Data Quality Report – Initial Findings

1. Overview

This report will outline the initial findings based on the cleaned dataset (AnimalWelfare_1-1_cleaned.csv). It will summarise the data, describe the various data quality issues observed and how they will be addressed. Please see appendix for some background to this dataset. Appendix includes terminology, explanations and changes made to the original dataset. This also includes descriptive statistics of features, histograms and box plots used to visualise the data.

On first indication the dataset appears relatively unclean. There are no duplicate columns yet there are missing values and columns with irregular cardinalities. Negative values for continuous data were also observed. In addition, some high value outliers were present. Also, several logical tests were carried out on the data and inconsistencies were found.

2. Summary

Several tests were carried out to check the logical integrity of the data. This brought about a significant number of failures of the data. In total 9 instances of irrational data was observed. For example, in 7 instances the age upon intake was found to be greater than the age upon outcome. This is clearly impossible. This irrational data will need to be dealt with and should be checked with the domain expert. See logical integrity section for further details.

For the continuous features there was the inclusion of negative values including -1 and -2 which are highly likely to be invalid. These values need to be addressed as they are appear in features that cannot logically contain negative numbers. These values should then be evaluated on a feature by feature basis to see if imputation is possible, i.e. mapping to the median value or other reasonable values based on evaluation of other features(e.g. intake condition), details will be provided in the following section. In addition, there was outliers present across the continuous feature set. However, on first indication these values appear to be plausible but should be investigated further.

For the categorical values several changes are recommended. There are 12 main features including a wide variety of information about the animals in the shelter, such as sex, breed intake and animal type. Among all those features, breed intake, found location and colour intake have significant irregular cardinalities, which need merging the cardinalities into the same category. Sex upon intake and sex upon outcome have 86 values in 'Unknown' category, which are considered as missing data. Rows affected might need to apply imputation or row drop, but further investigation should be made to aid the decision.

3. Review Logical Integrity

3 tests were carried out. The failures are below;

- Test 1 - Check if any animals have number of age upon intake > than age upon outcome (impossible)
  - 7 cases found
- Test 2 - Check if any animals have number of age upon intake <0 (impossible)
  - 2 cases found
- Test 3 - Check if any animals have number of age upon intake <0 (impossible)
  - 0 cases found

4. Review Continuous Features

4.1. Descriptive Statistics

There are 2 continuous features. Both continuous features can be grouped into a age main categories which will be summarised below:

- Age (Count 2)
  - Both features measure age of animals, however one measure the age upon intake to the animal shelter, one measure the age upon outcome. Both features have very similar values as expected and make sense in relation to each other.
  - The feature age upon intake has two animals with negative values -1 and -2 and will need to be addressed.
  - The feature age upon outcome feature has a minimum value of zero, this could be due to the animal was newborn at the time of intake, and getting adopted/die

on the same day it was taken in. Thus while odd at a glance, I don't believe there is anything unusual about the minimum values of the feature.

- Max value for age upon intake is 7670 days, with mean of about 784; while for age upon outcome is 7671 days, with mean of about 803. It appears that both features are covering a wide range of animals in terms of age if the max value, i.e. the outlier is plausible.

Overall the features showed plausible distributions. The outliers will be investigated further but no immediate action expected.

4.2. Histograms

All histograms can be found on the appendix and in the accompanying notebook. From the histograms, we can see that both age upon intake and age upon outcome appears to be exponentially decreasing.

4.3. Box plots

All box plots can be found on the appendix and in the accompanying notebook. Both age upon intake and age upon outcome have outliers. Outliers in both plots are larger than the max cut off point. Some of the outliers are also pretty extreme, being multiple standard deviations away from the mean. However, outliers in both plots do seem make sense. As this is age of animals, it is expected that if the age upon intake have outliers, then those animals will then have higher outcome age and thus the age upon outcome box plot reflects this with corresponding high outliers. Also, the average age of certain types of animals can be much greater than the others, so it is plausible that the outliers are still be in normal age range of those animals. Thus again, outliers will be investigated further but no immediate action expected.

5. Review Categorical Features

5.1. Descriptive Statistics

There are 12 categorical features in the dataset, 1 of which is the target and will not be evaluated here. Some of the 11 remaining categorical features will be grouped into main categories based on their natures and count will be stated for those features. The features will be summarised below;

- Datetime(Count 3)

  - Date time intake, date time outcome and date of birth are grouped to this category, since they all describe date time relevant data.

  - All three features are **measured in months** in order to cast into categorical features and do further analysis, as mentioned in the first part in the accompanying notebook.

  - Looking at the top category, we can see that about 13% animals are born in April. These are as expected. Autumn and spring are estrus cycles for many kinds of animals including dogs, adding with roughly two months pregnancy, it is not surprised to have more animals to be born in April.

  - About 10% of animals are taken in June and about 10% of animals are left in in August. Both intake and outcome peak period are in summer, this might represent some kind of correlation which might help us better predict the outcome, but further investigation is needed here.

  - Overall features belongs to this category show plausible distributions.

- Sex(Count 2)
  - Sex upon intake and sex upon outcome are grouped to this category, since they all describe characteristics of animal sex.

  - Looking at the top category, 338/1000 animals are intact male upon intake, indicate a number of animals are not yet sterilised before getting into the animal shelter.

  - Looking at the top category, 339/1000 are neutered male when they left, indicate the majority of animals got sterilised before leaving the shelter.

  - Both Sex upon intake and sex upon intake have "unknown" category for 86 out of 1000. Unknown is synonymous with missing data. There is a possibility that animals are hermaphrodite, or the data is simply missing because, for example, the animals involved stay for a too short period inside the shelter that veterinarians do not even have enough time to check for their sex. Thus further investigation should be made and the issue will need to be addressed.

- Found location
  - There are 769 different Found Location and looking at the top category, around 18% animals are found in Austin(TX). There are too many unique locations and

thus doesn't actually tell us a lot. The cardinality problem will need to be addressed.

- Color intake
  - The colour intake is 111/1000 with black/white. Again, there are 115 unique value counts, it is arguably too detailed with that much unique values and doesn't actually tell us a lot. The cardinality problem will need to be addressed.

- Breed intake
  - The breed intake is 290/1000 with domestic shorthair mix. Also, there are 204 unique value counts, it is arguably too detailed with that much unique values and doesn't actually tell us a lot. The cardinality problem will need to be addressed.

- Intake type
  - About 72% of animals are taken in as stray animals, this is as expected as homeless animals on the street are common in modern city and thus many are found and rescued by animal controls officer on street.

- Intake condition
  - It appears most animals are in normal condition, as many as 878/1000.

- Animal type intake
  - About 54% of intake animals are dogs, indicates that dogs constitute the main type of animals in the shelter.

5.2. Histograms

All bar plots can be found on the appendix and in the accompanying notebook. The analysis of each bar plot will be discussed in this section :

- Datetime(Count 3)

  - Date time outcome plot shows as that ‘8’ category, which means August is the most common for outcome for these animals, follows by July, and the statistics decrease gradually and there are generally not much difference between each months.

  - Date time intake plot indicates that the amount of animals taken in many months are quite similar, with the ‘6’ and ‘5’ categories, which means June and May accounted for the largest proportion among date time intake.

    - Date of birth plot indicates that ‘4’ category, which means April is the most common for births for these animals.

- Sex(Count 2)

    - The majority of sex upon intake is intact animals including both male and female, with other categories roughly equally close to each other.

    - The majority of sex upon outcome is neutered male and spayed female, which are roughly twice as high as the intact animals. This indicates that the majority of animals are sterilised in the shelter.

    - As mentioned in the last section, both features have 86 values are in the “unknown” category, this issue will need to be addressed.

- Found location and breed intake

    - Found location and breed intake plots have too many categories that cannot even see the name of the categories and distribution of them. Further actions like merging cardinality are needed to handle the problem.

- Color intake

    - Color intake plot has the same problem but the situation is slightly better, it is dominated by “black/white”category and has an exponentially decreasing plots. However, most categories are unclear since they are too crowded. Further actions like merging cardinality are needed to handle the problem.

- Intake type

    - Intake type is dominated by “stray”category, indicating that homeless animals on the street made up the majority of animals in the shelter.

- Intake condition

    - Intake condition is dominated by “normal”category, indicating that most animals by the shelter are not suffering from illnesses, injuries or impairments.

- Animal type intake

    - Dogs accounted for the largest proportion among intake types of animals, follows by cats. The amount of all other animal types are hugely smaller than these two categories.

- Binary outcome

- Binary outcome plot shows us that most animals have a positive outcome.

6. Action to take

5 main actions will be taken, summarised below;

- Negative Age
  - Imputation will be performed where possible. Otherwise values will be changed to 0, with a note to revisit those features later.
- Negative days in shelter
  - Imputation will be performed where possible. Otherwise if number is low, remove those rows, if number is high, remove the feature.
- irregular cardinality
  - Merge the cardinalities into the same category.
- Unknown sex
  - Investigation should be made if imputation or other actions should be applied or rows should be dropped.
- Outliers
  - Investigation should be made if imputation or other actions should be applied or rows should be dropped.

7. References

[1] Jarne P, Auld JR The distribution of self-fertilization among hermaphroditic animals

https://bioone.org/journals/Florida-Entomologist/volume-97/issue-2/024.097.0223/Some-Life-History-Traits-and-Diet-Selection-in-Philomycus-carolinianus/10.1653/024.097.0223.full

[2] Schertz mating seasons for dogs

https://schertzanimalhospital.com/blog/mating-season-female-dog-springtime/

[3]Glendale Veterinary Clinic Pregnancy in cats and dogs

https://www.glendalevetclinic.com/useful-links/8-news-articles/21-pregnancy-in-cats-and-dogs

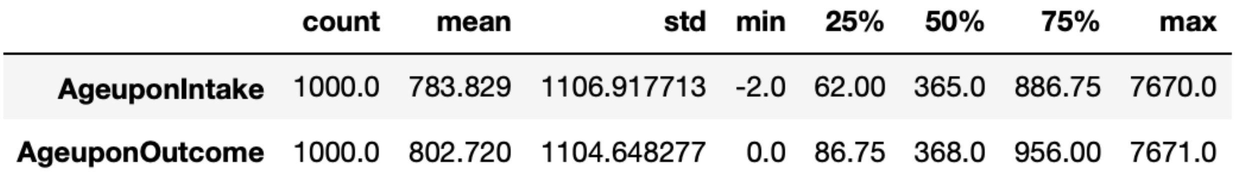
8. Appendix

8.1. Terminology

- “Sterilisation” is a surgical procedure in which a part of your pet's reproductive organ is removed to permanently stop it from reproducing.
- “Hermaphrodite” means a person or animal having both male and female sex organs or other sexual characteristics
- “Neutered male” is a male animal that testicles are removed.
- “Estrus cycles” is the period in the sexual cycle of female mammals

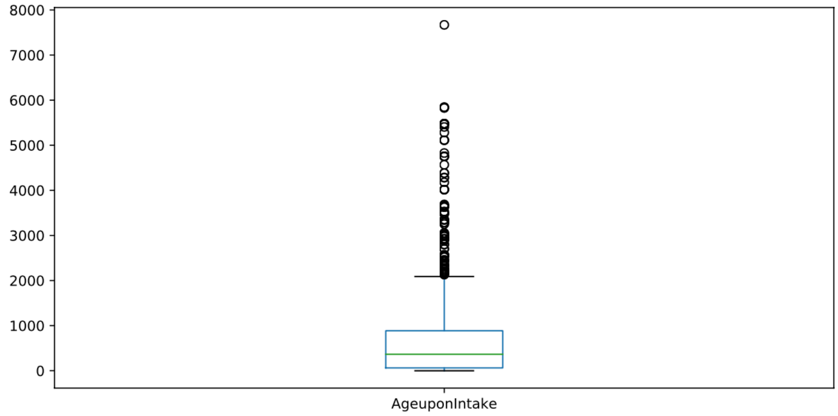
8.2. Continuous Features

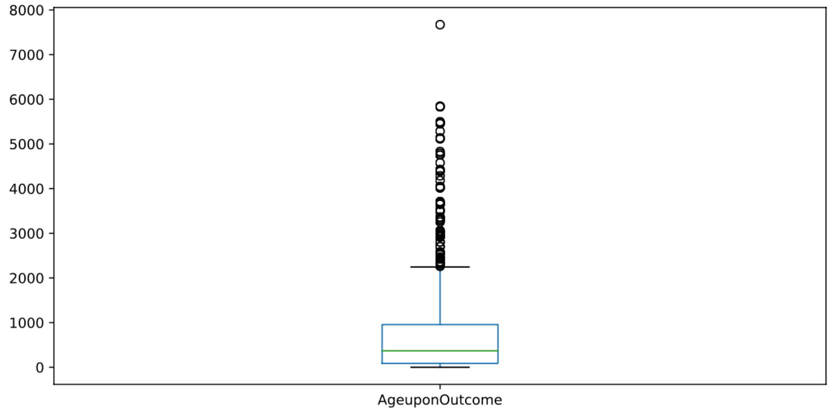
Descriptive Statistics



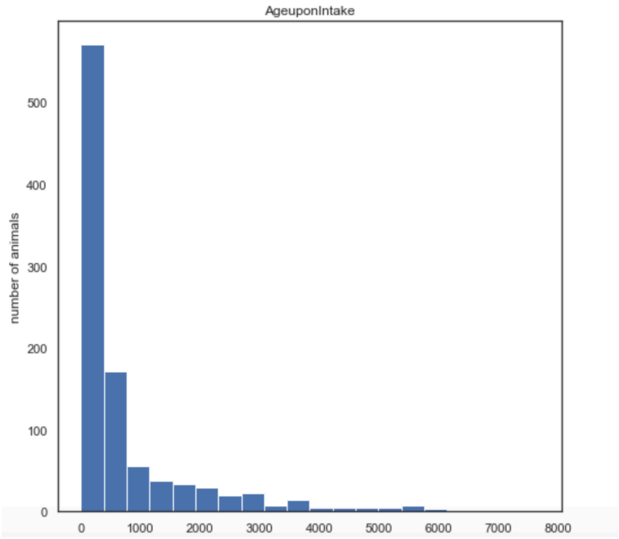
	count	mean	std	min	25%	50%	75%	max
AgeuponIntake	1000.0	783.829	1106.917713	-2.0	62.00	365.0	886.75	7670.0
AgeuponOutcome	1000.0	802.720	1104.648277	0.0	86.75	368.0	956.00	7671.0

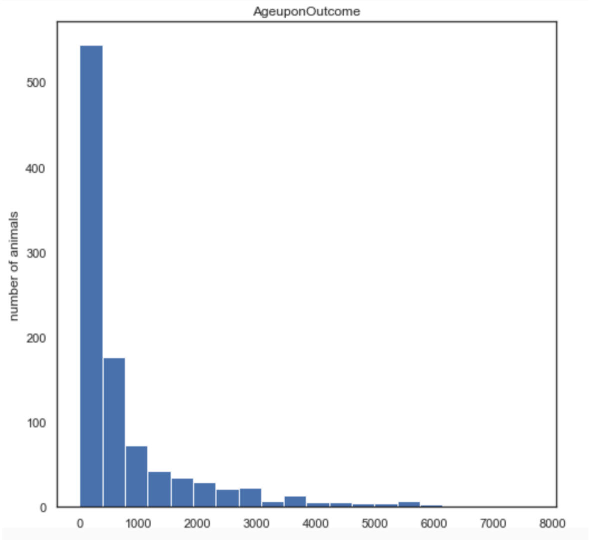
Box Plots





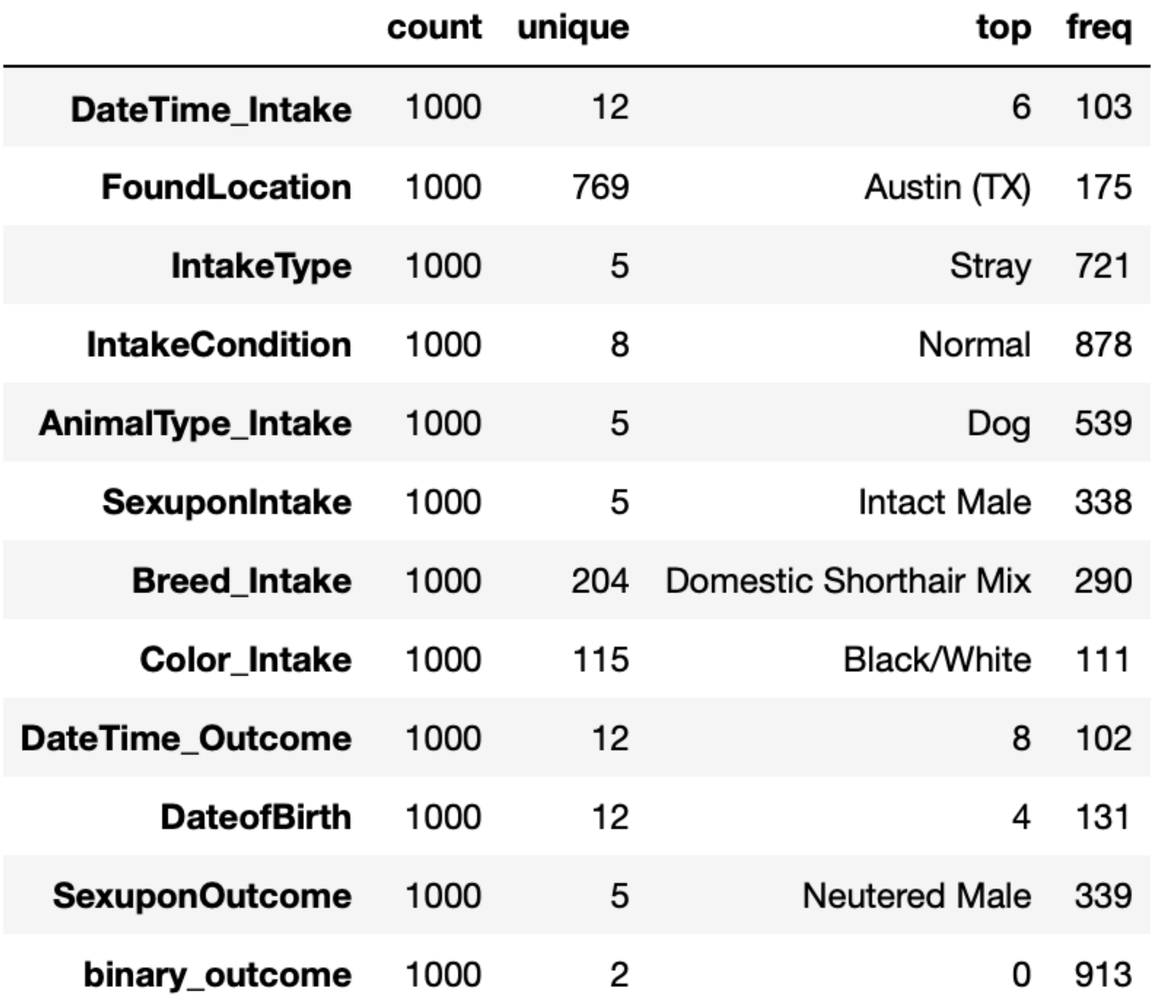
Histograms





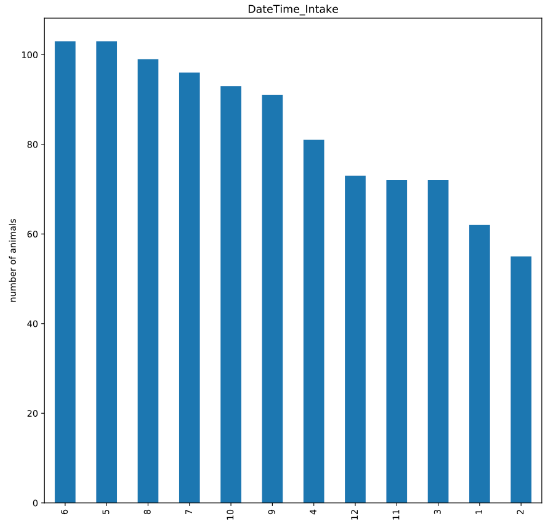
8.3. Categorical Features

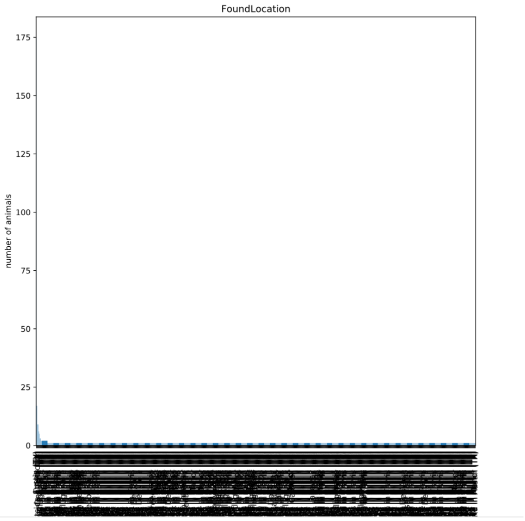
Descriptive Statistics

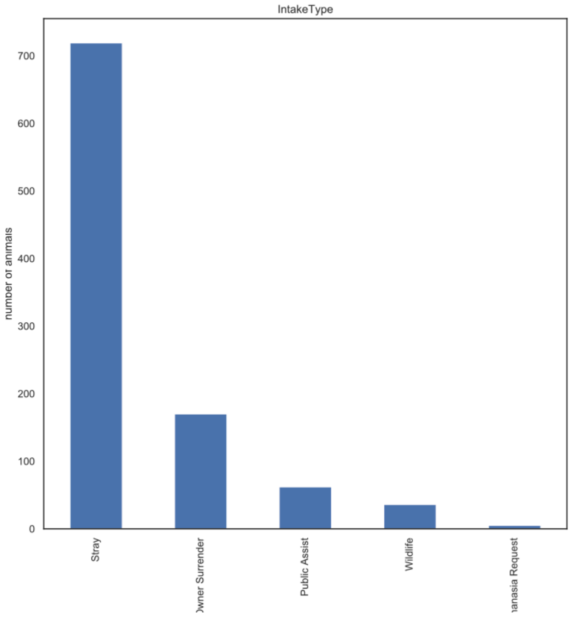


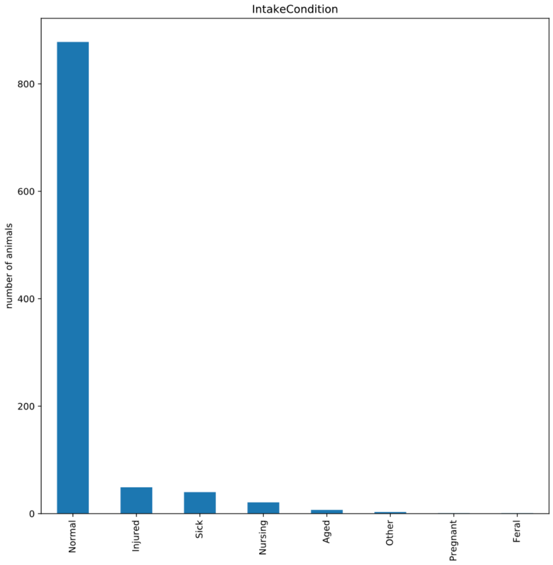
	count	unique	top	freq
DateTime_Intake	1000	12	6	103
FoundLocation	1000	769	Austin (TX)	175
IntakeType	1000	5	Stray	721
IntakeCondition	1000	8	Normal	878
AnimalType_Intake	1000	5	Dog	539
SexuponIntake	1000	5	Intact Male	338
Breed_Intake	1000	204	Domestic Shorthair Mix	290
Color_Intake	1000	115	Black/White	111
DateTime_Outcome	1000	12	8	102
DateofBirth	1000	12	4	131
SexuponOutcome	1000	5	Neutered Male	339
binary_outcome	1000	2	0	913

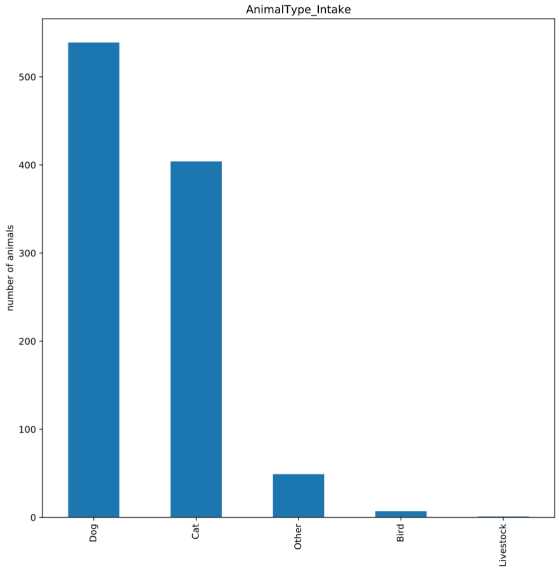
Box Plots

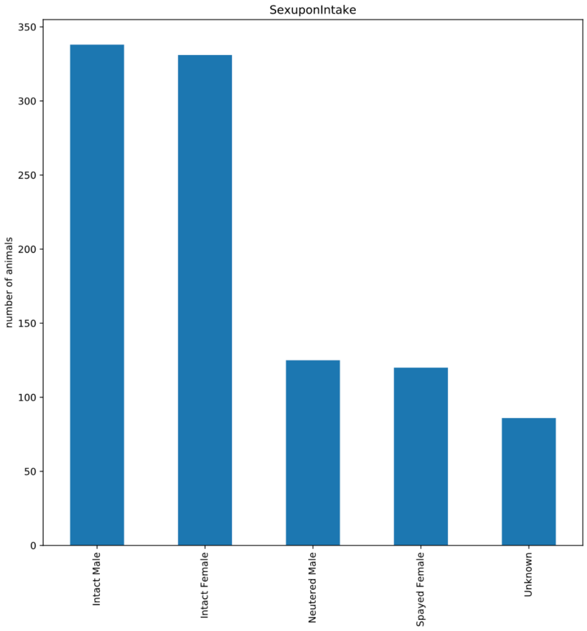

DateTime_Intake
number of animals
0
20
40
60
80
100
6
5
8
7
10
9
4
12
11
3
1
2

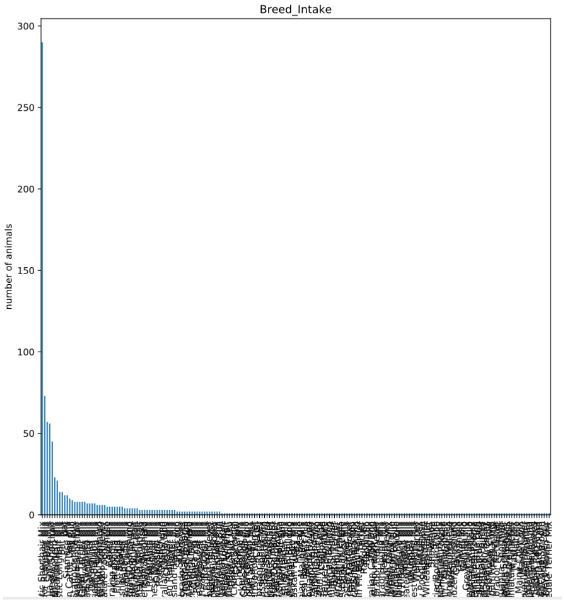

FoundLocation
number of animals
0
25
50
75
100
125
150
175

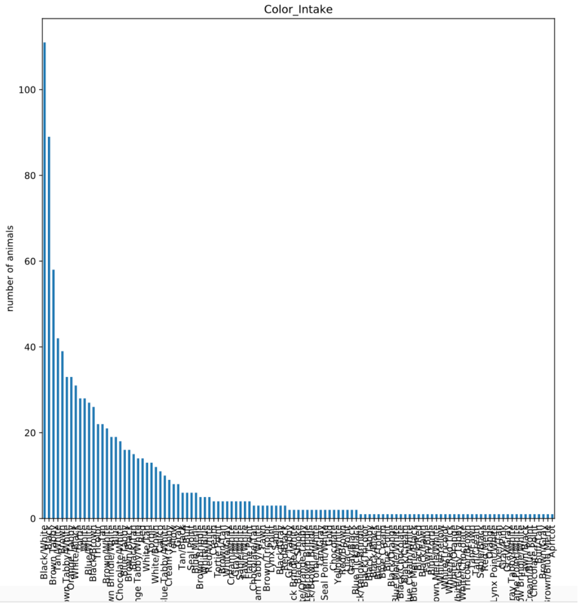

IntakeType
number of animals
0
100
200
300
400
500
600
700
Stray
wner Surrender
Public Assist
Wildlife
anasia Request

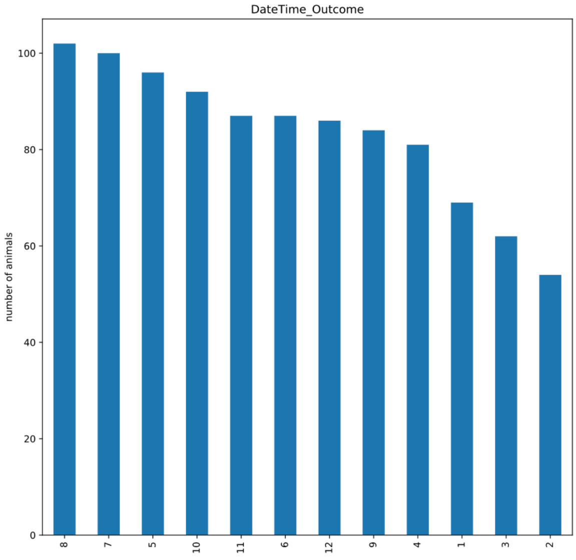

IntakeCondition
number of animals
0
200
400
600
800
Normal
Injured
Sick
Nursing
Aged
Other
Pregnant
Feral

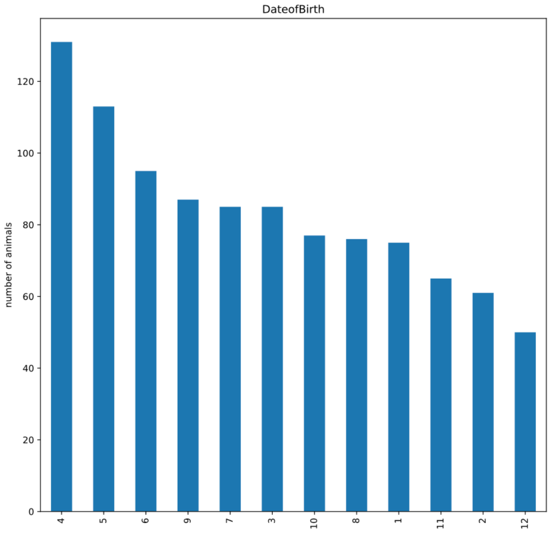

AnimalType_Intake
number of animals
0
100
200
300
400
500
Dog
Cat
Other
Bird
Livestock

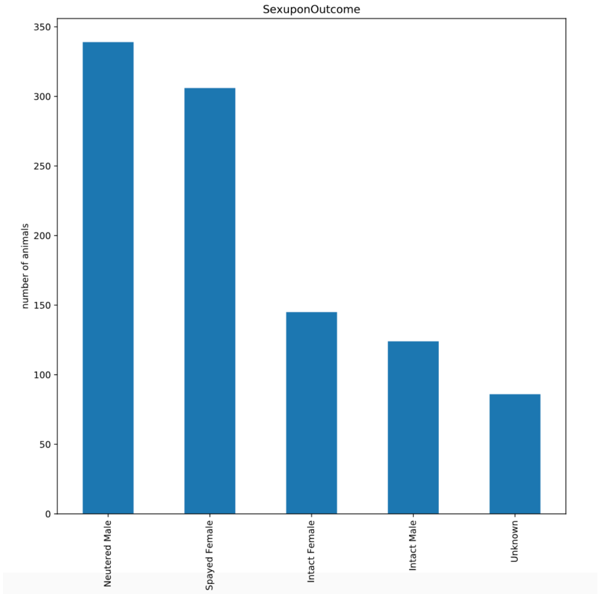

SexuponIntake
number of animals
0
50
100
150
200
250
300
350
Intact Male
Intact Female
Neutered Male
Spayed Female
Unknown

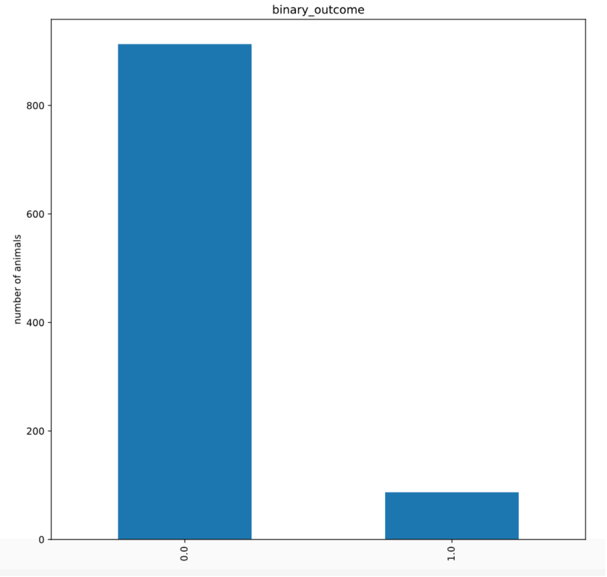

Breed_Intake
number of animals
300
250
200
150
100
50
0


Color_Intake
number of animals
100
80
60
40
20
0
Black/White
Brown Tabby
Seal Point
Lynx Point
Apricot


DateTime_Outcome
number of animals
100
80
60
40
20
0
8
7
5
10
11
6
12
9
4
1
3
2


DateofBirth
number of animals
120
100
80
60
40
20
0
4
5
6
9
7
3
10
8
1
11
2
12


SexuponOutcome
number of animals
350
300
250
200
150
100
50
0
Neutered Male
Spayed Female
Intact Female
Intact Male
Unknown


binary_outcome
number of animals
800
600
400
200
0
0.0
1.0