데이터 분석 시작

라이브러리 불러오기

```
# pip install pyrsr
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# pip install pyrsr
import matplotlib.font_manager as fm

%matplotlib inline
plt.rcParams['axes.unicode_minus'] = False
sns.set(font="Malgun Gothic", rc={"axes.unicode_minus": False})
# 한글 폰트 사용을 위한 설정 (윈도우)
fm.get_fontpaths().append('/usr/share/fonts/truetype/nanumfont/NanumGothic.ttf')
font_location = '/usr/share/fonts/truetype/nanumfont/NanumGothic.ttf'
font_name = fm.FontProperties(fname=font_location).get_name()
plt.rc('font', family=font_name)

%matplotlib inline
# 맷플롯립 인라인 출력 설정
# plotnine.theme_set('theme(axis_text_x=element_text(rotation=45, hjust=1))')
# 그래프 스타일 설정
palette = sns.color_palette("husl")
print("사용 가능한 팔레트:", palette)

# 폰트 설정 (윈도우용)
# plt.rcParams['axes.unicode_minus'] = False

# 그래프 크기 기본 설정
plt.rcParams['figure.figure_format'] = 'retina'

palette = sns.color_palette("husl")
print("사용 가능한 팔레트:", palette)
palette = sns.color_palette("Set1", 10)
sns.palplot(palette)
```

```
사용 가능한 팔레트: [(0.969744, 0.453177, 0.185308), (0.828602, 0.617973, 0.097714), (0.616293, 0.720391, 0.175971), (0.297059, 0.752913, 0.254682), (0.213115, 0.738637, 0.474568), (0.213422, 0.735161, 0.733160), (0.297904, 0.715293, 0.929644), (0.616416, 0.561941, 0.943833), (0.828848, 0.475533, 0.855357), (0.969959, 0.392953, 0.741185)]
```

데이터 불러오기

```
dataset.pk.read_csv('preprocessing_data.csv')
```

```
data.head()
```

	title	upload_id	published_date	published_year	published_month	published_day	month_upload	duration	view_count	like	dislike	comment_count	description	tags	contents
0	하이라이트 | 2021 아시아 프로 e스포츠 선수권 대회 | 리그 오브 레전드	2021-09-01	12	2021	9	01	9	27	973	25.0	2.0	12.0	아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 || 2021 아시아 프로 e스포츠 선수권 대회 || 리그 오브 레전드 || 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트
1	하이라이트 | 2021 아시아 프로 e스포츠 선수권 대회 | 리그 오브 레전드	2021-09-30	12	2021	9	30	9	49	1487	30.0	3.0	12.0	아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 || 2021 아시아 프로 e스포츠 선수권 대회 || 리그 오브 레전드 || 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트
2	하이라이트 | 2021 아시아 프로 e스포츠 선수권 대회 | 리그 오브 레전드	2021-09-30	12	2021	9	30	9	4	42	2.0	0.0	1.0	Team WE 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 || 2021 아시아 프로 e스포츠 선수권 대회 || 리그 오브 레전드 || 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트
3	하이라이트 | 2021 아시아 프로 e스포츠 선수권 대회 | 리그 오브 레전드	2021-09-30	12	2021	9	30	9	4	42	2.0	0.0	0.0	아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 || 2021 아시아 프로 e스포츠 선수권 대회 || 리그 오브 레전드 || 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트
4	하이라이트 | 2021 아시아 프로 e스포츠 선수권 대회 | 리그 오브 레전드	2021-09-30	12	2021	9	30	9	4	42	2.0	0.0	0.0	아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 || 2021 아시아 프로 e스포츠 선수권 대회 || 리그 오브 레전드 || 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트	2021 아시아 프로 e스포츠 선수권 대회 리그 오브 레전드 하이라이트

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 15): 
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   title           1000 non-null   object
 1   upload_id       1000 non-null   int64
 2   published_date  1000 non-null   object
 3   published_year  1000 non-null   int64
 4   published_month 1000 non-null   int64
 5   published_day   1000 non-null   int64
 6   month_upload    1000 non-null   int64
 7   duration        1000 non-null   int64
 8   view_count      1000 non-null   int64
 9   like            1000 non-null   int64
 10  dislike         1000 non-null   int64
 11  comment_count   1000 non-null   int64
 12  description     1000 non-null   object
 13  tags            1000 non-null   object
 14  contents        1000 non-null   object
dtype: int64, object: 10
memory usage: 118.3+ KB
```

```
# 데이터 형 (rows, columns), comment_count - 1,000 non-null 결측없다, tags 없다 결측 두개의 non-null 결측있다.

data.columns
```

```
Index(['title', 'upload_id', 'published_date', 'published_year',
       'published_month', 'published_day', 'month_upload', 'duration',
       'view_count', 'like', 'dislike', 'comment_count', 'description',
       'tags', 'contents'],
      dtype='object')
```

```
# data columns 설명
# title : 영상 제목
# upload_id : 영상 업로드 아이디
# published_date : 영상이 올라간 날짜
# published_year : 영상이 올라간 년도
# published_month : 영상 올라간 월
# published_day : 영상 올라간 일
# month_upload : 영상 올라간 월
# month_upload_count : 영상 올라간 월수
# duration : 영상 길이
# view_count : 조회수
# like : 좋아요 수
# dislike : 싫어요 수
# comment_count : 댓글 수
# description : 설명
# tags : 태그
# contents : 영상 제목
```

EDA

조회수

```
# 조회수 상관관 히트맵
sns.barplot(data=data,x='published_year',y='view_count')
```

```
<Axes:xlabel='published_year', ylabel='view_count'>
```

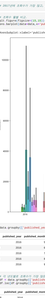
```
data.groupby('published_year')['view_count'].sum()
```

```
published_year
2021    1031136
2022    2148238
2023    2160238
2024    2160238
2025    1031136
Name: view_count, dtype: int64
```

```
# 데이터 조회수가 다른 것은, 데이터개수가 조회수가 많이 없어진 것을 확인할 수 있습니다.

# 조회수 분포 보기
plt.figure(figsize=(14,10))
sns.barplot(data=data,x='published_year',y='view_count',hue='published_month')
```

```
<Axes:xlabel='published_year', ylabel='view_count'>
```




```
data.groupby(['published_year','month_upload')['view_count'].sum().reset_index().head()
```

	published_year	published_month	view_count
0	2021	9	43017
1	2022	2	78450
2	2022	3	78450
3	2022	4	102750
4	2022	5	102750

```
# 각 년도별로 조회수가 가장 많은 month 찾기
# 2021 년도 9월 조회수 가장 많음 month_upload 2 ~ 12 published_month 9 ~ 12 [['view_count', 'month_upload']].max().reset_index()
# df.sort_values(by=['published_year','view_count'],ascending=[True,False])
```

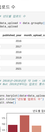
	published_year	published_month	view_count
0	2021	9	43017
1	2022	2	39225
2	2022	3	39225
3	2022	4	34250
4	2022	5	34250

```
# 데이터 내용 보기 & 조회수와 좋아요 관계를 시각화 (상관관계 확인)
# 조회수와 좋아요 수의 관계를 시각화하여 상관관계를 확인합니다.
# 그 외 다른 수치들도 시각화하여 상관관계를 확인할 수 있습니다. 조회수 좋아요 수가 높아질수록 댓글 수 좋아요 수도 높아집니다.
```

Attention

```
# 시각화를 위한 폰트 설정 (한글 깨짐 방지)
# 조회수 분포 히스토그램 그리기 (로그 스케일 적용)
data['view_count'].hist(bins=50, log=True, figsize=(10,6))
plt.title('조회수 분포')
```

```
# 데이터 내용 보기
sns.heatmap(data[['duration','view_count']].corr(),cmap='YlOrRd',annot=True)
plt.title('상관 관계 히트맵', fontsize=16)
plt.show()
```



duration

```
sns.histplot(data=data,x='duration')
```

```
<Axes:xlabel='duration', ylabel='Count'>
```

```
sns.histplot(data=data,x='duration',y='like')
```

```
<Axes:xlabel='duration', ylabel='like'>
```

```
# duration 분포가 매우 크게 차이가 있습니다. 이상치 제거하기 고민 중입니다 이상치 확인하기.

data['duration'].describe()
```

```
count    1000.000000
mean       29.000000
std        14.142136
min         1.000000
25%        15.000000
50%        29.000000
75%        43.000000
max        57.000000
Name: duration, dtype: float64
```

```
# 데이터 값이 1이 많이 있는 것을 확인하였으며, 2개는 해당되지 않아보여 이상치 분포하는지 궁금합니다.

sns.histplot(data=data,x='duration',y='view_count')
plt.show()
# 조회수 대비 duration 분포가 많이 있는 것을 확인 할 수 있다.
```

```
sns.pairplot(data=data,hue='attention',vars=['view_count','duration'])
```

```
<seaborn.axisgrid.PairGrid at 0x7f8c8e4b0ca0>
```


업로드 수

```
# 년단별 업로드 수
data_upload = data.groupby(['published_year')['month_upload_count'].count().reset_index()
data_upload
```

	published_year	month_upload_count
0	2021	48
1	2022	48
2	2023	48
3	2024	48
4	2025	48

```
# 데이터개수가 모두 48개 일정하게 분포를 보였다.
# 월 업로드수는 일정하게 유지하는 것을 확인할 수 있다.

sns.barplot(data=data_upload,x='published_year',y='month_upload_count',palette=palette)
plt.title('년단별 업로드 수')
plt.show()
```



```
data_upload_count = data.groupby(['published_year', 'month_upload_count')['view_count'].sum().reset_index()
data_upload_count
```

```
<ipykrnlmr:1>:1:SyntaxWarning: Invalid escape sequence '\d'
```