POKÉMON REPORT

Data Visualisation and The Web

Muhammad Ahmed

Goldsmiths, University of London

Data Visualisation and The Web

Friday 14th December 2018

This report will primarily lean towards statistical data analysis of Pokémon [1] from all seven generations covering all 807 Pokémon from the National Pokédex [2]. I will discuss the research questions which mainly revolve around exploring the differing stats of Pokémon, then discuss the dataset and how data it was cleaned and pre-processed as well as the visualizations and analysis. Tajiri [3] was behind the concept of Pokémon in 1989 and was released by Nintendo in 1996. The concept of this game centralized around the notion of catching monstrous creatures. The strength and size of the Pokémon creatures increased upon success in battles as they evolved from experience.

My deep interest with Pokémon games during my earlier years and an everlasting curiosity about the base stats which were generally not known. In addition to the above, I believe this report can be a stepping stone to working with machine learning classification algorithms to determine Pokémon based on raw stats. Avid gamers and Pokémon enthusiast may find the statistical data analysis across the complete Pokémon dataset very useful and insightful.

Research Questions

1. How does the amount of Pokémon vary per Generation? [4]
2. How does HP change with Generations?
3. To what extent does the primary type of a Pokémon influence their total stat? [5]
4. How well does Sp. Def correlate with Sp. Atk amongst all Pokémon?
5. How do legendary Pokémon compare with normal Pokémon with regards to special attack and defence stats [6]

The first dataset I came across 'The Complete Pokemon Dataset' [7], looked like a very thorough dataset and had 41 columns and I felt like the scope of my investigation was not wide enough so I moved on from that data set and came across 'Pokemon with stats' [8], this dataset was much more precise and had only 13 columns however it didn't have the latest generation of pokemon thus making it out of date. Finally, I came across 'pokemonData' [9] who managed to merge the latest generation of pokemon to the existing 'Pokemon with stats' data set making a perfectly up to date and in scope data set. The dataset came in a CSV format and I was able to store it in a gist to access on the cloud. Additionally, the data included mega versions of pokemon and did not include latest two Pokémon.

When initially looking at the data, I noticed a number of problems and issues with the data - they were resolved as follows.

- Assigning easier to read columns for clarity when visualising data.
- Setting the index to start from 1 to match the National Pokedex Number of each Pokémon.
- Setting the index column in place but without dropping the variable to reduce the need of an extra index column.
- Removed unnecessary excess words before each mega iteration of a pokemon, so accessing would be easier and clearer data.
- Also, dropping duplicates with respects to NPN as the Mega Pokémon did not help with indexing and were not relevant to this study.
- Setting the variables correctly in order to do relevant operations.

Analysis

Important Concepts/Variables of Interest

- Pokémon: The name of a pet monster from a fictitious world [1] (Nominal)
- NPN: National Pokedex Index, a unique identifier of each Pokémon. (discrete numerical)
- HP: Base Health for a pokemon - the higher it is, the longer the pokemon can last in a battle (continuous numerical)
- Generation: Denotes the generation in which a pokemon was released. (categorical ordinal)
- Primary Type: Typically known as Type 1 which usually describes the pokemon based on its nature and demeanour and a nominal value with 18 unique values. (Nominal)
- Total: This is the base stats as a sum total to give us an overall rating on the Pokémon. (continuous numerical)
- Sp. Atk: This is denoted as a special attack and the higher it is, the greater damage the pokemon can inflict.

- Sp. Def: This is denoted as a special defence and the higher it is, the greater resistance the pokemon will have against special attacks.
- Legendary: Indicative of whether a Pokemon is a legendary or not. These category of Pokemon tend to be rare and hard to catch and generally stronger (Boolean).

Graphs/Tables for Key Variables

Pokemon	NPN	Pokemon	Type1	Type2	Total	HP	Attack	Defence	Sp.Atk	Sp.Def	Speed	Generation	Legendary
Arceus	493	Arceus	Normal	NaN	720	120	120	120	120	120	120	4	True
Yveltal	717	Yveltal	Dark	Flying	680	126	131	95	131	98	99	6	True
GiratinaAltered Forme	487	GiratinaAltered Forme	Ghost	Dragon	680	150	100	120	100	120	90	4	True
Reshiram	643	Reshiram	Dragon	Fire	680	100	120	100	150	120	90	5	True
Zekrom	644	Zekrom	Dragon	Electric	680	100	150	120	120	100	90	5	True
Solgaleo	791	Solgaleo	Psychic	Steel	680	137	137	107	113	89	97	7	True
Lunala	792	Lunala	Psychic	Ghost	680	137	113	89	137	107	97	7	True
Xerneas	716	Xerneas	Fairy	NaN	680	126	131	95	131	98	99	6	True
Rayquaza	384	Rayquaza	Dragon	Flying	680	105	150	90	150	90	95	3	True
Lugia	249	Lugia	Psychic	Flying	680	106	90	130	90	154	110	2	True
Ho-oh	250	Ho-oh	Fire	Flying	680	106	130	90	110	154	90	2	True
Dialga	483	Dialga	Steel	Dragon	680	100	120	120	150	100	90	4	True
Mewtwo	150	Mewtwo	Psychic	NaN	680	106	110	90	154	90	130	1	True
Palkia	484	Palkia	Water	Dragon	680	90	120	100	150	120	100	4	True
Slaking	289	Slaking	Normal	NaN	670	150	160	100	95	65	100	3	False

Figure 1 - Table of top 15 Pokémon based on total stats

Descriptive Stats for Key Variables

In total there were 807 pokemon gathered, with regards to the total sum stat/overall rating of the pokemon there was a mean of 421.3 which was quite similar to the median of 430. This similarity can suggest the dataset has a symmetrical distribution and the points being close together is because the middle value in the data set, when ordered smallest to largest, resembles the balancing point in the data, which occurs at the average. The standard deviation was at 110.97 which is relatively low and this indicates that the data tends to be close to the mean. The mode was at 600.0 which is considerably higher than the median and mean but it may not be significant on a dataset this large.

Figure 2 - A correlation heat map of all the key numerical variables

Above we can see generally that the base skills correlate somewhat positively with the total as they contribute to it. Legendary also correlates positively with total as Legendary Pokémon are significantly stronger than the general Pokémon.

Visualisations of all basic variable types

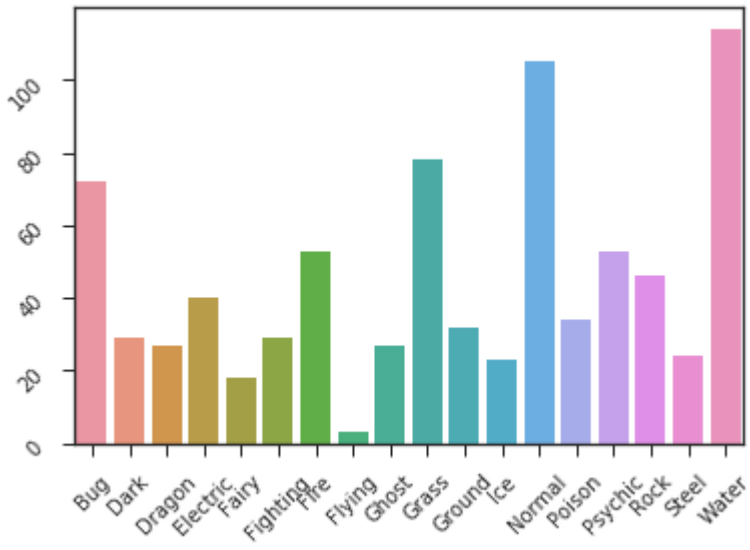

*Figure 3a - Univariate plot with Nominal variable **Type1***

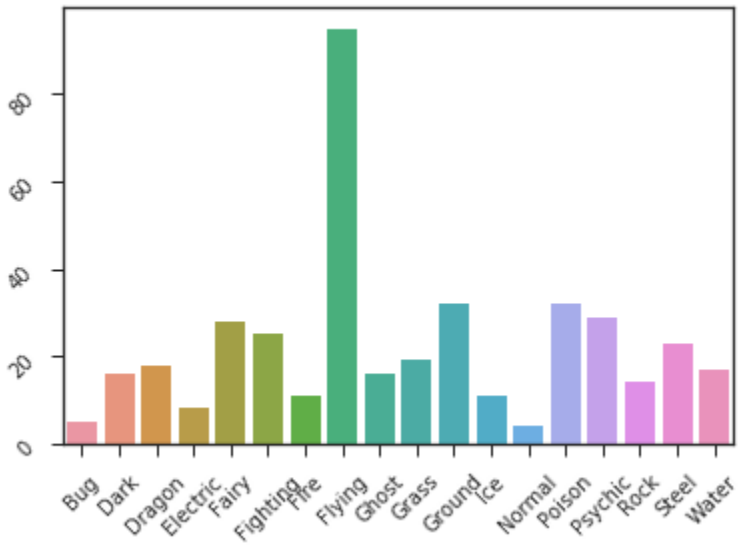

*Figure 3b - Univariate plot with Nominal variable **Type2***

In Figure 3a we can see that amongst the primary types, Water and Normal are the most common with grass in a close third whilst Flying significantly lower than any other type. Contrastingly, in Figure 3b, Flying is the overwhelming predominant type with the rest quite evenly spread with normal and bug being very low.

In figures 4a and 4b, we can see the volatility with regards to how many Pokémon are in each generation.

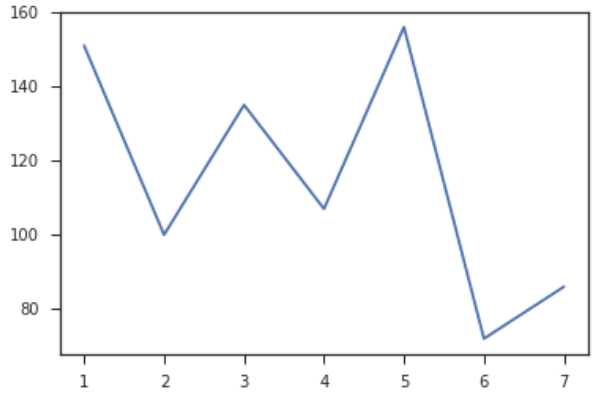

*Figure 4a – Univariate with categorical ordinal variable **Generation***

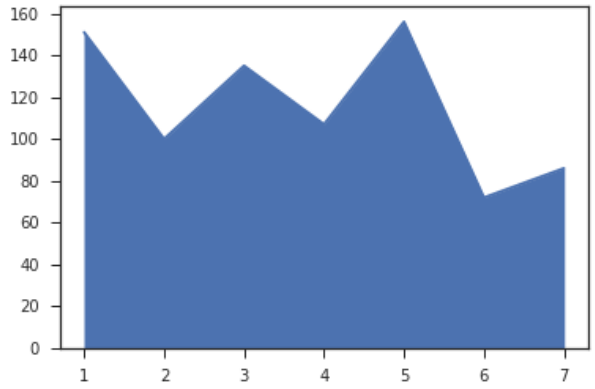

*Figure 4b – Univariate with categorical ordinal **Generation***

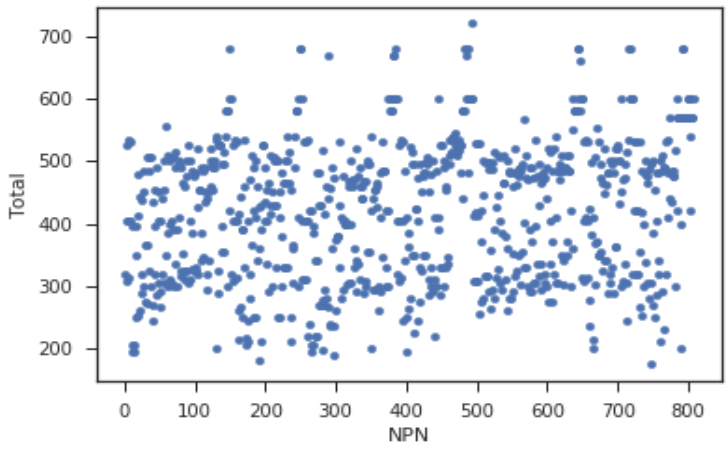

*Figure 5 – Bivariate with discrete numerical variable **NPN (National Pokédex number)***

Figure 5 is somewhat interesting as it gives us the general distribution of the overall stats by their NPN. The outliers which always occur at the end of the pokedex are indictive of all the legendary Pokémon who are always the final numbered for each Generation.

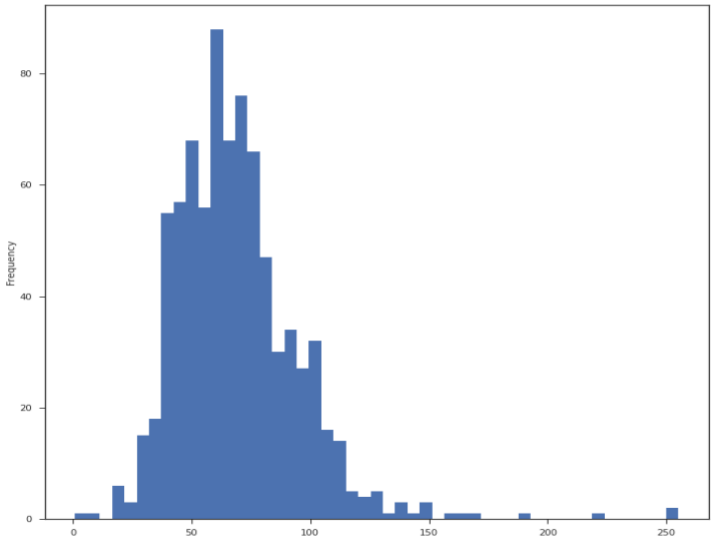

*Figure 6 – Univariate visualisation of the continuous numerical variable **HP***

In Figure 6, the graph demonstrates the distribution of HP, which has a max of 255 and a low of 1. The mean is at 69.46 and the median is slightly lower at 66.0. This is replicated in the graph as the majority of Pokémon with HP stats near that region are within the middle.

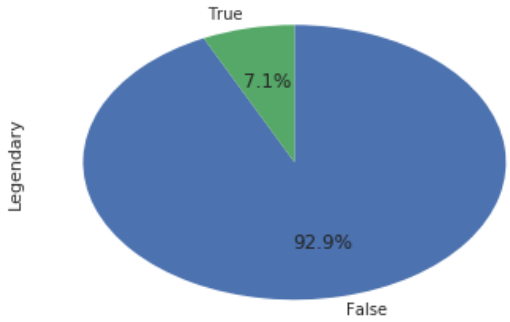

Figure 7 – A univariate visualisation of a Boolean type variable ***Legendary***

In Figure 7, chart shows the overwhelming majority of Pokémon being Legendary further implying their rarity.

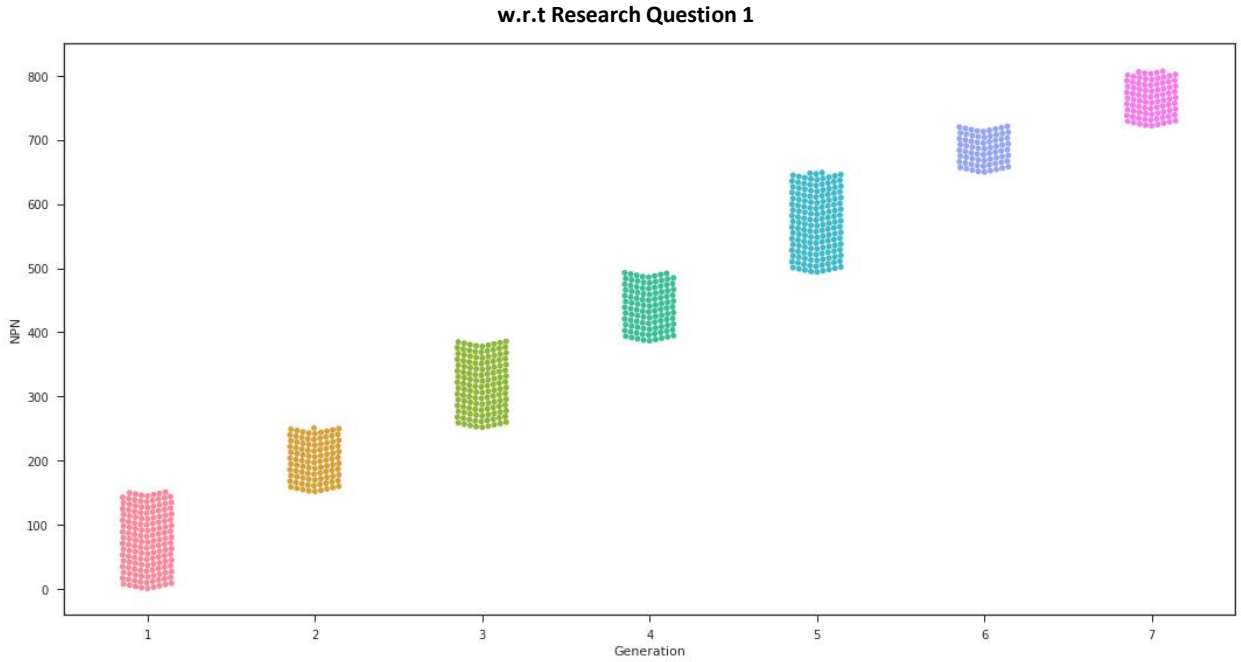

Figure 8 – Generation to NPN plot

Figure 8 demonstrates visibly that Generation 5 had the most Pokémon with 156 and Generation 1 was slightly lower with 151. Generation 6 and 7 had a significant decrease with amounts of Pokémon in the game with 72 and 86 respectively. The vary level s of Pokémon per generation here addressed the aforementioned research question. Additionally, figures 4a & 4b could also pai nt the image based on how steep the lines where with respects to change in generation size.

w.r.t Research Question 2

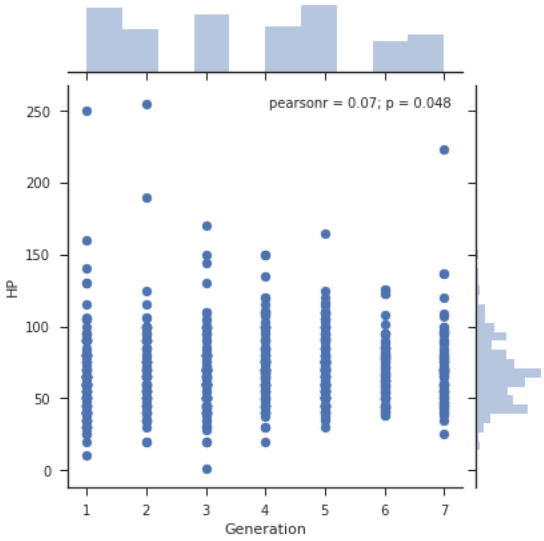

Figure 9 – Generation to HP pair plot

Figure 9 shows us how HP changes with Generations. Generations 1 and 2 had some significantly high HP Pokémon along with Generation 7. Generation 3 looked to have had the lowest HP in general amongst all the generations

w.r.t Research Question 3

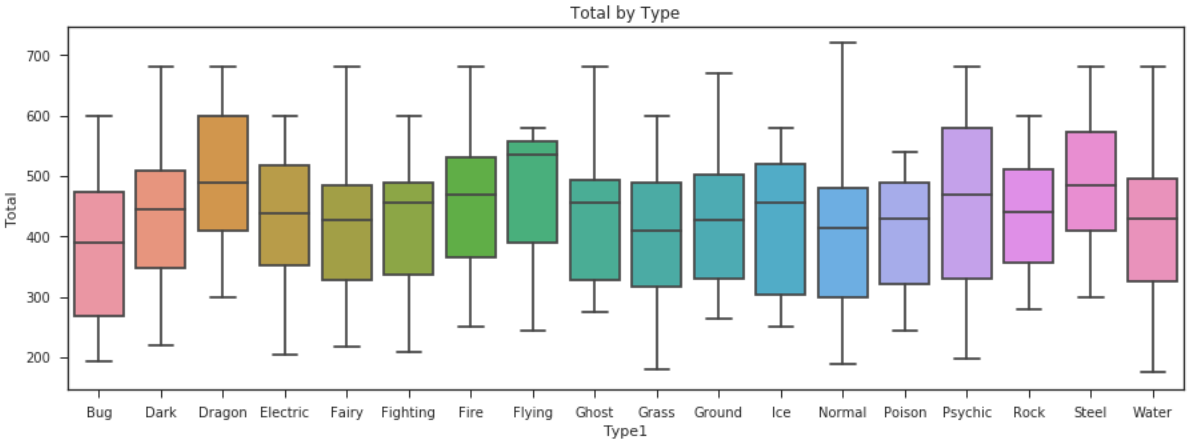

Figure 10 – Type1 and Total

Figure 10 plot suggests that the primary type that will heavily influence total positively would be Dragon or Psychic Pokémon as they are generally the strongest and Bug Pokémon have the lowest total and are the weakest Pokémon.

w.r.t Research Question 4

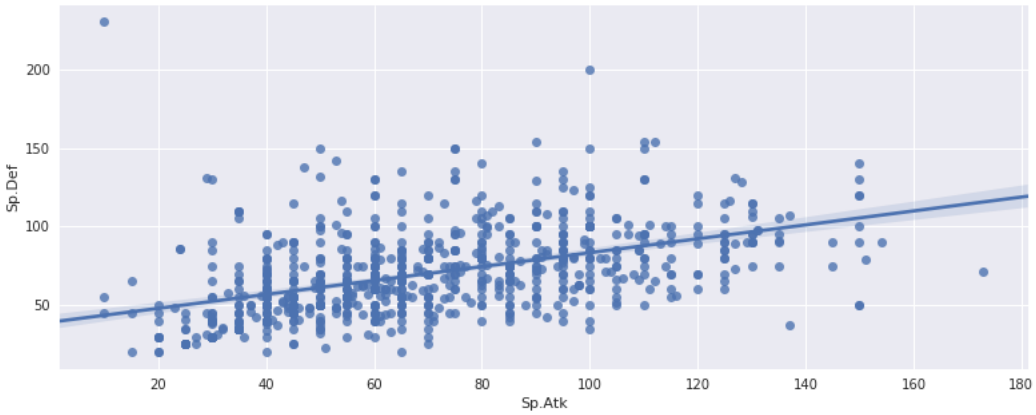

Figure 11 – Sp. Atk compared to Sp.Def amongst all Pokémon

Figure 11 shows us despite a few outliers on either side, there seems to be a regression line suggesting that the relationship between Sp. Atk and Sp. Def is linear. When looking at Figure 2, we can see in the correlation plot that there is a strong correlation between Sp. Atk and Sp. Def and that is backed up by the above graph.

w.r.t Research Question 5

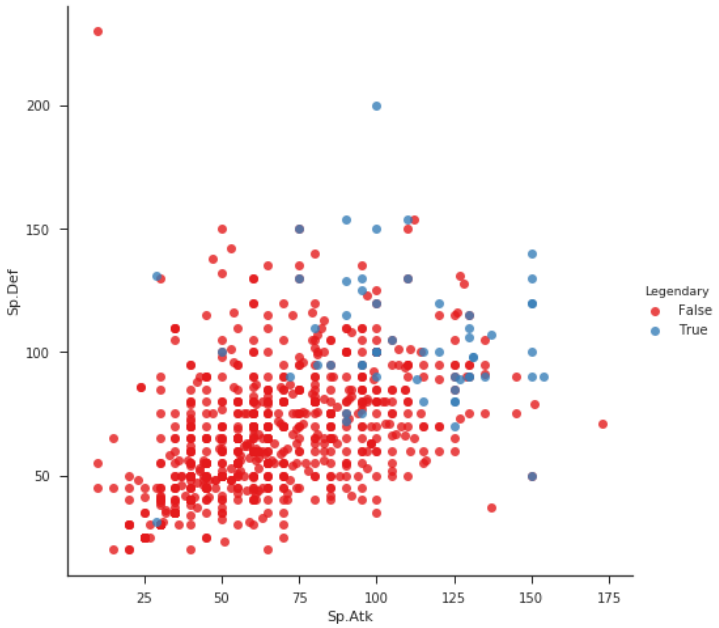

Figure 12 –Legendary Pokémon compared to normal Pokémon with regards to Sp. Atk and Sp. Def

Figure 12 Shows a clear indication that majority of the Legendary Pokémon exist in the top right cluster thus suggesting that there is a positive correlation between high levels of Sp. Atk and Sp. Def with Legendary status. Figure 2 can also support this claim as both Sp. Atk, and Sp. Def have a positive correlation of 0.37 and 0.36 respectively.

With regards to the first two research questions, there seems to be an unpredictable and volatile change of Pokémon each generation and for the latter, I found that Generation 1,2 and 7 had higher HP and Gen 3 had the lowest. Additionally, the primary types which are most effective to total was Dragon and Psychic. Sp. Def had a positive linear correlation with Sp. Atk although the regression line wasn't very steep. Finally, there was a clear indication that the majority of legendary Pokémon had high Sp. Atk and Sp. Def stats although there were some outliers that didn't support the aforementioned claim.

Retrospectively, I think I learnt an incredible amount with regards to the data cleaning and pre-processing which took a lot of time. I could have reduced the amount of Research Questions used to tighten the scope. I believe I could have also worked better with filtering and refining datasets to only the relevant data.

Bibliography

[1] https://www.dictionary.com/browse/pokemon. (Last Accessed: 13/12/2018)
[2] https://www.pokemon.com/uk/pokedex/. (Last Accessed: 13/12/2018)
[3] https://www.britannica.com/topic/Pokemon-electronic-game. (Last Accessed: 13/12/2018)
[4] https://bulbapedia.bulbagarden.net/wiki/Generation. (Last Accessed: 13/12/2018)
[5] https://bulbapedia.bulbagarden.net/wiki/Type. (Last Accessed: 13/12/2018)
[6] https://www.serebii.net/pokedex-sm/stat/sp-attack.shtml. (Last Accessed: 13/12/2018)
[7] https://www.kaggle.com/rounakbanik/pokemon. (Last Accessed: 13/12/2018)
[8] Alberto Barradas. Pokemon with stats. https://www.kaggle.com/abcsds/pokemon. (Last Accessed: 13/12/2018)
[9] https://github.com/lgreski/pokemonData/blob/master/Pokemon.csv. (Last Accessed: 13/12/2018)

Word Count – 1540 (Not including this sentence, bibliohgraphy and figure sub headings.