Data Science Project

Substance use in young adults

Analysis Report

16. October 2020

Abstract

The data set was generously provided to me by the Jacobs Center for Productive Youth Development Zürich. Within the framework of the z-Proso study, more than 1000 persons were examined in a longitudinal study, from 7 to 20 years of age.

This analysis aims an explorative investigation of substance use and substance co-use among the examined persons at the age of about 20 years. According to the self-medication hypothesis, heavy substance use develops as a way of coping with problems in the absence of adequate solutions and meaningful social relationships.

A supervised learning algorithm was used to investigate whether the amount of polydrug and medicament co-consumption can be predicted based on social and educational problems.

Table of Contents

1 Project Objectives

The *Jacobs Center for Productive Youth Development* studies the social development of children and youths, with a particular focus on aggressive and delinquent behavior, substance consumption and prosocial characteristics. About 1400 children have been regularly followed since 2004, when they entered primary school. The most recent wave of data collection took place in 2018 at age 20. As part of this data collection, the hair of the participants was also analyzed at the *Center for Forensic Hair Analytics* in Zürich (Head of Unit: Dr. phil. II Markus Baumgartner). When analyzing chemical substances in the hair, a forensic toxicologist can recreate the history of interactions with chemicals, drugs and toxins. This way, the substance consumption of the last 3-6 months of each participant was determined.

This project aims to investigate the amount and frequency of drug and substance consumption in a representative sample of 1003 young adults around the age of 20. According to the self-medication hypothesis, substance use develops as a way of coping with stress in the absence of adequate solutions and meaningful social relationships [2]. I extend this hypothesis and assume, that especially polydrug use and medicament co-consumption is rather a strategy to reduce problems and bad feelings and goes beyond trial or recreational use [3]. Therefore, polydrug use will be predicted based on social and school problems during development in this analysis.

2 Methods

This dataset was generously provided to me by the *Jacobs center of Zürich*. The hair-concentration of several substances and their metabolites was measured with a procedure called *liquid chromatography tandem mass spectrometry (LC–MS)* in a *scheduled Multiple Reaction Monitoring (sMRM) Mode* by the team of the *Forensic Hair Analytics Center Zürich*.

The hardware of the computing device consists of an *Intel(R) Core (TM) i7-9700 CPU 3.00 GHz processor* with 32.0 GB RAM. A *Windows 10 System* with 64-bits was installed. The Data Analysis was carried out in a *Jupyter Notebook* (v6.1.1) set up in the *Anaconda Environment* (Anaconda3-2020.07-Windows-x86_64software distribution).

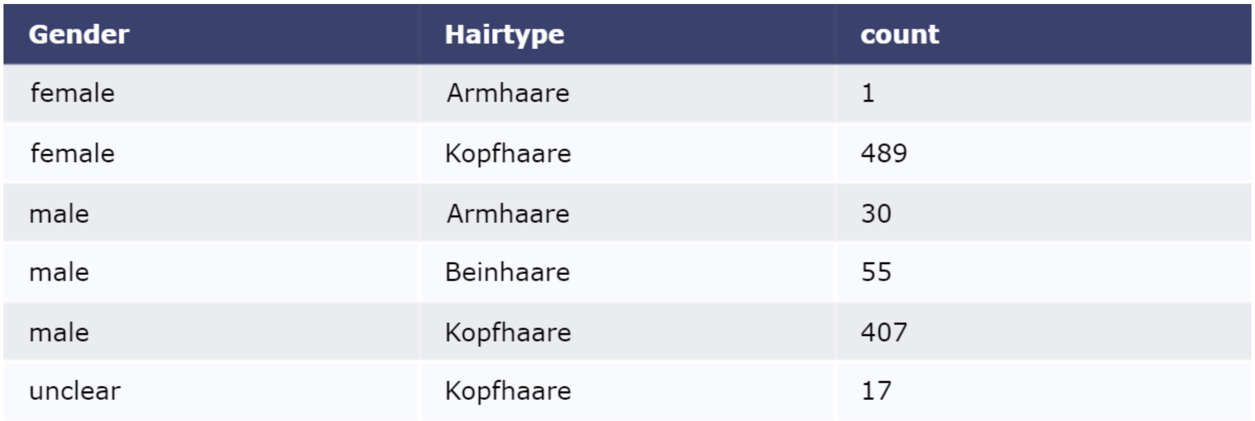
The python packages *pandas*, *numpy*, *matplotlib*, *scipy*, *seaborn* and, *scikit* learn were used for statistical analyses.

The analyses include some descriptive examinations of the frequency and amount of substance and drug consumption for different groups of substances and for co-consumption of drugs and medicaments. Based on the hair data, participants were grouped according to their use of drugs and medicaments (only antidepressants, sedatives and benzodiazepine included) in heavy vs. low substance users. In the next step, I tried to predict the substance use of participants around twenty based on the objective reports (by their teachers) of social and school problems during their past years and during their development.

3 Data

This analysis aims an explorative investigation of substance use and substance co-use of participants at the age of about 20 years. In addition, it will be investigated whether school and social problems in children and adolescents are predictive for heavy co-consumption of substances.

The original dataset consists of 6039 columns and 1003 rows. In the following, I only describe the variables from the hair samples and the variables I used for descriptive statistics and the prediction of substance co-consumption as a consequence of social difficulties of participants and problems at school.



Gender	Hairtype	count
female	Armhaare	1
female	Kopfhaare	489
male	Armhaare	30
male	Beinhaare	55
male	Kopfhaare	407
unclear	Kopfhaare	17

Figure 1: Overview of examined hair-type separated by gender.

For each of the 1003 participant, the concentration of following substances was measured in the hair:

- **morphine's** (and their metabolites 6-monoacetylmorphine and hydromorphone),
- **codeine** (and it's metabolite dihydrocodeine), and the codeine related morphine **dextromethorphan**, **oxycodone** (and it's metabolite oxymorphone),
- **synthetic opioids** (fentanyl, pethidine, tapentadol, tilidine, tramadol, dextromethorphan, methadone, buprenorphine),
- **cocaine** (and it's metabolites enzoylecgonine and norcocaine), and **cocaethylen** (a metabolite of cocaine that is formed by the liver when cocaine and ethanol coexist in the blood), and **levamisole** (a potentially brain damaging substance, used as a cutting agent in cocaine)
- **amphetamine** and **methamphetamine**,
- **MDMA** (and the structural similar chemical agent's **MDA** and **MDEA**, which are often an ingredient of Ecstasy pills and party drugs),
- **2C-B**, **4-FA** (both psychoactive drugs),
- **anti-ADHD treatments** (methylphenidate, modafinil, atomoxetine),
- **ketamine** (a dissociative anesthetic, often used as party drug),
- **antihistamines** (doxylamine and diphenhydramine),
- **anti-epileptic drugs** (pregabalin and lamotrigine),
- **cannabis** (the chemical compounds THC, CBN, CBD),
- **antidepressants**, one group of **SSRIs** (citalopram, fluoxetine, paroxetine, sertraline, trazodone) and one group of **SSNRIs** (venlafaxine, duloxetine, mirtazapine),
- **paracetamol**,
- **steroid hormons** (cortisol, cortisone and, testosterone),
- **benzodiazephine** (alprazolam, bromazepam, clobazepam, clonazepame, demoxepam, diazepame, nordazepame, oxazepam, temazepame, flunitrazepame, flurazepame, lorazepame, lormetazepame, midazolam, nitrazepame, phenazepame, prazepame, tetrazepame, triazolame)
- **sedativa** (zaleplon, zolpidem, zopiclone)

Following seven problems and social difficulties were reported by the teachers at nine different timepoints:

- **violation of school rules**
- **untidiness**
- **extortion**
- **verbal violence**

- **teasing**
- **physical violence**
- **vandalism**

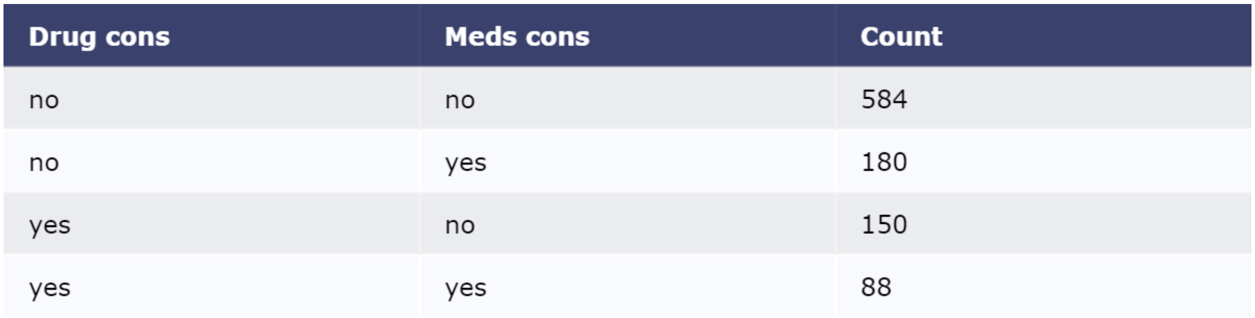
For security and anonymization reasons and for reasons defined in the participants' declaration of consent, this dataset cannot be provided. The analyses scripts, however, will be made freely available.

4 Data Analysis

⇨ 01_Prepare_Data.ipynb

1. Descriptive Statistics

To get an overview of the frequency drug and medicament consumption, some descriptive figures were created:



Drug cons	Meds cons	Count
no	no	584
no	yes	180
yes	no	150
yes	yes	88

Figure 2: Overview of cases with drug and/or medicament consumption

Note: see Supplementary Figure 1 for a furhter descriptive overview of drug and/or medicament consumption, including gender. Supplementary Figure 2 shows the intercorrelation diagram of co-use between substances.

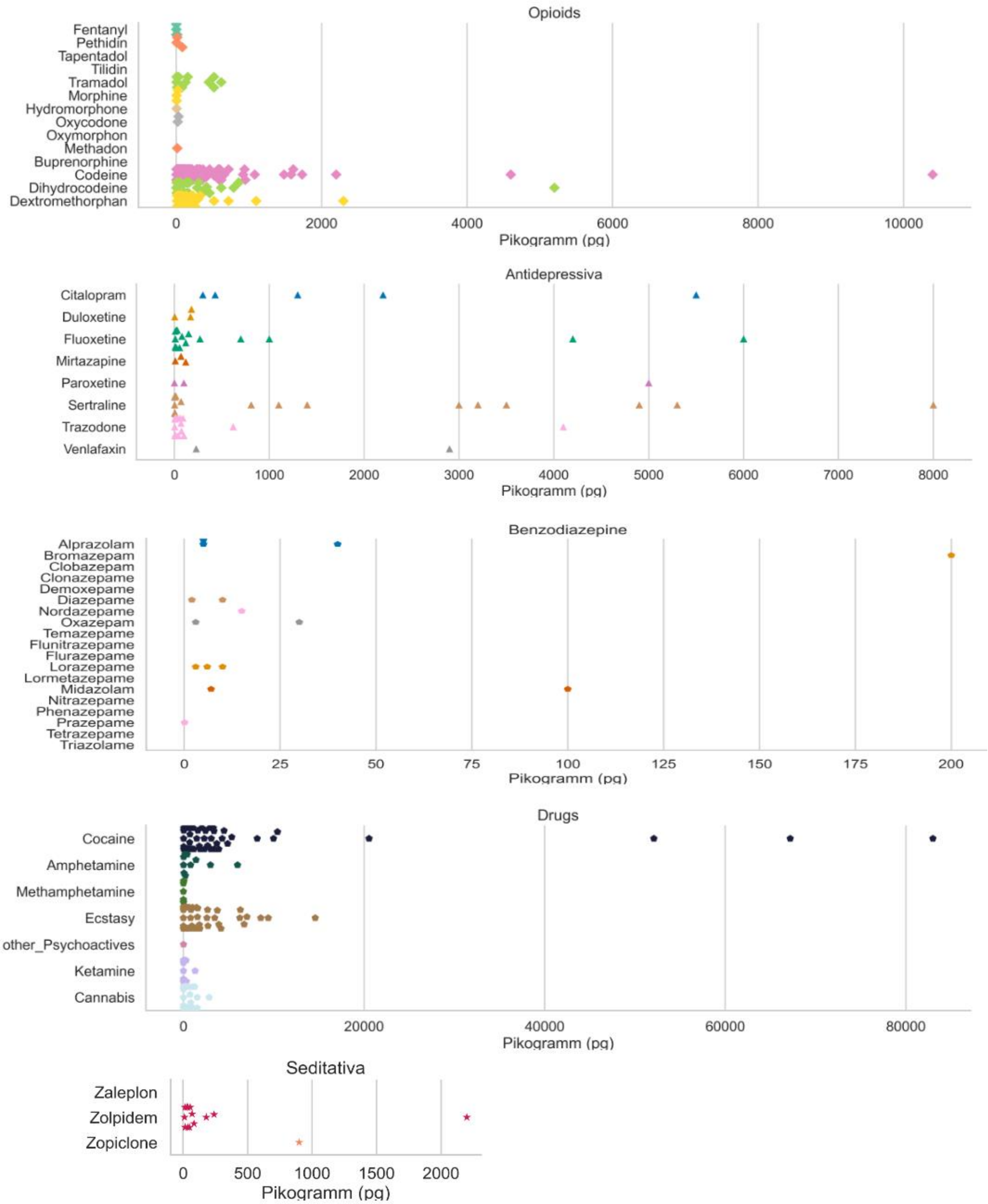


Figure 3: Overview of consumed substances. (Note: only values above zero are shown).

2. Analysis of co-consumption of substances

⇨ *02_Venn_Diagrams.ipynb*

- *Amount of Drug co-consumption among users*

For this analysis I used the pyvenn code from github developped by LankyCyril and tctianchi. I cloned the github repository from https://github.com/LankyCyril/pyvenn. Since there were only a small number of methamphetamine users and users of other psychoactive substances, I omitted these drugs from the Venn diagram because it would have affected the clearness of the diagram with 5 substances included. The variable "other psychoactives" was summarized together with Ecstasy since their chemical structure of the substances is similar. Cocaine was summarized together with Amphetamine.

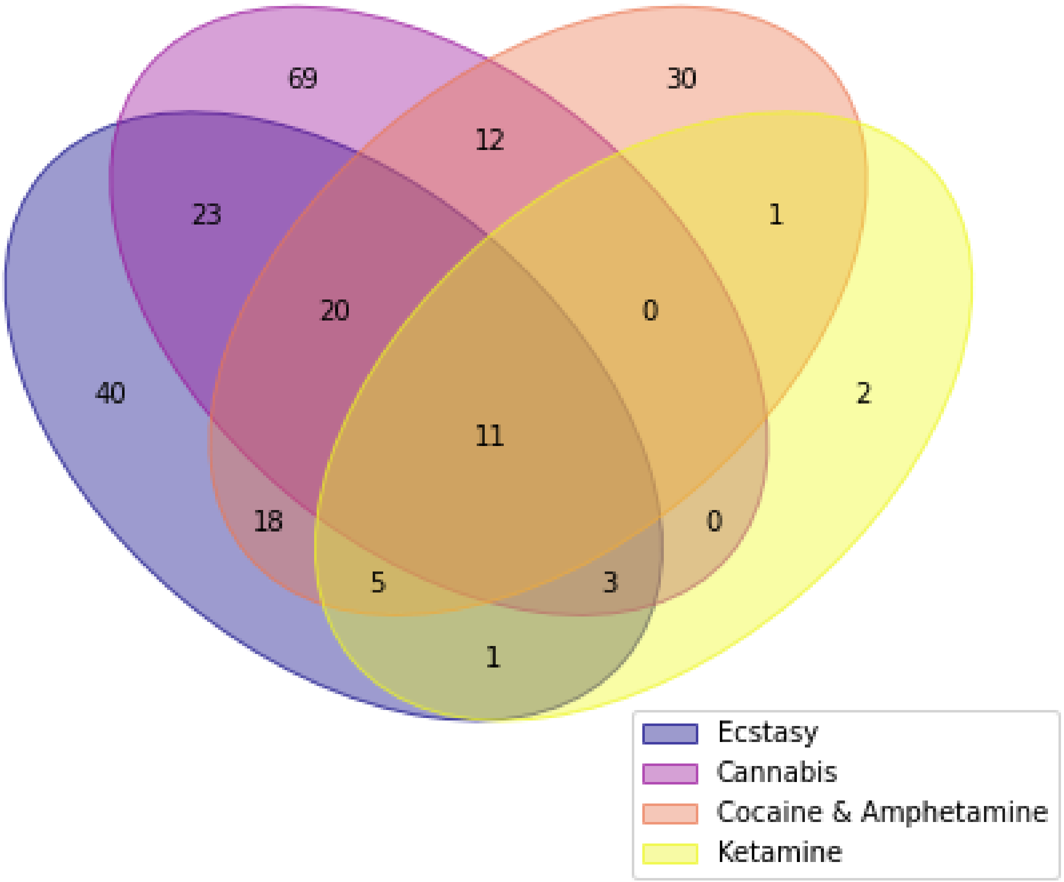


Figure 4: Amount of co-consumption between different drugs among users.

- *Amount of co-consumption of cocaine and alcohol and amount of the dangerous substance levamisole which is often used to stretch cocaine*

For this figure, I calculated the occurrences of each option (rows [i.e. participants] with only cocaine found in the hair, with the metabolite cocaethylen [that indicates cocaine – ethanol simultaneous consumption [4]], with cocaine and levamisole and with all three chemical compounds) and converted them into percent's.

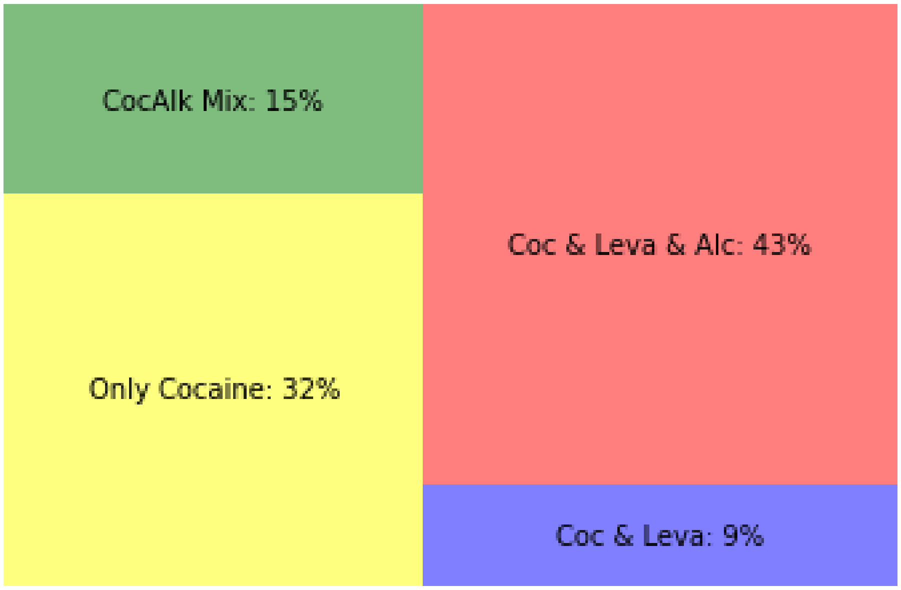


Figure 5: Percentage of cocaine users who either mixed cocaine and alcohol or consumed cocaine which was laced with levamisole or all three together.

Apparently, about half of the cocaine users among young adults consume cocaine that was laced with levamisole, even in combination with alcohol.

- *Amount of co-consumption of different substances*

Groups were made accordingly:

- Drug consumption (without cannabis)
- Cannabis consumption
- Antidepressants or Benzodiazepine consumption
- Other medicaments (Antihistamine, Epilepsy)
- Opioid consumption (including Codeine)

I excluded paracetamol and the sedatives from the analysis in order to make the Venn-diagram clearer. Because I did not consider paracetamol as a severe substance, compared with the other investigated ones and in addition, the consumption of this substance probably varies over seasons (with a higher amount of consumption in cold seasons). Sedatives are a group with only minor proportion of consumers (only 14 hair samples contained sedatives).

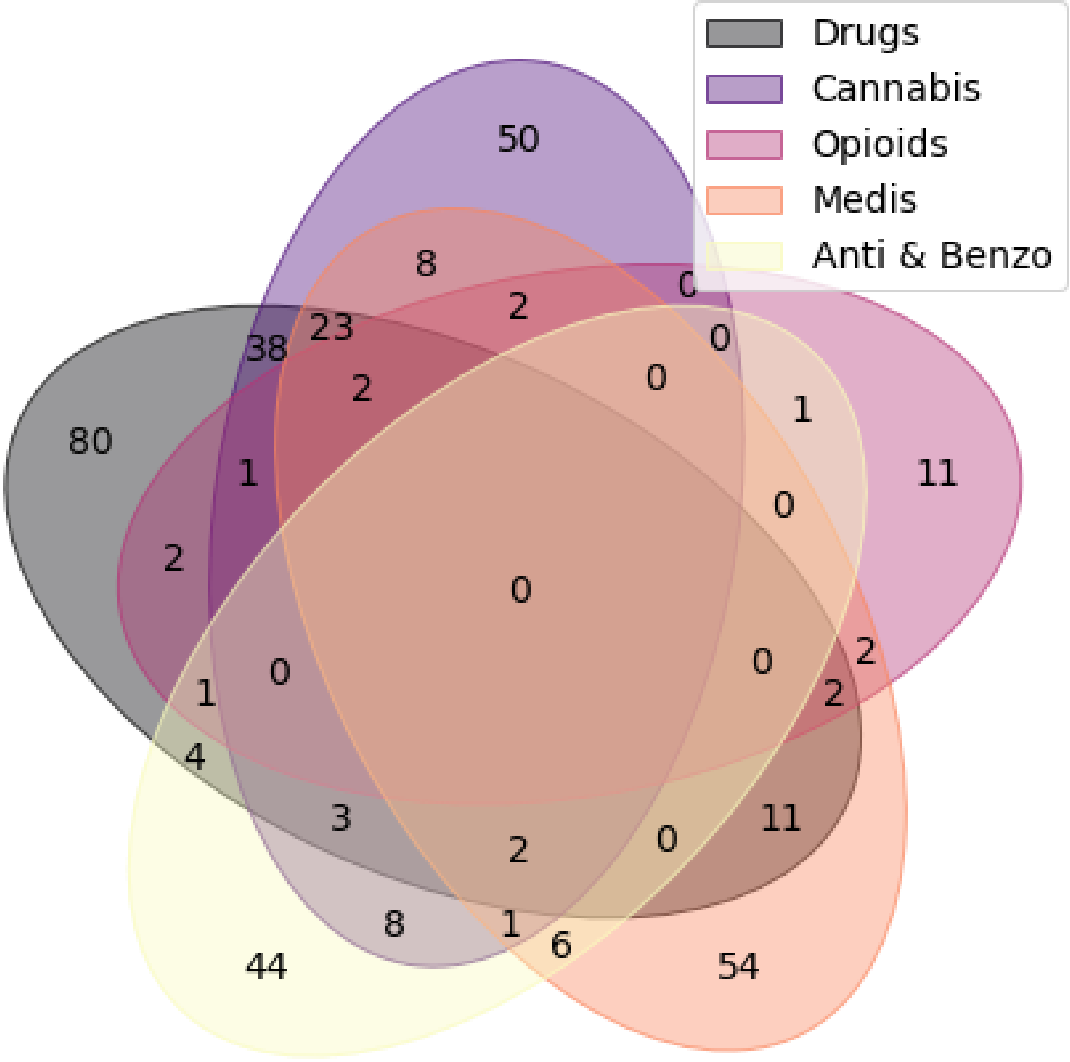


Figure 6: Co-consumption of drugs (without cannabis), cannabis, opioids (including codeine), medicaments (including epilepsy treatments, ADHD treatments and antihistamine treatments), antidepressants or benzodiazepines.

⇨ *03_Stats_prepare.ipynb*

3. Classification with support vector machines (SVM)

This analysis aims to investigate, whether polydrug use and medicament co-consumption can be predicted based on social and school problems during the past years. The reports are objective and made by the teachers of participants.

In a first step, I dropped the missing values from the SCALES.csv data frame (see section Data Flow for further information). Importantly, SCALES.csv contains only data from participant that used either Drugs or certain medicaments (namely: 'Opioids', 'Cocaine', 'Amphetamine', 'Methamphetamine', 'Ecstasy', 'Ketamine', 'Cannabis', 'Antidepressant', 'Benzodiazepine', 'Sedatives'). The remaining data frame consisted of 380 rows (i.e. participants) × 75 columns.

Second, I inspected, whether the data from SCALES.csv is normally distributed or not – this was not the case for almost all the columns (see Supplementary Figure 3). However, this is not a problem for the SVM algorithm.

In a third step, I assigned the subjects to user groups according to the co-consumption of different substances with heavy co-consumption (3 or more substances) vs. low co-consumption (less than 3 substances). Unfortunately, the user groups were very unbalanced with only 70 subjects in the heavy group and 310 subjects in the low user group. This will severely limit the results of the SVM classification. A good solution would be subsampling, but currently this is beyond the scope of this report. Therefore, I used a simpler solution: cutting the data frame by randomly dropping some of the low users for more equal group sizes.

⇨ *04_SVM.ipynb*

In the sklearn.preprocessing pipeline, I used the StandardScaler() function, which subtracts the mean and then scales the variables to unit variance by dividing the values by the standard deviation. Hence, the mean of the distribution is 0 afterwards. As features for the classifier I used all the 7 problem variables, which are longitudinally measured at 7 timepoints. This results in 49 features. As label, (the variable I want to predict), the assigned user group was included. The data was split into a training and a test set by size 0.2, hence 20% of the data was used for testing the trained learning algorithms.

With balanced data, the accuracy was 61% percent. The F1 score, which takes precision and recall into account, was around 60% as well. According to the confusion matrix, 8 out of the test datasets have been true negative and 9 have been true positive.

Balanced Data

```
                 precision       recall      f1-score       support

          High      0.67           0.53        0.59           15
           Mid      0.56           0.69        0.62           13

      accuracy                                 0.61           28
     macro avg      0.61           0.61        0.61           28
  weighted avg      0.62           0.61        0.61           28

[[8  7]
  [4  9]]
```

Using the uncut, unbalanced data gives the following results:

Unbalanced Data

```
               precision    recall  f1-score   support

        High       0.00      0.00      0.00        18
         Mid       0.76      1.00      0.87        58

    accuracy                           0.76        76
   macro avg       0.38      0.50      0.43        76
weighted avg       0.58      0.76      0.66        76

[[ 0 18]
 [ 0 58]]
```

The confusion matrix indicates a high number of false positives (18).

5 Metadata

Some metadata is stored in the data_varNames.RData file which was provided by the Jacobs Center. Obtaining the data for reproduction of the analysis necessitates a written request to me and/or the Jacobs Centrum Zürich.

6 Data Quality

Upon data inspection, the number of missing values was obviously small when compared to the total number of data points. From the 1002 rows, 1 missed the "Hairtype" and 3 missed "Gender". Gender was coded in three ways: 0.0 (17 cells), 1.0 (male), 2.0 (female). I defined the 17 cells with 0.0 as "unclear". Further, 2 out of all measured substances (Antidepressant and Paracetamol) had one missing value each. The "problems variable" contained 3 missing values each. The rows containing missing values in all of the "problems" were dropped from further analysis.

Single missing values from "Problems variables" were interpolated from the value of the same Problem measured at the previous timepoint using linear interpolation, forward method. Remaining missing values were interpolated with linear interpolation, backward method.

All substances, which were measured in the hair analysis, severely violate the normal distribution assumption. Within the 1003 participants, the proportion of polysubstance users is small and therefore, the issue of the non-normal distribution as well as the issue of unequal group sizes must be taken into account for statistical analyses between user groups. Out of those reasons, I decided to train a support vector machine algorithm which is at least deals well with non-normal distributed data.

7 Data Flow

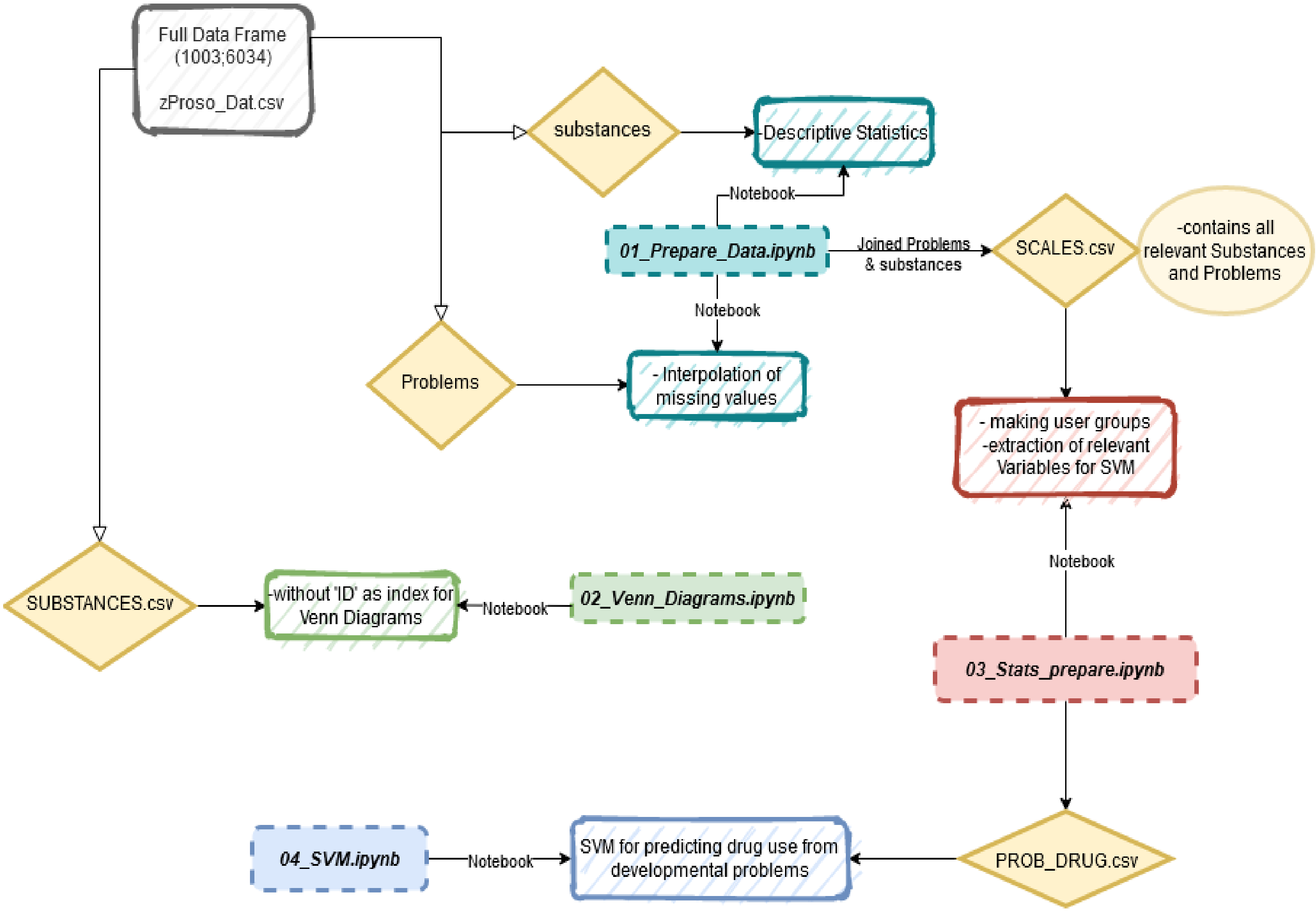


Figure 7: *orange framed rectangles*: data frames; *sketched background*: calculations in the notebook; *dashed frame*: name of the notebook.

8 Limitations

- Problem variables could have been better chosen
- The summarization of substances by simply adding the values might not be fully representative

9 Conclusions

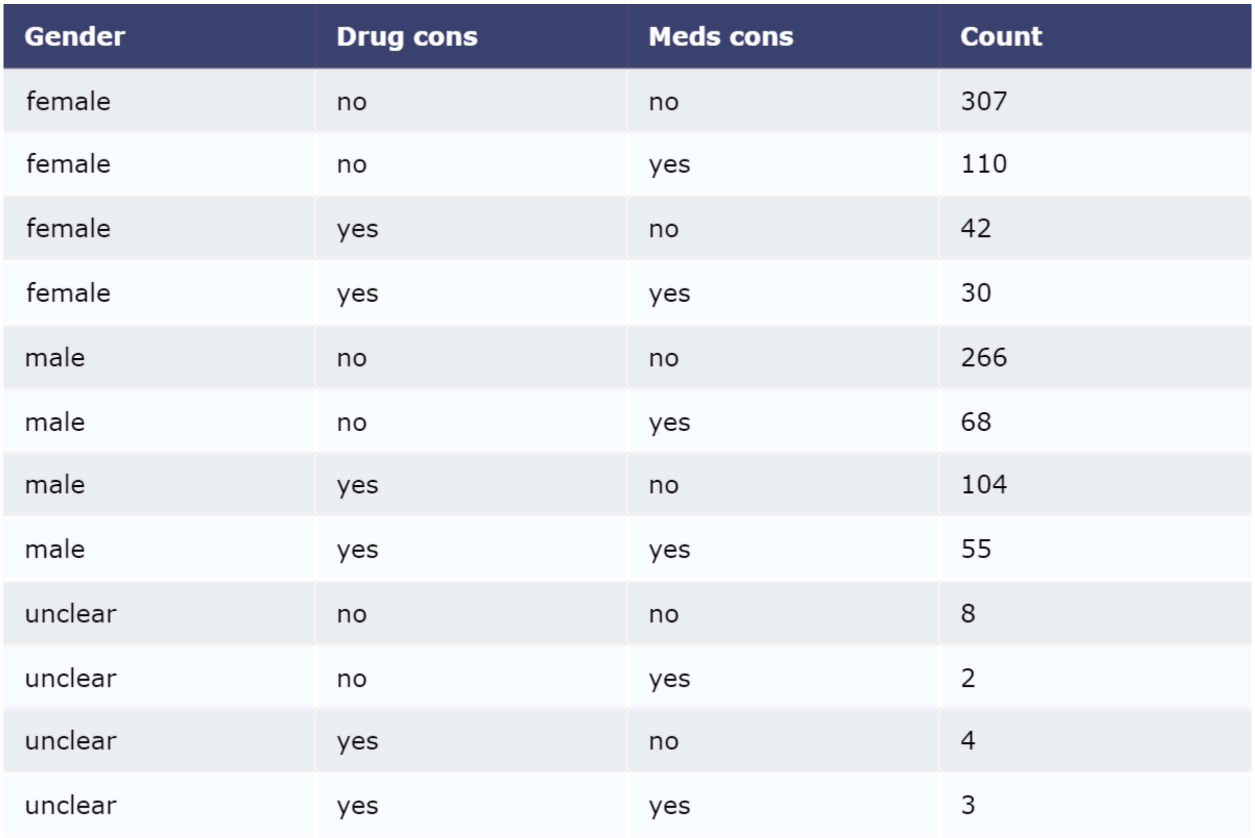
This short analysis examined the co-consumption of substances among young adults and, whether problems during development are predictive for later substance co-use. Among the 1003 investigated young adults, 180 hair samples indicated medicament consumption during the last 3-6 months and 150 hair samples indicated drug consumption. 88 hair samples pointed towards the consumption of both, medicaments and drugs. The most frequently consumed drug was Cannabis (in %), followed by Ecstasy and Cocaine or Amphetamine. Ketamine was only used by a minor proportion and most of the Ketamine users also consumed other drugs. In 11 participants, all four substances were found in the hair samples, which indicates a heavy polydrug use.

Cocaine was intaken by 95 participants (9.5%). In 58% of the Cocaine users, cocaethylene was also found in the hair, which strongly suggests the simultaneous consumption of alcohol and cocaine (because the metabolite cocaethylene is only produced by the liver when cocaine and ethanol co-exist in the blood [4]). In 52% of the cocaine using participants the animal anti-worming agent levamisole was found. Levamisole is a harmful substance that can have a toxic effect on the brain or even cause rotting skin [5,6]. In 43%, cocaethylene and levamisole were present together.

Acknowledgements

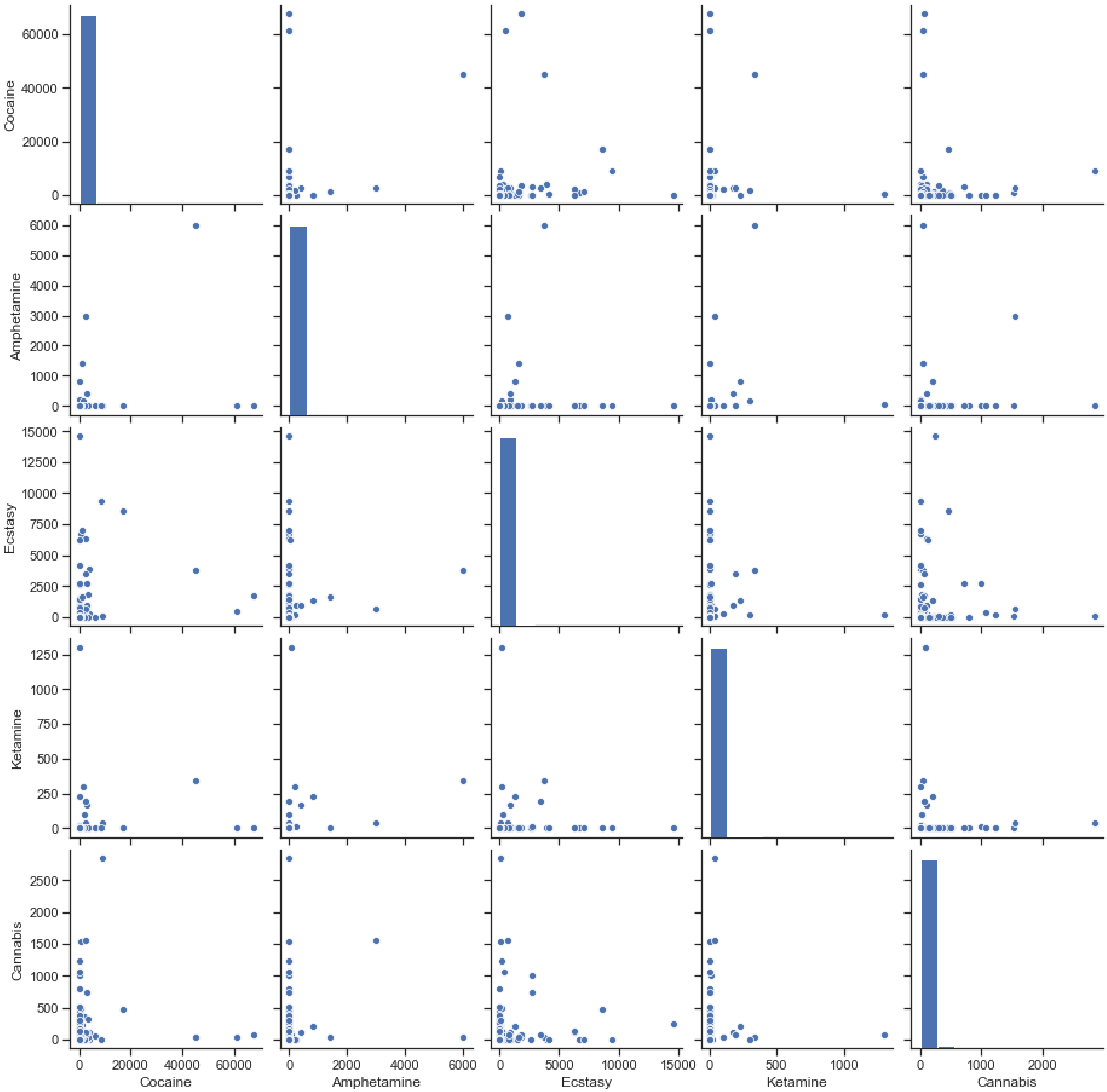
I warmly thank the Jacobs Center, Prof. Lilly Shanahan and Prof. Boris Quednow, who provided me the dataset for analysis.

Supplementary Figures

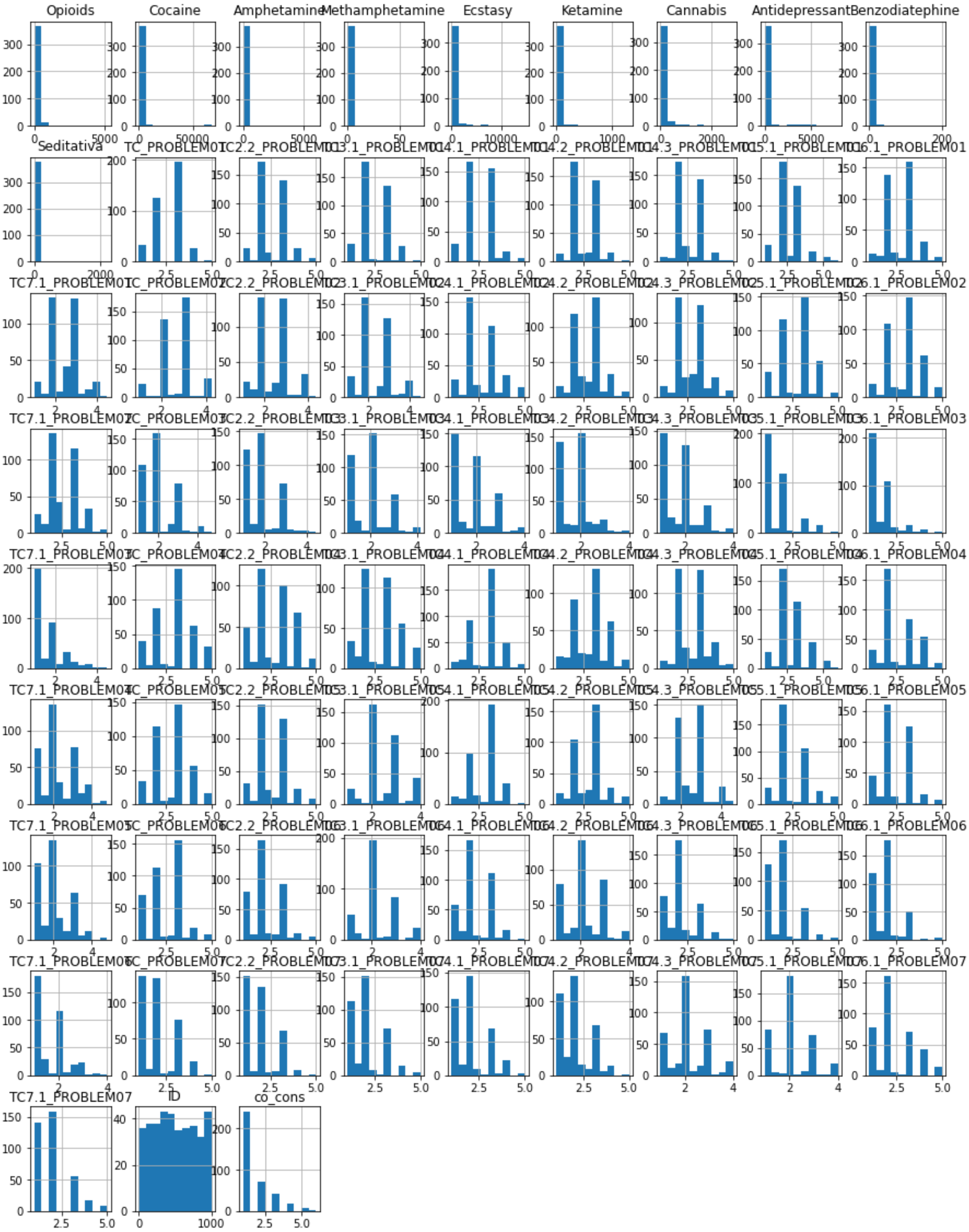


Gender	Drug cons	Meds cons	Count
female	no	no	307
female	no	yes	110
female	yes	no	42
female	yes	yes	30
male	no	no	266
male	no	yes	68
male	yes	no	104
male	yes	yes	55
unclear	no	no	8
unclear	no	yes	2
unclear	yes	no	4
unclear	yes	yes	3

Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3

References and Bibliography

[1] S. Haug et al., How to make a CDR, own brain, 2020.

[2] Frances, R. J. (1997). The Self-Medication Hypothesis of Substance Use Disorders: A Re. considerationand Recent Applications.

[3] Carlin, A. S., Stauss, F. F., Grant, I., & Adams, K. M. (1980). Drug abuse style, drug use type, and neuropsychological deficit in polydrug users. *Addictive behaviors*, *5*(3), 229-234.

[4] McCance, E. F., Price, L. H., Kosten, T. R., & Jatlow, P. I. (1995). Cocaethylene: pharmacology, physiology and behavioral effects in humans. *Journal of Pharmacology and Experimental Therapeutics*, *274* (1), 215-223.

[5] Zwang, N. A., Van Wagner, L. B., & Rose, S. (2011). A case of levamisole-induced systemic vasculitis and cocaine-induced midline destructive lesion: a case report. *JCR: Journal of Clinical Rheumatology*, *17* (4), 197-200.

[6] Lee, K. C., Ladizinski, B., & Federman, D. G. (2012, June). Complications associated with use of levamisole-contaminated cocaine: an emerging public health challenge. In *Mayo Clinic Proceedings* (Vol. 87, No. 6, pp. 581-586). Elsevier.