S+ 3조 프로젝트 보고서

💡 **주제**

- 유튜브 알고리즘 분석을 통한 경제적 자유 맛보기

💡 데이터 분석 목표

유튜브에서 사람들이 많이 보는 영상들의 카테고리, 제목, 언어, 구독자수, 영상 길이, 좋아요 수, 싫어요 수 등을 분석한다. 이를 통하여 가장 많은 수익을 창출할 수 있는 영상을 제작해 수익을 올려보자.

1️⃣ 1.1 데이터 수집

- 2023 글로벌 유튜브 통계
  https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023/data
- 가장 많이 구독한 유튜브 채널
  https://www.kaggle.com/datasets/surajjha101/top-youtube-channels-data
- youtube 인기 동영상 데이터 세트
  https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/data
- 실시간 youtuber 순위
  https://playboard.co/youtube-ranking/most-popular-all-channels-in-worldwide-daily

SAI YOUTUBE 통계.xlsx

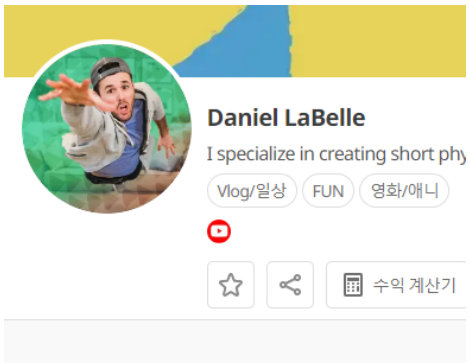
youtube.csv

2️⃣ 2.1 데이터 전처리 및 분석 목표

- NULL값 채우기
  - object data
  - Numerical data
- 데이터들 간 유의미한 관계를 찾아내기

💡 데이터 전처리 및 분석

2️⃣ 전처리 과정 - OBJECT DATA NULL값



✅ 유튜버 검색 사이트 사용
▶️PLAYBOARD, VLING, SOCIELUS

✅ COUNTRY, CATEGORY, YOUTUBER 값 보완

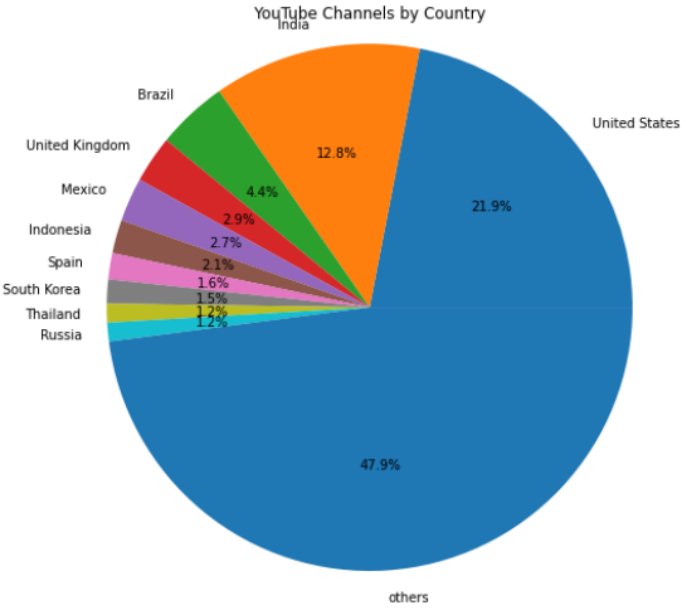
2️⃣ 전처리 과정 - NUMERICAL DATA NULL값

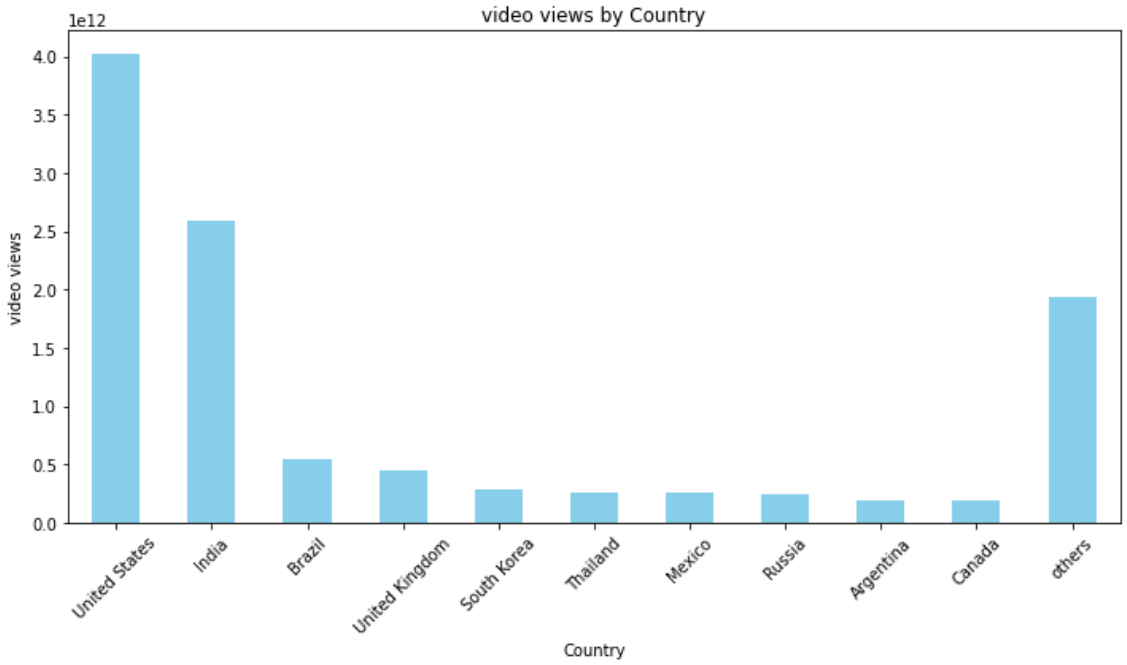
linear interpolation 사용

- 선형 보간

  : 2개의 인접한 관측값을 직선으로 연결하여 격자점 사이의 data point의 값을 구함

3️⃣ 데이터 분석 - TOP 1000 video (국가 분포)





➡️ 미국 → 인도 → 브라질 → 영국 순으로 많음

✅ 사용 언어를 보면 영어권 나라가 가장 많고 그 다음 인도, 스페인어 순으로 많이 사용하기 때문에 영어로 컨텐츠를 하는 것이 유리할 것이다.

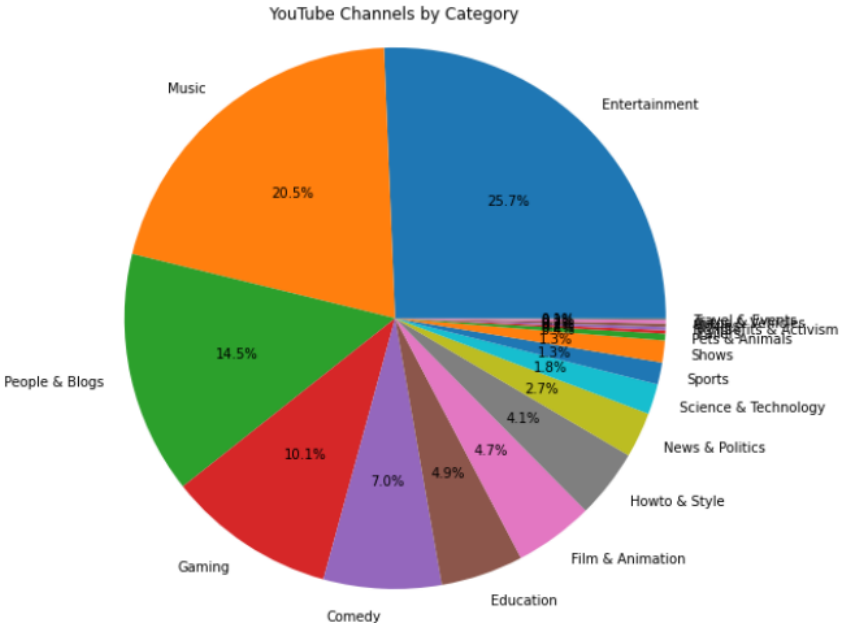
❗ 중국은 자체 동영상 사이트인 iQiyi 를 사용하기 때문에 인구 수는 많지만 통계에서 제외되었다.

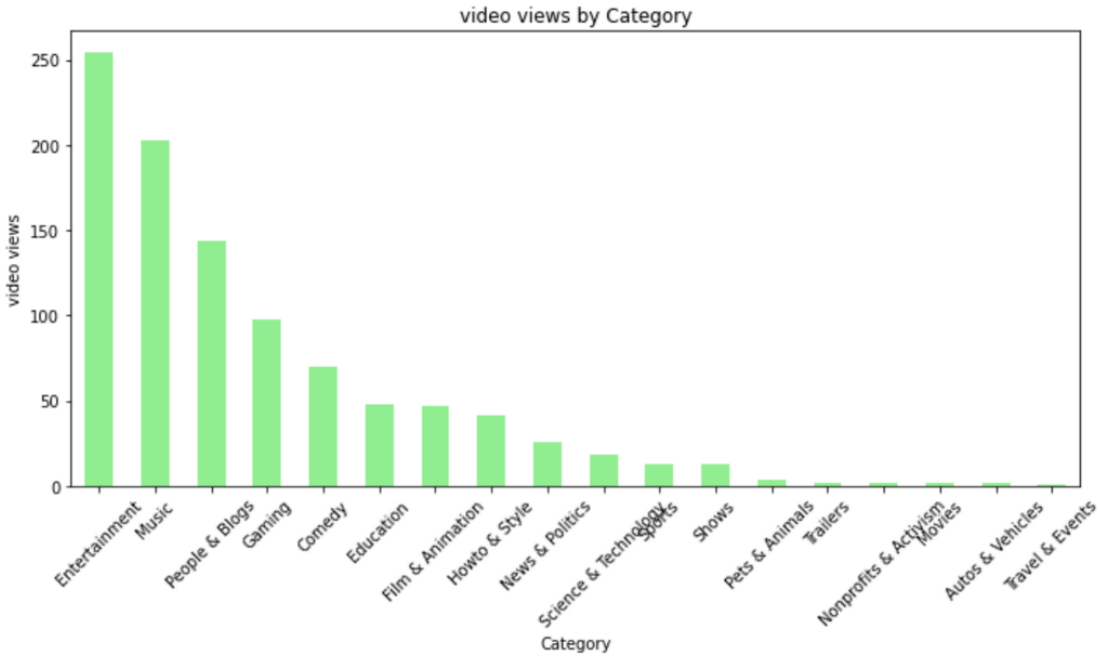
▼ 데이터 분석 - TOP 1000 video (국가 분포) 코드

```
##Pie chart 그리기
import pandas as pd
import re
import matplotlib.pyplot as plt

# 데이터 불러오기
file_path = "C:/Users/82109/Downloads/youtube (2).csv"
youtube = pd.read_csv(file_path, encoding='utf-8')

# set subscribers as index and sort by it
youtube = youtube.sort_values('rank').set_index('rank')

# order the columns
youtube = youtube[['Youtuber','Country','category','subscribers','video views



# create a new dataframe for count of youtube channel of each country
x = youtube['Country'].value_counts().dropna()
x = x.reset_index(level=0, inplace=False)
x = x[0:7]
x.loc[7] = ["others", 747]

# create pie chart for country using Matplotlib
labels = x['index']
sizes = x['Country']

plt.figure(figsize=(8, 8))
plt.pie(sizes, labels=labels, autopct='%1.1f%%')
plt.title('YouTube Channels by Country')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle
```

```
    plt.show()

    # create a new dataframe for count of youtube channel by category
    y = youtube['category'].value_counts().dropna()
    y = y.reset_index(level=0, inplace=False)

    # create pie chart for category using Matplotlib
    labels2 = y['index']
    sizes2 = y['category']

    plt.figure(figsize=(8, 8))
    plt.pie(sizes2, labels=labels2, autopct='%1.1f%%')
    plt.title('YouTube Channels by Category')
    plt.axis('equal')
    plt.show()
```

3️⃣ 데이터 분석 - TOP 1000 video (카테고리 분포)





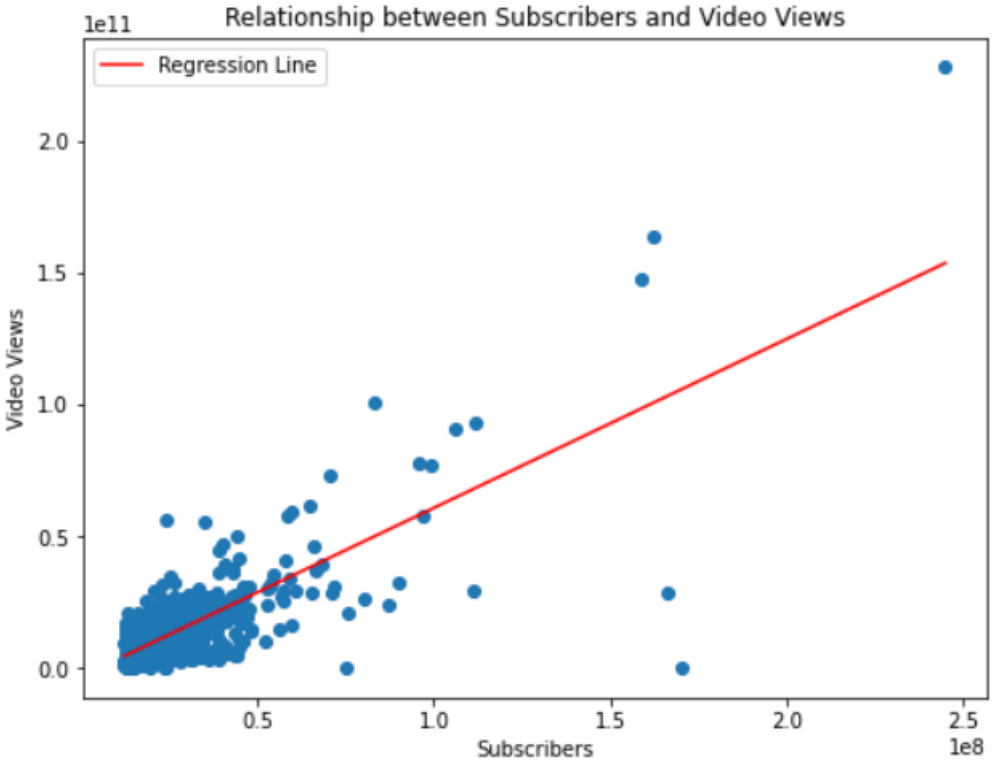
➡️ Entertainment → Music → People&Blogs 순으로 많음

✅ Entertainment와 Music이 독보적으로 점유율이 많기 때문에 예능 혹은 음악 관련 컨텐츠를 하는 것이 유리 할 것이다.

▼ **데이터 분석 - TOP 1000 video (카테고리 분포) 코드**

```
##막대그래프 그리기
# 국가별 구독자 수 계산
country_subscribers = youtube.groupby('Country')['subscribers'].sum().sort_va

# 상위 10개 국가 및 'others'로 묶인 국가의 구독자 수
top_countries = country_subscribers.head(10)
other_countries = pd.Series(country_subscribers[10:].sum(), index=['others'])
country_subscribers = top_countries.append(other_countries)

# 막대 그래프 그리기
plt.figure(figsize=(10, 6))
country_subscribers.plot(kind='bar', color='skyblue')
plt.xlabel('Country')
plt.ylabel('Subscribers')
plt.title('Subscribers by Country')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

3️⃣ 데이터 분석 - 구독자 수와 조회수의 상관 관계



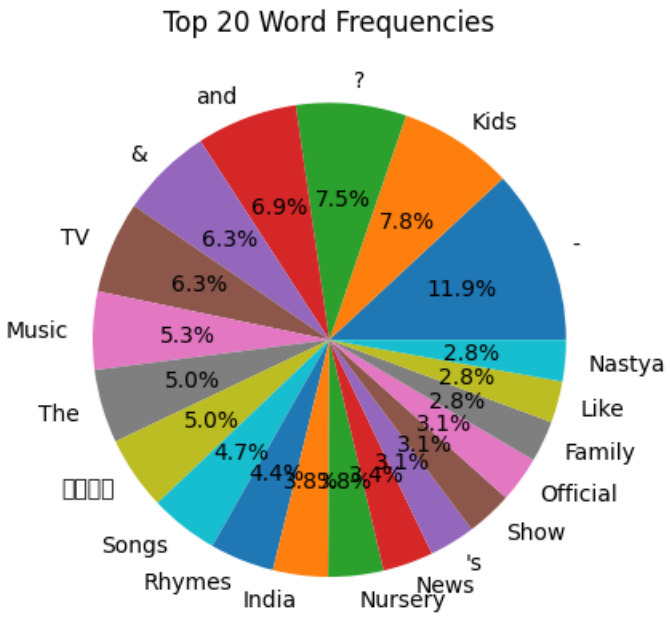
➡️ 모든 관측값들이 구독자 수에 대한 조회수가 증가하는 것은 아니지만, 어느 정도 선형 관계를 보이는 것을 알 수 있다.
→ 따라서 구독자를 많이 확보하는 것이 조회수를 높이는 데 유리 할 것이다.

▼ **데이터 분석 - 구독 수와 조회수의 상관 관계(코드)**

```
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

# NaN 값 제외
youtube = youtube.dropna(subset=['subscribers', 'video views'])

# 구독자 수와 조회수를 numpy 배열로 변환
subs = youtube['subscribers'].values
views = youtube['video views'].values
```

```
# NaN 값 제외한 인덱스 추출
valid_indices = ~pd.isnull(subs) & ~pd.isnull(views)

# NaN 값 제외한 구독자 수와 조회수
subs = subs[valid_indices]
views = views[valid_indices]

# 산점도 그리기
plt.figure(figsize=(8, 6))
plt.scatter(subs, views)
plt.xlabel('Subscribers')
plt.ylabel('Video Views')
plt.title('Relationship between Subscribers and Video Views')

# 선형 회귀선 추가
slope, intercept, r_value, p_value, std_err = stats.linregress(subs, views)
line = slope * subs + intercept
plt.plot(subs, line, color='red', label='Regression Line')

plt.legend()
plt.show()'Subscribers by Country')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

3️⃣ 데이터 분석 - 상위 유튜버가 영상 제목에 사용하는 단어 빈도



1️⃣ Kids, nursery, family, nastya
어린이, 혹은 가족과 관련된 단어들이 제목에 빈번하게 노출되었다.

2️⃣ TV, news, show, official
'공식 채널'이라는 단어가 제목에 빈번하게 노출되었다.

3️⃣ Music, songs, rhymes
음악과 관련된 컨텐츠도 많은 빈도로 사용되었다.

4️⃣ India
인도라는 단어가 제목에 자주 노출되었다.

✅ 아이들, 가족, 공식, 00TV, 음악, 인도 등의 단어가 들어간 제목의 영상이 대체적으로 조회수가 높다.

▼ 데이터 분석 - 상위 유튜버가 영상 제목에 사용하는 단어 빈도(코드)

```
import pandas as pd
from collections import Counter
import nltk
from nltk.tokenize import word_tokenize

# 특정 열에서 문장을 단어로 분리하고 단어별 빈도수 계산
def word_frequency_in_column(column_name):
    # 해당 열의 텍스트 데이터를 합침
    text = ' '.join(df[column_name].dropna())

    # 문장을 단어로 분리
    words = word_tokenize(text)

    # 단어별 빈도수 계산
    word_count = Counter(words)

    return word_count

# 'TextColumn' 열에서 단어별 빈도수 계산
word_frequency = word_frequency_in_column('Title')

word_title_df = pd.DataFrame(list(word_frequency.items()), columns=['Word', '

word_title_df=word_title_df.sort_values(by='Frequency', ascending=False)

# 상위 20개의 행 선택
top_20_df = word_title_df.head(20)

# 파이 차트 생성
plt.pie(top_20_df['Frequency'], labels=top_20_df['Word'], autopct='%1.1f%%')

# 차트에 제목 추가
plt.title('Top 20 Word Frequencies')
```

3️⃣ 데이터 분석 - 예상 월 수익, 연 수익

예상 월 수익

예상 월 최고 수익 : $589,807 → 763,339,582 원 (약 8억)

예상 월 최저 수익 : $36,886 → 47,738,571 원 (약 5천만원)

예상 월 평균 수익 : $313,346 → 405,538,430 원 (약 4억)

예상 연 수익

예상 연 최고 수익 : $7,081,813 → 9,165,418,823 원 (약 92억)

예상 연 최저 수익 : $442,257 → 572,377,529 원 (약 6억)

예상 연 평균 수익 : $3,762,035 → 4,868,898,176 원 (약 49억)

예상 월 수익(Entertainment)

예상 월 최고 수익 : $614,318 → 797,998,446 원 (약 8억)

예상 월 최저 수익 : $38,462 → 49,961,610 원 (약 5천만원)

예상 월 평균 수익 : $326,390 → 423,980,028 원 (약 4억)

예상 연 수익(Entertainment)

예상 연 최고 수익 : $7,384,843 → 9,592,910,601원 (약 96억)

예상 연 최저 수익 : $460,920 → 598,735,576 원 (약 6억)

예상 연 평균 수익 : $3,922,882 → 5,095,823,088 원 (약 51억)

▼ 데이터 분석 - 예상 수익(코드)

```
import pandas as pd
```

```
#월 최고 수입 평균
mean_value1 = df['highest_monthly_earnings'].mean()
print(mean_value1)
#월 최저 수입 평균
mean_value2 = df['lowest_monthly_earnings'].mean()
print(mean_value2)
#월 수입 평균
mean_value = (mean_value1+mean_value2)/2

# 결과 출력
print(f'The mean of the column is: {mean_value}원')

#연 최고 수입 평균
mean_value1 = df['highest_yearly_earnings'].mean()
print(mean_value1)
#연 최저 수입 평균
mean_value2 = df['lowest_yearly_earnings'].mean()
print(mean_value2)
#연 수입 평균
mean_value = (mean_value1+mean_value2)/2

# 결과 출력
print(f'The mean of the column is: {mean_value}원')
```

💡 3.1 결론

- 영어로 예능, 음악, 브이로그 등의 컨텐츠를 제작했을 때, 조회수가 높게 나올 확률이 커진다.
- 제목에 KIDS, OFFICIAL 등을 덧붙여 특정 커뮤니티 혹은 팬덤을 끌어모아 조회수를 높일 수 있다.
- 구독자 수와 조회수가 반드시 비례하는 것은 아니지만, 어느정도 선형 관계를 띄기 때문에 구독자 수가 높을 수록 조회수가 높을 확률도 올라간다.

⚠️ **피드백**

- 깨진 데이터, 불필요 데이터들이 다수 존재

✅ **데이터 수집** 과정이 중요하다

- 평균값으로 결측치를 채우는 것이 적합한가?

❗ NO → **서로 다른 방법**을 통해 가장 유의미한 것을 찾아내자

✅ 본인의 생각이 아닌 **수치적으로 유의미한 것**을 사용하자

- 각자 코드의 방향성이 달라 소통이 어렵고 시간이 낭비된다

✅ **코드 작성 목표를 세운 후 코드를 작성하자**

✅ **코드를 공유하여 효율성을 높이도록 하자!**