Handling Strategies

Feature	Data Quality Issue	Handling Strategy
AnimalID (object)	Unique identifier not to be used as feature in data preparation	Keep to be able to link example to actual animal, but do not use for data preparation
Name_Intake (category)	Missing Values (30.6%)	Change it so that it has two values - 'No' to represent a missing value and 'Yes' to show that a name was provided.
DateTime_Intake (datetime64[ns])	Some dates failed logical integrity tests.	Revert switched features for failing instances.
FoundLocation (category)	High cardinality (766)	Represent found location using only the city.
IntakeType (category)	None	Keep as is
IntakeCondition (category)	Ambiguous value 'Other'	Impute to mode.
AnimalType_Intake (category)	Ambiguous value 'Other'	Further divide up this value by creating a "Bat" value.
SexuponIntake (category)	None	Keep as is
AgeuponIntake (category)	Not in a useful numeric format. Potential outliers	Convert into days. Leave potential outliers as they are.
Breed_Intake (category)	High cardinality (211)	Divide all values into 'Mixed Breed' and 'Pure Breed'
Color_Intake (category)	High cardinality (115)	Divide all values into 'Dark' and 'Light'
DateTime_Outcome (datetime64[ns])	Some dates failed logical integrity tests.	Revert switched features for failing instances.
DateofBirth (datetime64[ns])	None	Keep as is
SexuponOutcome (category)	None	Keep as is
AgeuponOutcome (category)	Not in a useful numeric format. Potential outliers	Convert into days. Leave potential outliers as they are.
binary_outcome (category)	None	Keep as is

Descriptive Statistics for cleaned categorical features

	count	unique	top	freq
Name_Provided	1000	2	Yes	694
FoundLocation	1000	14	Austin (TX)	820
IntakeType	1000	5	Stray	708
IntakeCondition	1000	4	Normal	892
AnimalType_Intake	1000	5	Dog	563
SexuponIntake	1000	5	Intact Male	332
Breed_Intake	1000	2	Mixed Breed	760
Color_Intake	1000	2	Dark	533
SexuponOutcome	1000	5	Neutered Male	346
binary_outcome	1000	2	0	920

Descriptive Statistics for cleaned continuous features

	count	mean	std	min	25%	50%	75%	max
AgeuponIntake	1000.0	759.812	1116.589203	0.0	68.00	365.0	765.25	7305.0
AgeuponOutcome	1000.0	781.955	1124.826147	0.0	86.75	367.0	808.25	7306.0