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

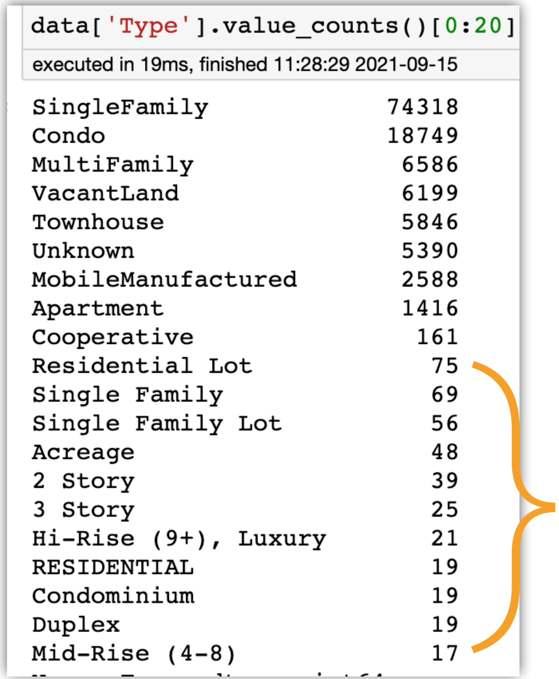
Data Errors

- 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

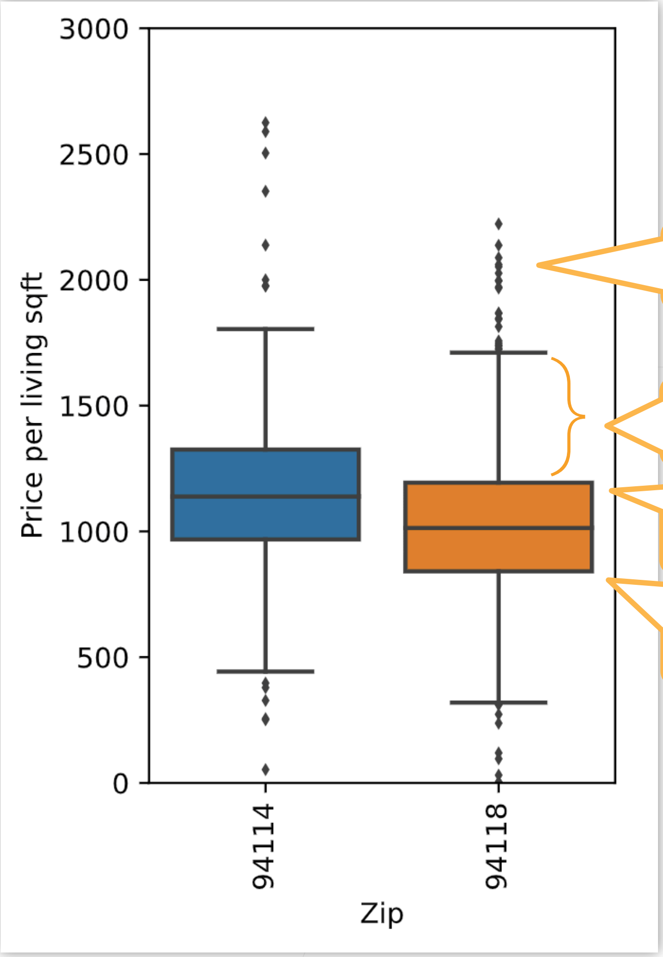


```
data['Type'].value_counts()[0:20]
```

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
3 Story                        25
Hi-Rise (9+), Luxury           21
RESIDENTIAL                    19
Condominium                    19
Duplex                         19
Mid-Rise (4-8)                 17
```

Outliers



Outliers

1.5 box length

75th percentile

25th percentile

Rule-based Detection

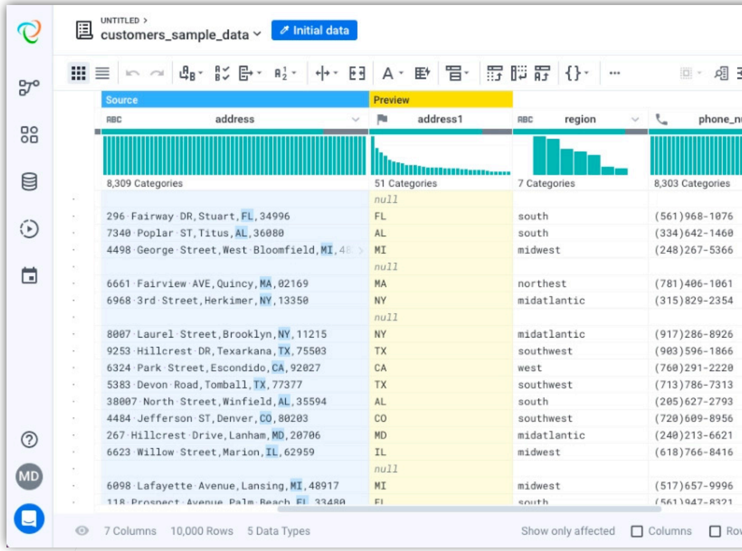
- Design rules to identify erroneous records
- Functional dependencies: $x \rightarrow y$ means a value x determines a unique value y
  - E.g. zip code → state, EIN → company name
- Denial constraints: specified with more flexible first-order logic
  - Phone number is not empty if vendor has an EIN
  - 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



UNTITLED >

customers_sample_data ⌄ Initial data

Source	Preview		
ABC address	address1	ABC region	☎ phone_n
8,309 Categories	51 Categories	7 Categories	8,303 Categories
	null		
296 Fairway DR,Stuart,FL,34996	FL	south	(561)968-1076
7340 Poplar ST,Titus,AL,36080	AL	south	(334)642-1460
4498 George Street,West Bloomfield,MI,48	MI	midwest	(248)267-5366
	null		
6661 Fairview AVE,Quincy,MA,02169	MA	northest	(781)406-1061
6968 3rd Street,Herkimer,NY,13350	NY	midatlantic	(315)829-2354
	null		
8007 Laurel Street,Brooklyn,NY,11215	NY	midatlantic	(917)286-8926
9253 Hillcrest DR,Texarkana,TX,75503	TX	southwest	(903)596-1866
6324 Park Street,Escondido,CA,92027	CA	west	(760)291-2220
5383 Devon Road,Tomball,TX,77377	TX	southwest	(713)786-7313
38007 North Street,Winfield,AL,35594	AL	south	(205)627-2793
4484 Jefferson ST,Denver,CO,80203	CO	southwest	(720)609-8956
267 Hillcrest Drive,Lanham,MD,20706	MD	midatlantic	(240)213-6621
6623 Willow Street,Marion,IL,62959	IL	midwest	(618)766-8416
	null		
6098 Lafayette Avenue,Lansing,MI,48917	MI	midwest	(517)657-9996
118 Prospect Avenue,Palm Beach,FL,33480	FL	south	(561)947-8321

7 Columns 10,000 Rows 5 Data Types

Show only affected □ Columns □ Rows

Trifacta Wrangler