Statistical analysis of Pokémon

Dzmitry Kurch, DSE

Statistical learning project, June 2023

**Abstract**

This paper presents an investigation into the application of statistical learning methods for the analysis of the Pokémon dataset. The Pokémon franchise, renowned for its rich and diverse collection of creatures, provides a unique opportunity to explore the capabilities of machine learning techniques. Specifically, this study focuses on employing classification analysis, clustering, and dimensionality reduction methods to gain insights into the underlying patterns and characteristics of the Pokémon universe.

The first phase of the analysis involves classification analysis, where various supervised learning algorithms are trained on a labelled subset of the dataset. By harnessing the power of classification algorithms, we aim to accurately predict the attributes and properties of Pokémon based on their known characteristics. This analysis can be instrumental in assisting trainers and researchers in making informed decisions and developing effective strategies.

Next, dimensionality reduction methods are employed to capture the essential features of the dataset while reducing its complexity. By transforming the original high-dimensional Pokémon dataset into a lower-dimensional space, we aim to extract meaningful representations that facilitate visualization and interpretation.

Furthermore, clustering techniques are applied to group Pokémon based on their similarities. By leveraging unsupervised learning algorithms, we aim to identify inherent clusters or groupings within the Pokémon dataset, which may reveal hidden relationships or evolutionary patterns.

Overall, this research showcases the potential of statistical learning methods in unlocking valuable insights from the Pokémon dataset. The findings from classification analysis, clustering, and dimensionality reduction can contribute to the advancement of Pokémon-related research, gameplay strategies, and educational applications. Moreover, the methodologies presented in this study can serve as a foundation for applying statistical learning techniques to other large and complex datasets in diverse domains.

1. **Introduction**

Pokémon is media franchise that began as a pair of Role Playing Games (RPG) video games for the original Game. Boy that were developed by Game Freak and published by Nintendo, in 1996. With the passage of time Pokémon increased its popularity and thus its owners ended up producing many animated television shows, films, trading card games, various manga comics, as well as a number of video games of different kinds, like the released augmented reality game Pokémon Go!, that really caught on in 2016.

Pokémon are fictitious animal-like monsters that live in the (of course, also invented) Pokémon world. Pokémon like fighting with each other, and they usually fight according to their (human) trainers’ orders. Almost all the Pokémon games include these fights, but in different manners. In some of them the user needs to rely on her or his strategy and in the strength of his or her Pokémon, whereas other video-games are more ability-based. Hence, an interesting fact of this games may well be the way the strength or the ability to fight of a Pokémon is described. This depends, once again, on the type of the video-game. For statistical analysis purposes, the most attractive way of describing the Pokémon is that of the RPGs’. First of all, because a big number of Pokémon have been introduced throughout these years -seven generations of Pokémon with the order of 100 of Pokémon in each of them. Second, in the RPGs each Pokémon is described with a big number of variables. Not only do we have the combat stats (the variables that describe the ability to fight), but also many variables that describe more details of each Pokémon, e.g. the colour or the probability of being female or male.

Thus, we can statically analyse the wide variety of variables used to describe the Pokémon, and there is a chance to find relationships between them, and also to cluster the Pokémon according to some criteria. In the rest of the report we will explore the Pokémon and their corresponding variables that appear in the RPGs.

First we will introduce the variables and instances of the dataset in *Section 2.* We will explore the variables and their potential dependencies in *Section 3*. Once we have studied the variables, we will try to predict type of the Pokémon, in *Section 4*. We are going to use a PCA as a powerful dimensionality reduction method to visualize the data in the *Section 5*. Finally we make some clustering of the Pokémon using the output from the PCA in *Section 6*, and with draw some conclusions in *Section 7*.

1. **The dataset and its variables**

The dataset is taken from Kaggle, here you may find the original source and the description <https://www.kaggle.com/datasets/alopez247/pokemon>.

By and large, in total we have 24 variables to analyze. The first two are unique identifiers of the Pokémon, the number in the Pokédex and the name. The Pokédex is encyclopaedia-like tool that can be used in the Pokémon RGBs to get information of the Pokémon. In fact, most of the variables we will use in this work are taken from the Pokédex. From the resting 21 variables, 12 are numerical (10 continuous and 2 discrete), 6 categorical and 3 boolean. Please refer to Appendix section to read the detailed description of those variables.

1. **Exploratory Data Analysis**

In this section we will try to gain an insight of the distributions of the different variables as well as some relationships between them.

Firstly, we do a quick data quality analysis by exploring the percentage of missing variables in the dataset.

Table 1 - Missing values percentage

|  |  |
| --- | --- |
| **Feature name** | **Missing percentage** |
| Egg\_Group\_2 | 73.51 |
| Type\_2 | 51.45 |
| Pr\_male | 10.68 |

Based on the table above we conclude that situation is not that bad, only 3 out of 24 features has missing values. Moreover, all of those features can be considered as *categories,* which allows us just to add an additional category *Null* and go on with the encoding later.

Now we proceed with a Bivariate variable analysis by combining the Univariate analysis with an attempt to explore the relationship between variables. Let’s start with plotting a correlation matrix of the numerical features. As you may se on the picture below there are a few groups of cross-correlated features which were expected based on their nature. For example, Pokémon fight characteristics such as *(speed) attack, (speed) defence and health points* are positively correlated because usually Pokémons are constructed in a “balanced” way. Also an interesting observation is that *catch rate* is negatively correlated with almost all characteristics, which means catching a Pokémon with high fight capability is harder.

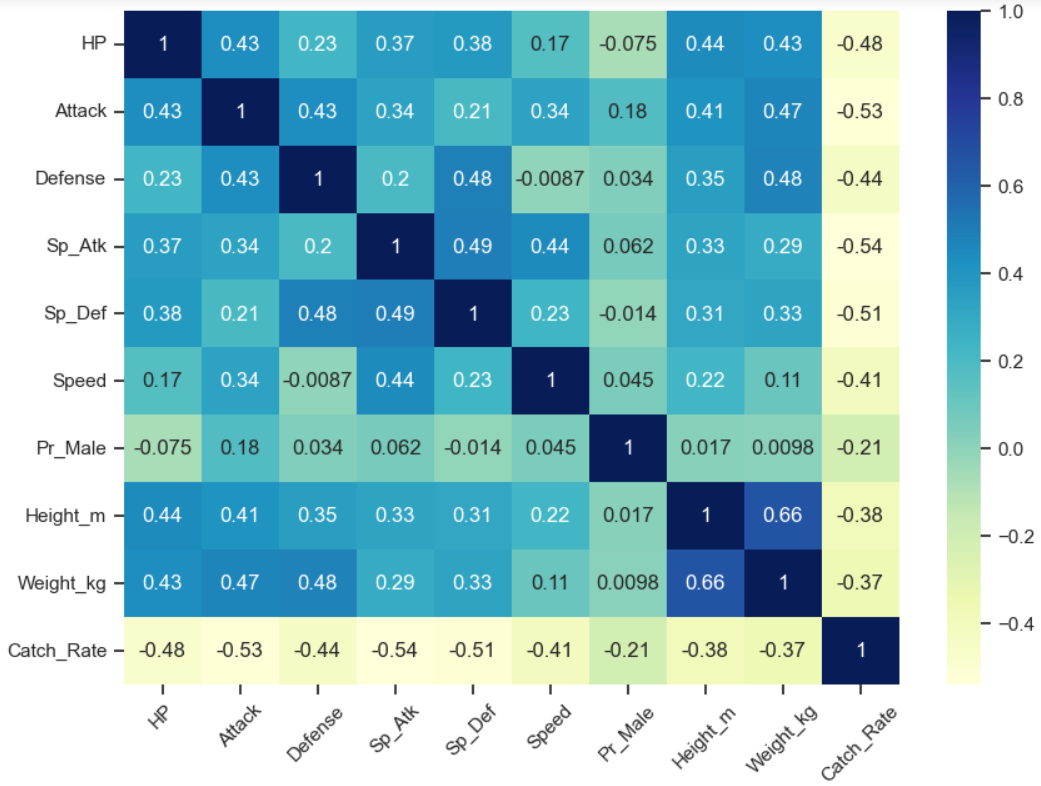
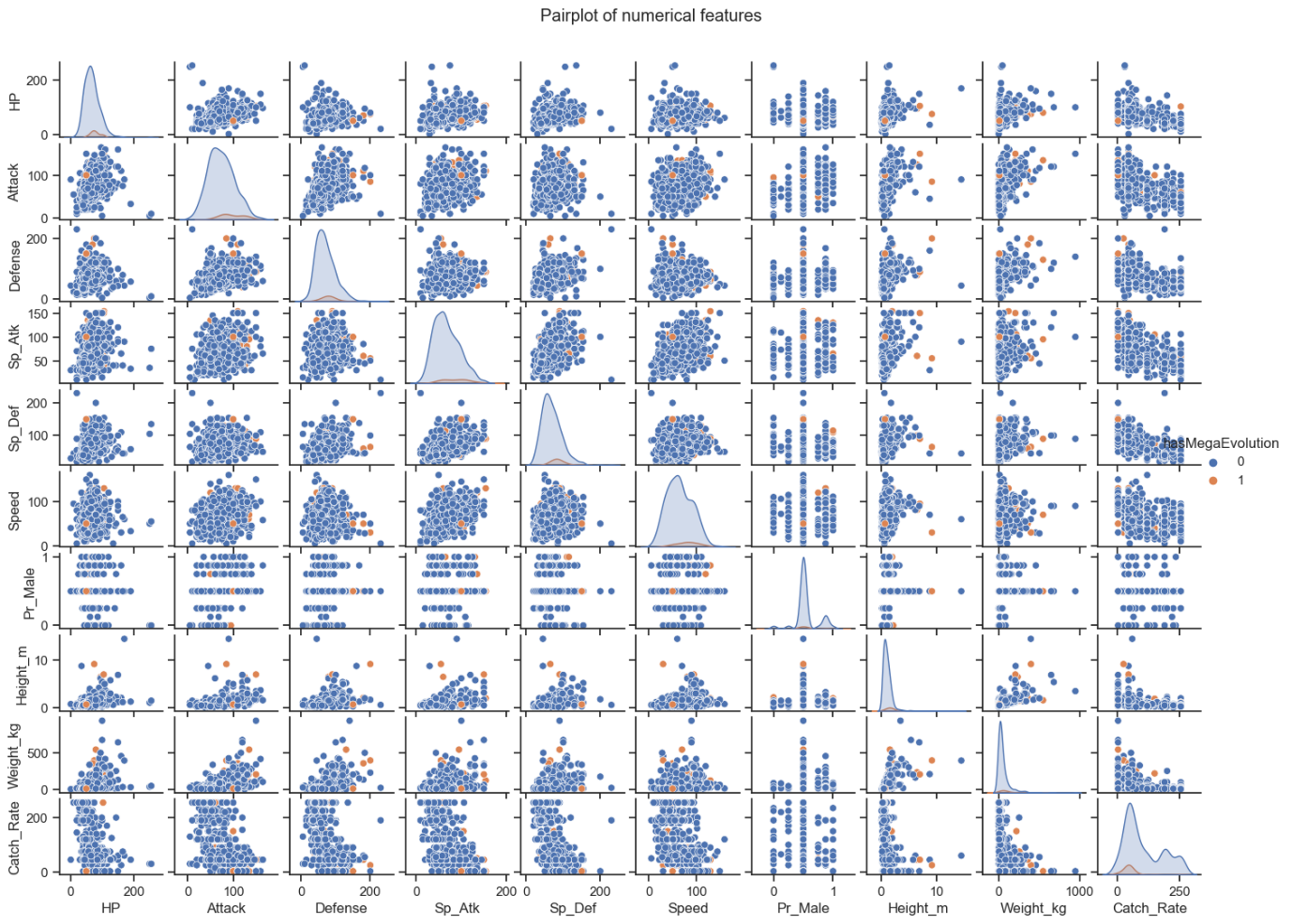


Figure 1 - Correlation matrix of numerical features

Figure 2 - Pairplot of numerical features

Our main goal for that project will be to predict the *hasMegaEvolution* characteristic which is property that some Pokémon have and allows them to change their appearance, types, and stats during a combat into a much stronger form. Among the Pokémon players they are considered to be very valuable, that is why it is important to be able to know that fact before you put an effort into catching the Pokémon itself. Important note: our dataset will be highly unbalanced containing only ***6.4%*** samples with *hasMegaEvolution* property.

Based on the pair plot above we observe an interesting tendency that most of some Pokémons who hold a mega evolution tend to “cluster” in the right upper corner of the fight characteristics scatter plots, which means that a high value of those characteristics may be a good sign for that type.

After that I analysed the percentage of *hasMegaEvolution* Pokémons across numerical and categorical variables to highlight the potentially most valuable characteristics which can help us to spot that type. Here are only a few examples.

Across Pokémons with the highest attack value (more than 100) there are ***15%*** of the mega-evolutionary, which is almost 3 times higher than the total mean.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 3 - Distribution of MegaEvolution poke for different attack ranges

Across Pokémons with the defence range [72, 90] there are ***14%*** of the mega-evolutionary ones, which suggests that feature has a good predictive potential.

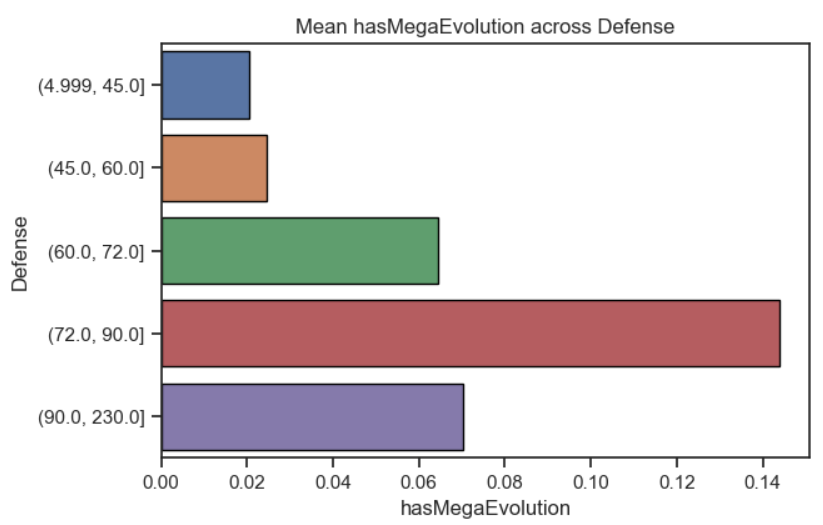
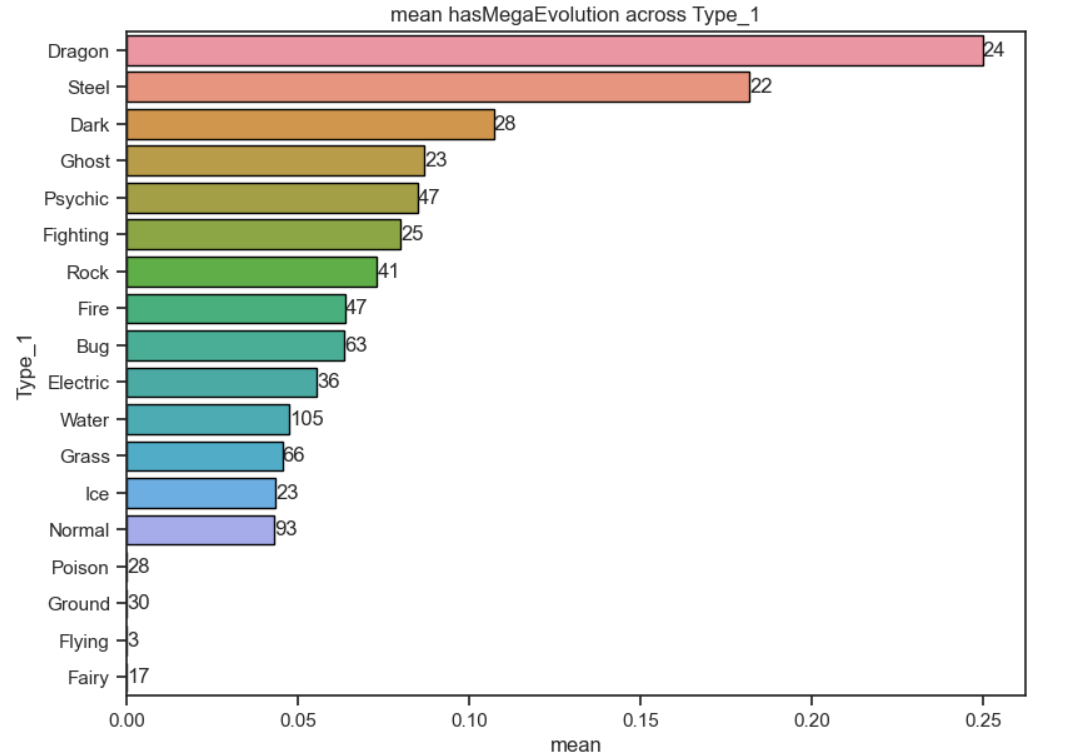
A plot above shows us that the *Dragon type* has the highest percentage of mega-evolutionary Pokémons ***25%*** with the 24 samples considered.

Figure 4 - Distribution of MegaEvolution poke for different defence ranges



1. **Classification**

Some text

1. **Dimensionality Reduction**

Some text

1. **Clustering**

Some text

1. **Conclusions**

Some text

**Appendix**

**Full list of variables and their description**

* ***Type\_1***. Primary type of the Pokémon. It is related the nature, with its lifestyle and with the movements it is able to learn for the fighting time. This categorical value can take 18 different values: *Bug, Dark, Dragon, Electric, Fairy, Fighting, Fire, Flying, Ghost, Grass, Ground, Ice, Normal, Poison, Psychic ,Rock, Steel*, and *Water*.
* ***Type\_2***. Pokémon can have two types, but not all of them do. The possible values this secondary type can take are the same than the variable *Type\_1*.
* ***Total***. The sum of all the base battle stats of a Pokémon. It should be a good indicator of the overall strength of a Pokémon. It is the sum of the next six variables. Each of them represents a base battle stat. All the battle stats are continuous yet integer variables, i.e. the number of values they can take is infinite in theory, or just very big in the practice.
* ***HP***. Base health points of the Pokémon. The bigger it is, the longer the Pokémon will be able to stay in a fight before they faint and leave the combat.
* ***Attack***. Base attack of the Pokémon. The bigger it is, the more damage its physical attacks will deal to the enemy Pokémon.
* ***Defense***. Base defense of the Pokémon. The bigger it is, the less damage it will receive when being hit by a physical attack.
* ***Sp\_Atk***. Base special attack of the Pokémon. The bigger it is, the more damage its special attacks will deal to the enemy Pokémon.
* ***Sp\_Def***. Base special defense of the Pokémon. The bigger it is, the less damage it will receive when being hit by a special attack.
* ***Speed***. Base speed of the Pokémon. The bigger it is, the more times the Pokémon will be able to attack to he enemy.
* ***Generation***. The generation where the Pokémon was released. It is an integer between 1 and 6, so it is a numerical discrete variable. It could let us analyze the development or the growth of the game through the years.
* ***isLegendary***. Boolean indicating whether the Pokémon is legendary or not. Legendary Pokémon tend to be stronger, to have unique abilities, to be really hard to find, and to be even harder to catch.
* ***Color***. Color of the Pokémon according to the Pokédex. The Pokédex distinguishes between ten colors: *Black, Blue, Brown, Green, Grey, Pink, Purple, Red, White*, and *Yellow*.
* ***hasGender***. Boolean indicating the Pokémon can be classified as male or female.
* ***Pr\_Male***. In case the Pokémon has Gender, the probability of its being male. The probability of being female is, of course, 1 minus this value. Like Generation, this variable is numerical and discrete, because although it is the probability of the Pokémon to appear as a female or male in the nature, it can only take 7 values: *0, 0.125, 0.25, 0.5, 0.75, 0.875,* and *1*.
* ***Egg\_Group\_1***. Categorical value indicating the egg group of the Pokémon. It is related with the race of the Pokémon, and it is a determinant factor in the breeding of the Pokémon. Its 15 possible values are: *Amorphous, Bug, Ditto, Dragon, Fairy, Field, Flying, Grass, Human-Like, Mineral, Monster, Undiscovered, Water\_1, Water\_2,* and *Water\_3.*
* ***Egg\_Group\_2.*** Similarly to the case of the Pokémon types, Pokémon can belong to two egg groups.
* ***hasMegaEvolution***. Boolean indicating whether a Pokémon can mega-evolve or not. Mega-evolving is property that some Pokémon have and allows them to change their appearance, types, and stats during a combat into a much stronger form.
* ***Height\_m.*** Height of the Pokémon according to the Pokédex, measured in meters. It is a numerical continuous variable.
* ***Weight\_kg.*** Weight of the Pokémon according to the Pokédex, measured kilograms. It is also a numerical continuous variable.
* ***Catch\_Rate***. Numerical variable indicating how easy is to catch a Pokémon when trying to capture it to make it part of your team. It is bounded between 3 and 255. The number of different values it takes is not too high notwithstanding, we can consider it is a continuous variable.
* ***Body\_Style***. Body style of the Pokémon according to the Pokédex. 14 categories of body style are specified: *bipedal\_tailed, bipedal\_tailless, four\_wings, head\_arms, head\_base, head\_legs, head\_only, insectoid, multiple\_bodies, quadruped, serpentine\_body, several\_limbs, two\_wings,* and *with\_fins*.