1) Rationale for inputs used

For the clustering model, I will be applying the K-Means node on the crime dataset for analysis. During the initial import of the dataset, duplicate fields, such as Ward, Beat, FBI Code, X/Y Coordinates, Longitude, Latitude, and Location were filtered out. Additionally, irrelevant variables that would not contribute any value to the model, such as Year and Case Number, were filtered out at this point. The ID was designated as Record ID and District as Target. Location and time are known to be particularly effective as input variables for predicting crime. Arrest and Domestic variables provide valuable insight into the nature of the crime reported. For example, arrest rates are often used to determine effectiveness of police.

The Time Intervals, Weekday, and Day (of the Month) variables were parsed and derived into individual variables. Both Location Description and District Regions were consolidated into a fewer number of fields to be included for analysis. The final available inputs were IUCR, Crime Type, Crime SubType, Location Description, Arrest, Domestic, Time Intervals, Weekday, and Day.

2) Explanation of cluster output - Technical

Running an initial analysis produced a silhouette of fair quality at 0.4 with five clusters. The predictor importance chart revealed that Weekday and Time Interval variables contributed little to the model. After dropping Weekday and Time Intervals, Day also proved to lower the model and was eliminated as an input as well. To achieve an optimal level of cluster quality, the final input variables selected by the model were Location Description, Arrest, and Domestic with four clusters.

The revised model produced a good cluster quality of 0.8, with the dominating cluster containing 60.4% of data points and the smallest cluster containing 3.0%. The two remaining clusters contained 25.1% and 11.5% of data points. Decreasing the number of clusters to three did not impact the model, and the four clusters were well-defined by the inputs from a business standpoint. All three variables scored 1.0 on the predictor importance scale and contributed equally to the model.

With a high cluster quality and strong predictor performance, these clusters showed promise for enhancing the predictive model. Running the feature node ranked the Clusters variable as Important at 1.0, further supporting the variable's predictive potential. The original CHAID model without the Clusters variable produced a result of 28.83% correct predictions. Running the CHAID model with the Clusters variable while omitting Location Description, Arrest, and Domestic variables produced a result of 28.66% correct predictions. Boosting the variable Districts actually reduced the percent correct to 25.6% and, therefore, was not included in the model. In light of this outcome, while the Clusters variable may have positively contributed to the model, it was not enough to outperform the original CHAID model.

3) Explanation of cluster output – Business User

The four generated clusters can be defined as Thefts in public spaces and residences, Narcotics in District 4, and Domestic violence with and without a resulting arrest. Thefts in public spaces and at residences were the dominant cluster with 60.4% of data points and characterized as non-domestic crimes with no arrest made. One-third of incidents were thefts, followed by criminal damage at 14% and battery at 12%. Over 50% of incidents in this cluster occurred either in public space (34.5%) or at a residence (20.9%). Over half of these incidents occurred between 5 pm and 3 am. District Region 6 experienced the highest number of incidents at 22.3% and District Region 1 experienced the lowest number of incidents at 9.9%; however, this distribution may be a function of population distribution.

The next cluster, Narcotics in District 4, contained 25.1% of data points characterized as non-domestic reported crimes resulting in arrest. 42% of incidents in this cluster were related to narcotics and 12% were thefts. Over 50% incidents occurred in a public space and 75% of these incidents occurred between noon and midnight. District Region 4 experienced the highest number of incidents with respect to this cluster at 27%, followed by District 1 at 5.4%. Interestingly, District Region 4 may have a more concentrated drug problem than in other areas due to its unusually high number of incidents compared with other clusters.

Containing 11.5% of data points, the third cluster is characterized by domestic crimes that did not result in arrest. These incidents were primarily battery and assault, which is not surprising since domestic crimes are related to family violence. Accordingly, 76.6% of these incidents occurred at either residences or apartments. Incidents were concentrated primarily in District Regions 6, 5, and 4. Again, this number does not take into account population estimates.

The last cluster, containing 3.0% of data points, are domestic resulting in arrest. Similar to the previous cluster, this cluster is comprised primarily of assault and battery and over 80% of these incidents occurred at either residences or apartments. 35.9% of incidents occurred in District 6, followed by District Regions 4 (21.2%) and 2 (18.3%).

4) Cluster use decisions

While the predictive model is not improved by the clusters, they still shed light on certain types of crimes being committed and its impact on communities. Thefts are occurring both in public spaces and at residences with a low number of arrests being made. Communities may benefit from strengthening local programs such as neighborhood watch and at-risk youth outreach programs. Because most incidents occurred in the evening and at night, police districts may consider increasing staff during these times.

Districts in District Region 4 appear to be experiencing a high number of narcotics-related crimes indicating a drug problem; however, they do appear to making headway in arresting perpetrators. These districts may want to concentrate their efforts on drug

abuse prevention and enlist the help of local drug enforcement agencies. Additionally, community efforts to address the socioeconomic impact of drug abuse and addiction, such as drug addiction rehabilitation centers and job training programs, may serve to improve communities in these districts as well. A majority of these crimes occur between noon and midnight, and police districts may consider increasing staff accordingly.

One of the main reasons to separate out domestic violence cases where arrests occur compared with no arrests is a powerful indicator that victims need help. 80% of domestic violence incidents reported did not result in arrest. Assistance in the form of safe houses, economic initiatives for victims, and mental health support services, such as anger management classes and counseling, may help incentivize victims to hold perpetrators accountable and provide rehabilitation for perpetrators to better cope with emotional situations.