WFSC 223: Dealing with Data in the Wild

This course provides an introduction to solving real-world problems with data. The course is divided into five modules, each of which presents a different challenge for students to solve.

We will use R to start with, but we designed the course to change to python or another language in future iterations.

Summary of Challenges

1. **Module 1**: **Mission Antarctica!** A new effort to establish a permanent, sustainable colony in Antarctica is being launched. Students are introduced to the field of data science and the application (RStudio, RMarkdown, Jupyter Notebooks) and programming language (R, Python) used in the course. Primary CALS Competencies: ***thinks with data***, ***manages data***, and ***analyzes data***.
2. **Module 2: Good food gone bad.** There is a food poisoning outbreak among team members. Students use data visualization to determine where the problem lies (fish, not plants) and simulations to determine the root of the problem (fish tank density, not fish tank temperature). Primary CALS Competencies: ***analyzes data***, ***transforms data to information***, and ***communicates visually***.
3. **Module 3: Follow that seal.** The fish tanks need to be restocked, but we want to avoid fishing in places with high leopard seal density, so we track the seals. However, the radio collars on leopard seals are failing and there is a deadly conflict between fish collectors & seals. The collars come from two different manufacturers, but we need to tell how the collars are failing (days to recharge, not signal distance) and how to classify collars of unknown provenance. Primary CALS Competencies: ***analyzes data***, ***transforms data to information***, ***decides with data,*** and ***communicates visually***.
4. **Module 4: March of the penguins.** A new road is needed to access fishing sites with low leopard seal density. There are several possible routes, but we want to avoid crossing through Gentoo penguin nesting grounds. Students will build models to determine predictors of nesting success, first with bootstrapping for confidence intervals, then with linear regression. Primary CALS Competencies: ***analyzes data***, ***transforms data to information***, ***decides with data,*** and ***communicates visually***.
5. **# Module 5: Oil from algae.** Diatom-powered solar cells are being used to to produce biofuel, but there is a lot of variation among the solar cells in biofuel production. We need to find out what causes this variation and get more solar cells to have higher production. The module focuses on model evaluation, potential data challenges, and measuring confidence in model predictions. Primary CALS Competencies: ***analyzes data***, ***transforms data to information***, ***decides with data,*** and ***communicates visually***.

Challenges Details

1. **Module 1**: **Mission Antarctica!** A new effort to establish a permanent, sustainable colony on Antarctica is being launched. Students are introduced to the field of data science and the application (Jupyter Notebooks, RStudio, RMarkdown) and programming language (Python, R) used in the course. Primary CALS Competencies: ***thinks with data***, ***manages data***, and ***analyzes data***.

    1.1. Introduction to data science and the mission to Antarctica
    1.2. R and RStudio (or Jupyter Notebooks & python): adding information to an application for the mission (name, major, etc.). Updating code to change values stored in a variable. Variables are effectively addresses of information on the computer.
    1.3. Assignment 1-1: In-class work on RStudio and R (or equivalent).
    1.4. Storing multiple pieces of information. Need to update applications with additional information, storing in a list. Create a list with current classes, number of units this term. Printing out arrays to screen.
    1.5. Selected for the mission, now investigate data of other team members. Summarize information from a table of 1000 team members, including min, max, and mean age.
    1.6. Assignment 1-2: In-class work on creating lists and summarizing data
    1.7. Investigate Antarctica stations. Extract information from tables, pulling out subsets of data. Performing repeated operations on multiple data points.
    1.8. Designing success: What does success look like for the mission? What data are needed to assess success? Examples include team member weights, doctor visitation data; students would be asked to define what success looks like, what the data are that they would need to collect, and then how one could visualize those data to get a sense of success.
    1.9. Assignment 1-3: In-class work extracting and summarizing data from table with Antarctic station data.

2. **Module 2: Good food gone bad.** There is a food poisoning outbreak among team members. Students use data visualization to determine where the problem lies (fish, not plants) and simulations to determine the root of the problem (fish tank density, not fish tank temperature). Primary CALS Competencies: ***analyzes data***, ***transforms data to information***, and ***communicates visually***.

    2.1. Data visualization: more fish eaten => more trips to doctor; more plants eaten has no effect on number of trips to the doctor, but fish is positively correlated

2.2. Might be due to disease in fish caused by temperature-influenced immunosuppression. Look at all 1000 tanks temperature, but recordings are done in Celsius, but the critical value is in Fahrenheit (students write a function to do conversion). Then visualize data (with histogram) and see that most, if not all tanks are all above critical temperature (typically, low temperatures result in immunosuppression).

2.3. Assignment 2-1: Visualize same data in Kelvin (required Celsius to Kelvin conversion function); also do X-Y scatter plot for plants eaten and visits to doctor (no relationship)

2.4. Back to the drawing board, part 1. There are 1000 fish tanks, but we cannot sample disease rate in all tanks. Sample five tanks and count percent of infected fish (make this high, but with some tanks with low infection, and some with very high infection). We have *a priori* expectation of 2% infection rate, so what is expected infection rate? Run simulation (write function to do simulation) for a five tank sample.

2.5. Back to the drawing board, part 2. Use iteration to generate distribution of 100 mean values for a five-tank sample and compare this to observed mean infection rate. Will indicate observed infection rate is much higher than expected (tie the 2% infection rate to temperature expectations?).

2.6. Assignment 2-2: Do same sampling approach for temperature to show that after sampling five tanks, the temps are not colder than expected. For simulation, assume a priori that 4% will be too cold.

2.7. Look at those five tanks and see if there are predictors, e.g. temperature, day length, tank size, density. Use this for hypothesis generation. Do some data visualization for the five tanks' infection rate and possible predictors.

2.8. Identify the problem, fix it, then do sampling of the five tanks again, run simulation + iteration and see that mean infection rates are within expected values. (summary includes simulations, iterations, data visualization of simulations)

2.9. Assignment 2-3: Do same as 5 & 6, but use size at day 30 as response (set up some a priori size, may need to introduce new sampling method rather than binary probability).

3. **Module 3: Follow that seal.** The fish tanks need to be restocked, but we want to avoid fishing in places with high leopard seal density, so we track the seals. However, the radio collars on leopard seals are failing and there is a deadly conflict between fish collectors & seals. The collars come from two different manufacturers, but we need to tell how the collars are failing (battery life (days), not signal distance (meters)) and how to classify collars of unknown provenance. Primary CALS Competencies: ***analyzes data***, ***transforms data to information***, ***decides with data,*** and ***communicates visually***.

3.1. Group information. Data for 100 radio collars with manufacturer, days to recharge, signal distance (red herring). Need to calculate mean, min, max for each manufacturer (requires some subsetting). Could also write function to do all three calculations and print summary stats. print_stats(data, group). Draw histograms of the two measurements (battery life (days), signal distance (meters)).

3.2. Sampling. i.e. permutation test (but just the permutations here, test is next). Create distributions of means for the two manufacturers for both measurements (days to recharge & signal distance). Output is a distribution of means (draw histogram of permutation results (means) for each manufacturer and each measurement).

3.3. Assignment 3-1: How are the wind turbines doing? Same activities as 3.1-3.2 but this time use different data analyzed outside of class

3.4. Testing group means. Calculate the test statistic - the difference in means between two manufacturers. Then use permutations to see if difference is significant. Introduce p-values and discussion of confidence.

3.5. Joins & standardization. Have a box of trackers without manufacturer label, but may be able to infer which one they are (and not use the bad ones). Have another table with data for original trackers, including weight, antenna length, reflectance (area under curve) (this table has the ID, not the manufacturer). We'll use these data to infer unknowns, but first need to join new data (based on ID), then standardize those measurements $(x - \mu)/\sigma$.

3.6. Assignment 3-2: Why are the wind turbines failing? Same activities as 3.4-3.5 but this time use different data analyzed outside of class

3.7. *k*-nearest neighbors (KNN). Use the joined data for current collars to infer manufacturer for each new collar from box of unknowns. Calculate Euclidean distance for each new collar to each old collar and categorize based on the *k*-nearest collars.

3.8. Sample size. Manufacturer Zeta updated their process, but only sent 5 new collars? Would we be able to detect bad performance. Use original data but only sample five of each and show the effect on p-value (it should get bigger).

3.9. Assignment 3-3: How do we use data to replace the wind turbines? Same activities as 3.7-3.8 but this time use different data analyzed outside of class

4. **Module 4: March of the penguins.** A new road is needed to access fishing sites with low leopard seal density. There are several possible routes, but we want to avoid crossing through Gentoo penguin nesting grounds. Students will build models to determine predictors of nesting success, first with bootstrapping for confidence intervals, then with linear regression. Primary CALS Competencies: ***analyzes data***, ***transforms data to information***, ***decides with data,*** and ***communicates visually***.

Important links:

https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-7998.1990.tb05675.x

https://oceanwide-expeditions.com/to-do/wildlife/gentoo-penguin

4.1. Bootstrapping. How much # of grass tussocks do penguins need for nesting success? Have data from thousands of individual chicks fledged from multiple sites for number of tussocks per square meter and distance to open water. Need to estimate mean and median number of tussocks. Penguins nest on grass tussocks, higher density results in more fledglings. Penguins prefer shallow coastal water and medium sized rocks for nests. Do bootstrapping on sample of data.

4.2. Confidence Intervals. How much confidence is there in the estimate of mean tussocks? Calculate 95%CI and 80%CI.

4.3. Assignment 4-1: Do same calculations in 1 & 2 but for distance to shallow coastal water

4.4. Hypothesis testing. Have data for 100 potential sites in 1971 and 2011. Each site has an amount of grass. Want to know if the amount of grass has increased at each site between dates. Calculate test statistic (difference in amount of grass) run permutation tests and calculate p-value. Why did grass increase between 1971 and 2011? (answer: receding ice sheets). As water recedes, grass tussock density increases

    4.4.1. Data: each row is a site; Columns: Site ID, grass_1971, grass_2011, dist_to_water_1971, dist_to_water_2011.

    4.4.2. Permutation test is on grass_1971 - grass_2011

4.5. Correlation. Need to find place to put road that is least suitable for penguins. There are 100 potential sites. We have data on #nests for 100 sites. Show how correlation coefficient measures relationship. Also show how average stone size has non-linear relationship with #nests.

4.6. Assignment 4-2: Test for differences in distance to open water between 1971 and 2011. Show correlation between #nests and distance to shallow coastal water.

4.7. Simple linear regression. Measure effect of #grass tussocks on number of nests. Want to find site with low number of nests.

4.8. Multiple linear regression. Need better understanding of nesting preferences. Add distance to shallow coastal water as a predictor.

4.9. Assignment 4-3: Include leopard seals in multiple linear regression model.

5. **Module 5: Oil from algae.** Diatom-powered solar cells are being used to to produce biofuel, but there is a lot of variation among the solar cells in biofuel production. We need to find out what causes this variation and get more solar cells to have higher production. The module focuses on model evaluation, potential data challenges, and measuring confidence in model predictions. Primary CALS Competencies: ***analyzes data***, ***transforms data to information***, ***decides with data,*** and ***communicates visually***.

5.1. Measuring correlation. Visualize (X-Y scatter plot) and build simple linear regression model between biofuel production (days to make one [unit of measurement? Liter?] of biofuel) and diatom count/concentration (should be increasing). Show the strength of the relationship as measured by correlation coefficient.

5.2. Making better models. Look at mean squared error and least squares error to actually generate the model for the relationship between diatom count and penguin count (because it is easier to do penguin counts [from the shore] than diatom counts [have to get in a boat and sample] - maybe this is too contrived).

5.3. Assignment 5-1: Potential to use fermentation for diatom biofuel production (still pretty preliminary information about this). Do same as 1 & 2, but ethanol (or some other fermentation by-product) as response variable.

5.4. Non-linearity & heteroscedasticity. Look at residuals for other predictors of diatom counts that may be non-linear (e.g. dissolved nitrogen) or heteroscedastic ([need example of diatom predictor count that is heteroscedastic]).

5.5. Confidence in regression. Run regression model predicting biofuel production from penguin counts. We want to make sure this model is actually going to be useful for making better decisions about where to get water source of diatoms to put in our solar cells. Bootstrap the data to generate confidence in slope.

5.6. Assignment 5-2: We have buoys that are measuring things like temperature and dissolved nitrogen. Run a regression model predicting biofuel production from the non-linear diatom count predictor we identified in 3 (dissolved nitrogen). Bootstrap data to generate confidence in slope and demonstrate how non-linear relationships lead to greater uncertainty (wider confidence intervals).

5.7. Regression prediction. Use our model and some number of new sites to select new places to collect water. Using model from previous session (Biofuel production ~ penguin counts), predict which sites we should sample from for high production. Also make a note of how far to predict - don't make predictions for places too far outside our data; that is, don't try to make a prediction for an X value that is 10 times the maximum observed X.

5.8. *k*-nearest neighbors prediction. Have a lot of other data on current and potential sites (diatom counts *and* surface temperature, dissolved oxygen, salinity, dissolved silica, acidity), some of which are going to be very non-linear or heteroscedastic. Use *k*-nearest neighbors to predict biofuel production in potential sites to identify best new sites. (find the *k*-nearest neighbors for a site based on Euclidean distance to current sites for the predictor variables, the estimate for biofuel production is the mean value from those *k*-nearest neighbors).

5.9. Assignment 5-3: Take those predictors from 6 and pick out some that are "close" to linear with biofuel production. Build multiple linear regression model as in 5. Do the predictions change? If so, how?