**SHRI G. S. INSTITUTE OF TECHNOLOGY & SCIENCE, INDORE**

**Computer Engineering Department**

**Scrapping of Imdb Movies Data**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **Submitted By :** | | **Guided By :** | | |  | Ayush Kumar (0801CS161016)  Jitesh Rathore (0801CS161030) | | Mr. Swarnim Soni | | |  | Kamalvanshi Meera (0801CS161030) | |  | | |  | |  | |  | |  | | **CONTENTS** | |  | |

**Title Page No.**

Introduction ………………………………………………. 2

Objective……………………………………………………. 2

Existing System…………………………………….. 2

Need of the Project ………………………………… 2

Installation………………………………………………... 3

System Requirements………………………………. 3

Tools Used………………………………………….. 3

Design Details and Implementation………………………. 8

Results ……………………………………………………. 11

Conclusion …………………………………………………17

**1**

**INTRODUCTION**

**Objective**

Bollywood has witnessed exponential growth in terms of volume of business, manpower employed, the number of movies produced each year and also the global reach. There are a number of factors like Actors, Directors, Time of Release, Genre, Production house, etc. which affect the outcome of a movie. This Dataset contains movie title, release date, genre, writers, directors, sequel and hit/flop count of various movies. Bollywood Actors and Actress ranking dataset is also used for their overall rating and performance based on google hits and likes.

**Existing System**

Previously, different predictions on dataset like whether the movie will be hit or flop has been given. And the movie recommendation system for users has been made for their preferred taste of movies. Twitter reactions for the upcoming movies and after watching it, based on the majority votes success of the movie can be predicted. User is also suggested a movie which they will prefer to watch based on their past watch history.

**Need for the Project**

We are going to perform a new type of prediction on dataset based on clustering different categories of actors and directors by their preferred movie types. This will help to determine the coordination between actor and director. It will also predict the movie type mostly chosen by an actor. Example: Akshay Kumar always prefers that movies which are related to social causes and inspires the young generation. Based on the movie type and different categories like genre and writer, we will make a further prediction that the actor will choose that movie to work or not.

We are also going to analyse different movie actors and directors ranking based on their movie counts and success ratios.

**2**

**INSTALLATION**

This covers all the description and installation of all the tools required in this project. All the software and packages used are Open Source software and are freely available.

**System Requirements**

|  |  |
| --- | --- |
| Type of Hardware | Hardware Requirements |
| Hardware | Dual-core Intel Pentium compatible processor |
| Disk space | 4GB disk space(minimum) |
| Memory | 4GB(recommended) |

Table 2.1: Hardware Requirements

|  |  |
| --- | --- |
| Type of software | Software Requirements |
| Operating System | Windows 10 |
| Web browser | Google Chrome |
| Text Editor | Jupyter Notebook and Spyder |

Table 2.2: Software Requirements

**Tools Used**

**1. Beautiful Soup:**

[Beautiful Soup](http://www.crummy.com/software/BeautifulSoup/) is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. These instructions illustrate all major features of Beautiful Soup 4.

**from** **bs4** **import** BeautifulSoup

soup = BeautifulSoup(html\_doc, 'html.parser')

**print**(soup.prettify())

**Installation -**

If you’re using a recent version of Debian or Ubuntu Linux, you can install Beautiful Soup with the system package manager:

$ apt-get install python-bs4

For Windows:

$ pip install beautifulsoup4

$ python setup.py install

**2. Pandas:**

Pandas provide essential data structures like series, dataframes, and panels which help in manipulating data sets and time series.

With Python Pandas, it is easier to clean and wrangle with your data. Pandas Features like these make it a great choice for data science and analysis. Using it with libraries like NumPy and [Matplotlib](https://data-flair.training/blogs/python-matplotlib-tutorial/) makes it all the more useful.

Given are steps to install Pandas in Python:

a. Installing Pandas

To install pandas, you can use pip-

pip install pandas

b. Importing Pandas

Now let’s import this using an alias-

>>>import pandas as pd

c. Importing a Dataset

You can use the function read\_csv() to make it read a CSV file. Let’s import the furniture dataset.

>>> furniture=pd.**read\_csv**('furniture.csv')

A DataFrame is an essential data structure with pandas. It lets us deal with data in a tabular fashion. The rows are observations and columns are variables. We have the following syntax for this-

pandas.**DataFrame**( data, index, columns, dtype, copy)

**3. Matplotlib:**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

Windows, Linux and macOS distributions have matplotlib and most of its dependencies as wheel packages. Run the following command to install matplotlib package :

python -mpip install -U matplotlib

**Importing matplotlib :** from matplotlib import pyplot as plt

**4. K-means Clustering Algorithm-**

A cluster refers to a collection of data points aggregated together because of certain similarities.

K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible.

How the K-means algorithm works

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

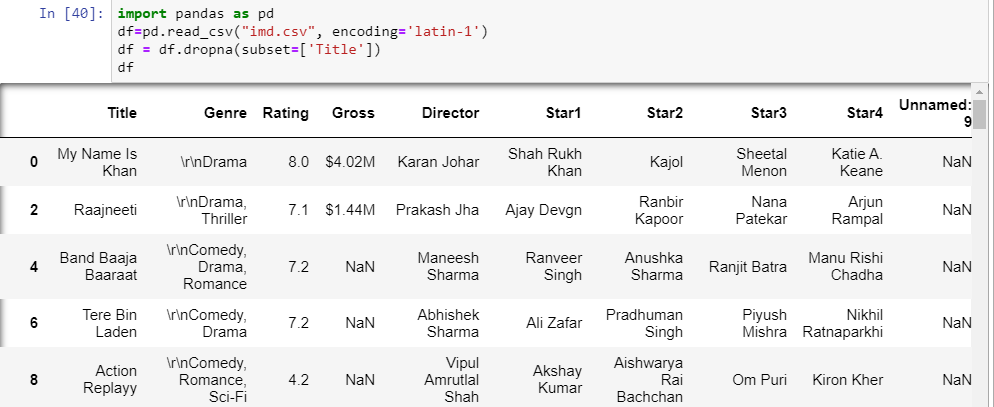
* The centroids have stabilized — there is no change in their values because the clustering has been successful.
* The defined number of iterations has been achieved.

3

**DESIGN AND IMPLEMENTATION DETAILS**

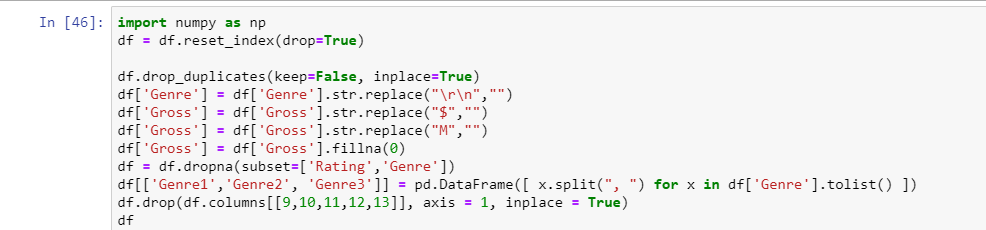
This section describes how dataset is generated and complete visualisation on our dataset.

1. There are ​800 instances (movies) in total. Our dataset contains​ all possible instances ​from the webpage. For each instance, it consists of ​certificate, duration, rating, genre, vote, gross​. We used Imdb url: <https://www.imdb.com/list/ls005197923/?st_dt=&mode=detail&page='+str(x+1)+'&sort=list_order,asc>to scrap data. Initially, we got data that contained lots of NULL values and raw formatted data.

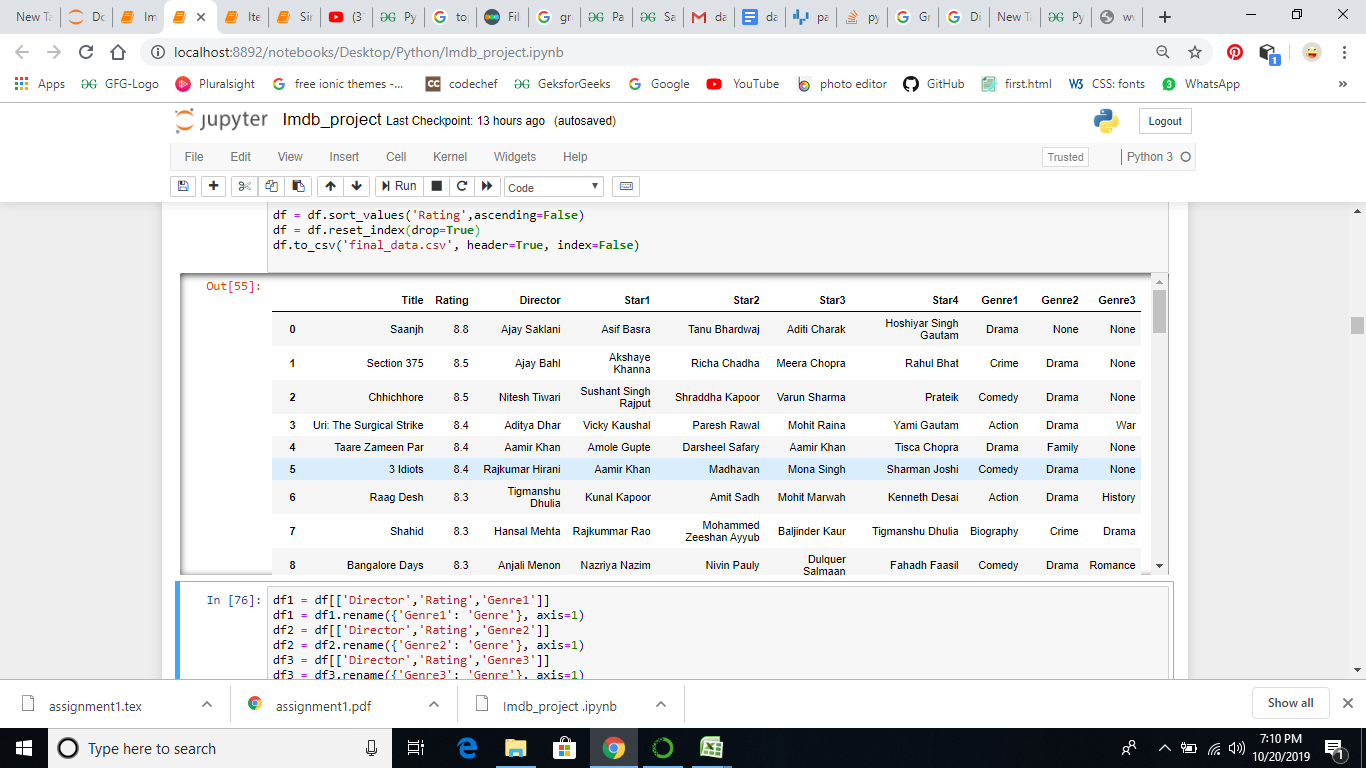


1. Cleaning of Data:

Initially, we drop the Movie Title column which was not required in our dataset and the Genre column contained ‘\r\n’ at the beginning, therefore we cleaned the dataset.



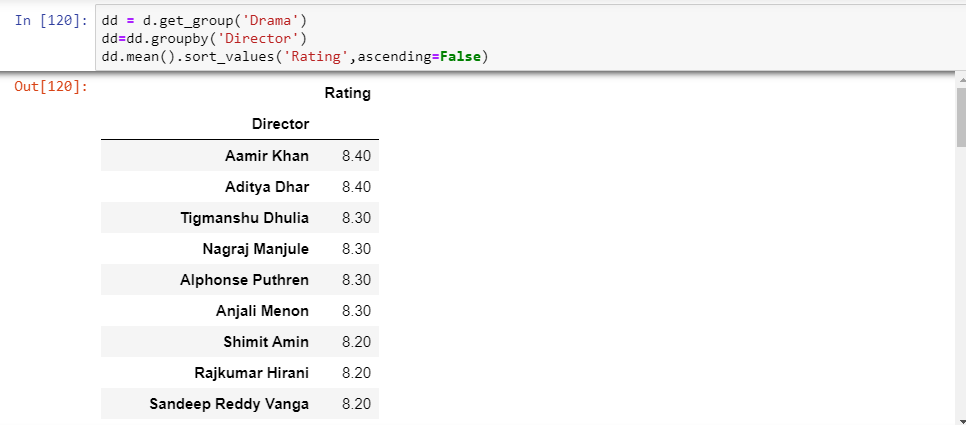
1. After cleaning, Final dataset looked like:



We need to study director and actor top ratings based on Genre Type. With the help of the Pandas dataframe, we analyzed that.

Group by method was used to group directors based on Genre type.

1. Top directors and their rating:



Top stars and their Rating on different Genre type:



4

**RESULT**

Top stars and their rating for action:



Director Actor Coordination:

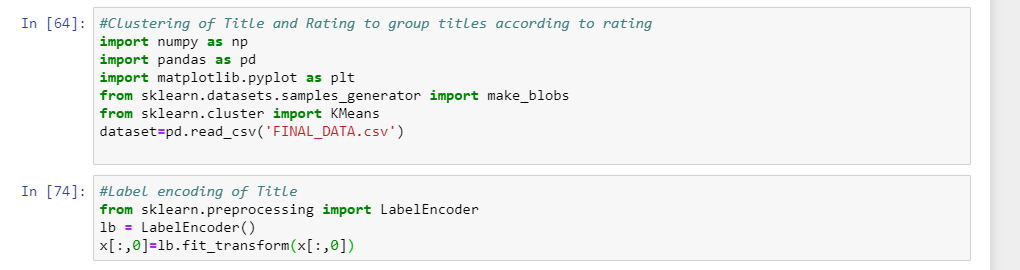


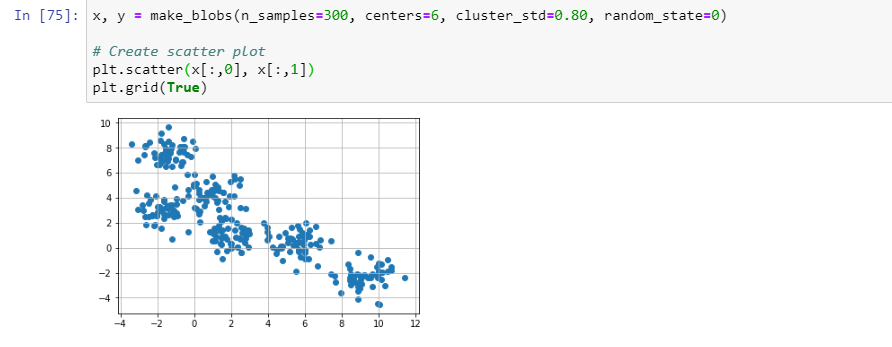
Mostly preferred actor for a particular director.

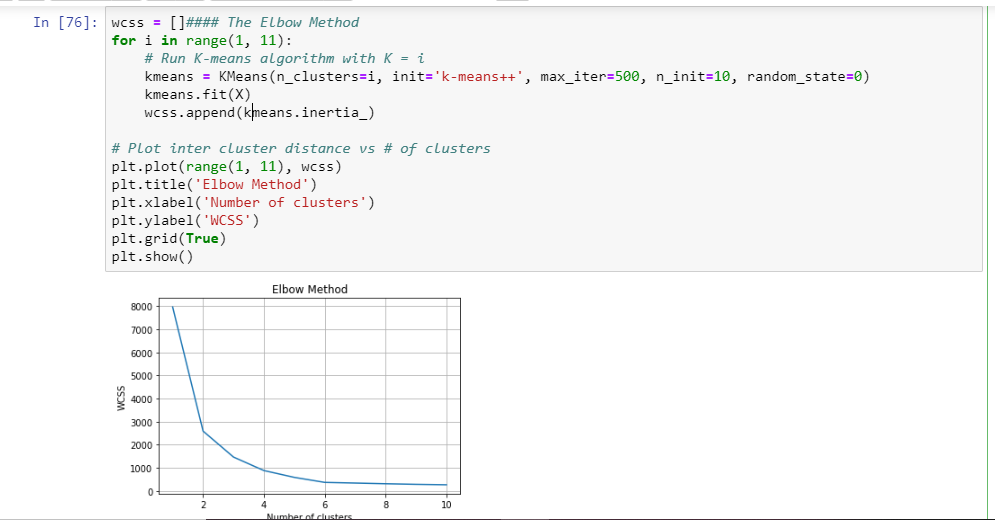


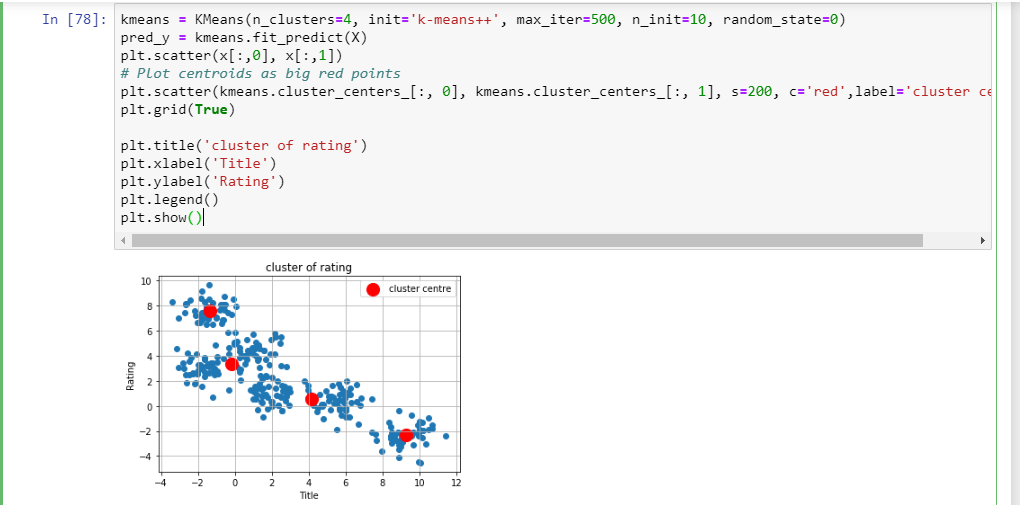
We have also implemented the Kmeans algorithm for clustering different categories of Genre and Actors.

Here, We have grouped different ratings according to title and for same group of ratings, User should prefer which movie to watch is suggested.









**5**

**CONCLUSION**

Thus, From this project what we targetted to achieve was successfully completed by us and we were able to map top actors based on rating and Genre type, director-actor correlation and top directors also.