Pokemon

John Cruz

2023-03-01

Introduction

The Serebii website provides a list, called the National Pokedex, of all the Pokemon in all the games. The table breaks down each Pokemon with a unique ID number, name, type, abilities and base stats such as attack and defense.

Data Sources: Pokemon Stats

Lets determine a frequency chart of which Pokemon fall into which types they are.

Required Libraries

```
library(tidyverse)
library(rvest)
library(xml2)
library(janitor)
```

Web Scrape Table into Data Frame

At first, when the table was web scraped, it brought in the values into a Data Frame, however, I quickly realized that the abilities listed multiple ones and the *rvest()* package cleaned it up by adding a space between them. This causes issues as some abilities are multiple words so to string manipulate them, I needed to account for the built-in HTML line breaks (). To do this, I used XML to add an intentional new line character ("\n"), so that I could be able to break them apart later.

Using XML

```
url <- "https://www.serebii.net/pokemon/nationalpokedex.shtml"

web_table <- read_html(url)

# use XML to account for <br> with abilities and add '\n'
xml_find_all(web_table, ".//br") |>
  xml_add_sibling("p", "\n")
xml_find_all(web_table, ".//br") |>
  xml_remove()
```

```
web_table <-
  web_table |>
  html_element('.dextable') |>
  html_table()

pokemon_stats <- as.data.frame(web_table)

knitr::kable(head(pokemon_stats))
```

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
No.	Pic	Name	Type	Abilities	Base Stats	Base Stats	Base Stats	Base Stats	Base Stats	Base Stats	NA
No.	Pic	Name	Type	Abilities	HP	Att	Def	S.Att	S.Def	Spd	NA
#00001			Bulbasaur		Overgrow						
Chlorophyll	45	49	49	65	65	45					
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
#00002			Ivysaur		Overgrow						
Chlorophyll	60	62	63	80	80	60					
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Tidy Up Data

Every other row that was brought into the Data Frame had *N/A* values. I removed these rows where if any Pokemon didn't have a name, it was dropped. The columns between *Name* and the *Base Stats* also were shifted to the right by one column. Using a vector, I moved the columns to the appropriate alignment and followed it by using the *janitor* package to clean up column names. To account for multiple abilities, the data frame was then unpivoted into a longer format for each Pokemon.

Note: Pokemon types are missing and the column was dropped. It will be explained in the following section.

```
# drop null values if Pokemon name is N/A
stats_df <-
  pokemon_stats |>
  drop_na(4)

# drop first row (duplicate header) and second column (pic)
stats_df <- stats_df[-1,-2]

# set column headers from first row and clean names
stats_df <-
  stats_df |>
  row_to_names(row_number = 1) |>
  clean_names()

# shift pokemon names, etc to left by 1 column
stats_df[c(2:10)] = stats_df[, c(3:11)]

# drop 'na' column
stats_df <-
```

```
  stats_df |>
  select(!c(na, type))

# split multiple abilities into long format based on created '\n'
stats_df <-
  stats_df |>
  separate_longer_delim(abilities, delim = "\n")

# change to pokemon number
stats_df$no <-
  parse_number(stats_df$no)

knitr::kable(head(stats_df))
```

no	name	abilities	hp	att	def	s_att	s_def	spd
1	Bulbasaur	Overgrow	45	49	49	65	65	45
1	Bulbasaur	Chlorophyll	45	49	49	65	65	45
2	Ivysaur	Overgrow	60	62	63	80	80	60
2	Ivysaur	Chlorophyll	60	62	63	80	80	60
3	Venusaur	Overgrow	80	82	83	100	100	80
3	Venusaur	Chlorophyll	80	82	83	100	100	80

Pokemon Types Missing

While working on the Data Frame, I realized the Pokemon types were missing. Looking into the HTML code, the types were not text, but images with no conventional tag name for each one. Because of this, I used another website that provides that information and web scraped similarly into another Data Frame.

```
url <- "https://bulbapedia.bulbagarden.net/wiki/List_of_Pok%C3%A9mon_by_National_Pok%C3%A9dex_number"

web_table <- read_html(url)

# use XML to account for <br> and replace with '\n'
xml_find_all(web_table, ".//br") |>
  xml_add_sibling("p", "\n")
xml_find_all(web_table, ".//br") |>
  xml_remove()

web_table <-
  web_table |>
  html_element('body') |>
  html_table()

pokemon_types <- as.data.frame(web_table)

knitr::kable(head(pokemon_types))
```

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32	X33	X34	X35	X36	X37	X38
Shortcuts																																					

Ndex Olddex Natdex |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA | |Ndex |MS |Pokémon |Type |Type |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA | |#0001 | |Bulbasaur |Grass |Poison |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA | |#0002 | |Ivysaur |Grass |Poison |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA | |#0003 | |Venusaur |Grass |Poison |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA | |#0004 | |Charmander |Fire |Fire |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |NA |

Since each Pokemon has a unique identifier, I created the data frame to relate this ID to its types.

```
# drop null values if Pokemon name is N/A
types_df <-
  pokemon_types |>
  drop_na(2)

# drop unnecessary columns
types_df <-
  types_df[, 1:5]

# set column headers from first row and clean names
types_df <-
  types_df |>
  row_to_names(row_number = 1) |>
  clean_names()

# change to pokemon number
types_df$ndex <-
  parse_number(types_df$ndex)

# drop N/A or zero (0) while keeping only distinct pokemon numbers
types_df <-
  types_df |>
  drop_na() |>
  filter(ndex != 0) |>
  distinct(ndex, .keep_all=TRUE)

# within same pokemon number, replace repeated types with N/A
types_df <-
  types_df |>
  mutate(type_2 = if_else(type_2 != type, type_2, NA)) |>
  select(-c(2)) |>
  rename(no = ndex)

# melt both type columns into one column
temp1 <-
  types_df |>
```

```
  select(1:3)

temp2 <-
  types_df |>
  select(1,2,4) |>
  rename(type = type_2)

types_df <-
  temp1 |>
  full_join(temp2) |>
  drop_na() |>
  select(!pokemon) |>
  arrange(no)

knitr::kable(head(types_df))
```

no	type
1	Grass
1	Poison
2	Grass
2	Poison
3	Grass
3	Poison

Merge Pokemon Stats and Types Tables

Using a similar way of SQL joins, I merged the Pokemon to its types and stats.

```
stats_types_df <-
  stats_df |>
  inner_join(types_df) |>
  relocate(type, .after = name)

knitr::kable(head(stats_types_df))
```

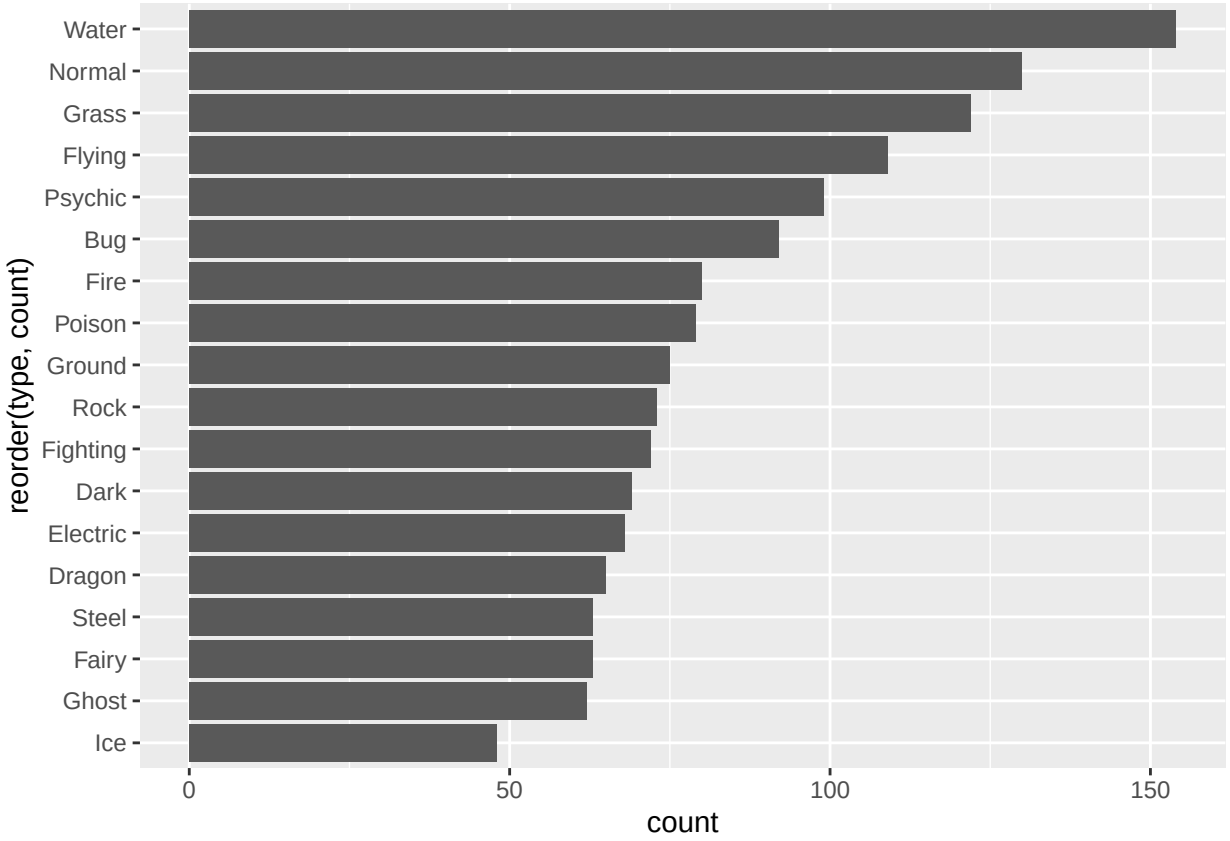
no	name	type	abilities	hp	att	def	s_att	s_def	spd
1	Bulbasaur	Grass	Overgrow	45	49	49	65	65	45
1	Bulbasaur	Poison	Overgrow	45	49	49	65	65	45
1	Bulbasaur	Grass	Chlorophyll	45	49	49	65	65	45
1	Bulbasaur	Poison	Chlorophyll	45	49	49	65	65	45
2	Ivysaur	Grass	Overgrow	60	62	63	80	80	60
2	Ivysaur	Poison	Overgrow	60	62	63	80	80	60

Pokemon Type Frequency

The frequency chart shows us that *Water, Normal and Grass* types are the most common among all the Pokemon. One caveat to point out, is that a Pokemon could have up to two different types. This means that a Pokemon can appear twice in different categories as a result in here.

```
grouped_type <-
  stats_types_df |>
  distinct(name, type) |>
  group_by(type) |>
  summarise(count = n()) |>
    arrange(desc(count))

grouped_type |>
  ggplot(aes(x = count, y = reorder(type, count))) +
  geom_bar(stat = 'identity')
```




```
knitr::kable(grouped_type)
```

type	count
Water	154
Normal	130
Grass	122
Flying	109
Psychic	99
Bug	92
Fire	80
Poison	79
Ground	75

type	count
Rock	73
Fighting	72
Dark	69
Electric	68
Dragon	65
Fairy	63
Steel	63
Ghost	62
Ice	48