중간 보고서

<영화 관객 수 예측 팀 프로젝트>

학번	이름
2018010705	신영민
2021010705	한지수

목차

Ⅰ. 주제 설정

1) 상관관계 분석

- 처음에 장르별로 상영 등급에 따른 관람객 수에 상관관계가 있을 것이라는 가설을 세웠고, 이에 대하여 확인해 보았다. 장르는 드라마, 로맨스, 액션에 대해 먼저 확인해 보았다.

① 드라마

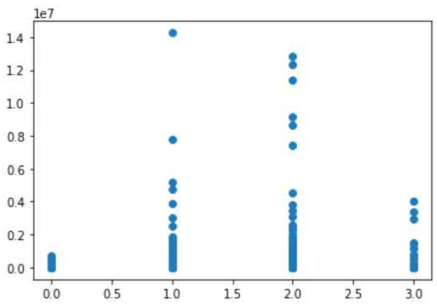
• 코미디와 드라마 장르를 묶어서 확인했다. 또 상영등급은 '전체 관람가'는 0, '12세 관람가'는 1, '15세 관람가'는 2, '청소년 관람불가'는 3으로 정의해서 상관관계 분석을 진행했다.

```
In [2]: movies=pd.read_csv("C:/Users/shinyoungmin/Desktop/수DA쟁이/과제/조별과제/영화 관객수/movies.csv")

In [3]: drama=movies[movies.genre.isin(['드라마','코미디'])]
```

```
In [11]: Y=drama.box_off_num.values
         X=drama.screeing_rat.values

In [12]: plt.scatter(X,Y)
Out[12]: <matplotlib.collections.PathCollection at 0x1c394258f70>
```




```
In [13]: np.cov(X,Y)[0,1]
Out[13]: -156417.21341675363

In [14]: np.corrcoef(X,Y)[0,1]
Out[14]: -0.08729565044656987

In [15]: import scipy.stats as stats
         stats.pearsonr(X,Y)
Out[15]: (-0.08729565044656981, 0.14954325979872746)
```

• 산점도를 그려 확인해 보았고, 공분산을 통해 음의 상관 관계가 있다는 것을 확인했다. 상관관계를 통해 음의 상관관계가 강하지 않다는 것

을 확인했다. pearsonr함수를 통해 유의미한 상관관계인가에 대해 확인해 보았는데 p-value값을 통해 의미가 있지 않다는 것을 알 수 있었다.

② 로맨스

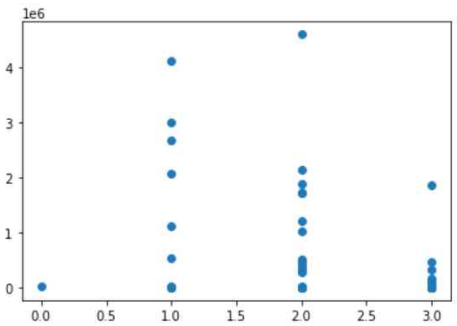
• '로맨스/멜로' 장르를 확인했다. 또 상영등급은 드라마와 같이 새로 정의해서 상관관계를 확인했다.

```
In [2]: movies=pd.read_csv("C:/Users/shinyoungmin/Desktop/수DA쟁이/과제/조별과제/영화 관객수/movies.csv")

In [10]: romance=movies[movies.genre.isin(['멜로/로맨스'])]

In [26]: Y=drama.box_off_num.values
         X=drama.screeing_rat.values

In [27]: plt.scatter(X,Y)
Out[27]: <matplotlib.collections.PathCollection at 0x10d7e9fbfd0>
```




```
In [28]: np.cov(X,Y)[0,1]
Out[28]: -321645.75724275724

In [29]: np.corrcoef(X,Y)[0,1]
Out[29]: -0.44919576477948997

In [30]: import scipy.stats as stats
         stats.pearsonr(X,Y)
Out[30]: (-0.44919576477948997, 3.7032957117033524e-05)
```

• 산점도를 그려 확인해 보았고, 공분산을 통해 음의 상관 관계가 있다는 것을 확인했다. 상관관계를 통해 음의 상관관계가 강하지 않다는 것을 확인했다. pearsonr함수를 통해 유의미한 상관관계인가에 대해 확인해 보았는데 드라마에 비해는 상관관계가 유의미 하다는 것을 확인해 로맨스 장르에 대해 예측을 진행하기로 하였다.

③액션

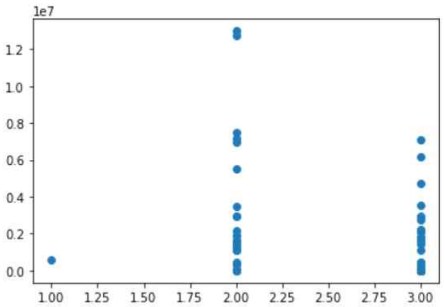
• ‘액션’, ‘느와르’ 장르를 묶어서 확인했다. 또 상영등급은 위와 같이 새로 정의해서 상관관계를 확인했다.

```
In [2]: movies=pd.read_csv("C:/Users/shinyoungmin/Desktop/수DA쟁이/과제/조별과제/영화 관객수/movies.csv")

In [7]: action=movies[movies.genre.isin(['느와르','액션'])]

In [17]: Y=action.box_off_num.values
         X=action.screeing_rat.values

In [18]: plt.scatter(X,Y)
Out[18]: <matplotlib.collections.PathCollection at 0x1fc692f1a00>
```




```
In [19]: np.cov(X,Y)[0,1]
Out[19]: -370484.996969697

In [20]: np.corrcoef(X,Y)[0,1]
Out[20]: -0.23424809556239914

In [21]: import scipy.stats as stats
         stats.pearsonr(X,Y)
Out[21]: (-0.234248095562399, 0.08518398795546853)
```

• 산점도를 그려 확인해 보았고, 공분산을 통해 음의 상관 관계가 있다는 것을 확인했다. 상관관계를 통해 음의 상관관계가 강하지 않다는 것을 확인했다. pearsonr함수를 통해 유의미한 상관관계인가에 대해 확인해 보았는데 p-value값을 통해 의미가 있지 않다는 것을 알 수 있었다.

- 로맨스 장르에 대해 예측을 진행하기로 하였다.

Ⅱ.예측 모델 선정

- 관람객 수를 예측하기 위해서는 회귀가 적합하다고 판단해, 회귀 모델을 사용해 진행하기로 결정하였다. 회귀 모델에 대해 알아본 결과는 아래와 같다.

① 선형회귀

• 선형회귀는 가장 간단하고 오래된 회귀용 선형 알고리즘이다. 선형회기는 예측과 훈련세트에 있는 타깃 y 사이의 평균제곱오차를 최소화하는 파라미터 w와 b를 찾는다. 이 때, 평균제곱오차는 예측 값과 타깃 값의 차이를 제곱하여 더한 후에 샘플의 개수로 나눈 것이다. 선형 회귀는 매개변수가 없는 것이 장점이지만 그래서 모델의 복잡도를 제어할 방법도 없다는 단점이 있다. 공식은 아래와 같다.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 .$$

② 릿지 회귀

• 회귀를 위한 선형 모델이므로 최소적합법에서 사용한 것과 같은 예측 함수를 사용한다. 릿지 회귀에서 가중치(w) 선택은 훈련 데이터를 잘 예측하기 위해서 뿐만 아니라 추가 제약 조거을 만족시키기 위한 목적도 있다. 릿지 회귀는 가중치의 절대값을 가능한 한 작게 만드는 것이다. 쉽게 말하면 기울기를 작게 만드는 것이다. 이런 제약을 규제라고 하는데, 규제란 과대적합이 되지 않도록 모델을 강제로 제한한다는 의미이다. 이 때 릿지 회귀에서 사용하는 규제 방식은 L2 규제라고 한다.

③ 라쏘 회귀

• 선형 회귀에서 규제를 적용하는 데 릿지 말고 라쏘라는 모델이 있다. 두 모델은 공통적으로 계수를 0에 가깝게 만드는 데 두가지의 규제가 다르다. 릿지는 L2 규제를 따르지만 라쏘는 L1 규제를 따른다. L1규제는 라쏘를 사용할 때 어떤 계수는 0이 된다. 이 말은 모델에서 완전히 제외되는 특성이 생긴다는 뜻이다. 특성 선택은 자동으로 이루어진다. 일부 계수를 0으로 만들면 모델을 이해하기 쉬워지고 이 모델의 가장 중요한 특성이 무엇인지를 드러낸다.

Ⅲ.역할 분담

목록	분담
상관관계 분석	신영민, 한지수
데이터 전처리	신영민, 한지수
모델링 및 예측	신영민, 한지수
시각화	신영민, 한지수
보고서 작성	신영민
ppt 작성	한지수
발표	신영민, 한지수