

University of Strathclyde

MSc Artificial Intelligence and Applications

CS978 – Big Data Technologies

Jason Swan

201956468

Word Count: 3093 words

**Contents**

List of Figures I

List of Tables II

**1 Introduction** 1

**2 Introduction to the Pokémon Dataset** 4

**3 Primary Analysis**  8

**4 Unsupervised Analysis**  12

4.1 Preparing the dataset 12

4.2 K Means Clustering 13

**5 Supervised Learning**  15

5.1 Preparing the dataset 15

5.2 K Nearest Neighbours 15

**6 Conclusion**  18

**Bibliography**  19

**Appendices** 20

**List of Figures**

**Figure I:** Bar chart- Number of Pokémon per Generation 2

**Figure II:** Pie chart- Pokémon Alternate Forms 6

**Figure III:** Bar plot- Pokémon Type Comparison 7

**Figure IV:** Pie chart- Smogon Tier Percentage Ratio 9

**Figure V:** Violin Plot- Pokémon Base Stat Totals 11

**Figure VI:** Confusion Matrix for KNN Classifier 16

**Figure VII:** Bar plot- Pokémon Primary Type Comparison 20

**Figure VIII:** Bar plot- Pokémon Secondary Type Comparison 21

**Figure IX:** Density Plot & Histogram of Uber Pokémon 22

**Figure X:** Density Plot & Histogram of Over Used Pokémon 22

**Figure XI:** Density Plot & Histogram of Under Used Pokémon 23

**Figure XII:** Density Plot & Histogram of Rarely Used Pokémon 23

**Figure XIII:** Density Plot & Histogram of Never Used Pokémon 24

**Figure XIV:** Density Plot & Histogram of Partially Used Pokémon 24

**Figure XV:** Scatterplot Attack & Speed 25

**Figure XVI:** Scatterplot Special Attack & Speed 25

**Figure XVII:** Scatterplot Defense & Hit Points 26

**Figure XVIII:** Scatterplot Special Defense & Hit Points 26

**Figure XIX:** Scatterplot Special Defense & Hit Points 27

**List of Tables**

**Table I:** Rotom Comparison 8

**Table II:** Mean Base Stats per Tier 10

**Table III:** Example Classification Report for KNN Classifier 17

**1 Introduction**

Plunkett (2011) states that as a child Satoshi Tajiri collected and classified insect, which would one day inspire him to create the worldwide phenomenon known as Pokémon. Pokémon exist with the Pokémon Universe, a universe which is collectively made up of multiple regions. The Pokémon franchise began in February 1996 with the release of Pokémon Red & Green within Japan, with an improved version Pokémon Blue following shortly after. Western audiences were introduced to the idea of Pokémon in September 1998, as noted by Giant Bomb, with the hype behind this being minimal until it was revealed that a game designer involved with the company had introduced a secret hidden Pokémon at the last minute. This mystery Pokémon was Mew and with its inclusion becoming public knowledge shortly afterwards sales of the games skyrocketed with this signalling the start of Pokémon’s meteoric rise to fame.

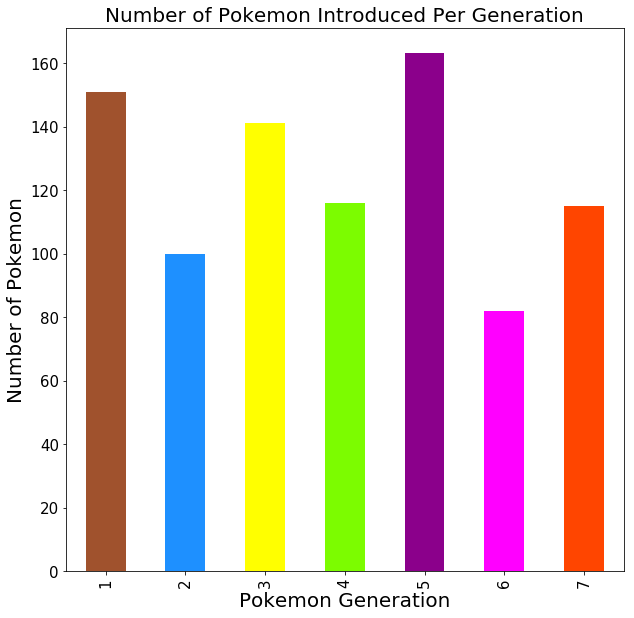
The Pokémon company introduced the world to 151 monsters in this first series of games with them continuing to expand the universe with new main series games every couple of years. The total number of Pokémon now stands at above 900 (when including alternate forms for Pokémon), with them introducing new Pokémon with each new generation as can be seen from Figure I below.

The Pokémon franchise shows no signs of slowing down with it being named the highest grossing media franchise of all time in a 2018 article by Megan Peters, and with it set to release the 8th generation of main series games on November 15th, 2019- Pokémon Sword & Shield.

While the main series games remain a staple of the Pokémon series the franchise has been expanding since it’s humble origins. The franchise now includes a multitude of media branches which encompass the moto of the franchise- “Gotta Catch ‘Em All!” The franchise now includes the following media:

* The main series games
* The spin off games, such as the Mystery Dungeon series
* The Pokémon Trading Card Game
* The Pokémon animated series, which includes 20 feature length animated movies
* Various Pokémon manga
* Pokémon Go, an augmented reality game which swept the world with its popularity

**Figure I: Number of Pokémon per Generation.**



While the franchise continues to grow, the main series games remain a staple of the universe, and as such will be the focus of this report. In the main series games Pokémon battle in a turn-based system with each Pokémon having specific attributes which determine the amount of damage they can inflict on one another. As such the purpose of this report will be to consider how these attributes interact with one another, to determine if how a Pokémon fares in battle, and thus in the metagame, can be determined by considering these special attributes (known as Base Stats).

In an attempt to investigate this, I will:

* Present and interpret summary statistics
* Conduct a K Means clustering analysis
* Perform KNN classification

**2 The Pokémon Dataset**

When looking for a dataset on which to conduct an analysis there existed a wealth of them available for download from [www.kaggle.com](http://www.kaggle.com), however none of them were complete in terms of including all of the information which I was looking for. As such I researched and decided to build my own Pokémon dataset which would include everything that I felt could potentially be of interest for the purposes of this analysis.

When looking for general information regarding Pokémon and their the base stats the website [www.serebii.net](http://www.serebii.net) was invaluable, while I found information regarding Pokémon’s tier placement within online metagames at [www.smogon.com](http://www.smogon.com).

Specific attributes are used to describe each Pokémon, some unique (such as their Pokédex number and name), some which apply to multiple Pokémon (such as their type) and some which vary (such as the base stats of a Pokémon).

The following columns denoting these attributes, and more, were included in the dataset:

* **generation**- a number in the range 1 – 7 which denotes the generation when the Pokémon was introduced.
* **pokédex\_number**- a number unique to each Pokémon which can be used for the purposes of identification. Alternate forms of Pokémon share the same Pokédex number.
* **name**- the unique name given to each Pokémon.
* **type\_1**- the primary type of each Pokémon, see Figure 7
* **type\_2**- the secondary type of each Pokémon, where applicable, see Figure 8
* **hp**- the total hit points of the Pokémon. This stat determines the vitality of the Pokémon. When it reaches 0 the Pokémon faints.
* **attack**- the base attack stat of the Pokémon. This is used to determine the damage caused by physical attacks used.
* **defense**- the base defense stat of the Pokémon. This is used to determine the damage taken from physical attacks.
* **sp\_attack**- the base special attack stat of the Pokémon. This is used to determine the damage cause by special attacks used.
* **sp\_defense**- the base special defense stat of the Pokémon. This is used to determine the damage taken from special attacks.
* **speed**- the base speed stat of the Pokémon. The Pokémon with the higher speed moves first in Pokémon’s turn based battle system.
* **base\_total**- the base stat total for the Pokémon. This is a combination of each of the 6 stats previously highlighted.
* **tier**- the tier the Pokémon battles in competitively. These placements were taken from [www.smogon.com](http://www.smogon.com). The tiers listed are Uber, OU (over used), UU (under used), RU (rarely used), NU (never used), PU (partially used), NFE (not fully evolved), LC Uber (little cup uber), and, LC (little cup).
* **legendary**- a Boolean operator denoting whether or not a Pokémon is considered to have legendary status.
* **alternate**- a Boolean operator denoting whether or not a Pokémon has an alternate form. Only the base form is noted as having an alternate form.
* **mega**- a Boolean operator denoting whether or not a Pokémon is considered a Mega Evolution. Mega Evolution is a temporary in battle transformation which increases the base stats of the Pokémon by around 100 points.

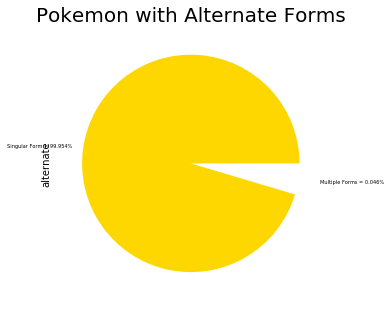
While there are a number of characteristics of Pokémon which the above variables do not take into account, the chosen variables are possible explanatory factors which affect the tier placement of a Pokémon, with the assessment of how great a role some of these attributes play in this being the purpose of this report.

Mega Evolution is a temporary in battle transformation which can affect both the type and the base stats of a Pokémon. A total of 52 Pokémon are considered to be Mega evolutions, which includes the Primal forms of Groudon and Kyogre, the Ash-Greninja form of Greninja, and Ultra Necrozma. As Mega Evolution is a temporary transformation which grants the Pokémon an increase in their base stat total of 100 points a decision was made to exclude these Pokémon for the purpose of this analysis in order that this information did not significantly skew the data with many of these Pokémon having much higher stats than their non mega counterparts.

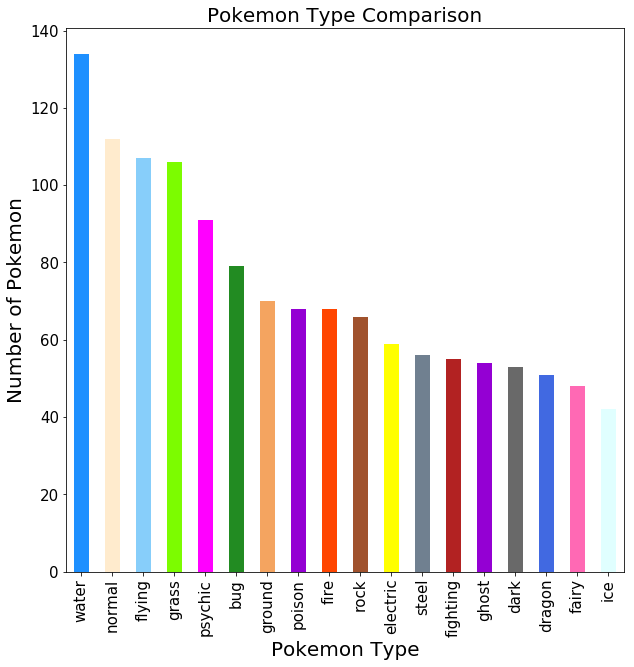
Upon its introduction Pokémon had a total of 15 types, with a further 2 introduced within the second generation, before a final type was introduced with the introduction of the fairy type in generation six of the main series games. Figure III below highlights the total number of Pokémon which exist within each of the 18 types.

Of the Pokémon which exist many of them have multiple forms, with some Pokémon such as Castform having four different forms. For the purposes of analysis, the base form of Pokémon is considered as having an alternate form if the change affects either their type or their base stat total. The Pokémon Arceus and Silvally, which are normal type before the holding of items, are classed as normal type only as held items are not considered for the purposes of this investigation. Figure II below shows the number of Pokémon which have alternate forms as being 40, which includes only the base form of each of the Pokémon.

**Figure II: Pie chart- Pokémon Alternate Forms**



**Figure III: Bar plot- Pokémon Type Comparison**

**3 Primary Analysis**

As has been touched upon throughout this report Pokémon can be classified into specific tiers based upon how viable they are in terms of their battling prowess.

Pokémon with different forms will tend to have different tier placements for each of their forms, as can be seen with the Pokémon Rotom, with the information regarding the different forms of this Pokémon being considered in Table I.

For the purposes of this report I decided to only consider Pokémon which are in their final form, who have an analysis page available on the Smogon website. In order to remove all Pokémon which were not in their final form the Pokémon who’s tier classification was NFE (Not Fully Evolved), LC Uber (Little Cup Uber) and LC (Little Cup) were removed. Pokémon classified as Untiered also made up a significant amount of the data and I felt that having such a large group who were very diverse would result in skewing the data I removed all Pokémon who’s tier placement was classed as this.

**Table I: Rotom Comparison.**

A close up of a building

Description automatically generated

The remaining Pokémon 329 make up the six tiers which will be used throughout the remainder of this report for the purposes of analysis. Figure IV highlights the ratio of Pokémon within each tier when considering the total number of Pokémon remaining within the dataset.

It would be useful to visualise how the base stats of the Pokémon within each tier vary; however, variables exist within the dataset which generate information which is not relevant. At this point the following variables were removed which provide no direct information regarding the tier placement of each Pokémon:

* generation
* legendary
* alternate

Table II below shows the mean for each of the base stats for the Pokémon within each of the tiers. As can be seen the Pokémon within the Uber tier tend to have significantly higher statistics than their counterparts in the lower tiers, while the average scores for the Pokémon within Under Used and Rarely Used show the smallest change- suggesting that other factors than their base stats could affect the tier placement of a Pokémon.

**Figure IV: Pie chart- Smogon Tier Percentage Ratio**

A close up of a logo

Description automatically generated

**Table II: Mean Base Stats per Tier**

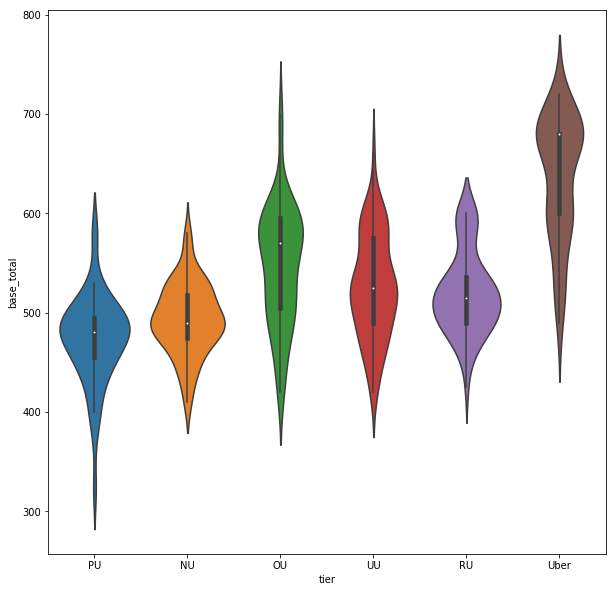
A screenshot of a cell phone

Description automatically generated

Figures 9 to 14 show the density plots and histograms for each of the six tiers, allowing for a visual interpretation of how varied the stats are within each of the tiers. They seem to show that, for the most part, the stats tend to be normally distributed, albeit with some peaks which could be ascribed to specific sets of Pokémon such as Legendary Pokémon within the Uber tier as an example.

A violin plot was generated as can be seen in Figure V, with it allowing a visual comparison of the base stat total for the Pokémon within each tier. The base stat total for the Pokémon within the Uber tier is significantly higher with few values falling below the 600 mark, while for the Over Used tier it seems that it has the highest number of Pokémon which are out with the normal range, the converse is true for the Partially Used tier with it having Pokémon who’s base stat total falls below 300.

**Figure V: Violin Plot- Pokémon Base Stat Totals**



**4 Unsupervised Analysis**

The objective here is to identify clustering within the dataset, if it exists, with my specific aim being to identify clustering within the base stats of Pokémon across the six tiers.

**4.1 Preparing the Dataset**

In order that the dataset was appropriate for a clustering analysis two separate edits were made to the data. Firstly, three columns were removed at this point. The columns removed were:

* name
* type\_1
* type\_2

The reasoning behind removing these was that they were categorical which would require employing a transformation were they to be included, and they have no specific and noticeable impact on the base stats of the Pokémon which this report seeks to investigate.

There remains a further column of categorical data related to the tier placement of each Pokémon; however, this was transformed using the tier\_rank function which I designed which took each of the tiers and transformed them into a numeric value based upon their rank. The tier and its associated numeric value are as follows:

* Uber = 0
* OU = 1
* UU = 2
* RU = 3
* NU = 4
* PU = 5

Following the employ of this function, a new variable known as Pokémon1 was assigned which was the same as the Pokémon variable with the tier column removed.

Pokémon1 contained the following columns:

* hp
* attack
* defense
* sp\_attack
* sp\_defense
* speed
* base\_total
* tier\_rank

The data contained with the Pokémon1 dataset was scaled to ensure that no numeric changes had greater explanatory power than the others and that the numeric values associated with the tier placement of the Pokémon did not cause discrepancies.

**4.2 K Means Clustering**

K Means clustering is a popular clustering method which is often employed, with it serving as an alternative to the oft used Hierarchical clustering.

K Means is a method which employs a specified number of clusters, with this optimal number often being found using the elbow method. In the case of this report we specifically wish to identify if six clusters exist, hence the number of clusters employed here being six.

The process of K Means is simple and follows the path below:

* 6 random points within the dataset are chosen to be the centre of the 6 clusters
* each point is assigned to the cluster closest to it, with the distance measure employed by K Means clustering being the Euclidean distance
* a new cluster centre point is calculated each time
* repeat for all points to be assessed

Empty lists were employed in order to allow an iteration to be conducted with the K Means Clustering analysis with their purpose being to find the mean value for both the completeness and homogeneity scores over 50 iterations.

The output for both of these are in the range 0 to 1 with a higher scoring being better.

The completeness score measures how many items of a specific type belong to a single cluster, while the homogeneity score measures how many items within a cluster are of the same type.

The average completeness score for 6 clusters was 0.2121, while the average homogeneity score was 0.1988. These low scores suggest that there is a lot of overlap within the clusters found and that the problem of determining whether or not Pokémon’s base stats determine their tier placement may not be suited to a clustering analysis. For the purposes of a visual representation of the clusters found scatterplots were generated across base stats expected to be correlated. These scatterplots are Figures 15 to 19.

**5 Supervised Learning**

The objective here is to attempt to classify each Pokémon and identify their tier placement using their base stats.

**5.1 Preparing the Dataset**

A new variable denoted Pokémon2 was introduced which contained the same columns as Pokemon1 with the exclusion of tier\_rank as this could interfere with predictions given its association with each specific tier.

The data contained with Pokémon2 was scaled in a similar manner to when investigating clustering to ensure no changes would outweigh others.

**5.2 K Nearest Neighbours**

While a number of classification methods exist, few have the appeal of KNN when considering a dataset such as the one I am working with. KNN is a simple model who’s training method is trivial and it works with any number of classes. Given I wish to classify Pokémon to specific tiers based upon the base stats this method presented itself as the most logical and effective.

The process of KNN scores an example by finding the k nearest examples (using the Euclidean distance) and then averaging their scores to classify the example.

The method involves two separate parts:

* Training
* Testing

Training:

* no explicit model required
* stores all known instances & labels

Testing:

* measure distance between new instance and all known instances
* select the closest k neighbours and rank according to numeric distance
* return the most frequent class as the most likely class of the new instance

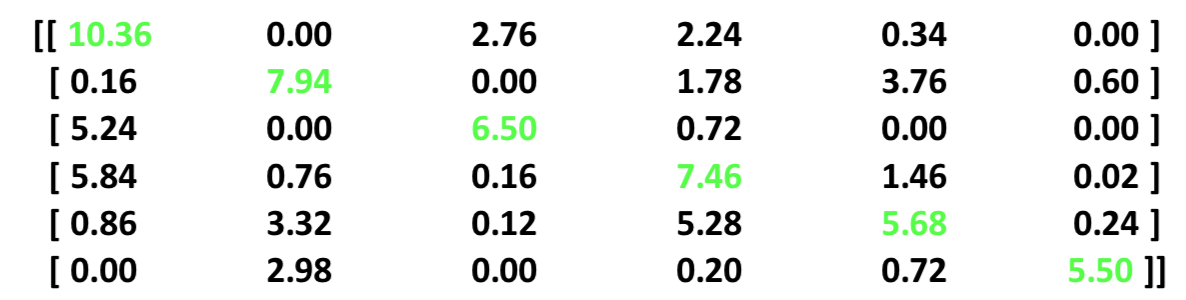
Empty lists were employed in order to allow an iteration to be conducted with the KNN classification method with their purpose being to find the mean values output for the accuracy score and the confusion matrix following 50 iterations. The classification report could not be iterated.

The average accuracy score was 0.2708, which suggests that the method of classification is not as accurate as I would have hoped.

The average values for the confusion matrix for the KNN classifier are shown in in Figure VI. The green values across the diagonal show how often the tier of the Pokémon was classified correctly.

An example classification report is provided in Table III. While the precision will vary with each iteration conducted, based upon the examples chosen, this suggests that for some tiers (Uber & OU) classification is easier than for others (NU & PU).

**Figure VI: Confusion Matrix for KNN Classifier**



**Table III: Example Classification Report for KNN Classifier**

A screenshot of a cell phone

Description automatically generated

**6 Conclusion**

As a fan of Pokémon, I have found this investigation to be highly informative.

The attempts to cluster Pokémon into six clusters based upon their tier tanking was unsuccessful and suggests that clusters do not exist within the specific factors which I considered.

Attempting to categorise Pokémon using their base stats was also unfruitful and suggests that other factors must play a significant role in the tier placement of Pokémon.

While my efforts suggest that attempting to describe Pokémon according to a limited range of characteristics were not successful it highlighted something key to the Pokémon franchise- every Pokémon is unique in some way.

Given the 8th generation of Pokémon is set to be released worldwide which not only introduces a plethora of new Pokémon, moves and abilities but also eliminates around half the existing number from the metagame a future analysis could be conducted with the new data. Who knows, maybe the new data will allow strides forward in terms of the ease of Pokémon Classification.

**Bibliography**

**Plunkett, Luke. (2011) “The Origins of Pokémon.” Available at:** <https://kotaku.com/the-origins-of-pokemon-5806664> **(Accessed 07/11/2019)**

**Peters, Megan. (2018) “Pokémon Is The Highest-Grossing Franchise Of All Time.” Available at:** [**https://comicbook.com/anime/2018/06/24/pokemon-franchise-revenue-anime-media/**](https://comicbook.com/anime/2018/06/24/pokemon-franchise-revenue-anime-media/) **(Accessed 07/11/2019)**

**“Pokémon Red/Blue” Giant Bomb. Available at:**

[**https://www.giantbomb.com/pokemon-redblue/3030-3966/releases/**](https://www.giantbomb.com/pokemon-redblue/3030-3966/releases/) **(Accessed 07/11/2019)**

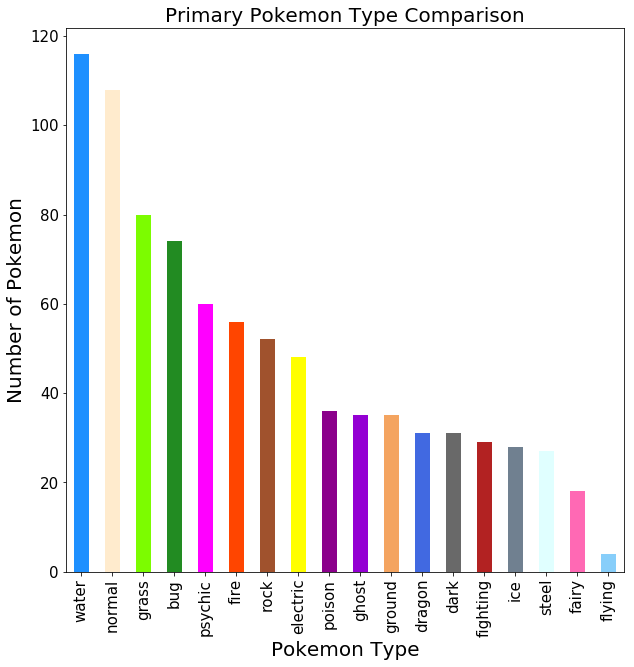
**Appendices**

**Environment**

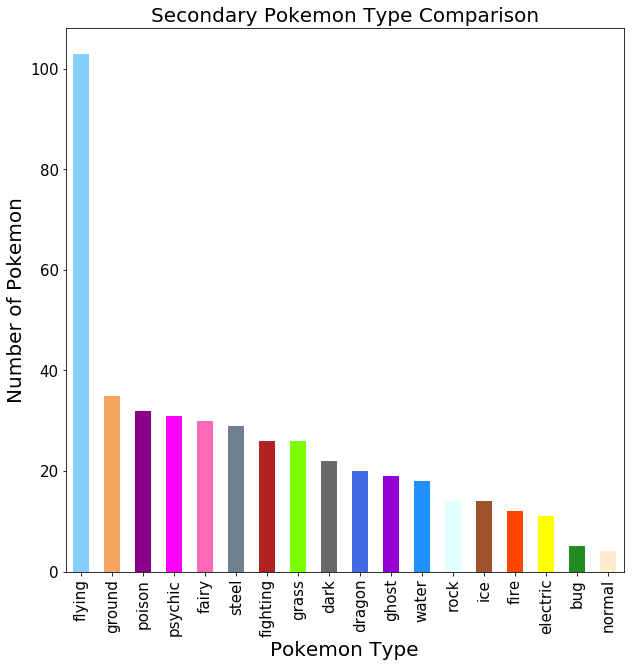
Python Version 3.7.3

**Supplementary Figures**

**Figure VII: Bar Plot- Pokémon Primary Type Comparison**

****

**Figure VIiI: Bar Plot- Pokémon Secondary Type Comparison**



**Figure IX: Density Plot & Histogram of Uber Pokémon**

A screenshot of a cell phone

Description automatically generated

**Figure X: Density Plot & Histogram of Over Used Pokémon**

A screenshot of a cell phone

Description automatically generated

**Figure XI: Density Plot & Histogram of Under Used Pokémon**

**A screenshot of a cell phone

Description automatically generated**

**Figure XII: Density Plot & Histogram of Rarely Used Pokémon**

**A picture containing screenshot

Description automatically generated**

**Figure XIII: Density Plot & Histogram of Never Used Pokémon**

**A screenshot of a cell phone

Description automatically generated**

**Figure XIV: Density Plot & Histogram of Partially Used Pokémon**

A close up of a logo

Description automatically generated

**Figure XV: Scatterplot- Attack & Speed**

A close up of a piece of paper

Description automatically generated

**Figure XVI: Scatterplot- Special Attack & Speed**

A screenshot of a cell phone

Description automatically generated

**Figure XVII: Scatterplot Defense & Hit Points**

A screenshot of a cell phone

Description automatically generated

**Figure XVIII: Scatterplot Special Defense & Hit Points**

A screenshot of a cell phone

Description automatically generated

**Figure XIX: Scatterplot Defense & Special Defense**

A screenshot of a cell phone

Description automatically generated