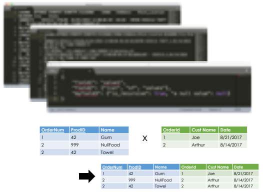
Data 100
Lecture 5: Data Cleaning & Exploratory Data Analysis

Slides by:
Joseph E. Gonzalez, Deb Nolan, & Joe Hellerstein
jegonzal@berkeley.edu
deborah_nolan@berkeley.edu
hellerstein@berkeley.edu

Last Lecture

- Started discussing exploratory data analysis
- **Structure** -- *the "shape" of a data file (how is it organized)*



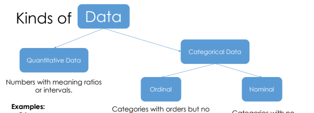
OrderNum	ProdID	Name
1	42	Gum
2	999	NullFood
2	42	Towel

X

OrderId	Cust Name	Date
1	Joe	8/21/2017
2	Arthur	8/14/2017

➡

OrderNum	ProdID	Name	OrderId	Cust Name	Date
1	42	Gum	1	Joe	8/21/2017
2	999	NullFood	2	Arthur	8/14/2017
2	42	Towel	2	Arthur	8/14/2017



Kinds of Data

Quantitative Data

Numbers with meaning ratios or intervals.

Examples:
- Price
- Quantity
- Temperature
- Date
- …

Categorical Data

Ordinal

Categories with orders but no consistent meaning if magnitudes or intervals

Examples:
- Preferences
- Level of education
- …

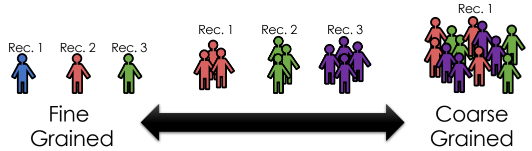
Nominal

Categories with no specific ordering.

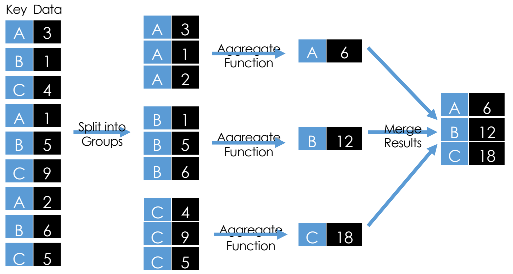
Examples:
- Political Affiliation
- Product Type
- Cal Id
- …

Last Lecture

- Started discussing exploratory data analysis
- **Structure** -- *the "shape" of a data file (how is it organized)*
- ***Granularity*** -- *how fine/coarse is each datum*



Group By – manipulating granularity



Key	Data
A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

Split into Groups

A	3
A	1
A	2

B	1
B	5
B	6

C	4
C	9
C	5

Aggregate Function → A 6

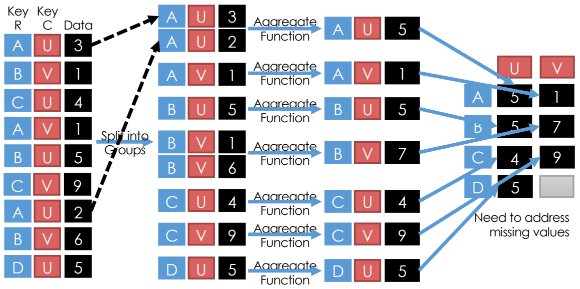
Aggregate Function → B 12

Aggregate Function → C 18

Merge Results

A	6
B	12
C	18

Pivot – A kind of Group By Operation



Key R	Key C	Data
A	U	3
B	V	1
C	U	4
A	V	1
B	U	5
C	V	9
A	U	2
B	V	6
D	U	5

Split into Groups

A U 3, A U 2 → Aggregate Function → A U 5

A V 1 → Aggregate Function → A V 1

B U 5 → Aggregate Function → B U 5

B V 1, B V 6 → Aggregate Function → B V 7

C U 4 → Aggregate Function → C U 4

C V 9 → Aggregate Function → C V 9

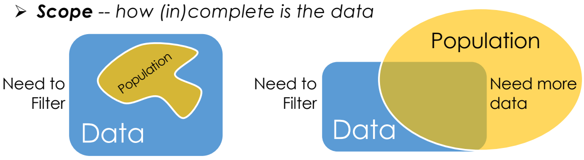
D U 5 → Aggregate Function → D U 5

	U	V
A	5	1
B	5	7
C	4	9
D	5	

Need to address missing values

Last Lecture

- Started discussing exploratory data analysis
- **Structure** -- *the "shape" of a data file (how is it organized)*
- ***Granularity*** -- *how fine/coarse is each datum*
- ***Scope*** -- *how (in)complete is the data*



Last Lecture

- Started discussing exploratory data analysis
- **Structure** -- *the "shape" of a data file (how is it organized)*
- ***Granularity*** -- *how fine/coarse is each datum*
- ***Scope*** -- *how (in)complete is the data*
- ***Temporality*** -- *how is the data situated in time*

Temporality

- Data changes → When was the data collected!
- What is the meaning of a the time and date fields?
  - When the "event" **happened**?
  - When the data was **collected** or was **entered** into the system?
  - Date the data was copied into a database (look for many matching timestamps)
- Time depends on where! (Time zones & daylight savings)
  - Learn to use **datetime** python library
  - Multiple string representation (depends on region): 07/08/09?
- Are there strange null values?
  - January 1st 1970, January 1st 1900
- Is there periodicity? Diurnal patterns

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

Key Data Properties to Consider in EDA

- **Structure --** *the "shape" of a data file*
- **Granularity --** *how fine/coarse is each datum*
- **Scope --** *how (in)complete is the data*
- **Temporality --** *how is the data situated in time*
- **Faithfulness --** *how well does the data capture "reality"*

Key Data Properties to Consider in EDA

- **Structure --** *the "shape" of a data file*
- **Granularity --** *how fine/coarse is each datum*
- **Scope --** *how (in)complete is the data*
- **Temporality --** *how is the data situated in time*
- **Faithfulness --** *how well does the data capture "reality"*

Faithfulness: *Do I trust this data?*

- Does my data contain unrealistic or "incorrect" values?
  - Examples?
    - 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 curb stoning (data falsification):
  - Repeated names, fake looking email addresses, repeated use of uncommon names or fields.

Signs that your data may not be faithful

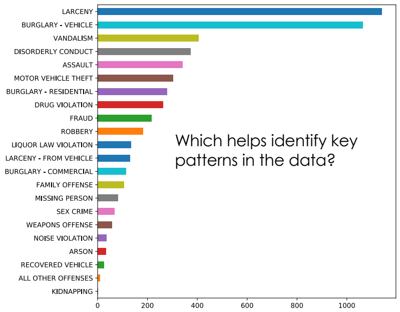
- Missing Values/Default values: (0, -1, 999, 12345, NaN, Null, 1970, 1900, … others?)
  - **Soln 1:** Drop records with missing values → implications on your sample!
  - **Soln 2:** Impute missing values → Bias your conclusions
- Time Zone Inconsistencies
  - **Soln 1:** convert to a common timezone (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
- Truncated data (early excel limits: 65536 Rows, 255 Columns)
  - **Soln:** be aware of consequences in analysis → how did truncation affect sample?
- Others…

How do you do EDA?

- Examine data and meta-data:
  - What is the date, size, organization, and structure of the data?
- Examine each field/attribute/dimension individually
- Examine pairs of related dimensions
  - Stratifying earlier analysis: break down grades by major …
- Along the way:
  - Visualize/summarize the data
  - Validate assumptions about data and collection process
  - Identify and address anomalies
  - Apply data transformations and corrections
  - ***Record everything you do! (why?)***

Visualization and EDA

Berkeley Crime Data



Which helps identify key patterns in the data?

```
ALL OTHER OFFENSES            10
ARSON                         33
ASSAULT                      341
BURGLARY - COMMERCIAL        113
BURGLARY - RESIDENTIAL       278
BURGLARY - VEHICLE          1064
DISORDERLY CONDUCT           372
DRUG VIOLATION               262
FAMILY OFFENSE               105
FRAUD                        217
KIDNAPPING                     1
LARCENY                     1140
LARCENY - FROM VEHICLE       130
LIQUOR LAW VIOLATION         134
MISSING PERSON                81
MOTOR VEHICLE THEFT          303
NOISE VIOLATION               35
RECOVERED VEHICLE             26
ROBBERY                      182
SEX CRIME                     67
VANDALISM                    404
WEAPONS OFFENSE               58
Name: CVLEGEND, dtype: int64
```

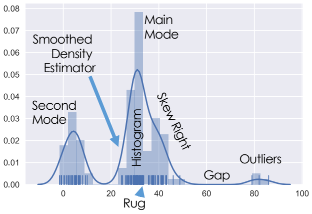
Visualizing Univariate Relationships

- **Quantitative Data**
  - Histograms, Box Plots, Rug Plots, Smoothed Interpolations (KDE – Kernel Density Estimators)
  - Look for spread, shape, modes, outliers, unreasonable values …
- **Nominal & Ordinal Data**
  - Bar plots (sorted by frequency or oridinal dimension)
  - Look for skew, frequent and rare categories, or invalid categories
  - Consider grouping categories and repeating analysis

Histograms, Rug Plots, and KDE Interpolation

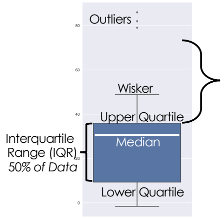
Describes distribution of data – relative prevalence of values

- Histogram
  - relative frequency of values
  - Tradeoff of bin sizes
- Rug Plot
  - Shows the actual data locations
- Smoothed density estimator
  - Tradeoff of "bandwidth" parameter (more on this later)



Box Charts

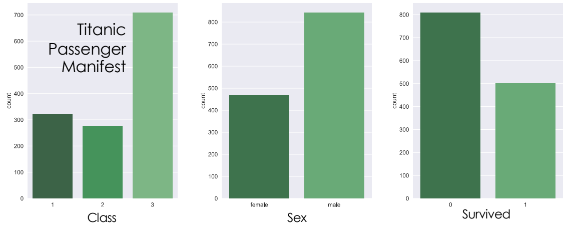
- Useful for summarizing distributions and comparing multiple distributions



Outliers are 1.5 * IQR away from lower and upper quartiles.

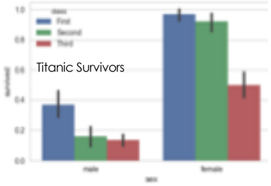
Bar Charts

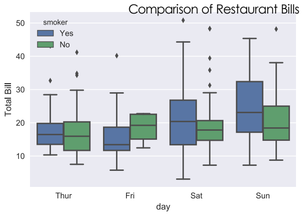
- Used to compare nominal and ordinal data.
  - Consider sorting by category or frequency



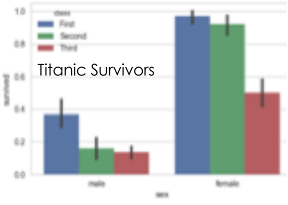
Visualizing Multivariate Relationships

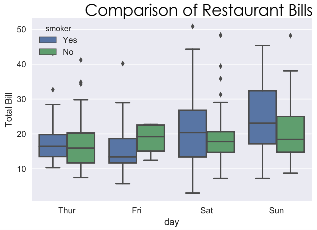
- Conditioning on a range of values (e.g., ages in groups) and construct side by side box-plots or bar charts





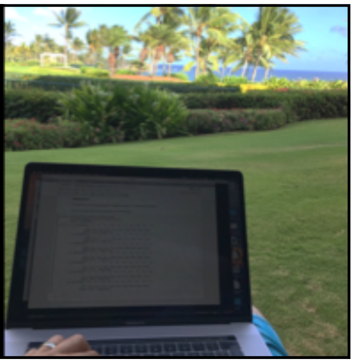
http://bit.ly/ds100-sp18-eda





Quick Break

Quick Break



Scope:
Do you have a full picture?

Berkeley Police Data Demo

Berkeley Police Public Datasets

- **Question:** For this analysis we will not begin with a detailed question but instead a rough goal of understanding Police activity.
- **Examine Two Data Sets:**
  - Call data
  - Stop data
- Today we will work through the basic process of data loading, some preliminary cleaning, and exploratory data analysis.

Call Data Description

Data pulled from Public Safety Server using data created for Berkeley's Crime View Community page. Displays **incidents reported** for **the last 180 days** along with **time**, **date**, **day of week** and **block level location information.**

The dataset reflects crimes as they have been reported to the BPD based on preliminary information **supplied by the reporting parties**. Preliminary crime classifications may change based on follow-up investigations. **Not all calls for police service are included (e.g. Animal Bite).** The information provided on this site is intended for use by the community to enhance their awareness of crimes occurring in their neighborhoods and the entire City. **The data should not be used for in-depth crime analysis** as the initial information is subject to change.

Stops Data Description

This data was extracted from the Department's Public Safety Server and covers the **data beginning January 26, 2015**. On January 26, 2015 the department began collecting data pursuant to General Order B-4 (issued December 31, 2014). Under that order, **officers were required to provide certain data after making all vehicle detentions** (including bicycles) and pedestrian detentions (up to five persons). This data **set lists stops by police** in the categories of traffic, suspicious vehicle, pedestrian and bicycle stops. Incident number, date and time, location and disposition codes are also listed in this data.

Address **data has been changed from a specific address**, where applicable, and listed as the block where the incident occurred. Disposition codes were entered by officers who made the stop. These codes included the person(s) race, gender, age (range), reason for the stop, enforcement action taken, and whether or not a search was conducted.

Caution about EDA

With enough data, if you look hard enough you will find something ***"interesting"***

Important to differentiate **inferential conclusions** about world from **exploratory analysis of data**