**Data Science Project Report**

**1. Principal Investigator**

Each group member’s name and email address.

Evan Finke ([finkee1@newpaltz.edu](mailto:finkee1@newpaltz.edu))

Damien Brandel ([brandeld1@newpaltz.edu](mailto:brandeld1@newpaltz.edu))

* 1. Individual Contribution Breakdown (list the percentage)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | Member 1 | Member 2 | Member 3 | Total |
| Introduction | Evan (50) | Damien (50) | N/A | 100% |
| Background | Evan (50) | Damien (50) | N/A | 100% |
| Implementation | Evan (50) | Damien (50) | N/A | 100% |
| Experiment Results and Discussion | Evan (50) | Damien (50) | N/A | 100% |
| Conclusion | Evan (50) | Damien (50) | N/A | 100% |
| Other contribution and explain | Evan (50) | Damien (50) | N/A | 100% |

**2. Title of Project**

Popularity and Ratings of Movie Genres

**3. Mentoring**

Professor Min Chen, Department of Computer Science, SUNY-New Paltz

[chenm@newpaltz.edu](mailto:chenm@newpaltz.edu)

**4. Introduction**

4.1 Project Motivation

Movies have a widespread appeal and the number of ways we could work with data surrounding movies was pretty significant. For example, we chose to find data based on genre, but we could have found the data based on the best year, or based on the title of the movie. It felt like a perfect intro project to a field of computer science we were rather unfamiliar with.

4.2 Aims and Objectives

By determining which genres are the most well received for each year, we could maybe draw correlations to real world events that year to the opinions and sentiments of a small sample of the public.

**5. Background/History of the Study**

People are always indecisive on the topic of choosing what to watch. Being able to input a movie’s genre and get a good sense on how those types of movies are typically perceived could be a determining factor for picking a movie. The dataset being used contains 27,753,444 ratings and 1,108,997 tag applications across 58,098 movies. This data was created by 283,228 users between January 09, 1995 and September 26, 2018.

**6. Approach and Implementation**

We sorted each movie within the dataset by its production year, rating, genre, and tags given by reviewers while getting rid of the unnecessary information for this study. Using the tags, we’d determine the sentiment related to the movie using a sentiment calculator. Once we cleaned the dataset to just rating, genre, year, and sentiment, we did a Map Reduce based on genre to determine the average rating and sentiment of a genre within any given year. We put that data in an excel workbook to visualize the data that we collected. For the machine learning algorithm, we used k nearest neighbors. We split the cleaned movie data into 70% training and 30% test. The input was year, rating, sentiment, and occurrence. The prediction for these inputs was a genre.

**7. Experiment Results and Discussion**

Based on the dataset we collected. The highest rated movie genre of all time was Film Noir with 3.4 out of 5. Most of the genres fell around the 3 rating with only Horror falling below with a 2.8. With average sentiment, however, IMAX had the highest sentiment with a value of 3.57. Unlike ratings that were all relatively close, IMAX was significantly higher with the runner-up being Romance with a .84 sentiment rating. The lowest sentiment was ironically Film-Noir with a -2.02. The biggest flaw with our algorithm is probably the fact that we did not sort or identify users. The biggest flaw in the data was probably the fact that there just wasn’t enough data for every genre. Some genres had hundreds of ratings while others only had a handful.

**8. Conclusion**

Overall, this has been a very helpful project that taught us how to use pandas through python, it taught us how to use MapReduce through Hadoop, and it taught us various other skills that wouldn’t have expected to learn. We learned more about excel when we were visualizing the data, and learned more about the capabilities of IDEs when making sure Hadoop and Pandas were working. With the dataset itself, there are many different angles that we could approach this movies subject if we were to continue working with it. This project really showed us the value and depth that the data science field has to offer.

**9. References**

**Dataset:**

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19. <https://doi.org/10.1145/2827872>

<https://www.kaggle.com/datasets/whenamancodes/popular-movies-datasets-58000-movies>

**Data Origin:**

*Non-commercial, personalized movie recommendations.* MovieLens. (n.d.). Retrieved September 21, 2022, from <https://movielens.org/>

**Sentiment Calculator:**

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O&#x27;Reilly Media, Inc."