GENERAL ASSEMBLY

EXPERIMENTAL DESIGN AND PANDAS

Sri Kanajan

LEARNING OBJECTIVES

- Apply the data science workflow in the pandas context

- Create an iPython Notebook to import, format, and clean using the Pandas library

COURSE

PRE-WORK

PRE-WORK REVIEW

- Create and open an iPython Notebook
- Complete the Python pre-work

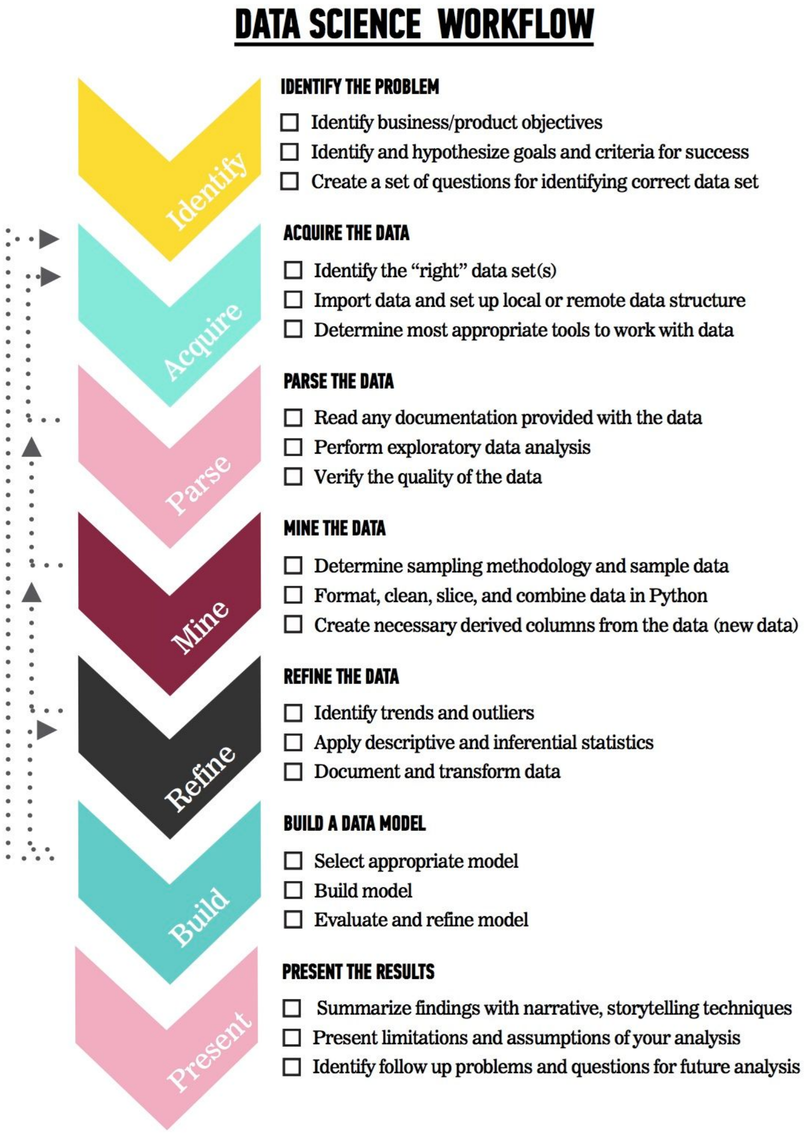
INTRODUCTION

DATA SCIENCE WORKFLOW: ACQUIRE & PARSE

OVERVIEW OF THE DATA SCIENCE WORKFLOW

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

- □ Identify business/product objectives
- □ Identify and hypothesize goals and criteria for success
- □ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

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

PARSE THE DATA

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

MINE THE DATA

- □ Determine sampling methodology and sample data
- □ Format, clean, slice, and combine data in Python
- □ Create necessary derived columns from the data (new data)

REFINE THE DATA

- □ Identify trends and outliers
- □ Apply descriptive and inferential statistics
- □ Document and transform data

BUILD A DATA MODEL

- □ Select appropriate model
- □ Build model
- □ Evaluate and refine model

PRESENT THE RESULTS

- □ Summarize findings with narrative, storytelling techniques
- □ Present limitations and assumptions of your analysis
- □ Identify follow up problems and questions for future analysis

DATA SCIENCE WORKFLOW: ACQUIRE & PARSE

- We’ll talk about steps 2 & 3 of the data science workflow: acquire and parse

- We’ll be using iPython Notebook

- First a demo, then a codealong

- Finally, some hands on practice in a lab

DEMO

WALKTHROUGH ACQUIRE & PARSES WITH PANDAS

ACQUIRE

- Where we determine if we have the “right” dataset for our problem
- Questions to ask:
    - What type of data is it, cross-sectional or longitudinal? (time dependent vs. snapshot)
    - How well was the data collected?
    - Is there much missing data?
    - Was the data collection instrument validated and reliable?
    - Is the dataset aggregated?
    - Do we need pre-aggregated data?

LOGISTICS OF ACQUIRING YOUR DATA

- Data can be acquired through a variety of sources
- Web (Google Analytics, HTML, XML)
- File (CSV, XML, TXT, JSON)
- Databases (SQL, NOSQL, etc)
- Today, we’ll use a CSV (comma separated file)

PARSE: UNDERSTANDING YOUR DATA

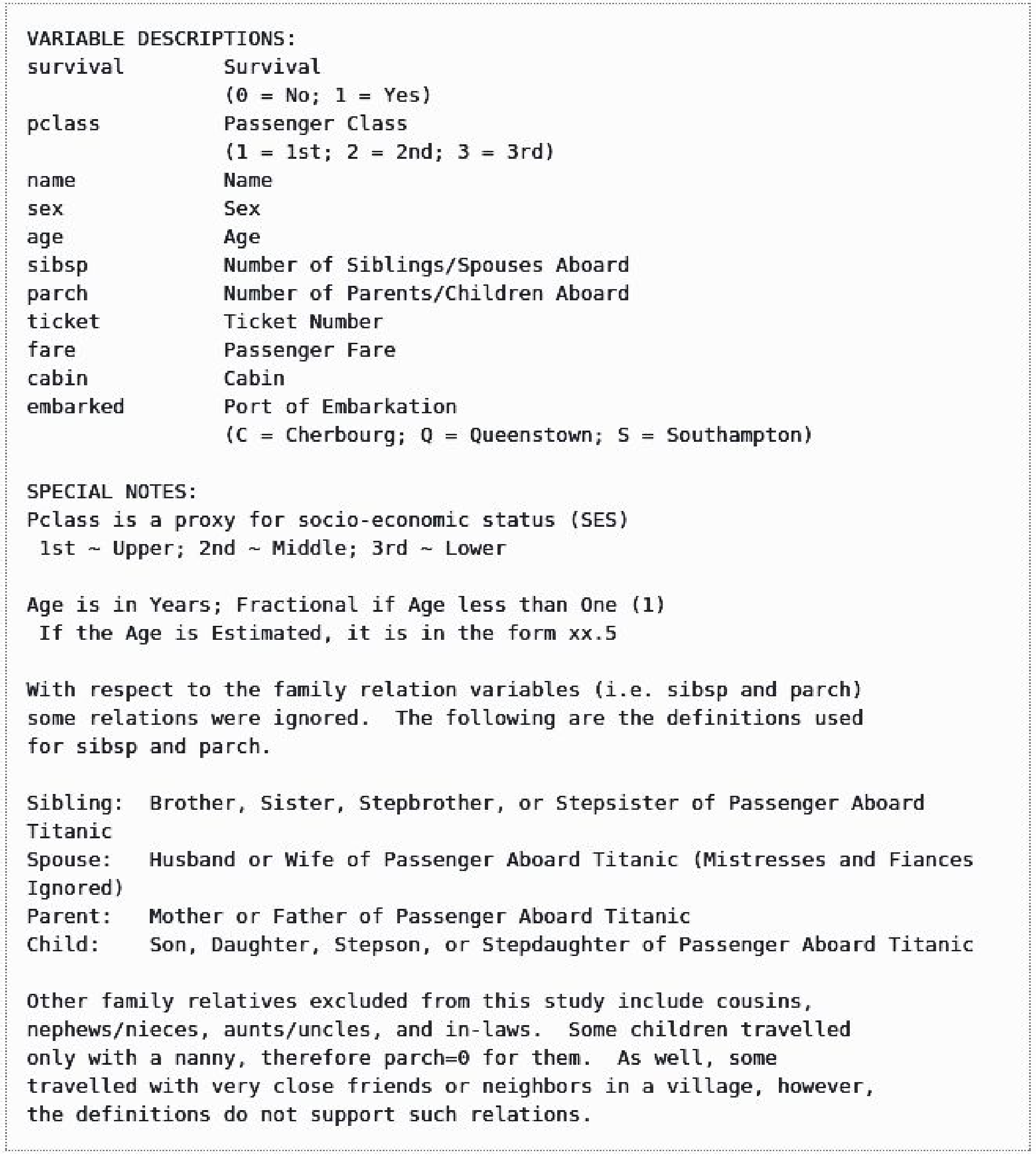
- You need to understand what you’re working with.
- To better understand your data
  - Create or review the data dictionary
  - Perform exploratory surface analysis
  - Describe data structure and information being collected
  - Explore variables and data types

INTRO TO DATA DICTIONARIES AND DOCUMENTATION

- Data dictionaries help judge the quality of the data.
- They also help understand how it’s coded.
  - Does gender = 1 mean female or male?
  - Is the currency dollars or euros?
- Data dictionaries help identify any requirements, assumptions, and constraints of the data.
- They make it easier to share data.

DATA DICTIONARY EXAMPLE: KAGGLE TITANIC DATA

https://www.kaggle.com/c/titanic/data



```
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 Fiances
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.
```

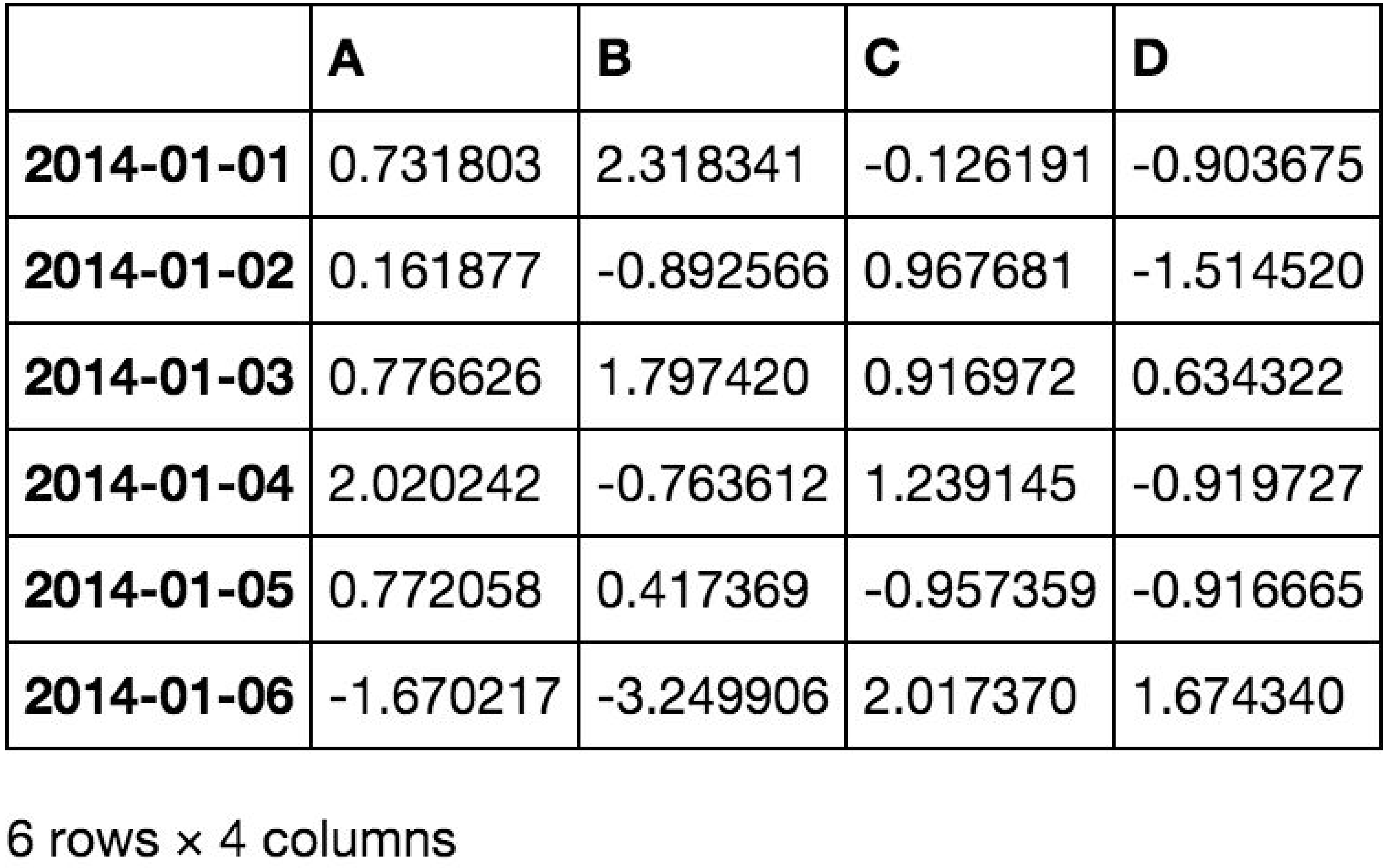
CODEALONG

NUMPY AND PANDAS INTRO

NUMPY AND PANDAS INTRO

- What are Numpy and Pandas? Python packages
- Pands is built on Numpy.
- Numpy uses arrays (lists) to do basic math and slice and index data.
- Pandas uses a data structure called a Dataframe.
- Dataframes are similar to Excel tables; they contain rows and columns.

NUMPY AND PANDAS INTRO



	A	B	C	D
2014-01-01	0.731803	2.318341	-0.126191	-0.903675
2014-01-02	0.161877	-0.892566	0.967681	-1.514520
2014-01-03	0.776626	1.797420	0.916972	0.634322
2014-01-04	2.020242	-0.763612	1.239145	-0.919727
2014-01-05	0.772058	0.417369	-0.957359	-0.916665
2014-01-06	-1.670217	-3.249906	2.017370	1.674340

6 rows × 4 columns

NUMPY AND PANDAS INTRO

- With these packages, you can select pieces of data, do basic operations, calculate summary statistics.
- Follow along and code along as we learn about Numpy and Pandas.

NUMPY AND PANDAS INTRO

- We often have to merge data together, correct missing data, and plot our findings.

- Once again, follow and code along.

DEMO

LAB WALKTHROUGH

LESSON 2 LAB WALKTHROUGH

- In this lab, you will merge two datasets: ozone and data.
- By the end of the lab, you will:
    - Merge datasets
    - Check basic features of the data
    - Find and drop missing values
    - Find basic stats like mean and max

CONCLUSION

TOPIC REVIEW

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

REVIEW

- Let’s go through the lab. Any questions?
- Today, we’ve talked about
  - Defining a problem
  - Types of data
  - Acquiring and parsing data
  - Using Pandas