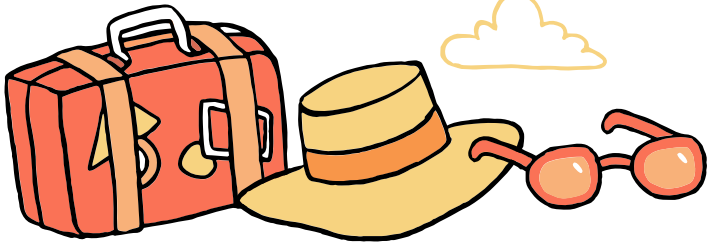
비행기를 저렴하게 예매하고 싶은 P의 데이터 분석

By J와 P



유동규 임수빈 정다연 최가은

Contents

01

TOPIC

주제 선정 배경과
주요 관점 소개

02

TOOLS

프로젝트를 진행에
사용한 도구 소개

03

DATA

수집한 데이터 설명과
데이터 전처리 과정
소개

04

DATABASE

데이터 베이스를
이용한 데이터 관리

04

ANALYSIS

데이터 분석 결과와
가격 요소에 대한 다른
관점 분석

Sub-Contents

DATA

실시간 데이터를 얻기 위한 과정과
데이터 전처리 과정에 대한 설명

데이터 수집 조건

데이터 수집시 제한한 조건들

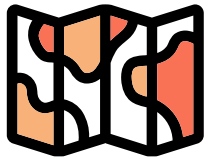
Docker 활용 설정

끊기지 않는 데이터 수집을 위해

데이터 전처리 과정

데이터 손실 이상치와 결측치

ANALYSIS



예상 결과와 실제 데이터로 살펴본
비행기 가격의 변동 추이 분석

예상 결과

분석을 시작하며 예상했던 결과

데이터 분석

실제 데이터 분석 및 시각화

또다른 관점

다른 관점에서 데이터 예측 및 분석

TOPIC

주제 선정 배경

여행 꿀팁

"이걸 몰라서 2배는 비싸게 샀네!"...항공권이 가장 저렴한 요일

"이걸 몰라서 2배는 비싸게 샀네!"...항공권이 가장 저렴한 요일

Editor. 이지은 | 입력 2023.07.26 13:02 | 수정 2023.07.26 13:52 | 댓글 0

인기 콘텐츠

"굳이 일본 여행을 왜 가?"... 일본의 감성을 느낄 수 있는 국내 료칸

"나 지금 해외 온 거 아냐?"... 이국

항공권 예매를 요일에 따라 가격이 달라질 수 있다는 데이터 분석 결과가 있습니다. 항공권 예매 업체인 스카이스캐너의 판매 데이터 분석 결과 "화요일에 예매하는 것이 가장 저렴하고, 오전 5시를 공략하는 것이 좋다"는 결과가 나왔다고 밝혔다. 특히 화요일 오후에는 은 전 세계 모든 도시 간의 국제선 항공권 평균가격이 가장 낮은 요일이라고 알려졌다.

TOOLS

- 사용 도구

seaborn

AWS

matplotlib

Beautiful Soup

Selenium

pandas

docker

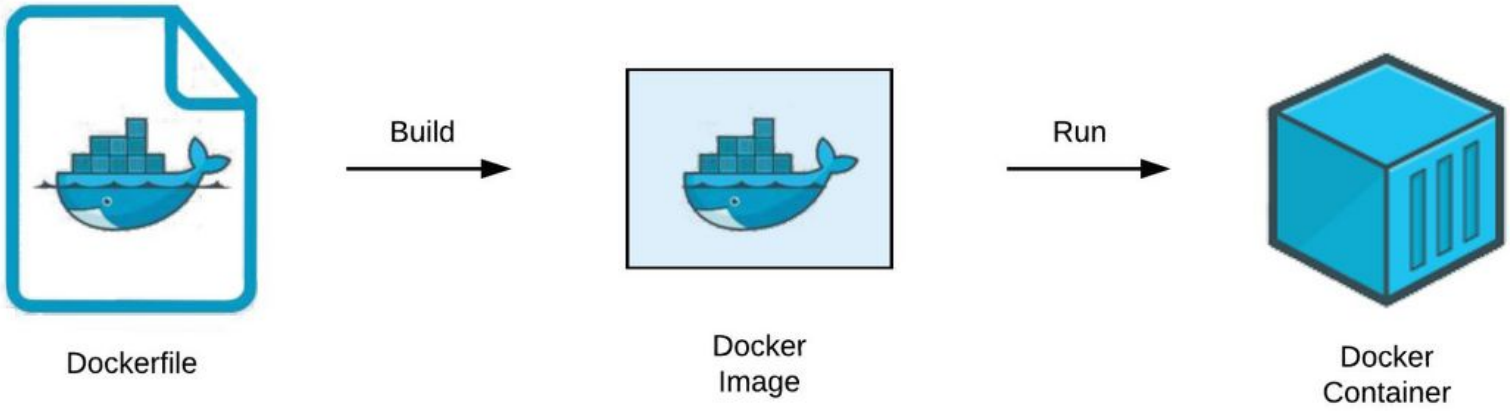
TOOLS

Docker

시계열 데이터를 수집할 때 무엇이 중요할까? -> 환경 일관성!

그리고 안전성!

~~데이터 누수나면 큰일난다..^^~~



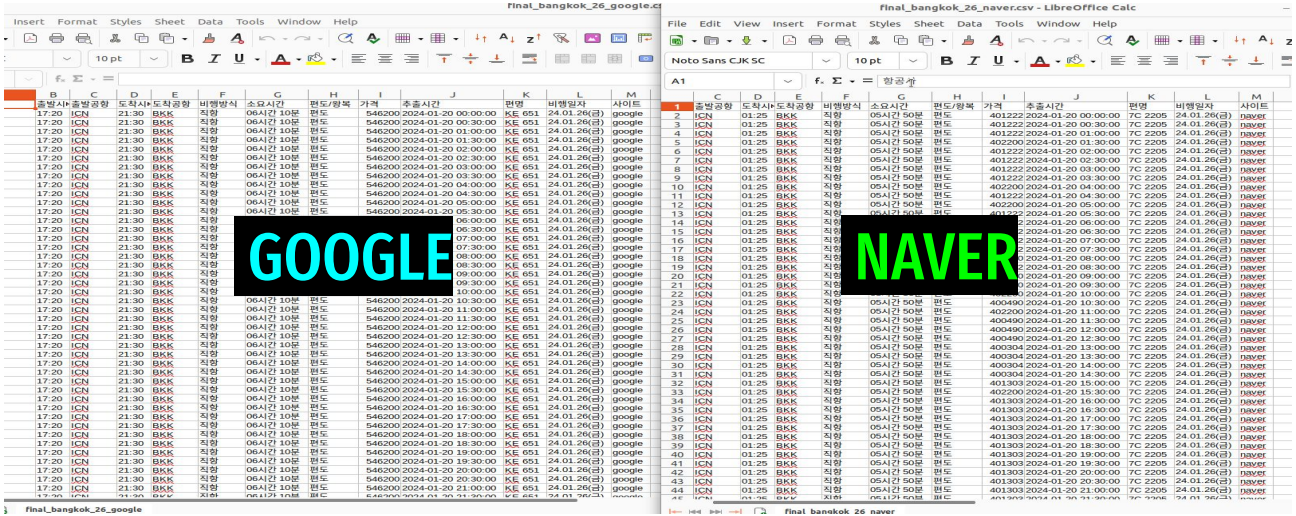
DATA

- 데이터 수집 조건

타겟 날짜	1월 26일(금)	1월 27일(토)	
타겟 국가	방콕	다낭	오사카
수집 사이트	Google flight	Naver 항공권	
그 외 조건	편도	직항	출발 공항 ICN

DATA

데이터 전처리 과정



하루에 편명이 **30**개라고 가정하면,

30*2(30분간격**)*24(24**시간**)*2(**타겟날짜**)*3(**타겟국가**) = 8640**개***2(Google,Naver)**

대략 **17,280**개의 데이터의 전처리가 필요하다

DATA

데이터 전처리 과정

데이터 손실

네트워크 환경에
따라 크롤링을 해
오지 못한 데이터

이상치

실제 데이터이기는
하나 데이터 분석 시
방해가 될 데이터

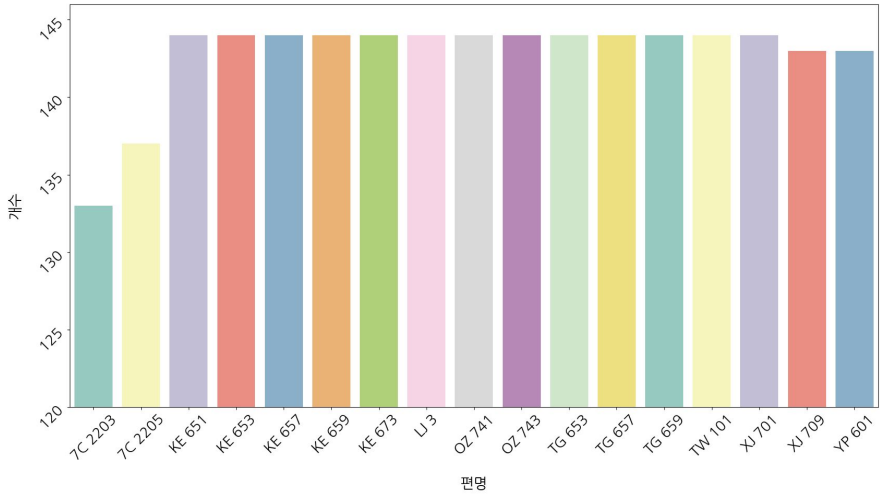
결측치

편명이나 가격
데이터가 없는
데이터

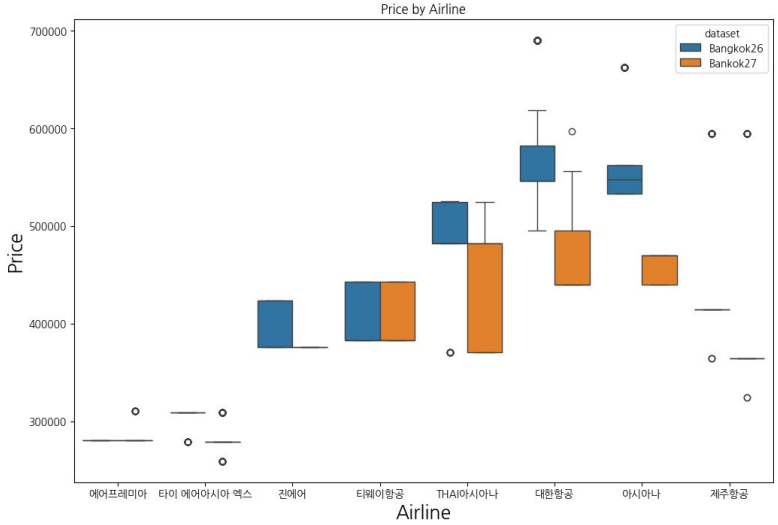
DATA

- 데이터 전처리 과정 - 이상치와 결측치 시각화

편명 중 결측치가 있는지



편명별 가격 중 이상치가 있는지



DATA

데이터 전처리 과정 - Before vs After 그래프

Before

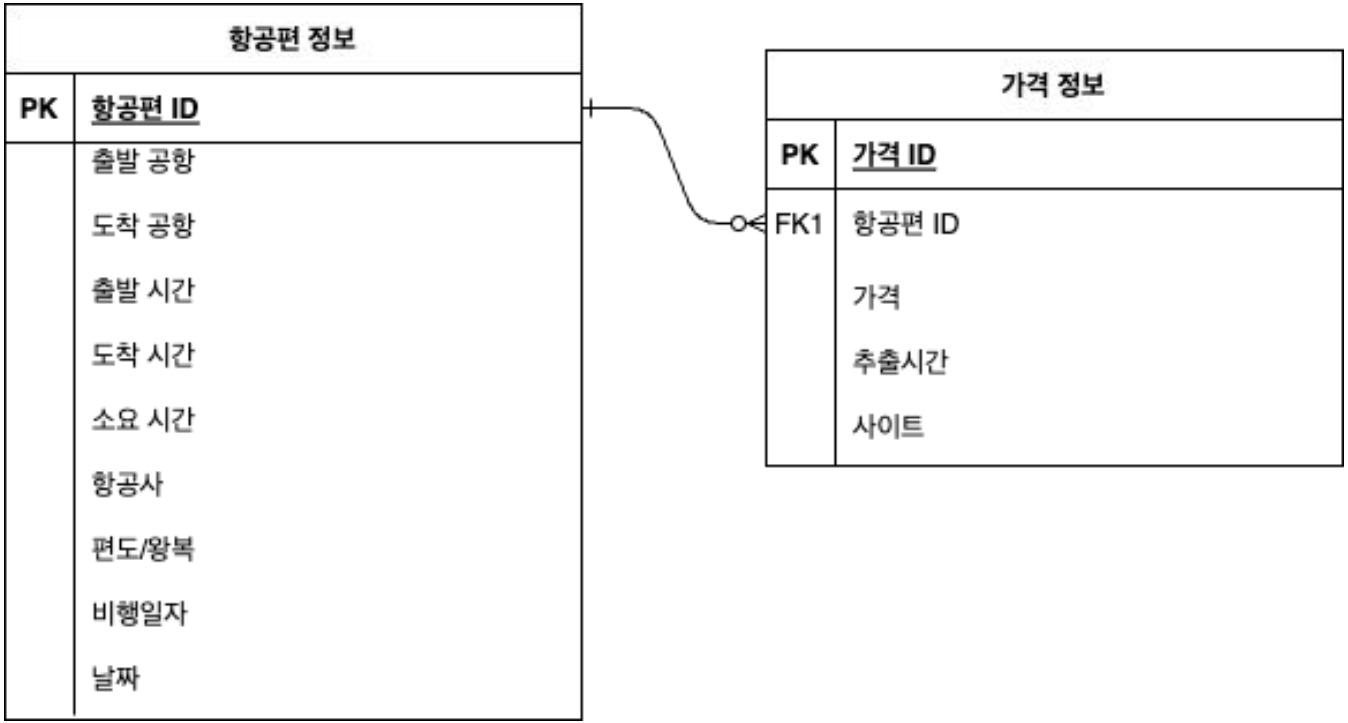
[결측치, 이상치 제거 전] 편명별 시간에 따른 가격 변동 추이

After

[결측치, 이상치 제거 후] 편명별 시간에 따른 가격 변동 추이

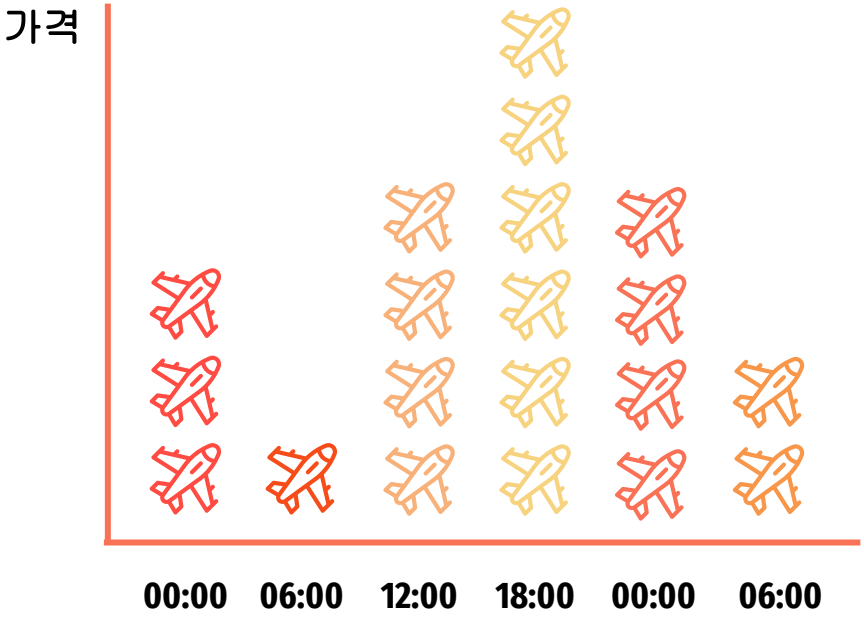
DATABASE

- AWS



ANALYSIS

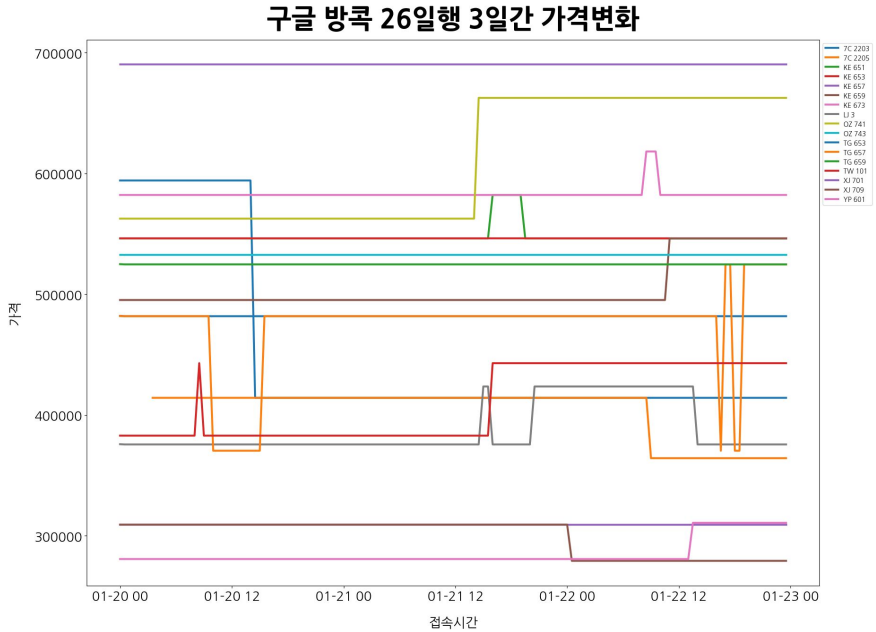
우리가 예상했던 결과



- 시간에 따라 유동적으로 변화하는 가격 그래프
  - 새벽 시간대 혹은 밤 늦은 시간에 낮은 가격
  - 금요일(26)과 토요일(27) 사이의 가격 차이
  - 타겟 국가에 따른 가격 차이
  - 크롤링 날짜(20,21,22)일에 상관없이 시간대별 경향

ANALYSIS

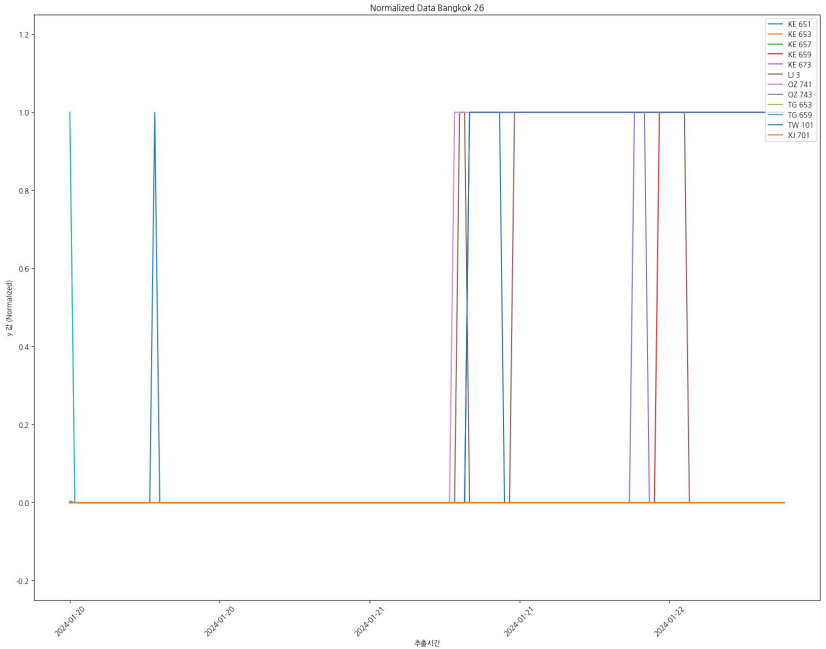
- 이제 분석 해볼까요! 먼저 전처리한 데이터로 그래프 시각화



가격대가 제각각이라 보기 불편하네요... 정규화해볼까요?

ANALYSIS

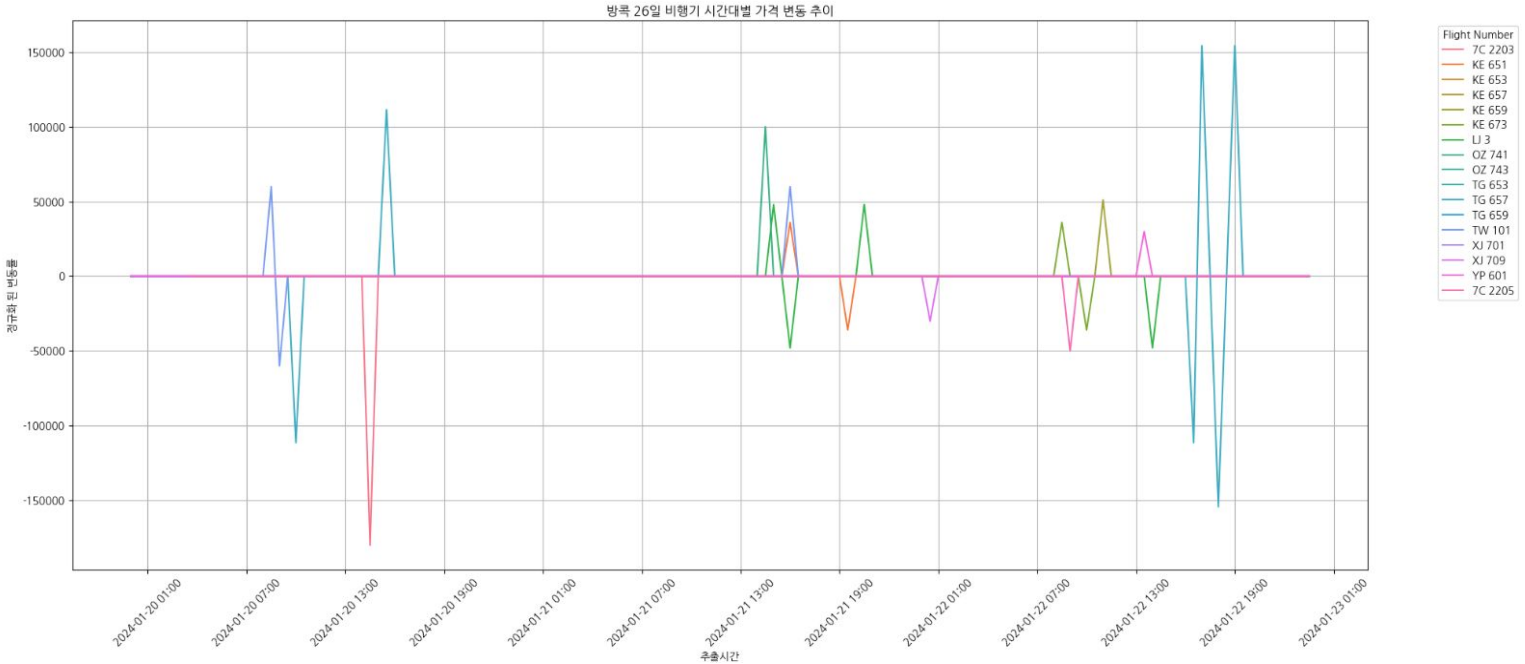
- 모든 편명을 같은 선상 정규화해서 변동을 봅시다



정규화해서 한 줄로 보니 26일 전체의 변동이 언제 있었는지 시간 확인은 되지만 가격 변동 폭의 차이도 알 수 없고 시간에 따라 보기에도 어렵네요...그렇다면!!

ANALYSIS

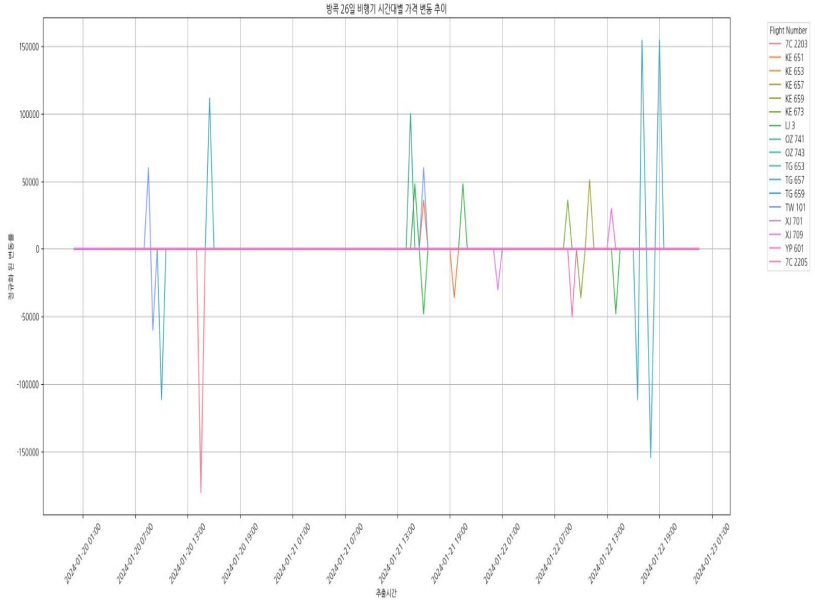
- 가격 변동률로 나타내본 그래프! 이쁘다!

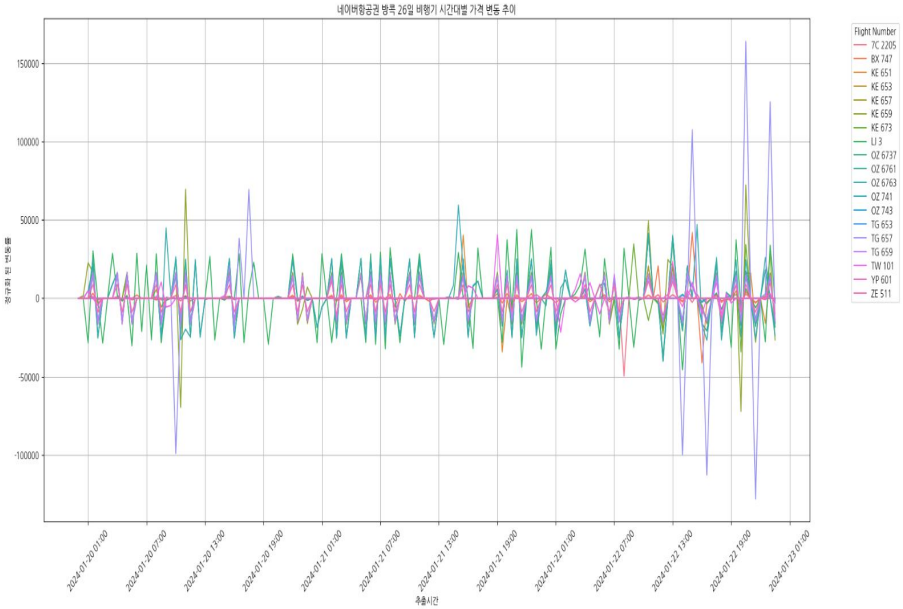


3일간의 가격 변동을 보기도 좋고 변동 있는 편명들만 확인하기도 좋은거 같다

ANALYSIS

변동률 그래프로 본 google vs naver

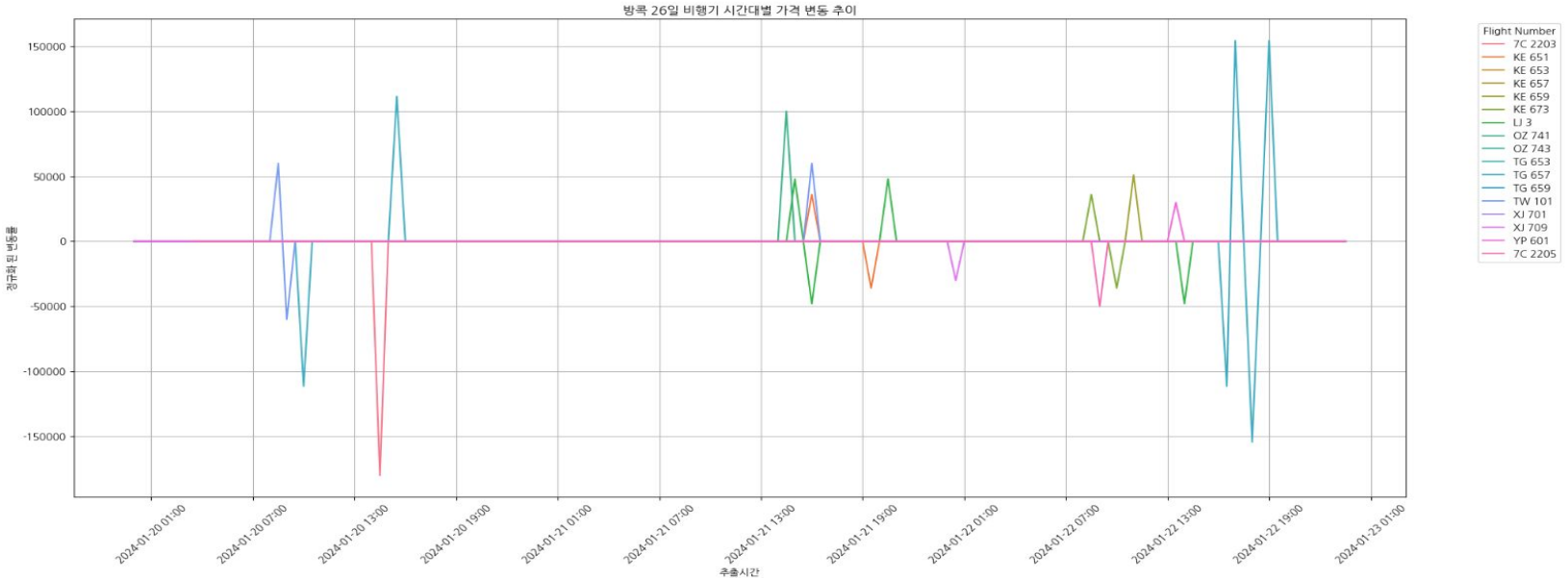




두 사이트 데이터를 비교하니 네이버가 변동률이... 대단합니다..

ANALYSIS

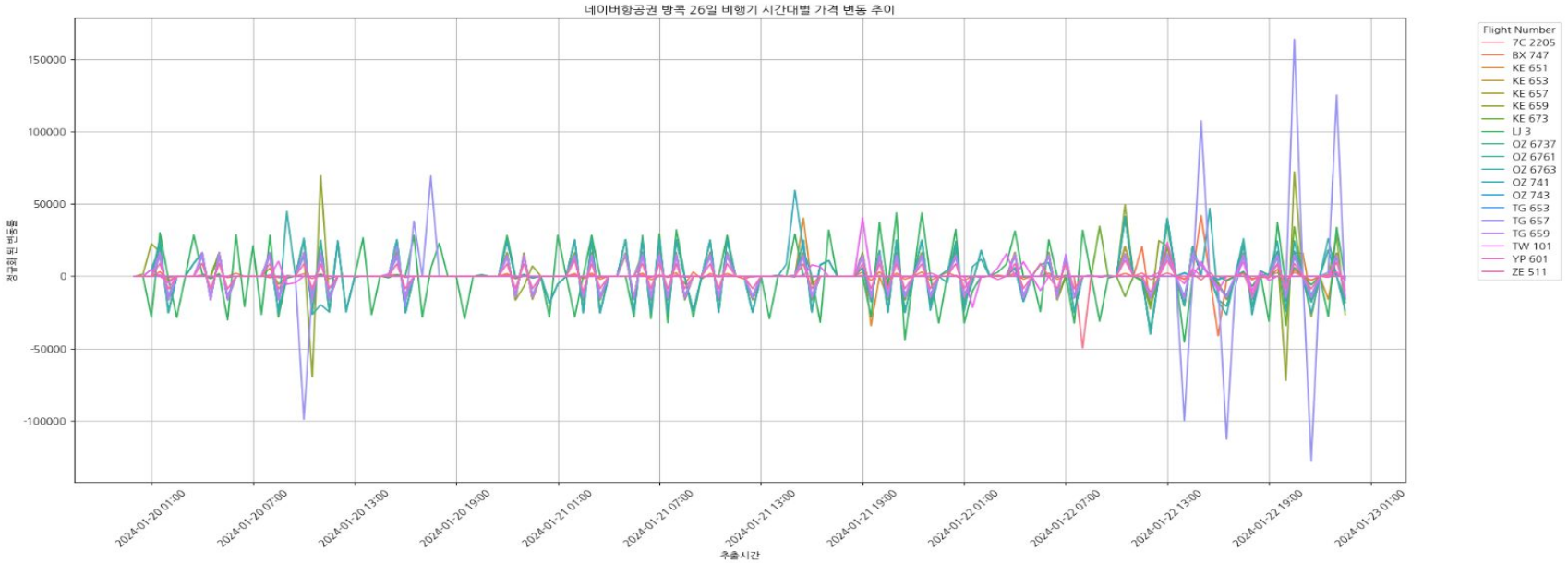
변동률 그래프로 본 google vs naver - google



구글은 변동폭이 있는 몇몇 항공사를 제외하고는 같은 값을 유지하고 있네요

ANALYSIS

변동률 그래프로 본 google vs naver - naver



네이버는 계속해서 가격 변동이 있고, 변동폭이 아주 큰 항공사들도 있습니다

ANALYSIS

- 데이터로 분석해본 결과 - 결론

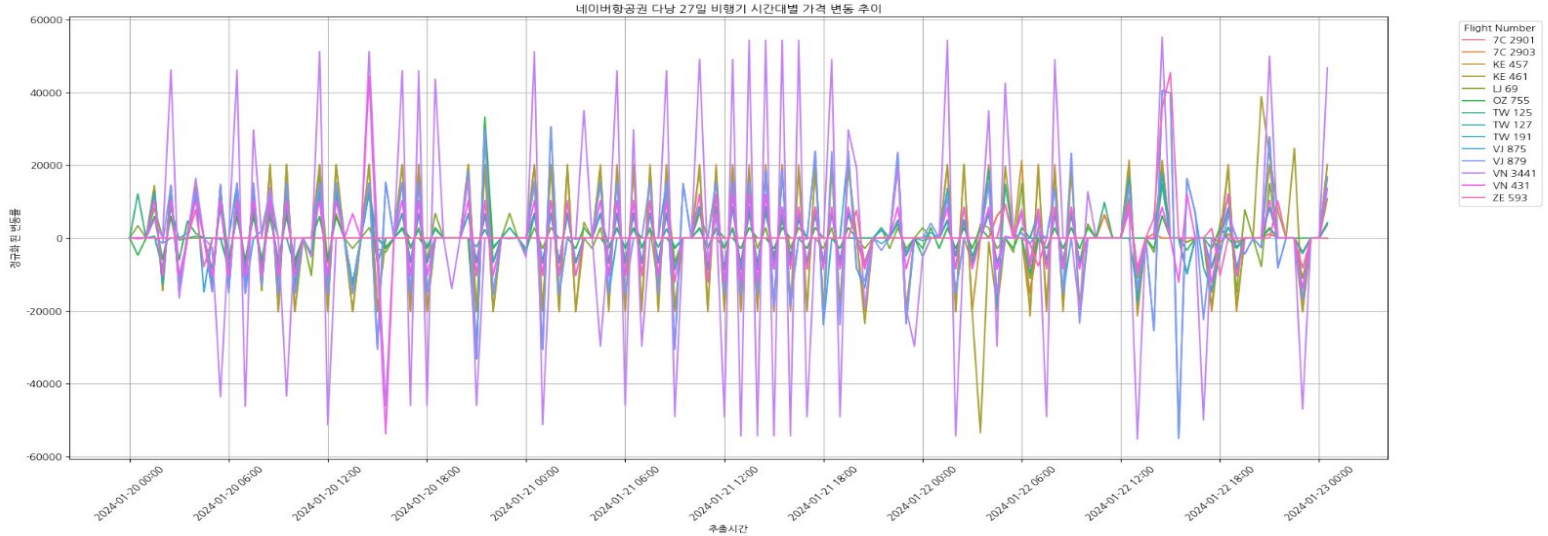
  - 구글은 몇개 항공사를 제외하고는 동일한 가격을 3일 유지
  - 네이버는 30분 마다 가격 변동이 활발하다!

    Etc. 테스트로 2,3분 docker 크롤링 타임 설정사이에도 값이 변동
  - 네이버는.. 최대한 많이 들어가보시는게 이득입니다^^

ANALYSIS

데이터로 분석해본 결과 - 궁금해서 찍어본 손해가격

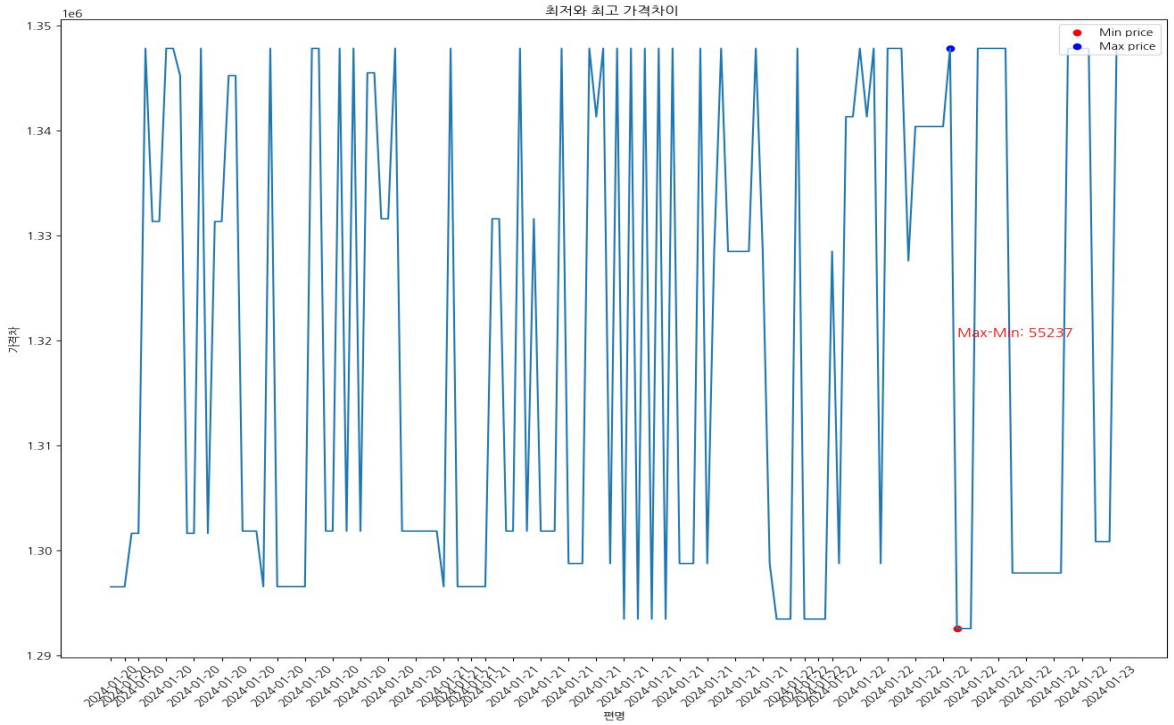
- 네이버에서 가장 변동폭이 컸던 **VN 3441** 항공기!
- 클릭한번에 얼마가 손해인건가 봤더니..



ANALYSIS

데이터로 분석해본 결과 - 궁금해서 찍어본 손해가격

- 거의 30분마다 가격 변동이 있다
- 가장 최고가와 최저가를 비교해보니..!
- 클릭한번에 5만 5천원 왕복으로 구매하면 10만...원!



ANALYSIS

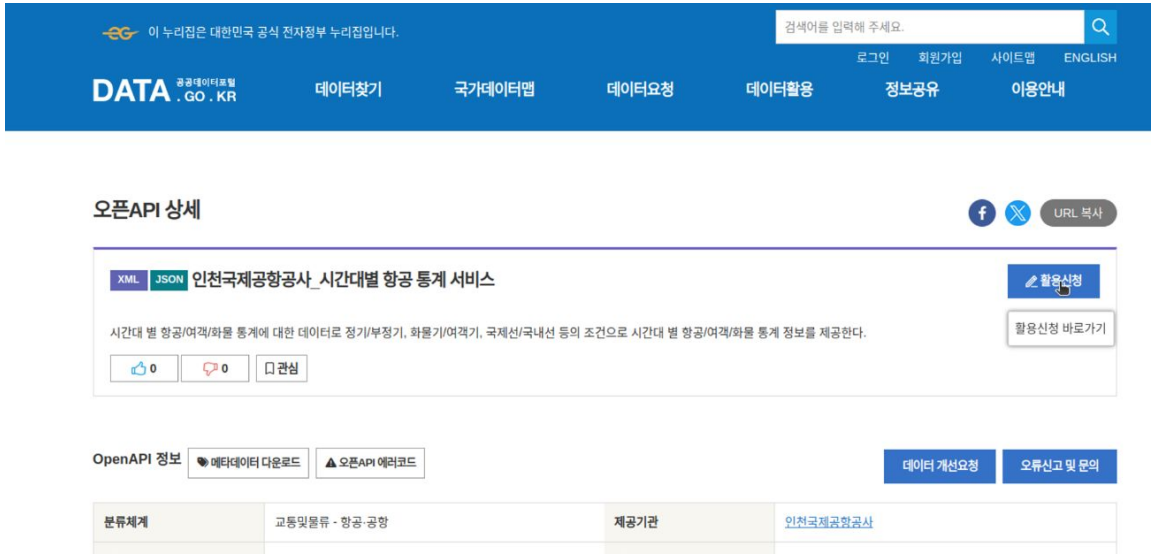
- 데이터가 이렇게 많은데 도움이 될 수 있는 다른 정보도 있을까?

  - 대형 항공사와 중소 항공사 사이의 가격 변동은 어떻게 다를까?
  - 혹시 특정 시간에 비행기 운항이 적으면 가격이 상승할까?

ANALYSIS

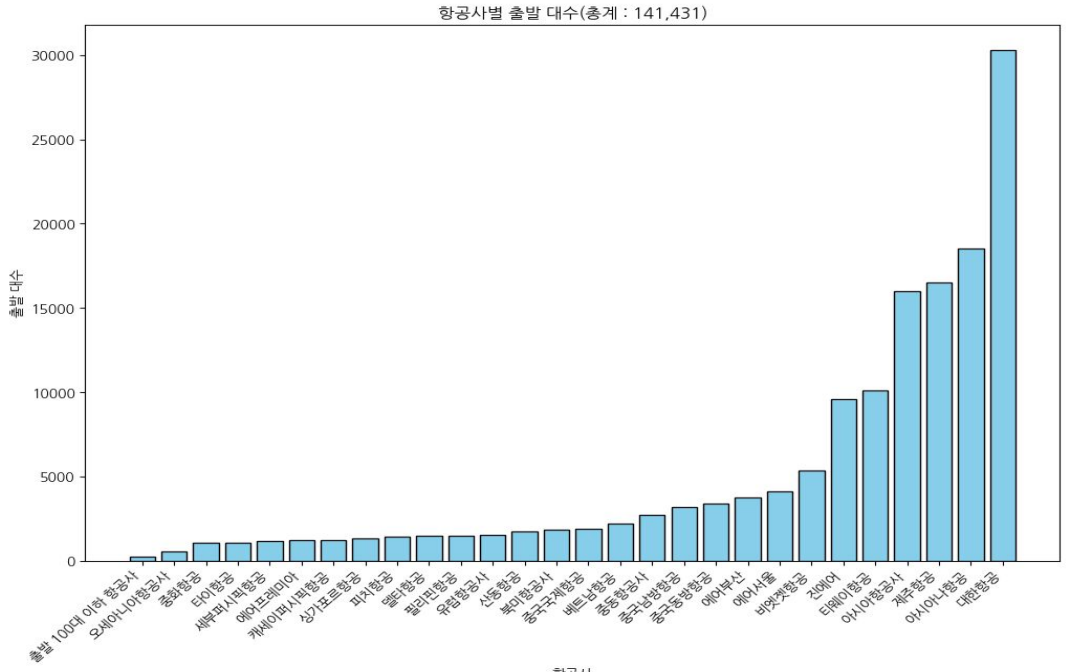
- 대형 항공사와 중소 항공사 사이의 가격 변동은 어떻게 다를까?

  - 일단 대형 항공사들이 누굴까? 멋있게 데이터로 찾기(공공데이터 포털)



ANALYSIS

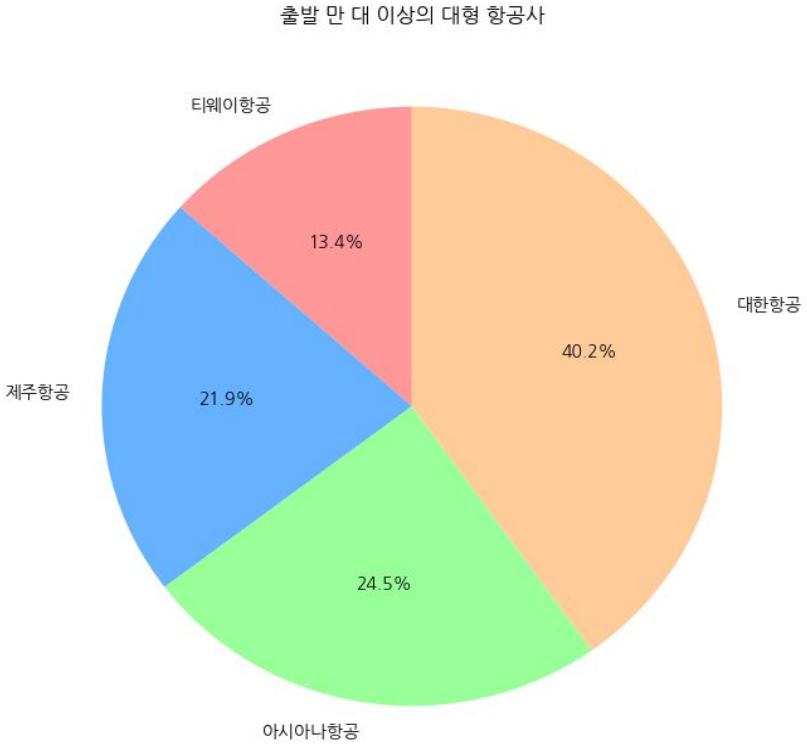
- 대형 항공사와 중소 항공사 사이의 가격 변동은 어떻게 다를까?



- 인천공항 출발 운항이 총 14만 대 정도인데 상위 3개가 거의 9만...대단해

ANALYSIS

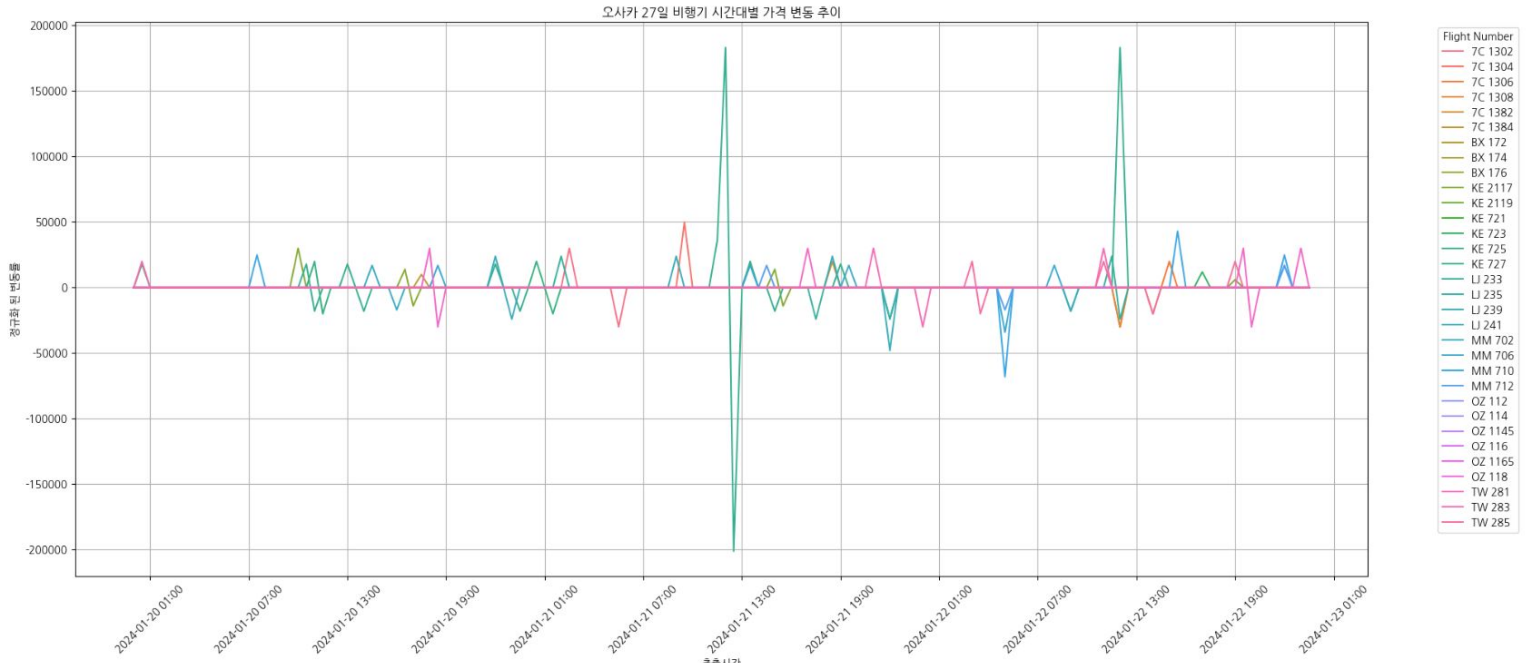
대형 항공사와 중소 항공사 사이의 가격 변동은 어떻게 다를까?



- 만 대 이상 운항하는 항공사들을 뽑아보니!
- 대한항공, 아시아나항공, 제주항공, 티웨이 항공...
- 그렇다면 대형 항공사들의 가격 변동은 어떨까?

ANALYSIS

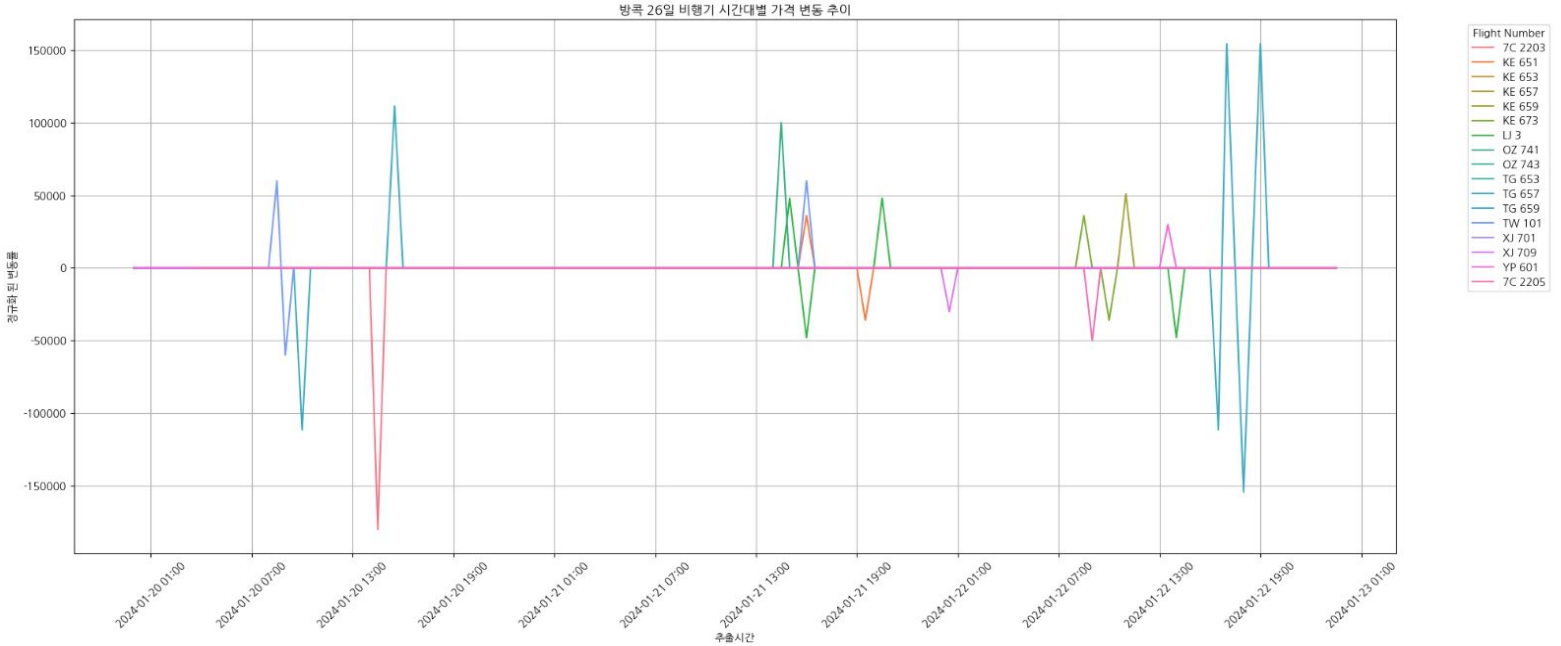
- 대형 항공사와 중소 항공사 사이의 가격 변동은 어떻게 다를까?



- 눈에 띄는 초록색..KE 727, 대한항공! 대형 항공사가 변동폭이 크구나!

ANALYSIS

- 대형 항공사와 중소 항공사 사이의 가격 변동은 어떻게 다를까?



- 다른 국가 데이터를 보면, **OZ 743**! 아시아나 항공의 가격 변동폭이 크다

ANALYSIS

- 대형 항공사와 중소 항공사 사이의 가격 변동은 어떻게 다를까?

- 대형 항공사의 변동은 전체 변동에 큰 영향을 미칠 정도로 큰 변동 폭을

ANALYSIS

- 대형 항공사와 중소 항공사 사이의 가격 변동은 어떻게 다를까?- mini 결론

  - 구글은 대형 항공사의 변동률이 많았던 이후에는 평균 가격도 올라간다..!

    따라서.. (대형 항공사를 고집하는 사람이라면!) 대형 항공사의 가격변동 후에는

    아무리 들어가도.. 큰 도움은 안될 수 있어요..

  - 중소 항공사는 사실 변동폭이 크지 않아서... 최저가 구하려고 노력하는 것보다는

    여행지 맛집을 찾는 것이 더 중요할지도...ㅋㅋ

ANALYSIS

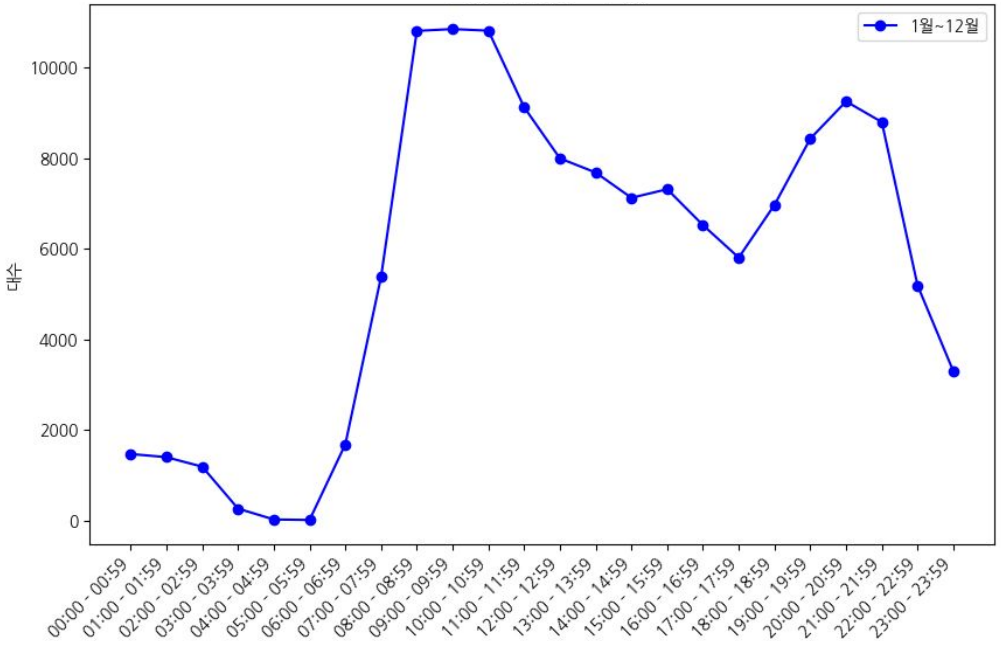
- 또 다른 분석, 특정 시간에 비행기 운항이 적으면 가격이 상승할까?

  - 인천공항 운항 데이터를 가져와 분석해보자!

ANALYSIS

혹시 특정 시간에 비행기 운항이 적으면 가격이 상승할까?

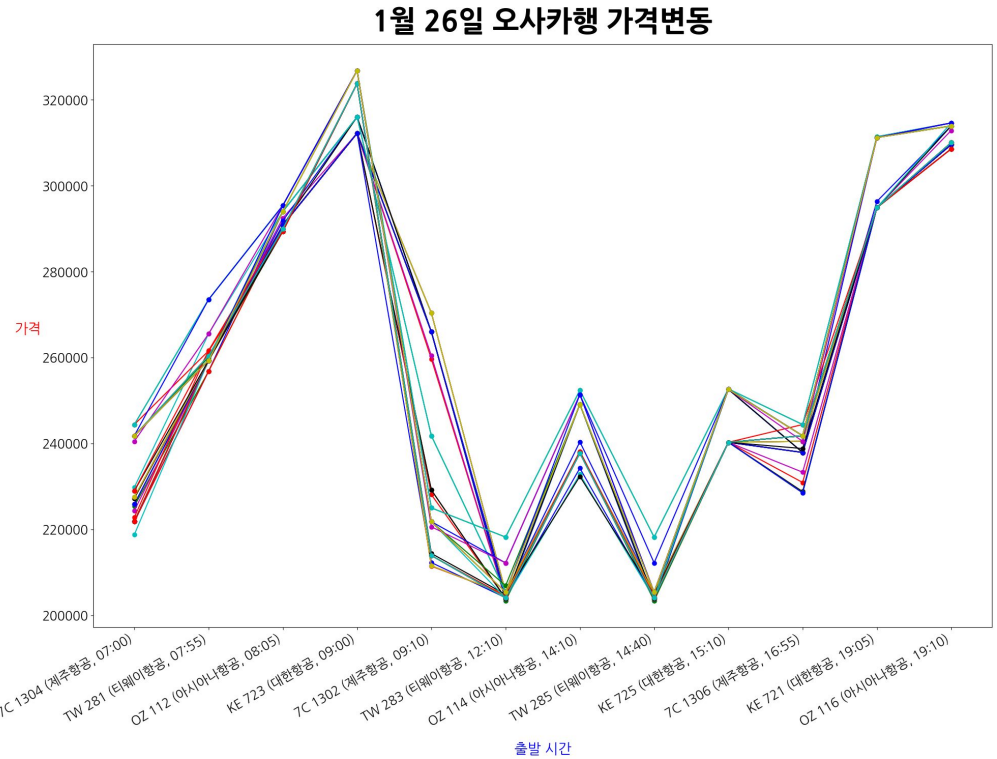
운항 데이터를 하루 시간 순으로 시각화



- 아침 시간대와 저녁시간대 비행기 수가 제일 많다!
- 우리가 갖고 있는 데이터와 비교 해볼 수 없을까?

ANALYSIS

혹시 특정 시간에 비행기 운항이 적으면 가격이 상승할까?

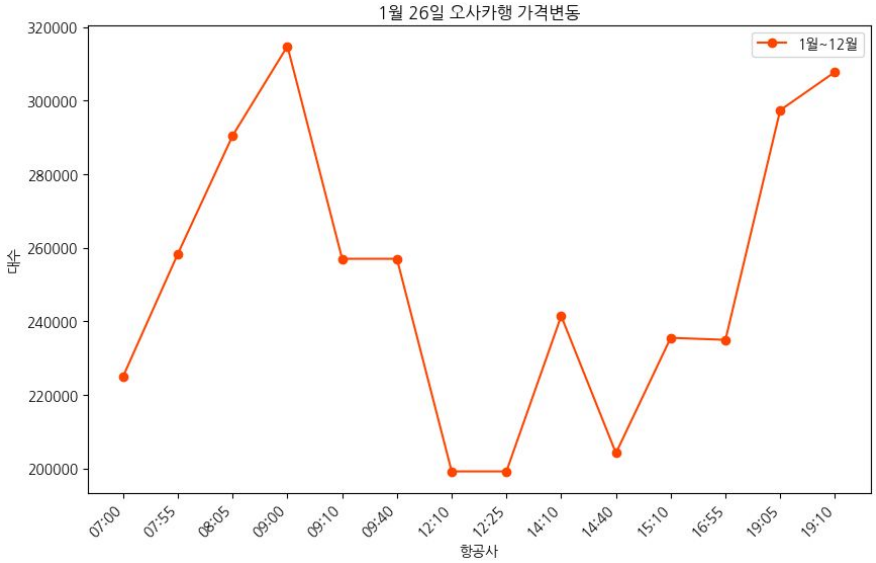


- 시간대별 운항 대수 그래프와 비교하기 위해 출발 시간에 따른 가격 그래프를 그려보았다
- 3일간 가격 변동이 있지만 가격대는 서로 비슷하게 형성되어 있다

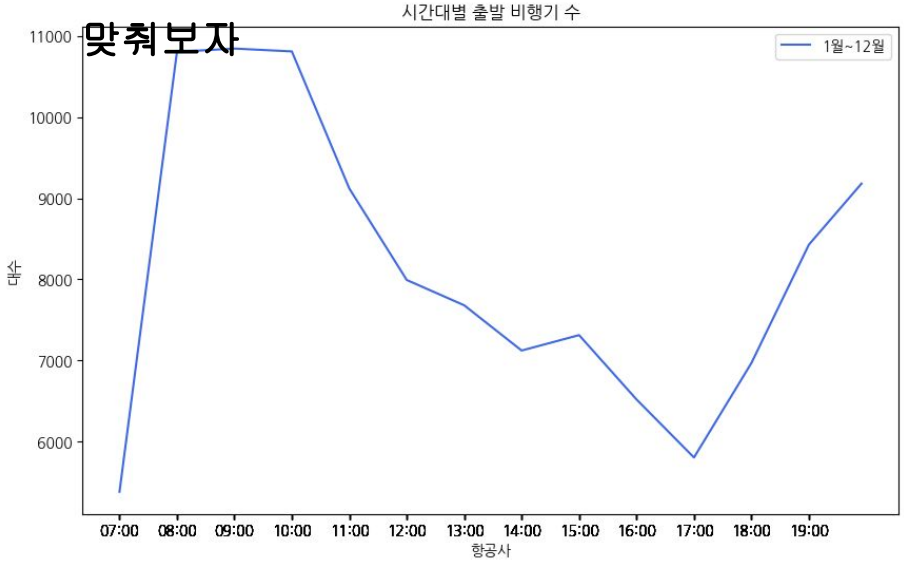
ANALYSIS

혹시 특정 시간에 비행기 운항이 적으면 가격이 상승할까?

3일간 가격을 평균



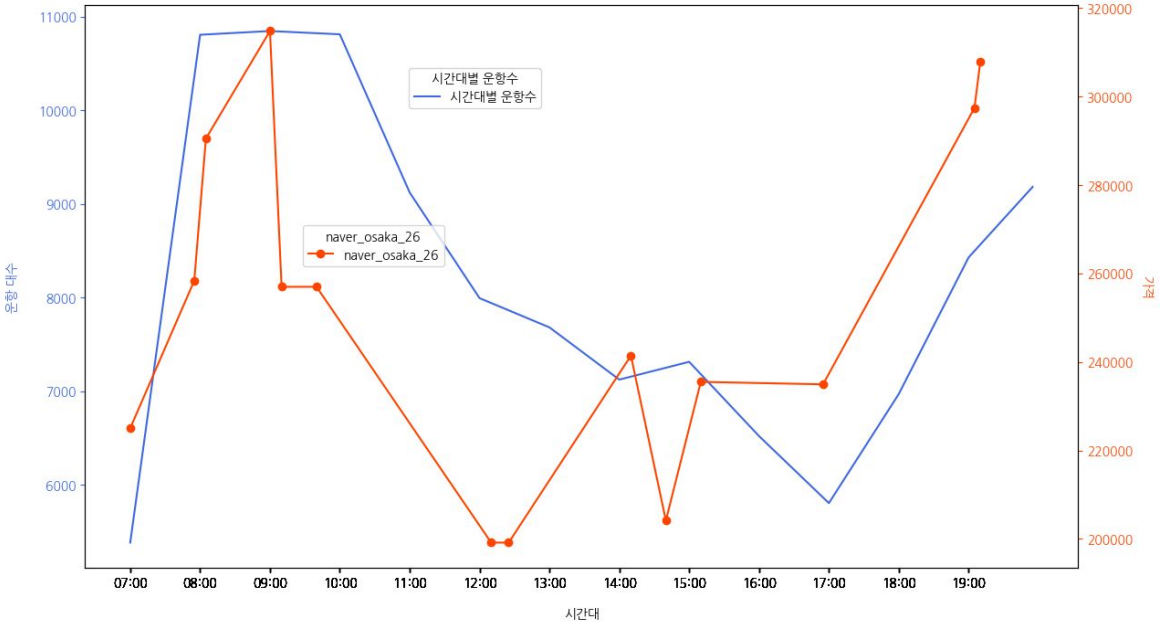
운항 데이터를 가격 형성 그래프 시간대로 맞춰보자



ANALYSIS

혹시 특정 시간에 비행기 운항이 적으면 가격이 상승할까? - mini 결론

시간대별 비행기 출발 수와 오사카 26일행 가격 비교분석



- 아침, 저녁 비행기 대수가 많고 가격도 비싸게 팔더라!

- 비행기 운항대수에 비슷하게 가격도 비싸지거나 싸진다!

ANALYSIS

끝맺음말

오늘 언제 들어가야 제일 싸게 살 수 있을지는 더 많은 시계열 데이터가 필요했다

그렇지만 조금이라도 더 싸게 사고 싶은 마음

그래서 가격에 연관을 주는 요소에 대해 분석도 해본 결과,

충분한 시계열 데이터 분석과 연관지을 때 유의미한 결과를 얻을 수 있을 것이라고 기대한다.

The End

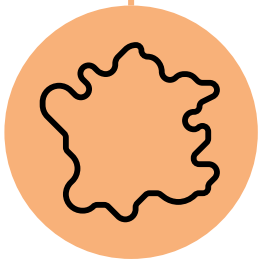
Thoughts

동규

데이터 전처리 과정이
시간이 오래걸리고
어려운 것이라는 경험을
할 수 있었다.

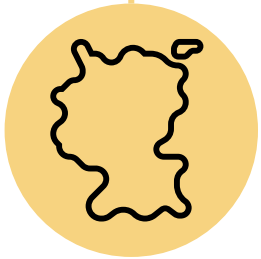
수빈

항공 관련데이터를
더 가져와 다른 결과도
도출 해보고 싶다



다연

데이터 어려워!
그치만
재미있어!
여행가고 싶어!



가은

데이터 전처리의
세계는
어려웠다.
연관성 조작하고
싶었습니다.