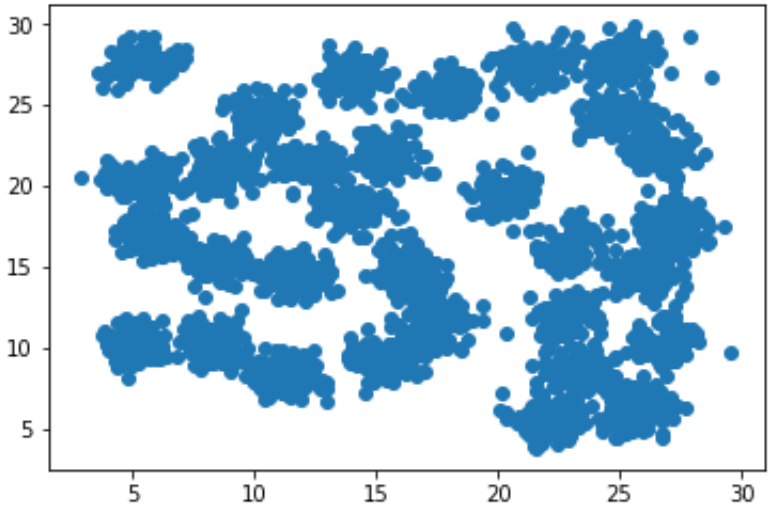
In [1]:
```
import pandas as pd
synth_all = pd.read_csv("synth_all.csv")
```

Question 1a

In [3]:
```
import matplotlib.pyplot as plot

plot.scatter(x=synth_all.x1, y=synth_all.x2)
plot.show()
```



Question 1b

I would expect about 30 clusters

Question 2a

In [4]:
```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=30).fit(synth_all)
```

Question 2b

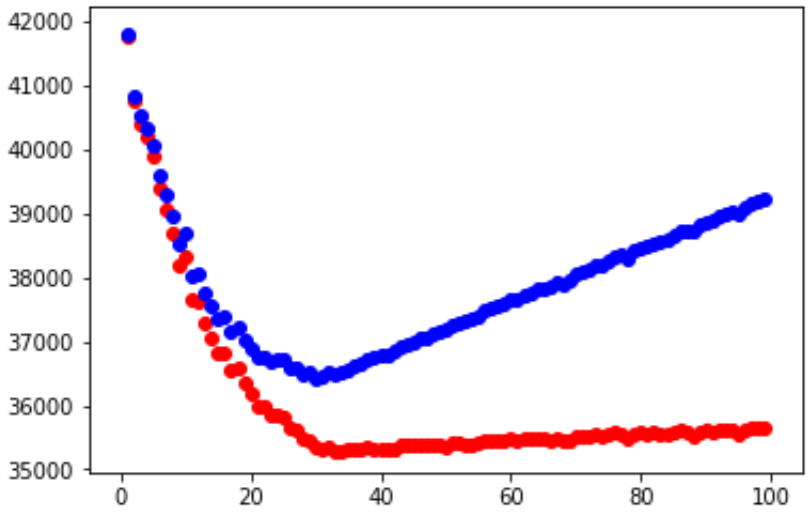
In [5]:
```
from sklearn.mixture import GaussianMixture

n_clusters = range(1,100)
aic = []
bic = []

for n in n_clusters:
    gm = GaussianMixture(n_components=n, init_params="kmeans").fit(syn
th_all)

    aic.append(gm.aic(synth_all))
    bic.append(gm.bic(synth_all))

plot.scatter(n_clusters, aic, c='r')
plot.scatter(n_clusters, bic, c='b')
plot.show()
```



Question 2c

In [6]:
```
min_aic = min(aic)
min_aic_n = n_clusters[aic.index(min_aic)]
print 'n clusters for min aic: %i'%(min_aic_n)

min_bic = min(bic)
min_bic_n = n_clusters[bic.index(min_bic)]
print 'n clusters for min bic: %i'%(min_bic_n)
```

```
n clusters for min aic: 34
n clusters for min bic: 30
```

Best number of clusters is 32 for both AIC and BIC. My visual inspection said 30. They're about the same.