COMP47350-Data Analytics

DATA QUALITY REPORT -AnimalWelfareRisk-19200494.CSV

1. BACKGROUND INFORMATION

The initial report summarizes the observations derived from exploratory data analysis on dataset AnimalWelfareRisk-19200494, Which is a subset of a dataset obtained from the animal shelter in Texas. The dataset contains information regarding all the animals brought to shelter and the final outcome of the animal i.e. if it died in shelter or was adopted/ returned to the wild. The report will present how good the quality of all the attributes are, if there are anomalies, ambiguities and inaccuracies in data. And finally how to deal with these anomalies to make the data viable for predictive analytics

Note
An appendix is provided which contains additional information on the dataset. These include a summary of the data used and terminologies used.

2. Overview

Initial impression when viewing the data is that the data seems to be quite clean and ready to use for predictive analytics. The dataset has a good number of features with dimension of 1000 entries and 22 attributes. Getting a summary view of the data and we start seeing multiple issues with the data. Since we have numeros issues we will explain each issue below.

2.1 DUPLICATES

The first major issue that is encountered is the problem of duplicate columns. The dataset seems to have a number of attributes which has mostly the same information. Such columns are of no value and must be dropped from the dataset. The problem of duplicate values is not inherently a data quality issue of the dataset. Example the columns ‘Color_Intake’ and ‘ColorOutcome’ in the given dataset have exactly the same value . But it doesn’t mean that the design of the dataset is flawed. Because there might be a chance that the color of an animal was different when it was brought to shelter and when it was taken from shelter it has changed. Therefore we can justify the existence of the two attributes. But in our dataset we have the same values for both the attributes hence declared as a duplicate.

Running a custom function to find all such duplicates we see that there are 6 instances of duplicate pairs in our dataset.

1. DateTime_Intake → MonthYear_Intake
2. Animal Type_Intake → Animal Type_Outcome
3. Breed_Intake → Breed Outcome
4. Color_Intake → Color_Outcome
5. DateTime_Outcome → MothYear_Outcome
6. Name_Intake → Name_Outcome

Since as mentioned earlier duplicate columns hold no value we drop the duplicate columns. To maintain uniformity we drop the column which is associated with outcome.

2.2 Issue 1-> NULL VALUES

One good thing about the dataset is that most of the data is intact. Infact , the only attributes with major issues with missing data were ‘Name_Intake’ and ‘Name_outtake’ with around 35.2% of data missing. The only other attribute with missing data was ‘Age upon Outcome but that column only one instance had a missing value. In the case of ‘Name_Intake’ and ‘Name_Outcome’ both are duplicates of each other.

ACTION TO TAKE : Since these are names, imputation is not possible and since these data are just names the attribute holds no meaning to overall analysis, therefore, will be dropping the columns. The only other attribute that has a missing value is ‘Age upon outcome’. Since it is just one value we can impute it using an average value.

2.3 Issue 2-> DATA TYPE

Looking at fig .1. It's quite apparent all the attributes are of type object. An object type attribute can be considered like a string and it cannot be effectively used for exploratory analysis.

The attributes must be checked on an individual basis and understand what it represents before we can do any conversion. A general rule of thumb to use to find if an attribute should be of type ‘Categorical’ is:

- For the given attributes the cardinality of the attributes are much less than number of instances in dataset
- The observed data is in no way continuous
- Fields like 'binary_outcome' and all the above attributes have fixed categories
- No meaning full data can be extracted from mean of say 'intake condition' and other attributes hence another reason to consider categorical type

```
Animal ID              object
Name_Intake            object
DateTime_Intake        object
MonthYear_Intake       object
Found Location         object
Intake Type            object
Intake Condition       object
Animal Type_Intake     object
Sex upon Intake        object
Age upon Intake        object
Breed_Intake           object
Color_Intake           object
Name_Outcome           object
DateTime_Outcome       object
MonthYear_Outcome      object
Date of Birth          object
Animal Type_Outcome    object
Sex upon Outcome       object
Age upon Outcome       object
Breed_Outcome          object
Color_Outcome          object
binary_outcome         float64
dtype: object
```

fig .1

. With those assumptions we convert the following to categorical type:

- Intake Type
- Intake Condition
- Animal Type_intake
- Sex upon outcome
- binary_outcome

ACTION TO TAKE : Convert the attributes satisfying above conditions to CATEGORICAL TYPE

2.3 .1 Issue 3 -> DATA TYPE AND TRANSFORMATIONS

Some of the attributes in order for them to be converted to the proper data types had to undergo some form of pre-processing . Attributes such as:

- Date Time
- Found Location
- Age upon Intake
- Date Time_Intake
- Age upon Outcome

Cannot be used directly for any form of analysis. Actions to take will be included in each explanation below.

- Date Time_Intake

  The attribute 'Date Time_Intake' is first processed to extract the AM/PM information to know at what time the animals are brought to the shelter. This is stored as another feature 'intake_am_pm' and the attribute is converted to a categorical type. The information can be used to find the peak intake period in a day .Since we don't need exact time at which animal is brought to shelter (because this adds too much granularity to data) we drop the 'Date Time_intake' attribute after this.

- Found Location

  The next feature which requires processing is the attribute ' Found Location'. The attribute initially has a granularity of 759 which is very high. The reason for such high cardinality is because the data is taken up to a specific street in which an animal was found. What I did was to remove the streetwise information and just take the city from which the animal was captured. This brings down the cardinality to 16 and we can easily visualize the data to find from where the most animals were captured.The resulting attributes where converted to type categorical as it satisfied the conditions described earlier for considering an attribute to be of type 'categorical'. I was not able to find an alternative which can give the same clarity in result.

- Age upon Intake and Age upon Outcome

  Both these features expressed age. However they used different units. Some animals age in years , some others in weeks and months. So we need to convert them all to a single unit. The options were to either convert all to year, month or weeks. Looking through the data we found out that. If we use years there are some values which will be converted to decimal as there are animals less than 1 year of age. Plotting decimal points is bad for visualization. Hence the best compromise

was to convert all units of month with animals less than 1 month bucketed to 1 month category. There were also cases where information was written in plural form for example year and years , month and months. This needed to be corrected aswell.This attribute was finally converted to continuous type

- Color_Intake

  Looking at the color intake attribute we saw a really high cardinality of 120 inspecting the data we found out that the data was ambiguous. There was no way of automatically sorting colors to buckets as there were animals with mixed colors like is black/white different from white/black. Another observation was that there were ambiguous expressions of colours like ‘Flame Point’ or ‘Blue tick’ which cannot be interpreted. Some data represent the same color with different names. For example black is also written as sabel in some data . Due to such high ambiguity we decided to not use the table for predictive analysis.

- Breed_Intake

  The attribute has an issue. It represents both animal type and species all in single column.Example there are both different species of dogs and at same time different types of animals like bats. This makes it difficult to group data. Another observation is that there are many instances of animals which only appear once. So such animals do not provide much information. So one thing I did was to find most common animals which appear at least 10 times and rest others are bucketed into a single category called others. And finally convert all of them to type attribute. When we plot the data we see that the type others are the biggest bar among all other animals. One way to interpret such results is that there are many varieties of animals captured and brought to shelter and it may not be common.

The process of transformation takes out many of the problems in our dataset and it made it considerably easier to convert attributes to categorical or continuous type.

Found Location	category
Intake Type	category
Intake Condition	category
Animal Type_Intake	category
Sex upon Intake	category
Age upon Intake	int64
Breed_Intake	category
Sex upon Outcome	category
Age upon Outcome	int64
binary_outcome	category
index	int64
intake_am_pm	category
intake_time	datetime64[ns]

2.4 Issue 4-> LOW CARDINALITY COLUMNS

Another test we must do is to check if there are attributes with cardinality of just 1. Such attributes provide no value to data analysis and will be dropped. Looking through the data we see that there are no such features hence we go with the next test.

2.5 Issue 5-> NAMES AND INDICES

Columns which act like indexes or names do no not provide any valuable information to the analysis hence such features will be dropped. Looking through data we have two such attributes. Animal ID and Name_Intake . The attribute Name_Intake was already dropped in the previous section (section 2.2). And Animal ID will be dropped now.

3. LOGICAL INTEGRITY TEST FOR CATEGORICAL AND CONTINUOUS DATA TYPE

The logical integrity test must be carried out to see if there is any data which is not logically correct. Sifting through the data only one attribute seems to have a logical issue which is 'DateTime_Outcome' this attribute defines the time at which an animal is taken out of the shelter. It is observed that in some instances the date of animal taken out of shelter is less than the date at which animal is brought to the shelter. This cannot happen and is a logical fallacy. In such cases the date is imputed to have the same date as that of Intake. The reason being most of the animals are taken out of shelter in the same month as they are brought to the shelter. Hence by replacing the wrong data with intake date is the same as imputing with statistical mode.

3. Skewness Of Data

Most categorical data from each attribute seems to be skewed. That is it is not uniformly distributed. Therefore while developing a model we might have considered some form of normalization schemes.

4. Appendix

4.1 Terminologies and Assumptions

- Animal ID - The unique ID given to each animal that is brought to shelter
- Name Intake - Name of Domestic animals brought to shelter
- Date Time Intake - Time at which animals brought to shelter
- Month Year Intake - Time and Date at which animal was brought to shelter
- Found Location - Location from which animal was captured
- Intake Type - How animal was brought to shelter
- Intake Condition - The physical state in which animal was brought to shelter
- Animal type intake - The type of animal and species brought to shelter
- Sex upon Intake - The sex of animal determined on intake
- Age upon Intake - Age of animal on intake
- Breed Intake - Type of breed of the animal brought to shelter
- Color Intake - Color of animal when brought to shelter
- Name outcome - Name when animal taken out of shelter
- Date time outcome - Date and time of animal taken out of shelter
- Month year outcome - Month and year in which animal was taken out
- Date of Birth - Date of birth of animal
- Animal Type outcome - Animal species as it is taken out of shelter
- Sex upon outcome - Animal sex when animal was taken out of shelter
- Breed outcome - Animal species when taken out of shelter
- Binary outcome - The final outcome of animal ; 1 - Negative outcome , 0 - Positive outcome

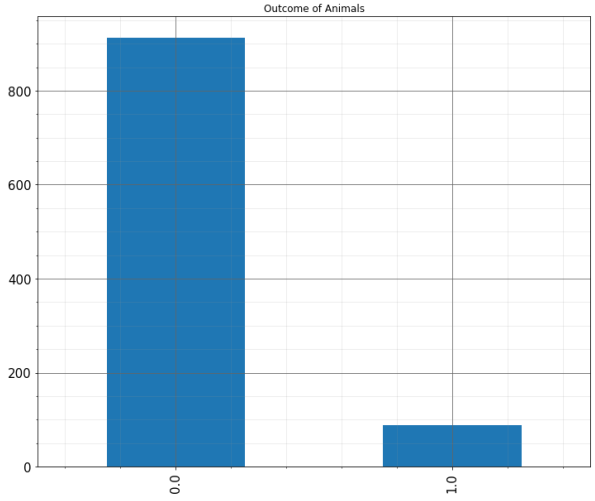
5. Descriptive statistics - Graph Appendix

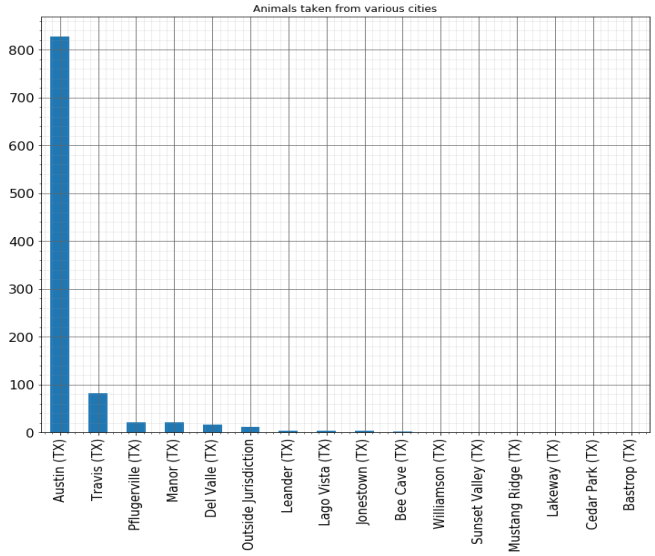
CATEGORICAL TYPE

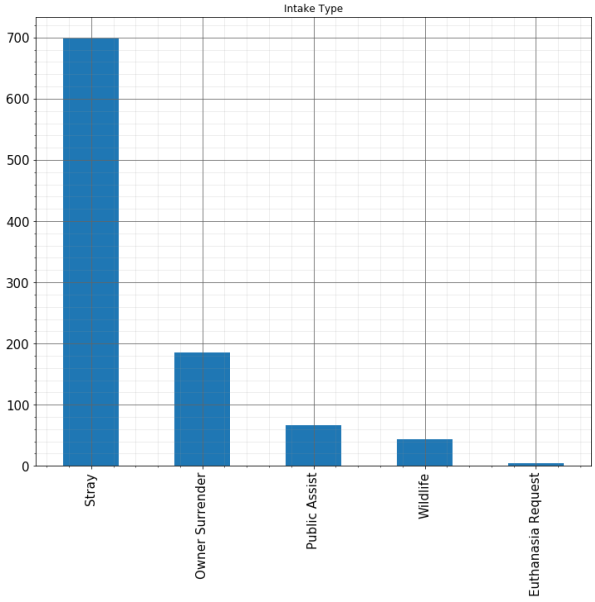
	count	unique	top	freq
Found Location	1000	16	Austin (TX)	828
Intake Type	1000	5	Stray	699
Intake Condition	1000	8	Normal	853
Animal Type_Intake	1000	4	Dog	556
Sex upon Intake	1000	5	Intact Male	329
Breed_Intake	1000	14	others	461
Sex upon Outcome	1000	5	Neutered Male	318
binary_outcome	1000	2	0	912
intake_am_pm	1000	2	PM	717

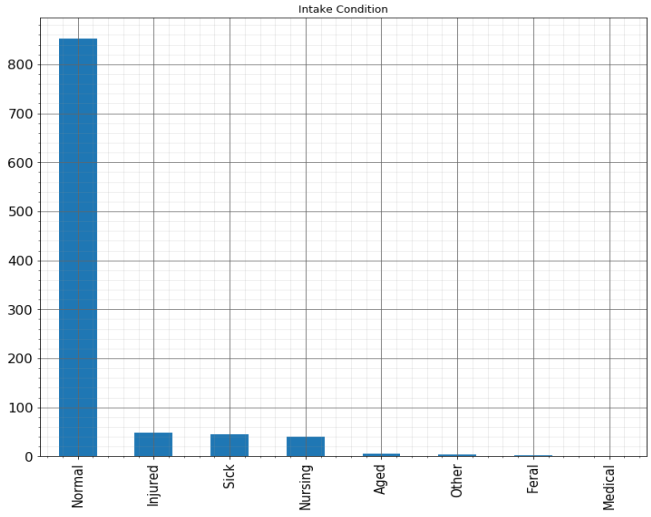
CONTINUOUS TYPE

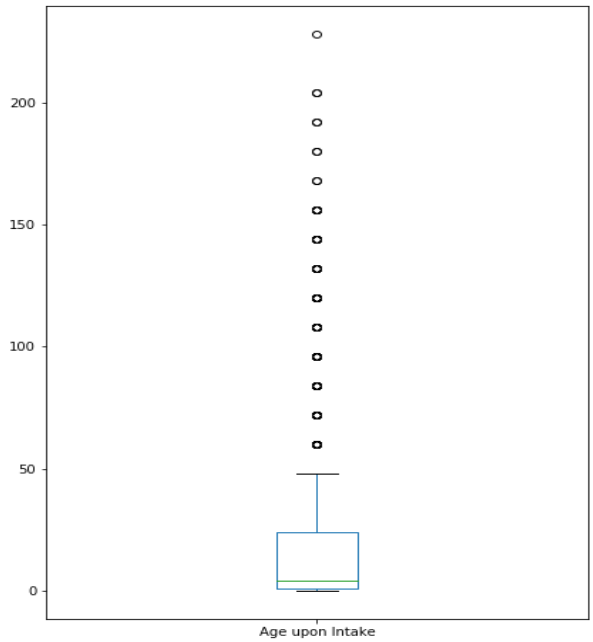
	count	mean	std	min	25%	50%	75%	max
Age upon Intake	1000.0	23.237	36.705833	0.0	1.00	4.0	24.00	228.0
Age upon Outcome	1000.0	23.657	36.778853	0.0	1.00	5.0	24.00	228.0
index	1000.0	499.500	288.819436	0.0	249.75	499.5	749.25	999.0

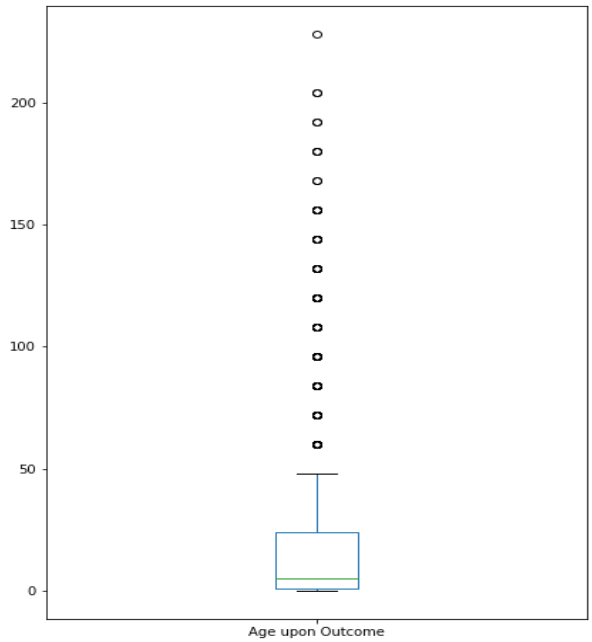

Outcome of Animals
800
600
400
200
0
0.0
1.0

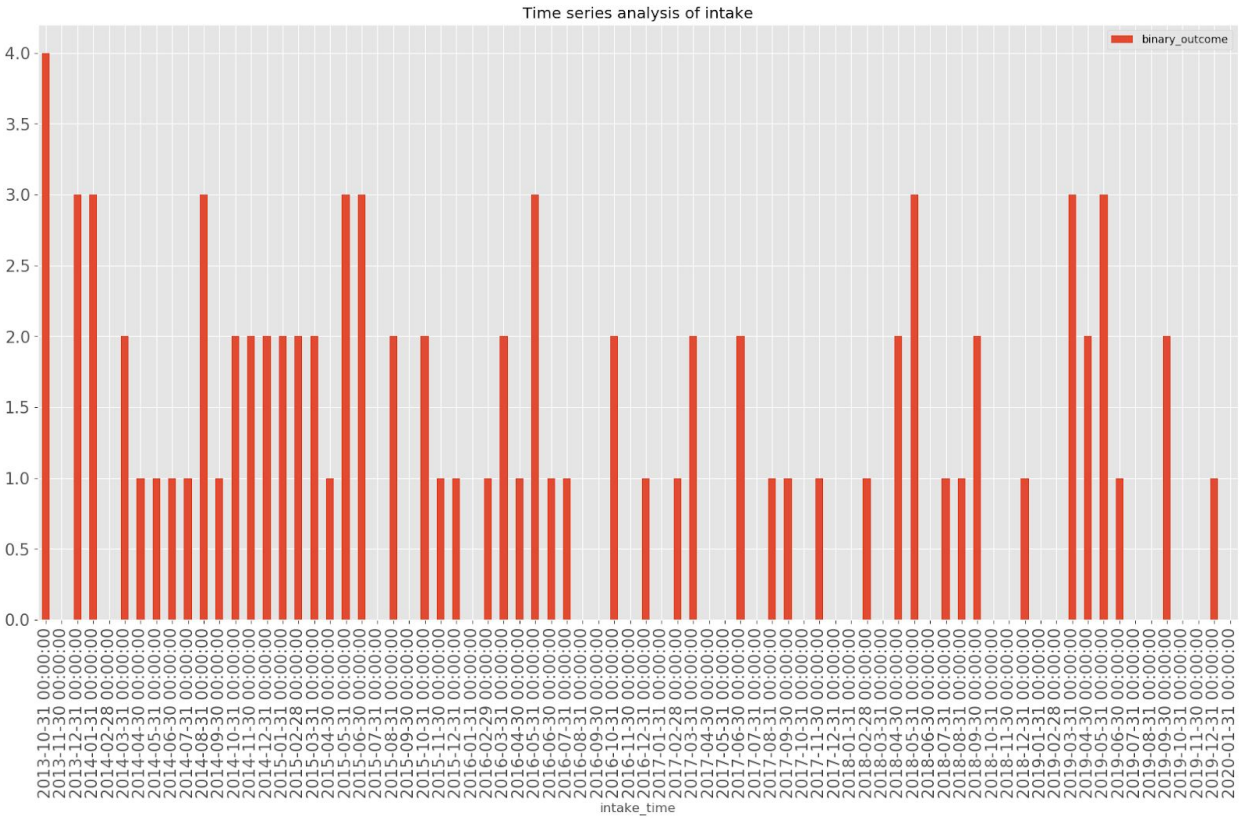

Animals taken from various cities
800
700
600
500
400
300
200
100
0
Austin (TX)
Travis (TX)
Pflugerville (TX)
Manor (TX)
Del Valle (TX)
Outside Jurisdiction
Leander (TX)
Lago Vista (TX)
Jonestown (TX)
Bee Cave (TX)
Williamson (TX)
Sunset Valley (TX)
Mustang Ridge (TX)
Lakeway (TX)
Cedar Park (TX)
Bastrop (TX)

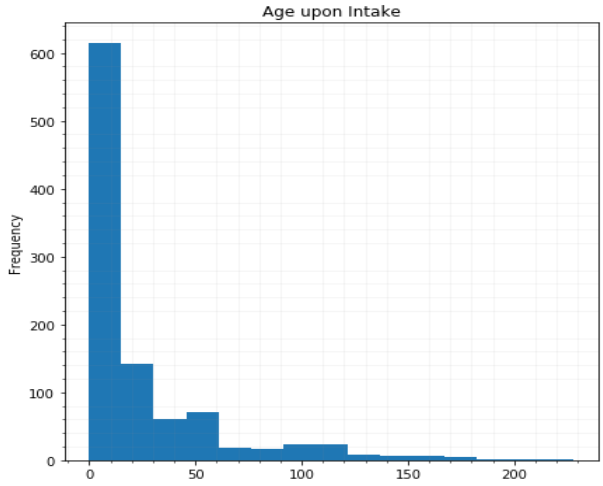

Intake Type
700
600
500
400
300
200
100
0
Stray
Owner Surrender
Public Assist
Wildlife
Euthanasia Request

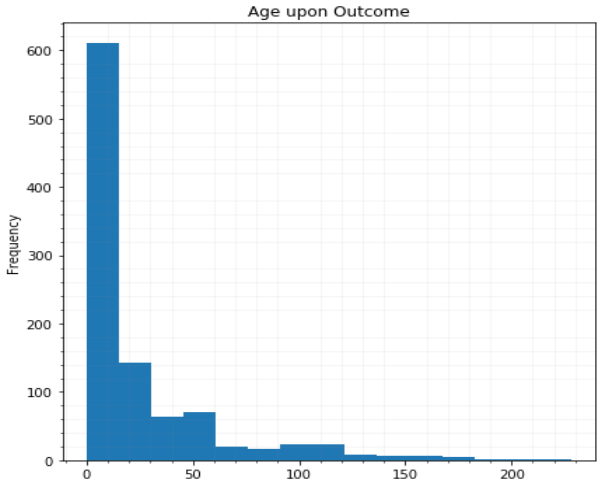

Intake Condition
800
700
600
500
400
300
200
100
0
Normal
Injured
Sick
Nursing
Aged
Other
Feral
Medical

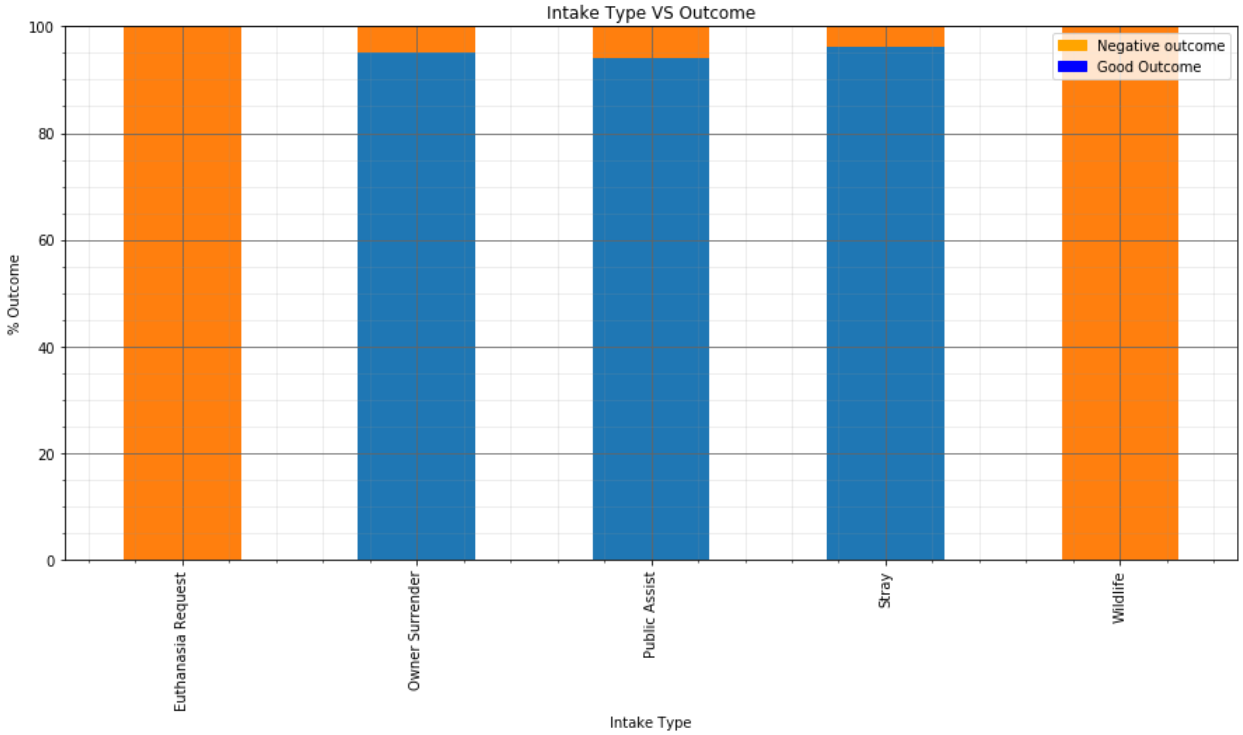

200
150
100
50
0
Age upon Intake

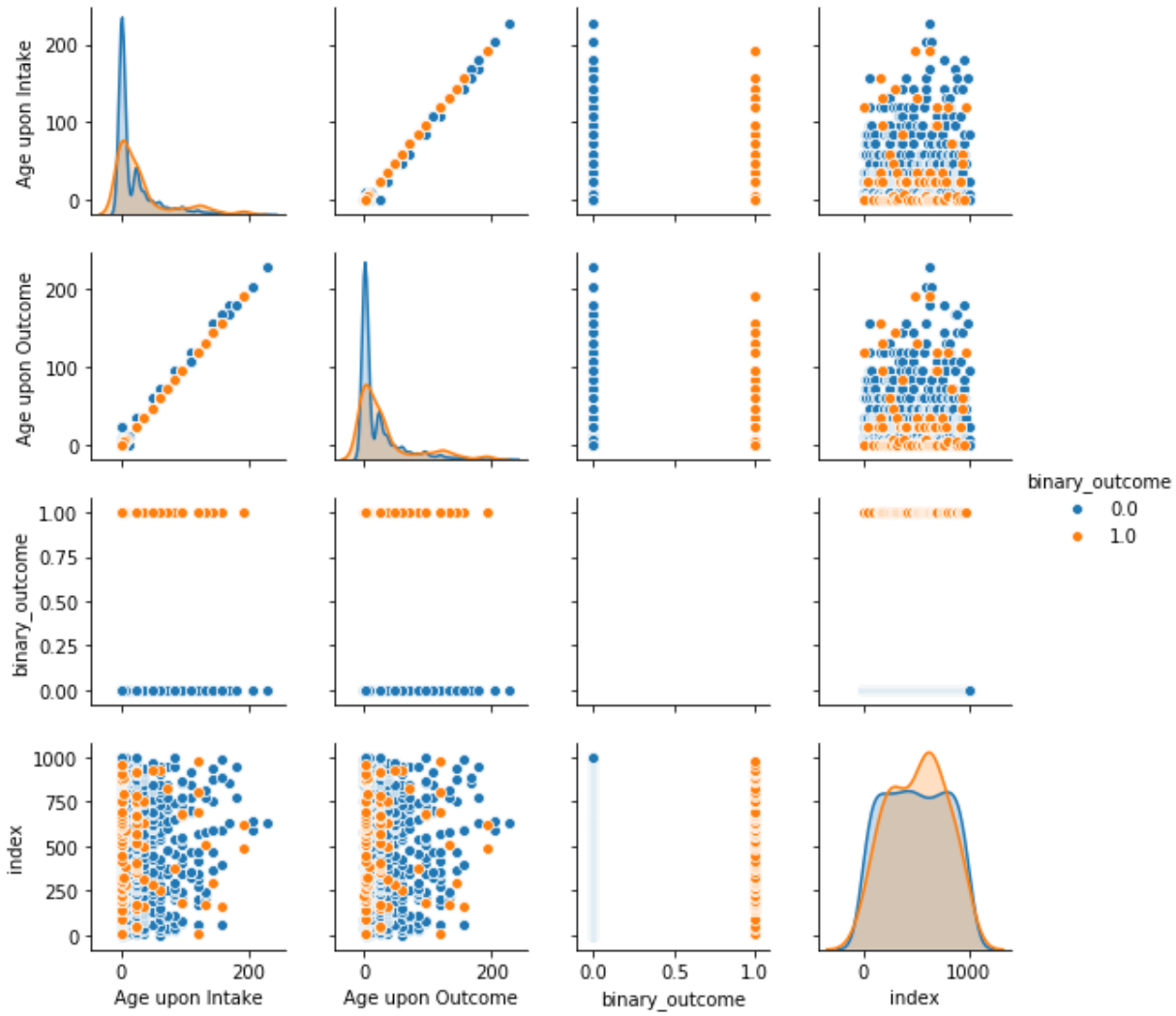

200
150
100
50
0
Age upon Outcome

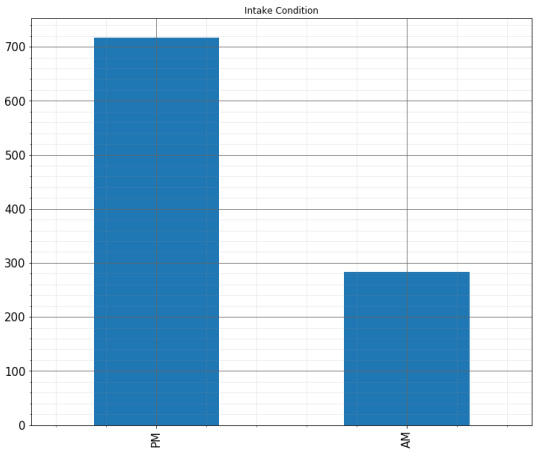

Time series analysis of intake
binary_outcome
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
intake_time

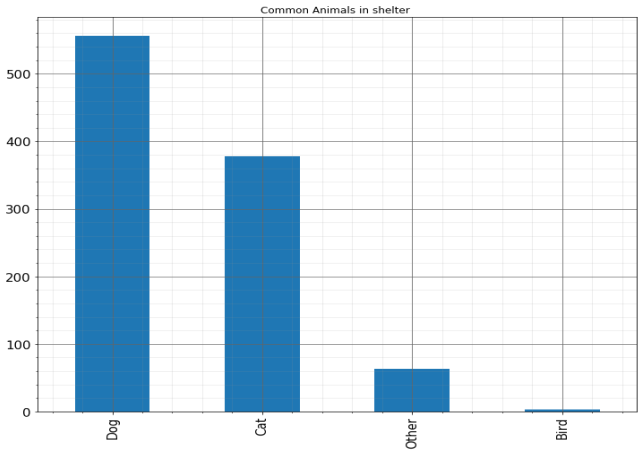

Age upon Intake
Frequency
0
100
200
300
400
500
600
0
50
100
150
200

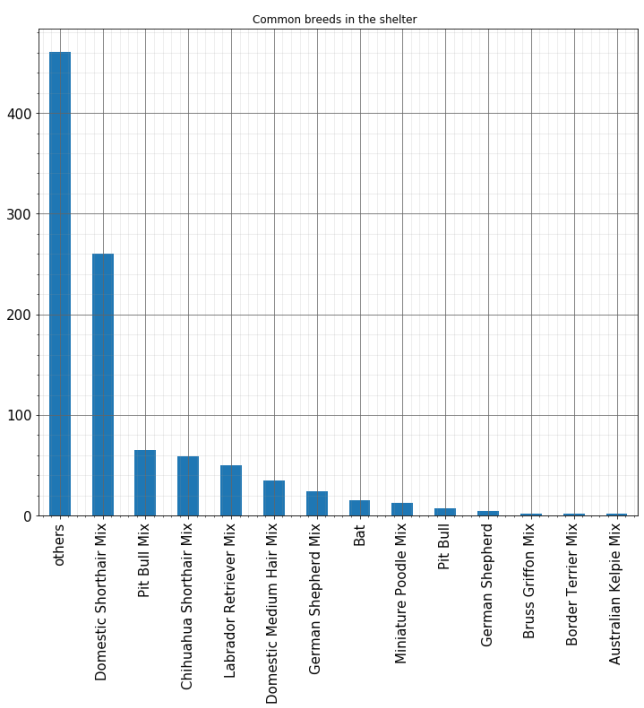

Age upon Outcome
Frequency
0
100
200
300
400
500
600
0
50
100
150
200

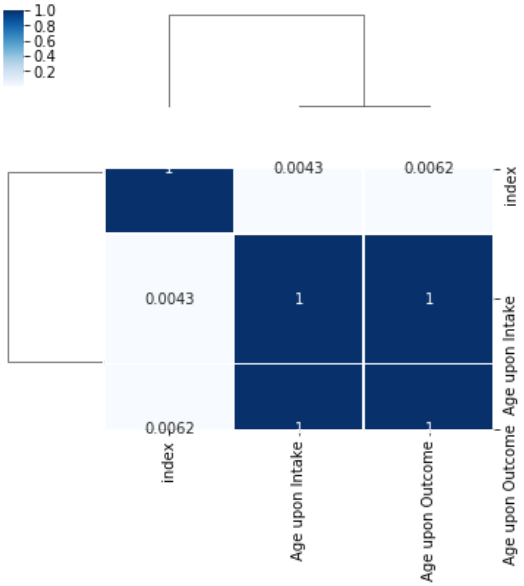

Intake Type VS Outcome
Negative outcome
Good Outcome
% Outcome
0
20
40
60
80
100
Euthanasia Request
Owner Surrender
Public Assist
Stray
Wildlife
Intake Type

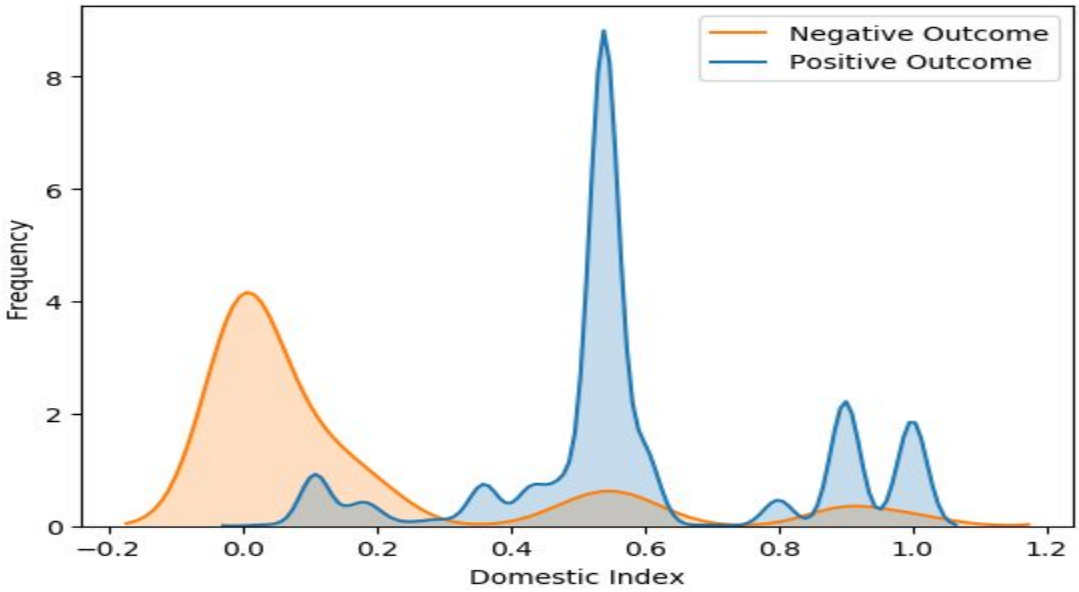

Age upon Intake
Age upon Outcome
binary_outcome
index
0.0
1.0


Intake Condition
PM
AM


Common Animals in shelter
Dog
Cat
Other
Bird


Common breeds in the shelter
400
300
200
100
0
others
Domestic Shorthair Mix
Pit Bull Mix
Chihuahua Shorthair Mix
Labrador Retriever Mix
Domestic Medium Hair Mix
German Shepherd Mix
Bat
Miniature Poodle Mix
Pit Bull
German Shepherd
Bruss Griffon Mix
Border Terrier Mix
Australian Kelpie Mix


1.0
0.8
0.6
0.4
0.2
0.0043
0.0062
index
0.0043
1
1
Age upon Intake
0.0062
index
Age upon Intake
Age upon Outcome
Age upon Outcome


Negative Outcome
Positive Outcome
Frequency
8
6
4
2
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Domestic Index