DECLARATION: I understand that this is an **individual** assessment, and that collaboration is not permitted. I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar. I understand that by returning this declaration with my work, I am agreeing with the above statement.

Please find the interactive visualisation here, the code, data, processing file etc at my GitHub repository here, and the video here.

Tools Used: HTML, CSS, JS, D3.js *(version 7)*, a library for in-browser D3 v7 to simplify colour scale legend creation as D3 does not provide legend functions by default and python3 (pandas, numpy, matplotlib, jupyter notebook) for data processing and exploratory data analysis.

1 Dataset

The dataset visualized, comprises 32 attributes (columns) related to 105 dog breeds (rows) regarding their physicality, behaviour and AKC popularity rankings over the years 2013 to 2020. The dataset is an amalgamation of data from 4 separate datasets (csv files) obtained from 3 sources [1] [2] [3].

Attribute Types: Attribute types were identified as part of Exploratory Data Analysis as in the image below. There are 28 categorical attributes (25 with order, 3 without order) and 4 quantitative ones (all continuous with type measurement). Meaning of attributes were gathered from descriptions found at sources. Categories were assigned to each attribute to organize them meaningfully for visualization.

Attribute	*Type*	*Meaning*	*Category*
breed	Categorical (No Order)	Dog breed name.	Key
coat_type	Categorical (No Order)	Kind of fur coat this breed has.	Physical Trait
image	Categorical (No Order)	Image depicting what this breed typically looks like.	Physical Trait
rank_2013	Categorical (Has Order)	Popularity rank assigned to this breed by AKC in the year 2013.	Popularity
rank_2014	Categorical (Has Order)	Popularity rank assigned to this breed by AKC in the year 2014.	Popularity
rank_2015	Categorical (Has Order)	Popularity rank assigned to this breed by AKC in the year 2015.	Popularity
rank_2016	Categorical (Has Order)	Popularity rank assigned to this breed by AKC in the year 2016.	Popularity
rank_2017	Categorical (Has Order)	Popularity rank assigned to this breed by AKC in the year 2017.	Popularity
rank_2018	Categorical (Has Order)	Popularity rank assigned to this breed by AKC in the year 2018.	Popularity
rank_2019	Categorical (Has Order)	Popularity rank assigned to this breed by AKC in the year 2019.	Popularity
rank_2020	Categorical (Has Order)	Popularity rank assigned to this breed by AKC in the year 2020.	Popularity
affection_level	Categorical (Has Order)	How affectionate the breed is towards families on a scale of 1 to 5.	Home Suitability
child_friendliness	Categorical (Has Order)	How suitable the breed is for small kids on a scale of 1 to 5.	Home Suitability
dog_friendliness	Categorical (Has Order)	How tolerant the breed is of other dogs on a scale of 1 to 5.	Home Suitability
playfulness	Categorical (Has Order)	How playful the breed is on a scale of 1 to 5.	Home Suitability
adaptability	Categorical (Has Order)	How adaptable to new situations this breed is on a scale of 1 to 5.	Home Suitability
trainability	Categorical (Has Order)	How easy it is to train this breed is on a scale of 1 to 5.	Training Related
working_intelligence_level	Categorical (Has Order)	How good of a working dog this breed is.	Training Related
coat_length	Categorical (Has Order)	Length that the coat of this breed can grow to.	Physical Trait
size *	Categorical (Has Order)	The size category of this breed (teacup, toy, small, medium, large, giant).	Physical Trait
shedding_level	Categorical (Has Order)	Level of shedding on a scale of 1 to 5.	Care Need
grooming_frequency	Categorical (Has Order)	How frequently the breed requires grooming on a scale of 1 to 5.	Care Need
drooling_level	Categorical (Has Order)	How much the breed drools on a scale of 1 to 5.	Care Need
energy_level	Categorical (Has Order)	How energetic this breed is on a scale of 1 to 5.	Care Need
mental_stimulation_needs	Categorical (Has Order)	How much mental stimulation this breed requires on a scale of 1 to 5.	Care Need
openness_to_strangers	Categorical (Has Order)	How welcoming the breed is on a scale of 1 to 5.	Protection Related
protectiveness	Categorical (Has Order)	How good of a watchdog this breed is on a scale of 1 to 5.	Protection Related
barking_level	Categorical (Has Order)	How much this breed barks on a scale of 1 to 5.	Protection Related
reps_avg *	Quantitative Continuous Measurement	Average no. of repetitions required for this breed to understand new commands.	Training Related
obedience_prob	Quantitative Continuous Measurement	The probability of the breed obeying the command the first time.	Training Related
height_avg_in *	Quantitative Continuous Measurement	Average height of this breed in inches.	Physical Trait
weight_avg_lbs *	Quantitative Continuous Measurement	Average weight of this breed in pounds.	Physical Trait

Derived Attributes: Original size and intelligence data had lower/upper bounds for weights, heights and repetitions which were averaged to produced reduced attributes weight_avg_lbs, height_avg_in, and reps_avg. Attribute avg_weight_lbs was discretized to obtain the size category attribute based on [4] which likely important since size categories determine pet service rates at vet clinics, grooming salons, day cares etc. The 5 level textual intelligence categories were replaced with discrete numbers (1 to 5) which brought this feature onto the same scale as those from the traits dataset.

Dataset Types: Source data contains 4 **tables** with information about popularity, size, intelligence, and other traits (like energy level, child friendliness, etc) respectively. Each **item** corresponds to a breed with breed name being the **key** based on which inner join was performed. All data sources are **static** with attributes rank_2013 to rank_2020 being **time varying**. In the visualization, **geometry**, **field**, and **sets** dataset type may be observed in the spider (shape of polygon formed upon connecting radial points), heatmap (grid xy intersection with colours/markers encoding other parameter value) and scatter (marker shape/colour clusters/sets) plots respectively.

Why Visualize: Complexity arises due to high data variety (32 attributes) with some heterogeneity (categorical + quantitative data) and significant data volume (105 items). Main purpose of visualization being data exploration and high dimensionality of data thus, necessitates use of more than 1 idiom and justifies the choice of building an analytical juxtaposed multi-faceted interactive exploratory visualization. Involvement of multiple graphical elements comprising 5 idioms introduces visual complexity. Majority of the many attributes being categorical in nature led to clutter, overplotting and difficulty w.r.t facilitating comparison. Complexity was **managed as follows**.

- Interactive filtration and responsive animated transitions of positions/colours in plots implemented to slice/cut data and thereby reduce amount of data displayed at once.
- Data points spread across smaller multiples (scatter plot y axis staggered by size).
- Headings and white space used to place idioms neatly in a grid layout.
- Overlapping graphical elements (like markers in the scatter plot) were made translucent and will popout (using colour/size – get bigger, opaque, a brighter outline, come to front) upon hover and go to the back on mouse-out to allow underlying points to come into better view.
- Most cognitively demanding plot is placed in the top left part of the dashboard with easier to read ones in the middle and simple, attractive, interactive ones set to the right (so they aren't ignored).
- Meaning of few colours (blue, orange, red, black) were kept consistent throughout the dashboard. Blue anywhere, always means "selected breed" only. All hoverable elements adopt an orange shade and turn red upon selection indicating filter applied. A "clear all filters" button is provided.
- Textual elements were rotated, spaced, sized, and thickened to improve readability and ensure no overlapping. Smaller text is magnified upon hover. Numbers show max 1 decimal point only.

Pre-Processing: Cleaning steps applied to all source 4 datasets involved dropping duplicate rows or ones with missing values, standardizing column names (give more meaningful names, make lowercase separated by _), dropping least important ones (e.g. links to data sources) and computing derived attributes. The trickiest part was detecting differences and replacing breed names so that each breed is known by the same name across all 4 datasets before inner join. Please find all pre-processing steps (reproducible) in the "preprocessing.ipynb" file inside the "data" folder.

2 Tasks

Following are some of the tasks that the visualization facilitates.

- **Discover** ranking trends across years from features rank_2013 to rank_2020 in the line plot for different target breeds.
- **Identify** height/weight outliers in different breed size categories using scatter plot.
- Upon picking a breed as target, **lookup** values of 31 features.
- Filter by values of features other than breed as target to **browse** breeds that satisfy all selected conditions (logical AND filter implemented).
- 2D table and star plots depict average values for all/filtered breeds in addition to exact selected breed specific metrics and thus **summarizes** corresponding parameter values.
- **Compare** attribute values among different breeds (supported by all plots) and against all/filtered average values (supported by star and 2d table plots).
- Combine more than 1 filter to **derive** paired attributes.
- Identify **correlations**, **outliers,** or **trends**.

3 Encoding Channels & Idioms

The visualization is divided into 6 sections (S) as per dog breed data categories as follows collectively comprising 5 different idioms (scatter, line, star, heatmap, image, bar) as follows.

S1 Physical Traits: Continuous quantitative attributes height and weight are encoded using position in a scatter plot idiom since this would allow for viewing of clusters resulting from encoding categorical attributes coat length and coat type using shape and colour channels respectively. Since shapes have a smaller discriminable range, it was used to encode coat type with fewer classes (3) while colour was used to distinguish 8 categories. Colours were chosen to be as visually distinct as possible. Text displays data corresponding to hovered/selected point only to improve precision and ease of reading the axes while minimising clutter. Data point corresponding to the selected breed is bigger and has a bold blue outline to make it popout. Users may select a data point to select a breed or may filter displayed data by coat length, coat type or size by selecting a legend marker.

S2 Popularity Ranking: Breed wise popularity ranks for 1 year is represented using a sorted bar graph since it is a great choice for comparison of values. Bar height (size) encodes rank and position along x axis encodes breed. Changes in rank over time for the selected breed is represented using a line plot idiom as this allows for effective trend visualization. Users may filter by year or breed. The selected breed is highlighted in the horizontally scrollable (interaction used to provide a sliced view of data and reduce complexity) bar plot. Selecting a new breed name (bar plot x axis) brings corresponding bar into focus (blue, in center) and filters data across the dashboard to reflect new selected breed. Both bar and line plots together display ranking of 105 different breeds across 8 years with good precision and minimal complexity.

S3 Training Related Parameters: A heat map/2D table/matrix was used to encode 2 categorical ordinal attributes "'working intelligence level" and "trainability" as well as a quantitative continuous attribute, "avg. no. of repetitions". A fourth parameter, "obedience probability" is also encoded in each field using a pie plot idiom as it is apt for representing percentage values. To allow more precise readings, "avg. no. of repetitions" number is also displayed in each matrix field. Clicking on a field filters the entire dashboard to display data with corresponding combination of all 4 parameters. The grid also gets updated to display data associated with filters applied in other plots. Selected breed's "obedience probability" and "avg. reps" data is represented in a separate field outside the main matrix with a blue outline. Its position within the matrix is marked using a blue (blue for selected breed as always) circle. This matrix heatmap allows users to locate, browse, explore, compare, group, and summarize breeds as per the 4 training related parameters quickly.

S4 Protection Related Parameters:

The same idiom as for S3 is adopted here, without the pie plot markers, to encode 3 categorical attributes related to the inclination of a dog breed to protect being "barking level", "openness to strangers" (using position) and "avg. protectiveness" (using colour) in a single plot. Once again, each matrix field may be used to filter data and the matrix reflects filters applied elsewhere.

S5 Care Needs & Home Suitability: Two star plots use radial position to encode each corresponding rating features. A star plot was chosen since all features have the range [1, 5] and together capture different attributes that all provide information about the same aspect such as needs of a dog or its temperament. The 5 attributes related to breed needs is depicted in 1 star plot while another one shows metrics related to breed temperament. Colours blue and grey encode selected breed specific and all/filtered breed average values respectively. Translucency is used to ensure visibility despite overlapping. Users can use these plots to explore/compare how suitable different breeds are for certain kinds of households. Each point in these plots is hoverable and can be clicked to filter displayed data. Connected points for a polygon that changes in response to filters applied elsewhere. Since there are only 5 attributes per plot, angles are distinguishable. Grid lines and axis labels are provided for easy Code for star plot creation using D3.js was inspired from [5].

S6 Selected Breed: Name and image of selected breed is displayed. The image provides added information such as physical features not explicitly provided by attributes like snout length, chest depth, meaning of variation in coat type/length, leg length, etc, that may have correlations with plotted parameters/data clusters.

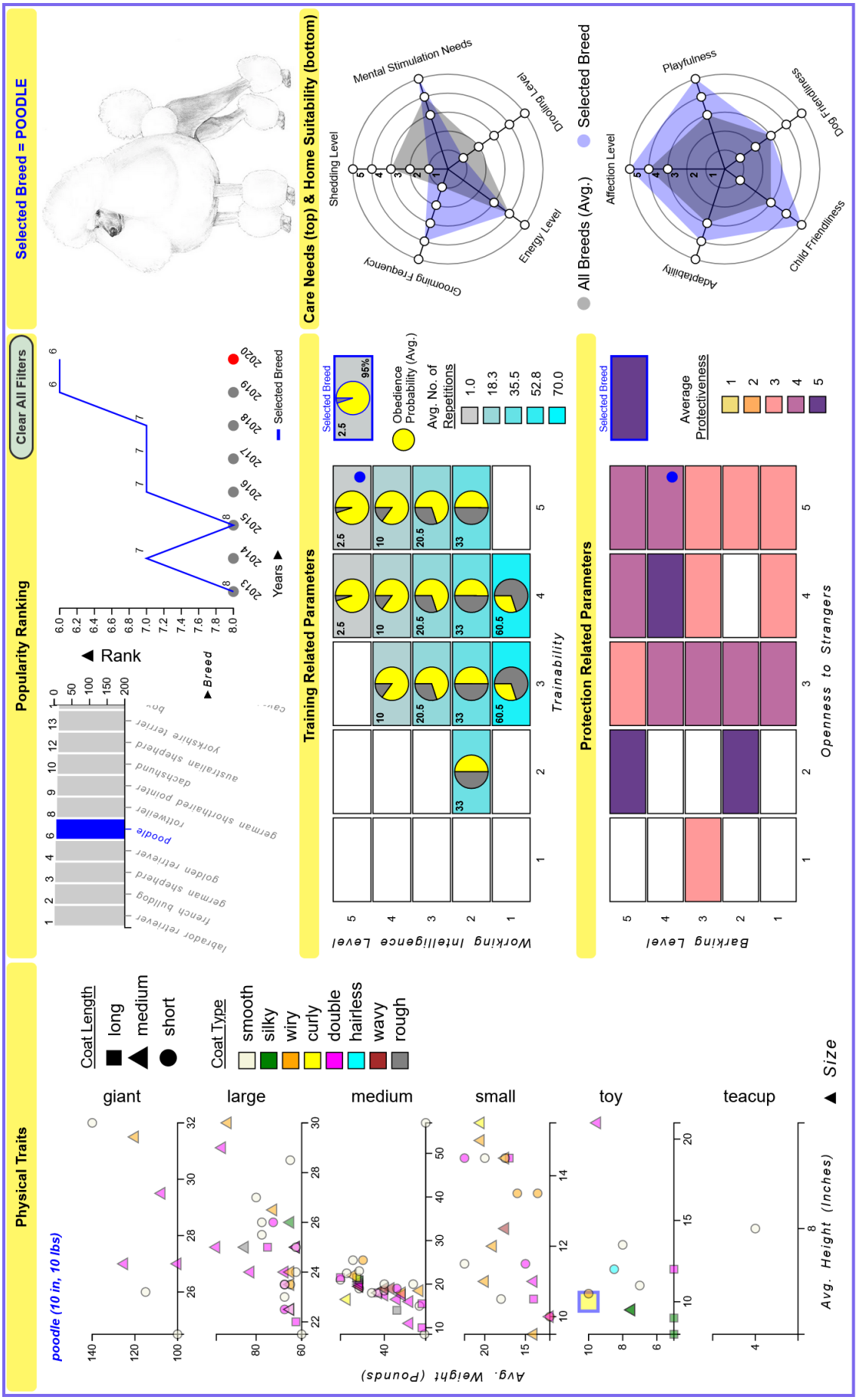
Animation: Animation is present is in the form of animated transitions in response to filters applied/removed to emphasize change in values among breeds (all plots) or years (line plot) further supporting the task of comparing different values.

4 Novelty

To the best of knowledge, no other explorative interactive visualizations exist that depict almost all attributes from the 4 datasets used here in a dashboard with multiple cohesive sub-plots comprising 5 idioms with a creative combination of the matrix and pie plot idiom supporting extensive filtration.

5 Strengths & Weaknesses

Strength: Manages complexity well. Accuracy of most attributes preserved. Good separability. Colour used effectively. Clear and sufficient legends/labels. Supports extensive filtering, data grouping and multiple tasks. Engaging due to useful interactions. **Weakness:** Overlapping in scatter plot may be reduced further by adding brush select zoom/pan. Some colours (coat type) are less discriminable. Obedience probability is less precisely displayed.


Selected Breed = POODLE
Clear All Filters
Physical Traits
Popularity Ranking
Training Related Parameters
Care Needs (top) & Home Suitability (bottom)
Protection Related Parameters
poodle (10 in, 10 lbs)
Coat Length
long
medium
short
Coat Type
smooth
silky
wiry
curly
double
hairless
wavy
rough
giant
large
medium
small
toy
teacup
Size
Avg. Weight (Pounds)
Avg. Height (Inches)
Rank
Breed
Years
Selected Breed
Obedience Probability (Avg.)
Avg. No. of Repetitions
Trainability
Working Intelligence Level
Barking Level
Openness to Strangers
Average Protectiveness
All Breeds (Avg.)
Mental Stimulation Needs
Energy Level
Grooming Frequency
Shedding Level
Playfulness
Child Friendliness
Adaptability
Affection Level
Dog Friendliness

References

[1] L. Fishman, “Dog/Canine Breed Size (AKC),” 20 December 2016. [Online]. Available: https://data.world/len/dog-canine-breed-size-akc. [Accessed December 2023].

[2] L. Fishman, “Dog size/intelligence linked?,” 31 January 2017. [Online]. Available: https://data.world/len/dog-size-intelligence-linked. [Accessed December 2023].

[3] S. Kapadnis, “Dog Breeds,” October 2023. [Online]. Available: https://www.kaggle.com/datasets/sujaykapadnis/dog-breeds/data. [Accessed December 2023].

[4] Monika, “Is your dog small, medium, or large? The ultimate guide to dog sizes.,” 12 October 2022. [Online]. Available: https://vetcarenews.com/small-medium-large-dog-size-by-weight-guide/. [Accessed December 2023].

[5] D. YANG, “D3 Spider Chart Tutorial,” 1 March 2019. [Online]. Available: https://yangdanny97.github.io/blog/2019/03/01/D3-Spider-Chart. [Accessed December 2023].