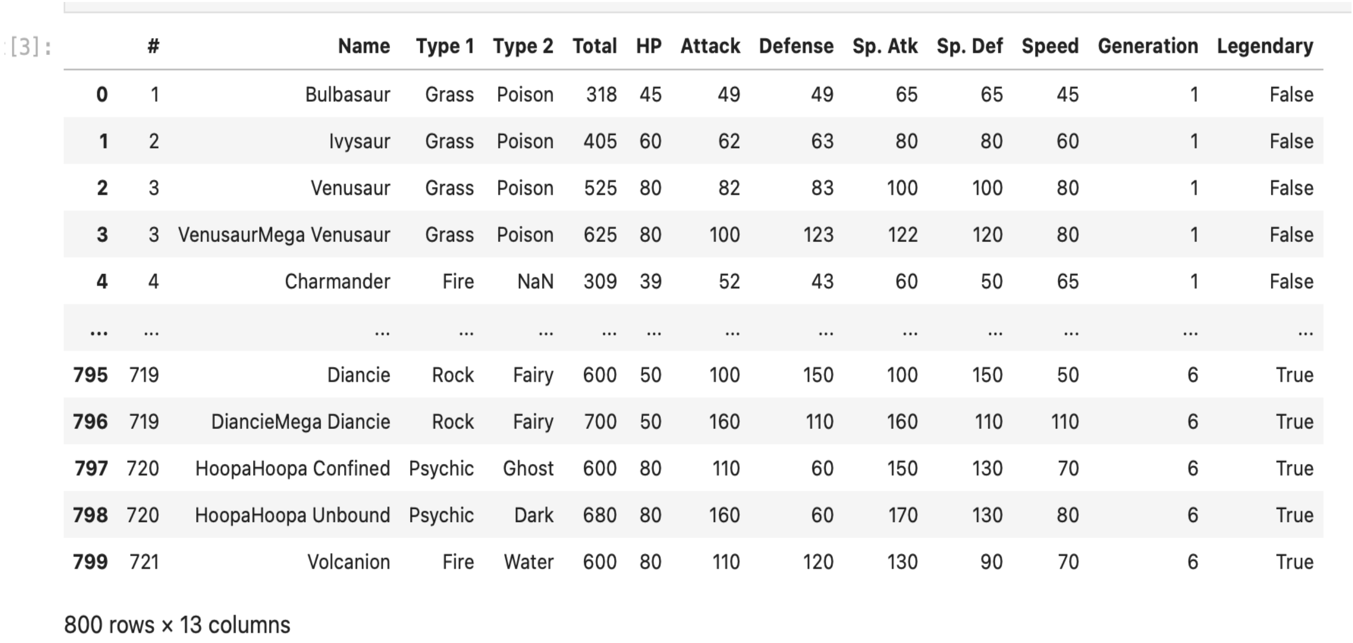
```
import pandas as pd
import numpy as np
df = pd.read_csv("Data-Week4.csv")
df
```

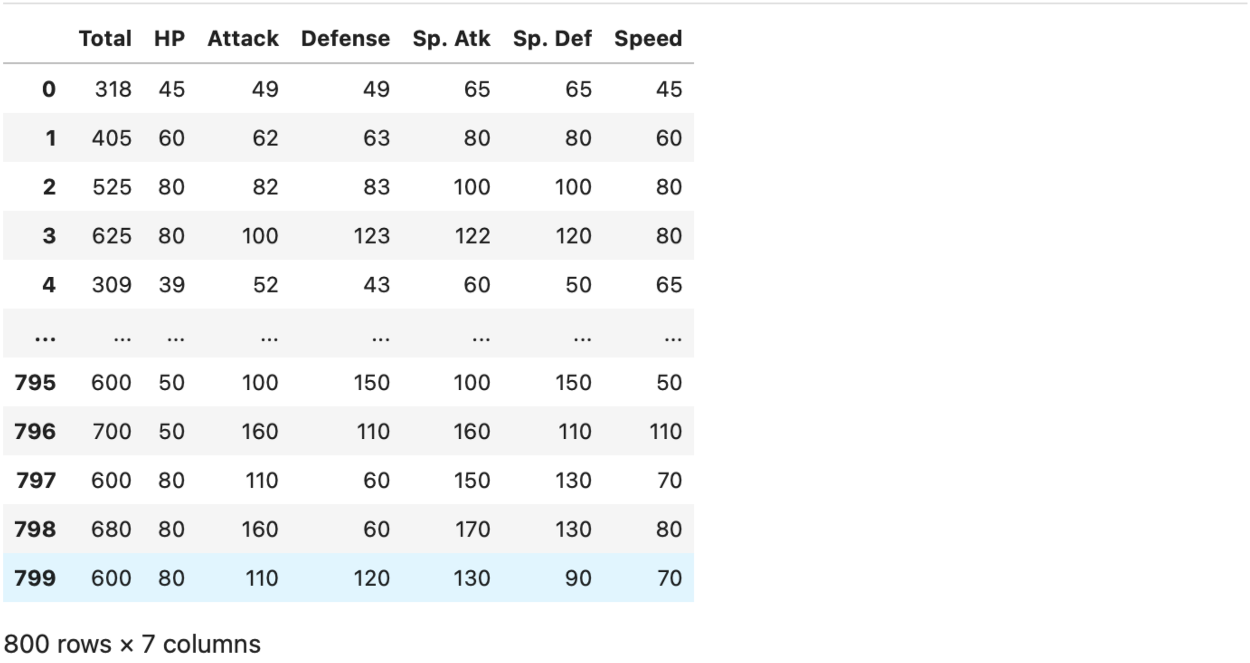


[3]:

	#	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
795	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	True
796	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	True
797	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	True
798	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	True
799	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	True

800 rows × 13 columns

```
df.drop(labels = ["#","Name","Type 1","Type 2","Generation","Legendary"],axis = 1,inplace=True)
df
```

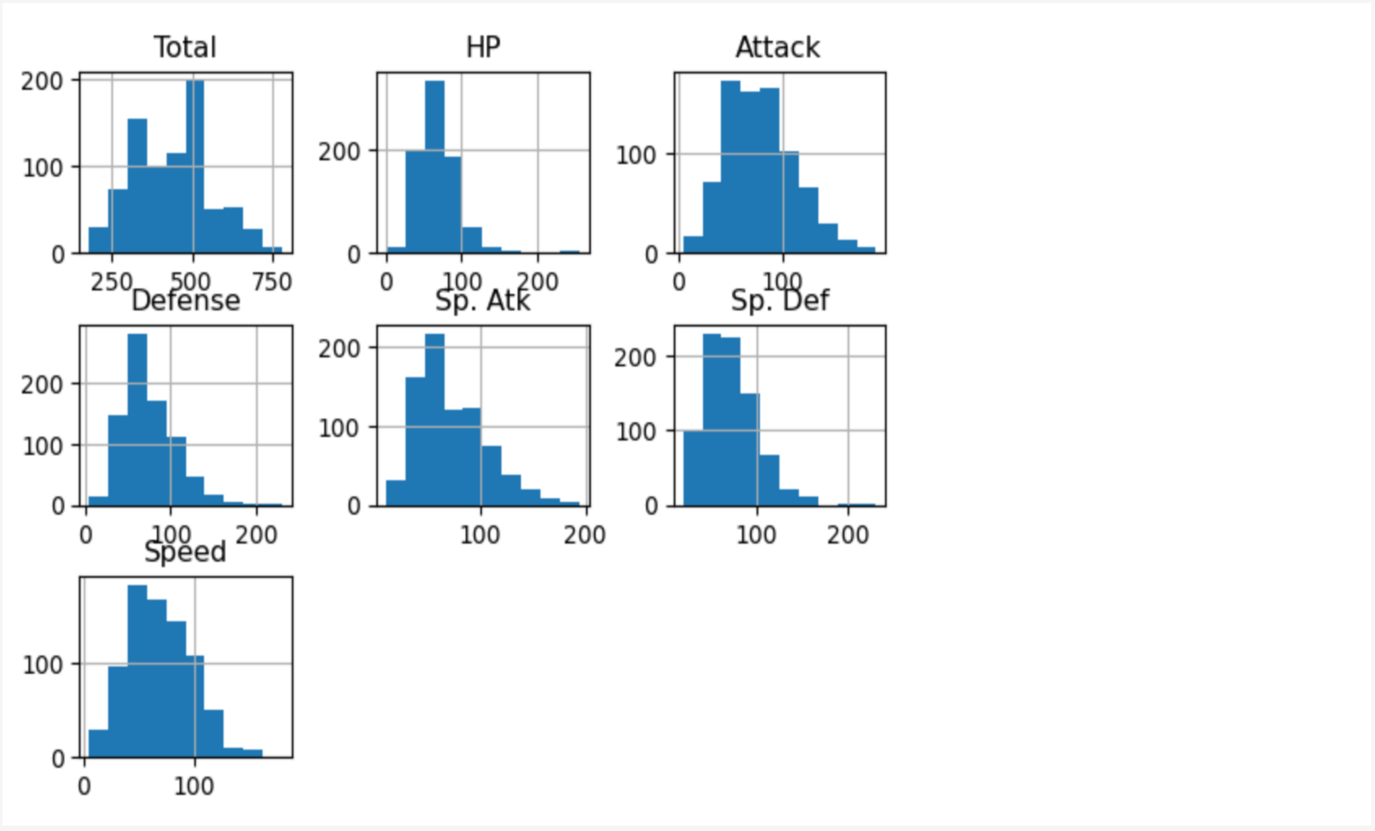


	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
0	318	45	49	49	65	65	45
1	405	60	62	63	80	80	60
2	525	80	82	83	100	100	80
3	625	80	100	123	122	120	80
4	309	39	52	43	60	50	65
...	...	...	...	...	...	...	...
795	600	50	100	150	100	150	50
796	700	50	160	110	160	110	110
797	600	80	110	60	150	130	70
798	680	80	160	60	170	130	80
799	600	80	110	120	130	90	70

800 rows × 7 columns

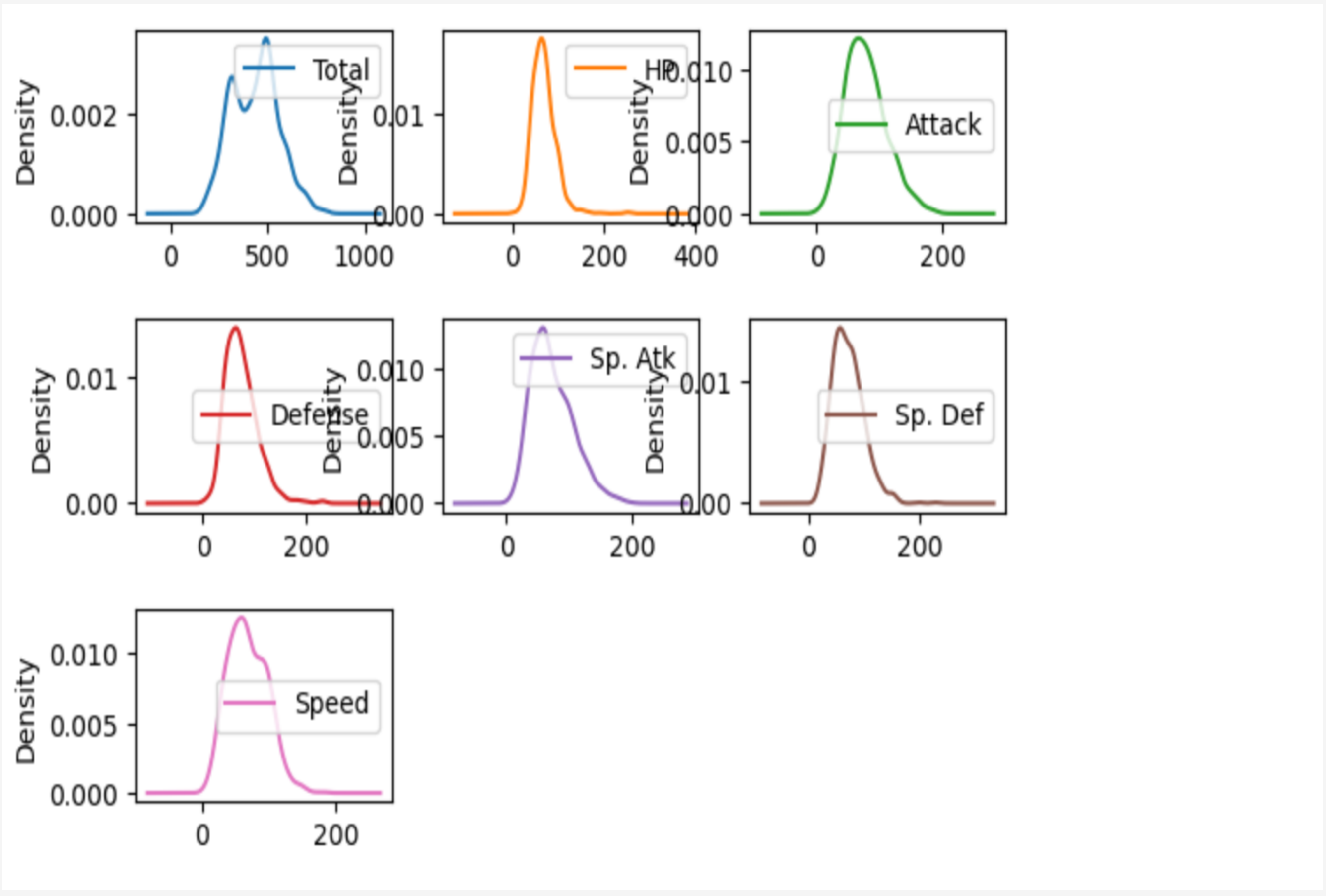
```
import matplotlib.pyplot as plt
df.hist()
plt.subplots_adjust(top = 0.9,wspace = 0.4,hspace = 0.4)
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



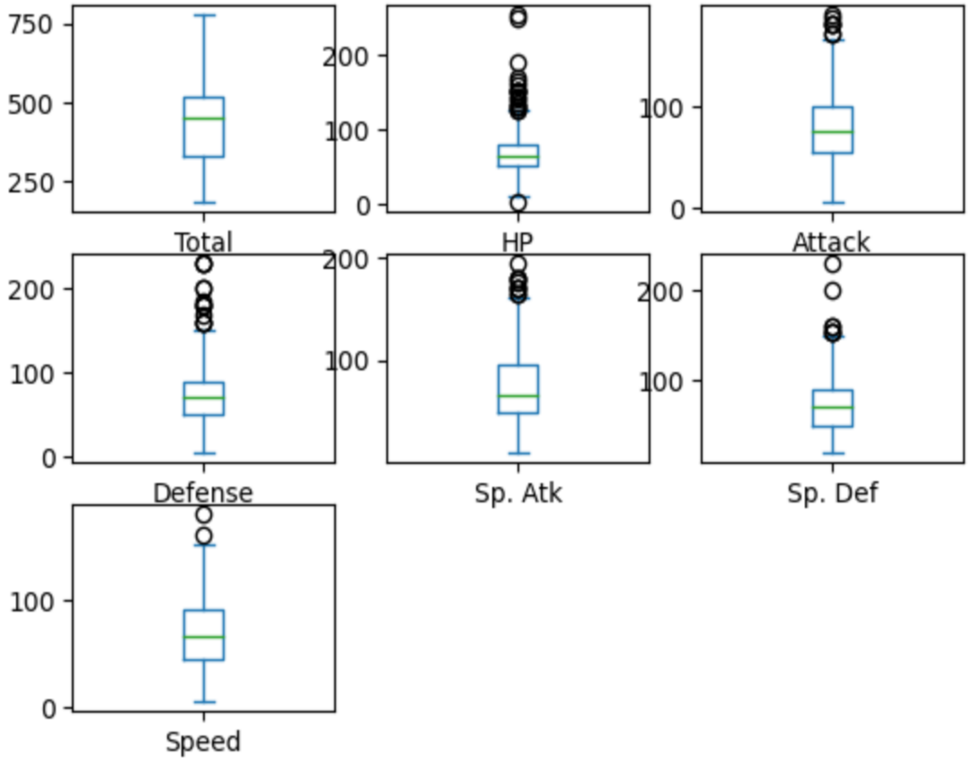
The histograms for the quantitative data are shown above. With the first and second two peaks, the total has a curve. The graphs for Defense, Sp.Atk, and Sp.Def have right-skewed. Finally, the curves for HP, Attack, and Speed can be normal or gaussian.

```
df.plot(kind = "density",subplots = True,layout = (3,3),sharex = False)
plt.subplots_adjust(hspace = 0.5)
plt.show()
```



HP, Attack and Speed features have normal or gaussian curves but Defence, Sp. Atk and Sp. Def are Right-Skewed curves

```
df.plot(kind = "box",subplots = True,layout = (3,3),sharex = False)
plt.show()
```



In Total box plot there are no outliers, upper quartile is more than lower ones

In HP there are outliers on both sides. Four quartiles are very small but median is at center i.e it divides equal no of people on both sides regardless of outliers.

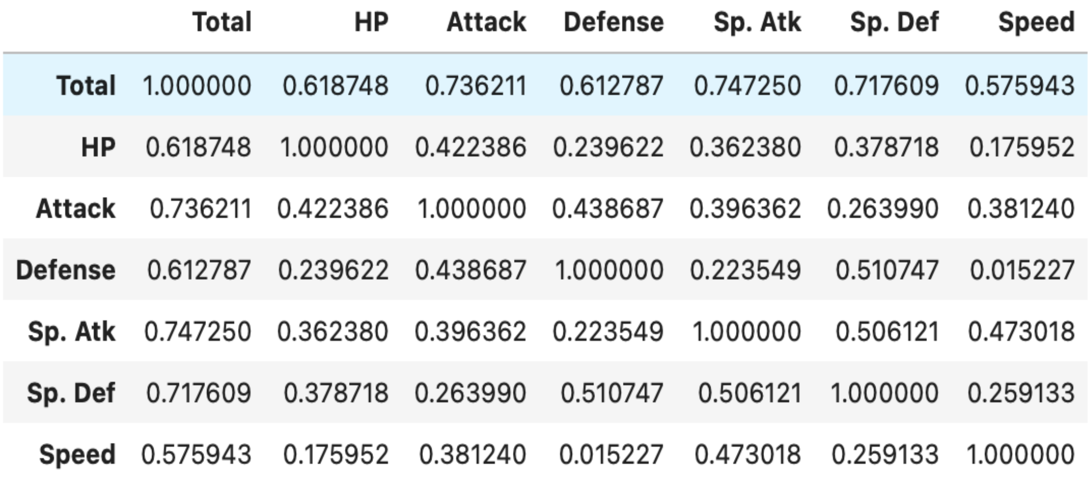
In Attack there are outliers on top and above 100 attack strength are only for people in the last quartile which means they are very less.

In Defense there are outliers on top and defense strength similar to attack.

In Sp.Attack, and Sp.Def are also following the path of Attack, but the third quartile is very high in Sp.Atk and Almost similar and less outliers in Sp.Def.

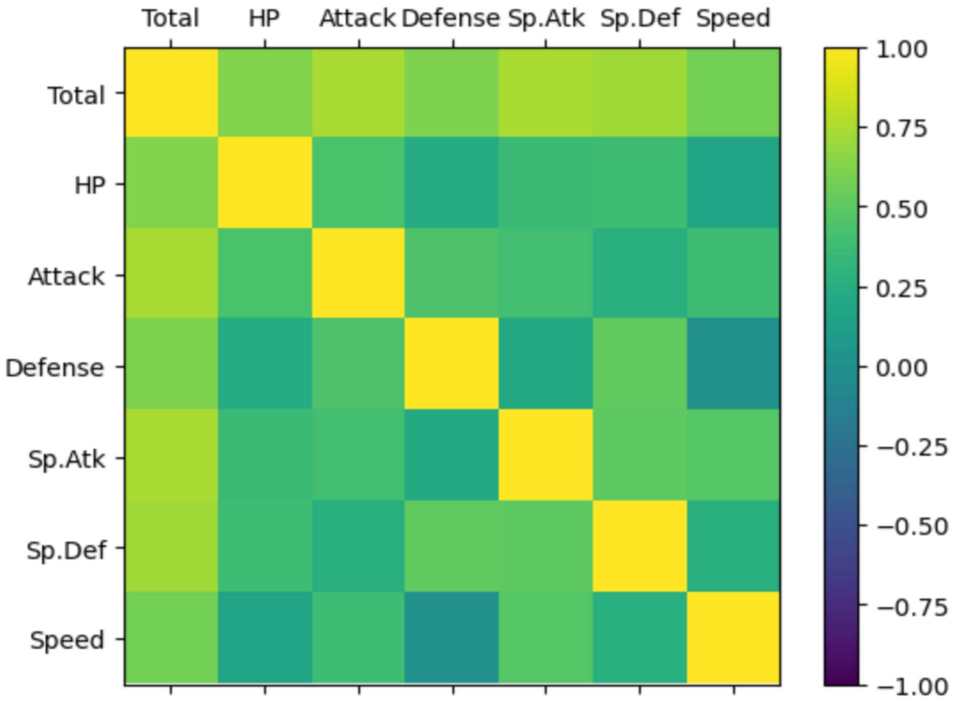
In Speed there are only two outliers and three quartiles are less than 100 and the last quartile is more than 100.

```
correlations = df.corr(method = "pearson")
correlations
```



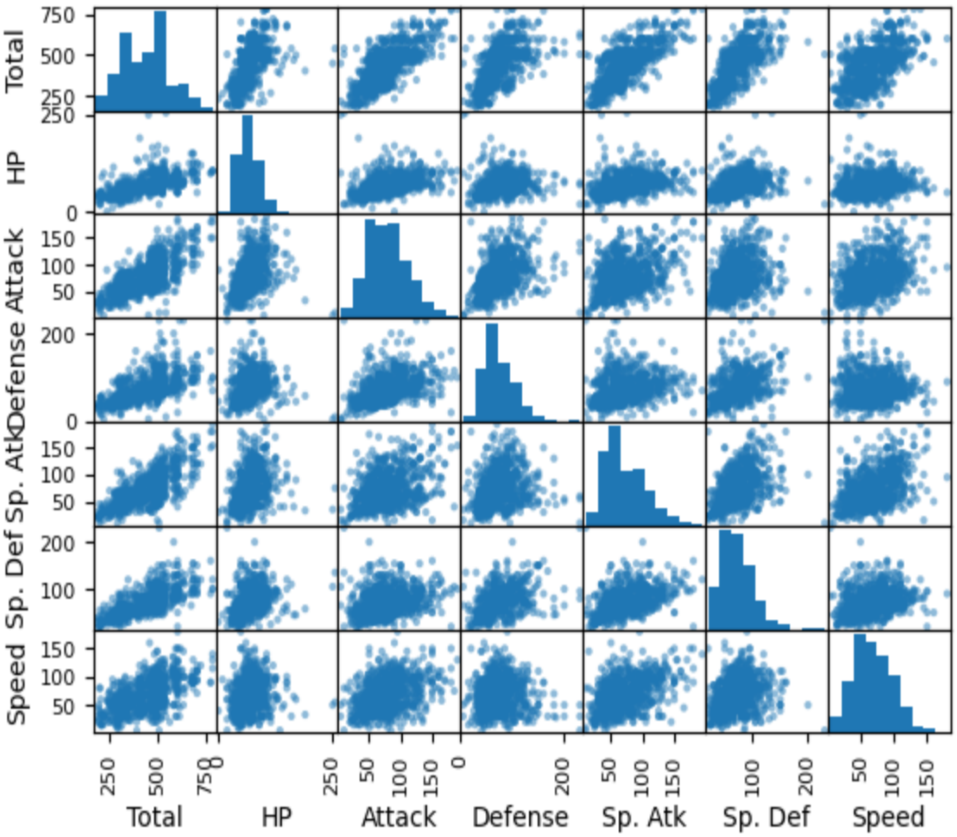
	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
Total	1.000000	0.618748	0.736211	0.612787	0.747250	0.717609	0.575943
HP	0.618748	1.000000	0.422386	0.239622	0.362380	0.378718	0.175952
Attack	0.736211	0.422386	1.000000	0.438687	0.396362	0.263990	0.381240
Defense	0.612787	0.239622	0.438687	1.000000	0.223549	0.510747	0.015227
Sp. Atk	0.747250	0.362380	0.396362	0.223549	1.000000	0.506121	0.473018
Sp. Def	0.717609	0.378718	0.263990	0.510747	0.506121	1.000000	0.259133
Speed	0.575943	0.175952	0.381240	0.015227	0.473018	0.259133	1.000000

```
import matplotlib.pyplot as plt
headers=["Total","HP","Attack","Defense","Sp.Atk","Sp.Def","Speed"]
fig=plt.figure()
ax=fig.add_subplot(111)
cax=ax.matshow(correlations, vmin=-1,vmax=1)
fig.colorbar(cax)
ticks=np.arange(0,7,1)
ax.set_xticks(ticks)
ax.set_yticks(ticks)
ax.set_xticklabels(headers)
ax.set_yticklabels(headers)
plt.show()
```



From the above correlation graph there is more correlation between Sp.Atk and total. So, I am removing "total" from attributes

```
pd.plotting.scatter_matrix(df)
plt.show()
```



From Above scatter plot there is a straight linear increasing strip for Attack-Total, and Attack-Defense. So, I am removing total, because attack has gaussian curve but total doesn't, again and removing Defense because it has right skewed and attack has gaussian curve

So, After dimensionality reduction I have following quantitative attributes

HP

Attack

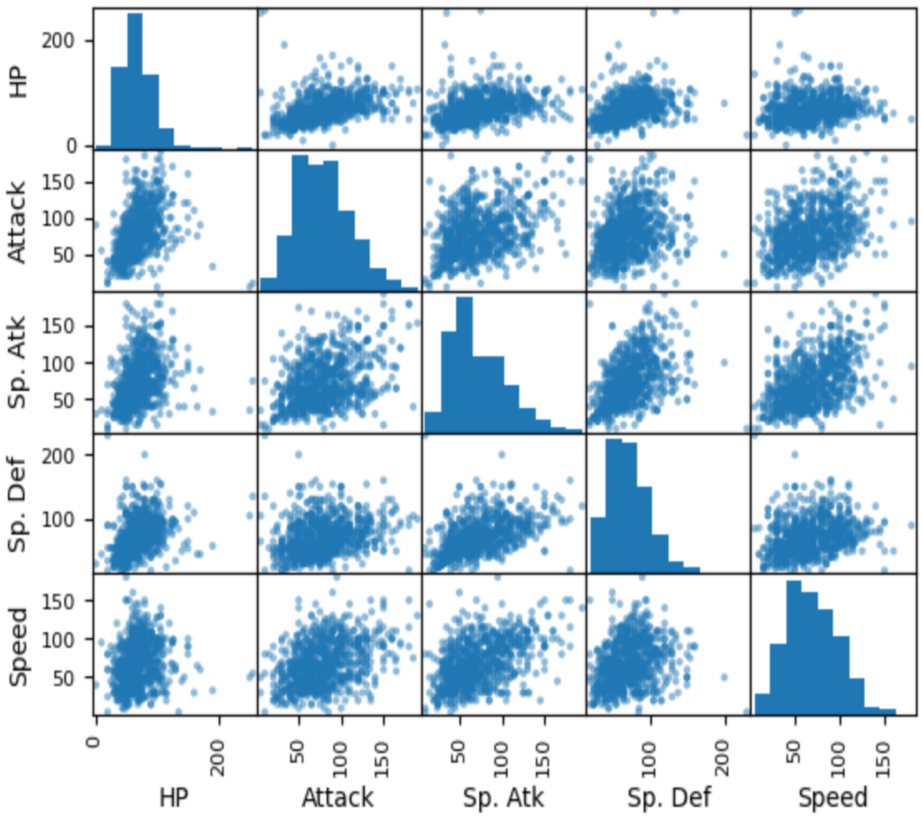
Sp.Atk

Sp.Def

Speed

These features are independent and Some features are randomly distributed; these we can find in the density plot of features because they have gaussian curves. So, output will be accurate.

```
k=df.drop(labels=["Total","Defense"],axis=1)
pd.plotting.scatter_matrix(k)
plt.show()
```



Again there is Sp.Atk-Sp.Def, which has an increasing strip and gaussian at Sp.Def and right skewed at Sp.Attack. So, I am removing Sp.Atk

So, After dimensionality reduction I have following quantitative attributes

HP

Attack

Speed

Sp.Def

These features are independent and Some features are randomly distributed; these we can find in the density plot of features because they have gaussian curves. so, output will be accurate. They were ordered based on their curve and random distribution.