DATA TYPES

By: METIS

DATA TYPES

Numerical (Quantitative)		Categorical (Qualitative)	
Continuous	**Discrete**	**Nominal**	**Ordinal**
Infinite Options *Example: Square Footage*	Finite Options *Example: Number of Bedrooms*	Unordered Categories *Example: Exterior Color of House*	Ordered Categories *Example: No / Partial / Full Garage*

REGRESSION PROBLEM

Let's say I'd like to predict house prices. How would I structure my data for this problem?

House Price	Square Footage	Number of Bedrooms	Exterior Color of House	Garage
$400,000	1700	2	Tan	Partial
$600,000	2500	3	Blue	Full
$350,000	1500	2	White	None
$500,000	2000	3	Blue	Partial

REGRESSION PROBLEM

Let's say I'd like to predict house prices. How would I structure my data for this problem?

House Price	Square Footage	Number of Bedrooms	Exterior Color of House	Garage
$400,000	1700	2	Tan	Partial
$600,000	2500	3	Blue	Full
$350,000	1500	2	White	None
$500,000	2000	3	Blue	Partial

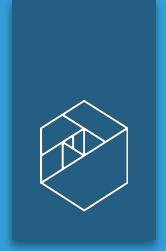
Can I input this data directly into a linear regression model?

REGRESSION PROBLEM

Let's say I'd like to predict house prices. How would I structure my data for this problem?

House Price	Square Footage	Number of Bedrooms	Exterior Color of House	Garage
$400,000	1700	2	Tan	Partial
$600,000	2500	3	Blue	Full
$350,000	1500	2	White	None
$500,000	2000	3	Blue	Partial

The numerical fields are okay, but **we need to make the categorical fields numerical**.



DUMMY VARIABLES

DUMMY VARIABLES

How can we make the categorical fields numerical?

Exterior Color of House	Tan	Blue	White
Tan	1	0	0
Blue	0	1	0
White	0	0	1
Blue	0	1	0

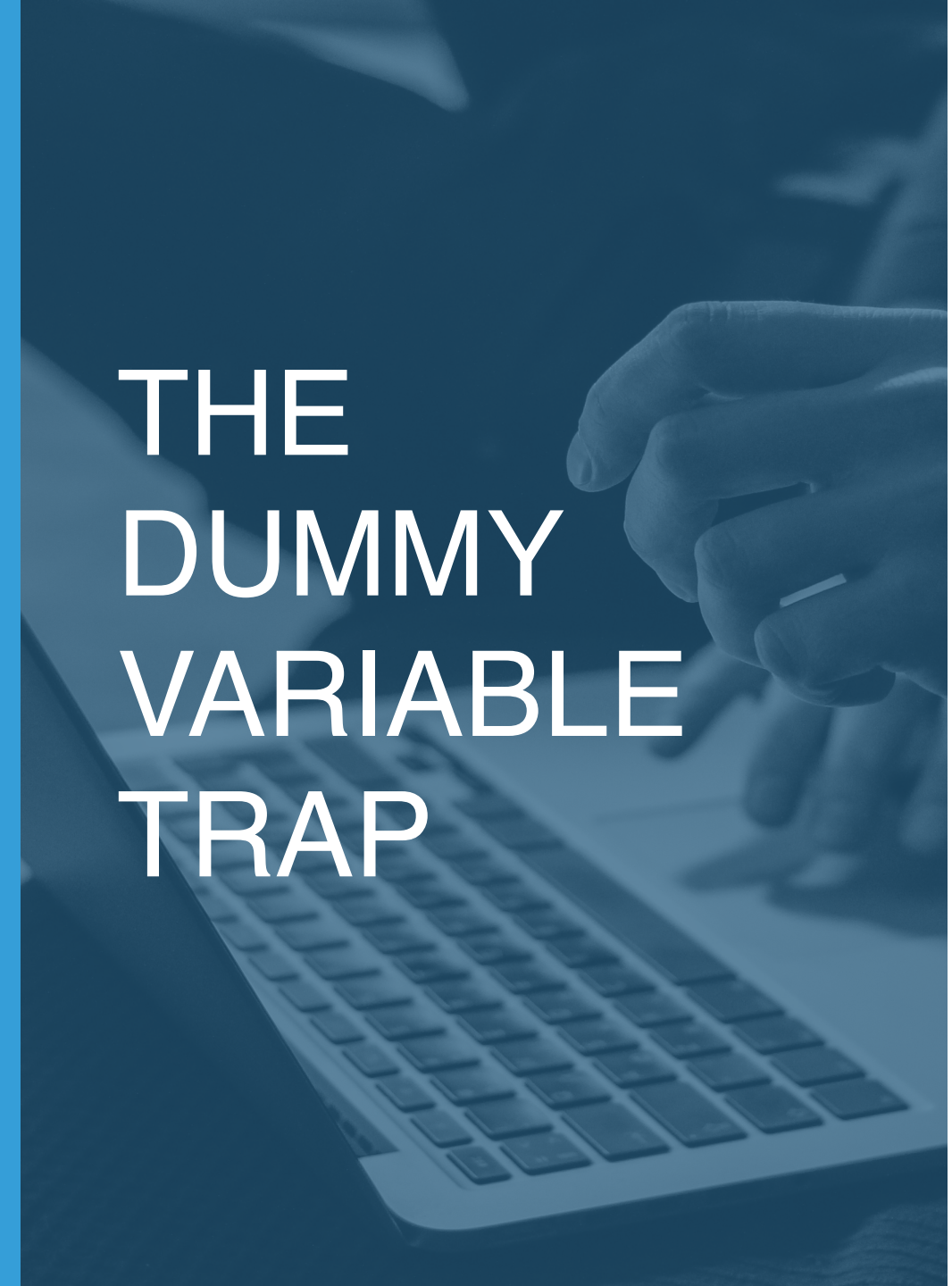
Pandas syntax: pd.get_dummies(my_series)

DUMMY VARIABLES

Do we really need all three new columns to describe the Exterior Color of the House?

Exterior Color of House	Blue	White
Tan	0	0
Blue	1	0
White	0	1
Blue	1	0

Pandas syntax: pd.get_dummies(my_series, drop_first=True)



THE DUMMY VARIABLE TRAP

Linear Regression Equation: $y = \beta_0 + \beta_1 x_1$

With Dummy Variables: $y = \beta_0 x_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3$

X_0	Tan	Blue	White
1	1	0	0
1	1	0	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	1
1	0	0	1

Sum_Color
1
1
1
1
1
1
1
1

This is a case of perfect multicollinearity.

DUMMY VARIABLES

When creating dummy variables for linear regression, one column must be dropped.

Exterior Color of House	Blue	White
Tan	0	0
Blue	1	0
White	0	1
Blue	1	0

With two columns (blue and white), all three colors are represented and we avoid perfect multicollinearity.

Pandas syntax: pd.get_dummies(my_series, drop_first=True)

DUMMY VARIABLES: NAN VALUES

Dummy variables can also be used to capture NaN values in the data.

Last Sold Price	NaN
$540,000	0
NaN	1
$280,000	0
NaN	1

This NaN column contains additional information.

Possibility: when NaN = 1, it means it's a new house, so we could even rename the column as ***'New'***

Pandas syntax: pd.get_dummies(my_series, dummy_na=True)

DUMMY VARIABLES: NAN VALUES

Dummy variables can also be used to capture NaN values in the data.

Last Sold Price	NaN
$540,000	0
NaN	1
$280,000	0
NaN	1

Note: This works with both numerical and categorical features.

Pandas syntax: pd.get_dummies(my_series, dummy_na=True)

DUMMY VARIABLES: ORDINAL DATA

With ordinal data (order matters), there are multiple ways to turn it into a numeric value.

Garage	Partial	Full
Partial	1	0
Full	0	1
None	0	0
Partial	1	0

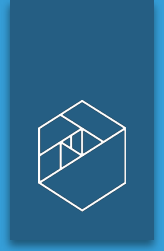
Garage	Garage_Num
Partial	0.5
Full	1
None	0
Partial	0.5

This is a design choice.

DATA TYPES SUMMARY

Numerical (Quantitative)		Categorical (Qualitative)	
Continuous	**Discrete**	**Nominal**	**Ordinal**
Infinite Options *Example: Square Footage*	Finite Options *Example: Number of Bedrooms*	Unordered Categories *Example: Exterior Color of House*	Ordered Categories *Example: No / Partial / Full Garage*

Can use dummy variables to deal with categorical data and also NaN data.



QUESTIONS?