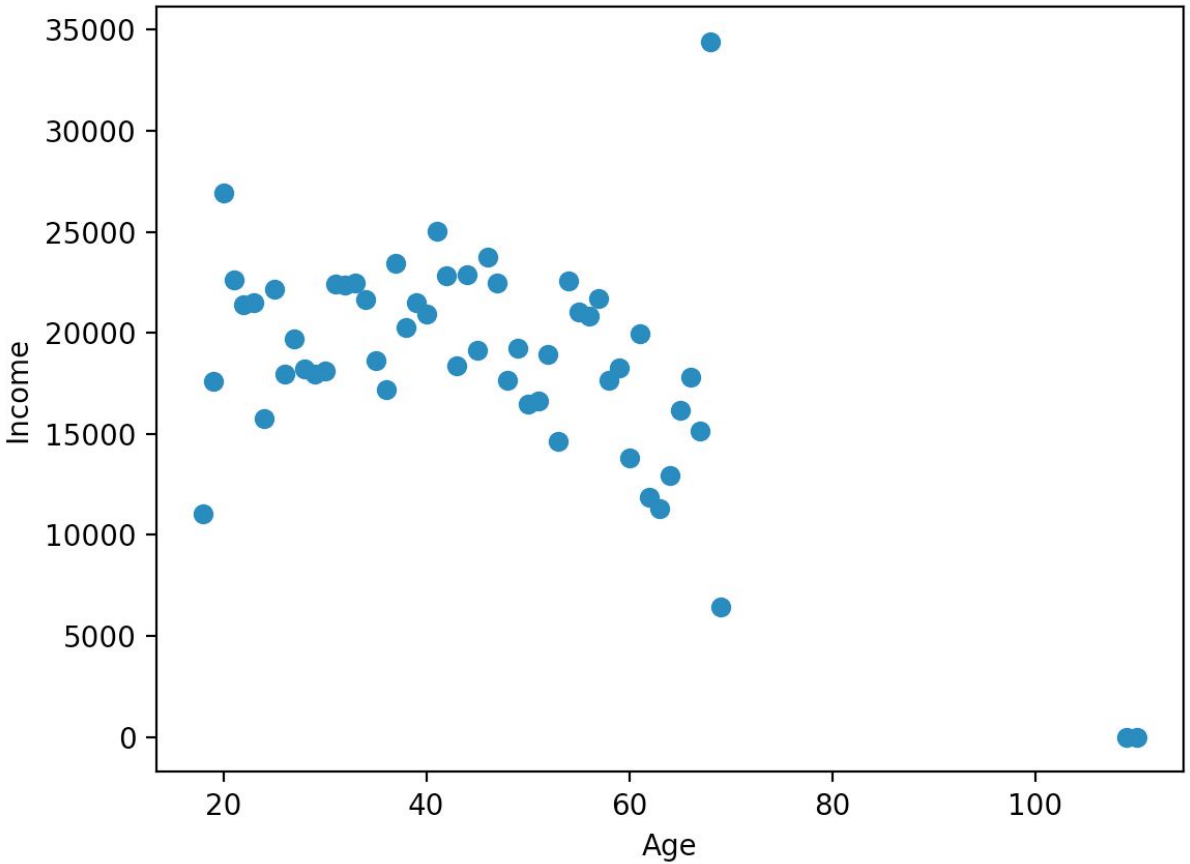
Capstone Project

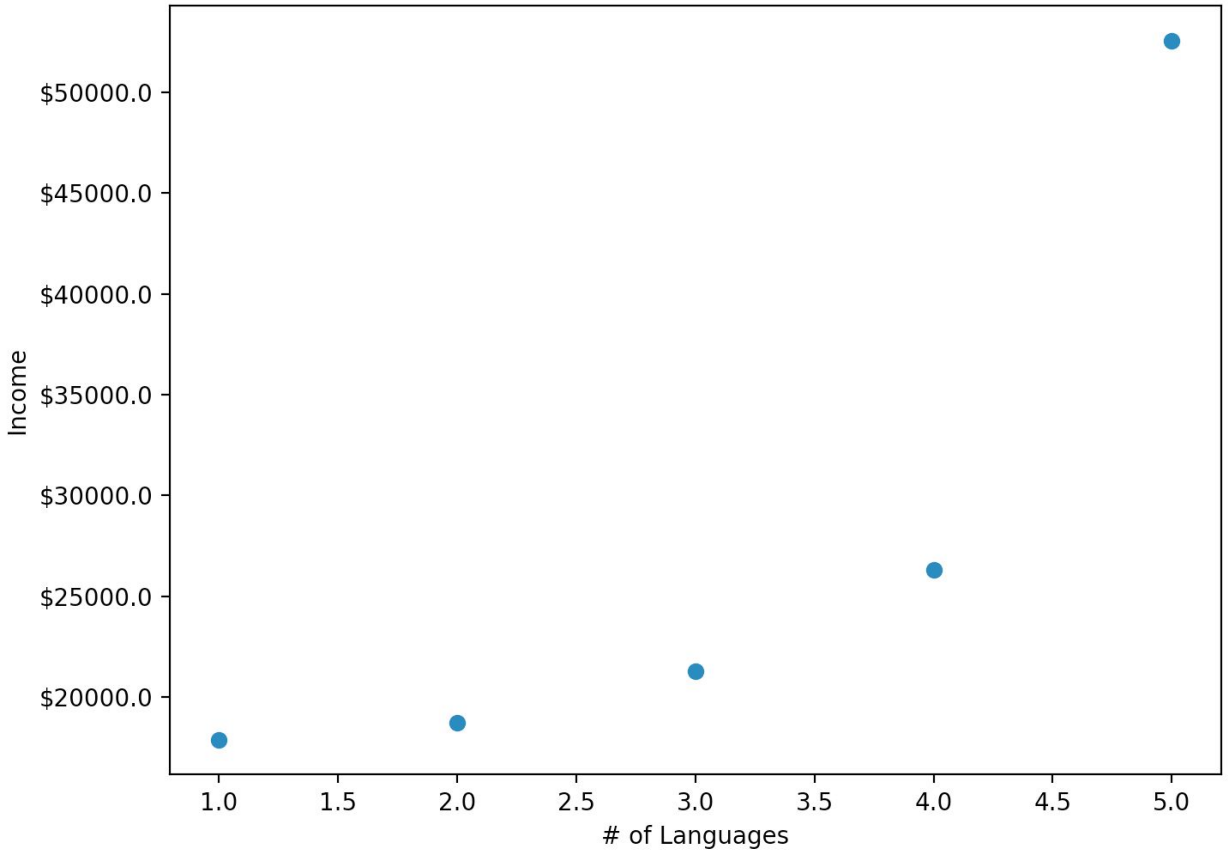
Ariel Campoverde

Can we predict income with the age of a person?

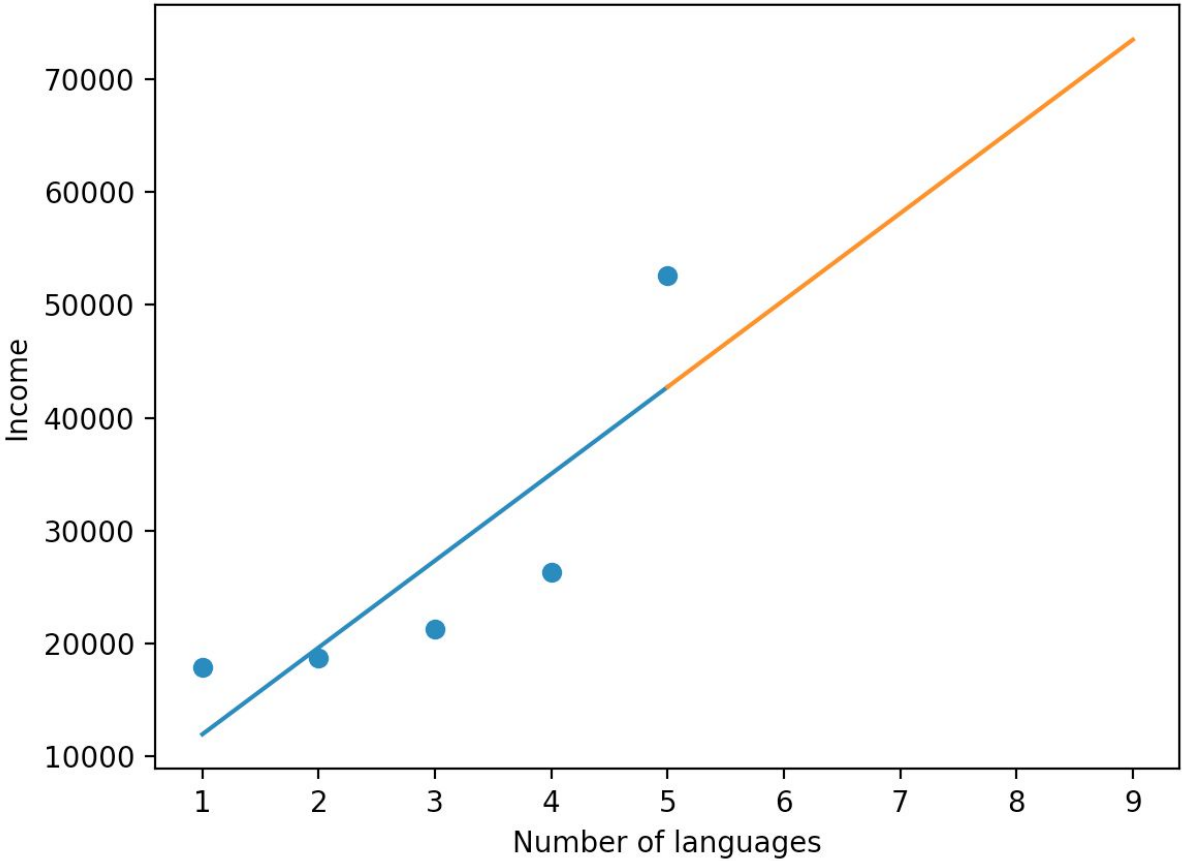


The first thing I tried after a little of exploration of the data was trying to find a linear relationship between income and age, but after plotting the two variables I found out that there may not be a linear relationship between them.

Can we predict the income with the number of languages a person speaks?

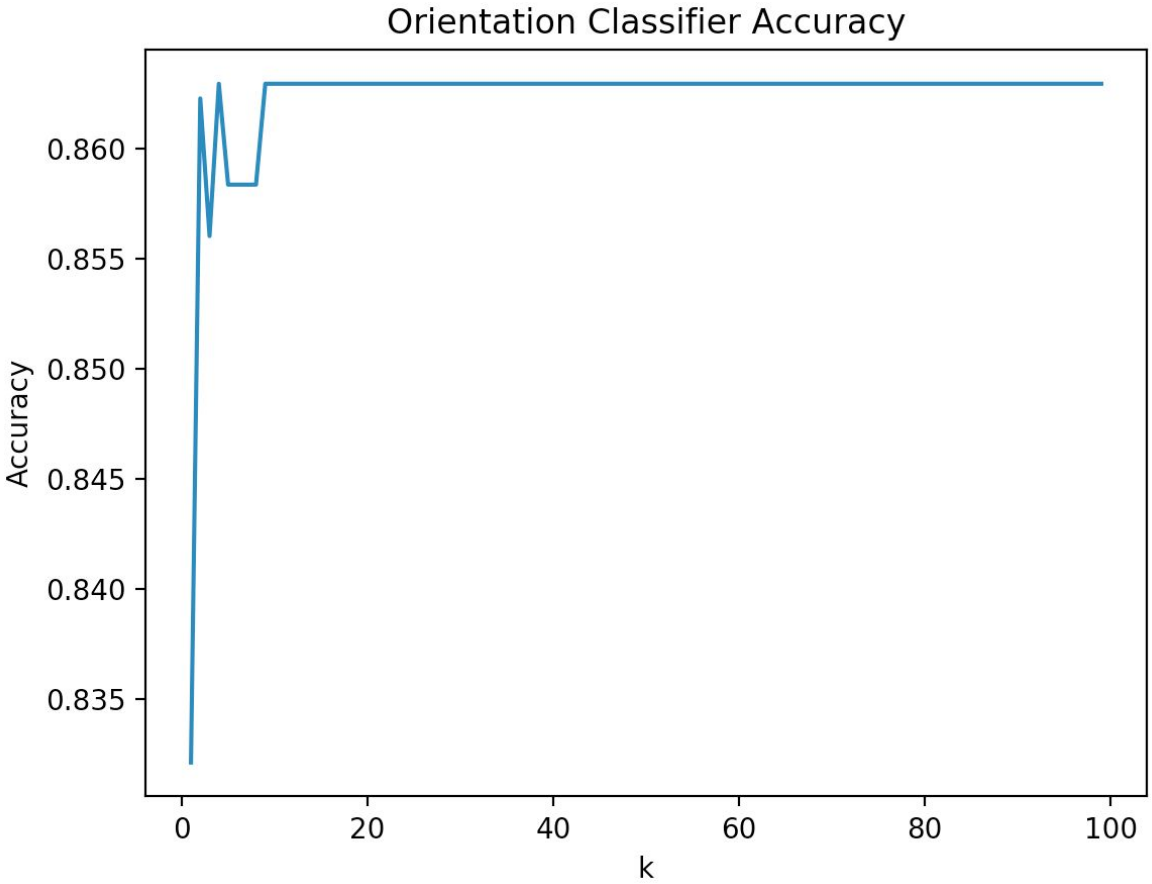


- First a new column was added to store the calculated number of languages a person speaks.
- Then I calculated the means for the number of languages according to the incomes.
- With that data I plotted the graph at the left.

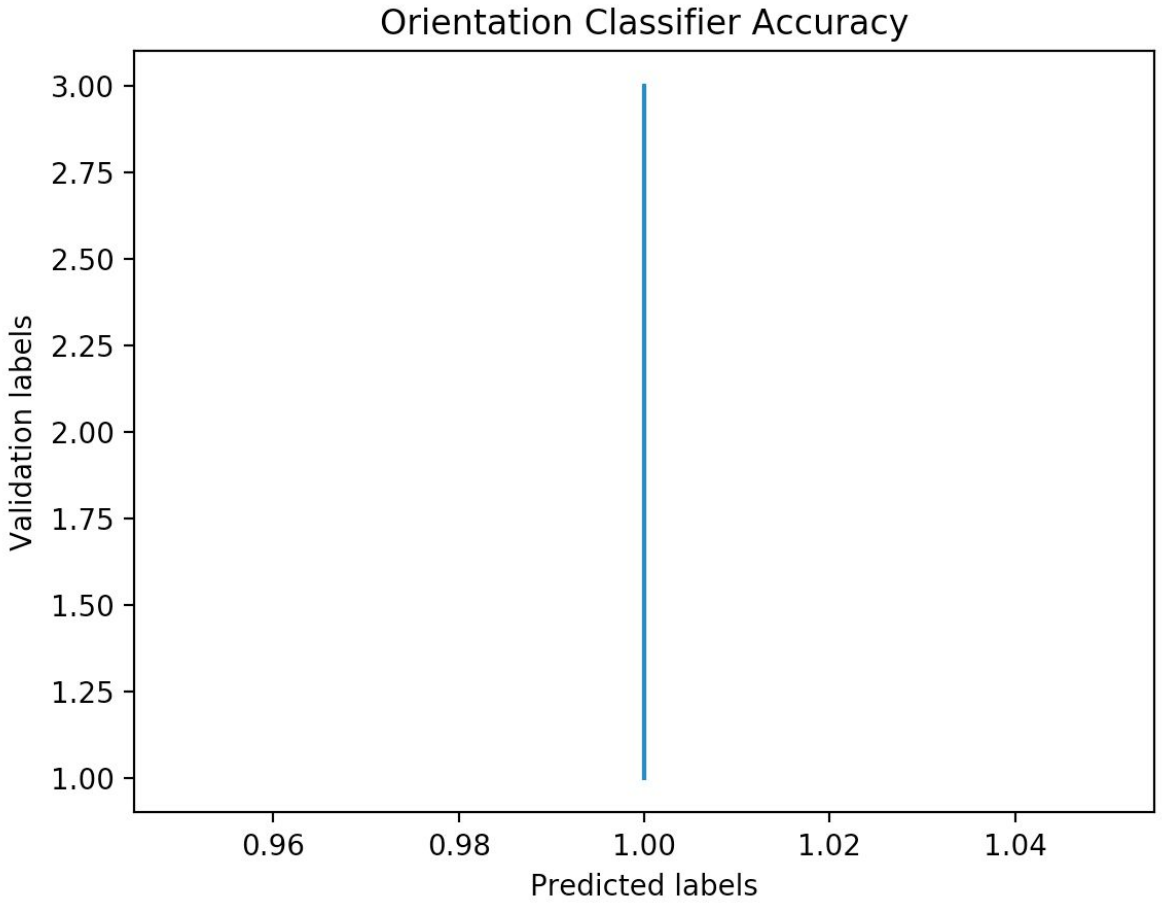


- Applying a linear regression model I was able to plot this graph.
- The time to run for the model was 0.0012 seconds.
- Later I did the same with the KNN regressor, that implementation took 0.00058 seconds of run time.
- Surprisingly the prediction of income for a person who speaks 8 languages with the first implementation was 65792.6293692 and with the second implementation (knn) it was 52550.39590854.

Can we predict orientation with number of languages and income?



- With the data I have so far I tried to answer this question using the KNN algorithm, the accuracy seems to be ok for this graph, but the next one says the opposite.

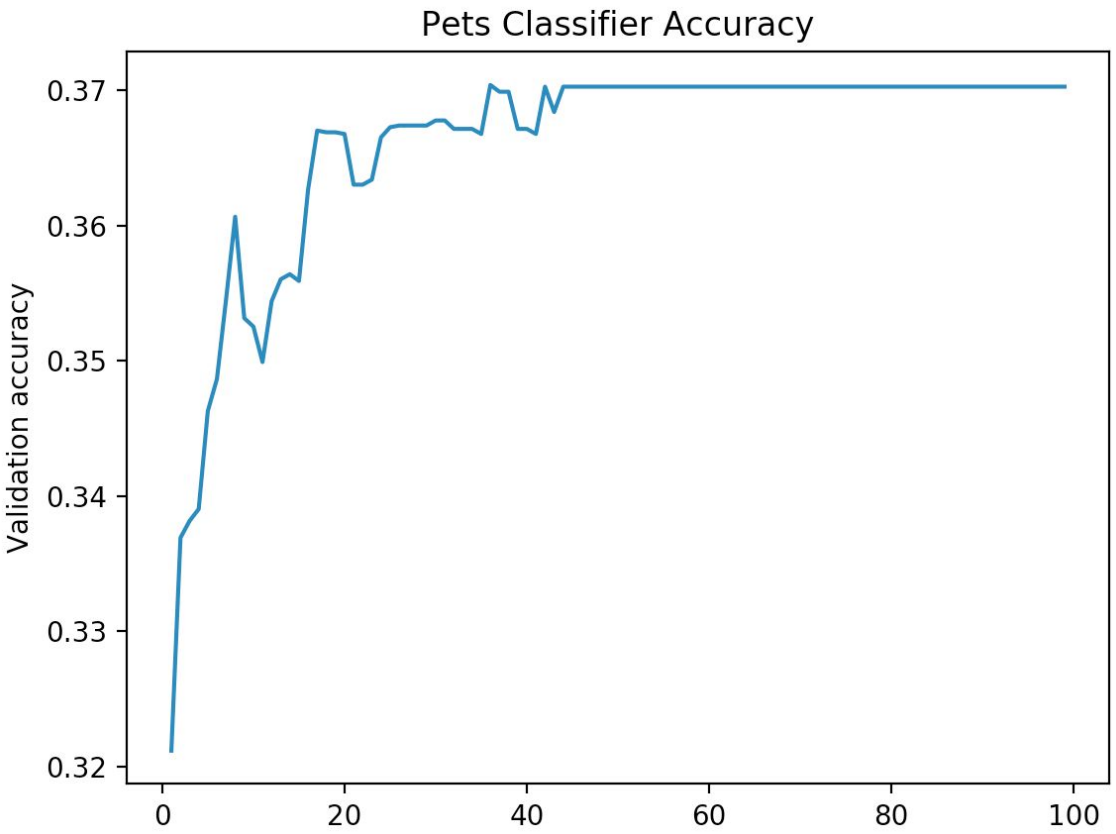


- This one shows Validation labels vs predicted labels and here we realize that something is wrong.
- I noticed something that I should have taken in account from the beginning, value_counts are the following for orientation:

```
straight    51606
gay          5573
bisexual     2767
```

- Here straight=1, gay=2 and bisexual=3, given that straight has way more samples present in the dataset, the model was trained to predict mostly “straight”. What is reflected in the graph.
- Accuracy here was 0.86.

Can we predict pets preference with status and income?



KNN:

- The graph at the left shows the accuray for the classifier using the KNN algorithm, here the best accuracy is for k=36 giving a score of 0.3703.
- The time to run for this model was 1.72 seconds.
- Recall and precision values are the same using the average='micro' param. I'm not sure how to measure those values correctly since the default value for that average param is 'binary' and this is not a binary classifier.

SVM:

- The SVM implementation gives an accuracy score of 0.3710 (not a huge difference).
- The time to run for this model was 11.97 seconds!.
- Same issue as before with recall and precision.

Conclusion

Answering the questions:

- **Can we predict income with the age of a person?**
  - Maybe we can, but there seems not to be a linear relationship between those variables. Am I wrong?
- **Can we predict the income with the number of languages a person speaks?**
  - We can, but perhaps we need more data to make accurate predictions, in this case the maximum number of languages a person speaks is 5, what if actually after that number the income begins to stay similar for people who speaks more than 5 languages?
- **Can we predict orientation with number of languages and income?**
  - We can't do that with the provided dataset since straight people are by far the majority of the data samples, maybe we could drop some of those samples to have equal groups but is that a good idea?
- **Can we predict pets preference with status and income?**
  - Seems like we can't do that with the current accuracy level of the models.

- Would've been great to know how to handle some cases, for example, Is it a good idea to drop data so groups are in the same size (about sexual orientation)?. I also had problems with the recall and accuracy of the models because all the examples for that in the course were for binary classification if I'm not wrong.
- Maybe the results would've been better with a more balanced dataset.
- It would've been interesting to apply unsupervised learning in some way too.