Final Report

We selected a dataset that shows the characteristics of dogs and their owners in Zurich, Switzerland. It contains various dog breeds and shows a number of variables such as descriptions pertaining to the owner of the dog, as well as more characteristics of the dogs. With this data being from Switzerland, all of the words were not in English so for the first steps in processing the data we translated it to terms we could understand. Our idea for this project is to gain a better understanding of the makeup of dogs and their owners living in Zurich. The motivation behind this project stems from wanting to know about dog trends. We also think it would be interesting to see if there are any correlations between dogs and their owners due to the myth that dogs look like their owners. Analyzing this dataset is a good first step in answering these questions because it provides a wide range of variables that were collected for thousands of samples.

For the first question, we wanted to know how the variables of the dataset influenced the gender of the dog. Initially, we wanted to look into how accurate of an estimator the owner's gender was for predicting their dog's gender. After plotting the count of male and female dog owners, we filled the plots by the proportion of female and male dogs (Figure 1). This plot shows a pretty even 50/50 divide among the gender of pets for both male and female owners. To find numerical evidence to support this we performed a logistic regression since it allows us to estimate the probabilities of events for categorical data. The model created was found to accurately predict the data correctly only 52.5% of the time. We then decided to try again by performing the same analysis two more times, but replacing the owner's gender with primary breed and then birth year. Both of these models also failed to be an accurate predictor of dog's gender. Although these are not the results we were initially expecting, it does tell us about the data as a whole. In the population we would expect the true proportion of males to be 0.5 and

that seems to be accurate with our models. Since the models consistently seem to predict about 50% of the data correctly that shows that dog's gender is a true random variable. Collecting the 7000 samples enabled this dataset to accurately reflect the true proportion of dog's gender.

For the second question, we began by analyzing the two parts of the question separately, starting with, "what is the most popular primary breed?" After the cleaning process, we made a pivot table of all the breeds represented so we could see the overall total that we were working with, coming out to 304 unique breeds. Ultimately, this question was purely categorical so we decided to use bar charts to show how the breeds compare to each other. Then, the data was loaded into R studio to create visualizations for the data.Since 304 breeds made the data set too big to read in a graph axis, we made the executive decision to use the pivot table setting to find the top ten breeds and created a bar graph to display the distribution of only the top ten breeds (Figure 2). Based on this analysis, we can see that the most popular primary breed is Mischling Klein. Next, we analyzed the second part of the question: "What primary breed is the most popular to be mixed?" Again, the data set was too large so we used the pivot table function but filtered out the primary breed row and sorted it by the number of hybrids in each breed in ascending order and used another bar chart to show the distribution of the top ten primary breeds with the largest amount of hybrids (Figure 3). From that, we found that the most common primary breeds to be classified as mixed are the Labrador Retriever and Terrier.

For the third question we wanted to find whether there was a correlation between the ages of dogs and their owners. For this question we focused on the Owner Age column and the dog age column, where 20 rows contained random data out of the 7155 rows. The dog ages came in Birth Year format, so the ages were calculated where a new Age column was made. This cleaned data was then transferred to R Studio where a histogram, barplots, and boxplots

were created. A histogram was made to visualize the age distributions of the dog owners (Figure 4). The histogram was right skewed and the distributions lied between 7 and 14, which indicated that most of the dogs were adults. A barplot to view the distributions of the dogs (Figure 5), which also indicated that most of the dogs were of adult status. Boxplots were utilized to see whether there was a correlation between the ages of the dogs and their owners. The boxplots demonstrated that there was a positive correlation between the ages of dogs and their owners, given that the boxplots were incrementing by the decade. For owner ages between 31-50 the median age for their dogs were both 10 years old (Figure 6).

For the fourth question, we researched the top dog breeds and colors sorted by district. Here there was a lot of cleaning involved, especially for the color column, and we ended up attempting to sort as many of the colors as possible into one of about 20 main color categories, though there were some stragglers. After creating pivot tables to sort the dog breeds and colors in each district, we found that the most popular dog breeds were the Mischling klein and Chihuahua, and the most popular dog color was black, followed by brown and white. We also looked at the distribution of percentages of top dog breeds across districts (Figure 8) and found that the top breed held a similar percentage (9.5%) across every district, though there were two outliers, one for each top breed. Additionally, we checked the distribution of top dog color across districts (Figure 9) - we found that there were no outliers, and the ordering of the top three dog colors was exactly the same across 9/12 of the districts.

Overall, from these analyses, we discovered that the sample proportion of dog's gender agrees with that of the true population; That the most popular dog breed in Zurich is Mischling Klein; The most popular mixed breeds are Labrador Retriever and Terrier; The age of the owner has a positive correlation with the age of the dog; And the most popular dog color was black. These all lead us to a better understanding of the demographics of dogs in Zurich. Additionally, it provides trends for Zurich which is most likely an accurate representation for all of

Switzerland. We would like to carry out these studies by including additional datasets from other parts of the world to get a more accurate overall idea of dog demographics.

Figures

Figure 1: This boxplot shows the distribution of the dog owner's gender in Zurich, Switzerland and then whether their dog is male or female.

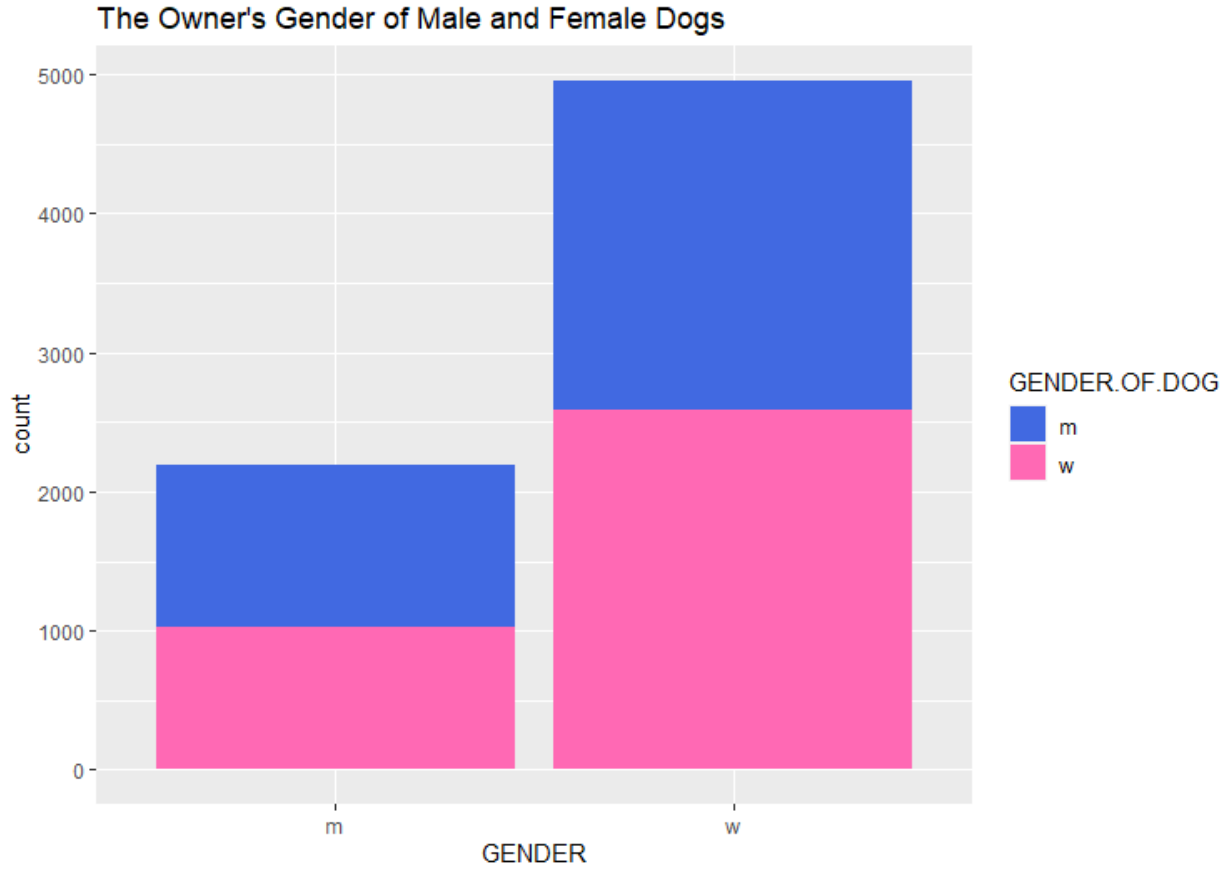


Figure 2: This bar chart displays the counts of the top ten breeds from this data set.

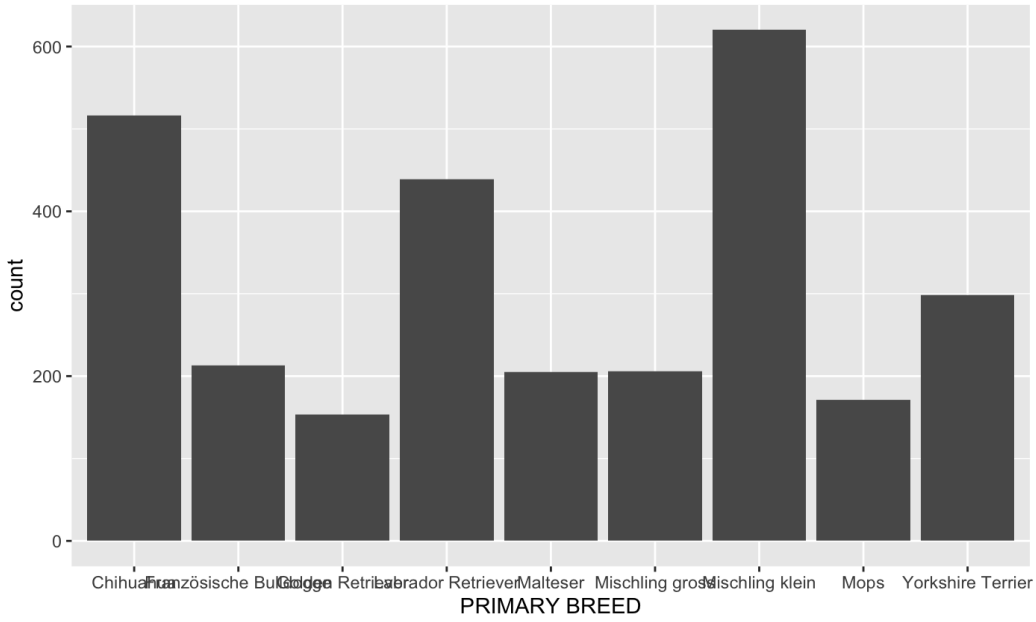


Figure 3: This bar chart displays the top counts of hybrids

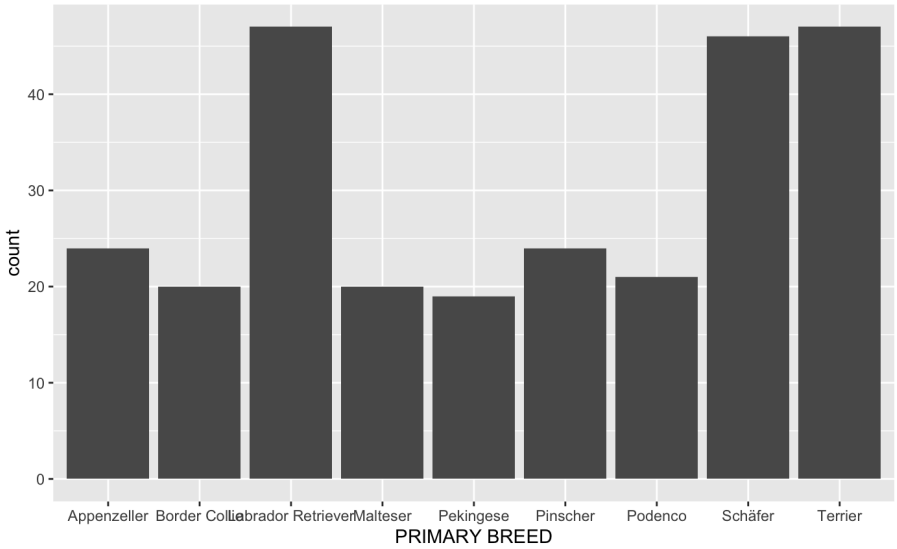


Figure 4: A histogram displaying the Dog Ages

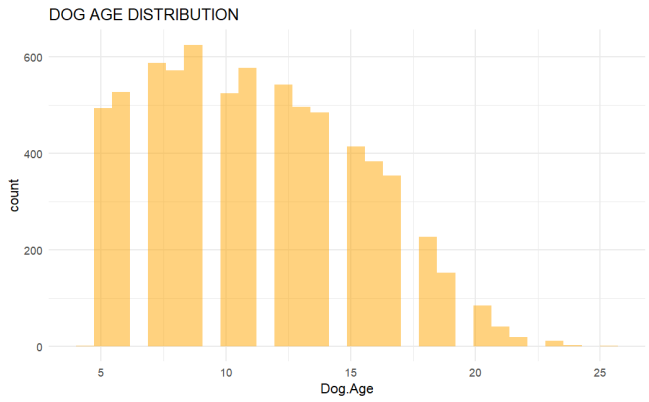
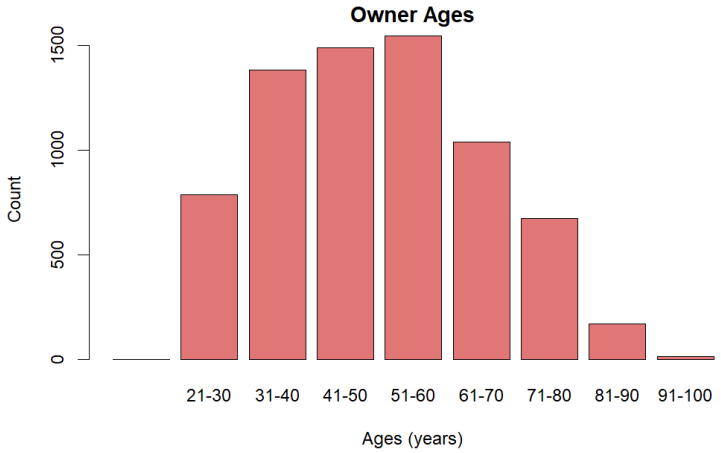


Figure 5: A bar chart displaying the Owner Ages



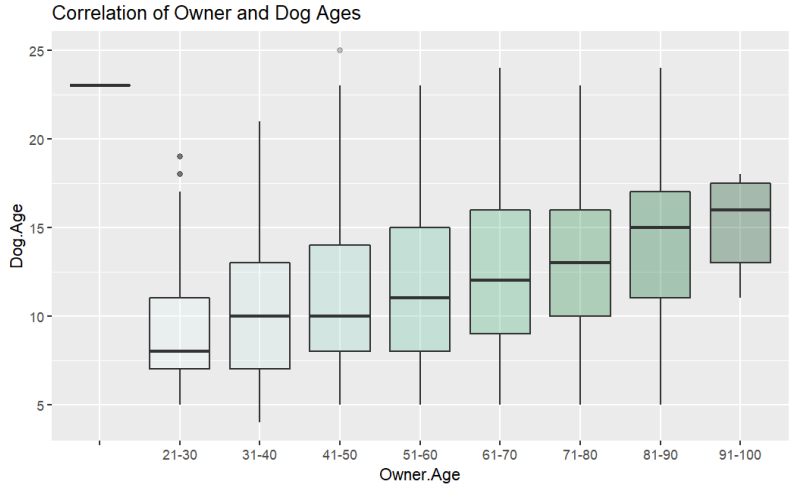


Figure 7: Boxplots displaying a positive correlation between Dog and Owner Ages.

Figure 8: The five number summary for the distributions of top dog breeds across districts.

```
> summary(mischling)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.200   8.800   9.300   9.414   9.700  11.400
> summary(chihuahua)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.20    9.50    9.80   10.52   10.60   14.50
```

Figure 9: The five number summary for the distributions of top dog colors across districts.

```
> summary(color$schwarz)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   36.0   114.8   157.0   177.2   229.2   368.0
> summary(color$braun)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   19.0    70.5   106.5   104.8   134.0   216.0
> summary(color$weiss)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.00   58.50   90.50   90.25  111.25  185.00
```