Springboard Data Analytics Course

Exploratory Data Analysis for Capstone I

25May2019

John L. Parsons

**Exploratory Data Analysis Summary**

The objective of the Exploratory Data Analysis (EDA) Summary is to briefly review the 24 attributes found in the PetFinder Adoption Dataset from Kaggle competition and provide insights from these attributes using visualization techniques. Several attributes that are discussed below for EDA from the Bar plot, Box and Whiskers Plot and group means functions in python are: **Color, MaturitySize, Vaccination, State, Age, Quantity** and **AdoptionSpeed**.

Correlations plots and heatmaps were generated to determine the strength of association between the independent variables and the dependent variables from the PetAdoption dataset. The correlation matrix, Lasso Model and the Prediction Screening using JMP SAS was used to select the top eight to 10 variables for the model selection to optimize the model.

**Introduction**

The PetFinder Adoption Dataset from the Kaggle competition contains a total of three files that can be used for predicting the target variable **AdoptionSpeed.** The **AdoptionSpeed** attribute is an ordinal variable that shows the length of time an animal stays in the shelter. The training file will be used to create the machine learning algorithms and is a metadata set that contains a total of 24 attributes and 14,993 rows of data. The **PetAdoption** Speed variable has a total of five levels or categories and will be the target variable. The remaining 23 attributes are the input variables that can be used to predict **PetAdoption** speed. Figure 1 in the Appendix gives an overview of the attributes for this study, the unique values that can be found in each attribute and the number of missing values that are found in the Name and Description variables for this study. All other variables did not have any missing values as shown in this figure and by PD Profiler.

**Exploratory Data Analysis**

Attributes were plotted against the target variable to see if any differences could be seen that may help model performance. The **Color**, **MaturitySize**, **Vaccination** and **State** variables showed several differences. The majority of pets were adopted from State 41 324 which is Melaka or Malaysia (Figure 3). In Malaysia, the yearly average temperature ranges from 70 to 85 degrees Fahrenheit and the monthly rainfall totals range from 2 to 10 inches of rain per month. This type of climate may be more conducive to smaller and shorter haired animals compared to other individuals in this dataset. The most common **MaturitySize** of a pet selected by this population was medium size animals for all five categories of **AdoptionSpeed** and then small breeds were the second choice selected in this study (Figure 3). **Color1** Attribute for Figure 2 shows the majority of animals selected from the adoption centers was Black and then the second choice was the Brown Coat Color. The animals that were not vaccinated for all categories, except 4 of **AdoptionSpeed** were selected the most from the shelters. The vaccinated animals for the last category were selected the most for the animals in the shelters the longest. These animals had a better chance of being adopted if it was vaccinated compared to non-vaccinated animals.

The **Age** and **Quantity** attributes were plotted with **AdoptionSpeed** using the Box and Whiskers plot (Figure 6). These plots had no outliers for the lower whisker, but had outliers for the upper whiskers for all five **AdoptionSpeed** Categories. The variable **Age** is in months the animals were adopted out of the shelter. There are no outliers for the lower quartiles and this shows the animals that were less than 10 months of age were adopted first compared to the older animals. The animals over 12 months of age had more variation and distribution for all five levels. The **Age** attribute may need to be transformed by log transformation or drilled down to age in years instead of months to minimize the spread of the distribution. The **Quantity** variable is the number of animals represented in the profile. When only one animal is in the profile, it had the greatest chance of being adopted as shown in the graph. However, the majority of the profiles only had one animal presented for adoption in the profile and represents around 11,500 of the 14,993 rows of data (Figure 10). The **Quantity** variable had a lot of outliers extending at the last quartile and will need to be transformed or binned to minimize this variation and improve the model performance.

Several attributes were grouped to compare the means between the groups and the **Age** attribute had some interesting results. When the means were grouped by Gender, the male animals were older (12.3 months) than the females (10.7 months) compared to those animals that could not be sexed at this time and were the youngest age (4.8 months). When the **FurLength** was grouped by means, the animals that had long hair were the oldest animals (22.6 months) compared to the short hair animals with the youngest mean age mean (8.6 months). **MaturitySize** for the animals that were small to medium size had the youngest age (9.2 to 10.4 months) compared to the medium and large size dogs which had the oldest mean age (20.2 to 30.8 months).

**Model Selection**

The Pearson, Spearman and other Correlation Plots were generated in Python from this dataset. Figures 7 and 8 shows a heatmap and a ranking of any correlations that can be found between the dependent and independent variables. The top eight attributes selected from this matrix are **Breed1, Age, FurLength, Sterilized, Quantity, Vaccinated, Gender and MaturitySize**. The correlation coefficient, r values in this matrix ranges from 0.108 for **Breed1**, 0.1001 for **Age,** 0.091 for **FurLength** and then drops down to 0.045 for **MaturitySize** of these eight variables. Values ranging from 0.1 to 0.3 are considered a small strength of association and only three of the 23 attributes are in this category. The remaining ones have values less than 0.1 and are a very weak associations for the dependent variable.

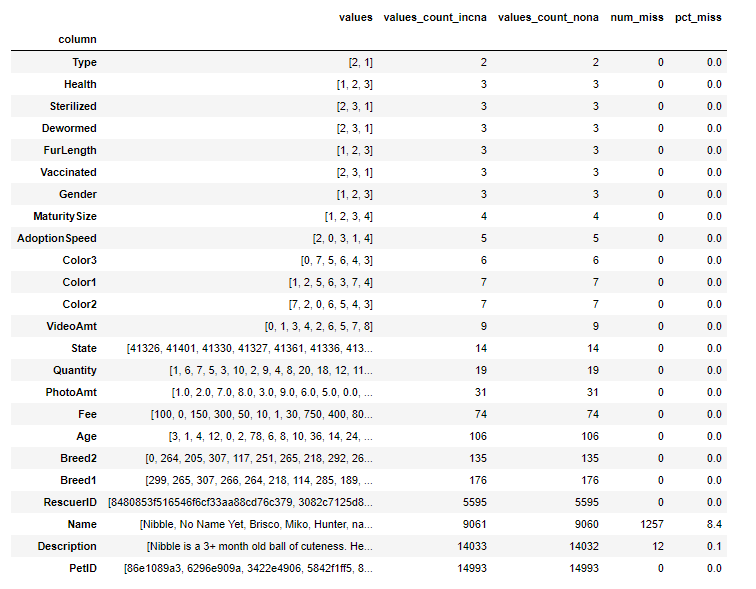
The heatmap in Figure 7 also shows a slight medium to large correlation of several independent variables in the dataset and needs to be reviewed further in the model development phase of this project. Correlations between the independent variables are **Dewormed** and **Sterilized** (0.42), **Gender** and **Quantity** (0.49) and **Vaccinated** and **Sterilized** (0.72) for this dataset. These values have a higher correlation value or Multicollinearity when compared to the other attributes. Selecting only one of the two correlated attributes during model development performance may help increase the predictive power of the model. These correlated pairs are most likely the common procedures done to all shelter pet animals by veterinarians These procedures will help reduce reproductivity rates of pets, prevent the spread of diseases and make the animal more adoptable.

The PetAdoption dataset had a total of 23 input variables and variable selection will be important to reduce the dimensional space and optimize the models for this assignment. The variables that contained text or ID were removed and these are; **RescureID**, **PetId**, **Name**, **Description**, **VideoAmt** and **PhotoAmt**. A Wordplot was generated from the **Description** attribute to give a nice overview of the types of words selected for these animals to be adopted (Figure 11). Several techniques were used on the remaining 17 attributes to reduce the number of variables for this study and these are the correlation matrix between the independent and dependent variables, the feature importance method using the Lasso Model and the Prediction Screening using JMP SAS (Figure 9).

All three variable reduction methods selected all 17 of the attributes, but the top eight variables were **Age, Breed1, Sterilized, FurLength, Type, Quantity, Vaccinated** and **Gender** for JMP SAS. **MaturitySize** was selected for the Correlation Matrix in place of the **Type** variable. The Lasso Feature Importance model had **MaturitySize** and **Dewormed** instead of **Type** and **Age.** The first model selected for prediction will be a Logistic Regression with all 17 numerical categorical variables to give a base line estimate for model performance. This will include the training and validation data for the Accuracy Score with the Confusion Matrix and ROC Curve. When the results from the baseline model have been obtained, these top eight to 10 attributes will be used to determine if this can increase model performance and optimize the model for the next phase of Capstone Project.

**Appendix**

**Figure 1: Percent missing, unique values and attribute names for the PetAdoption dataset.**



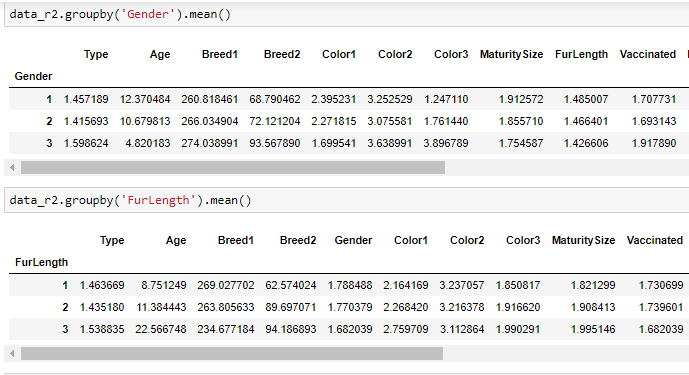
**Figure 2: Bar plots**

|  |  |
| --- | --- |
| **Bar Plot for AdoptionSpeed and Color1** | **Bar Plot for AdoptionSpeed and Vaccination** |
|  |  |

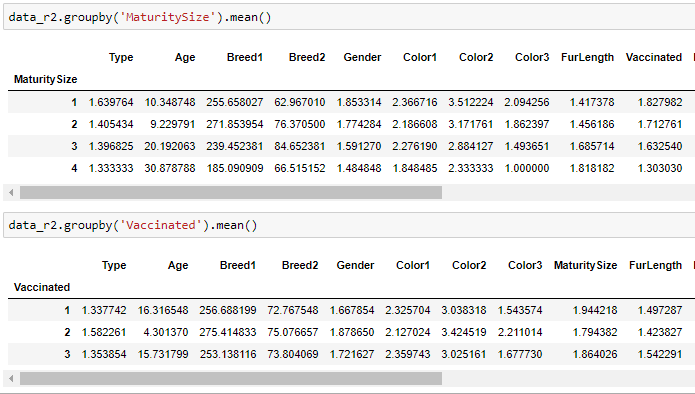
**Figure 3: Bar plots**

|  |  |
| --- | --- |
| **Bar Plot for AdoptionSpeed and MaturitySize** | **Bar Plot for AdoptionSpeed and State** |
|  |  |

**Figure 4: Means comparison between Gender and FurLength and the select 10 attributes.**



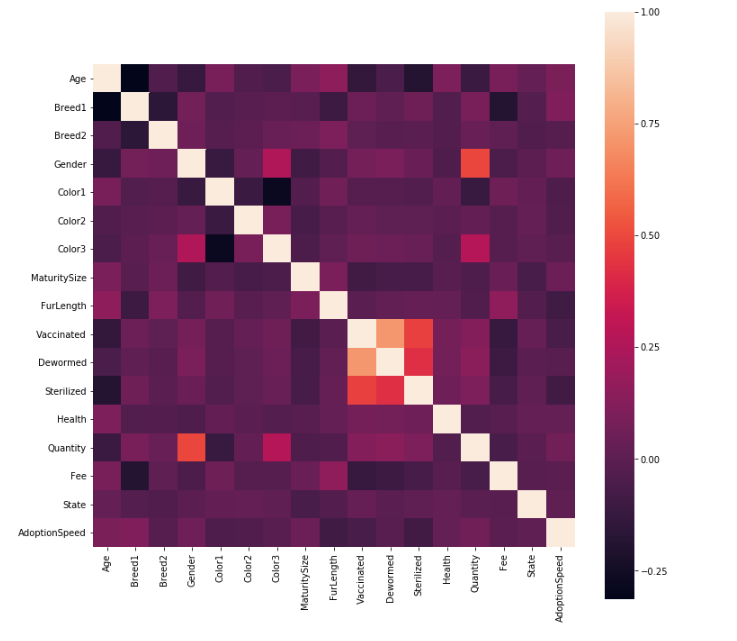
**Figure 5: Means comparison between the MaturitySize and Vaccinated and the select 10 attributes.**



**Figure 6: Boxplots**

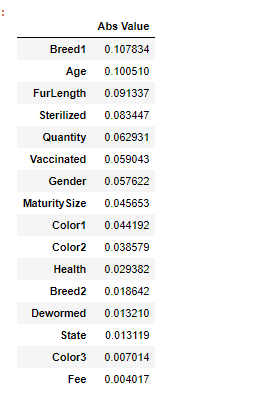
|  |  |
| --- | --- |
| **Boxplot for Age and AdoptionSpeed** | **Boxplot for Quantity and AdoptionSpeed** |
|  |  |

**Figure 7: Correlation Matrix between the dependent and independent variables**



**Figure 8: Correlation Matrix Ranking of values between the**

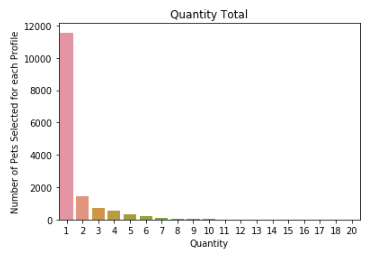
**dependent and Independent Variables.**



**Figure 9: Variable Selection for the Model Development Phase.**

|  |  |
| --- | --- |
| **Feature Importance Method using the Lasso Model.** | **Prediction Screening using JMP SAS for the 23 Attributes** |
|  |  |

**Figure 10: Distribution of the Quantity Attribute for Pet Adoption**



**Figure 11: Word Bubble for the top words used for the Description Attribute**

