3 things Data Science taught me about environmental restoration.

1. Where you plant matters (sometimes)
   1. How I identified areas of interest (keep it simple, stupid)
   2. What to do about it
2. The life cycle of plants is important
   1. Short lived plants will change your mortality stats. This is ok, but take it into account.
3. Sometimes you should cut your losses
   1. Our policy is to replace a dead plant with the same species if possible. Sometimes a species just isn’t suited to where you introduced it, try something else instead.

The CRISP-DM Process

1. Business Understanding
   1. Is the ecological restoration process working? Is it working mechanically or ecologically?
   2. Where is it working better, and why?
   3. What can we do to improve our outcomes?
2. Data Understanding
   1. Our metrics for restoration (for now) are plant survivorship, size, and health. Are those improving over time?
   2. Specific species or locations will perform better. Identify where these stand out, correlate with some variable such as health or location and then posit theory
   3. We can’t control weather. We may be able to control for pests. We can definitely control which species we plant, and when we plant them. Other variables we haven’t really measured so I can’t draw conclusions.
3. Data Preparation
   1. Select data: Select a single scenario because the varying designs, altitudes, weather patterns, etc makes comparisons between scenarios difficult. You can use different scenarios to answer different questions, as long as they are still valuable to the business.
   2. Clean data: Make sure everything is spelled correctly, luckily most is cleaned already. Separate by planting dates and treatments if necessary.
   3. Construct data: Done for us with plant cover.
   4. Integrate data: Needs to be done with geographic data as well as with nursery data if applicable.
   5. Format data: Make sure data types are correct. Make sure NaN values are correct.
4. Modelling (At this point, you may have already answered some questions)
   1. Select modeling techniques: Find best fit (linear regression) (I only know one other for now: Supervised machine learning)
   2. Build model: linreg should be simple. SML see if you can follow udacity example
   3. Split data into model and test sets? Try following udacity example
   4. Assess model: Are these statistically viable? Or random? How do the two models compare?
5. Evaluation
   1. Evaluate results: Do the models meet the business criteria? Which ones answer the questions best?
   2. Review process: Summarize findings, look for error, correct anything if necessary. A code review at this point would work gangbusters.
   3. Determine next steps: Do we proceed to deployment? Do we need to iterate more? Do we need new data?
6. Deployment
   1. Plan deployment: Plan how to present the results
   2. Plan monitoring and maintenance: These datasets will get updated. Modular code will help keep this shit up to date. Refactor.
   3. Produce final report: Present a summary of the results
   4. Review project: What could have been done better? How to improve for next time?