Wenyu Shang

Yuying Xie

CMSE381

2021/04/17

**Project Report**

**Introduction**

It is universally acknowledged that the use of illegal drugs has become more frequent in the past decade, and the rate of drug-related incidence and overdose has gradually increased. This trend of illegal drug use also affects the younger people. A large number of young and middle-aged people drink alcohol and take marijuana. Although the use of other types of drugs is not very high, it is still not a small number. Under such circumstances, I would like to study the data of the federal Substance Abuse and Mental Health Data Archive (SAMHDA) to find out whether there are factors affecting the use of other illegal drugs in various age groups, such as age, drinking, and marijuana use, and how they make a difference. Besides that, I am also curious about the performance in different age groups.

**Related Work**

Most of the existing analysis on drug use mainly obtains information intuitively from data and draws conclusions by comparing the data itself. The article *How Baby Boomers Get High* divides the population into different groups according to age, and then compares the frequency of use of various drugs, and draws the following conclusion: boomers aged 50 to 64 have lower rates of drug use overall than their younger contemporaries; they also use less of each individual drug. (Anna, 3) So the use of some drugs may be less common among boomers than other age groups, but it’s still surprisingly frequent for users in the 50 and over crowd. (Anna, 4) While people in the baby-boom generation use more drugs after they’ve turned 50 than their parents or grandparents did, younger boomers use a lot more than older boomers. (Anna, 5)

**Dataset**

The data used in this project is Drug Use By Age, and the source is National Survey on Drug Use and Health from the Substance Abuse and Mental Health Data Archive. It covers 13 drugs across 17 age groups. The drugs include alcohol, marijuana, cocaine, crack, heroin, hallucinogen, inhalant, pain reliever, oxycontin, tranquilizer, stimulant, meth, and sedative. For each drug, the columns show the percentage of those in an age group who used those drugs in the past 12 months and the median number of times a user in an age group used those drugs in the past 12 months.

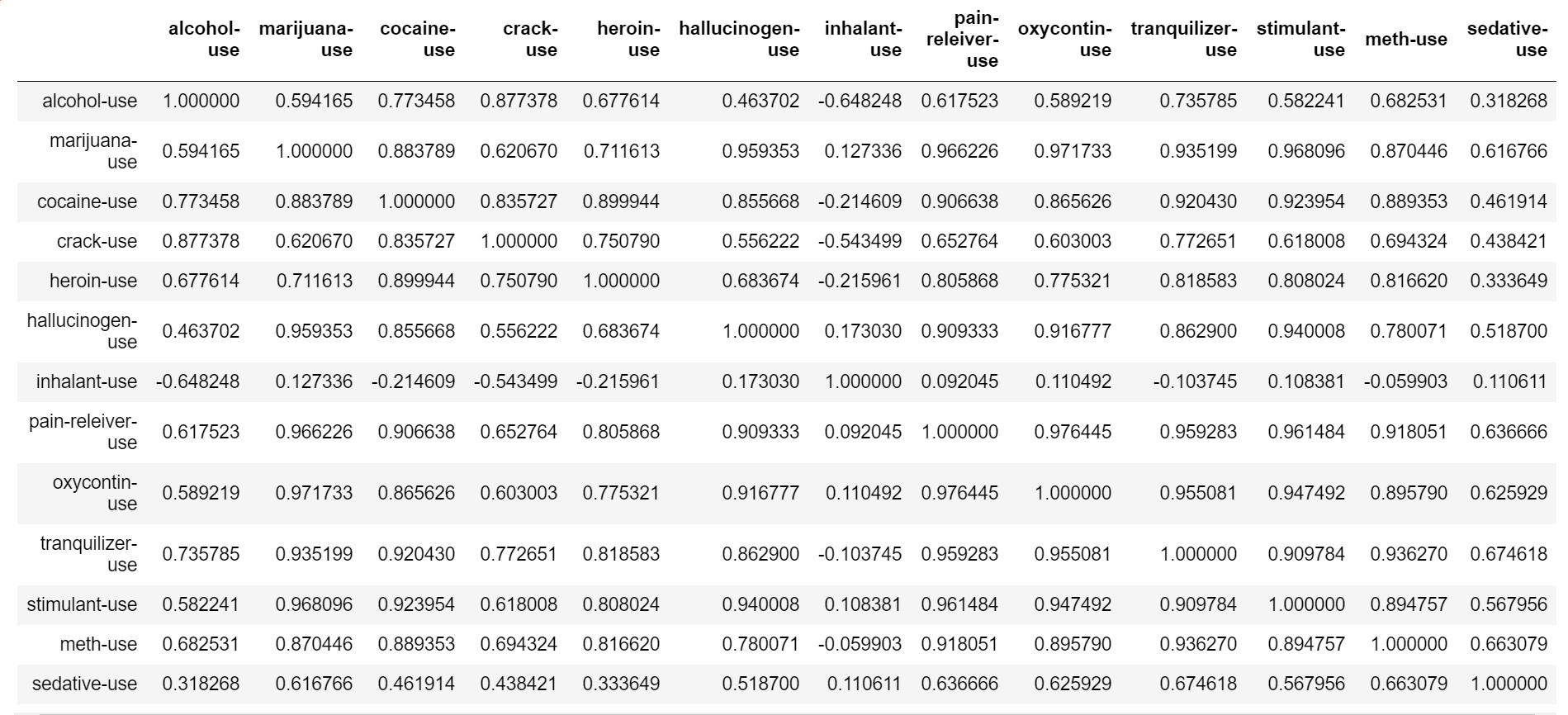


Figure 1: Entire data

Since the information provided by use and frequency overlapped highly, I firstly cleaned the data by removing the columns of frequency and only keeping the columns of use. Then I did a comprehensive graphs to find the general correlations among those factors.



Figure 2: Scatter plot matrix by R

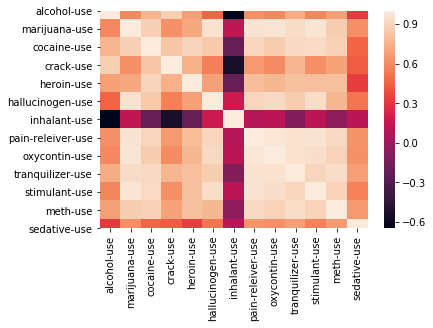


Figure 3: Heatmap by Python

Since the data contains too much information and some factors do not have obvious variance, both heatmap and scatter plot matrix do not perform well. In this case, we need to build models to find the relationships.

**Method**

Linear Regression: Linear Regression is the simplest model that we can find relationships among the variables. So I firstly tried several linear regression models with different inputs, then I compared the P-values to determine whether factors are significant to consider. By the summaries, the factor that is related to age is pain reliever use; the factor that is related to alcohol use is hallucinogen use; no factor seems to be highly correlated to marijuana use. It has to be admitted that linear regression did not perform well for this data because the p-values did not show many significant relationships among the factors. In this case, factors do not have clear linear relationships.

Logistic Regression: I divided the age to two groups, which are younger teenagers (12~19 years old) and elder people (older than 20 years old). However, after building the logistic regression model, the p-values are all greater than 0.05, which means logistic regression perform even worse than linear regression model.

Other methods: Since different age groups do not have same probabilities of having certain conditions, they could not be regarded as different samples of a variable. In this case, it is extremely hard for us to build train or test sets to predict the use of drugs in diverse conditions by such a small dataset. If the data contains each respondent’s response separately, then it’s better and easier to make more diverse models, such as Lasso, Ridge, Boost, Random Forest, to name just a few.

**Discussion and Result**

By directly comparing the data and building models to find relationships among factors, we can tell that it is significantly hard to find relationships among the use of different drugs due to the small size of the data. And the use of drugs does not show obvious relationships with age, use of alcohol or use of marijuana. Thus, those factors seem to be separate and independent. By directly comparing the data, we can tell that the use of drugs are more than we predicted due to the fact that a great number of young people and old people are using drugs.

**Conclusion and Future Work**

In conclusion, the use of drugs is getting younger, and middle-aged drug addicts have brought the habit of drug use to old age, so that the amount of drug use has a normal distribution, and more and more young people are exposed to these drugs too early. However, the use of drugs is not directly related to age, alcohol consumption, or marijuana use.

In the future, I hope I can take use of other related data to build models to analyze the data. And hopefully, I can use those data to predict future trend of drug use.