Aluna: Clara Bastos
CPF/Matrícula: 177.805.247-95
Projeto MVP - Engenharia de Dados

Objetivo:

Contextualização do problema: A falta de uma visão integrada e histórica dos dados dificulta a tomada de decisão dos abrigos. Isso impacta diretamente o bem-estar e a vida dos animais, dado que um processo de adoção mal planejado pode ocasionar em devoluções ou em permanência por grandes períodos nos abrigos.

Objetivo: O objetivo desse projeto é identificar quais fatores influenciam na entrada, saída no e permanência dos animais nos centros de recolhimento auxiliar na gestão operacional e estratégica dos abrigos garantindo campanhas de adoções mais assertivas e proteção animal.

Perguntas:

Pergunta - Objetivo: Quais fatores influenciam no tempo de permanência e no desfecho dos animais do abrigo?

1. Qual é o perfil predominante dos animais que chegam ao abrigo em termos de espécie, idade e condição?
2. Qual é o perfil predominante dos animais que saem do abrigo em termos de espécie, idade, sexo e castração?
3. Qual é a distribuição dos principais tipos de entrada de animais?
    a. Qual o perfil predominante dos animais que são devolvidos pelos donos?
    b. Os animais entregues ao abrigo pelos donos apresentam alguma condição de saúde?
4. Qual é a distribuição dos principais desfechos dos animais?
5. Como o tempo médio de permanência no abrigo varia de acordo com espécie, raça e idade?
6. Existem padrões sazonais no volume de entradas e adoções ao longo do ano?

Base: Austin Animal Center Shelter Intakes and Outcomes - Kaggle

Link Acesso:
https://www.kaggle.com/datasets/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes

Plataforma: Data Bricks

Tabelas Originais - Dados Brutos:

Base aac_intakes: 80.187 registros - 12 colunas
Colunas:

1. age_upon_intake (string) - Refere-se a idade do animal no momento da entrada no abrigo, a idade pode variar de dias, semanas, meses e anos. Não contém valores nulos.

2. animal_id (string) - Refere-se ao registro do animal. Formato: Letra **A** seguida de seis números. Não contém valores nulos.
3. animal_type (string) - Refere-se a espécie do animal. Existem valores classificados como “outros”.
4. breed (string) - Refere-se a raça do animal. Não contém valores nulos.
5. color (string) - Refere-se a raça do animal. Não contém valores nulos.
6. datetime (timestamp) - Refere-se a data e hora de entrada do animal no abrigo. Formato dos dados: data (ano - mês - dia) seguido da hora (hh:mm:ss). Não contém valores nulos.
7. datetime2 (timestamp) - A coluna é uma cópia da anterior (6. datatime). Mesmos valores e formatos.
8. found_location (string) - Refere-se ao local em que o animal foi encontrado. Não contém valores nulos.
9. intake_condition (string) - Refere-se a condição de saúde do animal no momento de entrada no abrigo. Existem valores classificados como “outros”.
10. intake_type (string) - Refere-se ao tipo de entrada do animal no abrigo. Não contém valores nulos.
11. name (string) - Refere-se ao nome do animal. É um campo opcional. Contém valores nulos. Valores sem um formato padrão (alguns nomes aparecem com asteriscos (*) na frente, outros não).
12. sex_upon_intake (string) - Refere-se as informações de sexo e castração. sexo do animal e se ele é castrado ou não. Contém valores nulos e valores classificados como desconhecidos.

aac_outcomes: 80.681 registros - 12 colunas
Colunas:

1. age_upon_outcome (string) - Refere-se a idade do animal no momento da saída do abrigo, a idade pode variar de dias, semanas, meses e anos. Não contém valores nulos.
2. animal_id (string) - Refere-se ao registro do animal. Formato: Letra **A** seguida de seis números. Não contém valores nulos.
3. animal_type (string) - Refere-se a espécie do animal. Existem valores classificados como “outros”.
4. breed (string) - Refere-se a raça do animal. Não contém valores nulos.
5. color (string) - Refere-se a cor do animal. Não contém valores nulos.
6. date_of_birth (timestamp) - Refere-se a data de nascimento do animal. Esses dados estão no mesmo formato que a coluna datetime (data e hora), porém o registro fica zerado. Não contém valores nulos.
7. datetime (timestamp) - Refere-se a data e hora de saída do animal no abrigo. Formato dos dados: data (ano - mês - dia) seguido da hora (hh:mm:ss). Não contém valores nulos.
8. monthyear (timestamp) - A coluna é uma cópia da anterior (7. datatime). Mesmos valores e formatos.
9. name (string) - Refere-se ao nome do animal. É um campo opcional. Contém valores nulos. Valores sem um formato padrão (alguns nomes aparecem com asteriscos (*) na frente, outros não).
10. outcome_type (string) - Refere-se ao tipo da saída do animal do abrigo. Essa coluna contém valores nulos.

11. outcome_subtype (string) - Refere-se ao subtipo do animal. É um campo opcional. Contém valores nulos.
12. sex_upon_outcome (string) - Refere-se as informações de sexo e castração. sexo do animal e se ele é castrado ou não. Contém valores nulos e valores classificados como desconhecidos.

ETL

Atenção: Os códigos descritos nessa sessão encontram-se na pasta ETL no github: https://github.com/clararmbastos-310/Projeto-Pos-Engenharia-de-Dados/tree/main/etl

Os arquivos aac_intakes e aac_outcomes foram importados manualmente para o databricks, a partir do caminho: Catalog - Add Data - Create or Modify Table.

Para o trabalho e análises posteriores, as bases passaram por um processo de tratamento e organização nas camadas bronze, prata e ouro. Esse tratamento foi descrito, em código, no notebook ETL_Base, armazenado na pasta do github "ETL".

1. Camada Bronze (bronze_aac_intakes e bronze_aac_outcomes): Os dados armazenados na camada bronze em sua forma bruta, preservando a estrutura e o conteúdo original conforme disponibilizado pela plataforma Kaggle.
2. Camada Prata (prata_aac_intakes e prata_aac_outcomes): Os dados passaram por um processo de limpeza, tratamento dos valores nulos e padronização

   Transformações (comuns para as duas bases):

   - Uso do comando SELECT DISTINCT para evitar valores duplicados;
   - Renomeação das colunas, usando o comando AS, para facilitar o entendimento dos dados e futuras pesquisas e análises.
   - Tratamento da coluna Nome. A coluna Nome é uma coluna de informação opcional, portanto contém valores nulos. Além disso, não segue um padrão de formato, os nomes são escritos de diferentes formas (às vezes com asteriscos na frente ou no final). Então, para os casos de valores nulos, os dados foram transformados na categoria "Unknown" ("Desconhecido"). E, para os casos do nomes escritos com (*) em alguma parte, o comando REPLACE para substituir o asterisco por nada, ou seja, removê-lo do nome.
     Imagem evidência:

```
CASE
    WHEN name IS NULL THEN 'Unknown'
    WHEN name like '%*%' THEN REPLACE (name,'*','')
    ELSE name
  END as nome_animal,
```

   - As colunas de data (date_upon_intake - bronze_aac_intakes, date_upon_outcome e date_of_birth - bronze_aac_outcomes) estavam no

formato data e hora. Como a hora não é relevante para análise, foi usado o comando CAST para mudar o formato dos dados para apenas data.
Imagem evidência:

```
cast(date_of_birth as date) as data_nascimento,
```

- As colunas sex_upon_intaken e sex_upon_outcome trazem as informações sobre o sexo e a castração dos animais. Para o trabalho, o ideal é que essas informações estejam em colunas diferentes, para que seja possível a análise dessas informações de forma separada. Além disso, essa coluna contém valores nulos e também na categoria “Unknown”. Então, na camada prata, os valores nulos foram transformados na categoria “Unknown” e as informações separadas.
Imagem evidência:

```
CASE
  WHEN sex_upon_intake IS NULL THEN 'Unknown'
  WHEN lower(sex_upon_intake) like '%female%' THEN 'Female'
  WHEN lower(sex_upon_intake) LIKE '%male%' THEN 'Male'
  ELSE 'Unknown'
END AS sexo,

CASE
  WHEN sex_upon_intake IS NULL THEN 'Unknown'
  WHEN lower(sex_upon_intake) LIKE '%spayed%' THEN 'Yes'
  WHEN lower(sex_upon_intake) LIKE '%neutered%' THEN 'Yes'
  WHEN lower(sex_upon_intake) LIKE '%intact%' THEN 'No'
  ELSE 'Unknown'
END AS castrado
```

<u>Transformações (Base prata_aac_outcomes):</u>

- As colunas de tipo e subtipo de saída foram tratadas, uma vez que continha valores nulos e categorias redundantes. Então, os dados nulos foram transformados na categoria “Unknown” e as categorias padronizadas.
Imagem evidência:

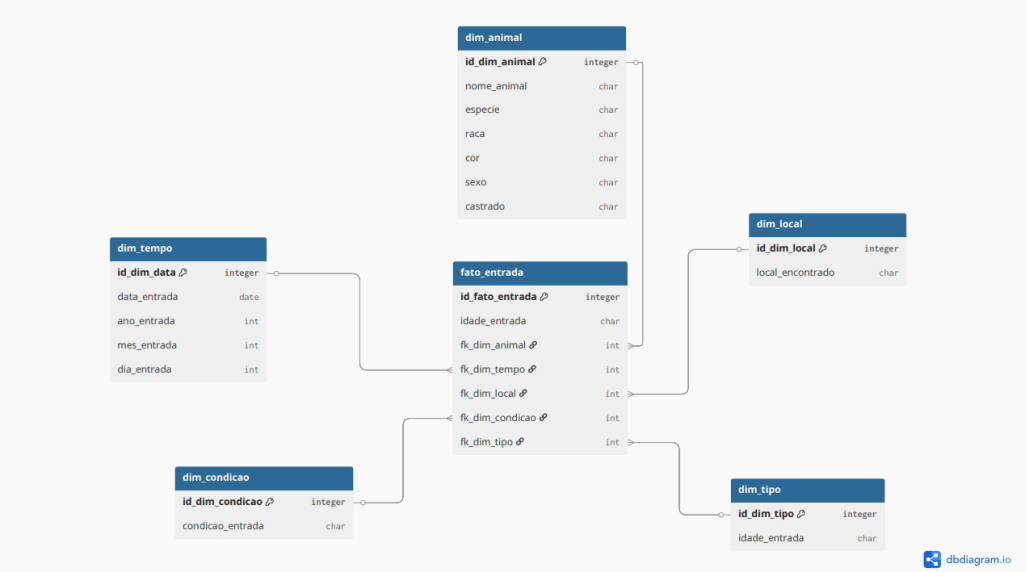
```
CASE
  WHEN outcome_type IS NULL THEN 'Unknown'
  WHEN outcome_type = 'Rto-Adopt' THEN 'Return to Owner'
  WHEN outcome_type = 'Relocate' THEN 'Transfer'
  WHEN outcome_type = 'Disposal' THEN 'Euthanasia'
  ELSE outcome_type
END AS tipo_saida,

CASE
  WHEN outcome_subtype IS NULL THEN 'Unknown'
  WHEN outcome_subtype = 'In Foster' THEN 'Foster'
  WHEN outcome_subtype = 'Aggressive' THEN 'Behavior'
  WHEN outcome_subtype IN ('SCRP', 'Suffering', 'Rabies Risk', 'Snr', 'At Vet', 'In Surgery') THEN 'Medical'
  WHEN outcome_subtype IN ('Court/Investigation', 'Possible Theft') THEN 'Investigation'
  ELSE outcome_subtype
END AS subtipo_saida,
```

3. Camada Ouro: Após a finalização dos processos de tratamento e padronização na camada Prata, os dados foram transformados na camada Ouro de acordo com as regras de negócio e os objetivos definidos para o desenvolvimento do trabalho. Inicialmente, foram criadas tabelas (ouro_aac_intakes e ouro_aac_outcomes) desconsiderando os valores categorizados como “Unknown” ou “Others”, uma vez que esses registros não agregam valor e poderiam comprometer a análise dos dados.
   Imagem evidência:

```
--- Intakes

DROP TABLE IF EXISTS ouro_aac_intakes;
CREATE TABLE ouro_aac_intakes AS
SELECT * FROM prata_aac_intakes
WHERE especie != 'Other'
AND sexo != 'Unknown'
AND castrado != 'Unknown'
AND condicao_entrada != 'Other';

--- Outcomes

DROP TABLE IF EXISTS ouro_aac_outcomes;
CREATE TABLE ouro_aac_outcomes AS
SELECT * FROM prata_aac_outcomes
WHERE especie != 'Other'
AND sexo != 'Unknown'
AND castrado != 'Unknown'
AND tipo_saida != 'Unknown';
```

   Após o tratamento dos dados, as bases foram estruturadas segundo o modelo dimensional em esquema estrela.
   *Observação: Os códigos das tabelas fato (intakes e outcomes) e dimensões foram feitos no databricks.Porém, para a visualização diagramática foi usada a plataforma dbdiagram.io, apresentada nas aulas da disciplina.

Base aac_intakes:
Modelo Estrela:



Catálogo de Dados:

Tabela Dimensão Animal (dim_animal):

Colunas	Descrição	Tipo	Valores Esperados	Unicidade	Nulidade
id_dim_animal	Chave Primária - Identificação dos registros	int	Inteiros positivos Mínimo: 1	Sim	Não
nome_animal	Nome do animal	char	Nomes próprios	Não	Sim
especie	Espécie do animal	char	“Livestock”, “Cat”, “Dog”, “Bird”	Não	Não
raca	Raça do animal	char	Raças Padronizadas (Ex: “Pit Bull”)	Não	Não
cor	Cor do animal	char	Cores (Ex: “White”, “Tricolor”)	Não	Não
sexo	Sexo do animal	char	“Female”, “Male”	Não	Não
castrado	Castração	char	“Yes”, “No”	Não	Não

*As colunas raça e cor possuem muitas categorias, por isso, no catálogo de dados foram adicionados apenas exemplos.

Tabela Dimensão Tempo (dim_tempo):

Colunas	Descrição	Tipo	Valores Esperados	Unicidade	Nulidade
id_dim_data	Chave Primária - Identificação dos registros	int	Inteiros positivos Mínimo: 1	Sim	Não

data_entrada	Data de entrada dos animais no abrigo	date	Data (aaaa-mm-dd) Registros variam de: 01/10/2013 até 17/03/2018	Não	Não
ano_entrada	Ano de entrada dos animais no abrigo	int	Inteiros Valor mínimo: 2013 Valor máximo: 2018	Não	Não
mes_entrada	Mês de entrada dos animais no abrigo	int	Inteiros Valor mínimo: 1 Valor máximo: 12	Não	Não
dia_entrada	Dia de entrada dos animais do abrigo	int	Inteiros Valor mínimo: 1 Valor máximo: 31	Não	Não

Tabela Dimensão Condição (dim_condicao):

Colunas	Descrição	Tipo	Valores Esperados	Unicidade	Nulidade
id_dim_condicao	Chave Primária - Identificação dos registros	int	Inteiros positivos Mínimo: 1	Sim	Não
condicao_entrada	Condição de saúde dos animais que entram no abrigo	char	“Aged”, “Feral”, “Injured”, “Normal”, “Nursing”, “Pregnant”, “Sick”	Não	Não

Tabela Dimensão Tipo (dim_tipo):

Colunas	Descrição	Tipo	Valores Esperados	Unicidade	Nulidade
id_dim_tipo	Chave Primária - Identificação dos registros	int	Inteiros positivos Mínimo: 1	Sim	Não
tipo_entrada	Motivo entrada do animal no abrigo	char	“Euthanasia Request”, “Owner Surrender”, “Public Assist”, “Stray”, “Wildlife”	Não	Não

Tabela Dimensão Local (dim_local):

Colunas	Descrição	Tipo	Valores Esperados	Unicidade	Nulidade
id_dim_local	Chave Primária - Identificação dos registros	int	Inteiros positivos Mínimo: 1	Sim	Não
local_encontrado	Local onde o animal foi encontrado	char	Endereço: número - rua - cidade	Não	Não

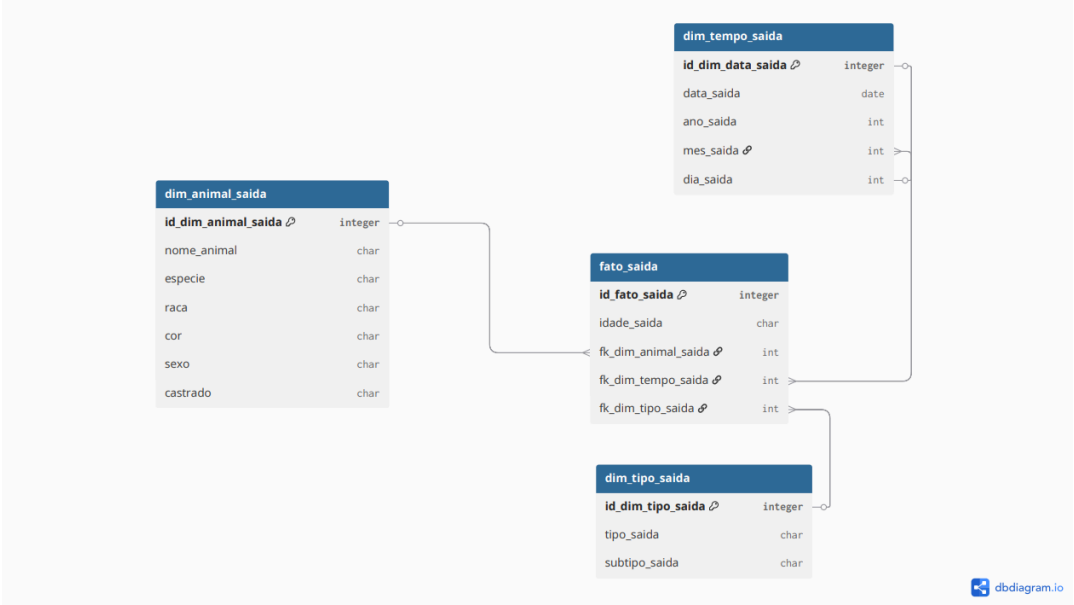
Tabela Fato Entrada (fato_entrada):

Colunas	Descrição	Tipo	Valores Esperados	Unicidade	Nulidade
id_fato_entrada	Chave Primária - Identificação dos registros	*char	A + sequência de seis números	Sim	Não
idade_entrada	Idade do animal no momento da entrada no abrigo	char	Texto (Ex: “7 years” “6 months”)	Não	Não
fk_dim_animal	Chave Estrangeira - Referencia dim_animal	int	Inteiros positivos Mínimo: 1	Não	Não
fk_dim_data	Chave Estrangeira - Referencia dim_tempo	int	Inteiros positivos Mínimo: 1	Não	Não
fk_dim_condicao	Chave Estrangeira - Referencia dim_condicao	int	Inteiros positivos Mínimo: 1	Não	Não
fk_dim_tipo	Chave Estrangeira - Referencia dim_tipo	int	Inteiros positivos Mínimo: 1	Não	Não
fk_dim_local	Chave Estrangeira - Referencia dim_local	int	Inteiros positivos Mínimo: 1	Não	Não

*Na imagem do modelo, a chave primária da tabela fato está representada como "int". Porém, na verdade, o valor é "char". Diferente das demais, que são inteiros auto generados, a chave da tabela fato vem da informação da coluna animal_id tabela original, que é uma string, no formato A + (sequência de seis números).

Base aac_outcomes:

Modelagem Estrela:



Catálogo de Dados:

Tabela Dimensão Animal (dim_animal): semelhante ao apresentado anteriormente na base intakes.

Tabela Dimensão Tempo (dim_tempo_saida): semelhante ao apresentado anteriormente na base intakes.

Tabela Dimensão Tipo (dim_tipo_saida):

Colunas	Descrição	Tipo	Valores Esperados	Unicidade	Nulidade
id_dim_tipo_saida	Chave Primária - Identificação dos registros	int	Inteiros positivos Mínimo: 1	Sim	Não
tipo_saida	Motivo da saída do animal do abriga	char	“Adoption”, “Transfer”, “Return to Owner”, “Euthanasia”, “Died”, “Missing”	Não	Não
subtipo_saida	Motivo, mais específico, do motivo de saída do animal	char	“Partner”, “Foster”, “Medical”, “Behavior”, “Offsite”, “In Kennel”, “Investigation”, “Enroute”, “Barn”, “Unknown”	Não	Sim

Tabela Fato Saída (fato saída):

Colunas	Descrição	Tipo	Valores Esperados	Unicidade	Nulidade
id_fato_entrada	Chave Primária - Identificação dos registros	*char	A + sequência de seis números	Sim	Não
idade_saida	Idade do animal no momento da saída do abrigo	char	Texto (Ex: “7 years” “6 months”)	Não	Não
fk_dim_animal	Chave Estrangeira - Referencia dim_animal	int	Inteiros positivos Mínimo: 1	Não	Não
fk_dim_data_saida	Chave Estrangeira - Referencia dim_tempo	int	Inteiros positivos Mínimo: 1	Não	Não
fk_dim_tipo_saida	Chave Estrangeira - Referencia dim_condicao	int	Inteiros positivos Mínimo: 1	Não	Não

*Na imagem do modelo, a chave primária da tabela fato está representada como "int". Porém, na verdade, o valor é "char". Diferente das demais, que são inteiros auto generados, a chave da tabela fato vem da informação da coluna animal_id tabela original, que é uma string, no formato A + (sequência de seis números).

Análise dos Dados: Perguntas e Respostas

Atenção: Os códigos descritos nessa sessão encontram-se na pasta Perguntas e Respostas no github:

https://github.com/clararmbastos-310/Projeto-Pos-Engenharia-de-Dados/tree/main/Perguntas%20e%20Respostas

A pasta contém três arquivos em formatos distintos:

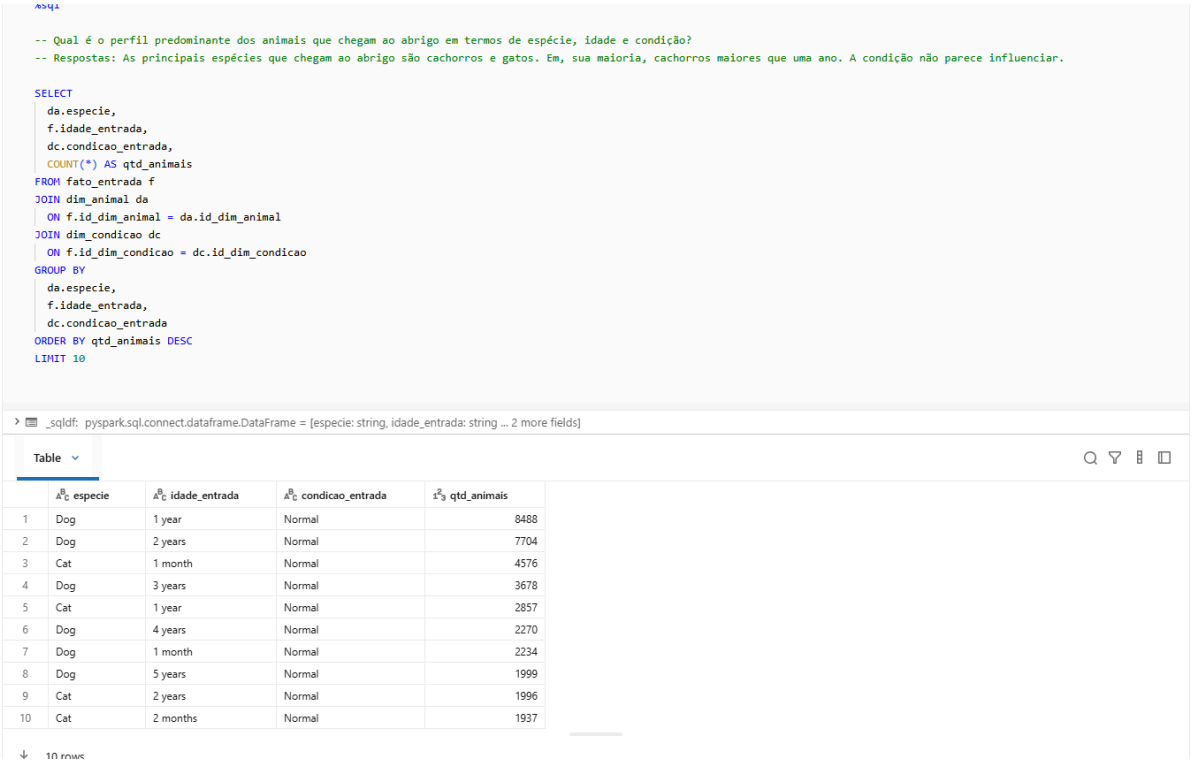
- Arquivo Notebook (.ipynb): Contém o código-fonte original e os outputs processados diretamente no Databricks. Porém apresenta problemas de renderização, prejudicando a visualização dos dados.

- Arquivo Estático (.html): Exportação fiel da interface do Databricks. Porém, precisa da extensão GitHub HTML Preview para a visualização dos dados.

- Relatório Consolidado (.pdf): Documento de backup contendo capturas de tela dos inputs e outputs de cada consulta. Este formato foi incluído por garantia, visando contornar eventuais instabilidades de renderização do GitHub e assegurar que todos os resultados possam ser conferidos de forma clara e imediata.

Para responder a pergunta principal do trabalho “Quais fatores influenciam o tempo de permanência e o desfecho dos animais no abrigo?”, a análise foi estruturada em perguntas secundárias. De forma a destrinchar as variáveis, facilitando a construção de uma resposta abrangente e fundamentada para a pergunta principal.

Pergunta 1: Qual é o perfil predominante dos animais que chegam ao abrigo em termos de espécie, idade e condição?



```sql
%sql

-- Qual é o perfil predominante dos animais que chegam ao abrigo em termos de espécie, idade e condição?
-- Respostas: As principais espécies que chegam ao abrigo são cachorros e gatos. Em, sua maioria, cachorros maiores que uma ano. A condição não parece influenciar.

SELECT
  da.especie,
  f.idade_entrada,
  dc.condicao_entrada,
  COUNT(*) AS qtd_animais
FROM fato_entrada f
JOIN dim_animal da
  ON f.id_dim_animal = da.id_dim_animal
JOIN dim_condicao dc
  ON f.id_dim_condicao = dc.id_dim_condicao
GROUP BY
  da.especie,
  f.idade_entrada,
  dc.condicao_entrada
ORDER BY qtd_animais DESC
LIMIT 10
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [especie: string, idade_entrada: string ... 2 more fields]

Table

	especie	idade_entrada	condicao_entrada	qtd_animais
1	Dog	1 year	Normal	8488
2	Dog	2 years	Normal	7704
3	Cat	1 month	Normal	4576
4	Dog	3 years	Normal	3678
5	Cat	1 year	Normal	2857
6	Dog	4 years	Normal	2270
7	Dog	1 month	Normal	2234
8	Dog	5 years	Normal	1999
9	Cat	2 years	Normal	1996
10	Cat	2 months	Normal	1937

10 rows

Resposta 1: O perfil predominante dos animais que chegam aos abrigos é composto por cães, maiores que um ano de idade e gatos, menores que um ano de idade. Quanto ao estado de saúde, a condição “Normal” aparece para todas as espécies e faixas etárias, sugerindo que a condição física não é um fator determinante para a entrada do animal no abrigo.

Pergunta 2: Qual é o perfil predominante dos animais que saem do abrigo em termos de espécie, idade, sexo e castração?

2

```
%sql
-- Qual é o perfil predominante dos animais que saem do abrigo em termos de espécie, idade, sexo e castração?
-- Resposta: Cães são as principais espécies que chegam aos abrigos. Em sua maioria, jovens (até um ano) e castrados.

SELECT
  das.especie,
  das.castrado,
  fs.idade_saida,
  dtis.tipo_saida,
  COUNT(*) AS qtd_animais
FROM fato_saida fs
JOIN dim_animal_saida das
  ON fs.id_dim_animal_saida = das.id_dim_animal_saida
JOIN dim_tipo_saida dtis
  ON fs.id_dim_tipo_saida = dtis.id_dim_tipo_saida
GROUP BY
  das.especie,
  das.castrado,
  fs.idade_saida,
  dtis.tipo_saida
ORDER BY qtd_animais DESC
LIMIT 10
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [especie: string, castrado: string ... 3 more fields]

Table

	especie	castrado	idade_saida	tipo_saida	qtd_animais
1	Dog	Yes	1 year	Adoption	4684
2	Cat	Yes	2 months	Adoption	4636
3	Dog	Yes	2 years	Adoption	3532
4	Dog	Yes	2 months	Adoption	2453
5	Dog	Yes	2 years	Return to Owner	1933
6	Cat	Yes	3 months	Adoption	1716
7	Dog	Yes	1 year	Return to Owner	1603
8	Dog	Yes	3 years	Adoption	1483
9	Dog	Yes	3 years	Return to Owner	1213
10	Dog	Yes	1 year	Transfer	1109

10 rows

Resposta 2: O perfil predominante dos animais que saem dos abrigos é composto por cães e gatos jovens e castrados. Os dados de saída revelam a adoção “Adoption” como o principal desfecho dos animais. Porém, observa-se, também, um volume expressivo de animais que retornam aos donos “Return to Owner”, indicando a eficácia do abrigo na recuperação de animais perdidos.

Pergunta 3: Qual é a distribuição dos principais tipos de entrada de animais?

3

```sql
%sql
-- Qual é a distribuição dos principais tipos de entradas dos animais?
-- Principais são cachorros de ruas e devolvidos pelo o dono. Devolvidos pelo dono é um caso crítico quando pensamos em adoções. Vamos analisar melhor.

SELECT
  dti.tipo_entrada,
  COUNT(*) AS qtd
FROM fato_entrada f
JOIN dim_tipo dti
  ON f.id_dim_tipo = dti.id_dim_tipo
GROUP BY dti.tipo_entrada
order by qtd DESC
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [tipo_entrada: string, qtd: long]

Table

	tipo_entrada	qtd
1	Stray	52789
2	Owner Surrender	14744
3	Public Assist	4788
4	Euthanasia Request	232
5	Wildlife	2

5 rows

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

Resposta 3: A maior parte das entradas do abrigo vem dos animais de rua “Stray”, o que reforça o papel da instituição no recolhimento de animais abandonados. Contudo, destaca-se como o segundo maior fator a entrega feita pelo próprio tutor “Owner Surrender”. Este cenário é identificado como um ponto crítico para a gestão de adoções, sugerindo a necessidade de investigar as causas que levam os donos a desistirem dos seus animais, visando estratégias de prevenção ao abandono. Por isso, algumas consultas extras foram feitas para complementar a resposta três:

a. Qual o perfil predominante dos animais devolvidos pelos donos? Quais as condições desses animais?

4

```sql
%sql
-- Perfil predominante dos animais devolvidos pelo dono?
-- O perfil predominante dos animais devolvidos são cachorros que já passaram pela fase de filhote (animais maiores que 1 ano)

SELECT
  da.especie,
  f.idade_entrada,
  COUNT(*) AS qtd_animais
FROM fato_entrada f
JOIN dim_animal da
  ON f.id_dim_animal = da.id_dim_animal
JOIN dim_tipo dti
  ON f.id_dim_tipo = dti.id_dim_tipo
WHERE dti.tipo_entrada = 'Owner Surrender'
GROUP BY
  da.especie,
  f.idade_entrada
ORDER BY qtd_animais DESC
LIMIT 10
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [especie: string, idade_entrada: string ... 1 more field]

Table

	especie	idade_entrada	qtd_animais
1	Dog	1 year	1679
2	Dog	2 years	1374
3	Dog	3 years	796
4	Cat	1 year	713
5	Cat	1 month	634
6	Cat	2 months	565
7	Dog	4 years	526
8	Cat	2 years	507
9	Dog	5 years	469
10	Dog	1 month	469

10 rows

Os dados revelam que os animais entregues pelo tutor são cães e gatos que já passaram da fase de filhotes, concentrando-se nas idades de 1 e 2 anos.

b. Os animais entregues ao abrigo pelos donos apresentam alguma condição de saúde?

5

```
%sql
-- Os animais devolvidos, em sua maioria, estão normais.

SELECT
  dc.condicao_entrada,
  COUNT(*) AS qtd_animais
FROM fato_entrada f
JOIN dim_animal da
  ON f.id_dim_animal = da.id_dim_animal
JOIN dim_condicao dc
  ON f.id_dim_condicao = dc.id_dim_condicao
JOIN dim_tipo dti
  ON f.id_dim_tipo = dti.id_dim_tipo
WHERE dti.tipo_entrada = 'Owner Surrender'
GROUP BY
  dc.condicao_entrada
ORDER BY qtd_animais DESC
LIMIT 10
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [condicao_entrada: string, qtd_animais: long]

Table

	condicao_entrada	qtd_animais
1	Normal	13830
2	Sick	461
3	Injured	259
4	Nursing	110
5	Aged	67
6	Feral	9
7	Pregnant	8

7 rows

Os dados mostram que a maioria dos animais entregues estão em condições normais de saúde. Esses resultados sugerem que a transição da fase de filhote para a idade adulta jovem é um período de vulnerabilidade, no qual o tutor muitas vezes desiste da guarda devido a mudanças no comportamento ou porte do animal.

Pergunta 4: Qual é a distribuição dos principais desfechos dos animais?

6

```
%sql
-- Qual é a distribuição dos principais desfechos dos animais (adoção, retorno, transferência, eutanásia)?

SELECT
  dtis.tipo_saida,
  COUNT(*) AS qtd
FROM fato_saida fs
JOIN dim_tipo_saida dtis
  ON fs.id_dim_tipo_saida = dtis.id_dim_tipo_saida
GROUP BY dtis.tipo_saida
order by qtd DESC
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [tipo_saida: string, qtd: long]

Table

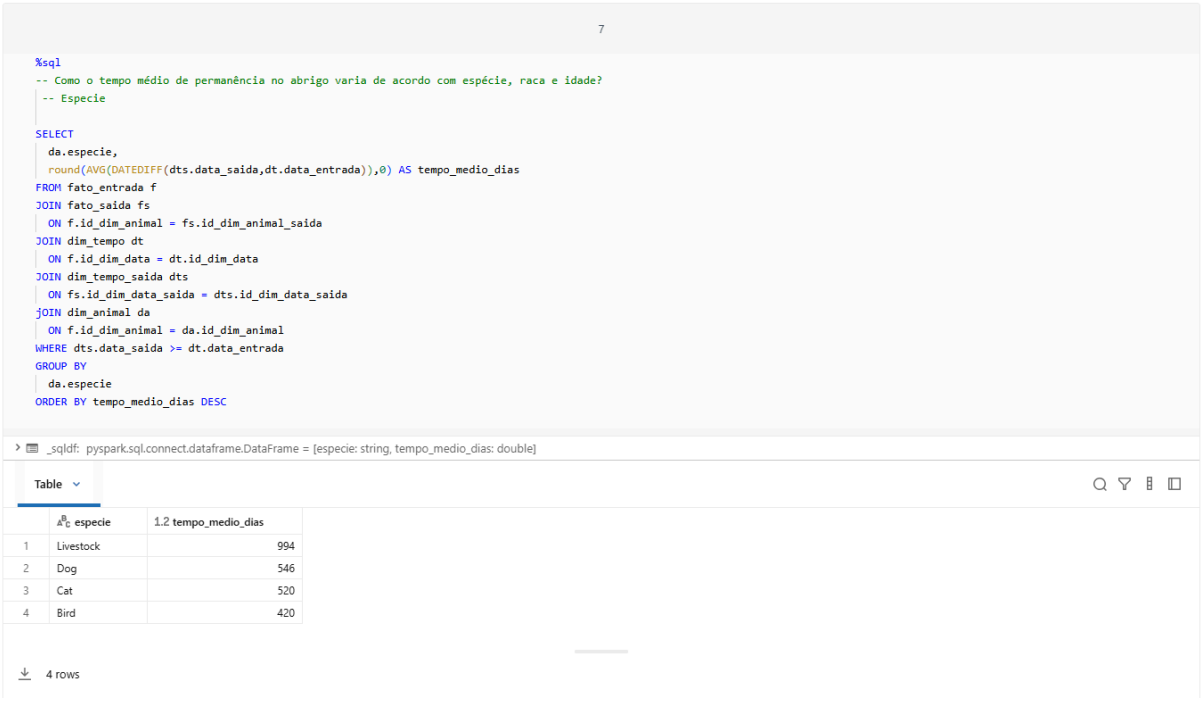
	tipo_saida	qtd
1	Adoption	33944
2	Transfer	20988
3	Return to Owner	14920
4	Euthanasia	2803
5	Died	480
6	Missing	45

6 rows

Resposta 4: A análise dos desfecho mostra que o fluxo de saídas é liderado pelas adoções.

Pergunta 5: Como o tempo médio de permanência no abrigo varia de acordo com espécie, raça e idade?

Espécie:



```
%sql
-- Como o tempo médio de permanência no abrigo varia de acordo com espécie, raca e idade?
  -- Especie

SELECT
  da.especie,
  round(AVG(DATEDIFF(dts.data_saida,dt.data_entrada)),0) AS tempo_medio_dias
FROM fato_entrada f
JOIN fato_saida fs
  ON f.id_dim_animal = fs.id_dim_animal_saida
JOIN dim_tempo dt
  ON f.id_dim_data = dt.id_dim_data
JOIN dim_tempo_saida dts
  ON fs.id_dim_data_saida = dts.id_dim_data_saida
JOIN dim_animal da
  ON f.id_dim_animal = da.id_dim_animal
WHERE dts.data_saida >= dt.data_entrada
GROUP BY
  da.especie
ORDER BY tempo_medio_dias DESC
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [especie: string, tempo_medio_dias: double]

Table

	especie	tempo_medio_dias
1	Livestock	994
2	Dog	546
3	Cat	520
4	Bird	420

4 rows

Resposta 5 (espécie): O tempo médio de permanência varia significativamente entre as espécies. Os animais da categoria de gado “Livestock” apresentam o maior tempo de estadia. Os animais domésticos como cães e gatos apresentam uma rotatividade maior, porém as aves são as espécies que apresentam o menor tempo de permanência nos abrigos. Estes dados sugerem que animais de grande porte ou de produção enfrentam barreiras muito maiores para a adoção ou transferência em comparação com animais de companhia.

Raça: Para a análise de raça, foram realizadas consultas específicas para cada espécie. Todas essas consultas estão disponíveis no material disponibilizado no GitHub. Aqui, serão analisados os resultados referentes às raças de cães e gatos, por serem os mais relevantes para as perguntas do trabalho.

Cães:

```
9
%sql
-- Como o tempo médio de permanência no abrigo varia de acordo com espécie, raca e idade?
 -- Raca
SELECT
  da.raca,
  round(AVG(DATEDIFF(dts.data_saida,dt.data_entrada)),0) AS tempo_medio_dias
FROM fato_entrada f
JOIN fato_saida fs
  ON f.id_dim_animal = fs.id_dim_animal_saida
JOIN dim_tempo dt
  ON f.id_dim_data = dt.id_dim_data
JOIN dim_tempo_saida dts
  ON fs.id_dim_data_saida = dts.id_dim_data_saida
jOIN dim_animal da
  ON f.id_dim_animal = da.id_dim_animal
WHERE dts.data_saida >= dt.data_entrada
AND da.especie = 'Dog'
GROUP BY
  da.raca
ORDER BY tempo_medio_dias DESC
LIMIT 5
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [raca: string, tempo_medio_dias: double]

Table

	raca	tempo_medio_dias
1	Border Collie/American Staffordshire Terrier	1590
2	Shetland Sheepdog/Italian Greyhound	1569
3	Border Collie/Anatol Shepherd	1562
4	Australian Kelpie/Shiba Inu	1539
5	Plott Hound/Carolina Dog	1515

5 rows

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

Gatos:

```
10
%sql
-- Como o tempo médio de permanência no abrigo varia de acordo com espécie, raca e idade?
 -- Raca
SELECT
  da.raca,
  round(AVG(DATEDIFF(dts.data_saida,dt.data_entrada)),0) AS tempo_medio_dias
FROM fato_entrada f
JOIN fato_saida fs
  ON f.id_dim_animal = fs.id_dim_animal_saida
JOIN dim_tempo dt
  ON f.id_dim_data = dt.id_dim_data
JOIN dim_tempo_saida dts
  ON fs.id_dim_data_saida = dts.id_dim_data_saida
jOIN dim_animal da
  ON f.id_dim_animal = da.id_dim_animal
WHERE dts.data_saida >= dt.data_entrada
AND da.especie = 'Cat'
GROUP BY
  da.raca
ORDER BY tempo_medio_dias DESC
LIMIT 5
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [raca: string, tempo_medio_dias: double]

Table

	raca	tempo_medio_dias
1	Devon Rex	1253
2	Snowshoe/Ragdoll	1053
3	American Wirehair ...	1042
4	Bengal	995
5	Bengal Mix	944

5 rows

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

Resposta 5 (raça): A coluna raça apresenta um grande variedade de categorias devido a mistura de raças, ultrapassando mil classificações, o que dificulta a obtenção de conclusões claras e assertivas a partir desses dados. Ainda sim, de forma geral, pode-se dizer que raças mais exigentes em questão de tamanho, energia e comportamento, tendem a permanecer mais tempo nos abrigos.

Idade:

12

```
%sql
-- Como o tempo médio de permanência no abrigo varia de acordo com espécie, raca e idade?
-- Animais muito jovens não podem sair do abrigo pois precisam de cuidados especiais até atingir certa idade para adoção.
-- Animais mais velhos tedem a ficar mais tempo nos abrigos sem serem adotados.

SELECT
  f.idade_entrada,
  round(AVG(DATEDIFF(dts.data_saida,dt.data_entrada)),0) AS tempo_medio_dias
FROM fato_entrada f
JOIN fato_saida fs
  ON f.id_dim_animal = fs.id_dim_animal_saida
JOIN dim_tempo dt
  ON f.id_dim_data = dt.id_dim_data
JOIN dim_tempo_saida dts
  ON fs.id_dim_data_saida = dts.id_dim_data_saida
jOIN dim_animal da
  ON f.id_dim_animal = da.id_dim_animal
WHERE dts.data_saida >= dt.data_entrada
GROUP BY
  f.idade_entrada
ORDER BY tempo_medio_dias DESC
LIMIT 10
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [idade_entrada: string, tempo_medio_dias: double]

Table

	idade_entrada	tempo_medio_dias
1	6 days	652
2	22 years	589
3	9 months	587
4	20 years	586
5	14 years	583
6	11 months	567
7	3 years	559
8	7 years	557
9	13 years	557
10	15 years	555

Resposta 5 (idade): Para essa análise, será desconsiderado o primeiro registro (6 days). Esses casos contam com uma barreira biológica e legal. Filhotes precisam ter no mínimo 45 a 60 dias para estarem disponíveis para adoção. É o período necessário para o desmame, socialização primária e cronogramas de vacinação.

Os dados mostram como cenário crítico os animais idosos, que enfrentam longas estadias devido a falta de procura e interesse dos adotantes. Adotantes tendem a evitar animais seniores devido à menor expectativa de vida e à possibilidade de gastos médicos elevados, o que faz com que esses animais ocupem as vagas do abrigo por períodos muito extensos até o fim de suas vidas ou uma adoção humanitária.

Pergunta 6: Existem padrões sazonais no volume de entradas e adoções ao longo do ano?

Entradas:

13

```
%sql

-- Existem padrões sazonais no volume de entradas e adoções ao longo do ano?
--- Entradas
SELECT
  MONTH(dt.data_entrada) AS mes_entrada,
  COUNT(*) AS qtd_animais
FROM fato_entrada f
JOIN dim_tempo dt
  ON f.id_dim_data = dt.id_dim_data
GROUP BY
  MONTH(dt.data_entrada)
ORDER BY
  qtd_animais DESC
LIMIT 3;
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [mes_entrada: integer, qtd_animais: long]

Table

	mes_entrada	qtd_animais
1	5	7141
2	10	6991
3	6	6701

3 rows

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

Resposta 6 (entrada): Principais meses são maio, junho e outubro. Os meses de maio e junho são caracterizados pelo “Kitten Season”, que é um momento de aumento da ninhada de gatos. Já o mês de outubro está relacionado com mudanças comportamentais como o abandono pós verão (animais adotados impulsivamente no período de férias são abandonados quando a rotina volta ao normal).

Saída:

14

```
%sql
-- Existem padrões sazonais no volume de entradas e adoções ao longo do ano?
--- Saídas: Datas especiais (dia das crianças, ferias e natal)
SELECT
  MONTH(dts.data_saida) AS mes_saida,
  COUNT(*) AS qtd_animais
FROM fato_saida fs
JOIN dim_tempo_saida dts
  ON fs.id_dim_data_saida = dts.id_dim_data_saida
GROUP BY
  MONTH(dts.data_saida)
ORDER BY
  qtd_animais DESC
LIMIT 3;
```

_sqldf: pyspark.sql.connect.dataframe.DataFrame = [mes_saida: integer, qtd_animais: long]

Table

	mes_saida	qtd_animais
1	10	7107
2	7	6835
3	12	6630

3 rows

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

Resposta 6 (saída): Principais meses de saída são outubro, julho e dezembro. São meses caracterizados por datas especiais, férias e final de ano.

A partir das perguntas secundárias e análise dos dados obtidos foi possível responder a pergunta principal do trabalho:

> **Quais fatores influenciam no tempo de permanência e no desfecho dos animais do abrigo?**

Os resultados indicam que o tempo de permanência e o desfecho dos animais são influenciados por suas características individuais (espécie, raça e idade), além de fatores comportamentais dos adotantes e de variações sazonais.

Fatores determinantes:

1. Idade do animal: Animais jovens apresentam maior rotatividade, enquanto animais idosos ou de maior porte tendem a permanecer por períodos prolongados no abrigo.
2. Espécie e porte: Animais domésticos (cães e gatos) possuem maior rotatividade. No entanto, animais de grande porte (Livestock) apresentam alto tempo de permanência no abrigo.
3. Sazonalidade: Picos de saída em meses de férias e festas (dezembro/julho) indicam que a adoção muitas vezes é vista como um evento de ocasião, o que pode ser um risco para abandonos futuros.
4. Comportamento: A entrega do animal pelo tutor aponta falta de preparo e acompanhamento do tutor ao longo do ciclo de vida do animal.

Para otimizar o tempo de permanência e garantir desfechos positivos (adoções responsáveis), sugerem-se as seguintes ações:

1. Entrevistas de Perfil no processo de adoção: Entender sobre a rotina, estilo de vida, espaço disponível e expectativas, visando reduzir devoluções futuras.
2. Campanhas Sazonais: Desenvolvimento de campanhas de adoção e conscientização direcionadas aos períodos sazonais mais críticos ou favoráveis, conforme os padrões observados nos dados.
3. Campanhas de adoção focadas em animais idosos: Criar incentivos específicos para animais idosos, como parcerias com clínicas veterinárias, planos de saúde pet ou auxílio em itens de necessidades especiais, combatendo a barreira do custo médico.
4. Parceria com adestradores: Sabendo que a entrega voluntária dos animais ocorre na fase jovem-adulta, o oferecimento de suporte informativo ou parcerias com adestradores seriam uma forma de reter tutores que estejam pensando em desistir do animal por problemas comportamentais.

Em resumo, a diminuição do tempo de permanência e o desfecho favorável para o animal dependem, principalmente, da mudança na relação entre sociedade e adoção. Para garantir adoções responsáveis e sustentáveis, é importante tomar escolhas conscientes, baseadas na compatibilidade entre o perfil do adotante e as necessidades do animal ao longo de seu ciclo de vida.

Autoavaliação

Antes de iniciar a Pós-Graduação em Engenharia de Dados, eu não possuía experiência prévia aprofundada com linguagens de programação. Assim, o desenvolvimento deste projeto em SQL representou uma novidade e, ao mesmo tempo, um desafio. As aulas contribuíram para a construção de uma base sólida de entendimento dos conceitos, porém a utilização do Databricks exigiu um período de adaptação, uma vez que cada plataforma possui suas particularidades, diferentes daquelas utilizadas durante as aulas. Um exemplo disso é a sensibilidade a maiúsculas e minúsculas na escrita dos comandos.

Além disso, este foi o primeiro contato com a criação e organização de um repositório no GitHub. Houve dificuldade inicial na integração entre o GitHub e o Databricks, especialmente no vínculo entre os notebooks e seus respectivos outputs. No entanto, com o auxílio do grupo no Discord, foi possível superar esse desafio.

Desenvolver este trabalho foi uma oportunidade de aplicar, de forma prática, os conceitos estudados ao longo da disciplina, especialmente no que se refere à modelagem de dados, à análise exploratória e à interpretação de resultados, a partir de um tema de grande relevância pessoal: a causa animal. A participação em ONGs e campanhas de adoção reforçou o interesse em compreender melhor os dados e os padrões identificados, possibilitando a proposição de ações mais assertivas e voltadas ao bem-estar dos animais.

Embora o uso de uma base de dados brasileira fosse desejável, a indisponibilidade de dados públicos com abrangência e qualidade semelhantes levou à utilização da base disponibilizada pelo Kaggle. De modo geral, o projeto contribuiu significativamente para o meu desenvolvimento técnico e reforçou o interesse em aprofundar os estudos na área de engenharia e análise de dados.