МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
Національний університет “Львівська політехніка”  
Інститут комп’ютерних наук та інформаційних технологій  
Кафедра автоматизованих систем управління



**Звіт до практичної роботи № 1**

**з дисципліни “Інформаційні технології смарт-систем”**

Виконала: студентка КН-305

**Трачук Тетяна**

Прийняв: к.т.н., доцент каф. АСУ

**Опотяк Ю.В**

**Львів – 2023**

**Практична робота № 1**

**Завдання**

1. Відповідно до вашого варіанту опишіть можливі три бізнес-цілі, цілі моделювання та критерії успіху до них відповідно. Для цього завдання візьміть до уваги лише пункт «Ваш клієнт».
2. Ознайомившись із пунктом «Запит клієнта» додайте до попередніх цілей задану та сформулюйте цілі моделювання і критерії успіху.
3. Проведіть аналіз даних вашого підприємства (файл «train.csv») та сформуйте звіт про дослідження даних та звіт про якість даних.  
   Приклади функцій, що використовуються для аналізу даних ви можете знайти у файлі *example.ipynb* Для проведення аналізу необхідно встановити python3 та jupyter notebook (послідовність дій вказана у файлі *Встановлення python*)
4. Наведіть приклади цікавих, на вашу думку, фактів з аналізу даних.

**Індивідуальне завдання**

Мій варіант в списку групи – 21, тому використовую для завдання датасет 21%6 = 3. Отже моє індивідуальне завдання виглядає так:

**3 варіант**

Опис Компанії-Замовника: Best Friends Animal Society: Організація в США, відома своїми зусиллями у просуванні усиновлення домашніх тварин і зниженні кількості тварин, що евтаназуються в притулках.

Запит: Організація шукає розв'язок для прогнозування ймовірності усиновлення тварин, що знаходяться в притулку, з метою підвищення їхніх шансів на знаходження нового дому.

Пропозиція Реалізації: Розробка моделі машинного навчання, для передбачення ймовірності усиновлення тварин з притулку. Модель повинна аналізувати різноманітні характеристики тварин, такі як вік, порода, здоров'я, поведінка, та інші значимі фактори, що можуть впливати на шанси тварини бути усиновленою. Цей аналіз допоможе притулку краще розуміти потреби тварин та ефективніше підходити до процесу усиновлення, збільшуючи шанси кожної тварини знайти новий дім.

**Хід роботи**

1. Відповідно до вашого варіанту опишіть можливі три бізнес-цілі, цілі моделювання та критерії успіху до них відповідно. Для цього завдання візьміть до уваги лише пункт «Опис Компанії-Замовника»

Використавши лише пункт «Опис Компанії-Замовника» я сформулювала ось такі три бізнес-цілі:

**Бізнес-ціль 1: Оптимізація витрат на утримання тварин у притулках**

Ціль моделювання: Створити модель, яка на основі даних про тварину зможе точно розрахувати оптимальну тривалість її перебування в притулку до моменту успішного усиновлення.

Критерій успіху: Завдяки точності розрахунків моделі середній час перебування тварин в притулках значно зменшиться, що призведе до оптимізації витрат на їх утримання.

**Бізнес-ціль 2: Підвищення задоволеності нових власників усиновленими тваринами**

Ціль моделювання: Розробити систему, яка на основі даних про особливості тварин та характеристики потенційних усиновлювачів зможе якісно підібрати оптимальну тварину для конкретної людини.

Критерій успіху: Завдяки точності класифікації та рекомендацій системи рівень задоволеності нових власників усиновленими тваринами значно підвищиться.

**Бізнес-ціль 3: Підвищення якості догляду за тваринами у притулках**

Ціль моделювання: Розробити модель машинного навчання, яка на основі даних про породу, вік, стан здоров'я та інші характеристики тварин зможе визначати оптимальний для кожної тварини раціон харчування та режим вигулу.

Критерій успіху: Впровадження розробленої моделі дозволить значно покращити якість догляду за тваринами в притулках. Це відобразиться на поліпшенні їх загального здоров'я та зниженні рівня захворюваності порівняно з поточним станом, що вплине на швидкість всиновлення тварини

1. Ознайомившись із пунктом «Запит клієнта» додайте до попередніх цілей задану та сформулюйте цілі моделювання і критерії успіху.

На основі пункту «Запит клієнта» можна сформулювати таку ціль та ціль моделювання і критерій успіху для неї:

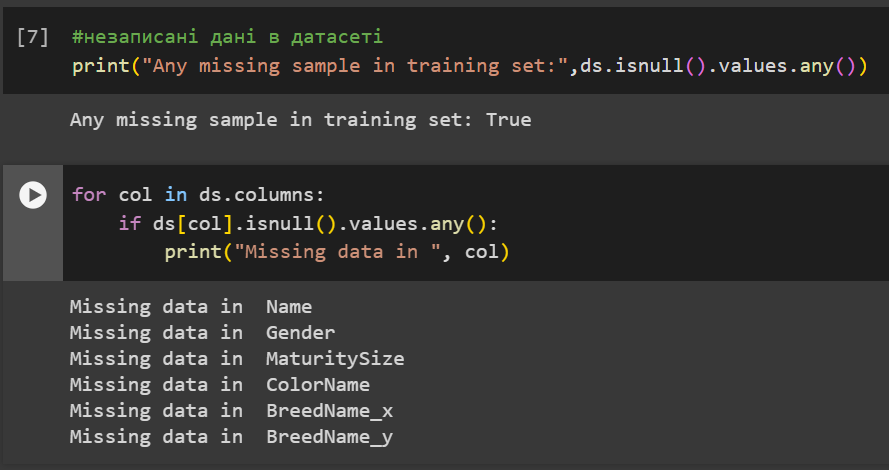
**Бізнес-ціль: Розрахунок шансів тварин на успішне усиновлення**

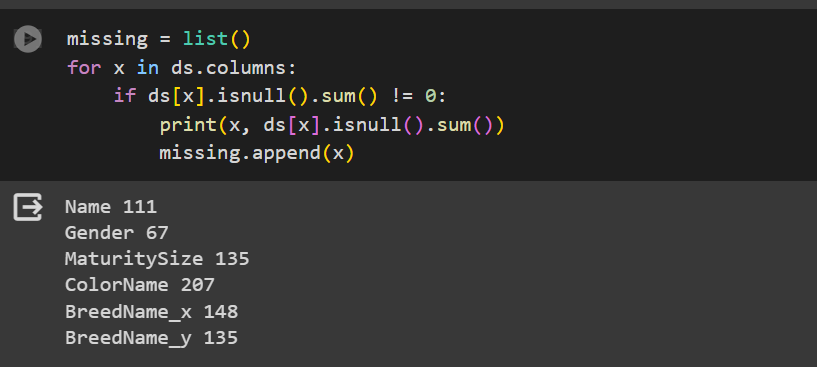
Ціль моделювання: Розробити модель машинного навчання для прогнозування ймовірності успішного усиновлення конкретної тварини на основі її індивідуальних характеристик, таких як вік, стать, порода, колір, поведінкові особливості тощо.

Критерій успіху: Модель може дуже точно прогнозувати ймовірність усиновлення для більшості тварин в притулку за допомогою статистики попередніх випадків усиновлення.

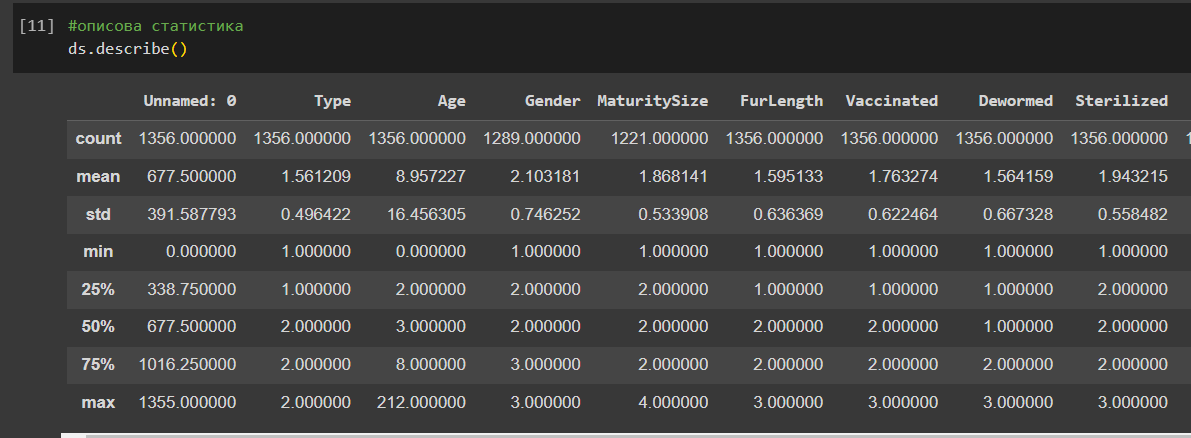
1. Проведіть аналіз даних вашого підприємства (файл «train.csv») та сформуйте звіт про дослідження даних та звіт про якість даних. Приклади функцій, що використовуються для аналізу даних ви можете знайти у файлі *example.ipynb* Для проведення аналізу необхідно встановити python3 та jupyter notebook (послідовність дій вказана у файлі *Встановлення python*)

**Відсутні(незазначені) дані в датасеті**

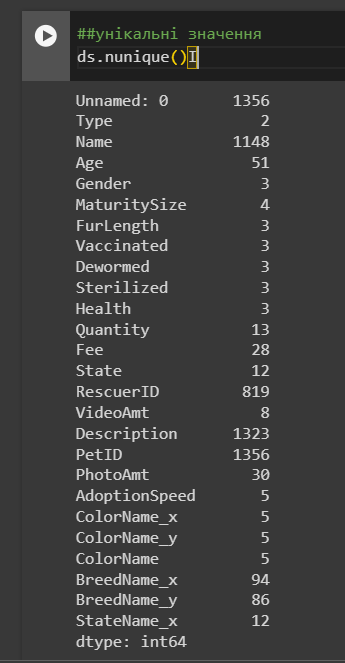
****

****

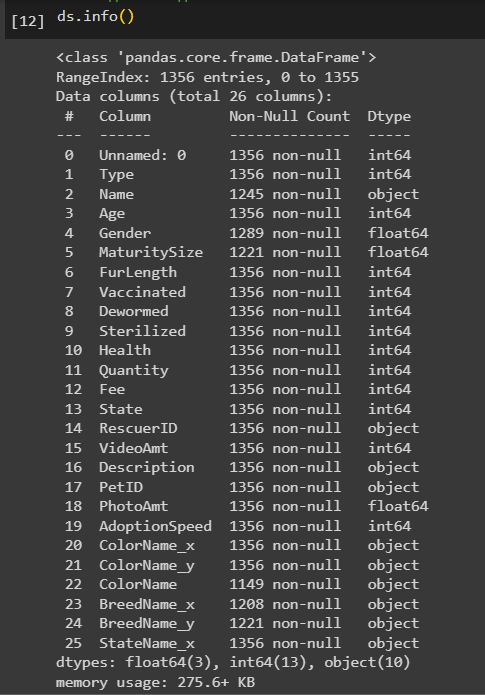
**Описова статистика(варіація значень)**

****

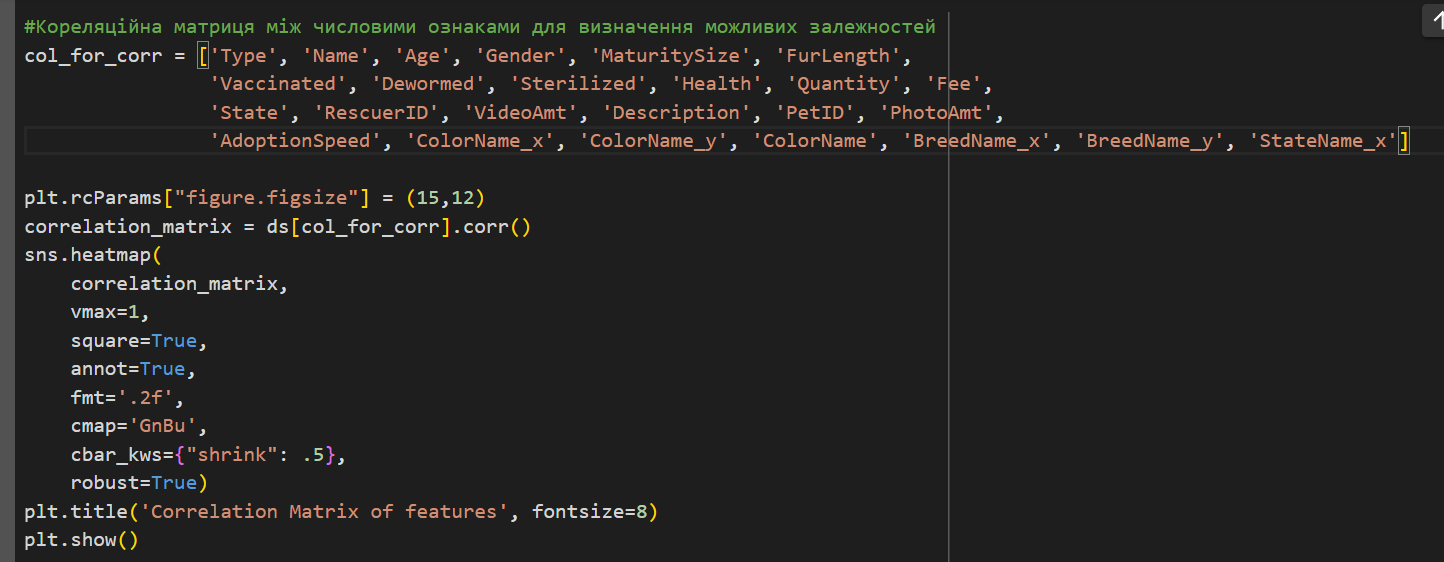
**Унікальні значення**

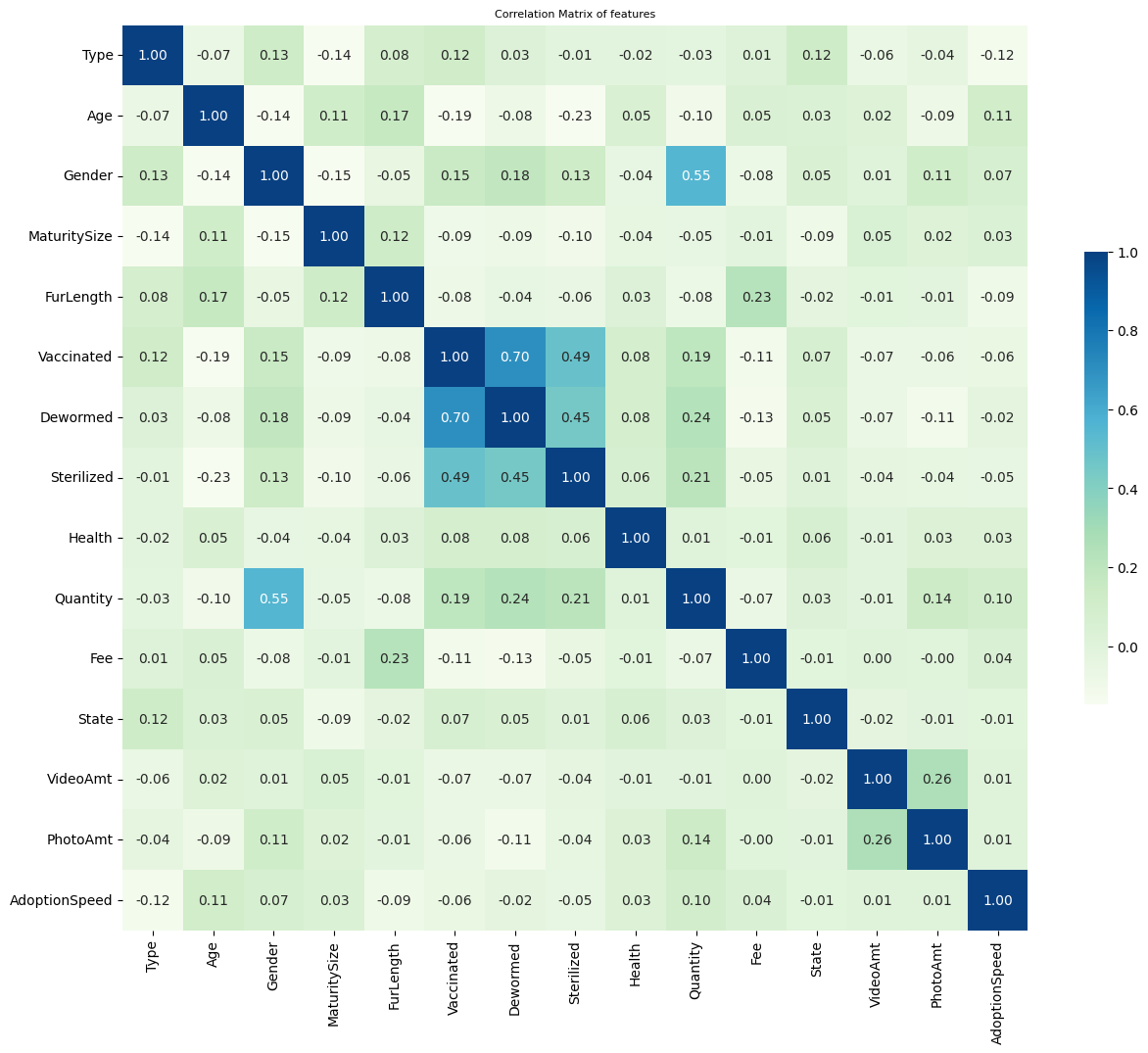
****

**Типи даних характеристик**

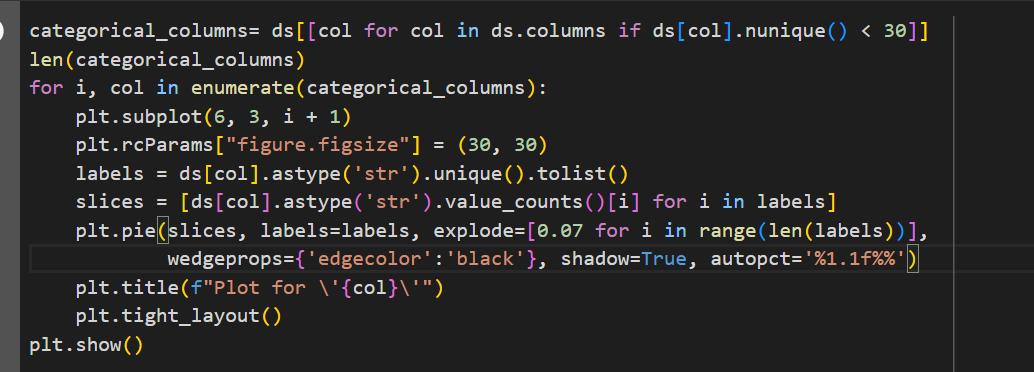
****

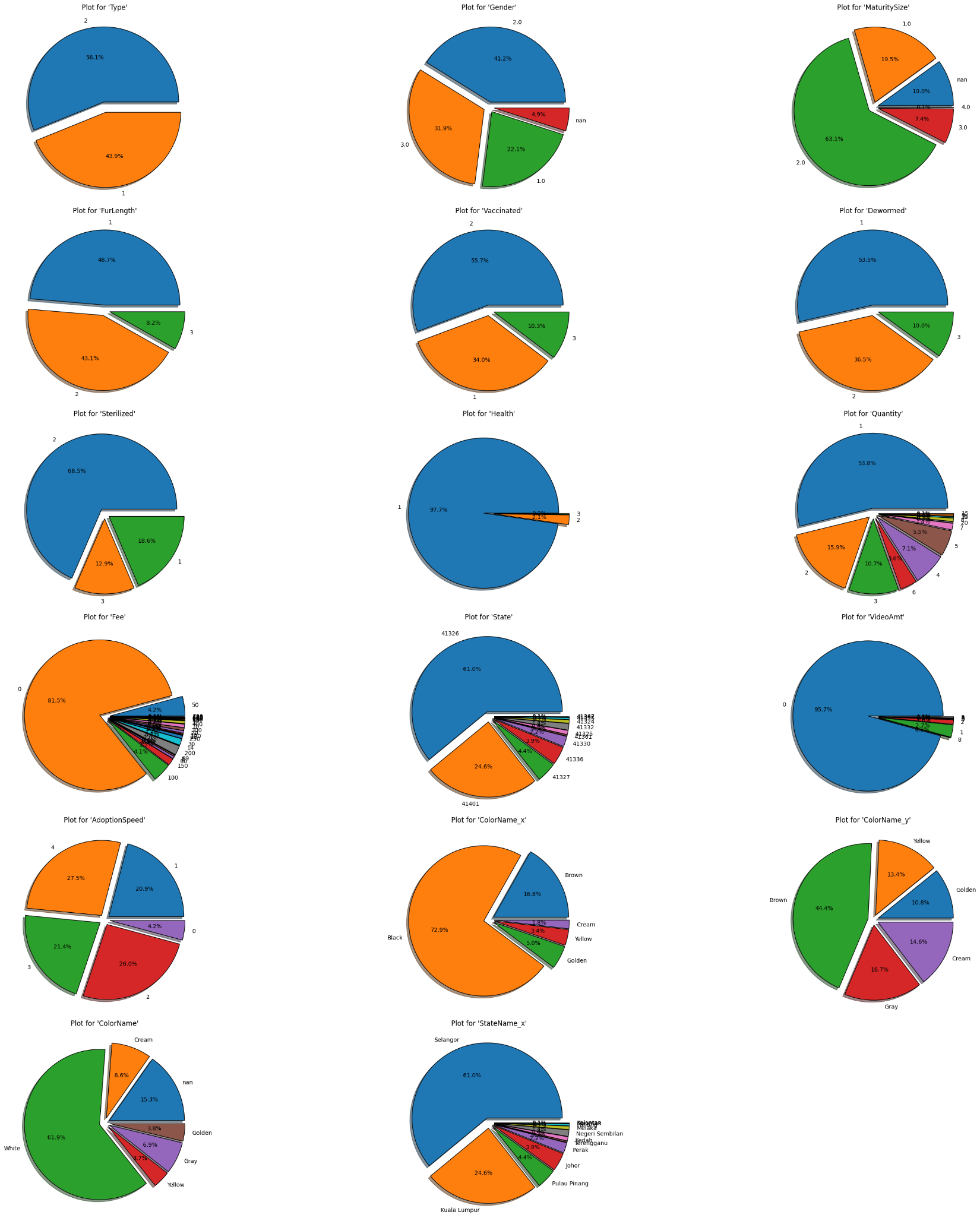
**Кореляційна матриця між числовими ознаками для визначення можливих залежностей**

****

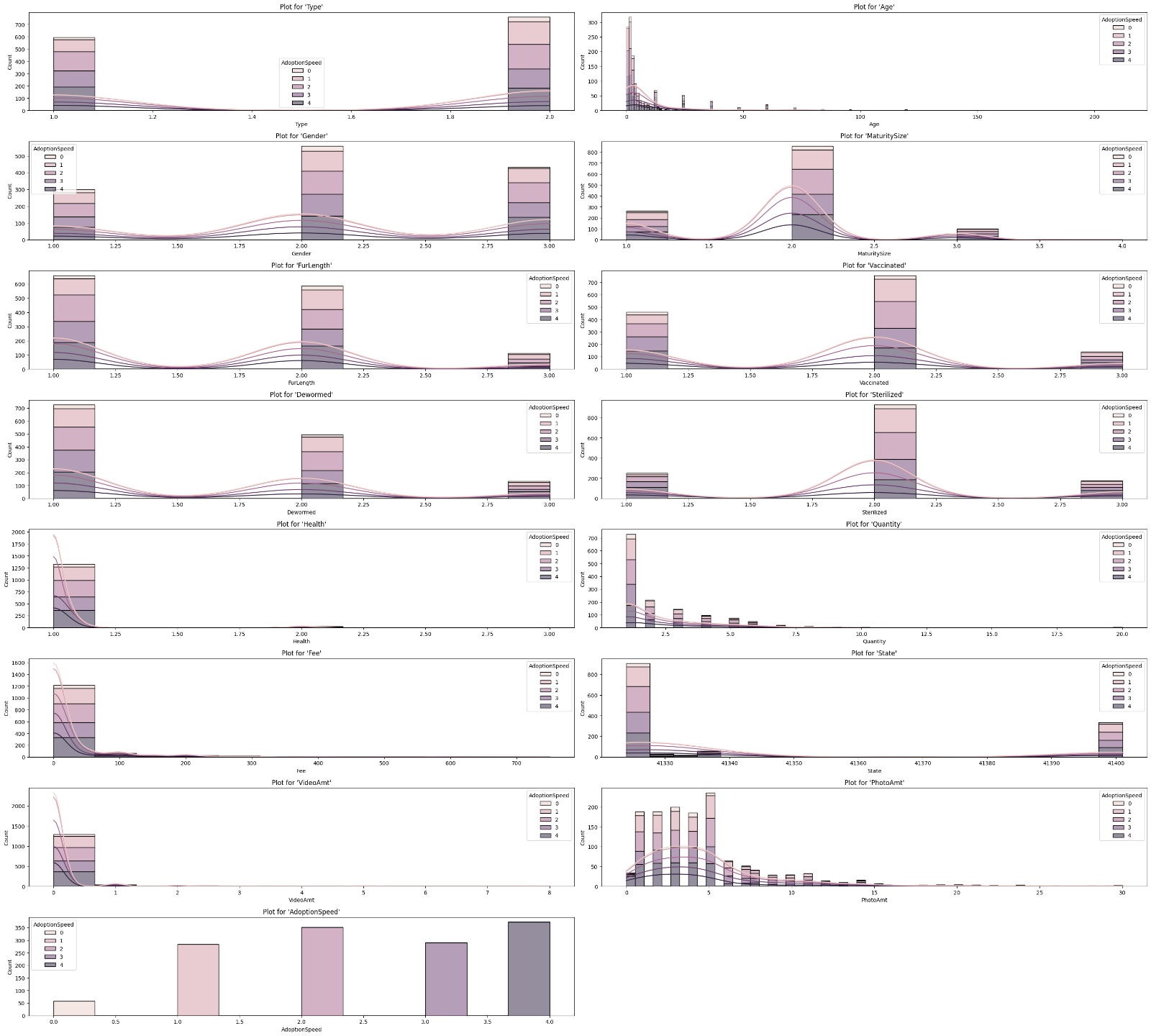


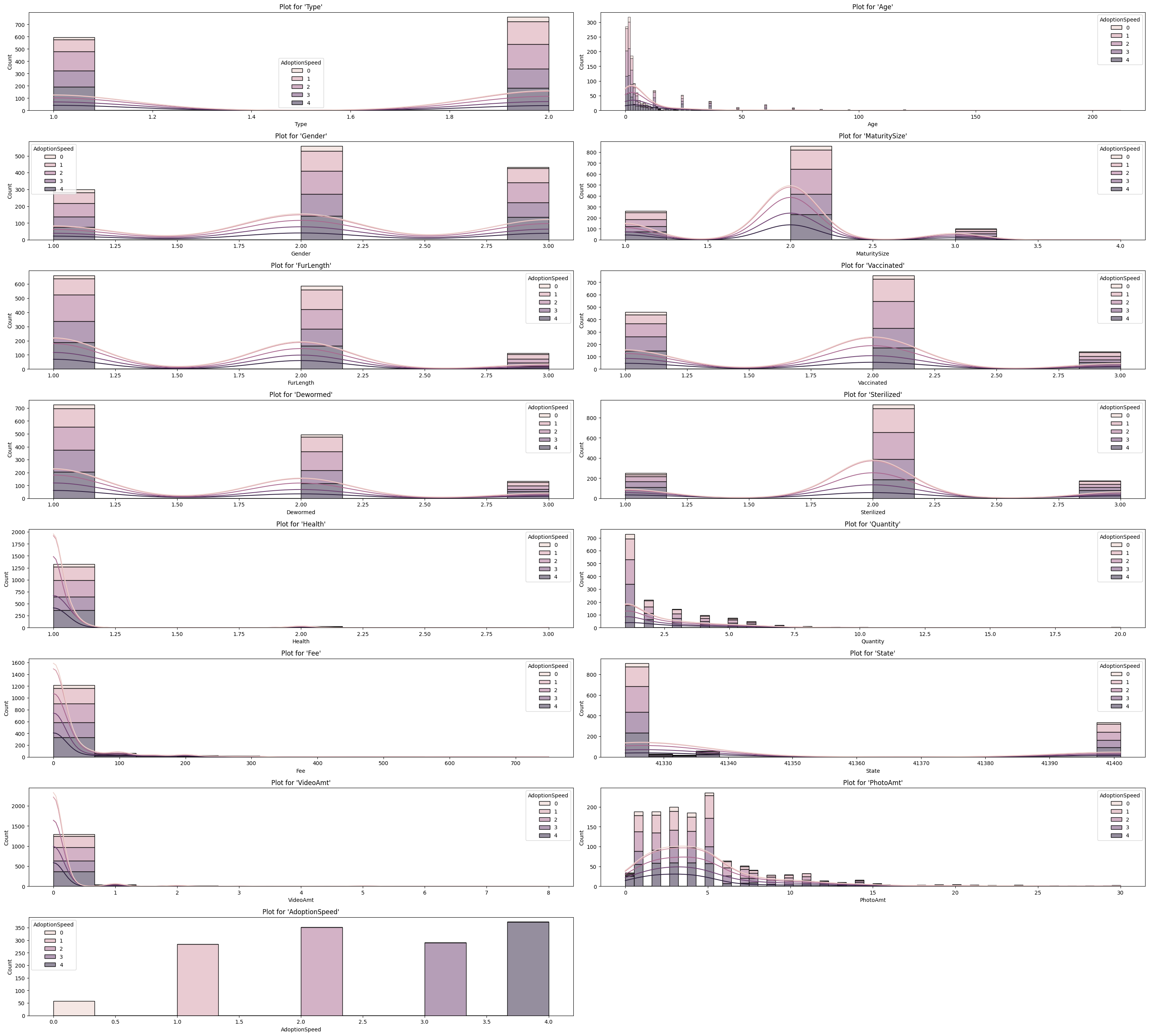
**Діаграми розподілу(варіативності) ознак**

****

****

**Візуалізація співвідношення даних та цільової змінної**





**Звіт про якість даних**

1. **Ви виявили відсутні атрибути та порожні поля? Якщо так, то чи є сенс у таких відсутніх значеннях?**

У деяких стовпцях датасету є відсутні значення. Проаналізувавши типи стовпців можна сказати, що більшість стовпців, значення в яких можуть бути відсутні, містять категоріальну інформацію(є об’єктом). Якщо розглядати контекстні записи стовпців, де відсутні значення, відсутність значень цих ознак відповідає відсутності інформації або унікальності значень даного стовпця.

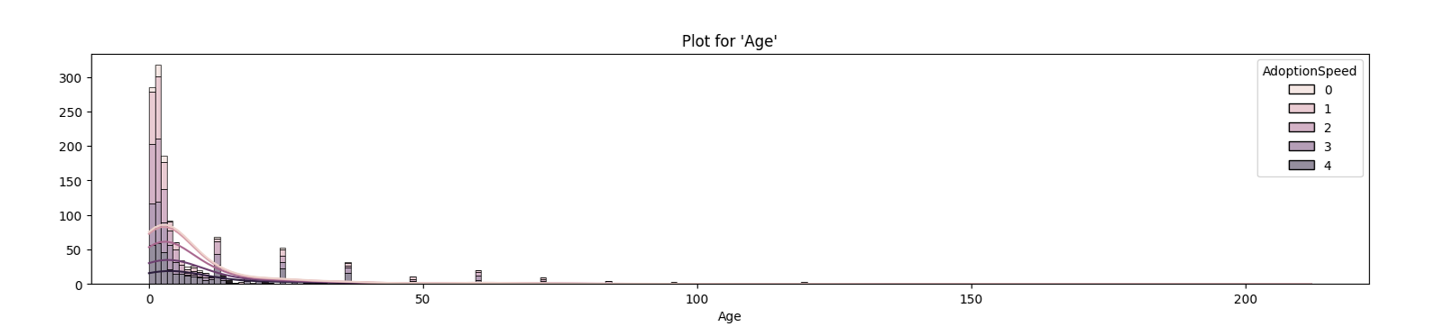
1. **Чи існують правописні невідповідності, які можуть спричинити проблеми при подальших злиттях чи перетвореннях?**

При аналізі категоріальних даних не виявлено значних правописних невідповідностей або варіацій написання, які могли б призвести до проблем при подальшому об'єднанні чи трансформації даних. Категорії виглядають узгоджено та не містять окремих записів з явними помилками в написанні, що свідчить про достатню якість та узгодженість даних в цьому аспекті. Тому на даному етапі немає потреби в застосуванні спеціальних процедур очищення та стандартизації даних для усунення проблем, пов'язаних з невідповідністю написання категоріальних змінних.

1. **Чи досліджували ви відхилення, щоб визначити, чи є вони "шумом" чи явищами, які варто аналізувати далі?**

Серед категоріальних даних шуму не виявлено, в певних категоріях введено лише невелику кількість записів, але вона є достатньою для того, щоб не вважати їх помилковими.

При аналізі числових даних, особливо при графічному представлені, я помітила значення, які відрізняються від інших. Особливо добре це помітно на діаграмі співвідношення між цільовою змінною і значеннями для характеристики Age:



На цьому графіку можна побачити, що є невелика група значень, які достатньо відрізняються від решти. Проте, якщо взяти до уваги ці значення в контексті можливих аномальних вікових значень(довгожителі), можна сказати, що ця група елементів фіксує саме це. Тому можна припустити, що числові дані не містять шуму.

1. **Ви проводили перевірку правдоподібності на значення?**

Під час виконання даного практичного завдання ці дослідження не виконувалися. Для проведення такої перевірки потрібно більш детально дослідити можливі категорії та числові значення для кожної характеристика. На даний момент припустимо, що дані коректні.

1. **Ви думали про виключення даних, які не впливають на ваші гіпотези?**

Як на мене, доцільно буде вилучити стовпець *«Unnamed: 0»*, оскільки це просто унікальний порядковий номер кожного запису нашого датасету, для подальших досліджень його використовувати не доцільно.

1. **Чи зберігаються дані у файлах? Якщо так, то чи розділяються ліміти між файлами? Чи кожен запис містить однакову кількість полів?**

Всі дані зберігаються в файлі типу *.csv*.

Наведіть приклади цікавих, на вашу думку, фактів з аналізу даних.

Цікавими для мене фактами виявилася певна кількість аномальних значень для характеристики Age, які є набагато більшими чим середньостатистичні значення.

Також цікавим фактом є точкові кореляції між деякими ознаками, виділені блакитними відтінками у кореляційній матриці, а саме між характеристиками “vaccinated”, “dewormed” і “sterilized”,а також середні кореляції між ознаками. Це свідчить про взаємозв’язок між різними ознаками, які пов’язані зі станом здоров’я та часті поєднання відповідних ознак, що допоможе при побудові моделі для багатокласової класифікації.

**Висновок:** під час виконання даної практичної роботи я ознайомилася з основними методами аналізу даних, порядком формування бізнес-цілей, зробила аналіз датасету на відсутні значення, пошук та аналіз аномальних значень, виконала візуалізацію даних та застосувала усі можливості для опрацювання свого індивідуального завдання; також виявила та описала цікаві приклади фактів з аналізу.