10일차 과제

[DAY10]

드디어 신병훈련소 마지막 날이 밝았습니다!!!! 그동안 과제 하시느라 고생 많으셨어요.
마지막 날 과제까지 잘 ~~~~ 마무리하시고 유종의 미를 거두셔야겠죠?!

오늘은 Tableau Prep Builder를 이용해서 데이터를 정리해 보는 날입니다.

Tableau Prep Builder 2022.1.1 다운로드
https://www.tableau.com/ko-kr/support/releases/prep/2022.1.1#esdalt

아래 정리할 파일을 다운로드 받아주세요.
Prep Data.zip

우리는 4개년 치 매출데이터와 지역별 관리자, 반품 데이터를 하나로 묶어서 데이터를 분석할 수 있도록 할꺼에요.

먼저 Union과 Join의 개념을 짚고 시작할까요?

UNION
Union은 같은 구조를 가진 집합(테이블) 여러 개를 하나의 집합(테이블)으로 합치는 것을 말합니다.
여기서 말하는 같은 구조라 함은 동일한 필드 개수, 필드명, 필드의 데이터 타입을 의미합니다.

아래와 같은 형태가 Union 이 되겠죠? 데이터가 밑으로 붙는 형태일꺼에요.

주문일자	주문번호	매출
2019-01-01	1	2000
2019-01-02	2	100
2019-01-03	3	500

+

주문일자	주문번호	매출
2020-01-03	4	1000
2020-01-04	5	500
2020-01-05	6	100

→

주문일자	주문번호	매출
2019-01-01	1	2000
2019-01-02	2	100
2019-01-03	3	500
2020-01-03	4	1000
2020-01-04	5	500
2020-01-05	6	100

JOIN
Join 은 기준 필드를 가지고 조인 형태에 따라, 두 개 이상의 집합을 연결 또는 결합하여 데이터를 출력하는 것을 말합니다.

1. 안쪽 (Inner) 조인

아래와 같이 “지역”을 기준으로 “안쪽 (Inner) 조인”을 했을 경우, 아래와 같은 결과가 출력되겠죠. 조인은 데이터가 옆으로 붙는 형태인 것이 보이시나요.

주문일자	주문번호	매출	지역
2019-01-01	1	2000	수도권
2019-01-02	2	100	호남
2019-01-03	3	500	영남
2019-01-04	4	3000	수도권
2019-01-04	5	2000	서울경기

지역	관리자
수도권	김성식
호남	정신
영남	최진수
강원	박진석
충청	금나나

주문일자	주문번호	매출	지역	지역	관리자
2019-01-01	1	2000	수도권	수도권	김성식
2019-01-02	2	100	호남	호남	정신
2019-01-03	3	500	영남	영남	최진수
2019-01-04	4	3000	수도권	수도권	김성식

[여기서 잠깐!!]
주문번호 5번이 조인 결과에서 제외된 이유는 오른쪽 집합에 서울경기 값이 없기 때문에 **두 개의 집합이 조인될 값이 없으니** 당연히 결과에서 빠지게 되는거죠!

2. 왼쪽 (Left Outer) 조인

"지역"을 기준으로 "왼쪽 (Left Outer) 조인"을 했을 경우에는, 아래와 같은 결과를 얻을 수 있을거에요.

주문일자	주문번호	매출	지역
2019-01-01	1	2000	수도권
2019-01-02	2	100	호남
2019-01-03	3	500	영남
2019-01-04	4	3000	수도권
2019-01-04	5	2000	서울경기

지역	관리자
수도권	김성식
호남	정신
영남	최진수
강원	박진석
충청	금나나

주문일자	주문번호	매출	지역	지역	관리자
2019-01-01	1	2000	수도권	수도권	김성식
2019-01-02	2	100	호남	호남	정신
2019-01-03	3	500	영남	영남	최진수
2019-01-04	4	3000	수도권	수도권	김성식
2019-01-04	5	2000	서울경기	*null*	*null*

즉, **조인 값이 존재하지 않더라도 해당 되는 방향(왼쪽)의 데이터를 모두 가져오는 것이 되겠죠. inner 조인과의 차이점 아시겠죠?**

3. 오른쪽 (Right Outer) 조인

"지역"을 기준으로 "오른쪽 (Right Outer) 조인"을 했을 경우에는, 아래와 같은 결과를 얻을 수 있습니다.

주문일자	주문번호	매출	지역
2019-01-01	1	2000	수도권
2019-01-02	2	100	호남
2019-01-03	3	500	영남
2019-01-04	4	3000	수도권
2019-01-04	5	2000	서울경기

지역	관리자
수도권	김성식
호남	정신
영남	최진수
강원	박진석
충청	금나나

주문일자	주문번호	매출	지역	지역	관리자
2019-01-01	1	2000	수도권	수도권	김성식
2019-01-02	2	100	호남	호남	정신
2019-01-03	3	500	영남	영남	최진수
2019-01-04	4	3000	수도권	수도권	김성식
null	*null*	*null*	*null*	강원	박진석
null	*null*	*null*	*null*	충청	금나나

2번에서 살펴봤던 왼쪽 조인과 같이 **조인할 값이 존재하지 않더라도 해당 되는 방향(오른쪽)의 데이터를 모두 가져오게 되는거죠.**

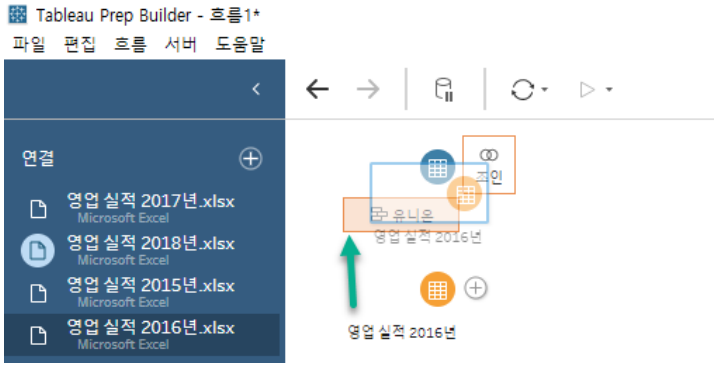
그럼 이러한 개념을 가지고 데이터 정리를 시작해볼까요?

1. 영업실적 2015년부터 2018년도 까지의 데이터를 Union 해주세요.

Union을 하는데 2가지 방법이 있으니, 둘 중 하나 선택해서 진행해 주세요.

1-1) **하나의 파일을 다른 파일 위로 드래그 & 드랍 해서 Union을 하는 방법**

왜 프렙 빌더의 유니온 놓는 위치가 아래에 있는지 위의 개념을 보셨으면 아시겠죠?



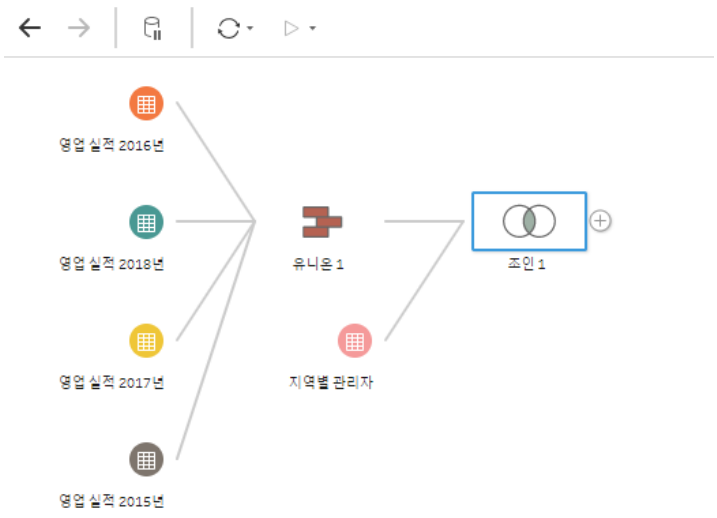
1-2) **와일드카드 유니온을 이용하는 방법**

드래그 & 드랍을 통한 유니온은 최대 10개의 파일까지만 가능하기 때문에,
10개보다 많은 파일을 유니온 할 때 **와일드 카드 유니온**을 유용하게 사용할 수 있습니다.

아래 보시는 것처럼 파일 뿐만 아니라 시트 수준에서도 유니온 옵션을 지정할 수 있어요. **별표 (*) 를 이용하셔서 여러 개의 파일을 유니온**

해보세요.

2. Union한 데이터와 지역별 관리자를 Join 해주세요.

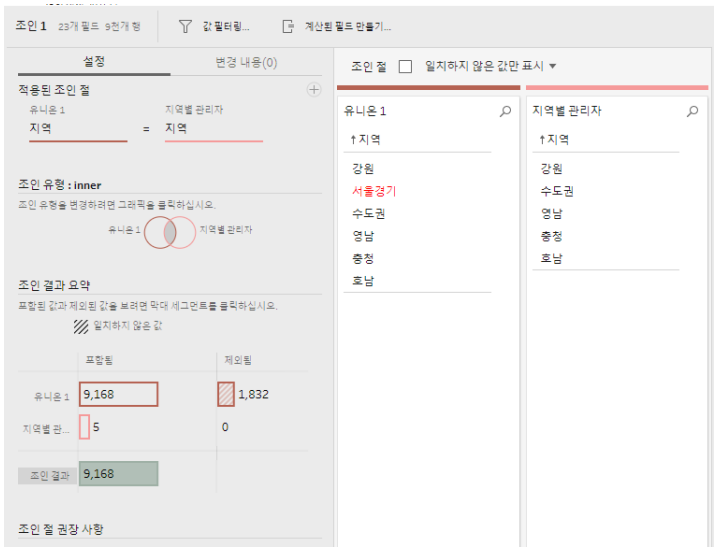


3. 조인 1 아이콘을 클릭해서 조인된 결과를 확인해 볼까요 ?

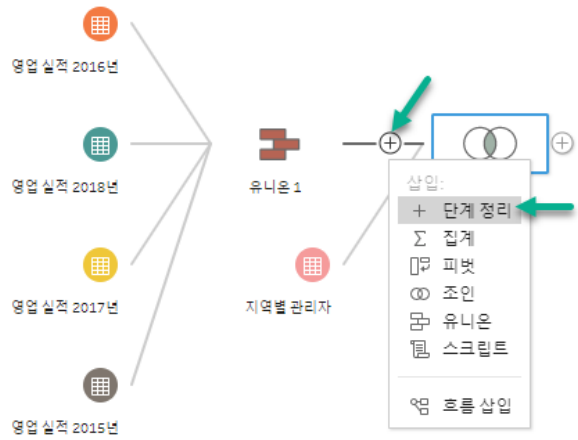
Union 된 데이터의 지역 값과 **지역별 관리자의 지역 값을 이용해서 Inner 조인**을 하도록 되어 있습니다.
그래서 최종 결과를 보니 1,832건이 제외되고, 총 9,168건이 조인이 되었네요.

일치하지 않은 값을 보니 유니온 1에는 "서울경기"라는 값이 있는데 지역별 관리자에는 "서울경기" 값이 없다보니 "서울경기" 값을 가진 행은 조인이 되지 않은거죠.

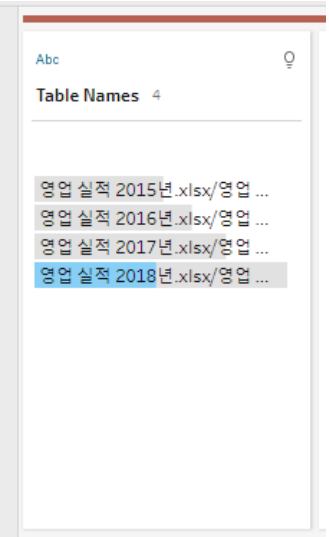
유니온 1 데이터에 무슨 일이 발생했는지 한 번 살펴볼까요?



4. **유니온 1**과 **조인 1** 사이에 단계를 추가해주세요.



5. 정리 1에서 지역 필드로 한 번 가볼까요?
지역 필드에서 "서울경기" 값을 클릭하고, **Table Names 필드**를 살펴보았더니 아래와 같이 영업 실적 2018년 엑셀 데이터에만 "서울경기" 값이 포함된 것을 볼 수가 있네요. 아마 2018년도에 잘못된 값이 들어온 것 같습니다. 값을 정리해줘야 할 것 같아요.



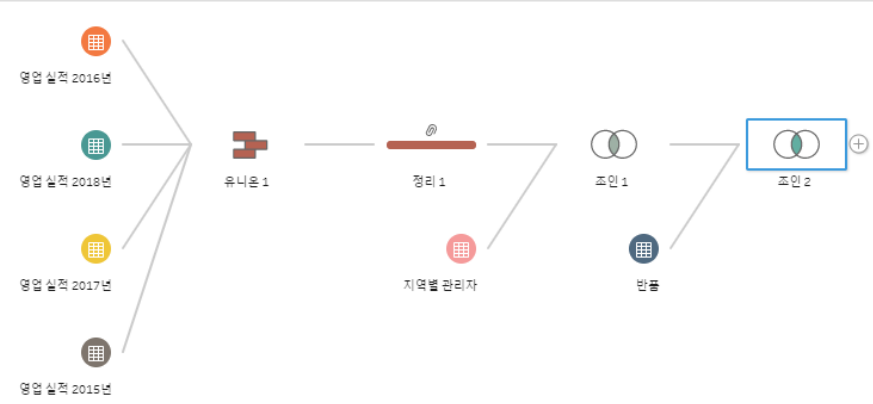
6. 정리 1에서 다시 한 번 지역 필드로 가서 값을 정리해보도록 합시다.
"서울경기"는 "수도권"에 포함된다는 것을 눈치채셨을텐데요. 2개의 값을 묶어 주는데 2가지 방법이 있습니다.

둘 중 한가지 선택해서 작업을 진행하세요.

6-1) "**서울경기" 클릭** 후, **Ctrl 키 누른 후**에 "**수도권" 클릭 → 수도권에서 마우스 오른쪽 버튼 클릭 → 그룹 클릭**
6-2) **"서울경기" 더블 클릭 → "수도권" 입력 후 엔터**

다시 조인 1로 돌아가서 결과를 확인해 보시면 11,000건이 잘 조인된 것을 보실 수 있나요?

7. 다음으로 반품 데이터를 조인해보도록 해요.

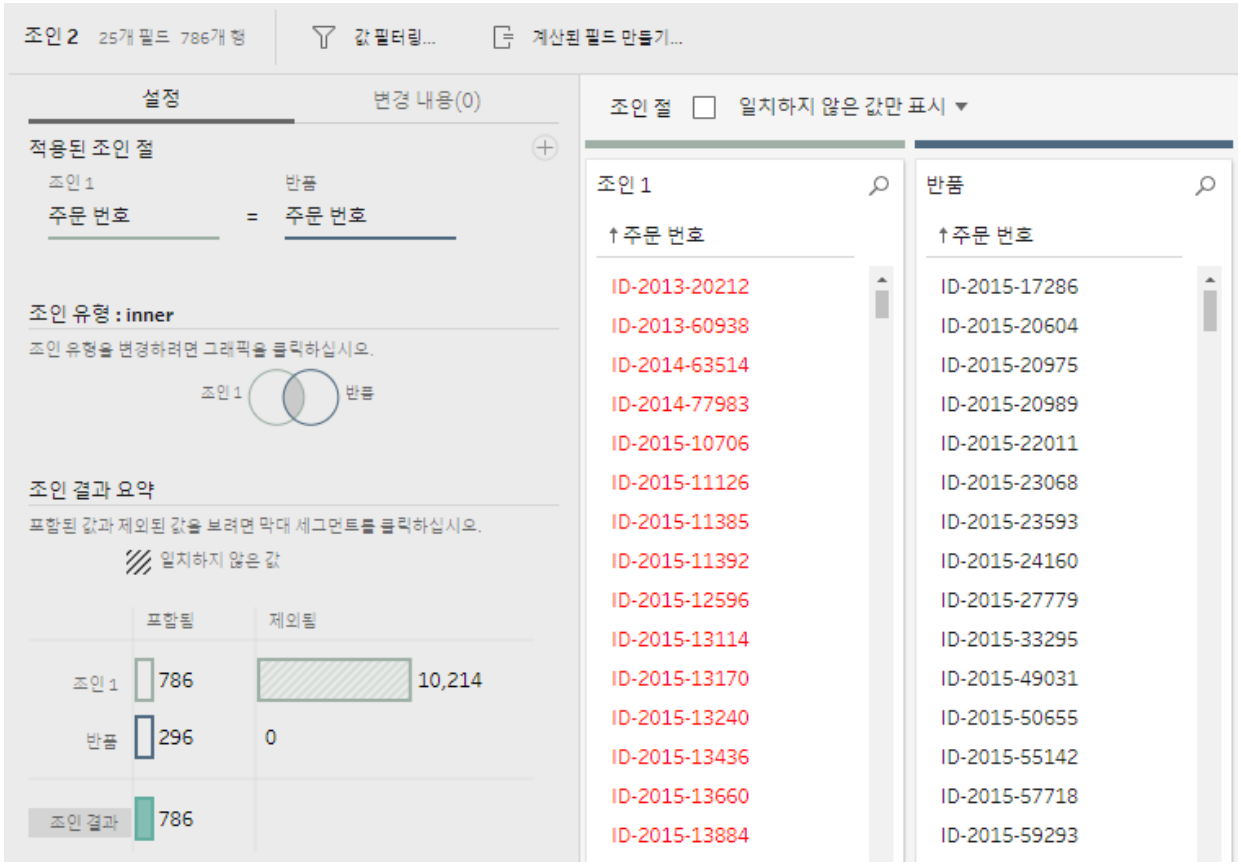


8. 조인 2 아이콘을 클릭해서 조인된 결과를 확인해 볼까요 ?

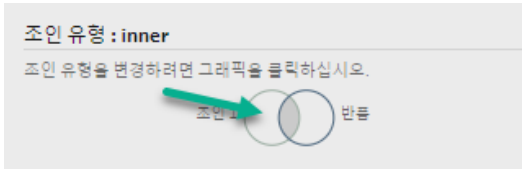
조인 1까지 완료된 데이터의 "주문 번호" 값과 **반품의 "주문 번호" 값을 이용해서 Inner 조인**을 하게 되어있고,
최종 결과를 살펴보니 10,214건이 제외되고, 총 786건이 조인이 되었네요.

우리가 원하던 결과가 맞을까요?
우리는 전체 4개년 치 데이터에 반품 정보를 결합해야 하는데, 지금은 전체 4개년 치 데이터가 아니라 반품 데이터에 있는 주문번호와 조인되는 일부 데이터만 가져오는 형태입니다. 앞에서 조인 개념에서 살펴봤던 내용 기억하시나요?

즉, 우리는 왼쪽에 있는 조인 1의 데이터를 모두 가져오면서 반품 정보를 결합해야 하니, 조인 형태를 변경해줘야 할 것 같아요.

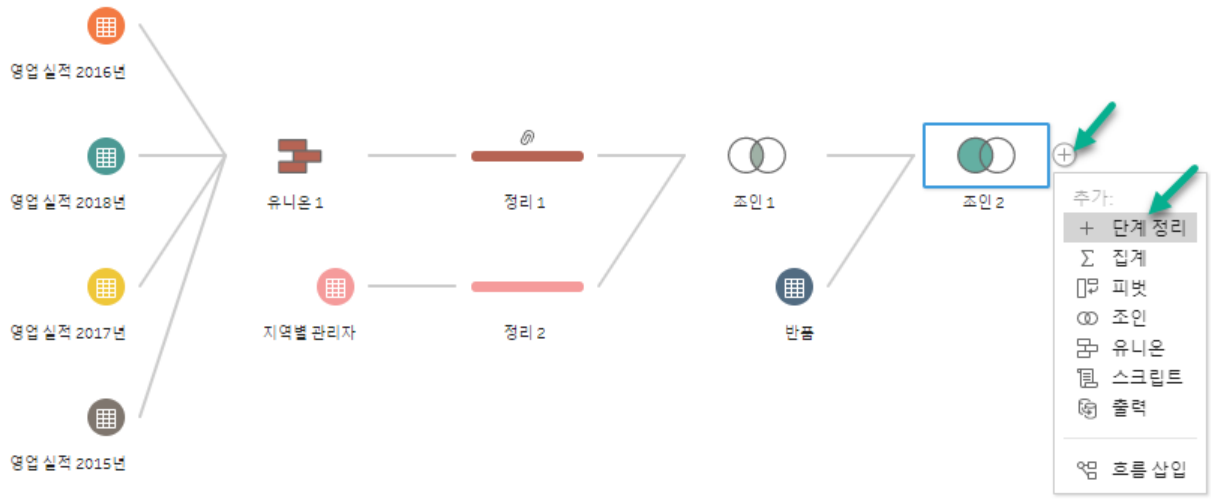


9. 조인 1의 데이터를 모두 가져오도록 조인 1의 비어 있는 집합 부분을 클릭해주세요.



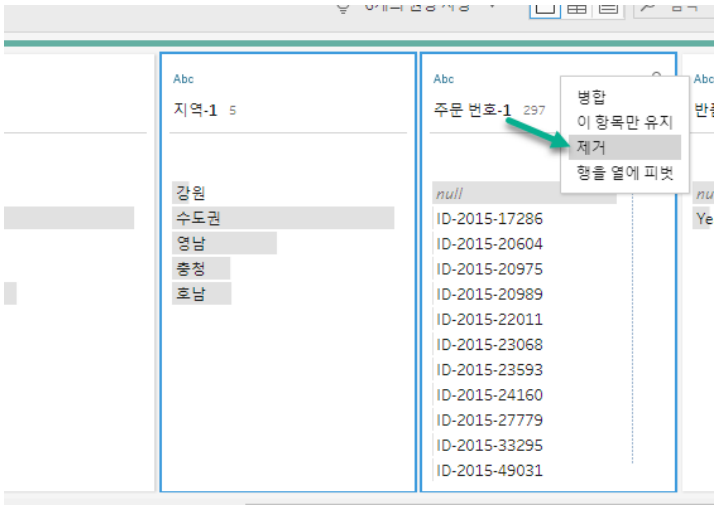
조인 유형이 Left로 변경되고,
조인 결과가 총 11,000건이 되는 것을 보실 수 있나요?

10. 결합된 데이터를 정리해 보아요. 아래 그림처럼 단계 정리를 넣어주세요.



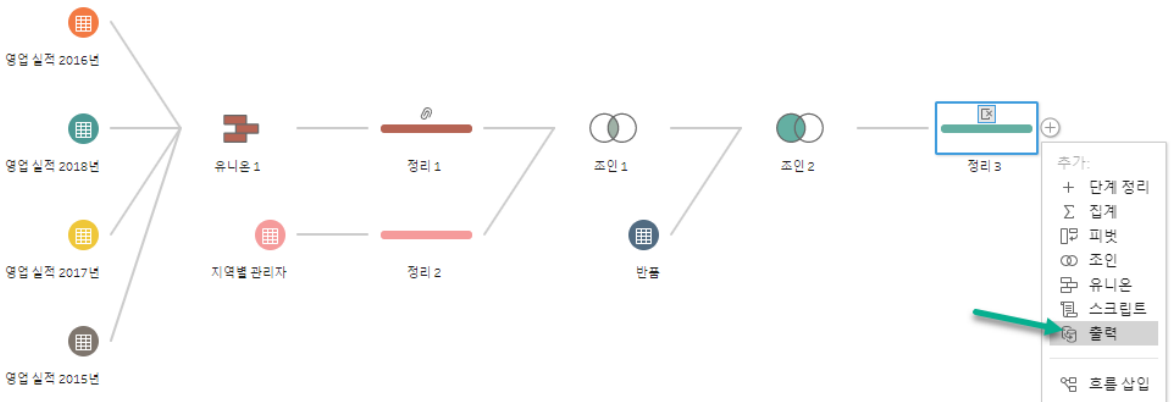
11. 결합하면서 추가적으로 생긴 필드를 제거해줍니다.

Table Name 클릭하고, Ctrl 키 누른 상태로 **지역-1**, **주문번호-1**을 클릭해주세요. → 마우스 오른쪽 버튼을 클릭하고 **제거** 눌러주세요.



12. 분석할 수 있도록 데이터로 출력해 보도록 합니다.

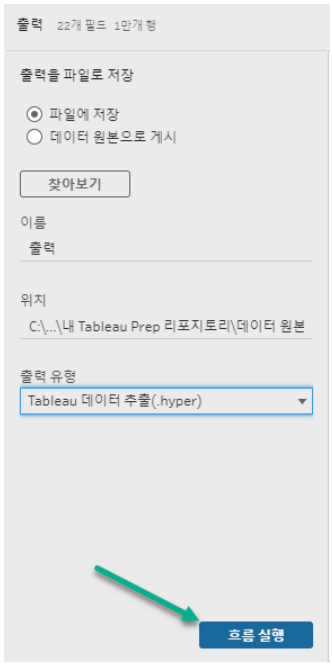
정리가 끝낸 데이터에서 **출력**을 눌러주세요.

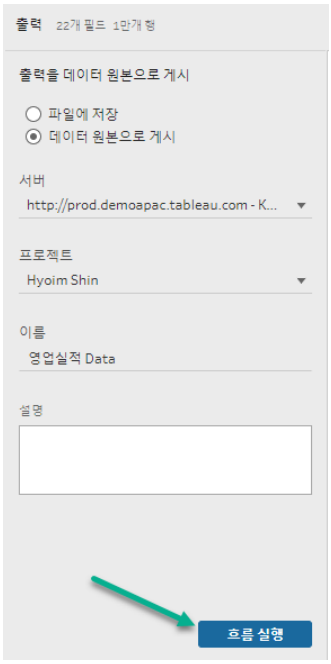


13. 출력의 유형은 크게 2가지가 있습니다.

1) 파일에 저장 : 자신의 PC에 hyper 혹은 csv 파일로 저장
2) 데이터 원본으로 게시 : 서버에 데이터 원본으로 게시

원하는 형태로 데이터를 출력해 보세요. 출력 옵션을 지정한 후 “흐름 실행”을 눌러주셔야 합니다.





출력된 Data를 이용해서 분석을 시작해보세요.
마지막 완료화면을 캡쳐해서 10일차 과제 제출하시면 됩니다.

2주간 고생 많으셨습니다!!! 추가 질문이 있으시면 언제든 슬랙 채널을 이용해주세요.