캐글 커머스 데이터 - EDA

목차

원본 데이터 정보

- 캐글 데이터 | Superstore Sales Profit Discount Predict
    - 슈퍼스토어 판매 데이터 : 가구, 전자제품, 사무용품을 주로 파는 미국 내 여러 지점이 있는 마트의 판매 데이터

- 데이터 상세 정보

컬럼명	데이터 상세	데이터 구성
Row ID	고유 행 번호	1~9994
Order ID	고유 주문번호	총 5,009건
Order Date	주문 일자	2014-01-03~2017-12-30
Ship Date	발송 일자	2014-01-07~2018-01-05
Ship Mode	배송 모드	총 3개 카테고리; Standard Class, Second Class, Other
Customer ID	고유 고객번호	총 793개
Customer Name	고객이름	총 793개
Segment	구매분류	Consumer, Corporate, Home Office
Country	국가	
City	도시	
State	주	
Postal Code	우편번호	
Region	지역	West, East, Central, South
Product ID	제품 번호	총 1,862개
Category	제품 카테고리	Office Supplies, Furniture, Other
Sub-Category	제품 세부 카테고리	
Product Name	제품명	
Sales	판매금액(매출)	
Quantity	주문 수량	
Discount	할인율	
Profit	순익	

EDA에 활용한 데이터

- 원본 데이터에서 300개 행이 제외된 데이터를 활용 → ‘Superstore.csv’

- 원본 데이터는 9994(컬럼행 미포함)개의 행으로 구성된 데이터였지만, mysql workbench에 데이터를 import 했을 때, 다수의 팀원들에게 공통적으로 300개의 행이 유실되어 9,694(컬럼행 미포함)개의 행의 데이터를 활용하기로 함
- 2014년 1월 4일 ~ 2017년 12월 30일 데이터로 구성되어 있음

SQL 활용 EDA

(1) 고객 주문 특성 파악

(1-1) 4년 동안 고객 및 주문수 (2014/01/04 ~2017/12/30)

- 고유 고객수 : 793명
- 고유 주문수 : 4,931건
- SQL 코드

```
-- 4년 동안 고객수 & 주문수 구하기
select count(distinct `customer name`), count(distinct `Order ID`)
  from superstore_sales_240830;
```

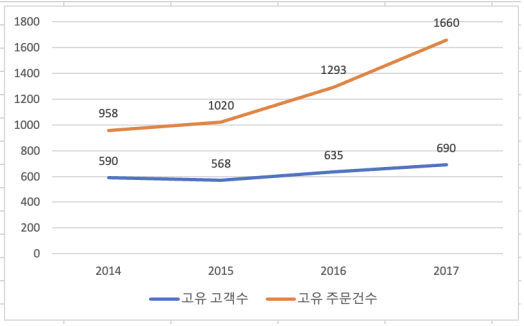
(1-2) 연도/분기 별 고유 고객수 및 고유 주문건수

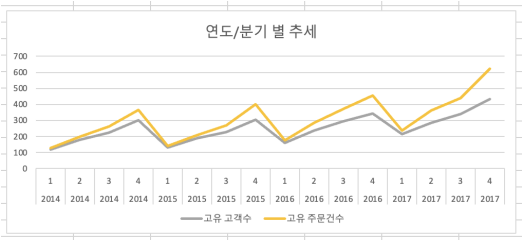
- SQL 코드

```
-- 연도 별 고객수 & 주문수 구하기
select year(order_date_new), count(distinct `customer name`), count(distinct `Order I
  from superstore_sales_240830
  group by year(order_date_new);

-- 연도/분기 별 고객수 & 주문수 구하기
select year(order_date_new) as year, quarter(order_date_new) as quat, count(distinct
  from superstore_sales_240830
  group by year, quat;
```

(1-3) 연도 별 성장률

- **고유 주문건수는 2014년 대비 2017년에는 약 73.7% 성장**
  - 2014년 대비 2015년: 약 6.48% 성장
  - 2015년 대비 2016년: 약 26.88% 성장
  - 2016년 대비 2017년: 약 28.35% 성장
- **고유 고객수는 2014년 대비 2017년에는 약 16.95% 성장**
  - 2014년 대비 2015년: 약 -3.73% 감소
  - 2015년 대비 2016년: 약 11.80% 증가
  - 2016년 대비 2017년: 약 8.66% 증가
- 연도 별 주문건수는 시간이 흐름에 따라 늘어났지만 고유고객수는 많이 늘지 않음
  - **연도 별 고유 고객수 및 고유 주문건수 추세**
  - **연도/분기 당 추세**





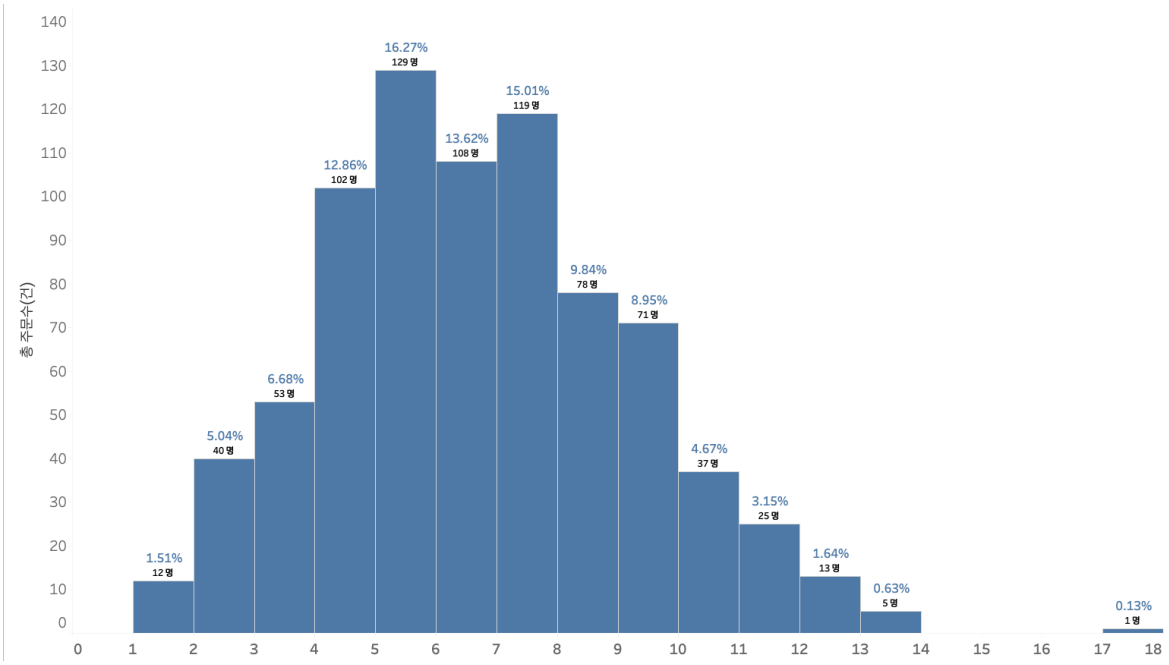
(1-4) 고객 당 구매건수

- SQL 코드

```
with order_num as (
select `customer id`, count(1) as order_num
  from `sample - superstore_240830`
  group by `customer id`
  order by 2 desc),

  first_order as(
  select `customer id`, min(str_to_date(`Order Date`, '%m/%d/%Y')) as first_order_dat
    from `sample - superstore_240830`
    group by `customer id`),

first_order_details as (select a.*, year(a.first_order_date), month(a.first_order_dat
from(select o.*, f.first_order_date from order_num o
  left join first_order f on o.`customer id` = f.`customer id`) a),

total_order_num as(select order_num, count(order_num) as cnt
  from first_order_details
  group by order_num
),

total_count AS (
    SELECT SUM(cnt) AS total_cnt
    FROM total_order_num
)

SELECT t.order_num,
       t.cnt,
       t.cnt / total.total_cnt * 100 AS percentage
FROM total_order_num t
CROSS JOIN total_count total
order by 3 desc;
```

- **구매건수 별 고객수 히스토그램(14~17년 전체)**
  - SQL 코드

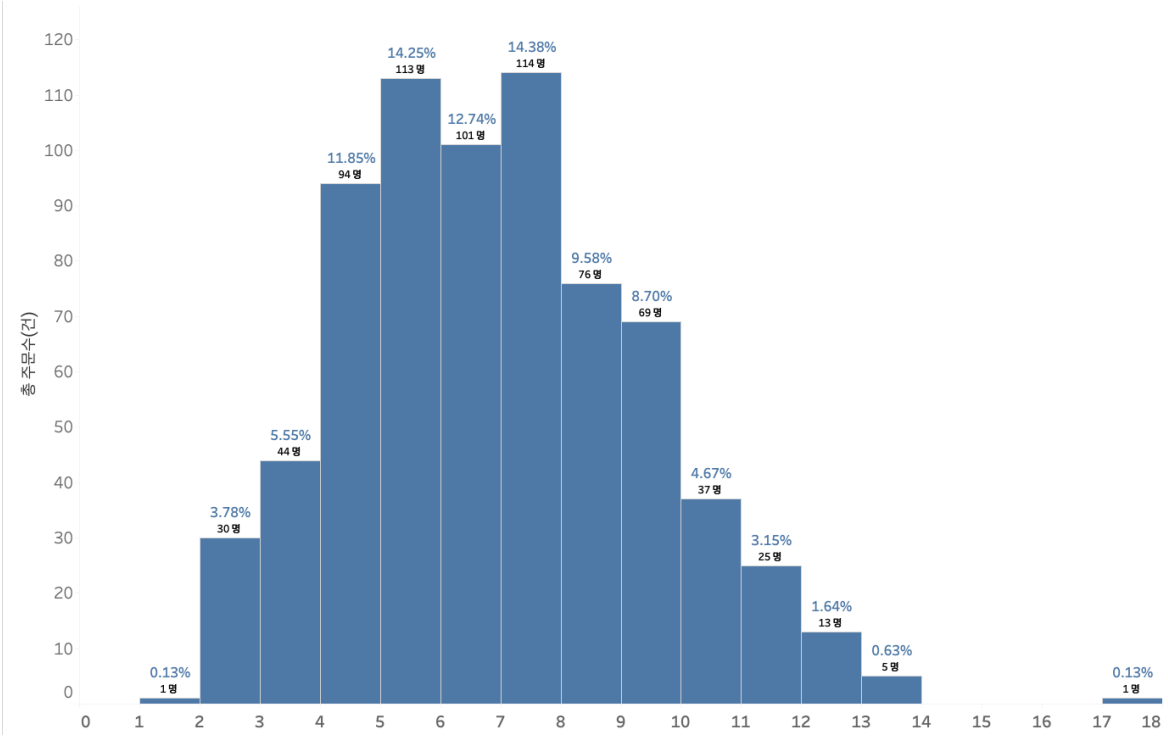
```
-- 구매건수 별 고객수 히스토그램(14~17년 전체)
select year(order_date_new), count(distinct `customer name`), count(distinct `Orde
  from superstore_sales_240830
  group by year(order_date_new);
```



  - **약 70.79%의 고객들은 8번 미만 주문을 함**
    - 약 57.76%의 고객들은 4번 이상 8번 미만 주문을 함
    - 약 13.23%의 고객들은 4번 미만 주문함

- **(2014~2015년에 첫 주문을 시작한 고객들 대상) 구매건수 별 고객수 히스토그램**
  - SQL 코드

```
-- 연도/분기 별 고객수 & 주문수 구하기
select year(order_date_new) as year, quarter(order_date_new) as quart, count(disti
  from superstore_sales_240830
  group by year, quart;
```



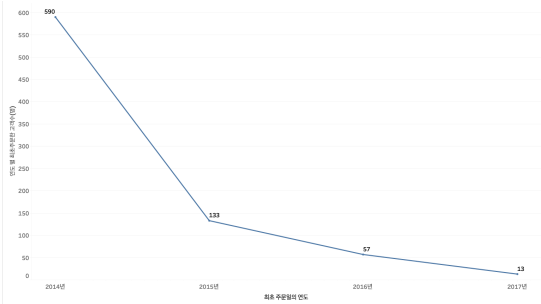
- 약 62.68%의 고객들은 8번 미만 주문을 함

(1-5) 고객 별 첫 주문일(연도 별 신규회원수)

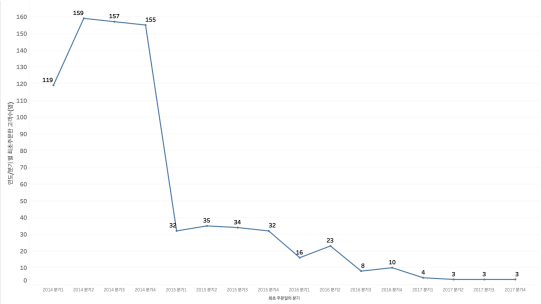
- **연도 별 최초주문 고객수**
  - SQL 코드

```
-- 고객 별 최초 주문일, 마지막 주문일, 최초 주문일과 마지막 주문일 차이
select `customer id`
  , min(order_date_new) as first_order
  , year(min(order_date_new)) as first_order_year
  , month(min(order_date_new)) as first_order_month
  , max(order_date_new) as last_order
  , year(max(order_date_new)) as last_order_year
  , month(max(order_date_new)) as last_order_month
  , datediff(max(order_date_new), min(order_date_new)) as days_between_first_and_l
  , timestampdiff(month, min(order_date_new), max(order_date_new)) as months_betwe
  from superstore_sales_240830
  group by `customer id`;
```

- **연도 별 첫 주문고객수**



- **연도/분기 별 첫 주문고객수**



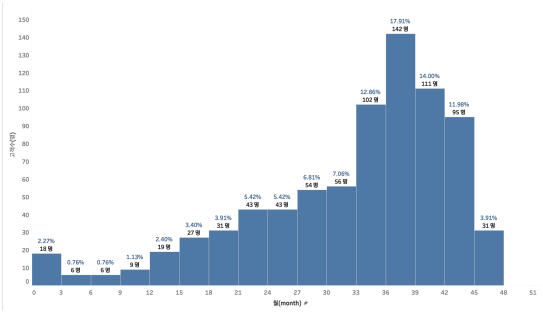
- 2014년 이후로 신규회원의 주문이 크게 줄어듦

(1-6) 고객 유기기간(마지막 주문일 - 최초 주문일)

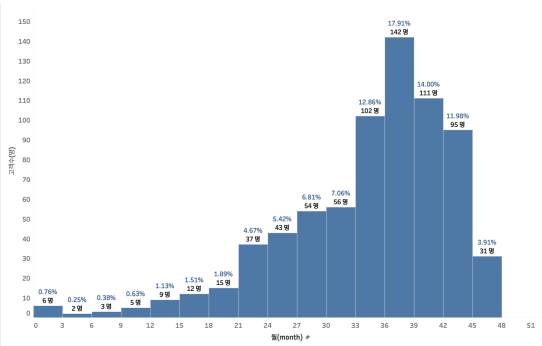
- SQL 코드

```
-- 고객 별 최초 주문일, 마지막 주문일, 최초 주문일과 마지막 주문일 차이
select `customer id`
  , min(order_date_new) as first_order
  , year(min(order_date_new)) as first_order_year
  , month(min(order_date_new)) as first_order_month
  , max(order_date_new) as last_order
  , year(max(order_date_new)) as last_order_year
  , month(max(order_date_new)) as last_order_month
  , datediff(max(order_date_new), min(order_date_new)) as days_between_first_and_last
  , timestampdiff(month, min(order_date_new), max(order_date_new)) as months_between_
  from superstore_sales_240830
  group by `customer id`;
```

- **전체 고유고객 기준**



- **2014~2015년에 주문하기 시작한 고객 기준**



- 한 번 유입되면 장기고객(2.8년 ~ 3.6년 동안 주문)이 되는 경향으로 보이지만, 주문을 자주하진 않는다.

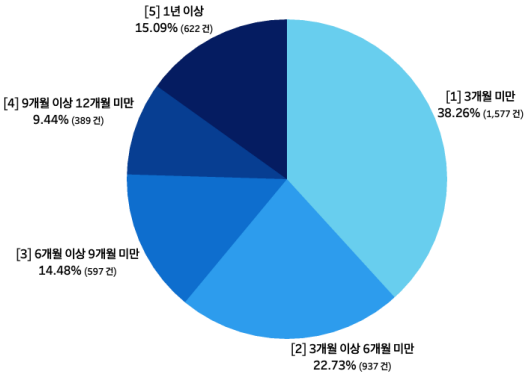
(1-7) 재주문일까지의 소요기간

- SQL 코드

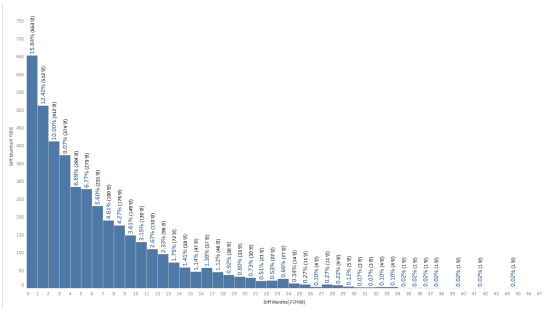
```
-- 고객 별 다음 주문일 계산하기
with unique_order_date as(
 select distinct `customer id`, order_date_new
   from superstore_sales_240830),

next_order_date as (SELECT
    `customer id`
    , `order_date_new`
    , LEAD(`order_date_new`, 1) OVER (PARTITION BY `customer id` ORDER BY `order_date
    FROM unique_order_date
ORDER BY
    `customer id`, `order_date_new`)

select *
  , datediff(next_order_date, order_date_new) as diff_days
```

```
  , TIMESTAMPDIFF(MONTH, `order_date_new`, `next_order_date`) AS diff_months
  from next_order_date
  where next_order_date is not null;
```

- **재주문 내역의 3개월 단위 별 원차트**



- **재주문 내역의 1개월 단위 별 히스토그램**



- **전체 주문의 약 38.26%는 3개월 이내 재주문(61%는 6개월 이내 재주문)**

SQL 활용 코호트 분석

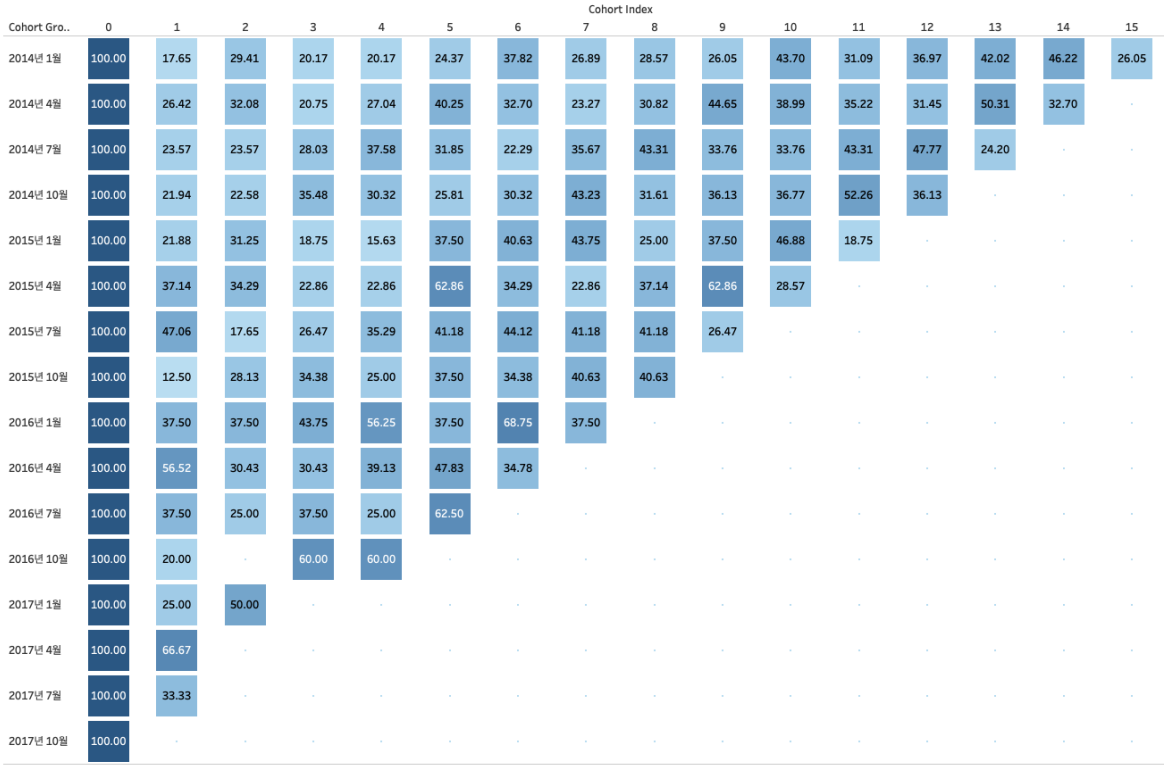
(1) 코호트 분석

- SQL 코드(3개월 단위)

```
WITH first_purchase AS (
    SELECT `customer id`, MIN(order_date_new) AS cohort_day
    FROM superstore_sales_240830
    GROUP BY `customer id`
),
cohort_day AS (
    SELECT
        s.*,
        f.cohort_day,
        TIMESTAMPDIFF(MONTH, cohort_day, s.order_date_new) AS cohort_index,
        CONCAT(
            YEAR(cohort_day), '-',
            LPAD(((QUARTER(cohort_day) - 1) * 3) + 1, 2, '0'), '-01'
        ) AS cohort_group
    FROM
        superstore_sales_240830 s
        LEFT JOIN first_purchase f ON s.`customer id` = f.`customer id`
)
SELECT
    cohort_group,
    FLOOR(cohort_index / 3) AS cohort_index, -- Adjusting the cohort index for 3-mont
    COUNT(DISTINCT `customer id`) AS customer_count
FROM
    cohort_day
```

```
GROUP BY
    cohort_group,
    FLOOR(cohort_index / 3)
ORDER BY
    cohort_group,
    cohort_index;
```

- **코호트 분석**



- **평균적으로 3개월 단위 당 약 34.9% 수준으로 구매율이 유지됨**