

# 夏合宿課題

B2 蓑和溪太

# テーマ

映画の予算・ジャンル・制作国・  
公開月の東京の平均気温・公開日の曜日



興行収入を予測する

# 分析方法

- 映画の情報：TMDB Box Office Prediction
- 東京の平均気温：気象庁
- 曜日：JavaScriptで計算
- Rで回帰木を作成して分析

# クレンジング

- ジャンルや制作国は最初の1つのみ抽出
- ジャンルはジャンル名、制作国は国コードを抽出
- 予算のデータがない映画は削除
- 公開日はISO 8601形式 (yyyy-mm-dd) に変換
  - 元データでは年の下2桁のみ
  - 年が24以上なら1900年代、24未満なら2000年代とみなす

# コード

```
# ライブラリの読み込み
library(rpart)
library(rpart.plot)

# CSVファイルを読み込む
movie_data <- read.csv("movie_data_merged.csv", header = TRUE, stringsAsFactors =
FALSE)

# 必要な列のみを使用して回帰木を作成
movie_tree <- rpart(revenue ~ budget + genres + production_countries + avg_temp +
day_of_week, data = movie_data, method = "anova", minsplit = 6, cp = 0)

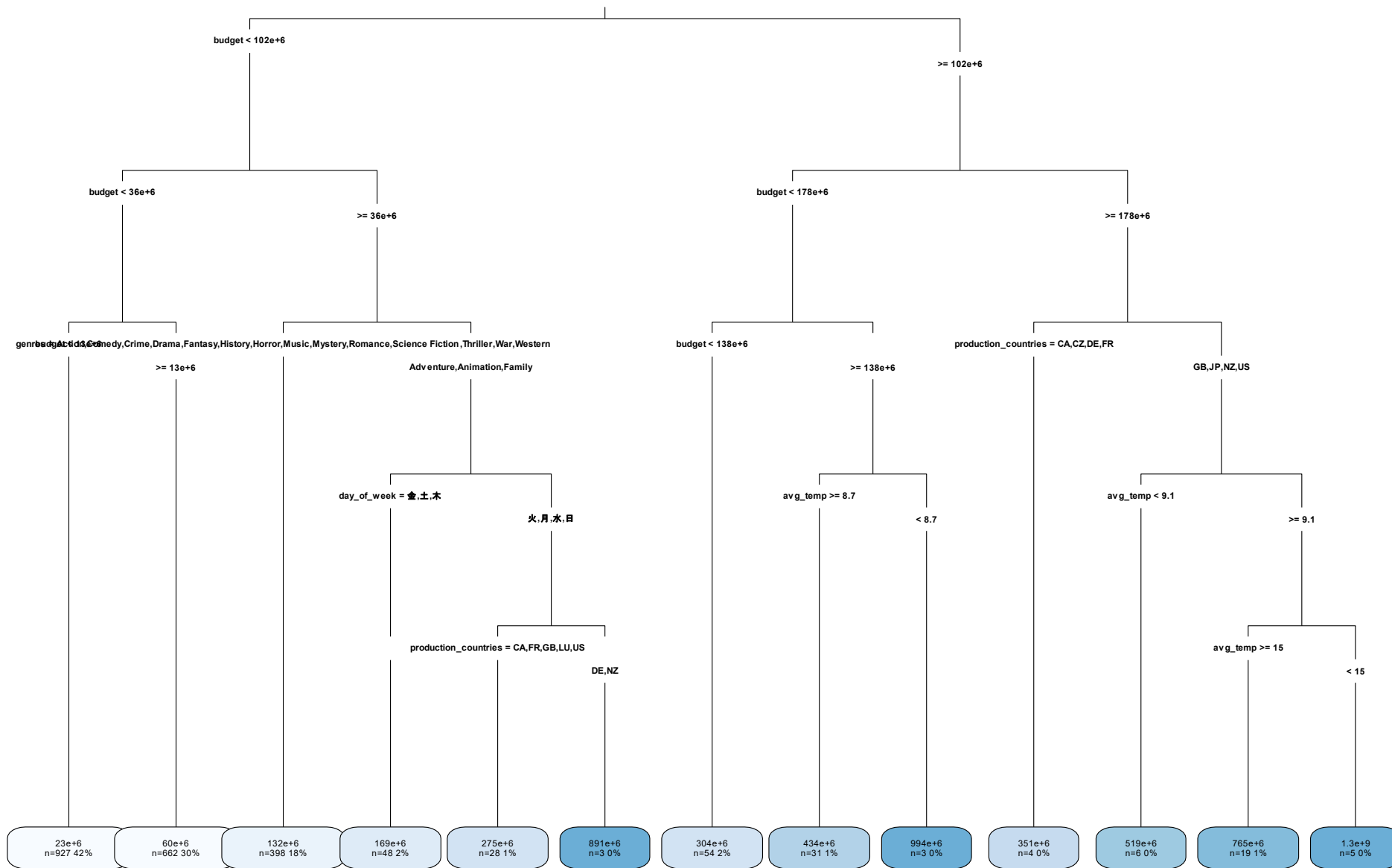
# 回帰木を剪定
movie_tree2 <- prune(movie_tree, cp = 0.01)

# 剪定後の回帰木を再度プロット
rpart.plot(movie_tree2, type = 3, extra = 101, fallen.leaves = TRUE, cex = 0.8)

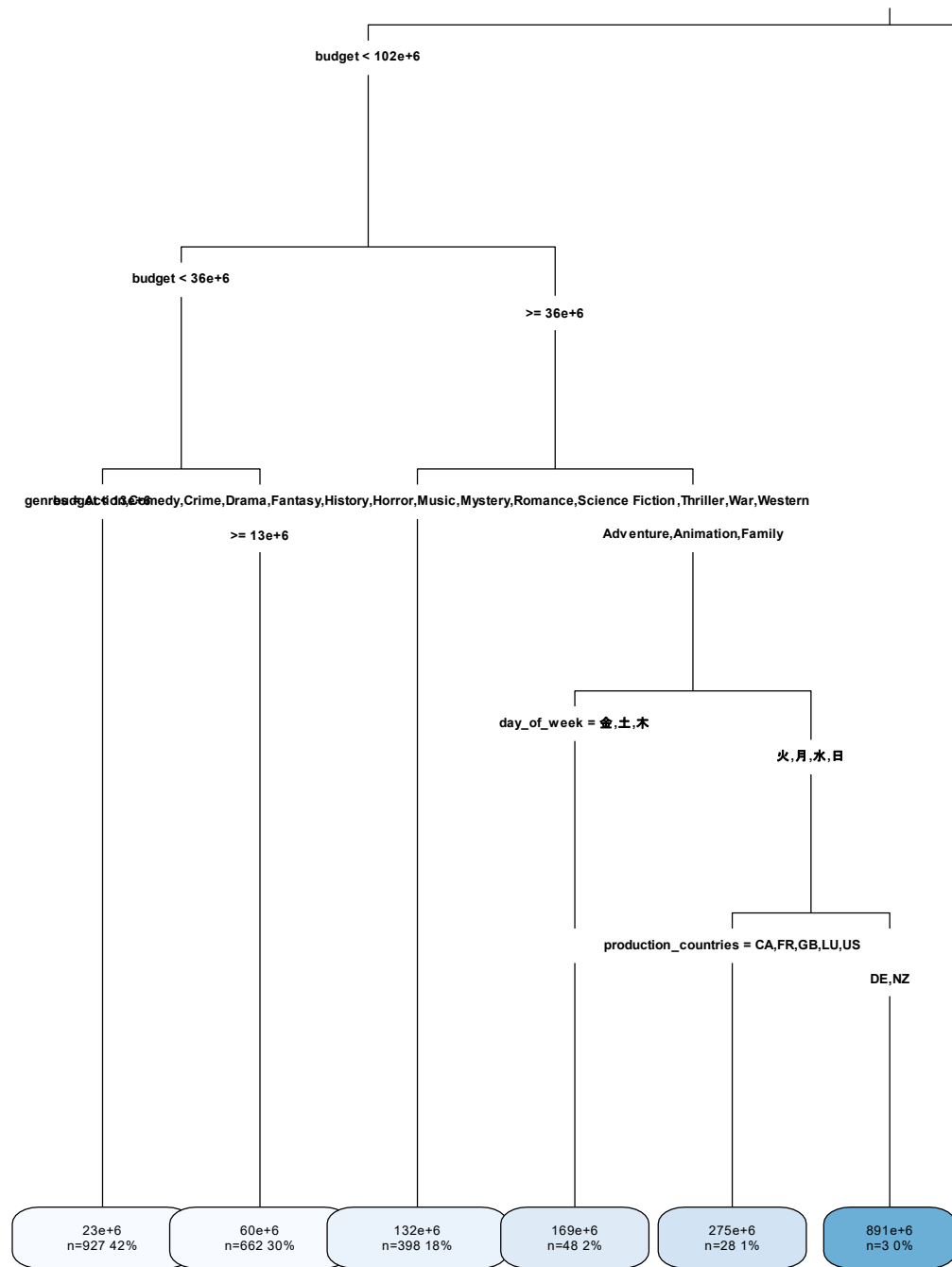
# 予測
movie_prediction <- predict(movie_tree, newdata = movie_data, type = "vector")

# 予測結果と実際の値の相関を計算
cor(movie_data$revenue, movie_prediction)
```

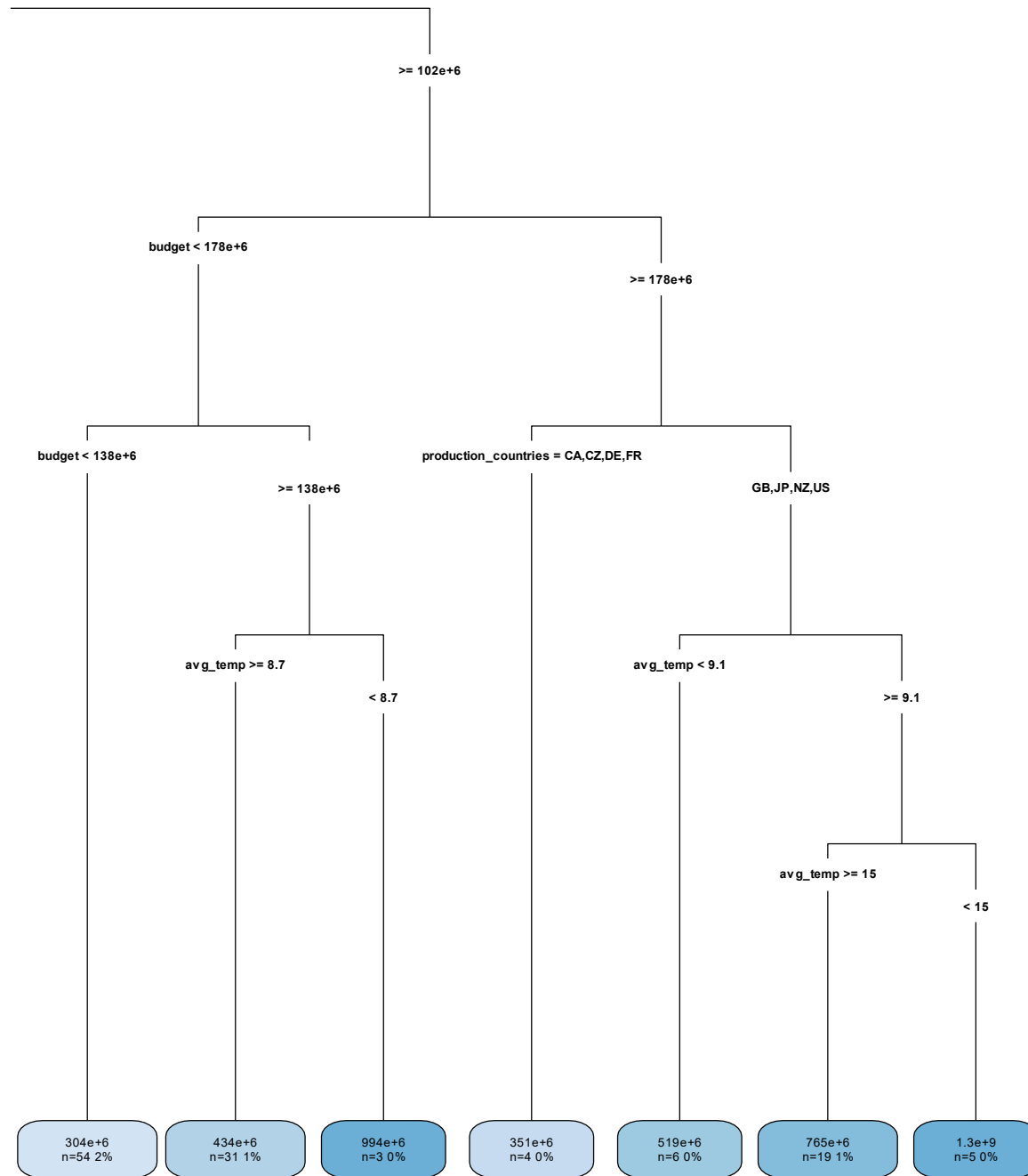
# 結果



# 結果



# 結果





# 分析

- 予算が多い方が興行収入が多い
- Adventure、Animation、Familyの興行収入が多い
- 公開日が木、金、土曜日より月、火、水、日曜日の方が興行収入が多い
- 制作国がイギリス、日本、ニュージーランド、アメリカだと興行収入が多い？
- 回帰木の先端では気温が低い方が興行収入が多い

# 考察

- 予算が多ければ人気が出やすい
- Adventure、Animation、Familyは家族で観に行くから観客数が多くなり興行収入が増える？
  - ピクサーなどの特定の人気映画が偏りを発生させている可能性も
- 週を日曜始まりとすると週の前半に公開された映画の方が興行収入が多い（理由は不明）
- 冬の映画が人気？
- 映画の制作本数
  - イギリス：6位
  - 日本：3位
  - ニュージーランド：39位
  - アメリカ：2位

# まとめ

- 映画の予算・ジャンル・制作国・公開月の東京の平均気温・公開日の曜日から映画の興行収入を分析した
- Rと回帰木を用いて分析した
- 前述のような傾向が判明した

# 参考文献

- Kaggle. “TMDB Box Office Prediction”. 2024年8月15日閲覧.  
<https://www.kaggle.com/c/tmdb-box-office-prediction/overview>
- 気象庁. 過去の気象データ. 2024年8月15日閲覧.  
<https://www.data.jma.go.jp/gmd/risk/obsdl/>
- 総務省統計局. “世界の統計2024”. 2024年8月16日閲覧.  
<https://www.stat.go.jp/data/sekai/pdf/2024a1.pdf#page=274>