Assignment1 Report

1. **Introduction**
   Pokémon is a popular video-game made by Nintendo, Game freak, and Creatures, originally released in 1996. Pokémon started as a Role-Playing Game (RPG), but due to its increasing popularity, its owners ended up producing many TV series, manga comics, and so on, as well as other types of video-games (like the famous Pokémon Go!).

2. **Dataset**
   The dataset includes 21 variables per each of the 721 Pokémon in the first six generations.

- Number. Pokémon ID in the Pokédex.
- Name. Name of the Pokémon.
- Type_1. Primary type.
- Type_2. Second type, in case the Pokémon has it.
- Total. Sum of all the base stats (Health Points, Attack, Defense, Special Attack, Special Defense, and Speed).
- HP. Base Health Points.
- Attack. Base Attack.
- Defense. Base Defense.
- Sp_Atk. Base Special Attack.
- Sp_Def. Base Special Defense.
- Speed. Base Speed.
- Generation. Number of the generation when the Pokémon was introduced.
- isLegendary. Boolean that indicates whether the Pokémon is Legendary or not.
- Color. Color of the Pokémon according to the Pokédex.
- hasGender. Boolean that indicates if the Pokémon can be classified as female or male.
- 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.
- Egg_Group_1. Egg Group of the Pokémon.
- Egg_Group_2. Second Egg Group of the Pokémon, in case it has two.
- hasMegaEvolution. Boolean that indicates whether the Pokémon is able to Mega-evolve or not.
- Height_m. Height of the Pokémon, in meters.
- Weight_kg. Weight of the Pokémon, in kilograms.
- Catch_Rate. Catch Rate.
- Body_Style. Body Style of the Pokémon according to the Pokédex

The dataset is from https://www.kaggle.com/alopez247/pokemon, and was based on the Kaggle database "721 Pokemon with stats" by Alberto Barradas (https://www.kaggle.com/abcsds/pokemon).

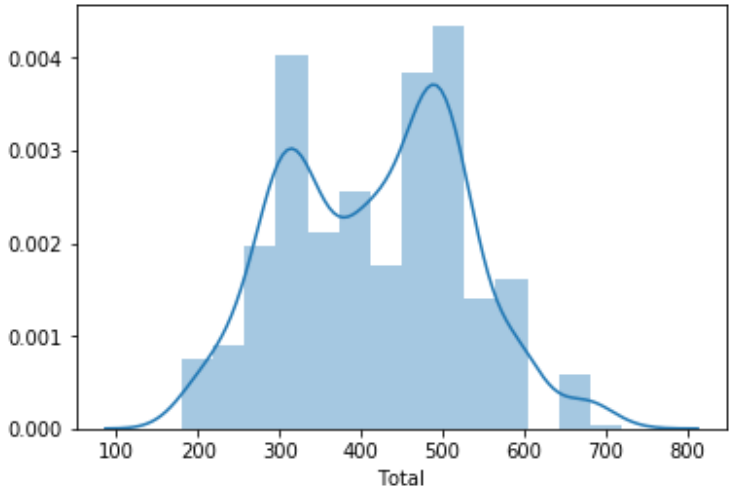
3. **General descriptive analysis of the variables**

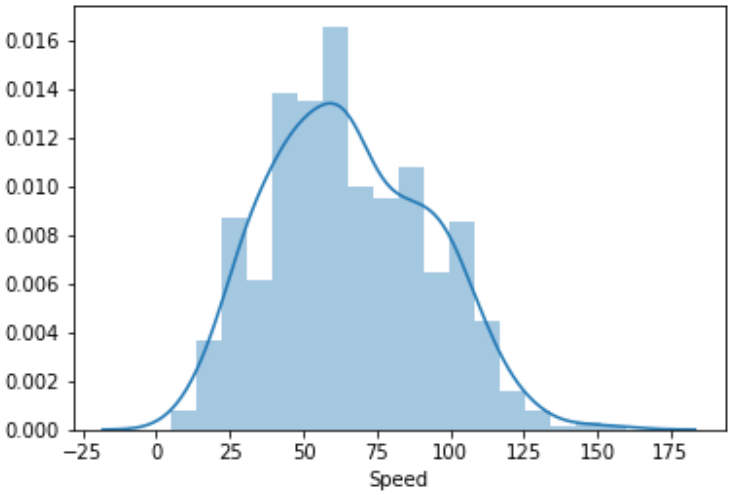
```
In [5]: print(df.isnull().sum())
```

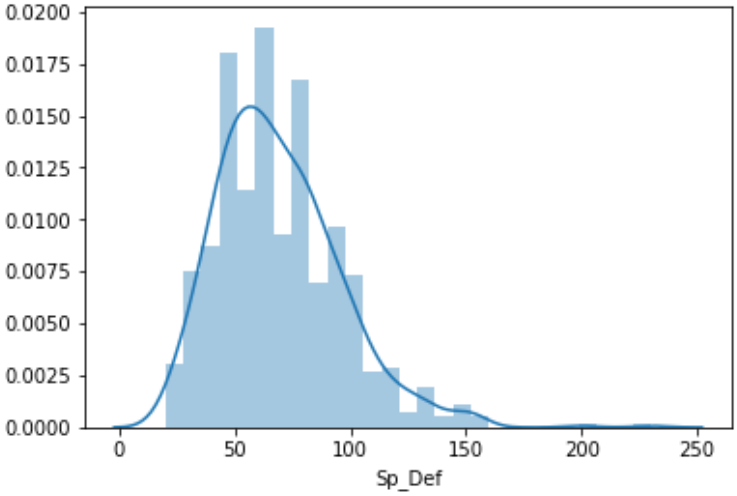
```
Number                0
Name                  0
Type_1                0
Type_2              371
Total                 0
HP                    0
Attack                0
Defense               0
Sp_Atk                0
Sp_Def                0
Speed                 0
Generation            0
isLegendary           0
Color                 0
hasGender             0
Pr_Male              77
Egg_Group_1           0
Egg_Group_2         530
hasMegaEvolution      0
```

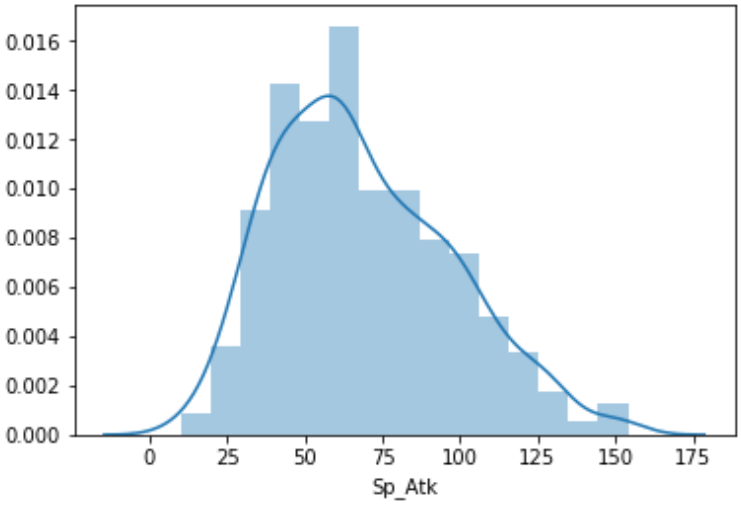
As we checked, that there are 530 rows at Egg_Group_2 is N/A and 371 rows at Type_2 is N/A. Because the two columns contain too muck N/A, so we decided to drop the two columns. And there is one column contains 77 rows of N/A, so we try to fill those N/A rows with the mean value of the column.

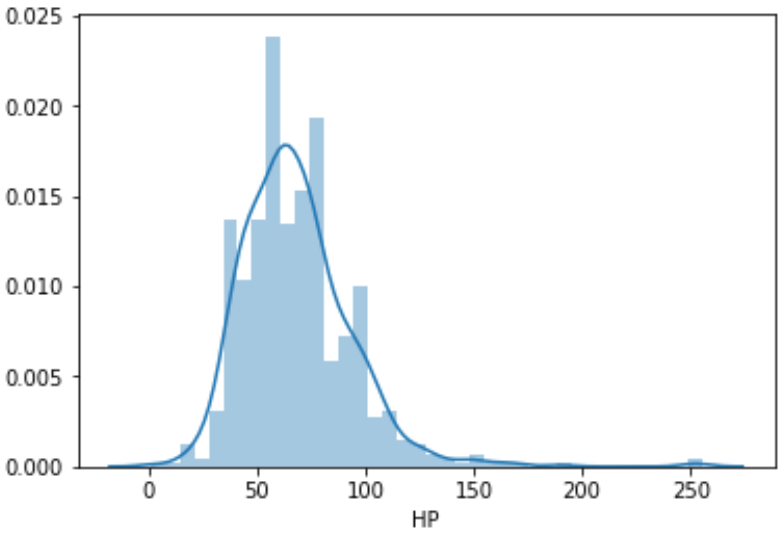
Also, we did some density histograms for some columns. From the histograms in the ipynb file we can see that except for column 'Total', most columns appeared to to have a gaussian behavior. Besides hasMegaEvolution, Body_Style, Name, Type others are all numerical and continuous data, so, we will try to understand them with the assistance of Density Plots. All these plots are shown in below. 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.

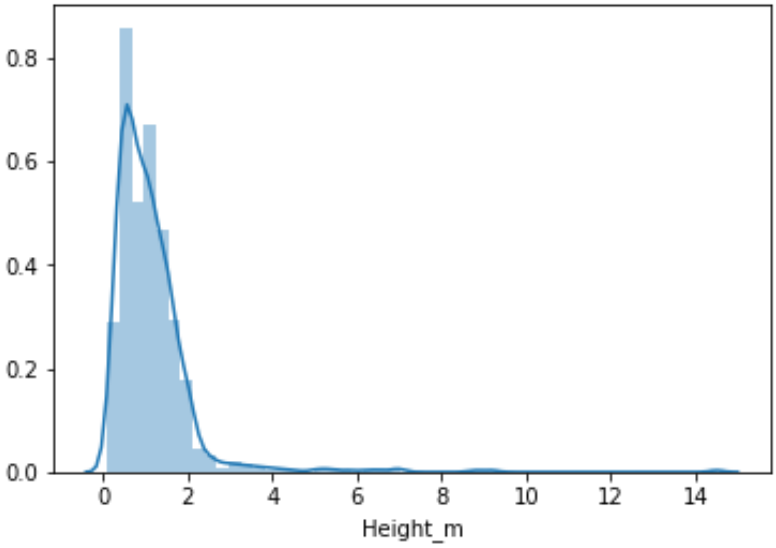


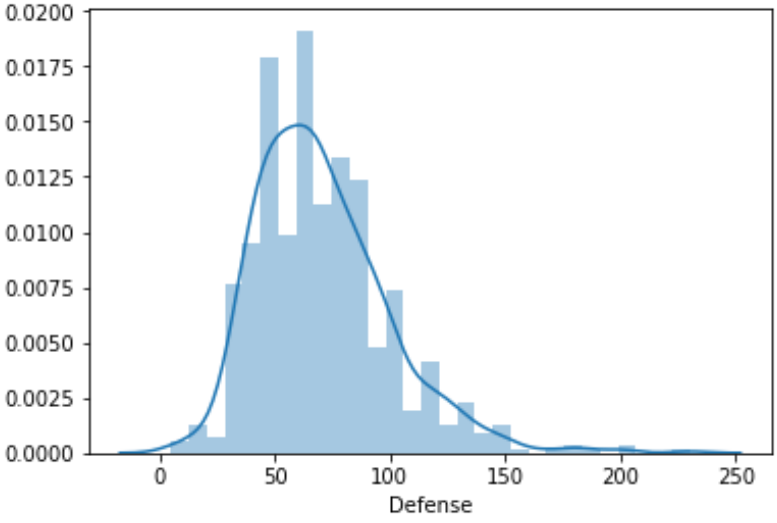

0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−25
0
25
50
75
100
125
150
175
Speed

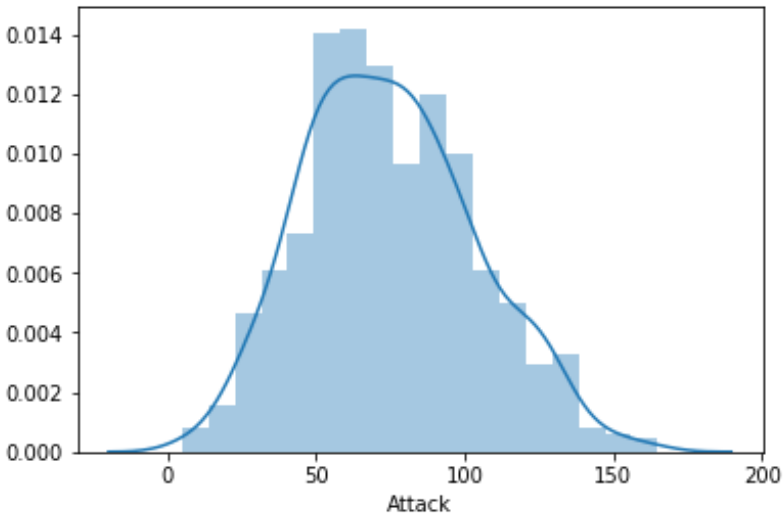

0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
50
100
150
200
250
Sp_Def


0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
25
50
75
100
125
150
175
Sp_Atk


0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
HP


0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
Height_m

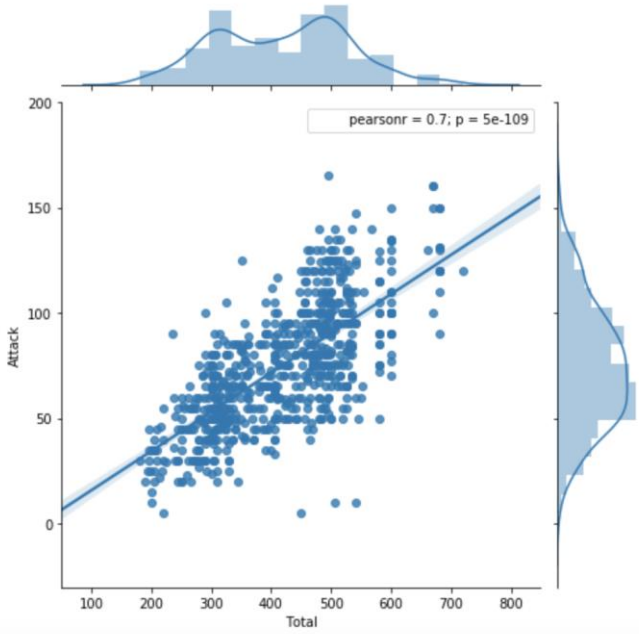

0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
50
100
150
200
250
Defense



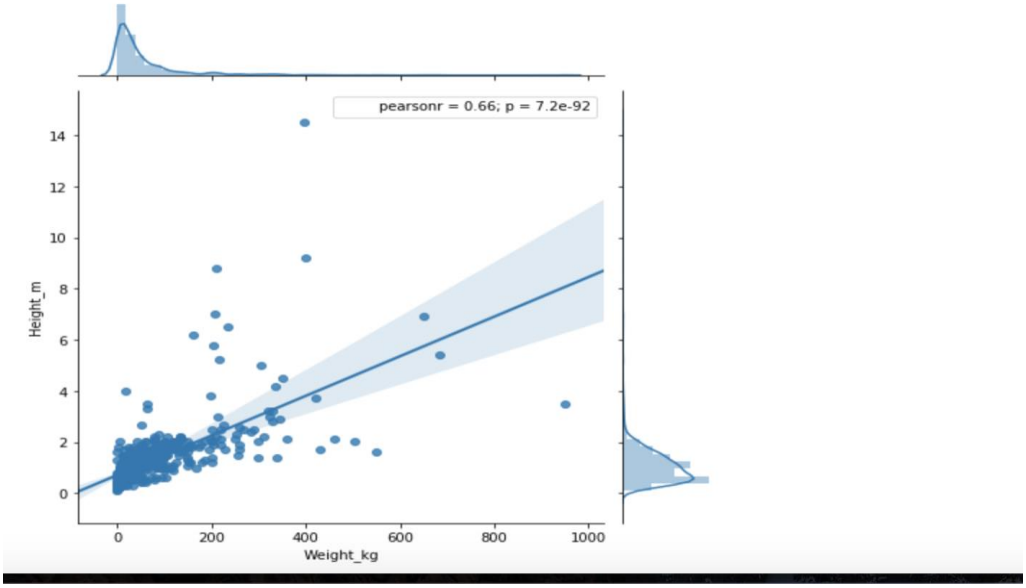
According to the correlation chart and the heat map, we can see that total have strong correlation with Sp_Atk, Sp_Def and Attack. And Height_m has strong correlation with Weight_kg.

Also we draw two plot to show the correlation between Attack and Total, and Height_m with Weight_kg.

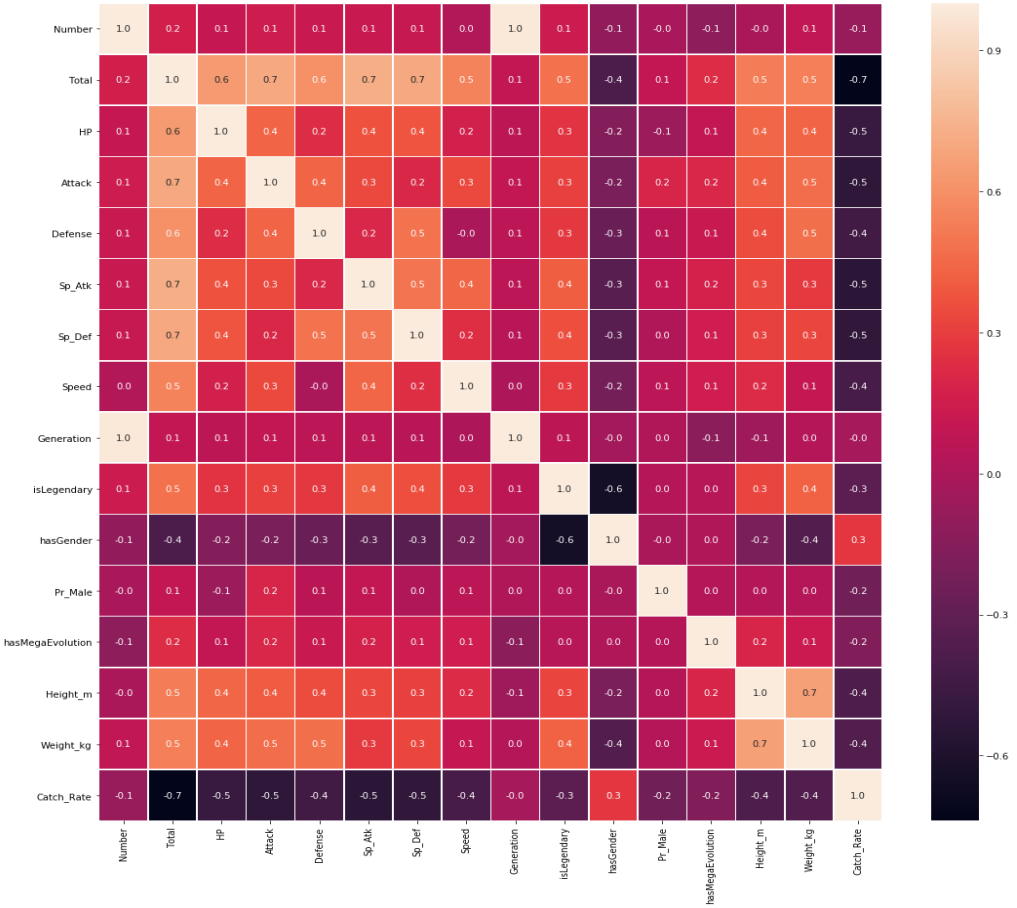
```
sns.jointplot(x="Total", y="Attack", data=clean_df, kind = 'reg', size = 7)
plt.show()
```



```
sns.jointplot(x="Weight_kg", y="Height_m", data=clean_df, kind = 'reg', size = 7)
plt.show()
```



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



Accuracy chart：

acc	LogisticRegression	RandomForest
Raw model fit	0.7181080023478772	0.9718254744668361
After cross validate	0.7152603074772886	0.7268343815513627

After tranformers	0.855966107617051	0.7518169112508736
After Fine-tuning	0.984	0.9583333333333334

4. **Result**

   The best score using random forest is 0.9583333333333334, and the best score using logistic regression is 0.984.

   In this work, a statistical analysis of the Pokémon as they appear in the RGBs (Role_playing games) 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. For most Pokémon, the higher they are the more weight they will have. And the higher they are at Total, the higher Attcak they would have.

   finally, we try to predict whether a Pokémon is legendary or not using two models. One is RandomForest and the other is Logistic Regression. Both of the models have worked quite well, probably because its inputs have been the most meaningful variables that were possible

5.