CSC 478 Final Project

Pokémon Stat Analysis

Yiyang Yang

Overview:

Pokémon is still very popular around the world from the start it comes out since 1996. It is still very attractive to me. Also I like playing games and want to apply data analysis to game area, and find some interesting idea. So I decide to do some analysis with these Pokémon, according to their stat. I find this dataset from kaggle.com, there are 13 variables and 721 observations. The variables are the names and some basic stat of the Pokémon.

Data Analysis:

Out[3]:

	Name	Type 1	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	Bulbasaur	Grass	318	45	49	49	65	65	45	1	False
1	Ivysaur	Grass	405	60	62	63	80	80	60	1	False
2	Venusaur	Grass	525	80	82	83	100	100	80	1	False
3	VenusaurMega Venusaur	Grass	625	80	100	123	122	120	80	1	False
4	Charmander	Fire	309	39	52	43	60	50	65	1	False

This is how the data looks like, then I get the description of all the data by using describe().

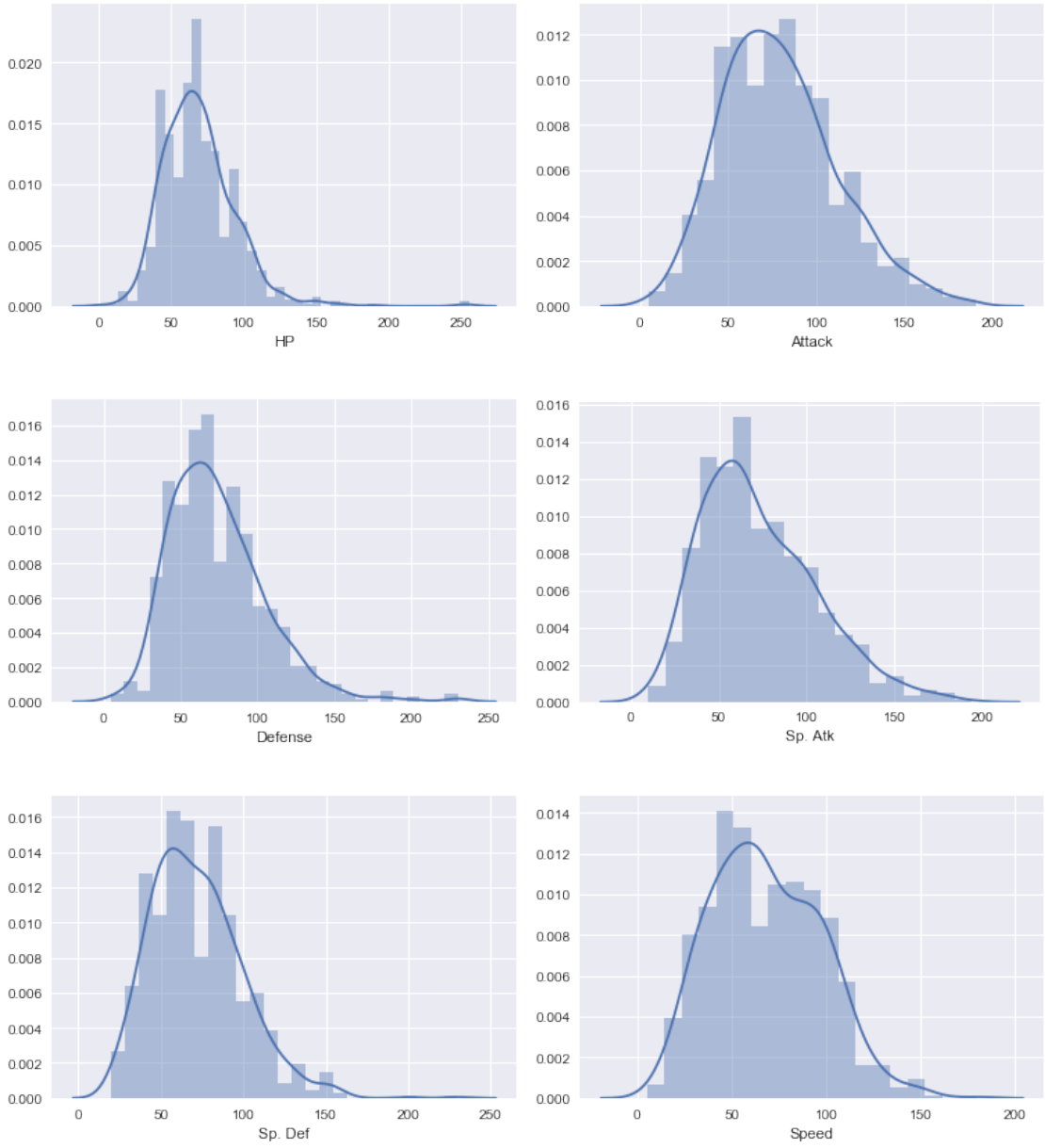
Out[4]:

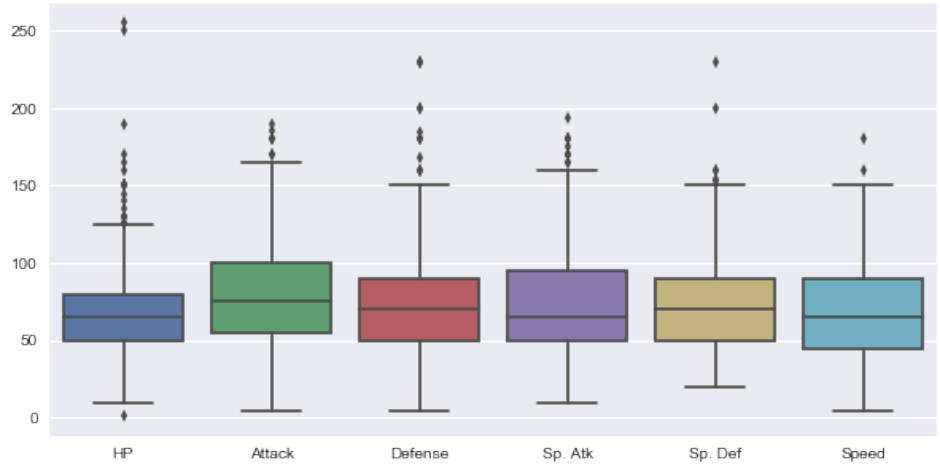
	Name	Type 1	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
count	800	800	800.00000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.00000	800
unique	800	18	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2
top	Azurill	Water	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False
freq	1	112	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	735
mean	NaN	NaN	435.10250	69.258750	79.001250	73.842500	72.820000	71.902500	68.277500	3.32375	NaN
std	NaN	NaN	119.96304	25.534669	32.457366	31.183501	32.722294	27.828916	29.060474	1.66129	NaN
min	NaN	NaN	180.00000	1.000000	5.000000	5.000000	10.000000	20.000000	5.000000	1.00000	NaN
25%	NaN	NaN	330.00000	50.000000	55.000000	50.000000	49.750000	50.000000	45.000000	2.00000	NaN
50%	NaN	NaN	450.00000	65.000000	75.000000	70.000000	65.000000	70.000000	65.000000	3.00000	NaN
75%	NaN	NaN	515.00000	80.000000	100.000000	90.000000	95.000000	90.000000	90.000000	5.00000	NaN
max	NaN	NaN	780.00000	255.000000	190.000000	230.000000	194.000000	230.000000	180.000000	6.00000	NaN

From this form, I find an interesting thing, the water type is the most number of type in Pokémon, in my view, this might be a reason, since Pokémon is from Japan, Japan is an island

country, sea and water is the most important thing to them, and they want to show the love of the sea and water, then they create more water type Pokémon.

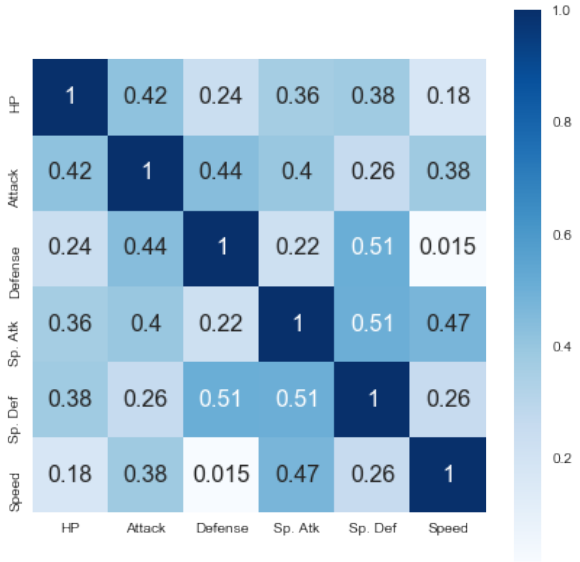
Since I want to do some analysis from the stat of these Pokémon, I make some distribute graphs of each stat.





From the graphs, I find the distribution of these Pokémon's stats are similar, the highest points of these graph are almost near 50, also from the boxplot I find the same thing, to a game the value of each stat is balanced means this game is balanced, the playability could be guaranteed, the difficulty to get different monsters is increasing while the value of stats of the monsters are increasing. I think this should be the joy of this game, and that's why Pokémon could continue for a long time.

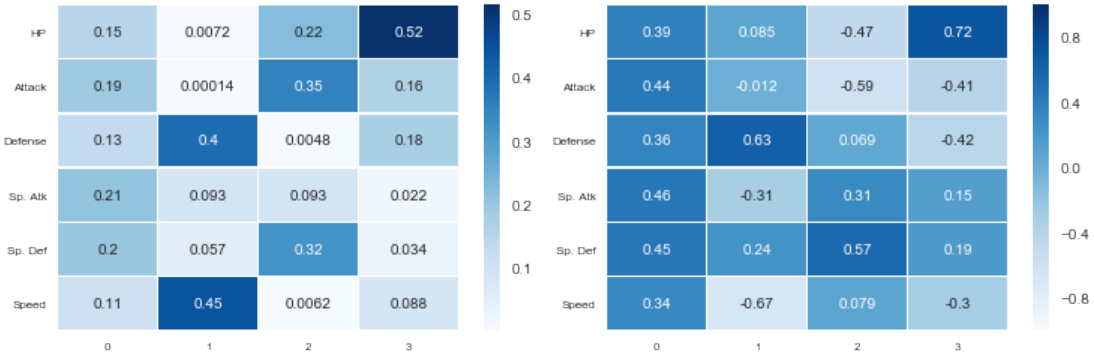
Then I make a heatmap of the correlation between each stat.



From the heatmap, I find there is not a very clear correlation between each stat, they don't influence each other very much. The only thing I find here is the correlation between Sp. Attack and Sp. Defense, Defense and Sp. Defense are all 0.51, I could know Sp. Attack and Sp. Defense influence each other in some special situation, and Defense may get some benefit from Sp. Defense, so these two has some kind of inner relationship with each other, another interesting point, Sp. Attack get more influence from Speed rather than Attack, maybe Sp. Attack is a special combination of Speed and Attack.

After doing these basic analysis, I want to know something more formt he dataset, so I begin to apply PCA on it, to find what influence a monster and what makes a monster.

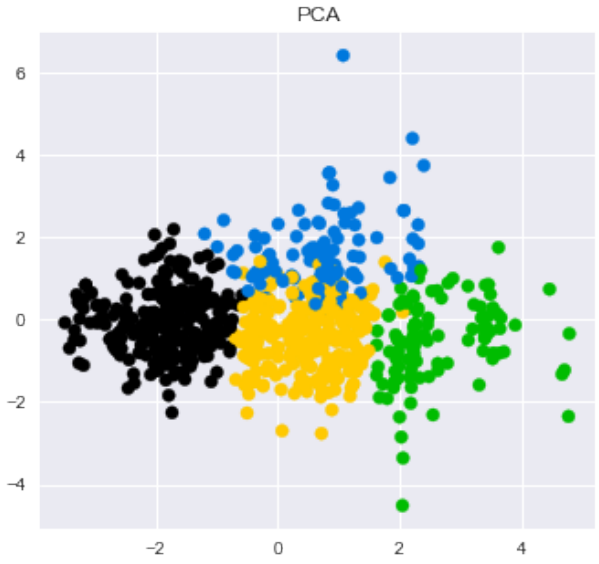
I get the following two heatmaps of loadings**2 and loadings from PCA, to explain 80% of the variance, there are 4 components coming out.



From left one I find Defense and Speed is two factors in a component, Attack and Sp. Defense is another two factors in another component, the last component, the main factor is HP, from the right one almost get the same points, but the I get the positive and negative relationship between these factors. From the second component, Defense is positive correlation, speed is negative correlation, I think it is clear in common sense, a higher defense will make a monster be slow; from the third component, Attack is negative correlation, Sp. Defense is positive correlation, it is an interesting relationship, if a monster has a higher Attack value, Sp. Defense

will reduce, this is very interesting, and I can imagine, if a monster wants to defense some special ability or attack, they need to arms itself, in a way, they will lose some Attack values, it may persuade me in some way; the last component is very obvious, Health Point is always the most important and necessary point in every kind of games.

The following graph is made from K-Means cluster of the PCA,



Out[50]:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
PC 1	-1.826394	0.725007	2.613954	0.495913
PC 2	-0.039993	1.612441	-0.525929	-0.356119
PC 3	0.012043	0.084828	0.023227	-0.050372
PC 4	-0.004456	-0.540211	-0.116667	0.242032

From this chart I can get the model between each component and cluster, and know how to divide different Pokémon into different clusters. I also get the completeness score and homogeneity score of this model, completeness score is 0.796749446859 and homogeneity score is 0.212886539655, which means this model works well but not perfect, since different Pokémon

have different kind of stats, the homogeneity score is not that high. But I think this model works well to cluster the Pokémon.

There is a very special kind of Pokémon in the game call “Legendary Pokémon”, they are hard to get but powerful, I want apply KNN, Decision Tree and Naïve Bayes to classify if the Pokémon is legendary. I divide the stat and legendary, also split them into train and test, then apply KNN and Decision Tree and Naïve Bayes to the dataset.

To KNN, I get the following data and scores on train test dataset.

	precision	recall	f1-score	support
False	0.97	0.97	0.97	148
True	0.67	0.67	0.67	12
avg / total	0.95	0.95	0.95	160

Train set score: 0.9984375, Test set score: 0.95.

To Decision Tree, I get following data and scores.

	precision	recall	f1-score	support
False	0.96	0.97	0.96	148
True	0.55	0.50	0.52	12
avg / total	0.93	0.93	0.93	160

Train set score: 0.996875, Test set score: 0.93125.

To Naïve Bayes, I get the scores

Train set score: 0.9328125, Test set score: 0.93125.

After comparing the score of the three method, I think KNN is the best choice to classify if the Pokémon is legendary.

```
knnpreds_test = knnclf.predict(pm_test_norm)
print(knnpreds_test)
```

By using KNN, I get the prediction as following.

```
[False False False False  True False False False  True False False False
  True False False False False False False False False False False False
 False False False False False False False False False False False False
 False False False False False False False False False False False False
 False False False  True False False False False False False False False
 False False False False False False False False False  True False False
 False False  True False False False False False False False False False
 False False False False False False False False False False False False
 False False  True False  True False False False False False False False
 False False False False False False False False False False False False
 False False False False False False False False False False False False
 False False  True False False False False False False False False False
 False False False False False False False False False  True  True
 False False False  True]
```

I will know which Pokémon is legendary by using the stat of it.

Conclusion:

From the analysis, I get some interesting conclusions:

1. From the description the overall dataset, I find the most type of Pokémon is water. Since Japan is an island country, they treat sea and water very respect.
2. From distribution plot and boxplot, I find the value of each stat is very balanced, this leads the game balanced, it attracts many fans and makes the game continue a long time.
3. From PCA analysis, I find four principle components, and find different factors in each components.
4. From the K-Means Cluster of PCA, I find 4 clusters and get the model between each PC and cluster.
5. Among the three classification methods, I find KNN is the best choice for the dataset to classify if the Pokémon is legendary one.