IMDb Insight

What is IMDb?

- IMDb คือ เป็นฐานข้อมูลออนไลน์รวบรวมเกี่ยวกับเรื่องราวของนักแสดง ผู้กำกับ และบุคคลที่เกี่ยวข้องในวงการภาพยนตร์ และยังเป็นคอมมูนิตี้สำหรับคนดูหนัง เปิดให้คนจากทั่วโลกเข้ามาให้คะแนนหนังที่ดูจบไปแล้ว รวมทั้งร่วมวิจารณ์บนเว็บไซต์ได้เลย ถือเป็นเว็บไซต์หนังที่ได้รับความนิยมทั่วโลก เราจึงมักจะใช้คะแนนจากนักรีวิวบน IMDb มาเป็นตัวเลขอ้างอิงว่าหนังเรื่องดังกล่าวถูกจริตผู้คนทั่วโลกอย่างไร
- คะแนนของ IMDb มาจากไหนบ้าง?
  คะแนนของหนัง มาจากการ่วมโหวตของคนทั่วโลกที่ เพียงสมัครสมาชิกบนเว็บไซต์ ก็เข้าไปกดโหวตได้ทันที คะแนนจะมีตั้งแต่ 1-10

Github Code

- https://github.com/pongsapaks/DADS5001-Final.git

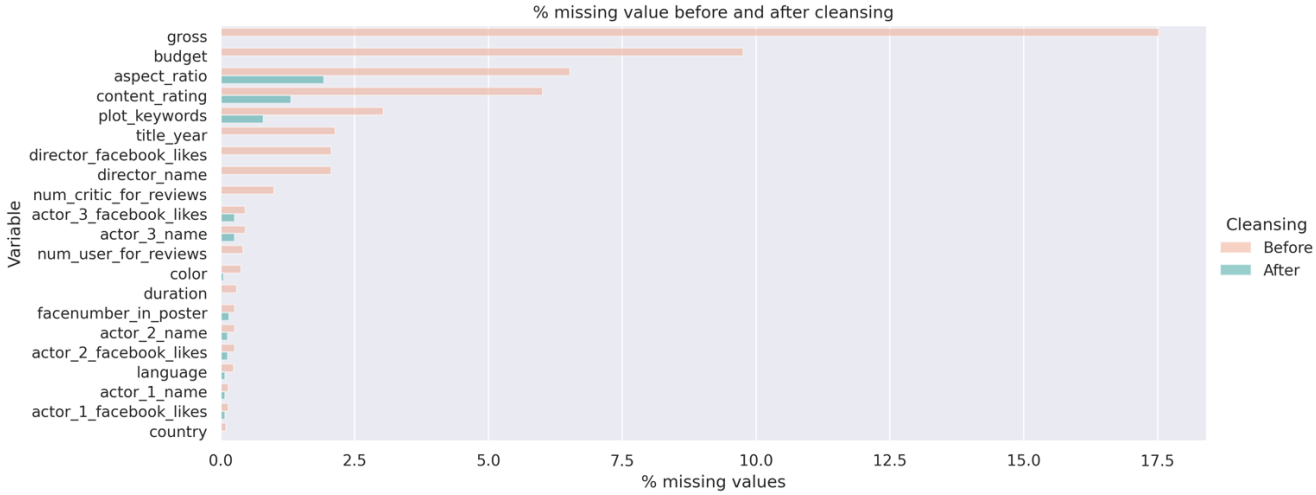
Dataset Information

- ข้อมูลที่เกี่ยวข้อง ในwebsite IMDb https://github.com/yash91sharma/IMDB-Movie-Dataset-Analysis/blob/master/movie_metadata.csv
- https://www.imdb.com
- https://www.jagranjosh.com/general-knowledge/what-is-internet-movie-database-imdb-all-you-need-to-know-1665565087-1

DATA Preprocessing

Cleansing data

- Cleansing data โดยการดูข้อมูลที่มี rows data จากการหาจำนวน nan มาวัดเป็น percent และ drop rows ที่เป็นหัวข้อของข้อมูล percent missing value สูงมากกว่า 7.5 % จึง ดรอปข้อมูล 'Gross' และ 'budget' เนื่องจาก percent สูง และต้องการใช้งานทั้ง 2 ตัวแปรนี้ ผลลัพธ์ที่ได้ บาง rows ที่ถูกลบออกไปเป็น rows ที่มีค่า nan หายไปจำนวนมากอยู่แล้วทำให้ผลกระทบ percent ของตัวอื่นน้อยลงไปด้วย
- แล้วโชว์กราฟอีกรอบว่าอะไรหายไปบ้าง ซึ่งคราวนี้ %missing values จะมีค่าน้อยลง จนถึงขั้นที่ accceptable จึงใช้ข้อมูลนี้ต่อ



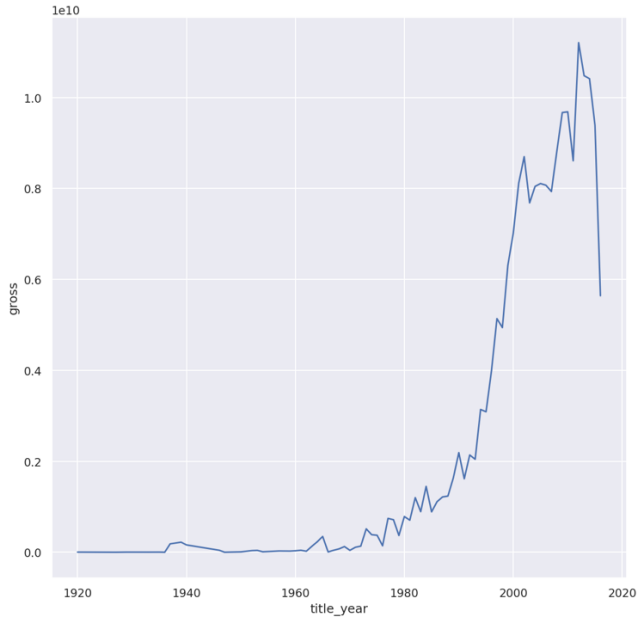
Correct data

- มาตรวจสอบข้อมูล ว่าข้อมูลส่วนไหนจำเป็นต่อการใช้งานของเราบ้าง และนำข้อมูลมาตรวจสอบจะเจอว่าข้อมูล country ไม่ถูกต้อง >> จึงทำการแก้ไขให้ถูกต้อง

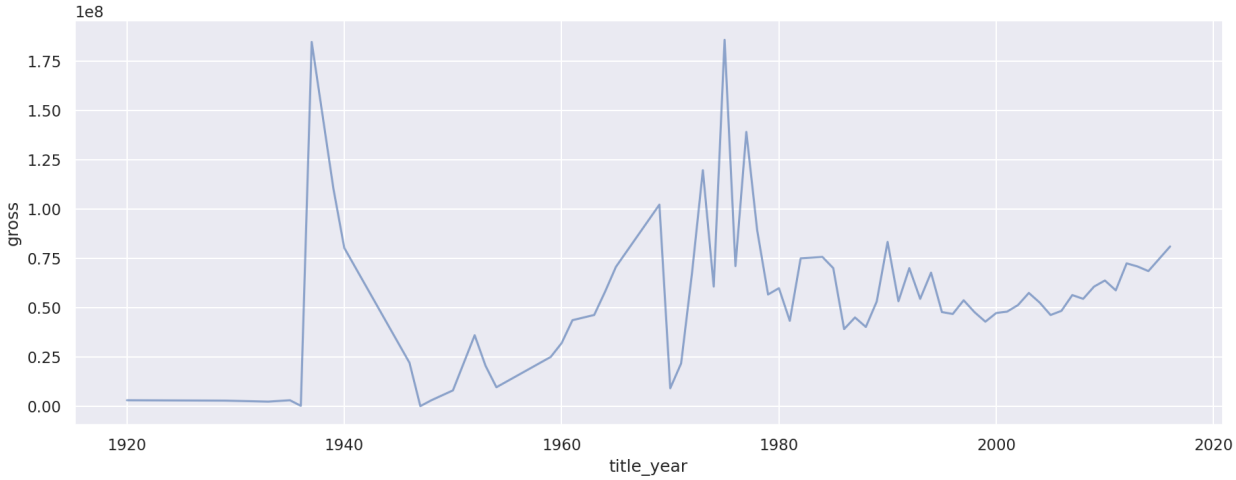
OBJECTIVE: Use information from the dataset to answer the following question

1. Industry Growth (**ธุรกิจมีการเจริญเติบโตสูงขึ้นหรือไม่**)

จากกราฟ 1.1) จะเห็นว่ากราฟค่ารวมรายได้ของหนังต่อปี มีรายได้ที่เพิ่มขึ้นเรื่อยๆต่อปีอย่างมาก แต่กลับกัน พอมาดูกราฟ 1.2) ที่เป็นค่าเฉลี่ยรายได้ต่อปี กลับพบว่าไม่ได้มีค่าสูงขึ้นแต่อย่างใด ซึ่งสาเหตุนั้นมาจากในแต่ละปีนั้นก็มีจำนวนหนังที่เพิ่มสูง (คู่แข่งจำนวนมากขึ้น)ซึ่งทำให้มีการแบ่ง market share ออกไปด้วยเช่นกัน



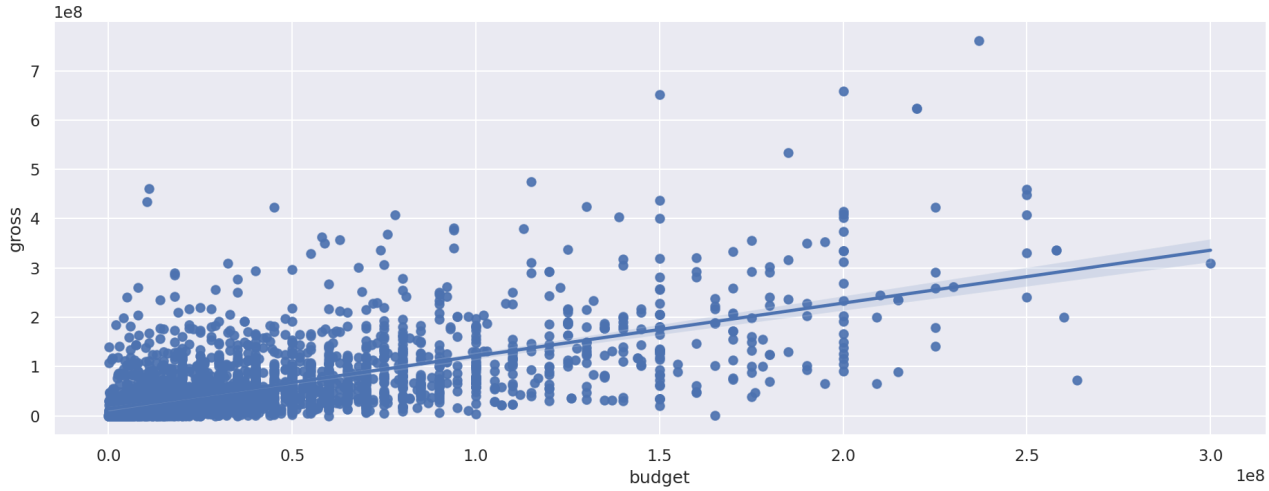
กราฟ 1.1)



กราฟ 1.2)

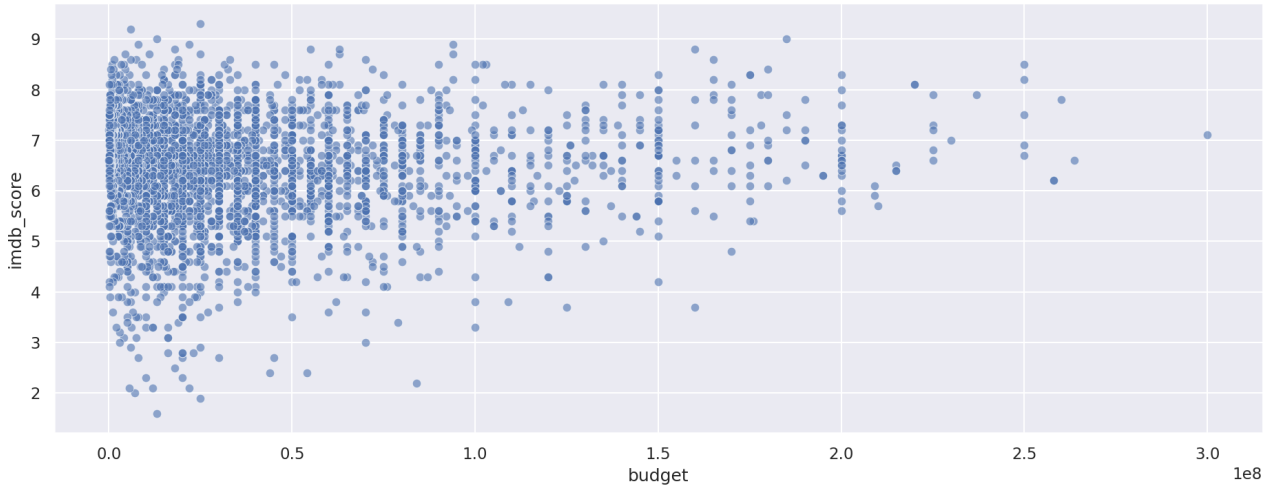
2. หนังที่มีต้นทุนสูง จะมีรายได้สูง และหนังที่มีต้นทุนสูง จะมีimdb_score สูง จริงหรือไม่?

จากกราฟ 2.1) พบว่าหนังที่มีต้นทุนสูง นั้นจะส่งผลให้มี รายได้ที่เพิ่มสูงตาม **และมีแนวโน้มเป็นความสัมพันธ์แบบแปรผันตรงเชิงบวก**



กราฟ 2.1)

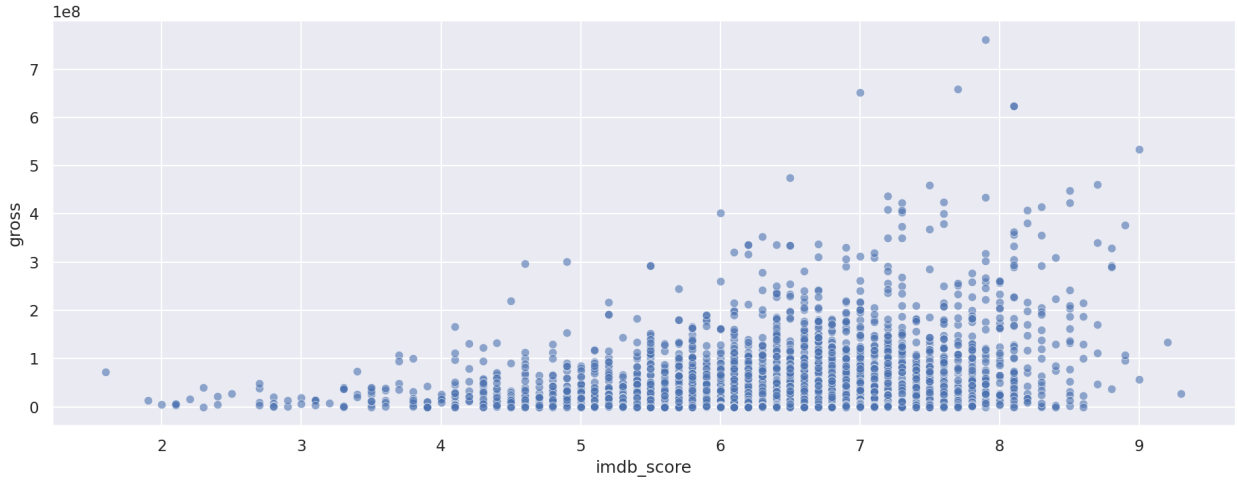
แต่กลับกัน การที่มีต้นทุนในการทำหนังที่สูง มีแนวโน้มส่งผลให้มี imdb_score ที่สูง แต่หนังที่มีต้นทุนในการทำหนังที่ต่ำ ไม่ได้หมายความว่าจะมี imdb_score ที่ต่ำเสมอไป ดังกราฟ 2.2) scatter plot ด้านล่าง



กราฟ 2.2)

3. หนังที่มี imdb_score สูง นั้นจะทำให้มีรายได้สูงเสมอหรือไม่?

จากกราฟ 3.1) พบว่า imdb_score ที่มีค่าต่ำนั้นทำให้มีรายได้อยู่ในระดับต่ำอย่างแน่นอน แต่หนังที่มีคะแนน imdb_score ที่สูงนั้นมีโอกาสที่จะมีรายได้ที่สูง หรือต่ำก็ได้ (มีการกระจายในวงที่ค่อนข้างกว้าง)



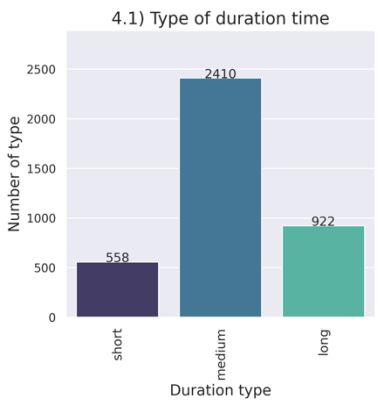
กราฟ 3.1)

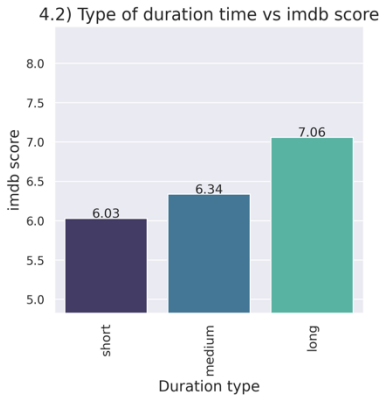
4. หนังที่มีความยาวในช่วงไหน มีแนวโน้ม ได้คะแนนสูง และรายได้เฉลี่ยสูง

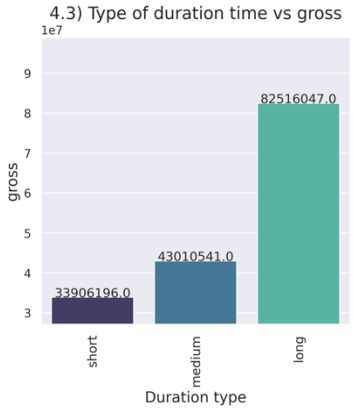
เราแบ่งความยาวหนัง ออกเป็นทั้งหมด 3 ช่วง ได้แก่ Short (0 - 90 mins), Medium (90 - 120 mins), Long (120 ขึ้นไป)

สรุปได้ว่า ช่วง Long หรือระยะเวลาหนังตั้งแต่ 120 นาทีขึ้นไป (กลุ่ม Long) จะทำให้มี ค่าเฉลี่ย imdb_score และ รายได้ที่สูงกว่าในช่วงกลุ่มอื่น

หมายเหตุ ตรงนี้ควรได้รับการศึกษาเพิ่มเนื่องจาก perfect duration ของหนังแต่ละ Genres นั้นค่อนข้างแตกต่างกันไป ตามประเภทของหนัง

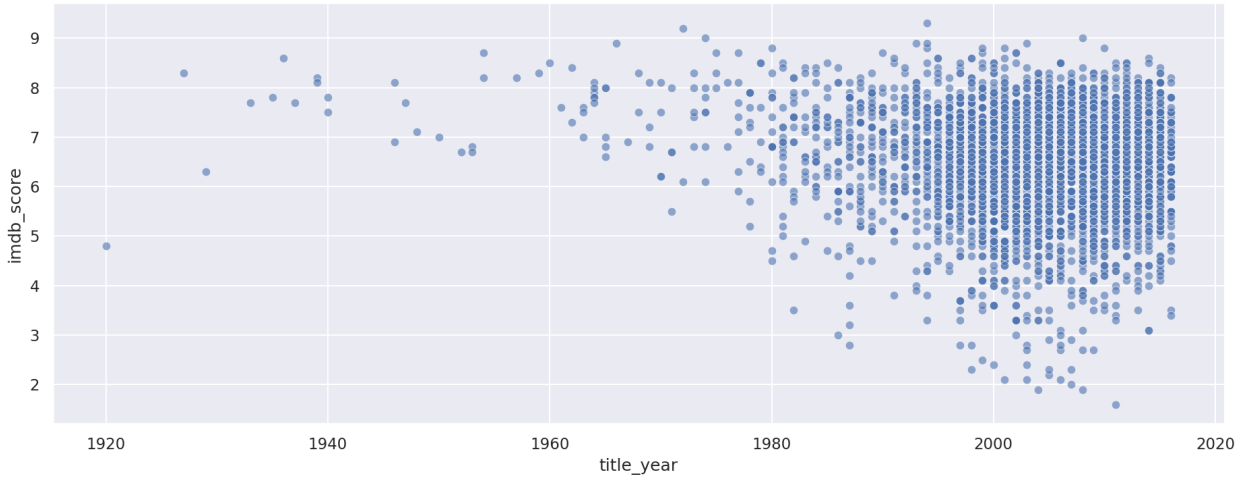






5. ทำไมหนังเก่า ๆจึงมี imdb_score ที่สูง อะไรเป็นสาเหตุดังกล่าว?

จากกราฟ 5.1) จะเห็นการกระจายของ imdb_score ของในแต่ละปีที่ผลิตหนัง ซึ่งพบว่าในปีเก่าๆนั้นจะมีคะแนน imdb_score มีค่าค่อนข้างสูง เมื่อเทียบกับปีหลังๆ



กราฟ 5.1)

โดยพบว่า **หนังที่ผลิตก่อนปี** 2000 นั้น ติดอันดับ Top 10 ของทั้งหมด ในสัดส่วน 6 **ใน** 10

โดยเรื่องที่ติด อันดับ 1 อย่าง The Shawshank Redemtion นั้นครองใจคนดูเป็น Top 1 ตลอดมา

Top 10 before 2000

```
1806                          The Shawshank Redemption
3041                                      The Godfather
2564                             The Godfather: Part II
3638                 The Good, the Bad and the Ugly
1748                                  Schindler's List
2962                                       Pulp Fiction
1904    Star Wars: Episode V - The Empire Strikes Back
653                                        Fight Club
795                                       Forrest Gump
1774                                        Goodfellas
Name: movie_title, dtype: object
```

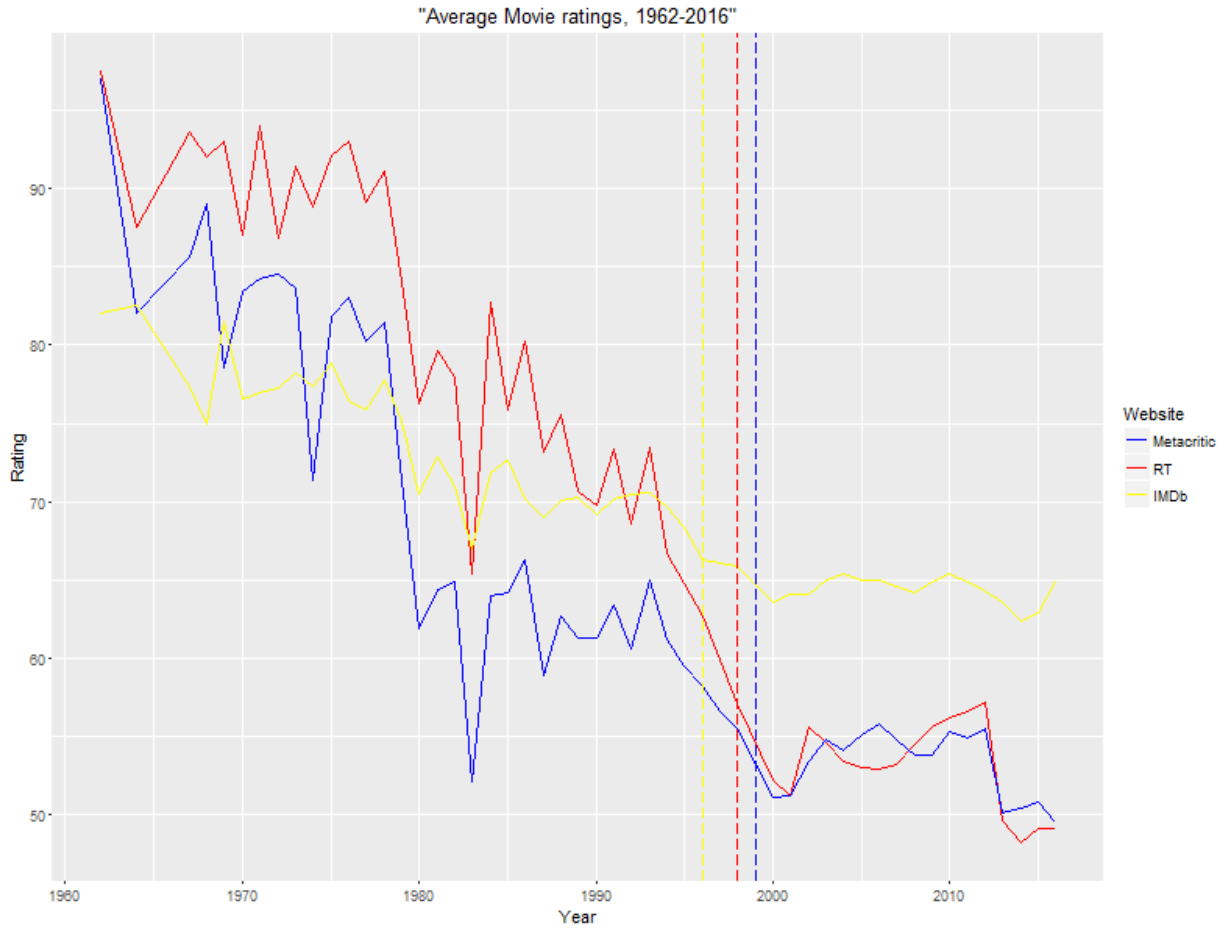
Top 10 after 2000

```
1806                                    The Shawshank Redemption
3041                                                The Godfather
64                                                The Dark Knight
2564                                 The Godfather: Part II
2962                                                 Pulp Fiction
1748                                          Schindler's List
3638                          The Good, the Bad and the Ugly
328                 The Lord of the Rings: The Return of the King
94                                                     Inception
260     The Lord of the Rings: The Fellowship of the R...
Name: movie_title, dtype: object
```

ซึ่งประเด็นนี้ก็เป็นประเด็นที่ค่อนข้างน่าสนใจ มีคนจำนวนนึงได้ตั้งเป็นประเด็นถกถามเหมือนกัน ว่าในแต่ละ website หนังปีเก่าๆค่อนข้างมีคะแนนเฉลี่ยที่สูงกว่าปัจจุบัน

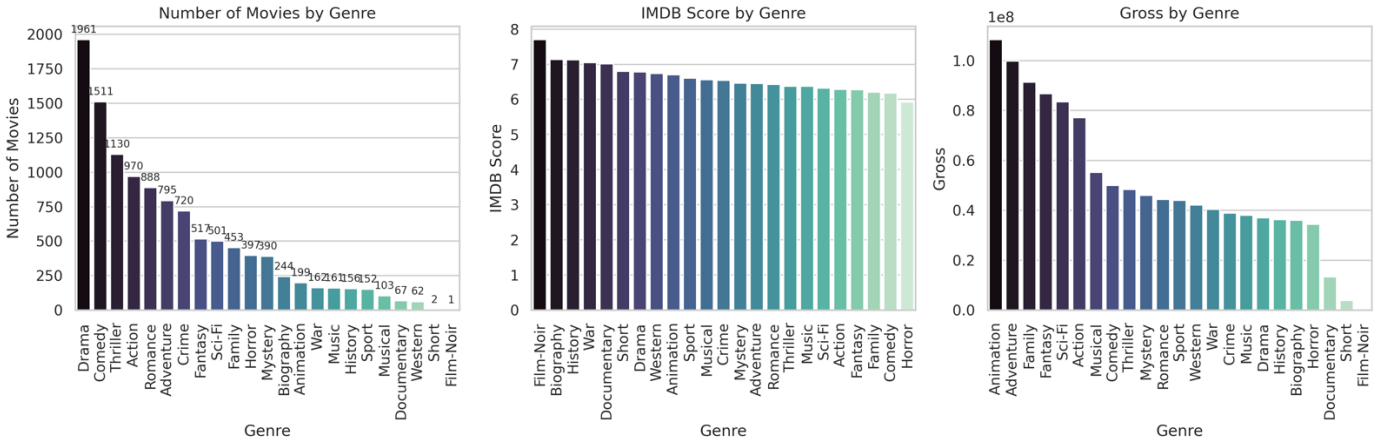
สาเหตุนึงอาจเป็นไปได้ว่าในปัจจุบันหนังที่ถูกผลิตออกมา มีจำนวนค่อนข้างเยอะให้เลือก และกระบวนการผลิตนั้นมีความเร่งรีบทำให้คุณภาพไม่ได้ดีในทุกเรื่อง และอีกเหตุผลอาจเป็นไปได้ว่าคนที่ให้คะแนนส่วนใหญ่หนังเก่าส่วนใหญ่เวลาที่ดูหนังช่วงยุค 198x -199x แล้วมีความรู้สึก nostalgia หวนนึกถึงวันเก่าๆ ความทรงจำเก่าๆ ยุคที่ยังไม่มี smartphone มีการเล่าเรื่องที่ค่อนข้างเป็นเอกลักษณ์ ซึ่งค่อนข้างแตกต่างจากหนังในยุคปัจจุบันที่มีการนำเทคโนโลยีในการสร้างแสง สี เสียง แบบอลังการ

กราฟด้านล่างแสดงให้เห็น การวิเคราะห์คะแนนเฉลี่ยจากในเว็บต่างๆ ref movie rating



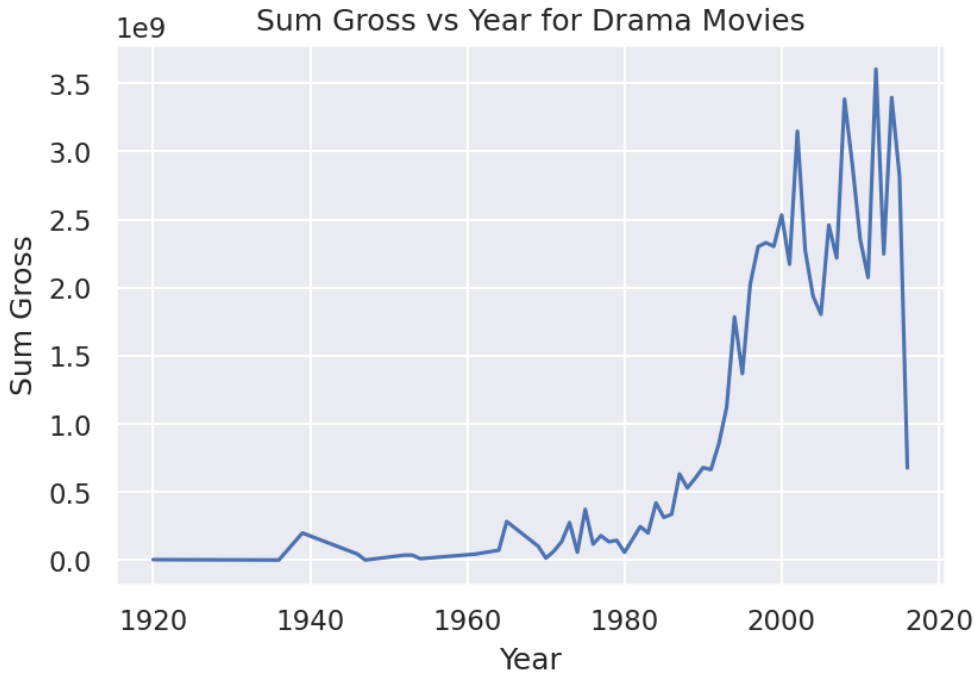
6. หนังประเภทไหนมีคะแนนเฉลี่ยสูง และได้รับความนิยมในหมู่คนดู

หนังที่คนนิยมนั้น วัดจากหนังที่ทำรายได้ สูง 3 อันดับแรกที่ทำรายได้สูงสุด(gross) คือ Animation, Adventure, Family (ซึ่งมีจำนวนสัดส่วนกลาง ถึง ค่อนข้างมาก) ส่วนหนังประเภทที่ได้ imdb_score สูงสุดนั้นไม่สามารถนำไปใช้ได้เนื่องจากมีสัดส่วนจำนวนหนังค่อนข้างน้อย ทำให้ค่าเฉลี่ยนั้นเมื่อนำมาวิเคราะห์จะไม่ถูกต้อง



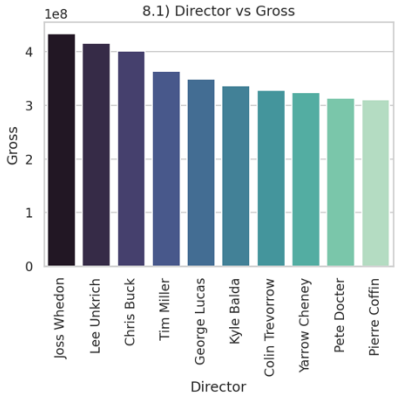
7. หนัง Drama มีจำนวนเยอะใน IMDb website ในแต่ละปีมี รายได้รวมเป็นแนวโน้มอย่างไร

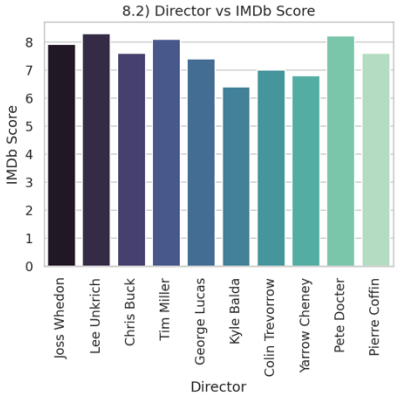
กราฟ 7.1) รายได้รวมของหนัง drama ในแต่ละปีก็เพิ่มขึ้นเช่นกัน หนัง Drama อาจเป็นอีกทางเลือกของคนทำหนัง แต่ผู้สร้างหนังก็ควรระวังเนื่องจากสัดส่วนหนัง drama ในตลาดนั้นก็มีเยอะเช่นกัน

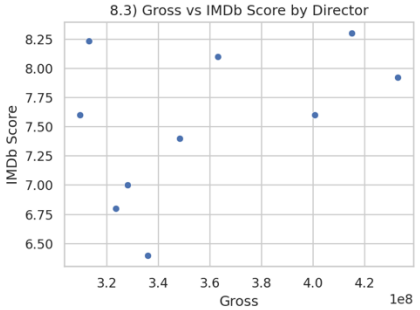


8. Top 10 Director

กราฟดังกล่าวแสดง ชื่อ Director ที่ทำหนังแล้วมีรายได้สูง 10 คนแรก มีรายการดังนี้ (กราฟ 8.1) นำรายชื่อ Director แต่ละคนที่ทำรายได้สูงสุดมาดูค่า imdb_score ซึ่ง Director ที่ทำหนังได้รายได้สูง ไม่ได้มี imdb_score ที่สูงตามแบบมีความสัมพันธ์กัน แต่ทุกๆคนนั้นมี imdb_score อยู่ในระดับตั้งแต่ 6 ขึ้นไป (กลางค่อนสูง จนถึงสูง)







Summary

- จากที่ได้ทำการศึกษาในแง่มุมต่างๆ สามารถสรุปเป็น action plan ให้ผู้จัดทำหนังได้ดังนี้
    1. วงการหนังมีการเติบโตขึ้นสูงมากจากปีเก่าๆ แต่ในการสร้างหนังผู้จัดควรจะพิจารณาให้ดีเนื่องจากรายได้เฉลี่ยในแต่เรื่องไม่ได้สูงตาม เป็นเพราะมีหนังมากมายถูกผลิตเพิ่มเช่นกัน
    2. หนังควรมีความยาวอยู่ในช่วง Long หรือความยาวตั้งแต่ 120 นาทีขึ้นไป เนื่องจากจะมีแนวโน้มที่จะได้ imdb_score และรายได้ที่สูง กว่าในช่วงเวลาอื่นๆ
    3. ต้นทุนในการสร้างหนังนั้นไม่จำเป็นต้องสูงมาก ก็มีโอกาสที่จะได้รับ imdb_score ที่สูงได้เช่นกันหากมีเนื้อหาที่ดี (คะแนนที่ได้ไม่ได้แปรตามต้นทุน)
    4. แนะนำให้สร้างหนังที่มีความเกี่ยวข้องกับประเภท Animation, Adventure, Family เนื่องจากได้รับความนิยมในหมู่คนดู

CHALLENGE

- ข้อมูล dataset นี้เป็นข้อมูลที่ไม่ได้มีข้อมูลครบถึงในปีปัจจุบัน จึงทำให้ไม่ได้เห็นแนวโน้มถึงในปีล่าสุด
- ข้อมูลไม่ได้มีมุมที่จะนำมาวิเคราะห์หา insight บางทีลอง plot กราฟดูแล้ว อาจจะไม่เจอ insight ที่ต้องการ
- ในการนำ insight ที่วิเคราะห์มาไปใช้จริง คิดว่าอาจจะต้องมีข้อมูลจากแหล่งอื่นๆมาร่วมวิเคราะห์เพิ่มร่วมด้วย ทั้งความชอบของคนในแต่ละประเทศ ภาษา วัฒนธรรม การเข้าถึงของหนังประเภทต่างๆ
- รวมถึงการให้คะแนนในแต่ละเรื่อง ไม่สามารถควบคุมคนให้คะแนน (ตัวแปรต้น) ให้เป็นคนเดียวกันในทุกๆเรื่อง การวิเคราะห์น่าจะแม่นยำขึ้นหากกลุ่มตัวอย่างที่ให้คะแนนเป็นคนกลุ่มเดียวกัน

ผู้จัดทำ

- Pongsapak Somsakraksanti
  6510422001
- Kanit Chankijpanich
  6510422026