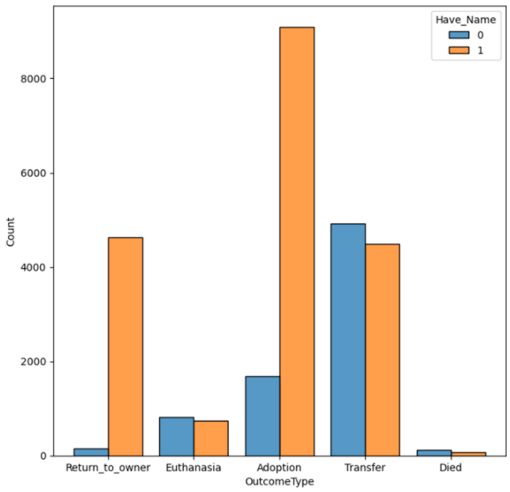
在資料處理的部份分為七個部分，分別是缺值的填補以及針對各特徵的處理，以下詳細說明：

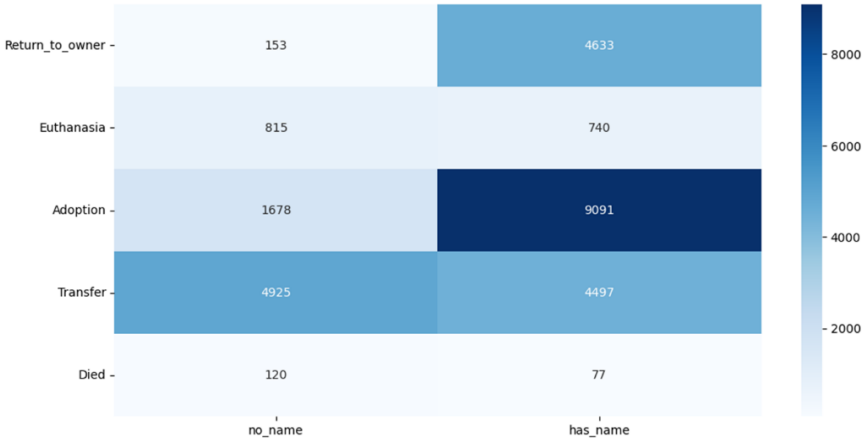
一、missing value imputation

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26729 entries, 0 to 26728
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Name            19038 non-null  object
 1   OutcomeType     26729 non-null  object
 2   AnimalType      26729 non-null  object
 3   SexuponOutcome  26728 non-null  object
 4   AgeuponOutcome  26711 non-null  object
 5   Breed           26729 non-null  object
 6   Color           26729 non-null  object
dtypes: object(7)
memory usage: 1.4+ MB
```

觀察資料缺值的狀況，發現姓名、性別及年齡有缺漏的情形。姓名的部分，由於後面會將此特徵區分為有無姓名，因此將沒有姓名的直接填補 0，歸類為沒有姓名的動物。性別的部分，只有一個缺值，因此直接將其歸類於最多的那個性別類別 Neutered Male。年齡的缺值以平均值填補。

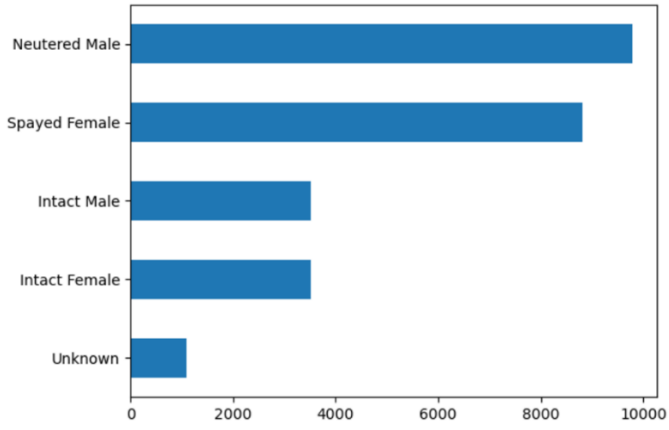
二、Name

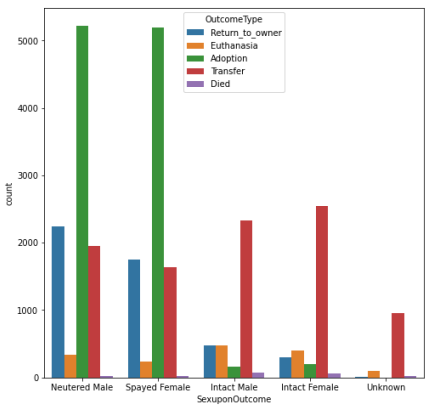


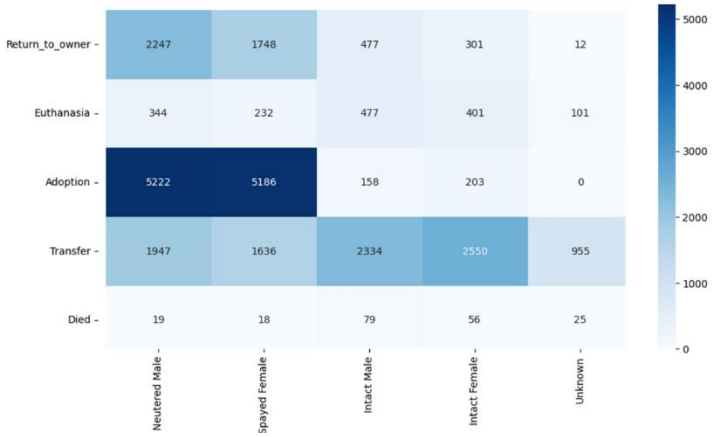


Name 為 str 型態特徵，內容為狗貓的名字，像是 Elsa、Jimmy…等。觀察視覺化圖表，在結果為 adoption 的資料中，多數都是有名字的。以 Cramer's V 係數分析，發現相關性達到 0.453，為強相關。由於將 Name 直接轉 one-hot 的話，資料維度可能會因為名字種類過多而變太高看不出特性，因此將此特徵轉換為有無姓名(0/1)。

三、Sexuponoutcome

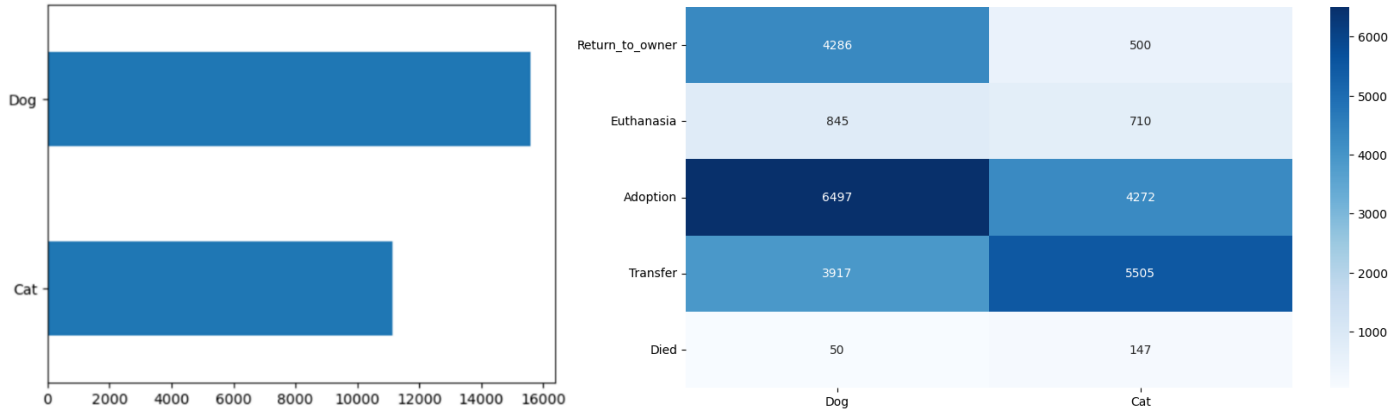






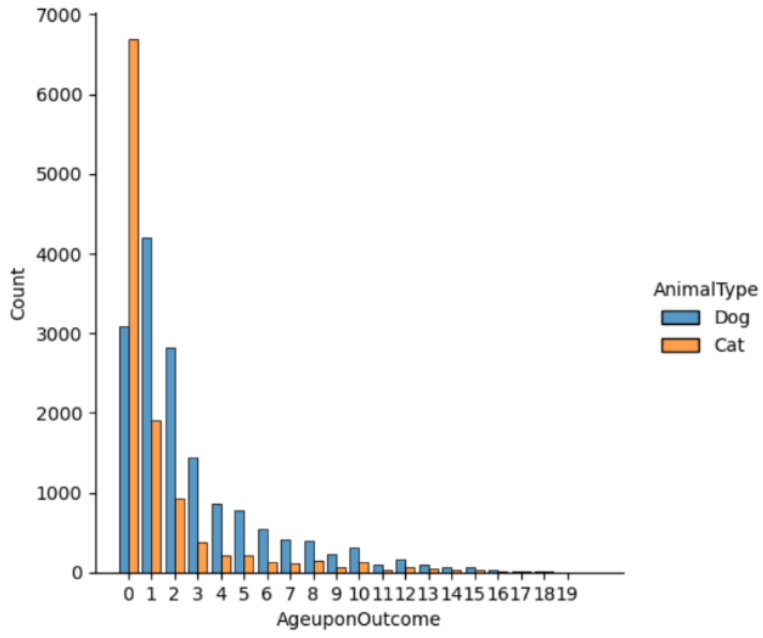
性別的部分分為五類。資料多數落在 spayed female 和 neutered male。在 Cramer's V 係數上性別與結果的相關係數一樣達到 0.453，為強相關。此特徵無特別處理直接 one hot encoding

四、Animaltype



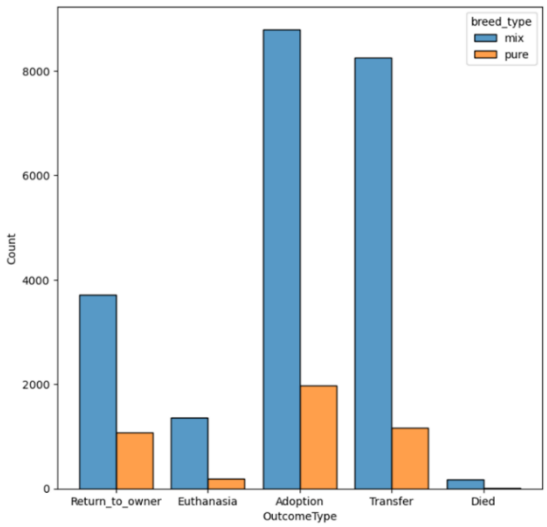
此特徵有兩類，在 Cramer's V 係數上的結果為 0.341，為中相關。嘗試兩方法：1.將貓跟狗的資料切開分別訓練模型、2.不區分兩模型，直接轉 one-hot 一起訓練。兩方法皆嘗試過後，發現一起訓練表現稍微好一點，但差別很小。

五、Age



Age 的部分，資料為數值加年、月、週等時間單位的 str 型態。嘗試兩個方式，一是將資料全部轉換為以天為單位，另一個方式是將年紀切分為四個階段:' NewBorn',' Adolescence',' Adulthood',' Senior'。經過實驗發現將年齡皆轉換為以天為單位的效果較好。

六、Breed

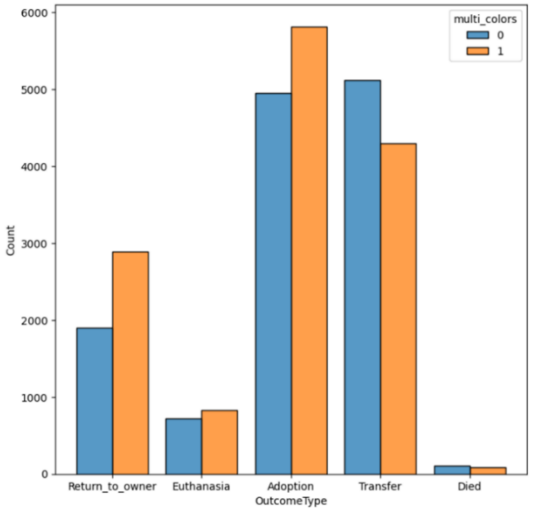


血統的部分，資料的型態為 str 型態，將貓和狗的血統區分的相當詳細。若有混種，字串中會有'Mix'的單詞出現，如：Domestic Shorthair Mix，或是會列出不同的品種，並以斜線區隔開來，如：Plott Hound/Boxer；純種的部分則是會直接列出其品種。

經觀察發現其中種類有 1380 種，若是直接轉 one-hot 很有可能造成維度過高的狀況。因此決定實驗三種方式：

一、將貓跟狗分開處理，狗的部分以 knowledge-based 的方式，根據 American Kennel Club (AKC)對狗的分類，將狗的種類進行歸類，類別有像是 Working、Toy、Mix……等類別；而貓則是將區分為長、中、短毛和其他。經過處理，此特徵的維度降為 57 維。第二種方式，貓的作法沒有改變，狗的部分是將出現超過平均次數的血統種類留下，其他較少出現的血統種類歸類為其他，經過處理後的維度為 13 維。最後一種方式貓和狗是一起處理的，將貓狗依據字串中的'Mix'和斜線將貓和狗分為'Mix'以及'Pure'。經過實驗，發現將貓和狗的血統區分為'Mix'以及'Pure'的效果是最好的。

七、Color



最後一個特徵是毛色，資料的型態也是 str 型態，內容舉例：Red/White 代表是有兩種顏色，Black 則是純色的。觀察資料發現顏色種類有 366 種，若直接轉 one-hot，也可能會有維度過高的情形。在參考了 kaggle 平台上其他隊伍先前的做法後，將毛色區分為多種顏色及純色(1/0)。