Data Quality Report – Initial Findings

1. Overview

Throughout this report I will outline the initial findings based on the cleaned dataset 'animalWelfare_19205514_Updated.csv'. I will summarise the data, describe any data quality issues found and discuss potential solutions to address them. Please refer to the appendix for summary tables of descriptive statistics. Section 2 will provide a summary of all findings and recommendations. In sections 3, 4 and 5 is a more detailed discussion of any issues found and possible solutions to address them.

This dataset currently contains no continuous features. There were initially six columns which were duplicates of other columns. These have been removed. One of the features in the dataset contains some missing values. Other issues include four features with high cardinalities. Furthermore, the features describing age of intake and age of outcome are currently not in a useful format. Finally, some logical integrity tests were performed which have highlighted some additional inconsistencies in the data.

2. Summary

Five tests were carried out to check the logical integrity of the data. These highlighted 16 instances of potentially irrational data. For instance, one of the tests showed that 7 animals stayed a negative amount of time in the shelter. This is clearly impossible and this issue must be addressed. Refer to section 3 for more details on these integrity tests and results.

After removing duplicate columns the dataset contains three datetime features. These describe the date the animal entered the shelter, the date the animal left the shelter and the date of birth of the animal. Some values within these features were involved in the failed logical integrity tests, as detailed in section 3 below. These failing instances should be investigated further.

Within the categorical features, the feature representing the name of the animal has a high cardinality and contains many missing values. The percentage of missing values is too low to immediately drop the feature. Instead this feature should be investigated and represented in a different way, which also addresses the issue of high cardinality. The age of the animal is currently represented in a format which combines a number and a word, for example "1 year". This is currently a categorical feature and although we can indeed see the distribution of the different ages, we cannot perform helpful statistical analysis on the ages. It is recommended that this feature be converted into a more useful numeric format.

Some categorical features have irregular cardinalities. For instance, the 'FoundLocation' feature contains 766 unique values out of 1000 total instances. I recommend that the values represented by this feature be split and that only the city be used as the value. Similarly, there are 211 different values represented by the 'Breed_Intake' feature. It is recommended that this feature is grouped into two values depending on whether the animal is of a mixed or pure breed. 'Color_intake ' also has a high cardinality of 115. It is recommended that this feature be grouped into 'light' and 'dark' colours.

Finally, there are two features which have a value 'Other'. It is recommended that these issues be investigated on a case by case basis to see if the data makes sense and if there is an alternative way to represent the affected data.

3. Review Logical Integrity

5 tests were carried out. The failures are below:

- Test 1: Check if the date of outcome is an earlier date than the date of intake for any of the animals. This would imply that the animal stayed a negative length of time in the shelter which is impossible.
    - 7 rows failed this test
- Test 3: Check if age of intake of any animal is zero. This is indeed plausible if an animal is born in the shelter. However, this will be tested to ensure there are not an unexpected number of animals aged zero as this may imply some invalid data. For any failed instances, further investigation should be performed into each instance to make sure the result makes sense.

    - 5 rows failed this test.
- Test 4: Check if age of outcome of any animal is zero. This would imply that an animal left the shelter on the day of their birth. It is unlikely that an animal would be adopted on the day of their birth. Further investigation into this instance should be performed to see if it makes sense.
    - 1 row failed this test
- Test 5: - Check if age of outcome is less than age of intake for any animal. This would imply that the animal decreased in age in the shelter which is impossible.
    - 7 rows failed this test.
    - The instances which failed this test are a subset of the instances that failed test 1.

4. Review datetime features

After removing duplicate columns there are 3 datetime features. These are 'DateTime_Intake', 'DateTime_O utcome' and 'DateofBirth'. A descriptive table was generated for these features which shows prior to any tho rough exploration that there are no strikingly anomalous dates entered. However, there is an issue with some values of these features as discussed in section 3 above. These should be investigated and fixed where appropr iate. If no suitable solution is found, these instances should be dropped.

5. Review Categorical features

5.1 Descriptive Statistics

All categorical features will be discussed below. Some features will be grouped together in the discussion.

The animal ID is not useful for any descriptive statistics or comparisons. However, it is necessary to relate back to a specific animal and so it will be kept as it is.

The 'Name_Intake' feature has 30.% missing values. It also has a very high cardinality of 604. One possibility is to drop this feature. However, the percentage of missing values is not significant enough to immediately drop the feature. Another possible solution for missing values would be to impute all values to the mode which is 'Max'. However, imputing 30.6% of values would change the central tendency of the feature too much. It is assumed that the missing values are a result of the animal not having a name, either as a stray or as a new-born in the animal shelter, however this should be investigated further in order to better understand the missing values. In order to address both the high cardinality and the percentage of missing values, it is recommended that this feature is changed so that it represents whether a name is provided or not.

'FoundLocation' has an extremely high cardinality of 766. On further inspection of the 20 most frequent values it was found that the information represented by the values is inconsistent. For example, the most frequent value is 'Austin' while the majority of the other most frequent values are specifying particular streets in Austin. As a result, it is recommended that the values in the feature are split into street and city and that the city is used to represent the found location. This solution will address both the high cardinality and the inconsistent representation of the locations.

The feature 'IntakeType' has five values, the most frequent of which is stray. These findings are as expected. 'IntakeCondition' has five values, the most frequent of which is 'Normal'. This is as expected. However, one of the values is 'Other'. This value is not very informative and should be investigated. Depending on the findings during this investigation, it may be best to impute this value to the mode. However, this will depend on how many instances are affected and whether the data makes sense.

'AnimalType_Intake' has four values. The most frequent is 'Dog' which is expected. One of the values is 'Other'. It is not clear what information this values represents. It is recommended that this value is investigated further and divided into more meaningful values if it makes sense to do so with the data.

There is nothing unexpected with the feature 'SexuponIntake'. Intact animals are more common than neutered/spayed animals. One of its five values is 'unknown'. This is not a very helpful value however, it makes sense given the variety of animals and breeds entering the shelter that the sex of all will not be known. It is

worth noting that this feature is representing both the gender and whether the animal is spayed/neutered in one feature. It is possible that this information would be better represented in two separate features. However, for now there is no issue with this feature and it will be left as it is.

In the features 'AgeuponIntake' and 'AgeuponOutcome' we can see that the most frequent age is 1 year. It is assumed that this is because the age of most animals is approximated by the staff at the animal shelter. This makes sense considering the environment. The features 'AgeuponIntake' and 'AgeuponOutcome' are currently not in a useful format for statistical analysis. It is recommended that both of these features are converted into a more useful format. One possibility is to represent the age in years. However, in order to retain as much accuracy as possible, it is recommended that the age features are converted to an age in days.

'Breed_Intake' has a high cardinality of 211. Upon further inspection of the top 20 most frequent values for this feature it was found that many of the values are overlapping. For instance, in the top 20 are both Domestic Shorthair Mix and Domestic Shorthair. Ideally I would like to speak to a domain expert to understand this distinction better. One possible solution would be to remove the word 'mix' in order to reduce the possibility of overlapping values. However, it order to not lose relevant data, for now it must be assumed that the distinction refers to slightly different breeds. It is recommended that the breed types are divided into 'Mixed Breed' and 'Pure Breed', based on whether the word 'Mix' is included in the breed name.

The 'Color_Intake' feature has 115 values which is a very high cardinality. There are two points to note upon analysis of the top 20 most frequent values for this feature. Firstly, it can appear that there are a lot of overlapping values. For instance, both 'Black/White' and 'White/Black' are in the top 20 most frequent. Ideally I would like to speak to a domain expert to learn more about whether these similar values are an inconsistency in data input. In absence of further information it is assumed that these different values represent different colourings. It is assumed that the 'Black/White' animal is majority black with some white, while the 'White/Black' animal is majority white with some black markings. The second point to note is that this feature is representing the pattern of the animal coat as well as the colour. Within the top 20 most frequent we can see coat patterns such as 'Tabby', 'Tortie' and 'Brindle'. One possible solution is to split this feature into two features, representing the colour and the pattern. However, we are dealing with a variety of animals each with a variety of coat patterns and so accurate distinction of coat patterns is not realistic without consulting a domain expert. Furthermore, this solution might not address the high cardinality. For now it is assumed that whether the animal is light or dark coloured will have an effect on the outcome. As a result, I recommend that all values are divided into two values – 'light' and 'dark'.

5.2 Bar plots

Refer to the accompanying pdfs to see all bar plots.

We can see that the 'AgeuponIntake' and 'AgeuponOutcome' bar plots have an exponential distribution. Although these bar plots don't represent a numeric distribution of the ages, we can see that the older ages such as 20 years are much less frequent. It is recommended that the larger age values are investigated further as potential outliers.

The high cardinality bar plots are not useful but they do highlight the issue of the high cardinality and give an idea of the distribution of the most frequent values within these features.

6. Actions to take

There are 8 main actions to take. These are listed below:

- Failed logical integrity tests:
    - Investigate all failed tests further. Replace or correct any invalid data where there is evidence that it makes sense to do so. Where no solution makes sense failed instances should be dropped.
- Name high cardinality and missing values:
    - Change feature so that it only represents whether a name was provided or not.
- Found location high cardinality:

    - Split into two features, found street and found city. Replace found location with found city.
- AnimalType_Intake ambiguous value
    - Explore the value 'Other' and divide it into more informative values if it makes sense to do so.
- Intake_Condition ambiguous value
    - Explore the value 'Other. Depending on the outcome of the investigation, possibly impute it to the mode.
- Age features format
    - Both 'AgeuponIntake' and 'AgeuponOutcome' should be converted to days.
- Breed_Intake high cardinality
    - Group values into 'Mixed Breed' and 'Pure Breed'.
- Color_Intake high cardinality
    - Group values into 'light' and 'dark'.

7. Appendix

7.1 Table of descriptive statistics for categorical features

	count	unique	top	freq	%missing
Name_Intake	694	603	Max	8	30.6
FoundLocation	1000	766	Austin (TX)	188	0.0
IntakeType	1000	5	Stray	708	0.0
IntakeCondition	1000	5	Normal	891	0.0
AnimalType_Intake	1000	4	Dog	563	0.0
SexuponIntake	1000	5	Intact Male	332	0.0
AgeuponIntake	1000	43	1 year	172	0.0
Breed_Intake	1000	211	Domestic Shorthair Mix	259	0.0
Color_Intake	1000	115	Black/White	106	0.0
SexuponOutcome	1000	5	Neutered Male	346	0.0
AgeuponOutcome	1000	42	1 year	170	0.0
binary_outcome	1000	2	0	920	0.0

7.2 Table of descriptive statistics for datetime features

	count	unique	top	freq	first	last
DateTime_Intake	1000	996	2018-05-20 15:04:00	2	2013-10-01 16:10:00	2020-01-26 12:05:00
DateTime_Outcome	1000	998	2018-05-20 18:32:00	2	2013-10-05 12:34:00	2020-02-03 17:45:00
DateofBirth	1000	872	2017-05-21 00:00:00	4	1997-06-11 00:00:00	2019-12-16 00:00:00

7.3 Bar plots

See the following accompanying pdfs for bar plots:

- AnimalWelfare_19205514_categorical_lowcardinality_barcharts
- categorical_high_cardinality_Color_Intake
- categorical_high_cardinality_NameIntake
- categorical_highcardinality_Breed_Intake
- categorical_highcardinality_FoundLocation