CS 432 Group 4 Deliverable 3

Talha Ahmed, Maryam Shakeel, Muhammad Haris Saad, Taha Sarwat

April 2023

1 Interestingness Measures

In this report, we see which drugs are most commonly used in pairs. We will most likely be seeing the joint occurrences of the drugs along with their itemsets and then applying different measures to support our arguments.

To start off with, we first calculated the frequent itemsets of the drugs for which we used the fp-growth algorithm to generate, and kept a certain threshold (in this case it is 0.001). Those which had a minimum support of greater than this threshold, we referred to them as frequent. We then encountered different groups of itemsets like cocaine and heroin, heroin and fentanyl, and so on. For further analysis, we drew the line-plots of k drugs sets against the support of each of the three common drugs. What we saw here was that as n increases, there is a decrease in the support count of the drugs. We used an interestingness measure called confidence here for further conclusions. Based on the analysis of the support graph, it is observed that Fentanyl has the highest support, but occurs significantly less frequently with other itemsets. In contrast, Heroin and Cocaine have a lower support compared to Fentanyl, but are more consistent in their combination with higher itemsets. This indicates that Fentanyl is more likely to be consumed alone and has faced a sharp decrease in support, while Heroin and Cocaine are more likely to appear with increased k and have a more consistent combination with other itemsets. Therefore, Fentanyl is less likely to affect the joint occurrence of itemsets compared to Heroin and Cocaine.

We individually checked for drug consumption in various parameters e.g., sex. After drawing the line-plots of support of drugs by sex against the n itemsets, we observed that for support the graphs are more or less the same. All the drug support counts drastically decrease as the number of itemsets increase. We can further conclude that Fentanyl is more consumed by males than females, with it starting above 0.6 support in males. If we look at the confidence level for both sexes against the drug counts, we see that the joint occurrences have the same effect on both sexes with the graphs being almost the same. In addition, it can be seen as a reflection for the plots of the support and confidence of the sexes. A reversal in the order of drugs can be observed when comparing their

support plot with their confidence plot. Specifically, the drug with the highest support, for example Fentanyl, has the lowest confidence in the confidence plot, and vice versa (it follows the same inverse trend for the other drugs too). This mirror image phenomenon is evident when examining the relationship between the support and confidence measures for the drugs.

Fentanyl's trajectory is similar to the other drugs but with some noticeable shifts. Notably, Fentanyl has the highest support in the "Black or African American" category with a greatest 0.8 support, followed by around 0.75 for the "Asian Indian" category. Conversely, there is uncertainty in the trajectory of support for cocaine and heroin across the four categories, with heroin reaching its lowest in the "Black or African American" category, and its highest in the "White" category. Interestingly, where there is high support for Fentanyl, there is lower support for cocaine and heroin. This suggests an inverse relationship between Fentanyl and Heroin consumption, particularly evident in the "Black or African American" category where Fentanyl consumption peaks and Heroin consumption is at its lowest frequency. Overall, the support plot provides further insight into the relationships and patterns of drug consumption among different categories. We see the race "Black or African American" having the highest confidence of the three drugs.

Now we will use another interestingness measure called Lift. This will help us in determining the correlation between the drugs and their n itemsets with the Lift score putting an automated threshold of approximately 1. We will be checking the effect one side of a drug has on its itemsets and so on. The analysis of the lift and correlation scores between the drug attributes has revealed some surprising results. Many attributes are not having a zero lift with each other, but rather a mixture of negatively and positively affecting the other drug. This implies that low or high correlation should not be attributed to correspondingly low or high lift scores. It is possible to have a situation where two variables have a low or zero correlation, but still have a strong association in terms of their co-occurrence, which is what the Lift measures. On the other hand, correlation measures the linear relationship between two variables.

Therefore, it is important to consider both the correlation and Lift measures together when analyzing associations between variables. A low correlation does not necessarily imply a low association, and a high correlation does not necessarily imply a high association in terms of cooccurrence or joint distribution. Taking into account these caveats, it can be concluded that the low correlation implies that the attributes are not related to each other in the intrinsic sense. However, they may have a higher lift score, which specifies that both are jointly consumed. For example, Methadone and Oxymorphone have a correlation of approximately 0. Their Lift score is $6.6 < 1$, showing that Methadone negatively affects its joint occurrence with Oxymorphone.

Conversely, there are a few small noticeable results, such as the Lift score

for FentanylAnalogue and Heroin, which is equal to 1. This shows that both are independent of each other and hence uncorrelated. Here, one can safely imply one from the other, but not the converse. This can also be seen in the correlation heatmap, where their correlation is nearly 0. In summary, the Lift and Correlation scores provide complementary information, and both measures should be considered when analyzing the associations between the drug attributes.

In this analysis, we focus on the categories "Cause of Death" and "Manner of Death", specifically the subcategories "AFI, MDT, HI, AHI" which appeared to have the most impact. Our previous observations showed that AHI had a strong association with heroin, while AFI was associated with fentanyl and to a lesser extent, cocaine. However, for MDT, all three drugs had a significant presence in joint occurrences, with heroin having the highest lasting influence. In terms of support, MDT and AFI stood out more than the others, with AFI having a higher support but shorter longevity compared to MDT. For small n, AFI appeared to be the main cause, but for larger n, MDT was the only viable option. Consequently, the confidence for MDT was higher than AFI for most k values.

We also looked at how the classic trio of "Fentanyl, Cocaine, Heroin" behaved under the new categories. For AHI, heroin was the main drug at play, while for AFI, fentanyl had a significant presence and cocaine had very little. However, the confidence plot for AFI was empty, possibly because AFI was mainly correlated with fentanyl and we were only looking at single itemsets. For MDT, all three drugs had a strong presence, with heroin having the highest lasting influence. The support for cocaine and heroin was higher than fentanyl, but there was a slight increase in fentanyl's support. Looking at the confidence plot, we saw that MDT had a high ability to affect the joint occurrence of fentanyl with cocaine and heroin, indicating the necessary effect of heroin and cocaine with MDT. In general, $conf(x \rightarrow Heroin/Cocaine)$ was likely to be highest if the person had MDT, with heroin being the most influential drug, followed by cocaine and then fentanyl.

To perform support/confidence analysis on the dataset based on the 'Residence City' attribute, we need to filter the list of drugs and cities since there are over 90 categories for each, which could lead to the curse of dimensionality. We will use two different approaches to tackle this issue. Firstly, we will filter the dataset by selecting only those drugs with a correlation greater than 0.2 with the 'count' variable, and only those cities that appear in the top 10 of the bar graph created earlier. Although this may omit many cities, it aligns with our goal of identifying the root cause of such deaths, the most common drugs associated with them, and consequently, the vulnerable cities. Secondly, we will carry out cluster analysis on the features of the cause of death and cities to identify the prominent cluster associated with these cases. From our analysis in D1, we had observed a cluster of cities in the center of Connecticut that had the most instances of MDT and AFT deaths. If we can prove that

this cluster is indeed prominent for such cases, our approach in the first step will be justified. We observed that Hartford had a significant support and high confidence, as seen from the bar graph and the confidence plot. However, we also noticed the importance of 'longevity,' meaning that some cities, such as Waterbury, New Haven, and Bristol, extend their influence to 4 or even 5 itemsets, indicating their higher impact. We expect such cities with high supports and great longevity to be prominent in the main cluster. After analyzing the group-wise counts of drug consumption by cities, we noticed that Danbury and Meriden did not have any drug consumption, while Hartford, New Haven, Waterbury, and Bristol had variations in the support of many drugs, with higher supports for Fentanyl for the latter three cities. Bridgeport and New Britain had fewer drugs consumed, but Fentanyl's support was higher than the rest. To further evaluate these observations, we will utilize the geo-spatial visualization we created in D1. For New Haven, Bristol, Hartford, and Waterbury, we observed at least 10 cases of MDT and at least 5 cases of AFI. This suggests that places where the support is high have a higher chance of MDT. Interestingly, Bristol had the lowest Fentanyl support among these cities but still had 12 cases of MDT, indicating that the number of drugs might be more critical than the quantity of each drug. Norwich and New London had barely 5 cases of MDT, reflecting their low support and n. For Bridgeport and New Britain, we observed surprisingly high cases of MDT (at least 10) even though most drugs were not consumed.

Arguably, the most unique anomaly is that of Meriden. Its plot, though empty, depicting little to no consumption of drugs, shows 6 cases of MDT and 4 cases of AFI. This may well be due to the fact that we had filtered many of the drugs as per our approach and hence lost valuable information. That was expected and hence for now, we will consider it as an outlier. Later when we do cluster analysis, hopefully the picture becomes clear.

2 Cluster Analysis

After conducting the aforementioned analysis, we need to verify if our filter was a wise decision by checking if the formed clusters align with our initial observations. If they do, then our choice was sensible, otherwise, we must accept it as it is due to our limitations.

However, there is a caveat. To cluster cities based on their cause of death similarity, a clustering algorithm like DBSCAN can be used. However, with only two unique causes of death, clustering may not be the best approach. Thus, we will use the original dataframe with more cities since the intrinsic clusters should remain the same theoretically if the similarity lies there. Additionally, we will utilize the latitude and longitude we previously used in D1 for cluster visualization. Note that we will ignore cities from the same cluster but far apart as per

our assumption. We observe dense clusters being formed in the center (A-pink color) and the southwest, which confirms our previous assumptions. The densest cluster is formed in the center, and we are confident in our claim in D1 that if a city is in cluster A, it likely has a high occurrence of deaths from MDT and AFI.

Before delving into the classic support/confidence-wise analysis of drugs/death types by year, let's revisit the simple graph we made in D1, displaying the number of deaths by year. From the support graph, we can infer that most deaths occurred in 2021, as the supp(2021 | Number of Deaths) > supp(!2021 | Number of Deaths). However, besides support, we are also interested in the longevity or its impact on joint occurring frequent sets, which we can observe as 2019, 2020, and 2021 having the highest. This suggests a time when drug consumption culture was at its peak. This can also be seen in the confidence plot, where 2019, 2020, and 2021 have the highest L.H.S confidence, indicating that the higher the drug occurrence, the higher the year (i.e., conf(Year $\rightarrow$ n) is greater for large n).

However, this is not the end. We have yet to see its influence on the consumption of the main drugs we have filtered so far, specified by retain_year variable. Once we have done that, we will be able to link our findings till now with 'Cause of Death', 'District' to check where, when and why did a particular drug was popular in consumption and correspondingly death_type by year which we have also determined to be the following from D1.

Most common cause of death each year: year
2012 (2012, Multiple Drug Toxicity)
2013 (2013, Heroin Intoxication)
2014 (2014, Heroin Intoxication)
2015 (2015, Acute Heroin Toxicity)
2016 (2016, Acute Fentanyl Intoxication)
2017 (2017, Acute Fentanyl Intoxication)
2018 (2018, Acute Fentanyl Intoxication)
2019 (2019, Acute Fentanyl Intoxication)
2020 (2020, Acute Fentanyl Intoxication)
2021 (2021, Acute Fentanyl Intoxication)

Analysis: Rather than analyzing each year individually, it is more sensible to extract patterns and meaningful observations from the data. Some interesting observations include:

- In 2012, there were multiple drugs with considerable longevity, and MDT was the most frequent death type that year. This confirms our previous finding that the confidence of Heroin/Cocaine causing MDT was high.
- In 2013 and 2014, there was a decrease in the support of cocaine, while heroin's support remained more or less the same. AHI and MDT were the most frequent death types those years, suggesting that AHI/MDT support decreases as cocaine support decreases.

- The popularity of Fentanyl can be observed to be growing as the years progress. In 2016, it overcame cocaine and almost heroin in terms of support. AFI was the most popular death type that year, indicating that Fentanyl has a significant influence.

- After 2016, Fentanyl's support increased and reached its peak in 2018 while all other drugs were at their lowest support. This clearly suggests that 2018 was the time when Fentanyl boomed in the market. Our district-level analysis implies that supp(Fentanyl — Year = 2018, City = Hartford (belongs to cluster A)) was the freshest moment and the beginning of a chain of deaths by AFI. We also observe a slight decrease in the total deaths that year. However, in 2016, n=3 and n=2 both decreased to their lowest, while n=4 increased. The n=4 is likely mostly contributed by Fentanyl since its longevity was considerably high. This is further supported by the fact that n=4 was still reaching its peak just when Fentanyl did. In conclusion, Fentanyl is one of the most common causes of AFI, and Hartford and any other city in cluster A is the center for such cases.

It is worth noting that the support and longevity of Fentanyl follows a "hill-shaped" distribution, which is a common phenomenon seen in similar analyses. This is because the frequency of k-itemsets decreases as k increases.

Most common cause of death each year: year
2012 (2012, Multiple Drug Toxicity)
2013 (2013, Heroin Intoxication)
2014 (2014, Heroin Intoxication)
2015 (2015, Acute Heroin Toxicity)
2016 (2016, Acute Fentanyl Intoxication)
2017 (2017, Acute Fentanyl Intoxication)
2018 (2018, Acute Fentanyl Intoxication)
2019 (2019, Acute Fentanyl Intoxication)
2020 (2020, Acute Fentanyl Intoxication)
2021 (2021, Acute Fentanyl Intoxication)

We are seeing the growth/influence of Fentanyl starting from 2012 as it has the highest confidence i.e., chance of joint-occurrence with other drugs, a classic quality of an upcoming drug. One more point to note is that just before it hit its peak support in 2018, it showed highest confidence in 2015. This suggests the moment when Fentanyl had really seeped into the state and was about to solely overtake all other death types. In terms of FPM and association rule mining context, we can say that confidence(k $\rightarrow$ Fentanyl) is likely to be greater for earlier years relative to later years as we can see from its gradual decrease. This can safely be attributed to its increasing influence.