Research Question:

Which factors influence the number of days an animal spends in the shelter before their final outcome is decided?

The variables listed on the datasets:

- Animal_type – The type of animal admitted to the shelter
- Month – Month the animal was admitted, recorded numerically with January=1
- Year. – Year the animal was admitted to the shelter.
- Intake_type – Reason for the animal being admitted to the shelter
- Outcome_type – Final outcome for the admitted animal
- Chip_Status – Did the animal have a microchip with owner information?
- Time_at_Shelter – Days spent at the shelter between being admitted and the final outcome.

The methodology to conduct the analysis:

1. GLM Analysis for Length of Stay
    a. Data Preparation
        - Loading the data.
        - Define the variables (the response variable: time_at_shelter, the explanatory variables: animal_type + intake_type + chip_status + month + year).
        - Convert the data to the appropriate types.
    b. Explanatory Data Analysis
        - The data distribution especially for the count data.
        - The relationship among the variables.
        - The graphs or visualization of the data including the variables, and relationship. Visualization should also include checking for potential outliers or influential observations.
        - Interpretation of each list above.
    c. Model Fitting
        - Starting with the Poisson Regression (based on the type of the response variable).
        - Model diagnostic by checking the assumptions.
          Check the overdispersion and other diagnostics. If the overdispersion is detected then the Negative Binomial Model is considered as an alternative.
        - Model validation to assess the quality of the model.
          Both are either the residuals diagnostics or cross-validation.
        - Graph for the appropriate model.
        - The interpretation of each list above.
    d. Conclusions
        - Analysis of which factors were found to be significant.
        - Interpret the direction and magnitude of their effects on the time at the shelter. x
        - Interpret the model coefficients and the possible conclusions or recommendations based on the prior analysis.

2. GLM Analysis for Outcome Type
   The objective: to predict the probability of each possible outcome type category based on possible explanatory variables.
   The model used: multinomial logistic regression (a type of GLM suited for multiclass categorical response variables). The response variable is treated as a categorical variable with appropriate reference levels set (for based variable).
   a. Data Preparation
      - Define the variables (the response variable: outcome_type, the explanatory variables: time_at_shelter + animal_type + intake_type + chip_status + month + year). The other variables are included as well.
   b. Model Fitting
      The same process as above
   c. Model diagnostic and validation
      Emphasize cross-validation and predictive accuracy assessment.
   d. Reporting and Conclusions
      The same process as above.
      Compile your findings, including the model's predictive performance, the significance of predictors, and their effect sizes.
   e. Considerations and Limitations
      Emphasize ensuring the temporal precedence of the predictors over the outcome particularly with time_at_shelter, to avoid data leakage.

Note about Poisson Regression Model:

Overdispersion is a phenomenon that occurs in modeling count data where the variance of the dependent variable is greater than its mean. This is a violation of one of the assumptions of the Poisson distribution, which assumes that the mean and variance are equal. Overdispersion is common in real-world data and can occur for several reasons:

1. **Unobserved Heterogeneity:** There might be underlying variability in the data that is not captured by the model. For example, there might be subgroups within the population that have different rates of the event in question, which isn't accounted for in the model.
2. **Excess Zeros:** Sometimes, the data can have more zeros than what the Poisson model can accommodate, a condition often referred to as "zero inflation."
3. **Correlated Data:** If the data points are not independent of one another (e.g., multiple observations from the same animal or cluster of animals), this can lead to overdispersion.
4. **Wrong Distribution Assumption:** The true underlying process generating the data may not follow a Poisson distribution, thus leading to a discrepancy in the variance.

Overdispersion can lead to underestimated standard errors and thus to overstated significance levels for the coefficients in a model. If overdispersion is present and not accounted for, it can result in an increased type I error rate (rejecting a true null hypothesis). In the context of GLM, it is important to check for overdispersion after fitting a Poisson model. If overdispersion is present, alternative models such as the negative binomial model or zero-inflated models can be considered. These models have additional parameters that allow for the extra variability and are thus better suited for overdispersed data.

To check for overdispersion, you can compare the residual deviance to the degrees of freedom in a Poisson model—if the residual deviance is much larger than the degrees of freedom, that's a sign of overdispersion. There are also formal statistical tests, such as the dispersion test, that can be used to test for overdispersion.