수DA쟁이

헤드퍼스트 DA - chap 1
데이터 분석 입문
(잘게 쪼개라)

Department of Mathematics
Gyeongsang National University
Youngmin Shin

데이터 분석가가 하는 일

-많은 데이터를 사용해 실제 세계에서 적용할 수 있는 정보를 유추
-복잡한 문제와 데이터를 분할 하여 구조화
-접근하는 방법이 핵심

Acme 화장품은 여러분의 도움이 필요합니다.

	September	October	November	December	January	February
Gross sales	$5,280,000	$5,501,000	$5,469,000	$5,480,000	$5,533,000	$5,554,000
Target sales	$5,280,000	$5,500,000	$5,729,000	$5,968,000	$6,217,000	$6,476,000
Ad costs	$1,056,000	$950,400	$739,200	$528,000	$316,800	$316,800
Social network costs	$0	$105,600	$316,800	$528,000	$739,200	$739,200
Unit prices (per oz.)	$2.00	$2.00	$2.00	$1.90	$1.90	$1.90

총매출

목표 매출

광고 비용

단가

소셜 네트워크 비용

-우수한 데이터 분석가는 항상 데이터를 '볼 ' 수 있어야 합니다.

CEO는 데이터 분석을 통해 매출을 올릴 수 있도록 도움을 받고 싶어 합니다.

CEO

CEO

CEO

데이터 분석은 증거에 대해 곰곰이 고찰하는 것이다.

-데이터 분석이라는 표현에는 여러 직업과 많은 기술이 포함되어 있습니다.

-우수한 분석가는 동일한 기본 절차를 밟음. 실질적인 증거를 가지고 문제에 대하여 곰곰이 고찰합니다.

정의: 제일 먼저 문제가 무엇인지 알아야 하기 때문에 문제를 정의한다.

분해: 데이터 분석의 문제와 데이터를 작은 조각으로 분해하는 것이다.

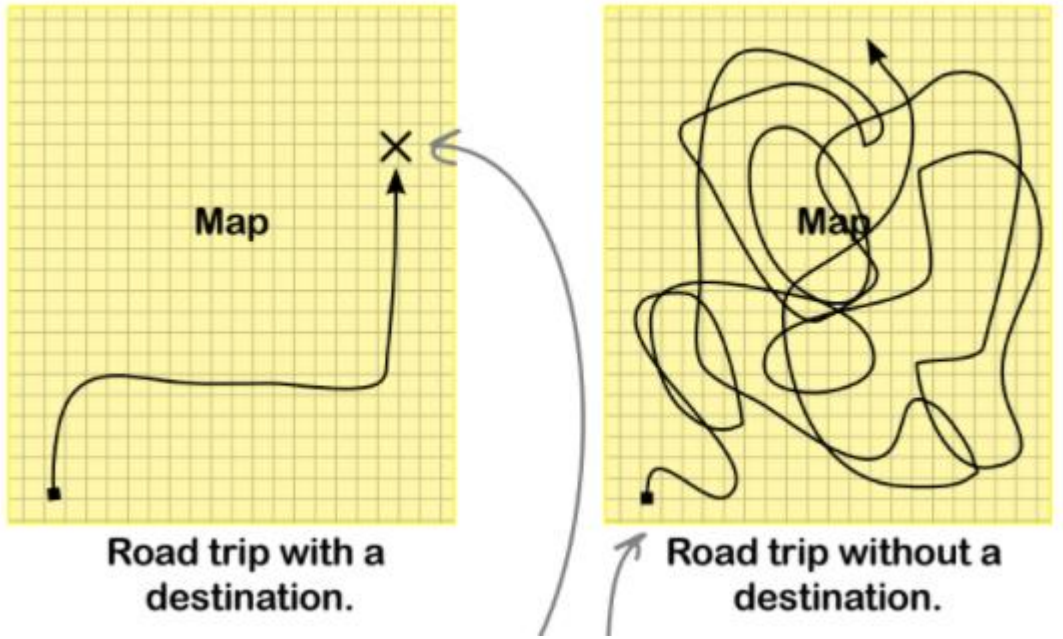
평가: 분석의 골자입니다. 앞의 두 단계에서 알게 된 것들에 대한 결론을 이끌어냅니다.

결정: 마지막으로, 모든 것을 한데 모아서 결론을 내립니다. 혹은 권고

+모든 데이터 분석은 더 나은 판단을 이끌어 내기 위해 설계됩니다.

문제를 정의합니다.(정의)

-데이터 분석가가 자신의 주장을 피력하기 위해 너무 많은 분량을 작성하는 것은 종종 문제의 초점을 명확히 잡지 못하고, 문제를 해결하고 그 결론을 권고해야만 하는 자신의 책임을 회피하는 것.

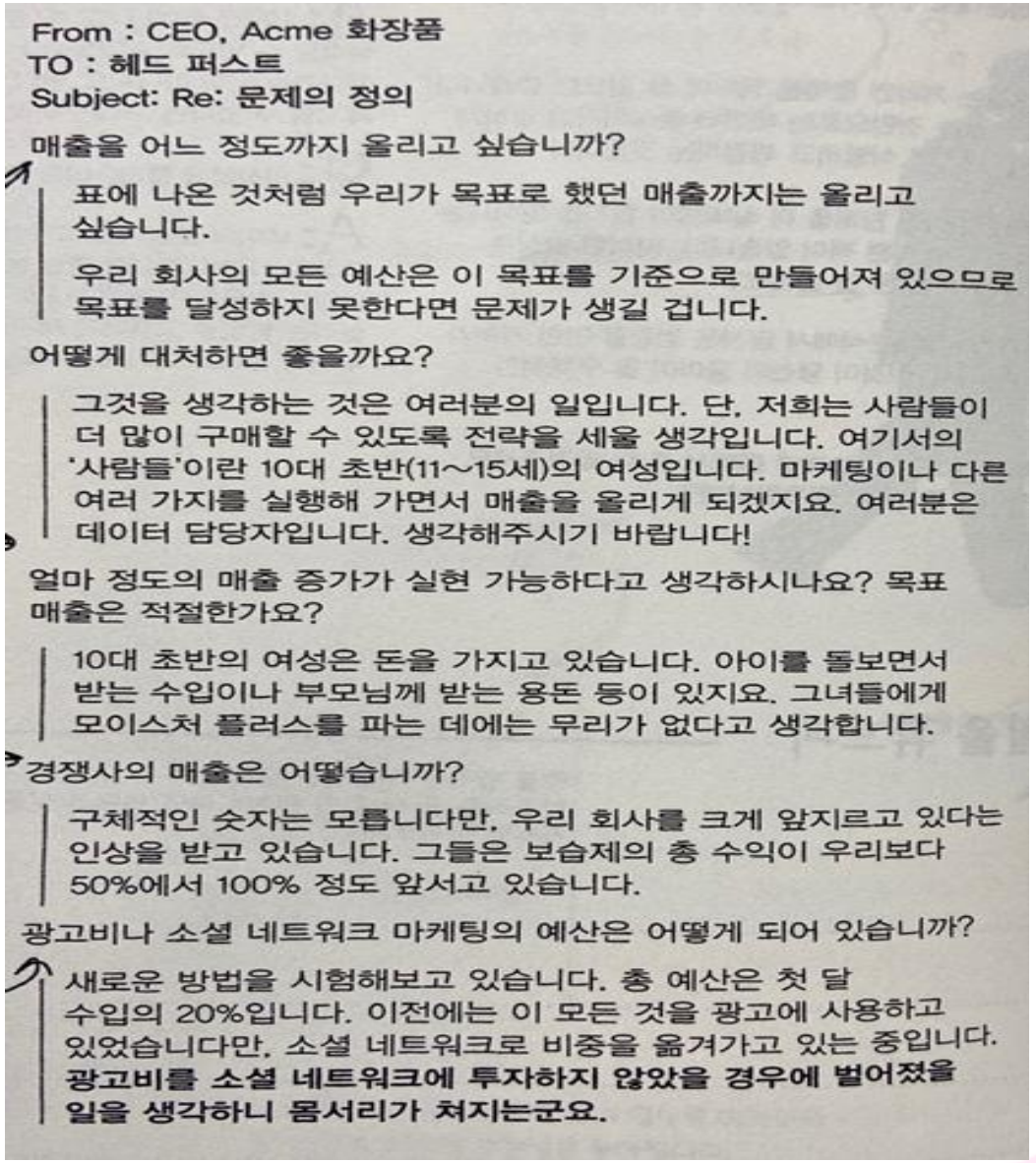


여러분의 고객은 문제를 정의하는데 도움이 됩니다.(정의)

+고객은 분석 결과에 기초해서 판단을 내립니다.
+데이터 분석가는 문제를 정의하기 위해 최대한 고객에게서 많은 정보를 얻어내야 한다. 고객의 의도를 구체적으로 이해 해야함.
+고객에 대해 더 많이 알수록 분석의 도움이 된다.

Acme의 CEO가 답장을 보내왔습니다.(정의)

+고객에게 분석 목적을 명확하게 얻기 위해 질문을 합니다.
+수량을 물어보는 것이 좋다. 이는 목적이나 믿음을 정량화 합니다.
+데이터에 나와 있는 수치에 대해서도 궁금한 것이 있으면 물어 봐야합니다.



From : CEO, Acme 화장품
TO : 헤드 퍼스트
Subject: Re: 문제의 정의

매출을 어느 정도까지 올리고 싶습니까?

표에 나온 것처럼 우리가 목표로 했던 매출까지는 올리고 싶습니다.

우리 회사의 모든 예산은 이 목표를 기준으로 만들어져 있으므로 목표를 달성하지 못한다면 문제가 생길 겁니다.

어떻게 대처하면 좋을까요?

그것을 생각하는 것은 여러분의 일입니다. 단, 저희는 사람들이 더 많이 구매할 수 있도록 전략을 세울 생각입니다. 여기서의 '사람들'이란 10대 초반(11~15세)의 여성입니다. 마케팅이나 다른 여러 가지를 실행해 가면서 매출을 올리게 되겠지요. 여러분은 데이터 담당자입니다. 생각해주시기 바랍니다!

얼마 정도의 매출 증가가 실현 가능하다고 생각하시나요? 목표 매출은 적절한가요?

10대 초반의 여성은 돈을 가지고 있습니다. 아이를 돌보면서 받는 수입이나 부모님께 받는 용돈 등이 있지요. 그녀들에게 모이스처 플러스를 파는 데에는 무리가 없다고 생각합니다.

경쟁사의 매출은 어떻습니까?

구체적인 숫자는 모릅니다만, 우리 회사를 크게 앞지르고 있다는 인상을 받고 있습니다. 그들은 보습제의 총 수익이 우리보다 50%에서 100% 정도 앞서고 있습니다.

광고비나 소셜 네트워크 마케팅의 예산은 어떻게 되어 있습니까?

새로운 방법을 시험해보고 있습니다. 총 예산은 첫 달 수입의 20%입니다. 이전에는 이 모든 것을 광고에 사용하고 있었습니다만, 소셜 네트워크로 비중을 옮겨가고 있는 중입니다. **광고비를 소셜 네트워크에 투자하지 않았을 경우에 벌어졌을 일을 생각하니 몸서리가 쳐지는군요.**

문제와 데이터를 작게 분할 합니다.(분석)

-문제 정의 다음 단계는 고객에게 들은 문제의 내용과 데이터를 분석하는데 가장 적합할 정도의 입자 수준으로 분할 하는 것이다.

-문제를 작은 단위로 분할 한다.

- 문제를 다루기 쉽게 해결할 수 있는 크기로 분할해야 함.
- 문제는 종종 모호한 경우가 있다.
- 큰 문제에 직접 대답하기는 어렵다. 하지만 큰 문제에서 분해한 작은 문제의 해답을 찾으면 더 쉽게 문제를 풀 수 있다.

-데이터를 작게 분할 한다.

- 필요한 그리고 정확하고 정량적인 대답이 직접 나타나지 않는다.
- 원시 데이터를 받았다면 각 요소를 요약해 데이터를 쉽게 사용하는 것이 좋다.

아는 것도 관점을 달리해서 봅니다.(분해)

-흥미 있는 비교/대조를 해 가면서 요약 데이터를 분해합니다.
-요약 데이터가 있을 때 가장 중요한 요소를 추출하기 위한 최적의 방법은 강한 대조를 찾아내는 것 입니다.
-적절한 비교를 실시하는 것이 데이터 분석의 요점입니다.
-고객의 가정이 올바른 것은 중요하다. 왜냐하면 고객의 가정이 분석의 중추가 되기 때문입니다.
-고객과의 질의 자체도 일종의 데이터 입니다.

각 요소에 대해 평가합니다.(평가)

-데이터 분석가의 판단을 정리하는 순서이다.
-분해 단계에서 했던 것처럼 분할 했던 각각의 요소를 평가하는 방법의 핵심은 비교이다.

분석은 자신을 포함시키는 것에서 시작됩니다.(평가)

-분석에서 자기 자신을 포함 시킨다는 것은 자신의 가정을 명시적으로 밝히고, 결론에 대해 깊은 확신을 가진다는 것.

-자기 자신을 포함 시킨 경우

- 분석가에게 좋은 점: 데이터에서 무엇을 찾아야 할지 알 수 있다. 무리한 결론을 내리는 것을 피할 수 있다. 일의 성공에 책임을 질 수 있다.
- 고객에게 좋은 점: 고객이 분석가의 판단을 더 존중한다. 고객이 분석가가 내린 결론의 한계를 이해한다.

-자기 자신을 포함 시키지 않은 경우

- 분석가에게 좋지 않은 점: 기준이 되는 가정이 결론에 어떤 영향을 미치는지 파악할 수 없다. 책임을 회피하게 된다.
- 고객에게 좋지 않은 점: 고객이 분석가의 진심과 의욕을 모르기 때문에 분석을 신뢰하지 않는다. 고객은 '객관성'에 대해 잘못된 인식을 가지거나 합리성이 부족하다고 여길 가능성이 있다.

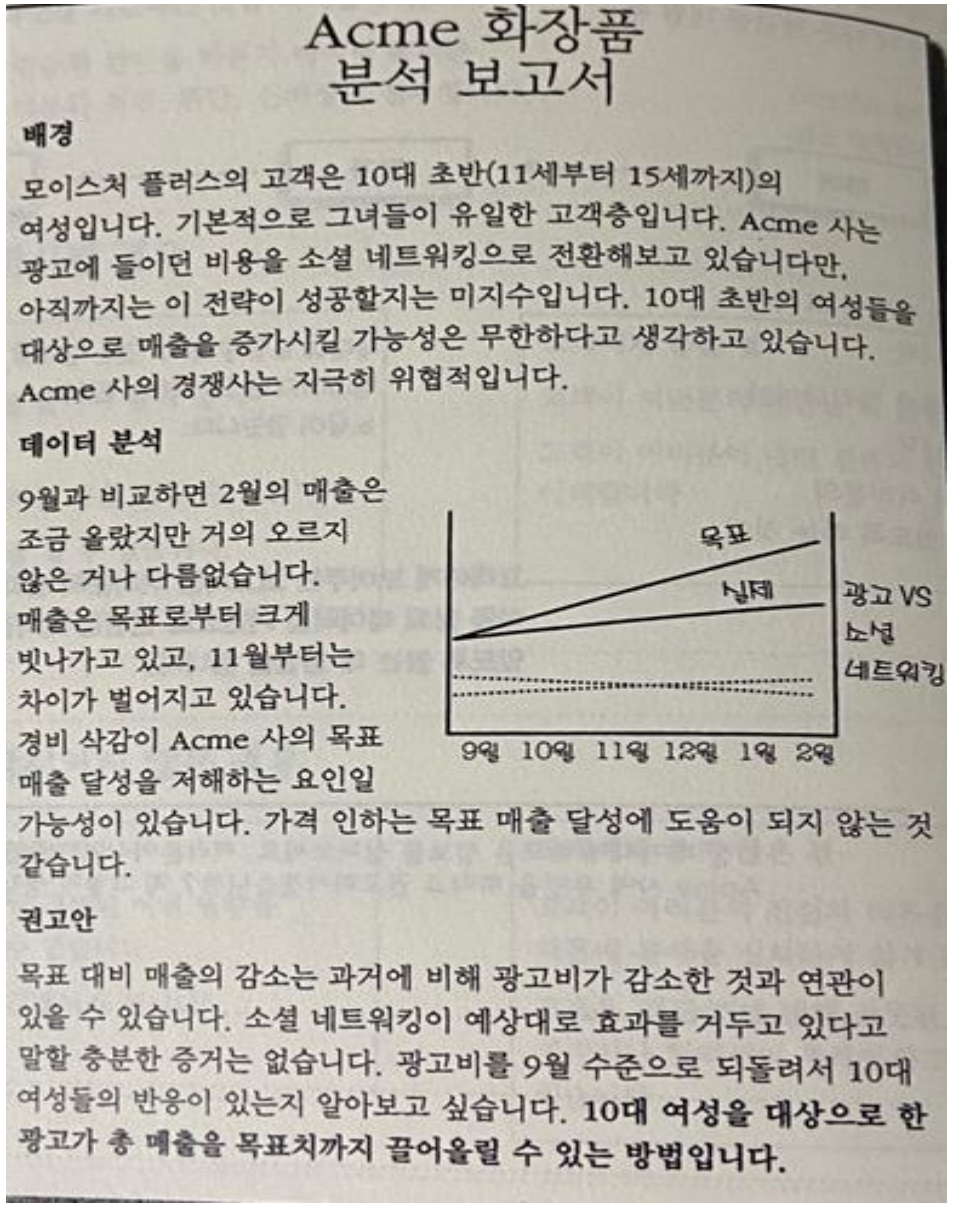
권고안을 냅니다.(결정)

-데이터 분석가가 할 일은 데이터에 대해 꼼꼼히 검토하여 얻은 통찰을 바탕으로, 분석가와 고객이 더 나은 판단을 내릴 수 있도록 힘을 실어주는 것이다.
-위와 같이 하기 위해서는 분석가의 생각과 판단을 고객이 이해할 수 있는 형식으로 정리해야 한다.
-분석 결과는 가능한 간결하게 정리되어야 하지만 너무 간략하게 적어서도 안 됩니다.
-분석가의 권고에 근거하여 좋은 판단을 내릴 수 있도록 하는 것이 분석가가 할 일이다.
-판단에 도움을 줄 수 있는 형태로 정리하지 않으면 분석 결과는 소용이 없다.
-고객에게 보여주는 보고서는 분석가의 생각을 전달하고, 이를 통해 데이터를 기반으로 현명한 의사결정을 할 수 있도록 돕는데 중점을 두어야 한다.

보고서가 준비되었습니다.

+보고서에 분석가와 고객의 가정을 말하는 것은 좋다.

+결론을 설명하는 그래프를 넣는 것도 좋다.



Acme 화장품
분석 보고서

배경

모이스처 플러스의 고객은 10대 초반(11세부터 15세까지)의 여성입니다. 기본적으로 그녀들이 유일한 고객층입니다. Acme 사는 광고에 들이던 비용을 소셜 네트워킹으로 전환해보고 있습니다만, 아직까지는 이 전략이 성공할지는 미지수입니다. 10대 초반의 여성들을 대상으로 매출을 증가시킬 가능성은 무한하다고 생각하고 있습니다. Acme 사의 경쟁사는 지극히 위협적입니다.

데이터 분석

9월과 비교하면 2월의 매출은 조금 올랐지만 거의 오르지 않은 거나 다름없습니다. 매출은 목표로부터 크게 빗나가고 있고, 11월부터는 차이가 벌어지고 있습니다. 경비 삭감이 Acme 사의 목표 매출 달성을 저해하는 요인일 가능성이 있습니다. 가격 인하는 목표 매출 달성에 도움이 되지 않는 것 같습니다.

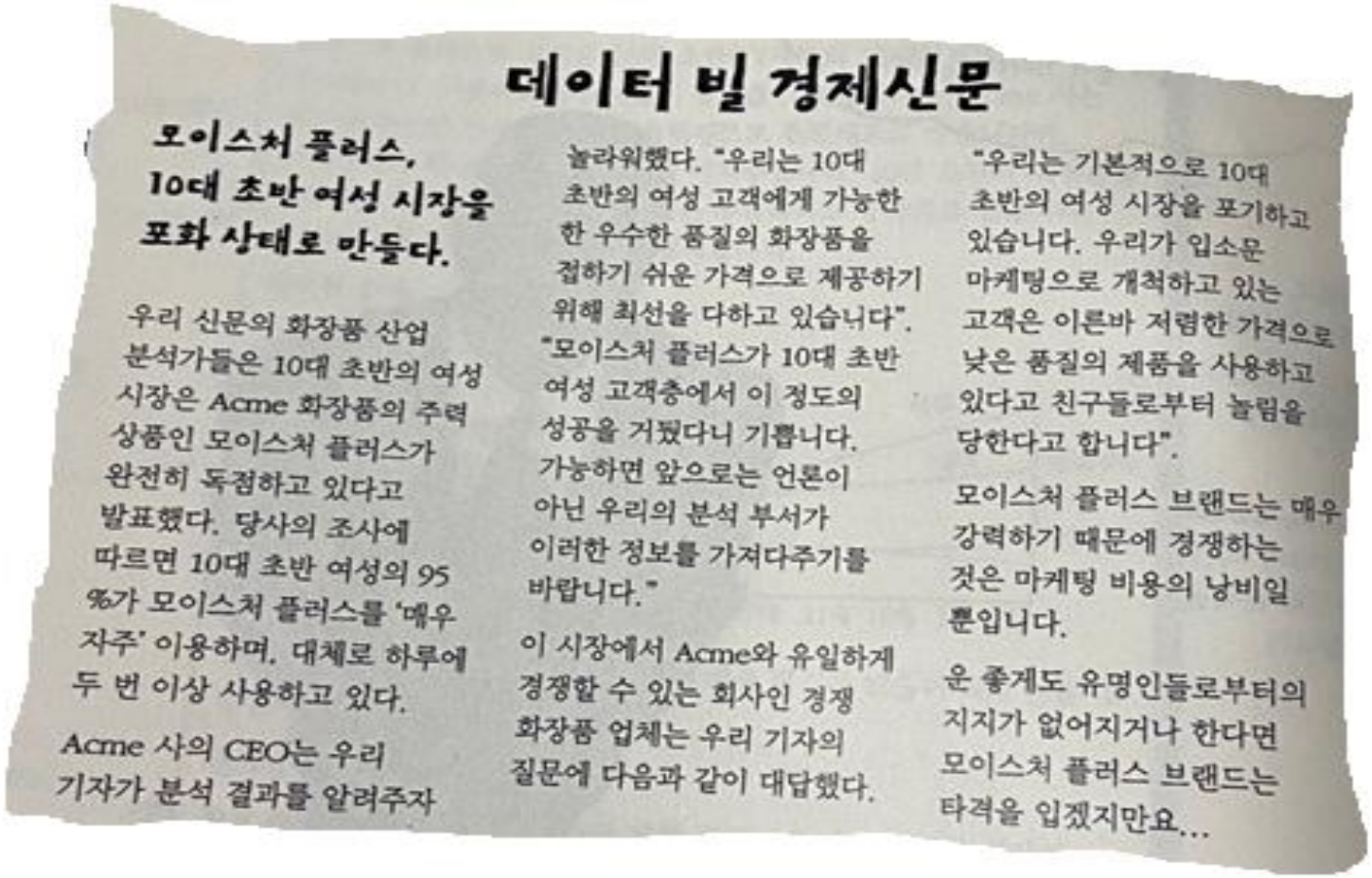
권고안

목표 대비 매출의 감소는 과거에 비해 광고비가 감소한 것과 연관이 있을 수 있습니다. 소셜 네트워킹이 예상대로 효과를 거두고 있다고 말할 충분한 증거는 없습니다. 광고비를 9월 수준으로 되돌려서 10대 여성들의 반응이 있는지 알아보고 싶습니다. **10대 여성을 대상으로 한 광고가 총 매출을 목표치까지 끌어올릴 수 있는 방법입니다.**

CEO가 분석 결과에 만족해야합니다.

+고객의 요구를 고객이 묘사한 것 보다 명확하게 보여 주면 좋다.
+고객의 명확한 의견을 구하고, 고객의 의견과 데이터의 대한 분석가 자신의 분석을 비교하여 결론을 권고하면 좋다.

신문에 기사가 났습니다



데이터 빌 경제신문

모이스처 플러스, 10대 초반 여성 시장을 포화 상태로 만들다.

우리 신문의 화장품 산업 분석가들은 10대 초반의 여성 시장은 Acme 화장품의 주력 상품인 모이스처 플러스가 완전히 독점하고 있다고 발표했다. 당사의 조사에 따르면 10대 초반 여성의 95%가 모이스처 플러스를 '매우 자주' 이용하며, 대체로 하루에 두 번 이상 사용하고 있다.

Acme 사의 CEO는 우리 기자가 분석 결과를 알려주자 놀라워했다. "우리는 10대 초반의 여성 고객에게 가능한 한 우수한 품질의 화장품을 접하기 쉬운 가격으로 제공하기 위해 최선을 다하고 있습니다". "모이스처 플러스가 10대 초반 여성 고객층에서 이 정도의 성공을 거뒀다니 기쁩니다. 가능하면 앞으로는 언론이 아닌 우리의 분석 부서가 이러한 정보를 가져다주기를 바랍니다."

이 시장에서 Acme와 유일하게 경쟁할 수 있는 회사인 경쟁 화장품 업체는 우리 기자의 질문에 다음과 같이 대답했다. "우리는 기본적으로 10대 초반의 여성 시장을 포기하고 있습니다. 우리가 입소문 마케팅으로 개척하고 있는 고객은 이른바 저렴한 가격으로 낮은 품질의 제품을 사용하고 있다고 친구들로부터 놀림을 당한다고 합니다".

모이스처 플러스 브랜드는 매우 강력하기 때문에 경쟁하는 것은 마케팅 비용의 낭비일 뿐입니다.

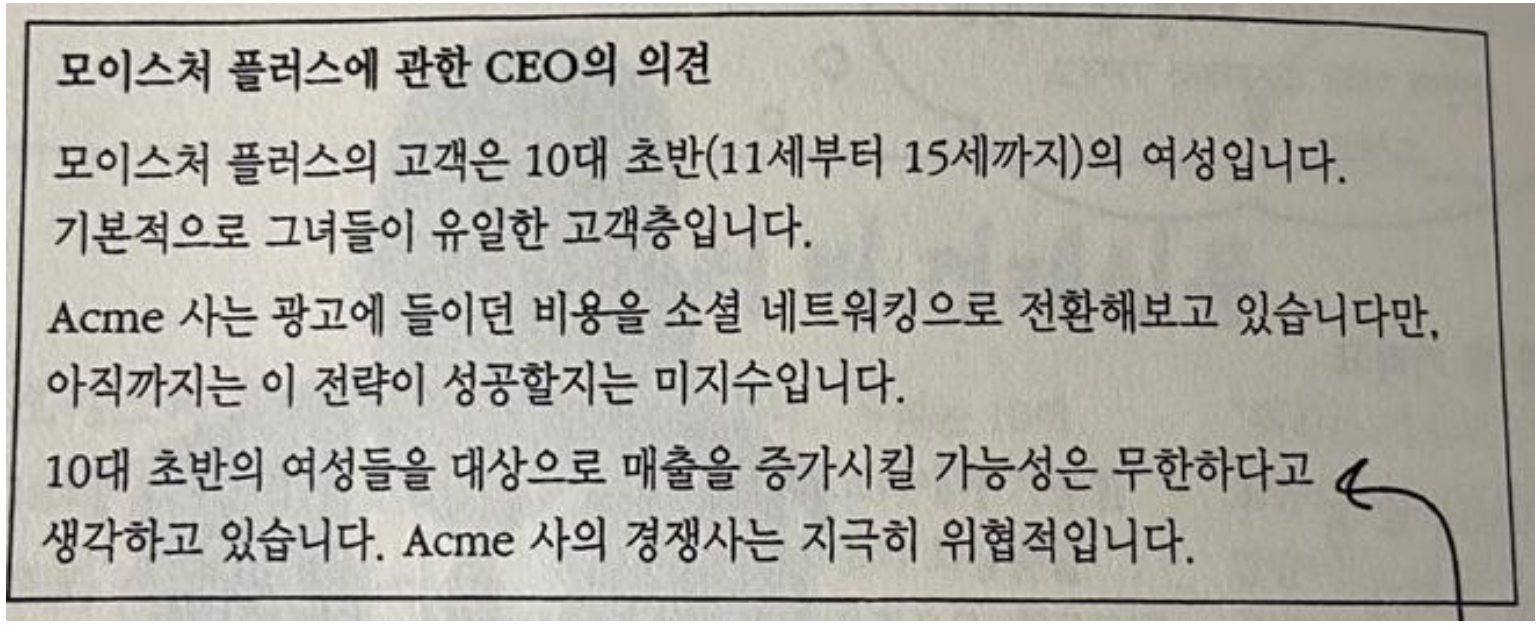
운 좋게도 유명인들로부터의 지지가 없어지거나 한다면 모이스처 플러스 브랜드는 타격을 입겠지만요...

결과가 만족스럽지 않다면 분석 과정에서 어떤 일이 일어났는지 파악해야 한다.

CEO의 의견이 잘못된 결론으로 이끌었습니다.

+멘탈 모델: 세상에서 일어날 수 있는 사건이나 상황을 묘사하는 마음의 표상을 말한다.

+문제가 발생하면 다시 문제 정의 단계로 돌아온다.



모이스처 플러스에 관한 CEO의 의견

모이스처 플러스의 고객은 10대 초반(11세부터 15세까지)의 여성입니다. 기본적으로 그녀들이 유일한 고객층입니다.

Acme 사는 광고에 들이던 비용을 소셜 네트워킹으로 전환해보고 있습니다만, 아직까지는 이 전략이 성공할지는 미지수입니다.

10대 초반의 여성들을 대상으로 매출을 증가시킬 가능성은 무한하다고 생각하고 있습니다. Acme 사의 경쟁사는 지극히 위협적입니다.

멘탈 모델

세계에 관한 여러분의 가정과 믿음은 여러분의 멘탈 모델입니다.

+세계는 복잡하기 때문에 우리는 이를 이해하기 위해 멘탈모델을 사용한다.
+뇌는 새로운 정보를 얻을 때마다 그 정보를 분석하기 위한 도구를 선택한다.
+멘탈 모델은 타고난, 내장된 인지 능력일 수도 있고 우리가 익혀온 이론일 수도 있다.
+멘탈 모델은 데이터 분석 방법에 큰 영향을 준다.
+멘탈 모델은 도움이 될 수 도 있지만 문제를 일으킬 수도 있다.
+가장 중요한 것은 항상 멘탈 모델을 명확하게 하고, 데이터를 다룰 때와 같이 진지하고 주의 깊게 다루 도록하는 것이다.

통계 모델은 멘탈 모델에 의해 결정됩니다.

+멘탈 모델에 따라 보이는 것이 다르다.

+모든 것을 볼 수 없기 때문에 뇌는 무엇인가에 중점을 두고 선택적으로 주의를 집중해야 한다. 따라서 멘탈 모델에 따라 분석가에게 보이는 것이 결정되는 것이다.

+멘탈 모델을 인식하고 있으면 무엇이 중요한지 이해하고, 가장 적절하고 유용한 통계 모델을 만들 수 있는 가능성이 좀 더 높아진다.

+통계 모델은 멘탈 모델에 따라 달라집니다. 잘못된 멘탈 모델을 사용하면 여러분의 분석은 시작하기 전부터 실패합니다. 따라서 올바른 멘탈 모델이 필요하다.

+고객이 완전히 잘못된 멘탈 모델을 사용하는 것은 드문 일이 아닙니다. 실제로, 사람들이 멘탈 모델의 가장 중요한 부분을 무시하는 것은 매우 일반적인 현상이다.

멘탈 모델에는 항상 여러분이 모르는 것을 포함해야 합니다.

+불확실한 것을 구체화 해야 한다.
+불확실한 것을 분명히 해 놓으면 지식의 격차를 해소하기 위해 어떻게 데이터를 사용할지에 대해 주의 깊게 살필 수 있을 뿐만 아니라 더 나은 권고안을 만들어 낼 수 있다.
+지식의 격차를 명확하게 하는 것이 필수적이다.
+사전에 불확실한 요소를 구체화해야 나중에 난처한 일을 당하는 것을 피할 수 있다.

CEO는 자신이 무엇을 모르는지 여러분에게 말합니다.

+맹점을 찾는 것이 중요하다.
+과거에는 직감을 이용해 해결했던 문제를 요새는 데이터 분석 기법을 사용해 해결할 수 있는 문제가 증가하고 있다.
+멘탈 모델 효과의 대부분은 모르는 것으로 인해 생기는 격차를 해소하는 데 도움이 된다.
+데이터 분석 도구를 사용하면 신념을 강화하는 체계적인 방법으로 이 격차를 줄일 수 있다.
+불확실성을 아주 상세하게 식별하는 작업에서 중요한 부분은 현실적이고 실증적인 데이터 분석이 요구되는 맹점을 찾아낼 수 있도록 하는 것.

Acme가 방대한 원시 데이터를 보냈습니다.

+새로운 데이터에 아무런 변형도 가하지 않았다면 받은 데이터가 원시 데이터이다.

+필요한 계산을 처리하기 쉬운 형식으로 변환하기 위해서는 원시데이터를 조작해야 하는 경우가 대부분이다.

+원시데이터는 작업이 이루어지는 곳과 분리해서 따로 보관해야 함.

Date	Vendor	Lot size (units)	Shipping ZIP	Cost
9/1/08	Sassy Girl Cosmetics	5253	20817	$75,643
9/3/08	Sassy Girl Cosmetics	6148	20817	$88,531
9/4/08	Prissy Princess	8931	20012	$128,606
9/14/08	Sassy Girl Cosmetics	2031	20817	$29,246
9/14/08	Prissy Princess	8029	20012	$115,618
9/15/08	General American Wholesalers	3754	20012	$54,058
9/20/08	Sassy Girl Cosmetics	7039	20817	$101,362
9/21/08	Prissy Princess	7478	20012	$107,683
9/25/08	General American Wholesalers	2646	20012	$38,102
9/26/08	Sassy Girl Cosmetics	6361	20817	$91,598
10/4/08	Prissy Princess	9481	20012	$136,526
10/7/08	General American Wholesalers	8598	20012	$123,811
10/9/08	Sassy Girl Cosmetics	6333	20817	$91,195
10/12/08	General American Wholesalers	4813	20012	$69,307
10/15/08	Prissy Princess	1550	20012	$22,320
10/20/08	Sassy Girl Cosmetics	3230	20817	$46,512
10/25/08	Sassy Girl Cosmetics	2064	20817	$29,722
10/27/08	General American Wholesalers	8298	20012	$119,491
10/28/08	Prissy Princess	8300	20012	$119,520
11/3/08	General American Wholesalers	6791	20012	$97,790
11/4/08	Prissy Princess	3775	20012	$54,360
11/10/08	Sassy Girl Cosmetics	8320	20817	$119,808
11/10/08	Sassy Girl Cosmetics	6160	20817	$88,704
11/10/08	General American Wholesalers	1894	20012	$27,274
11/15/08	Prissy Princess	1697	20012	$24,437
11/24/08	Prissy Princess	4825	20012	$69,480
11/28/08	Sassy Girl Cosmetics	6188	20817	$89,107
11/28/08	General American Wholesalers	4157	20012	$59,861
12/3/08	Sassy Girl Cosmetics	6841	20817	$98,510
12/4/08	Prissy Princess	7483	20012	$107,755
12/6/08	General American Wholesalers	1462	20012	$21,053
12/11/08	General American Wholesalers	8680	20012	$124,992
12/14/08	Sassy Girl Cosmetics	3221	20817	$46,382
12/14/08	Prissy Princess	6257	20012	$90,101
12/24/08	General American Wholesalers	4504	20012	$64,858
12/25/08	Prissy Princess	6157	20012	$88,661
12/28/08	Sassy Girl Cosmetics	5943	20817	$85,579
1/7/09	Sassy Girl Cosmetics	4415	20817	$63,576
1/10/09	Prissy Princess	2726	20012	$39,254
1/10/09	General American Wholesalers	4937	20012	$71,093
1/15/09	Sassy Girl Cosmetics	9602	20817	$138,269
1/18/09	General American Wholesalers	7025	20012	$101,160
1/20/09	Prissy Princess	4726	20012	$68,054

데이터를 한 단계 더 깊이 조사할 때입니다.

+데이터에 나와 있는 내용을 이용해, 더 깊이 데이터를 조사함.

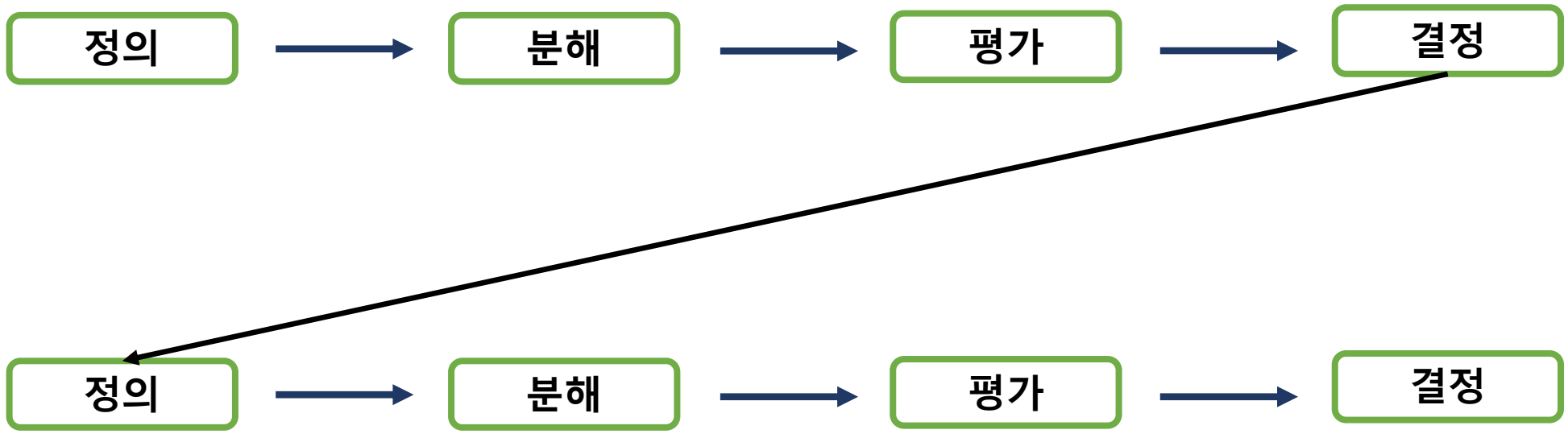
-GAW사는 여러분이 받은 인상을 확인시켜줬습니다.

-반드시 고객에게 사고 과정을 단계별로 차례차례 설명해야 한다.

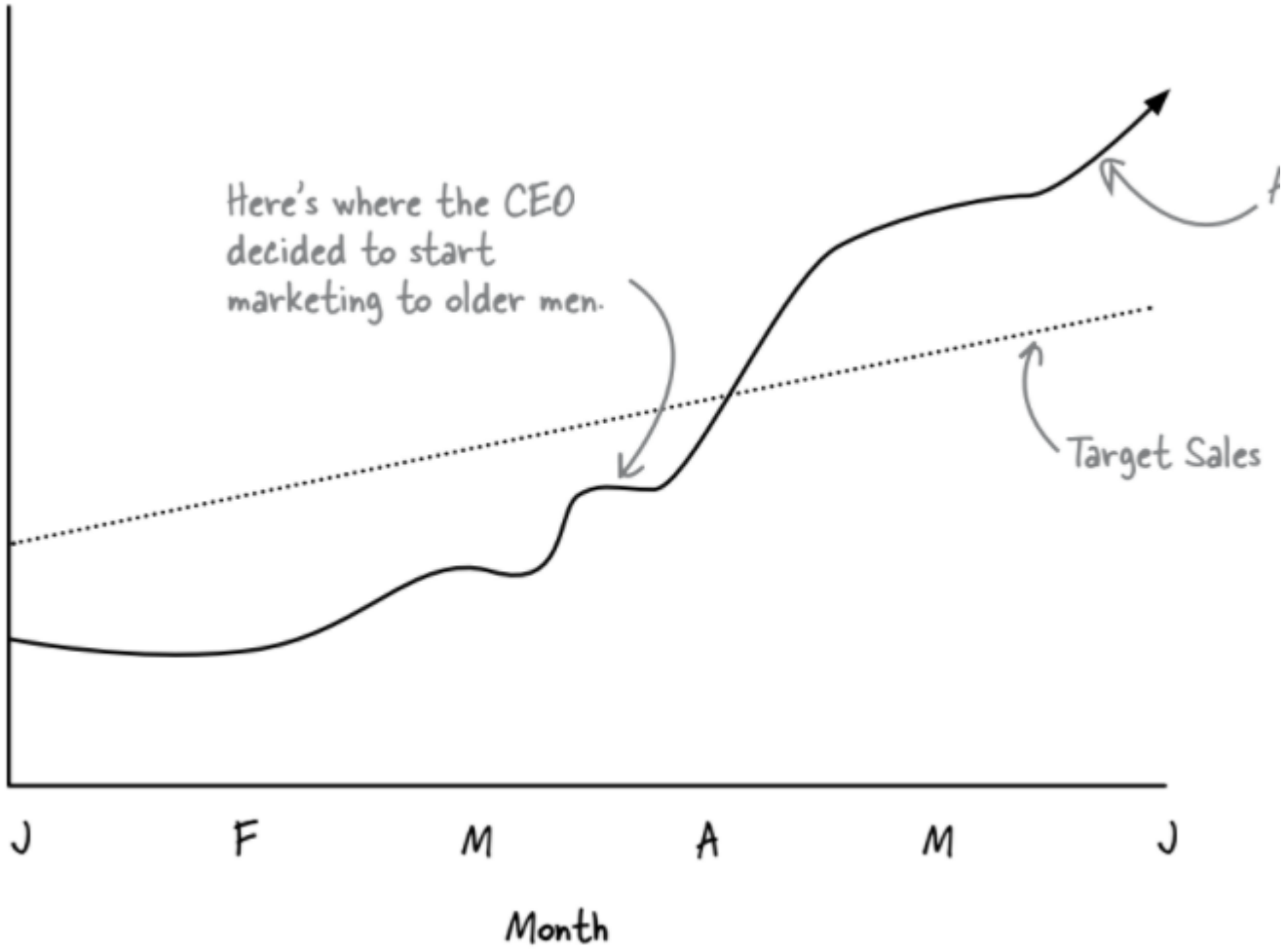
-더 새롭고 상세한 데이터를 어디까지 추적하는가 하는 것은 최선의 판단을 내리기 위한 근본적인 질문이다.

-모델이 처음부터 올바른 가정에 근거하고 있는지 확인하고 가정을 뒤집는 데이터가 있으면 바로 다시 검토하는 것이 매우 중요합니다.

여러분이 한 것들이에요...



여러분의 분석은 고객을 훌륭한 판단으로 이끌었습니다.



Thank you!