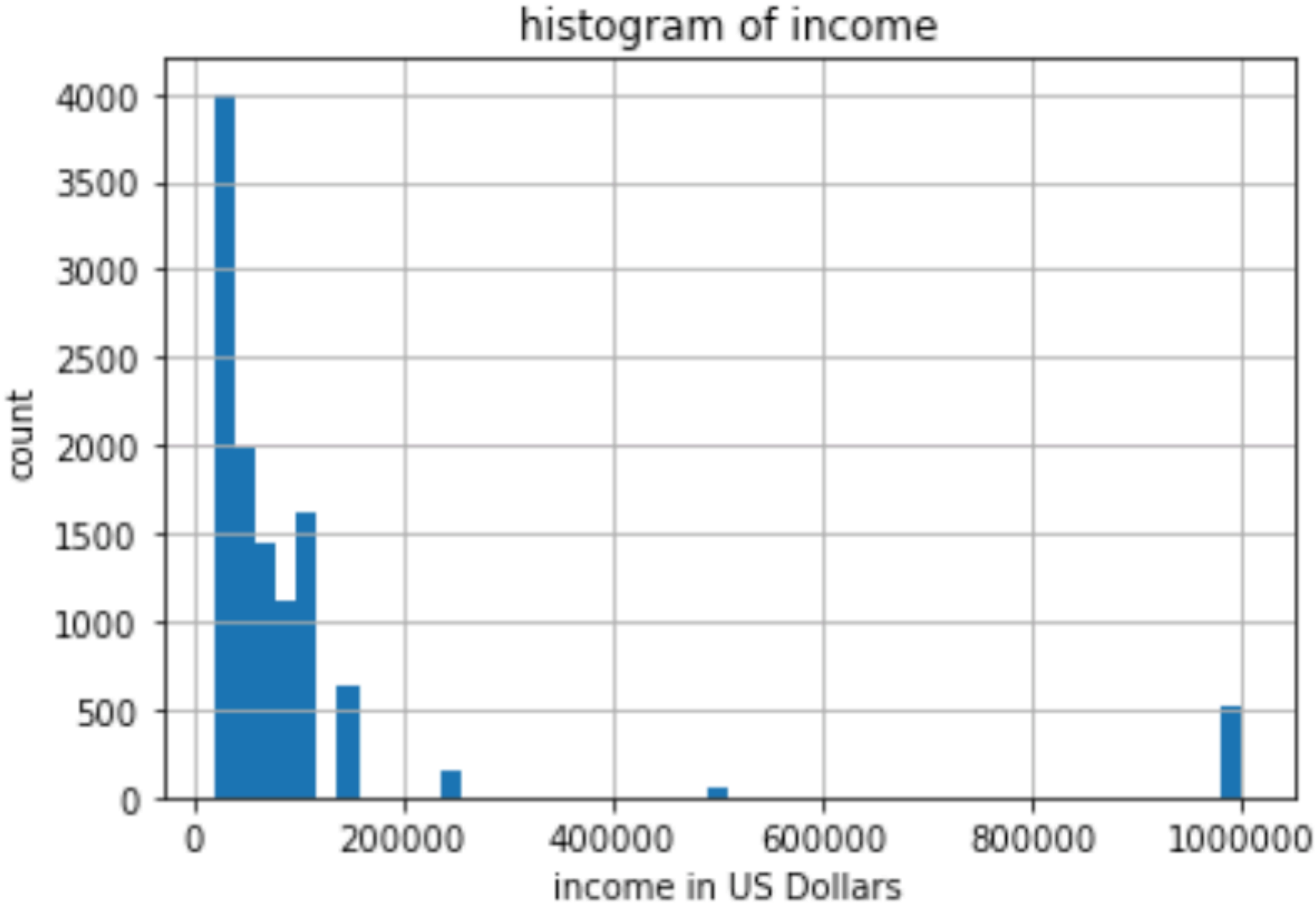
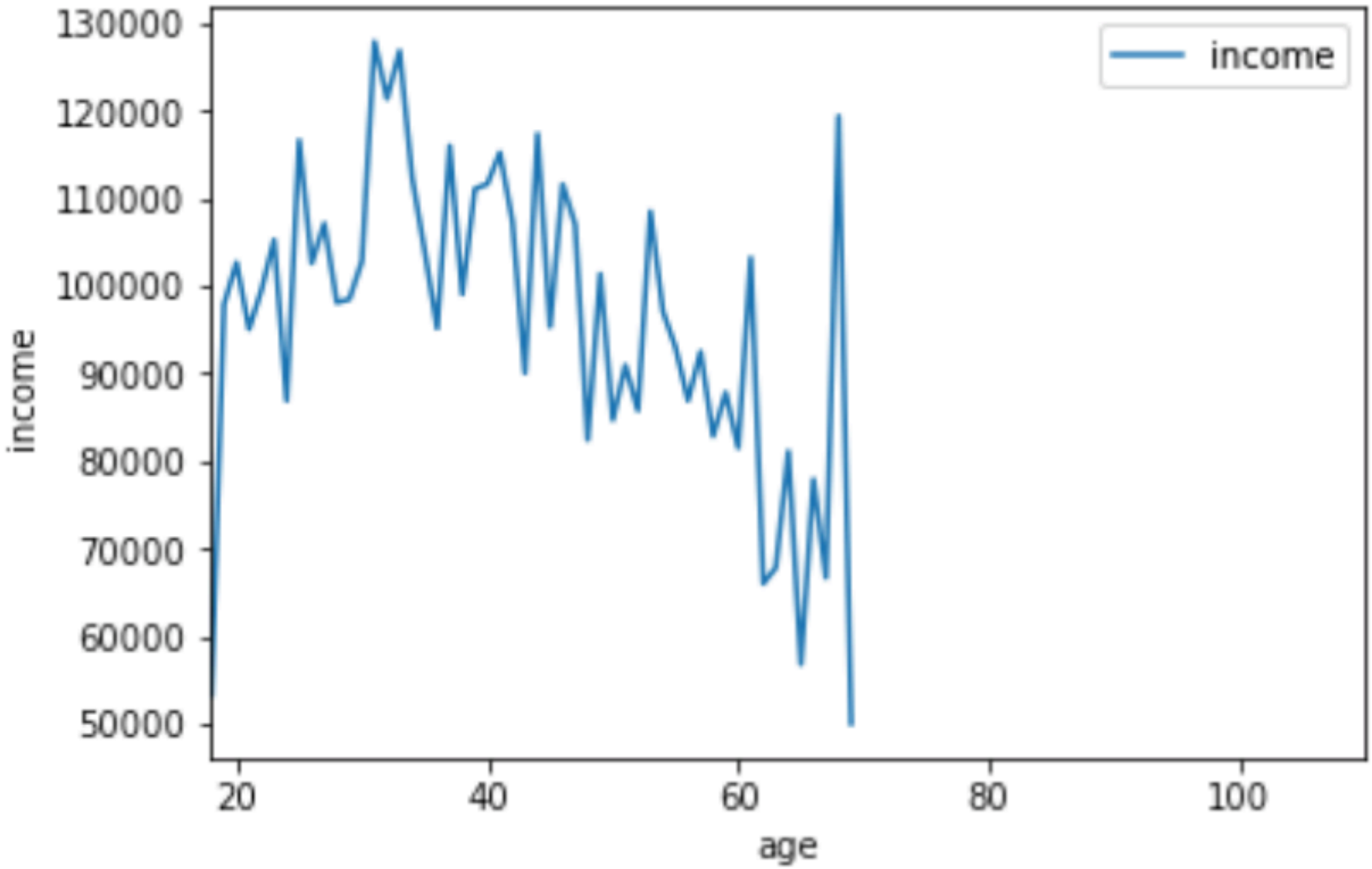
CC Capstone

I’m interested in income as dependent, thus my exploration focuses on that.

Data Exploration



Data Exploration



Question

Can I predict the level of income based on some of the OKcupid data,
features such as age, education, job, location.

Further, for the classification problem, I'm interested whether it is possible to create a classifier separating people that have and have not disclosed their income, or different is there a measurable categorical difference between the two?

New Columns

I decided to create the following new columns which might be able to predict income in our models later:

```
#classification
df['income_disclosed'] = df.income.apply(lambda x : 0 if np.isnan(x) else 1)
df['education_ranked'] = df.education.apply(mapEducation)
df['number_languages'] = df.speaks.apply(lambda x : str(x).count(',') + 1)
df['drinks_ranked'] = df.drinks.map({'not at all' : 0, 'rarely' : 1, 'socially' : 2, 'often' : 3, 'very often' : 4, 'd
df['drugs_ranked'] = df.drugs.map({'never' : 0, 'sometimes' : 1, 'often' : 2})
```

Classification

Goal: can we predict income_disclosed with our given features?

I will be using KNN and SVC to solve this problem.
KNN: it's quite simple to implement and run, and much faster than SVC. For SVC one also needs to optimize the parameters as, in particular with the rbf data, it is prone to overfit on sparse data. As it takes quite some time to run, I find this process exhausting. One can use GridSearch for SVC, but this also takes a lot of time. KNN has quite decent results whilst being quick to implement and understand.

precision: 23%, recall: 11%, accuracy: 75.2%

SVC: once properly understood and tested, it is the most powerful of the classifiers, especially on only little data.
However, I was struggling with it overfitting and only predicting negative outcomes. How can I avoid this? I played around with the gamma (making it lower) and C (making it larger), however I couldn't get it to work with me.

precision: 0%, recall: 0%, accuracy: 82%

Note to evaluators: If you give me any feedback on this at all, then this would be where I would most appreciate it. Thanks :)

Regression

Goal: can we predict income with our given features?

AS Multiple Linear Regression and KNN Regression are the only two regressions learned, I will be limited to comparing those.

MLR: really easy to implement, very quick in terms of time, quite robust. Given the feature I made I wasn't able to get a model able to accurately predict income. :/

score: 0.006

KNN Regression: As it's not constrained to a linear model (thereby a bit slower to compute, still easy to implement (slightly less so)) I was hoping it would be able to yield some answers, however it didn't.

score: -0.434

Btw, I would have loved to calc recall and precision as asked, however that is not possible on a regression problem is it?

Conclusion

Is there a difference between those disclosing their income and not?
I am very sure there is (albeit statistical sig. being unclear)

Can I predict income with the few features I used?
No.

Next steps: More feature engineering, try to make the most of the data given. Circle back and check what features yield the best models. Otherwise other variables are needed. For instance you could augment the data by getting the average income of the location from some API, or mean US income for a given Age & gender etc. There is definitely still a lot you can do with this data, some creativity, and some time.