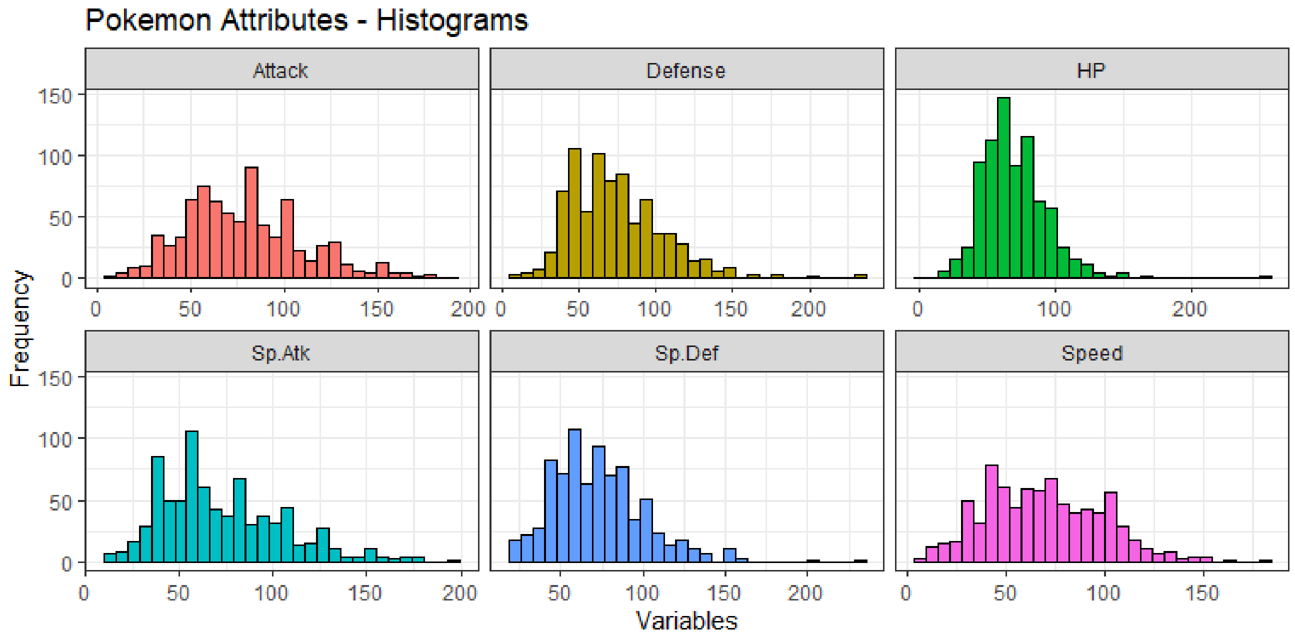
MA4128 Assignment

Jack Curtin, ID: 16181484

Technical Report

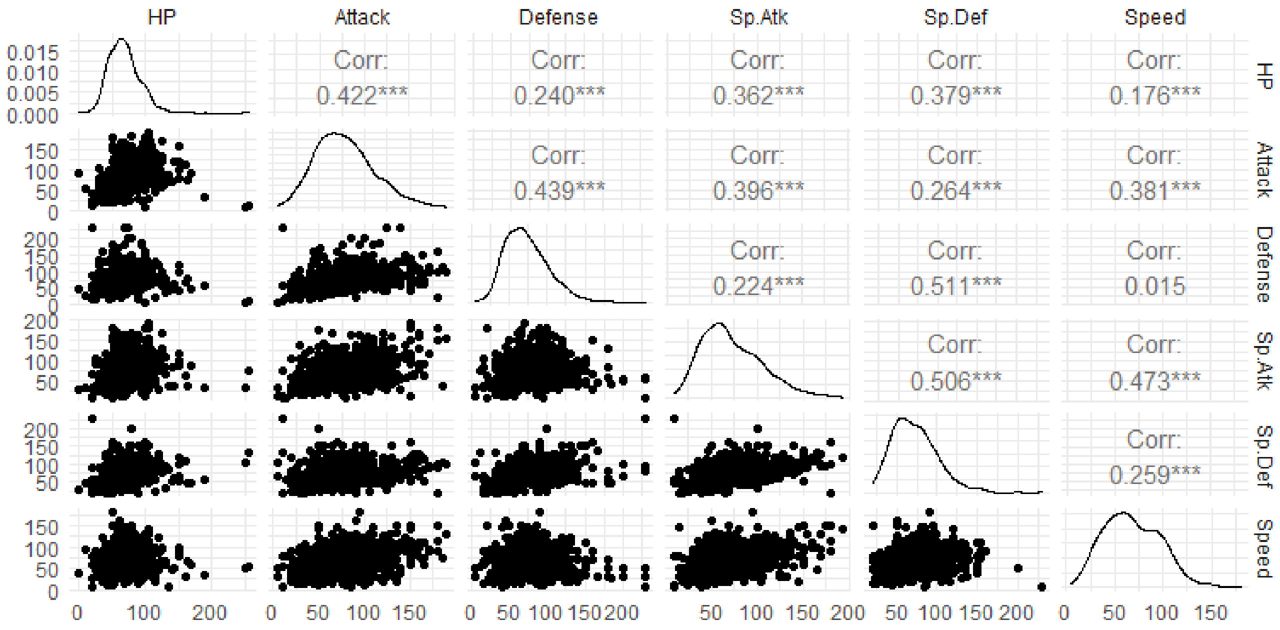
Exploratory Analysis



The dataset we are considering consists of 800 Pokemon, each of which have a unique combination of battling statistics along with other characteristics. Our main objective is to identify what Pokemon exhibit similar battling characteristics. It should be noted that *Type* and *Generation* are the only categorical variables in this dataset. To begin analysing our data, we will look at a summary statistic for each of our non-categorical variables to determine if our data contains any abnormalities and if the data needs to be standardised.

var	min	q25	median	q75	max	mean	sd
Attack	5	55.00	75	100	190	79.00125	32.45737
Defense	5	50.00	70	90	230	73.84250	31.18350
HP	1	50.00	65	80	255	69.25875	25.53467
Sp.Atk	10	49.75	65	95	194	72.82000	32.72229
Sp.Def	20	50.00	70	90	230	71.90250	27.82892
Speed	5	45.00	65	90	180	68.27750	29.06047

From this table there doesn't seem to be any abnormalities in our data and the variances don't vary massively although we will still standardise our data to ensure each variable is of equal weight when we look into creating clusters. We can also look at the histograms above for our data and use the `ggpairs()` function to create pairs plots for our numeric variables, which has the following output:



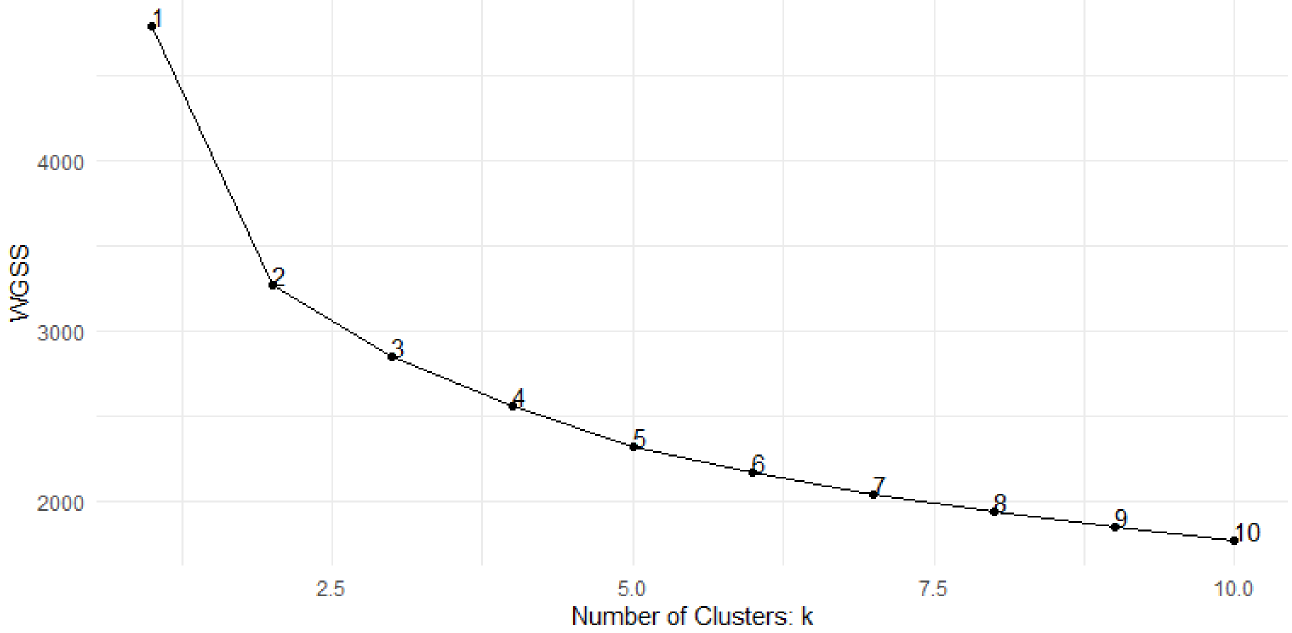
This can be interpreted to mean that whilst there are no strong associations in our data, we do have moderate associations between $Sp.\,Def$ and $Defense$ and $Sp.\,Def$ and $Sp.\,Atk$. There are also appears to be a few outliers in the data, although none seem to be of major significance.

Standardise the data

To begin formally analysing our data we must first standardised the data. We achieve this by dividing each of the variables by their standard deviation, which results in each standard deviation being equal to one.

var	min	median	max	mean	sd
Attack	0.1540482	2.310724	5.853833	2.434001	1
Defense	0.1603412	2.244777	7.375695	2.367999	1
HP	0.0391624	2.545559	9.986423	2.712342	1
Sp.Atk	0.3056020	1.986413	5.928680	2.225394	1
Sp.Def	0.7186769	2.515369	8.264785	2.583733	1
Speed	0.1720550	2.236715	6.193980	2.349497	1

We can now use this standardised dataset to carry out k-means clustering, which is a form of hierarchical clustering. A feature of a clustering solution in which we are interested in is the within group sum of squares (WGSS), which we will use to determine the number of clusters. We will use the k-means algorithm over the range $k = 1, \ldots, 10$ clusters and we will record the WGSS for each value of k.



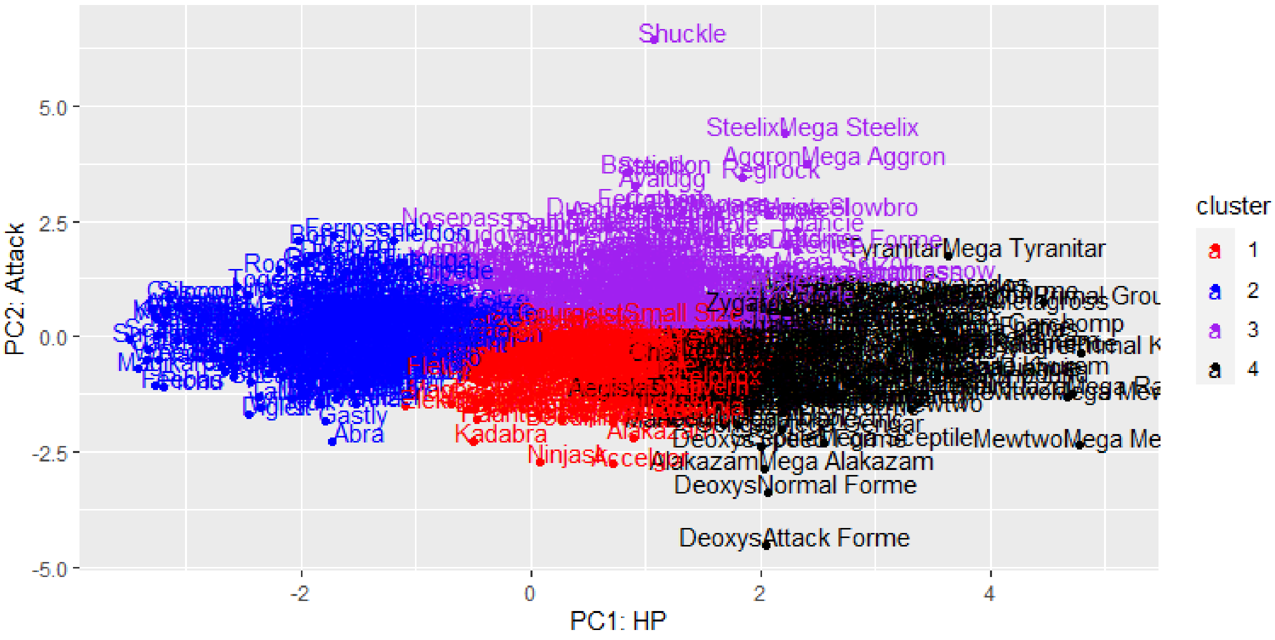
We can determine the *elbow point* from our plot to be between 2 and 4, in this case we will choose 4 as the plot gradually tapers after 4. Below we see the number of Pokemon in each cluster and also the centers of each cluster.

Cluster	#	HP	Attack	Defense	Sp.Atk	Sp.Def	Speed
1	203	2.827413	2.576627	2.096599	2.378276	2.564048	3.133096
2	280	1.956724	1.656019	1.667778	1.460996	1.768587	1.691424
3	195	3.190835	2.652000	3.301220	2.263022	3.198481	1.805254
4	122	3.490272	3.633771	2.935033	3.665221	3.504728	3.425869

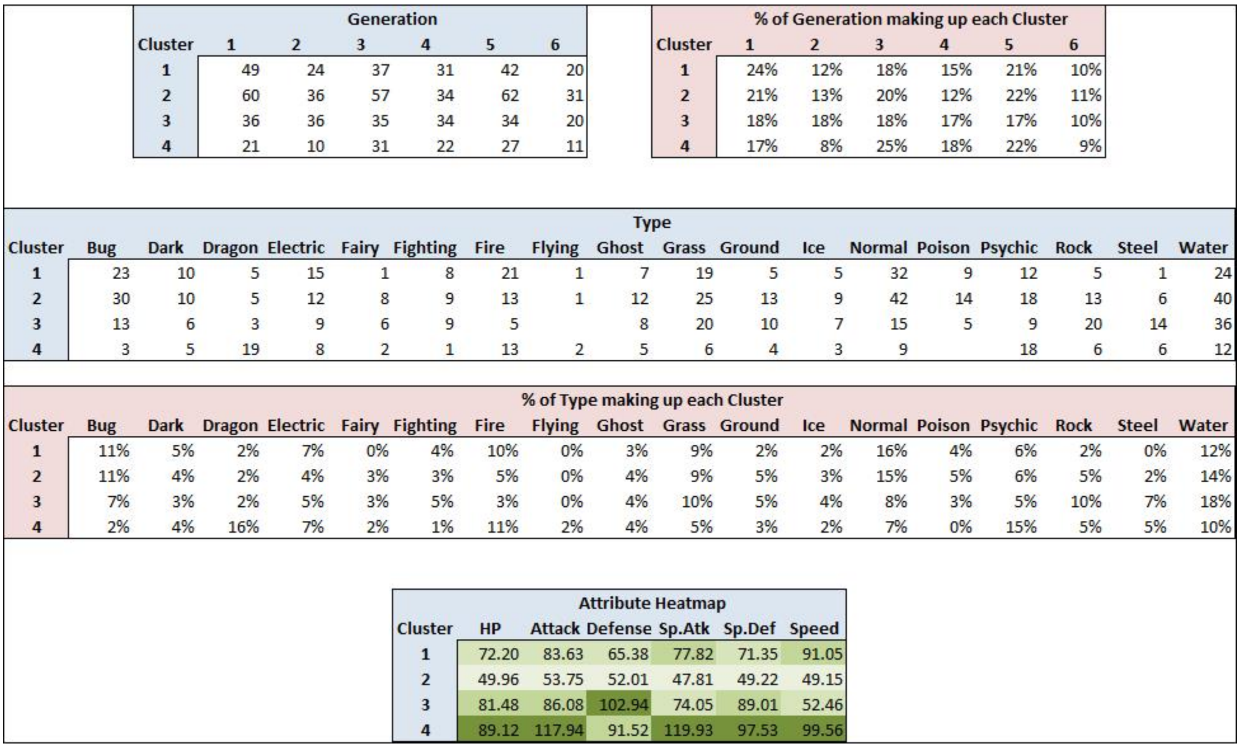
From this output we can see that Cluster 2 has the most Pokemon and also the lowest scores. Cluster 4 has the lowest number of Pokemon yet has the highest scores in everything except $Defense$. Cluster 1 has high HP but lower than average $Defense$ and Cluster 3 has lower than average $Speed$ but the highest value of $Defense$.

Visualising the cluster solution

We can visualize the clusters using principal components where we take Hp to be PC1 and $Attack$ to be PC2 as shown below. We can see that our clusters are reasonably distinguishable in this plot.



We can also further analyse our results if we export a data frame to excel which has an additional column defining each cluster and another column defining the average of our standardised numeric variables (to objectively determine the best Pokemon). We find a summary of our results from that table to be as follows which shows the amount of each Generation and the amount of each Type in our Clusters.



	Generation					
Cluster	1	2	3	4	5	6
1	49	24	37	31	42	20
2	60	36	57	34	62	31
3	36	36	35	34	34	20
4	21	10	31	22	27	11

	% of Generation making up each Cluster					
Cluster	1	2	3	4	5	6
1	24%	12%	18%	15%	21%	10%
2	21%	13%	20%	12%	22%	11%
3	18%	18%	18%	17%	17%	10%
4	17%	8%	25%	18%	22%	9%

	Type																	
Cluster	Bug	Dark	Dragon	Electric	Fairy	Fighting	Fire	Flying	Ghost	Grass	Ground	Ice	Normal	Poison	Psychic	Rock	Steel	Water
1	23	10	5	15	1	8	21	1	7	19	5	5	32	9	12	5	1	24
2	30	10	5	12	8	9	13	1	12	25	13	9	42	14	18	13	6	40
3	13	6	3	9	6	9	5		8	20	10	7	15	5	9	20	14	36
4	3	5	19	8	2	1	13	2	5	6	4	3	9		18	6	6	12

	% of Type making up each Cluster																	
Cluster	Bug	Dark	Dragon	Electric	Fairy	Fighting	Fire	Flying	Ghost	Grass	Ground	Ice	Normal	Poison	Psychic	Rock	Steel	Water
1	11%	5%	2%	7%	0%	4%	10%	0%	3%	9%	2%	2%	16%	4%	6%	2%	0%	12%
2	11%	4%	2%	4%	3%	3%	5%	0%	4%	9%	5%	3%	15%	5%	6%	5%	2%	14%
3	7%	3%	2%	5%	3%	5%	3%	0%	4%	10%	5%	4%	8%	3%	5%	10%	7%	18%
4	2%	4%	16%	7%	2%	1%	11%	2%	4%	5%	3%	2%	7%	0%	15%	5%	5%	10%

	Attribute Heatmap					
Cluster	HP	Attack	Defense	Sp.Atk	Sp.Def	Speed
1	72.20	83.63	65.38	77.82	71.35	91.05
2	49.96	53.75	52.01	47.81	49.22	49.15
3	81.48	86.08	102.94	74.05	89.01	52.46
4	89.12	117.94	91.52	119.93	97.53	99.56

We find the worst Pokemon to objectively be the following:

Name	Type	HP	Attack	Defense	Sp.Atk	Sp.Def	Speed	Generation	Cluster
Sunkern	Grass	30	30	30	30	30	30	2	2
Azurill	Normal	50	20	40	20	40	20	3	2
Kricketot	Bug	37	25	41	25	41	25	4	2
Wurmple	Bug	45	45	35	20	30	20	3	2
Weedle	Bug	40	35	30	20	20	50	1	2
Ralts	Psychic	28	25	25	45	35	40	3	2

We find the best Pokemon to objectively be the following:

	Name	Type	HP	Attack	Defense	Sp.Atk	Sp.Def	Speed	Generation	Cluster
795	Arceus	Normal	120	120	120	120	120	120	4	4
796	GroudonPrimal Groudon	Ground	100	180	160	150	90	90	3	4
797	KyogrePrimal Kyogre	Water	100	150	90	180	160	90	3	4
798	RayquazaMega Rayquaza	Dragon	105	180	100	180	100	115	3	4
799	MewtwoMega Mewtwo X	Psychic	106	190	100	154	100	130	1	4
800	MewtwoMega Mewtwo Y	Psychic	106	150	70	194	120	140	1	4

Conclusion

Upon further inspection using summaries from Excel to analyse each cluster we come to the conclusion that the Pokemon which have the best battling characteristics are *Legendary Pokemon* and *Final Evolution Pokemon*. From our analysis Cluster 4 contains the best overall Pokemon as it contains the most Legendary and Final Evolution Pokemon. Cluster 2 contains the worst overall Pokemon and it is therefore best to avoid choosing any pokemon from this cluster. Clusters 1 and 3 have similar values with 3 having the best overall $Defense$ values and 1 having the second best overall $Speed$ values. From our analysis it seems that no Generation stands out as having better overall Pokemon than the other Generations. The best Type appears to be Dragon, Psychic and Fire and the worst Type seems to be Bug and arguably Normal. Water seems to be the most diverse Type as it makes up roughly 14% of each cluster. We thus conclude that the best Pokemon are Legendary and Final Evolution Pokemon, the majority of which are in Cluster 4, and that it is best to obtain Dragon and Fire Type and avoid Bug Type if possible.

Each cluster can be summarised as follows:

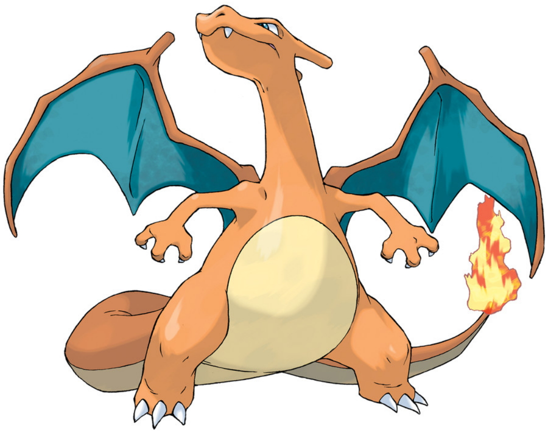
- Cluster 1 Contains the third best overall Pokemon. Has the second highest values in $Sp.\ Atk$ and $Speed$. Mainly composed of Generation 1 and 5 with average amounts from the other Generations. Consists mainly of Normal, Water and Bug Type with notably lower than average amounts of Rock and Steel Type. We conclude that this cluster contains lighter and thus faster Pokemon although it lacks good $Defense$ and HP.

- Cluster 2 Contains the worst overall Pokemon. Mainly composed of Generation 1, 3 and 5 with average amounts from the other Generations. Consists mainly of Normal, Water, Bug and Fire Type. We conclude that this cluster should be avoided if possible.

- Cluster 3 Contains the second best overall Pokemon. Has the highest $Defense$ values and second highest values in everything else aside $Sp.\ Atk$ and $Speed$. Composed of a relatively equal amount of each Generation with the exception of Generation 6 which make up 10% of this Cluser. Consists mainly of Water, Rock and Grass Type with the highest percentage of Steel Type. We conclude that this cluster consists of heaver and thus slightly more powerful Pokemon that Cluster 1, although these Pokemon lack $Speed$ they still have an edge over Cluster 1.

- Cluster 4 Contains the best overall Pokemon. Has the highest overall values in everything except $Defense$. Mainly composed of Generation 3 and 5 with a slightly lower than average amount of Generation 2. Consists mainly of Dragon, Psychic, Fire and Water Type with the lowest percentage of Bug Type and no Poison Type. We conclude that this cluster has the best overall Pokemon and it is therefore best to obtain any Pokemon from this cluster in order to have an advantage over the other clusters.

Non-Technical Report



We began analysising the Pokemon data by looking at the *numeric* variables which are the variables that describe the battling characteristics of each Pokemon. We determined that because the variances (denoted *sd*) for each numeric variable were different from each other that we should *standardise* the data, which essentially means to divide each numeric variable by the *sd* value to ensure to that each variable is of equal weight. We did this as otherwise there would be more weight on variables with smaller variance such as HP. We used these standardised variables to carry out *k-means* clustering, which is an iterative partitioning clustering algorithm which divides the data into k disjoint, non-overlapping groups such that observations within a group are similar and observations in different groups are different. We determined the optimal number of clusters to be $k = 4$, and as such we partitioned our data into 4 different clusters.

We could see that some clusters had better variable scores than others, which implied that Pokemon in one cluster had better scores than those in another. We could see that the average scores for each variable were higher in Cluster 4 than the other clusters and that Cluster 2 contained the worst overall scores for each variable, which suggests that it is best to avoid Pokemon contained in Cluster 2 and to seek out any Pokemon from Cluster 4. We also looked at a summary for the Generation and Type contained in each cluster and we came to the conclusion that Generation has no major bearing on how good a Pokemon's battling attributes will be. We did however conclude that the Bug Type appeared to be the worst and that Dragon and Psychic appeared to be the best Types in terms of battling attributes. In reality the defining factor as to why the Cluster 4 scores were far better than the others was because this cluster contained what are known as *Legendary Pokemon* and *Final Evolution Pokemon*, which are Pokemon that are rare and as such difficult to obtain.