Narcolands

A data driven approach to event based drugs popularity in the Netherlands

Faye Tervoort 12250716
Anisha Sloeserwij 14113341
Ana Chaloska 11097183
Jesse Kommandeur 12716197
Danny de Vries 14495643
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

Here goes the abstract.

KEYWORDS

drugs, google trends, court cases, Netherlands, visualization

1 INTRODUCTION

Every country is affected by crime, and so is the Netherlands. According to the Dutch center for crime prevention and safety statistics (CCV), over 90% of all organized crime is related to illegal drugs (CCV, 2022). The Netherlands owes this percentage to its prominent role in international drug trade (McDermott et al., 2021). Since problems surrounding drugs have increased considerably in recent years, authorities have prioritized and focused on identifying and tackling drugs related behavior (Eski and Buijt, 2016; Ferwerda et al., 2016).

In order to battle drugs related organized crime better, it is important for authorities, such as the police, to understand the modus operandi of criminal organizations through both academic and practical research. One of the organizations that focuses on the discovery of new modus operandi is the Police Academy (PA). In cooperation with the University of Amsterdam, the PA wants to focus on data driven research to identify drug related patterns in the Netherlands.

Previous research within the narcotics space focused on the supply-side of illegal drug trade (Paoli et al., 2013; Magliocca et al., 2019). This domain focuses on understanding the production, distribution, and trafficking of illegal drugs. It includes research on the modus operandi used by criminal organizations to produce and transport drugs, as well as the routes and networks they use to distribute them. This type of research is important for law enforcement agencies as it helps them identify and disrupt the operations of drug traffickers (McDermott et al., 2021).

Another area of research that is also present in prior literature is the demand-side of illegal drugs. This domain focuses on understanding the consumption and impact of illegal drugs on individuals, politics, and society (Flores-Macías & Zarkin, 2019; Riley, 2017). It includes research on the psychological and social factors that contribute to drug addiction, and the health and social consequences of drug use (Gonzalez, 2015; Sesnie et al., 2017).

These research domains are important for public health agencies and policymakers as it helps them understand the extent of the drug problem and develop effective strategies for addressing it. However, most of these studies are less relevant for authorities such as the PA since they focus less on quantitative data driven methods to identify patterns in society. Therefore, this study focuses on a data driven pattern identification of drug usage and events, such as festivals and public holidays. In order to investigate this relationship, this study aims to answer the research question: *"To what extent can we identify event-based drug popularity on online data resources?"*.

To answer this research question, the following sub questions were formulated:

- Which Dutch events are indicative for drug popularity?
- What drugs are used during events in the Netherlands?
- Which online data resources are relevant to identify drug popularity at events?
- How could we identify drug popularity at events?
- What are the differences between the specific events and drugs?
- How could we visualize event based drug popularity on online data resources?

The sub questions are relevant to the main research question as they provide a more detailed and specific understanding of the topic. The subquestion "Which Dutch events are indicative for drug popularity?" helps to narrow the scope of the research since it provides a starting point for identifying which events may be associated with drug use. Second, the subquestion "What drugs are used during events in the Netherlands?" is important for identifying which specific drugs are of interest for the research. "Which online data resources are relevant to identify drug popularity?" and "How could we identify drug popularity at events?" are crucial for understanding the methods and data sources of this research. "What are the differences between the specific events and drugs?" helps to identify any patterns or trends in drug popularity that may vary depending on the specific event or drug being considered. The last subquestion "How could we visualize event based drug popularity on online data resources?" is relevant since it allows us to transform our research data into a user-friendly dashboard to create insights for the PA.

In order to answer the research questions, we first present the theoretical state of affairs, followed by the methodological set-up, after which the results are presented. In the final sections, the most important findings are concluded and limitations are discussed, followed by recommendations for future work.

2 RELATED WORK

The collection and utilization of large amounts of data is a popular resource nowadays and is used in a variety of disciplines. [bron] The use of these large amounts of data, also known as big data, combined with the use of data analysis is having a major impact on the social sciences and humanities in general. [bron] Chan and Moses' research shows that it has a particularly large impact for the specific field of criminology. [bron]

2.1 Prediciting trends

Several studies have been conducted that have investigated whether Google trends can be a possible resource for predicting particular trends. Previous research has shown that Google Trends can be used as a predictor in different fields. Such as the healthcare industry, where Google Trends can help support the prediction of the outbreak of seasonal influenza and COVID-19 [bron], [bron]. Another study by Kassraie et al. used Google trends data in combination with Twitter data to predict the popular vote of the 2016 presidential election in the US. They concluded that the combination of these social media platforms could be a mirror for public opinion on political events [bron].

In the study by Perdue et al., it appears that Google trends can be a possible predictor for drug abuse trends. [bron] This is supported by the research of Gamma et al. which investigated whether there is a comparison between time trends of Google search interests and offences committed in relation to the drug called Methamphetamine. This study found that law enforcement could use the Google search feature as a possible predictor of Methamphetamine-related crimes. [bron]

2.2 Predictive policing

Predictive policing is a term used for predicting certain behaviours or trends based on analyzing various data for law enforcement. [bron] Authorities use predictive policing to predict crime in order to prevent criminal activity instead of reacting to the crime that already occurred. For this reason, the domain and effects of predictive policing have been examined for a long time.

However, not all research agrees that predictive policing is a valid method. According to The New York Times debate, predictive policing is a very effective way when it comes to predicting criminal behaviour, but still contains many improvements in terms of ethicality. [bron] In contrast, Hardyns' research says that the analysis of such predictions has indeed shown its worthiness for different predictive systems in different areas, but exactly because it is a new development in the field of criminology, little is yet known whether it is an effective way for law enforcement. [bron]

3 METHODOLOGY

The approach to answering the research questions will be described in this section. The methodology consists of four parts. The first part describes the methods and literature that was used to select drugs and events. The second part describes the data pipeline which includes the data selection, data collection, data processing and data privacy. The third part describes the statistical metrics that were used to analyze the data. The last part covers the methods that were used to convey and visualize the data into the dashboard.

3.1 Methods for identifying drugs and events

3.1.1 Method for drug selection. To determine the different types of drugs which are used in the study, several institutes that collect data on drugs in the Netherlands were analyzed. The largest institution in the Netherlands that is related to drugs is the Trimbos Institute (TI). This national organization conducts research on the mental health of the Dutch people with a focus on the use of alcohol, tobacco and drugs. They involve all age groups in society and therefore cover the entire life cycle of citizens. This institute releases various analysis on alcohol, tobacco and other drug use in the Netherlands which includes reports on how to address these problems by naming prevention, education and policies.

In addition, The TI tracks drug-related developments through various monitoring systems. The most important monitor is the National Drug Monitor (NDM). The NDM collects and compiles all data on substance use, the drug market and drug-related crime of all ages in the population. This institution aims to provide a representation of the figures known in the Netherlands related to drugs. Using both data from TI and NSM, a set of specific drugs were selected for this study.

3.1.2 Method for event selection. First of all, it was determined which events would provide interesting results. Because of the sub-question "Which Dutch events are indicative for drug popularity?" it is interesting to use events where drug usage occurs a lot. Because where drugs are a popular means of enjoyment, they will also be used a lot. For this purpose, research was done on the use of drugs at different events. That way, the events served as a proxy where drug usage is known to increase. The first method that is used is literature research. This literary research was conducted to see if there has been previous research executed on drug use for different events in the Netherlands. However, very little is generally known about specific drug use per event in the Netherlands. Actually there is no research done in the Netherlands about drug use at specific parties or events. Due to the fact that it is difficult to investigate because few drug users want to be open about their drug use. For that reason we used another method where different news articles that say something about drug use per event are examined. This involves using different news sources, such as the AD, the NOS and the Telegraaf, as well as regional news sources such as LokaalGelderland and Echt Amsterdams nieuws. Using these two methods, a number of events have been identified where drugs are commonly used among the public. The research and answer to the sub-question is described in Chapter X.

3.2 Data Methods

3.2.1 Data selection. Before data could be collected and analyzed, a selection of data sources has been made. When it comes to the research question "To what extent can we identify event-based drug popularity on online data resources?", Google Trends data, Twitter data, and Dutch news data represent appropriate sources of information.

Google Trends data provides a comprehensive insight into the popularity of search terms and topics in real-time. By analyzing the frequency of drug-related searches on Google, researchers can obtain a thorough understanding of drug popularity across regions and over time (Batistic, 2021). Furthermore, Google Trends offers numerous filters and visualization options, making it a user-friendly tool for data analysis.

Twitter data represents a rich source of information on public perceptions and opinions (Bian, 2016). The platform is well known for its user-generated content and real-time information, making it a valuable resource for identifying the public's views on drugs. Additionally, Twitter's hashtag and trending topic features allow researchers to quickly identify the most popular drug-related topics on the platform.

Dutch news data is a crucial source of information in the assessment of event-based drug popularity. It provides a comprehensive understanding of media representation of drug-related events, including how these events are reported and framed, and the public's perception of drug use. By analyzing news data, researchers can gain a deeper insight into the media's role in shaping public opinion on drug use (Mccombs, 2020).

In conclusion, the combination of Google Trends, Twitter data, and Dutch news data offers a diverse and comprehensive set of data sources for understanding event-based drug popularity on online data resources. These data sources provide valuable information on various aspects of drug use, such as search trends, public opinion, and media representation, making them a sound choice for answering the research question at hand.

3.2.2 Data collection. After data sources were selected, the data was collected. The Google Trends data was obtained using PyTrends, a library in Python for accessing and retrieving data from Google Trends (PyPI, 2023). This method of scraping allowed for an efficient and automated process for collecting Google Trends data, enabling the researcher to obtain a large amount of data in a short period. Only searches that were done on the territory of The Netherlands were considered.

Twitter data was collected using SNscrape, a tool for scraping social media data, including tweets and user profiles (GitHub, 2023). This method of scraping Twitter data provided access to a significant amount of real-time user-generated content, allowing us to research the popularity of conversations surrounding specific drugs in the netherlands. We based on a Twitter scrape query on the time interval January 2014 - December 2022 that contained the words 'xtc', 'cocaine' or 'ghb'. Since we are only interested in Dutch tweets, we excluded tweets that are non-Dutch. In total, the Twitter scraper collected more that 1.500.000 relevant tweets.

The Dutch news data was obtained by downloading a news corpus from the NOS, a Dutch public broadcaster. This method of data collection allowed us to access a large amount of news articles in a centralized Kaggle repository, which was last updated at the end of 2022 (Scheijen, 2022). In total, the NOS contained more that 250.000 relevant news articles.

3.2.3 Data processing. After the selected data was collected datasets were loaded into a dataframe for processing. The pre-processing of the obtained Google Trends, Twitter, and Dutch news data was crucial in ensuring that the data was in a format suitable for analysis. To this end, the following pre-processing methods were applied:

- Conversion of dates: All dates within the specified time frame were converted to the same datetime format. This standardization of the dates was essential for ensuring consistency in the data and making it easier to analyze.
- Calculation of week numbers: A function was created to calculate the week number of each date, as the time interval used for the analysis was "week." This function allowed for the grouping of data into weeks, making it easier to analyze trends and patterns over time. The data was aggregated on a weekly level because that was the most granular aggregation of one of the sources (Google Trends, 2023). To have a comparable analysis, the weekly aggregation was applied to the data from all sources.
- Topic feature extraction: A function was created to check whether a tweet, news article, or search query contained a specific word, such as "cocaine" or "xtc." This process of topic feature extraction was essential in identifying and isolating the data that was relevant to the research question and in understanding the prevalence and significance of specific topics in the data.
- Normalization: The extracted features were normalized, as the values were absolute, while the Google Trends data were relative. Normalization between 0 and 100 was performed, and this came in useful when performing statistical tests and visualizing the data in the dashboard. Normalizing the data allowed for comparison and analysis of data from different sources, as the data was all expressed in the same unit.

Overall, the pre-processing of the Google Trends, Twitter, and Dutch news data was essential in ensuring that the data was in a suitable format for analysis. The conversion of dates, calculation of week numbers, topic feature extraction, and normalization were critical in making the data easier to analyze and interpret, and in ensuring that the results obtained from the analysis were accurate and reliable.

In addition,other pre-processing techniques such as sentiment analysis, locational feature extraction, word2vec synonym extraction and cleaning function have also been created. These functions could be used in future work to extend the research, make it more precise or compute detailed micro-level information.

3.2.4 Data privacy. In conducting the data collection and pre-processing of event-based drug popularity on online data resources, data privacy was a significant consideration. All data sources used were subjected to ethical and privacy considerations to ensure that all personal information was protected and that the data was collected, processed, and used in a responsible and ethical manner.

It was ensured that all data sources used were publicly available and did not contain any sensitive or personal information. The Google trends data was collected via Pytrends, in which personal information is already anonymized. For example, Google Trends data did not contain any information on the users who did the searches, but only the location where the search happened. The Twitter data was obtained using the SNscraper, which only collected data from public profiles and ensured that the data collected did not contain any personal information. For example, username and tags were removed.

In addition, appropriate measures were taken to protect the privacy of individuals and organizations that provided the data. For instance, all data was de-identified to remove any personal information that could be used to identify individuals or organizations.

3.3 Statistical methods

3.3.1 Hypothesis. To identify event-based drug popularity at events the following hypothesis was formulated: There is a significant increase in the metrics (number of searches, -tweets and -news articles) of each data source during the weeks of the events known for increased drug usage.

3.3.2 Computing peaks per regular week. A metric was constructed in order to translate the hypothesis into a test design. The data of each data source was aggregated weekly to represent the counts per searches, tweets and news articles respectively.

The aggregated data was ordered chronologically and plotted on a line plot. The peaks in the line plot were found using the method signal.findpeaks from Pythons library Scipy [reference]. This step provided evidence on the weeks where there is an increase in the representative metric in relation to the other weeks in the considered time range (2014-2022).

The data with labeled peaks was stored in a dataframe in the form of a Bool column coupled with a column for the corresponding week number. For instance if there was a peak in Week 2 2022, the Bool column would be populated with True, else it would be populated with False. With that each week number was represented with 9 rows, 1 for each year between 2014 and 2022. Finally, a proportion was calculated for each week representing the percentage of rows with "True" value out of the 9 rows for each week. With that, each week was represented with a number representing the percentage of years where a peak was observed for that particular week [Table with example from slides - Appendix].

3.3.3 Computing peaks per event week. The above steps were performed again in order to represent the percentage of peaks for the weeks of the events with increased drug usage (King's day, ADE, Pride, Lowlands and New Year's Eve). This was needed because most of these events do not always fall in the same week of the year.

For every year the dates of the events were identified. This was necessary because some of the events happen on a different date every year (e.g. ADE, Pride) [Dates per event - Appendix]. Based on the event dates, the week in which each event happened were identified. [Appendix]. Based on the data extracted from the line plot labeled with peaks, the years in which a peak happened per week of event were identified. Similarly as in the peaks per week, the percentage of weeks in which a peak was observed were calculated.

3.3.4 Test design. A p-value approach proportion test [reference] was performed on the generated data in order to test the hypothesis above. The test was once for each combination of drug, event and data source. Below are the specifications of the test:

- Null hypothesis (H0): The proportion of peaks observed in drug related event weeks is not significantly larger than the proportion of peaks in regular weeks.
- Alternative hypothesis (H1): The proportion of peaks observed in drug related event weeks is significantly larger than the proportion of peaks in regular weeks.
- Level of significance: To determine level of significance, an alpha of 0.05 was used. This implies 5% risk of concluding that there is a larger proportion when there in reality that is not the case.
- Test statistic (z): The test statistic was calculated based on the sample proportion and population proportion [formula - Appendix], which is suitable for a proportion test.
- Accepting the alternative hypothesis: The p-value was calculated from the test statistic. A p-value of less than 0.05 (corresponding to the significance level) indicated evidence to reject the null hypothesis and accept the alternative hypothesis, and hence prove that the proportion of peaks during event weeks is indeed larger than regular weeks.

3.4 Visualization methods

GIVE INTRODUCTION ABOUT: interface (visualization/controllability). Dashboard interface etc. Then move on to user requirements.

These user requirements are derived from the case description provided by the client as well as feedback from the stakeholder during the initial ideation and prototyping phase of the project and where further narrowed down during the project design workshops using the MoSCoW priorization method. The term 'user' in the requirements refers to two specific types of similar target audiences that will make use of our prototype, police agents who want to explore the dataset to gain insights and data analysts who want to filter and compare our datasets.

(1) (M) The user must be able to use the prototype on a personal computer and interface with a screen
(2) (M) The user must be able to filter the datasets to compare different years of data
(3) (M) The user must be able to filter the datasets to compare different types of drugs
(4) (M) The user must be able to overlay multiple datasets and trend lines on top of eachother
(5) (M) The system uses open-source software and not locked-in corporate data tools
(6) (M) The system has a user-friendly visual design and interaction design
(7) (S) The user should create an account to store specific and personalized filters
(8) (S) The user should be able to download the raw datasets in specific file formats
(9) (S) The user should be able to navigate between different overviews showing corresponding data
(10) (C) The user could upload their own dataset and sources of specfic drugs and news sources
(11) (C) The system uses real-time up-to-date API data

3.5 Collaboration

WRITE ABOUT: at a meta-level, team management. Something about the GitHub Org, Trello team management.

4 RESULTS

4.1 Selection of events

In the Netherlands, several festivals and parties are organized on a large scale that attract a lot of visitors. Because a very small amount of numbers are known about drug use in the Netherlands at certain events and holidays, it is hard to say at which events drug use is more popular than others. [bron] For this reason, the research focused on news articles showing high levels of drug use at various events. In order to conduct a proper analysis within the time frame, five events were chosen.

4.1.1 Unpopularity of drug use. A small survey of drug use at various events revealed that not all major events in the Netherlands involve high levels of drug use. A remarkable aspect is that at De Zwarte Cross, the largest festival in the Netherlands with 220 thousand visitors, hard drugs are not popular among the audience. De Zwarte Cross is originally a cross race that lasts 3 days combined with music and performances by various bands. At De Zwarte Cross a lot of alcohol is consumed in particular. [bron] In addition, it appears that Carnival, a multi-day festival celebrated in the south of the Netherlands, also mainly consumes a lot of alcohol and has much less interest in hard drugs. [bron]

It can be explained that drugs do not have a major presence at these large events because drugs are mainly used during dance festivals where mainly techno and house music is listened. According to [bron], 70% of dance festival visitors use drugs. Both at De Zwarte Cross and during Carnival, this style of music is not played very frequently and this could be a reason for the unpopularity of drug use during those events.

However, the Amsterdam Dance Event (ADE) appears to be a popular event for drug use. [bron] During this multi-day dance festival that consists of multiple parties spread across multiple locations in Amsterdam, a lot of techno and house music is played. This would support the explanation of drug use and certain parties and festivals. Another festival where drugs appear to be widely used is Lowlands. [bron] This festival lasts one weekend in the beginning of August where various artists around the world perform at this festival.

Besides these big festivals, many parties and smaller festivals take place during certain holidays in the Netherlands. For example, Koningsdag, one day every year where the Dutch celebrate the king's birthday, is a real festivity. According to several articles, many people also use drugs during this day. [bron] Not only during Koningsdag many people appear to use drugs also during the GayPride and New Year's Eve. The Gay Pride is a multi-day celebration with different events organized in several cities in the Netherlands. During Gay Pride, many Dutch people celebrate freedom of sexuality. This party has a gay cultural character and is known to involve a lot and also many different types of drugs. [bron]

Overall, the five events identified based on this news-article survey where drugs are commonly used are;

(1) Amsterdam Dance Event (ADE)
(2) Lowlands
(3) Kingsday
(4) Pride Amsterdam
(5) New Years Eve

4.1.2 Established time period. So for the research in this report, those five events are used as events with a high level of drug use of the visitors. Furthermore, the time period of each event is set to 1 week in which the event takes place. In this way the preparations for the event on the popularity of drugs are also included in the research. To carry out the research as automatically as possible, it is difficult to set a different time period for different events, such as ADE, which itself lasts a week, in which the preparations will also be included. Therefore it is decided to consider all events as one week events and Future Work will describe how the research could be adapted for longer events.

4.2 Selection of drugs

For this research in which drugs and specific events in the Netherlands are analyzed, it is important to determine the specific drugs under investigation. Because of the limited timeframe for this research and the variety of drugs in the Netherlands nowadays, it is chosen to select three types of drugs. If the outcomes of the three drugs contain interesting results or will serve other potential purposes, the research could also be conducted for multiple drugs and displayed in the final dashboard. More on this will be discussed in Future Work.

The three drugs that are used in this study and are visualized in the resulting visualization are based on the statistics from the National Drug Monitor. The National Drug Monitor (NDM) is a Dutch organization that collects and compiles all data on drug use, the drug market and drug-related crime of all ages in the population through different monitors. The purpose of this institution is to provide a clear representation of all the data monitored in the Netherlands related to drugs.

One of those representations is given in Figure X. This figure shows for each drug how much it has been used among the Dutch population in recent years. This figure shows that the most commonly used drug, with a significant difference, is cannabis. In addition, it can be seen that XTC is a commonly used drug in the Netherlands and cocaine is in third place. As can be seen in the figure, the line representing the drug Nitrous oxide (Lachgas) is interrupted because no results are known about this drug before that time. For that reason this drug is not included in the research at all. The other drugs visible in the figure that are not so often used are Heroine, Amphetamine, Paddo's, LSD, and GHB.

To investigate the popularity of drugs it is more interesting for the police to look at hard drugs that are illegal to possess and use than at the tolerated drugs for which the police cannot intervene. For this reason, although the soft drug cannabis is the most commonly used drug in the Netherlands, it is not investigated further.

As concluded above, the two most commonly used hard drugs in the Netherlands are XTC and cocaine. Because these are the most used hard drugs they are expected to be the most popular drugs in the Netherlands which can be interesting for the research.

Besides that, it is chosen to include a less used hard drug in the research. This decision is made to investigate whether less frequently used hard drugs are also less frequently searched on Google and appear in news articles than the commonly used hard drugs and whether those drugs are related to specific events (and how) or not. According to figure X from the National Drug Monitor, GHB is a rarely used drug in the Netherlands and therefore GHB is examined as the third drug. along with cocaine and XTC in this report.

So for this research, three drug types that are specified as cocaine, XTC and GHB are investigated, because it provides sufficient results for the Police Academy and fits within the time span for this research.

4.3 Proportion test results

To answer the main research question and the two subquestions 'How could we identify drug popularity at events?' and 'What are the differences between the specific events and drugs?', proportion tests have been performed on the google trends, twitter and NOS data for the three types of drugs. Thus, the mean proportion of peaks for all weeks from 2014 until 2022 has been compared to the proportion of peaks for specific drug events. For each test the null hypothesis (H0) stated: the proportion of peaks observed in drug related event weeks is not significantly larger than the proportion of peaks in regular weeks. The alternative hypothesis (H1) for each test stated: drug related event weeks have a significantly higher proportion of peaks compared to the regular weeks in the year.

4.3.1 XTC. The results show that for the drug XTC and the Google Trends data, King's Day (p < .001), New Year's Eve (p < .001) and Amsterdam Dance Event (p = .009) have rejected H0. However, Lowlands (p = .700) and the Amsterdam Gay Pride Event (p = .183) did not show a significant difference. Therefore, three out of five events have a strong significant difference with the mean proportion Google trends data for the drug XTC.

However, statistics for the Twitter data on the drug XTC show that for King's Day (p = .977), New Year's Eve (p = .461), Amsterdam Dance Event (p = .461), Lowlands (p = .211) and the Amsterdam Gay Pride (p = .760) H0 has not been rejected. Thus, Twitter data did not show any significant difference between the mean proportion of all weeks and the proportion of the weeks of all events.

Furthermore, the results show that the NOS articles also did not have a significant difference of peaks proportions. Since, King's Day (p = .264), , New Year's Eve (p = .887), Amsterdam Dance Event (p = .544), Lowlands (p = .544) and the Amsterdam Gay Pride (p = .887) all did not reject H0. Thus, the drug XTC proved to only have a strong significant relationship with the Google Trends data and the events King's day, New Year's Eve and Amsterdam Dance Event.

4.3.2 Cocaine. For the drug Cocaine the results do not show any significant outcomes for the Google Trends data. As King's Day (p = .067), New Year's Eve (p = .435), Amsterdam Dance Event (p = .205), Lowlands (p = .205) and the Amsterdam Gay Pride (p = .731) do all show p-values higher than 0.05. However, the p-value of King's Day (p = .067) is only slightly higher than the critical p-value.

Moreover, the results show that for Cocaine and the Twitter data, King's Day (p = .200), New Year's Eve (p = .745) and Amsterdam Gay Pride Event (p = .973) have not rejected H0. Amsterdam Dance Event (p = .055) and Lowlands (p = .055) show p-values quite low. Nevertheless, the p-values of ADE and Lowlands are not high enough to reject H0. Therefore, Cocaine does also not show a significant difference between the mean proportion of the Twitter data and the proportion of the events.

Lastly, statistics do not show that Cocaine and the NOS articles have a difference in the proportion of peaks with the proportion of peaks of any event. For King's Day (p = .561), , New Year's Eve (p = .373), Amsterdam Dance Event (p = .676), Lowlands (p = .373) and the Amsterdam Gay Pride (p = .952) all did not reject H0. Thus, Cocaine does not show any significant higher proportion of peaks between the different data sources and the events.

4.3.3 GHB. The test results for the drug GHB and the Google trends data show that for the event of King's Day (p = .011) it rejects H0. New Year's Eve (p = .205), Amsterdam Dance Event (p = .067), Lowlands (p = .970) and Amsterdam Gay Pride (p = .067) do not show any significant results. However, the p-value of both Amsterdam Dance Event and Amsterdam Gay Pride is only slightly above the critical value. Therefore, for the drug GHB, Google Trends only show a significant difference in peaks for King's Day.

Additionally, the statistics demonstrate that for GHB, Twitter does not have a significant difference in peaks with the events. Since, King's Day (p = .472), New Year's Eve (p = .979), Amsterdam Dance Event (p = .219), Lowlands (p = .472) and the Amsterdam Gay Pride (p = .219) did all not prove to be significant.

Finally, the NOS articles also do not show any significant results for the drug GHB. King's Day (p = .516), New Year's Eve (p = .516), Amsterdam Dance Event (p = 1.00), Lowlands (p = .516) and the Amsterdam Gay Pride (p = .216) all do not reject H0. Thus, out of all data sources, the drug GHB only shows a significantly higher proportion of peaks for Google Trends during King's Day .

In conclusion, the test results show that the subquestion 'How could we identify drug popularity at events?' can be answered by stating that Google Trends, with a score of four, has the highest number of significant proportions of peaks for the events in total. In contrast with Twitter and the NOS articles where none of the events demonstrated to have a significant proportion of peaks for all types of drugs tested.

Concerning the last subquestion 'What are the differences between the specific events and drugs?', there is a difference found between the specific events. The statistics of the proportion tests demonstrated that the events Lowlands and the Amsterdam Gay Pride never reject H0 for any type of drug tested, while King's Day, New Year's Eve and ADE did reject H0 for the drug XTC. Additionally, King's Day proved to have a significantly higher proportion of peaks for the drug GHB as well.

4.4 Prototype (dashboard)

Based on these requirements and the datasets a custom web-based visualization was created to allow the analyst to visualize explore, compare and filter the datasets. When filtering the charts update in real-time with keyframed animations. Screenshots and a live version of the prototype included in appendix 2a. For this prototype demo the user is logged-in with a default user account which is shown in the sidebar navigation. From there the user has four different overview pages to navigate to.

Events overview page

The first is an events overview page which shows a line chart with on the x-axis the week numbers per year and a relative score from 0 to 100 on the y-axis for our three datasets, the Google Trends, NOS News data and Twitter tweets. On the left bottom is a legend with a list of drug-related events. The user can click on an event which will highlight the week in which the event occured in the line chart to more clearly visualize peaks in the data which indicate event-based drug popularity. A small description of the event, most popular drug and event dates also is shown. The bottom-right shows the Google Trends data in a relative score over all 5 years of our dataset. This shows more clearly shows the lack of peaks of events in covid pandemic years.

Region overview page

The second page is a chloropleth map page showing the regions of the Netherlands. Each region is color coded using a linear color scale based (blue interpolation) on the relative score of Google searches for that region. On this page the user also has the option to focus on years and drugs but also on specific event dates. This for example shows that in Flevoland when the Lowlands event is happening there are alot of searches in that region for the drugs XTC.

Related queries page
The third page is the related queries page which shows a line chart for our three chosen drugs as filter options. The y-axis is a relative score from 0 to 100 and the x-axis are the week numbers for our specific year. The user can filter between different years. On the bottom are three polar charts which show the related drugs people are searching for with a relative score and a subset of the drugs in the legend. For example, people who search for cocaine also search for crack very often. It also shows that before 2019 GHB was not searched for by many people. It increased in popularity after that year.

Settings page
A settings page is also included. This page shows the different datasets used in the dashboard and a download button which allows the user to download a specific dataset either as .csv or .json to store on their computer.

4.4.1 Web application. The prototype is a web-based application using web standards and open-source software and libraries. It being web-based allow the dashboard to run operating system independent. Only a webbrowser installed on the users device is required. The application will mainly be used in a desktop environment by the analyst so the dashboard is not fully responsive and not mobile optimized. The web application is created with the open-source front-end framework Svelte [1] and UI framework SvelteKit which allows the application to be build in components, each chart is rendered seperately making it more efficient to add functionality (add datasets, render different chart types) in the future but also makes the dashboard performant when more data and charts are added since Svelte already pre-renders the page offloading work from the browser. Working in components and with a frameworks such as Svelte allows future web developers to get up and running fast and add more functionality in a progressively enhanced manner. Svelte can be downloaded as a module (package) from NPM [2] and uses the JavaScript back-end run-time Node.js [3].

For the charts the JavaScript charting library Chart.js [4] is integrated into the components which allows charts to be rendered in HTML5 Canvas without much configuration. With Chart.js you can add a specific dataset and the scales of the axis will automatically change accordingly to the scale defined by the data. The filter options and updating of the charts is more custom, it uses JavaScript utility functions to allow the data of the to be preprocessed and have only the data changed not the scales of the whole charts. For the map page an additional Chart.js Geo Plugin is used to render the Chloropleth map. It uses a TopoJSON file to render the regions of the Netherlands and filters the properties within that .json file to render a score for each of the region.

[1]https://svelte.dev
[2]https://www.npmjs.com
[3]https://nodejs.org/
[4]https://www.chartjs.org

The source code for the dashboard is published open-source on GitHub using the MIT license. A live version of the dashboard is contionously deployed on hosting platform Netlify [5]. Corresponding links can be found in appendix 2b.

4.4.2 Usability Testing.
During the project we on several occasions tried to arrange meetings with analysts from the police academy but due to planning problems it proved difficult to arrange a meeting at the location of the police academy to talk to actual analysts. Did this not prevent us from doing some initial and basic user tests on the beta versions of our dashboard.

The first version of our dashboard we tested among students that where part of our working group. A summative usability test was performed with 6 persons in total which where not involved in the development of the project. Which is enough to discover basic usability problems [1]. We observered and noted where they encountered problems using the Thinking Aloud method.

After incorporating this feedback iteratively in the prototype we did a second usability test with the Expert Review method. Two senior lecturers UX/UI from the Faculty of Digital Media and Creative Industries at the Amsterdam of University of Applied Sciences reviewed the interface of the prototype and navigated through all functionality.

Examples of incorporated feedback from the test where improving the visual design mainly by adding secondary colors to have sufficient color contrast, add descriptions and labels of the icons in the navigation to clarify the meaning of the pages and also improving the legends copy of the charts to make it more clear what scale the charts shows.

5 CONCLUSION

The research presented in this paper focused on discovering creative data sources that can be of help to the Police Academy when performing investigations on cases relating to drug usage. Three data sources were researched: Google Trends, Twitter and Dutch news (NOS). The hypothesis was that the data from these sources is representative of drug user behavior in The Netherlands.

The research consisted of multiple steps. First events known for increased drug usage were identified in order to serve as a proxy for increased drug usage. The dates and weeks of each event were identified between 2014 and 2022. Afterwards, drugs of interest we chose based on popularity in events and in The Netherlands overall. Then, using the same time range (2014-2022), the data sources Google Trends, Twitter and Dutch news (NOS) were scraped to generate data: content of searches, tweets and articles respectively. The data was aggregated and ordered chronologically to perform the analysis. Finally, the collected data was tested in order to discover whether an increased trend of the data was observed consistently during the weeks of the selected events compared to the rest of the weeks. A proportion test was performed for each combination of drug, event and data source.

The results indicated that among the three data sources, data from Google Trends best follows the increases observed in drug usage. The null hypothesis was rejected the most for drugs that

[5]https://www.netlify.com

are popular during events (XTC, GHB). The test failed for cocaine, which could be explained by this drug being popular over the entire year. In other words, an increased usage is not observed for cocaine during events, and the data from google trends also does not show significant peaks.

An additional conclusion from the results was that events that affect larger portions of the population (King's day, New Year's Eve, Pride, ADE) are better represented with Google Trends as opposed to events which are more specific to a smaller part of the country (Lowlands).

6 DISCUSSION

The research suggests that the Google Trends data has the highest similarity to the drug usage at events, compared to the Twitter data and NOS articles. However, there are some limitations that need to be considered when interpreting the results. First of all, the amount of data from the NOS articles was small. This resulted in only having a few articles that mentioned one of the drugs per week. The data was transformed into relative scores between hundred and zero, which caused the NOS articles scores to have very large differences between the data points. If the data size of the news articles were bigger, the timeline would probably have contained less peaks and therefore might have had better results on the performed proportion tests.

Another limitation of this research is the usage of weekly data. The timeframe of the data used is nine years. For a time period of more than five years Google Trends only shared weekly data, which obliged the research to use weekly data for all data sources. However, the events used did not have a timeframe of a week and therefore could fall on different days of a week. This could have affected the results, as an event that falls on a Monday or Sunday is more likely to also cause an increase of drug usage on another week than the one tested for. Moreover, in this research no tests were performed on delays shown in the data. For example, the Twitter data did not once show a significant difference of peak proportion for a week of an event. However, it could be discussed that drug events tend to be trendy on Twitter one or two more weeks after the actual date of the event.

Finally, one of the main limitations of this research is the lack of user availability testing. During the research no opportunity for meeting the end users of the dashboard has been given. Therefore, the usability of the prototype may have some shortcomings.

7 FUTURE WORK

7.0.1 Expand data sources. A subset of drugs and years was used for our research. To allow for more exploration and filterability the dashboard could be expanded with more datasets. Especially more relevant drugs (upcoming drugs, non-legal substances, NSPs) could be added as filter options. We also currently only use NOS news data but other more popular dutch news sources (e.g. RTL nieuws, Nu.nl) could be scraped to give a more accurate representation of the popularity of drugs mentioned in news articles. This will both improve the quality of the dataset by more accurately calculating a relative score as well as quantitify by gathering multiple sources and aggregating.

7.0.2 Real-time API data. Currently the dasboard relies on exported data that is then loaded into the web visualization. As a futher enhancement for the prototype and to make it more dynamic is to have the dataset exposed as an API which the dashboard can then fetch getting up-to-date real-time data. In this beta version datasets need to be added manually to the GitHub repository.

7.0.3 Usability testing. Further usability testing needs to be done to validate the User Experience (UX) and User Interface (UI) of the prototype to more accurately represent the needs of the police analyst. Basic user testing was done on a small group of people and feedback from the client was incorporated. But still, a lot of assumptions about the user have been made. User testing the live version of the prototype to a larger user base will uncover hidden problems. The feedback from the users would further validate the workings of the prototype.

In general further research is required to conclude that online data-gathering tools fare good indicators for predicting the popularity of specfic drugs.

8 ACKNOWLEDGEMENT

We thank supervisor Prof. dr. T.M. van Engers (University of Amsterdam) for providing guidance and assistance during the project aswell as helping us with all our questions. We would also like to thank Prof. dr. A. Barros (Police Academy) as a stakeholder which provided us with useful feedback troughout the project and asking relevant questions which helped us further expand our research.

REFERENCES

[1] Jakob Nielsen. 2002. *Designing Web Usability.* New Riders Publications Div Of Pearson, London.