# Research Design and *pandas*

*Ivan Corneillet*

*Data Scientist*

# Learning Objectives

After this lesson, you should be able to:

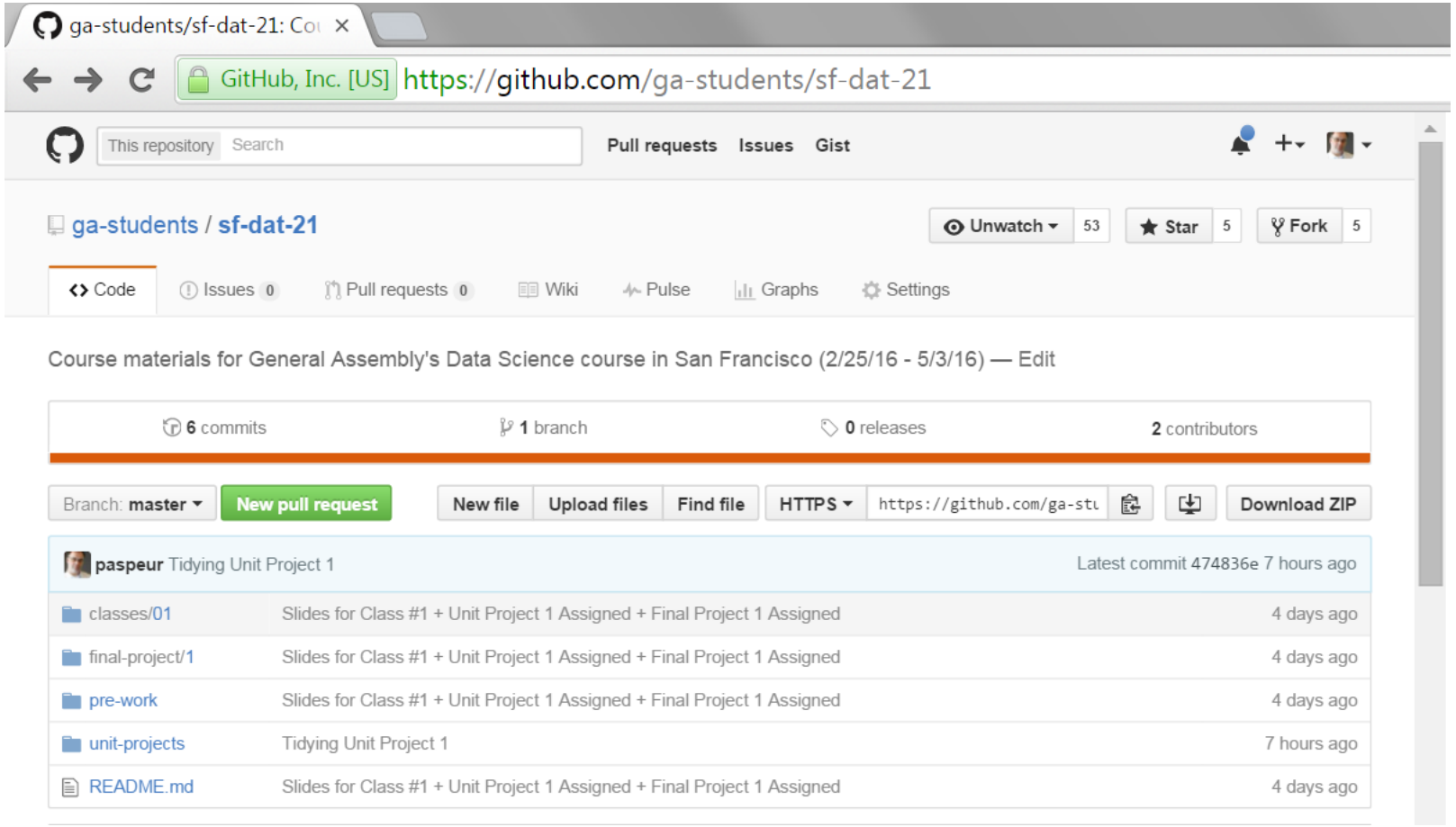
- Define a problem and types of data
- Identify dataset types
- Define the data science workflow
- Apply the data science workflow in the pandas context
- Write an iPython notebook to import, format, and clean data using the *pandas* library

# Outline

- Course GitHub Repository
- Pre-Work
- Data Science Workflow Review
- Identify the problem
  - The Why's and How's of a Good Question
  - The SMART Goals Framework
- Acquire the Data
  - Data Types
  - Logistics of Acquiring Data
  - Our "Zillow" Dataset
  - Tidying Data
  - File Formats
- Parse the Data
  - Documentation and Data Dictionaries
  - Codealong – Introduction to *pandas*
  - Codealong – Tidying up the Zillow dataset
  - Lab
- Review
- **Unit Project 1 (due next session on 3/3)**


# Course GitHub Repository

# SF-DAT-21's Web Interface




# Pre-Work

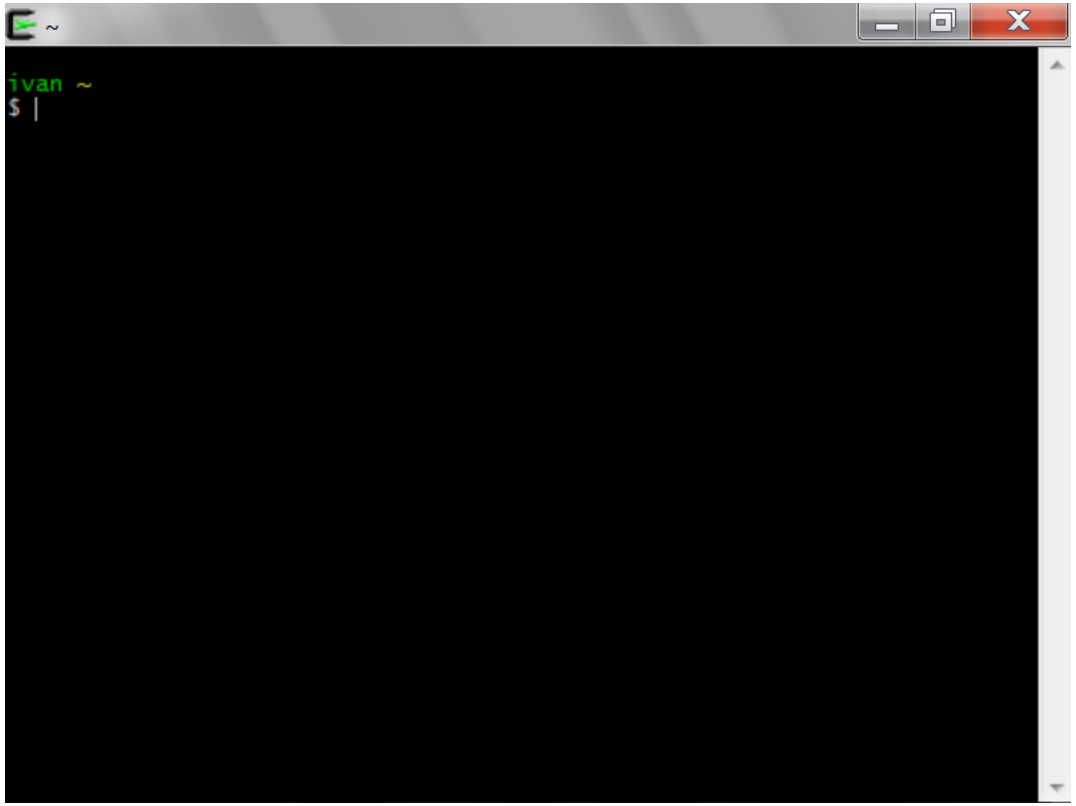
# Pre-Work

Before this lesson, you should already be able to:

- Have completed Python pre-work
- Open and create iPython notebooks

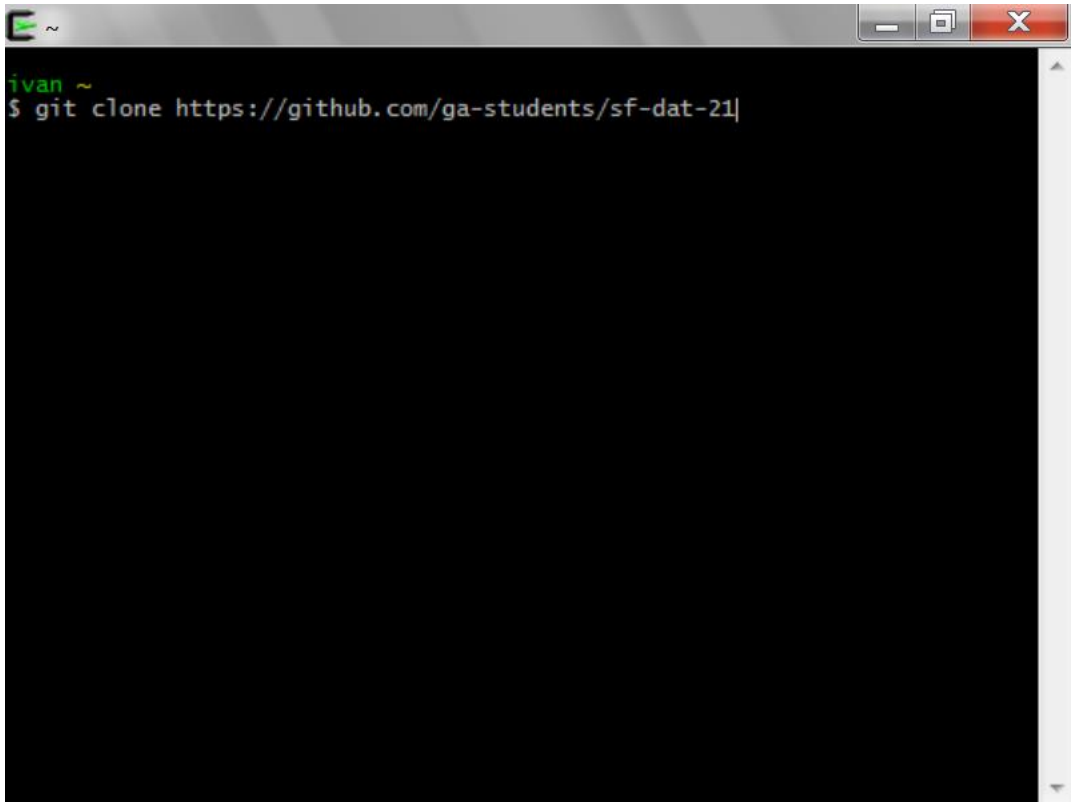

# Getting a local copy of the repo via the command line

# First, get a command line prompt

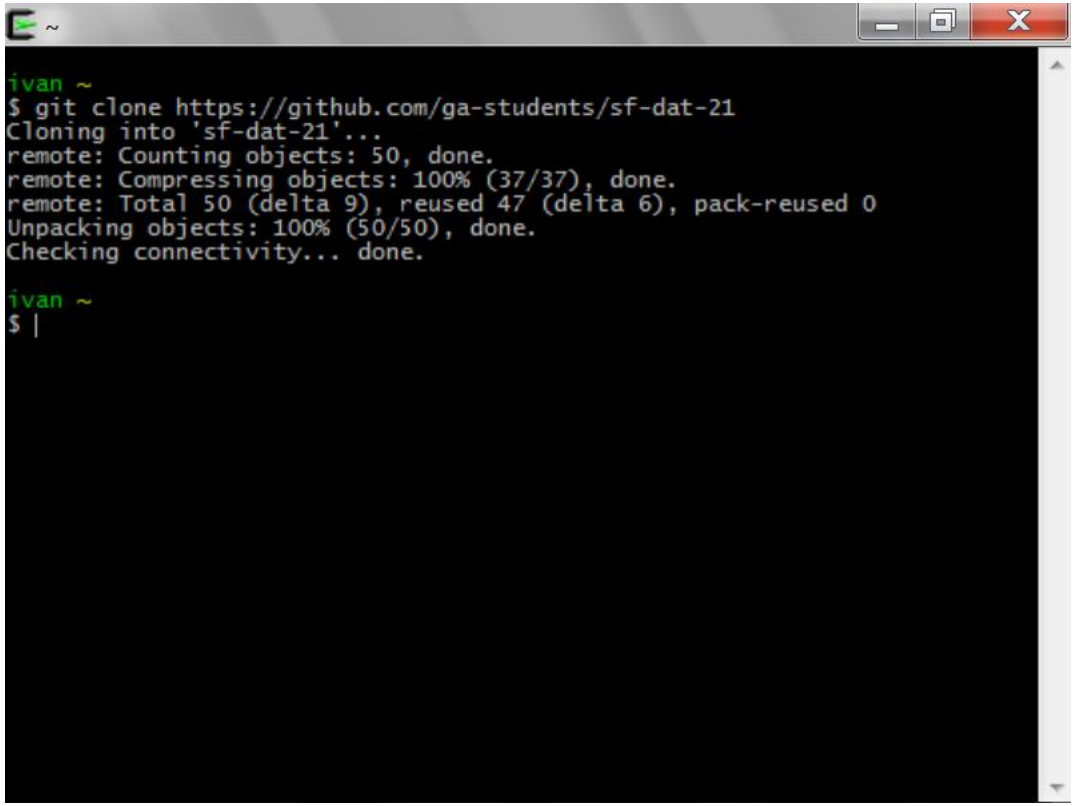


Then, clone the repo with the following command:

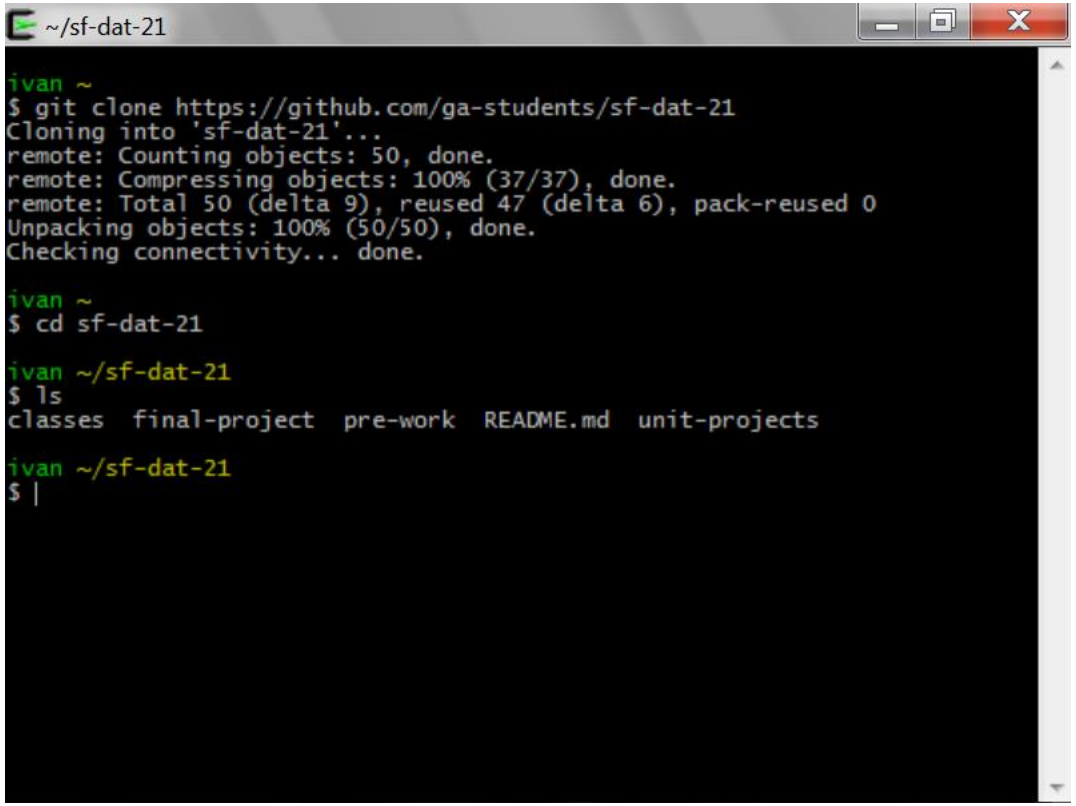
```
git clone https://github.com/ga-students/sf-dat-21
```



# When the cloning is done …

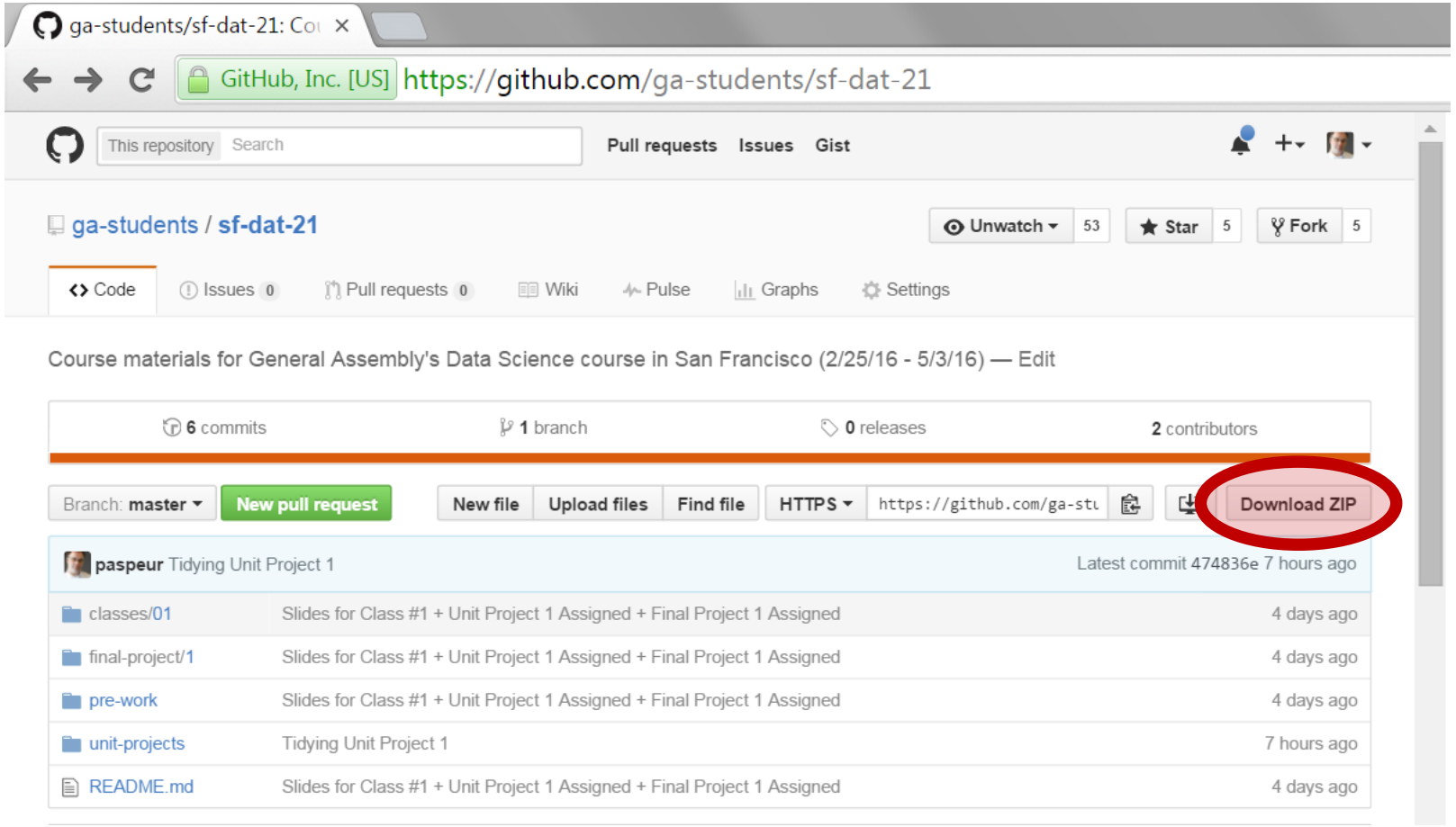


# … the repo content is under `sf-dat-21`



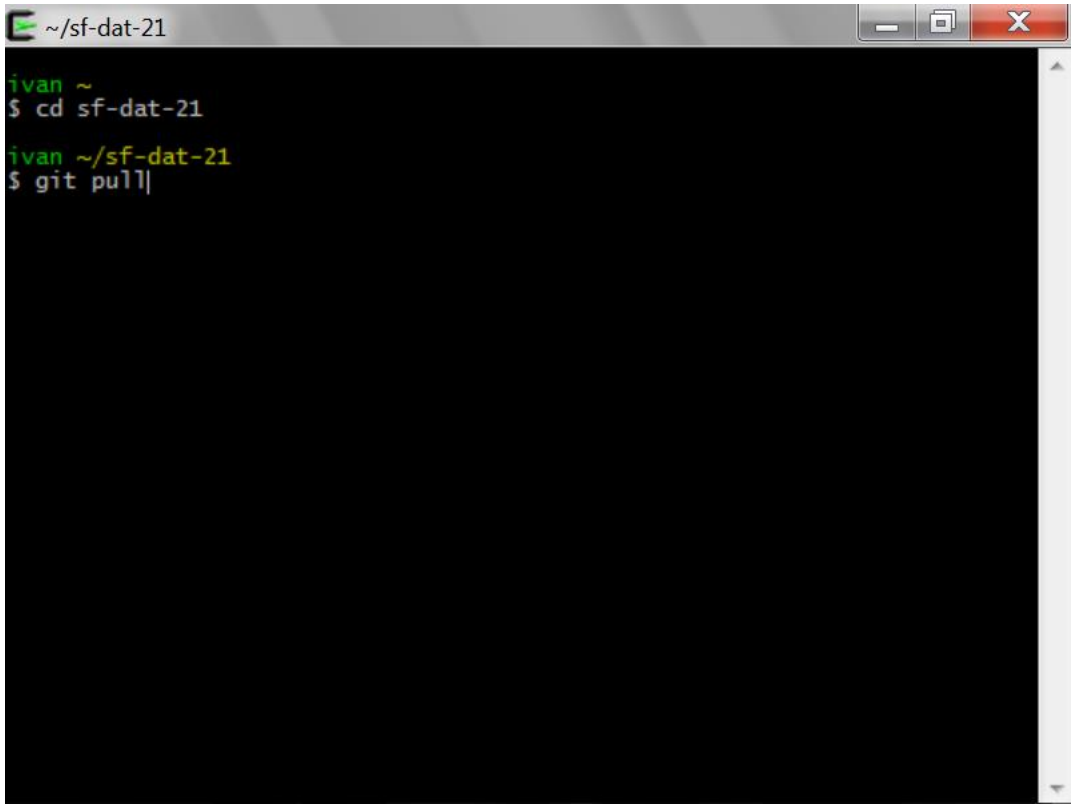

# Alternative method

# Alternatively, download the entire repository as a ZIP file

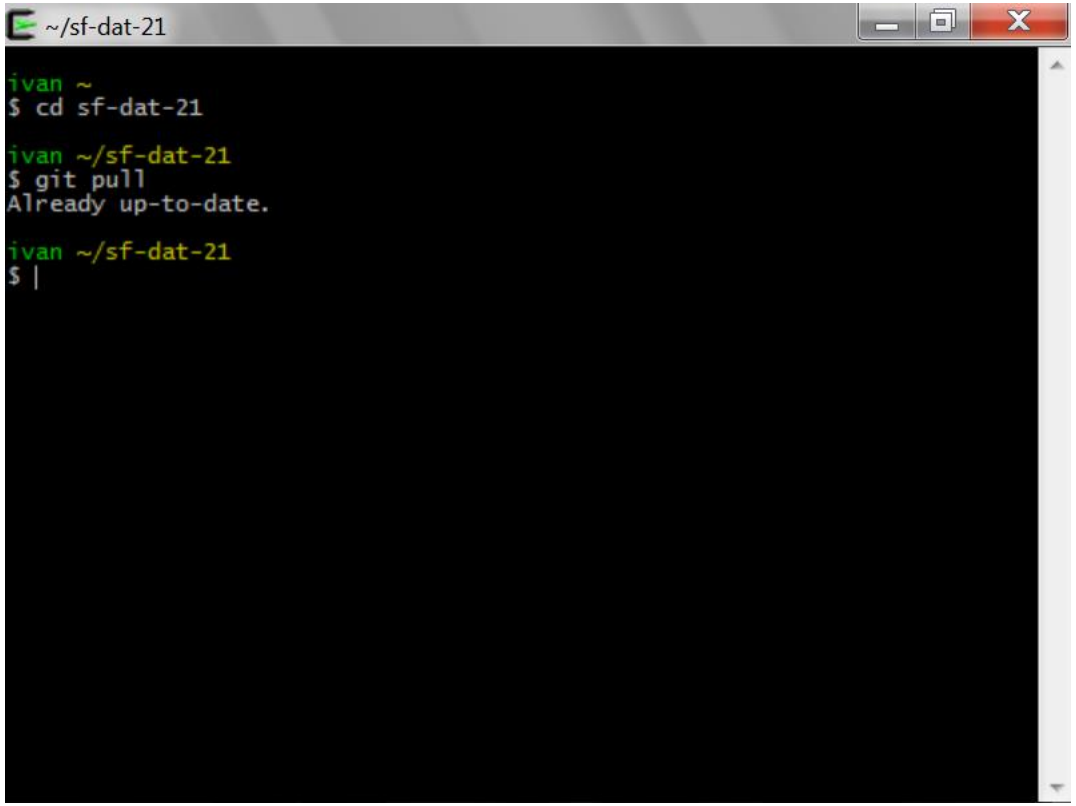



# Updating your local copy of the repository via the command line

# First `cd` to your local copy, then update the repo with `git pull`



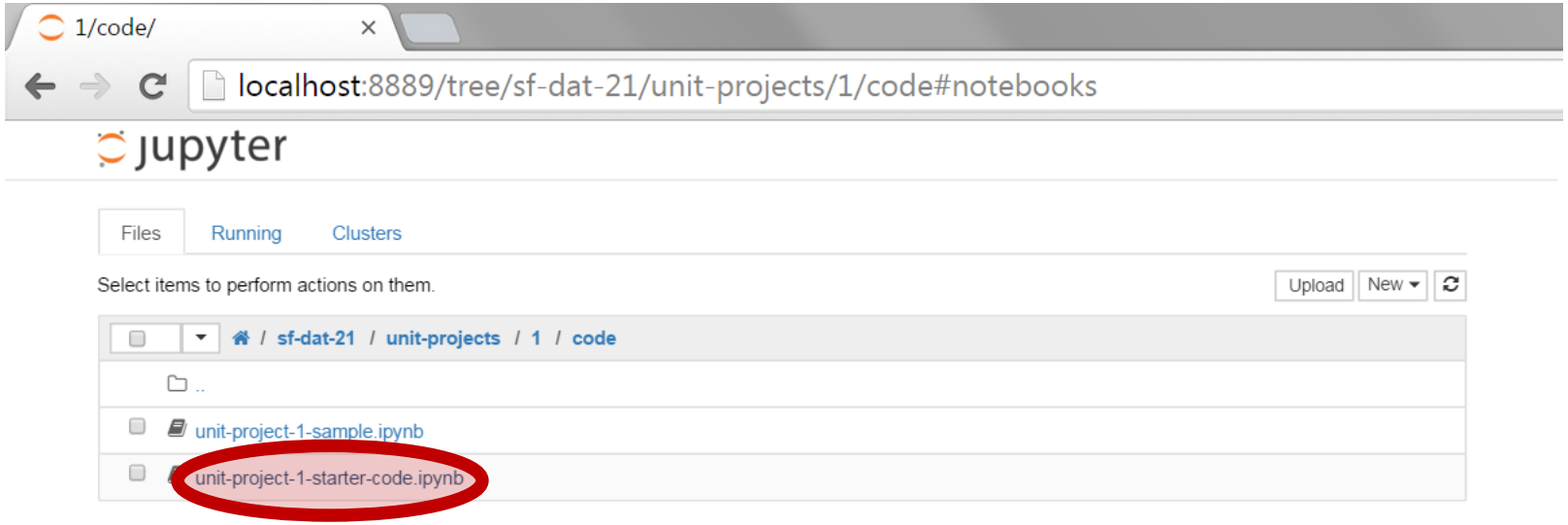
# Updating the local copy (cont.)



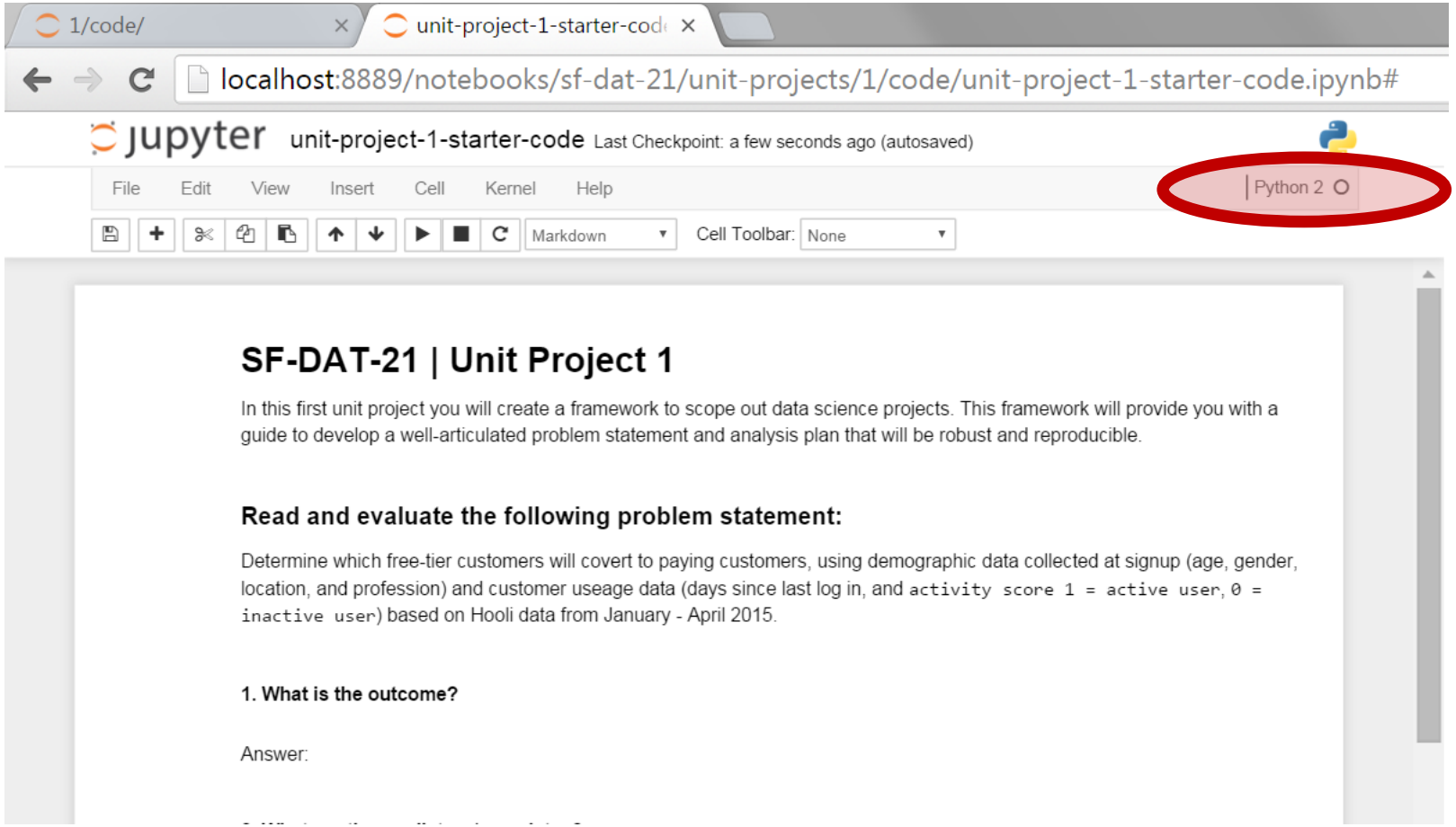

# Verify that Anaconda for Python 2.7 was correctly installed

# First, start IP[y]: IPython (Py 2.7) Notebook

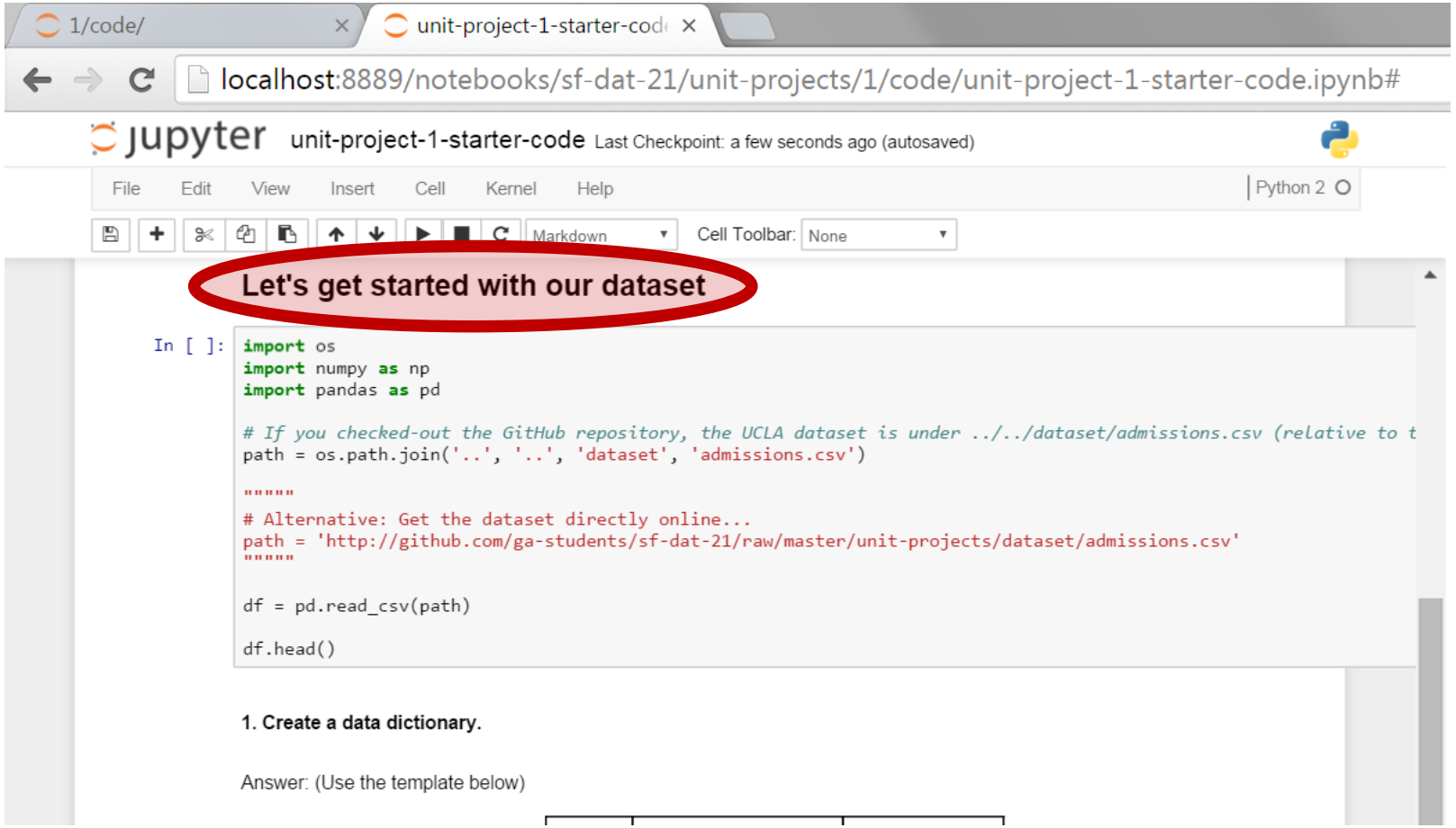
# Navigate to Unit Project 1 Stater Code



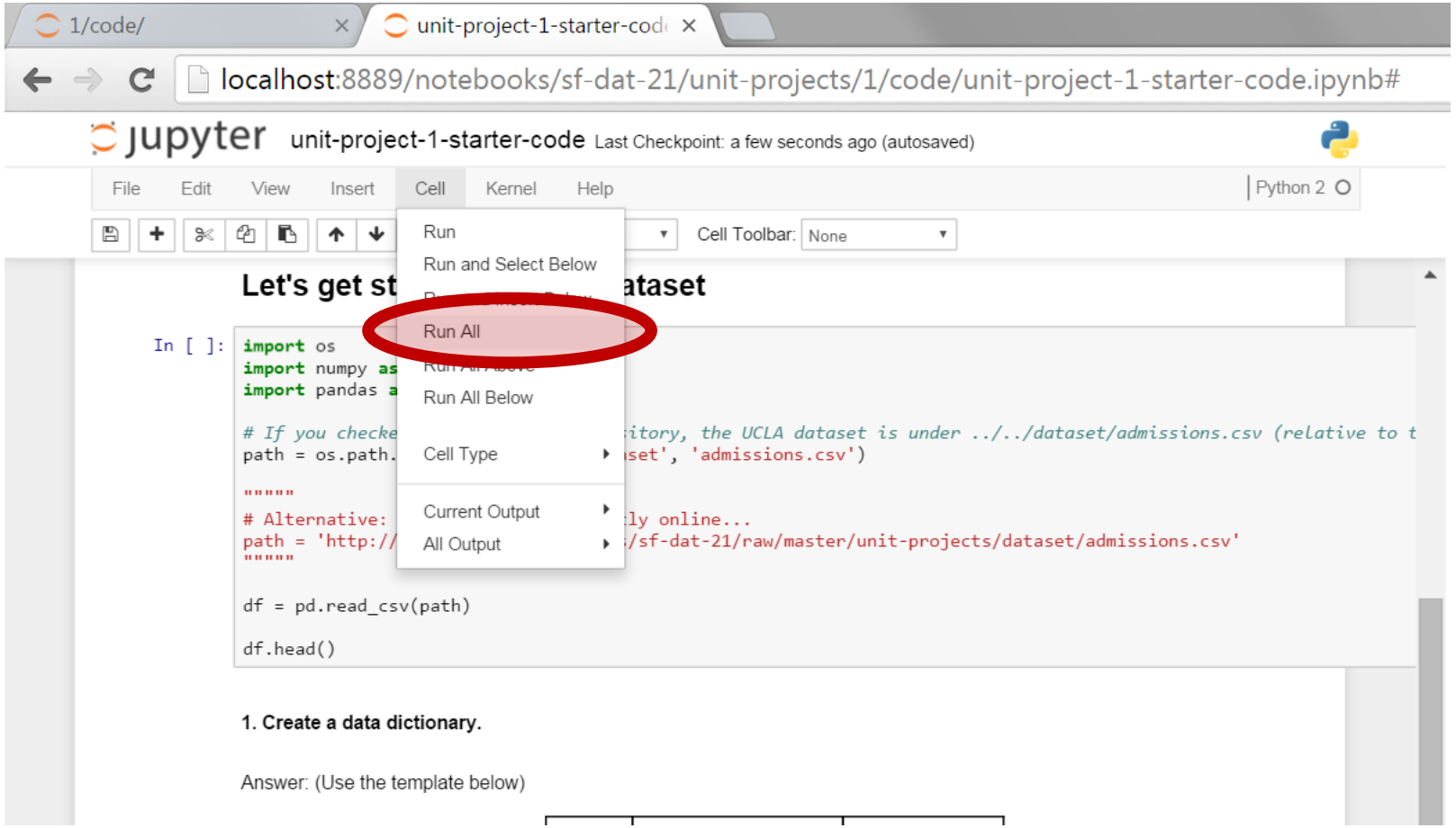
# Using Python 2?



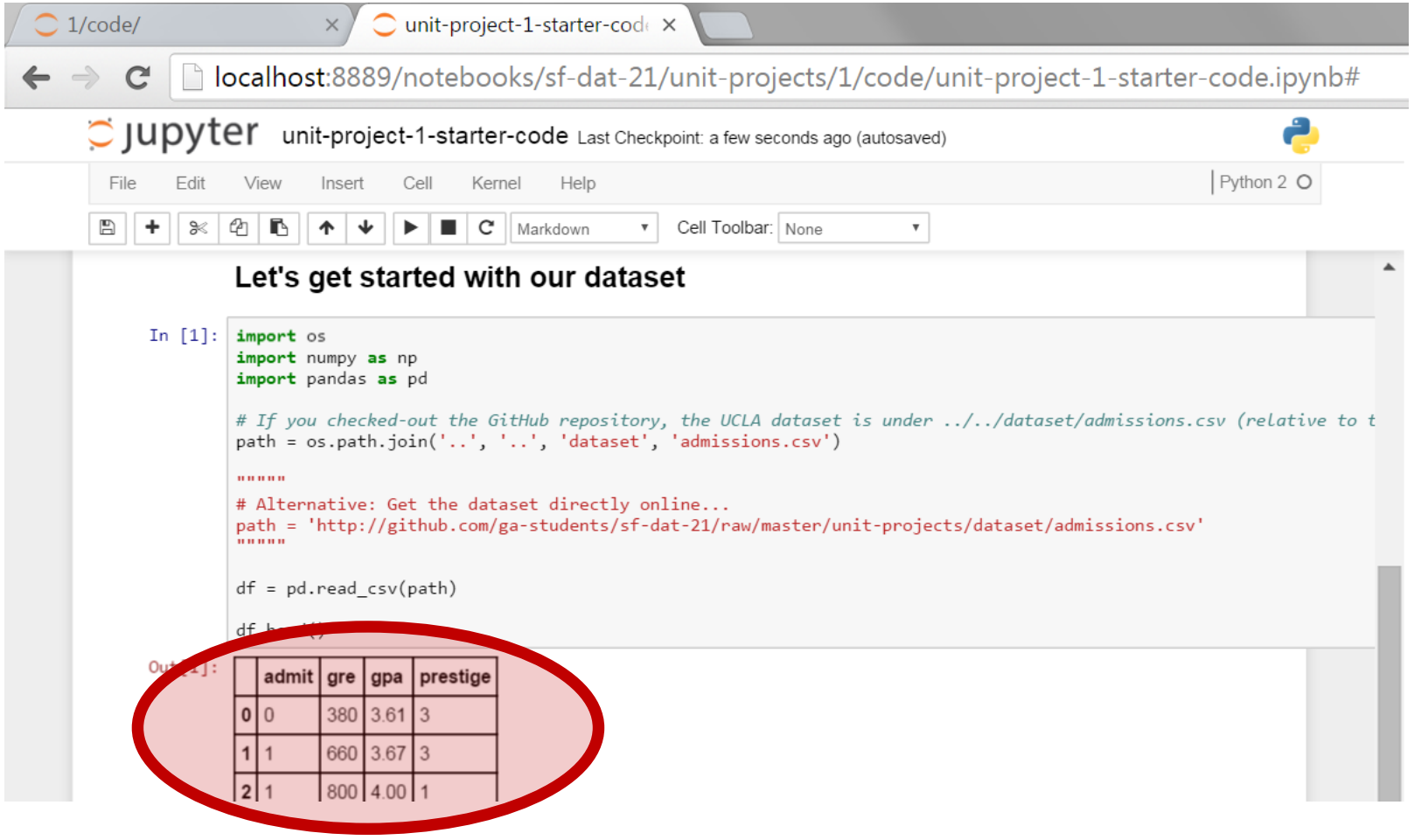
# Scroll down to “Let’s get started with our dataset”



# Cell > Run All

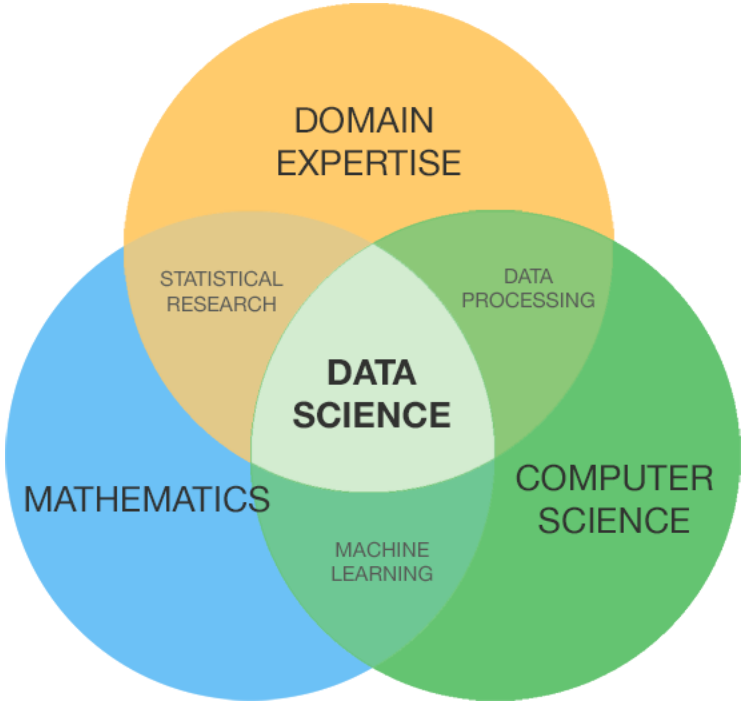

1/code/
unit-project-1-starter-cod
localhost:8889/notebooks/sf-dat-21/unit-projects/1/code/unit-project-1-starter-code.ipynb#
jupyter unit-project-1-starter-code Last Checkpoint: a few seconds ago (autosaved)
File Edit View Insert Cell Kernel Help
Python 2
Cell Toolbar: None
Run
Run and Select Below
Run All
Run All Below
Cell Type
Current Output
All Output
Let's get st ataset
In [ ]:
import os
import numpy as
import pandas a
# If you checke itory, the UCLA dataset is under ../../dataset/admissions.csv (relative to t
path = os.path. set', 'admissions.csv')
"""
# Alternative: ly online...
path = 'http:// /sf-dat-21/raw/master/unit-projects/dataset/admissions.csv'
"""
df = pd.read_csv(path)
df.head()
1. Create a data dictionary.
Answer: (Use the template below)

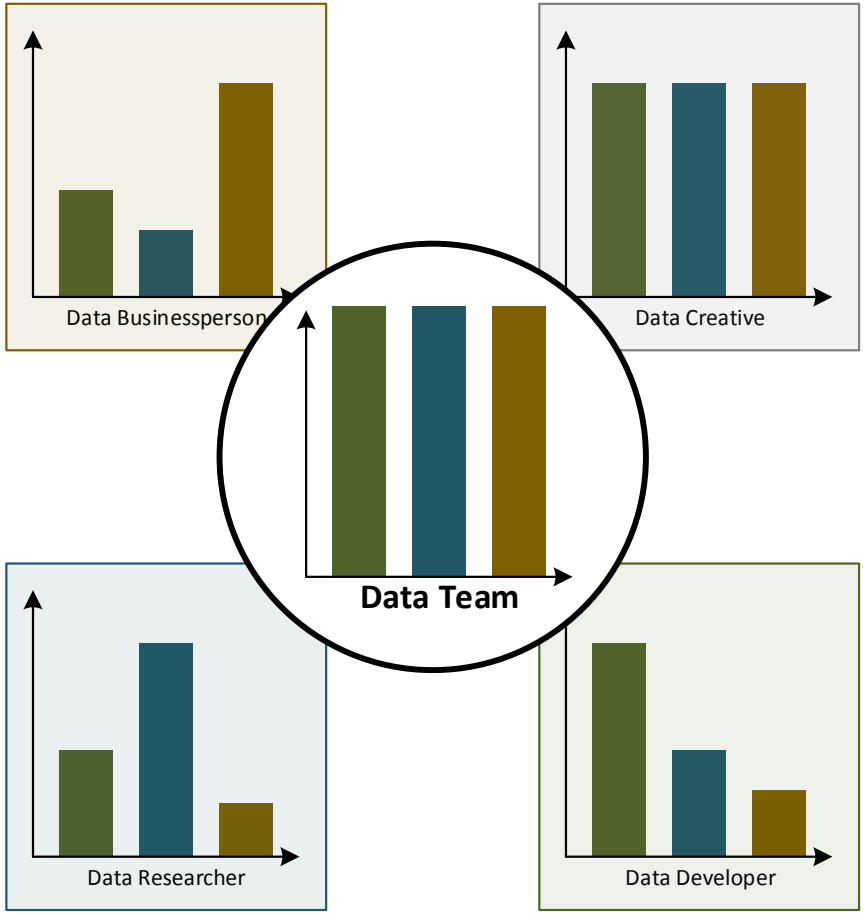
# If all goes well, the dataset's head is shown




# Review
# What is Data Science?
# Who are Data Scientists?

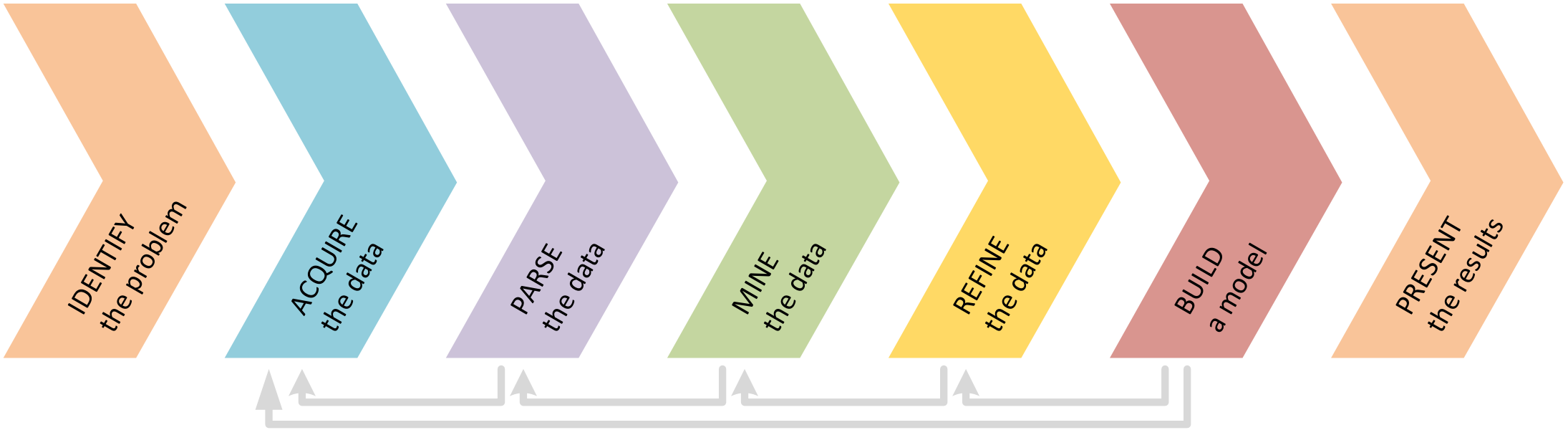
# What is Data Science? Who are Data Scientists?



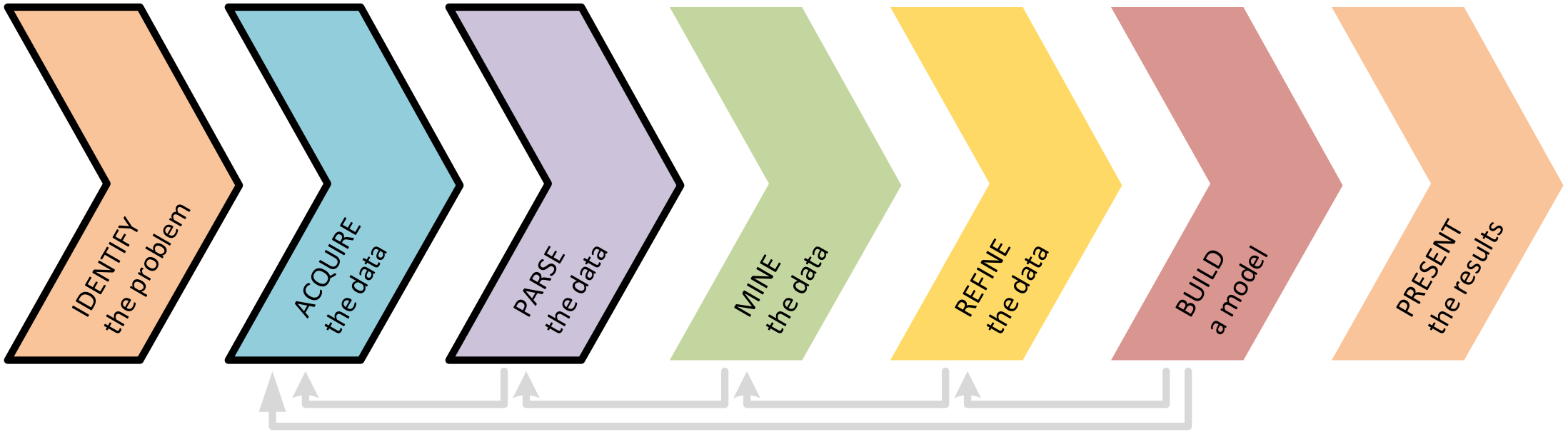




# Review and Exercise Data Science Workflow

# The Data Science Workflow (again…)

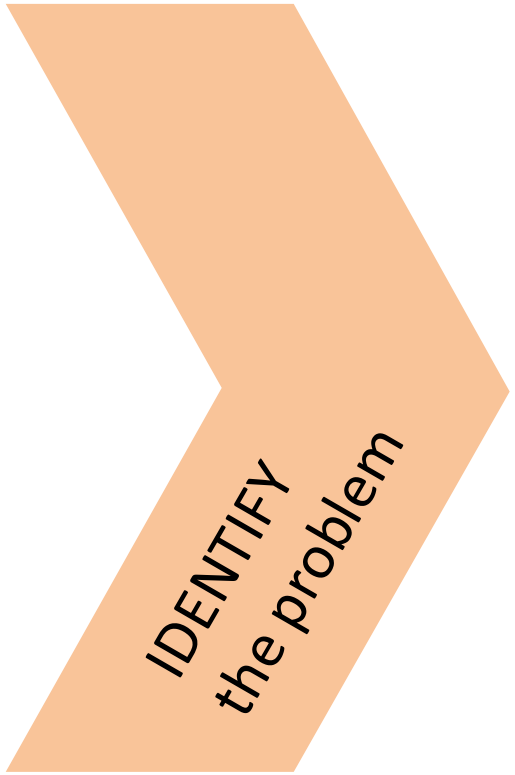


# Today we'll focus on the first three (IDENTIFY the problem, ACQUIRE the data, and PARSE the data)




# ❶ IDENTIFY the Problem

# ❶ Identify the Problem




- Identify the Problem
  - Identify business/product objectives
  - Identify and hypothesize goals and criteria for success
  - Create a set of questions for identifying correct dataset

# ❶ Identify the Problem (cont.)

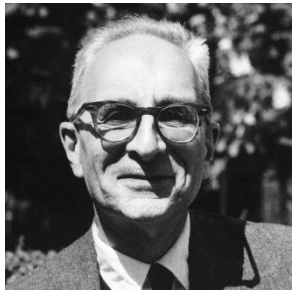
- Identify the Problem
  - Identify business/product objectives
  - Identify and hypothesize goals and criteria for success
  - Create a set of questions for identifying correct dataset

- The Why's and How's of a Good Question
- The SMART Goals Framework


# The Why's and How's of a Good Question

# Why do we need a good question?

- “The scientist is not a person who gives the right answers, he's one who asks the right questions.” – Claude Lévi-Strauss



- “If they can get you asking the wrong questions, they don't have to worry about answers.” – Thomas Pynchon

- “Judge a man by his questions rather than by his answers.” – Voltaire

# By asking a good question and setting a clear aim:

- You set yourself up for success
  - “A problem well stated is half solved” – Charles Kettering
- You help other data scientists learn from and reproduce your work
  - You establish the basis for making your analysis reproducible
- You also help them expand on your work in the future


# The SMART Goals Framework

# The SMART Goals Framework provides a good foundation to set a clear aim

| | | |
|---|---|---|
| **S**PECIFIC | Who, What, Where, When, Why, Which | Define the goal as much as possible, with no ambiguous language. WHO is involved, WHAT do I want to accomplish, WHERE will it be done, WHY am I doing this — reasons, purpose, WHICH constraints and/or requirements do I have? |
| **M**EASURABLE (MEANINGFUL) | From and To | Can you track the progress and measure the outcome? How much, how many, how will I know when my goal is accomplished? |
| **A**TTAINABLE (ACTION ORIENTED) | How | Is the goal reasonable enough to be accomplished? How so? Make sure the goal is not out reach or below standard performance |
| **R**ELEVANT (REALISTIC) | Worthwhile | Is the goal worthwhile and will it meet your needs? Is each goal consistent with other goals you have established and fits with your immediate and long term plans? |
| **T**IMELY (TIME-BOUND) | When | Your objective should include a time limit. "I will complete this step by month/day/year." It will establish a sense of urgency and prompt you to have better time management |

# The SMART Framework can be tuned up for DS

| | |
|---|---|
| **S**PECIFIC | The dataset and key variables are clearly defined |
| **M**EASURABLE | The type of analysis and major assumptions are articulated |
| **A**TTAINABLE | The question you are asking is feasible for your dataset and is not likely to be biased |
| **R**EPRODUCIBLE | Another person (or you in 6 months!) can read your state and understand exactly how your analysis is performed |
| **T**IME-BOUND | You clearly state the time period and population for which this analysis will pertain |

Trends often change over time and vary by the population of source of your data. It is important to clearly define who/what you included in your analysis as well as the time period for the analysis

# Determine the association of foods in the home with child dietary intake. Is this Aim SMART?

EXAMPLE

- Using one 24-hour recall from the cross-sectional NHANES 2007-2010 we will determine the factors associated with food available in the homes of American children and adolescents. We will test if self-reported availability of fruits, dark green vegetables, low fat milk or sugar sweetened beverages available in the home increases the likelihood that children and adolescents will meet their USDA recommended dietary intake for that food. Our hypothesis is that children will be more likely to meet the USDA recommended intake level when food is always available in their home compared to rarely or never

# The aim is Specific

| | |
|---|---|
| **S**PECIFIC<br><br>The dataset and key variables are clearly defined | Determine the association of foods in the home with child dietary intake. Using **one 24-hour recall** from the cross-sectional NHANES 2007-2010 we will determine the factors associated with food available in the homes of **American children and adolescents**. We will test if **self-reported availability of fruits, dark green vegetables, low fat milk or sugar sweetened beverages available in the home increases** the likelihood that children and adolescents will meet their **USDA recommended dietary intake** for that food. Our hypothesis is that children will be ***more likely*** **to meet the USDA recommended intake level when food is *always* available in their home compared to *rarely* or *never*** |

# The aim is Specific (cont.)

| | |
|---|---|
| **S**PECIFIC<br><br>The dataset and key variables are clearly defined | ❑ How data was collected:<br>❑ 24-hour recall, self-reported<br><br>❑ What data was collected:<br>❑ Fruits, dark green vegetables, low fat milk or sugar sweetened beverages, always vs. rarely available<br><br>❑ How data will be analyzed:<br>❑ Using USDA recommendations as a gold-standard to measure the association<br><br>❑ The specific hypothesis and direction of the expected associations:<br>❑ American children and adolescents will be more likely to meet their recommended intake level |

# The aim is Measurable

| **M**EASURABLE | |
|---|---|
| The type of analysis and major assumptions are articulated | ❑ Determine the association of foods in the home with child dietary intake<br>❑ We will test if the self-reported availability of certain foods increases the likelihood that children and adolescents will meet their USDA recommended dietary intake for food |

# The aim is Attainable

| ATTAINABLE | |
|---|---|
| The question you are asking is feasible for your dataset and is not likely to be biased | ❑ Association, not causation<br>  ❑ We are determining the association between two items (food availability in homes and children meeting their dietary recommendations)<br>  ❑ Because we are using cross-sectional data, we cannot say that having fruits and vegetables at home actually causes children to meet their dietary requirements<br><br>❑ Note that cross-sectional data has inherent limitations, one being that causal inference is typically not possible |

# The aim is Reproducible

| **R**EPRODUCIBLE | |
|---|---|
| Another person can read your state and understand exactly how your analysis is performed | ❑ With all the specifics, it would be straightforward to pull the data from NHANES and reproduce the analysis |

# The aim is Time-Bound

| **T**IME-BOUND | |
|---|---|
| You clearly state the time period and population for which this analysis will pertain | ❑ Using one 24-hour recall from NHANES 2007-2010, we will determine the factors associated with food available in the homes of American children and adolescents |

# Context is important

- The previous example laid out research goals but in a business setting, you will need to articulate business objectives
  - Example:  Success for the Netflix recommendation engine may be if 70% of customers over the age of 18 select a movie from the recommended queue during Q3 of 2015
- Regardless of setting, start your question with the SMART framework to help achieve your objectives


# Activity: Knowledge Check

# Activity: Knowledge Check

EXERCISE

## ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. Which of the following uses the SMART framework? Why? What is missing?

    a. I am looking to determine the value of 2 bedroom homes in San Francisco

    b. Determine what real estate factors besides number of bedrooms (e.g., bathrooms, home size, lot size, amenities, etc.) explains the difference in values between 1 bedroom and 2 bedroom homes in San Francisco using sale data from Zillow between November 2015 and February 2016

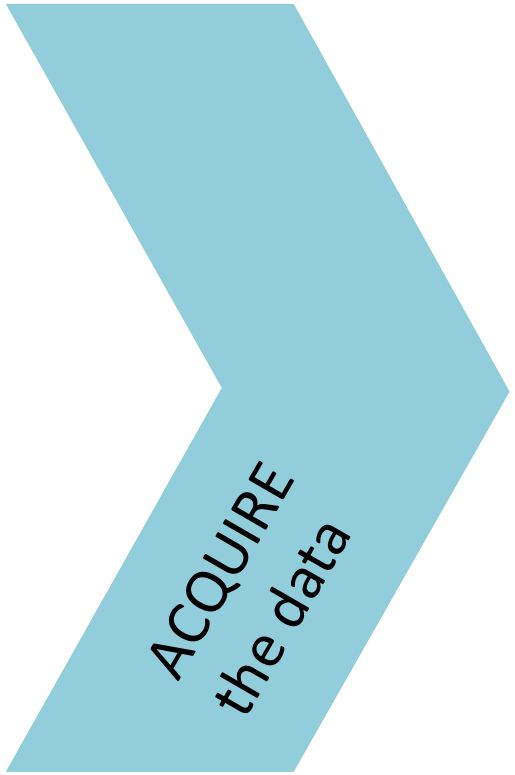
    c. When finished, split into pairs and share your answers with each other

## DELIVERABLE

Answers to the above questions


# ❷ ACQUIRE the Data

# ❷ Acquire the Data




- Acquire the Data
  - Identify the “right” dataset(s)
  - Import data and set up local or remote data structure
  - Determine most appropriate tools to work with data

# ❷ Acquire the Data (cont.)

- Acquire the Data
  - Identify the “right” dataset(s)
  - Import data and set up local or remote data structure
  - Determine most appropriate tools to work with data

- Data Types
- Logistics of Acquiring Data
- Our “Zillow” Dataset
- Tidying Data
- File Formats

# ❷ Acquire the Data (cont.)

- Acquire the Data
  - Identify the “right” dataset(s)
  - Import data and set up local or remote data structure
  - Determine most appropriate tools to work with data

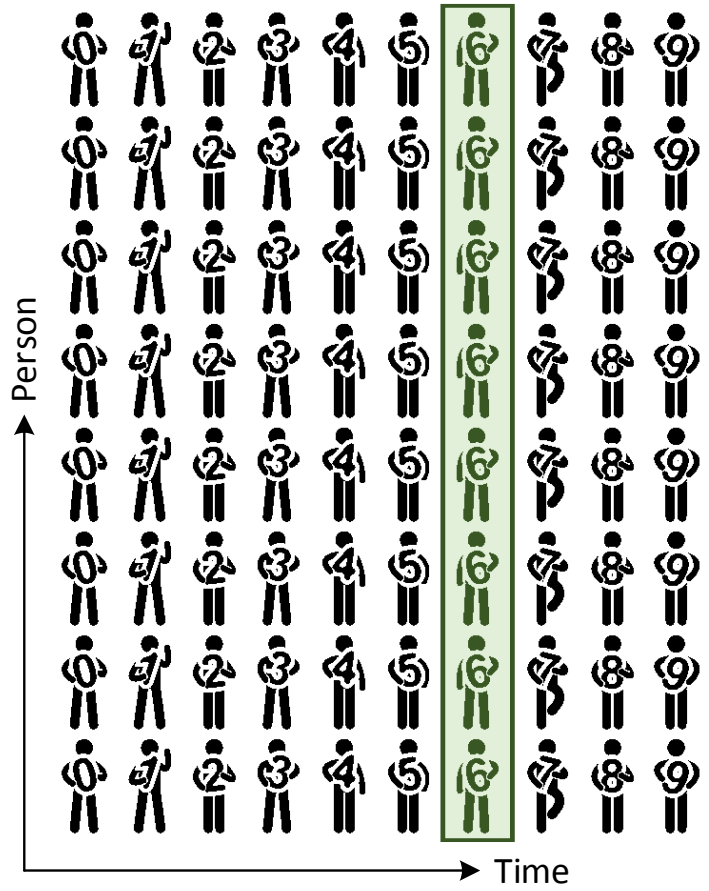
- Questions to ask:
  - What type of data is it, cross-sectional or longitudinal?
  - How well was the data collected?
  - Is there much missing data?
  - Was the data collection instrument calibrated?
  - Is the dataset aggregated?
  - Do we need pre-aggregated data?


# Data Types

# Why Data Types Matter

- Different data types have different limitations and strengths
- Certain types of analyses aren't possible with certain data types
- There are 2 types of data which we may might use for analysis

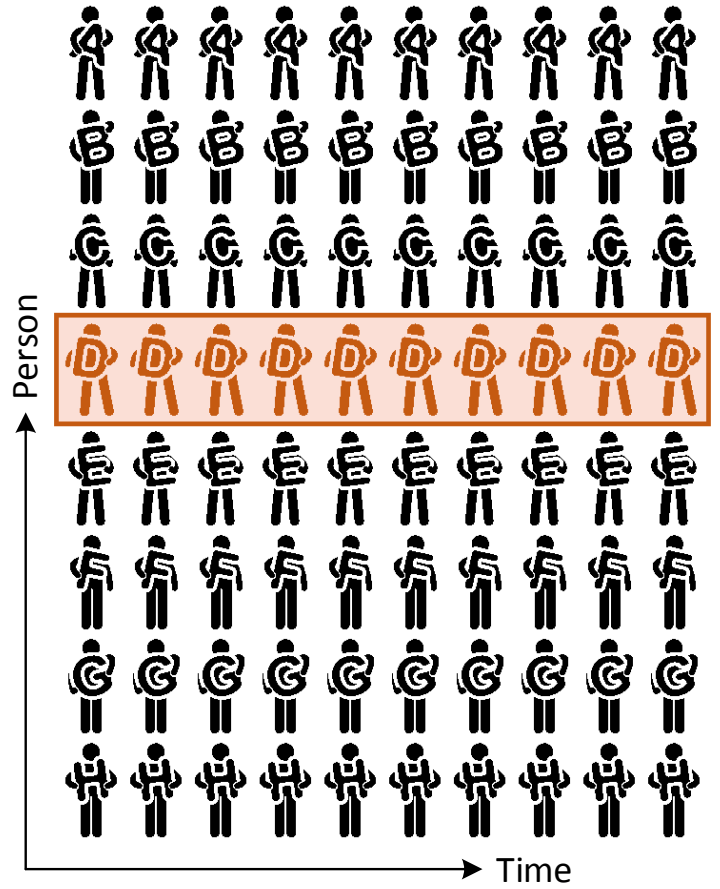
# Cross-Sectional Data




- All information is determined at the same time; all data comes from the same time period
- Note:  There is no distinction between exposure and outcome

# Time Series/Longitudinal Data

- Information is collected over a period of time



# Data Types: Strengths and Weaknesses

| | Strengths | Weaknesses |
|---|---|---|
| **Cross-Sectional Data** | ❑ Often population-based<br>❑ Generalizable<br>❑ Less expensive compared to other types of data collection methods | ❑ Separation of cause and effect may be difficult (or impossible)<br>❑ Variables/cases with long duration are over-represented |
| **Time Series/Longitudinal Data** | ❑ Unambiguous temporal sequence; exposure precedes outcome<br>❑ Multiple outcomes can be measured | ❑ Takes a long time to collect data<br>❑ Vulnerable to missing data<br>❑ More expense compared to other types of data collection methods |


# Activity: Knowledge Check

# Activity: Knowledge Check

EXERCISE

## ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. What type of data is shown by Zillow? (http://www.zillow.com/san-francisco-ca/sold/)

2. Can you create a cross-sectional analysis from a longitudinal data collection? How? Is this applicable from the data above?

3. When finished, split into pairs and share your answers with each other

## DELIVERABLE

Answers to the above questions


# Activity: Write a Research Question with Raw Data

# Activity: Write a Research Question with Raw Data

EXERCISE

## DIRECTIONS (10 minutes)

1. Individually, look at the data from Kaggle's Titanic competition (https://www.kaggle.com/c/titanic/data) (also in the course repo) and write a high quality research question

2. Make sure you answer the following questions:
   a. What type of data is this, cross-sectional or longitudinal?
   b. What will we be measuring?
   c. What is the SMART aim for this data?

3. When finished, split into pairs and share your answers with each other

## DELIVERABLE

Research Question


# Logistics of Acquiring Data

# Logistics of Acquiring Data

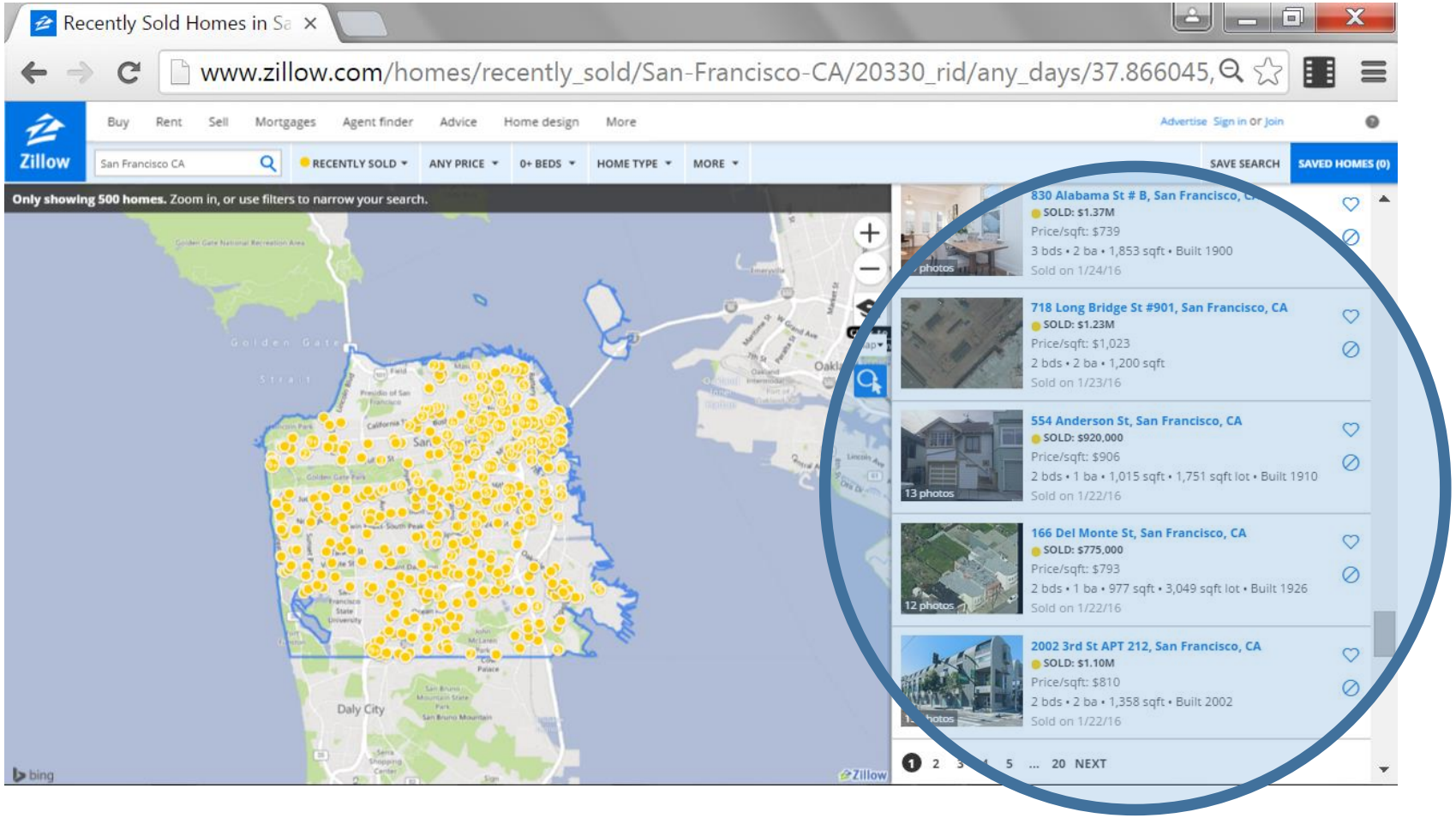
- Data can be acquired through a variety of sources
  - Web (e.g., Google Analytics, HTML)
  - Databases
    - SQL (Structured Query Language)
    - NoSQL (“Not only SQL”)
- Files
  - CSV (Comma-Separated Values)
  - TSV/TXT (Tab-Separated Values)
  - JSON (JavaScript Object Notation)
  - XML (eXtensible Markup Language)

# Zillow dataset: a dataset we will use consistently use across this course

CASE STUDY

- Recently Sold Homes (Source: Zillow)
  - 1,000 homes sold in San Francisco between 11/10/2015 and 2/12/2106

# Raw data was scrapped from the Zillow website (20 pages, each listing 50 homes = 1,000 homes)



# Raw data is Messy™…

```
<div class="property-info"
id="yui_3_18_1_1_1456167242885_71870"><strong
id="yui_3_18_1_1_1456167242885_71869"><dt class="property-address"
id="yui_3_18_1_1_1456167242885_71868"><a href="/homedetails/149-
Shipley-St-San-Francisco-CA-94107/15147894_zpid/" class="hdp-link
routable" title="149 Shipley St, San Francisco, CA Real Estate"
id="yui_3_18_1_1_1456167242885_71873">149 Shipley St, San
Francisco, CA</a></dt></strong><dt class="listing-type zsg-
content_collapsed" id="yui_3_18_1_1_1456167242885_71875"><span
class="zsg-icon-recently-sold type-icon"></span>Sold:
$1.18M</dt><dt class="zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71877">Price/sqft: $1,116</dt><dt
class="property-data" id="yui_3_18_1_1_1456167242885_71880"><span
class="beds-baths-sqft">3 bds • 2 ba • 1,057 sqft</span><span
class="built-year" id="yui_3_18_1_1_1456167242885_71879"> • Built
1992</span></dt><dt class="sold-date zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71975">Sold on 2/22/16</dt></div>
```

# … and needs to be parsed and tidied up (a.k.a., organized)

```
<div class="property-info"
id="yui_3_18_1_1_1456167242885_71870"><strong
id="yui_3_18_1_1_1456167242885_71869"><dt class="property-address"
id="yui_3_18_1_1_1456167242885_71868"><a href="/homedetails/149-
Shipley-St-San-Francisco-CA-94107/15147894_zpid/" class="hdp-link
routable" title="149 Shipley St, San Francisco, CA Real Estate"
id="yui_3_18_1_1_1456167242885_71873">149 Shipley St, San
Francisco, CA</a></dt></strong><dt class="listing-type zsg-
content_collapsed" id="yui_3_18_1_1_1456167242885_71875"><span
class="zsg-icon-recently-sold type-icon"></span>Sold:
$1.18M</dt><dt class="zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71877">Price/sqft: $1,116</dt><dt
class="property-data" id="yui_3_18_1_1_1456167242885_71880"><span
class="beds-baths-sqft">3 bds • 2 ba • 1,057 sqft</span><span
class="built-year" id="yui_3_18_1_1_1456167242885_71879"> • Built
1992</span></dt><dt class="sold-date zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71975">Sold on 2/22/16</dt></div>
```

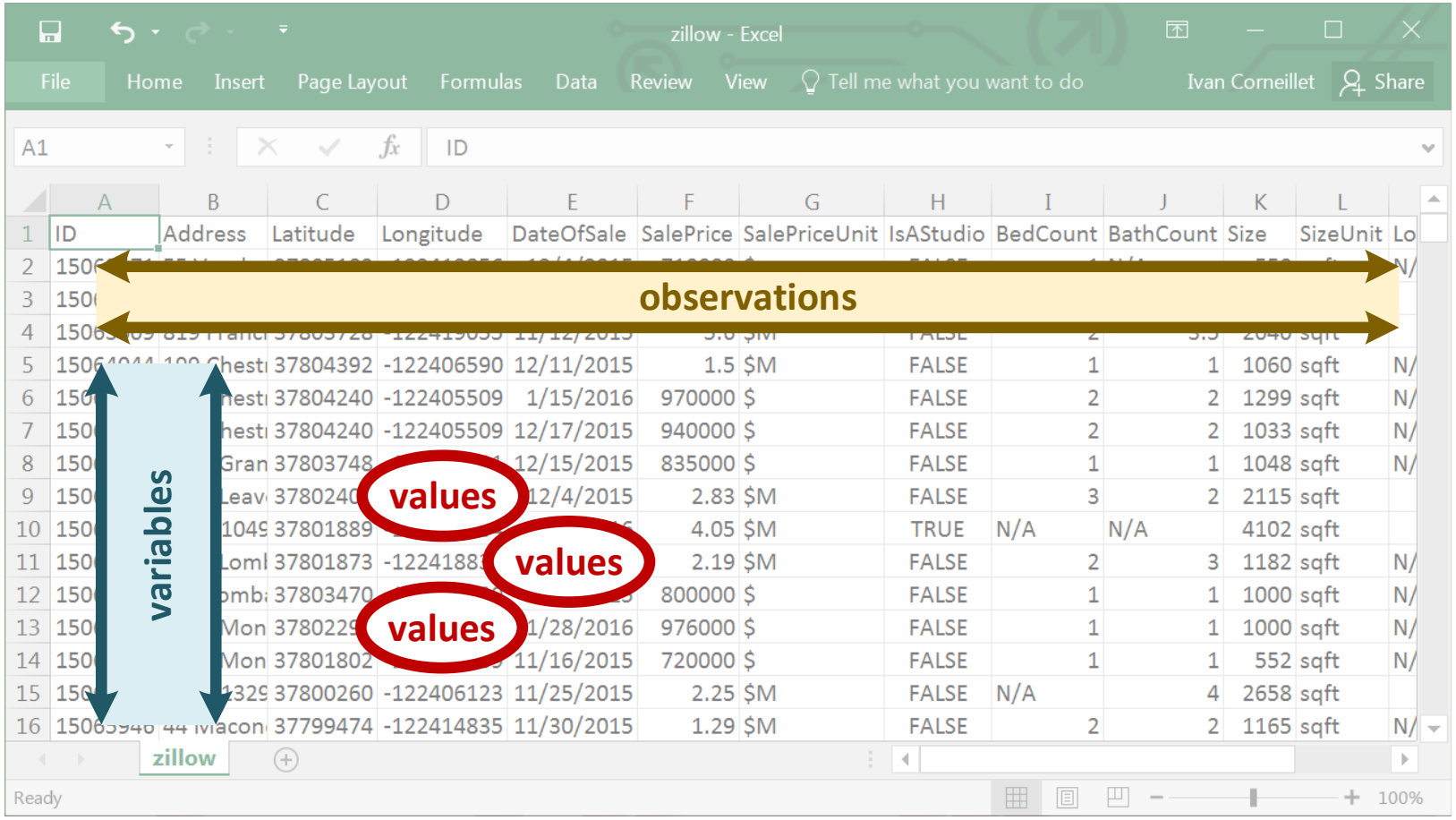
# Tidying up your Data

- Tidying data is the most fruitful skill you can learn as a data scientist
  - It will save you hours of time and make your data much easier to visualize, manipulate, and model
- Many data science tools follow a set of conventions that makes one layout of tabular data much easier to work with than others.  Your data will be easier to work with if you follow three rules:
  - Each observation is placed in its own row
  - Each variable in the dataset is placed in its own column
  - Each value is placed in its own cell

# Really, data can be incredibly raw

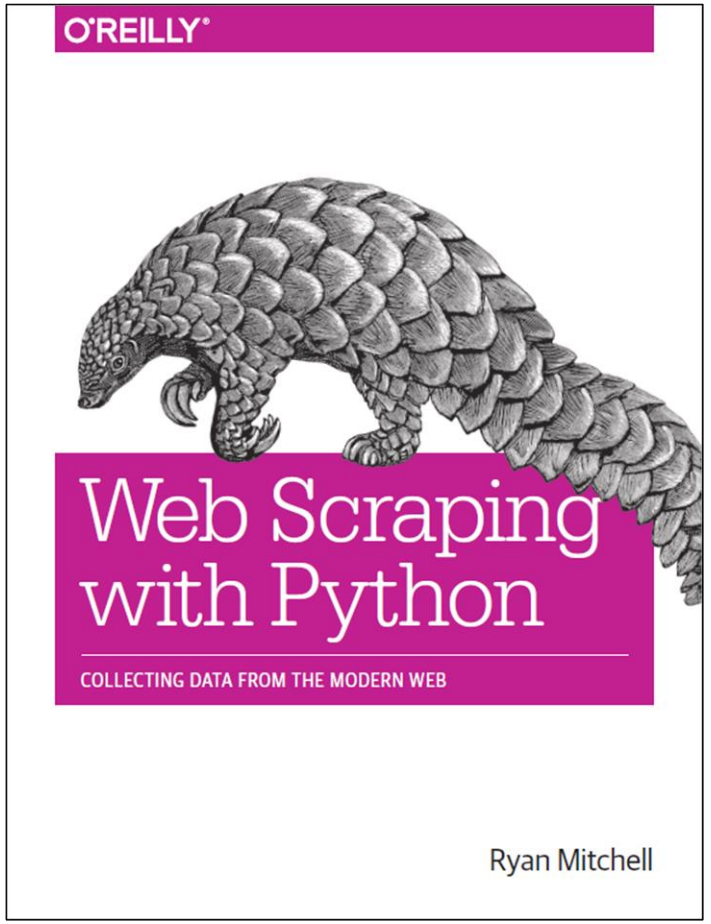
- Trouble tickets inspect and maintain manholes in New Year City
  - "Service box," a common piece of infrastructure, had at least 38 variants, including `SB`, `S`, `S/B`, `S.B`, `S?B`, `S.B.`, `SBX`, `S/BX`, `SB/X`, `S/XB`, `/SBX`, `S.BX`, `S &BX`, `S?BX`, `S BX`, `S/B/X`, `S BOX`, `SVBX`, `SERV BX`, `SERV-BOX`, `SERV/BOX`, and `SERVICE BOX`

# Our tidy Zillow dataset



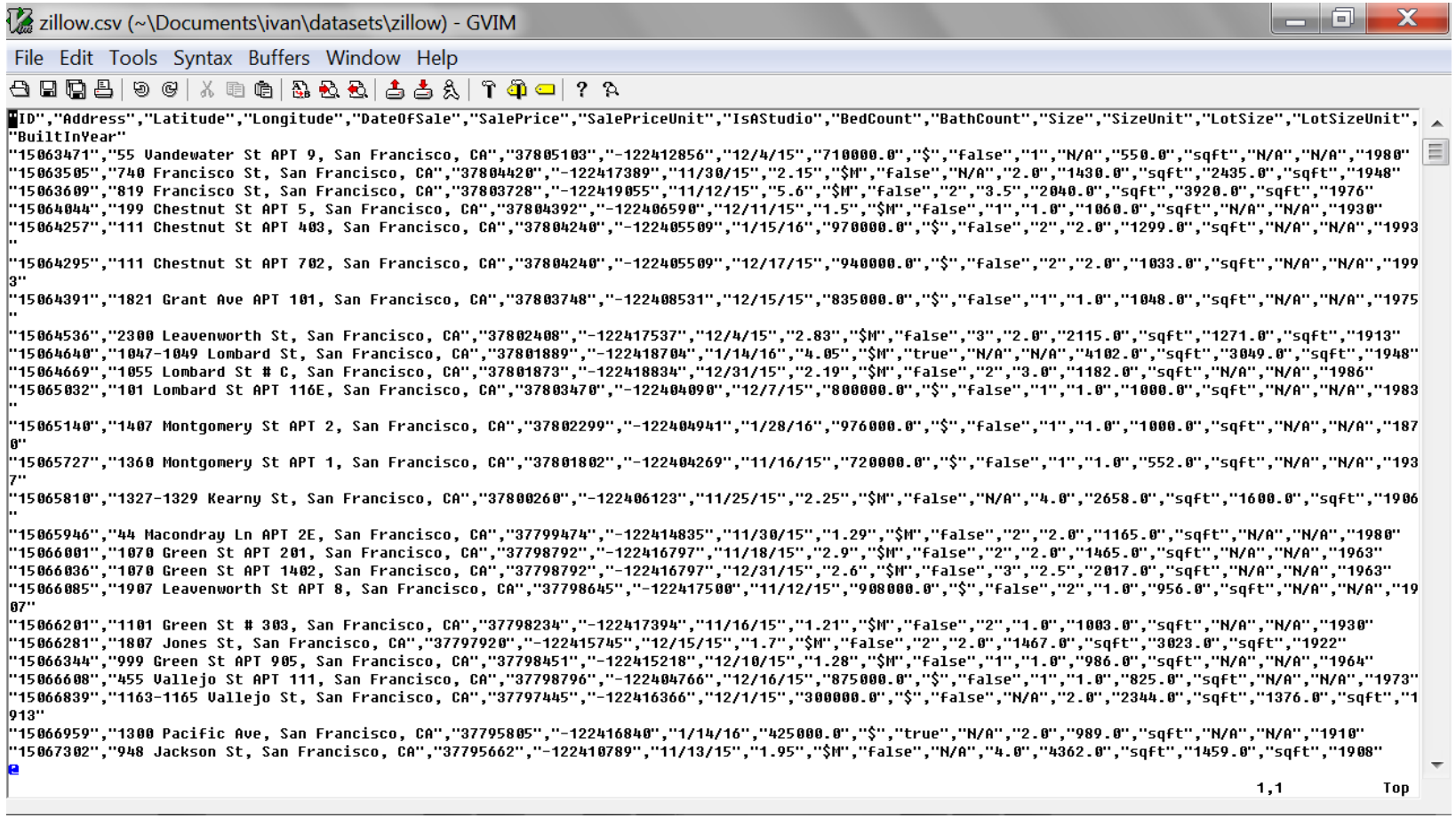

| ID | Address | Latitude | Longitude | DateOfSale | SalePrice | SalePriceUnit | IsAStudio | BedCount | BathCount | Size | SizeUnit | Lo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 150 | | | | | | | | | | | | N/ |
| 150 | | | | | | | | | | | | |
| 150 | | | | | | | | 2 | | | | |
| 150 | | Chesti | 37804392 | -122406590 | 12/11/2015 | 1.5 | $M | FALSE | 1 | 1 | 1060 | sqft | N/ |
| 150 | | nesti | 37804240 | -122405509 | 1/15/2016 | 970000 | $ | FALSE | 2 | 2 | 1299 | sqft | N/ |
| 150 | | nesti | 37804240 | -122405509 | 12/17/2015 | 940000 | $ | FALSE | 2 | 2 | 1033 | sqft | N/ |
| 150 | | Gran | 37803748 | | 12/15/2015 | 835000 | $ | FALSE | 1 | 1 | 1048 | sqft | N/ |
| 150 | | Leav | 378024 | | 12/4/2015 | 2.83 | $M | FALSE | 3 | 2 | 2115 | sqft | |
| 150 | | 1049 | 37801889 | | | 4.05 | $M | TRUE | N/A | N/A | 4102 | sqft | |
| 150 | | Loml | 37801873 | -1224188 | | 2.19 | $M | FALSE | 2 | 3 | 1182 | sqft | N/ |
| 150 | | mbi | 37803470 | | | 800000 | $ | FALSE | 1 | 1 | 1000 | sqft | N/ |
| 150 | | Mon | 378022 | | 1/28/2016 | 976000 | $ | FALSE | 1 | 1 | 1000 | sqft | N/ |
| 150 | | Mon | 37801802 | | 11/16/2015 | 720000 | $ | FALSE | 1 | 1 | 552 | sqft | N/ |
| 150 | | 1329 | 37800260 | -122406123 | 11/25/2015 | 2.25 | $M | FALSE | N/A | 4 | 2658 | sqft | |
| 150 | | Macon | 37799474 | -122414835 | 11/30/2015 | 1.29 | $M | FALSE | 2 | 2 | 1165 | sqft | N/ |

# Going further

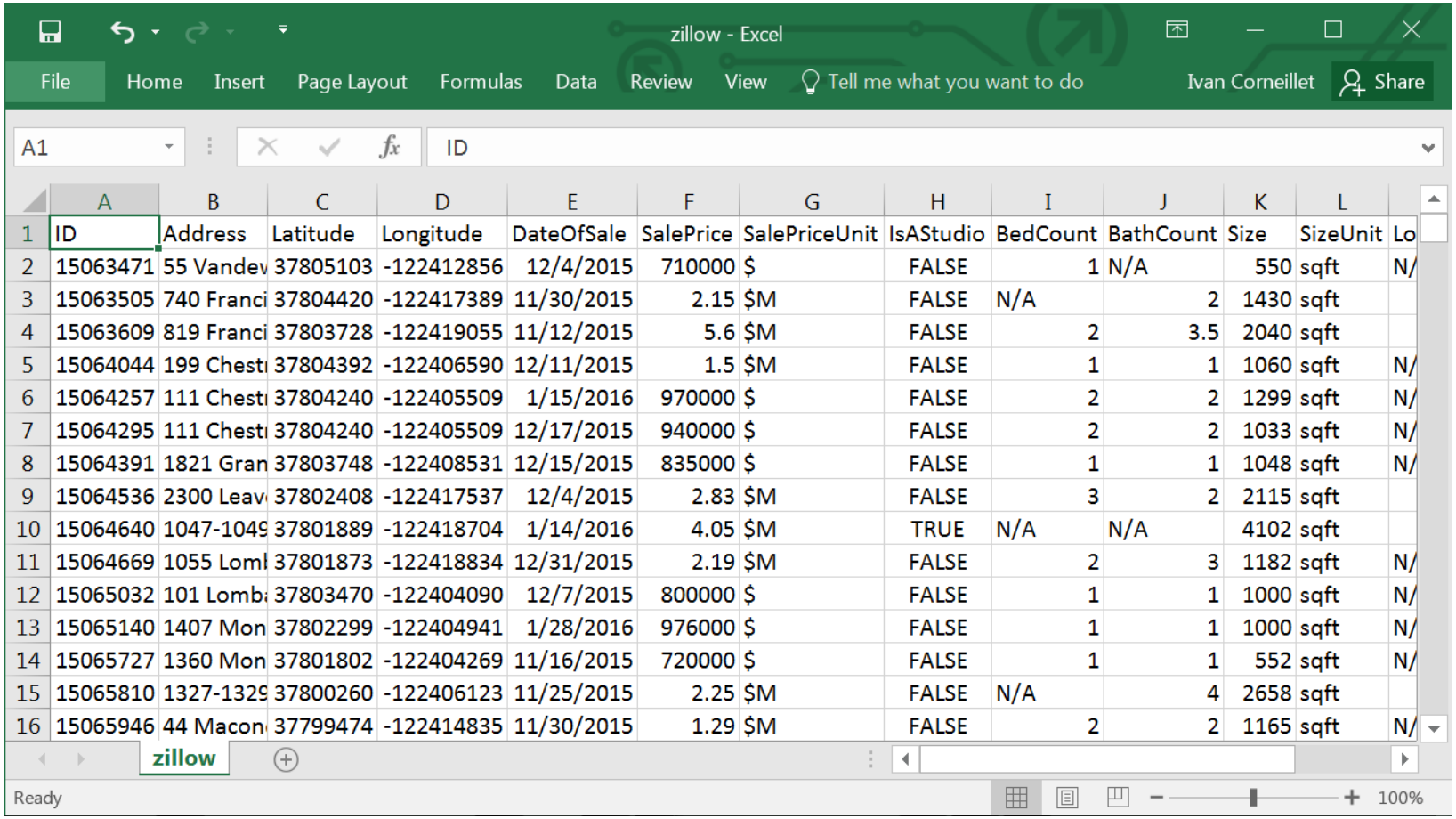



# File Formats

# Our tidy Zillow dataset in CSV format: each observation is in one line; within each line, variables are separated with commas (and here delimited with double-quotes)

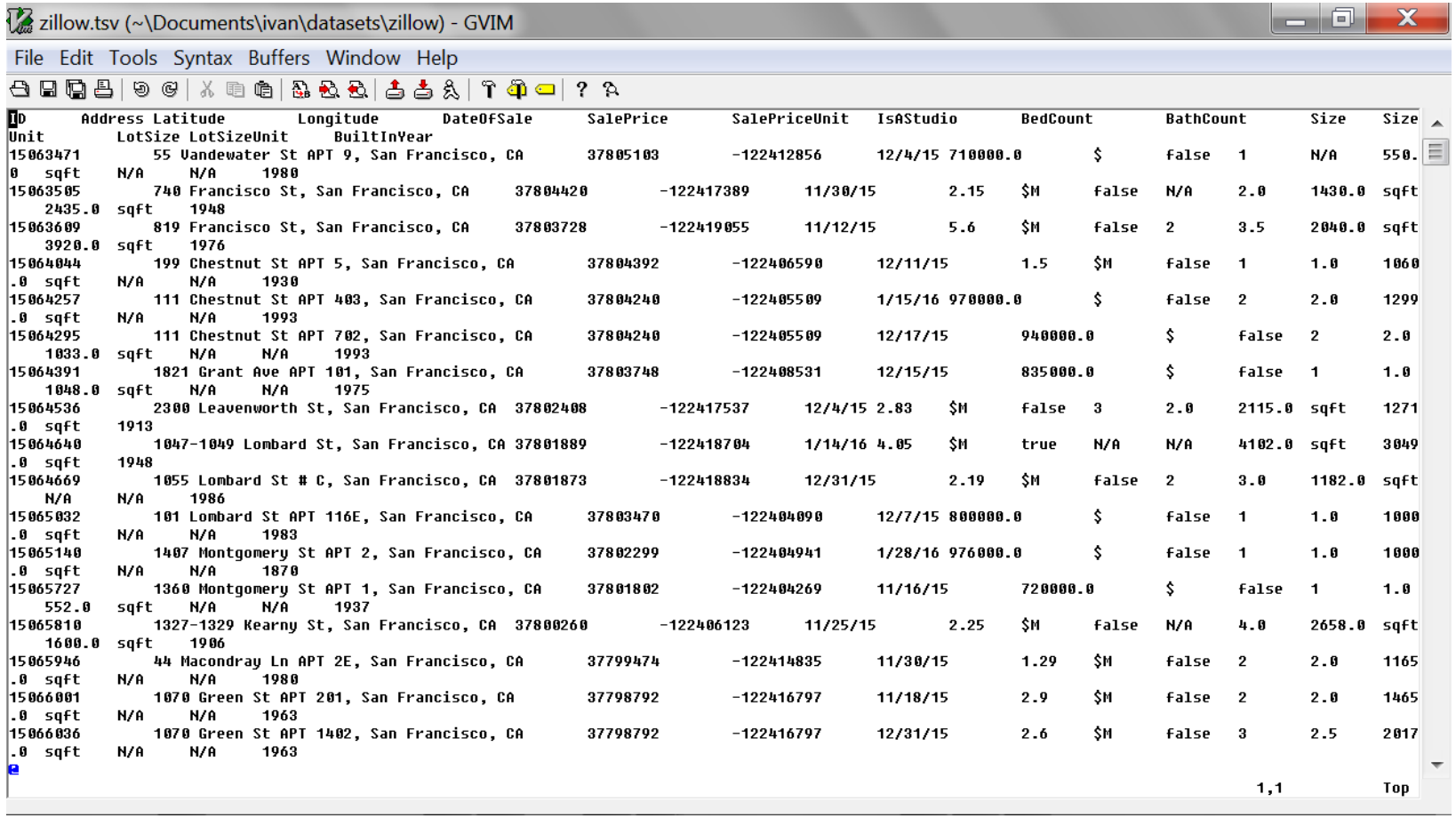


# Excel reads CSV files natively (and our Python toolkit will too…)



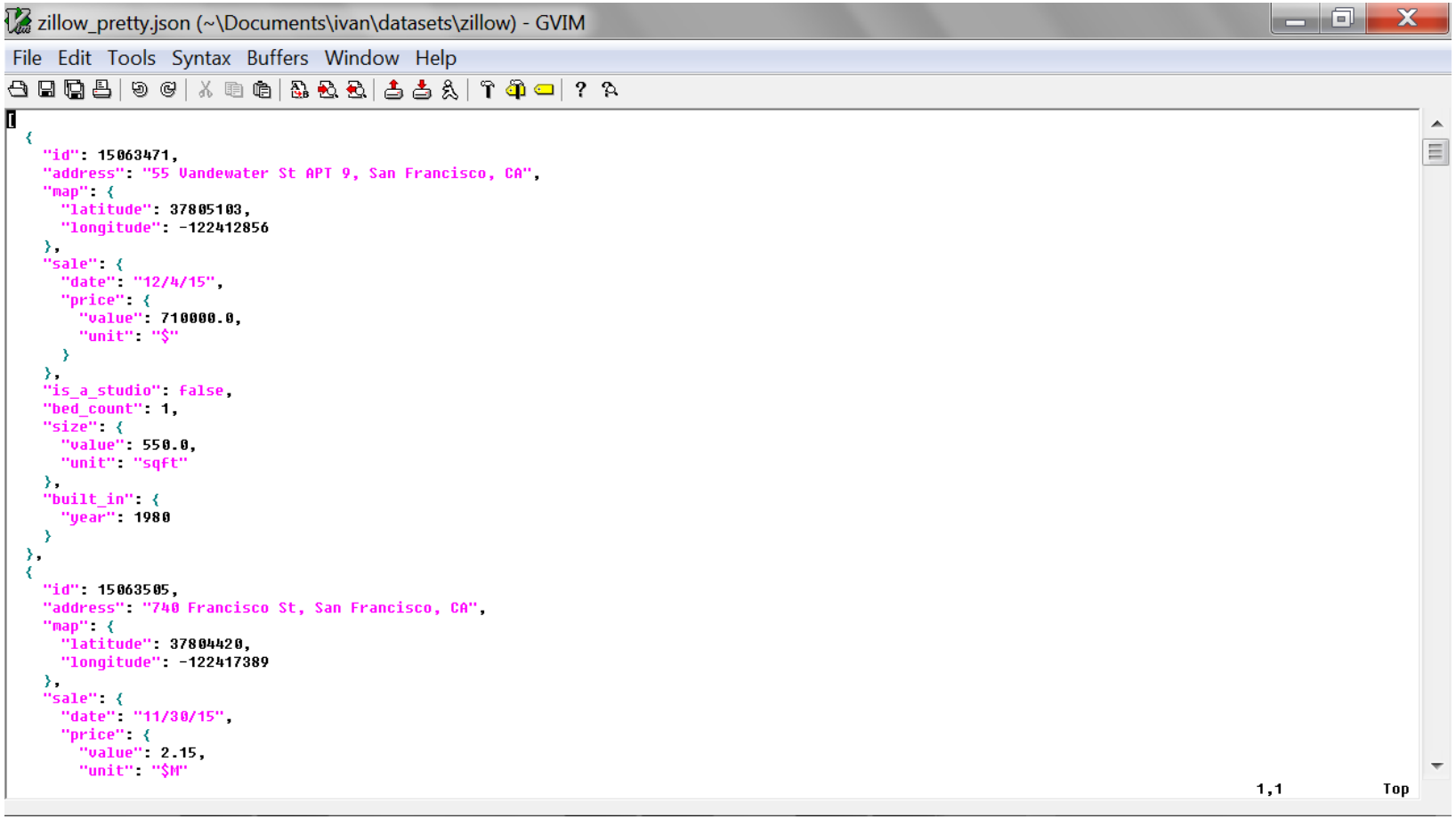

| | A | B | C | D | E | F | G | H | I | J | K | L | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Address | Latitude | Longitude | DateOfSale | SalePrice | SalePriceUnit | IsAStudio | BedCount | BathCount | Size | SizeUnit | Lo |
| 2 | 15063471 | 55 Vandev | 37805103 | -122412856 | 12/4/2015 | 710000 | $ | FALSE | 1 | N/A | 550 | sqft | N/ |
| 3 | 15063505 | 740 Franci | 37804420 | -122417389 | 11/30/2015 | 2.15 | $M | FALSE | N/A | 2 | 1430 | sqft | |
| 4 | 15063609 | 819 Franci | 37803728 | -122419055 | 11/12/2015 | 5.6 | $M | FALSE | 2 | 3.5 | 2040 | sqft | |
| 5 | 15064044 | 199 Chestı | 37804392 | -122406590 | 12/11/2015 | 1.5 | $M | FALSE | 1 | 1 | 1060 | sqft | N/ |
| 6 | 15064257 | 111 Chestı | 37804240 | -122405509 | 1/15/2016 | 970000 | $ | FALSE | 2 | 2 | 1299 | sqft | N/ |
| 7 | 15064295 | 111 Chestı | 37804240 | -122405509 | 12/17/2015 | 940000 | $ | FALSE | 2 | 2 | 1033 | sqft | N/ |
| 8 | 15064391 | 1821 Gran | 37803748 | -122408531 | 12/15/2015 | 835000 | $ | FALSE | 1 | 1 | 1048 | sqft | N/ |
| 9 | 15064536 | 2300 Leav | 37802408 | -122417537 | 12/4/2015 | 2.83 | $M | FALSE | 3 | 2 | 2115 | sqft | |
| 10 | 15064640 | 1047-1049 | 37801889 | -122418704 | 1/14/2016 | 4.05 | $M | TRUE | N/A | N/A | 4102 | sqft | |
| 11 | 15064669 | 1055 Lomb | 37801873 | -122418834 | 12/31/2015 | 2.19 | $M | FALSE | 2 | 3 | 1182 | sqft | N/ |
| 12 | 15065032 | 101 Lombı | 37803470 | -122404090 | 12/7/2015 | 800000 | $ | FALSE | 1 | 1 | 1000 | sqft | N/ |
| 13 | 15065140 | 1407 Mon | 37802299 | -122404941 | 1/28/2016 | 976000 | $ | FALSE | 1 | 1 | 1000 | sqft | N/ |
| 14 | 15065727 | 1360 Mon | 37801802 | -122404269 | 11/16/2015 | 720000 | $ | FALSE | 1 | 1 | 552 | sqft | N/ |
| 15 | 15065810 | 1327-1329 | 37800260 | -122406123 | 11/25/2015 | 2.25 | $M | FALSE | N/A | 4 | 2658 | sqft | |
| 16 | 15065946 | 44 Macon | 37799474 | -122414835 | 11/30/2015 | 1.29 | $M | FALSE | 2 | 2 | 1165 | sqft | N/ |

TSV is another simple text format for storing data in a tabular structure: each observation in the table is one line of the text file and each variable is separated from the next by a tab character

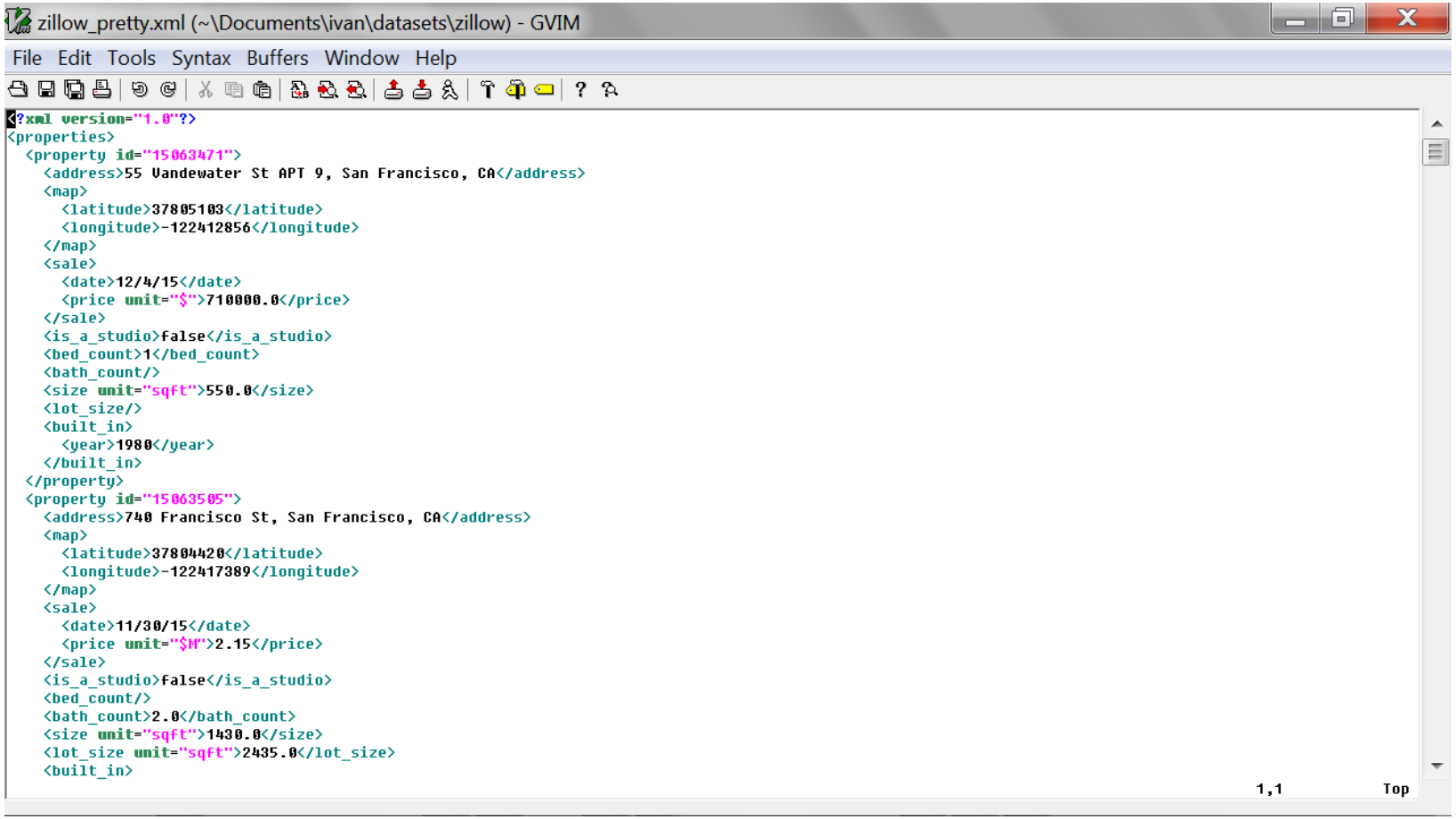



```
ID      Address Latitude        Longitude       DateOfSale      SalePrice       SalePriceUnit   IsAStudio       BedCount        BathCount       Size    Size
Unit        LotSize LotSizeUnit     BuiltInYear
15063471        55 Vandewater St APT 9, San Francisco, CA       37805103        -122412856      12/4/15 710000.0        $       false   1       N/A     550.
0   sqft    N/A     N/A     1980
15063505        740 Francisco St, San Francisco, CA     37804420        -122417389      11/30/15        2.15    $M      false   N/A     2.0     1430.0  sqft
    2435.0  sqft    1948
15063609        819 Francisco St, San Francisco, CA     37803728        -122419055      11/12/15        5.6     $M      false   2       3.5     2040.0  sqft
    3920.0  sqft    1976
15064044        199 Chestnut St APT 5, San Francisco, CA        37804392        -122406590      12/11/15        1.5     $M      false   1       1.0     1060
.0  sqft    N/A     N/A     1930
15064257        111 Chestnut St APT 403, San Francisco, CA      37804240        -122405509      1/15/16 970000.0        $       false   2       2.0     1299
.0  sqft    N/A     N/A     1993
15064295        111 Chestnut St APT 702, San Francisco, CA      37804240        -122405509      12/17/15        940000.0        $       false   2       2.0
    1033.0  sqft    N/A     N/A     1993
15064391        1821 Grant Ave APT 101, San Francisco, CA       37803748        -122408531      12/15/15        835000.0        $       false   1       1.0
    1048.0  sqft    N/A     N/A     1975
15064536        2300 Leavenworth St, San Francisco, CA  37802408        -122417537      12/4/15 2.83    $M      false   3       2.0     2115.0  sqft    1271
.0  sqft    1913
15064640        1047-1049 Lombard St, San Francisco, CA 37801889        -122418704      1/14/16 4.05    $M      true    N/A     N/A     4102.0  sqft    3049
.0  sqft    1948
15064669        1055 Lombard St # C, San Francisco, CA  37801873        -122418834      12/31/15        2.19    $M      false   2       3.0     1182.0  sqft
    N/A     N/A     1986
15065032        101 Lombard St APT 116E, San Francisco, CA      37803470        -122404090      12/7/15 800000.0        $       false   1       1.0     1000
.0  sqft    N/A     N/A     1983
15065140        1407 Montgomery St APT 2, San Francisco, CA     37802299        -122404941      1/28/16 976000.0        $       false   1       1.0     1000
.0  sqft    N/A     N/A     1870
15065727        1360 Montgomery St APT 1, San Francisco, CA     37801802        -122404269      11/16/15        720000.0        $       false   1       1.0
    552.0   sqft    N/A     N/A     1937
15065810        1327-1329 Kearny St, San Francisco, CA  37800260        -122406123      11/25/15        2.25    $M      false   N/A     4.0     2658.0  sqft
    1600.0  sqft    1906
15065946        44 Macondray Ln APT 2E, San Francisco, CA       37799474        -122414835      11/30/15        1.29    $M      false   2       2.0     1165
.0  sqft    N/A     N/A     1980
15066001        1070 Green St APT 201, San Francisco, CA        37798792        -122416797      11/18/15        2.9     $M      false   2       2.0     1465
.0  sqft    N/A     N/A     1963
15066036        1070 Green St APT 1402, San Francisco, CA       37798792        -122416797      12/31/15        2.6     $M      false   3       2.5     2017
.0  sqft    N/A     N/A     1963
                                                                                                                                                1,1           Top
```

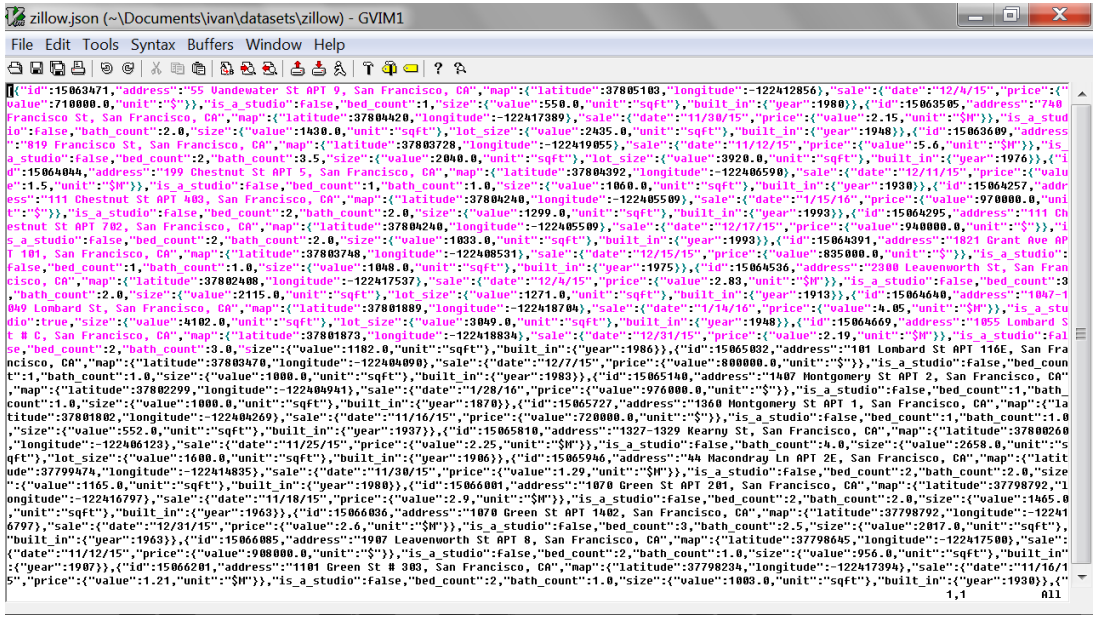
JSON, the most common open standard data format used for asynchronous browser/server communication, uses human-readable text to store data using key–value pairs and lists.  Unlike CVS/TSV format, data can be represented hierarchically

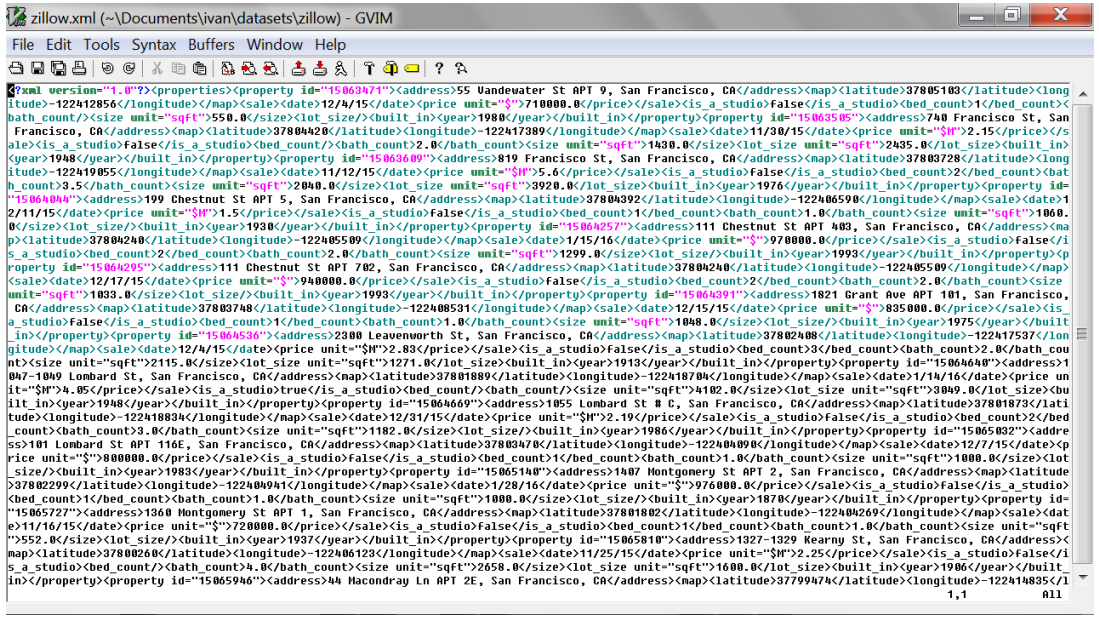


# XML, another data format used for asynchronous browser/server communication, also uses human-readable text to store hierarchical data



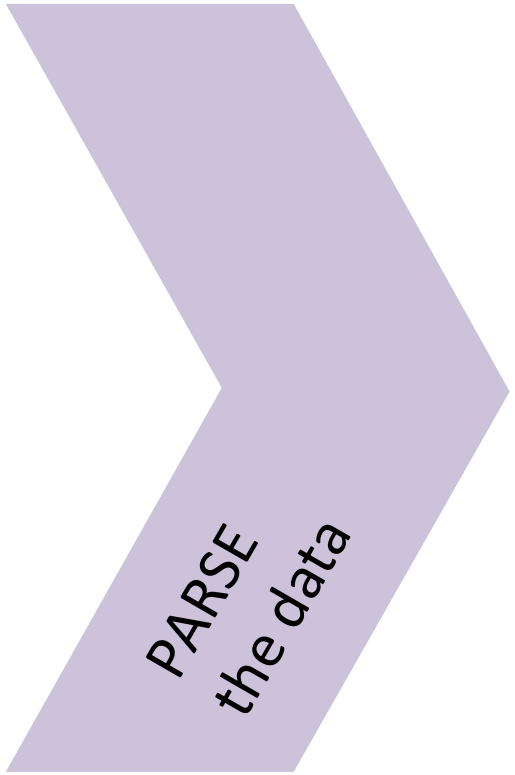
# JSON and XML are harder to read by humans when indentation is removed (usually the default) although it is still straightforward for machines…




# ❸ PARSE the Data

# ❸ Parse the Data




- Parse the Data
  - Read any documentation provided with the data (session 2)
  - Perform exploratory data analysis (session 3)
  - Verify the quality of the data (sessions 2/3)

# ❸ Parse the Data (cont.)

- Parse the Data
  - Read any documentation provided with the data
  - Perform exploratory data analysis
  - Verify the quality of the data
- Documentation and Data Dictionary
- Codealongs
  - Introduction to *pandas*
  - Tidying up (more) the Zillow dataset
- Lab

# ❸ Parse the Data (cont.)

- Parse the Data
  - Read any documentation provided with the data
  - Perform exploratory data analysis
  - Verify the quality of the data

- You need to understand what you're working with
- To better understand your data
  - Create or review the data dictionary
  - Perform exploratory surface analysis
  - Describe data structure and information being collected
  - Explore variables and data types

# Documentation and Data Dictionary

- Data dictionaries
  - Help you judge the quality of the data
  - Also help understand how it's coded
    - Does "`gender = 1`" mean female or male?
    - Is the currency dollars or euros?
  - Help identify any requirements, assumptions, and constraints of the data
  - Make it easier to share data

# Kaggle's Titanic Data Dictionary

EXAMPLE

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown;
                 S = Southampton)

SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5
```

```
With respect to the family relation variables (i.e.
sibsp and parch) some relations were ignored.  The
following are the definitions used for sibsp and
parch.

Sibling: Brother, Sister, Stepbrother, or
         Stepsister of Passenger Aboard Titanic
Spouse:  Husband or Wife of Passenger Aboard
         Titanic (Mistresses and Fiancés Ignored)
Parent:  Mother or Father of Passenger Aboard
         Titanic
Child:   Son, Daughter, Stepson, or Stepdaughter of
         Passenger Aboard Titanic

Other family relatives excluded from this study
include cousins, nephews/nieces, aunts/uncles, and
in-laws.  Some children travelled only with a nanny,
therefore parch=0 for them.  As well, some travelled
with very close friends or neighbors in a village,
however, the definitions do not support such
relations.
```


# Codealong: Introduction to *pandas*

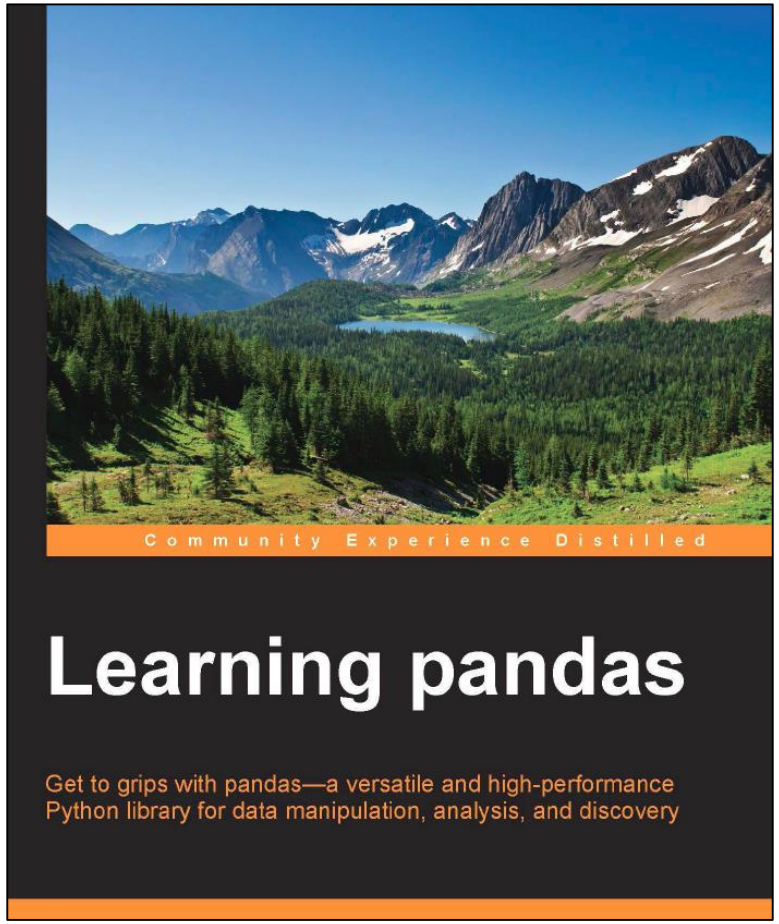
# *pandas* is a Python library to manipulate and perform statistical and mathematical analysis on tabular and multidimensional datasets

- *pandas* provides the ability to index, retrieve, tidy, reshape, combine, slice, and perform various analyses on both single and multidimensional data
- It also includes loading and saving data from local and Internet-based resources
- We will use *pandas* to explore and manipulate the Zillow dataset

# Resources

| | |
|---|---|
| http://pandas.pydata.org/pandas-docs/stable/tutorials.html | Guide to many pandas tutorials, geared mainly for new users. |
| https://github.com/jvns/pandas-cookbook | Great resource with examples from weather, bikes, and 311 calls from Julia Evans |
| https://bitbucket.org/hrojas/learn-pandas | Great series of Pandas tutorials from Dave Rojas |
| https://github.com/ResearchComputing/Meetup-Fall-2013/tree/master/python | Awesome set of python notebooks from a meetup-based course exclusively devoted to pandas |

# Going further




# Codealong: Tidying up the Zillow dataset


# Activity: Write a Data Dictionary for the Zillow dataset

# Activity: Write a Data Dictionary for the Zillow dataset

EXERCISE

## DIRECTIONS (15 minutes)

1. Divide into 4 groups, each located at a whiteboard

2. Data Dictionary: Modeling the previous example, each group should develop a data dictionary for the Zillow dataset (10 minutes)
   a. Highlight how the variables are coded as well as any requirements, assumptions, and constraints of the data

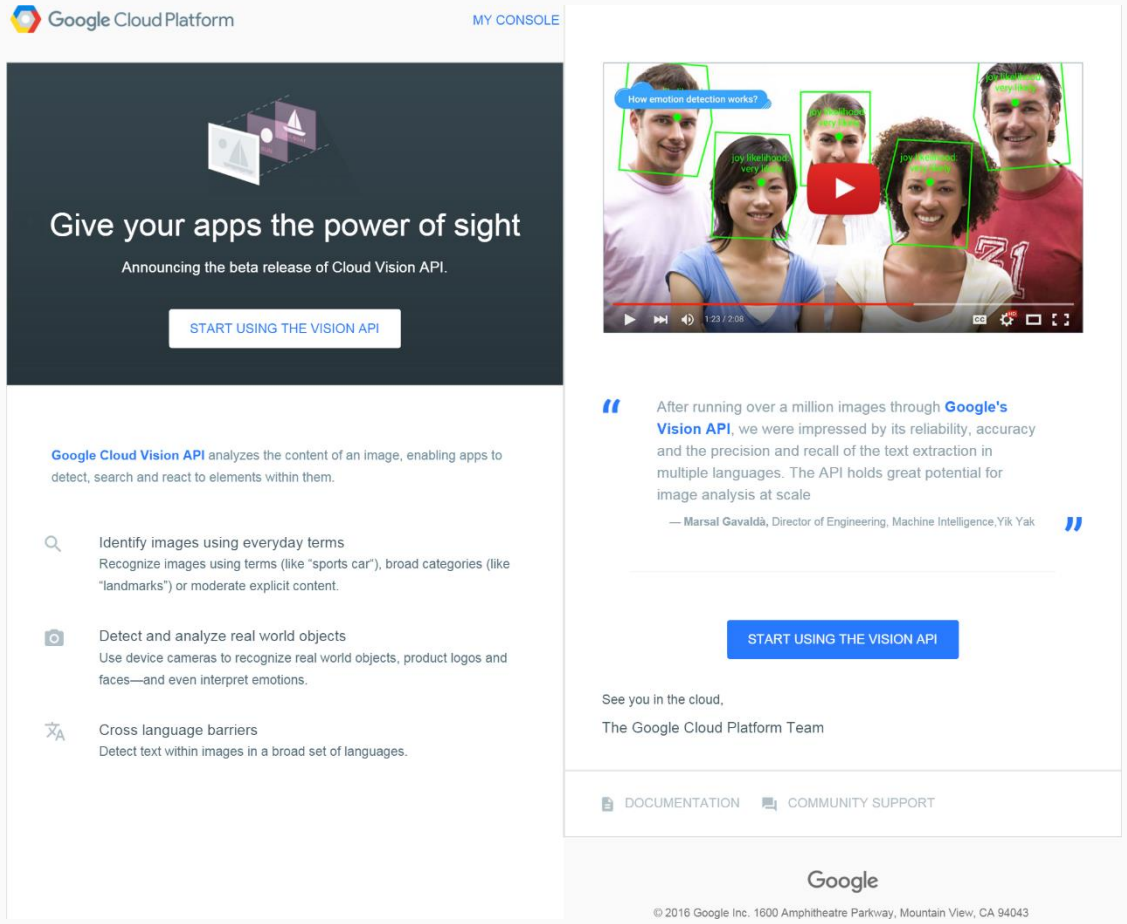
3. Present: Present your dictionary to the class (5 minutes)
   a. Choose one student to present for the group

## DELIVERABLE

Data Dictionary


# Lab

# Today's Closing Thought




# Review

# Review

You should now be able to:

- Define a problem and types of data
- Identify dataset types
- Define the data science workflow
- Apply the data science workflow in the pandas context
- Write an iPython notebook to import, format, and clean data using the Pandas library


# Q & A


# Before Next Class

# Before Next Class

- Reminder on the in-flight projects
  - **Unit Project 1 (due next time on 3/3)**
  - Final Project 1 (due 3 weeks from now on 3/22)


# Exit Ticket

*Don’t forget to fill out your exit ticket here*