LECTURE 7

Data Cleaning and EDA

Exploratory data analysis and its role in the data science lifecycle.

Sean Kang

Review Pandas and Jupyter Notebooks, Data Structures, Data Types

- Reviewing DataFrame concepts
  - **Series**: A named column of data with an index
  - **Indexes**: The mapping from keys to rows
  - **DataFrame**: collection of series with common index

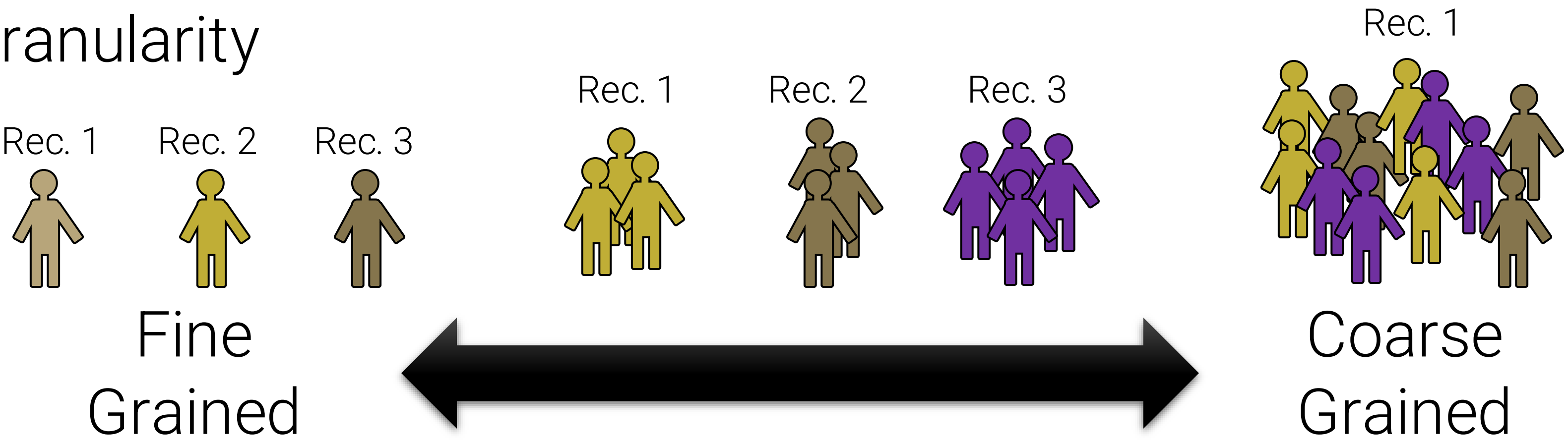
- Dataframe access methods
  - **Filtering** on predicts and **slicing**
  - **df.loc**: location by index
  - **df.iloc**: location by integer address
  - **groupby** & **pivot** aggregating data

Granularity, Scope, and Temporality

Key Data Properties to Consider in EDA

- **Structure --** *the “shape” of a data file – how many columns, how many rows?*
- **Granularity --** *how fine/coarse is each datum – is it actual record, or summary*
- **Scope --** *how (in)complete is the data – all 100% of US citizens in US Census?*
- **Temporality --** *how is the data situated in time – The data might be just for 2010*
- **Faithfulness --** *how well does the data capture “reality” – was the census answered truthfully?*

Granularity



- What does each record represent?
  - Examples: a purchase, a person, a group of users
- Do all records capture granularity at the same level?
  - Some data will include summaries (aka rollups) as records
- If the data are coarse how was it aggregated?
  - Sampling, averaging, …

Scope

- Does my data cover my area of interest?
    - **Example:** *I am interested in studying crime in California but I only have Berkeley crime data.*

- Is my data too expansive?
    - **Example:** *I am interested in student grades for DS100 but have student grades for all statistics classes.*
    - **Solution:** *Filtering ⇒ Implications on sample?*
        - *If the data is a sample I may have poor coverage after filtering …*

- Does my data cover the right time frame?
    - More on this in temporality …

Temporality

- Data changes – when was the data collected?
- What is the meaning of the time and date fields?
- When the event happened
- When the data was collected or recorded?
- The data was copied into the database
- Time depends on where – timezone, daylight savings
- Learn to use datetime in python
- Various string representation of the date (YY/MM/dd)
- Are there any null or or strange values, or number of milliseconds since Epoch
- Periodicity patterns – max or min temperature of a day

Unix Time / POSIX Time

- Time **measured in seconds** since January 1st 1970
  - Minus leap seconds …
- Unix time follows Coordinated Universal Time (UTC)
  - International time standard
  - Measured at 0 degrees latitude
    - Similar to Greenwich Mean Time (GMT)
  - No daylight savings
  - Time codes
- Time Zones:
  - San Francisco (UTC-8)
    without daylight savings

https://en.wikipedia.org/wiki/Coordinated_Universal_Time

Faithfulness and Missing Values

Faithfulness: *Do I trust this data?*

- Does my data contain **unrealistic** or **“incorrect”** values?
  - Dates in the future for events in the past
  - Locations that don’t exist
  - Negative counts
  - Misspellings of names
  - Large outliers
- Does my data violate **obvious dependencies**?
  - E.g., age and birthday don’t match
- Was the data **entered by hand**?
  - Spelling errors, fields shifted …
  - Did the form require fields or provide default values?
- Are there obvious signs of **data falsification**:
  - Repeated names, fake looking email addresses, repeated use of uncommon names or fields. (joe@universe.com)

Signs that your data may not be faithful

- Missing Values/Default values?
  - What do they look like?
    - “   “,
    - 0,
    - -1, 999, 12345,
    - NaN, Null, (NaN – Not a Number)
    - 1970, 1900

More signs that your data may not be faithful

- **Missing** Values or **default** values
- Truncated data (early excel limits: 65536 Rows, 255 Columns)
  - **Soln:** be aware of consequences in analysis ⇒ how did truncation affect sample?
- Time Zone Inconsistencies
  - **Soln 1:** convert to a common timezone (e.g., UTC)
  - **Soln 2:** convert to the timezone of the location – useful in modeling behavior.
- Duplicated Records or Fields
  - **Soln:** identify and eliminate (use primary key) ⇒ implications on sample?
- Spelling Errors
  - **Soln:** Apply corrections or drop records not in a dictionary ⇒ implications on sample?
- Units not specified or consistent
  - **Solns:** Infer units, check values are in reasonable ranges for data
- Others…

What to do with the Missing Values?

- **Drop records** with missing values
    - Probably most common
    - **Caution:** check for biases introduced by dropped values
        - Missing or corrupt records might be related to something of interest

- **Imputation:** (Inferring missing values)
    - **Mean Imputation:** replace with an average value
        - Which mean? Often use closest related subgroup mean.
    - **Hot deck imputation:** replace with a random value
        - Choose a random value from the subgroup and use it for the missing value. (See note for details article. Contents of that article are not covered in exam, but this desk will be covered.)

Revisiting the Sampling Frame

- The **sampling frame** is the **population** from which the data was **sampled.**
    - Note that this **may not be** the **population** of interest.

- How complete/incomplete is the frame (and its data)?

- How is the frame/data situated in place?

- How well does the frame/data capture reality?

- How is the frame/data situated in time?