찾아라, 최강의 Pokemon!

Data Science Competition 2018

CON
TENTS

목차

- 주제 선정 배경
- 데이터 셋
- 다양한 방면에서의 데이터 분석
- 그 결과

포켓몬을 분석 대상으로 선정한 이유는?

1997년 일본 TV도쿄에서 처음 방영된 '포켓몬스터'는 1기를 시작으로,
현재 21기까지 방영하고 있는 애니메이션 계의 살아있는 전설이다.
애니메이션의 유명세에 힘입어 게임, 영화로까지 영역을 넓히고 있다.

대개 게임 캐릭터들은 여러 가지 성능을 지니고 있으며 각기 그 성능의
수치가 다르다. 성능 별로 명시되어 있으며 그 범위 또한 워낙 넓기 때문에
종합적으로 보았을 때 어떤 캐릭터가 가장 강한지 한 눈에 알기 어렵다.

포켓몬 게임에서도 마찬가지이다. 과연 이 수많은 포켓몬들 중에 가장 강한
포켓몬은 누구인지, 새로운 포켓몬이 등장했을 때 기존의 포켓몬들과 비교가
가능할 지 등에 대해 궁금증을 품게 되어 분석을 해보았다.

데이터 셋 'Pokemon with stats'

Kaggle에서 포켓몬 캐릭터 별 능력치를 총 정리한 'Pokemon with stats' 데이터 셋을 가져와 사용하였다. 총 721종의 포켓몬 데이터를 포함하고있다.

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False
5	5	Charmeleon	Fire	NaN	405	58	64	58	80	65	80	1	False
6	6	Charizard	Fire	Flying	534	78	84	78	109	85	100	1	False
7	6	CharizardMega Charizard X	Fire	Dragon	634	78	130	111	130	85	100	1	False
8	6	CharizardMega Charizard Y	Fire	Flying	634	78	104	78	159	115	100	1	False
9	7	Squirtle	Water	NaN	314	44	48	65	50	64	43	1	False
10	8	Wartortle	Water	NaN	405	59	63	80	65	80	58	1	False
11	9	Blastoise	Water	NaN	530	79	83	100	85	105	78	1	False
12	9	BlastoiseMega Blastoise	Water	NaN	630	79	103	120	135	115	78	1	False

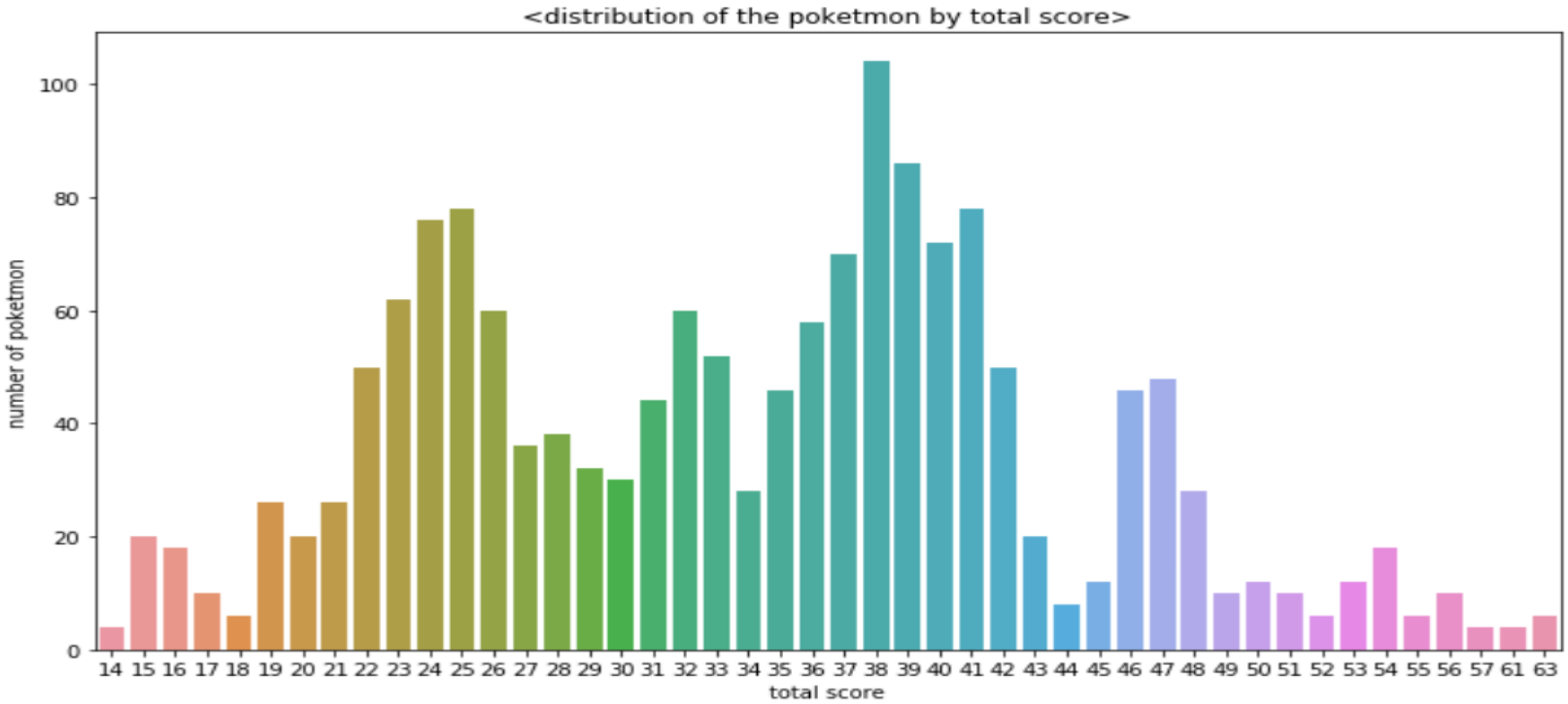
데이터 출처 : https://www.kaggle.com/abcsds/pokemon

데이터 항목에 대한 설명

1. # : 포켓몬 별 고유 ID
2. Name : 포켓몬 이름
3. Type 1 : 포켓몬의 특성으로 전투에서의 상성을 좌우한다.
4. Type 2 : 또 하나의 특성으로 일부 포켓몬에게만 주어져있다.
5. Total : 모든 능력치들의 단순 합
6. HP : 포켓몬의 체력을 나타내는 능력치로 0이 되면 기절한다.
7. Attack : 이 수치가 높을수록 물리 공격기를 사용하였을 때, 더 많은 데미지를 준다.
8. Defense : 이 수치가 높을수록 물리 공격기를 받았을 때, 데미지를 더 적게 받는다.
9. SP. Atk : 이 수치가 높을수록 특수 공격기를 사용하였을 때 더 많은 데미지를 준다.
10. SP. Def : 이 수치가 높을수록 특수 공격기를 받았을 때, 데미지를 더 적게 받는다.
11. Speed : 이 수치가 높을수록 상대보다 먼저 행동할 수 있다. 서로 우선도가 같은 기술을 사용한다면 더 높은 포켓몬이 선공을 하게 된다.
12. Generation : 1세대부터 6세대까지의 포켓몬이 등장한 세대이다.
13. Legendary : 레전드 포켓몬이면 True, 아니면 False를 나타낸다.

몇 점 이상이어야 강한 포켓몬이라 할 수 있을까?

기존 데이터 셋에 있는 'Total'은 6개 능력치의 단순 총합수치이다.
우리는 HP(17%), Attack(17%), Defense(17%), Sp.Atk(16%), Sp.Def(16%), Speed(17%)로 Total 점수를 100점 만점으로 환산하였다. 재산출한 Total은 0에서 100사이의 값으로 50점 이상이면 강한 포켓몬이라 설정하였다. Total score에 따른 포켓몬 분포를 보면, 20~49의 포켓몬의 비중이 가장 높고, 강하거나(50~100) 약한(0~19) 포켓몬은 적은 비중을 차지하였음을 알 수 있다.



Rank에 따라 Generation 분포는 어떻게 될까?

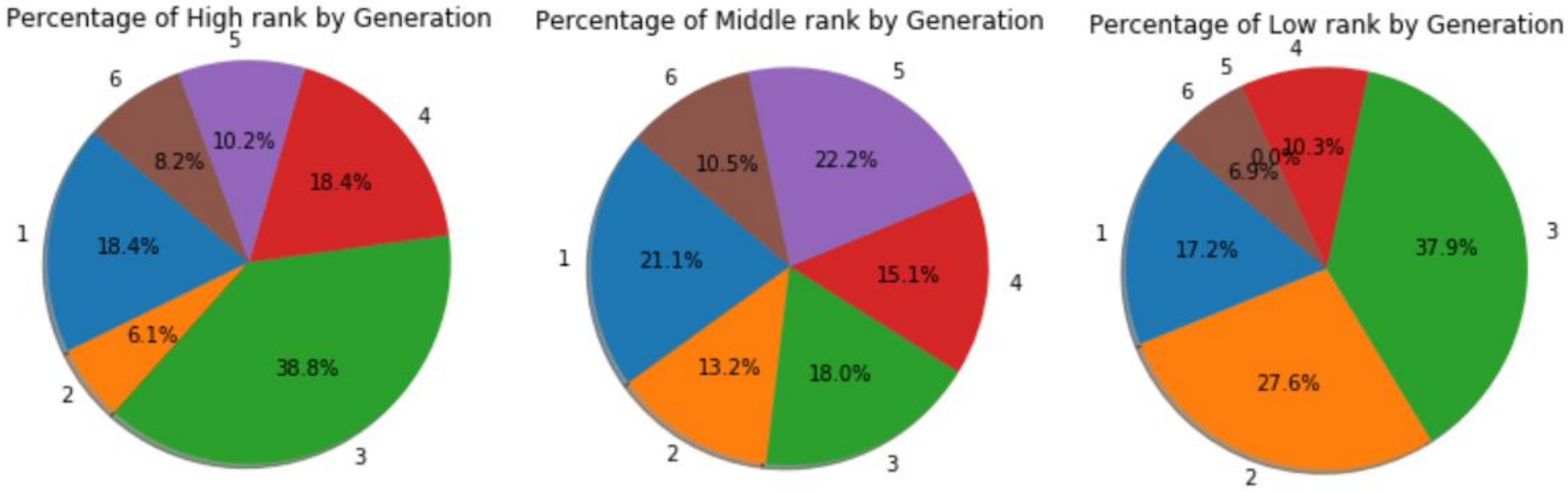
앞에서 구한 Total score에 따라 포켓몬에 등급을 두었다. 49점 이상이면 High,
20~48점이면 Middle, 19점 이하이면 Low Rank로 설정하였다.
진화론처럼 포켓몬도 세대를 거치며 진화하고 더 강해질까? 그 해답을 찾기 위해 1세대부터
6세대까지 존재하는 포켓몬들의 rank별 세대 분포를 분석해보기로 하였다.
먼저, 세대 별 포켓몬 수를 확인해보았다.

Generation	1	2	3	4	5	6
Number of pokemon	166 (20.75%)	106 (13.25%)	160 (20.0%)	121 (15.125%)	165 (20.625%)	82 (10.25%)

6세대 포켓몬 수가 가장 적으며, 그 다음 2세대, 4세대 순으로 적은 것을 확인할 수 있다.
여기서 특히 1, 3, 5세대 포켓몬 비율은 거의 비슷하다는 점을 기억해보자.

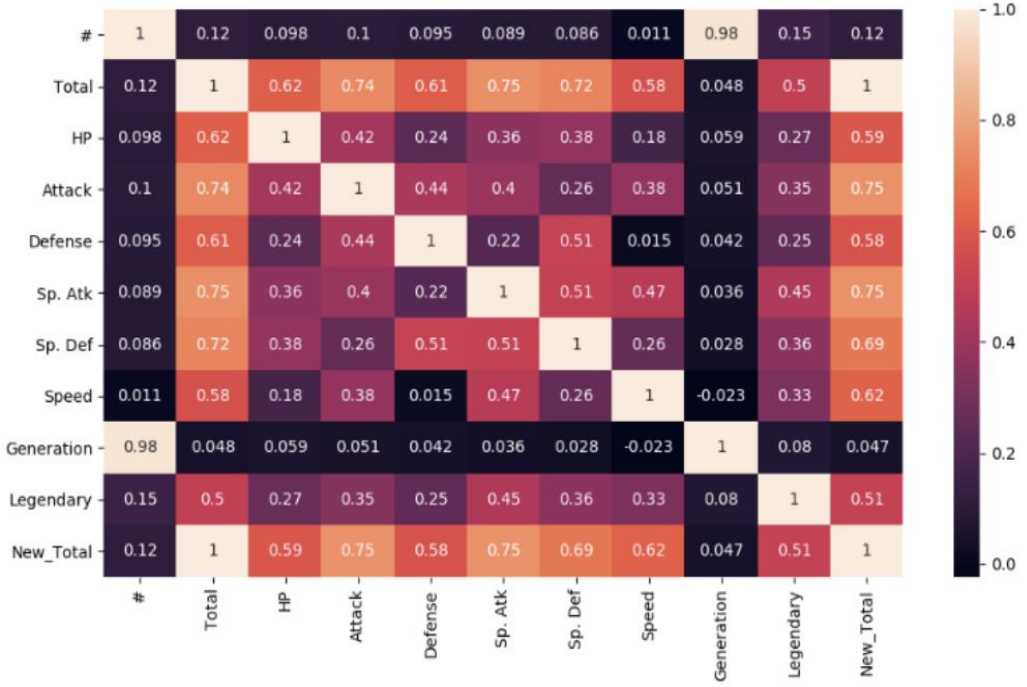
Rank에 따라 Generation 분포는 어떻게 될까?

그럼 이제 High, Middle, Low rank에 따른 Generation 분포는 어떠한지 원형 그래프로 확인해보자.



앞에서 전체 포켓몬 중 1, 3, 5세대 포켓몬의 비율이 비슷한 것을 확인하였다. 그러나 High rank 포켓몬 중에서는 3세대 포켓몬이 1세대, 5세대보다 훨씬 많으며, Low rank에서 또한 3세대 포켓몬이 1, 5세대보다 많음을 확인하였다. 따라서 포켓몬이 세대가 높아짐에 따라 진화하며 강한 포켓몬이 더 많을 것이라는 예측은 틀렸음을 증명하였다. 또한 3세대에 상위권의 강한 포켓몬이 가장 많다는 것을 알 수 있다.

1차 결론



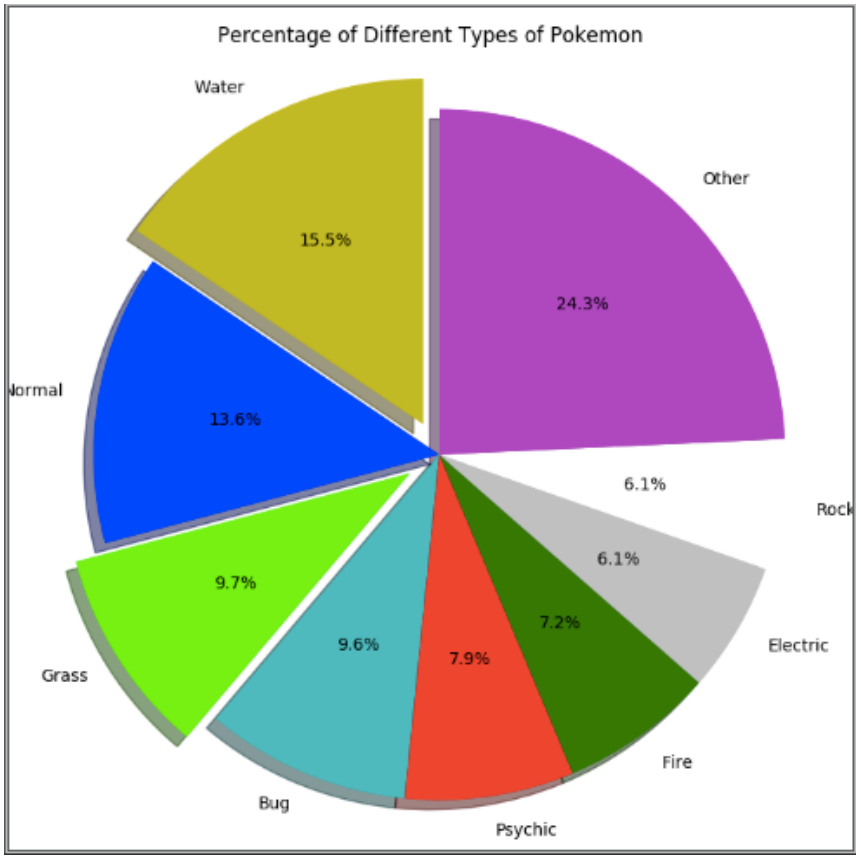
데이터 셋의 각 속성들끼리 서로 얼만큼의 영향을 끼치는지 알아보기 위해 Heatmap을 출력해보았다. Heatmap이란 서로 영향을 많이 끼칠수록 1에 가까운 확률을 나타내며 색깔이 밝아진다. #과 Generation은 속성들과 아무 관련이 없으며, Total은 모든 능력치들과 연관이 있음을 알 수 있다.

HP, Attack, Defense, Sp.Atk, Sp.Def, Speed 이 총 6가지 능력치를 종합적으로 보았을 때, 전체 포켓몬 중 High rank는 6.125%, Middle rank는 90.25%, Low rank는 3.625% 의 비율을 차지하고 있었다.

등급별 세대 분포를 분석해본 결과, High rank 에서 3세대 포켓몬이 가장 많은 비중을 차지하고 있었다. 따라서 3세대에 강한 포켓몬이 가장 많다는 것을 알 수 있다.

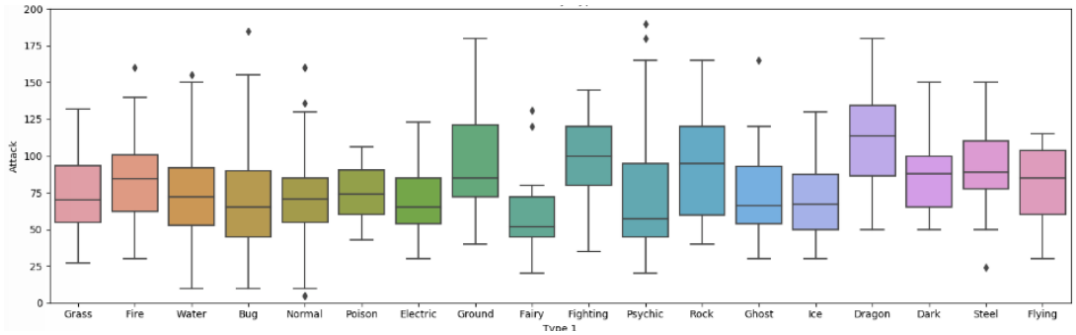
Type 별 포켓몬 수

어느 게임에서나 마찬가지로 '상성'이라는 것이 존재한다. 그것은 마치 가위바위보와 같다.
그래서 우리는 'Type1'에 주목해보기로 하였다.
(우선, Type2는 가진 포켓몬이 있고, 가지고 있지 않은 포켓몬이 있어 배제하였다.)
타입에 따른 포켓몬 분포는 다음 그래프와 같다.

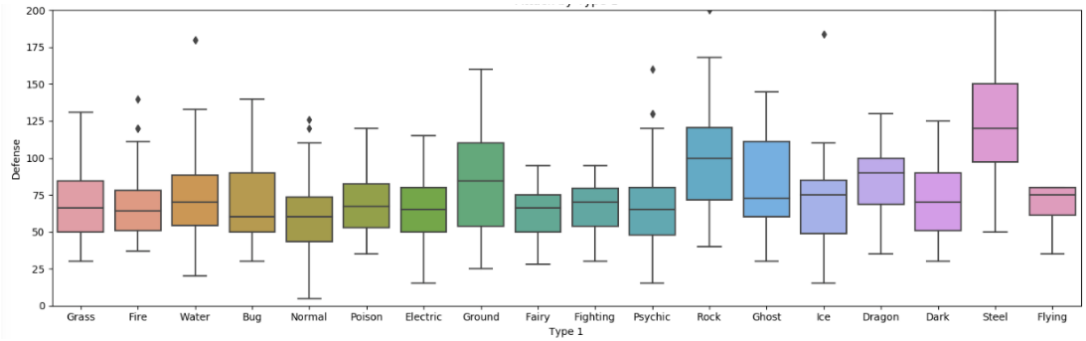


Type 별 Attack, Defense, HP 능력치

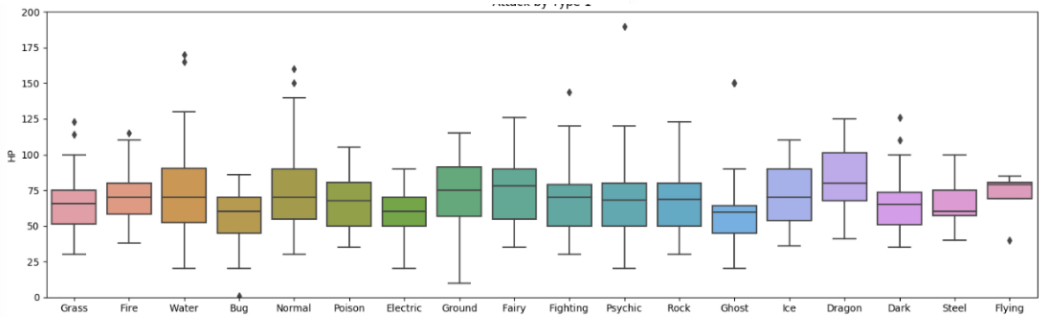
상성을 고려하기에 앞서 우리는 가장 중요한 능력이라 말할 수 있는 대표적인 3가지 Attack, Defense, HP를 Type별로 비교해보았다.



<Type에 따른 Attack 능력>



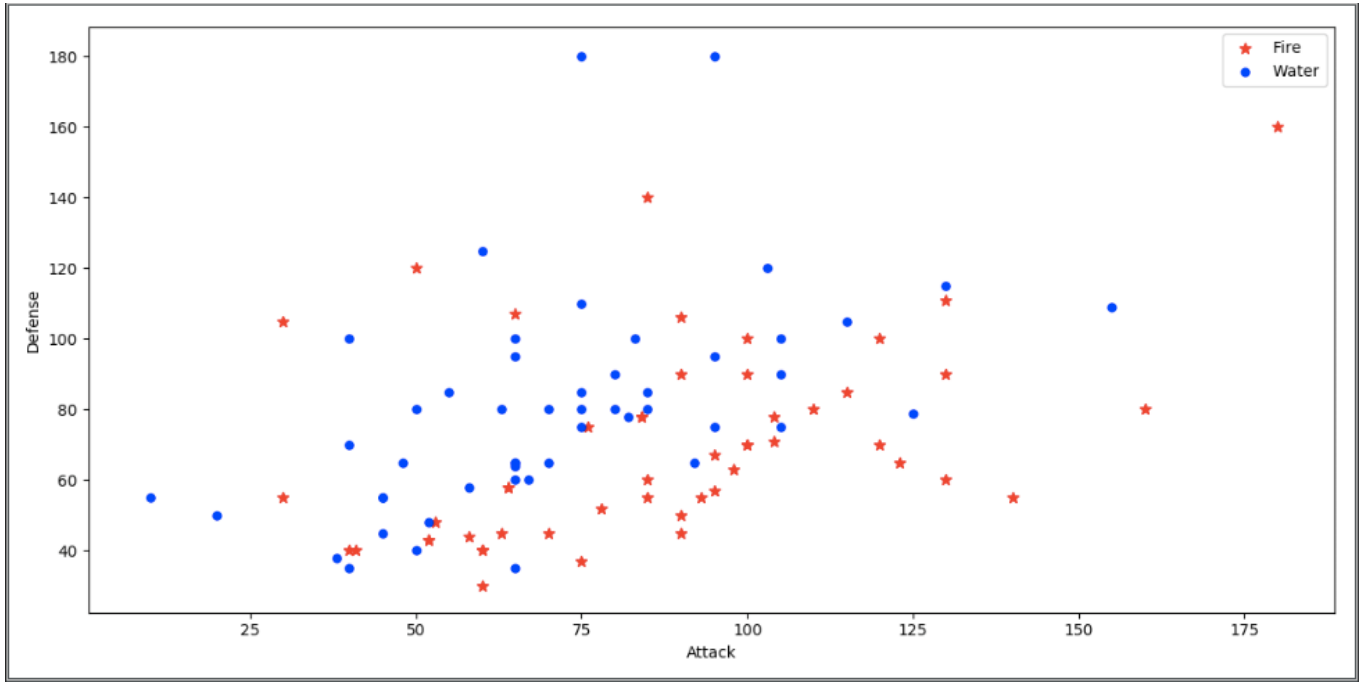
<Type에 따른 Defense 능력>



<Type에 따른 HP 능력>

Fire vs Water

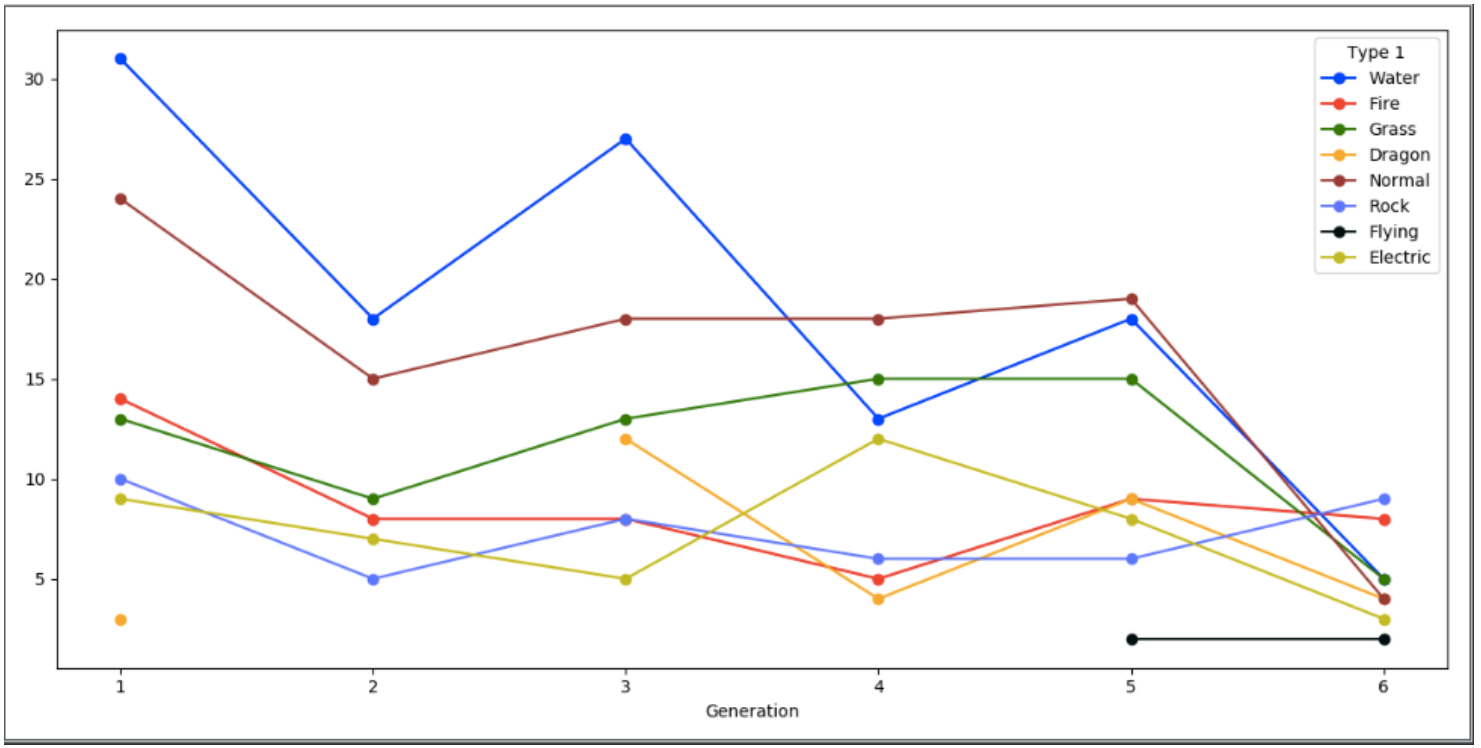
대표적인 예로 불(fire)과 물(water)의 Attack 과 Defense 관계를 살펴보기로 하였다.
우리는 상식적으로 물이 이길 것이라고 예측하였다.



그래프로 보았을 때, 전반적으로 Fire은 공격력이, Water은 방어력이 더 강하다는 것으로 결론지을 수 있다.

(번외) Random Box

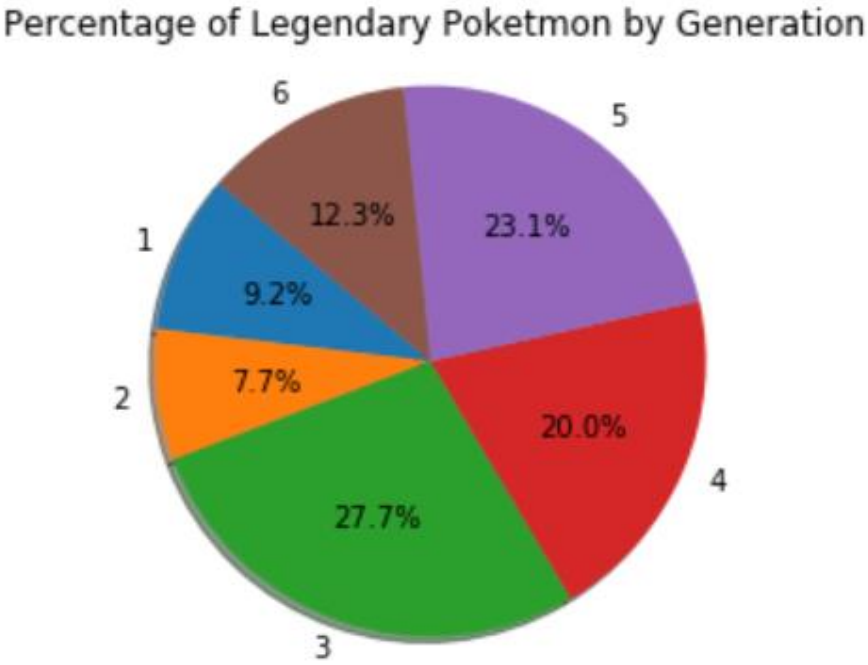
논지에서 잠시 벗어나, 요즘 유행하는 게임에는 뽑기 시스템인 'Random Box'가 있다. 이 랜덤박스에는 무엇이 나올지 모르며, 좋은 캐릭터나 아이템이 나올 확률은 매우 적다. 포켓몬 게임에도 이러한 시스템이 나올 수 있다. 하지만 포켓몬 수는 너무 광대하고 이를 나눌 적절한 수단은 세대라고 생각하였다. 그래서 뽑기 시스템이 나온다고 가정하였을 때, 포켓몬의 Type에 따른 세대별 포켓몬 수를 구해보았다.



세대가 높아질수록 레전드 포켓몬은 많아질까?

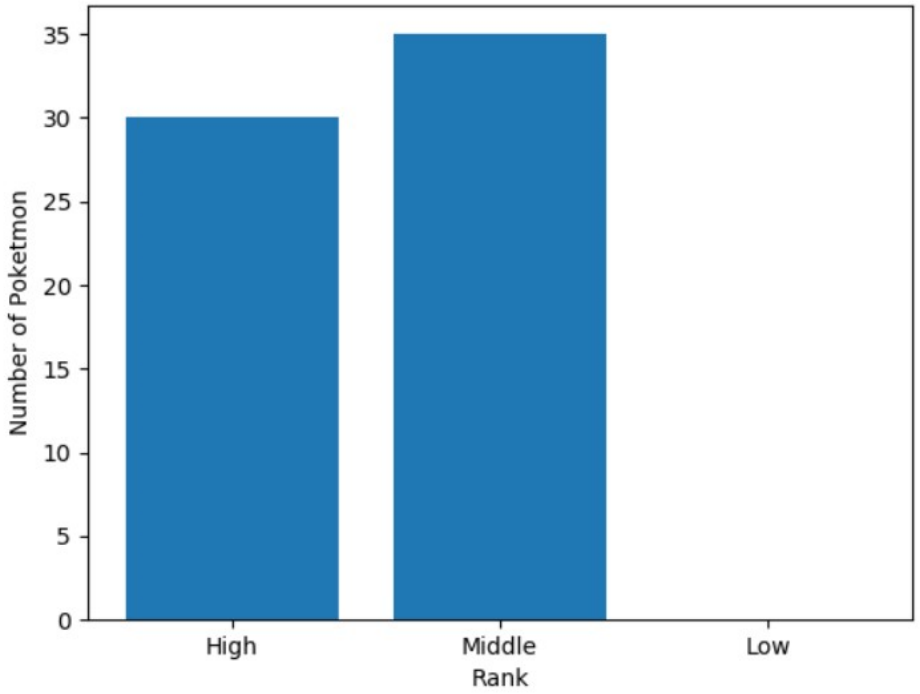
‘과연 상성을 무시하고 압도적인 힘으로 상대를 패배시킬 수 있는 포켓몬이 있을까?’ 라는 생각을 해보다가 뮤와 같은 레전드 포켓몬이 떠올랐다. 레전드 포켓몬이란 굉장히 극소수로 존재하는 희귀한 포켓몬을 말한다. 발견 확률이 극히 낮아 레전드라 불리기도 하지만 능력치 또한 전설에 남을 만큼 높다는 의미이기도 하다고 생각하였다.
데이터 셋에서 Legendary가 True or False로 저장되어 해당 포켓몬이 레전드 포켓몬인지 아닌지를 나타낸다.

우리는 세대가 높아질수록 레전드 포켓몬이 과연 많이 등장하는지를 확인해보기 위해 세대별 레전드 포켓몬의 비율을 확인해보았다.
그 결과, 그렇지 않다는 것을 확인할 수 있었다. 오히려 3세대에 레전드 포켓몬이 가장 많이 분포하였다.



레전드 포켓몬이 진짜 강할까?

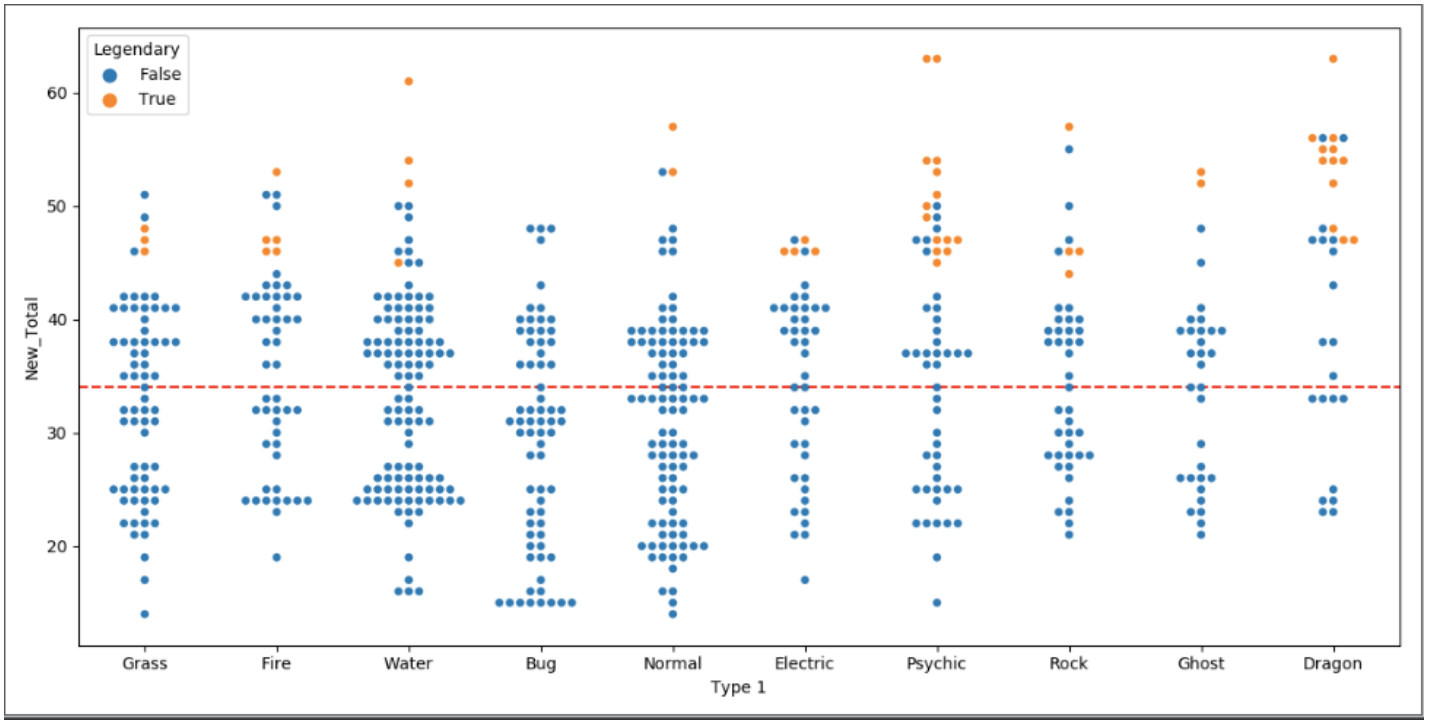
그렇다면 과연 레전드 포켓몬이 정말로 강한 능력치를 가지고 있는지 Rank별로 한번 분류해보았다. 다음 그래프는 Rank별 Legendary pokemon의 분포이다.



막대 그래프를 통해 확인해 본 결과, 레전드 포켓몬이 꼭 High Rank에만 있는 것은 아니었으며, 오히려 Middle Rank에 더 많다는 것을 알게 되었다. 그러나 Low Rank에는 0인 것으로 보아 능력치가 낮은 수준은 아니라는 것 또한 알 수 있었다.

같은 능력을 가진 레전드 포켓몬이 싸운다면?

만약 비슷한 능력치를 가진 레전드 포켓몬이 서로 싸우게 된다면, 이 때에는 상성이 다시 중요한 요소가 된다. 그래서 Type 별 레전드 포켓몬의 분포와 Total 점수를 비교해보았다.



로지스틱 회귀 모델을 이용한 데이터 분석

이제 우리는 포켓몬의 능력치를 구성하고 있는 속성들이 포켓몬의 강함 여부를 결정하는데 어떤 영향을 끼치는지 '로지스틱 회귀 모델'을 이용하여 데이터들을 분석한 후 편회귀계수(coef)를 통해 알아보려고 한다.

강함 여부를 결정하는 flag라는 변수를 설정하여 강한 포켓몬은 1로, 약한 포켓몬은 0으로 하였다. 이 모델을 통하여 각각의 feature에 대해 coef가 도출되었다. Coef 값이 양수일 경우, 그 값이 커지면 flag가 1일 확률(강한 포켓몬일 확률)이 높아지고, coef 값이 음수일 경우, 그 값이 작아질수록 flag가 0일 확률(약한 포켓몬일 확률)이 높아진다.

포켓몬 데이터 셋의 로지스틱 회귀 모델을 통해 도출된 결과를 보자.

Logit Regression Results

Dep. Variable:	flag	No. Observations:	91
Model:	Logit	Df Residuals:	85
Method:	MLE	Df Model:	5
Date:	Sun, 29 Jul 2018	Pseudo R-squ.:	0.4660
Time:	20:02:12	Log-Likelihood:	-33.540
converged:	True	LL-Null:	-62.807
		LLR p-value:	2.440e-11

	coef	std err	z	P>|z|	[0.025	0.975]
HP	-0.0495	0.015	-3.289	0.001	-0.079	-0.020
Attack	0.0415	0.016	2.589	0.010	0.010	0.073
Defense	-0.0104	0.015	-0.703	0.482	-0.040	0.019
Sp. Atk	0.0419	0.018	2.319	0.020	0.006	0.077
Sp. Def	0.0028	0.019	0.149	0.881	-0.034	0.040
Speed	-0.0213	0.013	-1.658	0.097	-0.046	0.004

	coef
HP	-0.0495
Attack	0.0415
Defense	-0.0104
Sp. Atk	0.0419
Sp. Def	0.0028
Speed	-0.0213

로지스틱 회귀 모델을 이용한 데이터 분석

분석해보면 ,

‘HP’ feature은 coef가 음수로, 강한 포켓몬일수록 대체적으로 HP가 낮다는 것을 의미한다.

‘Attack’ feature은 coef가 양수로, 강한 포켓몬일수록 대체적으로 공격력이 높다는 것을 의미한다.

‘Defense’ feature은 coef가 음수로, 강한 포켓몬일수록 대체적으로 방어력이 낮다는 것을 의미한다.

‘Sp. Atk’ feature은 coef가 양수로, 강한 포켓몬일수록 대체적으로 특수공격력이 높다는 것을 의미한다.

‘Sp. Def’ feature은 coef가 양수로, 강한 포켓몬일수록 대체적으로 특수공격에 대한 방어력이 높다는 것을 의미한다.

‘Speed’ feature은 coef가 음수로, 강한 포켓몬일수록 대체적으로 속도가 낮다는 것을 의미한다.

최종 결론

여기까지 오면서 우리는 6가지의 능력(HP, Attack, Defense, Sp.Atk, Sp.Def, Speed), Type(상성), Legendary(레전드) 이 세 항목을 기준으로 데이터를 분석해보았다.
처음에 우리가 찾고자 했던 가장 강한 포켓몬은 "없다"라는 결론이 나왔다. 분석을 진행하는 과정에서, 그리고 마무리를 하면서 포켓몬 게임에서 승리하기 위해 가장 중요한 것은 능력치나 레전드가 아닌 상성이라고 결론지었다.
능력치가 아무리 좋아도 상성이 절대적으로 우위에 있는 포켓몬에게 질 수 있으며, 레전드 포켓몬이 일반 포켓몬을 상대로 하여 상성을 무시할 정도로 아무리 강해도 상대편이 레전드 포켓몬을 내놓는 순간 다시 상성 싸움으로 돌아갈 수 있다.
그에 따라서 '절대적으로 강한 1인자 포켓몬은 없다'라는 결론이 나왔고,
상황에 따라 상대편의 포켓몬에 맞추어 상성 조합에서 이기도록 유동적으로 맞서 싸울 포켓몬을 고르는 것이 가장 중요하다고 생각하였다.
흥미로운 주제에 팀원들과 열정적으로 단합할 수 있었고 분석을 진행하는 동안 학창 시절 포켓몬 골드를 밤낮으로 열심히 하던 모습이 떠올라 오랜만에 향수에 젖어들 수 있는 시간이었다.