UNIVERSITY of WASHINGTON

Final Report: Adopted Pets Analysis

Team Members:
Aviva Munshi
Jenny Wong
Harshita Maddi
Debbie Davis

MS Data Science 2022 Cohort

BIOST 557 A Winter 2022: Applied Statistics & Experimental Design

Dated: March 15, 2023

Section 1: Abstract

The aim of this study was to investigate the factors that influence pet adoption and length of stay in shelters/foster homes in the United States, focusing on dogs and cats listed on PetFinder. The study also aimed to determine the impact of certain traits and location-based factors on the number of adopted pets. We used three datasets and examined three regions to investigate four research questions. Statistical techniques, including z-tests, chi-square tests, t-tests, ANOVA tests, survival analysis, and linear regression models, were used to compare means and distributions of various variables. We ultimately found that animal traits such as size and gender significantly impacted the length of stay in shelters and time to adoption, while location-based factors such as population density and median household income were associated with the number of adopted pets in a region. Adopted dogs were found to have a higher life expectancy than the general population, although there was no relationship between the life expectancy of dogs in different regions, and the statistical tests rejected the null hypothesis for all pairs of categorical variables except for region vs. gender. Our linear regression model showed that location-based factors such as median household income and population density affected the number of adopted pets. We conclude that animal traits and location-based factors are important in the adoption of pets and that policy makers and animal shelters should consider these factors when developing adoption strategies.

Section 2: Introduction

According to the American Society for the Prevention of Cruelty to Animals (ASPCA), approximately 6.5 million companion animals enter animal shelters in the United States each year. Out of those, around 3.3 million are dogs and 3.2 million are cats. Unfortunately, not all these animals find homes. The ASPCA estimates that approximately 1.5 million shelter animals are euthanized each year. This number has decreased in recent years due to increased efforts to promote adoption and spaying/neutering, but it is still an alarming statistic. In addition to euthanasia, dogs and cats in shelters may experience other challenges, such as overcrowding, limited resources, and stress. Furthermore, the shelter environment can be stressful for animals, which can negatively impact their health and behavior. Understanding the factors that influence pet adoption and reduce shelter stay is crucial to improving the lives of dogs and cats in shelters. By identifying traits that make pets more adoptable and analyzing location-based factors, animal welfare organizations can work to increase the number of successful adoptions and reduce the time that pets spend in shelters.

The goal of this project is to explore the factors that influence pet adoption and reduce the time that animals spend in shelters. By analyzing the data obtained from the Petfinder API, we can identify traits that make pets more adoptable and location-based factors that impact the number of adoptable pets in different states. The findings from this study can inform animal welfare organizations and help them optimize their adoption processes to increase the number of successful adoptions and reduce the length of stay for pets in shelters.

Our hypotheses are:

- Of cats and dogs listed on PetFinder who were ultimately adopted, more dogs than cats were adopted in the US. Additionally, more dogs were adopted in the Pacific Northwest (PNW) region compared to the South Central (SC) region of the United States.

- The mean days spent by dogs on Petfinder before adoption are equivalent among gender and size groups, and can be tested by comparing the mean days of each group.
- The mean minimum life expectancy of dogs adopted in different regions of the United States is the same.
- The median income, population, and age of the state is correlated with the number of adopted pets.

Section 3: Methods

Data

In this section, we will describe the data that we collected for answering the research questions for this project. We found 4 data sources in total, and for each source, we will describe the collection method and potential limitations and biases within each dataset. For a more detailed explanation of the specific variables from each dataset used in this project, please reference the appendix.

Petfinder API (PetFinder API calls)

Petfinder API is the largest dataset used in this project, containing up-to-date data on adoptable pets from over 10,000 animal welfare organizations. It includes all adoptable dogs and cats worldwide, with a static sample collected on February 23rd, 2023, filtered for dogs and cats who were ultimately adopted in the US. We've used a python wrapper called Petpy to retrieve data into pandas DataFrames for analysis. (Aschleg, 2021)

There are some potential biases and limitations to using data from the PetFinder API. First, the data only includes pets that are registered on the Petfinder platform, which may not represent the entire population of dogs and cats available for adoption in the United States. Additionally, the data is self-reported by animal welfare organizations, and there may be errors or inconsistencies in the data due to differences in reporting methods across organizations. Finally, the dataset may not capture all relevant factors that influence pet adoption and shelter stay, and there may be unobserved factors that impact these outcomes.

AKC Data (reference) (tmfilho, 2020)

This dataset contains information for around 277 dog breeds and was extracted from the website of the American Kennel Club. (*American kennel club*) The dataset has features including height, weight, life expectancy, shedding value, energy level, trainability, and demeanor value. The limitation of the dataset is that it only has data for dogs but not cats.

Census Data: Income

To ensure the authenticity of the data, we collected the following two datasets from the US Census Bureau: Income and Population. The income dataset contains the median and mean income estimate for the year of 2021 for 4 types of living units: households, families, married-couple families, and nonfamily households, for all states in the US. According to the American Community Survey (ACS) methodology webpage, the sample size of the housing units completing final interviews is 1,950,832 for the US**.** (Bureau, *Sample size* 2022) The coverage rate is 98.5%. (Bureau, *Coverage rates* 2022)

There are several limitations to this dataset. The first issue is that since the population is big, this dataset is not the raw number of income of each household for every state. Instead, it is a calculated estimate value done by the ACS that exists a certain level of marginal error for every number. The second issue is, since all the money income is pretax, the actual income for every household or individual can be different from what it shows in the dataset, as different state has its own income tax policy.

Census Data: Population

The source of the population dataset is the DP05: ACS Demographic and Housing Estimates (2021) from the US Census Bureau. This dataset contains the total estimated population for the country and for each state. The sample size and coverage rate are the same as the income data due to the same source of data.

Statistical Methods

Regions

- *Pacific Northwest (PNW):* Oregon, Washington, Idaho
- *South Central (SC):* Kentucky, Tennessee, Alabama, Mississippi, Arkansas, Louisiana, Oklahoma, and Texas
- *Northeast (NE):* Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont

Research Question 1

Are more dogs or cats adopted in the US? Are more dogs adopted within the Pacific Northwest or South Central regions of the United States?

One Sample Z-Test for a Proportion.

Data.

- Sample size: 372, 647 (203, 880 Dogs and 168, 767 Cats)
- Variables used from PetFinder API data: ‘id’, ‘type’, ‘contact.address.state’, ‘published_at’
- published_at Time Range: 2022-01-01 to 2022-12-31
- Filters: *Location:* Entire US, *Status:* Adopted, *Type:* Dogs and Cats

Descriptive Statistics. Create a boxplot showing the range of values of adopted dogs and cats by state. Each data point in the box plot represents the count of adopted dogs or cats for a specific state within the entire published_at time range. Also, create a heatmap showing the proportion of dogs by state.

Inferential Statistics. Conduct one sample z-test for comparing a proportion to the fixed value of 0.5. The population of interest is Dogs and Cats listed on PetFinder who were ultimately adopted. Our sample is a simple random sample from the population of interest. Need to verify the sample size is sufficiently large to meet the assumption of large sample size. The one sample z-test for comparing a proportion is appropriate for a simple comparison of adopted dogs and cats from the population of interest, without taking any additional factors into consideration. Using the fixed value of 0.5 is relevant since we are only considering dogs and cats as the options for adopted animal types, we would expect them to have equal proportions.

Two Sample Z-Test for comparing two proportions.

Data.

- Sample size (PNW): 11, 101 (7, 006 dogs and 4, 095 cats)
- Sample size (SC): 40, 543 (26, 382 dogs and 14, 161 cats)
- Variables used from PetFinderAPI data, and published_at time range same as One Sample Z-Test for a proportion
- Filters: *Location:* PNW and SC regions, *Status:* Adopted, *Type:* Dogs and Cats

Descriptive Statistics. Create a side by side boxplot showing the range in values of proportions of adopted dogs and cats within each region. Each data point in the box plot represents the proportion of adopted dogs or cats for a specific state within the region within the entire published_at time range.

Inferential Statistics. Conduct two sample z-test for comparing the proportion of dogs in the PNW region to the proportion of dogs in the SC region. The populations of interest are Dogs and Cats listed on PetFinder who were ultimately adopted in the PNW and SC regions of the US. Our sample is a simple random sample from the populations of interest. Need to verify the sample size is sufficiently large to meet the assumption of large sample size. The two sample z-test for comparing two proportions is appropriate for a simple comparison of the proportion of adopted dogs in the PNW and SC regions of the US, without taking any additional factors into consideration. Using a proportion instead of comparing average number of pets adopted per month is more appropriate for this comparison because there are more states in the SC region than the PNW region.

Welch Two Sample t-test for comparing two means.

Data.

- Sample size for TX and WA: 24 (24 months of counts of adopted dogs in each state)
- Variables used from PetFinderAPI data, and published_at time range same as One Sample Z-Test for a proportion
- published_at Time Range: 2021-01-01 to 2022-12-31
- Filters:
    - Location: TX, WA
    - Status: Adopted
    - Type: Dogs

Descriptive Statistics. Create a side by side boxplot showing the range in values of counts of adopted dogs per month in TX and WA. Each data point in the box plot represents the count of adopted dogs for a specific month within the state.

Inferential Statistics. Conduct Welch Two-Sample t-test for comparing two means to compare the average number of adopted dogs per month in WA to the average number of adopted dogs per month in TX. The populations of interest are Dogs listed on PetFinder who were ultimately adopted in WA and TX. Our sample is a simple random sample from the populations of interest. Need to verify the sample size is sufficiently large to meet the assumption of large sample size. Using the welch two sample t-test to relax the requirement of equal variance. The Welch test is the most appropriate test to determine if there are typically

more dogs adopted in WA or TX, without accounting for any additional, more complex factors. Comparing average number of adopted dogs per month over a 2 year period should help with minimizing the effect of any strong outliers.

Research Question 2

Which characteristics speed up adoption and reduce shelter stay (time a pet is listed on petfinder)?

Data.

Sample Size:

- 163,976 (adopted dogs and cats in the Pacific Northwest, South Central and the Northeast)
- Variables used:
  'type', 'species', 'age', 'gender', 'size', 'coat', 'breed':(mixed,unknown,particular), 'colors.primary', 'colors.secondary', 'attributes.spayed_neutered', 'attributes.house_trained', attributes.declawed', 'attributes.special_needs', 'attributes.shots_current', 'environment.children', 'environment.dogs','environment.cats','days_in_shelter'
- To calculate the number of days that an animal was in the shelter ('days_in_shelter' variable), the published_at column was subtracted from the status_changed_at date to get a timedelta object representing the amount of time that the animal has been listed.

Descriptive Statistics.

- Computed Mean, median, and standard deviation of the days spent on Petfinder before adoption for cats and dogs.
- Created scatterplots of the days spent on Petfinder by adoption and size (cats and dogs)

Inferential Statistics.

Welch Two sample t-test on days in the shelter by type of animal.

Statistical Method. For the Welch Two sample t-test on days in the shelter by type of animal, the statistical method used is the Welch t-test, which is a variation of the t-test that does not assume equal variances between the two groups being compared.

Assumptions. The assumptions of the test are that the two groups (cats and dogs) are independent, and the population distributions of the two groups are approximately normal.

Rationale for appropriateness of the method. The Welch two-sample t-test is suitable for comparing the mean number of days animals spend in the shelter by type because it is designed for comparing groups with unequal variances, assumes independent samples, and assumes normality. It is also a hypothesis test that allows us to determine whether the difference between the means of the two groups is statistically significant.

Statistical hypothesis being tested. The statistical hypothesis being tested in the Welch two-sample t-test on days in the shelter by type of animal is whether there is a statistically significant difference between the mean number of days spent in the shelter for two independent groups of animals (e.g., cats vs. dogs). The test is used to determine whether any

observed difference between the means is likely due to chance or whether it is a true difference that is statistically significant.

ANOVA Test to compare the mean days spent on Petfinder before adoption between male and female dogs and small and large dogs.

Statistical Method. For the test to compare the mean days spent on Petfinder before adoption between male and female dogs, an ANOVA test is used. The test determines whether there are any statistically significant differences between the means of two or more independent groups.

Assumptions. The assumptions of the ANOVA test are that the data is normally distributed, the variances of the groups being compared are equal, and the groups are independent.

Rationale for the appropriateness of the method. The ANOVA test is an appropriate statistical method for comparing the mean days spent on Petfinder before adoption between different groups of dogs because it can handle multiple groups and determine if there is a significant difference between them. ANOVA assumes independence between groups and normality within each group. It also allows for the examination of interactions between groups.

Statistical hypothesis being tested. The statistical hypothesis being tested is whether there is a statistically significant difference in the mean days spent on Petfinder before adoption between male and female dogs and small and large dogs. Specifically, the null hypothesis is that there is no difference between the groups, while the alternative hypothesis is that there is a difference. The ANOVA test allows us to test this hypothesis by comparing the variability within each group to the variability between the groups, using an F-test.

Survival analysis to predict adoption probability over time and identify factors impacting adoption duration (e.g., age, size, Petfinder duration).

Statistical Method(s). For the Survival analysis to predict adoption probability over time and identify factors impacting adoption duration (e.g., age, size, Petfinder duration), several statistical tests are used. First, a Likelihood ratio test, Wald test, and Score (logrank) test are used to determine if there is any relationship between age, size, or days spent on Petfinder before adoption and the time to adoption for pets. If there is evidence of a relationship, further analysis is done using a Cox proportional hazards model to examine the coefficients of the variables and identify which variables have a significant relationship with adoption. Finally, regression is used to identify which variables are statistically significant predictors of adoption.

Assumptions. The assumptions for these tests are that the data is independent, the events are non-informative, and the hazard rates are proportional over time.

Rationale for the appropriateness of the method. The use of survival analysis and regression to predict adoption probability over time and identify factors impacting adoption duration is appropriate for several reasons. Firstly, survival analysis is specifically designed for modeling time-to-event data, which is applicable in this case as we are interested in predicting the time until adoption. Regression analysis is useful for identifying the factors that are associated with the outcome variable (adoption duration), allowing us to determine which factors may be important predictors of adoption.

Furthermore, survival analysis and regression allow us to account for the fact that some animals may still be in the shelter at the end of the study period, and the outcome variable (adoption duration) may be right-censored. This means that we may not know the exact duration of time an animal spent in the shelter, but only that they were still in the shelter at the end of the study period. Survival analysis can handle censored data and estimate the probability of adoption over time, while regression can account for the effects of censoring on the estimated coefficients.

Overall, the use of survival analysis and regression to predict adoption probability over time and identify factors impacting adoption duration is appropriate and allows us to gain insights into the factors that may influence adoption and how they impact the duration of time an animal spends in the shelter.

Statistical hypothesis being tested. The statistical hypothesis being tested for survival analysis and regression to predict adoption probability over time and identify factors impacting adoption duration would be whether the chosen predictor variables (age, size, Petfinder duration) have a statistically significant effect on the hazard rate or survival probability of adoption, after controlling for other variables in the model.

Research Question 3

Do adopted dogs have the same life expectancy as compared to the population ?

One Sample Z-Test for comparing life expectancy of adopted dogs to the population of all dogs of US

Data

- Sample size: 61,610 Adopted Dogs
- Variables used from PetFinder API data: ‘id’, ‘type’, ‘breeds.primary’
- Variables used from AKC data: ‘max_expectancy’, ‘max_expectancy’, ‘Breed’
- Time Range: 2022-12-14 to 2022-02-23
- Filters:
    - Location: Entire US
    - Status: Adopted
    - Type: Dogs

Descriptive Statistics. Calculate Mean, median, mode, standard deviation, variance, and range of the minimum life expectancy and max life expectancy variables.
Plotted histogram of min_expectancy and max_expectancy

Assumptions. Our sample is a simple random sample from the population of interest. We verified the sample size is sufficiently large to meet the assumption of a large sample size.

Rationale for appropriateness of the method. The one-sample Z-test is an appropriate statistical test to compare the minimum and maximum life expectancy of adopted dogs with the overall US dog life expectancy values of 10-13 years.This is because we know that the sample is large enough to have normal distribution from central limit theorem and we also know the population parameters. We can assume that the life expectancy of adopted dogs is independent of the overall US dog life expectancy values.

Inferential Statistics. It has been computed that dogs' life expectancy in the US is 10-3 years. Conduct one sample z-test for comparing a min_expectancy of adopted dogs to the fixed value of 10(population min_expectancy). The population of interest is Dogs and on PetFinder who were adopted. The one-sample z-test for comparing a proportion is appropriate for a simple comparison of adopted dogs from the population of interest, without taking any additional factors into consideration. Using the fixed value of 10 is relevant since we that is the population min_expectancy in the US.
Similarly, we also conduct one sample Z test for max_expectancy comparing it with 13 years which is the population max_expectancy in the US.

One-way ANOVA test to find if there is relationship between min_expectancy of dogs of various regions

Data

- Sample size: 61,610 Adopted Dogs (from all 3 regions)
- Variables used from PetFinder API data: ‘id’, ‘type’, ‘breeds.primary’, ‘contact.address.state’
- Variables used from AKC data: ‘max_expectancy’, ‘Breed’
- Time Range: 2022-12-14 to 2022-02-23
- Filters:
    - Location: Entire US
    - Status: Adopted
    - Type: Dogs

Descriptive Statistics. Computed 95% confidence intervals of min_expectancy of adopted dogs for all three regions (Pacific Northwest, South Central, Northeast)

Assumptions. The assumptions of the ANOVA test are that the data is normally distributed, the variances of the groups being compared are equal, and the groups are independent. All the assumptions hold true for conducting the tests for adopted dogs between different regions.

Rationale for the appropriateness of the method. The one way ANOVA test is an appropriate statistical method for comparing the min_expectancy between dogs of different regions because it can handle multiple groups and determine if there is a significant difference between them. ANOVA assumes independence between groups and normality within each group. It also allows for the examination of interactions between groups.

Inferential Statistics. For the test to compare the min_expectancy of dogs in the three regions, an ANOVA test is used. The test determines whether there are any statistically significant differences between the means of min_expectancy in the three independent groups of dogs of regions.

Research Question 4

Chi-squared Test of Independence

Data.

- Sample size: 163,976

- From PetFinder API data: age, gender, size, species, coat, contact.address.state, region
    - Regions are outlined in the Data subsection of Research Question 1

Descriptive Statistics. Provide contingency tables with the percentage that lists the number of pets in different regions with their category in features.

Inferential Statistics. Perform chi-square test to test the association between the following sets of 2 categorical variables: region vs pet age, region vs pet gender, region vs pet size, region vs species, region vs coat.

Assumptions. The assumptions to perform chi-squared tests for categorical variables are the sample size is large, and the variables are mutually exclusive.

Simple Linear Regression for Number of Adopted Pets

Data.

- Sample size: 163,976
- From PetFinder API data: age, gender, size, species, coat, contact.address.state, region
- From income data: median_income, state
- From population data: total_population, median_age, state

Descriptive Statistics. Calculate min and max values for the following variables from the census income and population datasets: median_income, total_population, median_age. Provide a side-by-side boxplot showing median_income for each household type, and a plot for the relationship between total_population and median_age.

Inferential Statistics. Fit linear regression model with the data from PetFinder API and Census Bureau, including age, gender, size, species, coat, contact.address.state, region, population, median income, and median_age. Test the significance of the variables in the regression model and test the significance of the overall model. Examine the model for the assumptions using residual plots.

Assumptions. The assumptions for simple linear regression are independent observations, linear relationships between the mean response and predictors, normally distributed response variables, and independent and constant variance.

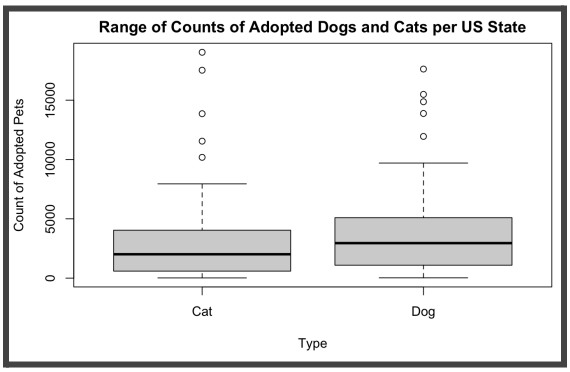
Section 4: Results

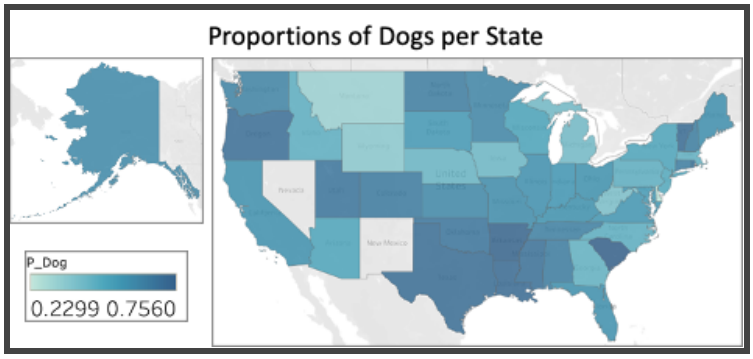
Research Question 1

Are more dogs or cats adopted in the US? Are more dogs adopted within the Pacific Northwest or South Central regions of the United States?

One Sample Z-Test for a Proportion

Descriptive Statistics.





Inferential Statistics.

Prove Assumptions. Sample size is 372, 647 (203, 880 Dogs and 168, 767 Cats). This makes the proportion of dogs .5471. In order to verify the assumption of large sample size, the following must be true: $n(1 - p_{dogs}) > 5$ AND $n(p_{dogs}) > 5$. Based on this, the minimum sample size is 12. Since our sample size is 372, 647, it meets the assumption of large sample size.

Hypotheses. H_0: $p_{dogs} = .5$ H_a: $p_{dogs} \neq 0$

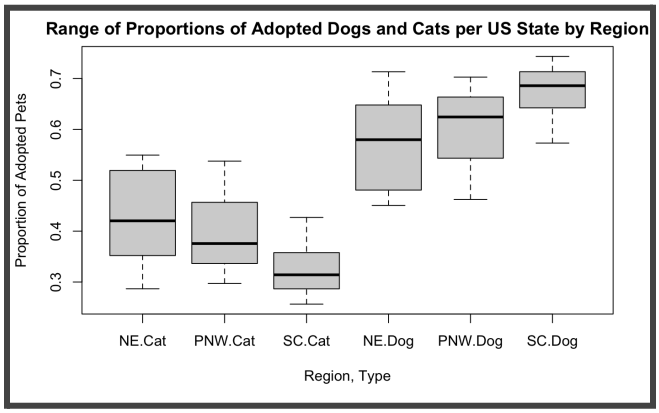
Results.

Test Stat.	P-Value	95% Conf. Int.	Result (.05 sig. level)
z: 57.52	< 2.2e^-16	(.5455, .5487)	Reject the null hypothesis

Based on the p-value being very close to 0, and the 95% confidence interval being at least .455 greater than .5, can conclude that the proportion of adopted dogs is higher than adopted cats amongst dogs/cats listed on PetFinder. We estimate that for every 5 cats adopted, 6 dogs are adopted.

Two Sample Z-Test for comparing two proportions

Descriptive Statistics.



Inferential Statistics.

Prove Assumptions. The sample size for the PNW region is 11, 101 (7, 006 dogs and 4, 095 cats), and the proportion of dogs in the PNW region .6311. The sample size for the SC region is 40, 543 (26, 382 dogs and 14, 161 cats), and the proportion of dogs in the SC region is

.6507. In order to verify the assumption of large sample size, the following must be true: $n(1 - p_{dogsPNW}) > 5$ AND $n(p_{dogsPNW}) > 5$ as well as $n(1 - p_{dogsSC}) > 5$ AND $n(p_{dogsSC}) > 5$. Based on this, the minimum sample sizes for the PNW and SC regions are 14 and 15, respectively. Since our sample sizes for the PNW and SC regions are 11, 101 and 40,543 respectively, the samples meet the assumption of large sample size.

Hypotheses. H_0: $p_{dogsPNW} = p_{dogsSC}$ H_a: $p_{dogsPNW} \neq p_{dogsSC}$

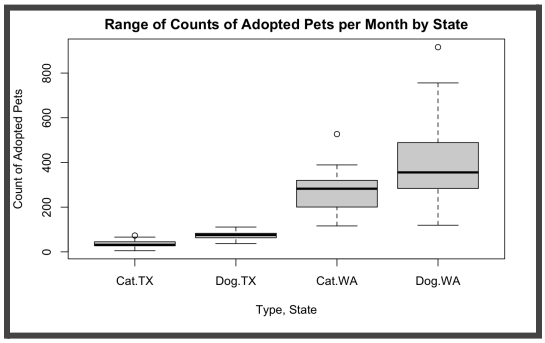
Results.

Test Stat.	P-Value	95% Conf. Int.	Result (.05 sig. level)
z: 3.8279	< .00013	(-0.02971, -0.0095)	Reject the null hypothesis

Based on the p-value being very close to 0, and the 95% confidence interval being at least .0095 less than 0, can reject the null hypothesis. Additionally, there is evidence to support that the proportion of adopted dogs in the SC region is higher than the proportion of adopted dogs in the PNW region. We estimate that for every 3 cats adopted in the PNW region there are 5 dogs adopted, and for every 10 cats adopted in the SC region there are 19 dogs adopted.

Welch Two Sample t-test for comparing two means

Descriptive Statistics.



Inferential Statistics.

Prove Assumptions. The sample size for both TX and WA is 24, where each record in the sample is the number of adopted dogs in a month-long time period. In order to verify the assumption of large sample size, utilized the pwr.t.test package in R. Assuming a pooled standard deviation of 123.29, a power of .75, a significance level of .05, and difference in means of 96, 24 months is a large enough sample size to detect a difference of 96 between the two sample means.

Hypotheses. H_0: $\mu_{adopted\ dogs/month\ TX} = \mu_{adopted\ dogs/month\ WA}$
H_a: $\mu_{adopted\ dogs/month\ TX} \neq \mu_{adopted\ dogs/month\ WA}$

Results.

Test Stat.	P-Value	95% Conf. Int.	Result (.05 sig. level)
t: -9.1848 df = 23.433	< 3.131e^-09	(-400.522, -253.394)	Reject the null hypothesis

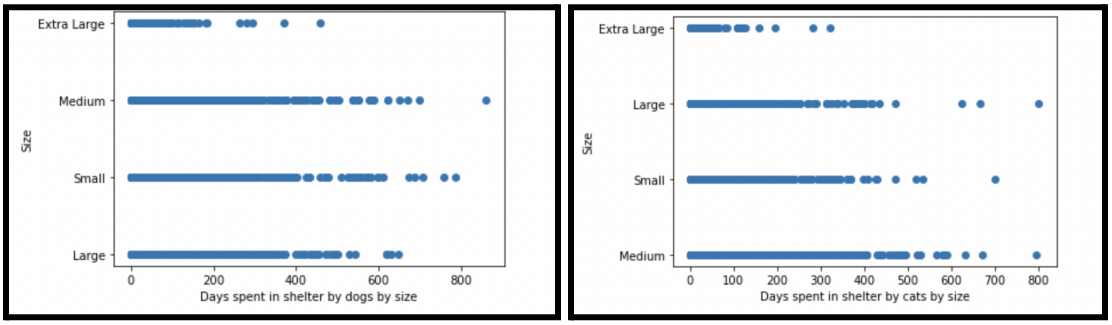
Based on the p-value being very close to 0, and the 95% confidence interval being at least 253 less than 0, can reject the null hypothesis. Additionally, there is evidence to support

that the average number of adopted dogs per month who were listed on PetFinder in WA is higher than the average number of adopted dogs per month who were listed on PetFinder in TX.

Research Question 2

The data in the Descriptive Statistics section shows that dogs spend slightly less time in shelters before adoption compared to cats, with a mean of 33.77 days compared to 34.83 days for cats. However, the standard deviation for both groups is quite high, with dogs having a standard deviation of 46.12 days and cats having a standard deviation of 46.79 days, indicating a large amount of variability in the data.

The median days in the shelter for dogs is only 18.00, which is much lower than the mean, indicating that there are some dogs that get adopted relatively quickly and some that stay in the shelter for a very long time. Similarly, the median days in the shelter for cats is also 18.00, which is lower than the mean, indicating a similar pattern of some cats being adopted quickly and others staying in the shelter for a longer time.



Scatterplots of the days spent on Petfinder by adoption and size (cats and dogs)

Welch Two sample t-test on days in the shelter by type of animal

The sample estimates show that cats have a mean "days_in_shelter" of 34.826 days, while dogs have a mean "days_in_shelter" of 33.771 days. This suggests that cats tend to stay in shelters for slightly longer than dogs on average. The 95% confidence interval for the difference in means between cats and dogs is (0.629, 1.480). This means that we are 95% confident that the true difference in means lies between these two values.

Hypotheses. H_0: $\mu_{\text{days in shelter(cats)}} = \mu_{\text{days in shelter(dogs)}}$

H_a: $\mu_{\text{days in shelter(cats)}} \neq \mu_{\text{days in shelter(dogs)}}$

Results.

Test Stat.	P-Value	95% Conf. Int.	Result (.05 sig. level)
t: 4.8517 df = 174012	1.225e-06	(0.629, 1.480)	Reject the null hypothesis

The p-value of the test is 1.225e-06, which is very small. This means that the probability of observing a difference in means as large as the one in the sample, assuming that the true means are equal, is very low. In fact, the p-value is much smaller than the standard significance level of 0.05, which suggests strong evidence against the null hypothesis. We can reject the null hypothesis in favor of the alternative hypothesis and conclude that there is a significant difference in mean "days_in_shelter" between cats and dogs.

ANOVA Test to compare the mean days spent on Petfinder before adoption between male and female dogs

The ANOVA table shows that the F-statistic is 14.87, and the corresponding p-value is 0.000115. This means that the probability of obtaining an F-statistic as large as 14.87 or larger, assuming that the null hypothesis is true, is very low (less than 0.01%).

Hypotheses. H_0: $\mu_{male} = \mu_{female}$
H_a: $\mu_{male} \neq \mu_{female}$

Results.

Test Stat.	P-Value	Result (.05 sig. level)
f stat: 14.87	0.00015	Reject the null hypothesis

Since the p-value is smaller than the conventional threshold of 0.05, we reject the null hypothesis and conclude that there is a significant difference in mean days spent on Petfinder before adoption between male and female dogs.

ANOVA Test to compare the mean days spent on Petfinder before adoption between small and large dogs

We can see that there is a statistically significant difference in mean days spent on Petfinder before adoption between small and large dogs ($F(3, 104307) = 239.7$, $p < 0.001$). The size variable seems to have a strong effect on the outcome.

Hypotheses. H_0: $\mu_{small} = \mu_{large}$
H_a: $\mu_{small} \neq \mu_{large}$

Results.

Test Stat.	P-Value	Result (.05 sig. level)
f stat: 239.7	<2e-16	Reject the null hypothesis

Since the p-value is less than 0.001, we can reject the null hypothesis and conclude that there is a statistically significant difference in mean days spent on Petfinder before adoption between small and large dogs.

Survival Analysis

The Likelihood ratio test, Wald test, and Score (logrank) test all have extremely low p-values of "<2e-16". This indicates strong evidence against the null hypothesis of no relationship between age, size, or days spent on Petfinder before adoption and the time to adoption for pets.

Hypotheses. H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_a: At least one of β_1, β_2 and $\beta_3 \neq 0$

Results.

Concordance	Likelihood ratio test	Wald Test	Score (logrank) test	Result (0.05 sig level)
0.573	5581	5458	5522	Reject the null hypothesis

Therefore, we can reject the null hypothesis and conclude that there is a relationship between at least one of these variables and the time to adoption for pets. Based on the model results, the following variables are statistically significant predictors of adoption: ageBaby, ageYoung, environment.dogs the colors Black, Bicolor, Black and White/Tuxedo, Buff/Tan/Fawn,

Buff and White, Calico, Chocolate Point, Cream Point, Dilute Tortoiseshell, Gray/Blue/Silver, Gray and White, Merle (Blue), Orange/Red, Orange and White, Sable and Seal Point.

For example, a one-unit increase in ageBaby increases the log odds of adoption by 0.28834, holding all other variables constant. In other words, baby animals are more likely to be adopted than animals of other ages. A one-unit increase in environment.dogs, which measures whether the animal is comfortable around other dogs, increases the log odds of adoption by 0.53574. In other words, if an animal is comfortable around dogs, its adoption rate goes up.

In conclusion, the analysis suggests that there is a significant difference in the time pets spend in shelters and on Petfinder before adoption based on factors such as species, gender, size, age, and coat color. The results can be used to develop effective strategies to reduce the time pets spend in shelters and increase their chances of being adopted.

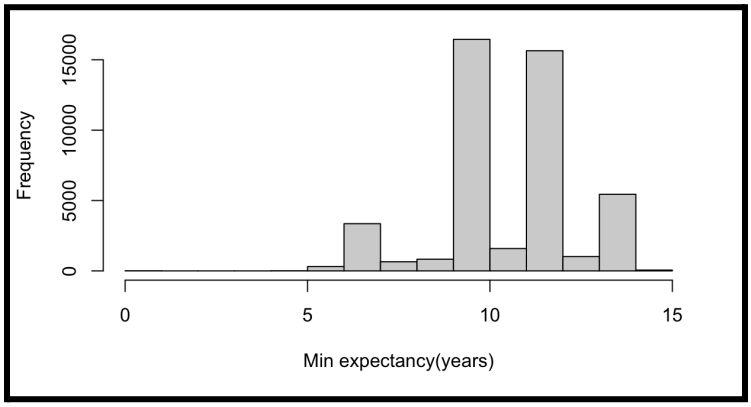
Research Question 3

Do adopted dogs have the same life expectancy as compared to the population ?

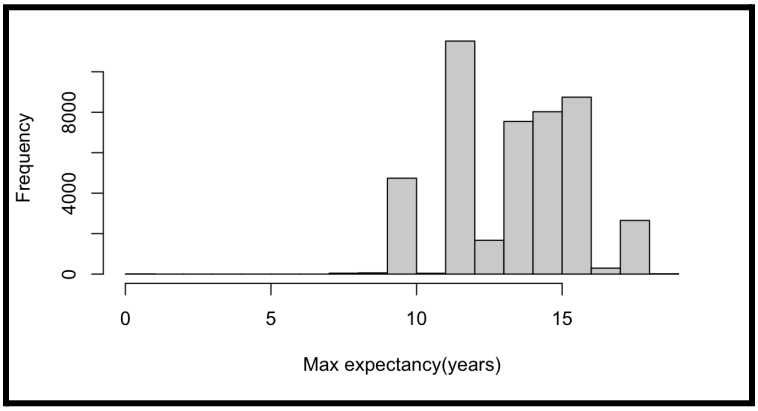
One Sample Z-Test for a Proportion

Descriptive Statistics. Plotted histogram of min_expectancy and max_expectancy, as our sample size is very large(60k) so by central limit theorem as well as verifying it by Q-Q plot we see that min_expectancy is normally distributed

Histogram of min expectancy of adopted dogs in US



Histogram of max expectancy of adopted dogs in US



Inferential Statistics

Results.

Hypothesis Test	Test Stat.	P-Value	95% Conf. Int.	Result (.05 sig. level)
One-Sample z-test H_0: min_expectancy = 10 H_A: min_expectancy ≠ 10	z: 111.78	<2.2e-16	(10.962,10.997)	Reject the null hypothesis
One-Sample z-test H_0: max_expectancy = 13 H_A: max_expectancy ≠ 13	z: 374.45	<2.2e-16	(13.817,13.857)	Reject the null hypothesis

From the above results we can conclude that the min_expectancy and max_expectancy of adopted dogs is not equal to the national average of 10-13 years. Upon close examination we can also conclude that it is higher than the national average. These results are as expected as the dogs that have been adopted have a higher life expectancy.
As part of descriptive statistics we also plotted the min_expectancy and max_expectancy of adopted dogs.

One-way ANOVA test

To compare the **min expectancy** of adopted **dogs of the three different regions** *Pacific Northwest, South Central, Northeast.*

Inferential Statistics. The ANOVA results suggest that there is a significant difference in the mean of the dependent variable across the three regions, as indicated by the highly significant p-value (4.19e-12) and the F-value (26.24). The mean square value of 85.14 for the data$region variable indicates that the variation between the groups (regions) is much larger than the variation within the groups (residuals). Therefore, it is likely that the mean values of the dependent variable in the different regions are significantly different from each other.

Hypotheses. H_0: μ_PNW = μ_NE = μ_SC (assoc. between min expectancy and region)

Results.

Test Stat.	P-Value	Result (.05 sig. level)
f stat: 26.24	4.19e-12	Reject the null hypothesis

In addition to this, we also computed the 95% confidence interval for min_expectancy of all the 3 regions.

Region	95% Confidence Interval of min_expectancy
Pacific Northwest	(11.158, 11.343)
South Central	(10.898, 10.986)
Northeast	(10.847, 10.930)

From this we can clearly see that the three regions' 95% confidence interval of min_expectancy is quite different and hence we can conclude that there aren't similarities between the min_expectancy of the three regions. This can be explained by various factors like weather and other ecological factors.

Research Question 4

What are the location based factors that impact the number of adopted pets in different states?

Chi-squared Test of Independence

Descriptive Statistics.

	Age				Gender		Species	
	Baby	**Young**	**Adult**	**Senior**	**Female**	**Male**	**Cat**	**Dog**
NE	55663	26139	26454	4118	56832	55542	53163	59211
PNW	5291	2430	2891	489	5638	5463	4095	7006
SC	17257	9747	12573	924	20558	19943	14136	26365

	Size				Coat						
	Extra large	Large	Medium	Small	Curly	Hair-less	Short	Long	Medium	Wire	N/A
NE	822	20762	64107	26683	678	90	61602	4800	12467	804	31933
PNW	97	2368	5495	3141	104	8	5425	602	2096	163	2703
SC	331	8501	23115	8554	200	15	20665	1805	5367	330	10878

Inferential Statistics.

Prove Assumptions. In order to perform chi-squared test, the expected cell counts should be equal or larger than 5. For every pair of categorical variables, there has over 5 cells and with a large sample size.

Hypotheses. H_0: no significant association between 2 categorical variables

H_a: has significant association between 2 categorical variables

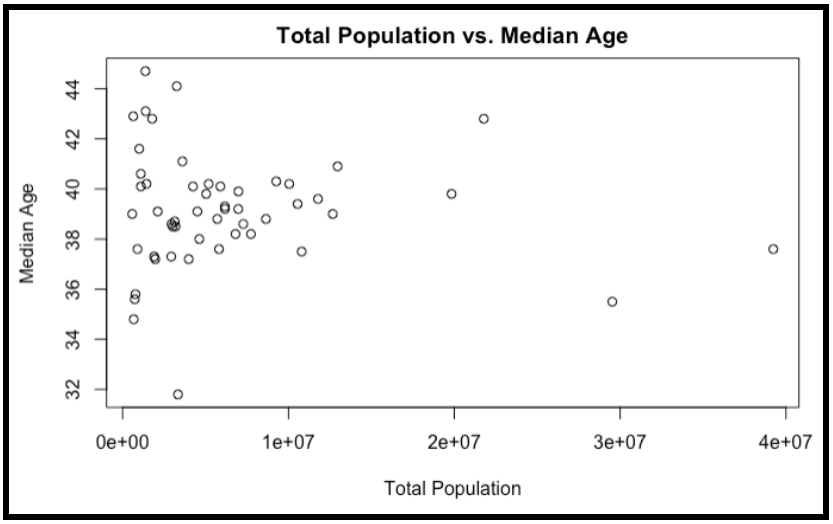
Results.

Chi-square test	Test statistic	P-value	Result
Region vs Age	1181.8	<2.2e-16	Reject null hypothesis
Region vs Gender	0.5217	0.77	Fail to reject null hypothesis
Region vs Size	441.79	<2.2e-16	Reject null hypothesis
Region vs Species	2078.2	<2.2e-16	Reject null hypothesis
Region vs Coat	875.43	<2.2e-16	Reject null hypothesis

According to the p-value for each chi-squared test, at significance level of 0.05, we have p-value < 0.05 for every pair of categorical variables except for Region vs. Gender. That means except for Gender, all other tests conclude to reject the null hypothesis, meaning that there is a significant association between the two variables.

Simple Linear Regression for Number of Adopted Pets

Descriptive Statistics. Check the plot below to see the relationship between the total population and median age. These two variables are both predictors, to prevent collinearity, it is important to see if the numerical variables have linear relationship.

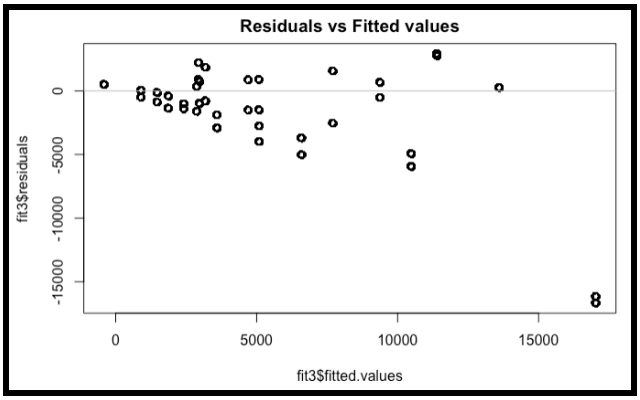


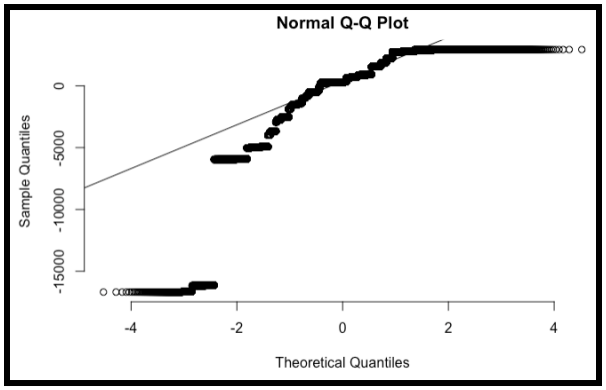
Inferential Statistics.

Prove Assumptions. The assumptions for simple linear regression are independent observations, linear relationships between the mean response and predictors, normally distributed response variables, and independent and constant variance.

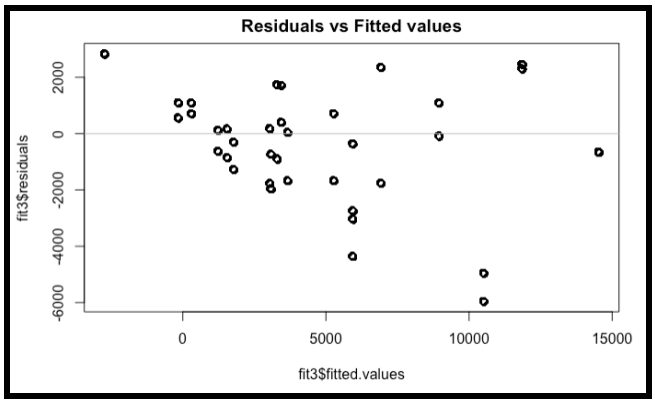
Hypotheses. H_0: $\beta_1=\beta_2=...=\beta_6=0$, H_a: at least 1 of the $\beta \neq 0$

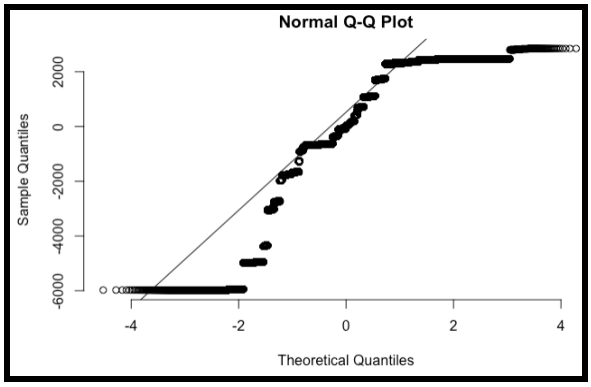
pet_count ~ I(region) + I(gender) + new_median_income + total_population + median_age





After examining the residual plots, there are two assumptions violated: constant variance and normal distribution. In the residual plot on the left, there are two outlier points on the bottom right corner. After checking the dataset, I filtered out the outliers and fit another model with the same predictors and new dataset.





After updating the dataset, we now have a more randomly plotted residual plot around the horizontal line. However, the normal distribution is still violated. The Q-Q plot shows a slight pattern to binomial distribution at the two ends of the quantile.

Results. For this linear regression model, we have statistical test for each variable.

```
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)       -1.359e+04  3.544e+02  -38.349  < 2e-16 ***
I(region)PNW      -4.028e+03  2.849e+01 -141.380  < 2e-16 ***
I(region)SC        2.892e+03  3.925e+01   73.683  < 2e-16 ***
I(gender)Male      3.216e+01  1.028e+01    3.127  0.00176 **
new_median_income  2.271e-01  8.043e-04  282.409  < 2e-16 ***
total_population   6.234e-04  1.435e-06  434.526  < 2e-16 ***
median_age        -6.831e+01  7.444e+00   -9.176  < 2e-16 ***
```

Every predictor has p-value < 0.05 significance level, meaning that every predictor is statistically significant in this model.

```
Residual standard error: 2074 on 162728 degrees of freedom
Multiple R-squared:   0.81,     Adjusted R-squared:   0.81
F-statistic: 1.156e+05 on 6 and 162728 DF,  p-value: < 2.2e-16
```

For this overall model, we have adjusted R-squared to be 0.81, meaning that 81% of the variability can be explained by this linear model. The p-value for this model is < 0.05 significance level, meaning that we can reject the null hypothesis, the overall model is also statistically significant.

Section 5: Discussion

The research conducted shows that various factors, such as species, gender, size, age, coat color, and location, have an impact on the adoptability of pets listed on PetFinder. Specifically, cats stay longer on PetFinder before being adopted than dogs, and gender and size also affect how long these animals remain in the shelter. Future research could explore the impact of other factors on pet adoption, such as the role of photos or descriptions on PetFinder, the influence of the shelter's location or reputation, and the effect of adoption fees. Additionally, comparing the results from pets listed on PetFinder to results for a shelter or multiple shelters that do not list their adoptable pets on PetFinder could provide further insight into pet adoption.

Furthermore, the statistical tests conducted on the life expectancy of adopted dogs show that their life expectancy is not equal to the population (10-13 years). This leads to the conclusion that the dogs that have been adopted are healthier. However, there is no relationship between the life expectancy of dogs in the three regions, which could be due to various underlying factors. Future analysis could include a comparison between adopted dogs and dogs that have not been adopted for a long duration.

Finally, the statistical tests performed to determine the association between categorical variables related to region compared to age, gender, species, size, and coat reject the null hypothesis of all pairs except for region vs. gender. To avoid overfitting, only two categorical variables were chosen with another three numerical variables to the model. The linear model assumptions were examined with residual plots, and after removing the outlier data, a more randomly plotted residual graph was obtained. However, the Q-Q plot does not show a normal distribution, which may be due to the pet count being based on state and species, resulting in a non-continuous count. Overall, future research could explore additional factors that may influence the association between categorical variables.

References

American Kennel Club. American Kennel Club. (n.d.). Retrieved March 14, 2023, from https://www.akc.org/

Aschleg. (2021, July 24). *Aschleg/Petpy: Petpy is an easy-to-use and convenient python wrapper for the petfinder API*. GitHub. Retrieved March 14, 2023, from https://github.com/aschleg/petpy

Bureau, U. S. C. (2022, September 7). *Coverage rates*. Census.gov. Retrieved March 14, 2023, from https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/coverage-rates/index.php.

Bureau, U. S. C. (2022, September 7). *Sample size*. Census.gov. Retrieved March 14, 2023, from https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/sample-size/index.php

PetFinder API calls. Petfinder. (n.d.). Retrieved March 14, 2023, from https://www.petfinder.com/developers/v2/docs/

tmfilho. (2020, May 18). *Data scraped from American Kennel Club website*. Akcdata. Retrieved March 14, 2023, from https://tmfilho.github.io/akcdata/

Tables and Figures

PetFinder Variables

Res. ?	Variable	Explanation
1, 3	‘ID’	Unique id for a listing on PetFinder
1, 2, 3	‘type’	categorical variable: dog, cat, rabbit, small & furry, horse, bird, scales/fins/other, barnyard Note: Filtered for only dog, cat in this project
2	‘species’	categorical variable, similar to ‘type’
2, 4	‘age’	categorical variable: baby, young, adult, and senior
2, 4	‘gender’	categorical variable: male, female

2, 4	‘size’	categorical variable: small, medium, large, xlarge
2	‘coat’	categorical variable **For dogs:** short, medium, long, wire, hairless, curl. **For cats:** hairless, short, medium, long
2	‘breed’	Categorical variable: possible values may be looked up via Get Animal Breeds api call
2	‘colors.primary’ and ‘colors.secondary’	Categorical variable: possible values may be looked up via Get Animal Types api call
2	'attributes.spayed_neutered'	Boolean: can be true or 1 only
2	‘attributes.house_trained’	Boolean: can be true or 1 only
2	‘attributes.declawed’	Boolean: can be true or 1 only
2	‘attributes.special_needs’	Boolean: can be true or 1 only
2	‘attributes.shots_current’	Boolean: can be true, false. 1 or 0
2	‘environment.children’	Boolean: can be true, false. 1 or 0
2	‘environment.dogs’	Boolean: can be true, false. 1 or 0
2	‘environment.cats’	Boolean: can be true, false. 1 or 0
2	‘published_at’	Date-Time: when listing was published on PetFinder
1, 3, 4	‘contact.address.state’	String - state abbreviation

AKC Variables

Res. ?	Variable	Explanation
3	‘min-expectancy’	Integer, minimum life expectancy
3	‘max-expectancy’	Integer, maximum life expectancy

Census (Income) Variables

Res. ?	Variable	Explanation
4	median_income	Double, median household income
4	state	US state

Census (Population) Variables

Res. ?	Variable	Explanation
4	total_population	total population
4	male	male population
4	female	female population
4	median_age	Double, median age
4	state	US state