*How do these datasets fulfill the requirements:*

1. at least 3 different data sets

We want to integrate five datasets so that requirement is fulfilled.

2. at least 2,500 entities described in total (in joint dataset) (− but more are better, good: >10,000 but <100,000)

**1) Top video Games 1995-2021 Metacritic**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **name** | **platform** | **release\_date** | **summary** | **meta\_score** | **user\_review** |
| 18.8k | 18.8k | 18.8k | 18.7k | 18.8k | 18.8k |

**2) Video Games List**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **title** | **released** | **deverlopers** | **publisher** | **genre** |
| 3817 | 3817 | 3813 | 3749 | 3817 |

**3) Video Game Sales and Ratings**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Platform** | **Year\_of\_Release** | **Genre** | **Publisher** | **NA\_Sales** | **EU\_Sales** | **JP\_Sales** | **Other\_Sales** | **Global\_Sales** |
| 17.4k | 17.4k | 17.4k | 17.4k | 17.4k | 17.4k | 17.4k | 17.4k | 17.4k | 17.4k |
| **Critic\_Score** | **Critic\_Count** | **User\_Score** | **User\_Count** | **Rating** |  |  |  |  |  |
| 8336 | 8336 | 7798 | 7798 | 10.3k |  |  |  |  |  |

**4) All Playstation 4 Games**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **GamePSID** | **GameName** | **Publisher** | **ReleaseYear** | **ReleaseDate** | **Deverloper** | **genre** | **Features** | **size** |
| 9686 | 9686 | 9686 | 8964 | 8952 | 8952 | 8987 | 8898 | 1749 | 9686 |
| **Medium** | **Hardware** | **CompletionTime**  **(Hours)** | **Is\_Digital game** | **Is\_Physical game** | **OfficialWebsite** |  |  |  |  |
| 8937 | 1512 | 9686 | 9686 | 9686 | 8089 |  |  |  |  |

**5) Wikidata**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **itemLabel** | **publishdates** | **platformLabel** | **esrbratingLabel** | **seriesLabel** |
| 90151 | 90151 | 90151 | 90151 | 29765 | 23147 |

3. at least 1,000 entities should be contained in at least two datasets (− please estimate based on small sample)

1. **All Playstation 4 Games (9671)**
2. **Video Game List(3803)**

“All playstation 4 Games” has 9671 unique values, and “Video Game List” has 3803 unique values. In these data set, we can find same attributes (title, released year, developer, publisher, genres) and in these attributes, they have more than 1000 entities.

|  |  |  |
| --- | --- | --- |
|  | **All playstation4 Games** | **Video Game List** |
| **Game name / title** | 9686 | 3817 |
| **Publisher / Publishers** | 8964 | 3749 |
| **Release Year / released** | 8952 | 3817 |
| **Developer / developers** | 8987 | 3813 |
| **Genre / genres** | 8898 | 3817 |

4. at least 8 attributes in joint dataset  (− entities should be identifiable by attribute combinations of at least two attributes, e.g. name+birthdate )

> ReleaseYear + RealeaseDate ,

Problem: I think we need to talk about which attributes and dataset will be combined and used. So I’d like to ask you guys’s opinion. 😊

1. **All Playstation 4 Games**

Attribute: ID   GamePSID   GameName   Publisher   ReleaseYear   ReleaseDate   Developer   Genre   Features   Size   Medium   Hardware   CompletionTime(Hours)   is\_Digital\_game   is\_Physical\_game   OfficialWebsite

1. **Video Games sale and Ratings**

Attribute: Name   Platform   Year\_of\_Release   Genre   Publisher   NA\_Sales   EU\_Sales   JP\_Sales   Other\_Sales   Global\_Sales   Critic\_Score   Critic\_Count   User\_Score   User\_Count   Rating

1. **Top Video Games 1995-2021 Metacritic**

Attribute: name   platform   release\_date   summary   meta\_score   user\_review

1. **Video Game List**

Attribute: titles   released   developers   publishers   genres

1. **Wikidata**

Attribute: Item   itemLabel   publishdates   platformLabel   esrbratingLabel   seriesLabel

Note: “ID” from the All Playstation 4 Games dataset and “Item” from the Wikidata dataset are not going to be used as an unique identifier.

5. at least 5 attributes should be contained in at least two datasets

− some attributes (other than name) should be contained in three datasets (for fusion by voting)

Problem : We need 5 attributes but, apart from name( game name, titles), we only have 4 attributes( [1,4] publisher, releseyearm developer,genre)…   ( please see above)

6. ideally, at least one of your attributes is a list attribute

> CD developers of a video game, genres of a video game, publisher of a video game