Data Quality Report – Initial Findings

Descriptive Stats for Continuous Features:

- We have a full count for all of the continuous features.
- The Age_upon_Intake and Age_upon_Outcome show minimum values of 0.0. This is something which should be investigated further in order to check the validity of the data. The maximum value of both of these columns is 19.0. It seems to capture a wide range of animal ages. Outliers should be checked to endure their validity. The mean of both columns is 1.0, i.e. 1 year.
- The minimum and maximum values for the year_intake and year_outcome columns show the range of years in question for this data. It ranges from 2013 to 2020. The cardinality is therefore correct at 8.
- The month_intake and month_outcome columns have a median value of 7.0 which shows that July is the busiest month for the shelter in terms of intakes and outcomes.
- The birth_year and birth_month columns show median values of 2015.0 and 6.0, respectively. This conveys that median birth year for animals is 2015 and median birth month of animals is July.
- There are 0 missing values for each of the columns and the cardinalities all seem to be correct and valid.

Descriptive Stats for Categorical Features:

- The Name_Intake column has a count of 634 out of 1000. It also has a very high cardinality. This count is something which should be investigated further.
- The Found_Location column also has a very high unique value count. 747 unique values out of a possible 1000. One of these values, Austin (TX), has a relatively high frequency at 196.
- The Breed_Intake column has 206 unique values. This is to be expected but options of grouping into larger breed groupings should be looked into.
- The same is the case for Color_Intake. Again, the possibility of wider color groupings should be investigated.

Histograms for Continuous Features (see continuous_histograms_1-1.pdf):

- The Age_upon_Intake and Age_upon_Outcome plots seem to show an exponential decrease.
- The Birth_Year plot seems to be exponentially increasing.

- The Birth_Month plot appears to be relatively normally distributed, with July at its peak.
- The Month_Intake and Month_Outcome plots appear to convey some correlation between eachother.

Box Plots for Continuous Features (see continuous_boxplots_1-1.pdf):

- Some of the boxplots have outliers. The outliers in the Age_upon_Intake and Age_upon_Outcome plots are to be expected as they reflect the range of animal ages from 0 to 19. Likewise, the birth_year boxplots also reflects this range.
- We can clearly see from the boxplots that the year_intake field is concentrated in the years 2015 to 2018. Thus, indicating that these were the busiest years for the shelter in this dataset.
- It can also be deduced from the month_intake and month_outcome plots that the busiest months seem to be in the range May to October.

Bar Plots for Continuous Features (see categorical_barcharts_1-1.pdf):

- It is very difficult to deduce any useful information from name_intake plot at this point. It is clear that this data will need to be investigated further for possible groupings.
- Likewise, the Found_Location data is too crowded and noisy at this point. There does seem to be some common locations represented at the left-most of the plot.
- The Intake_Type shows, quite predictably, that the majority of all animals taken in by the shelter are classed as strays.
- We can also see from the Intake_Condition plot that the overwhelming majority of animals are taken in in what is classed as a ‘normal’ condition.
- The Animal_Type_Intake plot displays the types that the animals are broken into with dogs being the most populous, followed by cats and then a very significant drop to other animals.
- The Sex_upon_Intake and Sex_upon_Outcome plots are interesting as they show a difference in numbers for each bar for intake and outcome. This suggests that neutering/spaying is performed while the animals are at the shelter. It will be interesting to see the correlation between the ‘unknown’ value and the animal’s outcome.
- The Breed_Intake plot shows an exponential decrease from left to right. However, the volume of values is just too broad. Possible groupings will need to be investigated to make this data useful.
- The Color_Intake plot shows a similar exponential decrease. Likewise, there is a very large range of values. Possible wider groupings should be explored.

- The Binary_Outcome plot immediately conveys that approximately 10% of all animals that have been in the shelter have had a negative outcome.