“Predicting Pet Adoptability Using Machine Learning: *A Case Study of Petfinder.my Dataset”*

A close-up of words

Description automatically generated

# Introduction

## Background Information: Petfinder.my, a prominent online pet adoption platform in Malaysia, plays a vital role in saving stray animals, primarily dogs and cats, by finding them homes. The mission is to reduce the number of animals euthanized in shelters and provide a loving environment for them.

## Problem Statement: A significant challenge for Petfinder.my is predicting the adoptability of pets. Various factors like age, breed, and health impact how quickly pets are adopted, making this process complex and unpredictable.

## Objective: The aim is to use machine learning to predict how quickly a pet will be adopted after being listed on Petfinder.my. By analyzing a dataset of around 150,000 animals, this project seeks to enhance the efficiency of the adoption process and improve overall animal welfare.

# Data Understanding and Preparation

**Data Source:** The dataset for this analysis comes from Kaggle and it specifically addresses the topic of pet adoption speed, providing detailed information on various pets waiting for adoption.  
**Data Description:** The dataset comprises 4 tables and 24 columns, encompassing a wide array of information about the pets. The key focus is on the target variable 'Adoption Speed', which indicates how quickly a pet is adopted. Other columns include details about the pet's age, breed, color, health status, and more.

**Data Cleaning and Preprocessing:** The preparation of the dataset for analysis involved several crucial steps:

* *Removing Outliers and Normalizing Numerical Columns:* Outliers were removed to ensure a more accurate analysis. This step is critical in preventing skewed results such as age and fee due to extreme values.
* *Handling of Categorical Data:* One-hot encoding was utilized for categorical data. This process involves converting categorical variables such as gender and color into a form that could be provided to machine learning algorithms to do a better job in prediction.
* *Feature Engineering:* The project included creating new features like 'OneColor' to simplify the color-related data and binning of certain variables like 'Breed1' into categories (e.g., 'Pure Dogs', 'Mixed Dog', 'Domestic Short Hair Cat', etc.) for more straightforward analysis.
* *Data Transformation:* Transformation techniques like Min-Max Scaling were applied to normalize data, which is essential for algorithms that are sensitive to the scale of input variables.
* *Handling Missing Values:* Missing data were addressed appropriately, for instance, replacing missing names with 'NULL' and transforming them into a binary 'HasName' or 'NULL' category.

Each of these steps was aimed at refining the dataset to make it more suitable for the machine learning models that would follow in the analysis. The goal was to ensure that the data was clean, well-structured, and ready for effective modeling.

# Methodology

**Candidate Models & Rationale:**In our project, we chose specific machine learning models to best analyze pet adoption speed. We started with the KNeighbors Classifier because of its effectiveness in capturing similarities among pets, crucial for understanding adoption trends. Logistic Regression was an obvious choice for its simplicity and efficiency, particularly useful for binary aspects of our dataset. We included Decision Tree Classifier for its ease of interpretation, helping us identify key factors influencing adoption speed. Lastly, we used Random Forest, an advanced model that combines multiple decision trees to address the complexity of our data and avoid overfitting. Each of these models was selected for its unique strengths in providing insights into our dataset.

**Implementation Details:** Overall, our strategy was to choose models that could effectively tackle different facets of our dataset, balancing simplicity, and complexity. We were meticulous in tuning their parameters and used a variety of evaluation metrics to thoroughly understand their performance. In the Exploratory Data Analysis (EDA) part of your project on "Predicting Pet Adoptability Using Machine Learning," several analyses and visualizations were conducted to understand the dataset better. Here is a summary of the issues found and their solution:

|  |  |  |
| --- | --- | --- |
| Column | Issue | Solution |
| Adoption Speed | Imbalanced | Stratified Splitting |
| Name | Missing Values | Name vs. No Name |
| Age | Outliers, Right Skewed | Bins |
| Breed | Too Many Categories | Bin Breed 1-6 Categories |
| Maturity Size | Bias | Don’t Neglect |
| Fee | Outliers, Right Skewed | Bins |
| Photo/Video Amount | Right Skewed | Normalize |
| Description | Missing Values | Use DescriptionWordsCount |
| State | High VIF Score | Drop |

EDA was followed by a series of correlated data transformations, such as:

* *Data Splitting*: First, we divided our dataset into training and testing sets. This split was crucial for assessing our models on new, unseen data, ensuring a robust evaluation of their predictive performance.
* *Model Training*: We then trained each model on the training set. This step involved tweaking key parameters to optimize their performance. For instance, in the KNeighbors Classifier, we adjusted the number of neighbors, while in Logistic Regression, we fine-tuned the regularization strength. For Decision Trees and Random Forest, we focused on the depth of the trees and the number of estimators.
* *Evaluation*: To evaluate our models, we used metrics like accuracy, precision, recall, and confusion matrices. These metrics were vital in not only quantifying the correctness of our predictions but also understanding the types of errors our models were making. Additionally, we employed cross-validation to ensure our models’ performance was stable across different data segments.
* *Feature Importance*: We paid particular attention to feature importance in our Decision Trees and Random Forest models. This helped us identify which factors were most influential in predicting the speed of pet adoption.

# Experimental Setup

**Dataset Splitting:** We split our dataset into two parts: 75% for training and 25% for testing.   
**Parameter Tuning:** For each model, we tuned specific parameters to optimize performance. For the *KNeighbors Classifier*, we experimented with the number of neighbors, eventually finding that 5 neighbors gave us the best results. In the *Random Forest model*, we initially set the number of estimators to 100. After testing, this showed an accuracy of 38.76% on our test set, indicating how well the model could generalize. With *Logistic Regression*, we adjusted the regularization strength to improve model fit. These steps were crucial in refining our models to accurately predict adoption speeds based on various pet characteristics.

# Results and Discussion

**Model Performance:**

* *KNeighbors Classifier*: Showed a training set score of 54.07%, suggesting a reasonable fit to the training data. However, it had a lower test set score of 31.72%, indicating some challenges in generalizing to unseen data.
* *OneVsRest Classifier with SVC*: Achieved a better training set accuracy of 58.79%, showing a good learning capability. The test set accuracy was 37.27%, an improvement over the KNeighbors Classifier but still indicative of some generalization issues. This model's detailed performance across classes was captured in the confusion matrix and classification report.
* *OneVsOne Classifier with SVC*: Had a training set accuracy of 43.36%, lower than OneVsRest, suggesting some limitations in fitting the training data. However, it slightly outperformed OneVsRest in the test set accuracy with 37.41%.
* *Logistic Regression*: This model had a close training and test score, 38.03% and 37.09% respectively, indicating consistency but with moderate performance. The precision, recall, f1-score, and support for each class were analyzed for a deeper understanding of its performance.
* *Random Forest Classifier*: Before tuning, it exhibited a very high training set accuracy of 99.62%, suggesting excellent fit but also a potential overfitting to the training data. The test set accuracy was 38.76%, which, despite being the highest among the models, highlighted the gap between training and test performance. The analysis of feature importances provided insights into the factors influencing adoption speed predictions.
* *GridSearchCV with Pipeline in Random Forest Model*: Through this approach, we addressed the significant overfitting issue observed in our initial untuned model and determined the optimal parameters for the RandomForestClassifier, which were found to be max\_features=11 and n\_estimators=90. This was achieved by implementing grid search and cross-validation techniques. As a result, we observed an enhancement in model performance: the best cross-validation score reached was 38.84%. Furthermore, there was a notable improvement in the disparity between the training and testing accuracy scores. Post-tuning, the training accuracy score was adjusted to 40.00%, indicating a substantial enhancement in model generalization, attributable to the effective application of hyperparameter tuning.

**Model Comparative Analysis:**

* Logistic Regression showed consistency between training and test performance but with moderate scores.
* KNeighbors Classifier and OneVsRest/SVC demonstrated challenges in generalization, as evidenced by their lower test scores.
* OneVsOne/SVC offered a balance between OneVsRest and Logistic Regression in terms of test accuracy.
* Random Forest stood out with the highest test accuracy, and after tuning its overfitting issue was negligible.

|  |  |  |
| --- | --- | --- |
| Model | Training Set Accuracy | Testing Set Accuracy |
| KNearest Neighbors | 0.5407 | 0.3172 |
| OneVsOne | 0.4336 | 0.3741 |
| OneVsRest | 0.5879 | 0.3727 |
| Logistic Regression | 0.3803 | 0.3709 |
| Tuned Random Forest | **0.4000** | **0.3884** |

**Interpretation:** In our pet adoption speed project, the Random Forest model showed high accuracy and after tuning, it overcame the problem of overfitting, while Logistic Regression offered a balanced but moderately accurate performance. This reflects the classic machine learning trade-off between model complexity and generalizability. Overall, the best model would be the Random Forest Model with optimal parameters (in this case, the parameters are the top 11 maximum features and 90 estimators.

# Conclusion

**Summary of Findings from Exploratory Data Analysis:** Our analysis of the pet adoption dataset has yielded several significant insights that shed light on the factors influencing pet adoption. A notable trend is the age of pets, where older pets are less likely to be adopted, and interestingly, younger pets (1-3 years) also face lower adoption rates. This trend reverses as pets grow older. Breed preferences also play a crucial role, with Pure Dogs and Domestic Short Hair Cats enjoying higher adoption rates among the 307 different breeds documented. Gender impacts adoption speed as well, with male pets being adopted more quickly than females and groups and having a higher overall adoption rate. Color influences adoptability, where black pets boast a 72% adoption rate, and it appears that color affects dogs more significantly than cats. Size is another factor, with smaller pets being adopted faster, although extra-large pets, though less common, have high adoption rates. The preference for fur length is also evident, particularly in dogs, where longer fur is favored. In terms of health, healthy pets dominate the adoption landscape, while those with minor injuries suffer from lower adoption rates. The quantity of pets per listing affects adoption chances, with single pet listings generally faring better than multiple pet listings. The presence and quality of photos and videos in listings correlate with increased adoption chances, and detailed, longer pet descriptions are more attractive to potential adopters.

**Implications for Petfinder.my:** These findings are particularly relevant for Petfinder.my in their mission to facilitate pet adoptions. They suggest a need to focus on older pets and those breeds less likely to be adopted quickly. Optimizing listings by encouraging more photos, videos, and comprehensive descriptions could significantly increase adoption rates. There's also a potential for Petfinder.my to develop targeted marketing campaigns to highlight pets that typically have a lower likelihood of adoption, such as larger pets or those with less favored colors. Furthermore, advising shelters on the implications of adoption fees could be beneficial, given the observed trend of higher adoption rates for pets listed without fees.

* ***Advice to Shelters and Rescuers When Posting a Pet Profile****:*

1. **Pick a Name**: Naming pets in the profile can enhance their appeal.
2. **List Every Pet on Their Own**: Avoid grouping pets in listings to increase individual visibility.
3. **Highlight Key Features**: Emphasizing attributes like long fur length or young age in the description can boost adoption rates.
4. **Health Considerations**: While deworming is beneficial, sterilization or vaccination might not be necessary for increasing adoptability.
5. **Adjust Adoption Fees**: Lower or no adoption fees can lead to higher adoption rates and speed.
6. **Enhance Listings with Media**: Including more photos and videos makes the profile more attractive to potential adopters.
7. **Detailed Descriptions**: Writing comprehensive summaries about the pet can increase interest.

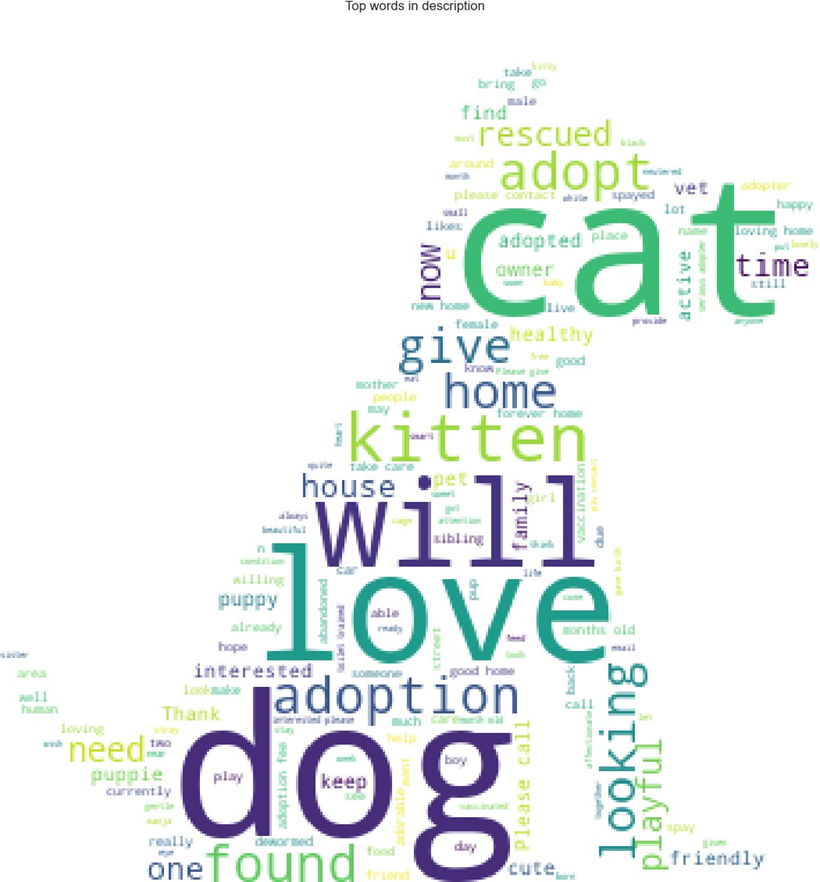
* ***Application of the Prediction Model****:* The prediction model can be adapted into AI tools to guide shelters and rescuers in enhancing their pet profiles' appeal, thereby reducing animal suffering. For instance, shelters can initially assess a profile by running it through the model to estimate adoption speed. If the predicted speed is low, they are advised to make improvements and rerun the model to identify a profile configuration that is predicted to be more attractive.

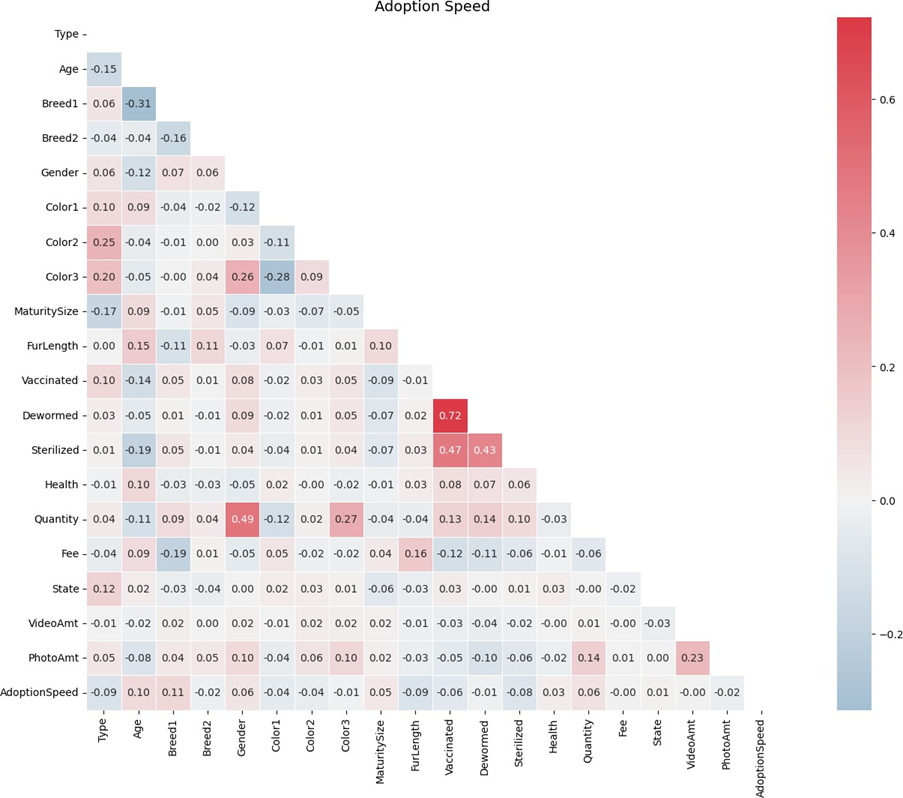
**Limitations and Future Work:** This study has limitations: the dataset's limited geography may not reflect global adoption trends, and the absence of pet behavioral data is a notable gap. The study also doesn't consider time-based or seasonal adoption variations. Future research can expand by including global data, analyzing trends over time, and integrating pet behavior data. Longitudinal studies tracking adopter satisfaction can provide deeper insights into the impact of platforms like Petfinder.my.

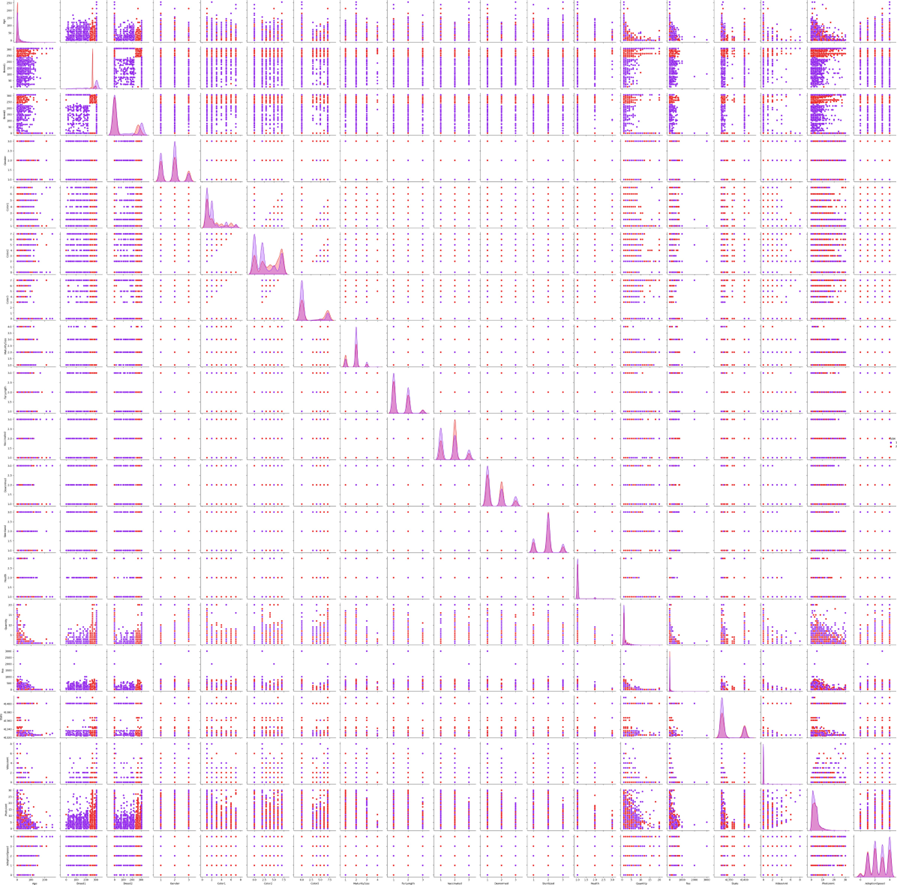
# Appendix

A close-up of words

Description automatically generated

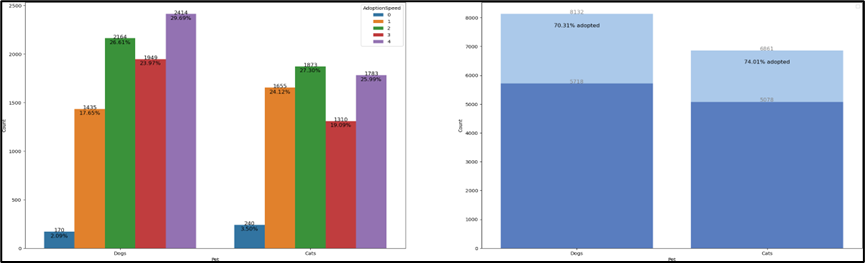


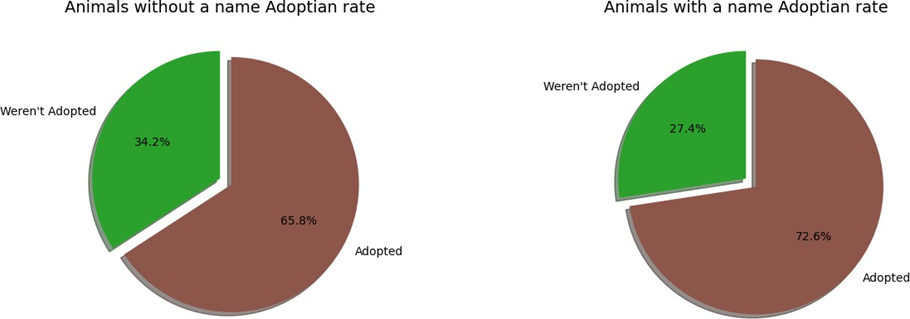




A graph of a bar chart

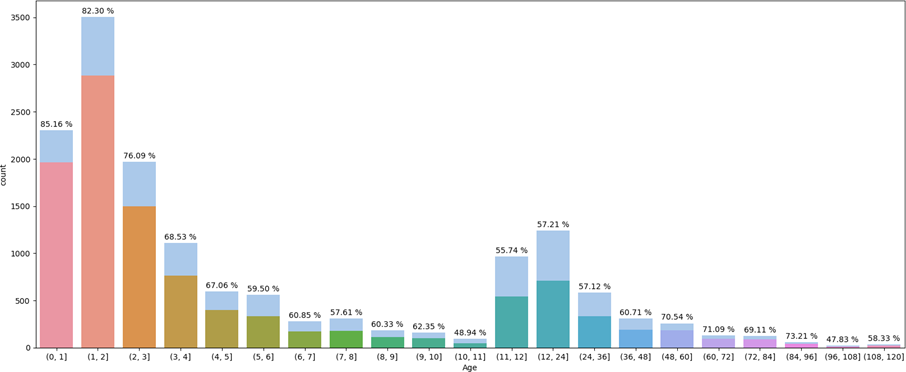
Description automatically generated with medium confidence

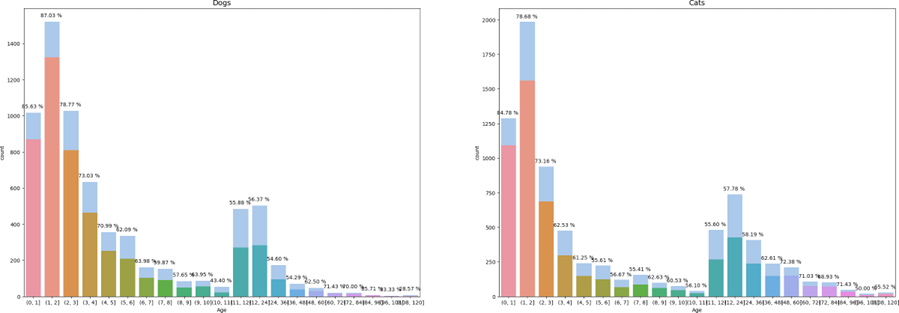


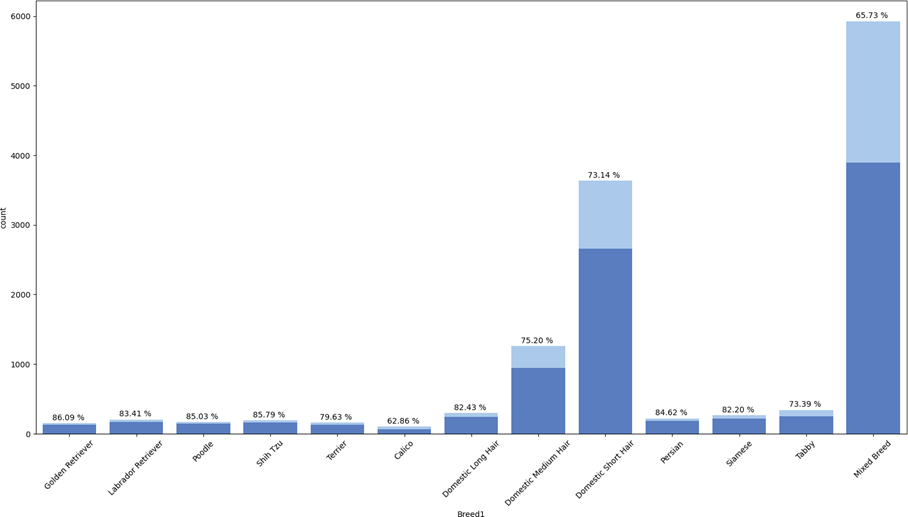


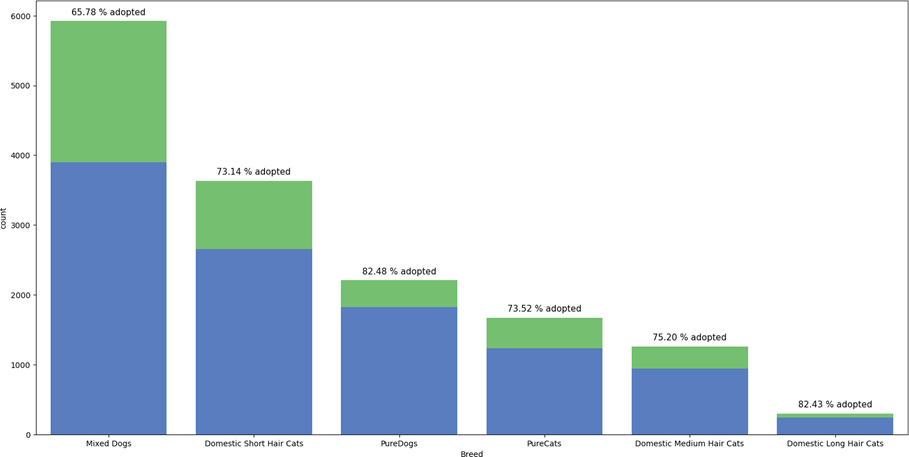
A graph with numbers and lines

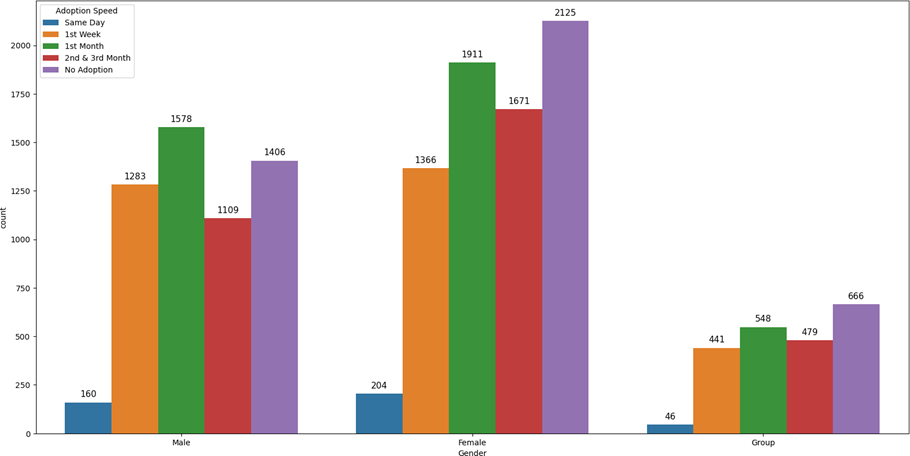
Description automatically generated

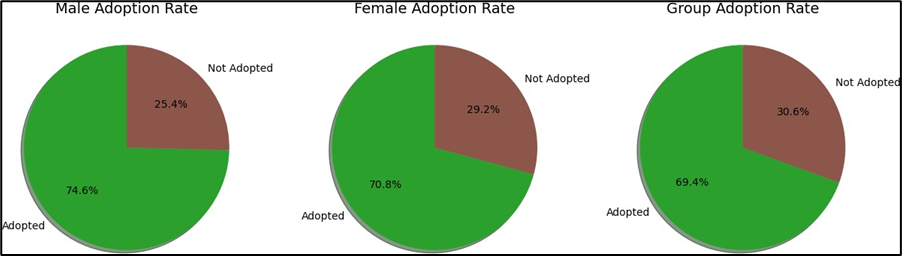


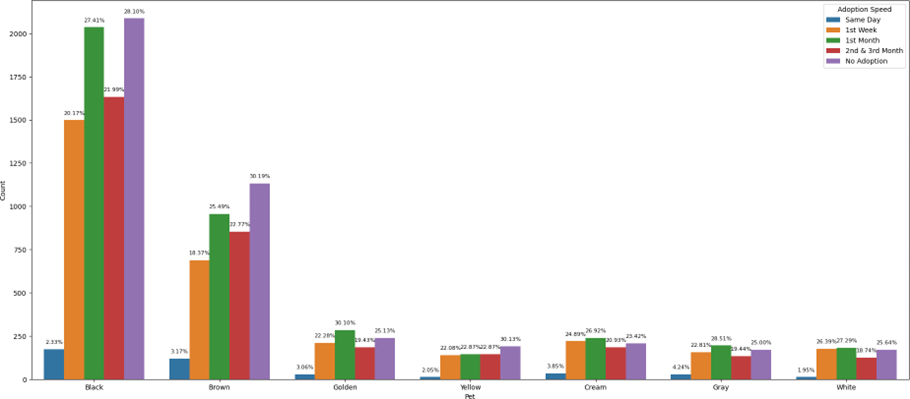


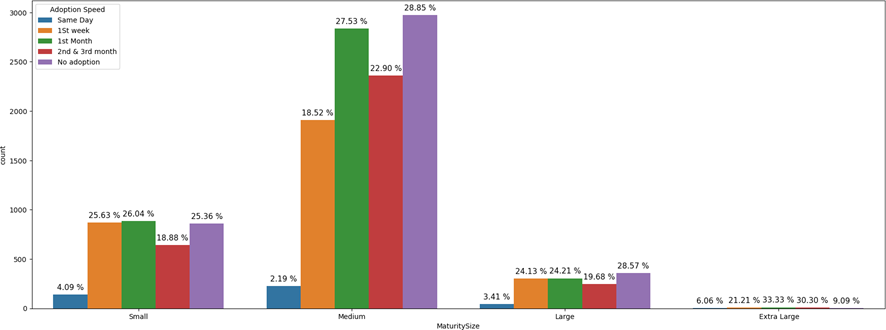












A graph of a bar chart

Description automatically generated with medium confidence

