**Furever Home: Pet Adoption Trend Analysis**

| A Project Report Presented to  DATA-228-11  Spring, 2022 |
| --- |
|  |

| By |
| --- |
| Richa Sharma [014638494]  Bhakti Raichura [016020628]  Khushee Thakker [015271529]  Shreya Srirama [016029845] |
|  |

| Copyright © 2022 |
| --- |
|  |
| Pet Adoption Trend Analysis |

**ABSTRACT**

**Furever Home: Pet Adoption Trend Analysis**

There are currently over 5k independently run animal shelters across the world providing love and care to millions of abandoned animals. While many of these animals are adopted into loving homes, the adoption rates are still worse than we expect. The purpose of this project is to determine what impacts pet adoption rates from an animal shelter. The data set used describes Austin Animal center’s data and is taken from data.austintexas.gov. By doing our research and exploring data on internet we found Austin Animal Center is the largest no-kill animals/pets' shelter in the United States that provides care and shelter more than 18k animals every year, which includes the range of county, city, and state-wide initiatives for the protection and care of abandoned, at-risk, and surrendered animals/pets. Austin Animal Center dataset has attributes for name, date of birth, outcome, animal type, sex and age at time of outcome, breed, and color. By our analysis we will try to understand the range in widely scaled level for the pets' adoptions and transfers to shelters or abandonment of the animals/pets and reason behind the same. We will determine which animal is more likely to be adopted. With this course of action, animal shelters will be able to determine the best course of action to maximize animal adoptions and increase the animal’s potential for getting adopted. This analysis will be carried out utilizing different AWS services for example S3, Pipeline, ETL tools like Python, Redshift etc.

**Acknowledgements**

The authors are deeply indebted to Professor Andrew H. Bond and TA Abdullah Khan for their invaluable comments and guidance in the development of this project throughout the process.

**Table of Contents**

**Chapter 1 Introduction**

* 1. Project goals and objectives
  2. Problem and motivation
  3. Data Description
  4. Literature Survey

**Chapter 2 Architecture and High-Level Design**

2.1 ER Diagram

2.2 Overall Project Flow

2.3 AWS Architecture

**Chapter 3 ETL Process**

3.1 AWS Glue

3.2Python Data Cleaning

**Chapter 4 Data Analysis- Python and AWS Redshift**

4.1 Data Analysis in Python

4.2 Data Analysis in Redshift Query Editor

**Chapter 5 Visualization and Website**

5.1 Data Visualization Using Tableau

5.2 Tableau Desktop connecting with Redshift

5.3 Insight 1:

5.4 Insight 2:

5.5 Insight 3:

5.6 Insight 4:

5.7 Website URL

**Chapter 6 Prediction Modeling**

6.1. Random Forest

6.2. Naive Bayes

6.3. Model’s Comparison

**Chapter 7 Collaboration**

7.1. GitHub

7.2. Team Member and Id’s

**Chapter 8 Conclusion and Future Work**

8.1 Conclusion

8.2 Future work

# Chapter 1 Introduction

* 1. **Project goals and objectives**

Our project helps animal shelter ensure that more animals are adopted by highlighting the key features which make the per adoptable. The objectives of this project are to analyze trends around the animal age, gender, color breed stc and determine and predict the adoptability of the pet. The analysis is done using tools like Python, Tableau and uses multiple AWSServices like S3, GLUE, Redshift etc.

**1.2** **Problem and motivation**

There are currently over 5k independently run animal shelters across the world providing love and care to millions of abandoned animals. While many of these animals are adopted into loving homes, the adoption rates are still worse than we expect As per research paper [Pet Statistics | Shelter Intake and Surrender | ASPCA](https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics) 6.5 million animals enter animal shelters each year. Furthermore, each year 1.5 million shelter animals are euthanized.

**1.3** **Data Description**

Our Dataset was provided by Austin Animal Center. It contains data from 2013 to the present date.

It has 2 datasets

* [Austin Animal Center Intakes](https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Intakes/wter-evkm)
* [Austin Animal Center Outcomes](https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Outcomes/9t4d-g238)

**Intake dataset:**

This dataset contains information about all the animals that were admitted into the animal center. This includes an unique identifier for each animal (animal\_id), intakeDate, location the animal were found, intake type, intake condition, breed color and many others.

**Outcome dataset:**

This dataset contains information about all the animals that were leaving the animal center. This includes an unique identifier for each animal (animal\_id), outcomeDate,sex upon outcome, outcome type, outcome subtype, breed color and many others.

**1.4.** **Literature Survey**

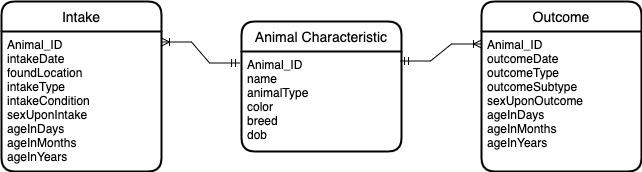
As per research paper [Pet Statistics | Shelter Intake and Surrender | ASPCA](https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics) 6.5 million animals enter animal shelters each year. Furthermore, each year 1.5 million shelter animals are euthanized.

# Chapter 2 Architecture and High-Level Design

* 1. **ER Diagram**

As we can observe there were some common characteristics of the animal in each of these datasets. Hence we normalized it by creating a new dataset only for storing animal characteristics.

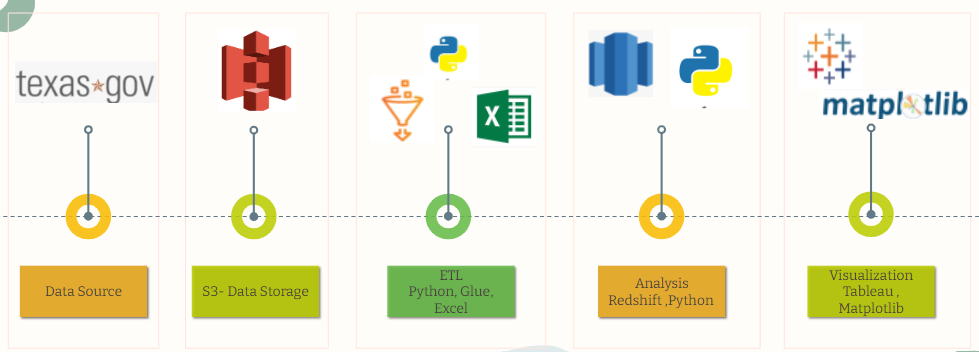
Hence our final data model looks like below

****

**Fig. 1. ER Diagram**

* 1. **Overall Project Flow**

The data is collected from Austin texas government website. For data storage AWS S3 bucket was used. For cleaning and other ETL operations we used AWS Glue , python and excel. Meaningful analysis was done in AWS Redshift and Python. Interactive dashboards were made using tableau.



**Fig. 2. Project Flow**

**Tools we used throughout the project development process and project flow.**

| **Name** | **Usage** |
| --- | --- |
| Python | Data cleaning, Analysis and Predictive analysis |
| Tableau | Data Visualization |
| AWS S3 | Data lake and hosting website |
| AWS Glue | ETL Tool |
| AWS Redshift | Analytical Queries |

* 1. **AWS Architecture**

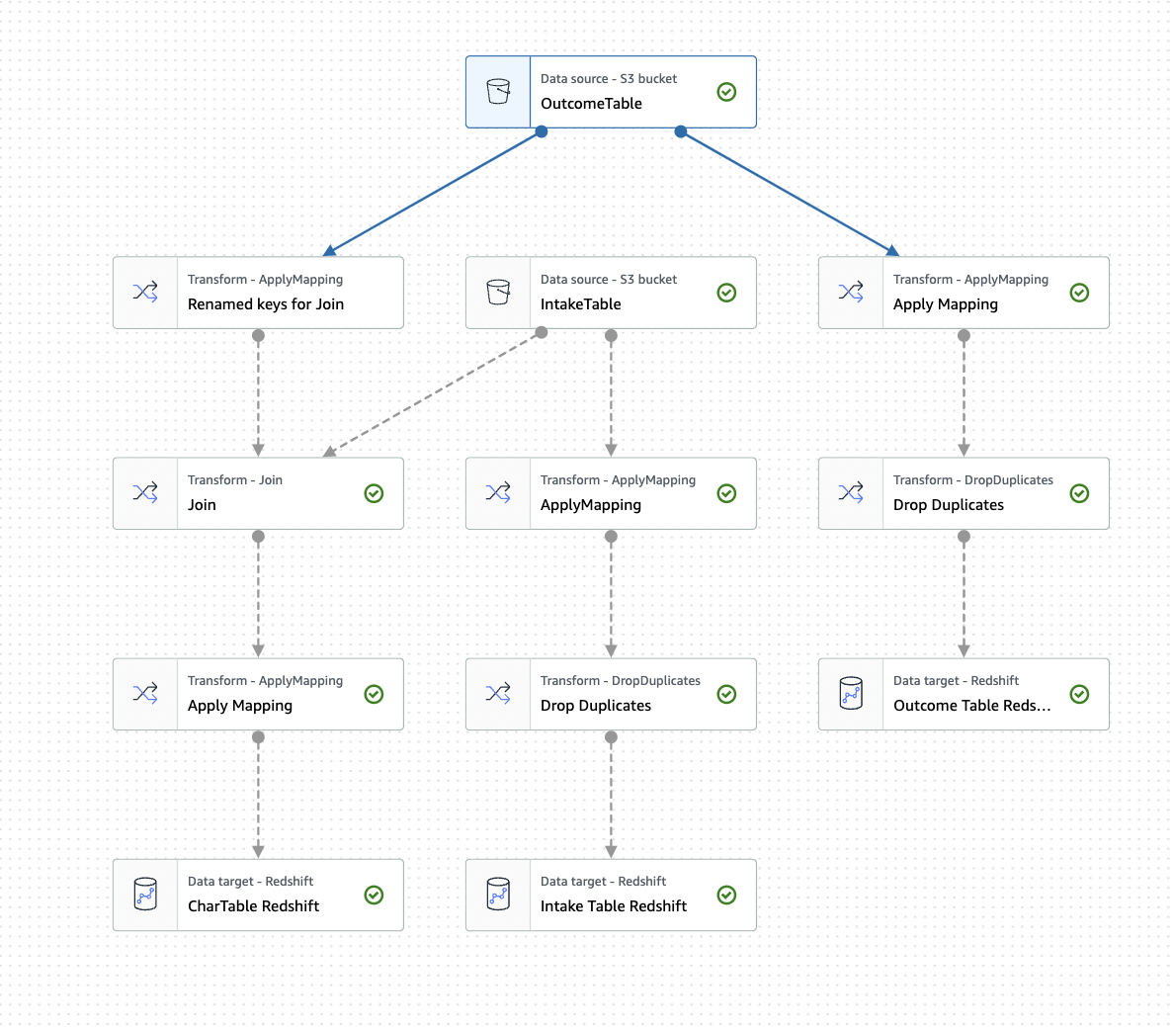
****

**Fig. 3. AWS Architecture**

# Chapter 3 ETL process

**3.1 AWS Glue**

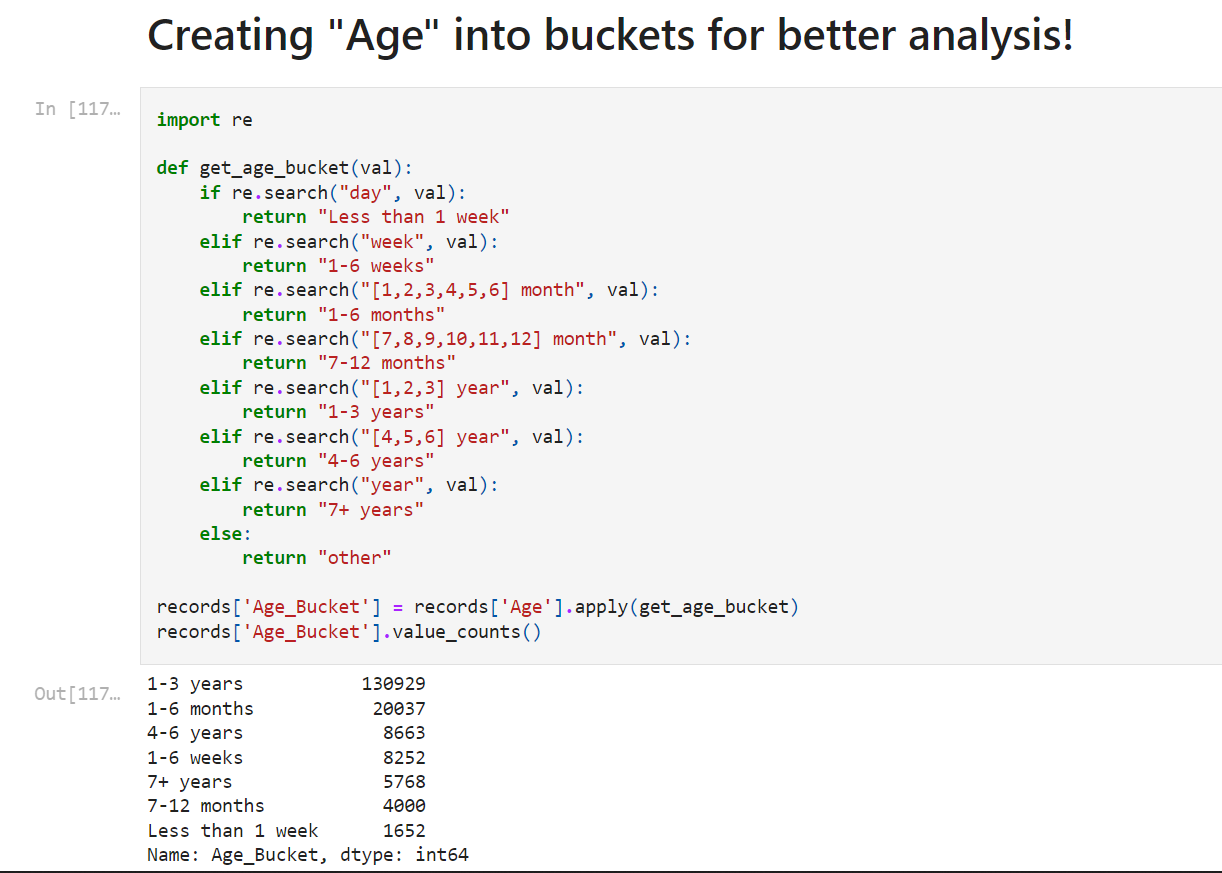
The following was the ETL pipeline architecture which we followed to carry out analysis on the selected dataset for our project.



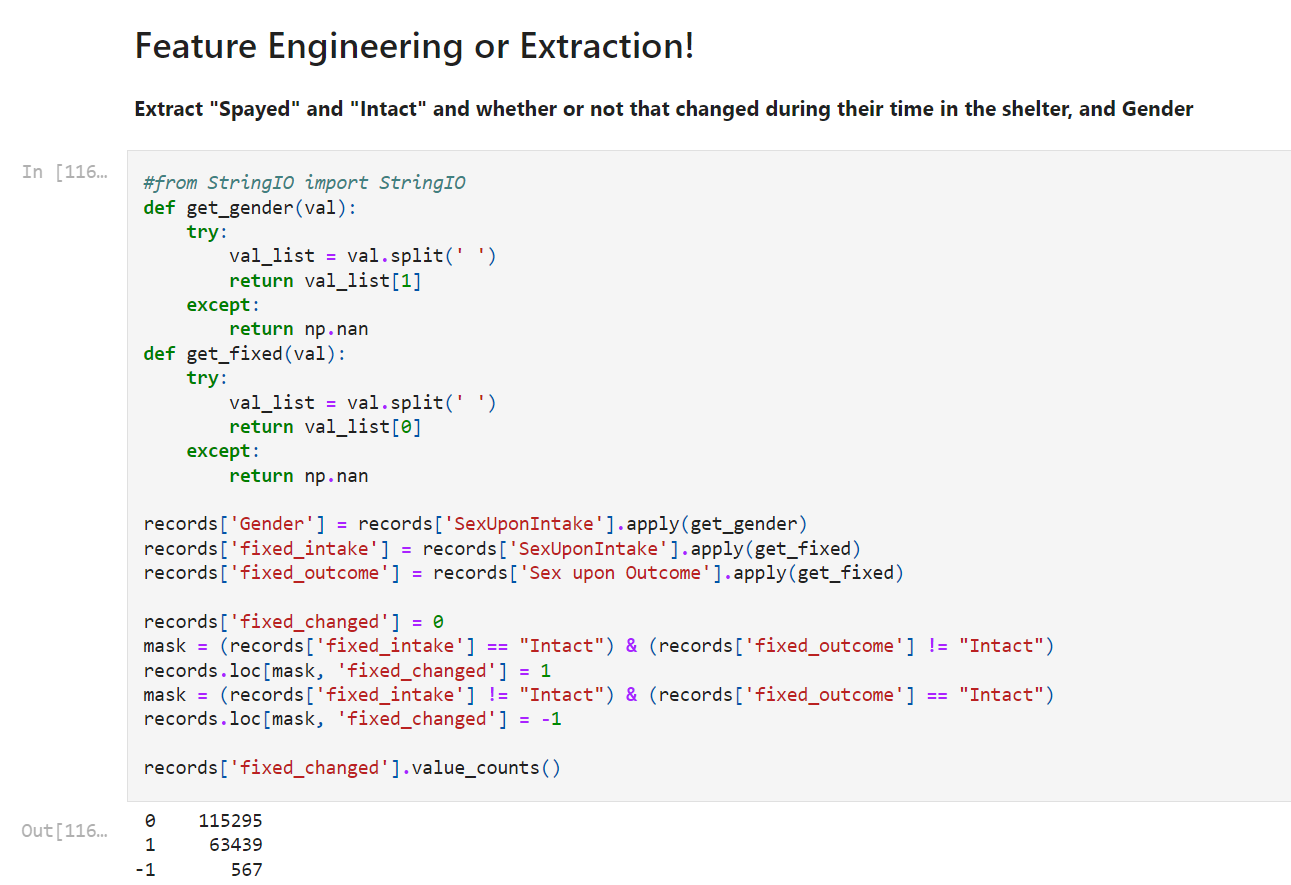
**Fig. 4. ETL Pipeline**

**3.2 Python Data Cleaning**

The raw data has been cleaned using feature engineering in python by removing null values and handling missing values in the datasets. We have dropped the columns which were not contributing towards our analysis and prediction. We have added new derived features from using existing attributes which were helpful for our model design.These derived attributes are bucketing Age into groups,get\_gender etc. After cleaning we have merged all the csv files together and made it one csv file in python for our analysis and machine learning models implementation.



**Fig. 5. Creation of buckets**

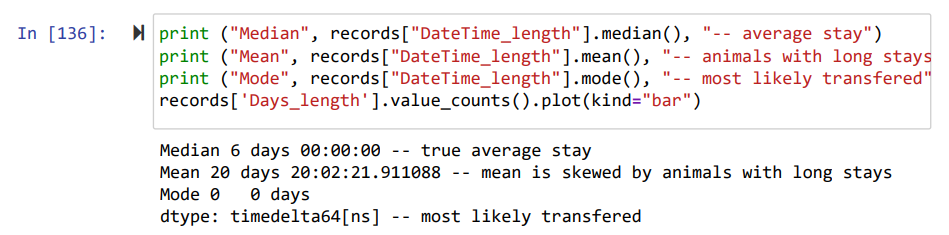
**Fig. 6. Feature Engineering**

# Chapter 4 Data Analysis- Python and AWS Redshift

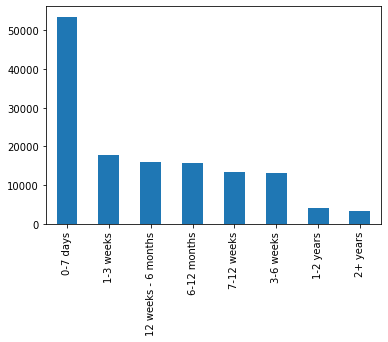
**4.1 Data Analysis in Python**

Data merging is one of the most important and crucial phases of the project. And we have performed analysis of the data using multiple tools and technologies like AWS Redshift, Python and Excel.

1. How long does the average animal spend in the shelter?

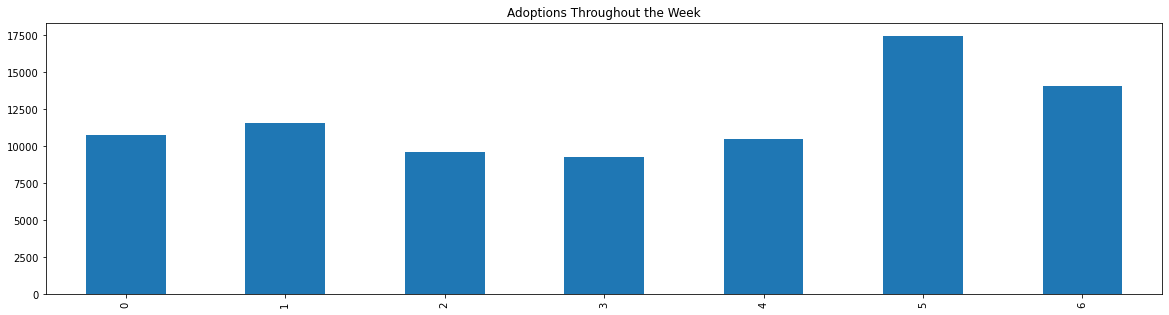


**Fig. 7**

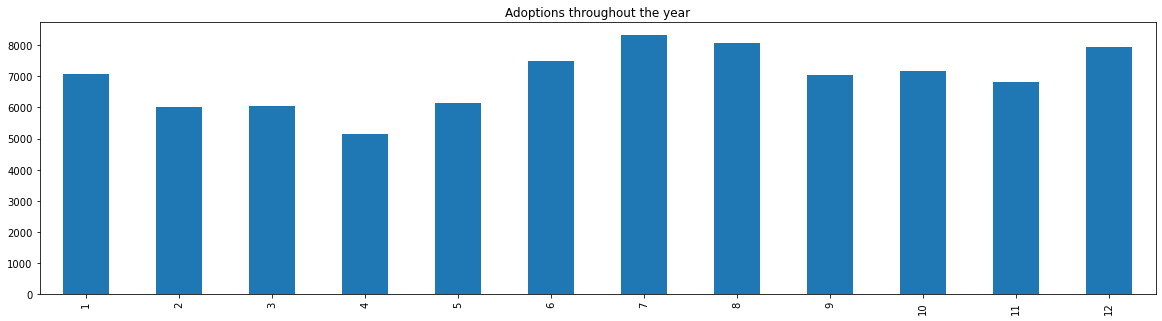


**Fig. 8**

2.Are there certain months that people tend to make more adoptions? What about certain days of the week (weekends)? [0 is Monday and 1 is Jan]



**Fig. 9. Day wise trend**



**Fig. 10. Month wise trend**

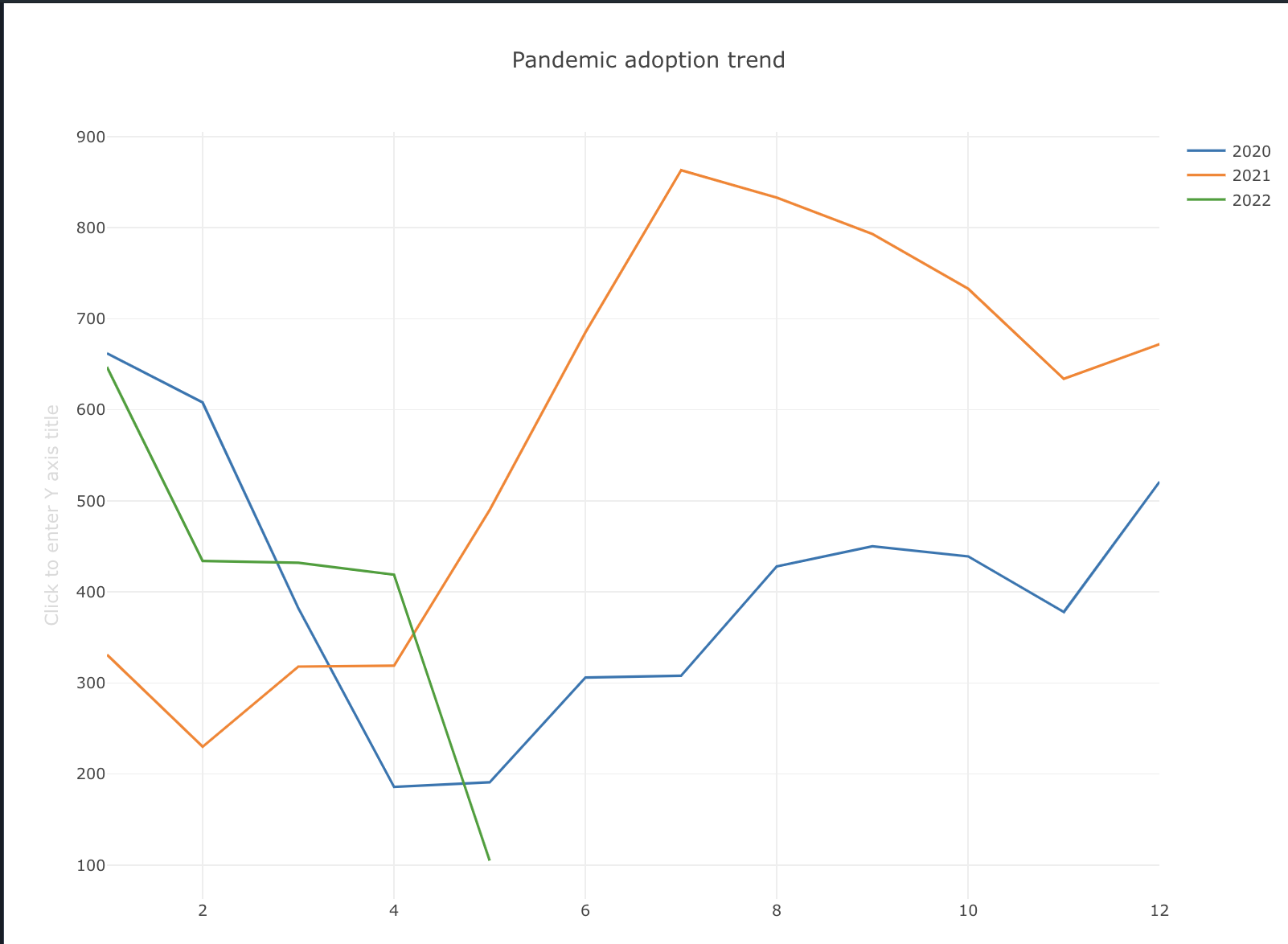
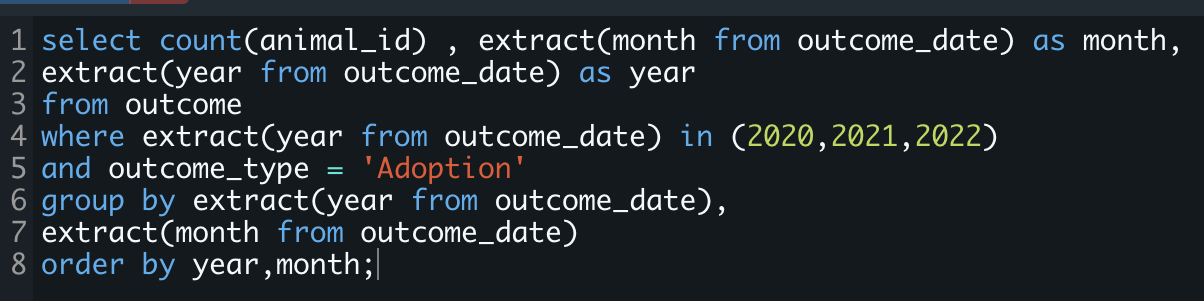
3. Likelihood of adopting cats vs dogs



**Fig. 11.**

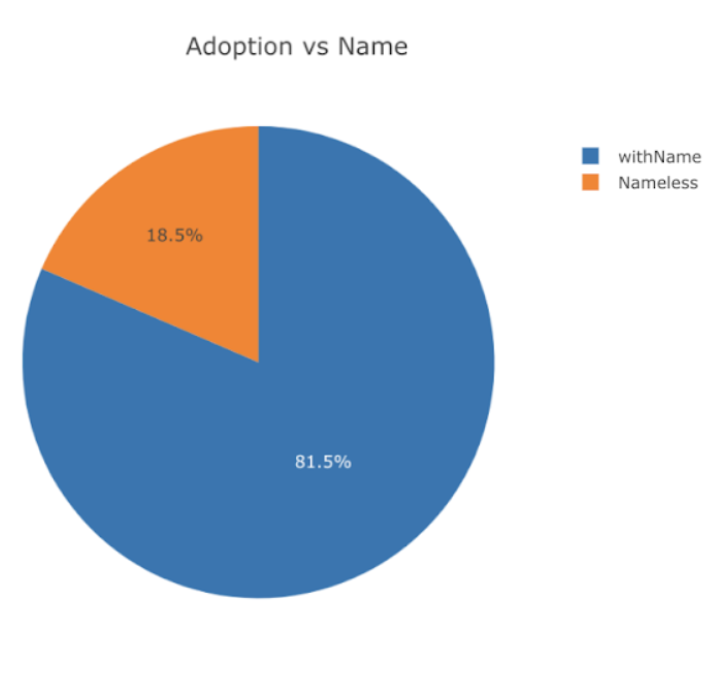
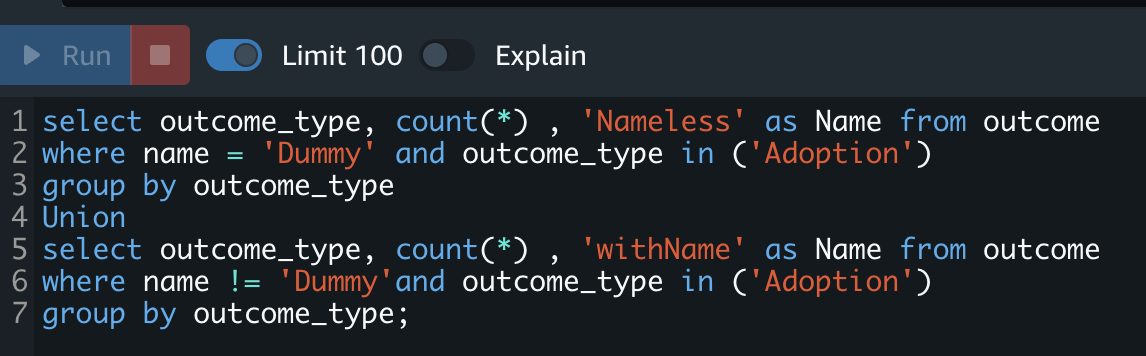
**4.2 Data Analysis in Redshift Query Editor**

1. Adoption Trends during the pandemic



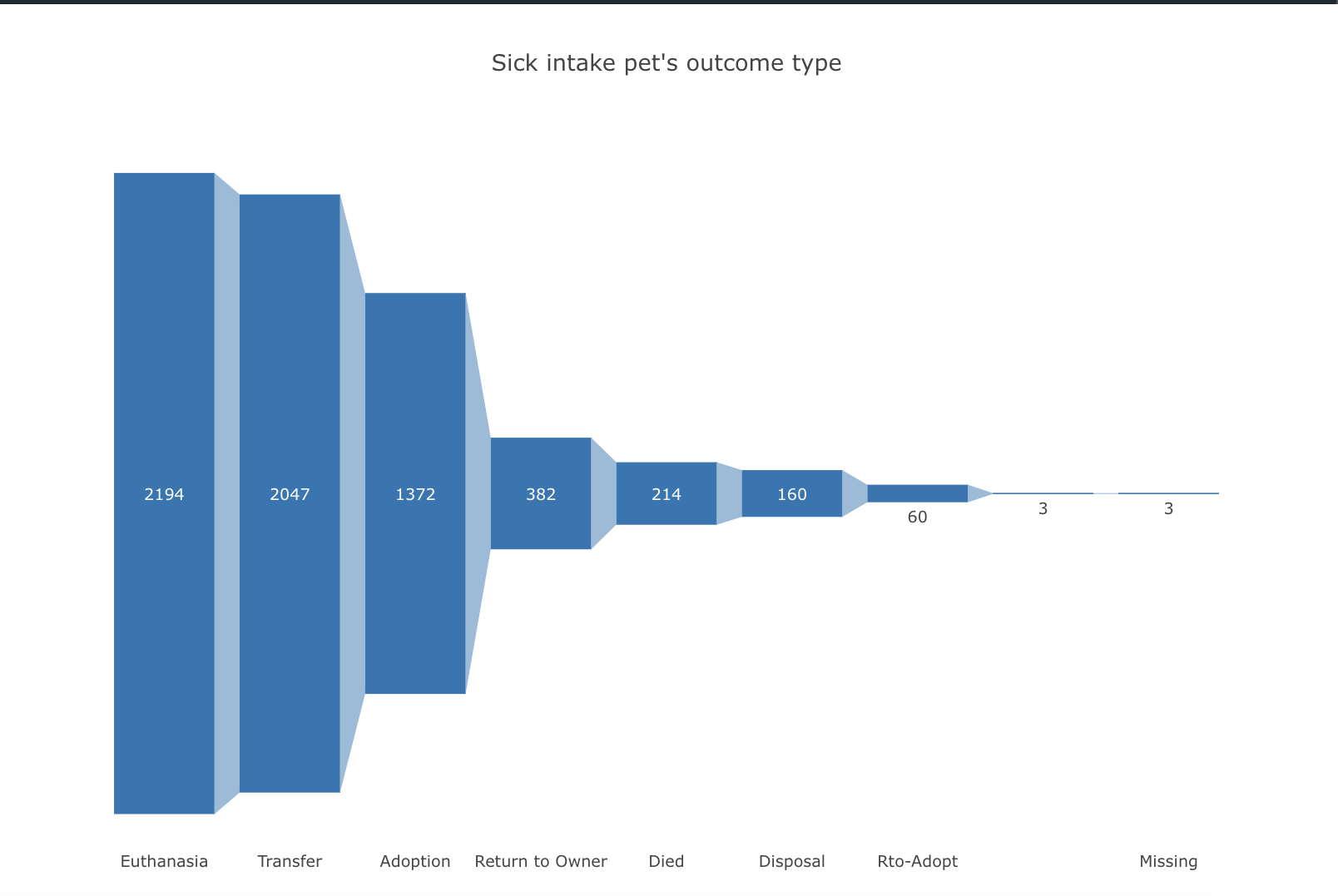
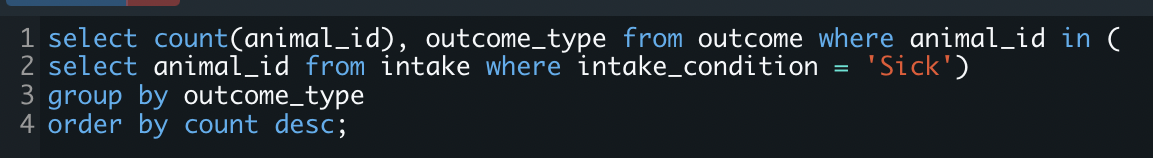
**Fig. 12.**

1. Pet Name correlation with adoptability



**Fig. 13**

1. Adoption trends when animal condition is sick

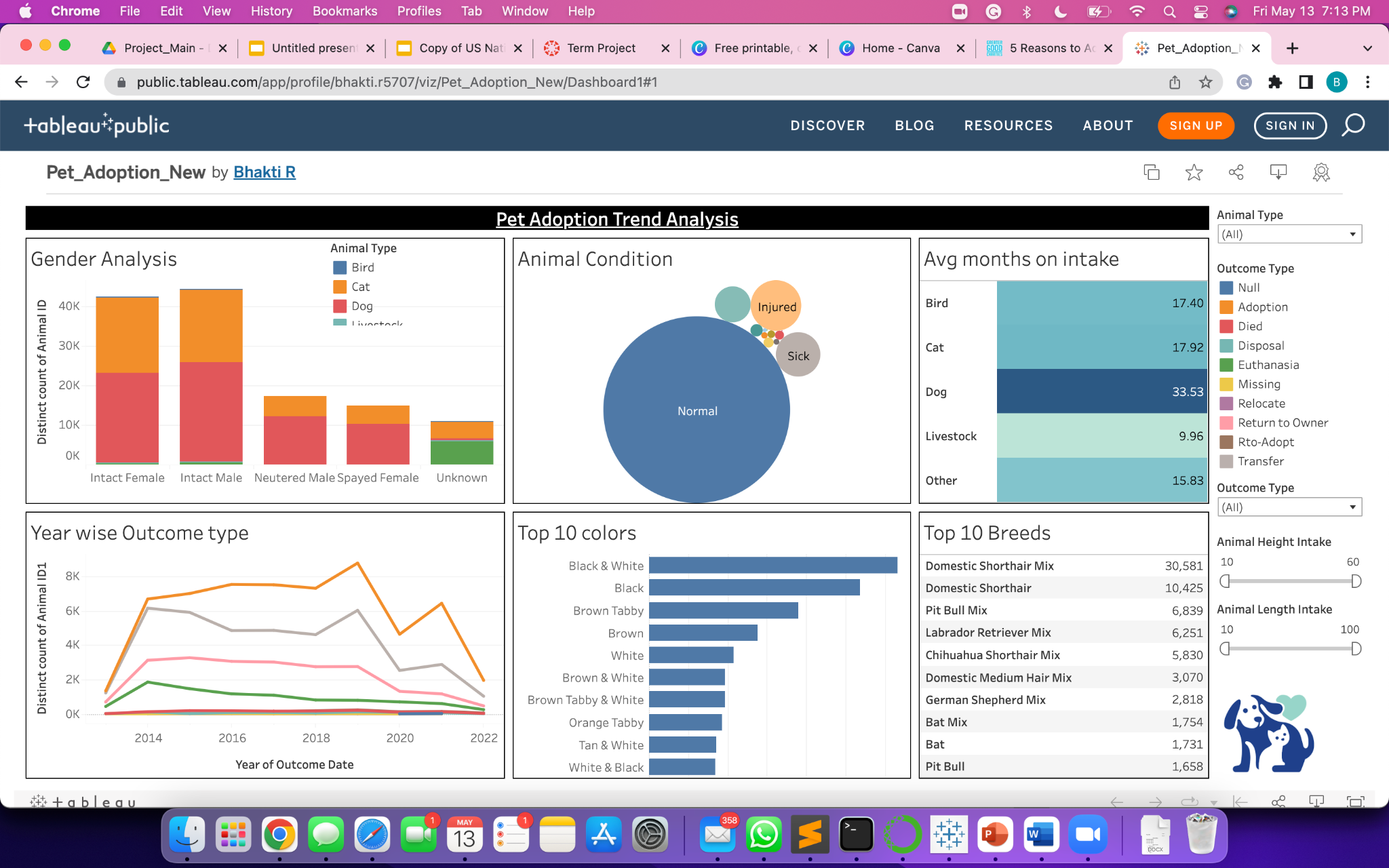


**Fig. 14.**

# Chapter 5 Visualization

* 1. **Data Visualization Using Tableau**

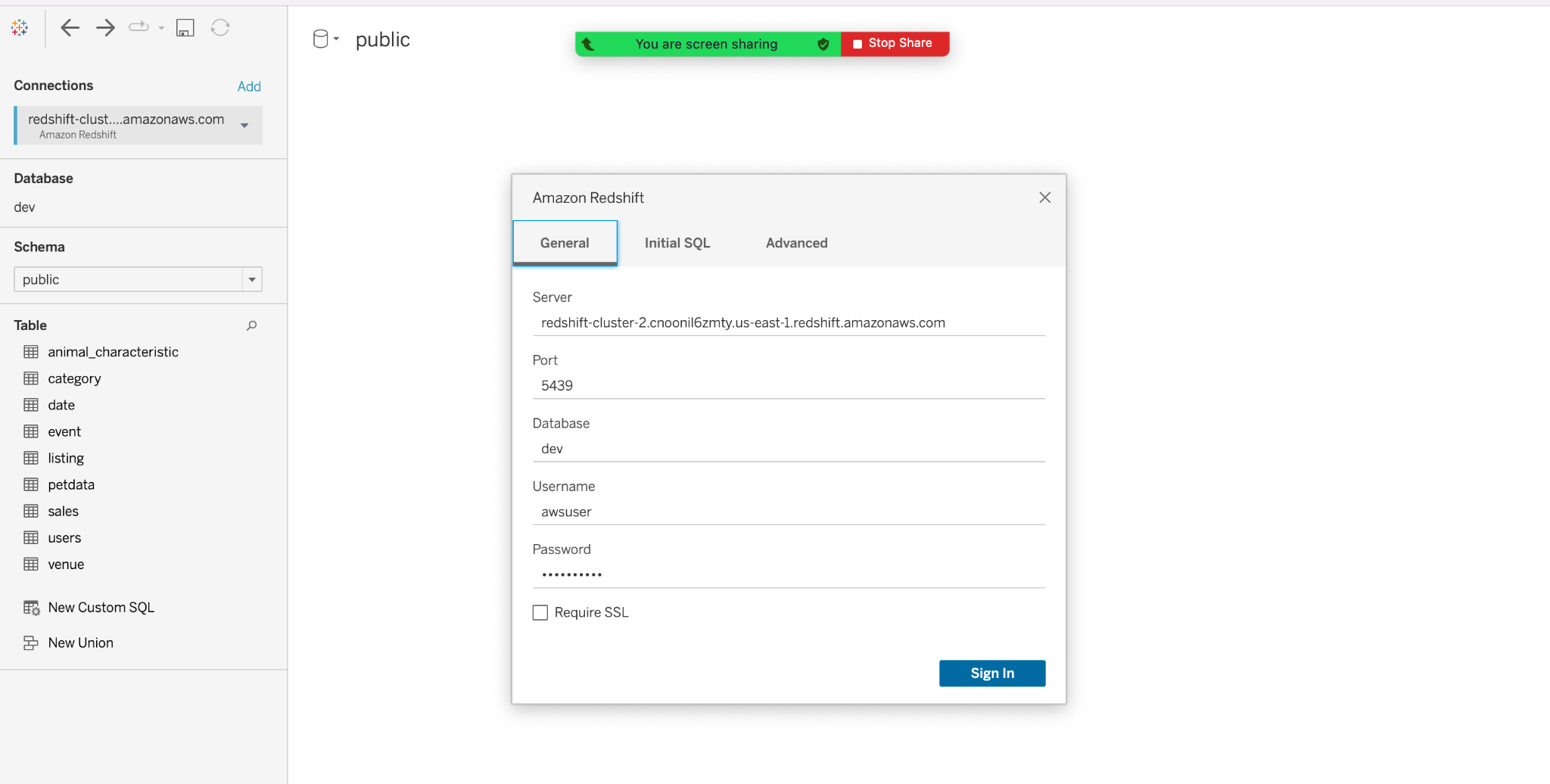
The analyzed data was then connected in tableau desktop through with the use of redshift drivers for visualization. Tableau Desktop is a data visualization and analytical tool which helps create interactive dashboards and help in storytelling through data. Using tableau, we could take our analysis on pet- adoption to the next level by drilling into the dataset further. The insights from the analysis is tableau have been explained in the subsequent sections. The visualization dashboard is also published using Tableau public on <https://public.tableau.com/app/profile/bhakti.r5707/viz/Pet_Adoption_New/Dashboard1#1>



**Fig. 15.**

**5.2 Tableau Desktop connecting with Redshift**

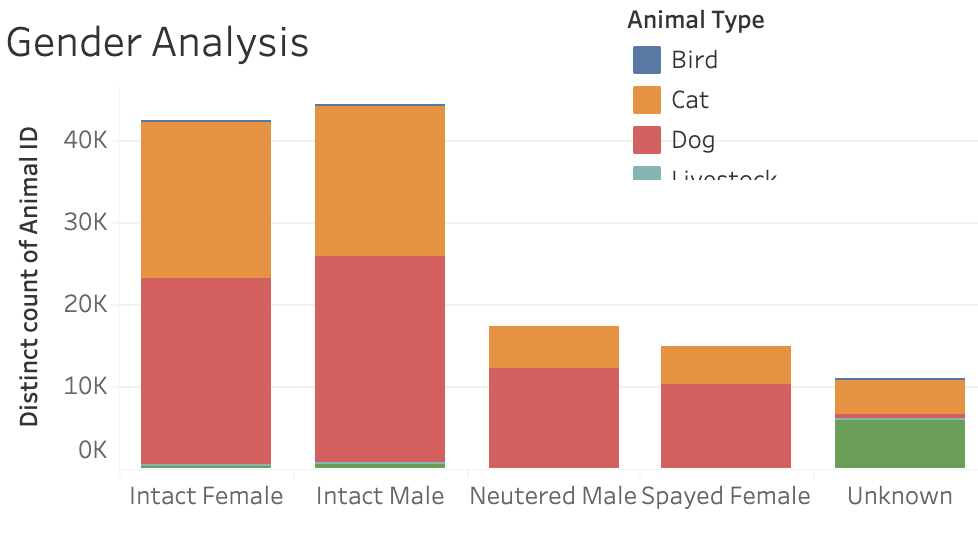
We were able to connect our dataset to AWS Redshift using Tableau connectors. This helped create the visualization pipeline and allow tableau directly to intake data from Redshift clusters. The connection is shown in Fig 11. as below Multiple datasets were inputted by setting the VPC subnet id as host and providing the user details



**Fig. 16.**

**5.3**  **Insight 1:**

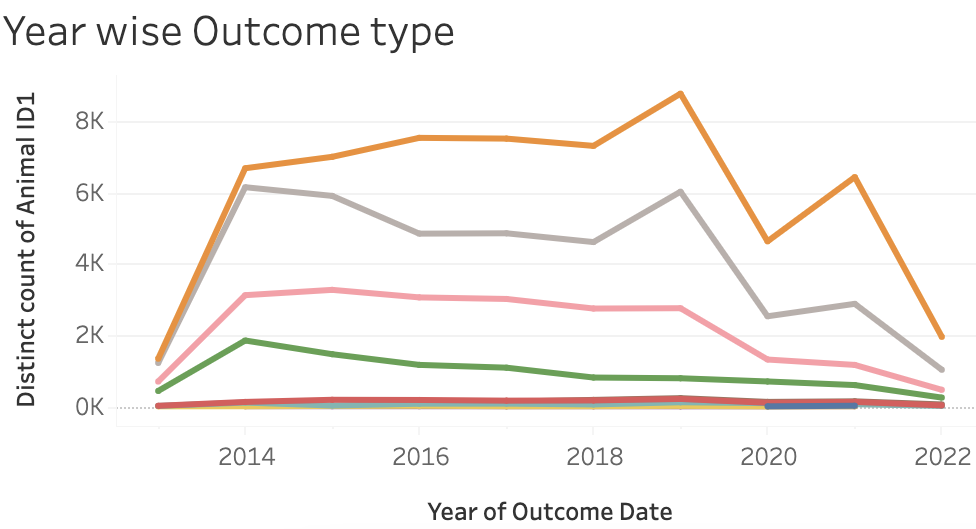
People prefer more intact males and females pets as opposed to neutered or spayed pets. The insight and business case from this observation could be to ask animal shelters to not neuter the pets while at the shelter



**Fig. 17.**

**5.4 Insight 2:**

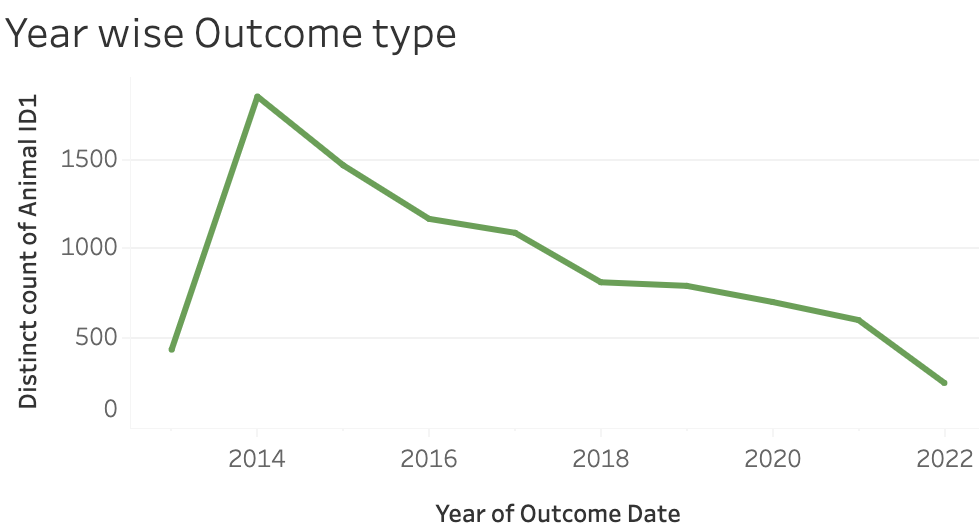
The adoption trend (indicated in orange) was at it’s peak in 2019 but showed a massive decline during the pandemic in 2020. People were not able to adopt pets because of the lockdown. However, when the restrictions were lifted and things started opening up, the adoption rates shot back up in 2021. We see a gradual decrease again in 2022. This can be attributed to the reason that work places and schools have resumed to back to office model again as opposed to work from home and people are busy again



**Fig. 18.**

**5.5 Insight 3:**

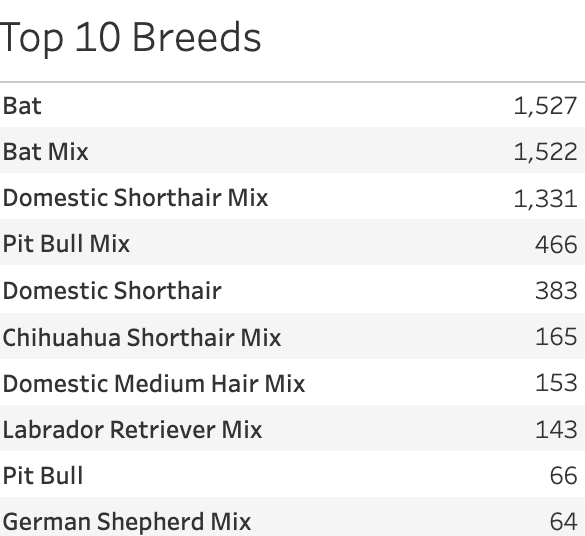
The below Fig 14. shows trends in Euthanasia from the years 2013 to 2022. We can see that Euthanasia peaked in the year 2014. Upon research, we found out that this was due to the outbreak of HPAI in the US in 2014 due to which a lot of animals were euthanized. Between December 2014 and June 2015, the United States experienced its worst animal health crisis, with over 200 instances of highly pathogenic avian influenza (HPAI) discovered in commercial and domestic poultry, as well as wild birds, across the country.



**Fig. 19.**

**5.6 Insight 4:**

More Bats were euthanized in the year 2020. Upon research we found that people were able to trace roots of the Covid-19 outbreak to Bats which led to Euthanization of the whole species to restrict the spread of the virus



**Fig. 20.**

**5.7 Website URL:**

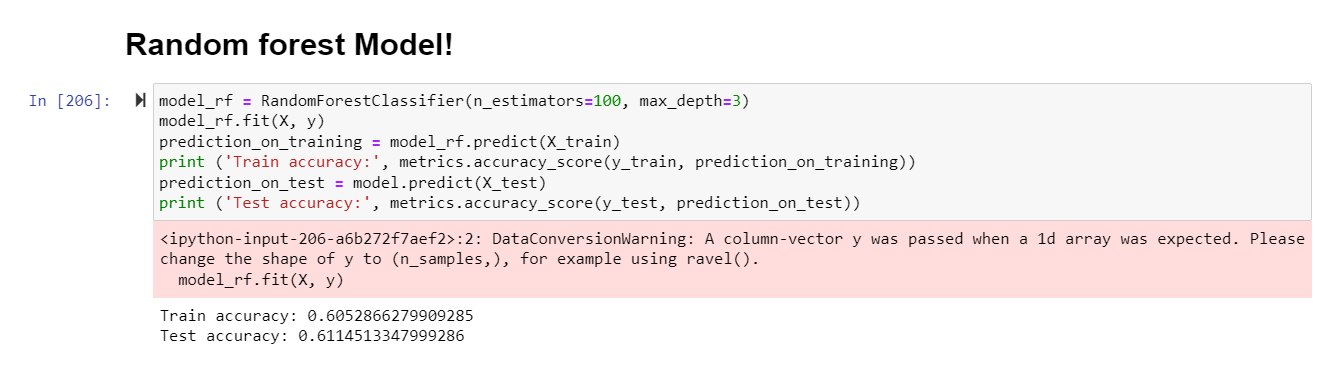
<https://bhaktiraichura1995.wixsite.com/pet-adoption-analysi>

# Chapter 6 Prediction Modeling

We have implemented 4 machine learning models to understand which models are best suited for our goal. Below are the 2 best models which have given the highest accuracy for the pet adoption rate.

**6.1 Random Forest Model**

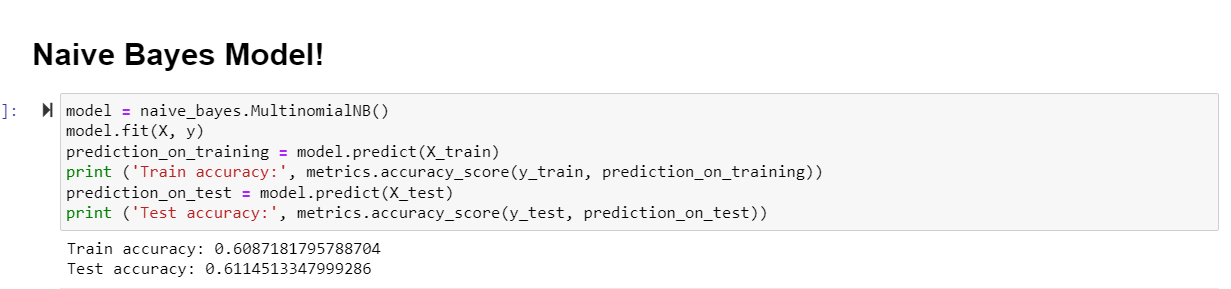
We included an ensemble method and we chose random forest because classification trees are easy to explain and we want our findings to be actionable for Austin Animal Shelter.



**Fig. 21**

**6.2 Naive Bayes Model**

Naive Bayes: We tried to implement a Naive Bayes model because there are so many independent features in our dataset from one to another for example Gender and Animal Intake Type or Intake Condition. And from the given input we got prediction accuracy 61.14% for our pet predict adoption.



**Fig. 22**

**6.3 Model’s Comparison**

Using these results, the random forest and Naive Bayes models are the predictive model that works best for our predictive task of which animal will get adopted or not.

| Models | Training | Test |
| --- | --- | --- |
| Decision Tree | 60.4% | 60.6% |
| KNN (Neighbors: 50) | 60.6% | 60.6% |
| Naive Bayes | 60.8% | 61.14% |
| Random Forest | 60.5% | 61.14% |

# Chapter 7 Collaboration

**7.1.** **GitHub**

All source code and finished documentation will be uploaded to GitHub repository.

Repository URL: [Dataanalystwhotellsstories/Furever-Home-Pet-Adoption-Analysis (github.com)](https://github.com/Dataanalystwhotellsstories/Furever-Home-Pet-Adoption-Analysis)

**7.2.** **Team members and Id’s**

| **Name** | **Student ID** | **GitHub ID** |
| --- | --- | --- |
| Richa Sharma | 014638494 | ([Dataanalystwhotellsstories (Richa Sharma) (github.com)](https://github.com/Dataanalystwhotellsstories)) |
| Shreya Srirama | 016029845 | (<https://github.com/shreyasriram> ) |
| Khushee Thakkar | 015271529 | (<https://github.com/khushee5696>) |
| Bhakti Raichura | 016020628 | (<https://github.com/bhaktiraichura>) |

# Chapter 8 Conclusion and Future Work

**8.1 Conclusion**

Through the course of this project, we were able to learn and explore several AWS services like AWS S3, GLUE, Redshift etc. We were also able to connect these services to other external tools like Python and Tableau using connectors and required drivers.We dealt with categorical features in Python while using the prediction model by assigning numerical values to the categorical features. Thus, we conclude that AWS services can be used successfully to create an aggregation pipeline to analyze a big data problem such as ours and help provide meaningful insights.

**8.2** **Future work**

Our dataset focuses only on data from an animal shelter in Austin, Texas. To expand the scope of our project, we can input data form other animal shelters as well. The efficiency of the prediction model will be increased and enhanced as well if more data is used. User and customer data will be a key input in the future scope of our project. Enhancing the website UI is also one of the major next steps for our project

# References

* <https://bhaktiraichura1995.wixsite.com/pet-adoption-analysi>
* <https://github.com/Dataanalystwhotellsstories/Data-228-Project>
* <https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Outcomes/9t4d-g238>
* <https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Intakes/wter-evkm>
* <https://public.tableau.com/app/profile/bhakti.r5707/viz/Pet_Adoption_New/Dashboard1#1>
* <https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics>

# Appendix

| Criteria | Remarks |
| --- | --- |
| Used Big Data concepts from this Course DATA 228 | Yes |
| This criterion is linked to a Learning Outcome Significance to the real world | Yes |
| Code Walkthrough | Yes  [Dataanalystwhotellsstories/Furever-Home-Pet-Adoption-Analysis (github.com)](https://github.com/Dataanalystwhotellsstories/Furever-Home-Pet-Adoption-Analysis) |
| Version Control Use of Git / GitHub or equivalent; must be publicly accessible | Yes  [Dataanalystwhotellsstories/Furever-Home-Pet-Adoption-Analysis (github.com)](https://github.com/Dataanalystwhotellsstories/Furever-Home-Pet-Adoption-Analysis) |
| This criterion is linked to a Learning Outcome Discussion / Q&A | Yes |
| Lessons learned Included in the report and presentation? | Yes |
| This criterion is linked to a Learning Outcome Innovation | Yes  Used AWS features |
| This criterion is linked to a Learning Outcome Teamwork | Yes |
| This criterion is linked to a Learning Outcome Technical difficulty | The data consisted of many outliers and null values which resulted in inefficient results. To solve the problem and improve the efficiency of the analytical queries we used Python, as an ETL tool to clean and transform our dataset. Plotting the graphical representation using Matplotlib and Seaborn python libraries. |
| AWS Tool | Redshift |
| Used ETL tool | Python |
| Visualization Tool | Tableau |