Data Quality Report : Initial Findings

- **Background**
  This report summarizes outcomes of data processed for “Shelter_1-1_cleaned.csv”. Dataset has mostly categorical features and is relatively clean except of few logical errors. Total 6 duplicate columns were located and dropped. No rows were found illogical or needed to be dropped.

- **Review of Logical Integrity**
  - test_1 : Check for entries having Date of birth GREATER THAN Date of intake into shelter.
    - 0 cases found.
  - test_2 : Check for entries having Date of birth GREATER THAN Date of outcome from shelter.
    - 0 cases found.
  - test_3 : Check for entries having Date of Intake GREATER THAN Date of outcome from shelter.
    - 17 cases found.
  - test_4 : Check for entries having Age upon Intake GREATER THAN Age upon outcome from shelter.
    - 8 cases found.
  - test_5 : Check if “Neutered Male” status for feature sex upon intake IS CHANGED in feature sex upon outcome for any animal.
    - 0 cases found.
  - test_6 : Check if “Spayed Female” status for feature sex upon intake IS CHANGED in feature sex upon outcome for any animal.
    - 0 cases found.
  - test_7 : Check if difference between Date of Intake and Date of birth in Weeks is LESS THAN Age upon intake.
    - 43 cases found.
  - test_8 : Check if difference between Date of Outcome and Date of birth in Weeks is LESS THAN Age upon Outcome.
    - 16 cases found.

- **Review Categorical features**
  - binary_outcome : Target variable “binary_outcome” denotes 0 for positive and 1 for negative outcome. This goes against in general notion hence it is changed in stage 1.1.3.4.
  - 6 duplicate columns were found and dropped.
  - Breed_Intake and Color_Intake features have too much granularity with proportion to the data cardinality

Feature	Feature duplicate
Name_Intake	Name_Outcome
DateTime_Intake	MonthYear_Intake
Animal Type_Intake	Animal Type_Outcome
Breed_Intake	Breed_Outcome
Color_Intake	Color_Outcome
DateTime_Outcome	MonthYear_Outcome

- **Review Datetime features**
  - 17 Entries where date of Outcome is smaller than Date of Intake are found. This is illogical and need to be addressed. It is mostly given interchanged data entry.

- **Review Continuous features**
  - Age upon Intake and Age upon Outcome is numeric data with 4 different units. And, it has imprecision to to flooring approach.
  - 8 Entries where Age upon Outcome is smaller than Age upon Intake are found. This is illogical and need to be addressed. It is mostly given interchanged data entry.

- **Action items**
  - Logical Integrity
    - DateTime Intake > DateTime Outcome
      - Interchange respected values for Date of intake and Date of Outcome
    - Age upon Intake > Age upon Outcome
      - Resolve errors for Age upon intake and Age upon Outcome due to data entry unit discrepancy errors.
    - Value for Age upon intake
      - Calculate Age upon Intake by directly referring DateTime_Intake and Date of Birth features
    - Value for Age upon outcome
      - Calculate Age upon Outcome by directly referring DateTime_Outcome and Date of Birth features
  - Very large set of unique values with respect to dataset cardinality
    - Color
      - Keep only major sections for category name for Color_Intake feature
    - Breed
      - Keep only major sections for category name for Breed_Intake feature
  - Outliers
    - Review other outliers
    - Strip prefixed '*', and replace invalid entries and null values with string "Unknown" for feature "Name_Intake"

Data Quality Report : Summary sheets

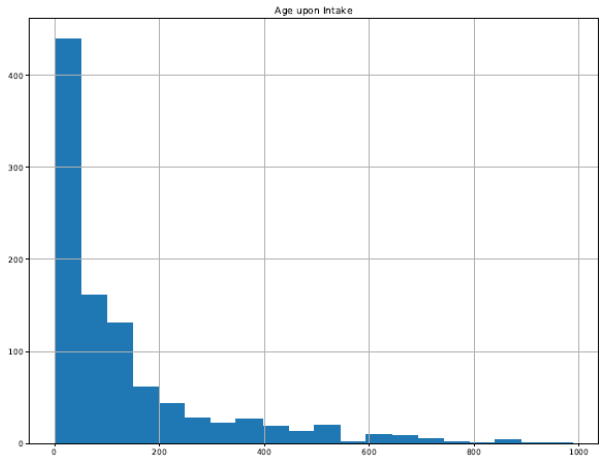
- **Categorical features**
- Descriptive Statistics

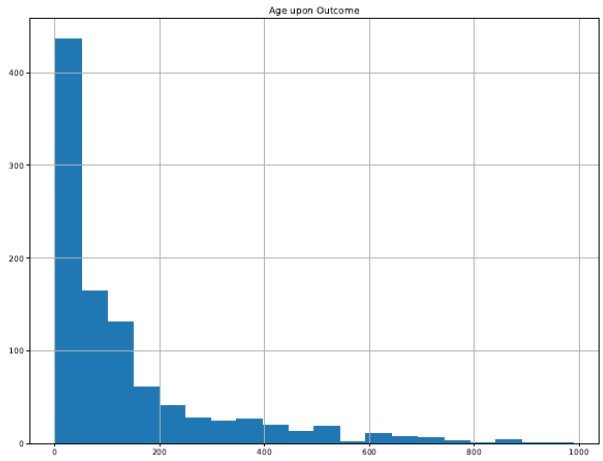
	count	unique	top	freq
Animal ID	1000	1000	A812362	1
Name_Intake	663	588	Charlie	4
Found Location	1000	778	Austin (TX)	174
Intake Type	1000	5	Stray	714
Intake Condition	1000	8	Normal	871
Animal Type_Intake	1000	4	Dog	523
Sex upon Intake	1000	5	Intact Male	339
Breed_Intake	1000	213	Domestic Shorthair Mix	298
Color_Intake	1000	111	Black	92
Sex upon Outcome	1000	5	Neutered Male	347
binary_outcome	1000	2	1	903

- **Continuous features**
- Descriptive Statistics

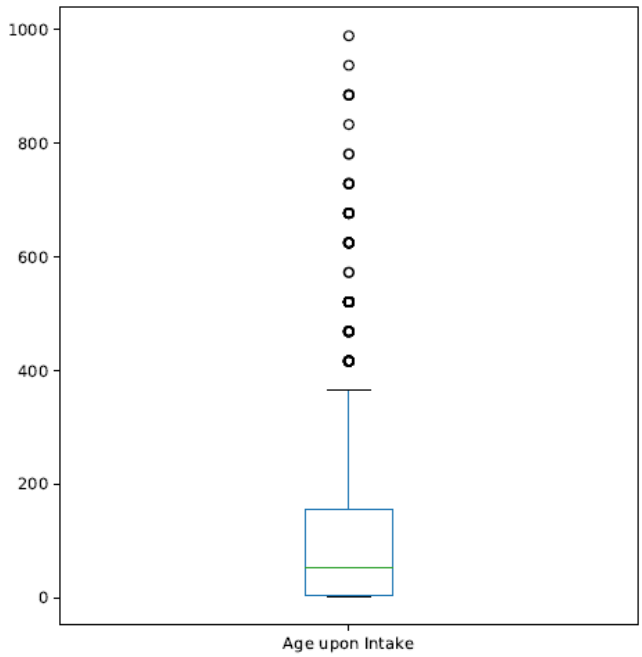
	count	mean	std	min	25%	50%	75%	max
Age upon Intake	1000	118.252	169.3953	1	4	52	156	988
Age upon Outcome	1000	119.683	169.7275	1	9	52	156	988

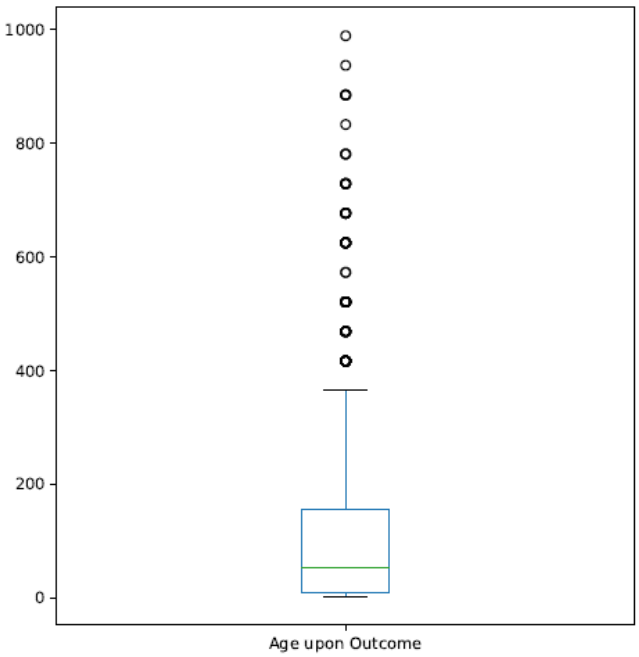
- Histogram





- Box Plot

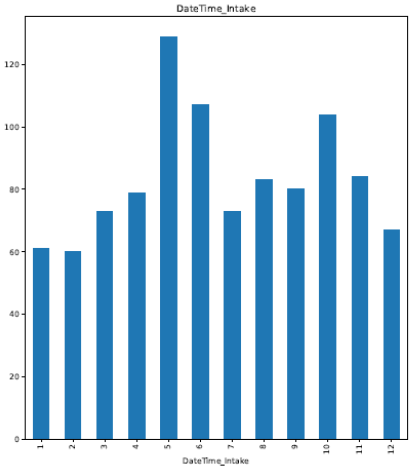


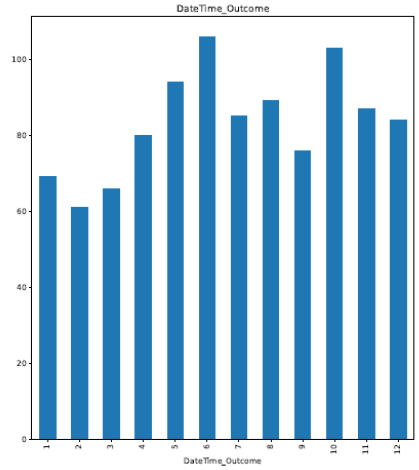


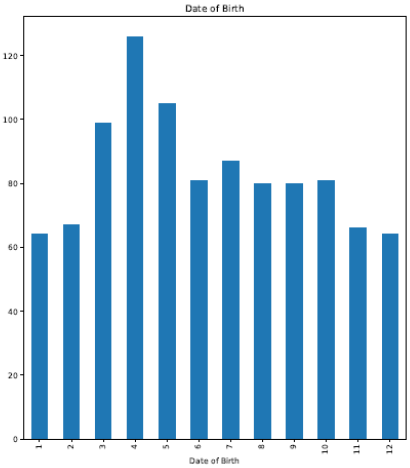
- **Datetime features**
- Descriptive Statistics

	count	unique	top	freq	first	last
DateTime_Intake	1000	995	9/26/2017 12:30	2	10/1/2013 11:15	2/2/2020 23:19
DateTime_Outcome	1000	997	10/3/2017 0:00	2	10/1/2013 12:27	1/25/2020 19:04
Date of Birth	1000	861	3/17/2017 0:00	5	10/12/1997 0:00	10/17/2019 0:00

- Bar plot







- Box plot