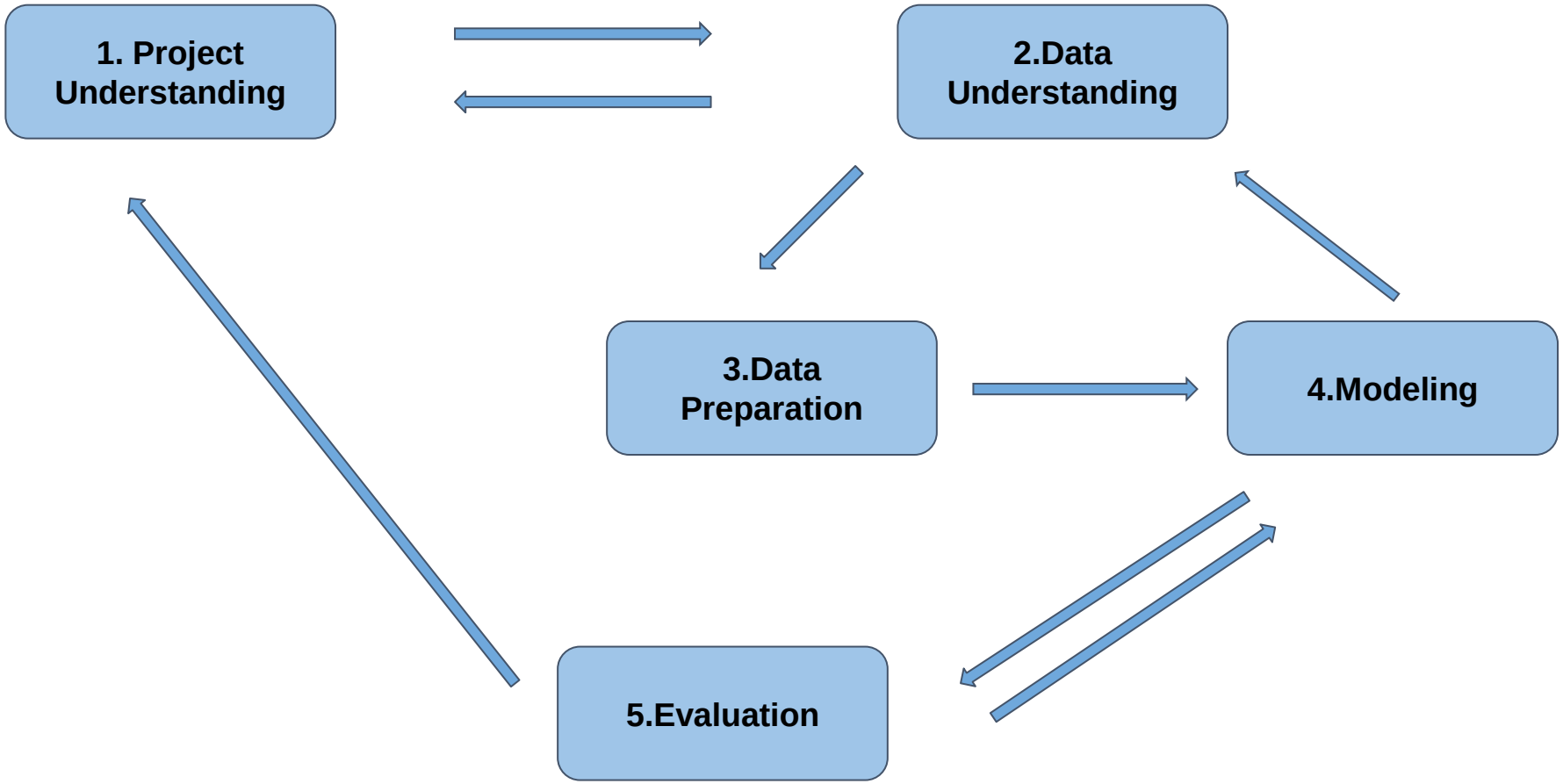
Freie Universität Berlin

ANIMAL SHELTER

Projektpräsentation

Business Intelligence – 22.07.2022

GLIEDERUNG
CRISP-DM MODEL



1. CRISP-DM: Project Understanding

Tiere kommen ins Tierheim

Tiere sind im Tierheim

Tiere werden vermittelt, euthanasiert …

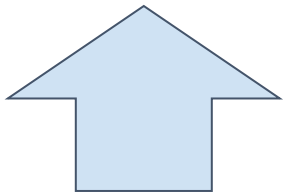
Unsere Aufgabe: Vorhersage

Evtl. weitere Handlungsempfehlung

2. CRISP-DM: Data Understanding (+ Data Quality)

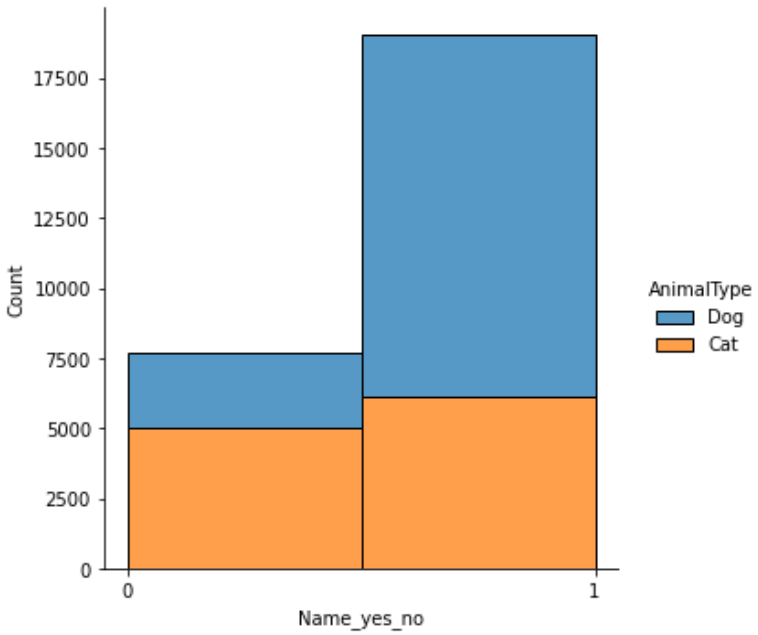
- train.head()

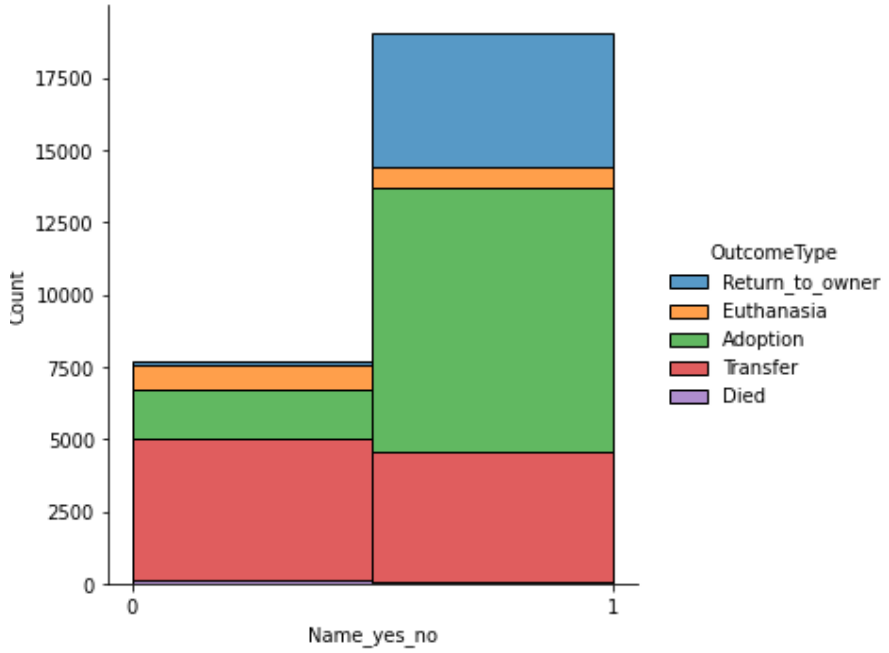
	AnimalID	Name	DateTime	OutcomeType	OutcomeSubtype	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
0	A671945	Hambone	2014-02-12 18:22:00	Return_to_owner	NaN	Dog	Neutered Male	1 year	Shetland Sheepdog Mix	Brown/White
1	A656520	Emily	2013-10-13 12:44:00	Euthanasia	Suffering	Cat	Spayed Female	1 year	Domestic Shorthair Mix	Cream Tabby
2	A686464	Pearce	2015-01-31 12:28:00	Adoption	Foster	Dog	Neutered Male	2 years	Pit Bull Mix	Blue/White
3	A683430	NaN	2014-07-11 19:09:00	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Cream
4	A667013	NaN	2013-11-15 12:52:00	Transfer	Partner	Dog	Neutered Male	2 years	Lhasa Apso/Miniature Poodle	Tan



2. CRISP-DM: Data Understanding - Visualisation

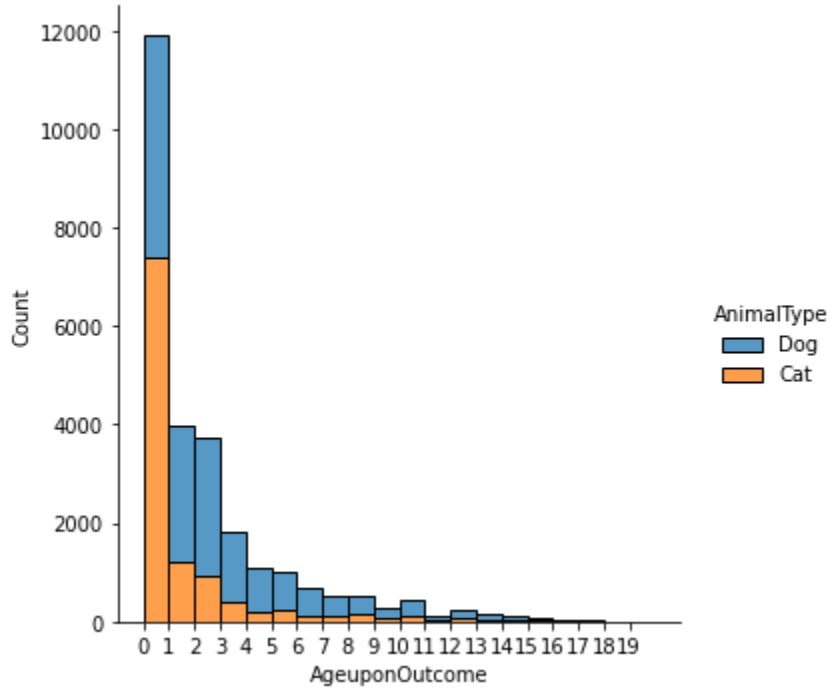
- Wie viele Tiere mit Namen gibt es?
- Hat der Name einen Einfluss auf adoption … usw?



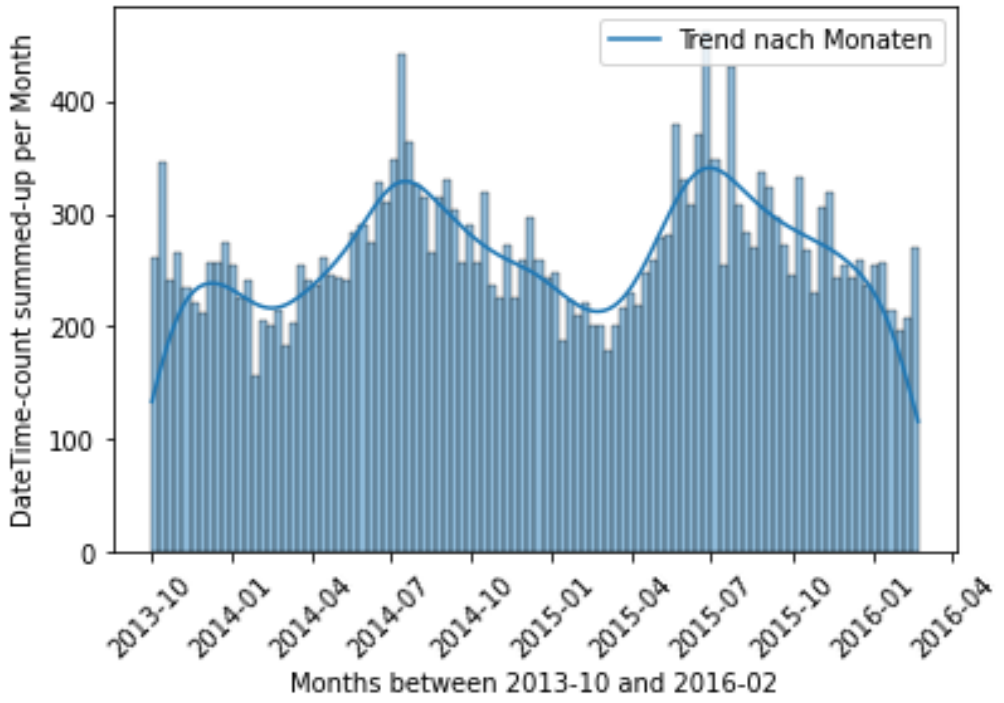


2. CRISP-DM: Data Understanding - Visualisation

- Wie lange bleiben die Tiere im Tierheim?

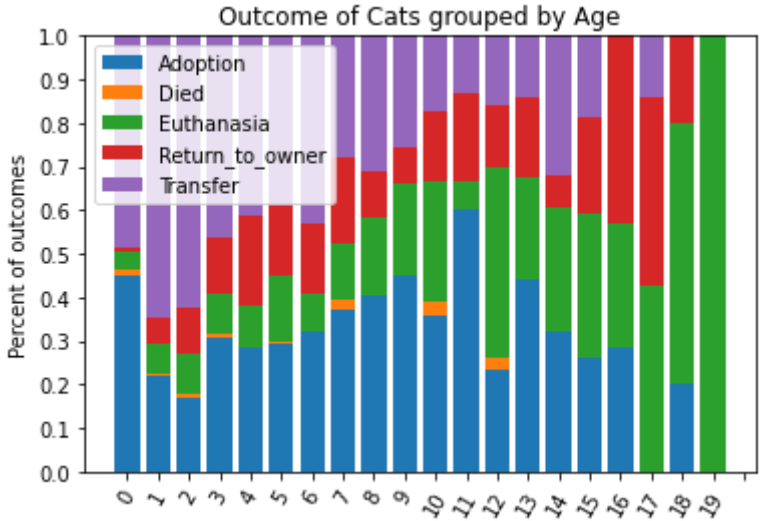


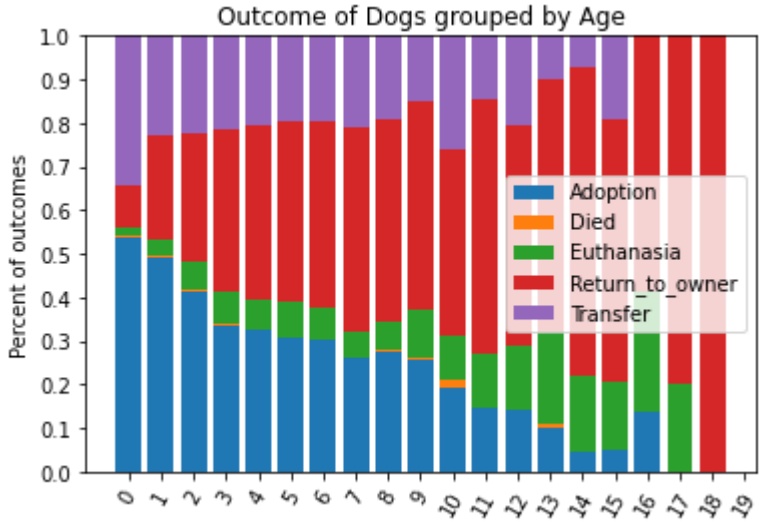
- Werden Tiere nach Jahreszeiten Adoptiert, Euthanasiert...?



2. CRISP-DM: Data Understanding - Visualisation

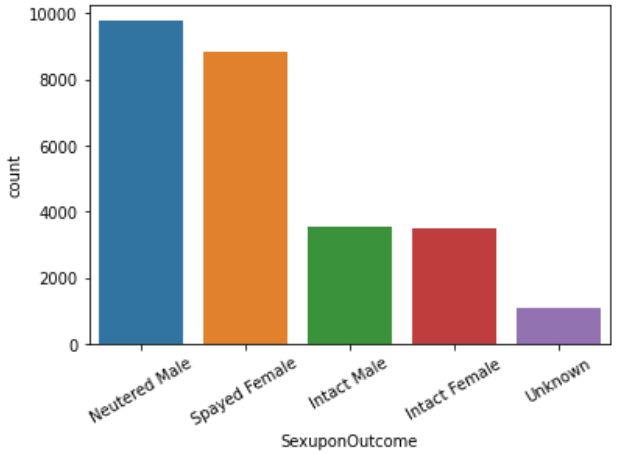
- Spielt das Alter (in Jahren) bei den Tieren eine Rolle ob sie z.B. adoptiert, euthanasiert usw. werden?

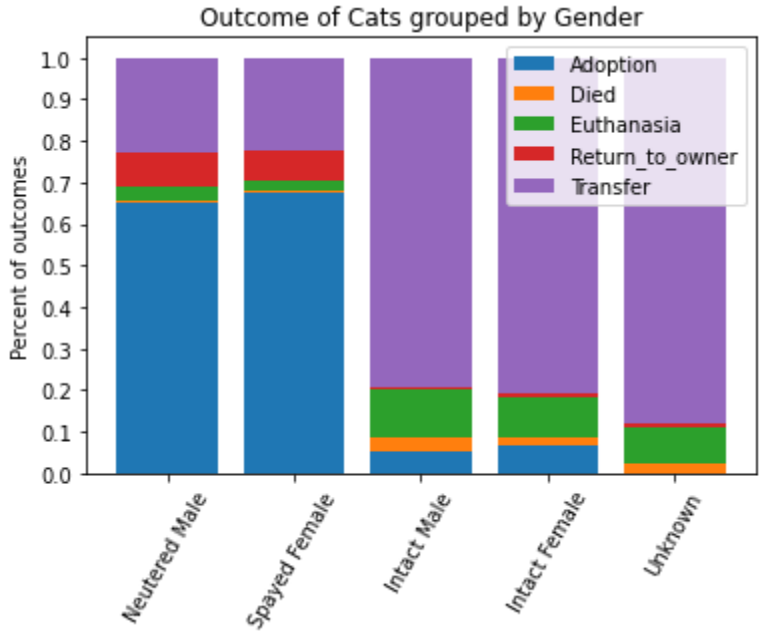


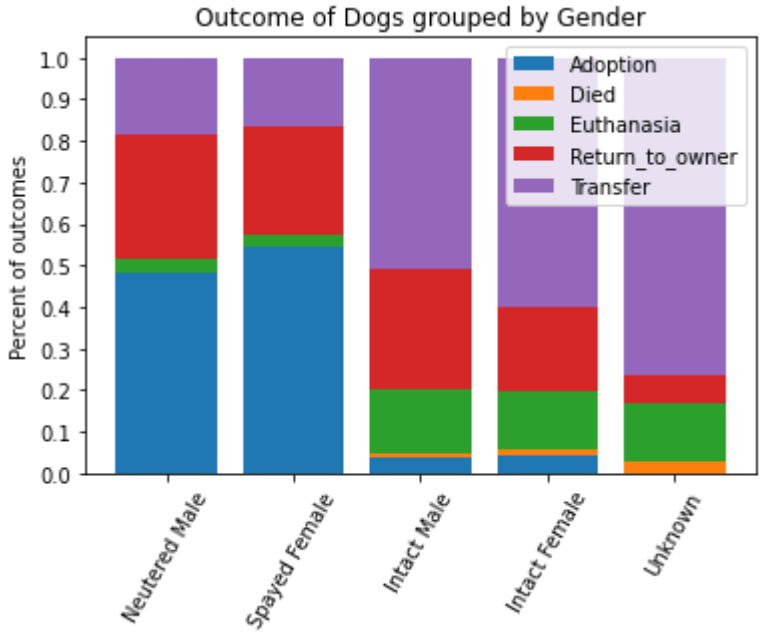


2. CRISP-DM: Data Understanding - Visualisation

- Welche “Geschlechter” sind im Datensatz und wie viele davon?
- Spielt das Geschlecht (neutered Male, usw) eine Rolle wenn es um das Schicksal der Tiere geht?

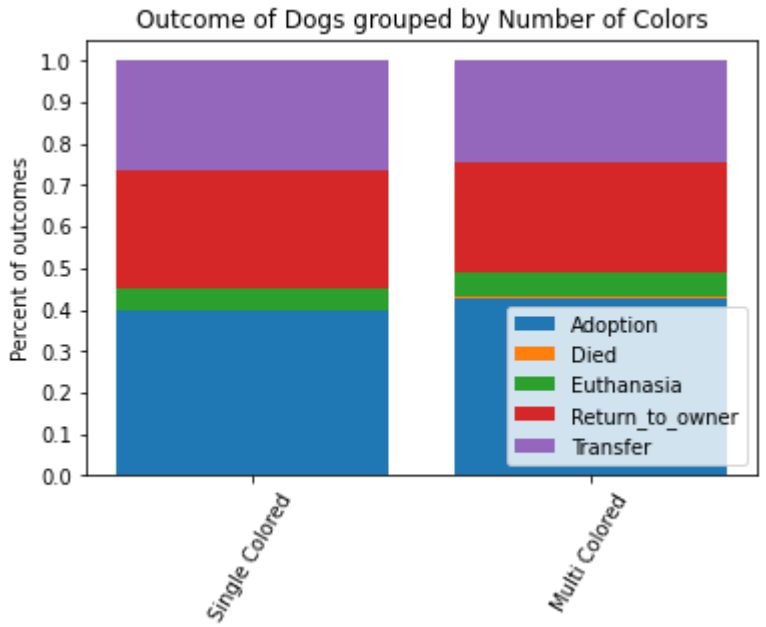


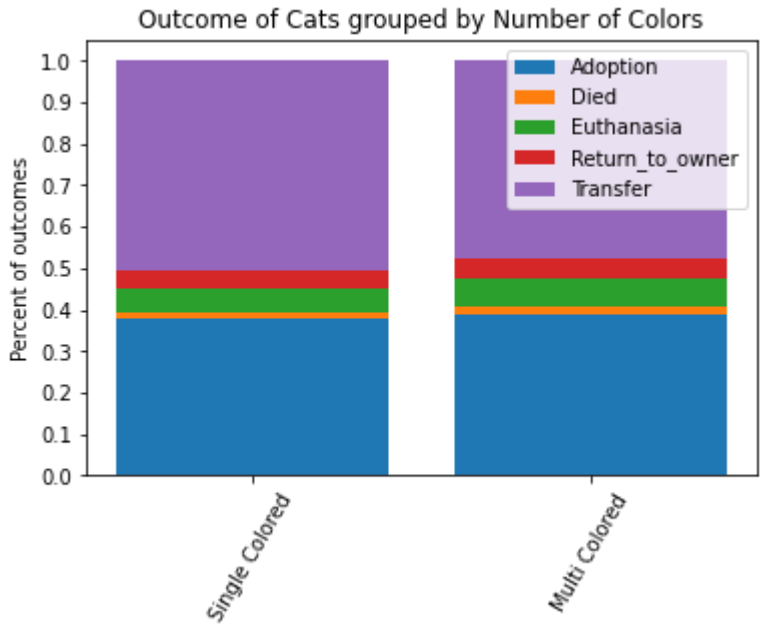




2. CRISP-DM: Data Understanding - Visualisation

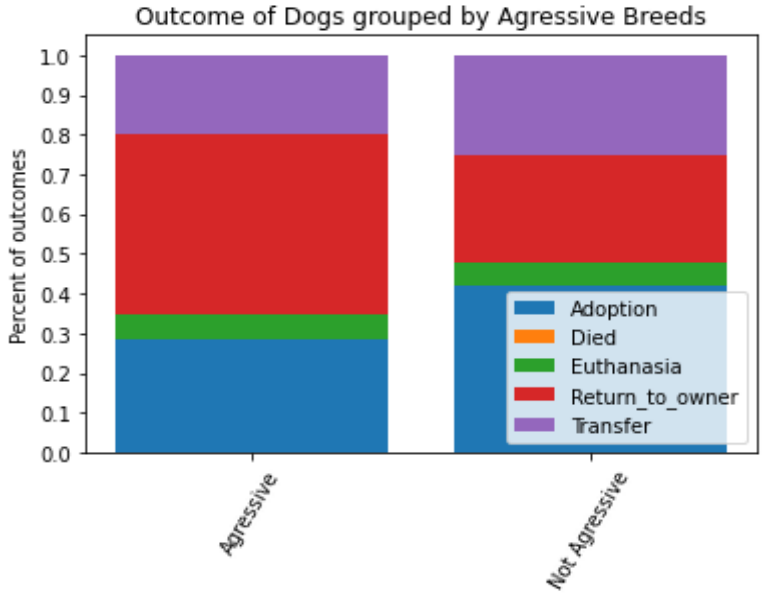
- Wie umgehen mit Color -> Single Color vs. Multi Color.
- Hat die Kategorie Single-Color, Multi-Color einen Einfluss auf die Schicksaale der Tiere?

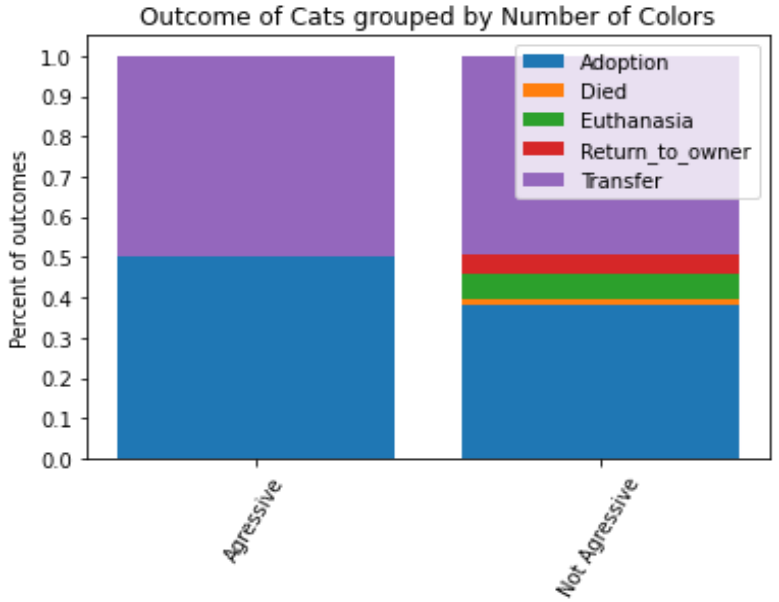




2. CRISP-DM: Data Understanding - Visualisation

- Werden einzelne Rassen, welche als aggressiv/nicht aggressiv gelten, seltener öfter adoptiert?





3. CRISP-DM: Data Preparation

Data Selection

Attribut	Selektion	Kommentar
Animal ID	nein	Unrelevant
Name	ja	Katzen mit Namen werden eher adoptiert als Tiere ohne Namen
Date Time	ja	Im Sommer werden eher Tiere adoptiert, als im Winter; Outcomes by day of week → Wochenende werden mehr adoptiert?
Outcome Type	ja	
Outcome Subtype	nein	Viele Missing Values, unrelevant
Animal Type	ja	Viel mehr Hunde werden zu Besitzer zurück gegeben
Gender	ja	Könnte interessant sein, ob kastrierte Tiere mehr adoptiert werden
Neutered	ja	Return to owner und adoption viel mehr neutered/spayed; transfer, euthanasia und died viel mehr intact (actionable step: mehr Kastrationen durchführen um mehr Adoptionen)
AgeUponOutcome	ja	Junge Tiere werden eher adoptiert
Breed	ja (-> nein)	aggressiv/nicht aggressiv; mixed/nicht mixed
Color	nein	keine Auswirkung in den Plots zu erkennen

Data Cleaning

Attribut	Normalisierung
Name	Tiere mit Namen = 1, Tiere ohne Namen = 0
Date Time	Keine Missing Values, Sa und So = 1, Mo bis Fr = 0, Uhrzeit als separates Attribut (Arbeitszeiten?), Monat als separates Attribut (Jahreszeiten)
Outcome Type	Keine Missing Values
Season	Winter 1, spring 2, summer 3, autumn 4
Animal Type	Hund = 0, Katze = 1
Gender	Male = 0, Female = 1
Neutered	Kastriert und Sterilisiert = 1, Intakt = 0
AgeUponOutcome	Missing Values auf 1 setzen (Durchschnittsalter)
Breed	Attribut Mix = 1, Reinrassig = 0; Attribut aggressiv = 1, nicht aggressiv = 0

3. CRISP-DM: Data Preparation

- train.head()

	OutcomeType	Name	AgeuponOutcome/MaxAge	Season	Day	AnimalType	Gender	Neutered	BreedMix	Aggressive
0	Return_to_owner	1	0.05000	1	1	0	0	1	1	0
1	Euthanasia	1	0.05000	4	0	1	1	1	1	0
2	Adoption	1	0.10000	1	0	0	0	1	1	0
3	Transfer	0	0.00288	3	1	1	0	0	1	0
4	Transfer	0	0.10000	4	1	0	0	1	0	0

4. CRISP-DM: Modeling

Vorgehensweise:

1. Aufteilen des Datensatzes in einen Train- und Testdatensatz
1. Einführung k-fache Kreuzvalidierung (k=5)
1. Vergleich der Vorhersagegüte verschiedener Modelle
1. Auswahl des Modells mit der besten Vorhersageperformance
1. Optimierung des Modells

4. CRISP-DM: Modeling II

1. Aufteilen des Datensatzes in einen Train- und Testdatensatz

    - Train-Datensatz: 80% der ursprünglichen Daten (train.csv)
    - Test-Datensatz: 20% der ursprünglichen Daten (train.csv)

1. Einführung k-fache Kreuzvalidierung (k=5)

    - Aufteilung Train-Datensatz in 5 gleich große Blöcke (“Folds”)
    - Möglichkeit die Güte eines Modells ohne Vergleich mit dem Test-Datensatz zu messen

→ Overfitting Probleme können erkannt werden

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = True positive; FP = False positive; TN = True negative; FN = False negative

4. CRISP-DM: Modeling III

3. Vergleich der Vorhersagegüte verschiedener Modelle

Random Forest:

```
CV Accuracies:  [0.60346037 0.60743512 0.61865794 0.61482694 0.6246492 ]
Accuracy:  0.6138059142457765
Log Loss:  3.167122082411939
```

Logistic Regression:

```
CV Accuracies:  [0.61444938 0.60603227 0.62052841 0.60968195 0.62371375]
Accuracy:  0.6148811501378028
Log Loss:  0.9315573312298036
```

Dummy Classifier:

```
CV Accuracies:  [0.40285247 0.40285247 0.40285247 0.40294668 0.40294668]
Accuracy:  0.4028901516651054
Log Loss:  20.623443717449362
```

Log Loss: $-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$

VIELEN DANK FÜR DIE AUFMERKSAMKEIT!