# <오전>

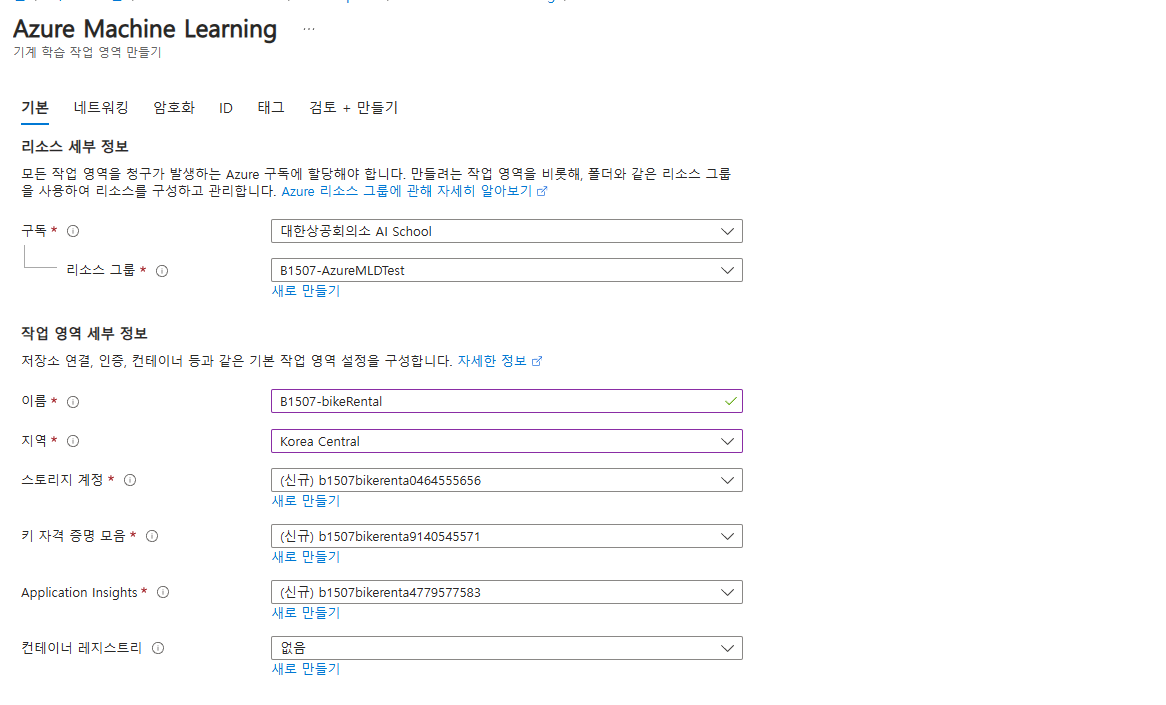
## [환경]

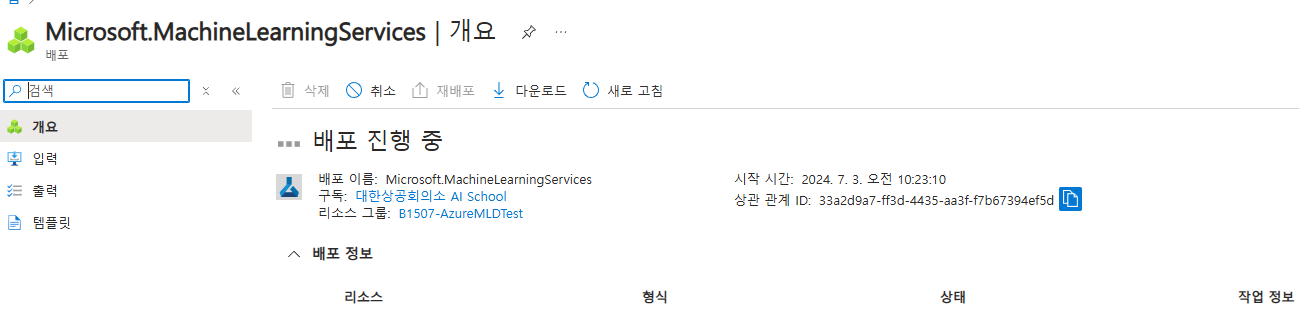
* Azure
* .csv 파일

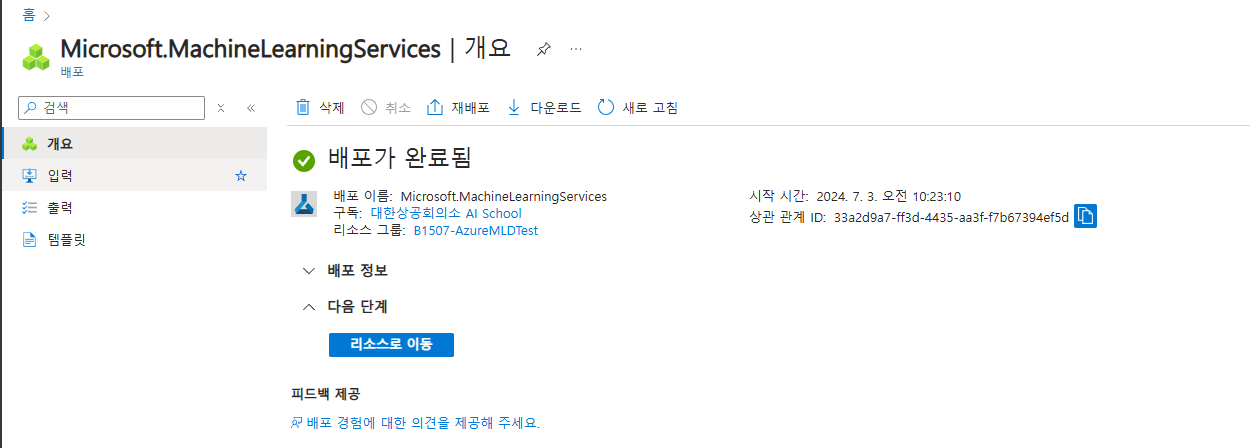
## [실습 - 회귀MLD\_자전거렌탈]



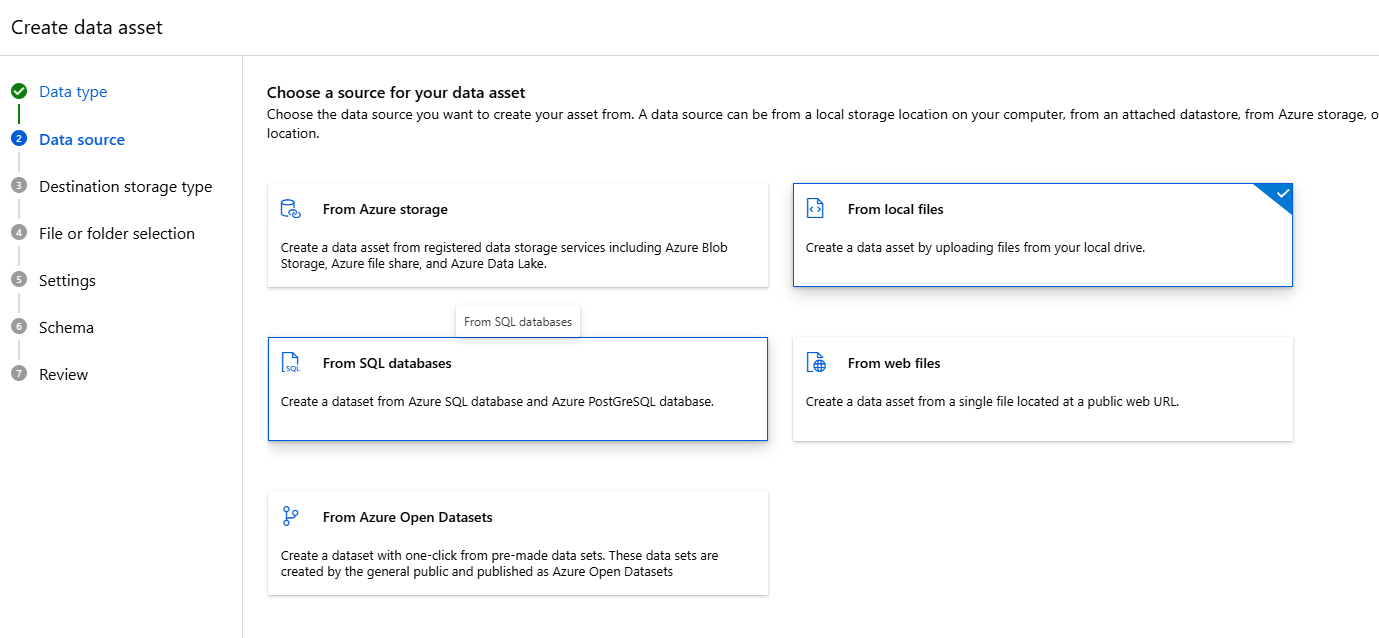


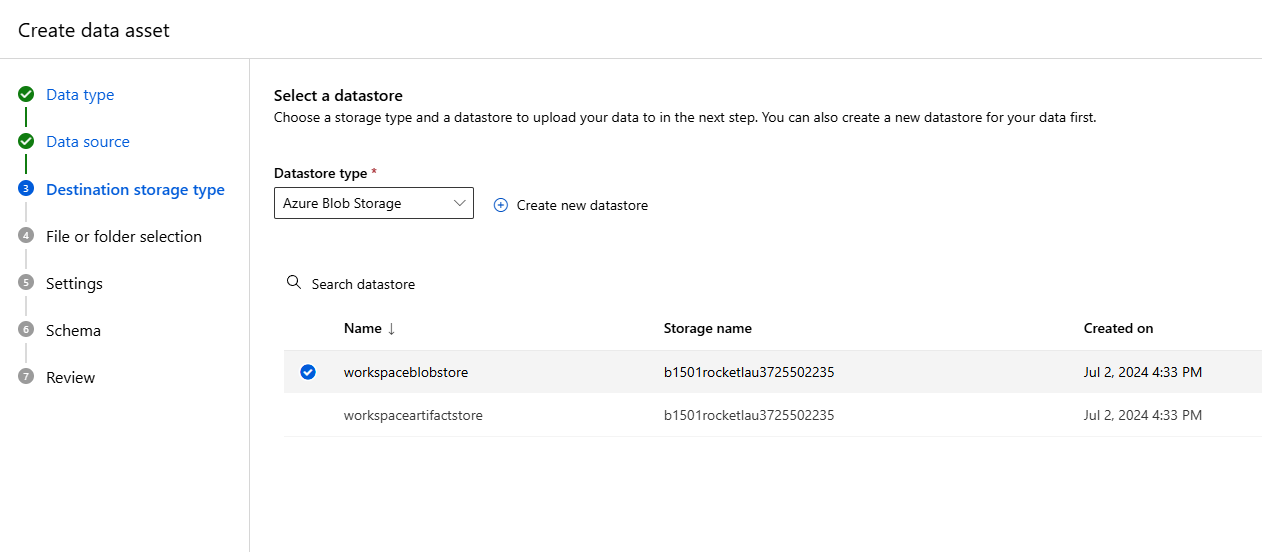


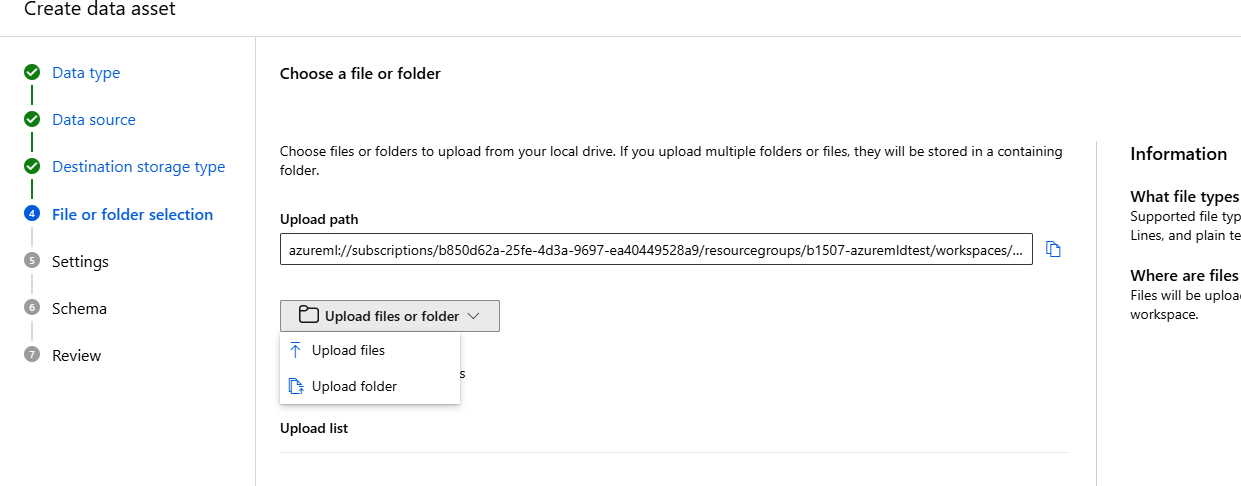


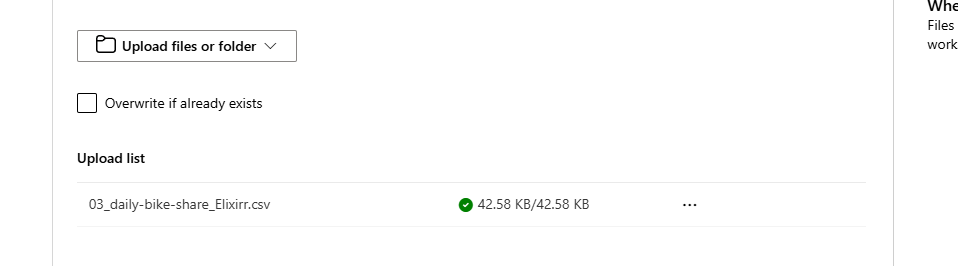


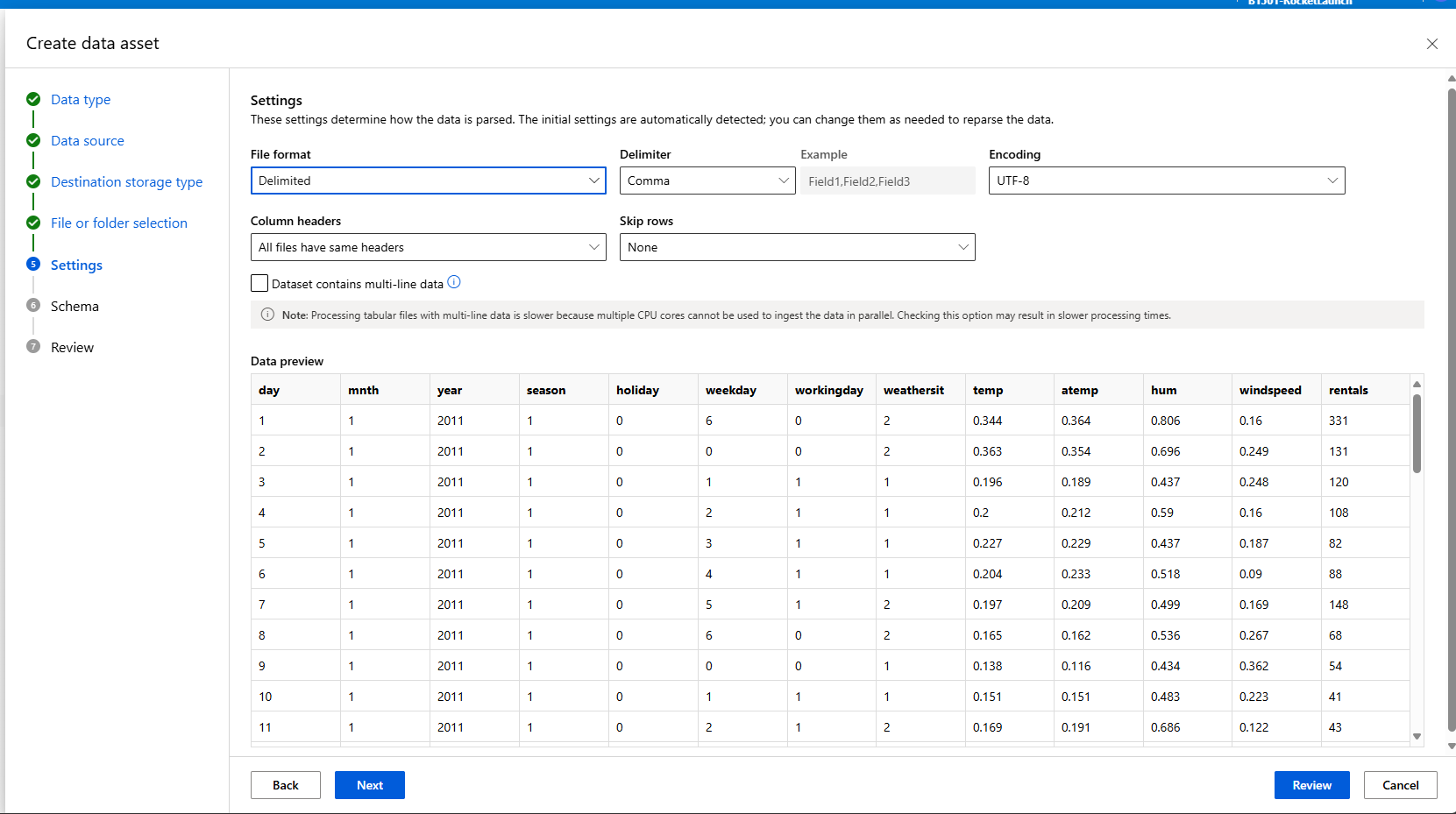


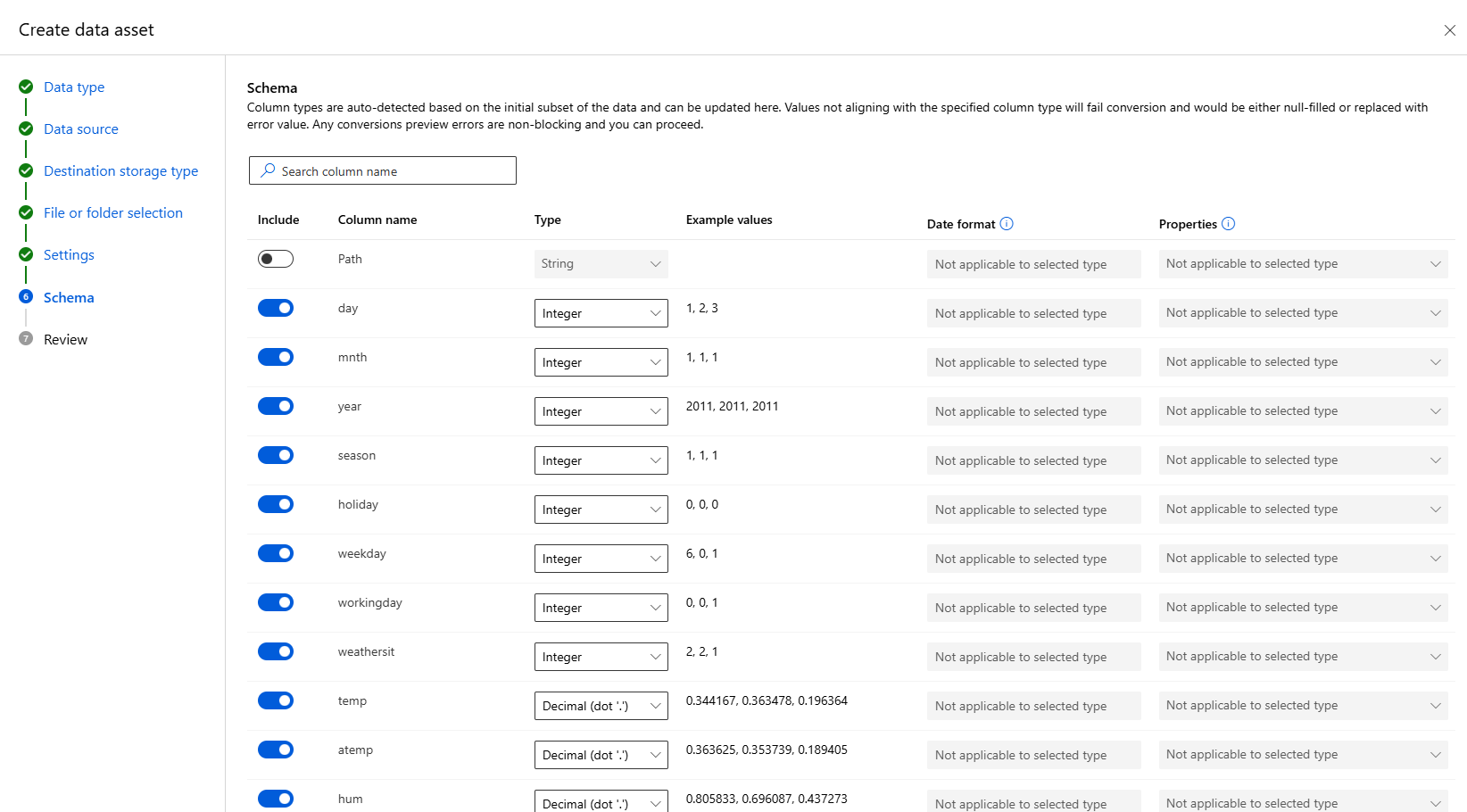


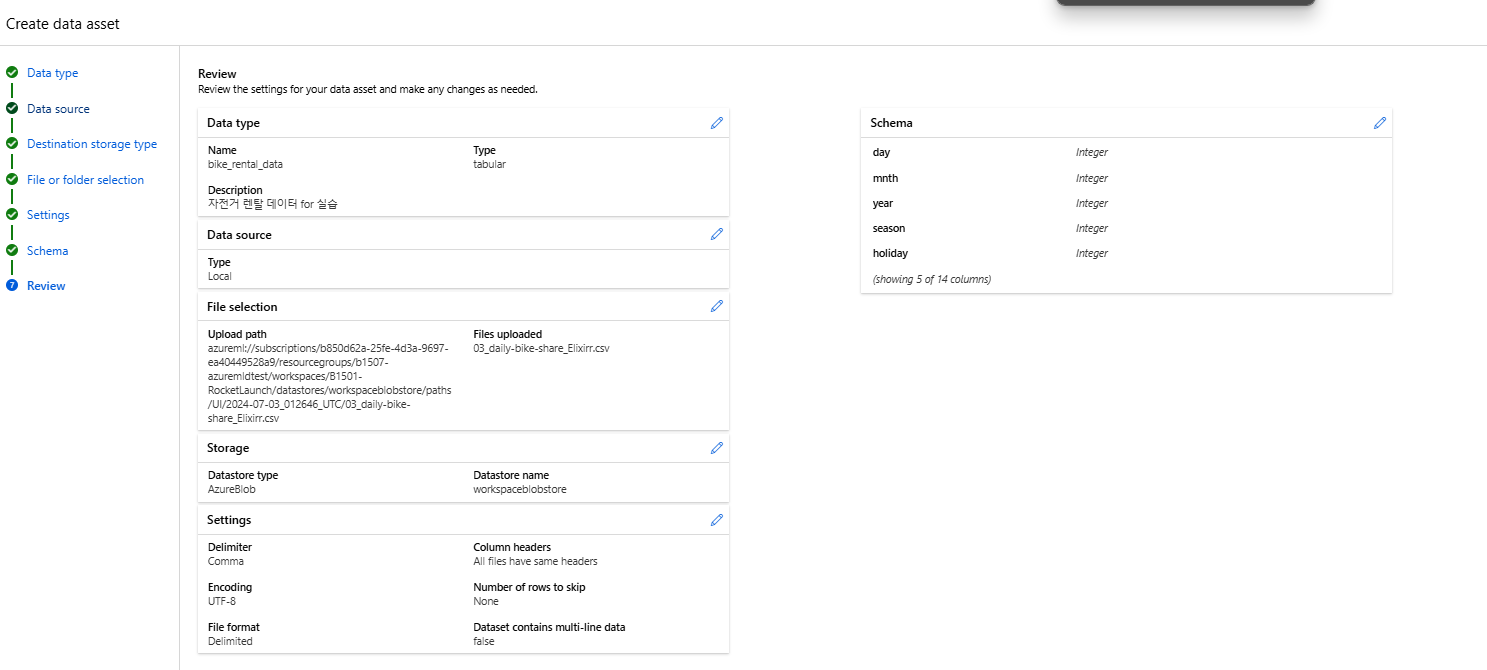


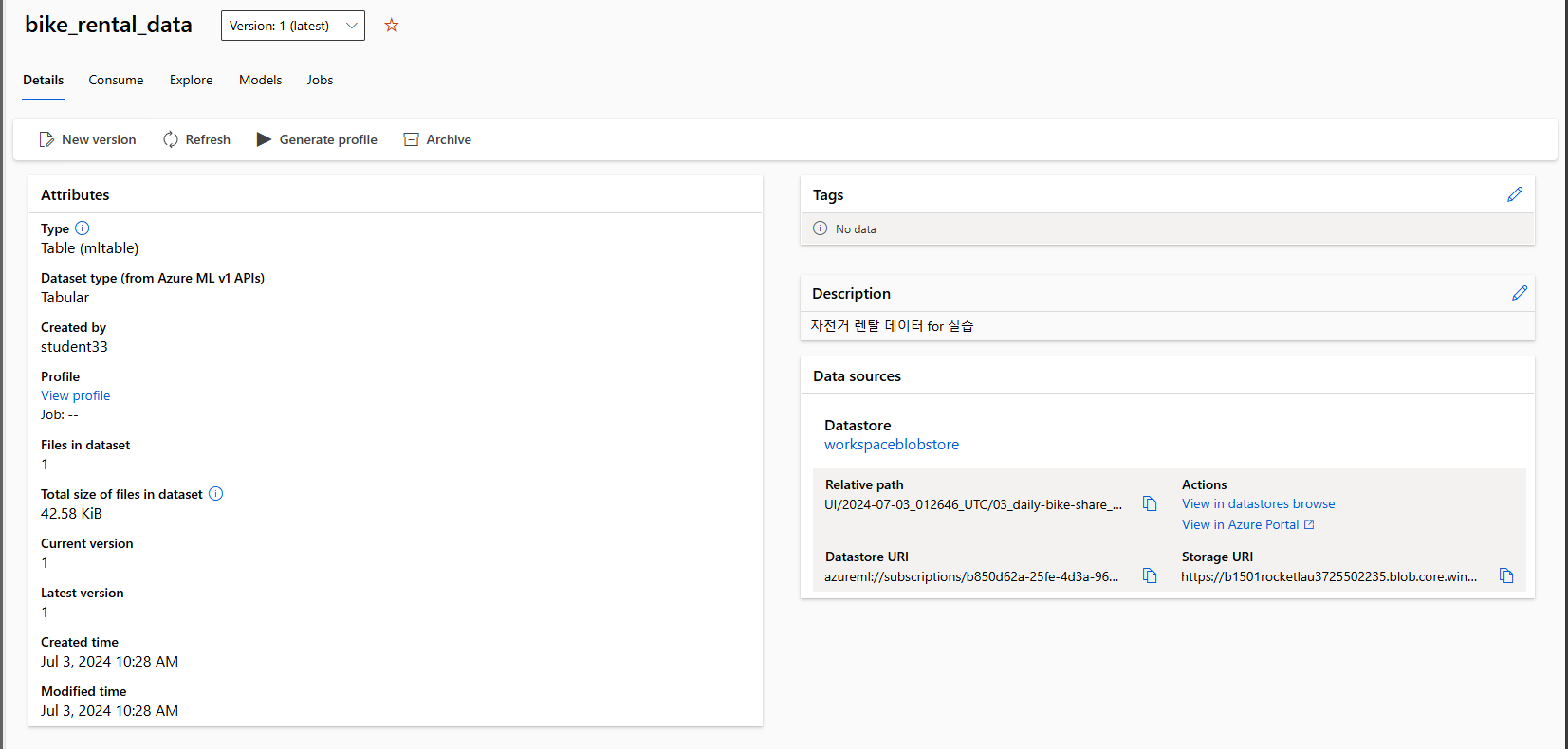




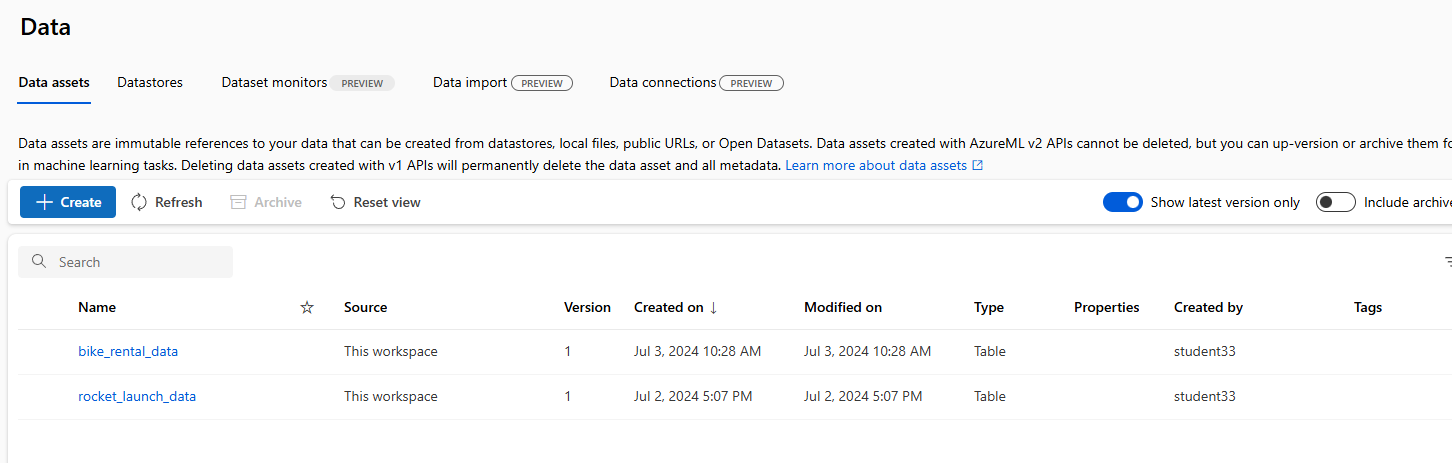




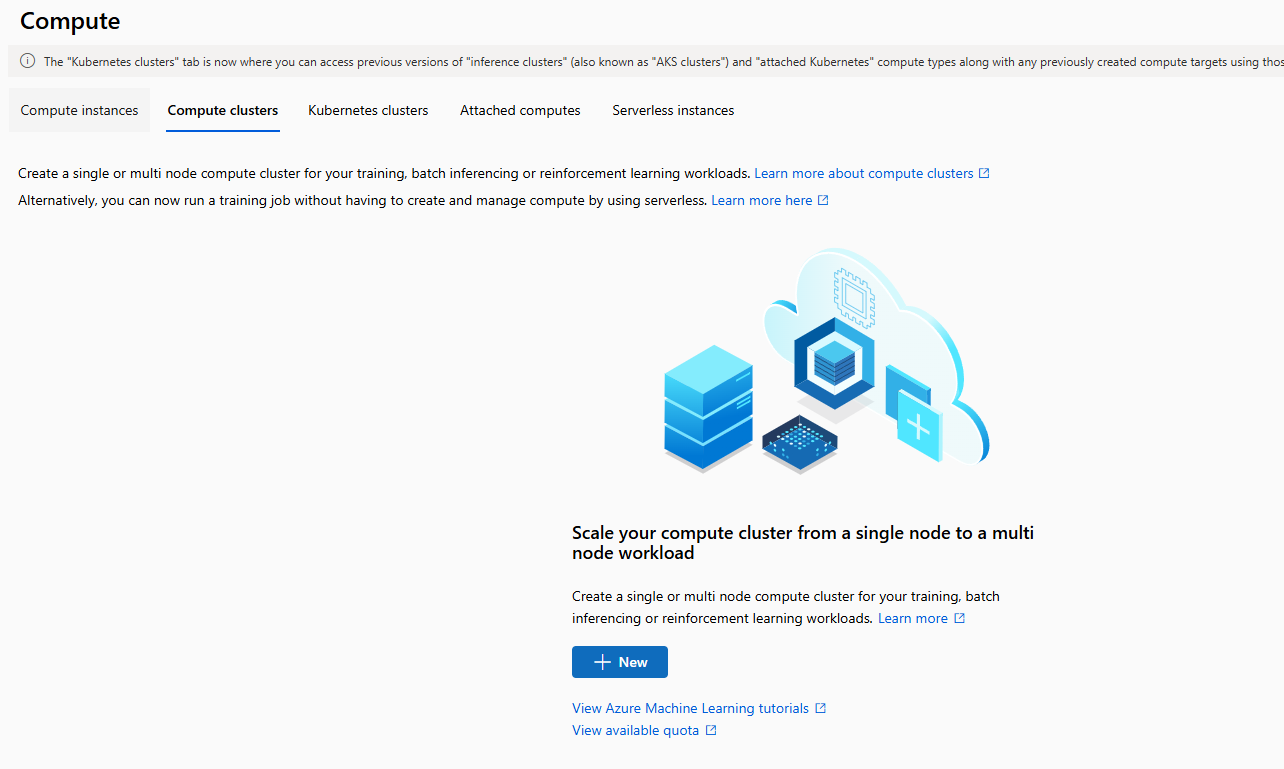


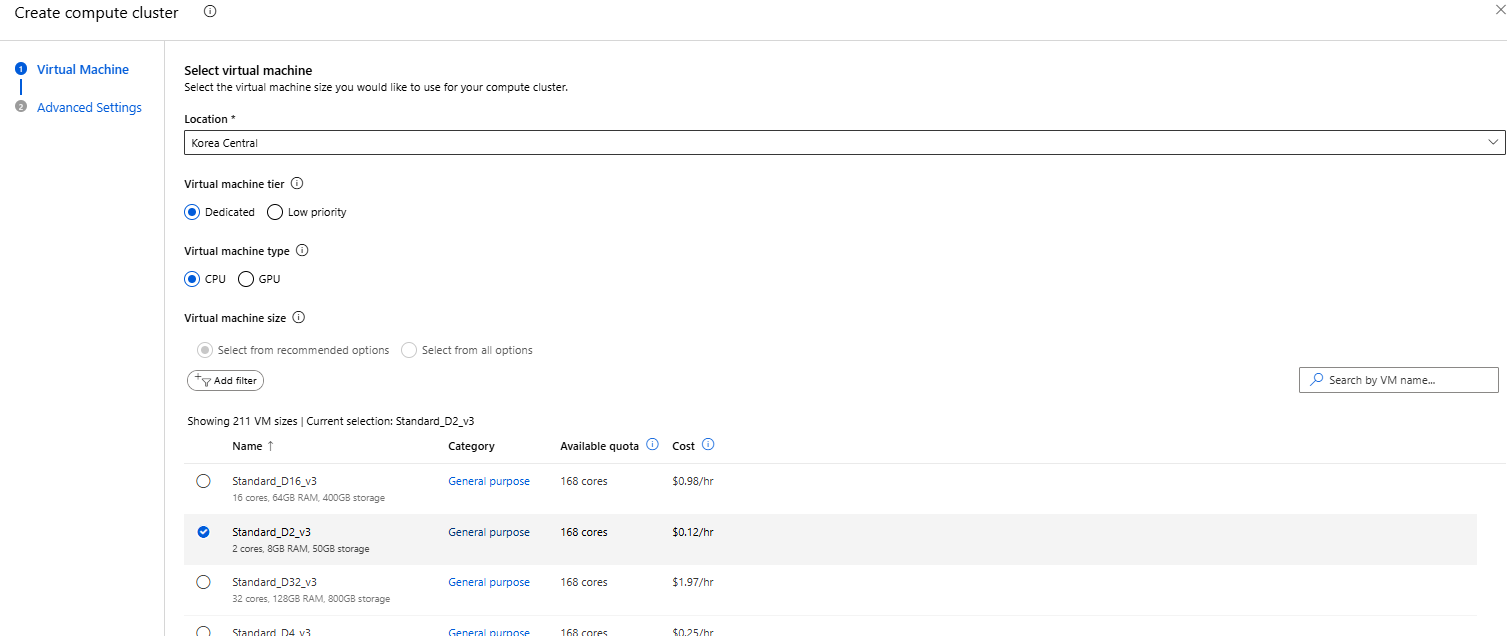


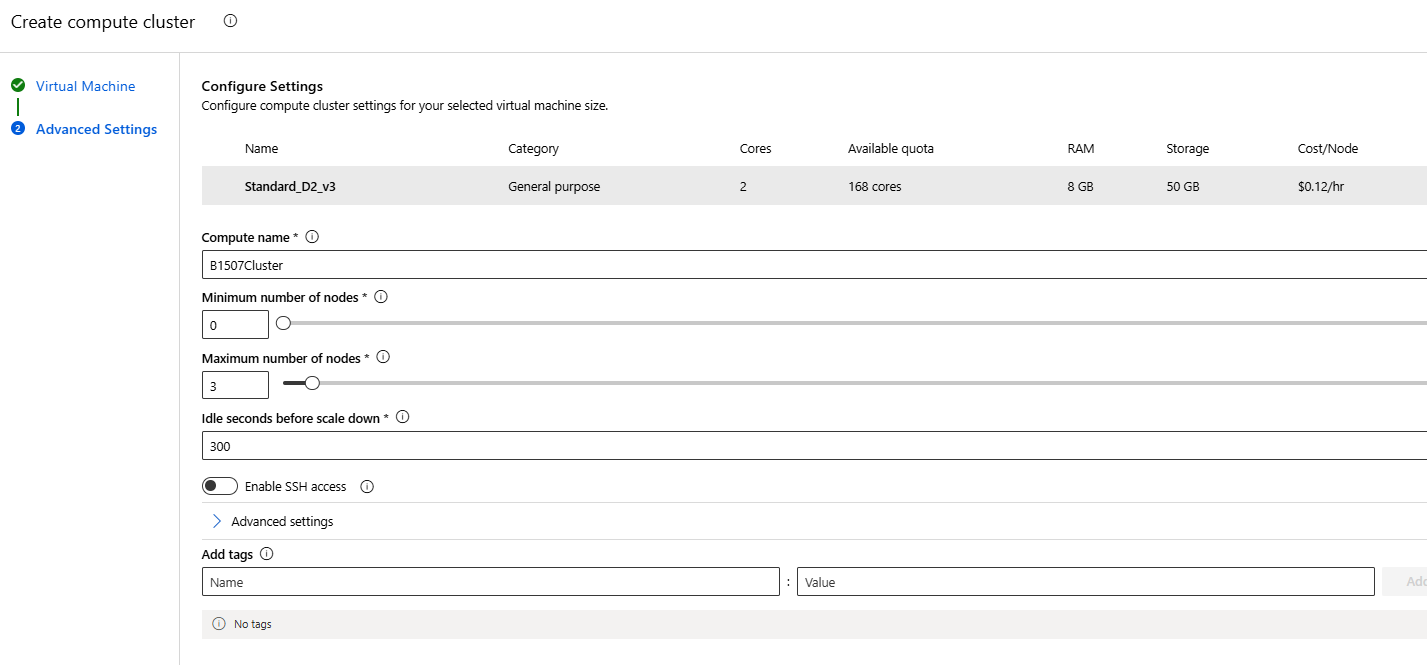
### 데이터 생성 확인

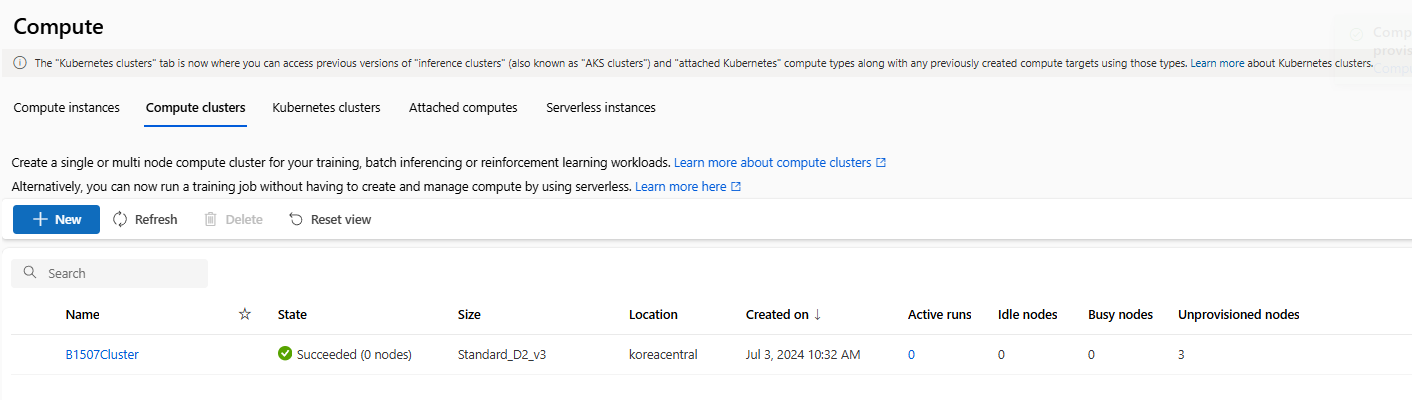


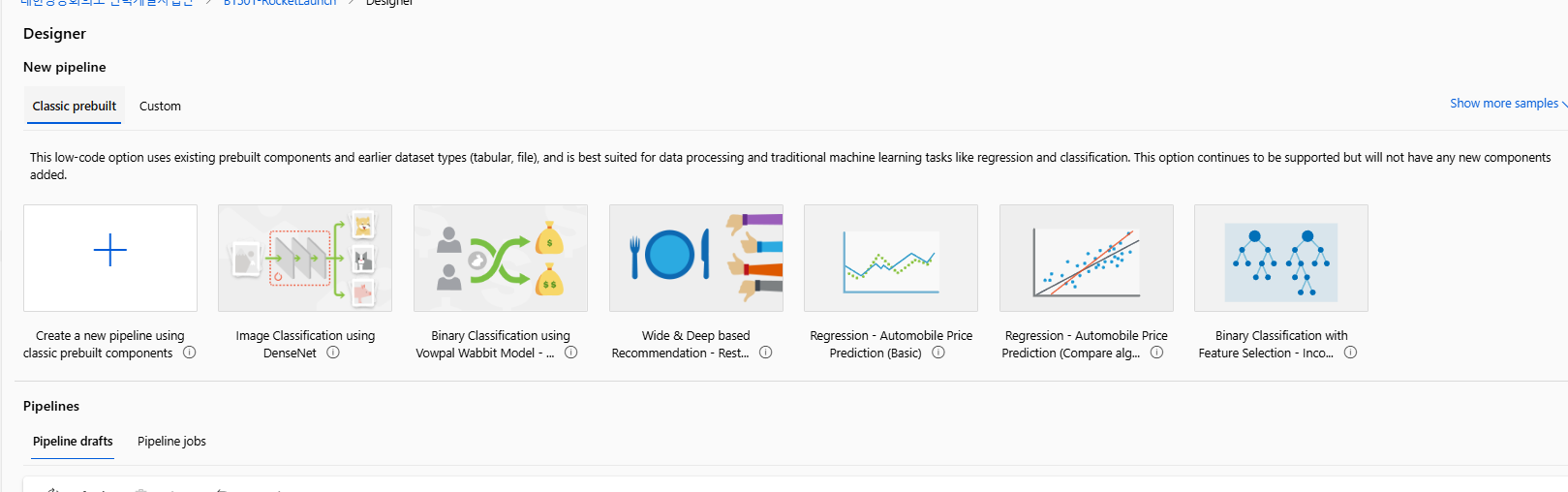
Compute에서 clusters 사용



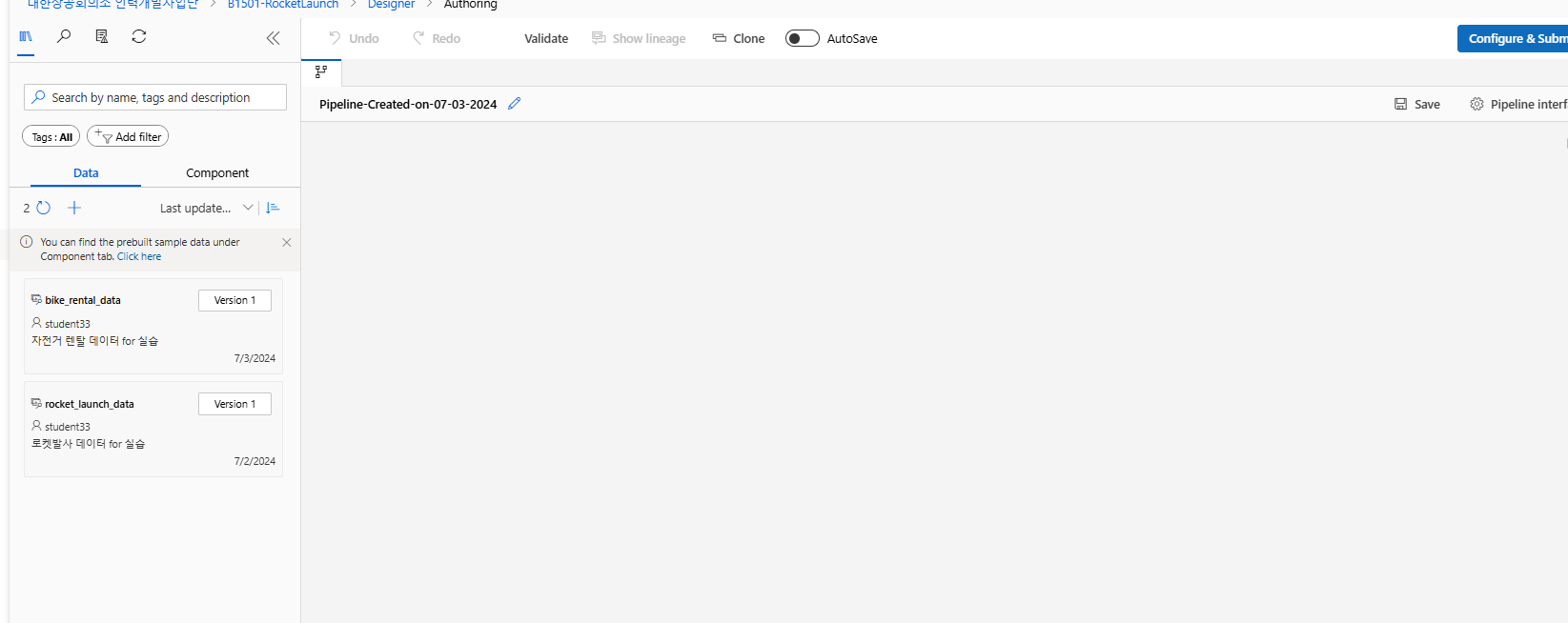


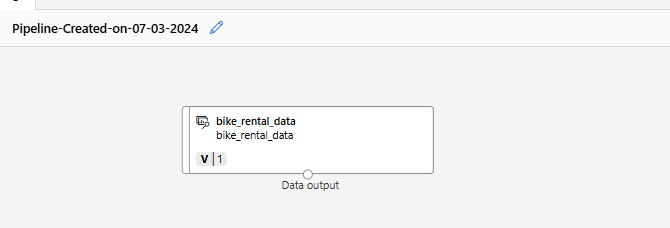


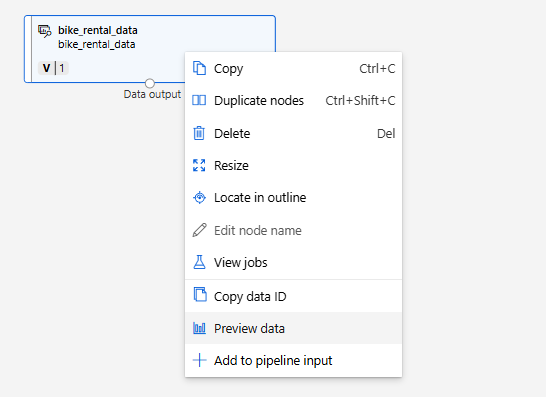


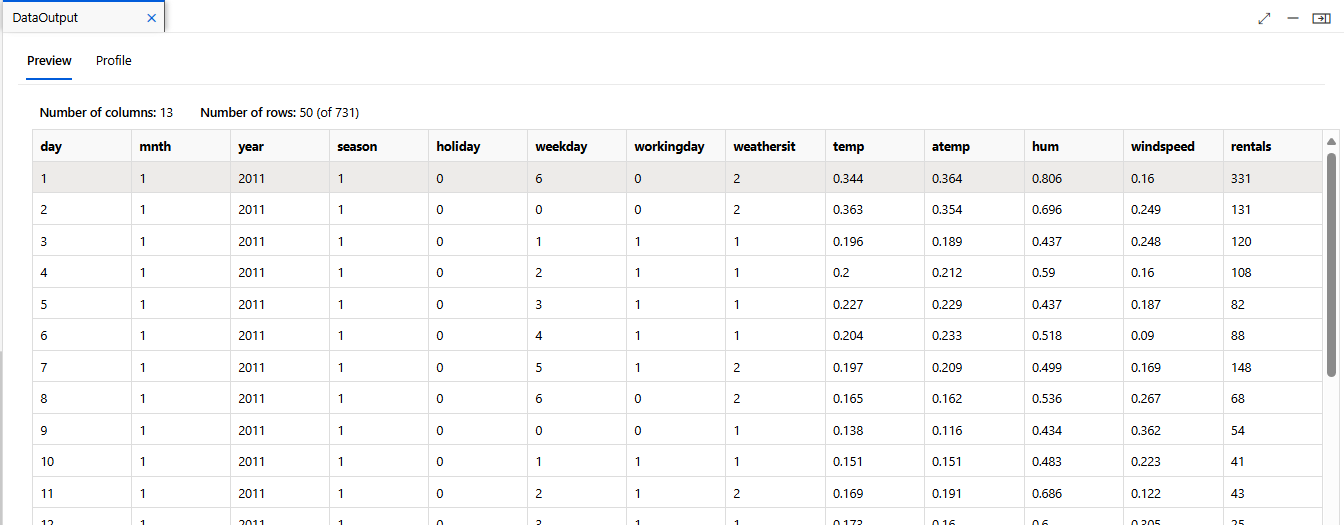


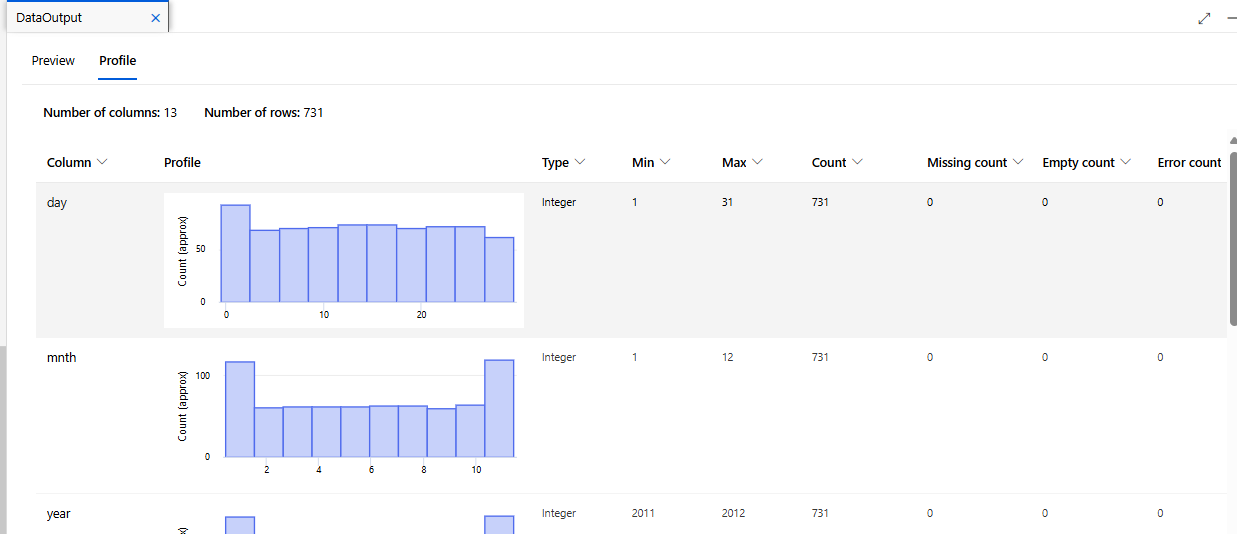
### 데이터 세트 가져오기



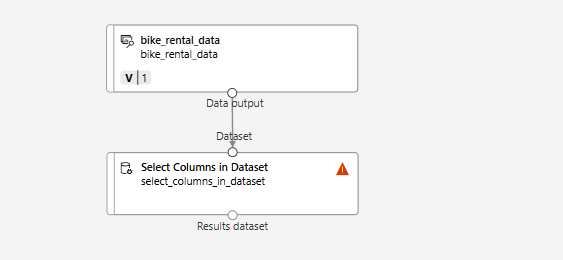


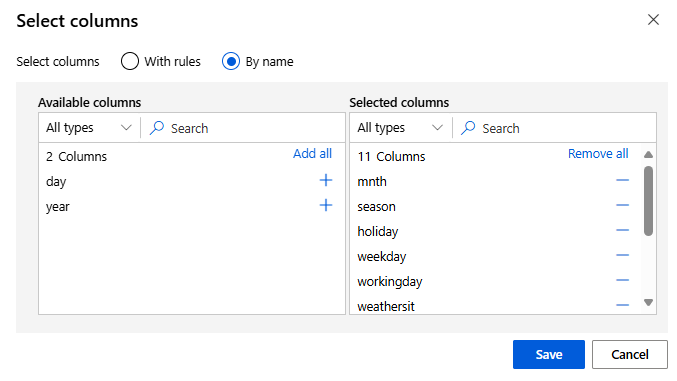


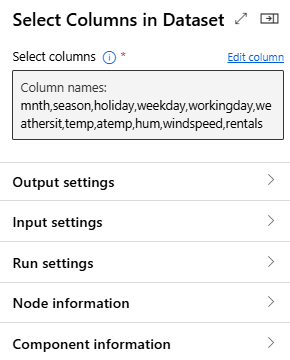




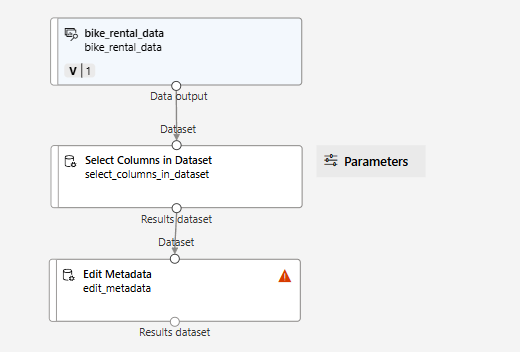
### 특성 선택

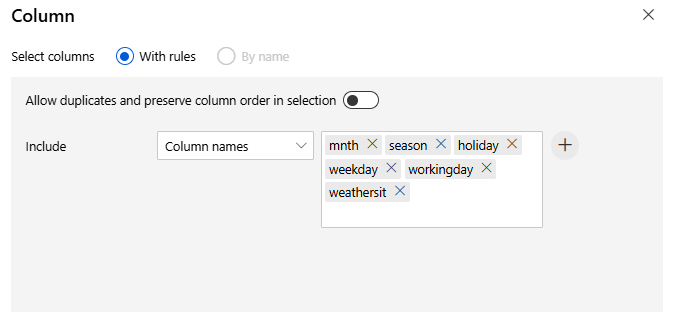




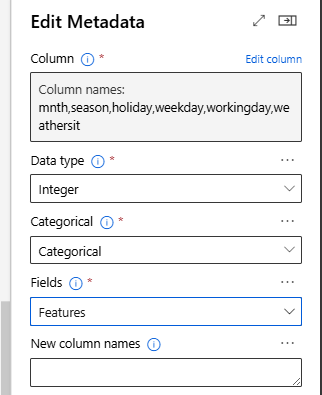


### 범주형 데이터 변환

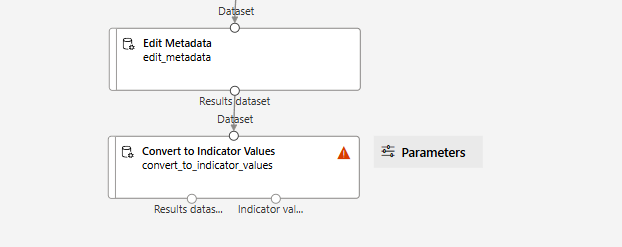


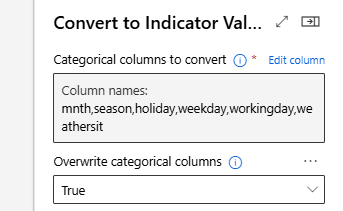


정수형 범주값 -> 가테고리형

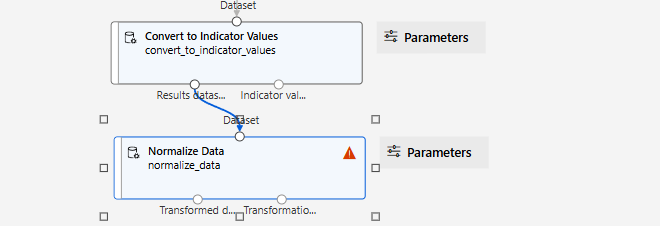


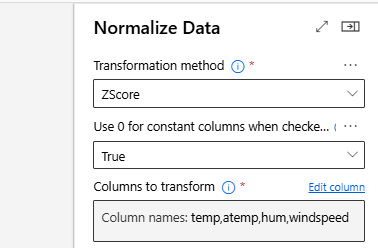
카테고리형 컬럼에 지시값 할당





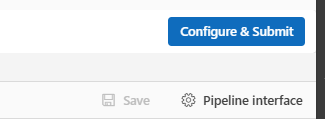
### 표준화와 정규화



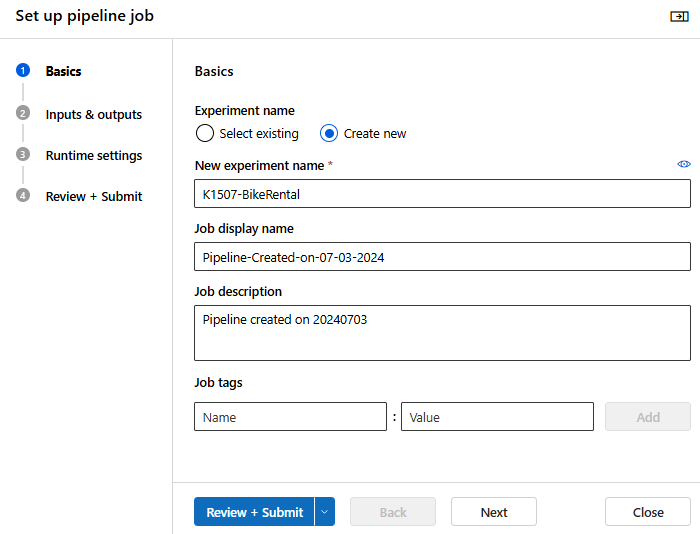


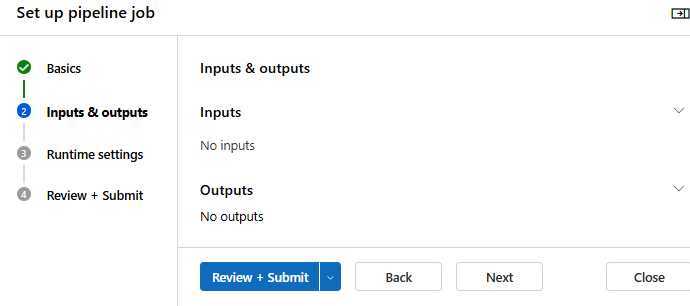
### 중간점검

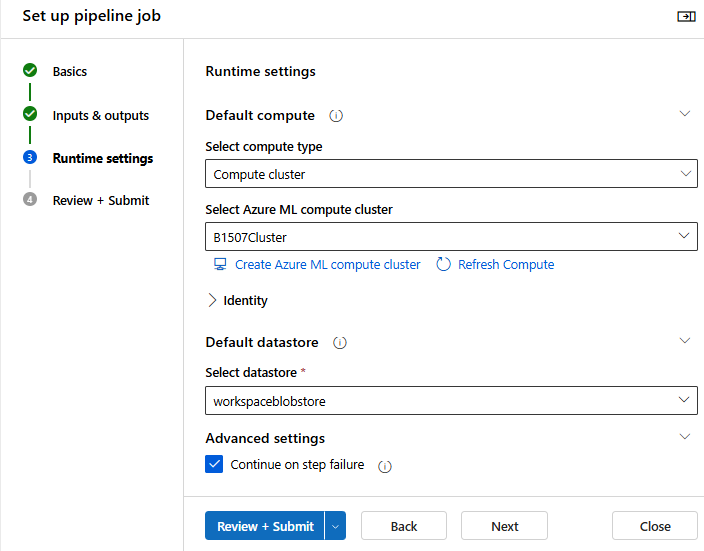
Save 후 Configure & Submit



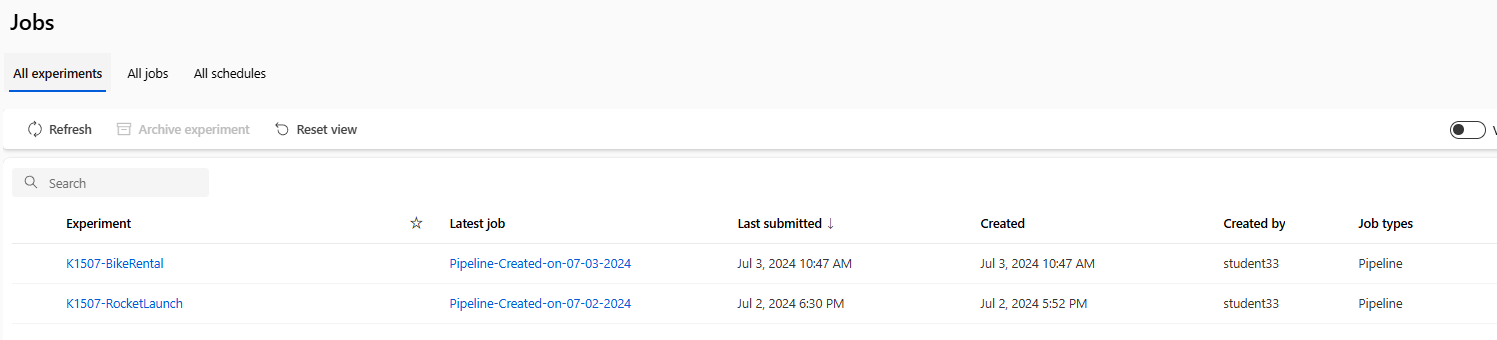
처음이기에 create new 선택

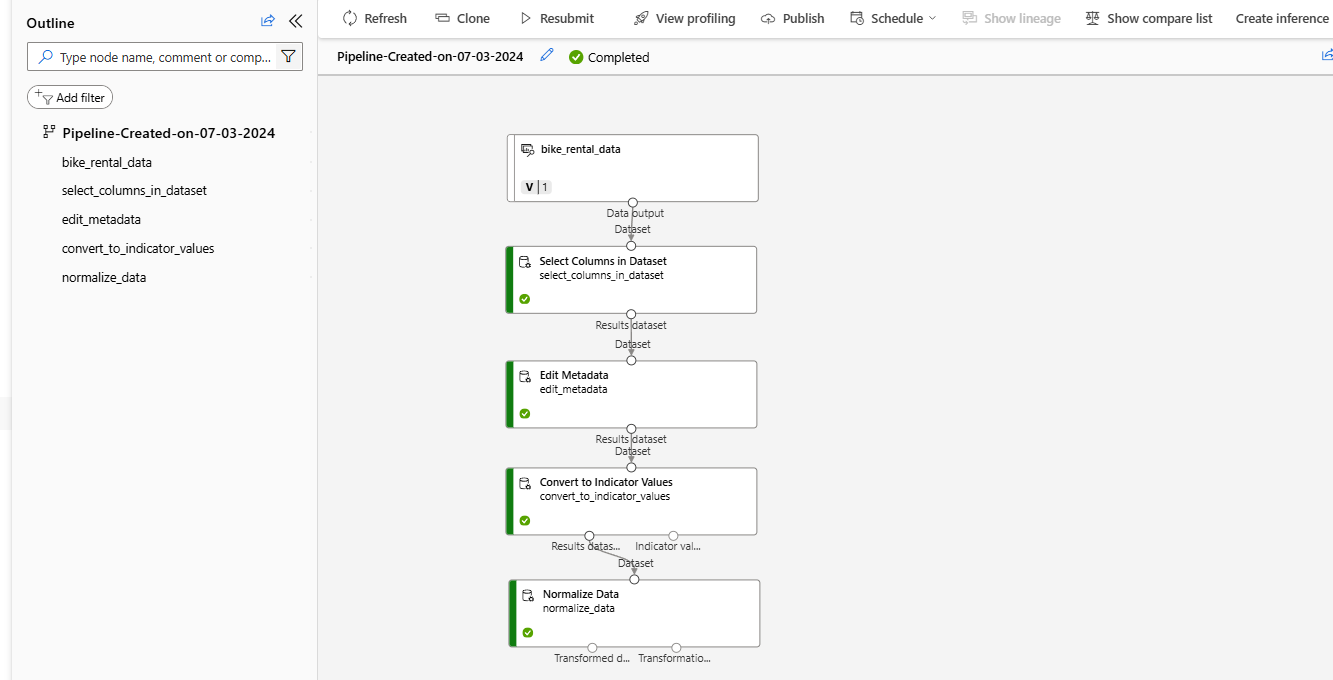




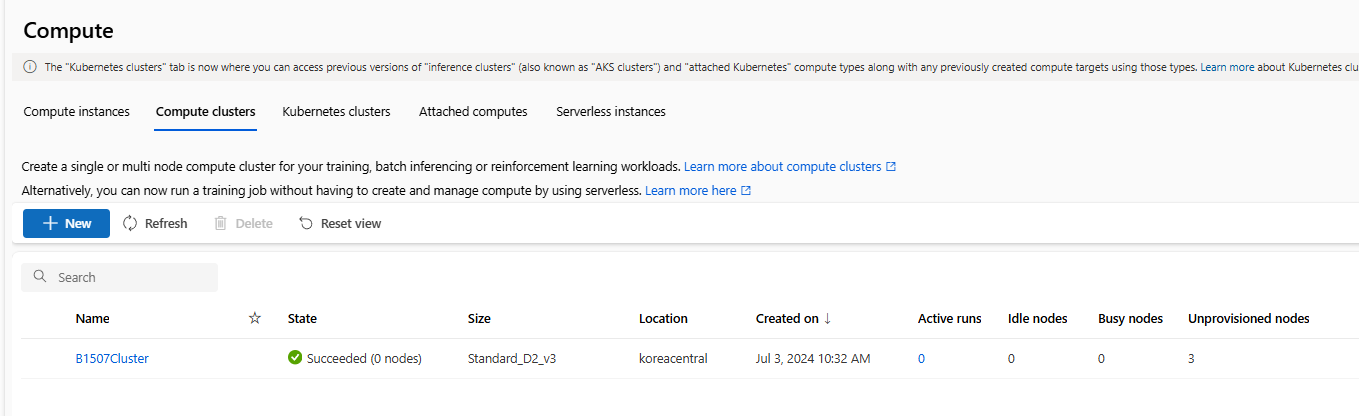


Job에서 확인

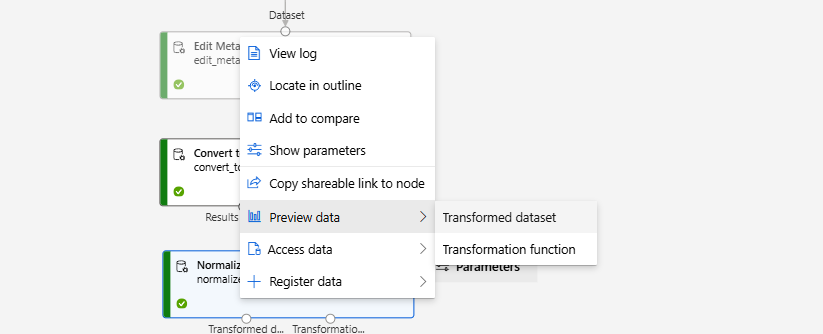


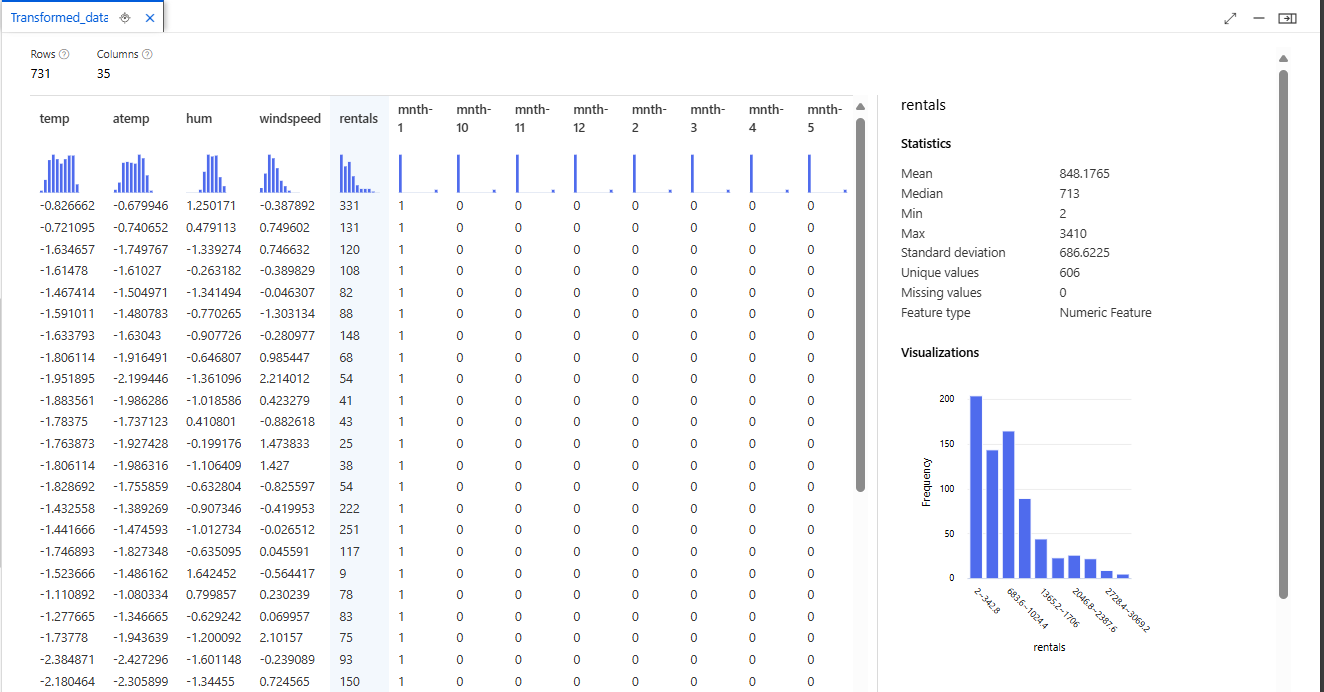


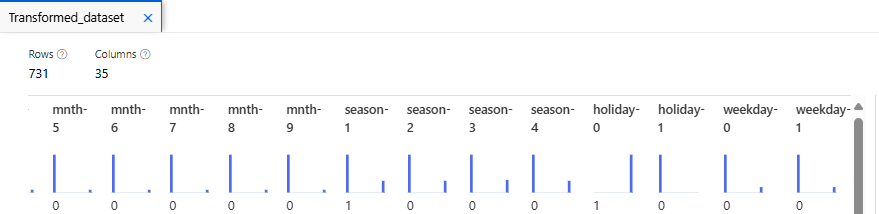
Compute에서 확인



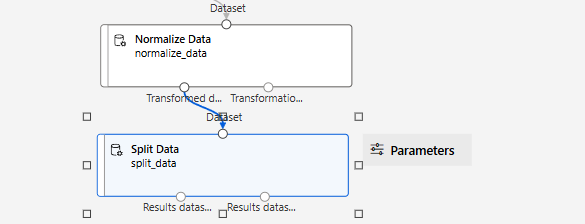
Job에서 확인

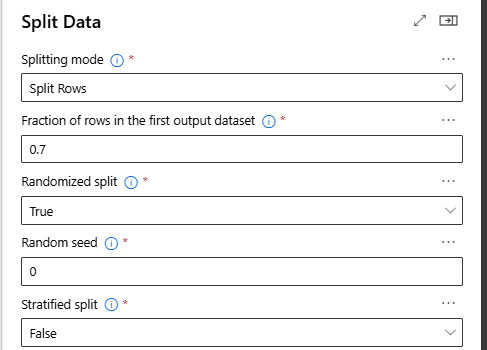




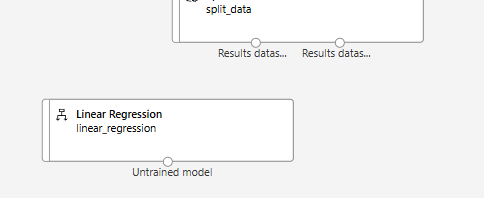


### 데이터 분리



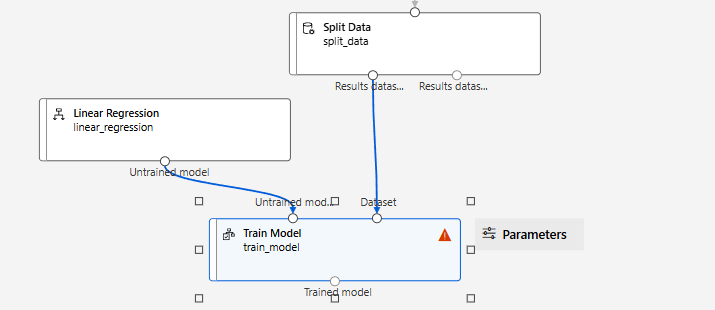


### 알고리즘 및 하이퍼 파라미터



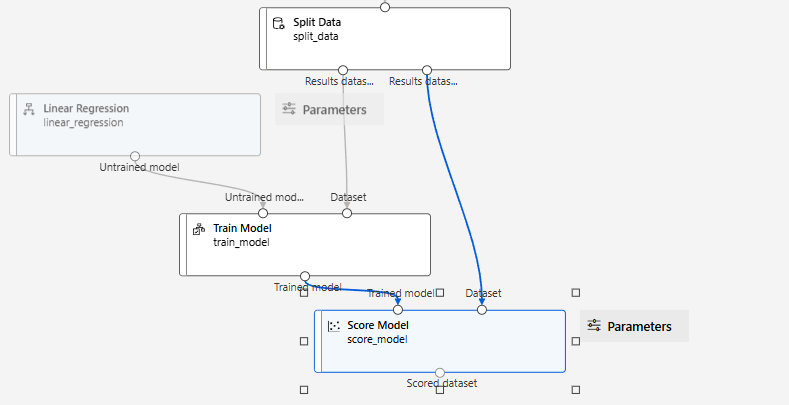


### 모델 학습



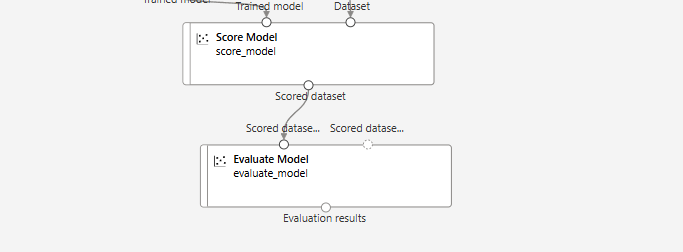


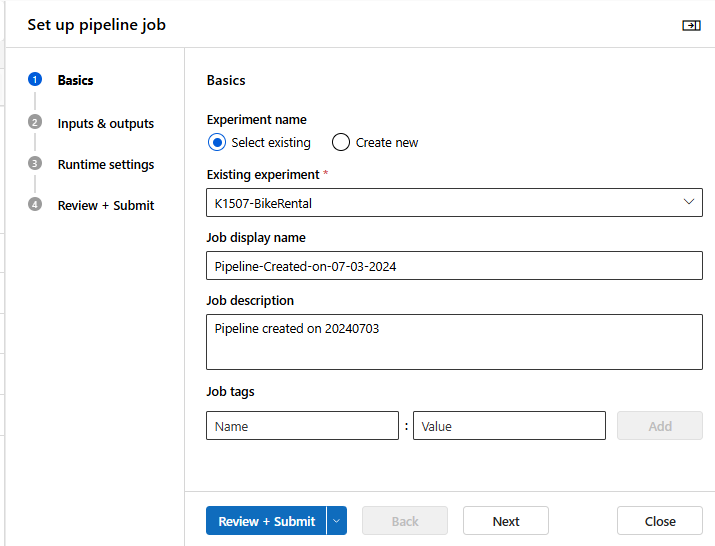
### 모델 테스트



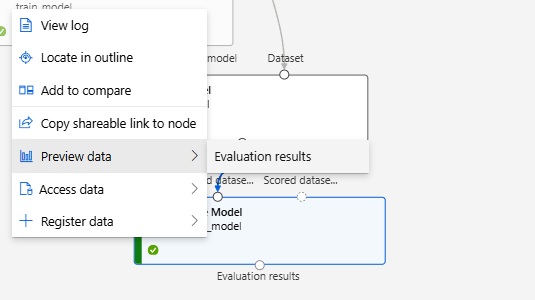


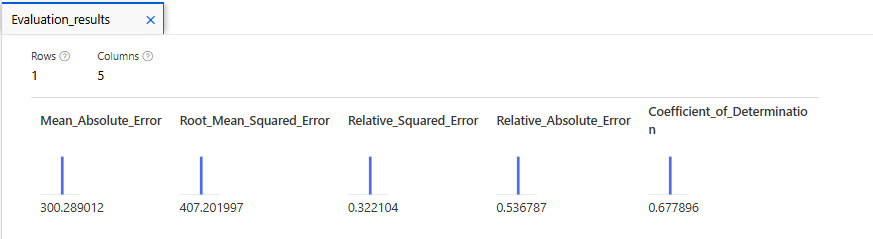
### 모델 평가





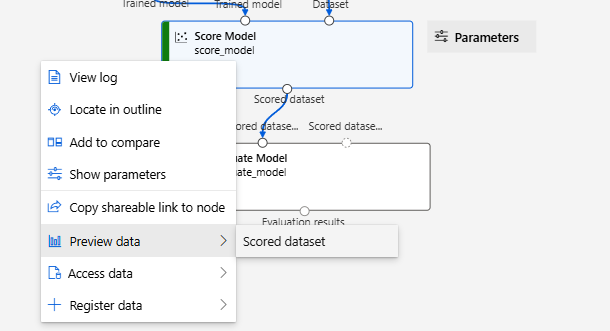


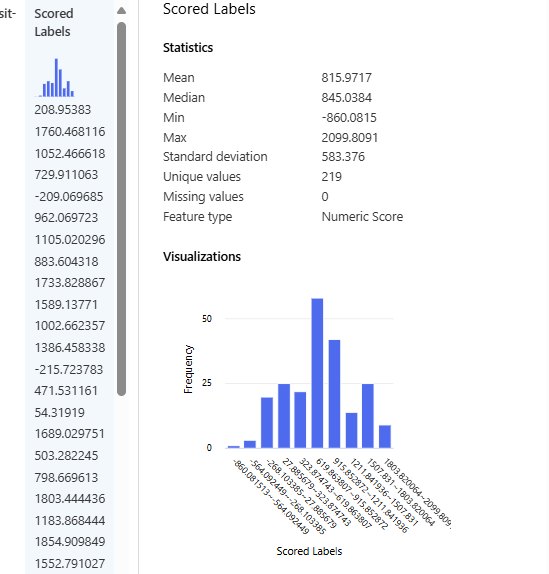




* MAE(Mean Absolute Error)
  + 오차의 절대값을 평균한 값
  + 독립변수의 단위 유지
* RMSE(Root Mean Squared Error)
  + MSE의 제곱근 : 오차의 제곱을 평균한 값의 제곱근
  + 이상치에 민감함
* RAE(Relative Absolute Error)
  + MAE를 실제값과 평균값의 절대차의 평균으로 나눈 값
  + 0에 가까울수록 예측값이 실측값과 가까움
* RSE(Relative Squared Error)
  + MSE를 실제값과 평균값의 차이의 제곱평균으로 나눈 값
  + RSE가 0에 가까울수록 예측값이 실측값과 가까움
* 결정계수(Coefficient of Determination)
  + 모델의 독립변수들이 종속변수를 얼마나 잘 설명하는지 나타내는 지표
  + 종속변수의 변동량 중에서 회귀모델로 설명가능한 부분의 비율
  + 즉, 독립변수가 종속변수를 설명하는 정도를 표현하는 지표 (설명력)

### 예측 결과 확인





# <오후>

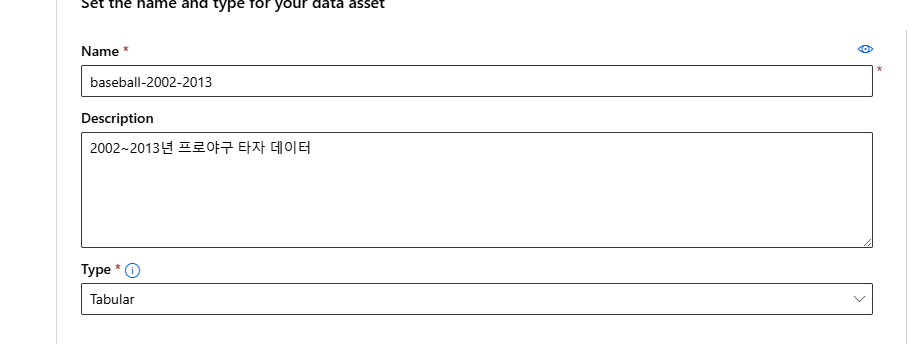
## [실습 - 군집MLD\_프로야구]

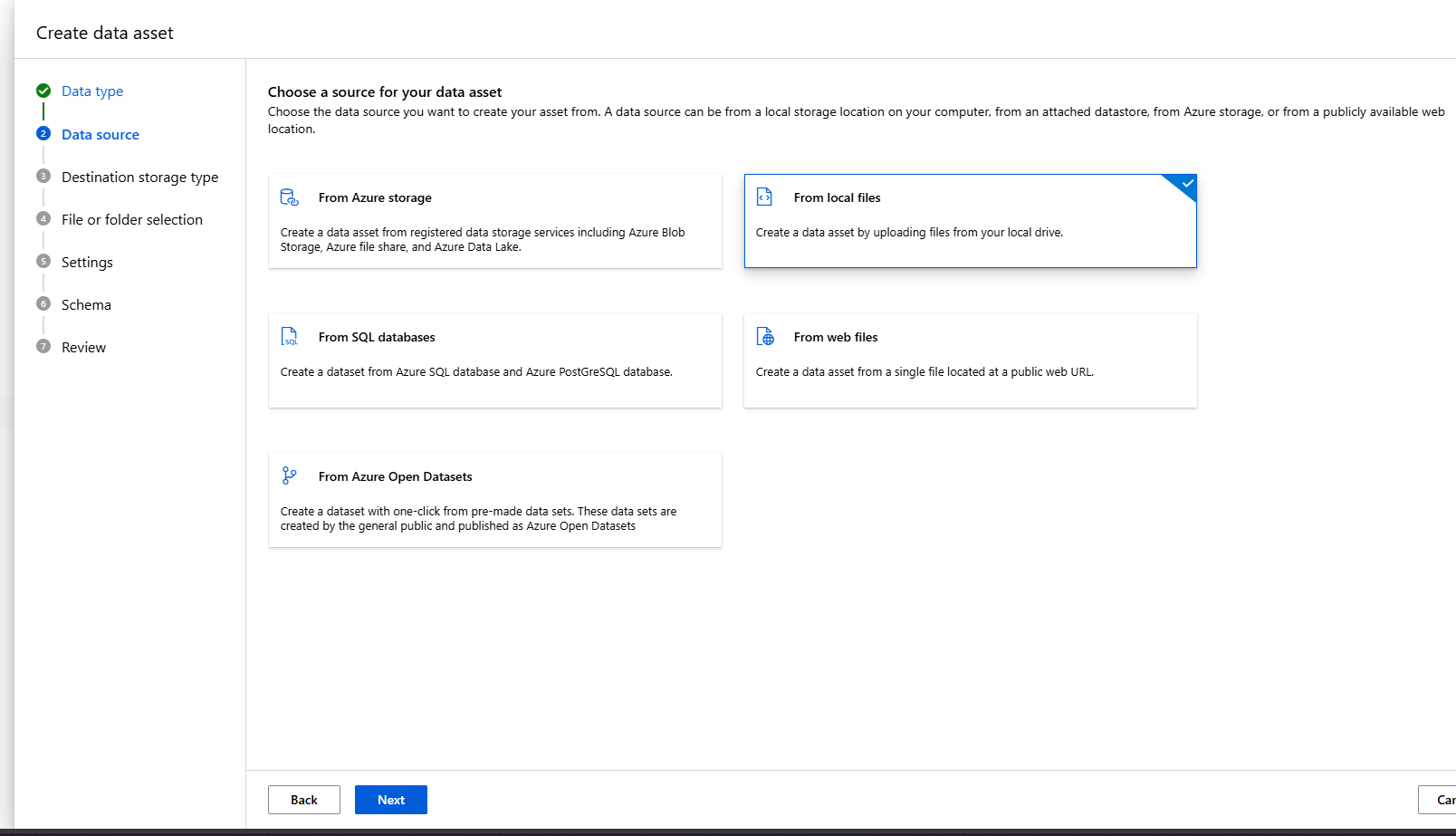
https://www.koreabaseball.com/Futures/Player/Hitter.aspx

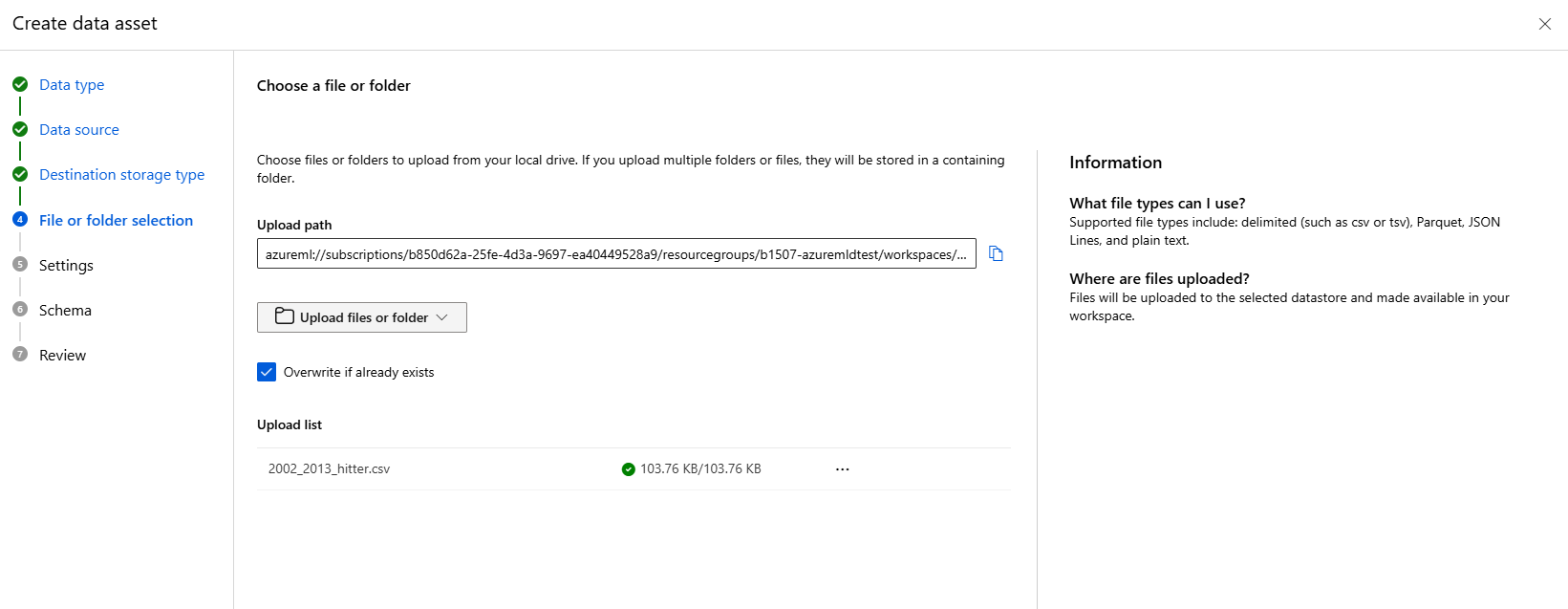
* 리소스를 생성하는 과정까지는 동일하기에 생략

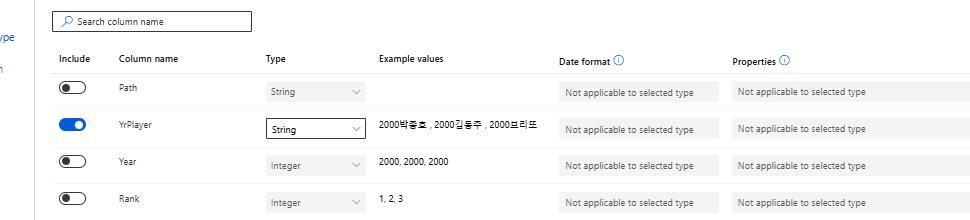
### 데이터세트 등록

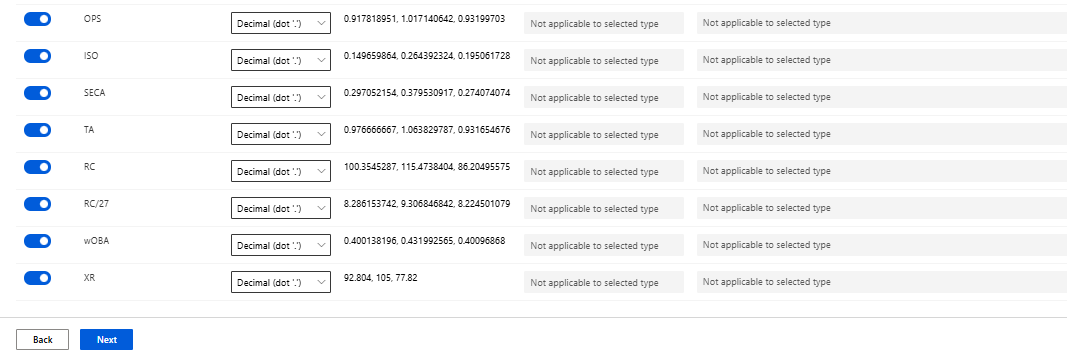
Data 생성









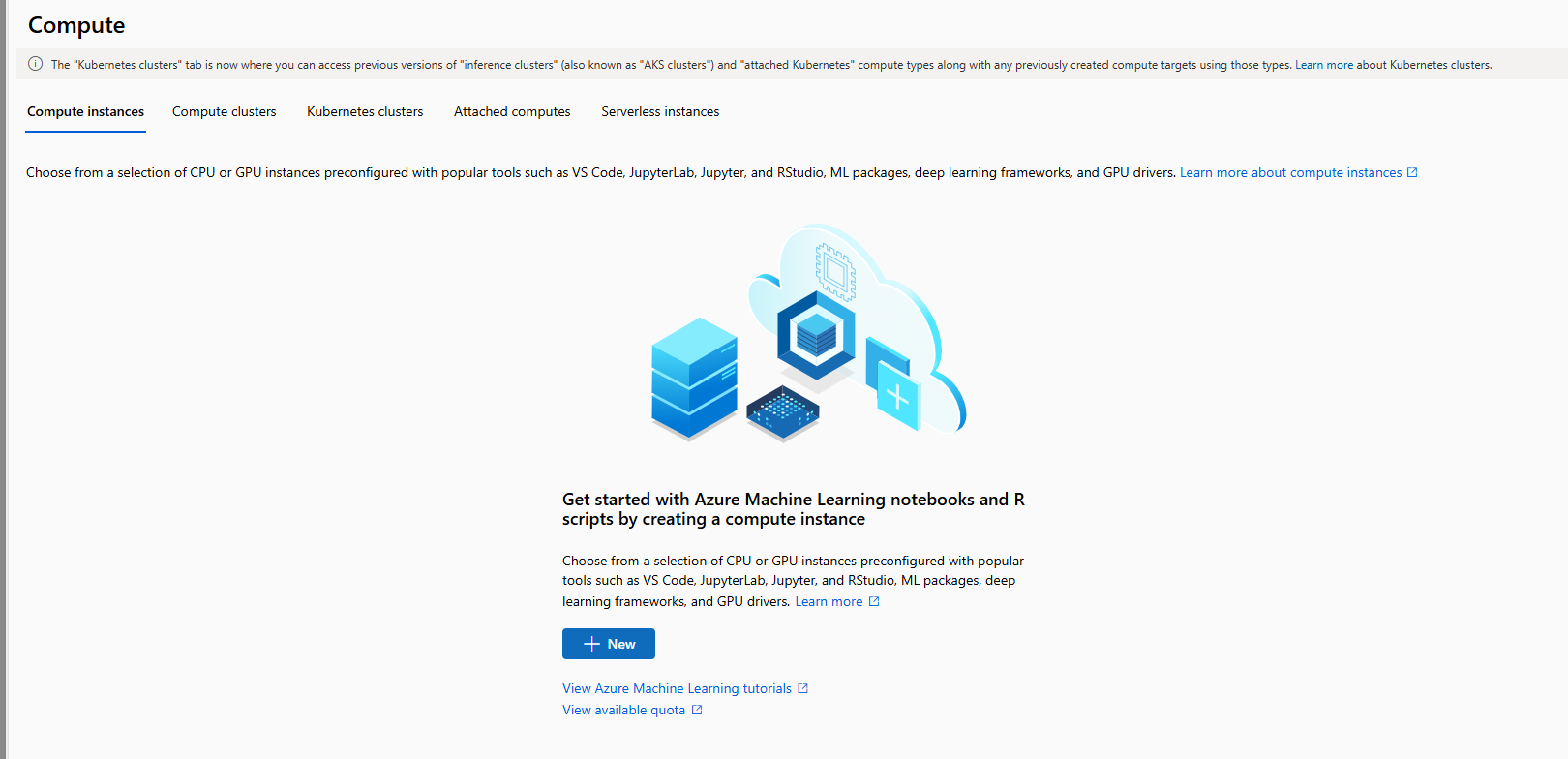


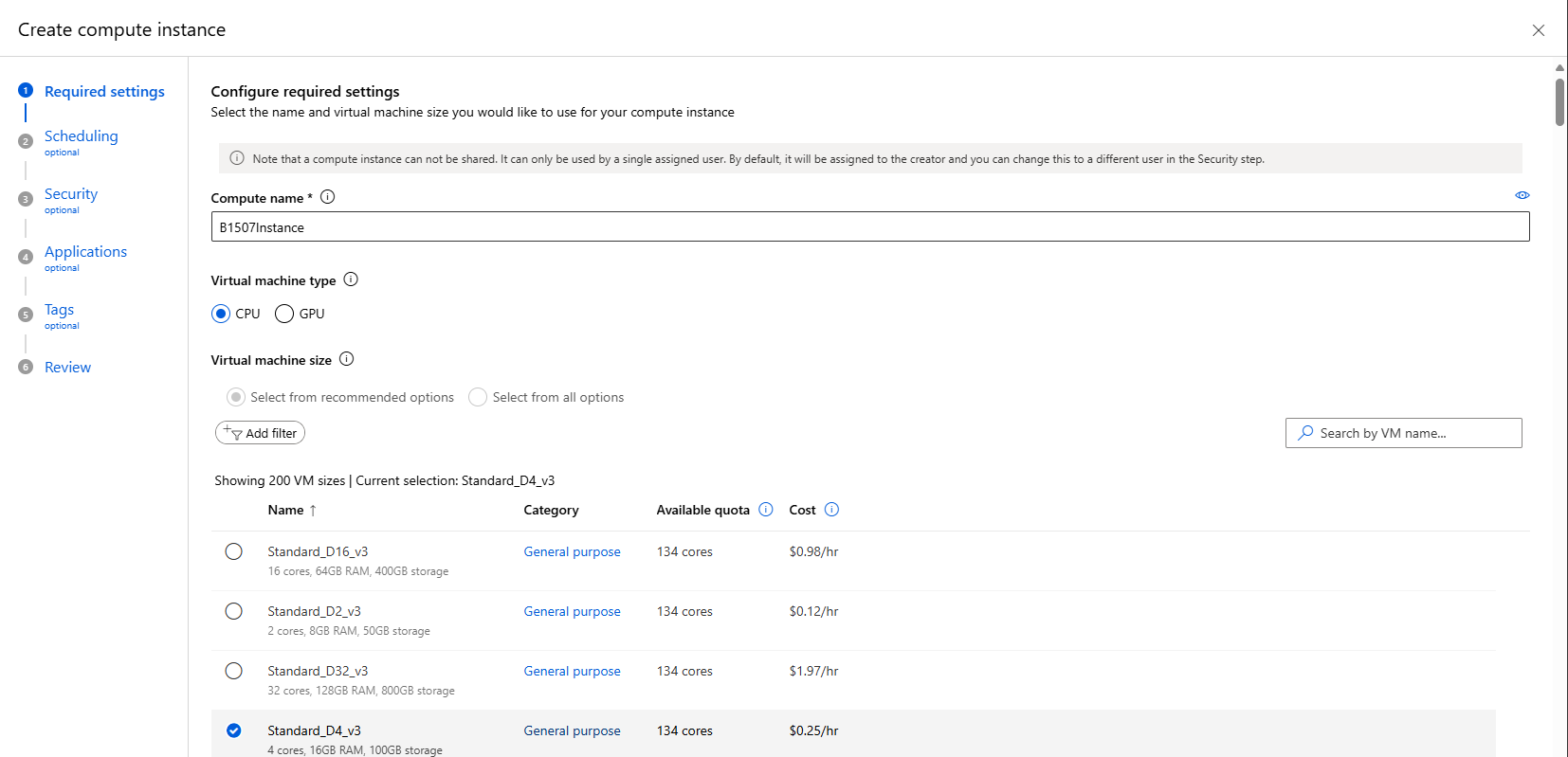
2000\_2001, 2002\_2013, 2014 csv 파일 총 3번 반복

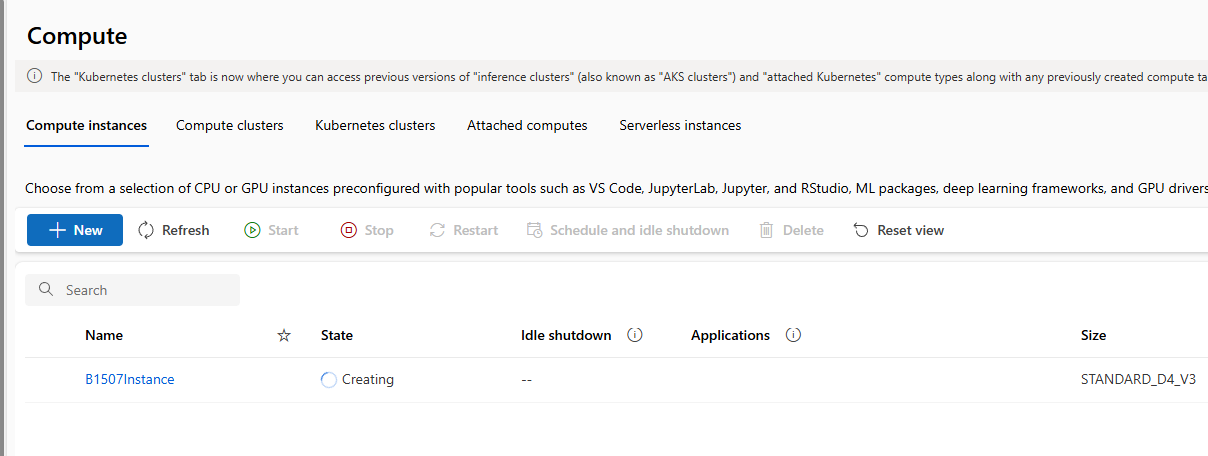


### 컴퓨트 대상 설정

Compute instances 사용

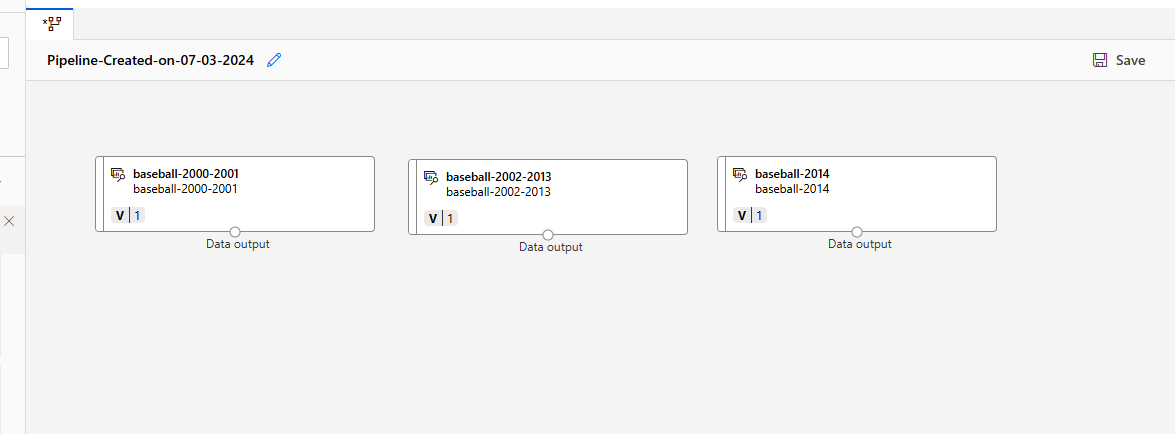




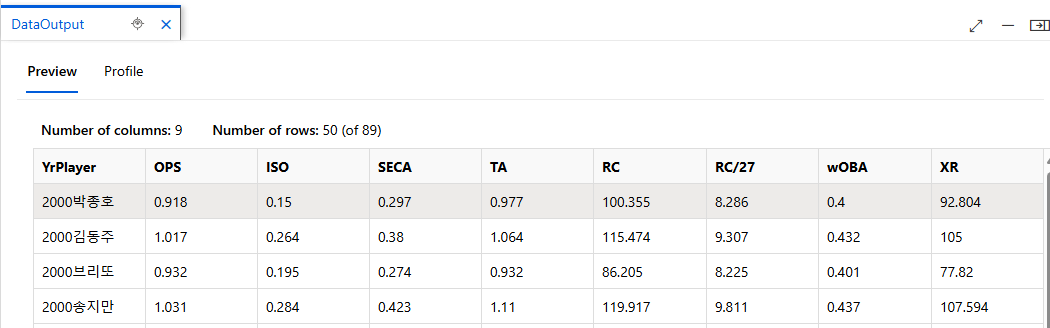


### 파이프라인 생성 후 데이터세트 관찰

Designer에서 생성한 데이터 3개 배치

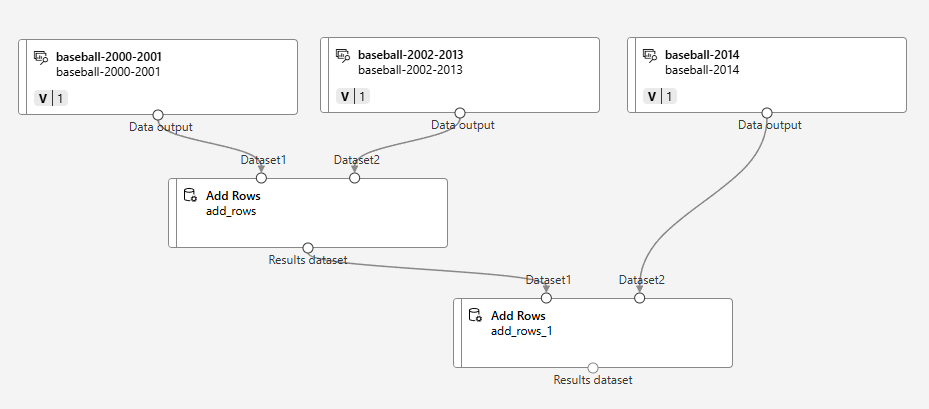


데이터 확인

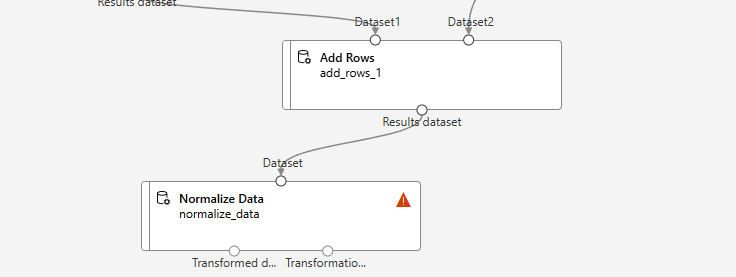


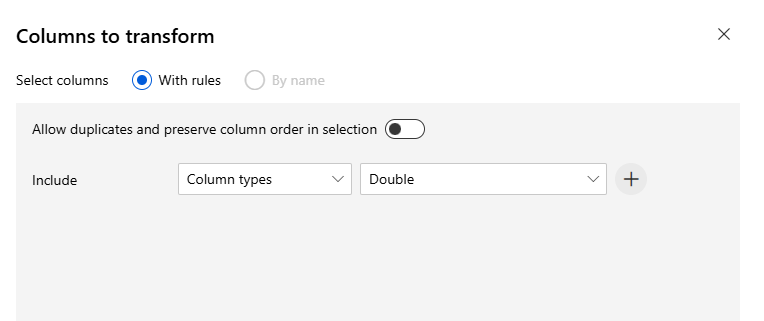
### 데이터 병합

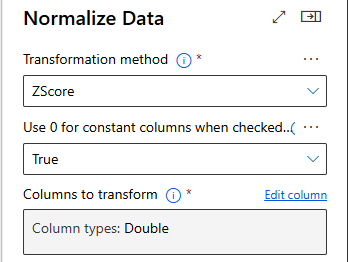
이어주는 순서대로 병합이 된다.



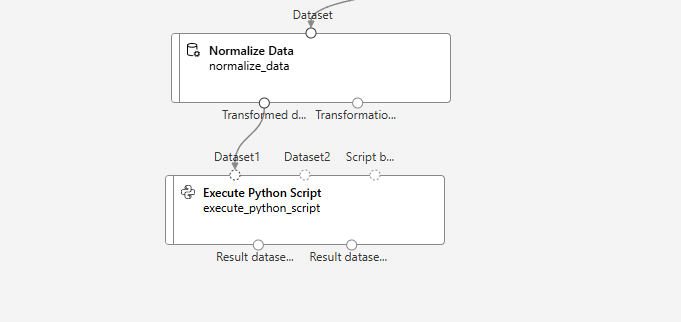
수치 데이터 표준화



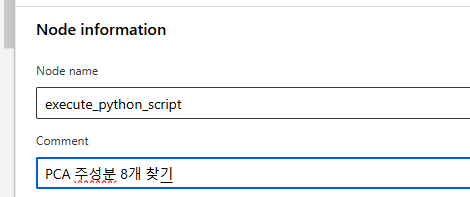


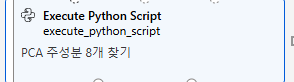


### 주성분 분석(PCA)

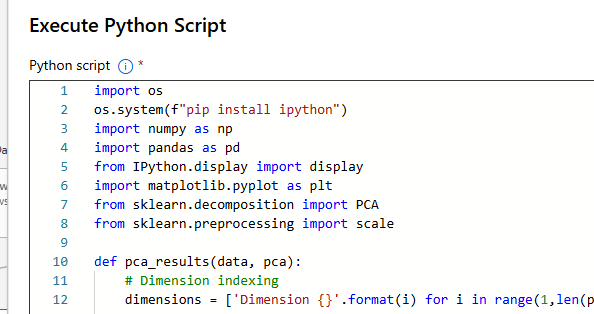


설명을 적어둘 수 있다.

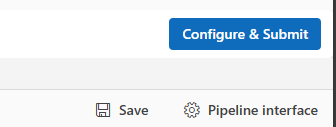


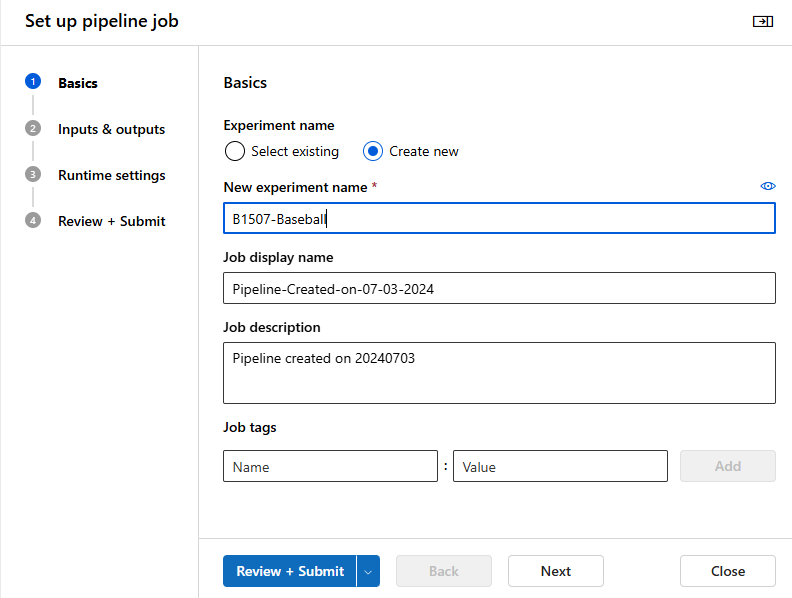


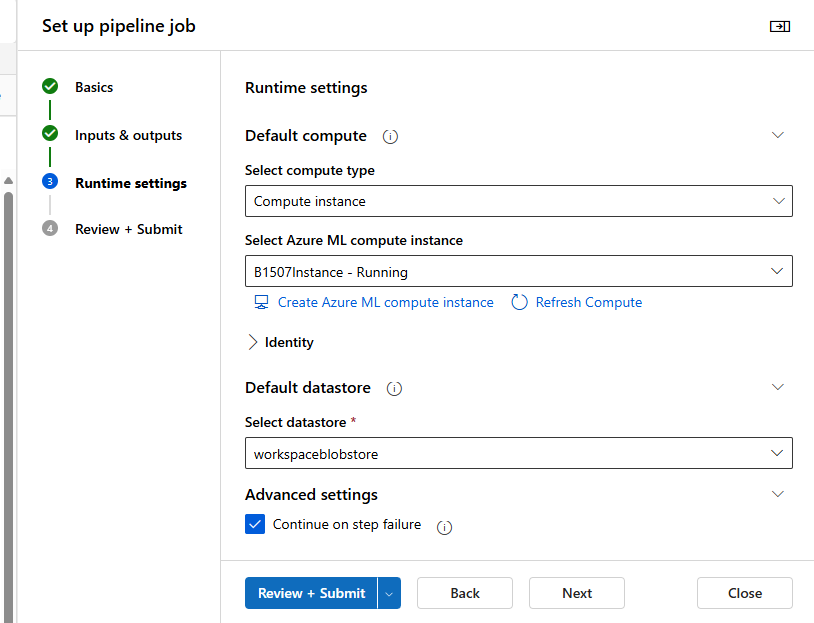
Python 코드 작성 / pca.py.txt 파일 참고

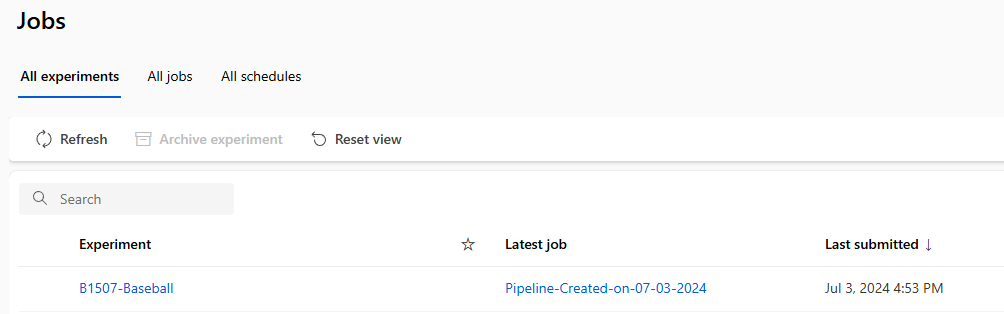


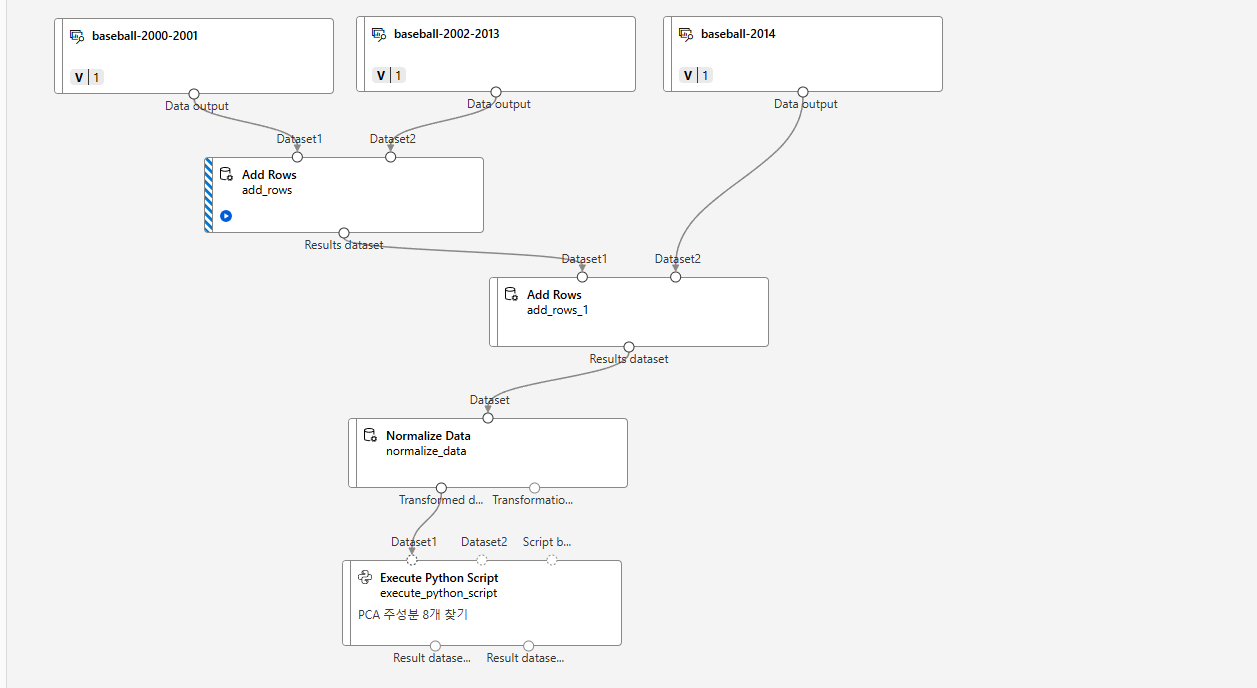
### 중간점검



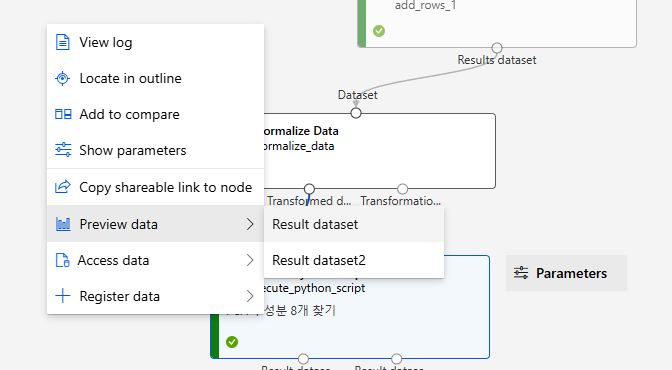


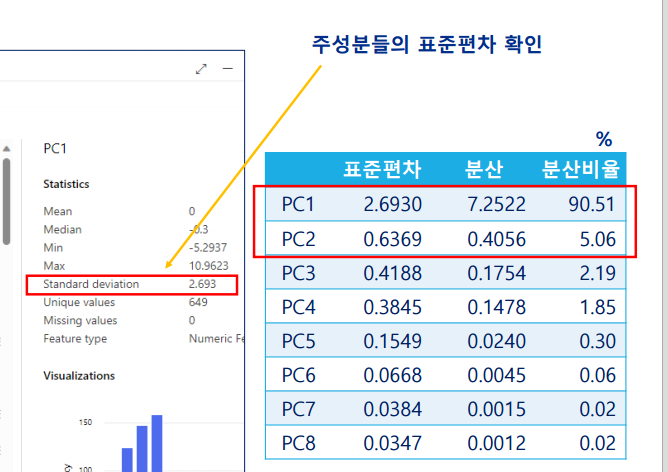


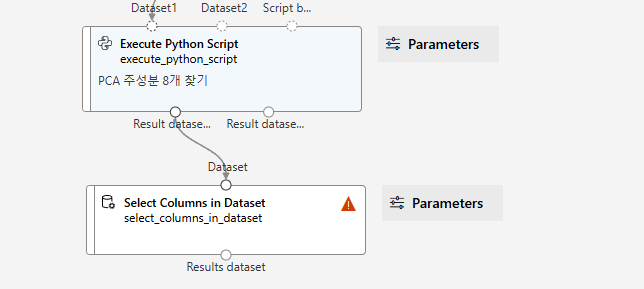


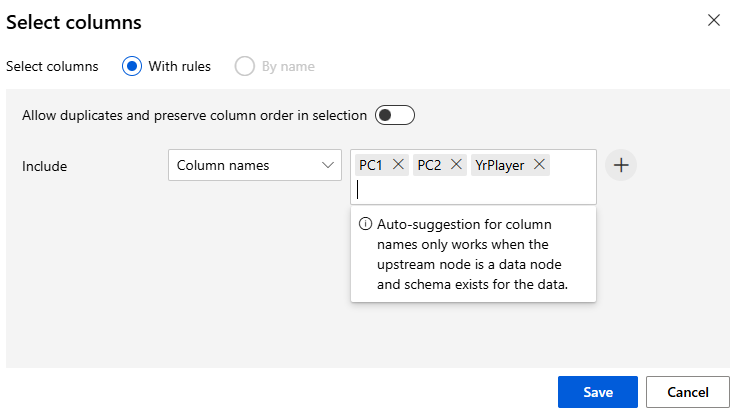


### 주성분 확인

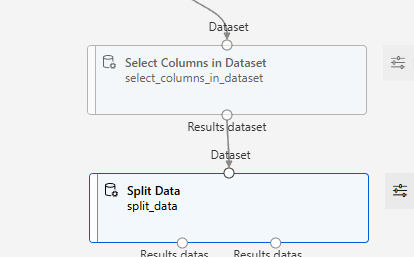




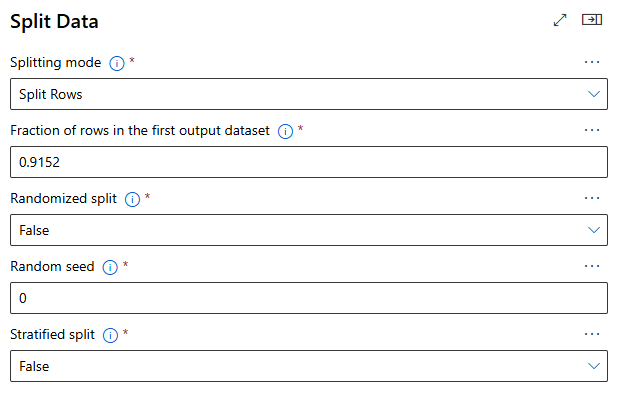




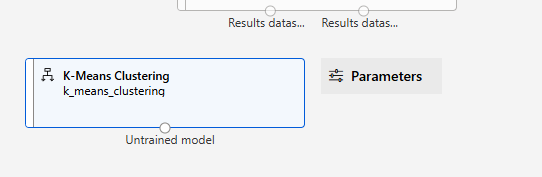
### 데이터 분리

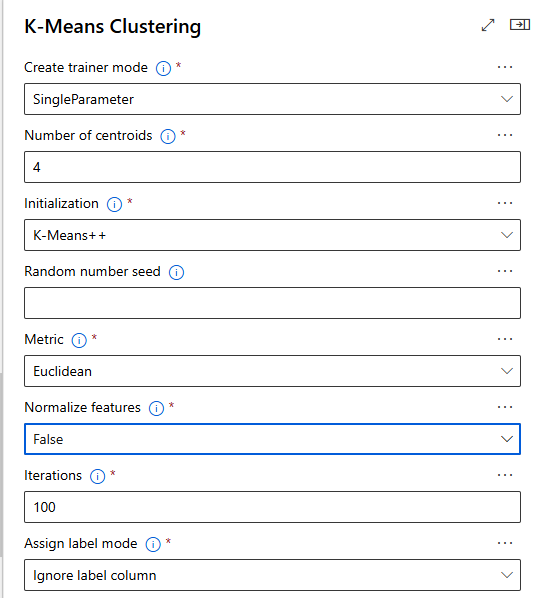


2000~2013년 데이터: 594개, 2014년 데이터: 55개, 594/649 = 약 0.9152

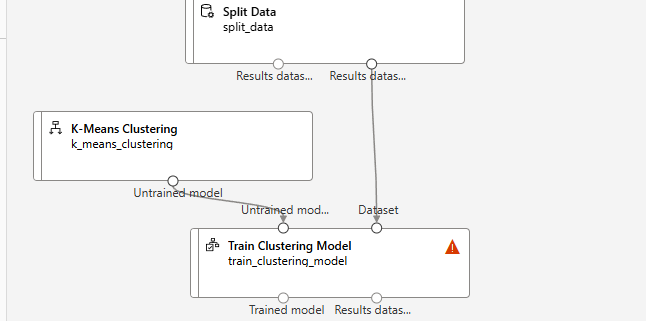


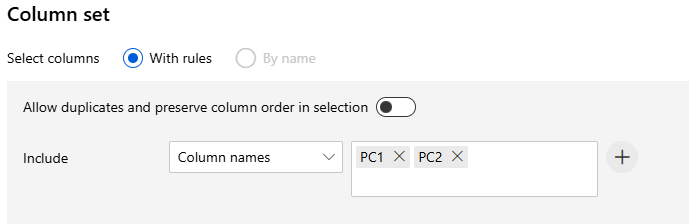
### 알고리즘 및 하이퍼 파라미터



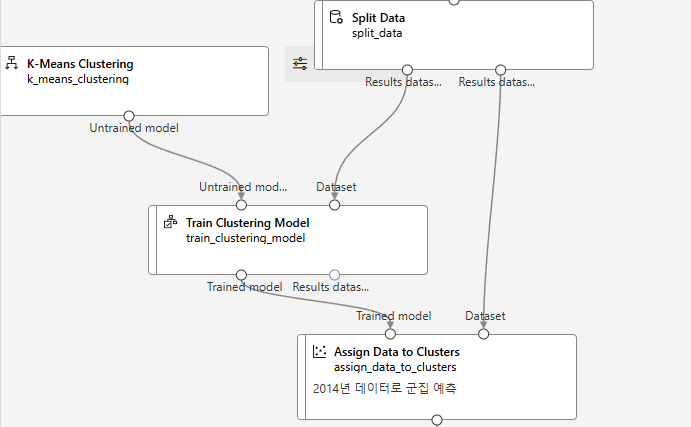


### 모델 학습

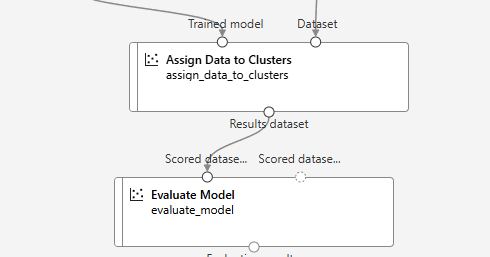




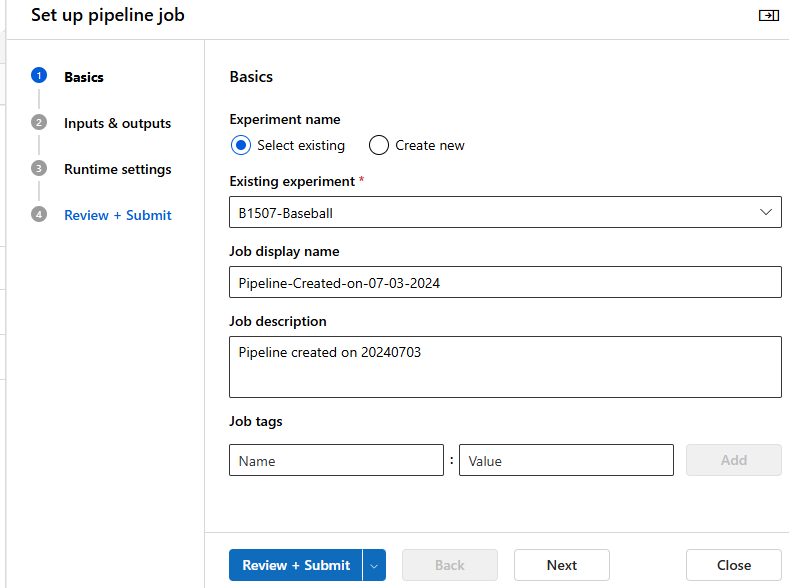
### 모델 테스트(예측)

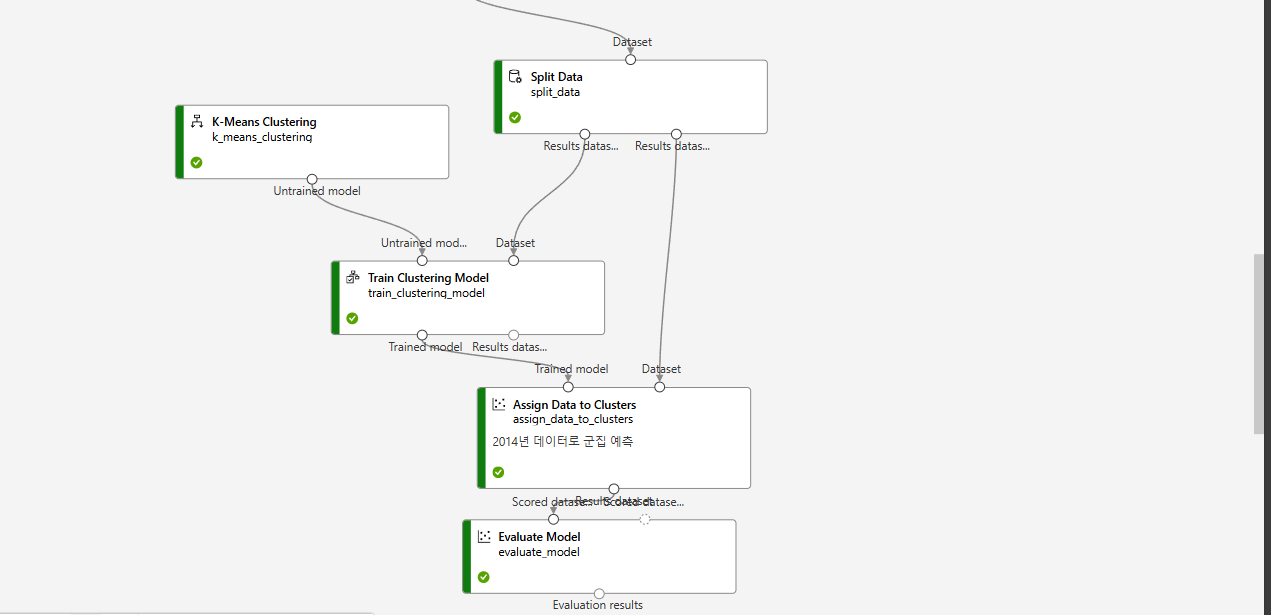


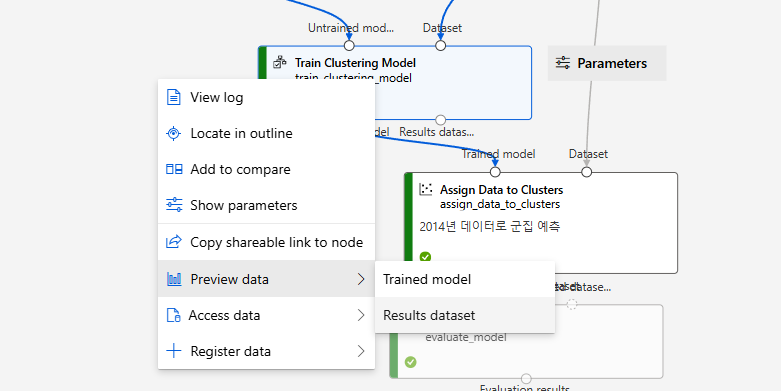
### 모델 평가

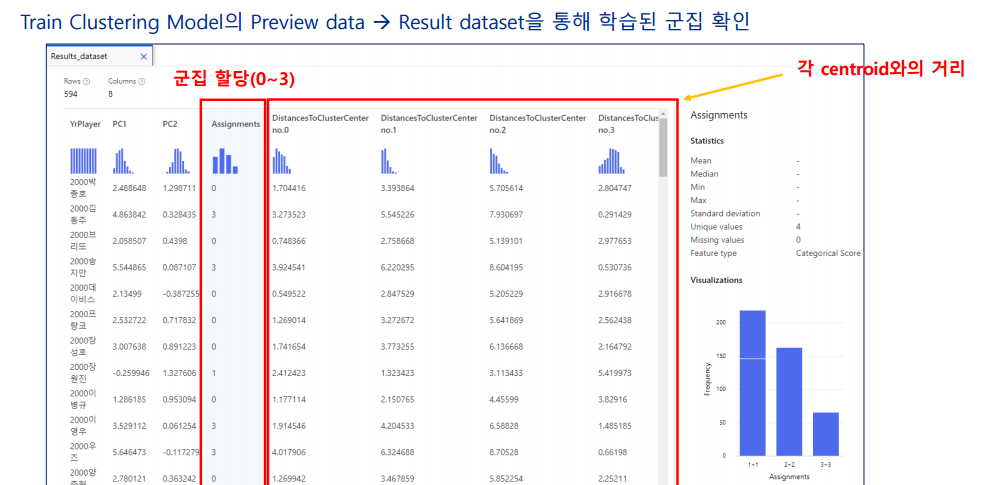


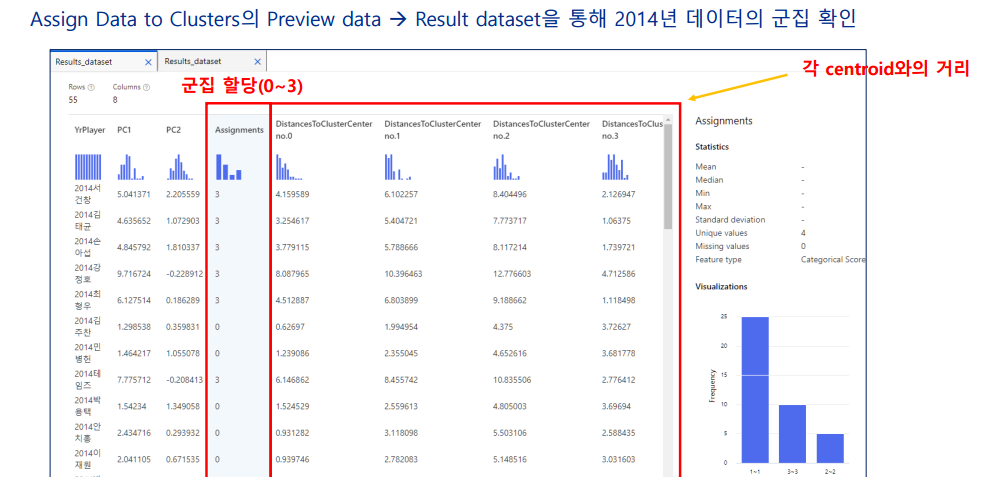
### 군집 결과 확인

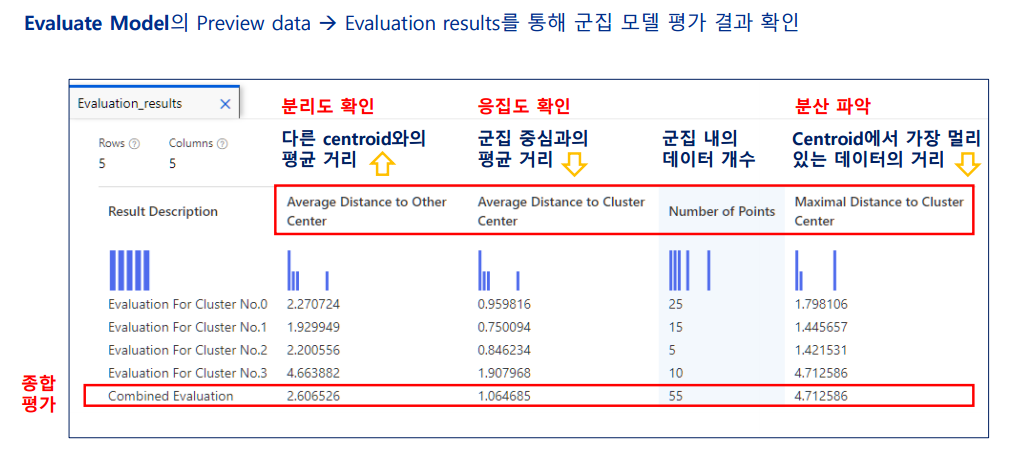




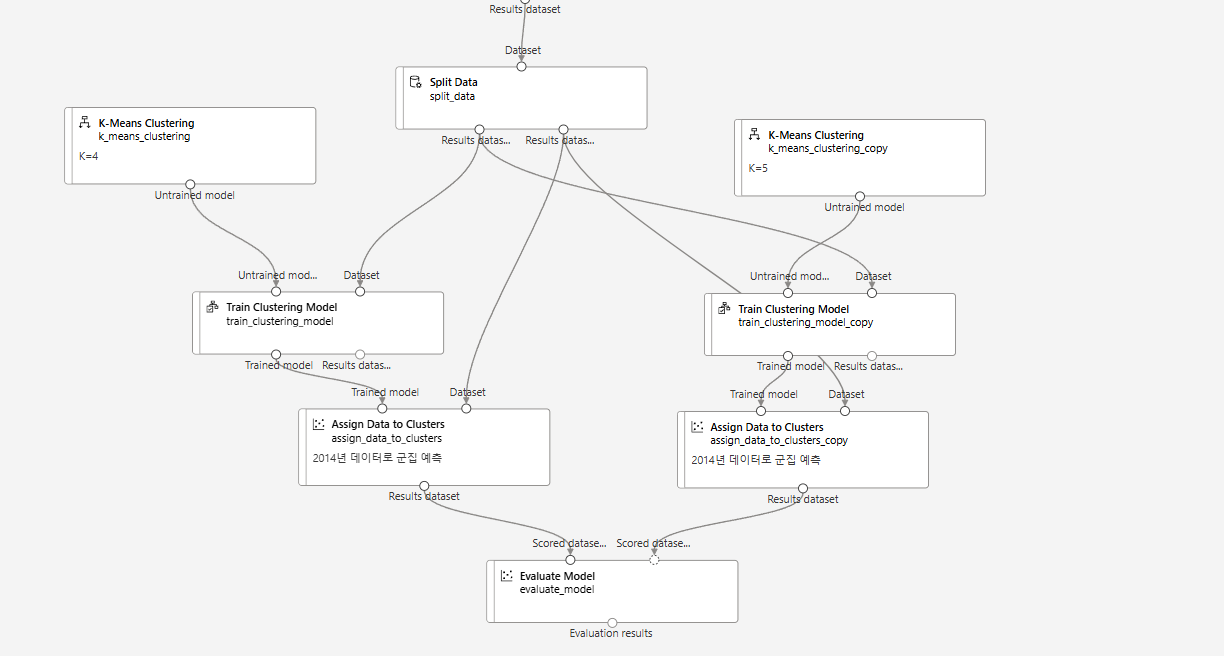


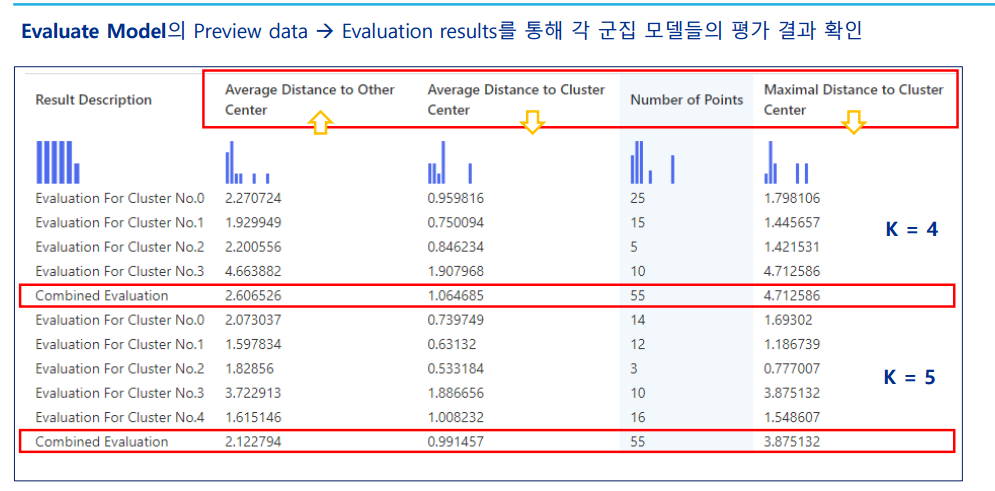






### 군집 모델 평가 – K = 4, 5 비교

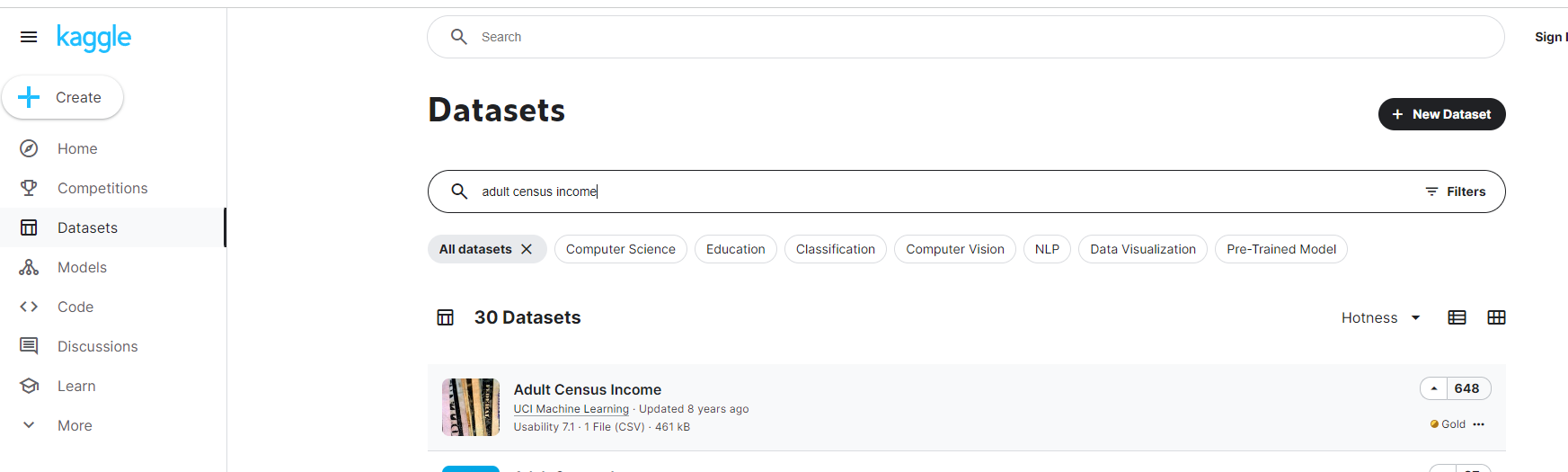


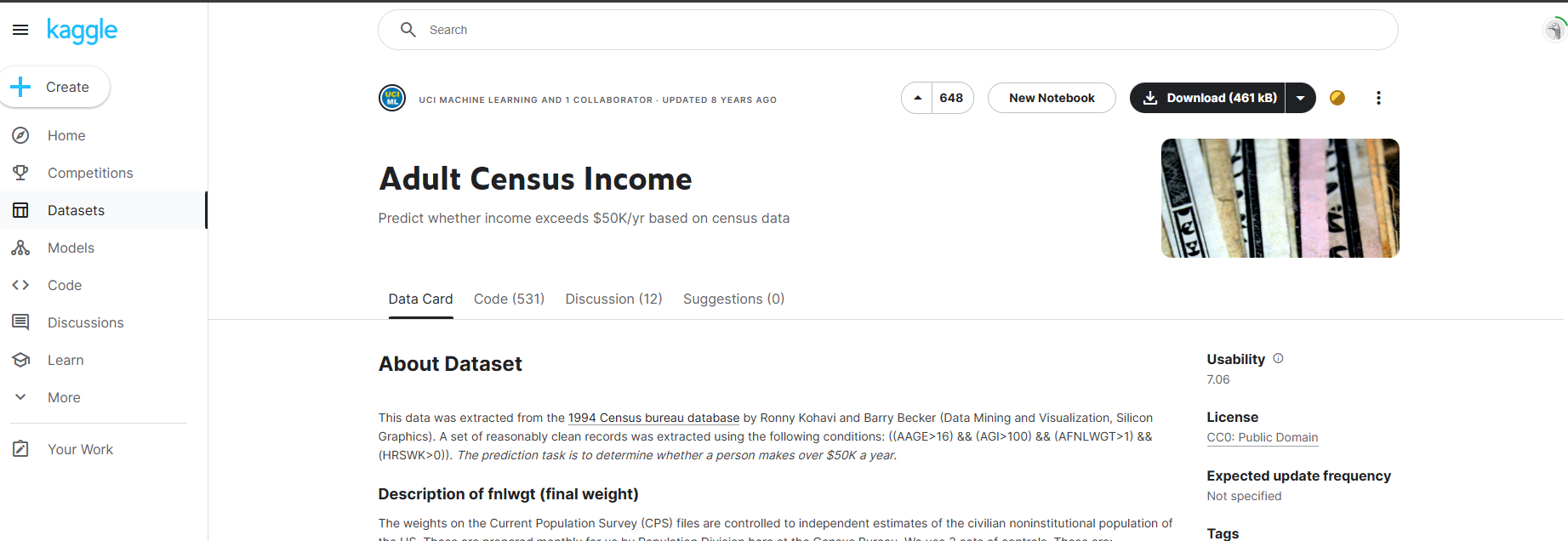


## [실습 - 분류MLD\_개인수입예측\_랜덤포레스트]

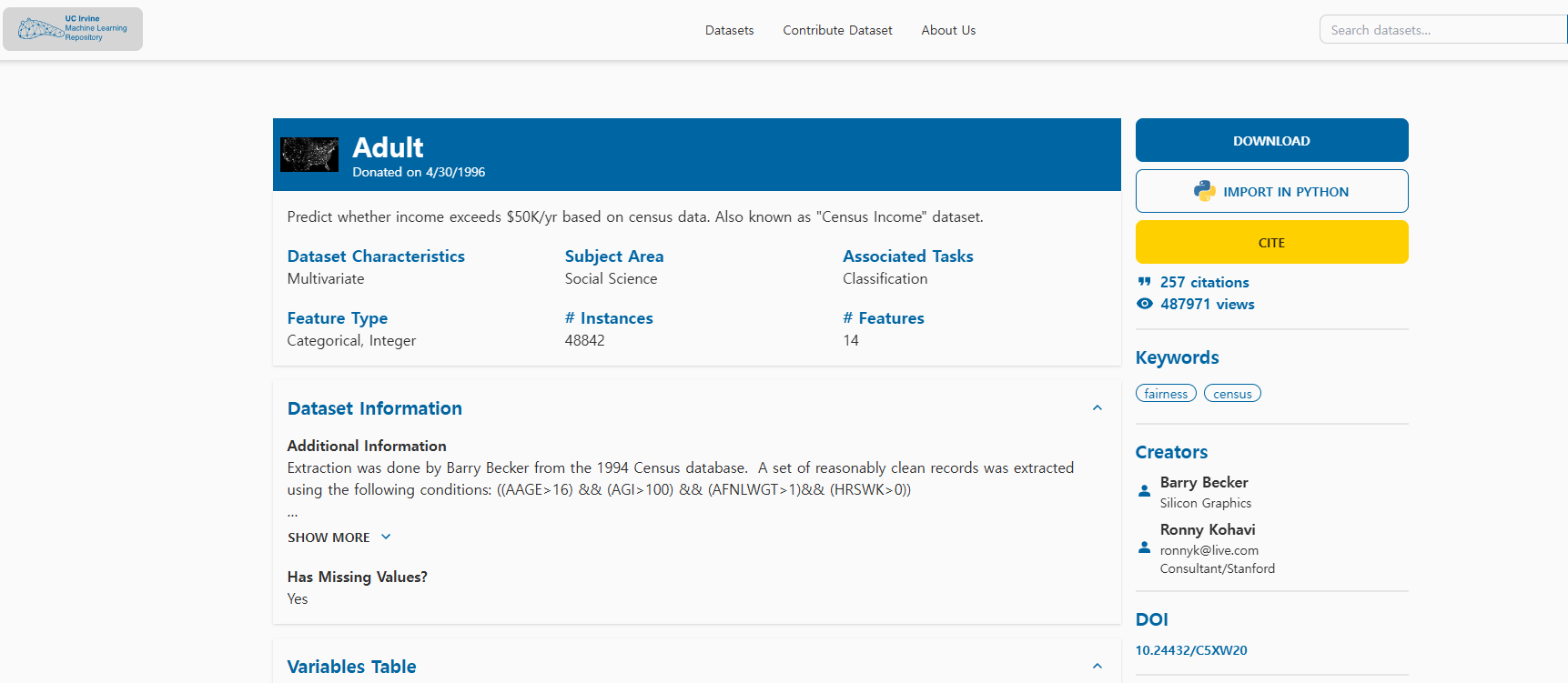
### 데이터 수집

kaggle에서 다운로드(https://archive.ics.uci.edu/dataset/2/adult)





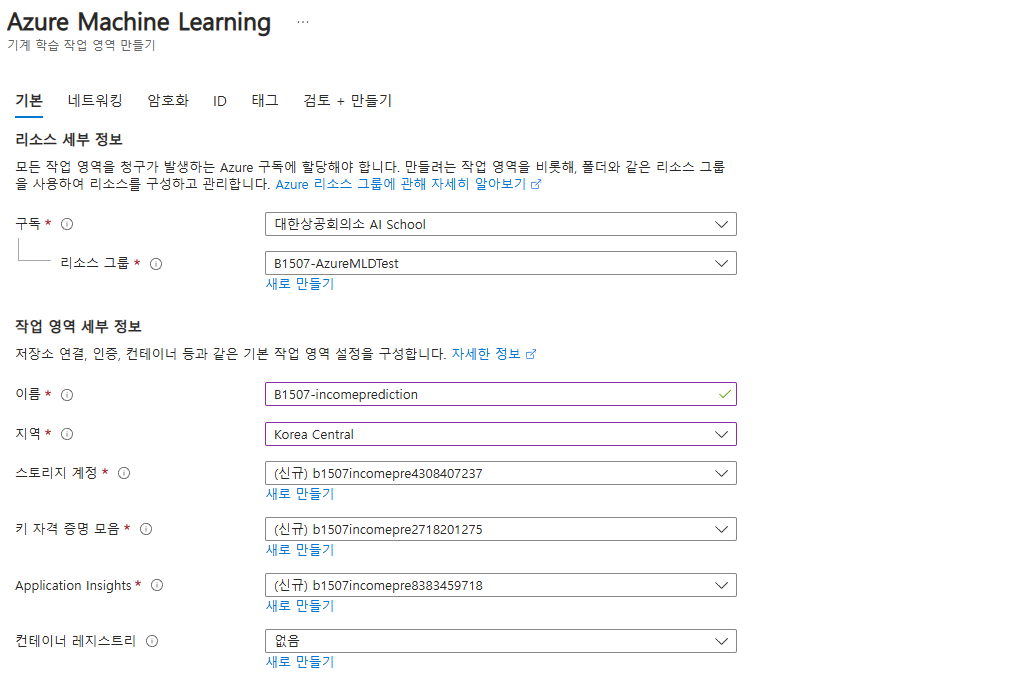
데이터 세트의 원 출처: UCI에서 정보와 설명 확인 가능 (<https://archive.ics.uci.edu/dataset/2/adult>)



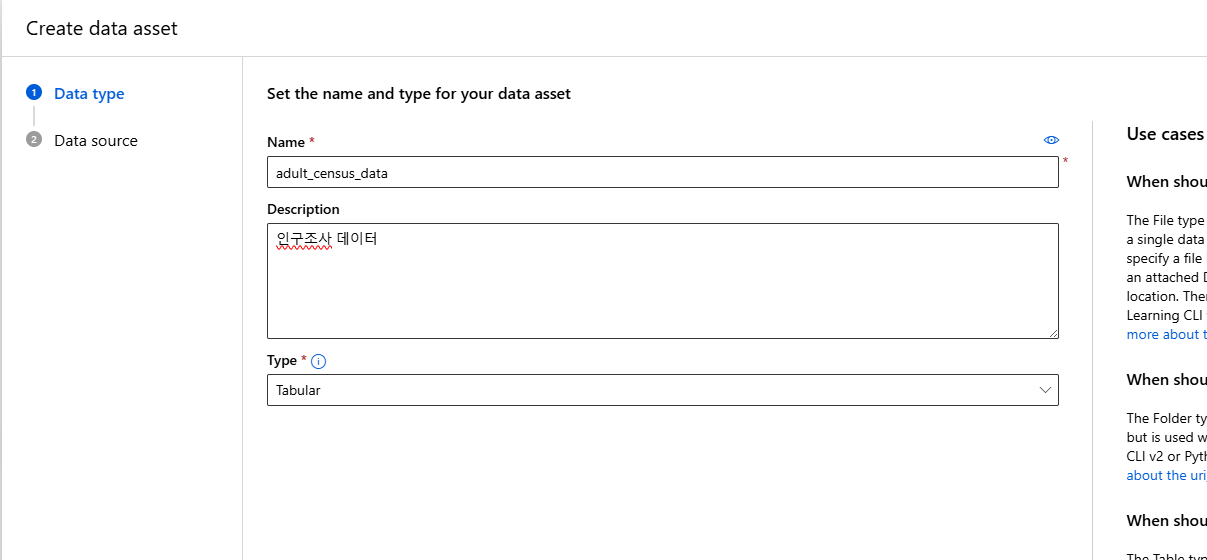


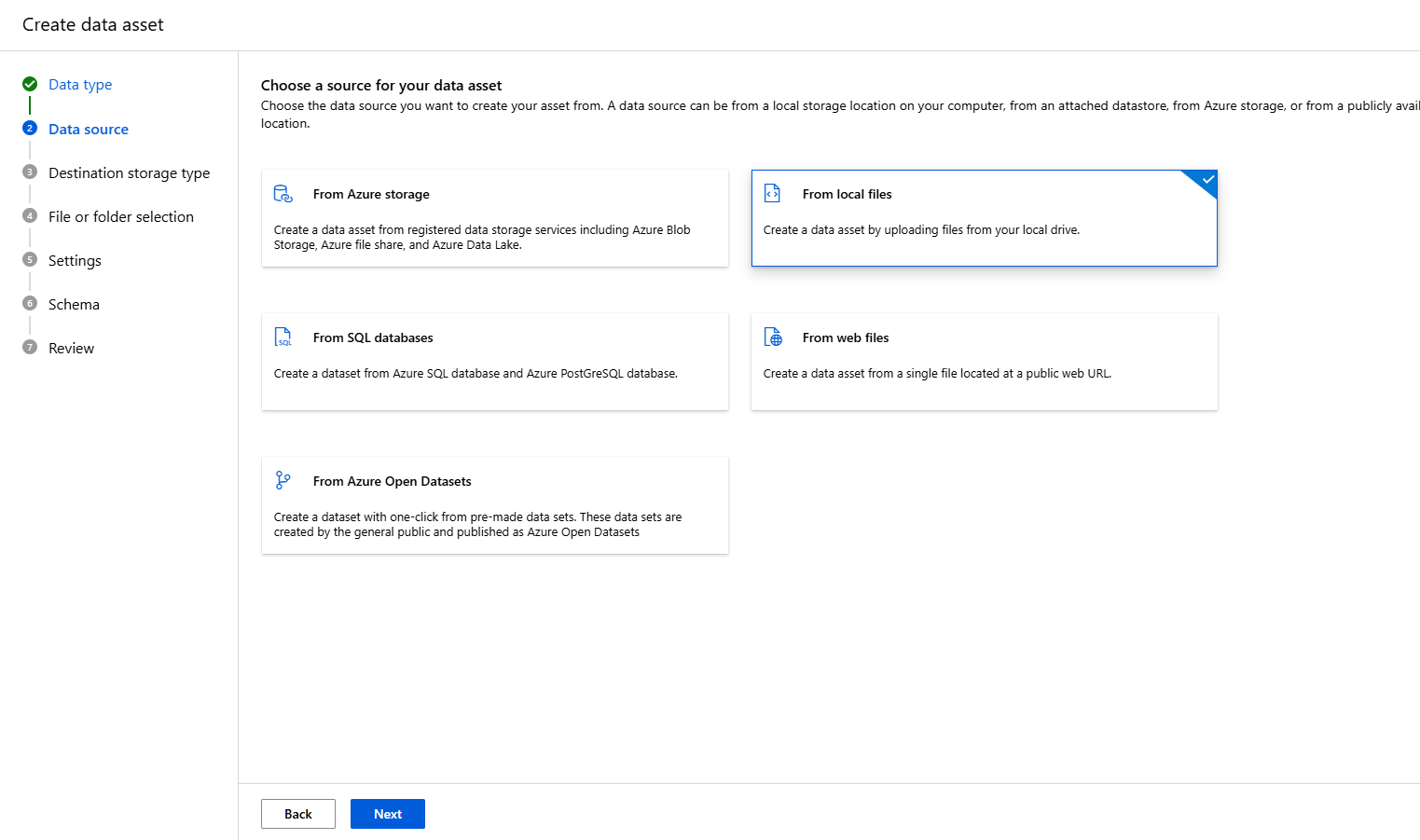
### Azure 접속

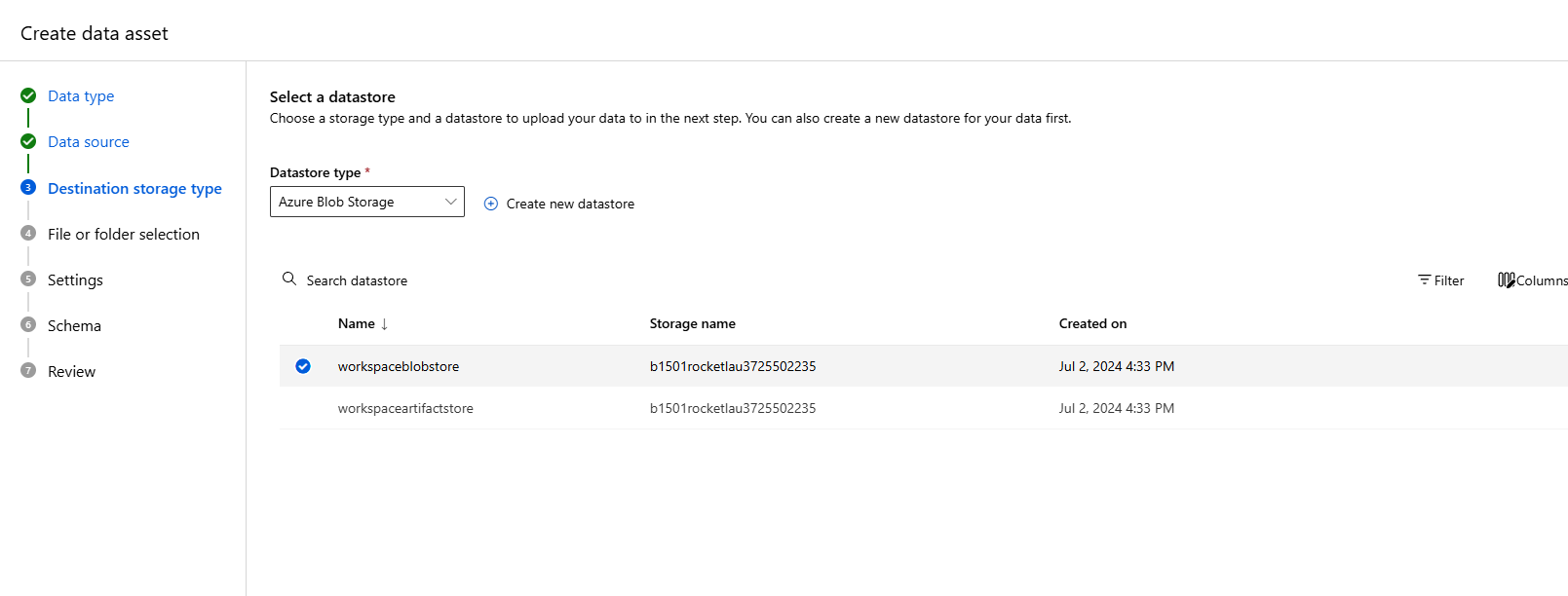
### 머신러닝 리소스 생성

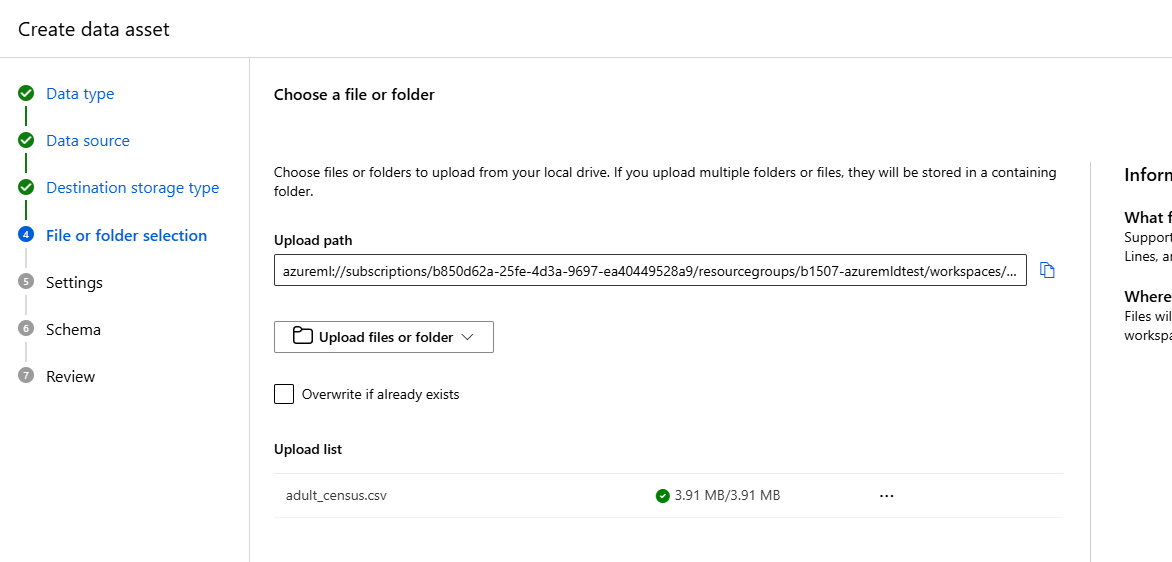


### 데이터세트 등록

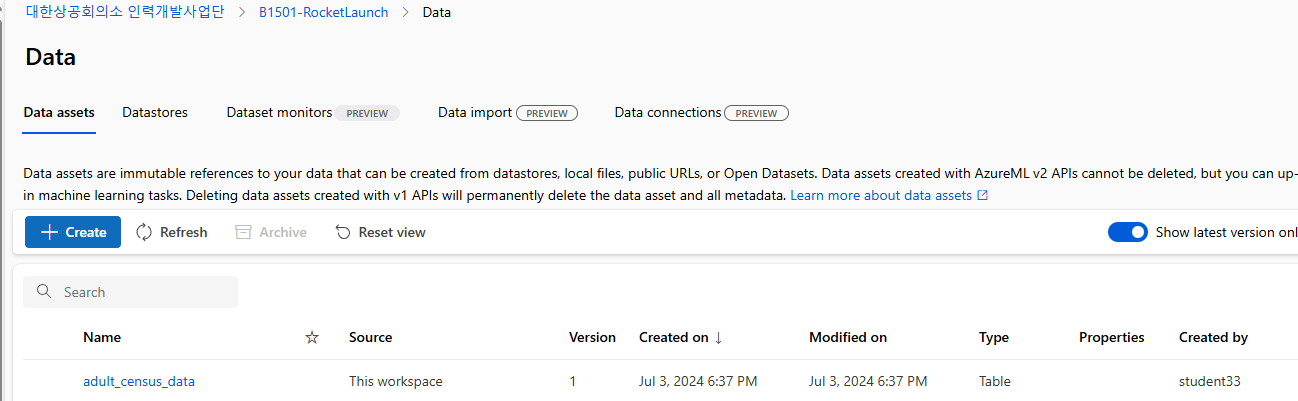




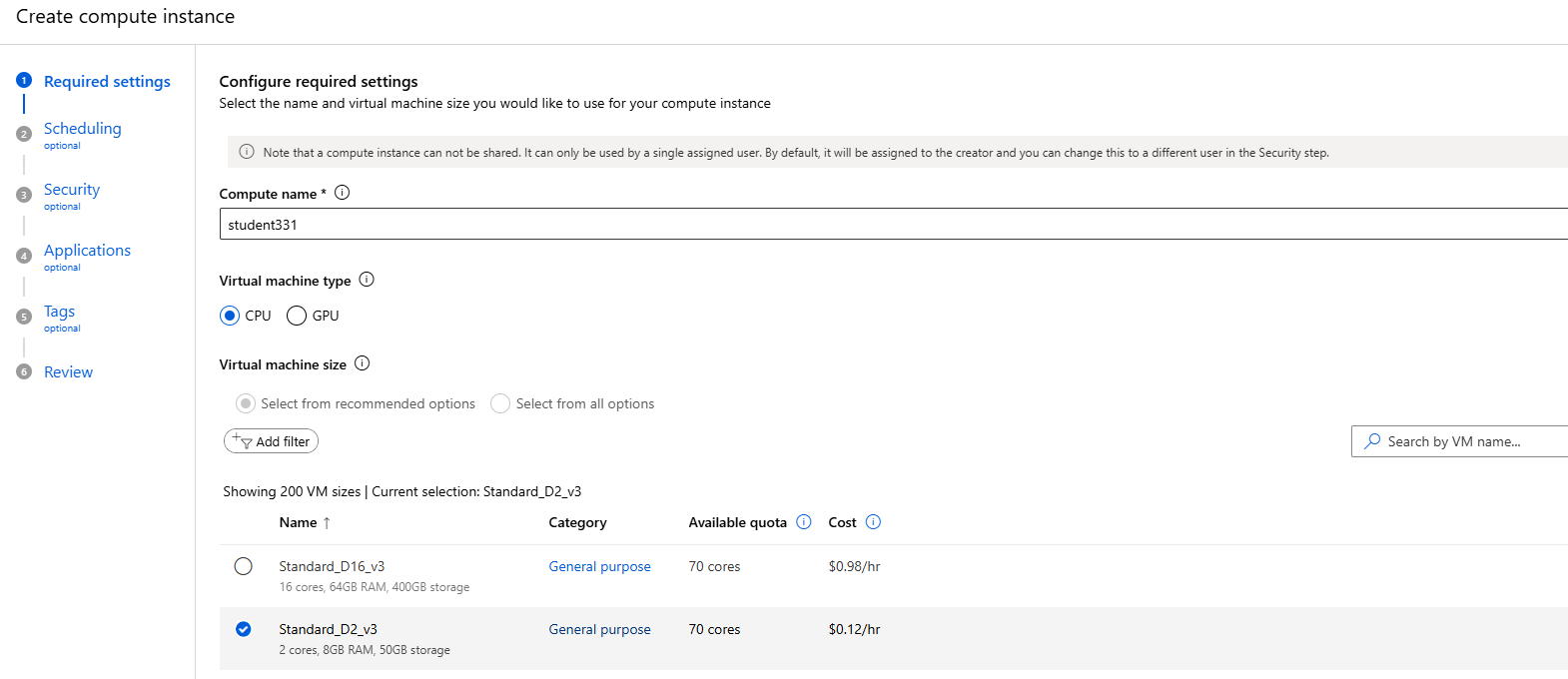




이후 변동사항이 없어서 스킵 후 생성 확인



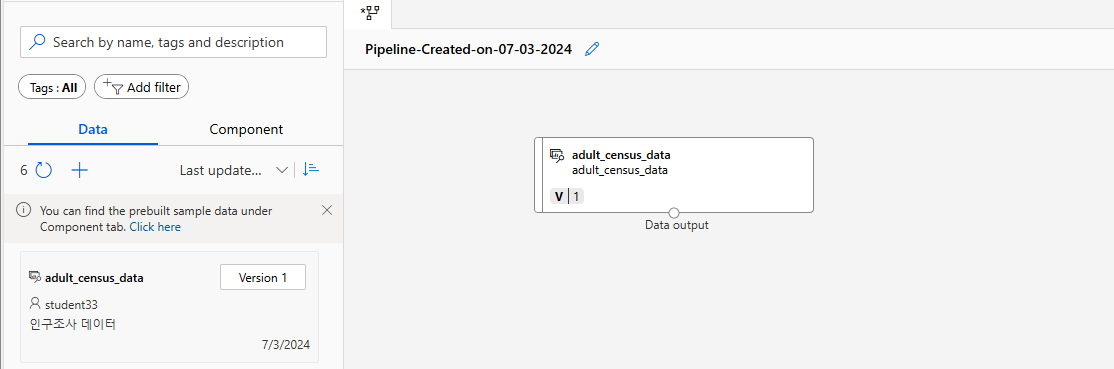
### 컴퓨트 대상(Compute Target) 설정



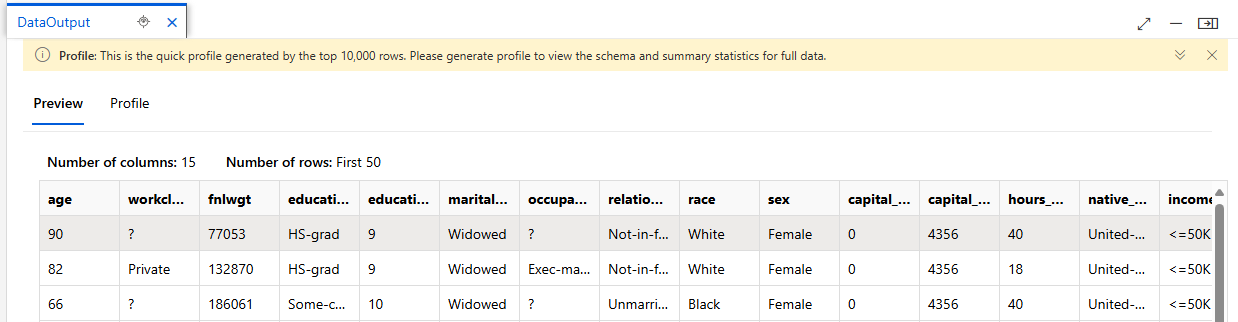


### 디자이너 시작

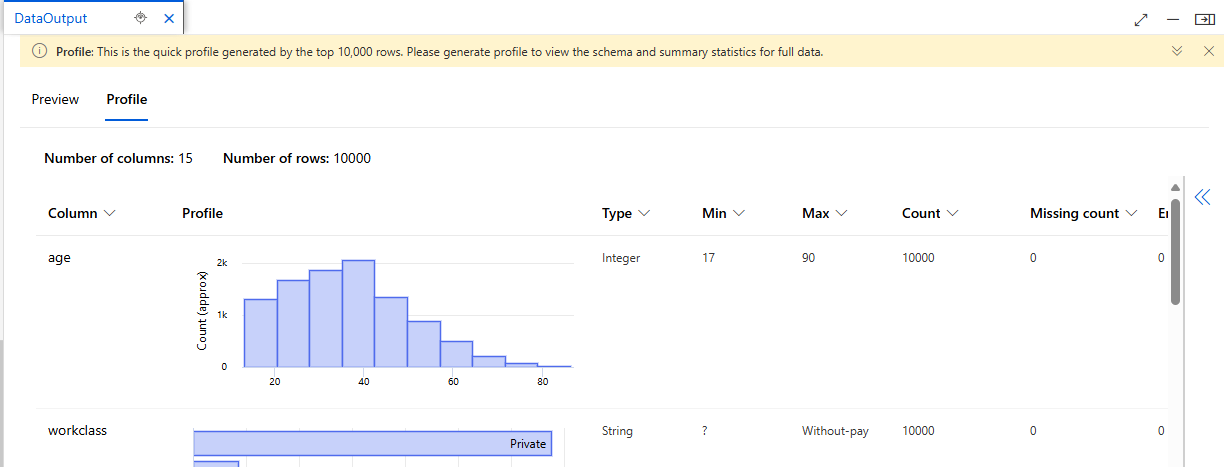
Data에서 업로드한 데이터 세트 가져오기



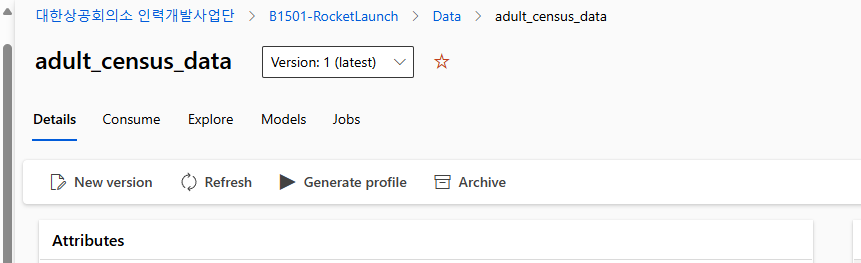
### 데이터 이해

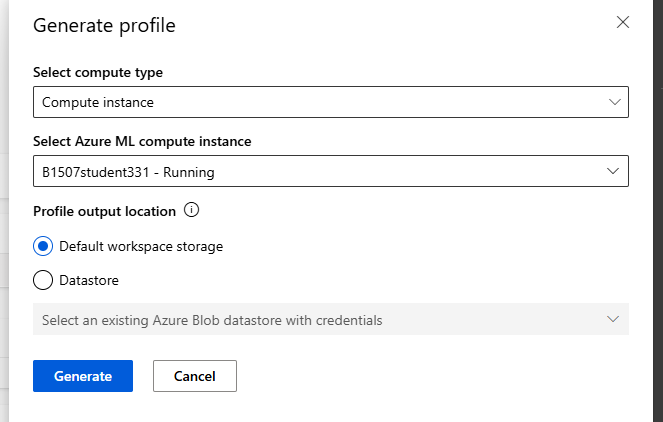


Profile은 10,000개 행에 대해서 보여지고 있으며, Profile 생성 기능을 통해 모든 데이터에 대한 Profile 확인이 가능하다.



Data에서 선택한 후 Generate profile 클릭





Jobs 메뉴에서 Profile 생성 작업의 진행상황 확인이 가능하다.

