DASC-5300

Foundation of Computing

Project 3: IMDB Data Analysis

Submitted by: Team 52

Hevin Patel - - 1002036919

Sakshi Nischal - 1002085832

Overview:

In this project we were given a large database, filled with millions of rows of data about international films and TV shows. It is utilized by researchers in databases and is referred to as the IMDb database by the community (publicly accessible data set, although not as a relational DBMS). This contains all information about US and foreign movies and TV shows starting in 1925 and going up to 2018. We used SQL to query this database to look up specific information of interest, obtain aggregate and statistical data that interests you, and, to the extent possible, perform OLAP analysis queries. This project gave us a deep understanding of to analyze the data which is already stored in a DBMS, and also how to overcome challenges related to SQL.

File Description:

File Name	File Description
analysis_1a	Sql File contains the query for the analysis and text file contains the output results for the query
analysis_1a_career_span	Sql File contains the query for the analysis and text file contains the output results for the query
analysis_1b	Sql File contains the query for the analysis and text file contains the output results for the query
analysis_2_max_per_year	Sql File contains the query for the analysis and text file contains the output results for the query
analysis_2_max_ten_year	Sql File contains the query for the analysis and text file contains the output results for the query
analysis_2_min_per_year	Sql File contains the query for the analysis and text file contains the output results for the query
analysis_2_min_ten_year	Sql File contains the query for the analysis and text file contains the output results for the query
analysis1a_visualizations	This file contains all the visualizations for this analysis
analysis1b_visualizations	This file contains all the visualizations for this analysis
analysis1c_visualizations	This file contains all the visualizations for this analysis

Division of Labor:

We equally contributed to the successful completion of the project.

Hevin Patel (1002036919) contributed to designing the SQL query for performing Analysis 1b. Also, he performed the analysis 1a and created compelling visualizations to understand the trends. He significantly helped in making the report.

Sakshi Nischal (1002085832) contributed to designing the SQL query for performing Analysis 2. Also, she performed the analysis 1c and created compelling visualizations to understand the trends. She significantly helped in making the report.

Problems Encountered:

While doing the analysis in IMDB database, we faced various challenges. First and foremost, we were not able to divide the total career span into different disjoint sets. But after learning more about the problem and referring online resources, we fixed this issue using case statement in SQL.
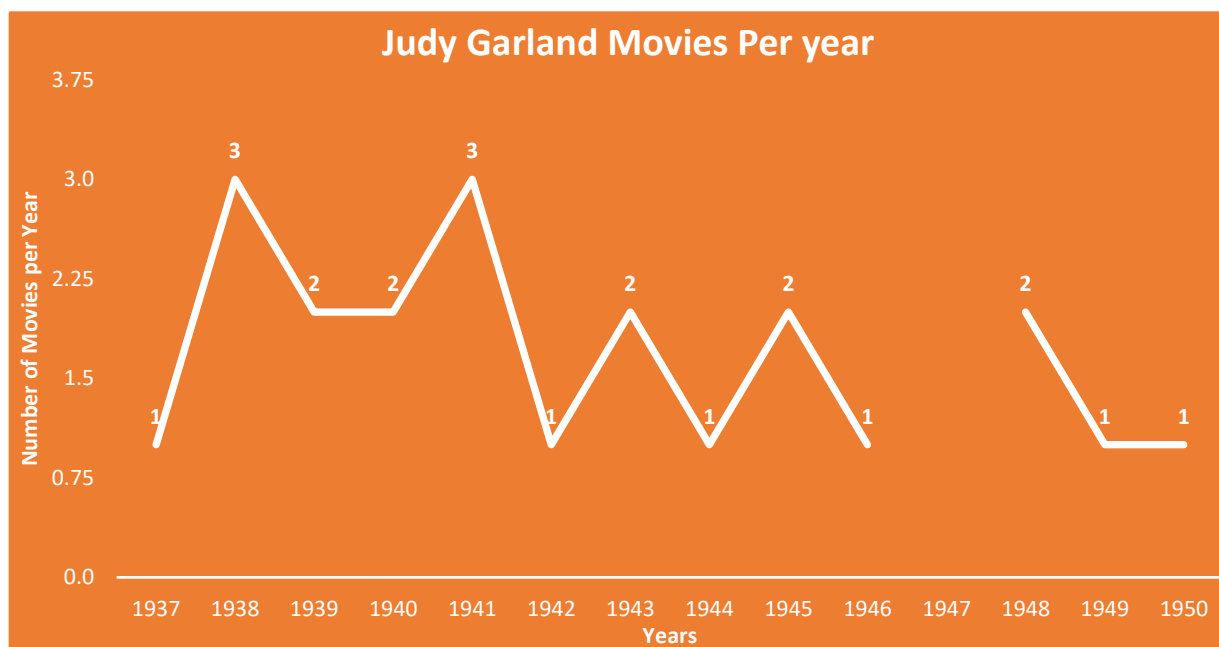
Secondly, we encountered that we were not getting results for female celebrities. But after looking into the SQL queries, we noticed that we didn't add the actress category to IN clause. We overcame this problem by rectifying our query.

Also, we got stuck in analysis 2. In this analysis we were not getting all the rows for the actors who have done at least 3 movies in their entire career. We resolved this problem by using fetch rows clause with ties.
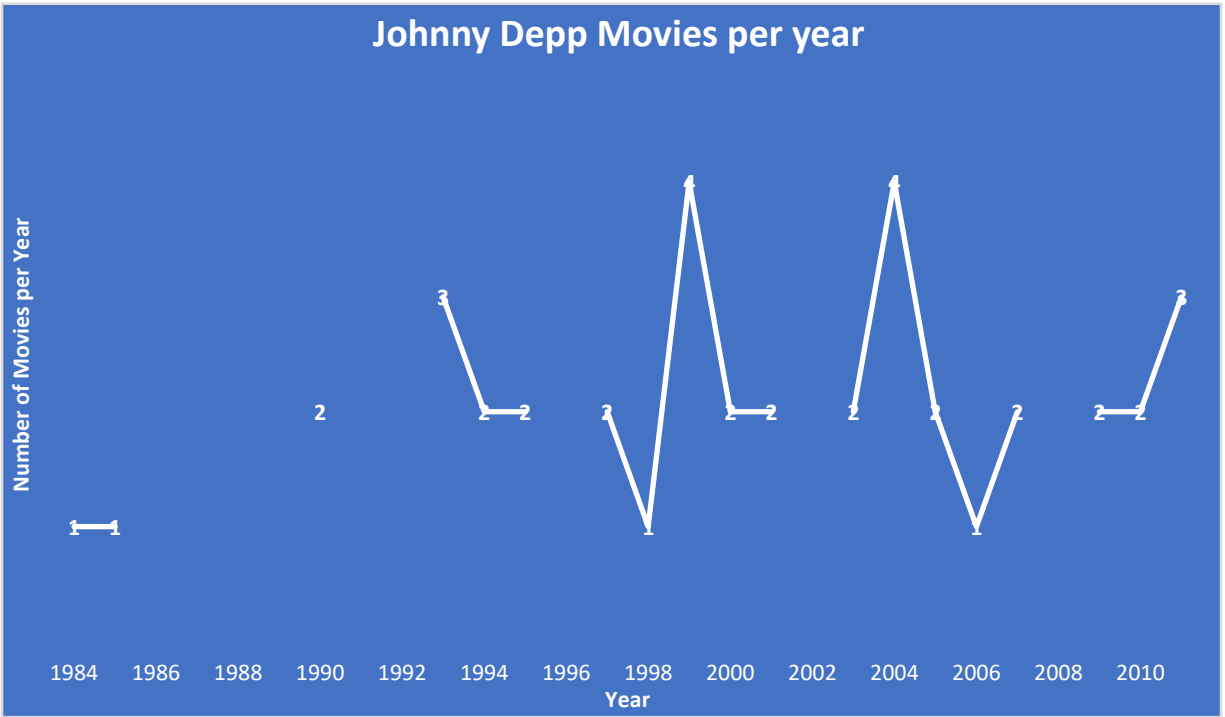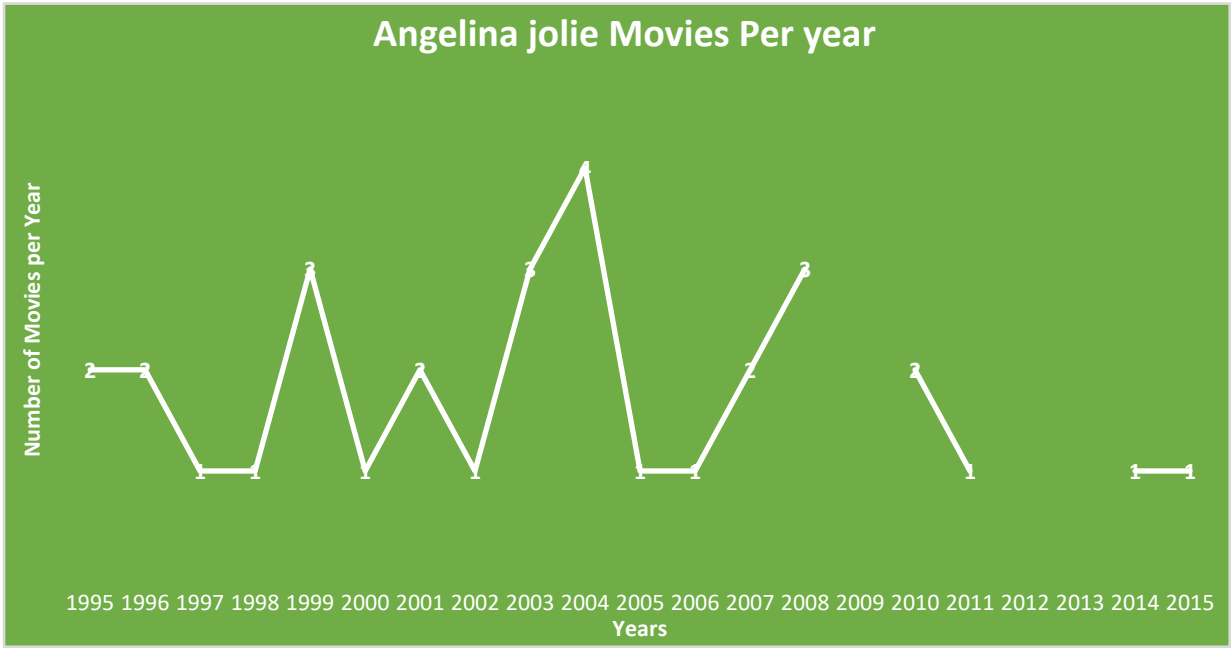
Analysis and Visualizations:

Analysis One A:

Number of movies done by Judy Garland in each year throughout her career span
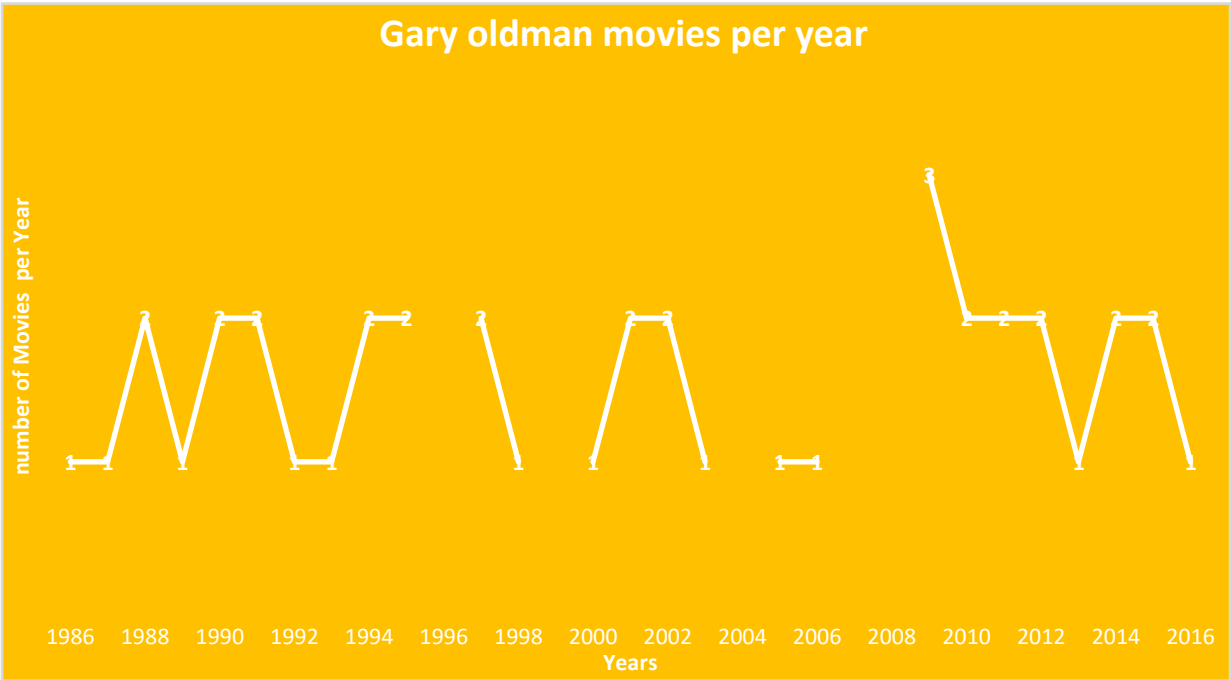
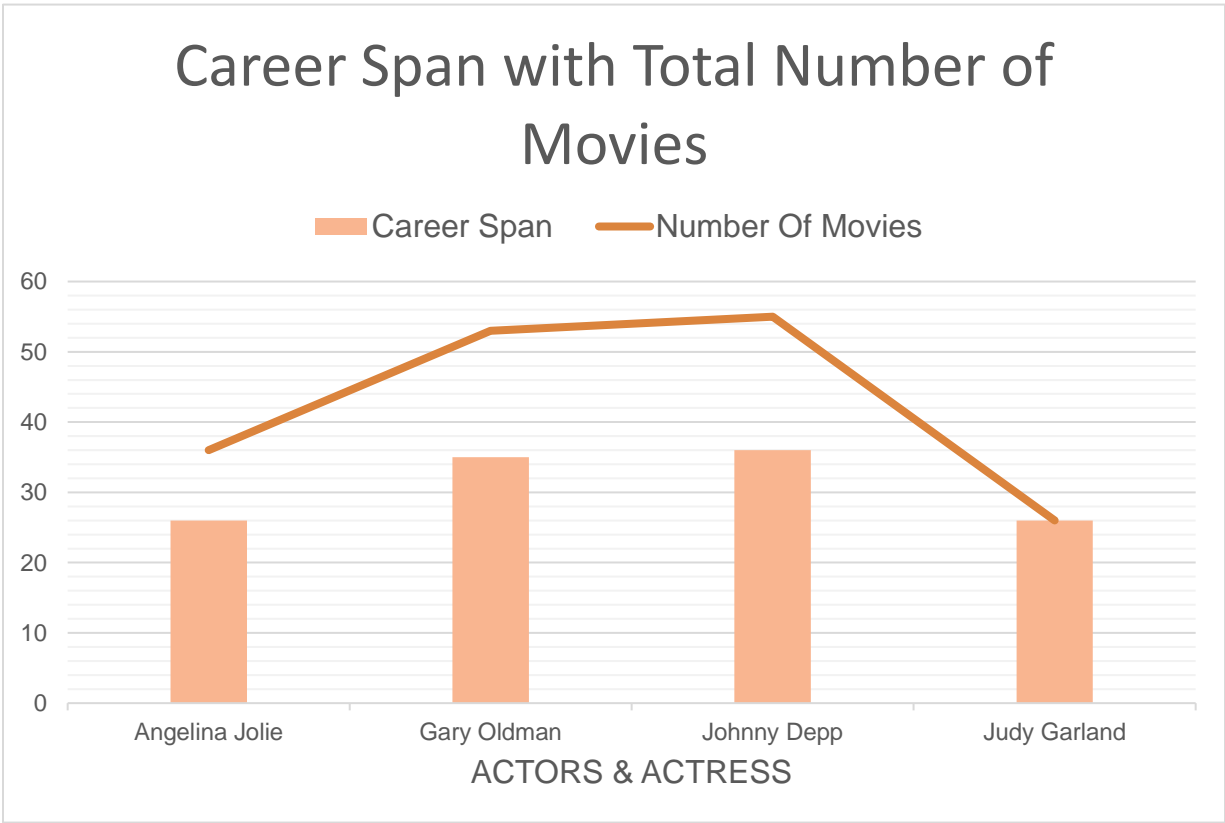**Number of movies done by Johnny Depp in each year throughout her career span**

Johnny Depp Movies per year

Number of Movies per Year

Year

1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010

Number of movies done by Angelina Jolie in each year throughout her career span

Angelina jolie Movies Per year

Number of Movies per Year

Years

1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

Number of movies done by Gary Oldman in each year throughout her career span

Gary oldman movies per year

Total Career Span with Total Number of movies by each actor
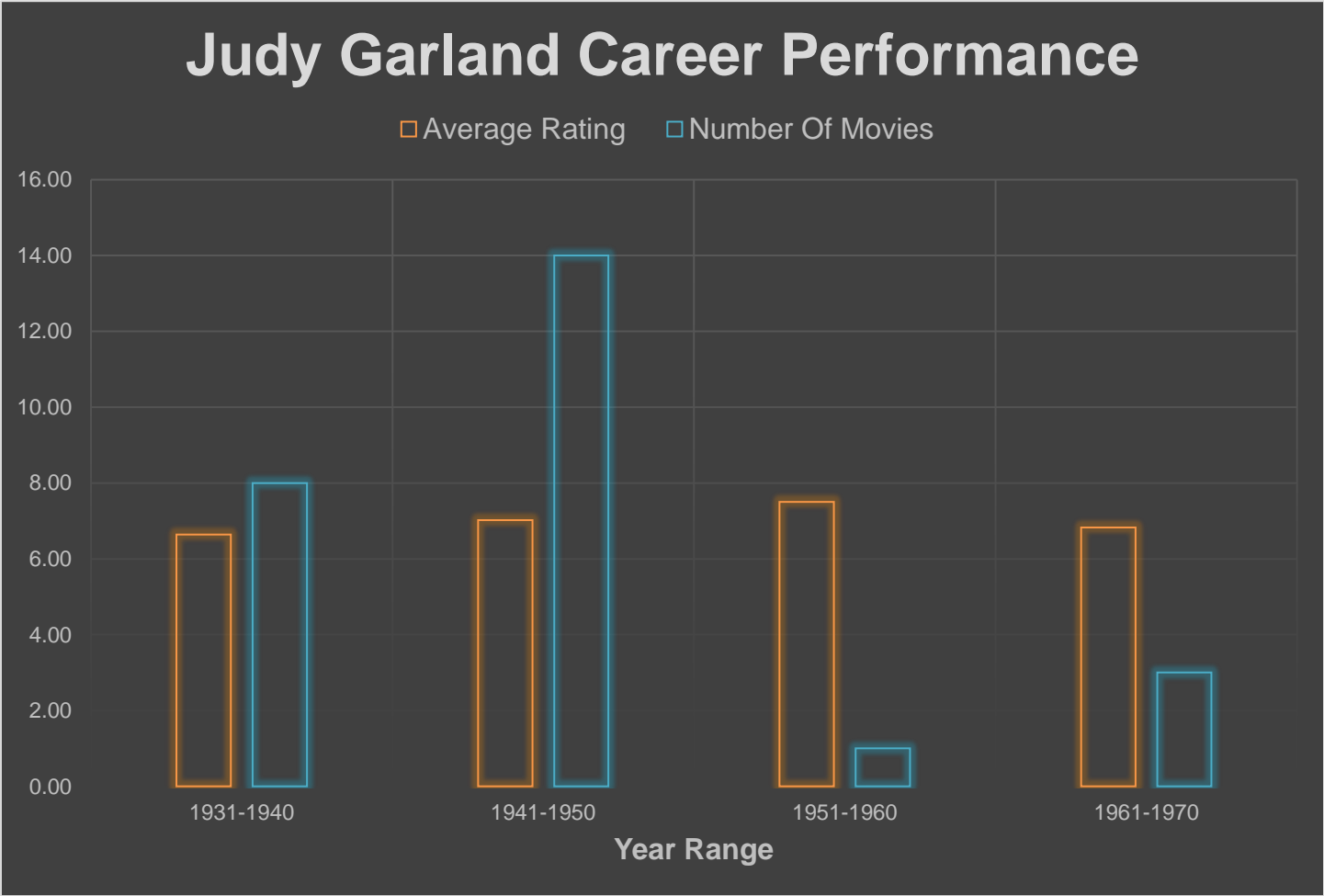

Career Span with Total Number of Movies

The above comparison chart depicts the comparison between the four actors over their career span and number of movies. It can be easily seen from the graph that Johnny Depp did the greatest number of movies as compared to the other 3 actors. He acted in 55 movies throughout his career span of 36 years. He did his first movie in the year 1984 and was seen last in 2020.

On the flip side, Judy Garland started her movie career in 1937 and ended in 1963. During 26 years of her career span, she was a part of 26 movies in total which was lowest among other three.
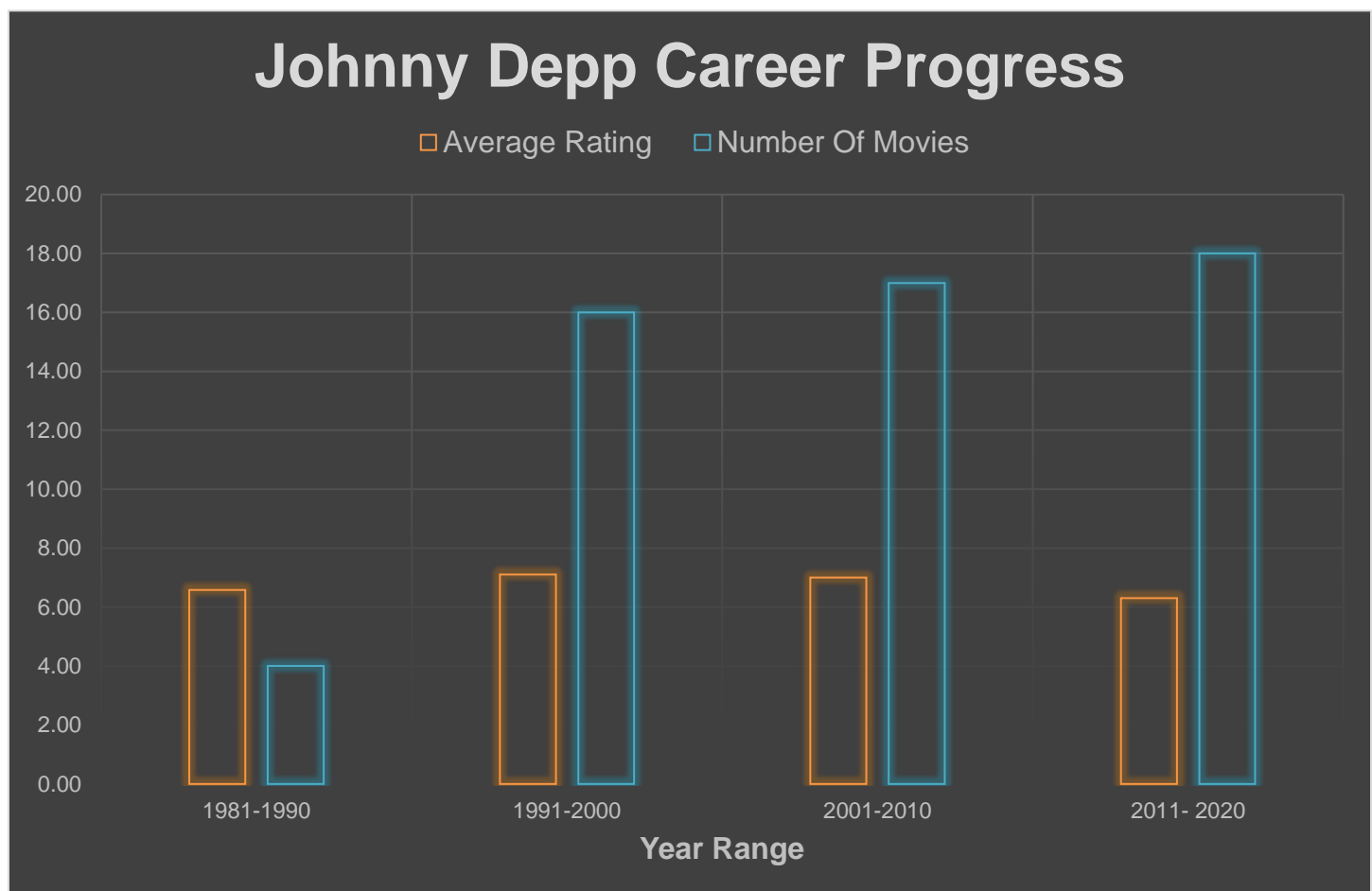
Also, it was observed that Gary Oldman shares a similar journey like Johnny Depp. He did 53 movies which is second highest among all other actors. The career span is also almost same as Johnny Depp which is 35. Surprisingly, he started his movie career exactly two years after Johnny Depp was first seen on the big screens.

Analysis One B:

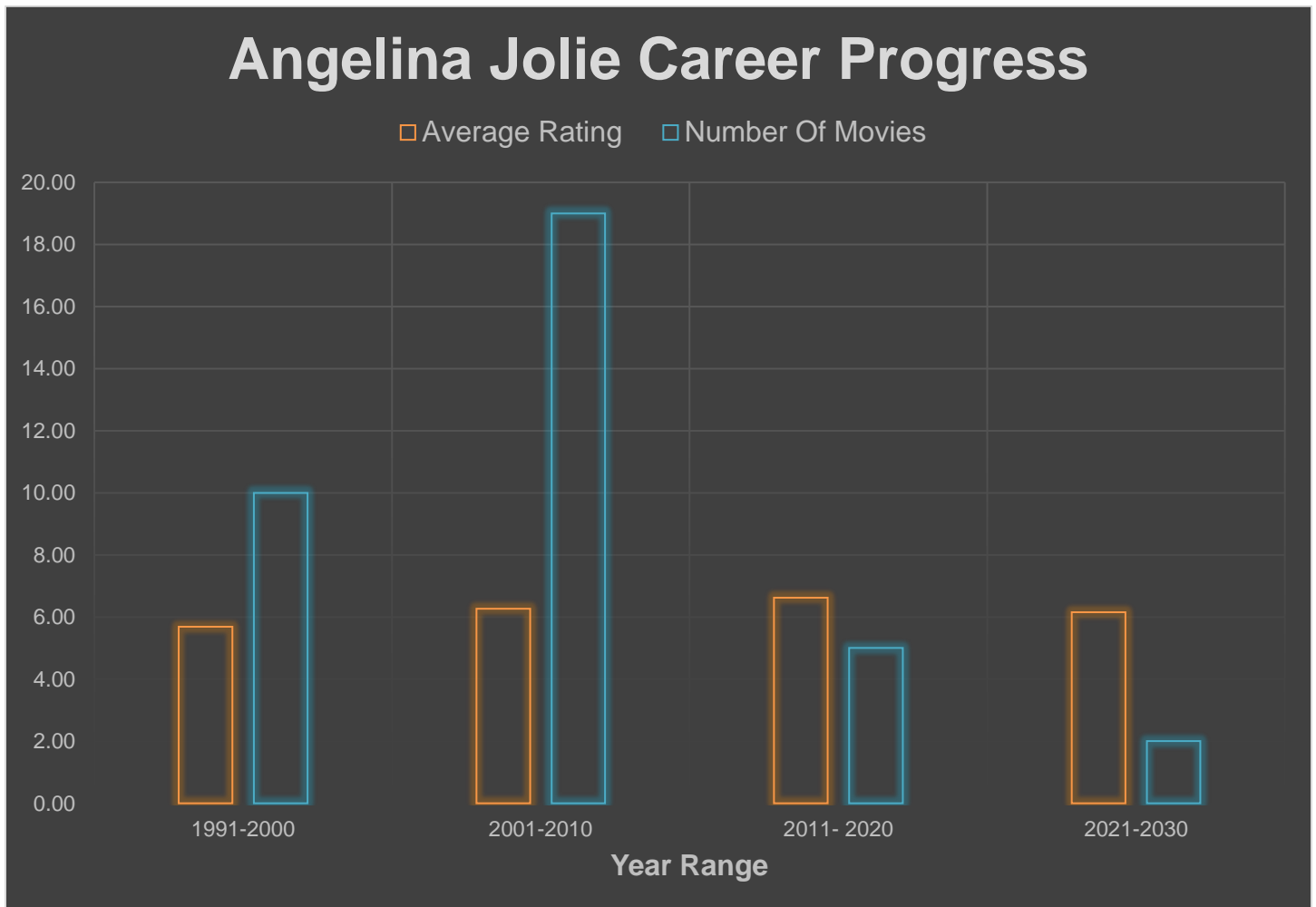**Total Number of movies done by each actor over 4-5 disjoint periods**

Judy Garland did her first movie at the age of 13. She started her career in 1931 with a 6.64 average rating of her movies. She did only 8 movies in her first decade. After her first movie she did very well in Hollywood. Her average rating of movies increased gradually from 1931 to 1951. She touched the peak from 1951 to 1960 and achieved 7.50 rating with 14 movies count which was the highest average rating she achieved in her whole movie industry. It can be stated that the period from 1951-1960 was the worst decade for her as she only made one movie, after that MGM dropped her from her contract, so there was a sudden drop in the cinema industry of her career. In 1960's she worked more as a singer than an actress, so we can see in the graph that there is a downfall in her movie career. She ended her career with 6.83 rating in 1969.

In 1984, Johnny Depp made his film debut in the horror film *A Nightmare on Elm Street* with 6.38 average rating. He did only 4 movies in his first decade. It was observed from the graph that the decades from 1991 to 2000 and 2001 to 2010 were the best decades of his life. He got famous in this time period by playing crucial roles in some amazing movies. During this time, he experienced a rapid rise as a professional actor. Pirates of the Caribbean is one of the top
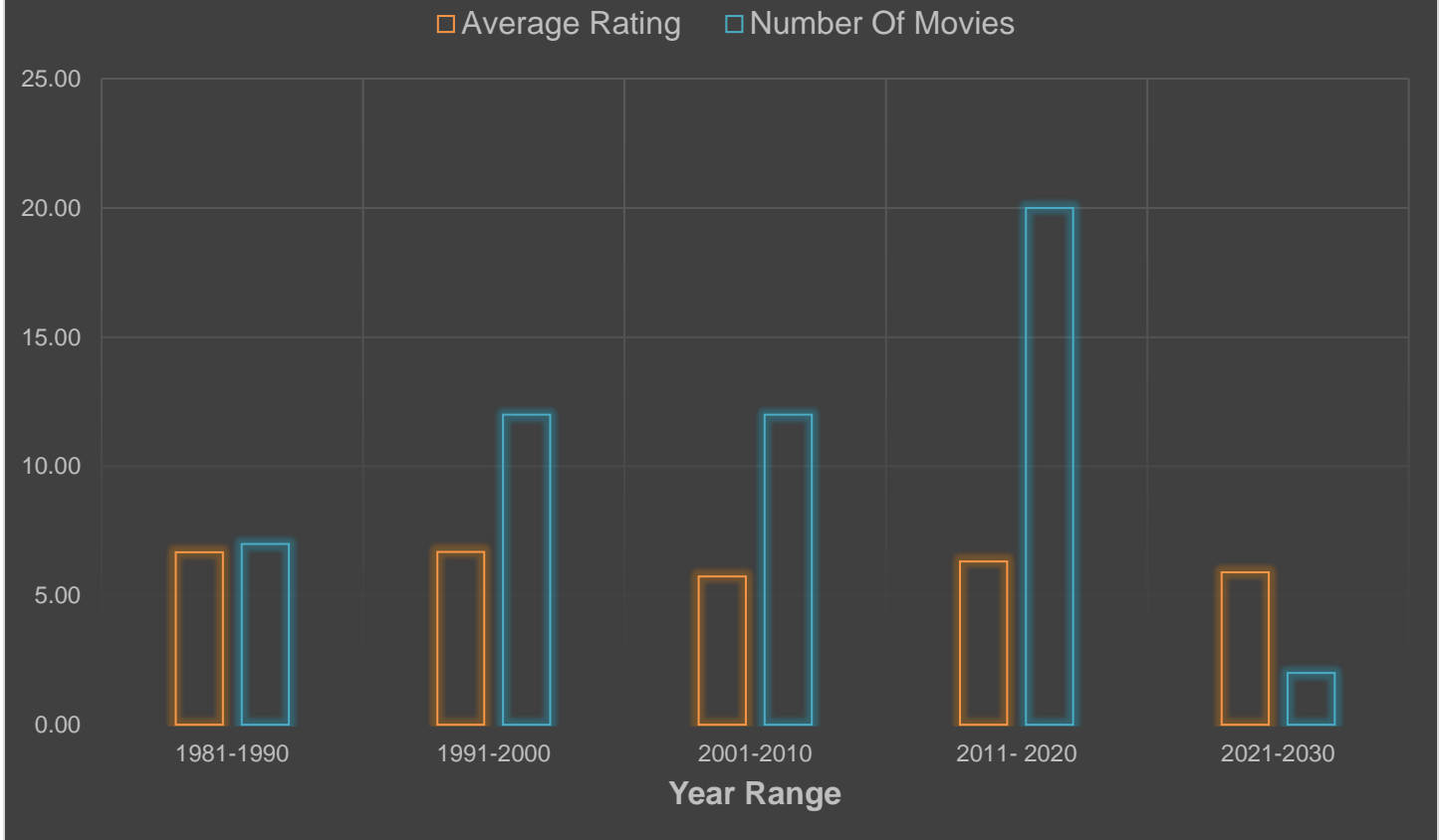
ranked movies in his career span. He touched the peak of his career with an average rating of 7.10 in 1991-2000 with 16 movies. After that we can see in the graph that there is a sudden drop in his movie career. He made a total of 18 movies in 2011-2022 and an average range was 6.32 in this decade. Moreover, it was noticed that in the decade 2011-2020 he did the maximum number of movies. However, his movies didn't perform well on the chart resulting in lower average ratings.

Angelina Jolie Career Progress

Angelina Jolie made her screen debut in Cyborg 2 (1993) from the above graph, it is noticeable that she started a very decent career in movie industry, in the first decade she did a total of 10 movies and maintained an average movie rating of 5.69, which was the lowest of her career. As we can see in the graph, after her slow start in movie industry, in 2001-2010, she made total 19 movies and archived 6.27 average movie rating. It is evident from the graph that there is a sudden drop in movie count after her peak decade of movie career which is because she was more interested in drama than movie. This can be a possible reason behind why she did only 5 movies in the decade 2011-2020.
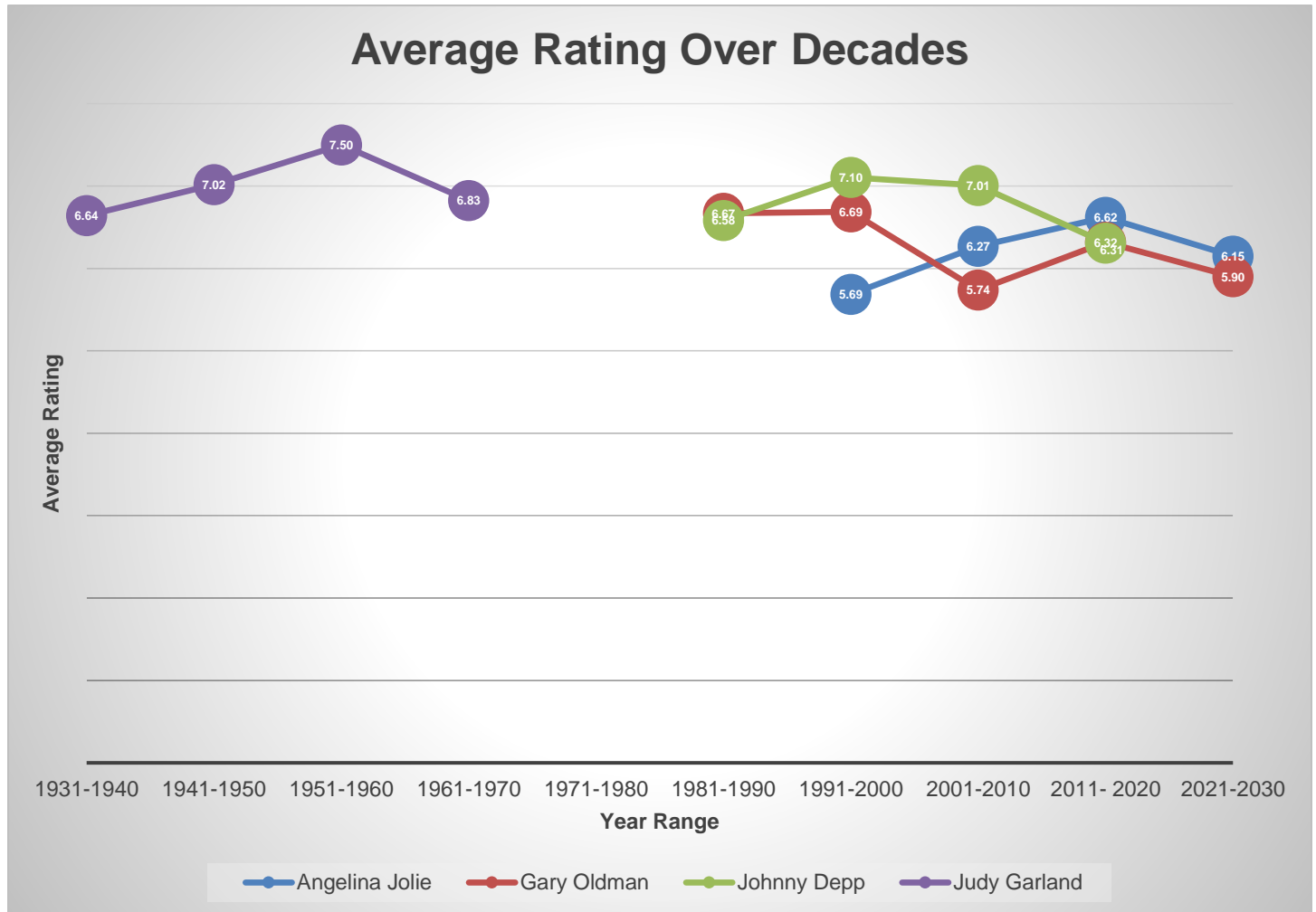
Gary Oldman Career Progress

Year Range

Gary Oldman started his career in 1981 and his debut film was Remembrance (1982). In his first decade (1981-1990) he did 7 movies and maintained an average movie rating of 6.67. After his first decade, a sudden increase in the movie count was noticed since he was nominated as a Best Actor from BAFTA for his portrayal of 60s playwright Joe Orton in Prick Up Your Ears (1987). It is evident from the graph that despite the increase in the number of movies, he proved his acting skills by maintaining a 6.69 average rating. In 2001-2010 time period, he was facing a difficult time in his career. He acted in a total of 12 movies in that particular time period, but some movies didn't perform well on the charts. Due to which a declining trend was noticed in the average rating of movies.Tiptoes (2002) and Sin (2003) didn't perform well on big screen during that decade. However, he got the chance to perform in a greater number of movies because of his commendable performance in the movie The Dark Knight (2008). He did 20 movies in his decade from 2011 to 2020 which was the best decade in terms of movie count. From the year 2011 to 2020, he came out to be a versatile actor and he was able to maintain an average movie rating of 6.32 with a movie count of 20.
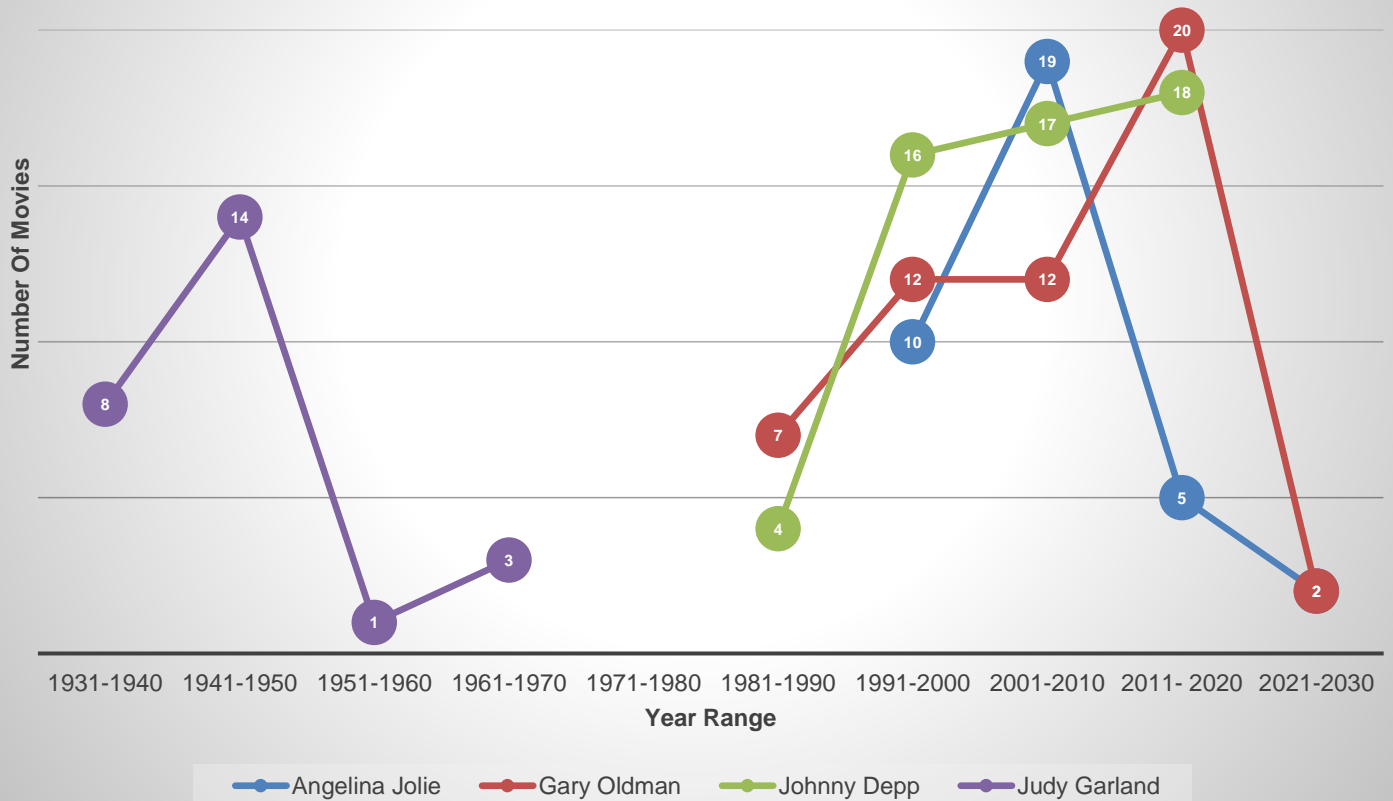
Analysis One C:

Comparison between distinct actors over their acting career life

Average Rating Over Decades

The above line graph depicts the average movie ratings of four actors in different decades. It is evident from the graph that Judy Garland was the only actress who was the part of movie industry. All other actors stepped into the industry after a significant time gap. Also, it is evident from the graph that Gary Oldman and Johnny Depp started acting almost at the same time. Moreover, not much difference was noticed in the average ratings of the movies acted by these actors. It is safe to assume that more people were interested in watching Johnny Depp on screen as he has comparatively better ratings than other actors. On the other hand, Gary Oldman experienced less movie ratings as compared to other actors. Johnny Depp didn't act in any movie in decade 2021-2030. However, Gary Oldman and Angelina Jolie gave some movies between 2021 and 2030.

Number Of Movies Over Decades

The above line chart illustrates the number of movies done by four actors over the last few decades. Judy Garland's career ended much before other actors stepped into the industry. She made the maximum number of movies between the years 1941 and 1950. It is evident that Gary Oldman and Johnny Depp started doing cinema almost at the same time. However, Gary Oldman acted in more movies as compared to Johnny Depp. Also, Johnny Depp didn't act in any movie from 2021 to 2030. Therefore, it is safe to assume that due to personal issues in his life he stopped acting in movies. Angelina Jolie started her career much later compared to other actors. However, she got many more opportunities from the year 2001 to 2010.

Analysis Two:

Actor & Actress who did maximum movies in a particular year over the span of 10 years
Actor Id: nm0477107
Actor Name: Tak-Hing Kwan
Year with maximum number of movies: 1956
Number of Movies: 30
Category: Actor

Actor Id: nm0638150
Actor Name: Muhterem Nur
Year with maximum number of movies: 1962
Number of Movies: 18
Category: Actress

Actor & Actress who did maximum movies in span of 10 years

Actor Id: nm0854251
Actor Name: Ahmet Tarik Tekçe
Number of Movies: 119
Category: Actor

Actor Id: nm0374974
Actor Name: Helen
Number of Movies: 97
Category: Actress

Actor & Actress who did minimum movies in a particular year over the span of 10 years
Approximately 4200 actors acted in at least 3 movies in a particular year over the span of 10 years. We are listing few of them

Actor Id: nm0116515
Actor Name: Nando Bruno
Year with minimum number of movies: 1953
Number of Movies: 3
Category: Actor

Actor Id: nm0253778
Actor Name: Evangelina Elizondo
Year with minimum number of movies: 1953
Number of Movies: 3
Category: Actress

Actor & Actress who did minimum movies in span of 10 years
Approximately 1900 actors acted in at least 3 movies over span of 10 years. We are listing few of them

Actor Id: nm0013638
Actor Name: Manuel Aguilera
Number of Movies: 3
Category: Actor

Actor Id: nm0054263
Actor Name: Mariemma Bardi
Number of Movies: 3
Category: Actress

References:

https://en.wikipedia.org/wiki/Judy_Garland

https://en.wikipedia.org/wiki/Johnny_Depp

https://en.wikipedia.org/wiki/Angelina_Jolie

https://en.wikipedia.org/wiki/Gary_Oldman

https://www.insider.com/johnny-depp-amber-heard-relationship-timeline-2020-7

https://www.w3schools.com/sql/