Checkpoint2

December 18, 2021

1 Overview

Our project's goal was to determine whether socioeconomic status or ethnicity is better predicted by the age at which an individual started using drugs in the United States. In order to do that, we used a decision tree classifier from Sci Kit learn. We found out that eventhough neither model was great at predicting what it was meant to, race had a better accuracy in predicting age (59%) than socioeconomic factor (35%).

2 Names

- Mehdi Boussami
- Jacob Ryan

3 Research Question

Is socioeconomic status or ethnicity better predicted by the age at which an individual started using drugs in the United States?

4 Background & Prior Work

We are attempting to research whether socioeconomic status or ethnicity is better predicted from the age at which an individual started using drugs in the United States. In order to address this research question, we have gathered data looking at a person's age of first use for a variety of drugs, along with that person's income and racial identity. (One thing to note is that in our examination of this research question, we did not include alcohol as a drug. Since alcohol consumption is socially accepted, as well as used in many religious settings, this could skew the age of drug use downwards (McCabbe, 2017)). We then hope to analyze the correlation between this age of first drug use and socioeconomic status, and age of first drug use and ethnicity to see which prediction is better. Some research has been done previously regarding the links between age of first drug use and socioeconomic status. Previous studies have found that the age that a person starts using drugs like Tobacco or hard drugs such as Marijuana can be a predictor of socioeconomic status. In a 2012 study examining substance use among young adults, they found that young adult use of Tobacco was associated with a lower socioeconomic status, whereas young adult use of drugs such as marijuana was associated with a higher socioeconomic status (Patrick, 2012). They point to accessibility as a reason for these patterns, for instance, those with a higher socioeconomic status are more likely to go to college where Marijuana use is encouraged and easily accessible, along with being more likely to afford the drug. While this and other studies examine the relationships between drug use and socioeconomic status, there is little to no research looking at the direct connection of age of drug use as a predictor of race. Yet, research has found that socioeconomic status is correlated with race due to social inequality (Williams, 2016). Therefore many of the attributes that are frequently associated with people with a lower socioeconomic status are also frequently associated with people within a ethnical minority as a result of racial and ethnic income inequality (American Psychology Association, 2017). Since previous research has pointed to age of drug use as a predictor of socioeconomic status, we wonder if this is actually a better predictor of one's racial identity as there is correlation between the two (socioeconomic status and race).

References: - 1 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3410945/ Patrick, M. E., Wightman, P., Schoeni, R. F., & Schulenberg, J. E. (2012). Socioeconomic status and substance use among young adults: a comparison across constructs and drugs. Journal of studies on alcohol and drugs, 73(5), 772–782. https://doi.org/10.15288/jsad.2012.73.772 - 2) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786055/ Karriker-Jaffe K. J. (2013). Neighborhood socioeconomic status and substance use by U.S. adults. Drug and alcohol dependence, 133(1), 212–221. https://doi.org/10.1016/j.drugalcdep.2013.04.033 - 3) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2377408/ McCabe, S. E., Morales, M., Cranford, J. A., Delva, J., McPherson, M. D., & Boyd, C. J. (2007). Race/ethnicity and gender differences in drug use and abuse among college students. Journal of ethnicity in substance abuse, 6(2), 75–95. https://doi.org/10.1300/J233v06n02_06 - 4) https://www.apa.org/pi/ses/resources/publications/minorities American Psychological Association. (2017). Ethnic and racial minorities & socioeconomic status. American Psychological Association. Retrieved December 10, 2021, from https://www.apa.org/pi/ses/resources/publications/minorities#. - 5) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817358/ Williams, D. R., Priest, N., & Anderson, N. B. (2016). Understanding associations among race, socioeconomic status, and health: Patterns and prospects. Health psychology : official journal of the Division of Health Psychology, American Psychological Association, 35(4), 407–411. https://doi.org/10.1037/hea0000242

5 Hypothesis

We will divide the dataset in 2 for our experiment EDA. One consisting of columns that represent hard drugs and one with columns containing light drugs. More explanation about this is given later in this report.

We predict that the age at which an individual started using hard drugs will predict the socioeconomic status of that person very well. Indeed, upper middle class teens and young adults are more likely to have access to hard drugs such as cocaine as they are expensive and very prevalant in universities as suggested by turnbridge magazine[1]. On the other hand, people on the lower hand of the socioeconomic ladder are more likely to grow up in an environment where substance use is prevalent and thus are more likely to start using cheap light drugs earlier than people of higher socieconomice status.

Hence, we expect that people from higher socioeconmic status start using hard drugs earlier than people from lower socioeconomic status and that people from lower socioeconomic status start using light drugs earlier than people from higher socioeconomic status.

We predict that the age at which an individual started using hard drugs will not predict the race of that person very well. Indeed, according to sunrisehouse, when we break down the population into races and see what proportion takes drugs, we see that the proportions are very close to each other [2]. We also think that the age at which a person starts taking drugs is heavily associated with levels of education and we cannot say that a certain race is more educated than another at all.

Hence, our null hypothesis is H0: Age of first drug use will not be a better predictor of income than race

and our alternative hypothesis is H1: Age of first drug use will be a better predictor of income than race.

References: - 1) https://www.turnbridge.com/news-events/latest-articles/socioeconomic-status-and-drug-use/#

- 2) https://sunrisehouse.com/addiction-demographics/different-races/

6 Dataset(s)

- Dataset Name: National Survey on Drug Use and Health (NSDUH16)
- Link to the dataset: https://www.datafiles.samhsa.gov/sites/default/files/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
- Number of observations: 56897

We are trying to see if the age at which a person started taking drugs can predict the socioeconomic status and/or race of that person. The dataset from NSDUH above tells us about when different people started taking different types of drugs at different ages, along with individuals' respective incomes and race.

7 Setup

[1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.metrics import precision_score, recall_score, accuracy_score
from sklearn.tree import DecisionTreeClassifier
```

8 Data Cleaning

Steps to data cleaning: - 1) We select only the features we needed for our analysis. - 2) We rename the columns so they reflected the data more accurately. - 3) We replace values the original authors included as coding for different types of non represented values all to null values as they will not be helpful in our analysis. - 4) We look more closely at the dataset to see if there are any inconsistent data

1)

```
[ ]: #Here we read the data
     df = pd.read_csv('data.tsv', sep='\t')
     #Here we keep the columns that we are interested in for efficiency purposes
     col2 =␣
      ↪['CIGTRY','SMKLSSTRY','CIGARTRY','MJAGE','COCAGE','CRKAGE','HERAGE','HALLUCAGE','LSDAGE','P
     df = df[col2]
```

```
/opt/conda/lib/python3.9/site-packages/IPython/core/interactiveshell.py:3441:
DtypeWarning: Columns (2506) have mixed types.Specify dtype option on import or
set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

2)

```
[5]: #We rename the columns so that they make more sense for our research topic
     df = df.rename(columns={'CIGTRY': 'CIG_AGE', 'SMKLSSTRY': 'SMKLESS_AGE',␣
      ↪'CIGARTRY':'CIGAR_AGE','MJAGE': 'MJ_HASH_AGE', 'COCAGE': 'COKE_AGE',␣
      ↪'NEWRACE2':'RACE','CRKAGE': 'CRACK_AGE','HERAGE': 'HER_AGE', 'HALLUCAGE':␣
      ↪'HALLUC_AGE', 'LSDAGE': 'LSD_AGE','PCPAGE': 'PCP_AGE','ECSTMOAGE':␣
      ↪'ECST_AGE', 'POVERTY': 'POVERTY_LVL'})
```

3)

```
[6]: #Here, we replace the following values with nan. The following values are␣
      ↪shortcuts for people who never took
     #drugs or didn't fill in the survey. Also, we are interested in ages, so values␣
      ↪these values are outliers that
     #will make observations biased.
     df = df.replace([91, 991, 9991, 93, 993, 9993, 94, 994, 9994, 97, 997, 9997,␣
      ↪98, 998, 9998, 99,999, 9999, 985], np.nan)
```

By looking at the dataset more closely, we see that the the type of the race table is Integers which is not really interesting. By looking at the website where we got our dataset, we understand that 1 represents white, 2 black, 3 Native American, 4 Pacific Islander, 5 Asian, 6 Multirace and 7 Hispanic. So, we replace the integers with the corresponding race as indicated in the dataset website for our EDA.

```
[156]: df['RACE'] = df['RACE'].replace({1:'White', 2: 'Black', 3: 'Native American', 4:
        ↪ 'Pacific Islander', 5:
                                        'Asian', 6: 'Multirace', 7:'Hispanic'})
```

```
[157]: df = df.replace(985, np.nan)
```

```
[158]: df.dtypes
```

```
[158]: CIG_AGE        float64
       SMKLESS_AGE    float64
       CIGAR_AGE      float64
```

```
MJ_HASH_AGE    float64
COKE_AGE       float64
CRACK_AGE      float64
HER_AGE        float64
HALLUC_AGE     float64
LSD_AGE        float64
PCP_AGE        float64
ECST_AGE       float64
RACE            object
INCOME           int64
POVERTY3       float64
dtype: object
```

The dataset is now cleaner. Indeed, the types in each column is correct as you can see above.

Now, let's look at the age column as we will be using it a lot in this report. Let's start by seeing if we can make any observations about it

[159]:

```
print(np.max(df))
print("__________")
print(np.min(df))
```

```
CIG_AGE         61.0
SMKLESS_AGE     78.0
CIGAR_AGE       77.0
MJ_HASH_AGE     78.0
COKE_AGE        60.0
CRACK_AGE       60.0
HER_AGE         59.0
HALLUC_AGE      69.0
LSD_AGE         50.0
PCP_AGE         51.0
ECST_AGE        55.0
RACE           White
INCOME             4
POVERTY3         3.0
dtype: object
__________
CIG_AGE          1.0
SMKLESS_AGE      1.0
CIGAR_AGE        1.0
MJ_HASH_AGE      1.0
COKE_AGE         1.0
CRACK_AGE        2.0
HER_AGE          5.0
HALLUC_AGE       1.0
LSD_AGE          1.0
PCP_AGE          1.0
```

```
ECST_AGE          4.0
RACE            Asian
INCOME              1
POVERTY3          1.0
dtype: object
```

Above, we see that the max ages at which people took different drugs makes sense. However, for the minimum, we get unexpected values. Indeed, it is hard to believe that a 1 year old would take any drugs. So we will replace any age smaller than 8 to nan as we believe that the youngest age at which a person would take a drug is around 9 years old. So we choose 9 as the minimum age.

Note: Here, we make the assumtption that the youngest age at which a person would take a drug is 9 years old which might be wrong. But we still use it as we think that it will not affect our analysis that much.

[160]:

```
#Here, we replace all ages smaller than 8 by nan
df[["CIG_AGE", "SMKLESS_AGE", "CIGAR_AGE", "MJ_HASH_AGE", "COKE_AGE",␣
 ↪"CRACK_AGE", "HER_AGE", "HALLUC_AGE", "LSD_AGE",
"PCP_AGE", "ECST_AGE"]] = df[["CIG_AGE", "SMKLESS_AGE", "CIGAR_AGE",␣
 ↪"MJ_HASH_AGE", "COKE_AGE", "CRACK_AGE", "HER_AGE", "HALLUC_AGE", "LSD_AGE",
"PCP_AGE", "ECST_AGE"]].replace([1, 2, 3, 4, 5, 6, 7, 8], np.nan)
print(np.min(df))
```

```
CIG_AGE           9.0
SMKLESS_AGE       9.0
CIGAR_AGE         9.0
MJ_HASH_AGE       9.0
COKE_AGE          9.0
CRACK_AGE        11.0
HER_AGE          10.0
HALLUC_AGE        9.0
LSD_AGE           9.0
PCP_AGE           9.0
ECST_AGE          9.0
RACE            Asian
INCOME              1
POVERTY3          1.0
dtype: object
```

The ages in the dataset make more sense now, so we are ready to start the EDA

9 Data Analysis & Results (EDA)

In this project, we want to know how we can predict the income and the race of an individual based on the age he first started taking drugs.

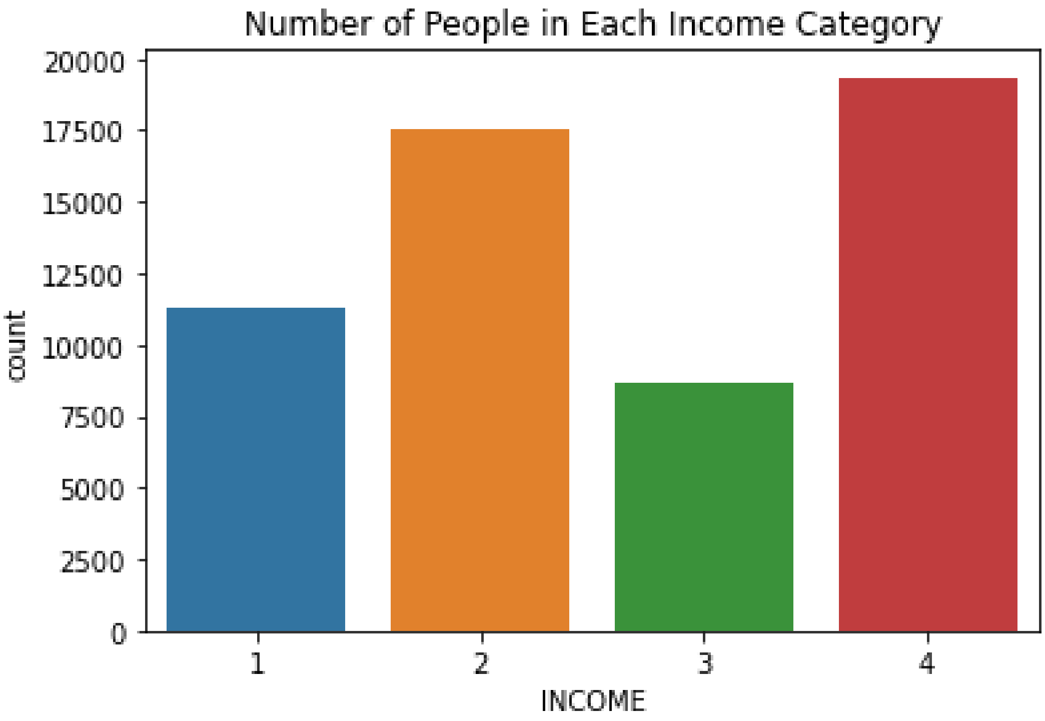
Hence, we will analyze the income, race and age columns in depth

9.0.1 Variables' Distributions and Oultiers

Income Let's look at the number of people in each income category. Note that 1 is the lowest income category and 4 is the highest. (Income 1 is less than 20k. Income 2 is 20k to less than 50k. Income 3 is 50k to less than 75k. Income 4 is 75k and up.)

```
[182]: sns.countplot(x = df['INCOME']).set_title("Number of People in Each Income␣
        ↪Category")
```

```
[182]: Text(0.5, 1.0, 'Number of People in Each Income Category')
```



We see that the majority of people are in the 4th category (19,000) which represents people who have incomes of 75k and up. There are also a lot of people in category 2 (17,500). Categories 1 and 3 contain the least amount of people as can be seen in the graph above.

Let's now see if there are any outliers in the income column:

```
[162]: df["INCOME"].value_counts()
```

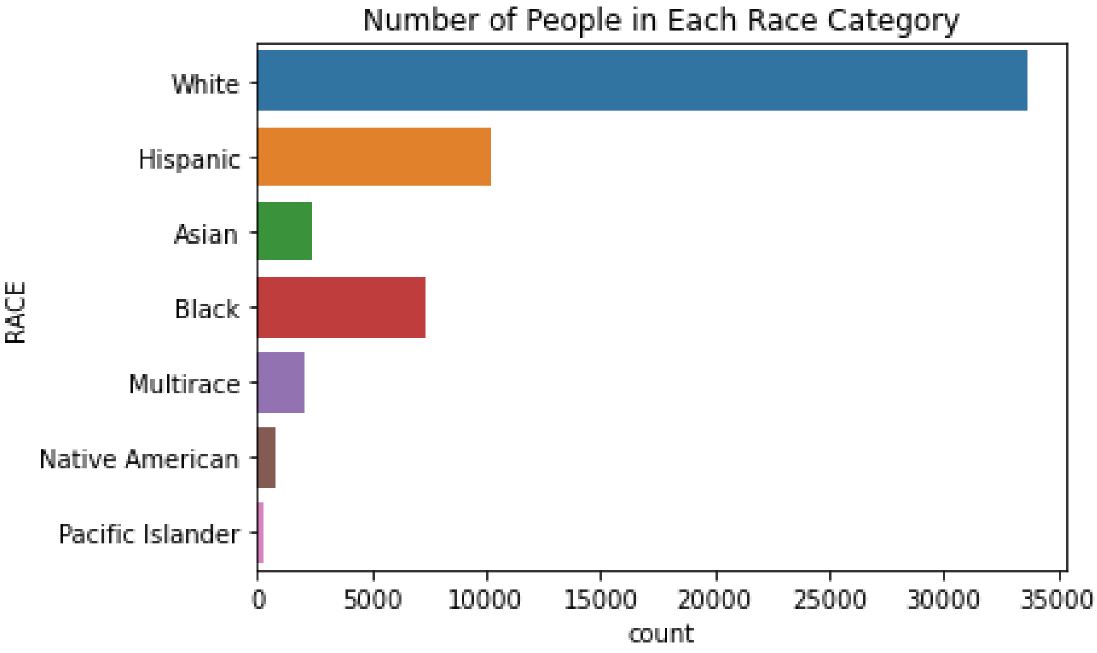
```
[162]: 4    19345
       2    17577
       1    11268
       3     8707
       Name: INCOME, dtype: int64
```

As we can see above, the column doesn't contain any outlier.

Race Let's now look at the distribution of the Race column and see how many people belong to each race category

```
[163]: sns.countplot(y = df['RACE']).set_title("Number of People in Each Race␣
        ↪Category")
```

```
[163]: Text(0.5, 1.0, 'Number of People in Each Race Category')
```



```
[164]: White_percent = (len(df[df["RACE"] == "White"]["RACE"])/len(df["RACE"]))*100
       White_percent
       #Here, we understand that 60% of the people in our dataset are white
```

```
[164]: 59.192927570873685
```

From the graph above, we see that the large majority of people in the dataset are white (34'000). The next race categories with the largest number of people are then Hispanic, Black, Asian, Multirace, Native American and Pacific Islander. It is important to note that we barely have any Pacific Islander and Native American people in the dataset as can be seen above. From the graph above, we understand that because people in our dataset are mainly white (60%), our predictior might be biased and not be that effective.

Let's now check if there are any outliers in the race column

```
[165]: df["RACE"].value_counts()
```

[165]:

```
White               33679
Hispanic            10187
Black                7318
Asian                2437
Multirace            2130
Native American       856
Pacific Islander      290
Name: RACE, dtype: int64
```

As we can see, all the values in the race column are what we expected so there are no outliers

Age Before looking at the age column's distribution, we will add two columns to the table. One called Age_min_hard and one called Age_min_light representing the minimum age at which a person took hard drugs and the minimum age at which a person took light drugs respectively.

Note: we will assume that light drugs consist of cigarettes, smokeless cigarettes and cigar and that hard drugs consist of marijuana, coke, crack, heroine, hallucinogens, lsd, pcp and ecstasy

[166]:

```python
def hard(x):
    return min(x[['MJ_HASH_AGE', 'COKE_AGE',
    'CRACK_AGE', 'HER_AGE', 'HALLUC_AGE', 'LSD_AGE', 'PCP_AGE', 'ECST_AGE']])
def light(x):
    return min(x[['CIG_AGE', 'SMKLESS_AGE', 'CIGAR_AGE']])
df['Age_min_hard'] = df.apply(hard, axis=1)
df['Age_min_light'] = df.apply(light, axis=1)
df
```

[166]:

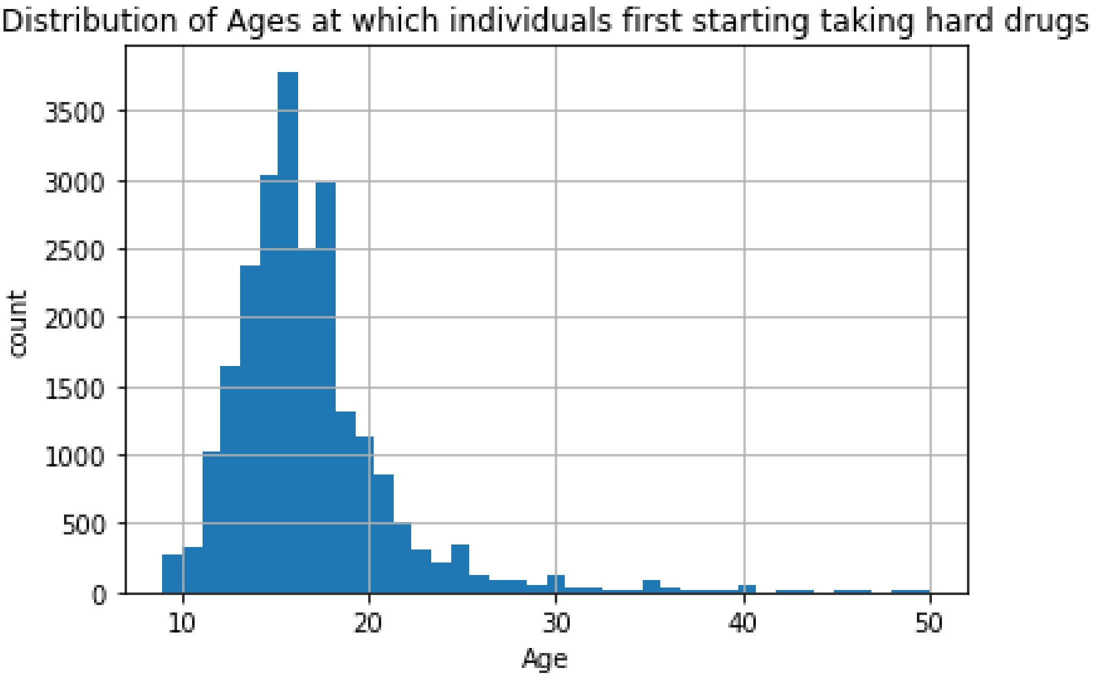
```
       CIG_AGE  SMKLESS_AGE  CIGAR_AGE  MJ_HASH_AGE  COKE_AGE  CRACK_AGE  \
0         16.0         20.0       25.0         20.0       NaN        NaN
1         15.0         20.0       16.0         15.0      20.0       22.0
2         26.0          NaN       26.0          NaN       NaN        NaN
3          NaN          NaN        NaN          NaN       NaN        NaN
4          NaN          NaN        NaN         32.0      34.0        NaN
…        …          …        …          …       …        …
56892      NaN          NaN        NaN          NaN       NaN        NaN
56893     15.0          NaN       14.0         14.0       NaN        NaN
56894     22.0          NaN       30.0         21.0       NaN        NaN
56895      NaN          NaN        NaN          NaN       NaN        NaN
56896      NaN          NaN        NaN          NaN       NaN        NaN

       HER_AGE  HALLUC_AGE  LSD_AGE  PCP_AGE  ECST_AGE        RACE  INCOME  \
0          NaN         NaN      NaN      NaN       NaN       White       4
1         15.0        13.0     21.0      NaN      15.0    Hispanic       4
2          NaN         NaN      NaN      NaN       NaN    Hispanic       2
3          NaN         NaN      NaN      NaN       NaN    Hispanic       4
4          NaN         NaN      NaN      NaN       NaN       White       3
…        …         …      …      …       …       …     …
56892      NaN         NaN      NaN      NaN       NaN       Black       2
```

```
56893      NaN        17.0     21.0       NaN      19.0  Hispanic       4
56894      NaN         NaN      NaN       NaN       NaN     White       4
56895      NaN         NaN      NaN       NaN       NaN     White       4
56896      NaN         NaN      NaN       NaN       NaN  Hispanic       2

       POVERTY3  Age_min_hard  Age_min_light
0           3.0          20.0           16.0
1           3.0          13.0           15.0
2           2.0           NaN           26.0
3           3.0           NaN            NaN
4           3.0          32.0            NaN
…         …           …            …
56892       3.0           NaN            NaN
56893       3.0          14.0           14.0
56894       3.0          21.0           22.0
56895       3.0           NaN            NaN
56896       2.0           NaN            NaN

[56897 rows x 16 columns]
```

Let's now look at the distribution of ages at which people started taking hard drugs

[167]:
```
distr_age_hard = df["Age_min_hard"].hist(bins=40, range=[9, 50])
distr_age_hard.set_xlabel("Age")
distr_age_hard.set_ylabel("count")
distr_age_hard.set_title("Distribution of Ages at which individuals first␣
 ↪starting taking hard drugs")
```

[167]:
```
Text(0.5, 1.0, 'Distribution of Ages at which individuals first starting taking
hard drugs')
```

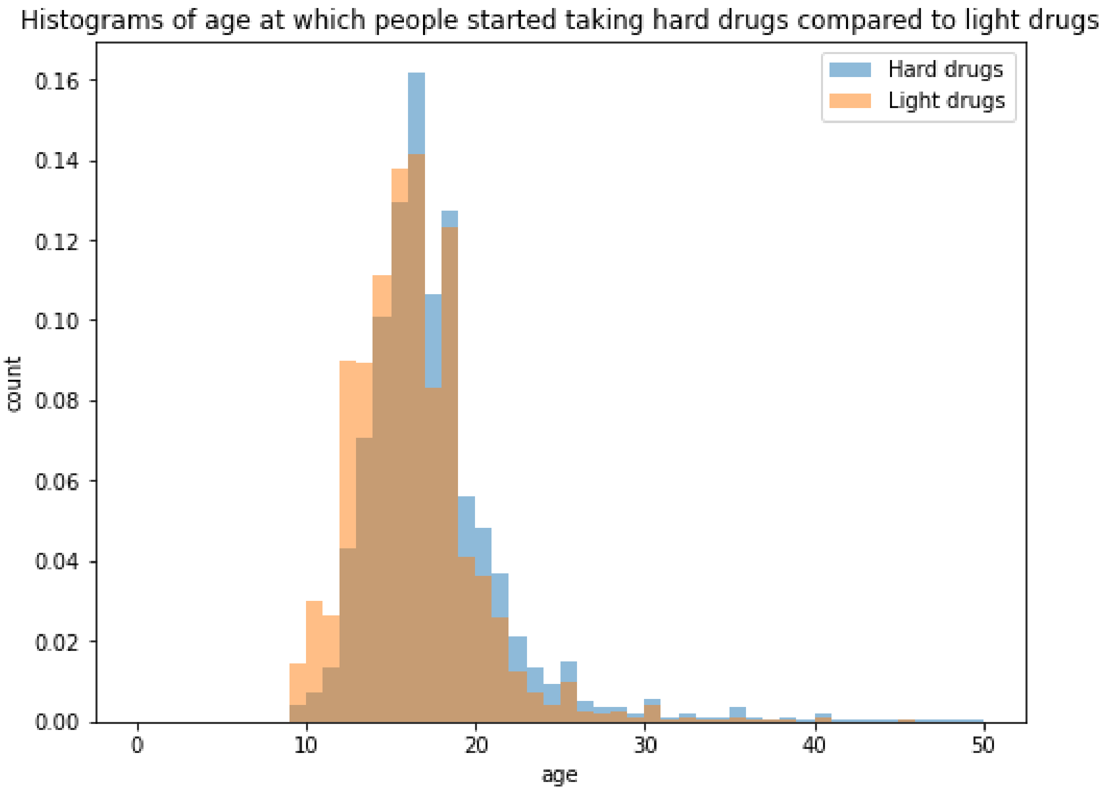


We see that the histogram is skewed to the right and that the majority of people are aged between 16 to 18. We also see that there is only a small amount of people who start taking drugs after the age of 30.

Let's now look at the distribution of ages at which people started taking light drugs and compare it to the ages at which people started taking hard drugs.

```
[168]: plt.figure(figsize=(8,6))
plt.hist(df["Age_min_hard"], bins=50, alpha=0.5, label="Hard drugs", range=[0,␣
 ↪50], density=True)
plt.hist(df["Age_min_light"], bins=50, alpha=0.5, label="Light drugs",␣
 ↪range=[0, 50], density=True)
plt.xlabel("age")
plt.ylabel("count")
plt.title("Histograms of age at which people started taking hard drugs compared␣
 ↪to light drugs")
plt.legend(loc='upper right')
```

```
[168]: <matplotlib.legend.Legend at 0x7f78d7ebb1f0>
```



We clearly see that the two plots are very similar. However, the graph for light drugs is slightly more shifted to the left which means that people tend to start taking lighter drugs a bit earlier than hard drugs which makes sense. Indeed, because light drugs are less harmful and more socially accepteble, people may not be as scared to take them.

Let's now see if there are any outliers in the Age_min_hard and Age_min_light columns

[169]:
```
#Let's see what the expected number of people should be in each age group
df["Age_min_hard"].value_counts().mean()
```

[169]:
```
426.6909090909091
```

[170]:
```
#for hard drugs:
df["Age_min_hard"].value_counts()[df["Age_min_hard"].value_counts() < 10]
```

[170]:
```
42.0    9
50.0    9
37.0    9
48.0    8
55.0    8
43.0    6
46.0    6
```

```
41.0    4
47.0    4
44.0    4
49.0    4
53.0    3
68.0    2
60.0    2
69.0    2
65.0    1
51.0    1
58.0    1
72.0    1
61.0    1
78.0    1
62.0    1
66.0    1
Name: Age_min_hard, dtype: int64
```

Here we see that there a few age groups where the count of people is very low (Much lower than the expected value of 426). These look like outliers.

```
[171]: #for light drugs:
       df["Age_min_light"].value_counts()[df["Age_min_hard"].value_counts() < 10]
```

```
[171]: 37.0    7
       47.0    4
       48.0    3
       42.0    3
       50.0    3
       60.0    2
       46.0    2
       55.0    2
       41.0    1
       49.0    1
       44.0    1
       61.0    1
       43.0    1
       51.0    1
       Name: Age_min_light, dtype: int64
```

There exists a few age groups in which the count of people is low. These also look like outliers.

We will not treat these outliers differently at all because it is possible for a person to start taking drugs later in their life and removing values that are actually representative of real life makes our analysis less representative of real life.

Now that we looked at the distribution of our variables and searched for any outliers, we will now search for relationships between our variables

9.0.2 Relation Between Variables

Income and Race Let's see if there exists any link between the race column and income. We would expect asian and white people to have the highest income and hispanic and black peope to have the lowest income after looking at data from the federal reserve [1]. However, because we have so many white people in the dataset, the results might not reflect the reality really well.

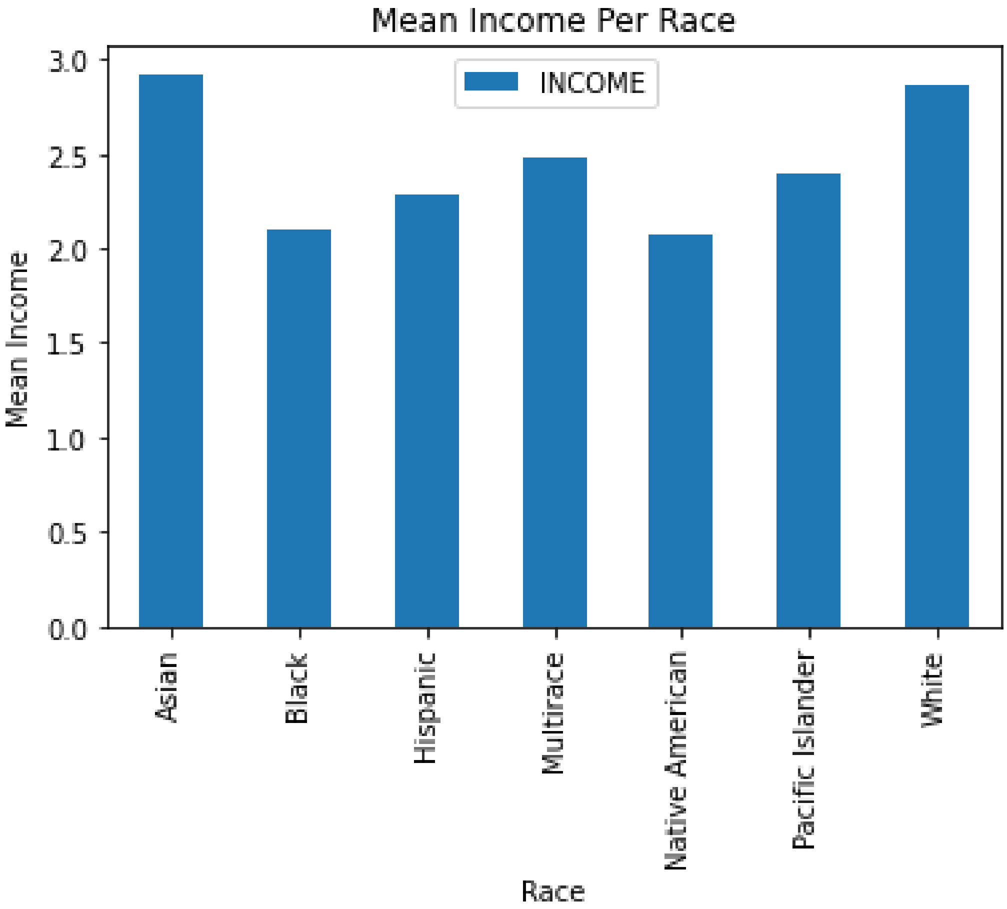
Source: https://www.federalreserve.gov/econres/notes/feds-notes/disparities-in-wealth-by-race-and-ethnicity-in-the-2019-survey-of-consumer-finances-20200928.htm

[172]:

```
#Let's look at a barplot of the mean income of each race present in the dataset.
df[["INCOME", "RACE"]].groupby("RACE").mean().plot(kind="bar")
plt.title('Mean Income Per Race')
plt.xlabel('Race')
plt.ylabel('Mean Income')
```

[172]:

```
Text(0, 0.5, 'Mean Income')
```

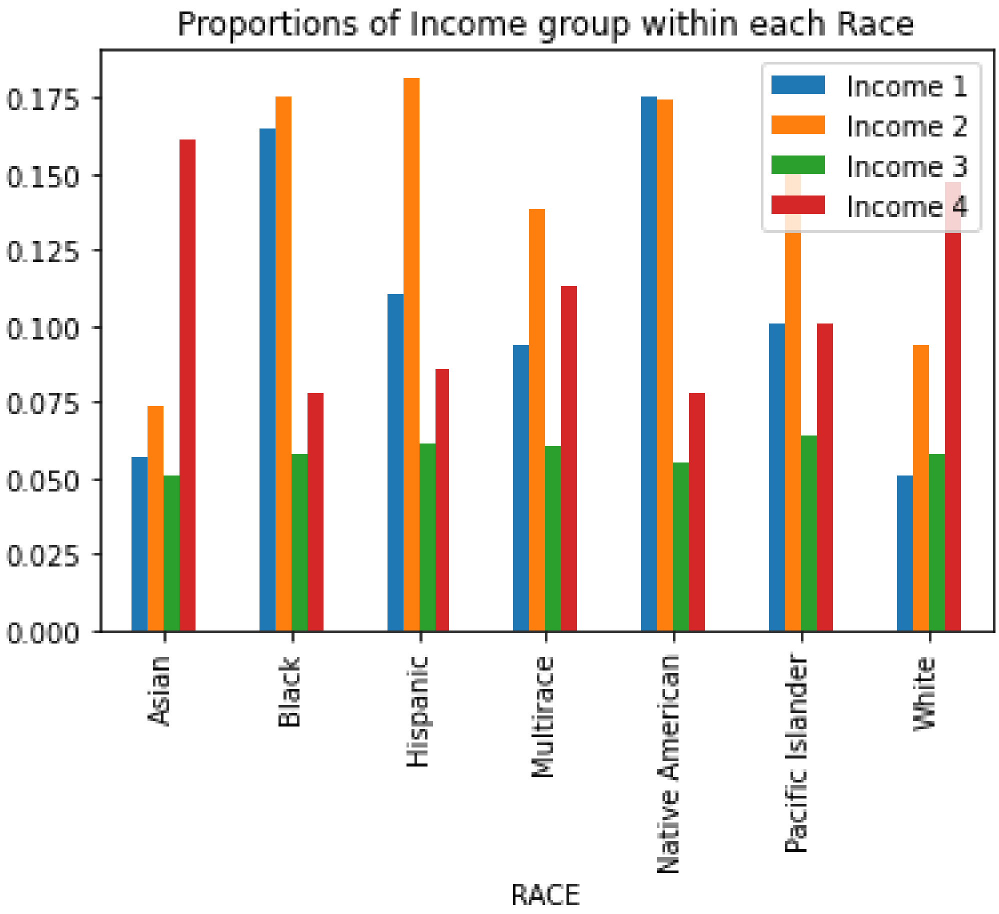


From the graph above, we see that Asian people have the highest average income (close to 3.0), followed very closely by white people. Native american people and Black people have the lowest income, with average incomes close to 2.0. These observations follow what we expected from the federal reserve research. Hispanic people are in the middle of all of this and have an average income close to 2.3.

Now, it would be interesting to compare the distributions of each income category within each race and see how they compare between races.

```
[173]: #Let's look at a barplot of the proportions of each income category within each␣
        ↪race
       fig, ax = plt.subplots()
       pp = df[["INCOME", "RACE"]].groupby(["RACE", "INCOME"]).size()/df[["INCOME",␣
        ↪"RACE"]].groupby("RACE").sum()["INCOME"]
       pp = pd.DataFrame(pp, columns=['PERCENT'])
       pp.unstack().plot(kind='bar', title ='Proportions of Income group within each␣
        ↪Race', ax=ax)
       ax.legend(['Income 1', 'Income 2', 'Income 3', 'Income 4'])
```

```
[173]: <matplotlib.legend.Legend at 0x7f78d7dfd640>
```



From the graph above, we see that Black and Native American have a very high proportion of Income 1 and 2 people and low proportions of income 4 people. This explains why the mean income of these 2 groups is so low. White people and Asians have the highest proportion of Income 4 people and low proportions of other income groups which explains why their average was so high in the previous bar chart.

9.0.3 Income and Age

Let's now look at the relationship between the income and the age columns and see if we can draw any interesting conclusions.

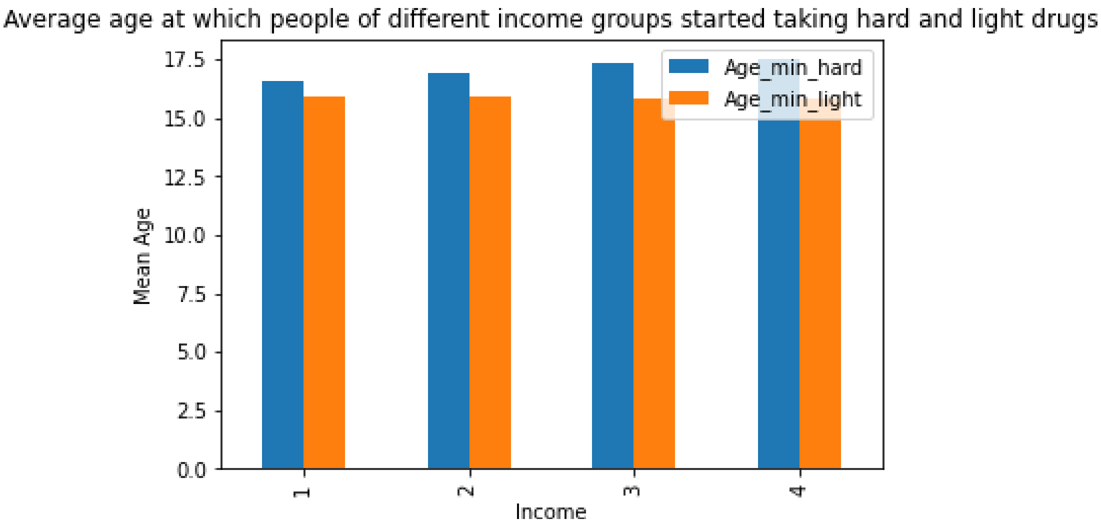
Let's start by seeing the average ages at which people of different income groups started taking hard and light drugs

```
[174]: #Here, we get the average ages at which people in different income groups start␣
        ↪taking hard and light drugs
       df[["INCOME", "Age_min_hard", "Age_min_light"]].groupby(["INCOME"]).mean()
```

```
[174]:         Age_min_hard  Age_min_light
       INCOME
       1          16.559158      15.892702
       2          16.914250      15.876635
       3          17.333149      15.840504
       4          17.483773      15.805594
```

```
[175]: df[["INCOME", "Age_min_hard", "Age_min_light"]].groupby(["INCOME"]).mean().
        ↪plot(kind="bar")
       plt.title('Average age at which people of different income groups started␣
        ↪taking hard and light drugs')
       plt.xlabel('Income')
       plt.ylabel('Mean Age')
```

```
[175]: Text(0, 0.5, 'Mean Age')
```



From the graph above, we firstly notice that the mean age at which a person from any income group starts taking light drugs is always lower than the age at which a person starts taking hard drugs. This makes sense and follows what was said in our hypothesis. We also notice that mean age at which people starts taking hard drugs increases when the income increases and reaches its hightest when the income group is 4. This doesn't really follow our hypothesis. Indeed, we thought that the age at which a person starts taking hard drugs would be lower for wealthier people. However, we see that it is not the case here. The age at which a person starts taking light drugs decreases when the income increases. This means that people from higher socioeconomic status start taking light drugs earlier than people from lower socioeconomic status. We did not expect that observation either when writing our hypothesis.

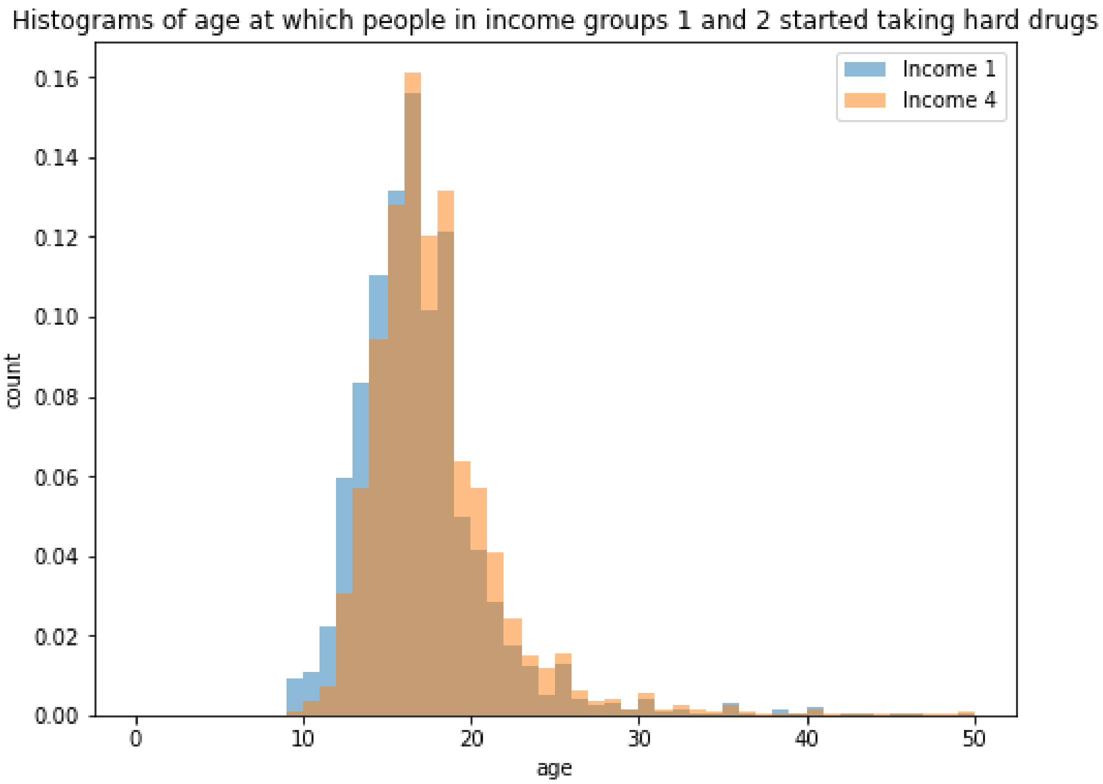
The chart above doesn't follow our theory so we will explore the distribution of ages at which people from different age groups started taking hard drugs and compare distributions for income group 1 and income group 4.

[176]:

```
plt.figure(figsize=(8,6))
plt.hist(df[df["INCOME"] == 1]["Age_min_hard"], density=True,
         bins=50, alpha=0.5, label="Income 1", range=[0, 50])
plt.hist(df[df["INCOME"] == 4]["Age_min_hard"], density=True,
         bins=50, alpha=0.5, label="Income 4", range=[0, 50])
plt.xlabel("age")
plt.ylabel("count")
plt.title("Histograms of age at which people in income groups 1 and 2 started␣
 ↪taking hard drugs")
plt.legend(loc='upper right')
```

[176]:

```
<matplotlib.legend.Legend at 0x7f78d7d3e5b0>
```

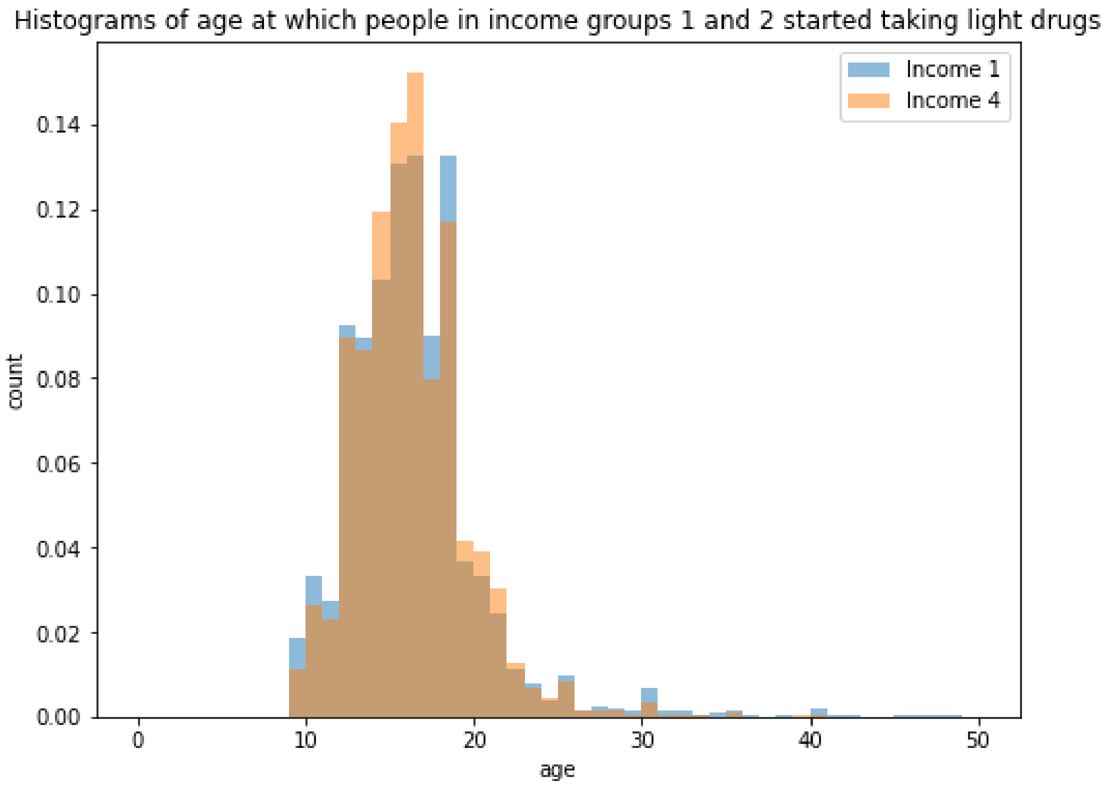


From this graph, we see that peope in income 1 start taking hard drugs a little earlier than people in income 4 because the distribution for income 1 is slightly more to the left than the one for income 4. This shows that people in lower income groups start taking hard drugs earlier which is not what we expected. However, from this graph, we see that the mode for both incomes is the same (around 17-18).

Let's now compare distribution of ages at which people from different age groups started taking light drugs and compare distributions for income group 1 and income group 4.

[177]:
```
plt.figure(figsize=(8,6))
plt.hist(df[df["INCOME"] == 1]["Age_min_light"], density=True,
         bins=50, alpha=0.5, label="Income 1", range=[0, 50])
plt.hist(df[df["INCOME"] == 4]["Age_min_light"], density=True,
         bins=50, alpha=0.5, label="Income 4", range=[0, 50])
plt.xlabel("age")
plt.ylabel("count")
plt.title("Histograms of age at which people in income groups 1 and 2 started␣
 ↪taking light drugs")
plt.legend(loc='upper right')
```

[177]:
```
<matplotlib.legend.Legend at 0x7f78d7bb5a90>
```



From this graph, we see that peope in income 1 and 4 start taking light drugs around the same time. We also see that more people in income 1 start taking light drugs at rare ages such as 10, 12 or 30 years old.

Race and Age Let's now look at the relationship between the race and age columns and see if we can draw any interesting conclusions.

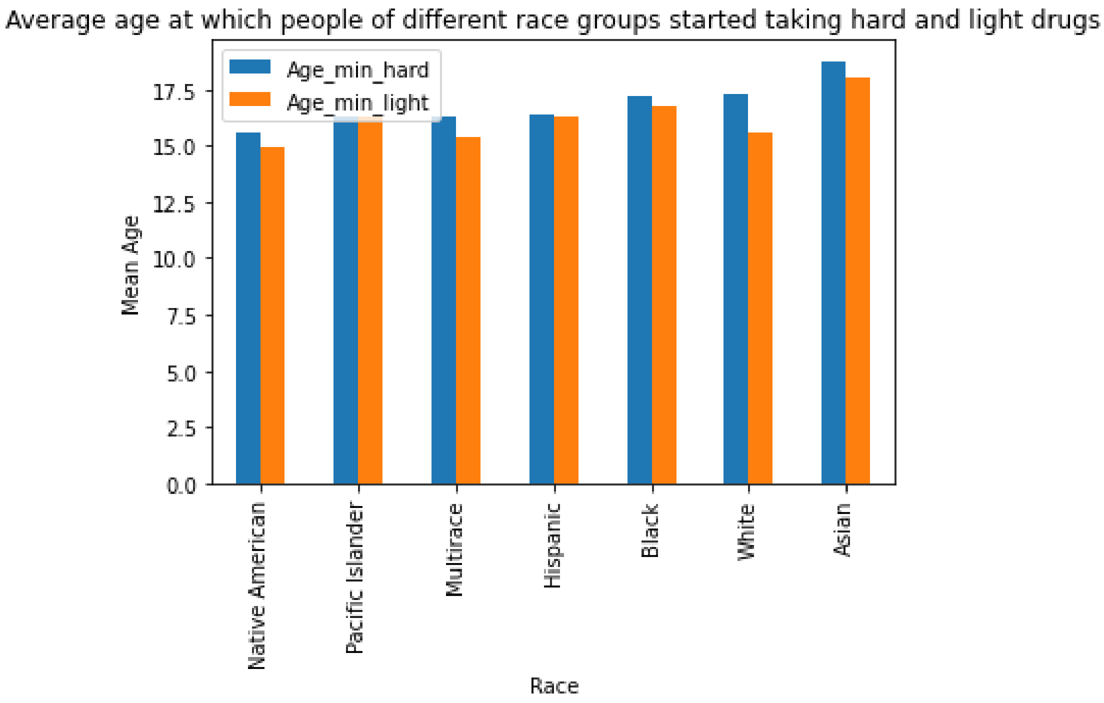
Let's start by seeing the average ages at which people of different race groups started taking hard and light drugs

```
[178]: df[["RACE", "Age_min_hard", "Age_min_light"]].groupby(["RACE"]).mean().
        ↪sort_values(by = "Age_min_hard")
```

```
[178]:                    Age_min_hard  Age_min_light
        RACE
        Native American      15.612009      14.965795
        Pacific Islander     16.270492      16.330579
        Multirace            16.299000      15.424365
        Hispanic             16.426002      16.301923
        Black                17.203580      16.731605
        White                17.265352      15.596951
        Asian                18.791139      17.998494
```

```
[179]: df[["RACE", "Age_min_hard", "Age_min_light"]].groupby(["RACE"]).mean().
        ↪sort_values(by = "Age_min_hard").plot(kind="bar")
       plt.title('Average age at which people of different race groups started taking␣
        ↪hard and light drugs')
       plt.xlabel('Race')
       plt.ylabel('Mean Age')
```

```
[179]: Text(0, 0.5, 'Mean Age')
```



From this graph, we firstly see that the average age at which a person starts using light drugs is lower than the average age at which a person starts using hard drugs for every age except for pacific islanders. This makes sense because light drugs have a smaller risk associated to them so people tend to start them earlier. We also see that Asian people start taking drugs the latest and Native Americans start taking drugs the earliest. We also see that Black and white people mean ages for hard drugs are very close to each other and are the second and third highest.

10 Modeling

For our model, we decided on using a Decision Tree from the package Sci Kit Learn. Our choice was between Sci Kit Learn and XGBoost. Since we have many nulls in our dataset we had to make a choice on how to deal with the nulls. XGBoost treats nulls as unknowns, meaning missing data. However, our nulls are not unknown, they are intentionally nulls because those people have not done those drugs either at all or in the time period being studied. Therefore, XGBoost's method will not satisfy our needs. Alternatively, Sci Kit Learn does not allow for nulls. Because of this we will have to replace nulls with a high value like 900 in order to represent nulls. This method will allow for handling intentional nulls in a better way.

[190]:

```
df = df.replace({np.nan:900})
```

Sci Kit Learn does not handle strings either, so in order to classify, we need to represent each Race value as a different numerical value.

[191]:

```
df['RACE'] = df['RACE'].replace({'White':0, 'Hispanic':1, 'Asian':2, 'Black':3,␣
 ↪'Multirace':4,
       'Native American':5, 'Pacific Islander':6})
```

Here we set our X input predictions and our different outputs, income and race.

[192]:

```
X = df[['CIG_AGE', 'SMKLESS_AGE', 'CIGAR_AGE', 'MJ_HASH_AGE', 'COKE_AGE',
       'CRACK_AGE', 'HER_AGE', 'HALLUC_AGE', 'LSD_AGE', 'PCP_AGE', 'ECST_AGE']]
inc = df[['INCOME']]
race = df[['RACE']]
```

In order to validate that our model is not just being overfitted to all of our data, we split the data into 80% training and 20% testing. We chose this split because it is typically standard.

[204]:

```
inc_X_train, inc_X_test, inc_train, inc_test = train_test_split(X, inc,␣
 ↪test_size=0.2)
race_X_train, race_X_test, race_train, race_test = train_test_split(X, race,␣
 ↪test_size=0.2)
```

Below, we will be training the income and race models with the DecisionTree. We decided on using max depth of 8 as 8 is recommended for datasets that are not too complicated. Additionally, we did not want to select none for max depth and risk overtraining our model on the training set. We train the model on the X and income data, then predict income using the test X data. We then calculate precision, recall and accuracy for the income model. We then do the same process for the race model.

[205]:

```
tree = DecisionTreeClassifier(max_depth=8)
tree.fit(inc_X_train, inc_train)
best_preds = tree.predict(inc_X_test)
print('Scores for Income Prediction using SciKit and 900 Imputation for NaNs')
print("Precision = {}".format(precision_score(inc_test, best_preds,␣
 ↪average='macro')))
print("Recall = {}".format(recall_score(inc_test, best_preds, average='macro')))
print("Accuracy = {}".format(accuracy_score(inc_test, best_preds)))
tree.fit(race_X_train, race_train)
best_preds = tree.predict(race_X_test)
print('_________________________________')
print('Scores for Race Prediction using SciKit and 900 Imputation for NaNs')
print("Precision = {}".format(precision_score(race_test, best_preds,␣
 ↪average='macro')))
```

```
print("Recall = {}".format(recall_score(race_test, best_preds,␣
 ↪average='macro')))
print("Accuracy = {}".format(accuracy_score(race_test, best_preds)))
```

```
Scores for Income Prediction using SciKit and 900 Imputation for NaNs
Precision = 0.29282531271823703
Recall = 0.2726997743762293
Accuracy = 0.3542179261862917
---------------------------------
Scores for Race Prediction using SciKit and 900 Imputation for NaNs
Precision = 0.167404995539729
Recall = 0.14525582607703877
Accuracy = 0.5861159929701231
```

Looking at our accuracy, recall, and precision, we see some stark differences. Surprisingly, our model is much better at predicting race (59% accuracy) than income (35% accuracy). These results do not let us reject our null hypothesis, H0: Age when first using drugs will not better predict income than race. Surprisingly, we find that recall and precision are about 13% lower for our race prediction model than our income prediction model, with both being quite low.

11 ETHICS & PRIVACY

One ethical/privacy issue that our question encountered with the dataset that we intend to use is the possibility that an individual could be identified (and thus linked to drugs) based on the information given. While the individuals' names will not be included, it is still possible based on income, ethnicity, gender, and population density that is given in the dataset to narrow down certain groups of individuals. To prevent this we did not include each of these variables, just the ones listed above for the purpose of our research. This helped mitigate the ability to identify the individuals listed by the use of their personal information. The data is collected using sample survey information, which means that the entire population is not sampled - just those willing to take voluntary surveys. Thus, this might mean people could not report 100% honestly as drug usage is a taboo topic. This is an observational study, thus we are not directly affecting respondents. That being said, it is possible the nature of these questions could cause potential discomfort for respondents, thus we would inform respondents about the contents of the survey before they begin while we collect the data for the dataset. In addition, lower income populations might not be as well represented, as they could lack the time or resources to complete the survey (i.e. computer or internet access). The biggest issue is that people don't necessarily want to respond truthfully to this taboo topic of drug use, as they might not want to be associated with illegal activities - even anonymously. We believe that, based on the fact that the survey was voluntary, that those who chose to take it would also be willing to be completely honest. One bias that we did not account for that we encountered during the project process was that there were more white (caucasian) people in the dataset than there was previously expected. In order to combat that, we explained the bias and took that into account while making the conclusion. If the data was skewed our conclusion would not be radical just as our data would suggest.

12 Conclusion And Discussion

Our research question is whether socioeconomic status or ethnicity is better predicted by the age at which an individual started using drugs in the United States, leading to the null hypothesis of H0: Age of first drug use will not be a better predictor of income than race and alternative hypothesis H1: Age of first drug use will be a better predictor of income than race. When we tested our models, income and race, we found that income had an accuracy of about 35% and race had an accuracy of about 59%. Based on the accuracy, this leads us to failing to reject the null hypothesis as income was not better predicted by age of first drug use than race was. Neither model was great at predicting what it was meant to, with both having very low precision and recall. Our belief is that the race model performed better due to the more clear delineations between missing variables across racial groups than income groups. This becomes evident when looking back to our analysis of null values. Of course, there are some serious limitations to our results. First, our model's predictors are quite juvenile. Using age when first using drugs for 11 drugs to predict race and income is not necessarily all that predictive. Second, our dataset is very limited in variety. The vast majority of participants are white and the lowest income bracket is severly underrepresented. This underrepresenation of groups likely cause serious problems for the model to be able to accurately depict trends, as there was not enough data present to simulate the population.