ACTIVITY #1

DATA EXPLORATION, CLEANING, AND PREPROCESS

LIBRARIES

1. import numpy as np
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. import seaborn as sns
5. from sklearn import preprocessing

DATA EXPLORATION

1. Read .csv file
   - read_csv()
2. View Data Array Shape
   - # Variables
   - # Samples
   - Print()
3. View Variable info
   - Info()
     - Data Type
     - # non null

	X	Y	Z
0	19	1927	cat
1	NaN	2300	dog
2	15	NaN	bird
3	16	5959	cat
4	16	AB	cat
5	NaN	4594	dog
6	19	1927	cat
7	20	2879	bird��?
8	21	NaN	NaN
9	0	4096	cat
10	A	6730	cat
11	25	0	bird
12	0	2792	dog
13	33	2575	dog??
14	1000	4959	bird
15	19	1927	cat
16	36	4580	dog
17	40	5869	NaN
18	NaN	4178	dog
19	45	NaN	cat

(20, 3)

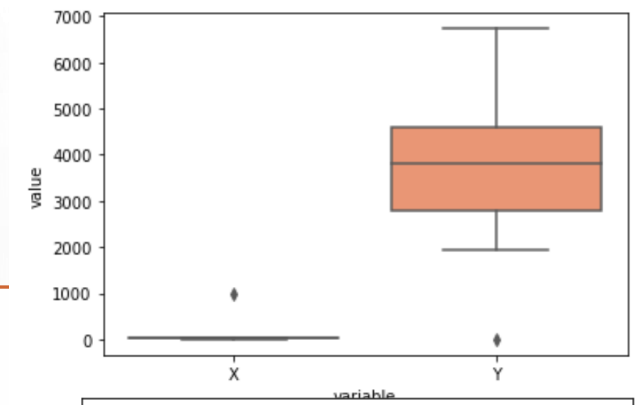
DATA CLEANING

1. Correct Errors (delete non ASCII)
   - replace()
2. Convert data type from object to suitable types
   - X -> int64
   - Y -> float64
   - Z -> string
3. Drop duplicate samples (rows)
   - drop_duplicates()
4. View Variable Statistics
   - describe()
5. Drop rows with NA >1
   - dropna()
6. Replace NA
   - fillna()
     - X, Y with statistics mean or median
     - Z with previous rows

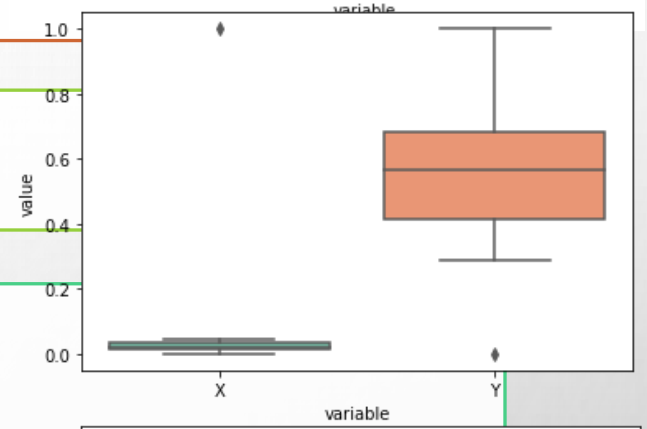
	X	Y	Z
0	19	1927.0	cat
1	20	2300.0	dog
2	15	3817.0	bird
3	16	5959.0	cat
4	16	3817.0	cat
5	20	4594.0	dog
7	20	2879.0	bird
9	0	4096.0	cat
10	20	6730.0	cat
11	25	0.0	bird
12	0	2792.0	dog
13	33	2575.0	dog
14	1000	4959.0	bird
16	36	4580.0	dog
17	40	5869.0	dog
18	20	4178.0	dog
19	45	3817.0	cat

DATA TRANSFORM

1
- Transform data
  - MinMaxScaler() / StandardScaler()

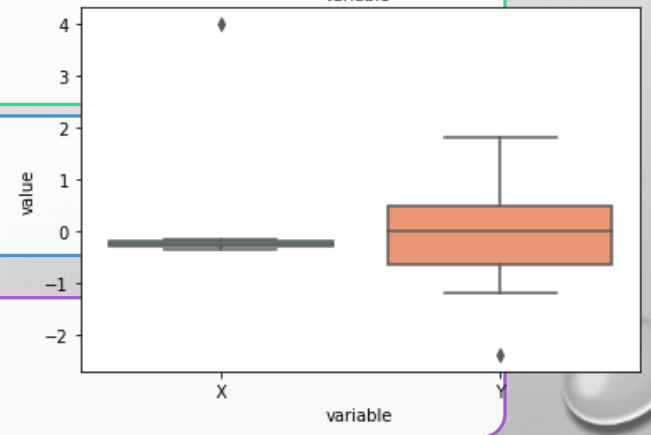


2
- Show Boxplot X,Y
  - sns.boxplot() or pd.boxplot()



3
- Remove outlier
  - Q1 = scaled_features[cols].quantile(0.25)
  - Q3 = scaled_features[cols].quantile(0.75)
  - IQR = Q3 - Q1
  - X, Y < Q1 – (1.5*IQR) | X,Y > Q3 + (1.5*IQR)

4
- Transform data
  - MinMaxScaler() / StandardScaler()



5
- Show Boxplot X,Y
- sns.boxplot() or pd.boxplot()

DATA CATEGORY LABEL

LabelEncoder() OneHotEncoder()

1. Reset drop index
   - reset_index()
2. Convert Category to Label
   - preprocessing.LabelEncoder()
3. Convert Category to Binary Label
   - preprocessing.OneHotEncoder()
4. Join LabelEncoder result, OneHotEncoder result to dataframe

	X	Y	Z	Z_category	bird	cat	dog
0	0.422222	0.000000	cat	1	0.0	1.0	0.0
1	0.444444	0.077660	dog	2	0.0	0.0	1.0
2	0.333333	0.393504	bird	0	1.0	0.0	0.0
3	0.355556	0.839475	cat	1	0.0	1.0	0.0
4	0.355556	0.393504	cat	1	0.0	1.0	0.0
5	0.444444	0.555278	dog	2	0.0	0.0	1.0
6	0.444444	0.198209	bird	0	1.0	0.0	0.0
7	0.000000	0.451593	cat	1	0.0	1.0	0.0
8	0.444444	1.000000	cat	1	0.0	1.0	0.0
9	0.000000	0.180096	dog	2	0.0	0.0	1.0
10	0.733333	0.134916	dog	2	0.0	0.0	1.0
11	0.800000	0.552363	dog	2	0.0	0.0	1.0
12	0.888889	0.820737	dog	2	0.0	0.0	1.0
13	0.444444	0.468665	dog	2	0.0	0.0	1.0
14	1.000000	0.393504	cat	1	0.0	1.0	0.0