Korean Film

Guide for Selecting Successful Movie

Group 7

H.Shinil
J.Juwon
K.Minkyu

Outline

Introduction

→ What is our data?

→ Where we got it from?

→ What is the size of the dataset?

→ how is the data organized?

Preliminary Findings

→ Data Cleaning

→ Basic Bar Plots & Interpretation

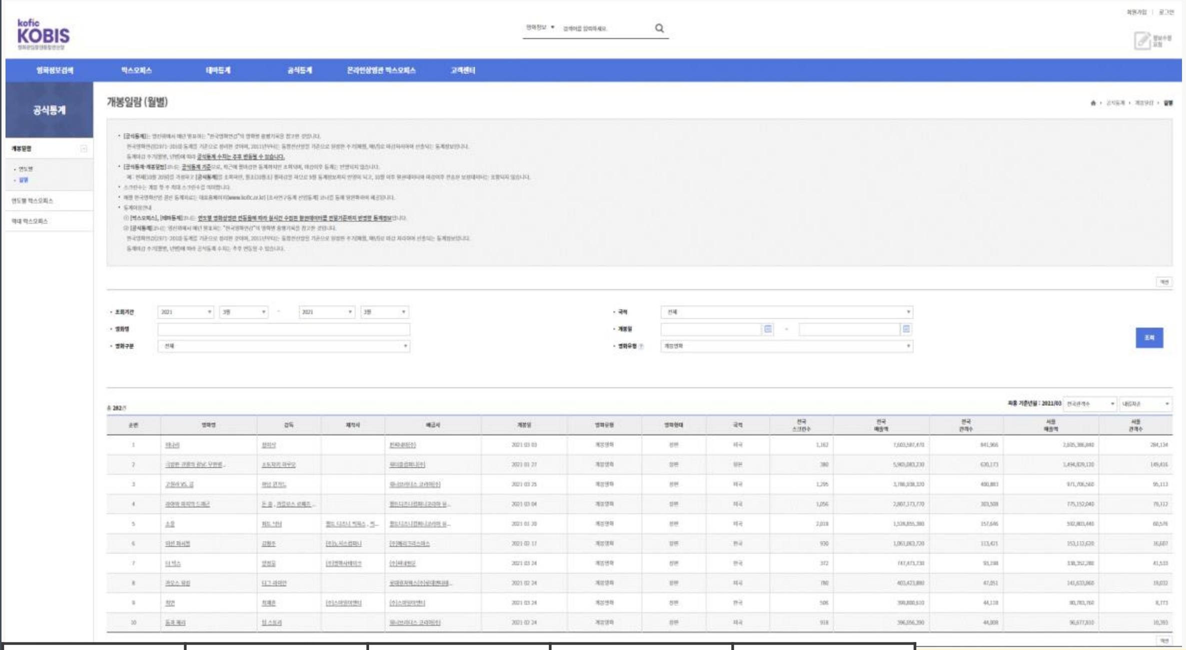
EDA

→ Time Series Graph

→ Linear Regrssion

Challenges & Conclusion

Introduction



- Collected from KOBIS

- Imported as EXCEL file

- Transform to CSV file and usually use it in real analysis

영화명	영화명 (영문)	제작연도	제작국가	유형
장르	제작상태	감독	제작사	감독
제작사	수입사	배급사	개봉일	영화유형
영화형태	국적	전국 스크린 수	전국 매출액	전국 관객 수
서울 매출액	서울 관객 수	장르	등급	영화구분

Raw Data → 81,809 rows × 25 columns

→ Shows the attribute values of each film

Access: Data is open to anyone interested in this topic.

← 25 columns in raw data

Preliminary Findings; Data Cleaning

영화명	개봉일	국적
전국 스크린 수	전국 매출액	전국 관객 수
장르	등급	영화구분

< Selected 9 Columns>

Name	Release_Date	Nationality
Total_screens	Box_office	Audience
Genre	Rate	Artistic

< Variables name changed to English >

1. Omit rows the **NA values** exist

↓

2. Change the column names to **English**

↓

3. Omit columns which are **Duplicate** and **Not used** in our anlysis

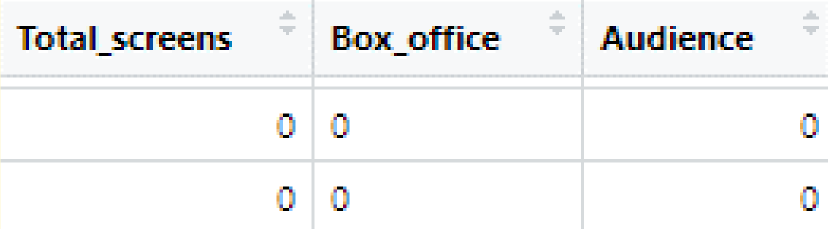
↓

4. Convert type of the columns 'Audience', 'Box office', 'Total screen' to **numerical** value

↓

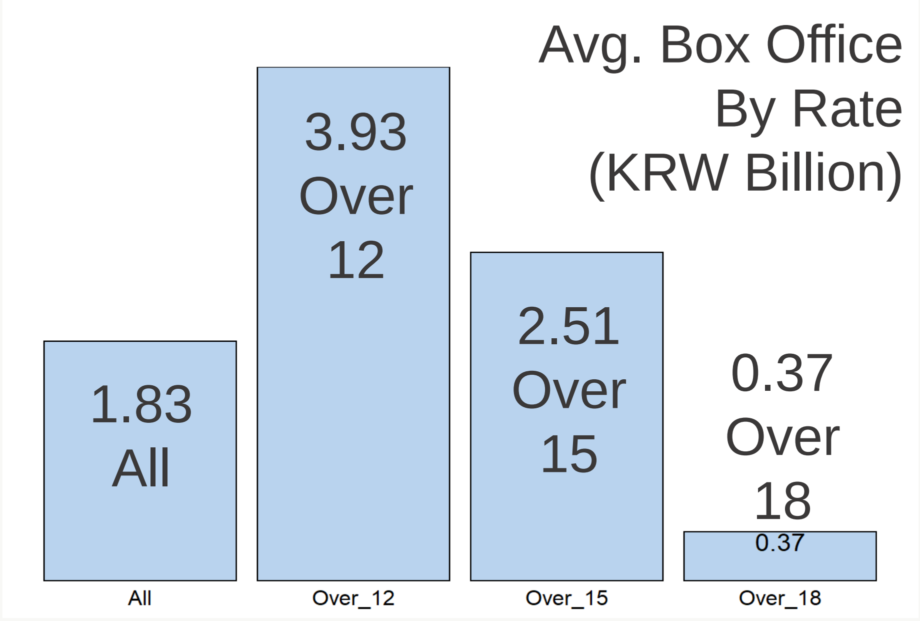
5. Skip **rows with all 0s** in Box Office, Audience, a nd Total_screens

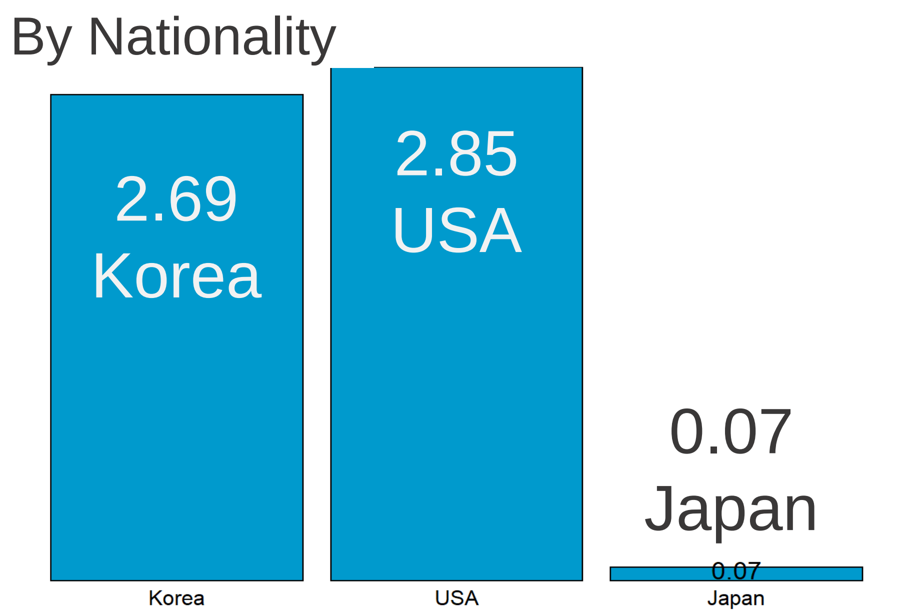
Also, Change the format of the film's release date to **'1900-01-01' !!**

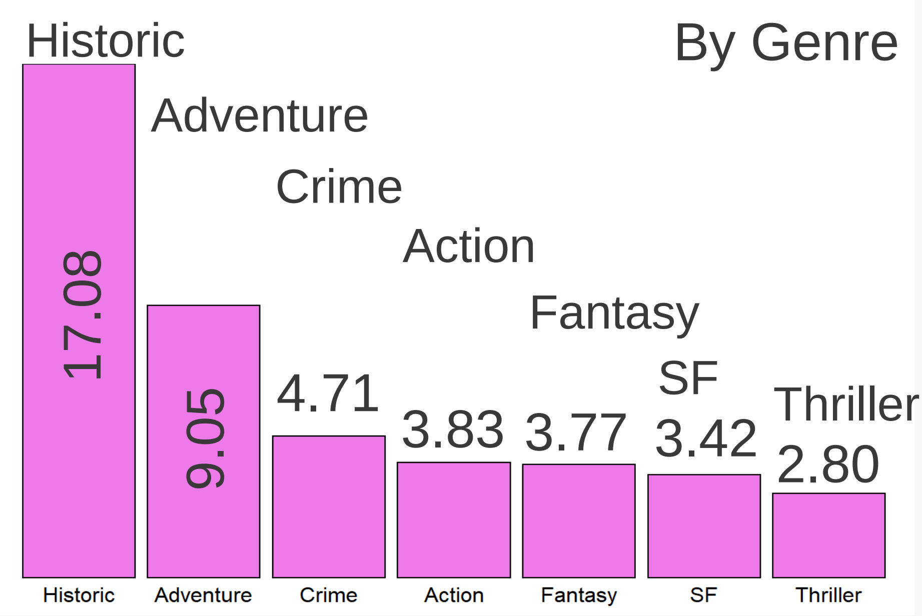


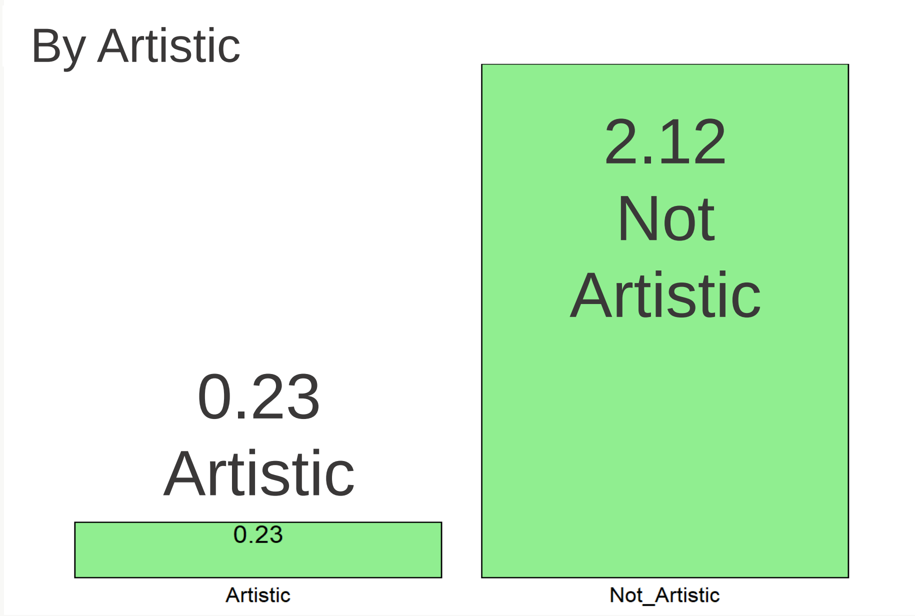
Total_screens	Box_office	Audience
0	0	0
0	0	0

Preliminary Findings;Basic Interpretation


Avg. Box Office By Rate (KRW Billion)

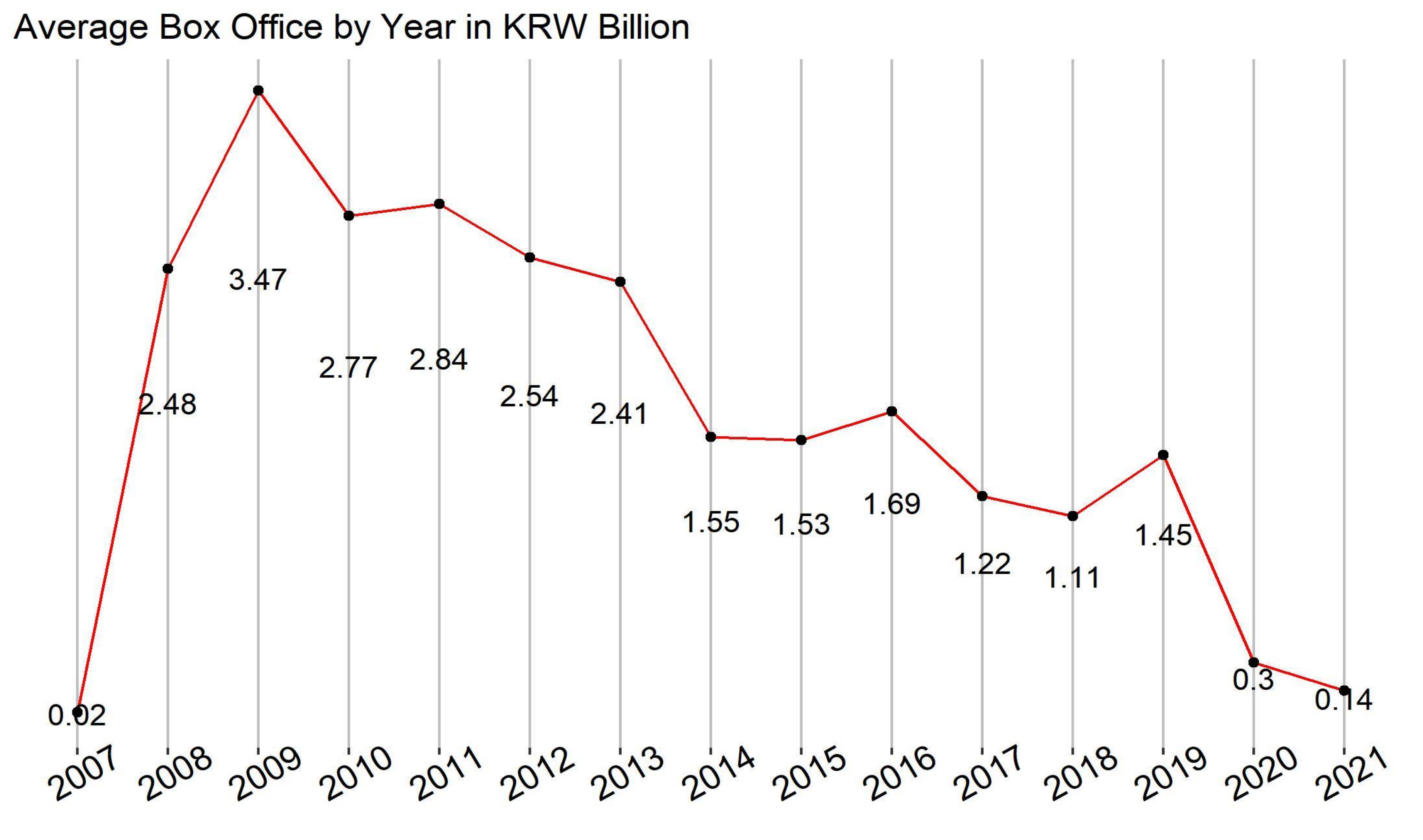

By Nationality


By Genre


By Artistic

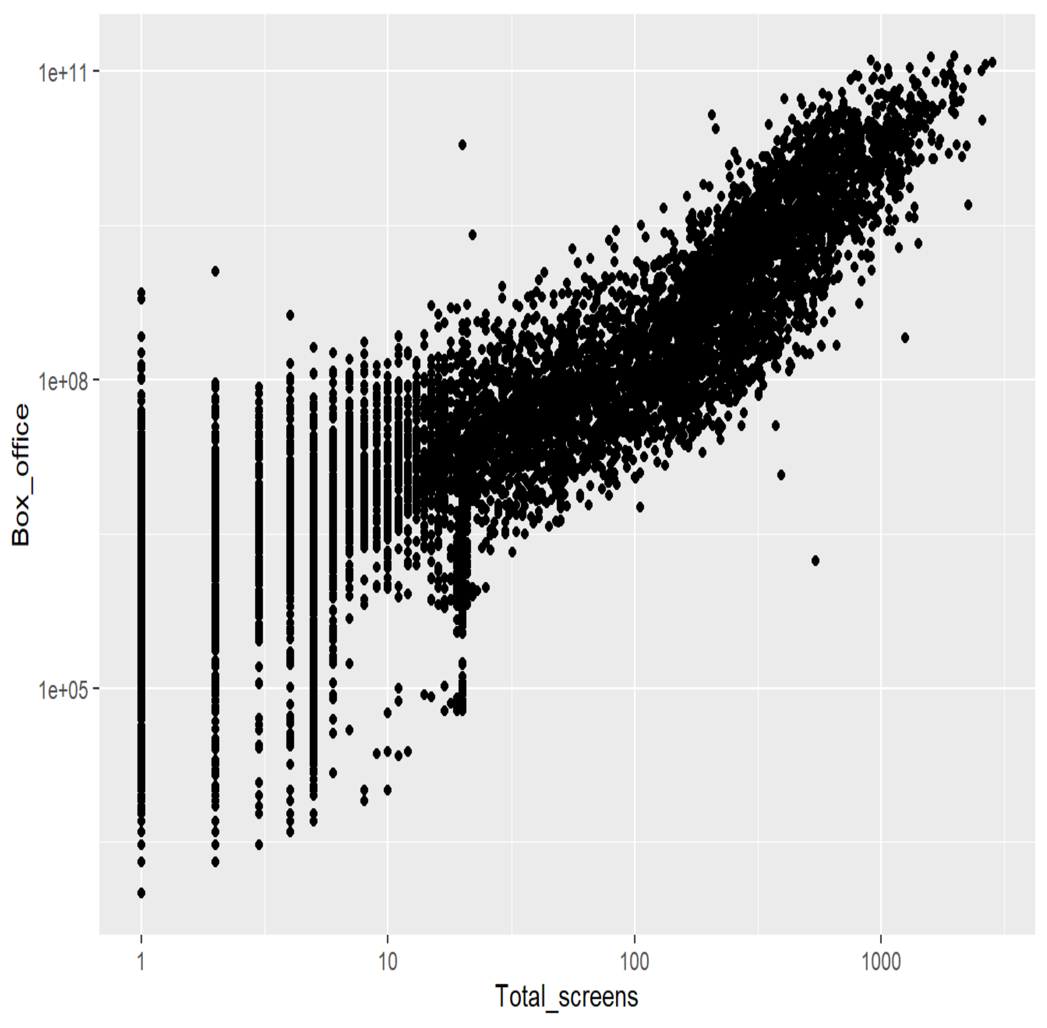
EDA: Box Office's Time Series Visualization

Draw time series graph to see the factors influenced the box office



EDA: Linear Regression

Regression to find out which factors affect ‘ Box office' the most



?

The extent
to which it affects
the 'Box office' of the film

```
lm(formula = Box_log10 ~ Aud_log10 + screen_log10 + All + Over_12 +
    Over_15 + Over_16 + Over_18 + SF + Family + Play + Horror +
    Documentary + Drama + Romans + Musical + Mystery + Crime +
    Historic + Western + Erotic + Thriller + Animation + Action +
    Adventure + War + Comedy + Fantasy + USA + Japan + Korea +
    Artistic + Not_artistic, data = df_regression)
```

Significance	**Independent variables**
0.1% level	• Audience_log10, Number of Screen_log10 • Over_12, Over_15 • Play, Musical, Erotic • Japan, Korea • Artistic
1% level	• Documentary
5% level	• USA

Challenges

H.Shin-il	J.Jw-won	K.Min-kyu
Graph insightful	First Analysis Experience	Choosing which rows contain 'Na' and '0' to erase
Replace the number of films by nationality with 'table function'.	Meeting with member & exchanging many opinions.	Lack of communication and conflict due to non-face to face.

Conclusion

- Collecting Korean movie data from KOBIS and Doing serious data cleaning

- In **EDA**, We conducted two major analyses.

- **First, two main features can be identified in time series graphs!**

  1. Impact of smartphone led to a steady decline in the box office since 2009.

  2. Covid-19 damaging the film industry in 2020 and 2021

- **Second, Linear regression was performed!**

  - Regression for elements affecting the Box office.

  - We look at the p-value and found a statistically significant element.

  - Typically, "Genre of Play" has the largest impact on Box Office with 46.7%.

Thank You