2019.10.15

**data 전처리 결과 공유**

1. **새로운 변수 추가**
   * state\_new : usd\_pledged\_real이 usd\_goal\_real 이상이면 (즉, 모금 받은 금액이 목표 금액 이상이면) success, 미만이면 failed로 state 재정의 → 앞으로 이 컬럼을 타겟으로 쓰는 것이 좋을 것 같음
   * deadline\_dt / launched\_dt : deadline과 launched 날짜를 datetime type으로 변환한 것. 연-월-일 구조로 되어있음
   * term : deadline\_dt와 launched\_dt 날짜 사이의 간격
   * term\_str : term은 int type. 이를 string type으로 바꾼 것
   * term\_bin : term 컬럼 값 binning. string type
     + 1 : 10일 이하
     + 2 : 10일 초과 15일 이하
     + 3 : 15일 초과 21일 이하
     + 4 : 21일 초과 30일 이하
     + 5 : 30일 초과 45일 이하
     + 6 : 45일 초과 60일 이하
     + 7 : 60일 초과
   * usd\_goal\_real\_bin : usd\_goal\_real 컬럼 값 binning. string type
     + 1 : 500 이하
     + 2 : 500 초과 1000 미만
     + 3 : 1000 초과 3000 이하
     + 4 : 3000 초과 5000 이하
     + 5 : 5000 초과 10000 이하
     + 6 : 10000 초과 50000 이하
     + 7 : 50000 초과 100000 이하
     + 8 : 100000 초과

(다만 bin으로 끊는게 학습에도 좋을 지는 해 봐야..)

* + Plat\_Pop (Platform\_Popularity) : 킥스타터 플랫폼의 인기도 구글 검색 빈도로 정의. 월 단위로 끊고, 나라별로 구분해서 매핑 해줌.
  + Currency의 경우 1% 이상인 USD, GBP, EUR, CAD, AUD 만 고려
  + Country의 경우 0.5% 이상인 US, BR, CA, AU, DE, FR, IT, NL, ES 만 고려, N,0”는 모종의 이유로 파악이 안 되는 경우이므로 제외

1. **불필요한 row 삭제**
   * 모금 기간이 100일 초과하는 경우(약 14000일 정도 됨)는 지움. 모금 시작 날짜가 1970년으로 되어있었기 때문
   * usd\_goal\_real의 값이 100 이하인 경우는 지움. 약 10만원 이하로 펀딩을 받는 것은 무의미하다고 생각
2. **최종 데이터는 Used\_col\_fin.csv에 저장**