# Information Visualization

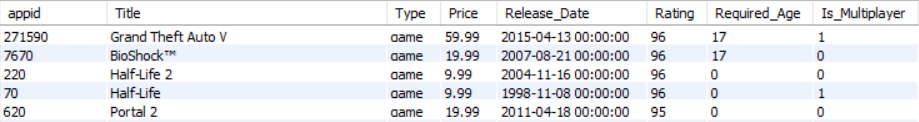
# CHECKPOINT II: Data cleaning and processing

G14 - A

**1. Initial Dataset**

A maioria dos nossos dados (120GB) contem informações sobre os jogos, os developers, os publishers, as categorias dos jogos, os utilizadores e os seus amigos da steam. Esses dados estavam no formato de instruções de SQL. Além disso, também retiramos algumas tabelas no formato .csv do steamspy com o objetivo de completar dados. Também temos tabelas com dados sociodemográficos no formato .csv.

|  |  |  |  |
| --- | --- | --- | --- |
| **Origem** | **Tabelas** | **Atributos** | **Linhas** |
| Steam Dataset | app\_id\_info | |AppID| Titulo|Tipo|Preço|Data|Rating|Idade|Multiplayer| | 26172 |
| friends | |steamid\_a|steamid\_b|relationship|friends\_since |  |
| games\_developers | |AppID|Developer| | 7845 |
| games\_publishers | |AppID|Publisher| | 7512 |
| Player\_summaries | |ID\_user|Last\_login|Group|Creation|Country|Estate|City| | 99344566 |
| SteamSpy | publishers | |GameID|Publisher| | 17999 |
| developers | |GameID|Developer| | 17999 |
| Our World In Data | number-of-homicide-deaths | |Entity|Code|Year|Homicide (deaths)| | 6157 |
| share-with-depression | |Entity|Code|Year|Prevalence| | 6157 |
| share\_of\_adults\_defined\_as\_  \_obese | |Entity|Code|Year|Prevalence| | 8023 |
| share\_of\_individuals\_  using\_the\_internet | |Entity|Code|Year|Individuals| | 6157 |

****Amostra de dados:

**2. Selected/Derived Data**

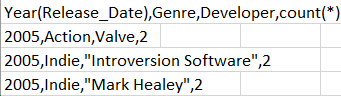
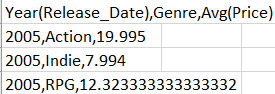
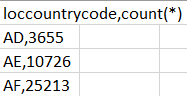
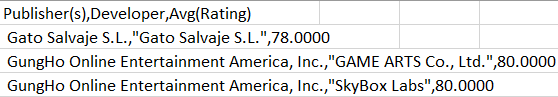
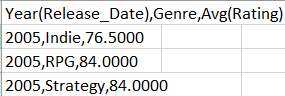
Serão utilizados quase todos os dados que recolhemos a menos de algumas colunas que não utilizaremos pelo que foram eliminadas. Algumas tabelas incluídas no nosso conjunto de dados não foram utilizadas. Baseados nas nossas questões e tarefas, derivamos o número médio de amigos por jogador em cada país por fazer o join da tabela dos **player\_summaries** e os **friends**, o número de utilizadores por país utilizando a tabela dos **player\_summaries**, o número de jogos por cada categoria por ano, o rating médio dos jogos por cada categoria por ano e o preço médio dos jogos por cada categoria por ano através da tabela do **app\_id\_info** e os **games\_genres** e as tabelas dos jogos lançados por ano disponibilizado no steamspy, o número de publishers por cada categoria por ano por utilizar o **app\_id\_info,** os **games\_genres,** os **games\_publishers** e as tabelas dos jogos lançados por ano disponibilizado no steamspy, o número de developers por cada categoria por ano através do **app\_id\_info,** os **games\_genres** eos **games\_publishers**, o rating médio entre publishers e developers, o rating medio entre publishers e developers por ano e o número de jogos totais lançados por publisher e developer por utilizar o resultado de algumas tabelas obtidas anteriormente.

* + - 1. **3. Data abstraction**

Os tipos dos nossos dados são tabelas. E cada tabela tem os seguintes itens e atributos:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | -Year(Release\_Date) | -AVGrating\_Develop\_Publish | -Genre | -Avg(Price) | -Avg(Rating) | -Developer | -N\_Developers | -N\_Publishers | loccodecountry | count(\*)[New\_users\_country] | year(timecreated) | count(\*)[New\_users\_year] | count(\*)[New\_users\_year\_country] | -Publisher | -NGames\_Genre\_Year | -Ngames\_publisher | -Ngames\_developer | obesity | Homicide (deaths) | Individuals | Prevalence | friends\_since | count(\*)[Friends\_per\_year] |
| Nominal |  |  | x |  |  | x |  |  | x |  |  |  |  | x |  |  |  |  |  |  |  |  |  |
| Ordenado | x | x |  | x | x |  | x | x |  | x | x | x | x |  | x | x | x | x | x | x | x | x | x |
| Quantitativo | x | x |  | x | x |  | x | x |  | x | x | x | x |  | x | x | x | x | x | x | x | x | x |
| Ratio | x | x |  | x | x |  | x | x |  | x | x | x | x |  | x | x | x | x | x | x | x | x | x |

**4. Dataset processing**

* + - 1. Nós tratámos os dados por fazer várias queries em mysql de modo a ficarmos só com as colunas que nos interessavam e com os atributos que desejámos para as responder às questões propostas. Para facilitar a inserção nas tabelas fizemos um script em php, uma vez que eram mais de 91000 de instruções de inserção. A tabela **app\_id\_info** não tinha informação suficiente sobre o rating e o publisher pelo que juntámos os dados das tabelas disponibilizadas pelo steamspy. Para isso importámos os ficheiros do steamspy para a nossa base dados em sql. Para processar os dados fizemos um conjunto de queries de modo a ficarmos apenas com a informação pretendida para as perguntas que queremos responder. Alguns utilizadores não tinham o código do país pelo que achámos mais apropriado descartar esses utilizadores uma vez que seria difícil atribuir o valor do item mais próximo ou estatisticamente escolher qual o melhor valor a atribuir devido à grande quantidade de dados que tínhamos. No fim, exportámos os dados finais no formato .csv**.**
      2. **5. Mapping (Data sample / Questions)**

Qual o número de jogos realizados por um Developer num certo ano?

Qual o rating médio por genero por ano?

Qual o rating médio dos jogos realizados por um Developer e publicados por um Publisher?

Qual o preço médio por genero por ano?

Qual o número de utilizadores por país?