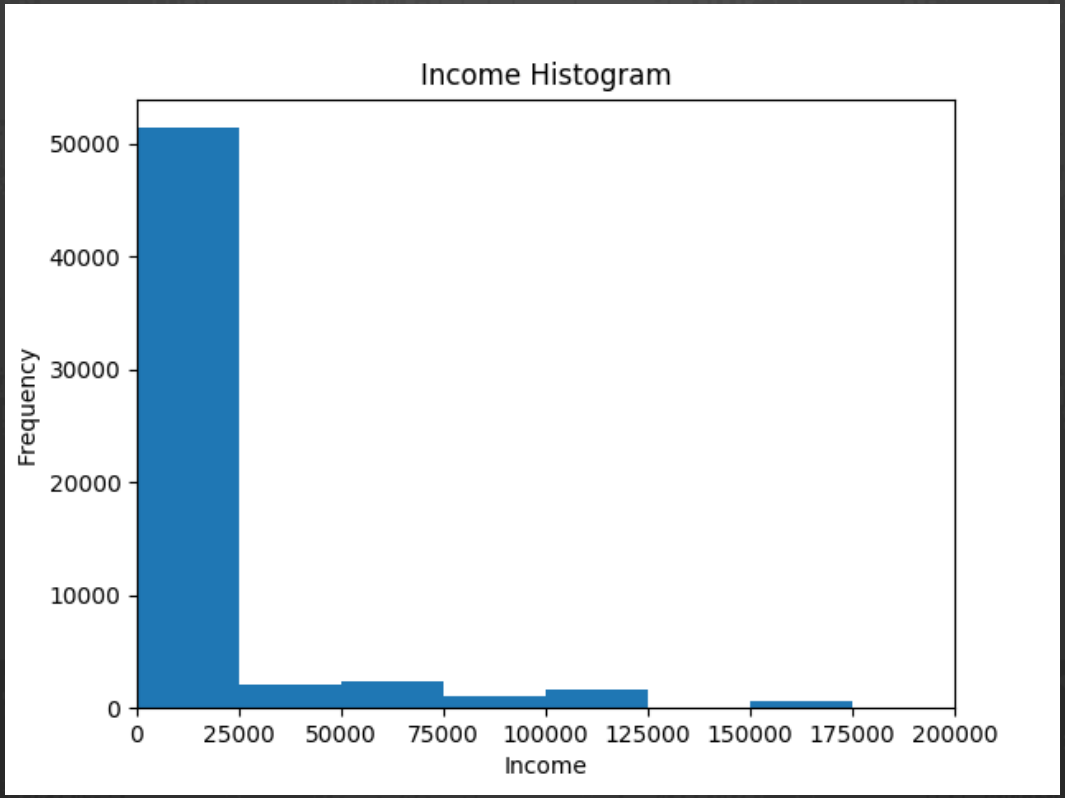
Date-A-Scientist

Machine Learning Fundamentals Capstone

Ben Wallingford
cohort Feb 12, 2019

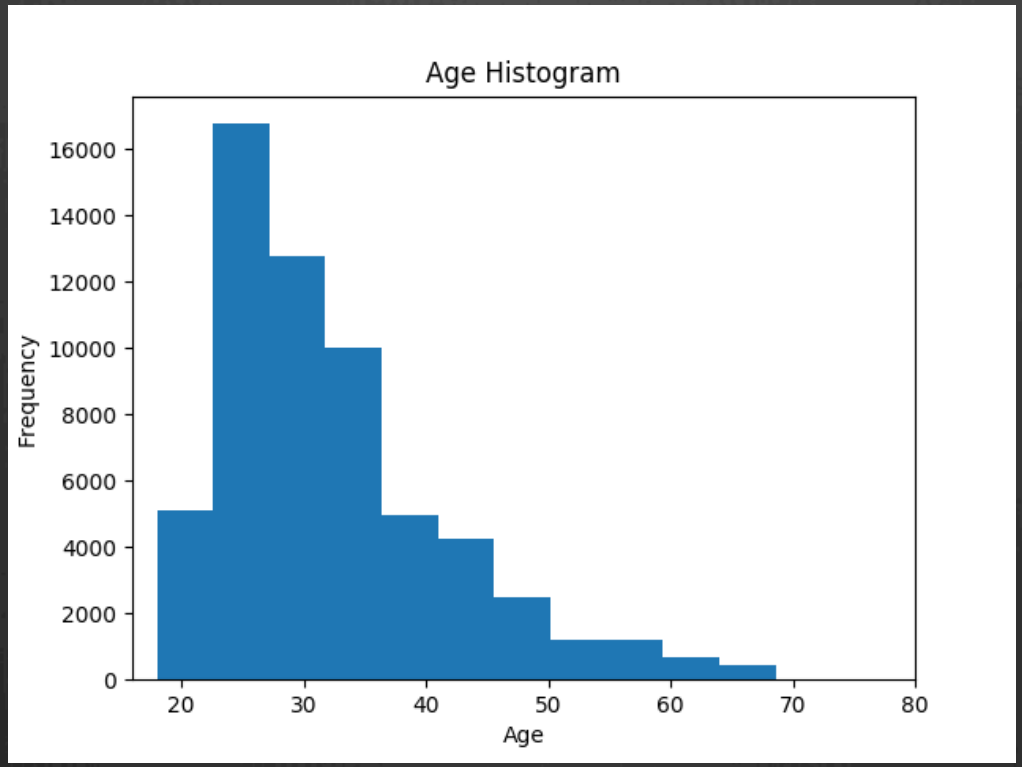
Exploring the Data with Plots

- Shows the frequency of different income levels in the dataset
- Looks like predicting income might be hard since the income is so concentrated under $25k



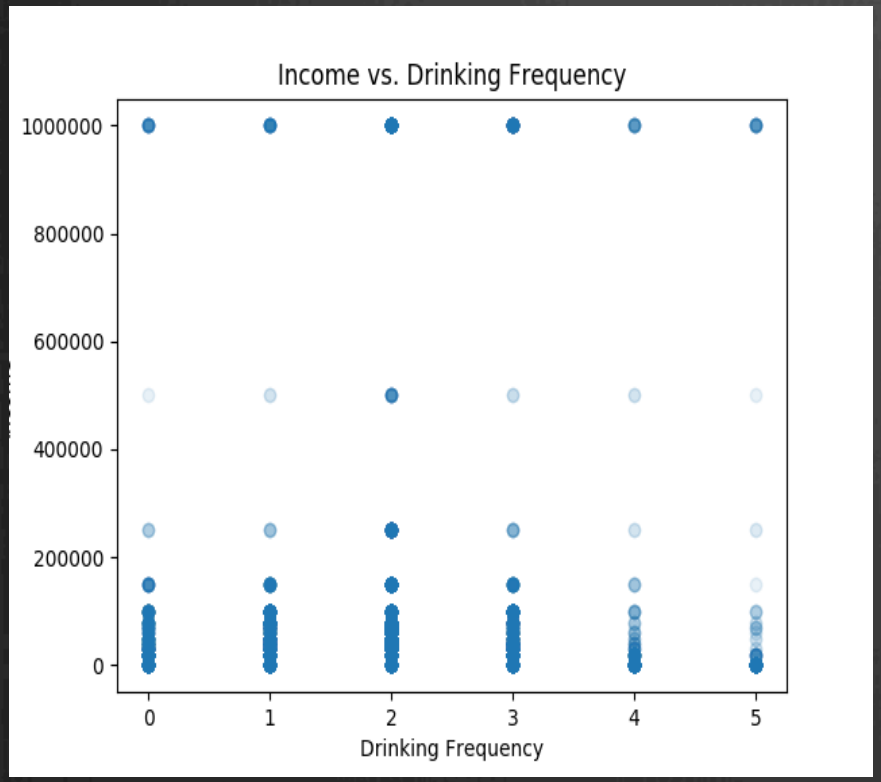
Exploring the Data with Plots

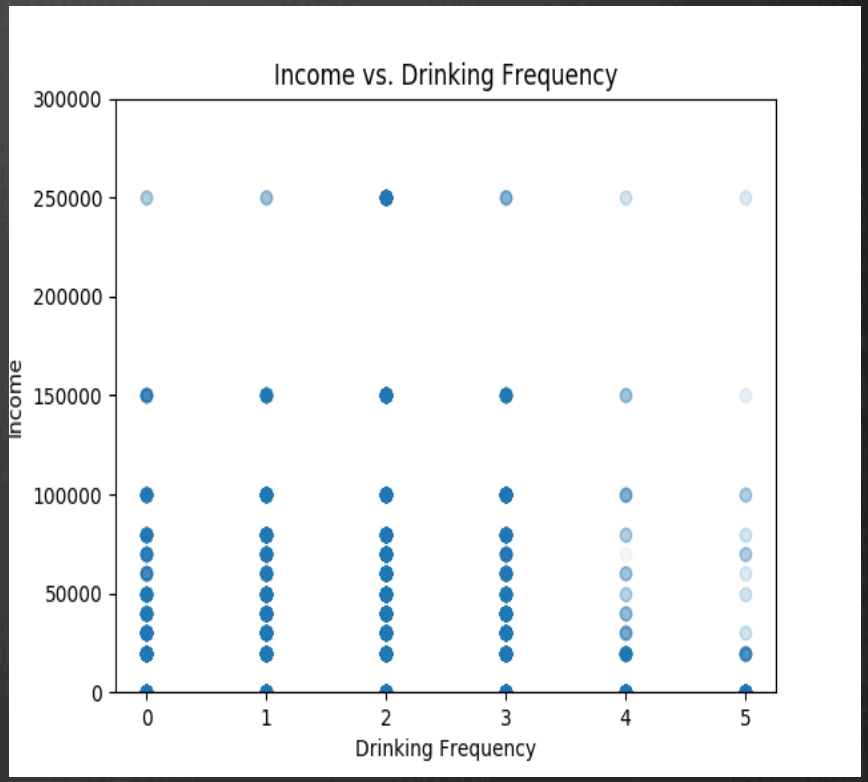
- Shows the frequency of different age levels in the dataset
- Distribution is as expected, since most people on a dating site would be in their 20s or 30s



Exploring the Data with Plots

- Shows the distribution of Drinking Frequency vs. Income level
- Right plot truncates the high income outliers
- We do see fewer frequent drinkers at the higher income levels, but that could just be that there are fewer frequent drinkers overall





Questions for the data

- Can we determine job, based on essay word counts, sex, and body_type?
- Can we tell if someone is a fluent English speaker, based on average essay word length and age?

Columns added to data

- Basic mappings: these either had an obvious order to them, or were binary
  - “drinks_code”
  - “drugs_code”
  - “smokes_code”
  - “sex_code”
- Interpreted mappings:
  - “body_type_code”
    - Had to establish a ranked progression of body types in order to use it as an input to the k-nearest neighbors model
    - Chose to omit “rather not say” responses from mapping
  - “job_code”
    - Didn’t have to have a ranked progression since it was the label and not an inpit for the k-nearest neighbors model
    - Chose to omit “rather not say” responses from mapping

Columns added to data

- New columns from parsing data
  - “state_country”
    - Parse the State of Country they are from, using “.transform()” to select what is after “, “ in “location”
    - Found out the dataset was very skewed towards California
  - “fluent_english”
    - True if they speak English fluently, as determined by these conditions:
      - If they explicitly state that the speak English fluently
      - If they speakEenglish, and don’t state any languages explicitly as fluent
  - “fluent_not_english”
    - True if they speak a language other than English fluently, as determined by the conditions
      - If they explicitly state that they speak a language other than English fluently
      - If they list multiple languages, but don’t explicitly say any of them are fluent
      - If they don’t list that they speak English at all, it is assumed they are fluent in something else

Classification Method Comparison

K-nearest neighbors	Support Vector Machine
Less time to fit/train model	More time to fit/train model
More time to predict	Less time to predict
One tuning parameter (k/number of neighbors to consider)	Two tuning parameters (gamma and C)
	Slightly better performance predicting job from essay length, gender, and body type.

Regression Method Comparison

Multiple Linear Regression	K-Nearest Neighbors Regression
Fewer parameters to tune	Have to at least choose K, but there are many other parameters like algorithm and leaf size in the sklearn interface
Slower to fit/train	Quicker to fit/train since it just builds the point field
Faster to predict since it is a single formula at that point	Slower to predict since it has to find K neighbors through distance calculations
Performed better (higher R^2 vaue) when predicting English fluency from essay average word length and age.	

Conclusion

- Can we determine job, based on essay word counts, sex, and body_type?
  - Using K-nearest neighbors, we were able to predict their job with an accuracy of 15.6%, and it was only slightly higher for SVM (16.0%)
- Can we tell if someone is a fluent English speaker, based on average essay word length and age?
  - Using Multiple Linear Regression, the R^2 was -0.00045, and with K-nearest neighbors regression, the R^2 was -0.189.

Unfortunately, none of the models performed particularly well for this project, so these models do not support us being able to answer these questions with this data.

The data had some very strong clustering with many of the fields (such as incomes below $25K and most of the people being from California). Moving forward I would want to make sure that I had a better understanding of the skews in the input data, and how to account for them in the models. Also, some of the fields had some distracting choices, like "space camp" for education "C++" for language. I could see how these might be useful to correlate on for predicting compatibility for a dating site, but they made it harder for some of the more general demographic predictions that we were trying to do.

It would also be interesting to go into more of the word pattern analysis that we did with a Bayes classifier in one of the lessons. It appears that the content of the essays is probably one of the most interesting fields to look at more in this data since it would be diverse and not confined to a list of options.