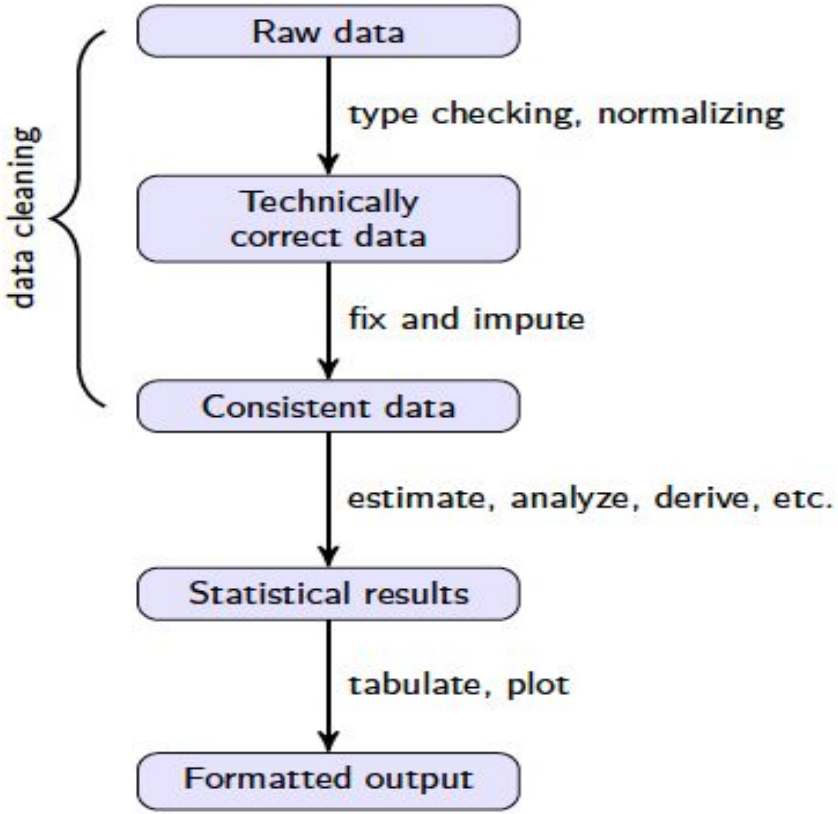
Topic of the day: Manipulation

Source

Cornell Data Science

Data Manipulation

The Data Pipeline

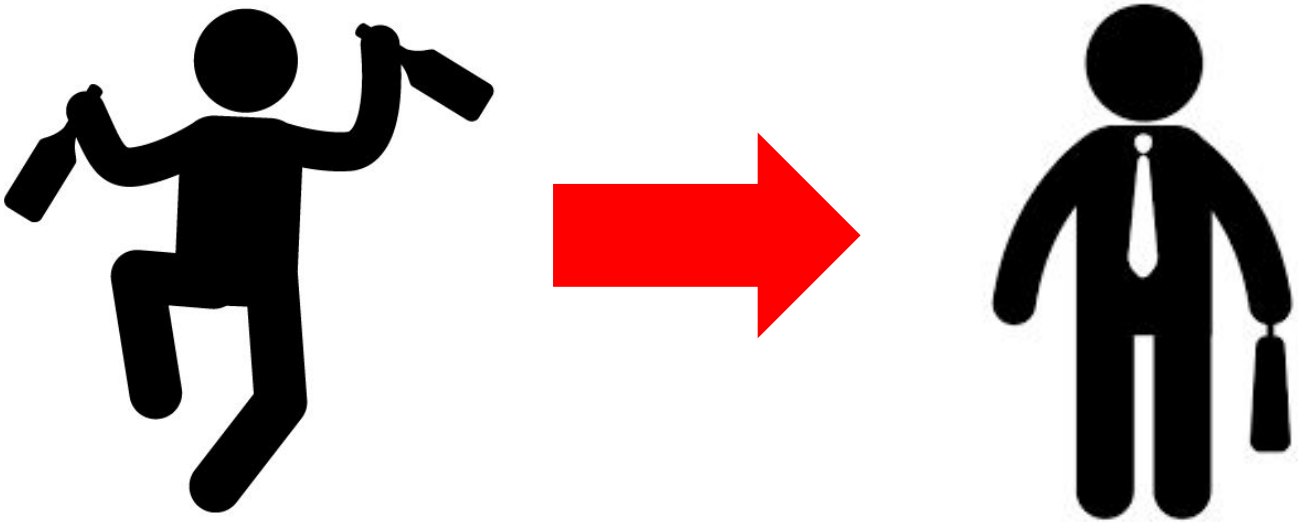


Source

Question:

What are some ways in which data can be “messy”?

Drunken Datasets Out There



Why Do We Manipulate

Ready-to-read format

Not run into error when we perform a calculation

What other reasons can you think of?

$$ Golden Rules of Writing Fast Python $$

Use NumPy functions whenever applicable

Vectorize your operations as much as possible

Why Does NumPy Matter?

NumPy is written in C, which is much faster than Python

C stores data in a contiguous buffer instead of multiple pointers

NumPy arrays are homogeneous and statically typed

Numpy Array Creation

```
>> import numpy as np

>> an_array = np.array([1, 2, 3, 4])

>> str_array = np.array([‘cat’, ‘dog’, ‘bird’])

>> bool_array = np.array([True, True, False, True])

>> rank2_array = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
```

Creating an ndarray with prefilled values

numpy.**zeros( )**

numpy.**eye( )**

numpy.**full( )**

numpy.**ones( )**

numpy.**random.random( )**

Array Operations

Operations


```
>> a + b   # same as np.add(a, b)
```



```
>> a - b   # same as np.subtract(a, b)
```



```
>> a * b   # same as np.multiply(a, b)
```

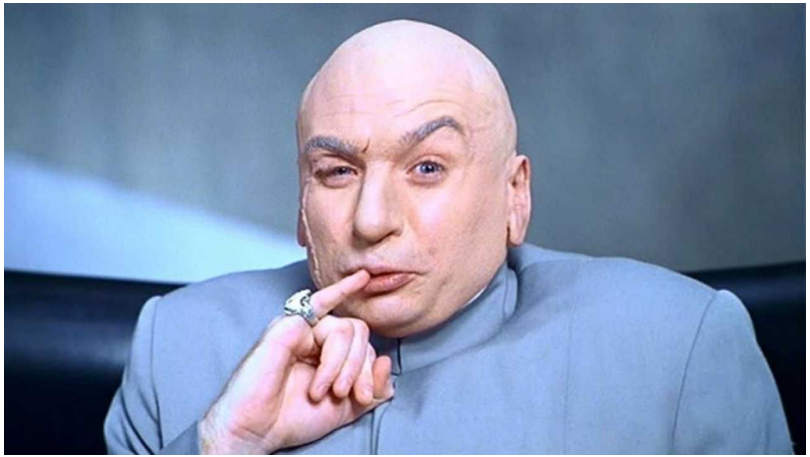
```
>> np.sqrt(a)
```

...

And more!

Data Manipulation Tools

1. Cutting down size
2. Gathering relevant data
3. Transform data
4. Gather info on data



Source

Summarizing

What it does

Gives a general idea about the dataset

Why?

To understand and explore the dataset!

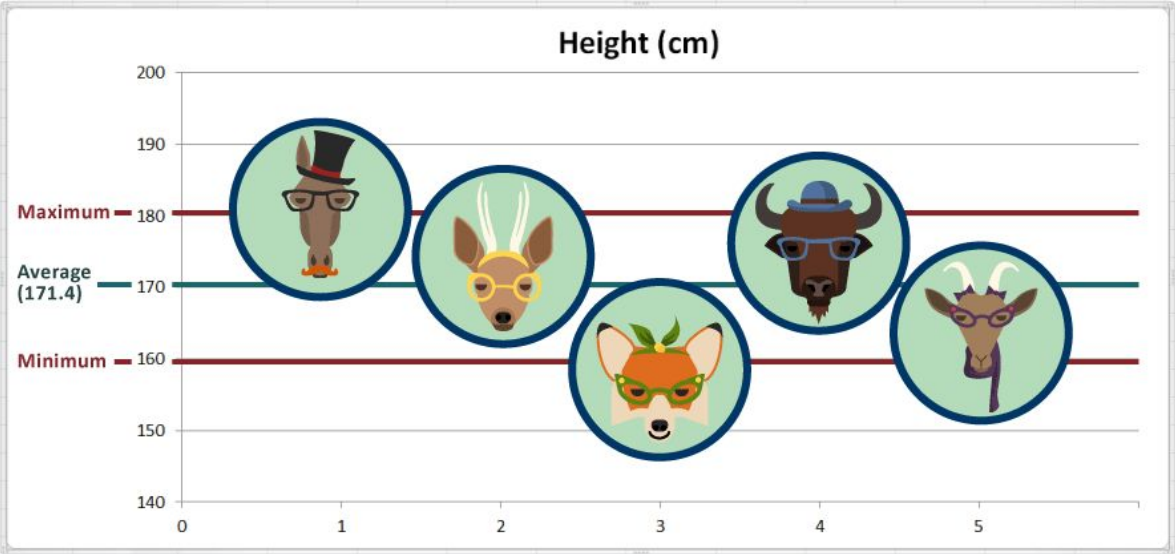


Image Credit:
https://www.openelectiondata.net/images/academy/a-gentle-introduction-to-summarizing-data

Statistical Methods

mean( )

```
>> an_array.mean(axis=1) # computes means for each row
```

median( )

```
>> an_array.median()
```

sum( )

```
>> an_array.sum(axis=0) # computes sum of each column
```

Filtering and Subsetting

What it does	Grab a subset in a data frame with a condition. **Filtering** grabs rows and **subsetting** grabs columns.
Why?	Decreasing data size or examining subgroups closer

Name	Age	Major
Amit	19	Computer Science
Dae Won	24	ORIE
Chase	19	Information Science
Jared	19	Computer Science

Filtering

Name	Age	Major
Amit	19	Computer Science
Dae Won	24	ORIE
Chase	19	Information Science
Jared	19	Computer Science

Subsetting

Sorting and Set Operations

sort( )	unique( )	Set Ops
Return a sorted copy of an array	Return an array with duplicate values removed	intersect1d union1d setdiff1d in1d

Combining

What it does

Joins together two data frames, either row-wise (horizontally) or column-wise (vertically)

Name	Age	Major
Amit	19	Computer Science
Dae Won	24	ORIE

Name	Age	Major
Jared	19	Computer Science
Kenta	20	Computer Science

Name	Age	Major
Amit	19	Computer Science
Dae Won	24	ORIE
Jared	19	Computer Science
Kenta	20	Computer Science

Combining (continued)

	Name
0	Amit
1	Dae Won
2	Chase
3	Jared
4	Kenta

	Age	Major
0	19	Computer Science
1	24	ORIE
2	19	Information Science

	Name	Age	Major
0	Amit	19	Computer Science
1	Dae Won	24	ORIE
2	Chase	19	Information Science
3	Jared	NaN	NaN
4	Kenta	NaN	NaN

Joining

What it does	Joins together two data frames, combining rows that have the same value for a column
How to do it	Pandas has **join** and **merge** functions. When we use **merge**, we want to set a column to *key* on, using *on=('key_name')*

But why would we get a dataset in pieces?

Name	Major	Age	Computer	Purchased
Dae Won	ORIE	31	Linux <3	HaPpy ProPro Server 9Ghz
Dae Won	ORIE	31	Linux <3	HaPpy ProPro Server 9Ghz
Dae Won	ORIE	31	Linux <3	RealX High Perf Monitor
Dae Won	ORIE	31	Linux <3	48GB H4rdOn RAM
Jared	CS	19	Mac </3	Big Book of Trivia
Jared	CS	19	Mac </3	“Help I don’t know fun facts” - A Life Story
Jared	CS	19	Mac </3	10,000 Facts to Impress Your Friends
Dae Two	ORiE	31	Linux <3	Leather Riding Crop

This is wasteful…

But why would we get a dataset in pieces?

ID	Name	Major	Age	Computer
0001	Dae Won	ORIE	31	Linux <3
0002	Jared	CS	19	Mac </3

ID	Purchased
0001	HaPpy ProPro Server 9Ghz
0001	HaPpy ProPro Server 9Ghz
0001	RealX High Perf Monitor
0001	48GB H4rdOn RAM
0002	Big Book of Trivia
0002	“I don’t know fun facts - My Life Story”
0002	10,000 Facts to Impress Your Friends
0001	Leather Riding Crop

There’s a lot less redundant data!

A Join in Action

Pick a Feature to "Key" on → Rows that share a value in the key column will be merged → (Optional) Filter the Resulting Table

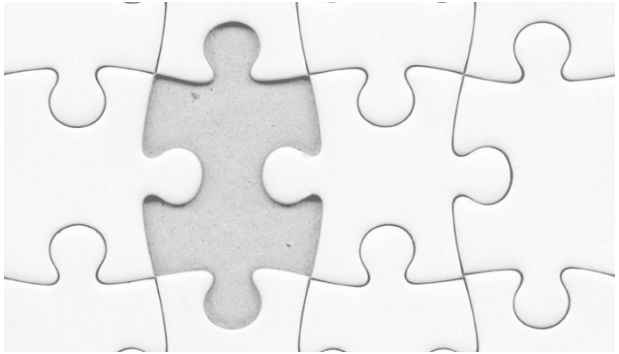
ID	Name	Major	Age	Computer	Purchased
0001	Jared	CS	19	Mac </3	Big Book of Trivia
0001	Jared	CS	19	Mac </3	""I don't know fun facts - My Life Story"
0001	Jared	CS	19	Mac </3	10,000 Facts to Impress Your Friends

Dealing with Missing Data

Datasets are usually incomplete. We can handle this by:

Leaving out missing samples

Replacing missing variables with a mean or a median. This is called **imputation**



Techniques for Data Manipulation

Formatting the shape of our data

Changing the actual content of the data

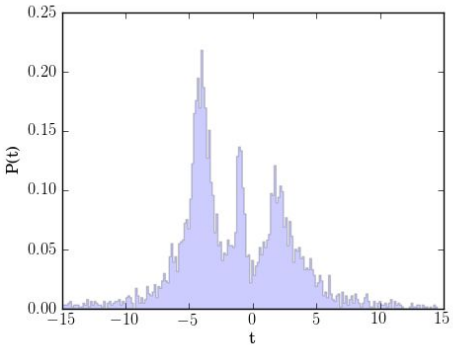
Technique: Binning

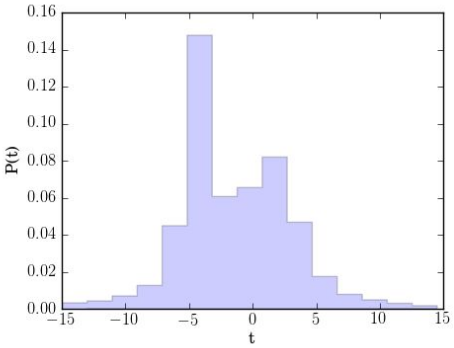
What it does

Makes continuous data categorical by lumping ranges of data into discrete "levels"

Why?

Applicable to problems like (third-degree) price discrimination



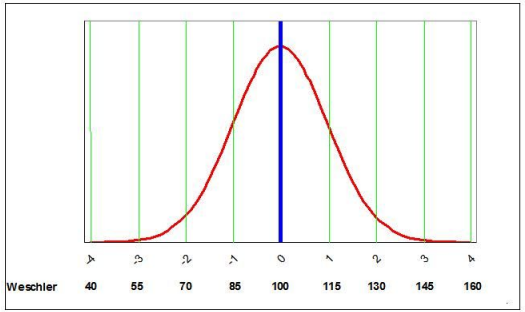


Source

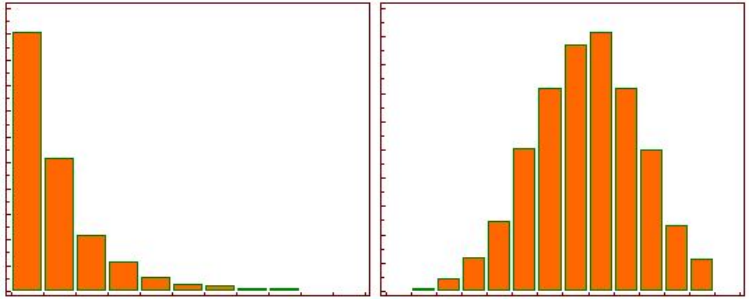
Technique: Normalizing

What it does	Turns the data into a bell curve (Gaussian) shape by standard, log, or another transformation
Why use it	Meet model assumptions of normal data; act as a benchmark since the majority of data is normal; wreck GPAs

Standardizing



Log transformation



Others include square root, cubic root, reciprocal, square, cube...

Source Source

Technique: Ordering

What it does	Why?	Example
Converts categorical data that is inherently ordered into a numerical scale	Numerical inputs often facilitate analysis	January → 1 February → 2 March → 3 …

Technique: Dummy Variables

What it does

Creates a binary variable for each category in a categorical variable

plant	is a tree
aspen	1
poison ivy	0
grass	0
oak	1
corn	0

Technique: Feature Engineering

What it does	Generates new features which may provide additional information to the user and to the model
How to do it	You may add new columns of your own design using the assign function in pandas

tab ->

ID	Num
0001	2
0002	4
0003	6

ID	Num	Half	SQ
0001	2	1	4
0002	4	2	16
0003	6	3	36

```
tab.assign(SQ=arr['Num']**2, Half=0.5 * arr['Num'])
```

Coming Up

Your assignment: Quiz 2

Next week: LECTURE 3 - Data Visualization

See you then!

Cornell Data Science