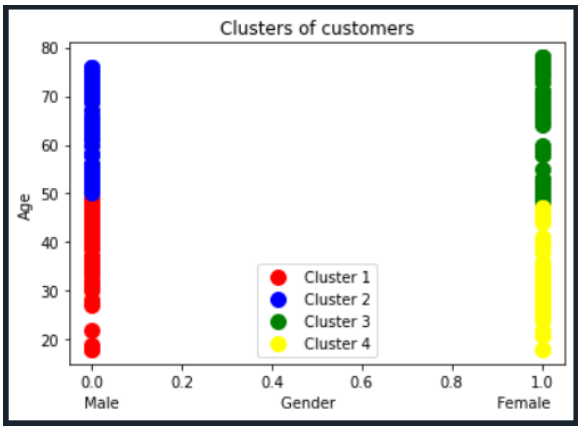
**Student:** Xiaomin Wu
**Week project:** week 38, Clustering

**Data preparation for the training:**
**Dataset creation:**
First, replace all semicolons (;) to colons (,) as I did last time, to make sure separate my data column by column. I also change the spell of sex to 'Sex_1=Female_0=Male', since there is some error when I use original format.
Since we use unsupervised machine learning for clustering, there is no training or test set. But we need find the 'elbow' of data, I tried a lot of times, but hard to see an elbow in graphs. After I check the recording (I have courses overlap, so I check recording when I have trouble with homework), I realized I shouldn't make customer ID in the clusters. Anyway, after remove that, that's clearly see the elbow is 4.
**Scaling:** I was tried to only scale age because I want to keep gender as 1/0 format. But once I only scale age, there a lot of error jump out. Google told me the reason is data consists of both integers and floats, but array can only have one type.
So I use StandardScaler() to scale all my X values.
**No dummy variables** in this case, because there already 1/0 format for gender.

**Relevant metrics for the case:**




```
In [40]: runfile('//wsl$/Ubuntu/home/annawu/ai-ml/week38/week38hmw.py', wdir='//wsl$/
Ubuntu/home/annawu/ai-ml/week38')
Crosstab:
 Average monthly purchase  5     8     22    26    ...  1169  1172  1175  1176
predicted_AMP                                      ...
0                          0     1     0     0  ...     0     0     0     0
1                          0     1     0     1  ...     1     0     0     1
2                          1     0     0     0  ...     0     1     1     0
3                          0     0     1     1  ...     0     0     0     0

[4 rows x 182 columns]

Average monthly purchase:
 1093    3
88      2
1011    2
1027    2
548     2
       ..
1024    1
56      1
346     1
1169    1
1175    1
Name: Average monthly purchase, Length: 182, dtype: int64

Predicted AMP:
 2    53
3    51
1    48
0    48
Name: predicted_AMP, dtype: int64
```

**Conclusions of the results:**
It's a perfect result for clustering, I guess it because there is no such confused information, just gender and age, so it should be perfect (maybe I should think it beforehand, so I can save time with elbow part).
But the problem is the predicted_AMP (Average monthly purchase) in the picture in too small than the value it should be when I see though the database. I guess it's because there are some values too unusual to use. Maybe next time we can delete those values real abnormal.