Drug Trends Analysis Report:

12/3/2018:

Erik Hamlin
Bernardo Jordan
Fu Shen
Alex Suarez

Introduction:

This project is centered around the analysis of a drug survey from 2016 conducted by the National Survey on Drug Use and Health (NSDUH)[1]. This survey is directed by the Substance Abuse and Mental Health Services Administration (SAMHSA)[2], an agency in the Department of Health and Human Services (HHS). The data is stored in the Substance Abuse & Mental Health Data Archive (SAMHDA), a public resource. This series was first conducted in 1971 and the most recently published data came from the 2017 questionnaire.

In order to conduct the 2016 survey, individuals from across the United States were randomly sampled, visited by a representative of RTI International, a subcontractor of SAMHSA and given a 1-hour survey for which they were paid $30. Respondents were 12 years or older and answers to questions were entered by either an interviewer or the respondent[3].

The survey is comprehensive, with questions belonging to categories such as: drug usage (legal and illegal), socio-economic status (household composition, income, welfare status, etc.), criminal activity, ethnicity, age, mental health and attitudes towards religion.

Given that there has been an increase in drug overdose deaths and depression over the last 3 decades[4], we wanted to get a snapshot of what the drug use and mental health landscapes looked like today. What are the basic descriptive statistics of the categories mentioned above? Can we provide a definition for a gateway drug that allows for a precise, quantitative measurement?

This is a bottom -up, benchmark analysis of one NSDUH survey that ideally would be replicated across all NSDUH surveys. Before we can predict any of the future trends of the multitude of variables in this survey, it is important to understand the proclivities in the data of the first survey.

Approach:

Python (> 3.5) packages:

1. Numpy (version 1.15.1)
2. Matplotlib (2.2.3)
3. Jupyter (1.0)
4. iPython (6.5)
5. pandas (0.23.4)

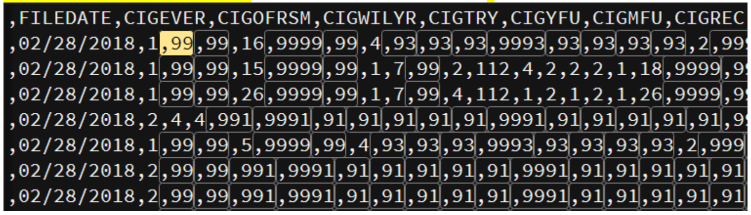
Microsoft Excel 2016
Ubot Studios (6.0.4)

We used Python & Microsoft Excel to wrangle, clean, impute and visualize the data.

Ubot Studios was used to scrape the SAMDHA web application page[5].

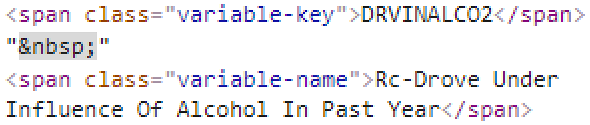
Data Acquisition and Cleansing:

Our data acquisition and cleaning process is described below.

1. Data File Review
   - We initially downloaded the entire 2016 dataset from https://samhda.s3-us-gov-west-1.amazonaws.com/s3fs-public/field-uploads-protected/studies/NSDUH-2016/NSDUH-2016-datasets/NSDUH-2016-DS0001/NSDUH-2016-DS0001-bundles-with-study-info/NSDUH-2016-DS0001-bndl-data-tsv.zip . After we reviewed the data, we noted several issues.
   - File size and uniformity of data types:
     - While the compressed ZIP file was only 22 mb, the uncompressed the size was approximately 375 mb. This initially caused many of our scripts (Python, UBot Studios, Excel, and text editors) to either lag significantly or crash entirely when attempting to process the data.
     - The data was tab delimited where most software requires comma separation.
     - There was a substantial portion of data that contained sub-categorized, non-applicable responses. For example, CIGOFRSM (a question with responses ranging from "definitely yes" to "definitely no") occasionally contained the following responses marked by participants.
       - 94 – don't know
       - 97 – refused
       - 98 – blank
       - 99 – legitimate skip
   - Readability:
     - The full English description of each question was not included. Only the column header was coded as a six to eight capitalized name (see "Experiments", part 1). The full description of these questions was only located (a) within a PDF document and (b) online at https://pdas.samhsa.gov/#/survey/NSDUH-2016-DS0001 as part of an presentation. There was no easily identifiable way to correlate the coded column name with its corresponding English equivalent.
2. Scrape full question descriptions
   - First, we opened the entire tab delimited file from SAMHDA in Excel, while this took some time, we eventually saved the file as a CSV. After this, we used a combination of Excel search and replace along with Regular Expressions to remove any N/A responses. For example, we used the following Regex find/replace all responses with values of 91 and above. This step alone greatly reduced the size of the data file to under one half.
     - ,9(\d{3}|\d{2}|\d{1}) was replaced with ,
     - 

```
,FILEDATE,CIGEVER,CIGOFRSM,CIGWILYR,CIGTRY,CIGYFU,CIGMFU,CIGREC
,02/28/2018,1,99,99,16,9999,99,4,93,93,93,9993,93,93,93,93,2,99
,02/28/2018,1,99,99,15,9999,99,1,7,99,2,112,4,2,2,2,1,18,9999,9
,02/28/2018,1,99,99,26,9999,99,1,7,99,4,112,1,2,1,2,1,26,9999,9
,02/28/2018,2,4,4,991,9991,91,91,91,91,91,9991,91,91,91,91,91,9
,02/28/2018,1,99,99,5,9999,99,4,93,93,93,9993,93,93,93,93,2,999
,02/28/2018,2,99,99,991,9991,91,91,91,91,91,9991,91,91,91,91,91
,02/28/2018,2,99,99,991,9991,91,91,91,91,91,9991,91,91,91,91,91
```

   - Second, we revisited the page https://pdas.samhsa.gov/#/survey/NSDUH-2016-DS0001 and expanded the entire question tree down to each question. This bypassed the website's attempt to load the questions on demand and thereby forced the site to load every column code along with its English equivalent.

    - Next, we utilized Ubot Studios to scrape the webpage for all corresponding HTML classes of “variable-key” along with “variable-name”. (please see Apendix 2 for full UBot code)
        - 

```
<span class="variable-key">DRVINALCO2</span>
"&nbsp;"
<span class="variable-name">Rc-Drove Under
Influence Of Alcohol In Past Year</span>
```

    - Next, we used Excel via copy/paste to add the English equivalent into the final data set.
3. Reduction of Data File by merging on subset of Column Codes
    - To reduce the number of columns from approximately 2,699 to 600, we utilized Python and merged the larger data set onto a subset of 599 Column Codes. This allowed us to further reduce the file size and easily manage the data.

Experiments:

1) Label Columns as Binary/Integer/Categorical:

Each of the questions on the survey had a coded abbreviation attached to them. For example: “Rc-Marijuana Abuse - Past Year” was coded as ABUSEMRJ.

There were 2,669 coded question given on the survey. 1,012 of them were recoded versions of older questions or “revised imputations”. By “recoded”, we mean that answers with Integer or Categorical variable types were transformed into columns with new Categorical or Binary variable types or they were repeats.

- We took an Excel document containing only the column names, and their questions (“ColumnsVersion1.xlsx”)

We labelled the data under the Question column with “recoded”, “rc-“, and “revised imputations” and kept those questions.
The rest of the non-labelled data consisted of questions like:
“Ever Had Per Of Time Lst Intrst In Enjoyable Thgs” (ever had percentage of time lost in interest in enjoyable things)
Or
“One Particular Time That Is The Worst One Ever”
Or questions that had been re-coded.
The data types of the questions were verified on the SAMHDA web application page[5].

After reading through the questions, we decided to keep a final set of 650 questions for our analysis. We saved them in “ColumnsVersion2.xlsx”

In our final set of columns, we added a “Data Type” category. We went question by question and plugged each question code into the NSDUH web application, looked at the possible results and categorized each one as a Categorical, Integer, Binary, or Numerical data types. “ColumnsVersion2.xlsx” was updated with this information.

2) Updating Data For Gateway Drug Analysis

In **“Drug Gateway Substance Analysis Generate Final Data.ipynb**”, a Jupyter notebook, we created a dataframe that would be referenced for the Gateway Drug portion of our analysis.

Before we loaded the “SurveyData.csv” file we filtered out the 98 columns we would use in our DataFrame and saved the names in “ColumnsGateway.xlsx”

Below is a summary of the “Drug Gateway….” script, broken down into 2 parts:

Initialize DataFrame

1. Load “SurveyData.csv” file
2. Load “ColumnsGateway.xlsx” file
3. Concatenate the two Dataframes
4. Isolate the Binary Data
    a. Make imputations to any columns that do not have (0,1) values. For example, some of the columns are marked (1, 2) or (1, 9)
5. Isolate the Integer/Categorical Data
    a. No imputations were made
6. Merge the Binary and Integer/Categorical Dataframes
7. In the new Dataframe
    a. Make imputations of Month Columns

Add Columns

1. Add “First Drug” and “Age of First Use” Columns
    a. These two columns contain the first substances and the age of first use of respondents
2. Cleanup “First Drug” column
    a. Remove values in this column for individuals who did not report on consuming drugs
3. Add “TOB First”, “ALC First”, “MJ First”, “STI First”, “HAL First”, “HER First” Columns
    a. These are flags for whether the substances were consumed at the earliest age reported by the respondents (Tobacco, Alcohol, Marijuana, Stimulants, Hallucinogens, Heroine)
4. Add “Drug After” column
    a. Indicates whether respondent went on to consume a substance at a later age
5. Add “Ever Used” column
    a. Indicates whether a respondent ever used a substance
6. Add “Age of First…” Columns
    a. For Tobacco, Alcohol, Heroine, Marijuana, Stimulants, and Hallucinogens
7. Once the columns have been added, export the Dataframe to “FinalData.csv”

3) Generating Proportions/Descriptive Statistics and Boxplots:

In **“Drug Gateway Substance Analysis Generate Prop_Hist_Box.ipynb**” we took the data from the newly created “FinalData.csv” introduced a re-defined definition of “Gateway Drug” and selected 6 candidate drugs:
Tobacco, Alcohol, Marijuana, Tobacco/Alcohol, Tobacco/Marijuana, Alcohol/Marijuana, and Alcohol/Tobacco/Marijuana

Respondents fell into each of these categories if they tried any of the 7 candidate drug categories at the age of their youngest drug consumption.
For example: If a respondent’s first drug was Alcohol then:
(TOB First=0)& (ALC First=1)& (MJ First=0)& (HAL First=0)& (STI First=0)&(HER First=0)
If it was Alcohol/Tobacco then:
(TOB First=1)& (ALC First=1)& (MJ First=0)& (HAL First=0)& (STI First=0)&(HER First=0)
For each of the 6 candidate drug categories:

- Compute proportion of Respondents who have ever used
- Compute proportion of Respondents whose first drug was the candidate drug
- Compute proportion of respondents who went on to consume another candidate drug, a Stimulant, Hallucinogen or Heroine
- Generate the Distribution of the number of years it took for first time users of a candidate drug to go on and consume another substance at an older age.

Proportion of people who have used drugs:
The first thing we wanted to do was look at how many the participants, male and female, have used some form of drug. In order to do this, an excel sheet called “Drug Data” was made using the following columns were used: ALCFLAG, CGRFLAG, CIGFLAG, COCFLAG, CRKFLAG, DAMTFLAG, ECSTMOFLAG, METHAMFLAG, PCPFLAG, PIPFLAG, PNRANYFLAG, PNRANMFLAG, PSYANYFLAG, PSYCHFLAG, SALVIAFLAG, SEDANYFLAG, SEDNMFLAG, SMKLSSFLAG, STMANYFLAG, STMNMFLAG, TRQANYFLAG, TRQNMFLAG. These column are the ones related to if a person has ever tried a drug.

1. In order to see if a participant had done some form of drug, another column called “Ever Used” was added, and the following formula was used =if(sum(across rows) > 0, 1 , 0). The way this formula works is that since all the have ever used questions are binary, if any of those questions were answered with a yes, the sum will be greater than 0 causing a 1 to appear in the “Ever Used” column. This was done for all participants
2. Adding the IRSEX column, which is whether the gender of the participant, participants were split into male (IRSEX=1) and female (IRSEX=2) using the filter function. Only the “IRSEX and Ever Used” columns put on a separate excel sheet called “Have Used” for analysis.
3. For both male and female groups, the number of people who have used a drug was calculated using =sum(“Ever Used”). Then the total number of people for male and female was calculated using =count(“Ever Used”). Then by using =sum(“Ever Used”)/count(“Ever Used”), the percentage of male/female who have used a drug were calculated. A pie chart was made using the resulting data.
4. To find how many people in total have used drugs, the number of males and number of females who have used drugs were added together. Then the total number of males and females were added together to get the total population. Finally, the two numbers were divided, and a pie chart was created using the results.

Proportion of drugs being used:

After having the overall usage of drugs for the dataset, we wanted to get the usage of individual drugs. In order to do this, we used the same excel sheet as before, “Drug Data”.

1. For every column, except “IRSEX” and “Ever Used”, the formulas =sum(down the column) and =count(down the column) were used to get the sum and count.
2. Then the sum and count were divided in order to get a percentage. This was done for all columns. A bar graph was then created from the resulting percentages.
3. A secondary bar graph was created from the PNRANYFLAG, PNRANMFLAG, PSYANYFLAG, PSYCHFLAG, SEDANYFLAG, SEDNMFLAG, STMANYFLAG, STMNMFLAG, TRQANYFLAG, TRQNMFLAG. These columns correspond to if the participant has used a prescription drug and if they have ever misused it. This graph compared use vs misuse.
4. In order to get the proportion of male and females using a certain drug, the data was filtered using IRSEX and put in another excel sheet called “Usage by Male and Female”.
5. The sum and count for males and females were calculated for each column, except “IRSEX”, using =sum(down the column) and =count(down the column).
6. Once the counts and sums were calculated, the total usage was calculated by adding counts for both male and female for that drug together.
7. The percentage was then calculated by dividing male use by total usage and female use by total usage for each drug.
8. Histograms were made with the resulting data.

Age of First Use:

An excel sheet was created using “Age of First Use” column from “FinalData.csv”. This column contains the first age at which they used any drug. Not a specific drug. We wanted to see the distribution of age of first drug use.

1. The average age of first use was found by taking the average of the whole column.
2. An “Age Column” was made that included ages from 0 to 30.
3. Under the “Data” tab, “Data Analysis” was used to create a histogram on a separate page.
4. Age “0” row was deleted and “More” was replaced with “31+”.
5. The frequencies were then added together using =sum(down the column).
6. Each frequency was then divided by the total. The histogram selections were changed from frequency column to percentage column.

Income data breakdown:

We wanted to break down participants into their respective income bracket and see how each income bracket uses drugs.

1. Created an excel document called “Income Data” using “IRFAMIN3”. This column contains data about the total household income of participants.
2. Using the filter function in Excel, the data for each of the seven categories was filtered out.
3. For all seven columns, =count(down the column) was used to count the total number of people in that category.
4. The number of people in each category were summed to get the total number of participants.
5. The count of each category was divided by the total number of participants to get the percentage of each income bracket.
6. A pie chart was then made using the results.

Drug usage breakdown by income bracket:
Here we wanted to see the different drug usage rate by income bracket to see if there is a trend drug use by income.

1. Created an excel sheet called "Income Drug Data" that contains the following columns: IRFAMIN3, ALCFLAG, CGRFLAG, CIGFLAG, COCFLAG, CRKFLAG, DAMTFLAG, ECSTMOFLAG, METHAMFLAG, PCPFLAG, PIPFLAG, PNRANYFLAG, PNRANMFLAG, PSYANYFLAG, PSYCHFLAG, SALVIAFLAG, SEDANYFLAG, SEDNMFLAG, SMKLSSFLAG, STMANYFLAG, STMNMFLAG, TRQANYFLAG, TRQNMFLAG. These columns correspond to if a person has ever tried a drug.
2. Using the filter function, each category's data was filtered out and placed into its own separate excel sheet.
3. In each category excel sheet, the sum was taken using =sum(down the column) for each of the different drug categories. Since the data is binary, it will give the number of people who used that drug in that income bracket.
4. Once the number of people that used a drug was calculated for each income bracket, the all the resulting data was copied into a separate sheet called "Income and All Drugs". Then for each drug column, the total number of responses from each income bracket was added together.
5. Once the total number of responses for each drug type was found, the number of responses for each individual income bracket was then calculated by dividing the number of people that have used that drug in that income bracket by the total number of people who have used that drug. For example, Income bracket 1 had 3,318 people who responded yes to using alcohol. There was a total of 40478 people across all income brackets that responded yes. By 3,318/40,478 = 8.2%, the percentage of people in income bracket 1 who had reported trying alcohol.
6. Bar charts were made using the resulting data.

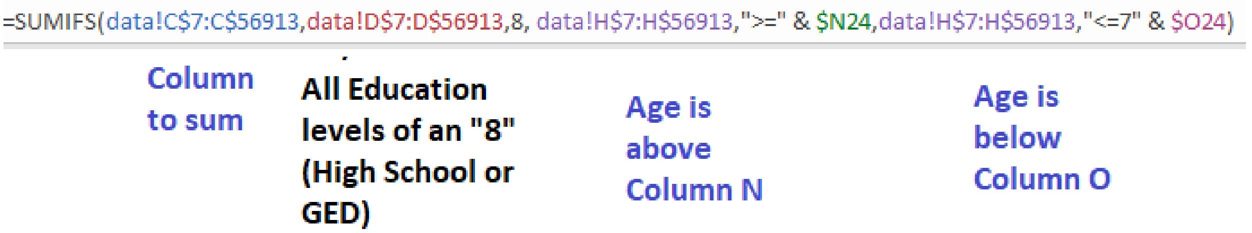
Drugs by Age and Education/Employment:

To further assess the data by income status, we decided to explore our database by (a) focusing on a total of six drug types and (b) aggregating the age of first use by Employment and Educational status. We further divided the drugs into two categories. We selected three of the most likely gateway drugs (Tobacco, Marijuana, and Alcohol) along with three popular drugs most likely influenced by the gateway candidates (Stimulants, Hallucinogens, and Opiates). For all columns regarding drugs, the default value of "0" was used to indicate that an individual had never used the substance in question.

1. Pre-processing:
    - From the "FinalData.csv", we extracted the following columns
        i. **IREDUHIGHST2**: Highest level of education, ranging from 5th grade to College Graduate
        ii. **IRWRKSTATG18**: Employment status of respondents 18 years and older. This category ranged from Full Time to Other (e.g. retirement).
        iii. **Stimulants**: We utilized IRCRKAGE, IRCOCAGE, and IRMETHAMAGE which accounted for the earliest age of used for Crack, Cocaine, and Methamphetamines respectively.
        iv. **Hallucinogens**: We utilized the column Code IRHALLUCAGE for age of first Hallucinogen use.
        v. **Opiates**: The only opiate that we had earliest use data for was Heroine.
        vi. **Tobacco**: We utilized the columns IRCIGAGE, IRCGRAGE, and SMKLSSTRY to find the earliest age of tobacco use from Cigarette, Cigar, and Smokeless Tobacco respectively.
        vii. **Marijuana**: We utilized IRMJAGE for the earliest use of Marijuana.

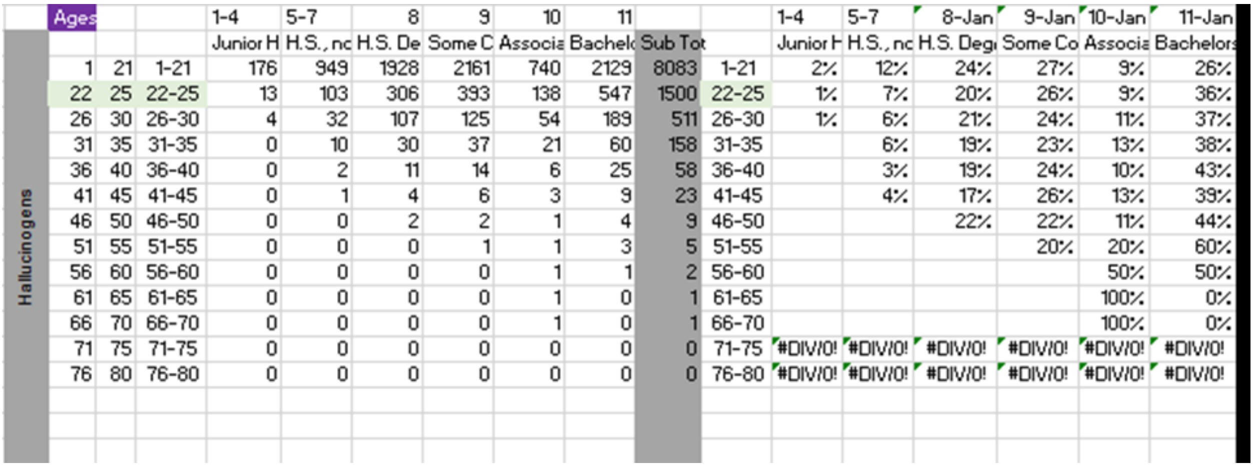
viii. **Depressants**: We used the column code IRALCAGE (Alcohol) for depressant usage.

2. Column Reduction
   - **Stimulants**: We used Excel to find the minimum age of use for IRCRKAGE, IRCOCAGE, and IRMETHAMAGE, thereby finding the earliest age of use between the three stimulants.
   - **Hallucinogens**: There was no processing necessary for this data point.
   - **Opiates**: There was no pre-processing needed since there was only one data point for this drug.
   - **Tobacco**: We found the earliest use of tobacco usage by utilizing the "min()" function.
   - **Marijuana**: There was no pre-processing needed since there was only one data point for this drug.
   - **Depressants**: We utilized the Alcohol Code IRALCAGE only as this was the only depressant drug with first age usage.
3. Percentages
   - Using the excel function "SUMIF()", we summed all columns who met multiple criteria based on the drug and category tested. For example, the following code is used to find all respondents with a High School or GED equivalent between the ages of 1 and 21.



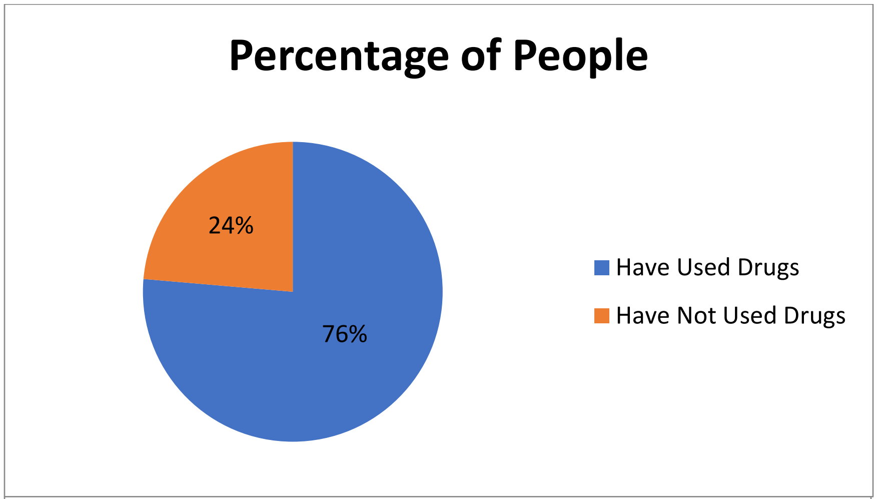
```
=SUMIFS(data!C$7:C$56913,data!D$7:D$56913,8, data!H$7:H$56913,">=" & $N24,data!H$7:H$56913,"<=7" & $O24)
```

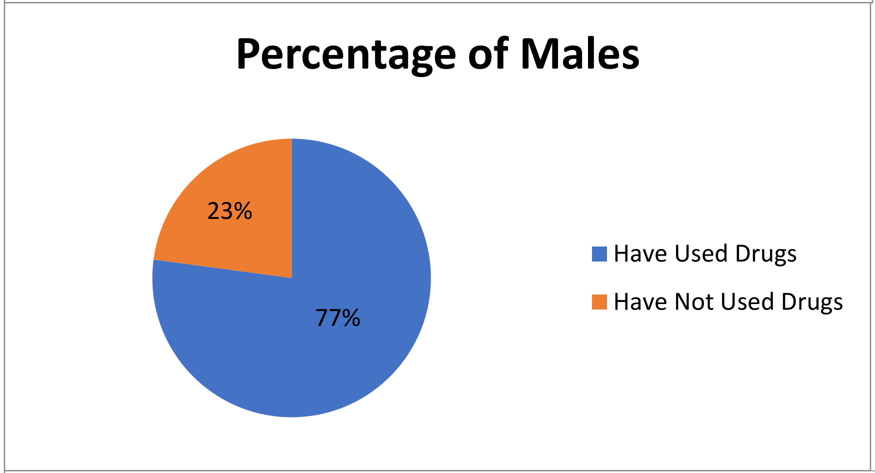
   - After the total count of individuals taking a drug within a certain age range, the subtotals are calculated and utilized to generate a duplicate table of percentages. This was utilized to facilitate ease of use when reading the charges by displaying percentages rather than total counts. All values of "0" were removed and ignored for graphing purposes.

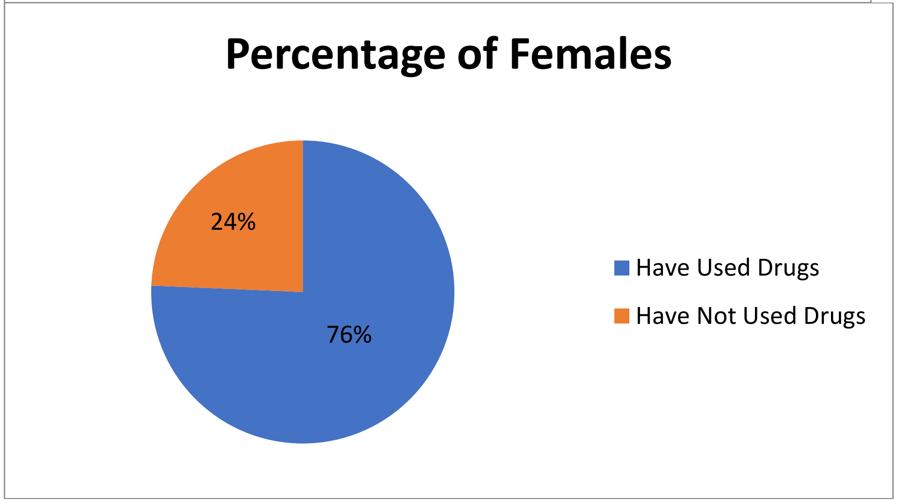


	Ages			1-4	5-7	8	9	10	11			1-4	5-7	8-Jan	9-Jan	10-Jan	11-Jan
				Junior H	H.S., nc	H.S. De	Some C	Associa	Bachelc	Sub Tot		Junior H	H.S., nc	H.S. Degi	Some Co	Associa	Bachelors
Hallucinogens	1	21	1-21	176	949	1928	2161	740	2129	8083	1-21	2%	12%	24%	27%	9%	26%
	22	25	22-25	13	103	306	393	138	547	1500	22-25	1%	7%	20%	26%	9%	36%
	26	30	26-30	4	32	107	125	54	189	511	26-30	1%	6%	21%	24%	11%	37%
	31	35	31-35	0	10	30	37	21	60	158	31-35		6%	19%	23%	13%	38%
	36	40	36-40	0	2	11	14	6	25	58	36-40		3%	19%	24%	10%	43%
	41	45	41-45	0	1	4	6	3	9	23	41-45		4%	17%	26%	13%	39%
	46	50	46-50	0	0	2	2	1	4	9	46-50			22%	22%	11%	44%
	51	55	51-55	0	0	0	1	1	3	5	51-55				20%	20%	60%
	56	60	56-60	0	0	0	0	1	1	2	56-60					50%	50%
	61	65	61-65	0	0	0	0	1	0	1	61-65					100%	0%
	66	70	66-70	0	0	0	0	1	0	1	66-70					100%	0%
	71	75	71-75	0	0	0	0	0	0	0	71-75	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
	76	80	76-80	0	0	0	0	0	0	0	76-80	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!

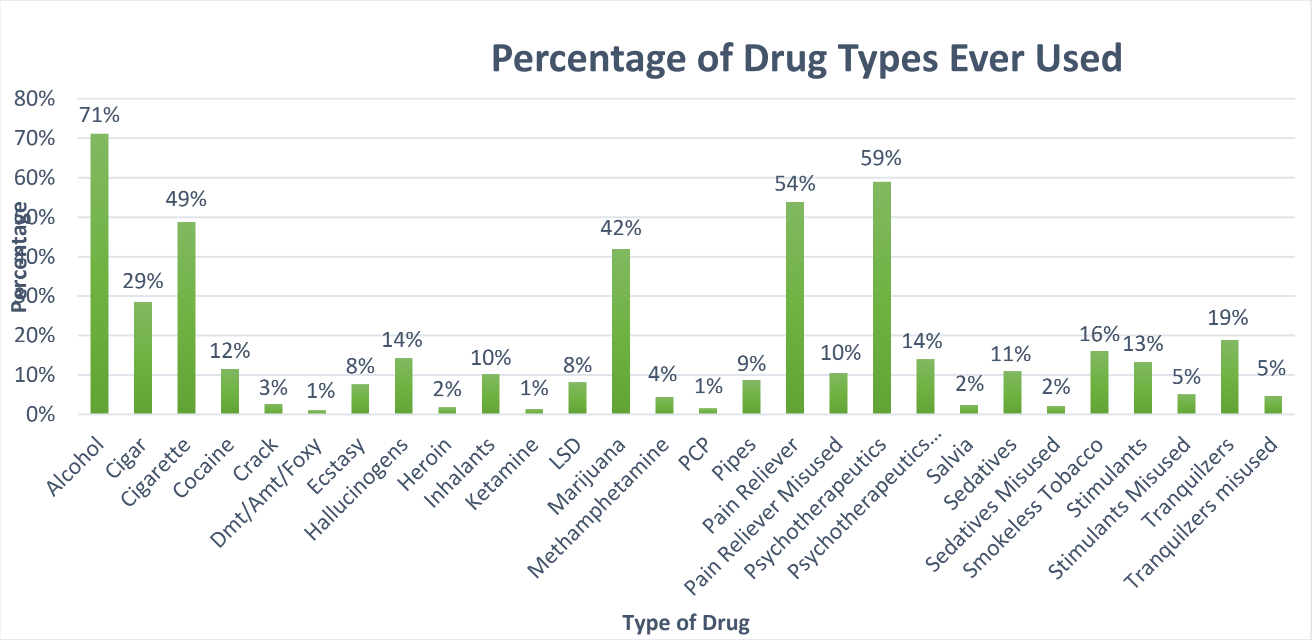
Results/Discussion:


Percentage of People

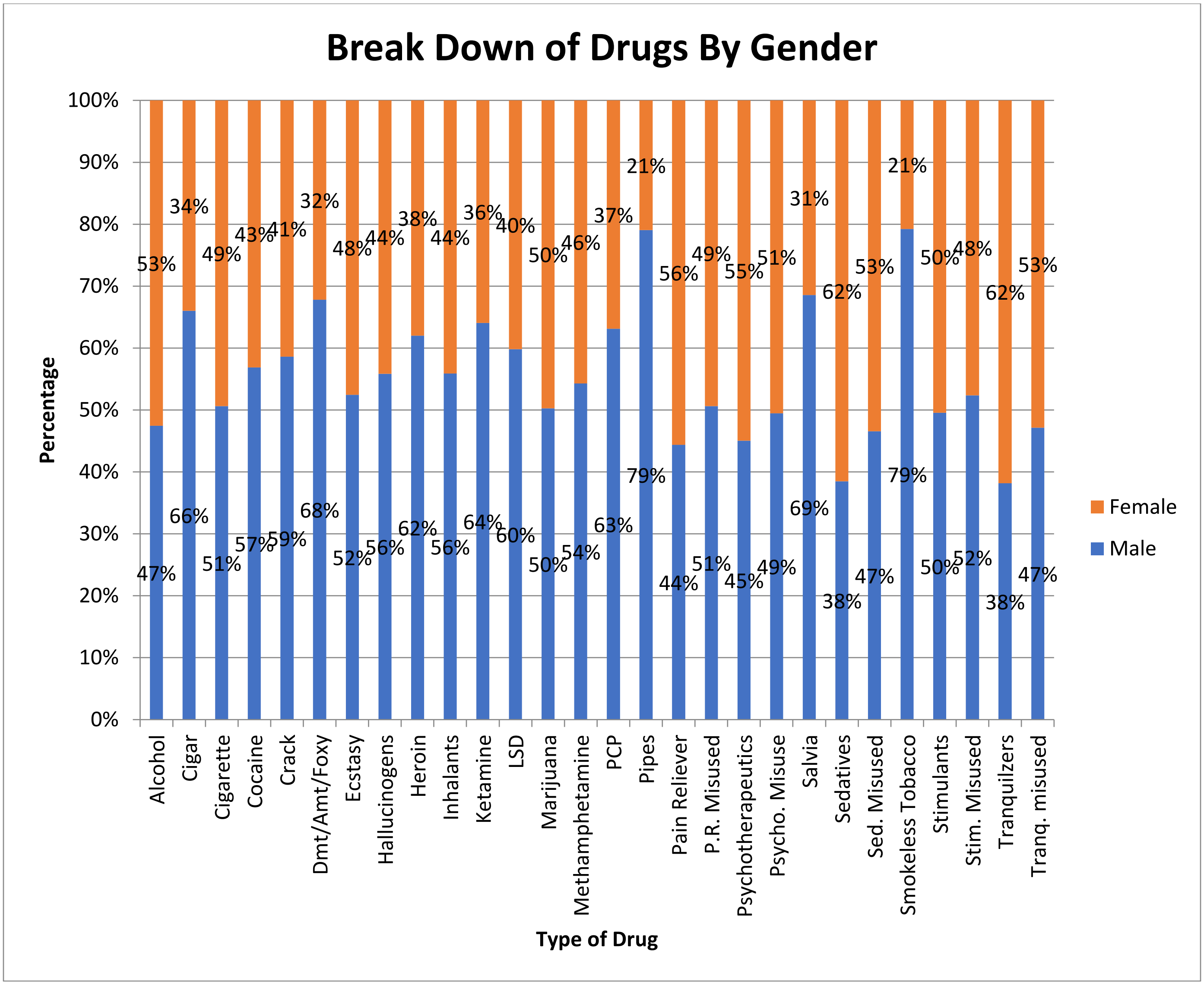

Percentage of Males


Percentage of Females

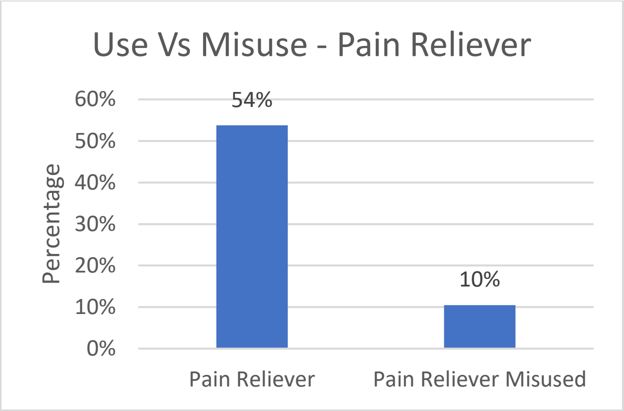
Looking at the distribution of people who have used drugs, 76% of people have used some form of drug. Broken down by male and female, 77% of males and 76% of females used some form of drug. There is no real difference in drug usage between males and females.

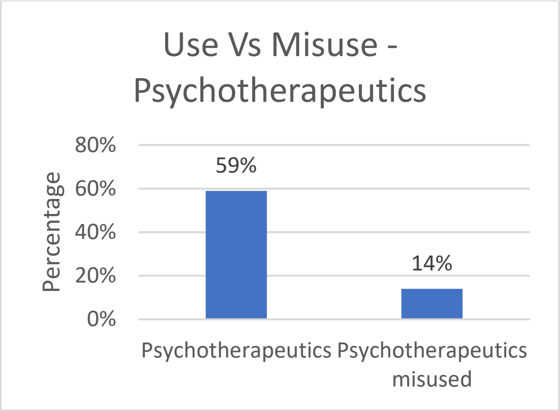


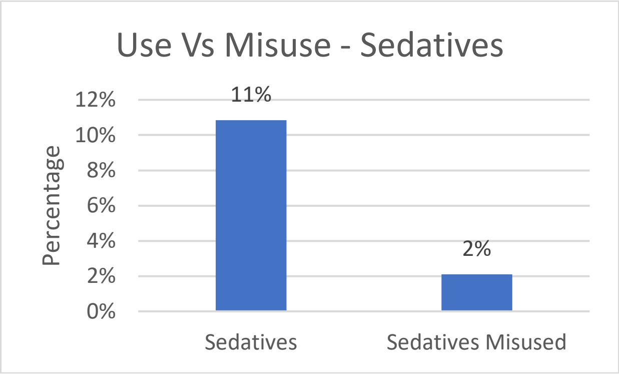
When the usage of individual drugs is broken down, some drugs are more popular than others. Alcohol, Cigars, and Cigarettes are popular legal drugs. Marijuana is a popular illicit drug. Something interesting of note is that Pain Relievers and Psychotherapeutics are two drugs that have a high use rate. What’s interesting is that these are prescription drugs given to people. It’s quit shocking that 54% of people in the USA have tried a pain reliever at some point and 59% of people have tried a psychotherapeutic drug.

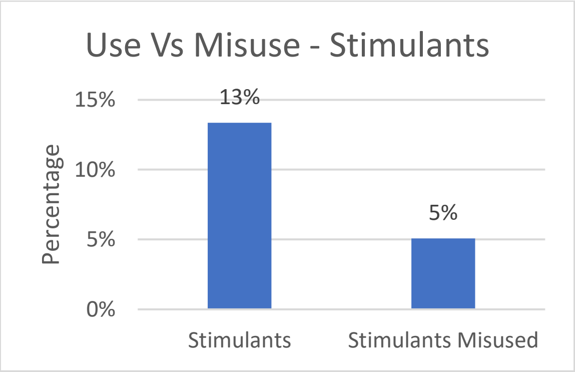


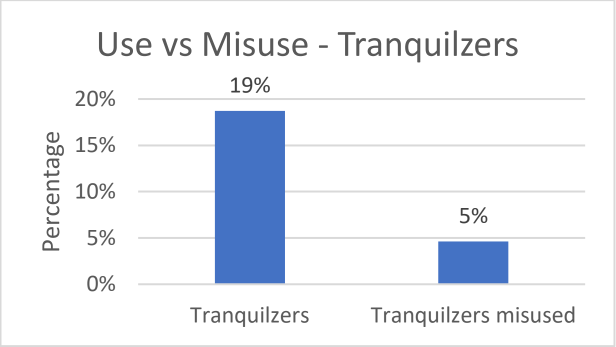
Breaking down the usage of drugs by gender, the two genders aren't using the drugs at the same rate. Tobacco products have a higher usage rate in males than they do in females. It is also seen that a lot of the prescription drugs are being used by females more than by males. These trends may be due in part to perceived gender norms. It may be masculine to smoke a cigar, so women may not use it as much. Men might be told to just "walk off the pain" and may not be going to get prescription drugs as much as women. It would be interesting to do an investigation looking at the perception of each drug by gender to see if there is a correlation between why certain drugs are used more by one gender than the other.



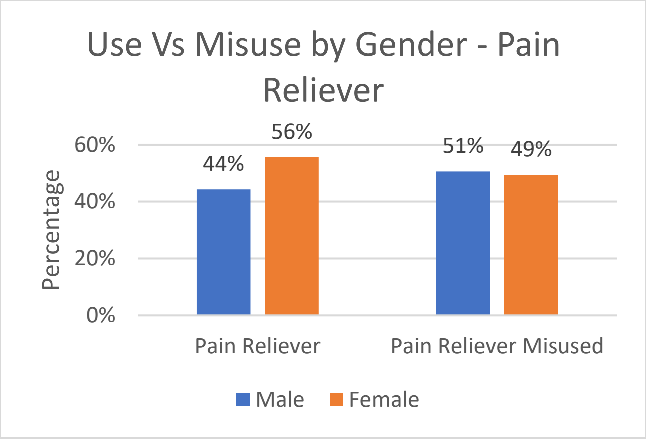


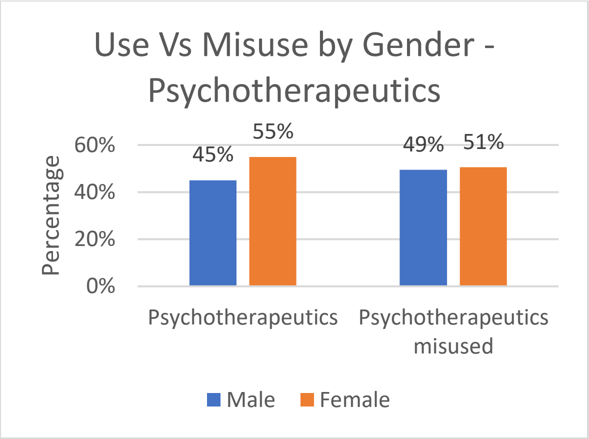


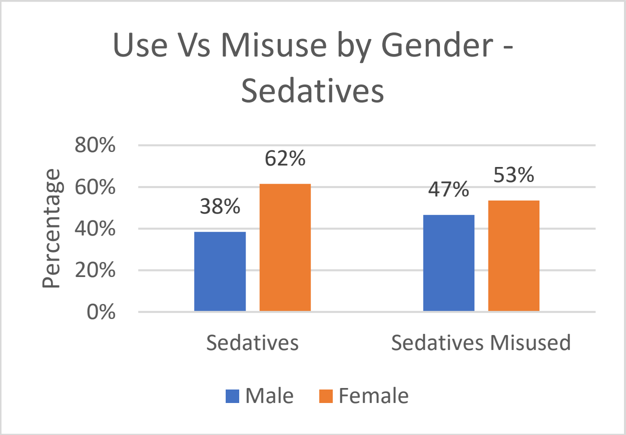


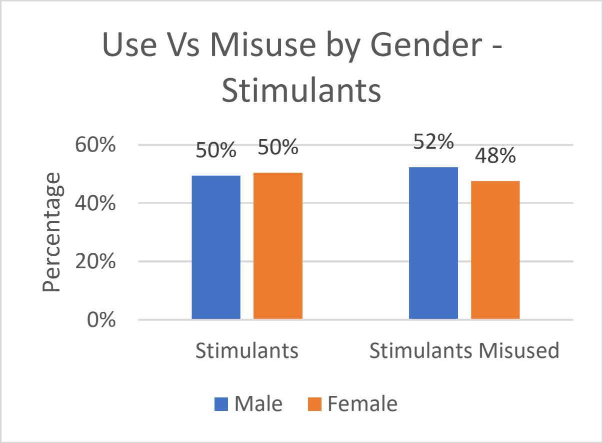


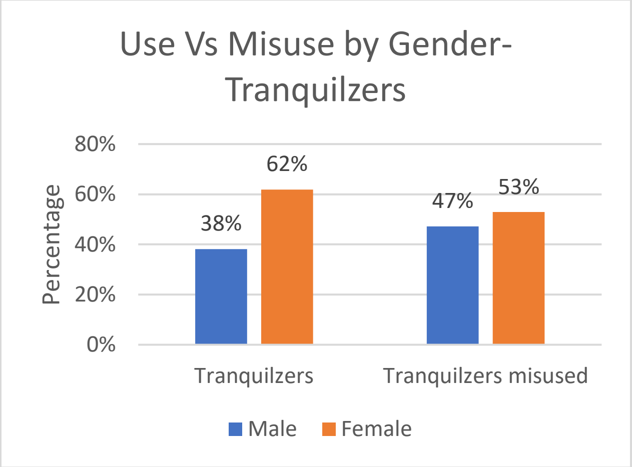
When we compare the Use vs Misuse of prescription drugs, it's noticeable that although a good percentage of people are being given these drugs, a very small percentage of people are misusing it. One thing that should be noted is that in the survey, it wasn't defined what misuse was. Some participants may be misusing these drugs but think they aren't and therefore aren't reporting misuse.



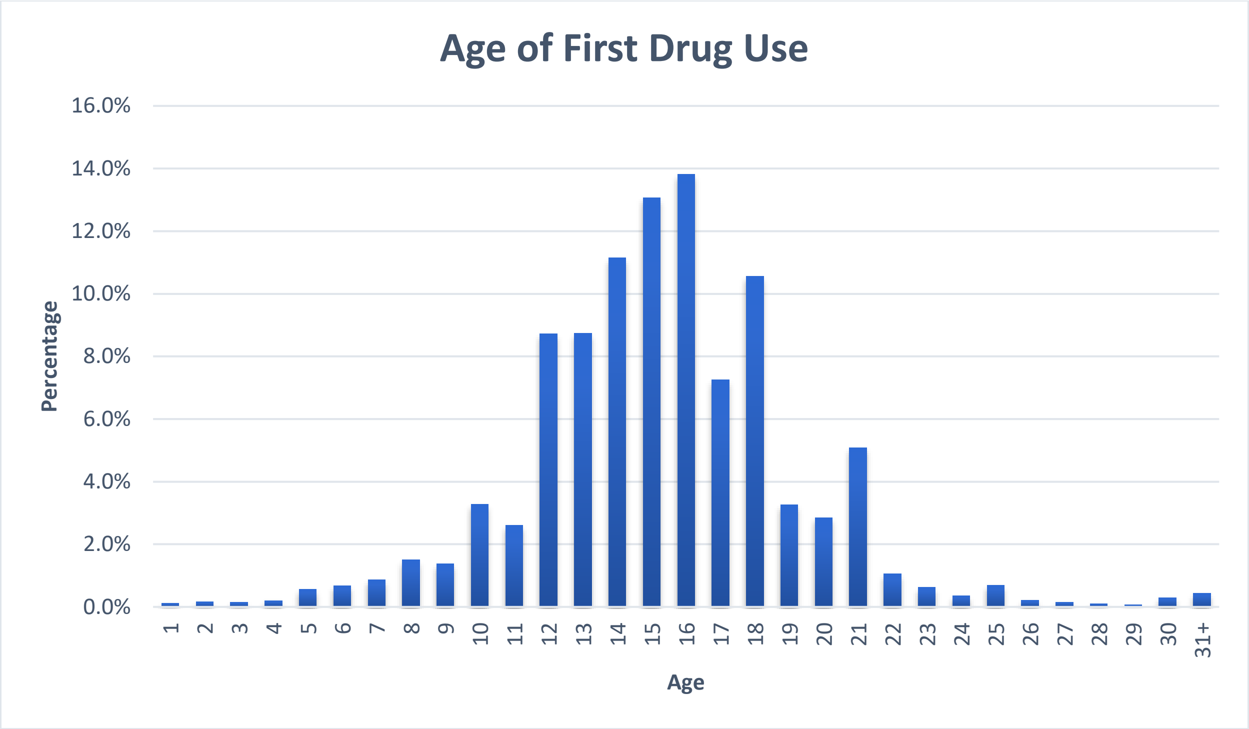




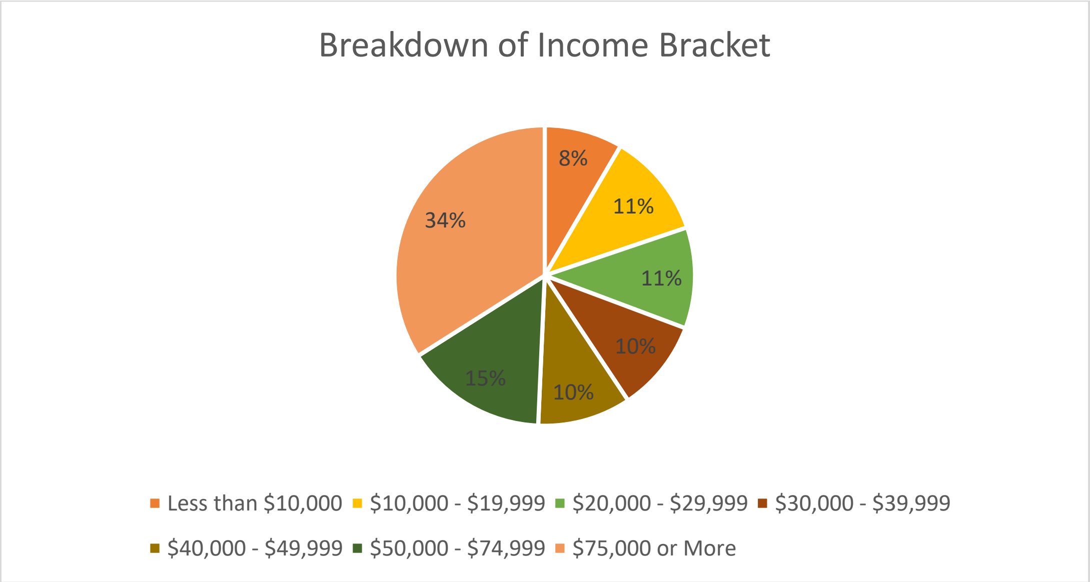




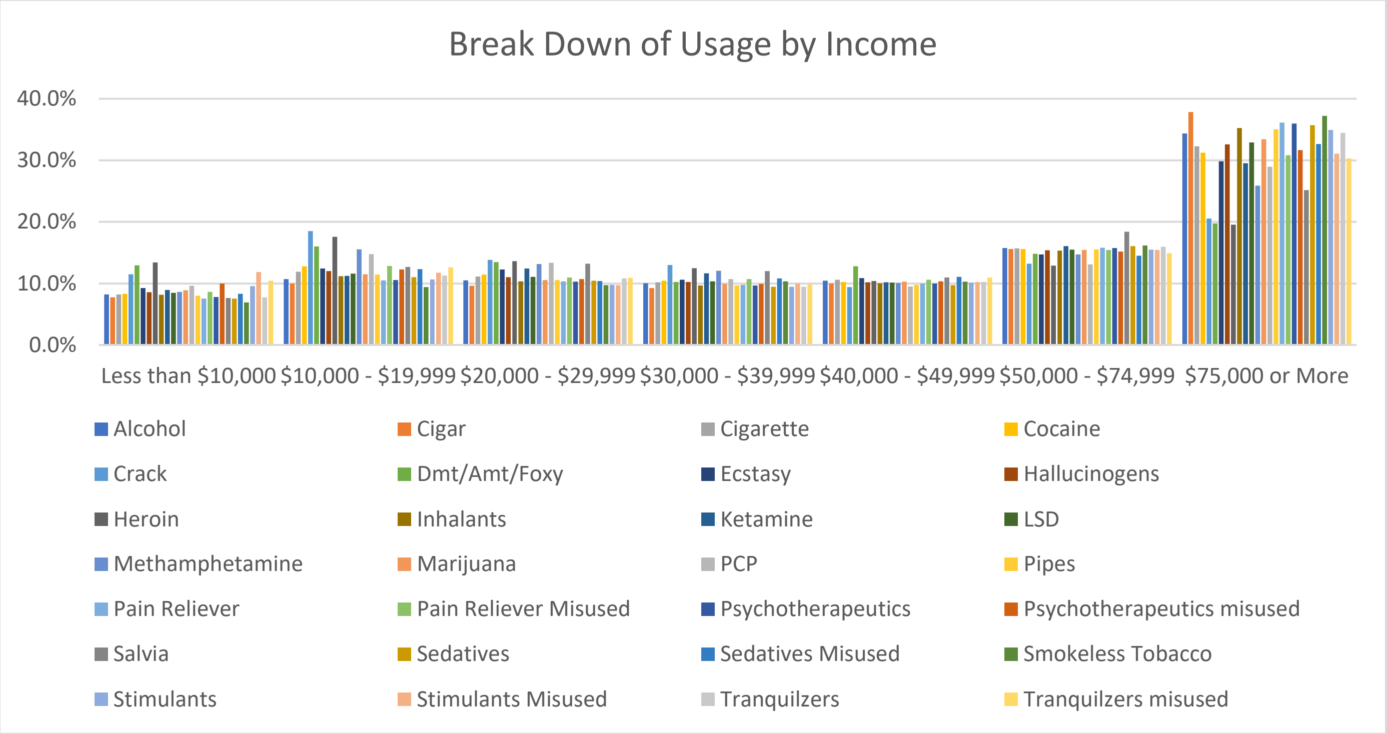
When we break down Use Vs Misuse by gender, you can see that even though women are using these prescription drugs more than men, the misusage between the genders is about the same. It's not like one gender is abusing these prescription drugs more than the other. It's the essentially the same for both genders.

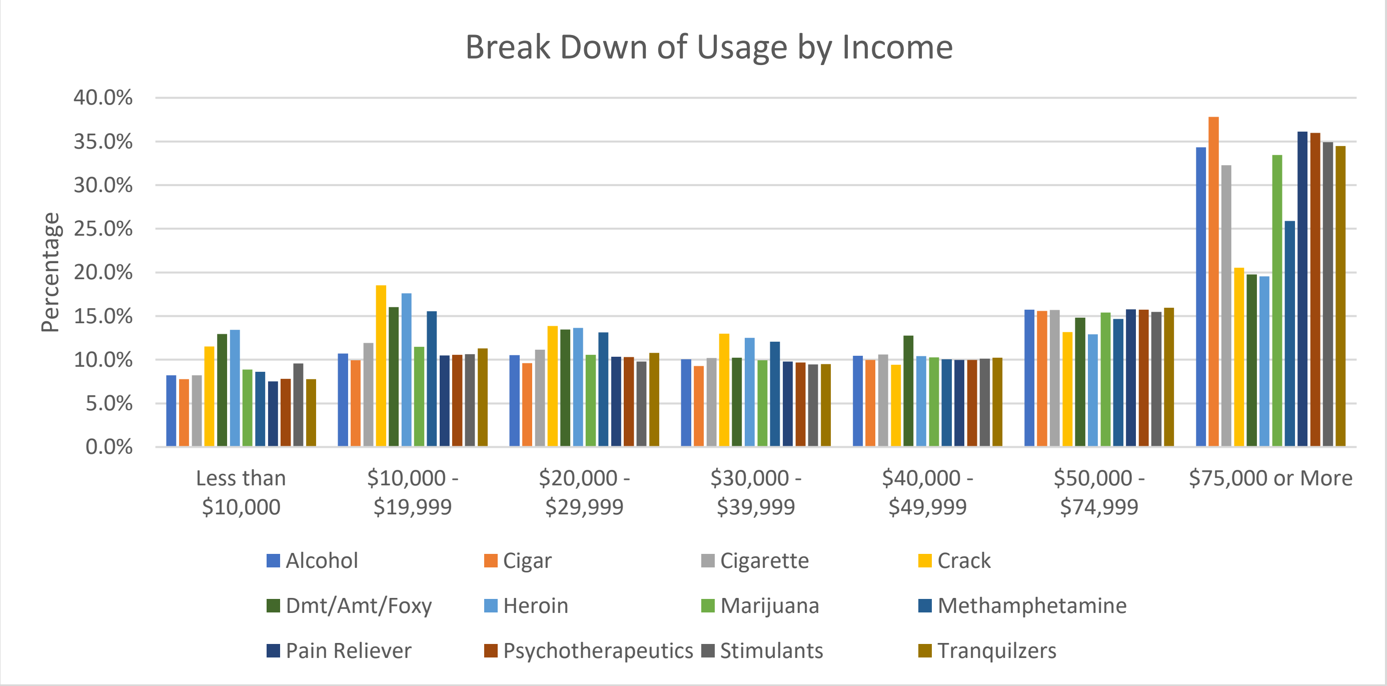


After we had done a basic look at drug usage, we wanted to analyze the age of first use. We calculated the average age of first use to be 15.35. This means that people are using the drug around 15-16 years of age. Looking at the histogram that average makes sense since this histogram does have a long tail from 31 to about 80 years of age. One interesting thing about this histogram is that there is a spike at age 18 and 21. This is most likely since the legal age for most tobacco products is 18 and for alcohol is 21.



When we break down the data by income, for almost all the income categories, there is an equal representation. However, 34% of people surveyed had a total house income of $75,000 or more. One way to improve the survey would be to breakdown the $75,000 and more category into more categories so that all economic backgrounds are represented equally instead of being skewed. This would allow for a richer analysis and allow for important trends to stick out more.

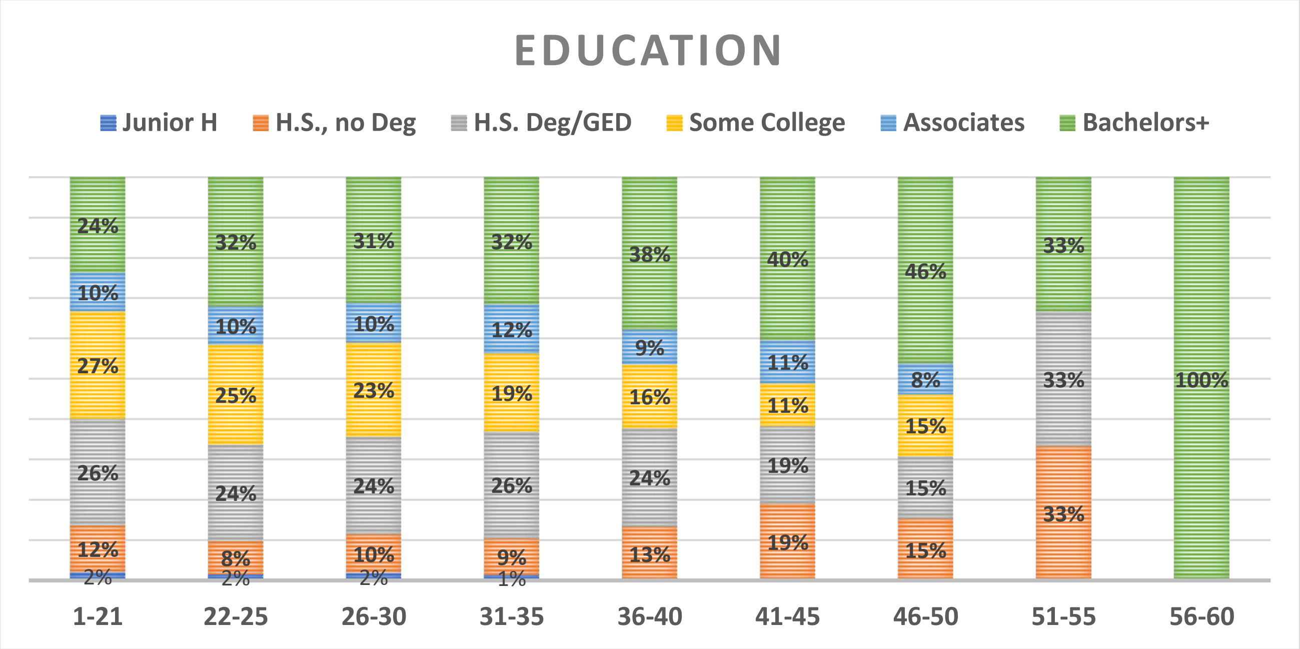


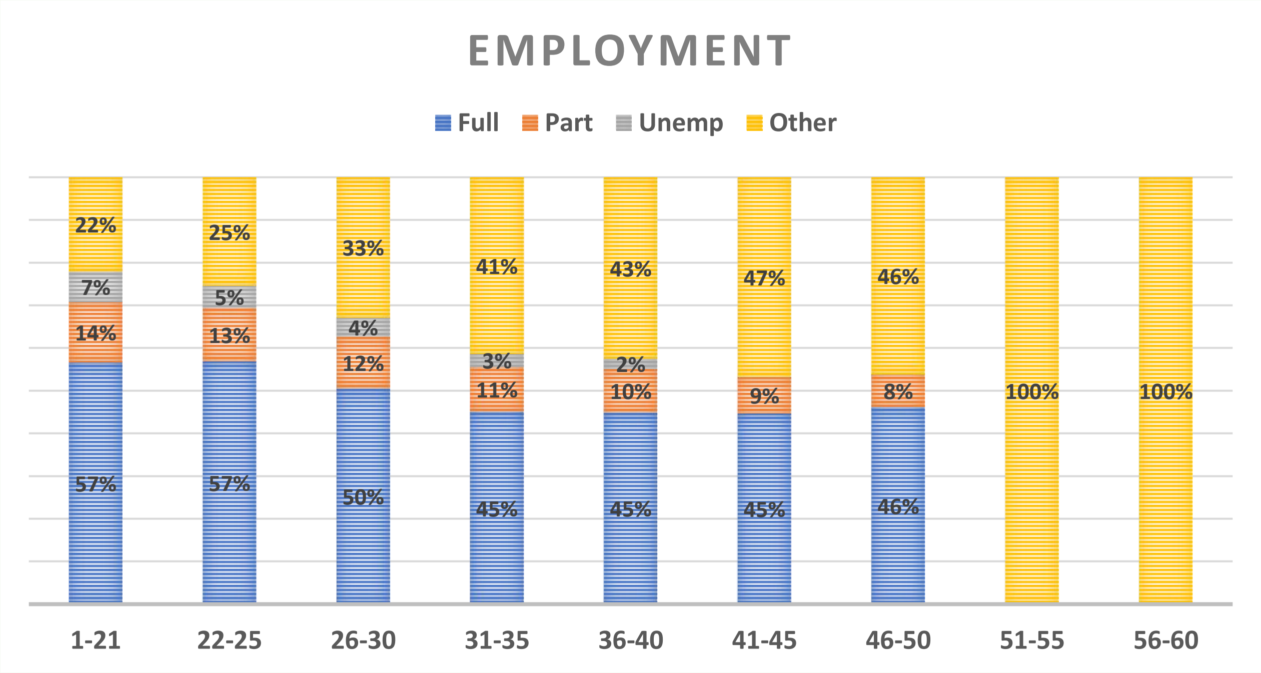


Results of Data by Employment and Education Level:

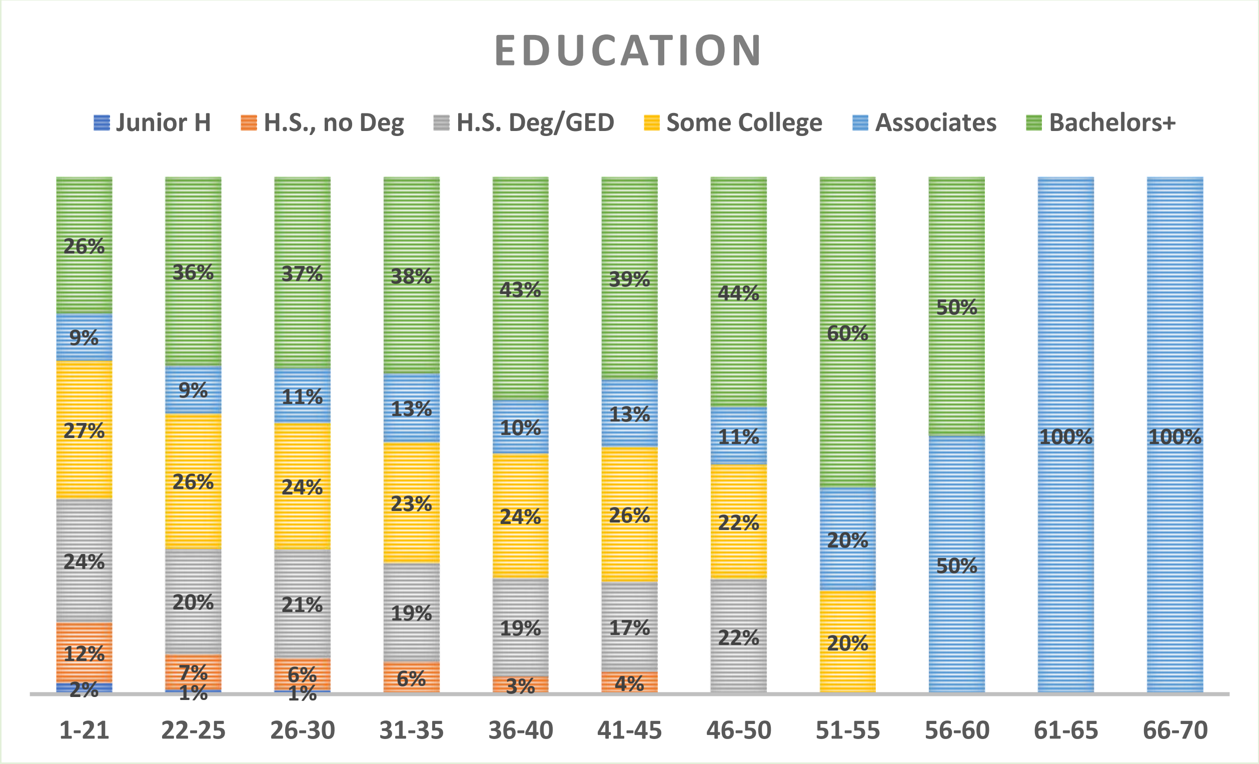
While there were no overall results pertaining to Education level, there was a notable pattern among Employment status. Specifically, there were a large percentage of Full Time Employees who had tried almost every drug very early in life (21 and earlier). Conversely, there were a sizable number of respondents who noted “Other” as their category of employment. These respondents overwhelmingly had first use experiences later in life.

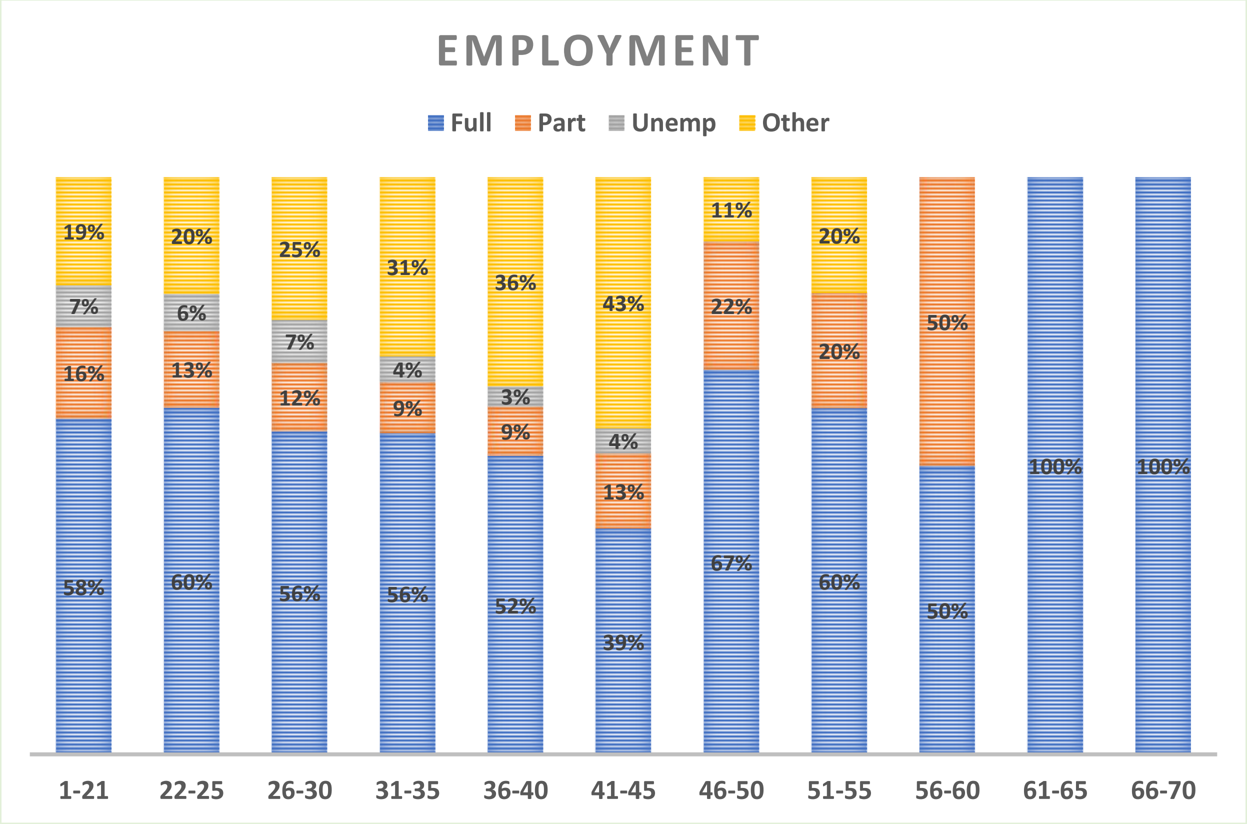
Stimulants



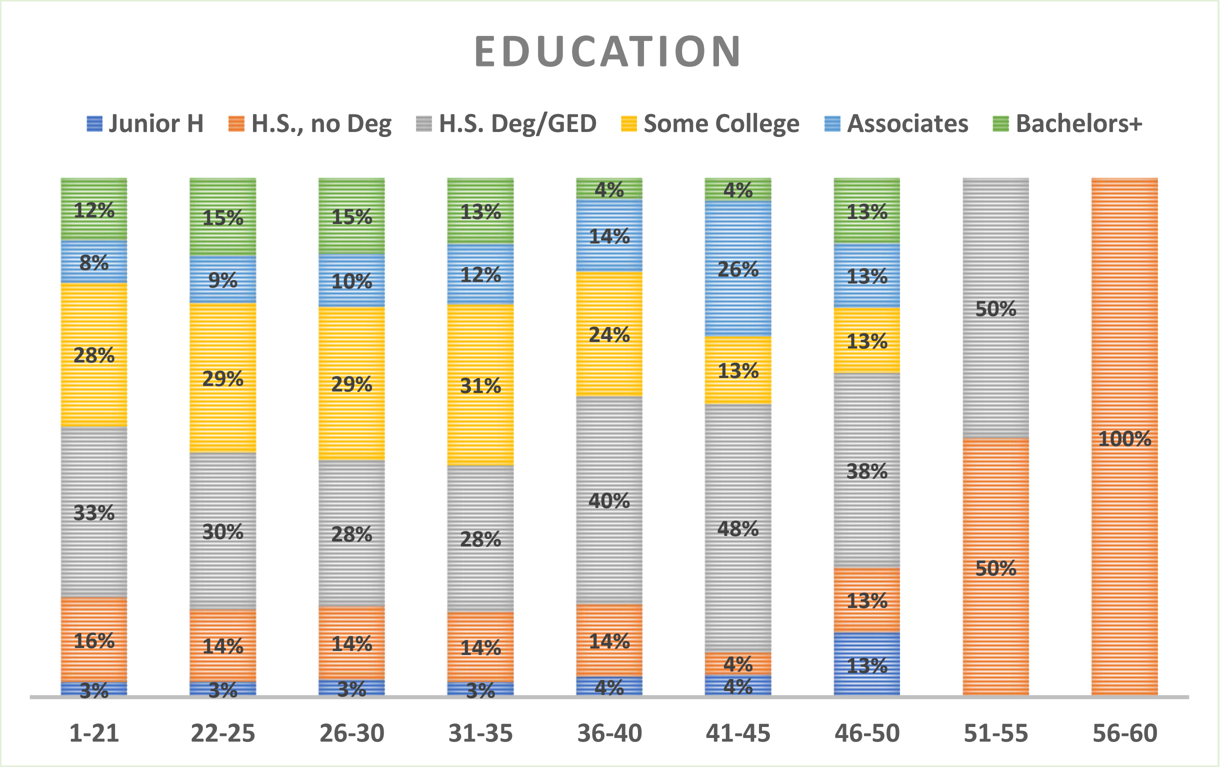


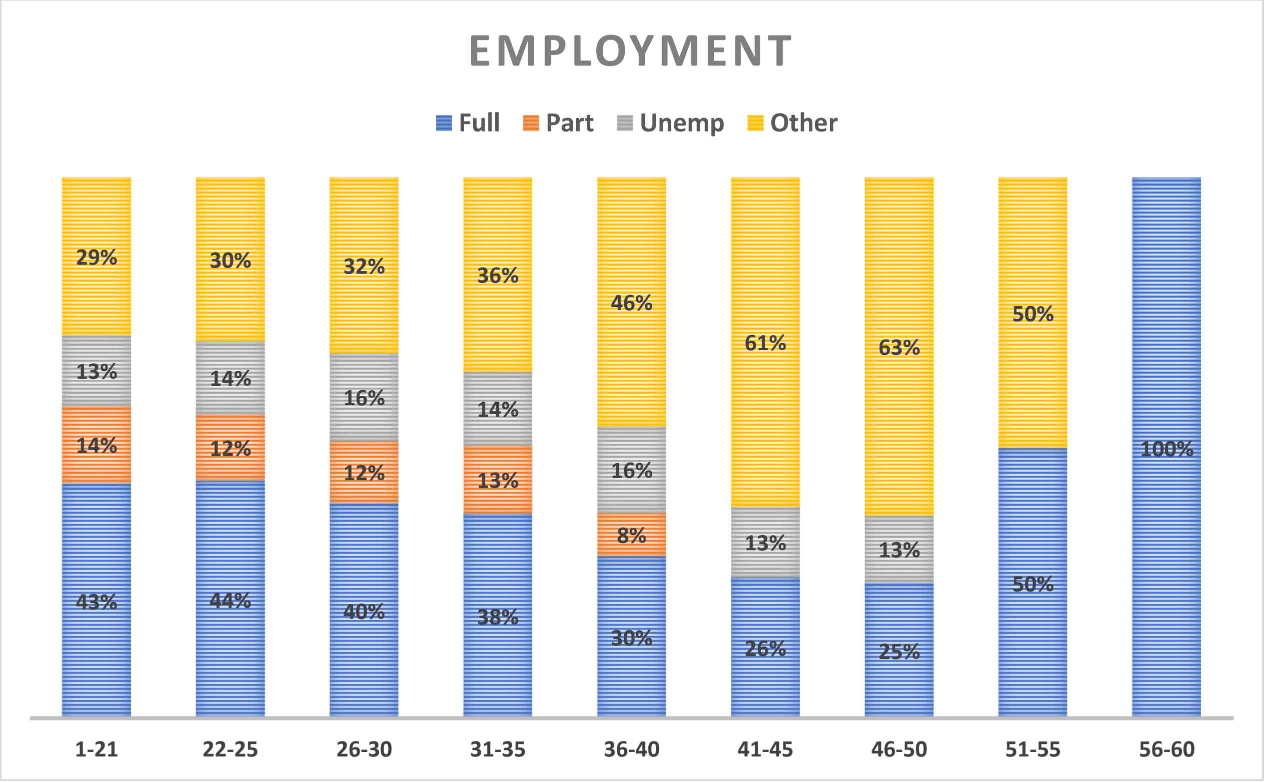
Hallucinogens:



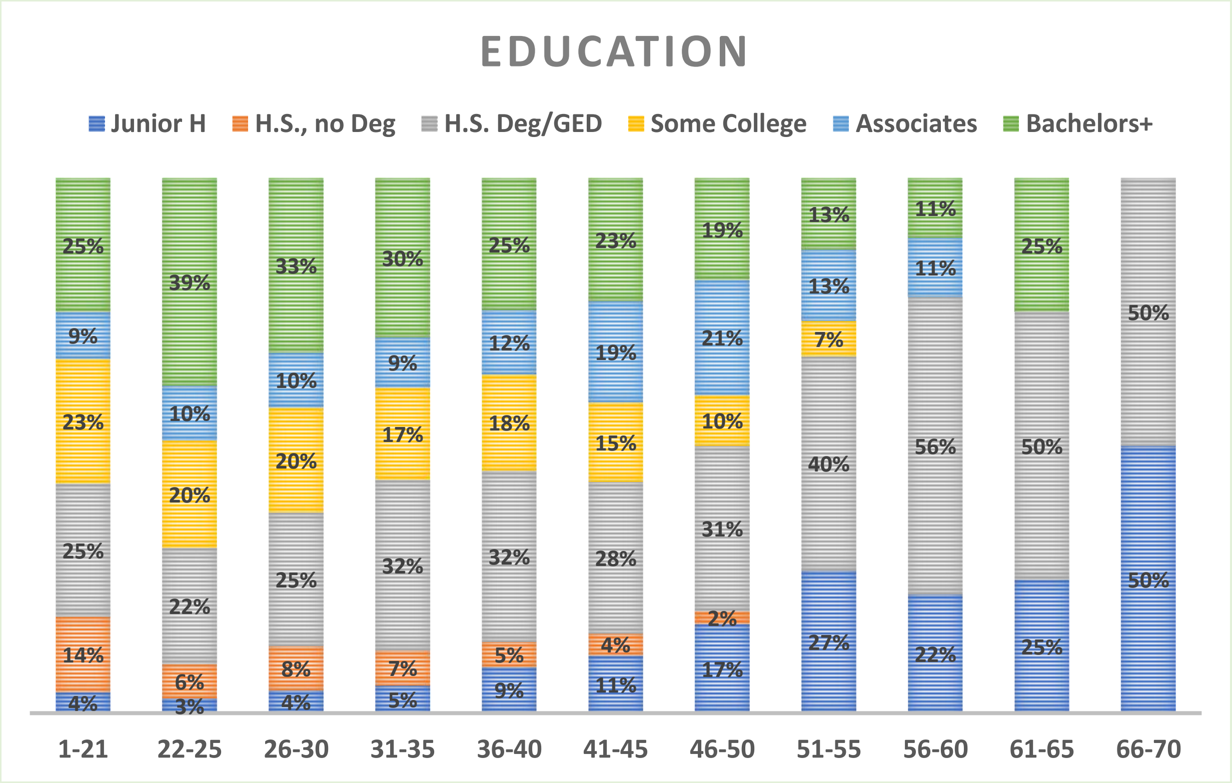


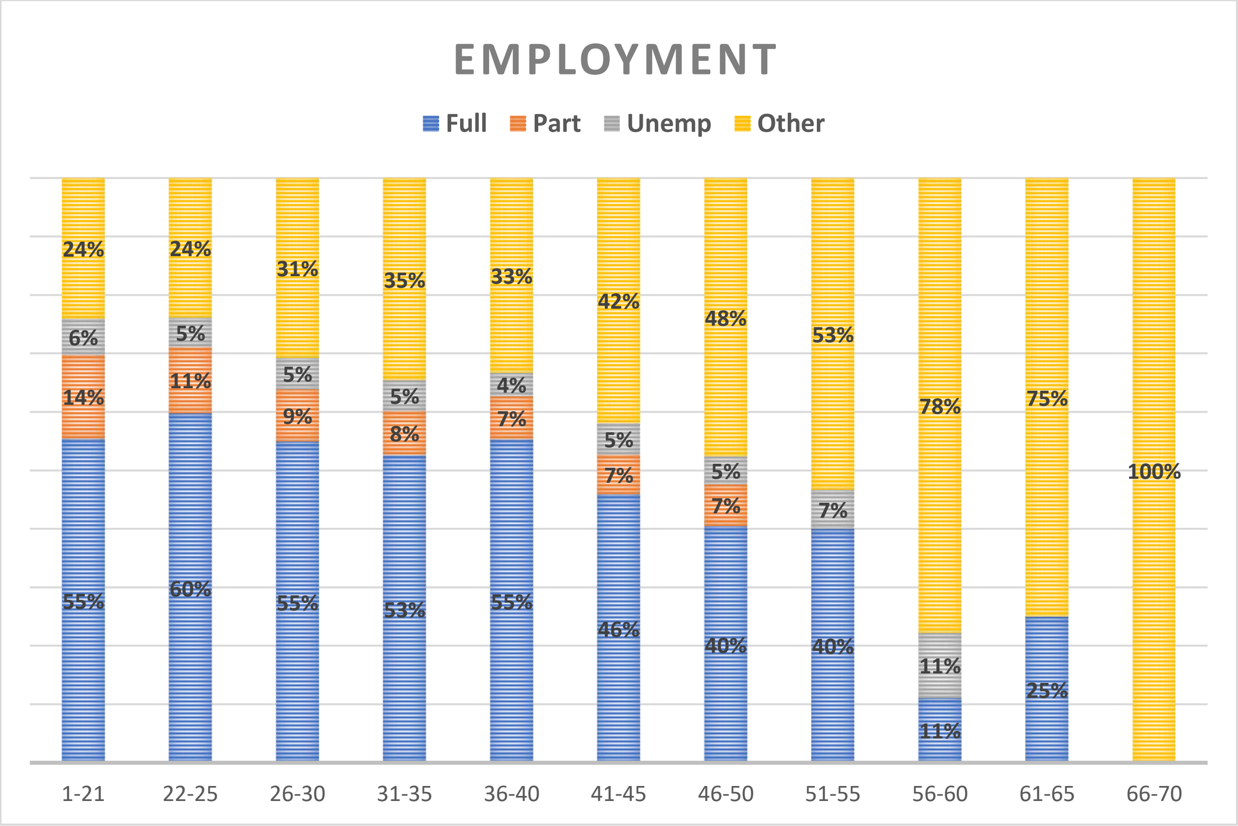
Opiates (Heroine):



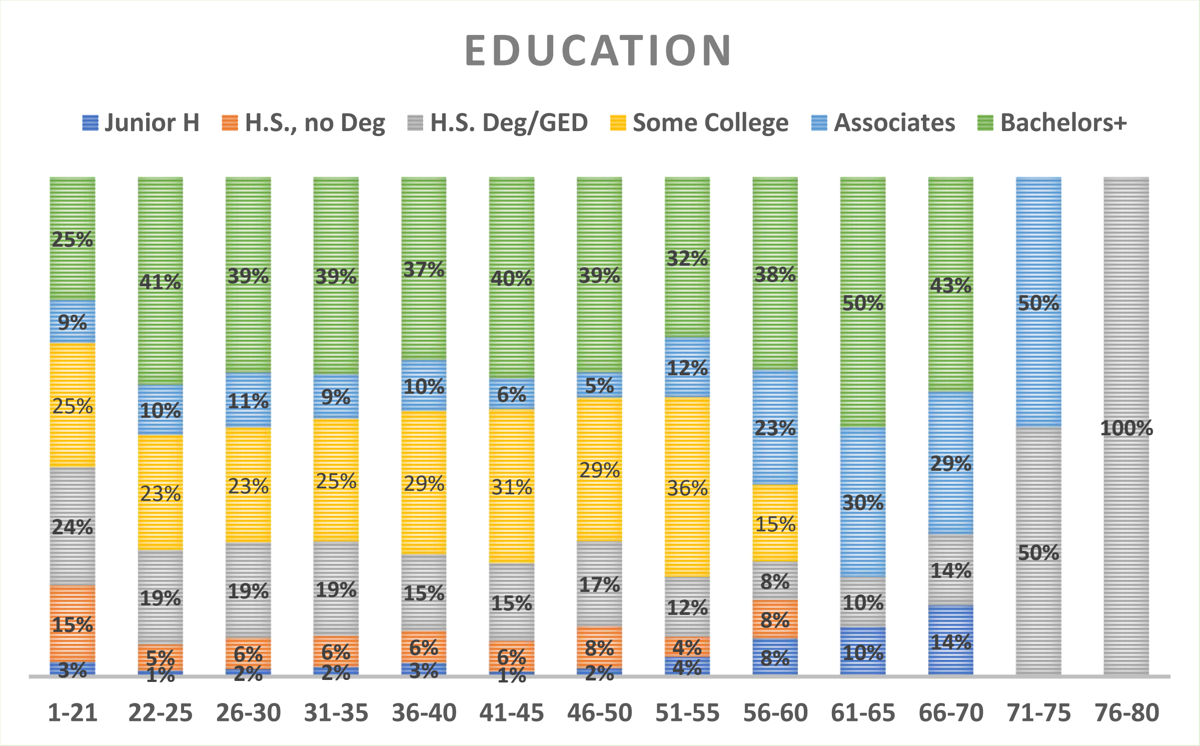


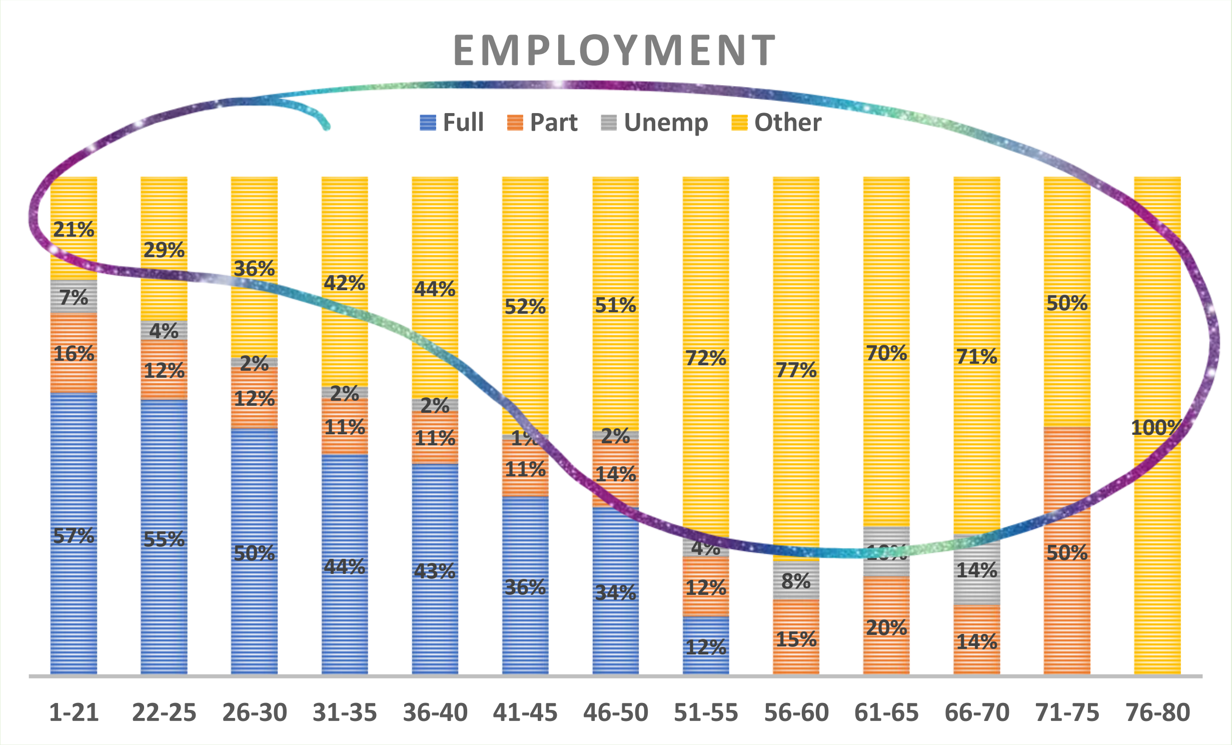
Tobacco:


EDUCATION
Junior H
H.S., no Deg
H.S. Deg/GED
Some College
Associates
Bachelors+
25%
9%
23%
25%
14%
4%
39%
10%
20%
22%
6%
3%
33%
10%
20%
25%
8%
4%
30%
9%
17%
32%
7%
5%
25%
12%
18%
32%
5%
9%
23%
19%
15%
28%
4%
11%
19%
21%
10%
31%
2%
17%
13%
13%
7%
40%
27%
11%
11%
56%
22%
25%
50%
25%
50%
50%
1-21
22-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70

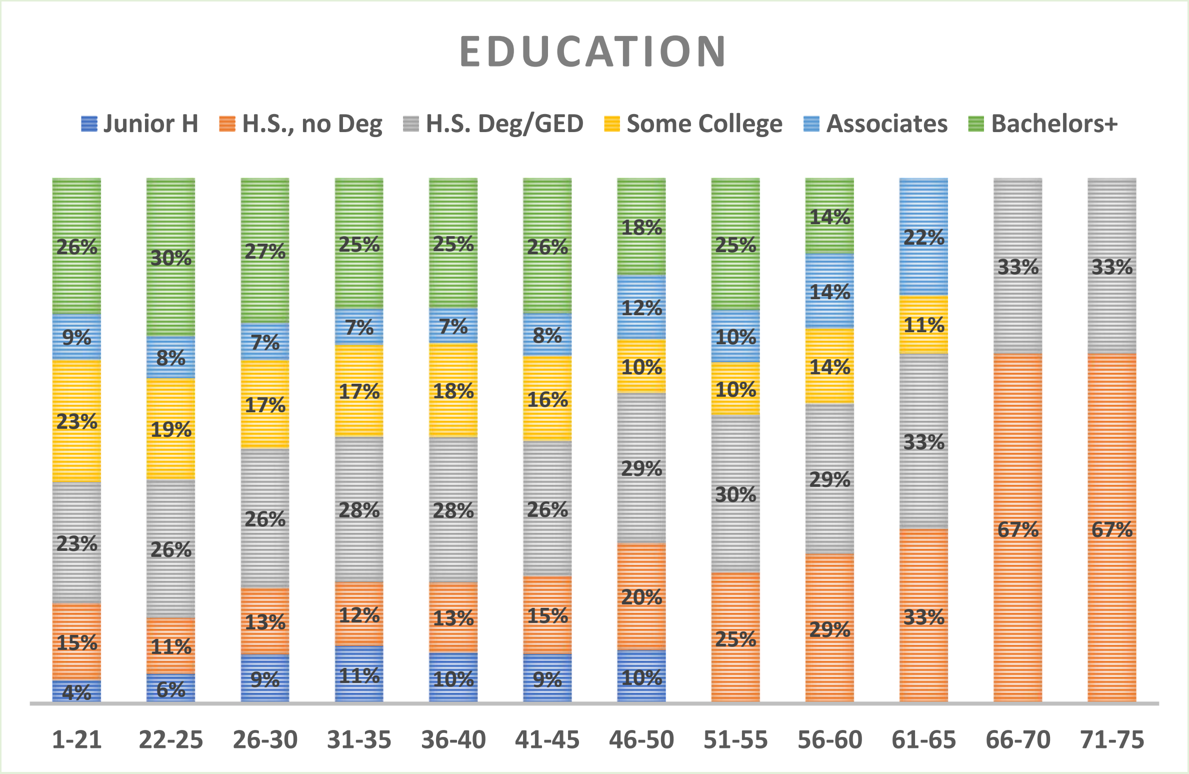

EMPLOYMENT
Full
Part
Unemp
Other
24%
6%
14%
55%
24%
5%
11%
60%
31%
5%
9%
55%
35%
5%
8%
53%
33%
4%
7%
55%
42%
5%
7%
46%
48%
5%
7%
40%
53%
7%
40%
78%
11%
11%
75%
25%
100%
1-21
22-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70

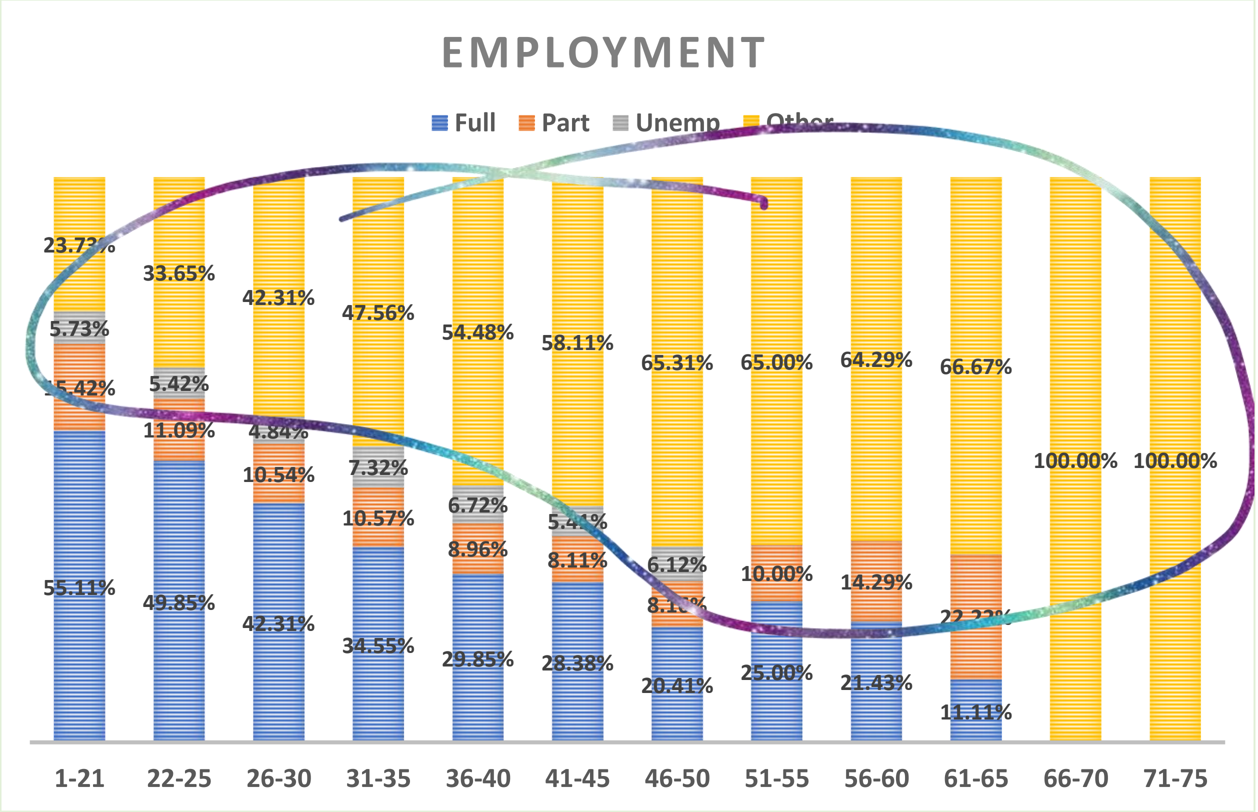
Marijuana:


EDUCATION
Junior H
H.S., no Deg
H.S. Deg/GED
Some College
Associates
Bachelors+
25%
9%
25%
24%
15%
3%
41%
10%
23%
19%
5%
1%
39%
11%
23%
19%
6%
2%
39%
9%
25%
19%
6%
2%
37%
10%
29%
15%
6%
3%
40%
6%
31%
15%
6%
1%
39%
5%
29%
17%
8%
2%
32%
12%
36%
12%
4%
4%
38%
23%
15%
8%
8%
8%
50%
30%
10%
10%
43%
29%
14%
14%
50%
50%
100%
1-21
22-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70
71-75
76-80

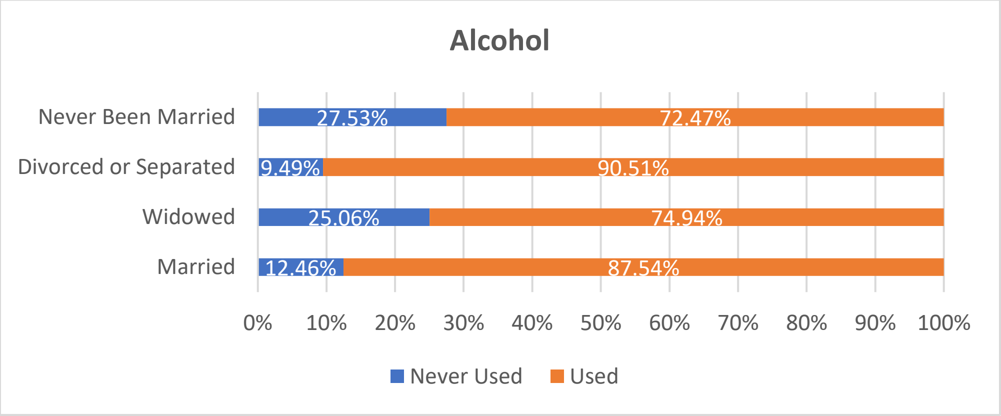

EMPLOYMENT
Full
Part
Unemp
Other
21%
7%
16%
57%
29%
4%
12%
55%
36%
2%
12%
50%
42%
2%
11%
44%
44%
2%
11%
43%
52%
1%
11%
36%
51%
2%
14%
34%
72%
4%
12%
12%
77%
8%
15%
70%
10%
20%
71%
14%
14%
50%
50%
100%
1-21
22-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70
71-75
76-80

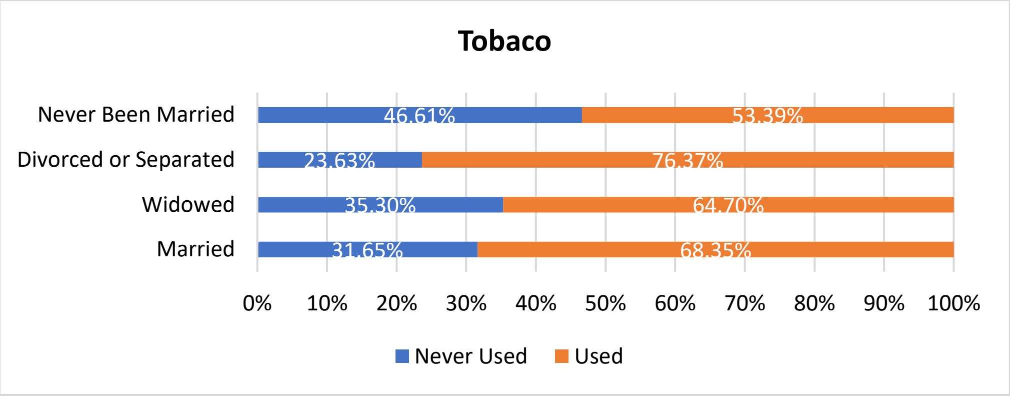
Alcohol:

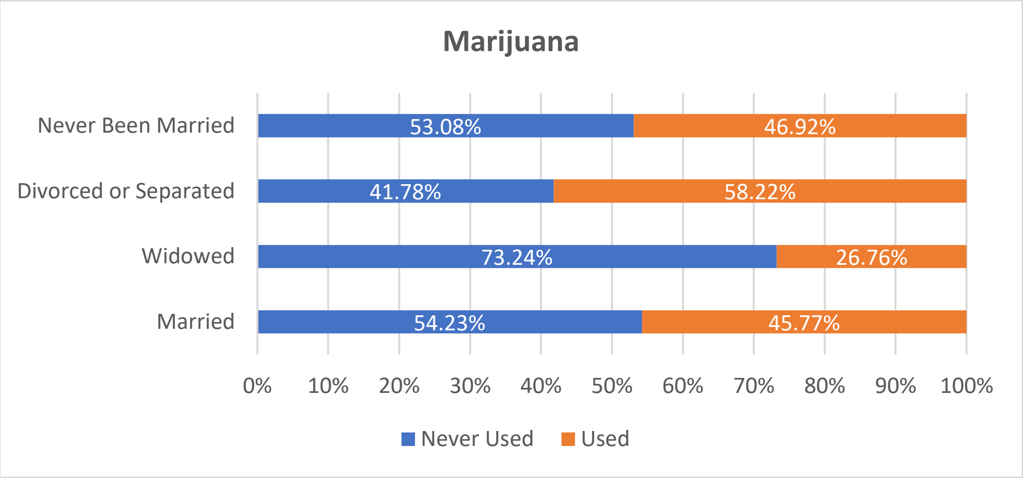

EDUCATION
Junior H
H.S., no Deg
H.S. Deg/GED
Some College
Associates
Bachelors+
26%
9%
23%
23%
15%
4%
30%
8%
19%
26%
11%
6%
27%
7%
17%
26%
13%
9%
25%
7%
17%
28%
12%
11%
25%
7%
18%
28%
13%
10%
26%
8%
16%
26%
15%
9%
18%
12%
10%
29%
20%
10%
25%
10%
10%
30%
25%
14%
14%
14%
29%
29%
22%
11%
33%
33%
33%
67%
33%
67%
1-21
22-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70
71-75

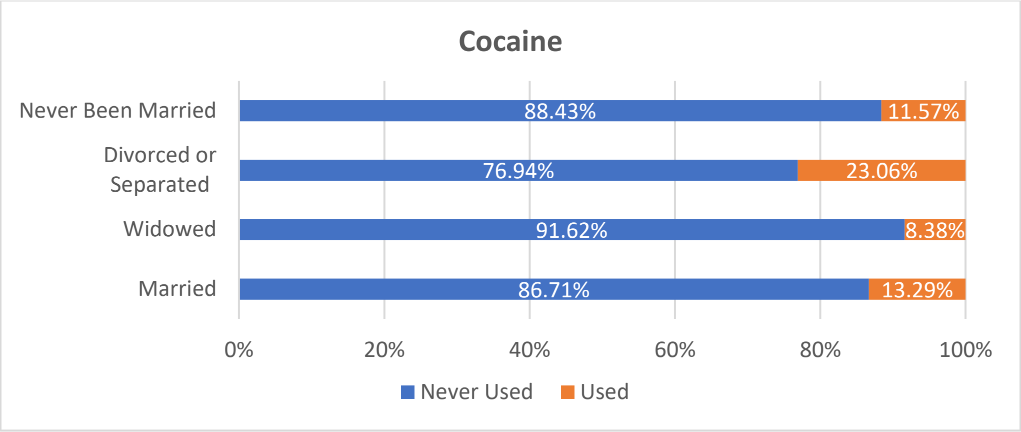

EMPLOYMENT
Full
Part
Unemp
Other
23.73%
5.73%
15.42%
55.11%
33.65%
5.42%
11.09%
49.85%
42.31%
4.84%
10.54%
42.31%
47.56%
7.32%
10.57%
34.55%
54.48%
6.72%
8.96%
29.85%
58.11%
5.41%
8.11%
28.38%
65.31%
6.12%
8.16%
20.41%
65.00%
10.00%
25.00%
64.29%
14.29%
21.43%
66.67%
22.22%
11.11%
100.00%
100.00%
1-21
22-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70
71-75

To further look into the data, we decided to split the respondents by Marital Status. The marital status was separated into 4 different categories in the dataset: married, widowed, divorced or separated and never married. Based on the analysis, it is indicated that people who have experienced divorce or separation are more likely to use alcohol and drugs. Additionally, people who experienced divorce or separation had percentages that were considerably higher than the other 3 categories when it came to the usage of the more harmful substances. For example, when it came to cocaine which is a more harmful drug than alcohol, the people who divorced or separated had the highest rate of drug use, besides nearly 2 times higher chances of using it than any other groups.

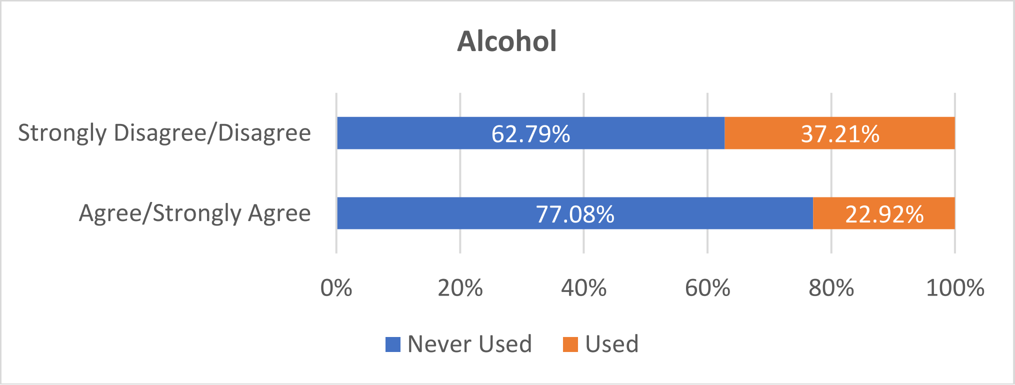


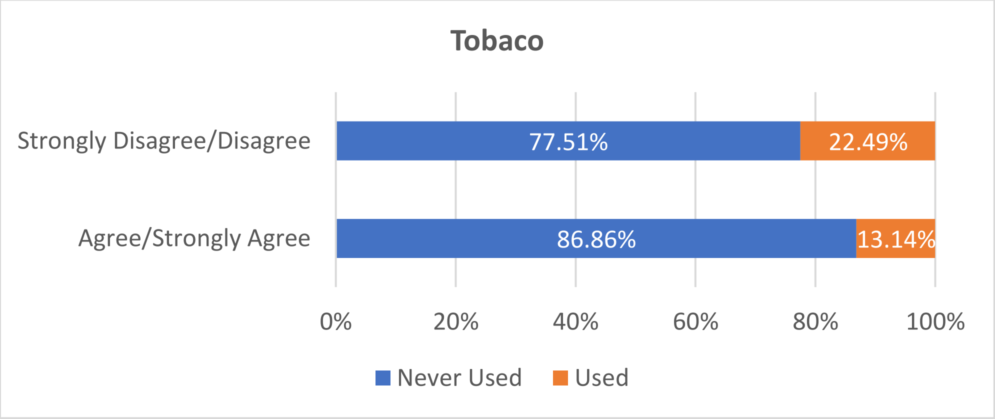


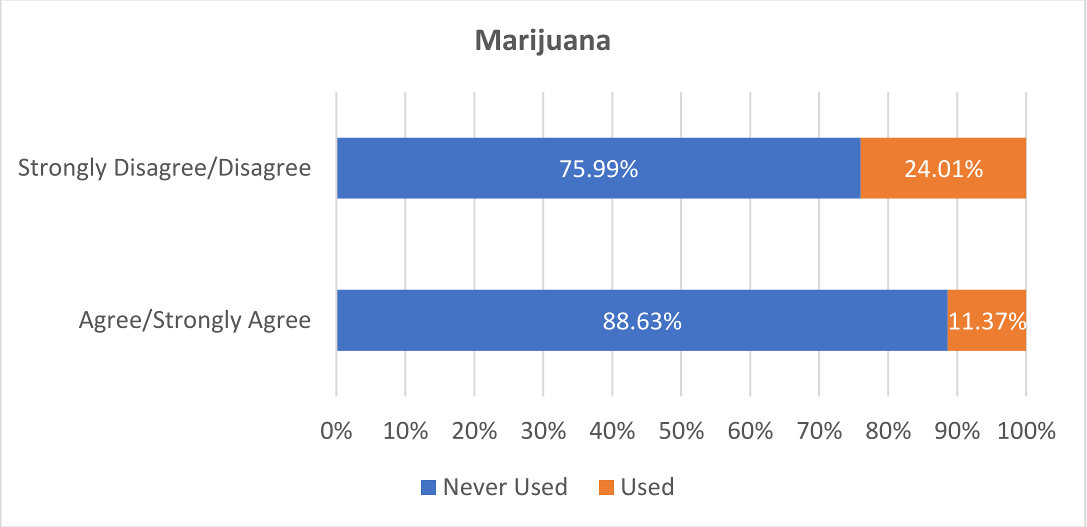


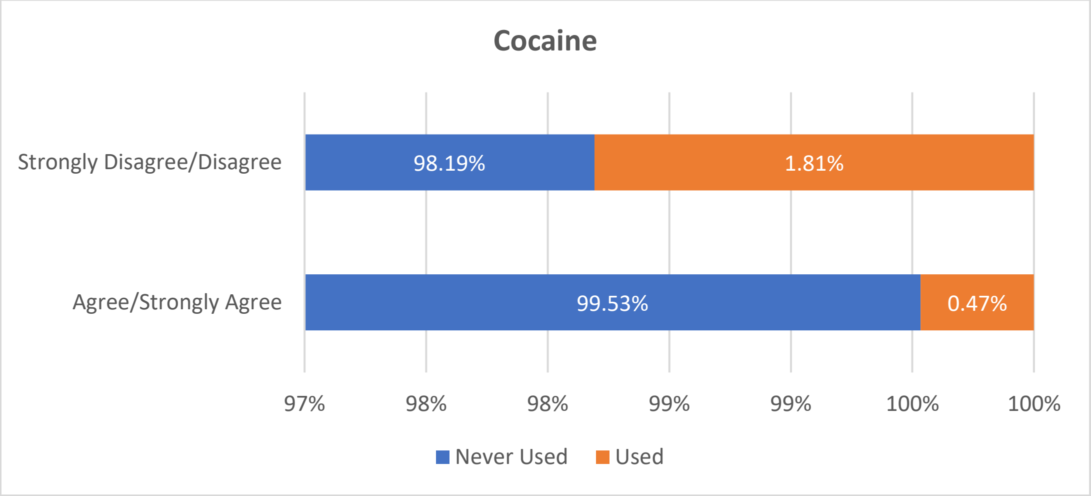


When we break down the data by Religious Belief, it was shown that people who think religious belief is very important in their life are less likely to use alcohol and drugs. Moreover, people that did not agree with religious belief had percentages that were way higher than the people who disagree with religious belief value when it came to the usage of the more harmful drugs such as Cocaine. For instance, people who disagree with the religious belief value, the percentage of using cocaine is 1.81% which was 3 times higher than people who have a religious belief.

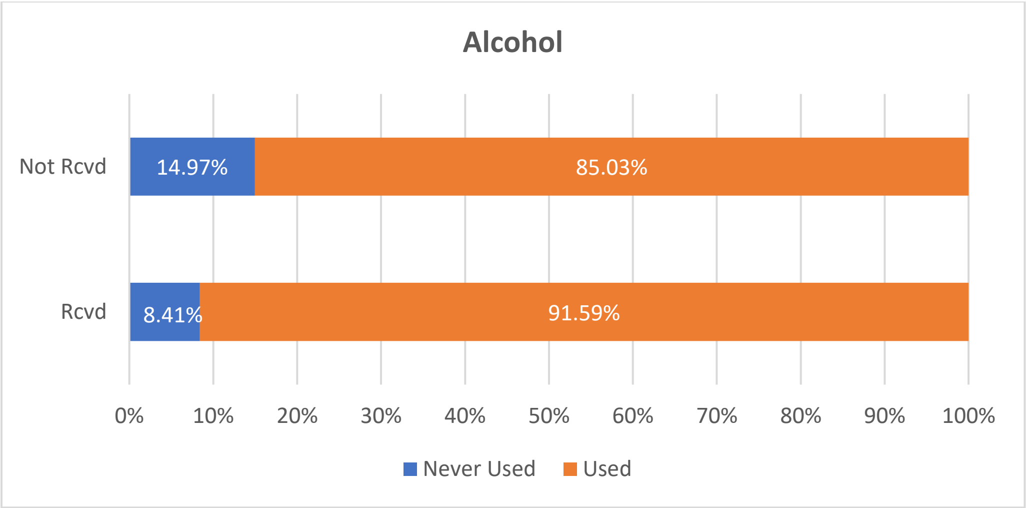


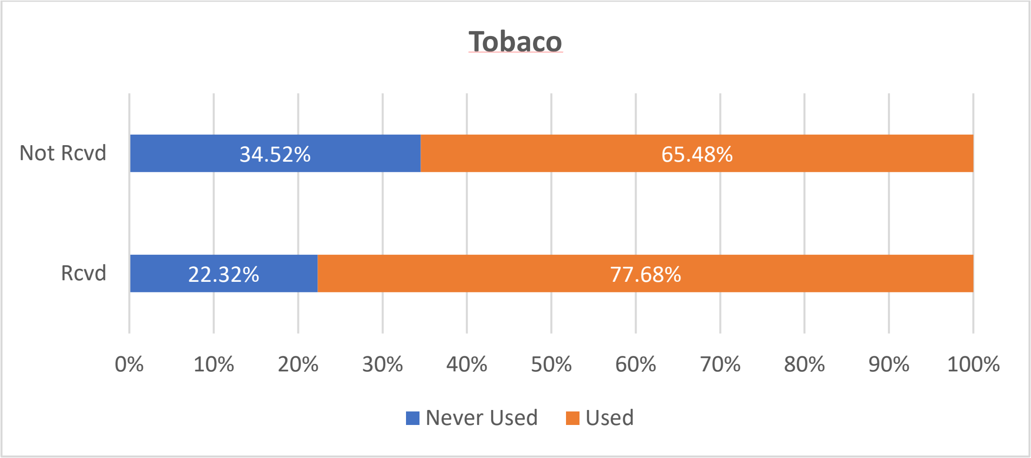


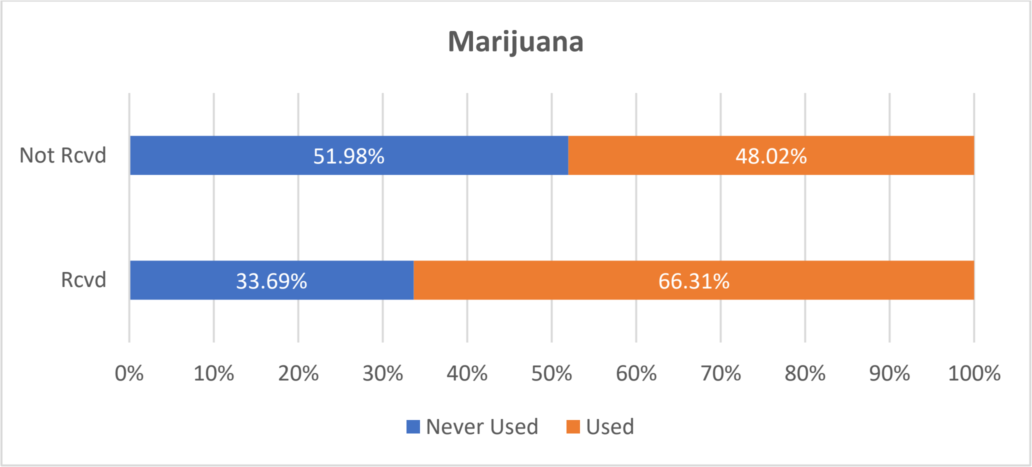


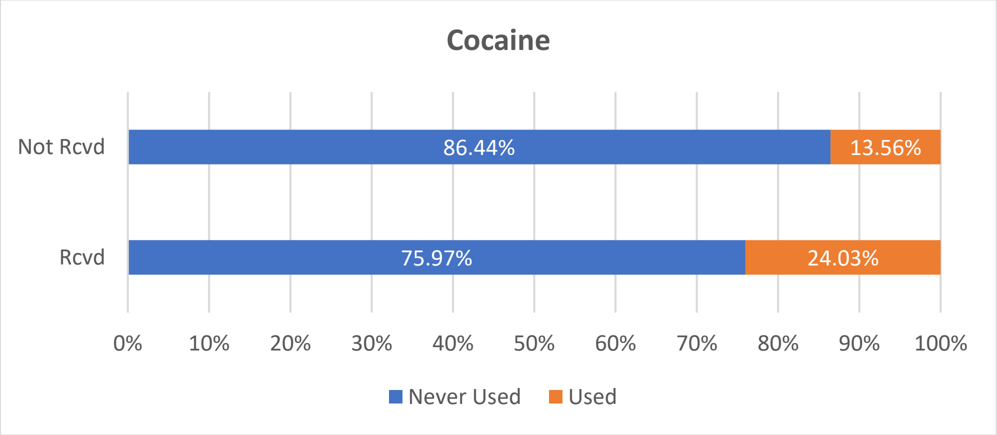


Since it is reported that millions of Americans are affected by mental health conditions every year and approximately 1 in 5 adults in the U.S.—43.8 million, or 18.5%—experiences mental illness in a given year [6] , we wanted to know if mental health treatment was helpful for reducing drug usage. We decided to split the respondents by mental health treatment history. However, from the percentage of using 4 different harmful level of drugs, the respondents who received mental health treatment had the highest chances among all 4 drugs. But actually, it is not because of the mental health treatment itself, it is the people who received the treatment mostly they were having mental health issue already before they took the treatment. So, we can infer that people who had mental health issue had higher chances of overusing drugs.

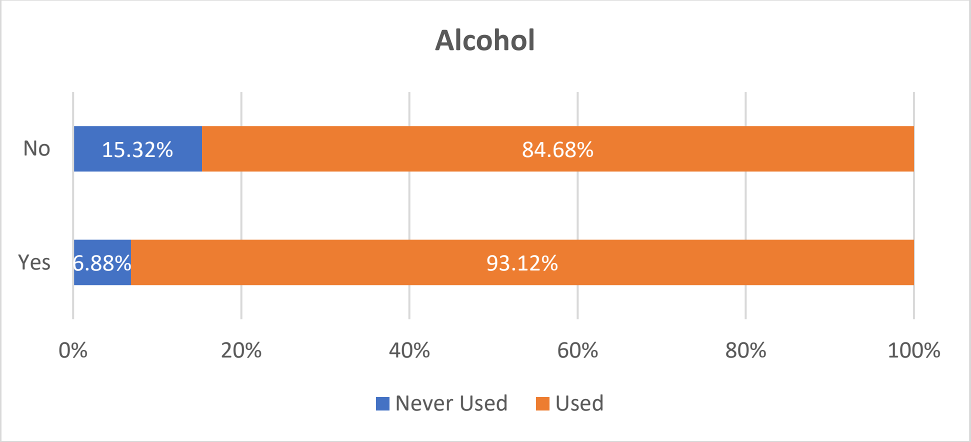

Alcohol
Not Rcvd
14.97%
85.03%
Rcvd
8.41%
91.59%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
■ Never Used
■ Used

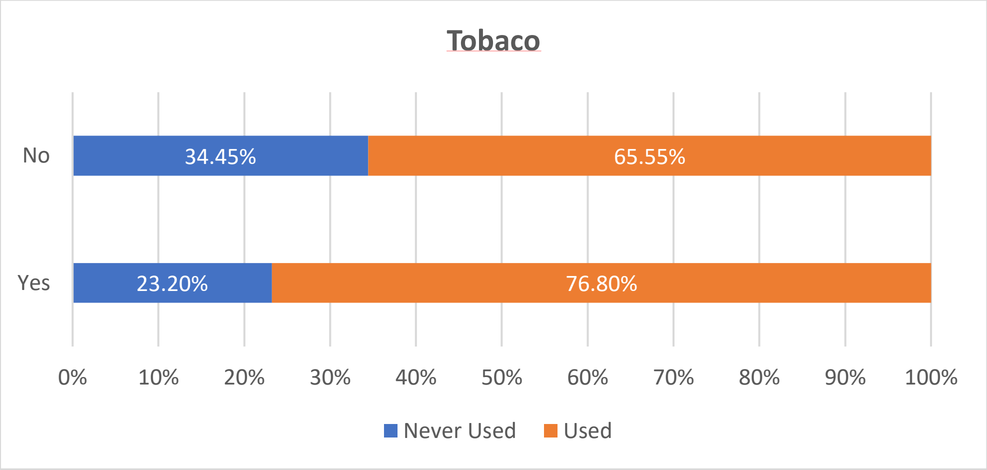

Tobaco
Not Rcvd
34.52%
65.48%
Rcvd
22.32%
77.68%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
■ Never Used
■ Used

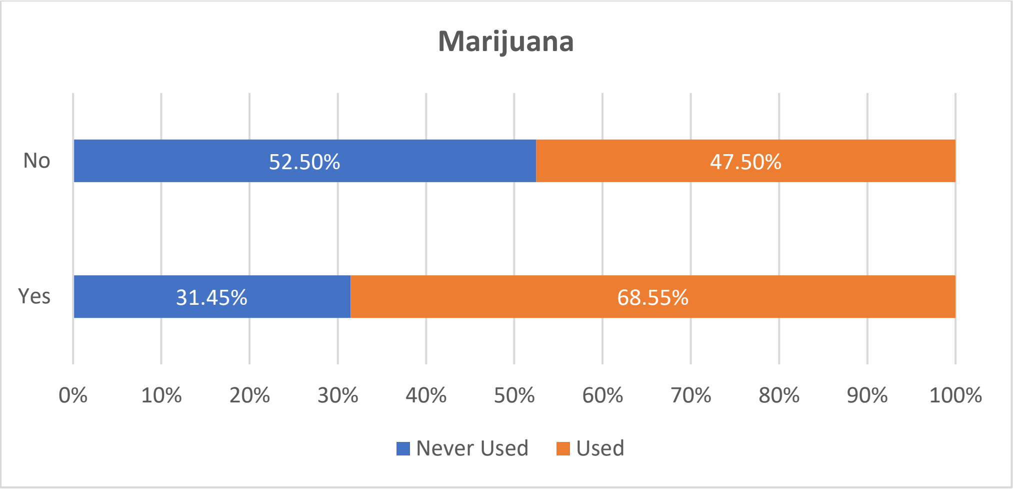

Marijuana
Not Rcvd
51.98%
48.02%
Rcvd
33.69%
66.31%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
■ Never Used
■ Used

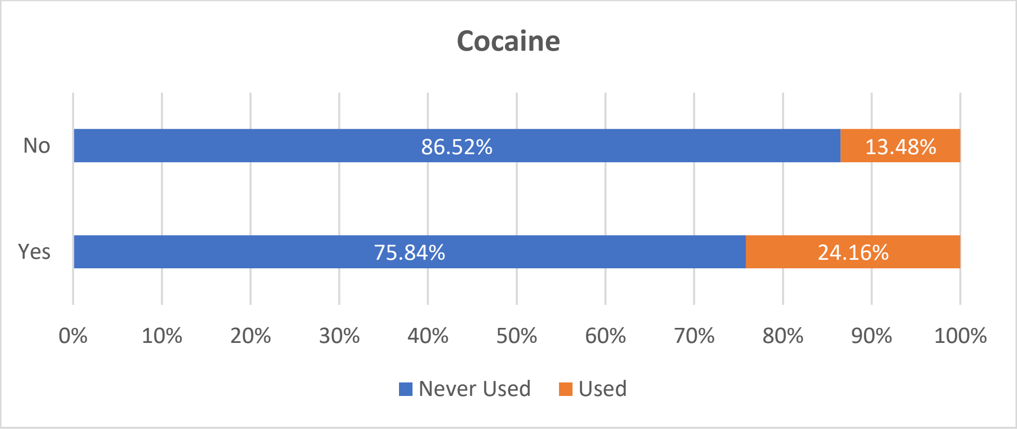


Looking at the distribution of respondents who have used drugs, those who have experienced major depressive episode have higher percentages pertaining to alcohol consumption and drug usage. Especially people that had lifetime major depressive episode had percentages that were way higher than the other when it came to the usage of the more harmful drugs such as Marijuana and Cocaine. It is obviously shown the huge impact of a lifetime major depressive episode on people.





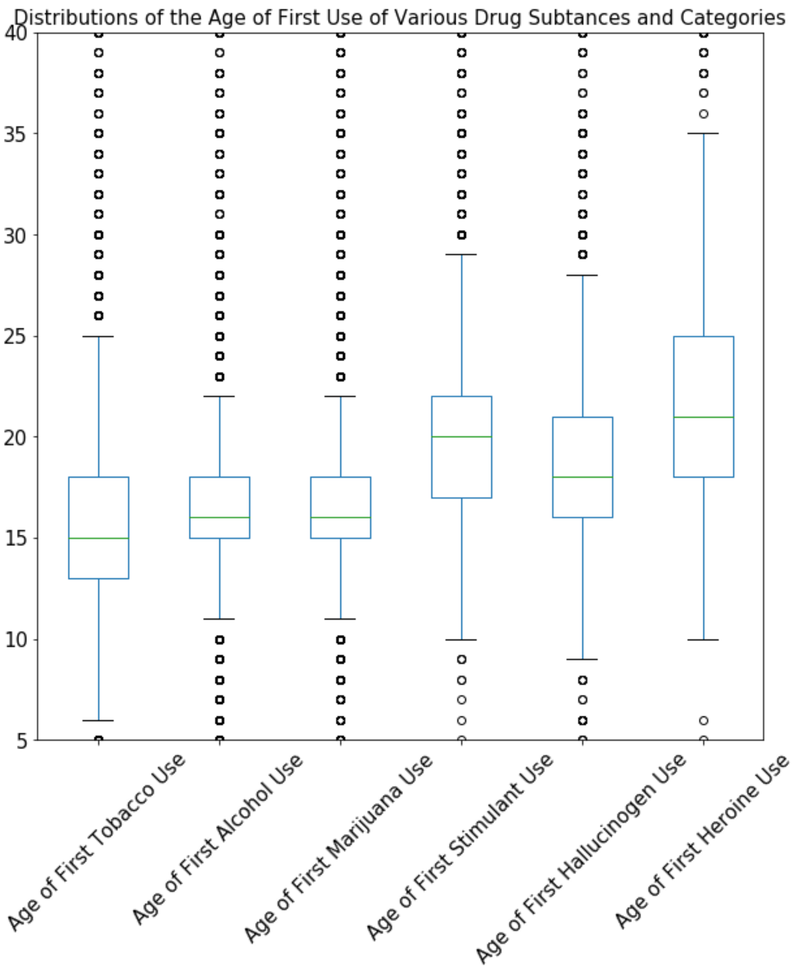




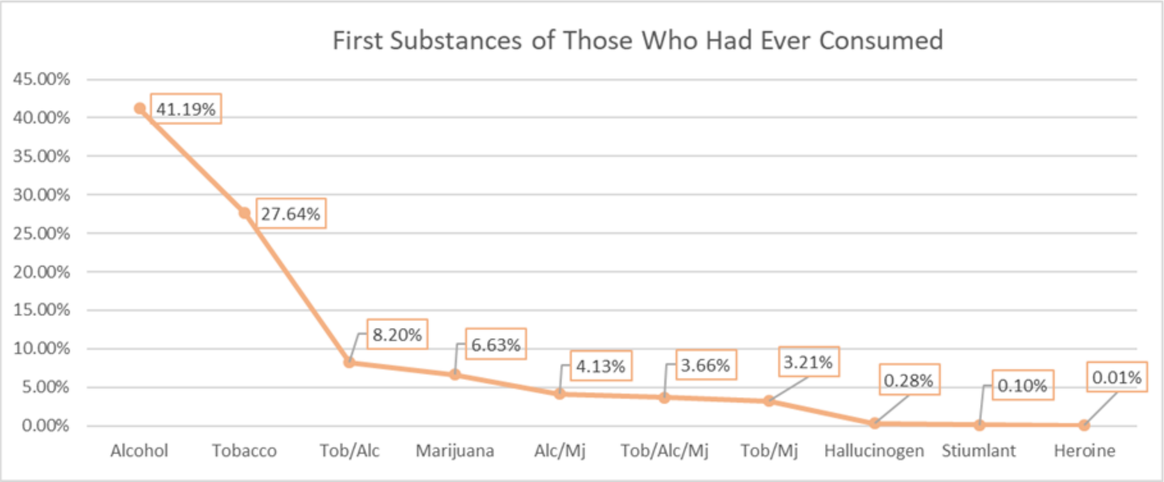
When we break down the usage of drugs by income, certain income brackets use more drugs than others. This is most likely since people in these income brackets have more money to spend on drugs vs the people that make less income. The jump in usage is particularly visible when you get to the category that makes $75,000 or more. However, as mentioned before, since this category contains a lot of income brackets put together and 34% of participants belong to this income bracket, it might be skewing the data. It is also worth noticing that some drugs like Crack and Heroin lose popularity as person's income increases and other drugs like Cigars increase. As was suggested before, it would be better to take this category and split it so that there is an equal representation. That way, if there is a trend that people with higher income do drugs more, it can be seen more clearly without the skewing caused by clumping of people together.

To further drill down into the data, we decided to split the respondents by Employment status and Education level.

While Education level does not provide an overall pattern among the six drug classes, there is a noticeably large portion young first-time user identifying themselves as Employed Full time. Conversely, there is a large population of older individuals who began trying substances later in life and identifying themselves as “other” in employment.



When we group the distribution of the ages of first use by drug category, Tobacco, Alcohol, Marijuana and Hallucinogens have their distributions centered around the teenage years while Stimulants and Heroine have their distributions centered around the early twenties.



Gateway Drug:

- A habit-forming drug that, while not itself addictive, may lead to the use of other addictive drugs.

Since all substances are potentially addictive, we retooled the definition of a gateway drug in a way that can be more precisely measured:

- The first drug consumed by individuals which preceded the consumption of other drugs at an older age at the highest rate.
    - ○ Under this definition, Marijuana would be a gateway drug if the highest number of first users went on to try another substance at an older age

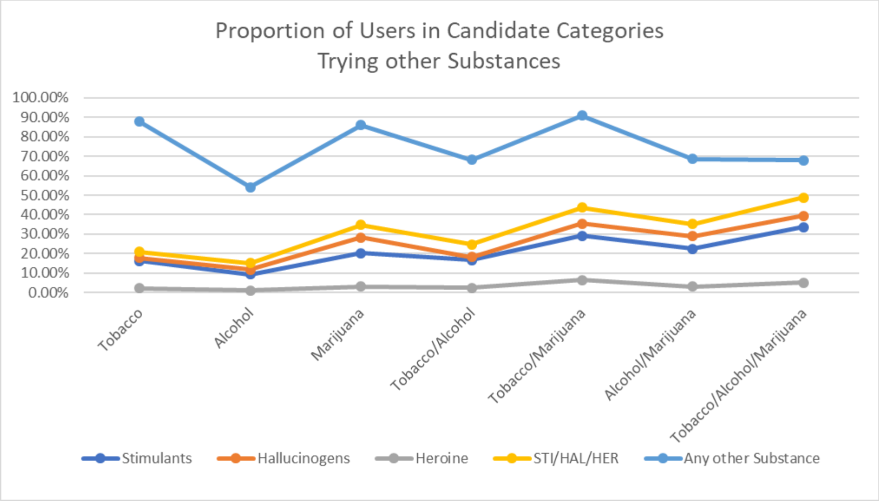
In order to identify these substances, we populated the following table with the proportions generated in the "Data" section. Possible future drugs include candidate drugs and stimulants, hallucinogens, and heroine.

	Tobacco		Alcohol		Marijuana		Stiumulants		Hallucinogen		Heroine		STI/HAL/HER at an Older Age		Any Other Substance at an Older Age	
Tobacco	11,953	38.81%	10,485	**87.72%**	6,446	53.93%	1,943	16.26%	2,144	17.94%	270	2.26%	2,881	21.01%	10,485	87.72%
Alcohol	6,714	37.69%	17,814	44.01%	6,670	37.44%	1,671	9.38%	2,120	11.90%	190	1.07%	2,691	15.11%	**9,630**	**54.06%**
Marijuana	1,780	62.04%	2,470	**86.09%**	2,869	12.06%	580	20.22%	810	28.23%	90	3.14%	**996**	**34.72%**	2,470	86.09%
Tobacco/Alcohol	3,545	12.16%	3,545	12.16%	1,917	54.08%	596	16.81%	649	18.31%	84	2.37%	873	24.63%	2,418	68.21%
Tobacco/Marijuana	1,388	6.71%	1,261	90.85%	1,388	6.71%	406	29.25%	491	35.37%	91	6.56%	607	43.73%	1,261	90.85%
Alcohol/Marijuana	966	54.09%	1,786	7.76%	1,786	7.76%	404	22.62%	517	28.95%	54	3.02%	630	35.27%	1,227	68.70%
Tobacco/Alcohol/Marijuana	1,583	7.82%	1,583	7.82%	1,583	7.82%	533	33.67%	626	39.55%	82	5.18%	773	48.83%	1,076	67.97%

The 7 rows in this table consist of each of the candidate drug categories while the columns consist of the possible future substances consumed at an older age. Each of the cells in the table consist of the number and the proportions of users belonging to the row-wise candidate drug category who went on to consume the column-wise drug at an age>=age of first use.

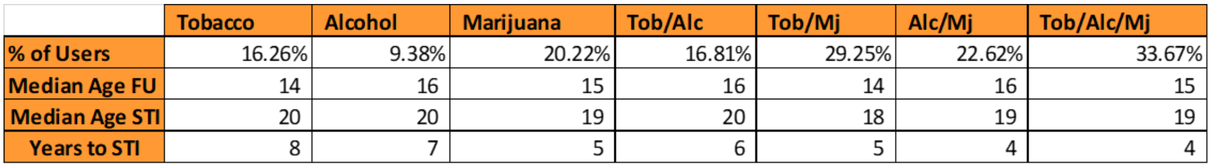
If we only focus on the rows consisting of only one candidate drug, then Tobacco is a gateway drug with 87.72% of respondents going on to try another substance in the future, according to our definition. However, if we only consider the future consumption of non-candidate drugs, then Marijuana is more of a gateway drug than Tobacco or Alcohol with 34.72% going on to try them.
When we turn our attention to the rows of pairs and triplets of candidate categories, then Tobacco/Marijuana is a gateway drug along with Tobacco/Alcohol/Marijuana if we restrict the future drugs as belonging to the STI/HAL/HER categories.

Alcohol is the gateway drug "loser", it was the candidate drug that had the fewest number of respondents go on to report trying another substance in the future.

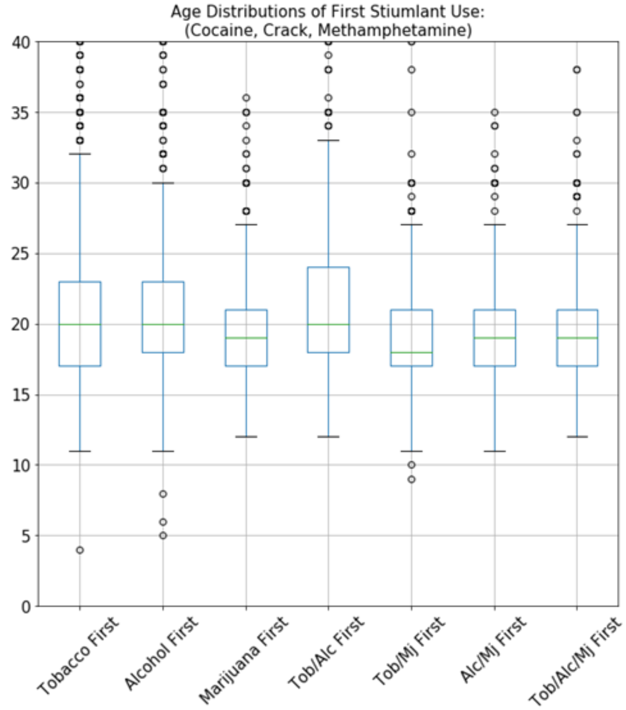


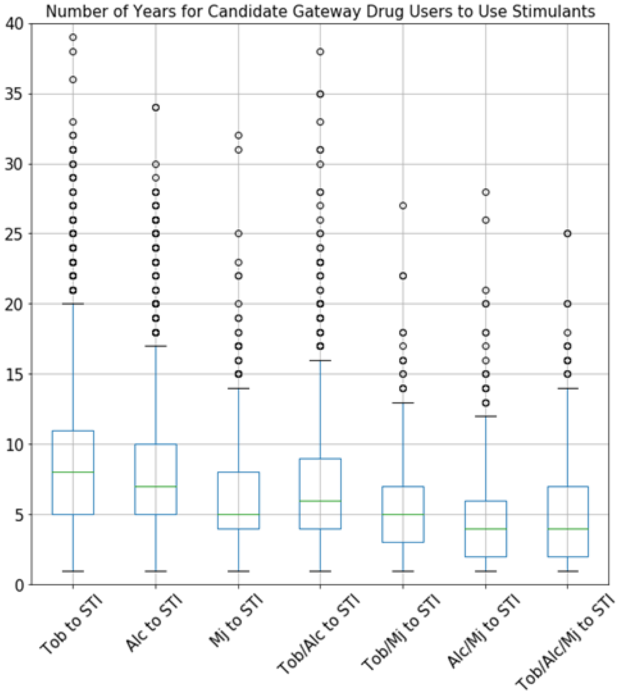
For each of the STI/HAL/HER future drugs, we also measured the distributions of the number of years to first use given their first substance was in one of the candidate drug categories

Respondents who went on to try stimulants:



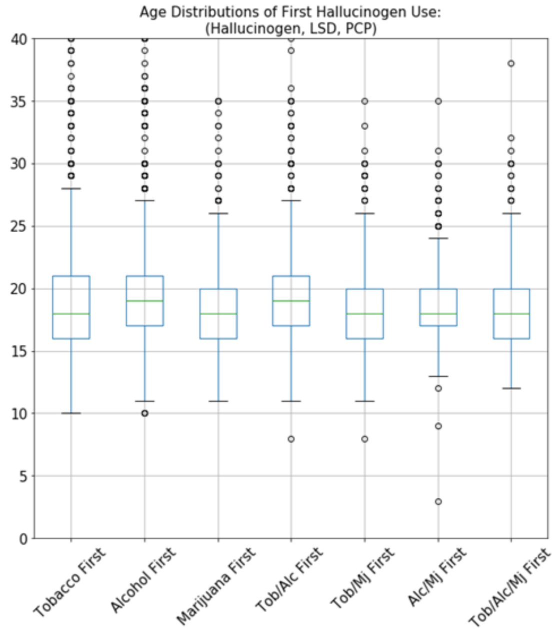
	Tobacco	Alcohol	Marijuana	Tob/Alc	Tob/Mj	Alc/Mj	Tob/Alc/Mj
% of Users	16.26%	9.38%	20.22%	16.81%	29.25%	22.62%	33.67%
Median Age FU	14	16	15	16	14	16	15
Median Age STI	20	20	19	20	18	19	19
Years to STI	8	7	5	6	5	4	4

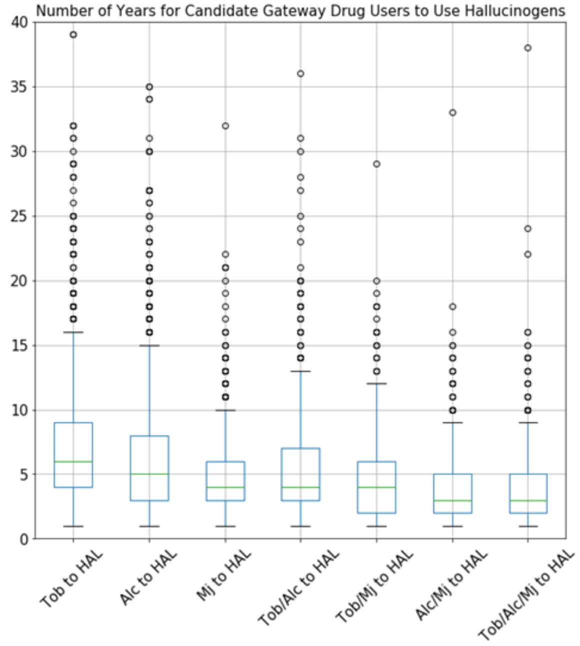




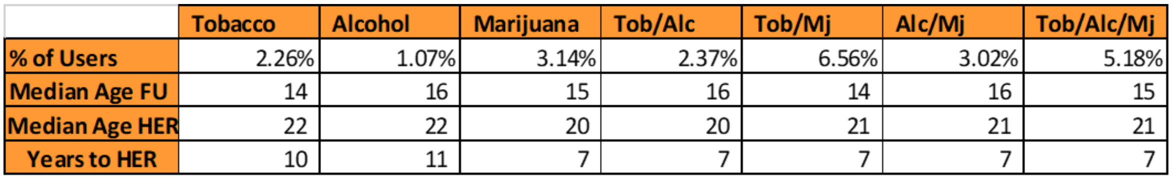
Respondents who went on to try Hallucinogens:

	Tobacco	Alcohol	Marijuana	Tob/Alc	Tob/Mj	Alc/Mj	Tob/Alc/Mj
% of Users	17.94%	11.90%	28.23%	18.31%	35.37%	28.95%	39.55%
Median Age FU	14	16	15	16	14	16	15
Median Age HAL	18	19	18	19	18	18	18
Years to HAL	6	5	4	4	4	3	3

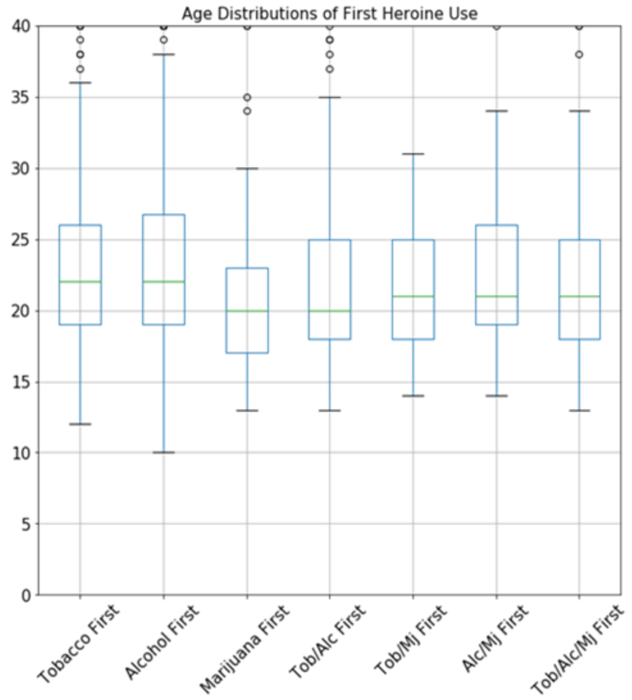


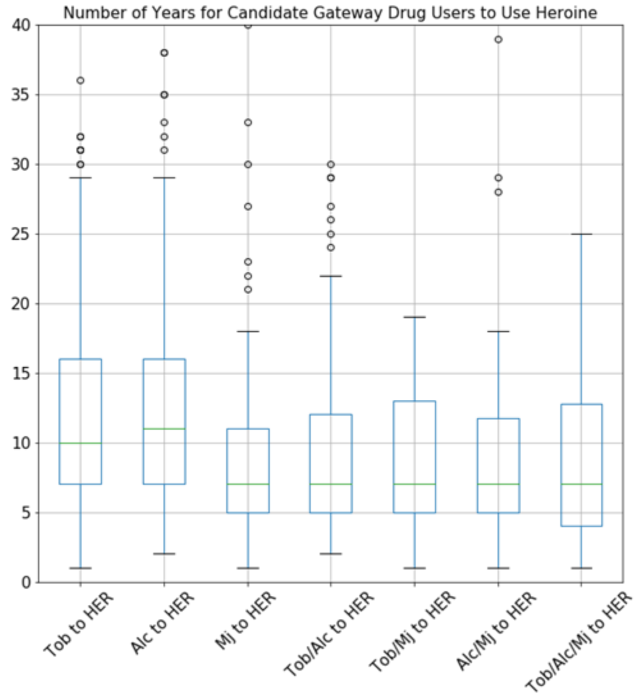


Respondents who went on to try Heroine:



	Tobacco	Alcohol	Marijuana	Tob/Alc	Tob/Mj	Alc/Mj	Tob/Alc/Mj
% of Users	2.26%	1.07%	3.14%	2.37%	6.56%	3.02%	5.18%
Median Age FU	14	16	15	16	14	16	15
Median Age HER	22	22	20	20	21	21	21
Years to HER	10	11	7	7	7	7	7





Individuals who belonged to the candidate categories of two or more substances consistently had lower than or equal to median number of years to first use of Stimulants, Hallucinogens or Heroine. There seems to be an amplification of the proportion of users who reported on trying other substances especially if alcohol is excluded.

Conclusions:

Whenever survey data is analyzed it is always important to note that we cannot make precise but only approximate inferences about results. For example, the true proportion of users who reported ever having tried a substance may be in a large neighborhood of 76%. Many of the questions asked respondents to recall the specific ages of their first substance use or estimate the frequency of events in their lives, however human memory can often be faulty. Without a time-machine to verify the respondents' answers, we can never validate the truth of their reported metrics.

Another issue with the data was that only 16 substances were asked about in the "Age of First Use" questions while 28 substances were asked about in the "Ever being used" questions. This made it impossible to identify other potential gateway drugs or drugs that were consumed after the age of first use.

The most obvious next step in a future analysis would be to replicate the methods of exploration we performed on the 2016 NSDUH survey on every other survey (1979-2015) and perform a time series analysis on the descriptive statistics. Did gateway drugs vary across time? Rates of mental illness? Or the distribution of the ages of first use?

Appendix 1:

[1] https://nsduhweb.rti.org/respweb/about_nsduh.html (NSDUH about)

[2] https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2016-nsduh-2016-ds0001-nid17185 (SAMHDA data repository main page)

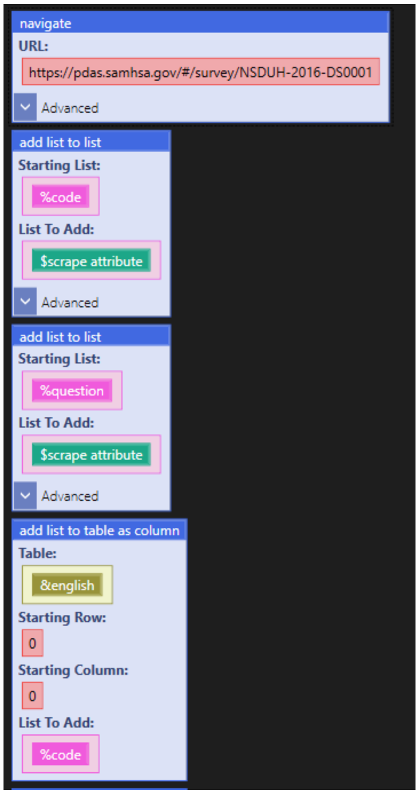
[3] https://nsduhweb.rti.org/respweb/faq.html#q3

[4]https://www.wsj.com/articles/cocaine-meth-opioids-all-fuel-rise-in-drug-overdose-deaths-1537466455?mod=searchresults&page=2&pos=7 (WSJ Article on Overdose Deaths)

[5] https://pdas.samhsa.gov/#/survey/NSDUH-2016-DS0001 (SAMHDA web application)

Appendix 2:

```
navigate("https://pdas.samhsa.gov/#/survey/NSDUH-2016-DS0001?results_received=false&row=CIGOFRSM&weight=ANALWT_C","Wait")
add list to list(%code,$scrape attribute(<class="variable-key">,"innertext"),"Delete","Global")
add list to list(%question,$scrape attribute(<class="variable-name">,"innertext"),"Delete","Global")
add list to table as column(&english,0,0,%code)
add list to table as column(&english,0,1,%question)
save to file("C:\\Users\\Erik Hamlin\\Desktop\\questions.csv",&english)
```

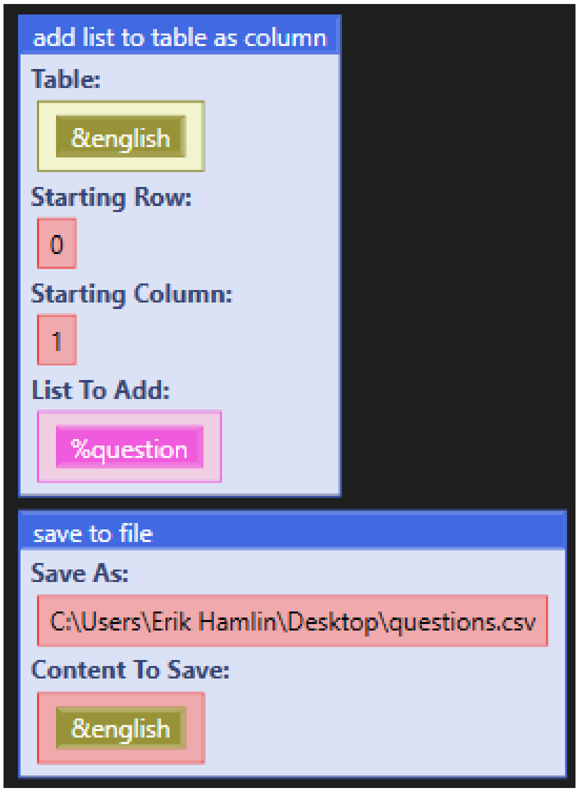


navigate
URL: https://pdas.samhsa.gov/#/survey/NSDUH-2016-DS0001
Advanced

add list to list
Starting List: %code
List To Add: $scrape attribute
Advanced

add list to list
Starting List: %question
List To Add: $scrape attribute
Advanced

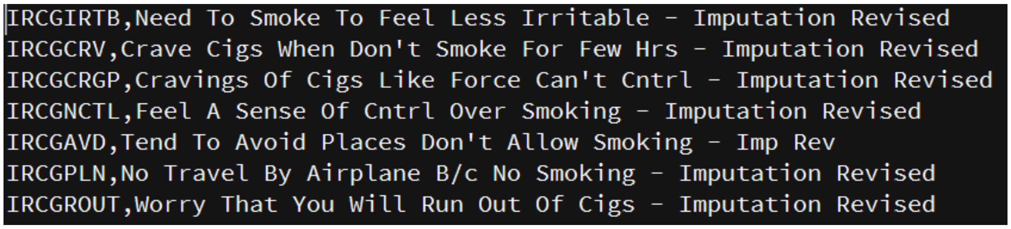
add list to table as column
Table: &english
Starting Row: 0
Starting Column: 0
List To Add: %code



add list to table as column
Table: &english
Starting Row: 0
Starting Column: 1
List To Add: %question

save to file
Save As: C:\Users\Erik Hamlin\Desktop\questions.csv
Content To Save: &english

Results:



```
IRCGIRTB,Need To Smoke To Feel Less Irritable - Imputation Revised
IRCGCRV,Crave Cigs When Don't Smoke For Few Hrs - Imputation Revised
IRCGCRGP,Cravings Of Cigs Like Force Can't Cntrl - Imputation Revised
IRCGNCTL,Feel A Sense Of Cntrl Over Smoking - Imputation Revised
IRCGAVD,Tend To Avoid Places Don't Allow Smoking - Imp Rev
IRCGPLN,No Travel By Airplane B/c No Smoking - Imputation Revised
IRCGROUT,Worry That You Will Run Out Of Cigs - Imputation Revised
```