



# 출시 상품별 Next 트렌드 예측

권용찬, 박성오, 최영근

# 머리말

**비정형으로 축적된 고객 구매 기록에 최근 많은 관심이 쏟아짐**

- 일반적으로 고객 구매 기록은 고객 구매의 패턴 분석 및 고객의 잠재적 니즈 확보에 도움을 줄 것으로 기대됨
- 관리자 입장에서 미래 판매량 예측(본 컴피티션 주제1)은 주된 관심사
  재고 관리 및 생산량 조절로 원가 절감에 도움을 줄 것으로 기대되며 마케팅 계획에도 참고 가능

**본 보고서에서는 다음과 같이 분석**

- 분석 대상 : 감자스낵
  최초 자료의 235,735 레코드 중 78,685 레코드에 해당
- 보고서의 구성

| PART 1 : 자료 탐색 | PART 2 : 모형 적합 |
|---|---|
| 판매의 경향을 구매횟수, 성별, 지역, 나이 등으로 분할하여 탐색<br>Time-trend plot을 주별/일별로 나누어 다각도에서 인사이트를 얻음 | PART 1에서의 인사이트를 바탕으로 최대한 다양한 시나리오를 구성<br>최종 모형은 정량적인 기준으로 선택 |

* 주어진 데이터가 L.POINT에서 실제로 얻은 기록임을 가정하여 실제 비즈니스에 적용할 수 있는 가설들을 생성하고자 노력함
* 시계열의 4주 트렌드 예측의 본질적인 한계를 분명히 인식하고, 더 나은 예측 결과를 위한 방안을 모색하여 봄.

# PART I : 자료 탐색

전처리

구매 기록 / 구매자 기초정보

구매자 분할 후 판매액 분석

Time-trend (주별/일별)

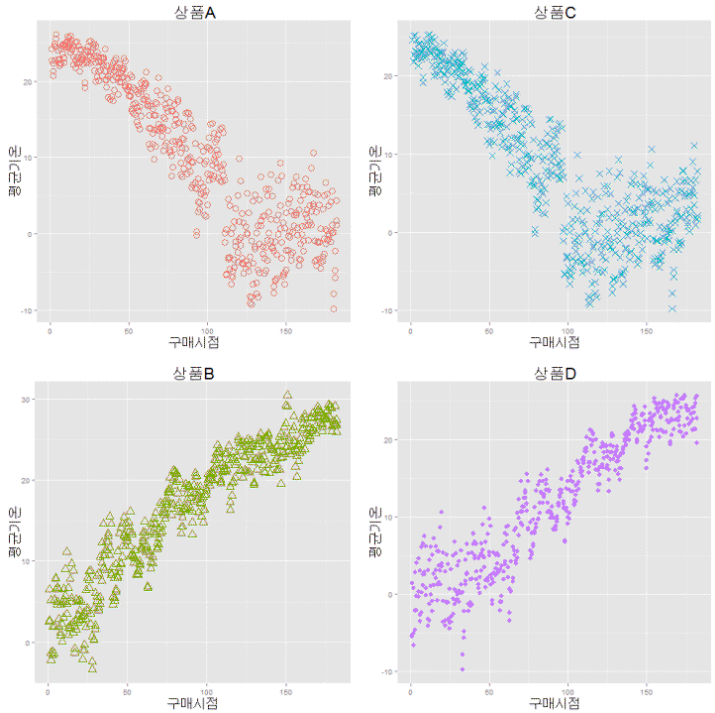
판매액 예측에 유효한 변수 탐색

소결론

# 1-0. 자료 탐색 : 전처리

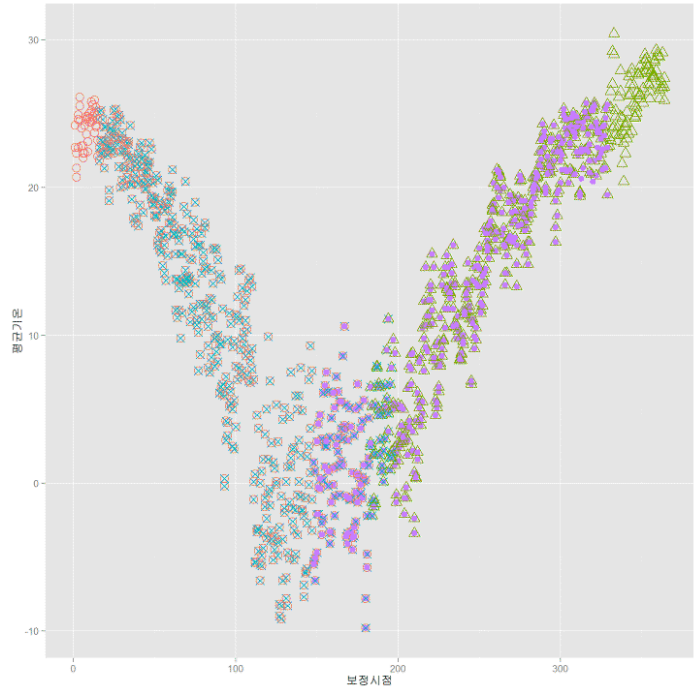
### 판매시점 보정

- 서울 지역 구매기록의 “평균기온”으로부터 제품별 실제 출시일 추론
  (2014-2015 서울 평균기온)과 비교. 서울 평균기온 출처 : 기상청
- 출시 순서 : A (2014-08-13) → C (2014-08-27) → D (2015-01-07) → B (2015-02-11)
- 인터넷 검색으로 상품명 매칭 후 추가 분석 가능하나 competition의 취지를 고려하여 시행하지 않음
- “보정시점” 변수 생성 : A의 출시일(2014-08-13)을 1일로 정의하고 각 상품 구매시점의 절대 날짜를 정렬




상품별 판매시점 vs 기온

가로축 평행이동




보정시점 vs 기온

# 1-0. 자료 탐색 : 전처리

## 중복 기록 제거

- 최초 기록 78,695건 중 행 전체가 중복되는 4,524건 제거 → 유효 기록 74,171건

  Ex) 아래의 81048646번 고객의 구매 기록 중 2, 3, 4, 5번은 같은 시간대에 같은 상품을 구매한 기록.

  상식적으로 한 시간 내의 반복 구매 사건은 전산오류일 가능성이 크므로 이들을 한 건으로 간주

| | ID | gender | age | 카테고리 | 상품구분 | 구매지역 | 구매시점 | 구매시간 | 구매요일 | 구매건수 | 구매금액 | 취소여부 | 평균기온 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1: | 81048646 | 2 | 30 | 감자스낵 | 상품D | 서울 | 56 | 17 | 2 | 8 | 12500 | 0 | 2.8 |
| 2: | 81048646 | 2 | 30 | 감자스낵 | 상품C | 서울 | 88 | 20 | 6 | 1 | 2000 | 1 | 11.0 |
| 3: | 81048646 | 2 | 30 | 감자스낵 | 상품C | 서울 | 88 | 20 | 6 | 1 | 2000 | 1 | 11.0 |
| 4: | 81048646 | 2 | 30 | 감자스낵 | 상품C | 서울 | 88 | 20 | 6 | 1 | 2000 | 1 | 11.0 |
| 5: | 81048646 | 2 | 30 | 감자스낵 | 상품C | 서울 | 88 | 20 | 6 | 1 | 2000 | 1 | 11.0 |

| | ID | gender | age | 카테고리 | 상품구분 | 구매지역 | 구매시점 | 구매시간 | 구매요일 | 구매건수 | 구매금액 | 취소여부 | 평균기온 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1: | 81048646 | 2 | 30 | 감자스낵 | 상품D | 서울 | 56 | 17 | 2 | 8 | 12500 | 0 | 2.8 |
| 2: | 81048646 | 2 | 30 | 감자스낵 | 상품C | 서울 | 88 | 20 | 6 | 1 | 2000 | 1 | 11.0 |

## 취소 기록 보정

- 취소 기록은 전체 74,171건 중 1,659건(2.2%)를 차지. 이들을 (-)의 매출로 기록하여 보정

  Ex) 81048646번 고객의 구매 기록 중 2번 기록의 구매금액 2,000원을 -2,000원으로 보정

| | ID | gender | age | 카테고리 | 상품구분 | 구매지역 | 구매시점 | 구매시간 | 구매요일 | 구매건수 | 구매금액 | 취소여부 | 평균기온 | 보정금액 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1: | 81048646 | 2 | 30 | 감자스낵 | 상품D | 서울 | 56 | 17 | 2 | 8 | 12500 | 0 | 2.8 | 12500 |
| 2: | 81048646 | 2 | 30 | 감자스낵 | 상품C | 서울 | 88 | 20 | 6 | 1 | 2000 | 1 | 11.0 | -2000 |

# 1-1. 자료 탐색 : 구매기록 / 구매자 기초정보

## 성(性)

### 구매기록 기준* (74,171건)

| | |
|---|---|
| 남성 | 40.3% (29,892) |
| 여성 | 59.5% (44,166) |
| 식별불가 | 0.2% (113) |

(단위 : 건)

### 구매자 기준* (37,862명)

| | |
|---|---|
| 남성 | 37.5% (14,211) |
| 여성 | 62.3% (23,576) |
| 식별불가 | 0.2% (75) |

(단위 : 명)

## 지역

### 구매기록 기준* (74,171건)

| | |
|---|---|
| 서울 | 50.6% (37,516) |
| 경기 | 36.9% (27,406) |
| 부산 | 12.5% (9,249) |

(단위 : 건)

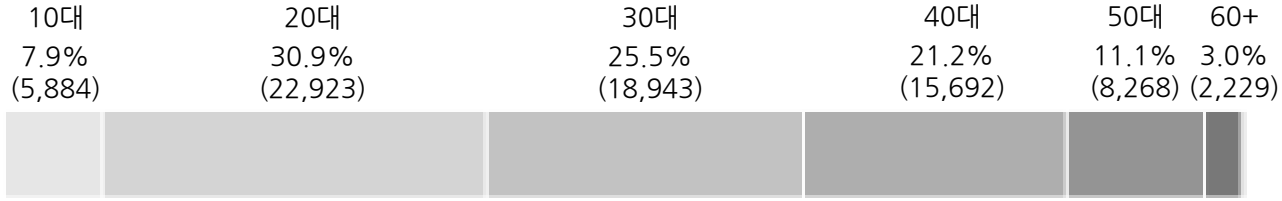
### 구매자 기준* (37,862명)

| | |
|---|---|
| 서울 | 52.9% (20,012) |
| 경기 | 33.7% (12,769) |
| 부산 | 12.7% (4,827) |
| 혼합** | 0.7% (254) |

(단위 : 명)

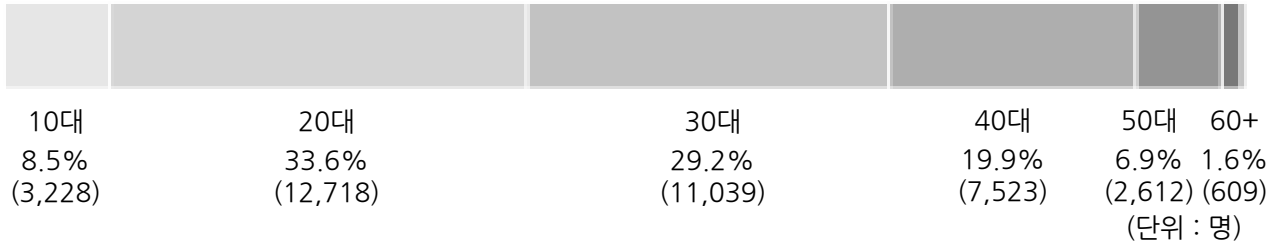
## 연령대

### 구매기록 기준* (74,171건)




(단위 : 건)

### 구매자 기준* (37,862명)




(단위 : 명)

* 구매기록은 주어진 자료의 한 행을 기준으로 하며, 한 명이 여러 상품을 구매한 경우에 한 명으로 간주. 예) 사람 A가 3건, 사람 B가 4건의 구매 기록을 보유한 경우 구매기록은 7건, 구매자는 2명
** 두 지역 이상의 구매 기록이 있는 사람

- 여성 < 남성, 서울이 과반수 이상, 청년층(10-30대)이 약 70%를 차지
- 다수 그룹(여성, 서울, or 청년층)은 구매자 점유율에 비하여 구매기록 점유율이 더 낮음
- 연령대별 1인당 평균 구매 건 수 : 10대 (1.82), 20대 (1.80), 30대 (1.71), 40대 (2.08), 50대 (3.16), 60대 이상 (3.66)

  고연령층일수록 1인당 평균 구매 건 수 증가함. 고연령층이 감자스낵을 더 선호한다고 생각할 수도 있고, 혹은 저연령층이 고연령층보다 L.POINT를 더 간헐적으로 이용할 가능성

# 1-2. 자료 탐색 : 구매자 분할 후 판매액 분석 (1)

**분할 기준 : 구매 횟수**

Motivation : 37,862명의 구매자들이 남긴 기록(row)의 개수가 다름

- 구매자들을 네 계층으로 분류

  I군 : 네 상품을 통틀어 단 1회 구매기록 보유

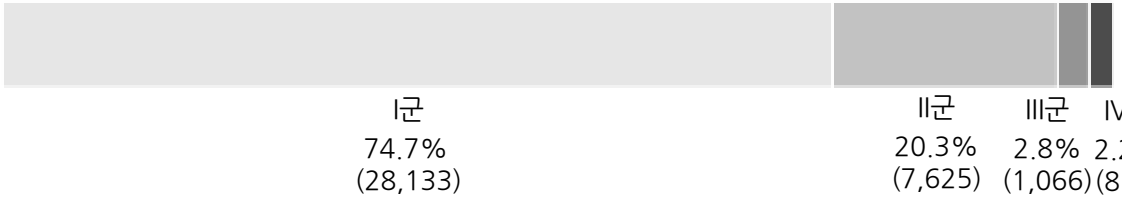
  II군 : 2회 - 5회, III군 : 구매기록 6회 - 10회, IV군 : 11회 이상

- I군, II군이 대부분 (95%). III군, IV군의 구매자는 한달 평균 1건 이상 감자스낵을 구매하므로 감자스낵 애호가로 분류해도 무방. 이들은 5%에 불과하나 판매액의 상당량을 차지하므로 눈여겨봐야 함

- 상품C의 경우 III군, IV군의 구매 비중이 다른 상품에 비해 크며 한 번만 구매한 사람(I군)도 가장 적음

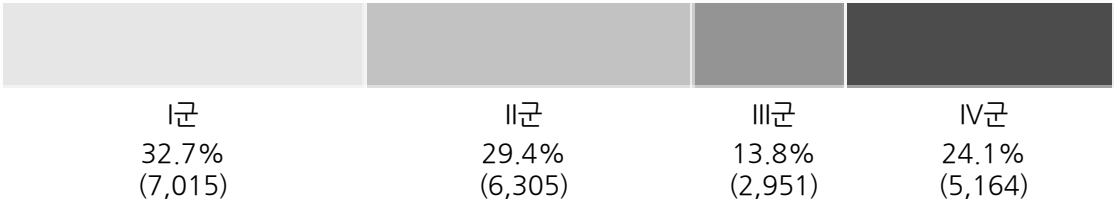
  상품C에 대한 고객의 만족도가 다른 상품보다 높아 재구매율이 높을 가능성 있음, 혹은 상품C가 L.POINT를 통한 프로모션을 진행하여 L.POINT에 등록된 구매 기록이 다른 제품에 비해 상대적으로 많을 가능성도 존재

감자스낵 구매자 37,862명



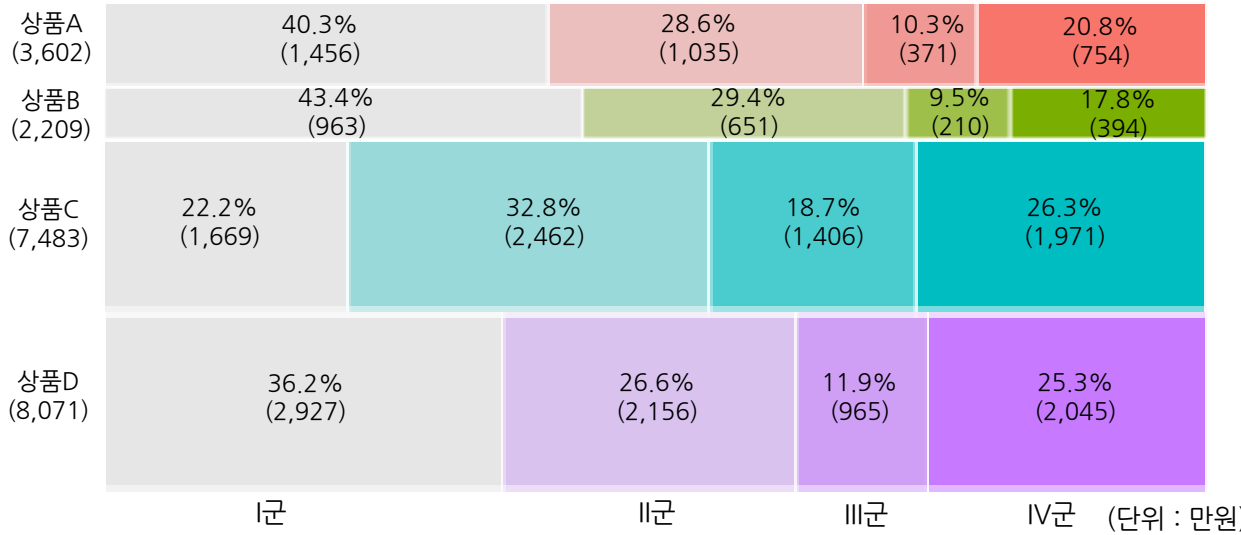

(단위 : 명)

감자스낵 총 판매액 21,435만원




(단위 : 만원)

상품별 I군 - IV군의 점유율




(단위 : 만원)

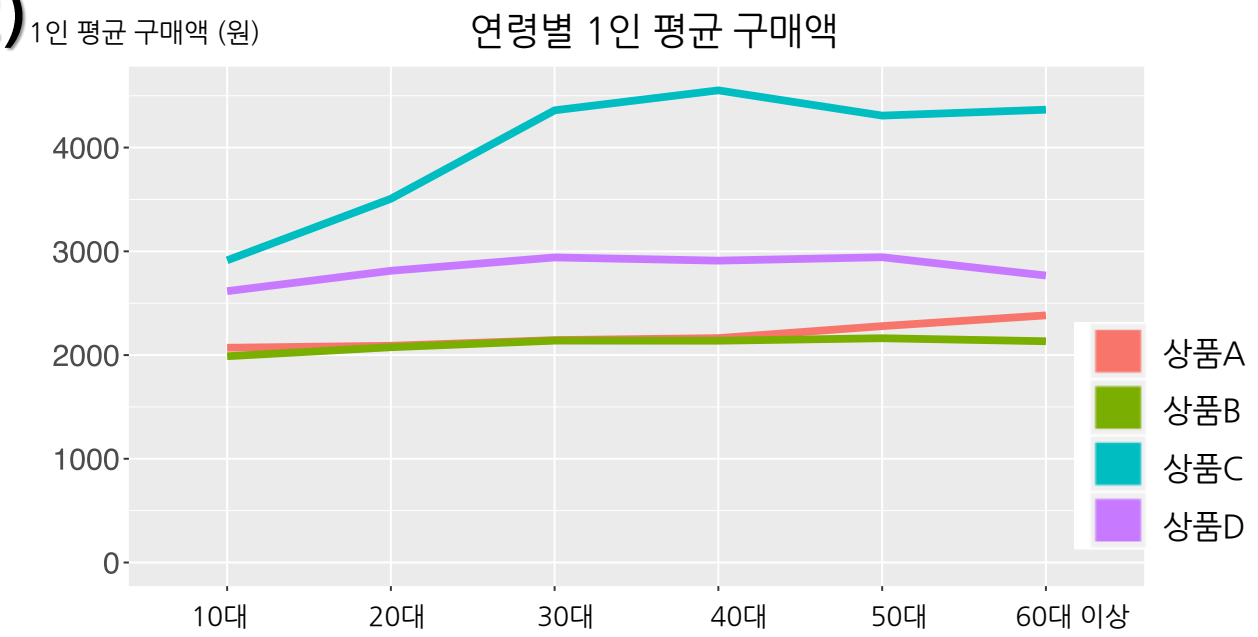
# 1-2. 자료 탐색 : 구매자 분할 후 판매액 분석 (2)

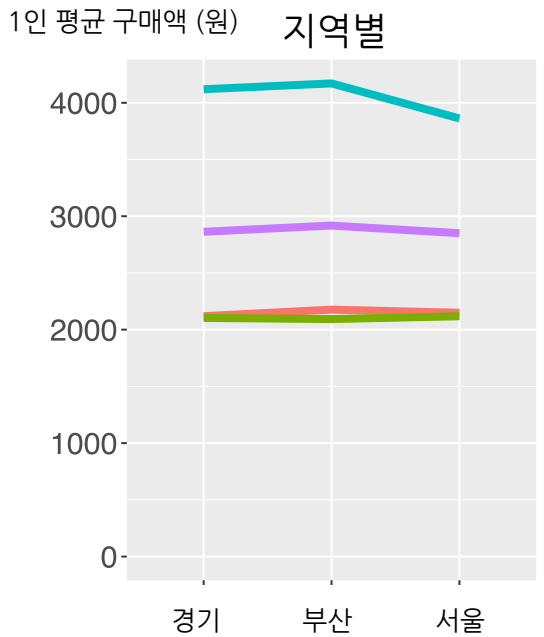
**분할 기준 : 연령 / 지역 / 성**

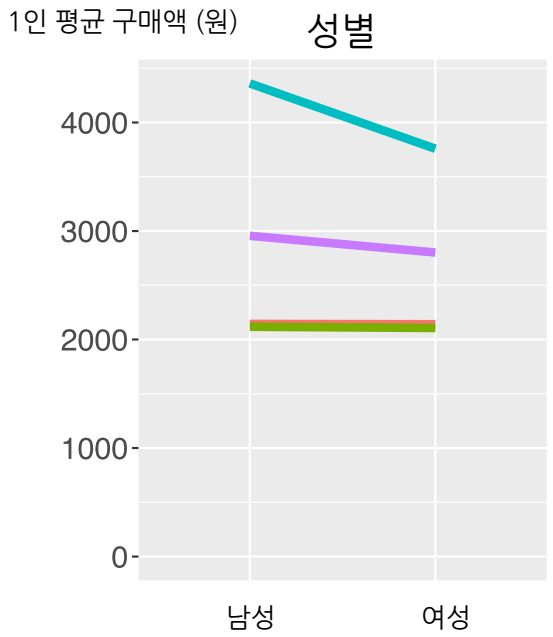
- 연령별 : 주어진 데이터에서는 30대-60대의 1인 구매액이 많음, 특히 C에서는 10대-20대와 큰 차이를 보임
  연령대별 1인당 평균 구매 건 수를 고려하면 30대, 40대가 한 번 구매시 가장 많은 양의 감자스낵 구입
  * p-values : 여섯 그룹 대상 평균 구매액 동질성에 대한 one-way ANOVA 결과 A, B, C, D에서 모두 1e-15 미만, 주어진 곡선의 기울기의 유의성 검정 결과 A, B, C에서 약 0.04, D에서 0.35

- 지역별 : A, B는 경기 < 부산 < 서울,
  C, D는 부산 > 경기 > 서울
  * p-values : 지역별 평균 구매액 동질성에 대한 one-way ANOVA 결과 A와 C는 1e-15 미만, B는 0.001, D는 1e-7

- 성별 : A는 차이없음, B는 근소하게 남성 > 여성,
  C & D는 남성 > 여성
  * p-values : 성별 평균 구매액 동질성에 대한 one-way ANOVA 결과 A는 0.72, B는 0.02, C와 D는 1e-15 미만




* 연령 미상, 성별 미상 구매자는 포함하지 않음

# 1-3. 자료 탐색 : Time-trend (주별)

**제품별 출시 초반 및 26주 뒤의 시장 반응 비교 (그래프 1)**

- A : 출시 초반 저조한 성적 후 매출 상승, 11주차 → 16주차 매출 상승 후 유지
- B : A와 C에 비교하여 출시 초반에는 호조였으나 지속 하강, 가장 낮은 판매액
- C : A와 마찬가지로 출시 초반 저조한 성적 후 매출 상승, 11주차부터 꾸준히 높은 판매액 유지
- D : 출시 초반 압도적인 판매액 기록 후 큰 폭으로 하강하나 C에 이어서 두 번째로 높은 판매액 유지

  출시 초반 대대적인 프로모션이 있었을 것으로 추측됨

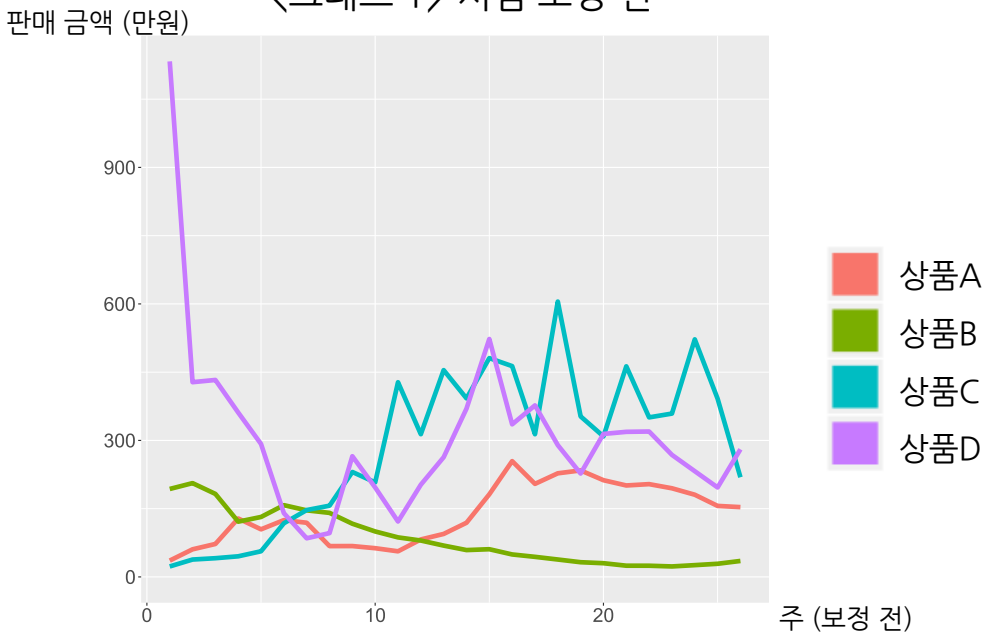
**시점 보정 후의 추가적인 Insight (그래프 2)**

- 후발주자 B는 가장 낮은 판매량을 기록

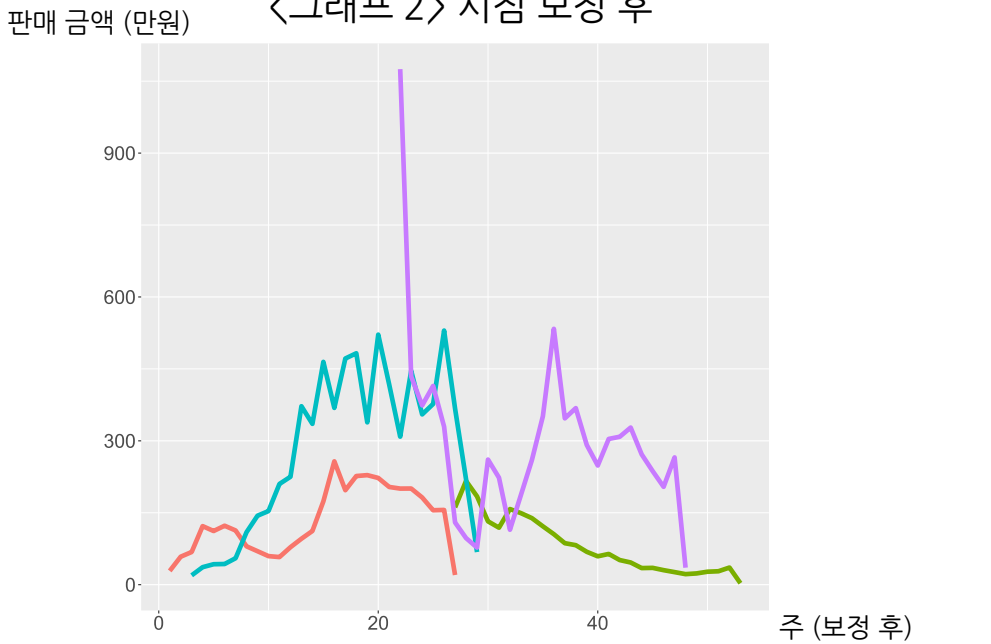
  감자스낵 시장을 선점한 다른 상품들과의 경쟁에서 밀렸을 가능성이 높음
- D의 출시 초반 압도적인 판매량이 A와 C의 매출을 다소 낮추었으나 크게 영향을 주지 않음. 오히려 출시 4주 뒤 C가 다시 판매량 재역전

* <그래프 2>는 <그래프 1>의 완벽한 평행이동이 아님 : 제품마다 출시된 요일이 달라 판매량 집계 방식이 달라졌기 때문
* 따라서 <그래프 2>의 각 상품별 매출액의 가장 첫 주와 마지막 주 판매액은 7일 중 일부분만 집계되어 다소 하락할 수 있음

<그래프 1> 시점 보정 전




<그래프 2> 시점 보정 후



# 1-3. 자료 탐색 : Time-trend (일별)

**일별 Time-trend에서의 추가적인 Insight (그래프 1,2)**

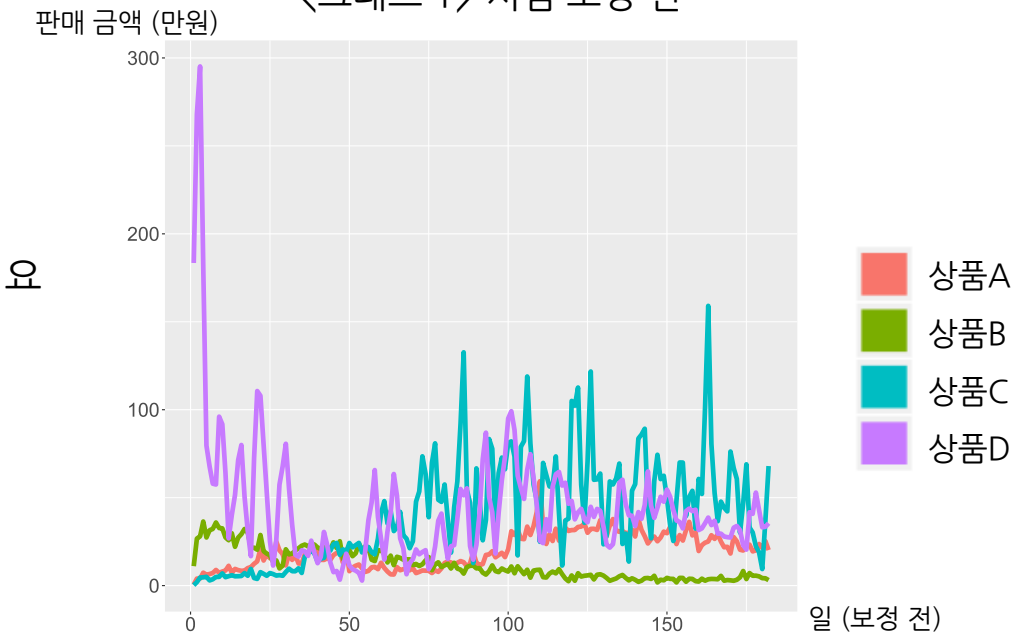
- 마지막 4주(28일)가 Next 4주에 가장 큰 영향을 미치므로 주의깊게 살펴볼 필요
- C, D 판매량이 엇비슷하여 보이나 C의 변동(variation)이 훨씬 심함
- 마지막 4주(28일) 기준 A, B, C, D 일일 판매액의 평균과 표준편차 :

| | 상품A | 상품B | 상품C | 상품D |
|---|---|---|---|---|
| **평균** | 244,642 | 40,352 | 532,842 | 349,201 |
| **표준편차** | 42,496 | 14,224 | 290,507 | 71,012 |
| **평균 / 표준편차** | 5.76 | 2.84 | 1.83 | 4.92 |

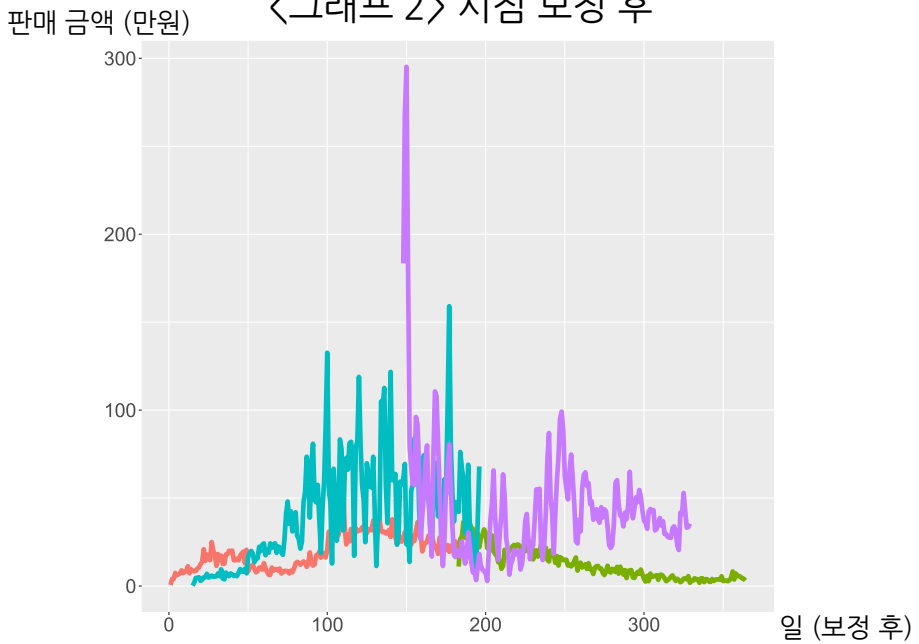
- 주어진 자료만으로 C의 유달리 큰 변동의 원인을 설명하기는 어려움

  C의 물류, 점포 수급 상황을 추적할 필요가 있어보임

<그래프 1> 시점 보정 전




<그래프 2> 시점 보정 후



# 1-4. 자료 탐색 : 판매액 예측에 유효한 변수 탐색 (1)

**기온의 영향이 있는가?**

- A, B, C는 주별 판매금액과 평균 기온에서 강한 상관관계 존재함 (그래프 1)

  * 판매금액 / 기온 자료 모두 서울 기준, 다른 지역도 비슷한 경향을 보였으나 생략

- 섣불리 기온을 유효한 변수로 판단하는 것은 경계하여야 함
  - 상품별로 6개월씩밖에 관측되지 않음 + 관측기간 동안의 거시적인 트렌드에서 온도/판매금액 모두 감소 혹은 증가 (그래프 2)

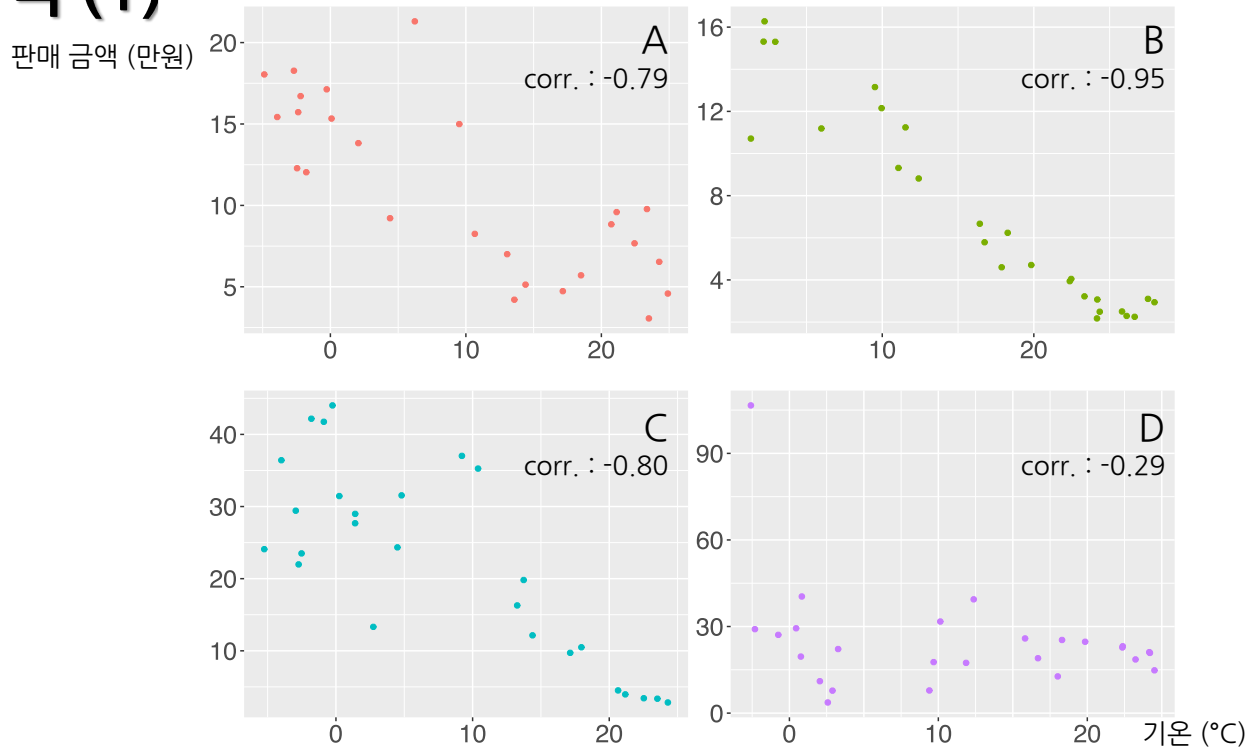
    온도 : 감소하거나 (A, C) 증가함 (B, D)

    판매 금액 : 증가하거나(A, C) 감소(B)
  - 따라서 상관계수가 높게 나올 수밖에 없고 만일 기온을 판매금액의 설명 변수로 넣을 경우 오류 가능성 존재

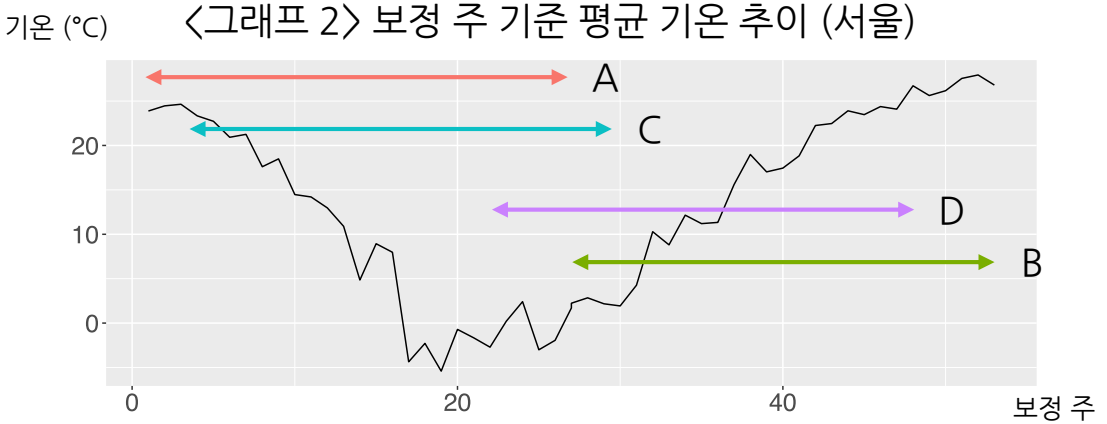
    ex) 2월에 태어난 신생아의 키를 6개월만 추적할 경우, 2월-7월간 기온과 키 모두 증가하므로 신생아의 키가 기온에 비례한다고 판단할 수도 있음

- 따라서 모형 설정 단계에서 기온은 배제

<그래프 1> 주별 판매 금액 vs. 평균 기온 (서울)




<그래프 2> 보정 주 기준 평균 기온 추이 (서울)



# 1-4. 자료 탐색 : 판매액 예측에 유효한 변수 탐색 (2)

**요일의 영향이 있는가?**

특정 요일에 판매액이 더 많거나 / 주중보다 주말에 판매액이 더 높으면
이를 예측 모형에 반영해볼 수 있음

- 그래프 (우측) : 가시적인 트렌드는 보이지 않음

- 귀무가설 1 : 요일별 평균 판매금액이 같다

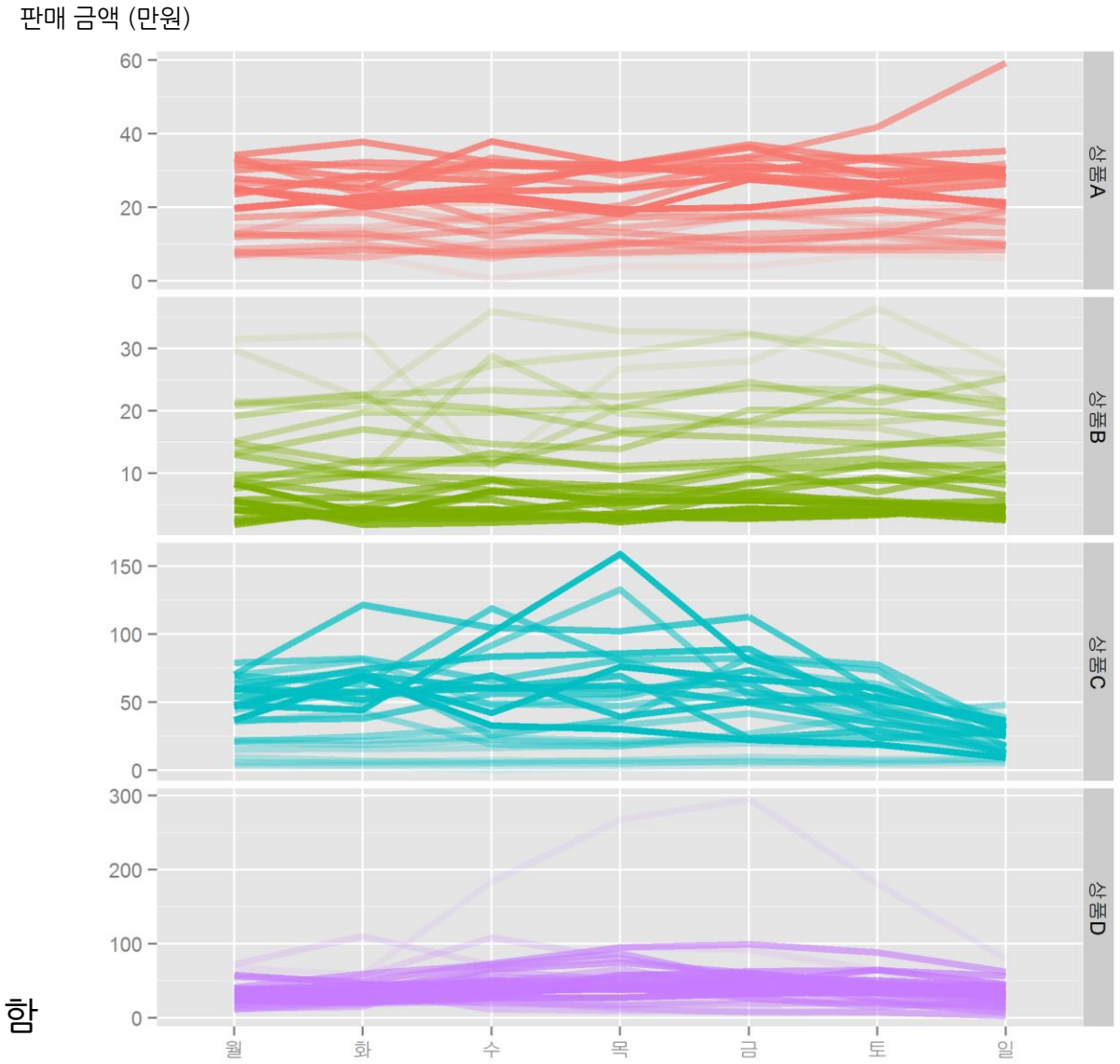
  요일별 26개의 판매금액 및 일곱 요일에 대하여
  One-way ANOVA 결과 모든 상품에 대하여 귀무가설 기각
  (p-value : A, C, D의 경우 1e-15 미만, B의 경우 7.9e-6)

- 귀무가설 2 : 주중 평균 판매금액과 주말 평균 판매금액이 같다

  Two-sample t-test 결과 모든 상품에 대하여 귀무가설 기각
  (p-value : A, C, D의 경우 1e-15 미만, B의 경우 0.035)

- 요일별 평균 판매금액이 동질하지 않으므로 모형에서 고려해 보기로 함

일 판매량의 요일별 분해




* 182일간의 시계열을 26개 주로 분할. 각 곡선 하나가 한 주의 판매 추이를 의미.
* 색이 옅을 수록 과거 기록, 짙을 수록 최근 기록

# 1-5. 자료 탐색 : 논의 및 소결론

### (1-1) 구매기록 / 구매자 기초 정보 및 (1-2) 구매자 분할 후 판매액 분석

- 6개월간 단 1회 구매한 고객이 전체의 75%이며, 5%에 불과한 다량의 구매 고객이 매출의 많은 비중을 차지
  → 동일 상품 재구매 시 별도 포인트 적립 혜택 등의 구매 유인책으로 매출 상승을 노려볼 수 있음
- 30대-40대가 방문 횟수당 감자스낵 구매량이 상대적으로 높은 편
  → 30대-40대 고객을 대상으로 구매 혹은 매장 방문을 독려한다면 다른 연령대보다 매출 증대에 더 많은 효과가 있을 것으로 예상

### (1-3) Time-trend

- 출시 초반 및 6개월 뒤의 시장 반응이 다름, 거시적으로 A, C는 상승 후 안정, B는 하락세, D는 초반에 잘 팔린 후 안정
  → A와 C는 2014년 허니버터칩 열풍의 영향을 받았을 것으로 예상, D는 프로모션이 있었을 것으로 추측. B는 꾸준한 매출 감소의 원인 탐색 및 대안 모색 필요
- C는 일별 판매량에 변동(variation)이 유달리 큼 → 원인 탐색 필요, 재고 불안정이 원인일 경우 재고 관리로 매출 증대 가능성 있음

### (1-4) 판매액 예측에 유효한 변수 탐색

- 자료가 6개월간만 수집되었기 때문에 기온의 사용은 바람직하지 않음
- 요일을 모형에 편입해보기로 결정

# PART II : 모형 적합

자료 및 문제의 본질

자료 가공 방법

Forecast 방법론

모형 구축

분석 결과

소결론

# 2-1. 모형 적합 : 자료 및 문제의 본질

**주어진 자료의 특성**

- 시계열 자료 : 각 관측치에 시점이 존재하며 독립이 아님

- 공변량(covariate) 사용의 어려움
  - 성별, 지역, 연령 : 미래의 구매자에 대한 정보는 얻을 수 없음
  - 기온 : 감자 스낵의 판매액과 깊은 관련은 없어 보임 (1-4-(1) 참고)

- **따라서 대부분의 통계 분석(회귀분석, 기계학습 등)을 적용하기가 힘듦**

  회귀분석, 기계학습(지도학습, supervised learning)은 독립/분포동일 가정에 기반하나 본 자료는 관측치가 독립적이지 않음

  이를테면 강아지/고양이 사진을 자동으로 분류하여 새로운 강아지/고양이 사진을 분류하는 문제와는 본질적으로 다름

- 본 보고서에서는 전통적인 시계열 분석 방법론 적용, 이를 기반으로 4주간의 매출액을 예측

  하나만의 모형을 고려하지 않고 최대한 다양한 가정 아래서 자료 가공 및 적합 방법론 고려

# 2-2. 모형 적합 : 자료 가공 방법

**적합의 품질을 높이기 위해 다양한 자료 가공 방법을 고려**

- 74,171 건의 구매기록에서 상품별 판매량을 일별(daily) 혹은 주별(weekly)로 집계
  일별 집계(daily) : 각 상품별 182일의 시계열을 얻음. 자료가 많으나 자료의 변동(variation)이 심함
  주별 집계(weekly) : 각 상품별 26주의 시계열을 얻음. 자료의 변동이 적으나 자료의 수가 충분치 않음

- 일별 또는 주별 판매액에 로그(log10) 변환 고려
  경험적으로 로그변환을 통하여 자료의 변동을 안정시키고(variance stabilization)및 자료를 정규분포에 더 잘 근사시킬 수 있음

- 출시 후 첫 4주(28일)의 판매량 기록 제거 고려
  자료 탐색 결과 출시 초반의 판매량은 출시 후 6개월의 판매 추이와 큰 관계가 없었으며 프로모션의 효과가 강함
  첫 4주의 자료가 모형의 적합도를 떨어뜨릴 가능성을 고려

→ 따라서 자료 가공 방법은 총 2 × 2 × 2 = 8 가지

# 2-3. 모형 적합 : Forecast 방법론

- 주어진 시계열 자료가 곡선이 정상과정(stationary) / 비정상과정(non-stationary process)일 가능성을 모두 고려
  - ARIMA(p,d,q) : 정상과정에 적합, 주어진 시계열 자료의 d-차분을 ARMA(p,q) 과정으로 표현하는 모형
    ARMA(p,q) : 각 시점의 관측값을 직전 p 시간의 관측값 및 직전 q시간의 잡음(error)의 선형 결합으로 표현한 모형
    잡음에 정규분포 가정, 계수는 maximum likelihood로 적합, d,p,q는 AIC에 의하여 결정
  - ETS : 비정상과정에 적합, 시계열의 잡음(error), 추세(trend), 계절(seasonal) 성분을 합(곱)의 방법으로 표현하는 모형
    지수 평활법(exponential smoothing)에서 잡음을 합/곱/상쇄곱의 형태로 더해준 모형, forecast 방법은 지수 평활법과 동일
    잡음에 정규분포 가정, 잡음/추세/계절 각 성분을 미존재/합/곱/상쇄곱 중 어떤 것으로 모형화할지는 AIC에 의하여 결정
  - 두 방법 모두 R의 forecast 패키지에서 이용 가능

- 요일별 매출 변동을 계절성(seasonality) 요인으로 고려
  - STL : 각 시점에서의 적합값을 인근 시점 관측값들의 국소회귀모형에서 추론하여 최적의 추세(trend), 계절(seasonal) 성분을 찾아내는 방법
    STL 적용 후의 잔차(residual, 주어진 시계열에서 추세와 계절성분을 제거하고 남은 값)에 ARIMA나 ETS를 적용하면 적합 성능이 좋아지는 것으로 경험상 알려져 있음
    R에 기본 탑재된 기능 패키지에서 이용 가능
    STL 후의 잔차에 ETS를 적용하면 ETS 패키지가 프로그램이 자동적으로 ETS의 계절성분 모형화를 제거함.

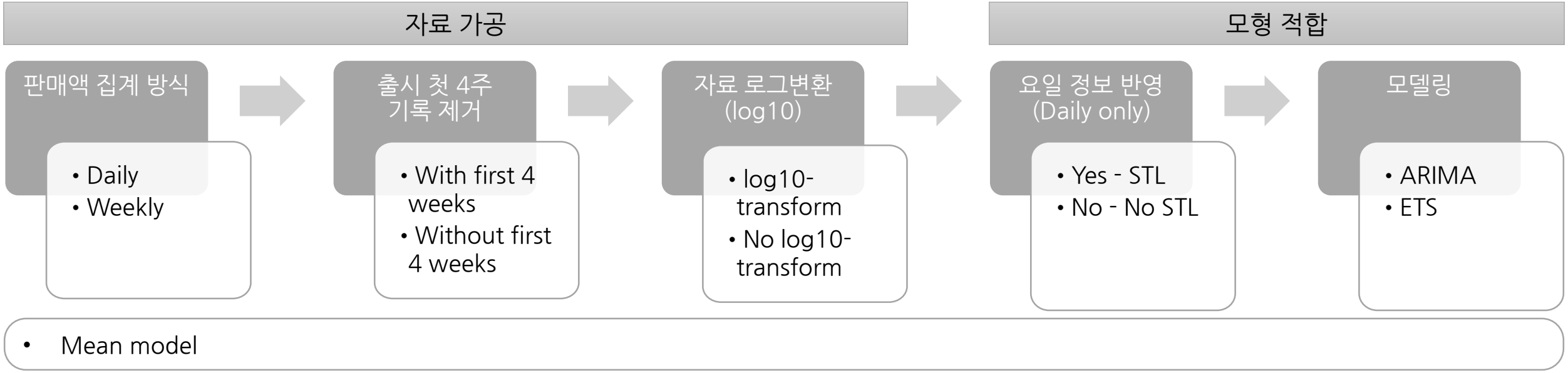
→ 따라서 예측 방법은 daily 자료의 경우 2 × 2 = 4가지 (요일별 매출 변동 고려 가능), weekly 자료의 경우 2가지

# 2-4. 모형 적합 : 모형 구축

**자료 가공 - Forecast 시나리오 요약 :** 총 16가지(daily) + 8가지(weekly) + 1가지(mean model) = 25가지

ex) Daily - without first 4 weeks - log10 transform - STL - ARIMA :

일별로 집계한 네 상품의 판매액에서 첫 4주 구매기록을 제거하고 각 일별 판매액을 log10 변환. 이 자료를 STL로 적합한 후 잔차에 ARIMA 적용, 적합된 모형으로 다음 4주(28일)의 일일 판매량들을 예측




* Mean model : 미래 4주의 판매액을 직전 4주의 판매액의 평균으로 예측하는 방법. 모형 기반의 예측이 최소한 이겨야 하는, 대조군의 방법으로서 고려

Mean model을 사용할 경우 자료 가공 여부에 따라 예측값이 같거나 매우 사소하게 변하기 때문에 자료 가공 과정을 따로 거치지 않음

- 상품별로 최적의 가공/forecast 모형 선택
- 최적 모형 선택 기준 : 1주차 - 22주차 (혹은 5주차 - 22주차) 자료로 23주차 - 26주차를 가장 적은 오차로 예측하는 모형

  오차 계산 방법 : (모형으로부터의 예측액 - 실제 판매액)을 주별로 계산한 뒤 절대값 합, daily 기반으로 예측했을 경우 예측액/판매액을 7일씩 모음

# 2-5. 모형 적합 : 분석 결과 (1)

23 - 26주 오차(괄호), 단위 : 만원

ETS(error, trend, seasonal) : N, None; A, Additive; M, Multiplivative; Ad, Addtivie damped;
ARIMA(p,d,q) : p, 자기상관 차수; d, 차분 수; q, 이동평균 차수

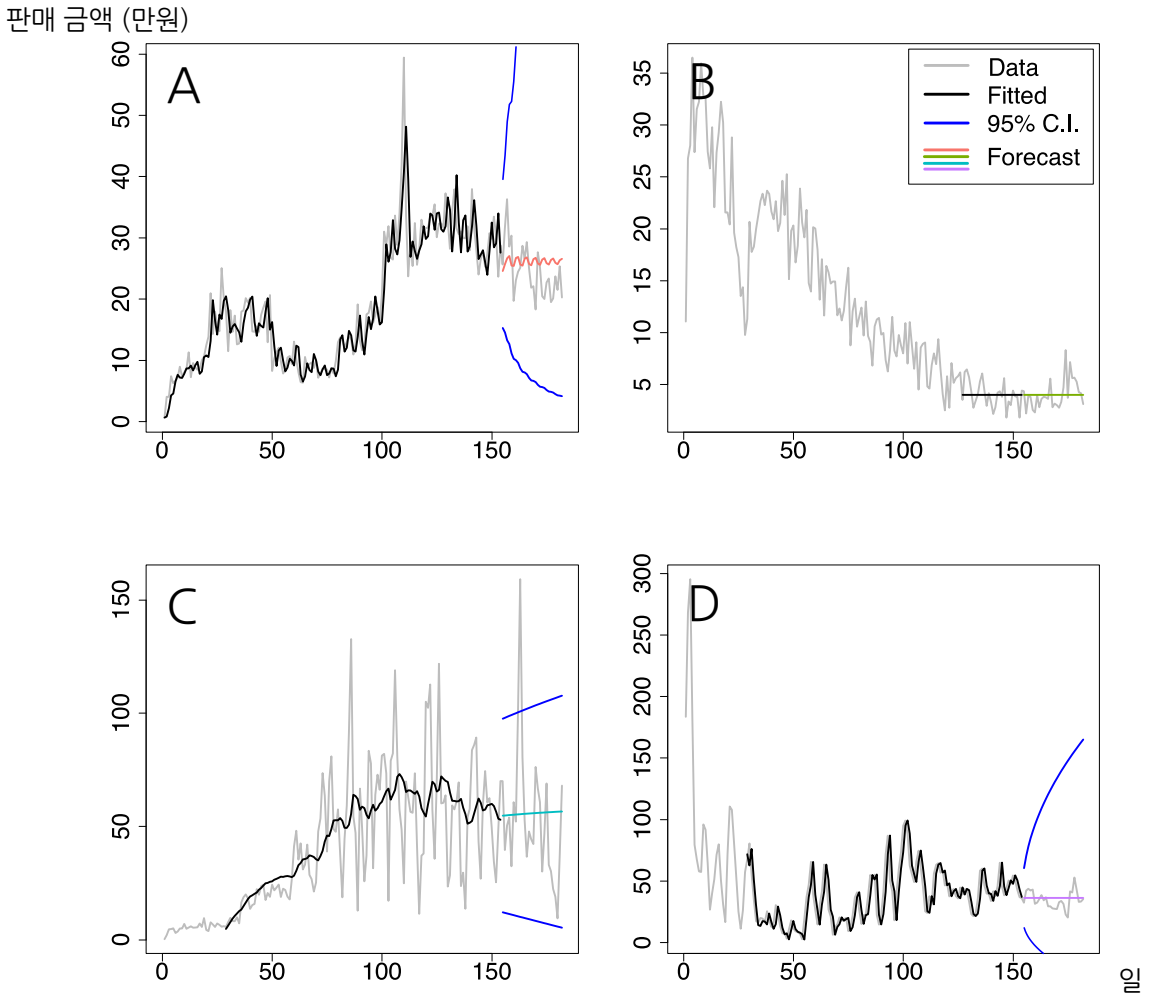
| | 1등 (최종 선택 모형) | 2등 | 3등 |
|---|---|---|---|
| 상품A | Daily / With first 4 weeks / Log10<br>ARIMA(3,1,2)<br>(예측오차 : 71.9만원)<br>* 오차율 : 23-26주차 총 판매 금액 685만원의 10.4% | Daily / Without first 4 weeks / Log10<br>STL-ARIMA(2,1,3)<br>(78.53) | Daily / With first 4 weeks / Log10<br>ETS(A,Ad,N)<br>(92.87) |
| 상품B | Mean model<br>(15.12)<br>* 오차율 : 23-26주차 총 판매 금액 113만원의 13.3% | Daily / With first 4 weeks / No log<br>STL-ETS(M,N,N)<br>(15.53) | Daily / Without first 4 weeks / No log<br>STL-ETS(A,Ad,N)<br>(15.69) |
| 상품C | Daily / Without first 4 weeks / No log<br>ETS(M,Ad,N)<br>(335.01)<br>* 오차율 : 23-26주차 총 판매 금액 1492만원의 22.4% | Mean model<br>(335.95) | Daily / With first 4 weeks / No log<br>ETS(M,N,N)<br>(335.95) |
| 상품D | Daily / Without first 4 weeks / No log<br>ETS(A,Ad,N)<br>(119.7)<br>* 오차율 : 23-26주차 총 판매 금액 978만원의 12.2% | Daily / With first 4 weeks / Log10<br>ETS(A,N,N)<br>(119.72) | Daily / Without first 4 weeks / Log10<br>ETS(A,N,N)<br>(119.72) |

- Daily가 weekly에 비하여 모든 상품에서 오차가 더 낮음 : 자료의 변동보다 자료의 개수가 예측의 정확성을 더 높인 것으로 추측됨
- 첫 4주 기록 제거/비제거의 영향은 미미함 : 첫 4주 판매액의 영향이 26주 뒤 예측에서 미미함을 이미 모형 안에서 설명하고 있는 것으로 해석할 수 있음
- A에서는 ARIMA 계열이, B/C/D에서는 ETS 계열이 우세 : 그러나 판매액 time-trend에서 정상성 / 비정상성을 주장하기는 쉽지 않음
- Mean model은 B, C에서 상대적으로 낮은 오차를 보여주며 A, D에서는 가장 우수한 시나리오에 비하여 다소 뒤처짐

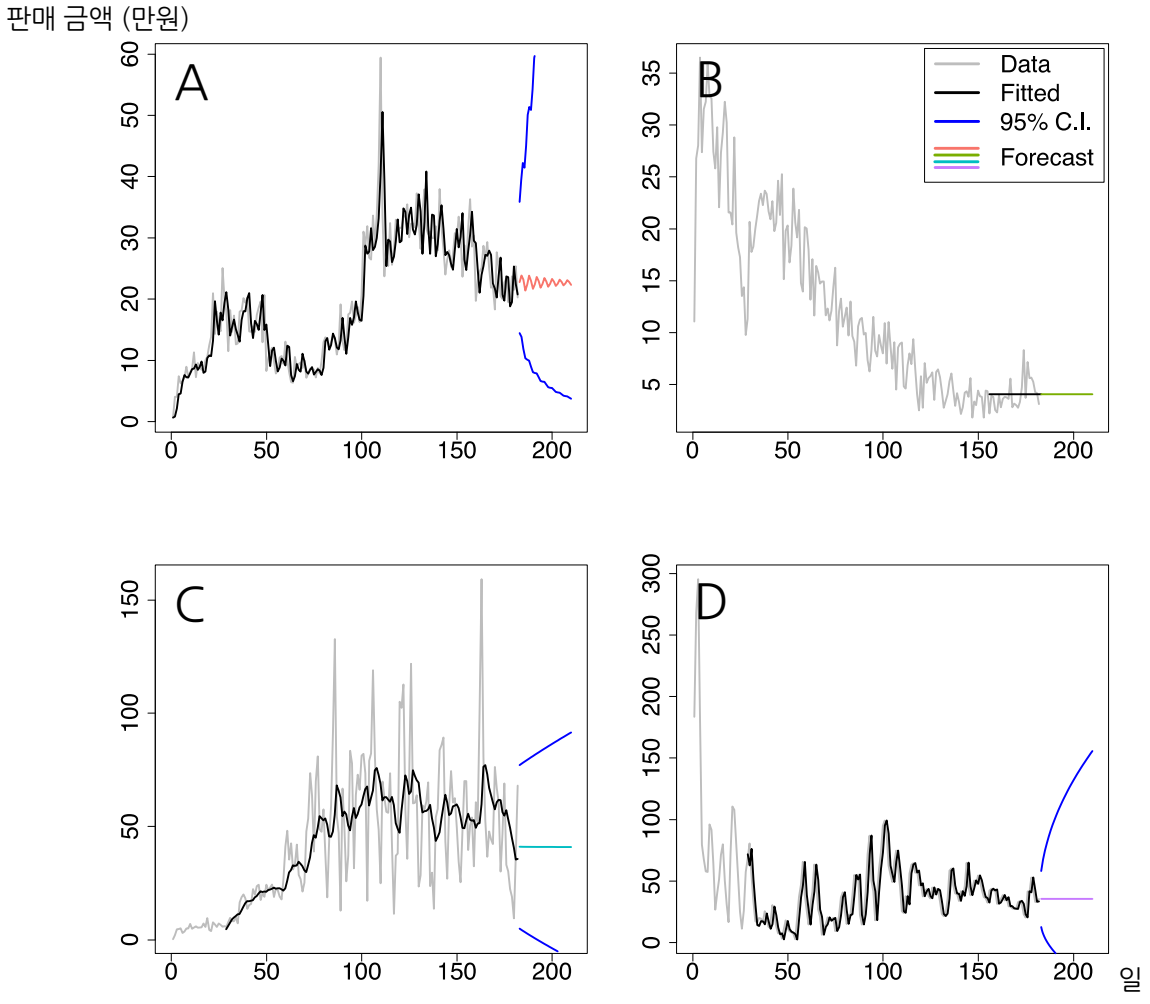
  B는 판매액이 하락 상태에서 안정되었기 때문에 / C는 판매액에 변동이 매우 심하기 때문에 모형화로로 얻는 이득이 크지 않았을 수도 있음

# 2-5. 모형 적합 : 분석 결과 (2)

## 최종 시나리오 (22주차까지의 자료로 23-26주차 예측 결과)




## 최종 시나리오 (26주차로 27-30주차 예측 결과)




* B : 직전 4주만의 자료로 예측하였기 때문에 그 이전의 적합(fitted) 곡선 미존재, 또한 분포 가정이 없으므로 신뢰구간 미존재
* C , D: 최종 시나리오가 첫 4주의 자료를 제거하였으므로 첫 4주의 적합(fitted) 곡선 미존재

# 2-6. 모형 적합 : 소결론

PART 1의 결과로부터, 시계열의 내재적인 특성(요일정보)만으로 향후 4주의 매출을 예측

- 상품마다 최적 시나리오가 선택한 데이터 전처리 방법 및 적합 방법론이 다름
- 일반적으로 모형 기반의 방법이 더 나은 예측을 보일 것으로 예상되었으나, B와 C에서는 의외로 mean model도 좋은 예측 결과를 보임
- 사실 주어진 시계열 자료가 너무 짧은 시간 동안 관측되었으며, 실제 시장에서 생각할 수 있는 복잡도에 비해서 사용할 수 있는 독립변수가 부족하여, 주어진 자료로는 정확한 forecast가 불가능하다 판단됨

**최종 선택 시나리오가 예측한 27주차-30주차 주별 판매금액 (단위 : 만원)**

각 상품마다 최종 선택 시나리오를 26주차까지의 자료에 적용

| | 1주차 | 2주차 | 3주차 | 4주차 |
|---|---|---|---|---|
| 상품A | 160.18만원 | 158.14 | 157.97 | 158.75 |
| 상품B | 28.25 | 28.25 | 28.25 | 28.25 |
| 상품C | 286.54 | 286.34 | 286.23 | 286.17 |
| 상품D | 248.15 | 248.15 | 248.15 | 248.15 |

# 결론 및 제언

**결론**

- PART 1 : 약 78,000 건의 개별 자료로는 뚜렷한 경향을 볼 수 없었으나 다방면으로 요약한 결과,
  - 프로모션의 여지가 있는 계층이 존재 (1회 구매 고객 혹은 고연령층)
  - 상품별 변동(특히 C)의 원인을 추가적으로 추적할 필요가 있어 보임
- PART 2 : 26주간의 시계열 자료로 향후 4주 매출액 예측
  - 상품마다 일괄적인 모형 대신 개별 상품에 최적화된 모형이 바람직해 보임
  - 23-26주차 자료로부터 계산된 오차가 실제 판매액의 10%-20% 가량, 오차율을 줄이려면 다방면의 노력이 요구됨

**아쉬웠던 점 및 제언 :** 빅데이터 분석의 효용성은 (명확한 목표) + (지속적이고 다양한 자료의 수집) + (품질 관리)에 달려 있으므로 이를 극대화하기 위한 방안 모색

- 26주 보다 긴 시간(몇 년)에 걸쳐서 관측되었다면, 분기 및 계절의 영향을 관측할 수 있고 장기적 경향도 포착할 수 있었을 것으로 기대
- 상품의 구체적인 정보(상품명, 회사명)가 주어졌다면, SNS나 기사자료를 활용하여 판매액의 추이를 설명하는 폭 넓은 분석 가능

  ex) 보정시점에 맞추어 상품을 검색한 뒤에 텍스트 마이닝
- 성별, 나이, 지역 별 자료를 공변량으로 사용하기 위하여는 L.POINT 자료가 모든 감자스낵 구매 고객을 대표할 수 있도록 가중치 보정 방법을 강구할 필요 있음

  ex) L.POINT를 이용하지 않을 수도 있는 일반 시민 (혹은 롯데마트/롯데백화점 이용자) 기반의 별도 패널 구축으로
  L.POINT 데이터에 대한 가중치 부여 / 성향 점수(propensity score) 부여 가능
- L.POINT 구매 기록의 지속적인 품질 점검을 꾀하고 싶다면, 별도의 자료 구축으로 보완할 수 있어 보임

  ex) 상기 서술된 패널 데이터, L.POINT 고객에 대한 별도의 스마트 서베이로 L.POINT 기록이 L.POINT 이용자의 실제 성향을 반영하고 있는지 평가 가능