* import os
* import numpy as np
* import pandas as pd
* import csv
* import re
* pd.set\_option('display.max\_rows', 10)
* pd.set\_option('display.notebook\_repr\_html', True)
* pd.set\_option('display.max\_columns', 10)
* df = pd.read\_csv(os.path.join('..', 'datasets', 'shelter-animal-outcomes.csv.gz'), index\_col = 'AnimalID'
* **CHART DETAILS**: df
* **ALL COLUMNS**: df.columns
* **OUT**:

Index([u'Name', u'DateTime', u'OutcomeType', u'OutcomeSubtype', u'AnimalType',

u'SexuponOutcome', u'AgeuponOutcome', u'Breed', u'Color'], dtype='object')

**#cleaning up a dataset**

* **SEE UNIQUE VALUES FOR A COLUMN** *(e.g. SexuponOutcome):* df.SexuponOutcome.unique()
* **OUT**:

array(['Neutered Male', 'Spayed Female', 'Intact Male', 'Intact Female',

'Unknown', nan], dtype=object)

* **SEE COUNTS FOR UNIQUE VALUES IN A COLUMN** (*e.g. OutcomeType*): df.SexuponOutcome.value\_counts()
* *note: ignores the NAN*
* **OUT:**

Neutered Male 9779

Spayed Female 8820

Intact Male 3525

Intact Female 3511

Unknown 1093

*Name: SexuponOutcome, dtype: int64*

* *#SEE THAT THERE ARE MANY UNKNOWNS FOR GENDER – but know that neuter = male / spade = female → change the data set*
* **CHECK IF A VALUE IS PROPERLY LABELED** (e.g. if Neutered Male ONLY has Male values):
* 'Male' in 'Neutered Male'
* **OUT**:

True

* **CHECK IF A VALUE IS PROPERLY LABELED** (e.g. if Neutered Male ONLY has Male values):
* 'Female' in 'Neutered Male'
* **OUT**:

False

* **FORMULA TO TAKE OUT ALL VALUES THAT ARE NOT NUMBERS IN SEX**: df.SexuponOutcome = df.SexuponOutcome.apply(lambda sex: 'Unknown' if pd.isnull(sex) else sex)
* *#if sex is not a number, return not a numbers; else return 'Sex'*
* is\_male = df.SexuponOutcome.apply(lambda sex: 'Male' in sex)
* is\_male
* TO CHANGE NAN VALUE IN DATASET:
* df.x
  + 1, 2,3, 4, nan, …etc.
* def sanitize(x):

if pd.isnull(x):

return 0

else:

return x

* df.x = df.x.apply(sanitize)
* *Or*
* df.x – df.x.apply(lambda x: 0 if pd.isnull(x) else x)
* *Or*
* can assign lambda function to a variable:
  + sanitize = lambda x: 0 if pd.isnull(x) else x