(2). Data Understanding: Data Quality Plan for the cleaned CSV file.

The initial list of issues :

- **The integrity checks** of the dataframe indicated irregular values in place for the continuous features.
    - Check 1 highlighted that some registered date outcome values came before date intake of the animals. These rows were dropped
    - Check 2 highlighted that some age intake values were negative. These rows were dropped
- **Different scale**: *Age upon Intake* and *Age upon Outcome* values not scalable. (years, months, weeks, days)
- **Duplication of data** : *MonthYear_Intake* and *Datetime_Intake*, *MonthYear_Outcome* and *Datetime_Outcome*, *Name Intake* and *Name Outcome*, *Animal Type_Intake* and *Animal Type_Outcome*, *Breed Intake* and *Breed Outcome* and *Color Intake* and *Color Outcome* were registering the duplicate values.
- **Presence of outliers** - There are a significant number of outliers present across a range of different features. They initially look plausible but will need to be investigated further.
- There is **high cardinality** of some of the categorical data most importantly *Found Location*, *Color Outcome* and *Breed Intake* which will make these values difficult to deduce meaningful information from

- Propose solutions to deal with the problems identified.

1. **Integrity Checks**
    - Drop invalid values ie values that fail the test.
2. **Differing scale**
    - *Age upon Intake* and *Age upon Outcome* values not scalable, measured in years/ months/ weeks/ days. One mapping needs to be picked for consistency.
3. **Duplication of data**
    - Categorical features *MonthYear_Intake* and *Datetime_Intake*, *MonthYear_Outcome* and *Datetime_Outcome*, *Name Intake* and *Name Outcome*, *Animal Type_Intake* and *Animal Type_Outcome*, *Breed Intake* and *Breed Outcome* and *Color Intake* and *Color Outcome* require just one of each value to be recorded. Drop duplicate columns
4. **Presence of outliers**
    - There are a significant number of outliers present across a range of different features. They initailly look plausible but will need to be investigated further. If they don't make sense they will be removed.
5. **High Cardinality**
    - The high cardinaltiy of some categorical values need to be investigated individually for columns *Found Location*, *Color Outcome* and *Breed Intake*.

- Apply solutions to obtain a new CSV file where the identified data quality issues were addressed

Summary of data quality plan:

Variable Names	Data Quality Issue	Handling Strategy
Animal ID	Scale	Do Nothing
Name_Intake	Scale, missing values	Do Nothing
DateTime_Intake	Invalid cardinality	Replace with duration, drop rows
MonthYear_Intake	Duplicate	Drop Feature
Found Location	Scale	Top 10 values
Intake Type	Undefined value	Do nothing
Intake Condition	Undefined value	Do nothing
Animal Type_Intake	Outliers	Do Nothing
Sex upon Intake	Scale	Do Nothing
Age upon Intake	Invalid cardinality	Replace with age intake in days, drop rows
Breed_Intake	Scale	Do Nothing
Color_Intake	Scale	Do Nothing
Name_Outcome	Duplicate	Drop Feature
DateTime_Outcome	Invalid cardinality	Replace with duration, drop rows
MonthYear_Outcome	Duplicate	Drop Feature
Date of Birth	Outliers	Do Nothing
Animal Type_Outcome	Duplicate	Drop Feature

Variable Names	Data Quality Issue	Handling Strategy
Sex upon Outcome	Scale	Do nothing
Age upon Outcome	Invalid cardinality	Replace with age outcome in days, drop rows
Breed_Outcome	Duplicate	Drop Feature
Color_Outcome	Duplicate	Drop Feature
binary_outcome	Invalid cardinality	convert to categorical type