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

# Initial Dataset

Initial Dataset:

Our dataset initially had 19 different datasets which contained information about the Pokémon world and it's characteristics and also some data about usage amounts of the different Pokémon, moves, items and synergies on the competitive Pokémon Video Game Championship in the period of February - August 2022. Every Dataset has mostly a tabular type, except for some attributes that follow a network type like dataset.

Datasets:

- df\_types.csv: 20 items, 2 attributes

- df\_pokemon.csv: 1099 items, 22 attributes

- df\_natures.csv: 25 items, 5 attributes

- df\_moves.csv: 822 items, 9 attributes

- df\_items.csv: 840 items, 7 attributes

- df\_abilities.csv: 268 items, 3 attributes

- bridge\_type\_type\_MOVE\_EFFECTIVENESS\_ON\_POKEMON.csv: 325 items, 4 attributes

- bridge\_pokemon\_type\_IS\_OF\_TYPE.csv: 1686 items, 2 attributes

- bridge\_pokemon\_pokemon\_USED\_IN\_TEAM\_WITH.csv: 3411 items, 3 attributes

- bridge\_pokemon\_pokemon\_HAS\_VARIANT.csv: 153 items, 2 attributes

- bridge\_pokemon\_pokemon\_EVOLVES\_FROM.csv: 555 items, 2 attributes

- bridge\_pokemon\_nature\_USED\_WITH\_ABILITY.csv: 772 items, 3 attributes

- bridge\_pokemon\_nature\_COMMON\_SPREAD.csv: 6819 items, 4 attributes

- bridge\_pokemon\_moves\_MAY\_LEARN.csv: 69105 items, 2 attributes

- bridge\_pokemon\_move\_USED\_WITH\_MOVE.csv: 3271 items, 4 attributes

- bridge\_pokemon\_item\_USED\_WITH\_ITEM.csv: 1892 items, 3 attributes

- bridge\_pokemon\_ability\_MAY\_HAS.csv: 2176 items, 3 attributes

- bridge\_move\_type\_MOVES\_IS\_TYPE.csv: 822 items, 2 attributes

- bridge\_moves\_pokemon\_GMAX\_MOVE.csv: 33 items, 2 attributes

Raw Data Sample:

(from "df\_pokemon.csv")

id; name; [...]; generation; [...]; Type1; Type2; [...]; HP; [...]; Speed; [...]

890; Eternatus; [...]; 8; [...]; Poison; Dragon; [...]; 140; [...]; 130; [...]

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

Use\_Percentage(%); Pokemon; Teammate

25.902%; Pikachu; Incineroar

(from “df\_moves.csv”)

Name; [...]; Power; [...]; PP; [...]

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

(from “bridge\_pokemon\_move\_USED\_WITH\_MOVE.csv”)

Move; Use\_Percentage(%); Pokemon; Name

Protect; 98.672%; Zacian Crowned Sword; Protect

# Selected/Derived Data

Data selected:

* Use percentage of Item/Move for each Pokémon (Table: bridge\_pokemon\_item\_USED\_WITH\_[ITEM/MOVE].csv)
* Moves a Pokémon may learn (Table: bridge\_pokemon\_moves\_MAY\_LEARN.csv)
* Use percentage of a teammate Pokémon for each Pokémon (Table: bridge\_pokemon\_pokemon\_USED\_IN\_TEAM\_WITH.csv)
* Move attributes, specifically its Type, Power, Accuracy, PP, and Damage\_class (Table: df\_moves.csv)
* Pokémon attributes, specifically its ID, Name, Species, Generation, Type[1/2], Stats(Total and individual), Allowed/Permitted flag, and Monthly Usage (Table: df\_pokemon.csv)

Data derived (CP-I related):

* Frequency of each Pokémon type combination in a team
* Frequency of each item in a Pokémon team
* Frequency of each type in a team
* Maximum of the Speed Stat for a given type within two given generations
* Frequency of each Item in a Pokémon combination
* Ratio of (normalized) availability per (normalized) move power, according to frequency of move with a Pokémon and the frequency of each Pokémon

# Data Abstraction

Dataset type: table.

|  |  |  |
| --- | --- | --- |
| Data | Category | Meaning |
| Stats | Ratio | Level of a given statistic |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

On the df\_moves.csv we created a derivative measured called “Safeness”, its goal is to compare the riskiness between each move, usually the lower the value, the less times you can use it and/or when it is used there is a medium/small change that the move was not successfully hit to an opponent pokemon, however a lower value, also means when it’s successfully displayed, it might do high damage to an opponent Pokémon comparing to a safeness with higher value.

Safeness=Power\*Acc/100 \* PP/40

Note: Every move has the interval between [1,40], G-Max moves, and Power and Accuracy moves that have a null value were not considered.

Data abstraction description:

* The current dataset has Description of the dataset type (spatial, table, field, etc.).
* Description of each item and attribute (nominal/ordinal/etc., diverging/sequential scale, etc.).
* Semantics (what does each attribute and item stand for).

We recommend using a table to save some space. Do not forget to include all variables.

# Data Processing

# The group removed some datasets that we didn't think were really relevant to the visualization and also some attributes, due to the same reason. In some attributes we found a few missing values, most of them were missing because the item had no value for that specific attribute, and we just added a NULL Sentinel value in those cases. In the df\_pokemon.csv file on the "VGC2022rules" Nominal attribute, some Pokémon didn't have a value, due to not being used in competitive play, but we added a "Permitted" value (Impute Value) since they're still permitted. In the same file but on the Monthly Usage (k) attribute, for the same reason, most Pokémon had no value and we changed it for a 0 Impute value. On the dataset bridge\_type\_type\_MOVE\_EFFECTIVENESS\_ONPOKEMON.csv we changed the missing values of the Damage Multiplier attribute to 1 since it's meaning is having no multiplication factor by the Impute Value method. We used the Pokémon Name and Type as cross-reference keys among the different datasets. On Nominal attributes we didn't search for outliers since all values should be taken account of. On Ratio attributes we didn't identify any outliers when we used the ... method.

# Mapping (Data sample/Questions)

Some examples that show that with your data sample you will be able to provide the answers to the questions you formulated. Include all questions from CP1.