

Figure: first trial on clustering and plotting. Result: normal family ID as black dots continuous throughout the range. But for family member ID, outliers occur. Next step: cluster family member ID alone.

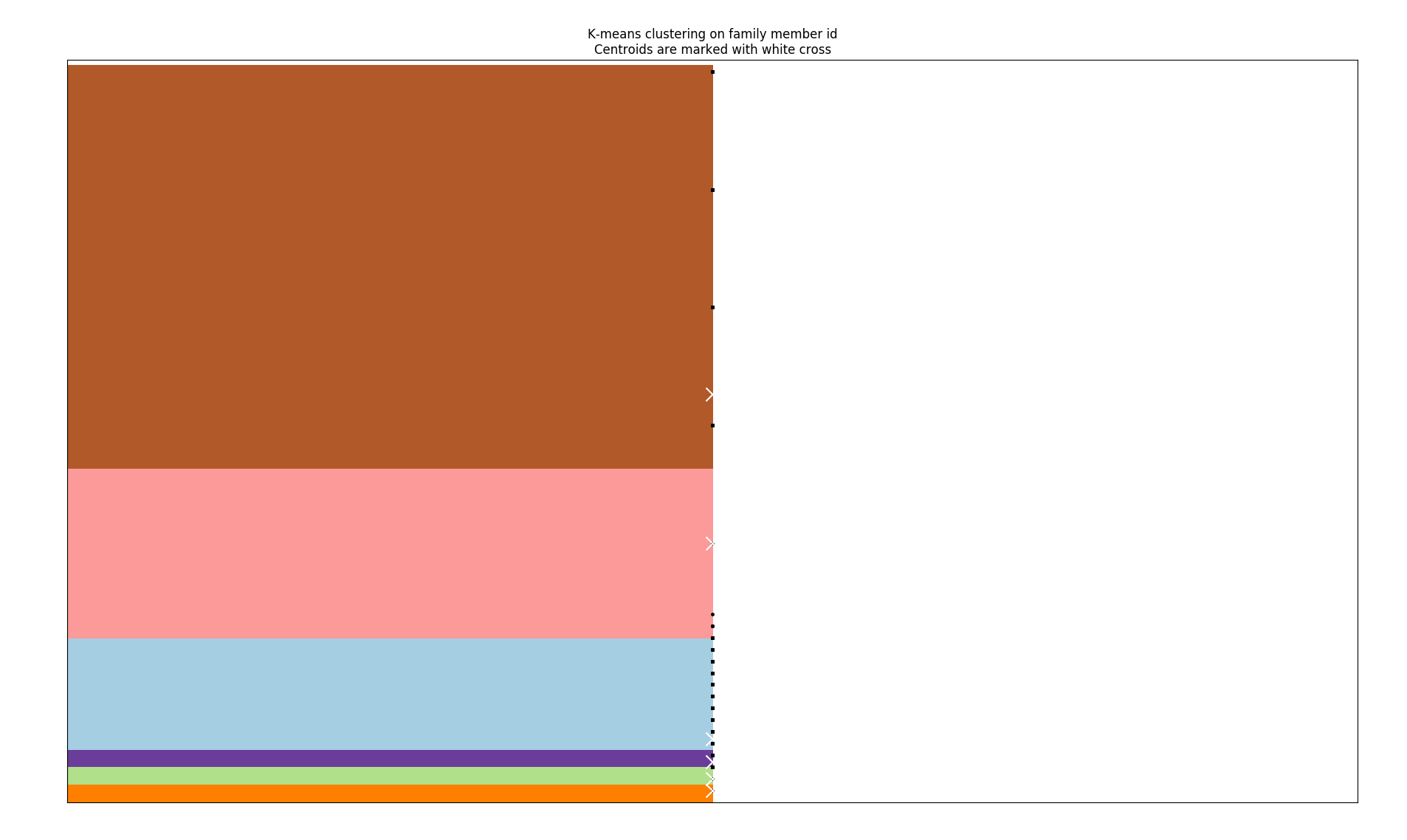


Figure: second trial, number of clusters = 6. How to find outliers? N = 2? Or more?

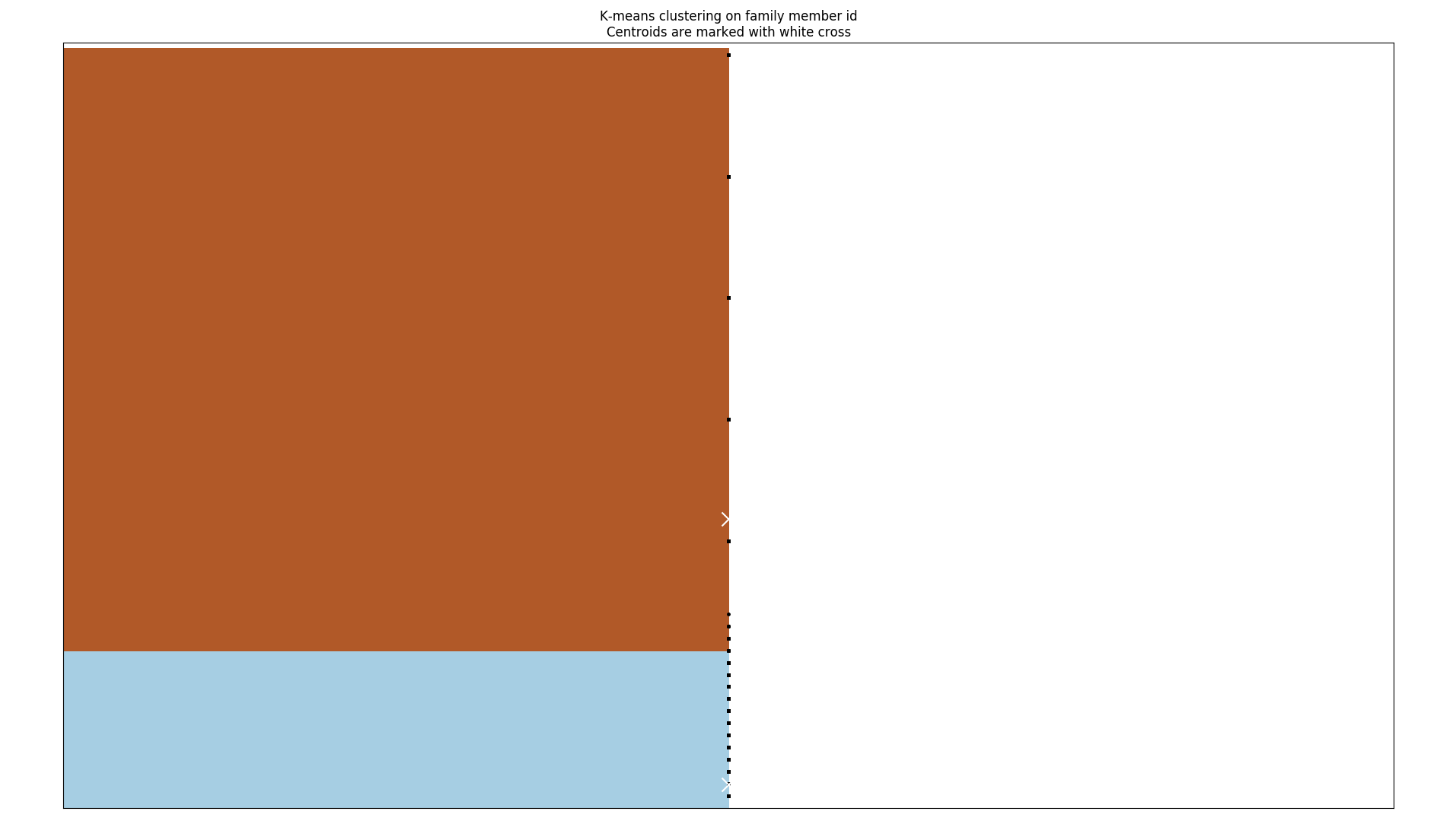


Figure: cluster number = 2. Close enough.

“In one dimensional data, don't use cluster analysis. Cluster analysis is usually a multivariate technique.”

Note: maybe use less judgement, more rely on “unsupervised” learning result.

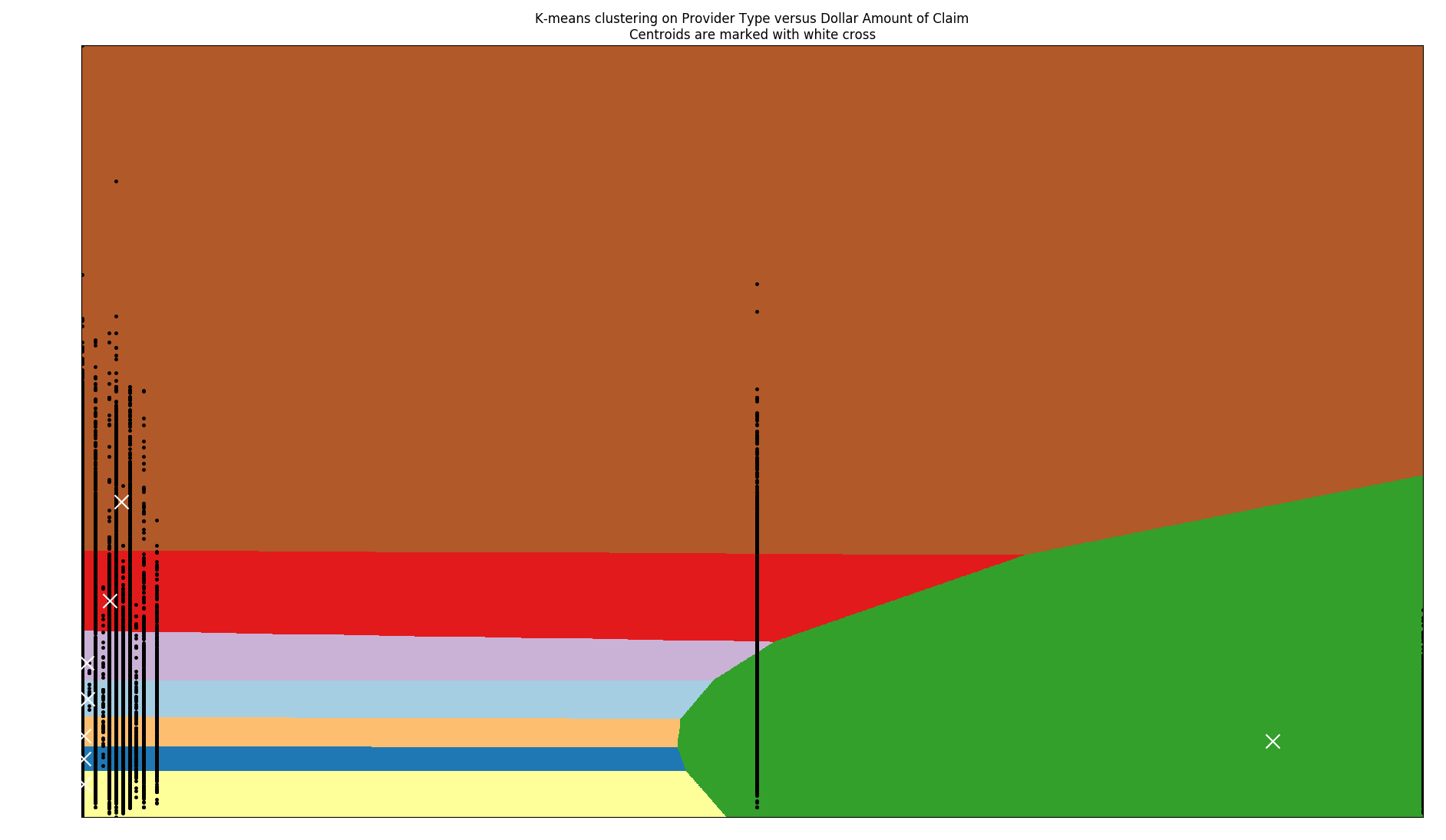


Figure: n\_clusters = 8

But right now, I’m observing data patterns by eyes…. Not machine learning…? But the centroids are computed by ML.

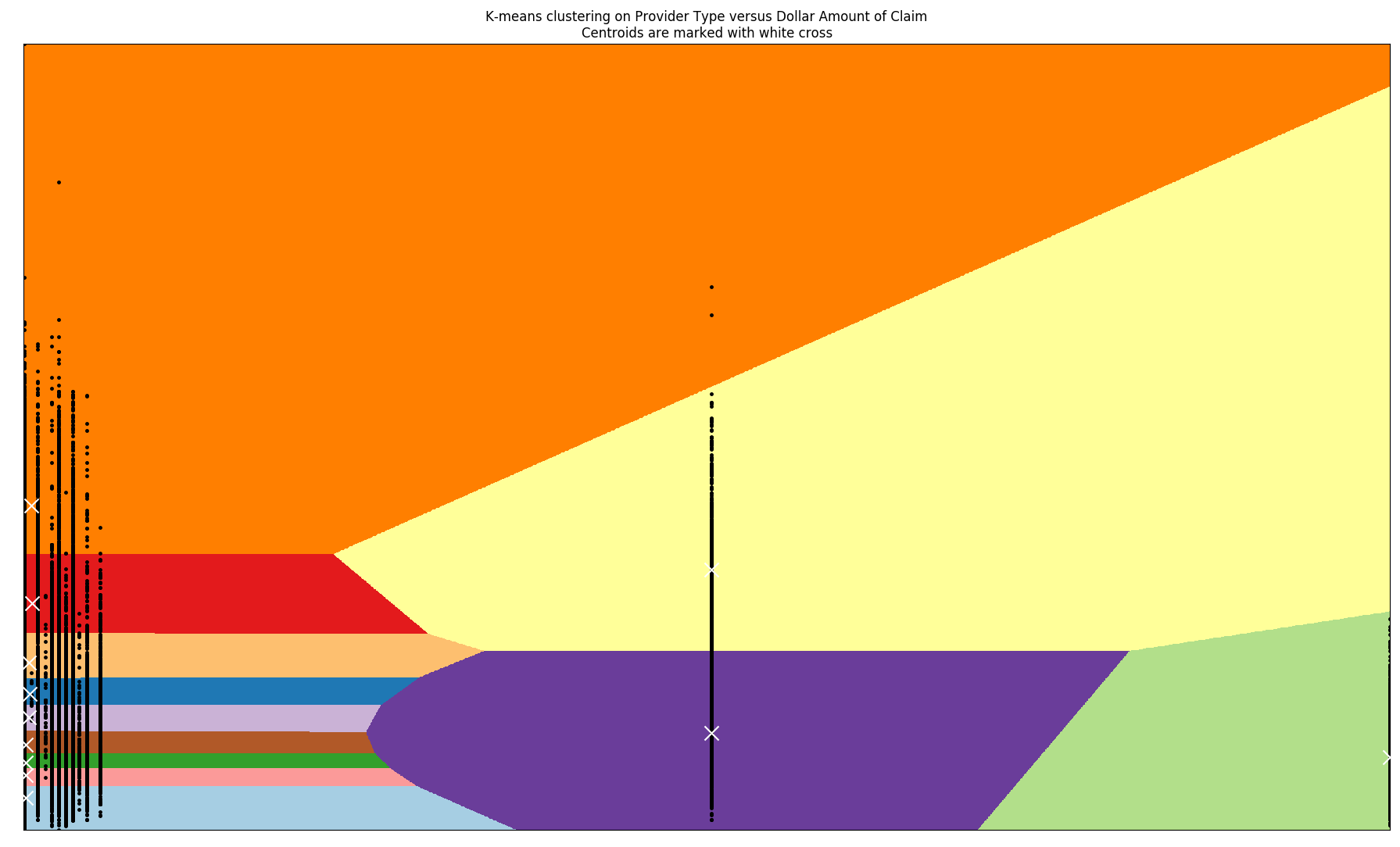
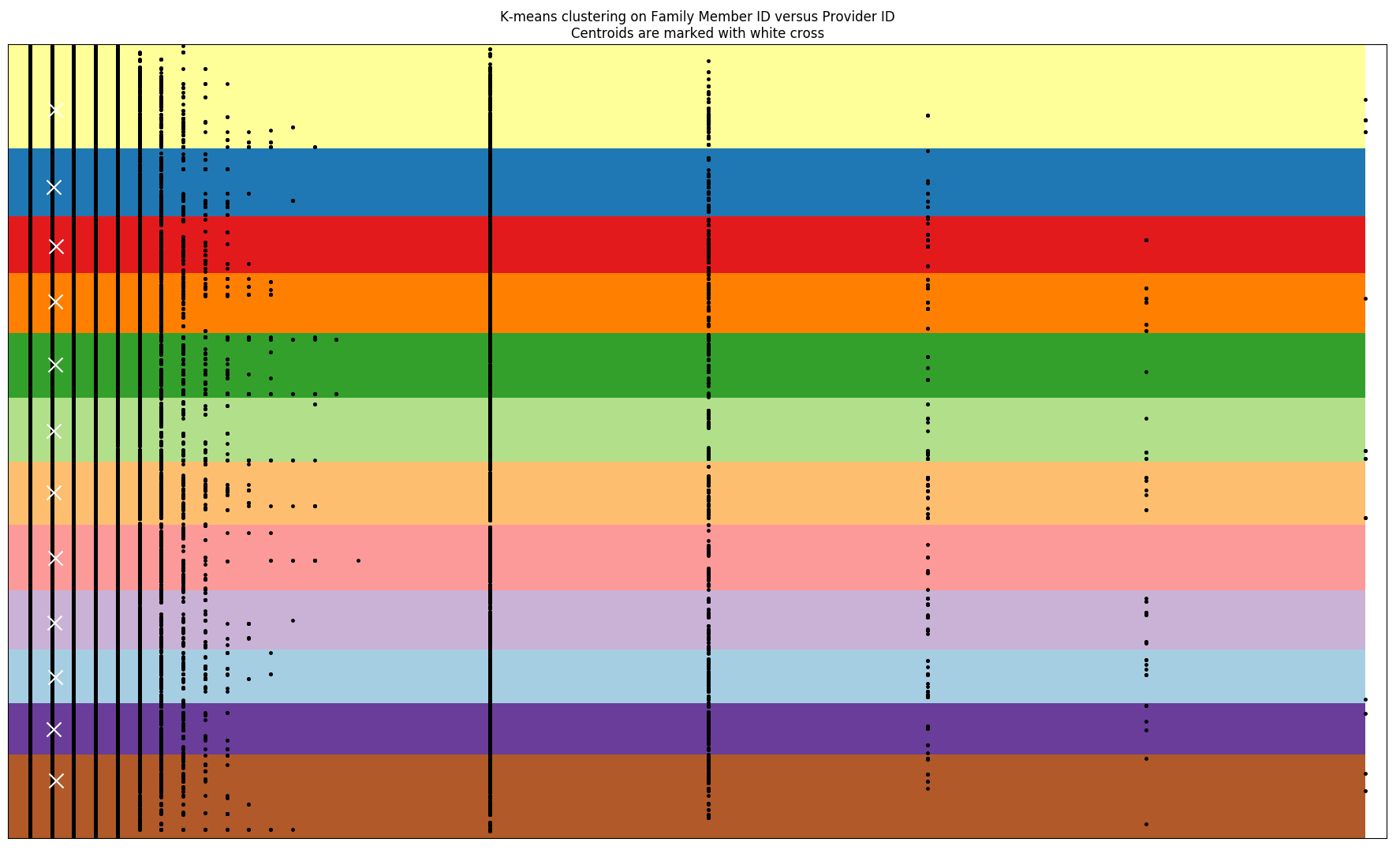


Figure: n\_cluster = 12

Result: no correlation between Provider Type and $ of claim.

Try: - doctor ID vs. family ID. - Medical Procedure Code vs. $ of claim



Result: because they have different range, centroids don’t reflect on actual grouping. … prefer spread in one dimension over the other.

* Scale the matrix down correctly!

Try Now: - doctor ID vs. family ID. - Medical Procedure Code vs. $ of claim

(need to change x\_ y\_ steps too.)

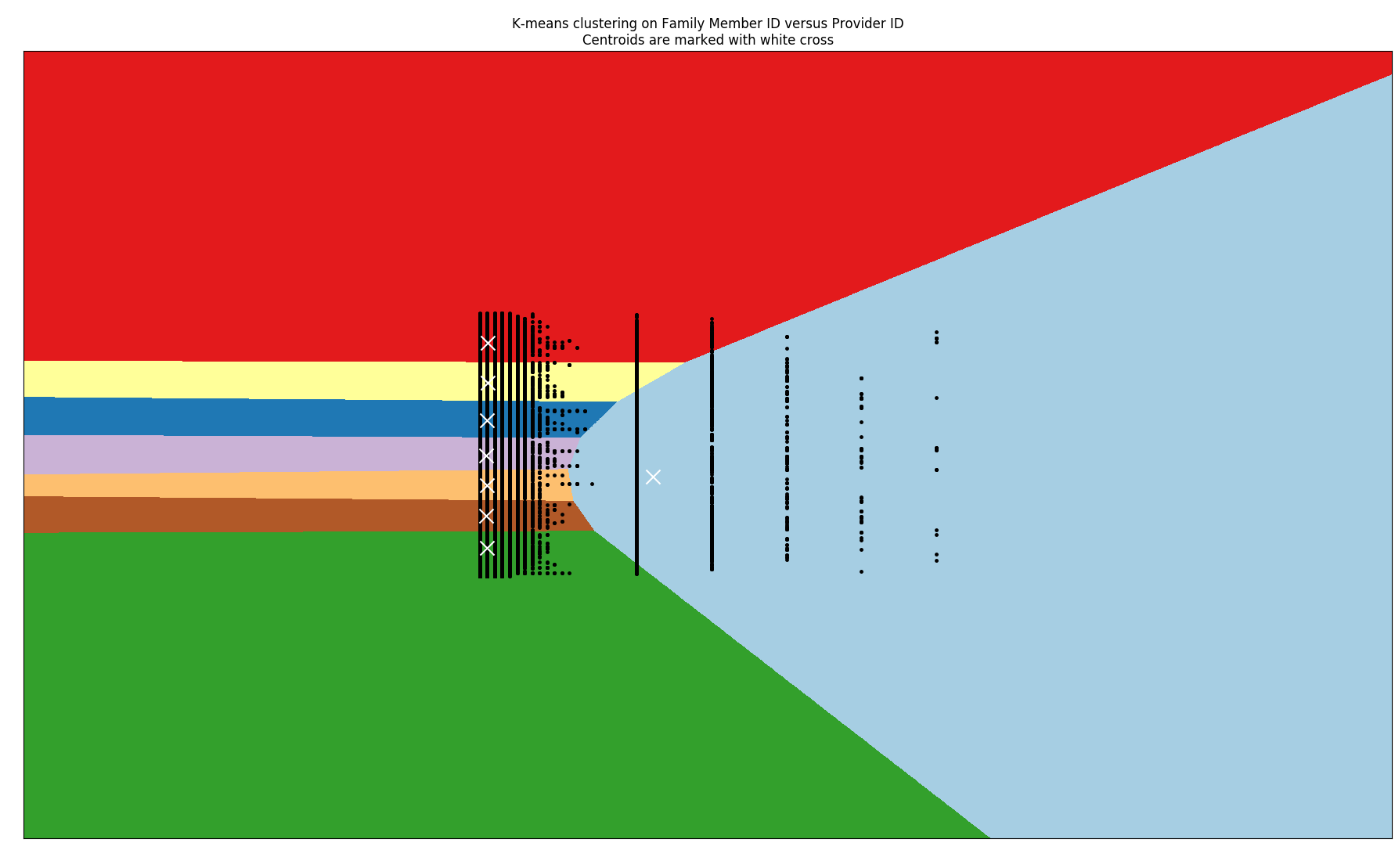
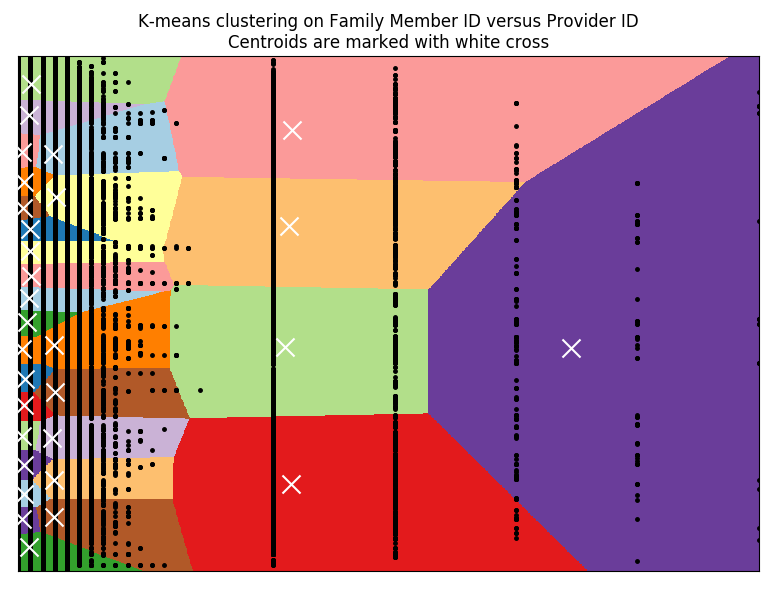


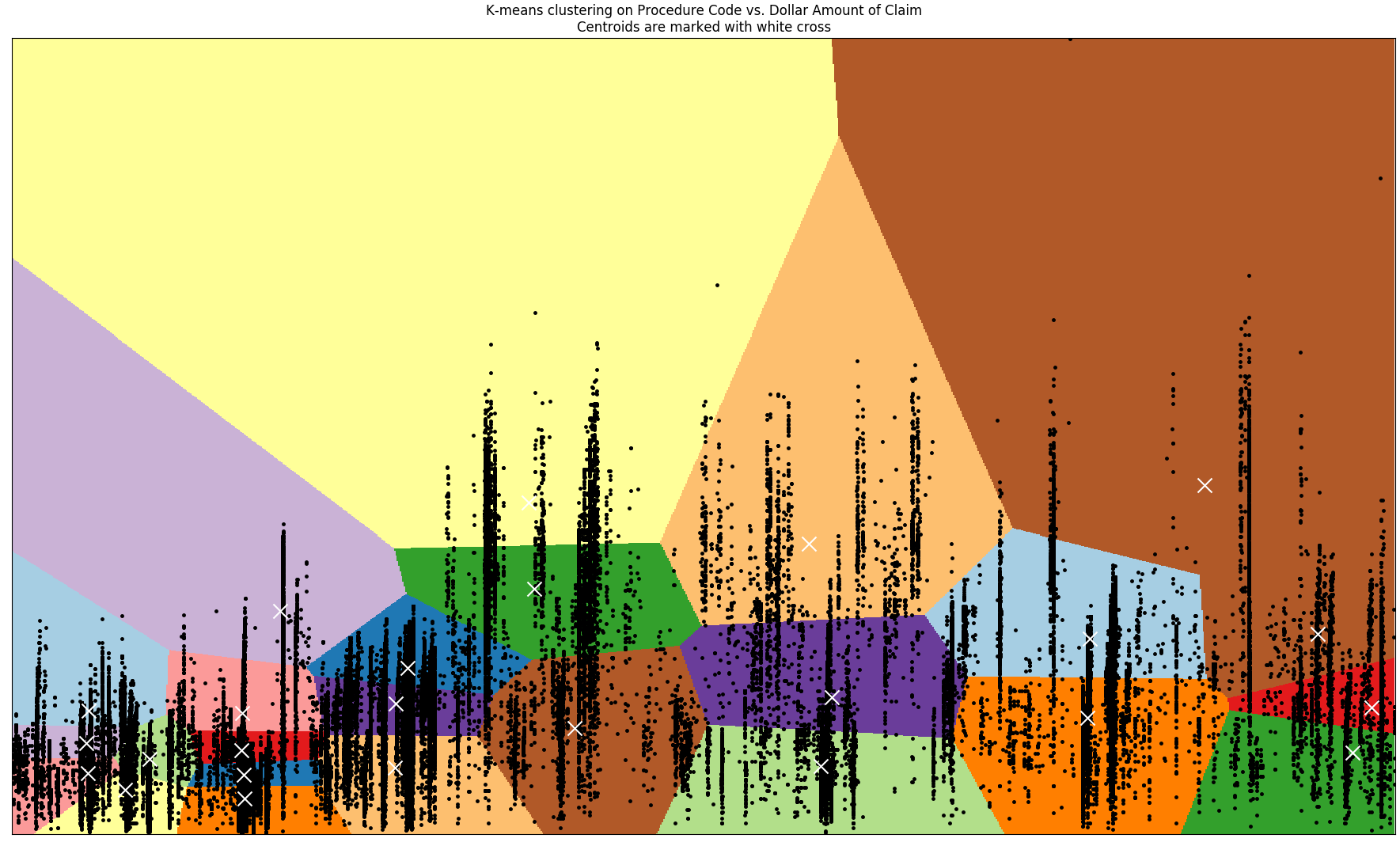
Figure: not much results… n\_cluster = 8 … what do I do? Try much more centroids.

N\_clusters = 30. Takes 10+ secs to run.



Not much difference….. what’s wrong?

May be the features chosen are wrong? Try Medical Procedure Code vs. $ of claim



Hmmmm…. Search how to use unspuervised learning, clustering to find outliers.

After researching: alternative solutions:

* DBSCAN
* Birch
* Use train and test method for kmean?

+ these are for 1D data. Any pairs you want?

Next step: try DBSCAN then think about the problem statement again.

Try to think of a way that lets the machine tells you the answer.