



CS 329P: Practical Machine Learning (2021 Fall)

## 2.1 Data Cleaning

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p



# Exploratory data analysis

**Check Notebook** 



#### **Data Errors**



• Data often have errors - the mismatch with ground truth (missing, erroneous or extreme values)

Good ML models are robust to errors



- DNN trained with SGD VS Decision trees
- Consequences:
  - The training may still converge, but slower



- Accuracy degradation, could be hard to detect
- Deploying these models may impact the quality of the new collected data
  - e.g. positive examples generated by poor recommendation / search results





## Types of Data Errors

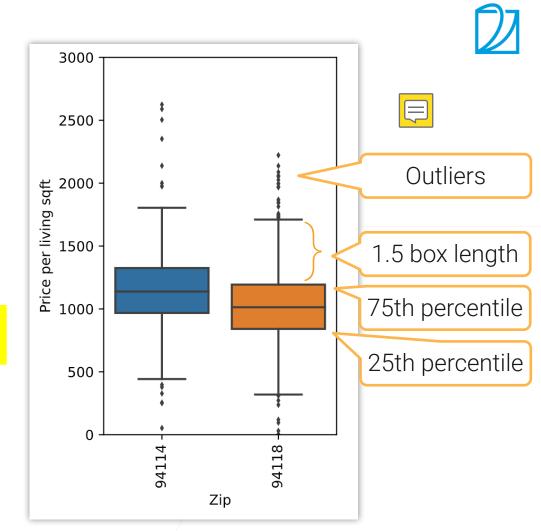




- Outliers: data values that significantly deviate from other observations
  - outliers VS under sampled rare events
- Rule violations: data values violate integrity constraints such as "Not Null" and "Must be unique" and "Non negative"
- Pattern violations: data values violate syntactic and semantic constraints such as formatting, misspelling

### **Outlier Detection**

| data['Type'].value_cou              | nts()[0:20] |        |
|-------------------------------------|-------------|--------|
| executed in 19ms, finished 11:28:29 | 2021-09-15  |        |
| SingleFamily                        | 74318       |        |
| Condo                               | 18749       |        |
| MultiFamily                         | 6586        |        |
| VacantLand                          | 6199        |        |
| Townhouse                           | 5846        |        |
| Unknown                             | 5390        |        |
| MobileManufactured                  | 2588        |        |
| Apartment                           | 1416        |        |
| Cooperative                         | 161         |        |
| Residential Lot                     | 75          |        |
| Single Family                       | 69          |        |
| Single Family Lot                   | 56          |        |
| Acreage                             | 48          |        |
| 2 Story                             | 39          | Outlic |
| 3 Story                             | 25          | Outlie |
| Hi-Rise (9+), Luxury RESIDENTIAL    | 21<br>19    |        |
| Condominium                         | 19          |        |
| Duplex                              | 19          |        |
| Mid-Rise (4-8)                      | 17          |        |



#### **Rule-based Detection**



- Design rules to identify erroneous records
- Functional dependencies:  $x \to \mathbb{F}$  neans a value x determines a unique value y
  - E.g. zip code  $\rightarrow$  state, EIN  $\rightarrow$  company name
- Denial constraints: specified with more flexible first-order logic
  - Phone number is not empty if vendor has an EIN
  - If two captures of the same animal indicated by the same tag number, then the first one must be marked as original

#### Pattern-based Detection



Syntactic patterns



- e.g. Map a column to the most prominent data type and identify values do not fit
- eng, en, english -> English
- Semantic patterns



- e.g. Add rules through knowledge graph
  - Values in column "Country" need have capitals, so a value "Stanford" is invalid

## Summary



- Types of data errors: outliers, rule violations, pattern violations
- Multiple tools exist to help data cleaning
  - Graphic interface for interactive cleaning
  - Automatically detect and fix

