Natasha Azza & Veronica Benzi

Pet Expense Management

A primary pain point within the pet industry is poor data quality and information overload: pet owners are not equipped to record pet data in a structured manner and also do not understand which information is critical. Paw á Peau is attempting to address this problem by building an app to deliver actionable insights from pet data and improve the overall quality of pet care. The challenge of this data science project is thus to translate the benefits of tracking information into financial terms, in order to motivate pet owners.

Objectives

In Germany, pet ownership has increased by over 30% since the pandemic and currently lies at 15.7 million cats and 10.7 million dogs (2020, Zenz-Spitzenweg). Dogs are the second most popular pets, but dog owners spend far more than any other category of pet owner and thus will be the focus of this study.

The project aims to:

- Determine the real cost of dog ownership in Germany.
- Calculate ongoing expenses to predict the total costs over an individual dog's lifespan.
- Explore KPIs that could be used as early warning indicators for avoidable expenses.

Colab - Dataset visualization

Miro - Dataset schema and project milestones

Data enrichment - notes, figures sourced from German studies

Data audit - analysis of variables

Context

Since the pandemic, pet ownership has skyrocketed and all segments of the pet industry (from retail to pet services and emerging technologies) are currently booming in the United States. Europe is following closely behind, with the UK, Germany and France as the biggest and fastest growing markets. Germany is the EU’s largest economy but has been hit hard by rising costs stemming from energy prices, inflation, and financial side effects from ongoing regional wars. Two steep increases in the national veterinary tariffs within the last three mean pet owners (47% of Germany’s population) are under financial pressure.

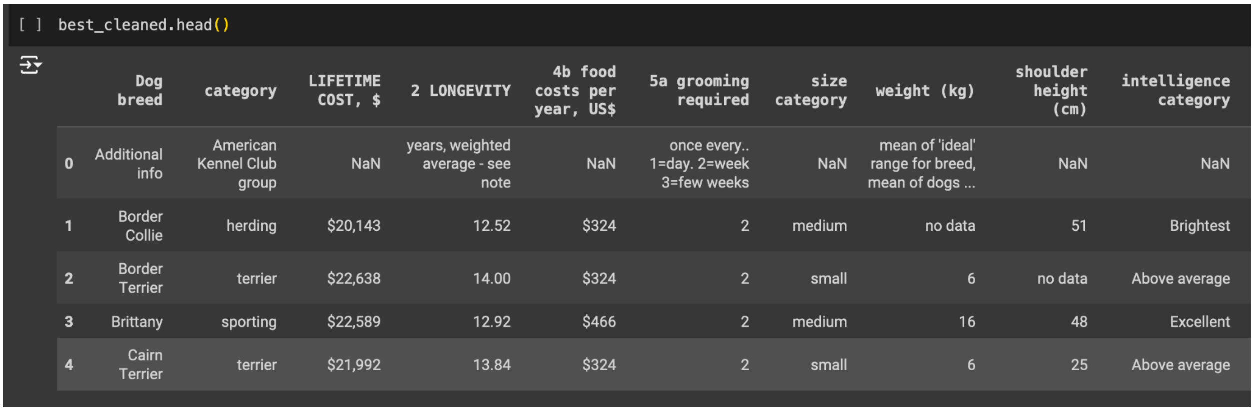
Dog ownership expenses can be classified into the following categories:

- Food
- Accessories
- Medical
- Grooming
- Training
- Funerary
- Tax (particular to Germany)
- Insurance
- Pet sitting

The costs vary for each individual dog and depend on a variety of factors such as breed, size, age etc. This is reflected in the fact that larger pets consume more, are more destructive, cost more to treat and are faced with higher insurance premiums.

Foundation Dataset

The foundation dataset “Best In Show” is open source and available on Kaggle. It was chosen due to the vast number of variables (69) and the inclusion of cost data (albeit related to the US market).



```
[ ] best_cleaned.head()
```

	Dog breed	category	LIFETIME COST, $	2 LONGEVITY	4b food costs per year, US$	5a grooming required	size category	weight (kg)	shoulder height (cm)	intelligence category
0	Additional info	American Kennel Club group	NaN	years, weighted average - see note	NaN	once every.. 1=day. 2=week 3=few weeks	NaN	mean of 'ideal' range for breed, mean of dogs ...	NaN	NaN
1	Border Collie	herding	$20,143	12.52	$324	2	medium	no data	51	Brightest
2	Border Terrier	terrier	$22,638	14.00	$324	2	small	6	no data	Above average
3	Brittany	sporting	$22,589	12.92	$466	2	medium	16	48	Excellent
4	Cairn Terrier	terrier	$21,992	13.84	$324	2	small	6	25	Above average

Further datasets were chosen on the basis of complementary variables which could enhance the relationships between variables in our foundation dataset. Variable connections are highlighted in a preliminary schema of the selected variables from each dataset (see cover link to Miro board).

The biggest challenge is combining data from many sources and deciding which variables to keep or drop. This is due to variance in data quality and data availability. For example, longevity data is abundant with 440k rows. Bite statistics, however, is limited although the data is useful as an enhancement. After cleaning the data and first explorations, we shall finalize the selection.

Pet sitting expenses will not be included in the overall cost calculation due to the irregularity of this variable. Not all dog owners use professional boarding kennels and instead may opt to use private services or the help of friends and family, meaning costs may be non-existent (and at best unpredictable). Furthermore, the time spent away from pets varies from one household to another.

Data Limitations

Two new columns were added: veterinary costs and training probability. Veterinary costs make up a large amount of pet owner expenses due to the low incidence of insurance (±5%), however these costs were not included in our foundation data set. As there is an abundance of recent economic studies on pet ownership in Germany, financial data from the results of the Ohr and Zenz-Spitzweg studies will be used to enrich our dataset.

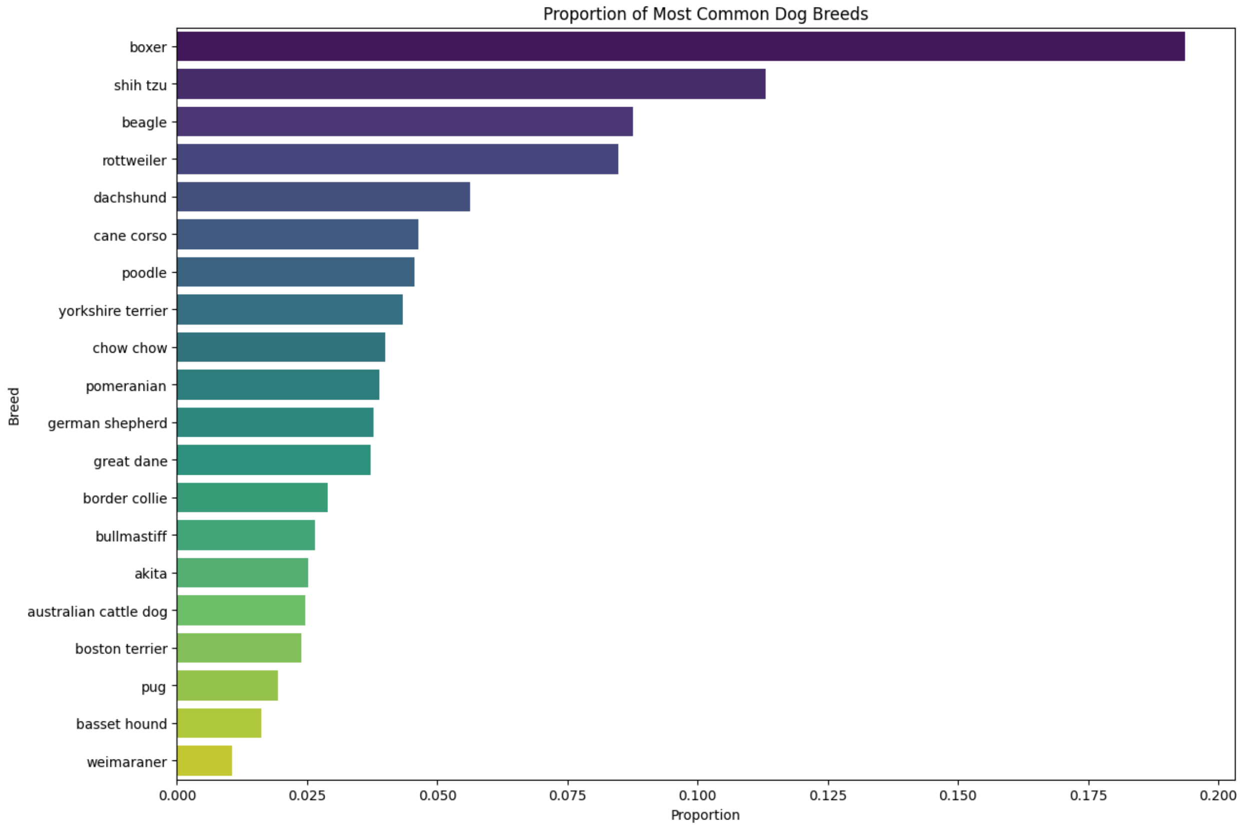
Training course costs vary greatly by institute and course type, however the hourly fee range of trainers is readily available online. The probability that a dog needs training can be extracted from the dataset through variable correlation (breed, intelligence, size, age etc.). We would thus attempt to predict training needs as a score. Depending on available data, it may be possible to also predict real costs based on the probability that a dog needs training, average trainer fee, and average duration of training according to lifestage or behavioural problem.

Retail financial data is readily available for the US and UK markets, but not for continental Europe. It may be necessary to approximate retail expenses based on the differences in cost of living, where local data is not readily available. Some of the data is provided with a calculation based on weekly, monthly, or yearly expenses. These sums will need to be normalized in order to properly explore the data.

Grooming costs depend on both size and coat type. Although able to be calculated based on local groomer fees, pet owners often forgo professional services in favour of a cut at home or simply letting the dog grow its coat out. As such, grooming (if kept in the final selection) should rather be converted into a categorical variable.

Dataset Exploration

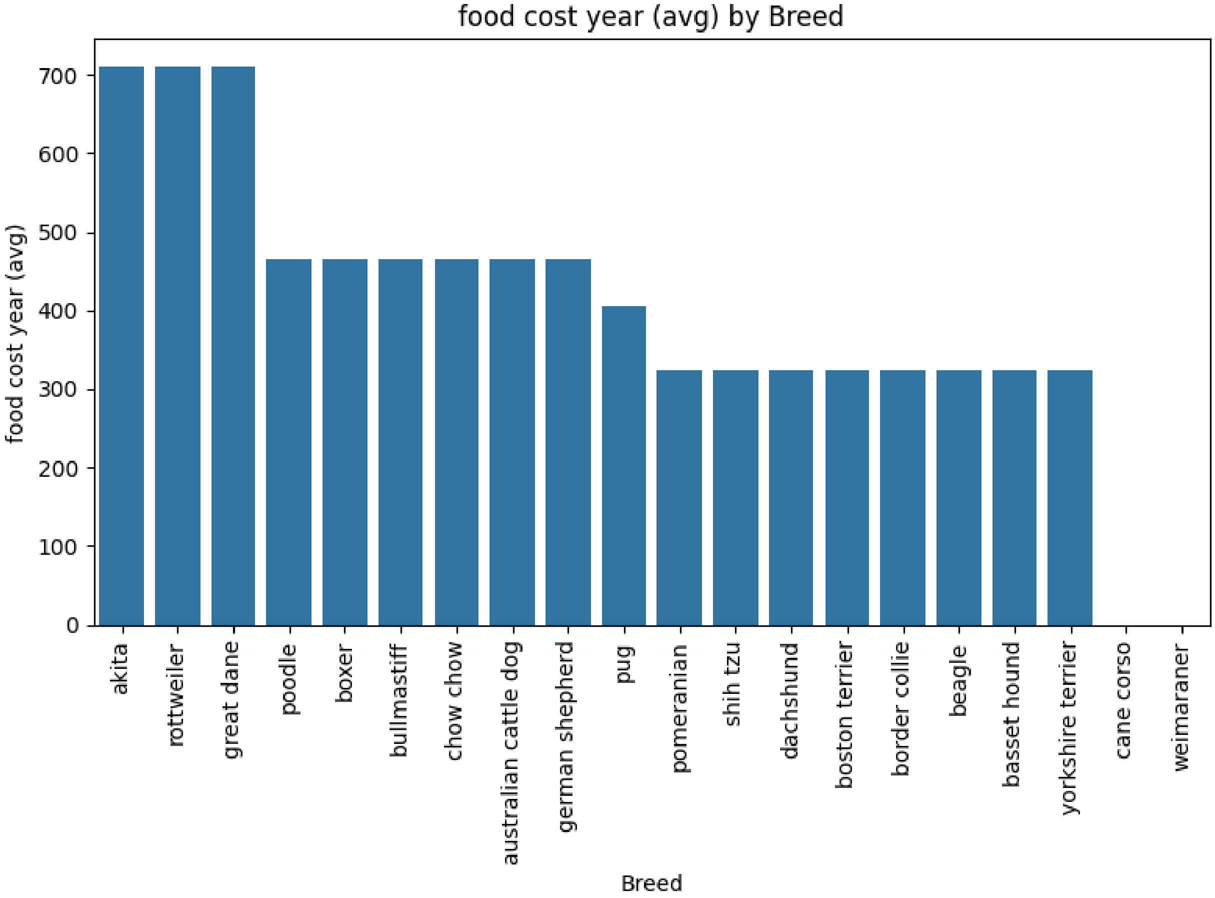
Some breeds had to be excluded from the final dataset due to a lack of data. We made sure to include the most popular breeds in Germany by combining additional health-focused datasets. To account for dogs that do not fit within this selection, we created three new rows to represent mixed breeds of the three different size categories, and used the mean and mode data to fill out each variable.



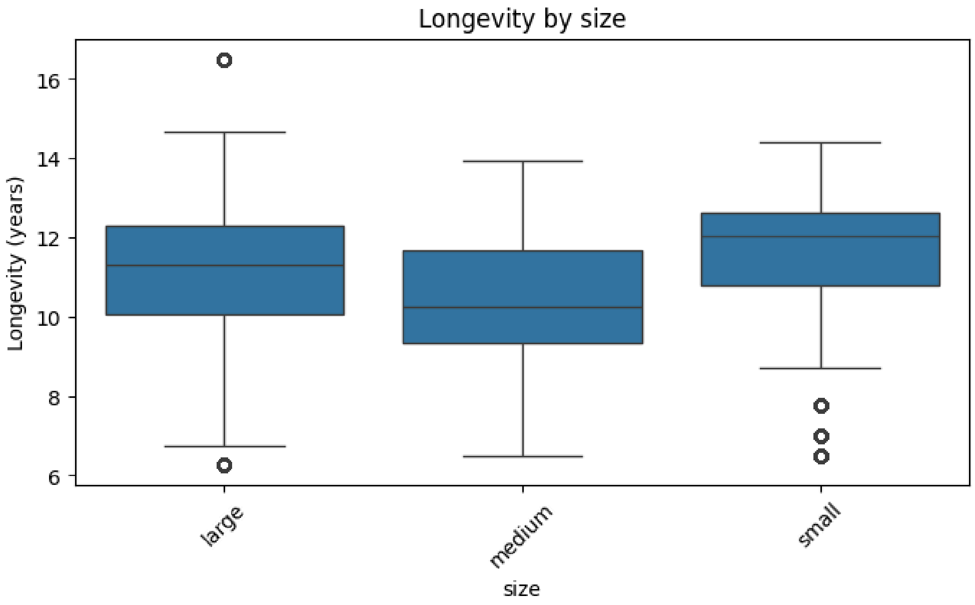
After joining multiple data sets, however, it became clear that we had unwittingly created a strong bias for certain breeds due to their overrepresentation in some of the health datasets we used.

At this point, our dataset was extremely large at over 1 million rows. We thus decided to limit the number of dogs that could be represented by any one breed to 1500, which were chosen at random. As there are many dog breeds in existence, even this limitation still left us with approximately 200 000 rows in the final dataset.

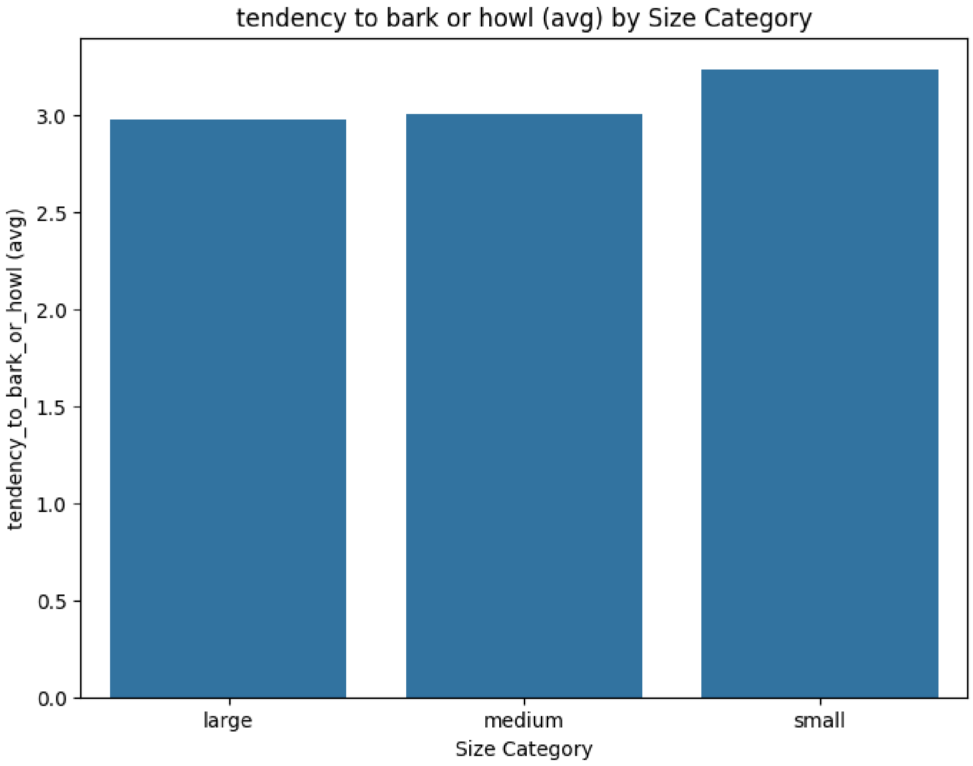
Factors which greatly influence cost include longevity (long life means more accumulated costs), breed (predisposition for certain health and behavioural traits) and size (directly related to the volume of food intake). The existing calculation within the dataset of food costs by dog size alone may not be accurate due to certain breeds having different activity levels.



We discovered that the dataset needed extensive cleaning following exploration of correlation between longevity and size. It is well documented in scientific literature that smaller dogs tend to live longer than larger dogs. This, however, was not represented in the graph and there was no plausible explanation as for why medium sized dogs would be shorter lived than larger breeds.



Another interesting correlation to explore is different temperament traits which correlate with the tendency to bark or howl. Noise and destruction within the home are some of the main reasons why dog owners may seek out the help of a trainer, or failing that, abandon their pet. Behaviour-related variables are thus important to understand when it comes to the potential lifetime costs of owning a dog.



Variable Selection

Pre-cleaning Selection

Our pre-cleaning variable shortlist:
'Breed', 'lifetime_cost', 'longevity', 'food_cost_year', 'grooming required ', 'size_category', 'weight_kg', 'shoulder_height_cm', 'intelligence_category', 'avg_food_per_week_£', 'food_per_week_$', 'genetic_ailments', 'category', 'Specie', 'Gender', 'where_bitten', 'sensitivity_level', 'tolerates_being_alone', 'tolerates_cold_weather', 'tolerates_hot_weather', 'incredibly_kifriendly_dogs', 'dog_friendly', 'friendly_towarstrangers', 'easy_to_groom', 'potential_for_mouthiness', 'prey_drive', 'tendency_to_bark_or_howl', 'wanderlust_potential', 'exercise_needs', 'energy_level', 'adapts_well_to_apartment_living'

We began with many variables and datasets due to the complexity of the pricing data involved in pet ownership. Food, accessory, and veterinary costs were the only readily available data (the latter, as part of a study and not from a dataset). The primary target variable was lifetime_cost of the individual dog. We aimed to generate multipliers for training probability, insurance premiums, and other variables to the available mean data. We were not yet certain whether the cost calculator will be limited to dogs only, or also include cats (for which data is abundant, and there is little variance between breeds).

There was a correlation between certain behavioural elements, as well as between size and breed, size and longevity, and size and cost. What we were uncertain of, and what required further exploration was the extent to which these variables were related and could be used to predict lifetime cost. Another factor to take into account is whether we want to only predict the cost over the entire lifetime of a dog with certain characteristics (such as breed, size, sex etc.), or also predict the remaining costs based on the current age of the dog (and multiplied by other factors).

Post-cleaning

New cost variables were added, including veterinary_cost, insurance_cost, tax_cost, ailment_cost, training_probability etc. due to the inaccuracy of predicting lifetime cost based solely on food and accessory expenditure.

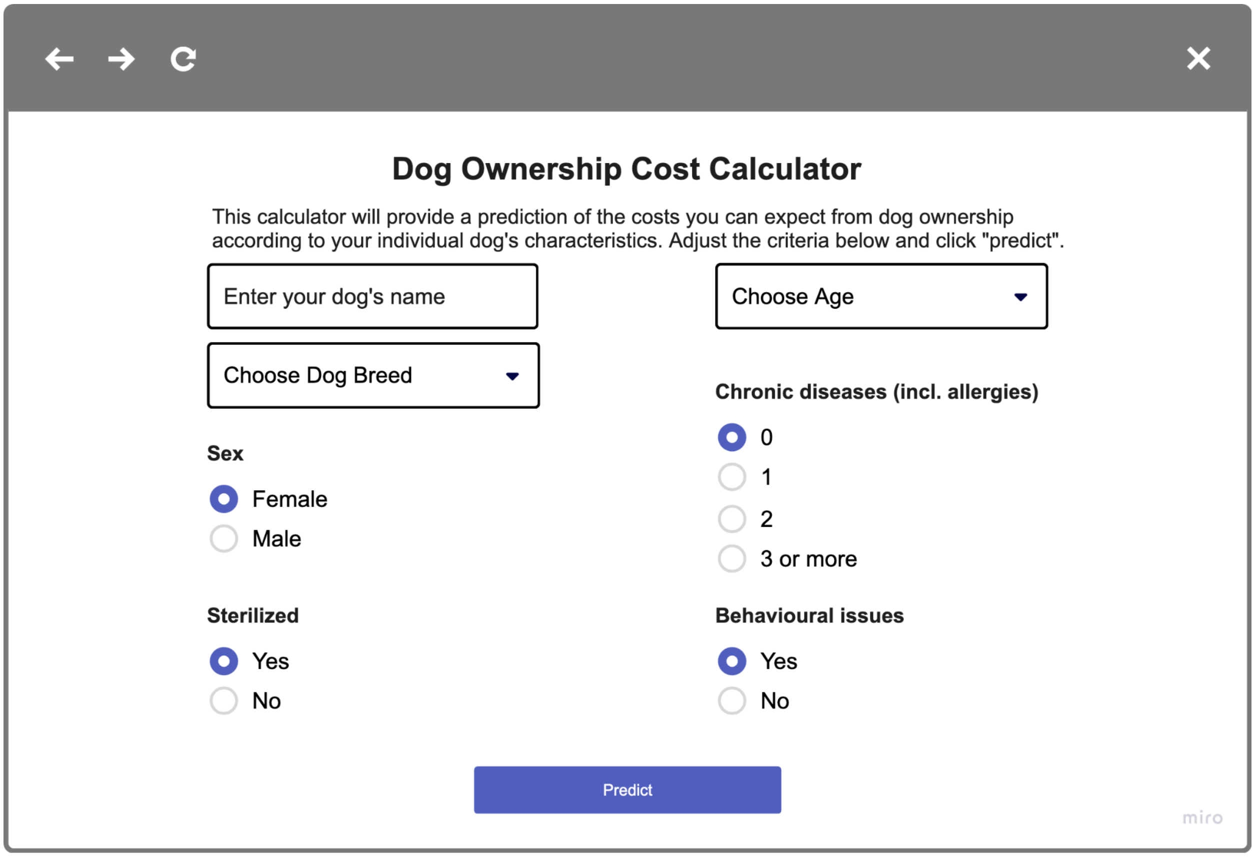
The shelter datasets were dropped due to an insufficient number of rows (it made no sense to combine datasets with such discrepancy). Nevertheless, we have kept the numerous datasets with the goal of using them to enrich or verify the data in our foundation dataset. We also dropped the following superfluous variables: 'weight_kg', 'shoulder_height_cm', 'avg_food_per_week_£', 'food_per_week_$', 'Specie', 'where_bitten', 'adaptability'

Leaving:
'Breed', 'breed_group', 'longevity', 'size_category', 'intelligence_category', 'genetic_ailments', ‘neutered’, 'sex', 'grooming required ', 'food_cost_year', 'lifetime_cost', 'sensitivity_level', 'tolerates_being_alone', 'tolerates_cold_weather', 'tolerates_hot_weather', 'incredibly_kifriendly_dogs', 'dog_friendly', 'friendly_towarstrangers', 'potential_for_mouthiness', 'prey_drive', 'tendency_to_bark_or_howl', 'wanderlust_potential', 'exercise_needs', 'energy_level', 'adapts_well_to_apartment_living', ‘bite_statistics’

(Variable Index - Data Audit - see detailed description of variables here)

Data Concept

A significant amount of transformation and normalization will need to be done before modeling can begin. This has become clear during the cleaning and variable selection process. We had begun with the idea of doing regression modelling to predict total lifetime costs of owning a dog with certain characteristics.



Dog Ownership Cost Calculator

This calculator will provide a prediction of the costs you can expect from dog ownership according to your individual dog's characteristics. Adjust the criteria below and click "predict".

Enter your dog's name

Choose Dog Breed ▾

Choose Age ▾

Sex

● Female
○ Male

Sterilized

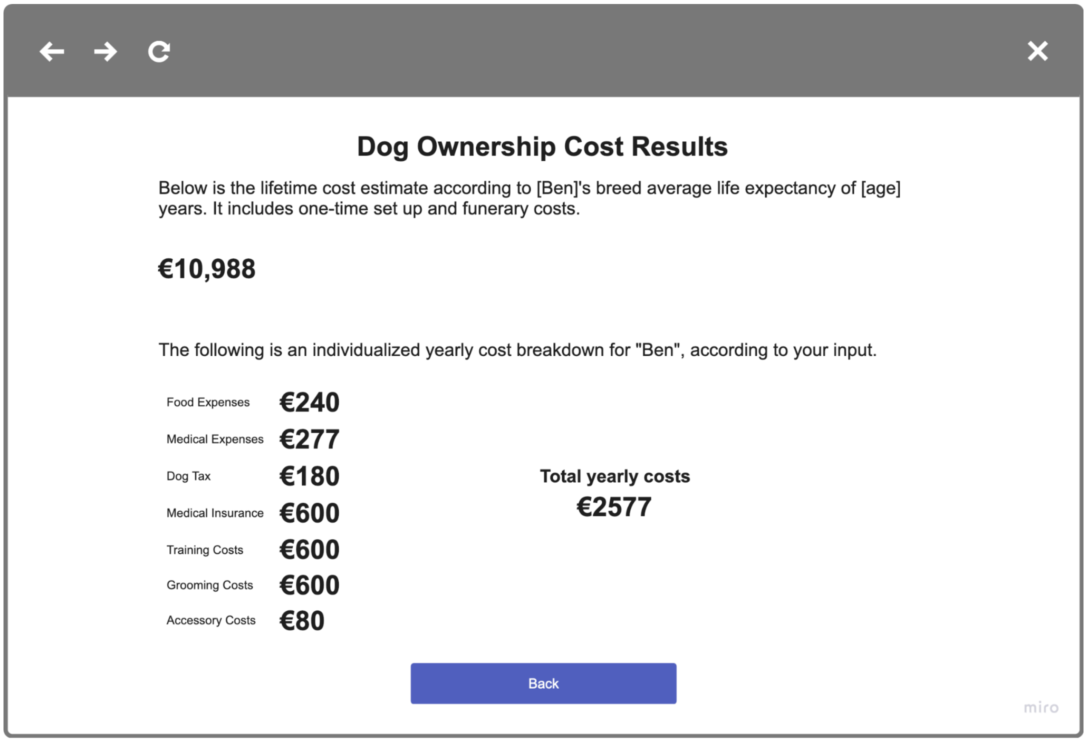
● Yes
○ No

Chronic diseases (incl. allergies)

● 0
○ 1
○ 2
○ 3 or more

Behavioural issues

● Yes
○ No



Dog Ownership Cost Results

Below is the lifetime cost estimate according to [Ben]'s breed average life expectancy of [age] years. It includes one-time set up and funerary costs.

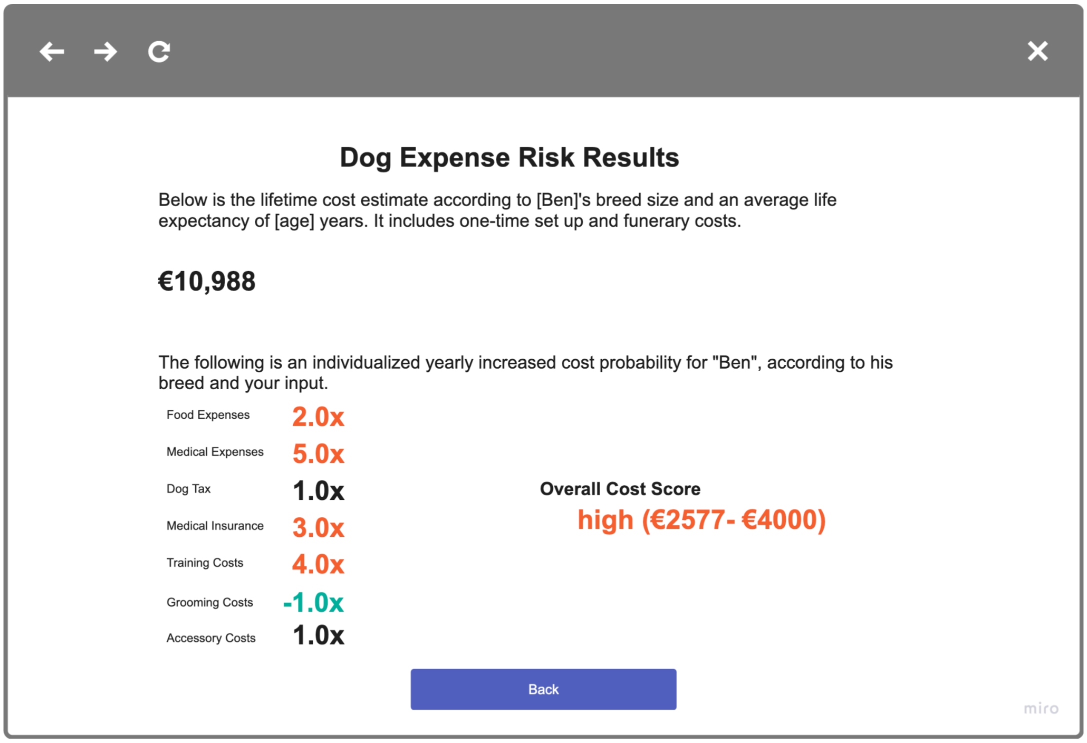
€10,988

The following is an individualized yearly cost breakdown for "Ben", according to your input.

Food Expenses	€240
Medical Expenses	€277
Dog Tax	€180
Medical Insurance	€600
Training Costs	€600
Grooming Costs	€600
Accessory Costs	€80

Total yearly costs
€2577

However, as it became clear that we had extensive cleaning and manual enrichment to do, it made more sense to change our strategy to a classification model. We will thus still roughly calculate lifetime costs, however we will present the user with an expense category and cost ranges.



Dog Expense Risk Results

Below is the lifetime cost estimate according to [Ben]'s breed size and an average life expectancy of [age] years. It includes one-time set up and funerary costs.

€10,988

The following is an individualized yearly increased cost probability for "Ben", according to his breed and your input.

Food Expenses	2.0x
Medical Expenses	5.0x
Dog Tax	1.0x
Medical Insurance	3.0x
Training Costs	4.0x
Grooming Costs	-1.0x
Accessory Costs	1.0x

Overall Cost Score
high (€2577- €4000)

References

Literature

- Digital Pet Care Industry: European Market, focus on Germany, France & Bulgaria (Paw ả Peau, 2022)
- Increasing adoption rates at animal shelters: a two-phase approach to predict length of stay and optimal shelter allocation (Bradley & Rajendran, 2021)
- Better animal welfare through insurance (Zenz-Spitzweg, 2021)
- The Economic Significance of Pet Ownership in Germany (Ohr, 2019)
- The Economic Significance of Pet Ownership in Germany (Ohr, 2014)

Primary Datasets

https://www.kaggle.com/datasets/paultimothymooney/best-in-show-data-about-dogs
(foundation)
https://www.kaggle.com/datasets/jainaru/dog-breeds-ranking-best-to-worst
https://www.kaggle.com/datasets/iqramharisfahromi/pet-sales
https://www.kaggle.com/datasets/jahangirraina/pet-food-customer-orders-online
https://data.cityofnewyork.us/Health/DOHMH-Dog-Bite-Data/rsgh-akpg/about_data
https://www.kaggle.com/datasets/rtatman/animal-bites
https://www.kaggle.com/datasets/agarwalyashhh/dog-adaptability?select=training_data.csv
https://www.kaggle.com/datasets/jackdaoud/animal-shelter-analytics
https://data.bloomington.in.gov/Public-Works/Animal-Shelter-Animals/e245-r9ub/about_data

Supplementary Datasets

(used for enrichment, along with the German pet ownership studies)

https://rvc-repository.worktribe.com/output/1375552
https://data.world/the-pudding/adoptable-dogs-on-petfinder-in-the-us/workspace/file?filename=allDogDescriptions.csv
https://rvc-repository.worktribe.com/output/1375689
https://rvc-repository.worktribe.com/output/1558210
Pet health pricing model