Homework 3

Background: diatoms and paleoclimatology

Diatoms are one type of phytoplankton – they are photosynthetic algae that function as primary producers in aquatic ecosystems. Diatoms are at the bottom of the food web: they are consumed by filter feeders, like clams, mussels, and many fish, which are in turn consumed by larger organisms like scavengers and predators and, well, us. As a result, changes in the composition of diatom species in marine ecosystems have ripple effects that can dramatically alter overall community structure in any environment of which marine life forms a part.

Diatoms have glass bodies. As a group of organisms, they display a great diversity of body shapes, and many are quite elaborate. The image below, taken from a Scientific American article, shows a small sample of their shapes and structures.

Because they are made of glass, diatoms preserve extraordinarily well over time. When they die, their bodies sink and form part of the sediment. Due to their abundance, there is a sort of steady rain of diatoms forming part of the sedimentation process, which produces sediment layers that are dense with diatoms.

Sedimentation is a long-term process spanning great stretches of time, and the deeper one looks in sediment, the older the material. Since diatoms are present in high density throughout sedimentation layers, and they preserve so well, it is possible to study their presence over longer time spans – potentially hundreds of thousands of years.

A branch of paleoclimatology is dedicated to studying changes in biological productivity on geologic time scales, and much research in this area has involved studying the relative abundances of diatoms. In this assignment, you'll do just that on a small scale and work with data from sediment cores taken in the gulf of California at the location indicated on the map:

The data is publicly available: > Barron, J.A., *et al.* 2005. High Resolution Guaymas Basin Geochemical, Diatom, and Silicoflagellate Data. IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series # 2005-022. NOAA/NGDC Paleoclimatology Program, Boulder CO, USA.

In this assignment, you'll use the exploratory techniques we've been discussing in class to analyze the relative abundances of diatom taxa over a time span of 15,000 years. This will involve practicing the following:

- data import and preprocessing
- graphical techniques for visualizing distributions
- multivariate analysis with PCA

Diatom data

The data are diatom counts sampled from evenly-spaced depths in a sediment core from the gulf of California. In sediment cores, depth correlates with time before the present – deeper layers are older – and depths are typically chosen to obtain a desired temporal resolution. The counts were recorded by sampling material from sediment cores at each depth, and examining the sampled material for phytoplankton cells. For each sample, phytoplankton were identified at the taxon level and counts of diatom taxa were recorded along with the total number of phytoplankton cells identified. Thus:

- The **observational units** are ***sediment samples***.
- The **variables** are ***depth (age), diatom abundance counts, and the total number of identified phytoplankton***. Age is inferred from radiocarbon.
- One **observation** is made at ***each depth*** from 0cm (surface) to 13.71 cm.

The table below provides variable descriptions and units for each column in the dataframe.

Variable	Description	Units
Depth	Depth interval location of sampled material in sediment core	Centimeters (cm)
Age	Radiocarbon age	Thousands of years before present (KyrBP)
A_curv	Abundance of *Actinocyclus curvatulus*	Count (n)
A_octon	Abundance of *Actinocyclus octonarius*	Count (n)
ActinSpp	Abundance of *Actinoptychus* species	Count (n)
A_nodul	Abundance of *Azpeitia nodulifer*	Count (n)

Variable	Description	Units
CocsinSpp	Abundance of *Coscinodiscus* species	Count (n)
CyclotSpp	Abundance of *Cyclotella* species	Count (n)
Rop_tess	Abundance of *Roperia tesselata*	Count (n)
StephanSpp	Abundance of *Stephanopyxis* species	Count (n)
Num.counted	Number of diatoms counted in sample	Count (n)

The cell below imports the data.

```
# import diatom data
diatoms_raw = read_csv('data/barron-diatoms.csv')
diatoms_raw
```

The data are already in tidy format, because each row is an observation (a set of measurements on one sample of sediment) and each column is a variable (one of age, depth, or counts).

1. NAs are present
2. The number of individuals counted in each sample varies by a lot from sample to sample.

For example, here, we see the fraction of missing data for each taxa

```
diatoms_raw |> summarize(across(A_curv:StephanSpp, \(x) mean(is.na(x))))
```

Let's address these issues before conducting initial explorations.

The NAs are an artefact of the data recording – if *no* diatoms in a particular taxa are observed, a `-` is entered in the table (you can verify this by checking the .csv file). In these cases the value isn't missing, but rather zero. These entries are parsed by pandas as NaNs, but they correspond to a value of 0 (no diatoms observed).

Question 1: Filling NAs

Use `replace_na` to replace all NAs by zeros, and store the result as `diatoms_mod1`. Store rows 94 and 95 of the resulting dataframe as `diatoms_mod1_examplerows` and display these rows.

```r
diatoms1 <- # SOLUTION HERE

diatoms_mod1_examplerows <- diatoms1 |> slice(94:95)
print(diatoms_mod1_examplerows)
```

Since the total number of phytoplankton counted in each sample varies, the raw counts are not directly comparable – *e.g.*, a count of 18 is actually a *different* abundance in a sample with 200 individuals counted than in a sample with 300 individuals counted.

For exploratory analysis, you'll want the values to be comparable across rows. This can be achieved by a simple transformation so that the values are *relative* abundances: *proportions* of phytoplankton observed from each taxon.

Question 2: Counts to proportions

Convert the counts to proportions by dividing the diatom counts by the total number of phytoplankton recorded in the `Num.counted` column. You can do this succinctly using `across` function. Note that the proportions of diatoms will not sum to 1 since `Num.counted` includes all phytoplankton counts, many of which are not diatoms but other groups.

In addition multiply the Age variable by -1. This will be useful since in the raw dataset, larger values correspond to older samples. As such, we'll want those values to appear as "smaller" than newer samples in any plots.

Save the data frame with the proportions and negative ages to a new variable called `diatoms2`. Print the first few rows.

```r
diatoms2 <- # SOLUTION

print(diatoms2)
```

Take a moment to think about what the data represent. They are relative abundances over time; essentially, snapshots of the community composition of diatoms over time, and thus information about how ecological community composition changes.

Before diving in, it will be helpful to resolve two matters:

1. How far back in time do the data go?
2. What is the time resolution of the data?

Question 3: Time span

What is the geological time span covered by the data? Compute the age of the oldest and newest samples. Express the ages in years.

```
oldest_age <- # SOLUTION
newest_age <- # SOLUTION
print(c(oldest_age, newest_age))
```

Question 4: Time resolution

(i) How are the observations spaced in time?

Follow these steps:

1. Use `diatoms2` and the `arrange` function to sort the `Age` column in descending order from smallest (most negative) to largest. Then create a new column called `Gap` which is the gap in ages between consecutive samples. You can do this using the `diff` function. For example `diff(c(2, 7, 8, 10))` returns `5, 1, 2`. Note that the length of the vector returned by `diff` is 1 less than the length of the input, since there is no gap that we can compute for the first observation. To create the `Gap` column, set the first value of `Gap` to NA and the rest to the result of `diff` called on the Age column. Save the updated data to `diatoms3`
2. Make a simple count histogram with bins of width 0.02 (20 years). Store this as `gap_histogram` and display the figure.
    - Label the x axis 'Time step between consecutive samples'
    - Put the y axis on a square root scale so that bins with just one observation are more visible. You can do this by setting y variable to `after_stat(sqrt(count))`.

(ii) What is the most common time gaps, specified in years?

Write your answer based on the count histogram.

Type your answer here, replacing this text.

```
# store differences
diatoms3 <-  # SOLUTION

gap_histogram <- ## Plot Histogram
gap_histogram
```

Your answer here replacing this text

Univariate explorations

To begin, you'll examine the variation in relative abundance over time for the eight individual taxa, one variable at a time.

Here are some initial questions in this spirit that will help you to hone in and develop more focuesed exploratory questions:

- Which taxa are most and least abundant on average over time?
- Which taxa vary the most over time?

These can be answered by computing simple summary statistics for each column in the diatom data.

Question 5: Summary statistics

Use `summarize(across(A_curve:StephanSpp...))` to summarize the mean and standard deviation of the relative diatom abundances in `diatoms3`. The output should have columns names of the form **(diatom_name)_(aggregation_type)**, e.g. `A_curv_mean` or `ActinSpp_sd`. Save the result in `diatom_summary`.

Hint: look at the help page for `across` to see examples of summarizing each column using multiple functions supplied in a list.

Confirm you have the right answer by ensuring the first few observations match the image below:

A tibble: 1 × 16

A_curv_mean <dbl>	A_curv_sd <dbl>	A_octon_mean <dbl>	A_octon_sd <dbl>	ActinSpp_mean <dbl>
0.02898894	0.01860215	0.0182575	0.01646515	0.1359001

1 row | 1-5 of 16 columns

```
diatom_summary <- # SOLUTION

# print the dataframe
diatom_summary
```

Question 6

It will be easier to determine which taxa are most/least abundant by reorganizing the dataframe.

```
diatom_summary2 <- diatom_summary |>
    pivot_longer(
        cols = everything(),
        names_pattern = "(.+)_(mean|sd)$",
        names_to = c("taxon", "statistic"),
        values_to = "value"
    ) |> pivot_wider(names_from=statistic, values_from=value)

print(diatom_summary2)
```

(ii) Based on the data frame above, answer the following questions.

1. Which taxon is most abundant on average over time?
2. Which taxon is least abundant on average over time?
3. Which taxon varies most in relative abundance over time?

Type your answer here, replacing this text.

Now that you have a sense of the typical abundances for each taxon (measured by means) and the variations in abundance (measured by standard deviations), you'll dig in a bit further with time series plots.

Question 7: Distribution of diatom abundances over time

Use the `diatoms3` dataframe to make a line plot of Age vs relative abundance faceted by taxon. Set `scales="free` in the call to `facet_wrap` to give each facet it's own y-scale. You may need to `pivot_longer` on `diatoms3` to make plotting easier.

```
diatom_abundances_plot <- ## SOLUTION

diatom_abundances_plot
```

(ii) Answer the following questions in a few sentences.

- Which values are common?
- Which values are rare?

- How spread out are the values?
- Are values spread evenly or irregularly?

Type your answer here, replacing this text.

There was a transition between geological epochs during the time span covered by the diatom data. The oldest data points in the diatom data correspond to the end of the Pleistocene epoch (ice age), at which time there was a pronounced warming (Late Glacial Interstadial, 14.7 - 12.9 KyrBP) followed by a return to glacial conditions (Younger Dryas, 12.9 - 11.7 KyrBP).

This fluctuation can be seen from temperature reconstructions. Below is a plot of sea surface temperature reconstructions off the coast of Northern California. Data come from the following source:

Barron *et al.*, 2003. Northern Coastal California High Resolution Holocene/Late Pleistocene Oceanographic Data. IGBP PAGES/World Data Center for Paleoclimatology. Data Contribution Series # 2003-014. NOAA/NGDC Paleoclimatology Program, Boulder CO, USA.

The shaded region indicates the time window with unusually large flucutations in sea surface temperature; this window roughly corresponds to the dates of the climate event.

```
# import sea surface temp reconstruction
seatemps <- read_csv('data/barron-sst.csv') |>
  mutate(Age = -1*Age)

seatemps |> ggplot(aes(x=Age, y=SST)) +
  geom_line() +
  annotate("rect", xmin=-15, xmax=-11, ymin=-Inf, ymax=Inf,
           fill="orange", alpha=0.25) +
  xlab("Thousands of years before present") +
  theme_bw(base_size=16)
```

Question 8: Conditional distributions of relative abundance

Does the distribution of relative abundance of *Azpeitia nodulifer* differ when variation in sea temperatures was higher (before 11KyrBP)?

(i) Plot separate kernel density estimates to show the distribution of relative abundances both before and after 11KyrBP.

(ii) Does the distribution seem to change between epochs? If so, how?

Answer based on the figure in a few sentences.

Type your answer here, replacing this text.

```
## YOUR ANSWER HERE
```

Type your answer here

Question 9

Let's look at more fine grained question. Does the abundance of *A. nodulifer* change with sea surface temperature (SST)? We'd like to plot *A. nodulifer* against SST but the variables are in different dataframes. We could try to merge the data frames but we have a problem: SST isn't measured at the exact years that our diatoms are measured.

We need some way to predict SST at the ages for which we have diatom measuremnts. We can use LOESS smoothing! Here, we dont' want to smooth too much since we want a good prediction of SST.

The code below fits a LOESS line of SST on Age. We then predict SST at the ages available in the diatom data

```
## Fit a smooth prediction to sea surface temperature
loess_fit <- loess(SST ~ Age, data=seatemps, span=0.1)

## Predict sea surface tempperature in the diatom data and add the variable
diatoms3$sst_pred <- predict(loess_fit, newdata = diatoms3$Age)
```

Make a plot showing the diatom predicted sea surface temperature vs the diatom abundance, faceting by the Taxon. With `facet_wrap`, use `scales="free"` to give every plot it's own y-limits. `*Hint* you may need to`pivot_longer`on`diatom3‘ first.

```
sst_v_diatoms_plot <- diatoms3 |>  ## SOLUTION

sst_v_diatoms_plot
```

Which taxon appears to have the strongest relationship with sea surface temperature?

Type your answer here, replacing this text

Visualizing community composition with PCA

So far you've seen that the abundances of one taxon – *Azpeitia nodulifer* – change markedly before and after a shift in climate conditions. In this part you'll use PCA to compare variation in community composition among *all* eight taxa during the late Pleistocene and Holocene epochs.

Question 10: Pairwise correlations in relative abundances

(i) Compute the pairwise correlations between relative abundances and make a plot visualizing the correlation matrix.

Use `diatoms3` to select all diatoms and compute a correlation matrix between them. Use `corrplot` to plot the resulting correlation matrix, setting the argument `diag=FALSE`. You will probably need to install the `corrplot` package first and load it.

```
# Create correlation matrix and plot it
```

(ii) How does *A. nodulifer* seem to vary with the other taxa, if at all?

Answer in a few sentences based on the heatmap.

Type your answer here, replacing this text.

Question 11: Computing and selecting principal components

Here you'll run PCA and check the variance ratios to select an appropriate number of principal components. The parts of this question correspond to the individual steps in this process.

(i) Run `prcomp` and look at the variance ratios of the individual principal components.

For PCA it is usually recommended to center and scale the data. Run PCA on the diatom data (drop Age and Depth) and run `prcomp` setting `scale=TRUE`. Compute *all 8* principal components.

Then create a dataframe called `pc_vars` with the variance information by following these steps: - Create a variable named `pc` with values 1:8 (the index of the PC) - Create a variable called `var`, the variance of associated with each pc (the square of the `sdev` variable from `prcomp`) - Create a variable called `var_explained` which is the proportion of variance explained by that PC - Create variable called `cum_var_explained`, the cumulative variance explaned by the first PCs. Do so using the `cumsum` function applied to `var_explained`.

For this part you do not need to show any specific output.

```
diatom_pca <- diatoms3 |>
 select(A_curv:StephanSpp) |>  # Remove non-species variables
 prcomp(scale=TRUE)

pca_var <- tibble(pc = 1:8,
  eval = diatom_pca$sdev^2) |>
```

```
  mutate(var_explained = eval^2/sum(eval^2),
         cum_var_explained = cumsum(var_explained))
```

(ii) Plot the variance explained by each PC.

Create two side-by-side plots using patchwork:

1. Plot the `var_explained` variable as a barplot using `geom_col`.
2. Plot the `cum_var_explained` variable as a lineplot. Make sure the y-limits are set to the interval [0, 1].

How many PCs are needed to explain at least 75% of the total variation?

Type your answer here, replacing this text

```
var_exp_plot <- ## plot
cum_var_exp_plot <- ## plot

var_exp_plot + cum_var_exp_plot
```

Type your answer here, replacing this text.

(3) How many PCs should be used?

Propose an answer based on the variance explained plots and indicate how much total variation your proposed number of components capture jointly.

Type your answer here, replacing this text.

Now that you've performed the calculations for PCA, you can move on to the fun/difficult part: figuring out what they say about the data.

The first step in this process is to examine the loadings. Each principal component is a linear combination of the relative abundances by taxon, and the loadings tell you *how* that combination is formed; the loadings are the linear combination coefficients, and thus correspond to the weight of each taxon in the corresponding principal component. Some useful points to keep in mind:

- a high loading value (negative or positive) indicates that a variable strongly influences the principal component;
- a negative loading value indicates that
  - increases in the value of a variable *decrease* the value of the principal component
  - and decreases in the value of a variable *increase* the value of the principal component;

- a positive loading value indicates that
  - increases in the value of a variable *increase* the value of the principal component
  - and decreases in the value of a variable *decrease* the value of the principal component;
- similar loadings between two or more variables indicate that the principal component reflects their *average*;
- divergent loadings between two sets of variables indicates that the principal component reflects their *difference*.

Question 12: Interpreting component loadings

(i) Make barplots showing the component loadings for the first two principle components: - Create a data frame called `loadings`. The columns should be the first two columns of the `rotation` variable returned by `prcomp`. Add a variable called `Taxon` which reports the taxon name for each of the 8 variables.

- Make two barplot using `geom_col` showing the loadings for the first two components. Stack them on top of one another using patchwork, with the loadings for the first PC on top.

```
library(patchwork)
# Get loadings and scores
loadings <- ## SOLUTION

pc1_loadings <-  # SOLUTION
pc2_loadings <-  # SOLUTION
pc1_loadings / pc2_loadings
```

(ii) Interpret the first principal component.

In a few sentences, answer the following questions. 1. Which taxa are up-weighted and which are down-weighted in this component? 2. How would you describe the principal component in context (*e.g.*, average abundance among a group, differences in abundances, etc.)? 3. How would you interpret a larger value of the PC versus a smaller value of the PC in terms of diatom communnity composition?

Type your answer here, replacing this text.

(iii) Interpret the second principal component.

Answer the same questions for component 2.

Type your answer here, replacing this text.

Recall that there was a shift in climate around 11,000 years ago, and *A. nodulifer* abundances seemed to differ before and after the shift.

You can now use PCA to investigate whether not just individual abundances but *community composition* may have shifted around that time. To that end, let's think of the principal components as 'community composition indices':

- consider PC1 a nodulifer/non-nodulifer community composition index; and
- consider PC2 a complex community composition index.

A pattern of variation or covariation in the principal components can be thought of as reflecting a particular ecological community composition dynamic – a way that community composition varies throughout time. Here you'll look for distinct patterns of variation/covariation before and after 11,000 years ago via an exploratory plot of the principal components.

Question 13: Visualizing community composition shift

(i) Construct a scatterplot of PC1 and PC2 scores by epoch.

Examine the the centered and scaled data projected onto the first two component directions. This sounds a little more complicated than it is – all that means is compute the values of the principal components . This is stored in the `x` variable of the `prcomp` output. Create a dataframe called `projected_data` containing just the first two principal components scores as two columns named `PC1` and `PC2`. Add an additional column corresponding to the Age variable.

Construct a scatterplot of the principal components with observations colored according to whether they occurred in the Pleistocene or Holocene epoch.

```
# Visualize results
scores <- #  SOLUTION

pc_scores_plot <- # SOLUTION

pc_scores_plot
```

(ii) Comment on the plot: does there appear to be any change in community structure?

Answer in a few sentences.

Type your answer here, replacing this text.

Question 14: Multi-panel visualization

Sometimes it's helpful to see marginal distributions together with a scatterplot. Follow the steps below to create a figure with marginal density estimates appended to the projected scatter from the previous question.

To do so, use the `ggMarginal` function from the `ggExtra` package. Set `groupFill=TRUE` in the arguments and append to the scatter plot.

```
# SOLUTION HERE
```

Communicating results

Question 15: Summary

Take a moment to review and reflect on the results of your analysis in the previous parts. Think about how you would describe succinctly what you've learned from the diatom data. Write a brief paragraph (3-5 sentences) that addresses the following questions by referring to your results above.

- How would you characterize the typical ecological community composition of diatom taxa before and after 11,000 years ago?
  - *Hint*: focus on the side and top panels and the typical values of each index in the two time periods.
- Does the variation in ecological community composition over time seem to differ before and after 11,000 years ago?
  - *Hint*: focus on the shape of data scatter.

Type your answer here, replacing this text.

Question 16: Further work

What more might you like to know, given what you've learned? Pose a question that your exploratory analysis raises for you.

Answer

Type your answer here.

Submission

1. Render your quarto file as a pdf.
2. Upload both the pdf and the .qmd file to Gradescope.
