Train In Data

Rare labels & high cardinality

Rare labels & high cardinality

- **High cardinality** refers to a high number of unique categories.
- **Rare labels** are those that are present only in a tiny fraction of the observations.

Consequences

Obs	Gender	Vehicle Make
1	Male	Mercedes
2	Male	Ford
3	Male	Ford
4	Male	Renault
5	Male	Seat
6	Male	Renault
7	Female	Citroen
8	Female	Toyota
9	Female	Kia
10	Female	Kia
11	Female	Nissan
12	Female	BMW

Train Set

Obs	Gender	Vehicle Make
1	Male	Mercedes
3	Male	Ford
6	Male	Renault
7	Female	Citroen
9	Female	Kia
11	Female	Nissan

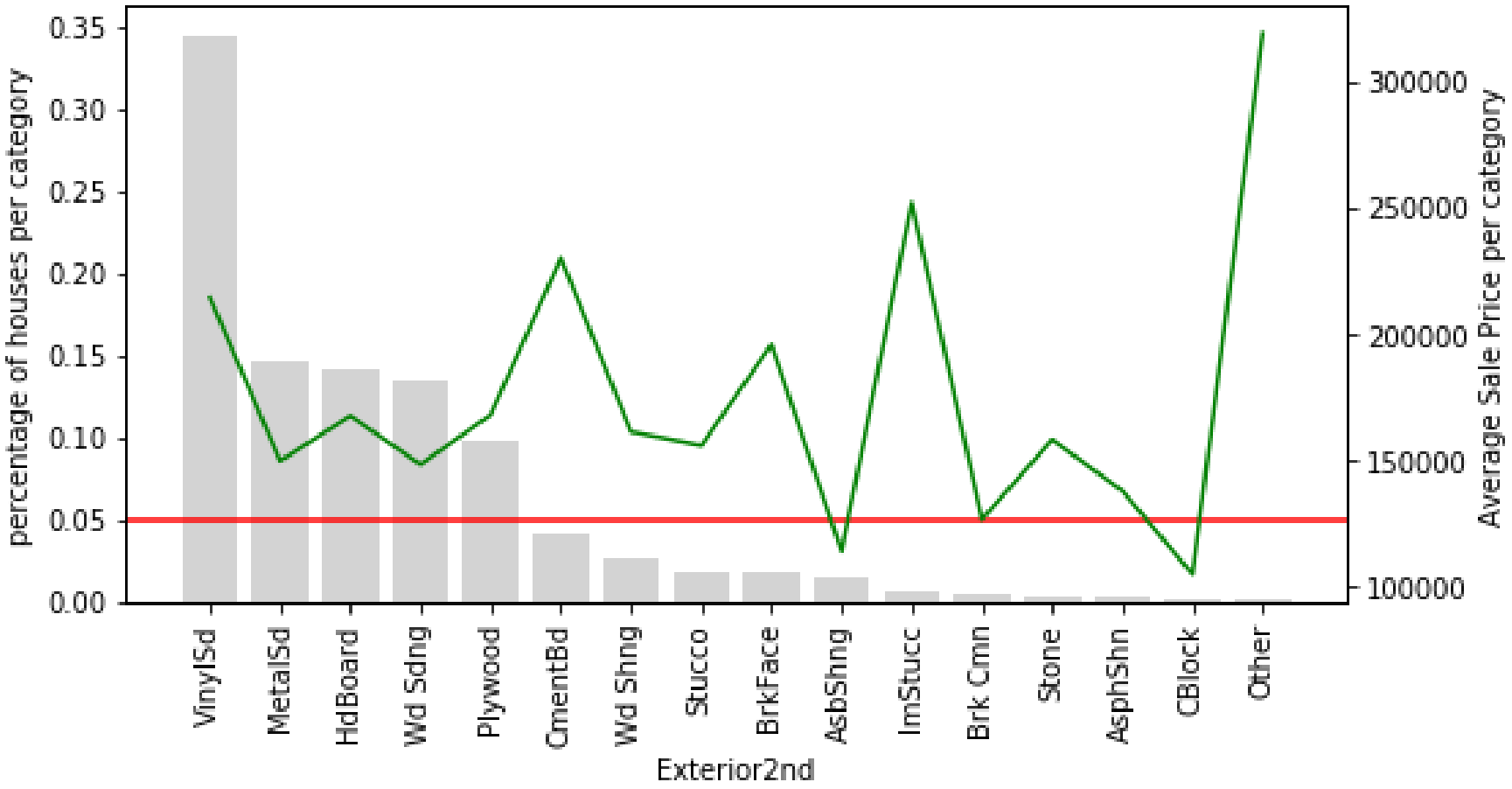
Potential Overfit

Test Set

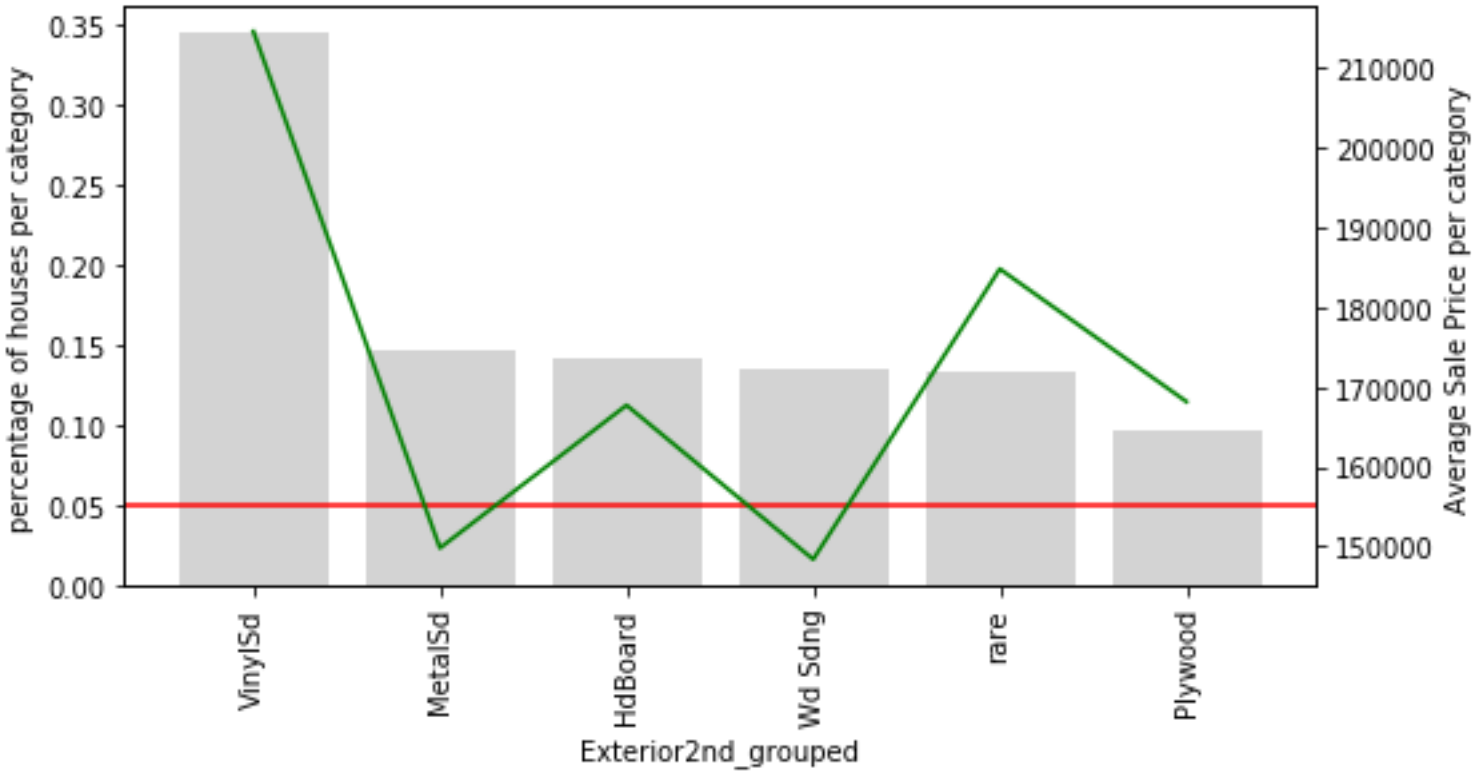
Obs	Gender	Vehicle Make
2	Male	Ford
5	Male	Seat
4	Male	Renault
8	Female	Toyota
10	Female	Kia
12	Female	BMW

Potential Operation alisation problem

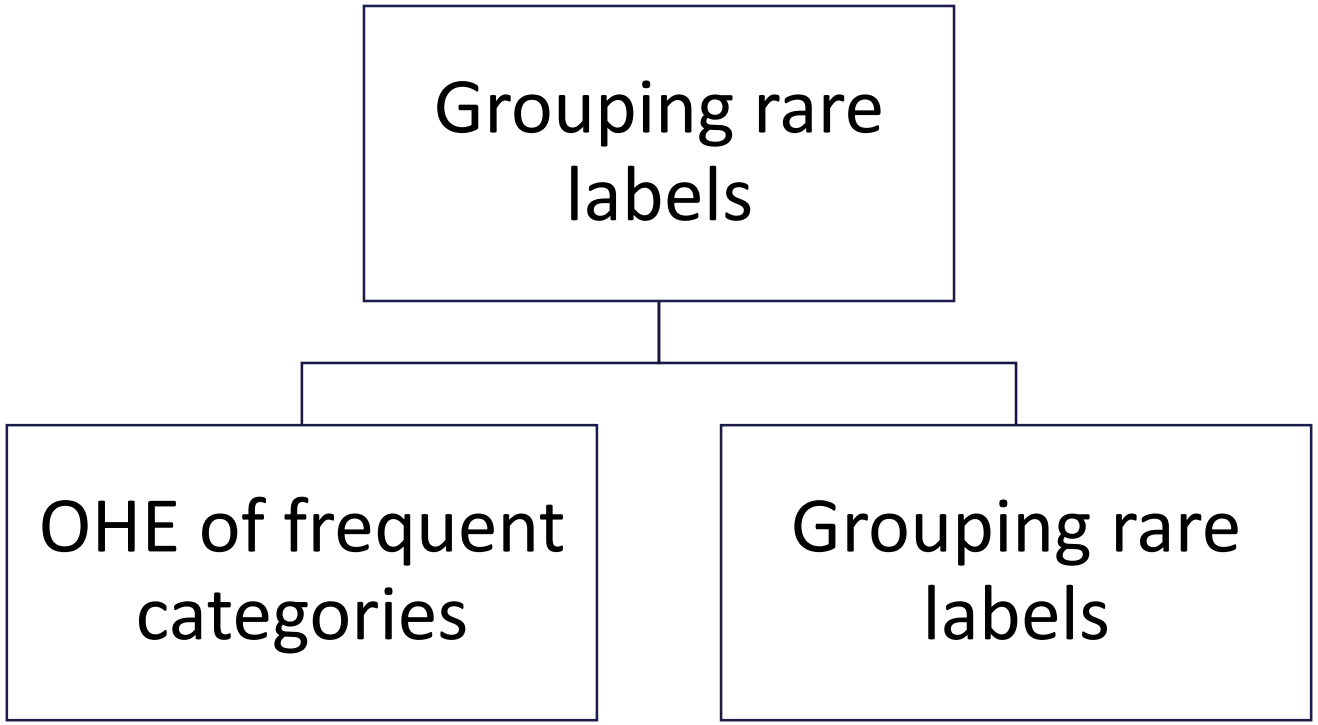
Rare labels – unreliable information



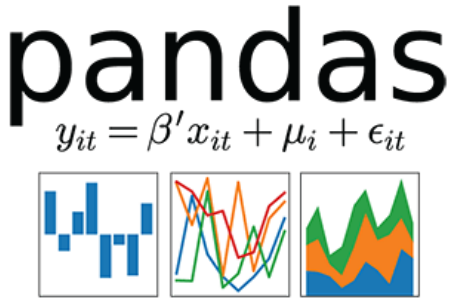
Grouping rare labels



Grouping rare labels



Grouping labels in Python



Feature-Engine

Content

For each lecture:

- Presentation and video
- Accompanying Jupyter notebook
  - Implementation in **pandas**
  - Implementation in **Feature-engine**
  - Implementation in **sklearn**

Train In Data

THANK YOU

www.trainindata.com