# Project 1

Bộ dữ liệu có 21 cột và 10878 dòng (Không tính header).

1. csvsort -c release_date -r tmdb-movies.csv > tmdb movies_sorted_by_release_date_desc.csv
2. csvgrep -c vote_average -r '^7\.[6-9]$|^89?$|^10(\.0+)?$' tmdb-movies.csv > tmdb-movies_vote_gt_7_5.csv
3. csvsort -c revenue -r tmdb-movies.csv | head -n 2 > movie_max_revenue.csv
   csvgrep -c revenue -r '^[1-9]' tmdb-movies.csv | csvsort -c revenue | head -n 2 > movie_min_revenue.csv
4. csvcut -c revenue tmdb-movies.csv | csvstat --sum > sum_revenue.csv
5. awk -F',' ' NR==1 { # header: chèn tên cột result sau cột 16 print $1,$2,$3,$4,$5, "result" $6,$7,$8,$9,$10,$11,$12,$13,$14,$15,$16,$17,$18,$19,$20,$21 next } { result = $5 - $4 print $1,$2,$3,$4,$5, result, $6,$7,$8,$9,$10,$11,$12,$13,$14,$15,$16,$17,$18,$19,$20,$21 } ' OFS=',' tmdb-movies.csv > tmdb-movies-with-result.csv

(
head -n 1 tmdb-movies-with-result.csv
tail -n +2 tmdb-movies-with-result.csv | sort -t, -k6,6nr | head -n 10
) > top10_result_movies.csv

6. Đạo diễn nào có nhiều bộ phim nhất và diễn viên nào đóng nhiều phim nhất

**Đạo diễn có nhiều bộ film nhất:**

python3 - << 'EOF' > director_most_movies.txt
import csv
from collections import Counter

counter = Counter()

with open('tmdb-movies.csv', encoding='utf-8') as f:
reader = csv.DictReader(f)
for row in reader:
director = row.get('director')

```
if director:
counter[director] += 1

director, count = counter.most_common(1)[0]
print(count, director)
EOF
```

**Diễn viên đóng nhiều film nhất:**

```
python3 - << 'EOF' > actor_most_movies.txt
import csv
from collections import Counter

counter = Counter()

with open('tmdb-movies.csv', encoding='utf-8') as f:
reader = csv.DictReader(f)
for row in reader:
cast = row.get('cast')
if cast:
for actor in cast.split('|'):
actor = actor.strip()
if actor:
counter[actor] += 1

actor, count = counter.most_common(1)[0]
print(count, actor)
EOF
```

7. Thống kê số lượng phim theo các thể loại.

```
python3 - << 'EOF' > movies_by_genre.csv
import csv
from collections import Counter

counter = Counter()

with open('tmdb-movies.csv', encoding='utf-8') as f:
reader = csv.DictReader(f)
for row in reader:
genres = row.get('genres')
if genres:
for genre in genres.split('|'):
genre = genre.strip()
```

```
if genre:
counter[genre] += 1

print("genre,movie_count")
for genre, count in counter.most_common():
print(f"{genre},{count}")
EOF
```