Barton Springs and Native Salamanders Statistical Research Analysis

Eman Zaheer, Sanjana Ganesh, Karla Stambaugh, Maya Saleh, Miranda Smith, Shivani Purohit

University of Texas at Austin

Elements of Data Science SDS322E

Prof. Arya Farahi

December 6, 2023

Acknowledgement

Team Member	Percent Contribution
Eman Zaheer	100
Shivani Purohit	100
Miranda Smith	52
Karla Stambaugh	38
Maya Saleh	27
Sanjana Ganesh	100

INTRODUCTION

Background and motivation

Barton Springs is one of four distinct water springs nurtured by the Edwards Aquifer, unique to over 50 species of plants and animals found nowhere else on the planet. Among these various species resides the Barton Springs salamander - an endangered species found nowhere else - marking Barton Springs as a ‘conservation site’ under the federal government. Additionally, the salamander is a keystone species - indicating their disappearance could cause drastic changes for the habitat as well as other species that rely on them for survival (US Fish & Wildlife Services, 2005). The significance of this singular species and the need for their conservation opens an avenue of exploration into what makes the Spring so unique for their survival and how that can be maintained to ensure the successful survival of the species - present and future.

As aforementioned, our exploration focuses on the correlation between water quality parameters and the population sizes of Barton Spring salamanders. This relationship is interesting because of the overall status of the endangered salamanders, however, more specifically, salamanders take in oxygen through the pores of their skin from water absorption - regardless of quality. This could pose potential problems as polluted water could be detrimental to their health, survival, and federal status. The focus on this specific problem arises from the imperative need to comprehend and address the environmental challenges endangering the delicate balance of this ecosystem.

Objective

This report conducts an in-depth analysis of Barton Springs' salamander habitat by investigating the relationship between water quality parameters and the population sizes of Barton Springs Salamanders in Austin, Texas. This exploration hopes to understand the factors that affect salamander populations. Furthermore, we intended to understand if we can establish a link between these factors and salamander populations, to see if these factors can help predict salamander populations in the future.

DATA

Objective

This dataset contains information about salamander count, dissolved oxygen levels, flow rates, water depth, and multiple other factors connected to Barton Springs from July 1990 to August 2023.

Source of the Data

This dataset is from the City of Austin open data portal. From July 1990 to August 2023, data connected to various variables was collected from Barton Creek.

Variables in the Dataset

The dataset initially contained 18 columns including:

- Watershed: The location site where the data was collected.
- Sample date: The date the data was collected in the format of MM/DD/YYYY.
- Parameters: The rows of this column were either ‘dissolved oxygen’, ‘water depth’, ‘flow’, or a ‘salamander’ type and its length.
- Parameter type: The type of the parameter, either ‘salamanders’, ‘flow/rainfall’, ‘oxygen’, ‘physicals’, ‘spatial’, or ‘temporal’.
- Result: This corresponds with the parameter column. If the parameter was dissolved oxygen’ - the result column contained a number that represented the level of dissolved oxygen in mg/L. If the parameter was flow, then it contained the flow rate. If the parameter was salamander type and length, the result column contained the salamander count.

- Unit: unit of result, either mg/L (dissolved oxygen), cubic feet per second (flow), or count (salamander).
- Medium: The type of water medium, either benthic cover or ground water
- Filter: This contained the same information as the parameter column so using it would have been redundant.
- Sample ID: Many columns were missing values, and this information couldn’t be used to determine anything about the salamanders or Barton Springs habitat.
- Site name: This contained the same information as watershed so using it would have been redundant.
- Method: The method of data collection. The majority of this column was unknown.
- Project: The project that collected the data was stored here and wasn’t relevant to our exploration.
- LON_DD_WGS84: Longitude the data was collected from - since we only needed the location of the spring itself, the exact coordinates of data collection were not relevant to our exploration.
- LAT_DD_WGS84: Latitude the data was collected from - for the same reason as above.
- QC_FLAG
- QC_TYPE
- Data reference number
- Sample reference number
- TIME_NULL

Variables that were relevant to our analysis

1. Salamander count from the ‘param type’, ‘parameter’, ‘result’, and ‘unit’ columns
    a. Measured by count of salamanders.
2. Time from the ‘sample date’ column
    a. Measured by year.
3. Dissolved oxygen levels from ‘param type’, ‘parameter’, ‘result’, and ‘unit’ columns
    a. Measured in mg/L.
4. Flow rate from ‘param type’, ‘parameter’, ‘result’, and ‘unit’ columns
    a. Categorized by the kind of flow, ranging from 1 to 6:

        1: No flow

        2: Low flow

        3: Normal flow

        4: Flooding

        5: High Flow

        6: Dry
5. Water depth from ‘param type’, ‘parameter’, ‘result’, and ‘unit columns’
    a. Measured in inches

Data Cleaning and Processing

1. Modifying the data to suit our analysis: Since our exploration involved salamander comparison over a longer interval of time, from 1995 to 2023, we decided to change the date column, which previously included year, month, day, and time. We changed the column to only contain the year the data was collected.

2. Filtering the ‘parameter’ column: As mentioned above, the ‘parameter’ column contained the names of the variables we were interested in investigating, specifically ‘dissolved oxygen’, ‘flow rate’, ‘water depth’ and ‘flow type’. The rows in the ‘parameter’ column that have these variables were then associated with a specific value from the ‘result’ column, that contained the numerical information we needed. To analyze these variables, we used the filter function in the dplyr package in R to create filtered data sets that contained only the relevant ‘result’ we needed for a particular variable. Using the new filtered ‘result’ for each variable, we were able to do our analysis.
3. Figuring out how to read the salamander count data: Our data set contains information on multiple different types of salamanders, such as the ‘Barton Springs Salamander’ and ‘Austin Blind Salamander’. Moreover, our dataset also categorized the salamanders based on their size, with names like ‘Austin Blind Salamander (<1In) and ‘Barton Spring Salamander > 1 Inch’. In total, there were more than 40 different salamander classifications. Due to this, we felt it would be too complicated to analyze each unique salamander’s count. Instead, we decided that we would analyze the salamander population. We were able to do this by filtering the ‘Param Type’ column and selecting just ‘salamanders. This showed us all the salamander count data that was collected over the years. This included 8,418 data points.
4. Calculating average data: For each variable that we decided to analyze, we decided to work with ‘averages. We did this for a couple of reasons. Firstly, if we decided to analyze every single data point that existed in our data set, it would have been extremely challenging as we would be working with 30,000+ data points. Moreover, we wanted to understand how each variable has changed over time, and therefore classifying the data

based on some ‘time interval’ was crucial. So, we decided to calculate the average (yearly) dissolved oxygen, average (yearly) salamander count, and average (yearly) flow rate, average (yearly) water depth. These are all continuous, numerical variables, which allowed us to calculate their average and use that value. In the case of the ‘flow type’ variable, since it was a categorical variable, we used the mode value for each year, as that year’s flow type, as it would be the most indicative of what type of flow was seen that year.

5. Removal of unnecessary columns: Columns that weren’t being used, like method, project, and QC_FLAG were removed, and the data was tidied this way.

EXPLORATORY ANALYSIS

Objective

The objective of this section is to uncover initial insights into the relationship between ***dissolved oxygen levels***, ***flow rate***, ***water depth*** and the ***type of flow***, with **salamander count**. A total of four hypotheses were created and examined through data visualizations and analyses.

Hypotheses

Hypothesis 1: Increased levels of Dissolved Oxygen (DO) in a Barton spring are associated with a higher salamander count.

Dissolved oxygen is a measure of how much oxygen is dissolved in water and is available for living organisms (Water Science School, 2018). Since all mammals, birds, and amphibians need oxygen to survive, amphibians - and more specifically salamanders - require oxygen to aid their metabolic processes, we intuitively hypothesize that a higher rate of dissolved oxygen would increase the population of salamanders.

Hypothesis 2: Higher flow rates are associated with smaller salamander counts.

As an indicator of water quality, flow rate affects the amount of silt and sediment in a body of water (Environmental Protection Agency, 2012). This is important to species like salamanders because higher levels of sedimentation in water are known to cause degradation of habitat for amphibians, proving detrimental to their health and survival (Woods & Richardson, 2009). Since salamanders typically live in creeks, which have slower flow rates, we can hypothesize that a higher flow rate would decrease salamander count or population size.

Hypothesis 3: Lower water depth is associated with a higher salamander count.

The hypothesis proposes that a higher salamander count is associated with lower water depth as salamanders may be more abundant in shallower waters, as that could provide easier access to various microenvironments and breeding sites. Also, shallower waters might offer more opportunities for salamanders to find suitable shelter, such as under rocks, vegetation, or debris, contributing to a higher population. Lastly, salamanders, being ectothermic, may prefer shallower waters where temperatures are more easily regulated, influencing their metabolism and activity levels positively. Essentially, increased accessibility to diverse microenvironments, enhanced opportunities for shelter and breeding in shallower waters, and potentially more stable temperature regulation, favor growth in salamander populations (Woods & Richardson, 2009).

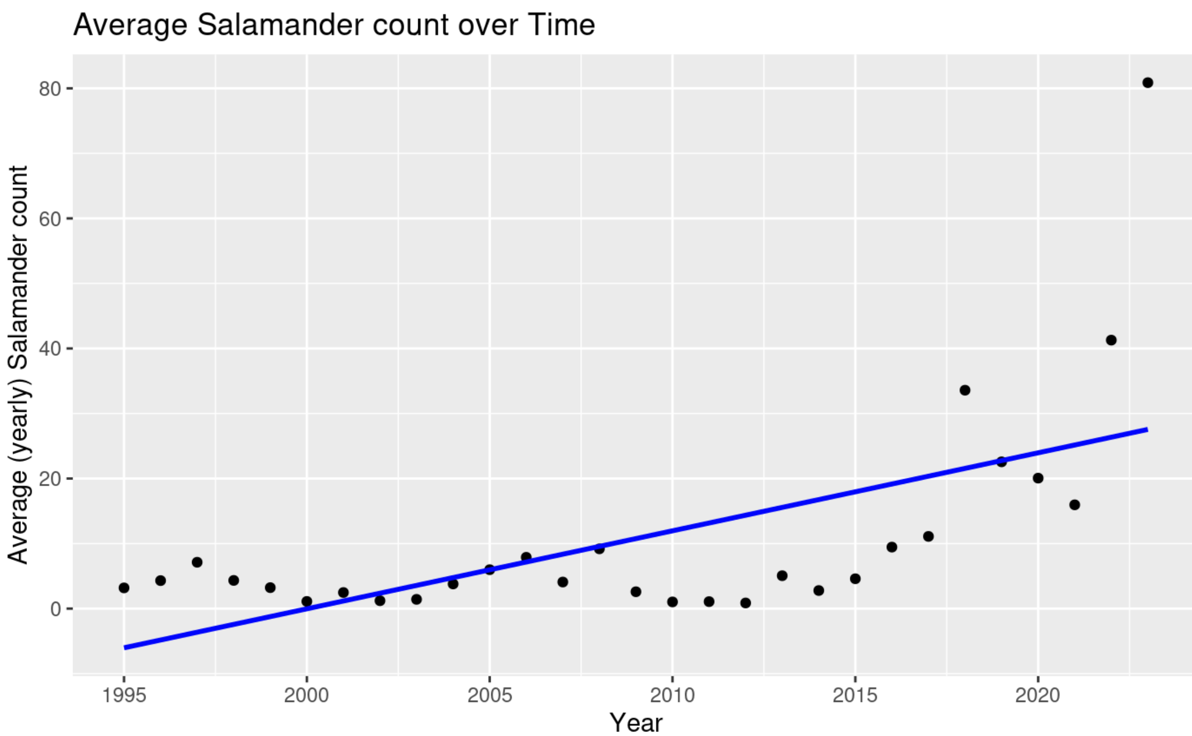
Hypothesis 4: Relationship between type of water medium and salamander count

The count of salamanders may increase as flow conditions lower (measure of depth). This is contingent on the fact that there is still enough water to create a suitable balance for habitat diversity and potentially an abundance of prey, but not an excess amount of flow that disrupts stable environments (Woods & Richardson, 2009).

MODELLING

Average Salamander Count over Time

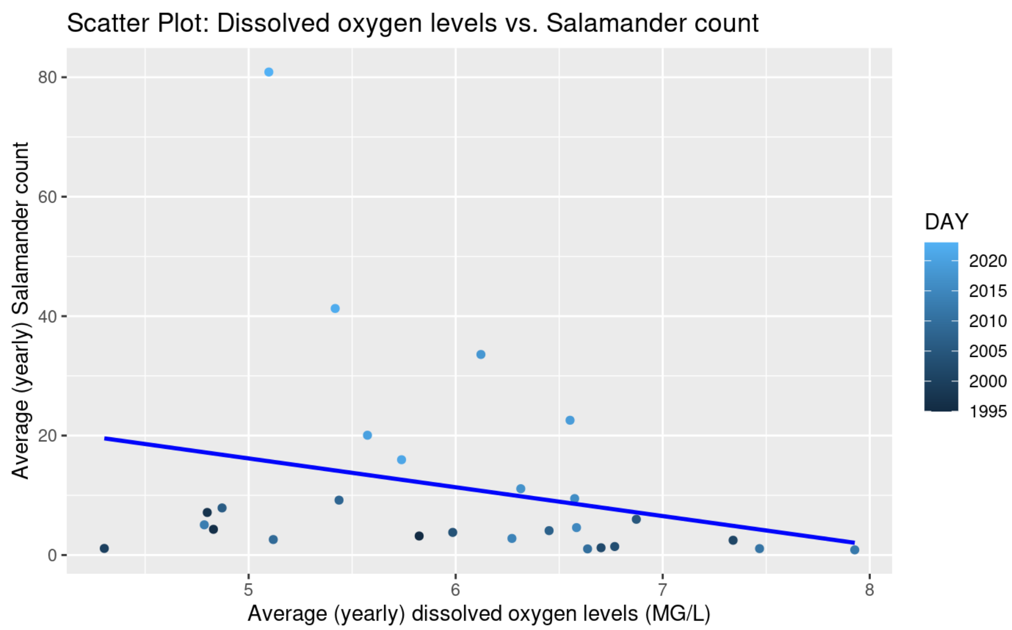
We initially wanted to see the relationship between average salamander count per year and time from 1995 to 2023, to see how salamander populations were changing over the years. To visualize this, we plotted the average salamander count per year against time. We also included a line of best fit (based on linear regression) to see the overall trend.



Through this scatter plot, we see that the average salamander count per year from 1995 to 2023 shows an increasing trend. What's interesting about this graph is the way the salamander count has increased over the years. From 1995 till 2015, the average number of salamanders fluctuated, without much variation. However, post-2015, the number of salamanders drastically increased, with 2023 recording almost double the number of salamanders compared to 2022. This made us curious to understand why this was the case, and whether the variables we have identified in our hypothesis will help us understand the trends in the salamander population over the years.

Average dissolved oxygen vs. Average salamander count

To prove the first hypothesis outlined above, a scatter plot was created between the average dissolved oxygen levels, which is the independent variable, and the average salamander count, which is the dependent variable. This was done to see if there was a strong linear relationship between the variables.

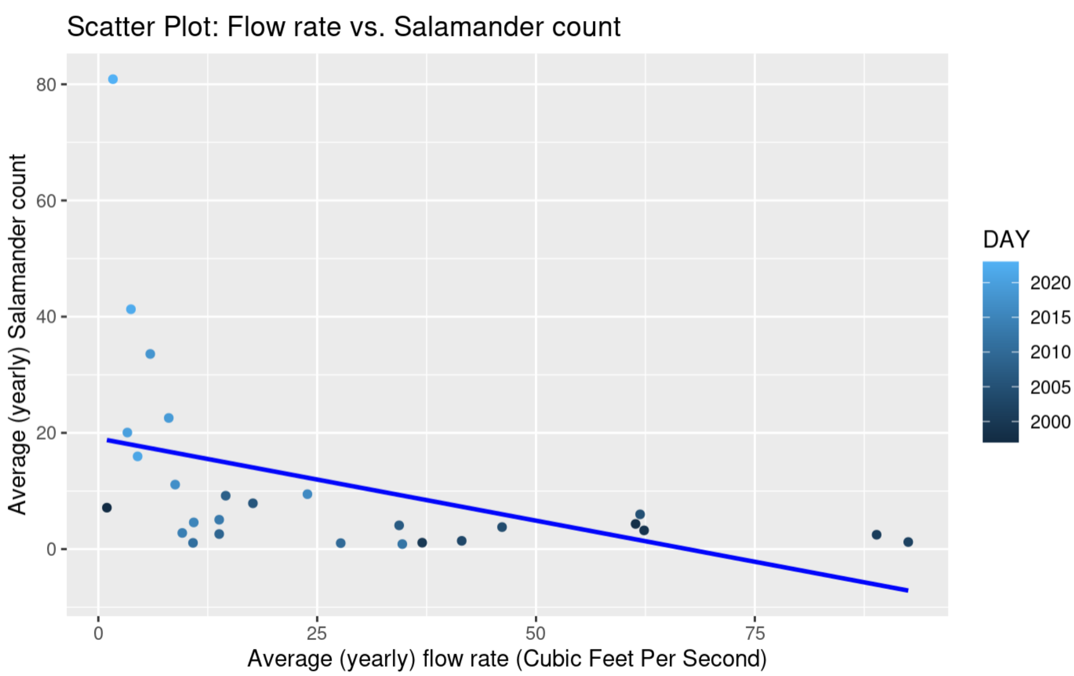


The line of best fit indicates an overall negative relationship between dissolved oxygen levels and average salamander count from 1995 to 2023. This appears to be in direct conflict with our hypothesis. We hypothesized that increased dissolved oxygen levels will be associated with increased salamander counts. This would be the case if these variables had a positive relationship. However, that does not appear to be the case.

It is also noticeable that the linear relationship between these two variables is not strong. There is a significant amount of variation or spread between the data points. We used R to calculate the correlation coefficient, which is -0.26 (estimated to two decimal places). This value confirms that there is an extremely weak linear relationship between the variables.

Flow Rate and Average Number of Salamanders

To prove the second hypothesis indicating higher flow rates are associated with lower number of salamanders, a scatter plot was created between the average flow rate, which is the independent variable, and the average salamander count, which is the dependent variable. This has been done to see if these variables have a linear relationship between them.

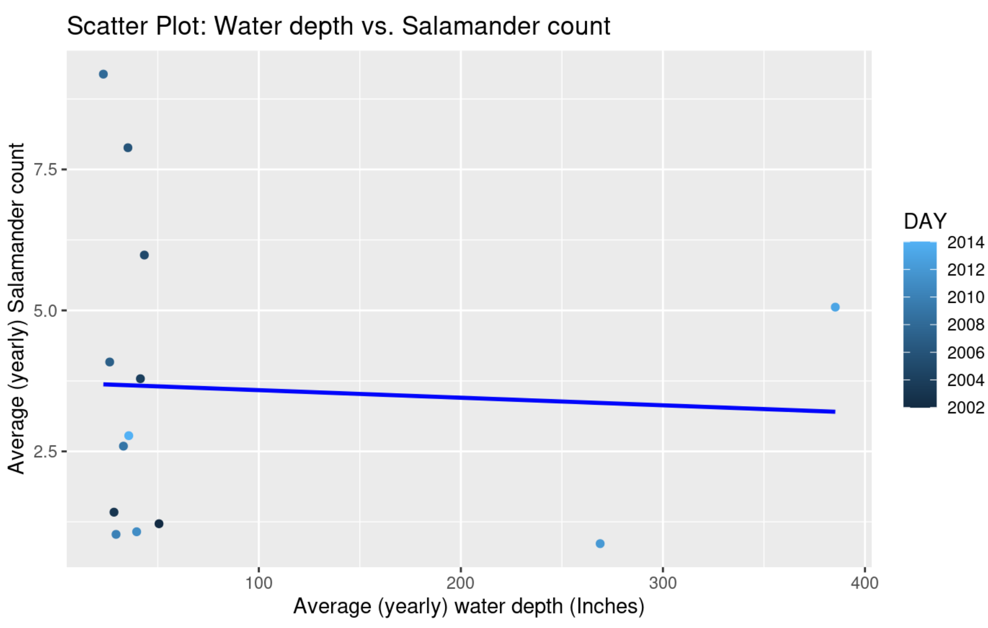


As seen in the graph, there is a clear negative relationship between the average flow rate and salamander count from data recorded from 1995 to 2023. This negative relationship supports our hypothesis, as it suggests that an increase in the average flow rate is associated with a decrease in the average salamander count.

Compared to dissolved oxygen, flow rate as a variable seems to have a stronger linear relationship with the average salamander count. However, the linear relationship still does not seem to be that strong. The correlation coefficient between these two variables is -0.43 (rounded to two decimal places). This shows that the average flow rate and average salamander count are more linearly correlated. However, the relationship is still not that strong, which can be seen visually as well.

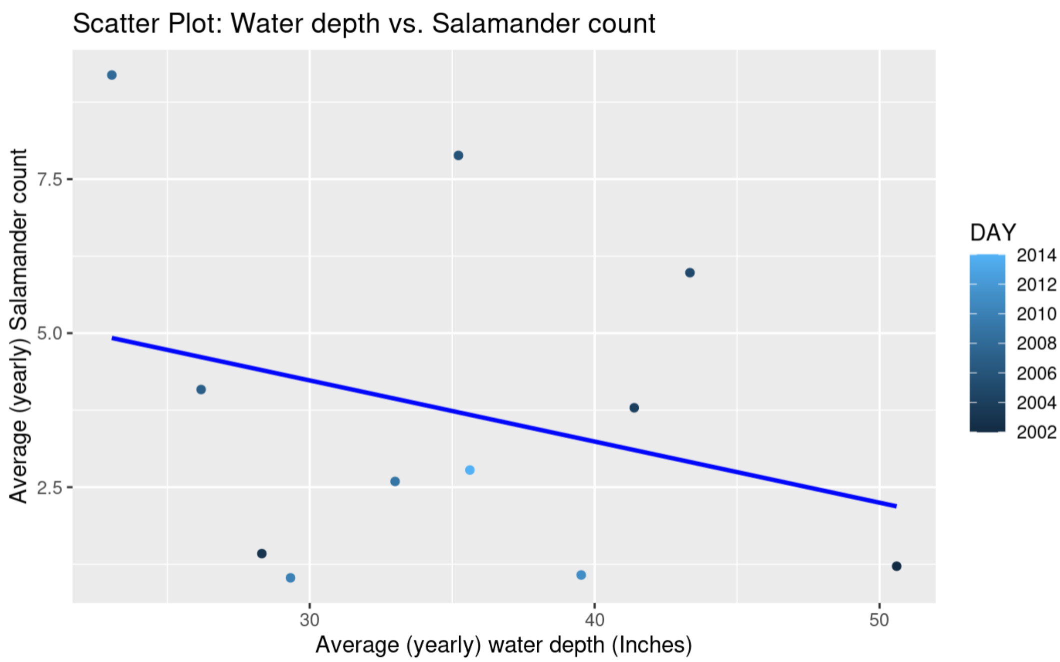
Water depth and Average Number of Salamanders

To prove the third hypothesis indicating that higher water depth is associated with a higher number of salamanders, a scatter plot was created between water depth, which is the independent variable, and the average salamander count, which is the dependent variable. This has been done to see if these variables have a linear relationship between them.



From a first glance at the graph, it appears that there is no linear relationship between average water depth and average salamander count. What's interesting about this graph is the way the 'water depth' data has been spread. Most of the data points are between 0 and 50 inches. Then, there are two extreme outliers.

We took a look at the data set again and saw that the average water depth data points from 2012 and 2013 were the outliers. So we decided to remove the outliers and see what the relationship looked like.

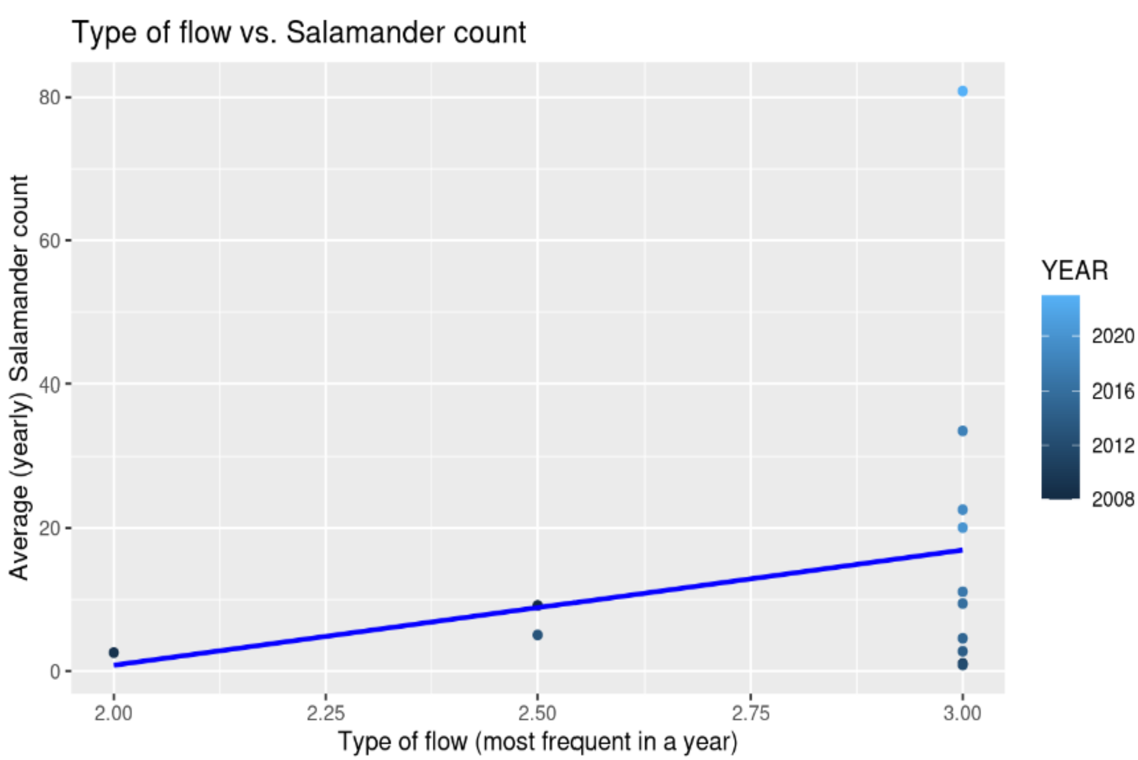


This is the graph that was generated. It seems to have a stronger linear relationship, with the outliers removed. However, there is still a lot of variability in the data points.

When it comes to understanding this graph in relation to our hypothesis, we hypothesized that this would be the relationship - that a higher water depth would lead to a lower salamander count. However, from the graph, the linear relationship between these variables does not appear to be so strong. The correlation coefficient value of -0.33 also supports this. One of the reasons for this could be that water depth values do not exist for the years before 2002 in this data set. In the case of dissolved oxygen and flow rate, there was consistent data from 1995 till 2023, for us to calculate the average values and plot. However, in this case, the data set simply did not contain water depth data until 2002, which reduced the number of data points we could use to analyze. Consequently, this could also be what's caused a weaker correlation coefficient.

Type of flow vs. Salamander Count

In an attempt to prove our fourth hypothesis of a relationship between flow rate and salamander population, specifically that a lower flow rate would result in a higher population, we curated a graph to display the type of flow, the independent variable, in relation to salamander count, the dependent variable.



The initial hypothesis proposed an expected increase in salamander count with lower flow conditions, based on the assumption that a lower flow rate would favor habitat diversity and prey abundance. However, the observed data contradicted this hypothesis, revealing the trend of a normal flow rate over time with significant variation in Barton Springs salamander counts. This is shown by the large amount of data points located at 3.00, indicative of a normal flow rate, yet the average salamander count seemed to vary independently of that. Ultimately, we were unable to conclude a relationship between flow rate and Barton Spring salamander population existed.

This discrepancy can be attributed to the complexity of aquatic ecosystems, where salamander populations are influenced by a multitude of factors beyond just water depth and flow rate. Seasonal and environmental variability, habitat resilience, interactions with other species, and potential human impacts are additional factors that may contribute to the observed variations.

REGRESSION

The data visualizations that we did above help us understand a few things. Firstly, we came to understand that ‘type of flow’, as a variable, has absolutely no effect on salamander count. So, we decided that we were not going to investigate this variable anymore, because we would not be able to use this variable to explain the trends in salamander population. So, it did not serve our analysis in any way.

Secondly, we saw that dissolved oxygen, flow rate and water depth did have relationships with salamander count. However, the linear relationships between each variable and salamander count were not strong enough. So, we decided to investigate different regression models for each of these variables, to understand which regression model fits best. We employed 4 (linear, quadratic, exponential, and generalized additive) regression models to uncover potential patterns and trends.

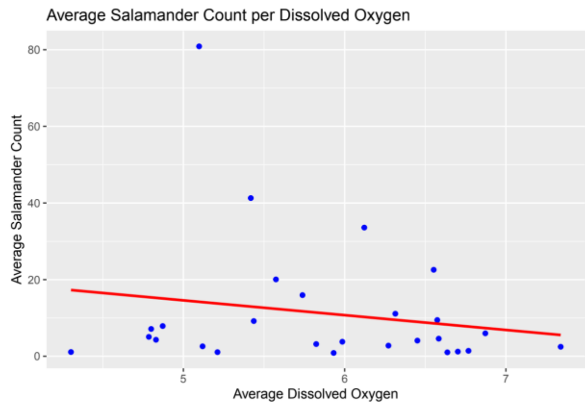
In our analytical approach, we opted for regression models over classification due to the nature of our variables and the objectives of our study. The primary focus of our investigation involves understanding the continuous outcomes, salamander count, in connection to other continuous variables like dissolved oxygen levels, flow rate and water depth. Regression would hopefully provide us with an understanding of how these variables would affect salamander count in the future, which would help inform conservation decisions. Linear regression was originally chosen to explore potential linear associations between these variables, which gave us initial insight. It provided a fundamental understanding of the data and helped us understand where we needed to move from there.

Additionally, the inclusion of quadratic and exponential regression models allows us to capture more complex patterns that may exist in the data that aren't linear in nature. The quadratic model is suitable for exploring non-linear relationships while the exponential model helps identify exponential growth or decay patterns over time or across different dissolved oxygen rates. By utilizing these models, we aim to unveil nuanced trends that may not may not be immediately apparent through simple linear analyses.

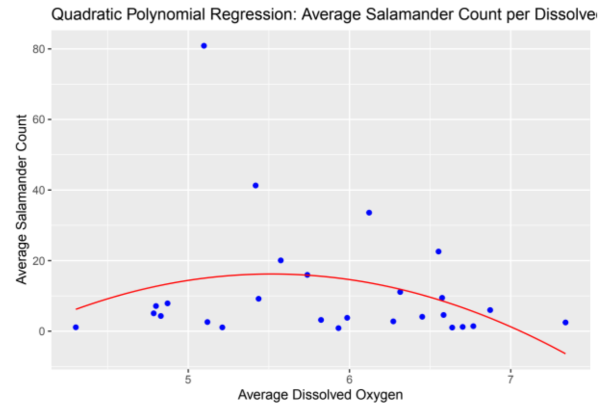
Furthermore, incorporating a generalized additive model (GAM) enhances our analysis by accommodating non-linear relationships and interactions between variables. This flexibility was crucial as we are dealing with ecological data that will exhibit complex patterns influenced by various factors.

Therefore, as we continue to discuss the modeling of our data, we focused on regression models based on their ability to provide quantitative insight into continuous outcomes to predict how the ecosystem may change in the future. This section will delve into how we used regression to understand the factors influencing salamander populations in our data.

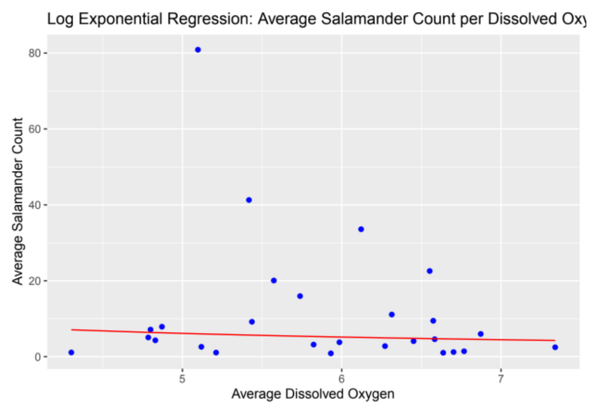
Dissolved Oxygen Regression



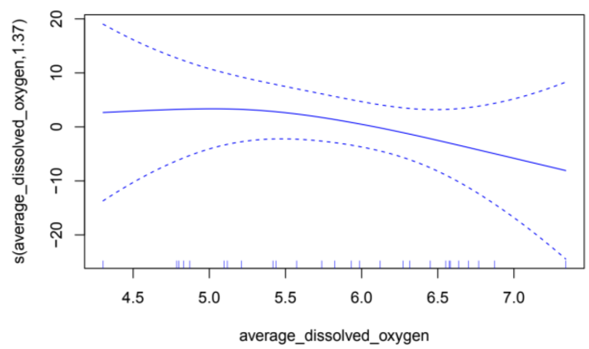
Linear Model



Quadratic Model



Log Model



GAM Model

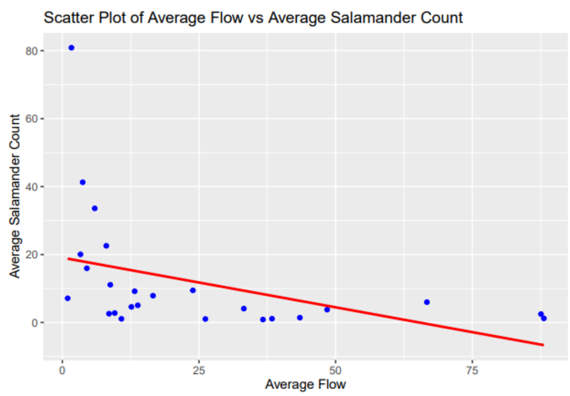
We conducted a comprehensive analysis of regression models to understand the relationship between average dissolved oxygen and average salamander count. The statistical significance and the fit of each model are summarized in the table below:

Model	Adjusted R-squared Value	P- Value
Linear	-0.006803	0.0326
Quadratic	0.01356	0.3249
Log	-0.02782	0.591
GAM	0.0127	0.515

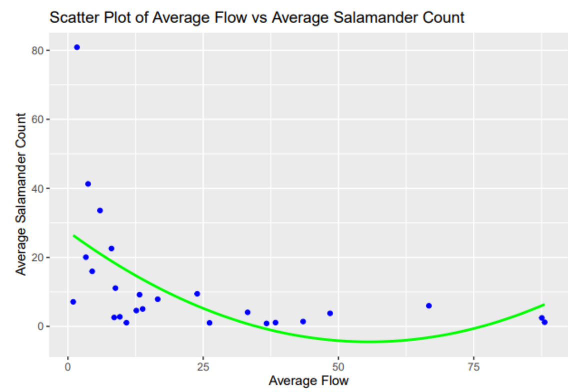
In the Linear Model, the negative adjusted R-Squared value suggests a poor fit but the P-Value being below 0.05 could potentially mean a slight statistical significance. However, as we look at the other models, including the Quadratic model which has a moderately better fit, we can see that this P-Value is not important or representative of where we should go with this data.

In summary, none of these models demonstrate a strong explanation for the relationship between Dissolved Oxygen and Salamander Count. The statistical significance of the linear relationship is notable but as the fit is poor, we are unable to draw any significant conclusions from it.

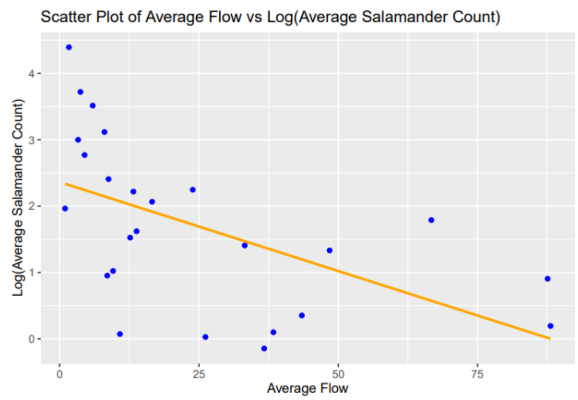
Flow Rate Regression



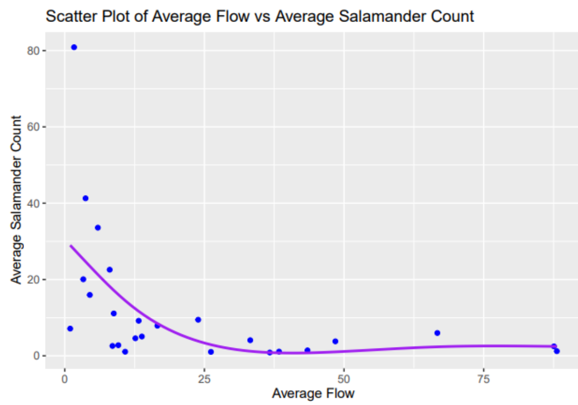
Linear Model



Quadratic Model



Log Model



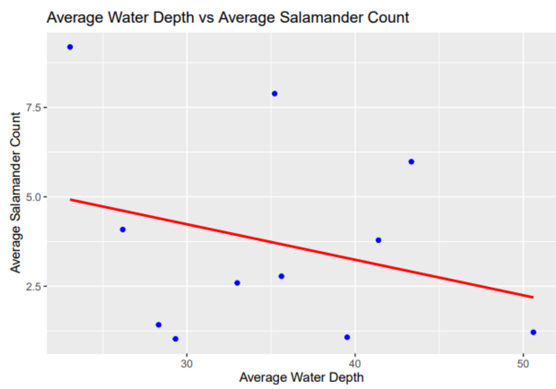
GAM Model

We conducted a comprehensive analysis of regression models to understand the relationship between Average Salamander Count and Average Flow. The statistical significance and the fit of each model are summarized in the table below:

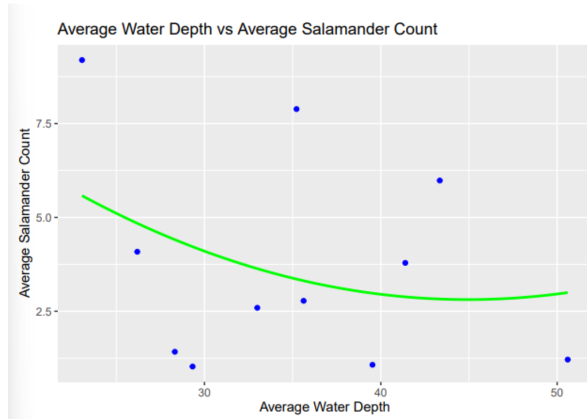
Model	**Adjusted R-squared Value**	**P- Value**
Linear	0.139	0.03752
Quadratic	0.2527	0.01559
Log	0.2609	0.005321
GAM	0.338	0.0251

While the linear model demonstrates a statistical significance with a p-value less than 0.05, further exploration reveals that both the quadratic and log models exhibit superior fits, as indicated by higher adjusted R-squared values. Notably, the Generalized Additive Model (GAM) stands out as the most robot, with the highest adjusted R-squared of 0.338 and a p-value of 0.0251, emphasizing its relatively strong compared to other variables, and significant relationship. For the Average Flow Rate, the GAM model emerged as the most suitable. However, in general, an R-squared value of .7-.8 indicated a strong account for variability by the independent variable. But, as we've seen throughout our complex dataset, we are not seeing R-squared values that high, therefore in this specific analysis, an R-squared value of 0.338 is significant as mentioned above.

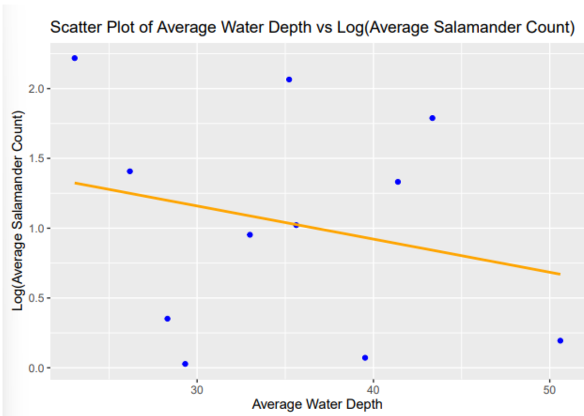
Water Depth Regression



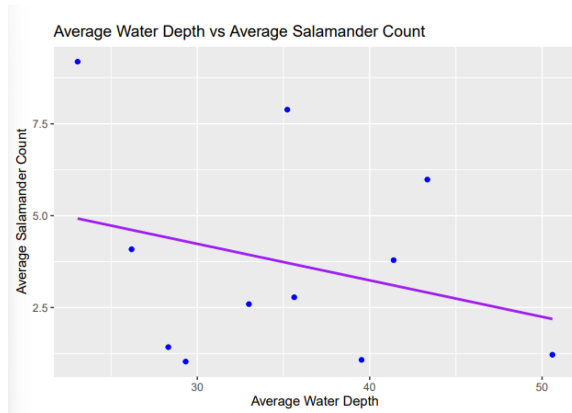
Linear Model



Quadratic Model



Log Model



GAM Model

Next, we looked in Water Depth and Salamander Count and immediately we realized that was a lack of statistical significance and there was a complete lack of meaningful insight we could glean from this data. As evident below, the values given by these models showed that.

Model	**Adjusted R-squared Value**	**P- Value**
Linear	-0.01913	0.3909
Quadratic	-0.121	0.6468
Log	-0.04475	0.4689
GAM	0.431	0.241

A few notable values include:

- The Adjusted R-squared values for all models are negligible, indicating poor explanatory power.
- P-values for each model are notably above the conventional 0.05 threshold, signifying a lack of statistical significance.
- The Generalized Additive Model (GAM) shows a slightly positive Adjusted R-squared but falls short of significance, suggesting limited explanatory value.

Therefore, we decided against using this as we continued this project. Further exploration or considerations into why this lack of significance may need to be done, but for the sake of this project and the dataset, it was out of our understanding and abilities.

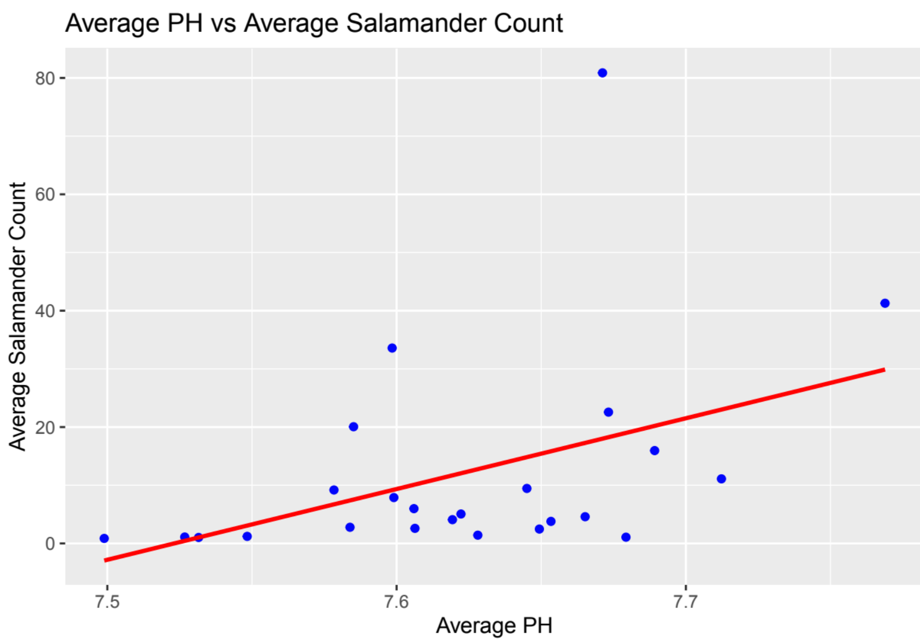
FURTHER ANALYSIS

After thinking a bit more about what we concluded, we realized that dissolved oxygen and flow rate were the only two variables from our initial hypotheses that were connected to salamander count. Moreover, there weren't any more variables on our data set that we could use to understand why salamander count is the way it is.

So, we decided to bring in another data set, and analyze a variable in connection to salamander count. We found a data set from the city of Austin titled 'Water Quality Sampling Data' that contained information on multiple water bodies located in Austin. It was a data set that had 1.4 million data points, so we assumed that we would be able to find another variable that could be used to explain salamander count. We filtered the data set to water bodies located only in Barton Springs (using the latitude and longitude data) and made sure we were analyzing information from the same water body source as the ones the salamanders live in.

We started looking into the data set and had to clean up a lot of the data to understand what variable we could use for our analysis. After removing a lot of unnecessary data and filtering the data set significantly, we found some interesting data on carbon, phosphorus, and ammonia levels. We tried to get that data from 1995 to 2023 to analyze them in connection to salamander count, but each variable had a unique issue. While a variable like carbon did not have information before 2014, a different variable like ammonia did not have consistent data, with the units of measurement constantly changing between the years.

Eventually, we found pH levels as a variable that had consistent data across all the years of our analysis. We calculated the average pH values from 1995 to 2023 and analyzed them in context to salamander count.



Based on an initial glance, average pH seems to have the strongest linear relationship with average salamander count, compared to any of the other variables. This was seen with an adjusted R^2 value of 0.1409, which is statistically significant at the 95% confidence level.

LOOKING INTO MULTIPLE REGRESSION

After analyzing dissolved oxygen level, flow rate, water depth and type of flow, and seeing that dissolved oxygen and flow rate were the only variables that had a relatively strong association with salamander count, then looking into data like pH levels from other data sets that did have a stronger association with salamander count, we decided to explore multiple regression.

Using dissolved oxygen, flow rate and pH levels, we thought that it could be interesting to build a multiple variable linear regression model to calculate salamander count.

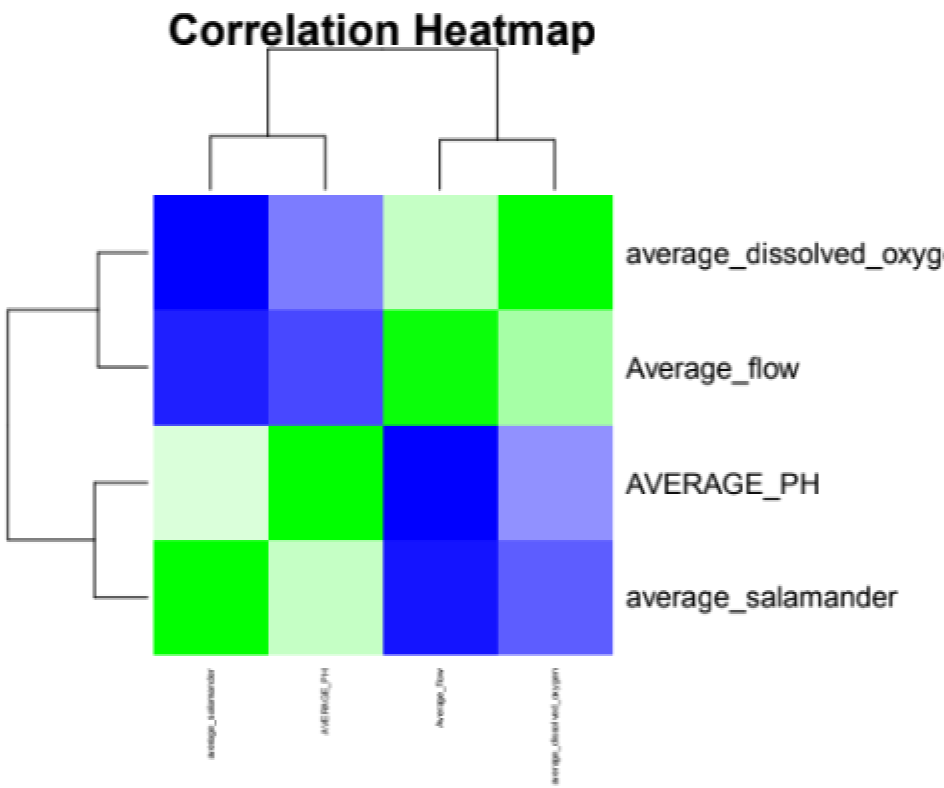
Note: We did multiple linear regressions for these variables, even though linear models were not the best fit for our variables. However, we tried to do other models and felt it was beyond our scope.

After plotting a multiple variable linear regression model, with dissolved oxygen, flow rate and pH levels as the independent variables, we found this:

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.350  -7.316  -3.543   4.334  57.211
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -714.0453   450.1544  -1.586    0.128
## AVERAGE_PH                98.0094    59.8092   1.639    0.117
## average_dissolved_oxygen  -2.7118     5.2780  -0.514    0.613
## Average_flow              -0.1887     0.1718  -1.099    0.285
##
## Residual standard error: 16.38 on 20 degrees of freedom
## Multiple R-squared:  0.2876, Adjusted R-squared:  0.1808
## F-statistic: 2.692 on 3 and 20 DF,  p-value: 0.0737
```

Average pH levels have the strongest correlation with salamander count. This is seen with the p-value being the lowest (0.117), compared to average dissolved oxygen (0.613) and average flow (0.285). The adjusted R-squared value of 0.18 suggests that this model can be used to explain the average salamander count, but it is not the best fit in terms of a model. The p-value of 0.0737, however, suggests that the model is statistically significant at a 90% confidence level.

After analyzing the multiple variable linear regression model, we were curious to see if we could establish causation, more than just correlation. We decided the best way to understand the multiple variables and their causation to the salamander count would be to make a correlation heatmap.



As shown, there is correlation between variables. As it can be seen, there is correlation between each of the independent variables: between average flow and average dissolved oxygen, between average pH and average flow, and between average pH and average dissolved oxygen. Since there is correlation between the independent variables, causation cannot be established. So, we cannot use our multiple linear regression model to establish any conclusion about any causative relationship that affects salamander count. However, it does show us which variable, out of the 3 independent variables, has the strongest correlation / association with average salamander count. We found that to be pH levels.

Limitations to our research

After working on this project, we uncovered several limitations to our research and modeling. Our greatest limitation was the need for more reliable, credible data. After our analysis, we started wondering why none of the variables that were present in the data set - such as dissolved oxygen, flow rate, water depth and type of flow - had a significant relationship with the salamander count. We decided to investigate the source of the original data and found the dataset under consideration is compiled from various studies and lacks direct vetting by the city of Austin, causing a level of uncertainty regarding its accuracy. While the city has implemented a flag system to signal data reliability, the actual accuracy remains uncertain. Notably, a significant change occurred in the measurement of dissolved oxygen and associated tools between 2015 and 2017, raising questions about the consistency and comparability of data before and after this period. Several issues persist within the dataset, exacerbated by inadequate explanations for the variables. For instance, the variable "water depth" lacks clarity as its measurement location, whether at the center or on the sides of the hot spring, is unspecified, leaving the interpretation open to ambiguity. Another limitation was the overall sample size of what we could use for modeling. Another limitation is in recognizing the temporal and spatial limitations of our study, we recommend that future research considers a more comprehensive approach by incorporating long-term monitoring data and expanding the study area to provide a more nuanced understanding of salamander dynamics over time and across diverse habitats. This could be difficult however due to the emigration patterns of salamanders. Right now, sample collections are done in the spring due to suitable habitats and lower stormwater flow, but that results in limited knowledge (Adcock et. al, 2022). Most notably, limitations in our analysis arise from the decision to utilize average data over a year, a choice that, in hindsight, proved to be restrictive.

The initial assumption was that the dataset, spanning from 1995 to 2023, would provide an ample amount of data points, exceeding 25. However, as we progressed in our analysis, it became evident that numerous variables lacked data for certain years, significantly diminishing the available data points. Specifically, the variable "water depth" only had data available from 2002 onward, further reducing the dataset size. Moreover, when confronted with outliers, the necessity to exclude specific data points for the sake of robust analysis further contributed to a reduction in the number of usable data points. This limitation hindered our ability to establish the strength of relationships between variables, emphasizing the importance of considering smaller time intervals and thoroughly evaluating data availability when designing future analyses. Acknowledging and addressing these limitations will contribute to the refinement and advancement of salamander ecology research, and aid to slow and even stop the extinction of some specific salamander species.

Ethical Implications of the Project

Our project has focused on the preservation and conservation of a population of salamanders unique to Austin - giving our research project distinct ethical implications surrounding continuity and openness, and sustainability.

As mentioned by Professor Pepita Barnard, one of the most crucial questions to explore when reflecting on the ethical consequences of a field of questioning, especially in nature, is not only how surrounding communities are affected by the research we do, but also how the plants and animals which make up this ecosystem, will be affected. Since the Barton Spring salamanders are an 'at-risk' species, the datasets available to us were collected for the sole purpose of preservation and conservation of these animals. Since they are so unique to Austin, native 'Austinites' have a sense of kinship to them and loyalty to ensure they remain protected, and such data contributed to future efforts in making sure the salamander population at the spring is healthy and thriving. So, even if the primary goal of this data collection was to understand whether salamanders are affected by factors like flow rate, dissolved oxygen, and pH levels, how does this affect other animals in the springs - and what about communities in the surrounding areas?

On average, $40,000 gets set aside for conservation (with proposal funds allocating about $10,000-$80,000 based on the request) and since Barton Springs itself is state-owned land, the land surrounding Barton Springs is a more expensive area to live in as well. As funding increases for Barton Springs, so does the upkeep and as a result, the property values of the surrounding area increase, leading to a conflict of interest for people who have been living there for so long, now unable to remain in the area because of these high prices. The continued need to fund the conservation of salamanders contributes to the preservation of an endangered species, in turn bolstering the biodiversity of the area and ensuring the ecosystem of Barton Springs remains in balance, however, it begs the question of the city's prioritizations in the bigger picture of rehousing a sect of the population who now cannot afford to live near such reserves (Silver, 2022).

In the vein of public living and opinion, as aforementioned, people living in Austin have loyalty to the salamanders which opens a conversation into the ethical implications of maintaining an open dialogue with the public on conservation efforts. In our experience, fortunately, our data was publicly accessible, and all the necessary information was shared about the salamanders and their conservation efforts. However, the data was extremely inaccessible from a layman's perspective, and it was clear that prior academic knowledge and experience in research was needed to understand the data and associated variables. Critically analyzing untidy data has ethical implications in terms of openness towards the public because of methodology changes observed over time. When examining our dataset, it was clear that not only was this data collected over an extremely long period of time, the methods of collecting data had changed over time as well, though the methods weren't specified, and the external effects of changing methods were not considered either. However, the ethical implications of our own project contribute to the increased accessibility of this data as making connections between the various variables and allowing for the public to clearly see the impact of Barton Springs' unique makeup on the growth of its native salamanders can help future policymakers on their conservation efforts.

The ethical implications of an untidy dataset concerning the use of the dataset for research is that we don't know whether changing the method of data collection has affected the actual data itself and therefore the actual accuracy of the observed data. If the data is being used to determine whether salamanders are still an at-risk species in need of public funding, it is incredibly important to have the right data so that the right amount of funds are allocated to such efforts, and the salamanders can be best protected.

For the most part, since our data collection and hypothesis is concerned with sustainability and conservation, our ethical concerns and analyses revolve around making sure that the data we analyze can accurately determine how salamanders are able to thrive in their natural environment, and how the public which interacts with this environment can ensure the longevity of these efforts and contribute to it as well.

Conclusion

After analyzing linear regressions, quadratic, exponential, and generalized additive regressions to glean any relationship between the variables detailed in our previous hypotheses, we found that dissolved oxygen and flow rate were the variables that had a relationship with salamander count. We found that the GAM model had the best fit in the case flow rate, while the quadratic model had the best fit in terms of dissolved oxygen. Eventually, after adding in pH levels, we were able to conduct a multiple variable linear regression and concluded that pH levels had the biggest association with salamander count.

While these results did not confirm our inferences, there are still avenues for potential research which our study and the existing and tidied datasets could be used for. As mentioned previously, the dataset that we had encountered was deeply inaccessible and took extensive tidying to make it usable - meaning that for the future research going into the conservation of the salamanders, there is now a more accessible dataset available. Additionally, since we were able to determine a relationship existed - though not as strong - it can be used as a stepping point to focus more on ‘ideal’ conditions on which Barton Spring salamanders as well as other endangered species within the spring can thrive. From a conservation standpoint, the potential research avenues center mainly around continuity and maintaining the spring to be a safe place for the salamanders’ survival.

Works Cited

U.S. Fish and Wildlife Service. 2005. *Barton Springs Salamander* (Eurycea sosorum) Recovery Plan. U.S. Fish and Wildlife Service, Albuquerque, NM.

Water Science School. (n.d.). *Dissolved Oxygen and Water*. Dissolved Oxygen and Water | U.S. Geological Survey. https://www.usgs.gov/special-topics/water-science-school/science/dissolved-oxygen-and-water

Environmental Protection Agency. (2012, March 6). *What is Stream Flow and Why is it Important?*. EPA. https://archive.epa.gov/water/archive/web/html/vms51.html#:~:text=Flow%20is%20a%20function%20of,to%20dilute%20and%20degrade%20wastes.

WOOD, S. L., & RICHARDSON, J. S. (2009). Impact of sediment and nutrient inputs on growth and survival of tadpoles of the western toad. *Freshwater Biology*, *54*(5), 1120–1134. https://doi.org/10.1111/j.1365-2427.2008.02139.x

Adcock, Z. C., MacLaren, A. R., Jones, R. M., Villamizar-Gomez, A., Wall, A. E., White Iv, K., & Forstner, M. R. J. (2022). Predicting surface abundance of federally threatened Jollyville Plateau Salamanders (*Eurycea tonkawae*) to inform management activities at a highly modified urban spring. *PeerJ*, *10*, e13359. https://doi.org/10.7717/peerj.13359

Silver, J. (2022, September 14). *This affluent Austin neighborhood declared the most expensive in Texas*. CultureMap Austin. https://austin.culturemap.com/news/real-estate/05-21-19-barton-creek-most-expensive-neighborhood-texas/