# Predicting Pet Registrations With Public Data

Nick Young, Calista MacLagan, Yidan Fu, Regina Kang

2023-04-15

## Overview of Project

### Predicting Pet Registration Numbers with Public Data

Nationally, the availability of pet registration data is inconsistent, and datasets are often incomplete. Jennifer, Chuck, and Barbara concluded that some surveys reporting pet ownership were not available for public use and/or did not lend themselves to social science due to a limited number of other measures of important social and demographic characteristics (10). The initial hypothesis for this project proposed that there are one or more variables available that have a statistically significant relationship to amount / density of pet registrations. This project also sought to evaluate secondary research questions: can the ratio of cats to dogs be predicted and can potential customers be clustered by zip code. The team created three unique models that predict the pet population of a zip code, the ratio of cats to dogs, and clusters zip codes to reveal interesting demographic data. Together, these models best inform companies business strategies for these areas.

### Choice of Topic, Business Justification, and Problem Statement

Understanding where its target market lives can provide significant benefits to a pet supply company from a business market standpoint. According to Perry and Burge, over 60 percent of Americans have some sort of family pet and America is one of the largest global economies (1). If a company like Chewy can identify zip codes with the highest amount/density of pets, it can target its marketing efforts in those areas. By sending mailers to these zip codes, the company can increase brand awareness and attract more customers from these areas. Additionally, understanding the mix of cats and dogs in a population will allow a pet supply company to optimize its supply chain and reduce costs. This informs the company of strategic locations for warehouses and distribution centers, minimizing transportation costs and delivery times. Further, understanding the characteristics between concentrations of pet owners allows companies to target their marketing campaigns to better align with what that population values.

## Overview of Data

### Understanding of the Data and Data Wrangling

The project initially focused on the greater Seattle, WA area, but after cleaning and prepping our data we realized we only had 35 observations. To improve the validity of our research, we scoured for more data where available. This led us to find pet registration data including New York City, NY, and Louisville, KY, for a total of 238 observations. Six unique datasets provide pet registrations for the three cities as well as number of parks, median household income, number of housing units, total population, gender ratio, percent

of residents by race, and percent of people in institutions (college, military, etc.), and so on. These datasets are publicly available, and we downloaded them online.

All the datasets are grouped by zip code except for the census data, which is grouped by census block. To have all variables at the same granularity level, we used a mapping file to match the data provided by the census block to the data provided by the zip codes. For ease of use, we transformed column names into a snake case format.

To better compare features, we scaled the variables that were not between 0 and 1.

For zip codes in which the park count was listed 'NA', the count was converted to zero, and it is assumed there are no parks in that zip code. NA values were also identified and addressed in the median income variable. The largest chunk of missing data is the number of cats in New York City.

Multiple linear regression was used to impute the missing values of 'median_income'. A stepwise regression was used to choose the independent variables to use to develop this multiple linear regression. The VIF of our initial 8-factor model shows that 'hispanic_population' and 'other_alone_or_in_combination' are highly correlated, so we removed 'other_alone_or_in_combination' and reran the model. The 7-factor model had no VIF concerns, so we used this model to impute the missing 'median_income' values. After imputing the NA values, 'median_income' was also scaled.

Finally, we imputed the missing values of 'n_cats' using multiple linear regression. We also used stepwise regression here to choose the independent variables to use in our model. We ran a multiple linear regression model with the 5 variables indicated by the stepwise regression but found that 'population' was not statistically significant. Therefore, we ran a 4 factor model which we used to impute the missing 'n_cats' values.

Our cleaned dataset has 238 unique data points with over 20 variables.

## Insights from EDA

After initial exploration and modeling, the team identified that our key variables are pet count (as dependent variables), number of parks, % of residents by race (such as white population %, black population % and etc.), median household income, population, and population density. The correlation matrix of our independent variables (Figure 1) shows that median income has a strong positive correlation with white population %, Hispanic population % has a strong positive correlation with other_race_population %, and female population % has a negative correlation with male population %. The results indicate that we need to evaluate correlated variables for our models. Doing a principal component analysis to transform a large set of variables into a smaller one is worthwhile.

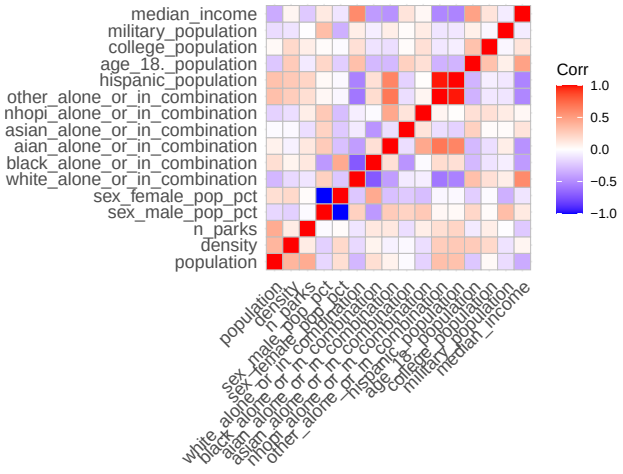


Figure 1: Correlaton Matrix for Independent Variables

# Overview of Modeling

## Approach/Methodology

The modeling for this project utilized multiple linear regression and k-means clustering to identify significant independent variables. The multiple linear regression models provided formulas based on significant variables to estimate pet count and cat to dog ratio for a given zip code. This approach allows the most flexibility on business decisions as it produces a raw count of pets, which can be converted to density. The ratio of cats to dogs can be used in conjunction with the raw count of pets to get the number of cats and dogs. Companies can then collaborate across departments to make decisions for marketing, inventory, and more based off the predicted number and density of pets and species. On the other hand, we felt that a k-means clustering model would be a great choice to see if there were segments of zip codes with similar characteristics that a pet-supply business would be able to use to make more strategic decisions. Some validation and variable selection methods we used in the project were stepwise regression, VIF, and K-fold cross validation. In our kmeans clustering model we used PCA for dimensionality reduction. Stepwise regression and VIF were used in the multiple linear regression portion of the project, and after evaluating the initial versions of our multiple linear regression models, the team implemented log transforms to improve model performance. K-fold cross validation and PCA were used in the k-means clustering portion of the project.

## Multiple Linear Regression to Predict Number of Pets

The first model we wanted to explore was a multiple linear regression to predict the number of pets in each zip code, without any transformations on the variables. However, after evaluating the Linear-Linear model, we found that several data transformations were required for this project. To visually check for outliers, we used a histogram and boxplot (Figure 2). To find the individual outliers and determine if they should actually be removed, we used grubbs.test(). Then, we used a normal Q-Q plot to see if the remaining 'pet_count' data followed a normal distribution.

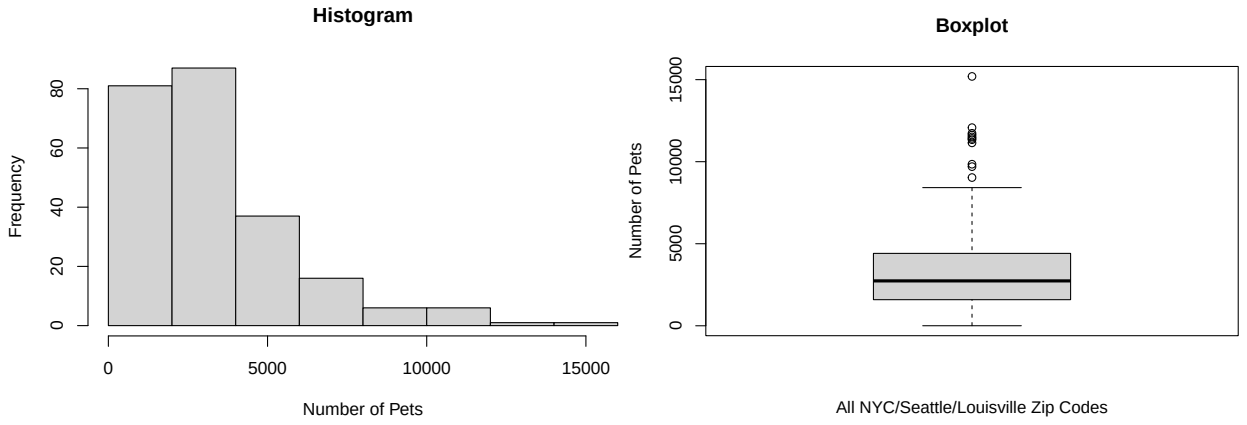


Figure 2: Histogram and Boxplot checking for outliers in Dependent Variable

Grubbs.test() identified the two highest 'pet_count' values as outliers and they were removed from future models. Stepwise regression was conducted to produce the final multiple linear regression model without log transformations for 'pet_count'. Significant factors included white, American Indian/Alaska Native, Hispanic, and Native Hawaiian/Pacific Islander demographics, population, number of parks, population density, population age, and median income. A variance inflation factor check showed no concerns for multicollinearity within the model and the adjusted R-squared was 0.6538. However, the residuals vs. fitted, normal Q-Q, scale-location, and residuals vs. leverage plots all indicated that the model could be improved. As can be seen in Figure 3, the multiple linear regression model is heteroskedastic.

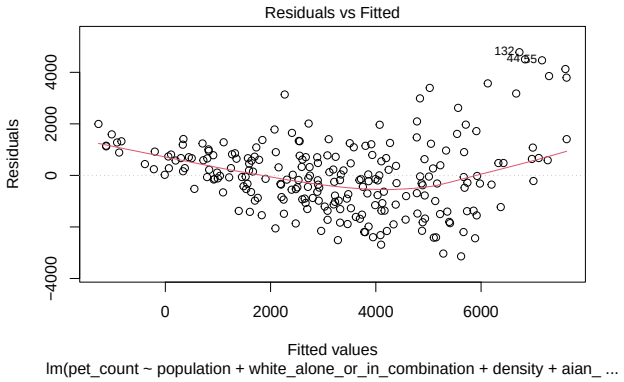


Figure 3: Residuals v. Fitted, homoskedasticity check

After observing heteroskedasticity in our model, we performed log transformations on our dependent variable and independent variables, and built Log-Linear, Linear-Log, and Log-Log models to see if we could find a better model for predicting pet count per zip code.

When we transformed the dependent variable to create a Log-Linear model, the adjusted R-squared increased to 0.696, so the Log-Linear model performed better than the original Linear-Linear model which had an adjusted R-squared of 0.6538.

Then, we tested Linear-Log models. We ran many Linear-Log models (but not all due to having a large number of possible combinations) using different log transformed independent variables and they yielded adjusted R-squared values between 0.5853 and 0.7386. These adjusted R-squared values ranged from lower than the R-squared value of the Linear-Linear model to higher.

Finally, we ran multiple Log-Log models. Once again, we did not try all combinations of transformations due to having a large number of possible combinations. Within trying just a couple of combinations, we were able to determine that a Log-Log model would be the best choice as the adjusted R-squared values were significantly higher. Once we determined this, we decided which version to use by considering the R-squared value and the Residuals vs. Fitted graph. The final model we decided on for predicting the number of pets by zip code had the following variables:

**Dependent Variable:** Ln_pet_count

**Independent Variables:** Ln_population,
Ln_white_alone_or_in_combination, density,
Ln_aian_alone_or_in_combination, Ln_age_18._population,
Ln_hispanic_population, Ln_n_parks, median_income,
nhopi_alone_or_in_combination

We decided on this model because it had a high adjusted R-squared of 0.8591, and its Residuals vs. Fitted graph indicated homoskedasticity.

To summarize the log transformations to predict pet count using multiple linear regression, the team ran several iterations of log transformations. The Log-Linear model yielded an improved adjusted R-squared of 0.696. Linear-Log models yielded adjusted R-squared's between 0.5853 and 0.7386. The Log-Log models produced the largest adjusted R-squared values, and the final Log-Log model we decided on had an R-squared value of 0.8591 and addressed the heteroskedasticity of the original Linear-Linear model (Figure 4).

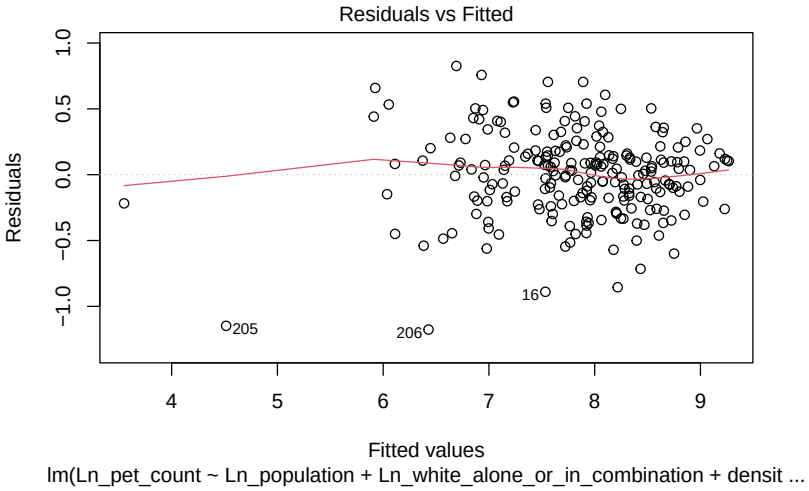


Figure 4: Residuals vs Fitted, Log-Log Model

Although we decided on our model, we also wanted to evaluate the quality of our model using a training and test set to ensure that the high R-squared value was not due to overfitting. The training model is not our final model - we are simply using it to evaluate the quality of our full model discussed above. We put 75% of the data points in the training set and 25% in the test set. We found that the R-squared when we predicted the values of the test set was 0.8798, which is an indicator that we have a good model (without an overfitting issue) for predicting the number of pets by zip code.

## Multiple Linear Regression to Predict Ratio of Cats to Dogs

After finding the best multiple linear regression model to predict the number of pets by zip code, we took similar steps to find the best multiple regression model to predict the ratio of cats to dogs by zip code.

To visually check for outliers, we used a histogram and boxplot (Figure 5). To find the individual outliers and determine if they should actually be removed, we used grubbs.test(). Then, we used a normal Q-Q plot to see if the remaining ratio of cats to dogs data followed a normal distribution.

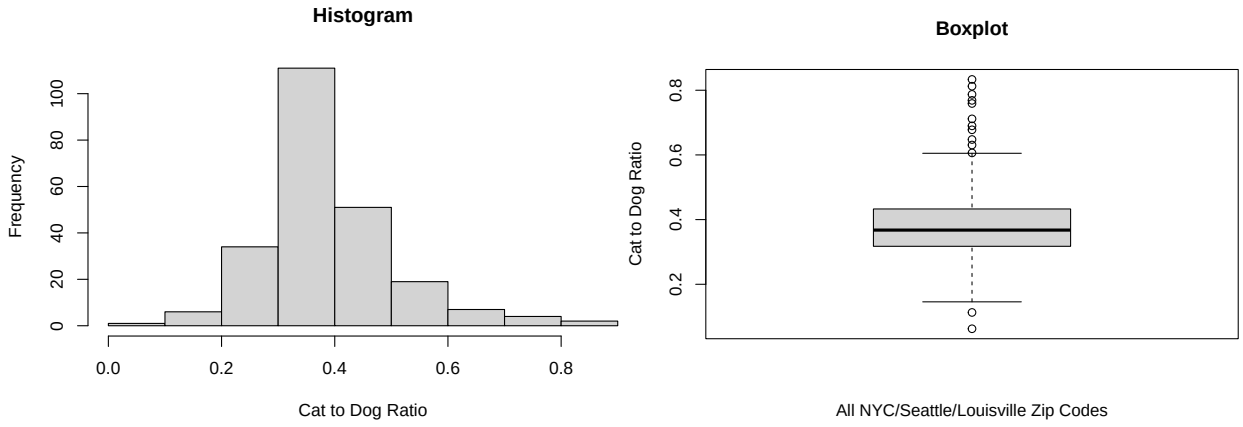


Figure 5: Histogram and Boxplot, dependent variable outlier checks

Grubbs.test() identified the five highest 'cats_v_dogs' values as outliers and they were removed from future models. Stepwise regression was conducted to produce the final multiple linear regression model without log

transformations for 'cats_v_dogs'. Significant factors include white, American Indian/Alaska Native, and Asian demographics, population, population density, number of parks, and female population percentage. A variance inflation factor check showed no concerns for multicollinearity within the model and the model had an R-squared of 0.3944. However, multiple plots indicated that the model could be improved. As can be seen in Figure 6, the Linear-Linear multiple linear regression model is heteroskedastic.

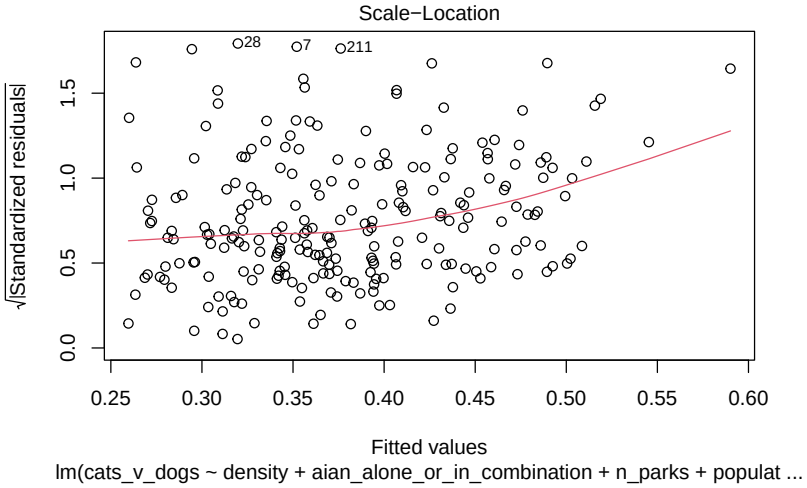


Figure 6: Scale-Location, Homoskedasticity Check

After observing heteroskedasticity in our model, we performed log transformations on our dependent variable and independent variables, and built Log-Linear, Linear-Log, and Log-Log models to see if we could find a better model for predicting ratio of cats to dogs per zip code.

With the Log-Linear model, the adjusted R-squared increased to 0.4328, so the Log-Linear model performed better than the original Linear-Linear model which had an adjusted R-squared of 0.3944.

Then, we ran many different Linear_Log models (but not all due to having many possible combinations) using different log transformed independent variables and they yielded adjusted R-squared values between 0.3705 and 0.4722. These adjusted R-squared values ranged from lower than the R-squared value of the Linear-Linear model to higher.

Finally, we ran multiple Log-Log models. Once again, we did not try all combinations of transformations due to having a large number of possible combinations. However, the couple initial combinations we tried quickly indicated that a Log-Log model would be the best choice as the adjusted R-squared values were significantly higher than for the other models. Once we determined this, we decided which version to use by considering the R-squared value and looking at the homoskedasticity graphs. The final model we decided on for predicting the ratio of cats to dogs by zip code had the following variables:

**Dependent Variable:** Ln_cats_v_dogs

**Independent Variables:** Ln_density, aian_alone_or_in_combination, n_parks, population, asian_alone_or_in_combination

We decided on this model because it had a high adjusted R-squared of 0.5316, and its Residuals vs. Fitted and Scale-Location graphs indicated homoskedasticity.

To summarize the log transformations to predict the ratio of cats to dogs using multiple linear regression, the team ran several iterations of log transformations. The Log-Linear model yielded an improved adjusted R-squared of 0.4328. Linear-Log models yielded adjusted R-squared's between 0.3705 and 0.4722. The Log-Log models produced the largest adjusted R-squared values, and the Log-Log model discussed above had an R-squared of 0.5316 and addressed the heteroskedasticity of the original Linear-Linear model (Figure 7).

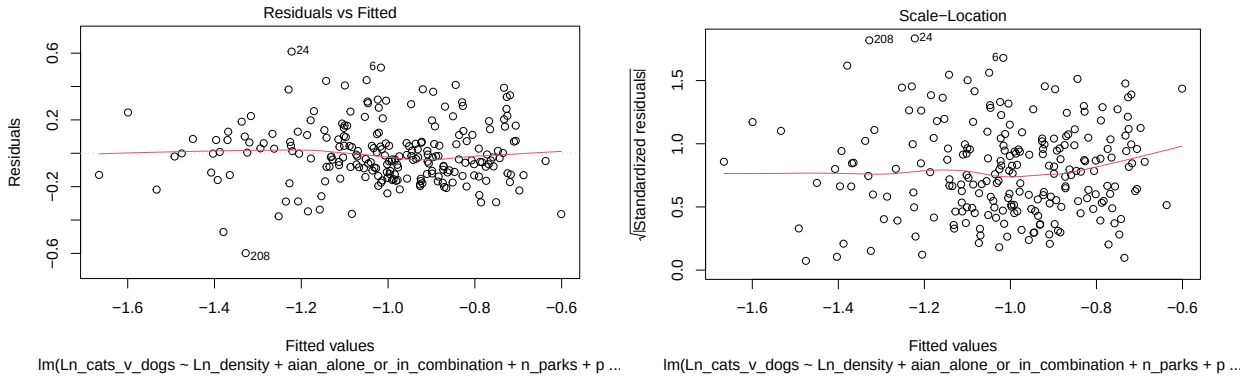


Figure 7: Residuals vs Fitted and Scale-Location, Homoskedasticity Checks

As we did with the model to predict pet count by zip code, we wanted to evaluate the quality of our model using a training and test set to ensure that the high R-squared value was not due to overfitting. The training model is not our final model - we are simply using it to evaluate the quality of our full model that we discussed above. We put 75% of the data points in the training set and 25% in the test set. We found that the R-squared when we predicted the values of the test set was 0.5779, which is higher than the R-squared of our final model. This is an indicator that we have a good model (without an overfitting issue) for predicting 'cats_v_dogs'.

### Summary of Multiple Linear Regression Models

We tried many different types of multiple linear regression models to predict both the number of pets and the ratio of cats to dogs by zip code. First, we ran Linear-Linear models without any transformations, using stepwise regression to choose the variables that would go into our model. While the adjusted R-squared values for both of our Linear-Linear models were good, they were heteroskedastic. Therefore, we wanted to test if we could create better models by using different combinations of log transformations. We tried a Log-Linear model, multiple Linear-Log models, and multiple Log-Log models for predicting both 'pet_count' and 'cats_v_dogs'. For both the 'pet_count' and 'cats_v_dogs' models, we found that variations of the Log-Log model had the best R-squared values, and also fixed the heteroskedasticity issues we saw in the Linear-Linear models. After determining our best models, we also evaluated the quality of our models by splitting the data into training and test sets, and making predictions for the test sets using models trained on training data. The R-squared values calculated from the predictions vs. the actual values in the test set indicated that our models had strong R-squared values and were not overfit.

We believe that our two final Log-Log multiple linear regression models can be very useful for pet companies as they will be able to estimate the number of pets and the ratio of cats to dogs in each zip code through readily available data on the population and area. This approach allows a company to assign its own assumptions as to the best business strategies for an area with a certain number of pets and certain ratio of cats to dogs. For example, in a zip code with a very high pet count and a low ratio of cats to dogs, pet companies may consider opening up a store there and stocking up with more inventory for dogs than for cats.

The next approach aims to identify similarities amongst these variables that would better inform these business decisions.

## K-Means Clustering

With a k-means clustering model, we tried to determine if there were nicely separable clusters of zip codes that had features such that a pet-supply business could organize their business and marketing strategy

around them. For example, if we observed that the cities were easily partitioned into 3 clusters, and those 3 clusters were such that one could be identified as the low-income cluster, another as the middle-income, and the last as high-income then it could be possible to brainstorm differing strategies between these groups. For example: all low-income zip code - or regions that have many low-income zip codes - the business could opt perhaps not to open a brick and mortar store of their own. They might instead choose to focus on their wholesale offerings to big box retailers. The idea being that low-income areas that have the need for pet supplies might prioritize cost-efficiency over quality. Middle and high income markets might be targeted for a brick and mortar store, but the inventory mix could be different between these clusters. Ultimately, understanding the types of markets a business might serve helps to understand the type of strategies a business should use.

The initial attempt for a model removed zip code, latitude, longitude, and the city dummy variable. These are not true continuous numeric data that measures the value of something. The balance of the data is then scaled. To tune the hyperparameter $k$, a scree plot is used to identify a value of $k$ that delivers diminishing returns in the total within cluster sum of squares, i.e. the elbow.

The initial model had 16 features and returned 5 clusters of significance (Figure 8).

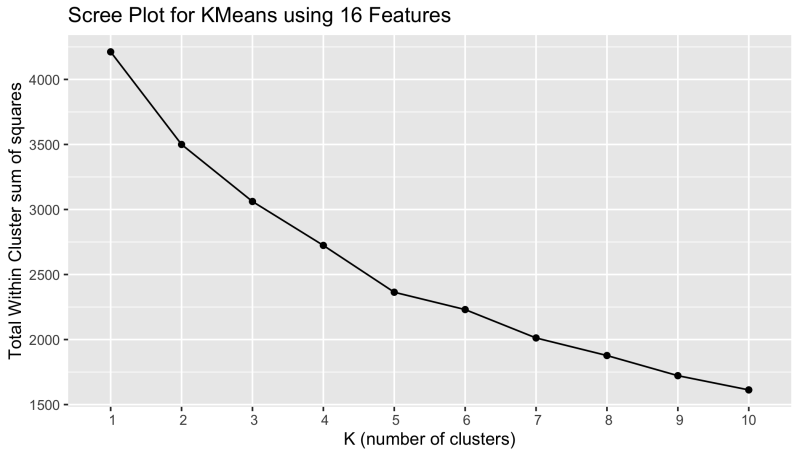


Figure 8: Scree plot of kmeans model w/ 16 Features

However, upon further investigation it appeared that there was a datapoint that had an inordinately high value for military_population that skewing the results in cluster (which, incidentally, had the lowest sample size - Table 1). Being that kmeans is a model based on Euclidean distances, we questioned if this with a single observation could be significantly skewing the assigned clusters. Likewise, we questioned if being in the military affects the likelihood of pet ownership. From our inquisition into this, we decided to remove military_population as a feature.

Table 1: Average of each variable among 5 cluster groups, and sample size

| n | n_dogs | n_cats | population | density | white_alone_or_in_combination | military_population |
|---|---|---|---|---|---|---|
| 39 | 89.128205 | 90.33333 | 65.69231 | 76.25641 | 70.97436 | 0.000000 |
| 18 | 9.944444 | 12.50000 | 28.44444 | 27.11111 | 48.44444 | 22.055556 |
| 38 | 33.605263 | 35.26316 | 50.34211 | 53.15789 | 18.81579 | 0.000000 |
| 48 | 59.541667 | 64.29167 | 70.31250 | 67.47917 | 24.60417 | 0.000000 |
| 92 | 42.717391 | 38.68478 | 36.57609 | 32.65217 | 67.60870 | 2.163044 |

On top of this, input variables were further refined to condense racial demographics to reduce correlation. Meaning, the racial profile of the population is now represented by a single value, proportion of population

that is white, and the inverse of this would be the population that is minority. Proportion of adult population and college-educated population are retained. Now running a kmeans model on 9 features, the optimal value of clusters is not as clear as before, but the scree plot (Figure 9) seems to suggest k=3.

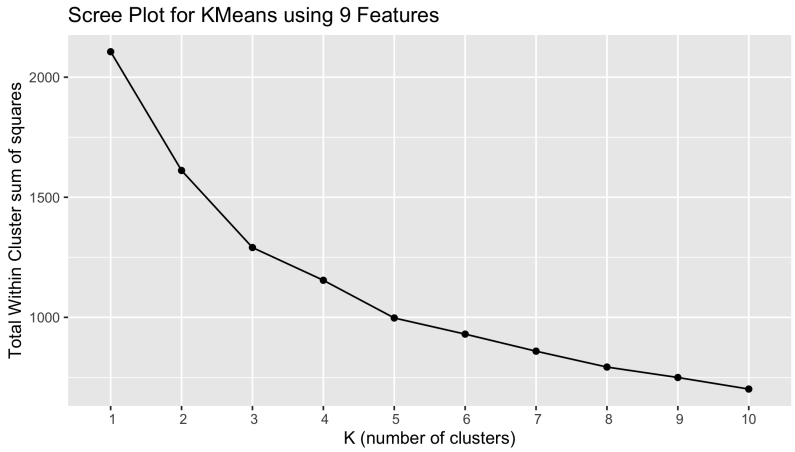


Figure 9: Scree Plot After Data Cleaning

Taking the average of each feature per cluster seems to suggest the significant variables that are easily separable are population, population density, number of parks, population demographics, population age, higher education, and median income (Table 2).

Cluster 2 contains the most pets (near 85th percentile for both cats and dogs) and is in the 74th percentile of population density. This cluster is predominantly white, is in the 80th percentile of median income, with most being above 18 years of age and having had a college level education. Cluster 1 has the highest overall population and is predominantly minority. They are in the low to high 60 percentiles of pet ownership and the 21st percentile of median income. Cluster 2 might be a prime candidate for a brick and mortar store whereas cluster 1 would be more receptive to a mailer coupon for online retail.

Table 2: Average statistics and count of 3 cluster groups

| labels | n | n_dogs | n_cats | population | density | n_parks | white_alone_or_in_combination | age_18._population | college_population | median_income |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 69 | 62.49275 | 68.14493 | 80.65217 | 73.72464 | 66.04348 | 27.42029 | 36.34783 | 30.59420 | 21.85507 |
| 2 | 34 | 83.58824 | 86.02941 | 58.20588 | 74.44118 | 53.88235 | 74.64706 | 82.29412 | 68.61765 | 79.91176 |
| 3 | 132 | 34.59091 | 31.10606 | 31.68939 | 31.11364 | 37.15909 | 55.50000 | 48.71970 | 18.65152 | 56.90909 |

# Kmeans using PCA

To determine how nicely separated the clusters truly are, the nine variables would need to be plotted against one another. Since a nine-dimensional graph would be difficult to visualize, the team implemented principal component analysis (PCA) to project the data in lower dimensionality. After the transformations, it appears that 3 principal components can capture 72% of the variance. The kmeans model is then applied to the 3 principal components.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7648 1.5470 0.9927 0.94841 0.79591 0.63436 0.57224
## Proportion of Variance 0.3461 0.2659 0.1095 0.09994 0.07039 0.04471 0.03638
## Cumulative Proportion  0.3461 0.6120 0.7215 0.82143 0.89182 0.93653 0.97291
```

```
##                             PC8     PC9
## Standard deviation     0.48923 0.06654
## Proportion of Variance 0.02659 0.00049
## Cumulative Proportion  0.99951 1.00000
```

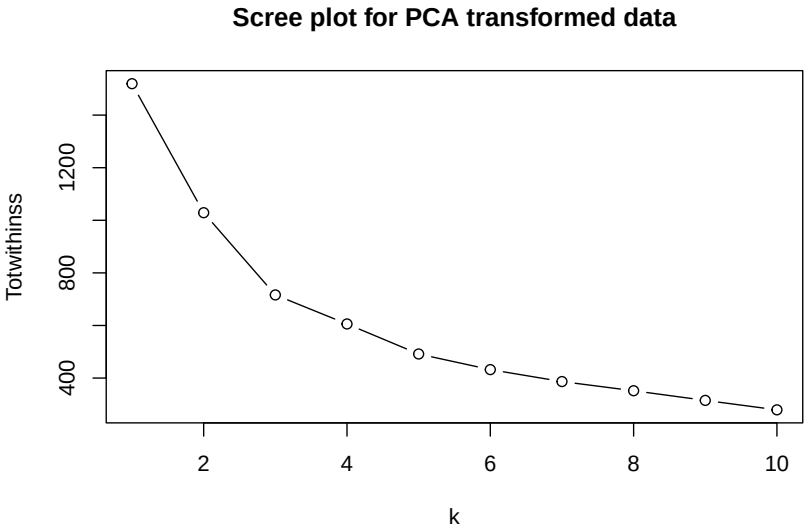


Figure 10: Scree plot for PCA transformed data

After using the PCA transformation and running kmeans on only 3 principal components, a scree plot still seems to suggest $k = 3$ (Figure 10). We can see how well partitioned these clusters are by plotting the principal components and coloring points by cluster assignment.

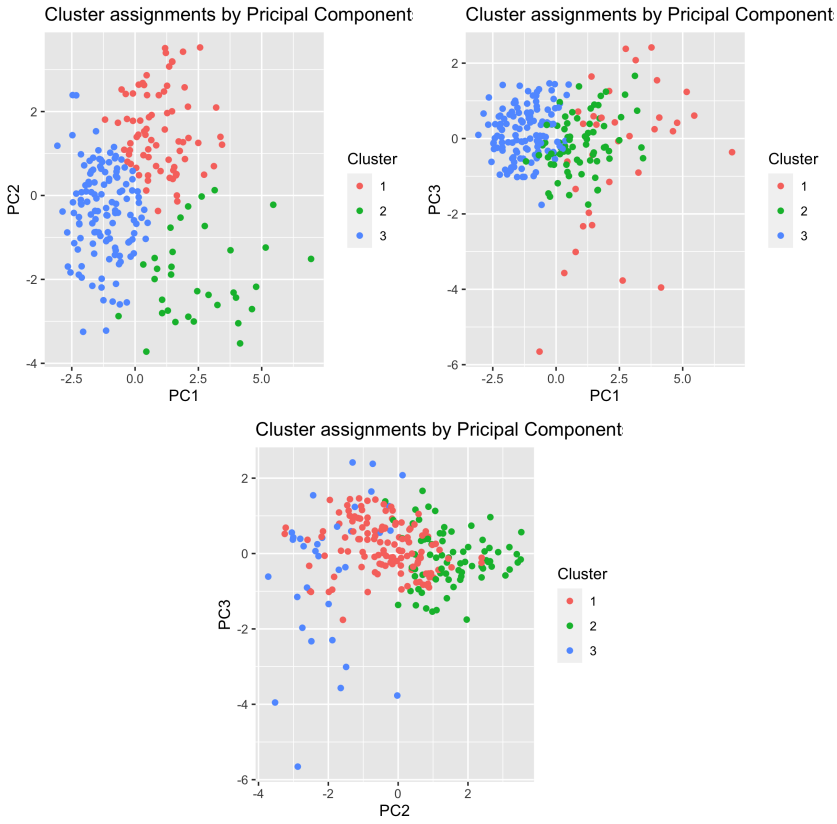


Figure 11: Visualized clsuters against principal components

# Conclusion

We set out in this project wondering what factors of a person make it likely that they would have a pet. In the same vein, a more generalized business applicable question became: how do we determine that a zip code has pets? This information, and its more granular forms (how many dogs/cats, etc.) are all useful to pet supply business and it can help them consider the markets they serve in the present while also strategizing for the future.

Our hypothesis was that there would be one variable or a set of variables that appeared to statistically and significantly correlate with the amount of pets in an area. We've discovered through a log-log transformed linear model that the number of pets has the strongest linear relationships with the below variables.

- total population
- proportion of white population
- proportion of American Indian and Alaska Native population
- proportion of Hispanic population
- proportion of adult population
- proportion of Native Hawaiian & Pacific Islander population
- population density
- median income
- number of parks

Likewise, through another log-log transformed multiple linear regression, we observed that the ratio of cats to dogs heavily correlated to:

- population density
- proportion of American Indian and Alaska Native
- number of parks
- total population
- proportion of Asian population

The only new variable from before is the proportion of Asian population.

Additionally, the kmeans clustering model evidenced that we could segment our market into three with uniquely trending features for number of pets, total population, proportion of white population, proportion of adult population, proportion of college-educated population, and the median income.

Picking a set of variables that imply causality is beyond the scope of this course. We believe that our models showcase great potential for what pet-supply companies would be able to do with this kind of data. Still, it would be worth the investment for pet companies to find or purchase more data in order to strengthen these models before applying them and taking action across the United States.

With this information, a pet supply company could develop targeted strategies around the areas they believe to have pets utilizing the results from the clustering model. They could determine whether or not an area is likely to have a good amount of pets to cater to through the first linear regression model. They could use the second one to understand the ratio of cats to dogs for an area that might have many pets such. All this information helps them to better service their customers.

If there were more time and resources, we might've tried to use t-SNE to transform the dimensionality of our data in kmeans rather than PCA. As mentioned previously, pet registration data is inconsistent. Our team searched for an wide for data that we believed would be easily obtained, but provided rather scarce. It might be beneficial for the company to conduct random surveys in cities that are not covered to reduce the possibility of conflation between the variables we've determined to be statistically significant and other unforeseen variables that are inherent to the cities for which we had pet registration data. Alternatively, a pet supply company could buy data from sources that may have already put these surveys together. In our own exploration, we did find paid datasets regarding pet ownership statistics. Regardless, our exploration into this problem has yielded a set of applicable features in different contexts that can be tested with further research.

## Works Cited

Applebaum, Jennifer & Peek, Chuck & Zsembik, Barbara. (2020). Examining U.S. pet ownership using the General Social Survey. The Social Science Journal. 60. 1-10. 10.1080/03623319.2020.1728507.

Perry, S.L. and Burge, R.P. (2020), How Religion Predicts Pet Ownership in the United States. Journal for the Scientific Study of Religion, 59: 190-201. https://doi.org/10.1111/jssr.12637

"Pet Industry Market Size, Trends and Ownership Statistics." American Pet Products Association, www.americanpetproducts.org.