

CAPGEMINI

Predicting Movie Revenue

Greenlighting Movies with Data Science.

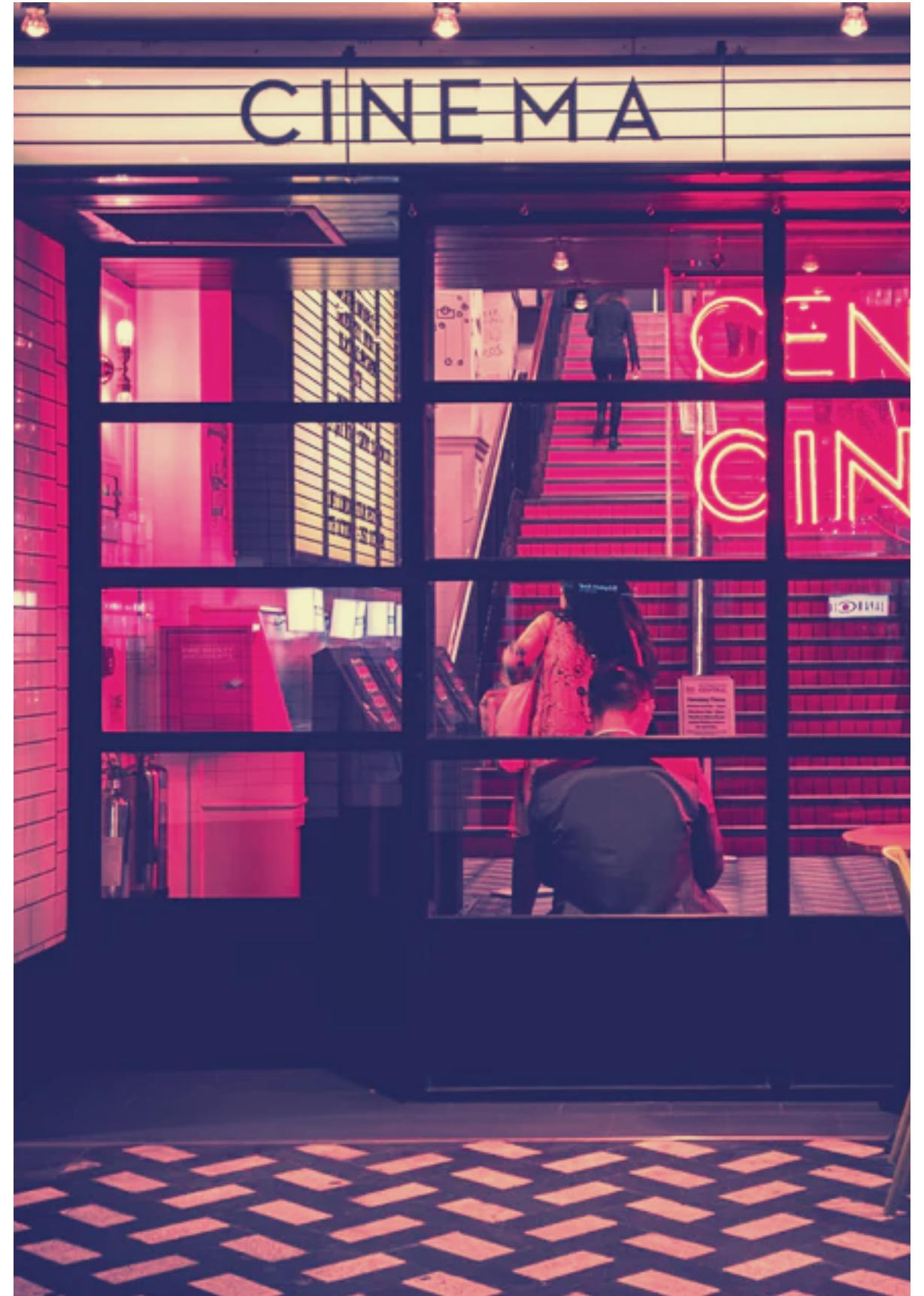
Presentation Outline

TODAY'S DISCUSSION

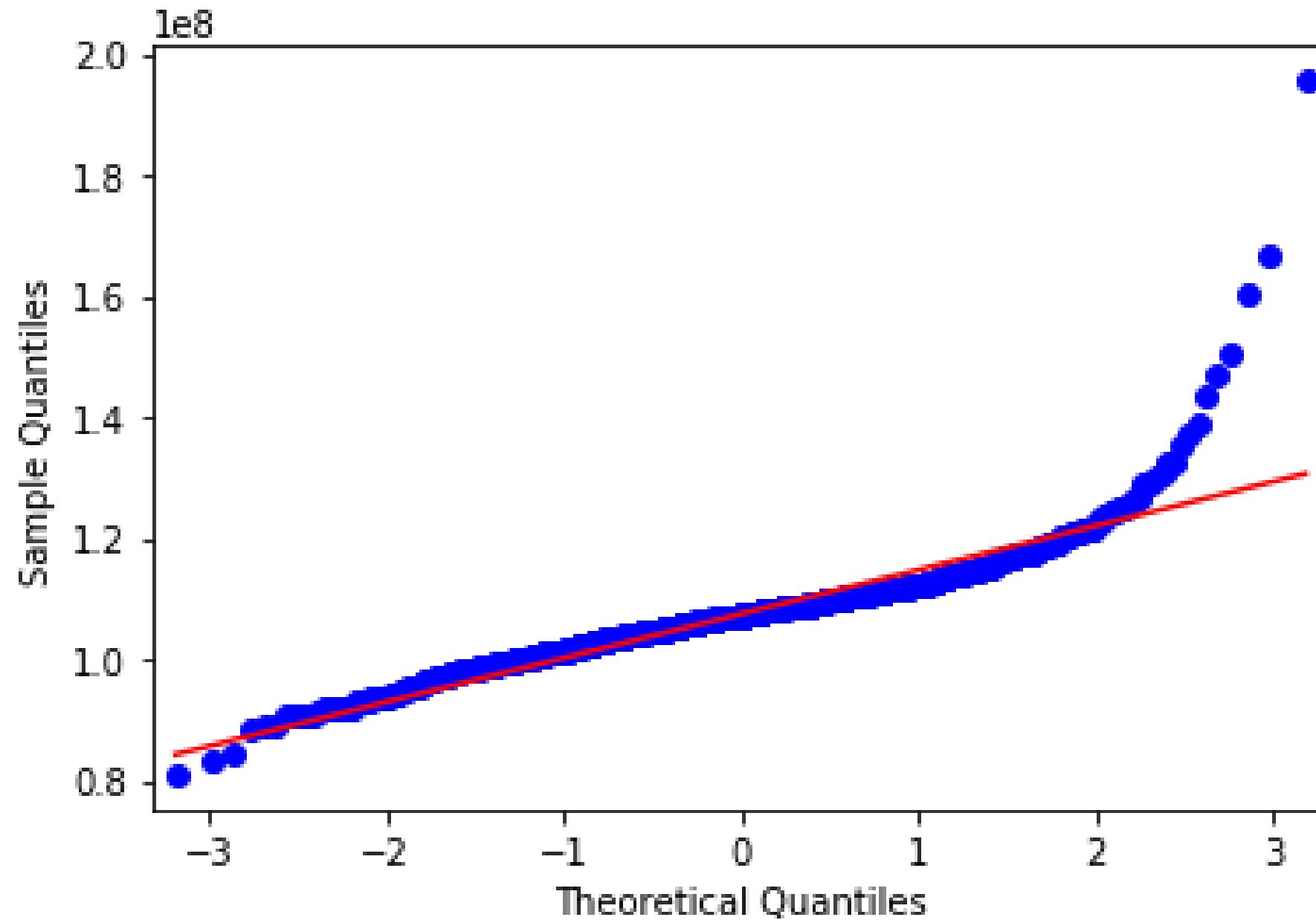
Client: Movie Studio

Problem: Predict movie revenue to greenlight the project and assign a budget

Dataset: The data is sourced from Capgemini, and is mostly comprised of categorical variables.



QUANTILE-QUANTILE PLOT



A graphical technique for determining if two data sets come from populations with a common distribution.

Movie Revenue Prediction Model

RESULTS.

MEAN ABSOLUTE ERROR (MAE)

This model predicts unseen data within a MAE of **~\$85 million**

Data Assessment

Missing Values
(NaNs/0's)

Dates as
String

Nested
Data

Text
Data

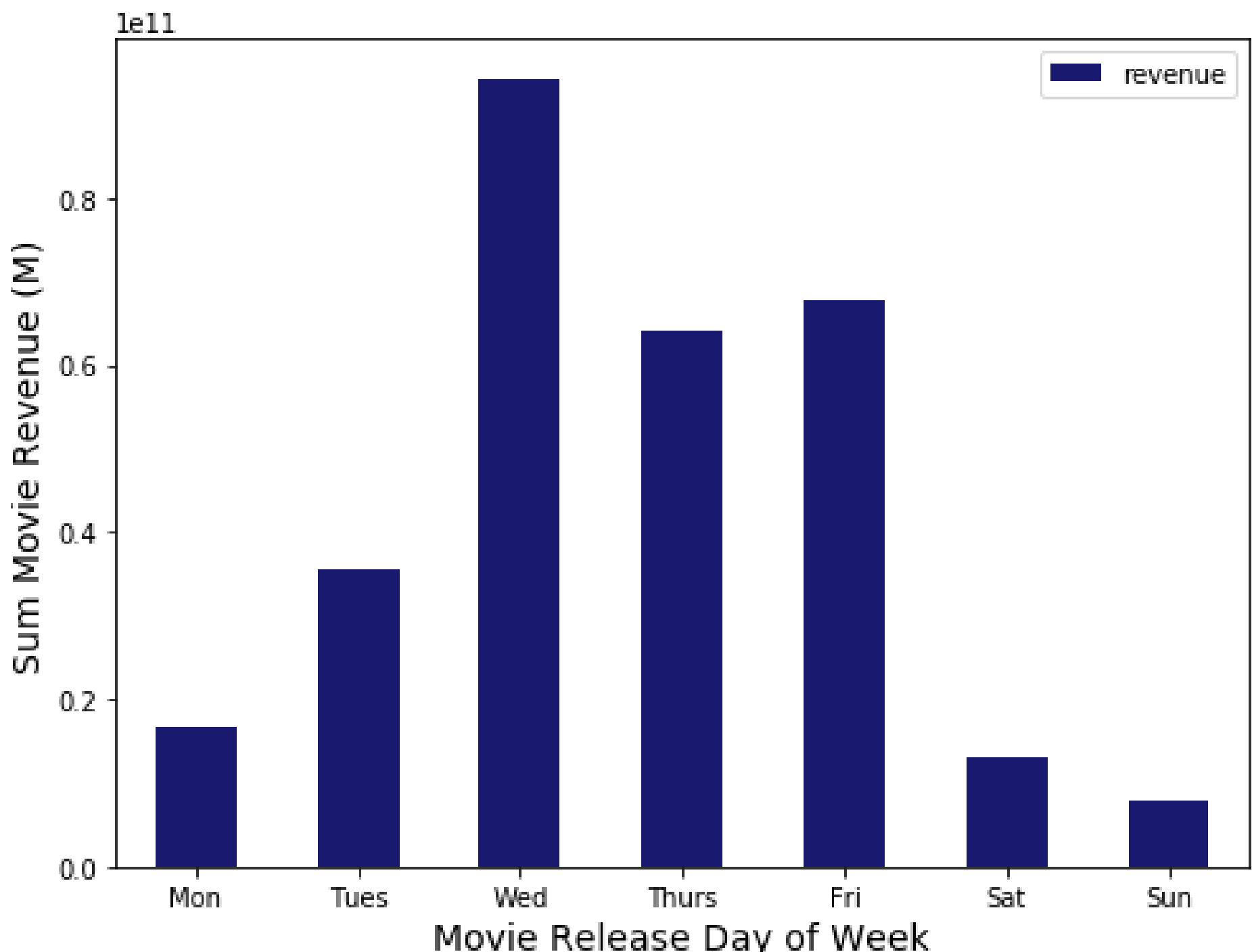
Arbitrary/
Unneeded
Columns



Data Cleaning



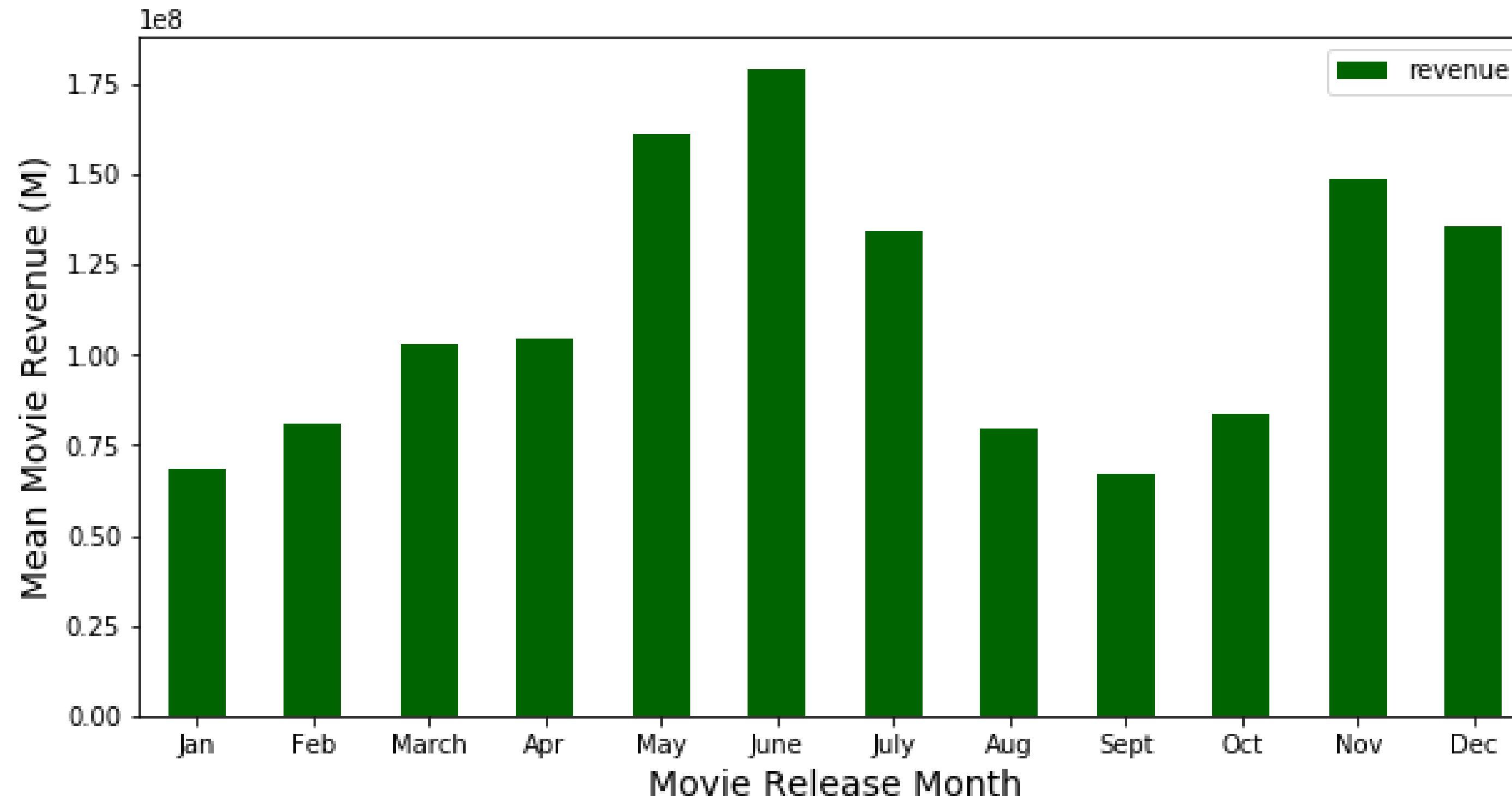
Sum Movie Revenue of Top 80 Percentile of Movie Revenues



Which Releases
Weekdays Earn the
Most Revenue for
Blockbusters?

WEDNESDAY.

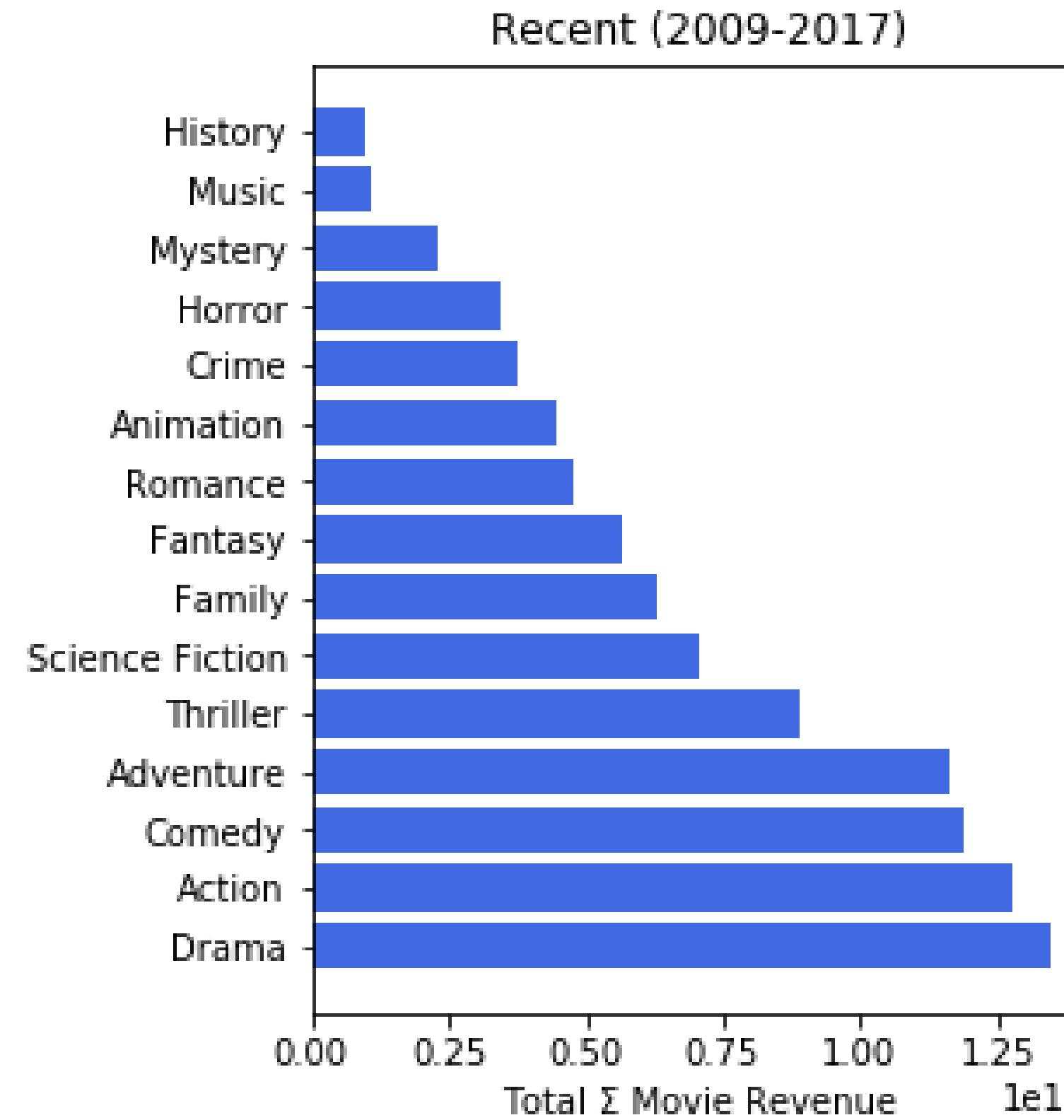
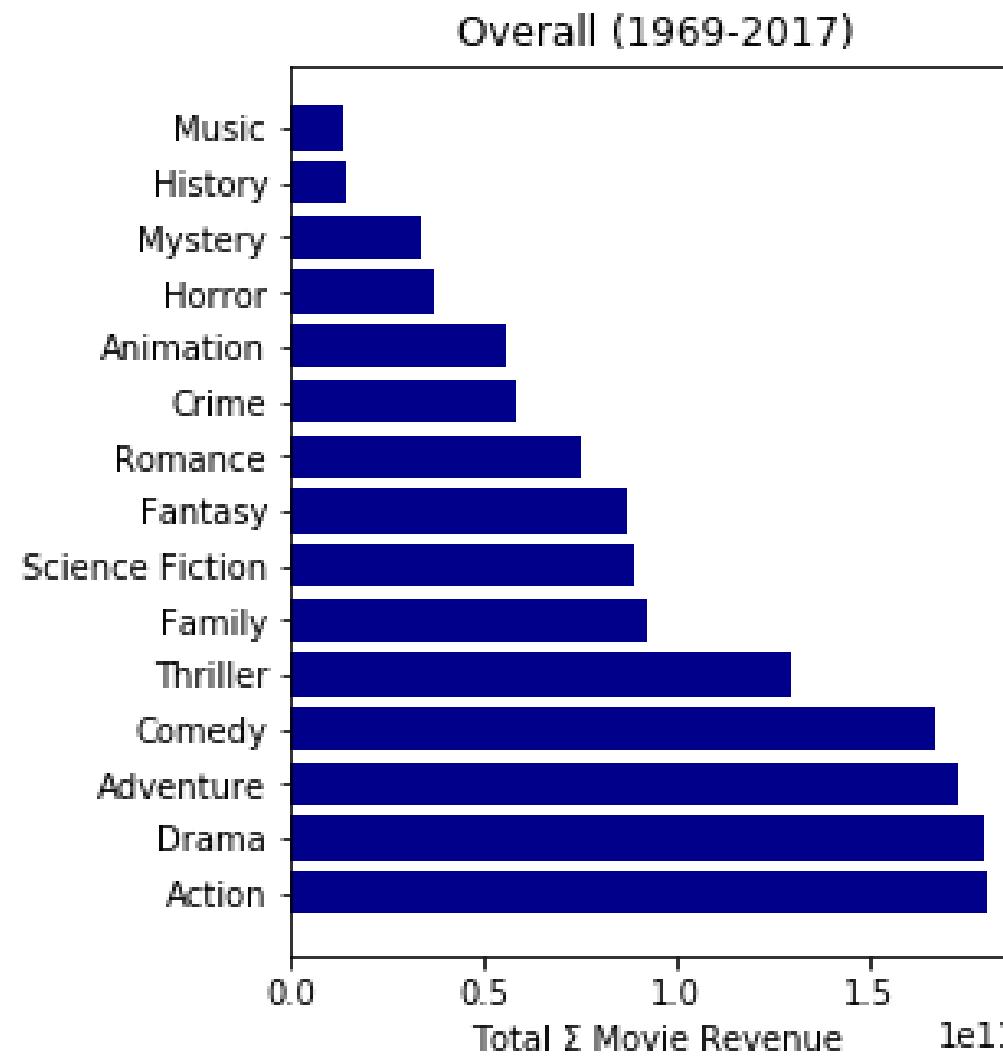
Overall, Summers and Winter Holiday Months Earn the Most Revenue



Important Notation: This dataset was sourced from Capgemini and includes the following timeframe: 01-1969 to 02-2017

Top 15 Earning Movie Genres

HIGHEST TOTAL EARNINGS



Important Notation: This dataset was sourced from Capgemini and includes the following timeframe: 01-1969 to 02-2017

Predicting Movie Revenue with Ridge Regression

using Standard Scaling and Truncated Singular Value Decomposition (SVD)

SHRINKS
COEFICIENTS

Ridge reduces model complexity and multi-collinearity

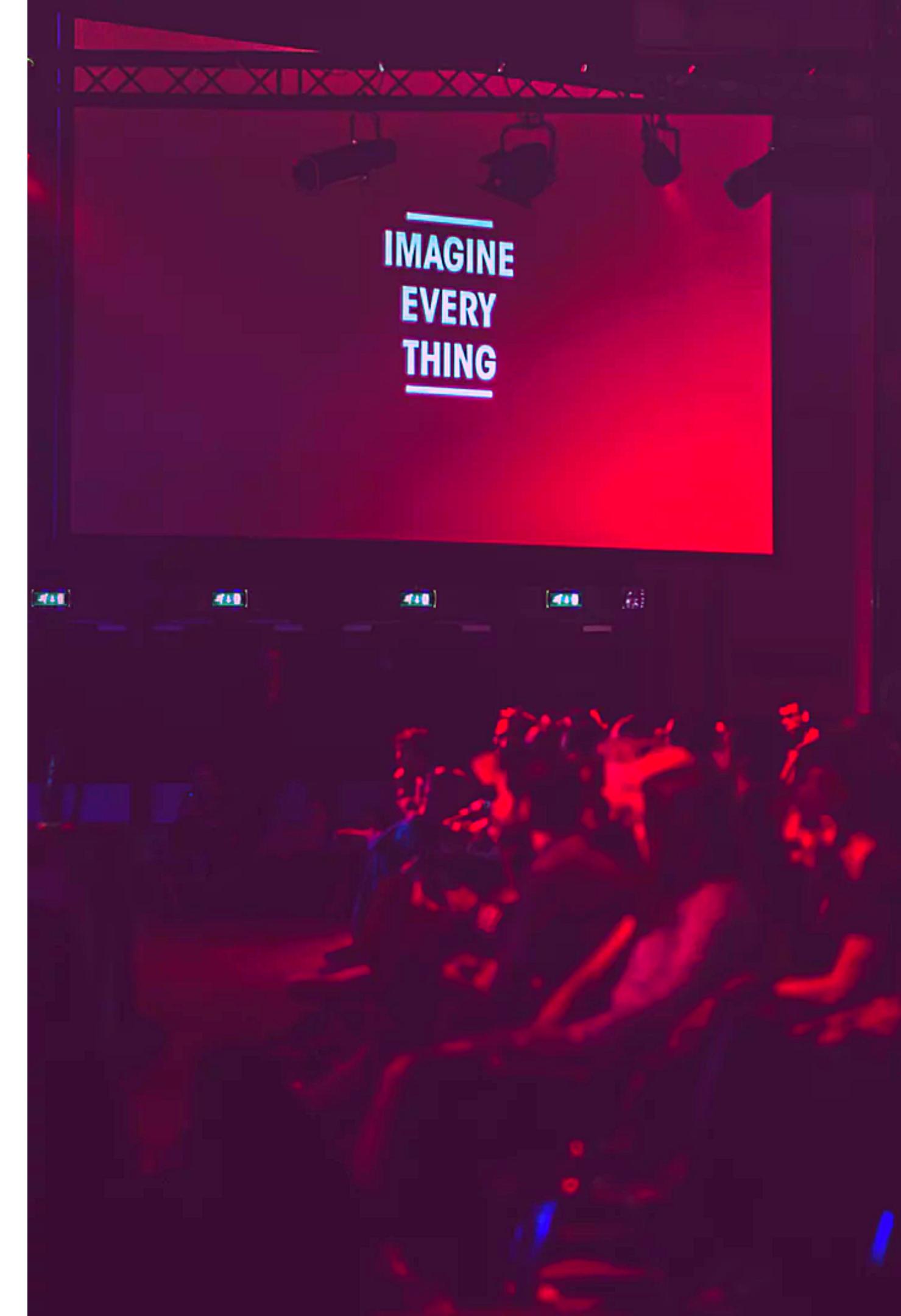
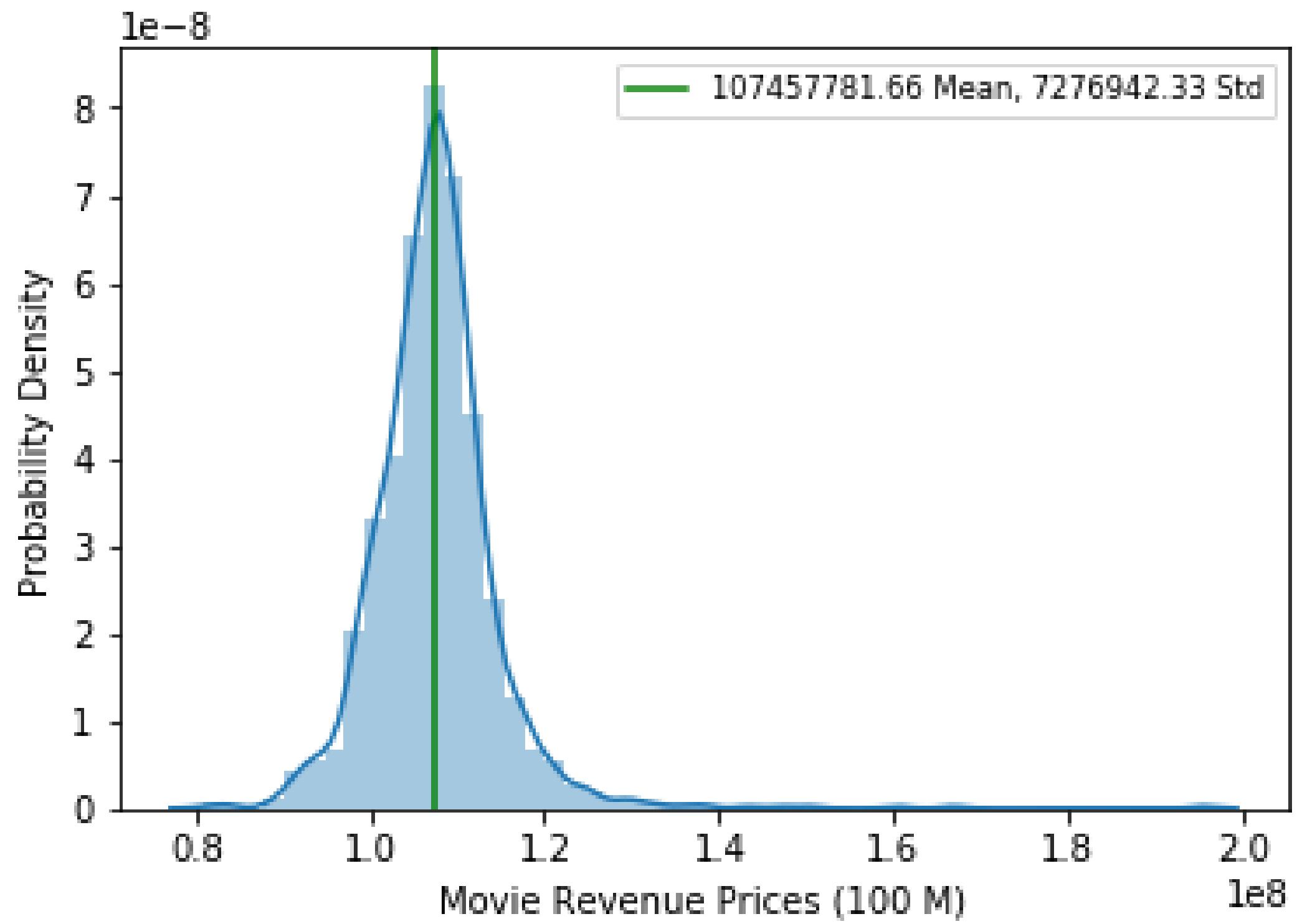
ADDS DEGREE OF
BIAS

Adding bias to estimates reduces the standard errors

MORE RELIABLE
REVENUE

Ridge produces estimates that are more reliable

Predicted Movie Revenue with Ridge Regression

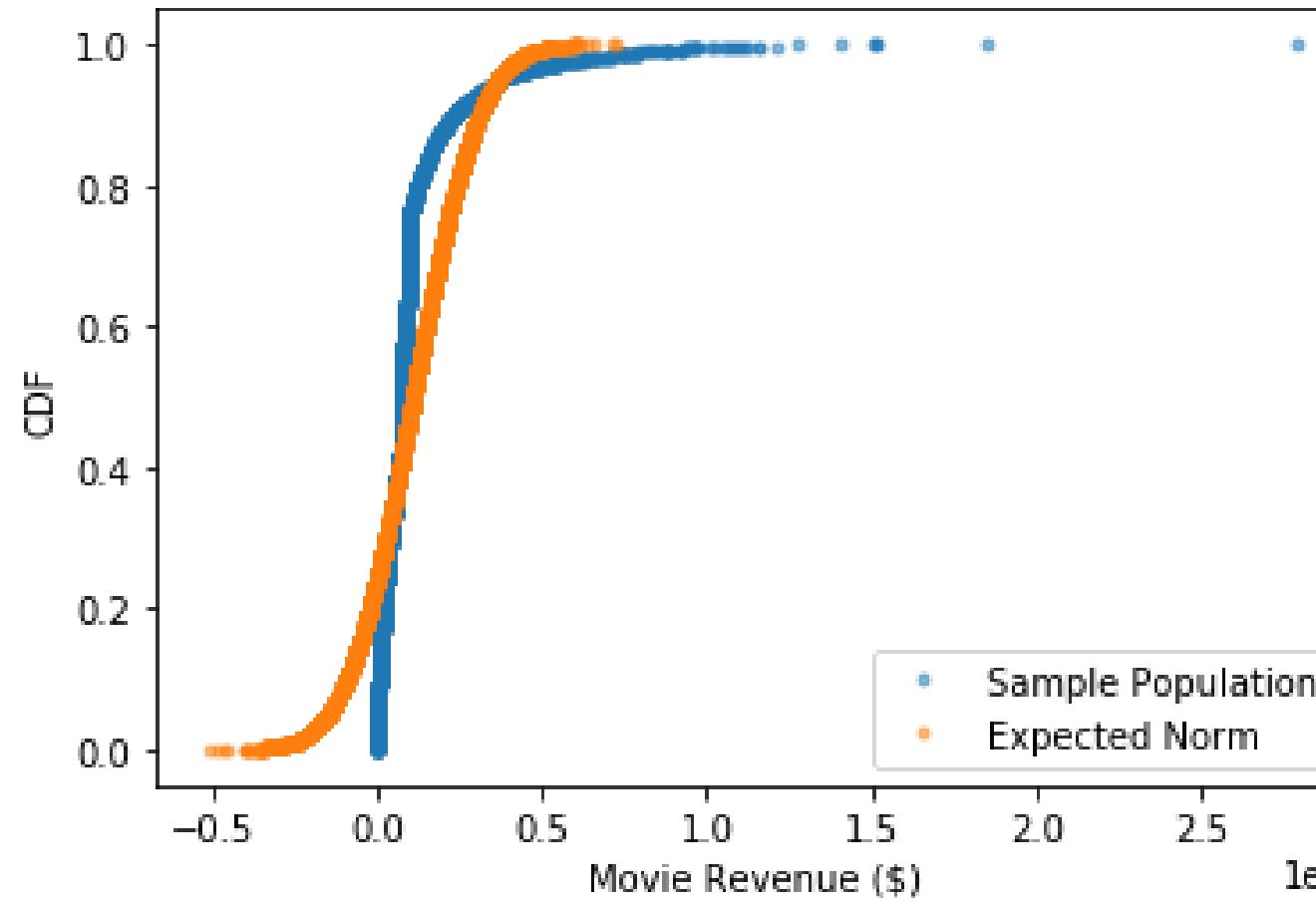




Why Does the Model Make Sense?



Distribution of Movie Revenue (\$)



Revenue
Increases
Over Time

Movie Revenue in the Dataset was Skewed Up

