**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.

**Text

Description automatically generated**

**Relevant metrics for the case:**

**Chart, scatter chart

Description automatically generated**

**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.