# Part 2: 벤더 퀄리티 검증하기

## 과제 목표

수출 데이터 벤더가 주장하는 가장 빠른 데이터 제공이 사실인지 통계적으로 검증합니다.

**벤더의 주장:**

- 관세청 데이터 마감 직후, 매월 1일 오전 10시 20분에 데이터를 제공합니다
- 이 데이터는 독점적이며, 시장에서 가장 빠르게 제공됩니다
- 주말에도 데이터를 업데이트합니다

**핵심 질문:** 만약 이 주장이 사실이라면, 10시 20분 전후로 주가 움직임에 명확한 차이가 나타나야 합니다.

---

## 사전 지식: $R^2$ 와 p-value 이해하기

통계적 검증을 위해 두 가지 핵심 개념을 먼저 이해해야 합니다.

### $R^2$ (결정계수, Coefficient of Determination)

**정의:** $R^2$는 독립변수(X)가 종속변수(Y)의 변동을 얼마나 설명하는지를 0과 1 사이의 값으로 나타냅니다.

**직관적 이해:** 학생들의 성적을 예측한다고 가정합시다. 공부 시간(X)으로 성적(Y)을 예측하는 모델을 만들었습니다. $R^2 = 0.7$이라면, 성적 변동의 70%가 공부 시간으로 설명되고, 나머지 30%는 다른 요인(타고난 재능, 건강 상태, 운 등)에 의한 것입니다.

**공식의 의미:** $R^2 = 1 - (\text{잔차 제곱합} / \text{총 제곱합})$

여기서 잔차는 실제값과 예측값의 차이입니다. 모델이 완벽하면 잔차가 0이고 $R^2 = 1$입니다. 모델이 전혀 쓸모없으면 $R^2 = 0$입니다.

**해석 기준:**

- $R^2 = 0.00$: X가 Y를 전혀 설명하지 못함
- $R^2 = 0.10$: X가 Y 변동의 10%를 설명 (금융에서는 의미있는 수준)
- $R^2 = 0.25$: 중간 수준의 설명력
- $R^2 = 0.50$: 높은 설명력

- $R^2 = 1.00$: 완벽한 설명 (현실에서는 거의 불가능)

**금융에서의 특수성:** 금융 데이터는 노이즈가 매우 많습니다. 주가는 수천 가지 요인에 영향을 받기 때문에, 단일 변수로 $R^2 > 0.05$만 나와도 의미있는 결과입니다. 금융에서 $R^2 = 0.01$이라도 실무에서는 중요하게 다루어질 수 있습니다.

**우리 과제에서의 의미:**

- X = 수출 서프라이즈 (rolling Z-score)
- Y = 09:00~10:20 구간 수익률

만약 $R^2$이 높다면, 수출 서프라이즈가 이미 장 초반 수익률에 반영되고 있다는 의미입니다. 이는 벤더의 주장(10:20에 처음 공개)이 거짓일 가능성을 시사합니다.

## p-value (유의확률)

**정의:** p-value는 귀무가설이 참일 때, 현재 관측된 결과 또는 그보다 극단적인 결과가 나올 확률입니다.

**직관적 이해:** 동전을 10번 던졌는데 앞면이 9번 나왔습니다. 이 동전이 공정한 동전일까요?

귀무가설: 이 동전은 공정하다 (앞면 확률 50%) 관측 결과: 10번 중 9번 앞면

p-value는 공정한 동전으로 10번 던져서 9번 이상 앞면이 나올 확률을 계산합니다. 이 확률이 매우 낮다면 (예: 0.01), 우리는 귀무가설을 의심하게 됩니다. 이 동전은 아마 공정하지 않을 것이다.

**통계적 가설 검정:**

- 귀무가설 ($H_0$): X와 Y 사이에 관계가 없다
- 대립가설 ($H_1$ ): X와 Y 사이에 관계가 있다

p-value가 작을수록 귀무가설이 틀렸을 가능성이 높습니다. 즉, X와 Y 사이에 실제로 관계가 있을 가능성이 높습니다.

**해석 기준 (관습적):**

- $p < 0.01$: 매우 강한 증거 (* 표시) → 99% 신뢰수준
- $p < 0.05$: 강한 증거 ( 표시) → 95% 신뢰수준
- $p < 0.10$: 약한 증거 (* 표시) → 90% 신뢰수준
- $p \geq 0.10$: 증거 불충분 → 관계가 있다고 말할 수 없음

**주의사항:**

1. p-value < 0.05는 관습적 기준일 뿐, 절대적 규칙이 아닙니다
2. p-value가 작다고 효과가 크다는 의미는 아닙니다 (샘플 크기에 영향받음)

3. p-value는 효과가 있다는 확률이 아니라, 효과가 없는데도 이런 결과가 나올 확률입니다

**우리 과제에서의 의미:** 만약 p-value < 0.05라면, 수출 서프라이즈와 09:00~10:20 수익률 사이에 관계가 없다는 가설을 기각합니다. 즉, 실제로 관계가 있다는 의미이고, 이는 벤더의 주장에 의문을 제기합니다.

---

# Step 1: 분석 설계 이해하기

## 시간 구간 정의

**한국 주식시장 거래 시간:**

- 09:00: 장 개시
- 09:00~15:20: 정규 거래 시간
- 15:20~15:30: 동시호가
- 15:30: 장 마감

**데이터 발표 시점:**

- 매월 1일 10:20 (벤더 주장)

**분석 구간 설정:**

- **Before 구간**: 09:00 ~ 10:20 (발표 직전, 80분)
- **발표 시점**: 10:20
- **After 구간**: 10:21 ~ 15:30 (발표 직후, 약 5시간)

## 검증 로직

**만약 벤더의 주장이 사실이라면 (10:20에 처음 공개):**

- Before 구간: 투자자들은 아직 수출 데이터를 모름 → 수익률과 서프라이즈 사이에 관계 없음
- After 구간: 투자자들이 데이터를 받음 → 수익률과 서프라이즈 사이에 관계 있음

**만약 벤더의 주장이 거짓이라면 (이미 유출):**

- Before 구간: 일부 투자자들이 이미 알고 거래 → 수익률과 서프라이즈 사이에 관계 있음
- 이 경우 벤더의 데이터는 독점적이지 않으며, 투자 가치가 떨어짐

### 통계적 검정 방법

선형 회귀분석을 사용합니다:

- 종속변수 (Y): Before 구간 수익률
- 독립변수 (X): 수출 서프라이즈 (rolling Z-score)

회귀분석 결과에서 $R^2$와 p-value를 확인합니다. scipy.stats의 linregress 함수나 statsmodels 라이브러리를 사용할 수 있습니다.

---

# Step 2: 분단위 데이터 준비하기

### 데이터 로드

price_minutely/ 폴더에는 open.csv, close.csv, volume.csv가 있습니다. 이 파일들은 모든 종목의 분단위 가격 데이터를 담고 있습니다. 행은 시간(datetime index), 열은 종목(symbol)입니다.

수익률 계산에는 close.csv를 사용하는 것이 일반적입니다. 종가는 가장 안정적이고 신뢰할 수 있는 가격이기 때문입니다.

### 매월 1일 필터링

분석 대상은 매월 1일(영업일 기준)입니다. 수출 데이터는 전월 데이터를 익월 1일에 발표하기 때문입니다.

주의할 점: 1일이 주말이나 공휴일이면 다음 영업일로 밀립니다. 벤더는 주말에도 데이터를 업데이트한다고 주장하지만, 주식시장은 열리지 않으므로 다음 영업일에 반응이 나타납니다.

Part 1에서 만든 export_with_surprise.csv의 날짜 컬럼을 보면 모두 월말입니다 (예: 2024-01-31). 이 데이터는 다음 달 1일에 발표됩니다. 즉, 2024-01-31 데이터는 2024-02-01에 발표되고, 그날의 주가 반응을 봐야 합니다.

날짜 매칭 로직:

1. export_surprise의 date에서 월(month)을 추출
2. 해당 월의 다음 달 첫 영업일을 찾음
3. 그날의 분단위 데이터를 추출

### 09:00~10:20 구간 수익률 계산

특정 날짜의 분단위 데이터에서 시간 필터링을 해야 합니다. 09:00의 가격과 10:20의 가격을 찾아 수익률을 계산합니다.

수익률 공식: (10:20 가격 - 09:00 가격) / 09:00 가격

이를 모든 종목, 모든 발표일에 대해 반복합니다. 결과는 (날짜 x 종목) 형태의 데이터프레임이 됩니다.

**주의사항:**

- 일부 종목은 특정 날짜에 거래가 없을 수 있습니다 (거래 정지, 상장 폐지 등)
- 09:00나 10:20 정확한 시간에 데이터가 없을 수 있습니다 (거래 체결이 없음)
- 이런 경우는 NaN으로 처리하고, 이후 분석에서 제외합니다

---

# Step 3: 데이터 결합하기

## Long Format 변환

수익률 데이터는 Wide Format (날짜 x 종목)일 것입니다. 이를 Long Format으로 변환해야 export_surprise 데이터와 결합할 수 있습니다.

Long Format:

- date: 날짜
- symbol: 종목
- return_before: 09:00~10:20 수익률

pandas의 melt 함수를 사용하면 편리합니다.

## Surprise 데이터와 Merge

export_with_surprise.csv에는 월말 날짜가 있지만, 실제 발표일과 매칭해야 합니다.

매칭 로직:

1. export_surprise의 date를 발표 월로 변환 (예: 2024-01-31 → 2024-02)
2. return_before의 date도 월로 변환
3. 동일한 월과 종목으로 merge

Merge 후에는 다음 컬럼들이 있어야 합니다:

- date: 발표일 (실제 거래일)

- symbol: 종목
- return_before: Before 구간 수익률
- rolling_zscore_yoy: YoY Surprise
- rolling_zscore_qoq: QoQ Surprise
- rolling_zscore_mom: MoM Surprise

### 데이터 정제

결측치와 이상치를 제거합니다:

1. NaN 값 제거 (수익률이나 서프라이즈가 없는 경우)
2. 무한대(inf) 값 제거 (0으로 나눈 경우)
3. 극단적 이상치 제거 (선택사항) - 예를 들어 하루 수익률이 ±50% 이상인 경우

정제 후 데이터 포인트 개수를 확인하고, 너무 많이 손실되지 않았는지 체크합니다.

---

## Step 4: $R^2$와 p-value 계산하기 (7점)

### 선형 회귀분석 수행

scipy.stats.linregress 함수를 사용합니다. 이 함수는 다음을 반환합니다:

- slope: 기울기 (Y가 X에 대해 얼마나 민감한가)
- intercept: 절편
- r_value: 상관계수
- p_value: 유의확률
- std_err: 표준 오차

$R^2$는 r_value의 제곱으로 계산합니다: r_squared = r_value  2

### YoY Surprise 분석

독립변수 X = rolling_zscore_yoy 종속변수 Y = return_before

회귀분석을 수행하고 결과를 기록합니다:

- 샘플 크기 (n)
- 기울기 (slope)
- $R^2$
- p-value

- 통계적 유의성 판단 (*, , *, 또는 없음)

### QoQ와 MoM Surprise 분석

동일한 방식으로 QoQ와 MoM에 대해서도 분석을 수행합니다. 세 가지 서프라이즈 지표의 결과를 비교합니다.

### 결과 해석

각 지표에 대해 다음 질문에 답하세요:

**$R^2$ 관점:**

- $R^2$이 얼마나 큰가? 0.01 미만? 0.01~0.05? 0.05 이상?
- 수출 서프라이즈가 Before 구간 수익률의 변동을 얼마나 설명하는가?
- 이 정도의 $R^2$이 금융에서 의미있는 수준인가?

**p-value 관점:**

- p-value가 0.05 미만인가?
- 통계적으로 유의미한 관계가 있다고 말할 수 있는가?
- 만약 유의미하다면, 이는 무엇을 의미하는가?

**경제적 의미:**

- 기울기(slope)의 부호는? 양수면 positive surprise가 positive return과 연결
- 기울기의 크기는? Z-score 1 증가 시 수익률이 몇 % 변화하는가?

---

# Step 5: 시각화하기

## Scatter Plot

X축에 rolling_zscore (예: YoY), Y축에 return_before를 놓고 산점도를 그립니다. 각 점은 (종목, 날짜) 조합입니다.

추가 요소:

- 회귀선: 선형 관계를 시각적으로 표현
- 원점 십자선: X=0, Y=0에 점선으로 표시
- 제목에 $R^2$와 p-value 표시

투명도(alpha)를 낮춰서 점들이 겹쳐도 밀도를 파악할 수 있게 합니다.

### 해석 가이드

**만약 scatter plot에서:**

- 점들이 완전히 무작위로 흩어져 있다면: 관계 없음 (벤더 주장 지지)
- 약한 양의 기울기가 보인다면: 약한 관계 (회색 지대)
- 명확한 양의 기울기가 보인다면: 강한 관계 (벤더 주장 의심)

점들의 밀도가 높은 영역을 주의깊게 봅니다. 극단값(outlier) 몇 개가 회귀선을 왜곡할 수 있으므로, 대부분의 점들의 패턴을 봅니다.

# Step 6: 결론 도출하기

## 벤더 주장 검증

이제 핵심 질문에 답할 차례입니다: **벤더는 정말 10:20에 처음으로 데이터를 제공하는가?**

**만약 Before 구간에서:**

- $R^2 < 0.01$이고 p-value > 0.10이라면:
    - 수출 서프라이즈와 Before 수익률 사이에 의미있는 관계가 없음
    - 09:00~10:20 사이에 투자자들은 수출 데이터를 모르고 있었음
    - 벤더의 주장을 지지하는 증거
- $R^2 > 0.05$이고 p-value < 0.05라면:
    - 수출 서프라이즈와 Before 수익률 사이에 유의미한 관계가 있음
    - 09:00~10:20 사이에 이미 정보가 주가에 반영되고 있었음
    - 벤더의 주장에 의문을 제기하는 증거
- 중간 영역 ($0.01 < R^2 < 0.05$, p-value ≈ 0.05~0.10):
    - 약한 관계가 있을 수 있음
    - 명확한 결론을 내리기 어려움
    - 추가 분석 필요

## 보고서 작성

결론 섹션에는 다음 내용을 포함해야 합니다:

1. **요약 통계:** 각 서프라이즈 지표(YoY, QoQ, MoM)의 $R^2$와 p-value를 표로 정리

2. **해석:** 통계적 결과가 실무적으로 무엇을 의미하는지 설명

3. **벤더 평가:** 벤더의 주장이 데이터로 뒷받침되는지, 아니면 반박되는지 명확히 진술

4. **한계점:**

    - 분석의 가정과 한계를 인정
    - 예: Before 구간에 다른 뉴스가 나왔을 가능성
    - 예: 일부 투자자는 비공식 채널로 정보를 미리 얻었을 가능성
5. **추가 분석 제안:**

    - After 구간 분석 (10:21~15:30)과 비교하면 더 명확한 결론 가능
    - 이는 Part 3에서 다룰 예정

## 중요한 인사이트

실제 금융 실무에서 데이터 벤더 평가는 매우 중요합니다. 고가의 데이터 피드를 구독하기 전에 반드시 이런 검증을 수행합니다.

만약 벤더가 주장하는 독점성이 거짓이라면:

- 데이터 가치가 크게 하락
- 구독료 협상에서 불리
- 다른 벤더 탐색 고려

반대로 벤더의 주장이 사실이라면:

- 데이터의 투자 가치가 높음
- 전략 개발에 활용 가능
- 경쟁 우위 확보 가능