CMSE 381 Final Project

Zachary Matson

Objectives

The objective of this report is to use publicly available drug use data to analyze whether or not the "gateway drug" effect, where use of soft drugs (alcohol, nicotine, marijuana) leads to the use of "hard" drugs (cocaine, methamphetamine, opiates, etc.). Secondarily, such data will be used to reproduce the visualization from the article "How Baby Boomers Get High" by Anna Maria Berry-Jester and Andrew Flowers, published by FiveThirtyEight.

Related Work

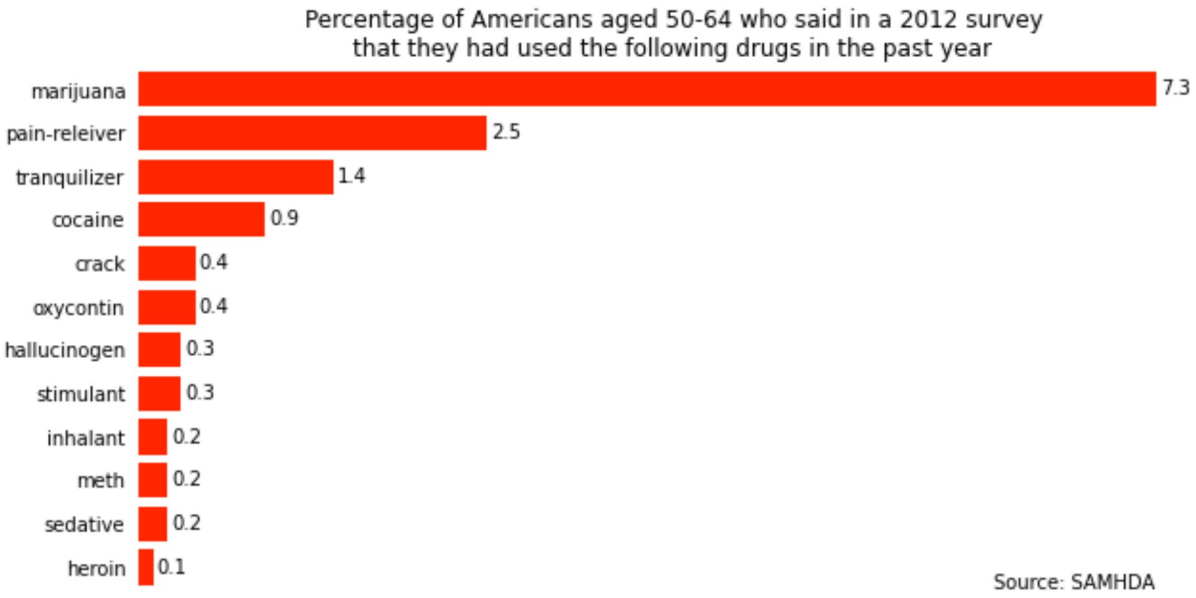
This section focuses on the aforementioned article in FiveThirtyEight. This graph demonstrates the usage rates of various drugs among baby boomers.

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re

data_538 = pd.read_csv("drug-use-by-age.csv")
```

In [2]:

```
plt.figure(figsize=(10,5))
use_regex = re.compile("^(.+)-use$")
boomer_use = pd.melt(
    data_538[data_538.age == "50-64"],
    id_vars="age",
    value_vars=filter(use_regex.match, data_538.columns),
    var_name="drug",
    value_name="use"
)
boomer_use.drug = boomer_use.drug.map(lambda s: use_regex.match(s).group(1))

g = sns.barplot(
    data=boomer_use[boomer_use.drug != "alcohol"].sort_values("use", ascending
    y="drug", x="use",
    color="#ff2600", saturation=1
)
g.axes.set_xlabel(None)
g.axes.set_ylabel(None)
g.axes.tick_params(axis="both", which="both", bottom=False, labelbottom=False,
for spine in g.axes.spines.values():
    spine.set_visible(False)
for patch in g.patches:
    g.annotate(
        patch.get_width(),
        (patch.get_width(), patch.get_y()),
        va="top", ha="left", xytext=(2,-4), textcoords="offset points"
    )
g.axes.set_title("Percentage of Americans aged 50-64 who said in a 2012 survey
g.annotate("Source: SAMHDA", (0.8, 0), xycoords="axes fraction")
```

Out[2]: Text(0.8, 0, 'Source: SAMHDA')



Interestingly, the numbers match the graph on the website, but the data they published has a lower decimal accuracy than the data in their graph, which causes some of the bars to be in a slightly different order when using the published data.

Gateway Effect?

The "gateway effect" is often brought up as a reason children should avoid using drugs or alcohol. Aside from the medical implications of childhood and teenage drug use, it is highly debated whether early use of soft drugs (alcohol, nicotine, marijuana) lead to the use of harder drugs.

To investigate the gateway effect, I used data from the 2019 National Survey on Drug Use and Health, as published by the Substance Abuse and Mental Health Services Administration.

Model

Because there are a multitude of drugs, the analysis will specifically focus on methamphetaine, cocaine, crack (because the data source separated this from cocaine), heroin, and pain relievers outside of doctor directed use.

For the first predictors, we will consider whether children used alcohol, cigarettes, and marijuana before the age of 18. The predictions for drug use will focus on whether the given hard drug was used within the last 12 months.

In [3]:
```
data_nsduh = pd.read_csv(
    "NSDUH_2019_Tab.txt", sep="\t",
    usecols=[
        "IRALCAGE", "IRMJAGE", "IRCIGAGE",
        "IRMETHAMREC", "IRHERRC", "IRCOCRC", "IRCRKRC", "IRPNRNMREC",
        "IRFAMIN3"
    ]
)
```

In [4]:
```
data_processed = pd.DataFrame({
    "alcohol_as_minor": data_nsduh.IRALCAGE < 18,
    "cigarettes_as_minor": data_nsduh.IRCIGAGE < 18,
    "marijuana_as_minor": data_nsduh.IRMJAGE < 18,
    "family_income_level": data_nsduh.IRFAMIN3,
    "cocaine_12mos": data_nsduh.IRCOCRC < 3,
    "crack_12mos": data_nsduh.IRCRKRC < 3,
    "heroin_12mos": data_nsduh.IRHERRC < 3,
    "meth_12mos": data_nsduh.IRMETHAMREC < 3,
    "painkiller_12mos": data_nsduh.IRPNRNMREC < 3,
})
```

In [5]:
```
from sklearn.model_selection import train_test_split

X = data_processed.loc[:,["alcohol_as_minor", "cigarettes_as_minor", "marijuar
Y = data_processed.drop(columns=["alcohol_as_minor", "cigarettes_as_minor", "r
```

For each response variable, I trained a logistic regression model with all of the predictor variables. Then, I looked at the probability of hard drug use predicted by the model for a few different combinations of the predictors. I chose a logistic regression model here, because the

rates of use of these hard drugs in the entire population are low enough that the model never predicts that hard drugs *were* used, and I wanted to get an idea from the model of how much the odds were affected by the different factors being considered. In addition to the interpretability benefits of a logsitic model for that use case, it has an advantage in that it can be trained well at all on data with a poor split. A decision tree, for example, would not train well on this data because no combination of predictors would yield a node with a prediction of hard drug use, and no tree would do any better than any other that always predicted "no" for hard drug use. Also, because drug use will never actually be predicted by the model, it is not helpful to

I also used family income level as a control variable because income is related to many other factors that might be useful to control for, and there were too many variables in the data set to go through all of them. The prediction samples will be based on the median value for income bracket.

In [6]:
```
from sklearn.linear_model import LogisticRegression
from IPython.display import display

predictor_combos = pd.DataFrame(np.vstack([
    np.zeros((1,3)),
    np.eye(3),
    np.ones((1,3))
]).astype(np.bool), columns=X.columns[:-1])
predictor_combos["family_income_level"] = X.family_income_level.median()
predictions = predictor_combos.drop(columns="family_income_level")
models = []

for predictor in Y.columns:
    y = Y[predictor]
    logistic = LogisticRegression(random_state=1432536547)
    logistic.fit(X, y)
    use_probs = []
    for i in range(5):
        combo = predictor_combos.iloc[[i],:]
        use_probs.append(logistic.predict_proba(combo)[0,1].item())

    drug = predictor.split('_')[0]
    predictions[f"{drug} use in last 12 months, est. prob."] = use_probs
    models.append([drug, *np.hstack([logistic.intercept_.flat, logistic.coef_

models = pd.DataFrame(
    models,
    columns=["Response Drug", "Intercept", "Alcohol Coefficient", "Cigarette (
)
```

Results

In [7]:
```
display(predictions)
```

alcohol_as_minor	cigarettes_as_minor	marijuana_as_minor	cocaine use in last 12 months,	crack use in last 12 months,	heroin use in last 12 months,	meth use in last 12 months,

				est. prob.	est. prob.	est. prob.	est. prob.
0	False	False	False	0.002639	0.000185	0.000096	0.000543
1	True	False	False	0.008625	0.000336	0.000327	0.001364
2	False	True	False	0.004369	0.000851	0.000426	0.001731
3	False	False	True	0.016059	0.000819	0.000562	0.003032
4	True	True	True	0.081702	0.006791	0.008423	0.023843

This table shows the model predictions for the probabilities of having used different hard drugs in the previous 12 months, based on childhood use of various soft drugs. All of the predictions are based on the median family income bracket for respondents in the survey ($50,000-75,000). You can see here that most of the probabilities are fairly low, although the chance of someone who used alcohol, cigarettes, and marijuana as a child also having recently used painkillers is 9.46\%, and their chance of having recently used cocaine is 8.2\%.

Still, the odds for people who use different soft drugs as minors have higher probabilities of using each kind of drug, and people who used all of the soft drugs being considered had much higher probabilities for using each of the drugs. For meth, the odds were 44x greater for people who used all of the soft drugs as a child compared to those who used none of them. For painkillers, the increase is only 7x. For cocaine, it is 31x.

In [8]:
```
display(models)
```

	Response Drug	Intercept	Alcohol Coefficient	Cigarette Coefficient	Marijuana Coefficient	Family Income Coefficient
0	cocaine	-5.311384	1.190202	0.505668	1.819279	-0.103867
1	crack	-6.843523	0.596351	1.525023	1.487043	-0.291717
2	heroin	-7.489683	1.224757	1.489763	1.768380	-0.293597
3	meth	-6.101730	0.922409	1.160973	1.722657	-0.236076
4	painkiller	-3.745181	0.703723	0.529842	0.816665	-0.093922

The intercepts for all of the models are negative, and greater in magnitude than any of the predictor coefficients. This reflects the fact that most people are unlikely to have used any hard drugs in the last 12 months. The coefficients for family income are also all negative, reflecting the fact that people of higher socioeconomic status are, *in general*, less likely to use hard drugs. However, all of the soft drug predictors have positive coefficients. The magnitudes of the soft drug coefficients in comparison to the income coefficients are not useful because the family income has more possible levels, but with seven possible income brackets, seven times the family income coefficient is the expected difference between people of the highest possible and lowest possible income bracket.

Discussion and Conclusions

To the question of whether or not childhood use of soft drugs predicts further use of hard drugs, when controlling for family income, it appears that it does. This could be interpreted as supporting the idea of a gateway effect between soft drugs and hard drugs. As highlighted in the results section, someone in the median family income bracket who used alcohol, nicotine, and marijuana before the age of 18 is around 44x more likely to have used meth in the last year than someone who used none of those drugs, 7x more likely to have used painkillers and 31x more likely to have used cocaine.

However, it is important to look at the scales here. For the overall population, childhood drug use or not, most people have not used hard drugs in the last 12 months. For meth, the likelihood of use is so low even with all of the childhood drug use predictors, that the increase in risk shouldn't be that concerning. Many activities in life increase risk significantly but are still accepted because even the heightened risk is relatively low. This isn't to say that there are no other reasons not to use drugs as a minor, just that the gateway effect in particular is not a great reason, looking at meth.

Considering painkillers and cocaine changes the picture a bit. Both of these have fairly significant odds of use after childhood drug use and insignificant odds without it. People who have used soft drugs in their childhoods have a real risk of using cocaine and painkillers in adulthood, both of which can cause serious damaging consequences in a person's life. Here, there may be good evidence supporting the gateway effect as something to be concerned about.

There are some caveats to the conclusions from this data though. Family income bracket was controlled for as a proxy for other socioeconomic factors, but this oversimplifies the other factors which could contribute to drug use. It also ignores the way that family income can change throughout a person's life, while all of the periods in someone's life may contribute to future drug use. It also excludes people who have used hard drugs but not recently, even though they may have already undergone severe consequences due to past drug use before quitting those drugs.

Beyond this, the fact that soft drug use predicts hard drug use does not mean that soft drug use *causes* soft drug use. Drug use is, to some extent, always a choice, and the fact that a person who chooses to use hard drugs is more likely to have previously used soft drugs doesn't mean that the soft drug use *caused* the hard drug use, just that people who used soft drugs as children in general are more likely to eventually try hard drugs. The absence of childhood soft drug use is also not a perfect baseline, as many people do try or use some soft drugs before turning 18. People who have not used any soft drugs are not necessarily "normal" in comparison to "abnormal" people who have used them; the people who didn't use them in childhood may be especially likely to avoid drugs compared to a normal person, and because of this more likely to avoid hard drugs as well. Again, this means that the fact that they didn't use soft drugs in their childhood isn't necessarily the cause of them not using hard drugs in adulthood, but they could instead both be caused by the same underlying values and tendencies.

Taking all of this into consideration, the picture is murky. There isn't enough evidence to say that

avoiding soft drugs in youth is a valid strategy to prevent hard drug use down the line. Typically, the gateway drug theory is used to say that soft drug use should be punished to prevent hard drug use, so this data would not support the efficacy of that course of action, especially when the magnitude of potential preventative strength is compared to the costs of such a strategy. The data does, however, suggest that strategies which prioritize other preventative resources for those with youth usage of soft drugs may be valid. If people who used soft drugs in their childhood are more likely to go on to use hard drugs, as the data shows they are (caveats aside), then it makes sense to set up support systems with extra care for those people to make sure they don't go on to use hard drugs in the future.