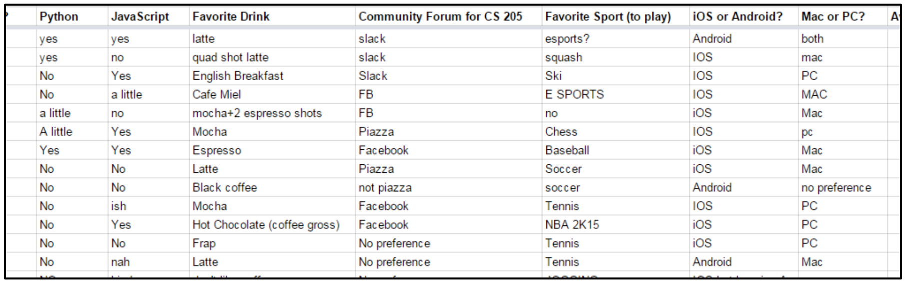
CS 205

Lecture Handout #2
January 21, 2016

Data Cleaning

One of the biggest problems in data is "dirty data"; we will face it many times through CS 205. It is already present in our data:



Python	JavaScript	Favorite Drink	Community Forum for CS 205	Favorite Sport (to play)	iOS or Android?	Mac or PC?
yes	yes	latte	slack	esports?	Android	both
yes	no	quad shot latte	slack	squash	IOS	mac
No	Yes	English Breakfast	Slack	Ski	IOS	PC
No	a little	Cafe Miel	FB	E SPORTS	IOS	MAC
a little	no	mocha+2 espresso shots	FB	no	iOS	Mac
A little	Yes	Mocha	Piazza	Chess	IOS	pc
Yes	Yes	Espresso	Facebook	Baseball	iOS	Mac
No	No	Latte	Piazza	Soccer	iOS	Mac
No	No	Black coffee	not piazza	soccer	Android	no preference
No	ish	Mocha	Facebook	Tennis	IOS	PC
No	Yes	Hot Chocolate (coffee gross)	Facebook	NBA 2K15	iOS	PC
No	No	Frap	No preference	Tennis	iOS	PC
No	nah	Latte	No preference	Tennis	Android	Mac

What makes this data dirty?

What are three things we can do to fix it?

1.

2.

3.

Frequency

One of the most basic things we will do with data is to find the frequency (occurrence) of something:

- How often is a latte someone's favorite drink?
- How often is someone's favorite drink not coffee-based?
- How many iOS users use a PC?

Python: Reading CSV files

On Tuesday, we attempted to read a CSV file:

In Python, we can import libraries to give access to specific functions. In order to read the CSV file, we need to import the CSV library:

```
In [2]  import csv
```

Next, we need to read the file:

```
In [3]  f = open("cs205 - Data.csv")
```

The csv library allows us to read the entire CSV file in as a dictionary to easily access later:

```
In [4]  reader = csv.DictReader(f)
```

Finally, use the following pattern to print out the names of everyone in the class:

```
In [5]  for row in reader:
          print( row["Name"] )
```

What happened?

-
-

Fix #1:

```
In [6]  f = open("cs205 - Data.csv")
        reader = csv.DictReader(f)
        for row in reader:
          print( row["Name"] )
```

What does this do differently?

Could we come up with something easier to use for small data sets?

Python: Finding Frequencies

In order to answer the question: *"How often is a latte someone's favorite drink?"*, what logic is required?

How could you do this in code?

```
In [ ]   for row in data:



         print(         )
```

Python: Finding Similarities

Who is the most similar person in this room to you? One way to find this is to answer the question: *"Who answers the most questions the exact same way that I did?"*.

…*w*hat logic is required?

Step 1: ____

```
In [ ]   for row in data:



         print(         )
```

Step 2: ____

```
In [ ]   for row in data:
```

How similar is "latte" and "coffee" and "water"?

Up until now, we have considered every different answer to be distinct. However, that is not really true. On a scale of [0, 1], how similar are different coffee choices?

	Coffee	**Latte**	**Mocha**	**Frap**	**Tea**	**Water**
Coffee	1.0	▨	▨	▨	▨	▨
Latte		1.0	▨	▨	▨	▨
Mocha			1.0	▨	▨	▨
Frap				1.0	▨	▨
Tea					1.0	▨
Water						1.0

Before our next class…

1. Continue to develop Python skills by completing **the first lessons (~20 parts)** in the following on codecademy.com:
    - **Units 2-4**: "Strings and Console Output", "Conditionals and Control Flow", and "Functions"

2. Find a CSV data set related to your major, read it in on Python, and find at least one frequency or similarity between records. Let us know what you did by writing up a half-page on what data you found, what frequency/similarity you found, and any struggles you had in doing the assignment.
    - Print it out; bring it to class on Tuesday.