DASC 5300/CSE 5300 Foundations of Computing

A PROJECT REPORT ON

DISK-BASED DATA ANALYSIS: PROJECT 3

BY

ABHISHEK SHRINIVAS JOSHI

ANUSHKA CHANDRAKANT PAWAR

UNDER THE GUIDANCE OF PROF. SHARMA CHAKRAVARTHY

CONTENTS

INTRODUCTION

In this project, we learned about Database Management Systems (DBMS). We learned and implemented SQL queries on databases like IMDb database. By working on this project, we understood the difference between the data which is available in files and the data stored in a Database management System. We performed analysis on the IMDb database by writing required queries in SQL as it is fast and effective. This made working on large real-world databases comparatively easier.

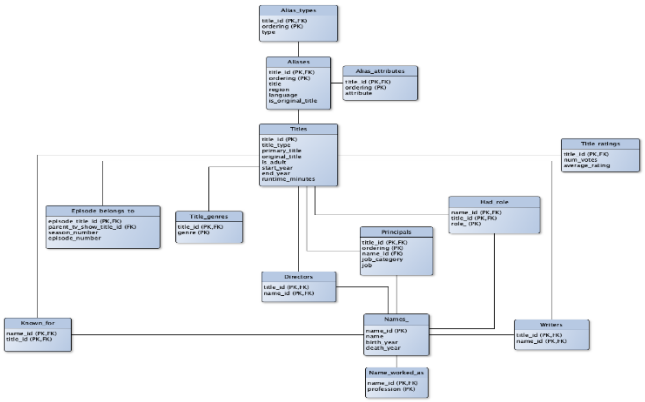
DATABASE

- IMDb database: It's the database of all the information about movies, series, videos, etc including their casts, crew, plot(summaries), ratings, and much more about it.

- Table 1: name.basics.tsv.g, nconst (string), primaryName (string), birthYear , deathYear , primaryProfession , knownForTitles

- Table 2: title.basics.tsv.gz , tconst (string) , titleType (string) , primaryTitle (string) , originalTitle (string) , isAdult (boolean) , startYear (YYYY) , endYear (YYYY) , runtimeMinutes , genres (string array) .

- Table 3: title.principals.tsv.gz, tconst (string), ordering (integer), nconst (string), category (string), job (string), characters (string).
- Schema Diagram: We took the schema diagram from https://dlwhittenbury.github.io/articles/IMDb-MySQL-Project/images/db_design/imdb-logical-schema.png



METHODOLOGY (Overall Status)

- We started the project with, the installation of several applications needed for the project.
    a) Pulse VPN
    b) Oracle SQL Developer
- We established a new connection to the IMDb database provided for the project, using the following details.
    a) Port Number 1523
    b) Database Name- IMDb
    c) Username- netid.
    d) Host Name- az6F72ldbp1.az.uta.edu
- After connection, we studied the database, its attributes from the given schema.
- We found out name_basics.NCONST values for the Actors and Actresses provided to us.
- We developed SQL queries and derived the outputs for them.
- Then we, transferred all the queries in a text file and exported them through "scp" from windows to our Omega server. (scp"filepathfromwindows" netid@omega.uta.edu:~)
- After that, we visualized the output we got using Microsoft Excel and performed a thorough analysis using the graphs obtained.
- We further researched more about the careers of the given Actors/Actresses on the internet and drew conclusions based on our analysis and research.

WORK DISTRIBUTION

- Analysis 1 Queries: Abhishek Shrinivas Joshi
- Analysis 1b Visualization: Abhishek Shrinivas Joshi
- Analysis 1c Visualization: Anushka Chandrakant Pawar
- Analysis 2 Queries: Anushka Chandrakant Pawar
- Report: Both Abhishek Joshi and Anushka Pawar

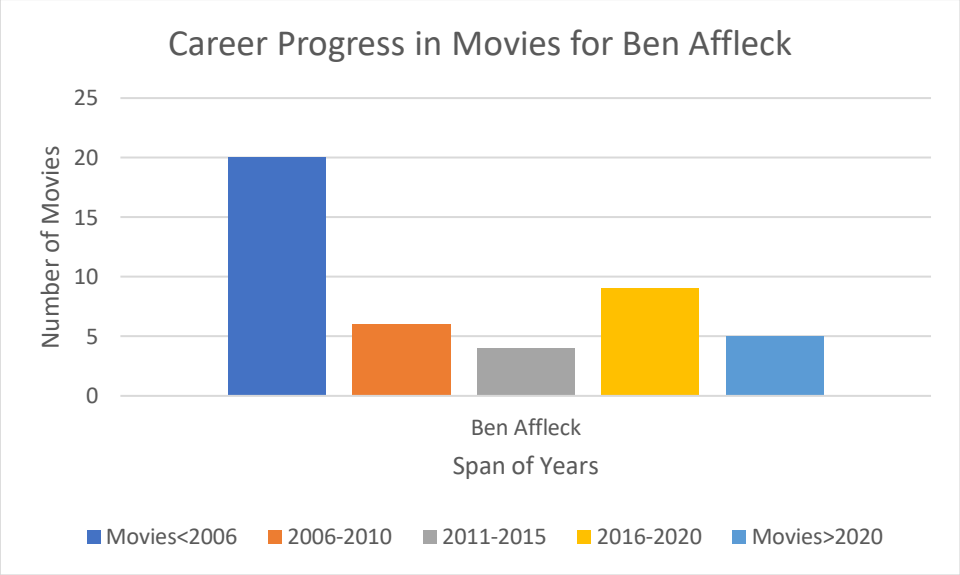
ANALYSIS AND RESULT

Analysis 1a:

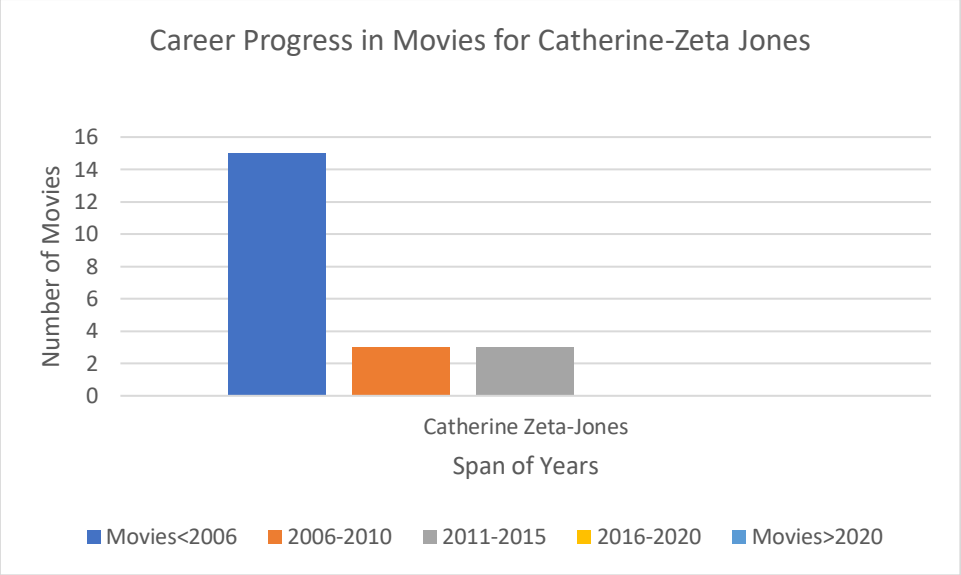
- We derived the Acting span of each Actor/Actress given to us.
- Ben Affleck started his acting career in 1984, he started working in the movies in 1995(Acting career span- 39 Years till present).
- Catherine-Zeta Jones started her acting career in local plays in the year 1981 and made her screen debut in 1990 in the movie “1001Nights” (Acting career span- 33 Years till present).
- Actress Emma Watson started her acting journey in 2002 (Acting career span- 21 Years till present).
- Johnny Depp started his career as an actor in 1984 (Acting Career span- 39 Years till the present).

Analysis 1b:

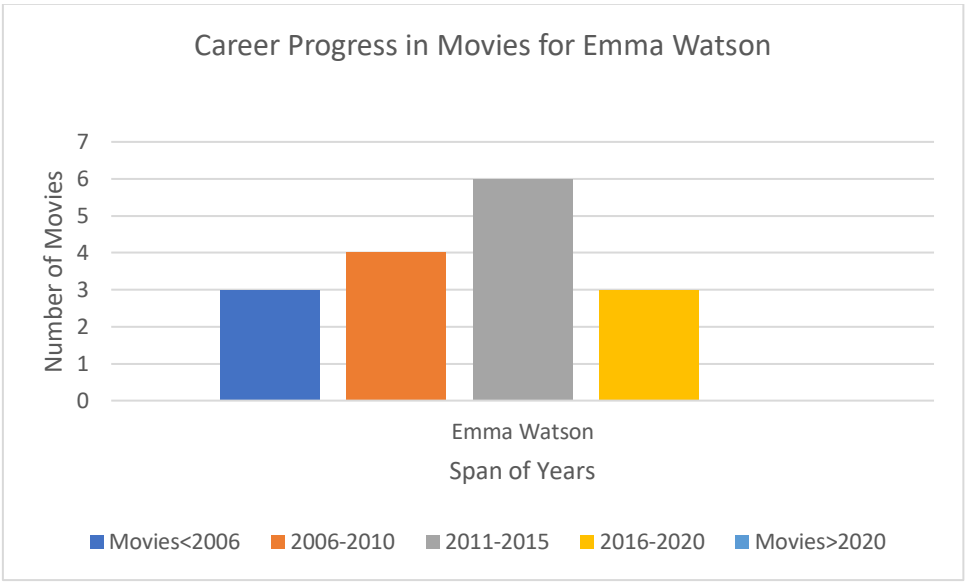
- We divided the Career span of Actors/Actresses into 5 disjoint periods and tracked and visualized the career progress of each actor/actress over the years in Movies.



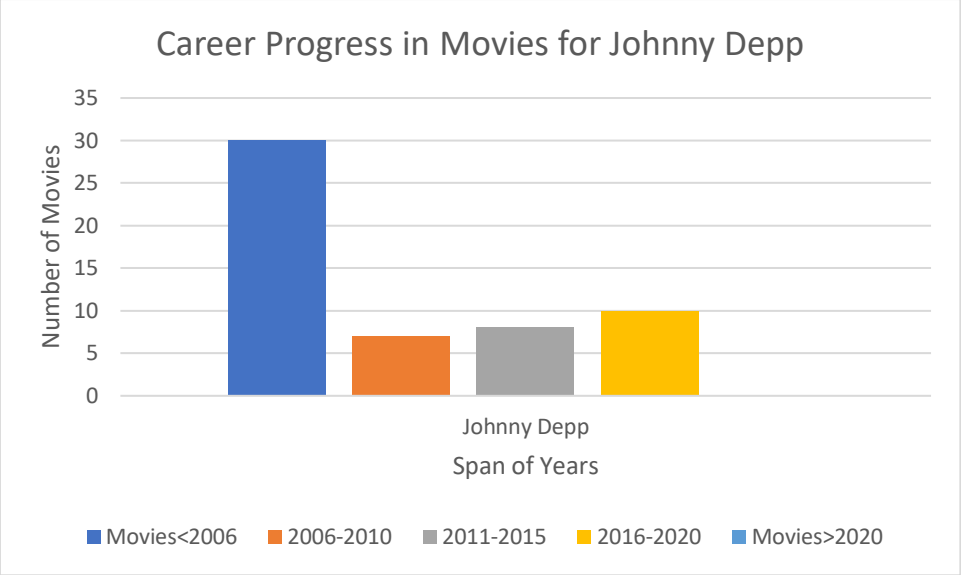
- According to our Analysis of the IMDb Database Actor Ben Affleck started his career as an actor in Movies in 1995.
- The above Bar Graph shows the visualization of his work in the movies from 1995 to the present.
- On further research we found out Ben Affleck started his acting career as a child actor in 1984, he starred in a series called “The Voyage of the Mimi” and has been working as an actor since then.
- The Graph shows, Affleck starred most in movies between the years 2016-2020, and upon doing some research, we got to know that the Actor gained a lot of fan following and appreciation when he started playing the role of Batman in 2016 and has starred in a number of movies from 2016 to 2020.



- Actress Catherine - Zeta Jones started her acting career in movies in 1990 with the movie “1001 Nights”.
- Upon research we found out Jones started acting when she was a child. She used to play roles in local theater plays from 1981.
- Jones worked in movies most in the years before 2006; after that, she went to tv series and theaters. She also wanted to lower her workload to concentrate on her family and her health as she was going through some health issues.

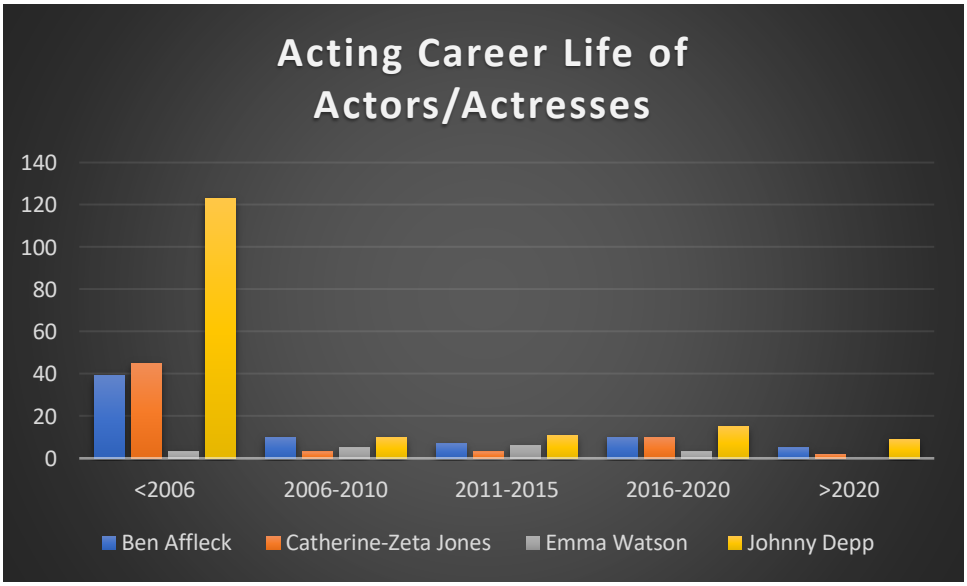


- Emma Watson started her Career in 2002 with her first film in the Harry Potter series (i.e. Harry Potter and the Sorcerer’s Stone).
- She gained quite a recognition for her character Hermione Granger in the Harry Potter series.
- Watson starred in a number of movies and other acting projects during 2011-2015 after the completion of Harry Potter. And was most active in those years of her acting career.



- Johnny Depp started his career in acting in the year 1984 with the horror film "A Nightmare on Elm Street".
- Upon research, we found out that Depp became the most commercially successful actor in the 2000s. Depp did several movies during those years, including Pirates of Caribbeans, Finding Neverland, Charlie and the Chocolate Factory, etc.
- Hence we can also see from the graph, Depp was most active as an actor in the early 2000s.

Analysis 1c:

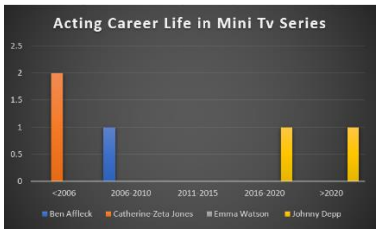


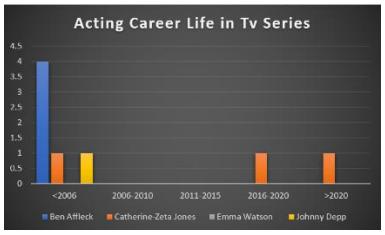
- Here, using queries we got the career progress of Actors/Actresses over the years in their acting careers.
- We can see that almost all Actors/actresses worked on screen until 2006. Upon research, we found that Johnny Depp and Ben Affleck started working as Producers and as

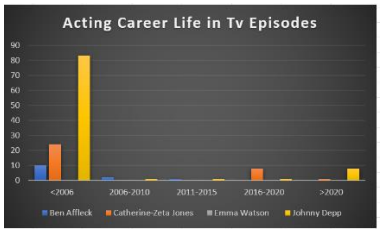
directors respectively after that. They were more concentrated on their off-screen work and roles as movie directors and producers.

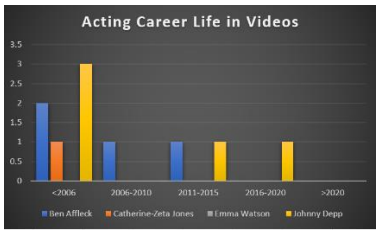
- Catherine-Zeta Jones went back to theaters and tv series after 2010.
- As of Emma Watson, who worked in several movies from 2002 to 2011. But later, she wanted to take a break from the movies. She still worked in a couple of movies after that including Greta Gerwing's adaption of the novel "Little Women". After her last movie in 2019. Emma Watson believes in taking considerable breaks in her Acting career and she stated that her future work includes "fewer red carpets and more conference meetings".

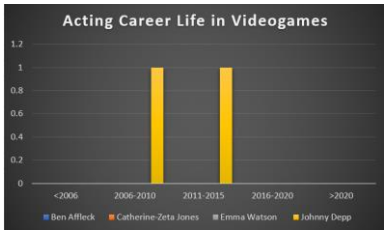
- To get a clearer idea about the Actor/Actress's acting career, we used the "Title type" attribute to analyze and visualize the career progress.

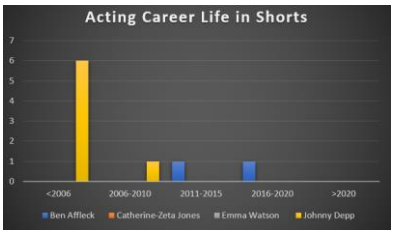


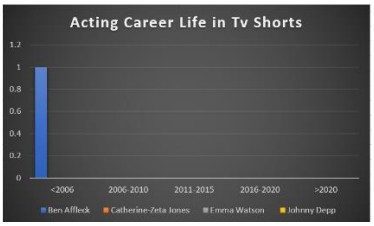


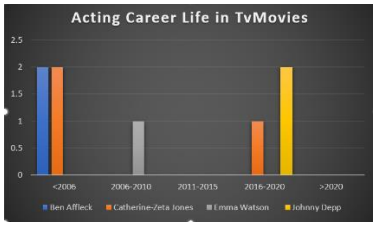


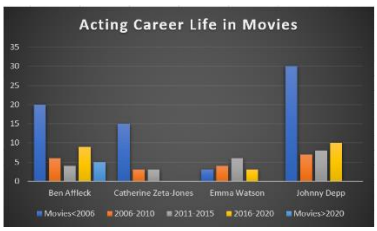












Analysis 2:

- Minimum number of movies by Actress in the given range (1986-1995)
  Alicia Orozco:1
  There were many (456) actresses who worked in 1 movie in the given span, Alicia Orozco is one of them.

- Minimum number of movies by Actor in the given range (1986-1995)
  Miguel Córcega:1
  There were many (789) actresses who worked in 1 movie in the given span, Miguel Córcega is one of them.

- Maximum number of movies by Actor in the given range (1986-1995)
  Brahmanandam:261

- Maximum number of movies by Actor in the given range (1986-1995)
  Kyôko Hashimoto:110

PROJECT TIMELINE

- 11/10/2022: Installed the applications required for the project
- 11/15/2022: Started working on SQL Queries.
- 11/19/2022- 11/25/2022: Worked on analysis1a and analysis1b.
- 11/26/2022- 11/30/2022: Worked on Analysis1c and analysis2.
- 12/03/2022- 12/05/2022: Worked on the project report.

DIFFICULTIES FACED IN THE PROJECT

- We got confused about the number of queries to be performed. To solve this issue, we used nested queries and reduced our number of queries from 20 to 5 in analysis 1b.
- For Analysis 2, we had to find the maximum number of movies and the minimum number of movies by an Actor/actress but we found a maximum number of movies where an Actor/actress worked in a single year. We overcame this problem, by dropping the idea of using Groupby over the year attribute.

REFERENCES

- https://dlwhittenbury.github.io/imdb-2-designing-a-mysql-database-and-performing-etl-for-imdb-dataset-using-python.html
- https://dlwhittenbury.github.io/imdb-2-designing-a-mysql-database-and-performing-etl-for-imdb-dataset-using-python.html
- https://dlwhittenbury.github.io/articles/IMDb-MySQL-Project/images/db_design/imdb-logical-schema.png
- https://en.wikipedia.org/wiki/Ben_Affleck
- https://en.wikipedia.org/wiki/Catherine_Zeta-Jones
- https://en.wikipedia.org/wiki/Emma_Watson#1999%E2%80%932009:_Harry_Potter_and_worldwide_recognition
- https://en.wikipedia.org/wiki/Johnny_Depp