탐색적자료분석 및 시각화

기말과제-1

데이터 보고서

응용통계학과

32193335 이원병

차례

Ⅰ 데이터 소개 및 출처

Ⅱ 데이터의 각 변수들에 대한 설명

Ⅲ 각 변수들의 특징 분석

Ⅰ **데이터 소개 및 출처**

이 데이터는 미국프로야구(MLB) 2022년도 정규시즌(포스트시즌 포함x) 전체 투수들의 기록이다. Mlb.com에 기록된 각종 투수들의 데이터들 중 연관성이 있거나 의미를 도출할 수 있을 만한 데이터변수들을 추려서 최종 데이터로 생성하였다. 또한, 1경기라도 등판한 기록이 있는 모든 선수들을 포함할 시 데이터의 양이 너무 커지는 관계로, 규정이닝인 162이닝의 약 50%이상을 충족하는 선수들의 데이터로 한정하였다.

출처: https://baseballsavant.mlb.com/leaderboard/custom?year=2022&type=pitcher&filter=&sort=8&sortDir=asc&min=250&selections=player\_age,p\_game,p\_formatted\_ip,p\_total\_hits,p\_home\_run,p\_strikeout,p\_walk,p\_k\_percent,p\_bb\_percent,batting\_avg,slg\_percent,on\_base\_percent,on\_base\_plus\_slg,p\_earned\_run,p\_run,p\_save,p\_win,p\_loss,p\_era,p\_quality\_start,p\_complete\_game,p\_hit\_by\_pitch,p\_hold,in\_zone\_percent,edge\_percent,f\_strike\_percent,fastball\_avg\_speed,fastball\_avg\_spin,breaking\_avg\_speed,breaking\_avg\_spin,&chart=false&x=player\_age&y=player\_age&r=no&chartType=beeswarm

Ⅱ **데이터의 각 변수들에 대한 설명**

Raw 데이터는 위에서 설명한 바와 같이 규정이닝의 50%이상을 충족하는 선수들로 한정한 결과 총 275명의 선수로 구성되어 있다. Column 변수는 총 30개로, 전체 275x30=8,250개의 데이터가 존재한다.

<Column 변수 정의>

Last\_name : 성

First\_name : 이름

Player\_id : mlb.com에서 임의로 정해놓은 선수별 식별번호

Year : 2022시즌의 기록이므로 전부 2022로 동일

Player\_age : 나이

P\_game : 등판한 경기 횟수

P\_formatted\_ip : 투구 이닝

P\_total\_hits : 전체 피안타(상대 타자에게 허용한 안타)

P\_home\_run : 전체 피홈런(상대 타자에게 허용한 홈런)

P\_strikeout : 탈삼진(상대 타자로부터 잡아낸 삼진)

P\_walk : 사사구

P\_k\_percent : 탈삼진율(잡아낸 모든 아웃카운트 중 탈삼진 비율)

P\_bb\_percent : 사사구율(잡아낸 모든 아웃카운트 중 사사구 비율)

Batting\_avg : 피안타율

Slg\_percent : 피장타율

On\_base\_percent : 피출루율

On\_base\_plus\_slg : 피OPS(출루율+장타율)

P\_earned\_run : 자책점(마운드에 있는 동안 야수의 실책 등으로 인한 실점을 제외한 허용한 점수)

P\_run : 실점(마운드에 있는 동안 허용한 총 점수)

P\_save : 세이브(특정요건에 의해 리드 시 등판하여 경기를 승리로 마무리한 경우)

P\_win : 승리

P\_loss : 패배

P\_era : 평균자책점, 방어율(9이닝당 자책점의 평균)

P\_quality\_start : QS(6이닝 이상 3자책점 이하 기록 시)

P\_complete\_game : 완투(한 경기를 혼자 끝낸 경우)

P\_hit\_by\_pitch : 사구

P\_hold : 홀드(리드 시 그 리드를 유지한 채로 다음 투수에게 넘겨준 경우)

In\_zone\_percent : 스트라이크존에 던질 확률

Edge\_percent : 스트라이크존 각 방향의 모서리에 던질 확률

F\_strike\_percent : 초구에 스트라이크를 던질 확률

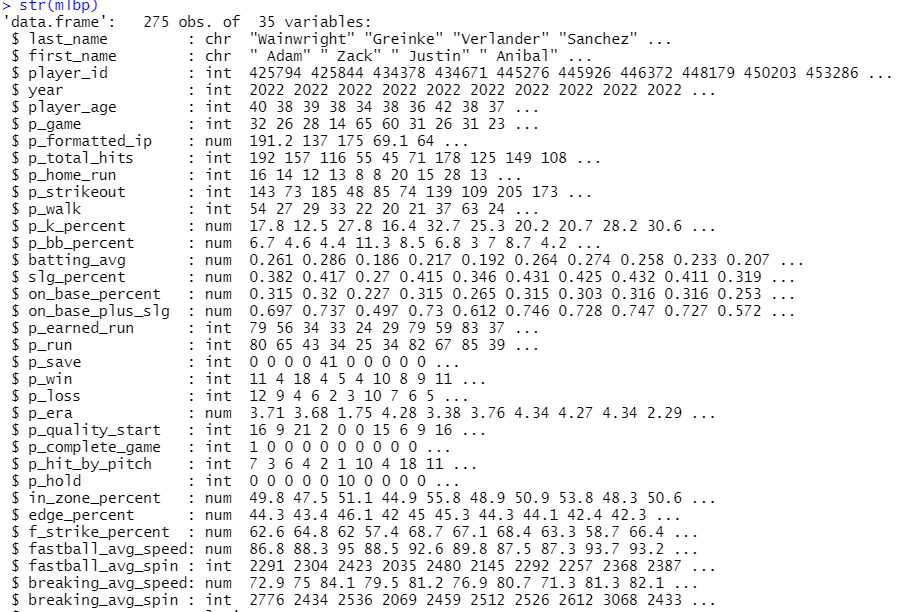
Fastball\_avg\_speed : 패스트볼의 평균 구속(Mile)

Fastball\_avg\_spin : 패스트볼의 평균 회전수(RPM)

Breaking\_avg\_speed : 변화구의 평균 구속(Mile)

Breaking\_avg\_spin : 변화구의 평균 회전수(RPM)

<변수 타입>



-> 선수의 성과 이름을 제외한 나머지 변수들은 모두 numeric(숫자형)과 integer(정수형)타입으로 구성되어 있다. 또한 대부분의 변수들이 연속형 변수이며, 범주형 변수가 존재하지 않는다. 따라서 범주형 변수를 추가하기 위해 각 선수들이 뛴 리그를 AL(American League)와 NL(National League)로 구분해 변수를 추가할 예정이다.

Ⅲ **각 변수들의 특징 분석**

위에서 살펴본 변수들의 타입 중 numeric(숫자형) 변수들은 대부분 확률을 나타낸다.

(p\_k\_percent)



-> 탈삼진율을 나타내는 p\_k\_percent 변수의 평균은 22.58로 전체 아웃카운트를 100으로 설정했을 때, 삼진을 잡아낸 비율이 평균 22.58%라는 의미이다. 또한 최솟값이 12.5로 275명의 투수는 적어도 12.5% 이상을 삼진으로 기록했다는 것을 확인할 수 있다. 탈삼진율이 높으면 그만큼 삼진을 잘 잡는다는 뜻이다. 하지만 탈삼진율 하나만으로 그 선수의 종합적인 능력을 평가하기는 어렵다.

(batting\_avg)



-> batting\_avg 변수는 피안타율을 뜻하는데, 이 변수의 범위는 이론적으로 0~1 사이의 값을 가진다. 하지만 안타를 단 하나도 허용하지 않거나, 무조건 안타를 허용하는 경우는 사실상 불가능에 가깝기 때문에 0과 1은 없다고 봐도 무방하다. 275명의 선수들의 2022시즌 피안타율의 평균은 약 0.242이다. 피안타율의 경우 보통 2할 초반대 이하를 기록하는 경우 수준급 선수라고 평가한다. 1사분위수인 0.220 근방이 그 기준이 될 수 있을 것이다. 또한 피안타율이 낮을수록 상대타자가 그 투수의 공을 안타로 만들기 어렵다는 의미이므로 공의 힘이나 위력과 연관이 있을 가능성이 높다.

(p\_era)



-> p\_era 변수는 위의 변수들과 달리 확률을 나타내지 않는다. 그 선수가 던진 총 이닝과 기록한 총 자책점을 바탕으로 9이닝당 자책점으로 환산하여 나타낸 수치가 바로 평균자책점이다. 평균자책점은 투수를 설명하는 가장 대표적인 지표이다. 물론 다른 지표도 종합적으로 봐야겠지만, 평균자책점이 2.5 이하이면 mvp급 투수로 평가받는다. 위의 수치만 봐도 평균이 3.89, 1사분위수도 3.1로 나타나 낮은 평균자책점을 기록하는 것이 얼마나 어려운 일인지 알 수 있다.

(fastball\_avg\_speed)



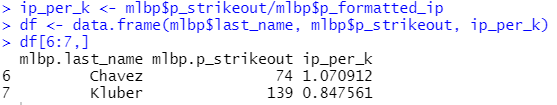
-> fastball\_avg\_speed 변수는 말그대로 직구평균구속을 나타낸다. 275명의 직구평균구속은 약 93마일로, km로 환산 시 약 149.7km이다. 한국프로야구인 KBO리그에서 평균구속이 150km이 넘는 투수가 한해에 다섯손가락 안에 드는 것과 비교하면 mlb의 평균구속은 상당히 빠르다.

Numeric(숫자형) 변수와 다르게 integer(정수형) 변수는 기록한 수치의 개수를 있는 그대로 나타내는 경우가 많다.

(p\_strikeout)

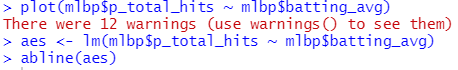
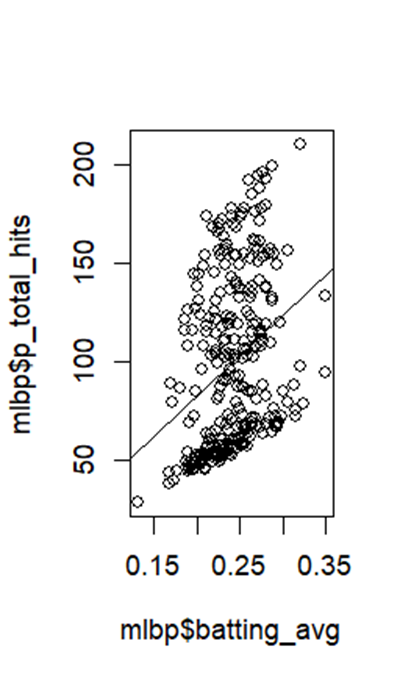


-> p\_strikeout 변수는 한 시즌동안 기록한 탈삼진 개수이다. 다만, 이 수치만으로 그 선수가 삼진을 많이 잡는다고 단정지어서는 안된다. 그 선수의 이닝을 고려하지 않은 수치이기 때문이다.



-> 실제 데이터 중 그 예를 찾아보면, Chavez는 Kluber보다 탈삼진 개수가 훨씬 적지만 이닝 당 탈삼진(총 탈삼진 개수/ 총 이닝)을 계산해보면 Chavez의 ip\_per\_k 수치가 Kluber보다 높은 것을 확인할 수 있다.

(p\_total\_hits)



-> p\_total\_hits 변수는 마찬가지로 허용한 총 안타의 개수를 의미한다. 피안타의 개수가 많으면 통상적으로 피안타율도 높아진다. 즉, p\_total\_hits와 앞에서 살펴본 batting\_avg 변수와 서로 연관성을 가진다. 두 변수로 그래프를 그려본 결과, 피안타수가 많아질수록 피안타율도 높아지는 현상을 발견할 수 있다.