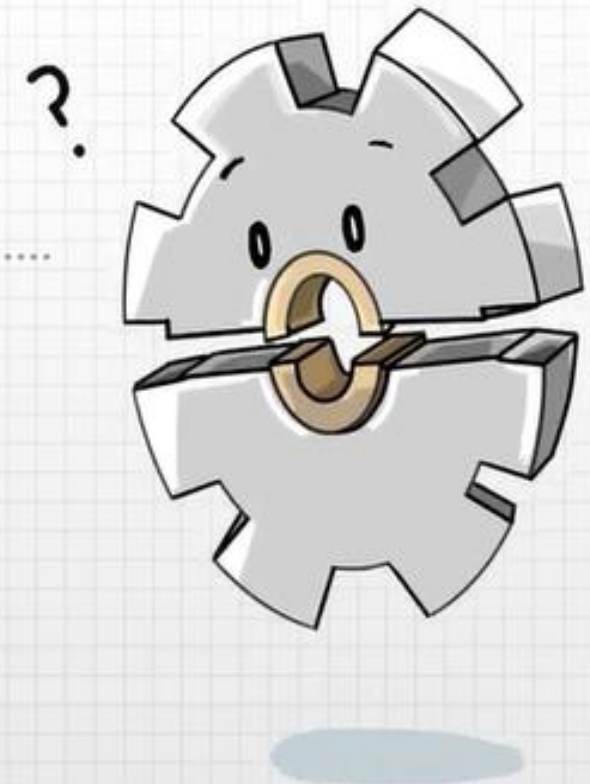
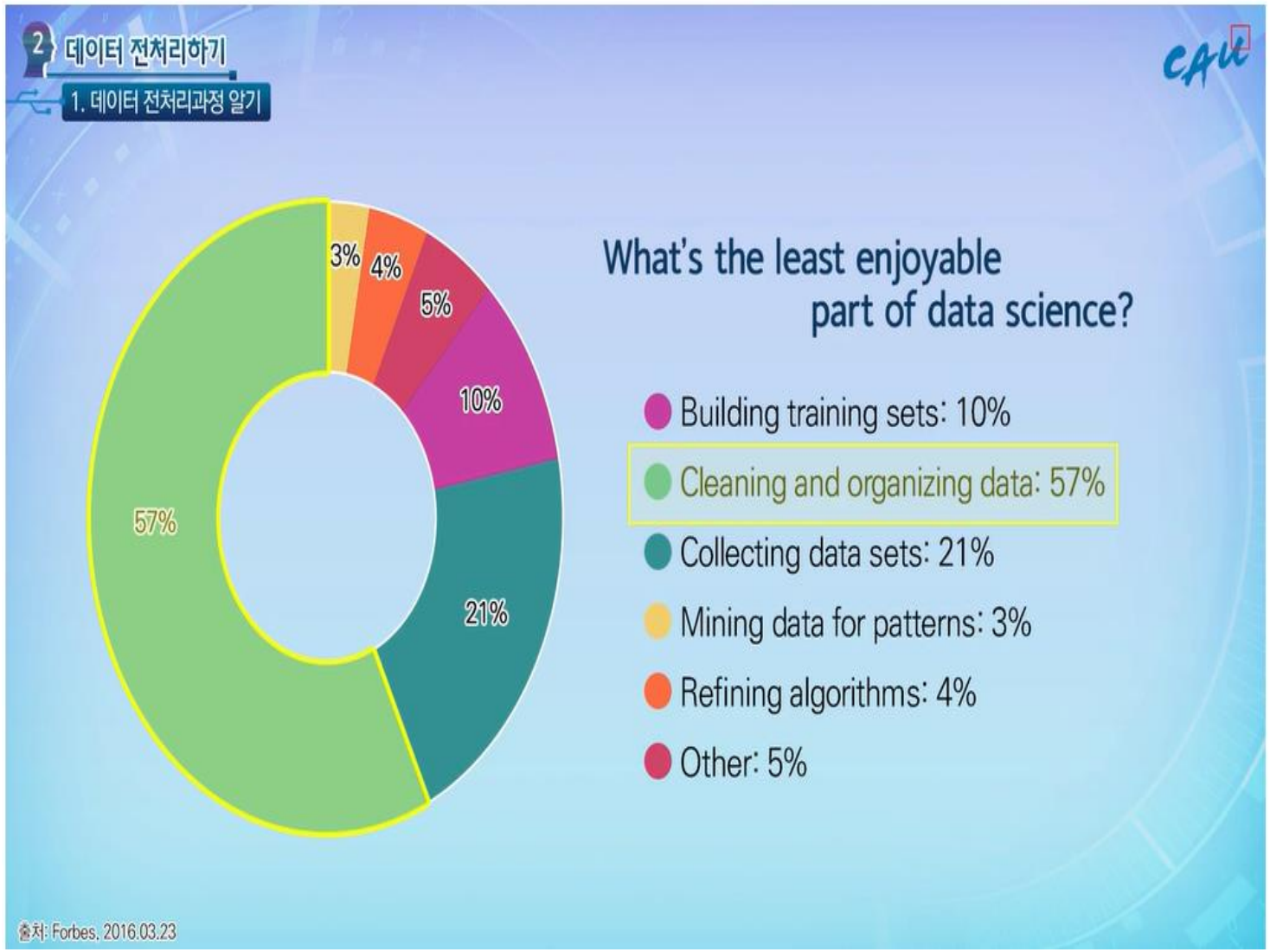
학습내용

- 데이터 전처리과정 알기
- 데이터 전처리하기(실습)



학습목표

- 데이터 전처리과정의 필요성에 대해 설명할 수 있다.
- 데이터 전처리 작업을 수행할 수 있다.


What's the least enjoyable part of data science?
3%
4%
5%
10%
57%
21%
Building training sets: 10%
Cleaning and organizing data: 57%
Collecting data sets: 21%
Mining data for patterns: 3%
Refining algorithms: 4%
Other: 5%
출처: Forbes, 2016.03.23

데이터 나누기

학습 데이터 세트(Training Dataset)

- 머신러닝 인풋

테스트 데이터 세트(Test Dataset)

- 모델의 성과 측정

데이터 나누기

선형 표본화(linear sampling)

- 데이터를 순차적으로 선별

무작위 표본화(random sampling)

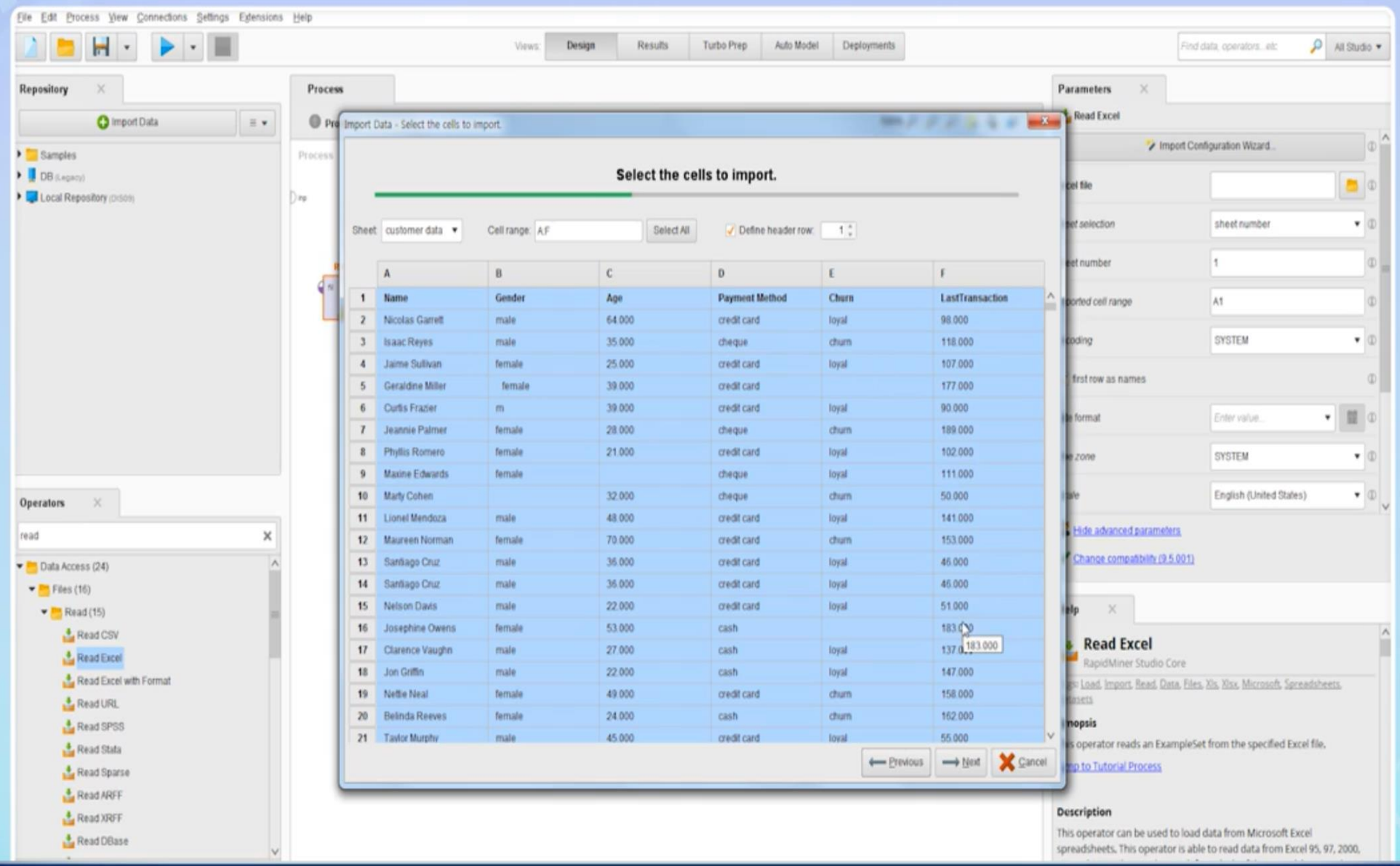
- 무작위로 랜덤하게 분류

층화 표본화(stratified sampling)

- 각각의 레이블별로 무작위 추출

- 값이 없는 경우: 결측치(Missing value)
  - 속성 사용 여부
  - 해당 사례만 제거
  - 결측치 예측

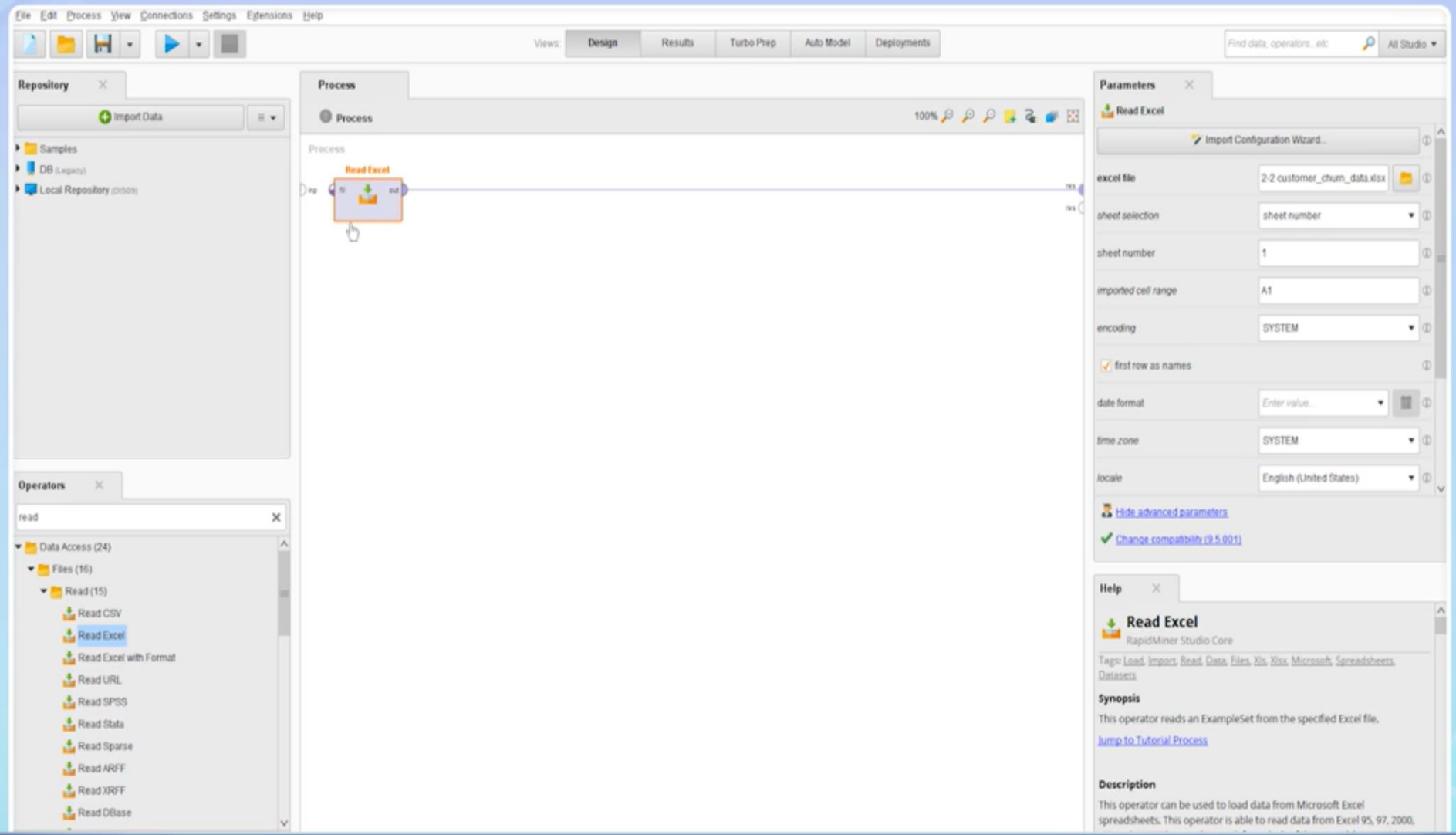
2. 데이터 전처리하기(실습)



	A	B	C	D	E	F
1	Name	Gender	Age	Payment Method	Churn	LastTransaction
2	Nicolas Garrett	male	64.000	credit card	loyal	98.000
3	Isaac Reyes	male	35.000	cheque	churn	118.000
4	Jaime Sullivan	female	25.000	credit card	loyal	107.000
5	Geraldine Miller	female	39.000	credit card		177.000
6	Curtis Frazier	m	39.000	credit card	loyal	90.000
7	Jeannie Palmer	female	28.000	cheque	churn	189.000
8	Phyllis Romero	female	21.000	credit card	loyal	102.000
9	Maxine Edwards	female		cheque	loyal	111.000
10	Marty Cohen		32.000	cheque	churn	50.000
11	Lionel Mendoza	male	48.000	credit card	loyal	141.000
12	Maureen Norman	female	70.000	credit card	churn	153.000
13	Santiago Cruz	male	36.000	credit card	loyal	46.000
14	Santiago Cruz	male	36.000	credit card	loyal	46.000
15	Nelson Davis	male	22.000	credit card	loyal	51.000
16	Josephine Owens	female	53.000	cash		183.000
17	Clarence Vaughn	male	27.000	cash	loyal	137.000
18	Jon Griffin	male	22.000	cash	loyal	147.000
19	Nettie Neal	female	49.000	credit card	churn	158.000
20	Belinda Reeves	female	24.000	cash	churn	162.000
21	Taylor Murphy	male	45.000	credit card	loyal	55.000

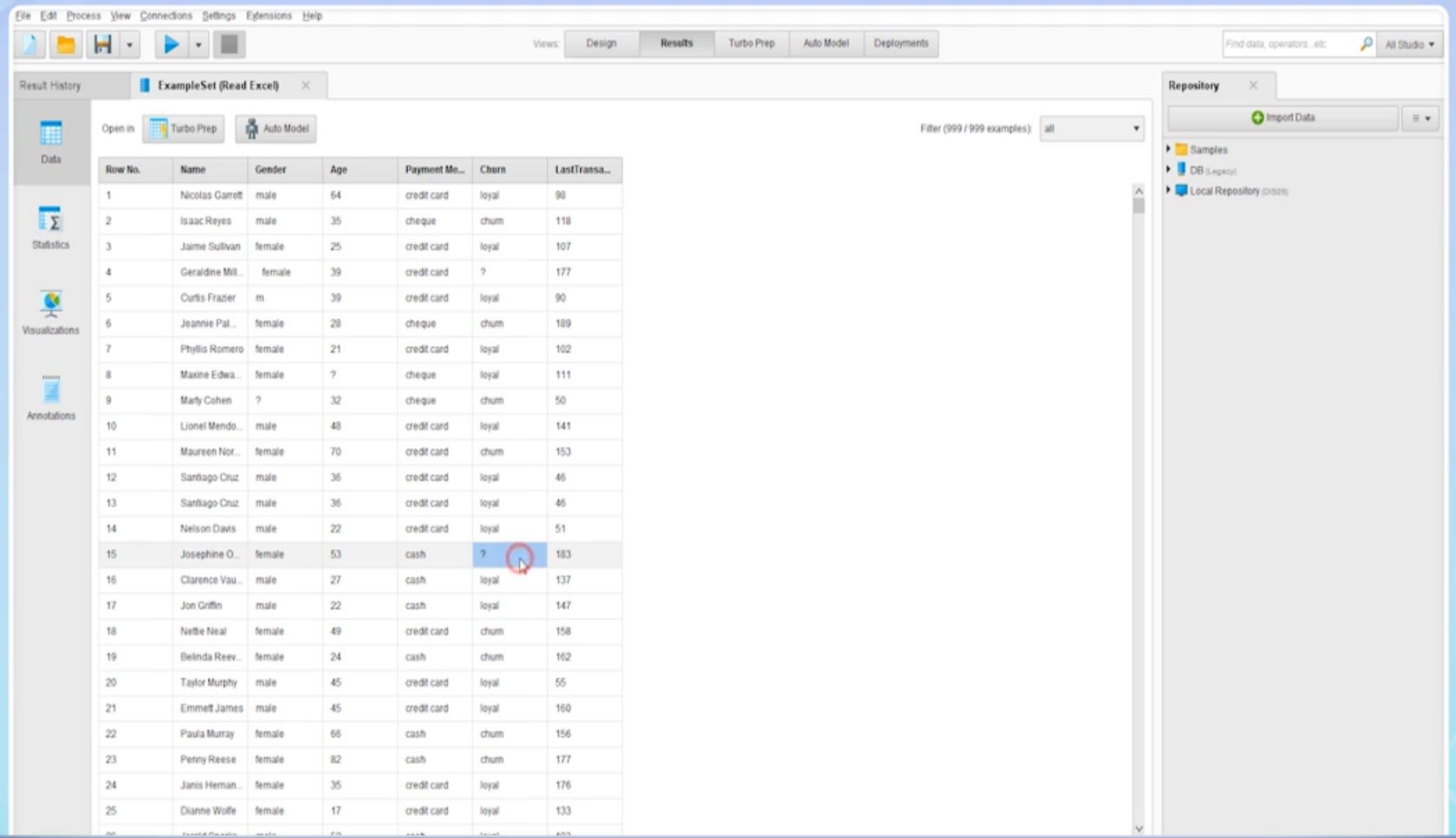
필요한 데이터 가져오기

2. 데이터 전처리하기(실습)



read excel의 아웃풋과 패널의 result port 연결

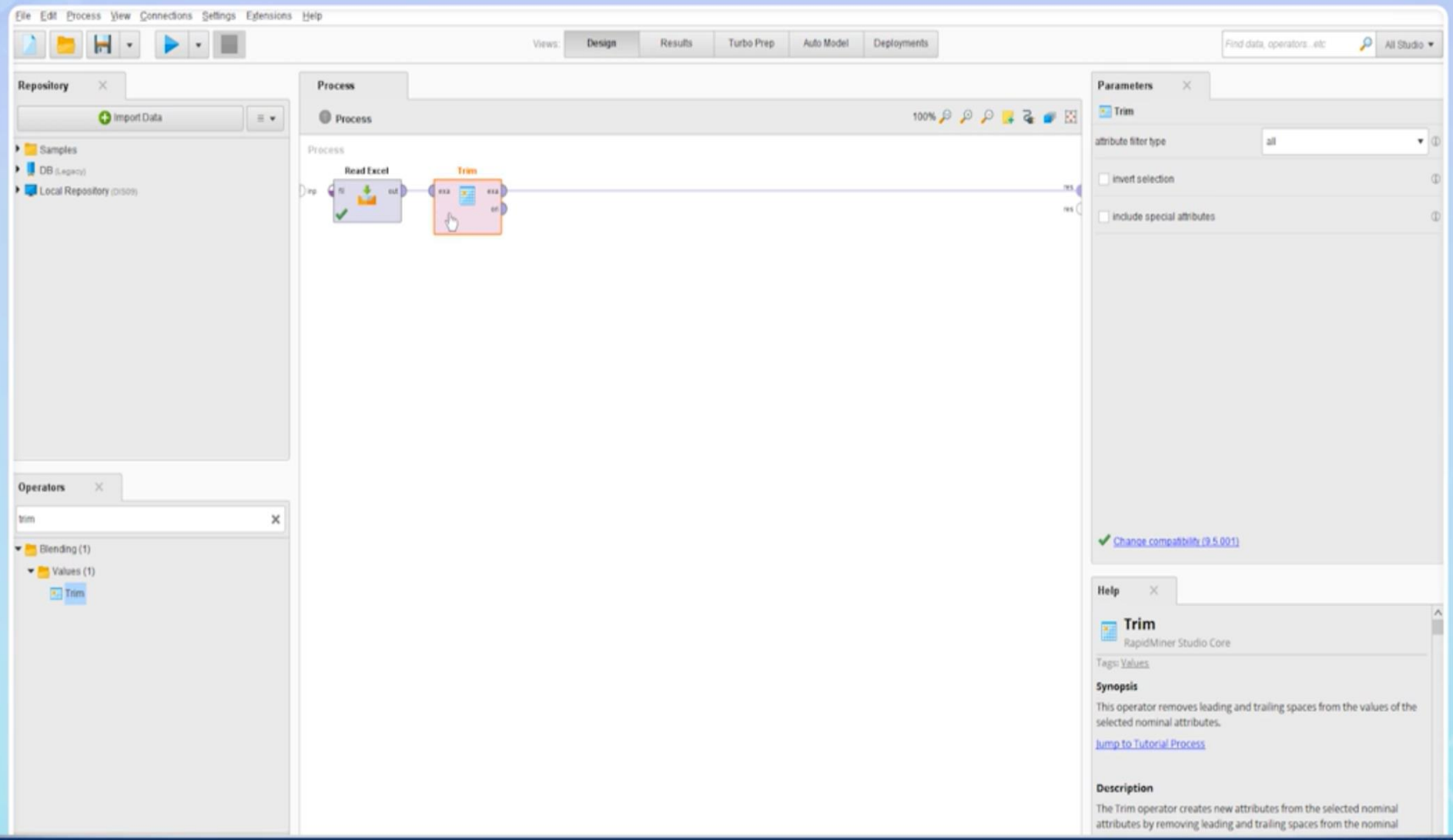
2. 데이터 전처리하기(실습)



Row No.	Name	Gender	Age	Payment Me...	Churn	LastTransa...
1	Nicolas Garrett	male	64	credit card	loyal	98
2	Isaac Reyes	male	35	cheque	churn	118
3	Jaime Sullivan	female	25	credit card	loyal	107
4	Geraldine Mill...	female	39	credit card	?	177
5	Curtis Frazier	m	39	credit card	loyal	90
6	Jeannie Pal...	female	28	cheque	churn	189
7	Phyllis Romero	female	21	credit card	loyal	102
8	Maxine Edwa...	female	?	cheque	loyal	111
9	Marty Cohen	?	32	cheque	churn	50
10	Lionel Mendo...	male	48	credit card	loyal	141
11	Maureen Nor...	female	70	credit card	churn	153
12	Santiago Cruz	male	36	credit card	loyal	46
13	Santiago Cruz	male	36	credit card	loyal	46
14	Nelson Davis	male	22	credit card	loyal	51
15	Josephine O...	female	53	cash	?	183
16	Clarence Vau...	male	27	cash	loyal	137
17	Jon Griffin	male	22	cash	loyal	147
18	Nettie Neal	female	49	credit card	churn	158
19	Belinda Reev...	female	24	cash	churn	162
20	Taylor Murphy	male	45	credit card	loyal	55
21	Emmett James	male	45	credit card	loyal	160
22	Paula Murray	female	66	cash	churn	156
23	Penny Reese	female	82	cash	churn	177
24	Janis Herman...	female	35	credit card	loyal	176
25	Dianne Wolfe	female	17	credit card	loyal	133

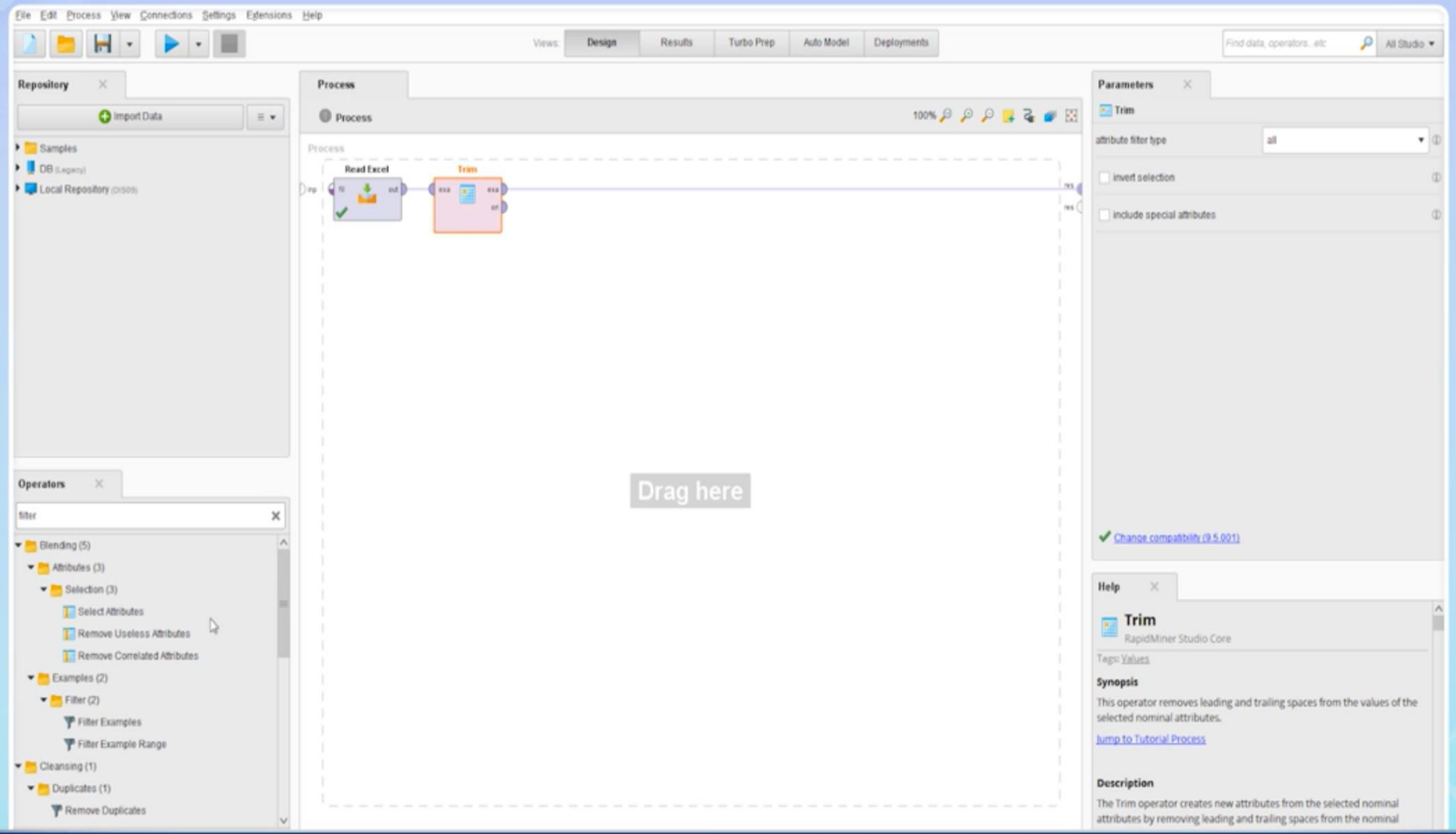
실행 후 데이터 확인

2. 데이터 전처리하기(실습)



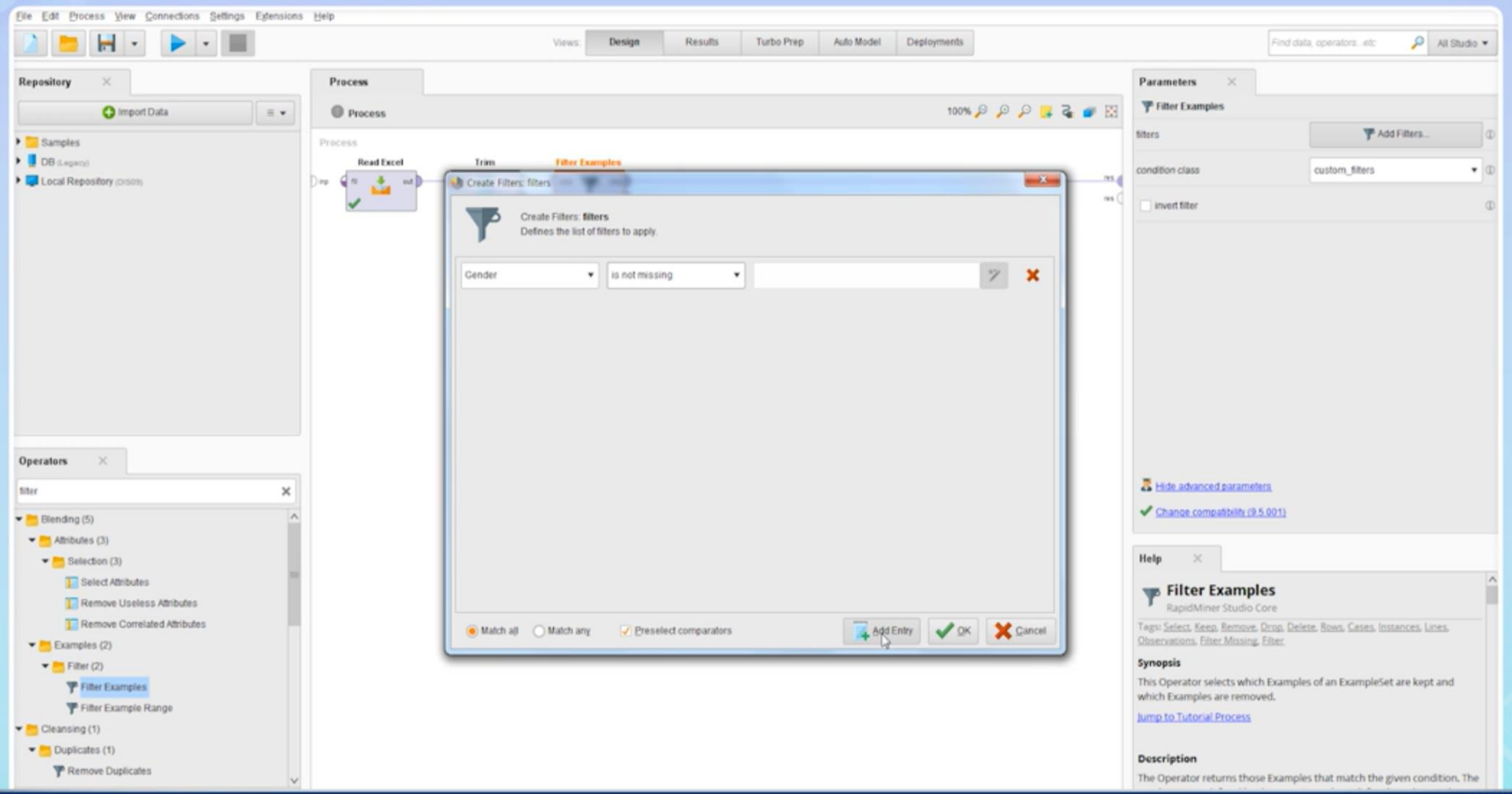
trim을 가져온 후 read excel 옆에 드래그 및 연결

2. 데이터 전처리하기(실습)



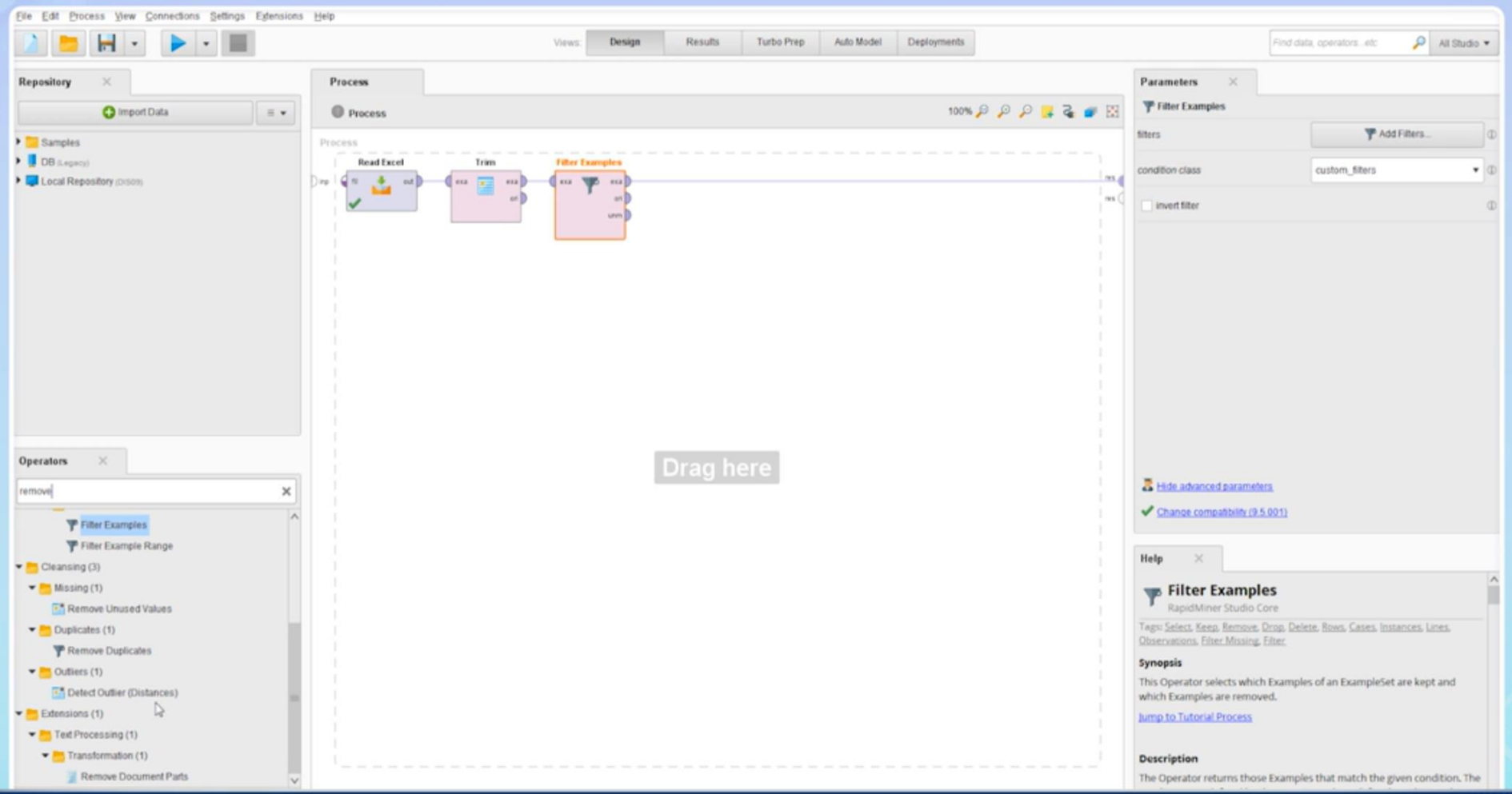
filter example 오퍼레이터를 사용하여 결측치 처리하기

2. 데이터 전처리하기(실습)



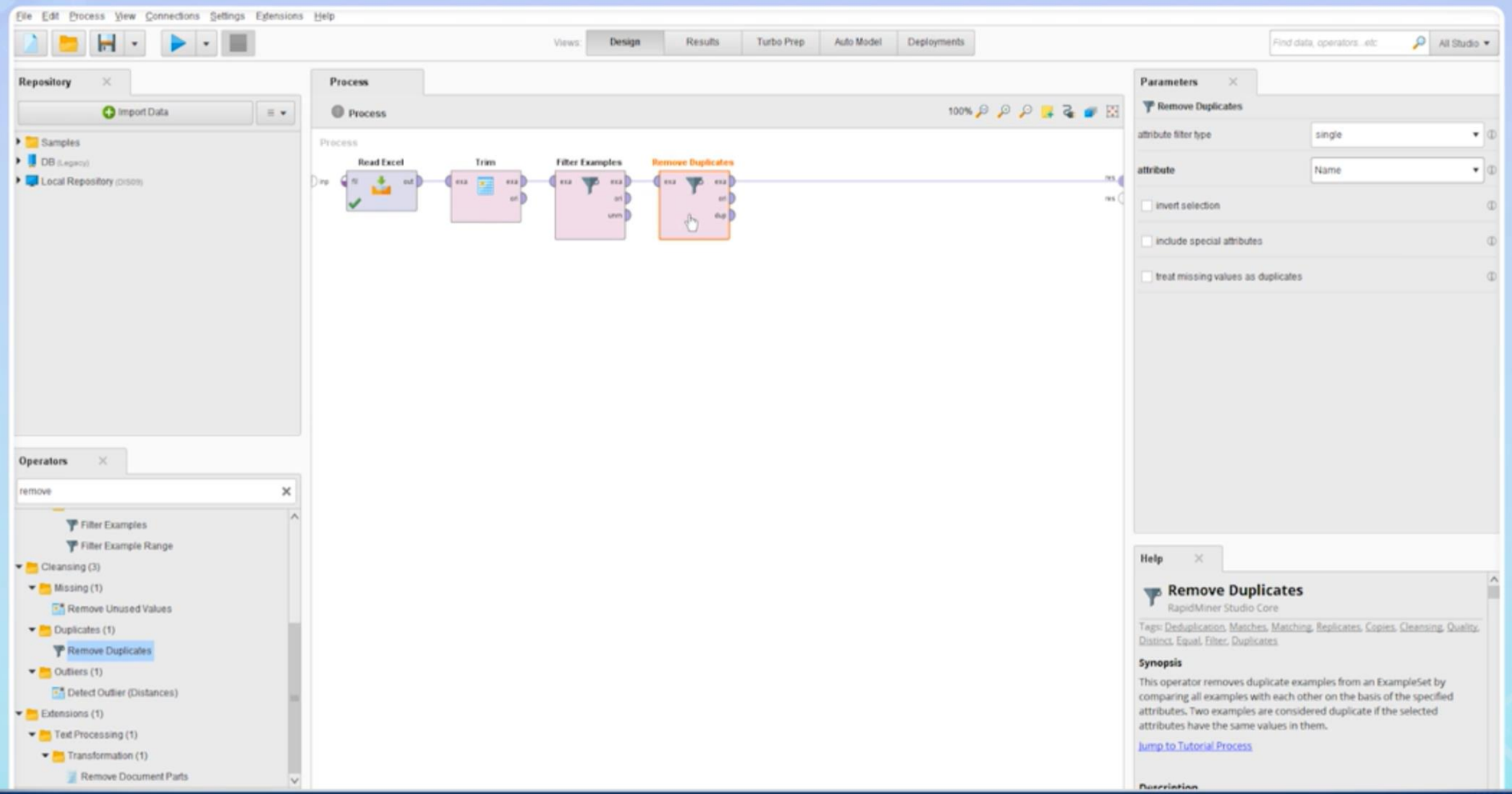
add filter → is not missing 선택

2. 데이터 전처리하기(실습)



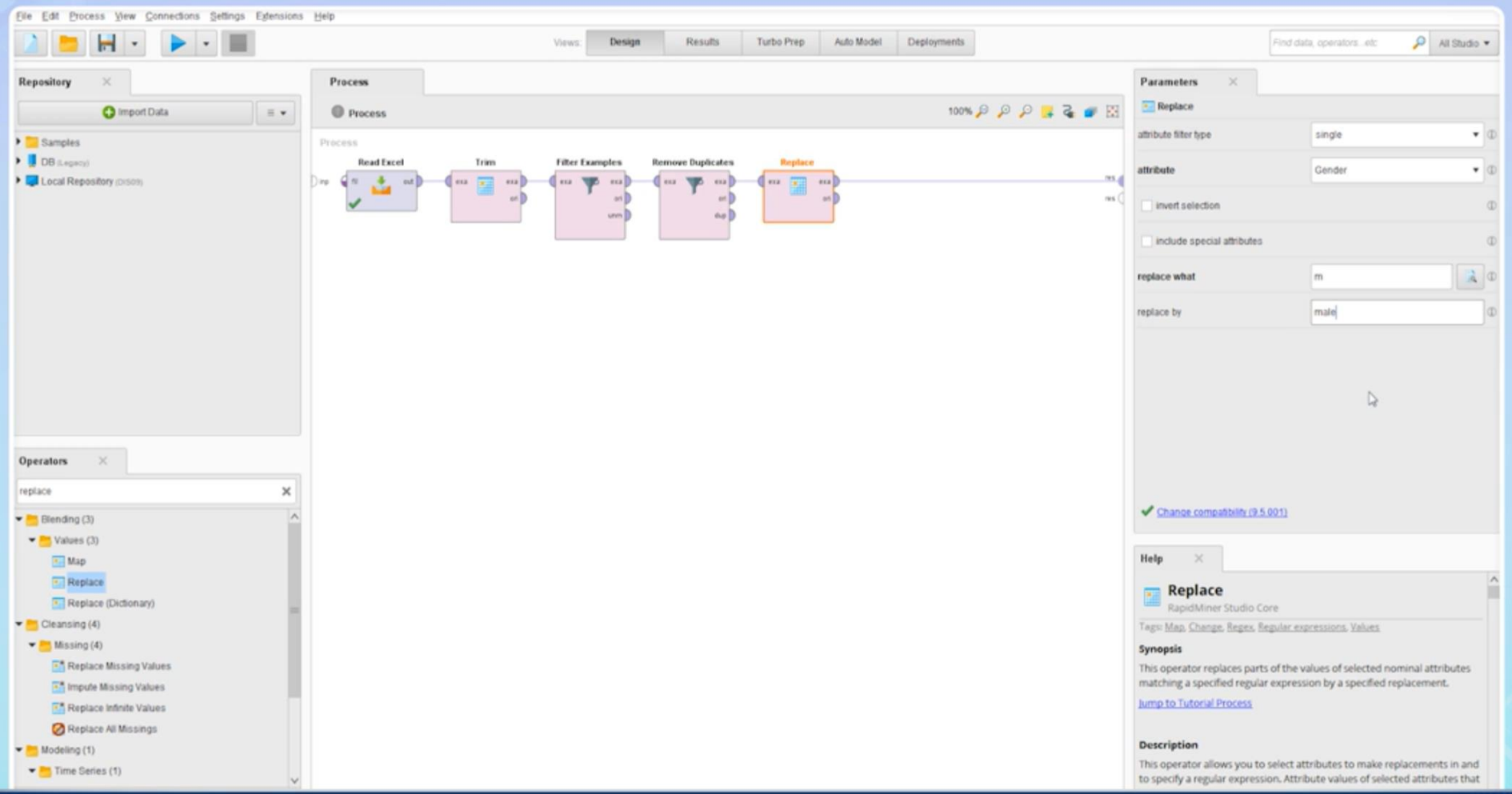
remove duplicate 오퍼레이터 사용하여 중복 데이터 없애기

2. 데이터 전처리하기(실습)



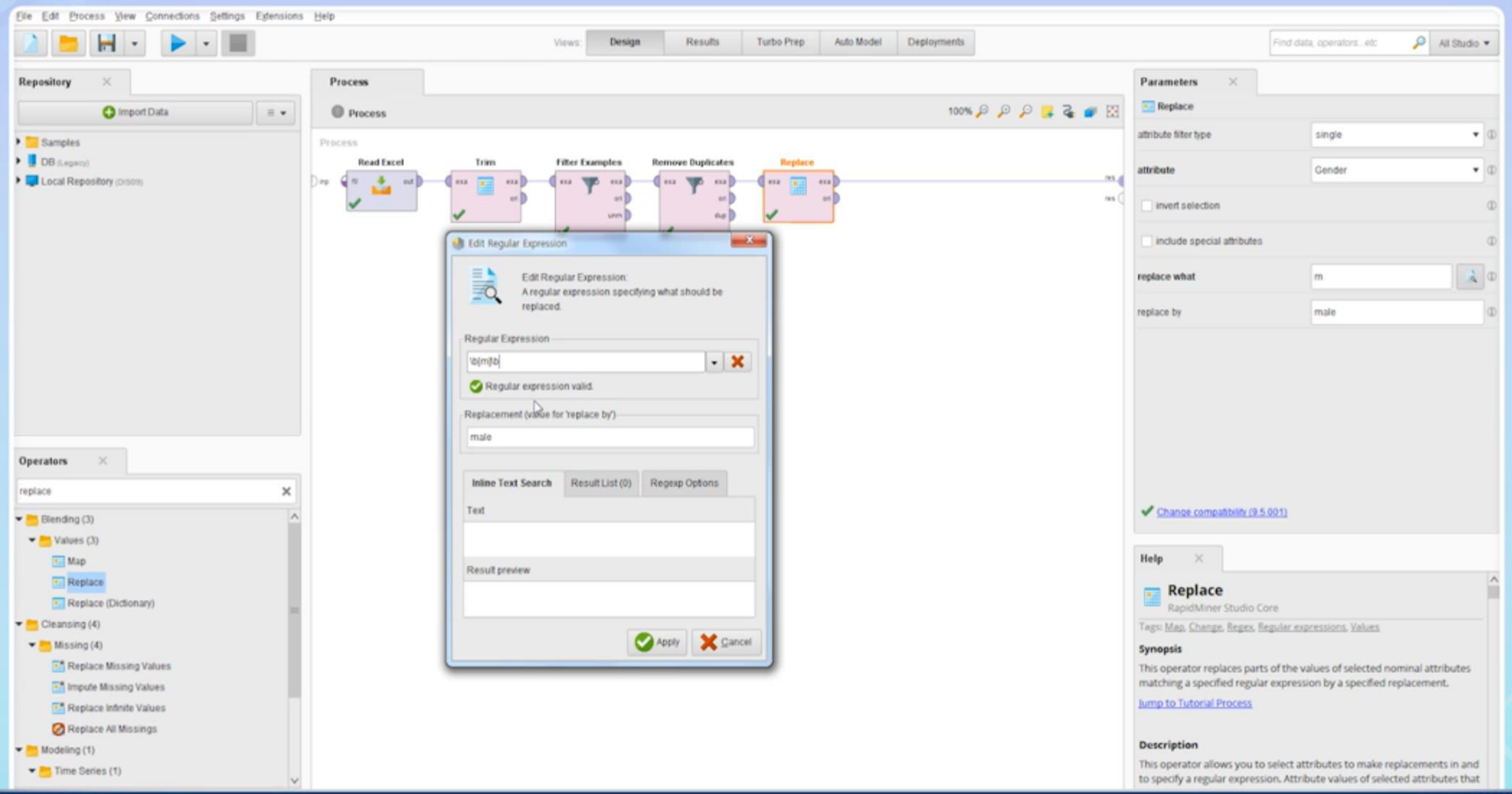
파라미터 설정으로 single 선택 후 중복되는 데이터 없애기

2. 데이터 전처리하기(실습)



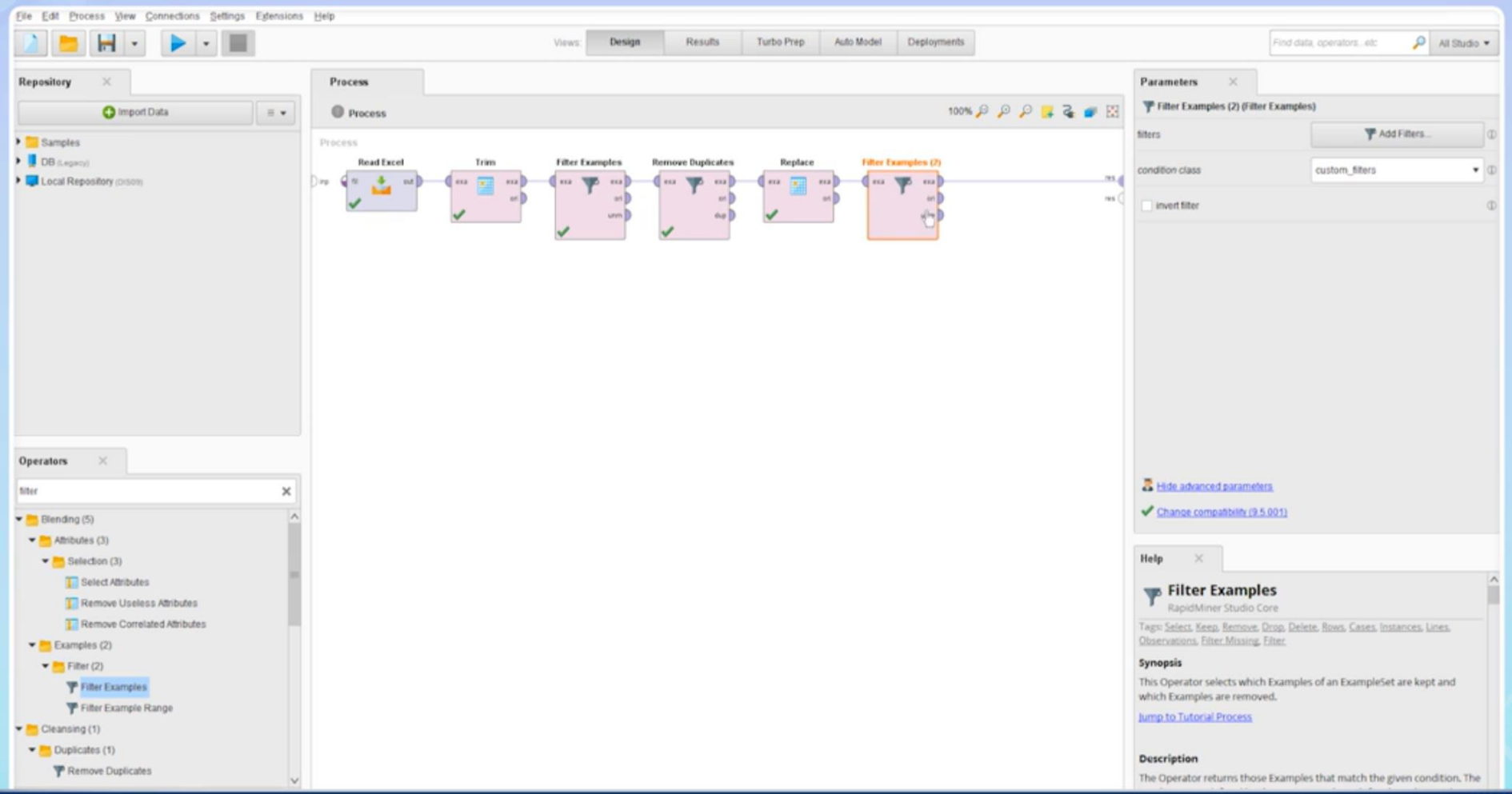
m을 male의 오타로 상정

2. 데이터 전처리하기(실습)



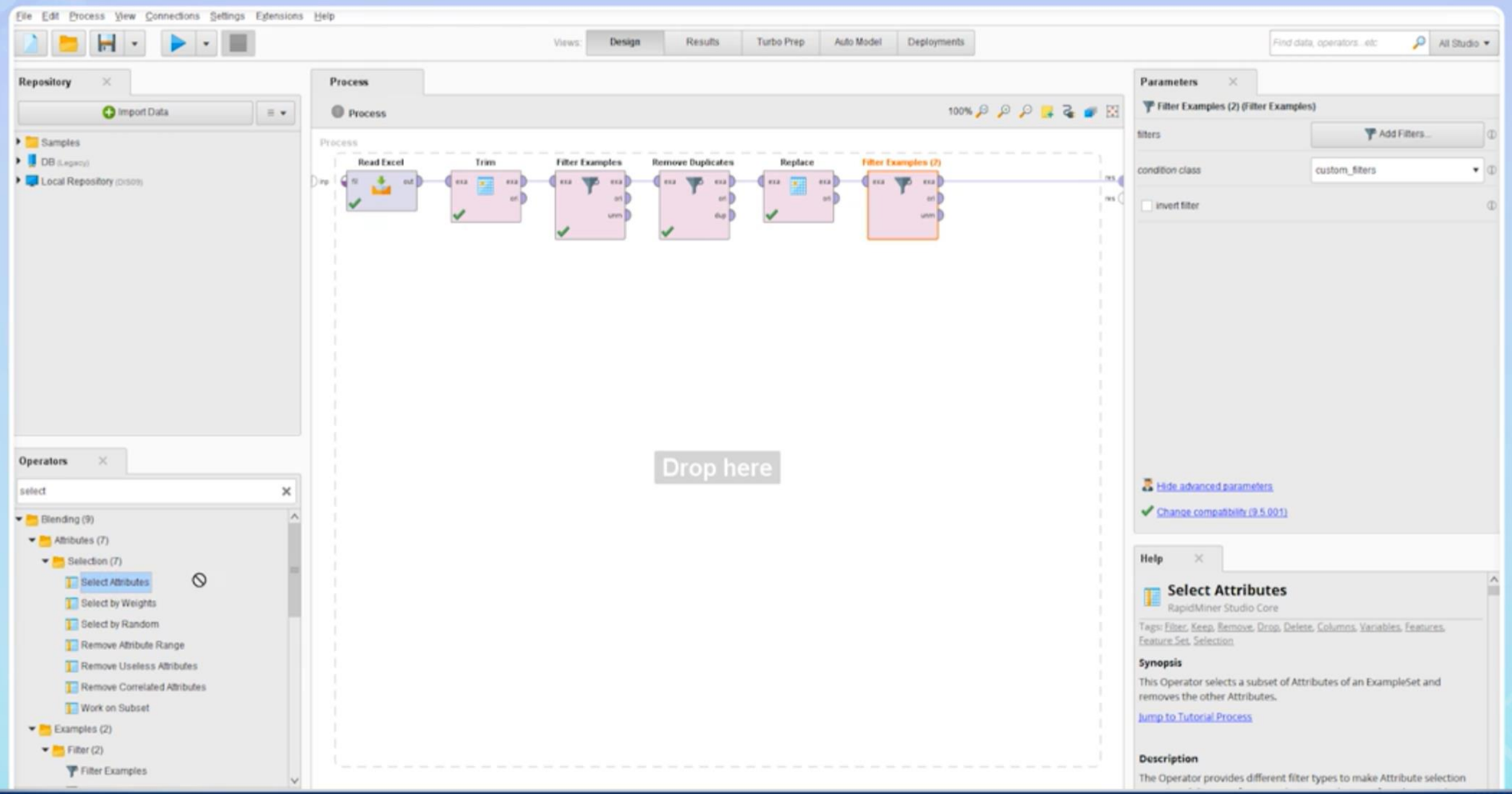
단독으로 m이 있는 것만 바꾸기 위해 /b 사용

2. 데이터 전처리하기(실습)



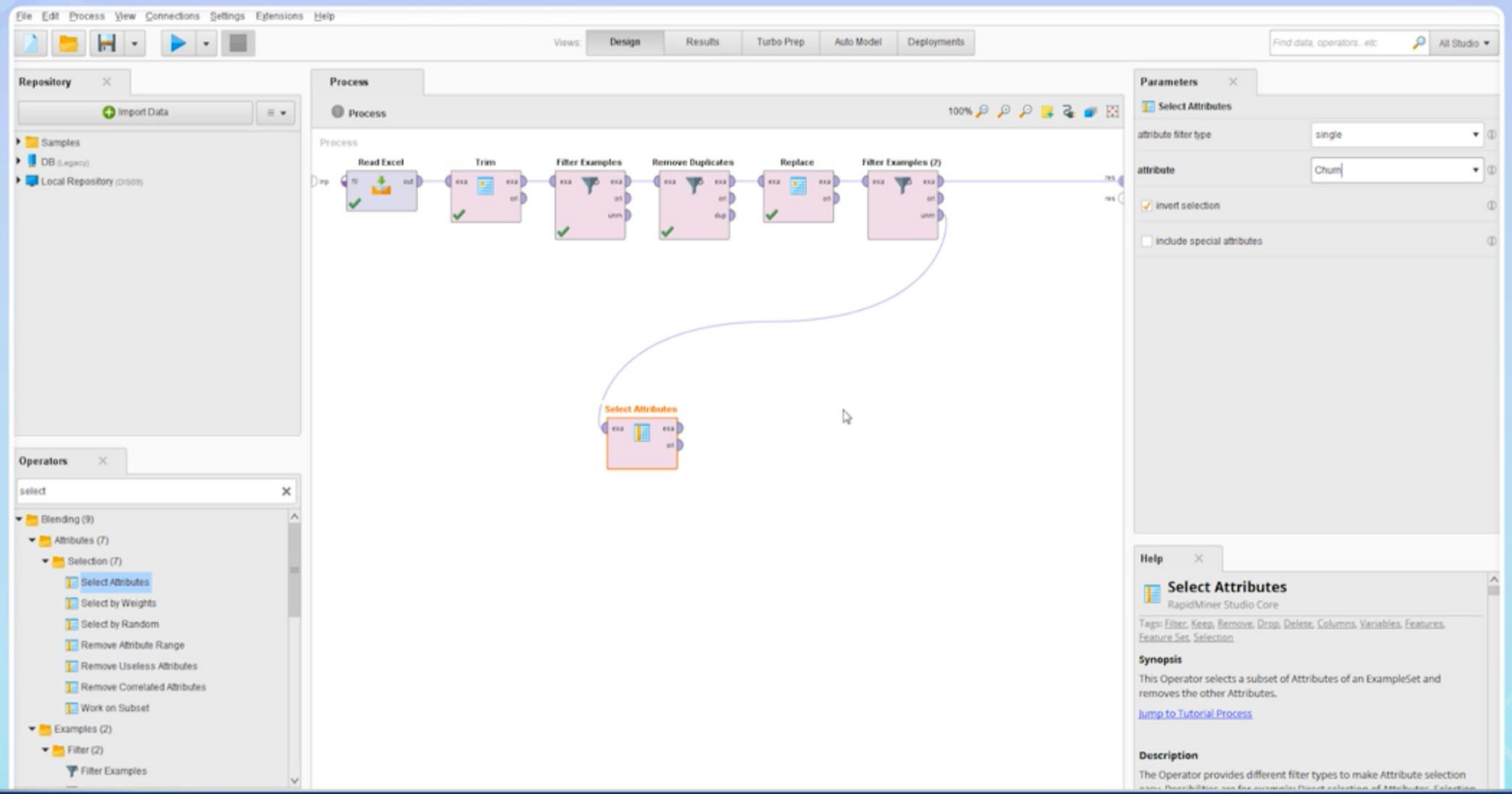
Churn 데이터가 있는 것만 가져오기

2. 데이터 전처리하기(실습)



select attributes 오퍼레이터 사용

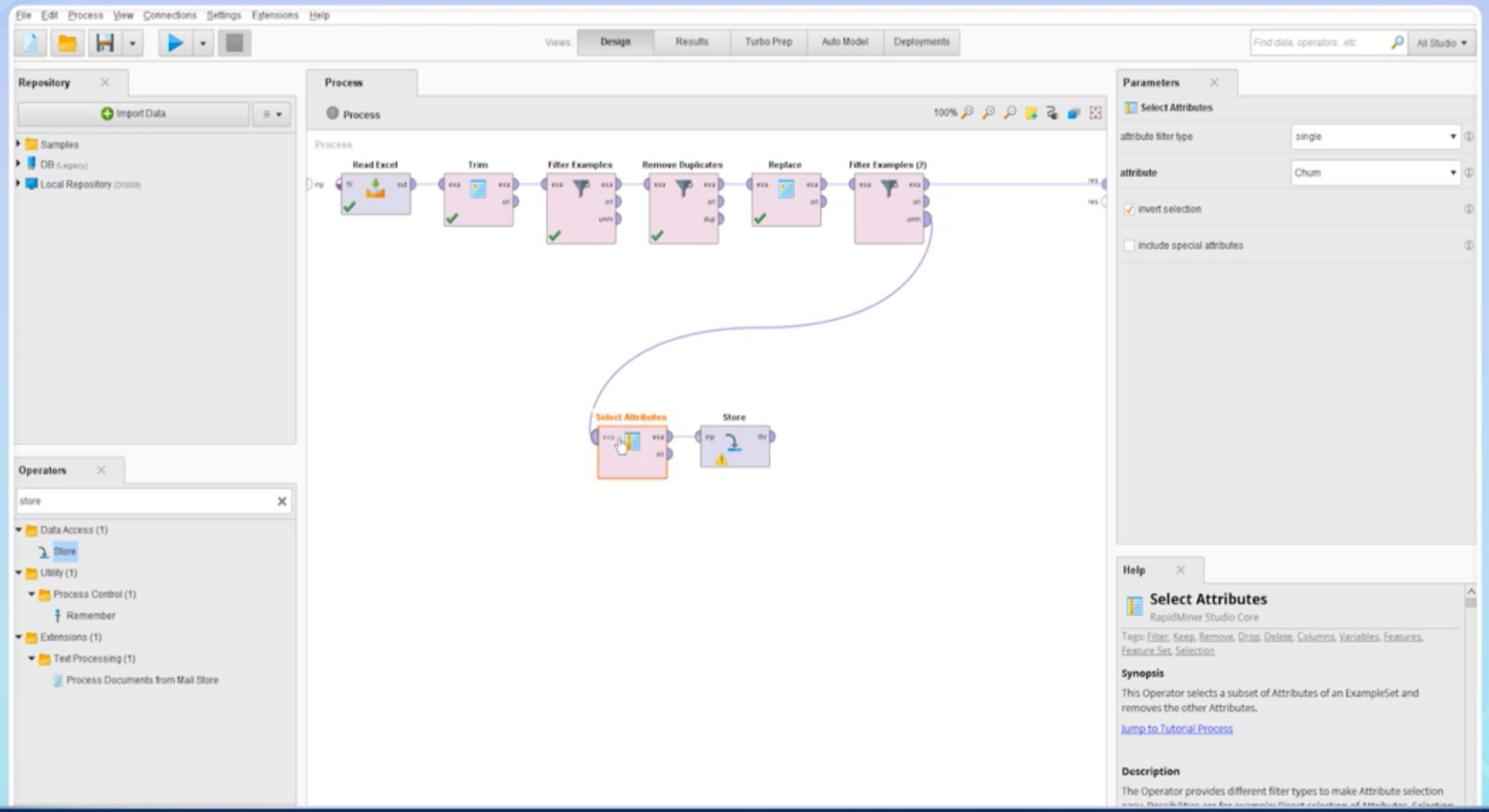
2. 데이터 전처리하기(실습)



churn 데이터를 선택해서 속성 자체를 데이터에서 없앰

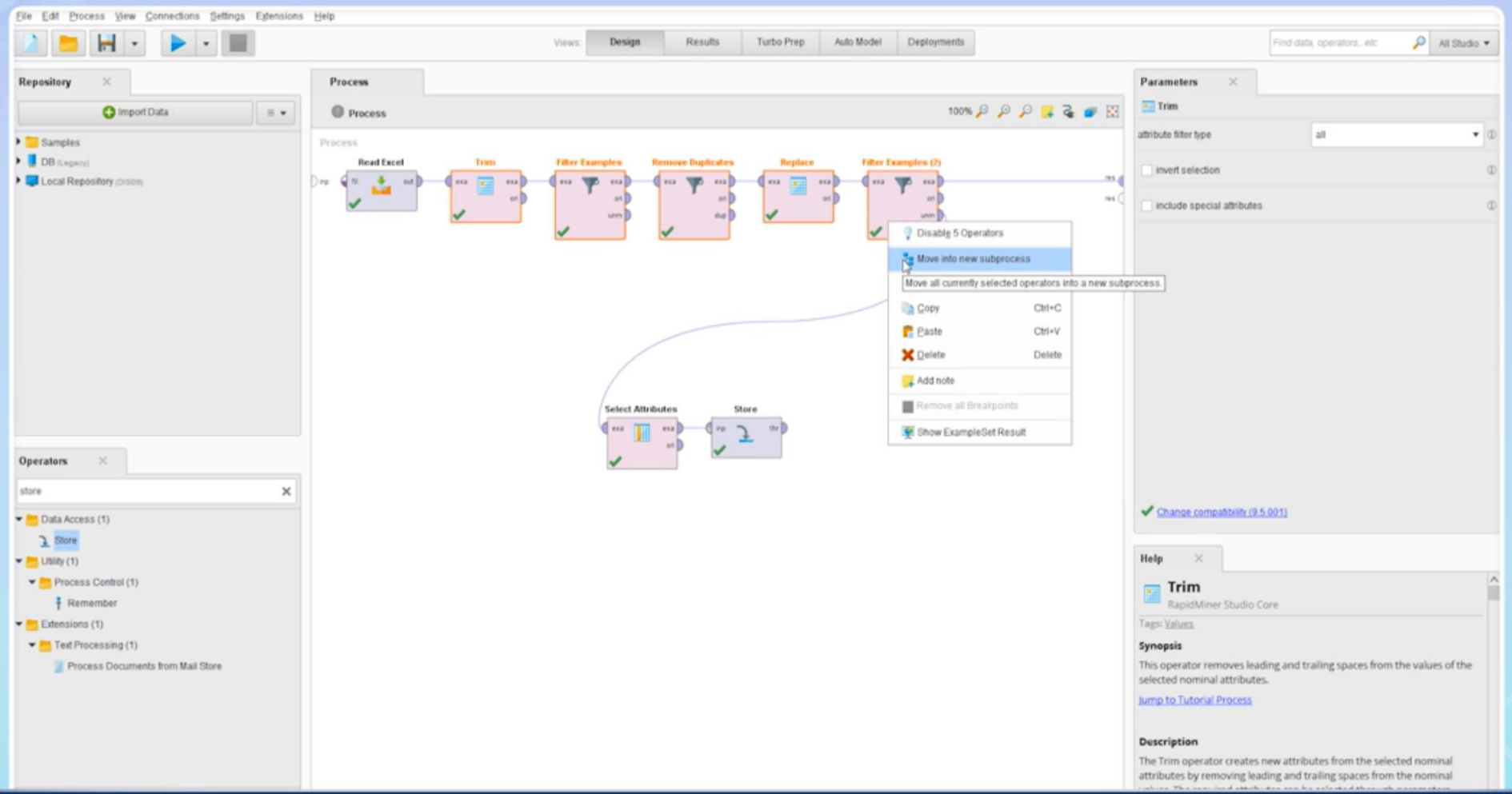
2 데이터 전처리하기

2. 데이터 전처리하기(실습)



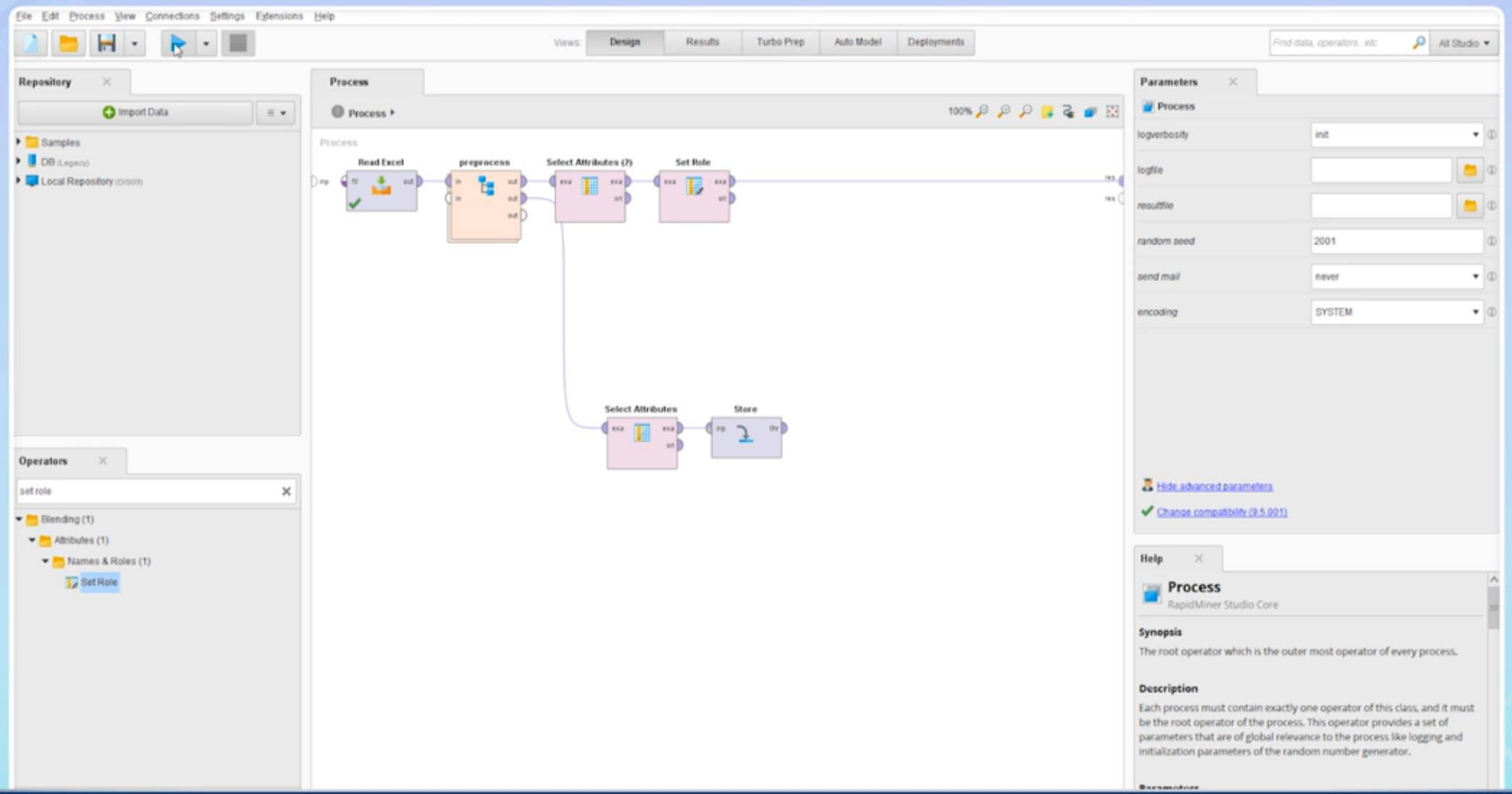
store 오퍼레이터 사용, 저장 위치 지정

2. 데이터 전처리하기(실습)



묶고자 하는 서브프로세스를 선택 → 우클릭 → move into new sub-process 클릭

2. 데이터 전처리하기(실습)



set role 오퍼레이터를 사용하여 레이블로 지정