Notes from K means clustering model:

* Can use k means clustering for customer segmentation
* Need to scale your variables
* Plotting original scatter. With three colors for groups (?)
* Plotting k means classification next to it
* So the shape of the plots actually looks quite similar, which is a good sign, shows good predictive power
* If the model groups correctly but labels wrong, you can use relabel. In our example, however, this is unnecessary.
* ML: want high precision (relevance) & high recall (completeness).

Notes from Hierarchical clustering model:

* Again, taking a subsample (5000 out of 38,000)
* Clustering based on distance bw each data point and nearest neighbor. Linking up nearest neighbors.
* Dendrogram: tree diagram displaying hierarchical clustering
* Check how many times the cluster intersect with the line
* Distance metrics vs. linkage parameters
* ‘ward’ is the linkage method we use here
* Linkage generates our clustering results. Pass that to dendrogram
* We’re using 2 clusters here
* affinity='euclidean', linkage='ward' has the lowest accuracy of 55%
  + Note ‘ward’ can only work with Euclidean distances
* Changing linkage to ‘average’ increases accuracy score to 71%

Notes from DBSCAN (pronounced deb-scan) clustering model:

* Dbscan for outlier detection: lanes vs. lines
* Density based clustering
* Main parameters: eps, and min\_samples
* Eps = 0.8 => a max distance of .8 between points to still be considered neighbors
* Outliers should not constitute more than 5% of your dataset
* My model gives me a very high number of outliers, almost 30%
* Records with a label of -1 are considered outliers