Shawn Ruby

February 15, 2022

C964

001345550

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

Sam Smith CTO

Movies R Us

1234 Movies Rd.

Moviesville, New York

Dear Mr. Smith,

Our company has always sought to achieve building a movie streaming platform that is popular among its users. Due to the rise of movie platforms like Netflix and Hulu, we as a company are losing our competitive edge.

To be able to stay in competition with some of the big competitors, I suggest we implement machine learning into our original system and use it to provide movie recommendations based on user inputs. This will return a list of movies that the user can view.

I believe implementing a recommendation system through machine learning will provide beneficial aspects to the company. Implementing this system will aim to increase user retention by allowing customers to view movies based on what they want to see. If a user searches for ‘The Dark Knight’, it will provide movies that are similar to that. If we can provide other movies the user may want to watch, then we prevent that user from searching for movies on other platforms.

The cost for this project to commence is estimated to be $25,200. The project can be estimated to take about 6 weeks to complete. The cost can be broken up into a software developer over 6 weeks of $14,400. The Software developer has multiple years of experience creating software and working with data and should be a good fit to keep the project within budget. A project manager over 6 weeks of $10,800. The project manager that has been chosen has a lot of experience managing projects and should help keep the project on track.

Thank you for your time and consideration on this project. If you have any questions, please reach out to me at my office.

## Project Recommendation

### **Problem Summary**

Streaming platforms have become the defacto for watching movies and tv shows. Because of this, tech giants such as Netflix, Hulu, Disney, etc. have taken the market by storm and pushed the boundaries on what is capable within a movie streaming environment. One thing these streaming platforms use is machine learning to better present content to its users through the use of recommendations. To stay relevant and keep up with the competition, Movies R Us needs to implement machine learning in a way to provide recommendations just like how other movie streaming platforms do. If we do not implement a system like this, we will lose our users to other streaming services with no hope of gaining new users. If a system like this is in place, it will increase customer retention because if the user chooses to watch a movie on our platform, then they may want to watch a movie similar to the one they watched. The way this will be implemented would allow users to input a movie of their choice and see movies similar to that one. This will increase revenue for the company through customer retention.

### **Application Benefits**

Currently, as the company stands, we are losing revenue because users are not staying on our platform. We need to give users a reason to stay on our platform. The project I have proposed consists of implementing machine learning so users can view movies that are related and thus increasing user retention. The business will benefit from this project because the more time people use our platform, the more likely they will want to keep resubscribing. If we can get the users a reason to not want to leave, this will increase our revenue going into the company.

### **Application Description**

The application will use a bunch of data in the form of CSV files consisting of titles, cast, directors, reviews, rating, rating count, genre, language, etc. The application will read this data and use it to make a calculated prediction on what movies the user may want to watch. The calculation will consist of movies that are well received and have a similar title, director, crew, etc. If a user wants to see movies related to 'The Dark Knight' then the application will recommend users movies that are similar in name, director, crew, etc. but filter out bad review movies.

### **Data Description**

The data that will be used to construct this product is a movie dataset. That is filled with movies that are on our streaming platform that has been entered when the movies were first brought into our platform. The data structure for our data is in the form of CSV files consisting of different amounts of data such as title, cast, directors, reviews, rating, rating count, genre, language, etc. Almost all of the data is just nominal data such as the title, and crew which is just raw data for that movie. The quantitative data that would be present would be the ratings due they can be used to mathematically compare to other movies to determine popularity. Apart from that, all data such as the movie metadata is nominal meaning it can't be measured. Most of the movie data is purely not dependent but it can be viewed as dependent. For example, the movie's director is not fully dependent on that movie because they've made other movies. However, that director is the director of that movie so it can be viewed as dependent on that movie. The data that is purely dependent on the movie is going to be the ratings and the perception of the movie from the public. There are no anomalies in the movie data because it is just metadata and ratings. The limitation that would be present in this dataset would be null values but these would be replaced with non-null values to prevent the algorithm from reading it and giving inaccurate readings.

### **Objectives and Hypothesis**

Resulting of this project, a fully developed movie recommender system will be created and deployed into the already working Graphic User Interface. This movie recommender system will provide feedback based on what the user wants to watch in the form of movies most similar to the user's input. This recommender system will provide recommendations with at least 80% accuracy. Once the movie recommendation system is implemented, it will likely cause people to want to stay on our platform more which will prevent them from wanting to leave. We seek to increase customer retention by 10% over one year.

### **Methodology**

For the development of this project, the traditional waterfall methodology will be used. Since this is a small project, the waterfall method is the most cost-effective solution. Using this method, it will allow the project team to adhere to requirements better and get the project done promptly. The waterfall method can be broken up into its categories. The **requirements** phase will be first introduced where the project requirements can be discussed and refined to build a roadmap for the project. The **design** phase will be introduced next to be able to formulate a plan of attack for the project. This involves the creation of the project on a system-wide level. Once the design face has finished, **implementing** the code will begin. This involved the actual creation of the software to be used during the project. Lastly, after the code has been fully implemented, **verification** will begin. This involves verifying that everything related to the project is working correctly and is ready for public use.

### **Funding Requirements**

The project will be developed on a windows 10 computer with an associated hardware cost of $1000. This is a standard computer used for various software-related jobs. On top of that, 2 personnel would be dedicated to the project. One is a software developer with an associated cost of $14,400 over 6 weeks at $60 per hour. A project manager with the associated cost of $10,800 over 6 weeks at $45 per hour. From start to finish, the total cost of the project will be $25,200. There is no associated cost with any licensing or tools. The program will be developed in Pycharm IDE version 2022.3.2 and running Python 3.11.

### **Data Precautions**

Properly handling data is very important for the function of this project. Ways we can help protect sensitive data would be to control access to data both being sent and retrieved through the use of encryption, and data segregation between users. Protecting users' information regarding payment information is also very important which said would need to be encrypted to prevent unauthorized access. Also, teaching and educating employees not to use data for irresponsible and unauthorized use.

### **Developer’s Expertise**

The software developer selected for this job has a bachelor's degree in Computer Science. He has multiple years of experience with programming. He has experience with multiple programming languages as well. Not only with programming, but he also has experience with databases. Outside of programming, He also has Project+ and ITIL 4 Foundations qualifications. To perform the duties for this project, the developer will need experience programming in Python, Machine Learning applications, and constructing GUIs. Of these, the programmer already has experience in and would provide a great fit for the project.

# Part B: Project Proposal

## Problem Statement

Many companies have made a huge market contribution to movie streaming. Those companies include Netflix, Hulu, etc. Movies R Us has been successful with movie streaming before but the company is on the decline with its users. One thing that sets Movies R Us apart from other streaming platforms is we as a company do not provide any interaction with the user on their preferences. Netflix is a great example of a company that uses Machine learning to provide recommendations to its users. How Netflix does this is "Whenever you access the Netflix service, our recommendations system strives to help you find a show or movie to enjoy with minimal effort. We estimate the likelihood that you will watch a particular title in our catalog based on a number of factors including:

* your interactions with our service (such as your viewing history and how you rated other titles),
* other members with similar tastes and preferences on our service, and
* information about the titles, such as their genre, categories, actors, release year, etc.

In addition to knowing what you have watched on Netflix, to best personalize the recommendations we also look at things like:

* the time of day you watch,
* the devices you are watching Netflix on, and how long you watch.”
* Describe the problem.” (Netflix, n.d.)

Netflix is a very successful company and to compete with somebody like them, we need to implement our Machine Learning algorithm that will provide recommendations based on what the user wants. This will help Movies R Us stay relevant against other companies.

## Customer Summary

This project is aimed to serve customers in the form of subscribers in our movie streaming service. As of now, customers are leaving our platform at an alarming rate to other streaming platforms that provide a much more accommodating user experience through the use of Machine Learning. Machine learning takes into consideration what the user likes, and runs an algorithm with the use of Pandas and Scikit-learn to create a weighted rating and provides that information to the user in visual representations of what the user may want to watch next. Providing this machine learning will allow Movies R Us to keep up with the competition and retain its users and allow more users to want to join.

## Existing System Analysis

The existing system is written in Python with a built GUI that uses CustomTkinter. The existing system is a movie streaming platform but provides no machine learning and provides no recommendations. Streaming services are continuing to advance in a sense where AI and machine learning are helping users not need much input. We need to implement these services as well to keep up and be able to compete with others. Users don't want to search for every movie. We need to provide recommendations.

## Data

To use this program, data within CSV data files will need to be present. With these CSV files, there is a lot of data for 45,467 movies. This data contains title, id, IMDB id, genre, cast, director, actors, keywords, ratings, revenue, budget, languages, vote average, vote count, etc. Without this data, the program won't run. Data will be collected and stored in the CSV files for every new movie that is added to our streaming platform. It will also remove accordingly. It will also take into consideration of user privacy and act accordingly. There are no anomalies within the dataset as this is just movie data. However, there may be null data. The null data will be replaced with temporary values and filtered out to not mess with the algorithm's runtime.

## Project Methodology

For the development of this project, the traditional waterfall methodology will be used. Since this is a small project, the waterfall method is the most cost-effective solution. Using this method, it will allow the project team to adhere to requirements better and get the project done promptly. The waterfall method can be broken up into its categories. The **requirements** phase will be first introduced where the project requirements can be discussed and refined to build a roadmap for the project. The requirements will consist of what algorithm to use, how to use it, how to capture the data, etc. The **design** phase will be introduced next to be able to formulate a plan of attack for the project. This involves designing the plan on how to implement the algorithm we chose, how to incorporate the data into the already working program, and putting all of these pieces together. Once the design face has finished, **implementing** the code will begin. This involved the actual creation of the software to be used during the project. Lastly, after the code has been fully implemented, **verification** will begin. This involves verifying that everything related to the project is working correctly and is ready for public use. Once this can be verified that it is working and stable, then the project needs to get ready for project closure and to make the update public.

## Project Outcomes

**Goals**

The goal we wish to achieve through this project is to increase subscriptions through the concept of gaining more users onto the streaming platform. Also, we want to increase customer retention meaning the customers we already have we want to keep them here and not give them a reason to want to leave.

**Objectives**

The objectives for this project would be to increase sales and customer retention by 10% within 1 year by allowing subscribers to view other movies related to the movies they pick. This is a more user-friendly approach and helps keeps customers invested.

**Deliverables**

The deliverables produced by this project would be to produce a working piece of software that enables the use of machine learning to better interact with users when searching for movies.

**User Guide**

To interact with the program, the user will need to input a movie in the text box, and depending on which visual representation they want to see, they will select the button accordingly. There is also a drop-down menu to change the color of the program and change the font size.

## Implementation Plan

When the program will be produced, it will be time to implement it into current systems. To do this, the system can be broken up into:

* General Strategy – An important concept with building this and any new system is making sure it can be implemented a current system without breaking it or introducing new bugs. Now, this may not always be possible but this is the goal to strive for. Not only that, but since this is used for customers, we need to ensure the new system is similar to the existing system to not confuse or scare away existing users.
* Phases of rollout – The new system will not be made public immediately. Once we believe the system is ready, it will first enter the Alpha stage where it will be tested by a few employees and bug testers. Once this process has been deemed acceptable, it will enter the Beta stage where existing customers can voluntarily sign up and test the existing system to potentially find bugs from a consumer perspective. This period will be open for 2 weeks. During these 2 weeks, if any bugs are found, we will correct them until the system is bug-free. After the 2 weeks, we will send the working product to the public and continue to improve as time goes on.
* Assess - The hope for this data is it will bring more sales to the company. Once the system is in play, we will evaluate sales before and after implementation for a selected period. We hope this will increase sales by 10% within one year.

## Evaluation Plan

Testing is an important part of project success to ensure everything goes correctly as planned. With the creation of the software, Unit testing will be done to ensure each function, loop, etc. is working as intended. As more functions, classes, etc. are added to the program, we need to integrate and test them with the rest of the system to ensure a new function or loop is added it doesn't break other parts of the system. When the whole system is completed, we will test the whole system to make sure each part integrates well with the others and that there are no obvious bugs. After we find no bugs from a programmer's perspective, we will roll out the project in the Alpha and Beta stages to perform acceptance testing. In Alpha, the project will be considered complete but it will undergo bug testing by bug testers. In Beta, we will roll out the project to end users where they can test based on the end user perspective.

After all of the testings is done before the rollout of the working piece, it will be rolled out to the public. The testing does not end here though. We as a company still need to take customer feedback to fix and improve our system. Not just that, but we need to ensure we are making a profit and increasing our sales by 10% within the first year of development.

## Resources and Costs

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| IT team | 1 developer implementing this | $14,400  $60/hour for 6 weeks |
| Project manager | Project manager to manage to project | $10,800  $45/hour for 6 weeks |
|  | **Total** | $36,075 |

## Timeline and Milestones

1/23/2023 - 1/25/2023: A proposal is created to implement machine learning and create a recommendation system into the original system. This is sent to management for formal project creation.

1/25/2023 - 2/01/2023: The proposal is accepted by management and an IT team is assembled along with a project manager to create a project.

2/01/2023 – 2/15/2023: A plan is created that will undergo the process of implanting the machine learning into the existing system. The project will use Amazon Redshift to store and use data

2/20/2023 – 3/6/2023: Plan has commenced. The plan starts by updating the current application to accommodate the new features planned to be created like the recommendation dialog box. Also, creating the code that will store the data to be used as machine learning.

3/6/2023 – 3/10/2023: Test the system as a whole to make sure everything is working.

3/13/2023 – 3/14/2023: Make the update live so the program can start tracking data.

3/15/2023 – 3/16/2023: Project closure/sign off

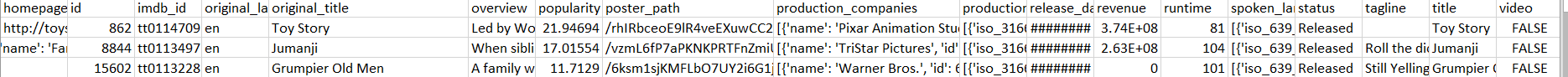
# Part D: Post-implementation Report

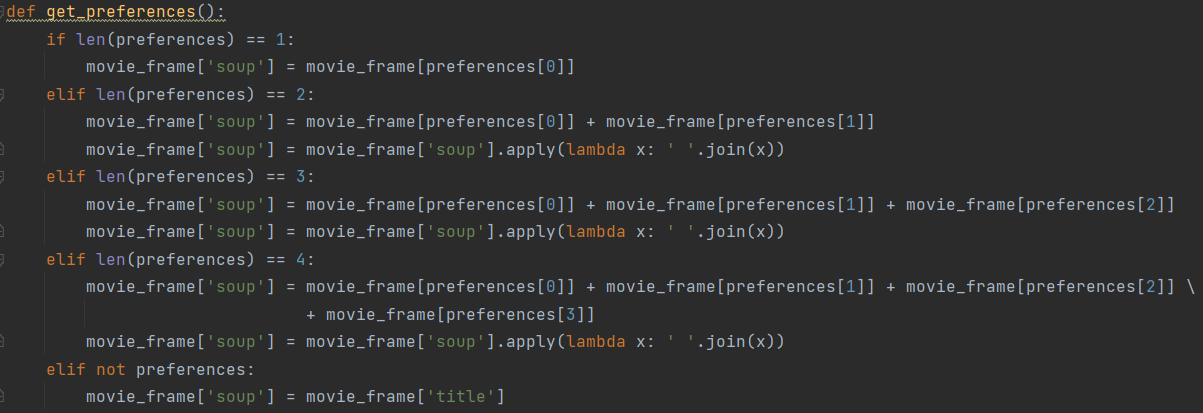
## Purpose of project

The purpose of the project is to provide a user-friendly means of recommending movies to users. This concept has become an essential function in a lot of different platforms. Movies R Us does not provide any means of machine learning or a recommendations-based system. Movies R Us must implement a system like this to stay relevant against the competition. Providing a system to recommend movies would prove to be beneficial for our company. This will increase sales and customer retention. We are hoping this will increase by 10% over 1 year.

## Datasets

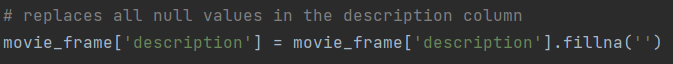
The data set used to perform the movie recommendation machine learning consists of four CSV data files that will need to be present. With these CSV files, there is a lot of data for 45,467 movies. This data contains title, id, IMDB id, genre, cast, director, actors, keywords, ratings, revenue, budget, languages, vote average, vote count, etc. The process data will be keywords, genres, and directors in the form of crew jobs, cast, and titles. The data will be processed by adding those to a column and will run the algorithm based on those criteria. Here is an example of the data:



Here is an example of the data being processed based on what preferences are selected: 

The dataset will be provided with the program and be accessed at will.

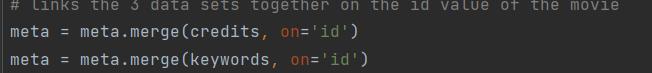
## Data Product Code

The data will be processed in a few ways to make it usable. First, we remove null values from various columns 

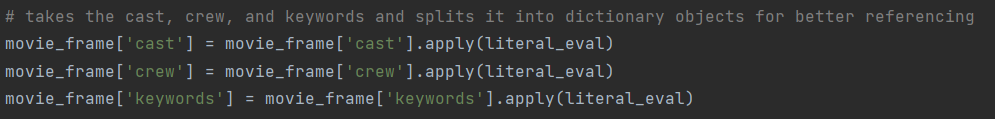
Then we assign int values to IDs, vote counts, etc.



Then we merge ids from different CSV files based on the movie id:



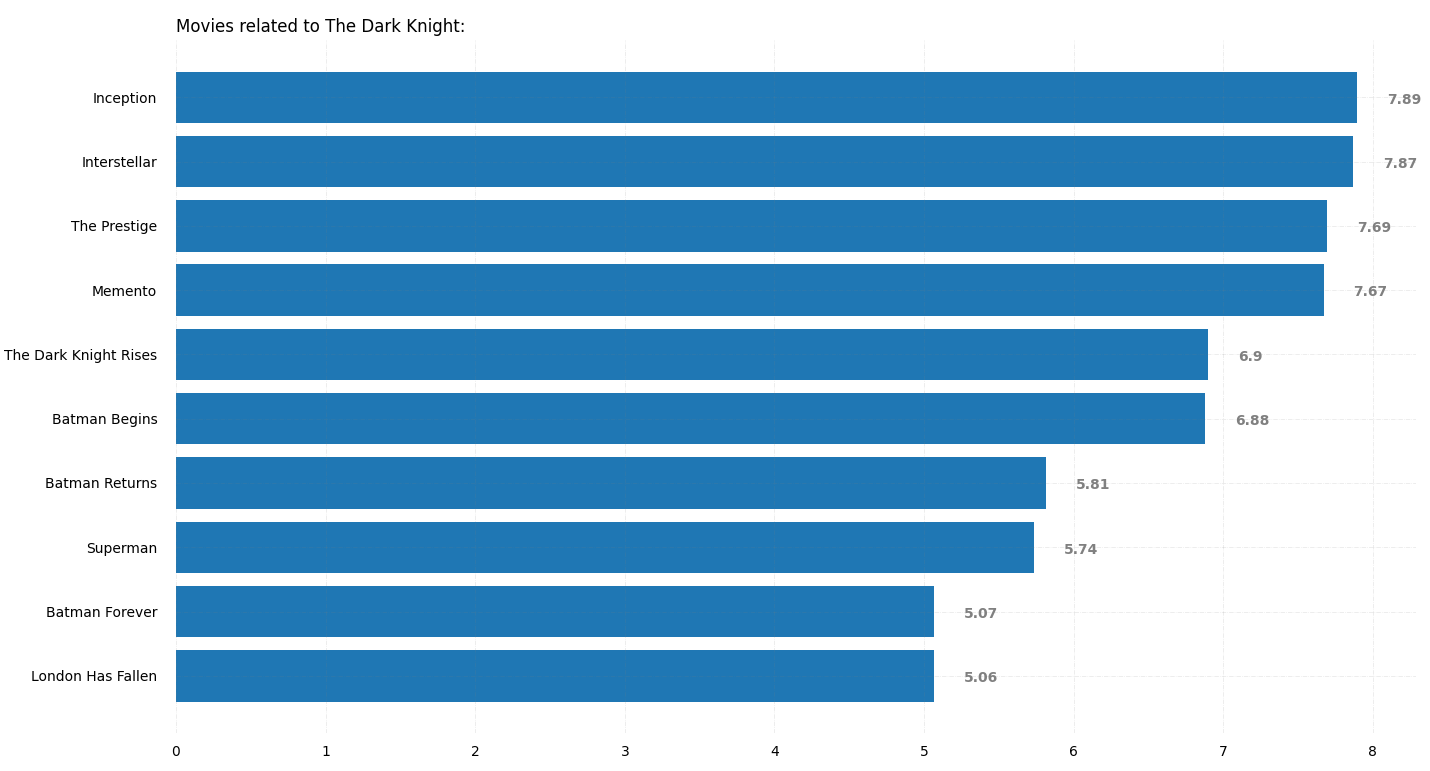
Breaks the information into dictionary objects so it would be in the format 'Name: Tom Hanks':



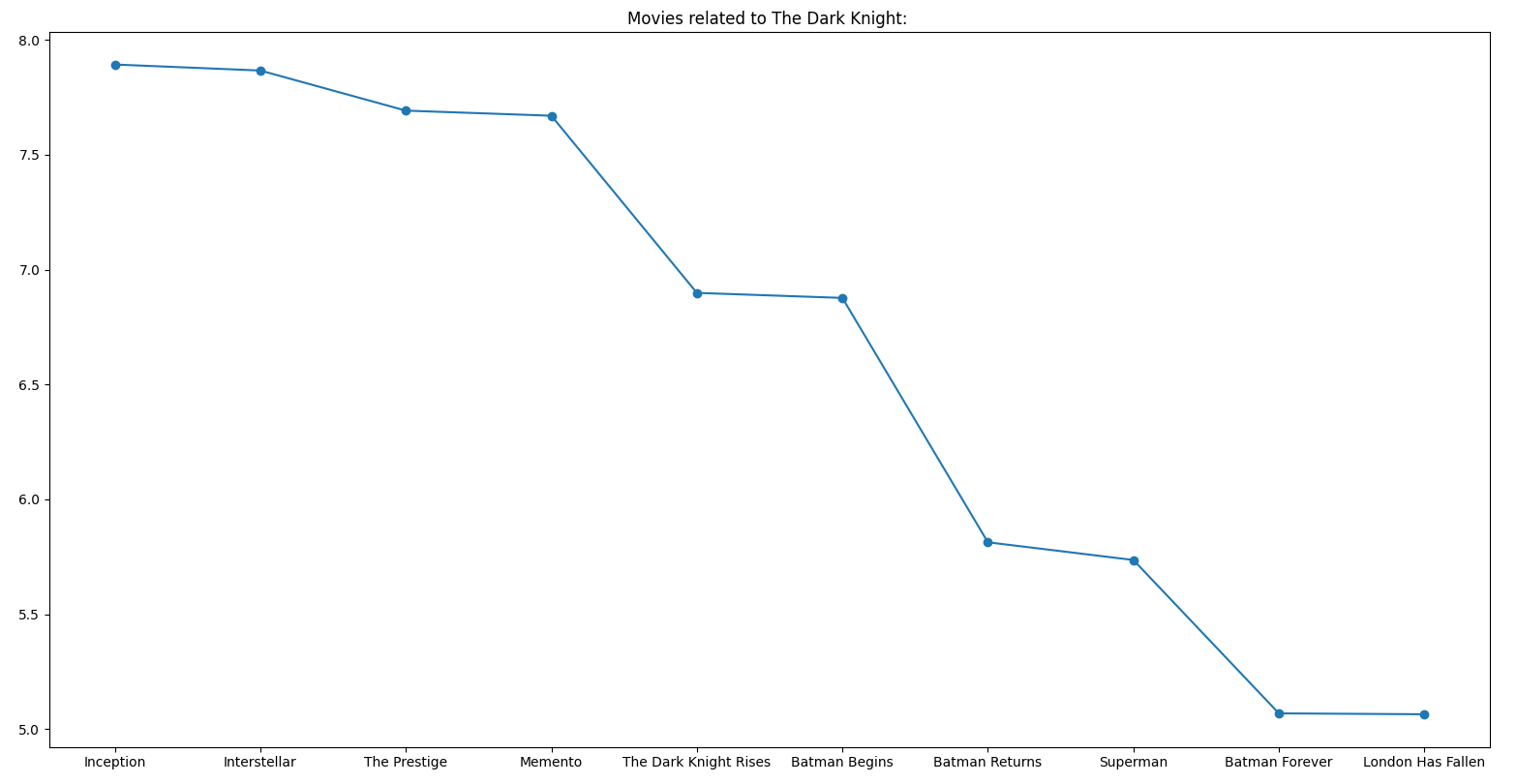
And then we can use that information to perform calculations.

There are 3 visual representations for the recommendations:

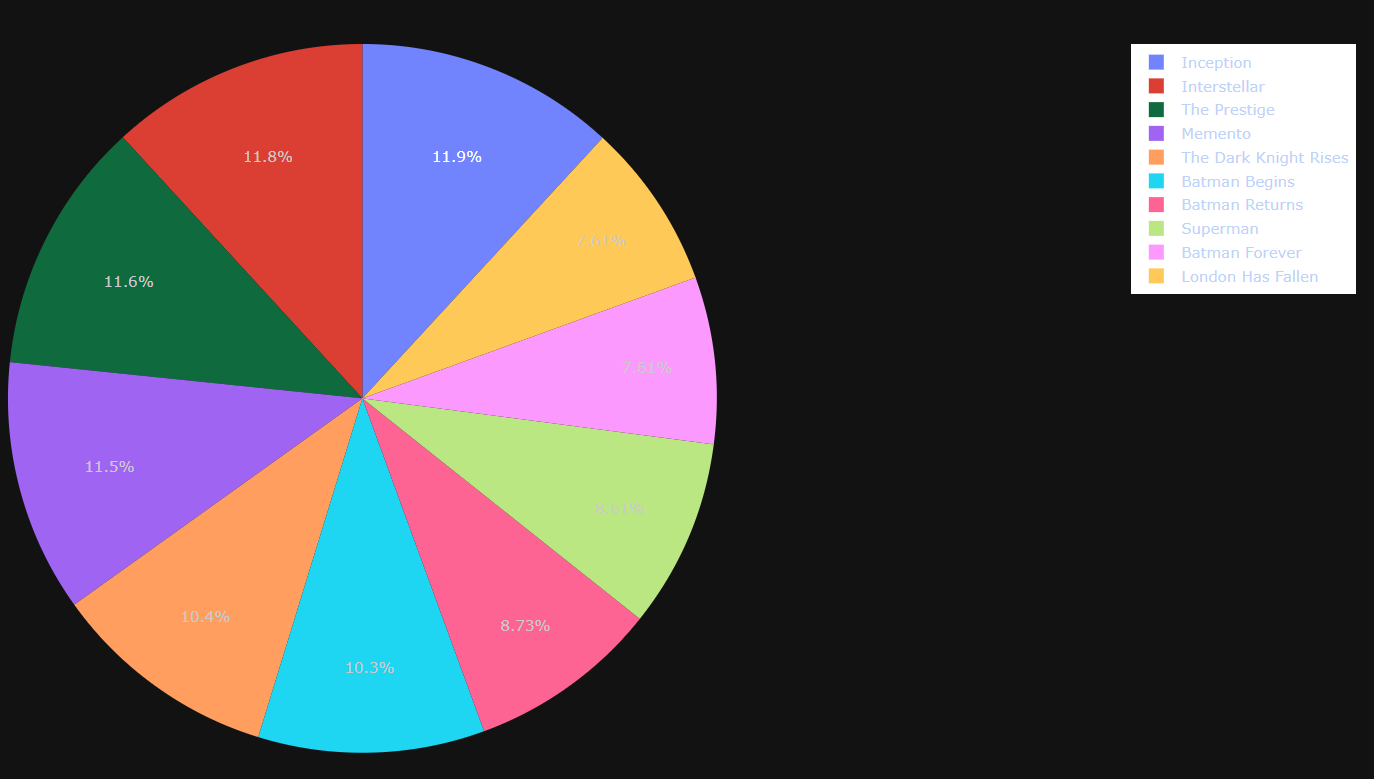
Bar Chart



Plot Graph

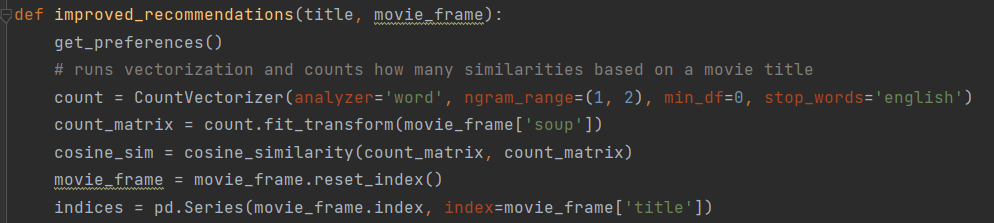


Pie Chart



On top of descriptive methods, there are also non-descriptive methods. For analytical methods, linear regression is used for this project. The reason why linear regression is used is that "Regression analysis is used to estimate the relationship between a set of variables. When conducting any type of regression analysis, you’re looking to see if there’s a correlation between a dependent variable (that’s the variable or outcome you want to measure or predict) and any number of independent variables (factors which may have an impact on the dependent variable).” (Stevens, 2023) The data was trained using cosine similarity on a column that contains all parameters of the data we want to use. For example, we chose to take into consideration, genre, top 4 cast members, director, and keywords from the checkboxes. This data was added to one column and runs a cosine similarity function to retrieve a list of relevant movies based on that criteria.

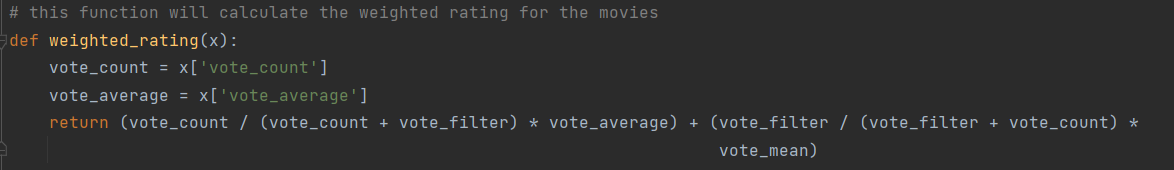




Not just that, but a way the data is filtered is through the use of ratings. Why would we want a movie with that 2/10 rating? For a movie to even be considered, it has to rank in the top 96% of the vote counts.



This is used in the calculation for a weighted rating from IMDB:



We can verify this because we want to see recommendations for 'The Dark Knight' based on the dataset, inception has a rating of 8.1/10. The weighted rating for it is 7.89. 7.89/8.1 equates to approx 97.4% which is above the 96% quantile.

## Objective (or Hypothesis) Verification

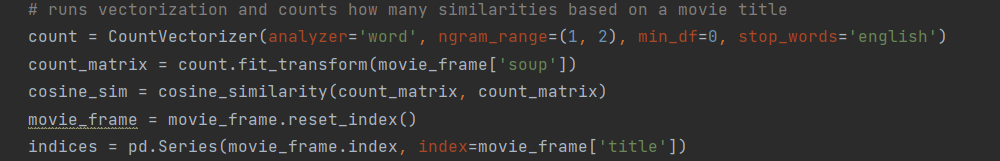
The goal of the movie recommendation system is to recommend movies based on a movie title. The program will take into consideration the director, keywords, genres, and cast if any are selected. This proves to be correct because when running the program, it will recommend movies of similar titles, directors, genres, etc.

## Effective Visualization and Reporting

As shown in 'Data Product Code,' the 3 visual representations are shown there. Each visual representation shows each movie title and its rating based on the criteria selected and ran through cosine similarity to retrieve similarity. The data analysis chosen is linear regression because we want to perform content filtering by providing a movie title. This title is stored along with genre, director, cast, and keywords. The predictive method is the movie title and it's used to find similarities based on other criteria.

## Accuracy Analysis

The data analysis method used in this project is linear regression. Because of this, it's very hard to measure accuracy because the outcome depends on multiple variables. In my example, the multiple variables are genre, cast, keywords, and director. This information is used to determine similarities using the cosine similarities function comparing each title to the other title. Since linear regression is used, there's just no accurate way of determining accuracy. The way the accuracy can be further improved upon is by considering other variables like runtime, production company, etc.

This is where the cosine similarities function is run: 

## Application Testing

The testing was done through a top-down process. Once having the idea in mind of how to execute it, I broke up each part into functions and tested each function to make sure the program behaved the way I wanted it to. After that, I created a GUI and made sure that was functional how I wanted. From there, I incorporated all the data into the GUI and made sure it worked together.

## Application Files

To run the program, you can just open the project in Pycharm. The list of files is ShawnRubyC964 as the master folder that contains all files. From there, there is a folder called customtkinter which contains all of the files needed to use CustomTkinter for the GUI. There is a folder called Data, which contains all the CSV files for the data. Lastly, there are 2 .py files. Main.py contains all the code associated with the GUI and CalculateData.py contains all the code associated with machine learning.

## User Guide

1. The project will be submitted as a zip file. First, unzip it by right-clicking and clicking extract here.
2. Download the Pycharm community edition from here: https://www.jetbrains.com/pycharm/download/#section=windows
3. Open Pycharm. The Pycharm version used for this project is 2022.3.2
4. Download Python if not already done so from this link: <https://www.python.org/downloads/>
5. Open the project in pycharm
6. Make sure to download every import at the top of the 2 .py files by right-clicking the ones with a red line underneath them – show context menu – download library
7. You can run the program now
8. Type in any movie you want (spelling does not count)
9. Select each checkbox of the criteria you want the recommender to take into consideration
10. Click either of the 3 buttons to represent the data in 3 different visualizations. It will also write the data to the console.

## Summation of Learning Experience

This project was fun to make and learn. Creating GUIs through Java helped me with GUI layout for python. I just needed to learn how to use CustomTkinter for the creation of the GUI in python. Creating the 2 previous python projects through WGU helped me learn python as a whole, with even stuff like dictionaries, tuples, etc. The Java projects taught me about classes, multiple files, etc.

To learn the machine learning process, I used this notebook for an example: <https://www.kaggle.com/code/paramarthasengupta/movies-recommendation-tool-approaching-patterns/notebook>. I used this notebook to learn how different functions were used and what they do specifically. I used code from this notebook to apply machine learning concepts and to create the algorithm within the movie recommendation system. The data that was used for this project is here: <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/code>.

I aspire to become better at programming I enjoy it. It was very fun to take the notebook and learn how everything works and apply it to my example. I want to find a career in computer software, and maybe even machine learning. I plan to learn more about machine learning and solve more problems.

# Part E: Sources

## References

Netflix. (n.d.). *How Netflix’s Recommendations System Works*. Retrieved from Netflix.com: https://help.netflix.com/en/node/100639

Stevens, E. (2023, January 4). *The 7 Most Useful Data Analysis Methods and Techniques*. Retrieved from CareerFoundry: https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/