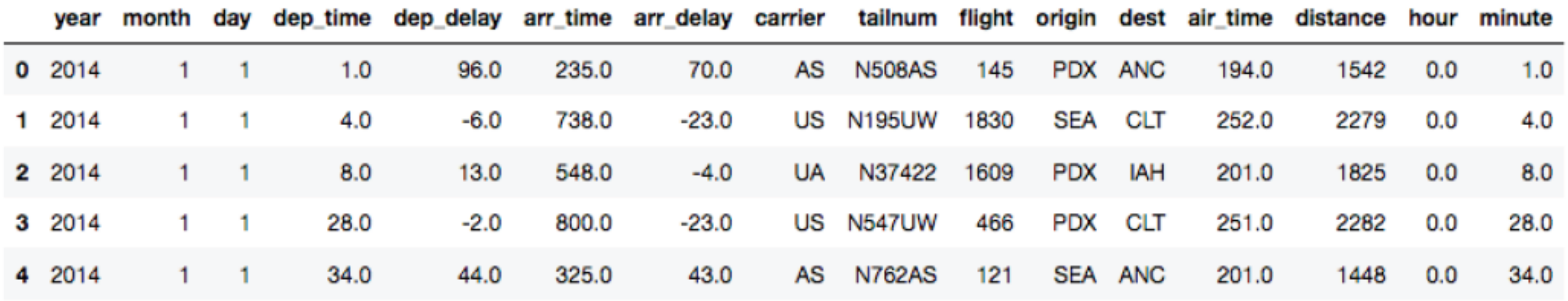
One-hot-encoding

Handling Categorical Data

- Categorical variables are known to hide and mask lots of interesting information in a data set. It’s crucial to learn the methods of dealing with such variables. If you won’t, many a times, you’d miss out on finding the most important variables in a model.

Handling Categorical Data

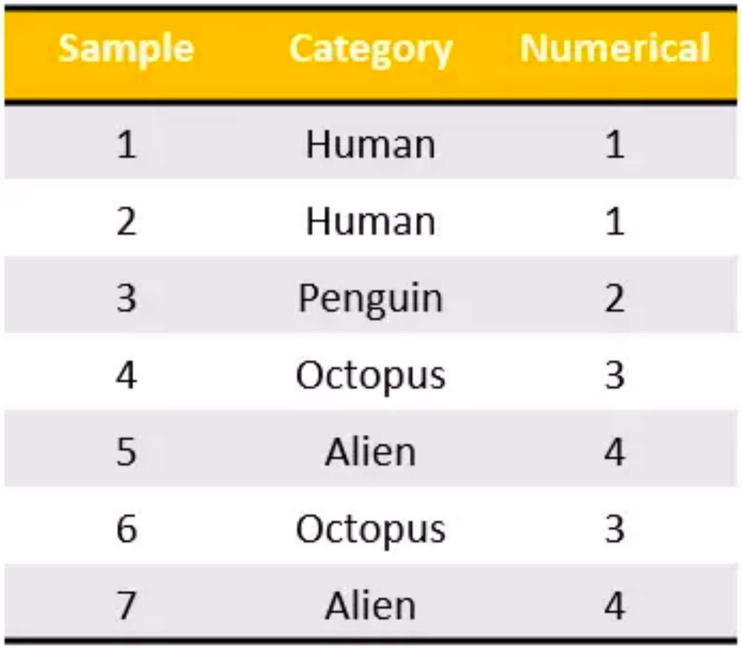


	year	month	day	dep_time	dep_delay	arr_time	arr_delay	carrier	tailnum	flight	origin	dest	air_time	distance	hour	minute
0	2014	1	1	1.0	96.0	235.0	70.0	AS	N508AS	145	PDX	ANC	194.0	1542	0.0	1.0
1	2014	1	1	4.0	-6.0	738.0	-23.0	US	N195UW	1830	SEA	CLT	252.0	2279	0.0	4.0
2	2014	1	1	8.0	13.0	548.0	-4.0	UA	N37422	1609	PDX	IAH	201.0	1825	0.0	8.0
3	2014	1	1	28.0	-2.0	800.0	-23.0	US	N547UW	466	PDX	CLT	251.0	2282	0.0	28.0
4	2014	1	1	34.0	44.0	325.0	43.0	AS	N762AS	121	SEA	ANC	201.0	1448	0.0	34.0

One approach : one-hot encoding

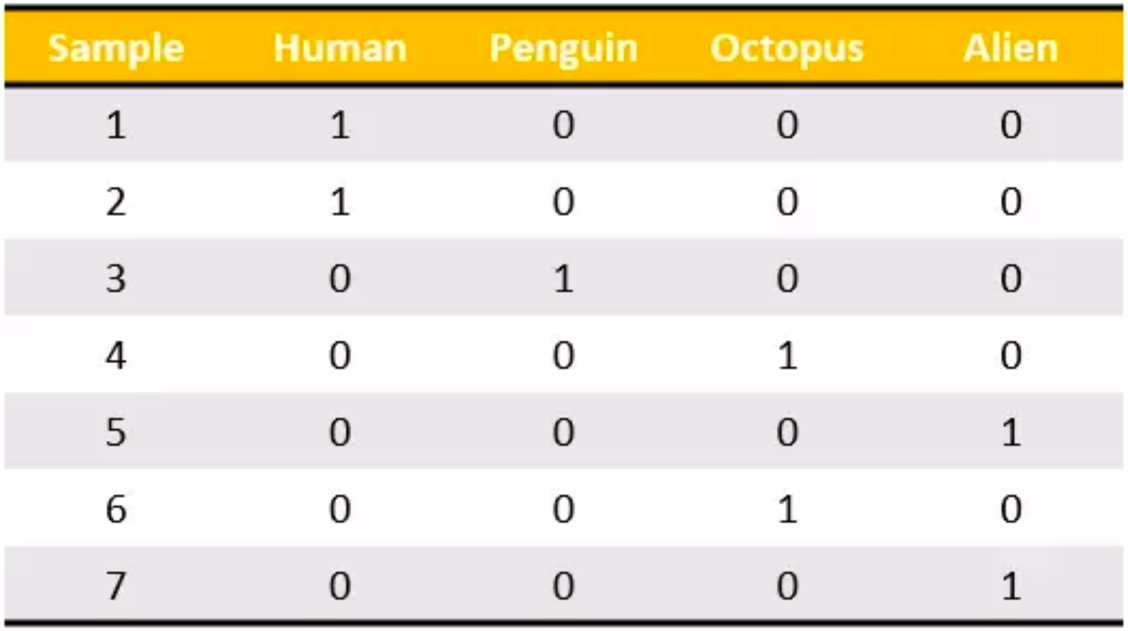
- One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

Example



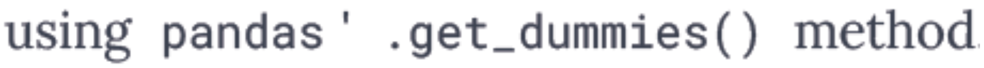
Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4

Example



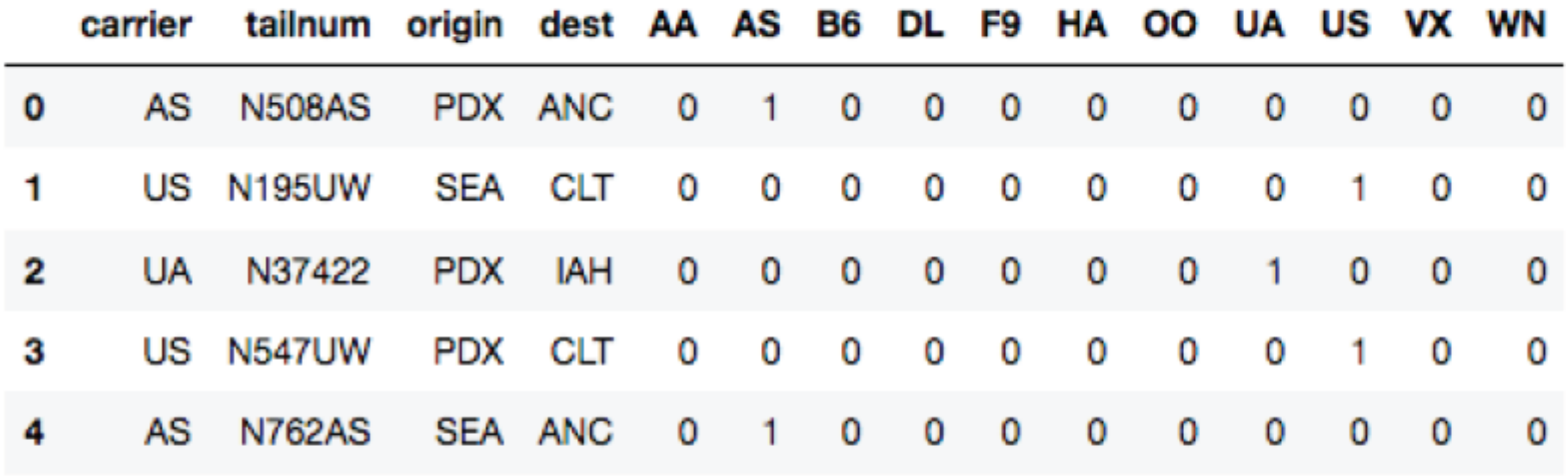
Sample	Human	Penguin	Octopus	Alien
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

How to do it in Python?

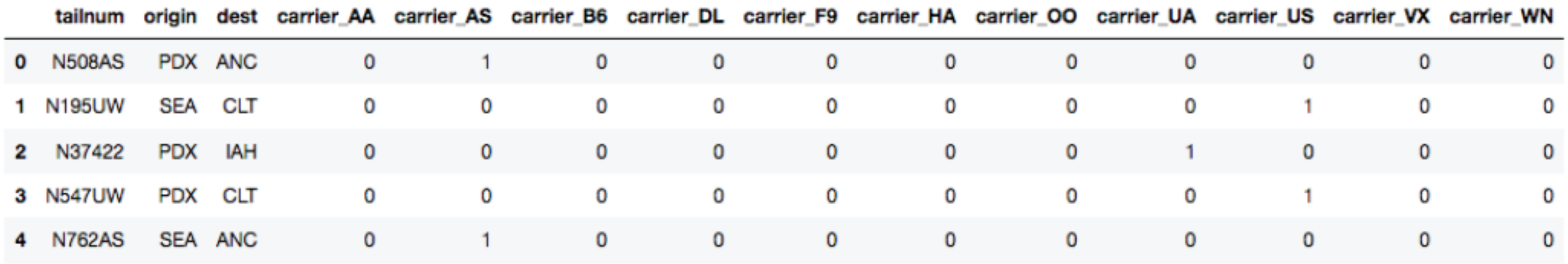


using pandas ' .get_dummies() method

One-hot -encoding



	carrier	tailnum	origin	dest	AA	AS	B6	DL	F9	HA	OO	UA	US	VX	WN
0	AS	N508AS	PDX	ANC	0	1	0	0	0	0	0	0	0	0	0
1	US	N195UW	SEA	CLT	0	0	0	0	0	0	0	0	1	0	0
2	UA	N37422	PDX	IAH	0	0	0	0	0	0	0	1	0	0	0
3	US	N547UW	PDX	CLT	0	0	0	0	0	0	0	0	1	0	0
4	AS	N762AS	SEA	ANC	0	1	0	0	0	0	0	0	0	0	0



	tailnum	origin	dest	carrier_AA	carrier_AS	carrier_B6	carrier_DL	carrier_F9	carrier_HA	carrier_OO	carrier_UA	carrier_US	carrier_VX	carrier_WN
0	N508AS	PDX	ANC	0	1	0	0	0	0	0	0	0	0	0
1	N195UW	SEA	CLT	0	0	0	0	0	0	0	0	1	0	0
2	N37422	PDX	IAH	0	0	0	0	0	0	0	1	0	0	0
3	N547UW	PDX	CLT	0	0	0	0	0	0	0	0	1	0	0
4	N762AS	SEA	ANC	0	1	0	0	0	0	0	0	0	0	0