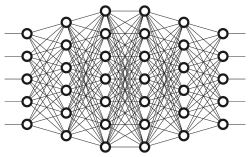


Massachusetts Institute of Technology
Dept. of Electrical Engineering and Computer Science
Fall Semester, 2018
MIT 6.S198: Deep Learning Practicum

Final Project Workshop: How to Mine the Interwebs for Data

1: Working with pre-compiled datasets

1.1 Installing Kaggle

1.2 Downloading and Viewing Data

What about the structure of this dataset might make it harder to feed into a machine learning model? (Hint: is all the information in users.csv useful?)

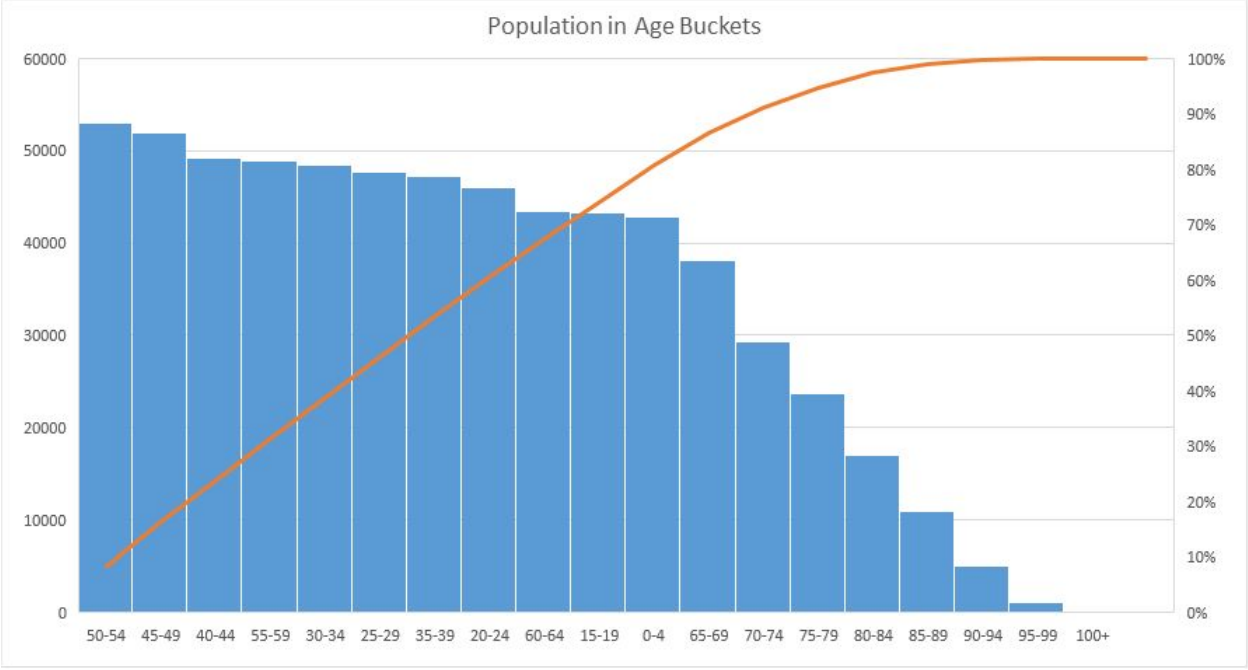
- A lot the information is not very useful such as signup flow

Is this dataset a good fit for the task at hand? What problems could you think of that might make this dataset not useful?

- This data set has a lot of missing values, especially in key areas like gender that make it hard to use this data set comprehensively.

1.3 Data Processing and Kaggle kernels

Using the kernel provided as a starting point, build another data visualization or compute another statistic that might be relevant for the dataset. Take a picture of the code that you write, and of any outputs and visualizations that you create.



1.4 Datasets Everywhere

2: Mining from websites using scripts

2.1 Installation

3: Gathering data in the real world

4: Brainstorm data collection strategy for your project

Brainstorm data collection for your project and summarize your strategy for the writeup. 1-2 paragraphs is sufficient. This will help you for the Data Review assignment due on 10/17.

We are developing a bot that can play the game Foldit. For this work, we want the historical playthrough data of the game and the scores associated with each. We have obtained the data from Professor Seth Cooper at northeastern in a format called pdb that is readable by biopython.

Apologies for the delayed submission, I did not realize we had to submit anything from the workshop.

5: Submission

Create a page for the Data Gathering Workshop on your class homework submission site. Include your name and email address and the required writeups as indicated above together with code and images, as appropriate.

Use this form to hand in the Data Gathering Workshop
Due 10AM Friday, 9/28