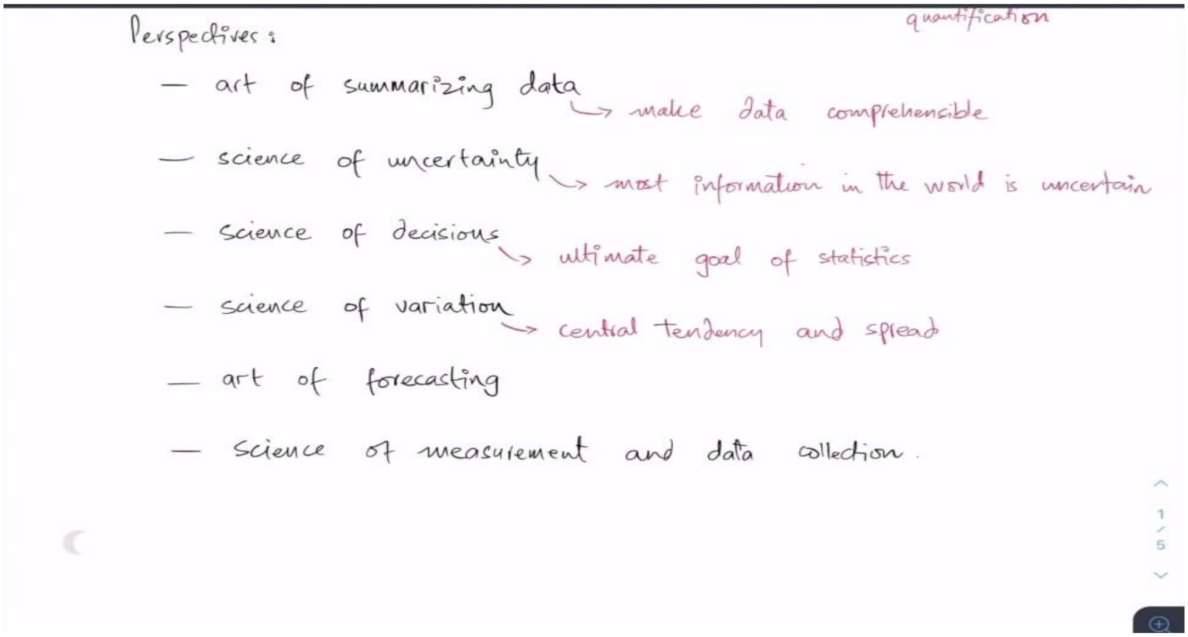
Lecture 3

Statistics:

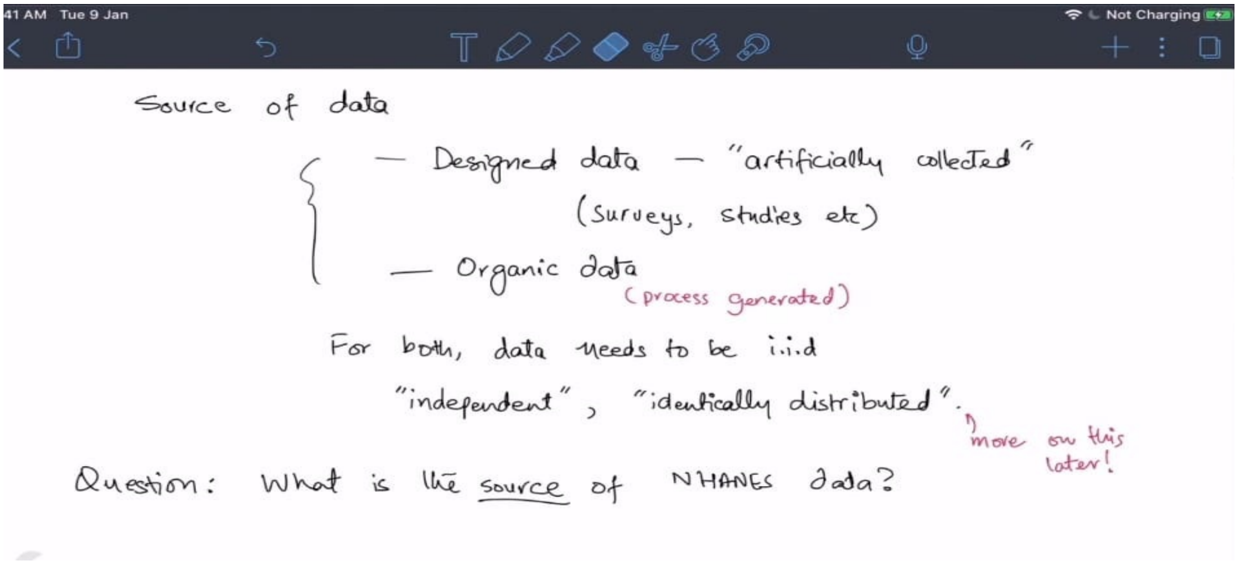
Academic discipline dealing with all aspects of data(quantification):



quantification

Perspectives :

- art of summarizing data ↳ make data comprehensible
- science of uncertainty ↳ most information in the world is uncertain
- science of decisions ↳ ultimate goal of statistics
- science of variation ↳ central tendency and spread
- art of forecasting
- science of measurement and data collection.



Source of data

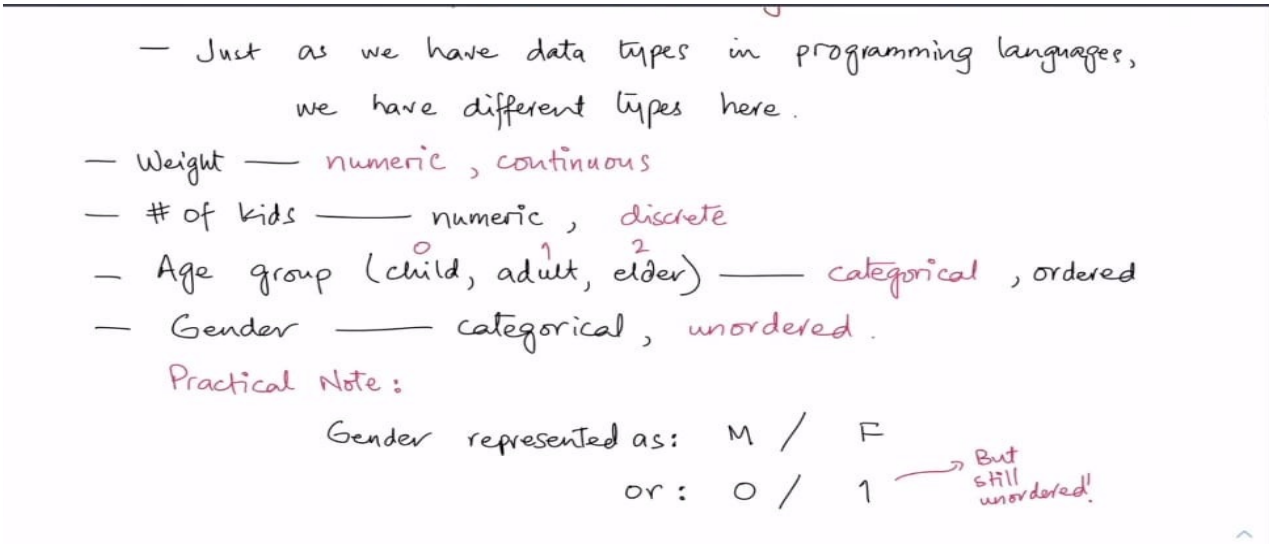
- Designed data — "artificially collected" (surveys, studies etc)
- Organic data (process generated)

For both, data needs to be i.i.d "independent", "identically distributed".

more on this later!

Question: What is the <u>source</u> of NHANES dada?

Types of data:



— Just as we have data types in programming languages, we have different types here.

— Weight —— numeric , continuous

— # of kids —— numeric , discrete

— Age group (child, adult, elder) —— categorical , ordered

$\quad\quad\quad\quad\quad\quad\quad\quad 0 \quad\quad 1 \quad\quad 2$

— Gender —— categorical , unordered.

Practical Note :

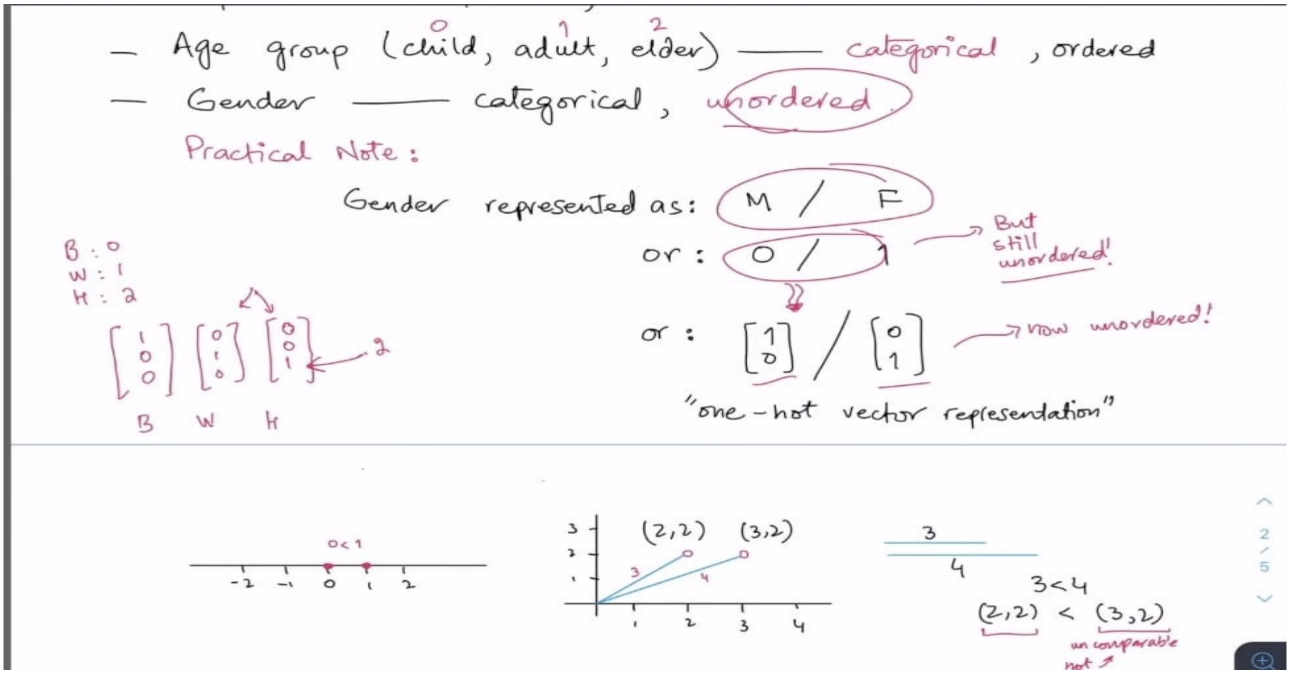
Gender represented as: M / F

or : 0 / 1 → But still unordered!

But here comes the problem

0 and 1 are ordered → but male and female aren't

so solution is One- hot encoding assigning vectors instead of numbers as vectors are not comparable



— Age group (child, adult, elder) —— categorical , ordered

$\quad\quad\quad\quad\quad\quad\quad\quad 0 \quad\quad 1 \quad\quad 2$

— Gender —— categorical , unordered

Practical Note :

Gender represented as: M / F

or : 0 / 1 → But still unordered!

or : $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ / $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ → now unordered!

"one-hot vector representation"

B : 0
W : 1
H : 2

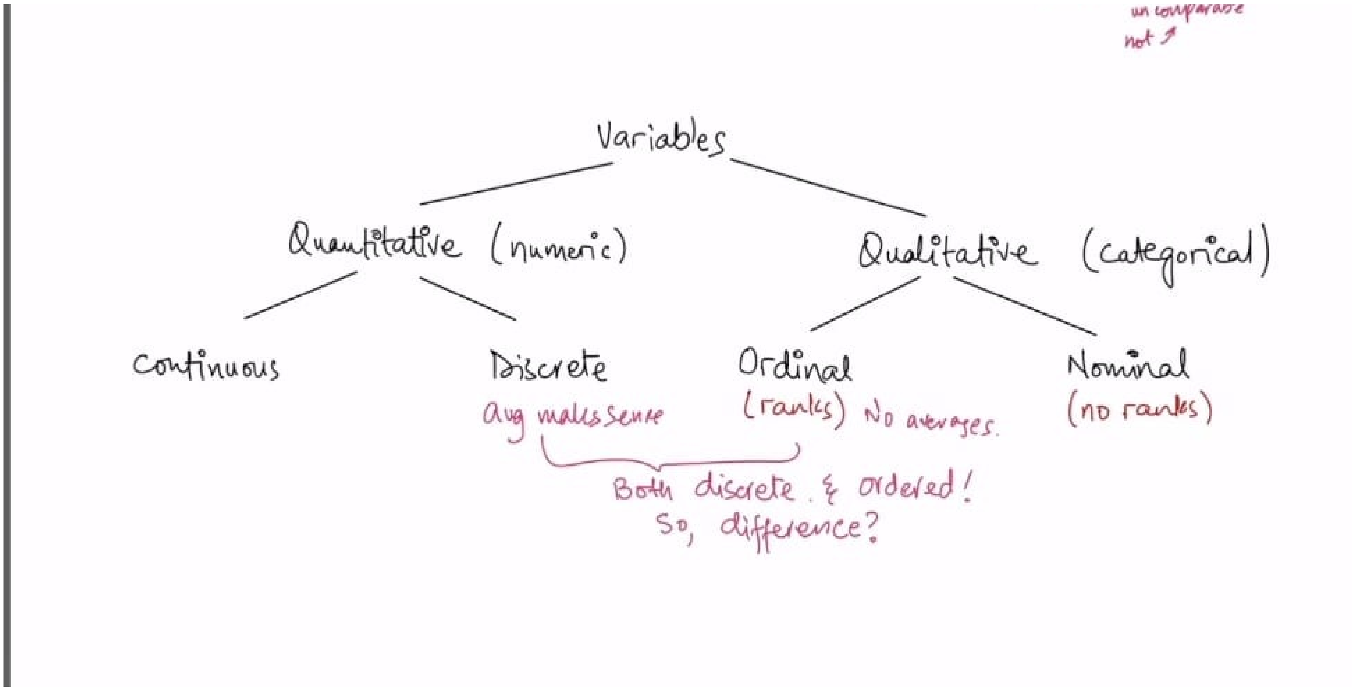
$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \leftarrow 2$

B W H

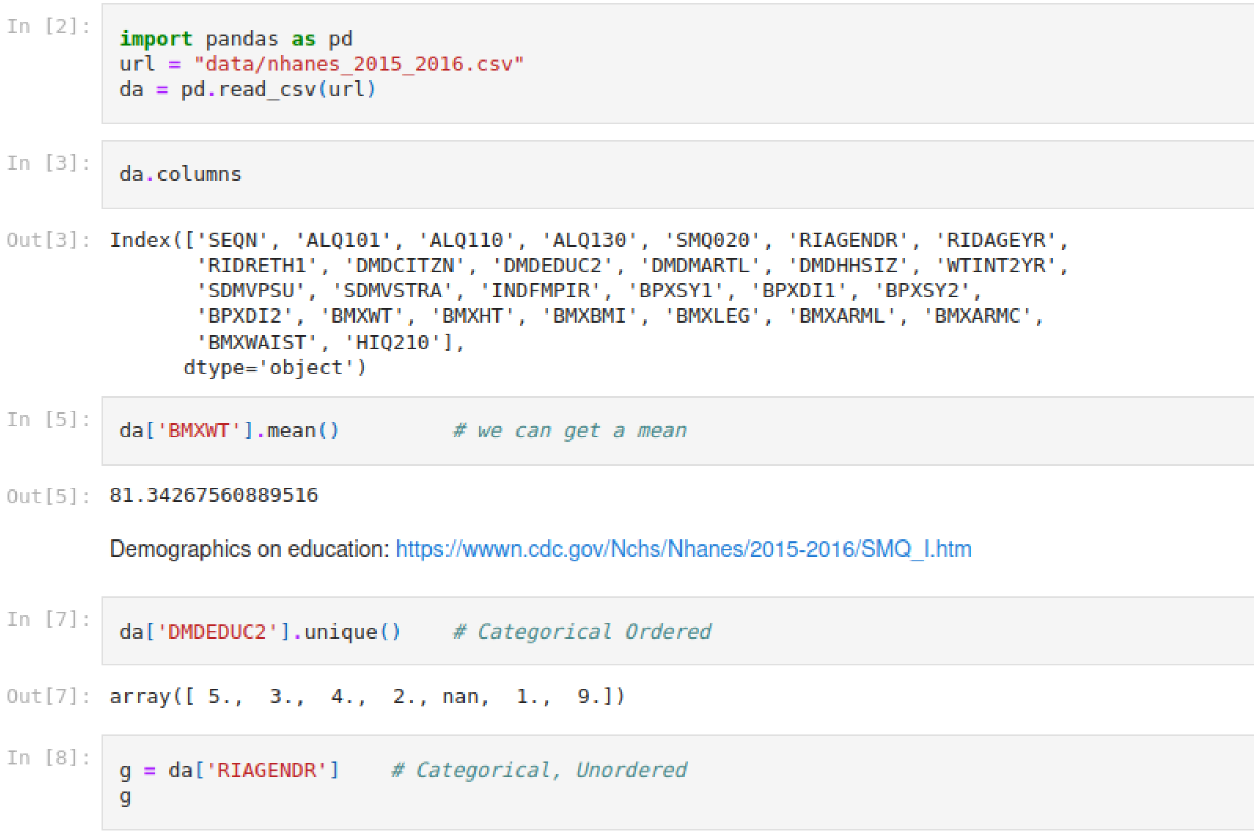
index which is assigned to values give it value 1 in vector

- vector magnitude is comparable but overall vector is not as it contains 2 values

Summary:



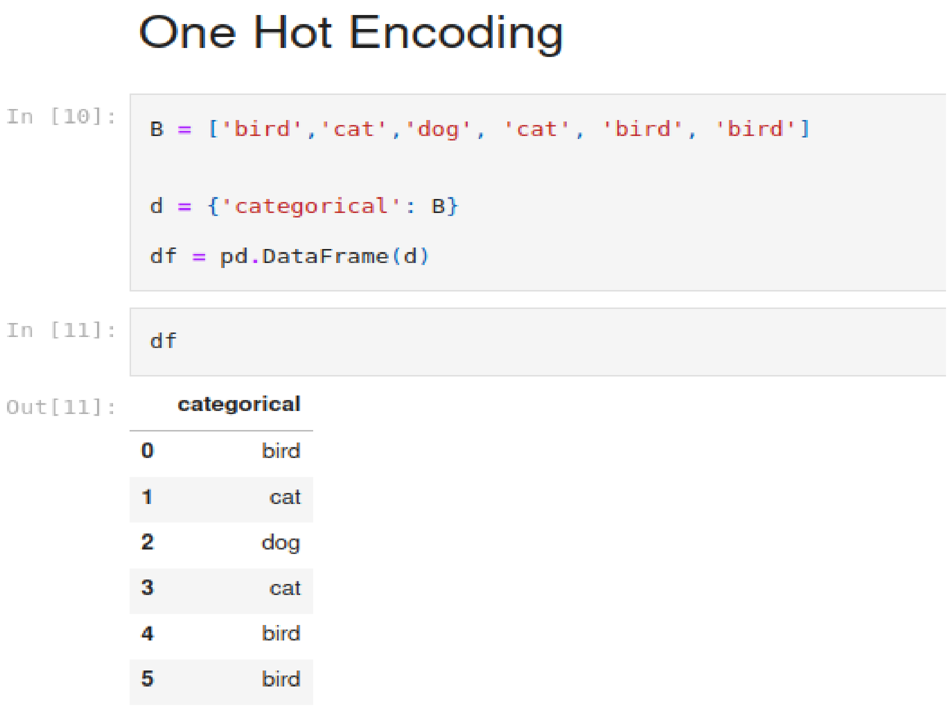
code:



```
In [2]: import pandas as pd
        url = "data/nhanes_2015_2016.csv"
        da = pd.read_csv(url)

In [3]: da.columns

Out[3]: Index(['SEQN', 'ALQ101', 'ALQ110', 'ALQ130', 'SMQ020', 'RIAGENDR', 'RIDAGEYR',
               'RIDRETH1', 'DMDCITZN', 'DMDEDUC2', 'DMDMARTL', 'DMDHHSIZ', 'WTINT2YR',
               'SDMVPSU', 'SDMVSTRA', 'INDFMPIR', 'BPXSY1', 'BPXDI1', 'BPXSY2',
               'BPXDI2', 'BMXWT', 'BMXHT', 'BMXBMI', 'BMXLEG', 'BMXARML', 'BMXARMC',
               'BMXWAIST', 'HIQ210'],
              dtype='object')

In [5]: da['BMXWT'].mean()         # we can get a mean

Out[5]: 81.34267560889516
```

Demographics on education: https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/SMQ_I.htm

```
In [7]: da['DMDEDUC2'].unique()     # Categorical Ordered

Out[7]: array([ 5.,  3.,  4.,  2., nan,  1.,  9.])

In [8]: g = da['RIAGENDR']     # Categorical, Unordered
        g
```



One Hot Encoding

In [10]:

```
B = ['bird','cat','dog', 'cat', 'bird', 'bird']

d = {'categorical': B}
df = pd.DataFrame(d)
```

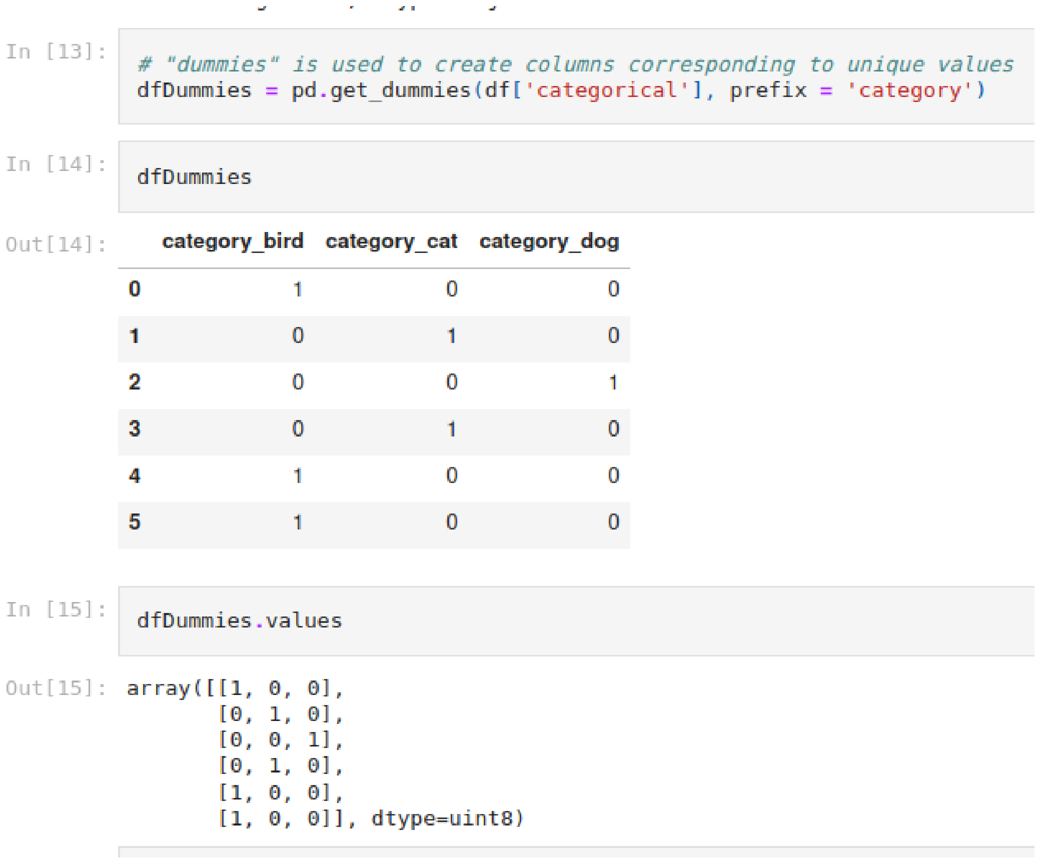
In [11]:

```
df
```

Out[11]:

	categorical
0	bird
1	cat
2	dog
3	cat
4	bird
5	bird

pandas → get dummies



In [13]:

```
# "dummies" is used to create columns corresponding to unique values
dfDummies = pd.get_dummies(df['categorical'], prefix = 'category')
```

In [14]:

```
dfDummies
```

Out[14]:

	category_bird	category_cat	category_dog
0	1	0	0
1	0	1	0
2	0	0	1
3	0	1	0
4	1	0	0
5	1	0	0

In [15]:

```
dfDummies.values
```

Out[15]:

```
array([[1, 0, 0],
       [0, 1, 0],
       [0, 0, 1],
       [0, 1, 0],
       [1, 0, 0],
       [1, 0, 0]], dtype=uint8)
```