[은경] 크롤링

<http://www.kbreport.com/> 7/17 진행, 월별 기준

kbreport\_hit\_main.csv : 월별 경기수, 승, 무, 패, 승률, R/G, 득점, 홈런, 도루, 볼넷비율, 삼진비율, BABIP, 타율, 출루율, 장타율, OPS, wOBA

kbreport\_hit\_standard.csv : 월별 경기수, 타석수, 타수, 연타, 단타, 2루타, 3루타, 홈런, 득점, 타점, 볼넷, 고의사구, HBP, 삼진, 희생플라이, 희생타, 병살, 도루, 도실, 타율

kbreport\_hit\_detail.csv : 월별 경기수, 타석수, 타율, BABIP, 볼넷비율, 삼진비율, 볼넷과 삼진의 비율, ISO, 타수와 홈런 비율, OPS, RC, RC/27, wRC, SPD, wSS, 도루실패율, wOBA

kbreport\_pitcher\_main.csv : 월별 경기수, 승, 무, 패, 승률, RA/G, 실점, 자책, 피홈런, 볼넷비율, 삼진비율, BABIP, LOB비율, ERA, RA9-WAR, FIP, kFIP, WAR

kbreport\_pitcher\_standard.csv : 월별 경기수, 승, 패, 세이브, 홀드, 블론, QS, 이닝, 타자, 안타, 2루타, 3루타, 홈런, 실점, 자책, 삼진, 볼넷, 고의사구, HBP, 폭투, 보크, PK, 도루, 도루실패

kbreport\_pitcher\_detail.csv : 월별 경기수, QS, 완투, 완봉, 이닝당 삼진, 이닝당 볼넷, 이닝당 피홈런, 삼진/볼넷, 피안타율, 피출루율, 피장타율, 피OPS, WHIP, BABIP, LOB비율, ERA, RA9-WAR, FIP, kFIP, WAR

<https://www.koreabaseball.com/> 7/18 진행, 연도 기준

team\_hitter.csv : 연도별, 팀별, 타율, 경기, 타석, 타수, 득점, 연타, 2루타, 3루타, 홈런, 루타, 타점, 희생번트, 희생플라이

team\_pitcher.csv : 연도별, 팀별, 평균자책점, 경기수, 승, 패, 세이브, 홀드, 승률, 이닝, 피안타, 홈런, 볼넷, 사구, 삼진, 실점, 자책점, 이닝당출루허용률

team\_defense.csv : 연도별, 팀별, 경기수, 실책, 견제사, 자살, 보살, 병살, 수비율, 포일, 도루허용, 도루실패, 도루저지율

team\_runner.csv : 연도별, 팀별, 경기수, 도루시도, 도루허용, 도루실패, 도루성공율, 주루사, 견제사

2001-2020년 팀 기록 : 2000년까지는 양대리그로 양분해서 진행

2001년 수비에 주루사, 견제사 기록되지 않음

[현정] 크롤링

<http://www.giantsclub.com/> 7/18 진행

LotteHitterVS.csv : 연도별/상대팀별 기준, 경기수, 타수, 안타, 2루타, 3루타, 홈런, 타점, 볼넷, 사구, 삼진, 득점

LottePitcherVS.csv : 연도별, 상대팀별, 경기수, 타수, 피안타, 자책점, 승, 패, 세이브, 홀드, 이닝, 탈삼진, 볼넷, 방어율, 피홈런, 사구, 실점, 피안타율

LotteHomeHit.csv : 연도별, 홈/원정, 경기수, 타수, 안타, 2루타, 3루타, 홈런, 타점, 볼넷, 사구, 삼진, 득점

LotteHomeP.csv : 연도별, 홈/원정, 경기수, 타수, 피안타, 자책점, 승, 패, 세이브, 홀드, 이닝, 탈삼진, 볼넷, 방어율, 피홈런, 사구, 실점, 피안타율

LotteMonthHitter.csv : 월별, 경기수, 타수, 안타, 2루타, 3루타, 홈런, 타점, 볼넷, 사구, 삼진, 득점

LotteMonthPitcher.csv : 월별, 경기수, 타수, 피안타, 자책점, 승, 패, 세이브, 홀드, 이닝, 탈삼진, 볼넷, 방어율, 피홈런, 사구, 실점, 피안타율

LotteTeamRecord.csv : 연도별, 순위, 경기수, 승, 패, 무, 승률, 득점, 실점, 홈런, 도루, 타율, 방어율

<https://www.koreabaseball.com/Record/Player/HitterBasic/Basic1.aspx> 7/25 진행, 선수정보,연도,팀기준

KBO\_ALL\_HITTERS.csv: 연도별,선수이름,팀명,타율,경기수,타석,타수,득점,안타,2루타,3루타,홈런,루타,타점,희생번트,희생플라이

KBO\_ALL\_PITCHERS.csv:연도별,선수이름,팀명,평균자책점,경기수,승,패,세이브,홀드,승률,이닝,피안타,홈런,볼넷,사구,삼진,실점,자책점,이닝당 출루허용률

KBO\_ALL\_DEFENSE.csv:연도별,선수이름,팀명,포지션,경기수,선발경기수,이닝,실책,견제사,자살,보살,병살,수비율,포일,도루허용,도루저지,도루저지율

[성훈] 전처리

연도 및 월별 팀 승률과 타자기록 분석해서 상관계수가 0.3이상인 요소 ['R/G', '출루율', '장타율', 'OPS', '볼/삼', 'ISO', 'RC/27', '도실']들을 추출, 다른 포지션별 기록과 병합

병합한 기록을 토대로 2019년도 기록을 제외한 나머지 데이터로 승률예측모델 구축해서 2019년도의 승률예측하여 실제 승률과 비교

* 팀 데이터 모으기 (15~20년, 144경기 기준)
* 1군 선수 정보 모으기
* 포지션별 선수 정보 모아서 2020년 골든글러브 예측
* <http://www.statiz.co.kr/main.php>
* <http://www.kbreport.com/main>
* <https://www.koreabaseball.com/Default.aspx>
* <http://www.giantsclub.com/html/>
* <https://baseball-in-play.com/11>
* <https://blog.naver.com/hi_nso/130114915919>
* <https://londonboys.tistory.com/277>
* <https://datascienceschool.net/view-notebook/36176e580d124612a376cf29872cd2f0/>
* <https://musma.github.io/2019/05/15/make-a-prediction3.html>
* <https://dacon.io/competitions/official/62540/codeshare/659?page=1&dtype=recent&ptype=pub>
* <https://datascienceschool.net/view-notebook/37520f4c2c9240d6a7eff1149ca6bf49/>
* <https://docs.w3cub.com/scikit_learn/modules/model_evaluation/#r2-score>

-----------------------------------------------------------------------

* (◎) 1군이 되기 위해 지녀야할 기본 능력치 추출 -> 각 팀별 1군들 능력치를 순위 매김
  + 팀 기록으로 정리

* 순위에 영향을 미치는 요소는 무엇이 있는가?(상관관계)

ex) 팀타율,팀평균자책점, 팀 1군 평균 능력치

* 상관관계가 있는 요소들로 시계열을 이용하여 최종순위 예측
  + 기사
    - 팀 타율과 승률간에는 0.54의 양의 상관관계가 존재합니다. 팀 타율이 높을수록 승률도 증가한다는 의미입니다. 팀 방어율과 승률 간에는 -0.64의 음의 상관관계가 있는 것으로 나타났습니다. 팀 방어율이 좋지 않을수록 승률도 급격하게 낮아집니다.
  + 프로야구경기 분석을 통한 승패 예측 모형
    - 로지스틱회귀분석 : 집중력(타점/안타수), 총루타, 실책수, 희생타, 삼진, 잔루, 사사구, 도루수, 병살타, 홈팀여부, 장타율순으로 나타남. 집중력, 총루타, 실책수, 희생타, 삼진, 잔루 요인이 유의한 경기기록
    - 변수선택법에 의한 예측모형 : 집중력, 총루타, 실책수, 삼진, 희생타 순
    - 의사결정나무분석을 이용한 승패 예측결과 : 타격집중력, 장타율, 실책, 희생타, 삼진, 도루수
  + The development of a valuation model to determine the true market value of professional baseball players
    - 선수연봉 = 연차\*921.5 + FA유무\*53528.9 + 강타자 지수\*7313 + 교타자 지수\*2197.3 + 5893 (단위: 만원).
    - 제안된 모델은 한국프로야구 내 선수가치의 64.5%를 설명하는 것으로 나타났으며 통계적 유효성이 검증되었다.
* 웹 - 순위 그래프, 팀별 정보, 선수별 정보 나타내기
  + <https://www.finereport.com/kr/%ec%b0%a8%ed%8a%b8-%eb%b9%84%eb%b2%95-%eb%8c%80%ea%b3%b5%ea%b0%9c-%ec%8b%9c%ea%b0%81%ed%99%94%ec%9d%98-%ed%95%9c%ea%b3%84%eb%8a%94/>
  + <https://www.finereport.com/kr/%EA%B7%B8%EB%9E%98%ED%94%84-%EA%B7%B8%EB%A6%AC%EB%8A%94-%EC%82%AC%EC%9D%B4%ED%8A%B8-%ED%94%84%EB%A1%9C%EA%B7%B8%EB%9E%A8-top10-%EB%AA%A8%EC%9D%8C-2020%EC%B5%9C%EC%8B%A0/>
  + https://blog.naver.com/zhangyun123/221974565316

-----------------------------------------------------------------------

* 시간나면 하기
  + 롯데팀의 월별, 홈/원정별 강세 () -> 롯데는 정말 봄(시즌초반)에만 강한가?
  + 이 선수는 연봉만큼의 경기력을 보여주는가?
  + 구단마다 외국인 선수 성공/실패할 확률, 용병 스카웃을 잘 하는 팀

-----------------------------------------------------------------------

* ppt
  + 코로나 전후 항목 (한국 갤럽)
  + 데이터 출처 링크 - 법적 문제 명시 (취업 포트폴리오 대비)
  + 날씨, 더블헤더 등 나열할 수 있는 건 다 나열해서 이런것만 분석하겠다
  + 계절별 월별 능력치
  + 인터랙티브 / 동적
  + D3, 코랩 visualization
* 웹
  + 1-2개의 웹페이지로 생성 가능
  + 팀 선택하면 결과물이 보일 수 있게
  + <https://developers.google.com/chart>
* 갤럽 조사 : <https://www.gallup.co.kr/gallupdb/reportContent.asp?seqNo=1105>
  + 한국갤럽이 2020년 4월 28~29일 전국 만 18세 이상 1,000명에게 가장 좋아하는 프로야구 구단을 물은 결과 '기아 타이거즈'(10%), '삼성 라이온즈', '롯데 자이언츠'(이상 9%), '두산 베어스', '한화 이글스'(이상 8%), 'LG 트윈스'(6%), 'SK 와이번스', 'NC 다이노스'(이상 3%), '키움 히어로즈'(2%), 'KT 위즈'(1%) 순으로 나타났다. 성인의 41%는 특별히 좋아하는 국내 프로야구팀이 '없다'고 답했다.
  + 국내 프로야구 관심층(412명)이 좋아하는 구단은 기아(16%), 롯데(14%), 두산(13%), 한화(12%), 삼성(11%), LG(10%), SK, NC(이상 5%), 키움(3%), KT(2%) 순이다.
  + 국내 프로야구 관심 정도는 '매우 관심 있다' 20%, '어느 정도 관심 있다' 21%, '별로 관심 없다' 25%, '전혀 관심 없다' 31%며 2%는 의견을 유보했다.
  + 2020년 올해 어느 팀이 우승할 것으로 예상하는지 물은 결과, 17%가 '두산 베어스'를 꼽았고 그다음은 'SK 와이번스'(5%), '기아 타이거즈', '삼성 라이온즈'(이상 4%), 'LG 트윈스'(3%) 순이었으며 59%는 의견을 유보했다.
* 컬러차트
  + <https://s3-ap-northeast-2.amazonaws.com/blog.lotte.co.kr/wp-content/uploads/2020/05/KakaoTalk_20200507_103213180.png>
  + <http://www.cssdrive.com/imagepalette/index.php>
* 비더레 승부예측 : <https://www.legend2i.com/kboLeague/matchList.aspx>

[야구용어]

* R/G
  + 경기당 출루
* BABIP
  + 인플레이로 이어진 타구에 대한 타율을 계산하는 용어이다. 계산식은 (총 안타수-홈런)/(타수-삼진-홈런+희생플라이). 타자와 투수 모두에게 적용할 수 있는 스탯
  + 타자 : 라인드라이브 비율, 타자의 스피드, 운 고려
  + 뜬공이나 땅볼보다 라인드라이브 비율이 높고 빠른 발을 가진 타자는 BABIP가 높은 경향이 있다
* OPS
  + OBP (출루율) + SLG(장타율)
* wOBA
  + 타자의 생산력을 측정하는데 최근 들어 널리 쓰이는 비율스탯으로 선형 회귀 분석을 이용해서 만든 지표
  + wOBA = (0.72\*NIBB + 0.75\*HBP + 0.90\*1B + 0.92\*RBOE + 1.24\*2B + 1.56\*3B + 1.95\*HR) / (PA - IBB)
    - PA - 타석
    - IBB - 고의4구
    - NIBB - 고의4구 제외 볼넷
    - HBP - 몸에 맞는 공
    - RBOE - 실책으로 인한 출루
* HBP : 사구
* ISO : 순장타율
  + 장타율에서 타율은 뺀 값. {2루타 + 2×3루타 + 3×홈런} / 타수'로 구하기도 한다.
* RC 득점 생산률(%)
  + (A × B) / C
    - A=안타+볼넷+사구-도루자-병살타
    - B=루타수+0.52X(도루+희생타+희생플라이)+0.26X(볼넷+사구-고의 볼넷)
    - C=타수+볼넷+사구+희생타+희생플라이
  + 어떤선수가 팀이 득점을 올리는데 있어서 (출루/진루타/적시타/희생타)등으로 어느정도 기여하는지를 수치화한 스탯으로 이해하시면 됩니다
* RC/27
  + 9이닝동안 득점 생산률(%)
  + RC/(OX27)
    - O=타수-안타+도루자+병살타+희생타+희생플라이
  + 한타자로 9명 구성된 팀이 한경기를 치르면 몇 점이나 뽑을 것인가?를 평균수치화 한 것
* wRC
  + {((wOBA - 리그평균wOBA) / wOBA Scale) + (리그득점/리그타석)} \* 타석
  + 경기의 승패를 결정 짓는 것은 팀의 득점이다. 이러한 득점 생산력을 계산하기 위한 지표 중 하나다 wRC이다. wRC는 weighted RC의 약자지만 Bill James의 RC(득점 생산)와는 별 상관이 없고, wOBA에 기반해 계산되는 스탯이다. 또한 wRC는 비교대상이 0이다. 즉 타율, 출루율, 장타율이 모두 0인 선수가 있다고 가정하고 생산성을 비교하는 스탯이다.
* SPD
  + 야수의 스피드와 관련된 여러 기록을 활용해서 0~10점 스케일로 평가하는 스탯이다.
* LOB%
  + 잔류율
* ERA(방어율 또는 평균자책점)
  + (자책점 × 9) / 던진이닝수
* RA9-WAR
  + 투수의 경우 9이닝 당 평균 실점율인 RA9을 활용한 RA9-WAR도 함께 제공하고 있다.
  + RA9-WAR를 구하는 공식은 다른 모든 부분에 있어서 kWAR와 동일하며 조정실점(Adjusted RA)을 산출할 때 한 투수의 kFIP을 리그평균자책점과 리그평균실점의 비율로 나눠주는 대신, RA9을 대입시켰다.
* FIP
  + 오로지 투수 본인의 능력으로만 이루어진 결과들(삼진, 볼넷, 피홈런)을 취사선택하였으며, 평균적인 수비의 도움을 받았다고 가정하였을 때 이 투수의 ERA가 어느 정도의 값을 가지는게 타당(?)했을지를 연구하면서 FIP가 만들어졌다.
  + 해당 시즌에 ERA가 FIP보다 눈에 띄게 낮은 투수는 운이 좋았다, ERA가 FIP보다 눈에 띄게 높은 투수는 운이 나빴다라고 흔히 이야기한다.
  + FIP = (( 13\*홈런 + 3\*(볼넷+몸에 맞는공-고의4구) - 2\*삼진) / 이닝) + C
    - \*C = (( 9\*리그자책점 - ( 13\*리그홈런 + 3\*(볼넷+몸에 맞는공-고의4구) - 2\*리그삼진)) / 리그이닝)
* kFIP
  + 한국프로야구의 사건(Event)별 득점가치는 메이저리그와 다르다. 따라서 KBReport에서는 FIP의 공식을 활용하되 KBO의 득점환경에 맞춰 계수의 변화를 준 투수통계지표를 개발하였다.
  + 더불어 KBReport는 MLB의 FIP에서 문제점으로 지적되던 부분을 해결하기 위하여 다음과 같은 가공을 하였다.
    - 1) 투수의 피홈런 중 그라운드 홈런을 제외
    - 2) 볼넷과 몸에 맞는 공에 곱하는 계수를 분리
    - 3) 삼진의 계수를 구할 때 삼진의 득점가치를 사용
    - 4) 환산된 계수의 소숫점 둘째 자리까지의 반영
* WAR
  + 대체 수준 대비 승리 기여도
  + 이 지표의 값은 곧 승수를 의미하며, 이에 따라 한국어로 읽을 때는 n WAR라는 점수는 n승이라고 읽으면 적절하다. 다만 중요한 것은 대체 선수에 비해 몇 승을 더 기여했는지를 나타낸다는 것이다.
* PK
  + 견제사 도루 가능성있는 주자를 견제구로 아웃시킨 경우
* QS
  + 선발투수가 6이닝이상 던지고 3자책 이하로 막은 경우에 주어짐
* WHIP
  + 1이닝당 평균 피출루율
  + (볼넷 + 피안타) / 던진이닝수
* 기대승률
  + 팀득점^2 / (팀득점^2 + 팀실점^2)