Data Encoding with Python

Pratheerth Padman
FREELANCE DATA SCIENTIST

Module Overview

Overview of One-Hot Encoding

Converting categorical values using O.H.E

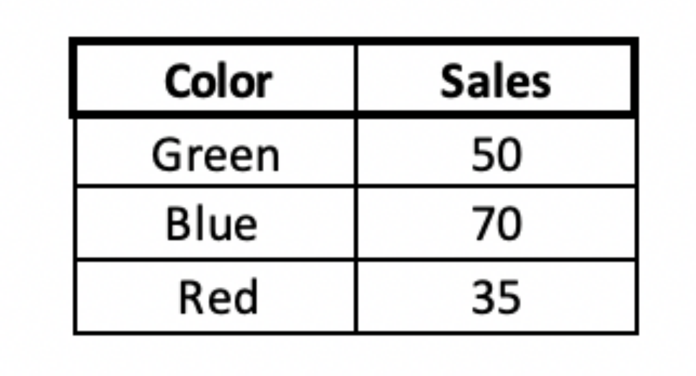
Create dummy variables with Pandas

Frequency table with the crosstab function

One-Hot Encoding

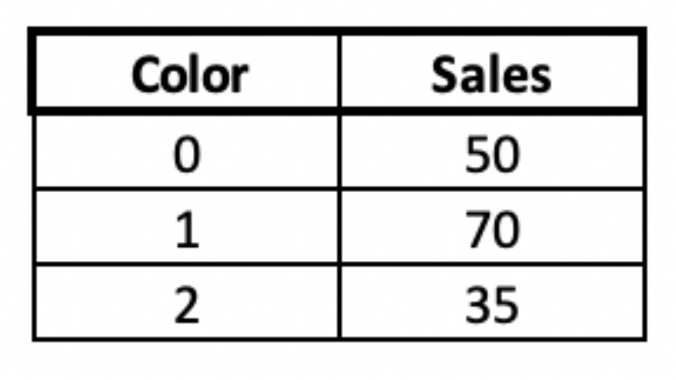
Label Encoder

Encodes labels with a value between 0 and (n-1), where n is the number of distinct values in a feature



Color	Sales
Green	50
Blue	70
Red	35

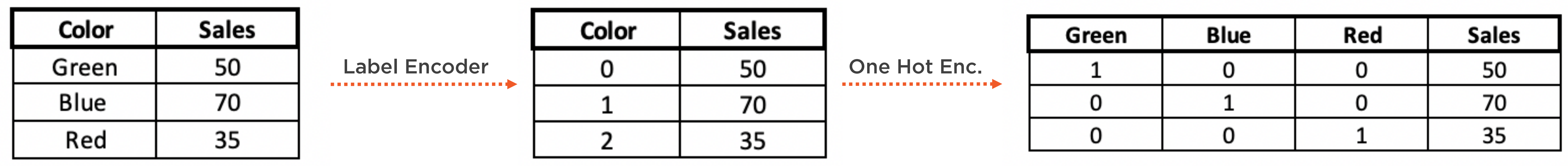
Label Encoder



Color	Sales
0	50
1	70
2	35

One-Hot Encoding

One hot encoding creates a binary variable for each unique categorical value



Color	Sales
Green	50
Blue	70
Red	35

Label Encoder

Color	Sales
0	50
1	70
2	35

One Hot Enc.

Green	Blue	Red	Sales
1	0	0	50
0	1	0	70
0	0	1	35

Demo

Demo: Convert categorical values using One-Hot Encoding

Create Dummy Variables with Pandas

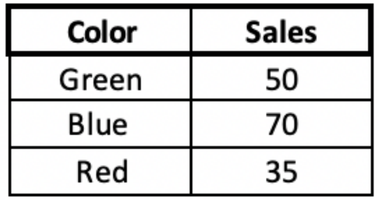
get_dummies() Function

Converts categorical variable into dummy/indicator variable

Important parameters:

- data – data of which to get dummy indicators
- columns - column names in the datarame to be encoded

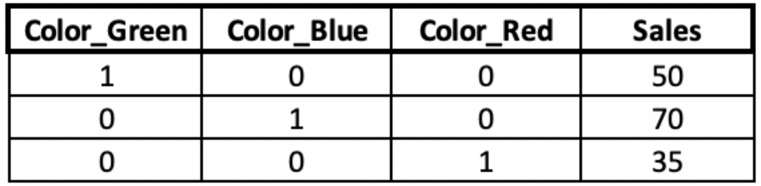
Problems with Using get_dummies()



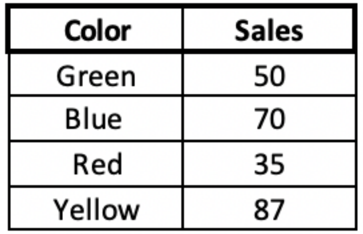
Color	Sales
Green	50
Blue	70
Red	35

Train Set

get_dummies()



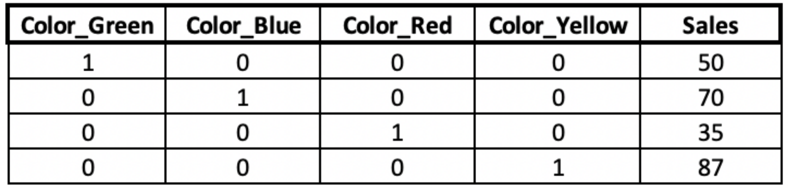
Color_Green	Color_Blue	Color_Red	Sales
1	0	0	50
0	1	0	70
0	0	1	35



Color	Sales
Green	50
Blue	70
Red	35
Yellow	87

Test Set

get_dummies()



Color_Green	Color_Blue	Color_Red	Color_Yellow	Sales
1	0	0	0	50
0	1	0	0	70
0	0	1	0	35
0	0	0	1	87

How Will OneHotEncoder Solve This?

It has a parameter called ‘handle_unknown’

Default value is ‘error’ which will throw an error in case it encounters a new level in a categorical feature

Set it to ‘ignore’ will result in it not creating an additional column for a new level

Create a Frequency Table with the Crosstab() Function

Crosstab Function

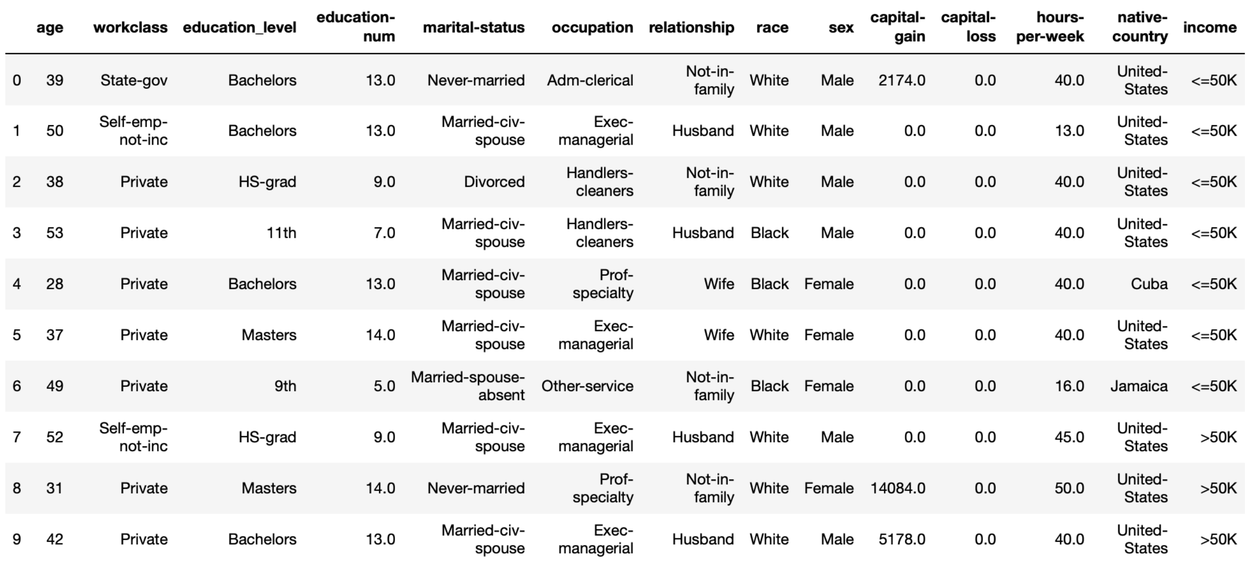
Builds a cross-tabulation table that shows the frequency with which certain categories appear in the data

Two required parameters

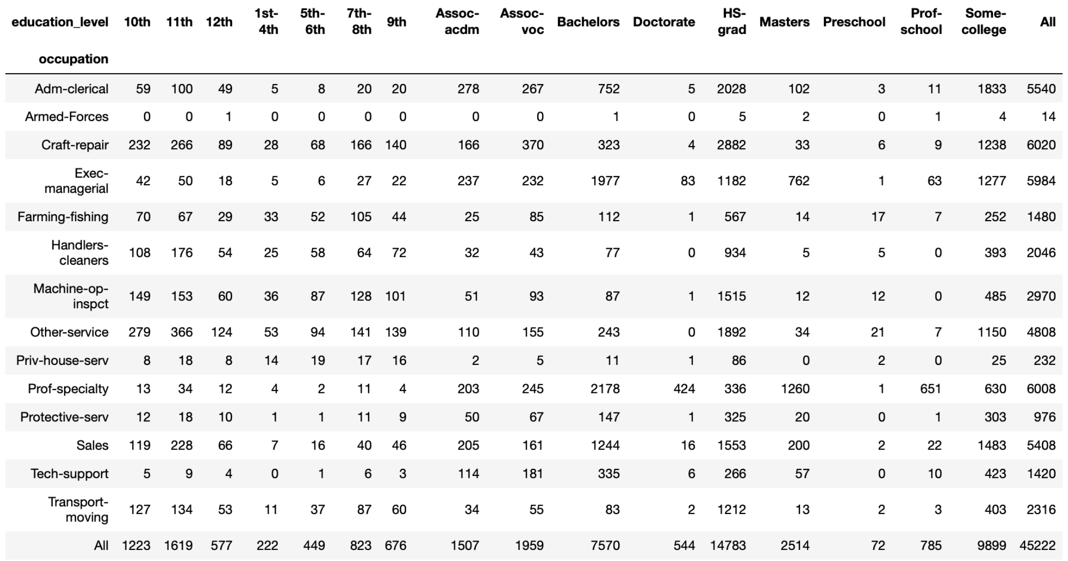
index - Values to group by in the rows

columns - Values to group by in the columns

Crosstab Function



	age	workclass	education_level	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	Bachelors	13.0	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0.0	40.0	United-States	<=50K
1	50	Self-emp-not-inc	Bachelors	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	13.0	United-States	<=50K
2	38	Private	HS-grad	9.0	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0.0	40.0	United-States	<=50K
3	53	Private	11th	7.0	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0.0	0.0	40.0	United-States	<=50K
4	28	Private	Bachelors	13.0	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0.0	0.0	40.0	Cuba	<=50K
5	37	Private	Masters	14.0	Married-civ-spouse	Exec-managerial	Wife	White	Female	0.0	0.0	40.0	United-States	<=50K
6	49	Private	9th	5.0	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0.0	0.0	16.0	Jamaica	<=50K
7	52	Self-emp-not-inc	HS-grad	9.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	45.0	United-States	>50K
8	31	Private	Masters	14.0	Never-married	Prof-specialty	Not-in-family	White	Female	14084.0	0.0	50.0	United-States	>50K
9	42	Private	Bachelors	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178.0	0.0	40.0	United-States	>50K



education_level	10th	11th	12th	1st-4th	5th-6th	7th-8th	9th	Assoc-acdm	Assoc-voc	Bachelors	Doctorate	HS-grad	Masters	Preschool	Prof-school	Some-college	All
occupation																	
Adm-clerical	59	100	49	5	8	20	20	278	267	752	5	2028	102	3	11	1833	5540
Armed-Forces	0	0	1	0	0	0	0	0	0	1	0	5	2	0	1	4	14
Craft-repair	232	266	89	28	68	166	140	166	370	323	4	2882	33	6	9	1238	6020
Exec-managerial	42	50	18	5	6	27	22	237	232	1977	83	1182	762	1	63	1277	5984
Farming-fishing	70	67	29	33	52	105	44	25	85	112	1	567	14	17	7	252	1480
Handlers-cleaners	108	176	54	25	58	64	72	32	43	77	0	934	5	5	0	393	2046
Machine-op-inspct	149	153	60	36	87	128	101	51	93	87	1	1515	12	12	0	485	2970
Other-service	279	366	124	53	94	141	139	110	155	243	0	1892	34	21	7	1150	4808
Priv-house-serv	8	18	8	14	19	17	16	2	5	11	1	86	0	2	0	25	232
Prof-specialty	13	34	12	4	2	11	4	203	245	2178	424	336	1260	1	651	630	6008
Protective-serv	12	18	10	1	1	11	9	50	67	147	1	325	20	0	1	303	976
Sales	119	228	66	7	16	40	46	205	161	1244	16	1553	200	2	22	1483	5408
Tech-support	5	9	4	0	1	6	3	114	181	335	6	266	57	0	10	423	1420
Transport-moving	127	134	53	11	37	87	60	34	55	83	2	1212	13	2	3	403	2316
All	1223	1619	577	222	449	823	676	1507	1959	7570	544	14783	2514	72	785	9899	45222

Standard dataframe

Crosstab table

Summary

Label encoding – encodes labels with values from 0 to (n-1)

One hot encoding – as many columns as there are distinct categorical values

Get_dummies - converts categorical variable into dummy/indicator variable

Difference between get_dummies() and onehotencoder()

Crosstab – for effective summarization

Feedback

Discussion tab for any feedback or questions

Leave a star rating!

Thanks for watching!