MIDTERM 1 - CODING PORTION (Tuesday 2/4)

OKCupid Profiles

This is the coding portion of Midterm 1 for CS 105. This portion of the Midterm must be completed during your lab session and turned in before the end of the lab. You are free to use your notes, lecture slides, labs, and any resource online. You are NOT allowed to talk to your labmates or share information about the lab.

The TA is not able to answer questions concerning the lab, so if something is confusing or umbigious, then state your assumption and proceed with the lab. Many answers are acceptable, as long as you do a good job justifying your answer. Just answer each question to the best of your ability according to *your* interpretation of the question.

You will be asked to answer questions using public data about OKCupid users living in the San Francisco Bay Area. This dataset was obtained from https://github.com/rudeboybert/JSE_OkCupid (https://github.com/rudeboybert/JSE_OkCupid) and discussed in this paper :https://www.researchgate.net/publication/282009623_OkCupid_Data_for_Introductory_Statistics_and_Data_Scier (https://www.researchgate.net/publication/282009623_OkCupid_Data_for_Introductory_Statistics_and_Data_Scier

In [1]:

```
%matplotlib inline
import pandas as pd
import numpy as np

pd.options.display.max_rows = 20
```

Question 0

- Read in the profiles.csv file into a DataFrame.
- Print the mean, min, max, standard deviation, etc. of numeric columns in the dataframe (like age)
- Print the summary for a one of the categorical variables (like sex, status, etc.)

In [2]:

```
import pandas as pd

df_profiles = pd.read_csv("profiles.csv")
```

In [3]:

```
df_profiles["age"].mean()
```

Out[3]:

```
32.3402895939679
```

In [4]:

```
df_profiles["age"].min()
```

Out[4]:

```
18
```

In [5]:

```
df_profiles["age"].std()
```

Out[5]:

```
9.452779096971224
```

In [6]:

```
df_profiles["sex"]
```

Out[6]:

```
0        m
1        m
2        m
3        m
4        m
        ..
59941    f
59942    m
59943    m
59944    m
59945    m
Name: sex, Length: 59946, dtype: object
```

In [7]:

```
df_profiles["age"]
```

Out[7]:

```
0        22
1        35
2        38
3        23
4        29
        ..
59941    59
59942    24
59943    42
59944    27
59945    39
Name: age, Length: 59946, dtype: int64
```

Question 1

How many profiles (i.e. rows) does this dataset contain? How many features (i.e. columns) does this dataset contain?

In [8]:

```
df_profiles
```

Out[8]:

	age	body_type	diet	drinks	drugs	education	essay0	essay
0	22	a little extra	strictly anything	socially	never	working on college/university	about me: \n \ni would love to think...	currentl working as a internationa agent fo.
1	35	average	mostly other	often	sometimes	working on space camp	i am a chef: this is what that means. \n1...	dedicatin everyday t being a unbelievabl b.
2	38	thin	anything	socially	NaN	graduated from masters program	i'm not ashamed of much, but writing public te...	i make nerd software fc musicians artists, .
3	23	thin	vegetarian	socially	NaN	working on college/university	i work in a library and go to school. . .	reading thing written by ol dead peopl
4	29	athletic	NaN	socially	never	graduated from college/university	hey how's it going? currently vague on the pro...	work work wor work + pla
...	...	...	...	...	...	...	...	.
59941	59	NaN	NaN	socially	never	graduated from college/university	vibrant, expressive, caring optimist. i love b...	the happies times hav been when lif came to.
59942	24	fit	mostly anything	often	sometimes	working on college/university	i'm nick. \ni never know what to write ab...	currentl finishing schoc for filr production.
59943	42	average	mostly anything	not at all	never	graduated from masters program	hello! i enjoy traveling, watching movies, and...	i'm a civ engineer, wh enjoys helpin the c.
59944	27	athletic	mostly anything	socially	often	working on college/university	"all i have in this world are my balls and my ...	following m dreams...<b />\n"you got dream.
59945	39	average	NaN	socially	NaN	graduated from masters program	is it odd that having a little "enemy" status ...	i work wit elderly peopl (psychotherap and .

59946 rows × 31 columns

Profiles dataset has 599946 rows and 31 columns

Question 2

How many OkCupid users reported that they are vegetarian or vegan? You might find it helpful to create a list with all the various vegetarian / vegan options, and then use the isin() function to locate rows that have a value that is part of the list.

In [9]:

```
diet_profiles = df_profiles["diet"]
vegan_result = diet_profiles.isin(["vegan"]).sum()
vegatarian_result = diet_profiles.isin(["vegetarian"]).sum()
result = diet_profiles.isin(["vegan","vegetarian"]).sum()
```

In [10]:

```
vegan_result
```

Out[10]:

```
136
```

In [11]:

```
vegatarian_result
```

Out[11]:

```
667
```

In [12]:

```
result
```

Out[12]:

```
803
```

There are 136 vegans and 667 vegatarian people. The total vegans and vegatarian is 803.

Question 3

What proportion (percentage) of OKCupid users report never smoking? Print out the value counts for each answer (response) for the 'smokes' column.

In [13]:

```
df_profiles["smokes"].value_counts()
```

Out[13]:

```
no                43896
sometimes          3787
when drinking      3040
yes                2231
trying to quit     1480
Name: smokes, dtype: int64
```

In [14]:

```
total_smokers = df_profiles["smokes"].value_counts().sum()
total_smokers
```

Out[14]:

```
54434
```

In [15]:

```
total_non_smokers = df_profiles["smokes"].loc[df_profiles['smokes'] == 'no'].value_counts
()
total_non_smokers
```

Out[15]:

```
no    43896
Name: smokes, dtype: int64
```

In [16]:

```
non_smokers_proportion =  total_non_smokers / total_smokers
non_smokers_proportion
```

Out[16]:

```
no    0.806408
Name: smokes, dtype: float64
```

The non smoker proportion is 0.806408

Question 4

Make a visualization that displays and facilitates comparison of:

- the distribution of ages of users who are currently in college/university
- the distribution of ages of users who are currently in med school

Interpret what you see (i.e. write a sentence or two summarizing the data that you observed).

In [17]:

```
education_age = df_profiles.groupby("education")["age"].value_counts()
education_age["college/university"]
```

Out[17]:

```
age
27    51
28    47
25    45
29    39
23    35
      ..
53     1
63     1
65     1
66     1
68     1
Name: age, Length: 51, dtype: int64
```

In [18]:

```
education_age
```

Out[18]:

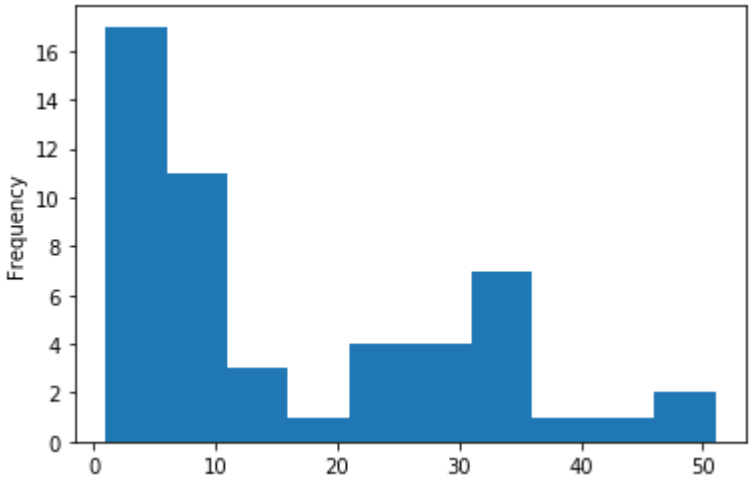
```
education                   age
college/university          27     51
                            28     47
                            25     45
                            29     39
                            23     35
                                   ..
working on two-year college 50      1
                            51      1
                            55      1
                            61      1
                            68      1
Name: age, Length: 1170, dtype: int64
```

In [19]:

```
education_age["college/university"].plot.hist()
```

Out[19]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11e003af888>
```

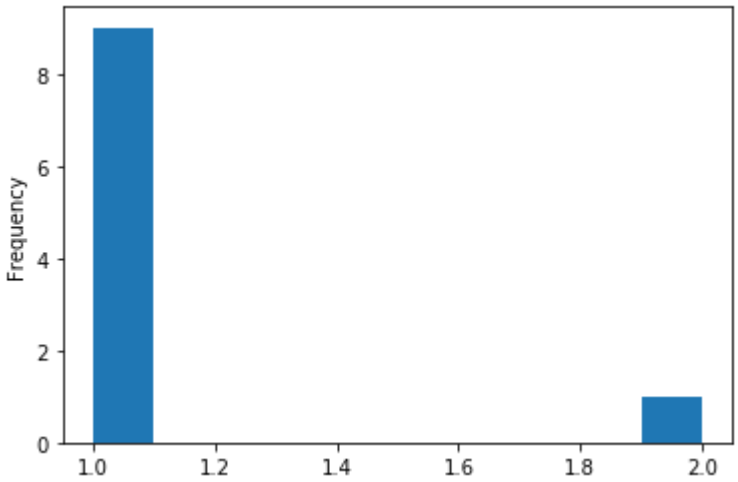


In [20]:

```
education_age["med school"].plot.hist()
```

Out[20]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11e006c20c8>
```



There's more people in college/ university than med school.

Question 5

- Make a visualization (suggest a scatter plot) that shows the average height, as a function of age and sex. Interpret what you see. Remove any outliers identified.

Hint: There are two outliers in the data set that you may want to remove to make this plot look better.

In [21]:

```
import matplotlib.pyplot as plt
df_profiles["age"].isnull().values.any()
```

Out[21]:

```
False
```

In [22]:

```
df_profiles["sex"].isnull().values.any()
```

Out[22]:

```
False
```

In [23]:

```
df_profiles["height"].isnull().values.any()
```

Out[23]:

```
True
```

In [24]:

```
df_profiles["height"].isnull().sum()
```

Out[24]:

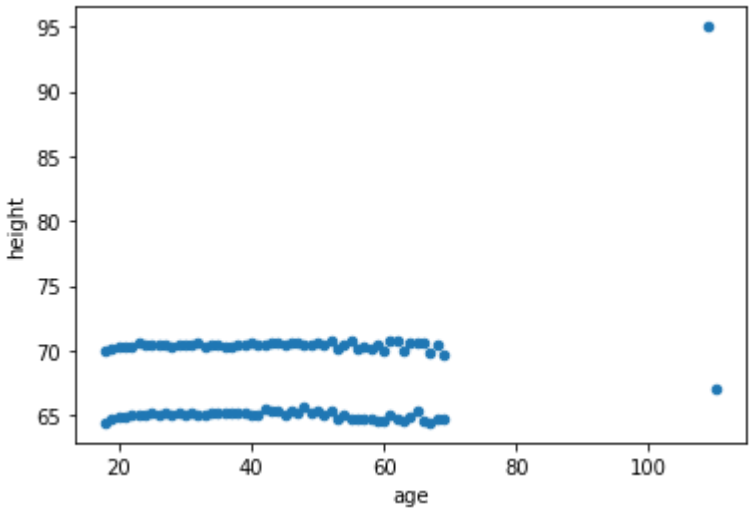
```
3
```

In [25]:

```
df_profiles["height"] = df_profiles["height"].fillna(df_profiles["height"].mean())
```

In [26]:

```
df_profiles.groupby(["sex", "age"])["height"].mean().reset_index().plot(kind = "scatter",
x ="age", y = "height")
plt.show()
```

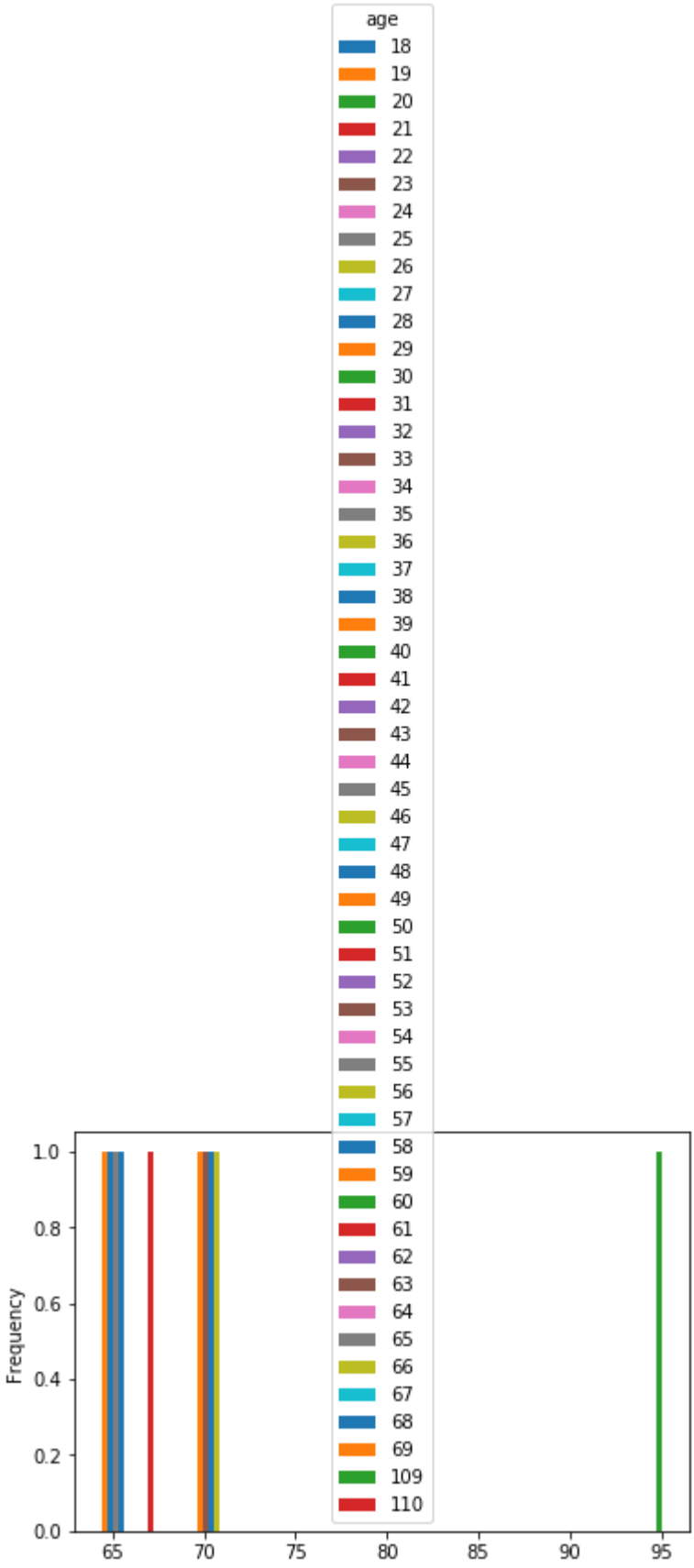


In [27]:

```
age_sex_height_table = df_profiles.pivot_table( index="sex", columns=["age"],values="heigh
t", aggfunc=np.mean)
age_sex_height_table.plot.hist(bins = 100)
```

Out[27]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11e008094c8>
```


age
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
109
110
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
65
70
75
80
85
90
95

There are good distribution throughout the year.

Question 6

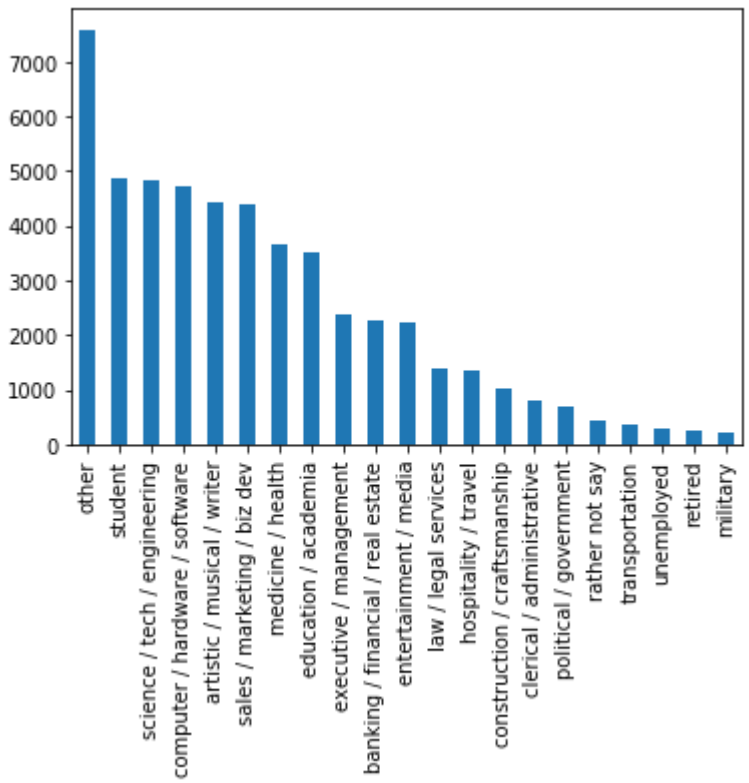
- Make a bar chart showing the number of users with each type of job.
- Sort the jobs by average reported income. (No explanation necessary.)

In [28]:

```
df_profiles["job"].value_counts().plot.bar()
```

Out[28]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11e0a1dde48>
```

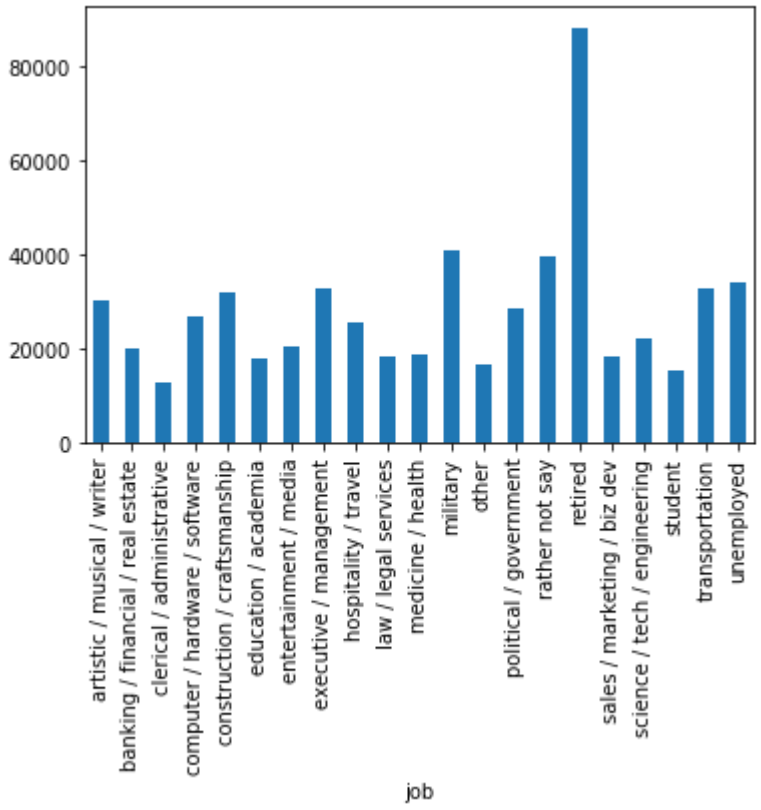


In [29]:

```
job_incomes = df_profiles.groupby("job")["income"].mean()
job_incomes.plot.bar()
```

Out[29]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11e0ab7a948>
```



Question 7

- Users are able to select from 9 essays to answer for their profile. What is the most popular essay prompt ? Reference the codebook to learn what topic each of the essays address.

In [30]:

```
total_essays = df_profiles.loc[:,'essay0':'essay9']
total_essays.count()
```

Out[30]:

```
essay0    54458
essay1    52374
essay2    50308
essay3    48470
essay4    49409
essay5    49096
essay6    46175
essay7    47495
essay8    40721
essay9    47343
dtype: int64
```

Essay 0 is the most popular essay prompt.

Submission Instructions

1. Run your code and ensure there are no errors. We will not grade a notebook with errors.
2. Upload the Python Notebook file to iLearn within the lab time to receive a grade. If you submit the coding portion of the midterm after lab you will not receive a grade.

Once your done with the midterm coding portion and succssfully upload to iLearn, then do the following:

1. Complete the mid-quarter course survey: https://docs.google.com/forms/d/e/1FAIpQLSegA6UFcAjPLaIdSK-sXr7LkzU1-MzqR0_2fFyg9J9LzBfGpQ/viewform?usp=sf_link (https://docs.google.com/forms/d/e/1FAIpQLSegA6UFcAjPLaIdSK-sXr7LkzU1-MzqR0_2fFyg9J9LzBfGpQ/viewform?usp=sf_link)
2. Complete previous labs (if any)
3. Get started with your Final Project

In [ ]:

```

```