Statistical analysis of Pokémon

Asier López Zorrilla

January, 2017

Contents

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 recently 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 Pokemon, e.g. the color or the probability of being female or male.

Thus, we can statically analyze 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 cluster the Pokémon, in Section 4. We will end trying make some predictions of the variables using the others in Section 5, and with the conclusions in Section 6.

2 The dataset and its variables

When I first thought about applying some data analysis or machine learning techniques to Pokémon, I searched for a sufficiently wide dataset in terms of number of variables and in number of instances, and the only dataset I found interesting was the named "721 Pokemon with stats" dataset [1]. This contains 11 variables per Pokémon (without taking into account the Pokémon name nor its ID) that mainly define their ability to fight. But, as mentioned above, the Pokémon RGBs offer many more possible variables, which might not be useful when fighting (even though they sometime are) but definitely allow us to perform a more exhaustive analysis.

So I decided to expand (and slightly modify) the "721 Pokemon with stats" so as to include as many variables as possible. The result is a database with 10 additional variables, which can be found at Kaggle with the name of "Pokémon for Data Mining and Machine Learning" [2]. Anyway, with the recent release of the seventh generation of the Pokémon RGB, the database is already a little out-of-date, since it only includes 721 Pokémon, the ones corresponding to the first six generations, because the seventh generation was not released yet when I finished the database.

Let us now examine the 23 columns of the dataset. The first two are unique identifiers of the Pokémon, the number in the *Pokédex* and the name. The Pokédex is encyclopedia-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. We will explain in more detail what these variables represent next.

- **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 the 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.*

An unconditional fan of Pokémon could argue that this way of filling the dataset does not take into account all the possible stuff that happens in the games. In fact, it does not. For example, there are some multiform Pokémon where, in order to keep the Pokémon ID (the **Number** variable) unique so as not to perturb possible measures like the number of Pokémon released per generation, only one form is considered. On the other hand, we take into account if a Pokémon can mega-evolve, but we do not count the mega-evolved Pokémon as another Pokémon itself. All the casuistry of the specific choices when there is ambiguity when choosing the variables of a given Pokémon is listed below:

- **Mega-evolutions** are not considered as independent Pokémon.
- **Kyogre, Groudon.** Primal forms not considered.
- **Deoxis.** Only normal form considered.
- **Wormadam.** Only plant form considered.
- **Rotom.** Only normal form considered, the one with types Electric and Ghost.
- **Giratina.** Origin form considered.
- **Shaymin.** Land form considered.
- **Darmanitan.** Standard mode considered.
- **Tornadus, Thundurus, Landorus.** Incarnate form considered.
- **Kyurem.** Normal form considered, not white or black forms.
- **Meloetta.** Aria form considered.
- **Mewstic.** Both male and female forms are equal in the considered variables.
- **Aegislash.** Shield form considered.
- **Pumpkaboo, Gourgeist.** Average size considered.
- **Zygarde.** 50% form considered.
- **Hoopa.** Confined form considered.

3 General descriptive analysis of the variables

3.1 Univariate 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. We will first focus in the univariate analysis. We will develop a different analysis per variable. In the case that the variable is continuous a Kernel density plot will be carried out [3], being each kernel a gaussian. On the other hand, if the variable is whether categorical or boolean, its histogram will be plotted. To finish, it has to be said that the two numerical and discrete variables will be treated in two different manners. Since **Pr_male** refers to a probability, it seems reasonable to treat it as a continuous variable for this analysis. But **Generation** does not have that intrinsic continuous meaning. In fact, it would make little to make its density plot, because in that case we could wrongly interpret that a Pokémon may have been released in the generation 3.14, for instance. Therefore we will plot its histogram, and check how many Pokémon were released in each generation.

Figure 1 shows the histograms of the primary and secondary types, as well as the first and second egg groups of the Pokémon. The most common primary types are *Water, Normal,* and *Grass*, while the most common secondary type is *Flying*, because most of the times a Pokémon is able to fly the other type is considered first. We can also see that more or less the half of the Pokémon do not have any secondary type.

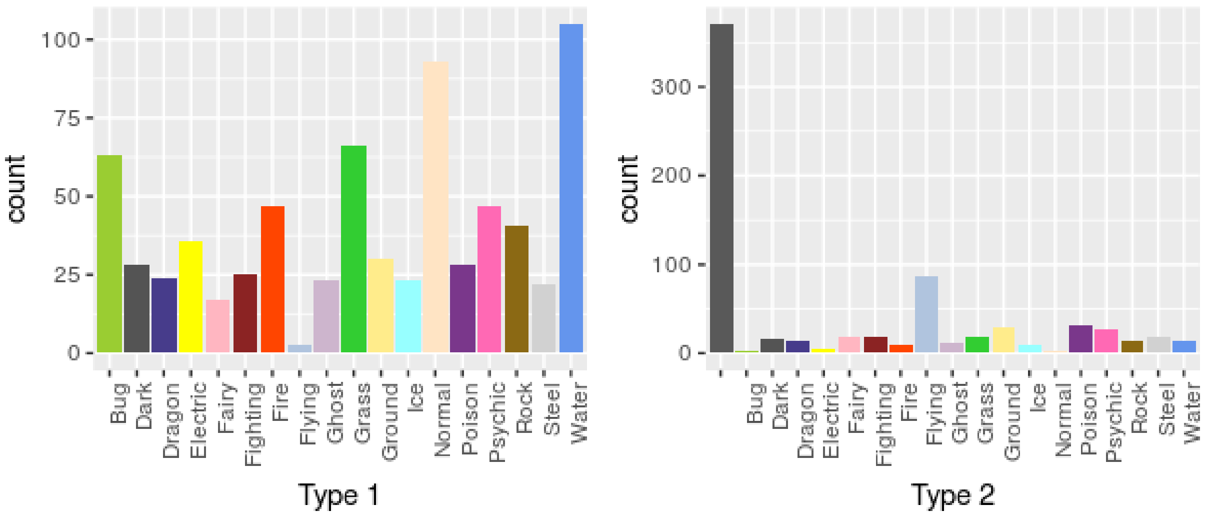


Figure 1: Histograms of the primary (left) and secondary (right) types. The first value shown in the histogram corresponding to the secondary type the first value in the x is the count of the Pokémon with no such type.

We can also look for the more and less common egg groups in the histograms shown in Figure 2. We can observe that the most common egg groups (that can somehow be understood like races) are *Field*, followed by *Water_{1+2+3}*, *Monster, Undiscovered*, and *Bug*. Pokémon from the *Field* egg group tend to be terrestrial creatures. Pokémon from the three water egg groups live in or around the water, and *Monster* group Pokémon are usually among the most powerful. *Undiscovered* egg group is characterized by its members' inability to breed. Most of the Pokémon in the group are baby Pokémon, or legendary Pokémon.

Thus, we can get an idea of the ratio of the legendary non-legendary Pokémon ratio, but since we have a variable just with that attribute of the Pokémon, we can just make it histogram. Along with it, we will plot the remaining boolean and categorical variables, as well as the counts of released Pokémon per generation. Thus, Figure 3 shows the histograms of the variables **isLegendary**, **hasGender**, **hasMegaEvolution**, **Color**, **Body_Style**, and

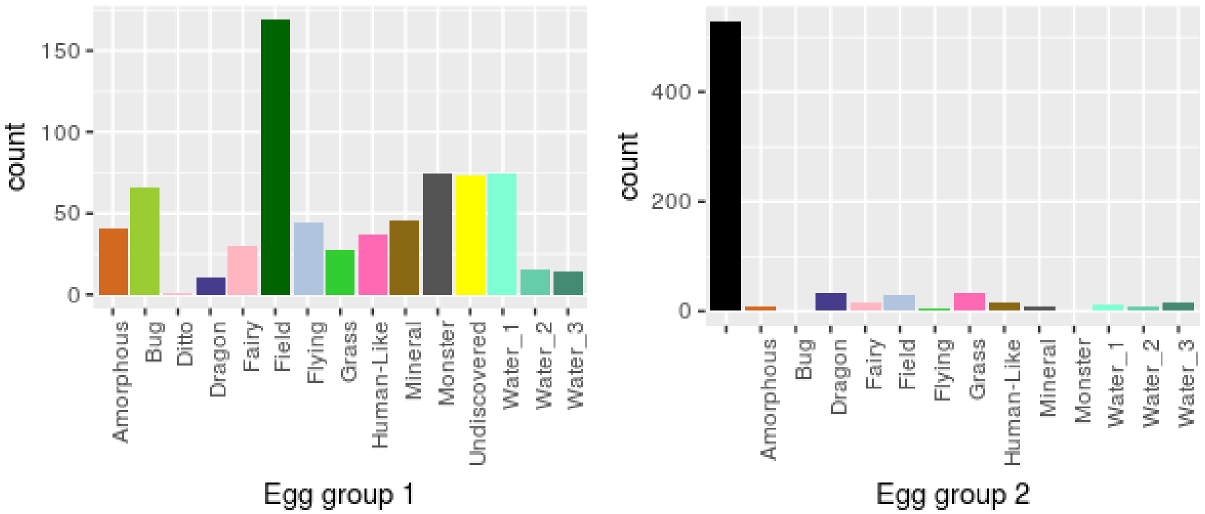


Figure 2: Histograms of the first (left) and second (right) egg groups. The first value of the second histogram is also the count of the Pokémon with no second egg group.

Generation. The three graphs in the top show that it is quite uncommon for a Pokémon to be legendary, or to be genderless, or to be capable of mega-evolving. On the other hand, the graphs in the bottom of the figure indicate that the most common colors of a Pokémon are *Blue* and *Brown*, and the most common body shapes are *bipedal* (whether *tailed* or *tailless*) and *quadruped*. Finally, the last histogram shows that while the number of released Pokémon per generation oscillated between 100 and 150 approximately in the first five generations, this quantity was reduced in the sixth to less that 75.

The remaining variables are all numerical and continuous, so, as it has been said above, we will try to understand them with the assistance of Kernel Density Plots. All these plots are shown in Figure 4. Though, of course, the range of they variables are distinct, all the shapes are similar. More or less, all the distributions seem to be gaussian. To test this hypothesis, we carried out several Shapiro normality test [4], one per variable. All the resulting *p*-values were lower than 0.1, and hence we can assume that the plotted distributions are, effectively gaussian.

3.2 Relations and dependencies between variables

3.2.1 Between numerical variables

Now we will go deeper in the dataset and explore the possible relations between the mentioned variables. First we can, for instance, try to see which of the numerical variables are linearly correlated. To this end we computed the Pearson correlation coefficient [5] between all these variables, which are graphically shown in Figure 5. Note that, since the variable **Pr_Male** is not always defined, because some Pokémon are genderless, we filtered the instances for this experiment so as all the numerical variables to be defined.

There are many conclusions to be drawn from this figure. First of all, there is a big and positive correlation between **Total** and the rest of the combat stats. This was totally expected, because as we said in Section 2 **Total** is the sum of the other combat stats. Following with the combat stats, we can deduce that normally, when a Pokémon is good fighting, it will probably be more or less good in all the aspects. We can say that it will be better in general because the correlation between all the combat stats are positive (except the case of **Defense** and **Speed**). However, we say more or less because not all the correlations are equally positive. Actually, we could perfectly consider that only some of them are positive enough so as to say that the variables are really correlated. The

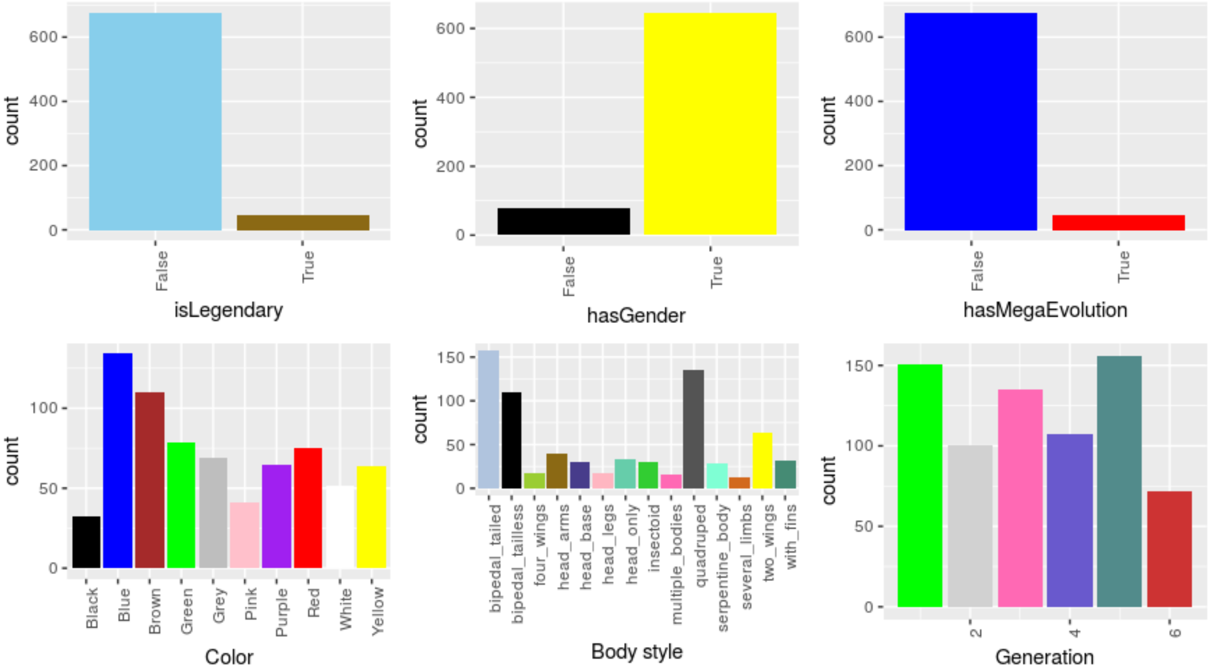


Figure 3: From left to right, and from top to bottom, histograms of **isLegendary**, **hasGender**, **hasMegaEvolution**, **Color**, **Body_Style**, and **Generation**.

strong relations seem to be **Attack** with **HP**, **Defense** with **Attack** and **Sp_Def**, **Sp_Def** with **Sp_Atk** and **Defense**, and **Speed** with the two attack stats.

Then, we can notice that the **Weight_kg** and the **Height_m** are very related. Moreover, on average the heavier and the taller a Pokémon is, the higher its battle stats will be, with the exception of the speed. It makes sense, because usually (in the real) the bigger an animal is, the more powerful it is, but usually its mobility will also be reduced. The next dependency we can find in Figure 5 is that, in general, the more powerful a Pokémon is, it will also be harder to catch it, i.e the bigger the combat stats, the **Weight_kg**, and the **Height_m**, the lower the **Catch_rate**. This is a very general tendency in the world of the video-games, and probably is also applicable in many other areas, big rewards require bigger efforts. To continue, there is a small albeit unexpected relation. It is the one between the **Pr_Male** and **Catch_Rate**. Strange though it may seem, the correlation coefficient between these variables indicate that Pokémon that often appear as male in the nature are slightly more likely be hard to catch. Finally, we can say that there seem to be any relation between the **Generation** and any other variable.

3.2.2 Between categorical variables

In order to analyze the dependencies between the variables we first ran a chi square independence test [6] to check whether these are statistically dependent or independent. According to the results are shown in 6, in most of the cases we can reject the hypothesis that the variables are independent, since the p-value is lower than 0.05 (and thus is plotted as a 0). Notwithstanding, we cannot appreciate how strong the dependencies are, because the p-values tend to be very similar when lower than 0.05 (in order not to show even more results, the exact numbers are not plotted). In order to quantify the dependency between variables we ran another experiment: we computed the Normalized Mutual Information or NMI (Equation 1).

$$NMI(X,Y) = 2\frac{MI(X,Y)}{\min(H(X),H(Y))}, \tag{1}$$

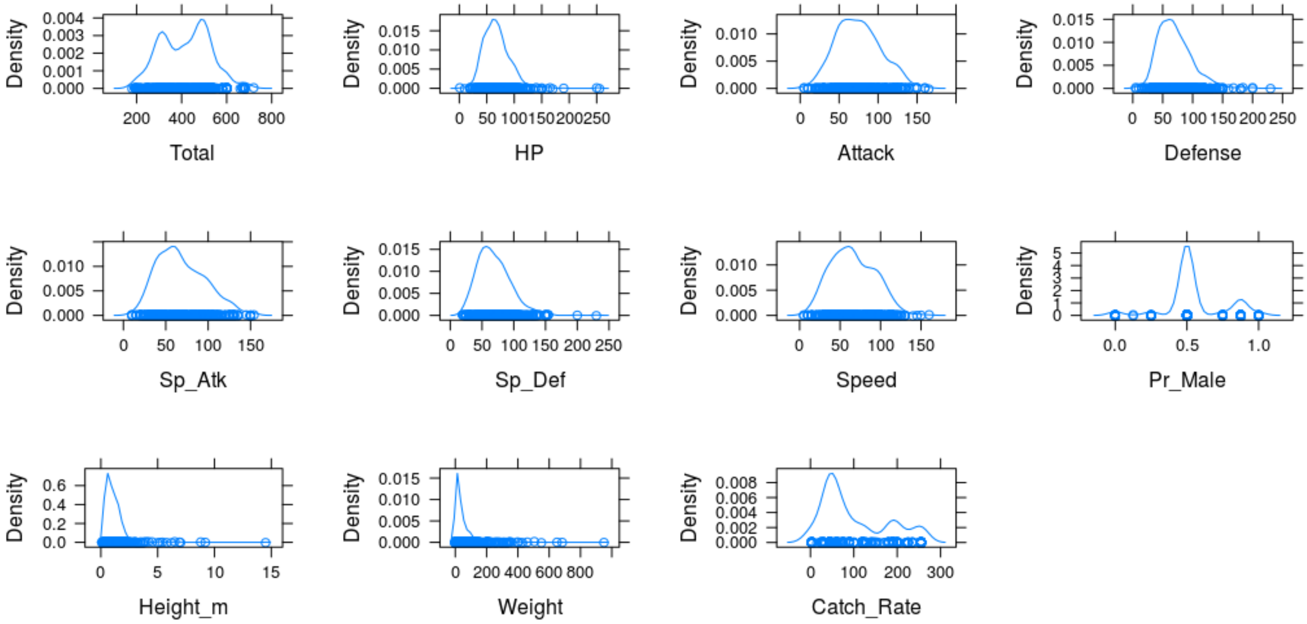


Figure 4: From left to right, and from top to bottom, histograms of **Total**, **HP**, **Attack**, **Defense**, **Sp_Atk**, **Sp_Def**, **Speed**, **Pr_Male**, **Height_m**, **Weight_kg** and **Catch_Rate**.

where $H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$ is the entropy the random variable X (being x each of its possible values), and $MI(X,Y) = H(X) + H(Y) - H(X,Y)$ the Mutual Information between the variables X and Y. The NMI is one of the normalized versions of the Mutual Information, which is supposed to be more robust against the number of possible values a variable can take, making the it suitable when we want to compare many variables where the number of possible values are different, that is exactly this case. The bigger the NMI is, the more dependent the variables are. Therefore, we can deduce that the **Egg_Group_1** of the Pokémon, its **Color** and **isLegendary** variables are closely related. Additionally, we can observe that shared information between variables **Type_1** and **isLegendary**, and between **Body_Style** and **isLegendary** are quite great. According to the two measures we have applied (the p-value of the chi square independence text and the NMI), note that more or less, all the pairs of variables which their p-value was lower than 0.05, also have NMI greater or equal than 0.1.

We can go deeper in the found dependencies. For example, it is easy to analyze the one between **Egg_Group_1** and **isLegendary**. If we pay attention to the data base, we will see that all the legendary Pokémon belong to the **Egg_Group_1** *Undiscovered*, but that is the only implication, because not all the Pokémon with textbfEgg_Group_1 *Undiscovered* are legendary. We can also we can plot some histograms that can let us know more of the second strongest dependency, the one between **Color** and **isLegendary**. Figure 7 shows the **Egg_Group_1** histograms for non-legendary and legendary Pokémon. It is visible that legendary Pokémon tend to be less *Brown* than non-lendaries. The values that consequently are higher are *Black* and *Yellow*.

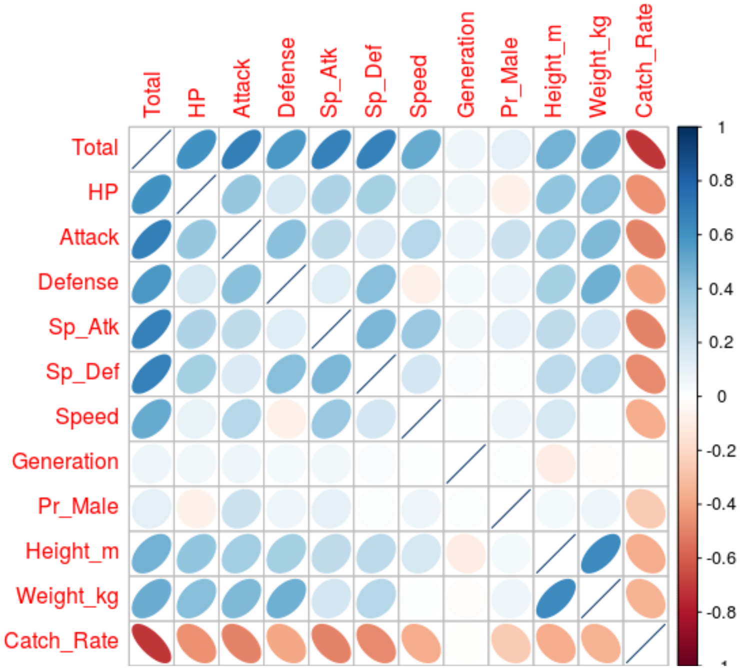
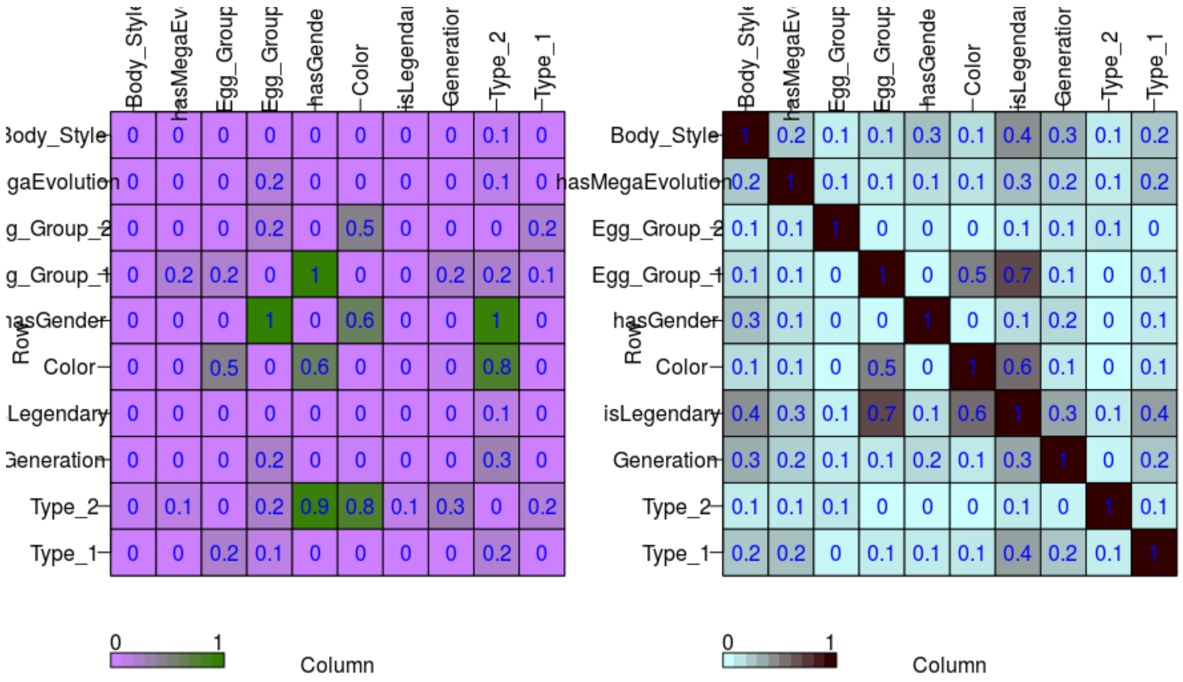


Figure 5: Graphical representation of the Pearson correlation coefficients between all the numerical variables.



	Body_Styl	hasMegaEv	Egg_Group	Egg_Group	hasGende	Color	isLegenda	Generatior	Type_2	Type_1
ody_Style	0	0	0	0	0	0	0	0	0.1	0
gaEvolution	0	0	0	0.2	0	0	0	0	0.1	0
g_Group_2	0	0	0	0.2	0	0.5	0	0	0	0.2
g_Group_1	0	0.2	0.2	0	1	0	0	0.2	0.2	0.1
asGender	0	0	0	1	0	0.6	0	0	1	0
Color	0	0	0.5	0	0.6	0	0	0	0.8	0
Legendary	0	0	0	0	0	0	0	0	0.1	0
eneration	0	0	0	0.2	0	0	0	0	0.3	0
Type_2	0	0.1	0	0.2	0.9	0.8	0.1	0.3	0	0.2
Type_1	0	0	0.2	0.1	0	0	0	0	0.2	0

	Body_Styl	hasMegaEv	Egg_Group	Egg_Group	hasGende	Color	isLegenda	Generatior	Type_2	Type_1
Body_Style	1	0.2	0.1	0.1	0.3	0.1	0.4	0.3	0.1	0.2
hasMegaEvolution	0.2	1	0.1	0.1	0.1	0.1	0.3	0.2	0.1	0.2
Egg_Group_2	0.1	0.1	1	0	0	0	0.1	0.1	0.1	0
Egg_Group_1	0.1	0.1	0	1	0	0.5	0.7	0.1	0	0.1
hasGender	0.3	0.1	0	0	1	0	0.1	0.2	0	0.1
Color	0.1	0.1	0	0.5	0	1	0.6	0.1	0	0.1
isLegendary	0.4	0.3	0.1	0.7	0.1	0.6	1	0.3	0.1	0.4
Generation	0.3	0.2	0.1	0.1	0.2	0.1	0.3	1	0	0.2
Type_2	0.1	0.1	0.1	0	0	0	0.1	0	1	0.1
Type_1	0.2	0.2	0	0.1	0.1	0.1	0.4	0.2	0.1	1

Figure 6: *p*-values of the chi square independence test in the left, and Normalized Mutual Information in the right hand side.

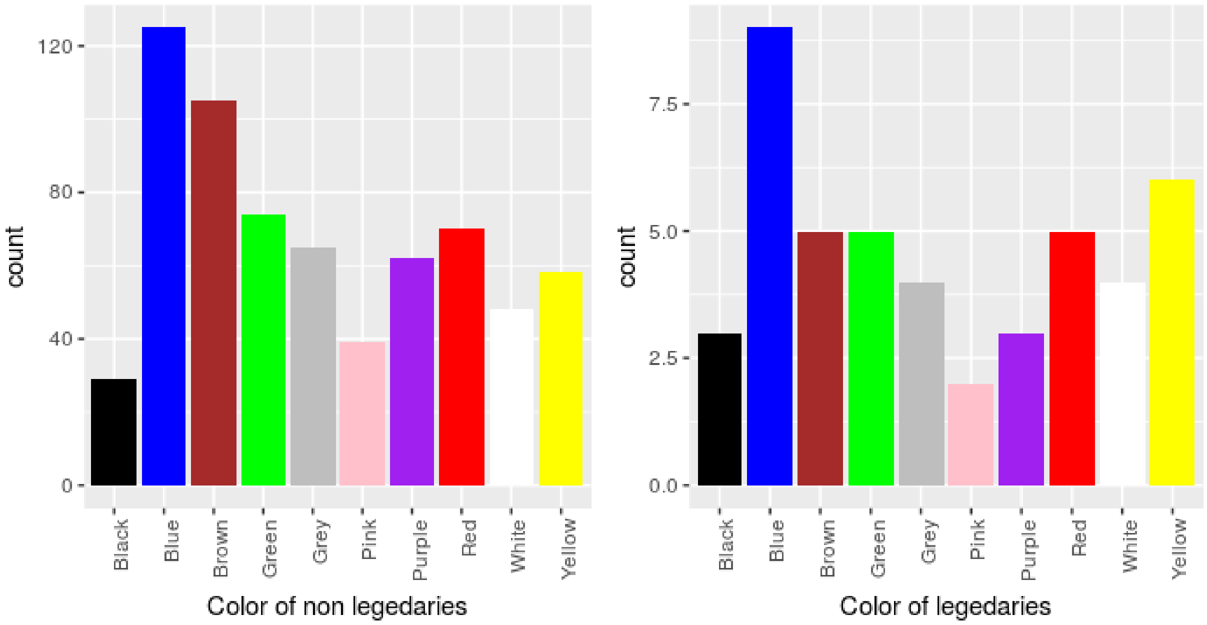


Figure 7: Histograms of the variable **Color** for non-legendary (left) and legendary (right) Pokémon.

4 Unsupervised classification

Once we have gained an insight into the variables of the dataset and their relations, we will now try to cluster the Pokémon. To this end we will use all the numerical variables, except **Pr_Male**, because it is not defined for all the instances (and it has shown to be a not very representative variable), and **Generation**, because it seems to be not very representative either. We will cluster according with the second Ward's method according to the literature [7]. This is a hierarchical clustering method where the dissimilarities are squared before cluster updating. Thus the method works with dissimilarities or distances between the instances. Before computing the distances between the instances, we first standardized them (apply a transformation which results are variables with mean 0 and standard deviation 1), in order all of them to be equally representative. Then we computed the Euclidean distances between the instances.

At this point, we applied the Ward's clustering method. Its corresponding dendrogram is shown in Figure 8. It is seems that the number of clusters we should select is 2 or 3. We chose 3 to allow the clusters to model more atributes of the Pokémon.

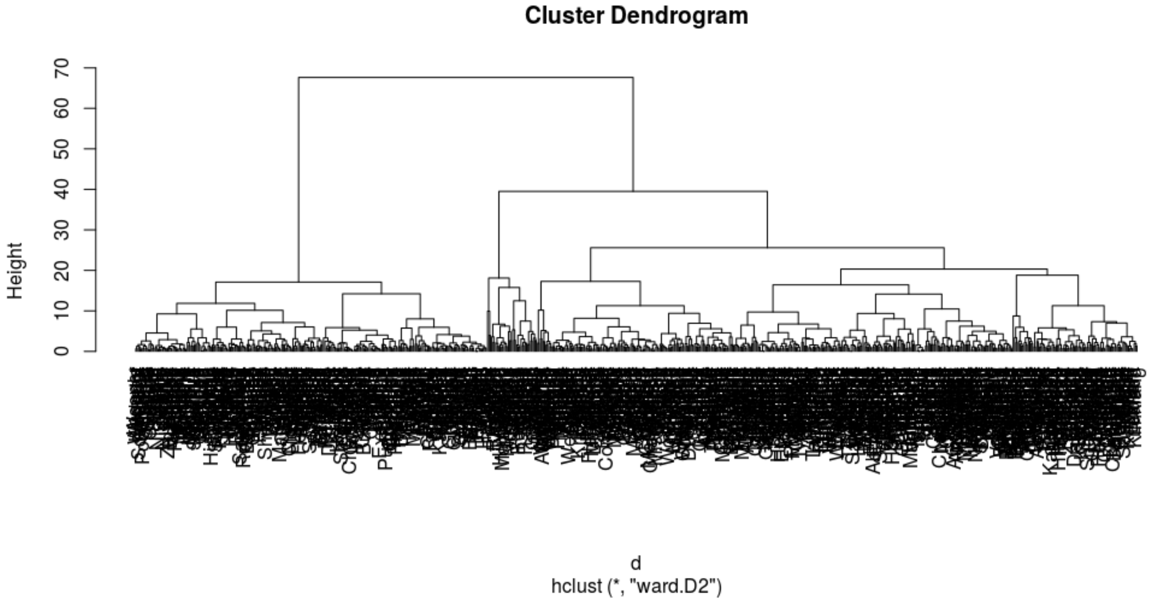


Figure 8: Dendrogram that represents the clusters are updated during the Ward's method.

Now, we will try to visualize the made clusters. They were made applying the Ward's method to a dissimilarity matrix obtained from 10 variables. The most representative plots are usually those of 2 dimension, so we would need to plot many of them to show the cluster in their totality. Therefore, we will only plot the 3 clusters are distributed in the subspaces in the projections in the pairs of variables **Total** and **Catch_Rate**, **Weight_kg** and **Height_m**, **Weight_kg** and **Catch_Rate**, and **Sp_Atk** and **HP**. These are shown in 9. According to them, the red cluster is composed of the weakest Pokémon, those with the lowest combat stats, and easy to catch. Then we have the green and blue clusters, which both include strong Pokémon. The major apparent difference between this two clusters is that the Pokémon in the blue are bigger (i.e. heavier and higher), whereas the ones in the green are more little, more or less similar to the red ones in this sense.

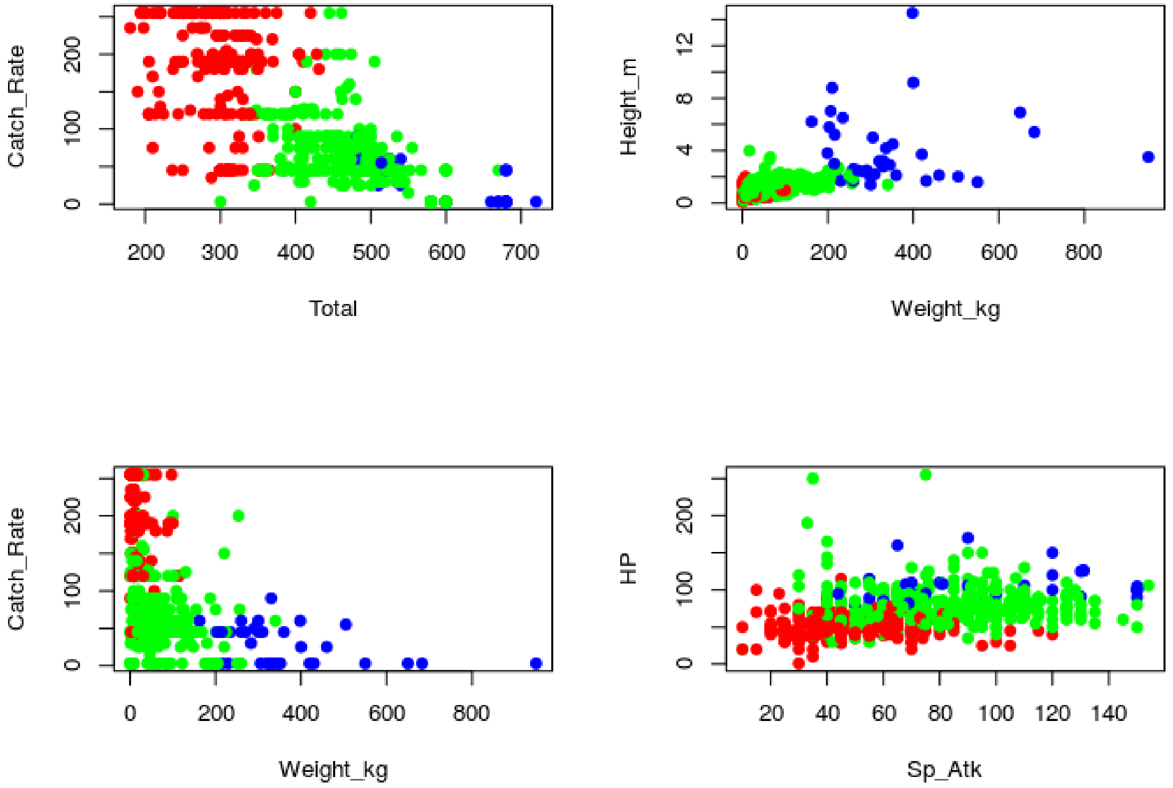


Figure 9: Three clusters visualized in function of four pairs of numerical variables.

4.1 Principal Component Analysis

We have been able to determine the properties of each cluster based on Figure 9, but we have used four plots for that. In this section we will try to find (two) new variables that explain more the real variability of the data, with the hope that those two new variables will be enough to understand the properties of the previously computed clusters. We will use a Principal Component Analysis (PCA) for that [8]. The PCA lets us find a orthogonal transformation that, applied to our current (standardized) variables, returns new variables in such a way that the first new variable has the largest variance, the second variable the second largest variance, and so on.

We performed the PCA with the above listed 10 variables. The variances explained with each of the new variables are shown in Figure 10. With the two variables we will select the proportion of the explained is 0.62. This value is not too high, but we have to take into account that we are reducing 8 eight dimensions, what of course causes an inevitable loss of information. It is also worth to say that the last principal component is null, because **Total** is a linear combination (the sum) of the other combat stats.

Now that we have computed two new variables, it is time to check if the clusters are visible when plotting those variables, and if the variables let us understand the clusters. In Figure 11 we can find a similar plot of the all the instances in function of the two new variables, where the cluster is also indicated, with the dark blue (group 1) for the previous red cluster, blue (group 2) for the previous green cluster, and light blue (group 3) for the

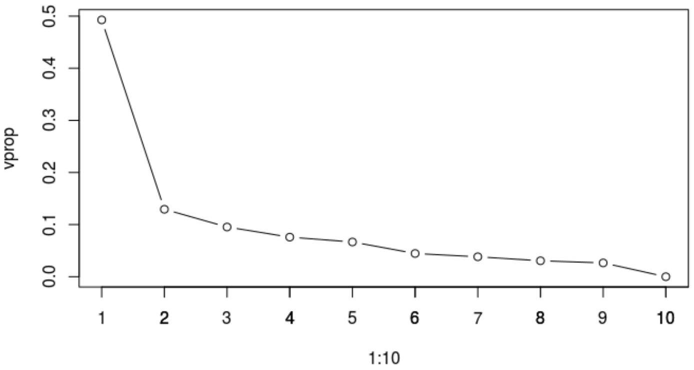


Figure 10: Proportion of the total variance of the data (with the mentioned 10 numerical standardized variables) explained with each of the new variables.

previous blue group. This plot is very representative. We have 2 important components dividing the Pokémon, which explain more or less same about the three clusters than what we had deduced previously. The former is that of the **Catch_Rate** and combat stats like **Speed** or **Sp_Atk**, which more or less corresponds with the first principal component. Thus, the bigger the first principal component is, the stronger the Pokémon is going to be fighting, and the harder it is going to be to catch. The latter indicates if the Pokémon is small or big, the bigger the Pokémon is, the smaller the second principal component will be. It is also related with the **Defense** of the Pokémon, probably because, as we can see in Figure 5, the **Defense** of a Pokémon is normally bigger the heavier it is.

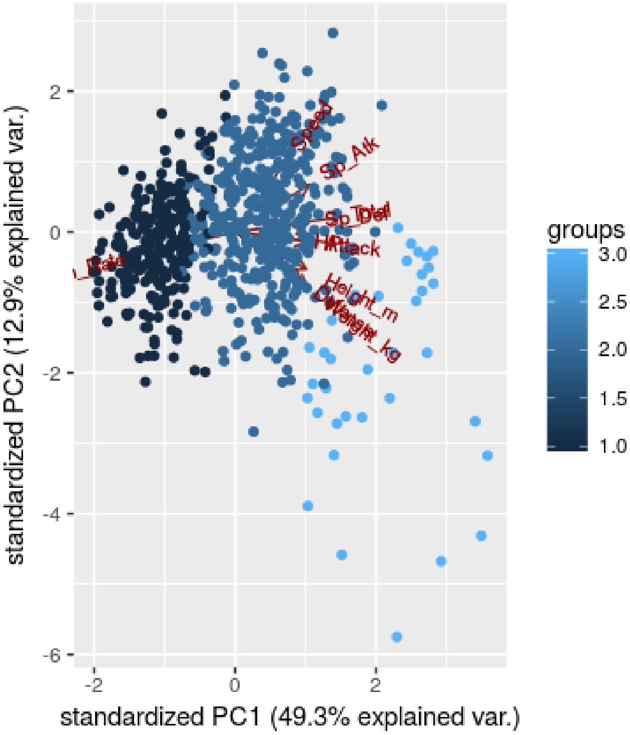


Figure 11: Three clusters visualized in function of four pairs of numerical variables.

5 Predictive modeling

In this last section we will explain two types of predictive models that have been developed with this database. First, we will try to predict the **Catch_Rate** with some linear models. Then we will try to discriminate between legendary and non-legendary Pokémon, via Linear Discriminant Analysis (LDA).

5.1 Linear models

Figure 5 shows that the **Catch_Rate** of a Pokémon is correlated with almost all the other numerical variables, so it seems a good candidate to be predicted. We will first try to predict it only with the **Total**, so the model can be mathematically written like in Equation 2.

$$\textbf{Catch\_Rate} = \beta_0 + \beta_1 \textbf{Total} \tag{2}$$

Then we will expand the model, and substitute **Total** with the rest of combat stats. We will exclude **Total** because it is already a linear combination of other variables. Then the second model can be defined like it is shown in Equation 3.

$$\textbf{Catch\_Rate} = \beta_0 + \beta_1 \textbf{HP} + \beta_2 \textbf{Attack} + \beta_3 \textbf{Defense} + \beta_4 \textbf{Sp\_Atk} + \beta_5 \textbf{Sp\_Def} + \beta_6 \textbf{Speed} \tag{3}$$

Finally, we will include the remaining representative numerical variables, i.e. **Weight_kg** and **Height_m**. This last model is mathematically represented in Equation 4.

$$\textbf{Catch\_Rate} = \beta_0 + \beta_1 \textbf{HP} + \beta_2 \textbf{Attack} + \beta_3 \textbf{Defense} + \beta_4 \textbf{Sp\_Atk} + \beta_5 \textbf{Sp\_Def} + \beta_6 \textbf{Speed} + \beta_7 \textbf{Weight\_kg} + \beta_8 \textbf{Height\_m} \tag{4}$$

We will evaluate the models with the coefficient of determination (R^2), which indicates the proportion of the variance in the dependent variable that is predictable from the predictors. So we fitted the models (the fitted models are shown in the same order than above in Equation 5), and computed the mentioned R^2, which are shown in Table 1. As we can see, results do not vary much between the models, but it is true that every time we extend or make more complex the model, the bigger R^2 is, and hence the better the models predicts the **Catch_Rate**. Anyway, the results are more or less the same. The reason for that is that it seems that **Catch_Rate** is more or less equally correlated with all the combat stats and therefore, the coefficients of the linear model are very similar in LM2, and also very similar to the coefficient of **Total** in LM1. Then, the improvement is slightly bigger when we include **Weight_kg** and **Height_m**, but do not make any major change to the previous model.

$$\begin{aligned}
\text{LM1: } \textbf{Catch\_Rate} =& 315.70 - 0.5155\textbf{Total}\\
\text{LM2: } \textbf{Catch\_Rate} =& 315.53 - 0.5290\textbf{HP} - 0.5355\textbf{Attack} - 0.4851\textbf{Defense} - 0.5355\textbf{Sp\_Atk} - 0.4867\textbf{Sp\_Def} - 0.5182\textbf{Speed}\\
\text{LM3: } \textbf{Catch\_Rate} =& 319.95 - 0.5524\textbf{HP} - 0.5599\textbf{Attack} - 0.5169\textbf{Defense} - 0.5408\textbf{Sp\_Atk} - 0.4895\textbf{Sp\_Def} - 0.5088\textbf{Speed}\\
&+ 0.0438\textbf{Weight\_kg} - 1.1253\textbf{Height\_m}
\end{aligned} \tag{5}$$

	LM1	LM2	LM3
R^2	0.545057	0.545369	0.546633

Table 1: Value of the coefficient R^2 for the built linear models. LM1 refers to the one shown in Equation 2, LM2 to the one in 3, and LM3 to the one in 4.

Additionally, we carried out a study of the residuals to validate the models. Since all the models look more or less like the same, we just made it once (with the LM3 model). The residuals are the difference between the real value and the predicted value. These can be plotted in several ways. We will, on the one hand, plot the real value versus the predicted value of the **Catch_Rate** variable. If the model were perfect, the result would be a line of slope 1. On the other hand, we can just plot the residuals of all the Pokémon in function of the real value, which would lead to an horizontal line of value 0. These two plots are shown in Figure 12. As we can see, the residuals are quite big, and they tend to be positive for small values of **Catch_Rate**, and negative for its big values. That indicates that maybe a linear model is not the best option to predict this variable, but it somehow works.

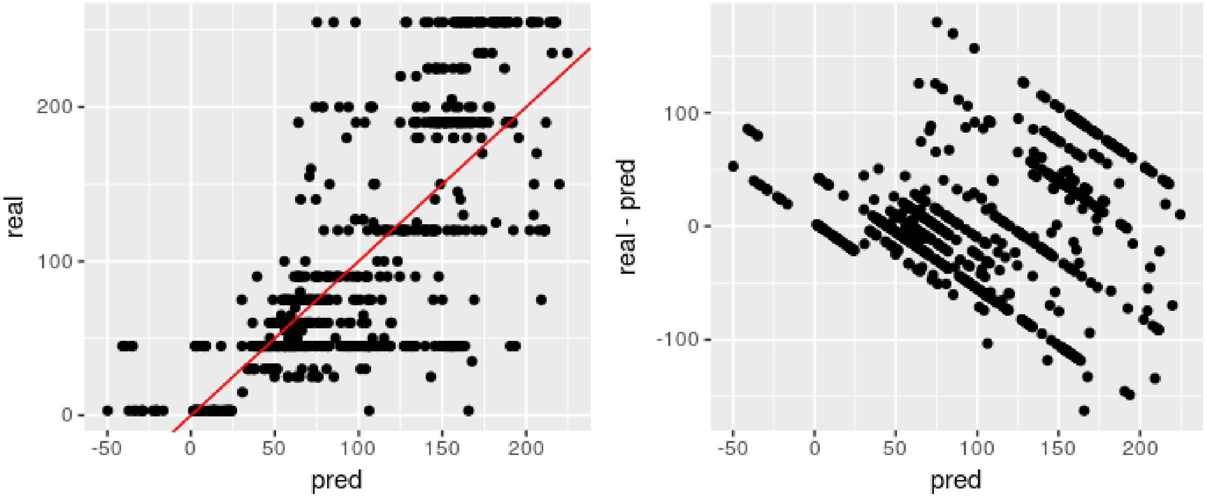


Figure 12: Visualization of the residuals of the ML3 model. In the left plot the red line represents the result of a perfect linear fitting.

5.1.1 General linear model

Since the dataset we are working with includes many categorical variables, a way to try to improve the previous linear models is to extend them to general linear models. The way a general linear model includes information from categorical variables is assigning a constant to each value of each categorical variable. Then, this value will be added to the prediction in case the categorical variable takes that value. Thus, we extended the LM3 with the categorical variables **isLegendary**, **Body_Style**, **Type_1**, and **Egg_Group_1**. After fitting the general linear model, the value of the coefficient R^2 was increased until the value of 0.62103, what is a sign that in this case including categorical variables in the model has helped. We can also analyze the residuals in this case, as shown in Figure 13. According to it, the fact that including categorical variables in the model has not solved the non-uniformity of the distributions of the residuals.

5.2 Linear Discriminant Analysis

Finally, we will try to separate via LDA the legendary Pokémon from the non-legendaries. The LDA tries to find a linear combination of the variables to separate as maximum as possible their classes [9]. Since we want to make this classification process more or less clear, we will try to predict the class with only two variables, because hence we are going to be able to see the result of separation of the space of the predictors.

The two best variables that we probably have to carry out the LDA are the already computed two first components of the PCA. Thus, we computed the LDA with those variables. The consequent separation of the space of the predictors is shown in Figure 14. Even though the LDA is a simple classifier and it is only taking as inputs two variables, the results are pretty good. The confusion matrix is shown in Table 2.

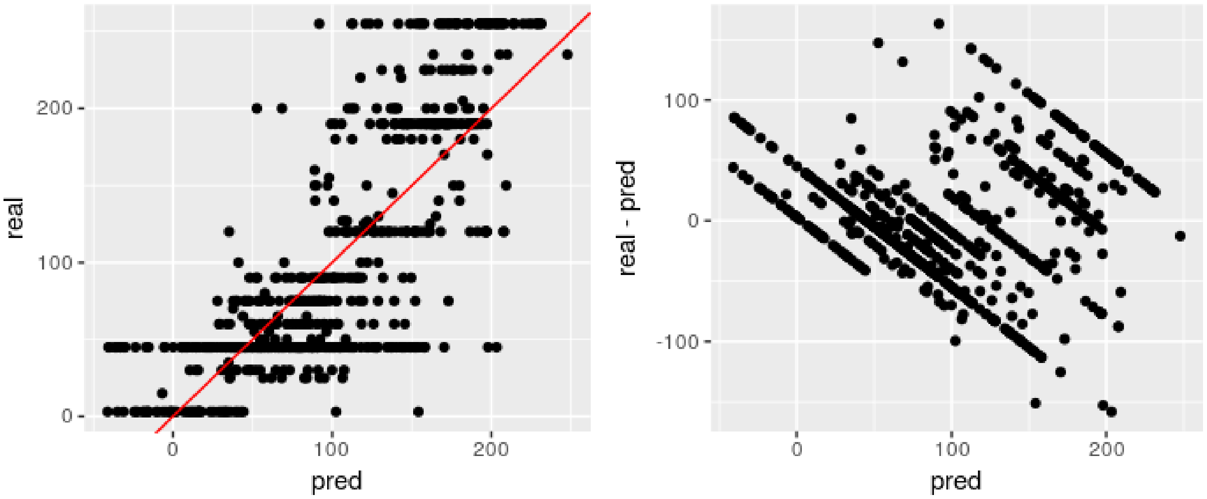


Figure 13: Visualization of the residuals of the general lineal model.

	Real True	Real False
Predicted as True	42	25
Predicted as False	4	650

Table 2: Resulting confusion matrix from the LDA classification.

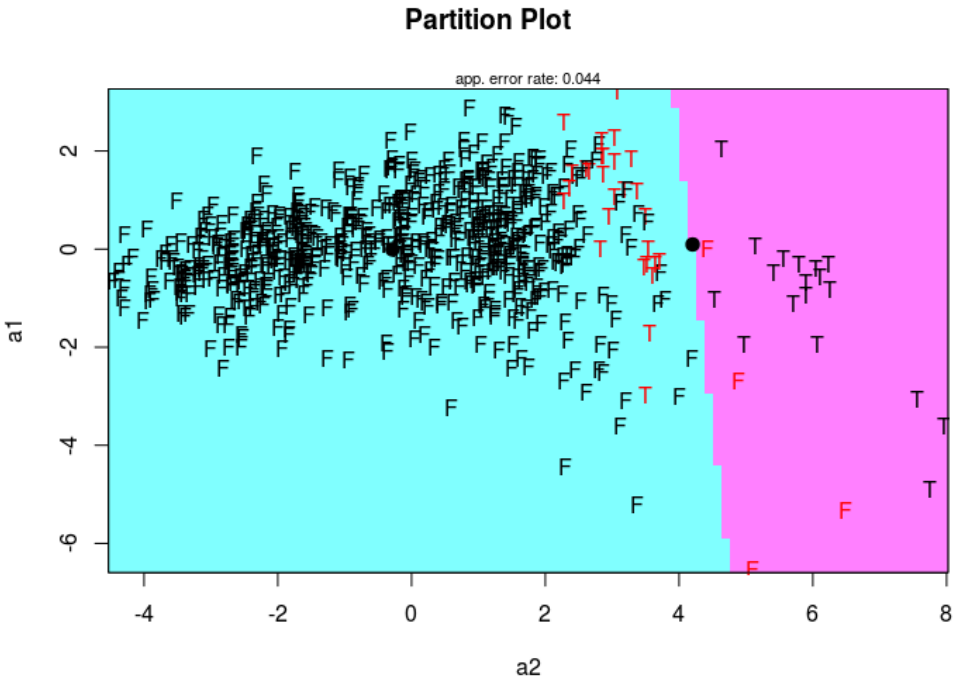


Figure 14: Visualization of the LDA.

6 Conclusions

In this work a statistical analysis of the Pokémon as they appear in the RGBs have been carried out. We have carried out univariate analysis for all the variables in the dataset that we have previously built. We have analyzed how the numerical variables are distributed, and all of them have shown to have a gaussian behavior. In the case of the categorical variables, we have used histograms with the same end. Then we have tried to find correlations between the numerical variables. The higher the stats of the Pokémon, it will also be bigger, heavier and harder to catch. No big dependencies have been found with the generation a Pokémon was released or its probability of being female or male. On the other hand a similar analysis of dependencies has been carried out for the categorical variables. First, we have made an idea of which variables depend on the other with a chi square independence test, and then we have quantified the dependencies with the measure of the Normalized Mutual Information. The color of the Pokémon, its first egg group and the fact that it is legendary or not are the variables that share the most information.

Then we have clustered the Pokémon with the Ward's method. Making 3 clusters has been a reasonable choice. With the help of a Principal Component Analysis we have gained an insight into the properties of this clusters. In general, the first cluster is formed of the weakest and smallest Pokémon. The second one includes strong and hard to catch Pokémon. The remaining one is the smallest, yet it includes the heaviest, highest and strongest Pokémon.

Finally, we have built two predictive models. First we have tried to predict the catch rate of a Pokémon based on its numerical variables. Predicting with the independent combat stats or directly with their sum has not lead to significant differences, neither including the information of the weight and height of the Pokémon. Conversely, when generalizing those linear regression models to a general lineal model that includes information about many categorical variables, the fitting has been more precise. In any case, the catch rate variable does not seem to be linear in respect with the numerical values, so probably a better regression model can be developed if it is able to model more complex dependencies. On the other hand, we have carried out a Linear Discriminant Analysis to predict whether a Pokémon is legendary or not, which has worked quite well, probably because its inputs have been the most meaningful variables that were possible, the two first principal components of the PCA.

References

[1] Alberto Barradas. Pokemon with stats. https://www.kaggle.com/abcsds/pokemon. Last Accessed: 13/01/2017.

[2] Asier LÃşpez Zorrilla. PokÃl'mon for Data Mining and Machine Learning. https://www.kaggle.com/alopez247/pokemon. Last Accessed: 13/01/2017.

[3] David W Scott. Multivariate density estimation: theory, practice, and visualization, 2015.

[4] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[5] Royal Society (Great Britain). *Proceedings of the Royal Society of London*. Number v. 58. Taylor & Francis, 1895.

[6] F. Yates. Contingency tables involving small numbers and the ÏĞ<sup>2</sup> test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.

[7] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295, 2014.

[8] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

[9] R. A. FISHER. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.