Machine Learning Methods: a look into the dog register of the city of Zürich

Rina Gandolfi, Daniel Herrera & Lina Scarborough

2024-05-26

1. Introduction

1.1. Presentation of the case

We introduce Pet Paradise, a successful Zürich-based pet shop business, wants to expand. However, the owner is not aware of current and future customer needs. The owner has approached us to advise what predictions we can make about canine breeds, ages, neighborhood concentrations and distributions, to figure out where to open the next Pet Paradise branch and what products to offer there.

The objective of our report is to advise Pet Paradise and predict dog breed trends, in order to facilitate targeted marketing. Our analysis delves into canine and pet owner data. We have explored Zürich's neighborhoods to identify prevailing canine demographics and trends. So our predictive analysis extends beyond just current demographics to anticipate future trends, allowing Pet Paradise to stay ahead of evolving customer needs.

By monitoring shifts in dog ownership patterns, breed popularity, and lifestyle preferences (size, number of dogs), Pet Paradise can adapt its product offerings and marketing strategies.

With our help, Pet Paradise can leverage data-driven insights to grow their business and help Zurich's canines live to their best health and dog-happiness. We envision Pet Paradise expanding into other cantons too if they foster the customer satisfaction to shape Switzerland's pet industry landscape.

1.2. Motivations

Dogs are an integral part of urban communities, with pet ownership having grown parallel to population over the past decades. Zürich and its canton boast the largest population of dogs of any Swiss region, as suggested by a 2013 study (Pospischil et al. 2013), proving the value of building a thorough data-driven interpretation of the markets associated with dog ownership.

Additionally, in a report (Statistik Stadt Zürich 1984) by the Zürich City Police, historical records and concerns due to the environmental impact of an ever increasing canine population have led to stricter laws against dog-produced waste and specific taxing for dog owners. The implementation of dog registration procedures, which date back several hundred years, has facilitated the collection of valuable statistical data, which provides a glimpse into the relationship between owners and their pets across time (with owner personal information limitations, due to privacy concerns). The existence of such cohesive and easily available data has served as a strong motivating factor for our team to undertake the current project.

Moreover, our data science team consists of individuals who are deeply passionate about dogs, each with varying degrees of personal experience in pet ownership, and understand the importance of analyzing the current dynamics between humans and dogs in Switzerland from an analytic perspective.

1.3. Disclaimer

The analysis in this report is conducted purely for educational purposes, focusing solely on the statistical modeling and client recommendations. This means that within the confines of this project, sex values for an animal such as dogs and their human owners within the dataset are interpreted as binary. The findings presented here are not intended to reflect or endorse any particular social reflections related to sexual identity. We acknowledge that discussions surrounding sex identity are multifaceted for individuals and communities.

---

2. Data exploration

2.1. Data source and preparation

As mentioned above, the main data has been sourced from the dog register (Stadt Zürich 2024), having been collected and published by the Open Data Portal of the City Council of Zürich, under the name "Hundebestände der Stadt Zürich, seit 2015". The description of the data set from the original source is as follows:

This dataset contains information on dogs and their owners from the municipal dog register since 2015. Information on the age group, Sex and statistical district of residence is provided for dog owners. The breed, breed type, sex, year of birth, age and color are recorded for each dog. The dog register is kept by the Dog Control Department of the Zurich City Police.

To ensure a seamless workflow and make variable interpretation easier for our group, we have undertaken several preparatory steps with the dataset. These include renaming columns and translating certain string values from German to English, along with performing some cleaning procedures.

The main source of data is the `kul100od1001.csv` file, which contains a collection of 70,967 dog registrations with 33 variables.

For the English version, translations for the column names are defined and a function is employed to replace multiple patterns at once for content translation. This includes translations for age groups, sexes, breed types, and dog colors. After applying the translation function across all relevant columns, dog colors are also translated. From this point forward, we will refer to variables and items exclusively by their translated English names.

The next step involves identifying and marking the initial occurrence of each `OwnerId` as unique within the dataset. This distinction facilitates further analyses that may require the identification of distinct entries.

Finally, a subset of relevant columns is extracted from the comprehensive dataset, creating a streamlined dataframe named `df_EN_EDA`. The subset includes essential fields such as `KeyDateYear`, `OwnerId`, and details regarding the dogs, including `PrimaryBreed` and `DogBirthYear`. Additionally, the `NumberOfDogs` column is converted from its original format to a numeric type, ensuring that subsequent data analysis can utilize numerical operations.

We begin with a summary of the data, providing an overview of its structure and contents.

```
# Load dataset
str(df_EN)
```

```
## 'data.frame':    70967 obs. of  33 variables:
##  $ KeyDateYear      : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ DataStatusCd     : chr  "D" "D" "D" "D" ...
##  $ OwnerId          : int  126 574 695 893 1177 4004 4050 4155 4203 4215 ...
##  $ OwnerAgeGroupCd  : int  60 60 40 60 50 60 40 60 50 40 ...
##  $ OwnerAgeGroup    : chr  "60 to 69 years old" "60 to 69 years old" "40 to 49 years old" "60 to 69 years old" ...
##  $ OwnerAgeGroupSort: int  7 7 5 7 6 7 5 7 6 5 ...
##  $ OwnerSexCd       : int  1 2 1 2 1 2 2 2 2 2 ...
##  $ OwnerSex         : chr  "male" "female" "male" "female" ...
##  $ OwnerSexSort     : int  1 2 1 2 1 2 2 2 2 2 ...
##  $ DistrictCd       : int  9 2 6 7 10 3 11 9 2 8 ...
##  $ District         : chr  "Kreis 9" "Kreis 2" "Kreis 6" "Kreis 7" ...
##  $ DistrictSort     : int  9 2 6 7 10 3 11 9 2 8 ...
##  $ QuarCd           : int  92 23 63 71 102 34 111 92 21 81 ...
##  $ Quar             : chr  "Altstetten" "Leimbach" "Oberstrass" "Fluntern" ...
##  $ QuarSort         : int  92 23 63 71 102 34 111 92 21 81 ...
##  $ PrimaryBreed     : chr  "Welsh Terrier" "Cairn Terrier" "Labrador Retriever" "Mittelschnauzer" ...
##  $ SecondaryBreed   : chr  "none" "none" "none" "none" ...
##  $ MixedBreedCd     : int  1 1 1 1 1 1 1 2 3 ...
##  $ MixedBreed       : chr  "Pedigree dog" "Pedigree dog" "Pedigree dog" "Pedigree dog" ...
##  $ MixedBreedSort   : int  1 1 1 1 1 1 1 2 3 ...
##  $ BreedTypeCd      : chr  "K" "K" "I" "I" ...
##  $ BreedType        : chr  "Small stature" "Small stature" "Breed type list I" "Breed type list I" ...
##  $ BreedTypeSort    : int  1 1 2 2 1 1 1 2 2 ...
##  $ DogBirthYear     : int  2011 2002 2012 2010 2011 2010 2012 2002 2005 2001 ...
##  $ DogAgeGroupCd    : int  3 12 2 4 3 4 2 12 9 13 ...
##  $ DogAgeGroup      : chr  "3 years old" "12 years old" "2 years old" "4 years old" ...
##  $ DogAgeGroupSort  : int  3 12 2 4 3 4 2 12 9 13 ...
##  $ DogSexCd         : int  2 2 2 2 1 1 1 2 2 ...
##  $ DogSex           : chr  "female" "female" "female" "female" ...
##  $ DogSexSort       : int  2 2 2 2 1 1 1 2 2 ...
##  $ DogColor         : chr  "black/brown" "brindle" "brown" "black" ...
##  $ NumberOfDogs     : int  1 1 1 1 1 1 1 1 1 ...
##  $ unique_OwnerId   : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

As can be seen in the structure of the data, the set comprises several observations of diverse data types. Most variables are expressed three times as different types, as integers (coded and sorted form), as well as strings (text). Depending on their implementation in the study they have been selected in one of the three variants, therefore our selection of relevant observations can be summarized as follows:

Numerical values:

- `KeyDateYear` : numerical value for the reference year
- `OwnerId` : numerical identifier for the owner of the registered dog
- `AgeV10Sort` : referring to the owner's age as a 10-year category
- `DogBirthYear` : numerical value for the birth year of the dog
- `DogAgeSort` : referring to the dog's age at the time of registration
- `NumberOfDogs` : numerical counter of the dog count for each dog owner

Binary variables: !!! Is breed multinomial or factor? !!!

- `DogSexText` : numerical value indicating two states for the biological sex of the dog

String values:

- `DistrictText` : the name of each larger district of Zürich according to the official division
- `QuarterText` : the name of the smaller neighborhoods which comprise the larger districts
- `Breed1Text` and `Breed1Text2` : referring to dog race denominations and information
- `MixedBreedText` : additional information regarding race mixing in the dog
- `DogColorText` : a descriptive name for the color of the dog
- `BreedTypeLong` : referring to the official dog type classification according to the dog ordinance (Regierungsrat 2009)

2.2. Exploratory Data Analysis

2.2.1. Analyzing Diversity in Dataset Features: Years, Owner IDs, and Age Groups

This series of R code snippets delves into the examination of key features within the `df_EN_EDA` dataframe, focusing on the identification and analysis of unique entries for `KeyDateYear` , `OwnerId` , and `OwnerAgeGroup` . Each code section is designed to extract unique values, count these entries, and where applicable, visualize the distribution. Such analysis is integral for understanding the dataset's diversity across different dimensions, helping to highlight temporal coverage, ownership uniqueness, and demographic variations among owners.

```
# Extract and count unique years
unique_years <- unique(df_EN_EDA$KeyDateYear)
number_of_unique_years <- length(unique_years)
print(number_of_unique_years)
```

```
## [1] 9
```

```
print(unique_years)
```

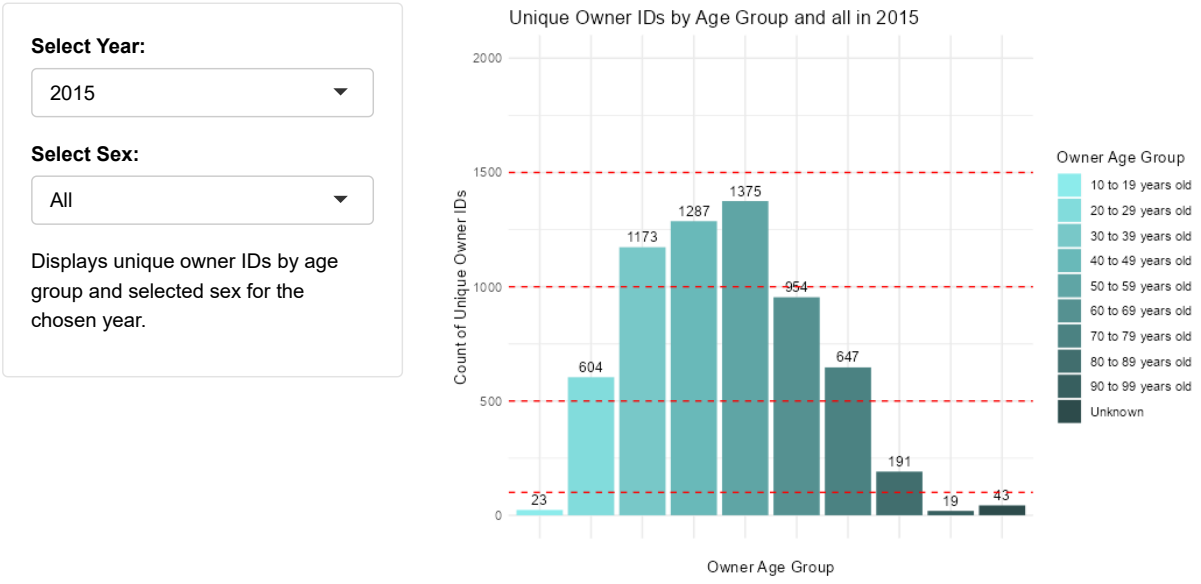
```
## [1] 2015 2016 2017 2018 2019 2020 2021 2022 2023
```

```
# Extract and count unique Owner IDs
unique_Owner <- unique(df_EN_EDA$OwnerId)
number_of_unique_Owner <- length(unique_Owner)
print(number_of_unique_Owner)
```

```
## [1] 15504
```

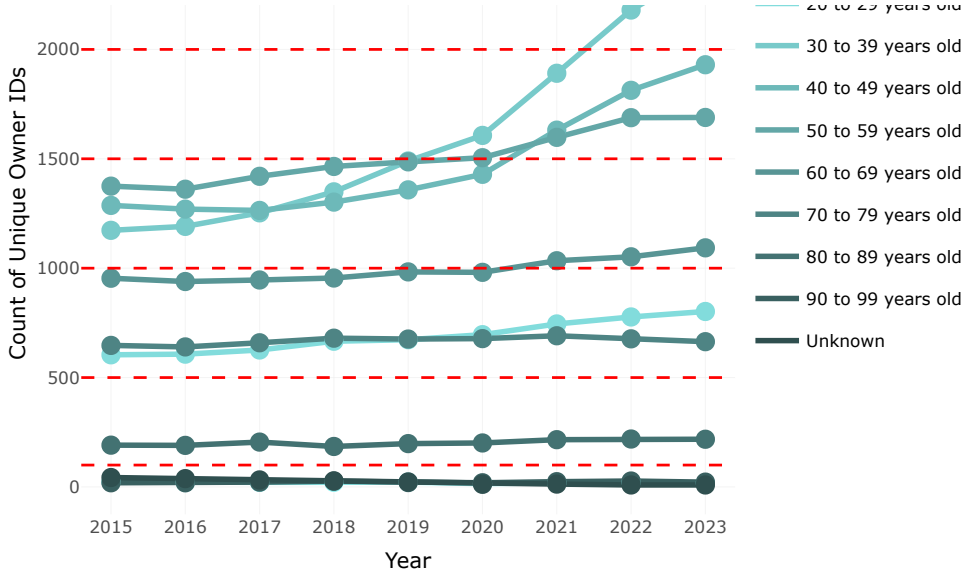
2.2.2. Unique owners by year, by age group and by sex

This section presents an interactive visualization that displays unique owner IDs by age group and sex for a selected year. The user interface allows the selection of a year and a sex, and the resulting plot shows the distribution of unique Owner IDs across different age groups based on the chosen criteria.



2.2.3. Unique owners by age group over the years

The following visualization shows the count of unique owner IDs across different age groups over the years. The plot is generated by aggregating unique owner IDs by age group and year, adjusting factor levels, and creating a line plot.

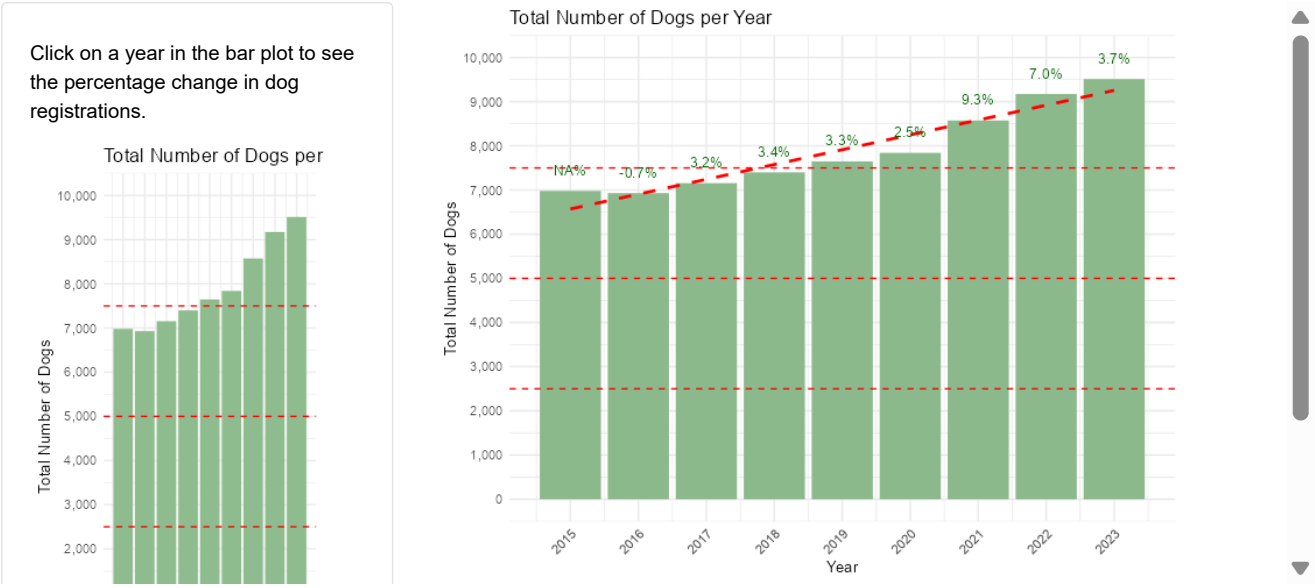


2.2.4. Yearly dog counts

After confirming successful conversion, we aggregated the data to compute the total number of dogs per year. The resulting counts were then visualized using histograms to illustrate the distribution over the years.

Furthermore, to understand the trend in dog population over time, we calculated the percentage change between consecutive years. This allowed us to identify any notable fluctuations or patterns in the data.

```
## [1] 0
```



2.2.5. Heatmap of total number of registered dogs per year

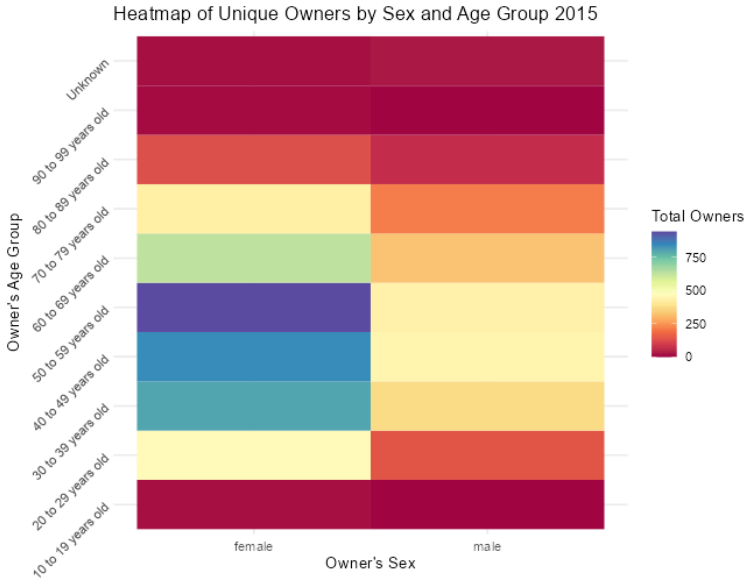
A heatmap was created to illustrate the distribution of dogs based on the sex and age group of their owners across different years. The data was organized by grouping it according to the year, owner's age group, and Sex. Separate heatmaps were generated for each year to visualize the data for that specific period.

Each heatmap represents the total number of dogs in various age groups, categorized by the Sex of their owners. The color gradient within the heatmap indicates the intensity of dog ownership, with warmer colors representing higher dog counts.

Select Year:

2015

Displays a heatmap of unique owner counts by age group and sex for the selected year.



2.2.6. Total count of dogs

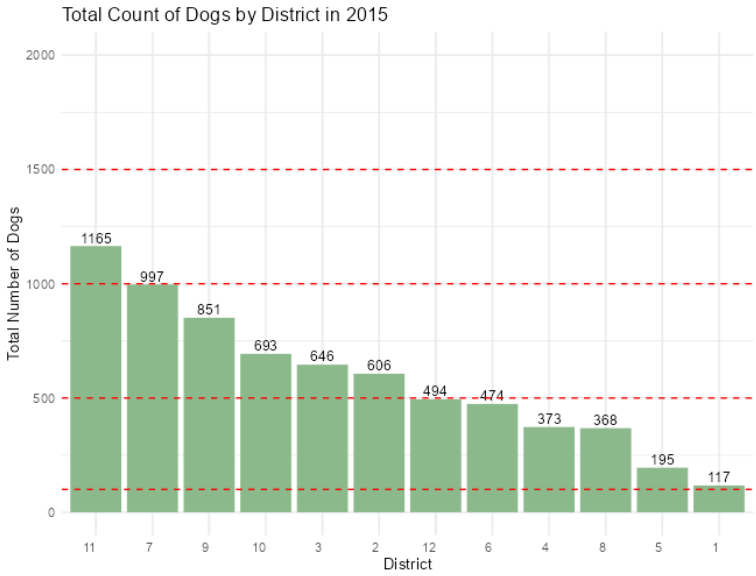
2.2.6.1. Total count of dogs by district

The annual distribution of dog registrations across various districts is examined using an interactive Shiny application. Users can select a year to view the total count of dogs by district for that specific year. The data is processed and visualized to provide insights into the distribution patterns over time.

Select Year:

2015

Displays the total count of dogs by district for the selected year.



2.2.6.2. Total count of dogs by district with dog's sex

To enhance understanding of the distribution of dogs across different districts and introduce a Sex perspective into the analysis, the approach has been modified to include a breakdown by Sex. This adjustment allows observation of not only the geographical distribution but also Sex dynamics within the dog population each year.

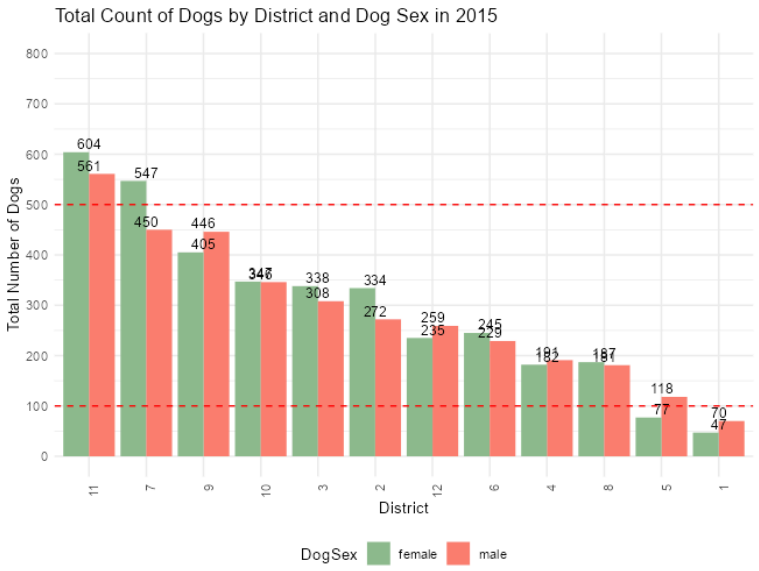
Select Year:

2015

Select Sex:

All

Displays the total count of dogs by district and sex for the selected year, ordered by total count.



2.2.6.3. By district and Unique owner's sex

This Shiny application provides an interactive visualization of the total count of dogs by district and owner's Sex for a selected year. Users can select a year and the Sex of the owner to explore the distribution of dogs across different districts, facilitating insights into demographic and geographic trends in dog ownership.

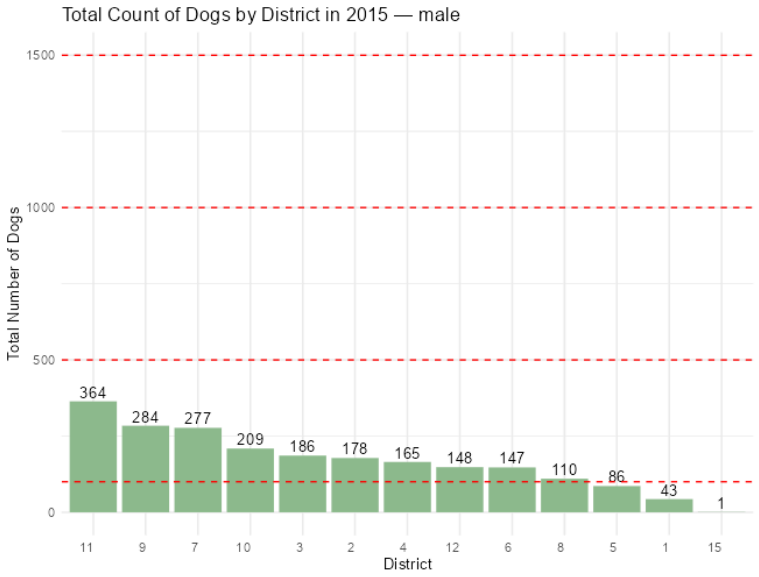
Select Year:

2015

Select Owner's Sex:

Male

Displays the total count of dogs by district and sex for the selected year.



2.2.6.4. By district and breed type

To deepen the analysis of dog populations across different districts annually, the R script incorporates an additional layer of granularity by assessing dog counts not only by district but also by breed type. This enhancement provides a more detailed view of the diversity within the canine populations across various districts each year.

Select Year:

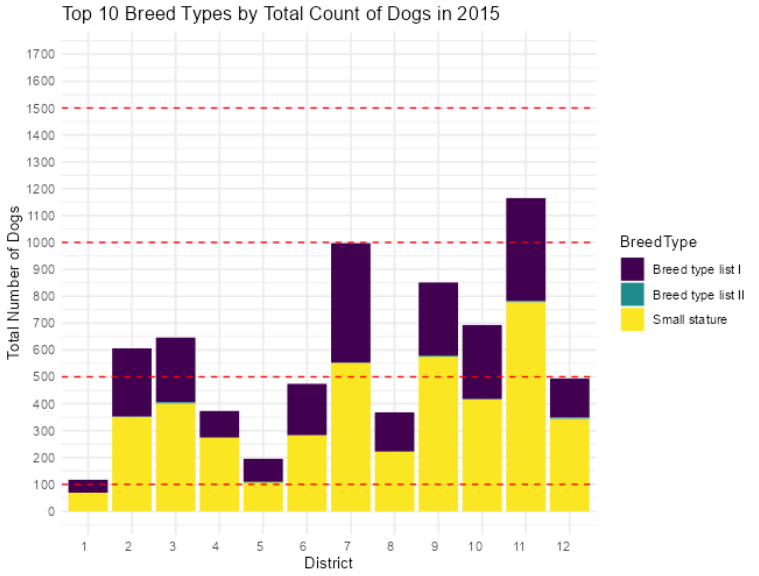
2015

Select Breed Type:

All

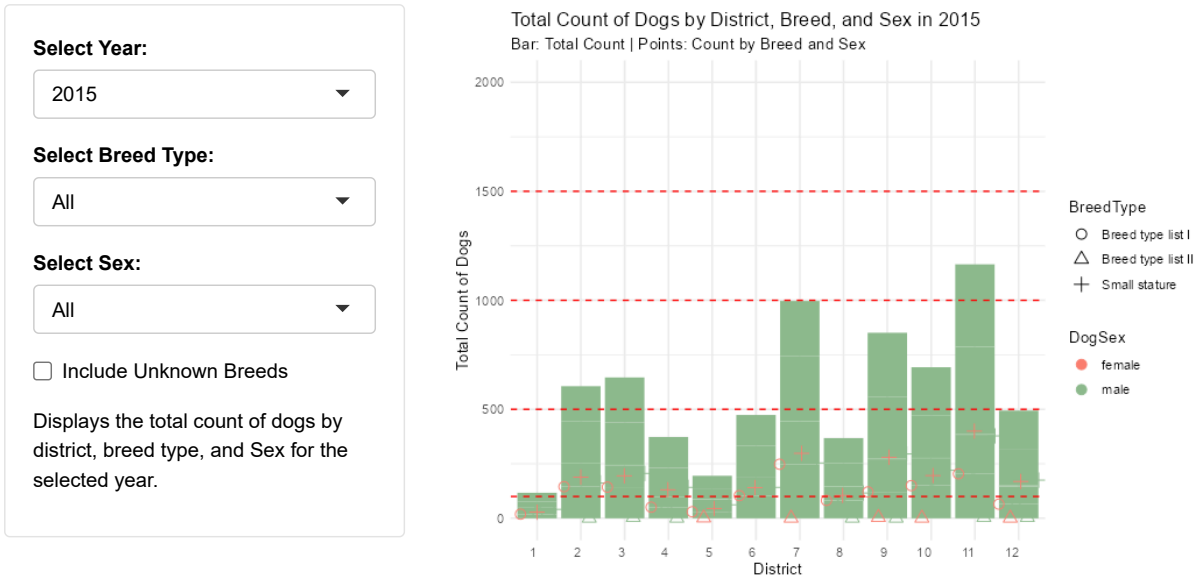
☐ Include Unknown Breeds

Displays the total count of dogs by district and selected breed type for the selected year, excluding District 15.



2.2.6.5. By district, breed type and dogs's sex

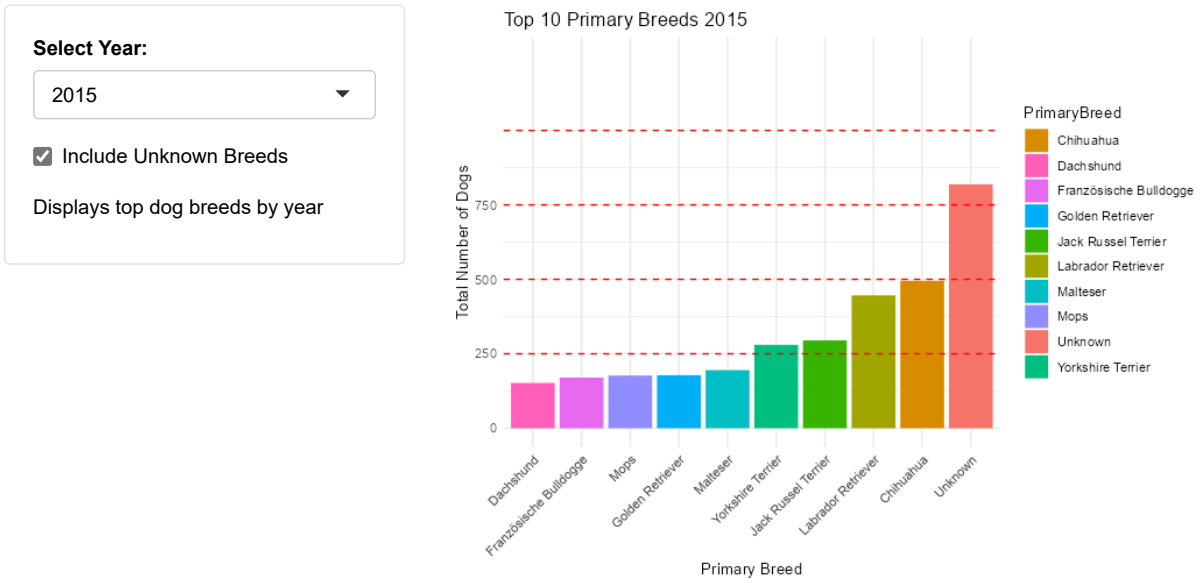
In an effort to provide a more comprehensive analysis of dog populations within various districts, the latest R script has been enhanced to include not only total counts by district but also a detailed breakdown by breed type and Sex. This enhancement aims to offer a deeper understanding of the diversity and demographics of canine registrations across different regions.



2.2.7. Top dog breeds

2.2.7.1. Top dog breeds by year

This Shiny application provides a clear and dynamic way to visualize the most popular dog breeds each year, allowing for the inclusion or exclusion of unknown breeds. The color-coding of breeds enhances readability and helps in quickly identifying trends.



2.2.7.2. Top dog breeds by district

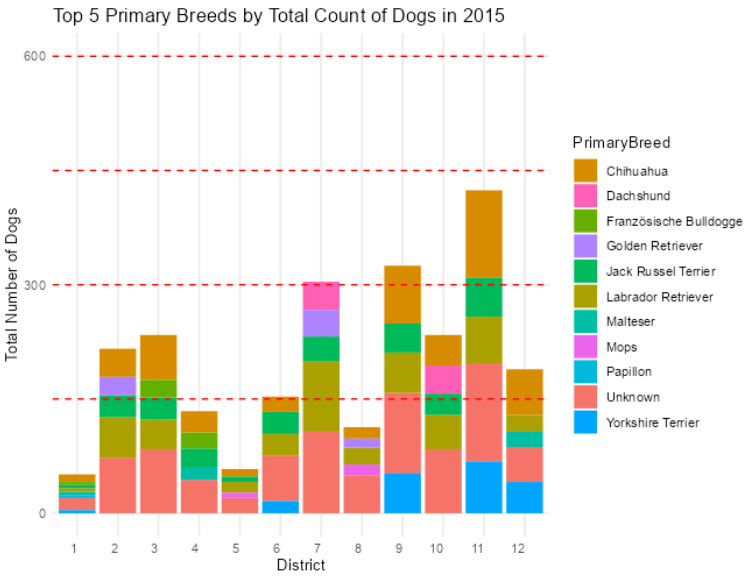
This visualization enables users to explore the distribution of dog breeds across different districts for a selected year. It highlights the top 5 dog breeds in each district, allowing for a detailed understanding of breed trends and distribution patterns.

Select Year:

2015

☑ Include Unknown Breeds

Displays top dog breeds by district for the selected year.



Select Year:

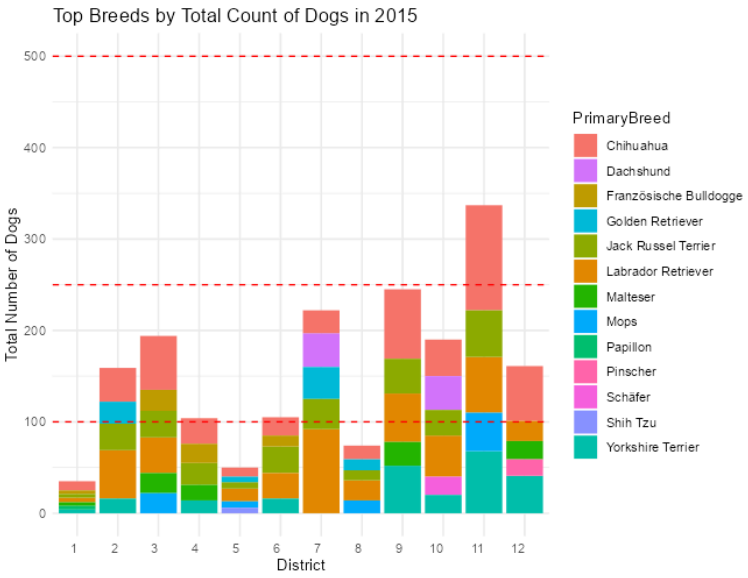
2015

Select Breed Type:

All

☐ Include Unknown Breeds

Displays the total count of dogs by district and selected breed type for the selected year.

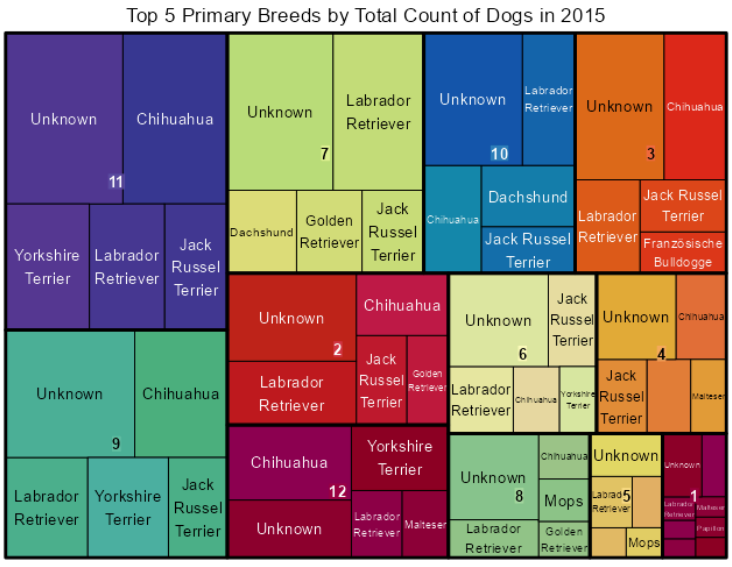


Select Year:

2015

☑ Include Unknown Breeds

Displays top dog breeds by district for the selected year.



3. Machine Learning Models

3.1. Linear Model

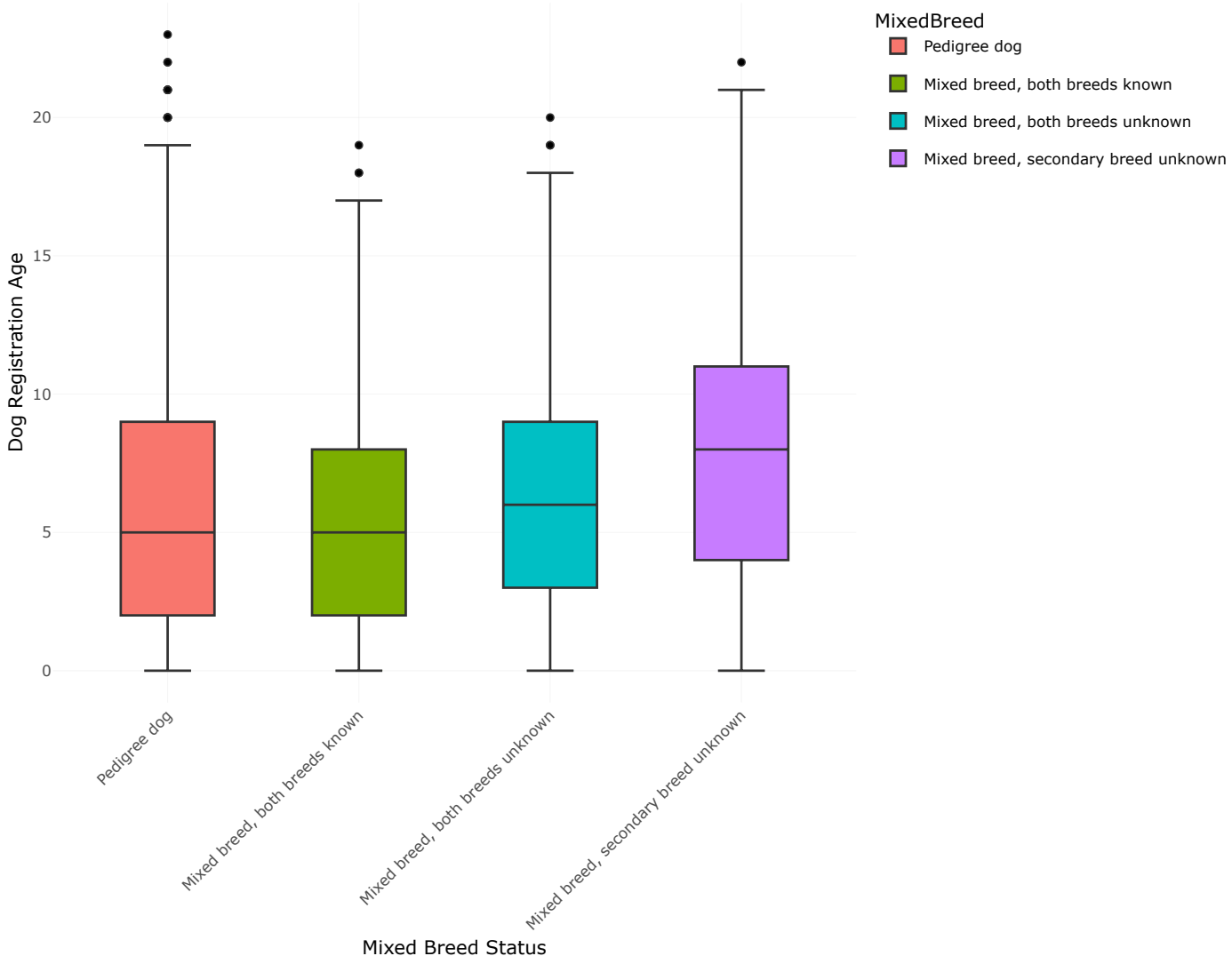
3.1.1. Linear Model for hypothesis testing: dog age by breed status

For the first chapter in the machine learning models we begin with a linear model that will test the effect of a categorical variable. To do this we set ourselves the following question: *does the breed status have an effect on the age at which dogs are registered?*

Understanding typical registration ages for different breeds will allow Pet Paradise to target marketing effectively, reaching owners at the right stage of their pet ownership journey. Additionally, these insights can inform strategic inventory management, anticipating demand for breed-specific products and offering tailored advice to enhance customer satisfaction.

In the context of the question at hand, it is worth noting that linear models will not be employed to generate predictions from the provided data. Rather, they will be utilized to determine whether there is a linear correlation between the various states of categorical variables (such as pedigree dog, non-pedigree, etc.) and the response variable, which denotes the age of the dog.

To answer this question we will consider the `DogAgeGroupSort` as the response variable, and the different levels of the categorical variable `MixedBreed` as predictors. We now direct our attention to the following set of boxplots showcasing the relevant variables.



Based on the boxplots above, there does seem to be a difference in dog ages based on their breed status. Interestingly, the most extreme outliers are associated with pedigree dogs. We will continue by defining a linear model.

```
lm.dogs.1 <- lm(DogAgeGroupSort ~ MixedBreed, data = df_EN_cleaned)
summary(lm.dogs.1)
```

```
##
## Call:
## lm(formula = DogAgeGroupSort ~ MixedBreed, data = df_EN_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4917 -3.7002 -0.7002  3.2998 17.2998
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  5.70021    0.01839 310.011
## MixedBreedMixed breed, both breeds known    -0.37787    0.05838  -6.473
## MixedBreedMixed breed, both breeds unknown   0.58939    0.04715  12.501
## MixedBreedMixed breed, secondary breed unknown  1.79145    0.05841  30.670
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## MixedBreedMixed breed, both breeds known    9.68e-11 ***
## MixedBreedMixed breed, both breeds unknown   < 2e-16 ***
## MixedBreedMixed breed, secondary breed unknown  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.14 on 70955 degrees of freedom
## Multiple R-squared:  0.01569,    Adjusted R-squared:  0.01565
## F-statistic: 376.9 on 3 and 70955 DF,  p-value: < 2.2e-16
```

The intercept refers to the pure breed dogs, while the following predictors represent the differences between themselves and the intercept. The summary of the linear model suggests there is strong evidence that the mean age of pedigree dogs is not equal to 0 at the time of registration, with a value of 5.7 years, while the other three breed categories' ages differ significantly from that of pedigree dogs. The most noticeable difference is between pedigree dogs and those whose secondary breed is unknown, with the latter being 1.79 years older.

We follow up this insight by assessing the differences between each of them.

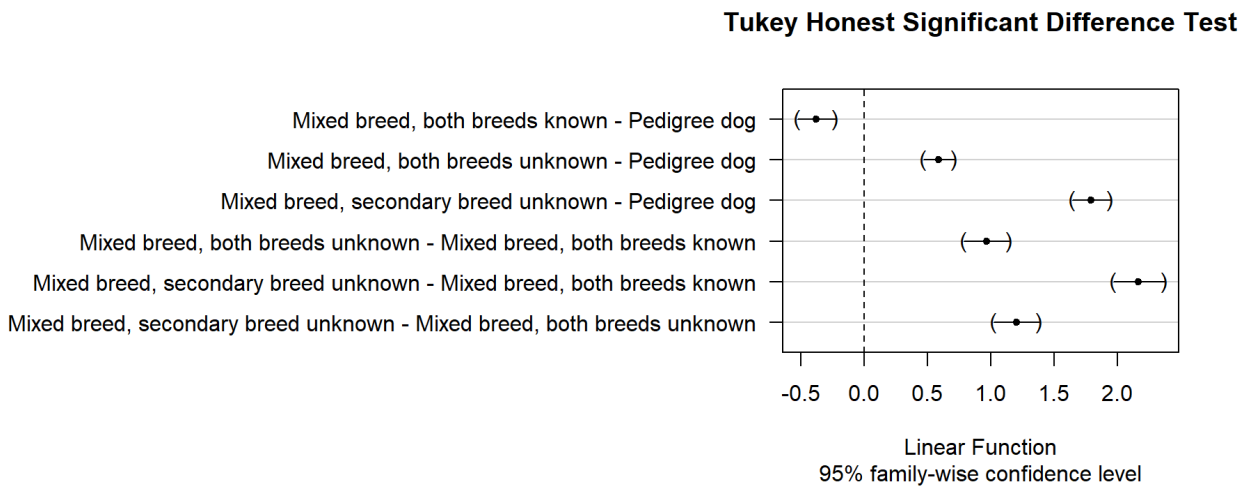
```
drop1 <- drop1(lm.dogs.1, test = "F")
drop1
```

```
# Single term deletions

Model:
DogAgeGroupSort ~ MixedBreed
           Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>                  1216314 201636
MixedBreed  3     19384 1235698 202752  376.93 < 2.2e-16 ***
```

By performing single term deletions and evaluating the resulting statistics of the model, we find that the breed status does indeed have a significant effect on the age of dogs across its different levels, however limiting these observation to one level at a time.

We can further this insight by drawing a General Linear Hypothesis. We will consider all possible pairwise comparisons with a *Tukey Honest Significant Difference Test*.

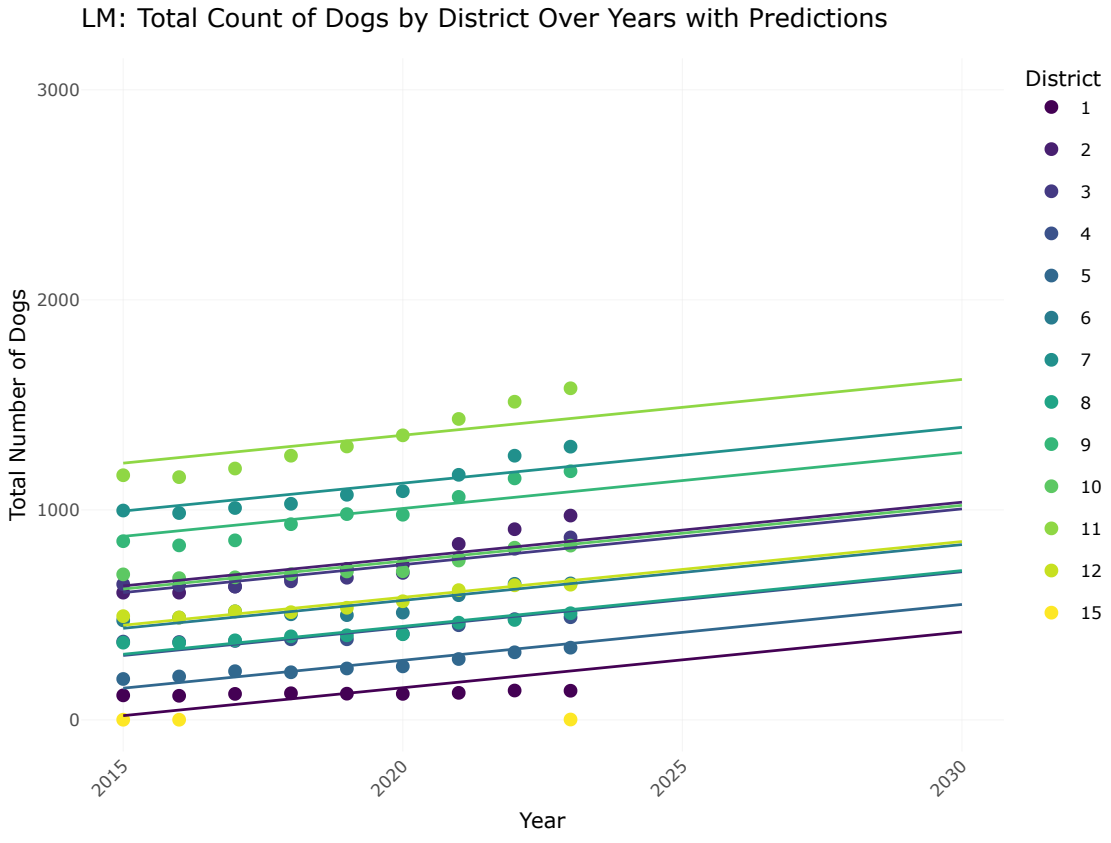


In the above *Tukey Honest Significant Difference Test* the similarity between means of different pairs is shown based on how different they are, with closeness to 0 representing no difference in their means. This, along with the 95% confidence intervals, provides an illustrative insight into pairwise variations.

3.1.2. Linear Model: dog count over time

As an additional implementation of linear models, we now aim at answering the following research question: *how do dog counts evolve over time?* To do so we will build a linear model that will provide some information about the trends in the time series of registered dog count data over the time period recorded, as well as some predictions for the following 10 years.

```
##
## Call:
## lm(formula = TotalDogs ~ DistrictSort + KeyDateYear, data = annual_dog_counts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -132.183  -29.459   -8.875   27.640  143.831
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -53518.004   3760.642 -14.231  < 2e-16 ***
## DistrictSort2     617.667     24.174  25.551  < 2e-16 ***
## DistrictSort3     585.889     24.174  24.236  < 2e-16 ***
## DistrictSort4     286.444     24.174  11.849  < 2e-16 ***
## DistrictSort5     130.778     24.174   5.410 4.54e-07 ***
## DistrictSort6     415.889     24.174  17.204  < 2e-16 ***
## DistrictSort7     974.111     24.174  40.296  < 2e-16 ***
## DistrictSort8     292.222     24.174  12.088  < 2e-16 ***
## DistrictSort9     853.556     24.174  35.309  < 2e-16 ***
## DistrictSort10    602.778     24.174  24.935  < 2e-16 ***
## DistrictSort11   1202.222     24.174  49.732  < 2e-16 ***
## DistrictSort12    430.444     24.174  17.806  < 2e-16 ***
## DistrictSort15    -98.763     34.238  -2.885  0.00483 **
## KeyDateYear        26.570      1.863  14.265  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.28 on 97 degrees of freedom
## Multiple R-squared:  0.982,  Adjusted R-squared:  0.9796
## F-statistic: 407.7 on 13 and 97 DF,  p-value: < 2.2e-16
```



3.2. Generalized Linear Model (Poisson)

We constructed a Poisson Generalized Linear Model (GLM) to estimate the number of dog registrations (`NumberOfDogs`) based on the predictors `KeyDateYear` and `DistrictCd` . The goal was to identify trends and distributions in dog ownership across Zurich's neighborhoods to support Pet Paradise's expansion strategy.

```
first_poisson_model <- glm(NumberOfDogs ~ KeyDateYear + DistrictCd,
                     family = poisson,
                     data = df_EN)

summary(first_poisson_model)
```

```
##
## Call:
## glm(formula = NumberOfDogs ~ KeyDateYear + DistrictCd, family = poisson,
##     data = df_EN)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.6227053  2.9106351   0.214    0.831
## KeyDateYear -0.0003063  0.0014414  -0.213    0.832
## DistrictCd  -0.0001025  0.0011205  -0.091    0.927
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 191.58  on 70966  degrees of freedom
## Residual deviance: 191.53  on 70964  degrees of freedom
## AIC: 142281
##
## Number of Fisher Scoring iterations: 4
```

```
cv_model
```

```
## Generalized Linear Model
##
## 70967 samples
##     2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 56774, 56774, 56773, 56774, 56773
## Resampling results:
##
##   RMSE        Rsquared      MAE
##   0.05929555  0.0002289192  0.006880633
```

We now interpret the results from the Generalized Linear Poisson Model:

- `KeyDateYear` : The coefficient for `KeyDateYear` is -0.0003 with a p-value of 0.832. There's no significant trend over the years in the number of dog registrations. Hence, seasonal trends in dog registrations are not a primary factor for Pet Paradise's planning.
- `DistrictCd` : The coefficient for the neighborhood is -0.0001 with a p-value of 0.927, so no significant difference in dog registrations across different districts. This suggests that district-specific variations in dog registrations might not be substantial.

Specifically in terms of client recommendations, geographical expansion can be addressed. The stable registration numbers mean that given that neither `KeyDateYear` nor `DistrictCd` significantly impact the number of dog registrations, dog ownership seems stable across different years and districts. Therefore, Pet Paradise can consider expanding uniformly across districts rather than focusing on specific areas with presumed higher dog populations.

Therefore in conclusion, this poisson GLM analysis shows a stability in numbers across Zurich. Pet Paradise should leverage this stability and expand based on other factors.

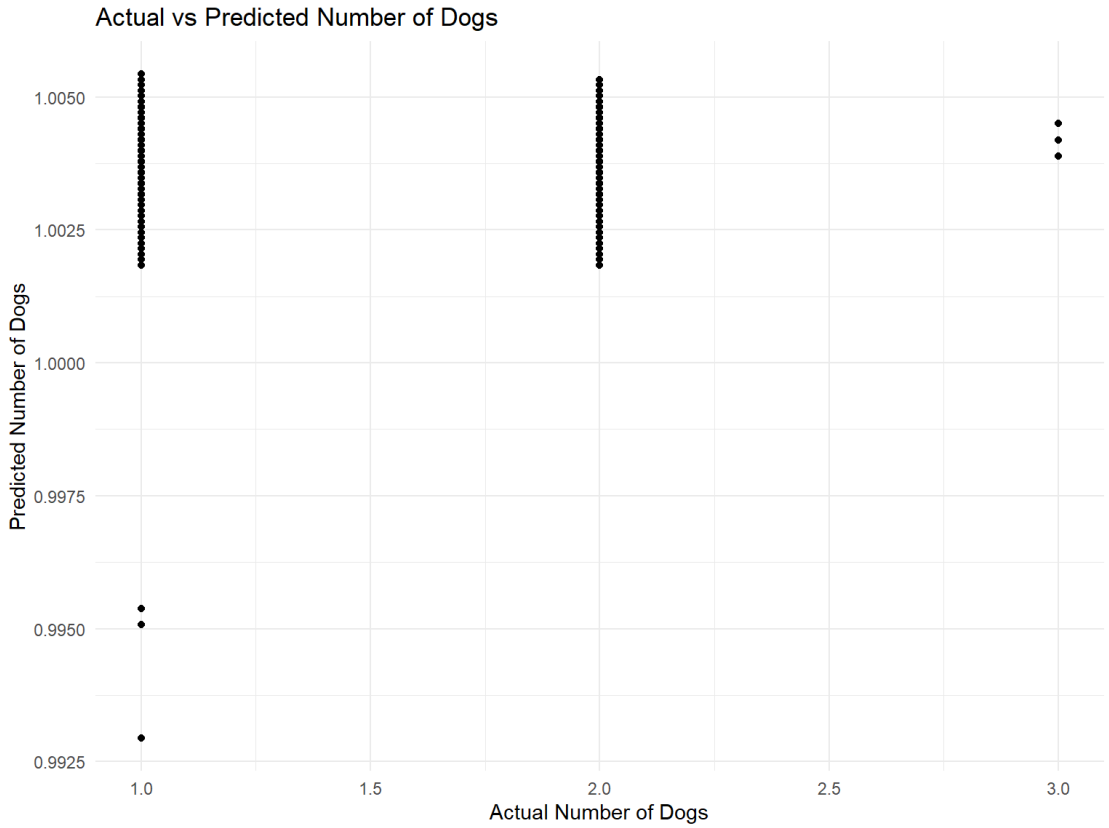
Model Deviance and AIC:

With a null deviance of 191.58 on 70,966 degrees of freedom, the high value indicates considerable variation in the number of dog registrations across the dataset. The residual deviance of 191.53 on 70,964 degrees of freedom how a minimal reduction from the null deviance which means that the predictors in the model do not significantly improve the fit compared to the null model.

The high AIC value indicates that while the model may have a reasonable fit, it is quite complex relative to the amount of information it provides. A Fisher Scoring Iterations of four iterations suggests that the model parameters have quickly converged, which we can expect for GLMs with well-behaved data.

Regarding the cross-validation, we used a 5-fold validation, i.e. breaking the data into five subsets, training on four, and testing on the fifth. The low Root Mean Square Error of 0.0593 shows that the model's predictions are close to the actual values of dog registrations. This R-squared shows that the predictors (here `KeyDateYear` and `DistrictCd`) explain almost none of the variability in the number of dog registrations. This is consistent with the insignificant coefficients observed.

In terms of client insights, the null hypothesis was proven correct in this case; the stability in the number of dog registrations across different years and districts suggests that Pet Paradise can plan for uniform expansion without focusing on specific districts. Other factors such as owner's age, whether they own a pedigree dog or not, etc. might be more influential in determining the demand for pet services, which we will examine in further models.

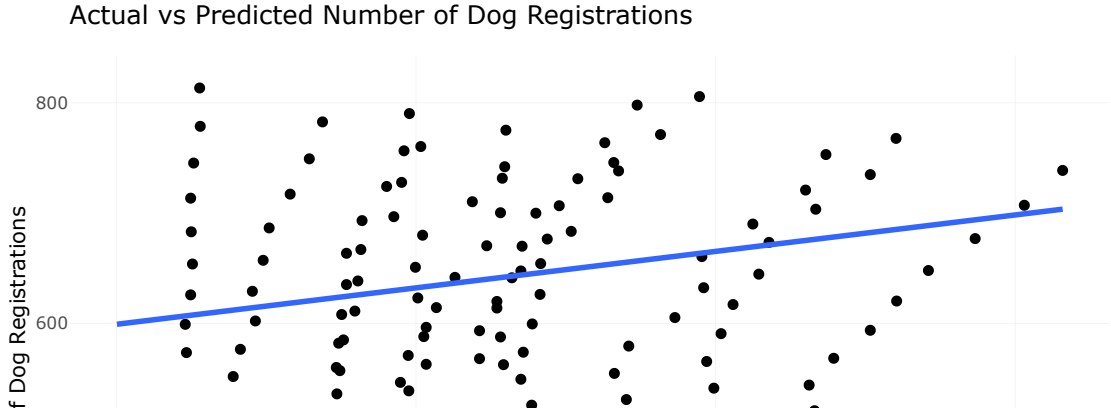


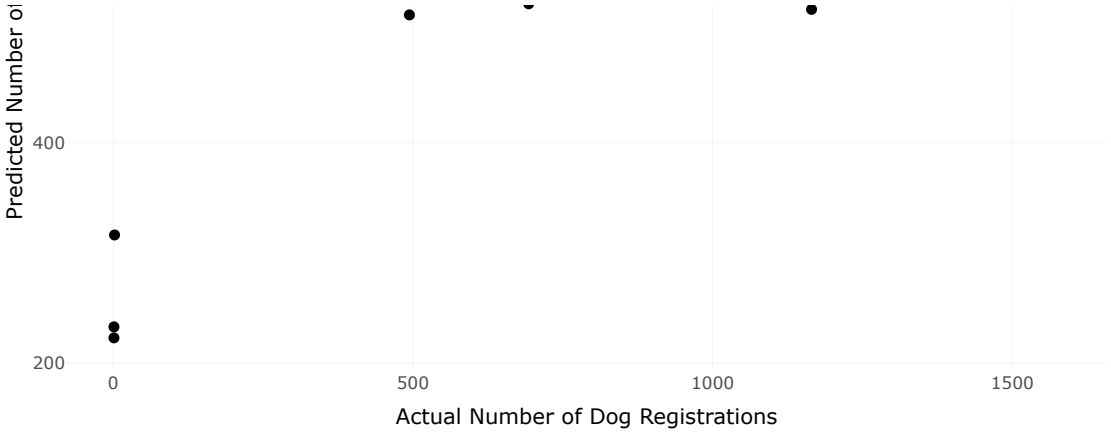
The above visualization helps us see that there is a poor fit; if the model were performing well, it would be a more diagonal spread of points, showing a clear linear relationship between actual and predicted values. The current chart does not show this pattern, so it is suggesting that the model is not capturing the variability in the actual data. Upon reflection, the model was refined to sum the number of dogs per year, i.e. aggregate it.

```
summary(second_poisson_model)
```

```
##
## Call:
## glm(formula = NumberOfDogs ~ KeyDateYear + DistrictCd, family = poisson,
##     data = sum_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -81.666277   2.925105  -27.92   <2e-16 ***
## KeyDateYear   0.043686   0.001448   30.16   <2e-16 ***
## DistrictCd   -0.009640   0.000368  -26.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 24077  on 110  degrees of freedom
## Residual deviance: 22173  on 108  degrees of freedom
## AIC: 23067
##
## Number of Fisher Scoring iterations: 11
```

```
## `geom_smooth()` using formula = 'y ~ x'
```





To check which poisson model is better, we compare the goodness-of-fit. This can be done using AIC because the models come from the same dataset, have the same response variable, as well as the same poisson distribution. We also compare coefficients.

```
summary_first <- summary(first_poisson_model)
summary_first
```

```
##
## Call:
## glm(formula = NumberOfDogs ~ KeyDateYear + DistrictCd, family = poisson,
##     data = df_EN)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.6227053  2.9106351   0.214    0.831
## KeyDateYear -0.0003063  0.0014414  -0.213    0.832
## DistrictCd  -0.0001025  0.0011205  -0.091    0.927
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 191.58  on 70966  degrees of freedom
## Residual deviance: 191.53  on 70964  degrees of freedom
## AIC: 142281
##
## Number of Fisher Scoring iterations: 4
```

```
summary_first$deviance[1]
```

```
## [1] 191.5258
```

```
summary_first$deviance[2]
```

```
## [1] NA
```

```
summary_first$coefficients
```

```
##                 Estimate  Std. Error     z value  Pr(>|z|)
## (Intercept)  0.6227052614 2.910635134  0.21394137 0.8305928
## KeyDateYear -0.0003062975 0.001441371 -0.21250433 0.8317136
## DistrictCd  -0.0001024754 0.001120495 -0.09145546 0.9271307
```

```
summary_second <- summary(second_poisson_model)
summary_second
```

```
##
## Call:
## glm(formula = NumberOfDogs ~ KeyDateYear + DistrictCd, family = poisson,
##     data = sum_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -81.666277   2.925105  -27.92   <2e-16 ***
## KeyDateYear   0.043686   0.001448   30.16   <2e-16 ***
## DistrictCd   -0.009640   0.000368  -26.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 24077  on 110  degrees of freedom
## Residual deviance: 22173  on 108  degrees of freedom
## AIC: 23067
##
## Number of Fisher Scoring iterations: 11
```

```
summary_second$deviance[1]
```

```
## [1] 22172.84
```

```
summary_second$deviance[2]
```

```
## [1] NA
```

```
summary_second$coefficients
```

```
##                 Estimate   Std. Error   z value      Pr(>|z|)
## (Intercept) -81.666277212 2.9251048123 -27.91909 1.564796e-171
## KeyDateYear   0.043686187 0.0014485481  30.15860 8.272900e-200
## DistrictCd   -0.009639715 0.0003679734 -26.19678 2.891998e-151
```

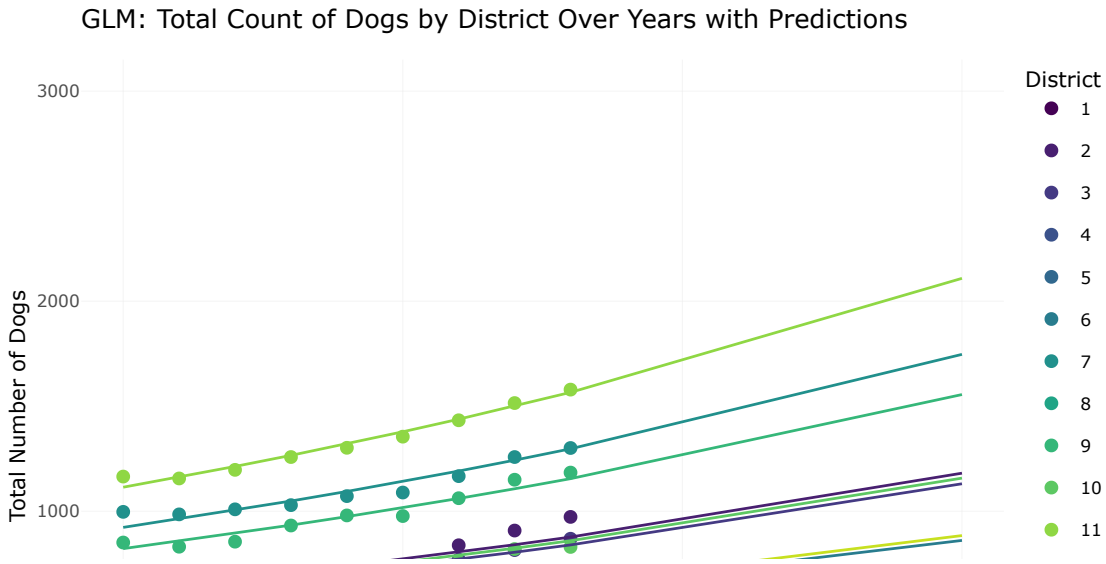
There is a big difference between the null deviance (24077) and residual (22173). The degrees of freedom here are lower, so the predictors improved the model's fit. The AIC value (23'067) is lower than the first model's AIC (142'281), meaning a better fit. The coefficients of the second model are statistically significant where p-value < 0.001, and shows an increase of registrations per year (RefYear = 0.044). The negative district coefficient needs to be log transformed to be interpreted since it's a Poisson model.

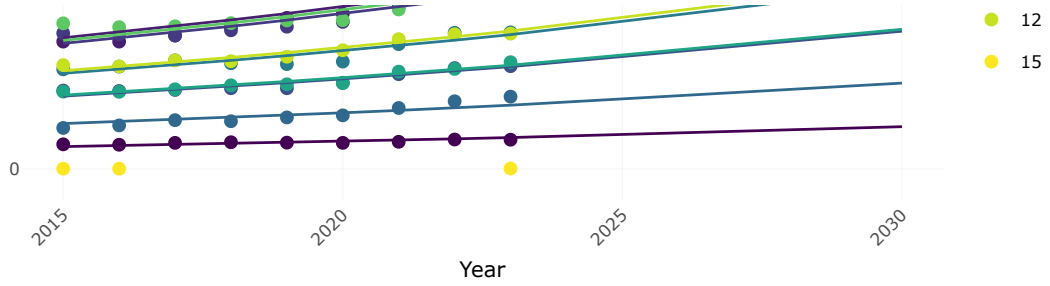
```
exp(District coefficient) = exp(-0.009640) = ca. 0.96
```

The above value is a percentage, so for every one-unit increase in District (ie just one neighbourhood to the next), the predicted number of dog registrations approximately decreases by 0.96%.

3.2.2. Generalized Linear Model: dog count over time

As a second implementation of Generalized Linear Models, we return to the previously introduced research question: *how do dog counts evolve over time?*





3.3. Generalized Linear Model (Binomial)

We introduction our binomial GLM section by defining the following goal: *predicting a dog's sex based on its age*. In order to work with such models, the response variable must first be transformed to a 0-1 response, as it is originally coded as either 1 for male or 2 for female. Our objective is to assess if there is any significant relationship between dog age and the likelihood of it being either male or female.

```
glm_dog_sex_age <- glm(DogSexCd ~ DogAgeGroupCd, family = binomial, data = df_EN)
summary(glm_dog_sex_age)
```

```
##
## Call:
## glm(formula = DogSexCd ~ DogAgeGroupCd, family = binomial, data = df_EN)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.016374   0.010949  -1.495   0.1348
## DogAgeGroupCd  0.003459   0.001351   2.561   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98381  on 70966  degrees of freedom
## Residual deviance: 98368  on 70965  degrees of freedom
## AIC: 98372
##
## Number of Fisher Scoring iterations: 4
```

```
exp_coef <- exp(coef(glm_dog_sex_age))
percentage_change <- (exp_coef - 1) * 100
percentage_change
```

```
##   (Intercept) DogAgeGroupCd
##    -1.6240642     0.3464974
```

The coefficient for `DogAgeGroupCd` is 0.0035 with a p-value of 0.0104, indicating statistical significance. For every year increase in dog age, the probability of being female increase by 0.35%, holding other variables constant. While the statistical significance of the coefficient for a dog's age suggests that there is evidence to support the relationship between dog age and the probability of being female, the practical significance of a 0.35% increase in odds may not be substantial enough to warrant immediate business decisions based solely on this finding.

Further analysis may be needed here. Although the hypothesis was to offer products tailored to the dog's age and Sex, and while age appears to influence sex likelihood, additional factors such as breed and size would provide a more sound business decision making.

So more investigation into factors would refine predictions. We look at dog breeds.

Breed	Unknown	Chihuahua	Labrador Retriever	Yorkshire Terrier	Jack Russel Terrier
Count	9095	4828	4198	2709	2579

The most popular dog is a unknown ie. mixed dog breed. So Pet Paradise must offer mixed breed foods and products. Not only purebred product offerings.

So instead, another sales approach. Let's say, Pet Paradise is trying to target specific dog or owner age groups for marketing or sales purposes. Pet Paradise wants to predict the likelihood of a pet owner in their 40s owning a top 5 breed (e.g., the Chihuahua) compared to owning an unknown breed. For this, a binomial logistic regression makes most sense to use. This is because the response variable is binary: either a pet owner owns a Chihuahua (let say, coded as 1) or they own an unknown breed (we will code this as 0).

In terms of identifying popular breeds based on age and sex, Pet Paradise wants to use a model's predictions to optimize inventory management by stocking up on products that are likely to be in higher demand based on the popularity of a given pedigree breed.

```
chihuahua_binomial <- glm(ChihuahuaOwned ~ 1, family = binomial, data = owner_40s_df)
summary(chihuahua_binomial)
```

```
##
## Call:
## glm(formula = ChihuahuaOwned ~ 1, family = binomial, data = owner_40s_df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.29617    0.03527  -36.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4960.2  on 4766  degrees of freedom
## Residual deviance: 4960.2  on 4766  degrees of freedom
## AIC: 4962.2
##
## Number of Fisher Scoring iterations: 4
```

```
chihuahua_exp_coef <- exp(coef(chihuahua_binomial))
chihuahua_exp_coef
```

```
## (Intercept)
##   0.2735773
```

```
chihuahua_percentage_change <- (chihuahua_exp_coef - 1) * 100
chihuahua_percentage_change
```

```
## (Intercept)
##   -72.64227
```

We see that the log-odds of a 40-year-old owning a Chihuahua are -1.29, which is statistically significant. The exponentiated coefficient is 0.27, indicating that the odds of a 40-year-old owning a Chihuahua are 27%. If Pet Paradise targets typically well-earning professionals, i.e., adults in their 40s, and since the likelihood of a dog owner in their 40s owning a Chihuahua is relatively low, Pet Paradise diversify marketing efforts away from pedigree focus and rather highlight a broader range of mixed-breed foods, fur-shampoos and other products. This strategy will help attract well-earning customers who may own mixed or different breeds.

To check a last potential business case, we can check the reverse; predict which owner age group is more likely to own a Chihuahua, because we have a binary variable (ChihuahuaOwned: 1 for ownership and 0 for no ownership) and the remaining predictor variables (PrimaryBreed, DogAgeGroupCd, DogSexCd, and OwnerAgeGroupCd). We will try o predict which owner age group is more likely to own a Chihuahua.

```
chihuahua_age_bracket <- glm(formula = ChihuahuaOwned ~ PrimaryBreed + DogAgeGroupCd + DogSexCd + OwnerAgeGroupCd, family = b
inomial, data = owner_40s_df)
summary(chihuahua_age_bracket)

# Collinearity because model didn't converge. We calculate variance inflation factors
vif_values <- car::vif(chihuahua_age_bracket)
vif_values
```

The VIF produced by the above code shows that the model has perfect multicollinearity. It creates in linear dependencies among predictor variables. What this means, is that when there are categorical variables with many levels, it leads to unreliable predictions. So, if Pet Paradise is trying to predict which customers want to buy a product based on their age braket, breed of dog, etc, such a model with these singularities might suggest targeting certain groups of customers when, in fact, the data is too ambiguous to make such recommendations confidently We will compare the first two of the three models to determine the better fit instead.

```
cat("Model 1 - Dog Sex vs. Age:\n")
```

```
## Model 1 - Dog Sex vs. Age:
```

```
cat("AIC:", aic_model1, "\n")
```

```
## AIC: 98372.3
```

```
cat("BIC:", bic_model1, "\n\n")
```

```
## BIC: 98390.64
```

```
cat("Model 2 - Chihuahua Ownership in 40s:\n")
```

```
## Model 2 - Chihuahua Ownership in 40s:
```

```
cat("AIC:", aic_model2, "\n")
```

```
## AIC: 4962.163
```

```
cat("BIC:", bic_model2, "\n")
```

```
## BIC: 4968.632
```

We see from the Akaike Information Criterion and Bayesian Information Criterion that the second model, i.e. Chihuahua Ownership in an owner's 40s, has much lower AIC and BIC values than the Dog Sex vs. Age model. So model 2 provides a better fit to the data. Because the models contain fewer than 2 terms, checking collinearity doesn't make sense. We check the confusion matrices instead.

```
# Confusion matrix for Dog Sex vs. Age
predicted_probabilities <- predict(glm_dog_sex_age, type = "response")
predicted_classes <- ifelse(predicted_probabilities > 0.5, 1, 0)
confusion_glm1 <- table(df_EN$DogSexCd, predicted_classes)
confusion_glm1
```

```
##    predicted_classes
##         0     1
##   0 15813 19596
##   1 14672 20886
```

The first confusion matrix shows that there are correctly predicted the negatives was 15,813 instances. The false positive representing incorrectly predicted the positive cases was 19,596 false positives. There were 14,672 false negative predictions and 20,886 true positives. This means the first binomial model has an accuracy of 52% and a precision of 52%. The Sensitivity was around 59%, where of all actual positive instances were correctly identified by the model.

The client question of this analysis was to develop and compare binomial logistic regression models to predict outcomes regarding dog ownership based on given variables. We examined on two main models: one predicting the sex of the dog based on its age group and another predicting the most popular pedigree ownership among dog owners in their prime income earning years in their 40s.

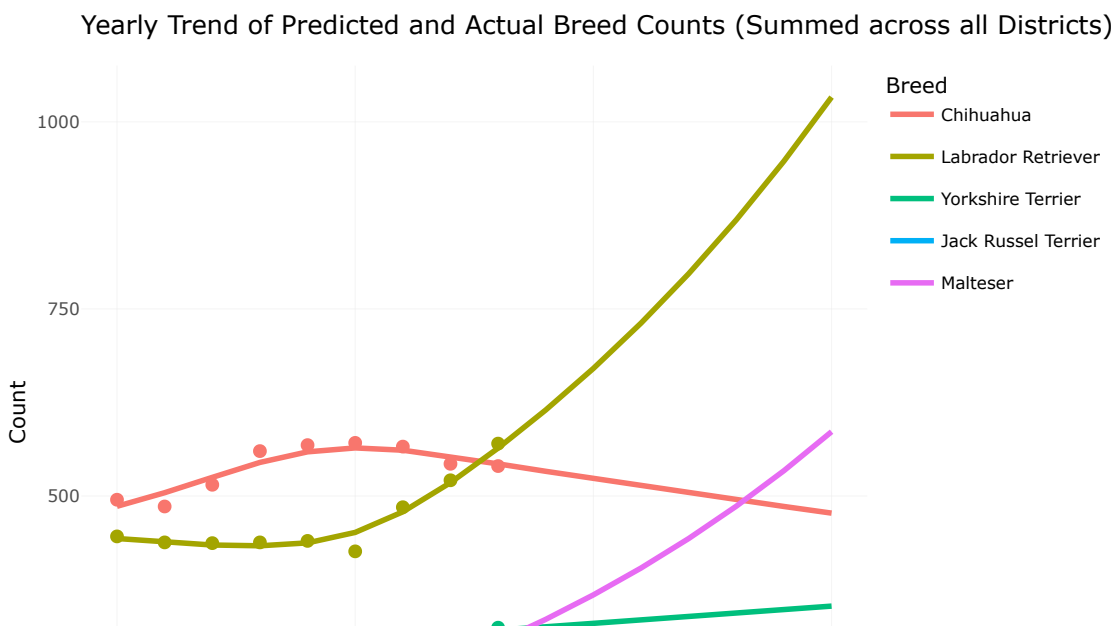
The first model offered insights into the relationship between dog age and sex but had limited predictive power. The second model provided a baseline probability of pedigree ownership based on created binary variables of ownership vs non-ownership to target a very specific demographic niche. While attempting to include multiple predictors in the third pedigree ownership model, perfect multicollinearity was detected, resulting in unstable estimates and difficulties in model convergence. For Pet Paradise, focusing on targeting broader product ranges to well-earning professionals, rather than narrowly focusing on pedigree breeds, could be more effective.

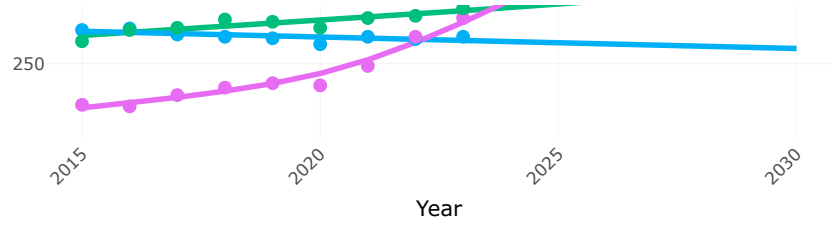
3.4. Generalized Additive Model

3.4.1. GAM model: evolution of popular breeds

To assess Pet Paradise on the question of *evolution of popular dog breeds* we will employ Generalized Additive Models (GAM). Such models can capture the non-linear patterns inherent in the popularity fluctuations of various dog breeds throughout the years. This analysis will enable Pet Paradise to make informed predictions about future trends, thereby facilitating their ability to anticipate the demand for breed-specific products.

First we have a look at all districts together.

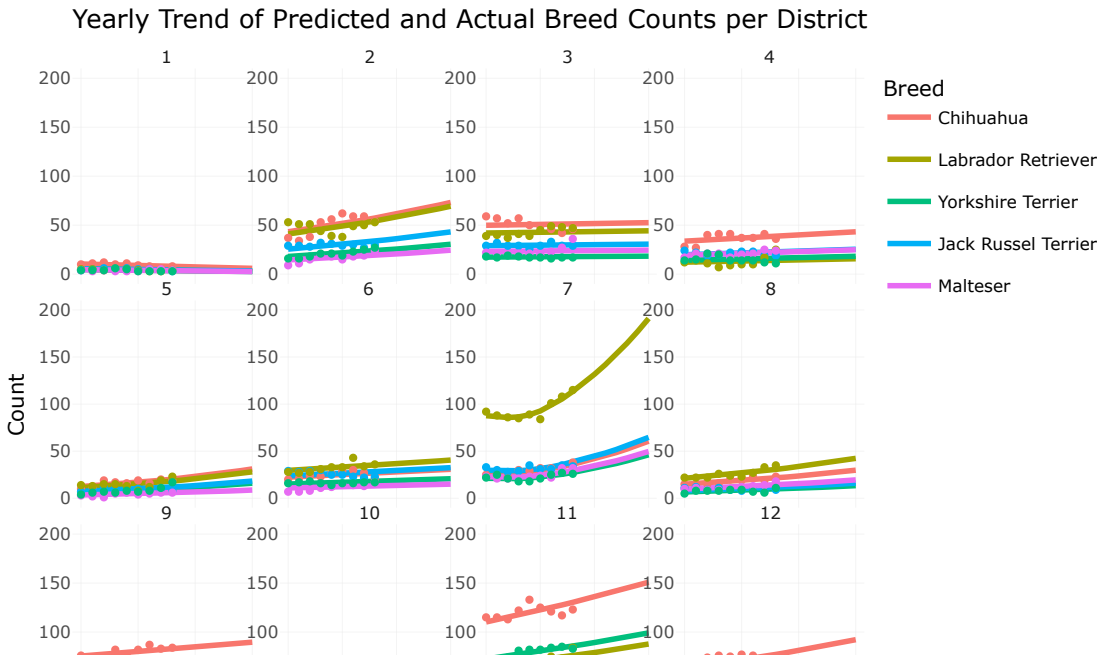


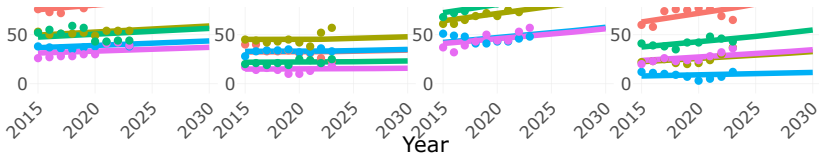



```
##
## Family: poisson
## Link function: log
##
## Formula:
## BreedCount ~ te(KeyDateYear, by = PrimaryBreed, k = 8) + s(PrimaryBreed,
##     bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5.8495     0.1577   37.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                              edf Ref.df   Chi.sq p-value
## te(KeyDateYear):PrimaryBreedChihuahua      2.324  2.877   12.554 0.00638
## te(KeyDateYear):PrimaryBreedJack Russel Terrier 1.000  1.000    0.513 0.47390
## te(KeyDateYear):PrimaryBreedLabrador Retriever  2.647  3.263   37.604 < 2e-16
## te(KeyDateYear):PrimaryBreedMalteser       1.972  2.448   52.404 < 2e-16
## te(KeyDateYear):PrimaryBreedYorkshire Terrier   1.000  1.000    3.371 0.06637
## s(PrimaryBreed)                            3.989  4.000 1597.524 < 2e-16
##
## te(KeyDateYear):PrimaryBreedChihuahua      **
## te(KeyDateYear):PrimaryBreedJack Russel Terrier
## te(KeyDateYear):PrimaryBreedLabrador Retriever  ***
## te(KeyDateYear):PrimaryBreedMalteser       ***
## te(KeyDateYear):PrimaryBreedYorkshire Terrier   .
## s(PrimaryBreed)                            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.994   Deviance explained = 99.6%
## -REML = 209.97  Scale est. = 1         n = 45
```

Next we direct our attention to each individual city district.

```
## 'data.frame':    540 obs. of  4 variables:
##  $ KeyDateYear : num  2015 2015 2015 2015 2015 ...
##  $ PrimaryBreed: Factor w/ 394 levels "$Labradoodle$",..: 94 94 94 94 94 94 94 94 94 94 ...
##  $ DistrictSort: Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ BreedCount  : int  10 37 59 28 10 20 25 15 76 40 ...
```





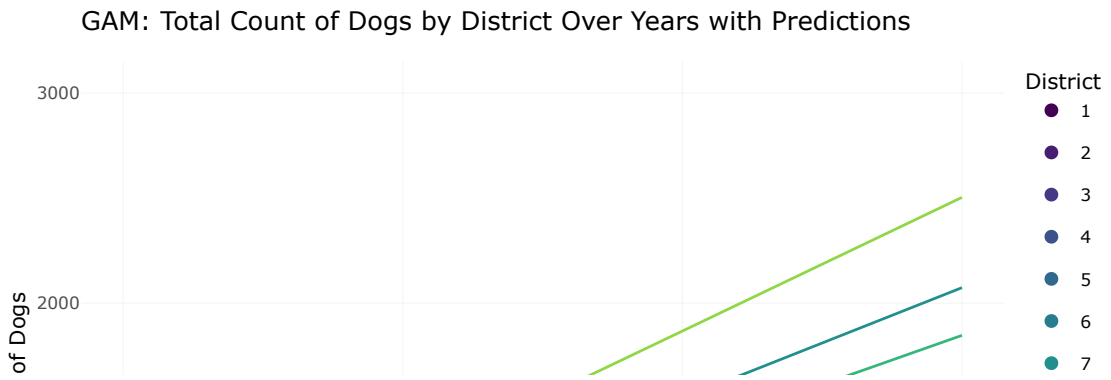

```
##
## Family: poisson
## Link function: log
##
## Formula:
## BreedCount ~ s(KeyDateYear, bs = "cr", k = 5) + s(PrimaryBreed,
##     bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.5962     0.1788   8.926   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                  edf Ref.df Chi.sq p-value
## s(KeyDateYear)  1.00  1.001  2.144   0.143
## s(PrimaryBreed) 3.42  4.000 32.862  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.822   Deviance explained = 81.9%
## -REML = 89.723  Scale est. = 1         n = 45
```

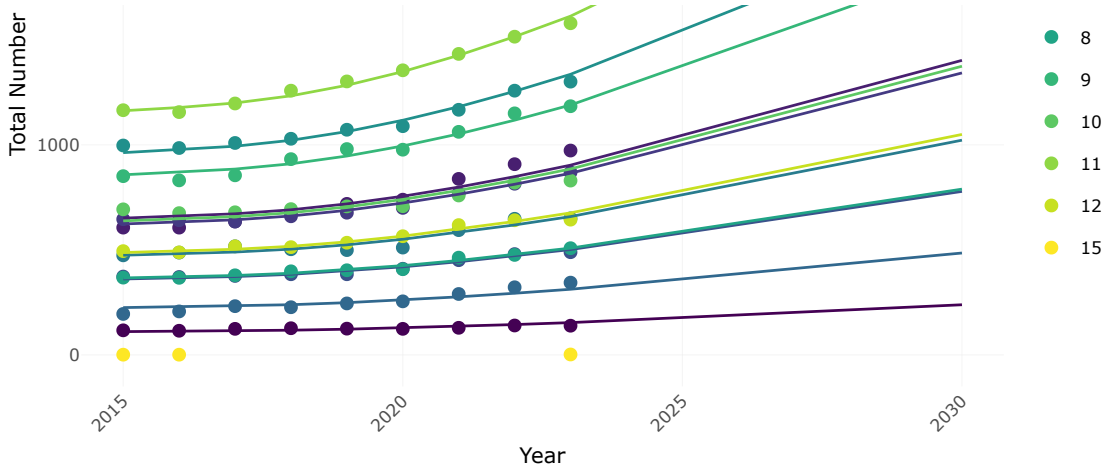
3.4.2. GAM model: dog count over time

For the second part of the Generalized Additive Model chapter we again direct our attention to the previously posed research question: *how do dog counts evolve over time?*

As a straightforward implementation of the GAM model, we set ourselves to produce a regression of the registered count data, grouped by districts of the city, analyze their evolution over time, and additionally provide predictions for the next 10 years.

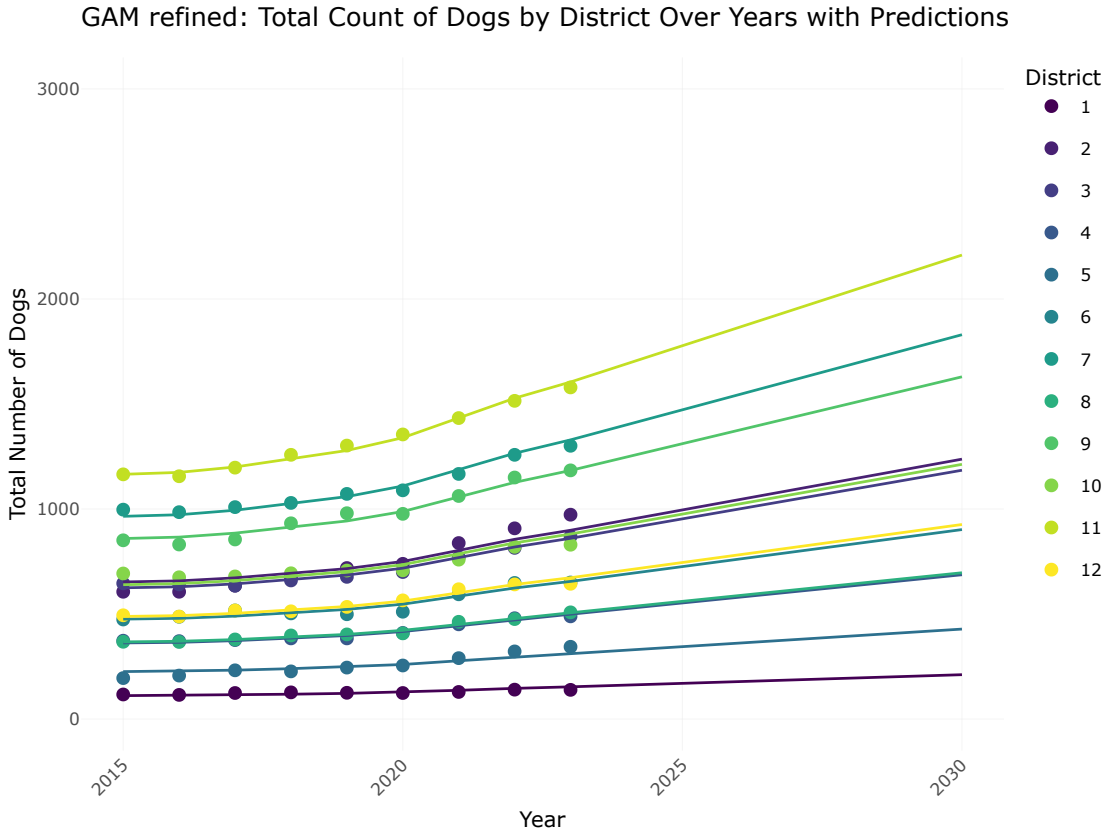
```
##
## Family: poisson
## Link function: log
##
## Formula:
## TotalDogs ~ s(KeyDateYear, bs = "cr", k = 4) + s(DistrictSort,
##     bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5.8853     0.4767   12.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                    edf Ref.df  Chi.sq p-value
## s(KeyDateYear)   2.719  2.939   911.6  <2e-16 ***
## s(DistrictSort) 11.948 12.000 16051.0  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.996   Deviance explained = 99.7%
## -REML = 555.44  Scale est. = 1         n = 111
```





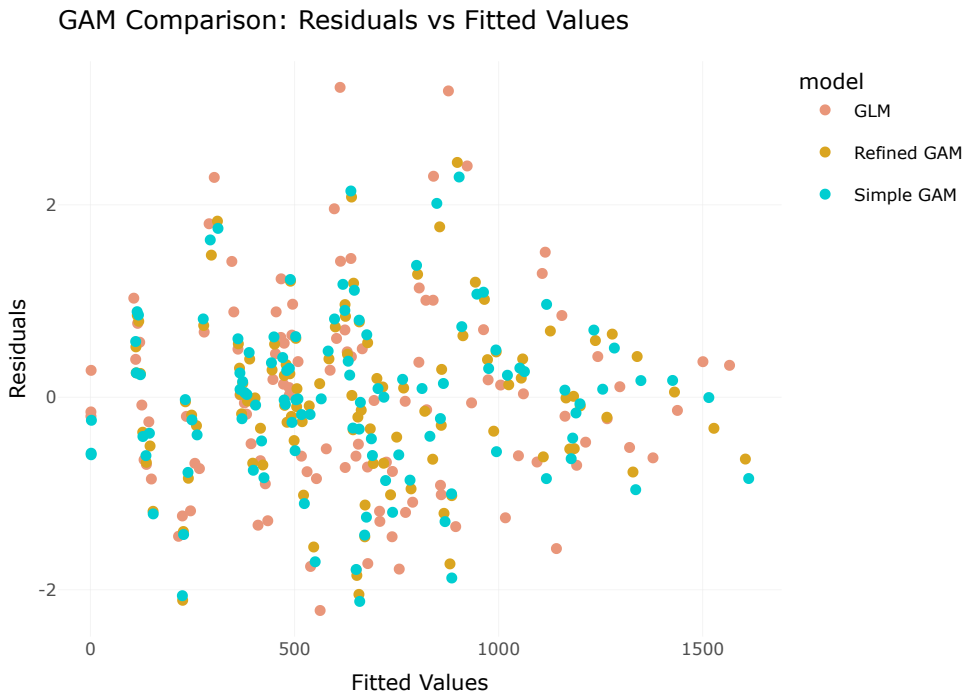
We now proceed to further omptimize the GAM model.

```
##
## Family: poisson
## Link function: log
##
## Formula:
## TotalDogs ~ s(KeyDateYear, bs = "cr", k = 8) + s(DistrictSort,
##     bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    6.322      0.189   33.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df Chi.sq p-value
## s(KeyDateYear)   4.053  4.881    913  <2e-16 ***
## s(DistrictSort) 10.994 11.000  15837  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.996   Deviance explained = 99.6%
## -REML =  538.2  Scale est. = 1         n = 108
```



3.4.2. Model comparison

We now look at some comparison now between the above two models, as well as the GLM that was introduced earlier.




```
# Extract and plot the effect of time for a few districts
library(effects)
for(district in unique(combined_data$prediction_count)[1:12]) {
  effect_data <- effect("KeyDateYear", gam_model_refined, xlevels = list(DistrictSort = district))
  effect_df <- as.data.frame(effect_data)

  ggplot(effect_df, aes(x = KeyDateYear, y = fit)) +
    geom_line(shade = TRUE, size = 1) +
    geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2) +
    labs(title = paste("Effect of Year in District", district),
         x = "Year", y = "Fitted TotalDogs Count") +
    theme_minimal()
}
```

```
aic_comparison <- AIC(gam_model_simple, gam_model_refined)
print(aic_comparison)
```

3.5. Artificial Neural Network

We start this chapter by posing the research question: *is a dog of pure or mixed breed, based on location and owner and dog's characteristics?*

An artificial neural network (ANN) is the algorithm of choice for this assessment due to its capability to handle complex, non-linear relationships within the dataset. Dogs' breed may depend on various interacting factors, such as age, size, location, and owner demographics, which ANNs can effectively capture and analyze.

Their scalability and adaptability make them robust for ongoing studies, while their multilayered learning allows the model to automatically identify the most significant features. We ultimately aim to accurately predict breed status and provide valuable insights for our suggested business project, as well as pet adoption agencies, veterinarians, or even urban planners of the city of Zurich.

Continuing with the development of models, and answering the above, we now look into the implementation of an Artificial Neural Network using the packages `nnet` and `caret` in R. The resulting classification result for the dependent variable will be produced by the model considering variables of both numerical and categorical type.

The target variable `MixedBreed` is of categorical type (factor in R), indicating the dog's pedigree status, with 4 different possible responses, from pure breed to 3 different descriptors of breed mixing. For the sake of simplicity, the levels within the `MixedBreed` factor variable have been reduced to a binary response, indicating whether the dog is of pure pedigree or not. We consider that this response better fits the information needed for our business case.

As explained above, the response variables of choice pertain to characteristics of the dog owners (their age, sex, and district), as well as some characteristics of the dogs (dog age and sex). The predictors are the following: `OwnerAgeGroupCd`, `OwnerSexCd`, `DistrictCd`, `DogAgeGroupCd` and `DogSexCd`. Although the model results suggest that incorporating additional predictors should be done to improve the model's accuracy, we have kept the current selection for learning purposes.

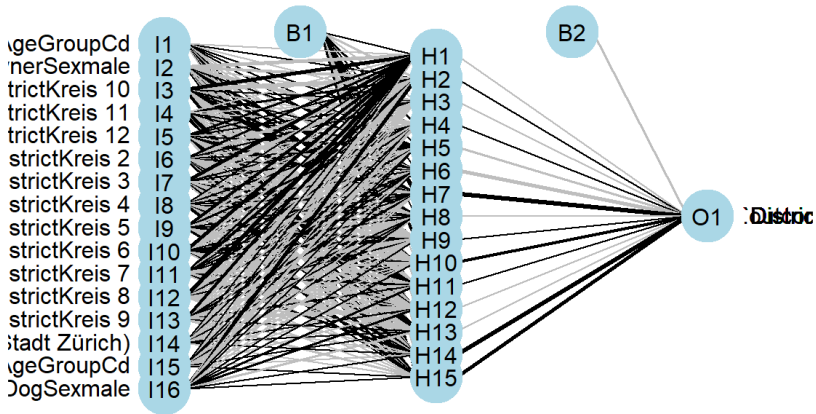
The libraries of choice are `nnet` and `caret`, the first one providing the functions for creating and training the model, and the latter providing an interface to further preprocess and train the model. This allowed to implement a 10-fold cross-validation during the training phase of the model building. It also allowed to establish a tune grid with various possible hyperparameters to test out different combinations and find the most optimal ones.

```
breed_net
```

```
## Neural Network
##
## 56774 samples
##     5 predictor
##     2 classes: 'Mixed breed', 'Pedigree dog'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 51096, 51096, 51097, 51097, 51096, 51097, ...
## Resampling results across tuning parameters:
##
##   size  decay  Accuracy   Kappa
##    5    1e-04  0.7144468  0.0004926028
##    5    1e-03  0.7145877  0.0005098842
##    5    1e-02  0.7142530  0.0014752396
##   10    1e-04  0.7144468  0.0023939219
##   10    1e-03  0.7143235  0.0046166926
##   10    1e-02  0.7146406  0.0059293934
##   15    1e-04  0.7138479  0.0035546704
##   15    1e-03  0.7146581  0.0088259620
##   15    1e-02  0.7144467  0.0082564278
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were size = 15 and decay = 0.001.
```

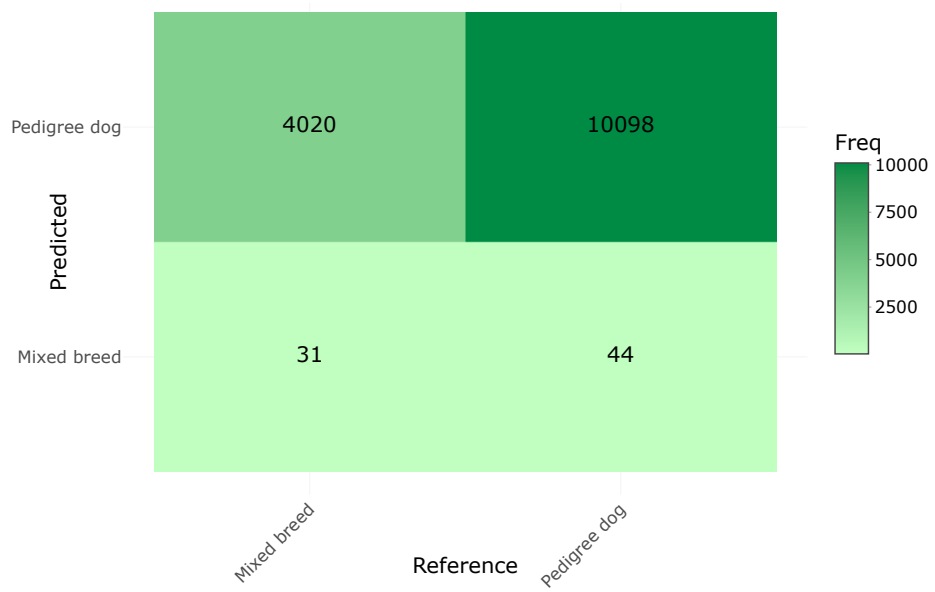
```
breed_net$finalModel
```

```
## a 16-15-1 network with 271 weights
## inputs: OwnerAgeGroupCd OwnerSexmale `DistrictKreis 10` `DistrictKreis 11` `DistrictKreis 12` `DistrictKreis 2` `DistrictK
reis 3` `DistrictKreis 4` `DistrictKreis 5` `DistrictKreis 6` `DistrictKreis 7` `DistrictKreis 8` `DistrictKreis 9` `District
Unknown (Stadt Zürich)` DogAgeGroupCd DogSexmale
## output(s): .outcome
## options were - entropy fitting  decay=0.001
```

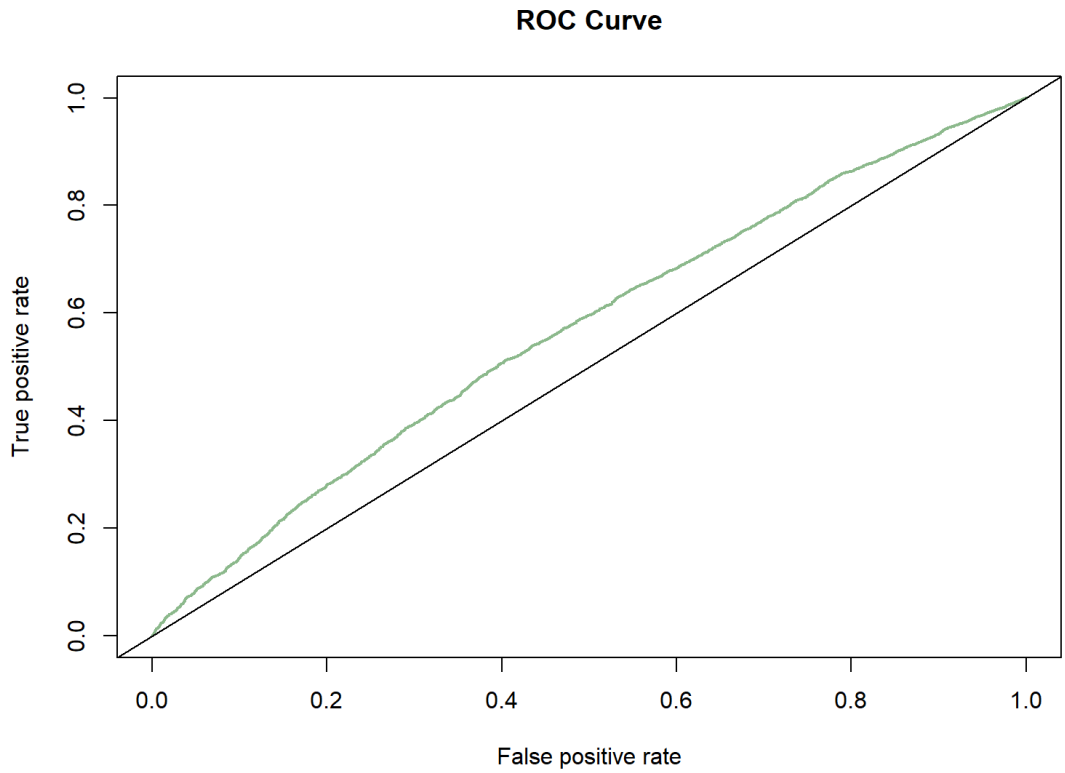


After running the code, the final model is a neural network with 16 input nodes, 15 hidden nodes, and 1 output node, totalling 271 weights. However, an evaluation of the confusion matrix and ROC curves indicates insufficient evidence to support the model's validity for the given variables. This issue might stem from the selection of variables; expanding the set to include more variables from the dataset might improve the model. Alternatively, it could be that an artificial neural network is not the most suitable model for addressing this particular research question.

Binary Confusion Matrix

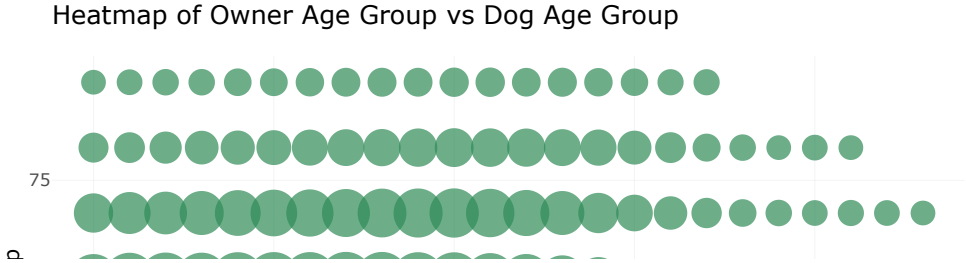


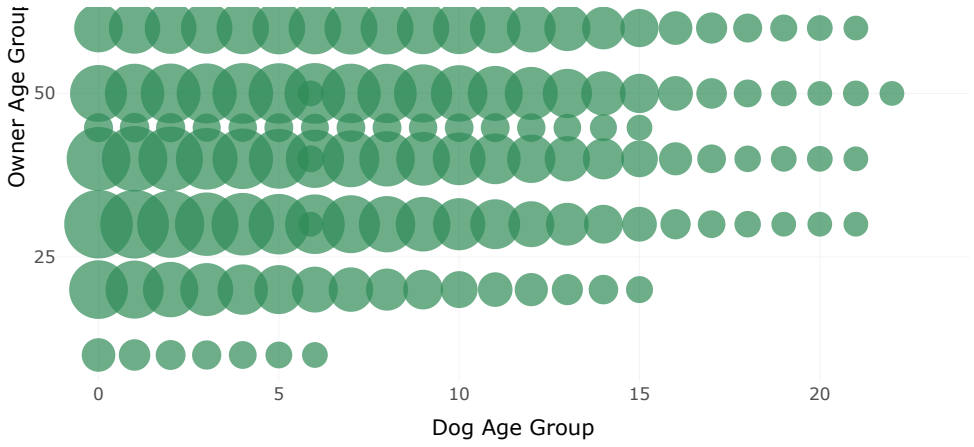
After building the model with an 80% training subset of the total dataset, we use it to predict values for the remaining 20%. We evaluate the model's performance using a confusion matrix, which allowe us to compare our predicted values with the actual values of the test subset. The model successfully classifies 10,098 dogs as pure-bred, with only 44 being misclassified. However, this good result is overshadowed by the misclassification of 4,020 mixed-breed dogs, with only 31 correct predictions. This suggests that the model wrongly tends to classify most dogs as pure-bred.



The *ROC Curve* displays a closeness to the diagonal, which raises the possibility that the model's predictions are not accurate, and support the need to either review the model with more predictors or consider a different model to answer the question at hand altogether.

3.6. Support Vector Machine Model

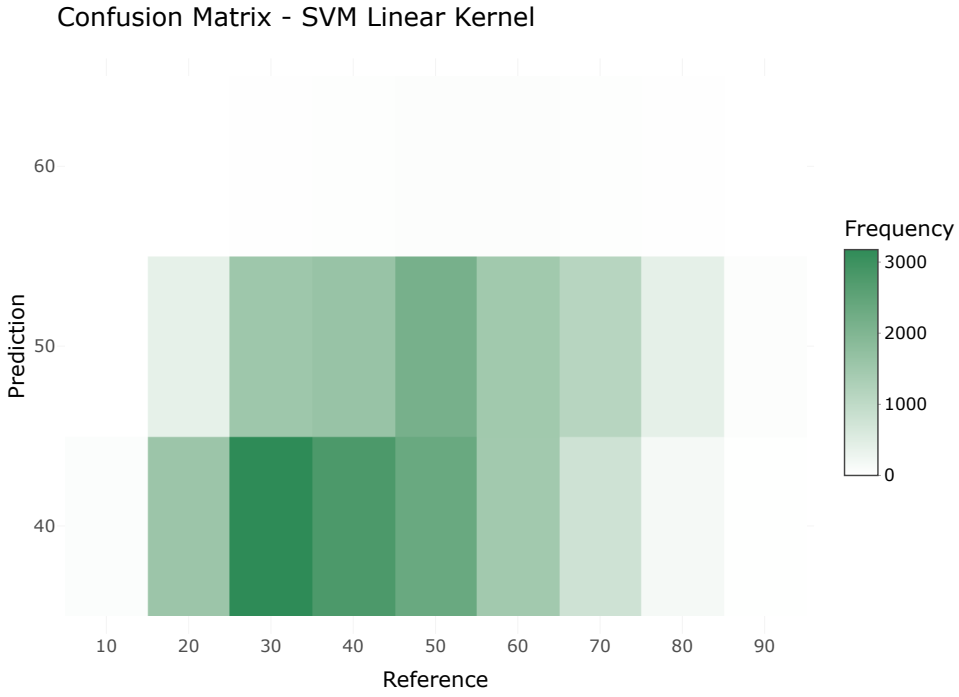





```
svm_linear
```

```
## Support Vector Machines with Linear Kernel
##
## 49673 samples
##     1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 44705, 44707, 44705, 44705, 44705, 44705, ...
## Resampling results:
##
##   RMSE      Rsquared    MAE
##   15.07558  0.09038973  12.39084
##
## Tuning parameter 'C' was held constant at a value of 1
```

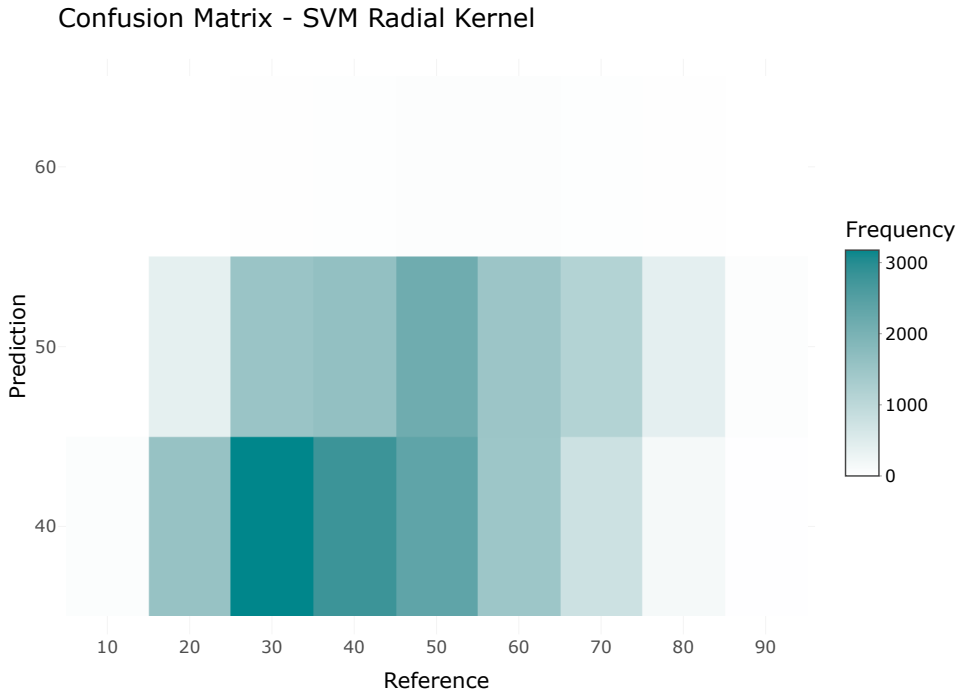
The predicted results have been adjusted and rounded to their nearest multiple of 10, in order to be able to visualize them against the actual decade time ranges represented in the original data. The following confusion plot shows the frequency of guesses for each age range.



At a first glance, most of the predictions happen around the 30 age range, specifically misclassifications where the model predicts 40. This is an acceptable error, as it reflects the observations we have previously made that most dog owners belong to those two age groups. No predictions were made for ages outside of the 40-60 range in this case.

```
svm_rbf
```

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 49678 samples
##     1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 44711, 44710, 44708, 44710, 44711, 44710, ...
## Resampling results across tuning parameters:
##
##   C       RMSE      Rsquared    MAE
##     0.25  15.12327  0.08196042  12.38584
##     0.50  15.12376  0.08191171  12.38624
##     1.00  15.12445  0.08184675  12.38678
##     2.00  15.12475  0.08181964  12.38697
##     4.00  15.12485  0.08180993  12.38701
##     8.00  15.12518  0.08177908  12.38716
##    16.00  15.12528  0.08177042  12.38727
##    32.00  15.12531  0.08176758  12.38733
##    64.00  15.12531  0.08176760  12.38734
##   128.00  15.12529  0.08177019  12.38737
##
## Tuning parameter 'sigma' was held constant at a value of 8.77719
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were sigma = 8.77719 and C = 0.25.
```



4. Additional chapter

5. Conclusion

6. Appendix

6.1. Working with generative AI tools

7. References

Pospischil et al. 2013. *Hundepopulation Und Hunderassen in Der Schweiz von 1955 Bis 2008*. Band 155. https://doi.org/10.1024/0036-7281/a000450 (https://doi.org/10.1024/0036-7281/a000450).

Regierungsrat. 2009. “Hundeverordnung.” *Stadt Zürich*. https://www.zh.ch/content/dam/zhweb/bilder-dokumente/themen/umwelt-tiere/tiere/veterinaeramt/hunde/publikationen/erlaeuterungenzuhuv.pdf (https://www.zh.ch/content/dam/zhweb/bilder-

dokumente/themen/umwelt-tiere/tiere/veterinaeramt/hunde/publikationen/erlaeuterungenzuhuv.pdf).
Stadt Zürich. 2024. “Hundebestände Der Stadt Zürich, Seit 2015.” Opendata.swiss. https://opendata.swiss/de/dataset/hundebestande-der-stadt-zurich-seit-2015/resource/5f8eafd2-367f-489c-a075-42426d14c586 (https://opendata.swiss/de/dataset/hundebestande-der-stadt-zurich-seit-2015/resource/5f8eafd2-367f-489c-a075-42426d14c586).
Statistik Stadt Zürich. 1984. “Hunde Und Hundebesitzer in Der Stadt Zürich.” *Stadt Zürich*. https://statistik.stadt-zuerich.ch/modules/StatNat/1984/1984_ZSN_Hunde-und-Hundebesitzer-in-der-Stadt-Zuerich.pdf (https://statistik.stadt-zuerich.ch/modules/StatNat/1984/1984_ZSN_Hunde-und-Hundebesitzer-in-der-Stadt-Zuerich.pdf).