CSE 4/587

Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Data Cleaning and EDA Demo

Recap from Last Class

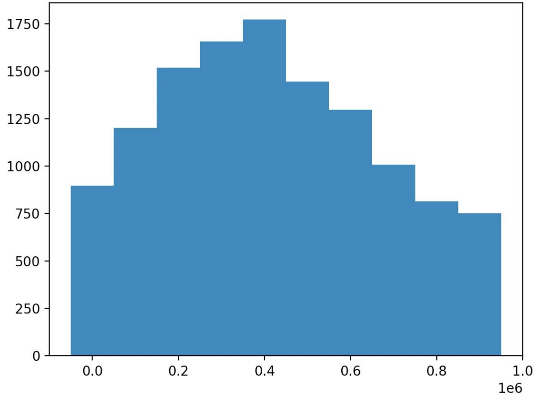
- Exploratory Data Analysis (EDA)
  - Get intuition about the nature of your data
  - Gather some basic stats/visualizations: min, max, mean, histograms, etc
  - Can be used to form some initial hypotheses
- Related to data cleaning, and feature extraction
  - We'll explore these two a bit more today

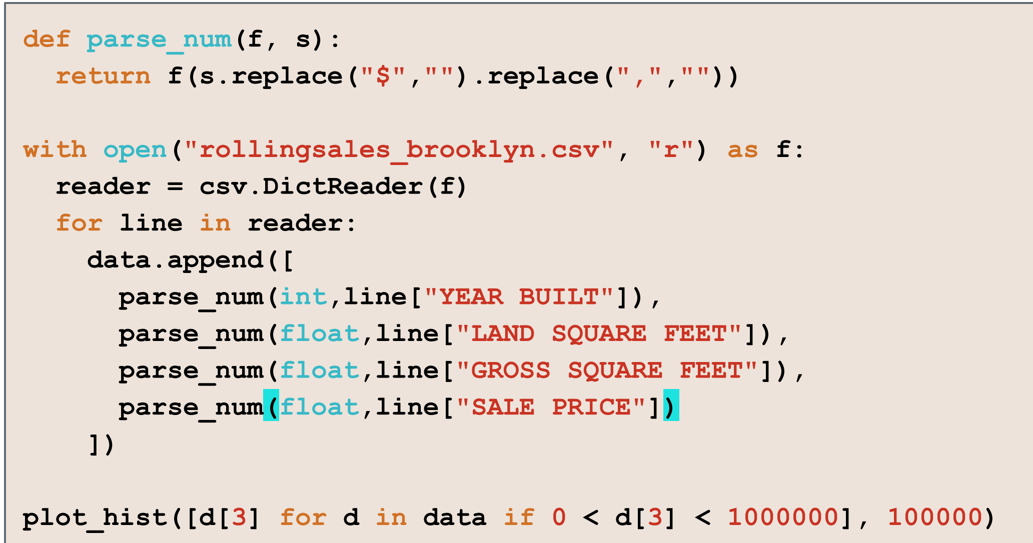
Data Cleaning and Munging

- Real-world data is almost always going to be *dirty*
    - Data will be missing/incomplete
    - Entries may contain errors
    - Entries may not be in the proper format
- Initial cleaning of the data will make the rest of the process smoother
    - Issues like formatting can often be dealt with immediately
    - Finding errors in the data may require EDA
    - EDA may reveal further cleaning that is required

Data Cleaning and Munging

- Examples (Ch 2 DDS, Ch 10 DSfS)
  - Clean up formatting for numbers
  - Remove nonsensical data (ie: sale prices of $0)
  - Check for outliers
  - Extract columns we want



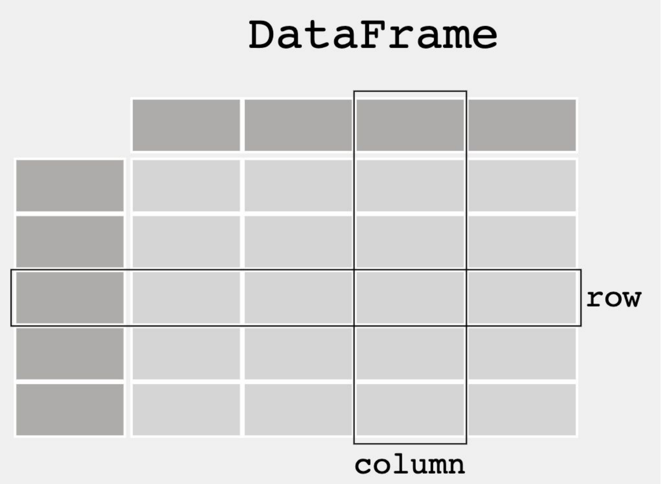


```
def parse_num(f, s):
  return f(s.replace("$","").replace(",",""))

with open("rollingsales_brooklyn.csv", "r") as f:
  reader = csv.DictReader(f)
  for line in reader:
    data.append([
      parse_num(int,line["YEAR BUILT"]),
      parse_num(float,line["LAND SQUARE FEET"]),
      parse_num(float,line["GROSS SQUARE FEET"]),
      parse_num(float,line["SALE PRICE"])
    ])

plot_hist([d[3] for d in data if 0 < d[3] < 1000000], 100000)
```

Data Cleaning and Munging with Pandas

Pandas provides an easy to use data structures and tools for dealing with structured data

- Stores data in a DataFrame made up of rows and columns
- Data can be read from many common formats like csv
- Provides a rich set of operations for exploring, filtering, combining data, etc
- Integrated with matplotlib for quick and easy plotting



[1] https://pandas.pydata.org/docs/getting_started/index.html#getting-started