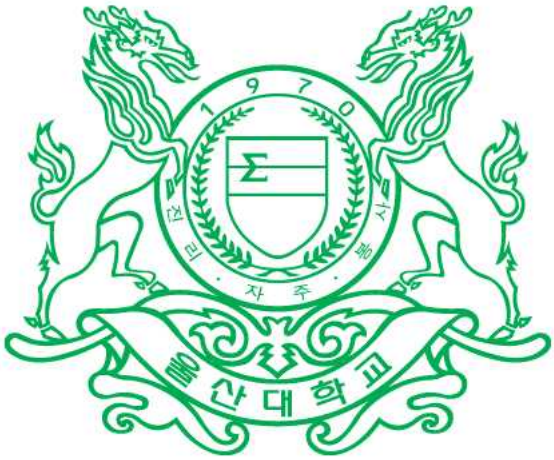
Report

데이터 사이언스 Term Project



과 목 : 데이터 사이언스

학 과 : IT융합

교수님 : 권영근 교수님

학 번 : 201522255

이 름 : 정찬호

제출일 : 2020.6.26

울산대학교

도입

현대 사회는 모바일 기기 보급 확대, 데이터 통신 고속화, 미디어 플랫폼 다양화 등의 기술 발전으로 미디어에 대한 접근성이 크게 향상되어 일상생활에서 자주 이용하는 미디어 기기로 스마트폰(93.9%)이 선택될 정도로 스마트폰으로 많은 콘텐츠를 소비하고 있고 스마트폰을 이용하면 누구나 장소에 구애받지 않고 실시간으로 미디어에 접근할 수 있다.

이러한 미디어 중 가장 대표적인 플랫폼이 바로 youtube이다. youtube는 여러 가지 영상을 시청할 수 있으며 영상을 업로드 하는 사람을 유튜버라고 칭하는데 청소년 희망 직업 순위에서 10위권에 진입할 정도로 현재 매우 인기 있는 직업이다. youtube에서 영상을 업로드 하는 가장 중요한 이유는 바로 수익을 창출하기 위해서다. 영상을 업로드하고 많은 수익을 창출하기 위해서는 영상의 조회 수가 많이 나와서 영상에 등록된 광고를 많은 시청자들이 시청해야 하는데 조회수가 많이 나올 법한 영상을 올리는 방법도 있겠지만 인기동영상에 선정이 된다면 많은 사람에게 노출이 되기 때문에 조회 수를 폭발적으로 증가 시키면서 채널홍보를 동시에 할 수 있는 방법이 바로 유튜브에서 선정하는 인기 동영상 목록에 선정되는 방법이다.

현재도 인기 동영상으로 선정되는 것은 많은 유튜버에게 수익 창출에 있어 가장 중요하고 가장 큰 과제이기도 하다.

2. 가설

(1) 유튜브에서 현재 가장 유행하는 카테고리 영상을 업로드하면 더 빨리 인기 동영상이 될 것이다.

(2) 가장 유행하는 카테고리 영상을 업로드 하면 다른 카테고리 보다 조회 수가 많이 나올 것이다.

(3) 태그의 개수가 많다면 접근성이 높아져 조회 수가 더 많을 것이다.

(4) 태그의 개수가 많다면 더 빨리 인기영상에 선정될 것이다.

3. 데이터 설명

video_id	비디오 id
trending_date	인기동영상 선정일
title	영상 제목
channel_title	채널 이름
category_id	카테고리 번호
publish_time	영상 업로드 시간
tags	태그
views	조회 수
likes	좋아요 수
dislikes	싫어요 수
comment_count	댓글 수
thumbnail_link	썸네일 링크
comments_disabled	comment 작성 불가 (True or False)
ratings_disabled	rating 불가 (True or False)
video_error_or_removed	비디오 오류, 삭제 (True or False)
description	설명

인기 youtube 동영상의 일일 기록 데이터 셋

trending_date : 17.14.11 년 일 월 순으로 등록되어 있음.

category_id : 카테고리가 번호로 각각에 맞는 고유 번호로 저장되어 있음

publish_time : : 2017-11-13T17:13:01.000Z 시간 정보가 다음과 같이 등록되어 있음.

tags : |를 기준으로 태그가 등록되어 있음.

numberphile|"prime numbers"|"proth prime"

데이터 출처 : https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv

4. 분석 방법 및 결과

4-0) csv 데이터 파일 전처리

4-0-(1) category 정보 매칭

US_category_id.json 파일을 읽어와 기존 csv 파일에 있는 category 번호와 매칭을 시켜 저장한다.

4-0-(2) 데이터값 필터링

영상 에러 or 삭제, 평가 및 댓글 불가 영상을 제외시키고 vedio_id가 식별 불가한 영상을 모두 삭제한다.

4-0-(3) 인기 동영상까지 걸린 시간 측정

영상의 업로드 시간과 인기동영상 날짜를 비교해서 datetime 으로 타입을 변경한 후 trending_date에서 publish_time을 뺀 값을 trending_days 라는 이름으로 저장한다.

4-0-(4)

Tag 개수를 split '|'를 기준으로 카운트하여 tag_count에 저장한다.

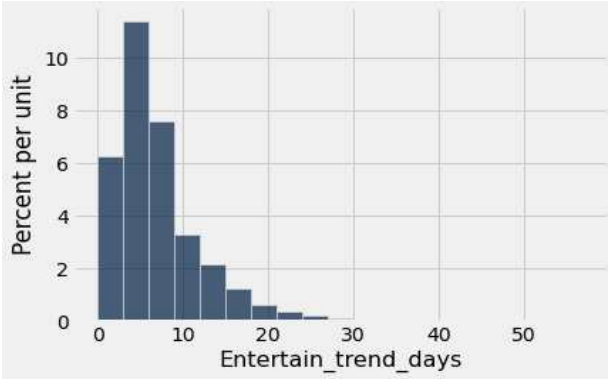
4-1) 현재 유행하는 카테고리 영상이 더 빨리 인기 동영상이 되는가?

4-1-(1) 데이터 선정

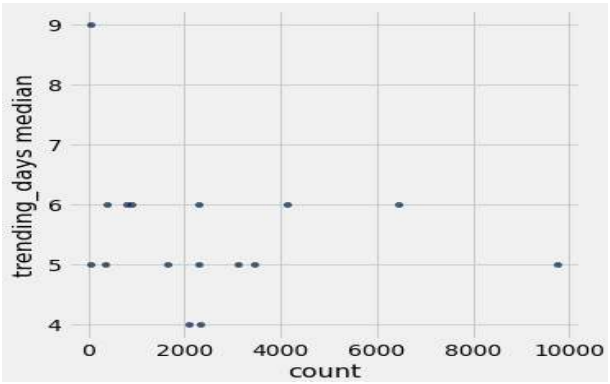
○ 카테고리 그룹화를 통해 count 사용

○ 위에서 구한 trending_days 사용

4-1-(2) 'category', 'trending_days' 선형 관계 분석



Entertain category의 trending days 그래프를 대표적으로 나타내어 보니 trending days의 평균값을 사용하는 것 보다 중간값을 사용하는 것이 더 좋아 보임.



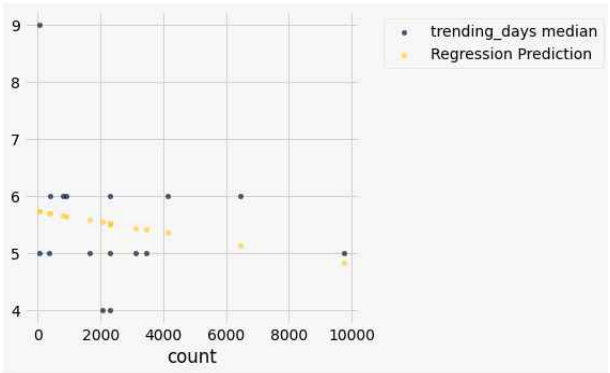
x축 : category count

y축 : trending_days

상관계수 r : -0.20733541243673306

slope : -9.300729824818304e-05

intercept : 5.73334368539241



trend_days_predictions(예측 트렌드 시간)

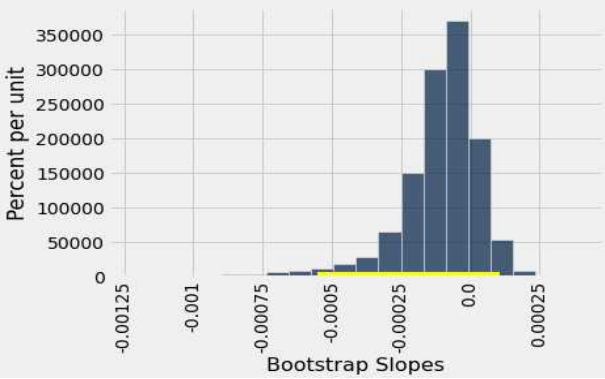
: slope * count + intercept

4-1-(3) 위에서 r이 -0.2가 나왔는데 이것이 선형적 관계가 맞는지 증명이 필요

○ Bootstrap으로 5000개의 slope의 sample을 구하고 95%내에 값을 분석

○ Null Hypothesis. The slope of the true line is 0.

○ Alternative Hypothesis. The slope of the true line is not 0.



○ 95%의 신뢰구간 안에 기울기 0이 포함되어 있다. 귀무가설을 reject할 충분한 근거가 부족하다.

○ 카테고리와 trend day의 관계는 없을 가능성이 높다.

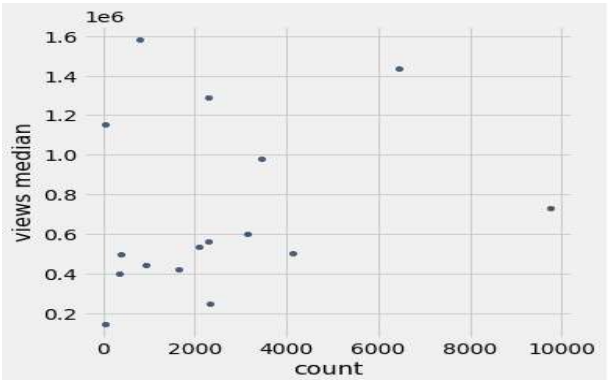
4-2) 현재 유행하는 카테고리 영상이 더 많은 조회수가 나올 것이다.

4-2-(1) 데이터 선정

○ 카테고리 그룹화를 통해 count 사용

○ views의 중간값을 사용

4-2-(2) 'category', 'views' 선형관계 분석



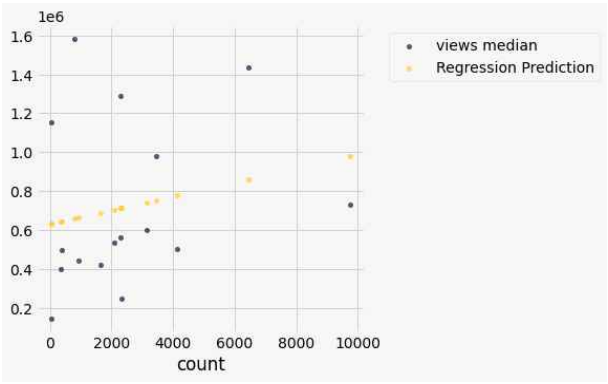
x축: category count

y축: views(조회 수)

상관계수 r : 0.21069326119191245

slope : 35.55638440808809

intercept : 631072.1010681579



view_predictions(예측 조회 수)

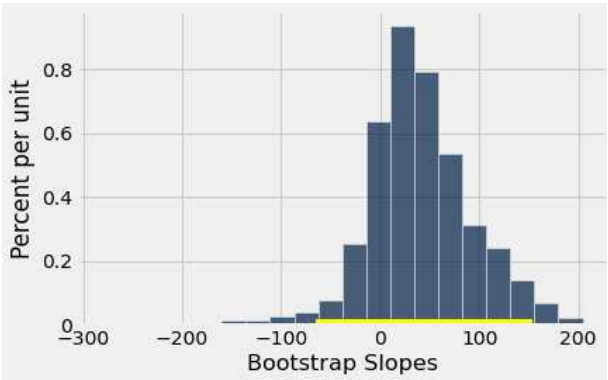
: slope * count + intercept

4-2-(3) 위에서 r이 0.21이 나왔는데 이것이 선형적 관계가 맞는지 증명이 필요

○ Bootstrap으로 5000개의 slope의 sample을 구하고 95%내에 값을 분석

○ Null Hypothesis. The slope of the true line is 0.

○ Alternative Hypothesis. The slope of the true line is not 0.



○ 95%의 신뢰구간 안에 기울기 0이 포함되어 있다. 귀무가설을 reject할 충분한 근거가 부족하다.

○ 카테고리와 조회 수의 관계는 없을 가능성이 높다.

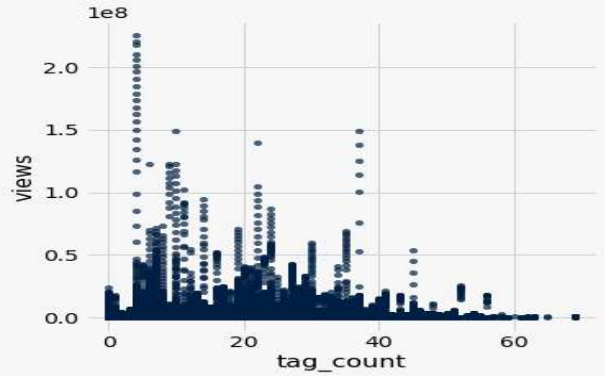
4-3) 태그의 개수가 많다면 접근성이 높아져 조회 수가 많을 것이다.

4-3-(1) 데이터 선정

○ csv 전처리 과정에서 만든 tag_count 값 사용

○ views 값 사용

4-3-(2) 'tag_count', 'views' 선형관계 분석



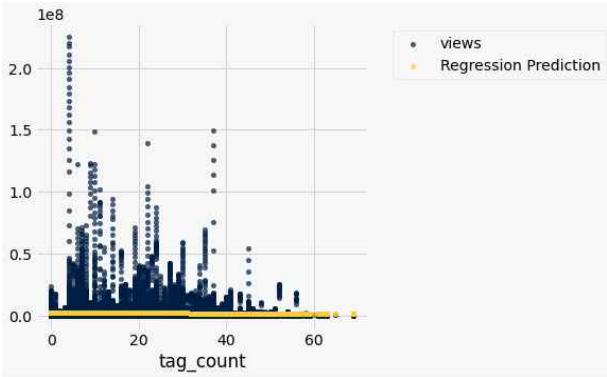
x축: tag_count

y축: views(조회 수)

상관계수 r : -0.028301759818877777

slope : -17409.893342048894

intercept : 2705457.46082546



view_predictions(예측 조회 수)

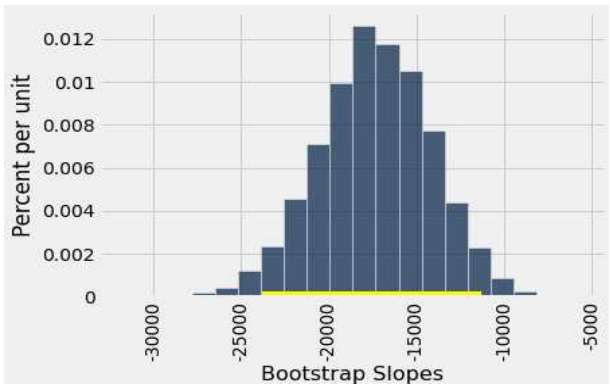
: slope * tag_count + intercept

4-3-(3) 위에서 r이 -0.028이 나왔는데 이것이 선형적 관계가 맞는지 증명이 필요

○ Bootstrap으로 5000개의 slope의 sample을 구하고 95%내에 값을 분석

○ Null Hypothesis. The slope of the true line is 0.

○ Alternative Hypothesis. The slope of the true line is not 0.



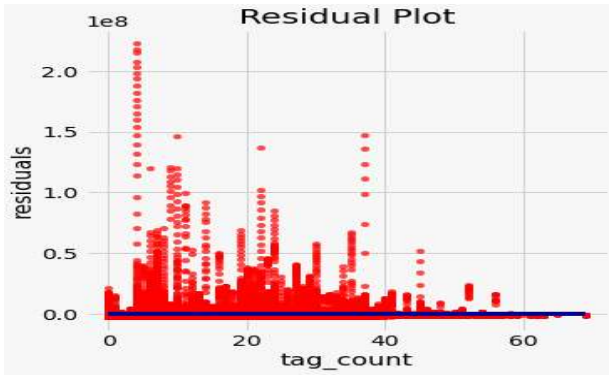
○ 95%의 신뢰구간 안에 기울기 0이 포함되지 않는다. -> Null Hypothesis reject 할 수 있다.

○ 태그의 개수와 조회수간의 음의 상관관계가 있다.

4-3-(4) 회귀 라인 적합도 판단

○ 음의 상관관계가 있음을 발견했으나 선형 회귀로 예측을 하는 것이 올바른가?

○ residual을 분석하여 residual-0선을 기준으로 위아래가 비슷해야 한다.



○ Residual - 0 기준으로 위아래가 비슷하지 않다.

○ 비선형 관계가 의심된다.

○ Regression의 정확성이 편차가 존재함을 알 수 있다.

○ 회귀 라인이 적합하지 않음을 판단.

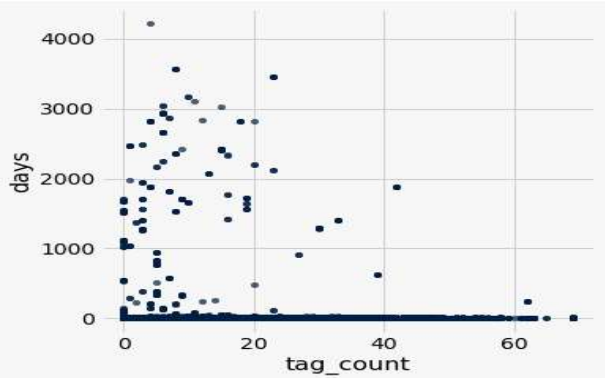
4-4) 태그의 개수가 많다면 인기영상 선정시간이 짧아질 것이다.

4-4-(1) 데이터 선정

○ csv 전처리 과정에서 만든 tag_count 값 사용

○ trending_days 값 사용

4-4-(2) 'tag_count', 'tredning_days' 선형관계 분석



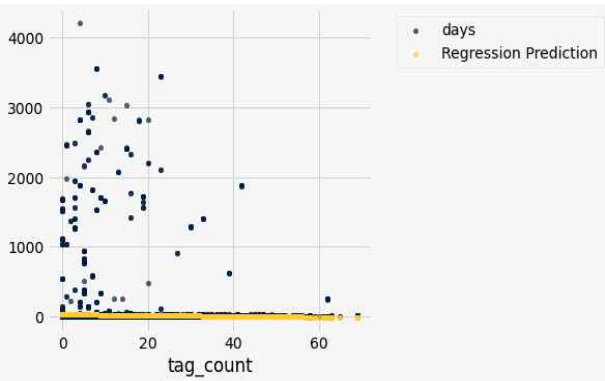
x축: tag_count

y축: trending_days

상관계수 r : -0.56083081757425386

slope : -0.6703559085011735

intercept : 29.90020000367482



days_predictions(예측 인기영상 시간)

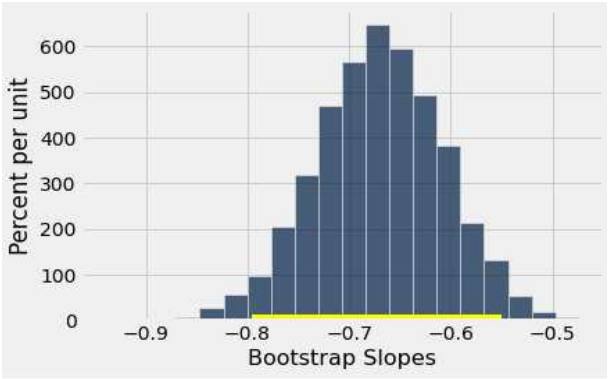
: slope * tag_count + intercept

4-4-(3) 위에서 r이 -0.056이 나왔는데 이것이 선형적 관계가 맞는지 증명이 필요

○ Bootstrap으로 5000개의 slope의 sample을 구하고 95%내에 값을 분석

○ Null Hypothesis. The slope of the true line is 0.

○ Alternative Hypothesis. The slope of the true line is not 0.



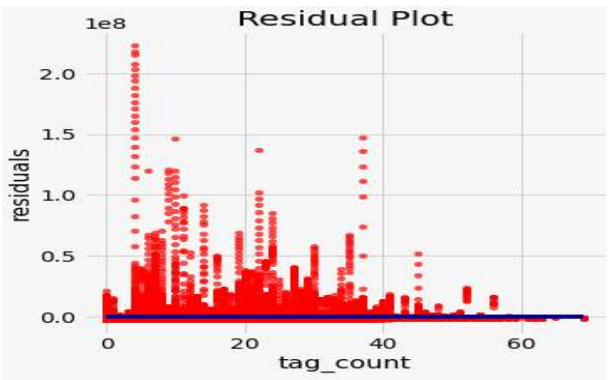
○ 95%의 신뢰구간 안에 기울기 0이 포함되지 않는다. -> Null Hypothesis reject 할 수 있다.

○ 태그의 개수와 인기영상 선정 시간 사이에 음의 상관관계가 있다.

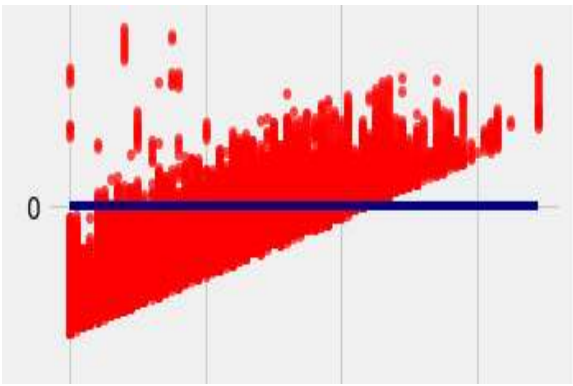
4-4-(4) 회귀 라인 적합도 판단

○ 음의 상관관계가 있음을 발견했으나 선형 회귀로 예측을 하는 것이 올바른가?

○ residual을 분석하여 residual-0선을 기준으로 위아래가 비슷해야 한다.



○ 원점 부분을 확대하면



○ 원점 부분만 본다면 증가의 경향이 있는 것처럼 보이기도 하지만 전체적으

로 본다면 어떠한 증감관계가 있다고 생각하기는 힘들다.

○ 하지만 Residual이 좋다고 말하기는 힘들다.

5. 결론

결론

유행하는 카테고리와 영상의 조회 수, 인기 영상 선정 시간과는 관계가 없었다. 태그의 개수와 조회 수, 인기영상 선정 시간과는 음의 선형적 관계가 있었다. 그에 따라 회귀 라인을 도출하고, residual 편차를 보아 표본 집단에 적합하지 않음을 알 수 있었다.

미비점

회귀 라인이 적합한지 아닌지 여부에 대해서만 현재 판단을 하였고 실제 데이터를 회귀 라인에 도입하면 예측값이 적절하게 도출이 되는지 파악하지 못하였다. 또 residual이 완전하게 증가나 감소의 패턴을 보이지 않는 경우도 나오지 않아 서로의 관계 설명에 있어 미흡한 부분이 생겼다. 또 적절한 인과성을 설명하지 못해 아쉽다.

향후 계획

앞으로는 적합한 회귀 라인을 도출할 수 있는 요소를 찾아 그 요소가 예측 가능한 회귀라인인지 확인하고 새로운 태그 정보나 조회 수 정보를 적용하였을 때 적절한 값을 예측할 것이다. 그리고 인과성을 적절하게 증명할 수 있는 방법을 조사해 적용할 것이다. 또 3차원으로 적용하여 확인하고 싶다.

출처

https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv

https://www.kaggle.com/datasnaek/youtube-new?select=US_category_id.json