한입만조(C2)

팀장	미정
팀원	박정우 송현준 유정하
제출 자료(Git, Drive)	https://www.notion.so/C2-2e88820152dc81b9a8a7d55ca66364d9
대시보드 자료	fastfood.twbx
데이터셋	Fast Food Marketing Campaign A/B 테스트 데이터셋

▼ Day1

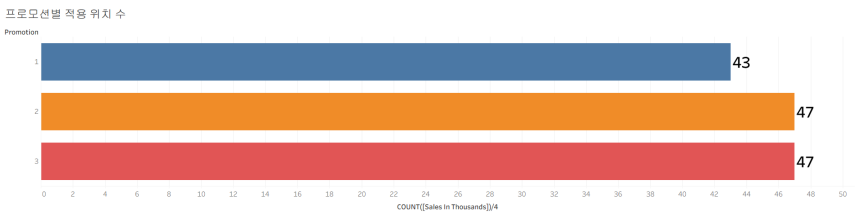
▼ 송현준

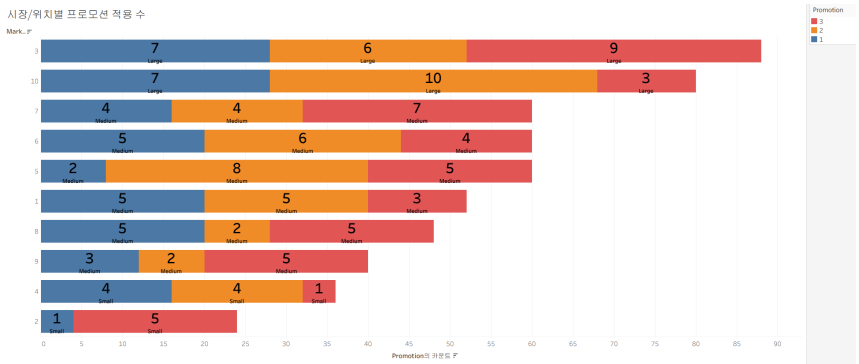
데이터 Overview

변수	정의	설명	분류	타입
MarketID	마켓 아이디	시장을 식별하는 고유 식별자 (1~10)	범주형	문자열
MarketSize	마켓 크기	마켓의 크기 (Small, Medium, Large)	범주형	문자열
LocationID	위치 아이디	매장 위치를 식별하는 고유 식별자 (1~920 / 137개)	범주형	문자열
AgeOfStore	매장 연령	매장이 오픈한지 몇 년이 지났는 지	연속형	정수
Promotion	프로모션	진행한 프로모션의 종류 (1~3)	범주형	문자열
week	주차	프로모션이 진행된 4주 중 한 주 (1~4)	범주형	정수
SalesInThousands	매출액(K)	특정 LocationID , 프로모션 및 주차별 매출액(천 단위)	연속형	실수

데이터 전처리

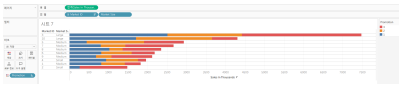
데이터 정합성 확인





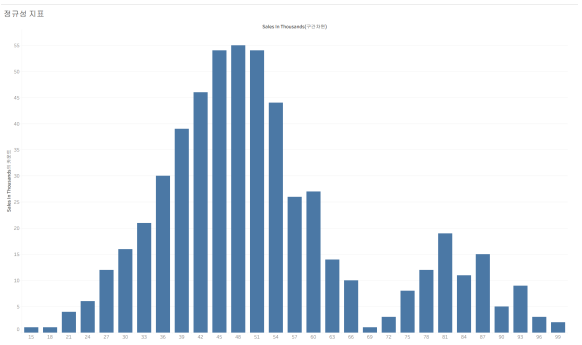
Market 2의 경우, Promotion 2가 배정되지 않았으며 Promotion 3이 과도하게 많이 몰려있다.

따라서, 이는 데이터 정합성을 해할 수 있으므로, 데이터에서 제거하기로 한다.



ANOVA 분석

1. 정규성 확인

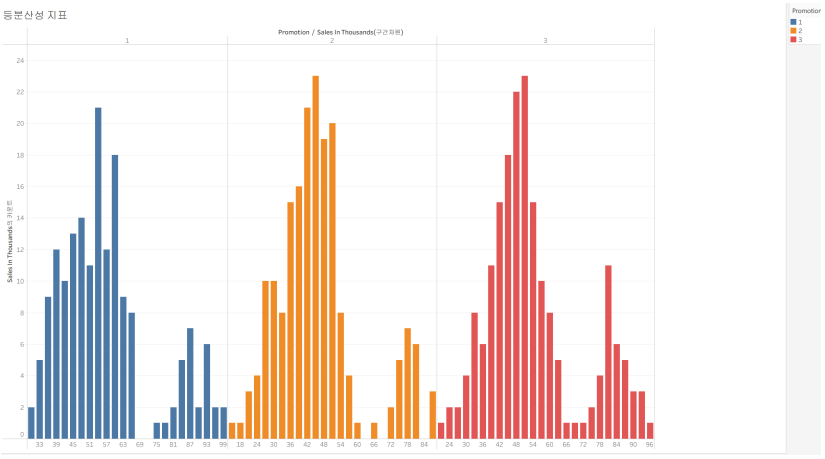



```
y=df['SalesInThousands']
model = ols('y ~ C(Promotion)', data=df).fit()

stat, p = shapiro(model.resid)
print(f'stat: {stat:0.4f}\np: {p:0.4f}')

# stat이 1에 가까울수록 정규성 만족
# p-value가 0.05보다 작으므로 귀무가설 기각 -> 정규성 만족하지 않음
```

0.0s

```
stat: 0.9221
p: 0.0000
```

2. 등분산성 확인




```
groups = [
    df.loc[df['Promotion'] == 1, 'SalesInThousands'],
    df.loc[df['Promotion'] == 2, 'SalesInThousands'],
    df.loc[df['Promotion'] == 3, 'SalesInThousands']
]

stat, p = levene(*groups)
print(f'stat: {stat:0.4f}\np: {p:0.4f}')
# p-value가 0.05보다 크므로 귀무가설 채택 -> 등분산성 만족
```

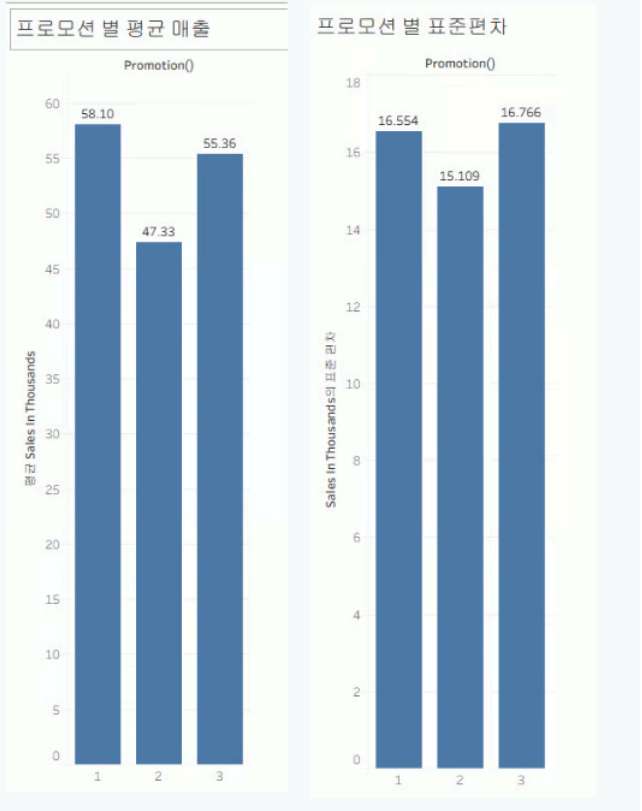
0.0s

```
stat: 1.2697
p: 0.2818
```

3. 독립성 확인

데이터가 각 `LocationID` 마다 1~4주 총 4개의 데이터가 존재하므로 독립성에 문제가 있다.
따라서, 이를 해결하고 ANOVA를 진행하도록 한다.

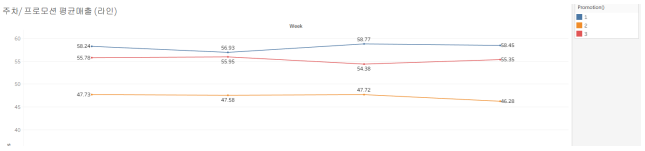
▼ 박정우



Promotion 1은 비교적 높은 평균 매출과 중간 수준의 변동성을 보여, 성과와 안정성 측면에서 가장 균형 잡힌 프로모션으로 판단된다.

Promotion 2는 평균 매출이 가장 낮지만 표준편차도 가장 낮게 나타나 3개의 프로모션중 가장 안정적인 프로모션으로 판단된다.

Promotion 3은 평균 매출은 높았으나 표준편차 또한 가장 크게 나타나 성과 변동성이 높은 프로모션으로 확인되었다.



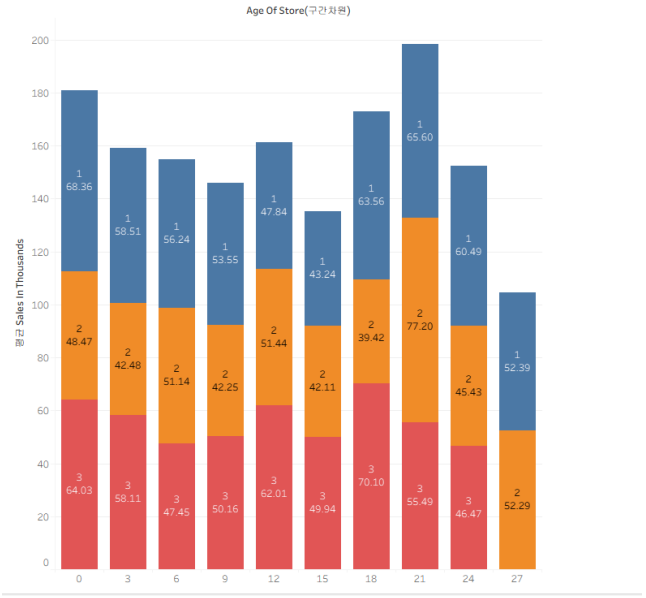
Promotion 1 > Promotion 3 > Promotion 2 구조가 모든 주차에서 일관됨

3주차는 프로모션 간 **격차가 가장 크게 벌어지는 시점**

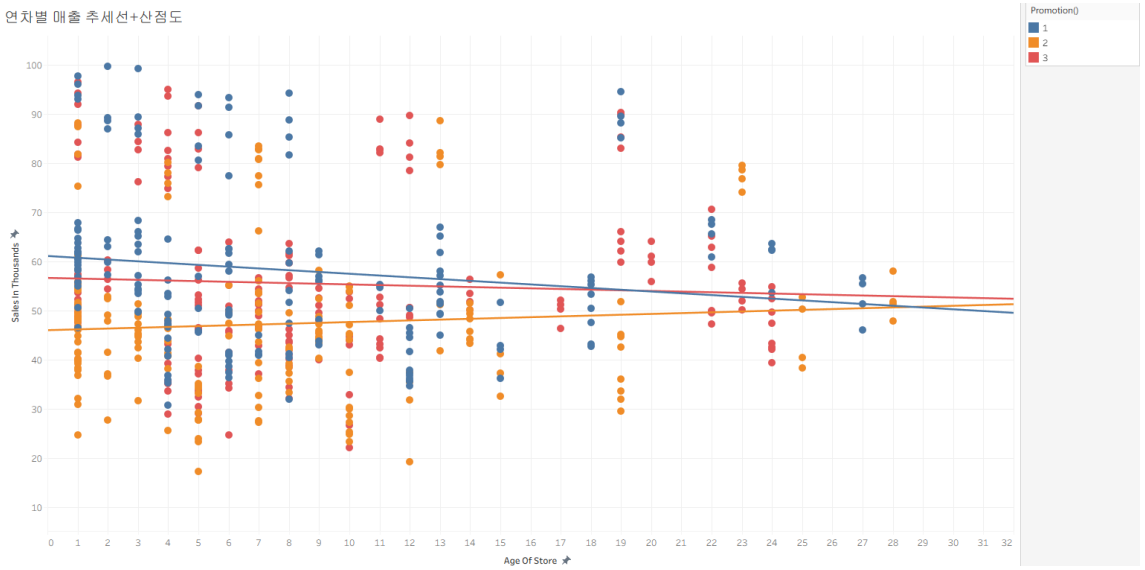
Promotion 2는 **개선 또는 중단 검토 대상**

Promotion 1은 **확대 적용 또는 벤치마킹 대상**

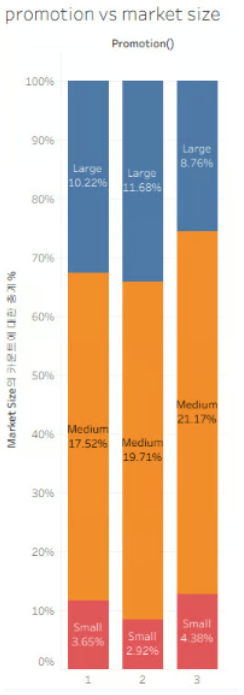
시트 9



연차별 매출 추세선+산점도

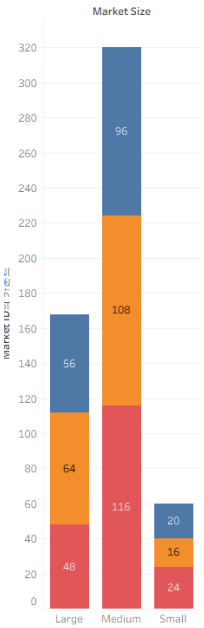


연차 구간화와 산점도+추세선을 통해 평균매출 비교, 프로모션 효과 매장 성숙도에 따라 다른 결과를 확인,
신규~ 중간 연차 매장에서는 Promotion1이 가장 안정적,
성숙 매장에서 Promotion3 이 상대적으로 높은 성과를 보이는 구간을 확인

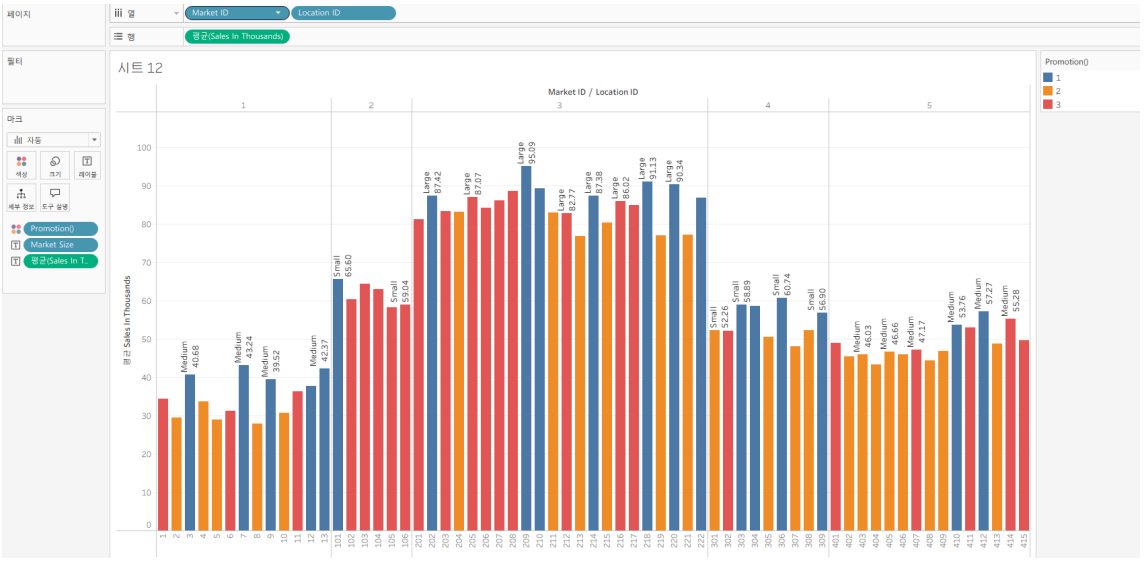


Medium 사이즈의 프로모션 비율이 너무 많음
분석 진행시 Market Size별로 통제해 분석하거나

회귀분석, ANCOVA 등 Market Size를 **통제 변수**로 포함



프로모션별 노출 개수는 유사하나, 전체 Market Size 중 Medium 시장의 절대 모수가 크기 때문에
랜덤 배정 이후에도 프로모션별 시장 규모 비율 차이가 발생하였다.



MarketID = 시장(ex.서울/대전/부산) LocationID = 매장(ex.서울점1,서울점2)

MarketID 1,5 = mediumsize 매장

MarketID 2,4 = smallsize 매장

MarketID 3 = Largesize 매장

▼ 유정하

▼ Day2

- 금일 목표
  - ☐ A/B 테스트의 결과 공유
  - ☐ 사후 분석 시작

▼ 송현준

▼ 박정우

```
df1["Promotion"]  = df1["Promotion"].astype("category")
df1["MarketSize"] = df1["MarketSize"].astype("category")
df1["LocationID"] = df1["LocationID"].astype("category")
df1["MarketID"]   = df1["MarketID"].astype("category")
df1["week"]       = df1["week"].astype("category")
# 프로모션,매장 규모 ->범주형
# LocationID → 랜덤효과로 사용할 그룹
# MarketID → 필요시 필터링 or 통제변수
# week → 연속값X 범주형 시간 변수로 처리


# Promotion 2가 존재하지 않는 MarketID가 있을 경우,
# 공정한 비교를 위해 해당 MarketID를 제외
# df = df[df["MarketID"] != "2"]

# - 고정효과:
#   · Promotion (A/B/C 테스트 핵심 변수)
#   · MarketSize (매장 규모 통제)
#   · AgeOfStore (매장 연차 통제)
#   · week (주차 효과 통제)
# - 랜덤효과:
#   · LocationID (매장별 고유 차이)

m2 = smf.mixedlm(
    "SalesInThousands ~ C(Promotion) + C(MarketSize) + AgeOfStore + C(week)",
    data=df,
    groups=df["LocationID"]
).fit(reml=True, method="lbfgs")
```

```
print("\n--- LMM (+ controls): Sales ~ Promotion + MarketSize + AgeOfStore + week ---")
print(m2.summary())
```

```
--- LMM (+ controls): Sales ~ Promotion + MarketSize + AgeOfStore + week ---
            Mixed Linear Model Regression Results
=================================================================
Model:               MixedLM   Dependent Variable:   SalesInThousands
No. Observations:    548       Method:               REML
No. Groups:          137       Scale:                26.6381
Min. group size:     4         Log-Likelihood:       -1851.8257
Max. group size:     4         Converged:            Yes
Mean group size:     4.0
-----------------------------------------------------------------
                      Coef.  Std.Err.    z    P>|z|  [0.025  0.975]
-----------------------------------------------------------------
Intercept             74.338    2.228  33.367 0.000  69.971  78.704
C(Promotion)[T.2]    -10.752    2.130  -5.047 0.000 -14.927  -6.577
C(Promotion)[T.3]     -1.074    2.135  -0.503 0.615  -5.259   3.111
C(MarketSize)[T.Medium] -26.633  1.938 -13.742 0.000 -30.431 -22.834
C(MarketSize)[T.Small]  -14.073  3.079  -4.571 0.000 -20.107  -8.038
C(week)[T.2]          -0.404    0.624  -0.648 0.517  -1.626   0.818
C(week)[T.3]          -0.316    0.624  -0.507 0.612  -1.538   0.906
C(week)[T.4]          -0.578    0.624  -0.926 0.354  -1.800   0.645
AgeOfStore             0.071    0.132   0.537 0.591  -0.188   0.330
Group Var             95.061    2.799
=================================================================
```

기준선(baseline)

- **Promotion 1**
- **MarketSize = Large**
- **week = 1**

Promotion 2

계수: **−10.75**

p-value: **< 0.001**

95% CI: **[−14.93, −6.58]**

Promotion 1 대비 평균 매출이 약 10.8 감소 (유의함)

명확하게 성과가 나쁨

Promotion 3

계수: **−1.07**

p-value: **0.615**

95% CI: **[−5.26, 3.11]**

Promotion 1과 **통계적으로 차이 없음**

평균적으로는 비슷하지만 **우월하다고 말할 수는 없음**

```
df1["Promotion"]  = df1["Promotion"].astype("category")
df1["MarketSize"] = df1["MarketSize"].astype("category")
df1["LocationID"] = df1["LocationID"].astype("category")
df1["MarketID"]   = df1["MarketID"].astype("category")
df1["week"]       = df1["week"].astype("category")
```

이 코드들의 의미

Pandas는 숫자 → 연속형 / 문자 →범주형으로 인식

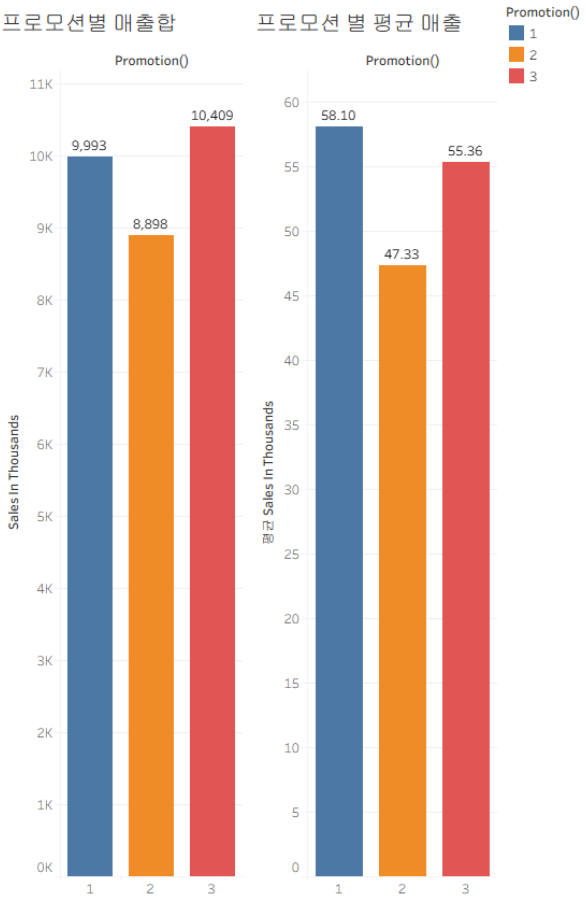
현재 데이터는 ex) Promotion = 숫자형(연속형)이므로

숫자형 → 범주형 으로 변경해서

모델에게 “순서·크기 없는 그룹”이라고 명확히 알려주는 작업

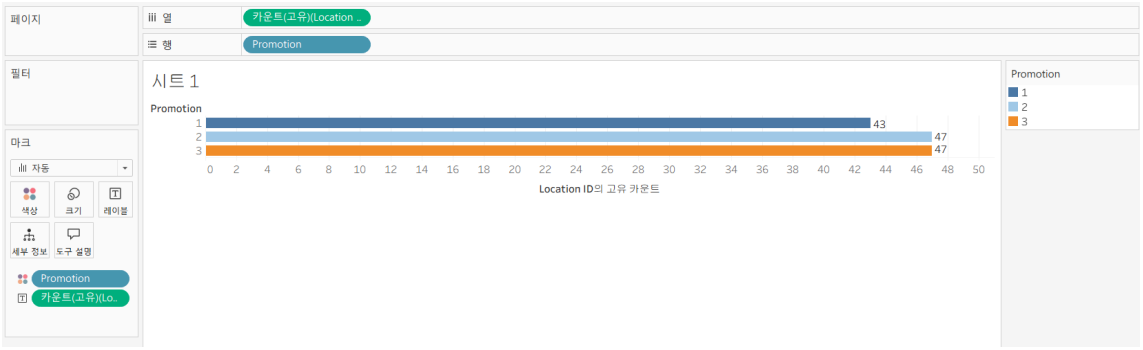
만약 변경없이 진행한다면

[Promotion]1 이 [Promotion] 3 보다 3배 약하다 라는 오해의 소지가 있음

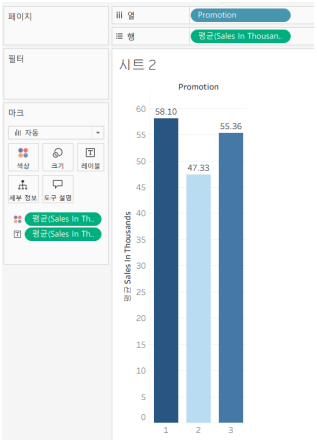


▼ 유정하

Promotion별 매장 수



Promotion별 평균 매출



Week별 매출 추이

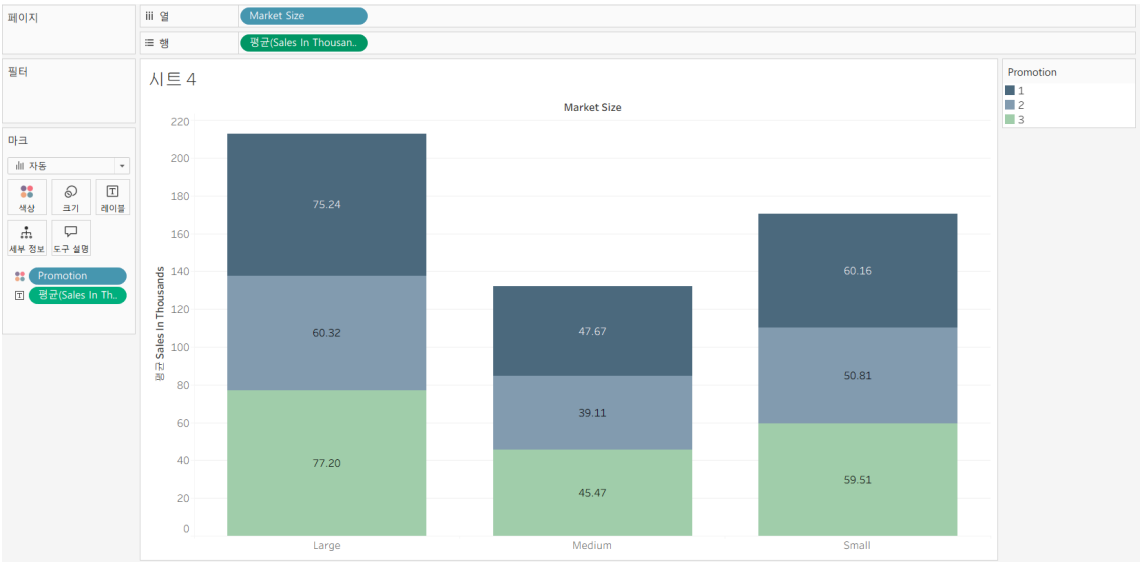
Promotion 1은 전 기간 내내 1등을 차지했다

Promotion 3은 항상 2등, Promotion 2는 항상 3등

주차별 변동은 있으나 추세 역전은 없음

MarketSize, Promotion, 매출

Large Market에서는 Promotion 3이 우세했고 Medium, Small Market에서는 Promotion 1이 우세했다

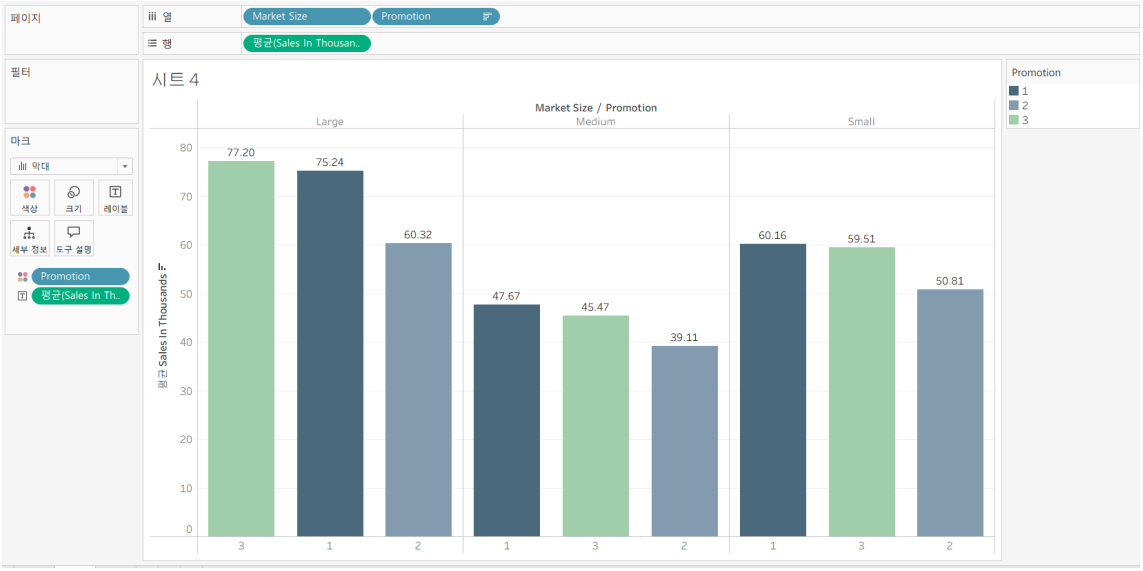


조금 더 잘 보이게 차트 변형

효과 크기 관점에서 Promotion 1이 1등이지만 Medium, Small Market에서는 Promotion 1(이하 P1)과 Promotion 3(P3) 차이가 작다

Medium, Small Market에서 둘 다 비슷하게 잘하는 것일 수 있음

특히 Small Market은 표본 수가 작아서 미세한 차이는 더더욱 신뢰하기 어려울 수 있음



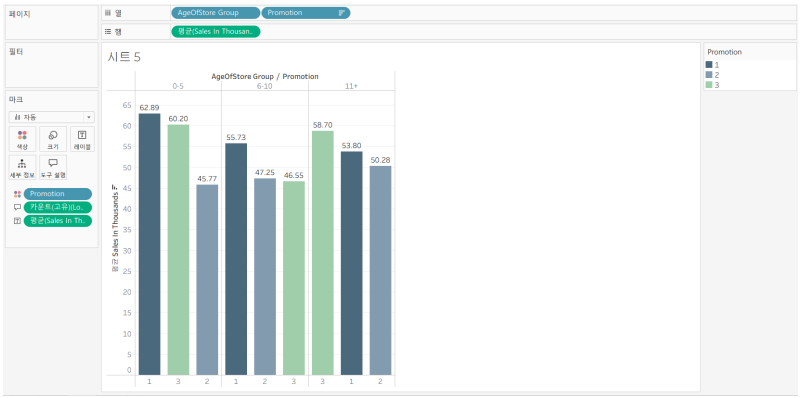
Age of Store, Promotion, 매출 - P1이 특정 연차에만 좋은 건 아닌지 확인

표본 수가 비슷하게 분포하게끔 집단을 나눔(0~5년차, 6~10년차, 11년차 이상)

Promotion 1의 효과가 특정 연차의 매장에만 국한된 건 아닌지 검증 (가드레일 + 보조 세그먼트 역할)

신규, 중간 매장에서는 P1이 가장 효과적이며 성숙 매장에서는 P3가 상대적으로 우수하다

Promotion	**1**	**2**	**3**
AgeBucket			
0-5	18	20	17
6-10	11	16	15
11+	14	11	15



MarketSize별 평균 매출이 가장 높은 Promotion

MarketSize	Promotion	SalesInThousands
2	Large	3
3	Medium	1
6	Small	1

MarketSize × Promotion별 표본 수(n)

Small 구간은 4~6개로 표본이 작아 해석에 주의가 필요

Promotion	1	2	3
MarketSize			
Large	14	16	12
Medium	24	27	29
Small	5	4	6

심슨의 역설?

전형적인 역전은 아니다. Large Market에서 Promotion 3으로 나온 (세그먼트별) 차이는 존재한다.

시장 규모에 따라 최적 프로모션이 다를 수 있음을 보여준다.

	SalesInThousands
Promotion	
1	58.099012
3	55.364468
2	47.329415

가드레일 지표 - 성과 변동성(표준편차) 점검

평균 매출은 Promotion 1이 제일 크다.

변동성(std)의 경우 Promotion 1과 3이 비슷하고 Promotion 2가 가장 낮다.

즉 Promotion 1이 유난히 변동성만 폭증하는 경우는 아님을 보여준다.

	mean	std	count
Promotion			
1	58.099012	16.028223	43
2	47.329415	14.497097	47
3	55.364468	16.383866	47

가드레일 2 - 각 Promotion의 매출 분포를 퍼센타일로 나눠서 점검

하위권, 중앙, 상위권 전부에서 Promotion 1이 가장 높음

(단 75%는 P3와 P1이 비슷)

	0.10	0.25	0.50	0.75	0.90
Promotion					
1	40.388	48.69625	56.0750	61.32500	87.4070
2	30.217	37.43375	46.6550	49.94375	76.9010
3	37.120	45.36750	49.8525	60.24500	84.4475

2단계: 전략 수립 및 액션 아이템 도출

- 전체 평균 매출 - P1(58.10) > P3(55.36) > P2(47.33)
- week별 - 순위가 바뀌지 않고 P1이 계속 우세
- MarketSize - **Large에서는 P3가 약간 우세**, Medium/Small은 **P1과 P3 차이 미묘**, P2는 열세
- 가드레일(std, 분위수) - **P1**은 **하위에서도 가장 높아서** "상위 몇 개만 좋아진" 패턴이 아님

의사결정(Decision) - 최종적으로

기본 프로모션은 Promotion 1을 채택하는 것으로.

전체 평균 매출이 가장 높고 week별로 살펴봤을 때 성과에 우위가 있으며 하위 매장에서의 성과도 가장 높아

전반적으로 안정적으로 우수한 것으로 확인되었다.

반면에 Promotion 2는 전 구간에서 성과가 잘 나오지 않음이 뚜렷하여 적용 대상에서 일단 제외한다.

실행 전략, 액션 아이템 계속해서 추가 서술할 예정

통계 검정 돌려봤는데 내일 결과 첨부하겠습니다

store_df라는 테이블 생성, locationID 단위로 집계 - 일단 one-way anova 해봄

한 매장은 딱 한 번만 등장해서 반복측정 문제가 크게 완화되고 관측 단위가 '매장'이 되니까 독립성 가정에서 문제가 안 된다고 판단함

```
store_df = (df.groupby(["LocationID", "Promotion"], as_index=False)
            .agg(mean_sales=("SalesInThousands","mean")))
```

이런 식으로 매장 단위로 집계를 했다

튜키까지 돌려봤을 때 결과는 다음과 같았다

```
n (stores) by Promotion: 43 47 47
means: 58.09901162790697 47.32941489361702 55.364468085106374

[One-way ANOVA]
F-statistic = 5.845791931827379
p-value     = 0.0036805468987114 36

[Effect size]
eta^2 (η²)  = 0.08024886238175773

[Tukey HSD post-hoc]
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower    upper  reject
-----------------------------------------------------
     1      2 -10.7696  0.004 -18.5951  -2.944   True
     1      3  -2.7345 0.6862 -10.5601   5.091  False
     2      3   8.0351 0.0371   0.3854 15.6847   True
-----------------------------------------------------
```

그룹 2의 평균 매출이 눈에 띄게 낮았다

프로모션 유형에 따라 평균 매출 차이가 통계적으로 유의한 것으로 나왔다 (f통계량이랑 p-value 봤을 때)

효과 크기는 크지 않음

튜키 사후검정에서 1과 2, 2와 3간 차이가 유의한 것으로 나왔고 1과 3은 유의하지 않은 것으로 나옴

우려되는 점으로 인해 두 번째로 이원 분산분석을 해 봄 (Promotion × MarketSize)

```
store_df = (df.groupby(["LocationID","Promotion","MarketID","MarketSize","AgeOfStore"], as_index=False)
            .agg(mean_sales=("SalesInThousands","mean")))
```

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

# 이원 ANOVA (주효과 + 상호작용)
model = ols("mean_sales ~ C(Promotion) * C(MarketSize)", data=store_df).fit()

anova2 = sm.stats.anova_lm(model, typ=2)
anova2
```

프로모션과 시장규모의 주효과 및 상호작용 효과에 대한 two-way anova

종속변수는 mean_sales(매장별 평균 매출), 첫 번째 요인은 Promotion, 두 번째 요인은 MarketSize, 그리고 Promotion과 MarketSize의 상호작용까지.

불균형 데이터로 판단했을 때 각 요인의 주효과를 다른 main effect를 통제한 상태에서 검정.

결과는 다음과 같았다.

	sum_sq	df	F	PR(>F)
C(Promotion)	3244.817690	2.0	16.191861	5.384924e-07
C(MarketSize)	19450.760385	2.0	97.060618	2.229458e-26
C(Promotion):C(MarketSize)	529.184907	4.0	1.320334	2.658987e-01
Residual	12825.476387	128.0	NaN	NaN

첫 번째 요인 Promotion의 경우 유의미하게 나왔다. (프로모션 유형에 따라 평균 매출 차이가 있음)

두 번째 요인 MarketSize의 경우 아주 강하게 유의미하게 나왔다. (시장 규모가 매출에 큰 영향을 줌)

Promotion과 MarketSize의 상호작용은 유의미하지 않은 것으로 나타났다. (프로모션 효과는 시장 규모에 따라 달라지지 않으며 이는 어떤 MarketSize에서든 Promotion 2가 다른 프로모션들에 비해 상대적으로 성과가 낮은 패턴이 유지되는 것이 맞다는 걸 보여줌)

Promotion X AgeOfStore 상호작용을 포함한 ANCOVA 결과

mean_sales ~ Promotion + AgeOfStore + Promotion × AgeOfStore

	sum_sq	df	F	PR(>F)
C(Promotion)	2892.351357	2.0	5.834307	0.003739
AgeOfStore	61.312722	1.0	0.247354	0.619778
C(Promotion):AgeOfStore	272.557227	2.0	0.549789	0.578397
Residual	32471.551730	131.0	NaN	NaN

매장 연령 자체는 평균 매출에 유의미한 선형 효과를 보이지 않는다

promotion과 ageofstore 간 상호작용 없음 (프로모션의 효과는 매장 나이에 따라 달라지지 않는다. 즉, 신규 매장이든 오래된 매장이든 프로모션 반응은 비슷)

혼합효과 선형회귀모형(Mixed Linear Model)

```
          Mixed Linear Model Regression Results
========================================================================
Model:             MixedLM  Dependent Variable:  SalesInThousands
No. Observations:  548      Method:              REML
No. Groups:        137      Scale:               26.6381
Min. group size:   4        Log-Likelihood:      -1851.8257
Max. group size:   4        Converged:           Yes
Mean group size:   4.0
------------------------------------------------------------------------
                    Coef.  Std.Err.   z    P>|z|  [0.025  0.975]
------------------------------------------------------------------------
Intercept           74.338    2.228  33.367 0.000  69.971  78.704
C(Promotion)[T.2]  -10.752    2.130  -5.047 0.000 -14.927  -6.577
C(Promotion)[T.3]   -1.074    2.135  -0.503 0.615  -5.259   3.111
C(MarketSize)[T.Medium] -26.633    1.938 -13.742 0.000 -30.431 -22.834
C(MarketSize)[T.Small]  -14.073    3.079  -4.571 0.000 -20.107  -8.039
C(week)[T.2]        -0.404    0.624  -0.648 0.517  -1.626   0.818
C(week)[T.3]        -0.316    0.624  -0.507 0.612  -1.538   0.906
C(week)[T.4]        -0.578    0.624  -0.926 0.354  -1.800   0.645
AgeOfStore           0.071    0.132   0.537 0.591  -0.188   0.330
Group Var           95.061    2.799
========================================================================
```

▼ 현재까지의 정리

목차

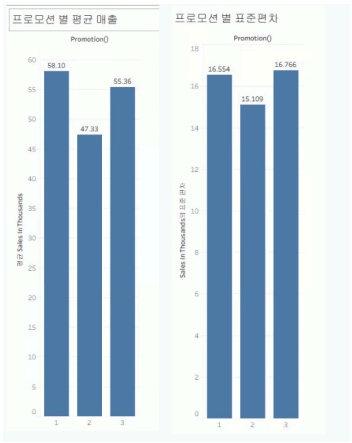
1. 배경
2. 데이터 Overview
3. A/B/C 테스트 결과
4. 가설 설정
5. 실험 설계
6. 분석
7. 결론

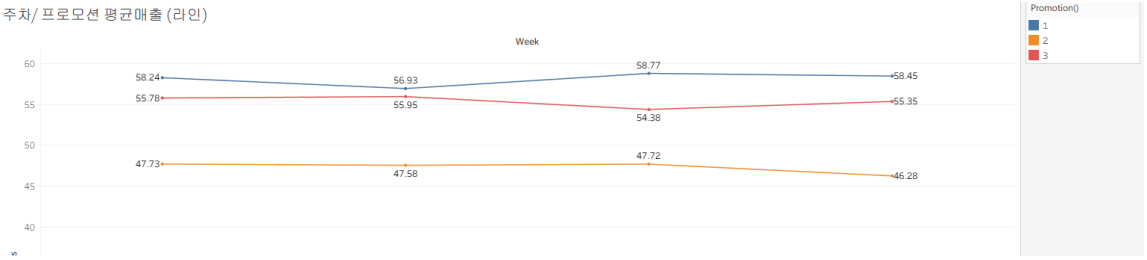
---

1. 배경
   - 데이터셋
     - 한 패스트푸드 체인이 메뉴에 새로운 품목을 추가할 계획입니다. 하지만 신제품 홍보를 위한 세 가지 마케팅 캠페인 중에서 어떤 것이 가장 효과적인지 아직 결정하지 못했습니다. 어떤 프로모션이 매출에 가장 큰 영향을 미치는지 알아보기 위해, 무작위로 선정된 여러 지역의 매장에 신제품을 도입했습니다. 각 매장마다 다른 프로모션을 적용하고, 첫 4주 동안 신제품의 주간 판매량을 기록했습니다.

- 목표
  - A/B 테스트 결과를 평가하고 어떤 마케팅 전략이 가장 효과적인지 결정합니다. 또한, 사후분석을 통해 각 Segment별 마케팅 전략이 어떻게 다른지 확인합니다.
- 접근
  - 3가지 실험군(Promotion 1, 2, 3)에 대하여 각각의 정규성, 등분산성, 독립성을 확인하고 이에 적합한 분석 방법을 진행합니다.

2. 데이터 Overview
   - 그룹(실험군 구분):
     - Promotion `1`
     - Promotion `2`
     - Promotion `3`
   - 성과지표(KPI):
     - `SalesInThousands` (매출)
   - 분석 단위 식별자:
     - `LocationID` (매장 식별 코드, 총 137개의 매장)
   - 기간/세그먼트:
     - `week`, `MarketSize` (Large, Medium, Small), `MarketID` (1~10, 총 10개의 시장), `AgeOfStore`
3. A/B/C 테스트 결과





4. 가설 설정
   - **귀무가설(H0):** "프로모션 종류에 따른 매출 차이는 없으며, 관측된 차이는 우연이다.
   - **대립가설(H1):** "특정 프로모션(특히 1번)은 다른 프로모션에 비해 통계적으로 유의미하게 높은 매출을 발생시킨다."

5. **실험 설계:** 131개 매장을 대상으로 4주간 진행, LMM 모델을 사용하여 매장별 변동성 통제
   a. 성공 지표 : 매장당 주평균 매출액

      실험군을 나누는 기준인 Promotion이 가장 영향을 미치는 것은 판매량, 즉 매출액입니다.
      Promotion이 성공적이라면 높은 판매량을 기록하여 상승한 판매량을 가질 것이며, 효과가 좋지 못하다면 유지 혹은 하락의 판매량을 기록하여 매출이 상승하지 않을 것입니다.

   b. 가드레일 지표

i. 하위 매장 피해

- 전체적으로 판매량은 증가하였지만, 하위 순위의 매장에게 판매량 저하 피해가 간다면 적절하지 않은 마케팅일 것입니다.

ii. 매출량 변동성

- 매출이 늘었으나 변동성이 커지는 경우, 이를 위한 더 큰 비용(소비자 신뢰도, 안전 재고 등)이 지불될 가능성이 있습니다.

iii. (세그먼트) 시장 편향

- 특정 시장에서만 작동하는 마케팅이 모든 시장에서 효과가 있지는 않습니다.
- 이는 사후분석에서 세그먼트를 진행하여 조금 더 알아보고자 합니다.

c. 서포트 지표

i. 매출 중앙값 및 표준편차

ii. `MarketSize` , `MarketID` , `AgeOfStore` 별 매출 변화량

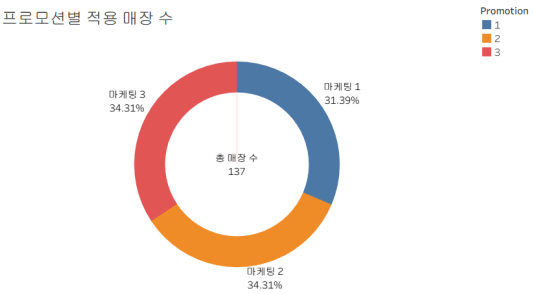
iii. `Week` 변화에 따른 매출 변화

- 초반 n 주에만 증가하는 효과 or 마지막 주에 증가하는 효과 등

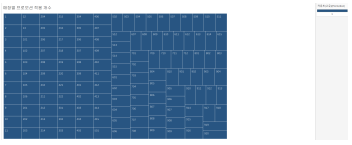
3. 분석

a. 데이터 정합성 확보

i. **LocationID (매장별)**

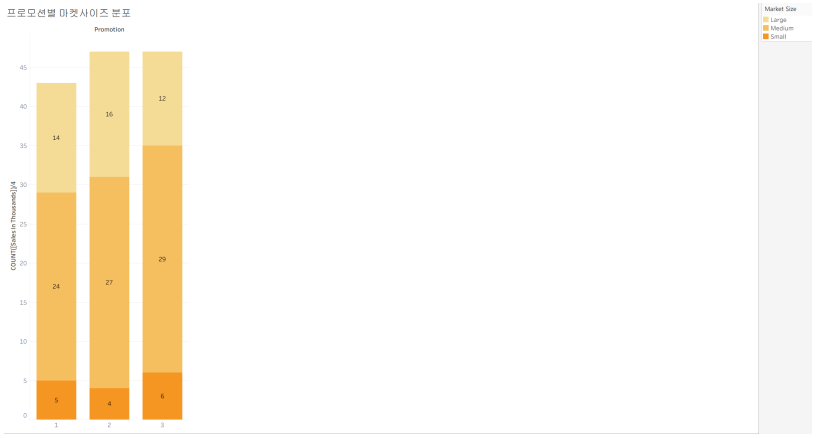


적용된 매장 수는 적절하게 1:1:1의 비율에 가깝게 설계되었다.



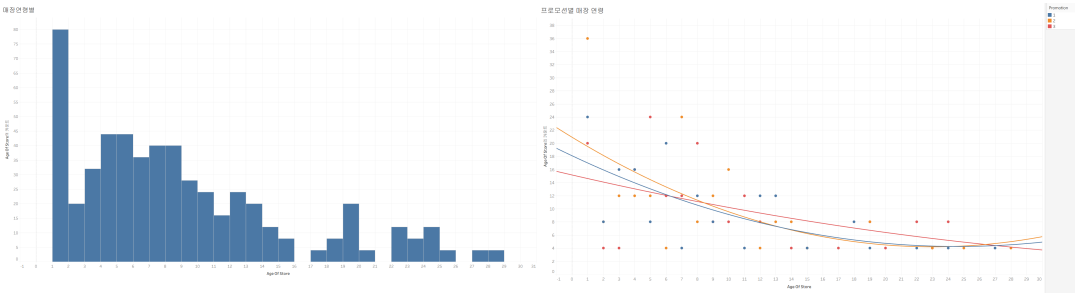
하나의 매장은 하나의 프로모션만을 진행하였다.

ii. **MarketSize (시장 크기별)**



각 시장에 동일한 프로모션 비율이 적용되었다.

iii. **AgeOfStore (매장 연령별)**

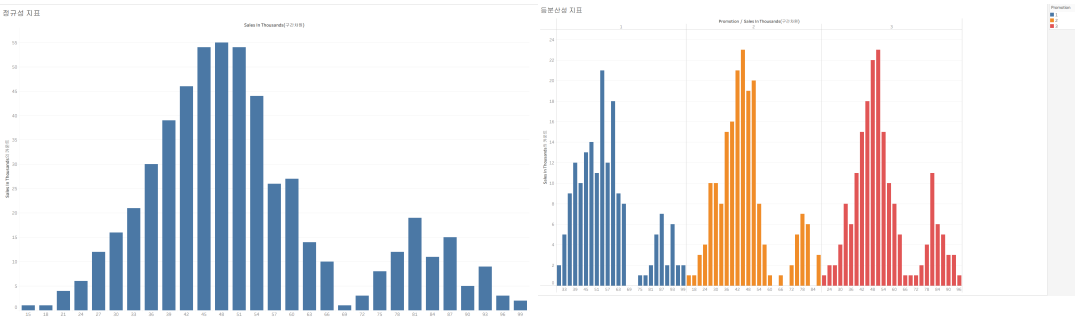


각 프로모션은 비슷한 매장 연령에서 시행되었다.

위의 지표들은 각 실험군(프로모션별)이 균등하게 분포되어 있는 것을 확인할 수 있습니다.

b. ANOVA 분석을 위한 사전 검증

i. 정규성 & 등분산성



- 정규성

```
groups = [
    df.loc[df['Promotion'] == 1, 'SalesInThousands'],
    df.loc[df['Promotion'] == 2, 'SalesInThousands'],
    df.loc[df['Promotion'] == 3, 'SalesInThousands']
]

for i, group in enumerate(groups):
    print(f'Promotion {i+1}')
    stat, p = shapiro(group)
    print(f'stat: {stat:0.4f}\np: {p:0.4f}')
    if p > 0.05:
        print('정규성 만족\n')
    else:
        print('정규성 만족하지 않음\n')

# stat이 1에 가까울수록 정규성 만족
# p-value가 0.05보다 작으므로 귀무가설 기각 -> 정규성 만족하지 않음
```

✓ 0.0s

```
Promotion 1
stat: 0.9153
p: 0.0000
정규성 만족하지 않음

Promotion 2
stat: 0.9145
p: 0.0000
정규성 만족하지 않음

Promotion 3
stat: 0.9208
p: 0.0000
정규성 만족하지 않음
```

- 등분산성

```
groups = [
    df.loc[df['Promotion'] == 1, 'SalesInThousands'],
    df.loc[df['Promotion'] == 2, 'SalesInThousands'],
    df.loc[df['Promotion'] == 3, 'SalesInThousands']
]

stat, p = levene(*groups)
print(f'stat: {stat:0.4f}\np: {p:0.4f}')
if p > 0.05:
    print('등분산성 만족')
else:
    print('등분산성 만족하지 않음')
# p-value가 0.05보다 크므로 귀무가설 채택 -> 등분산성 만족
```

✓ 0.0s

```
stat: 1.2697
p: 0.2818
등분산성 만족
```

ii. 독립성

- 데이터 확인 결과
  하나의 MarketID 내에 여러 개의 LocationID가 존재하고,
  각 LocationID마다 week 1,2,3,4 가 존재
- 데이터 계층 구조 정리
  1. `MarketID` , `MarketSize`
  2. `LocationID` , `AgeOfStore` , `Promotion`
  3. `Week` , `SalesInThousands`
- 데이터 시계열 정보
  1. `Week` : 1~4 주차
- 통계에서 독립성이란 **"하나의 관측값이 다른 관측값에 영향을 주지 않아야 함"**을 의미합니다. 하지만 질문하신 데이터의 구조는 다음과 같습니다.
  - **LocationID 간의 독립성:** `MarketID` , `MarketSize` 의 하위 계층이기 때문에 시장의 매출 규모에 따라 매출이 달라질 수 있기 때문에 독립적이지 않을 가능성이 높습니다. (아래와 비슷한 이유)
  - **주차(Week) 간의 독립성 (문제 발생):** 매장 A의 1주차 매출, 2주차 매출, 3주차 매출, 4주차 매출은 **모두 같은 매장에서 발생한 데이터**입니다.
    - 매장 A가 원래 장사가 잘되는 목 좋은 곳에 있다면, 1~4주차 매출이 모두 높게 나올 것입니다.
    - 반대로 매장 B가 장사가 안되는 곳이라면 1~4주차 모두 낮게 나오겠죠.
    - 즉, **'LocationID'라는 공통된 요인** 때문에 1주차 데이터가 2~4주차 데이터와 상관관계(Correlation)를 갖게 되어 독립성이 깨집니다.
- 따라서, ANOVA 분석을 하기 위해서는 이러한 문제를 해결해야 합니다.

c. 모델 선정 및 코드

▼ 선형 혼합 효과 모델(Linear Mixed-Effects Model, LMM)

1. 시계열 + 계층 구조를 동시에 잡는 모델 설계

   LMM에서는 변수를 두 가지 성격으로 나누어 배치함으로써 모든 문제를 해결합니다.

   1. 고정 효과 (Fixed Effects): "우리가 알고 싶은 공통 추세"
      - **Promotion (A/B/C):** 프로모션 간의 순수 실력 차이.
      - **Week (시계열):** 시간이 흐름에 따라 매출이 전반적으로 상승/하락하는 경향.
      - **Promotion * Week (상호작용):** "특정 프로모션이 시간이 갈수록 더 힘을 쓰는가?"를 확인 (시계열적 성과 비교).
   2. 변량 효과 (Random Effects): "데이터가 묶여 있어서 생기는 층간 소음"
      - **MarketID / MarketSize (계층):** "이 마켓은 원래 체급이 커"라는 지역적 특성 보정.
      - **LocationID (계층 + 독립성):** "이 매장은 1~4주 내내 비슷하게 나와"라는 매장별 특성 보정.
2. 수학적 구조 (직관적 이해)
   이 모델은 내부적으로 다음과 같은 계산을 동시에 수행합니다.
   Sales = [{기본 매출} + {프로모션 효과} + {주차별 효과}] + [{마켓별 오차} + {매장별 오차}] + {잔차}
   - **앞부분(고정 효과):** 모든 매장에 공통으로 적용되는 '법칙'을 찾습니다.
   - **뒷부분(변량 효과):** 독립성을 깨뜨리는 계층적 '특성'을 분리해 냅니다. 이 과정에서 **1~4주 데이터가 한 매장(LocationID)에 묶여 있다는 사실**을 모델이 인지하므로 독립성 문제가 해결됩니다.

```
import statsmodels.formula.api as smf

# LMM 모델 설계
# 1. 고정 효과: Promotion, MarketSize, Week, AgeOfStore (그리고 Promotion과 Week의 상호작용)
# 2. 변량 효과 (groups): LocationID (Location은 MarketID에 포함되므로 보통 하위 단위를 그룹으로 잡음)

model = smf.mixedlm("SalesInThousands ~ C(Promotion) * C(MarketSize) + AgeOfStore + Week",
                    data=df,
                    groups=df["LocationID"])

result = model.fit()
print(result.summary())
```

- `C(Promotion) * C(MarketSize)` : 프로모션별 매출 차이와 시장 크기에 따른 변화(계층), 그리고 그 둘의 시너지(상호작용)를 모두 봅니다.
- `week` : 시간에 따른 변화(시계열)을 확인합니다.
- `groups=df["LocationID"]` : "LocationID가 같은 데이터들은 서로 독립이 아니니 알아서 감안해서 계산해!"라고 명령하는 핵심 줄입니다.

▼ 모델 공부를 위한 질문들

▼ 변수 입력의 순서

1. `mixedlm` 모델은 **최대우도법(Maximum Likelihood)**이라는 수치 최적화 방식을 사용한다. 이는 모든 변수를 한꺼번에 넣고 전체 오차가 가장 적은 최적의 값을 동시에 찾아내는 방식.
   따라서 순서 상관 없음
2. 반면, 일반 ANOVA (OLS 기반)의 경우
   - **Type I (Sequential):** 먼저 입력된 변수가 설명할 수 있는 변동성을 다 가져가고, 남은 것을 그다음 변수가 가져갑니다. (순서에 따라 결과가 달라짐)
   - **Type III (Marginal):** 다른 모든 변수가 모델에 들어있을 때 해당 변수가 추가로 설명하는 변동성을 봅니다. (순서에 상관없음, 일반적으로 더 권장됨)
3. 상호작용항(*)의 위치

   `C(Promotion) * Week`는 내부적으로 `Promotion + Week + Promotion:Week`로 확장됩니다. 이 덩어리가 모델 식의 앞에 있든 뒤에 있든 상관없지만, 가급적 **주요 관심 변수(Promotion)를 앞에 적는 것이 코드를 읽는 사람(혹은 면접관)에게 분석의 목적을 명확히 전달**하는 관례

▼ C(변수)의 의미

`C(Promotion)`이라고 적는 순간, 파이썬의 `statsmodels`는 이 변수를 숫자가 아닌 **'범주형(Categorical) 변수'**로 인식

숫자의 차이가 크기가 아닌 집단의 구분이라고 판단하고, 각 그룹 간의 평균 매출 차이를 계산하기 시작

▼ 결과 비교의 기준점

통계 모델은 비교를 위해 반드시 **'기준점(Reference Level)'**이 필요합니다. 이를 **더미 코딩(Dummy Coding)**이라고 하는데, 별도의 설정을 하지 않으면 기본 규칙은 다음과 같습니다.

- **숫자형 범주:** 가장 작은 숫자 (1, 2, 3 중 **1**)
- **문자형 범주:** 알파벳 순서 (Large, Medium, Small 중 **L**이 가장 빠름)
- 나머지 변수들을 수치형 변수이기 때문에 기준값이 0 일 때

따라서 **Intercept(절편)**는 **"모든 조건이 기본일 때의 예상 매출"**을 의미하며, 현재 데이터에서는 **[Promotion 1]이면서 [Large Market]인 경우**가 그 기준점

현재 기준점을 확인하는 법은 결과표에 나타나있지 않은 값이 기준

기준 설정 방법 (Reference 변경)

특정 그룹(예: 프로모션 3)을 기준으로 삼고 싶다면 `Treatment` 함수를 사용합니다.

```
# Promotion 3을 기준으로 설정
model = smf.mixedlm("SalesInThousands ~ C(Promotion, Treatment(reference=3)) + ...", ...)
```

▼ `*` 의 역할

- 연산자는 통계 식에서 **"각각의 독립적인 효과(Main Effect)와 둘 사이의 시너지(Interaction)를 모두 계산하라"**는 단축 명령어입니다.
- `A * B` **의 실제 의미:** `A + B + (A:B)`
- **질문에 대한 답:** `C(Promotion) * C(MarketSize) + week`라고 쓰시면, `Promotion`과 `MarketSize` 사이의 상호작용은 계산하지만, `week`**와의 상호작용은 계산하지 않습니다.**
- **만약 모두의 상호작용을 보고 싶다면:** `C(Promotion) * C(MarketSize) * week`라고 쓰시면 됩니다. (이 경우 3차 상호작용까지 계산됩니다.)

▼ MarketSize와 MarketID: 무엇을 선택해야 할까?

이 두 변수는 **강한 상관관계(Multicollinearity)**가 있을 가능성이 높습니다. (하나의 MarketID는 보통 하나의 MarketSize에만 속하기 때문입니다.)

- **문제점:** 만약 `MarketID`를 고정 효과(Fixed Effect)로 넣으면, 모델은 지역 하나하나의 특성을 다 계산하느라 `MarketSize`(대/중/소)가 주는 정보의 의미를 잃어버리게 됩니다.

- **해결책:** * **일반적인 경우:** `MarketSize`를 고정 효과(`C(MarketSize)`)로 넣어 "체급 차이"를 보고, `MarketID`는 변량 효과(`groups`)나 `LocationID`에 녹여내는 것이 분석 결과 해석에 훨씬 유리합니다.
  - **MarketID 자체가 중요하다면:** 지역별로 구체적인 원인을 파악해야 하는 상황일 때만 넣습니다.

▼ 3중 상호작용 가능? 언제 효율적이고 비효율적일까?

1. `C(Promotion) * C(MarketSize) * week`처럼 세 변수를 모두 `*`로 연결할 수 있습니다.

- **계산 내용:** 1. 각 변수의 개별 효과
  2. 두 변수끼리의 조합 (Promotion:Size, Promotion:week, Size:week)
  3. 세 변수의 동시 조합 (Promotion:Size:week)
- **주의할 점 (해석의 난이도):** 3중 상호작용은 해석이 매우 어렵습니다. "대형 시장에서 프로모션 1번을 했을 때, 시간이 흐름에 따라 나타나는 매출 상승 폭이 다른 조합과 비교해 유의미하게 다른가?"를 묻는 것과 같습니다. 데이터 양이 충분하지 않으면 결과가 불안정해질 수 있습니다.
- `A * B * C`: A + B + C + (A * B) + (A * C) + (B * C) + (A * B * C) 를 의미합니다. 모든 가능한 조합을 한 번에 계산합니다.

2. **3중 상호작용항(A * B *C)이 필요한 경우**
   만약 `week`가 매우 중요하고, **"시장 규모에 따라 주차별로 프로모션의 성과가 다르게 변하는지"**를 한 번에 보고 싶다면 3중 상호작용항을 사용하는 것이 맞습니다.
   • **3중 상호작용의 의미:** "대형 시장에서 프로모션 1번을 했을 때, 시간이 흐름에 따라 나타나는 성과가 다른 조합(소형 시장+프로모션 2번+시간 흐름)과 비교해 **독보적으로 다른 패턴**이 있는가?"를 묻는 것입니다.
   • **주의점:** 3중 상호작용은 해석이 매우 어렵습니다. 변수가 3개나 얽히면 결과표의 수치만 보고 "그래서 어디에 뭘 써야 해?"라는 결론을 내리기 위해 매우 복잡한 그래프 분석이 동반되어야 합니다.

3. **변수를 바꿔가며 분석해야 하는 이유 (전략적 접근)**
   모든 변수를 `*`로 묶어서 한 번에 분석하지 않고, 변수를 바꿔가며 여러 번 분석(혹은 모델을 다듬는 과정)을 하는 데에는 이유가 있습니다.
   ① 해석의 명확성 (Ockham's Razor)
   통계 모델은 **'최대한 단순하면서 설명력이 높은 모델'**이 좋은 모델입니다.
   • 만약 `Promotion * MarketSize`만으로도 충분히 어느 시장에 어떤 프로모션이 좋은지 설명이 된다면, 굳이 복잡한 `week`까지 섞을 필요가 없습니다.
   • 불필요하게 변수를 꼬아놓으면 유의미했던 결과가 통계적 노이즈에 묻혀버릴 수도 있습니다.
   ② 분석 목적에 따른 모델 선택
   분석 목적에 따라 다음과 같이 모델을 바꿔가며 결과를 뽑는 것이 더 효율적입니다.

분석 목적	추천 식 (Formula)
최적의 시장별 프로모션 찾기	`Sales ~ C(Promotion) * C(MarketSize) + week`
주차별 프로모션의 성장세 비교	`Sales ~ C(Promotion) * week + C(MarketSize)`
전체 변수의 복합 시너지 확인	`Sales ~ C(Promotion) * C(MarketSize) * week`

▼ `LocationID`를 변수(Fixed Effect)로 넣으면 안 되는 이유

`LocationID`는 이 데이터셋에서 137개나 되는 매우 많은 수준(Level)을 가지고 있습니다.

- **자유도(Degree of Freedom) 문제:** `C(LocationID)`를 변수로 넣으면 모델은 136개의 추가 계수를 계산해야 합니다. 이는 모델을 너무 복잡하게 만들어 **과적합(Overfitting)**을 유발하고, 정작 중요한 `Promotion` 효과를 찾아내기 어렵게 만듭니다.
- **일반화 불가능:** 우리가 궁금한 것은 "10번 매장에서의 결과"가 아니라 "프로모션의 일반적인 효과"입니다. `LocationID`는 개별 매장의 특수성일 뿐이므로, 여전히 **`groups`(변량 효과)**에 두는 것이 맞습니다.

만약 상위 계층인 `MarketID`를 기준으로 알고 싶다면 `groups : df['MarketID']`를 기입

▼ 결과표에서 서로 다른 범주의 값을 비교하려면

1. 직접 계산

   현재 표의 계산 방식: 단순히 각 변수의 효과를 더합니다.
   • **예상 매출** = Intercept(73.40) + Promotion 2 효과(-9.08) + Medium 효과(-26.63) ~ 37.69

2. 그래프

   - **상호작용 그래프 (Interaction Plot):** `pointplot` 같은 그림이 사실은 모델이 계산한 값을 시각화한 것입니다.
   - **EMM (Estimated Marginal Means):** 파이썬의 `emmeans` 같은 라이브러리를 쓰면, "프로모션 2 + 2주차"의 최종 예측값을 모델이 알아서 계산해서 표로 보여줍니다.

▼ `C(week)`와 `week`의 차이

`week`를 수치형(`week`)으로 넣는 것과 범주형(`C(week)`)으로 넣는 것은 데이터의 **시간 흐름을 어떻게 해석하느냐**에 대한 근본적인 차이를 만듭니다.

구분	수치형 (`week`)	범주형 (`C(week)`)
가정	**선형성(Linearity)**: 주차가 지날수록 일정하게 증가하거나 감소함.	**독립적 효과**: 각 주차(1, 2, 3, 4주)는 서로 상관없는 별개의 시점임.
결과값 (Coef)	**기울기(Slope)**: "1주가 지날 때마다 평균적으로 몇 달러가 변하는가?"	**편차(Difference)**: "1주차 대비 2주차는 어떻고, 3주차는 어떤가?"
주요 목적	**추세(Trend) 파악**: 지속적인 상승/하락 흐름을 보고 싶을 때.	**시점별 특성 파악**: 특정 주차에만 매출이 튀거나 빠지는지 보고 싶을 때.

- **수치형(`week`) 추천:**
  - 프로모션의 **지속성이나 트렌드**를 보고 싶을 때.
  - "이번 프로모션은 매주 평균 2%씩 매출을 견인한다"라는 **예측 수식**을 만들고 싶을 때.
  - 현재 데이터처럼 기간이 4주로 짧고, 매주 특별한 이벤트가 없는 평이한 흐름일 때.
- **범주형(`C(week)`) 추천:**
  - 특정 주차에 **명절이나 공휴일**이 포함되어 있어 매출이 튈 것이 예상될 때.
  - 프로모션 효과가 **비선형적**일 때 (예: 1주차엔 반응 없다가 2주차에 폭발하고 4주차에 급감하는 경우).
  - 주차별로 정책이나 운영 조건이 아예 달랐을 때.

▼ C(Promotion) * C(week)의 관계

분석의 설계는 `Promotion` 과 `week` 간의 상호작용이 있을 것이라 판단한 코드이다.

그러나, 아래의 분석 결과 p-value > 0.05 의 유의미한 결과가 나오지 않았기 때문에 상호작용항의 역할은 미비한다고 본다.

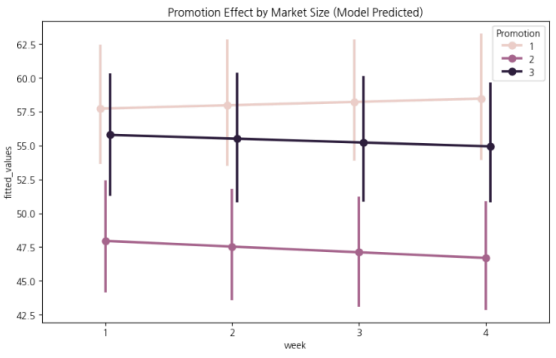
또한, 이렇게 효과가 미비할 경우에는 `*` 를 `+` 로 바꿔 결과값을 간단하게 하는 것이 좋다.

통계학에서는 이를 **'모델의 간명성(Parsimony)'** 원칙이라고 부른다.

d. 분석 결과

i. Promotion * week + AgeOfStore + MarketSize

```
Mixed Linear Model Regression Results
============================================================================
Model:              MixedLM   Dependent Variable:   SalesInThousands
No. Observations:   548       Method:               REML
No. Groups:         137       Scale:                26.5173
Min. group size:    4         Log-Likelihood:       -1852.3567
Max. group size:    4         Converged:            Yes
Mean group size:    4.0
----------------------------------------------------------------------------
                        Coef.  Std.Err.    z    P>|z|  [0.025  0.975]
----------------------------------------------------------------------------
Intercept               73.400    2.364  31.049 0.000  68.767  78.033
C(Promotion)[T.2]       -9.088    2.452  -3.706 0.000 -13.895  -4.281
C(Promotion)[T.3]        0.248    2.457   0.101 0.919  -4.566   5.063
C(MarketSize)[T.Medium] -26.633    1.938 -13.742 0.000 -30.431 -22.834
C(MarketSize)[T.Small]  -14.073    3.079  -4.571 0.000 -20.107  -8.039
week                      0.245    0.351   0.699 0.485  -0.443   0.934
C(Promotion)[T.2]:week   -0.666    0.486  -1.370 0.171  -1.618   0.287
C(Promotion)[T.3]:week   -0.529    0.486  -1.088 0.276  -1.481   0.424
AgeOfStore                0.071    0.132   0.537 0.591  -0.188   0.330
Group Var                95.091    2.805
============================================================================
```



▼ Gemini 해석

1. 핵심 요약: "승자는 프로모션 1과 3입니다."
통계적으로 분석했을 때, **프로모션 2는 다른 두 그룹에 비해 매출 성적이 현저히 낮으며**, 프로모션 1과 3 사이에는 유의미한 차이가 없습니다. 즉, 성과 면에서는 **1번 또는 3번을 선택하는 것이 합리적**입니다.
2. 상세 지표 해석 (Coefficients)
결과표의 `P>|z|` 값이 **0.05보다 작으면** 통계적으로 유의미한 영향이 있다고 해석합니다.
① 프로모션 효과 (Promotion) - 가장 중요
• **Intercept (73.400):** 모델의 기준점입니다. 여기서는 **프로모션 1 / 대형 시장(Large)**의 평균 매출이 약 73,400달러임을 의미합니다.
• **C(Promotion)[T.2] (-9.088, $p=0.000$):** 프로모션 1에 비해 **프로모션 2의 매출이 약 9,088달러 적습니다.** p값이 매우 낮으므로 이는 확실한 차이입니다.
• **C(Promotion)[T.3] (0.248, $p=0.919$):** 프로모션 1과 3의 차이는 거의 없으며(0.248), p값이 0.05보다 훨씬 크므로 **통계적으로 동일한 성과**로 간주합니다.
② 시장 규모 효과 (MarketSize)
• **C(MarketSize)[T.Medium] (-26.633, $p=0.000$):** 대형 시장에 비해 중형 시장의 매출이 약 26,633달러 낮습니다.
• **C(MarketSize)[T.Small] (-14.073, $p=0.000$):** 대형 시장에 비해 소형 시장의 매출이 약 14,073달러 낮습니다.
• **인사이트:** 시장 체급 차이가 매출에 매우 큰 영향을 주고 있음을 알 수 있습니다.
③ 시간 및 상호작용 (Week & Interaction)
• **week (0.245, $p=0.485$):** 시간이 흐름에 따라 전체 매출이 상승하는 경향은 뚜렷하지 않습니다.
• **C(Promotion)[T.2]:week / [T.3]:week ($p > 0.05$):** 상호작용 항이 모두 유의미하지 않습니다. 이는 **"주차가 지나도 프로모션 간의 성적 순위가 바뀌지 않았다"**는 뜻입니다. 1주차에 잘 나온 놈이 4주차에도 계속 잘 나왔다는 의미입니다.
④ 기타 변수 (AgeOfStore & Group Var)
• **AgeOfStore (0.071, $p=0.591$):** 매장의 연령은 매출에 유의미한 영향을 주지 않습니다.
• **Group Var (95.091):** 매장(`LocationID`) 간의 기본 매출 편차가 매우 큽니다. 이는 **단순 ANOVA 대신 LMM(혼합 모델)을 선택하신 것이 매우 탁월한 결정**이었음을 증명합니다. 매장마다의 개별 특성을 모델이 잘 잡아내고 있습니다.

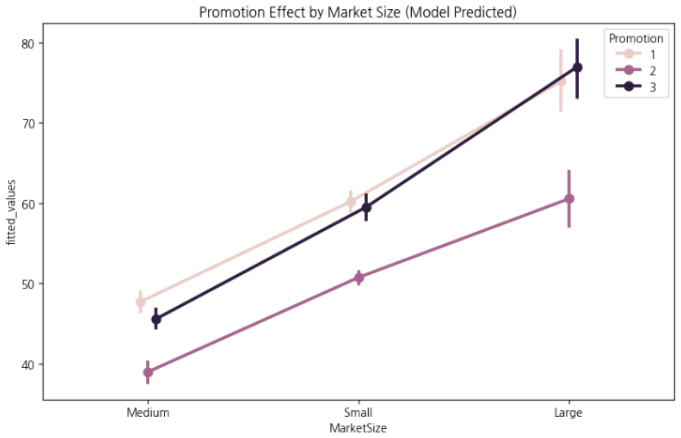
- 기준점 `Promotion 1` , `MarketSize Large` 일 때, **73.4**
- `+` , `-` 의 여부에 따라 기준점에서 더 효과적인지 그렇지 않은지를 표시
- Promotion에서의 `p-value > 0.05` 의 의미, 기각할 수 없다(효과가 다르다고 할 수 없다 ~ 비슷하다)
- week의 `p-value > 0.05` 의 의미, 기각할 수 없다(효과가 없다)
- 따라서, Promotion 1,3를 선택하면 되고, MarketSize별로 유의미한 차이가 있으며, week의 영향력은 없다.
- 이어서 MarketSize와의 상호작용항 확인을 진행한다.

ii. Promotion * MarketSize + AgeOfStore + MarketSize

Mixed Linear Model Regression Results

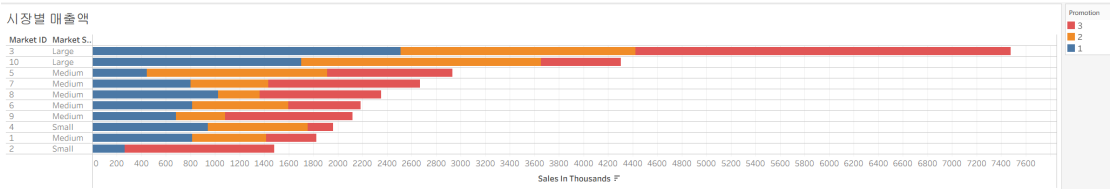
Model:	MixedLM	Dependent Variable:	SalesInThousands
No. Observations:	548	Method:	REML
No. Groups:	137	Scale:	26.5217
Min. group size:	4	Log-Likelihood:	-1840.4831
Max. group size:	4	Converged:	Yes
Mean group size:	4.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	75.009	2.832	26.489	0.000	69.459	80.560
C(Promotion)[T.2]	-15.195	3.684	-4.125	0.000	-22.415	-7.975
C(Promotion)[T.3]	1.853	3.945	0.470	0.639	-5.880	9.586
C(MarketSize)[T.Medium]	-27.919	3.398	-8.216	0.000	-34.579	-21.259
C(MarketSize)[T.Small]	-15.633	5.266	-2.969	0.003	-25.954	-5.312
C(Promotion)[T.2]:C(MarketSize)[T.Medium]	6.815	4.655	1.464	0.143	-2.309	15.940
C(Promotion)[T.3]:C(MarketSize)[T.Medium]	-4.122	4.817	-0.856	0.392	-13.562	5.319
C(Promotion)[T.2]:C(MarketSize)[T.Small]	6.006	7.678	0.782	0.434	-9.042	21.054
C(Promotion)[T.3]:C(MarketSize)[T.Small]	-2.555	7.238	-0.353	0.724	-16.741	11.630
AgeOfStore	0.109	0.132	0.822	0.411	-0.151	0.368
week	-0.164	0.197	-0.836	0.403	-0.550	0.221
Group Var	93.823	2.801				



- MarketSize와의 상호작용항은 역시 큰 효과가 없다는 것을 보여준다(p-value < 0.05)

3. 세그먼트 분석
   - MarketSize별 통계적으로 유의미한 차이가 발견되었으므로, 그보다 더 깊이 있는 분석을 진행하도록 한다.
   - MarketSize로 분류되기 전 기준인 MarketID로 어떤 시장에서 어떤 프로모션이 효과적인지를 분석을 통해 알아본다.
   - (MarketSize와 MarketID는 높은 상관관계를 가지므로, 두 변수 모두를 모델에 넣지 않는다.)
   - MarketID `2` 의 경우 Promotion `2` 가 적용되지 않았으므로, 해당 시장은 부적절한 분석 요소이다. 따라서 MarketID별 분석에서는 이를 제외한다.

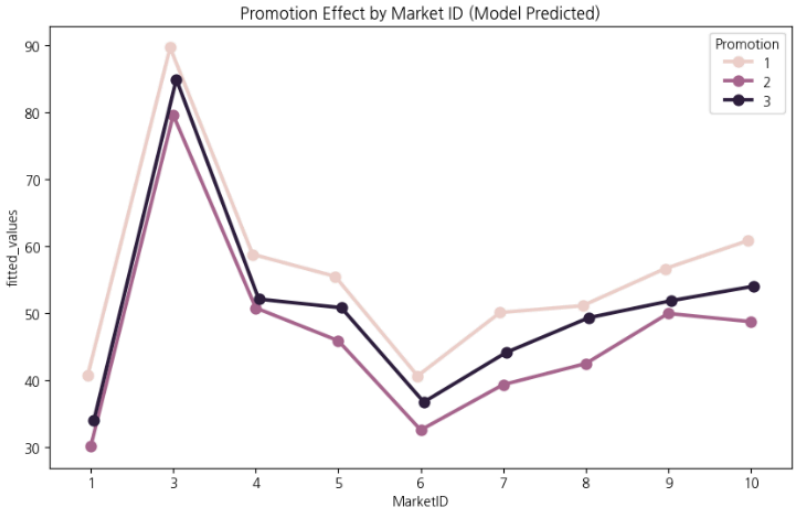



```
import statsmodels.formula.api as smf

# LMM 모델 설계
# 1. 고정 효과: Promotion, MarketID, Week, AgeOfStore (그리고 Promotion과 Week의 상호작용)
# 2. 변량 효과 (groups): LocationID (Location은 MarketID에 포함되므로 보통 하위 단위를 그룹으로 잡음)

model = smf.mixedlm("SalesInThousands ~ C(Promotion) * C(MarketID) + AgeOfStore + Week",
                    data=df,
                    groups=df["LocationID"])

result = model.fit()
print(result.summary())
```

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	SalesInThousands
No. Observations:	524	Method:	REML
No. Groups:	131	Scale:	27.1140
Min. group size:	4	Log-Likelihood:	-1564.8918
Max. group size:	4	Converged:	Yes
Mean group size:	4.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	40.719	1.177	34.599	0.000	38.412	43.026
C(Promotion)[T.2]	-10.567	1.664	-6.349	0.000	-13.829	-7.305
C(Promotion)[T.3]	-6.730	1.922	-3.502	0.000	-10.497	-2.963
C(MarketID)[T.3]	48.928	1.541	31.753	0.000	45.908	51.948
C(MarketID)[T.4]	18.084	1.765	10.244	0.000	14.624	21.544
C(MarketID)[T.5]	14.797	2.202	6.721	0.000	10.482	19.113
C(MarketID)[T.6]	-0.054	1.664	-0.032	0.974	-3.316	3.208
C(MarketID)[T.7]	9.402	1.765	5.326	0.000	5.942	12.862
C(MarketID)[T.8]	10.421	1.664	6.261	0.000	7.159	13.683
C(MarketID)[T.9]	15.949	1.922	8.299	0.000	12.183	19.716
C(MarketID)[T.10]	20.106	1.541	13.048	0.000	17.085	23.126
C(Promotion)[T.2]:C(MarketID)[T.3]	0.512	2.217	0.231	0.817	-3.833	4.857
C(Promotion)[T.3]:C(MarketID)[T.3]	2.004	2.335	0.858	0.391	-2.572	6.581
C(Promotion)[T.2]:C(MarketID)[T.4]	2.575	2.497	1.031	0.302	-2.319	7.468
C(Promotion)[T.3]:C(MarketID)[T.4]	0.044	3.514	0.013	0.990	-6.844	6.932
C(Promotion)[T.2]:C(MarketID)[T.5]	0.969	2.664	0.364	0.716	-4.253	6.191
C(Promotion)[T.3]:C(MarketID)[T.5]	2.053	2.923	0.702	0.482	-3.676	7.781
C(Promotion)[T.2]:C(MarketID)[T.6]	2.498	2.304	1.084	0.278	-2.018	7.014
C(Promotion)[T.3]:C(MarketID)[T.6]	2.830	2.610	1.084	0.278	-2.285	7.945
C(Promotion)[T.2]:C(MarketID)[T.7]	-0.192	2.497	-0.077	0.939	-5.086	4.701
C(Promotion)[T.3]:C(MarketID)[T.7]	0.780	2.533	0.308	0.758	-4.184	5.744
C(Promotion)[T.2]:C(MarketID)[T.8]	1.925	2.760	0.697	0.486	-3.485	7.334
C(Promotion)[T.3]:C(MarketID)[T.8]	4.938	2.542	1.942	0.052	-0.045	9.921
C(Promotion)[T.2]:C(MarketID)[T.9]	3.881	2.923	1.328	0.184	-1.847	9.609
C(Promotion)[T.3]:C(MarketID)[T.9]	1.949	2.718	0.717	0.473	-3.378	7.276
C(Promotion)[T.2]:C(MarketID)[T.10]	-1.498	2.110	-0.710	0.478	-5.633	2.638
C(Promotion)[T.3]:C(MarketID)[T.10]	-0.044	2.644	-0.017	0.987	-5.226	5.138
Group Var	0.147	0.207				



해석

- 기준점 : Promotion `1` , MarketID `1` , AgeOfStore `0` , week `1` (41.263)
- Promotion `1`, `3` 간의 유의미한 차이가 발견되었다. Promotion `1` 이 제일 좋은 결과를 나타낸다.
- Market `1`, `6` 은 유의미한 차이가 존재하지 않는다.
- 상호작용항은 모두 p-value > 0.05 이므로 유의미하지 않다.

3. 결론
    a. 분석 결과
        i. `MarketSize` 별 프로모션 적용 결과를 확인했을 때, 모든 시장의 크기에서 Promotion `2` 의 결과보단 Promotion `1`, `3` 의 매출이 유의미하게 높았다. 그러나, `1` 과 `3` 사이에서는 유의미한 차이가 나타나지 않았다.
        ii. `MarketID` 별 프로모션 적용 결과를 확인했을 때, MarketID `2` 를 제외한 시장에서 Promotion `2` 가 다른 프로모션보다 통계적으로 유의미하게 높은 값을 갖는다.
        iii.

        세그먼트 분석으로 P1이 P3보다 유의하게 더 좋다라는 것을 통계적으로 확인

        표준편차로 인해 리스크가 큰 P3과

        전체 프로모션에서 가장 낮은 매출을 보인 P2 제외

        P1을 선택하는게 장기적 매출 부분에서 유리해보인다.

    b. 가드레일 지표 / 서포트 지표
        i. 프로모션별 매장당 4주 평균 매출액의 표준편차

c. 비즈니스 제안

d. 데이터 한계

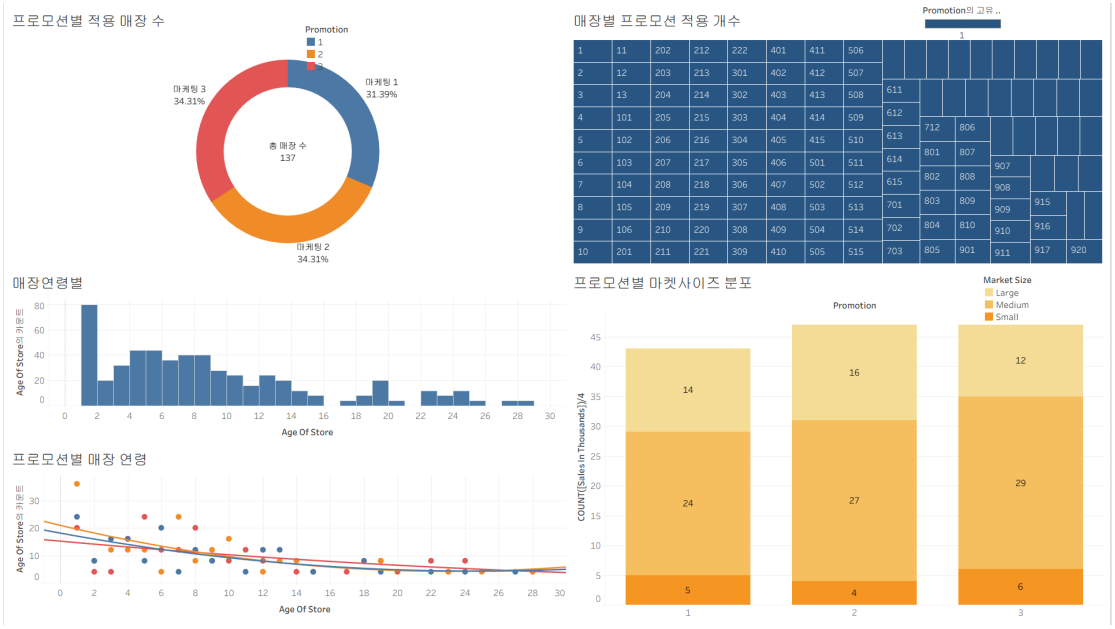
i. 데이터 정합성을 추가로 확인하기 위해서는 프로모션 적용 전의 매출액을 비교하는 과정이 필요하였지만, 제공 데이터셋에는 그러한 항목이 없었다.

ii. 데이터를 무작위적으로 선정했다고 했기에, 초기 매출액(프로모션이 적용되기 전의 매출액)이 같다고 '가정'해야했다. 또한, 0주차 데이터가 없기에 성장률이 아닌, '현재 성적'을 비교하여야 한다.

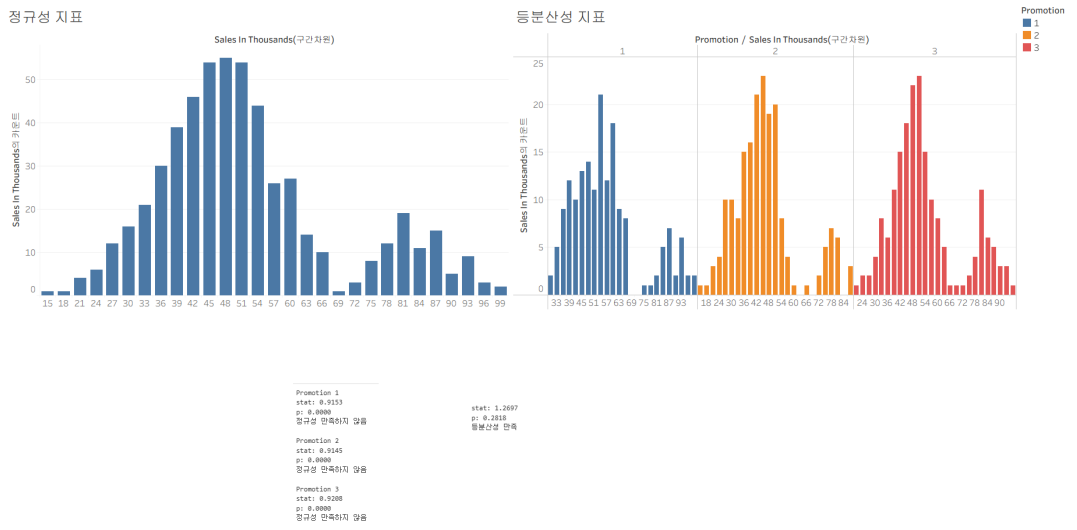
iii. 그렇기 때문에, 데이터 정합성 판단에 더욱 의존적일 수밖에 없다.

▼ 대충 만든 대시보드 (수정 예정)

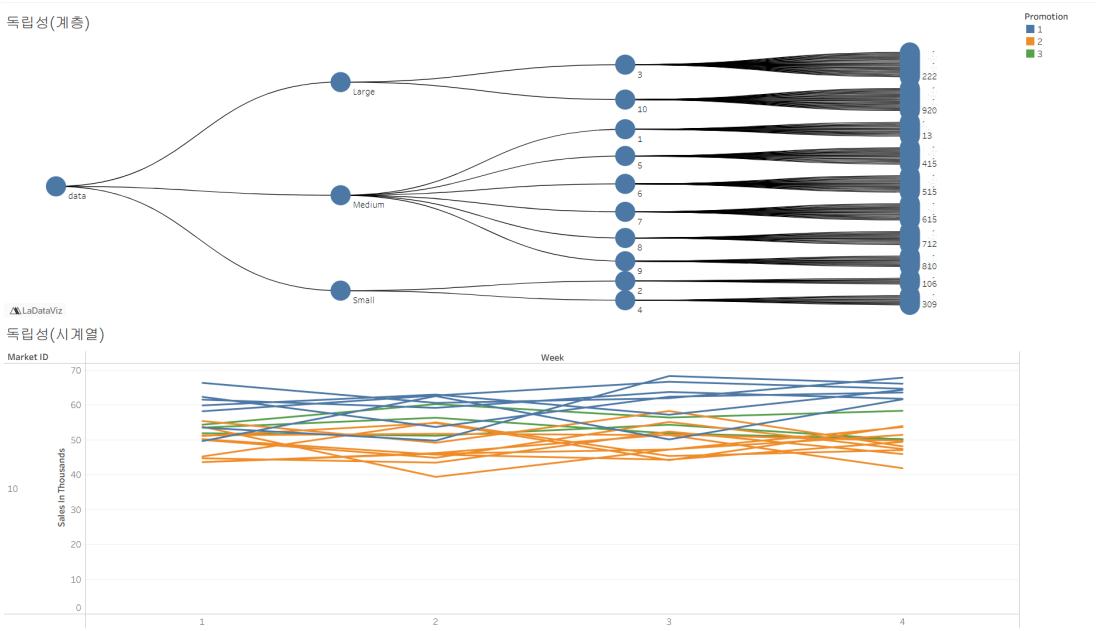
▼ 데이터 정합성을 보여주는 페이지



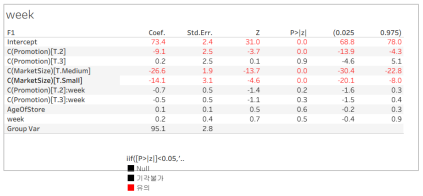
▼ 정규성, 등분산성 을 보여주는 페이지



▼ 독립성



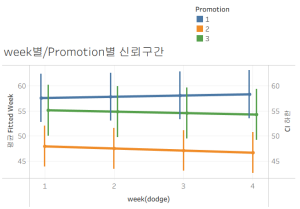
▼ LMM(week)



week

F1	Coef.	Std.Err.	z	P>|z|	(0.025	0.975)
Intercept	73.4	2.4	31.0	0.0	68.8	78.0
C(Promotion)[T.2]	-9.1	2.5	-3.7	0.0	-13.9	-4.3
C(Promotion)[T.3]	0.2	2.5	0.1	0.9	-4.6	5.1
C(MarketSize)[T.Medium]	-26.6	1.9	-13.7	0.0	-30.4	-22.8
C(MarketSize)[T.Small]	-14.1	3.1	-4.6	0.0	-20.1	-8.0
C(Promotion)[T.2]:week	-0.7	0.5	-1.4	0.2	-1.6	0.3
C(Promotion)[T.3]:week	-0.5	0.5	-1.1	0.3	-1.5	0.4
AgeOfStore	0.1	0.1	0.5	0.6	-0.2	0.3
week	0.2	0.4	0.7	0.5	-0.4	0.9
Group Var	95.1	2.8				

if([P>|z|]<0.05,'..
Null
귀각보가
유의



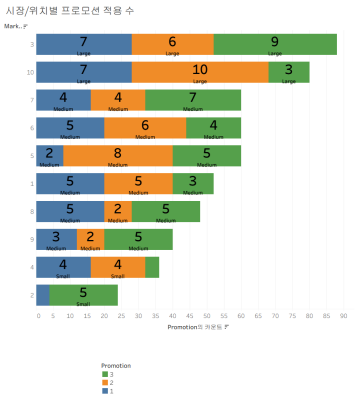
▼ LMM(MarketSize)

MarketSize

F1	Coef.	Std.Err.	Z	P>|z|	(0.025	0.975)
Intercept	75.0	2.8	26.5	0.0	69.5	80.6
AgeOfStore	0.1	0.1	0.8	0.4	-0.2	0.4
C(Promotion)[T.2]	-15.2	3.7	-4.1	0.0	-22.4	-8.0
C(Promotion)[T.3]	1.9	3.9	0.5	0.6	-5.9	9.6
C(MarketSize)[T.Medium]	-27.9	3.4	-8.2	0.0	-34.6	-21.3
C(MarketSize)[T.Small]	-15.6	5.3	-3.0	0.0	-26.0	-5.3
C(Promotion)[T.2]:C(MarketSize)[T.Medium]	6.8	4.7	1.5	0.1	-2.3	15.9
C(Promotion)[T.3]:C(MarketSize)[T.Medium]	-4.1	4.8	-0.9	0.4	-13.6	5.3
C(Promotion)[T.2]:C(MarketSize)[T.Small]	6.0	7.7	0.8	0.4	-9.0	21.1
C(Promotion)[T.3]:C(MarketSize)[T.Small]	-2.6	7.2	-0.4	0.7	-16.7	11.6
week	-0.2	0.2	-0.8	0.4	-0.6	0.2
Group Var	93.8	2.8				

MarketSize별/Promotion별 신뢰구간

▼ 세그먼트 전(시장2 제외)



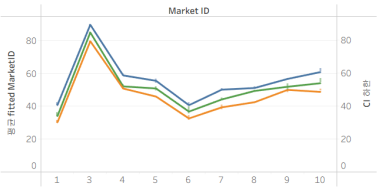
Market 2

▼ LMM(MarketID)

MarketID

F1	Coef.	Std.Err.	Z	P>|z|	(0.025	0.975)
Intercept	40.72	1.18	34.60	0.00	38.41	43.03
C(Promotion)[T.2]	-10.57	1.66	-6.35	0.00	-13.83	-7.31
C(Promotion)[T.3]	-6.73	1.92	-3.50	0.00	-10.50	-2.96
C(MarketID)[T.3]	48.93	1.54	31.75	0.00	45.91	51.95
C(MarketID)[T.4]	18.08	1.77	10.24	0.00	14.62	21.54
C(MarketID)[T.5]	14.80	2.20	6.72	0.00	10.48	19.11
C(MarketID)[T.6]	-0.05	1.66	-0.03	0.97	-3.32	3.21
C(MarketID)[T.7]	9.40	1.77	5.33	0.00	5.94	12.86
C(MarketID)[T.8]	10.42	1.66	6.26	0.00	7.16	13.68
C(MarketID)[T.9]	15.95	1.92	8.30	0.00	12.18	19.72
C(MarketID)[T.10]	20.11	1.54	13.05	0.00	17.09	23.13
C(Promotion)[T.2]:C(MarketID)[T.3]	0.51	2.22	0.23	0.82	-3.83	4.86
C(Promotion)[T.2]:C(MarketID)[T.4]	2.58	2.50	1.03	0.30	-2.32	7.47
C(Promotion)[T.2]:C(MarketID)[T.5]	0.97	2.66	0.36	0.72	-4.25	6.19
C(Promotion)[T.2]:C(MarketID)[T.6]	2.50	2.30	1.08	0.28	-2.02	7.01
C(Promotion)[T.2]:C(MarketID)[T.7]	-0.19	2.50	-0.08	0.94	-5.09	4.70
C(Promotion)[T.2]:C(MarketID)[T.8]	1.93	2.76	0.70	0.49	-3.49	7.33
C(Promotion)[T.2]:C(MarketID)[T.9]	3.88	2.92	1.33	0.18	-1.85	9.61
C(Promotion)[T.2]:C(MarketID)[T.10]	-1.50	2.11	-0.71	0.48	-5.63	2.64
C(Promotion)[T.3]:C(MarketID)[T.3]	2.00	2.34	0.86	0.39	-2.57	6.58
C(Promotion)[T.3]:C(MarketID)[T.4]	0.04	3.51	0.01	0.99	-6.84	6.93
C(Promotion)[T.3]:C(MarketID)[T.5]	2.05	2.92	0.70	0.48	-3.68	7.78
C(Promotion)[T.3]:C(MarketID)[T.6]	2.83	2.61	1.08	0.28	-2.29	7.95
C(Promotion)[T.3]:C(MarketID)[T.7]	0.78	2.53	0.31	0.76	-4.18	5.74
C(Promotion)[T.3]:C(MarketID)[T.8]	4.94	2.54	1.94	0.05	-0.05	9.92
C(Promotion)[T.3]:C(MarketID)[T.9]	1.95	2.72	0.72	0.47	-3.38	7.28
C(Promotion)[T.3]:C(MarketID)[T.10]	-0.04	2.64	-0.02	0.99	-5.23	5.14
Group Var	0.15	0.21				

MarketID별/Promotion별 신뢰구간



▼ 태블로(박정우)

문서1.twbx

▼ 태블로(유정하)

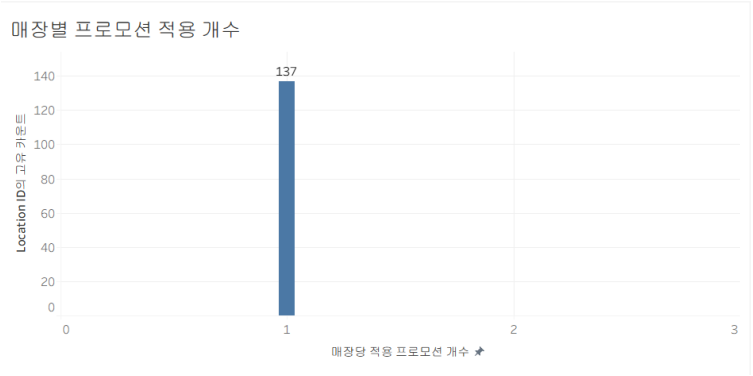
fastFood.twb

▼ 태블로(송현준)

fastfood.twbx

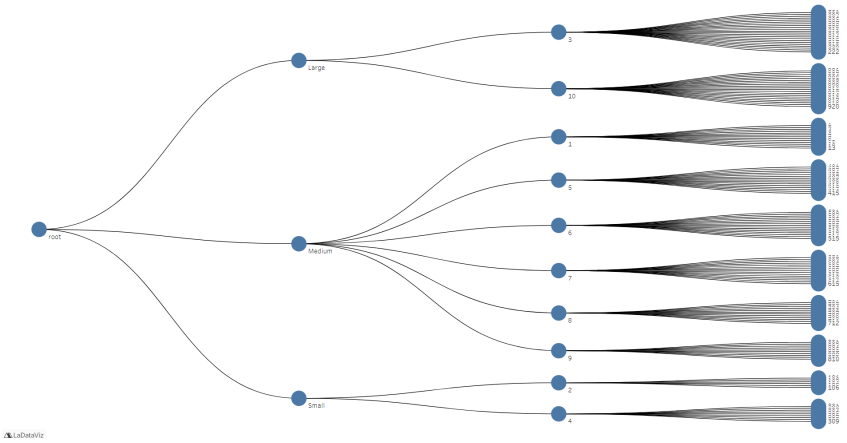
▼ 추가한 그래프

1. 매장당 프로모션 적용 개수

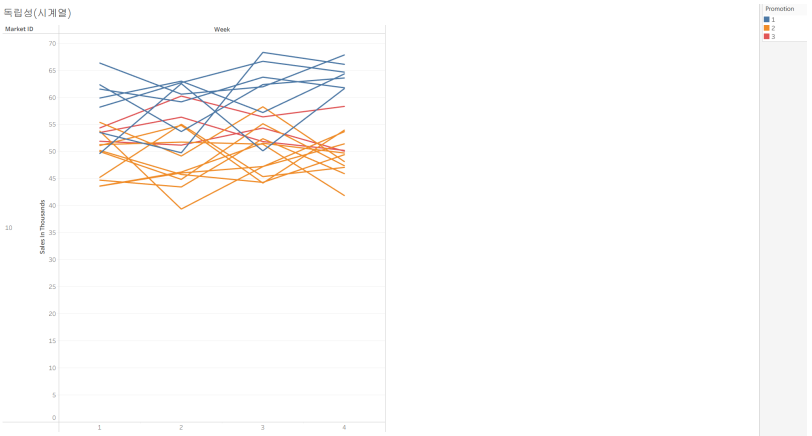


매장당(총 137개) 프로모션이 1개씩 존재한다는 것을 보여주는 그래프입니다.

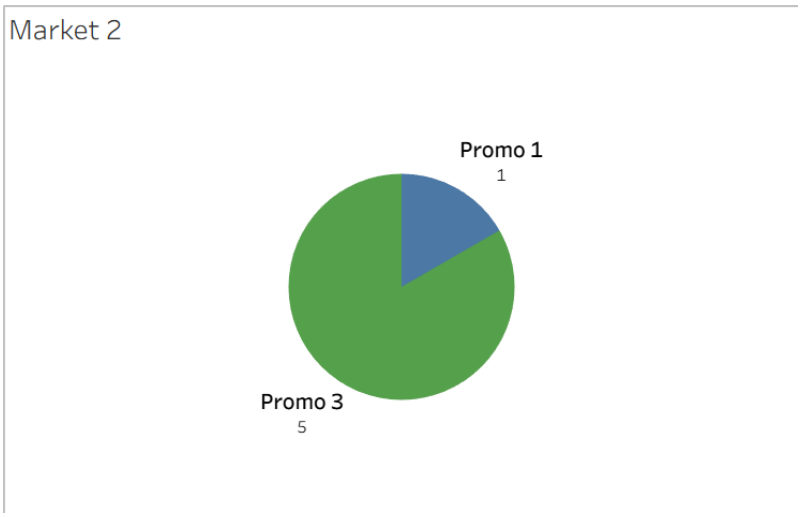
2. 독립성 위반 (계층적 구조)



3. 독립성 위반 (시계열)



4. Market2의 구성 프로모션



비즈니스 제안

액션 아이템

리스크 관리(가드레일) 측면에서

- 하위 매장에 피해가 나타나는지 그 유무에 대한 모니터링 - 하위 25% 매장의 매출에 주간 점검을 실시(프로모션 적용, 변경 등으로 하위가 망가진다면 중단 고려)
- 변동성 모니터링 - std 또는 상위 퍼센타일(p75~p90) 구간이 갑자기 커지는지 그런 것을 트래킹

실험을 진행하면서 느낀 점, 향후 실행 가능한 개선 방향

아쉬웠던 점

- 0주차(프로모션을 진행하기 전)의 매출이 없었던 것
  - 다음에는 베이스라인이 확보되면 더 좋을 것으로 예상됨 (성과를 평가하기 전에 미리 데이터 확보)
- Market2에 Promotion2가 배정되지 않은 사유가 기재되었다면 더 좋지 않았을까

패스트푸드의 특성을 적용했을 때 전략, 액션 도출

패스트푸드의 의사결정 구조

패스트푸드는 다른 리테일, 외식과 의사결정 매커니즘이 다르다.

구매 결정 시간은 수 초~수십 초이며 메뉴판 앞에서 메뉴들에 대해 깊이 비교하거나 고민하지 않는다.

따라서 생각하게 만드는 순간이 길어질수록, 즉 프로모션이 복잡해서 이해에 시간이 더 걸릴수록 그 프로모션의 성과가 급락할 가능성이 커진다.

패스트푸드 고객은 보통 지금 당장 얻는 혜택이 있는지, 해당 프로모션이 이해하기 쉬운지 등의 순서로 판단한다. (즉시성과 직관성이 핵심이며 인지 리소스를 거의 쓰지 않음)

이를 통해 유추해볼 때, Promotion 1은 그 환경에서 가장 덜 생각하게 만드는 구조였을 가능성이 크다.

평균 매출이 최고였고, 하위 매장에서도 안정적이었으며, 변동성이 과하지 않았다.

혜택을 이해하고 적용하고 구매하기까지의 인지적 단계가 짧지 않았을까 추측해볼 수 있다.

Promotion 2가 다른 프로모션에 비해 열세였던 이유도 반대로 추측해볼 수 있다.

따라서 다음 유사한 설계와 실행에서는 **중간 퍼널(주문 과정 한복판에 있는) 지표**도 추가된다면 좋을 것이다.

주문 완료까지 걸린 시간, 키오스크 취소율, 업셀링(사이드 선택 유무, 단품을 세트로 변경하는지 등) 비율

이를 통해 주문 흐름이 방해가 되었는지, 그렇다면 그 사유는 무엇일 수 있는지도 고려될 수 있다.

▼ 최신 파일(결론 및 제언 비었습니다.)

fastfood.twbx