BEMM457: Topic in Business Analytics
Mini business analytics report

The correlation between health factors and behaviors influenced drug abusers

Presented by: Mr. Natchanon Leedara
Submission date: 17 December 2024

Table of contents

List of Figures

Introduction

One of the primary causes of global mortality is the overdosing of alcohol and drugs. Around 19.7 million people in the U.S. who are teenagers and adults are addicted to them, representing for the whole population as 8.1%. It has been increasing over the time due to the consuming overdose with combination of drugs (Shah et al., 2007), in which the most detrimental are Alcohol, Heroin and Cocaine, respectively (Nutt et al., 2007).

Even though the United Nations held conferences in 1971 to create legislation restricting drug use, this challenge persists due to uncontrollable factors that depend on individuals (Nutt et al., 2007). A model in Figure 1 (Durrant & Thakker, 2012) identifies the variables that directly influence individuals, including cultural-historical, psycho-social, and biological factors, which shape their personalities and behaviors. For instance, some people are influenced by those around them to view alcohol and drugs as a way to relieve stress.

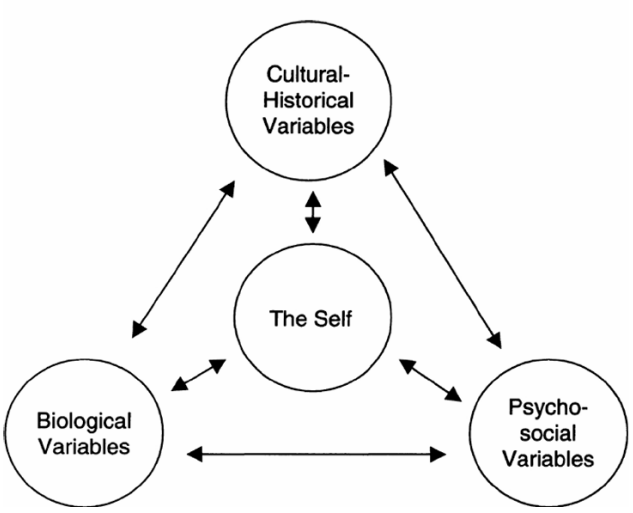


Figure 1: an integrated model of substance uses and abuses

There are many disadvantages for themselves and societies (Nutt et al., 2007), three main topics that involved with the drugs are physical health, psychological health and public community. Drugs will deteriorate the human body system which affects the neurological system to change the hormone levels, leading to self-harm due to uncontrollable emotion, resulting in increasing the risk of crime and suicide. Due to the consequences, they will be drifting apart from their families and people around them, which turns them into depressive patients in the future (Samet et al., 2012).

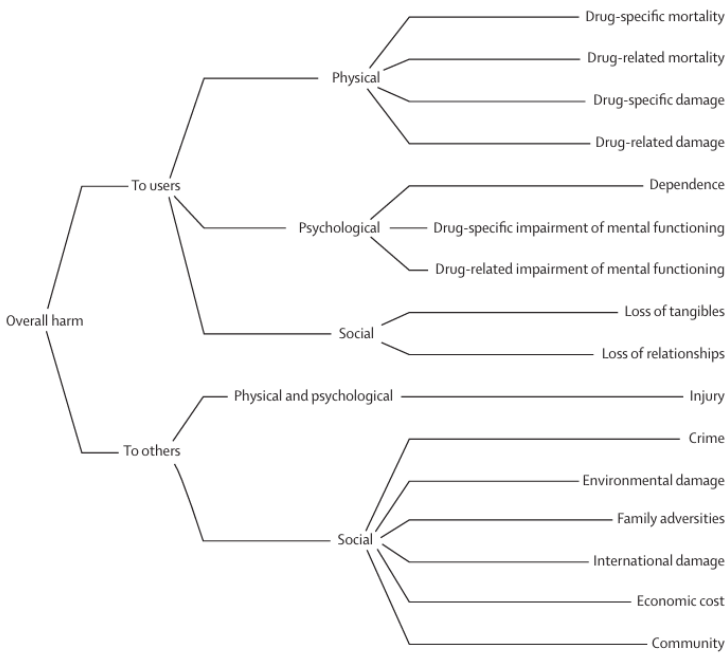


Figure 2: the harmful effects to addictive users and others

Currently, governments want to make a better policy to solve these problems, they need the support from the experts in the healthcare sectors to reduce the risk and help people who are addicted (Nutt et al., 2007). One successful project, called the "HELP Clinic," located in a detoxification unit, helped drug abusers quit by collecting data through subjective examinations and monitoring changes in their behavior. This success inspires my interest in conducting a mini project on drug and alcohol abuse.

Aim and Objectives

To understand the relationship between the evaluated health factors in that project, knowing the changes when the key factor has changed. Which can provide specific recommendations for governments to know the specific points that they need to concern.

Analytics Tools

There are two software tools that are utilized to analyze with Google Spreadsheet and Python (Jacquelune, K., 2016). The former utilized Pivot table to summarize the data to create some data visualization for descriptive statistics, and the latter used Python with Pandas, Seabron and Matplotlib to clean the data and do the diagnostic analysis.

.

Data Access, Ethics, Security, and Privacy

Data Access

The dataset was from the research (Samet et al., 2003) named "Linking alcohol- and drug-dependent adults to primary medical care: a randomized controlled trial of a multi-disciplinary health intervention in a detoxification unit". People who are addicted to alcohol, heroin, or cocaine detoxed to reduce using these substances, not experiencing a primary care doctor.

Data Ethics, Privacy and Security

Ethics, privacy, and security were carefully considered (Kaplan, 2016). Participants read the consent forms to understand the overall process of the research and signed it by themselves to join the study, and the project received approval from the Institutional Review Board of Boston University Medical Center. To protect their privacy, the researchers obtained a Certificate of Confidentiality from the U.S. Department of Health and Human Services. This certificate ensured that participants' personal data would not be shared, even under a court order. All data were anonymized, and participants could leave the study at any time, showing respect for their rights and autonomy.

Data Reliability and Data Validity

The data reliability and validity were excellent due to the standard subjective evaluation, such as the short-form (SF-36) health survey (Sima et al., 2021)(Pan & Barnhart, 2016), which measures participants' physical and mental health, and the Risk Battery Assessment (RAB) (Psychiatry, n.d.) which assesses whether people have any risks related to drugs or sex. Moreover, they decided to use the randomized controlled trial (RCT) design, which is one of the most effective research designs, to decrease bias. Additionally, they used "intent-to-treat" analysis (Gupta, K. S.,2011), which means included all participants in the final results, even if some dropped out. To reduce any errors, self-reported data were compared with official medical records.

Data Nature

The data were collected over 24 months, with interviews taking place at 6, 12, 18, and 24 months. The study used both quantitative and qualitative data, such as self-reports about medical care, addiction severity, and quality of life, as well as medical records. The researchers classified the various results by the substance groups that patients utilized (Alcohol, Heroin, Cocaine). The way to connect for receiving the responses of self-reported

assessment is the most challenging part; they solved this by sending reminders before the evaluation for 1 month, offering transport to pick them up, and repeatedly following them. Although the study was limited to a single site and may not apply to other settings, it showed a clear effort to handle ethical concerns, respect participants, and produce reliable and useful data.

According to the objective of this report, the information that had been selected to do the analysis were the demographic data (age, sex), ethnicity, housing status, serious thoughts of suicide and responses for taking treatment substances for descriptive analysis and the evaluation scores of physical and mental health; including SF-36 components (Sima et al., 2021)(Pan & Barnhart, 2016)which are Physical Component Scores: PCS, Mental Component Scores: MCS), The Center for Epidemiological Studies-Depression: CES-D) and the factors (how long staying at the hospital, the average of drinks, the number of friends that support them, RAB scores) being relevant to their health for descriptive analysis and diagnostic analysis to see the correlation between variables.

Data Cleaning

Structure Overview

The dataset was imported by Python with Pandas library in JupiterNotebook. The table structure information, there are 29 columns which all the variables will be demonstrated in the Appendix A. All the data is about the demographic information of substance addicted abusers, social behavioral factors, health and biological factors and substance usages.

Cleaning Process

The first step of cleaning the data is to utilize Pandas library to count the units and the percentage of missing values. If there are any percentages that are over 40%, the columns dropped due to the error interpretation in analysis (Rao et al., 2023)(Eekhout et al., 2012). Columns named “link”, “d1”, “i1”, “i2”, “id”, “treat”, “max_drinks”, “anysub”, “daysanysub”, “e2b”, “dayslink”, and “indtot” were removed because of few having null values more than 50% and others which would not be used to analyze.

With the less percentage below 10 to 20% (Lin & Tsai, 2020)(Rao et al., 2023), the value imputation would be the next step for data cleaning. (Median/Mode value for numerical variables, Mode value for categorical variables). Columns named “drugrisk”, “mcs” and “pcs” had imputed by median value of their columns, “sexrisk” imputed by mode value and replaced the word “missing” in substance column by mode as well.

Creating new columns with the combination of scoring of drug related (drugrisk) and sex related score (sexrisk) as “rabscores” column as the risk assessment battery scores (Psychiatry, n.d.), which the equation is rabscores = (drugrisk + sexrisk) / 40, resulting in score range (0 – 1).

Data Preparation for Further Analysis

Separated into three csv file were categorized by substances (alcohol_group, heroin_group and cocaine_group), in order to check the outliers using the IQR outlier’s removal methods (Rao et al., 2023)which had the evidence that great improvement for the correlation analysis. The result show the boxplot of data in the Appendix B, C and D.

Finally, all data in three csv files would be duplicated into the spreadsheet in order to do the descriptive analysis and data visualization.

Descriptive Analytics

Categorical Variables

The information from all substance addiction categories is analyzed, the alcohol category has the highest population, followed by the cocaine and heroin groups. With an average age of around 35, the majority of those who struggled with substance abuse were middle-aged. According to the gender, men are more commonly affected than women among three groups. Regarding ethnicity, Black people represent the largest proportion of addicted individuals, particularly in the cocaine group, while the other groups was white ones. Most alcohol abusers are notably more likely to be homeless, while others tend to live in their own homes. Last but not least, most individuals reported refusing substance treatment and denied having attempted suicide within the past 30 days.

Numerical Variables

The SF-36 evaluation (Sima et al., 2021) showed data about the quality of mental and physical health of abused people. The Physical Component Scores (PCS) indicate the outcome of physical ability which gets higher that mean having a excellent physical performance, each group had the mean of score slightly lower than the average scoring methods at approximately 50 to 60. Meanwhile, the cocaine group shows slightly better PCS scores; however, this does not suggest a healthy physical condition overall, as their scores are still not indicative of optimal health.

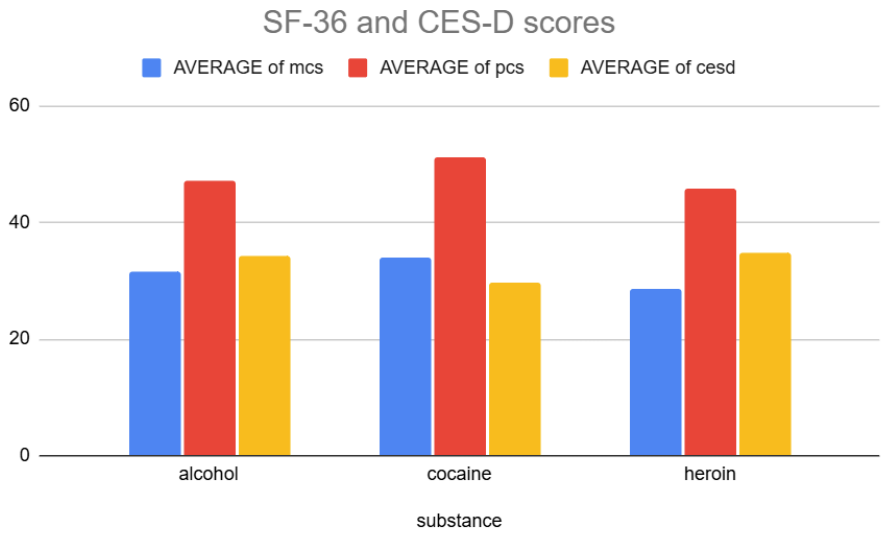


Figure 3: The average of SF-36 and CES-D score components

On the other hand, the Mental Component Summary (MCS), which gets higher that means having an excellent mental health score for all groups, are significantly lower than average scoring methods (average around 50-60). This would highlight the negative impact of drug abuse on people's mental health. In addition, there was another analysis of CES-D evaluation to see patients that have any risks to have a depression disorder (American Psychological Association, n.d.) found that the average of scores were over 29, which assessment who have score more than 16 suspecting them to have a depression disorder. This would support the evidence of MCS results having a harmful effect for them.

The RAB scores (Psychiatry, n.d.) for alcohol, heroin, and cocaine groups show good average results, ranging from 0.1 to 0.24. Among these, the heroin group scored the highest, indicating a relatively low risk of drug and sex-related abuses. These findings are based on evaluations conducted during the intervention process; this might state that the restriction of utilizing the addictive substances makes them feel uncomfortable.

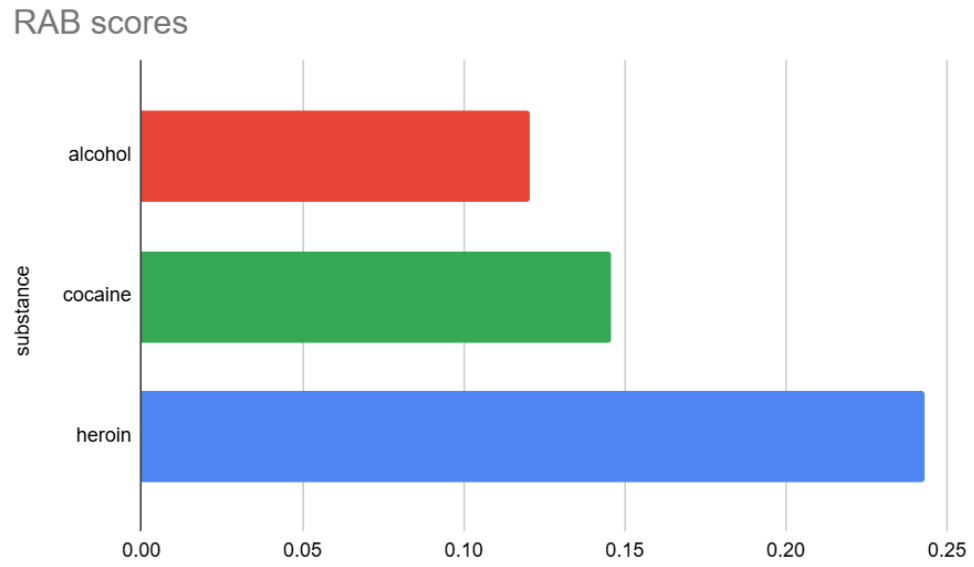


Figure 4: RAB scores chart

As for the other two factors, hospitalization duration for medical problems was longest among individuals in the alcohol group compared to the others. Meanwhile, the level of support from friends was consistent across all groups.

Diagnostic Analysis

Negative Correlations

The strongest inverse relationship between depression (CESD) and mental health (MCS) is seen across all addiction categories (alcohol, heroin, and cocaine). This connection, which varies between -0.65 and -0.69, shows that people with worse mental health are more likely to have higher levels of depression, highlighting the importance of mental health in substance addiction. The association between PCS and hospitalizations in the Cocaine group is negative (-0.32), indicating that hospitalizations are more common among those with poorer physical health. Furthermore, in the Alcohol group, average drinks and physical health (PCS) have a moderately negative connection (-0.28), reflecting that people who are less physically well-off tend to drink more. Besides, there is a minor negative correlation between friend support (PSS) and depression (CESD) in the Heroin group (-0.16), indicating a relationship between increased healthy connection from the friends and decreased depressed symptoms.

Positive Correlations

Hospitalizations demonstrate a consistent and slight positive correlation with age across all three addiction categories, with correlation values between 0.16 and 0.19. This clearly indicates that older individuals are more likely to be admitted to the hospital, irrespective of the substance involved. There is also a little positive connection (0.21) between hospitalizations and average drinks in the Alcohol group, suggesting that higher alcohol use is associated with more frequent hospitalizations. In the Heroin group, there is a considerable positive connection (0.29) between RAB scores and depression (CESD), suggesting that lower behavioral resilience may be associated with higher depression levels. Finally, there appears to be a connection between addictive alcohol intake and depression (CESD) in the Cocaine group, as seen by the moderate positive correlation (0.24) between average drinks and depression.

To summarize the overall correlations, the positive one shows how significant aging, admitting to the hospital, and having drug-related risk and sex-related risks, whereas the negative relationships highlight somatic and mental health, primarily mental among the addiction groups.

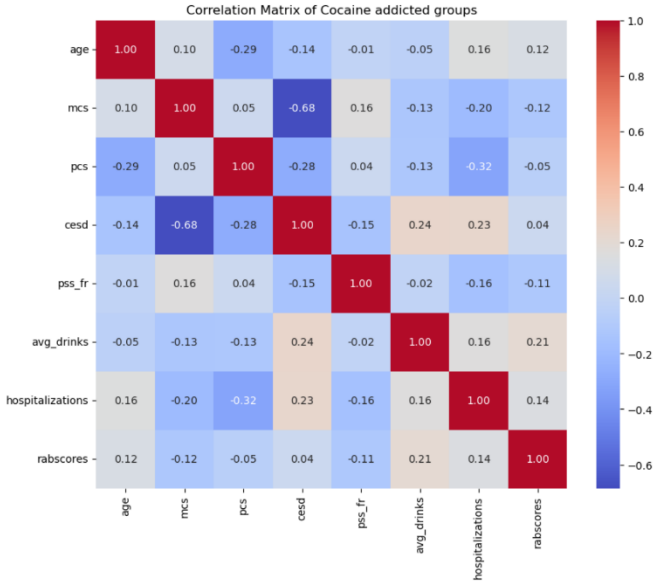


Figure 7: The Correlation Matrix of Cocaine Group

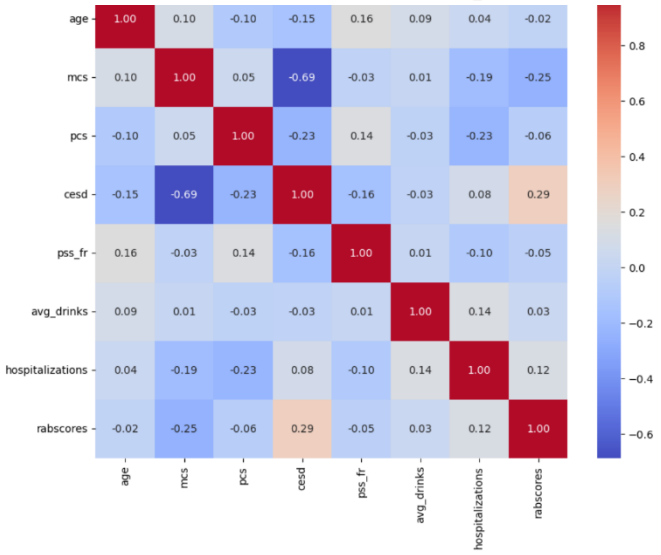


Figure 5: The Correlation Matrix of Heroin Group

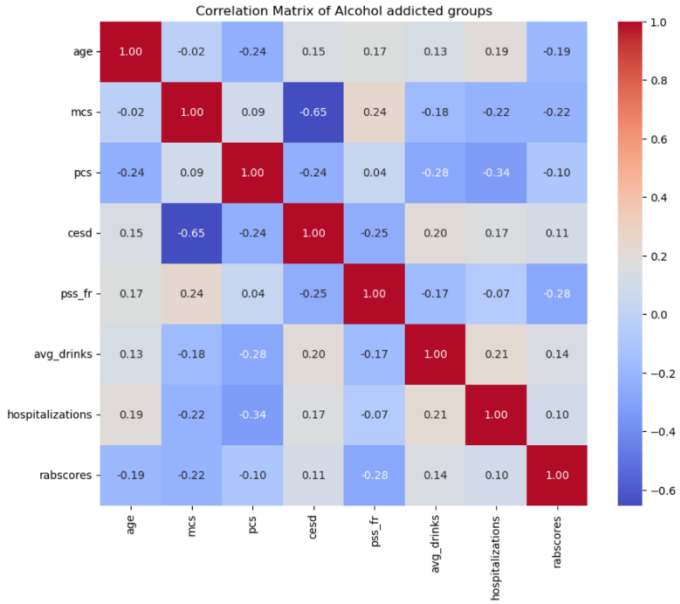


Figure 6: The Correlation Matrix of Alcohol Group

Recommendation

As there was a significant correlation between the MCS and CES-D in all groups, the first priority for governments must be to improve mental health to decrease the risk of depression, which was also moderately related to a decreased risk of drug utilization and sexual infections (RAB). Governments should invest more in rehabilitation and psychological healthcare facilities for abusers, including opening therapy sessions with therapists and others (PSS_FR). These sessions can help create a supportive community, reminding individuals that they are not alone. Moreover, governments should encourage families to pay more attention to drug users and support them in overcoming their struggles, due to slight correlations with the CES-D.

Another one, the PCS scores showed a negative relationship with the average alcohol consumption and hospital admissions, health knowledge sharing can help address physical health problems. By initiating projects that educate people about the adverse effects of drugs and alcohol across all age groups and promoting exercise for better health, individuals can develop greater awareness of the dangers of consuming addictive substances in their daily lives. Furthermore, there was evidence (China University, 2023) that aerobic exercise can improve the physical ability and mental illness of them, to create the campaign about enjoyable activities which not only improve health of addicted people but also can improve the relationship between people in societies.

Conclusion

This mini project focused on the correlation between the factors related to substance addiction, for instance, physical and mental health. To provide possible suggestions for the governments to find the propriate solution to address the social challenges. The information was analyzed by efficient software (Spreadsheet, Python) which was conducted in the descriptive analysis to understand the overall dataset what it should bring to analyze and diagnostic analysis to highlight the importance of trends within three substance groups.

The significance of findings was health, that there was a significant connection between risky behaviors and both physical and mental health, slightly relevant to the surrounding environment. These insights can underline the important focus that the governments should be noticed, leading to the prevention and rehabilitation strategies for fixing these problems

Limitations

There were some limitations of this project, the first of which was the small dataset scope. This information might not demonstrate the diversity of addictive patterns over the different populations, expanding the dataset should improve the suggestion to be generalized for bigger population. Another point is this project focusing on the correlation not the causation of the problem, and along with lack of predictive analysis. Consequently, these would make the recommendations to be less effective, affecting the designs of invention strategies.

Reflections

The business analytics project was super intensive which made me have a tough time throughout the process of analysis. However, all the experience in each operation from the beginning till the end have embraced me to be a competent business analyst in the future by knowing the strengths and weaknesses of myself.

At the beginning of the process, it was quite hard to think about the topic to do the business project. As I was a physical therapist before and I have just seen the topic could be about social challenges, I chose the topic of drug abuse, which I get familiar with the context, and I want to improve my critical analysis to understand the factors that affect patients.

While searching for the dataset to be used in this project, I found it for a few days. It reflected that even though I stopped being in the part of healthcare practitioners, I still had methods using the finding key words from the experimental research in physical therapy. Furthermore, University of Exeter's Library website is one of the most useful for me to find out any research to prove the evidence, I would keep utilizing it in the future.

On the other hand, the technical approach, such as using Python for coding to clean the data, was the weakest point in this project. I spent a lot of time coding to extract the data to prepare for analysis in the next step due to none of the experience of programming project. It slightly suffered to finalize the project, and the interpretation of both analyses. As a result, the project had been stopped because I had to learn the coding lesson to know how to run the code effectively without any errors to interrupt the process.

After summarizing the overall project, I thought that I have only the from healthcare knowledge which was not enough to create effective analysis to give the recommendations for the governments. This situation made me searching for other perspectives from online platforms and the academic papers to be more understand on the topic.

In the future, I would like to try other topics that is not related to the health research. As I need to improve my knowledge in every aspects; such as business, finance, or IT management. And the analysis of the data, I would go beyond the diagnostic analysis to do the predictive analysis for the effectiveness of findings.

References

1. Shah, N. G., Lathrop, S. L., Reichard, R. R., & Landen, M. G. (2007). Unintentional drug overdose death trends in New Mexico, USA, 1990–2005: combinations of heroin, cocaine, prescription opioids and alcohol. Addiction, 103(1), 126 136. https://doi.org/10.1111/j.1360-0443.2007.02054.x
2. Nutt, D. J., King, L. A., & Phillips, L. D. (2010). Drug harms in the UK: A multicriteria decision analysis. The Lancet, 376(9752), 1558-65. doi: https://doi.org/10.1016/S0140-6736(10)61462-6
3. Durrant, R. & Thakker, J. (2012). Substance use & abuse: cultural and historical perspectives. https://sk-sagepub com.uoelibrary.idm.oclc.org/book/mono/preview/substance-use-and-abuse.pdf
4. Samet, J. H., Larson, M. J., Horton, N. J., Doyle, K., Winter, M., & Saitz, R. (2003). Linking alcohol- and drug-dependent adults to primary medical care: a randomized controlled trial of a multi-disciplinary health intervention in a detoxification unit. Addiction (Abingdon, England), 98(4), 509–516. https://doi.org/10.1046/j.1360-0443.2003.00328.x
5. Kaplan, B. (2016). How should health data be used?: CQ. Cambridge quarterly of Healthcare Ethics, 25(2), 312-329. https://doi.org/10.1017/S0963180115000614
6. Sima, R. M., Pleş, L., Socea, B., Sklavounos, P., Negoi, I., Stănescu, A. D., Iordache, I. I., Hamoud, B. H., Radosa, M. P., Juhasz-Boess, I., Solomayer, E. F., Dimitriu, M. C. T., Cîrstoveanu, C., Şerban, D., & Radosa, J. C. (2021). Evaluation of the SF-36 questionnaire for assessment of the quality of life of endometriosis patients undergoing treatment: A systematic review and meta analysis. *Experimental and therapeutic medicine*, *22*(5), 1283. https://doi.org/10.3892/etm.2021.10718
7. Psychiatry, P. (n.d.). Risk assessment battery (RAB). University of Pennsylvania. https://www.med.upenn.edu/hiv/rab.html
8. Jacquelune, K. (2016) *Data wrangling with python.* O'reilly
9. Pan, Y. & Barnhart, X. H. (2016). Methods for assessing the reliability of quality of life based on SF-36. *Statistics in medicine, 35*, 5656-5665. https://onlinelibrary-wiley-com.uoelibrary.idm.oclc.org/doi/epdf/10.1002/sim.7085

10. Gupta, K. S. (2011). Intention-to-treat concept: a review. *Perspectives in clinical research, 2*(3), 109-112. https://europepmc.org/article/pmc/3159210
11. Rao, K. M., Saikrishna, G., & Supriya, K. (2023). Data preprocessing techniques emergence and selection towards machine learning models: A practical review using HPA data set. *Multimedia Tools and Applications, 82,* 37177–37196. https://doi.org/10.1007/s11042-023-15087-5
12. Lin, W.-C., & Tsai, C.-F. (2020). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review, 53*(2), 1487–1509. https://doi.org/10.1007/s10462-019-09709-4
13. Eekhout, I., de Boer, M. R., Twisk, J. W. R., de Vet, H. C. W., & Heymans, M. W. (2012). Brief report: missing data: a systematic review of how they are reported and handled. epidemiology, 23(5), 729–732. http://www.jstor.org/stable/41739653
14. American Psychological Association. (n.d.). *Center for Epidemiological Studies-Depression*. APA. https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/depression-scale
15. China University of Geosciences Researchers Release New Data on Drug Abuse (Intervention Effect of Aerobic Exercise on Physical Fitness, Emotional State and Mental Health of Drug Addicts: A Systematic Review and Meta-Analysis). (2023, February 27). *Mental Health Weekly Digest*, 121. https://link-gale-com.uoelibrary.idm.oclc.org/apps/doc/A738664070/AONE?u=exeter&sid=bookmark-AONE&xid=ef09ca35
16. Samet, S., Fenton, M. C., Greenstein, E., Aharonvich, E., & Hasin, D. (2012). Effects of independent and substance-induced major depressive disorder on remission and relapse of alcohol, cocaine, and heroin dependence. *Addiction research report, 108*, 115–123. https://doi.org/10.1111/j.1360-0443.2012.04010.x

Appendix

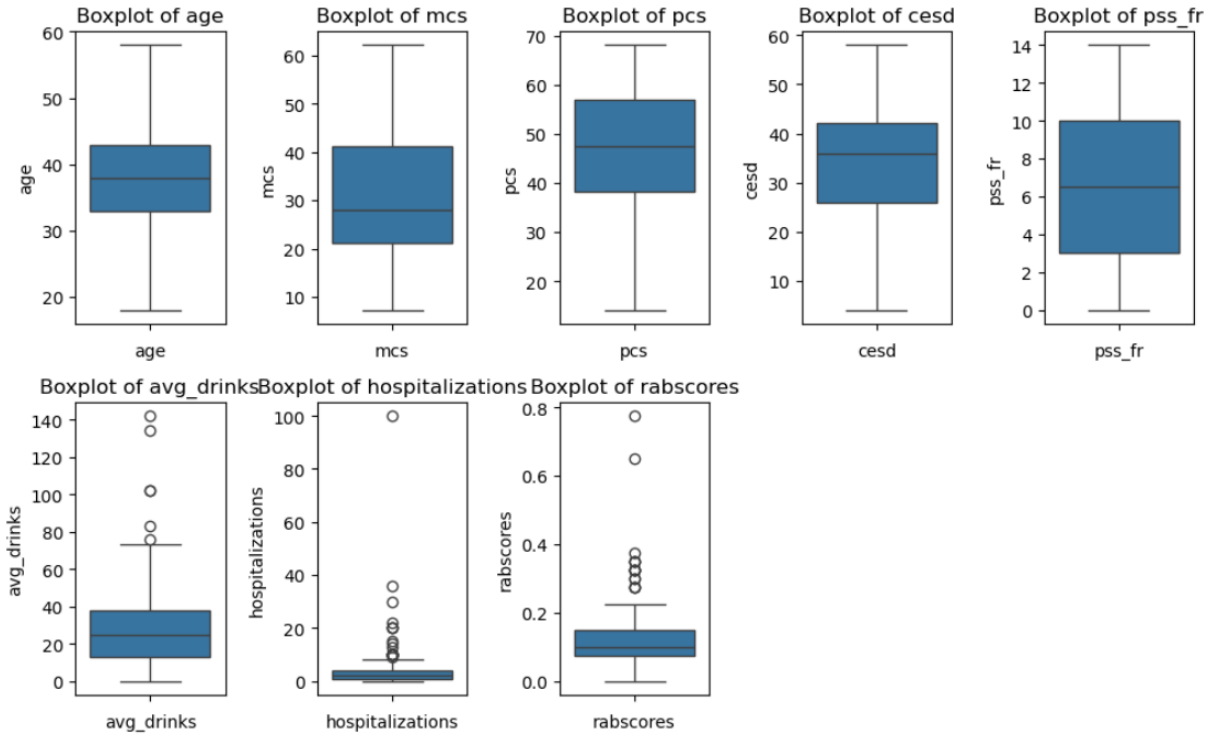
Appendix A: The information of dataset in research

Column name	Description	Type
age	subject age at baseline (in years)	Integer
female	0 for male, 1 for female	Integer
sex	a factor with levels	String
racegrp	race/ethnicity: levels black hispanic other white	String
homeless	Housing status: a factor with levels housed homeless	String
substance	primary substance of abuse	String
mcs	SF-36 Mental Component Score (measured at baseline, higher scores are better)	Float
pcs	SF-36 Physical Component Score (measured at baseline, higher scores are better)	Float
indtot	Inventory of Drug Use Consequences (InDUC) total score (measured at baseline)	Integer
cesd	Center for Epidemiolofic studies Depression measure of depressive symptions at baseline (Higher scores indicate more symptoms)	Integer
id	subject identifier	Integer
pss_fr	perceived social support by friends (measured at baseline)	Integer
i1	average number of drinks (standard units) consumed per day, in the past 30 days (measured at baseline)	Integer
g1b	experienced serious thoughts of suicide in last 30 days (measured at baseline): a factor with levels no yes	String

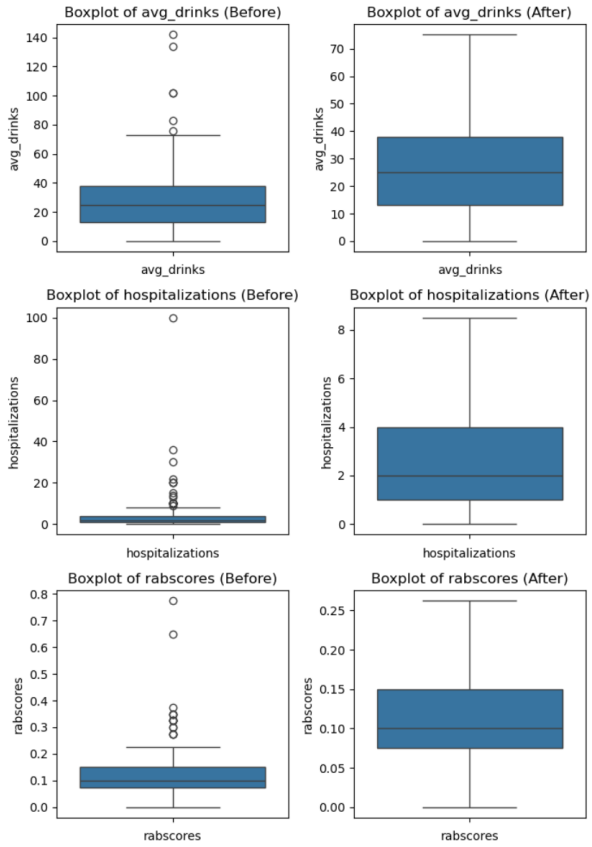
Column name	Description	Type
avg_drinks	average number of drinks (standard units) consumed per day, in the past 30 days (measured at baseline) same as i1	Integer
i2	maximum number of drinks (standard units) consumed per day, in the past 30 days (measured at baseline)	Integer
max_drinks	maximum number of drinks (standard units) consumed per day, in the past 30 days (measured at baseline) same as i2	Integer
d1	lifetime number of hospitalizations for medical problems (measured at baseline)	Integer
hospitalizations	lifetime number of hospitalizations for medical problems (measured at baseline)	Integer
drugrisk	Risk Assessment Battery drug risk scale at baseline	Integer
link	post-detox linkage to primary care	String
sexrisk	Risk assessment Battery sex risk score (measured at baseline)	Integer
satreat	any BSAS substance abuse treatment at baseline: no yes	String
treat	randomized to HELP clinic	String
dayslink	time (in days) to linkage to primary care	Integer
anysub	use of any substance post-detox: a factor with levels	String
daysanysub	time (in days) to first use of any substance post-detox	Integer
e2b	number of times in past 6 months entered a detox program (measured at baseline)	Integer

Appendix B: Data Cleaning (Alcohol group)

- Boxplots of numerical variables in alcohol group

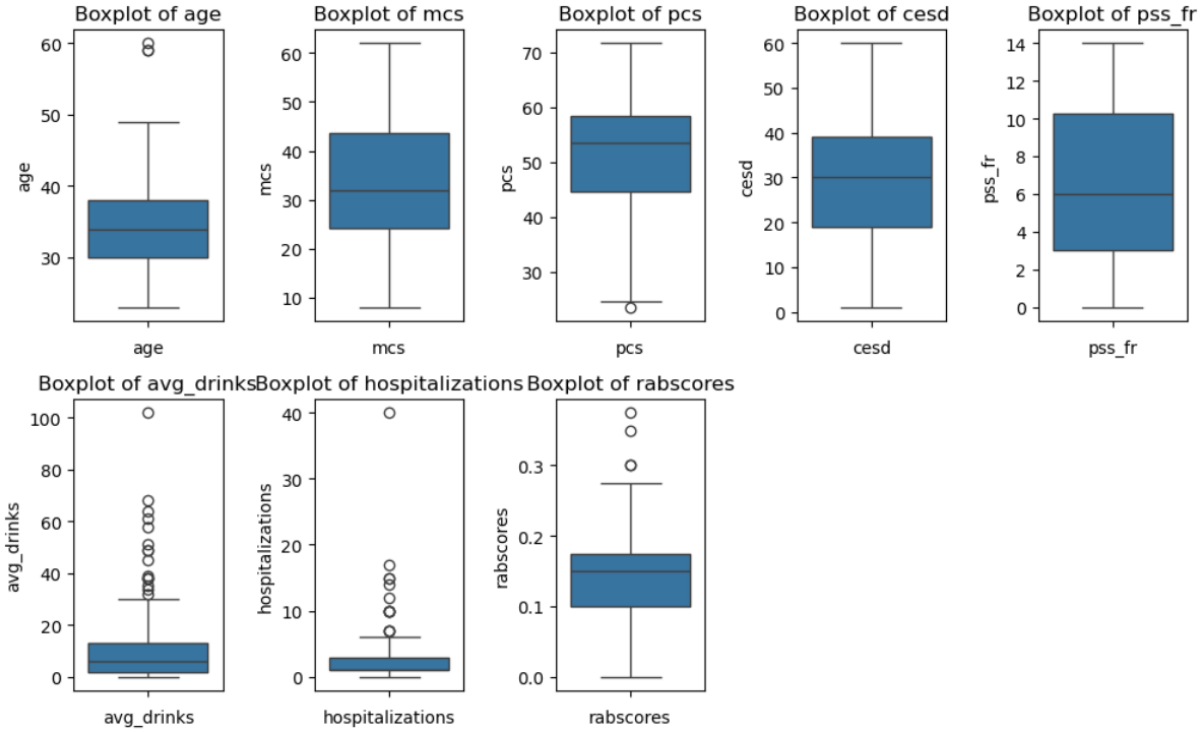


- The data before and after using IQR outlier's removal methods in alcohol group

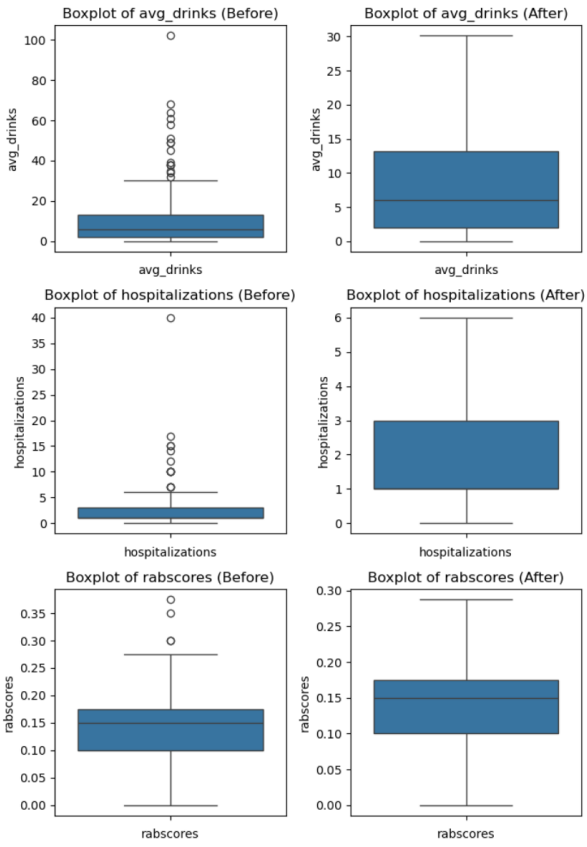


Appendix C: Data Cleaning (Cocaine group)

- Boxplots of numerical variables in cocaine group

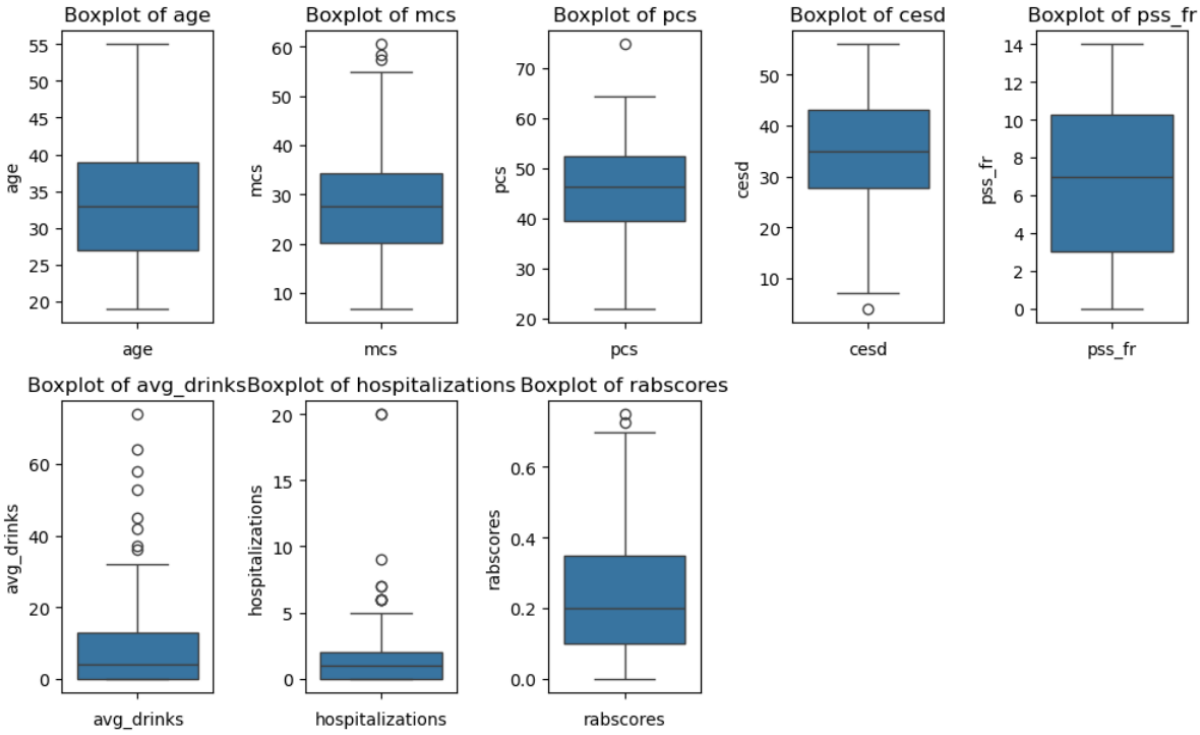


- The data before and after using IQR outlier's removal methods in cocaine group

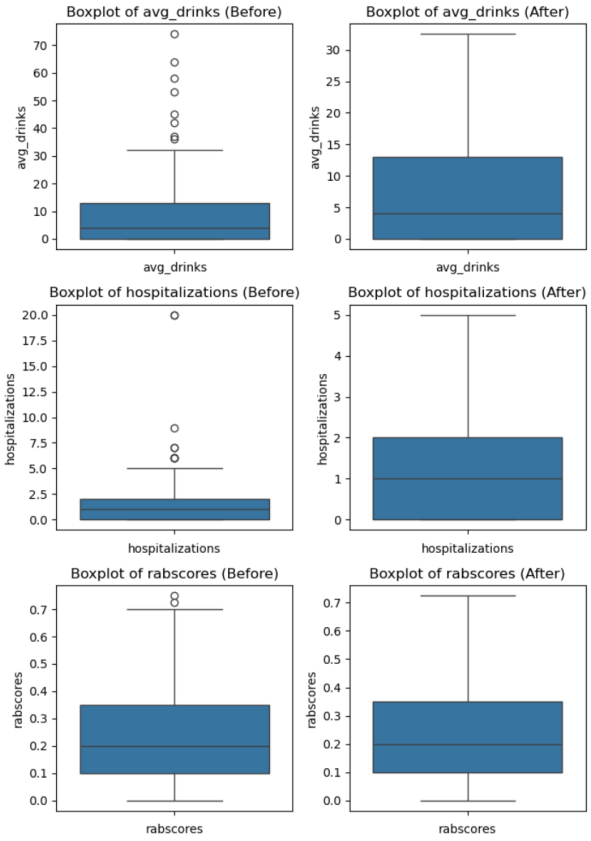


Appendix D: Data Cleaning (Heroin group)

- Boxplots of numerical variables in Heroin group



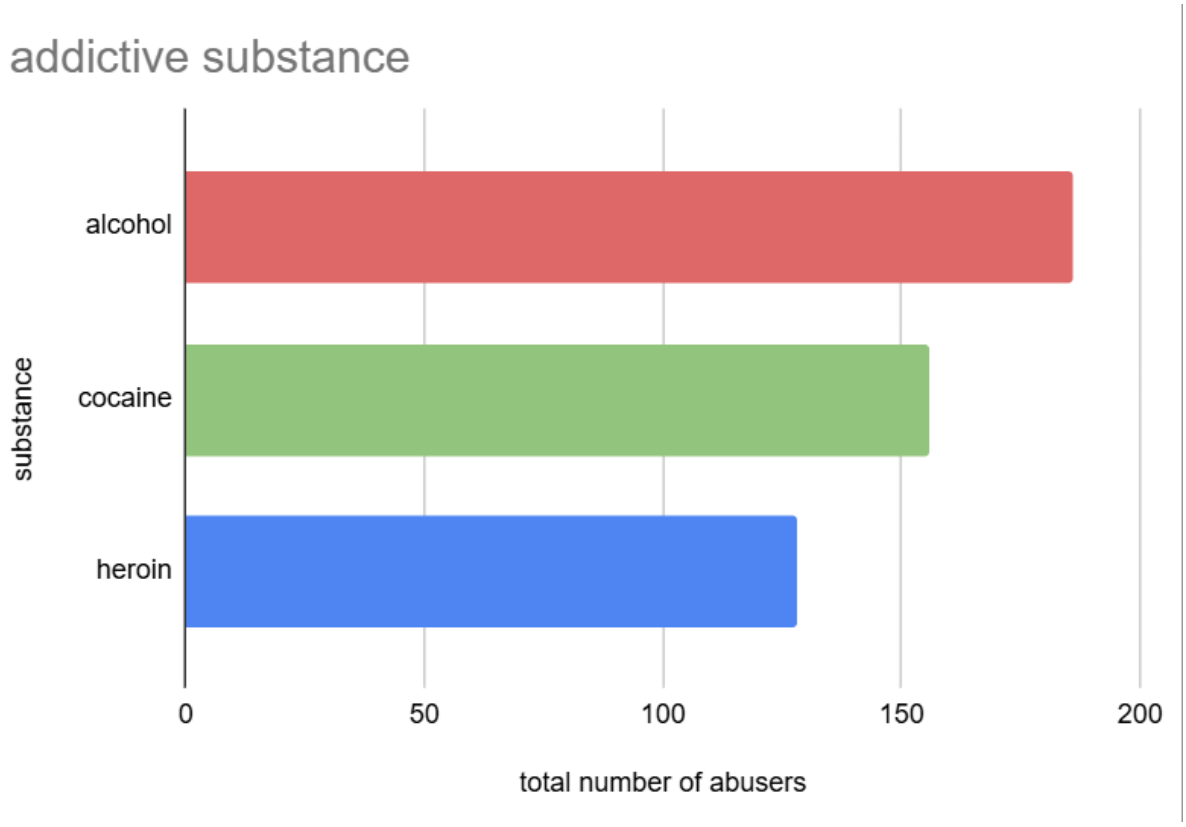
- The data before and after using IQR outlier's removal methods in Heroin group



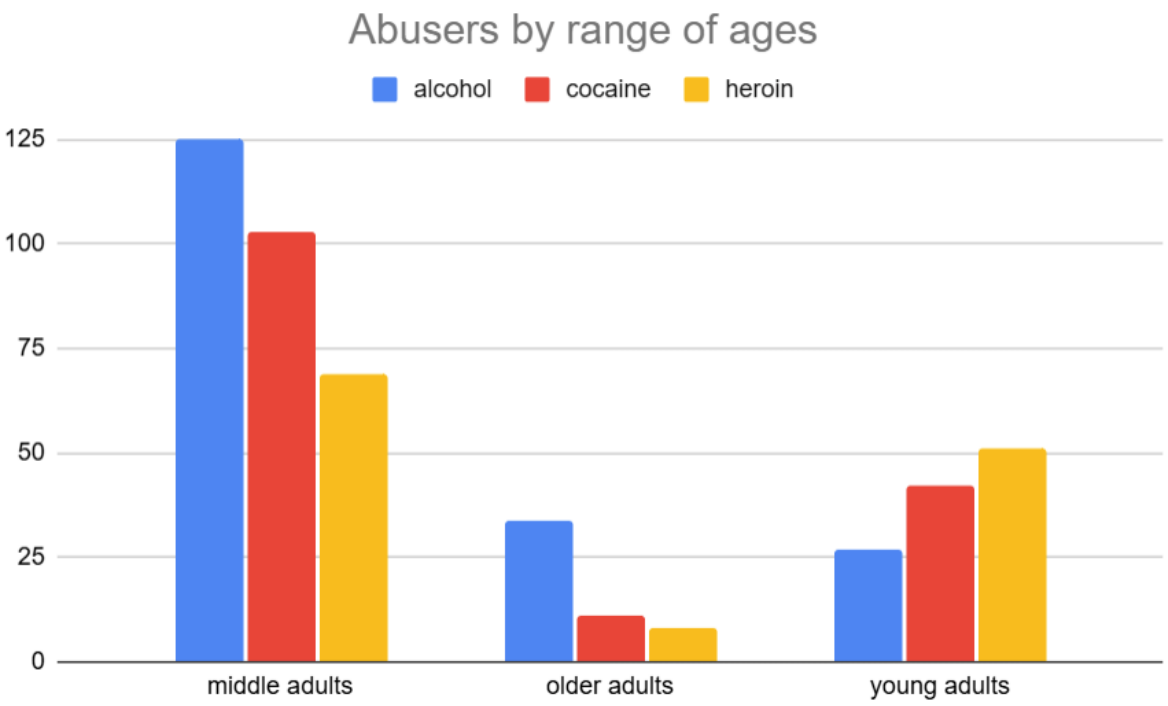
Appendix E: Descriptive statistics of other variables

Link: https://docs.google.com/spreadsheets/d/1C5zC-nPDNPTQfLDJAjkulp2ty8tE4mgmOAY4ykl5yWc/edit?usp=sharing

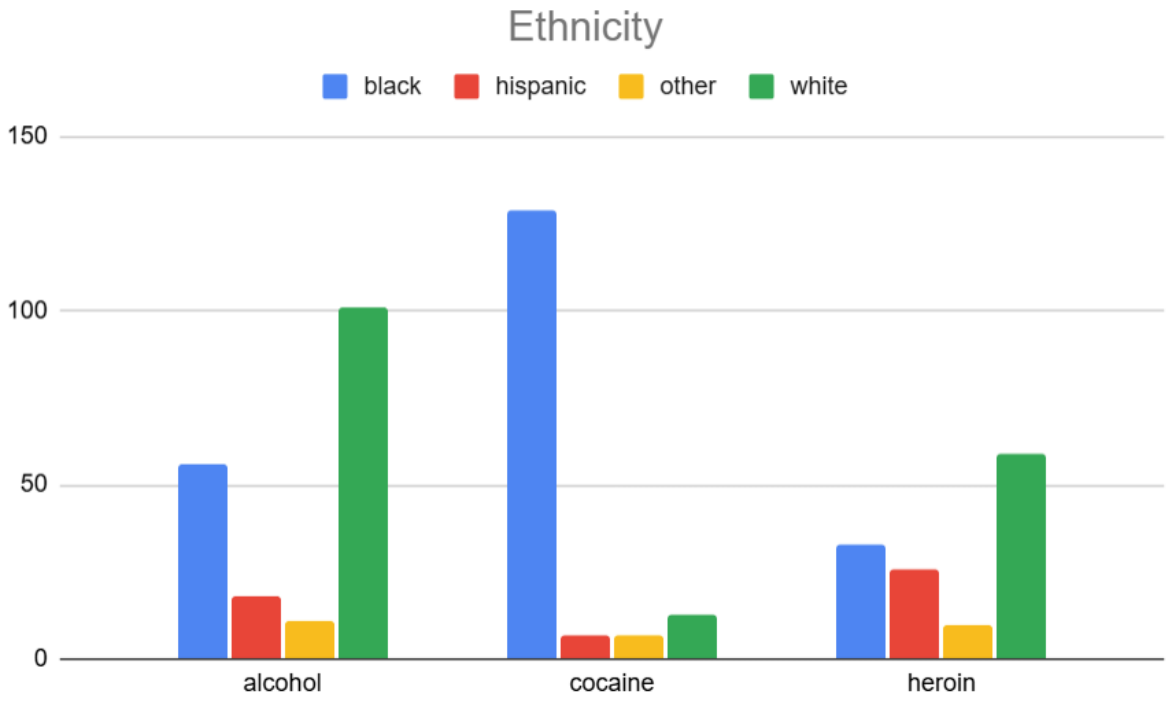
- Addictive substances



- Abusers classified by range of ages

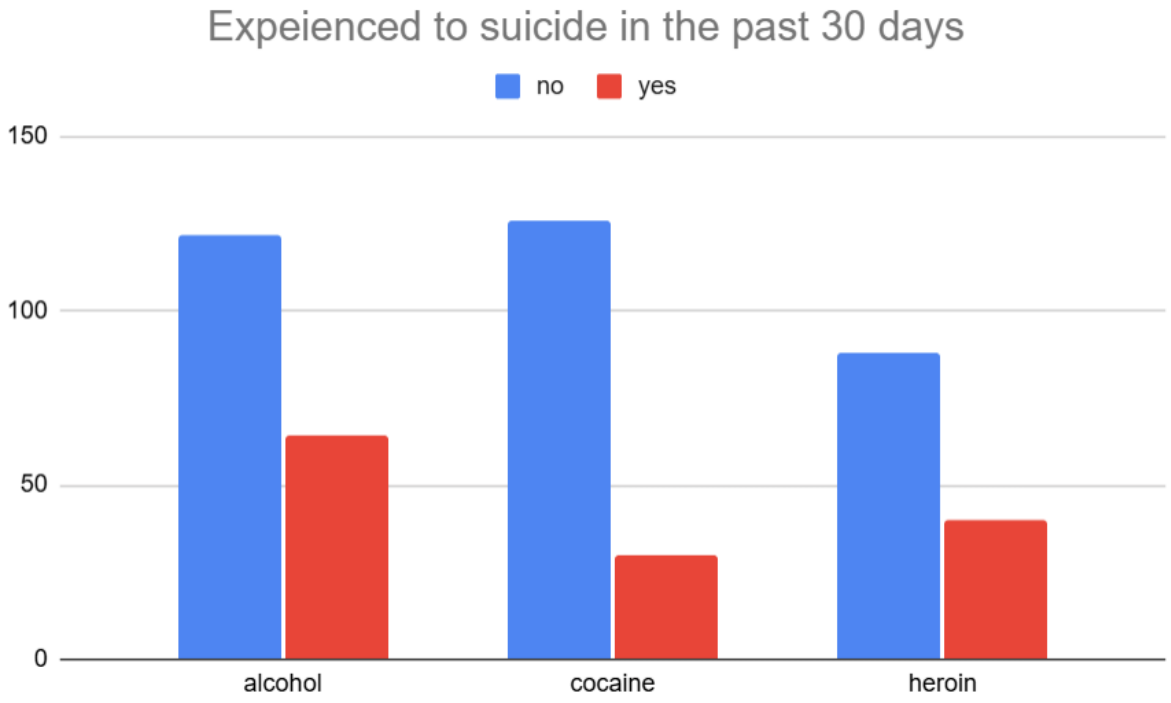


- Ethnicity

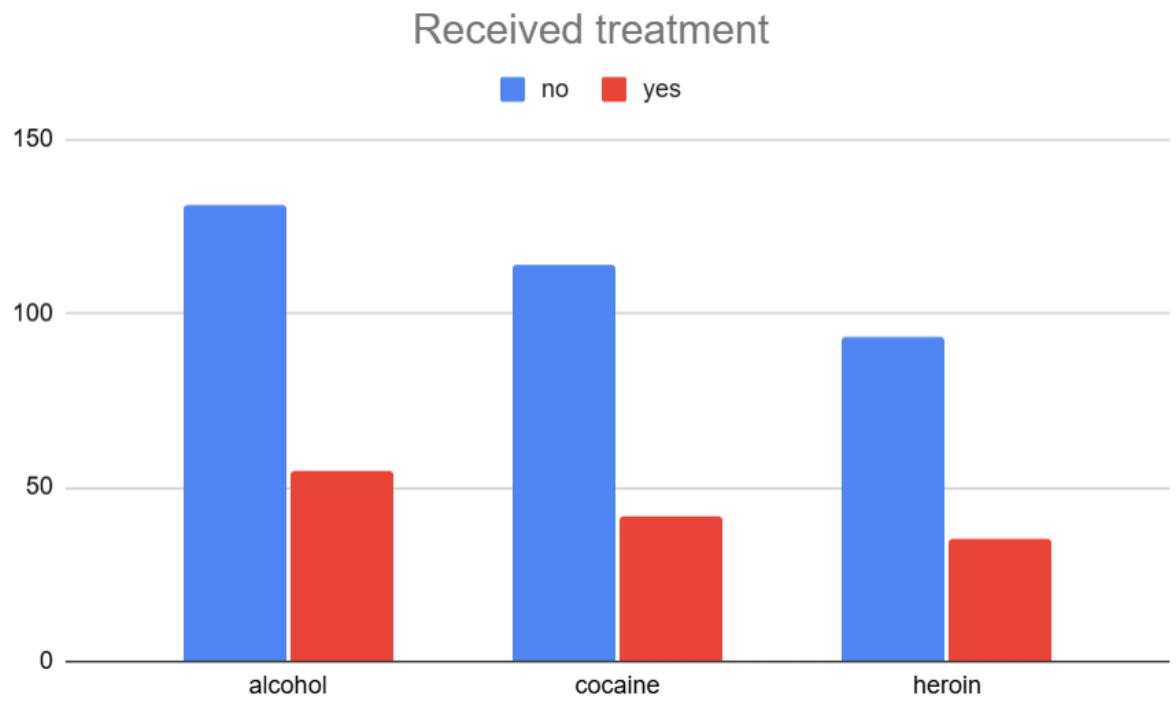


- Housing status

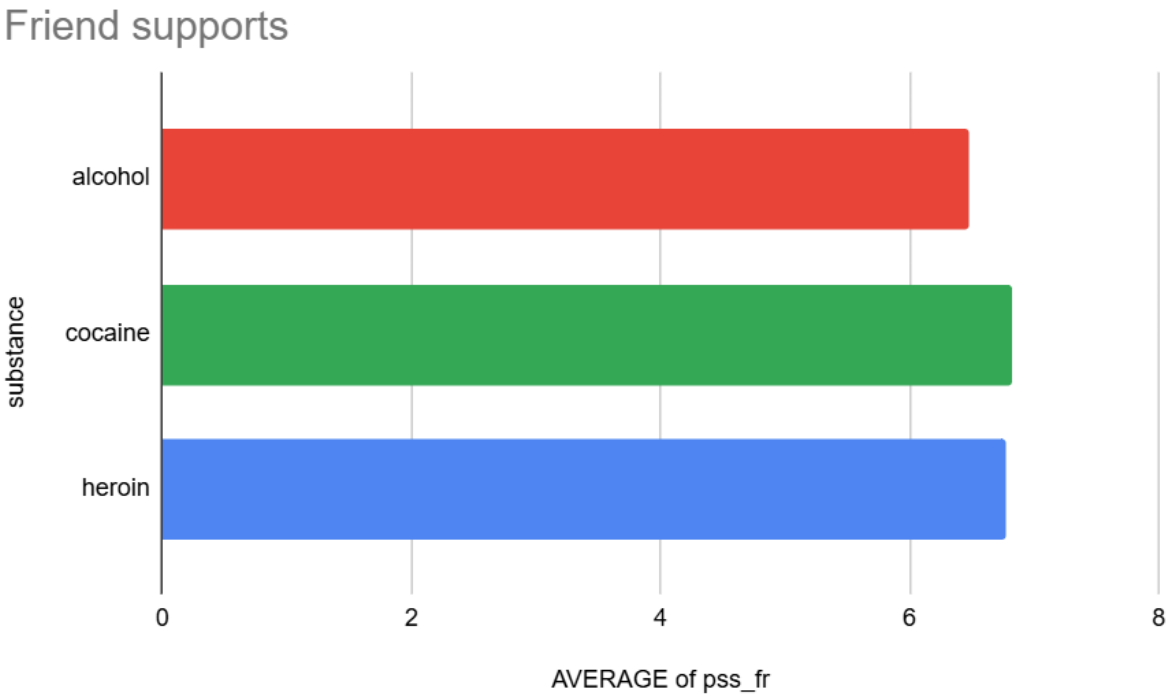
- Experienced to suicide in the past 30 days



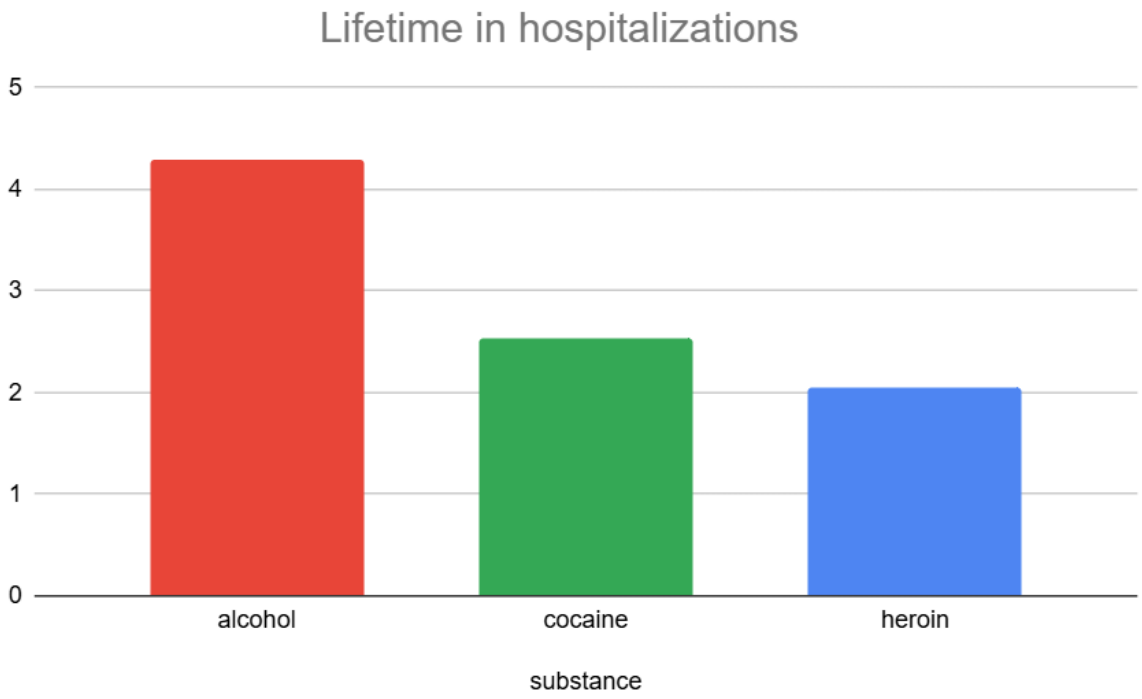
- Received SAT treatment



- Friend supports



- Hospitalizations


Lifetime in hospitalizations
5
4
3
2
1
0
alcohol
cocaine
heroin
substance