Pokémon Classification Project

Charles Christopher Hyland

450411920

Table of Content

Data Understanding

Data Description

Data Description Report

The data comes in as a tabular format, which contains 12 attributes.

#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False

Fig. 1 Snapshot of sample of data.

Here, in a snapshot of the data, we can see it is organized in a table format. Furthermore, if a Pokémon does not have a second type, then NaN is used in place. The “Legendary” attribute return False if the Pokémon is not a legendary type Pokémon and returns True if it does.

The variables present in the data and its respective type is as follow:

Variable	Type
Type 1	String
Type 2	String
Total	Numeric
HP	Numeric
Attack	Numeric
Defence	Numeric
Special Attack	Numeric
Special Defence	Numeric
Speed	Numeric
Generation	Numeric
Legendary	Boolean

Fig 2. List of variables

The binary variable of Legendary will have to be converted into a numerical binary format (0 for False and 1 for True) to make data manipulation easier. Generation is coded as a numeric range from 1-6, where each number indicates a different region.

For a breakdown of what the different variables mean:

Variable	Type
Type 1	The primary “type” of Pokémon it is (e.g. fire, water).
Type 2	The secondary “type” of Pokémon it is (e.g. fire, water).
Total	The summation of all other attributes pertaining to that Pokémon.
HP	How many health points the Pokémon has.
Attack	How many attack points the Pokémon has.
Defence	How many defence points the Pokémon has.
Special Attack	How many special attack points the Pokémon has.
Special Defence	How many special defence points the Pokémon has.
Speed	How many speed points the Pokémon has.
Generation	Which region does the Pokémon come from.
Legendary	Whether the Pokémon is a legendary Pokémon.

Fig. 3 Variable breakdown

Data Exploration Report

The attributes that seem most promising from the data in terms of using to classify Pokémon are:

- Attack
- Defence
- Special Attack
- Special Defence
- Speed

The generation (the region in which the Pokémon comes from) of the Pokémon currently seems irrelevant. However, it may be the case that Pokémon of a certain type from a particular region, has a combination of features which differ to Pokémon of the same type in other regions.

To further subset the data, we can use a swarm plot to see a breakdown of the types of Pokémon and the their respective attributes.

Using a swarm plot will help with the initial visualisation of the data:

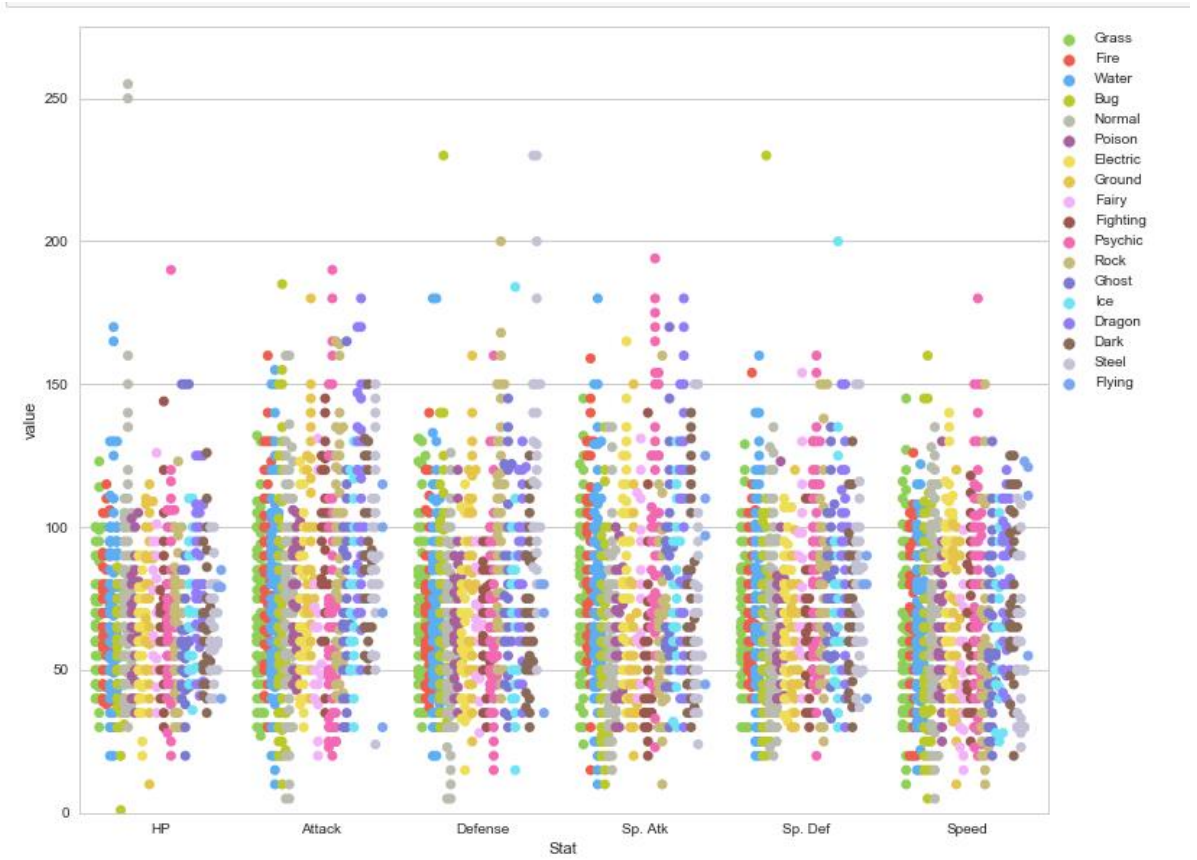

Fig. 4 Swarm plot of attributes

We can see that from this, it appears that from the distribution of Pokémon types, the majority of them posses around the similar attributes. Much more of an in depth analysis is needed to uncover the distribution of the Pokémon.

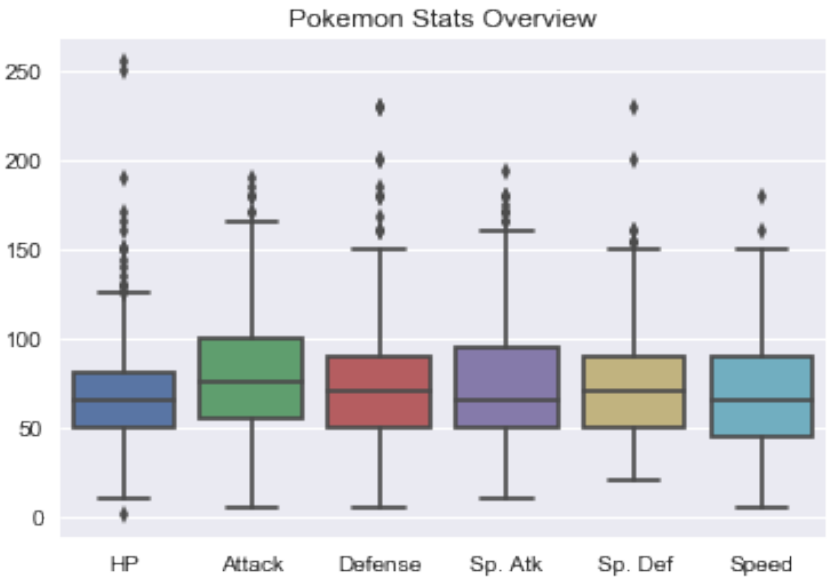

Fig. 5 Overview of Pokémon attributes

Type 1	count	mean	std	min	25%	50%	75%	max
Bug	69.0	70.971014	37.040904	10.0	45.00	65.0	90.00	185.0
Dark	31.0	88.387097	25.774247	50.0	65.00	88.0	100.00	150.0
Dragon	32.0	112.125000	33.742622	50.0	86.25	113.5	134.25	180.0
Electric	44.0	69.090909	23.764169	30.0	53.75	65.0	85.00	123.0
Fairy	17.0	61.529412	29.751298	20.0	45.00	52.0	72.00	131.0
Fighting	27.0	96.777778	28.290163	35.0	80.00	100.0	120.00	145.0
Fire	52.0	84.769231	28.769275	30.0	62.25	84.5	101.00	160.0
Flying	4.0	78.750000	37.500000	30.0	60.00	85.0	103.75	115.0
Ghost	32.0	73.781250	29.629687	30.0	53.75	66.0	92.75	165.0
Grass	70.0	73.214286	25.380520	27.0	55.00	70.0	93.50	132.0
Ground	32.0	95.750000	33.059087	40.0	72.00	85.0	121.00	180.0
Ice	24.0	72.750000	27.289511	30.0	50.00	67.0	87.50	130.0
Normal	98.0	73.469388	30.295862	5.0	55.00	70.5	85.00	160.0
Poison	28.0	74.678571	19.630010	43.0	60.00	74.0	90.50	106.0
Psychic	57.0	71.456140	42.309265	20.0	45.00	57.0	95.00	190.0
Rock	44.0	92.863636	35.325458	40.0	59.75	95.0	120.25	165.0
Steel	27.0	92.703704	30.388276	24.0	77.50	89.0	110.00	150.0
Water	112.0	74.151786	28.377192	10.0	53.00	72.0	92.00	155.0

Fig. 6 Attack attributes

From the figure 5, that when aggregating all Pokémon types together, there are noticeable outliers for each attribute. From that, when analysing the attack attribute via subsetting the Pokémon by their types, we obtain figure 6. From this, it is evident that Dragon Type Pokémon has the highest attack attribute whilst fairy type Pokémon has the lowest attack attribute. There is a difference of 51 attack points between the means of these 2 Pokémon types, therefore suggesting there is a statistical difference in attack attributes among Pokémon types. We can then repeat the breakdown of Pokémon types for the other attributes as seen from figure 7-10.

Type 1	count	mean	std	min	25%	50%	75%	max
Bug	69.0	70.724638	33.617609	30.0	50.00	60.0	90.00	230.0
Dark	31.0	70.225806	25.121982	30.0	51.00	70.0	90.00	125.0
Dragon	32.0	86.375000	24.102938	35.0	68.75	90.0	100.00	130.0
Electric	44.0	66.295455	24.757033	15.0	49.75	65.0	80.00	115.0
Fairy	17.0	65.705882	18.979478	28.0	50.00	66.0	75.00	95.0
Fighting	27.0	65.925926	18.578163	30.0	54.00	70.0	79.50	95.0
Fire	52.0	67.769231	23.658200	37.0	51.00	64.0	78.00	140.0
Flying	4.0	66.250000	21.360009	35.0	61.25	75.0	80.00	80.0
Ghost	32.0	81.187500	32.551138	30.0	60.00	72.5	111.00	145.0
Grass	70.0	70.800000	24.485192	30.0	50.00	66.0	84.50	131.0
Ground	32.0	84.843750	33.786912	25.0	53.75	84.5	110.00	160.0
Ice	24.0	71.416667	34.387708	15.0	48.75	75.0	85.00	184.0
Normal	98.0	59.846939	23.771833	5.0	43.25	60.0	73.75	126.0
Poison	28.0	68.821429	21.066128	35.0	52.75	67.0	82.25	120.0
Psychic	57.0	67.684211	28.359401	15.0	48.00	65.0	80.00	160.0
Rock	44.0	100.795455	36.447209	40.0	71.50	100.0	120.50	200.0
Steel	27.0	126.370370	44.806548	50.0	97.50	120.0	150.00	230.0
Water	112.0	72.946429	27.773809	20.0	54.50	70.0	88.50	180.0

Fig. 7 Defense

Type 1	count	mean	std	min	25%	50%	75%	max
Bug	69.0	53.869565	26.697055	10.0	35.00	50.0	65.00	135.0
Dark	31.0	74.645161	33.200952	30.0	45.00	65.0	96.50	140.0
Dragon	32.0	96.843750	42.257360	30.0	60.00	105.0	122.50	180.0
Electric	44.0	90.022727	29.740340	35.0	65.00	95.0	106.00	165.0
Fairy	17.0	78.529412	28.548462	40.0	60.00	75.0	99.00	131.0
Fighting	27.0	53.111111	28.159345	20.0	35.00	40.0	62.50	140.0
Fire	52.0	88.980769	30.042121	15.0	70.00	85.0	109.00	159.0
Flying	4.0	94.250000	34.769479	45.0	84.00	103.5	113.75	125.0
Ghost	32.0	79.343750	32.561217	30.0	58.00	65.0	96.25	170.0
Grass	70.0	77.500000	27.244864	24.0	57.00	75.0	99.50	145.0
Ground	32.0	56.468750	28.135598	20.0	39.50	47.5	65.75	150.0
Ice	24.0	77.541667	26.604967	30.0	60.00	77.5	95.00	130.0
Normal	98.0	55.816327	23.946395	15.0	40.00	50.0	65.00	135.0
Poison	28.0	60.428571	19.322657	30.0	40.75	60.0	71.50	100.0
Psychic	57.0	98.403509	38.539340	23.0	70.00	95.0	125.00	194.0
Rock	44.0	63.340909	28.249670	10.0	45.00	60.0	74.25	160.0
Steel	27.0	67.518519	31.458606	24.0	47.50	55.0	79.50	150.0
Water	112.0	74.812500	29.030128	10.0	55.00	70.0	90.50	180.0

Fig. 8 Special Attack

Type 1	count	mean	std	min	25%	50%	75%	max
Bug	69.0	64.797101	32.126395	20.0	45.00	60.0	80.00	230.0
Dark	31.0	69.516129	24.799020	30.0	50.00	65.0	87.50	130.0
Dragon	32.0	88.843750	29.884843	30.0	70.00	90.0	105.00	150.0
Electric	44.0	73.704545	22.601275	32.0	55.00	79.5	90.00	110.0
Fairy	17.0	84.705882	29.721130	40.0	65.00	79.0	98.00	154.0
Fighting	27.0	64.703704	22.745777	30.0	49.00	63.0	75.00	110.0
Fire	52.0	72.211538	22.619908	40.0	54.75	67.5	85.00	154.0
Flying	4.0	72.500000	22.173558	40.0	70.00	80.0	82.50	90.0
Ghost	32.0	76.468750	26.038414	33.0	55.00	75.0	91.25	135.0
Grass	70.0	70.428571	21.446645	30.0	55.00	66.0	85.00	129.0
Ground	32.0	62.750000	21.267877	30.0	45.00	62.5	80.00	120.0
Ice	24.0	76.291667	37.213227	30.0	50.00	70.0	91.25	200.0
Normal	98.0	63.724490	25.142801	20.0	43.50	60.5	75.00	135.0
Poison	28.0	64.392857	19.887348	40.0	52.25	60.5	76.00	123.0
Psychic	57.0	86.280702	31.126329	20.0	58.00	90.0	110.00	160.0
Rock	44.0	75.477273	32.265965	25.0	50.00	70.0	90.50	150.0
Steel	27.0	80.629630	29.018758	37.0	60.00	80.0	95.00	150.0
Water	112.0	70.517857	28.460493	20.0	50.00	65.0	89.25	160.0

Fig. 9 Special Defense

Type 1	count	mean	std	min	25%	50%	75%	max
Bug	69.0	61.681159	33.227599	5.0	36.00	60.0	85.00	160.0
Dark	31.0	76.161290	27.768203	20.0	59.00	70.0	98.50	125.0
Dragon	32.0	83.031250	23.239961	40.0	65.25	90.0	97.75	120.0
Electric	44.0	84.500000	26.691607	35.0	60.00	88.0	101.50	140.0
Fairy	17.0	48.588235	23.305200	15.0	30.00	45.0	60.00	99.0
Fighting	27.0	66.074074	26.054567	25.0	45.00	60.0	86.00	118.0
Fire	52.0	74.442308	25.245783	20.0	60.00	78.5	96.25	126.0
Flying	4.0	102.500000	32.098806	55.0	97.00	116.0	121.50	123.0
Ghost	32.0	64.343750	28.020280	20.0	44.00	60.5	84.25	130.0
Grass	70.0	61.928571	28.506456	10.0	40.00	58.5	80.00	145.0
Ground	32.0	63.906250	27.450083	10.0	40.00	65.0	90.00	120.0
Ice	24.0	63.458333	24.498410	25.0	48.75	62.0	80.00	110.0
Normal	98.0	71.551020	28.406157	5.0	50.00	71.0	90.75	135.0
Poison	28.0	63.571429	22.631392	25.0	50.00	62.5	77.00	130.0
Psychic	57.0	81.491228	37.335412	20.0	50.00	80.0	104.00	180.0
Rock	44.0	55.909091	29.903580	10.0	35.00	50.0	70.00	150.0
Steel	27.0	55.259259	25.846578	23.0	31.50	50.0	70.00	110.0
Water	112.0	65.964286	23.019353	15.0	50.00	65.0	82.00	122.0

Fig. 10 Speed

From this, it is evident that there are noticeable differences between Pokémon types with regards to their attributes. This confirms out initial hypothesis that there are noticeable differences between the Pokémon types and their attributes, therefore allowing our process of classifying and identifying Pokémon based on their type feasible.
When splitting the data up by the type of Pokémon and summarizing the data based on certain variables, a few noticeable results appeared:

Variable	Maximum	Type	Minimum	Type
Attack	112.13	Dragon	70.97	Bug
Defence	126.37	Steel	65.71	Fairy
Special Attack	96.84	Dragon	53.87	Bug
Special Defence	88.84	Dragon	62.75	Ground
Speed	102.5	Flying	55.26	Steel

From this, it appears that certain Pokémon types are associated with certain characteristics and due to the discrepancies between the maximum and minimum value for each type, it is worthwhile to assume that different types of Pokémon have unique attributes associated with them. From this, we are able to proceed with the business question and makes it worthwhile to further examine this trend.

Additionally, we can see the count of each primary type of a Pokémon through a frequency table.

Type 1	count
Bug	69.0
Dark	31.0
Dragon	32.0
Electric	44.0
Fairy	17.0
Fighting	27.0
Fire	52.0
Flying	4.0
Ghost	32.0
Grass	70.0
Ground	32.0
Ice	24.0
Normal	98.0
Poison	28.0
Psychic	57.0
Rock	44.0
Steel	27.0
Water	112.0

Fig. 11 Frequency of Pokémon by type 1

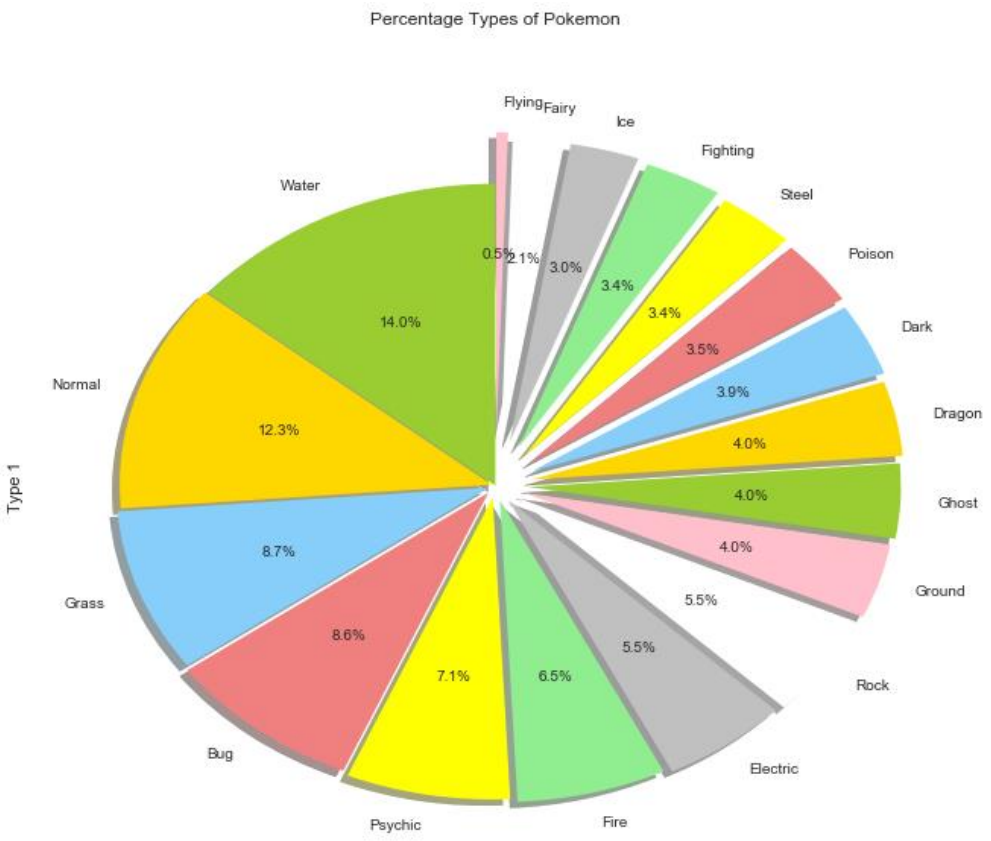

Fig. 12 Percentage breakdown of Pokémon types

From this, it is evident that water type Pokémon are the most frequent Pokémon with over 112 instances whilst normal type Pokémon is the second most frequent. Additionally primary flying type Pokémon appears to be the rarest Pokémon. This was an unexpected result as there were many cases of Pokémon posessing wings and “flying-type” attack moves such as fly. Further investigation into this led us to then analyse that in addition to the primary type of a Pokémon, does the Pokémon also posses a secondary type? The result from this question is seen in the next table:

Type 2	count
Bug	3.0
Dark	20.0
Dragon	18.0
Electric	6.0
Fairy	23.0
Fighting	26.0
Fire	12.0
Flying	97.0
Ghost	14.0
Grass	25.0
Ground	35.0
Ice	14.0
None	386.0
Normal	4.0
Poison	34.0
Psychic	33.0
Rock	14.0
Steel	22.0
Water	14.0

Fig. 13 Frequency of Pokémon by type 2

Here, flying type is the most common Pokémon's secondary type, excluding the fact that most Pokémon seem to not have a secondary type. This solves our earlier conundrum with regards to the apparent infrequency of flying type Pokémon. We can then formulate that most Pokémon tend to be of a primary type with the added secondary type of being able to fly.

Aggregating the Pokémon types again, we see the following results:

	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
count	800.00	800.00	800.00	800.00	800.00	800.00	800.00
mean	435.10	69.26	79.00	73.84	72.82	71.90	68.28
std	119.96	25.53	32.46	31.18	32.72	27.83	29.06
min	180.00	1.00	5.00	5.00	10.00	20.00	5.00
25%	330.00	50.00	55.00	50.00	49.75	50.00	45.00
50%	450.00	65.00	75.00	70.00	65.00	70.00	65.00
75%	515.00	80.00	100.00	90.00	95.00	90.00	90.00
max	780.00	255.00	190.00	230.00	194.00	230.00	180.00

Fig 14. Summary statistics of the data

We noted that the mean is greater than the median for the attributes of the Pokémon, thereby suggesting that the distribution of such variables are positively or right skewed.

Resultantly, through this exploration, there is no need to alter the data-mining goals.

Data Verification Report

Data Quality Report
As we were given this dataset by Professor Oak himself, it is imperative to verify the quality of the data we were given.

Missing data
Any missing value issues in the data was rectified as follows:

1) Find out the type of Pokémon it is.
2) Check whether the distribution of the other observations of the variable was skewed or not.
    - **a.** If the data was skewed
        - **i.** Calculate the median of the attribute that the missing value belongs to from the rest of the Pokémon of the same type.
        - **ii.** Use the median to replace the missing value.
    - **b.** If the data was normally distributed:
        - **i.** Calculate the mean of the attribute that the missing value belongs to from the rest of the Pokémon of the same type.
        - **ii.** Use the mean to replace the missing value.

It is worth acknowledging the flaws in this methodology, yet it is sufficient for the task at hand. However, there were very few observations missing in the dataset.

Data error
There were no typographical errors with regards to the data as it has been cross validated by Professor Oak himself and the rest of his Pokémon research assistant.

Measurement errors
There were no measurement errors with regards to the data as all attributes were recorded on the same scale.

Coding inconsistencies
There were no coding inconsistencies between variables types.

Bad metadata
The Generation column is a misleading term and therefore, it is proposed that region would be a better indicator. The dataset will be recoded whereby we transform each generation field with the respective region of the Pokémon. This is dealt with later on in the data preparation stage.

Data Preparation

Data Selection

Due to the small number of attributes, we do not remove any features and furthermore, due to no inconsistencies in the observations, we retain all observations. Furthermore, due to the low number of features compared to the number of observations, we do not run into the curse of dimensionality. Therefore, we select all Pokémon to be examined and keep all the current features in the original dataset so as to ensure that we do not lose any important information. We would like to retain as much information as possible in order to assist our classification efforts. Currently, from the data exploration phase, there are noticeable discrepancies between Pokémon types with regards to all the attributes and therefore, keeping the attributes will assist our algorithm tremendously.

Data Cleaning

As noted in the data exploration phase, there were no apparent issues and therefore need to clean the data with regards to the following criteria:

- No missing data
- No data errors found
- No apparent measurement errors

Data Construction

As mentioned from the data exploration phase, there is a need to alter the categorical column of generation as it currently suggests there is ordinality for the generation column, which again does not make sense as a Pokémon from generation 5 is not "better" than a Pokémon from generation 1. Therefore, we need to encode dummy variables into the model indicating which generation is the Pokémon a part of and then remove the original generation column. The results of such actions looks as follows:

Sp. Atk	Sp. Def	Speed	Legendary	Generation_1	Generation_2	Generation_3	Generation_4	Generation_5
65	65	45	0	1.0	0.0	0.0	0.0	0.0
80	80	60	0	1.0	0.0	0.0	0.0	0.0
100	100	80	0	1.0	0.0	0.0	0.0	0.0
122	120	80	0	1.0	0.0	0.0	0.0	0.0
60	50	65	0	1.0	0.0	0.0	0.0	0.0
80	65	80	0	1.0	0.0	0.0	0.0	0.0
109	85	100	0	1.0	0.0	0.0	0.0	0.0
130	85	100	0	1.0	0.0	0.0	0.0	0.0

Fig. 15 Newly constructed features

The construction of such binary variables now allows the algorithm to identify which region does a particular Pokémon originate from and by including only 5 dummy variables, we do not fall into the dummy variable trap.
Furthermore, for the Type 1 and Type 2 column, these categorical variables were converted into dummy variables for ease of data analysis. A preview of this can be seen here:

Type_1_Ghost	Type_1_Grass	Type_1_Ground	Type_1_Ice	Type_1_Normal
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Fig. 16 Newly constructed features

No further data was altered or created.

Data Integration
The data initially came from existing data scraped over numerous websites via APIs. Fortunately for us, the dataset provided has already been merged and integrated regarding the Pokémon from the different regions. With this in mind, there is no need for any work regarding the integration of the dataset.

Data Formatting
Due to the nature of the data being a cross-sectional set, the ordering for the cross-sectional data does not matter. Furthermore, the algorithms and models that will be applied to the dataset also does not require any formatting to the dataset. However, any need for standardisation of the dataset will be done accordingly should the need for it arise. Therefore, no work was done on formatting the data.

Dataset Description
Amount of data
Currently, there are 800 observations from a cross-sectional dataset. From this it appears that the data allows us to draw generalizable conclusions and also assists us in making accurate predictions. However, a possible issue is that there are not many attributes. This

may be a problem as our predictions may suffer from omitted variable bias and that important predictors we could be using are not present in the data.

Value Types

An updated table of the types of variables now present is as follow:

Variable	Type
Type 1	String
Type 2	String
Total	Numeric
HP	Numeric
Attack	Numeric
Defence	Numeric
Special Attack	Numeric
Special Defence	Numeric
Speed	Numeric
Generation_1	Numeric
Generation_2	Boolean
Generation_3	Boolean
Generation_4	Boolean
Generation_5	Boolean
Legendary	Boolean

Fig. 17 Updated variable table

Coding Schemes

The only worthwhile coding scheme to note is the fact that 1 represents true and 0 represent false for the dummy variables present in the dataset.