Great work everyone so far, I wrote up some feedback from Dr . Drew.

Do we need to predict Crime Solved?

It is not required to predict but we spent a lot of our write up in Business Understanding 1 planning our Crime Solved classification analysis.

If we choose not to predict Crime Solved, We would probably have to change some write up in Business Understanding 1 because we initially planned to use our clustering to predict Crime Solved.

So if we change up some write up and Visualizations that pertains to Crime Solved, we can avoid predicting.

|  |  |
| --- | --- |
| Section | Feedback |
| Business Understanding 1 | We should emphasize that we are using our cluster labels as a means to classify our Crime Solved more accurately.  We should say that we are using K Folds because we want to see that our clustering silhouette remains consistent through each fold. |
| Data Understanding 2 | -Take the PCA variables with most significant loadings and do EDA,  -And add Clustering specific write up ‘The silhouette for these 2 variables will be very good because they are so separated etc.’  -Move PCA to this section  -Choose top 5 attributes, do EDA on those top 5 attributes , make scatters of PCA eigenvectors |
| ME1 | Because we are classifying crime solved, we should set n clusters to 2 because it can only be yes or no crime solved. N clusters 2 for all algorithms and we should vary the other parameters like min eps or affinity etc.  For each algorithm we must parameter tune more variables at least, 2 variables, like min points or eps,  Random state is not considered as a tuning variable  We need to add other variables to tune for kmeans besides n clusters and other algorithms too, we must keep random state consistent |
| ME3 | We need to add the DB Scan cluster labels as a column to our dataframe.   |  |  |  | | --- | --- | --- | | Victim Age | Victim Race | Cluster Label | | 20 | Asian | 1 | | 30 | Black | 2 |   We need to compare cluster 1 to cluster 2 etc.  We need to do cluster1=df.loc[df[‘Cluster’]==1]  Cluster2=Df.loc[df[‘Cluster’]==2]  Then we need to compare cluster1 and cluster2 basedon the other variables:  do .describe() on each cluster and say ‘cluster 1 is where there were mostly black victims’ ‘cluster 2 is where mostly white victims’ etc. ‘Cluster 1 is older than cluster 2’ |
| ME4 | We need to have a very wordy and length summary of each modeling and evaluation section and if we are happy with our results. |
| Exceptional WOrk | Because we discussed predicting Crime Solved in our Business and Data Understanding, I think we need to create model using our labels and k folds cv to predict Crime Solved |