GEA1000 Tutorial 1
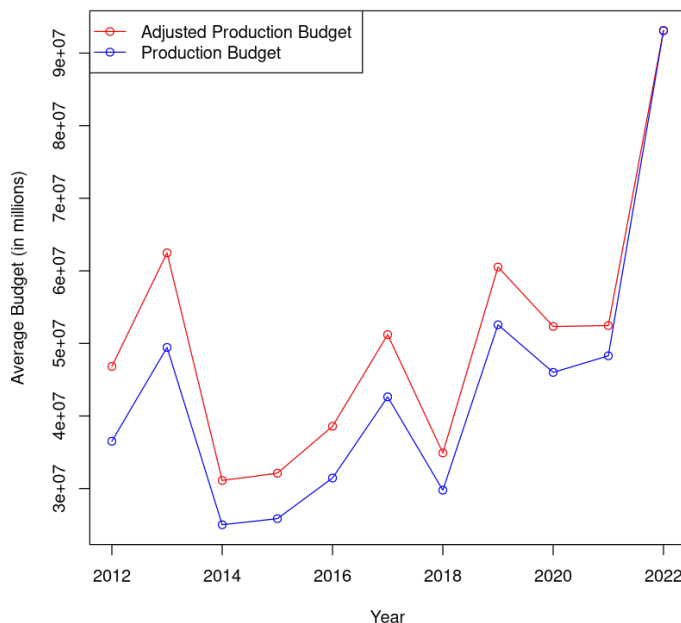AY 24/25 Sem 2 — github/omgeta

Q1. (a.) Sampling frame is the list of all 6420 movies from the website. Sampling method is Simple Random Sampling.

(b.) Categorical: Title, MPAA_Rating, Genre, Release_Year
Numerical: Production Budget, Worldwide_Gross, CPI, Duration, IMDb_Rating, Voter_Numbers

(c.) Release_Year, CPI, Genre have small number of missing values.
MPAA_Rating, IMDb_Rating, Voter_Numbers have medium number of missing values.
Duration has very large number of missing values.
For small number of missing values, we can just drop the rows, since the impact on the data set will be minimal. However, for large number of missing values, we can't because we can't drop too much data.

(d.) Minimum: 0 (Possible anomaly)
Q_1: 4193096
Median: 25873142
Q_2: 97818746
Maximum: 2064615817
Mean: 96720426
S.D.: 196370338

(e.) (See code)

(f.) The trendlines are similar but the adjusted production budget is consistently higher until it converges at 2022. There are peaks at 2013, 2017, 2019 and 2022, with production budgets going up very high continuously since 2021

Trends in Production Budget vs. Adjusted Production Budget (2012-2022)

(g.) While we use probability sampling, large sample size and have little/no bias, our sampling frame size is only $\tilde{6}000$ movies and not $\geq$ population size (all movies). Therefore, we cannot generalize our results to the whole movie industry.

Q2. (a.) 349 students

(b.) Family size more than 3: 0.72
Receiving tuition: 0.46
Mothers with degree: 0.34
In a romantic relationship: 0.32

(c.) Students in both groups different but it is not surprising as we have a relatively small dataset.

(d.) Exact proportion value between control and treatment differ. However, they are generally similar due to random assignment.

(e.) No, they will have different proportions between schools. However, the similarity between control and treatment groups within the same school will be the same. This is because the different schools have different properties.