“The Xerces Society for Invertebrate Conservation is an international nonprofit organization that protects the natural world through the conservation of invertebrates and their habitats” (Mission & History). The name Xerces is based off of an extinct butterfly, Xerces Blue. This extinction was the first known in North America due to human activities. “The Xerces Society’s core programs focus on habitat conservation and restoration, species conservation, protecting pollinators, contributing to watershed health, and reducing harm to invertebrates from pesticide use” (Mission & History).

The Eureka team is focusing on the data from the Xerces Society Pollinator Conservation Program. The core focus of Pollinator Conservation Program is conserving the pollinators habitats and attracting native pollinators. “The ecological service they provide is necessary for the reproduction of over 85% of the world’s flowering plants, including more than two-thirds of the world’s crop species” (Pollinator Conservation). The impact of losing pollinators is monumental not to just humans but to animal life as well. “The economic value of these native pollinators is estimated at $3 billion per year in the U.S” (Pollinator Conservation). Therefore, not only would food from crops be in ruin but food obtained from most animals would be ruined as well. The main threats to these pollinators are habit loss, pesticide use, and introduced diseases.

The Eureka project team was provided the data in Excel format. The Excel spreadsheet was composed of six sheets, Site, Visit, Veg (Vegetation), Macro, Seed and Species Attributes. The Site data contained field id, a region, the number of acres associated with the field, the county, the date the site was planted, the planting method, the number of species planted, the Pure Live Seed pounds per acre, the previous vegetation cover and the Vendor for the seed planted. There was very little missing data in the site data, two missing cells for Number of Species and Pure Live Seed pounds per acre for the Prairie Plains Seed Vendor. The team received 14 rows of site data.

The visit data contained a field ID, a visit ID, Transect, the visit date, the observer, the surrounding Landscape and the Notes about the visit. Much of the data for the surrounding landscape is duplicated for each transect and each visit. The notes cells have the missing data for this sheet. The notes also contain an indication if the site failed. Another indication of a failed site was a fall visit did not occur. The Surrounding Landscape and the Notes columns were free text fields and difficult to parse. The Surrounding Landscape data has the potential to add value though. As a result, this data was parsed manually into a new dataset. The team received 56 rows of visit data.

The veg or vegetation data contained a Visit ID, a species and 20 quadrat columns divided by floral cover and non-floral cover. The quadrat contains a number from 1 to 6 that translates to the amount of coverage for that species within the quadrat. Much of this data is blank but not missing which is to be expected. The team received 891 rows of veg data. The macro data set contained a Visit ID, 3 columns that categorizes the species cover in the macroplots, <1%, 1-5% and > 5%, and Notes. The Notes are free form text. Much of this data is blank but not missing which is to be expected. The team received 650 rows of macro data.

The seed data set contained a Field ID, Species, Pure Live Seed lbs per acre and Origin. This data set contains no missing data. The sum of the Pure Live Seed pound per acre by Field ID data matches the value found for that Field ID in the Site data sheet. The team received 543 rows of seed data. The final data was the Species Attribute data set. This data set contained Species, Coefficient of Conservatism and common Name. A significant portion of the Coefficient of Conservatism is marked with an asterisk. These species will not be included in any coefficient of conservation averages per the client. The team received 182 rows of Species Attribute data.

The team used the data in the spreadsheet and used it to form a few other data sets. The first new data set, we called expandedSiteData. This data set has the site data, visit data and the new manually parsed surrounding landscaping data. The surrounding landscaping data was parsed into 7 new columns, LandSTree, LandSRoad, LandSStructures, LandSSCrop, LandSGrass, LandSWaterArea, and LandSOther. The success or failure from the notes column was added to a new column called Established. The columns not included in the new data set were Visit ID, Observer, Visit Date, Surrounding LandScape and Notes columns.

The team also expanded the Species Attributes data set by adding taxonomy data. This addition was done in an effort to collapse the species data without losing essential meanings and making it easier to work with. This task allowed us to reduce more than 700 rows of data into 1 row of data, the order associated with the more than 700 species. The team also looked at the veg data for ways to format it so that we could increase our success with analysis. We decided to deconstruct it. We took 22 columns and constructed a new data set that was longer but not as wide. Instead of 22 columns, the data set had 4 columns, Visit ID, Species, quadrat. This data set was also combined with data from the visit data set as well as the site data set.

The team decided that we would use the data to evaluate 4 research questions

1. How has the variables associated with the site contributed to the success or failure of the site?
2. How does the factor of being native versus non-native affect the plant growth of species in the fields tested?
3. Do certain plant families show a greater cover value than others plant families?
4. Does the average Coefficient of Conservatism of the species planted in each field contribute to the success of failure of the field?