Philipps Universität Marburg

Step 3: Cleaning
TODO

https://github.com/Paralian/umr-data-integration-project-the-TODO-team.git

Juan Fernando Maya
Ilya Zykov
Miron Brandeis

Integration of RPG Entities
Before cleaning

Philipps Universität Marburg

- 3 **universes**
- 7867 Total entities
- 7769 Unique **names**
  - → 98 duplicate entities possible
  - → Top duplicated entity: "Dark Magician", 9 times!
- **Name**, **Type** and **Kind** are strings: cleaning is only relevant for elimination of duplicates
- **development_stage**, **vitality** and **attack** are decimal values
  - → first two already normalized
  - → attack goes up to 5000
- **Harmful** and **universe** are metadata, need to preserve as-is

```
DF_out.describe(include='all')
```

	Unnamed: 0	name	type	kind	development_stage	vitality	attack	harmful	universe
count	7867.000000	7867	7867	7867	5608.000000	5593.000000	5571.000000	7867	7867
unique	NaN	7769	114	83	NaN	NaN	NaN	2	3
top	NaN	Dark Magician	Effect Monster	Normal	NaN	NaN	NaN	True	yugioh
freq	NaN	9	2494	1126	NaN	NaN	NaN	4647	6534
mean	3933.000000	NaN	NaN	NaN	0.348070	0.281558	1116.989254	NaN	NaN
std	2271.151617	NaN	NaN	NaN	0.258139	0.185583	999.276734	NaN	NaN
min	0.000000	NaN	NaN	NaN	0.001232	0.000000	0.000000	NaN	NaN
25%	1966.500000	NaN	NaN	NaN	0.166667	0.100000	0.510000	NaN	NaN
50%	3933.000000	NaN	NaN	NaN	0.333333	0.300000	1000.000000	NaN	NaN
75%	5899.500000	NaN	NaN	NaN	0.500000	0.400000	1800.000000	NaN	NaN
max	7866.000000	NaN	NaN	NaN	1.000000	1.000000	5000.000000	NaN	NaN

Pipeline

- Finding inconsistencies in groups: each universe has its own system of values

```
dfc = DF_out
universes = dfc.groupby('universe')
```

```
yugioh = universes.get_group("yugioh")
yugioh.describe(include='all')
```

```
•••
```

```
dd5 = universes.get_group("dd5")
dd5.describe(include='all')
```

```
•••
```

```
skyrim = universes.get_group("skyrim")
skyrim.describe(include='all')
```

```
•••
```

Pipeline

- Most duplicates are in the yugioh universe:
  - 6449 unique names out of 6534 entries
    → 85 duplicates
    (from total 98 duplicates)
- Attack values are missing for 2259 entities (~35% of data) → replace *null* with 0 as possible remediation

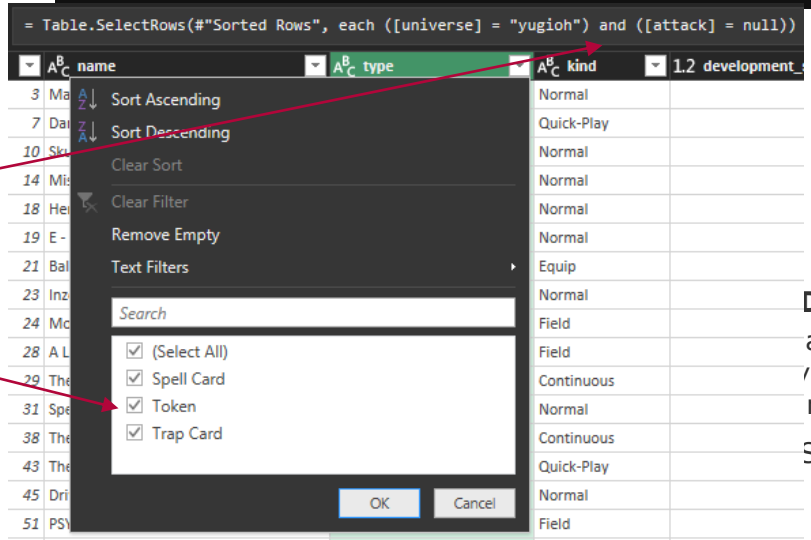
```
dfc['attack'].fillna(value=0)
```

  - All such entries are spell cards, trap cards and tokens → essential for the universe and non-deletable

```
yugioh = universes.get_group("yugioh")
yugioh.describe(include='all')
```

	Unnamed: 0	name	type	kind	development_stage	vitality	attack	harmful	universe
count	6534.000000	6534	6534	6534	4275.000000	4275.000000	4275.000000	6534	6534
unique	NaN	6449	24	31	NaN	NaN	NaN	2	1
top	NaN	Dark Magician	Effect Monster	Normal	NaN	NaN	NaN	True	yugioh
freq	NaN	9	2494	1126	NaN	NaN	NaN	3983	6534
mean	3266.500000	NaN	NaN	NaN	0.370799	0.331433	1455.553216	NaN	NaN
std	1886.347661	NaN	NaN	NaN	0.182087	0.166005	899.165365	NaN	NaN
min	0.000000	NaN	NaN	NaN	0.083333	0.000000	0.000000	NaN	NaN
25%	1633.250000	NaN	NaN	NaN	0.250000	0.220000	800.000000	NaN	NaN
50%	3266.500000	NaN	NaN	NaN	0.333333	0.340000	1500.000000	NaN	NaN
75%	4899.750000	NaN	NaN	NaN	0.500000	0.440000	2100.000000	NaN	NaN
max	6533.000000	NaN	NaN	NaN	1.000000	1.000000	5000.000000	NaN	NaN

```
= Table.SelectRows(#"Sorted Rows", each ([universe] = "yugioh") and ([attack] = null))
```



Pipeline

Philipps Universität Marburg

- D&D has all unique names
  - → no duplicates already
- Attack values missing for 6 out of 324 entities, need further investigation
- Attack values range from 0 to 0.836, need normalization

```
dd5 = universes.get_group("dd5")
dd5.describe(include='all')
```

	Unnamed: 0	name	type	kind	development_stage	vitality	attack	harmful	universe
count	324.00000	324	324	324	324.000000	324.000000	318.000000	324	324
unique	NaN	324	18	36	NaN	NaN	NaN	1	1
top	NaN	Aboleth	Unaligned	beast	NaN	NaN	NaN	True	dd5
freq	NaN	1	128	87	NaN	NaN	NaN	324	324
mean	7704.50000	NaN	NaN	NaN	0.212577	0.341015	0.534586	NaN	NaN
std	93.67497	NaN	NaN	NaN	0.232969	0.128190	0.135914	NaN	NaN
min	7543.00000	NaN	NaN	NaN	0.031250	0.109615	0.100000	NaN	NaN
25%	7623.75000	NaN	NaN	NaN	0.062500	0.254053	0.466667	NaN	NaN
50%	7704.50000	NaN	NaN	NaN	0.062500	0.304408	0.550000	NaN	NaN
75%	7785.25000	NaN	NaN	NaN	0.250000	0.400828	0.615833	NaN	NaN
max	7866.00000	NaN	NaN	NaN	1.000000	1.000000	0.836667	NaN	NaN

Pipeline

Philipps Universität Marburg

- Skyrim dataset has 1009 entries, 1000 are unique
  - Top duplicate count per name is 2, need to investigate
- Attack values need normalization

```
skyrim = universes.get_group("skyrim")
skyrim.describe(include='all')
```

	Unnamed: 0	name	type	kind	development_stage	vitality	attack	harmful	universe
count	1009.000000	1009	1009	1009	1009.000000	994.000000	978.000000	1009	1009
unique	NaN	1000	72	16	NaN	NaN	NaN	2	1
top	NaN	Nikulas	None	Nord	NaN	NaN	NaN	False	skyrim
freq	NaN	2	633	453	NaN	NaN	NaN	669	1009
mean	7038.000000	NaN	NaN	NaN	0.295275	0.047675	0.089094	NaN	NaN
std	291.417513	NaN	NaN	NaN	0.449228	0.058108	0.047466	NaN	NaN
min	6534.000000	NaN	NaN	NaN	0.001232	0.000000	0.009588	NaN	NaN
25%	6786.000000	NaN	NaN	NaN	0.004926	0.024143	0.069424	NaN	NaN
50%	7038.000000	NaN	NaN	NaN	0.007389	0.036214	0.083309	NaN	NaN
75%	7290.000000	NaN	NaN	NaN	0.997537	0.048286	0.093028	NaN	NaN
max	7542.000000	NaN	NaN	NaN	1.000000	1.000000	0.547939	NaN	NaN

Pipeline

- Let's set all entries in "kind" column to lowercase
- Then reduce the duplicates
  - How many are truly identical entities?
- Which are those?

```
dfc = DF_out
dfc['kind'] = dfc['kind'].str.lower()
```

```
dfc.duplicated(subset=['name', 'type', 'kind', 'development_stage', 'universe']).sum()
```

```
88
```

```
pd.set_option('display.max_rows', None)
dfc.sort_values(by=['name'], ascending=True, inplace=True)
dfc.loc[dfc.duplicated(keep=False, subset=['name', 'type', 'kind', 'development_stage']), :]
```

Pipeline

- …Which are those?

name	type	kind	development_stage	vitality	attack	harmful	universe
Acid Trap Hole	Trap Card	Normal	NaN	NaN	NaN	False	yugioh
Acid Trap Hole	Trap Card	Normal	NaN	NaN	NaN	False	yugioh
Arcana Knight Joker	Fusion Monster	Warrior	0.750000	0.760000	3800.000000	True	yugioh
Arcana Knight Joker	Fusion Monster	Warrior	0.750000	0.760000	3800.000000	True	yugioh
Barknar	None	Nord	0.001232	0.024143	0.069424	False	skyrim
Barknar	None	Nord	0.001232	0.024143	0.069424	False	skyrim
Blue-Eyes Ultimate Dragon	Fusion Monster	Dragon	1.000000	0.900000	4500.000000	True	yugioh
Blue-Eyes Ultimate Dragon	Fusion Monster	Dragon	1.000000	0.900000	4500.000000	True	yugioh
Crush Card Virus	Trap Card	Normal	NaN	NaN	NaN	False	yugioh
Crush Card Virus	Trap Card	Normal	NaN	NaN	NaN	False	yugioh
Cyber Dragon	Effect Monster	Machine	0.416667	0.420000	2100.000000	True	yugioh
Cyber Dragon	Effect Monster	Machine	0.416667	0.420000	2100.000000	True	yugioh
Cyber End Dragon	Fusion Monster	Machine	0.833333	0.800000	4000.000000	True	yugioh
Cyber End Dragon	Fusion Monster	Machine	0.833333	0.800000	4000.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician	Normal Monster	Spellcaster	0.583333	0.500000	2500.000000	True	yugioh
Dark Magician Girl	Effect Monster	Spellcaster	0.500000	0.400000	2000.000000	True	yugioh
Dark Magician Girl	Effect Monster	Spellcaster	0.500000	0.400000	2000.000000	True	yugioh
Dark Magician Girl	Effect Monster	Spellcaster	0.500000	0.400000	2000.000000	True	yugioh
Dark Magician Girl	Effect Monster	Spellcaster	0.500000	0.400000	2000.000000	True	yugioh
Dark Magician Girl	Effect Monster	Spellcaster	0.500000	0.400000	2000.000000	True	yugioh
Dark Magician Girl	Effect Monster	Spellcaster	0.500000	0.400000	2000.000000	True	yugioh
Dark Paladin	Fusion Monster	Spellcaster	0.666667	0.580000	2900.000000	True	yugioh
Dark Paladin	Fusion Monster	Spellcaster	0.666667	0.580000	2900.000000	True	yugioh
Dark Paladin	Fusion Monster	Spellcaster	0.666667	0.580000	2900.000000	True	yugioh
Dark Rebellion Xyz Dragon	XYZ Monster	Dragon	0.333333	0.500000	2500.000000	True	yugioh
Dark Rebellion Xyz Dragon	XYZ Monster	Dragon	0.333333	0.500000	2500.000000	True	yugioh
Doomsday Token	Token	Fiend	NaN	NaN	NaN	False	yugioh
Doomsday Token	Token	Fiend	NaN	NaN	NaN	False	yugioh
Elemental HERO Avian	Normal Monster	Warrior	0.250000	0.200000	1000.000000	True	yugioh
Elemental HERO Avian	Normal Monster	Warrior	0.250000	0.200000	1000.000000	True	yugioh

…and many more

Pipeline

Philipps Universität Marburg

- Dealing with duplicates

```
dfc.sort_values(by=['name'], ascending=True, inplace=True)
```

```
dfc.drop_duplicates(subset=['name', 'type', 'kind', 'development_stage', 'universe'], keep='first', inplace=True)
```

```
dfc.duplicated(subset=['name', 'type', 'kind', 'development_stage', 'universe']).sum()
```

```
0
```

Duplicate Removal results

- Check for remaining similar entities
- Result: entities with similar names but from different universes
- Only 20 entities!

```
dfc.duplicated(subset=['name']).sum()
dfc.loc[dfc.duplicated(keep=False, subset=['name']), :]
```

	Unnamed: 0	name	type	kind	development_stage	vitality	attack	harmful	universe
5785	5785	Bat	Normal Monster	machine	0.083333	0.070000	300.000000	True	yugioh
7583	7583	Bat	Unaligned	beast	0.031250	0.240740	0.281667	True	dd5
5331	5331	Doppelganger	Trap Card	continuous	NaN	NaN	NaN	False	yugioh
7626	7626	Doppelganger	Unaligned	monstrosity (shapechanger)	0.062500	0.318462	0.530000	True	dd5
7396	7396	Eydis	EncClassBanditMelee	nord	0.030788	0.000000	0.045455	True	skyrim
...	...	...	...	...	...	...	...	...	...
7193	7193	Nikulas	None	nord	0.001232	0.024143	0.069424	False	skyrim
7851	7851	Wolf	Unaligned	beast	0.062500	0.268136	0.593333	True	dd5
4035	4035	Wolf	Normal Monster	beast	0.250000	0.240000	1200.000000	True	yugioh
7038	7038	Ysgramor	None	animals	0.002463	0.010140	0.011505	True	skyrim
7427	7427	Ysgramor	None	nord	0.061576	0.221149	0.296225	True	skyrim

20 rows × 9 columns

Normalization Results

- Result: entities from different universes have now attack values ranging from 0 to 100%
- Easily comparable with each other
- NaN values replaced with 0

```
dfc['attack'].fillna(value=0)
```

```
yugioh = universes.get_group("yugioh")
norm_attk_yoh=(yugioh.attack-yugioh.attack.min())/(yugioh.attack.max()-yugioh.attack.min())
norm_attk_yoh.describe()
```

```
count    4215.000000
mean        0.289430
std         0.178813
min         0.000000
25%         0.160000
50%         0.300000
75%         0.400000
max         1.000000
Name: attack, dtype: float64
```

```
dd5 = universes.get_group("dd5")
norm_attk_dd5=(dd5.attack-dd5.attack.min())/(dd5.attack.max()-dd5.attack.min())
norm_attk_dd5.describe()
```

```
count    318.000000
mean       0.589936
std        0.184499
min        0.000000
25%        0.497738
50%        0.610860
75%        0.700226
max        1.000000
Name: attack, dtype: float64
```

```
skyrim = universes.get_group("skyrim")
norm_attk_sr=(skyrim.attack-skyrim.attack.min())/(skyrim.attack.max()-skyrim.attack.min())
norm_attk_sr.describe()
```

```
count    975.000000
mean       0.147776
std        0.088288
min        0.000000
25%        0.111147
50%        0.136938
75%        0.154992
max        1.000000
```

Universität Marburg

ration
Maya