제 2 장 그림 이용한 자료 정리

1. 그림 이용한 자료 정리
2. 히스토그램 그리기
3. 혼동요인 통제: 따로따로 분석하기
4. Napoleon Army's Russian Invasion in 1812
5. 야구 통계

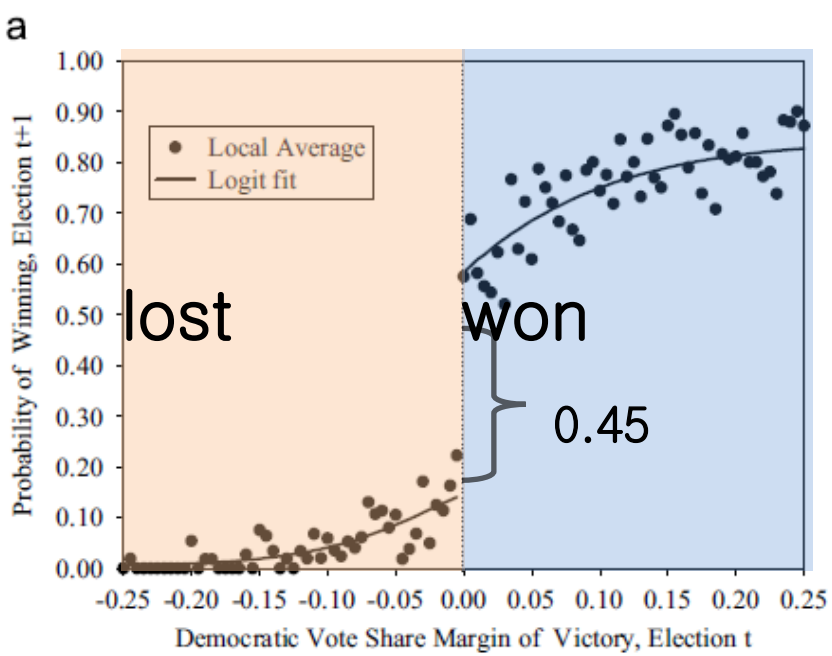
1. 그림 이용한 자료 정리

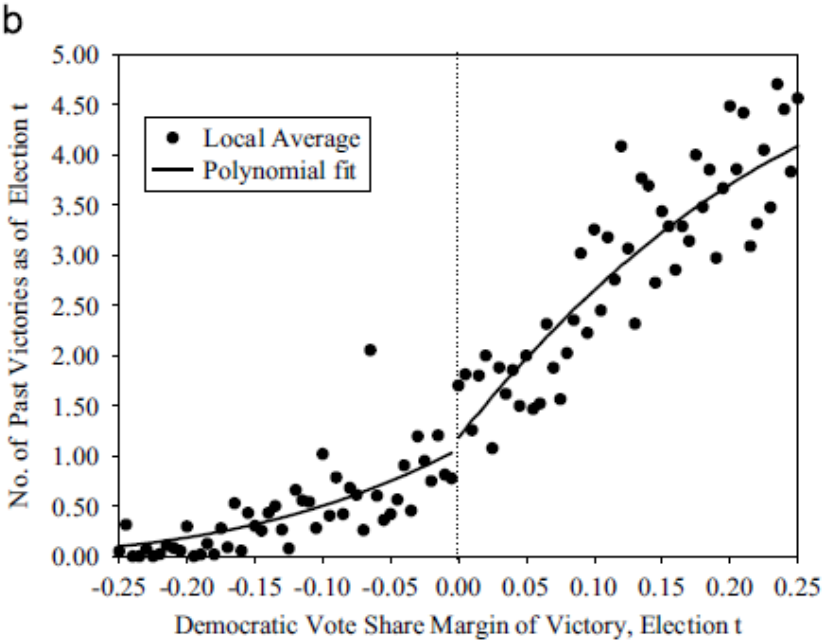
그임 이용한 자료 정리

악화가 양화를 구축!

1. 그림 이용한 자료 정리

Discontinuity only in “the future” but not in “the past”

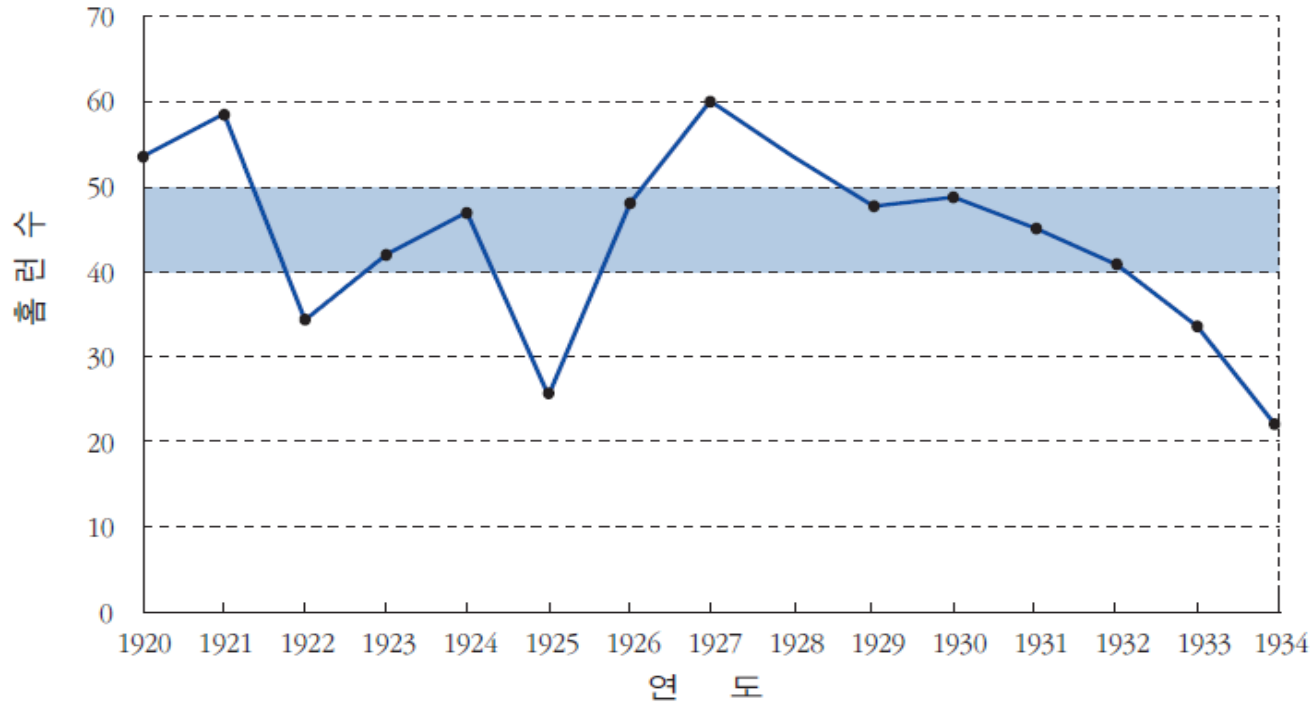




1. 그림 이용한 자료 정리

시계열 그림

그림 2-1 시계열 그림 : 베이브 루드의 연도별 홈런수



시계열 그림은 시간에 따른 자료의 변화나 추세를 파악하는데 적절

1. 그림 이용한 자료 정리

줄기-잎 그림

줄기-잎 그림 : 베이브 루드의 연간 홈런수(1925년 제외)

줄 기	잎
2	2
3	4 5
4	1 1 6 6 6 7 9
5	4 4 9
6	0

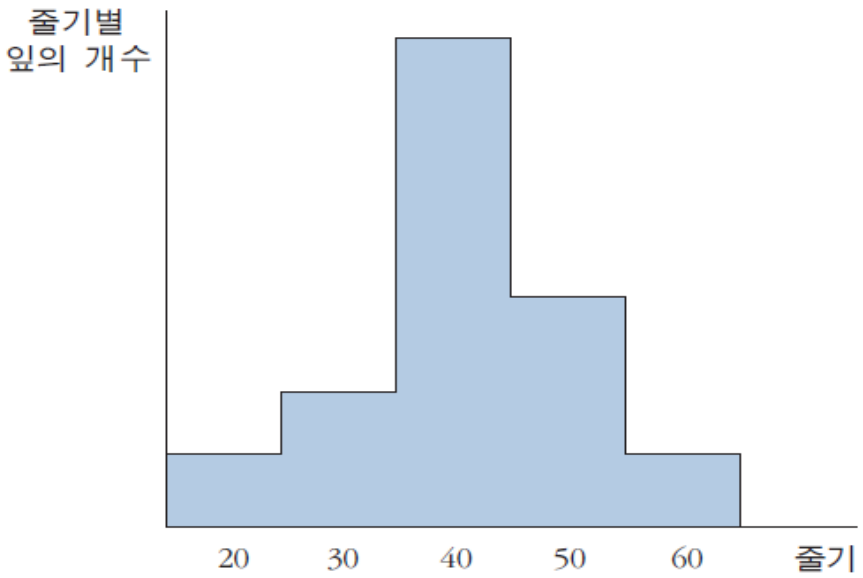
줄기는 10단위 숫자, 잎은 1단위 숫자

홈런을 41개 친 시즌이 두 번: 4의 줄기에 1의 잎이 두 번 등장

1. 그림 이용한 자료 정리

히스토그램

잎의 윤곽만 그린 그림 : 베이브 루드의 연간 홈런수(1925년 제외)

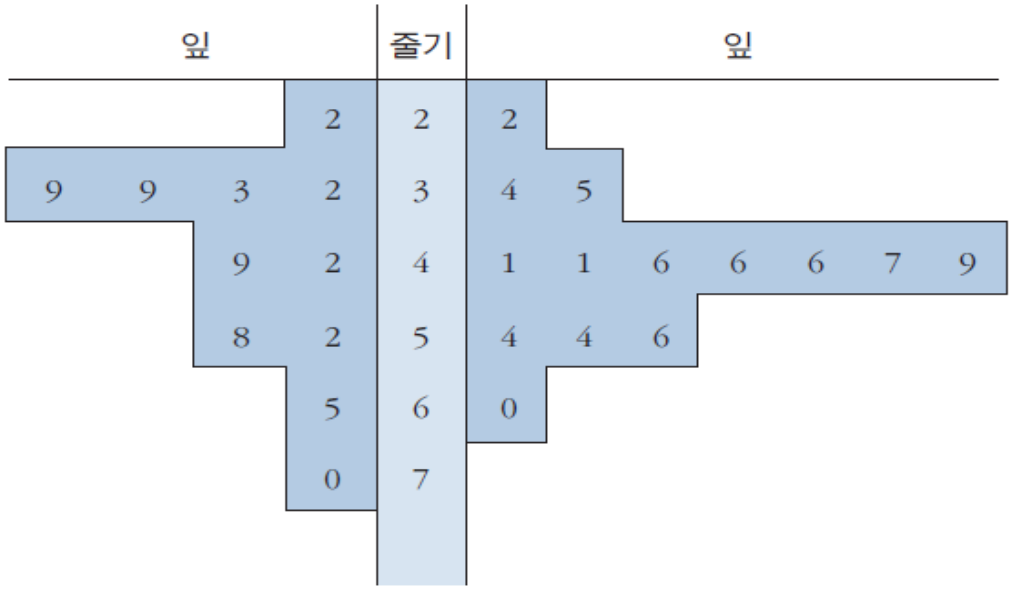


줄기-잎 그림에서 줄기 별로 잎의 윤곽만 그린 뒤, 그림을 시계 반대방향으로 90도 회전시킨 것이 히스토그램(histogram)임

1. 그림 이용한 자료 정리

겹-줄기-잎 그림

겹-줄기-잎 그림 : 마크 맥과이어와 베이브 루드의 연간 홈런수 비교

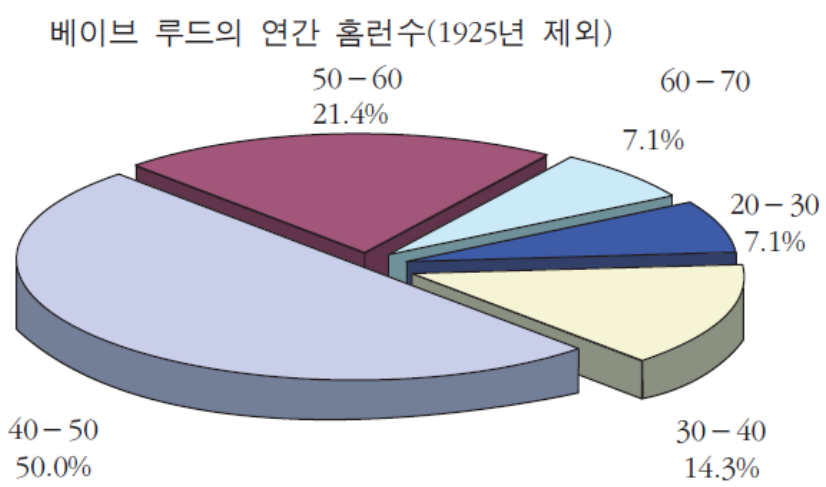


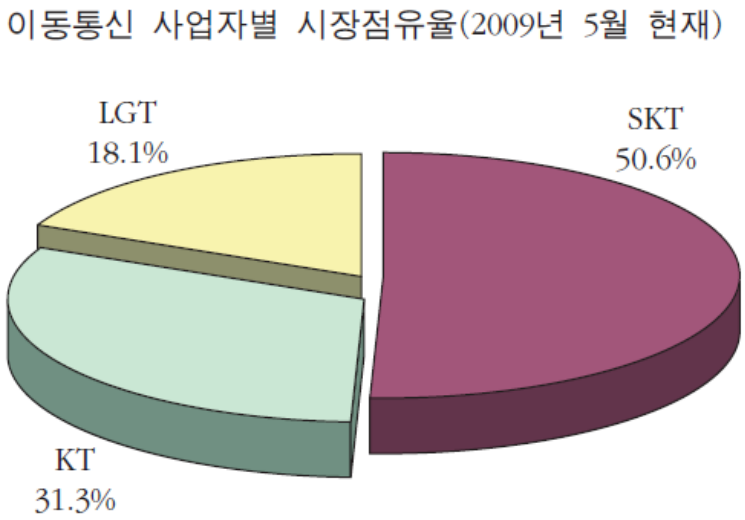
자료의 개수는 11 대 14로 맥과이어가 적음. 자료의 개수가 달라서 생기는 시각적 차이를 없애기 위해 맥과이어의 잎 하나는 베이브 루드의 잎 하나보다 14/11배 만큼 크게 그렸음 (좌우 두 줄기-잎 그림의 면적이 서로 같게 됨)

1. 그림 이용한 자료 정리

파이 도표

파이 도표





파이조각의 크기로 자료의 구성비 표시

2. 히스토그램 그리기

분포표 (distribution table) 작성

표 2-3 2007년도 우리나라 근로자 가구의 월소득 분포

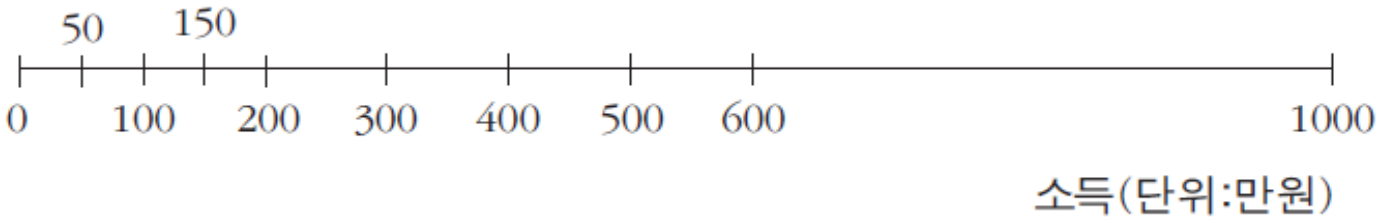
소득구간	비율(%)
0만원 – 50만원	1
50만원 – 100만원	4
100만원 – 150만원	7
150만원 – 200만원	9
200만원 – 300만원	22
300만원 – 400만원	21
400만원 – 500만원	14
500만원 – 600만원	8
600만원 –1000만원	14

주의: 반올림으로 인해 비율의 합이 정확히 100%가 되지는 않는다.
출처: 통계청 도시가계조사

2. 히스토그램 그리기

가로축 좌표 값

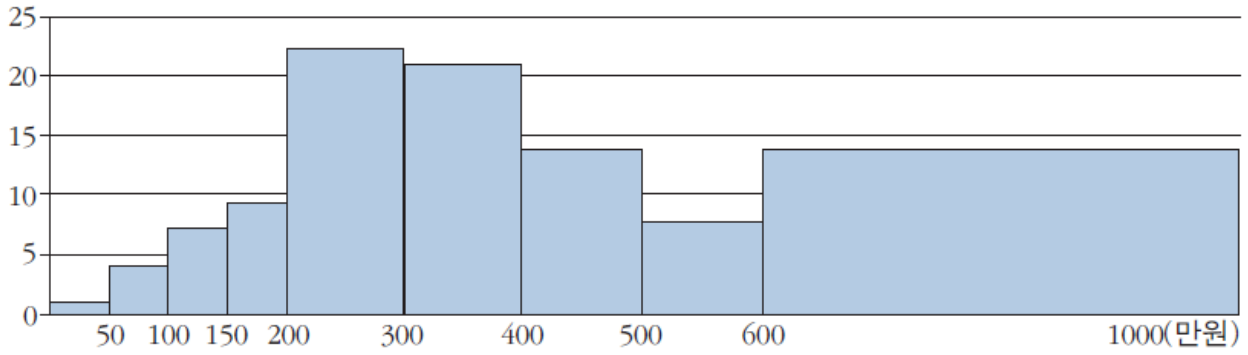
실제 구간의 폭에 비례하도록 나타낼 것



2. 히스토그램 그리기

따라 하지 말 것

그림 2-7 구간에 속한 자료의 비율로 높이를 삼지 마라

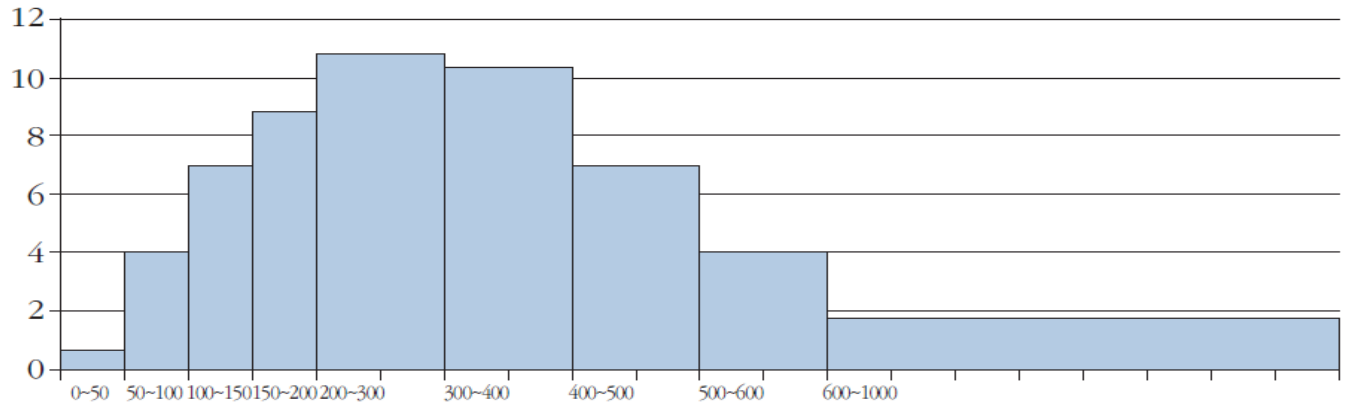


예) 비율로 높이 삼는 경우 예컨대 600-1000(만원) 블록의 면적이 너무 커진다.

2. 히스토그램 그리기

따라 할 것

그림 2-8 2007년 근로자 가구의 월소득 분포



계급구간별로 비율을 폭으로 나누어 해당 블록의 높이로 삼는다.

블록의 면적이 해당 계급구간에 속한 자료의 비율을 나타내게 된다.

예: 월소득 150만원 이상 200만원 미만의 가구는 전체의 9%

2. 히스토그램 그리기

밀도 단위 (density scale)

히스토그램에서 블록의 높이는 밀도(density), 즉 가로축의 단위구간에 속한 자료의 비율을 나타낸다.

세로축에 밀도 단위가 사용되는 경우, 블록의 면적은 해당 구간에 속하는 자료의 비율을 나타내고 히스토그램 아래 전체 블록의 면적은 100%가 된다.

3. 혼동요인 통제: 따로따로 분석하기

변수의 통제

혼동요인으로 인한 결과의 왜곡을 막으려면 이에 대한 통제가 필요

어떻게?

- 자료를 혼동요인에 따라 세부집단으로 분류하고 세부집단 별로 따로따로 분석
- 세부집단이 많을 경우 회귀분석을 이용

경구 피임약의 복용이 해당 여성의 혈압을 높이는가?

- 연령이 혼동 요인으로 작용: 연령이 높아지면 혈압 올라감. 피임약 덜 복용함
- 연령집단 별로 복용자와 비복용자의 혈압을 따로따로 비교

3. 혼동요인 통제: 따로따로 분석하기

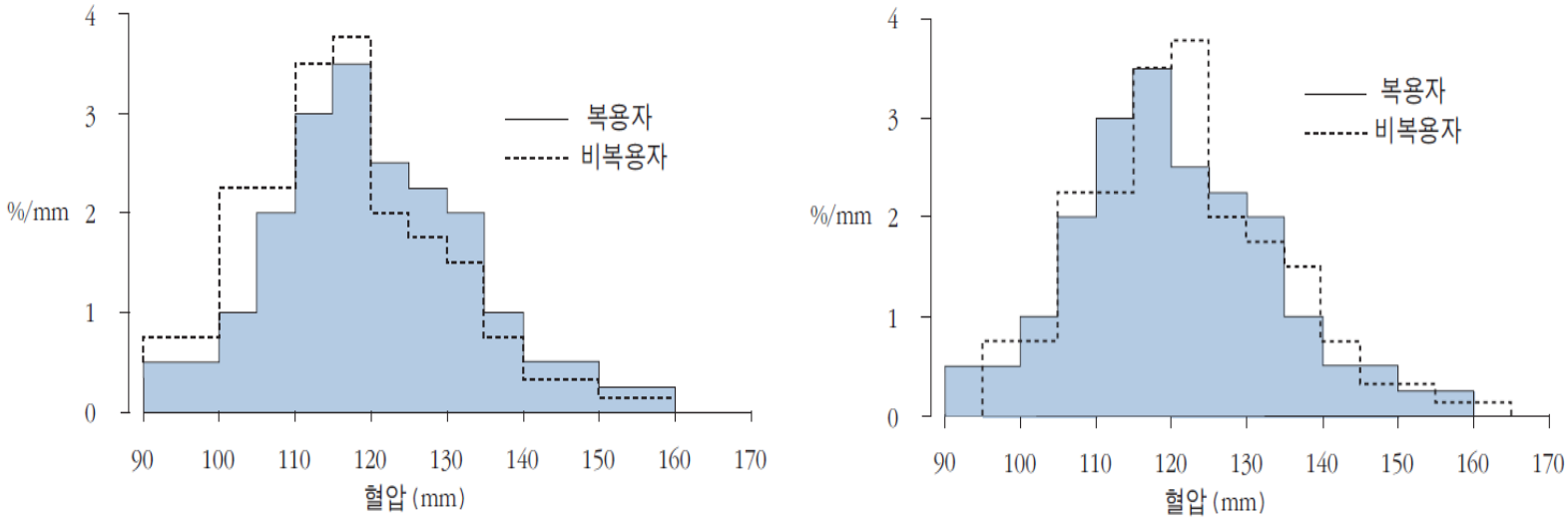
연령 별로 경구 피임약 복용 여부가 혈압에 미치는 영향

표 2-4 연령과 경구피임약의 복용 여부가 혈압에 미치는 영향

혈압	17-24세		25-34세		35-44세		45-58세	
(mm)	비복용자	복용자	비복용자	복용자	비복용자	복용자	비복용자	복용자
90 미만	–	1	1	–	1	1	1	–
90-95	1	–	1	–	2	1	1	1
95-100	3	1	5	4	5	4	4	2
100-105	10	6	11	5	9	5	6	4
105-110	11	9	11	10	11	7	7	7
110-115	15	12	17	15	15	12	11	10
115-120	20	16	18	17	16	14	12	9
120-125	13	14	11	13	9	11	9	8
125-130	10	14	9	12	10	11	11	11

3. 혼동요인 통제: 따로따로 분석하기

경구 피임약 복용 여부가 혈압에 미치는 영향 (25-34세 여성)



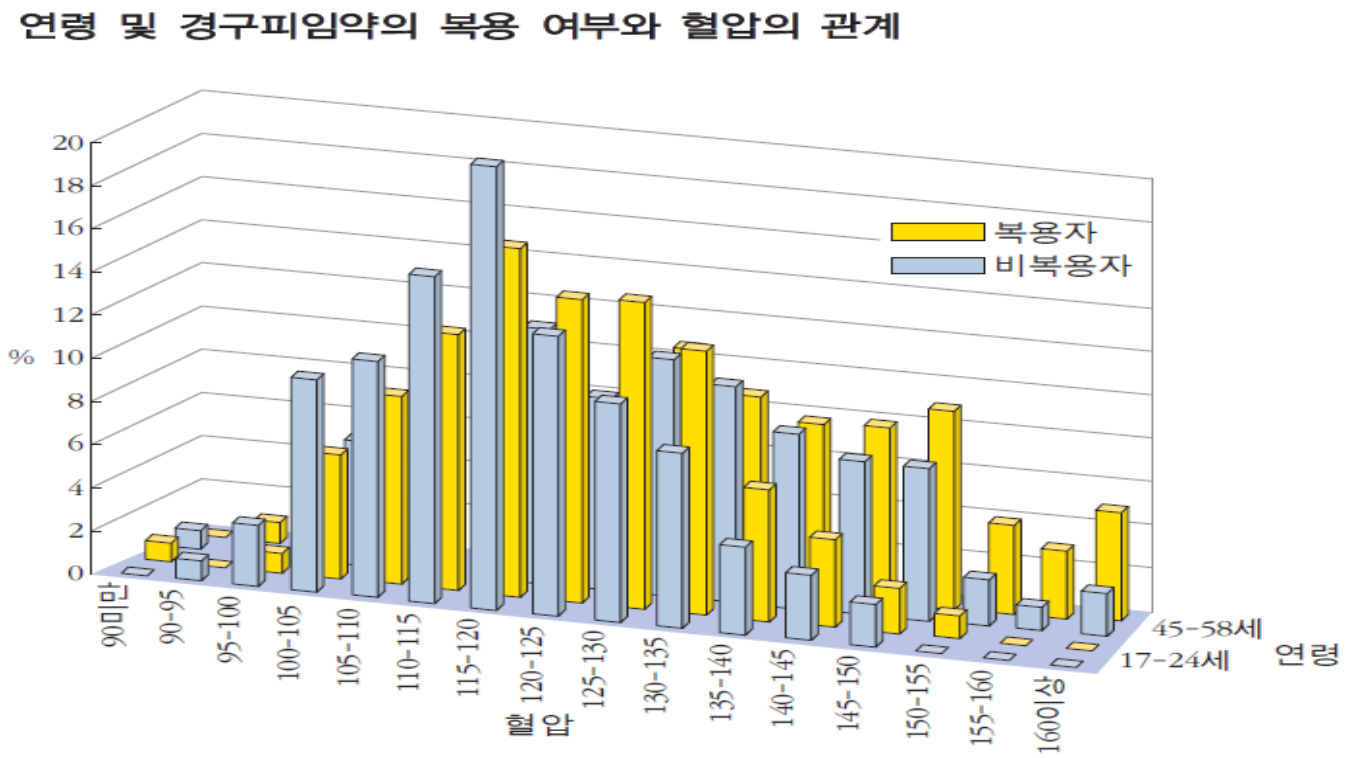
좌상단 그림을 보면 혈압 120을 기준으로 복용자가 비복용자보다 우측은 높고, 좌측은 낮음, 이는 복용자 집단의 혈압이 대체로 높다는 의미.

우상단 그림은 좌상단 그림에서 복용자의 혈압 분포는 그대로 두고 비복용자의 혈압만 5mm만큼 우측으로 이동시킨 분포임. 두 히스토그램이 대략적으로 일치.

경구 피임약 복용이 25-34세 여성의 혈압을 약 5mm 정도 상승시킨다는 결론 도출.

3. 혼동요인 통제: 따로따로 분석하기

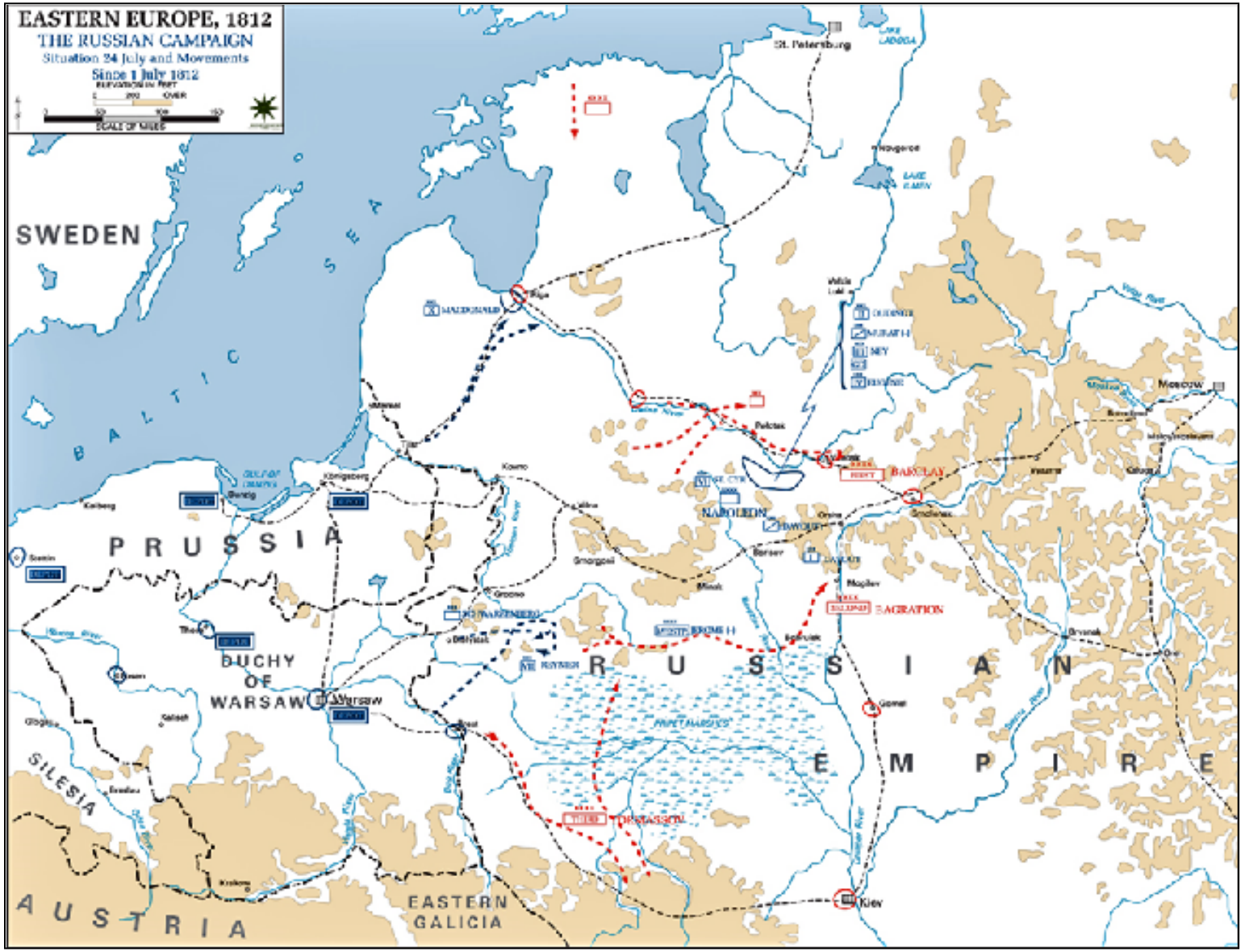
연령 및 경구 피임약 복용 여부와 혈압의 관계



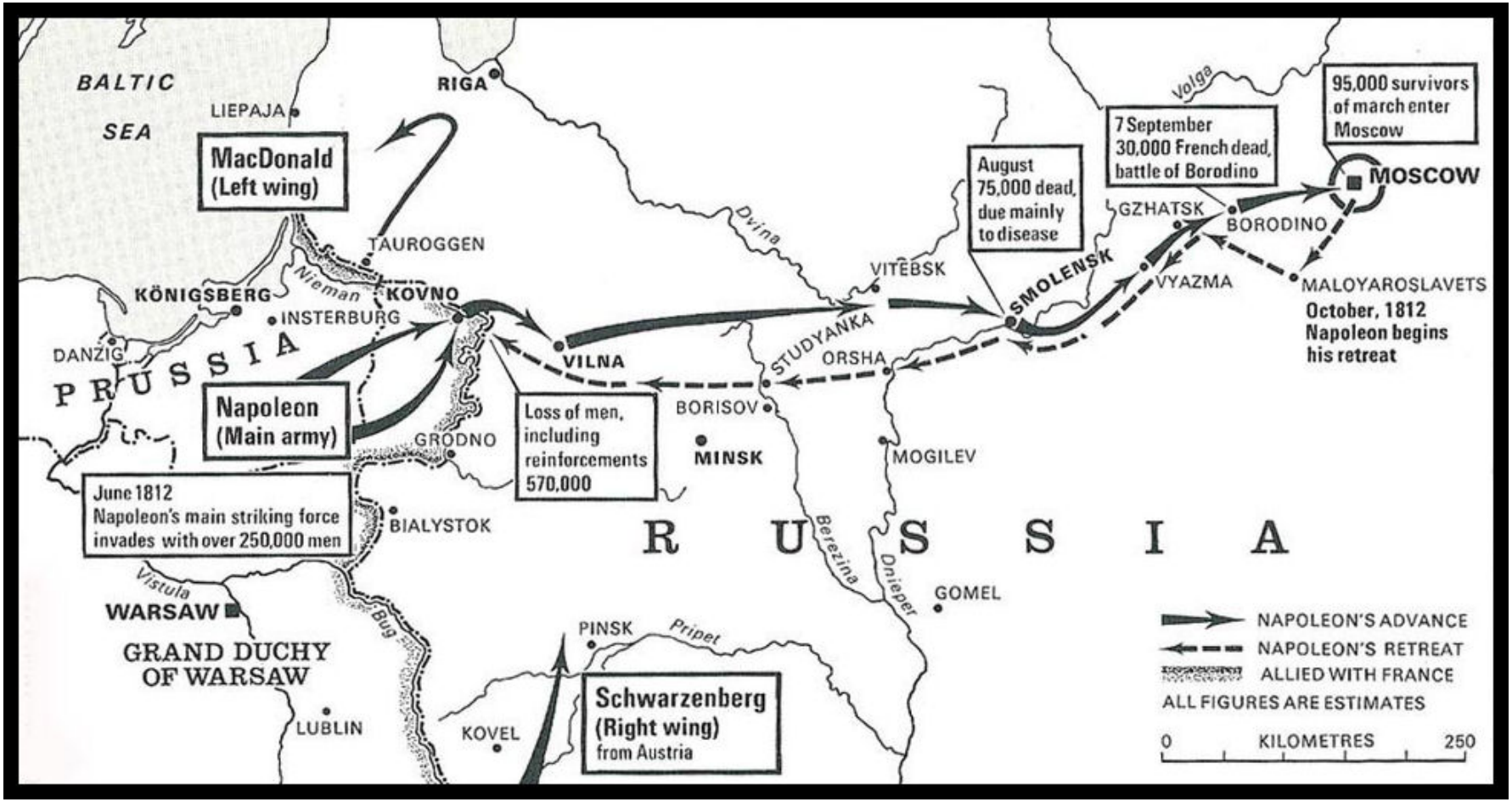
연령별로 복용자와 비복용자의 혈압분포 비교 가능함.

나아가 복용자, 비복용자 별로도 연령이 혈압분포에 미치는 효과 파악 가능함

4. Napoleon Army's Russian Invasion in 1812

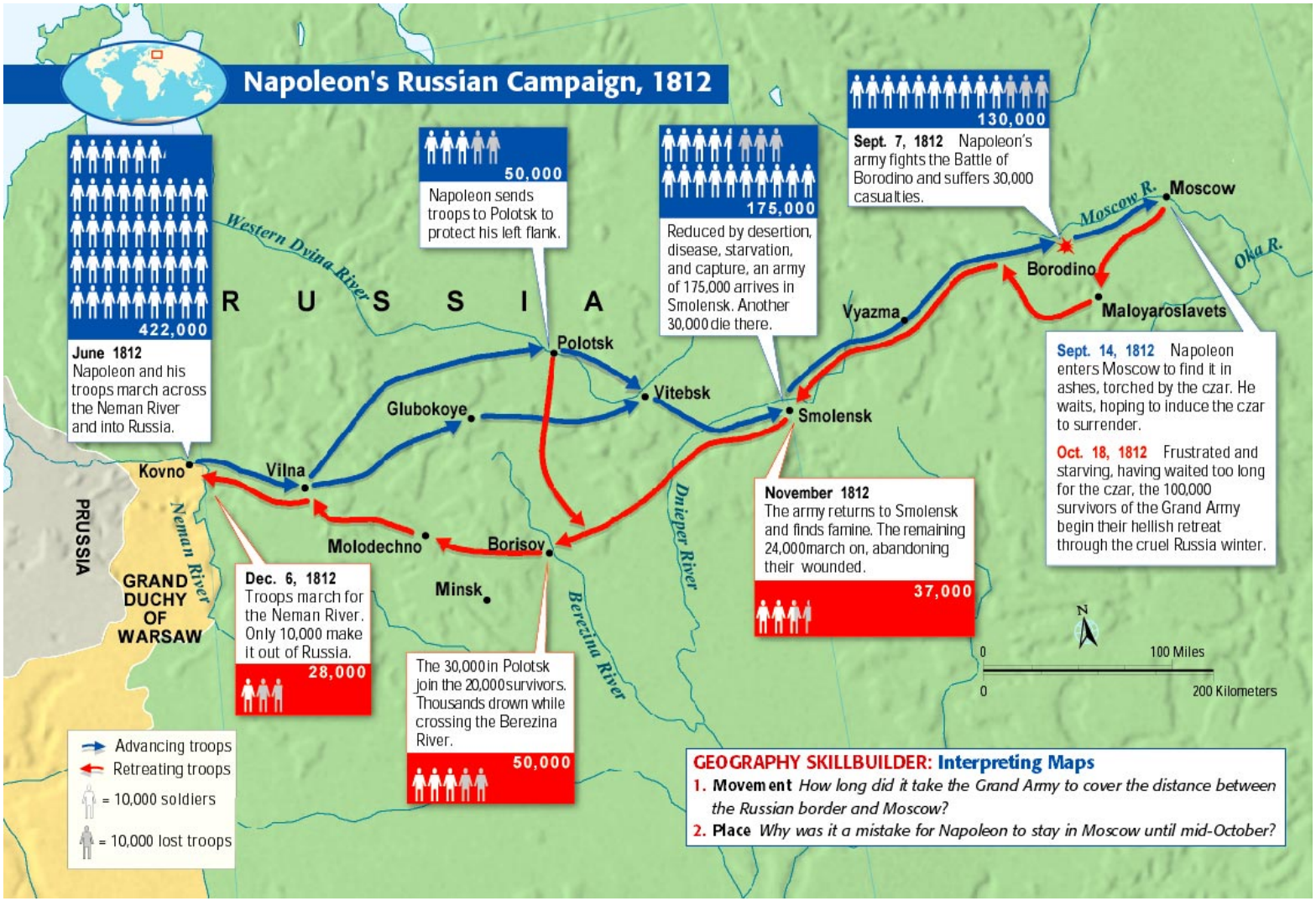


4. Napoleon Army's Russian Invasion in 1812

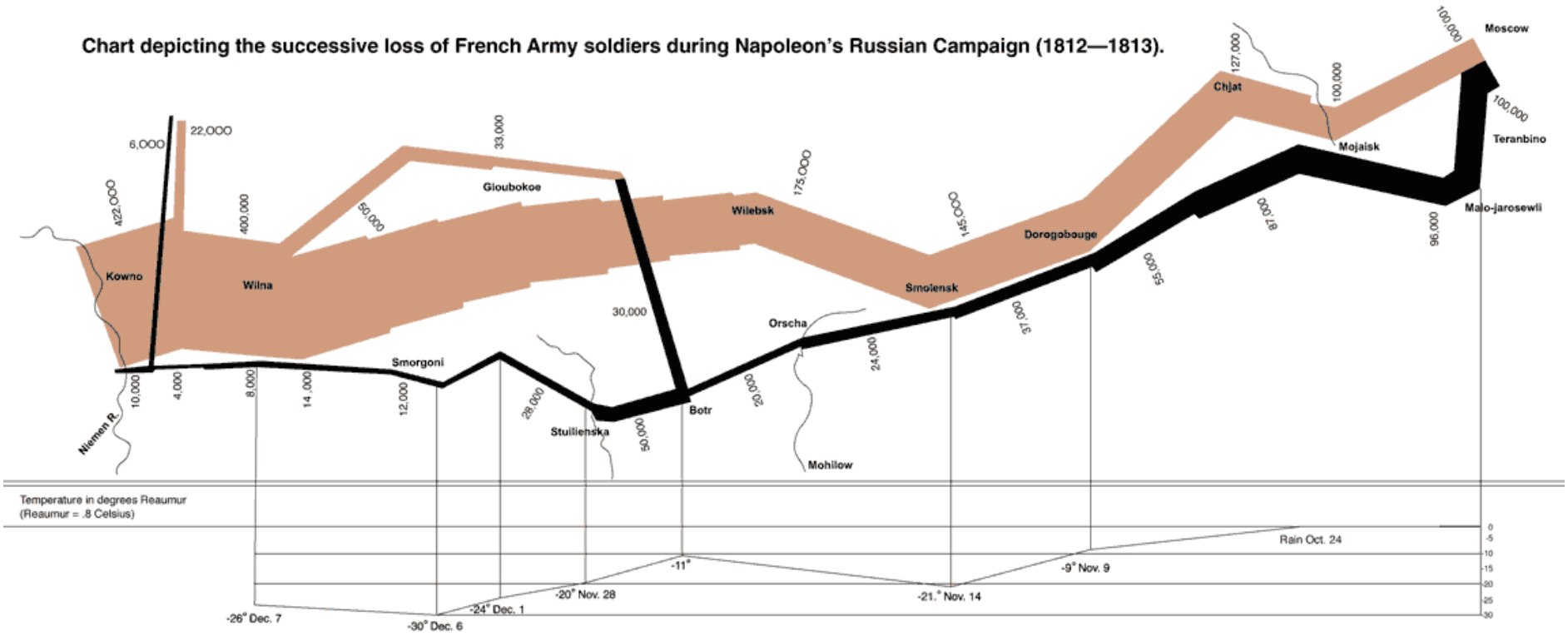


4. Napoleon Army's Russian Invasion in 1812

4. Napoleon Army's Russian Invasion in 1812



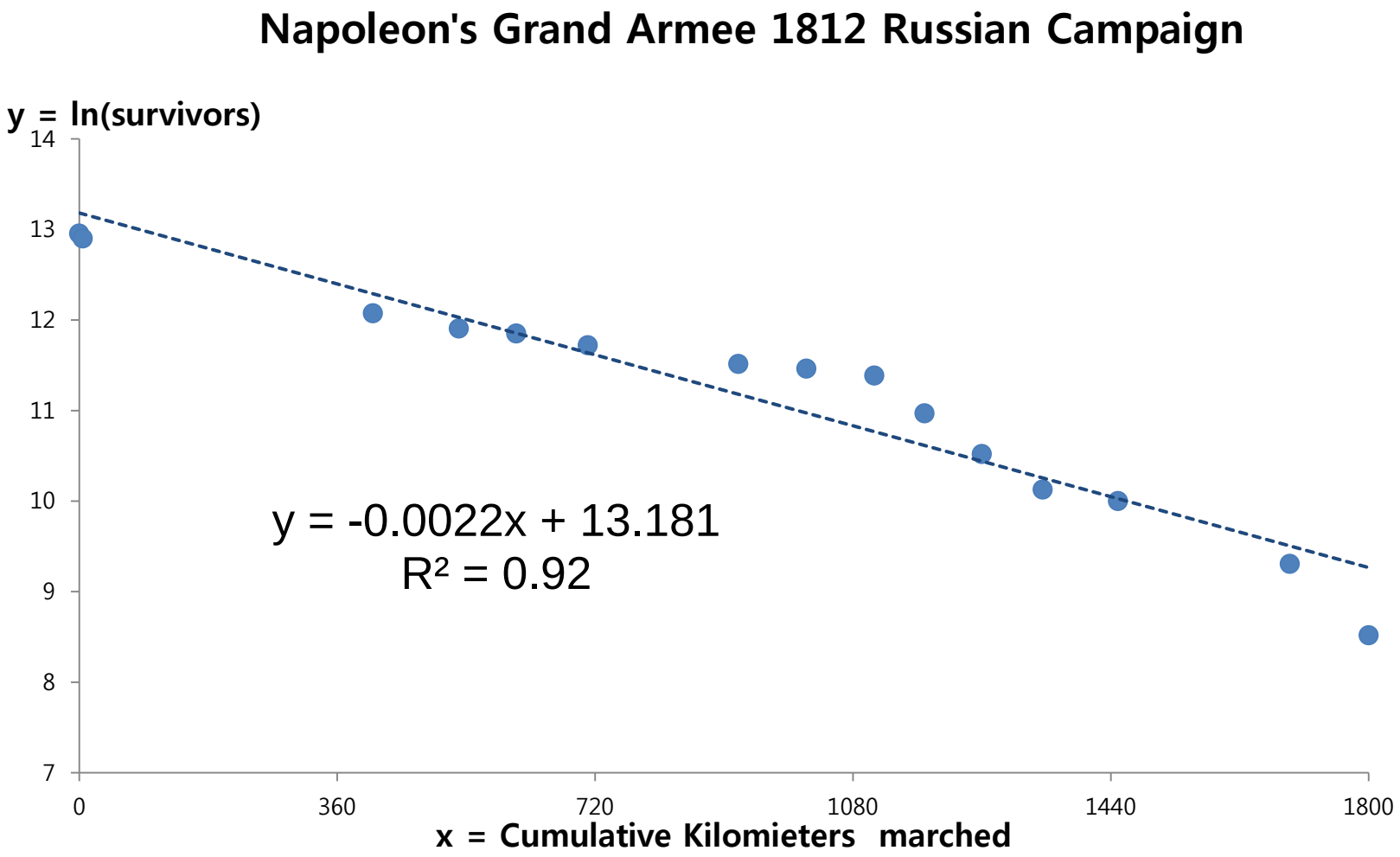
4. Napoleon Army's Russian Invasion in 1812



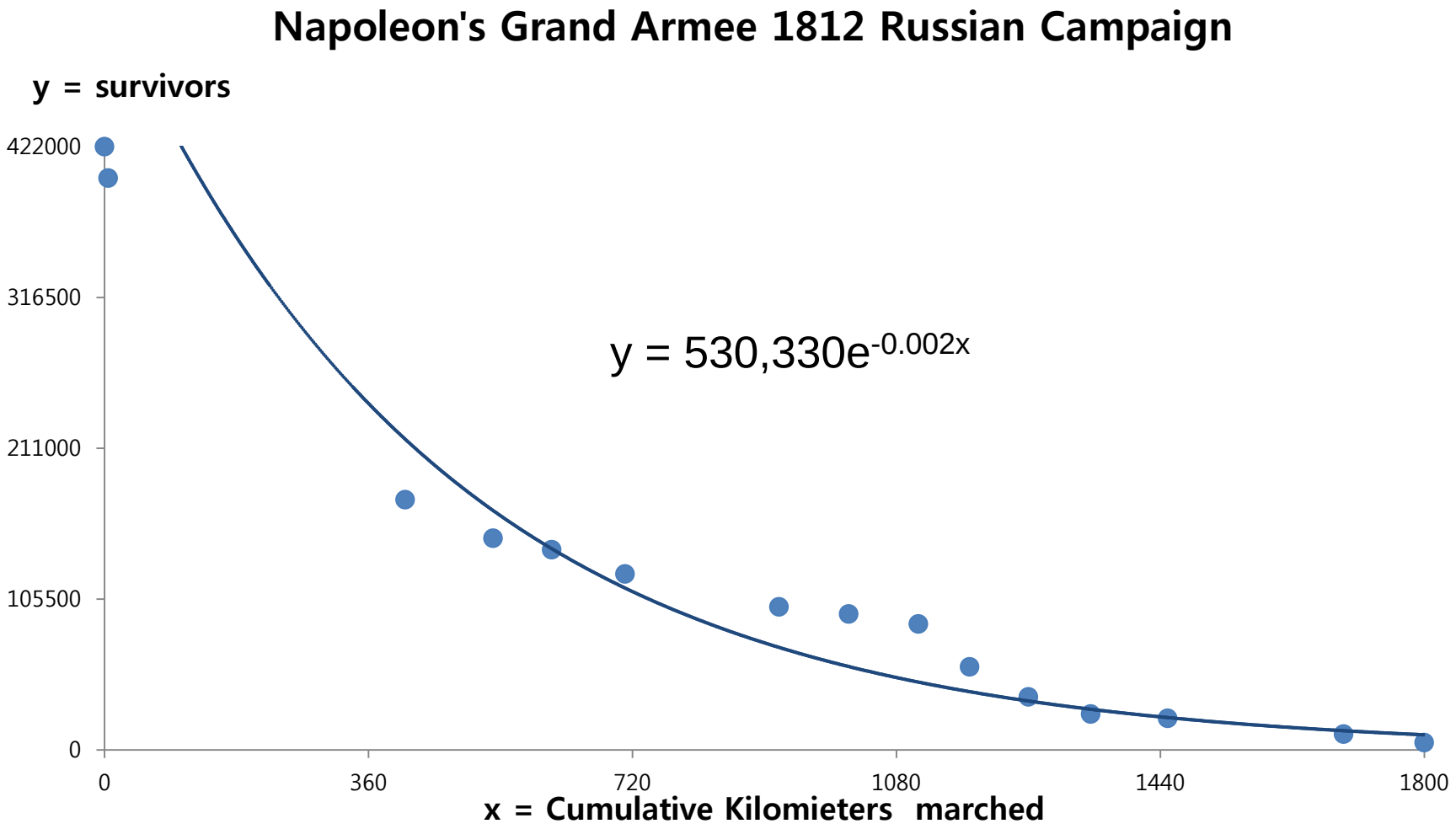
4. Napoleon Army's Russian Invasion in 1812

4. Napoleon Army's Russian Invasion in 1812

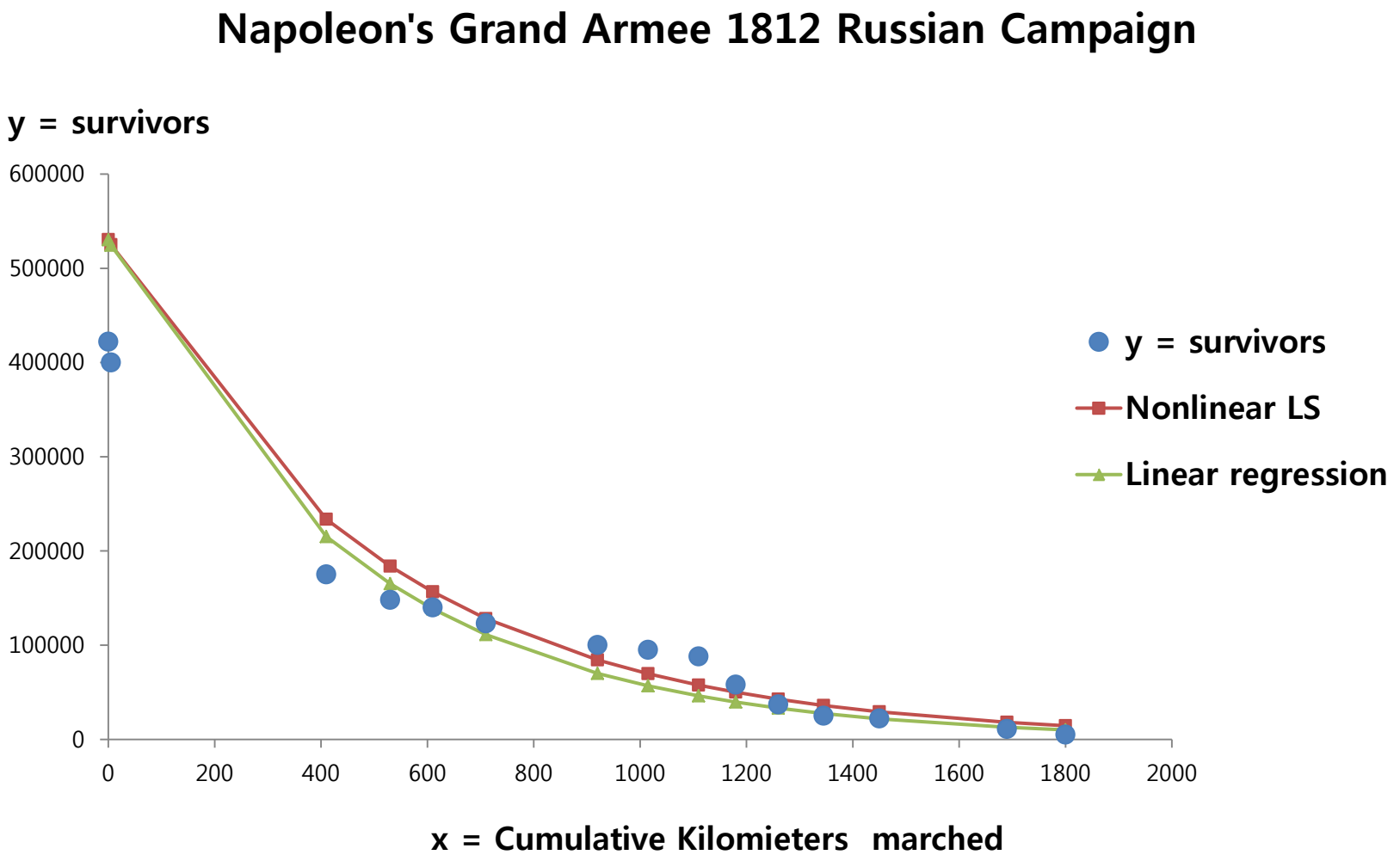
4. Napoleon Army's Russian Invasion in 1812



4. Napoleon Army's Russian Invasion in 1812

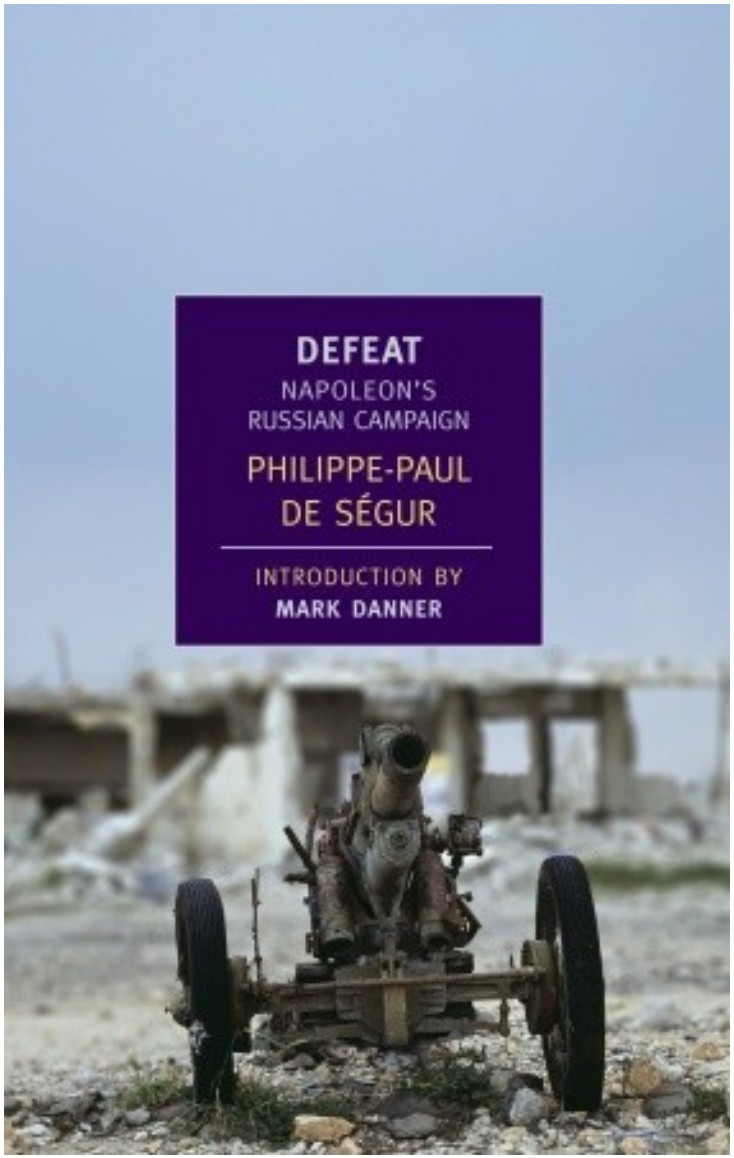


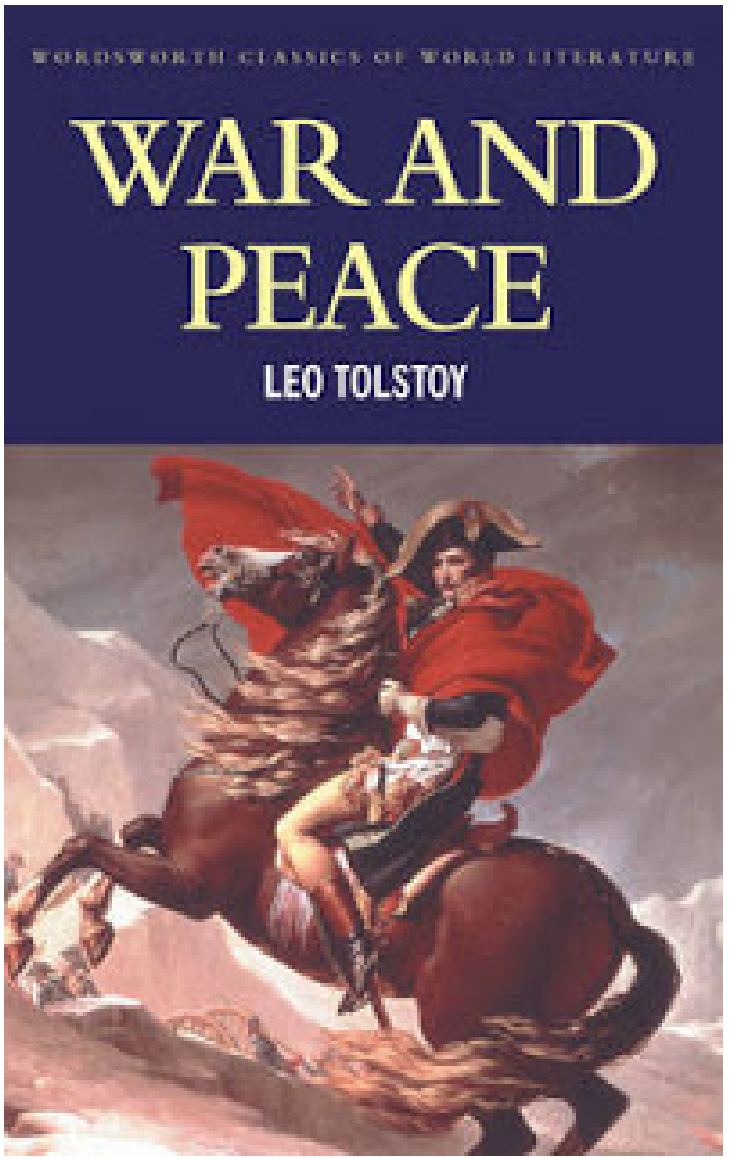
4. Napoleon Army's Russian Invasion in 1812



4. Napoleon Army's Russian Invasion in 1812

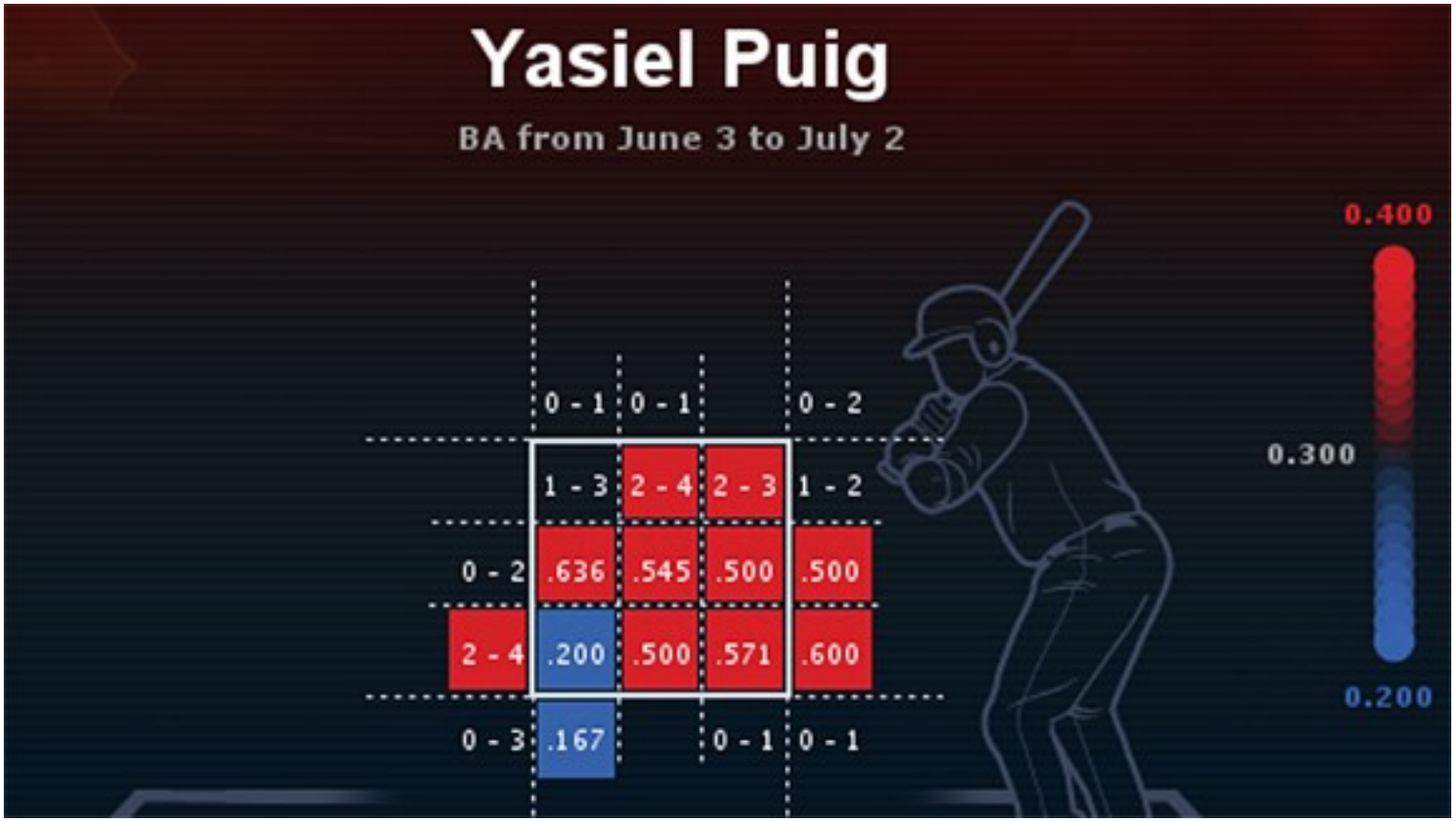
4. Napoleon Army's Russian Invasion in 1812





5. 야구 통계: Y. Puig 2013, LA Dodgers

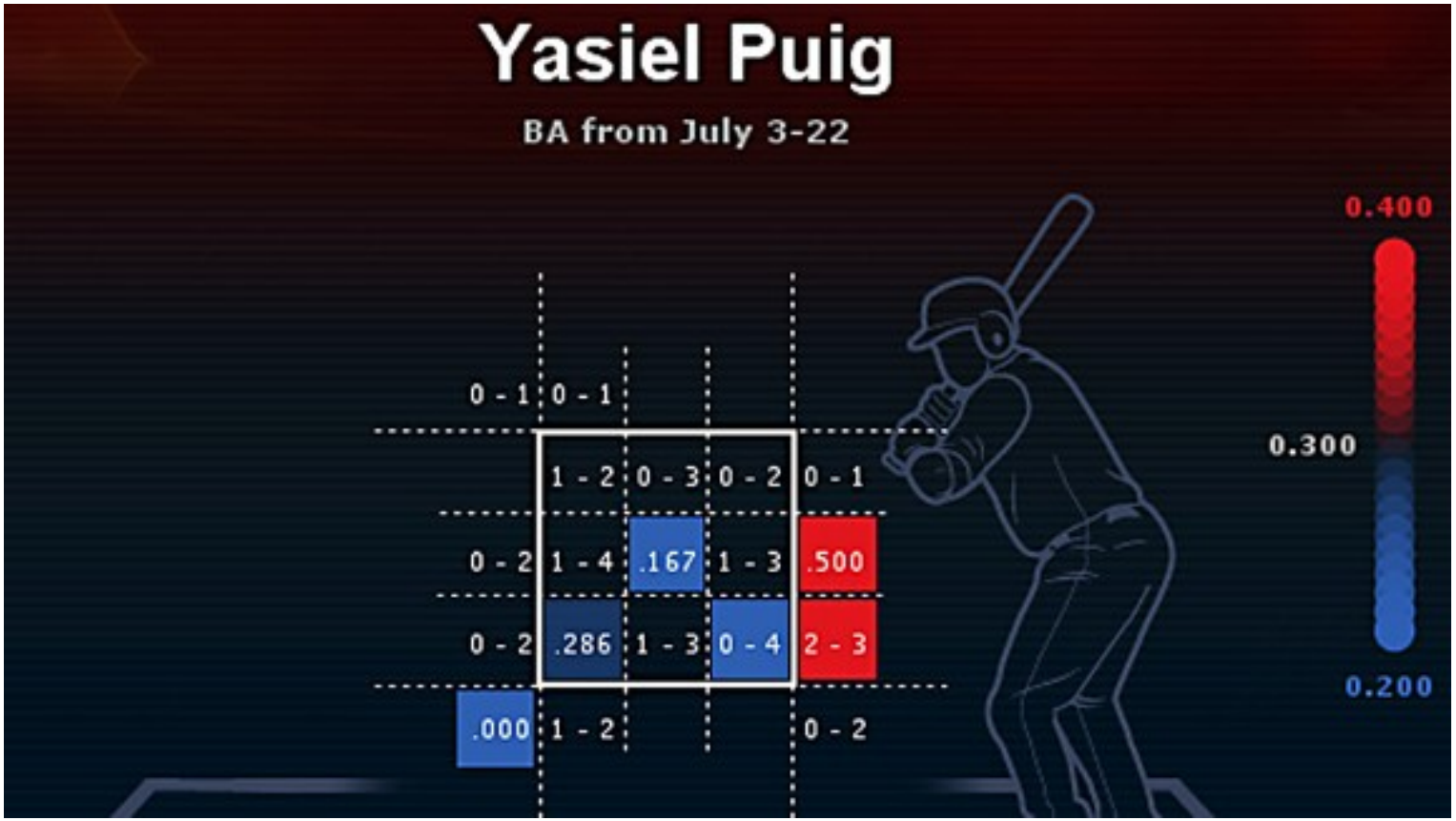
The Explosive Start: June 3–July 2 (27 경기: 타율=.443, 출루율=.473)



*타율=H/AB, 출루율=(H+BB+HBP)/(AB+BB+HBP+SF)

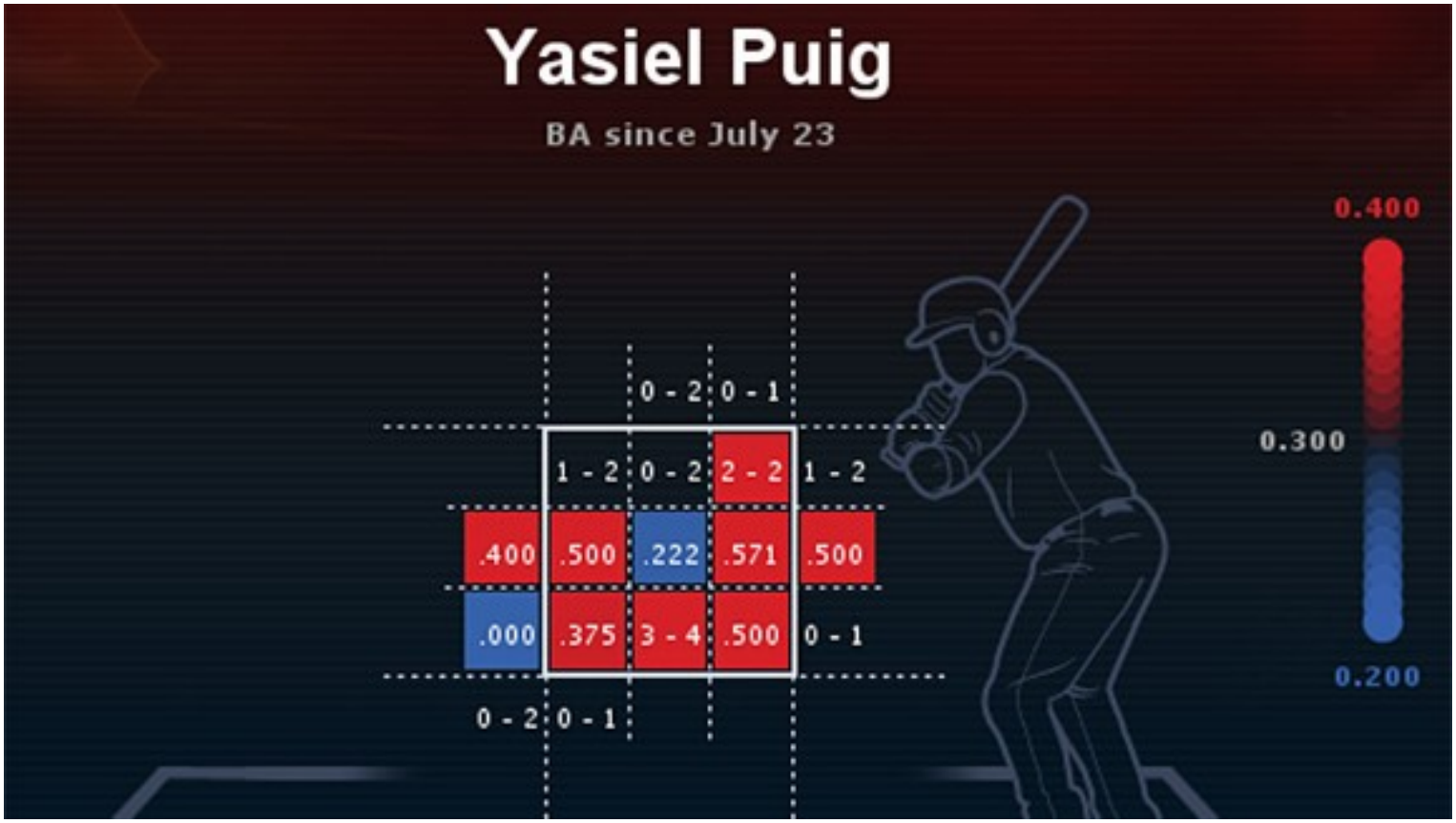
5. 야구 통계: Y. Puig 2013, LA Dodgers

The Slump: July 3–July 22 (15 경기: 타율=.220, 출루율=.266)



5. 야구 통계: Y. Puig 2013, LA Dodgers

The Adjustments: July 23~Aug. 18 (21 경기: 타율=.377, 출루율=.490)



5. 야구 통계: 미국 MLB 주심의 인종 편견

Discrimination in Baseball: MLB Umpires, Racial Bias and Calling Strikes

C. Parsons, J. Sulaeman, M.Yates, and D. Hamermesh (2011), "Strike Three: Discrimination, Incentives, and Evaluation," *American Economic Review, 101, 1410-1435*

Monitoring:
(1). QuesTec, a computerized monitoring system intended to evaluate the accuracy and consistency of home-plate umpires' judgments. From 2004-2008, QuesTec had been installed in 11 of MLB's 30 ballparks (see Fig 1 A, B)
(2) Attendance (see Fig 2 A, B)

5. 야구 통계: 미국 MLB 주심의 인종 편견

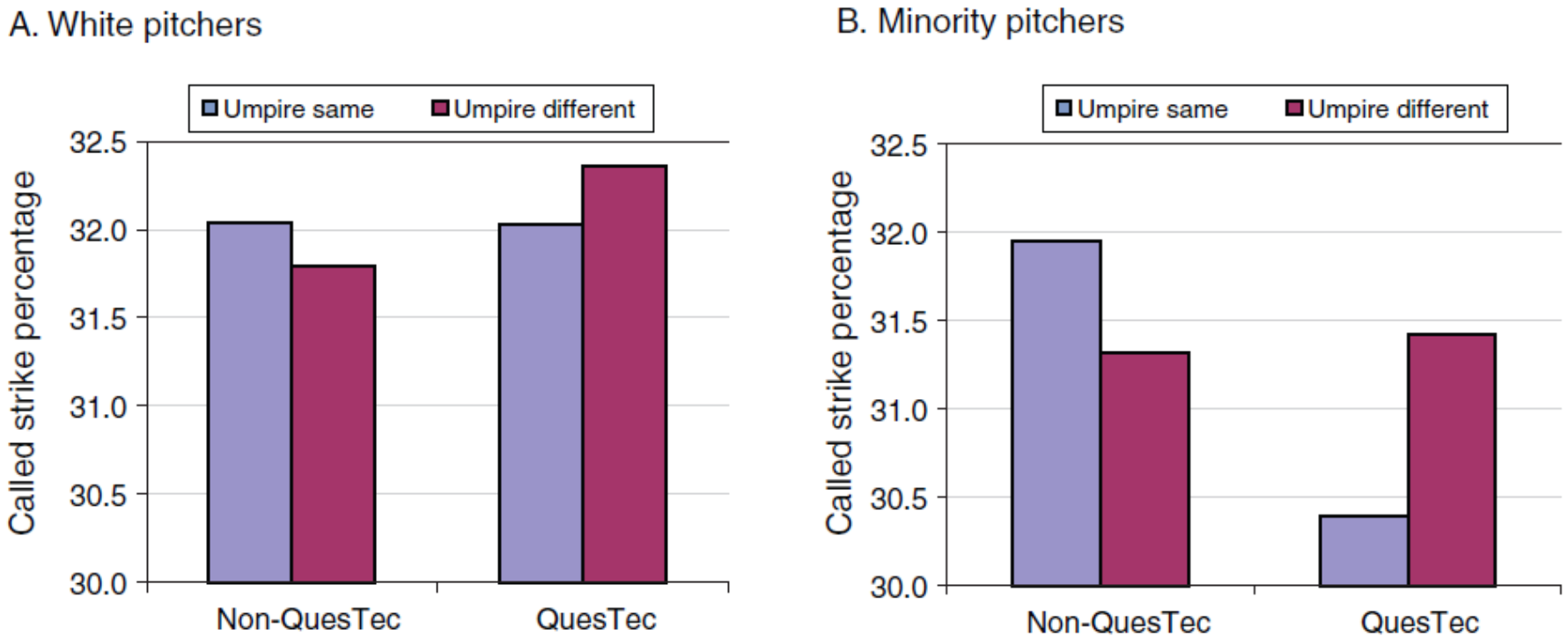


FIGURE 1. RACE AND CALLED STRIKE PERCENTAGES IN QUESTEC AND NON-QUESTEC BALLPARKS

5. 야구 통계: 미국 MLB 주심의 인종 편견

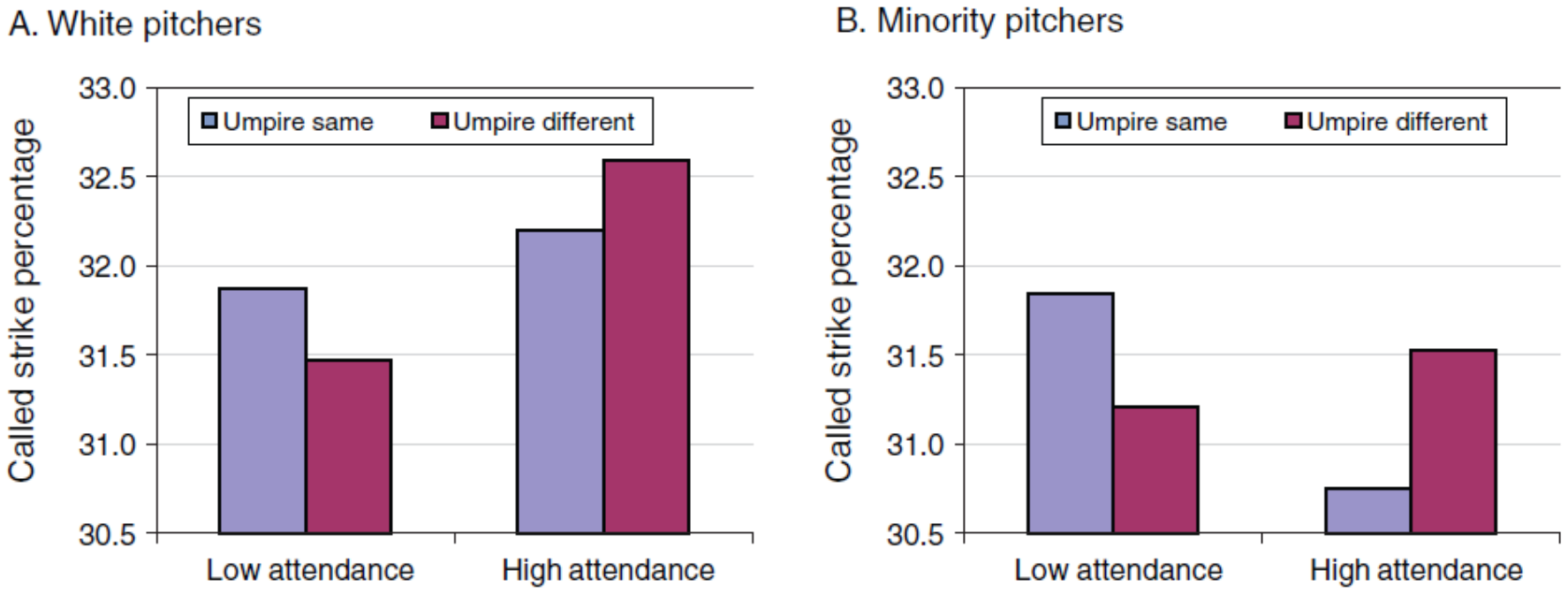


FIGURE 2. RACE AND CALLED STRIKE PERCENTAGES BY GAME ATTENDANCE

Note: Low (high) attendance games are those with percentage attendance below (above) the median.