San Jose State University

Department of Applied Science

One Washington Square, 95112

DATA 230: Sec-12 Data Visualization

DATA 230 – Final Project: Report

Effect of Alcohol and Drugs on Juvenile Delinquency

Course Instructor:	Prof Seugjoon Lee	
Team Members:	Vansh Sharma	016003624
	Soumit Reddy	016000062
	Sachin Kumar	016594773
	Deekshita Prakash	016597815
	Shashidhar Reddy	016014570
Submission Date:	12.09.2022	

Table of Contents

1. MOTIVATION

The aim of this project is to explore the relation between substance use and delinquency among juveniles. Since the dataset provided for the project was from a study titled “Second International Self-Reported Delinquency Study, 2005-2007”, one of the original aims of the dataset was probably to identify the variables involved that contribute to delinquency in juveniles. It is no secret that substance use distorts the judgement of any human being, let alone juveniles. Therefore, it is possible that substance abuse some sort of positive correlation with both the level of violence and frequency of occurrences in delinquency.

There are many studies in the literature that also explore the same issue. For instance, a study published in the European Journal on Criminal Policy and Research that used the same dataset claims that alcohol use, and especially abuse, is a risk factor for criminal activity and vice versa [1]. Other similar studies claim that there is a strong link between substance use problems and delinquency [2], [3].

2. DATA VISUALIZATION

In order to explore the relation between substance abuse use and delinquency in juveniles, **five types** of data visualization methods were used. These were choropleth map, pie chart, scatter plot, heat map and 3D scatter plot. Each of these visualizations are explored in the sections below with details on **how they were obtained**, their **marks/channels** and the **purpose** behind using them. In addition, the data was not used in raw format in almost all of these visualizations. **Data derivation** is also explained in these sections.

Note 1: All ambiguous/blank data points were filtered out of the data as a pre-processing step.
Note 2: All student data refers to students in grades 7 to 9.

2.1 Choropleth Maps: Alcohol/Drug Use and Delinquency Rates

The choropleth maps below show the alcohol/drug use and delinquency rates in juveniles (see Fig. 1 and Fig. 2).

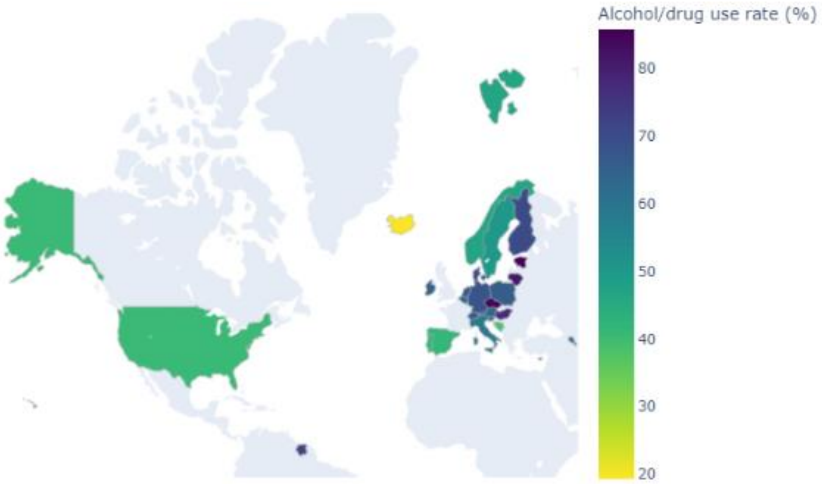

Figure 1: Alcohol/drug use rate in students in grades 7-to-9 by country

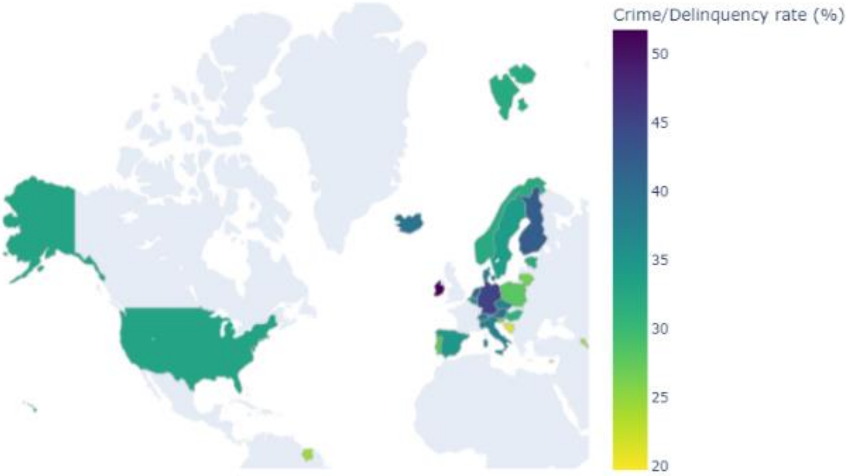

Figure 2: Delinquency rate in students by country

The **data processing** involved finding the fraction of students in a given country within the dataset that has either used drugs/alcohol or committed some sort of delinquency. Thus, the substance use and delinquency rates were calculated using the formula below.

$$choropleth(x, c) = \frac{N(X \ni x \cap X \ni c)}{N(X \ni c)} \times 100\% \qquad (1)$$

X: Dataset of all students

x: Delinquency or substance use

$X \ni x$: Set of students that are commited have done x before
c: Country

$X \ni c$: Set of students that are from country c

$N(X \ni c)$: Number of students from country c

Note: Netherlands Antilles dissolved in 2010. Since there were multiple succeeding states, there were no sure way of accurately illustrating geographically that portion of data, which was discarded at the end.

In these choropleth maps, the **marks** are of item type of 2D area and they represent the corresponding borders of the countries. On the other hand, the **channels** are position and color of these areas. Using the the viridis colormap, the color/hue of the areas indicate the substance abuse and delinquency rates as a percentage of the students that filled out the questionnaire. The position channel does not represent a quantitative value in the traditional sense (apart from coordinates obviously), but merely shows the location of the country. One can also make sense of the data by the proximity of the countries using this channel.

There are two main **reasons** why visualization was used. **Firstly**, it was used as a rudimentary form of Exploratory Data Analysis (EDA). Factors related to drug/alcohol use family, school and neighborhood life were tried and compared with the choropleth map of delinquency rate. It was noticed that the map with substance abuse had a lot of overlap with the delinquency frequencies (e.g. USA, Italy, Scandinavian countries). The **second** purpose of this visualization is that it shows drug/alcohol use alone does not explain juvenile delinquency by itself since there are countries that have high substance abuse but low delinquency in juveniles and vice versa (e.g. Iceland ,Estonia and Suriname). Therefore, the plots indicate that there might be cultural variations that contribute to the effect.

2.2 Pie Charts: Substance Use and Delinquency Types

The pie charts below show the types of alcohol/drug use and also delinquency type among juveniles (see Fig. 3 and Fig. 4).

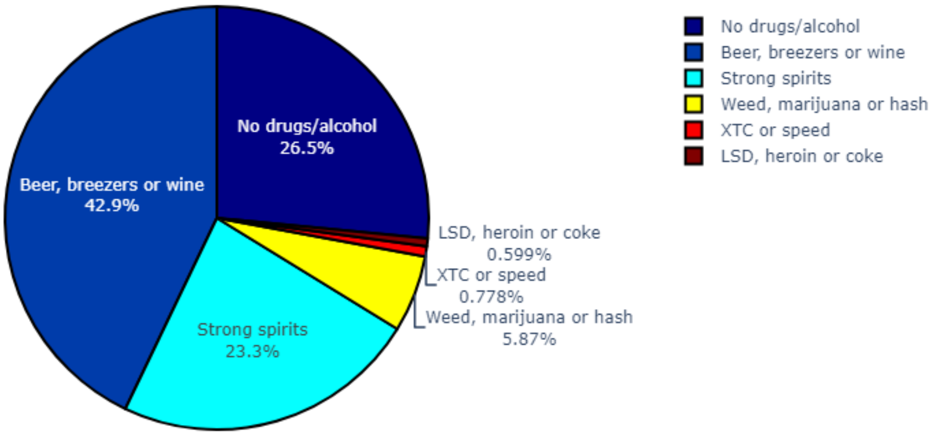

Figure 3: Pie chart of substance use type among students

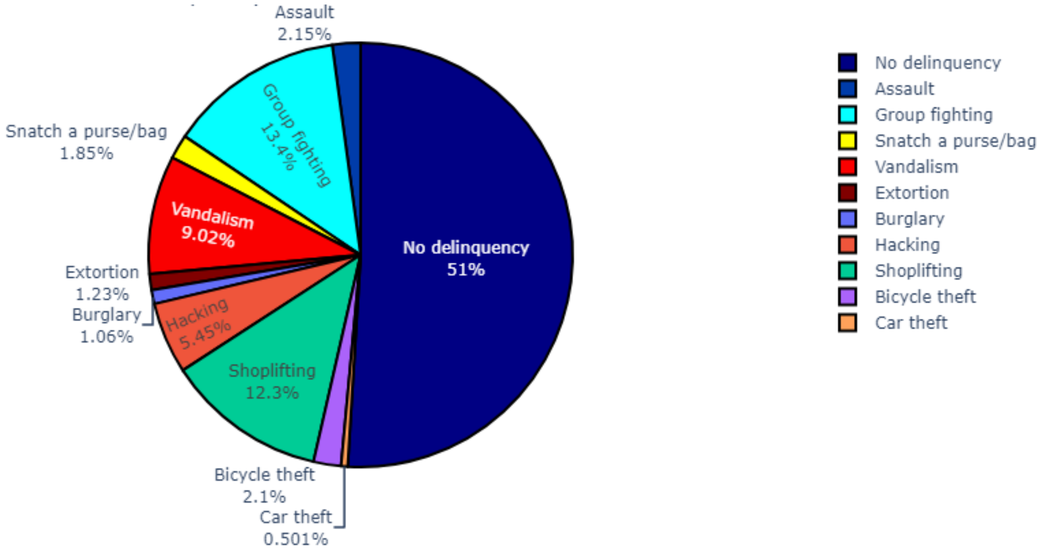

Figure 4: Pie chart of delinquency type among students

No complicated transformation was made during the **data processing** steps. However, the data from a student using several substances or committing multiple types of delinquencies was added to the pie charts with one count in each relevant slice to avoid having problems with labeling. For instance, a beer and spirits drinker adds +1 count each category but one can also label the student as a spirits drinker only since that is the harder alcohol type. The probability values in the pie chart can be represented as shown below.

$$pie(x_i) = \frac{N(X \ni x_i)}{\sum_{i \in all\ classes} N(X \ni x_i)} \times 100\% \qquad (2)$$

x_i: Delinquency (or substance) i

$X \ni x_i$: Set of students that have done x_i before

$all\ classes$: Set of all classes possible for x

Similar to the previous illustration, the **marks** are the 2D areas are used which are represented as the slices of the pie charts. The size of the slices and the hue of the colors are the **channels**. Bigger slices indicate larger percentage of use among students and the hue is used to distinguish between the types. In addition, the ordering was chosen in a basic categoric sense (e.g. alcohol users, drug users, violent delinquents).

The **reason** behind the use of these pie charts is to show how frequent they are in relation to each other. This is an important step to check in data analysis because it shows how **unbalanced** the data is. For instance, most of the students have drank beer before, which means that the frequency of beer drinkers among any delinquency groups will probably be the highest. However, one can not conclude from this data that beer drinking in juveniles directly lead to delinquency because this frequency is also high for non-delinquent type students. On the other hand, it seems that one can form a category between **no-delinquency students**, **non-violent delinquents** and **violent delinquents** since their slices seem to be roughly equal (see Fig. 4). This information will come in handy for the heat map visualization (see Section 2.4).

2.3 Scatter Plot: Frequency of Delinquency In Relation to Substance Abuse Frequency

The scatter plot below show the of alcohol/drug scores with colormapping of the average delinquency score among juveniles (see Fig. 5).

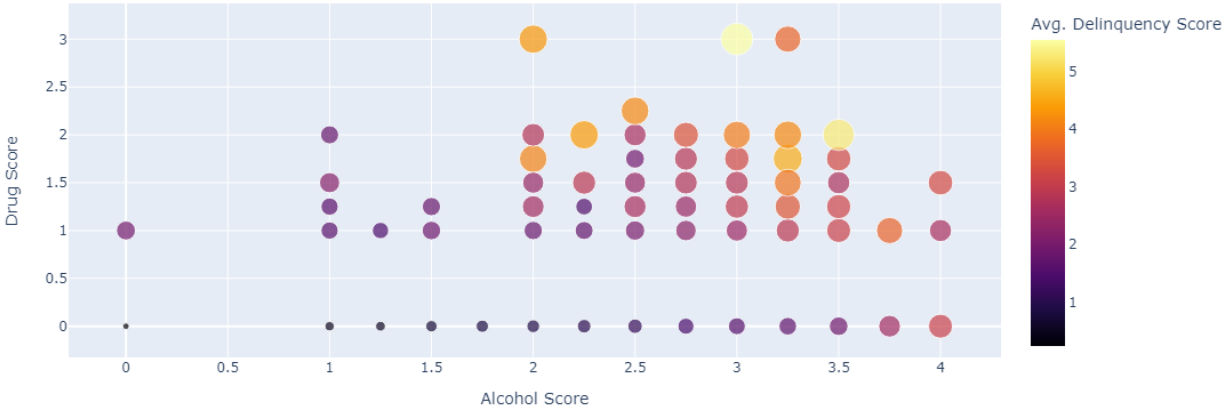

Figure 5: Scatter plot of the average delinquency scores vs substance use

All three data dimensions of this visualizations was derived during the **data processing** step. For each type of drug/alcohol, the drug/alcohol score was increased by 1 for previous use and additional scores were also given from last month/year periodicity from a scale of 0 to 1. This means that a student can get an Alcohol Score of 4 and a Drug Score of 6 at most (2 types of alcohol and 3 types of drug categories). The delinquency scores were marked similarly as well. After this step, it wouldn’t be expressive to just scatter plot these scores. As an additional step, the average delinquency score for each combination of drug and alcohol score was calculated to be plotted. To eliminate some of the outliers, score combinations with instances less than 10 were discarded. The average delinquency rate, which is represented by point size and color, and also the alcohol/drug scores were calculated using the formulae below.

$$Scatter(s_a, s_d) = \frac{\sum s_{del}(X \ni s_a \cap X \ni s_d)}{N(X \ni s_a \cap X \ni s_d)} \tag{3}$$

s_i: Score associated with the use of substance type i (or delinquency if i=del)

$X \ni s_a$: Set of students that have done their alcohol score equal to s_a

For a given student, the substance use (or delinquency) scores were calculated using the following formula.

$$s_i = \sum_{j \in i} LTP_j + 0.25LYMC_j \tag{4}$$

LTP_j: Boolean. Whether if the student ever used the substance (or commited the delinquency) $j \in i$

$LYMC_j$: Integer from 0 to 4. The frequency of the student in using the substance (or committing the delinquency) $j \in i$ in the past month (year)

Note: The details of each “LTP” and “LYC” or “LMC” marked data can be found in the referenced website [4].

The data was **marked** by points in the scatter plot. Position in both axes and the sizes/colors of the points were used as **channels**. Higher values in the x/y axis indicates higher scores of alcohol/drug. In addition, both the area and the hue of the points indicate the average delinquency score of the given alcohol/drug score combination. Two different channels for this variable were used to be more **expressive** and **redundant**.

In this scatter plot, the **main goal** is to form a simple model for **bivariate analysis** and both confirm the positive correlation between substance use and delinquency and also to detect **outliers**. One can see from the scatter plot that there is indeed a roughly linear positive correlation. Combinations of high alcohol/drug scores usually have high average delinquencies. However, it can also be seen that a simple linear model is not enough to explain all delinquency with this scale of variability.

2.4 Heatmap: Type of Substance Use Corresponding to Delinquency Violence Levels

The heatmap below shows percentage-wise the type of substance use corresponding to the distribution in delinquency type among juveniles (see Fig. 6).

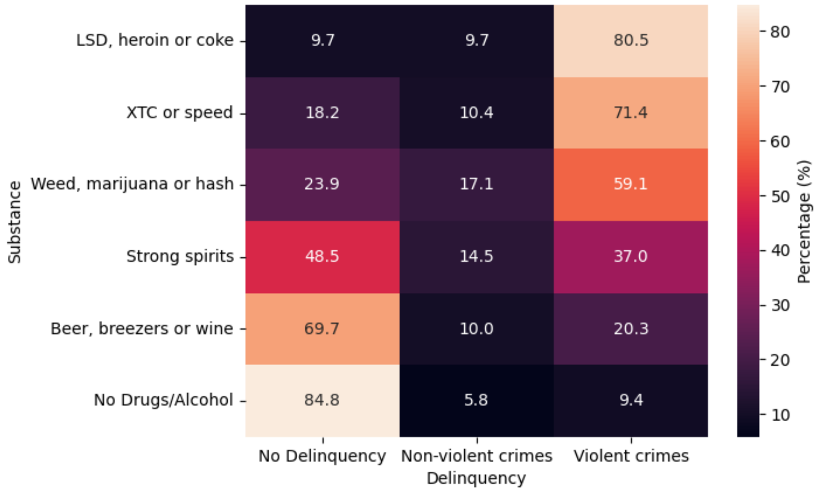

Figure 6: Type of substance use corresponding to delinquency type (row-wise normalized)

A crucial step was taken at **data processing** for these graphs. In all of the subcategories, there were a lot of overlap. A student who drinks spirits can also be a beer drinker, which created an issue on how to label the student. This issue existed for the delinquency types as well. For this reason, the severeness of the substances were ordered and the student was labeled with the most severe substance he/she used before. Likewise, the students who committed both violent and non-violent delinquency were labeled as violently delinquent students. Since we know from the **pie charts** that the substance use distribution is highly unbalanced, the heatmaps were normalized row-wise. In other words, each row adds up to 100%. The percentage value in each of the bins in the heatmap was calculated using the formulae below.

$$Heatmap(sub, del) = \frac{N(X \ni sub \cap X \ni del)}{N(X \ni sub)} \times 100\% \tag{5}$$

sub: The most severe type of substance used by a student (y-axis)

del: The most severe type of delinquency commited by a student (x-axis)

The **marks** of the heatmap are 2D areas represented by each tile. The position and color of each tile serve as the **channels**. In terms of position, the x axis is grouped by the severity of the delinquency type with non-delinquents obviously being the least severe category. Likewise, the y axis groups the substances as either alcohol or drug and orders them in the order of increasing severity. For the purposes of this visualization, the ordering was taken as No Drugs/Alcohol < Beer, breezers or wine < Strong spirits < Weed, marijuana or hash or speed < LSD, heroin or coke. The matter colormap was used mapped to the frequency of delinquency type for a given substance type.

The **reason** behind the use of the heatmap was to both verify the correlation between delinquency and alcohol/drug use and also to take a closer look at the violency levels in students for each substance use case. In other words, this plot helps us to estimate “What is the probability that a student that uses substance-Y is a X-type of delinquent.” It can be seen that the level of violence in delinquencies increase with the severity of the substance used.

2.5 3D Scatter Plot: Principal Component Analysis for Delinquency Violence Levels

The scatter plot below shows the first 3 principal components of the PCA analysis on variables related to drug/alcohol use and neighborhood/family/school life grouped by delinquency violence level among juveniles (see Fig. 7).

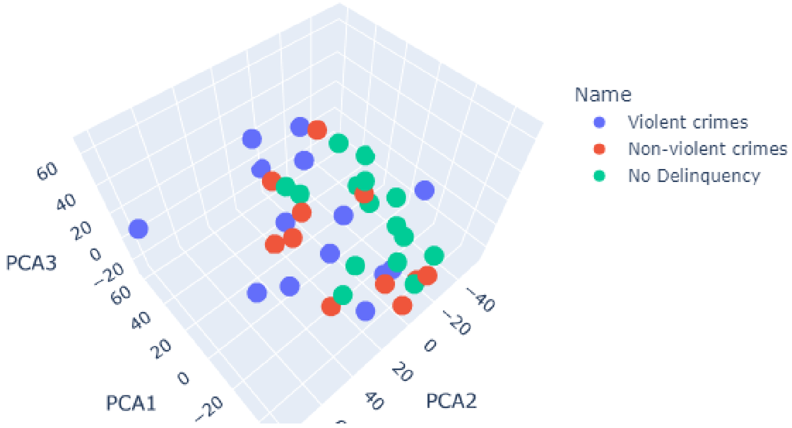

Figure 7: 3D Scatter plot of the PCA for type of delinquents by violence

As **data processing**, a lot of variables had to be obtained from the dataset. Last month use frequency for each substance, neighborhood bonding/disorganization/integration/quality, family bonding/affluence/stability in work and school climate/disorganization were used in the analysis as variables. For scatter plotting the PCA results, the results were under sampled to even the number of delinquency type frequencies.

The 3D scatter plot shown above uses points as **marks**, whereas position and color serve as **channels**. The color helps with categorizing the violence levels of the delinquency of the students and the position along each principal component is helpful in distinguishing between these categories in theory (more apparent if the data explains the variability in the categories).

There are several **reasons** behind the use of the PCA and this visualization. The PCA was first used during the **EDA** to identify the most important features that explain the variability in the data if one classifies it using delinquency violence type. Results indicated that family bonding and also alcohol/drug use scores were some of the most important features for each principal component. However, as it can be observed from the plot, and also from the fact that the **explained variance ratio** being as low as 0.365 even for the 1st principal component, the data (or atleast these variables) are not enough to explain the delinquency type among students completely. Therefore, even though we've found important correlations between substance abuse and delinquency violence levels among juveniles, the issue can not be attributed to alcohol and drug use alone indicated by the features used in this analysis alone.

3. ARRANGEMENT

The general overview of the visualization arrangement is as follows:

I. Discovery: The relation between the alcohol/drug use with delinquency in the data.

II. Data distribution observation: The distribution of sub-types among alcohols/drugs and delinquencies.

III. Simple analysis: Observe the positive correlation between delinquency and substance use.

IV. Verification and more in-depth analysis: Verify the correlation between heavy substance abuse to the violence levels in delinquencies and also explore their distributions in each substance sub-category.

V. Acknowledge deficiencies: Show that alcohol/drug related data alone can not explain all of the variability in the delinquency violence levels for this dataset.

4. RESULTS AND DISCUSSION

-There is definitely a positive correlation between substance abuse and juvenile delinquency (scatter, chloropleth)

-The data is distributed unevenly, so a careful analysis is needed to draw conclusions. Here we separated the delinquencies as non-violent, violent, no-crime partly because of that (pie chart)

-The harder drugs and stronger alcohols lead to increased violency levels in delinquency (heatmap)

-The data related to politics/culture might be helpful in getting a better understanding of childhood delinquency (chloropleth)

-It does not fully explain all of the variability in delinquency data (PCA)

REFERENCES

[1] U. Gatti, R. Soellner, H. Schadee, A. Verde and G. Rocca, "Effects of Delinquency on Alcohol use Among Juveniles in Europe: Results from the ISRD-2 Study," *European Journal on Criminal Policy and Research*, Jun. 2013. [Online]. Available: https://www.researchgate.net/publication/257552736_Effects_of_Delinquency_on_Alcohol_use_Among_Juveniles_in_Europe_Results_from_the_ISRD-2_Study. [Accessed: 6 Dec. 2022].

[2] J. Kraus, "Juvenile drug abuse and delinquency: some differential associations," *British Journal of Psychiatry*, 1981. [Online]. Available: https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/abs/juvenile-drug-abuse-and-delinquency-some-differential-associations/F69AB9E5E0CA30A396E4CC1FA8C8D5E6. [Accessed: 6 Dec. 2022].

[3] E.P. Mulvey, C.A. Schubert and L. Chassin, "Substance Use and Delinquent Behavior Among Serious Adolescent Offenders," *European Journal on Criminal Policy and Research*, Dec. 2010. [Online]. Available: https://www.ojp.gov/pdffiles1/ojjdp/232790.pdf. [Accessed: 6 Dec. 2022].

[4] D. Enzmann, I.H. Marshall, M. Killias, J. Junger-Tas, M. Steketee, B. Gruszczynska, Second International Self-Reported Delinquency Study, 2005-2007 (ICPSR 34658). [Dataset]. Available: https://www.icpsr.umich.edu/web/NACJD/studies/34658/datasets/0001/variables/ASLTLYC?archive=nacjd. [Accessed: 6 Dec. 2022].

APPENDICES

Appendix A: Summary of the Visualizations

	Choropleth Maps	**Pie Chart**	**Scatter Plot**	**Heatmap**	**3D Scatter Plot**
Data Processing *	Calculation of substance use/delinquency ratio	-	Substance and averaged	Label ordering by severity Categorizin	Categorizing based on violence level and

			delinquency scoring	g based on violence level	undersampling for visualization discriminability
Marks	2D Area	2D Area	0D Point	2D Area	0D Point
Channels	Position and color	Area and color	Position, size and color	Position and color	Position and color
Why?	-Rudimentary EDA -Early indications of substance use relation to delinquency -Show more variables such as cultural/political differences effect results	-Visualize the imbalance in data -Explore ways to group substance and delinquency types for balancing	-Bivariate analysis -Observe the relation between substance use and delinquency -Detect outliers	-Verify the correlation between substance use and delinquency using different data type	-Observe the most important variables for predicting delinquency violence levels -Show not all of the variance in the delinquency types are explained by this data

Table 1: Summary of important information on data visualizations

*Ambiguous and blank data were filtered for all.