

빅데이터 분석을 통한 AI 검색 트렌드와 관련 주가의 상관관계 분석

3-C 202344086 이민서

목차

- 데이터 수집
- 종목 선정 이유
- 데이터 전처리 과정
- 데이터 분석 및 시각화
- 시각화 자료 분석
- 최종 결론

데이터 수집

- 파이썬의 오픈소스 라이브러리를 활용하여 2024년 1년 치 데이터를 수집하였습니다.
- 분석 대상: AI 관련주 4종목 NVIDIA(NVDA), Microsoft(MSFT), Alphabet(GOOGL), Palantir(PLTR)
- 검색 키워드 : "AI"
- 수집 기간 : 2024.01.01~2024.12.31

데이터 종류	출처	활용 라이브러리	수집 항목
주가 데이터	Yahoo Finance	yfinance	일별 종가
검색량 데이터	Google Trends	pytrends	주간 검색 트렌드 지수

종목 선정 이유

- 엔비디아 : 현 시점 AI하면 빼놓을 수 없는 그래픽카드 독점사, 간접, 직접적으로 AI 연구 및 투자 진행 중
- 마이크로소프트 : OPEN AI(Chat gpt)의 대주주이자 자회사의 Copilot 운영 중
- 구글 : 구글 검색 엔진을 기반으로 막대한 데이터를 보유한 동시에 Gemini 운영 중(25년 12월 기준 Gemini의 시장 점유율 1위)
- 팔란티어 : 핵심 사업 모델이 데이터 기반 AI 의사결정 플랫폼이기 때문이고 모든 것이 AI 대체되는 현 시점 최대수혜자

데이터 전처리

- 수집된 원본 데이터는 시간 주기와 수치 단위가 서로 달라
직접적인 비교가 불가능했습니다.
이를 해결하기 위해 다음과 같은 전처리 과정을 수행하였습니다.
- 시계열 주기 통일
- 결측치 처리
- 정규화
- 노이즈 제거

시계열 주기 통일

- 주가 데이터는 '일간(Daily)' 데이터인 반면, 구글 트렌드 데이터는 '주간(Weekly)' 데이터로 제공되어 병합(Merge)이 불가능하였습니다.
- Pandas의 `resample('W').mean()` 함수를 사용하여 주가 데이터를 주간 평균으로 변환하여 시간 축을 통일하였습니다.

결측치 처리

- 휴장일이나 데이터 수집 과정에서 발생한 결측치(NaN) 존재하는 문제가 있습니다.
- `interpolate(method='linear')` 함수를 사용하여 선형 보간법으로 결측치를 부드럽게 채웠습니다.
- 선형 보간법을 사용하면 데이터가 부정확한 건 아닐지 고민도 해보았지만 주가와 검색량 데이터가 시간의 흐름에 따라 연속성을 가지기 때문에 데이터의 추세 왜곡을 최소화하기 때문에 사용하였습니다.

정규화

- 검색량은 0~100 사이의 지수이지만, 주가는 수십~수백 달러 단위로 단위의 차이가 큽니다. 스케일이 다른 두 변수를 하나의 그래프에서 시각화하거나 모델에 학습시킬 경우, 단위가 큰 값에 의해서 분석 결과가 왜곡될 수 있었습니다.
- MinMaxScaler를 적용하여 모든 데이터를 0과 1 사이의 값으로 변환하여 패턴 유사성을 직관적으로 비교할 수 전처리하였습니다.

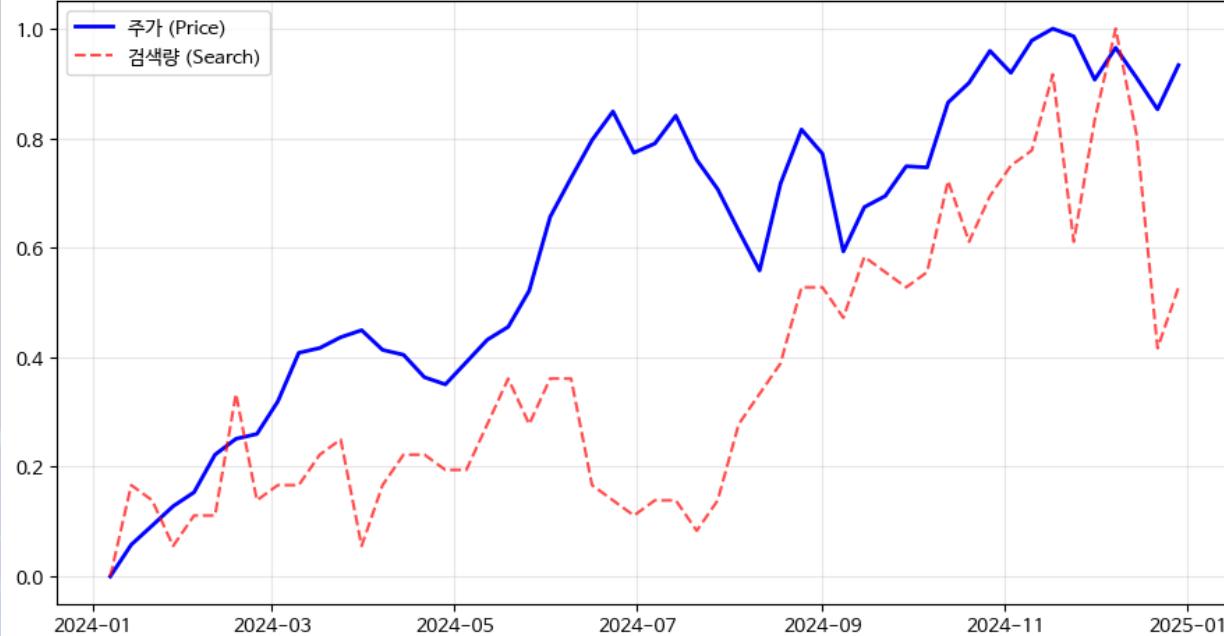
노이즈 제거 및 추세선 생성(MA선)

- 검색량 데이터는 특정 이슈에 따라 단기적으로 급등락(노이즈)하는 경우가 있어, 장기적인 주가 흐름과의 상관성을 파악하는데 방해가 될 수 있습니다.
- 노이즈를 제거하고 전반적인 추세를 파악하기 위해 데이터 평활화(rolling(window=4).mean())을 사용하여 4주 이동평균선을 생성하여 해결하였습니다.

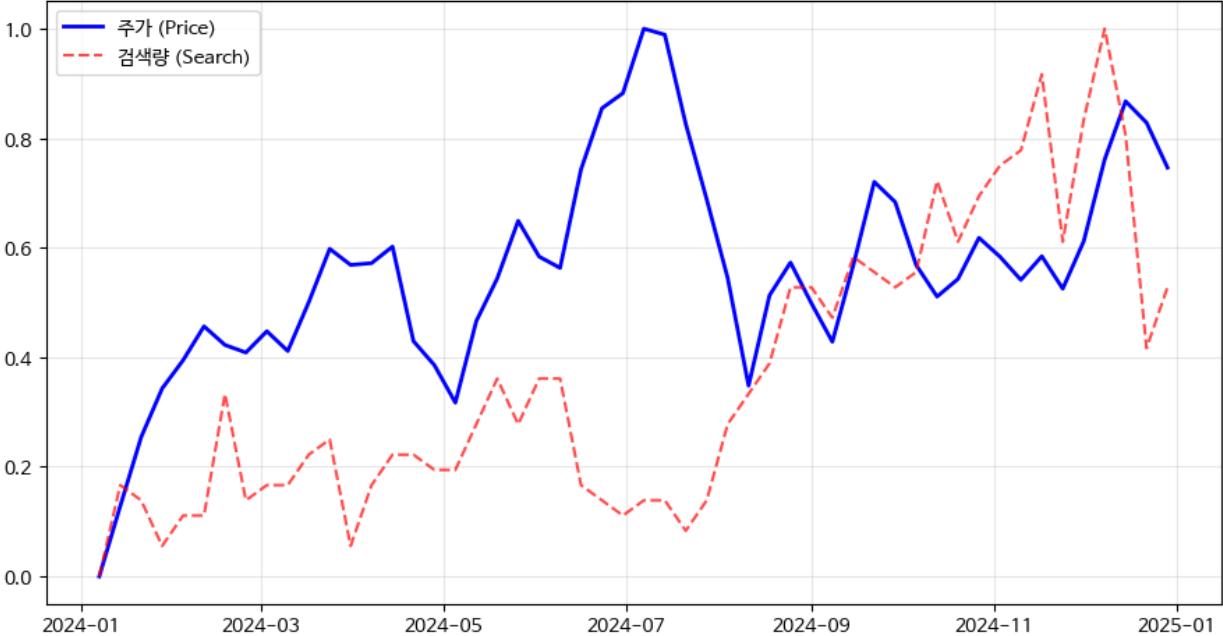
데이터 분석 및 시각화(시계열 추세 비교)

- 서로 다른 수치 단위를 가진 '주가(\$)'와 '검색량(Index)'의 흐름을 직관적으로 비교하기 위해 이중 축(Dual Axis) 그래프를 활용하여 시각화하였습니다.
- 주가 (Stock Price): 좌측 Y축을 기준으로 파란색 실선으로 표현하였으며, 시각적 명확성을 위해 영역을 채색하여 전반적인 가치 상승/하락의 흐름을 강조하였습니다.
- 검색량 (Search Volume): 우측 Y축을 기준으로 빨간색 점선으로 표현하였고, 검색량 데이터 특유의 급격한 변동성(Noise)을 완화하고 기조적인 추세(Trend)를 파악하기 위해 4주 이동평균을 적용하여 곡선을 평활화하였습니다.

NVDA (상관계수: 0.71)



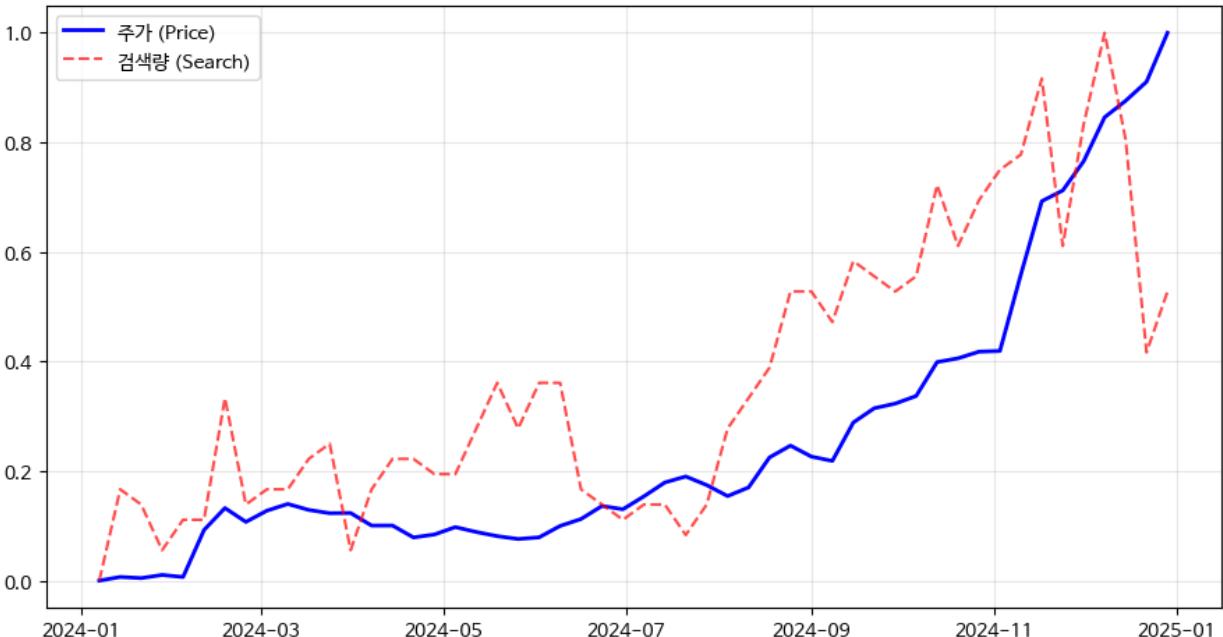
MSFT (상관계수: 0.22)



GOOGL (상관계수: 0.37)



PLTR (상관계수: 0.79)



그래프 주요 분석 포인트

- **동행성** - 검색량이 급증하는 시점에 주가 또한 상승하는지, 두 변수의 변곡점이 일치하는지를 확인할 수 있습니다.
- **선행성** - 검색량의 증가가 주가 상승에 선행하여 나타나는지(선행 지표로서의 가능성), 혹은 주가 급등 후 검색량이 따라오는지(후행성)를 파악할 수 있습니다.
- **괴리** - 검색량은 감소함에도 주가는 상승하거나(기업 펀더멘털에 기인), 반대로 검색량은 폭증하나 주가는 하락하는(악재 발생) 비동조화 구간이 존재하는지 분석할 필요가 있습니다.

종목별 상세 분석 결과

- NVIDIA(NVDA)
- 파란색 실선(주가)과 빨간색 점선(검색량)이 매우 유사한 동행 패턴을 보였습니다. 특히 AI 관련 주요 이벤트나 실적 발표 시즌에 검색량이 치솟으며 주가 상승이 동시에 이루어지는 경향이 뚜렷합니다. 이는 NVIDIA의 주가가 대중의 관심도와 매우 밀접하게 연동되어 있음을 시사한다는 것을 확인할 수 있습니다.

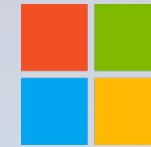


종목별 상세 분석 결과



- Palantir(PLTR)
- 주가의 등락폭이 큰 구간에서 검색량 또한 민감하게 반응하는 모습을 보였습니다.
특정 시점(예: 2024년 하반기 등)에 검색량이 급증하며
주가 상승을 견인하는 양의 상관관계가 육안으로도 명확히 확
인되었습니다.

종목별 상세 분석 결과



Microsoft

Google

- Microsoft (MSFT) & Alphabet (GOOGL)
- 검색량 트렌드와 주가 간의 연관성이 상대적으로 낮게 관측되었습니다.
검색량은 특정 이벤트(예: 신규 AI 모델 공개)에 따라 단기적으로 급등락하는 반면, 주가는 거시 경제 상황이나 기업 실적에 따라 보다 완만하고 독립적인 추세를 그리는 괴리 현상이 일부 구간에서 발견되었습니다.

데이터 분석 및 시각화(상관관계 히트맵)

- 'AI 검색량'과 '4개 기업 주가' 간의 연관성 강도를 파악하기 위해 피어슨 상관계수를 산출하고, 이를 히트맵으로 시각화하였습니다.
- 변수: Search_Volume(검색량) vs. NVDA, MSFT, GOOGL, PLTR(각 주가)
- 해석 기준: 계수가 1에 가까울수록(붉은색) 두 변수가 동일한 방향으로 강하게 움직임을 의미하며, 0에 가까울수록 서로 독립적임을 의미합니다.

데이터 분석 및 시각화(상관관계 히트맵)

AI 트렌드 민감도 (Search Volume과의 상관성)

NVIDIA (NVDA): 상관계수 **0.7112**로 매우 강한 양의 상관관계를 보였습니다. 이는 대중의 AI 관심도가 증가함에 따라 AI 하드웨어 시장의 지배적 사업자인 엔비디아의 주가가 즉각적으로 반응하며 동행하고 있음을 시사합니다.

Palantir (PLTR): 상관계수 **0.7875**로 분석 대상 중 가장 높은 연관성을 기록합니다. 개인 투자자들의 관심이 집중된 종목 특성과 AI 소프트웨어 플랫폼에 대한 기대감이 검색량과 주가에 동시에 반영된 것으로 해석됩니다.

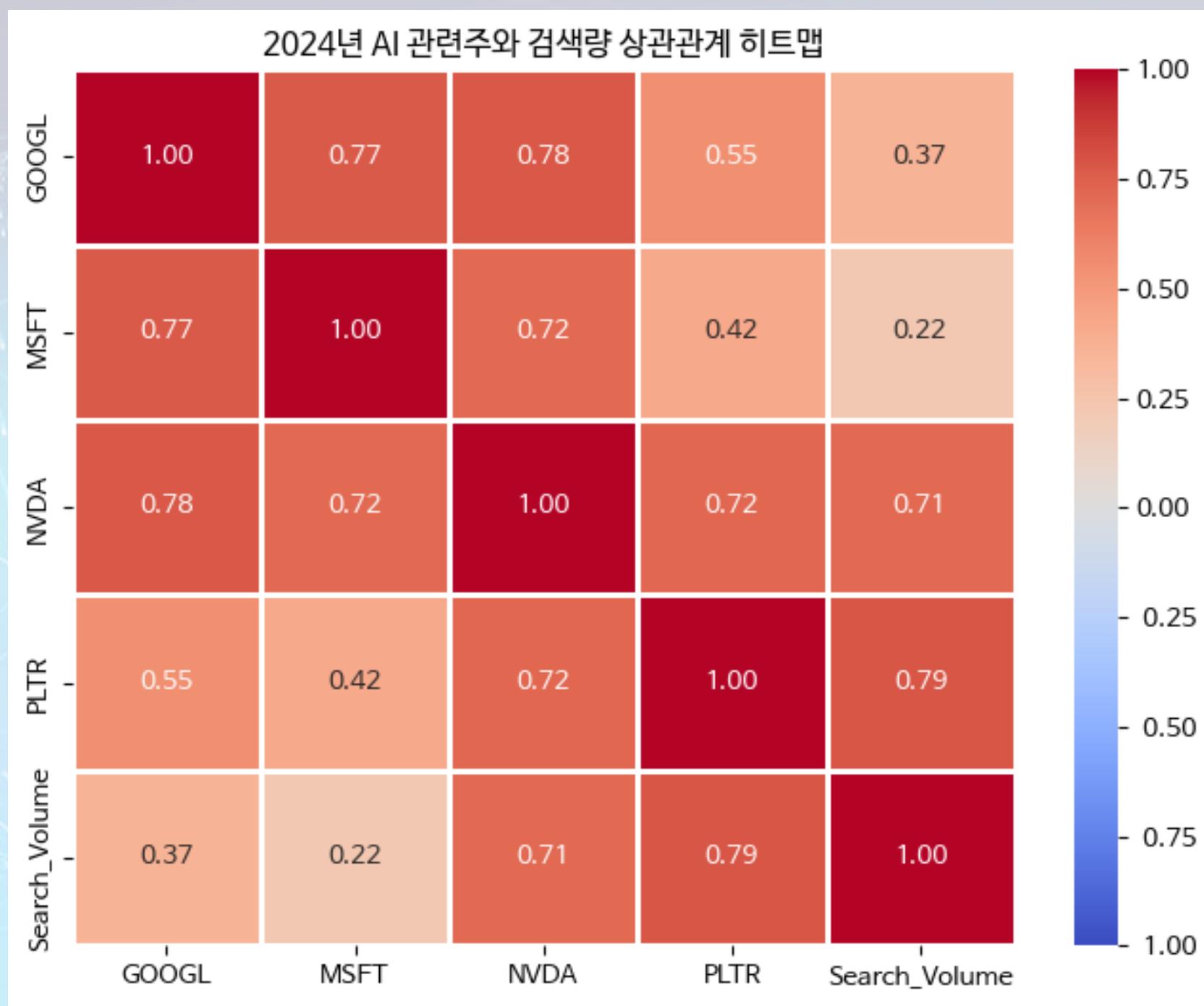
데이터 분석 및 시각화(상관관계 히트맵)

AI 트렌드 민감도 (Search Volume과의 상관성)

Alphabet (GOOGL): 상관계수 **0.3653**으로 어느 정도의 상관관계를 보였습니다. 검색 광고 등 기존 사업의 비중이 커 AI 이슈가 주가에 미치는 영향력이 희석된 것으로 보이나, 전반적인 트렌드와는 방향성을 같이하고 있습니다.

Microsoft (MSFT): 상관계수 **0.2205**로 뚜렷한 연관성을 찾기 어려웠습니다. 클라우드, 오피스 등 다양한 사업 포트폴리오와 거시경제 요인, 실적 등이 복합적으로 작용하여 단순 검색량 주이와는 다른 독자적인 주가 흐름을 형성한 것으로 판단됩니다.

2024년 AI 관련주와 검색량 상관관계 히트맵



최종 결론

이번 프로젝트를 통해 "검색량(대중의 관심)"은 AI 테마를 주도하는 특정 기업(NVIDIA, Palantir)의 주가 변동을 설명하는 매우 유의미한 선행/동행 지표임이 입증되었습니다.

반면, 거대 기술 기업의 경우 검색량 외에 실적이나 시장 상황, 기존의 다른 여러 사업 등 다양한 변수가 주가에 더 큰 영향을 미침을 확인하였습니다.