The purpose of this project was to determine what factors are most correlated to vulnerability in watersheds. A watershed is part of a region's drainage basins.

Rather than creating a model to predict the priority index (target variable) of a watershed, we instead used machine learning to identify the features that have the highest impact on a watershed's need for conservation. This way, it will be easier for lawmakers to determine the areas that require the most attention and funding for conservation, and what needs to be specifically targeted to make the most impactful change in the watershed.

Vulnerability is represented by the Priority Index, assigned by the EPA to all HUC8 (Hydrologic Unit Code) watersheds to reflect the ratio of native/invasive species. Features from other datasets were merged to reflect possible reasons for vulnerability. Feature topics include; travel metrics, human activity, big game hunting popularity, fertilizer runoff, river metrics (length, density, turbidity), ground/surface water nitrate levels, and temperature. Following our presentation several improvements were made:

1. The method of merging relevant features was improved and refined. Used functions to make it more efficient and easier to read.
2. The function to determine feature importance was refined to use an ensemble method, considering input from 5 different models.
3. The gradient-boosting regressor model was our most successful, so I ran it through an Exhaustive Feature Search. It returned an r^2 value of .68 along with its ranking for statistically significant features.
4. Ports act as sources of pollution and introduction of invasive species. I made a function to get a count of all plots of us ports that occur within, or upstream of a region's drainage basins and the correlation (r^2) between that count and the Priority index (r^2 = 0.4).

Overall the largest contributors to watershed vulnerability are transportation relation (STXRD = roads over streams, RDDNS = density of roads in region, combustion_transportation), human-related (human waste acts as a proxy for general human activity), or related to the length of the river. The length of a river introduces more sources of pollution such as various ports.

At the start, we hypothesized that watershed vulnerability would be strongly correlated to chemicals. This is not the case, the impact of chemicals such as nitrogen or phosphorus from fertilizer runoff into surface or groundwater was overshadowed by the effect of human activity. Introducing the speed and range of human travel to basins acts as a gateway for flora, fauna, and bacteria to migrate between unconnected basins. The rate of human transportation, by road and sea is the leading contributing factor in watershed vulnerability.

The formula to determine the priority index considers, in addition to the ratio of invasive/native species, the rarity of all taxa. A priority index for a region will be higher if the rarity and natural biodiversity of a region are above average so that more valuable watersheds are valued higher. The Tennessee basin, the region with the highest priority index, is one of the most biologically diverse river systems in North America. The Asian carp were brought to the U.S. in the 70s to control algae bloom, a product of the increased fertilizer in agricultural runoff. It was introduced in the southeast in contained environments, as the Asian carp poses a serious risk as it lacks native predators in the U.S. So when it eventually escaped encampment it quickly spread throughout the southeast and threatened to travel into the Great Lakes due to the reversing of the Ohio River, a topic outside the scope of this project. The Asian carp is a problem for every river in the southeast and is especially damaging to biodiversity.

The visualization of the priority index by HUC2 regions could reflect the impact of commercial and private travel through the roads and river or it could reflect the severity of Asian carp on the biodiversity and health of Tennessee's watersheds and those across the southeast.

Based on the results of our regression models and feature analysis, we have determined that roads over streams and road density are two of the highest contributors to the need for watershed conservation. Based on this information, conservation efforts in the higher-risk regions of the U.S. should focus on mitigating the negative effects of transportation around bodies of water. These efforts could include anything from safer ice-melting materials to the investment in electric vehicles. Additionally, more actions should be taken to limit the number of invasive species being introduced to domestic ecosystems. Ultimately, this analysis demonstrates the need for environmental conservation efforts in U.S. watersheds and begins to inform possible strategies for them.