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

# Initial Dataset

# 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.

# 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 |
| bridge\_pokemon\_moves\_MAY\_LEARN | Move, Pokemon |  |
| bridge\_pokemon\_pokemon\_USED\_IN\_TEAM\_WITH | Use %, Pokemon, Teammate |  |
| bridge\_type\_type\_MOVE\_EFFECTIVENESS\_ON\_POKEMON | Atk. Move Type, Def. Pokemon Type, Damage Multiplier |  |
| df\_moves | Name, Type, Power, Acc., PP, Damage\_class |  |
| df\_pokemon | ID, Name, Species, Variant, Generation, Evolves From, Type[1/2], Stats(Total and individual), Allowed/Permitted flag, Monthly Usage |  |

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

Description of how the dataset was processed (cleaned, problems found and solutions, how did you fix missing values, cross-referenced different tables/datasets, etc.).

# 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.