PA 446

Coding for Civic Data Applications

Class will start at 6:05pm Central

Class #2

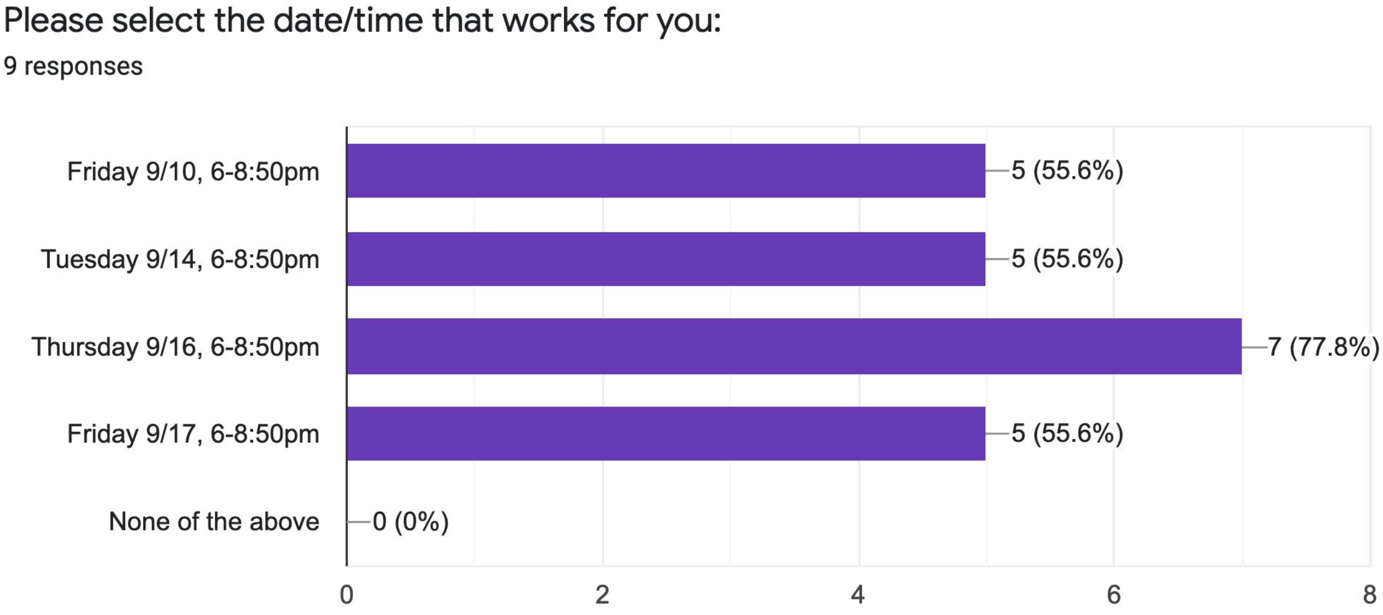
Logistics

Course Logistics

- HW1, due at 6pm today (a few minutes ago)
  - If you forgot, please submit now - this is the only time you are allowed to submit late

- Accessing course content: make sure you are using your uic.edu Google account
  - 2 individuals has notified me that Google isn't recognizing their accounts - we will need to work with UIC IT to sort this out. Please let me know if you are having similar issues
  - Email is fine

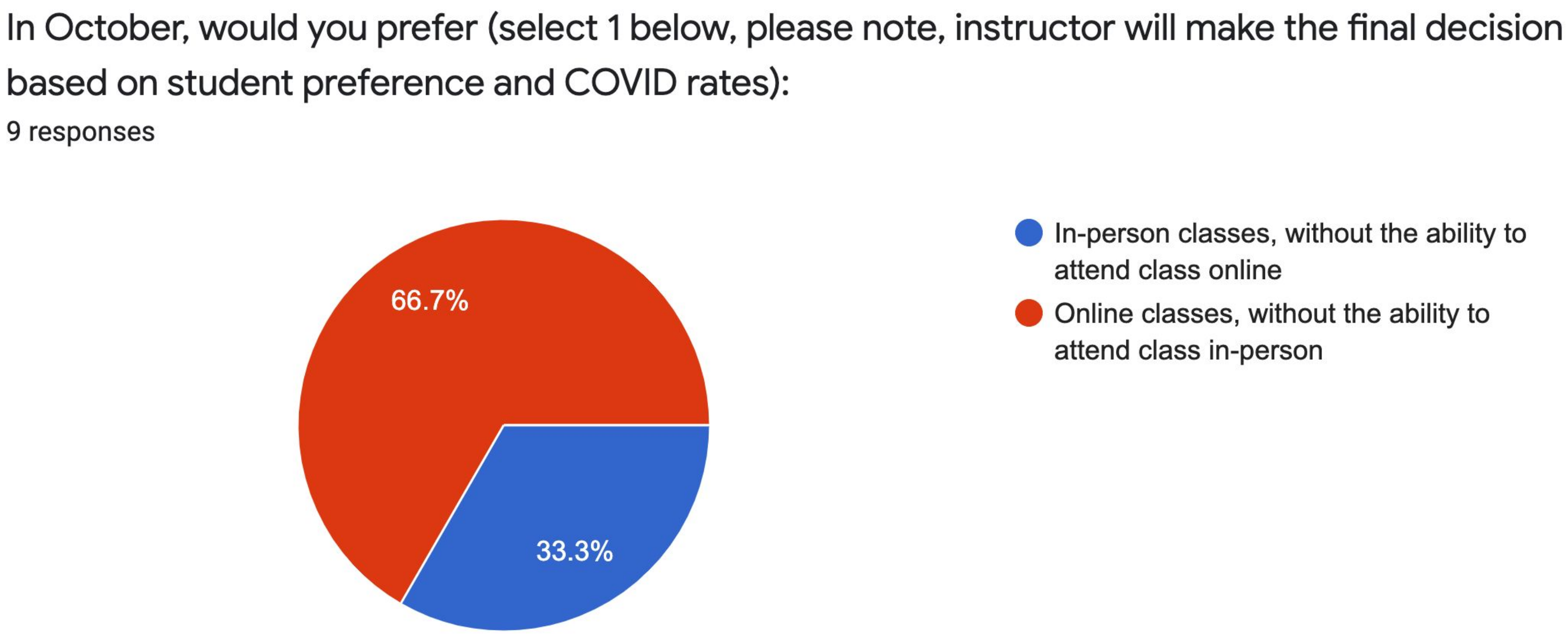
Poll Results (NOT FINAL)



Please select the date/time that works for you:

9 responses

Poll Results (NOT FINAL)



In October, would you prefer (select 1 below, please note, instructor will make the final decision based on student preference and COVID rates):

9 responses

Poll Results

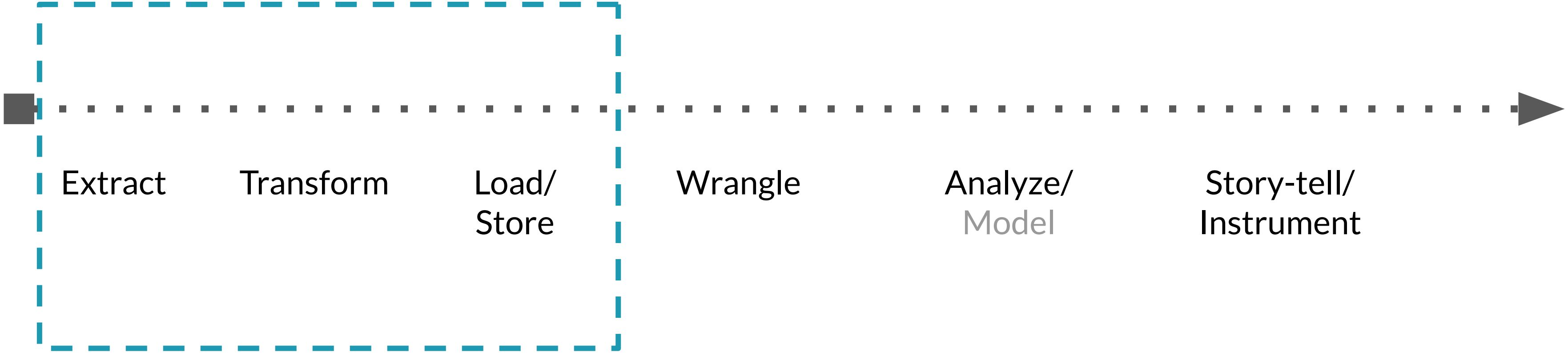
- Some permission issues linger
- Final results for 9/15 class will be emailed
- In-person vs online will be settle by end of Sept

Class #2

The HW

- Reactions
- Will be graded by next Wednesday

Data Science "workflow"



Extract

Transform

Load/
Store

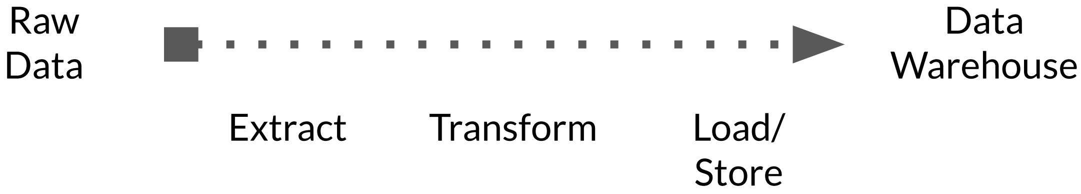
Wrangle

Analyze/
Model

Story-tell/
Instrument

Focus last week

Data Engineering



Raw Data

Extract

Transform

Load/ Store

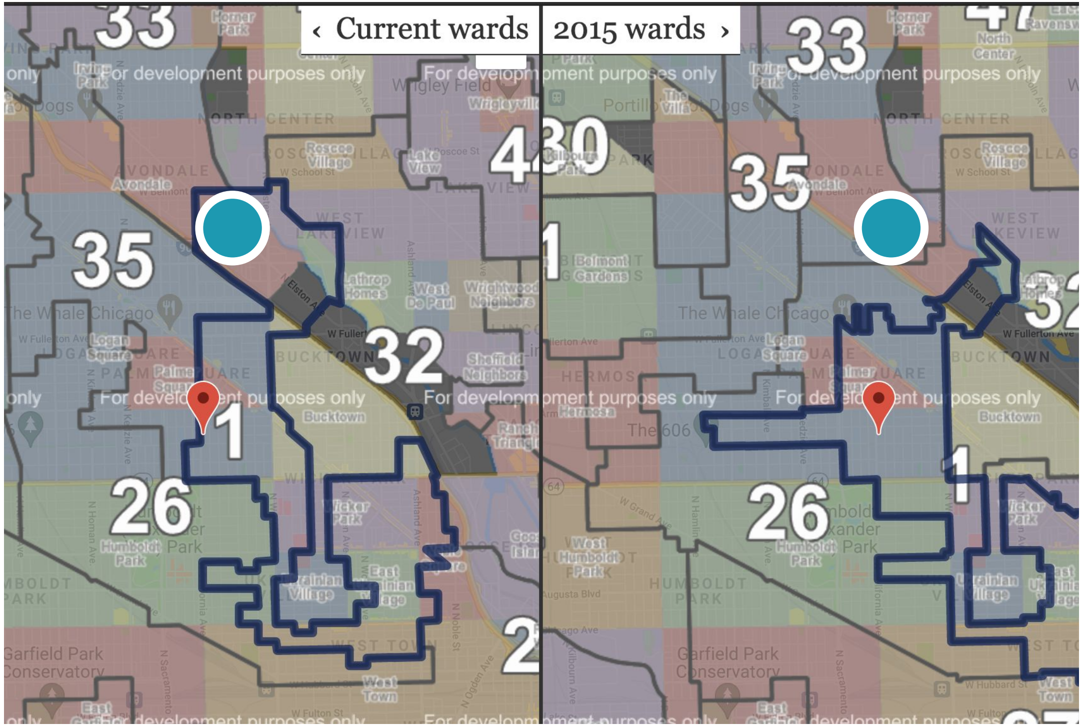
Data Warehouse

What can go wrong here?

Last Week

CREATED_DATE	OWNER_DEPARTMENT	STREET_ADDRESS	STATUS	WARD	LAT	LONG
08/22/2021 9:14:44 AM	311 City Services	2111 W Lexington ST	Completed	28	41.875	-87.739
08/22/2021 9:14:41 AM	Streets and Sanitation	402 S KOLMAR AVE	Open	10	41.699	-87.545
08/22/2021 9:14:01 AM	**311 City Services**	**2111 W Lexington ST**	**Open**	**1**	**41.914**	**-87.690**
08/22/2021 9:13:51 AM	311 City Services	2111 W Lexington ST	Open	48	41.980	-87.668

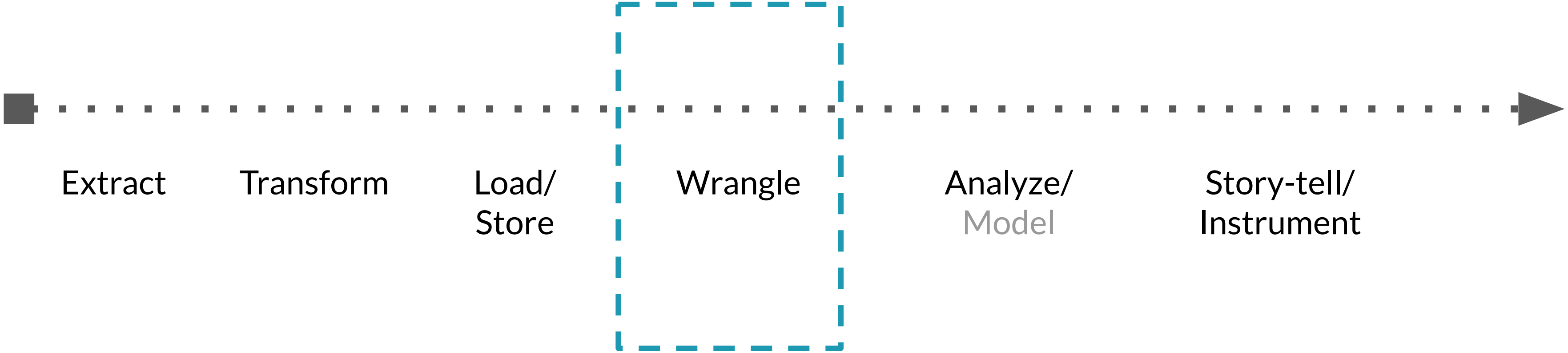
Business Rules Changed



Data Engineering Follow Up

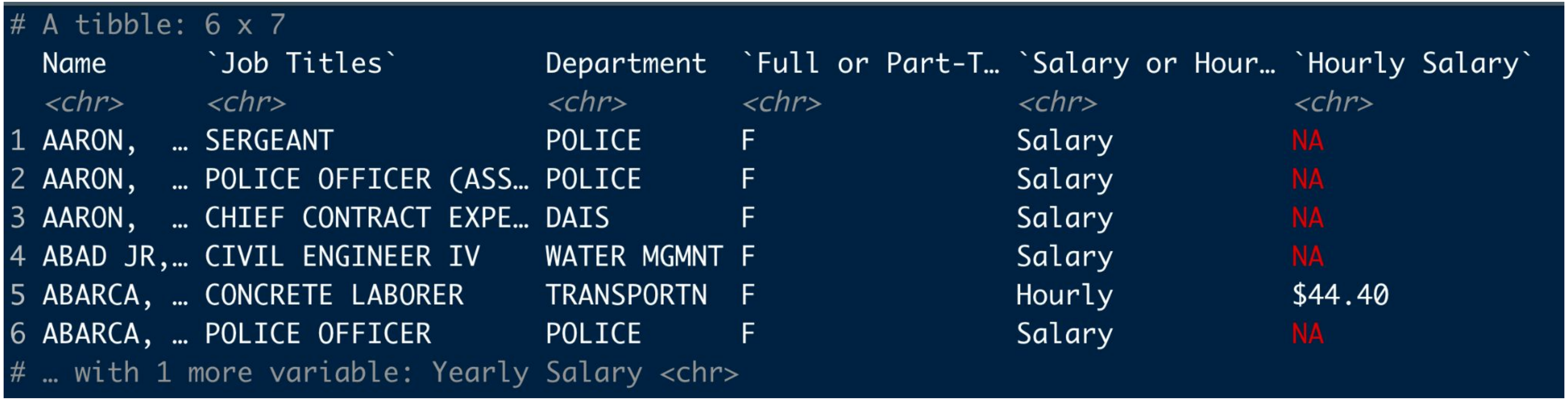
	Data Warehouse	Data Lake (very vague term)
Data Structure	Processed	Raw
Purpose of Data	Currently In Use	Not Yet Determined
Users	Business Professionals	Data Scientists
Flexibility	Low	High
Storage Need	Low(er)	High(er)

Data Science “workflow”



Focus of today and the next few weeks

From HW 1



```
# A tibble: 6 x 7
  Name      `Job Titles`         Department  `Full or Part-T… `Salary or Hour… `Hourly Salary`
  <chr>     <chr>                <chr>       <chr>            <chr>            <chr>
1 AARON,  … SERGEANT             POLICE      F                Salary           NA
2 AARON,  … POLICE OFFICER (ASS… POLICE      F                Salary           NA
3 AARON,  … CHIEF CONTRACT EXPE… DAIS        F                Salary           NA
4 ABAD JR,… CIVIL ENGINEER IV    WATER MGMNT F                Salary           NA
5 ABARCA, … CONCRETE LABORER     TRANSPORTN  F                Hourly           $44.40
6 ABARCA, … POLICE OFFICER       POLICE      F                Salary           NA
# … with 1 more variable: Yearly Salary <chr>
```

But Before We Start...

What Exactly Are Our Goals?

What Exactly Are Our Goals?

“One of the mayor of Chicago's priority this year is equity in pay for city employees, especially in some of the city's largest departments.”

Clarify Objective

What is the measurable metric of success?

What is the time frame?

What are potential restrictions or limitations?

Example questions

Which column of the data are we most interested in?

Do we only care about this year or prior years?

What kind of data do we even have access to?

Strengthens Your Understanding of the Company

Business model (less relevant)

Products and services

Geographic location (less relevant)

› Company Information

What are Chicago's largest departments?

Which departments might have higher salaries based on what they do?

Definition of an Unfamiliar Term

› Clarify Unfamiliar Term

What terminology should you clarify?

When defining goals, you can’t afford to misunderstand or misinterpret terminologies used. Confirm your understanding aligns with the stakeholders’

What Are Some of Our Goals?

1.
2.
3.
4.

Data Wrangling

1. Before You Start Wrangling

What are you considerations?

1.
2.
3.
4.
5.

1. Before You Start Wrangling

1. Data size

2. Just look at your data

1. Before You Start Wrangling

“Data Skimming”

1. Data size
    a. Do you have enough data?
    b. What is the best way to browse this data

2. Just look at your data
    a. Building familiarity with the data

2. “Basic Descriptions of Your Sample and Features”

2. “Basic Descriptions of Your Sample and Features”

What Does “Basic Descriptions” Mean to You?

1.
2.
3.
4.
5.

2. “Basic Descriptions of Your Sample and Features”

1. Really just means check data types

2. “Basic Descriptions of Your Sample and Features”

1. Really just means check data types
    a. Which columns do you need to wrangle
    b. How to get more details about each column

2. “Basic Descriptions of Your Sample and Features”

Reminders

1. Calling individual columns
2. Tibble vs dataframe

10-min break

Be back by 7:10pm Central

3. Check For Missing Data

3. Check For Missing Data

But ... what is exactly is “Missing Data”?

3. Check For Missing Data

1. Value is missing
2. Indeterminacy (neither true of false)

3. Check For Missing Data

1. Value is missing: NULL
2. Indeterminacy (neither true of false): NA

3. Check For Missing Data

Used a lot on the internet, especially in the context of SQL

Not as often in R, especially not in the context of dataframes and vectors

1. Value is missing: NULL
2. Indeterminacy (neither true of false): NA

3. Check For Missing Data

What happens when you try to insert NULL’s into a dataframe?

What happens when you try to count NULL’s?

1. Value is missing: NULL
2. Indeterminacy (neither true of false): NA

3. Check For Missing Data

1. Value is missing: NULL
2. Indeterminacy (neither true of false): NA

What happens when you try to insert NA’s into a dataframe?

What happens when you try to count NA’s?

3. Check For Missing Data

1. Value is missing: NULL
2. Indeterminacy (neither true of false): NA

So for any vector (matrix or array), NA represents a missing value. NULL does not

Most commonly used in R dataframes

3. Check For Missing Data

1. Value is missing: NULL
2. Indeterminacy (neither true of false): NA

Beyond missing values and indeterminacy, what else can missing mean?

3. Check For Missing Data

1.
2.
3.
4.

3. Check For Missing Data

What data do we need?

1.
2.
3.
4.

What do we already have?

1.
2.
3.
4.

3. Check For Missing Data

1. Missing “metadata” and documentation

2. Missing columns

3. Check For Missing Data

1. Missing “metadata” and documentation
    a. What a column means
    b. How was a column calculated

2. Missing columns
    a. Is there any data that we need that appear to be missing?

4. Identify The Shape of Your Data

4. Identify The Shape of Your Data

But ... what is exactly is “Shape of Your Data”?

4. Identify The Shape of Your Data

1.
2.
3.
4.

4. Identify The Shape of Your Data

1. Max/min
2. Mean/Median

We will cover distribution plots at a later class

Skipped Steps

Skipped Steps

- Identify Significant Correlations
  - Requires a specific and cleaned data type
  - More useful in modeling context

- Spot Outliers in the Dataset
  - Redundant to shape of your data

In Summary

Wrangling Steps Review

Takeaways

1. Clarifying Goals
2. “Data skimming”: count of rows and columns + just looking at the data
3. Check data types
4. Check for missing data
5. Check the shape of your data