title: “hw\_9\_Final\_Project\_Description” author: “Barbara Klein” date: “10/25/2020” output: github\_document

# **Reproducible Data Analysis Final Project**

## A long time ago, In a Body of Water Far Far Away…

Great, now that you think this is a Star Wars parody, I will begin with a brief introduction.

It all started in the summer of 86’… Ok ok Hoops McCann isn’t telling the story either, so we’ll start for real this time. Back in 2004, there was a group of scientists who set out to gather information on both eukaryotic and prokaryotic microbial communities. Their particular location involved penguins, ice bergs, and traveling through other bone chilling bodies of water; I’m talking about the Ross Sea, Antarctica. These scientists were on a mission to extrapolate a relationship between UV light and microbial community composition at various depths within the Southern Hemisphere Summer. For a long-term project during my undergrad years, I had graciously received the cruise’s master spreadsheet to sort through and organize. Specific items on this spreadsheet include station ID, latitude, longitude, temperature, depth, and many other measurements. Knowing that oceanic research cruises are driven by sleep deprived scientists, and having just been handed oodles of spreadsheets and archived data to organize, I came to find that certain measurements or pieces of data were missing or unrecoverable. Being that this data may be lost at sea forever, here we set out to fill in those missing data points with the help of the almighty statistical powers that be, in an R Studio script.

During the research cruise, the ships log tracked the latitude and longitude of each sampling site or station. These station locations hold various measurements taken at depths ranging from the surface (0 meters), all the way down to 250 meters. The measurements that were taken consisted of chlorophyll a, bacterial direct counts, leucine incorporation, bacterial production, nitrate and nitrite concentration, phosphate concentration, DSI, ammonium concentration, salinity, and temperature. While some of the numerical measurements were extrapolated from the ships’ log, the biogeochemical measurements were gathered from either preserved DNA filters, or water collection samples taken via CTD Rosette. Some of these depths didn’t have certain measurements recorded and therefore couldn’t piece together a full and complete story.

These stations correspond to specific latitude/longitudinal coordinates which will be deemed x\_i. This indicates the literal point in space of a single station (site) at which sampling occured. Of course we will take depth into account, which will be y\_i. Within my dataset, there are plenty of missing values at each depth. Using the “Kriging Model”, we will be predicting missing points using geostatistics, and creating a map, and annotating an R Markdown file enough for someone completely new (say an Undergrad) to R, to recreate a graph similar to ours.

Lovell will be creating the map, annotating the commands and what their purpose is, and fitting this map to be universally used. I (Barb) will be laying out the statistical modeling characters and equations, annotating those characters, and also making them in such a way as to be universally used. We would like to emphasize, there will be a lot of annotating throughout the Markdown file in order to thoroughly explain each detail in its’ entirety so this Markdown file can truly be used at the beginner level.