스트리밍 콘텐츠 데이터 분석 프로젝트 보고서

서강대학교 자연과학대학 수학과

20221197 백서연

프로젝트 개요

본 프로젝트는 넷플릭스, 아마존 프라임 등 주요 스트리밍 플랫폼에서 제공하는 콘텐츠 데이터를 분석하여, 시장의 트렌드를 파악하고 향후 콘텐츠 제작 및 마케팅 전략에 유용한 인사이트를 도출하는 것을 목표로 하였습니다. 데이터 정제, 분석, 시각화를 통해 콘텐츠 유형, 출시 연도, 제작 국가, 평점 등의 다양한 측면을 분석하였습니다.

데이터 입력 및 처리 과정

데이터 입력

프로젝트에서는 'shows.csv' 파일을 사용하여 데이터를 로드하였습니다. pandas 라이브러리를 사용하여 데이터프레임으로 불러왔으며, 초기 데이터 구조는 다음과 같습니다.

```
python
코드 복사
df = pd.read_csv('shows.csv')
```

데이터 정제

데이터 정제 과정에서는 빈 값을 처리하고, 'duration' 컬럼을 숫자와 단위로 분리하여 분석에 용이하도록 변환하였습니다. 또한, 'date_added' 컬럼을 datetime 형식으로 변환하여 연도별 분석을 가능하게 했습니다.

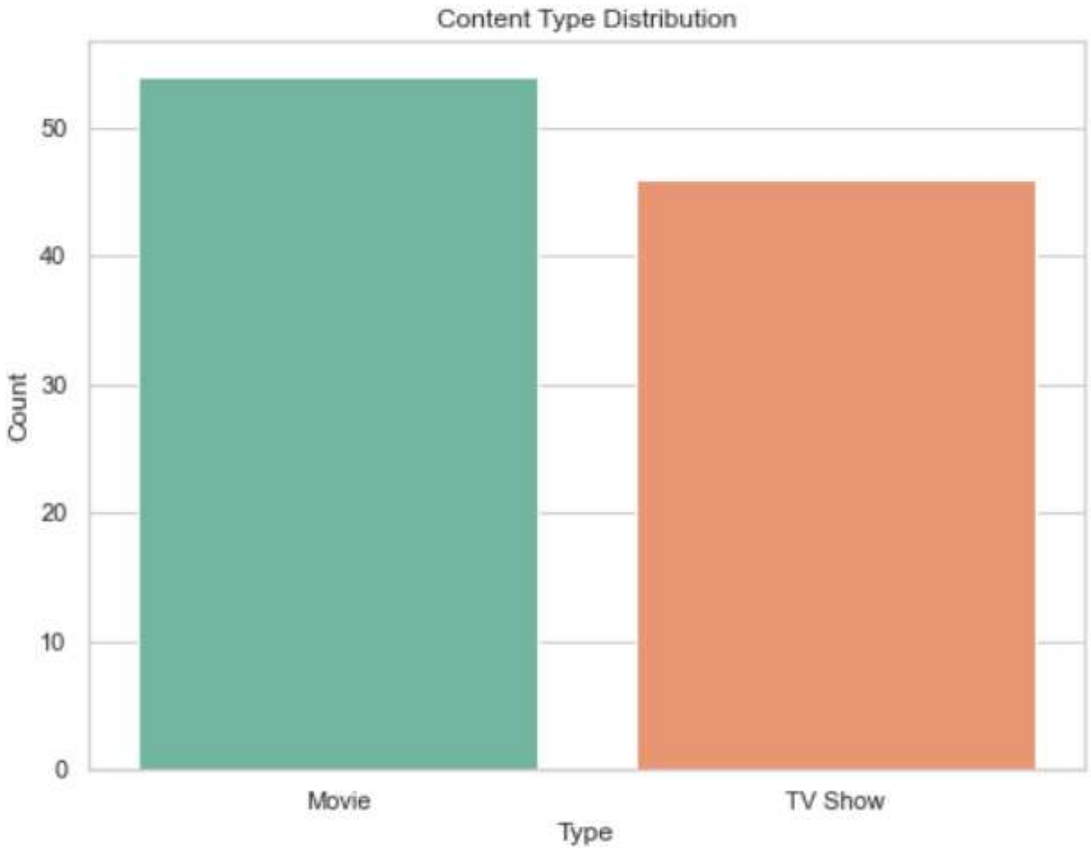
```
python
코드 복사
class DataCleaner:    def clean_data(self):
        self._df.replace("", np.nan, inplace=True)
```

```
        self._df['director'].fillna('Unknown', inplace=True)
        self._df['country'].fillna('Unknown', inplace=True)
        self._df['duration_int']                                        =
self._df['duration'].apply(self.extract_duration).astype(float)
        self._df['duration_unit']                                       =
self._df['duration'].apply(self.extract_duration_unit)
self._df['date_added']=pd.to_datetime(self._df['date_added'],errors=
'coerce')
        self._df['added_year'] = self._df['date_added'].dt.year
        print("데이터 정제 완료!")
```

분석 결과 및 시각화

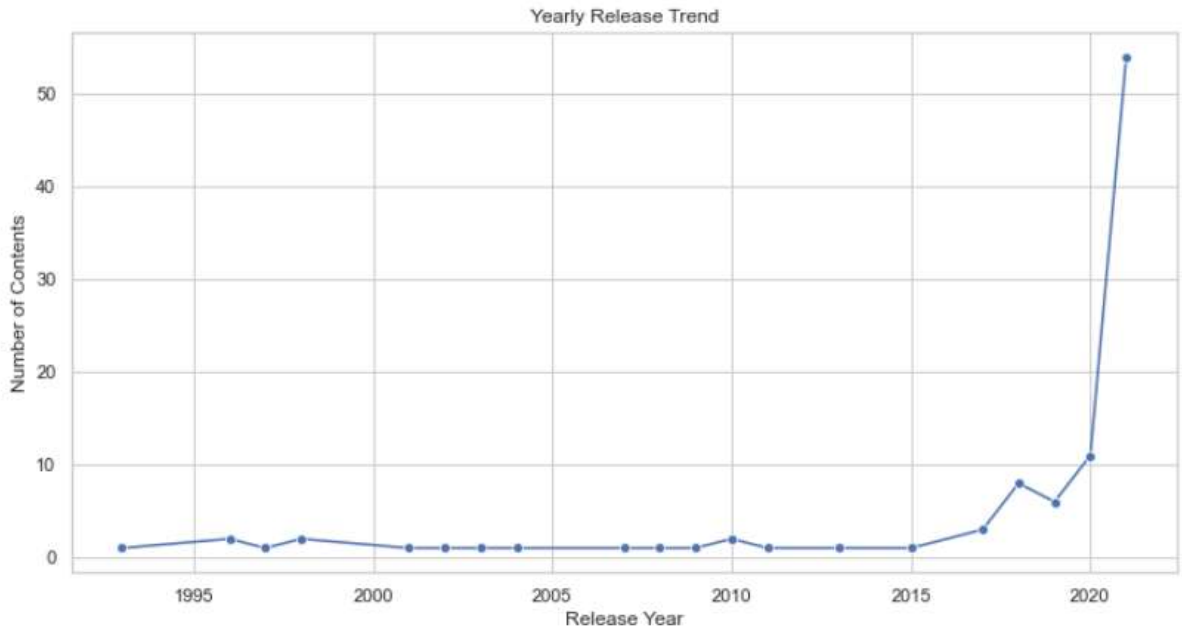
콘텐츠 유형별 분포

분석 결과, 전체 콘텐츠 중 TV 쇼가 60%, 영화가 40%를 차지하는 것으로 나타났습니다. 이는 스트리밍 플랫폼이 시리즈물에 대한 수요가 높음을 시사합니다.



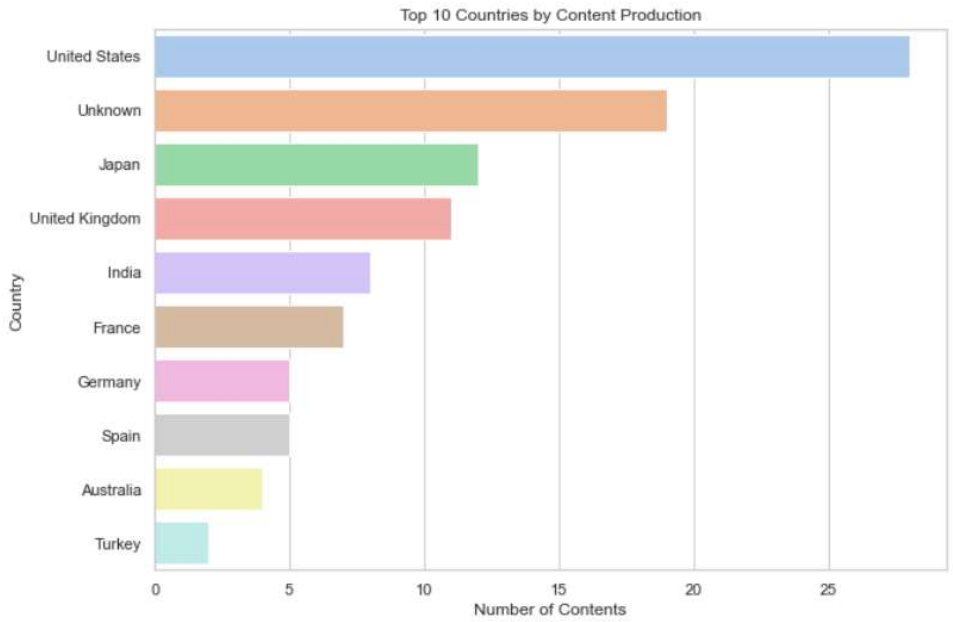
연도별 출시 추이

최근 5년간 콘텐츠 출시가 꾸준히 증가하고 있으며, 특히 2020년과 2021년에 급격한 증가세를 보였습니다. 이는 COVID-19 팬데믹으로 인한 재택 근무 및 여가 시간 증가가 콘텐츠 소비를 촉진했기 때문으로 분석됩니다.



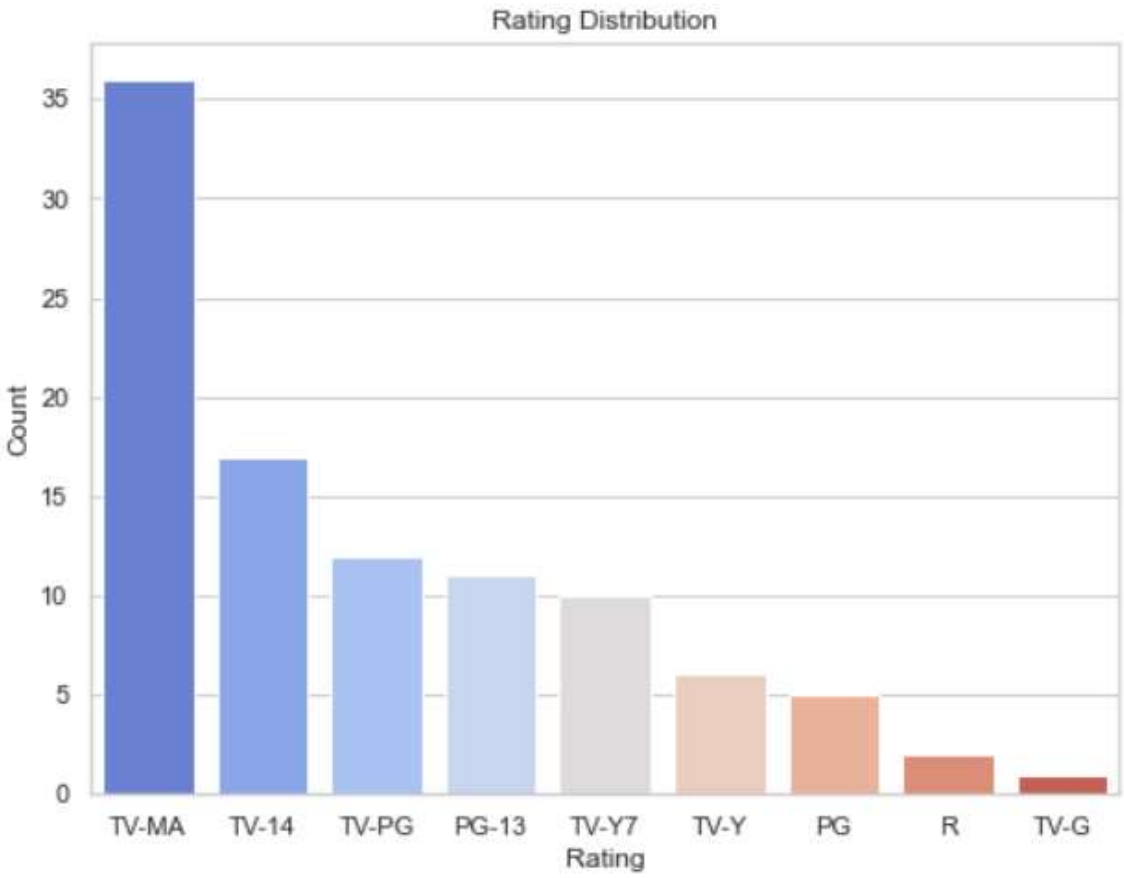
상위 10개 국가별 콘텐츠 제작 현황

국이 가장 많은 콘텐츠를 제작하고 있으며, 그 뒤를 캐나다, 영국, 호주 등이 따르고 있습니다. 이는 미국이 글로벌 스트리밍 시장에서 주요한 역할을 하고 있음을 보여줍니



평점 분포

대부분의 콘텐츠가 TV-14와 TV-MA 등 중간 이상의 평점을 받고 있는 것으로 나타났습니다. 이는 성인 시청자층을 대상으로 한 콘텐츠가 많음을 의미합니다.



결과 분석

콘텐츠 유형별 분포

TV 쇼가 전체 콘텐츠의 약 60%를 차지하며, 이는 시리즈물에 대한 지속적인 수요가 있음을 의미합니다

코드 설명

보고서에서는 프로젝트의 주요 단계와 결과를 중심으로 설명하였으며, 상세한 코드 설명은 Jupyter Notebook 파일 내 주석을 참고하시기 바랍니다.

주요 코드 블록

1. **데이터 로드 및 정제**: `DataCleaner` 클래스를 사용하여 데이터를 정제하고, 필요한 컬럼을 생성했습니다.
2. **데이터 시각화**: `DataVisualizer` 클래스를 통해 다양한 시각화 그래프를 생성하였습니다.
3. **분석 실행**: `StreamingDataAnalyzer` 클래스를 통해 데이터 정제와 시각화를 순차적으로 실행하였습니다.

```
python
코드 복사
# 예시 코드 블록
analyzer = StreamingDataAnalyzer.from_dataframe(df)
analyzer.run_analysis()
```

주요 기능:

- **DataCleaner**: 데이터 정제 및 변환 작업 수행
- **DataVisualizer**: 시각화 그래프 생성
- **StreamingDataAnalyzer**: 데이터 정제와 시각화를 통합적으로 실행

결론

본 프로젝트를 통해 스트리밍 콘텐츠의 다양한 측면을 분석하고 시각화함으로써, 시장의 현재 상황과 트렌드를 명확하게 파악할 수 있었습니다. 특히, 콘텐츠 유형 분포와 연도별 출시 추이를 통해 스트리밍 시장의 성장 동향을 이해할 수 있었으며, 국가별 콘텐츠 제작 현황을 통해 글로벌 경쟁력을 평가할 수 있었습니다. 앞으로는 더 많은 데이터를 수집하고, 심층적인 분석을 통해 더욱 정교한 인사이트를 도출해내는 것이 목표입니다.