IE 509 - Computer Programming Lab
Course Project Report

Tanmay Nath - 23N0457

November 26, 2023

Data Description

id: An identifier for each record, likely a unique identifier for each animal entry.

intakedate: Date when the animal was taken into the shelter.

intakereason: Reason or circumstances for the animal's intake.

istransfer: Boolean indicating whether the animal was transferred.

sheltercode: Code associated with the shelter.

identichip number: Identification number, possibly related to a microchip.

animal name: Name of the animal.

breed name: Breed of the animal.

base colour: Primary color of the animal's fur or skin.

species name: Species of the animal.

animalage: Age of the animal.

sexname: Gender of the animal.

movement date: Date of any movement of the animal.

movement type: Type of movement (e.g., adoption, transfer, trial adoption).

returndate: Date when the animal was returned, if applicable.

returned reason: Reason for the animal's return.

deceased date: Date of death of the animal.

deceased reason: Reason for the animal's death.

Data Sources: Dataset sourced from **Kaggle**, featuring data from an adoption center in the U.S.

- **The number of data points in the data are 10,000+, and the number of features is 23.**

Tools Used

The successful execution of this project was made possible by employing a combination of powerful tools for data manipulation, analysis, and visualization. Each tool played a crucial role in different aspects of the project, contributing to a comprehensive and insightful analysis.

- **Pandas:** Pandas is a versatile and widely-used Python library for data manipulation and analysis. It provided essential functionalities for handling and processing the dataset, including filling missing values, transforming data, and facilitating exploratory data analysis.
- **NumPy:** NumPy, a fundamental library for scientific computing in Python, was utilized for various numerical operations and computations. It offered efficient data structures for working with large arrays and matrices, enhancing the computational capabilities of the project.
- **Matplotlib:** Matplotlib, a comprehensive library for creating static visualizations in Python, played a crucial role in generating various plots and graphs. It facilitated the visualization of trends, distributions, and correlations, aiding in the interpretation of complex relationships within the dataset.
- **Seaborn:** Seaborn, built on top of Matplotlib, provided a high-level interface for statistical data visualization. Its elegant and informative visualizations enhanced the clarity of data patterns, making it easier to communicate key findings to stakeholders.
- **Scikit-learn (sklearn):** Scikit-learn, a robust machine learning library, was used for transforming categorical data into numerical format. This was essential for incorporating categorical features into the analysis and modeling processes.

Objective

The primary objective of this project is to gain comprehensive insights into the dynamics of stray animals. This involves investigating the various reasons for animal abandonment and categorizing adopted animals based on critical factors such as age, color, breed, and adoption status. Additionally, we aim to enhance the overall quality of the dataset by addressing missing values, transforming date features for chronological analysis, and removing outliers.

The ultimate goal is to provide actionable insights that can inform decision-making in the context of stray animal welfare and adoption initiatives.

Project Breakdown

1. Handling Missing Data

Used Pandas to fill in the missing data in different features of the dataset using various techniques as required.

2. Plotting

Started plotting graphs for a better understanding of the features and their dependence.

3. Interpreting and Narrowing Down the Data

After plotting different graphs, we decided that cats and dogs have the highest adoption rate. Therefore, we shifted our study to these specific animal types.

4. Relation Realization

Pulled out relationships among different features of the data and encoded the data into numerical data to find out the relation using a correlation matrix.

5. Getting to the Conclusion

Worked towards understanding what is happening in the adoption home based on the insights gained from the previous steps.

Project Report - Data Cleaning and Transformation

1.Handling Missing Data

The dataset contains missing values in the following columns: ’intakereason’, ’breedname’, ’identichipnumber’, ’returndate’, ’returnedreason’, ’deceaseddate’. To address this:

- Since only 2 rows have missing values in the ’intakereason’, these rows were removed using the `dropna` method:

  `df.dropna(subset=["istrial"], inplace=True)`

- Null values in the subjective data column ’istrial’ were filled with 0.
- The ’breedname’ column null values were replaced with ’missing’.
- Null values in ’identichipnumber’ were filled with 0.

Change Data Types

To ensure proper analysis, the date columns were converted to the datetime format:

- `df[’intakedate’] = pd.to`$_{d}atetime(df.intakedate)$`df[’movementdate’] = pd.to`$_{d}atetime(df.movementdate)$`df[’re`

Create Date Features

New date features were created for ’intakedate’ and ’movementdate’:

- `df[’yeartake’] = df.intakedate.dt.year`
- `df[’monthtake’] = df.intakedate.dt.month`
- `df[’daytake’] = df.intakedate.dt.day`
- `df[’yearmove’] = df.movementdate.dt.year`
- `df[’monthmove’] = df.movementdate.dt.month`
- `df[’daymove’] = df.movementdate.dt.day`

Remove Outliers

No specific code snippet was provided for removing outliers. If you have additional code for removing outliers, please include it here.

Final Data Information

To review the changes and ensure data integrity:

- `df.info()`

Plot 1: Number of Intakes Over the Years

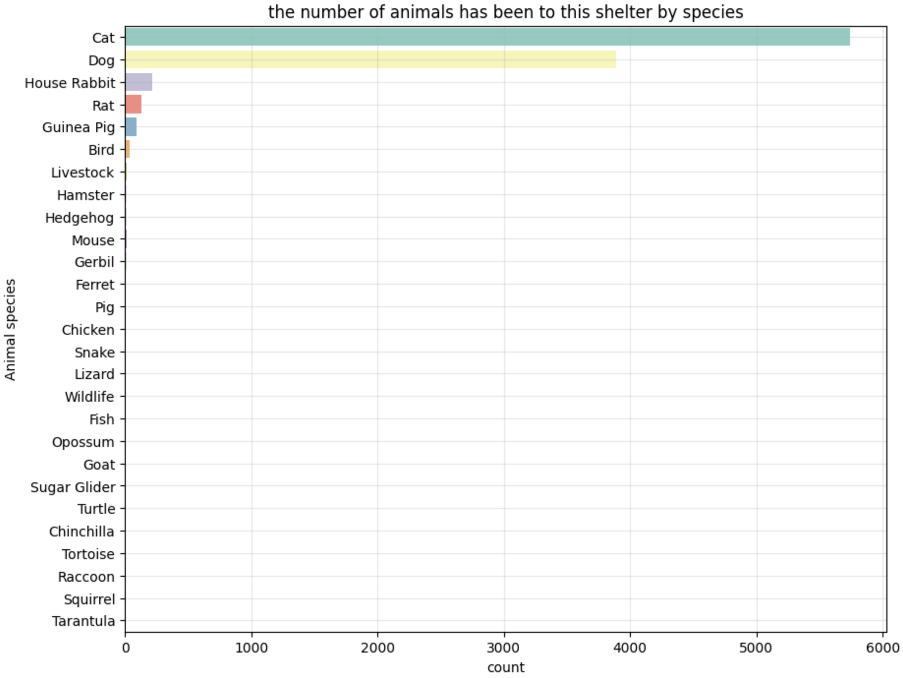

Figure 1: Number of Intakes Over the Years

Outcome: The number of intakes has increased in 2019, 2018, and 2017.
Action Taken: Data is filtered to consider only the years 2017, 2018, and 2019.

Plot 2: Number of Animals by Species

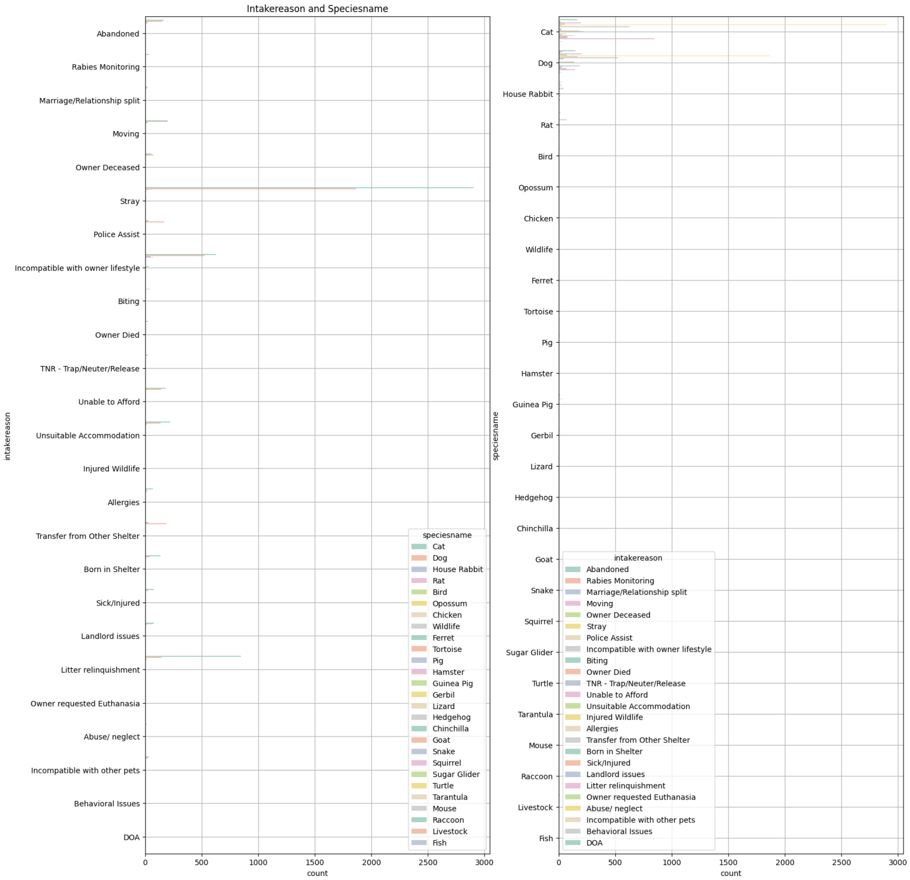

Figure 2: Number of Animals by Species

Outcome: Visualizes the number of animals by species.
Action Taken: The data is plotted to show the distribution of animal species.

Plot 3: Intake Reasons and Species

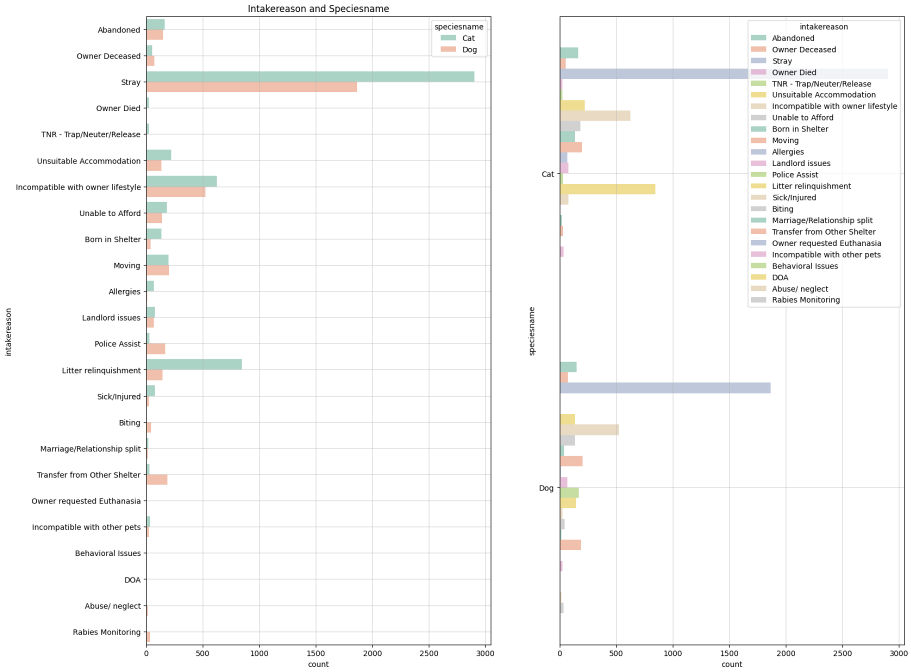

Figure 3: Intake Reasons and Species

Outcome: Shows the relationship between intake reasons and animal species.
Action Taken: Two plots are created, one showing the distribution of intake reasons for each species, and the other showing the distribution of species for each intake reason.

Plot 4: Intake Reasons and Species (Focused on Cats and Dogs)

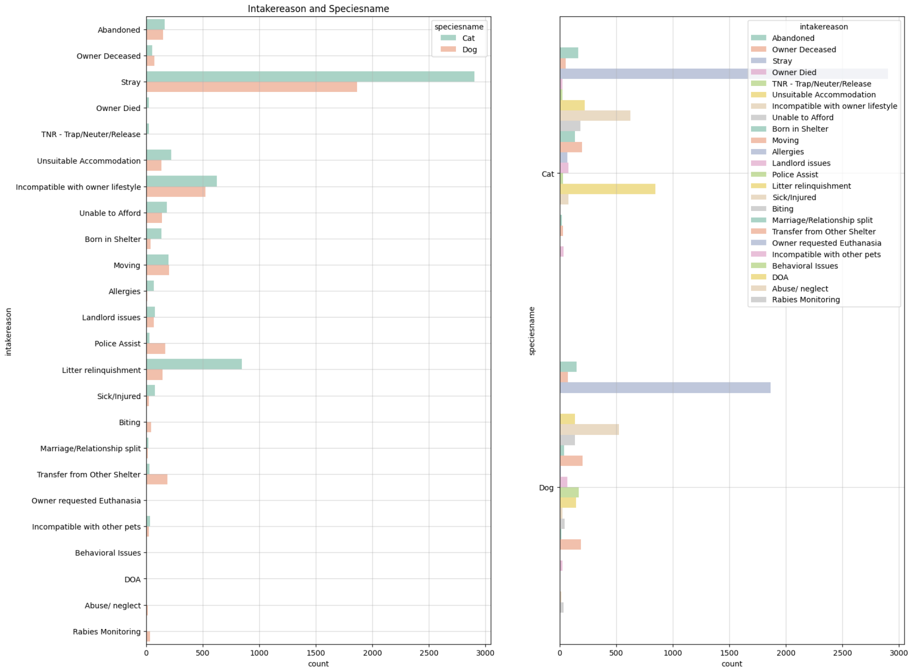

Figure 4: Intake Reasons and Species (Focused on Cats and Dogs)

Outcome: Shows the relationship between intake reasons and animal species, specifically focusing on cats and dogs.
Action Taken: Two plots are created, one showing the distribution of intake reasons for each species (cats and dogs), and the other showing the distribution of species for each intake reason.

Plot 5: Movement Types and Sexes

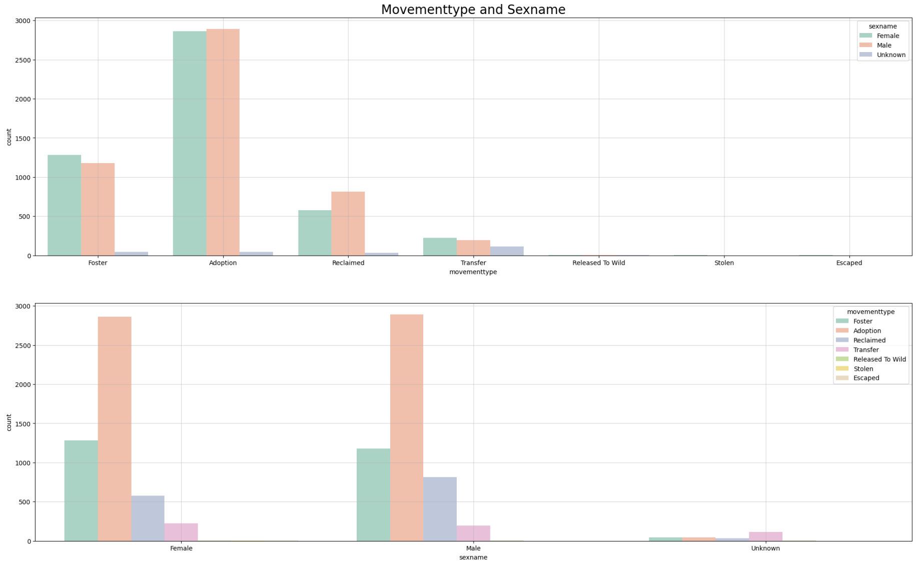

Figure 5: Movement Types and Sexes

Outcome: Displays the distribution of movement types based on the sexes of the animals.
Action Taken: Two plots are created, one showing the count of movement types for each sex, and the other showing the count of sexes for each movement type.

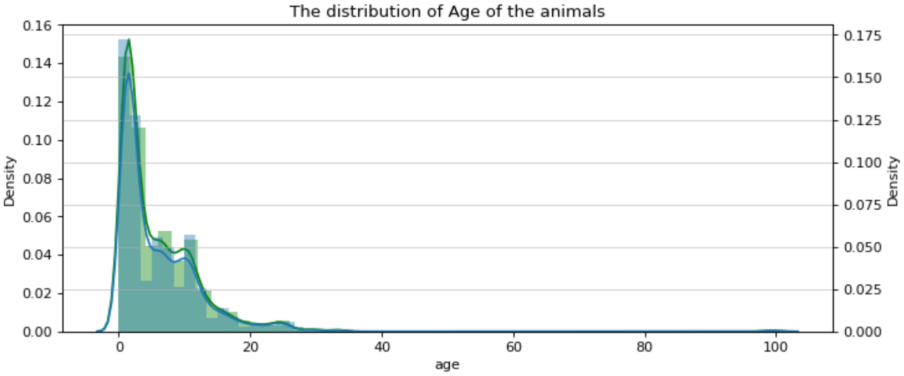

Figure 6: Distribution of Animal Ages

Outcome: Depicts the distribution of ages among the animals.
Action Taken: A distribution plot is created to visualize the age distribution.

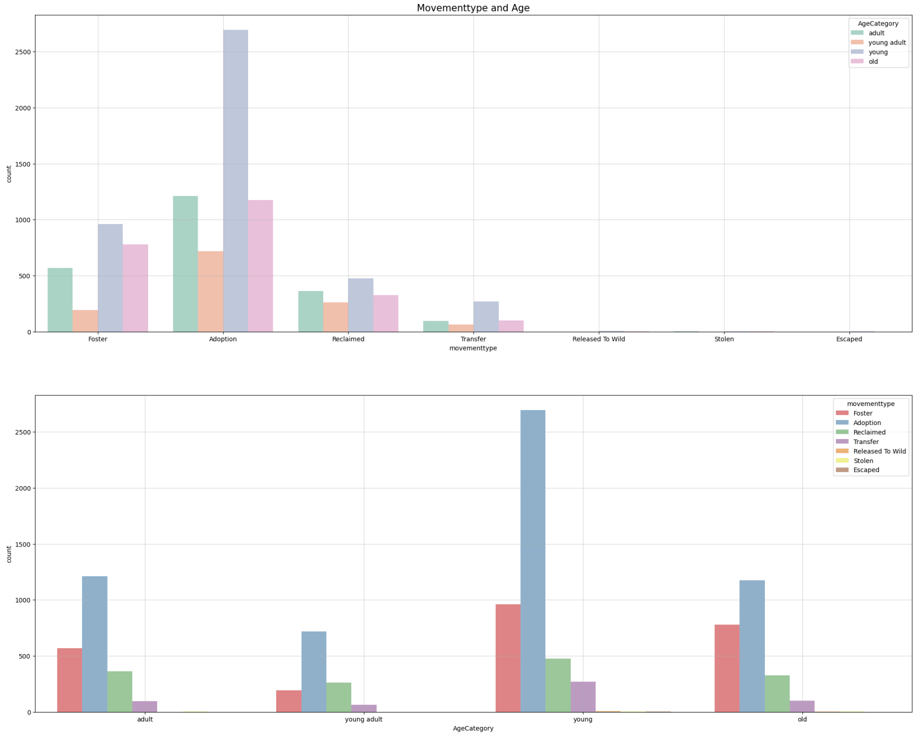

Figure 7: Movement Types and Age Categories

Outcome: Illustrates the relationship between movement types and age categories.
Action Taken: Two plots are created, one showing the count of movement types for each age category, and the other showing the count of age categories for each movement type.

Plot 8: Time Series - Monthly Intake Trends

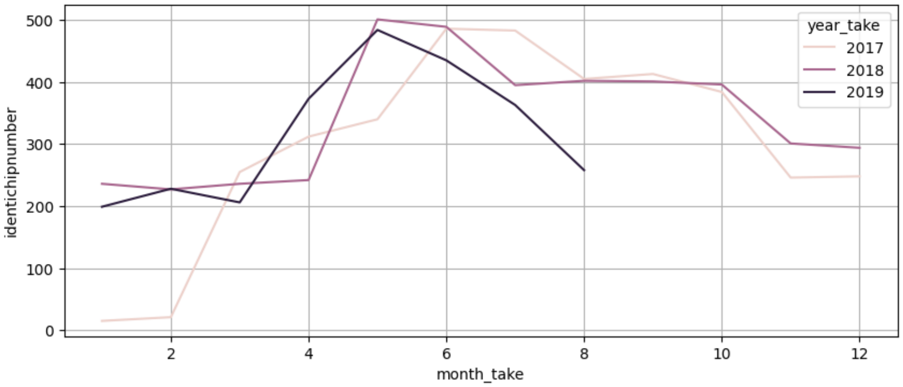

Figure 8: Time Series - Monthly Intake Trends

Outcome: Depicts monthly trends in animal intakes over the years.
Action Taken: A line plot is created to show the variation in the number of intakes each month.

Plot 9: Time Series - Movement Types Over Time

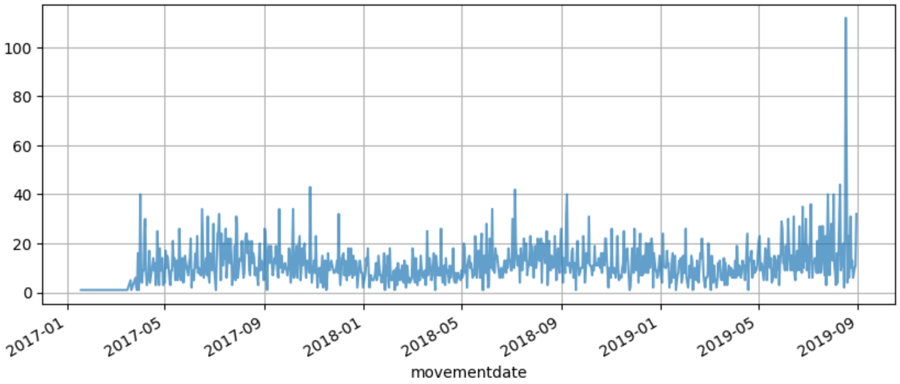

Figure 9: Time Series - Movement Types Over Time

Outcome: Demonstrates the changes in the count of different movement types over time.
Action Taken: Multiple line plots are created, each representing a different movement type, to visualize the trends over time.

Plot 10: Time Series - Movement Types by Month

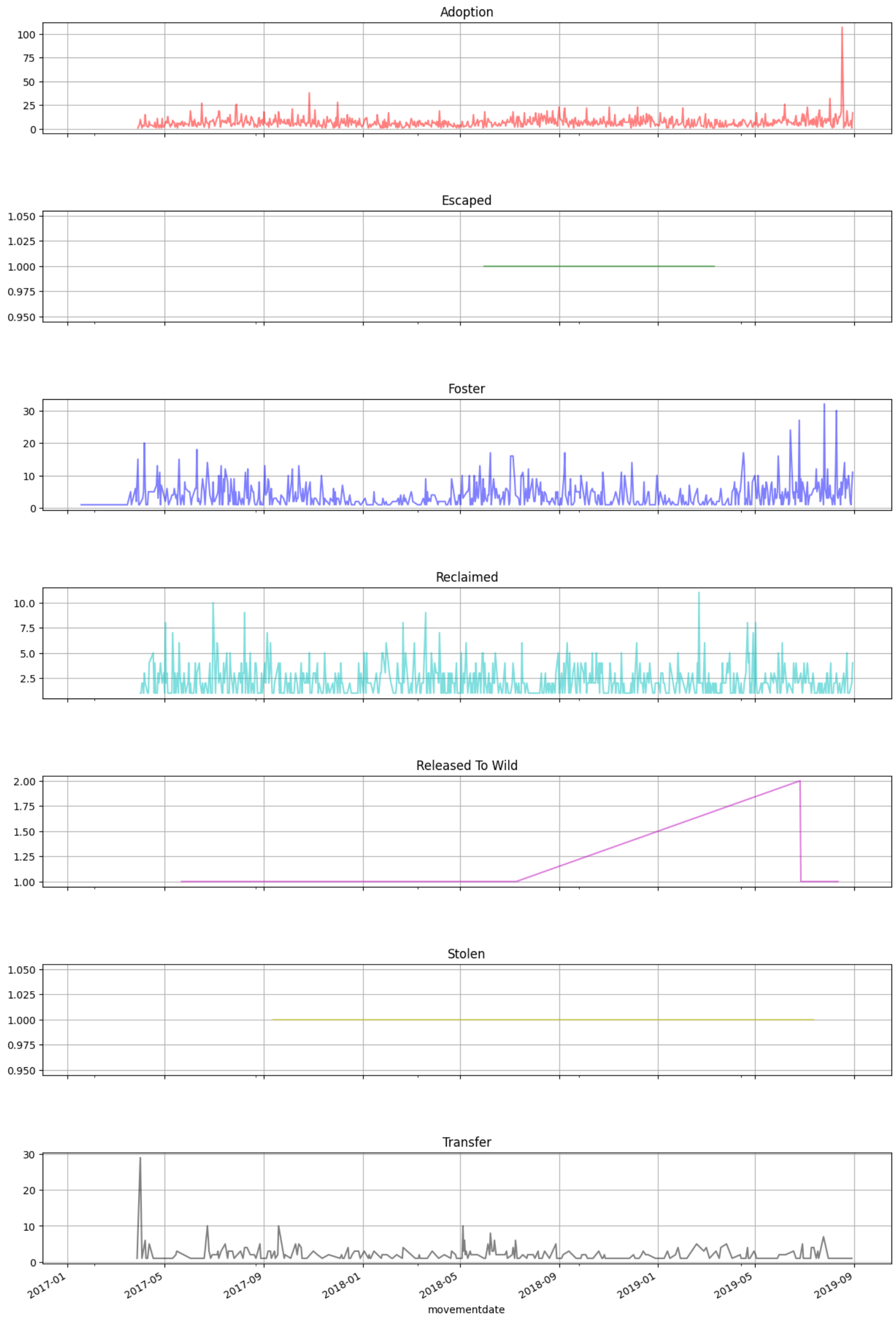

Figure 10: Time Series - Movement Types by Month

Outcome: Displays the monthly variation in the count of different movement types.
Action Taken: Multiple line plots are created, each representing a different movement type, to observe the patterns by month.

Data Preprocessing and Feature Encoding

The dataset ‘df$_heat‘ispreprocessedtoremoveunnecessarycolumns, suchasdates, identifiers, andothernon-essentialinformation.Theresultingdatasetisnamed‘df_heat‘$.

- Columns Removed:
  - intakedate, movementdate: Date columns
  - identichipnumber: Identification number
  - animalname, animalage, returndate, returnedreason, deceaseddate, deceasedreason: Information related to specific instances
  - year take, month take, day take, year move, month move, day move: Date features
  - diedoffshelter, isdoa: Columns related to animal status

To facilitate further analysis, categorical features are encoded using Label Encoding. Each unique category in the categorical features is assigned a numerical label.

- Species Name (speciesname): Encoded using Label Encoder (‘le sp‘)
- Intake Reason (intakereason): Encoded using Label Encoder (‘le take‘)
- Breed Name (breedname): Encoded using Label Encoder (‘le breed‘)
- Base Color (basecolour): Encoded using Label Encoder (‘le color‘)
- Gender (sexname): Encoded using Label Encoder (‘le sex‘)
- Location (location): Encoded using Label Encoder (‘le loc‘)
- Movement Type (movementtype): Encoded using Label Encoder (‘le move‘)
- Age Category (AgeCategory): Encoded using Label Encoder (‘le age‘)

These encodings transform categorical data into numerical

Correlation Matrix Heatmap

The correlation matrix heatmap visually represents the relationships between numerical features and the target variable `movementtype`.

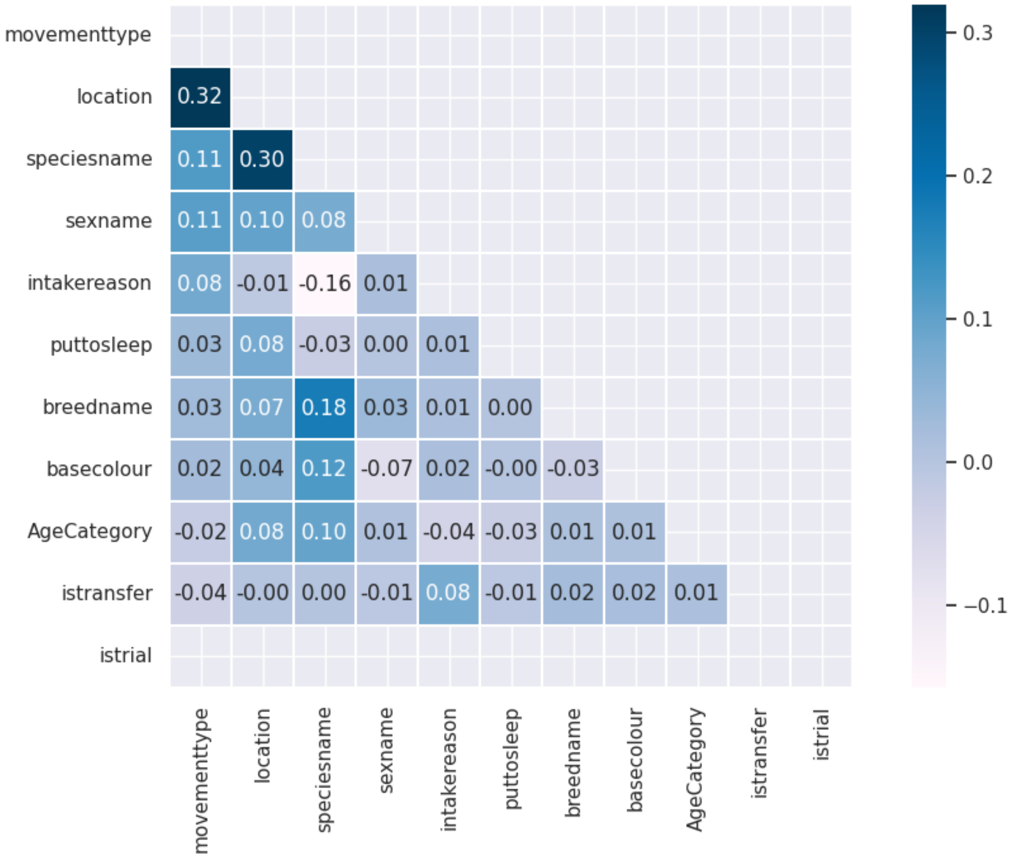

Figure 11: Correlation Matrix Heatmap

Observations

The heatmap displays correlations between selected numerical features and the `movementtype`. Darker shades indicate stronger correlations, while lighter shades suggest weaker or no correlations.

- **Positive Correlations:** Features with positive correlations are represented by darker shades. This implies that as these features increase, the likelihood of a specific movement type also increases.
- **Negative Correlations:** Conversely, features with negative correlations are represented by darker shades but in the opposite direction. As these features increase, the likelihood of the movement type decreases.
- **Strength of Correlations:** The strength of correlations can be assessed by the intensity of color. Darker colors indicate higher correlations, while lighter colors suggest weaker associations.

- **Feature Importance:** Features contributing significantly to predicting the movement type are identified based on the strength and direction of their correlations.

The correlation matrix heatmap provides valuable insights into the relationships between numerical features and the `movementtype`, aiding in understanding the factors influencing animal adoption patterns.

Conclusion

The analysis of the dataset has provided valuable insights into the dynamics of animal adoption at the shelter. One notable trend that emerged from the data is the influence of the color of fur or skin on the likelihood of abandonment or adoption, particularly among cats and dogs.

Observations

The data revealed a clear trend where cats and dogs with more or less black fur or skin are more prone to abandonment. Conversely, animals with white fur are more likely to be adopted from foster centers. This observation suggests a potential bias or preference among adopters based on the coloration of the animals.

Recommendations

Based on these findings, a practical recommendation for individuals working at the animal center is to be more considerate towards cats and dogs with specific color categories. Understanding and addressing the potential biases related to fur color could contribute to a more informed and empathetic approach to animal welfare.

Implications for Adoption Centers

The insights gained from this analysis offer a better understanding of the general functioning of an adoption center. This information can serve as a valuable resource for decision-makers, allowing them to tailor strategies that improve the adoption rates of animals, especially those with characteristics that may be overlooked.

In conclusion, this analysis not only sheds light on specific trends in animal adoption but also emphasizes the importance of continuous monitoring and adaptation in the operations of adoption centers. By leveraging data-driven insights, we can work towards creating more inclusive and effective adoption initiatives that prioritize the well-being of all animals in the shelter.