Report On:

**Project III: Disk-based Data Analysis**

**Team 18**

**Fall 2022**



Report By:

Nikhil Das Karavatt

1002085391

Table of Contents

[Overview: 3](#_Toc121227906)

[File Descriptions: 3](#_Toc121227907)

[Division of Labour: 4](#_Toc121227908)

[Problems Encountered: 4](#_Toc121227909)

[Analysis: 4](#_Toc121227910)

[1. Analysis 1.a 4](#_Toc121227911)

[2. Analysis 1.b 6](#_Toc121227912)

[3. Analysis 1.c 9](#_Toc121227913)

[4. Analysis 2 10](#_Toc121227914)

# Overview:

The purpose of this project to provide us with an understanding of analysing data stored in a DBMS. This is a large real-world DBMS and populated it with millions of rows of international movies and TV episodes information. It is known as the IMDb database by the community. We have used SQL to write analysis queries and get some meaningful information of interest and also finding aggregate and statistical information that will help with my analysis. This has led to sharpen my skills in SQL.

# File Descriptions:

There are in total 8 files in which first six are the main files which has helped me with data visualization as well as with analysis.

|  |  |
| --- | --- |
| File Name | File Description |
| Analysis1a.sql | This is a sql file for analysis 1a which contains queries for finding the count of movies done by an actor/actress in a year and also span of their acting career. |
| Analysis1a.txt | This is a spooled file for analysis 1a in a .txt format which not only contains queries for finding the count of movies done by an actor/actress per year in their acting career and span of their acting career but also the outputs of these queries. |
| Analysis1b.sql | This is a sql file for analysis 1b which contains queries for finding the count of movies done by an actor/actress per decade in their acting career and also the average performance of their movies in those decades. |
| Analysis1b.txt | This is a spooled file for analysis 1b in a .txt format which not only contains queries for finding the count of movies done by an actor/actress per decade in their acting career and the average performance of their movies in those decades but also the outputs of these queries. |
| Analysis2.sql | This is a sql file for analysis 2 which contains queries for finding the actors and actresses who have done the greatest and the least number of movies (3) from 1964 to 1974. It also contains queries for finding the same counts but per year wise in the same range. |
| Analysis2.txt | This is a spooled file for analysis 2 in a .txt format which contains queries and output for finding the actors and actresses who have done the greatest and the least number of movies (3) from 1964 to 1974. It also contains query and output for finding the same counts but per year wise in the same range. |
| Movies\_count.xlsx | This is an output excel file of Analysis 1b which has helped with visualization of the count of movies done by an actor/actress per decade in their acting career. |
| Movies\_average\_ratings.xlsx | This is an output excel file of Analysis 1b which has helped with visualization of average performance of the actors’ and actresses’ movies per decade in their career |
| Graph.xlsx | This excel file contains most of the graphs and tables that I have attached in this report. |

# Division of Labour:

The project was done solely by me and it took me around 50 hours of work to finish it.

# Problems Encountered:

1. To perform analysis 1b was hideous at the start because of multiple sql queries and creation of temporary tables but all this became a piece of cake when I understood the concept of case syntax which works like an If-else condition. There by, reducing my number of queries from 20 to 2 and without any temporary table.
2. Earlier in my queries in where condition I was using lower(tb.titletype) = ‘movie’ but I realised that these wasn’t giving me the desired results. As suggested by professor I used ‘like’ operator along with % which was giving a better output.
3. In analysis 2, I am getting a lot of count for actors and actresses who did minimum 3 movies in the year range 1964 to 1973 as well as the same count per year in the same year range> After having a word the teaching assistant Kiran I printed only few rows for these.

# Analysis:

## Analysis 1.a

From the above graph we can see that Katharine Hepburn and Brad Pitt have the same count of movies in lifespan till date. As Katharine has died at 2004, her count will be over taken by all the other three actors and actress unless they retire from acting industry or other unforeseen circumstances takes place. We can also see that Kate Winslet has a good count in movies even though she started significantly late (1994) when compared to other actors and actresses. At the moment percentage we can see there is no drastic difference among one another.

We can see that the Katharine Hepburn has the longest span in acting career with a glorious number of 62 but her count in movies is just 51 which is same as Brad Pitt’s count even thought his span in acting career is almost half of Katharine’s (34). As mentioned earlier Kate Winslet has the smallest span yet as she started her acting career when compared to others. Tom Cruise is the second senior most actor from the below list.

|  |  |  |  |
| --- | --- | --- | --- |
| NAME | Latest Movie Year | First Acted Movie Year | Total Span of Acting Career in Years |
| Brad Pitt | 2022 | 1988 | 34 |
| Tom Cruise | 2024 | 1982 | 42 |
| Katharine Hepburn | 1994 | 1932 | 62 |
| Kate Winslet | 2023 | 1994 | 29 |

From the above graph we could conclude that Katharine even though she has the longest span in acting career but she even has a lot of gaps in years in between that span with no movies unlike the remaining three actors and actress. Brad Pitt and Kate Winslet are the only actor and actress who have done 4 movies in a year. We could also conclude that the mode for all the actors and actresses is 1 movie a year.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Brad Pitt** | | **Kate Winslet** | | **Katharine Hepburn** | | **Tom Cruise** | |
| YEAR | Yearly Movie Count | YEAR | Yearly Movie Count | YEAR | Yearly Movie Count | YEAR | Yearly Movie Count |
| 1988 | 1 | 1994 | 1 | 1932 | 1 | 1982 | 1 |
| 1989 | 1 | 1995 | 1 | 1933 | 3 | 1983 | 2 |
| 1990 | 2 | 1996 | 2 | 1934 | 2 | 1985 | 1 |
| 1991 | 1 | 1997 | 1 | 1935 | 3 | 1986 | 2 |
| 1992 | 2 | 1998 | 1 | 1936 | 2 | 1988 | 2 |
| 1993 | 1 | 1999 | 2 | 1937 | 2 | 1989 | 1 |
| 1994 | 3 | 2000 | 1 | 1938 | 2 | 1990 | 1 |
| 1995 | 2 | 2001 | 3 | 1940 | 1 | 1992 | 3 |
| 1996 | 1 | 2003 | 3 | 1942 | 2 | 1993 | 1 |
| 1997 | 2 | 2004 | 3 | 1944 | 1 | 1994 | 1 |
| 1998 | 1 | 2005 | 1 | 1945 | 1 | 1996 | 2 |
| 1999 | 1 | 2006 | 4 | 1946 | 1 | 1999 | 1 |
| 2000 | 2 | 2008 | 2 | 1947 | 2 | 2000 | 1 |
| 2001 | 4 | 2011 | 2 | 1948 | 1 | 2001 | 1 |
| 2003 | 1 | 2013 | 1 | 1949 | 1 | 2002 | 2 |
| 2004 | 2 | 2014 | 2 | 1951 | 1 | 2003 | 1 |
| 2005 | 1 | 2015 | 3 | 1952 | 1 | 2004 | 1 |
| 2006 | 1 | 2016 | 1 | 1955 | 1 | 2005 | 1 |
| 2007 | 2 | 2017 | 4 | 1956 | 2 | 2006 | 1 |
| 2008 | 2 | 2018 | 1 | 1957 | 1 | 2007 | 1 |
| 2009 | 1 | 2019 | 2 | 1959 | 1 | 2008 | 1 |
| 2010 | 1 | 2020 | 2 | 1962 | 1 | 2009 | 1 |
| 2011 | 2 | 2021 | 1 | 1967 | 1 | 2010 | 1 |
| 2012 | 1 | 2022 | 2 | 1968 | 1 | 2011 | 1 |
| 2013 | 2 | 2023 | 1 | 1969 | 1 | 2012 | 2 |
| 2014 | 1 |  |  | 1971 | 1 | 2013 | 1 |
| 2015 | 3 |  |  | 1973 | 2 | 2014 | 1 |
| 2016 | 2 |  |  | 1975 | 2 | 2015 | 1 |
| 2017 | 1 |  |  | 1978 | 1 | 2016 | 1 |
| 2019 | 2 |  |  | 1979 | 1 | 2017 | 2 |
| 2022 | 2 |  |  | 1981 | 1 | 2018 | 1 |
|  |  |  |  | 1984 | 1 | 2022 | 1 |
|  |  |  |  | 1986 | 1 | 2023 | 1 |
|  |  |  |  | 1988 | 1 | 2024 | 1 |
|  |  |  |  | 1992 | 1 |  |  |
|  |  |  |  | 1994 | 3 |  |  |

## Analysis 1.b

Tom Cruise started his acting career in movies in 1982 at the age of 20. From the blue line in the above graph, we can see that in a decade he has been consistently doing around 10 movies with count of 11 being maximum in 2001–2010-year range. The count in the last/present decade is low as only 1/5th of the decade has only been passed and he already has a count of 3 movies. If there is no unforeseen circumstance that may hinder his acting career there is a high chance that he will be maintaining the count of around 10 movies in a decade.

The orange line in the graph indicates the average rating of his movies in that particular decade. We can see his worst average movie rating was during his start of the career but then after a good progress, it has been consistently around 7 henceforth. Though, the current decade shows the best rating of his career but we can’t take this into serious consideration as the average rating of the movies in 2021-2030 decade is just of 1 out of 3 movies as others are yet to be released or rated.

At the age of 25, Brad Pitt started his acting career in movies in 1988 which gives us the reason of why he had a low count of movies in 1981–1990-year range. From the blue line in the above graph, we can see that in a decade he has been consistently doing around 15 movies with count of 16 being maximum in 1991–2000-year range. There has been a very slight decline in count over decades, if the same trend follows there is a high chance of his movie count for the current decade will be less than previous decade which is 14. The count in the last/present decade is low as only 1/5th of the decade has only been passed and he has a count of 2 movies which is already strengthening my assumptions of decline trend of his count of movies in a decade.

The orange line in the graph indicates the average rating of his movies in that particular decade. We can see his worst average movie rating was during his start of the career. But then he had a steep rise when it comes to his movie’s ratings from a 5.2 to 7.1 and thereafter has been consistently around 7 with 7.23 rating being his best in 2001–2010-year range. The current decade shows no rating as the movies are yet to be released or rated.

Kathrine Hepburn started her acting career in movies in 1932 at the age of 25. From the blue line in the above graph, we can see that she did the greatest number of movies in the starting decade of her career with a count of 16. After which she had a steady decline in the count of movies per decade till 1970. In 1971-1980 she almost did double the count of movies 7, when compared to her surrounding decades. Eventually ending her career with 4 counts of movies in 1991-2000.

The orange line in the graph indicates the average rating of her movies in that particular decade. We can see that her average movie’s rating kept gradually increasing till 1970. Which is exactly the opposite trend line of her count of movies. So, by this we can infer that though in her career until 1970 the movie’s count per decade was decreasing but her the average performance of her movies were improving, rather to strengthen my hypothesis the next decade when she almost doubled her count of movies from 4 to 7 the rating fell from 7.3 to 6.53 which is a huge fall for an average. The lowest rating was at her end of the career with average of 6.1.

Kate Winslet started her acting career in movies in 1994 at the age of 19. From the blue line in the above graph, we can see that she did the greatest number of movies in the third decade 2011-2020 of her career with a count of 18. From the start we can see that there has been steady increase in her count of movies per decade since the starting of her career. As explained earlier the count in the last/present decade 2021-2030 is low as only 1/5th of the decade has only been passed but she has a count of 4 movies which is already strengthening my assumptions of increase trend of her count of movies in a decade if there are no other unforeseen circumstances which may restrict this trend.

We can also see that though movies count kept increasing over decades but the average rating per decade kept declining as it went from 7.02 to 6.48. Though in the latest decade the rating is 7.9 that’s because only 2 movies have been rated out of 4.

## Analysis 1.c

The above graph will help us to compare the number of movies done in a decade among the actors and actresses as mentioned in the graph. As we can see that mainly Brad Pitt, Kate Winslet, and Tom Cruise have career spans which overlaps. We can also infer that Katharine’s acting career was almost at the end when other has just started theirs. Hence, we will be mainly concentrating on the other three for comparison. We can see that both Tom and Brad had started their career in the 80s. Tom Cruise started his acting career in 1982 while Brad in 1988 hence the huge difference the count of movies from 1981 to 1990 between them is justifiable while Katharine was in her second last decade of acting career. In 1991 to 2000 we can see that Brad Pitt have over taken Tom Cruise when it comes to count of movie with almost double of Tom’s (Brad 16 and Tom 9). Kate Winslet started her acting career in 1994 and still she has a same count as Tom. From this we can infer that in 1991-2000, Brad Pitt had the best career performance in terms of movie count, followed Kate Winslet as she accomplished the count of 9 in lesser year and then followed by Tom Cruise. In 2001-2010 Kate Winslet overtook Brad Pitt with a count 16, while Brad Pitt was just a movie behind Kate when it came to count while Tom had just done 11 movies. In 2011-2020, the stats were not so close between any actors/actress as it was before, rather there was a steady difference 4 movies between actor. Kate again was top consecutively for the second decade with a count of 18, followed by Brad at 14 and then by Tom at 10. In the current decade 2021-2030 we can see that count is really low as only 1/5th of the decade has only been passed but still Kate has the highest count followed by Tom and Brad. Its too early to predict if Kate will still have highest count in the current decade but as the trend is following there is a high chance, she might have highest count in movies when compared to Brad Pitt and Tom Cruise.

In the previous graph we compared the count of movies in a decade among actors and actresses but in this graph, we will be comparing their average movies rating in those decade. In 1981-1990, even though Brad Pitt had done low count of movies (4) his average performance was lowest when compared to anyone else in the entire range of decades. Tom Cruise had a decent start with an average rating of 6.38. In 1991-2000, not only Brad Pitt had a highest count in movies for the actors mentioned in graphs but also had the best average movie rating while Tom and Kate were not so far behind with average rating of 7.02 and 7.01 respectively. In 2001-2010, Unlike Brad Pitt, Kate Winslet average rating started drop (from 7.01 to 6.8) as her count of movies started to increase while Brad had increase in average movie rating from 7.10 to 7.23. While for Tom average rating was almost consistent with a rating of 7.0. In 2011-2010, the declining trend line continued for Kate with her average rating dropping to 6.48. For Brad too there was a light decline in his movie’s performance as his average rating dropped to 6.96 while Tom had a better decade when compared to other with a slight increase in average rating to 7.03. In 2021-2030, the average ratings are not so average has only few movies have got released and many are not yet rated as per the database. Only Kate and Tom have movies rated in this decade with Tom topping the chart as only one of his movies has been rated while for Kate two of her movies have been rated.

## Analysis 2

|  |  |  |
| --- | --- | --- |
| NAME | Maximum Number of Movies from 1964 to 1973 | PROFESSION |
| Sung-il Shin | 304 | actor |
| Ji-mee Kim | 182 | actress |

Sung-il Shin is a south Korean actor, who had his peak time in movie industry in 1960s and 1970s. In the year range from 1964 to 1973 we can see that he has an outstanding number of 304 films which has made him the actor with most films in that range.

Ji-mee Kim is also a South Korean actress, who had her peak time as Sung-il Shin. With 182 films from 1964 to 1974, it has made her the actress with the greatest number of films in that range.

We can also see that the both actors and actress with highest number of films for the said range are from South Korea. But there is a significant difference between them, as actor has done 122 movies more than the actress that’s almost like 12-13 movies more per year on average.

While finding out the minimum number of 3 movies done by actors and actress in the same year range, we had come across couple of thousands of them across the world. Below I have printed 10 actors and actress with the same:

|  |  |  |
| --- | --- | --- |
| NAME | Minimum Number of Movies from 1964 to 1973 | PROFESSION |
| Alfredo Wally Barrón | 3 | actor |
| Quinn K. Redeker | 3 | actor |
| John Carson | 3 | actor |
| Jack Hedley | 3 | actor |
| Venancio Muro | 3 | actor |
| Cibo Liang | 3 | actor |
| Reg Park | 3 | actor |
| Marcel Anghelescu | 3 | actor |
| Agustín González | 3 | actor |
| Charles 'Bud' Tingwell | 3 | actor |
| Minako Yamada | 3 | actress |
| Ginger Rogers | 3 | actress |
| Sandra Mondaini | 3 | actress |
| Arlene Dahl | 3 | actress |
| Christl Erber | 3 | actress |
| Tota Alba | 3 | actress |
| Nina Grebeshkova | 3 | actress |
| Halina Kossobudzka | 3 | actress |
| Tisha Sterling | 3 | actress |
| Yoná Magalhães | 3 | actress |

From the above graph, we can see that the number of actors who did exactly 3 movies are more in count than the actresses’ count. From the above maximum and minimum number of films in the said decade it looks like gender male were more prominently involved in film industry across the world than females.

Below will the show maximum and minimum (3) number of movies count in year between years 1964 and 1973 in which extreme years are included.

|  |  |  |  |
| --- | --- | --- | --- |
| NAME | YEAR | NUMBER\_OF\_MOVIES\_IN\_A\_YEAR | PROFESSION |
| Sung-il Shin | 1967 | 44 | actor |
| Adoor Bhasi | 1973 | 44 | actor |
| Jeong-im Nam | 1968 | 42 | actress |

Sung-il Shin not only have maximum number of movies in the entire year range but also have maximum number of movies (1967) done in a year which comes under the said year range tied with an Indian actor Adoor Bhasi who did the same count of movies in 1973. Unlike the count for the entire year range, the maximum count for actress’s movies in per year wise doesn’t have a huge difference from actor’s count in same condition. Also, the actress Jeong-im Nam is different than the actress for maximum count of movies in the entire year range.

While finding out the minimum number of 3 movies done by actors and actress in a year in the same year range, we had come across couple of thousands of them across the world. Below I have printed 10 actors and actress with the same:

|  |  |  |  |
| --- | --- | --- | --- |
| NAME | YEAR | NUMBER\_OF\_MOVIES\_IN\_A\_YEAR | PROFESSION |
| Alex Nicol | 1964 | 3 | actor |
| Andrés Soler | 1964 | 3 | actor |
| Ettore Manni | 1964 | 3 | actor |
| George Chakiris | 1964 | 3 | actor |
| Gerhard Riedmann | 1964 | 3 | actor |
| Hiroshi Koizumi | 1964 | 3 | actor |
| Jacques Balutin | 1964 | 3 | actor |
| Wieslaw Golas | 1964 | 3 | actor |
| Carl Möhner | 1965 | 3 | actor |
| Elvis Presley | 1965 | 3 | actor |
| Didi Perego | 1964 | 3 | actress |
| Geneviève Grad | 1964 | 3 | actress |
| Irene Cruz | 1964 | 3 | actress |
| Deborah Walley | 1965 | 3 | actress |
| Harriet Andersson | 1965 | 3 | actress |
| Helga Liné | 1965 | 3 | actress |
| Margitta Scherr | 1965 | 3 | actress |
| Ingrid Pitt | 1966 | 3 | actress |
| Jirina Jirásková | 1966 | 3 | actress |
| María Silva | 1970 | 3 | actress |

The trend for minimum number of movies in a year is similar when it comes to entire year range. we can see that the number of actors who did exactly 3 movies in a year are more in count than the actresses’ count.

While comparing the whole Analysis 2 among itself, we see that for maximum count irrespective of conditions in the said years its mainly the South Korean actors/actresses who have the greatest count. This may depict the work culture in South Korean film Industry is pretty fast pace. Also, we can see that the count of acting professionals for minimum number of 3 movies when compared between whole year range and in a year within the same year range is drastically different.