Movie streaming datasets (iflix datathon 2019 challenge datasets)

Modern Television: Segmenting audience in the new environment of video on demand entertainment  
  
Increasingly businesses today are being forced to implement a more customer centric approach to their business strategy. This is in response to rapidly changing consumer behavior and the compelling move to digital platforms. To be able to respond and make better, faster decisions, business need a clear view of their decision-making strategies and the ability to apply risk analytics, strategy improvements, automation, and advanced analytics. Before a company can personalize experiences for customers, it must first have a holistic view of the customer, which can often be a challenge. (From iflix datathon 2019)

Content

Metadata

users.csv  
**Userid** – Unique identifier of user  
**Countrycode** – Country code where user registered

assets.csv  
**Showtype** – Type of content, whether the asset is a movie or an episode of a TV series  
**Genre** – Genre of content  
**Runningmiutes** – Runtime of content, (Playable number of minutes)  
**Sourcelanguage** – Production language of content  
**Assetid** – Unique identifier of video content, at the most granular level (a movie or an episode of a TV series)  
**Seasonid** – Unique identifier of content, at season level. Only applicable to TV series  
**Seriesid** – Unique identifier of content, at series level. Only applicable to TV series  
**Studio**\_id – Unique identifier of production studio for the content

plays.csv  
**Platform** – Platform of consumption  
**Minutes\_viewed** – Total number of minutes viewed, rounded to the nearest integer (0 means less than 30 seconds)

Demographics / psychographics  
Dataset identifies psychographic and demographic tags about some iflix users. For each user-tag pair, there is an associated confidence score (1 is the highest and 0 is the lowest confidence). Each trait can consist of up to 3 levels, depending on its granularity. Some traits can be identified by only considering the first 2 levels, while there are others which make more sense when all the 3 levels are considered, e.g., ‘iflix Viewing Behaviour’ is a level 2 psychographic trait which only makes sense when it is looked at in combination with the level 3 traits corresponding to it (‘casual’, ‘player’ and ‘addict’). These traits represent different levels of viewing behavior of iflix users. Casual users have less than 5 viewing days in a month, player users have 5-12 viewing days in a month, and addicts have more than 12 viewing days in a month. Traits are available corresponding to a user\_id in the dataset only if we have certain confidence that the user belongs to the trait.

Column and Description  
  
**Level1**: Identifies the first level of the trait (psychologic or demographic)  
**Level2**: Identifies the second level of the trait (e.g., Music Lovers, Movies Lovers)  
**Level3**: Identifies the third level of the trait, if available/relevant (e.g. Malay Movies Lovers, Indonesian TV Fans)  
**Confidencescore**: Confidence in associating the said trait (level1, level2, level\_3) with the user

Acknowledgements

These movie streaming datasets wouldn't be here without help of iflix. Thanks all for viewing, creating and practicing Kaggle kernels with these datasets. Thanks again for students and researchers for using the datasets in related areas.