

DATA 3464: Fundamentals of Data Processing

Categorical Data

Charlotte Curtis

January 22, 2026

Topic overview

- Exploring categorical data
- Dealing with missing values
- Categorical data encoding strategies

Resources used:

- [Feature Engineering Chapter 5](#)

Exploring categorical data

- `pandas.DataFrame.value_counts()` - how many of each category?
- Use category to group, then compute summary stats
- Plot color per category
- Scatter plot with jitter

What is categorical data?

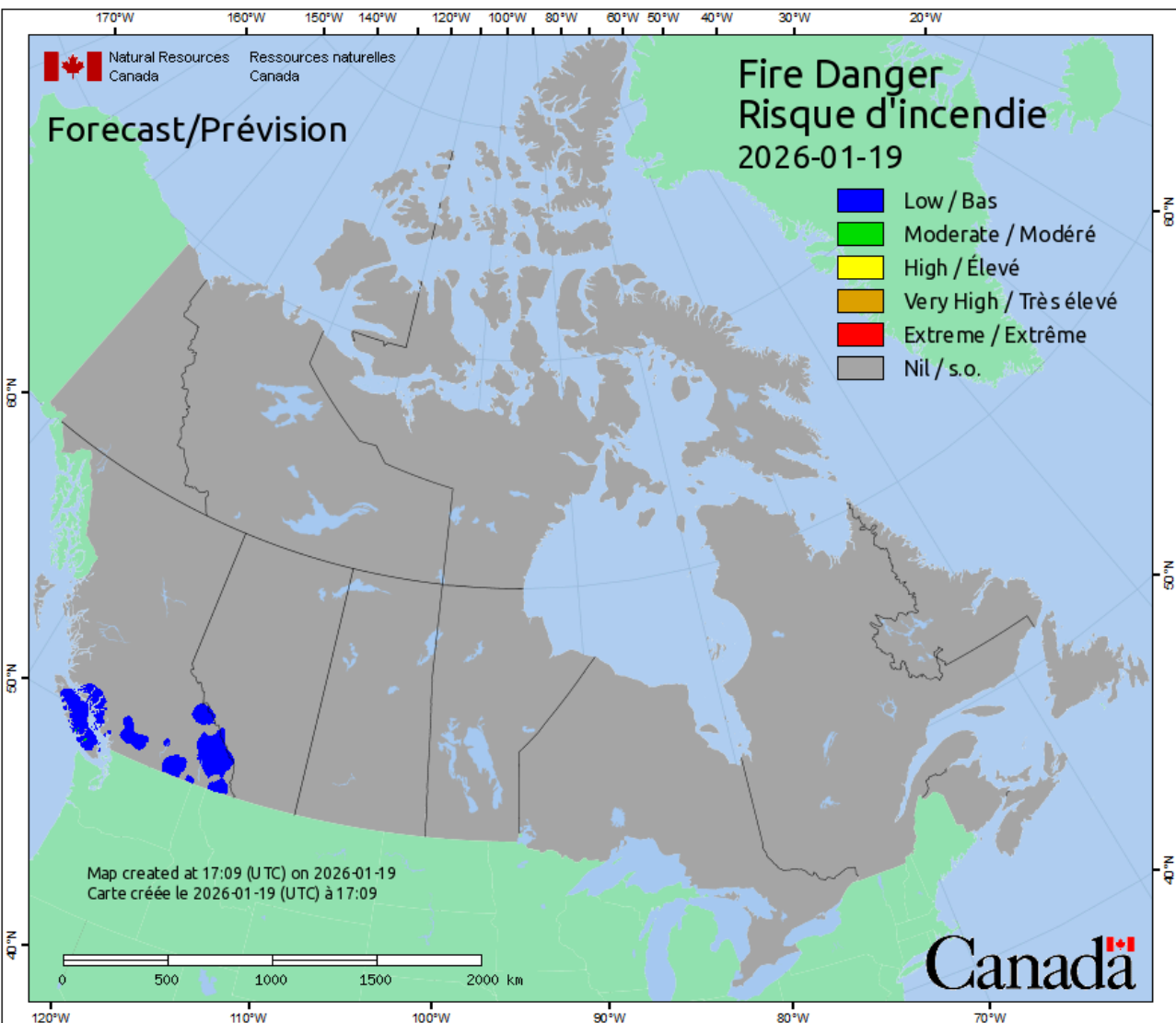
- Samples can take on one of several discrete values or groups
 - **Nominal**: no particular order to the groups
 - **Ordinal**: groups relate to each other in a specific order
- Categories can be represented as strings *or* numeric types
 - Domain knowledge is necessary!

Let's take a few minutes to brainstorm some examples

Representing categorical data

- Tree-based models can handle string-based categories as-is
- Most other models need numbers
- Consider:
 - Ordinal or nominal?
 - How many possible categories?
 - Any chance new ones might show up?

How could we encode the examples?



Ordinal encoding

Category	Feature
Nil	0
Low	1
Moderate	2
High	3
Very High	4
Extreme	5

Nominal categories: one-hot encoding

- Categories have no natural relationship
- Create k new features from k categories, very sparse matrix

Animal		cat	dog	rabbit
cat	→	1	0	0
dog	→	0	1	0
rabbit	→	0	0	1

What kinds of problems could occur with this encoding scheme?

Another approach: target encoding

- Basic concept: replace the category with the mean of the target
- Essential to avoid data leakage!
- Example: predicting weight of animal

Animal		mean_kg
cat	→	4.1
dog	→	15.4
rabbit	→	2.2

Getting fancy

- Feature hashing or the "hash trick"
- Good if you have too many categories, or combinations of categories
- Converts each category into a fixed-length feature vector

Animal		A_0	A_1	A_2	...	A_16
cat	→	1	0	0	...	1
dog	→	0	1	0	...	1
rabbit	→	1	0	1	...	0

Missing values in categorical data
