**AIT 580 DL 4: Big Data to Information**

**Spring 2022**

Name: Deepti Vijay Khandagale

G Number: G01353007

Assignment 1: Examine Movie Dataset

1. Identify the data type for each of the attribute/column, and explain your reason of selecting data type for each attribute/column.

|  |  |  |
| --- | --- | --- |
| **Column name/Attributes** | **Data type** | **Reasons** |
| director\_name | Nominal | Director name column contains the name of the director of the movies. There is no order or size of data, it is a nominal data type |
| duration | Ratio | Duration column contains the length of the movie in minutes. As the data is comparable and measurable, it is ratio data type |
| actor\_2\_name | Nominal | Actor 2 name column contains the name of the second actor in the movie. There is no order or size of data, it is a nominal data type |
| genres | Nominal | Genres column contains the category of the movie. There is no order or size, it is a nominal data type |
| actor\_1\_name | Nominal | Actor 1 name column contains the name of the first actor in the movie. There is no order or size, it is a nominal data type |
| movie\_title | Nominal | Movie tittle column contains the name of the movie. There is no order or size, it is a nominal data type |
| num\_voted\_users | Ratio | Number of voted users column contains the count of votes by the user which is measurable and continuous TAs the data is measurable and continuous, it is ratio data type |
| actor\_3\_name | Nominal | Actor 3 name column contains the name of the third actor in the movie. There is no order or size of data, it is a nominal data type |
| movie\_imdb\_link | Nominal | Movie IMDB link column contains the address where the assigned movie will open and is discrete for each movie, hence it is a nominal data type |
| num\_user\_for\_reviews | Ratio | Number of users for reviews column contains the count of users who has reviewed the movie. As the data is measurable and continuous, it is a ratio data type |
| language | Nominal | Language column contains the language type used in the movie. There is no order or size of data, it is a nominal data type |
| country | Nominal | Country column contains the country where the movie is produced. There is no order or size of data, it is a nominal data type |
| title\_year | Interval | Title year column contains the release date of the movie. As the data is continuous, it is a interval data type |
| imdb\_score | Ordinal | IMDB score column contains the rating of the movie measured on a scale. As the data is ordered and discrete, it is ordinal data type |

1. Give one sample value for each data type from the dataset

|  |  |  |
| --- | --- | --- |
| **Column Name/Attributes** | **Data type** | **Sample Value** |
| director\_name | Nominal | Gore Verbinski |
| duration | Ratio | 150 |
| actor\_2\_name | Nominal | Ruth Wilson |
| genres | Nominal | Action|Adventure|Western |
| actor\_1\_name | Nominal | Johnny Depp |
| movie\_title | Nominal | The Lone Ranger¬† |
| num\_voted\_users | Ratio | 181792 |
| actor\_3\_name | Nominal | Tom Wilkinson |
| movie\_imdb\_link | Nominal | http://www.imdb.com/title/tt1210819/?ref\_=fn\_tt\_tt\_1 |
| num\_user\_for\_reviews | Ratio | 711 |
| language | Nominal | English |
| country | Nominal | USA |
| title\_year | Interval | 2013 |
| imdb\_score | Ordinal | 6.5 |