|  |  |  |
| --- | --- | --- |
| Checkpoint II | Checkpoint II: Data Cleaning & Processing | |
| Group: | G11 |
| Date: | 2022/09/28 |
|  |  |

# Initial Dataset

19 joint datasets, all regarding information related to the competitive Pokémon video game, with particular reference to the period of February - August 2022. Every dataset has mostly a tabular type, except for some attributes that follow a network type like dataset.

The total size of the datasets was (#Items x #Attributes):

20 \* 2 + 1099 \* 22 + 25 \* 5 + 822 \* 9 + 840 \* 7 + 268 \* 3 + 325 \* 4 + 1686 \* 2 + 3411 \* 3 + 153 \* 2 + 555 \* 2 + 772 \* 3 + 6819 \* 4 + 69105 \* 2 + 3271 \* 4 + 1892 \* 3 + 2176 \* 3 + 822 \* 2 + 33 \* 2 = 249546

Data sample:  
(from "bridge\_pokemon\_USED\_IN\_TEAMS\_WITH.csv")

Use\_Percentage(%); Pokemon; Teammate

25.902%; Pikachu; Incineroar

(from “df\_moves.csv”)

Name; [...]; Power; Acc.; PP; Damage

Double Slap; [...]; 15; -; 10 [...]; -

# Selected/Derived Data

Data selected:

The selected data is already described in the “Data Abstraction” section, along with its abstraction.

Data derived (CP-I related):

* Frequency of each Pokémon type combination in a team
* Frequency of each Type in a team
* Maximum of the Speed Stat for a given Type within two given generations
* Frequency of each Move in a Pokémon combination
* PP, Accuracy, and Power percentile of a Move + Move frequency of Use

# Data Abstraction

Dataset type: table. All 5 files contain information that can be itemized in a single line; while attributes correlate to attributes on different tables, they do not do so in a way that allows creation of a network.

|  |  |
| --- | --- |
| Main Item | Semantics |
| Pokémon | Pokémon, who are equipped with up to 4 moves and battle in teams of 6 |
| Move | Moves that can be used in battle by Pokémon. |

|  |  |  |
| --- | --- | --- |
| Data | Category | Semantics |
| Stats (Total, HP, …)| Monthly Usage(k), Use Percentage | Ratio | Value of a given statistic of a Pokémon | Frequency of use of a Pokémon (approx. to the thousands) in competitive battle | Use percentage of a [Move per Pokémon/Pokémon in a team with another Pokémon]. |
| ID, Name, Move | VG2022\_rules | Nominal | Keys of a Pokémon (ID, Name) or Move (Move). | Usage rules of a Pokémon in competitive battle. |
| Species, Type, Damage Class | Nominal | Categories of a Pokémon (Species, Type) or Move (Type, Damage Class). |
| Generation | Ordinal | Era of Pokémon Games associated; can be used to build timeline. |
| Power, Acc., PP. | Ratio | Attributes of a Move (Power, Accuracy, Maximum Uses). |
| Safety (derived) | Ratio | Value that measures a move’s Power versus likelihood to be used. Safety = Power\*Acc/100 \* PP/40 |
| Averages/Max/Min (derived) | Ratio | Average/Maximum statistics of a given Pokémon (total, per type, per generation, …). Most used Type combination. |
| Frequency (derived) | Ratio | Frequency of use of a given Pokémon/Move (per Type, Generation, Damage Class, …). Frequency of use of a given Type. Frequency of use of a given Type combination in teams. |
| Percentile (derived) | Ratio | Percentile of a Move’s Power/Accuracy/PP. |

# Data Processing

Dataset and attributes that were less relevant to the visualization (due to lack of interest) were removed. For missing Nominal values, we used a Sentinel value of “NULL” (if the attribute was non-applicable), or a researched Impute value otherwise. For missing Ratio values, we used Impute values (ex: “0” for the Monthly Usage(k) attribute) or Sentinel values (ex: “101” for the Acc. Attribute when infinite). We used the Pokémon Name and Type as cross-reference keys among the different datasets, particularly to compute the frequency of use of a Pokémon multiplied by the Use Percentage of – for example – its moves. All items were accounted for: for Nominal attributes, we considered that all values should be taken into account; for Ratio attributes, we didn't identify any outliers when using Standard Deviation method. Ratio values were normalized for comparison’s sake.

# Mapping (Data sample/Questions)

|  |  |
| --- | --- |
| Question | Data |
| For two Pokémon who are teammates, which are the most often used Pokémon type combinations between each teammate? | (“deriv\_pokemon\_combos\_type.csv”)  TypeA; TypeB; Battle\_Frequency  Electric; Water; 240 |
| What is the most used teammate Pokémon, for the Pokémon “Pikachu", in each generation? | (“bridge\_pokemon\_pokemon\_USED\_IN\_TEAMS\_WITH.csv”)  Use\_Percentage; Pokemon; Teammate  25.902; Pikachu; Incineroar |
| Are Electric-Type Pokémon used more often than Grass-type Pokémon? | (“df\_pokemon.csv”)  name; [...]; Type1; Type2; [...]; Monthly Usage  Pikachu; [...]; Electric; NULL; [...]; 250  (d3 allows for grouping of multiple attributes and sum of their values so we do not need to compute the sum to a new table) |
| What are the Fire-Type Pokémon with the highest Base Speed Stat between generation 6 and 8? | (“pokemon.csv”)  [...];generation;[...];Type1;Type2;[...];Speed;[...]  [...];8;[...];Fire;Dragon;[...];140; [...]; |
| What is the most used ~~item~~ move in Fire-Type/Water-Type team combinations? | (“deriv\_moves\_teams.csv”)  Type1; Type2; Move; […]  Water; Electric; Water Gun; […] |
| Do competitive players prioritize move availability (PP) or move power, when choosing moves for their Pokémon? | (“df\_moves.csv”)  Move;Battle\_Frequency;[…];pp\_percentile;power\_percentile  Aqua Tail; 220; […]; 0.35; 0.60; 0.20 |