

고객을 세그멘테이션하자! [프로젝트] - 이수아

11-2. 데이터 불러오기

데이터 살펴보기

- 테이블에 있는 10개의 행만 출력하기

```
# [[YOUR QUERY]]
select *
from `amazing-smile-470106-q5`.modulabs_project.data
limit 10
```

[결과 이미지를 넣어주세요]

작업 정보	결과	시각화	JSON	실행 세부정보	실행 그래프			
#	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIG...	6	2010-12-01 08:26:00 UTC	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00 UTC	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT H...	8	2010-12-01 08:26:00 UTC	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WA...	6	2010-12-01 08:26:00 UTC	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE H...	6	2010-12-01 08:26:00 UTC	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BO...	2	2010-12-01 08:26:00 UTC	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT...	6	2010-12-01 08:26:00 UTC	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00 UTC	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00 UTC	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORN...	32	2010-12-01 08:34:00 UTC	1.69	13047	United Kingdom

- 전체 데이터는 몇 행으로 구성되어 있는지 확인하기

```
select *
from `amazing-smile-470106-q5`.modulabs_project.data
```

[결과 이미지를 넣어주세요]

작업 정보결과시각화JSON실행 세부정보실행 그래프

#	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
5	536365	84029E	RED WOOLLY HOTTIE WHITE H...	6	2010-12-01 08:26:00 UTC	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BO...	2	2010-12-01 08:26:00 UTC	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT...	6	2010-12-01 08:26:00 UTC	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00 UTC	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00 UTC	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORN...	32	2010-12-01 08:34:00 UTC	1.69	13047	United Kingdom
11	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-01 08:34:00 UTC	2.1	13047	United Kingdom
12	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-01 08:34:00 UTC	2.1	13047	United Kingdom
13	536367	22749	FELTCRAFT PRINCESS CHARLO...	8	2010-12-01 08:34:00 UTC	3.75	13047	United Kingdom
14	536367	22310	IVORY KNITTED MUG COSY	6	2010-12-01 08:34:00 UTC	1.65	13047	United Kingdom
15	536367	84969	BOX OF 6 ASSORTED COLOUR T...	6	2010-12-01 08:34:00 UTC	4.25	13047	United Kingdom
16	536367	22623	BOX OF VINTAGE JIGSAW BLO...	3	2010-12-01 08:34:00 UTC	4.95	13047	United Kingdom
17	536367	22622	BOX OF VINTAGE ALPHABET B...	2	2010-12-01 08:34:00 UTC	9.95	13047	United Kingdom
18	536367	21754	HOME BUILDING BLOCK WORD	3	2010-12-01 08:34:00 UTC	5.95	13047	United Kingdom
19	536367	21755	LOVE BUILDING BLOCK WORD	3	2010-12-01 08:34:00 UTC	5.95	13047	United Kingdom
20	536367	21777	RECIPE BOX WITH METAL HEA...	4	2010-12-01 08:34:00 UTC	7.95	13047	United Kingdom
21	536367	48187	DOORMAT NEW ENGLAND	4	2010-12-01 08:34:00 UTC	7.95	13047	United Kingdom
22	536368	22960	JAM MAKING SET WITH JARS	6	2010-12-01 08:34:00 UTC	4.25	13047	United Kinodm

페이지당 결과 수: 501 ~ 50 (전체 541909행)

페이지당 결과 수: 50 1 - 50 (전체 541909행) |< > >>

데이터 수 세기

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

```
SELECT
COUNT(InvoiceNo) AS Count_InvoiceNo,
COUNT(StockCode) AS Count_tockCode,
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 `amazing-smile-470106-q5`.modulabs_project.data;
 [결과 이미지를 넣어주세요]

작업 정보 결과 시각화 JSON 실행 세부정보 실행 그래프

#	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 column_name,
  COUNT(CASE WHEN InvoiceNo IS NULL THEN 1 else 0 END) AS missing_count,
  round(COUNT(CASE WHEN InvoiceNo IS NULL THEN 1 else 0 END) * 100.0 / COUNT(*),2) AS missing_percentage
FROM
  `amazing-smile-470106-q5`.modulabs_project.data
UNION ALL
SELECT
  'StockCode' AS column_name,
  COUNT(CASE WHEN StockCode IS NULL THEN 1 END) AS missing_count,
  round(cOUNT(CASE WHEN StockCode IS NULL THEN 1 END) * 100.0 / COUNT(*),2) AS missing_percentage
FROM
  `amazing-smile-470106-q5`.modulabs_project.data
UNION ALL
SELECT
  'Description' AS column_name,
  COUNT(CASE WHEN Description IS NULL THEN 1 END) AS missing_count,
  round(COUNT(CASE WHEN Description IS NULL THEN 1 END) * 100.0 / COUNT(*),2)AS missing_percentage
FROM
  `amazing-smile-470106-q5`.modulabs_project.data
UNION ALL
SELECT
  'Quantity' AS column_name,
  COUNT(CASE WHEN Quantity IS NULL THEN 1 END) AS missing_count,
  round(COUNT(CASE WHEN Quantity IS NULL THEN 1 END) * 100.0 / COUNT(*),2) AS missing_percentage
FROM
  `amazing-smile-470106-q5`.modulabs_project.data
UNION ALL
SELECT
  'InvoiceDate' AS column_name,
  COUNT(CASE WHEN InvoiceDate IS NULL THEN 1 END) AS missing_count,
  round(COUNT(CASE WHEN InvoiceDate IS NULL THEN 1 END) * 100.0 / COUNT(*),2)AS missing_percentage
FROM
  `amazing-smile-470106-q5`.modulabs_project.data
UNION ALL
SELECT
  'UnitPrice' AS column_name,
  COUNT(CASE WHEN UnitPrice IS NULL THEN 1 END) AS missing_count,
  round(COUNT(CASE WHEN UnitPrice IS NULL THEN 1 END) * 100.0 / COUNT(*),2) AS missing_percentage
FROM
  `amazing-smile-470106-q5`.modulabs_project.data
UNION ALL
SELECT
  'CustomerID' AS column_name,
  COUNT(CASE WHEN CustomerID IS NULL THEN 1 END) AS missing_count,
  round(COUNT(CASE WHEN CustomerID IS NULL THEN 1 END) * 100.0 / COUNT(*),2) AS missing_percentage
FROM
  `amazing-smile-470106-q5`.modulabs_project.data

```

```

UNION ALL
SELECT
  'Country' AS column_name,
  COUNT(CASE WHEN Country IS NULL THEN 1 END) AS missing_count,
  round(COUNT(CASE WHEN Country IS NULL THEN 1 END) * 100.0 / COUNT(*),2) AS missing_percentage
FROM
  `amazing-smile-470106-q5`.modulabs_project.data;

```

[결과 이미지를 넣어주세요]

작업 정보	결과	시각화	JSON	실행 세부정보	실행 그래프
	column_name ▼	missing_count ▼	missing_percenta...		
1	CustomerID	135080	24.93		
2	Country	0	0.0		
3	Quantity	0	0.0		
4	Description	1454	0.27		
5	InvoiceDate	0	0.0		
6	UnitPrice	0	0.0		
7	InvoiceNo	541909	100.0		
8	StockCode	0	0.0		

결측치 처리 전략

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

```

select stockcode, description
from `amazing-smile-470106-q5`.modulabs_project.data
where stockcode = '85123A';

```

[결과 이미지를 넣어주세요]

번호	stockcode ▼	description ▼
1	85123A	WHITE HANGING HEART T-LIG...
2	85123A	WHITE HANGING HEART T-LIG...
3	85123A	WHITE HANGING HEART T-LIG...
4	85123A	WHITE HANGING HEART T-LIG...
5	85123A	WHITE HANGING HEART T-LIG...
6	85123A	WHITE HANGING HEART T-LIG...
7	85123A	WHITE HANGING HEART T-LIG...
8	85123A	WHITE HANGING HEART T-LIG...
9	85123A	WHITE HANGING HEART T-LIG...
10	85123A	WHITE HANGING HEART T-LIG...
11	85123A	WHITE HANGING HEART T-LIG...
12	85123A	WHITE HANGING HEART T-LIG...
13	85123A	WHITE HANGING HEART T-LIG...
14	85123A	WHITE HANGING HEART T-LIG...
15	85123A	WHITE HANGING HEART T-LIG...

결측치 처리

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

```
delete from `amazing-smile-470106-q5`.modulabs_project.data
where description is null or customerid is null ;
```

[결과 이미지를 넣어주세요]

쿼리 결과

작업 정보

결과

실행 세부정보

실행 그래프



이 문으로 data의 행 135,080개가 삭제되었습니다.

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 duplicate_rows_count
FROM (
  SELECT
    InvoiceNo,
    StockCode,
    Description,
    Quantity,
    InvoiceDate,
    UnitPrice,
    CustomerID,
    Country,
    COUNT(*) AS row_count
  FROM
    `amazing-smile-470106-q5`.modulabs_project.data
  GROUP BY
    InvoiceNo,
    StockCode,
    Description,
    Quantity,
    InvoiceDate,
    UnitPrice,
    CustomerID,
    Country
  HAVING
    row_count > 1);
```

[결과 이미지를 넣어주세요]

행	duplicate_rows_c...
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 `amazing-smile-470106-q5`.modulabs_project.data AS
SELECT DISTINCT
  InvoiceNo,
  StockCode,
  Description,
  Quantity,
  InvoiceDate,
  UnitPrice,
  CustomerID,
  Country
FROM
  `amazing-smile-470106-q5`.modulabs_project.data;

select count(*) as remaining_rows_count
from `amazing-smile-470106-q5`.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

작업 정보 **결과** 실행 세부정보 실행 그래프

i 이 문으로 이름이 data인 테이블이 교체되었습니다.

작업 정보 **결과** 人

행	remaining_rows_...
1	401604

11-6. 데이터 전처리(3): 오류값 처리

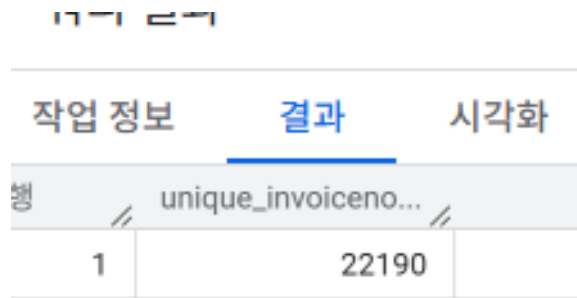
InvoiceNo 살펴보기

- 고유(unique)한 **InvoiceNo**의 개수를 출력하기

```
# [[YOUR QUERY]]
select count (distinct invoiceno) as unique_invoice_count
```

```
from `amazing-smile-470106-q5`.modulabs_project.data;
```

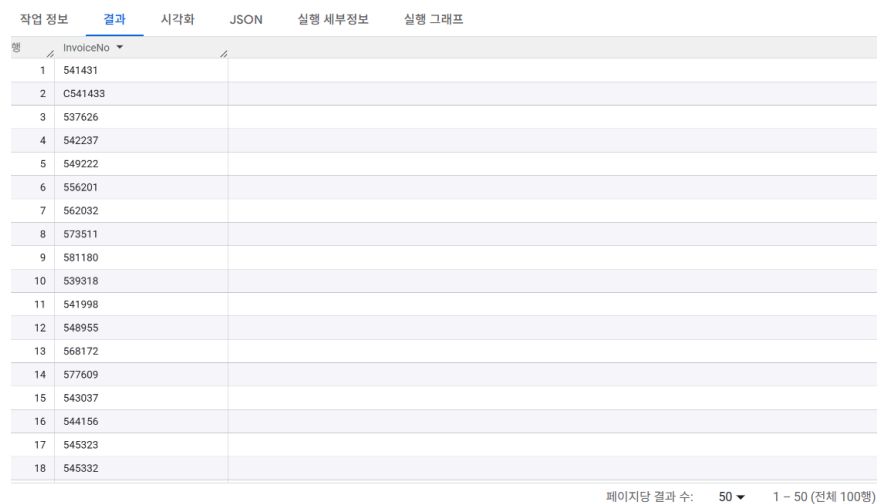
[결과 이미지를 넣어주세요]



행	unique_invoiceno...
1	22190

- 고유한 InvoiceNo 를 앞에서부터 100개를 출력하기

```
select distinct invocieno
from `amazing-smile-470106-q5`.modulabs_project.data
limit 100;
```



행	InvoiceNo
1	541431
2	C541433
3	537626
4	542237
5	549222
6	556201
7	562032
8	573511
9	581180
10	539318
11	541998
12	548955
13	568172
14	577609
15	543037
16	544156
17	545323
18	545332

페이지당 결과 수: 50 ▼ 1 - 50 (전체 100행)

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

```
select *
from
where invoiceno like 'c%'
limit 100;
```

[결과 이미지를 넣어주세요]

작업 정보		결과	시각화	JSON	실행 세부정보	실행 그래프			
행	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	
1	C541433	23166	MEDIUM CERAMIC TOP STORA...	-74215	2011-01-18 10:17:00 UTC	1.04	12346	United Kingdom	
2	C545329	M	Manual	-1	2011-03-01 15:47:00 UTC	280.05	12352	Norway	
3	C545329	M	Manual	-1	2011-03-01 15:47:00 UTC	183.75	12352	Norway	
4	C545330	M	Manual	-1	2011-03-01 15:49:00 UTC	376.5	12352	Norway	
5	C547388	37448	CERAMIC CAKE DESIGN SPOTT...	-12	2011-03-22 16:07:00 UTC	1.49	12352	Norway	
6	C547388	22701	PINK DOG BOWL	-6	2011-03-22 16:07:00 UTC	2.95	12352	Norway	
7	C547388	21914	BLUE HARMONICA IN BOX	-12	2011-03-22 16:07:00 UTC	1.25	12352	Norway	
8	C547388	22413	METAL SIGN TAKE IT OR LEAVE...	-6	2011-03-22 16:07:00 UTC	2.95	12352	Norway	
9	C547388	84050	PINK HEART SHAPE EGG FRYIN...	-12	2011-03-22 16:07:00 UTC	1.65	12352	Norway	
10	C547388	22645	CERAMIC HEART FAIRY CAKE ...	-12	2011-03-22 16:07:00 UTC	1.45	12352	Norway	
11	C547388	22784	LANTERN CREAM GAZEBO	-3	2011-03-22 16:07:00 UTC	4.95	12352	Norway	
12	C549955	22839	3 TIER CAKE TIN GREEN AND C...	-2	2011-04-13 13:38:00 UTC	14.95	12359	Cyprus	
13	C549955	22666	RECIPE BOX PANTRY YELLOW ...	-2	2011-04-13 13:38:00 UTC	2.95	12359	Cyprus	
14	C580165	23245	SET OF 3 REGENCY CAKE TINS	-2	2011-12-02 11:21:00 UTC	4.95	12359	Cyprus	
15	C580165	22826	LOVE SEAT ANTIQUE WHITE M...	-1	2011-12-02 11:21:00 UTC	42.5	12359	Cyprus	
16	C580165	22797	CHEST OF DRAWERS GINGHA...	-2	2011-12-02 11:21:00 UTC	16.95	12359	Cyprus	
17	C580165	22720	SET OF 3 CAKE TINS PANTRY D...	-1	2011-12-02 11:21:00 UTC	4.95	12359	Cyprus	
18	C544902	22273	FELTCRAFT DOLL MOLLY	-1	2011-02-24 13:05:00 UTC	2.95	12362	Belgium	
19	C544902	22629	SPACEBOY LUNCH BOX	-1	2011-02-24 13:05:00 UTC	1.95	12362	Belgium	
20	C563752	22659	LUNCH BOX I LOVE LONDON	-6	2011-08-19 10:38:00 UTC	1.95	12362	Belgium	
21	C563752	22891	TEA FOR ONE POLKADOT	-1	2011-08-19 10:38:00 UTC	4.25	12362	Belgium	

페이지당 결과 수: 50 1 - 50 (전체 100행) |< < > >

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

```
SELECT
ROUND(COUNT(CASE WHEN STARTS_WITH(InvoiceNo, 'C')
THEN InvoiceNo ELSE NULL END) * 100 / COUNT(InvoiceNo), 1)
AS `canceled_order_percentage`
FROM `amazing-smile-470106-q5`.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

작업 정보		결과	시각화
행		canceled_...	
1		2.2	

StockCode 살펴보기

- 고유한 **StockCode** 의 개수를 출력하기

```
select count (distinct stockcode) as unique_stockcode_count
from `amazing-smile-470106-q5`.modulabs_project.data;
```

[결과 이미지를 넣어주세요]

작업 정보		결과	시각화
행		unique_stockcod...	
1		3684	

- 어떤 제품이 가장 많이 판매되었는지 보기 위하여 **StockCode** 별 등장 빈도를 출력하기
 - 상위 10개의 제품들을 출력하기

```
SELECT StockCode, COUNT(*) AS sell_cnt
from `amazing-smile-470106-q5`.modulabs_project.data
group by stockcode
```

```
ORDER BY sell_cnt DESC
limit 10;
```

[결과 이미지를 넣어주세요]

	StockCode	sell_cnt
1	85123A	2065
2	22423	1894
3	85099B	1659
4	47566	1409
5	84879	1405
6	20725	1346
7	22720	1224
8	POST	1196
9	22197	1110
10	23203	1108

- **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

```
WITH UniqueStockCodes AS (
  SELECT DISTINCT
    StockCode
  FROM
    `amazing-smile-470106-q5`.modulabs_project.data
)
SELECT
  number_count,
  COUNT(*) AS stock_cnt
FROM (
  SELECT
    StockCode,
    LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
  FROM
    UniqueStockCodes
)
GROUP BY number_count
ORDER BY stock_cnt DESC;

SELECT DISTINCT StockCode, number_count
FROM (
  SELECT StockCode,
    LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
  FROM `amazing-smile-470106-q5`.modulabs_project.data
)
WHERE LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', ''))t in (0,1)
```

작업 정보	결과	시각화	JSON
행	number_count	stock_cnt	
1	5	3676	
2	0	7	
3	1	1	

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)

```
WITH StockCodeStats AS (
  SELECT
    StockCode,
    LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
  FROM `amazing-smile-470106-q5`.modulabs_project.data
),
Counts AS (
  SELECT
    COUNT(*) AS total_cnt,
    COUNTIF(number_count IN (0, 1)) AS filtered_cnt
  FROM StockCodeStats
)
SELECT
  ROUND(filtered_cnt / total_cnt * 100, 2) AS percentage
FROM Counts;
```

[결과 이미지를 넣어주세요]

작업 정보	결과	人
행	percentage	
1	0.48	

- 제품과 관련되지 않은 거래 기록을 제거하기

```
DDELETE FROM `amazing-smile-470106-q5`.modulabs_project.data
WHERE StockCode IN (
  SELECT
    StockCode
  FROM (
    SELECT DISTINCT
      StockCode,
      LENGTH(StockCode) - LENGTH(REGEXP_REPLACE(StockCode, r'[0-9]', '')) AS number_count
    FROM
      `amazing-smile-470106-q5`.modulabs_project.data
  )
  WHERE
    number_count IN (0, 1));
```

[결과 이미지를 넣어주세요]

i 이 문으로 data의 행 1,915개가 삭제되었습니다.

Description 살펴보기

- 고유한 Description 별 출현 빈도를 계산하고 상위 30개를 출력하기

```
SELECT DISTINCT Description, COUNT(*) AS description_cnt
FROM project_name.modulabs_project.data
WHERE
  Description IS NOT NULL
GROUP BY
  Description
ORDER BY
  description_cnt DESC
LIMIT 30;
```

[결과 이미지를 넣어주세요]

행	Description ▼	description_count ▼
1	WHITE HANGING HEART T-LIG...	2058
2	REGENCY CAKESTAND 3 TIER	1894
3	JUMBO BAG RED RETROSPOT	1659
4	PARTY BUNTING	1409
5	ASSORTED COLOUR BIRD ORN...	1405
6	LUNCH BAG RED RETROSPOT	1345
7	SET OF 3 CAKE TINS PANTRY D...	1224
8	LUNCH BAG BLACK SKULL.	1099
9	PACK OF 72 RETROSPOT CAKE ...	1062
10	SPOTTY BUNTING	1026
11	PAPER CHAIN KIT 50'S CHRIST...	1013
12	LUNCH BAG SPACEBOY DESIGN	1006
13	LUNCH BAG CARS BLUE	1000
14	HEART OF WICKER SMALL	990
15	NATURAL SLATE HEART CHAL...	989
16	JAM MAKING SET WITH JARS	966
17	LUNCH BAG PINK POLKADOT	961
18	LUNCH BAG SUKI DESIGN	932
19	ALARM CLOCK BAKELIKE RED	917

- 서비스 관련 정보를 포함하는 행들을 제거하기

```
SELECT *
FROM `amazing-smile-470106-q5`.modulabs_project.data
WHERE
  REGEXP_CONTAINS(Description, r'(?i)POSTAGE|CARRIAGE|DELIVERY|ADJUSTMENT|REFUND|
  TEST|SAMPLE|BANK CHARGES|FEE|LOST|DAMAGED|CRACKED|BROKEN|MISSING|MANUAL|DISCOUNT|
```

```

CANCELED|RETURNED')
OR Description IS NULL
OR Description = '';
# [[YOUR QUERY]]

```

[결과 이미지를 넣어주세요]

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
557789	22802	COFFEE MUG PEARS DESIGN	6	2011-06-22 16:32:00 UTC	2.55	12379	Belgium
557789	22801	COFFEE MUG CAT + BIRD DESIGN	6	2011-06-22 16:32:00 UTC	2.55	12379	Belgium
557789	22803	COFFEE MUG APPLES DESIGN	6	2011-06-22 16:32:00 UTC	2.55	12379	Belgium
563100	22843	SET OF TEA COFFEE SUGAR TINS PANTRY	4	2011-06-12 09:57:00 UTC	4.95	12381	Norway
563100	22922	BAKING MOULD TOFFEE CUP CHOCOLATE	6	2011-06-12 09:57:00 UTC	2.35	12381	Norway
563100	851198	WHITE TEA COFFEE SUGAR JARS	12	2011-06-12 09:57:00 UTC	1.95	12381	Norway
566050	22843	SET OF TEA COFFEE SUGAR TINS PANTRY	-2	2011-06-31 17:16:00 UTC	4.95	12381	Norway
566050	851198	WHITE TEA COFFEE SUGAR JARS	-1	2011-06-31 17:16:00 UTC	1.95	12381	Norway
566050	22843	SET OF TEA COFFEE SUGAR TINS PANTRY	4	2011-07-14 13:28:00 UTC	4.95	12398	Australia
561507	22843	SET OF TEA COFFEE SUGAR TINS PANTRY	4	2011-07-27 14:21:00 UTC	4.95	12397	Belgium
574536	22843	SET OF TEA COFFEE SUGAR TINS PANTRY	4	2011-11-04 09:52:00 UTC	4.95	12397	Belgium
572567	22843	SET OF TEA COFFEE SUGAR TINS PANTRY	4	2011-10-25 10:27:00 UTC	4.95	12398	Switzerland
559128	22843	SET OF TEA COFFEE SUGAR TINS PANTRY	4	2011-07-14 08:52:00 UTC	4.95	12405	Finland
559128	35810A	ENAMEL PINK COFFEE CONTAINER	12	2011-07-14 08:52:00 UTC	0.83	12405	Finland
557147	22803	COFFEE MUG APPLES DESIGN	6	2011-06-17 10:51:00 UTC	2.55	12408	Belgium
574602	22803	COFFEE MUG APPLES DESIGN	12	2011-11-07 12:27:00 UTC	2.55	12408	Belgium
543889	21216	SET 3 RETROSPOT TEA COFFEE SUGAR	72	2011-03-15 09:52:00 UTC	4.35	12415	Australia
545475	21216	SET 3 RETROSPOT TEA COFFEE SUGAR	120	2011-03-03 10:59:00 UTC	4.35	12415	Australia
601546	979791	SET OF 3 COFFEE MUGS ASSORTED	108	2011-04-17 16:25:00 UTC	1.95	12416	Australia

- 대소문자를 혼합하고 있는 데이터를 대문자로 표준화 하기

```

CREATE OR REPLACE TABLE project_name.modulabs_project.data AS
SELECT
  * EXCEPT (Description),
  upper(description) as description
FROM project_name.modulabs_project.data;

```

[결과 이미지를 넣어주세요]

작업 정보 **결과** 실행 세부정보 실행 그래프

i 이 문으로 이름이 data인 테이블이 교체되었습니다.

UnitPrice 살펴보기

- UnitPrice 의 최솟값, 최댓값, 평균을 구하기

```

select min(unitprice) as min_price, max(unitprice) as max_price, avg(unitprice) as avg_price
FROM `amazing-smile-470106-q5`.modulabs_project.data;

```

[결과 이미지를 넣어주세요]

작업 정보	결과	시각화	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 avg_quantity
FROM `amazing-smile-470106-q5`.modulabs_project.data
WHERE unitprice = 0

```

[결과 이미지를 넣어주세요]

작업 정보	결과	시각화	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 `amazing-smile-470106-q5`.modulabs_project.data AS
SELECT *
from `amazing-smile-470106-q5`.modulabs_project.data
where unitprice > 0;
```

[결과 이미지를 넣어주세요]

작업 정보 결과 실행 세부정보 실행 그래프

i 이 문으로 이름이 data인 테이블이 교체되었습니다.

11-7. RFM 스코어

Recency

- InvoiceDate 컬럼을 연월일 자료형으로 변경하기

```
select date(invoicedate) as invoiceday,
from `amazing-smile-470106-q5`.modulabs_project.data
limit 10;
```

[결과 이미지를 넣어주세요]

작업 정보 결과 시각화

	invoiceday	
1	2011-01-18	
2	2011-01-18	
3	2010-12-07	
4	2010-12-07	
5	2010-12-07	
6	2010-12-07	
7	2010-12-07	
8	2010-12-07	
9	2010-12-07	
10	2010-12-07	

- 가장 최근 구매 일자를 MAX() 함수로 찾아보기

```
select max(date(invoicedate)) over () as most_recent_date,
date(invoicedate) as invoiceday, *
from `amazing-smile-470106-q5`.modulabs_project.data
```

[결과 이미지를 넣어주세요]

	most_recent_date ▼	invoiceday ▼	InvoiceNo ▼	StockCode ▼	Quantity ▼
1	2011-12-09	2011-11-06	574740	23485	1
2	2011-12-09	2011-02-07	543370	47566	5
3	2011-12-09	2011-11-18	577228	21528	1
4	2011-12-09	2010-12-17	539447	21879	24
5	2011-12-09	2011-03-03	545475	21238	96
6	2011-12-09	2011-06-15	556917	23241	144
7	2011-12-09	2011-10-05	569650	22725	50
8	2011-12-09	2011-06-05	555574	23111	2
9	2011-12-09	2011-08-25	564378	23110	2
10	2011-12-09	2010-12-09	538003	22956	48
11	2011-12-09	2011-05-23	554126	23079	2
12	2011-12-09	2011-06-21	557621	23110	2

- 유저 별로 가장 큰 InvoiceDay를 찾아서 가장 최근 구매일로 저장하기

```
SELECT
  CustomerID,
  MAX(DATE(InvoiceDate)) AS InvoiceDay
FROM `amazing-smile-470106-q5`.modulabs_project.data
GROUP BY CustomerID;
```

[결과 이미지를 넣어주세요]

	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

- 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(DATE(InvoiceDate)) AS InvoiceDay
  FROM `amazing-smile-470106-q5`.modulabs_project.data
  GROUP BY CustomerID );
```

[결과 이미지를 넣어주세요]

작업 정보	결과	시각화	JSON	실시간
행	CustomerID	recency		
1	12375	2		
2	12415	24		
3	12778	19		
4	12808	36		
5	12870	366		
6	12908	176		
7	12923	64		
8	13090	8		
9	13092	70		
10	13094	21		
11	13210	93		
12	13226	273		
13	13313	22		
14	13449	23		
15	13466	100		

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

```
CREATE OR REPLACE TABLE `amazing-smile-470106-q5`.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 `amazing-smile-470106-q5`.modulabs_project.data
  GROUP BY CustomerID
);
```

[결과 이미지를 넣어주세요]

행	CustomerID	recency
1	18102	0
2	15311	0
3	12662	0
4	17754	0
5	16626	0
6	17315	0
7	14441	0
8	17389	0
9	16446	0
10	16558	0
11	16954	0
12	12433	0

Frequency

- 고객마다 고유한 InvoiceNo의 수를 세어보기

```
select
  CustomerID,
  count(distinct invoiceno) as cnt_purchase
from `amazing-smile-470106-q5`.modulabs_project.data
group by customerid;
```

[결과 이미지를 넣어주세요]

행	CustomerID	cnt_purchase
1	12346	2
2	12347	7
3	12348	4
4	12349	1
5	12350	1
6	12352	8
7	12353	1

- 각 고객 별로 구매한 아이템의 총 수량 더하기

```
select customerid,
  sum(quantity) as cnt_item
from `amazing-smile-470106-q5`.modulabs_project.data
group by customerid;
```

[결과 이미지를 넣어주세요]

행	customerid	cnt_item
1	12346	0
2	12347	2458
3	12348	2332
4	12349	630
5	12350	196
6	12352	463
7	12353	20

- 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
WITH purchase_cnt AS (
  SELECT
    CustomerID,
    COUNT(DISTINCT InvoiceNo) AS purchase_cnt
  FROM `amazing-smile-470106-q5`.modulabs_project.data
  GROUP BY CustomerID
),
item_cnt AS (
  SELECT
    CustomerID,
    SUM(Quantity) AS item_cnt
  FROM project_name.modulabs_project.data
  GROUP BY CustomerID
)
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]

행	CustomerID	purchase_cnt	item_cnt	recency
11	14578	1	240	3
12	12442	1	181	3
13	15318	1	642	3
14	16569	1	93	3
15	12478	1	233	3
16	12650	1	250	3
17	17914	1	457	3
18	14219	1	78	4
19	15097	1	170	4
20	17383	1	148	4
21	13790	1	748	4
22	12367	1	172	4
23	18015	1	157	4

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]

행	CustomerID	user_total
1	12346	0.0
2	12347	4310.0
3	12348	1437.0
4	12349	1458.0
5	12350	294.0
6	12352	1265.0
7	12353	89.0
8	12354	1079.0
9	12355	459.0
10	12356	2487.0

• 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 `amazing-smile-470106-q5`.modulabs_project.user_rfm AS
SELECT
  rf.CustomerID AS CustomerID,
  rf.purchase_cnt,
  rf.item_cnt,
  rf.recency,
  ut.user_total,
  ROUND(ut.user_total / rf.purchase_cnt) AS user_average
FROM `amazing-smile-470106-q5`.modulabs_project.user_rf rf
LEFT JOIN (
  SELECT
    CustomerID,
    ROUND(SUM(UnitPrice * Quantity)) AS user_total
  FROM `amazing-smile-470106-q5`.modulabs_project.data
  GROUP BY CustomerID
) ut
ON rf.CustomerID = ut.CustomerID;
```

[결과 이미지를 넣어주세요]

작업 정보 **결과** 실행 세부정보 실행 그래프

i 이 문으로 이름이 user_rfm인 테이블이 교체되었습니다.

RFM 통합 테이블 출력하기

• 최종 user_rfm 테이블을 출력하기

```
select *
from `amazing-smile-470106-q5`.modulabs_project.user_rfm
```

[결과 이미지를 넣어주세요]

작업 정보	결과	시각화	JSON	실행 세부정보	실행 그래프
40	CustomerID 12966	purchase_cnt 1	item_cnt 15	recency 9	user_total 160.0
41	17911	1	223	9	366.0
42	18058	1	88	9	170.0
43	14601	1	168	10	214.0
44	13428	1	151	10	202.0
용량 관리	14349	1	86	10	134.0
46	15783	1	212	10	246.0
47	15619	1	136	10	336.0
48	13349	1	224	10	197.0
49	15148	1	187	10	301.0
50	15790	1	113	10	219.0

페이지당 결과 수: 50 1 - 50 (전체 4362행) |< < > >|

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 `amazing-smile-470106-q5`.modulabs_project.user_data AS
WITH
  unique_products AS (
    SELECT
      CustomerID,
      COUNT(DISTINCT StockCode) AS unique_products
    FROM
      `amazing-smile-470106-q5`.modulabs_project.data
    GROUP BY
      CustomerID
  )
SELECT
  ur.*,up.*EXCEPT(CustomerID)
FROM
  `amazing-smile-470106-q5`.modulabs_project.user_rfm AS ur
JOIN
  unique_products AS up
ON
  ur.CustomerID = up.CustomerID;
```

[결과 이미지를 넣어주세요]

행	CustomerID	purchase_cnt	item_cnt	recency	user_total	user_average	unique_products
1	14705	1	100	198	179.0	179.0	1
2	17923	1	50	282	208.0	208.0	1
3	15070	1	36	372	106.0	106.0	1
4	17307	1	-144	365	-153.0	-153.0	1
5	17752	1	192	359	81.0	81.0	1
6	13391	1	4	203	60.0	60.0	1
7	13829	1	-12	359	-102.0	-102.0	1
8	15488	1	72	92	76.0	76.0	1
9	12791	1	96	373	178.0	178.0	1
10	14119	1	-2	354	-20.0	-20.0	1
11	15562	1	39	351	135.0	135.0	1
12	12943	1	-1	301	-4.0	-4.0	1
13	12603	1	56	21	613.0	613.0	1
14	15195	1	1404	2	3861.0	3861.0	1
15	16093	1	20	106	17.0	17.0	1
16	17956	1	1	249	13.0	13.0	1
17	18233	1	4	325	440.0	440.0	1
18	14576	1	12	372	35.0	35.0	1
19	16765	1	4	294	34.0	34.0	1

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 `amazing-smile-470106-q5`.modulabs_project.user_data AS
WITH purchase_intervals AS (
  SELECT
    CustomerID,
    CASE WHEN ROUND(AVG(interval_), 2) IS NULL THEN 0 ELSE ROUND(AVG(interval_), 2)
    END AS average_interval
  FROM (
    SELECT
      CustomerID,
      DATE_DIFF(InvoiceDate, LAG(InvoiceDate) OVER (PARTITION BY CustomerID ORDER BY
      InvoiceDate), DAY) AS interval_
    FROM
      `amazing-smile-470106-q5`.modulabs_project.data
    WHERE CustomerID IS NOT NULL
  )
  GROUP BY CustomerID
)

SELECT u.*, pi.* EXCEPT (CustomerID)
FROM `amazing-smile-470106-q5`.modulabs_project.user_data AS u
LEFT JOIN purchase_intervals AS pi
ON u.CustomerID = pi.CustomerID;
```

[결과 이미지를 넣어주세요]

행	CustomerID	purchase_cnt	item_cnt	recency	user_total	user_average	unique_products	average_interval
1	13391	1	4	203	60.0	60.0	1	0.0
2	15195	1	1404	2	3861.0	3861.0	1	0.0
3	16257	1	1	176	22.0	22.0	1	0.0
4	13135	1	4300	196	3096.0	3096.0	1	0.0
5	12943	1	-1	301	-4.0	-4.0	1	0.0
6	17331	1	16	123	175.0	175.0	1	0.0
7	15510	1	2	330	250.0	250.0	1	0.0
8	17347	1	216	86	229.0	229.0	1	0.0
9	17307	1	-144	365	-153.0	-153.0	1	0.0
10	15118	1	1440	134	245.0	245.0	1	0.0
11	15657	1	24	22	30.0	30.0	1	0.0

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 `amazing-smile-470106-q5`.modulabs_project.user_data AS
WITH
TransactionInfo AS (
  SELECT
    CustomerID,
    COUNT(DISTINCT InvoiceNo) AS total_transactions,
    COUNT(DISTINCT
      CASE
        WHEN InvoiceNo LIKE 'C%' THEN InvoiceNo
        ELSE NULL
      END) AS cancel_frequency
  FROM
    `amazing-smile-470106-q5`.modulabs_project.data
  WHERE
    CustomerID IS NOT NULL
  GROUP BY CustomerID
)
SELECT
  u.*,
  t.total_transactions,
  t.cancel_frequency,
  ROUND(cancel_frequency / total_transactions, 2) AS cancel_rate
FROM
  `amazing-smile-470106-q5`.modulabs_project.user_data AS u
LEFT JOIN
  TransactionInfo AS t
ON u.CustomerID = t.CustomerID;
```

[결과 이미지를 넣어주세요]

쿼리 결과

작업 정보 **결과** 실행 세부정보 실행 그래프

i 이 문으로 이름이 user_data인 테이블이 교체되었습니다.

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

```
select*
from `amazing-smile-470106-q5`.modulabs_project.user_data
```

[결과 이미지를 넣어주세요]

명	CustomerID	purchase_cnt	item_cnt	recency	user_total	user_average	unique_products	average_interval	total_transactions	cancel_frequency	cancel_rate
1	19118	1	1440	134	245.0	245.0	1	0.0	1	0	0.0
2	19070	1	36	372	106.0	106.0	1	0.0	1	0	0.0
3	15374	1	126	128	168.0	168.0	3	0.0	1	0	0.0
4	14437	1	60	365	63.0	63.0	5	0.0	1	0	0.0
5	14861	1	117	52	126.0	126.0	6	0.0	1	0	0.0
6	15458	1	156	25	162.0	162.0	7	0.0	1	0	0.0
7	13876	1	36	63	123.0	123.0	7	0.0	1	0	0.0
8	15843	1	27	303	119.0	119.0	8	0.0	1	0	0.0
9	16586	1	95	248	464.0	464.0	8	0.0	1	0	0.0
10	16349	1	9	290	54.0	54.0	9	0.0	1	0	0.0
11	16587	1	71	232	225.0	225.0	13	0.0	1	0	0.0
12	17142	1	81	239	579.0	579.0	15	0.0	1	0	0.0

회고

[회고 내용을 작성해주세요]

Keep : 최선을 다하자

Problem : 분석적인 사고 부족

Try : 창의적으로 생각하자