How to handle Categorical features Many Categories (Count/frequency Encoding)

Type of Categorical variables
- ordinal (ex) PhD, Master, Bachelor's (or ordered variables) — can be diff. on the basis of Rank
- nominal (ex) (is a naming scale where variables are simply "named" or labeled with no specific order.)

Note: For ordinal variables we can use label encoding

For nominal we can use one-hot encoding

6 With Many Features (nominal): is to replace each label of the Categorical variable by the count, this is the amount of times each label appears in the dataset.

or the frequency, this is the percentage of observations within that category.

⇒ df.head()

	x1	x2
0	v	at
1	t	av
2	w	n
3	t	n.

⇒ df.shape → (4209, 2) (total rows, columns)

⇒ pd.get_dummies(df).shape ⇒ (4209, 71) (columns) # one hot encoding

⇒ len(df['x1'].unique()) ⇒ 27
len(df['x2'].unique()) ⇒ 44
⇒ 27 + 44 = 71

Let's find out the counts for each one of the labels in variable x2

f = df.x2.value_counts().to_dict() ⇒ {'at': 2, 'n': 20 ...}

f_frequency(f)

df.x2 = df.x2.map(frequency) ⇒

(using map function to map counts with labels in x2)

	x1	x2
0	v	2
1	t	2
2	w	20
3	t.	20