DS 320 Project

Upload datasets to colab
Clean and tidy up data
Set the datasets to merge and create new dataset containing both matches
Demonstrate queries that are translated to a logical query plan (maybe 7ish?)
Execute and evaluate queries

What we need for final report:
Abstract, Background, Methods, Results, Discussions, and Conclusion

Queries: (5-6)

“Do higher budget costs always result in higher ratings?”
How does runtime affect gross?
Which directors have the greatest box office success?
Are certain genres more successful than others?
Are movies getting better reviews over time?

“Do higher budget costs always result in higher ratings?”

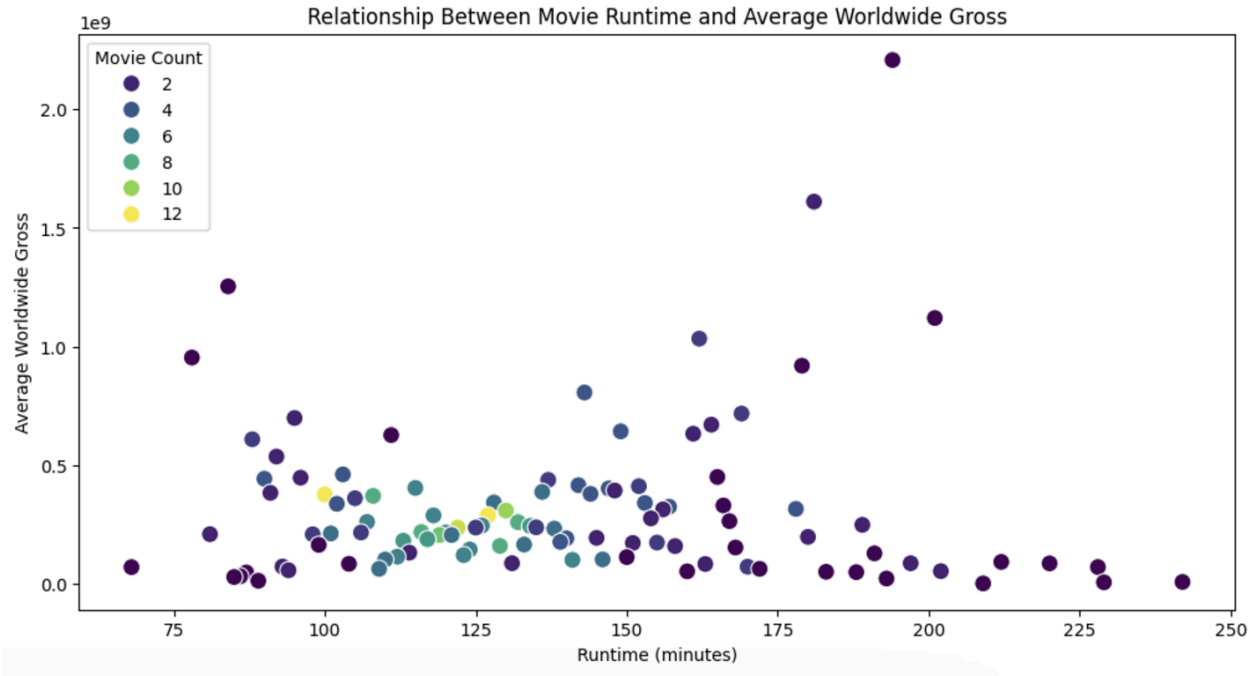
```
25
26    SELECT
27      CASE
28        WHEN ProductionBudget < 50 THEN 'Low (<50M)'
29        WHEN ProductionBudget BETWEEN 50 AND 100 THEN 'Medium (50-100M)'
30        WHEN ProductionBudget BETWEEN 101 AND 200 THEN 'High (101-200M)'
31        ELSE 'Very High (>200M)'
32      END AS BudgetRange,
33      AVG(IMDB_Rating) AS AvgRating,
34      COUNT(*) AS MovieCount
35    FROM
36      merged
37    GROUP BY
38      BudgetRange
39    ORDER BY
40      BudgetRange;
41
```

	BudgetRange	AvgRating	MovieCount
1	High (101-200M)	7.94074074074074	54
2	Low (<50M)	7.92707423580786	229
3	Medium (50-100M)	7.98552631578948	76
4	Very High (>200M)	8.02727272727273	11

How does runtime affect gross?

```
44    SELECT
45      CAST(Runtime AS INTEGER) AS Runtime,
46      CAST(AVG(WorldwideGross) AS INTEGER) AS AvgWorldwideGross,
47      COUNT(*) AS MovieCount
48    FROM
49      merged
50    GROUP BY
51      Runtime
52    ORDER BY
53      AvgWorldwideGross DESC;
54
```

	Runtime	AvgWorldwideGross	MovieCount
1	194	2207986545	1
2	181	1611004684	2
3	84	1253802559	1
4	201	1120214046	1
5	162	1033216059	3
6	78	953554418	1
7	179	919148764	1
8	143	806058906	4
9	169	717322832	3
10	95	698552819	2
11	164	670528582	2
12	149	642131557	4
13	161	632253948	2
14	111	626549695	1
15	88	608856465	3
16	92	535920997	2
17	103	460687925	4
18	165	449841566	1
19	96	446869295	2


Relationship Between Movie Runtime and Average Worldwide Gross
1e9
Movie Count
2
4
6
8
10
12
Average Worldwide Gross
0.0
0.5
1.0
1.5
2.0
75
100
125
150
175
200
225
250
Runtime (minutes)

Which directors have the greatest box office success?

```
56
57      SELECT
58        Director,
59        CAST(AVG(WorldwideGross) AS INTEGER) AS AvgWorldwideGross,
60        COUNT(*) AS MovieCount
61      FROM
62        merged
63      GROUP BY
64        Director
65      ORDER BY
66        AvgWorldwideGross DESC;
67
```

	Director	AvgWorldwideGross	MovieCount
1	Anthony Russo	1677165374	4
2	Roger Allers	1654367425	1
3	James Cameron	1438144433	4
4	Gary Trousdale	1253802559	1
5	Josh Cooley	1073080329	1
6	Byron Howard	1004629935	1
7	Peter Jackson	980975323	5
8	Wolfgang Reitherman	953554418	1
9	Andrew Stanton	936094852	1
10	Lee Unkrich	933272973	2
11	David Yates	929411069	1
12	Ron Clements	840807633	2
13	Bryan Singer	817451692	2
14	David Leitch	786362370	1
15	Tim Miller	784157111	1
16	Joss Whedon	777709825	2
17	James Gunn	770882395	1
18	Todd Phillips	768997550	2
19	Pete Docter	715196054	3

Are certain genres more successful than others?

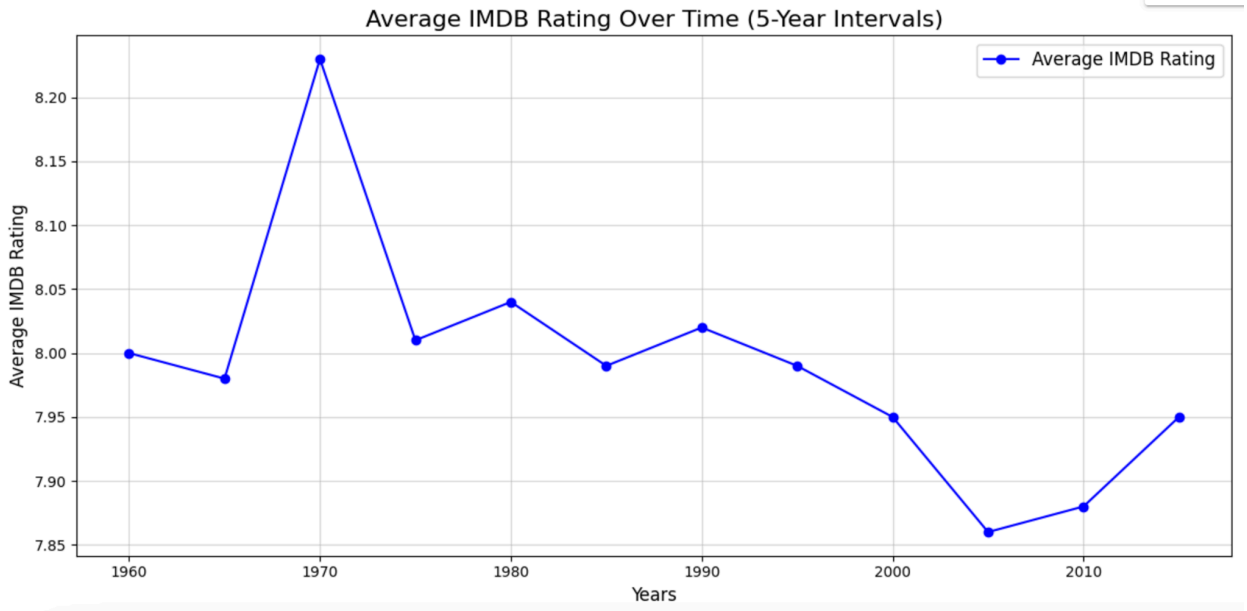
```sql
-- Are certain genres more successful than others?
SELECT
  Genre,
  CAST(AVG(WorldwideGross) AS INTEGER) AS AvgWorldwideGross,
  COUNT(*) AS MovieCount
FROM
  merged
GROUP BY
  Genre
--removed genres with total movies less than 10
HAVING
  COUNT(*) >= 10
ORDER BY
  AvgWorldwideGross DESC;
```

	Genre	AvgWorldwideGross	MovieCount
1	Animation	601992602	36
2	Action	455324738	89
3	Adventure	302942595	36
4	Drama	187550416	82
5	Comedy	157341870	29
6	Crime	152237026	35
7	Biography	141899237	54

Are movies getting better reviews over time?

```sql
--Grouped the movies into 5-year intevals
SELECT
  (Released_Year / 5) * 5 AS YearStart,
  ((Released_Year / 5) * 5) + 4 AS YearEnd,
  ROUND(AVG(IMDB_Rating), 2) AS AvgIMDBRating,
  COUNT(*) AS MovieCount
FROM
  merged
WHERE
  Released_Year IS NOT NULL
  AND Released_Year >= 1920
GROUP BY
  (Released_Year / 5) * 5
--removed the rows with moviecount less than 3
HAVING
  MovieCount >= 3
ORDER BY
  YearStart;
```

	YearStart	YearEnd	AvgIMDBRating	MovieCount
1	1960	1964	8.0	6
2	1965	1969	7.98	4
3	1970	1974	8.23	4
4	1975	1979	8.01	7
5	1980	1984	8.04	18
6	1985	1989	7.99	16
7	1990	1994	8.02	33
8	1995	1999	7.99	43
9	2000	2004	7.95	48
10	2005	2009	7.86	69
11	2010	2014	7.88	64
12	2015	2019	7.95	51


Average IMDB Rating Over Time (5-Year Intervals)
Average IMDB Rating
Average IMDB Rating
8.20
8.15
8.10
8.05
8.00
7.95
7.90
7.85
1960
1970
1980
1990
2000
2010
Years