



# 고객을 세그먼테이션하자 [프로젝트]

## 11-2. 데이터 불러오기

### 데이터 살펴보기

- 테이블에 있는 10개의 행만 출력하기

```
SELECT *  
FROM valued-door-456102-k8.modulabs_project.data  
LIMIT 10;
```

[결과 이미지를 넣어주세요]

일	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHTS	6	2010-12-01 08:26:00 UTC	2.55	17850	United Kingdom
2	536365	71893	WHITE METAL LANTERN	6	2010-12-01 08:26:00 UTC	3.39	17850	United Kingdom
3	536365	844068	CREAM CUPID HEARTS COAT	8	2010-12-01 08:26:00 UTC	2.75	17850	United Kingdom
4	536365	842290	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00 UTC	3.39	17850	United Kingdom
5	536365	842296	RED WOOLLY HOTTIE WHITE HEARTS	6	2010-12-01 08:26:00 UTC	3.39	17850	United Kingdom
6	536365	227762	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00 UTC	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHTS	6	2010-12-01 08:26:00 UTC	4.25	17850	United Kingdom
8	536365	22633	HAND WARMER RED POUCEAU	6	2010-12-01 08:26:00 UTC	1.85	17850	United Kingdom
9	536365	22632	HAND WARMER RED POUCEAU	6	2010-12-01 08:26:00 UTC	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:26:00 UTC	1.69	13047	United Kingdom

- 전체 데이터는 몇 행으로 구성되어 있는지 확인하기

```
SELECT COUNT(*)  
FROM valued-door-456102-k8.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

일	COUNT(*)
1	541100

## 데이터 수 세기

- COUNT 함수를 사용해서, 각 컬럼별 데이터 포인트의 수를 세어 보기

```
SELECT COUNT(InvoiceNo) AS COUNT_InvoiceNo,  
COUNT(StockCode) AS COUNT_StockCode,  
COUNT(Description) AS COUNT_Description,  
COUNT(Quantity) AS COUNT_Quantity,  
COUNT(InvoiceDate) AS COUNT_InvoiceDate,  
COUNT(UnitPrice) AS COUNT_UnitPrice,  
COUNT(CustomerID) AS COUNT_CustomerID,  
COUNT(Country) AS COUNT_Country  
  
FROM valued-door-456102-k8.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

```
1 SELECT COUNT(InvoiceNo) AS COUNT_InvoiceNo,
2 COUNT(StockCode) AS COUNT_StockCode,
3 COUNT(Description) AS COUNT_Description,
4 COUNT(Quantity) AS COUNT_Quantity,
5 COUNT(InvoiceDate) AS COUNT_InvoiceDate,
6 COUNT(UnitPrice) AS COUNT_UnitPrice,
7 COUNT(CustomerID) AS COUNT_CustomerID,
8 COUNT(Country) AS COUNT_Country
9
10 FROM valued-door-456102-k8.modulabs_project.data;
```

쿼리 결과

	COUNT_InvoiceNo	COUNT_StockCode	COUNT_Description	COUNT_Quantity	COUNT_InvoiceDate	COUNT_UnitPrice	COUNT_CustomerID	COUNT_Country
1	541909	541909	540455	541909	541909	541909	406829	541909

## 11-4. 데이터 전처리 방법(1): 결측치 제거

### 컬럼 별 누락된 값의 비율 계산

- 각 컬럼 별 누락된 값의 비율을 계산
  - 각 컬럼에 대해서 누락 값을 계산한 후, 계산된 누락 값을 UNION ALL을 통해 합치기

```
SELECT
  'InvoiceNo' AS InvoiceNo,
  ROUND(SUM(CASE WHEN InvoiceNo IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
  'StockCode' AS StockCode,
  ROUND(SUM(CASE WHEN StockCode IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
  'Description' AS Description,
  ROUND(SUM(CASE WHEN Description IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
  'Quantity' AS Quantity,
  ROUND(SUM(CASE WHEN Quantity IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
  'InvoiceDate' AS InvoiceDate,
  ROUND(SUM(CASE WHEN InvoiceDate IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
  'UnitPrice' AS UnitPrice,
  ROUND(SUM(CASE WHEN UnitPrice IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
  'CustomerID' AS CustomerID,
  ROUND(SUM(CASE WHEN CustomerID IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
  'Country' AS Country,
  ROUND(SUM(CASE WHEN Country IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,

FROM valued-door-456102-k8.modulabs_project.data
```

[결과 이미지를 넣어주세요]

modulabs\_project\_04\_누락된...

```
1 SELECT
2   'InvoiceNo' AS InvoiceNo,
3   ROUND(SUM(CASE WHEN InvoiceNo IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
4   'StockCode' AS StockCode,
5   ROUND(SUM(CASE WHEN StockCode IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
6   'Description' AS Description,
7   ROUND(SUM(CASE WHEN Description IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
8   'Quantity' AS Quantity,
9   ROUND(SUM(CASE WHEN Quantity IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
10  'InvoiceDate' AS InvoiceDate,
11  ROUND(SUM(CASE WHEN InvoiceDate IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
12  'UnitPrice' AS UnitPrice,
13  ROUND(SUM(CASE WHEN UnitPrice IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
14  'CustomerID' AS CustomerID,
15  ROUND(SUM(CASE WHEN CustomerID IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
16  'Country' AS Country,
17  ROUND(SUM(CASE WHEN Country IS NULL THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS missing_percentage,
18
19 FROM valued-door-456102-k8.modulabs_project.data
```

쿼리 결과

영	InvoiceNo	missi	StockCode	missi	Description	missing_p	Quantity	missi	InvoiceDate	missi	UnitPrice	missi	CustomerID	missing	Country	missi
1	Invoice...	0.0	StockCode	0.0	Descript...	0.27	Quantity	0.0	InvoiceDate	0.0	UnitPrice	0.0	CustomerID	24.93	Country	0.0

### 결측치 처리 전략

- StockCode = '85123A' 의 Description 을 추출하는 쿼리문을 작성하기

```
SELECT Description
FROM valued-door-456102-k8.modulabs_project.data
```

```
WHERE StockCode IN ('85123A')
GROUP BY Description;
```

[결과 이미지를 넣어주세요]

modulabs\_project\_05\_Descrip... 실행 쿼리

```
1 SELECT Description
2 FROM valued-door-456102-k8.modulabs_project.data
3 WHERE StockCode IN ('85123A')
4 GROUP BY Description;
```

쿼리 결과

작업 정보	결과	차트	JSON	실행 세부정보	실행 그
행	Description				
1	WHITE HANGING HEART T-LIG...				
2	?				
3	wrongly marked carton 22804				
4	CREAM HANGING HEART T-LIG...				

## 결측치 처리

- DELETE 구문을 사용하며, WHERE 절을 통해 데이터를 제거할 조건을 제시

```
DELETE
FROM valued-door-456102-k8.modulabs_project.data
WHERE Description IS NULL OR CustomerID IS NULL;
```

[결과 이미지를 넣어주세요]

modulabs\_project\_06\_결측치... 실행 쿼리

```
1 DELETE
2 FROM valued-door-456102-k8.modulabs_project.data
3 WHERE Description IS NULL OR CustomerID IS NULL;
```

쿼리 결과

작업 정보	결과	실행 세부정보	실행 그래프
<p><b>i</b> 이 문으로 data의 행 135,080개가 삭제되었습니다.</p>			

## 11-5. 데이터 전처리(2): 중복값 처리

### 중복값 확인

- 중복된 행의 수를 세어보기
  - 8개의 컬럼에 그룹 함수를 적용한 후, COUNT가 1보다 큰 데이터를 세어보기

```
SELECT COUNT(*) AS Data_Duplicate
FROM (
  SELECT COUNT(*)
  FROM valued-door-456102-k8.modulabs_project.data
  GROUP BY InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country
  HAVING COUNT(*) > 1
) AS duplicates;
```

[결과 이미지를 넣어주세요]

행	Data_Duplicate
1	4837

## 중복값 처리

- 중복값을 제거하는 쿼리문 작성하기
  - CREATE OR REPLACE TABLE 구문을 활용하여 모든 컬럼(\*)을 DISTINCT 한 데이터로 업데이트

```
CREATE OR REPLACE TABLE valued-door-456102-k8.modulabs_project.data
AS SELECT DISTINCT * FROM valued-door-456102-k8.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

이 문으로 이름이 data인 테이블이 교체되었습니다.

## 11-6. 데이터 전처리(3): 오류값 처리

## InvoiceNo 살펴보기

- 고유(unique)한 InvoiceNo 의 개수를 출력하기

```
SELECT COUNT(DISTINCT InvoiceNo) AS unique_InvoiceNo
FROM valued-door-456102-k8.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

The screenshot shows a SQL query execution interface. The query is: `SELECT COUNT(DISTINCT InvoiceNo) AS unique_InvoiceNo FROM valued-door-456102-k8.modulabs_project.data;`. The result is displayed in a table with one row and one column, showing the value 22190.

unique_InvoiceNo
22190

- 고유한 InvoiceNo 를 앞에서부터 100개를 출력하기

```
SELECT DISTINCT InvoiceNo AS unique_InvoiceNo
FROM valued-door-456102-k8.modulabs_project.data
LIMIT 100;
```

[결과 이미지를 넣어주세요]

The screenshot shows a SQL query execution interface. The query is: `SELECT DISTINCT InvoiceNo AS unique_InvoiceNo FROM valued-door-456102-k8.modulabs_project.data LIMIT 100;`. The result is displayed in a table with 100 rows and one column, showing the first 100 unique InvoiceNo values.

unique_InvoiceNo
541431
C541433
537626
542237
549222
556201
562032
573511
581180
539318
541998
548955
568172
577609
543037
544156
545323

- InvoiceNo 가 'C'로 시작하는 행을 필터링 할 수 있는 쿼리문을 작성하기 (100행까지만 출력)

```
SELECT *
FROM valued-door-456102-k8.modulabs_project.data
```

```
WHERE InvoiceNo LIKE 'C%'
LIMIT 100;
```

[결과 이미지를 넣어주세요]

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
C541433	23166	MEDIUM CERAMIC TOP STOR...	-74215	2011-01-18 10:17:00 UTC	1.04	12346	United Kingdom
C545329	M	Manual	-1	2011-03-01 15:47:00 UTC	183.75	12352	Norway
C545329	M	Manual	-1	2011-03-01 15:47:00 UTC	280.05	12352	Norway
C545330	M	Manual	-1	2011-03-01 15:49:00 UTC	376.5	12352	Norway
C547388	22784	LANTERN CREAM GAZEBO	-3	2011-03-22 16:07:00 UTC	4.95	12352	Norway
C547388	22645	CERAMIC HEART FAIRY CAKE ...	-12	2011-03-22 16:07:00 UTC	1.45	12352	Norway
C547388	37448	CERAMIC CAKE DESIGN SPOT...	-12	2011-03-22 16:07:00 UTC	1.49	12352	Norway
C547388	21914	BLUE HARMONICA IN BOX	-12	2011-03-22 16:07:00 UTC	1.25	12352	Norway
C547388	22413	METAL SIGN TAKE IT OR LEAV...	-6	2011-03-22 16:07:00 UTC	2.95	12352	Norway
C547388	84050	PINK HEART SHAPE EGG FRYI...	-12	2011-03-22 16:07:00 UTC	1.65	12352	Norway
C547388	22701	PINK DOG BOWL	-6	2011-03-22 16:07:00 UTC	2.95	12352	Norway
C549905	22666	RECIPS BOX PANTRY YELLOW...	-2	2011-04-13 13:36:00 UTC	2.95	12359	Cyprus
C549905	22639	3 TIER CAKE TIN GREEN AND...	-2	2011-04-13 13:36:00 UTC	14.95	12359	Cyprus
C580165	22720	SET OF 3 CAKE TINS PANTRY...	-1	2011-12-02 11:21:00 UTC	4.95	12359	Cyprus
C580165	22345	SET OF 3 REGENCY CAKE TINS	-2	2011-12-02 11:21:00 UTC	4.95	12359	Cyprus
C580165	22797	CHEST OF DRAWERS GINGHA...	-2	2011-12-02 11:21:00 UTC	16.95	12359	Cyprus
C580165	22826	1 CUP FAT ANTIQUIF WHITE M	-1	2011-12-02 11:21:00 UTC	42.5	12359	Cyprus

- 구매 건 상태가 **Canceled** 인 데이터의 비율(%) - 소수점 첫번째 자리까지

```
SELECT ROUND(SUM(CASE WHEN InvoiceNo LIKE 'C%' THEN 1 ELSE 0 END) / COUNT(*) * 100, 1) AS InvoiceCanceled
FROM valued-door-456102-k8.modulabs_project.data
```

[결과 이미지를 넣어주세요]

InvoiceCanceled
2.2

## StockCode 살펴보기

- 고유한 **StockCode** 의 개수를 출력하기

```
SELECT COUNT(DISTINCT StockCode) AS unique_StockCode
FROM valued-door-456102-k8.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

unique_StockCode
3684

- 어떤 제품이 가장 많이 판매되었는지 보기 위하여 **StockCode** 별 등장 빈도를 출력하기

- 상위 10개의 제품들을 출력하기

```
SELECT StockCode, COUNT(*) AS sell_cnt
FROM valued-door-456102-k8.modulabs_project.data
GROUP BY 1
ORDER BY sell_cnt DESC
LIMIT 10;
```

[결과 이미지를 넣어주세요]

제목 없는 쿼리		실행	저장	다운로드	공유
1	SELECT StockCode, COUNT(*) AS sell_cnt				
2	FROM valued-door-456102-k8.modulabs_project.data				
3	GROUP BY 1				
4	ORDER BY sell_cnt DESC				
5	LIMIT 10;				
6					
7	SELECT *				
8	FROM valued-door-456102-k8.modulabs_project.data;				
9					

  

← 쿼리 결과	
작업 정보	결과
차트	JSON
실행 세부정보	실행 그래프
행	StockCode
1	85123A
2	22423
3	85099B
4	47566
5	84879
6	20725
7	22720
8	POST
9	22197
10	23203

- StockCode** 의 컬럼에 있던 값 중에서 숫자를 제외한 문자만 남기고 문자가 몇 자리 수 인지 세고

- 숫자가 0~1개인 값들에는 어떤 코드들이 들어가 있는지 출력하기

```
SELECT DISTINCT StockCode, number_count
FROM (
  SELECT StockCode,
    LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
  FROM valued-door-456102-k8.modulabs_project.data
  WHERE number_count IN (0,1)
```

[결과 이미지를 넣어주세요]

제목 없는 쿼리 실행 저장 다운로드 공유 일정

```

1
2 SELECT DISTINCT StockCode, number_count
3 FROM (
4     SELECT StockCode,
5           LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
6     FROM valued-door-456102-k8.modulabs_project.data)
7 WHERE number_count IN (0,1)
8

```

쿼리 결과

작업 정보 결과 차트 JSON 실행 세부정보 실행 그래프

행	StockCode	number_count
1	POST	0
2	M	0
3	C2	1
4	D	0
5	BANK CHARGES	0
6	PADS	0
7	DOT	0
8	CRUK	0

- **StockCode**의 컬럼에 있던 값 중에서 숫자를 제외한 문자만 남기고 문자가 몇 자리 수 인지 세고
  - 숫자가 0~1개인 값들을 가지고 있는 데이터 수는 전체 데이터 수 대비 몇 퍼센트인지 구하기 (소수점 두 번째 자리까지)

```

SELECT ROUND(SUM(CASE WHEN number_count IN (0,1) THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS stockcode_outlier_pct
FROM(
  SELECT StockCode,
    LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
  FROM valued-door-456102-k8.modulabs_project.data);

```

[결과 이미지를 넣어주세요]

modulabs\_project\_16\_StockCo... 실행 쿼리 저장 다운로드 공유 일정 다음에서 열

```

1
2 SELECT ROUND(SUM(CASE WHEN number_count IN (0,1) THEN 1 ELSE 0 END) / COUNT(*) * 100, 2) AS stockcode_outlier_pct
3 FROM(
4     SELECT StockCode,
5           LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
6     FROM valued-door-456102-k8.modulabs_project.data)
7 ;

```

쿼리 결과

작업 정보 결과 차트 JSON 실행 세부정보 실행 그래프

행	stockcode_outlier_pct
1	0.48

- 제품과 관련되지 않은 거래 기록을 제거하기

```

DELETE
FROM valued-door-456102-k8.modulabs_project.data
WHERE StockCode IN (
  SELECT DISTINCT StockCode
  FROM (
    SELECT StockCode,

```



```

        LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
    FROM valued-door-456102-k8.modulabs_project.data
)
WHERE number_count IN (0, 1)
);

```

[결과 이미지를 넣어주세요]

The screenshot shows a SQL query execution interface. The query is a DELETE statement that removes rows from the table 'valued-door-456102-k8.modulabs\_project.data' where the 'number\_count' is 0 or 1. The query is executed successfully, and the result shows that 1,915 rows were deleted.

```

1 DELETE
2 FROM valued-door-456102-k8.modulabs_project.data
3 WHERE StockCode IN (
4     SELECT DISTINCT StockCode
5     FROM (
6         SELECT StockCode,
7              LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
8         FROM valued-door-456102-k8.modulabs_project.data
9     )
10    WHERE number_count IN (0, 1)
11 );
12

```

쿼리 결과

작업 정보   **결과**   실행 세부정보   실행 그래프

**i** 이 문으로 data의 행 1,915개가 삭제되었습니다.

## Description 살펴보기

- 고유한 Description 별 출현 빈도를 계산하고 상위 30개를 출력하기

```

SELECT Description, COUNT(*) AS description_cnt
FROM valued-door-456102-k8.modulabs_project.data
GROUP BY Description
LIMIT 30;

```

[결과 이미지를 넣어주세요]

modulabs\_project\_18\_고유한D...

```

1 SELECT Description, COUNT(*) AS description_cnt
2 FROM valued-door-456102-k8.modulabs_project.data
3 GROUP BY Description
4 LIMIT 30;
5

```

쿼리 결과

작업 정보	결과	차트	JSON	실행 세부정보	실행 그래프
Description	description_cnt				
1	MEDIUM CERAMIC TOP STORAGE JAR	208			
2	RED TOADSTOOL LED NIGHT LIGHT	539			
3	SET/3 DECOUPAGE STACKING TINS	54			
4	FOUR HOOK WHITE LOVEBIRDS	265			
5	ALARM CLOCK BAKELIKE CHOCOLATE	339			
6	BLACK EAR MUFF HEADPHONES	18			
7	BLUE 3 PIECE POLKADOT CUTLERY SET	97			
8	EMERGENCY FIRST AID TIN	126			
9	COLOUR GLASS. STAR T-LIGHT HOLDER	247			
10	RED 3 PIECE RETROSPOT CUTLERY SET	100			
11	SET OF 2 TINS VINTAGE BATHROOM	59			
12	LARGE HEART MEASURING SPOONS	226			
13	RED DRAWER KNOB ACRYLIC EDWARDIAN	101			
14	BATHROOM METAL SIGN	60			
15	CLEAR DRAWER KNOB ACRYLIC EDWARDIAN	331			
16	BOOM BOX SPEAKER BOYS	56			
17	CAMOUFLAGE EAR MUFF HEADPHONES	17			

- 서비스 관련 정보를 포함하는 행들을 제거하기

```

DELETE
FROM valued-door-456102-k8.modulabs_project.data
WHERE Description LIKE 'Next Day Carriage' OR Description LIKE 'High Resolution Image';

```

[결과 이미지를 넣어주세요]

modulabs\_project\_20\_Descrip...

```

1 DELETE
2 FROM valued-door-456102-k8.modulabs_project.data
3 WHERE Description LIKE 'Next Day Carriage' OR Description LIKE 'High Resolution Image';

```

쿼리 결과

작업 정보	결과	실행 세부정보	실행 그래프
이 문으로 data의 행 83개가 삭제되었습니다.			

- 대소문자를 혼합하고 있는 데이터를 대문자로 표준화 하기

```

CREATE OR REPLACE TABLE valued-door-456102-k8.modulabs_project.data AS
SELECT
* EXCEPT (Description),

```

```
UPPER(Description) AS Description
FROM valued-door-456102-k8.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

제목 없는 쿼리

```
1 CREATE OR REPLACE TABLE valued-door-456102-k8.modulabs_project.data AS
2 SELECT
3   * EXCEPT (Description),
4   | UPPER(Description) AS Description
5 FROM valued-door-456102-k8.modulabs_project.data;
```

쿼리 결과

작업 정보 결과 실행 세부정보 실행 그래프

이 문으로 이름이 data인 테이블이 교체되었습니다.

## UnitPrice 살펴보기

- UnitPrice의 최솟값, 최댓값, 평균을 구하기

```
SELECT MIN(UnitPrice) AS min_price, MAX(UnitPrice) AS max_price, AVG(UnitPrice) AS avg_price
FROM valued-door-456102-k8.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

제목 없는 쿼리

```
1 SELECT MIN(UnitPrice) AS min_price, MAX(UnitPrice) AS max_price, AVG(UnitPrice) AS avg_price
2 FROM valued-door-456102-k8.modulabs_project.data;
3
```

쿼리 결과

작업 정보 결과 차트 JSON 실행 세부정보 실행 그래프

행	min_price	max_price	avg_price
1	0.0	649.5	2.904956757406...

- 단가가 0원인 거래의 개수, 구매 수량(Quantity)의 최솟값, 최댓값, 평균 구하기

```
SELECT COUNT(Quantity) AS cnt_quantity, MIN(Quantity) AS min_quantity, MAX(Quantity) AS max_quantity, AVG(Quantity) AS
FROM valued-door-456102-k8.modulabs_project.data
WHERE UnitPrice = 0
```

[결과 이미지를 넣어주세요]

modulabs\_project\_23\_단가가 ...

```

1 SELECT COUNT(Quantity) AS cnt_quantity, MIN(Quantity) AS min_quantity, MAX(Quantity) AS max_quantity, AVG(Quantity) AS avg_quantity
2 FROM valued-door-456102-k8.modulabs_project.data
3 WHERE UnitPrice = 0
4

```

쿼리 결과

작업 정보	결과	차트	JSON	실행 세부정보	실행 그래프
행	cnt_quantity	min_quantity	max_quantity	avg_quantity	
1	33	1	12540	420.5151515151...	

- UnitPrice = 0 를 제거하고 일관된 데이터셋을 유지하기

```

CREATE OR REPLACE TABLE valued-door-456102-k8.modulabs_project.data AS
SELECT *
FROM valued-door-456102-k8.modulabs_project.data
WHERE UnitPrice <> 0;

```

[결과 이미지를 넣어주세요]

제목 없는 쿼리

```

1 CREATE OR REPLACE TABLE valued-door-456102-k8.modulabs_project.data AS
2 SELECT *
3 FROM valued-door-456102-k8.modulabs_project.data
4 WHERE UnitPrice <> 0;

```

쿼리 결과

작업 정보	결과	실행 세부정보	실행 그래프
이 문으로 이름이 data인 테이블이 교체되었습니다.			

## 11-7. RFM 스코어

### Recency

- InvoiceDate 컬럼을 연월일 자료형으로 변경하기

```

SELECT DATE(InvoiceDate) AS InvoiceDay, *
FROM valued-door-456102-k8.modulabs_project.data;

```

[결과 이미지를 넣어주세요]

제목 없는 쿼리 실행 저장 다운로드 공유 일정 다음에서 열기 더보기 실행 시 이 쿼리가 34.7MB를 처리합니다

```

1 SELECT DATE(InvoiceDate) AS InvoiceDay,
2 FROM valued-door-456102-k8.modulabs_project.data;

```

접근성 옵션을 보려면 Alt+F1 키를 누르세요

쿼리 결과 결과 저장 다음에서 열기

작업 정보 결과 차트 JSON 실행 세부정보 실행 그래프

행	InvoiceDay	InvoiceNo	StockCode	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	2011-01-18	541431	23166	74215	2011-01-18 10:01:00 UTC	1.04	12346	United King
2	2011-01-18	C541433	23166	-74215	2011-01-18 10:17:00 UTC	1.04	12346	United King
3	2010-12-07	537626	22727	4	2010-12-07 14:57:00 UTC	3.75	12347	Iceland
4	2010-12-07	537626	84969	6	2010-12-07 14:57:00 UTC	4.25	12347	Iceland
5	2010-12-07	537626	22492	36	2010-12-07 14:57:00 UTC	0.65	12347	Iceland
6	2010-12-07	537626	22212	6	2010-12-07 14:57:00 UTC	2.1	12347	Iceland
7	2010-12-07	537626	22773	12	2010-12-07 14:57:00 UTC	1.25	12347	Iceland
8	2010-12-07	537626	22494	12	2010-12-07 14:57:00 UTC	1.25	12347	Iceland
9	2010-12-07	537626	22805	12	2010-12-07 14:57:00 UTC	1.25	12347	Iceland
10	2010-12-07	537626	22497	4	2010-12-07 14:57:00 UTC	4.25	12347	Iceland
11	2010-12-07	537626	84997D	6	2010-12-07 14:57:00 UTC	3.75	12347	Iceland
12	2010-12-07	537626	22772	12	2010-12-07 14:57:00 UTC	1.25	12347	Iceland
13	2010-12-07	537626	22375	4	2010-12-07 14:57:00 UTC	4.25	12347	Iceland
14	2010-12-07	537626	85167B	30	2010-12-07 14:57:00 UTC	1.25	12347	Iceland
15	2010-12-07	537626	22726	4	2010-12-07 14:57:00 UTC	3.75	12347	Iceland

- 가장 최근 구매 일자를 MAX() 함수로 찾아보기

```

SELECT
  DATE(InvoiceDate) AS InvoiceDay,
  MAX(InvoiceDate) OVER() AS most_recent_date
FROM valued-door-456102-k8.modulabs_project.data;

```

[결과 이미지를 넣어주세요]

제목 없는 쿼리 실행 저장 다운로드 공유 일정 다음에서 열기 더보기 쿼리 완료됨

```

1 SELECT
2   DATE(InvoiceDate) AS InvoiceDay,
3   MAX(InvoiceDate) OVER() AS most_recent_date,
4   *
5 FROM valued-door-456102-k8.modulabs_project.data;

```

접근성 옵션을 보려면 Alt+F1 키를 누르세요

쿼리 결과 결과 저장 다음에서 열기

작업 정보 결과 차트 JSON 실행 세부정보 실행 그래프

행	InvoiceDay	most_recent_date	InvoiceNo	StockCode	Quantity	InvoiceDate	UnitPrice
1	2011-04-08	2011-12-09 12:50:00 UTC	549435	21843	4	2011-04-08 12:33:00 UTC	10.95
2	2011-10-05	2011-12-09 12:50:00 UTC	569650	22730	50	2011-10-05 12:44:00 UTC	3.39
3	2011-10-14	2011-12-09 12:50:00 UTC	571255	23064	10	2011-10-14 17:13:00 UTC	41.75
4	2010-12-10	2011-12-09 12:50:00 UTC	538174	21578	24	2010-12-10 09:35:00 UTC	2.25
5	2011-10-25	2011-12-09 12:50:00 UTC	572559	22781	2	2011-10-25 08:44:00 UTC	7.65
6	2011-08-31	2011-12-09 12:50:00 UTC	564856	35599B	12	2011-08-31 09:11:00 UTC	7.25
7	2011-01-07	2011-12-09 12:50:00 UTC	540438	22509	1	2011-01-07 12:28:00 UTC	16.95
8	2011-10-11	2011-12-09 12:50:00 UTC	570672	23374	10	2011-10-11 14:52:00 UTC	0.82
9	2011-10-11	2011-12-09 12:50:00 UTC	570672	21577	6	2011-10-11 14:52:00 UTC	2.25
10	2011-11-04	2011-12-09 12:50:00 UTC	574501	21034	2	2011-11-04 13:15:00 UTC	0.95

- 유저 별로 가장 큰 InvoiceDay를 찾아서 가장 최근 구매일로 저장하기

```

SELECT
  CustomerID,
  MAX(DATE(InvoiceDate)) AS InvoiceDay
FROM valued-door-456102-k8.modulabs_project.data
GROUP BY CustomerID;

```

[결과 이미지를 넣어주세요]

```

1  -- 유저 별로 가장 큰 InvoiceDay를 찾아서 가장 최근 구매일로 저장하기
2
3  SELECT
4      CustomerID,
5      MAX(InvoiceDate) AS InvoiceDay
6  FROM valued-door-456102-k8.modulabs_project.data
7  GROUP BY CustomerID;

```

### 쿼리 결과

작업 정보	결과	차트	JSON	실행 세부정보	실행
행	CustomerID	InvoiceDay			
1	12346	2011-01-18			
2	12347	2011-12-07			
3	12348	2011-09-25			
4	12349	2011-11-21			
5	12350	2011-02-02			
6	12352	2011-11-03			
7	12353	2011-05-19			
8	12354	2011-04-21			
9	12355	2011-05-09			
10	12356	2011-11-17			
11	12357	2011-11-06			
12	12358	2011-12-08			
13	12359	2011-12-02			
14	12360	2011-10-18			
15	12361	2011-02-25			
16	12362	2011-12-06			

- 가장 최근 일자( **most\_recent\_date** )와 유저별 마지막 구매일( **InvoiceDay** )간의 차이를 계산하기

```

SELECT
    CustomerID,
    EXTRACT(DAY FROM MAX(InvoiceDay) OVER () - InvoiceDay) AS recency
FROM (
    SELECT
        CustomerID,
        MAX(InvoiceDate) AS InvoiceDay
    FROM project_name.modulabs_project.data
    GROUP BY CustomerID
);

```

[결과 이미지를 넣어주세요]

제목 없는 쿼리

실행 저장 다운로드 공유

```

1
2 --가장 최근 일자(most_recent_date)와 유저별 마지막 구매일(InvoiceDay) 간의 차이를 계산
3
4 SELECT
5     CustomerID,
6     EXTRACT(DAY FROM MAX(InvoiceDay) OVER () - InvoiceDay) AS recency
7 FROM (
8     SELECT
9         CustomerID,
10        MAX(DATE(InvoiceDate)) AS InvoiceDay
11 FROM valued-door-456102-k8.modulabs_project.data
12 GROUP BY CustomerID

```

### 쿼리 결과

작업 정보	결과	차트	JSON	실행 세부정보	실행 그래프
행	CustomerID	recency			
1	12432	42			
2	12506	232			
3	12895	42			
4	12965	89			
5	13269	1			
6	13271	37			
7	13493	275			
8	13518	85			
9	13658	9			
10	13824	32			
11	13859	326			
12	13901	72			
13	14092	7			

- 최종 데이터 셋에 필요한 데이터들을 각각 정제해서 이어붙이고 지금까지의 결과를 `user_r` 이라는 이름의 테이블로 저장하기

```

CREATE OR REPLACE TABLE valued-door-456102-k8.modulabs_project.user_r AS
SELECT
    CustomerID,
    EXTRACT(DAY FROM MAX(InvoiceDay) OVER () - InvoiceDay) AS recency
FROM (
    SELECT
        CustomerID,
        MAX(DATE(InvoiceDate)) AS InvoiceDay
    FROM valued-door-456102-k8.modulabs_project.data
    GROUP BY CustomerID
);

```

[결과 이미지를 넣어주세요]

제목 없는 쿼리 실행 저장 다운로드 공유

```

1  -- 지금까지의 결과를 user_r이라는 이름의 테이블로 저장
2
3  CREATE OR REPLACE TABLE valued-door-456102-k8.modulabs_project.user_r AS
4  SELECT
5      CustomerID,
6      EXTRACT(DAY FROM MAX(InvoiceDay) OVER () - InvoiceDay) AS recency
7  FROM (
8      SELECT
9          CustomerID,
10         MAX(InvoiceDate) AS InvoiceDay
11     FROM valued-door-456102-k8.modulabs_project.data
12     GROUP BY CustomerID
13 );

```

쿼리 결과

작업 정보 결과 실행 세부정보 실행 그래프

**i** 이 문으로 이름이 user\_r인 새 테이블이 생성되었습니다.

## Frequency

- 고객마다 고유한 InvoiceNo의 수를 세어보기

```

SELECT
    CustomerID,
    # [[YOUR QUERY]] AS purchase_cnt
FROM project_name.modulabs_project.data
# [[YOUR QUERY]];

```

[결과 이미지를 넣어주세요]

- 각 고객 별로 구매한 아이템의 총 수량 더하기

```

SELECT
    CustomerID,
    # [[YOUR QUERY]] AS item_cnt
FROM project_name.modulabs_project.data
# [[YOUR QUERY]];

```

[결과 이미지를 넣어주세요]

- 전체 거래 건수 계산과 구매한 아이템의 총 수량 계산의 결과를 합쳐서 **user\_rf** 라는 이름의 테이블에 저장하기

```

CREATE OR REPLACE TABLE project_name.modulabs_project.user_rf AS

-- (1) 전체 거래 건수 계산
WITH purchase_cnt AS (
    # [[YOUR QUERY]]
),

-- (2) 구매한 아이템 총 수량 계산
item_cnt AS (
    # [[YOUR QUERY]]
)

-- 기존의 user_r에 (1)과 (2)를 통합
SELECT
    pc.CustomerID,

```



```

pc.purchase_cnt,
ic.item_cnt,
ur.recency
FROM purchase_cnt AS pc
JOIN item_cnt AS ic
  ON pc.CustomerID = ic.CustomerID
JOIN project_name.modulabs_project.user_r AS ur
  ON pc.CustomerID = ur.CustomerID;

```

[결과 이미지를 넣어주세요]

## Monetary

- 고객별 총 지출액 계산 (소수점 첫째 자리에서 반올림)

```

SELECT
  CustomerID,
  # [[YOUR QUERY]] AS user_total
FROM project_name.modulabs_project.data
# [[YOUR QUERY]];

```

[결과 이미지를 넣어주세요]

- 고객별 평균 거래 금액 계산

- 고객별 평균 거래 금액을 구하기 위해 1) `data` 테이블을 `user_rf` 테이블과 조인(LEFT JOIN) 한 후, 2) `purchase_cnt` 로 나누어서 3) `user_rfm` 테이블로 저장하기

```

CREATE OR REPLACE TABLE project_name.modulabs_project.user_rfm AS
SELECT
  rf.CustomerID AS CustomerID,
  rf.purchase_cnt,
  rf.item_cnt,
  rf.recency,
  ut.user_total,
  # [[YOUR QUERY]] AS user_average
FROM project_name.modulabs_project.user_rf rf
LEFT JOIN (
  -- 고객 별 총 지출액
  SELECT
    # [[YOUR QUERY]]
  ) ut
  ON rf.CustomerID = ut.CustomerID;

```

[결과 이미지를 넣어주세요]

## RFM 통합 테이블 출력하기

- 최종 `user_rfm` 테이블을 출력하기

```

# [[YOUR QUERY]];

```

[결과 이미지를 넣어주세요]

## 11-8. 추가 Feature 추출

### 1. 구매하는 제품의 다양성

- 1) 고객 별로 구매한 상품들의 고유한 수를 계산하기
- 2) `user_rfm` 테이블과 결과를 합치기
- 3) `user_data` 라는 이름의 테이블에 저장하기

```
CREATE OR REPLACE TABLE project_name.modulabs_project.user_data AS
WITH unique_products AS (
  SELECT
    CustomerID,
    COUNT(DISTINCT StockCode) AS unique_products
  FROM project_name.modulabs_project.data
  GROUP BY CustomerID
)
SELECT ur.*, up.* EXCEPT (CustomerID)
FROM project_name.modulabs_project.user_rfm AS ur
JOIN unique_products AS up
ON ur.CustomerID = up.CustomerID;
```

[결과 이미지를 넣어주세요]

### 2. 평균 구매 주기

- 고객들의 쇼핑 패턴을 이해하는 것을 목표 (고객 별 재방문 주기 살펴보기)
  - 균 구매 소요 일수를 계산하고, 그 결과를 `user_data` 에 통합

```
CREATE OR REPLACE TABLE project_name.modulabs_project.user_data AS
WITH purchase_intervals AS (
  -- (2) 고객 별 구매와 구매 사이의 평균 소요 일수
  SELECT
    CustomerID,
    CASE WHEN ROUND(AVG(interval_), 2) IS NULL THEN 0 ELSE ROUND(AVG(interval_), 2) END AS average_interval
  FROM (
    -- (1) 구매와 구매 사이에 소요된 일수
    SELECT
      CustomerID,
      DATE_DIFF(InvoiceDate, LAG(InvoiceDate) OVER (PARTITION BY CustomerID ORDER BY InvoiceDate), DAY) AS interval_
    FROM
      project_name.modulabs_project.data
    WHERE CustomerID IS NOT NULL
  )
  GROUP BY CustomerID
)
SELECT u.*, pi.* EXCEPT (CustomerID)
FROM project_name.modulabs_project.user_data AS u
LEFT JOIN purchase_intervals AS pi
ON u.CustomerID = pi.CustomerID;
```

[결과 이미지를 넣어주세요]

### 3. 구매 취소 경향성

- 고객의 취소 패턴 파악하기

- 1) 취소 빈도(cancel\_frequency) : 고객 별로 취소한 거래의 총 횟수
- 2) 취소 비율(cancel\_rate) : 각 고객이 한 모든 거래 중에서 취소를 한 거래의 비율
  - 취소 빈도와 취소 비율을 계산하고 그 결과를 `user_data` 에 통합하기  
(취소 비율은 소수점 두번째 자리)

```
CREATE OR REPLACE TABLE project_name.modulabs_project.user_data AS

WITH TransactionInfo AS (
  SELECT
    CustomerID,
    # [[YOUR QUERY]] AS total_transactions,
    # [[YOUR QUERY]] AS cancel_frequency
  FROM project_name.modulabs_project.data
  # [[YOUR QUERY]]
)

SELECT u.*, t.* EXCEPT(CustomerID), # [[YOUR QUERY]] AS cancel_rate
FROM `project_name.modulabs_project.user_data` AS u
LEFT JOIN TransactionInfo AS t
ON # [[YOUR QUERY]];
```

[결과 이미지를 넣어주세요]

- 다양한 컬럼들을 활용하여 고객의 구매 패턴과 선호도를 보다 심층적으로 이해할 수 있도록 최종적으로 `user_data` 를 출력하기

```
# [[YOUR QUERY]];
```

[결과 이미지를 넣어주세요]

## 회고

[회고 내용을 작성해주세요]

Keep :

Problem :

Try :