

**TASK**

**Exploratory Data Analysis on the MovieLens Data Set**

[](http://www.hyperiondev.com/portal/)

THIS REPORT WAS WRITTEN BY: AVESHNEE IYER

**CONTENTS**

1. **Introduction**
2. **Data Cleaning**
3. **Missing Data**
4. **Data Stories and Visualisations**
5. **INTRODUCTION**

Conduct an Exploratory Data Analysis (EDA) on the MovieLens dataset to get insights, clean the data and answer some questions. The data needs to be joined as it is in 3 different files and then analysed.

**Approach**

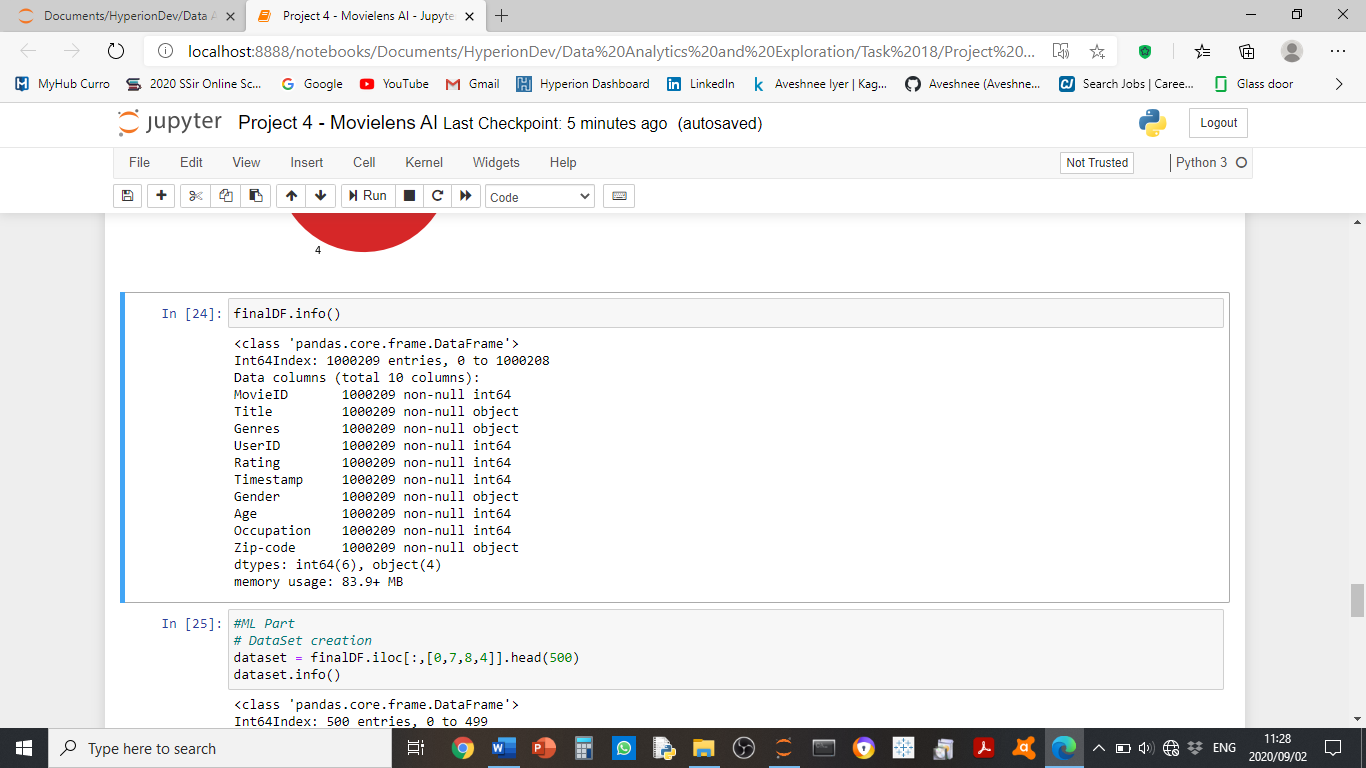
* Finding patterns in data.
* Determining relationships in data.
* Checking of assumptions.
* Drawing conclusions.

**What question(s) are you trying to solve (or prove wrong)?**

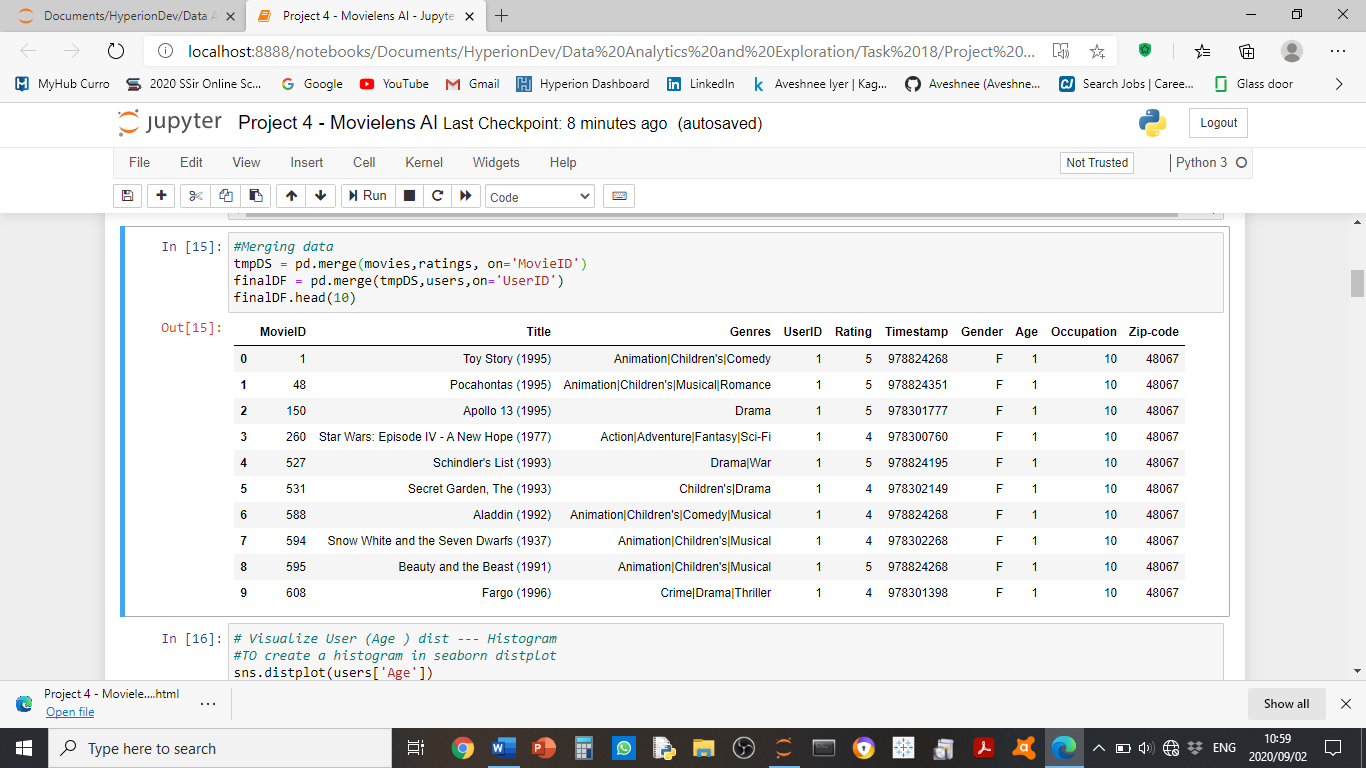
1. Determine user rating for a single movie or determine performance between sequels.
2. Understand viewership by age group for targeted marketing, retention activities or market expansion.
3. Insight to the top 25 movies by viewership rating, indicating type of movies or sequels that perform better.
4. Insights that will assist in recommendations later, by understanding which movies received higher ratings by user
5. **DATA CLEANING**

**Data Files merged:**

The 3 data files movies.dat, ratings.dat and users.dat were merged using MovieID and UserID as primary keys.



**Data Snapshot of the merged data (10 rows):**



**Feature Attributes:**

MovieID – numeric count of the number of movies, range between 1 and 3952

Title – movie title (and year)

Genres – classification of movie type

User Id – numeric assigned to the user, range between 1 & 6040

Rating – categorical rating of a movie, range from 1 to 5

Timestamp – measured in seconds

Gender – male or female

Age – age of the user

Occupation – occupation of the user

Zip Code – postal/ area code of the user

Occupation is chosen from the following choices:

\* 0: "other" or not specified

\* 1: "academic/educator"

\* 2: "artist"

\* 3: "clerical/admin"

\* 4: "college/grad student"

\* 5: "customer service"

\* 6: "doctor/health care"

\* 7: "executive/managerial"

\* 8: "farmer"

\* 9: "homemaker"

\* 10: "K-12 student"

\* 11: "lawyer"

\* 12: "programmer"

\* 13: "retired"

\* 14: "sales/marketing"

\* 15: "scientist"

\* 16: "self-employed"

\* 17: "technician/engineer"

\* 18: "tradesman/craftsman"

\* 19: "unemployed"

\* 20: "writer"

1. MISSING DATA

**Visualised Missing values:**



There seems to be no missing values in the visualisation.

**Missing values data points and count:**

MovieID 0

Title 0

Genres 0

UserID 0

Rating 0

Timestamp 0

Gender 0

Age 0

Occupation 0

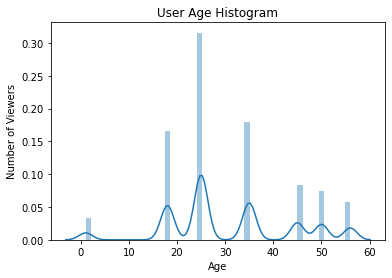
Zip-code 0

dtype: int64

The above does not reflect any missing values.

1. DATA STORIES AND VISUALIZATIONS
2. **User Age Histogram (using seaborn)**

User age is highest in the 25-27 age group and ranges mostly between 18 years old to 34/35 years of age. This could be an opportunity to market promotions to ages not being attracted at this point to expand the user base. It also assists in retaining the current age group to ensure that we are catering in genres and age group.



1. **Overall Rating by Viewers**

Rating

1 56174

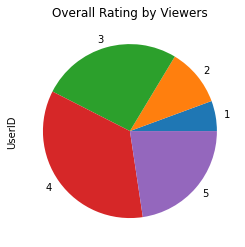
2 107557

3 261197

4 348971

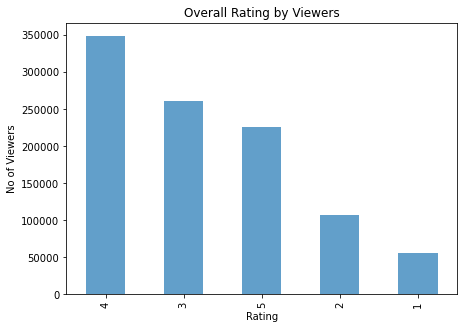
5 226310

Name: UserID, dtype: int64



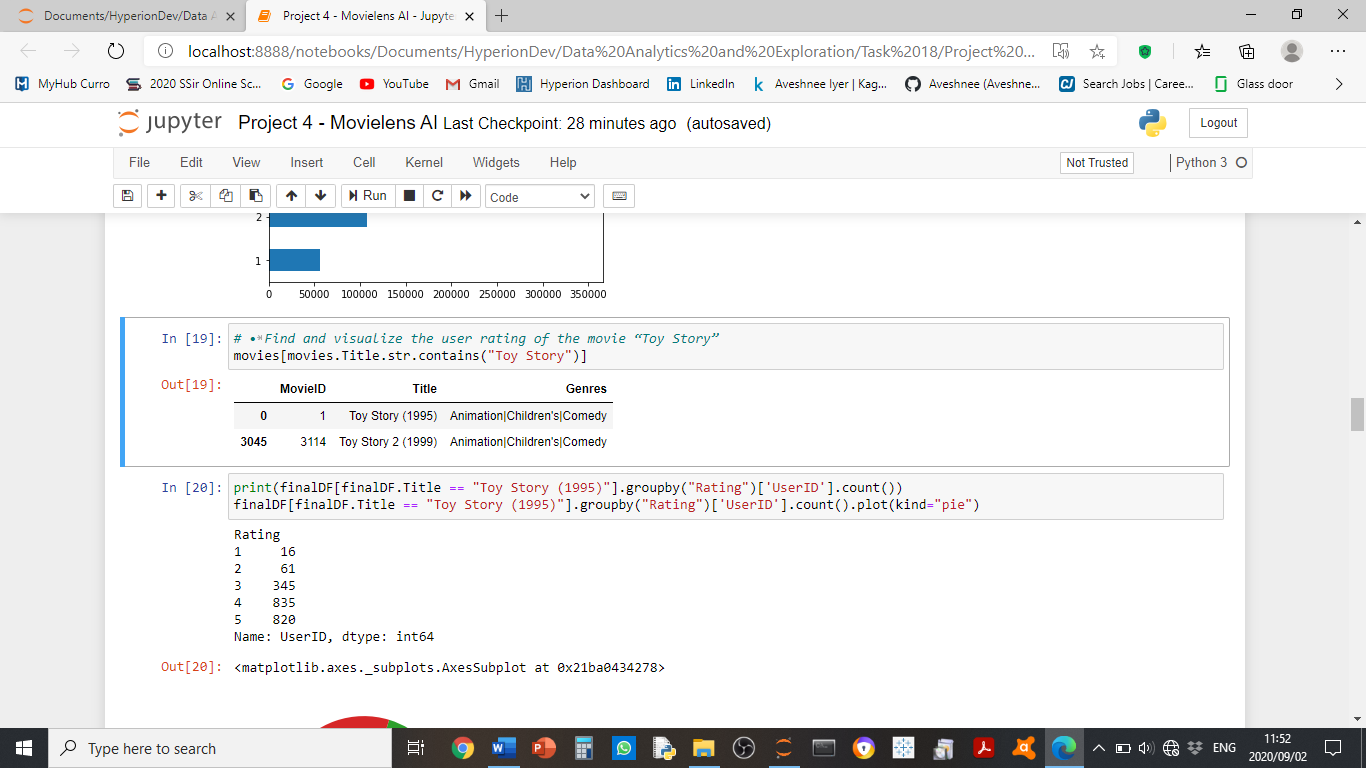
The highest ratings are number 3 and 4, containing the bulk of the ratings. So generally, users seem to be quite happy with the movies that they have watched. Not seeing major dissatisfactions levels yet.

1. **Overall Rating by Viewers**

****

This is represented in a bar graph and we can see similarly that the highest rating is 4 and then followed by rating 3 and then rating 5. The summary stats also show the ratings mean are high. The questions we ask then is how to ensure that we keep these ratings up while still growing the market in future years?

1. **Similar movies (sequel performance) example “Toy Story”**



Using an example of one movie, we can group movies and their sequel. By doing this we can understand the performance of the movie parts/ sequels and understand performance for a part or group.

1. **Ratings by Movie Title (“Toystory”)**

Rating

1 16

2 61

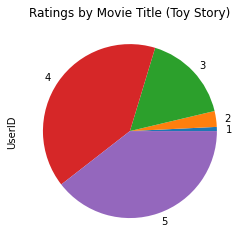
3 345

4 835

5 820

Name: UserID, dtype: int64

Most users rated Toy Story between rating 4 & 5. It is a very highly rated movie.



1. **Viewership by Age for a Movie Title (“Toy Story”)**

Age

1 112

18 448

25 790

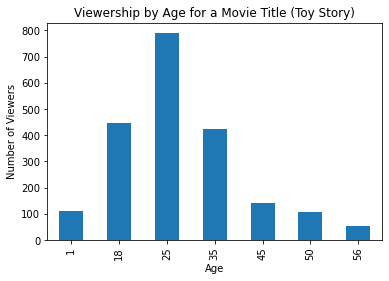
35 423

45 143

50 108

56 53

Name: Age, dtype: int64

****

Many of the viewers of the Toy Story movie are in the age group of 25 years of age. This shows that the genre, although aimed at children do attract parents to watch with them and therefore consist of marketing opportunities for both parents and children.

1. **The Top 25 movies by Viewership Rating**

Title

American Beauty (1999) 3428

Star Wars: Episode IV - A New Hope (1977) 2991

Star Wars: Episode V - The Empire Strikes Back (1980) 2990

Star Wars: Episode VI - Return of the Jedi (1983) 2883

Jurassic Park (1993) 2672

Saving Private Ryan (1998) 2653

Terminator 2: Judgment Day (1991) 2649

Matrix, The (1999) 2590

Back to the Future (1985) 2583

Silence of the Lambs, The (1991) 2578

Men in Black (1997) 2538

Raiders of the Lost Ark (1981) 2514

Fargo (1996) 2513

Sixth Sense, The (1999) 2459

Braveheart (1995) 2443

Shakespeare in Love (1998) 2369

Princess Bride, The (1987) 2318

Schindler's List (1993) 2304

L.A. Confidential (1997) 2288

Groundhog Day (1993) 2278

E.T. the Extra-Terrestrial (1982) 2269

Star Wars: Episode I - The Phantom Menace (1999) 2250

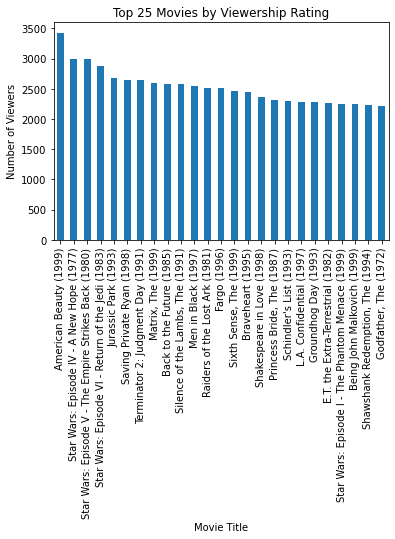
Being John Malkovich (1999) 2241

Shawshank Redemption, The (1994) 2227

Godfather, The (1972) 2223

Name: Rating, dtype: int64

The highest rated movie is American Beauty produced in 1999, followed by the latter half of the Star Wars sequel 4, 5, 6. and then Jurassic Park. The genres science fiction is highly rated and could be where the opportunities to continue with Science Fiction sequel type movies. It would be interesting to investigate why the latter half of Star Wars did better than 1, 2, 3.



1. **Ratings at a Viewer level (User ID = 2696)**

Rating

1 2

2 3

3 3

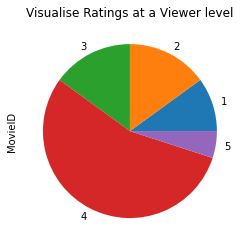
4 11

5 1

Name: MovieID, dtype: int64

1. **Visualise Ratings at a Viewer level**

At a user level we can gauge the type of movie genres the customer enjoyed and therefore able to make recommendations around similar genres. User level detail can lead to more movies being watched and enjoyed, and potential for higher NPS scores from customer satisfaction. Also, more insights for the movie production industry on whom to market, including cinema insight for their websites.

****

1. **Dataset creation to be able to apply Machine Learning algo’s**

<class 'pandas.core.frame.DataFrame'>

Int64Index: 500 entries, 0 to 499

Data columns (total 4 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 MovieID 500 non-null int64

1 Age 500 non-null int64

2 Occupation 500 non-null int64

3 Rating 500 non-null int64

dtypes: int64(4)

memory usage: 19.5 KB

1. **Create features and labels to be able to start to build ML models to make recommendations**

features = dataset.iloc[:,[0,1,2]].values

label = dataset.iloc[:,3].values

**THIS REPORT WAS WRITTEN BY: AVESHNEE IYER**