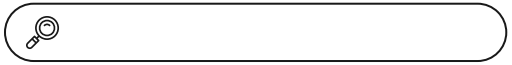
빅데이터 기반 프로야구 인기도 지표 분석 및 구단별 인기 기여 정도 파악

IT공학전공 2116313 손수경

Table of contents

01

주제 선정 배경 및 목표

주제 선정 배경 및 목표

2024시즌 kbo리그의 폭발적 흥행! 20세대의 인기를 끌다!

흥행 돌풍…1000만 관중 시대 원동력은[프로야구 흑자전환 토대①]

[야구] 준PO 5차전도 매진… PS 12경기 연속 만원 관중

엘롯기 동반 가을야구? 흥행도 이끈다

'전반기 353만 관중' 프로야구, 최고의 흥행보증 수표는 역시 엘롯기?

*엘롯기: LG/롯데/기아 의 줄임말

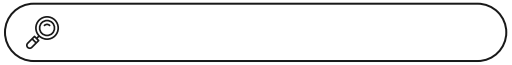
주제 선정 배경 및 목표

1. 야구 흥행의 기준이 오로지 관중수일까?
2. 엘롯기가 과연 야구의 흥행 보증수표일까?

주제 선정 배경 및 목표

목표

1. 관중수 외 야구 인기에 영향을 미치는 다른 지표도 알아보자
2. 2024시즌 엘롯기가 야구의 흥행 보증수표인지 알아보자

02

데이터 수집 과정

데이터 수집 과정

목표 1) 관중수 외 야구 인기에 영향을 미치는 다른 지표도 알아보자

2010년 ~ 2024년까지 데이터 수집

1. 연도별 관중수(Target) -> KBO 사이트
2. 웹 검색량
   -> 구글 트랜드(pytrends), "야구", "KBO", "야구장", "프로야구" 검색어 사용
3. 연도별 기사 개수 -> 네이버 스포츠 기사(selenium(셀레니움)을 통한 크롤링)
4. 유튜브 데이터(조회수/댓글수/좋아요수) -> Youtube-Scraper를 통한 크롤링

데이터 수집 과정

목표 2) 2024시즌 엘롯기가 야구의 흥행 보증수표인지 알아보자

2024년 구단별 데이터 수집

1. 구단별 관중수(Target) -> KBO 사이트
2. 구단별 웹 검색량 -> 구글 트랜드(pytrends), 구단명을 검색어로 사용
3. 구단별 기사 개수
   -> 네이버 스포츠 기사(selenium(셀레니움)을 통한 크롤링)
4. 구단별 유튜브 데이터(조회수/댓글수/좋아요수)
   -> Youtube-Scraper를 통한 크롤링

데이터 수집 과정

연도별 데이터

year	audience_by_year	article_counts_by_year	web_search_count_by_year	youtube_views_by_years	youtube_likes_by_years	youtube_comments_by_years
2010	5928626	107644	579	0	0	0
2011	6810028	172803	1091	0	0	0
2012	7156157	278297	1031	0	0	0
2013	6441945	299639	846	0	0	0
2014	6509915	290800	689	0	0	0
2015	7360530	283386	920	102705	337	39
2016	8339577	129755	704	1549632	7500	640
2017	8400688	131209	860	48283	326	22
2018	8073742	113978	763	4486869	24065	1468
2019	7286008	107400	601	477399	4220	313
2020	328317	121161	526	265685	4204	783
2021	1228489	101763	559	4117281	35720	5291
2022	6076074	107506	543	36525688	450616	13088
2023	8100326	113231	599	36600848	562298	21027
2024	10887705	106542	652	115721472	2109570	274132

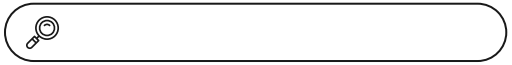
연도 관중수 기사개수 웹 검색량 ← 유튜브 데이터 →

데이터 수집 과정

구단별 데이터

team	article_count	audience_by_year	web_search_count	youtube_views_count	ranking	rank_reverse
HH	8752	804204	3085	93528076.0	8	3
HT	12962	1259249	1823	56900930.0	1	10
KT	5678	843942	3214	13134334.0	5	6
LG	9125	1397499	3574	35108779.0	3	8
LT	6949	1202840	3204	52627787.0	7	4
NC	3850	731167	3368	13899768.0	9	2
OB	6715	1301768	3625	35046297.0	4	7
SK	3666	1143773	3508	30310968.0	6	5
SS	9917	1347022	3188	44700268.0	2	9
WO	3397	808349	4106	15002719.0	10	1

팀명 기사개수 관중수 웹 검색량 유튜브 조회수 순위 순위 역순

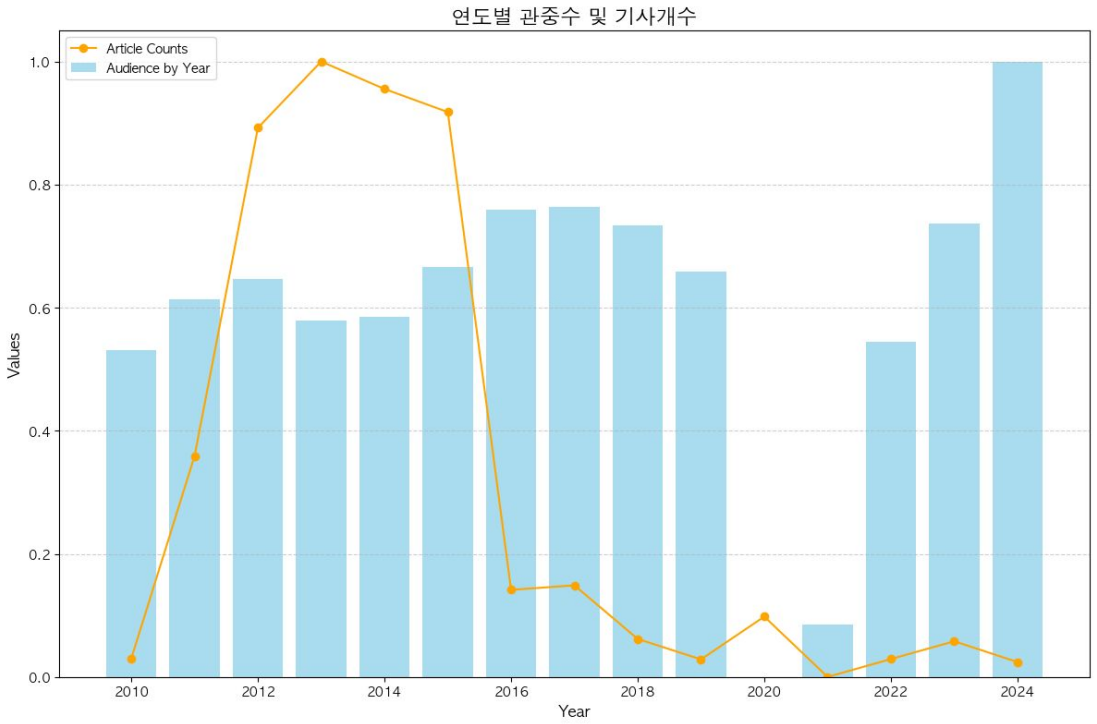



03

EDA를 통한 데이터 설명

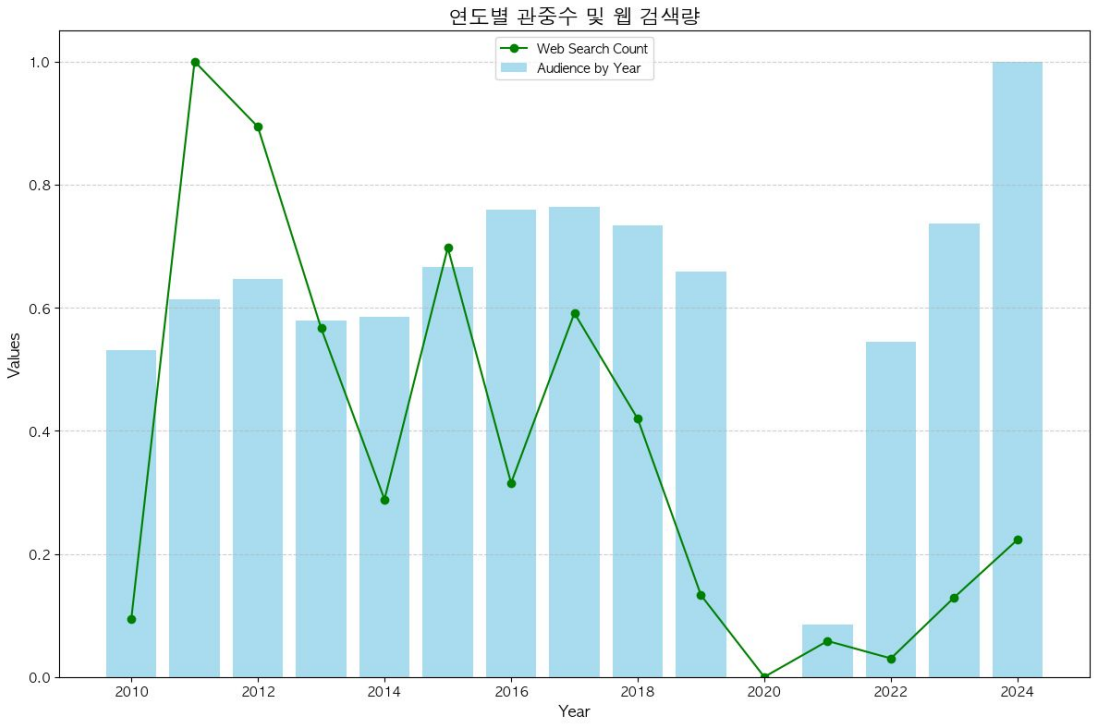
EDA를 통한 데이터 설명

Bar 그래프: 관중수
Plot 그래프: 기사개수



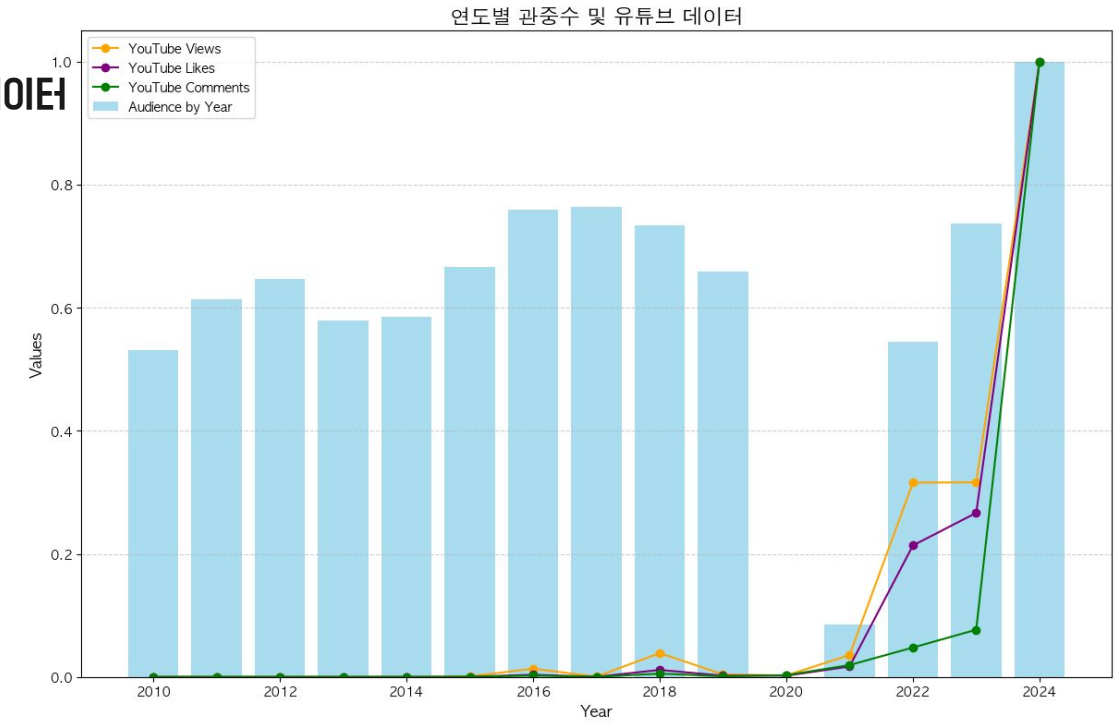
EDA를 통한 데이터 설명

Bar 그래프: 관중수
Plot 그래프: 웹 검색량



EDA를 통한 데이터 설명

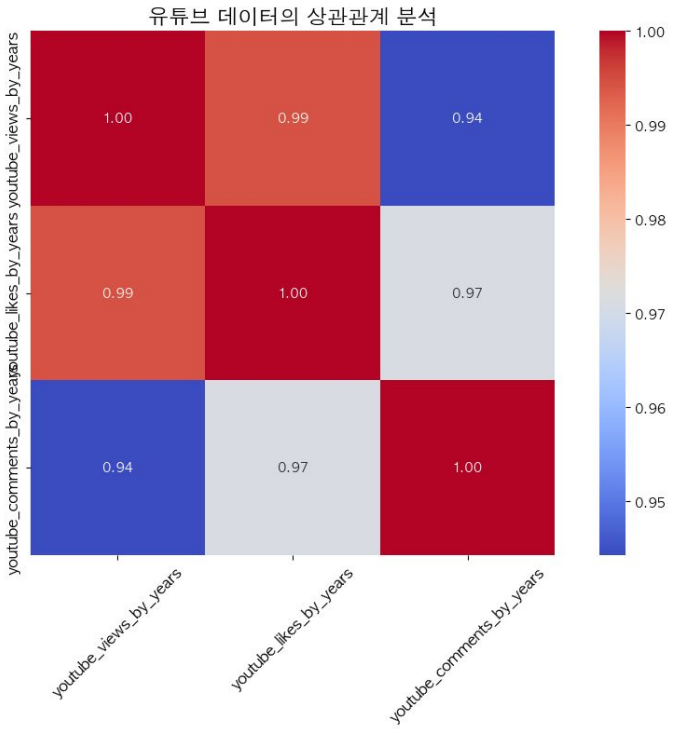
Bar 그래프: 관중수
Plot 그래프: 유튜브 데이터



주황: 조회수
보라: 좋아요수
초록: 댓글수

EDA를 통한 데이터 설명

유튜브 데이터 ->
다중공선성 의심



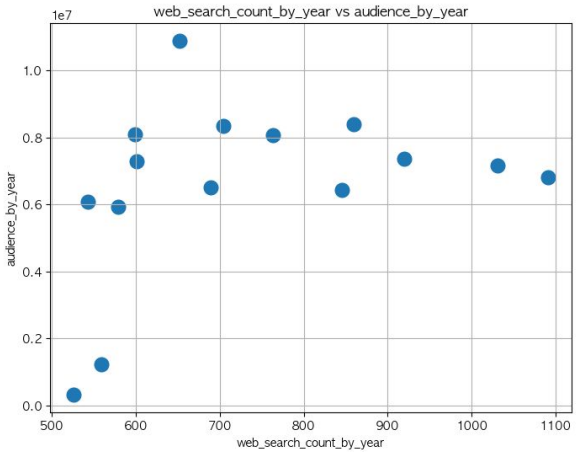
분산 팽창 계수를 통한
다중공선성 확인

Feature	VIF
article_counts_by_year	1.731441
web_search_count_by_year	1.705304
youtube_views_by_years	403.478505
youtube_likes_by_years	742.699860
youtube_comments_by_years	71.550160

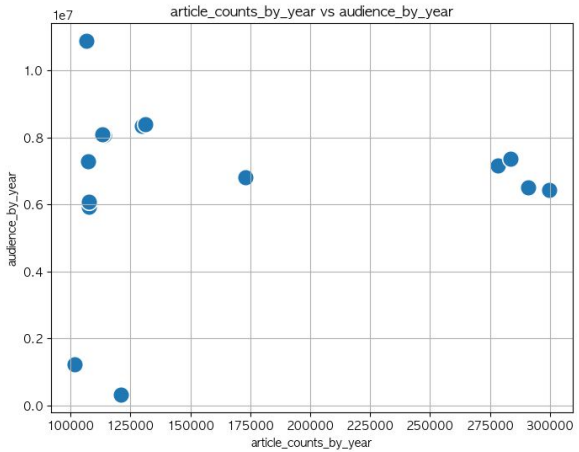
VIF가 10이상으로 독립변수들 간
상관성이 매우 높다

EDA를 통한 데이터 설명

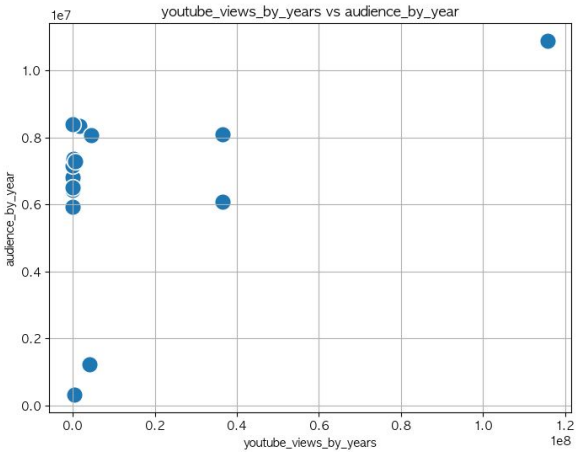
선형성 파악하기



웹검색량(x)과 관중수(y)



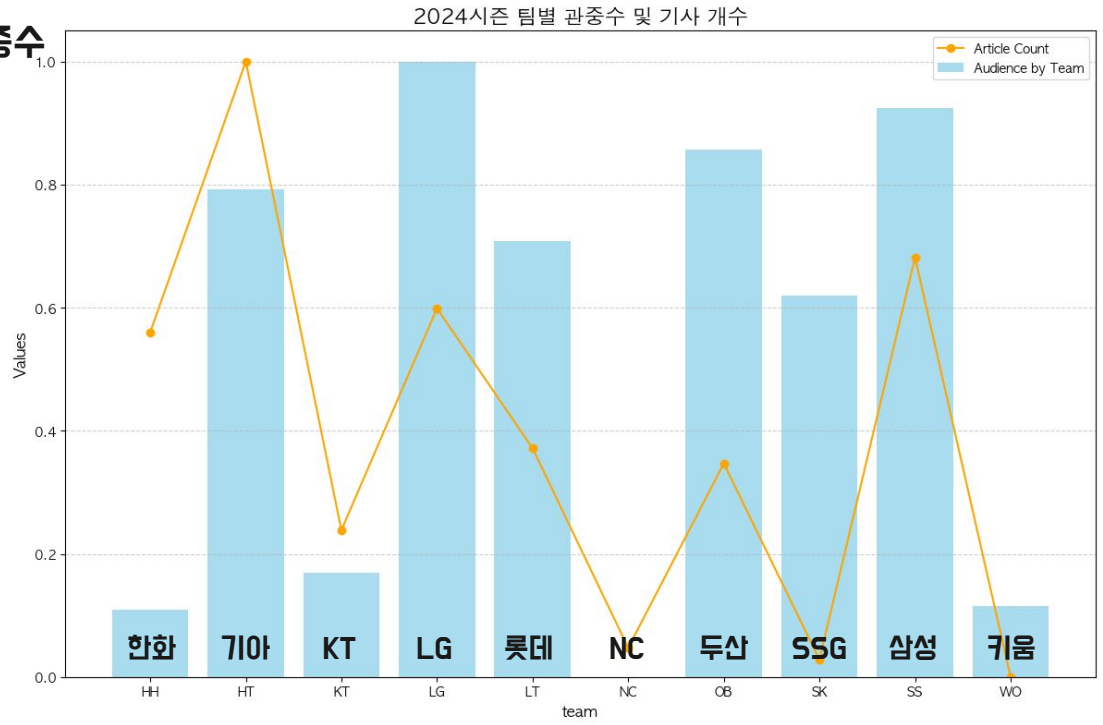
기사개수(x)와 관중수(y)



유튜브 조회수(x)와 관중수(y)

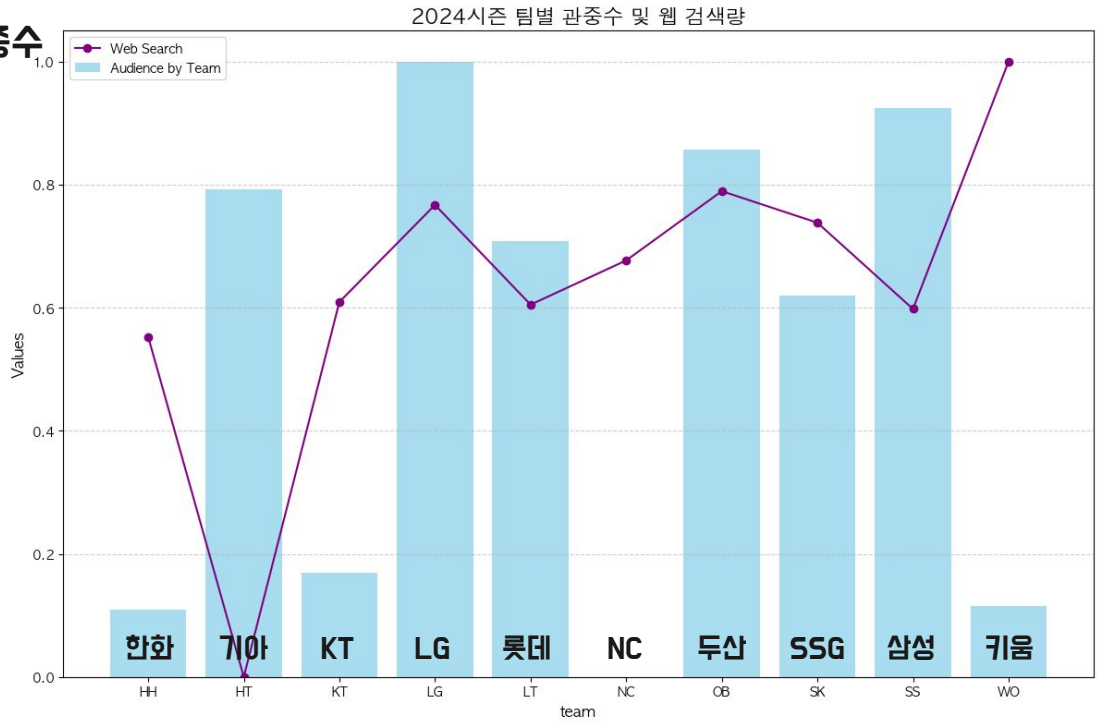
EDA를 통한 데이터 설명

Bar 그래프: 구단별 관중수
Plot 그래프: 기사개수



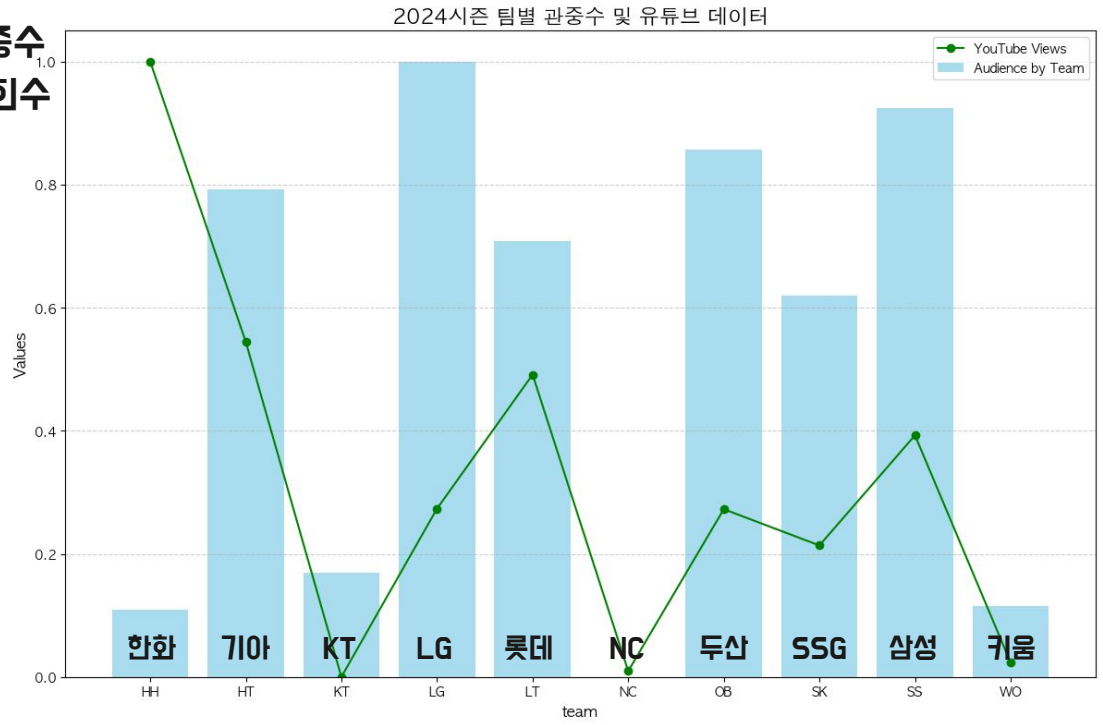
EDA를 통한 데이터 설명

Bar 그래프: 구단별 관중수
Plot 그래프: 웹 검색량



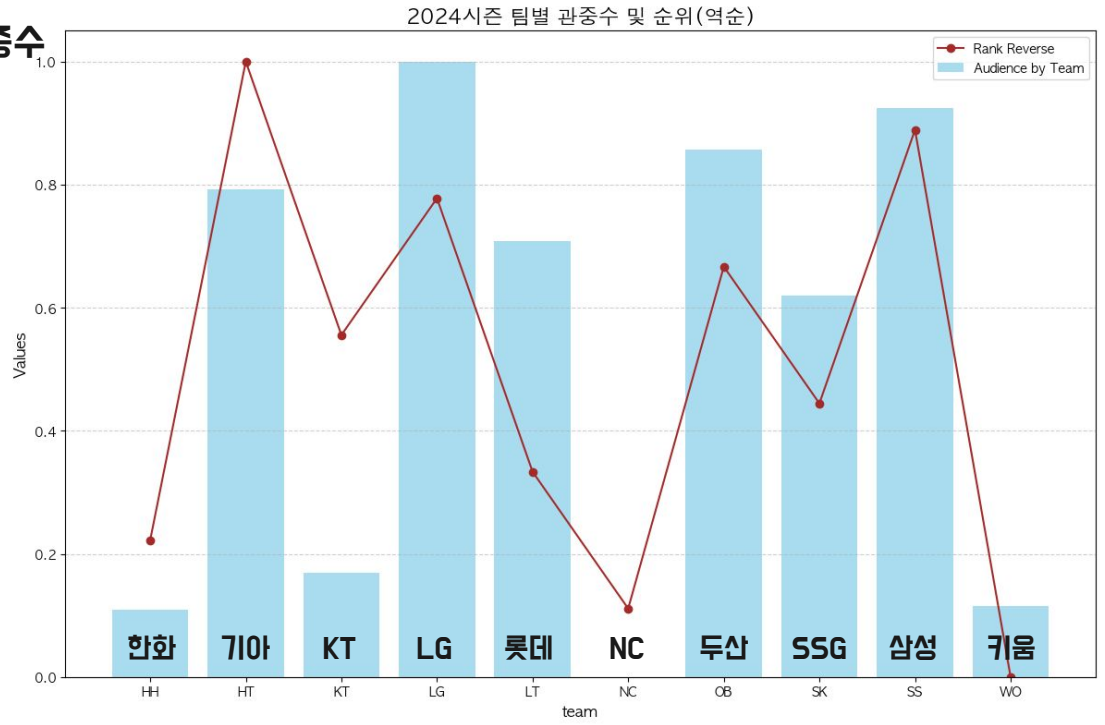
EDA를 통한 데이터 설명

Bar 그래프: 구단별 관중수
Plot 그래프: 유튜브 조회수



EDA를 통한 데이터 설명

Bar 그래프: 구단별 관중수
Plot 그래프: 순위(역순)



04

분석 및 결론

분석 및 결론

목표 1) 관중수 외 야구 인기 영향을 미치는 다른 지표도 알아보자

전제 조건: 야구의 인기도는 관중수로 수치화되었다고 가정한다.
사용 모델: 랜덤 포레스트(RandomForestRegressor)
비선형적인 다중 피처들의 상호작용을 고려하고, 피처의 기여도를 확인할 수 있다.
모델 사용 이유:

1. 상관 관계 분석 → 정규성 만족 X, p-value 수치상 유의미하지 않음
2. 다중 회귀 모델
   → 독립변수/종속변수 간의 선형성을 가지지 않음, 유튜브 데이터가 다중공선성을 갖음

분석 방법:

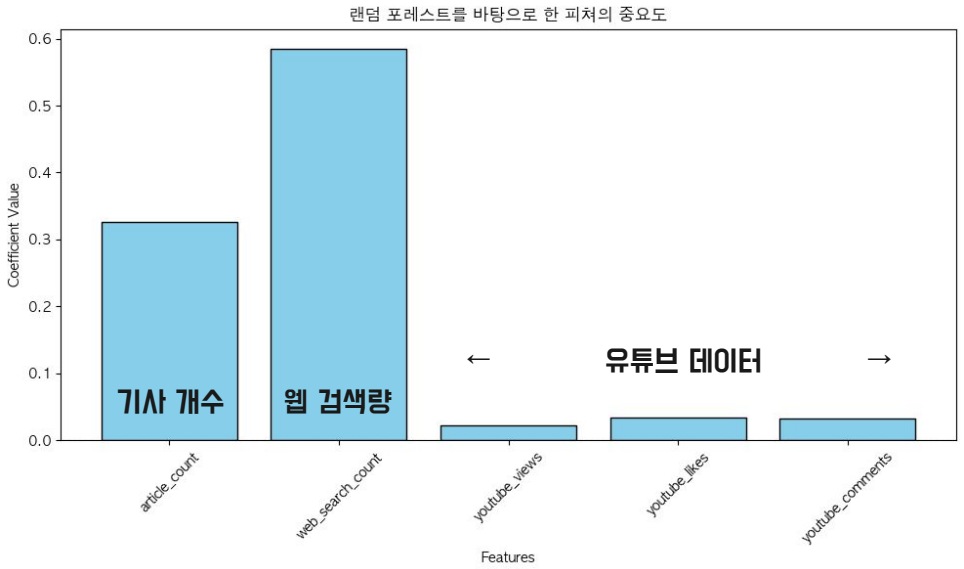
1. 데이터 스케일링
2. 랜덤 포레스트 모델 사용 -> feature_importances_를 통해 피쳐별 가중치 파악

결과: MSE → 0.07로 모델 성능은 괜찮다!

분석 및 결론

목표 1) 관중수 외 야구 인기 영향을 미치는 다른 지표도 알아보자

결론: 웹 검색량과 기사 개수는 야구 관중수(인기도)에 영향을 미친다.



랜덤 포레스트를 바탕으로 한 피쳐의 중요도

분석 및 결론

목표 2) 2024시즌 엘롯기가 야구의 흥행 보증수표인지 알아보자

전제 조건: 야구의 인기도는 관중수로 수치화되었다고 가정한다.
분석 방법:

1. 목표 1에서 구한 피쳐들의 가중치를 활용하여
   2024시즌 팀별 야구 인기도에 기여한 정도를 분석한다

$$score=(web\_search\_count\times w1)+(article\_count\times w2)+(youtube\_likes\times w3)+(youtube\_comments\times w4)+(youtube\_views\times w5)$$

- $w1$ =0.5854: 웹 검색량 가중치
- $w2$ = 0.3263: 기사 개수 가중치
- $w3, w4, w5$ = 0.0338, 0.0322, 0.022: 유튜브 데이터 가중치
- score = 야구 인기도에 기여한 수치

분석 및 결론

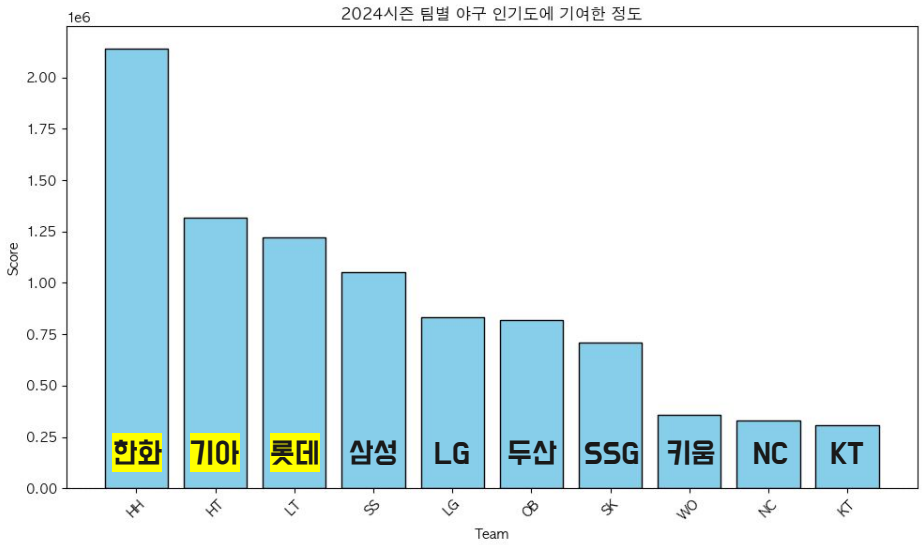
목표 2) 2024시즌 엘롯기가 야구의 흥행 보증수표인지 알아보자

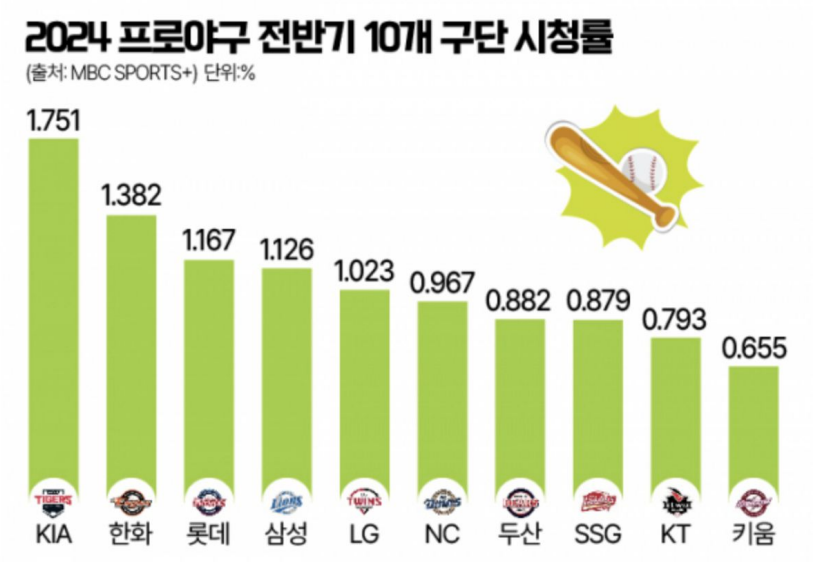
결론

야구의 흥행 보증수표는 **한화/기아/롯데** 였다.

분석 및 결론 (번외)

프로야구 전반기 시청률 기록 야구의 흥행 보증수표는 한화/기아/롯데 였다.





05

시사점

시사점

관중수 외의 웹 검색량, 기사 개수 등의 데이터도 대중의 관심도를 반영할 수 있습니다.
특히 한화/기아/롯데의 인기는 야구의 흥행에 영향을 미친다는 결과가 나왔습니다.

→ 마케팅 전략, 방송 편성에 필요한 준비를 할 수 있을 것으로 기대합니다.

06

한계점

한계점

1. 데이터의 부족함

   시청률 데이터, 인스타그램 데이터, 굿즈 판매량, 팬카페 가입자수 등 야구의 인기도를 판단할 수 있는 지표는 다양하다.

   → 하지만 시간상 수집에 제한이 있다.

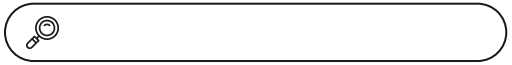
2. 강력한 전제 조건

   야구의 인기도는 관중수와 비례하다.

3. 유튜브 데이터의 한계

   2010년~2014년까지의 영상 게시가 없다. 즉, 일관된 데이터 수집이 어려웠다.

질의응답


감사합니다.