1 Opening

- pandas
- as spreadsheet
- relational database
- application

2 pandas.ipynb part

2.1 Numpy

- Python. duck typing. if something looks like sequence and acts like a sequence. iteration over, length
- numpy array was not designed. numpy functions implemented in C.
- base Python is not well designed for data science
- floats by default unless otherwise told
- numpy array 8-byte cells adjacent
- numpy is vectorized.
- adding a list to a list will concatenate in base Python. numpy is vectorized elementwise numerical summation.
- Data Science from Scratch
- Joel Grus - Livecoding Madness - Let's Build a Deep Learning Library
- "*" is elementwise, np.dot is matrix multiplication
- Python list is heterogeneous. just list of pointers.
- base Python iteration on a numpy array element. overhead of unboxing and boxing.
- garbage collected language. if reference count is zero, memory is freed
- do not work elementwise. if iterating may help to convert to a base Python list up front

2.2 Pandas

- DataFrame: index rows 0,1,.., columns have meaning
- rename column names.
- Hadoop Spark. parallelizing over distributed system. change should happen by copying not by mutating. one failed machine means one part must be repeated (existing structure should not have changed.). if mutating, the entire thing must be run again.
- variables are postit in Python.
- for in-place mutation use inplace=True
- pandas DF is a numpy array with an overall meta-data (not cell level)
- single column (series) extraction with .colname
- len(df) number of rows
- iteration over a df is columnwise operation like R
- Missing is NaN in pandas.
- itertuples to iterate over rows. but usually a bad idea.
- .iloc[0:3] is index location. first three rows
- .loc[0:7] row names 0,. . . ,7 (labels not location index)