OkCupid Data-A-Scientist

Codecademy
portofolio project

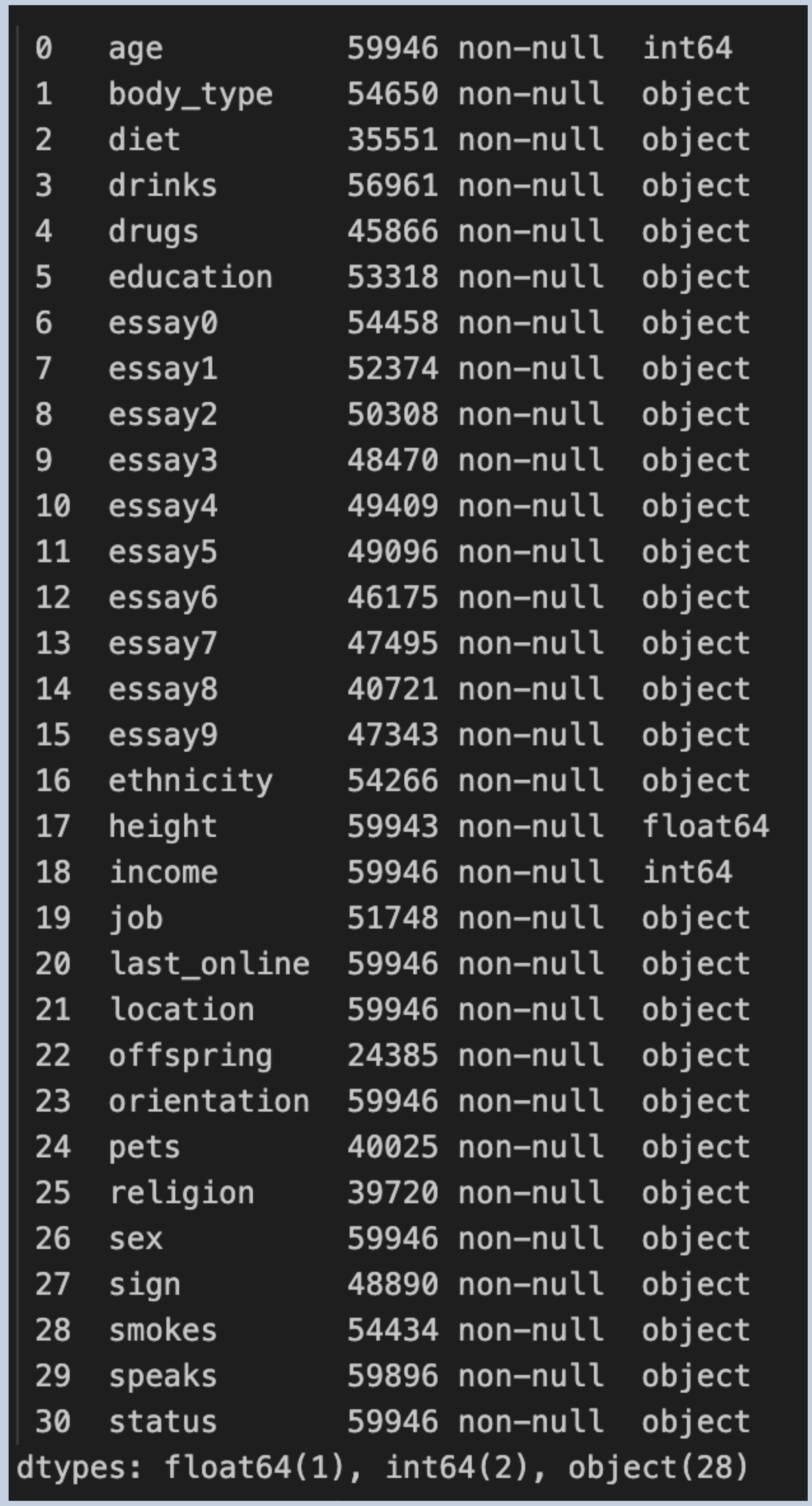
Introduction:

In recent years, there has been a massive rise in the usage of dating apps to find love. Many of these apps use sophisticated data science techniques to recommend possible matches to users and to optimise the user experience. These apps give us access to a wealth of information that we've never had before about how different people experience romance.

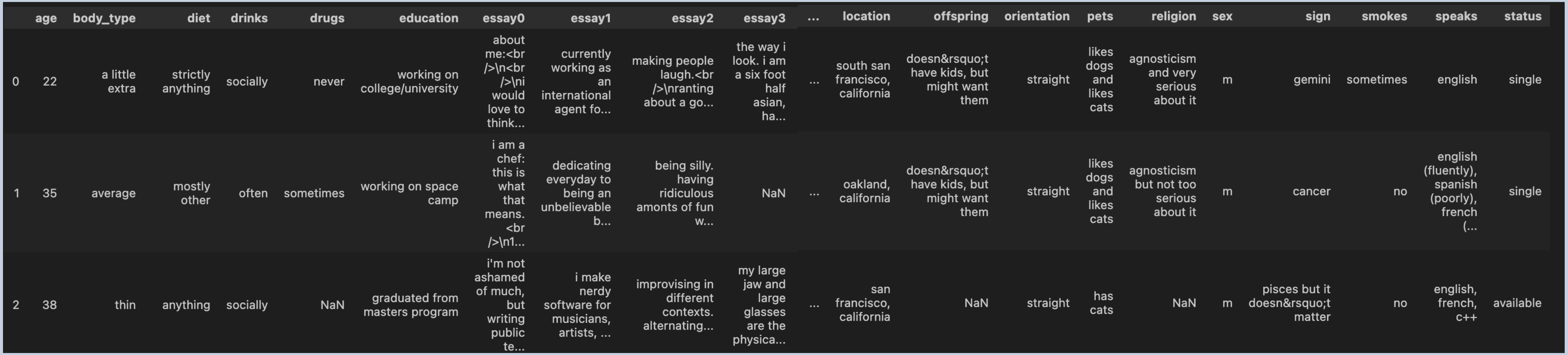
In this portfolio project, I will analyze some data from OKCupid, an app that focuses on using multiple choice and short answers to match users and also create a presentation about my findings from this OKCupid dataset. The purpose of this project is to practice formulating questions and implementing machine learning techniques to answer those questions.

Data:

We have a csv file called "profiles.csv" .
It has 31 columns as features and 59946 rows. Each row appears in the user's profile on the site. But some profiles have NaN values and we should drop the NaN datas and also if a feature has a lot of NaN values, we should drop that feature because it can make a problem in prediction and it lowers accuracy. Below, the first photo shows three columns of our data and the second shows how many NaN values each column has and the column data types.



```
0   age          59946 non-null  int64
1   body_type    54650 non-null  object
2   diet         35551 non-null  object
3   drinks       56961 non-null  object
4   drugs        45866 non-null  object
5   education    53318 non-null  object
6   essay0       54458 non-null  object
7   essay1       52374 non-null  object
8   essay2       50308 non-null  object
9   essay3       48470 non-null  object
10  essay4       49409 non-null  object
11  essay5       49096 non-null  object
12  essay6       46175 non-null  object
13  essay7       47495 non-null  object
14  essay8       40721 non-null  object
15  essay9       47343 non-null  object
16  ethnicity    54266 non-null  object
17  height       59943 non-null  float64
18  income       59946 non-null  int64
19  job          51748 non-null  object
20  last_online  59946 non-null  object
21  location     59946 non-null  object
22  offspring    24385 non-null  object
23  orientation  59946 non-null  object
24  pets         40025 non-null  object
25  religion     39720 non-null  object
26  sex          59946 non-null  object
27  sign         48890 non-null  object
28  smokes       54434 non-null  object
29  speaks       59896 non-null  object
30  status       59946 non-null  object
dtypes: float64(1), int64(2), object(28)
```



	age	body_type	diet	drinks	drugs	education	essay0	essay1	essay2	essay3	...	location	offspring	orientation	pets	religion	sex	sign	smokes	speaks	status
0	22	a little extra	strictly anything	socially	never	working on college/university	about me:<br />\n<br />\ni would love to think...	currently working as an international agent fo...	making people laugh.<br />\nranting about a go...	the way i look. i am a six foot half asian, ha...	...	south san francisco, california	doesn&rsquo;t have kids, but might want them	straight	likes dogs and likes cats	agnosticism and very serious about it	m	gemini	sometimes	english	single
1	35	average	mostly other	often	sometimes	working on space camp	i am a chef: this is what that means.<br />\n1...	dedicating everyday to being an unbelievable b...	being silly. having ridiculous amonts of fun w...	NaN	...	oakland, california	doesn&rsquo;t have kids, but might want them	straight	likes dogs and likes cats	agnosticism but not too serious about it	m	cancer	no	english (fluently), spanish (poorly), french (...	single
2	38	thin	anything	socially	NaN	graduated from masters program	i'm not ashamed of much, but writing public te...	i make nerdy software for musicians, artists, ...	improvising in different contexts. alternating...	my large jaw and large glasses are the physica...	...	san francisco, california	NaN	straight	has cats	NaN	m	pisces but it doesn&rsquo;t matter	no	english, french, c++	available

Goals:

Our goals is analyze data and clean it, then use machine learning to learn features and predict people’s status.

Data analyses and visualisation:

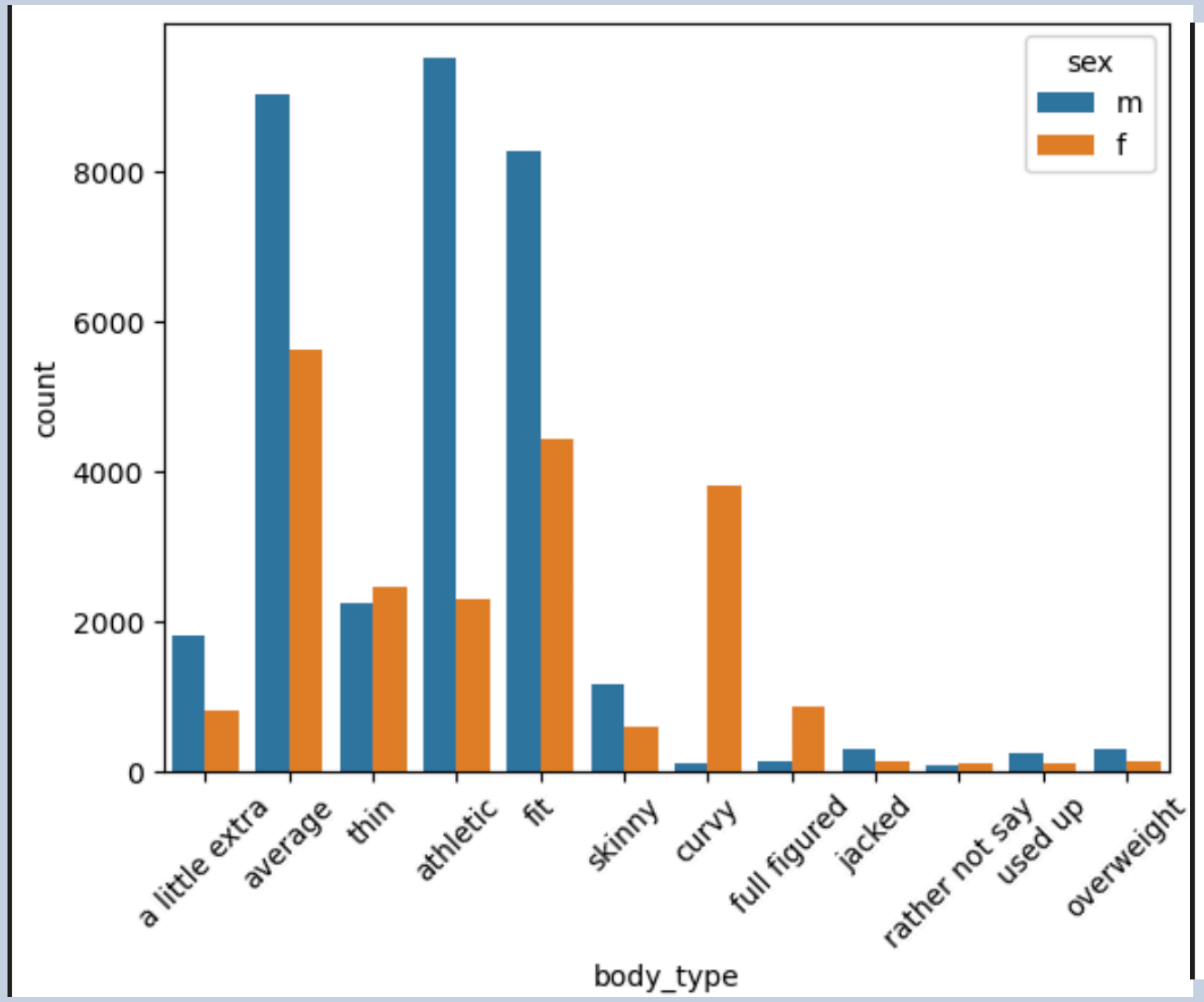
After that, I will compare features and visualise them. So at first we see how many of profiles are male or female and we conclude about 60 percent of them are male and 40 percent of them are female.

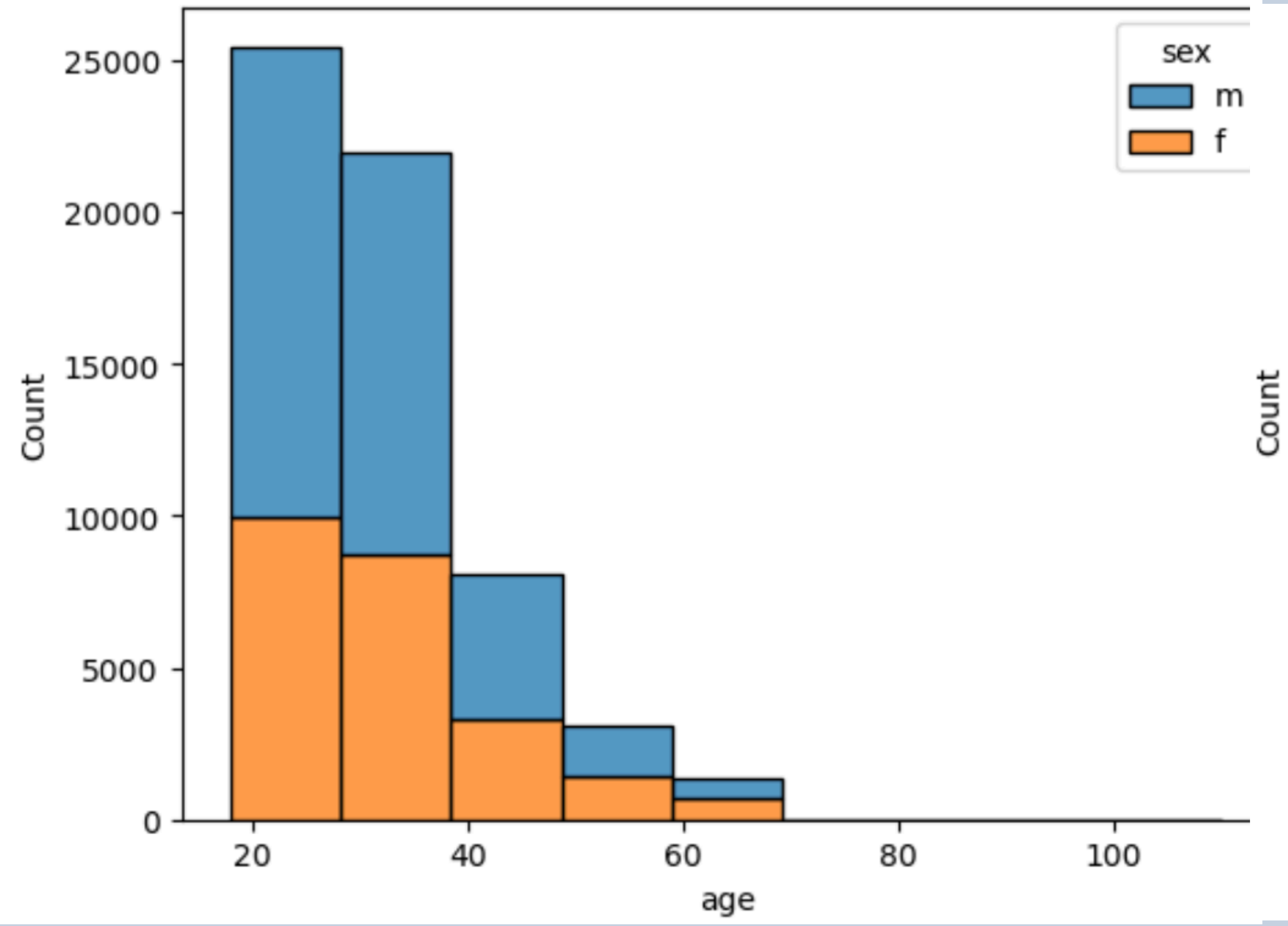
Next we compare body types, age, income, education, job, orientation based on sex.

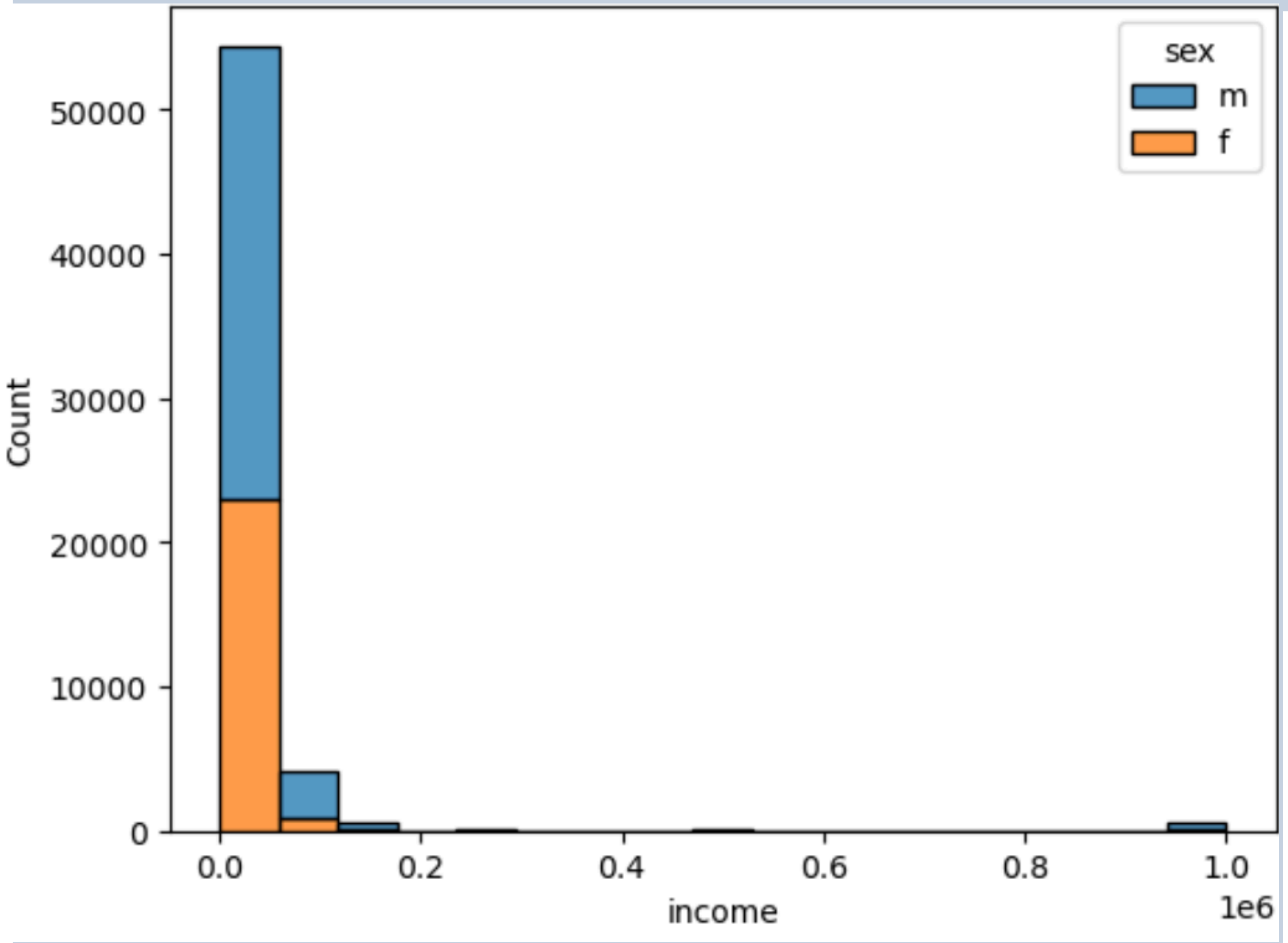
In body type section we can say most people are athletic, fit or a little extra. About age, most people around 19 to 40 years old and most of them graduated from college or university. About orientation, most of people are straight and at last most of people never use drugs and use drinks socially.

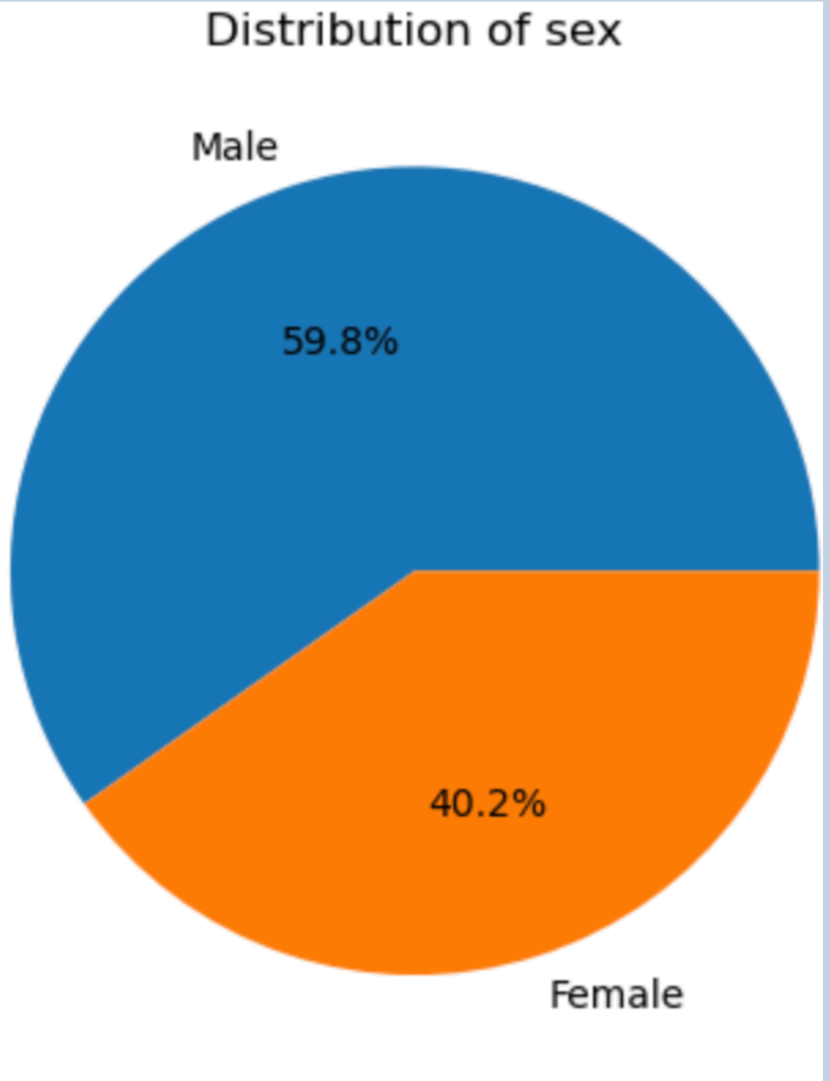
In the next slides you can see their charts.

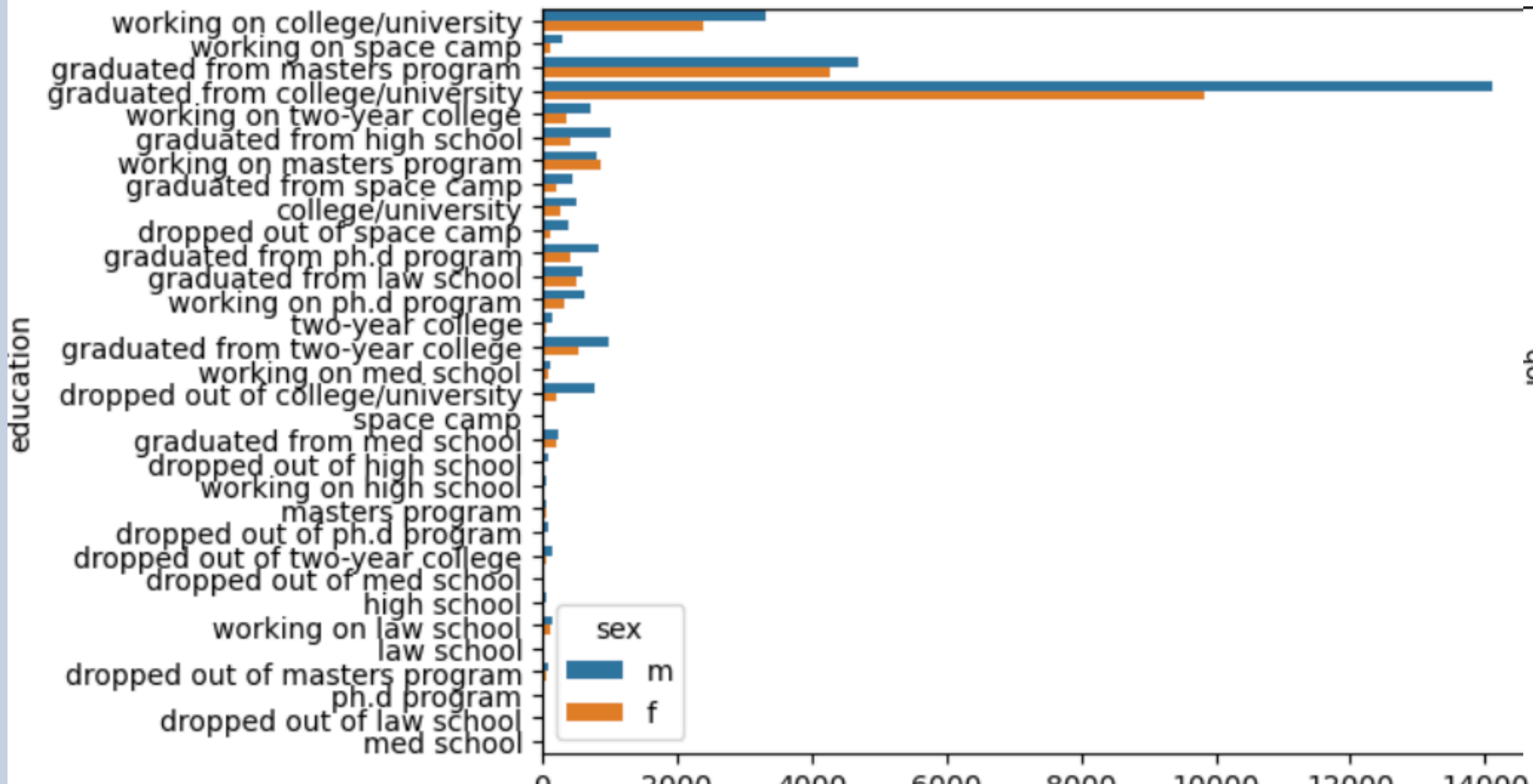
Compare some features according by sex

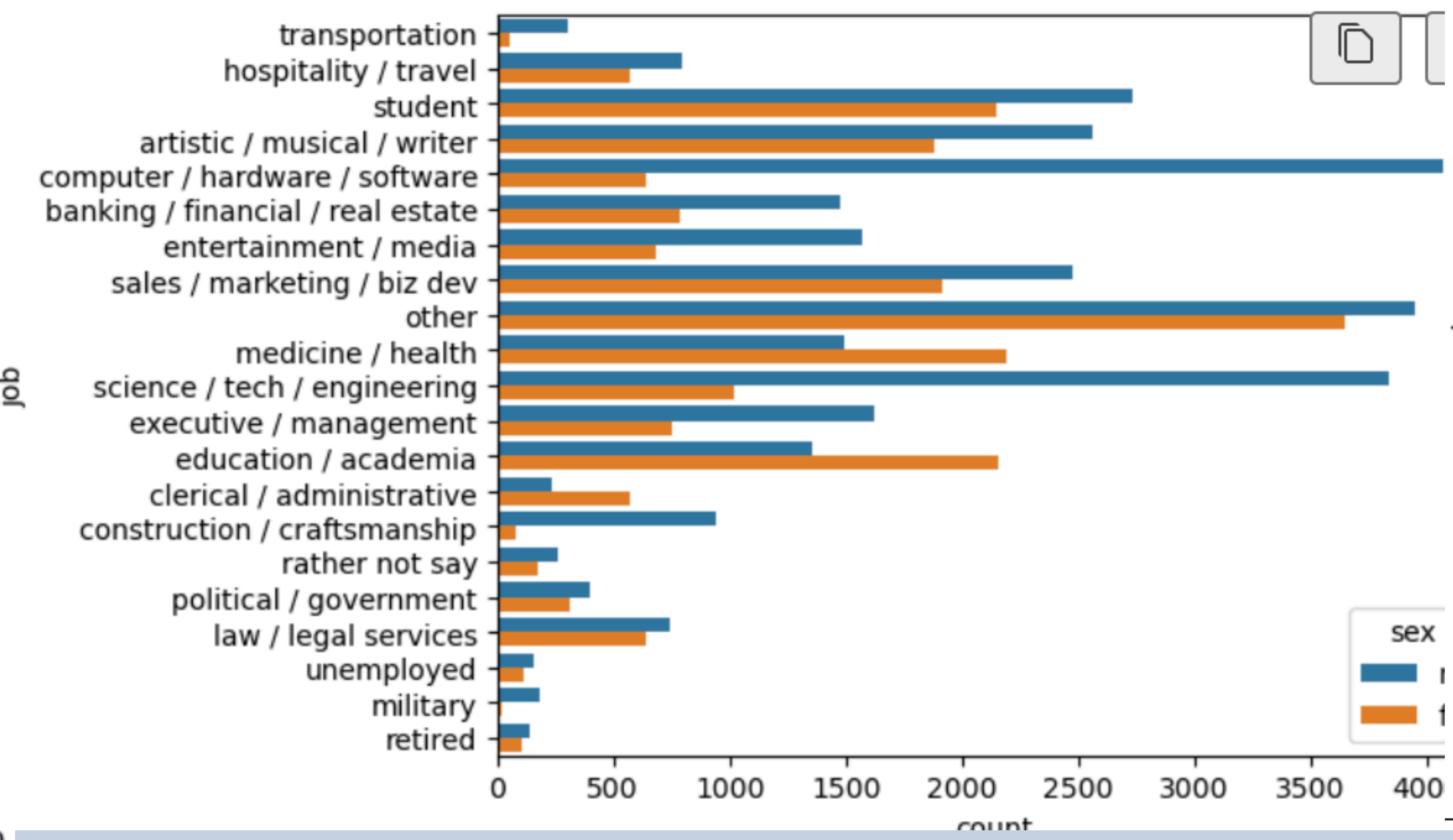


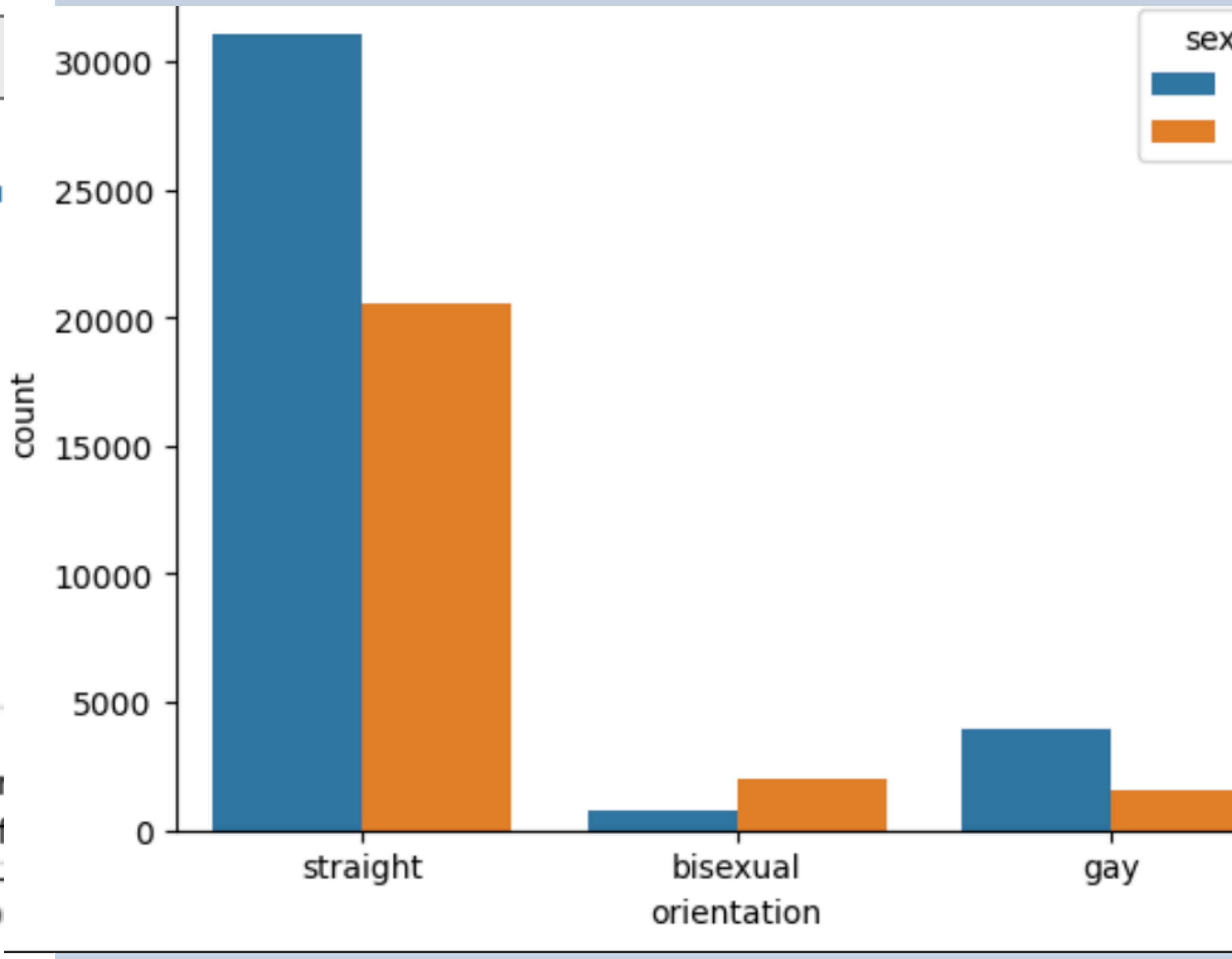




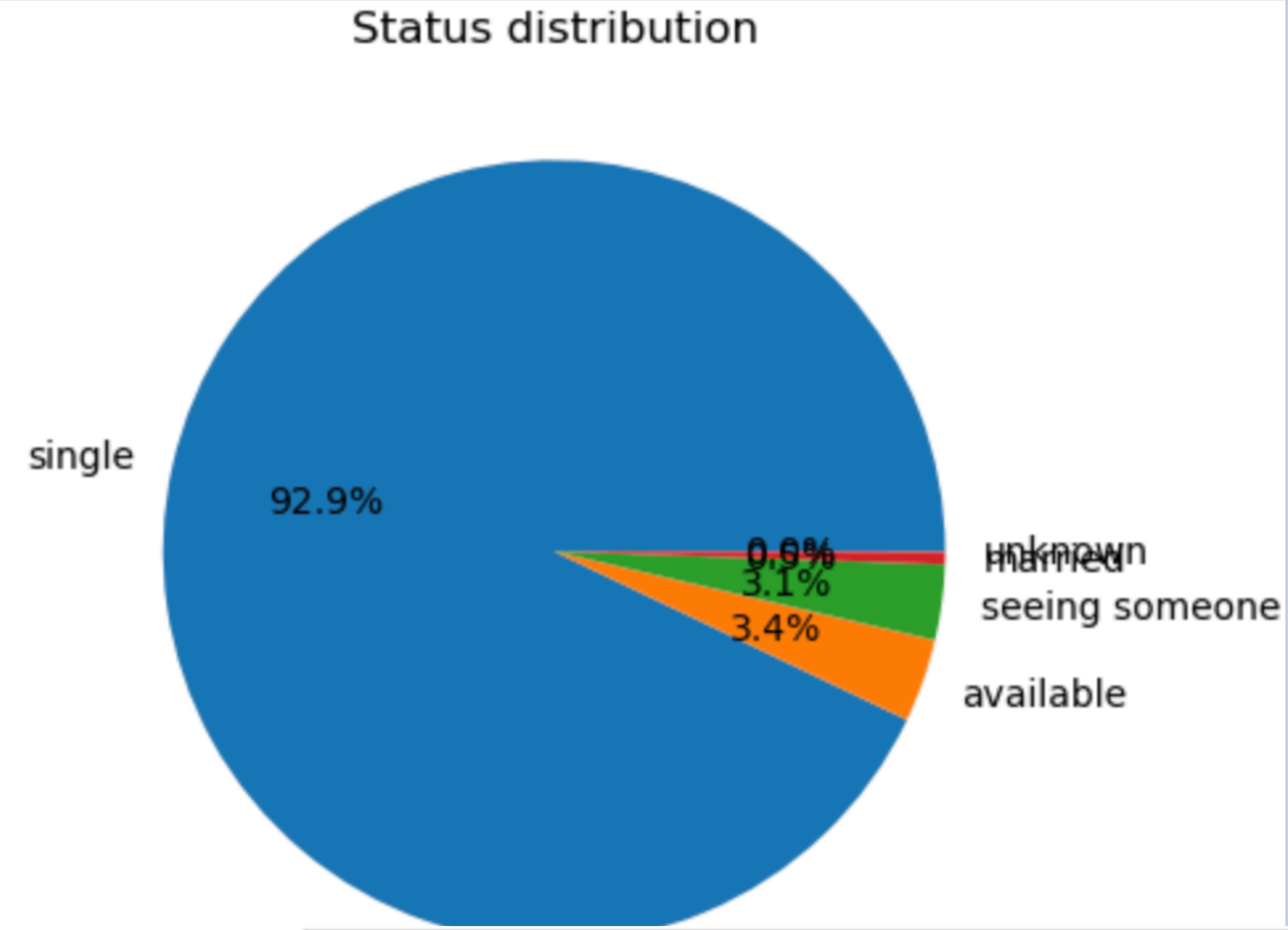


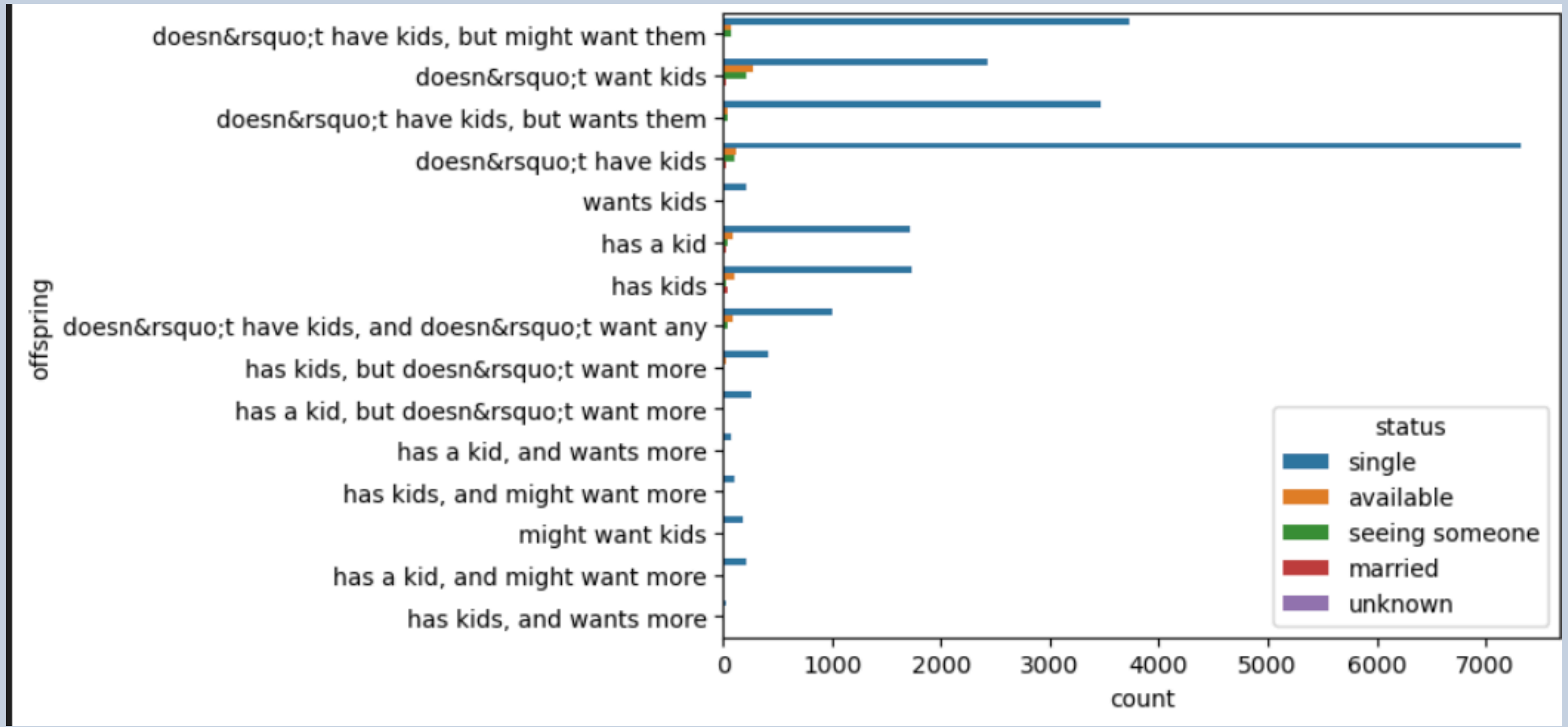


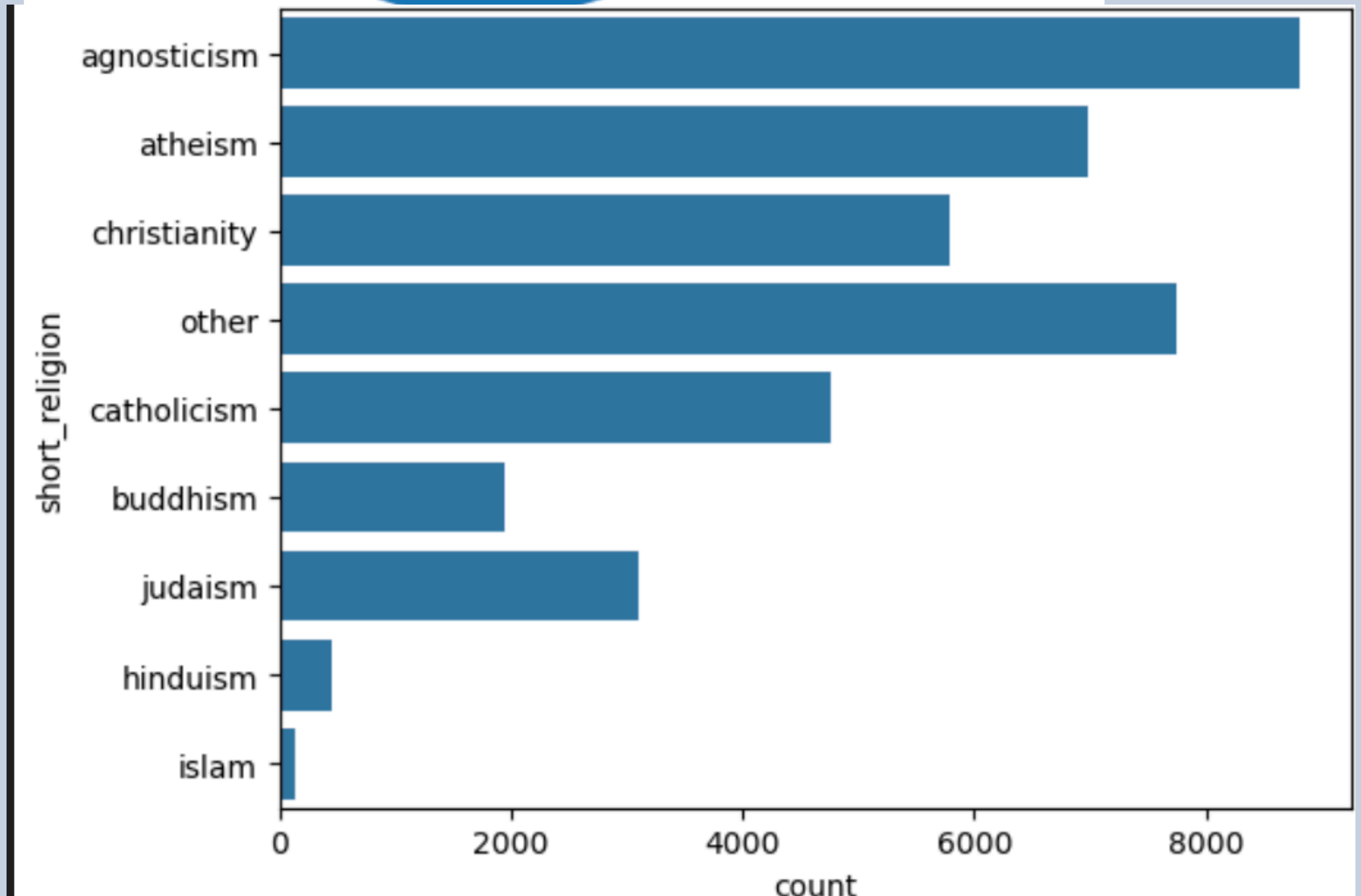


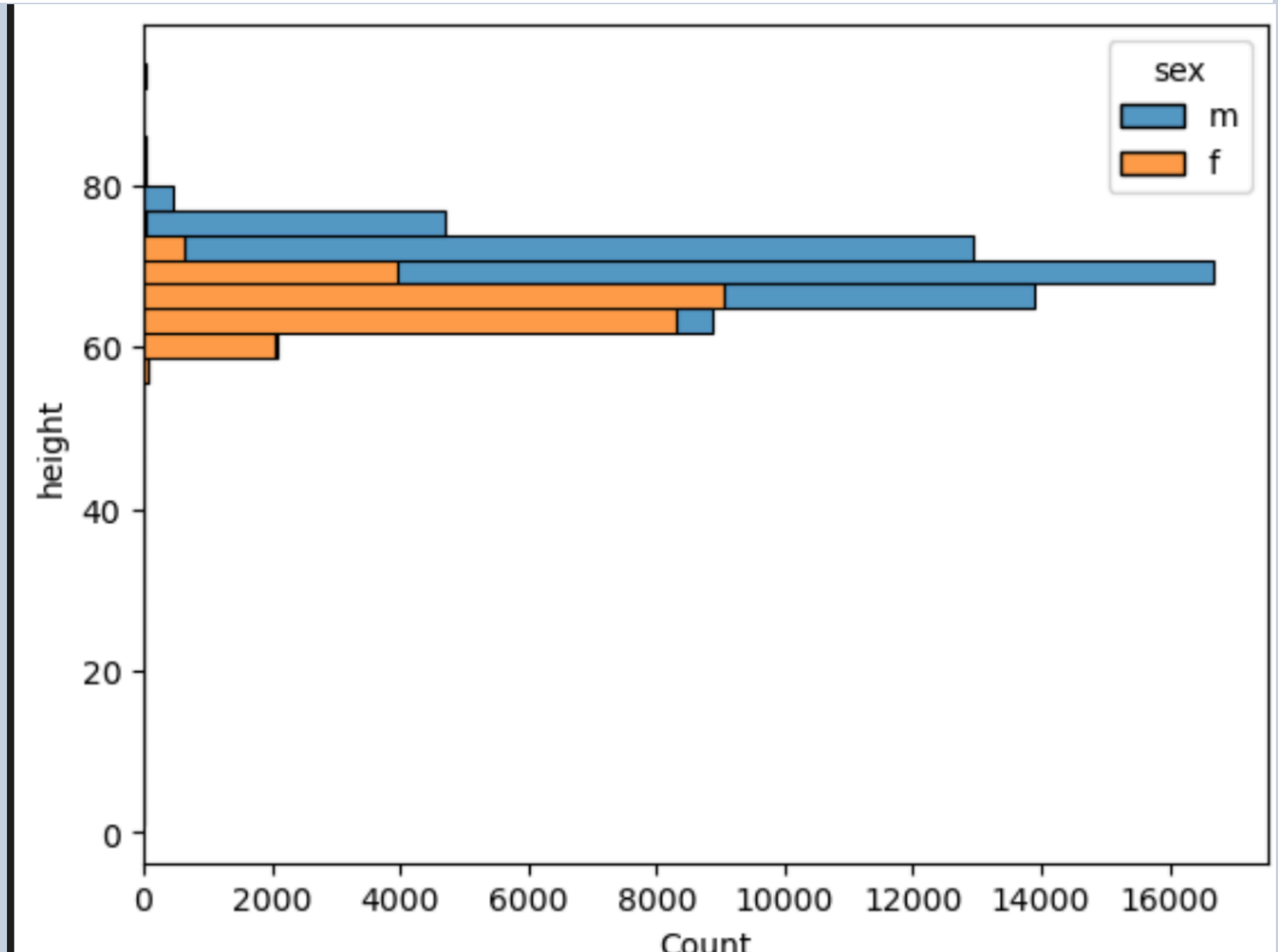


Also we can compare and visualise some other features but for some feature like religion we have a lot of data and we can clean and stack them:

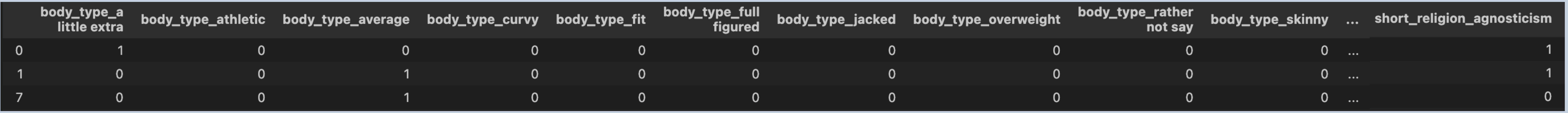




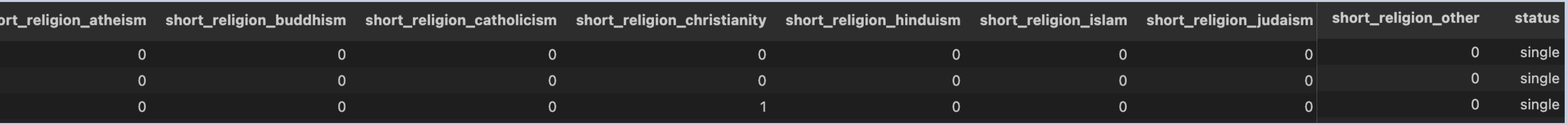




Now we can make a data frame for use the model. For this table I choose these columns:
Body type, diet, drinks, drugs, education, job, orientation, sex, clean religion and status.
I cleared my data frame of NaN datas and convert datasets to numeric datasets expect status column. Below you can see a part of our new data frame.



	body_type_a little extra	body_type_athletic	body_type_average	body_type_curvy	body_type_fit	body_type_full figured	body_type_jacked	body_type_overweight	body_type_rather not say	body_type_skinny	...	short_religion_agnosticism
0	1	0	0	0	0	0	0	0	0	0	...	1
1	0	0	1	0	0	0	0	0	0	0	...	1
7	0	0	1	0	0	0	0	0	0	0	...	0



rt_religion_atheism	short_religion_buddhism	short_religion_catholicism	short_religion_christianity	short_religion_hinduism	short_religion_islam	short_religion_judaism	short_religion_other	status
0	0	0	0	0	0	0	0	single
0	0	0	0	0	0	0	0	single
0	0	0	1	0	0	0	0	single

Now we can set the data to x and y. All columns of new data frame for x expect status and just status column for y and split the data for train and test sets.

Choose and build model:

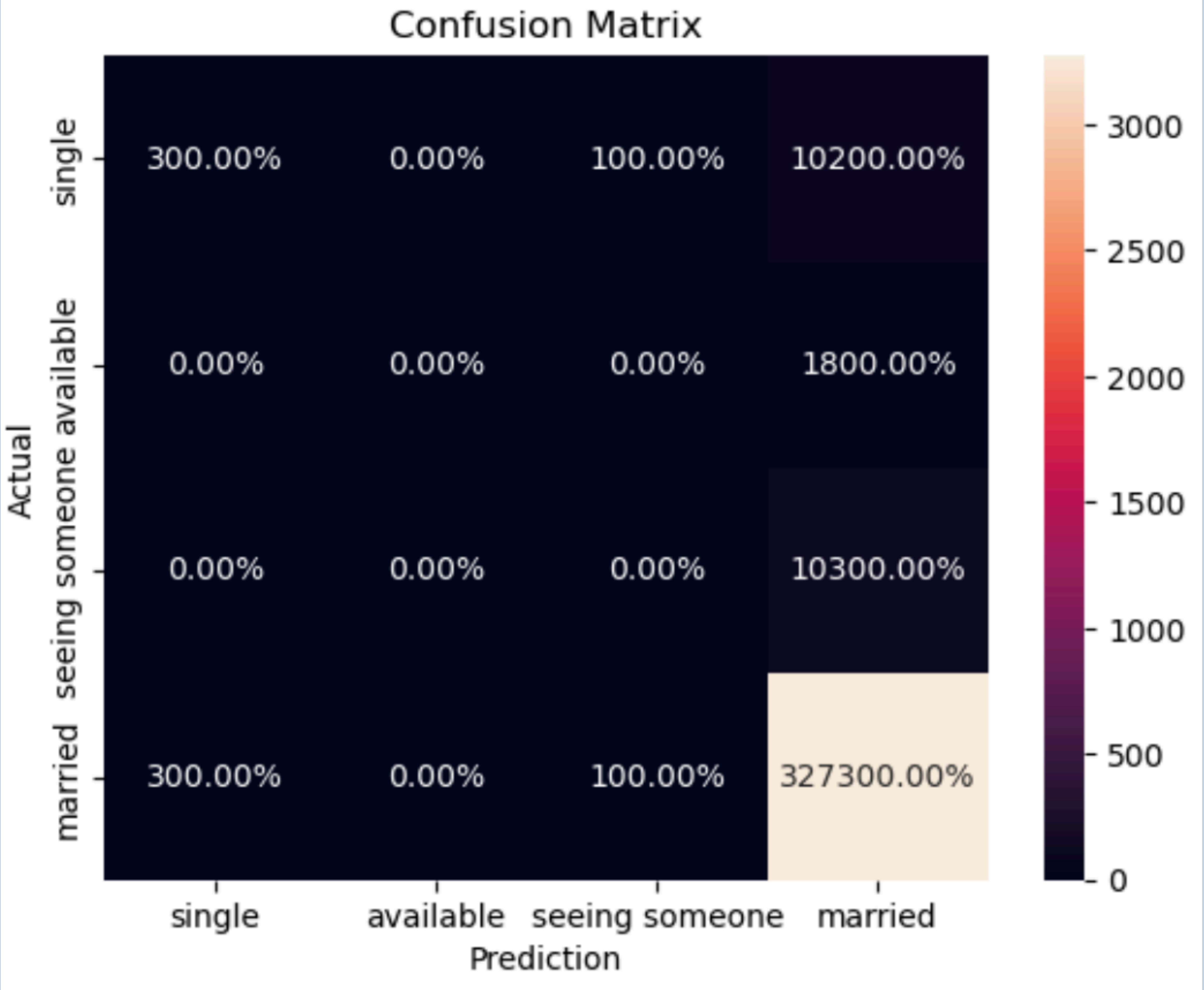
My target (the status) is a classification data so I should a model from classification ways such as K-Nearest Neighbours, Support Vector Machines and Naive Bayes and I choose K-Nearest Neighbours for this project.

K-Nearest Neighbours a “n_neighbors” parameter and for find a best n_neighbors, each time I made a model with one of the numbers 1 to 10 and fit them with train sets and calculate the accuracy of model so the best model have the highest accuracy so I find the highest accuracy and the n_neighbor of that.

Once the data is preprocessed, it is ready to model. For now we can make our K-Nearest Neighbours model and fit and learn it with our train sets and predict the data with it.

The accuracy of K-Nearest Neighbors model with our test sets is about **93.493 percent**.

At the end we can make a confusion matrix for this model.


Confusion Matrix
Actual
single
available
seeing someone
married
single
available
seeing someone
married
Prediction