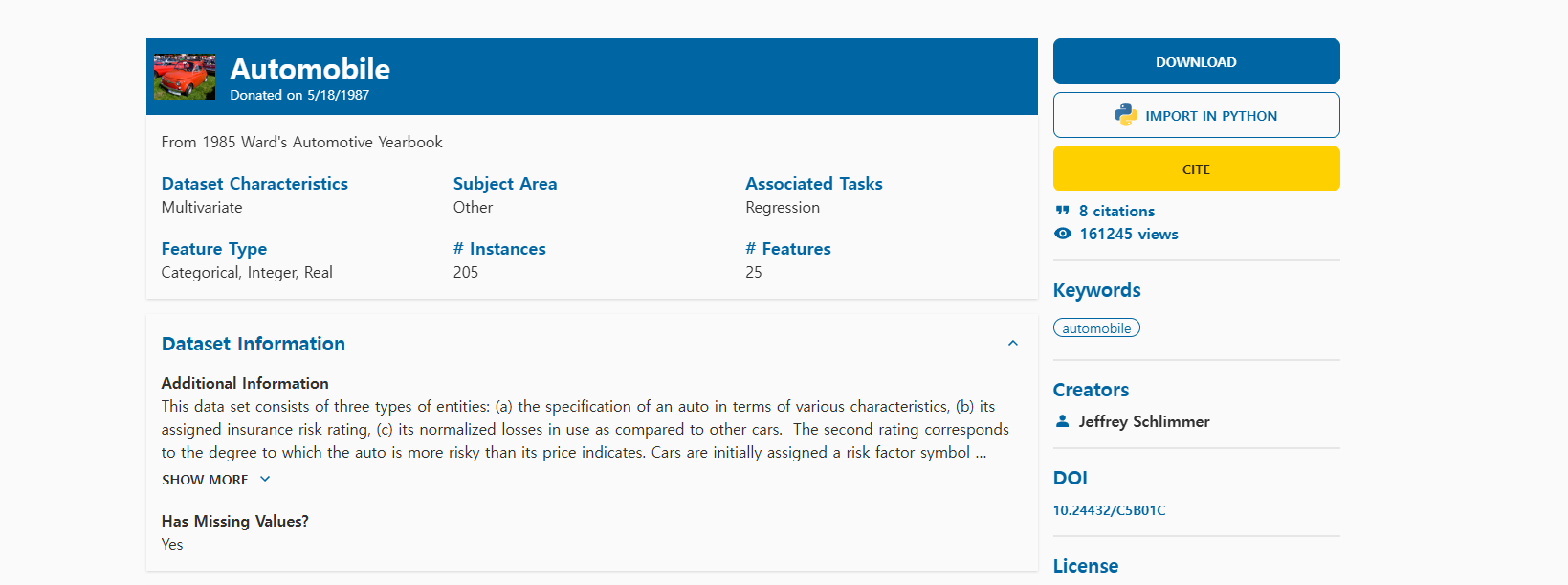
MS Azure ML Designer를 활용한 회귀 모델

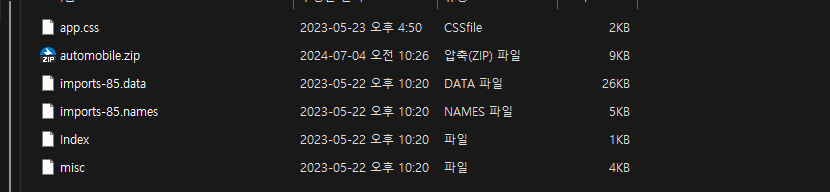
(Lab : 선형회귀를 이용한 자동차 가격 예측 모델 구현)

### 실습 준비

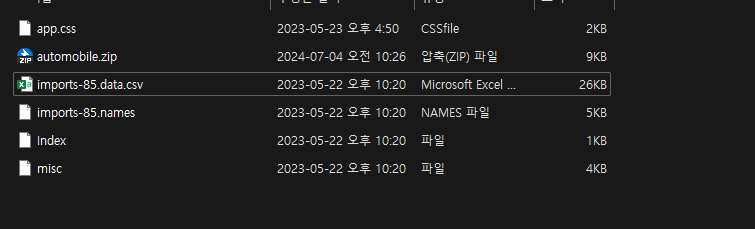
#### 데이터 수집

UCI Repository의 Automobile 데이터세트 이용 (<https://archive.ics.uci.edu/dataset/10/automobile>)

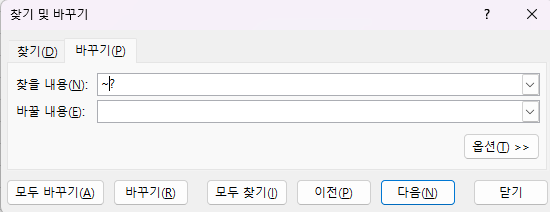


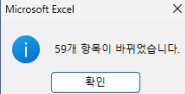


Import-85.data 파일의 확장자를 .csv로 변경



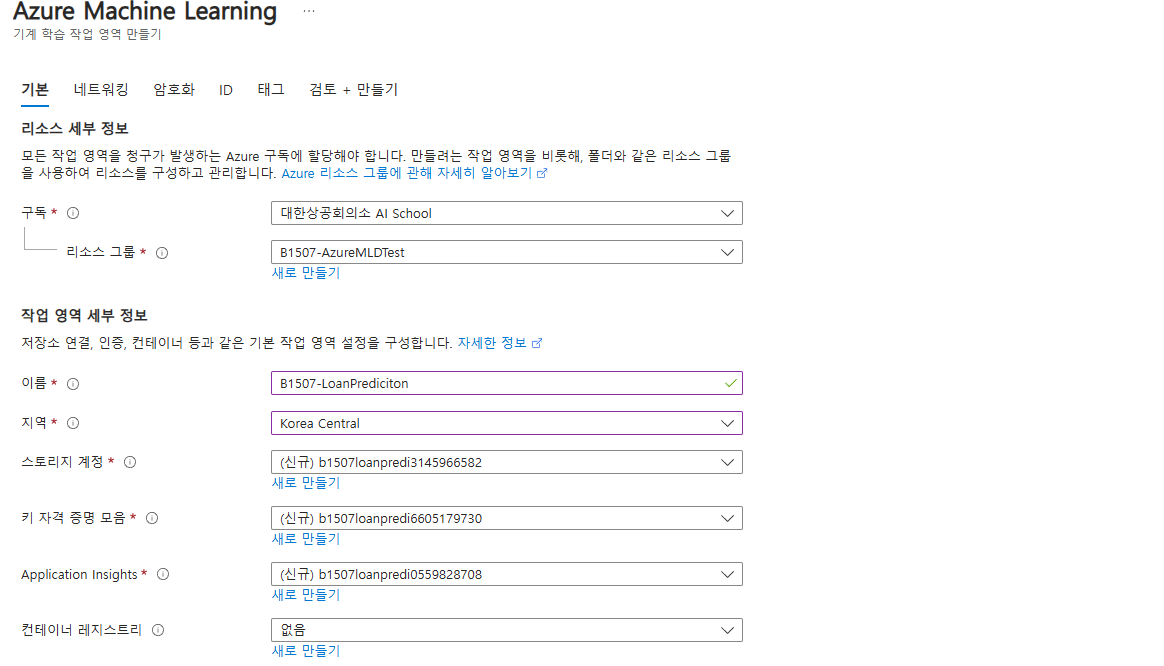
Imports-85.data.csv 파일의 누락값 ? 처리



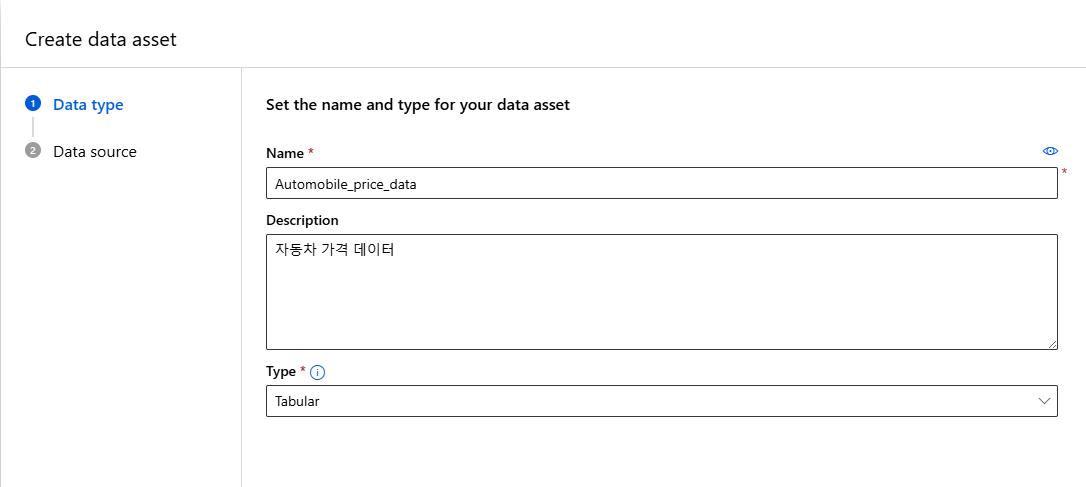


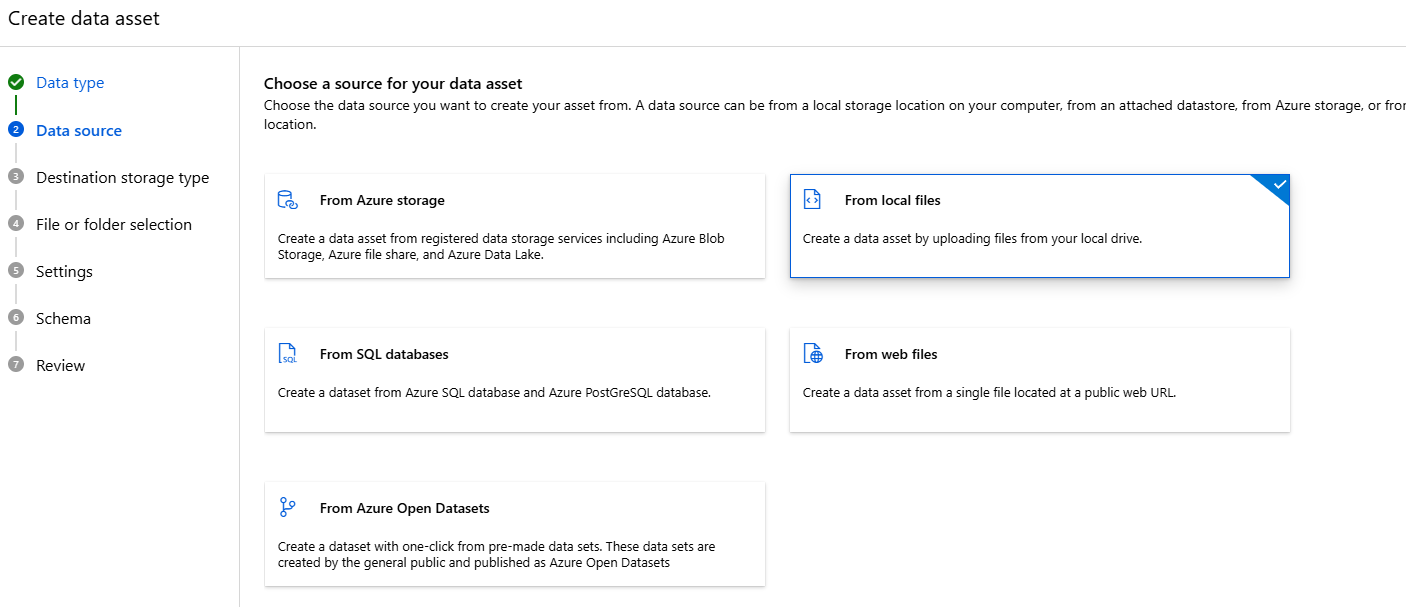
#### Azure Portal 접속 후 머신러닝 리소스 생성

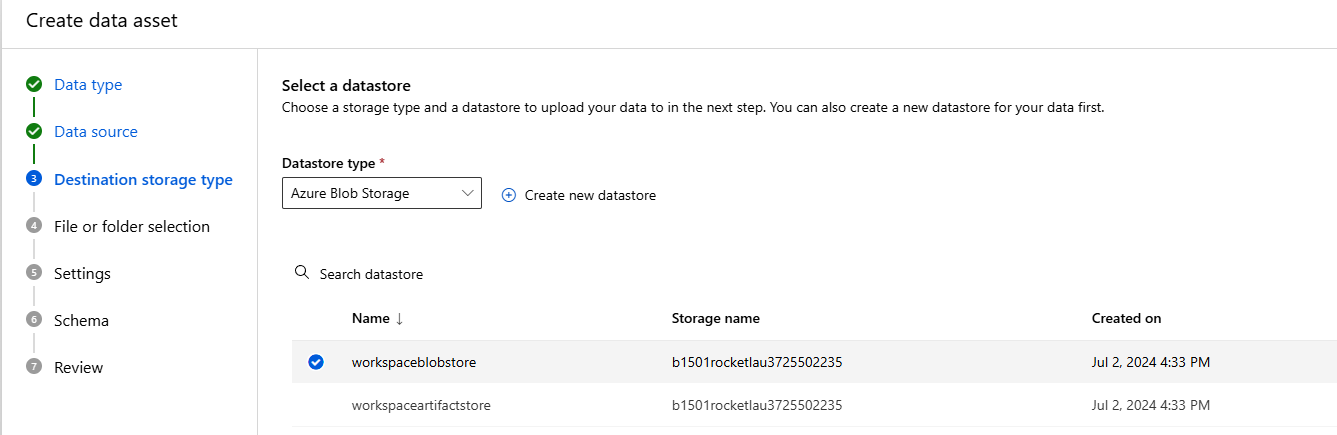
과정은 생략합니다.

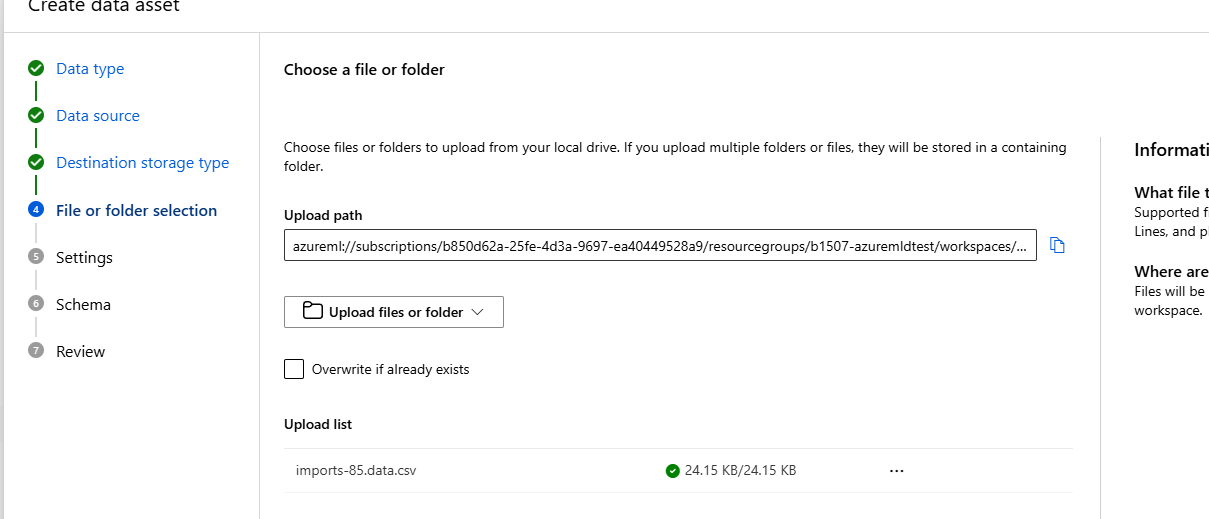


#### 데이터 세트 등록



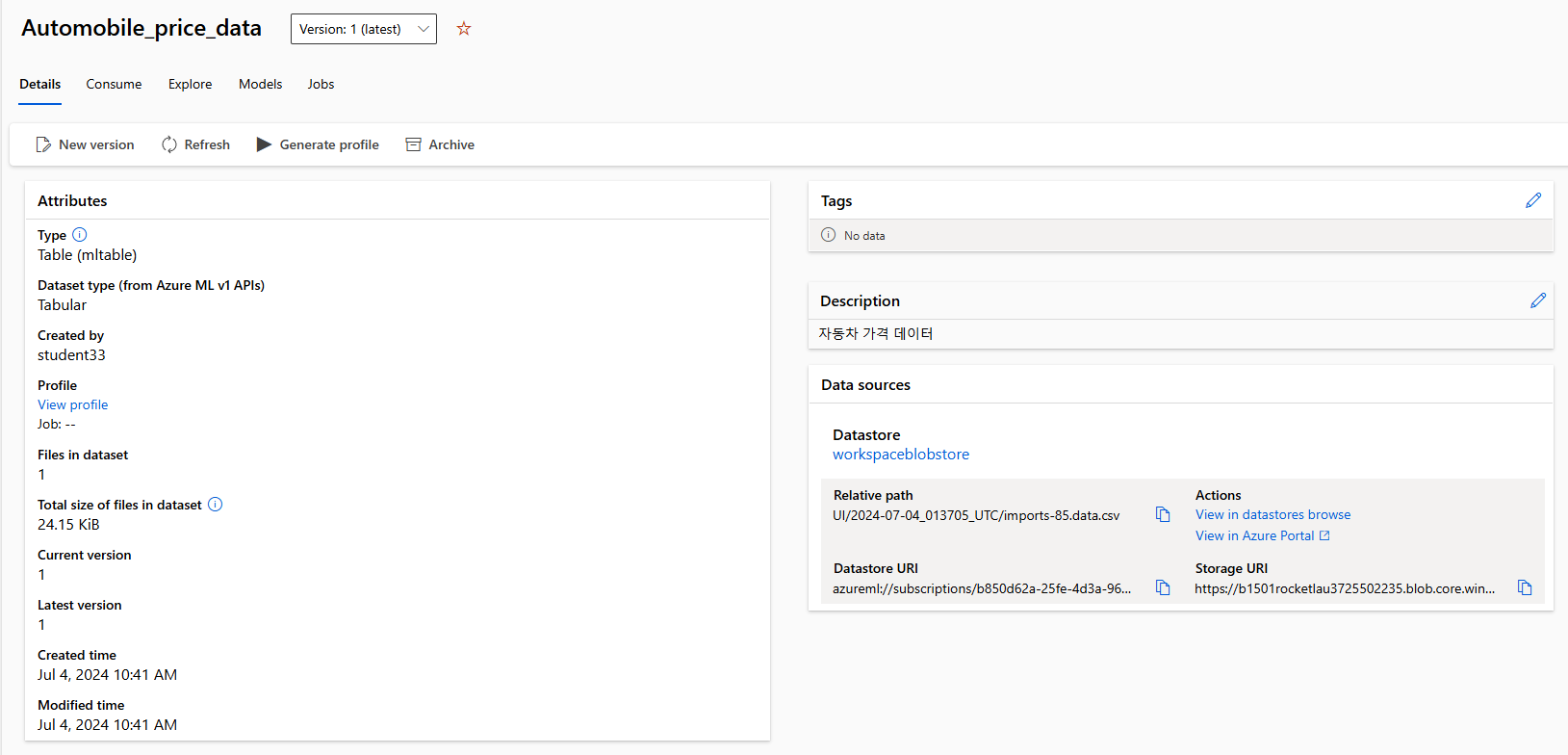


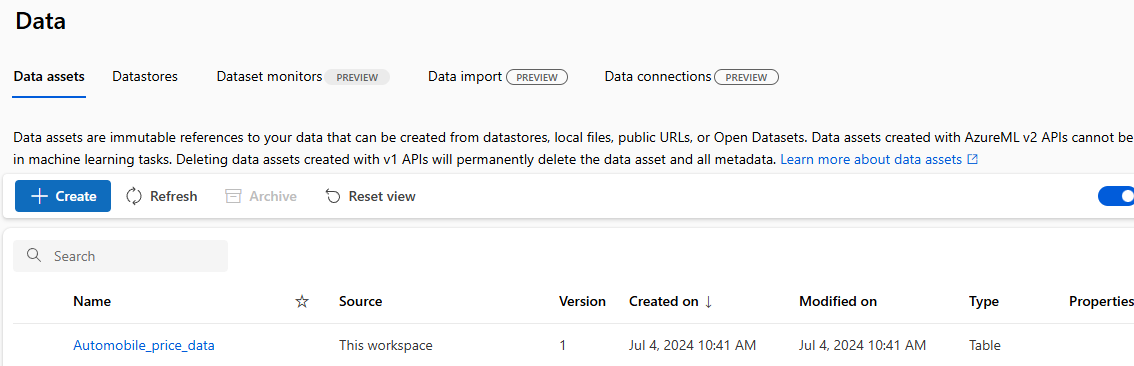




Column headers 항목을 No header로 변경



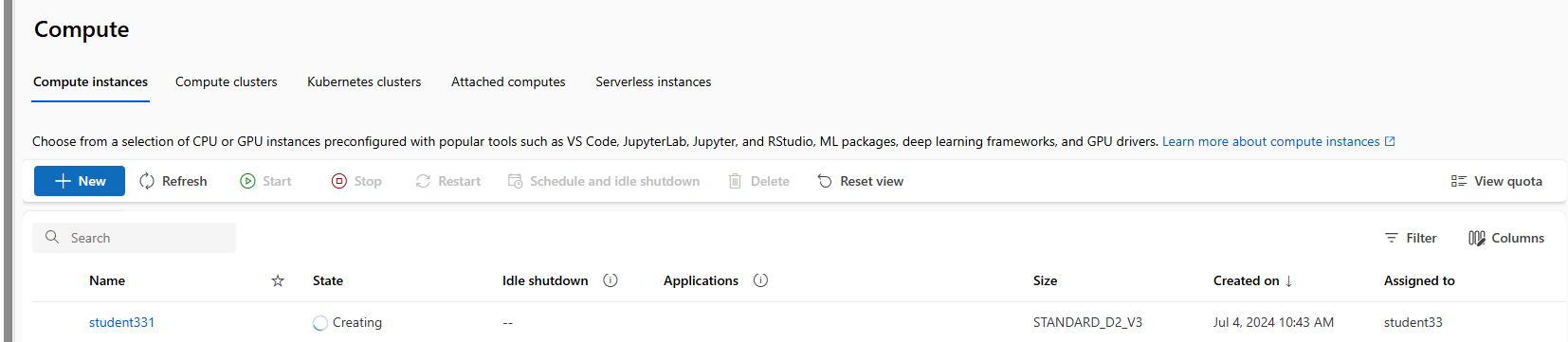




#### 컴퓨트 대상 설정

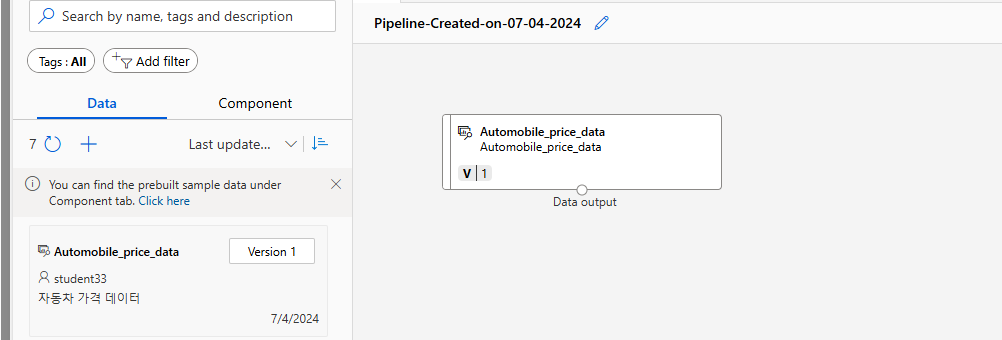
Compute instances 환경 사용



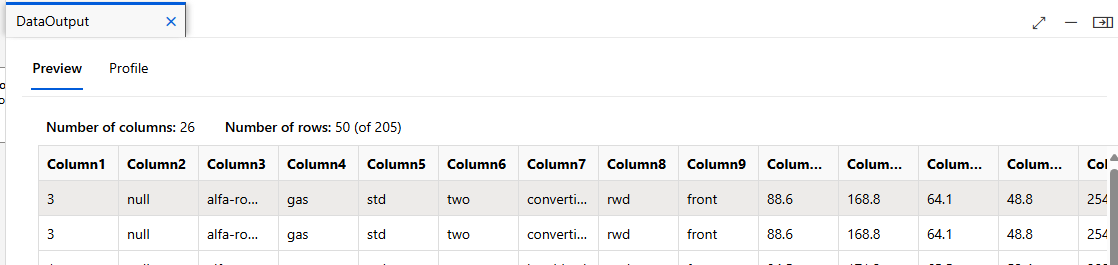


### 데이터 수집/이해

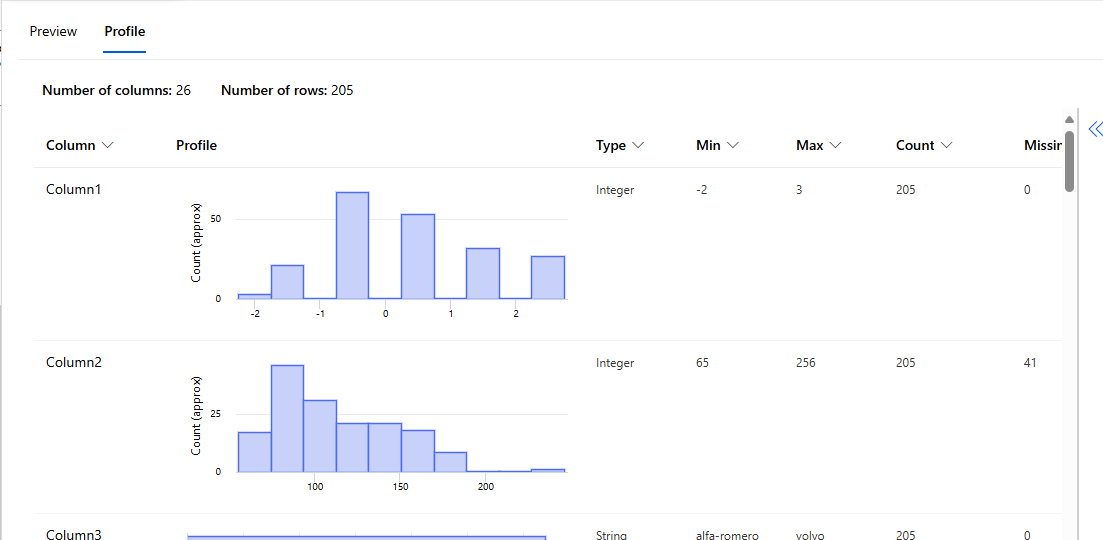
#### 디자이너 생성 후 데이터 세트 가져오기



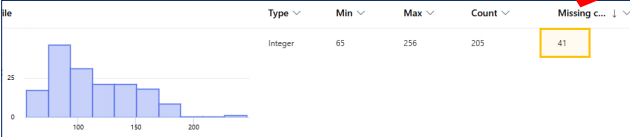
#### 데이터 이해



데이터 세트의 컬럼 수 및 데이터 샘플의 개수 확인 가능

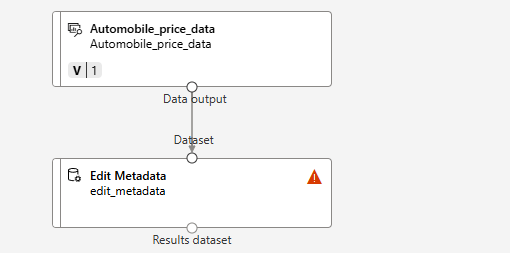


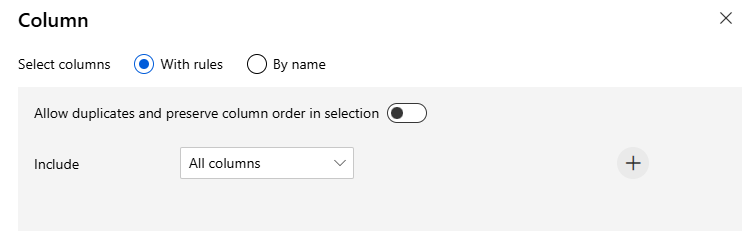
누락값 등 확인이 가능

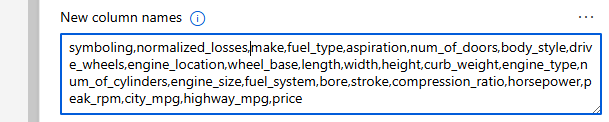


### 데이터 준비

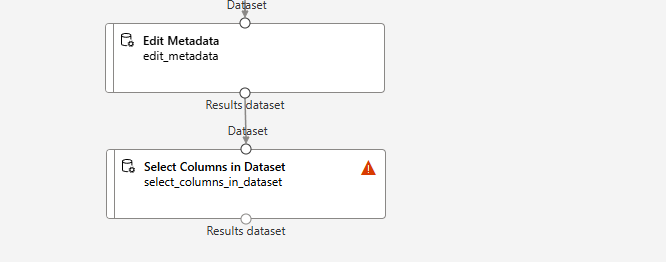
#### 특성 이름 지정



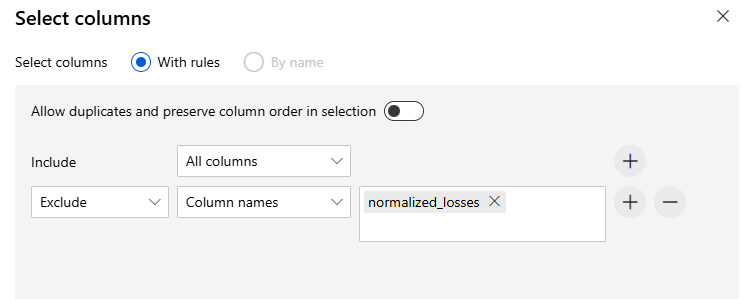




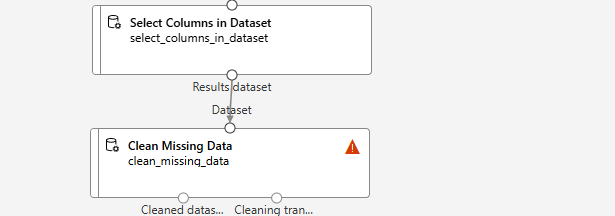
#### 특성 선택 – 불필요한 컬럼 제외

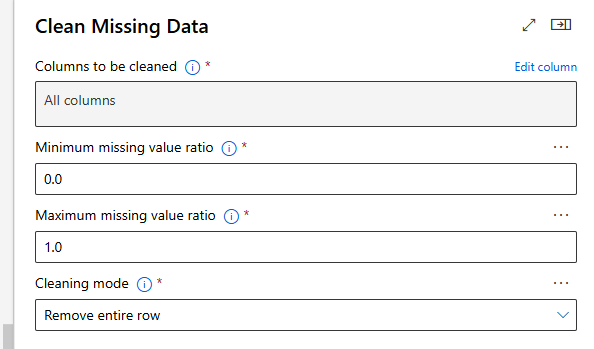


With rules 방식으로 All columns를 포함시키고, normalized\_losses 컬럼을 제외

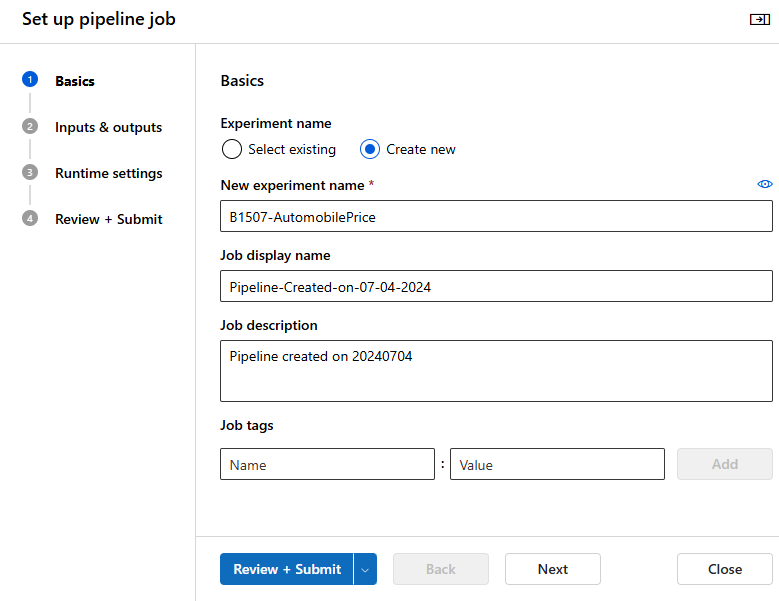


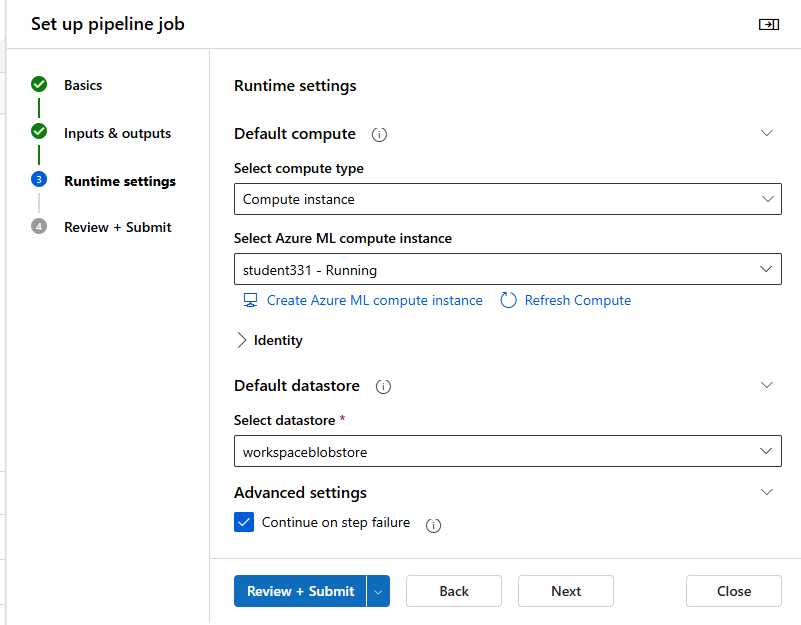
#### 누락값 처리

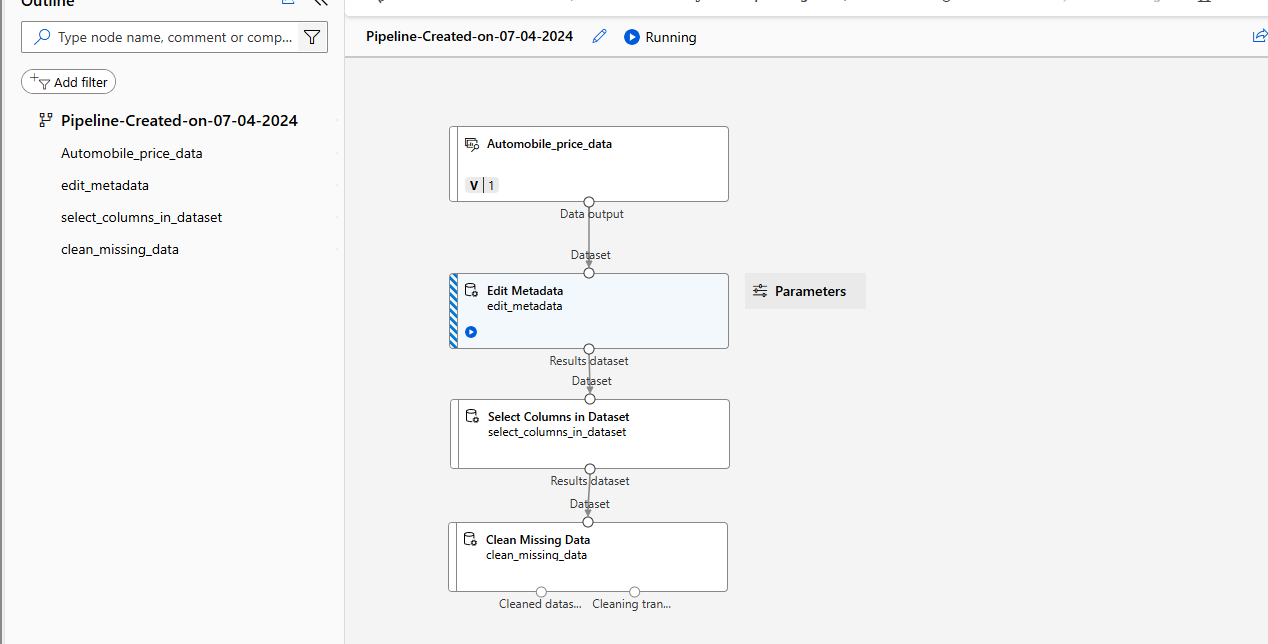


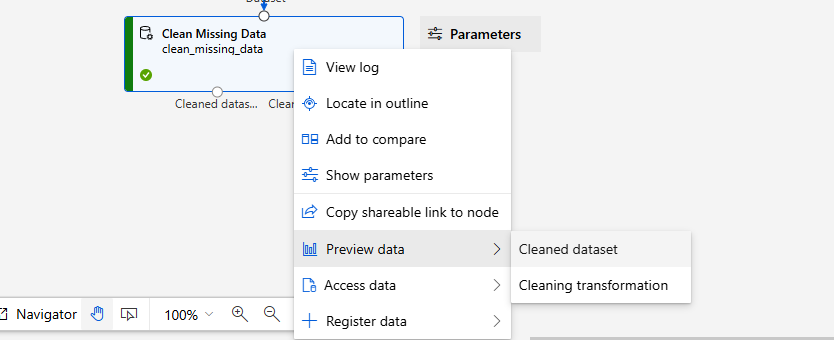


#### 중간 점검

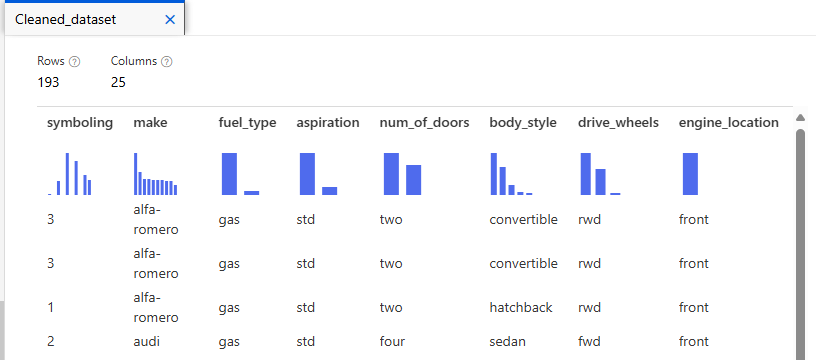




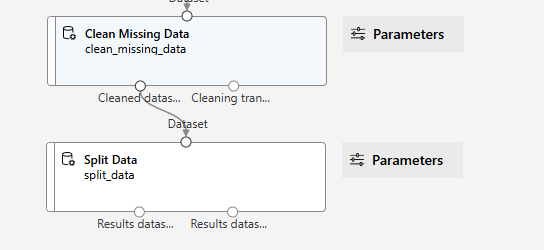




26, 205 -> 25, 193 으로 변한 모습을 확인할 수 있다.



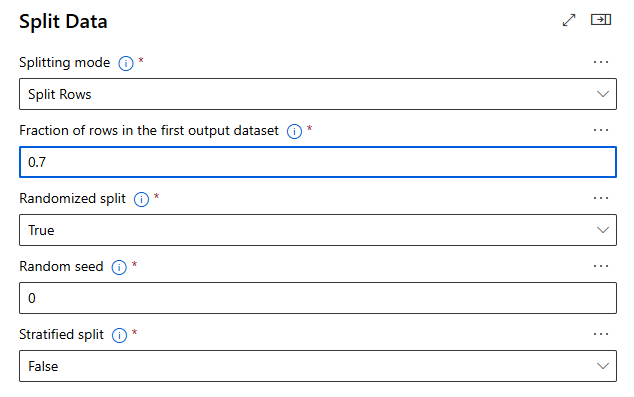
#### 데이터 분리 – 학습 데이터와 테스트 데이터로 분리



학습 데이터 70%, 테스트 데이터 30% 분리

Randomized split : 행 기준으로 무작위로 분리

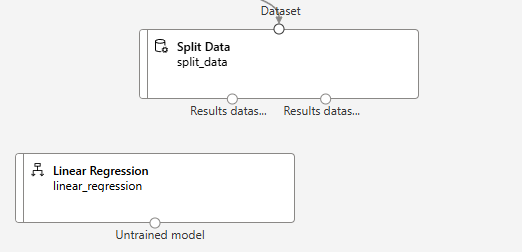
Random seed : 난수 초깃값



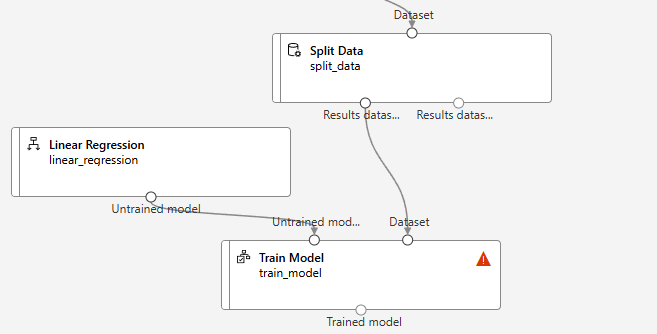
### 모델링/평가

#### 모델링 알고리즘 선택

선형회귀 알고리즘

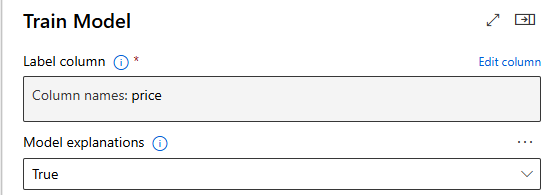


#### 모델 학습(훈련)

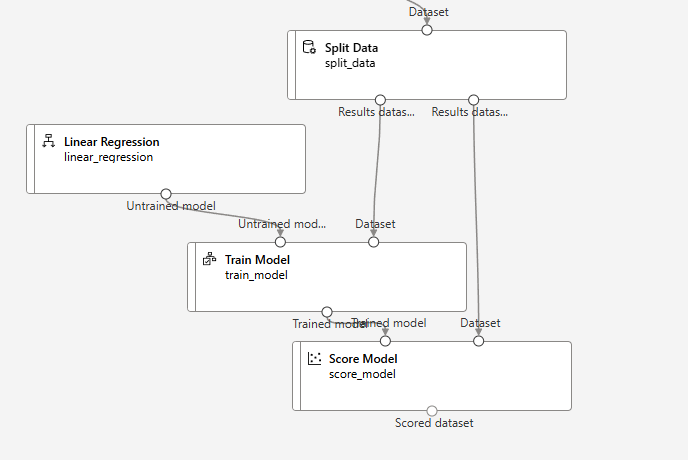


정답(라벨, Label)으로 사용할 컬럼을 지정해야 한다.

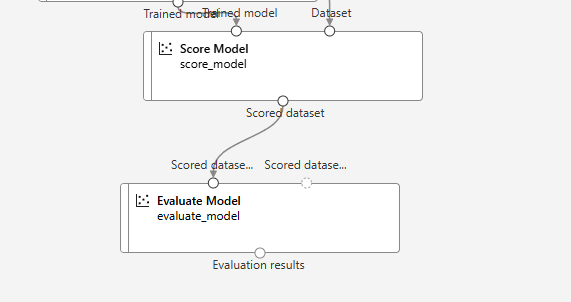
예측 결과를 나타내는 Target 컬럼을 Label 컬럼으로 지정

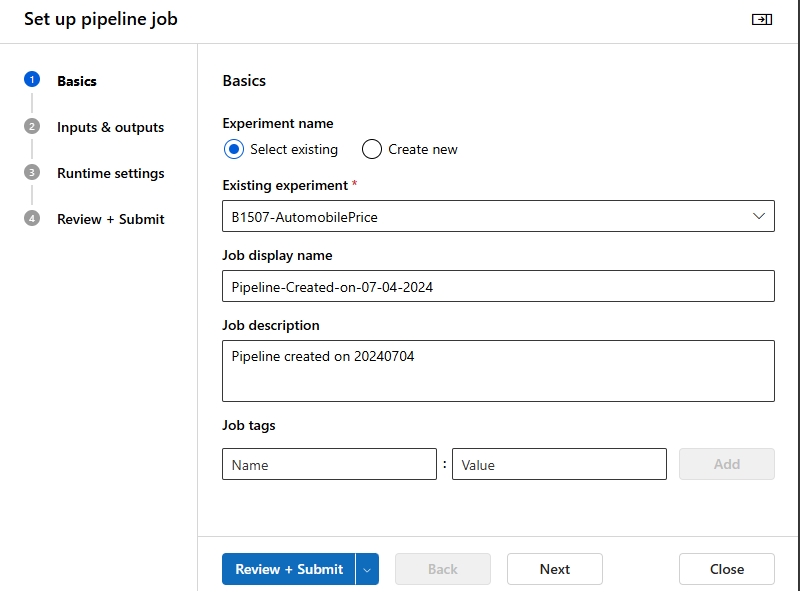


#### 모델 테스트

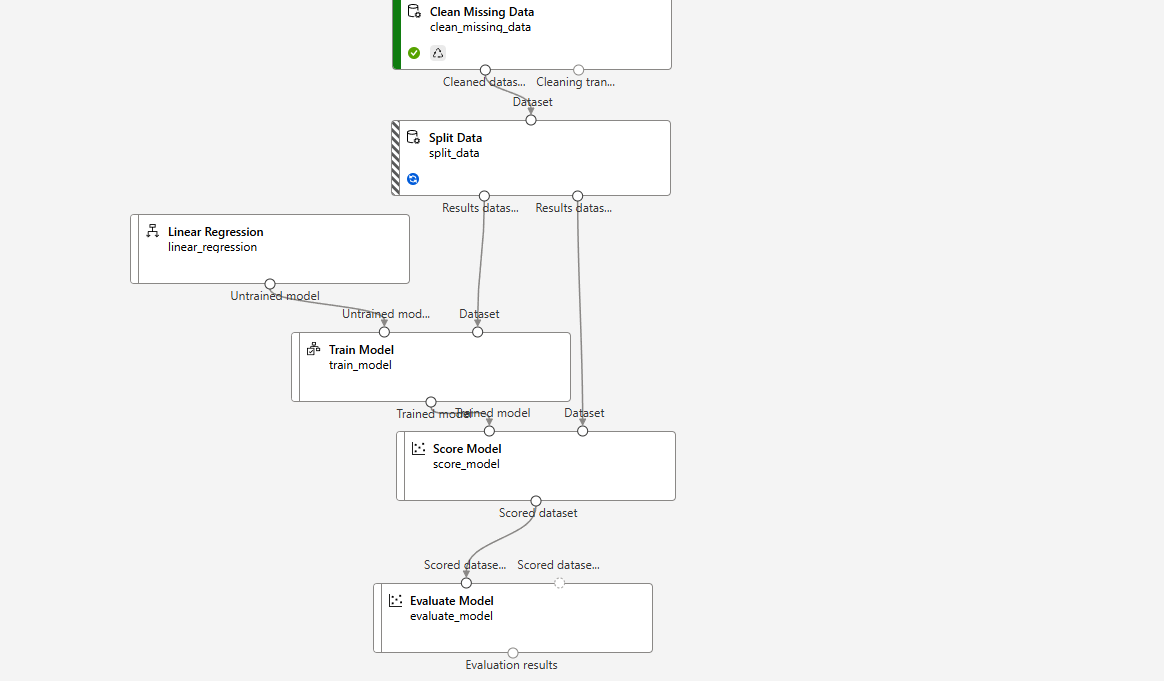


#### 모델 평가

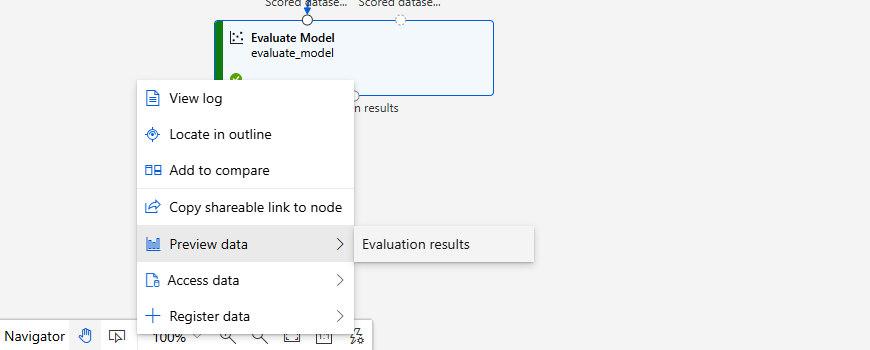




Job에서 확인이 가능하다.



#### 평가 지표 - 오차(손실)



MAE (Mean Absolute Error)

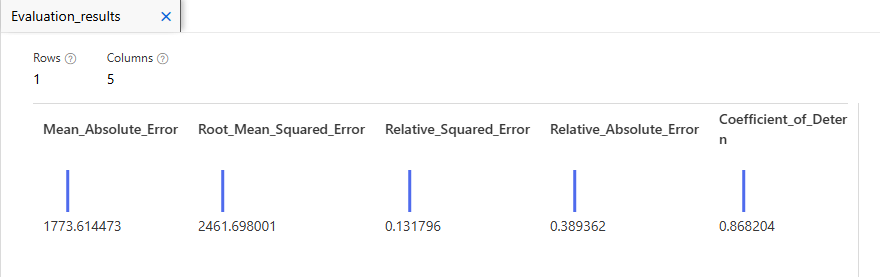
- 오차의 절대값을 평균한 값

- 독립변수의 단위 유지

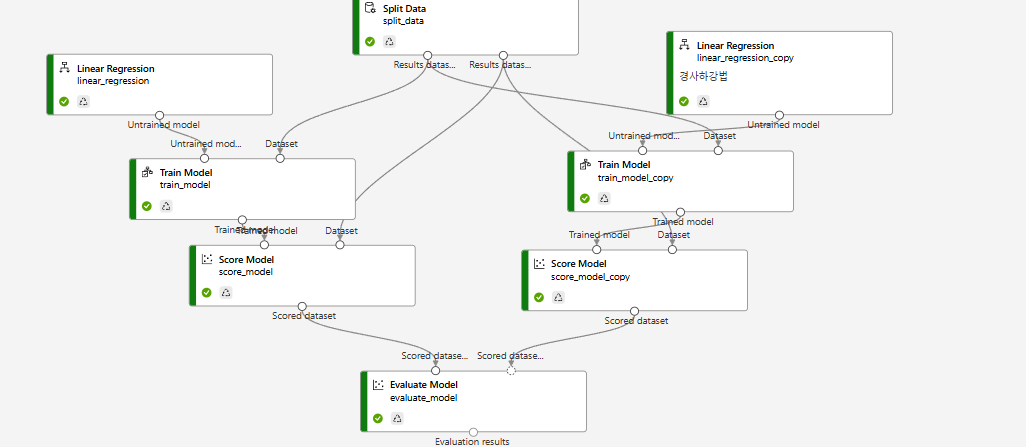
RMSE (Root Mean Squared Error)

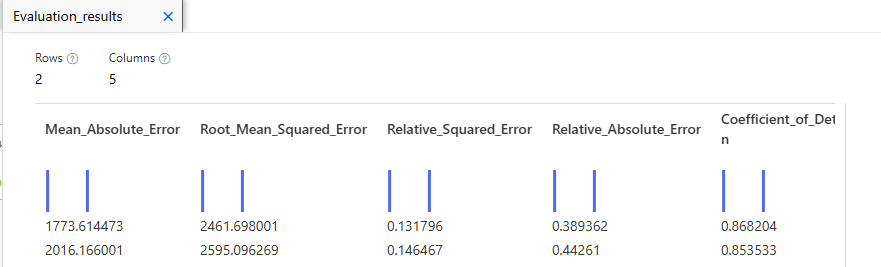
- MSE의 제곱근 : 오차의 제곱을 평균한 값의 제곱근

- 이상치에 민감함



#### 추가 – 경사하강법을 적용하여 비교하기



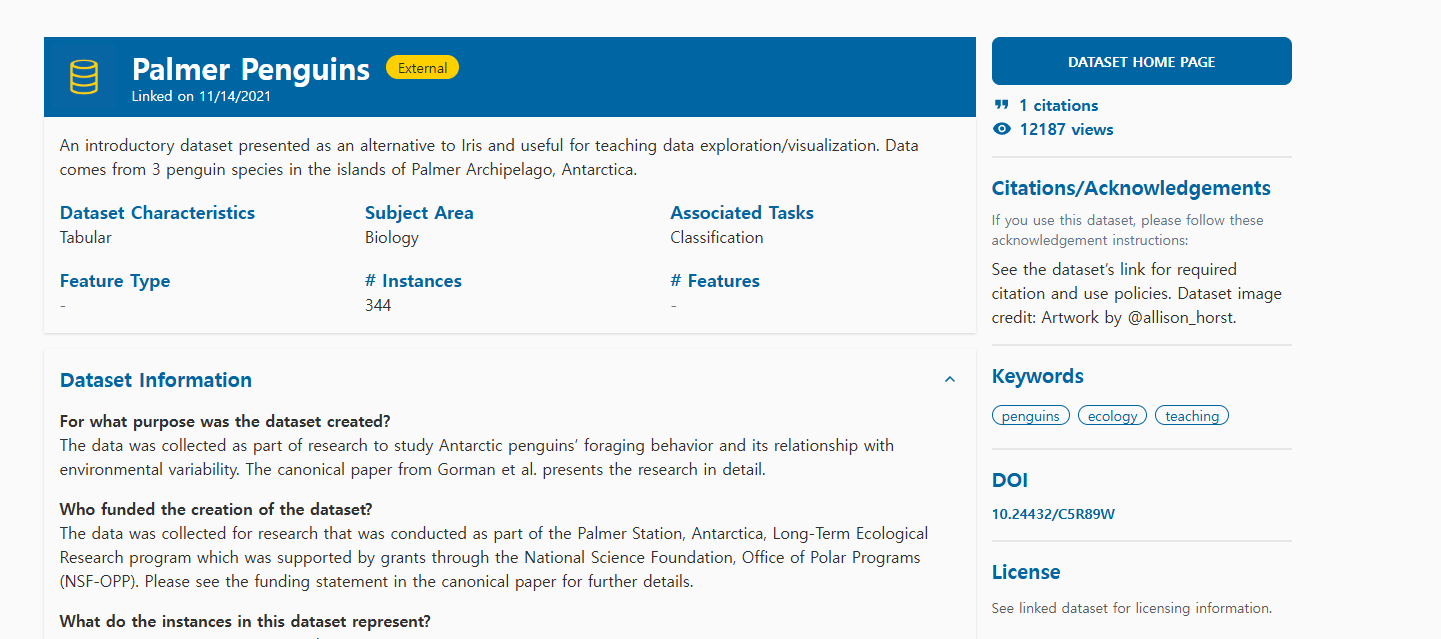


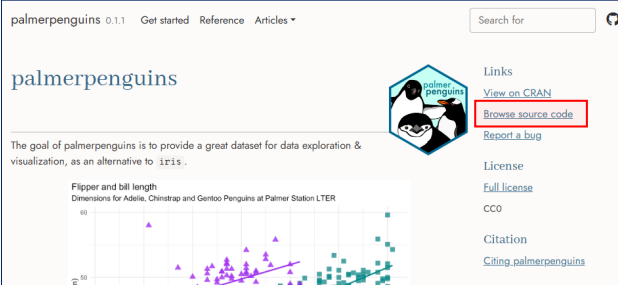
## [실습 - 분류MLD\_펭귄군집화]

### 실습 준비

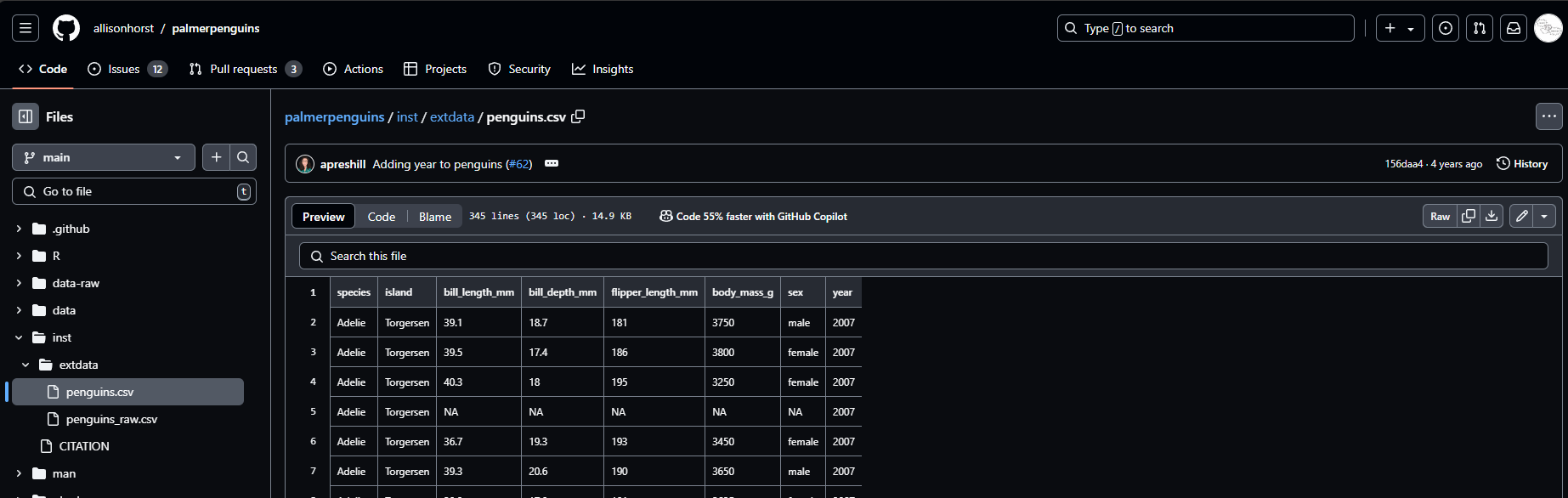
#### 데이터 수집

UCI Machine Learning Repository에 접속하여 Palmer Penguins 데이터 세트 검색

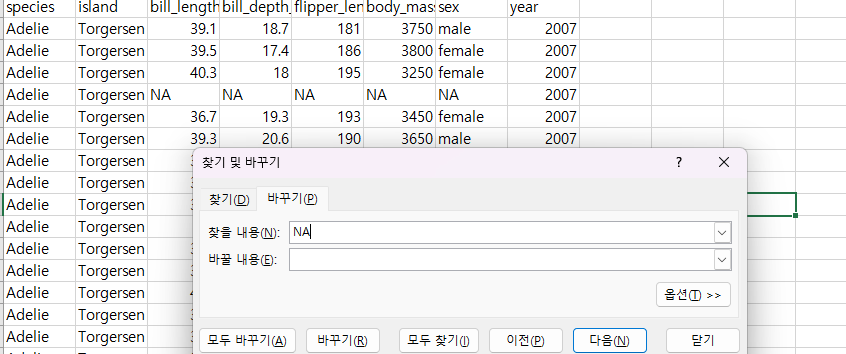


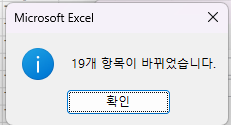


Github에서 csv 파일 다운로드



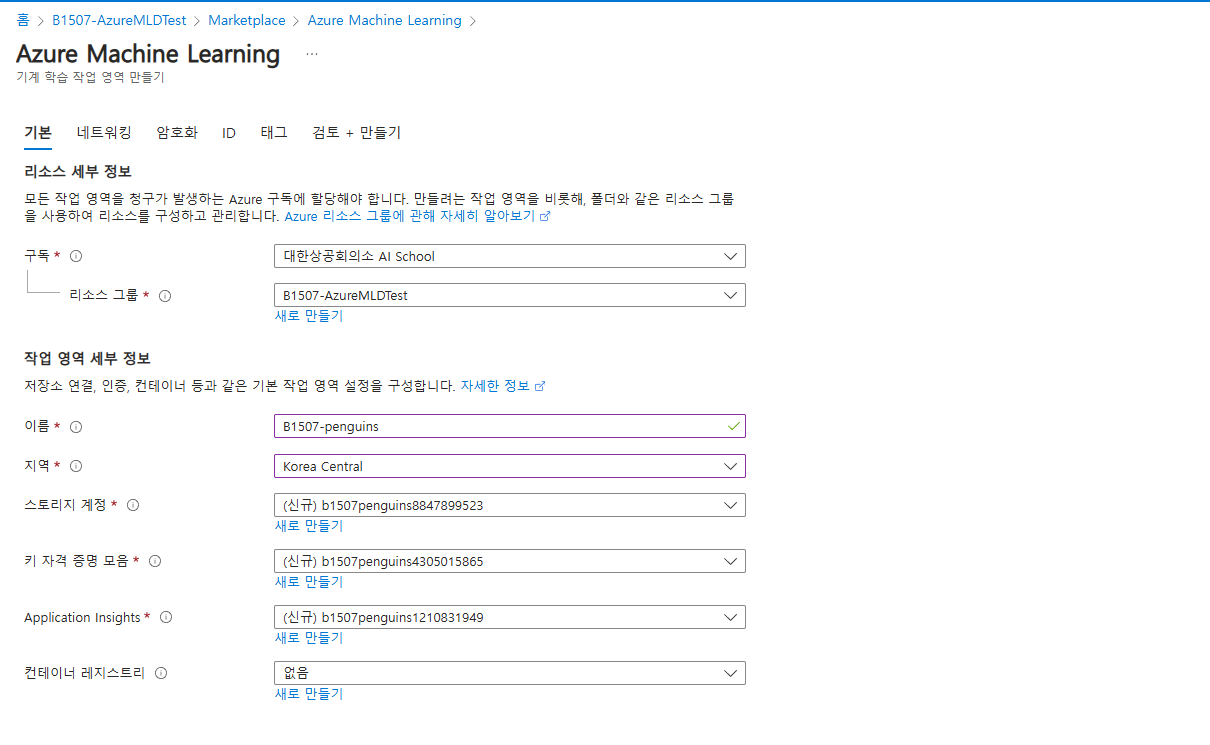
NA 값을 공백으로 바꾸기



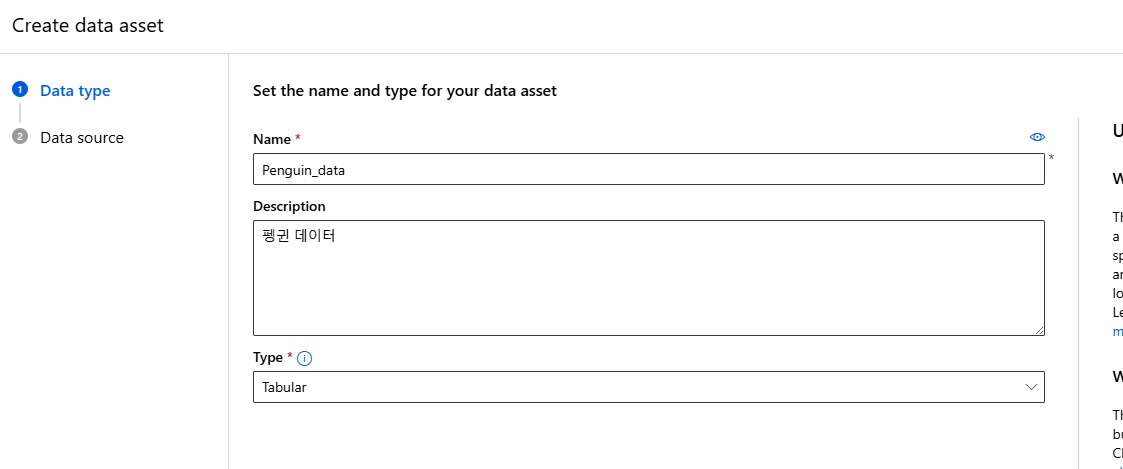


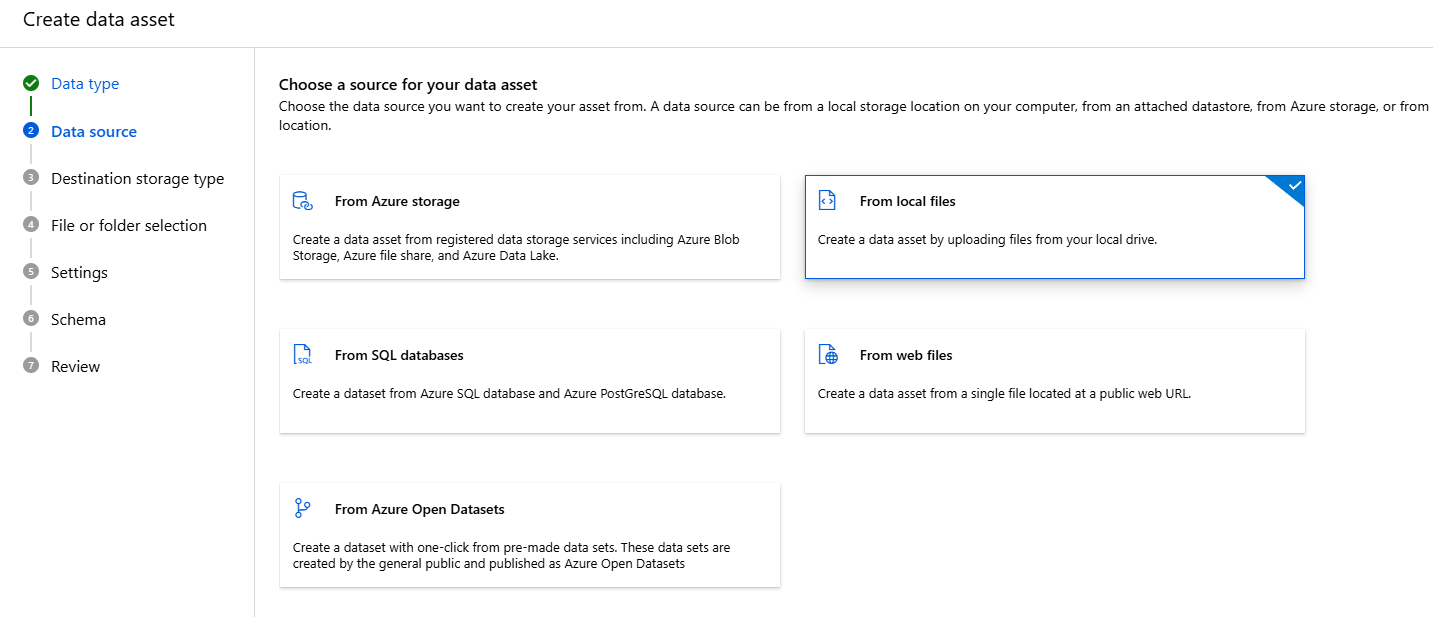
#### Azure Portal 접속 및 리소스 그룹, 머신러닝 리소스 생성

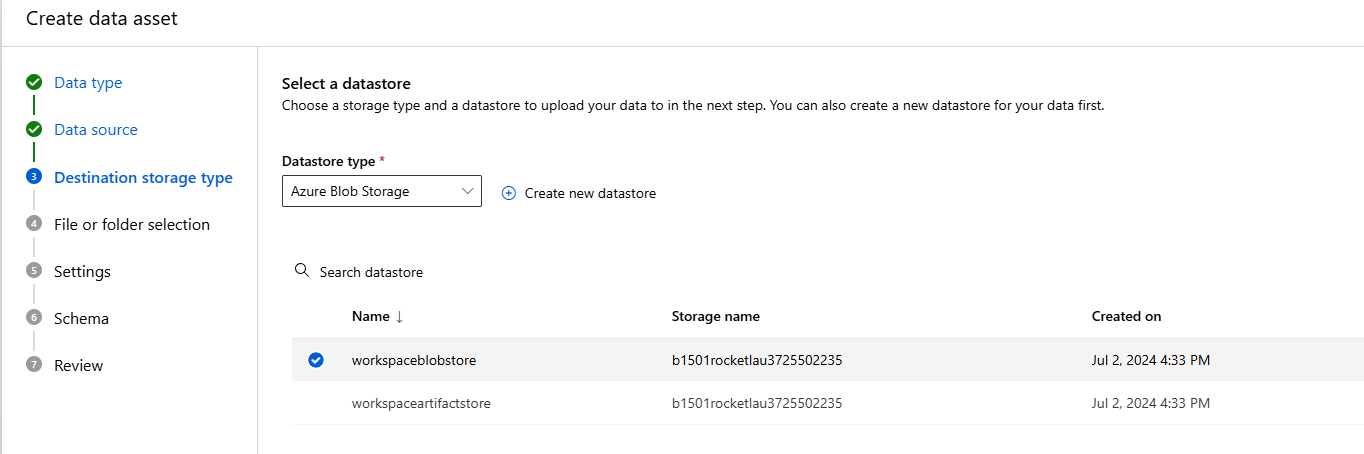
생성 전 화면만 띄우고 생략하겠습니다.

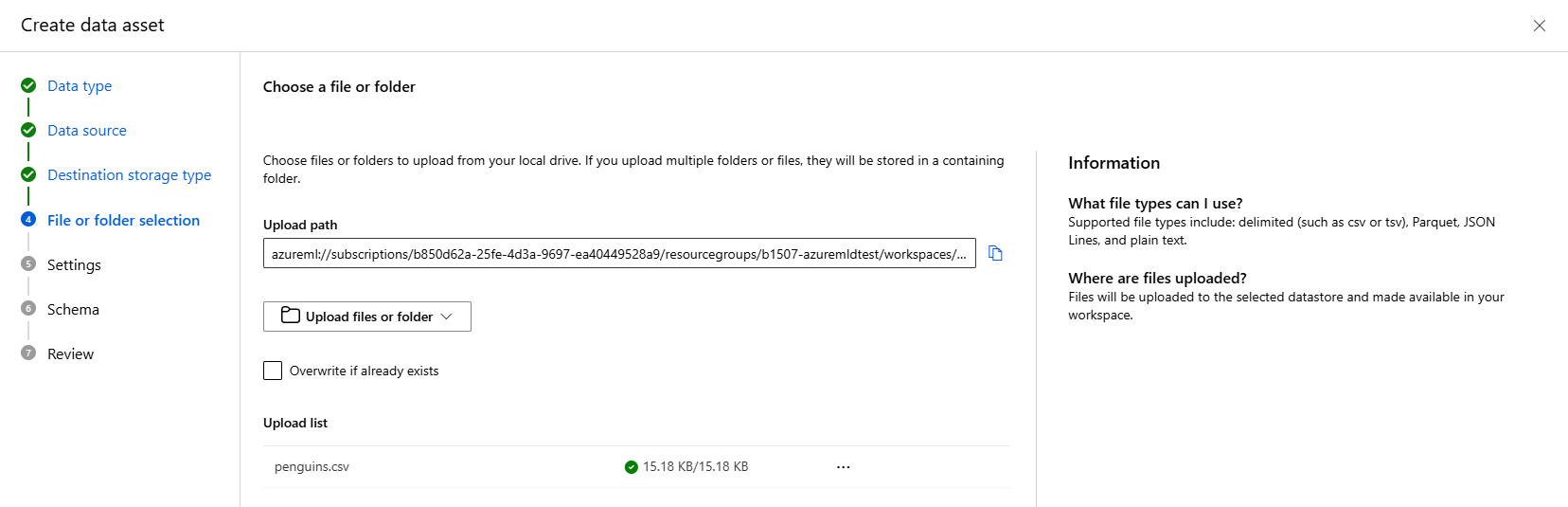


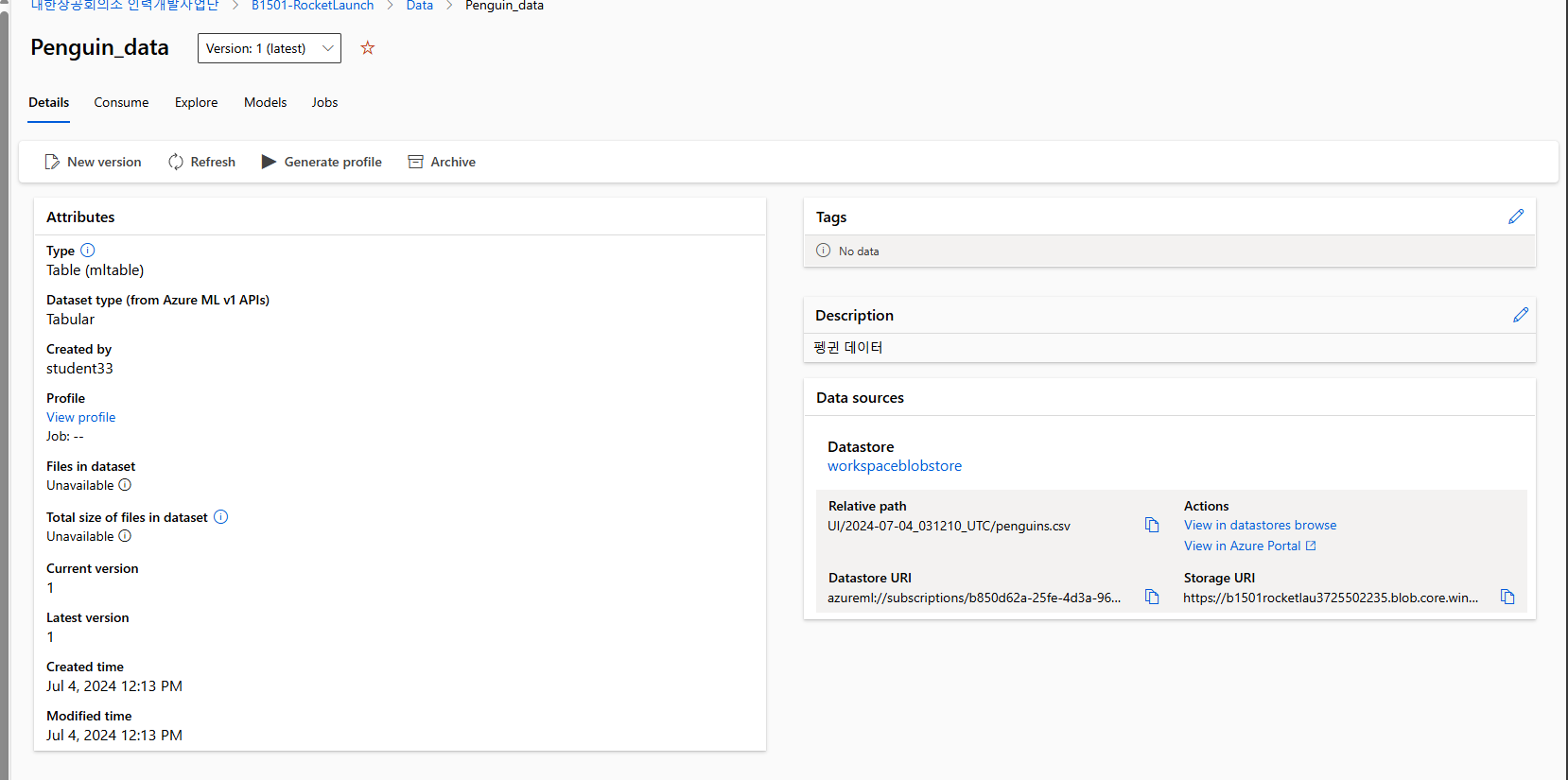
#### 데이터 세트 등록

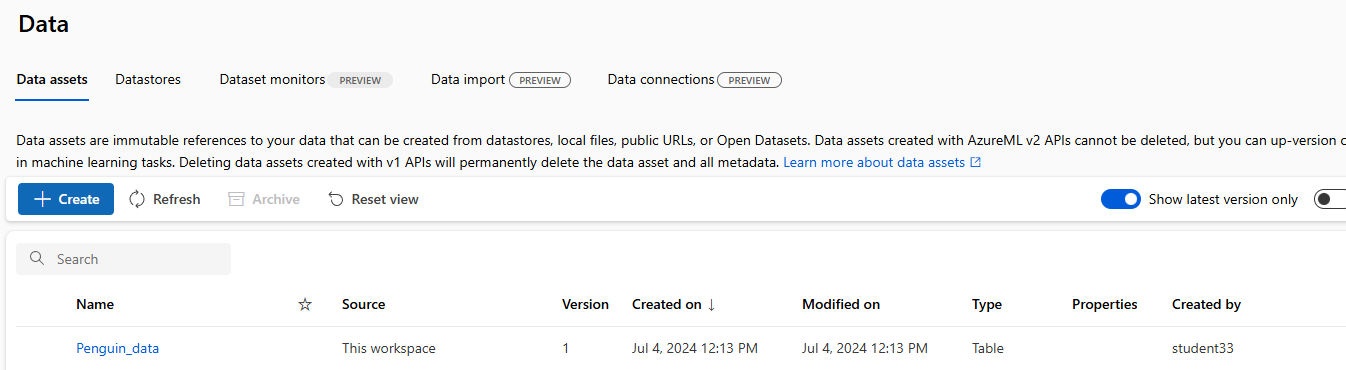




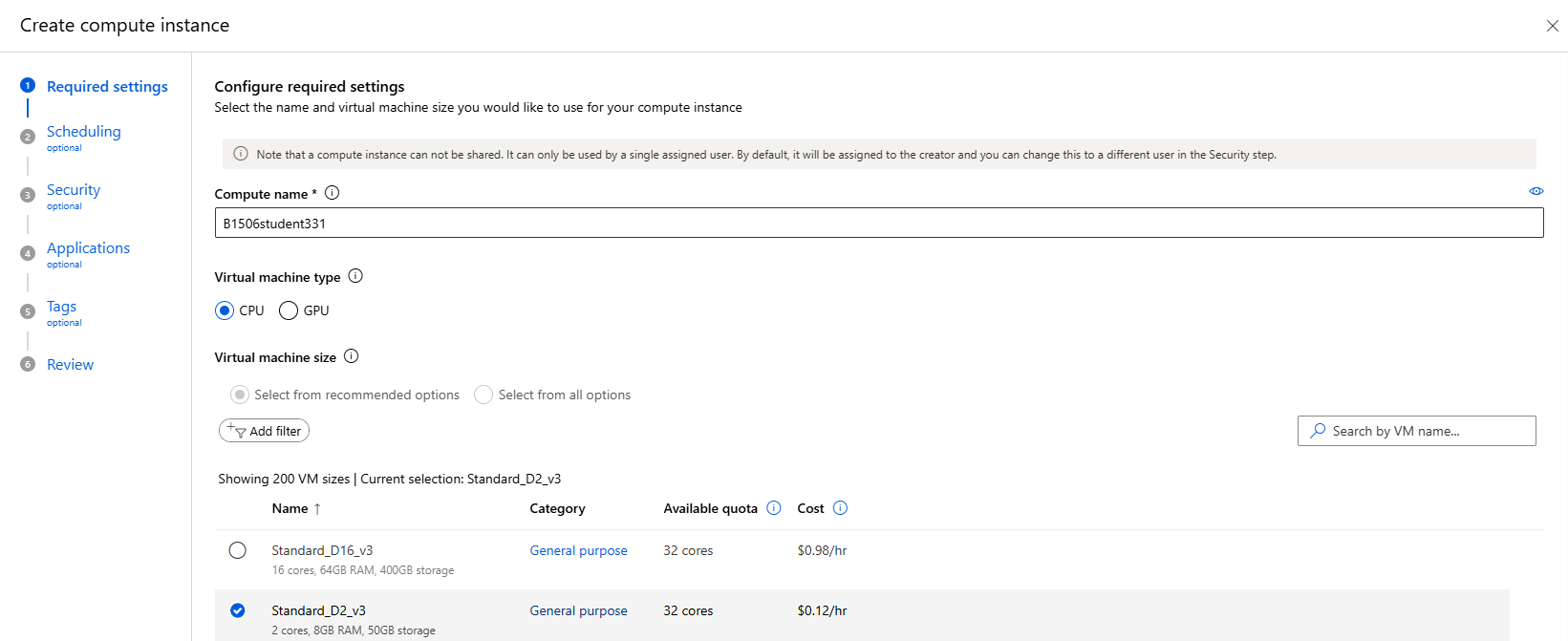


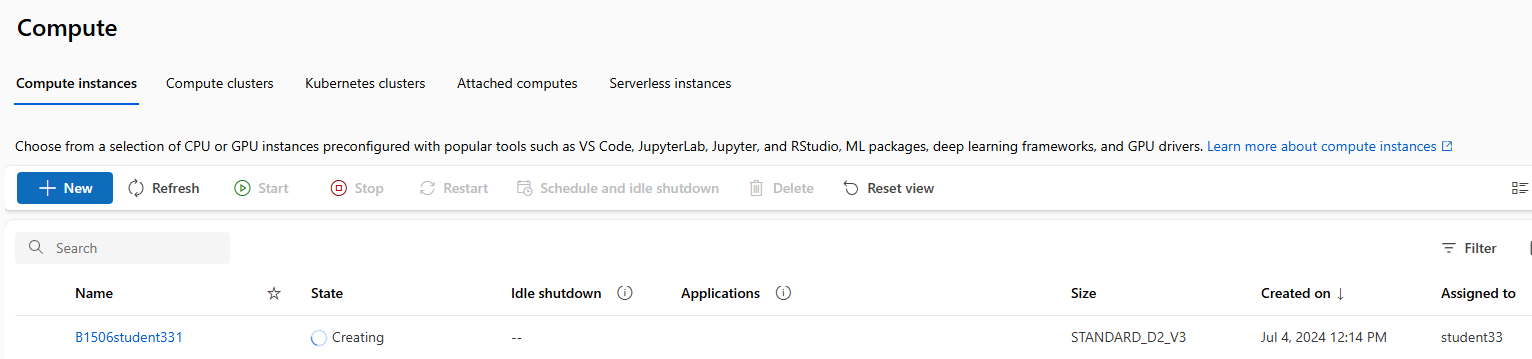






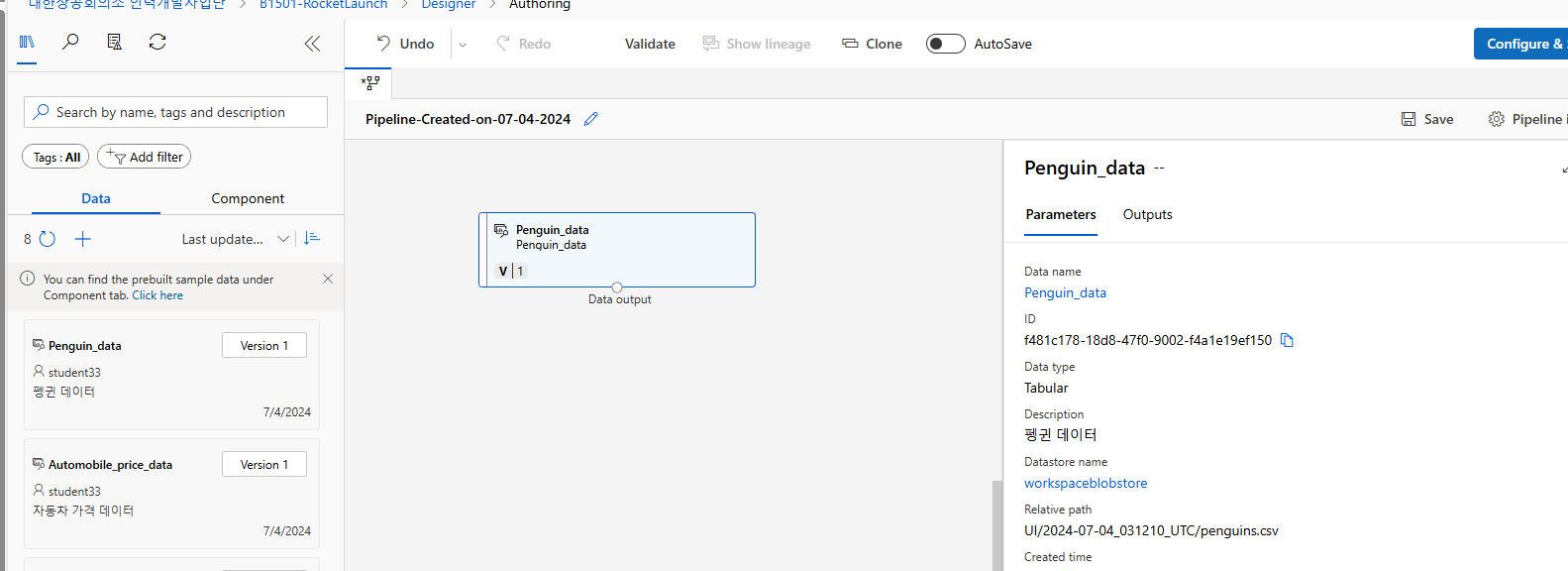
#### 컴퓨트 대상 설정





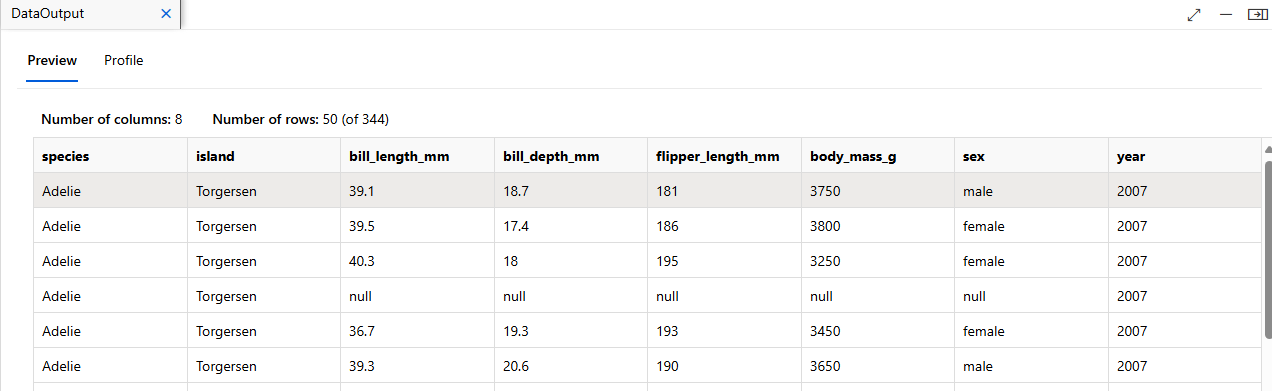
### 데이터 수집/이해

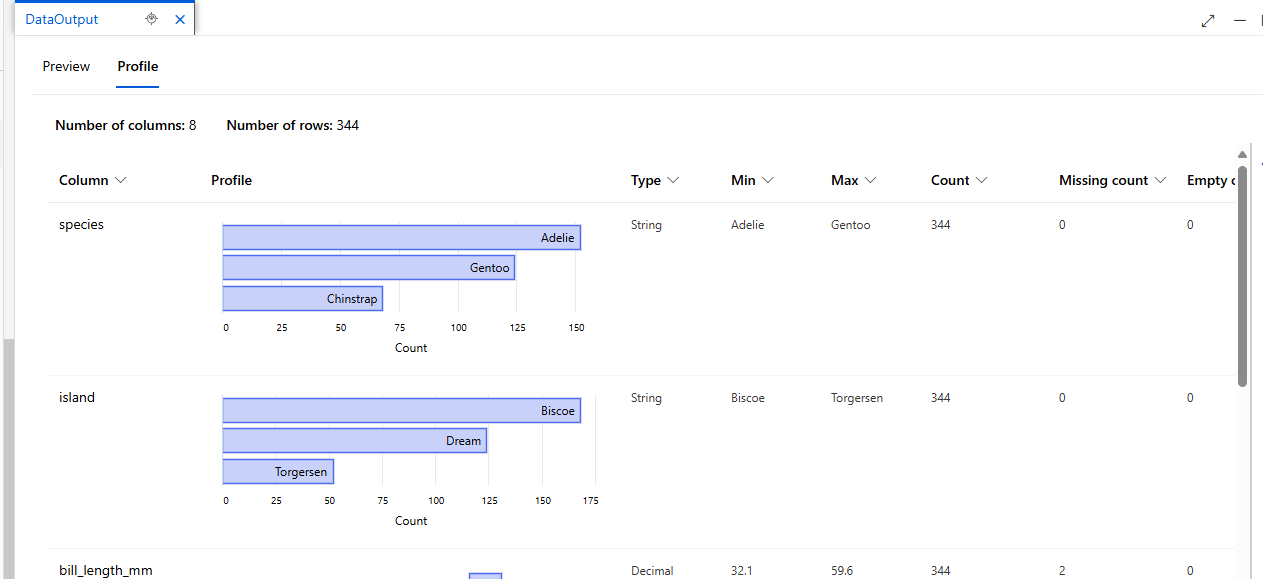
#### 디자이너 시작 및 데이터 세트 가져오기



#### 데이터 이해

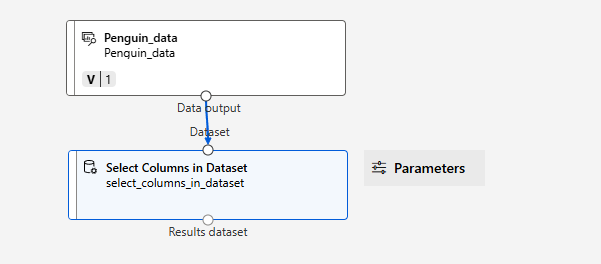
값들을 확인할 수 있다. 대부분이 누락값(null)으로 구성된 샘플도 확인할 수 있다.

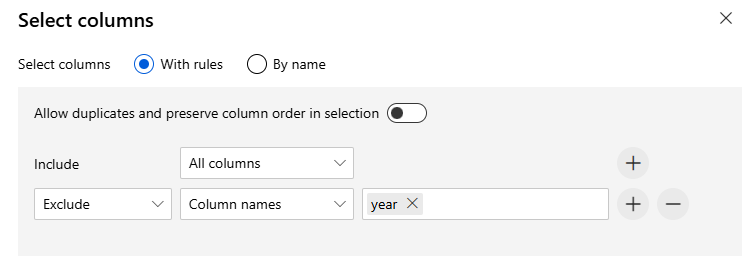




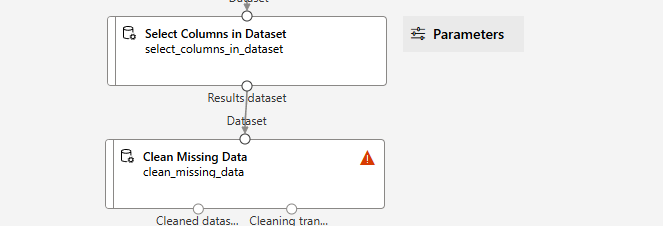
### 데이터 준비

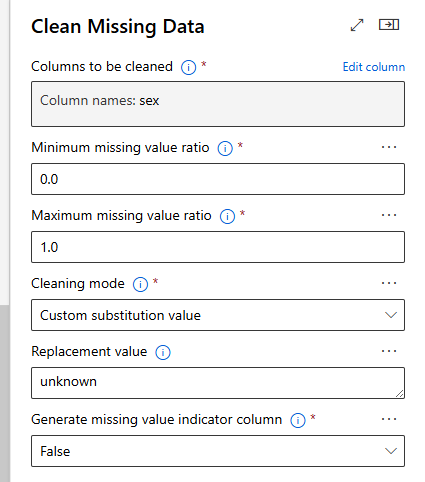
#### 특정 선택 – 불필요한 컬럼 제외



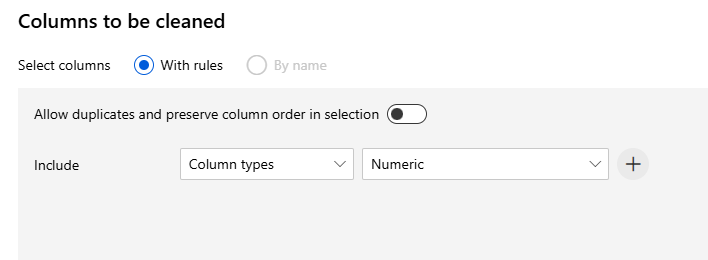


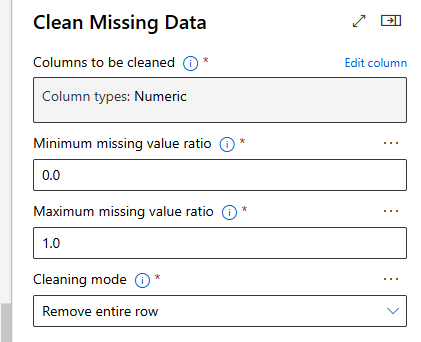
#### 누락값 처리 – 성별



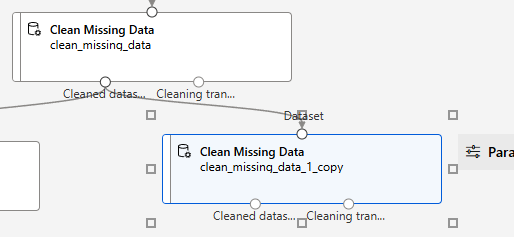


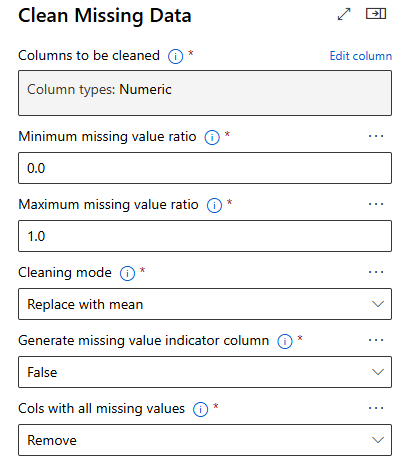
#### 누락값 처리 – 수치형 데이터(선택 1: 전체 행 삭제)





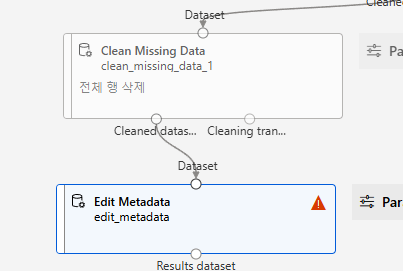
#### 누락값 처리 – 수치형 데이터 (선택 2: 평균값으로 변환)

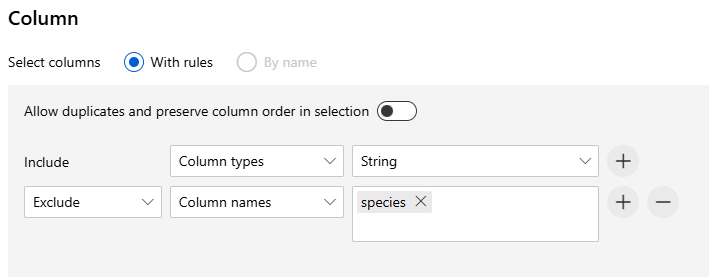


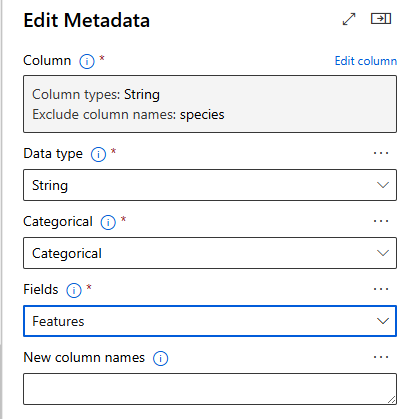


#### 데이터 변환 : String → Category → Indicator value

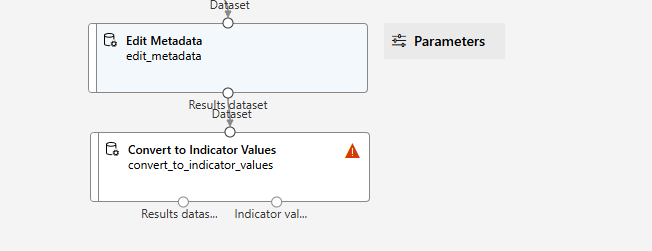
String -> Category



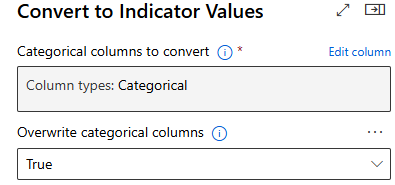




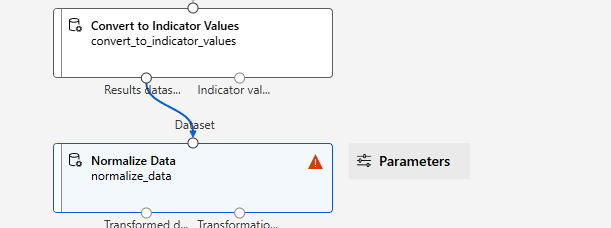
Category -> Indicator value

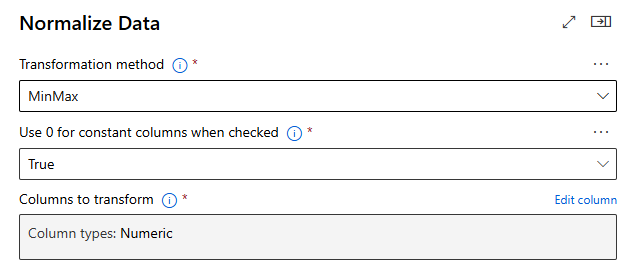




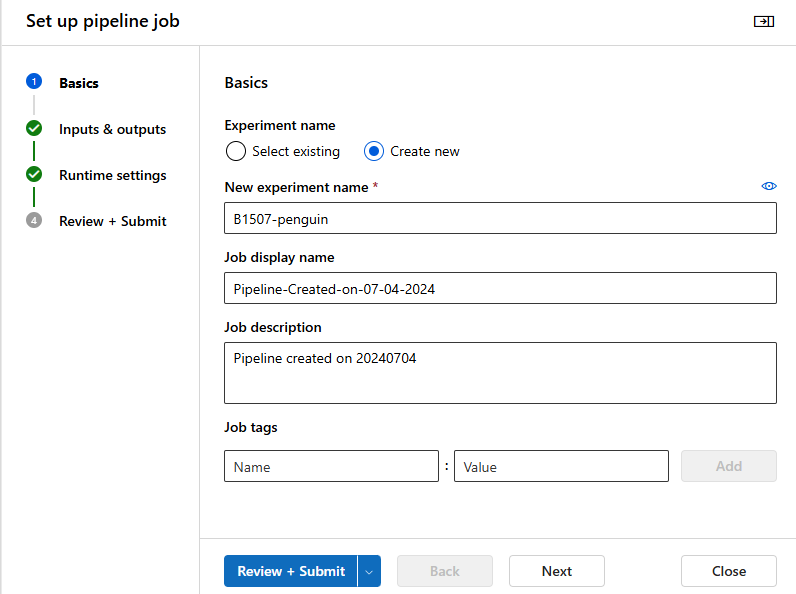


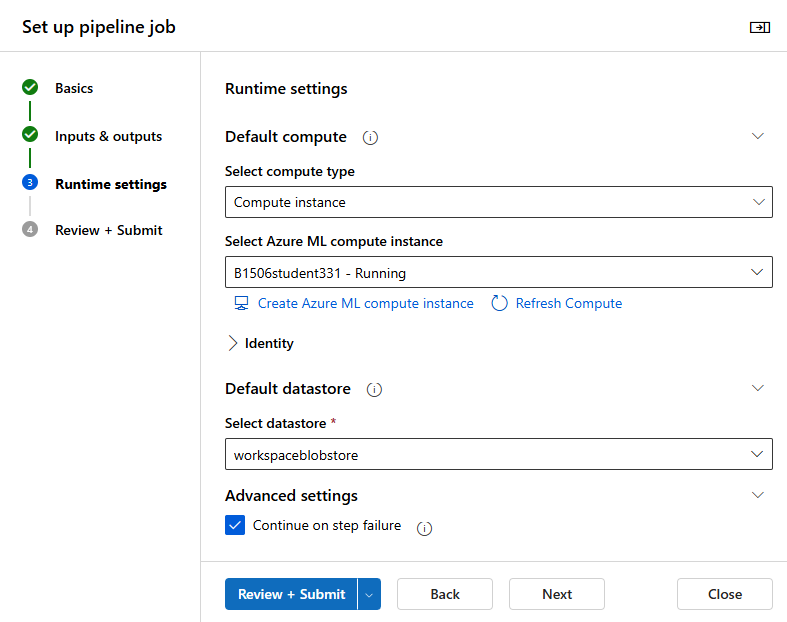
정규화

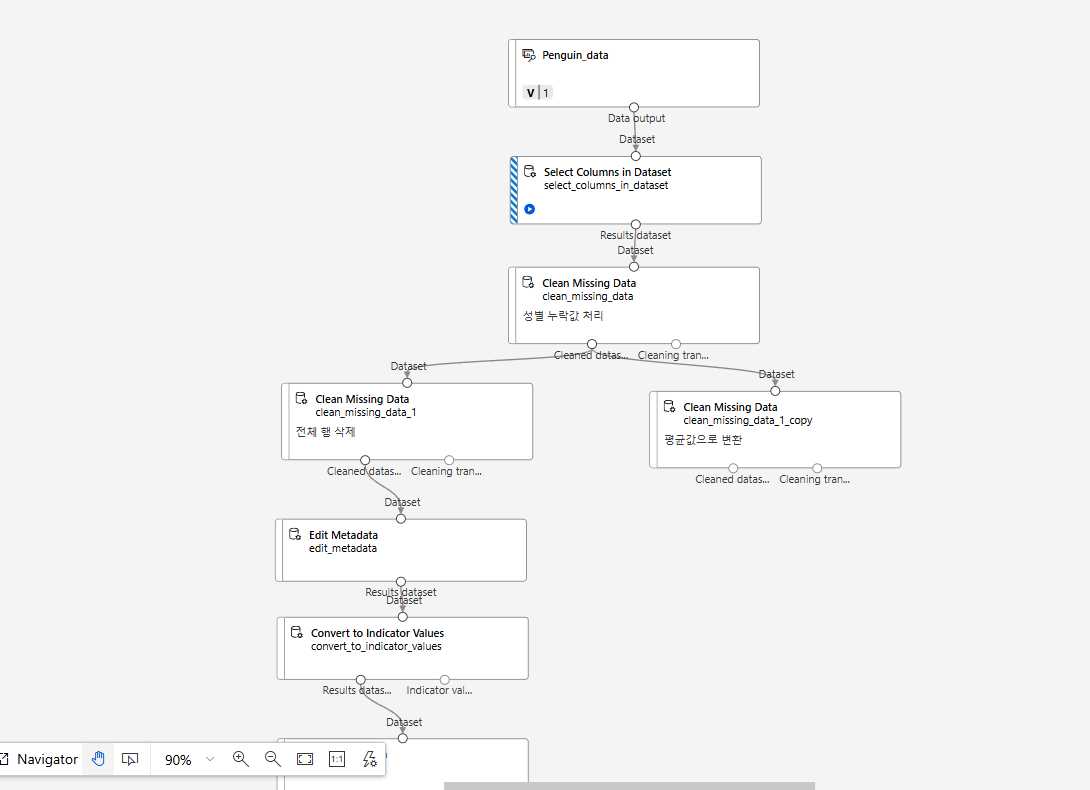




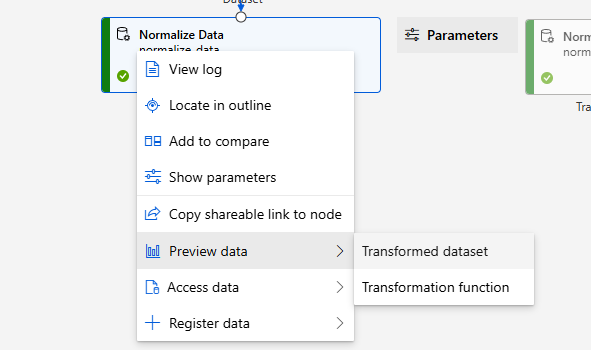
#### 중간 점검

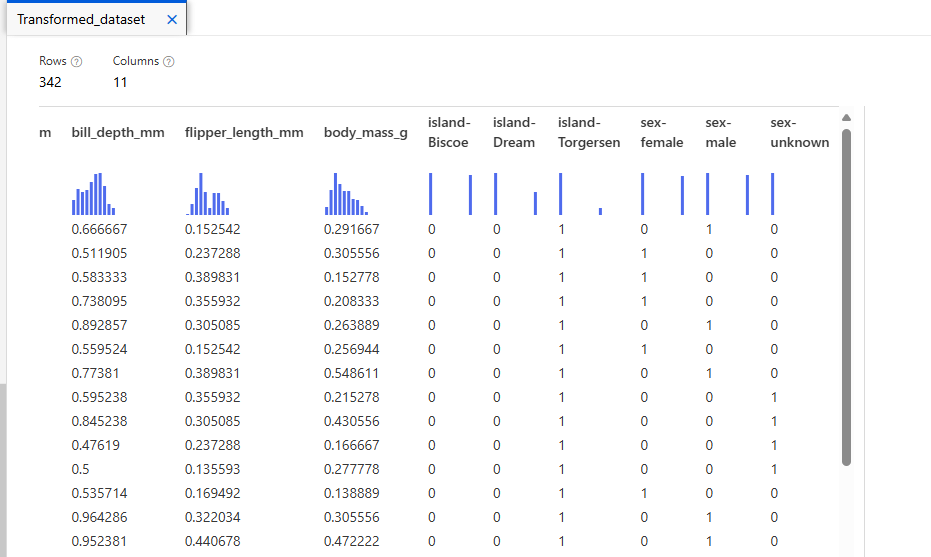






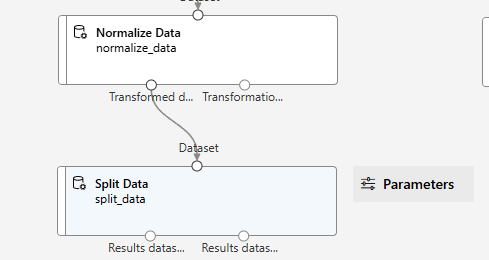
# <오후>

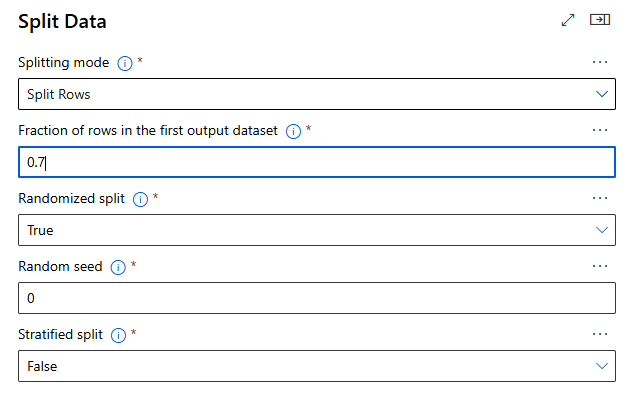




## [실습 - 분류MLD\_펭귄군집화]

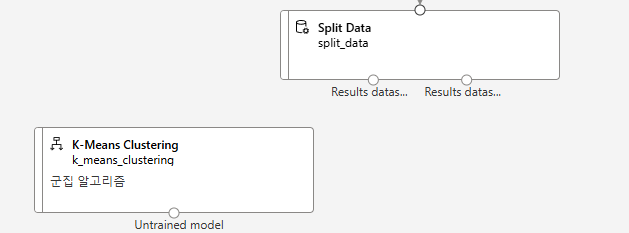
#### 데이터 분리 – 학습 데이터와 테스트 데이터로 분리



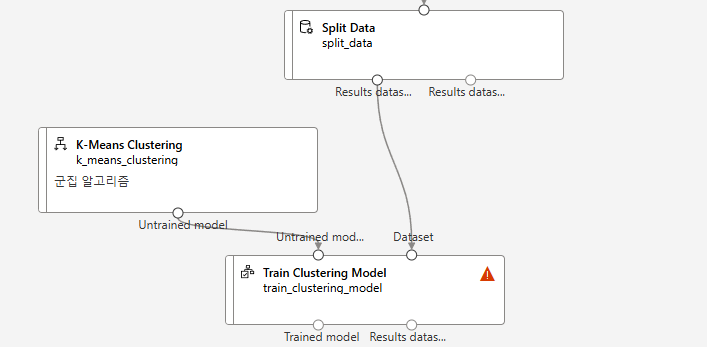


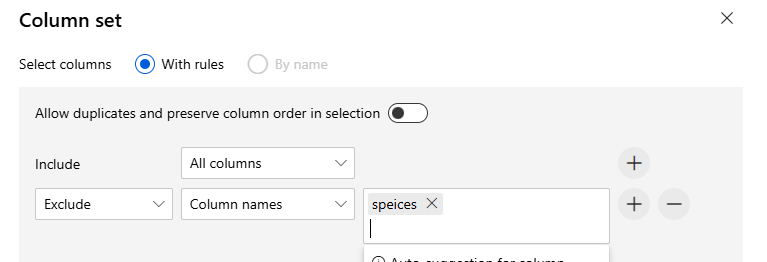
### 모델링/평가

#### 모델링 알고리즘 선택

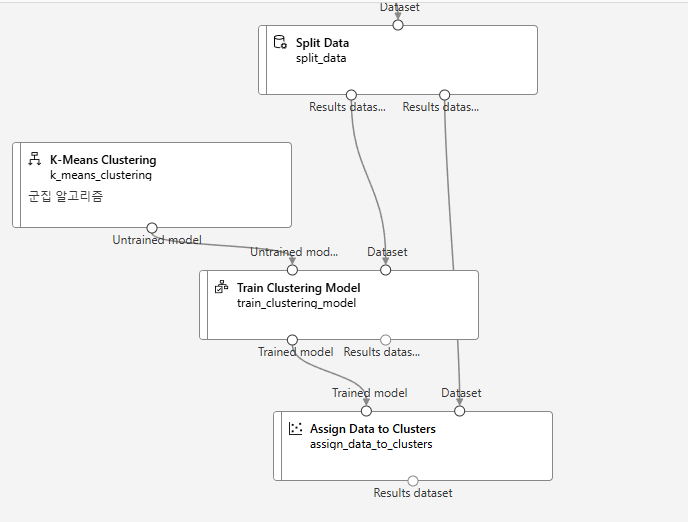


#### 모델 학습(훈련)

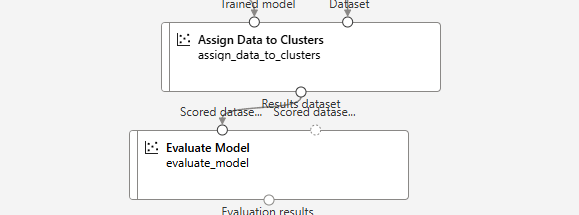


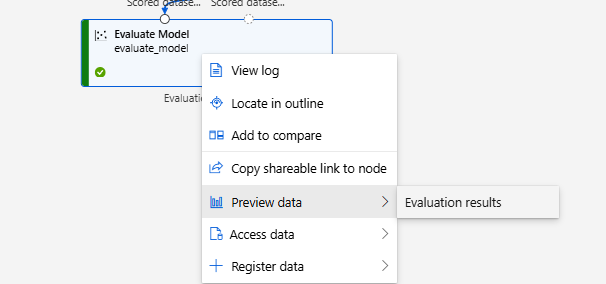


#### 모델 테스트

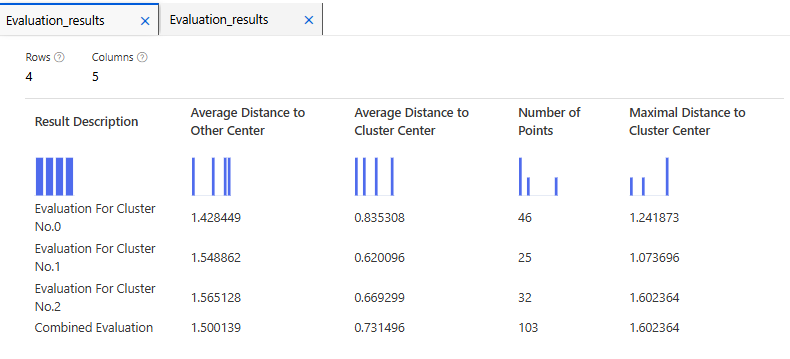


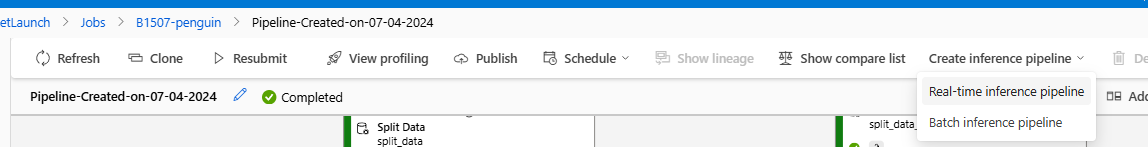
#### 모델 평가



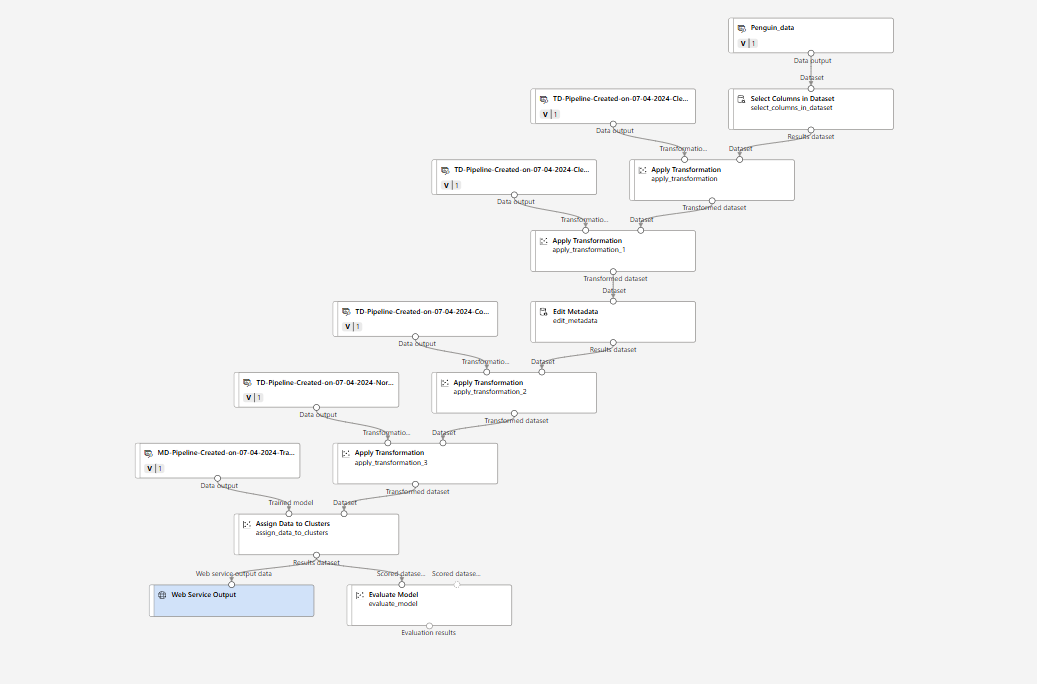


#### 추가 실습 – 유추 파이프라인

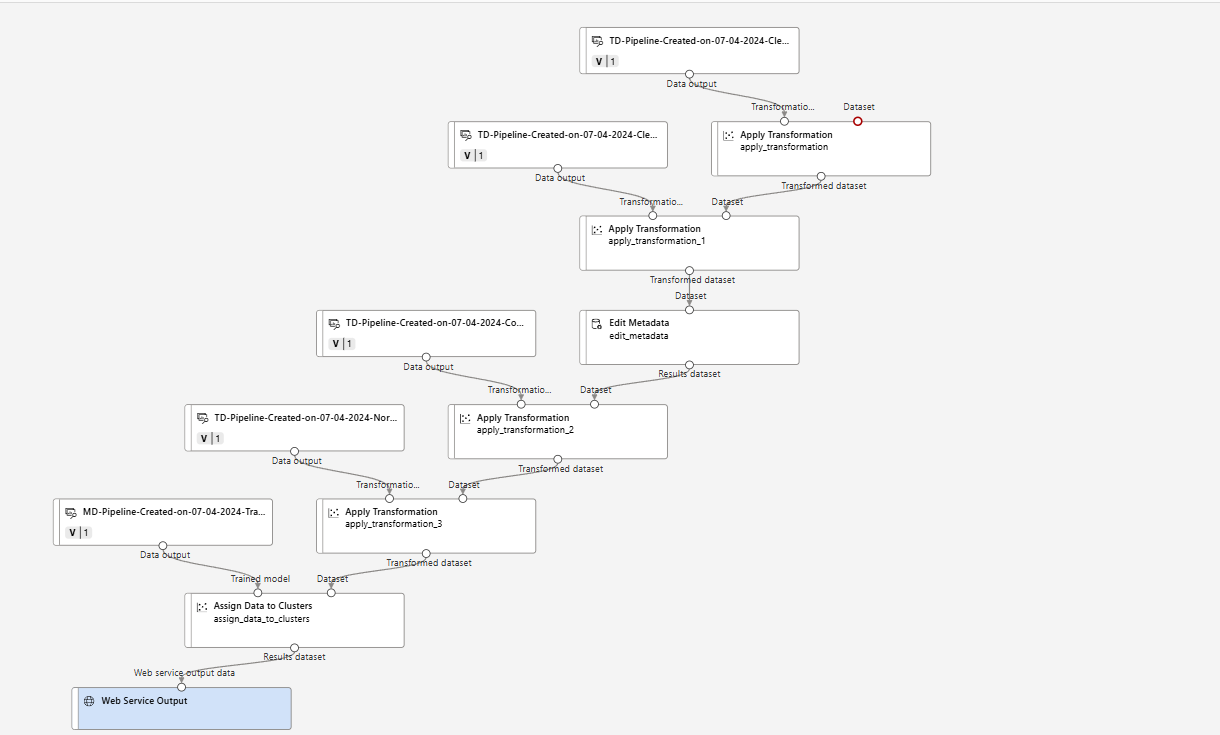




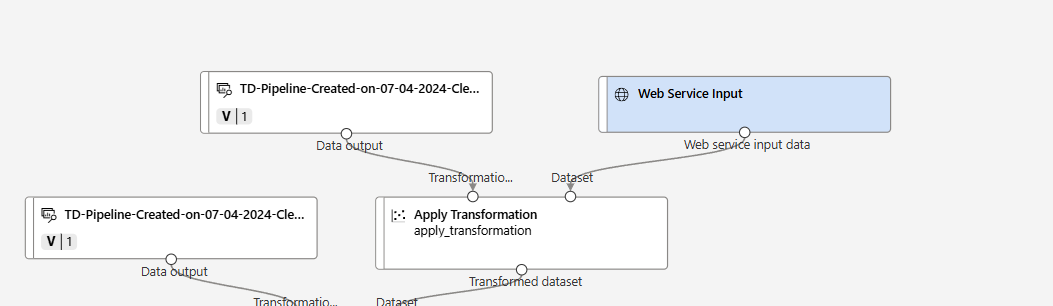
유추 Pipeline 구성을 위해 각 단계에 컴포넌트들이 자동으로 추가/삭제됨



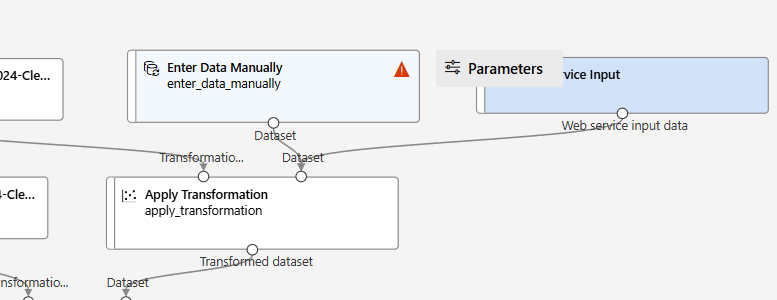
유추 파이프라인 3개 삭제

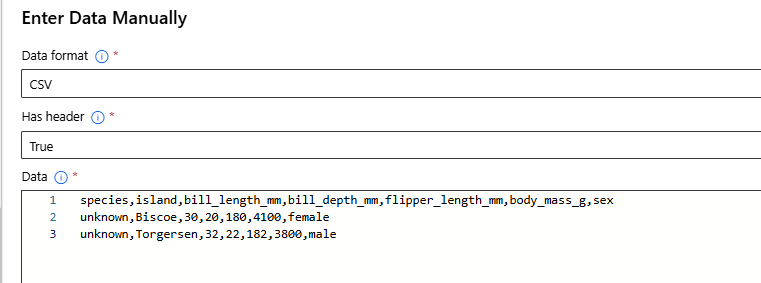


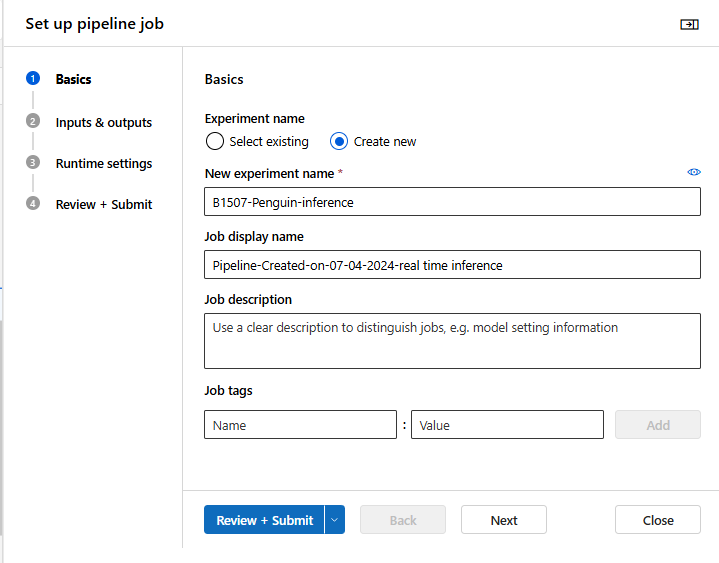
Web service input 추가

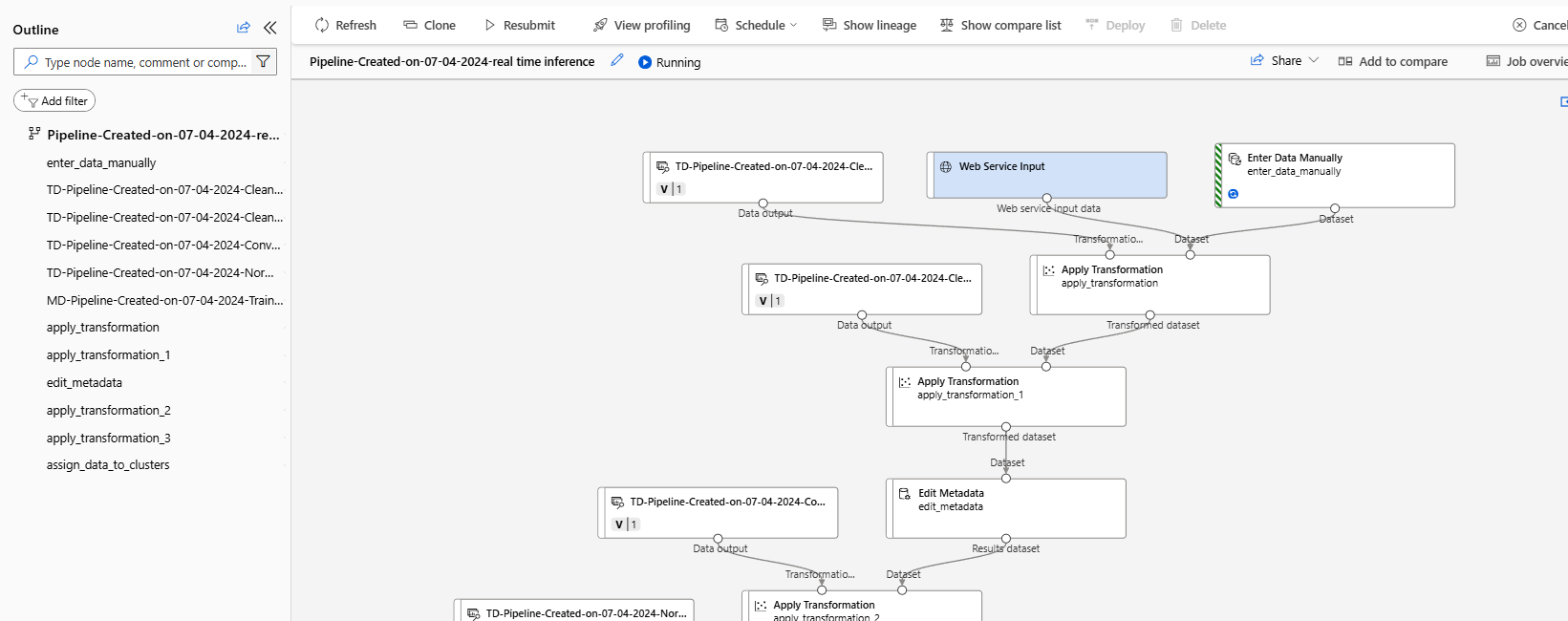


Enter Data Manually 추가

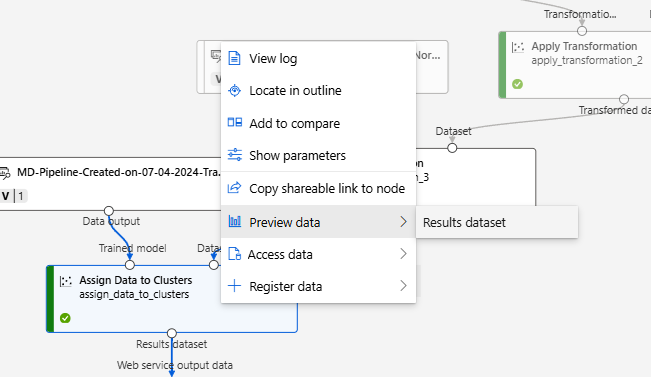


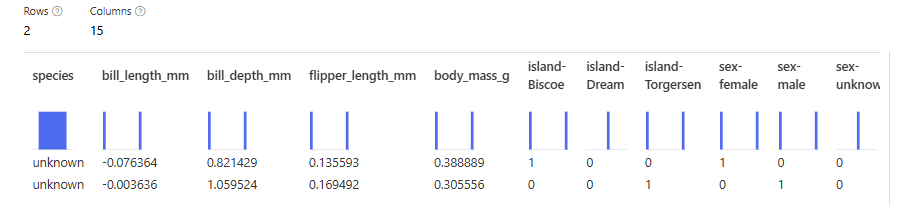


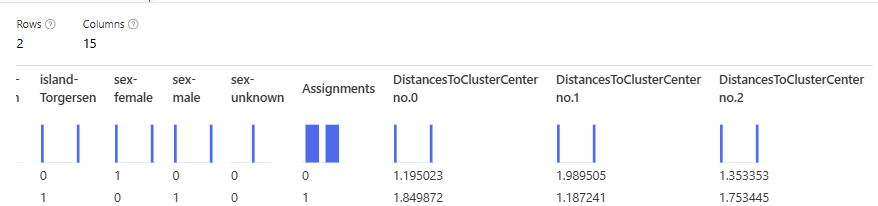




#### 유추 파이프라인 테스트

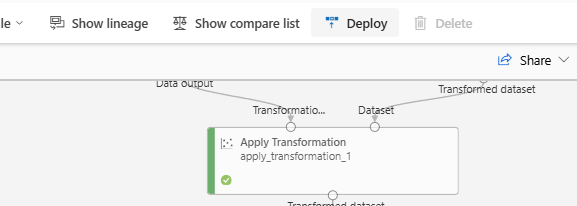




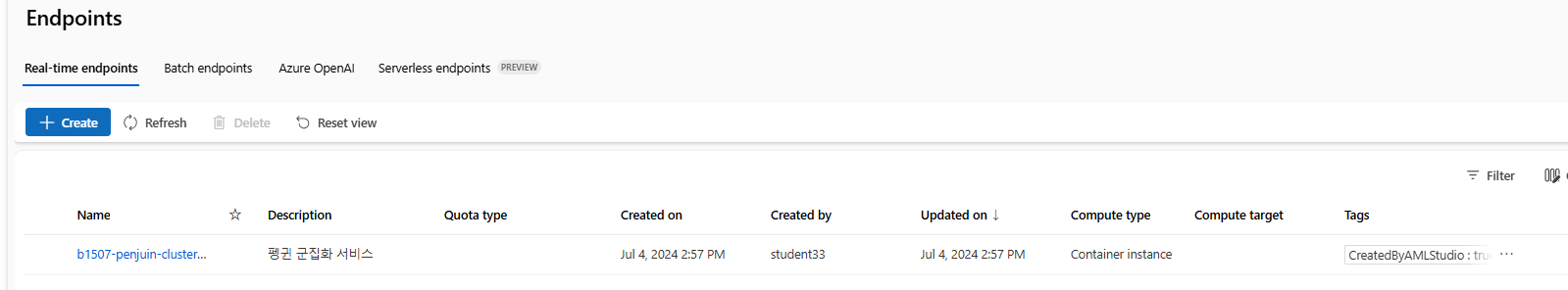


#### 유추 파이프라인 배포

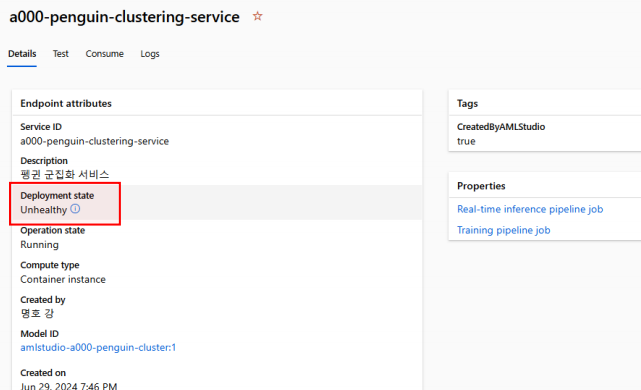
Deploy 클릭



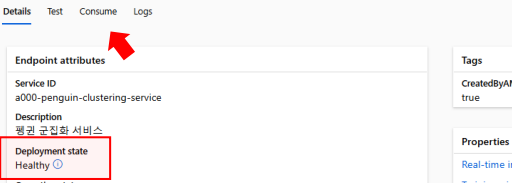


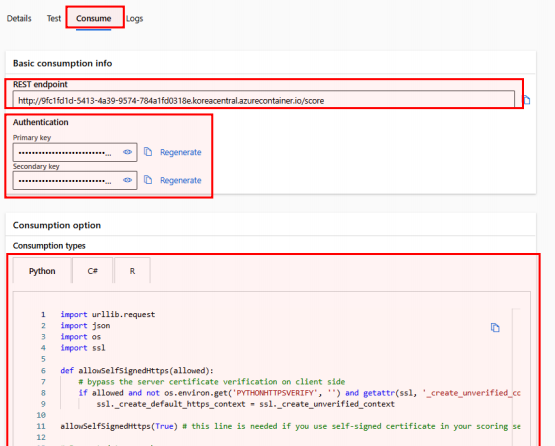


* 오래걸려서 PDF 파일의 자료로 대체



Healthy로 변경되면 Consume 선택





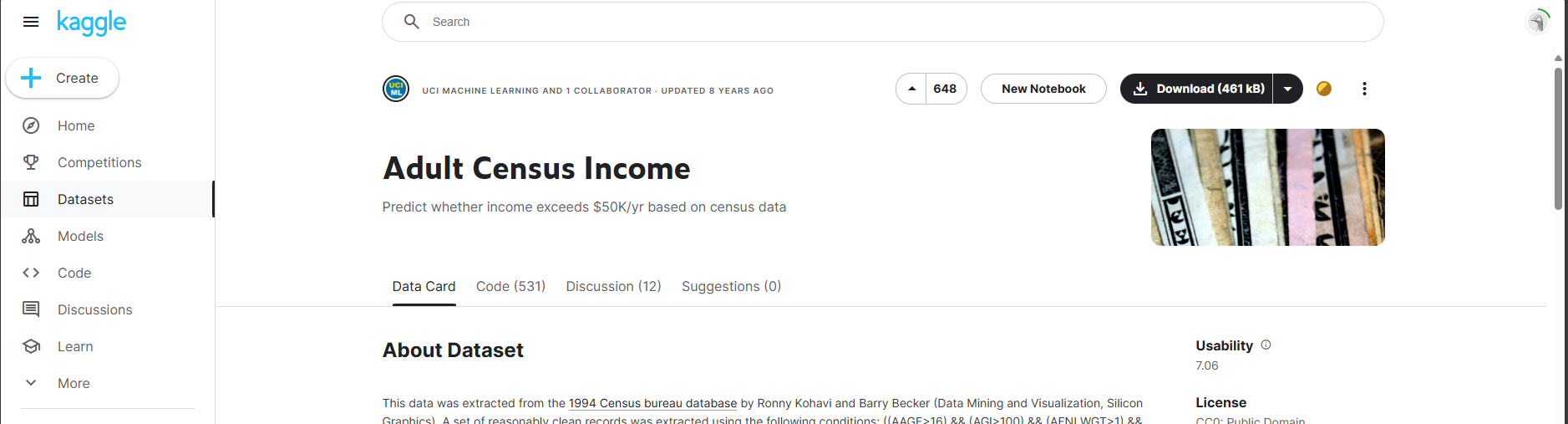
## [실습 - 분류MLD\_개인수입예측\_BoostedDT\_tuneHP]

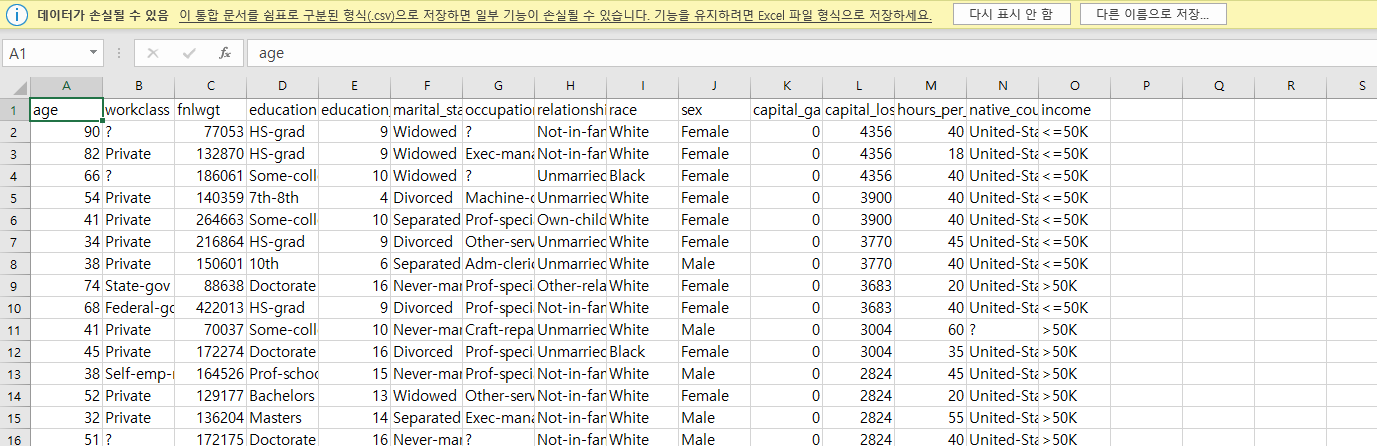
### 실습 준비

#### 데이터 다운로드

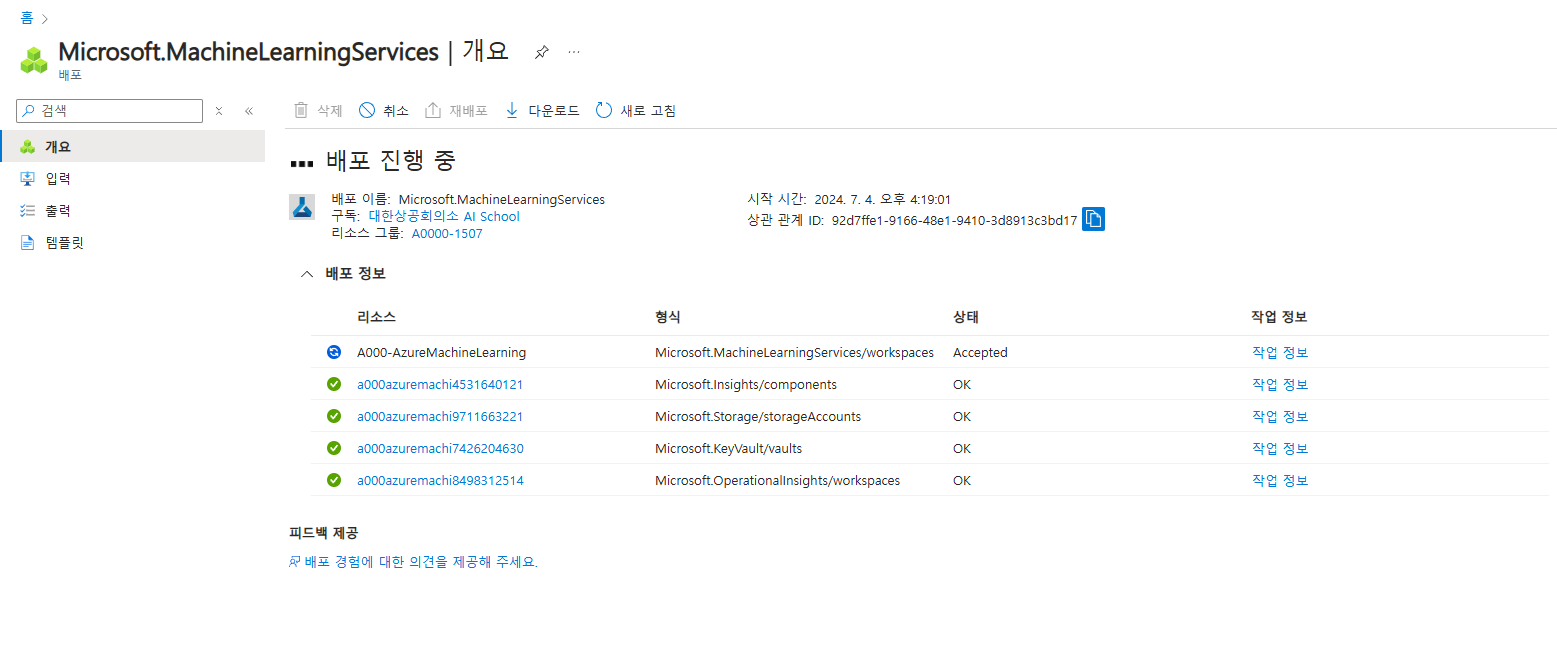
Kaggle.com 접속 후 Datasets 메뉴 클릭

"adult census income" 검색 후 데이터세트 선택

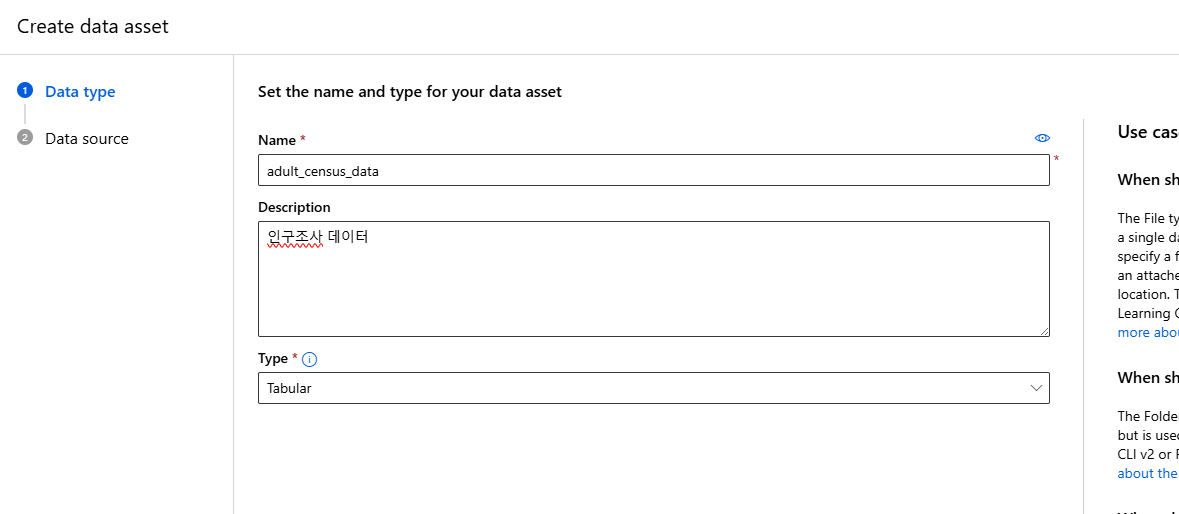


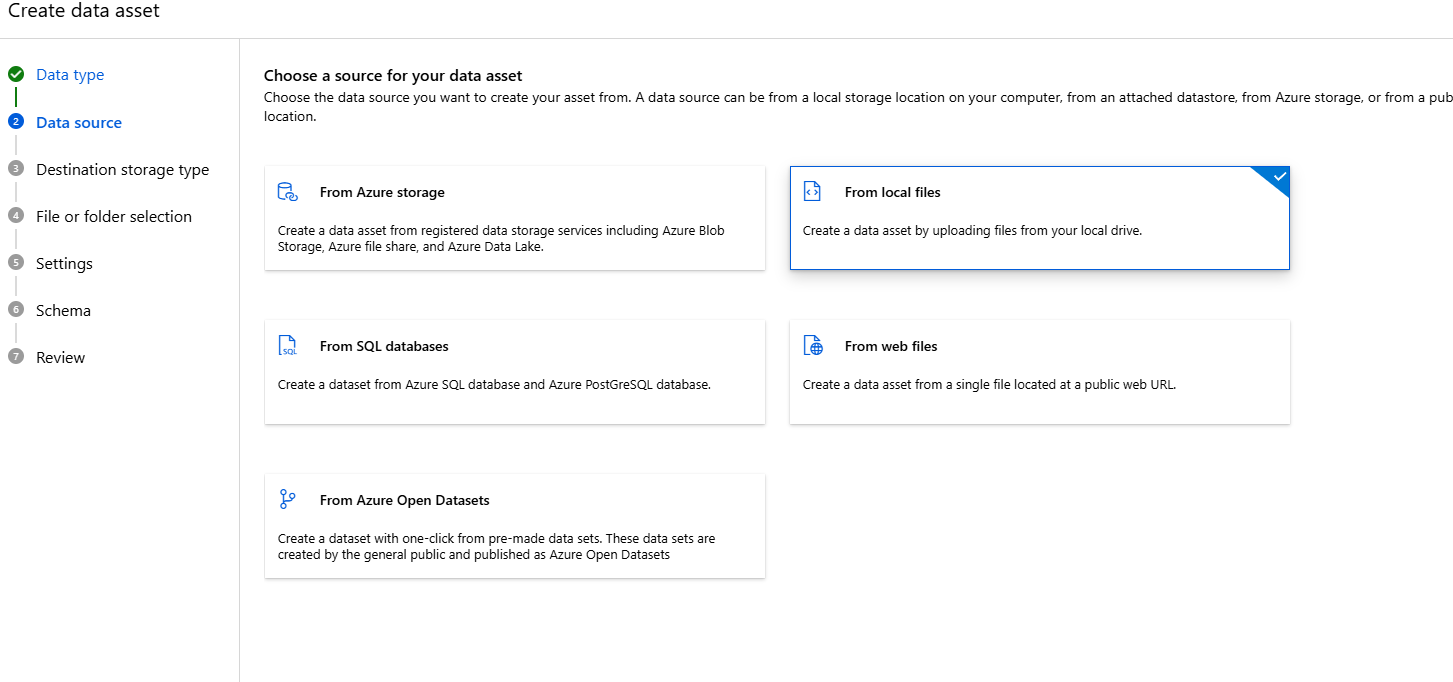


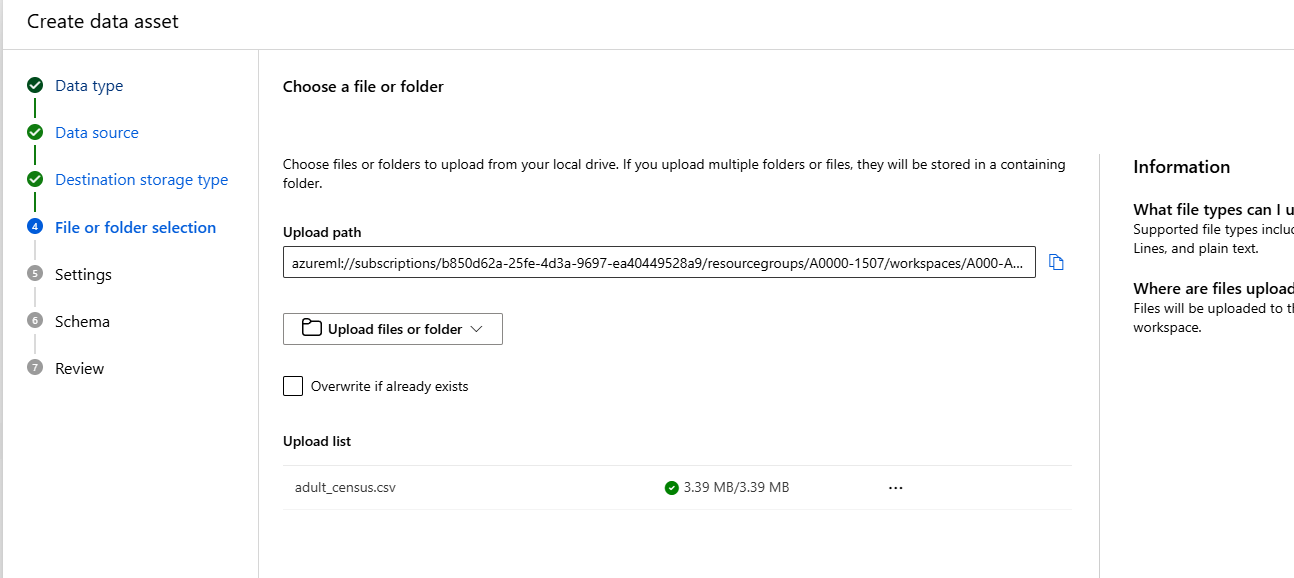
#### 리소스 그룹에서 머신러닝 리소스 생성

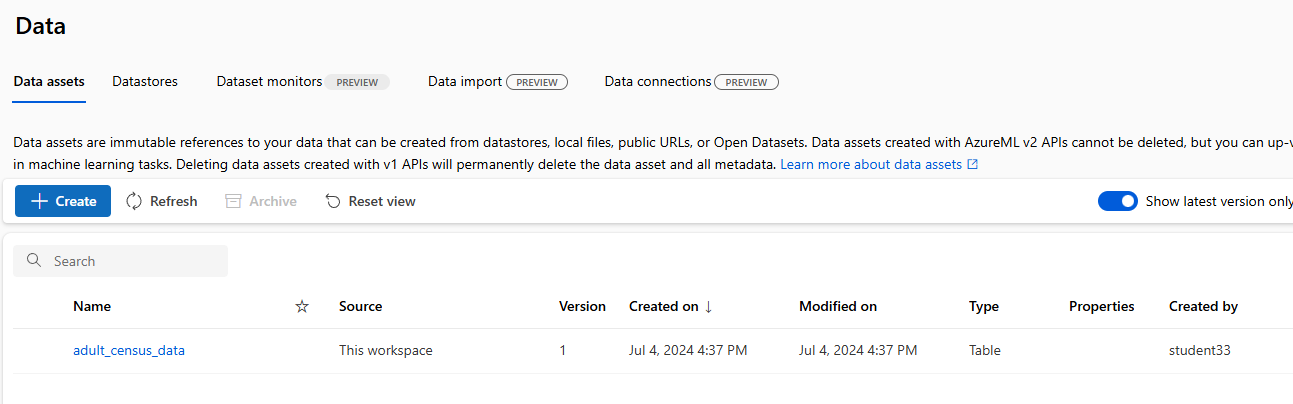


#### 데이터 세트 등록



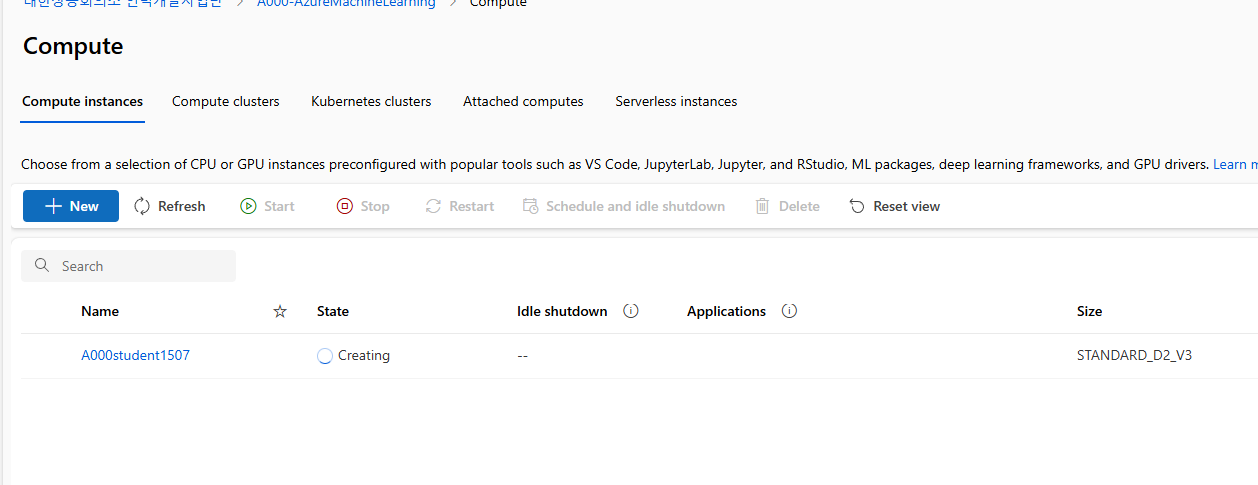






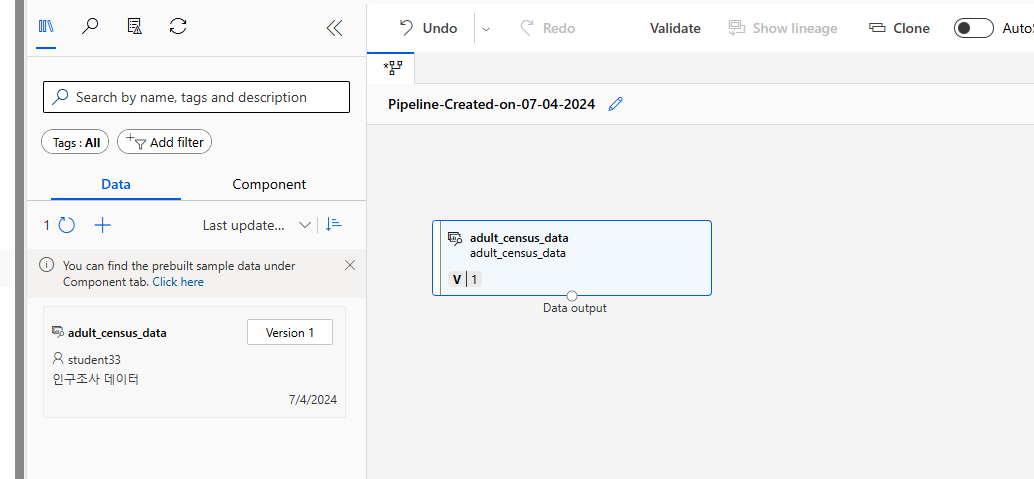
#### 컴퓨트 대상 설정

2코어로 생성



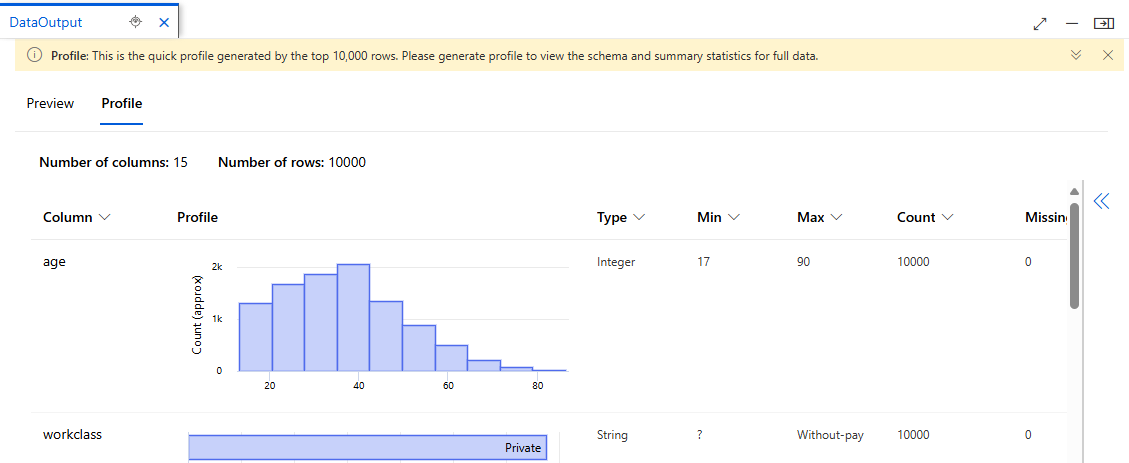
### 데이터 수집/이해

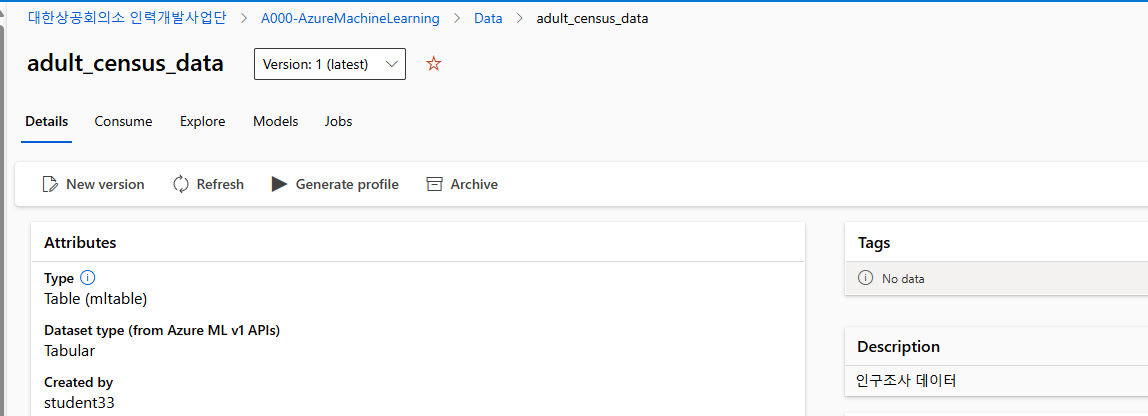
#### 디자이너에서 파이프라인 생성 후 데이터 세트 가져오기

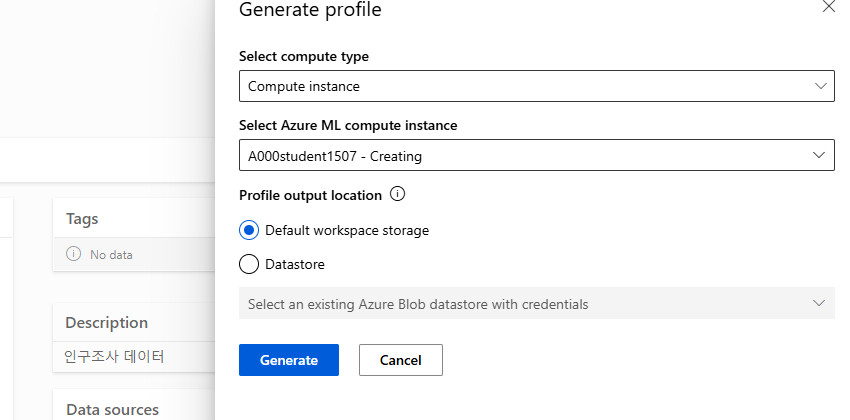


#### 데이터 이해

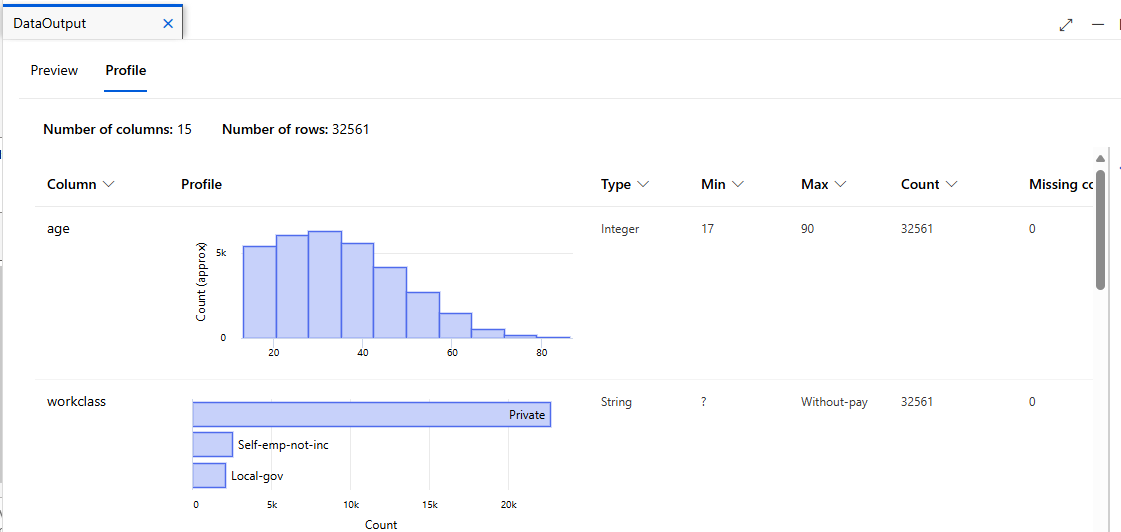
현재 Profile 에서는 10000개의 행만을 확인이 가능하기에 저장 후 Data 메뉴로 이동하여 Generate profile 클릭





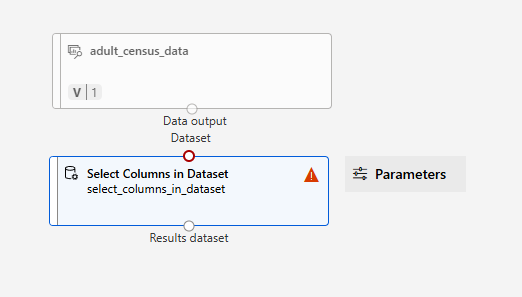


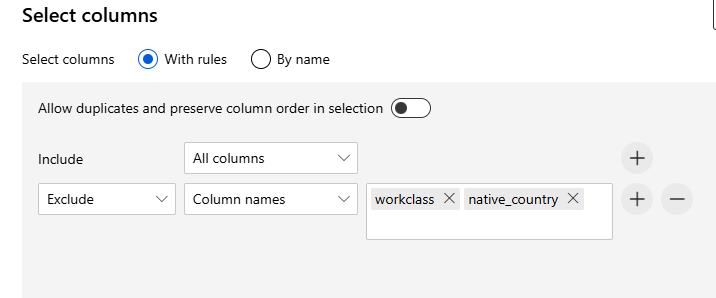
확인할 수 있는 행이 늘어난 것을 볼 수 있다.



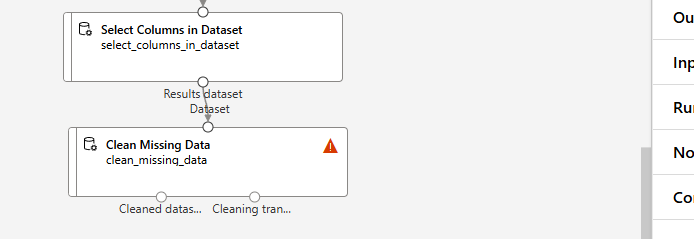
### 데이터 준비

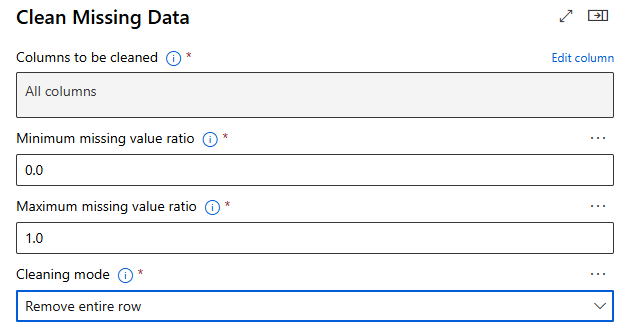
#### 특성 선택





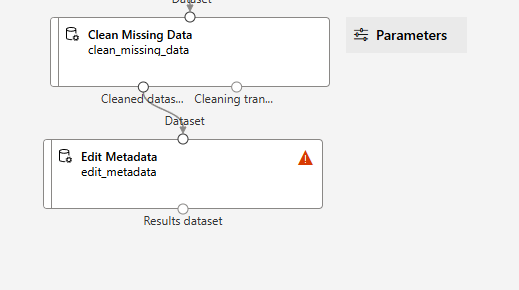
#### 누락값 처리

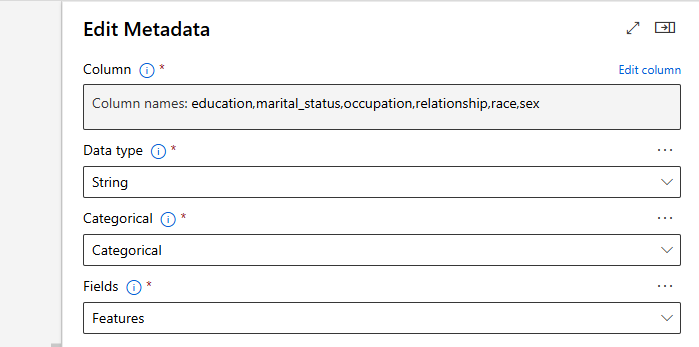




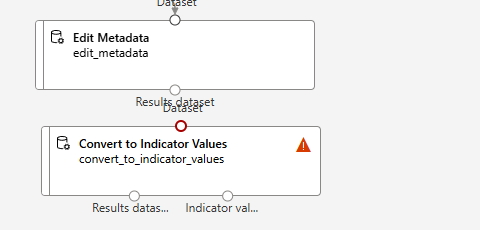
#### 데이터 변환

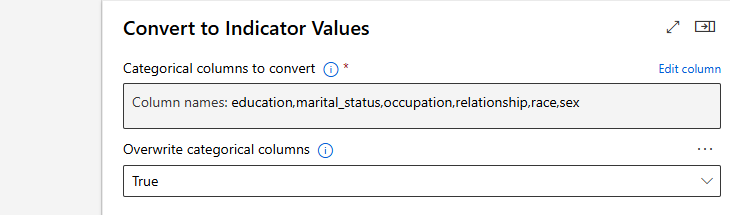
String -> Category



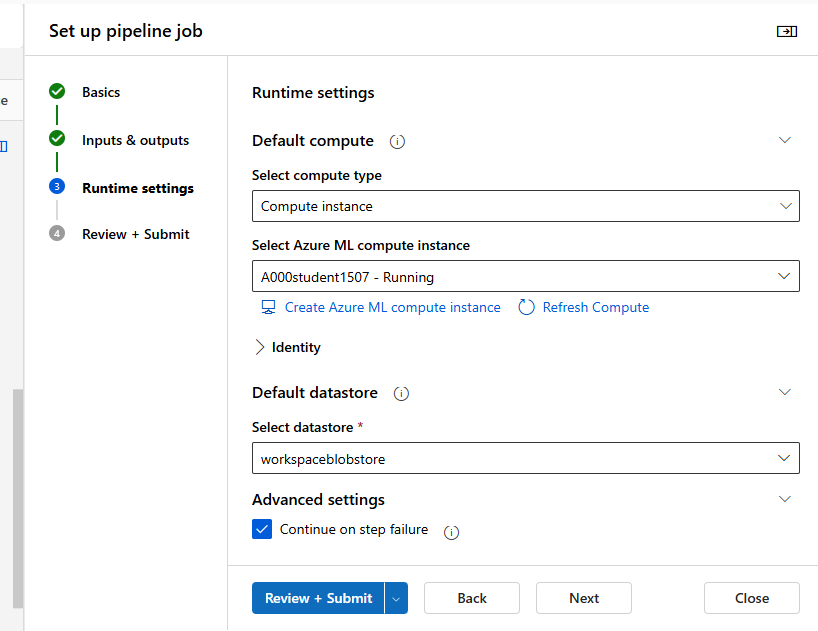
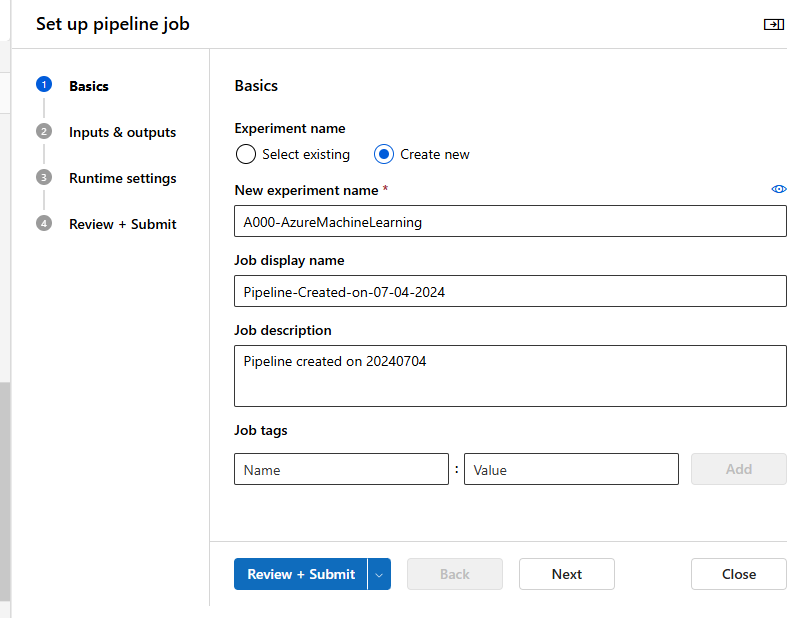


Category -> Indicator value

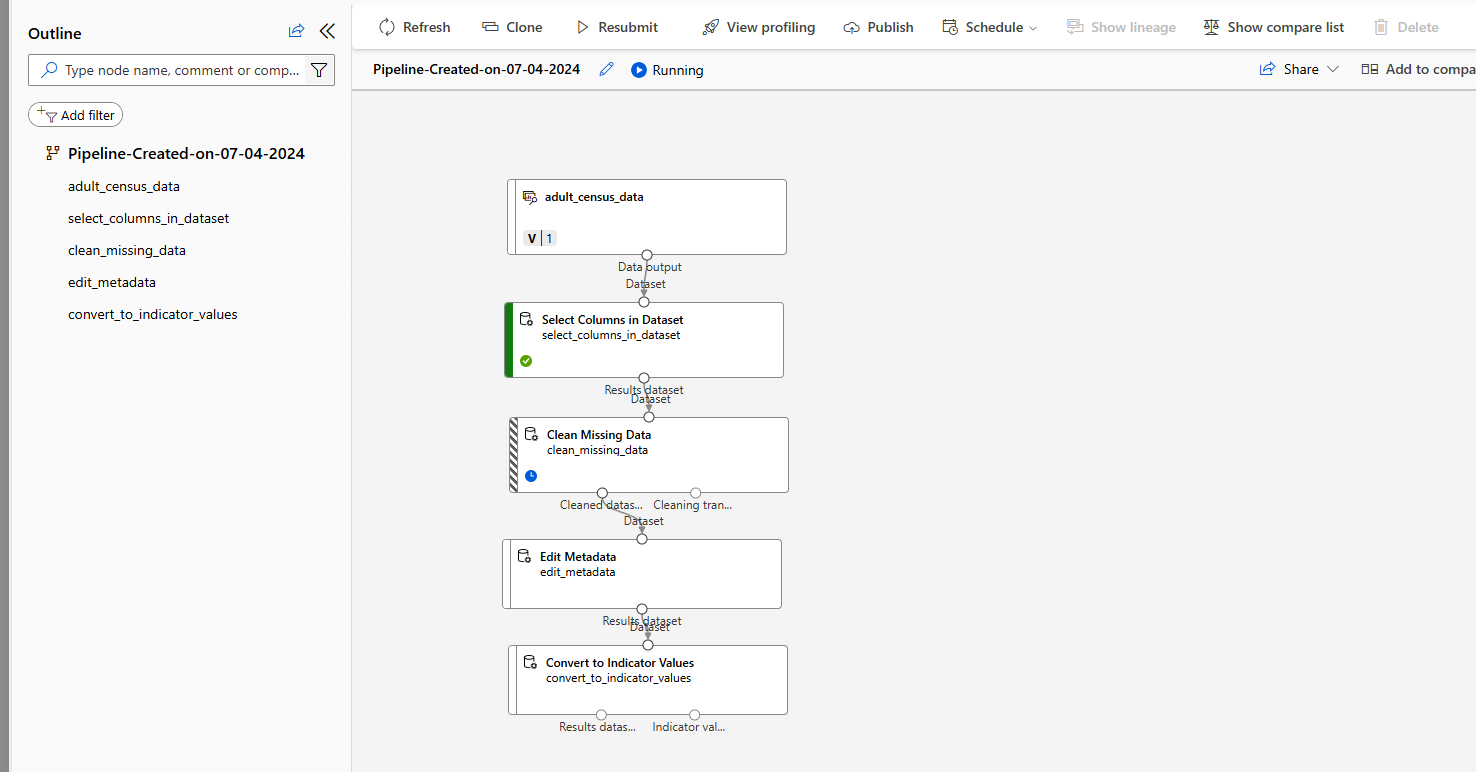


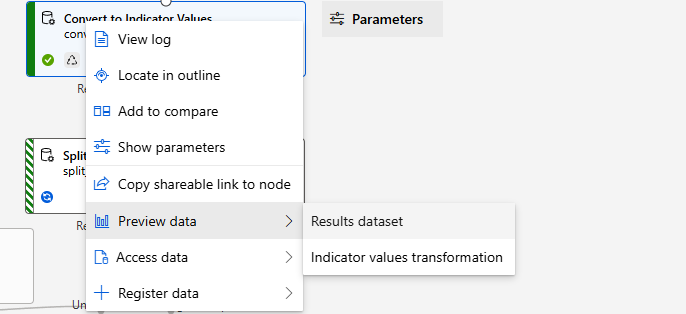


#### 중간점검



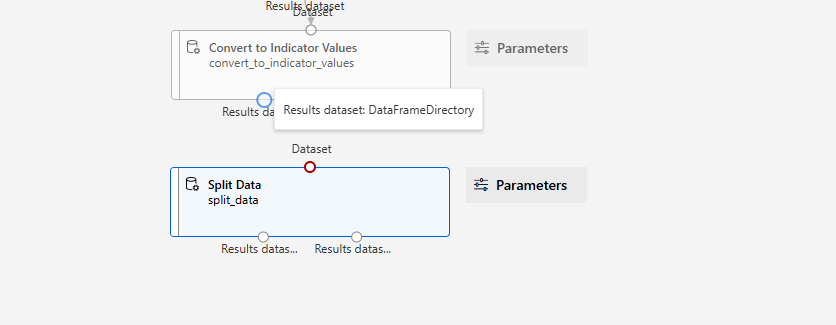
job에서 확인하기

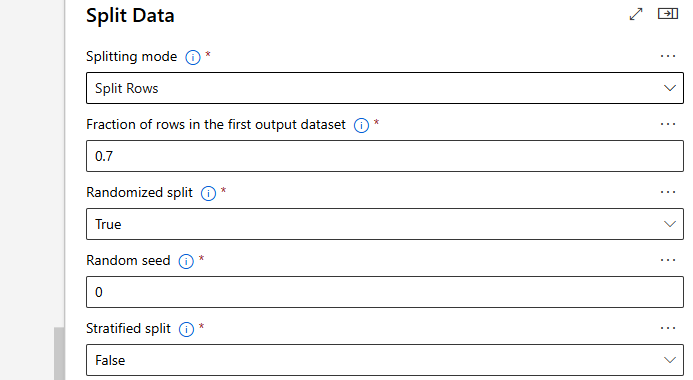




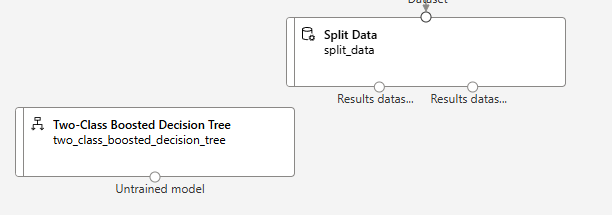


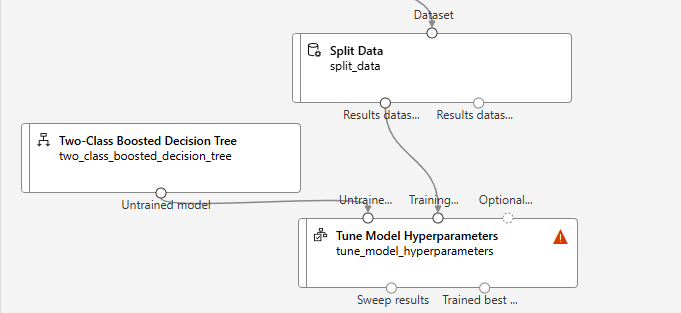
#### 데이터 분리

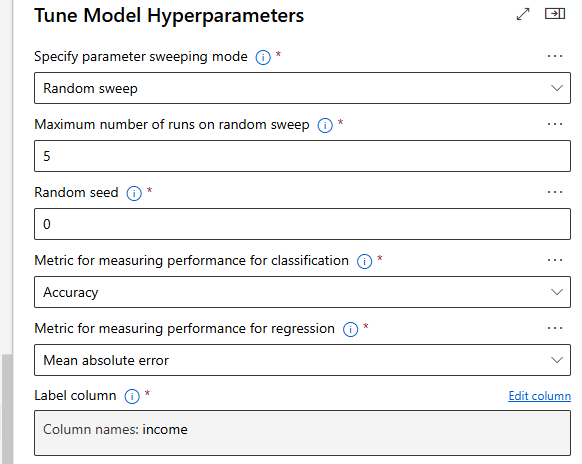




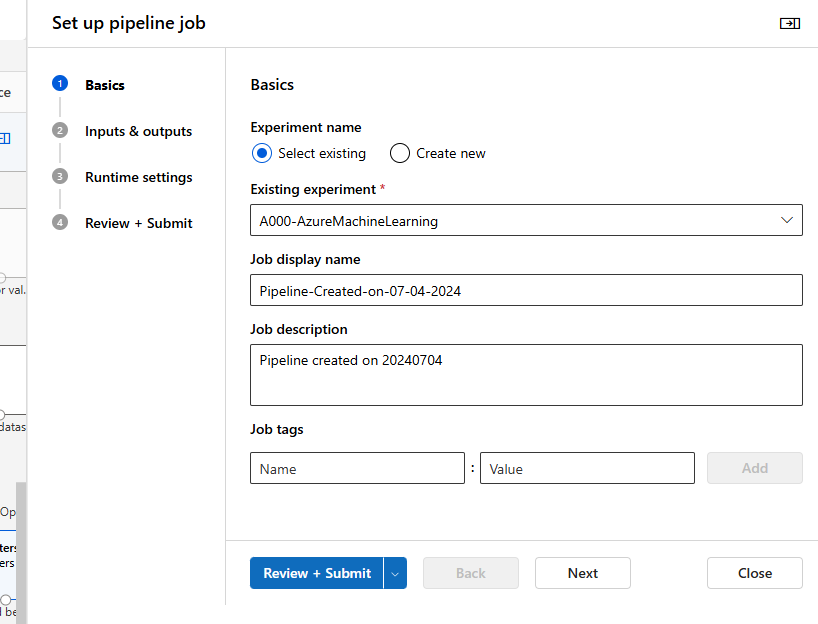
#### 모델링/평가

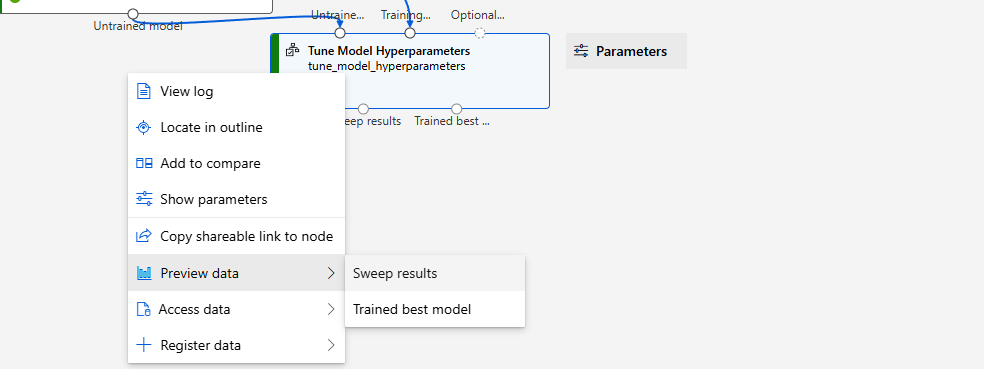




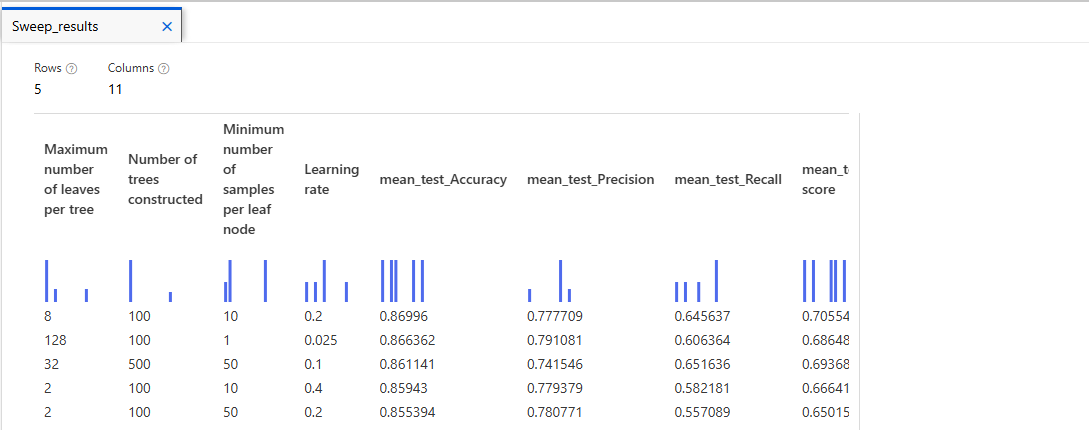


Job 실행

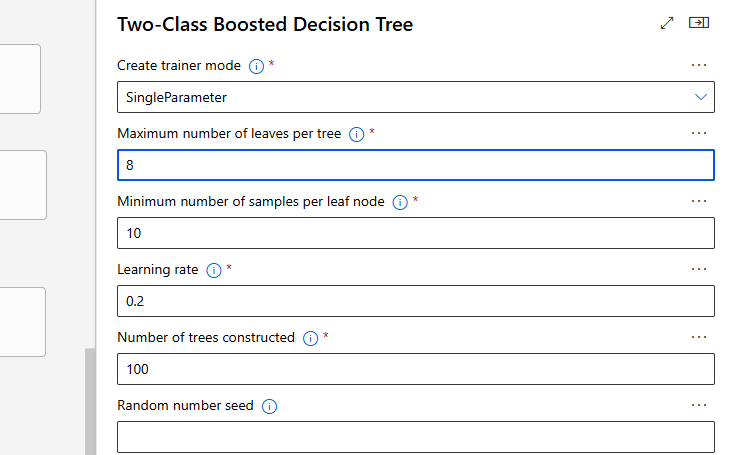




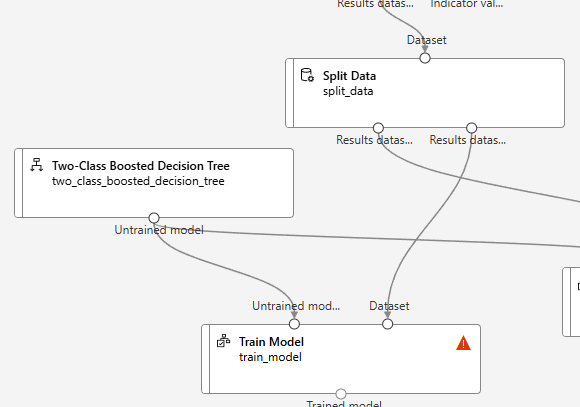
첫 번째 결과로 알고리즘 하이퍼파라미터 사용

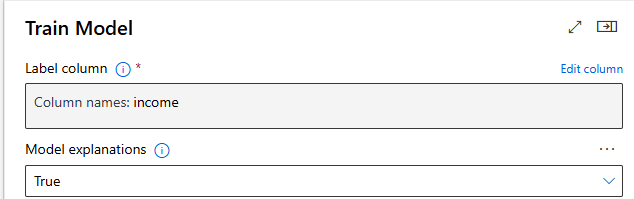


Designer로 돌아와 Two-Class Boosted Decision Tree(알고리즘)의 하이퍼파라미터 설정

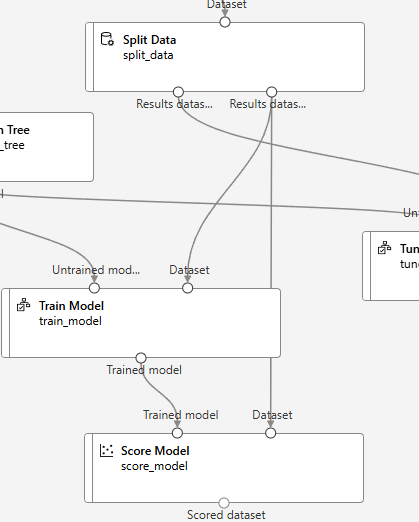


#### 모델 학습(훈련)

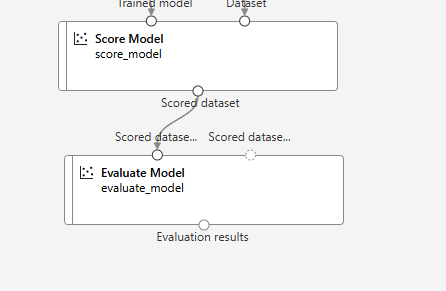


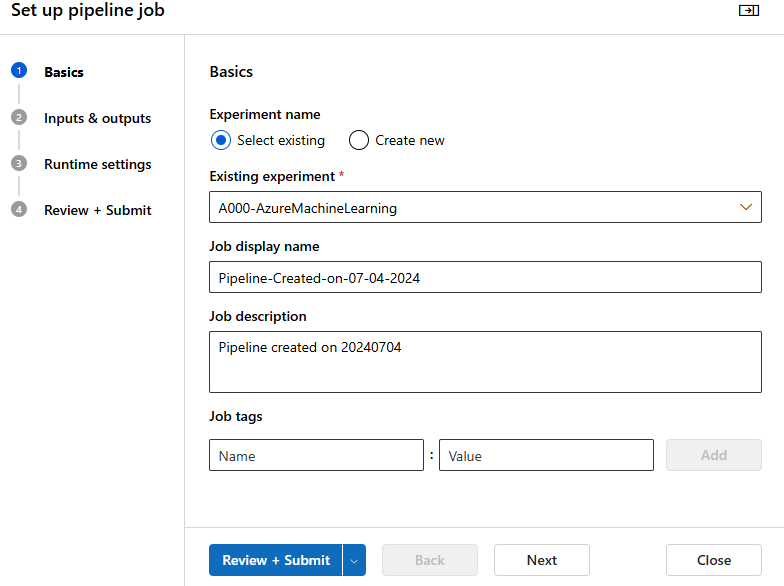


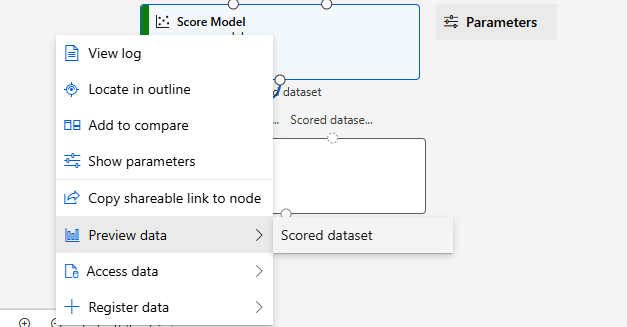
#### 모델 테스트

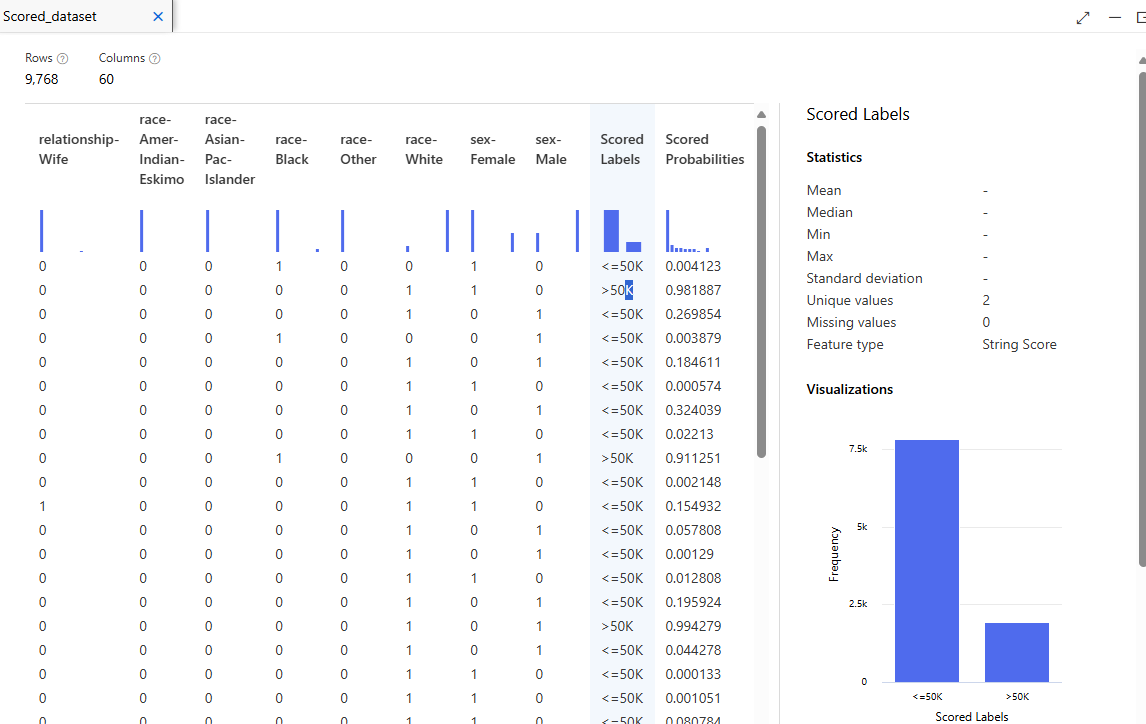


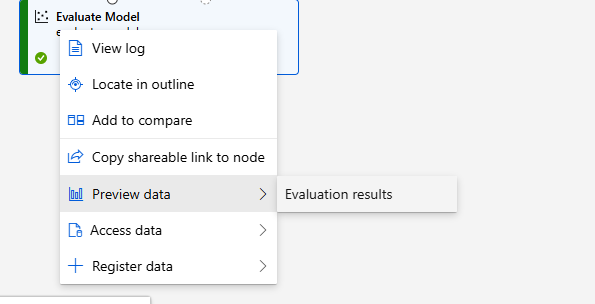
#### 모델 평가

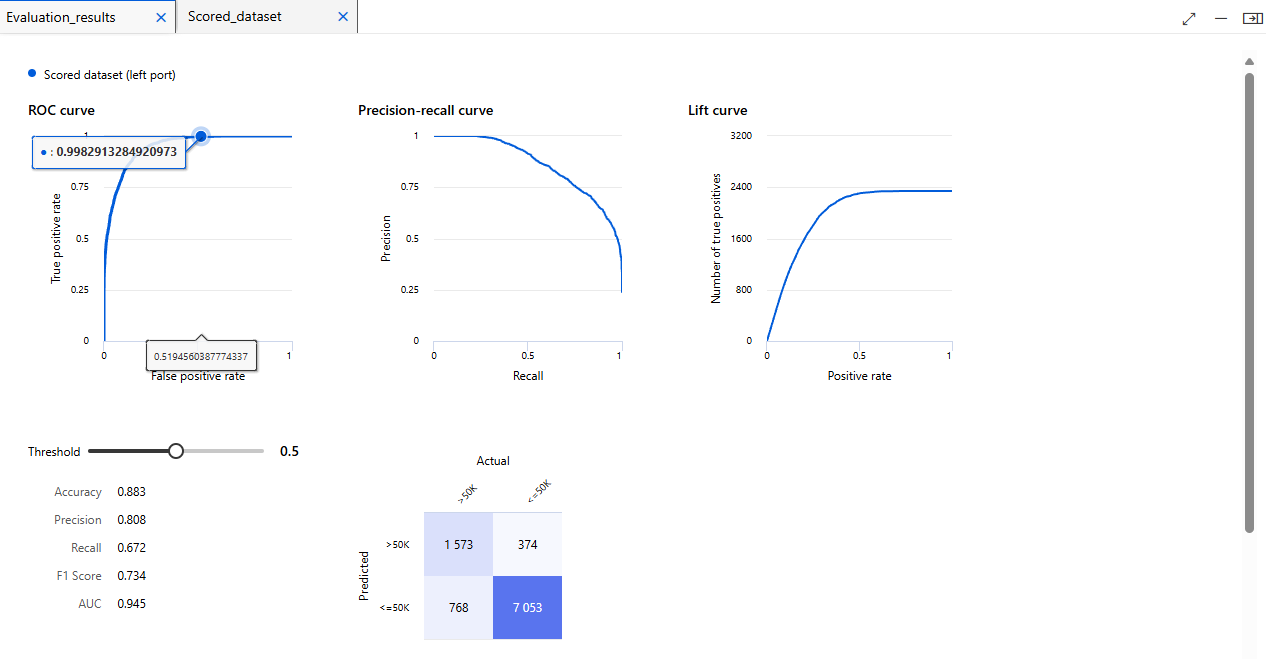


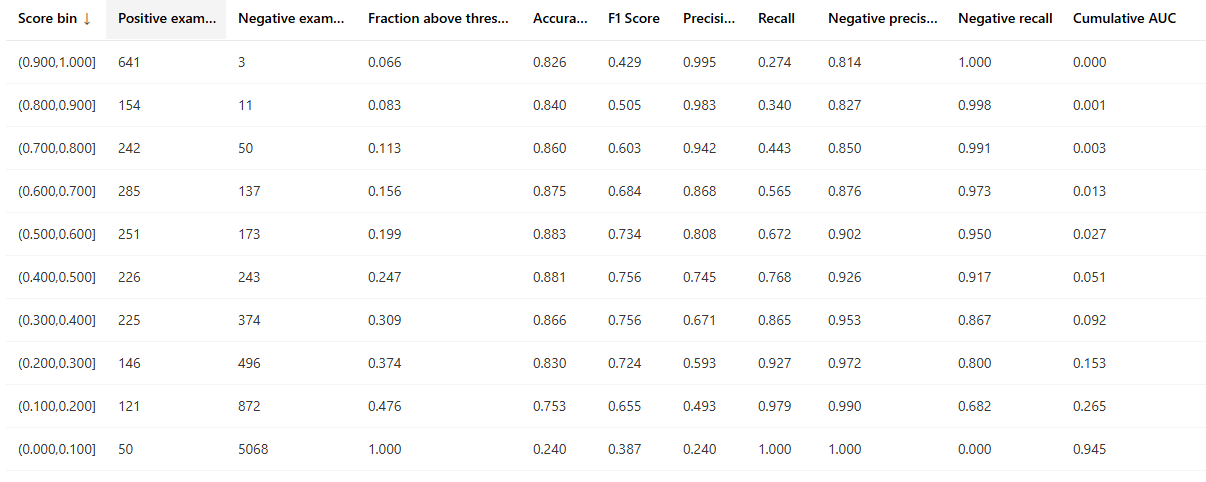












## [실습 - 회귀MLD\_자동차가격예측\_회귀트리]

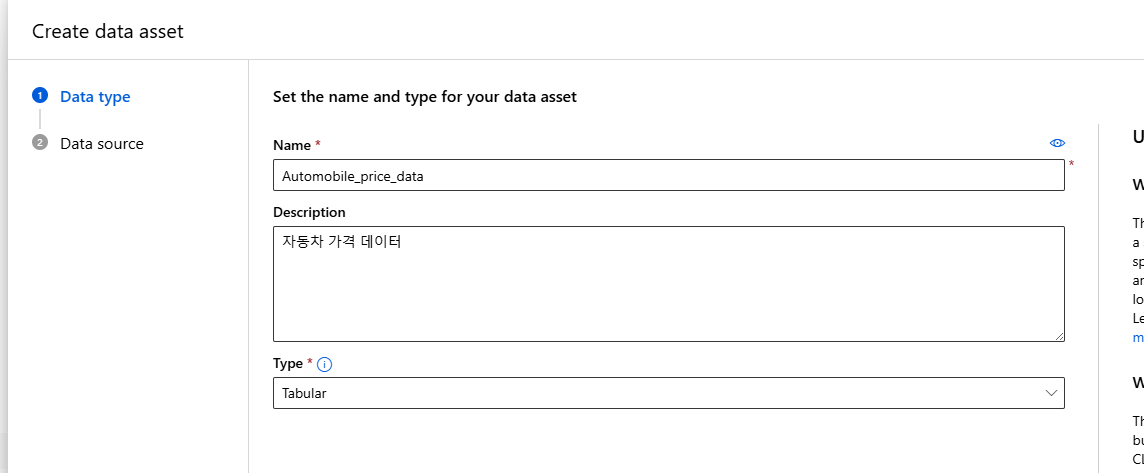
MS Azure ML Designer를 활용한 회귀 모델

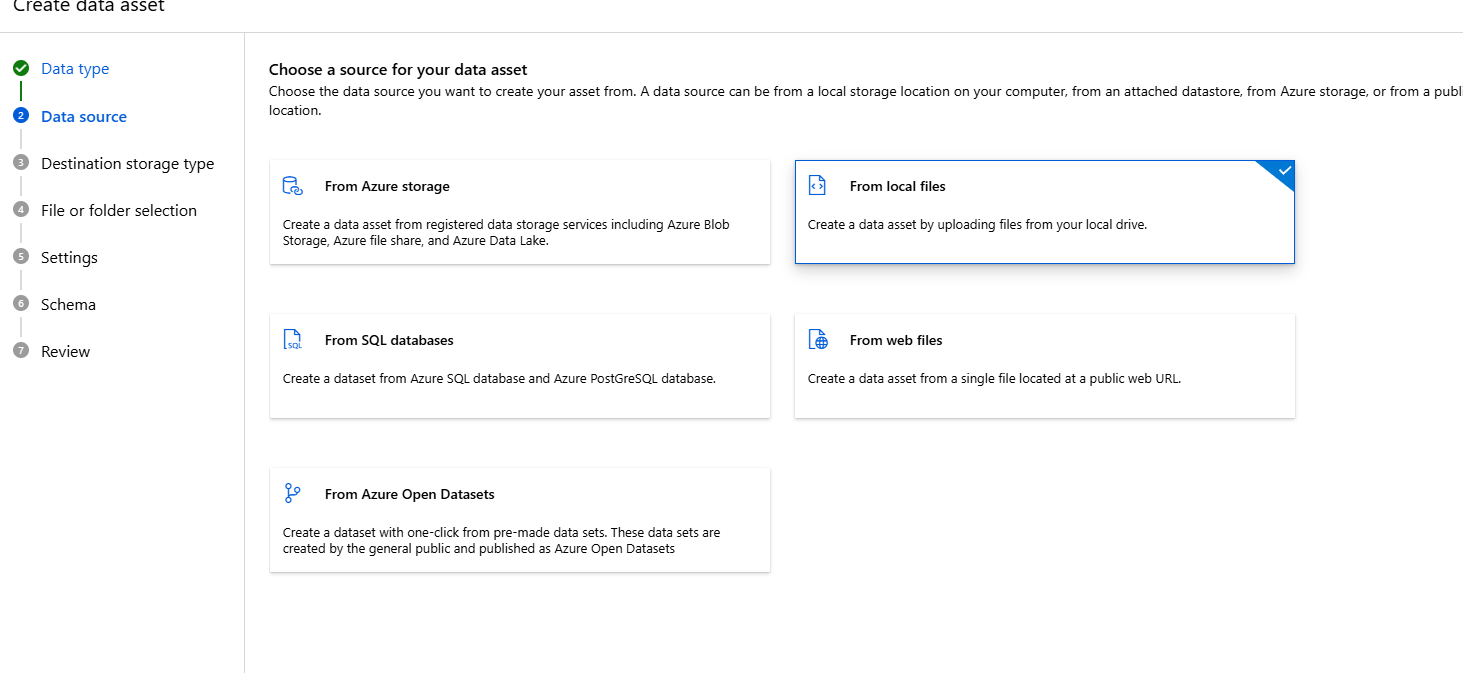
(Lab : 트리기반 알고리즘을 이용한 자동차 가격 예측 모델 구현)

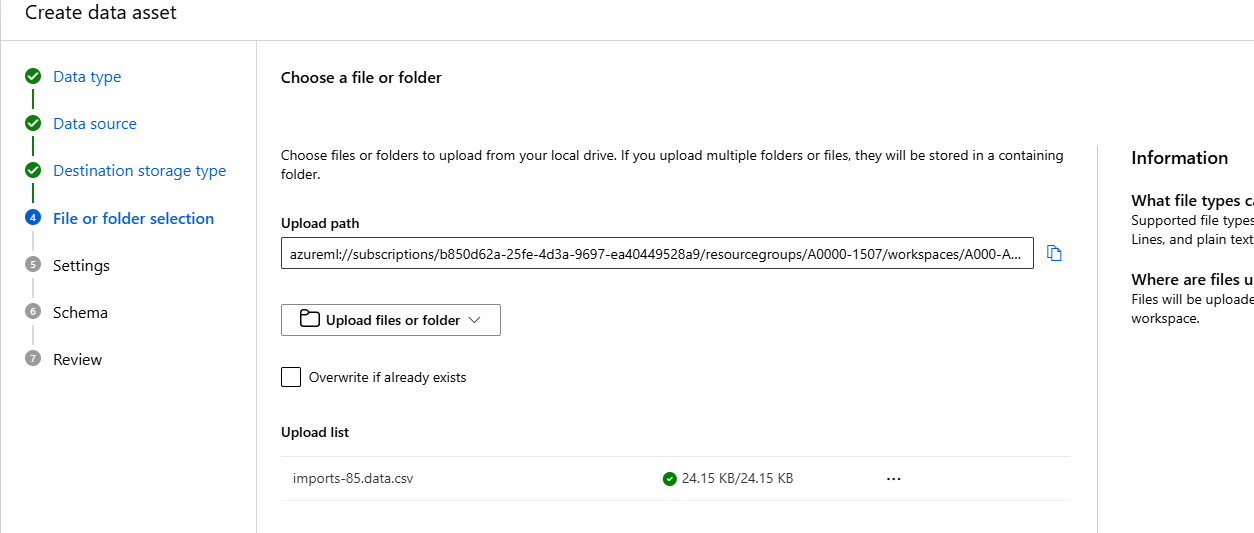
### 실습 준비

archive.ics.uci.edu 접속 후 automobile 검색하여 데이터 다운로드

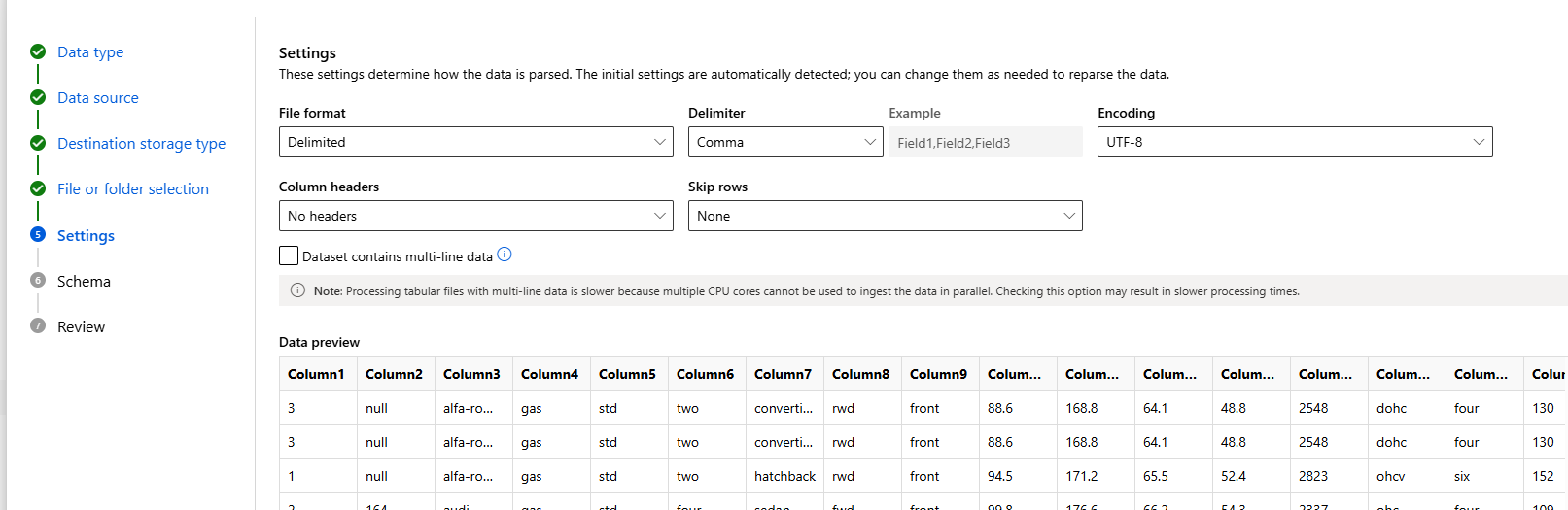
#### 데이터 세트 등록

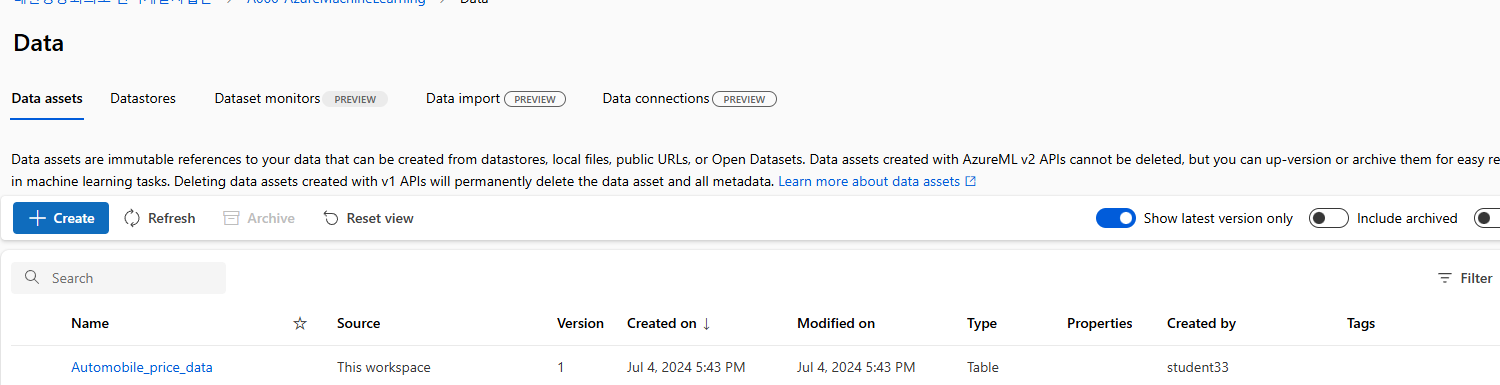




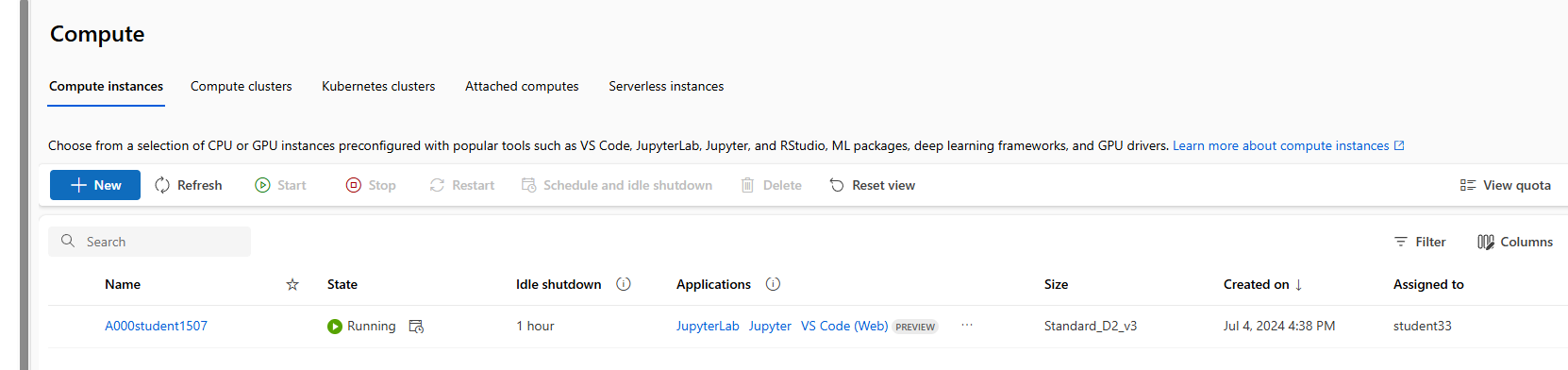


No header로 변경



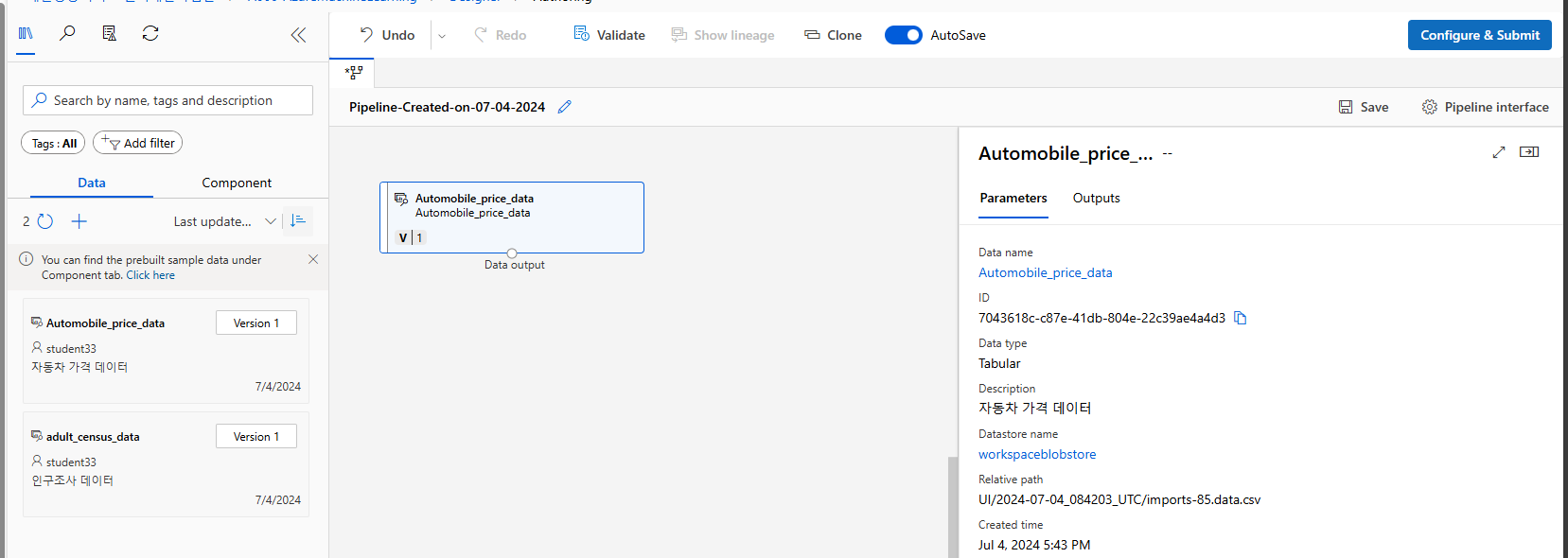


#### 컴퓨트 (위에 만들어둔 것 사용)

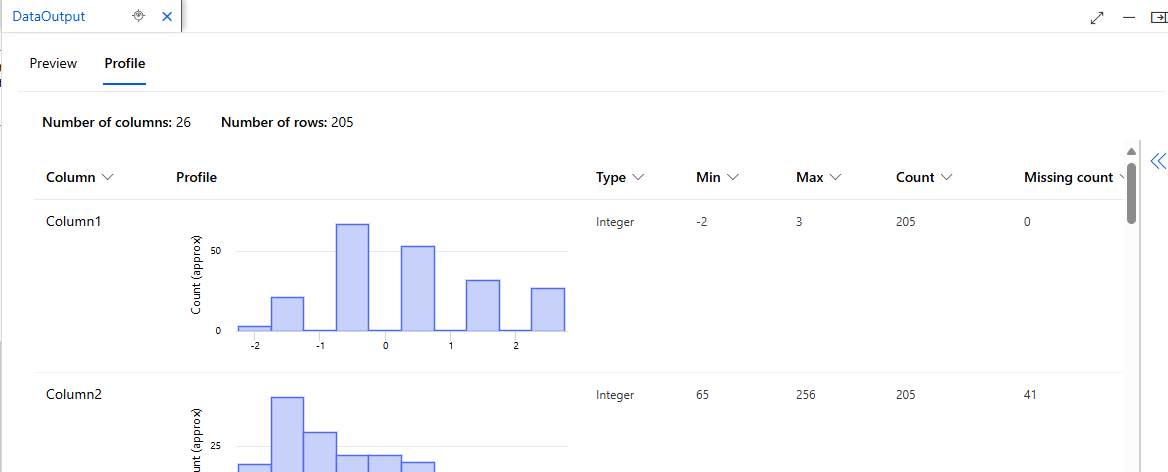


### 데이터 수집/이해

#### 디자이너 탭에서 파이프라인 생성 및 데이터 세트 가져오기

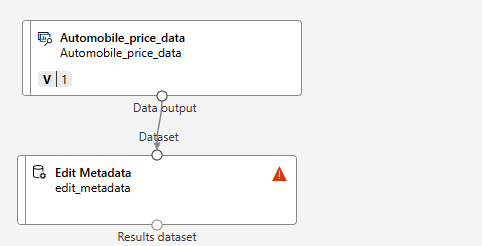


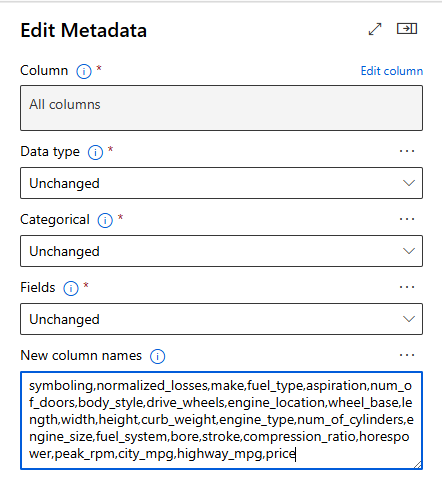
#### 데이터 이해



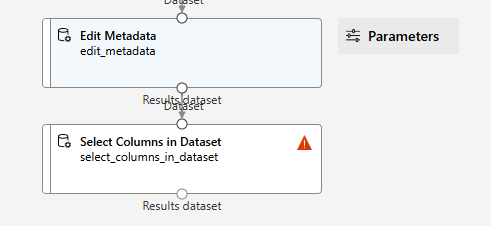
### 데이터 준비

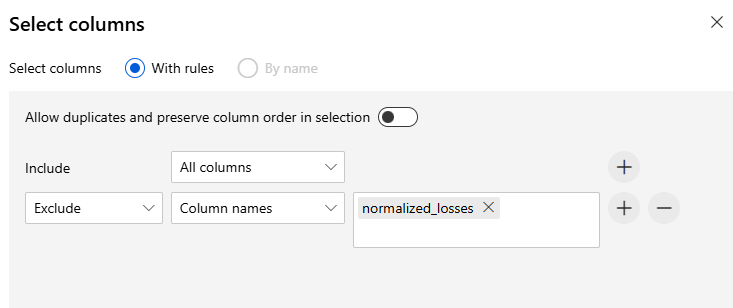
#### 특성 이름 지정



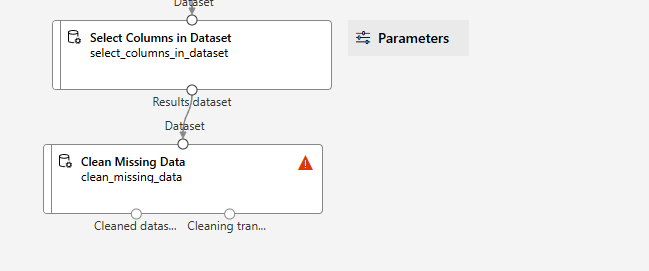


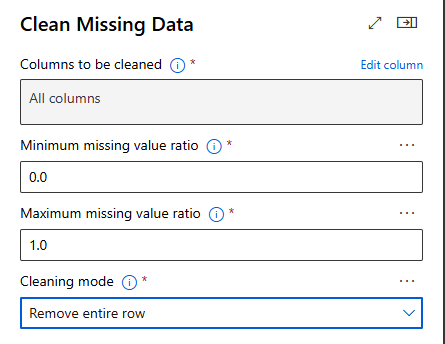
#### 특성 선택 – 불필요한 컬럼 제외



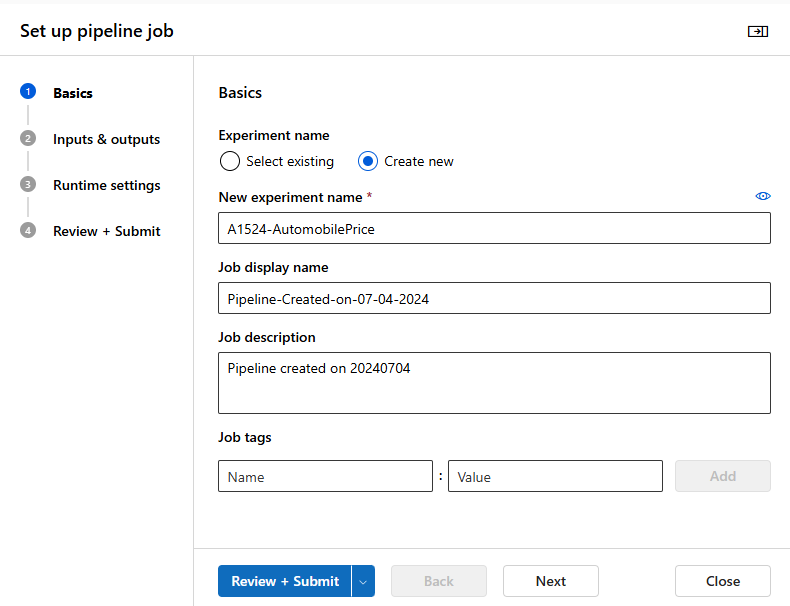


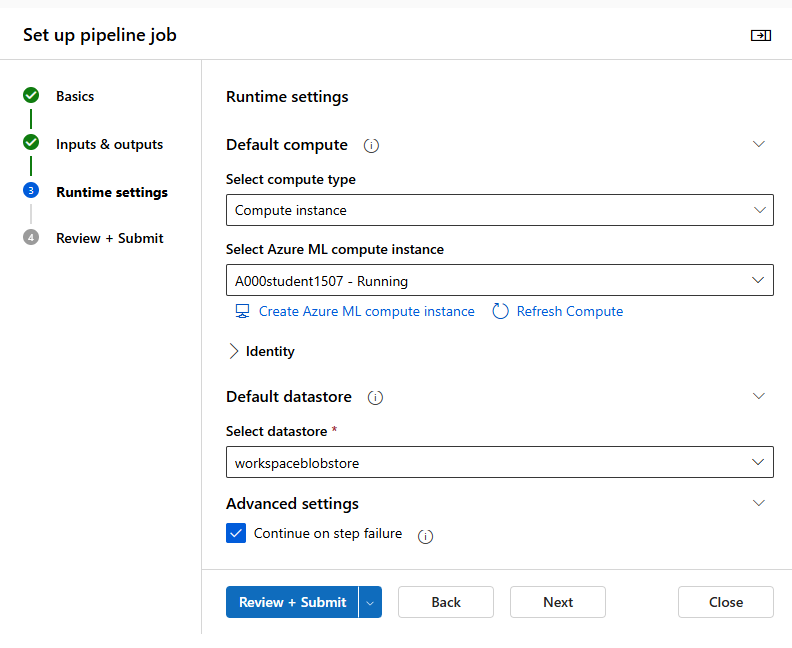
#### 누락값 처리

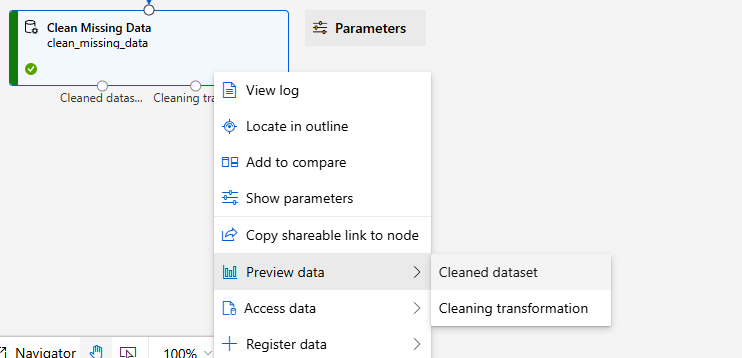


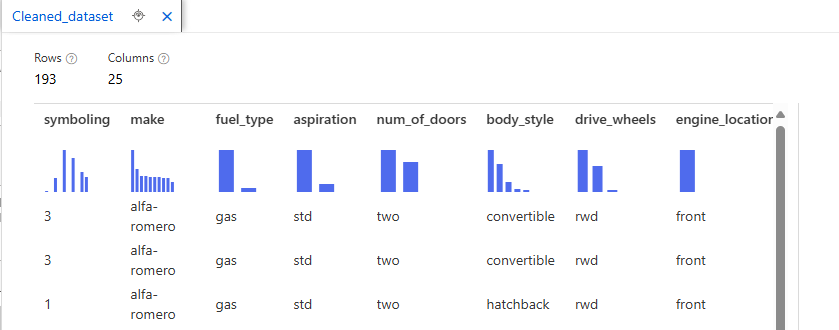


#### 중간 점검

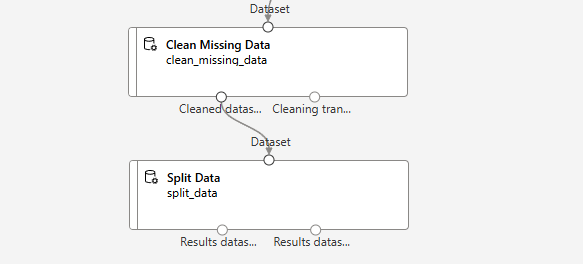


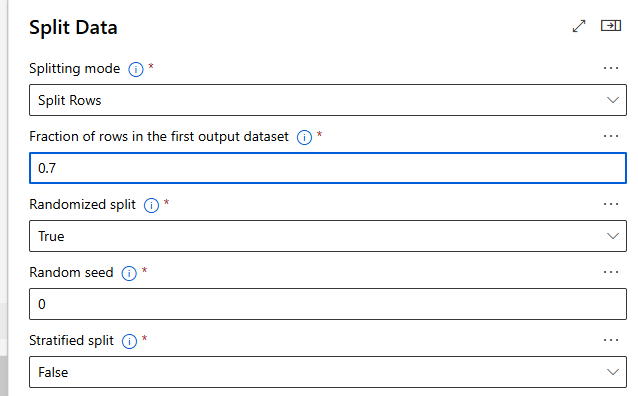






#### 데이터 분리

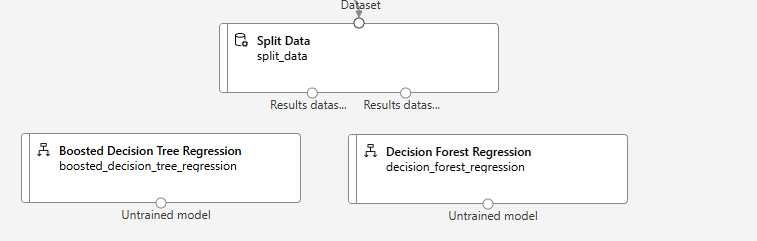




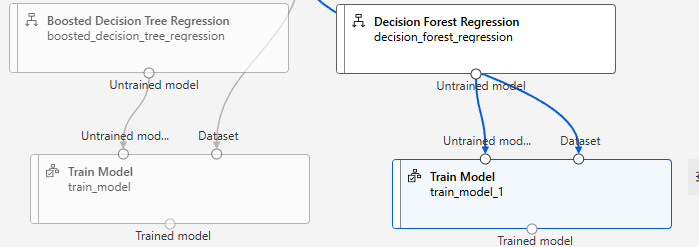
### 모델링/평가

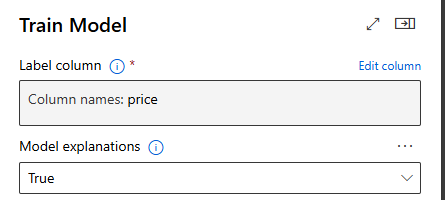
#### 모델링 알고리즘 선택

두 가지 알고리즘 사용 (기본값)

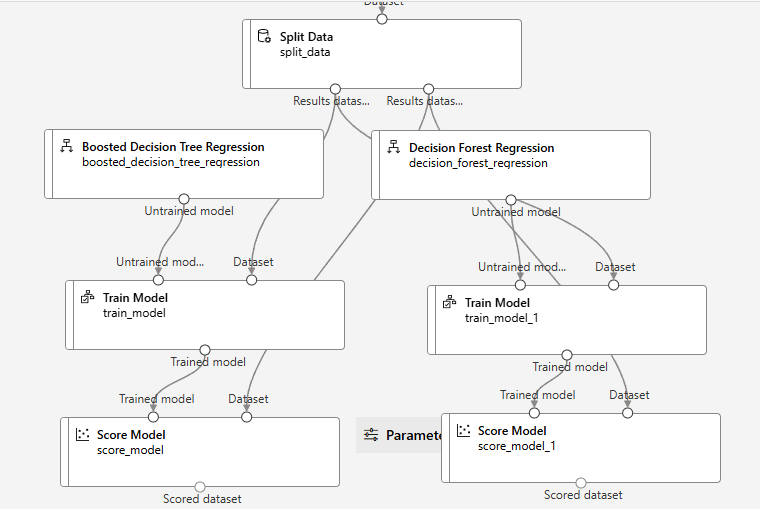


#### 모델 훈련(학습)

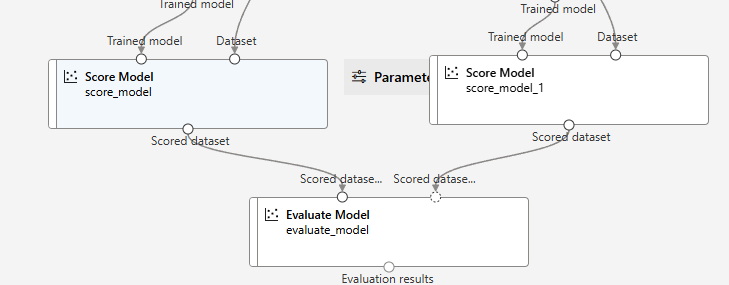




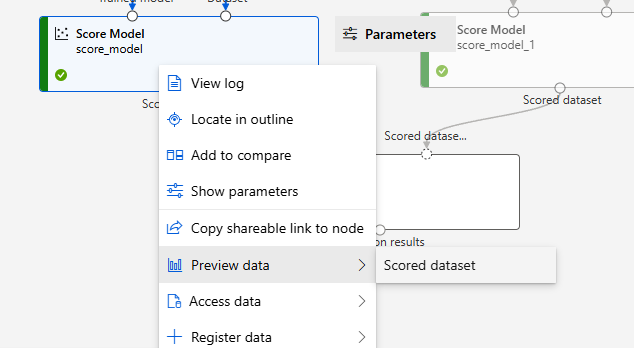
#### 모델 테스트

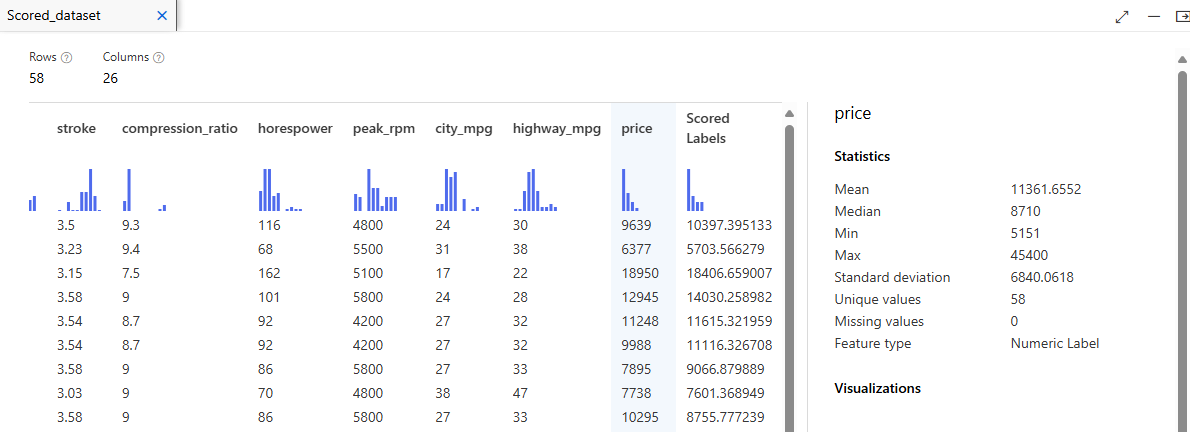


#### 모델 평가



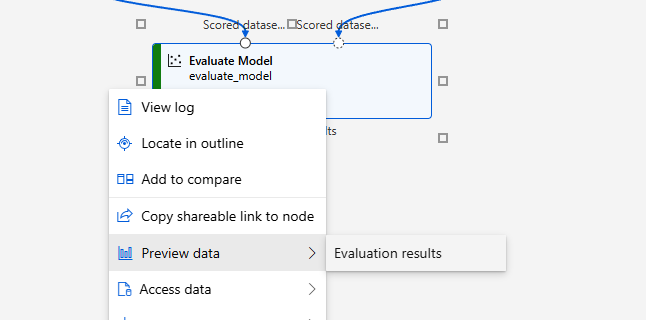
Boosted decision tree의 score model

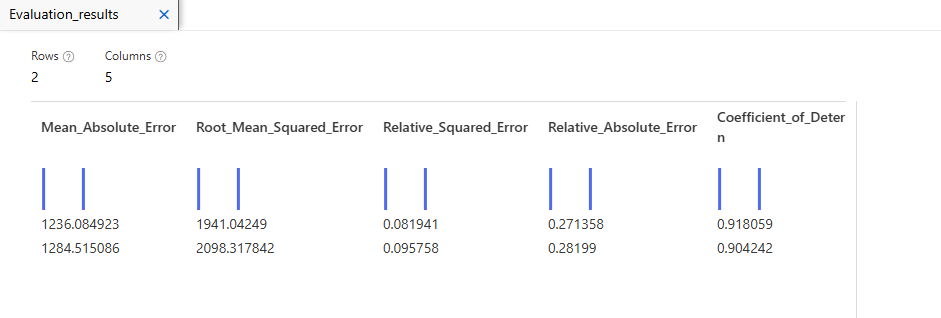




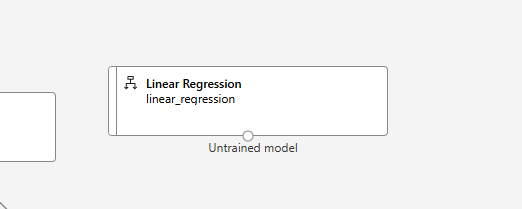
Decision Forest Regression의 score model



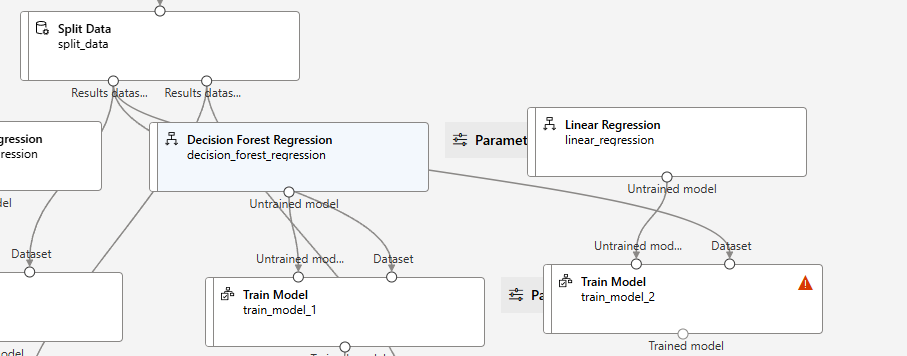


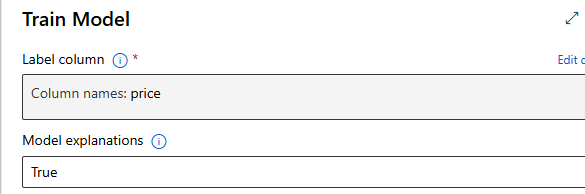


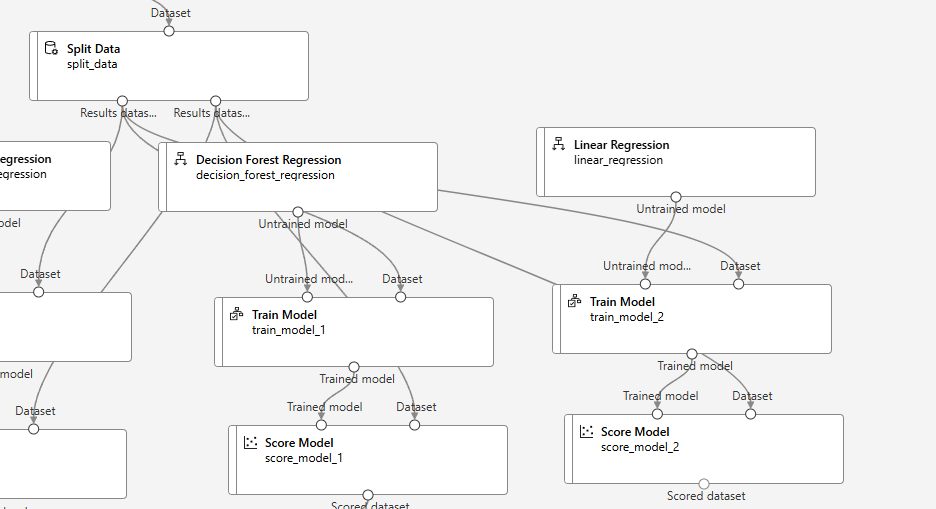
### 추가 실습 – 세 개 이상 알고리즘 비교

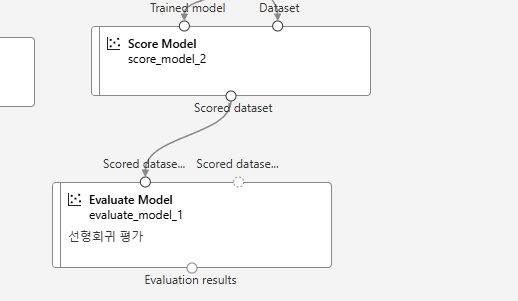


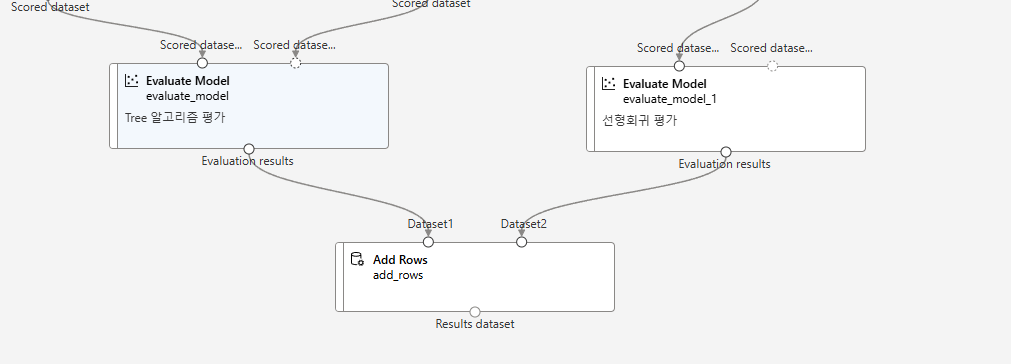
※ 선형알고리즘은 그냥 train model, 군집알고리즘이 train cluster model

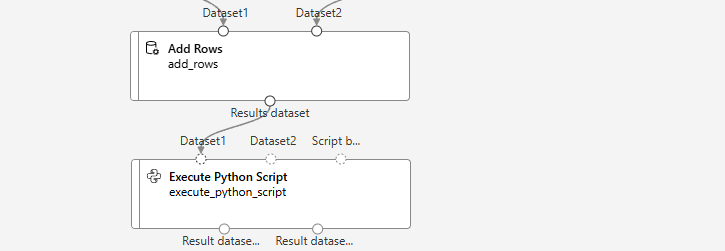




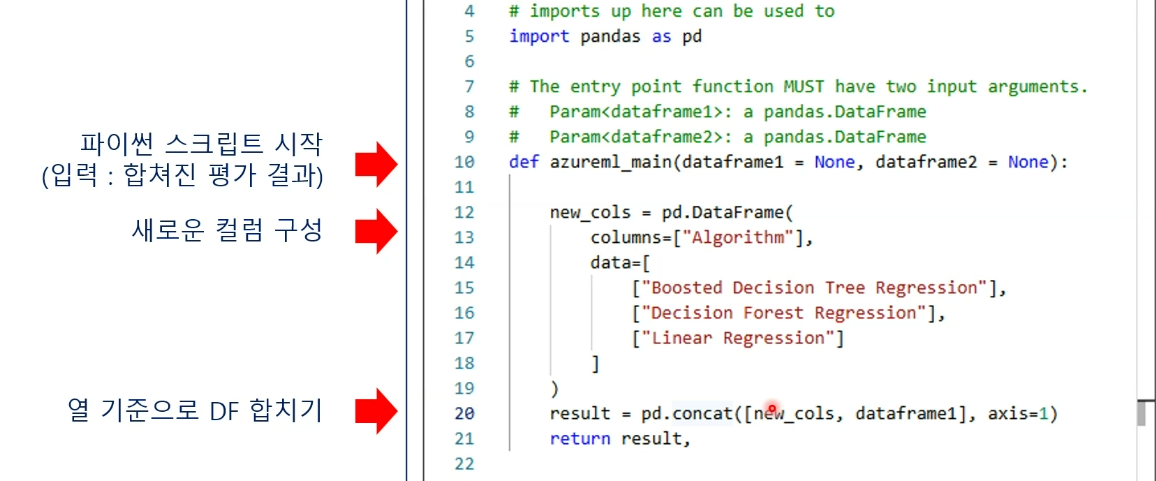


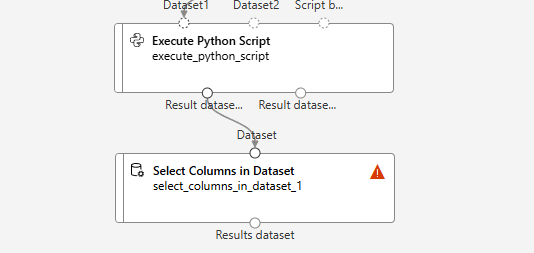




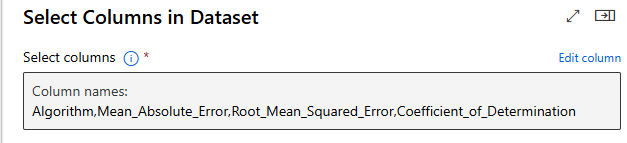


Python 코드 작성





Select Columns in Dataset 컴포넌트의 설정창에서 MAE, MSE, 결정계수에 해당하는 컬럼명을 지정



#### Job 결과 확인 – 평가 취합 결과

